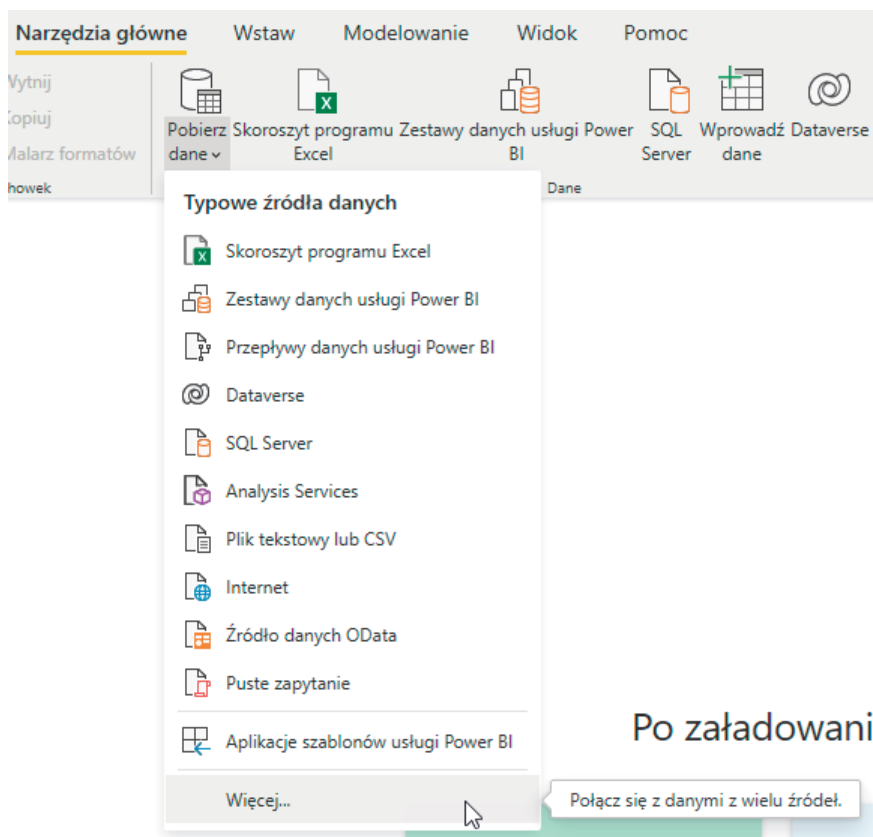
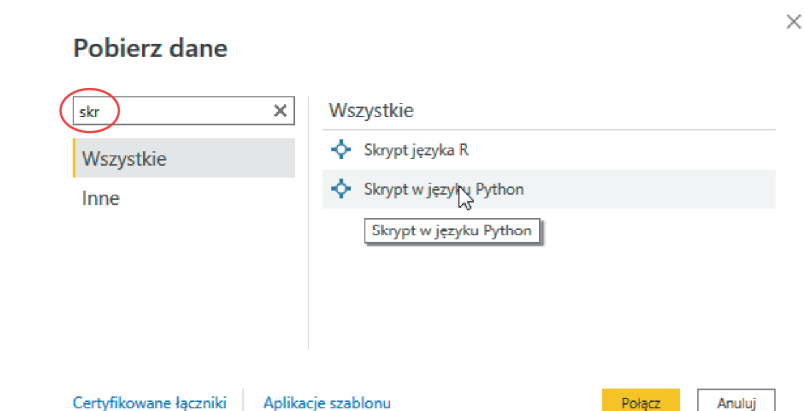


Kolorowe wykresy i zrzuty z polskiego wydania książki „Dodaj mocy Power BI!”

Rozdział 1. Gdzie i jak używać w usłudze Power BI skryptów języka R i Python?



Rysunek 1.1. Przeglądanie dodatkowych łączników pozwalających na załadowanie danych



Rysunek 1.2. Skrypty języka R i języka Python w oknie Pobierz dane



Rysunek 1.3. Okno z edytorem skryptów Pythona

Skrypt języka R



Skrypt

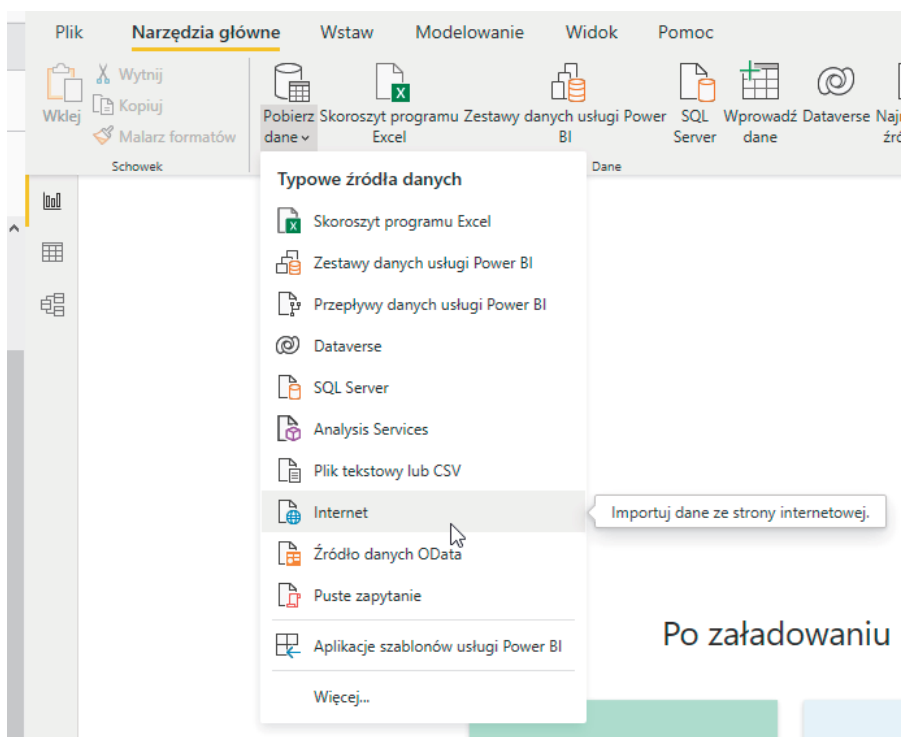
Język R nie został zainstalowany. Aby zainstalować język R i rozpocząć wykonywanie skryptów, wybierz pozycję Opcje i ustawienia > Opcje > Skrypty języka R.

[Jak zainstalować język R](#)

OK

Anuluj

Rysunek 1.4. Okno z edytorem skryptów R



Rysunek 1.5. Aby zaimportować dane ze strony internetowej, wybierz łącznik Internet

Z sieci Web

☒ Podstawowy ☐ Zaawansowane

Adres URL

bit.ly/iriscsv

OK

Anuluj

Rysunek 1.6. Importowanie danych tęczówek z Internetu

bit.ly/iriscsv

Pochodzenie pliku: 65001: Unicode (UTF-8) | Ogranicznik: Przecinek | Wykrywanie typu danych: Na podstawie pierwszych 200 wiersz...

Column1	Column2	Column3	Column4	Column5
sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3	1.4	0.1	setosa
4.3	3	1.1	0.1	setosa
5.8	4	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa

Dane w podglądzie zostały obcięte z powodu ograniczeń rozmiaru.

Wyodrębni tabelę przy użyciu przykładów | Załaduj | **Przekształć dane** | Anuluj

Rysunek 1.7. Podgląd zaimportowanych danych

Bez tytułu — Edytor Power Query

Przekształć

Ustawienia zapytania

Właściwości

ZASTOSOWANE KROKI

Zmieniło typ

Rysunek 1.8. Narzędzia skryptowe języka R i Python w programie Power Query Editor

Uruchom skrypt języka Python

Wprowadź skrypty języka Python do edytora, aby przekształcać i kształtować dane.

Skrypt

```
# Element 'dataset' zawiera dane wejściowe dla skryptu
```

Skrypt jest uruchamiany przy użyciu następującej instalacji języka Python: C:\Users\User\Anaconda3.

Aby skonfigurować ustawienia i zmienić instalację języka Python, która ma być uruchamiana, przejdź do obszaru Opcje i ustawienia.

OK

Anuluj

Rysunek 1.9. Edytor skryptów w Pythonie

Uruchom skrypt języka R

Wprowadź skrypty języka R do edytora, aby przekształcać i kształtować dane.

Skrypt

```
# Element 'dataset' zawiera dane wejściowe dla skryptu
```

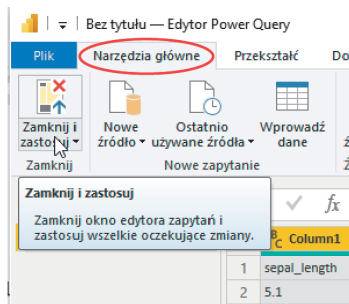
⚠ Język R nie został zainstalowany. Aby zainstalować język R i rozpocząć wykonywanie skryptów, wybierz pozycję Opcje i ustawienia > Opcje > Skrypty języka R.

[Jak zainstalować język R](#)

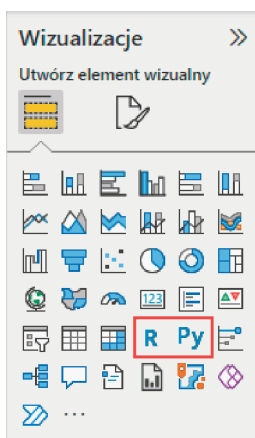
OK

Anuluj

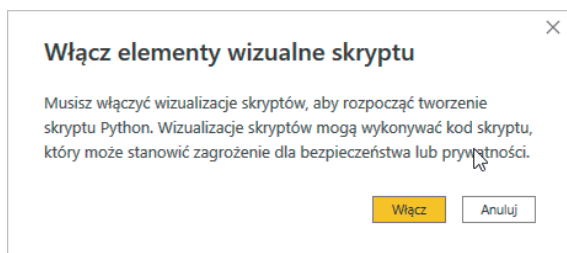
Rysunek 1.10. Edytor skryptów języka R



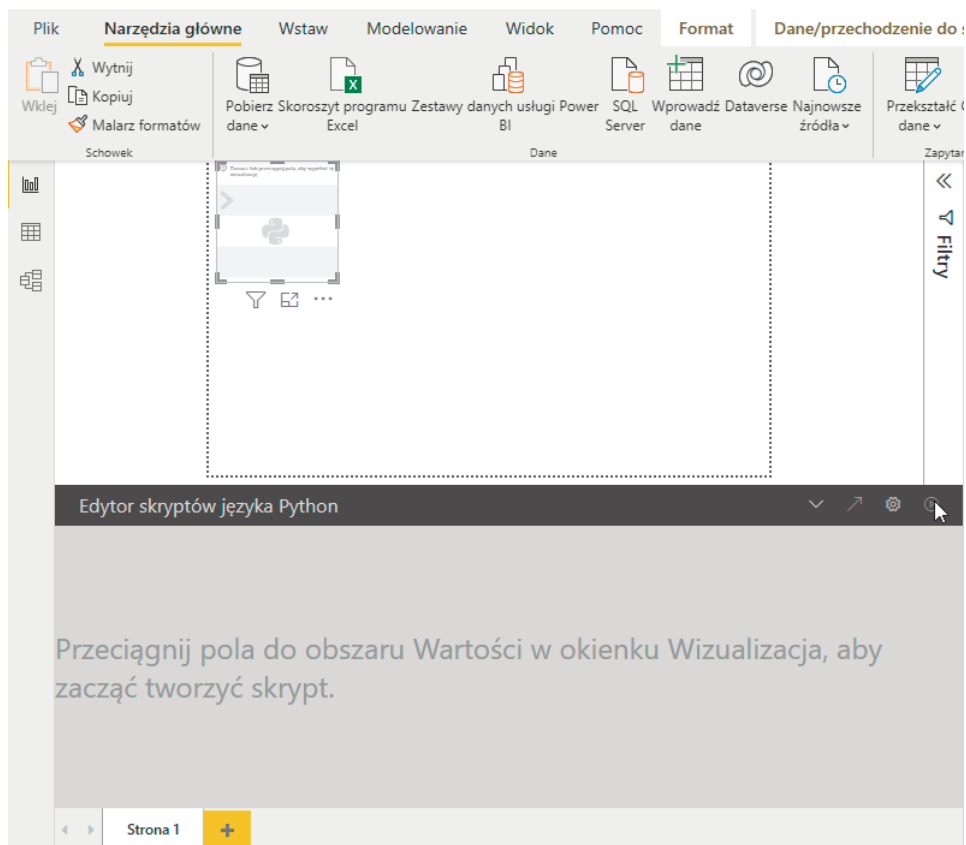
Rysunek 1.11. Aby zaimportować dane z tabeli, kliknij Zamknij i zastosuj



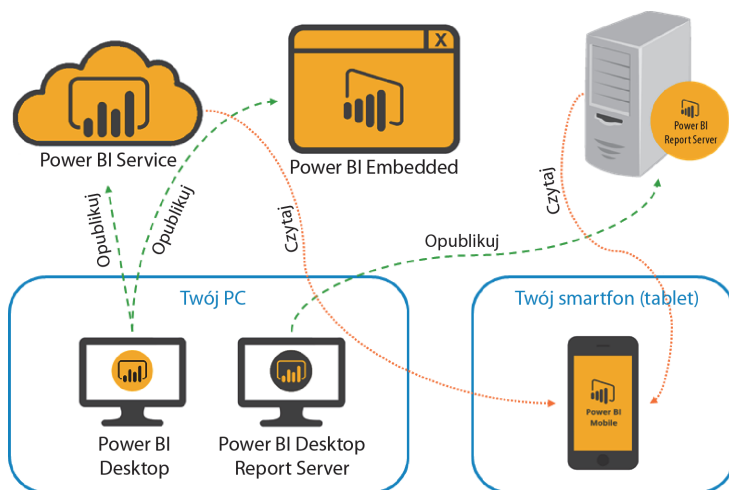
Rysunek 1.12. Wizualizacje za pomocą skryptów w językach R i Python



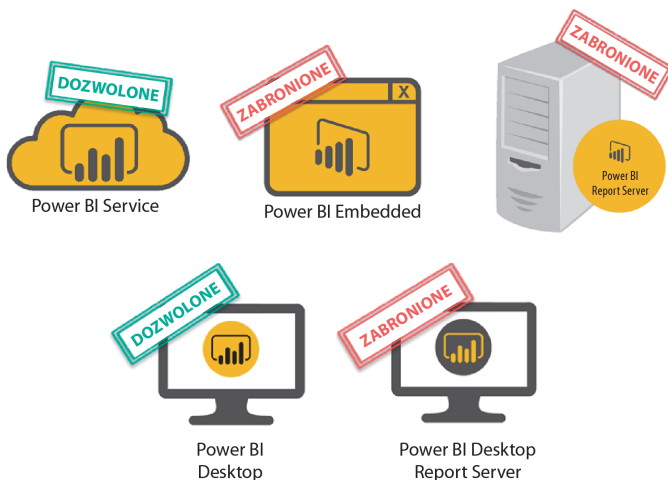
Rysunek 1.13. Włączenie wykonywania kodu skryptów



Rysunek 1.14. Układ ekranu tworzenia wizualizacji w Pythonie

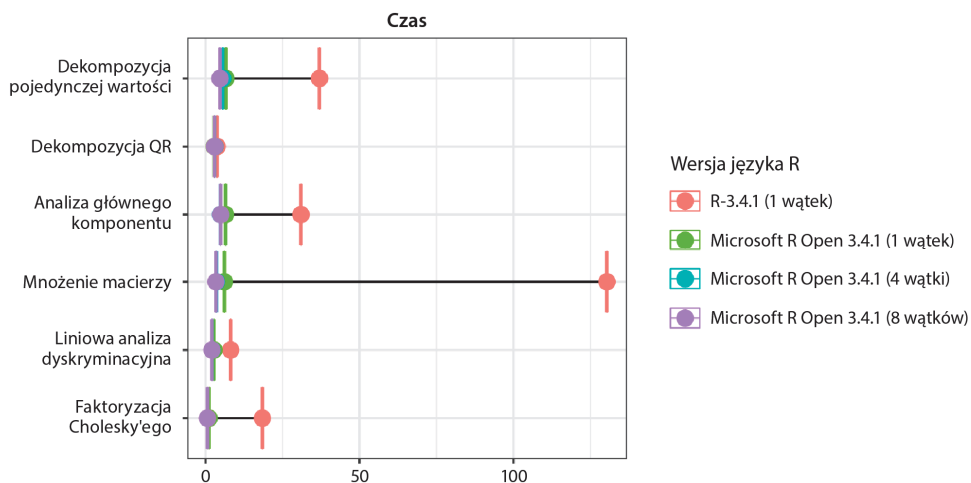


Rysunek 1.15. Interakcje między produktami usługi Power BI

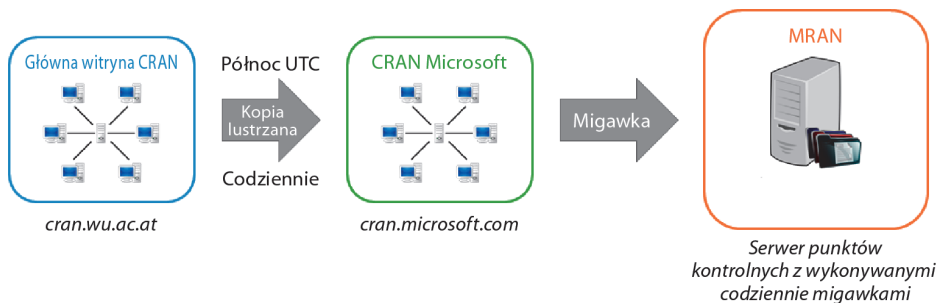


Rysunek 1.16. Zgodność produktów usługi Power BI z językami R i Python

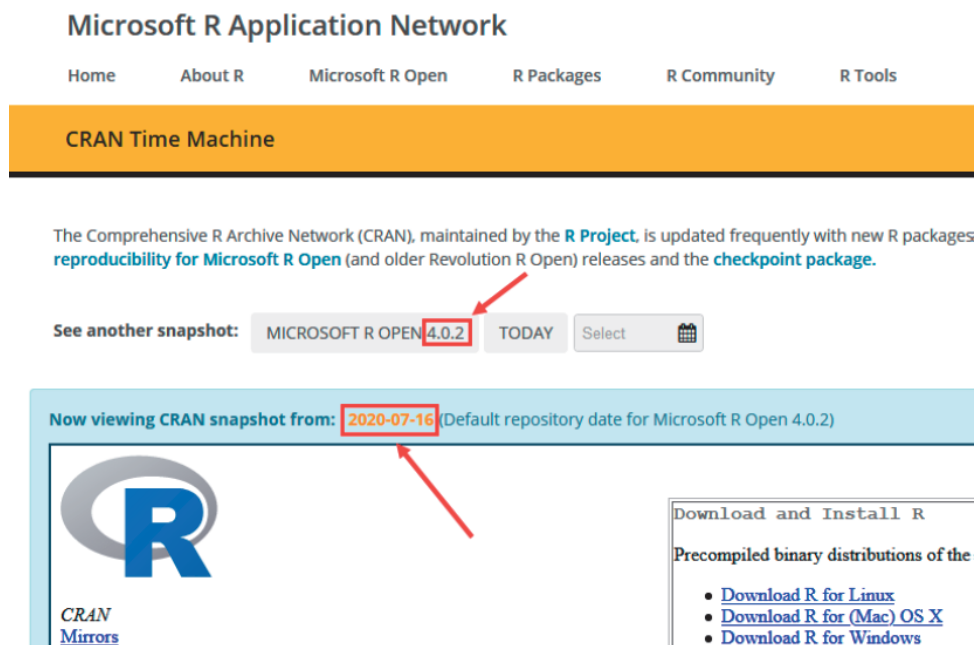
Rozdział 2. Konfigurowanie języka R na potrzeby usługi Power BI



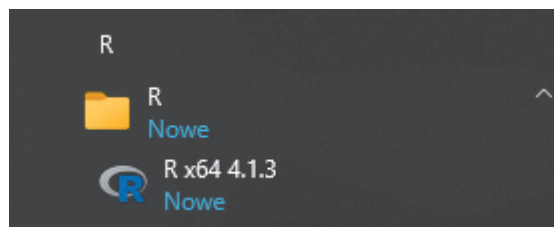
Rysunek 2.1. Całkowity czas, jaki upłynął dla każdego z testów porównawczych na tej samej maszynie



Rysunek 2.2. Proces utrwalania codziennej migawki repozytorium CRAN w repozytorium MRAN



Rysunek 2.3. Proces utrwalania codziennej migawki repozytorium CRAN w repozytorium MRAN



Rysunek 2.4. Program rGUI zainstalowany wraz z dystrybucją CRAN R

Requirements and Limitations of R packages

There are a handful of requirements and limitations for R packages:

- Current R runtime: Microsoft **R 3.4.4**

Rysunek 2.5. Wersja MRO używana przez usługę Power BI

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R and the [R](#)

- [Patches to this release are incorporated in the r-patched snr](#)
- [A build of the development version \(which will eventually](#)
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current [CRAN MIRROR](#) [bin/windows/base/release.html](#).

Rysunek 2.6. Link do poprzednich wersji dystrybucji CRAN R dla systemu Windows

See another snapshot:

MICROSOFT R OPEN 3.4.4

TODAY

Select



Now viewing CRAN snapshot from: **2018-04-01** (Default repository date for Microsoft R Open 3.4.4)

Rysunek 2.7. Domyślna data repozytorium dla MRO 3.4.4

```
1 # Things you might want to change
2
3 # options(papersize="a4")
4 # options(editor="notepad")
5 # options(pager="internal")
6
7 # set the default help type
8 # options(help_type="text")
9 # options(help_type="html")
10
11 # set a site library
12 # .Library.site <- file.path(chartr("\\", "/", R.home()), "site-library")
13
14 # set a CRAN mirror
15 # local({r <- getOption("repos")
16 #   r["CRAN"] <- "http://my.local.cran"
17 #   options(repos=r)})
18
19 # Give a fortune cookie, but only to interactive sessions
20 # (This would need the fortunes package to be installed.)
21 # if (interactive())
22 #   fortunes::fortune()
```

Rysunek 2.8. Domyślna zawartość pliku Rprofile.site

RStudio Desktop

Open Source License

Free

DOWNLOAD

[Learn more](#)

Rysunek 2.9. Przejdź do strony pobierania środowiska RStudio

RStudio Desktop 1.4.1103 - [Release Notes](#)

- 1.** Install R. RStudio requires R 3.0.1+.
- 2.** Download RStudio Desktop. Recommended for your system:

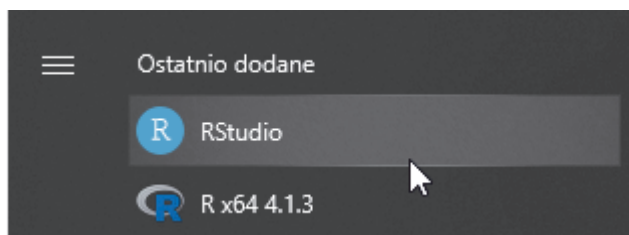


DOWNLOAD RSTUDIO FOR WINDOWS

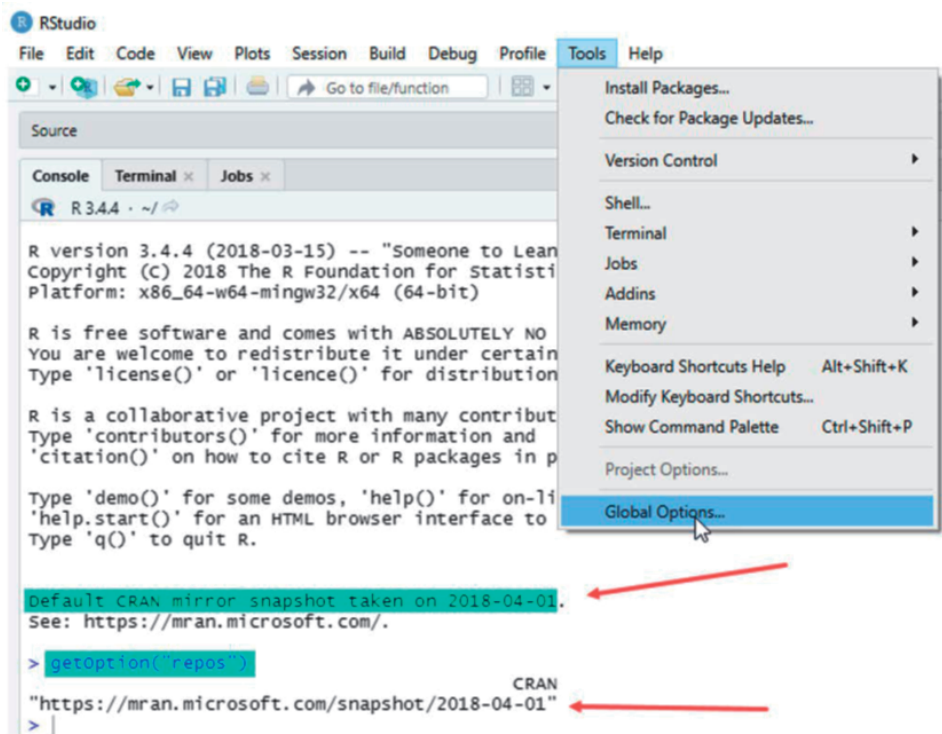
1.4.1103 | 156.96MB

Requires Windows 10/8/7 (64-bit)

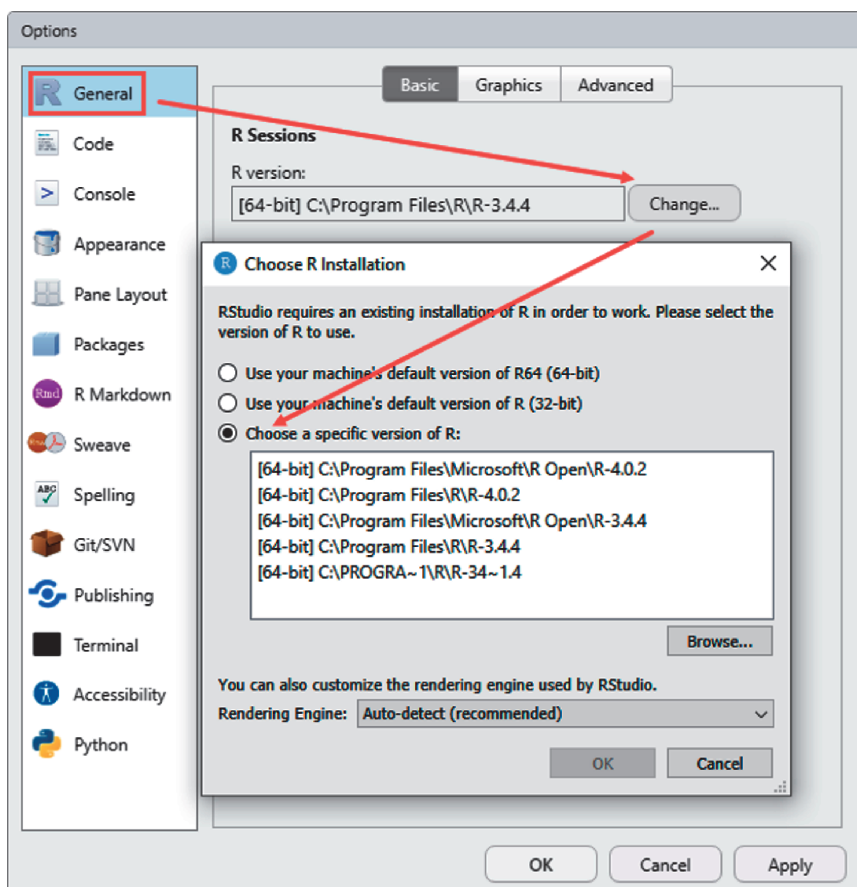
Rysunek 2.10. Pobieranie środowiska RStudio dla systemu Windows



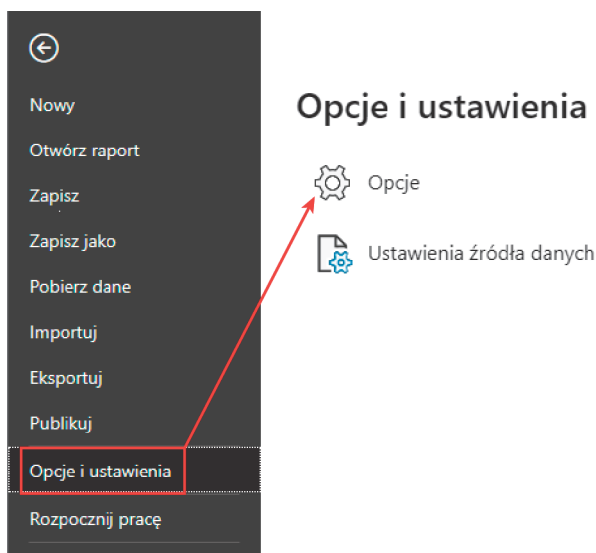
Rysunek 2.11. Uruchom środowisko RStudio z menu Start



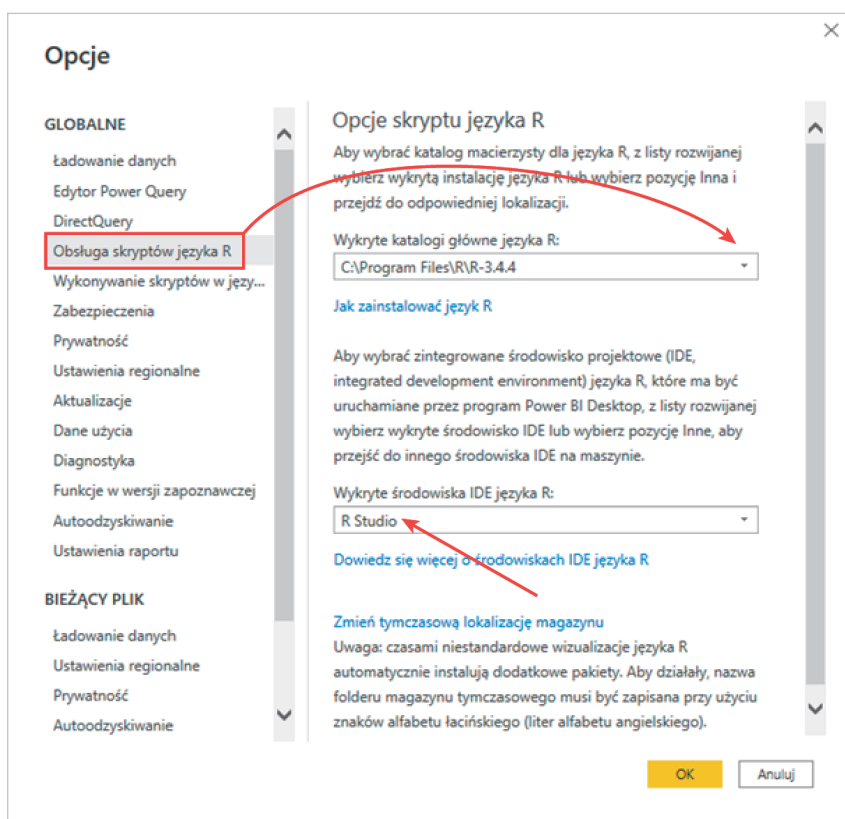
Rysunek 2.12. Domyślny silnik języka R wybrany w programie RStudio



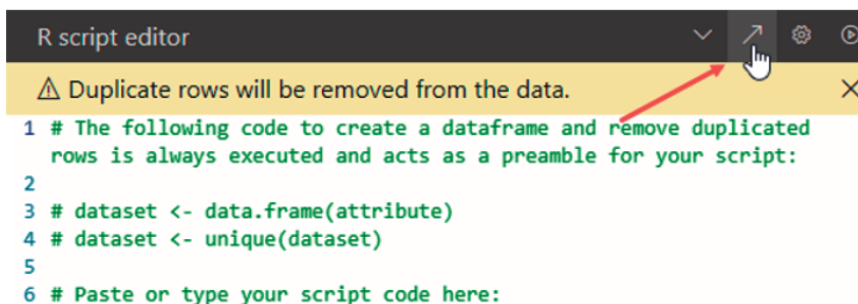
Rysunek 2.13. Wybierz preferowany silnik języka R do wykorzystania w środowisku RStudio



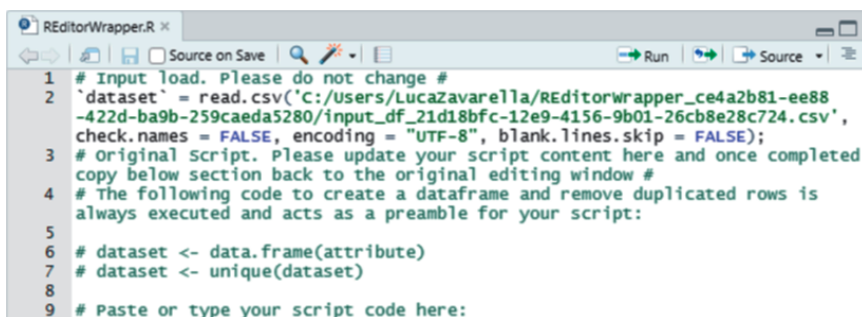
Rysunek 2.14. Otwieranie okna Opcje i ustawienia programu Power BI Desktop



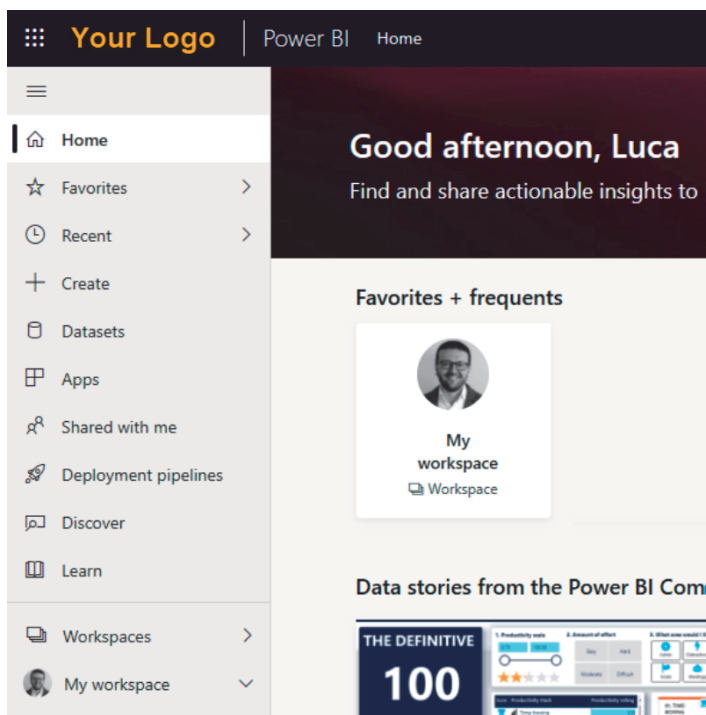
Rysunek 2.15. Wybieranie silnika języka R i środowiska IDE do pracy w programie Power BI Desktop



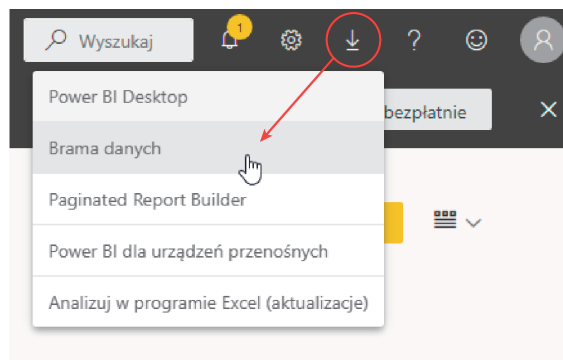
Rysunek 2.16. Otwieranie kodu wizualizacji języka R w programie RStudio



Rysunek 2.17. Debugowanie kodu wizualizacji języka R w programie RStudio



Rysunek 2.18. Strona główna usługi Power BI



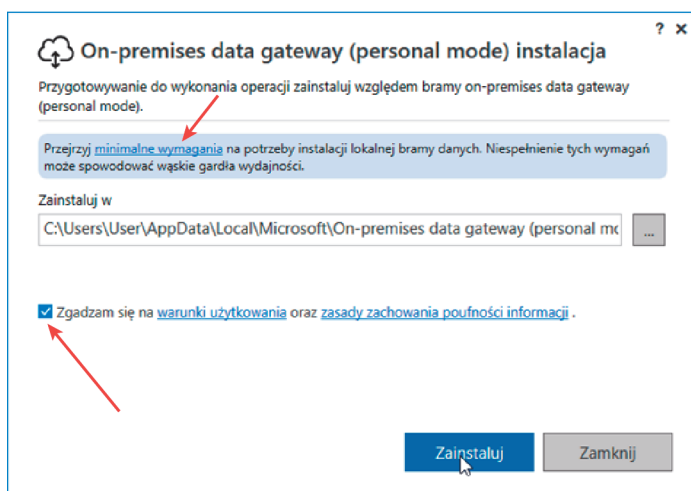
Rysunek 2.19. Menu pobierania na stronie głównej usługi Power BI

Łączenie z lokalnymi źródłami danych przy użyciu bramy usługi Power BI

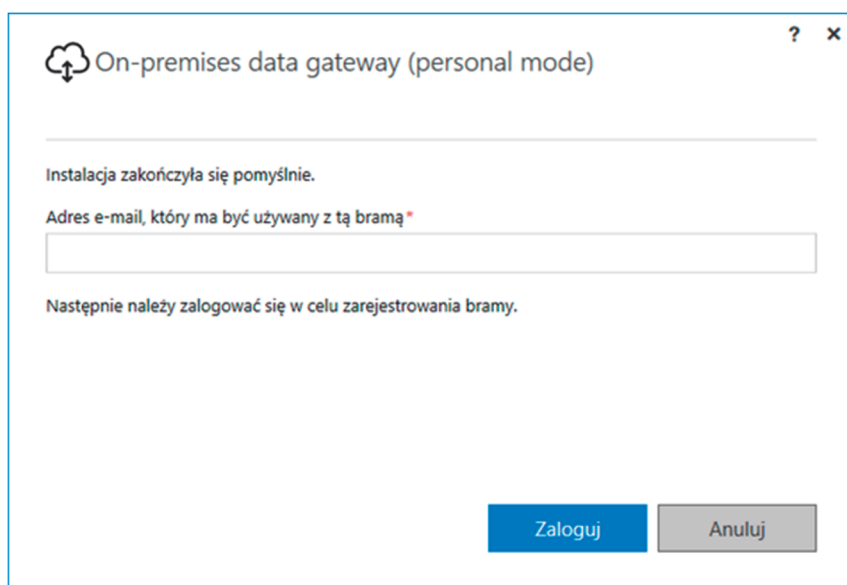
Zachowaj aktualność pulpitów nawigacyjnych i raportów, łącząc się z lokalnymi źródłami danych bez konieczności przenoszenia danych. Wykonuj zapytania względem dużych zestawów danych oraz korzystaj z istniejących inwestycji. Uzyskaj elastyczność, której potrzebujesz do spełnienia potrzeb Twoich i Twojej organizacji.

[Pobierz tryb standardowy >](#) [Pobierz tryb osobisty >](#)

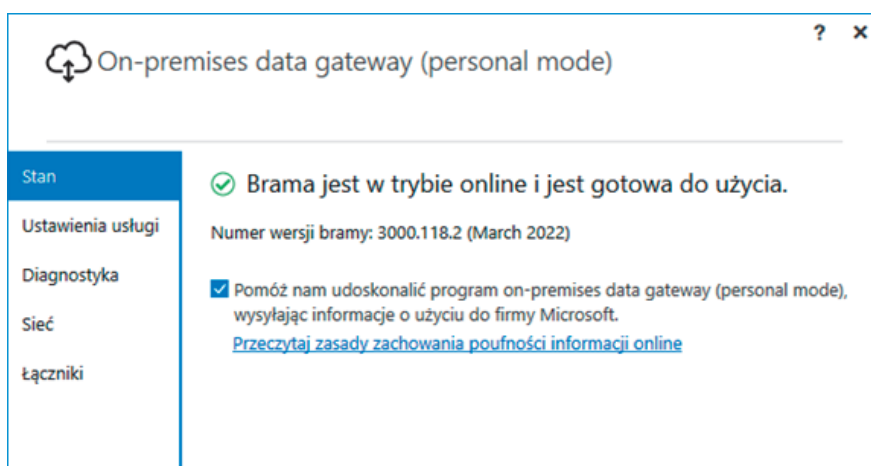
Rysunek 2.20. Pobierz wersję bramy danych dla trybu osobistego



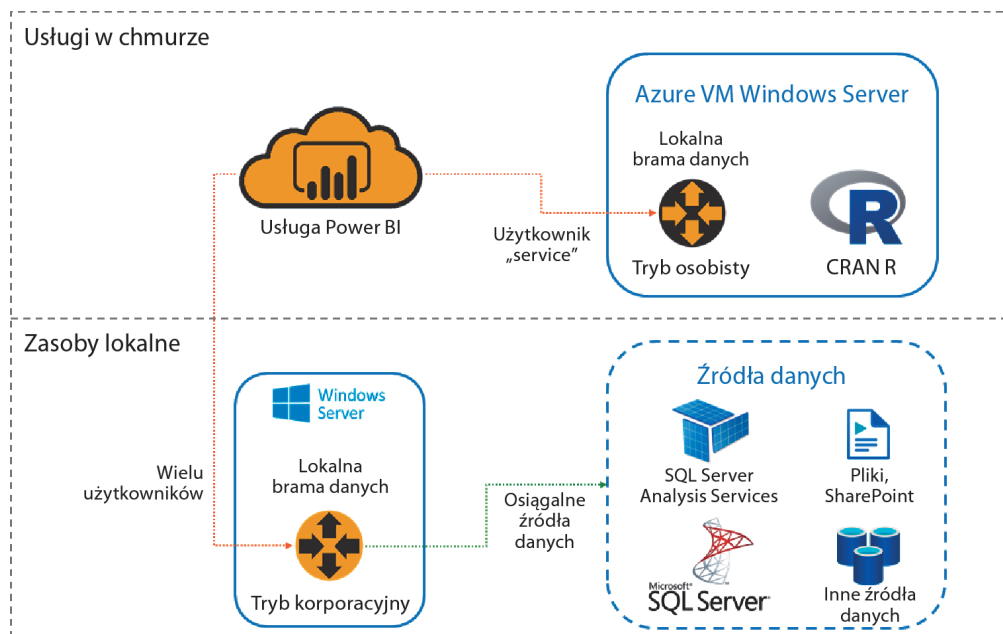
Rysunek 2.21. Okno instalacji bramy danych



Rysunek 2.22. Okno logowania do bramy danych



Rysunek 2.23. Uruchomiona brama danych

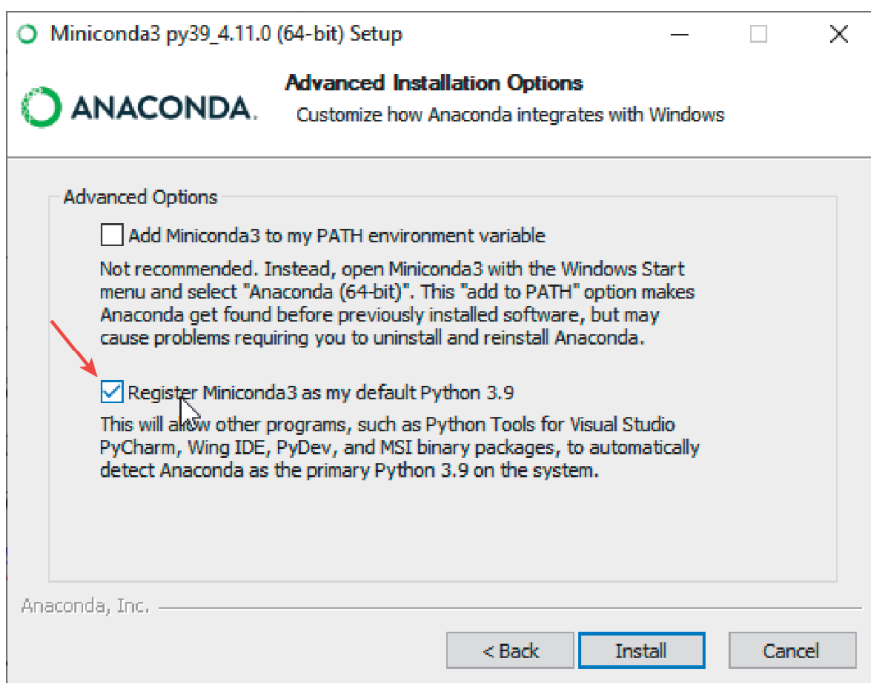


Rysunek 2.24. Architektura korporacyjna umożliwiająca używanie języka R do transformacji danych

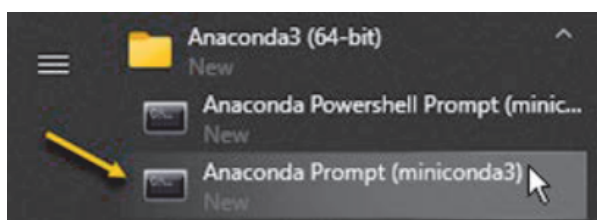
Rozdział 3. Konfigurowanie języka Python na potrzeby usługi Power BI

Python version	Name	Size
Python 3.9	Miniconda3 Windows 64-bit	70.4 MiB
Python 3.8	Miniconda3 Windows 64-bit	69.8 MiB
Python 3.7	Miniconda3 Windows 64-bit	68.1 MiB
Python 3.9	Miniconda3 Windows 32-bit	66.5 MiB
Python 3.8	Miniconda3 Windows 32-bit	65.6 MiB
Python 3.7	Miniconda3 Windows 32-bit	64.2 MiB

Rysunek 3.1. Pobieranie najnowszej dostępnej wersji Minicondy



Rysunek 3.2. Ustaw Minicondę jako domyślny silnik Pythona 3.9



Rysunek 3.3. Narzędzia wiersza polecenia dystrybucji Anaconda przydatne do interakcji z Minicondą

```
Anaconda Prompt (miniconda3)

(base) C:\Users\LZavarella>conda search python
Loading channels: done
# Name          Version      Build      Channel
python          2.7.13       h1b6d89f_16 pkgs/main
python          2.7.13       h9912b81_15 pkgs/main
python          2.7.13       hb034564_12 pkgs/main
python          2.7.14       h2765ee6_18 pkgs/main
python          3.8.5        h3rd99cc_1  pkgs/main
python          3.8.5        he1778fa_0  pkgs/main
python          3.9.0        h6244533_2  pkgs/main
python          3.9.0        h8aef87e_1  pkgs/main
python          3.9.1        h6244533_2  pkgs/main

(base) C:\Users\LZavarella>
```

Rysunek 3.4. Lista wszystkich dostępnych wersji Pythona

```
#
# To activate this environment, use
#
#     $ conda activate pbi_powerquery_env
#
# To deactivate an active environment, use
#
#     $ conda deactivate
#

(base) C:\Users\LZavarella>
```

Rysunek 3.5. Po utworzeniu nowego środowiska wirtualnego nadal jesteś w starym, zwanym „base”

```
(base) C:\Users\LZavarella>conda env list
# conda environments:
#
base * C:\Users\LZavarella\miniconda3
pbi_powerquery_env C:\Users\LZavarella\miniconda3\envs\pbi_powerquery_env

(base) C:\Users\LZavarella>
```

Rysunek 3.6. Lista środowisk conda w systemie

```
(base) C:\Users\LZavarella>conda activate pbi_powerquery_env
(pbi_powerquery_env) C:\Users\LZavarella>
```

Rysunek 3.7. Aktywacja nowego środowiska


```
(pbi_powerquery_env) C:\Users\LZavarella>python --version
Python 3.9.1
(pbi_powerquery_env) C:\Users\LZavarella>
```

Rysunek 3.8. Sprawdzanie wersji Pythona zainstalowanej w nowym środowisku

```
(pbi_powerquery_env) C:\Users\LZavarella>conda list
# packages in environment at C:\Users\LZavarella\miniconda3\envs\pbi_powerquery_
env:
#
# Name                                Version                                Build                                Channel
beautifulsoup4 ←                      4.9.3                                pypi_0                              pypi
ca-certificates                       2021.1.19                             haa95532_0                          pypi
certifi                               2020.12.5                             py39haa95532_0                      pypi
chardet                               4.0.0                                pypi_0                              pypi
idna                                  2.10                                  pypi_0                              pypi
numpy ←                               1.20.0                               pypi_0                              pypi
openssl                               1.1.1i                                h2bbff1b_0                          pypi
pandas ←                              1.2.1                                pypi_0                              pypi
pip                                    20.3.3                               py39haa95532_0                      pypi
python                                3.9.1                                h6244533_2                          pypi
python-dateutil                       2.8.1                                pypi_0                              pypi
pytz                                   2021.1                                pypi_0                              pypi
pyyaml ←                              5.4.1                                pypi_0                              pypi
requests ←                             2.25.1                               pypi_0                              pypi
scipy ←                                1.6.0                                pypi_0                              pypi
setuptools                             52.0.0                               py39haa95532_0                      pypi
six                                    1.15.0                               pypi_0                              pypi
soupsieve                             2.1                                   pypi_0                              pypi
sqlite                                 3.33.0                               h2a8f88b_0                          pypi
tzdata                                2020f                                 h52ac0ba_0                          pypi
urllib3                                1.26.3                               pypi_0                              pypi
vc                                      14.2                                  h21ff451_1                          pypi
vs2015_runtime                        14.27.29016                          h5e58377_2                          pypi
wheel                                  0.36.2                               pyhd3eb1b0_0                        pypi
wincertstore                           0.2                                   py39h2bbff1b_0                      pypi
zlib                                    1.2.11                               h62dcd97_4                          pypi
(pbi_powerquery_env) C:\Users\LZavarella>
```

Rysunek 3.9. Sprawdzanie, czy wszystkie wybrane pakiety Pythona zostały zainstalowane

Requirements and Limitations of Python packages

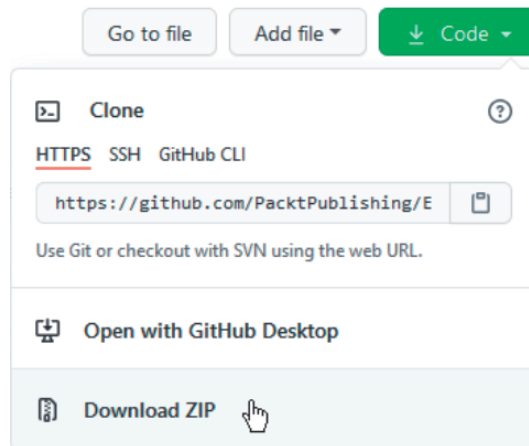
There are a handful of requirements and limitations for Python packages:

- Current Python runtime: Python 3.7.7

Rysunek 3.10. Obsługiwana wersja języka Python dla wizualizacji w usłudze Power BI

```
(pbi_powerquery_env) C:\Users\LZavarella>
```

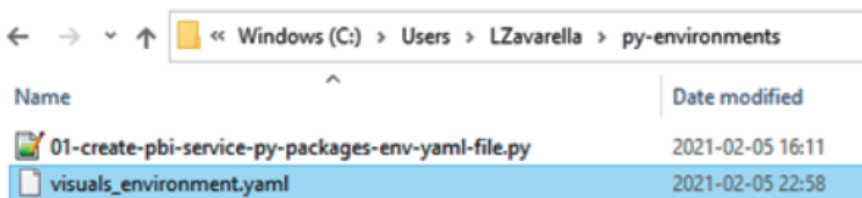
Rysunek 3.11. Domyślna ścieżka w programie Anaconda Prompt



Rysunek 3.12. Pobieranie całego spakowanego repozytorium

```
(pbi_powerquery_env) C:\Users\LZavarella\py-environments>python 01-create-pbi-service-py-packages-env-yaml-file.py
dependencies:
- python==3.7.7
- pip
- pip:
  - matplotlib==3.2.1
  - numpy==1.18.4
  - pandas==1.0.1
  - scikit-learn==0.23.0
  - scipy==1.4.1
  - seaborn==0.10.1
  - statsmodels==0.11.1
  - xgboost==1.1.0
name: pbi_visuals_env
```

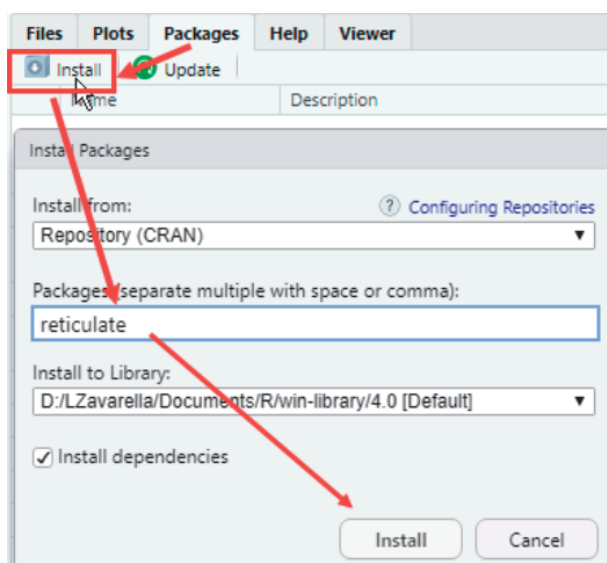
Rysunek 3.13. Uruchom skrypt Pythona, aby wygenerować plik YAML



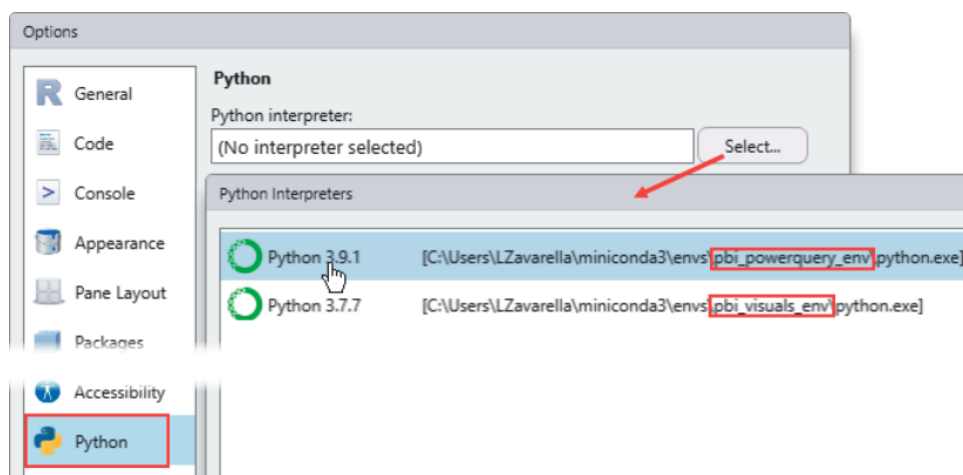
Rysunek 3.14. Poprawnie utworzony plik YAML

```
(pbi_visuals_env) C:\Users\LZavarella\py-environments>python --version
Python 3.7.7
```

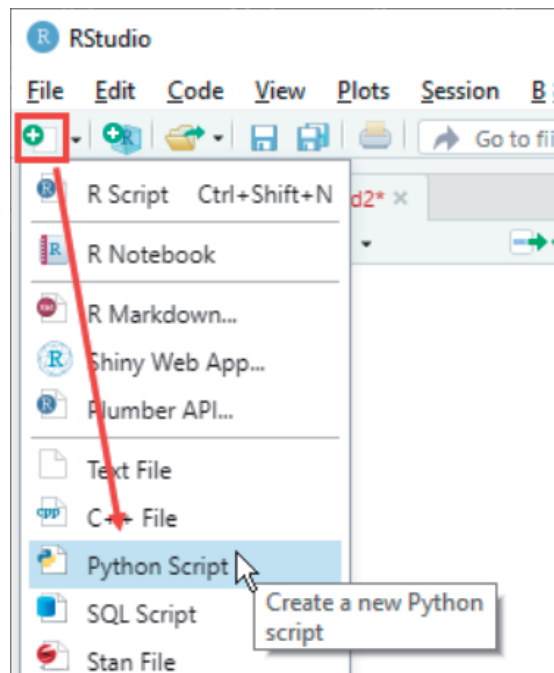
Rysunek 3.15. Nowe środowisko wirtualne zawiera prawidłową wersję Pythona



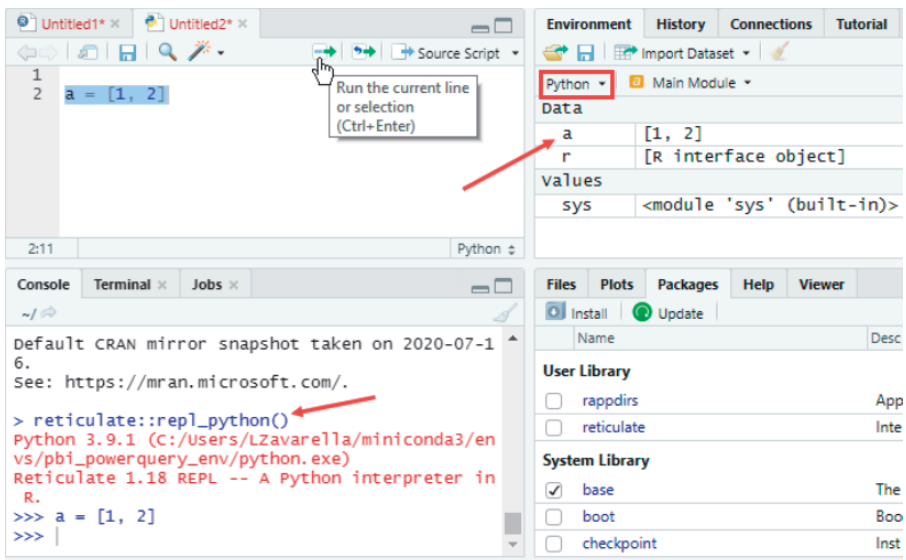
Rysunek 3.16. Instalowanie pakietu reticulate



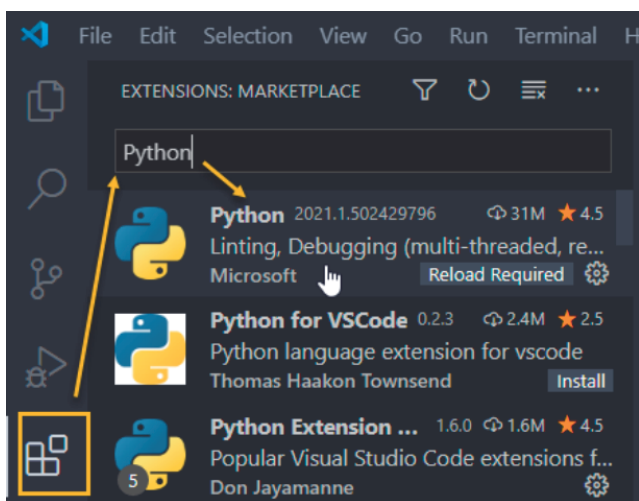
Rysunek 3.17. Ustaw preferowany interpreter języka Python w RStudio



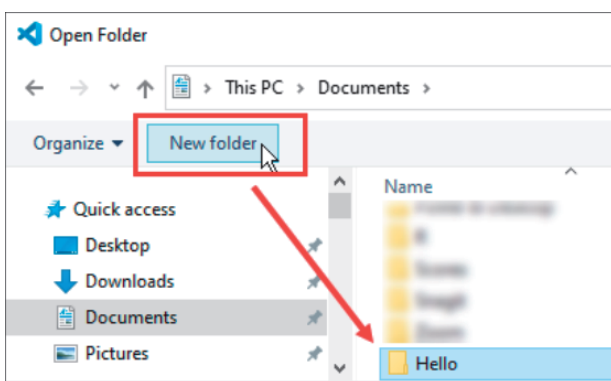
Rysunek 3.18. Tworzenie nowego skryptu Pythona w RStudio



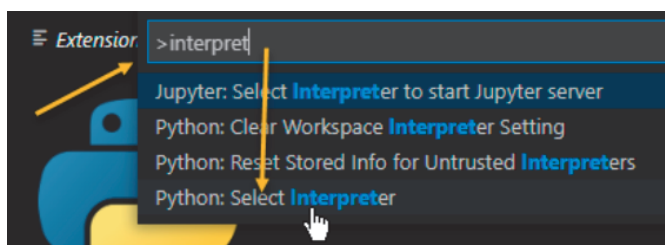
Rysunek 3.19. Uruchom pierwszy skrypt Pythona w RStudio



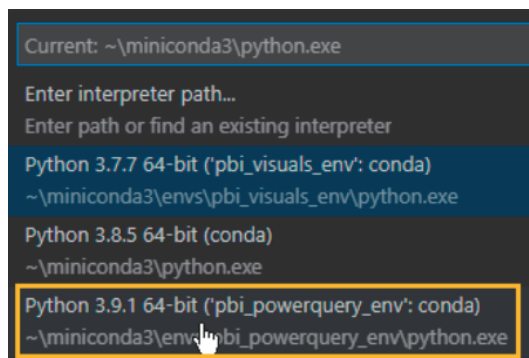
Rysunek 3.20. Uruchom pierwszy skrypt Pythona w Visual Studio Code



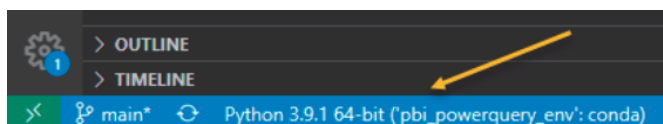
Rysunek 3.21. Utwórz nowy folder w locie i wybierz go w oknie Open Folder



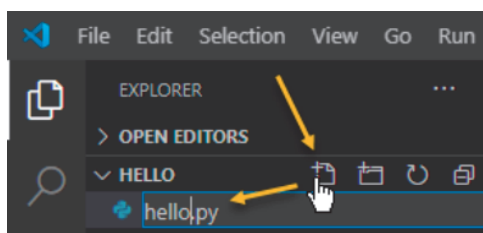
Rysunek 3.22. Na palecie poleceń wybierz opcję Python: Select interpreter



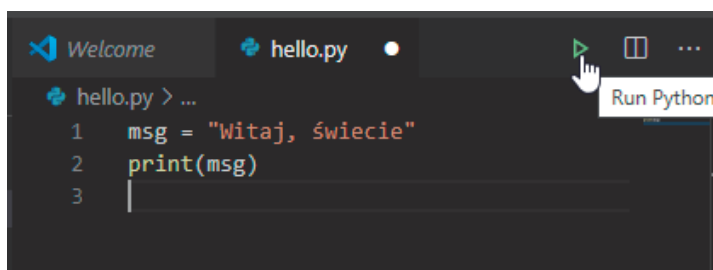
Rysunek 3.23. Wybierz preferowane środowisko



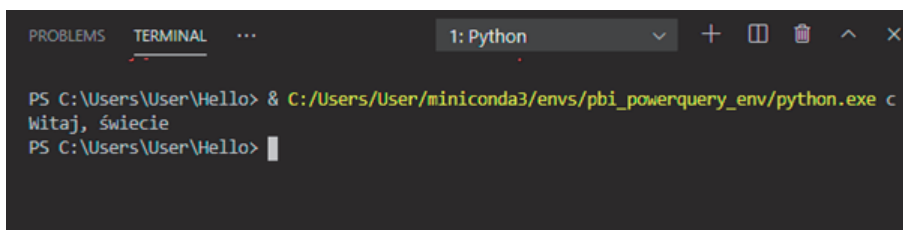
Rysunek 3.24. Sprawdź wybrane środowisko na pasku stanu



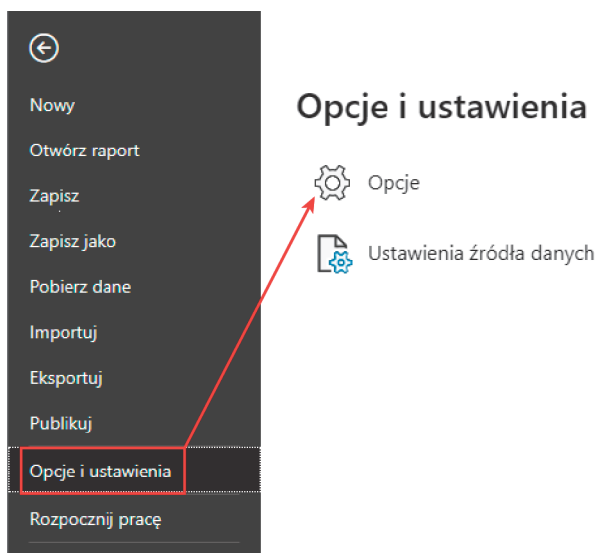
Rysunek 3.25. Utwórz w folderze Hello nowy plik o nazwie hello.py



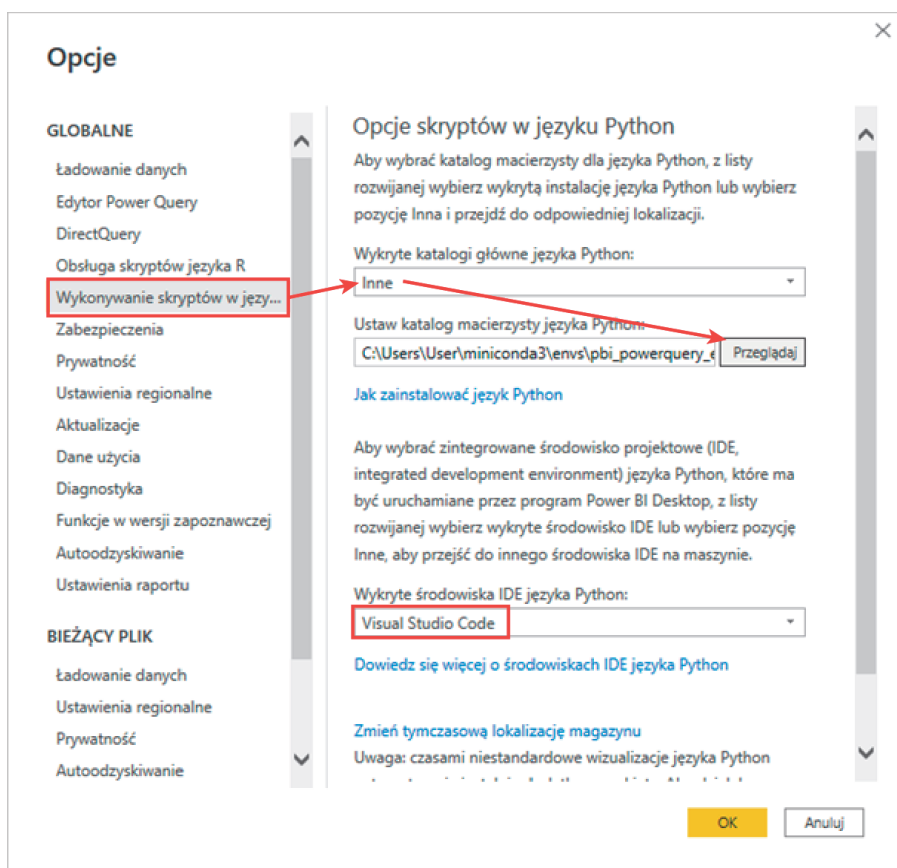
Rysunek 3.26. Wprowadź przykładowy kod i uruchom go



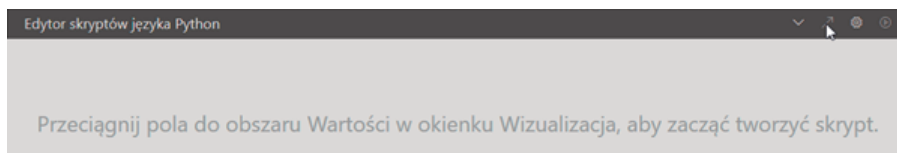
Rysunek 3.27. Twój pierwszy skrypt Python uruchomiony w VSCode



Rysunek 3.28. Otwieranie okna Opcje i ustawienia w programie Power BI Desktop



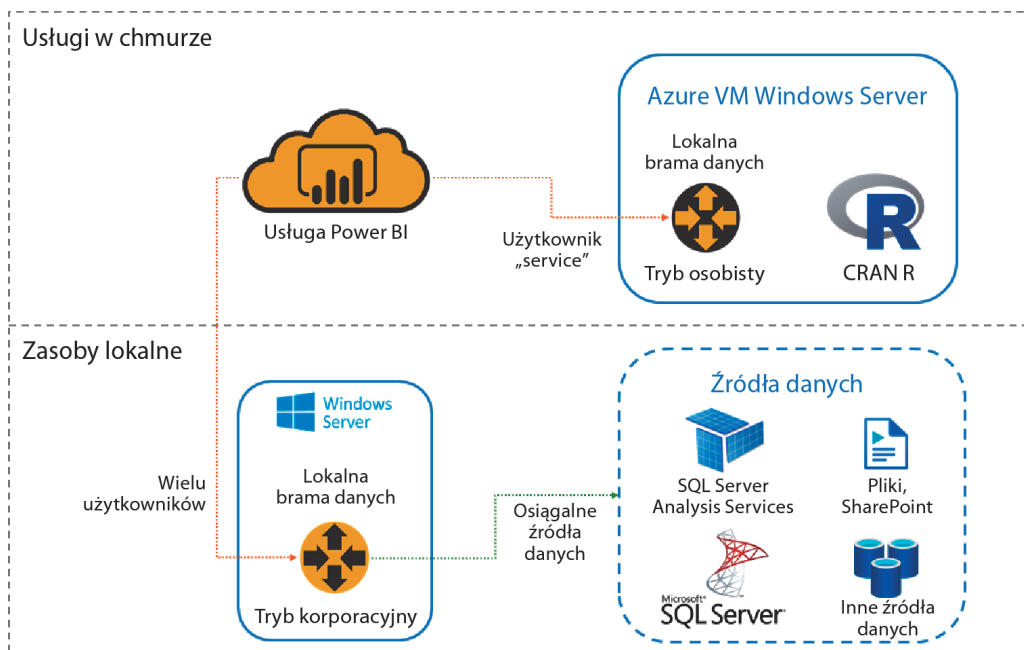
Rysunek 3.29. Konfigurowanie środowiska Pythona i środowiska IDE w usłudze Power BI



Rysunek 3.30. Otwieranie kodu wizualizacji w Pythonie w środowisku VSCode

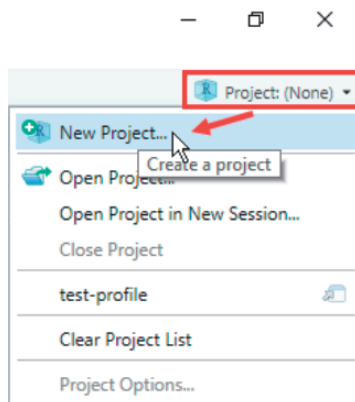
```
PythonEditorWrapper.PY X [Run] [Debug Console] [Output]
C: > Users > LucaZavarella > PythonEditorWrapper_5cacc6d5-9953-47b5-9e5f-526d75ec4c5b > PythonEditorWrapper.PY
1  # Prolog - Auto Generated #
2  import os, uuid, matplotlib
3  matplotlib.use('Agg')
4  import matplotlib.pyplot
5  import pandas
6
7  import sys
8  sys.tracebacklimit = 0
9
10 os.chdir(u'C:/Users/LucaZavarella/PythonEditorWrapper_5cacc6d5-9953-47b5-9e5f-526d75ec4c5b')
11 dataset = pandas.read_csv('input_df_bf3c1343-a225-459e-aa93-b6a6a973d912.csv')
12
13 matplotlib.pyplot.figure(figsize=(5.55555555555556,4.16666666666667), dpi=72)
14 matplotlib.pyplot.show = lambda args=None,kw=None: matplotlib.pyplot.savefig(str(uuid.uuid1()))
15 # Original Script. Please update your script content here and once completed copy below section
   back to the original editing window #
16 # The following code to create a dataframe and remove duplicated rows is always executed and
   acts as a preamble for your script:
17
18 # = pandas.DataFrame()
19 # = .drop_duplicates()
20
21 # Paste or type your script code here:
22
23 # Epilog - Auto Generated #
24 os.chdir(u'C:/Users/LucaZavarella/PythonEditorWrapper_5cacc6d5-9953-47b5-9e5f-526d75ec4c5b')
```

Rysunek 3.31. Debugowanie kodu wizualizacji Pythona w środowisku VSCode

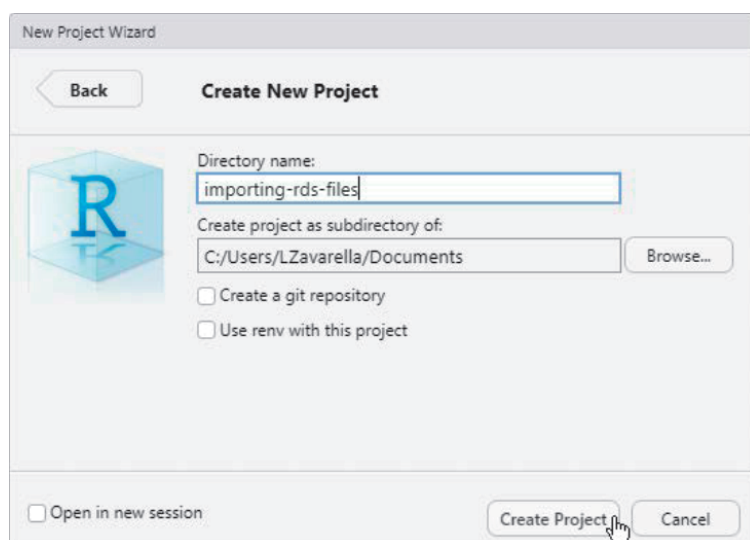


Rysunek 3.32. Architektura korporacyjna do wykorzystywania języków Python i R w transformacjach danych

Rozdział 4. Importowanie nieobsługiwanych obiektów danych



Rysunek 4.1. Tworzenie nowego projektu w RStudio



Rysunek 4.2. Tworzenie nowego folderu projektu

```
> data("population")
> # Rzućmy okiem na obiekt tibble z danymi populacji
> population
# A tibble: 4,060 x 3
  country      year population
  <chr>      <int>    <int>
1 Afghanistan 1995    17586073
2 Afghanistan 1996    18415307
3 Afghanistan 1997    19021226
4 Afghanistan 1998    19496836
5 Afghanistan 1999    19987071
6 Afghanistan 2000    20595360
7 Afghanistan 2001    21347782
8 Afghanistan 2002    22202806
9 Afghanistan 2003    23116142
10 Afghanistan 2004    24018682
# ... with 4,050 more rows
```

Rysunek 4.3. Załaduj obiekt tibble z danymi populacji

```
> population %>%
+ distinct(country) %>%
+ pull()
[1] "Afghanistan"
[3] "Albania"
[5] "Algeria"
[7] "American Samoa"
[9] "Angola"
[11] "Antigua and Barbuda"
[13] "Argentina"
[15] "Armenia"
[17] "Australia"
[19] "Austria"
[21] "Azerbaijan"
[23] "Bahamas"
[25] "Bangladesh"
[27] "Barbados"
[29] "Belarus"
[31] "Belgium"
[33] "Belize"
[35] "Benin"
[37] "Bhutan"
[39] "Bolivia"
[41] "Bosnia and Herzegovina"
[43] "Botswana"
[45] "Brazil"
[47] "Bulgaria"
[49] "Burkina Faso"
[51] "Burundi"
[53] "Cambodia"
[55] "Cameroon"
[57] "Canada"
[59] "Cape Verde"
[61] "Cayman Islands"
[63] "Central African Republic"
[65] "Chad"
[67] "Chile"
[69] "China"
[71] "Colombia"
[73] "Comoros"
[75] "Congo"
[77] "Costa Rica"
[79] "Croatia"
[81] "Cuba"
[83] "Cyprus"
[85] "Czechia"
[87] "Democratic Republic of the Congo"
[89] "Denmark"
[91] "Dominica"
[93] "Dominican Republic"
[95] "East Asia"
[97] "Ecuador"
[99] "Egypt"
[101] "El Salvador"
[103] "Equatorial Guinea"
[105] "Eritrea"
[107] "Estonia"
[109] "Ethiopia"
[111] "Faroe Islands"
[113] "Fiji"
[115] "Finland"
[117] "France"
[119] "French Polynesia"
[121] "Gabon"
[123] "Gambia"
[125] "Georgia"
[127] "Germany"
[129] "Ghana"
[131] "Greece"
[133] "Greenland"
[135] "Grenada"
[137] "Guatemala"
[139] "Guinea"
[141] "Guinea-Bissau"
[143] "Guyana"
[145] "Haiti"
[147] "Honduras"
[149] "Hungary"
[151] "Iceland"
[153] "India"
[155] "Indonesia"
[157] "Iran"
[159] "Iraq"
[161] "Ireland"
[163] "Israel"
[165] "Italy"
[167] "Jamaica"
[169] "Japan"
[171] "Jordan"
[173] "Kazakhstan"
[175] "Kenya"
[177] "Kiribati"
[179] "Korea"
[181] "Kuwait"
[183] "Kyrgyzstan"
[185] "Laos"
[187] "Latvia"
[189] "Lebanon"
[191] "Lesotho"
[193] "Liberia"
[195] "Liechtenstein"
[197] "Lithuania"
[199] "Luxembourg"
[201] "Madagascar"
[203] "Malawi"
[205] "Malaysia"
[207] "Maldives"
[209] "Mali"
[211] "Malta"
[213] "Marshall Islands"
[215] "Mauritania"
[217] "Mauritius"
[219] "Mexico"
[221] "Micronesia"
[223] "Moldova"
[225] "Monaco"
[227] "Mongolia"
[229] "Montenegro"
[231] "Morocco"
[233] "Mozambique"
[235] "Myanmar"
[237] "Namibia"
[239] "Nauru"
[241] "Nepal"
[243] "Netherlands"
[245] "New Caledonia"
[247] "New Zealand"
[249] "Nicaragua"
[251] "Niger"
[253] "Nigeria"
[255] "North Macedonia"
[257] "Norway"
[259] "Oman"
[261] "Pakistan"
[263] "Palau"
[265] "Palestine"
[267] "Panama"
[269] "Papua New Guinea"
[271] "Paraguay"
[273] "Peru"
[275] "Philippines"
[277] "Poland"
[279] "Portugal"
[281] "Puerto Rico"
[283] "Qatar"
[285] "Romania"
[287] "Russia"
[289] "Rwanda"
[291] "Saint Kitts and Nevis"
[293] "Saint Lucia"
[295] "Saint Vincent and the Grenadines"
[297] "Samoa"
[299] "San Marino"
[301] "Saudi Arabia"
[303] "Senegal"
[305] "Serbia"
[307] "Sierra Leone"
[309] "Singapore"
[311] "Slovakia"
[313] "Slovenia"
[315] "Solomon Islands"
[317] "South Africa"
[319] "South Asia"
[321] "South Korea"
[323] "South Sudan"
[325] "Spain"
[327] "Sri Lanka"
[329] "Sudan"
[331] "Suriname"
[333] "Swaziland"
[335] "Sweden"
[337] "Switzerland"
[339] "Taiwan"
[341] "Tajikistan"
[343] "Tanzania"
[345] "Thailand"
[347] "Timor-Leste"
[349] "Togo"
[351] "Tonga"
[353] "Trinidad and Tobago"
[355] "Tunisia"
[357] "Turkey"
[359] "Turkmenistan"
[361] "Turkmenistan"
[363] "Uganda"
[365] "Ukraine"
[367] "United Kingdom"
[369] "United States"
[371] "Uruguay"
[373] "Uzbekistan"
[375] "Vanuatu"
[377] "Venezuela (Bolivarian Republic of)"
[379] "Viet Nam"
[381] "West Bank and Gaza Strip"
[383] "Zambia"
[385] "Zimbabwe"
> |
```

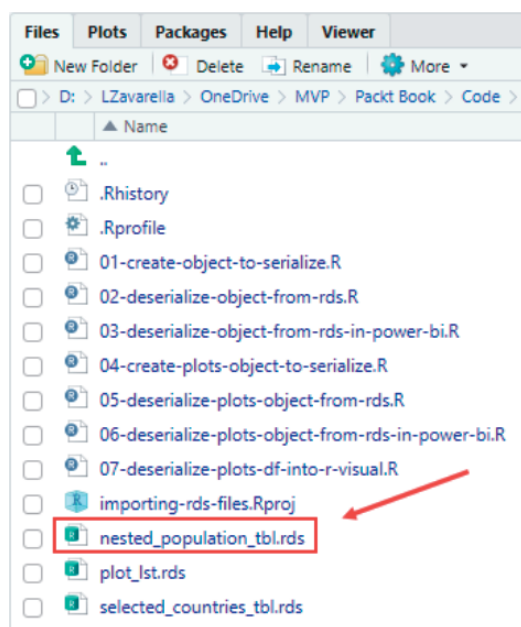
Rysunek 4.4. Lista różnych krajów

```

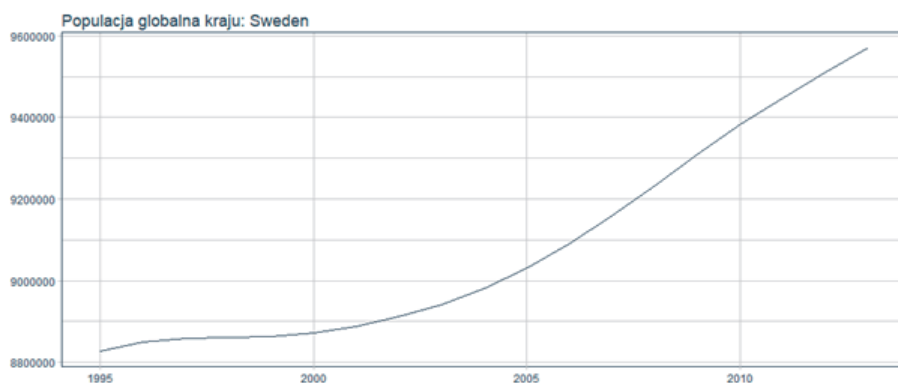
> nested_population_tbl <- population %>%
+   tidyr::nest( demographic_data = ~country )
> nested_population_tbl
# A tibble: 219 x 2
  country      demographic_data
  <chr>        <list>
1 Afghanistan <tibble [19 x 2]>
2 Albania     <tibble [19 x 2]>
3 Algeria     <tibble [19 x 2]>
4 American Samoa <tibble [19 x 2]>
5 Andorra     <tibble [19 x 2]>
6 Angola      <tibble [19 x 2]>
7 Anguilla    <tibble [19 x 2]>
8 Antigua and Barbuda <tibble [19 x 2]>
9 Argentina   <tibble [19 x 2]>
10 Armenia    <tibble [19 x 2]>
# ... with 209 more rows

```

Rysunek 4.5. Obiekt tibble z zagnieżdżonymi danymi demograficznymi



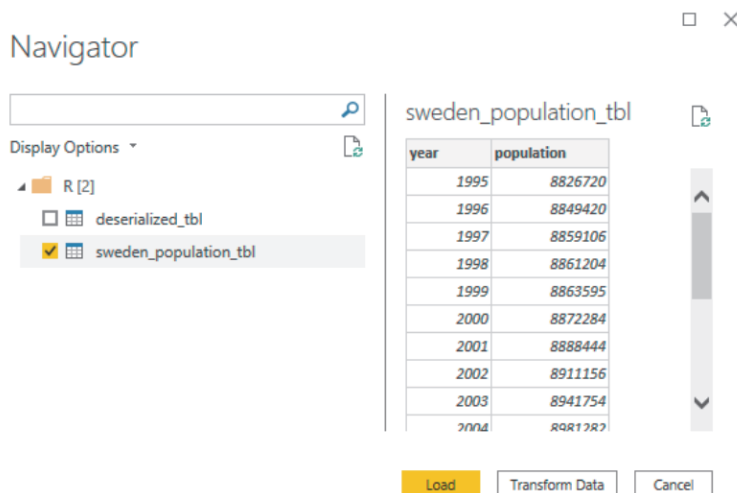
Rysunek 4.6. Poprawnie utworzony plik RDS



Rysunek 4.7. Wykres szeregów czasowych wzrostu liczby ludności dla Szwecji

```
> sweden_population_tbl <- deserialized_tbl %>%
+   filter( country == "Sweden" ) %>%
+   pull( demographic_data ) %>%
+   pluck(1)
> sweden_population_tbl
# A tibble: 19 x 2
  year population
  <int>      <int>
1  1995    8826720
2  1996    8849420
3  1997    8859106
4  1998    8861204
5  1999    8863595
6  2000    8872284
7  2001    8888444
8  2002    8911156
```

Rysunek 4.8. Zawartość danych demograficznych Szwecji zorganizowanych w obiekcie tibble



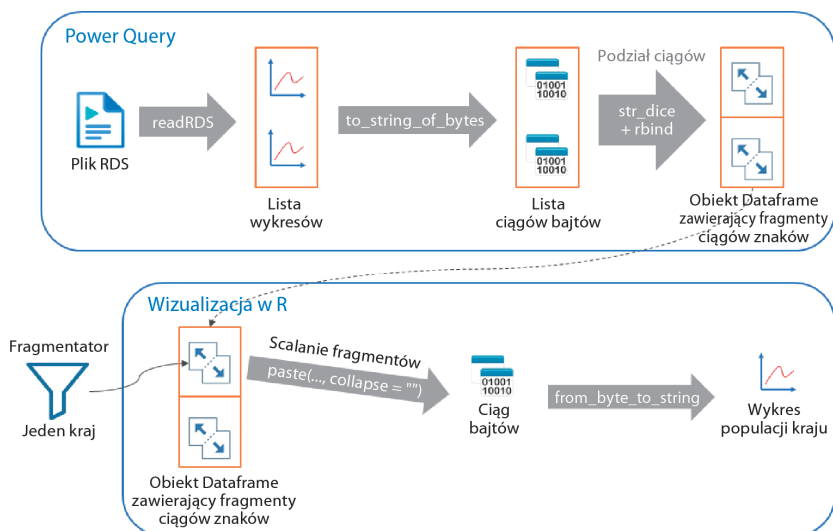
Rysunek 4.9. Importowanie zdeserializowanej ramki danych do usługi Power BI

Name sweden_population...

Structure

year	population
1995	8826720
1996	8849420
1997	8859106
1998	8861204
1999	8863595

Rysunek 4.10. Ramka danych została poprawnie zaimportowana



Rysunek 4.11. Deserializowanie zawartości pliku RDS do wizualizacji języka R

Nawigator

Opcje wyświetlania ▾

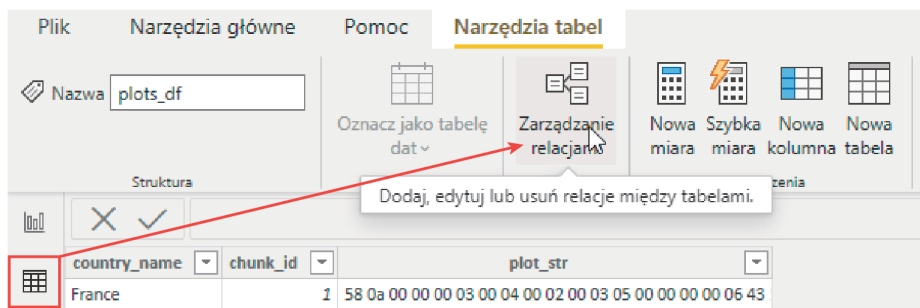
R [3]

- ☒ plots_df
- ☒ selected_countries_df
- ☐ tmp_df

plots_df

country_name	chunk_id	plot_str
France	1	58 0a 00 00 00 03 00 04 0
France	2	0 00 00 0e ff 00 00 00 02
France	3	65 73 73 69 6f 6e 73 49 6
France	4	00 00 fe 00 00 00 f4 00 0
France	5	0 00 00 f3 00 00 00 07 00
France	6	00 00 00 1a 00 00 00 01 0

Rysunek 4.12. Wybierz dwie ramki danych zawierające przydatne dane

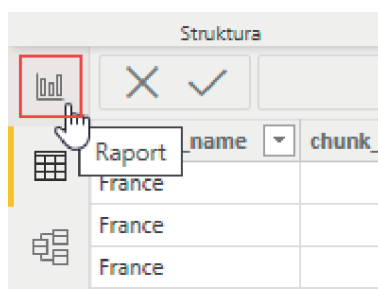


Rysunek 4.13. Przycisk Zarządzanie relacjami

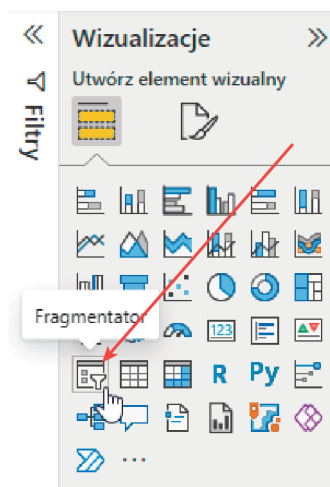
Zarządzanie relacjami

Aktywne	Z: tabela (kolumna)	Do: tabela (kolumna)
<input checked="" type="checkbox"/>	plots_df (country_name)	selected_countries_df (country_name)

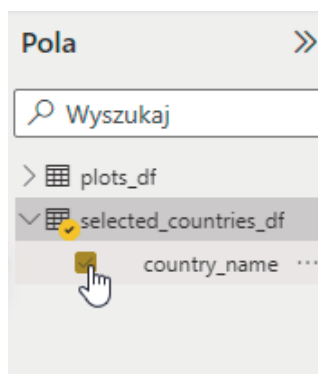
Rysunek 4.14. Automatycznie wykryta relacja między tabelami



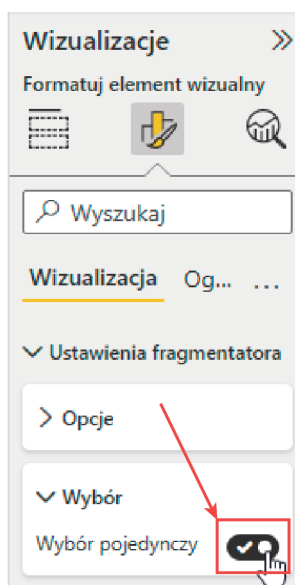
Rysunek 4.15. Ikona Raport



Rysunek 4.16. Ikona fragmentatora



Rysunek 4.17. Wybierz kolumnę country_name dla wizualizacji fragmentatora

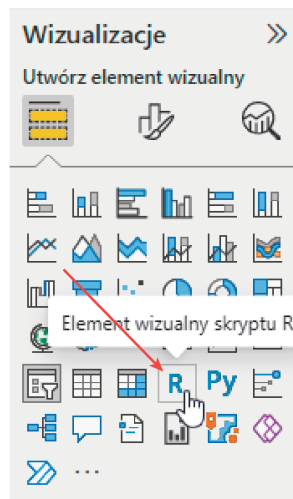


Rysunek 4.18. Zezwalaj tylko na pojedynczy wybór

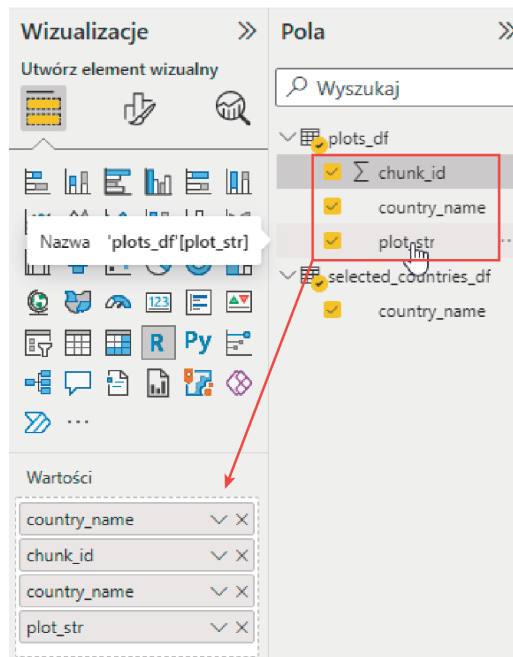
country_name

- ☒ France
- ☐ Germany
- ☐ Italy
- ☐ Sweden

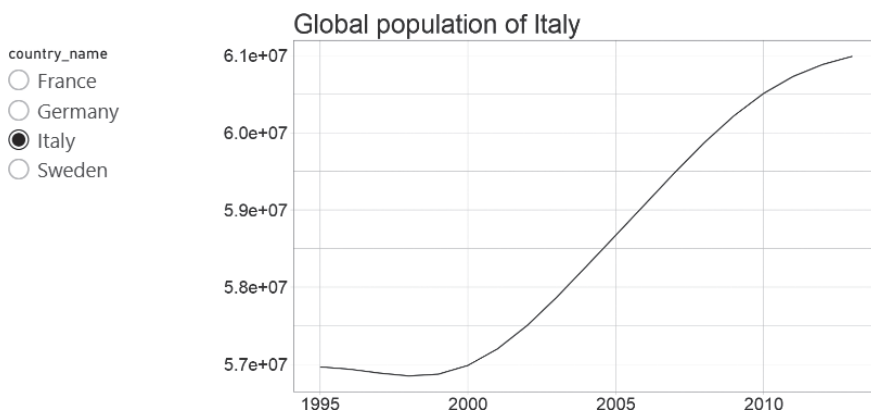
Rysunek 4.19. Tak wygląda wizualizacja fragmentatora



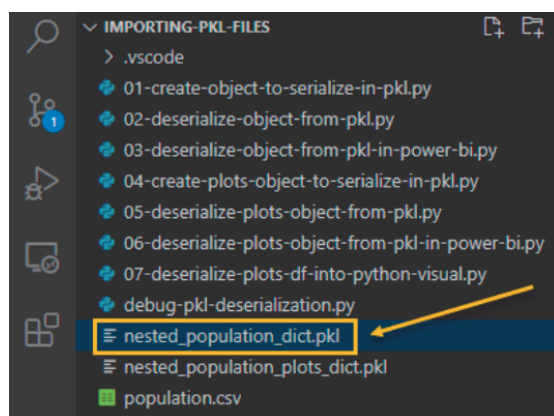
Rysunek 4.20. Ikona elementu wizualnego skryptu R



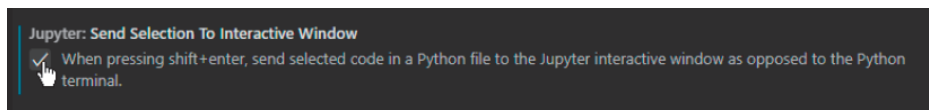
Rysunek 4.21. Wybieranie pól do wykorzystania w elemencie wizualnym języka R



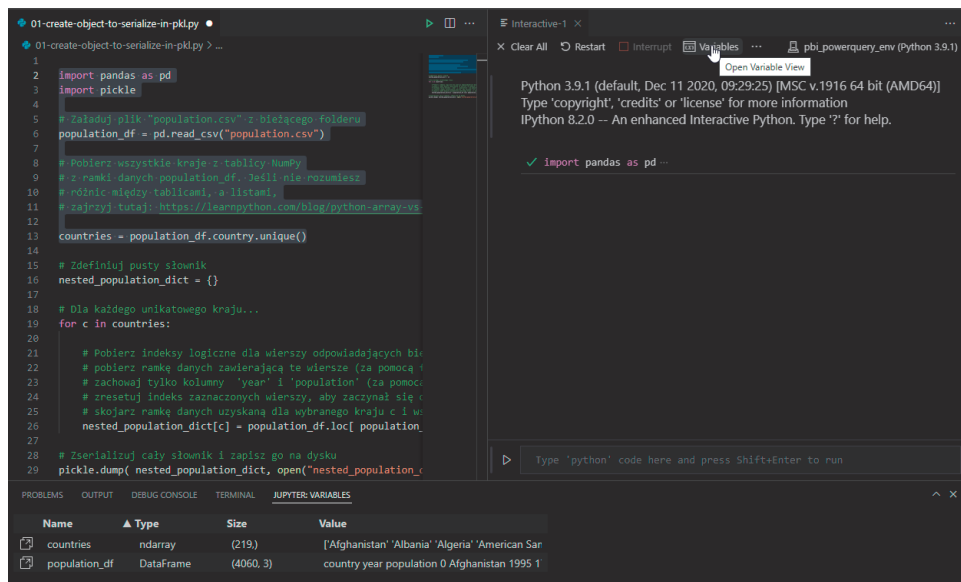
Rysunek 4.22. Wyświetlanie danych populacji we Włoszech



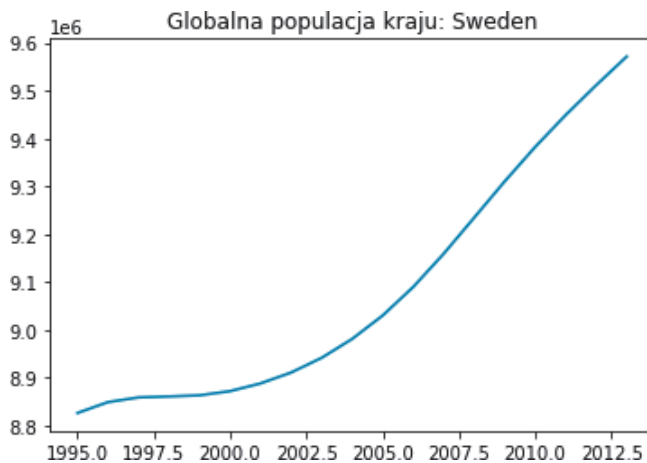
Rysunek 4.23. Twój pierwszy plik PKL został poprawnie utworzony



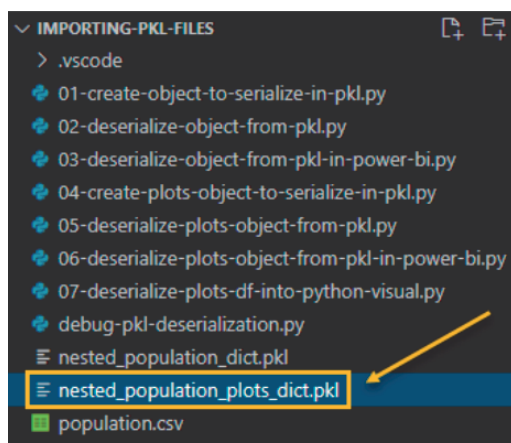
Rysunek 4.24. Włącz wykonywanie fragmentów kodu Pythona w interaktywnym oknie Jupyter



Rysunek 4.25. Uruchamianie zaznaczonych fragmentów skryptu w programie Visual Studio Code

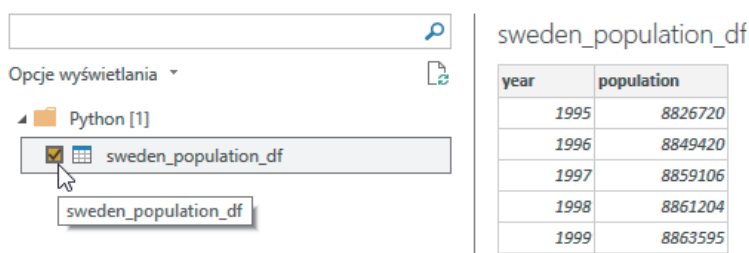


Rysunek 4.26. Wykres szeregów czasowych dla Szwecji w oknie Interactive

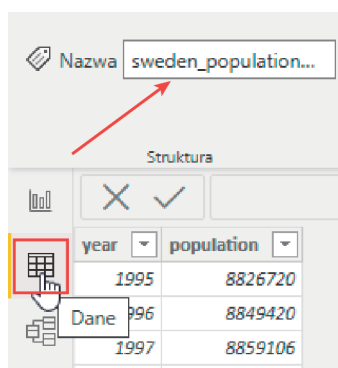


Rysunek 4.27. Nowy słownik został poprawnie zserializowany w pliku PKL

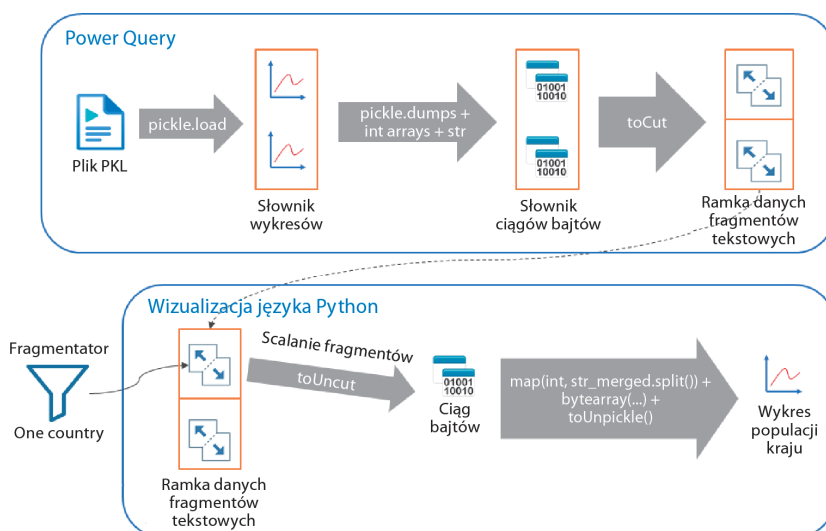
Nawigator



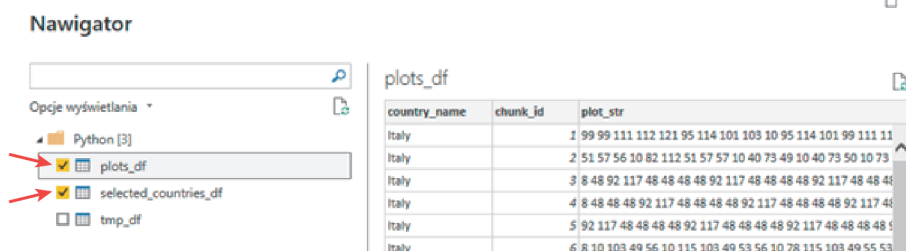
Rysunek 4.28. Importowanie zdeserializowanej ramki danych do usługi Power BI



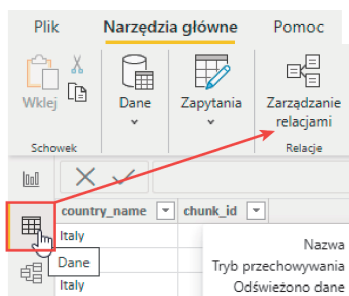
Rysunek 4.29. Ramka danych została poprawnie zaimportowana



Rysunek 4.30. Deserializacja zawartości pliku PKL do wizualizacji Pythona



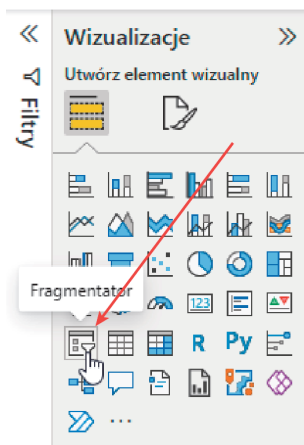
Rysunek 4.31. Zaznacz dwie ramki danych zawierające przydatne dane



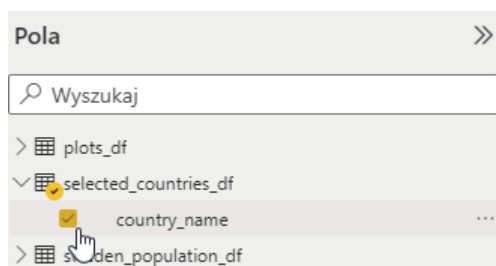
Rysunek 4.32. Przycisk Zarządzanie relacjami

Aktywne	Z: tabela (kolumna)	Do: tabela (kolumna)
<input checked="" type="checkbox"/>	plots_df (country_name)	selected_countries_df (country_name)

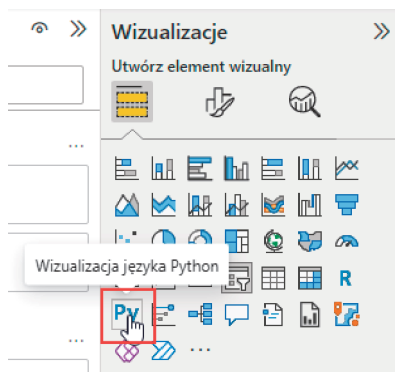
Rysunek 4.33. Automatycznie wykryta relacja między tabelami



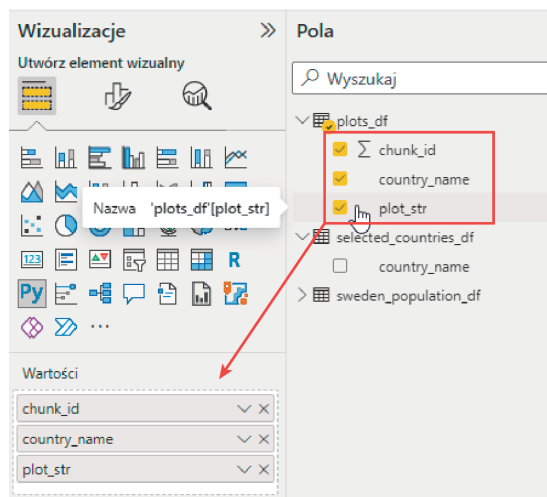
Rysunek 4.34. Ikona fragmentatora



Rysunek 4.35. Wybierz kolumnę country_name dla wizualizacji fragmentatora

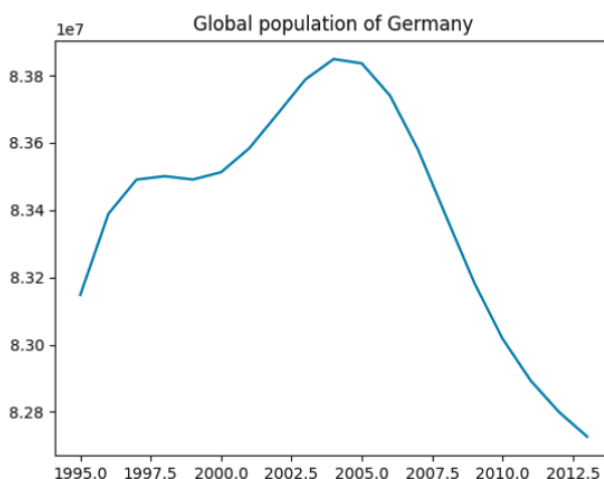


Rysunek 4.36. Ikona Wizualizacja języka Python



Rysunek 4.37. Wybierz pola do wykorzystania w wizualizacji języka Python

- country_name
- ☐ France
 - ☒ Germany
 - ☐ Italy
 - ☐ Sweden

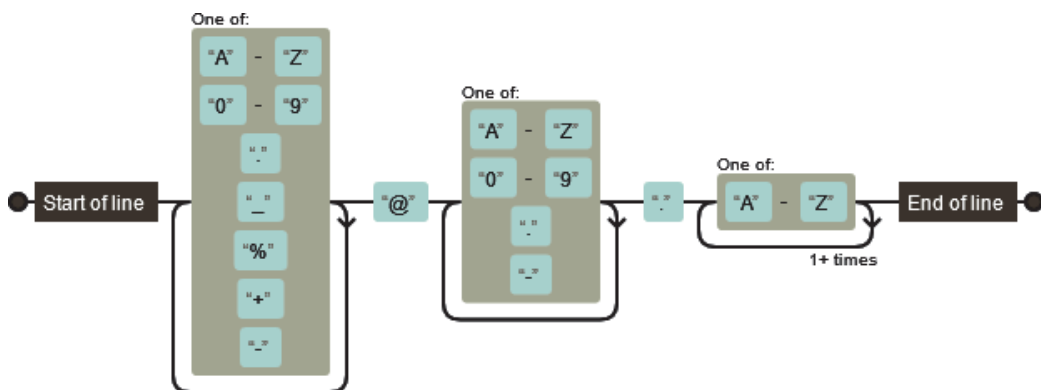


Rysunek 4.38. Wykres danych populacji dla Niemczech

Rozdział 5. Korzystanie z wyrażeń regularnych w usłudze Power BI

`^[A-Z0-9._%+-]+@[A-Z0-9.-]+\.[A-Z]{2,}$`

Rysunek 5.1. Przykład wzorca regeksa



Rysunek 5.2. Wizualizacja regeksa

Regular Expression JavaScript flags

`/ie/g` 2 matches

Test String

Niech moc rozszerzenia Power BI za pomocą Pythona i R bądź*ie* z Tobą!

Rysunek 5.3. Wyszukiwanie ciągu „ie” za pomocą regeksów

Regular Expression JavaScript flags

`/niech moc/ig` 2 matches

Test String

Niech moc rozszerzenia Power BI za pomocą Pythona i R będzie z Tobą i niech moc trwa wiecznie!

Rysunek 5.4. Wyszukiwanie globalne bez uwzględniania wielkości liter

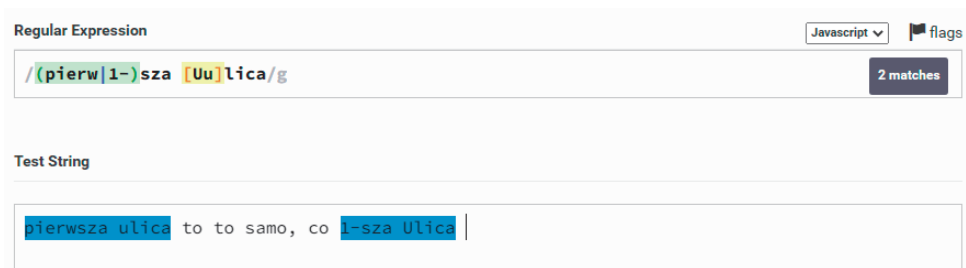
Regular Expression JavaScript flags

`/^niech moc/ig` 1 match

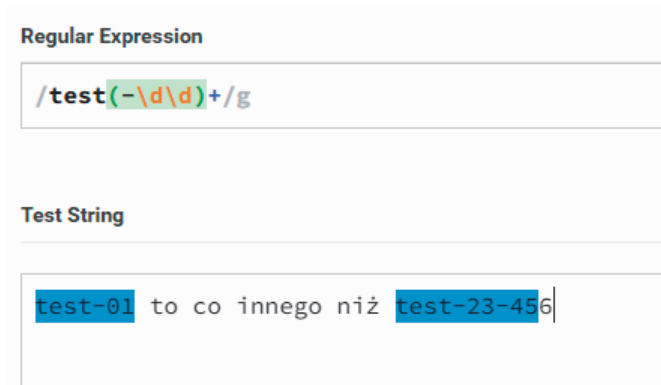
Test String

Niech moc rozszerzenia Power BI za pomocą Pythona i R będzie z Tobą i niech moc trwa wiecznie!

Rysunek 5.5. Wyszukiwanie globalne bez uwzględnienia wielkości liter oraz z zastosowaniem symbolu karetki (^)



Rysunek 5.6. Przykład zastosowania operatorów OR



Rysunek 5.7. Powtarzanie grupy znaków za pomocą kwantyfikatora +



Rysunek 5.8. Chciwość symbolu .+

Regular Expression

```
<.+?>/g
```

Test String

```
<em>Power BI wymiata</em>
```

Rysunek 5.9. Przekształcenie chciwego kwantyfikatora w leniwy dzięki metaznakowi ?

Regular Expression

```
<[^>]+>/g
```

Test String

```
<em>Power BI wymiata</em>
```

Rysunek 5.10. Wykorzystanie zanegowanej klasy znaków zamiast leniwego kwantyfikatora

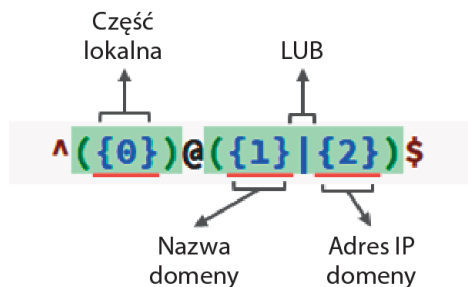
Regular Expression

```
/^.+@.+\..+$ /g
```

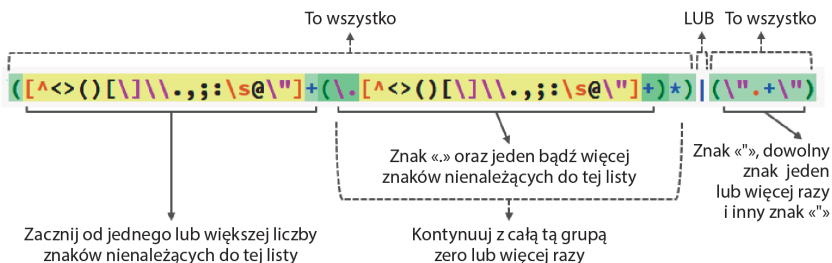
Test String

```
example@example.c
```

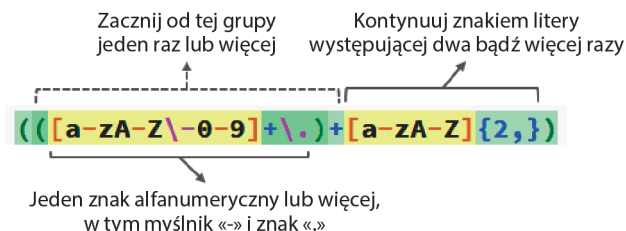
Rysunek 5.11. Wykorzystanie prostego regeksa do sprawdzenia poprawności nieprawidłowego adresu e-mail



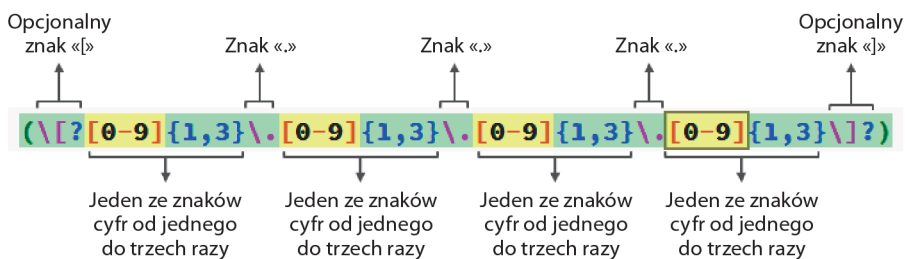
Rysunek 5.12. Struktura złożonego regeksa do sprawdzania poprawności adresów e-mail



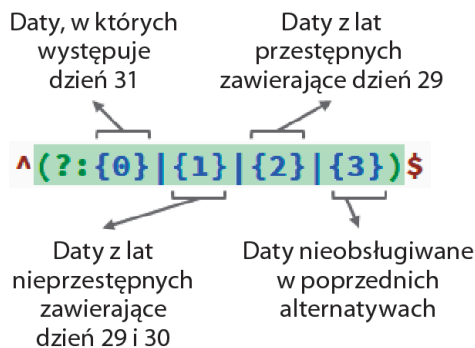
Rysunek 5.13. Szczegółowe wyjaśnienie regeksa części lokalnej



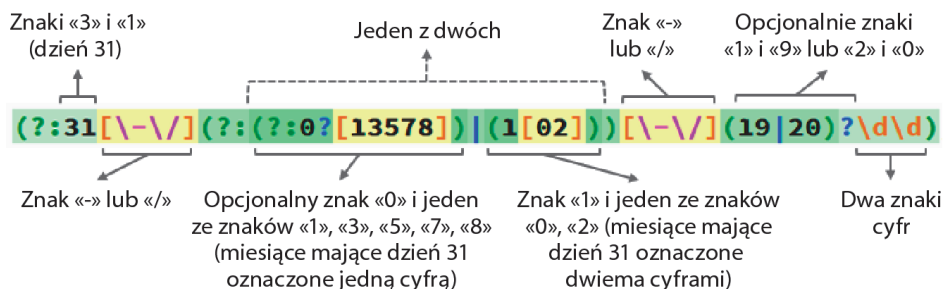
Rysunek 5.14. Szczegółowe wyjaśnienie regeksa nazwy domeny



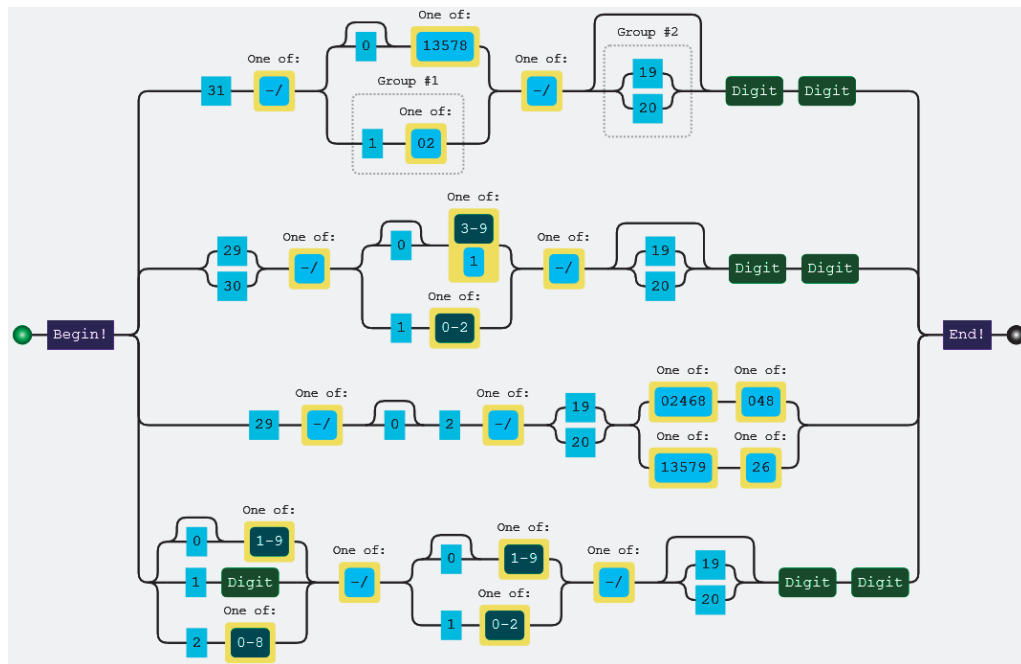
Rysunek 5.15. Szczegółowe wyjaśnienie regeksa adresu IP domeny



Rysunek 5.18. Struktura złożonego regeksa do walidacji daty



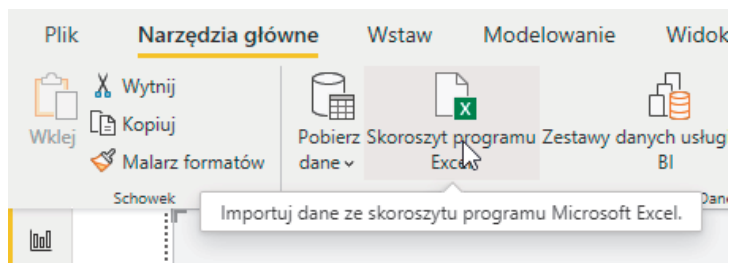
Rysunek 5.19. Szczegółowy opis części regeksa dla dat, które mają dzień 31



Rysunek 5.20. Wizualizacja całego regeksa sprawdzania poprawności dat

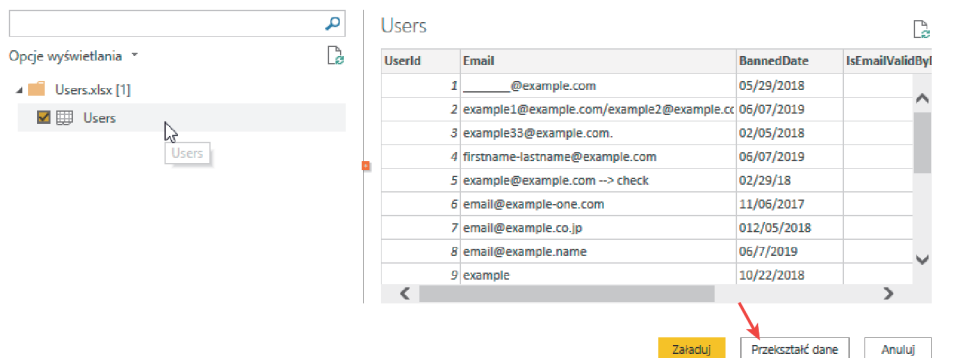
	A	B	C	D	E
1	UserId	Email	BannedDate	IsEmailValidByDefinition	IsDateValidByDefinition
2	1	@example.com	05/29/2018	1	1
3	2	example1@example.com/example2@example.com	06/07/2019	0	1
4	3	example33@example.com,	02/05/2018	0	1
5	4	firstname-lastname@example.com	06/07/2019	1	1
6	5	example@example.com --> check	02/29/18	0	0
7	6	email@example-one.com	11/06/2017	1	1

Rysunek 5.21. Zawartość pliku Users.xlsx

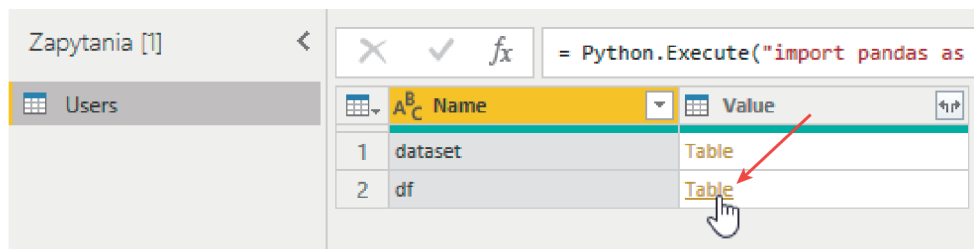


Rysunek 5.22. Importowanie danych z Excels

Nawigator



Rysunek 5.23. Wybierz arkusz Users i kliknij przycisk Przekształć dane



Rysunek 5.24. Wybieranie zestawu danych df uzyskanego w wyniku transformacji za pomocą skryptu Pythona

A ^B Email	A ^B BannedDate	1 ² IsEmailValidByDefinition	1 ² IsDateValidByDefinition	1 ² IsEmailValidFromRegex
@example.com	05/29/2018	1	1	1
example1@example.com/example2@exampl...	06/07/2019	0	1	0
example33@example.com.	02/05/2018	0	1	0
firstname-lastname@example.com	06/07/2019	1	1	1
example@example.com -> check	02/29/18	0	0	0
email@example-one.com	11/06/2017	1	1	1
email@example.co.jp	012/05/2018	1	0	1
email@example.name	06/7/2019	1	1	1
example	10/22/2018	0	1	0
email@example.com	9/04/2017	1	1	1
email@subdomain.example.com	11/22/2018	1	1	1
email@123.123.123	07/31/17	1	1	1
.@example.com	04/24/018	0	0	0
firstname.lastname@example.com	11/22/18	1	1	1
firstname+lastname@example.com	05/16/2018	1	1	1
example@example.c	6/7/2019	0	1	0
1234567890@example.com	05/16/2018	1	1	1
email@example.museum	03/26/2018	1	1	1
email@[123.123.123.123]	06/31/2017	1	0	1

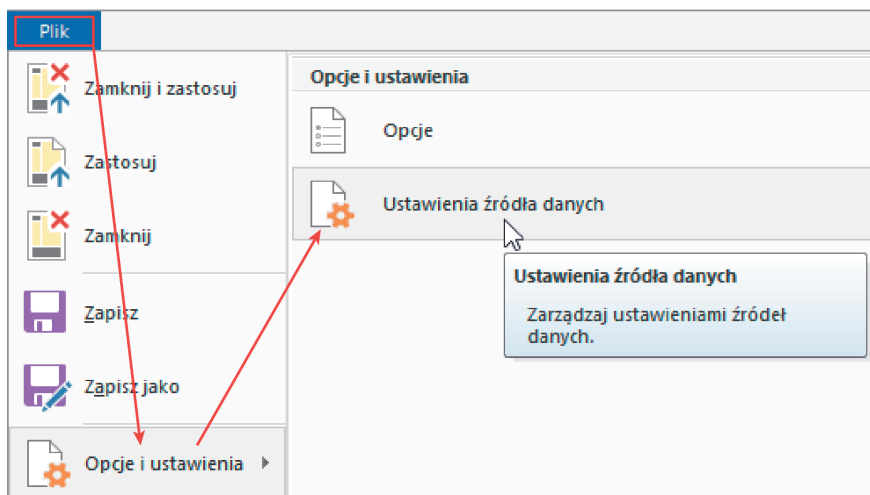
Rysunek 5.25. Wyniki weryfikacji adresów e-mail za pomocą regeksa

Python: `[^<>()\\[\]\\. ,;: \\s@\\"]`

R: `[^<>()\\[\\]\\. ,;: \\s@\\"]`

↑

Rysunek 5.26. W języku R należy poprzedzić otwarty nawias kwadratowy znakiem ucieczki, aby był interpretowany jako literał



Rysunek 5.27. Otwieranie okna ustawień źródła danych dodatku Power Query

☒ Źródła danych w bieżącym pliku ☐ Uprawnienia globalne

Wyszukaj ustawienia źródeł danych

c:\users\user\desktop\extendin...nd-r-main\chapter05\users.xlsx

R

Zmień źródło...
Edytuj uprawnienia...
Wyczyść uprawnienia
Kopiuj ścieżkę do Schowka
Usuń

Rysunek 5.28. Edytowanie uprawnień prywatności dla źródeł danych

Users	A ^B _C Name	ABC 123 Value
1	df	Table

Rysunek 5.29. Wybieranie zestawu danych df uzyskanego w wyniku transformacji za pomocą skryptu języka R

A ^B _C BannedDate	1 ² ₃ IsEmailValidByDefinition	1 ² ₃ IsDateValidByDefinition	1 ² ₃ isValidDateFromRegex
05/29/2018	1	1	1
06/07/2019	0	1	1
02/05/2018	0	1	1
06/07/2019	1	1	1
02/29/18	0	0	0
11/06/2017	1	1	1
012/05/2018	1	0	0
06/7/2019	1	1	1
10/22/2018	0	1	1
9/04/2017	1	1	1
11/22/2018	1	1	1
07/31/17	1	1	1
04/24/018	0	0	0
11/22/18	1	1	1
05/16/2018	1	1	1
6/7/2019	0	1	1
05/16/2018	1	1	1
03/26/2018	1	1	1
06/31/2017	1	0	0

Rysunek 5.30. Wyniki regaksa walidacji dat

```

1 83.149.9.216 - - [17/May/2015:10:05:03 +0000] "GET
/presentations/logstash-monitorama-2013/images/kibana-search.png HTTP/1.1" 200 203023
"http://semicomplete.com/presentations/logstash-monitorama-2013/" "Mozilla/5.0 (Macintosh;
Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.77 Safari/537.36"
2 83.149.9.216 - - [17/May/2015:10:05:43 +0000] "GET
/presentations/logstash-monitorama-2013/images/kibana-dashboard3.png HTTP/1.1" 200 171717
"http://semicomplete.com/presentations/logstash-monitorama-2013/" "Mozilla/5.0 (Macintosh;
Intel Mac OS X 10_9_1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.77 Safari/537.36"

```

Rysunek 5.31. Przykład logu dostępu serwera Apache

Nawigator

Opcje wyświetlania ▾

Python [1]

df

df

hostName	userName	requestDateTime	requestContent
df=83.149.9.216	-	17/May/2015:10:05:03 +0000	GET /presentations/k
83.149.9.216	-	17/May/2015:10:05:43 +0000	GET /presentations/k
83.149.9.216	-	17/May/2015:10:05:47 +0000	GET /presentations/k
83.149.9.216	-	17/May/2015:10:05:12 +0000	GET /presentations/k
83.149.9.216	-	17/May/2015:10:05:07 +0000	GET /presentations/k

Rysunek 5.32. Wybieranie ramki danych df zwróconej przez skrypt Pythona

hostName	userName	requestDateTime	requestContent	requestStatus	responseSizeBytes	requestReferrer	requestAgent
36.38.8.174	-	17/May/2015:12:05:39 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537
92.108.120.46	-	17/May/2015:13:05:12 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537
75.97.9.59	-	17/May/2015:19:05:12 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537
217.140.110.23	-	17/May/2015:19:05:40 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537
201.76.90.77	-	17/May/2015:20:05:59 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537

Rysunek 5.33. Log dostępu serwera Apache załadowany w usłudze Power BI za pomocą języka Python

Nawigator

Opcje wyświetlania ▾

R [1]

df

df

hostName	userName	requestDateTime	requestContent
83.149.9.216	-	17/May/2015:10:05:03 +0000	GET /presentations/k
83.149.9.216	-	17/May/2015:10:05:43 +0000	GET /presentations/k
83.149.9.216	-	17/May/2015:10:05:47 +0000	GET /presentations/k

Rysunek 5.34. Wybieranie ramki danych df zwróconej przez skrypt języka R

hostName	userName	requestDateTime	requestContent	requestStatus	responseSizeBytes	requestReferrer	requestAgent
36.38.8.174	-	17/May/2015:12:05:39 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537
92.108.120.46	-	17/May/2015:13:05:12 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537
75.97.9.59	-	17/May/2015:19:05:12 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537
217.140.110.23	-	17/May/2015:19:05:40 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537
201.76.90.77	-	17/May/2015:20:05:59 +0000	GET /favicon.ico HTTP/1.1	200	3638	-	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537

Rysunek 5.35. Log dostępu serwera Apache załadowany w usłudze Power BI za pomocą języka R

	A	B
1	OrderNumber	Notes
2	ORD000001	5,00 EUR Kradzież przy dostawie opłaconej przelewem 11/02/2021
3	ORD000002	29,00 EUR Zwrot za kradzież w dostawie 04/06/2020
4	ORD000003	53,00€ Zwrot za kradzież w dostawie 24/09/2020
5	ORD000004	29/10/2020 EUR45,00 Zwrot za kradzież w dostawie
6	ORD000005	EUR 522,00 Zwrot za kradzież w dostawie 20/08/2020
7	ORD000006	€ 266,00 - Kradzież w dostawie opłaconej przelewem w dniu 10/12/2020
8	ORD000007	EUR68,50 - Zwrot za kradzież w dostawie 02/07/2020
9	ORD000008	EUR 50,00 - Zwrot za kradzież przy dostawie z dnia 30/07/2020
10	ORD000009	30/07/2020 209,00 € - Zwrot za kradzież w dostawie

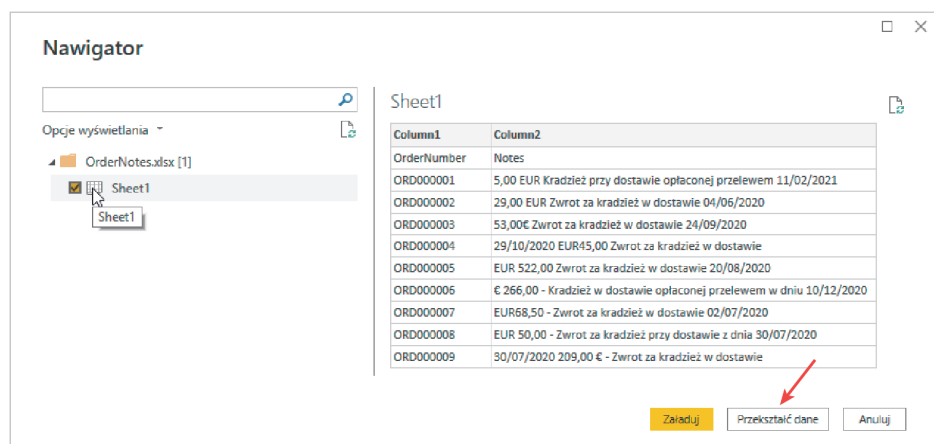
Rysunek 5.36. Notatki tekstowe wprowadzone przez operatora dla niektórych zamówień

```

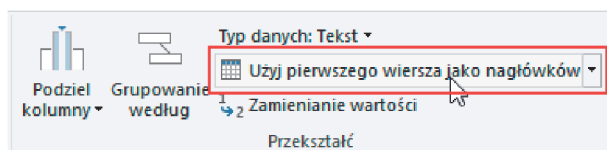
^(?:
({waluta}{separator}{kwota}{separator}{powód}{separator}{data})
OR
({kwota}{separator}{waluta}{separator}{powód}{separator}{data})
OR
({data}{separator}{waluta}{separator}{kwota}{separator}{powód})
OR
({data}{separator}{kwota}{separator}{waluta}{separator}{powód})
)$

```

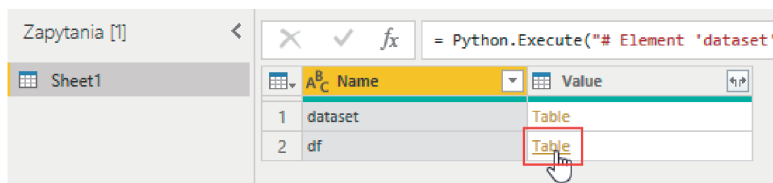
Rysunek 5.37. Pełna struktura regeksa do wyodrębniania informacji z notatek



Rysunek 5.38. Ładowanie notatek dotyczących zamówień z arkusza Excela i przekształcanie danych



Rysunek 5.39. Przycisk Użyj pierwszego wiersza jako nagłówek



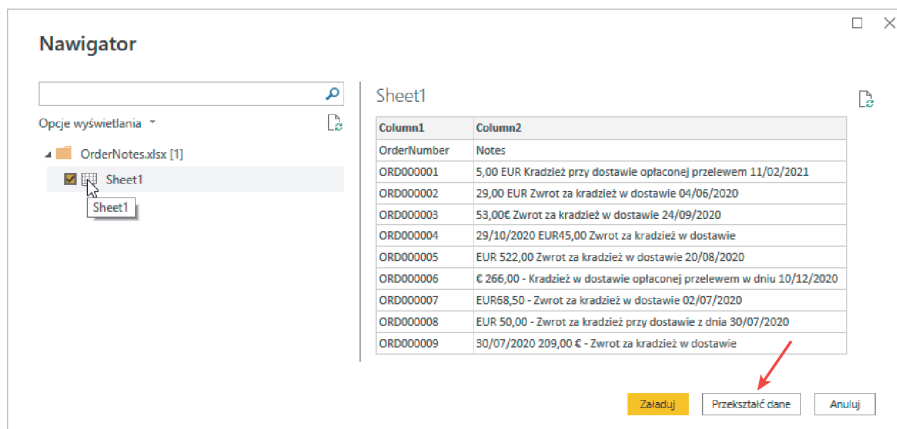
Rysunek 5.40. Wybieranie zestawu danych df otrzymanego w wyniku transformacji za pomocą skryptu Pythona

	OrderNumber	Notes	RefundAmount	RefundReason	RefundDate
1	ORD000001	5,00 EUR Kradzież przy dostawie opłaconej przelewem 11/02/2021	5	Kradzież przy dostawie opłaconej przelewem	2021-02-11
2	ORD000002	29,00 EUR Zwrot za kradzież w dostawie 04/06/2020	29	Zwrot za kradzież w dostawie	2020-06-04
3	ORD000003	53,00€ Zwrot za kradzież w dostawie 24/09/2020	53	Zwrot za kradzież w dostawie	2020-09-24
4	ORD000004	29/10/2020 EUR45,00 Zwrot za kradzież w dostawie	45	Zwrot za kradzież w dostawie	2020-10-29
5	ORD000005	EUR 522,00 Zwrot za kradzież w dostawie 20/08/2020	522	Zwrot za kradzież w dostawie	2020-08-20
6	ORD000006	€ 266,00 - Kradzież w dostawie opłaconej przelewem w dniu 10/12/20...	266	Kradzież w dostawie opłaconej przelewem w dniu	2020-12-10
7	ORD000007	EUR68,50 - Zwrot za kradzież w dostawie 02/07/2020	68,5	Zwrot za kradzież w dostawie	2020-07-02
8	ORD000008	EUR 50,00 - Zwrot za kradzież przy dostawie z dnia 30/07/2020	50	Zwrot za kradzież przy dostawie z dnia	2020-07-30
9	ORD000009	30/07/2020 209,00 € - Zwrot za kradzież w dostawie	209	Zwrot za kradzież w dostawie	2020-07-30

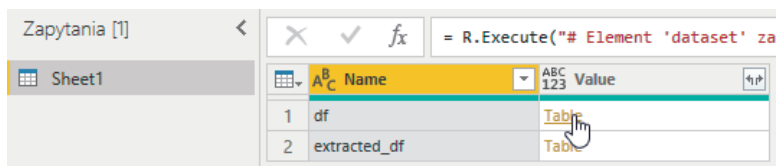
Rysunek 5.41. Wartości wyodrębnione z notatek zapisanych w formacie dowolnego tekstu za pomocą regaksa oraz kodu w Pythonie

	RefundAmount	RefundReason	RefundDate	RefundAmount.1	RefundReason.1	RefundDate.1	RefundDate.2	RefundAmount.2
1				5,00	Kradzież przy dostawie opłaconej przelewem	11/02/2021		
2				29,00	Zwrot za kradzież w dostawie	04/06/2020		
3				53,00	Zwrot za kradzież w dostawie	24/09/2020		
4							29/10/2020	45,00
5	522,00	Zwrot za kradzież w dostawie	20/08/2020					
6	266,00	Kradzież w dostawie opłaconej przelewem w dniu	10/12/2020					
7	68,50	Zwrot za kradzież w dostawie	02/07/2020					
8	50,00	Zwrot za kradzież przy dostawie z dnia	30/07/2020					
9								

Rysunek 5.42. Funkcja str_match_named() zwraca tyle kolumn, ile razy została użyta nazwana grupa



Rysunek 5.43. Ładowanie notatek dotyczących zamówienia z Excela i przekształcanie danych



Rysunek 5.44. Wybieranie zestawu danych df uzyskanego w wyniku transformacji za pomocą skryptu języka R

OrderNumber	Notes	RefundAmount	RefundDate	RefundReason
ORD000001	5,00 EUR Kradzież przy dostawie opłaconej przelewem 11/02/2021	5,00	11/02/2021	Kradzież przy dostawie opłaconej przelewem
ORD000002	29,00 EUR Zwrot za kradzież w dostawie 04/06/2020	29,00	04/06/2020	Zwrot za kradzież w dostawie
ORD000003	53,00€ Zwrot za kradzież w dostawie 24/09/2020	53,00	24/09/2020	Zwrot za kradzież w dostawie
ORD000004	29/10/2020 EUR45,00 Zwrot za kradzież w dostawie	45,00	29/10/2020	Zwrot za kradzież w dostawie
ORD000005	EUR 522,00 Zwrot za kradzież w dostawie 20/08/2020	522,00	20/08/2020	Zwrot za kradzież w dostawie
ORD000006	€ 266,00 - Kradzież w dostawie opłaconej przelewem w dniu 10/12/20...	266,00	10/12/2020	Kradzież w dostawie opłaconej przelewem w dniu
ORD000007	EUR68,50 - Zwrot za kradzież w dostawie 02/07/2020	68,50	02/07/2020	Zwrot za kradzież w dostawie
ORD000008	EUR 50,00 - Zwrot za kradzież przy dostawie z dnia 30/07/2020	50,00	30/07/2020	Zwrot za kradzież przy dostawie z dnia
ORD000009	30/07/2020 209,00 € - Zwrot za kradzież w dostawie	209,00	30/07/2020	Zwrot za kradzież w dostawie

Rysunek 5.45. Wartości wyodrębnione z notatek zapisanych w formacie dowolnego tekstu za pomocą regleksa oraz kodu w R

Rozdział 6. Anonimizacja i pseudonimizacja danych w usłudze Power BI

Name	Email	Gender	Country	NumOfOrders
Johann Lucas Idler	j.lucas.idler@bluewin.ch	male	SWITZERLAND	15
Ruth Connelly	clevecon@outlook.com	female	UNITED STATES	1
Francisco Javier Diez Gonzalez	francis1971diez@gmail.com	male	MEXICO	1
Alexei Fabius Malenkiy-Fransk	fabius-fransk@hotmail.com	male	UKRAINA	1
Hilma Kitzmann	hilmakit234@gmx.de	female	GERMANY	6
Luke Romba	rombaluke77@gmail.com	male	UNITED STATES	1
John Romkanid	romkaj69@gmail.com	male	UNITED STATES	2



Gender	Country	NumOfOrders
male	SWITZERLAND	15
female	UNITED STATES	1
male	MEXICO	1
male	UKRAINA	1
female	GERMANY	6
male	UNITED STATES	1
male	UNITED STATES	2

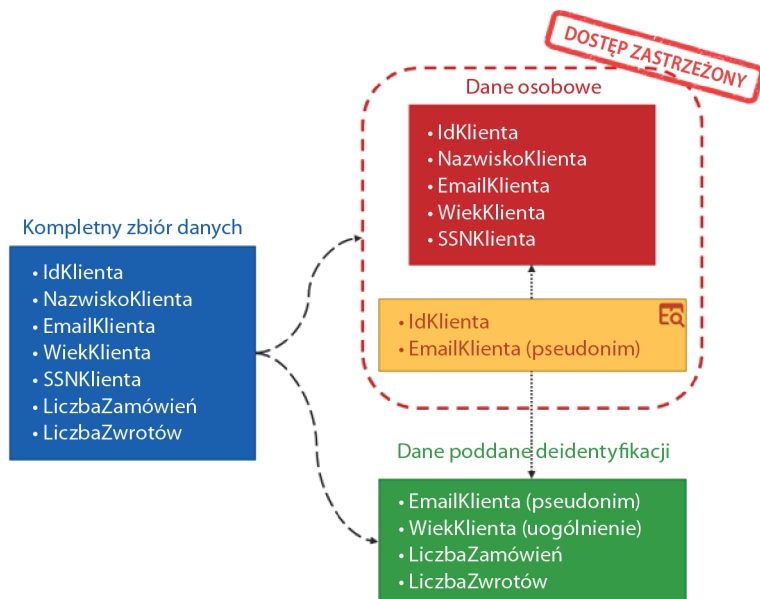
Rysunek 6.1. Anonimizacja, usuwanie informacji

Name	Email	Gender	Country	NumOfOrders
Johann Lucas Idler	j.lucas.idler@bluewin.ch	male	SWITZERLAND	15
Ruth Connelly	clevecon@outlook.com	female	UNITED STATES	1
Francisco Javier Diez Gonzalez	francis1971diez@gmail.com	male	MEXICO	1
Alexei Fabius Malenkiy-Fransk	fabius-fransk@hotmail.com	male	UKRAINA	1
Hilma Kitzmann	hilmakit234@gmx.de	female	GERMANY	6
Luke Romba	rombaluke77@gmail.com	male	UNITED STATES	1
John Romkanid	romkaj69@gmail.com	male	UNITED STATES	2



Name	Email	Gender	Country	NumOfOrders
Roberto Hall	*@bluewin.ch	male	SWITZERLAND	15
Amber Cole	*@outlook.com	female	UNITED STATES	1
Christopher Patel	*@gmail.com	male	MEXICO	1
Douglas Singleton	*@hotmail.com	male	UKRAINA	1
Brittany Harrison	*@gmx.de	female	GERMANY	6
John Trujillo	*@gmail.com	male	UNITED STATES	1
Stephen Thomas	*@gmail.com	male	UNITED STATES	2

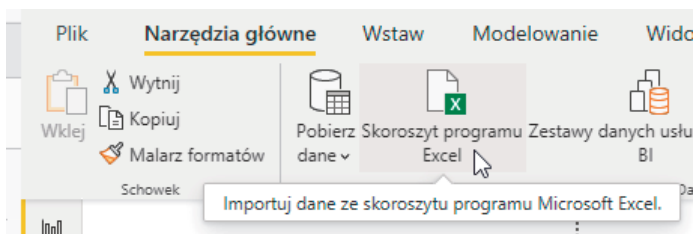
Rysunek 6.2. Anonimizacja, maskowanie danych



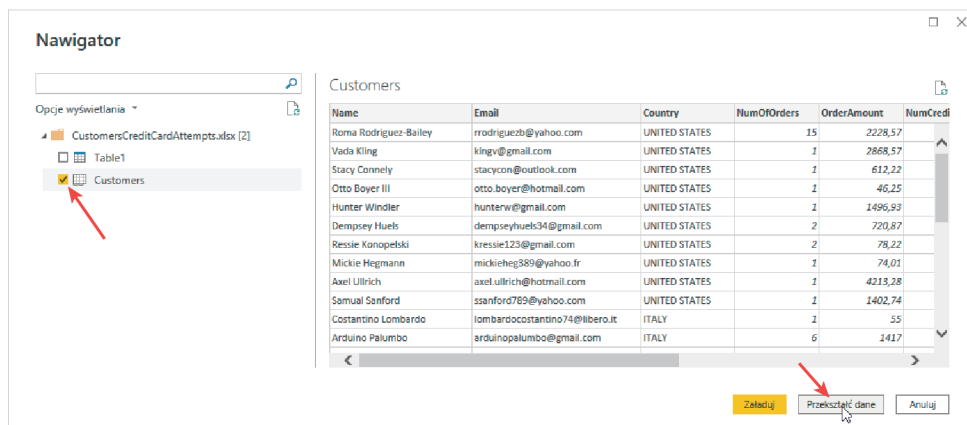
Rysunek 6.3. Proces pseudonimizacji

	A	B	C	D	E	F	G
1	Name	Email	Country	NumOfOrders	OrderAmount	NumCreditCardAttempts	Notes
2	Roma Rodriguez-Bailey	rrodriguezb@yahoo.com	UNITED STATES	15	2228.57	21	It seems the customer has another account with the name Ashley Stevenson and email kevin123b@hotmail.com
3	Vada Kling	kingv@gmail.com	UNITED STATES	1	2868.57	3	
4	Cleveland Connelly	clevecon@outlook.com	UNITED STATES	1	612.22	2	

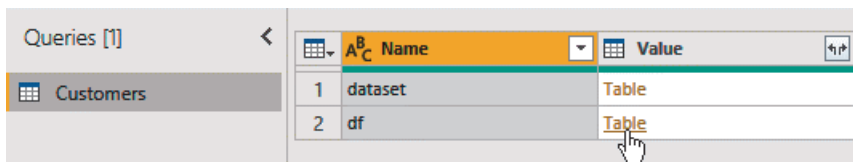
Rysunek 6.4. Dane z arkusza Excela do anonimizacji



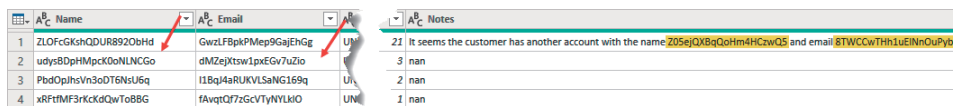
Rysunek 6.5. Importowanie danych z Excela



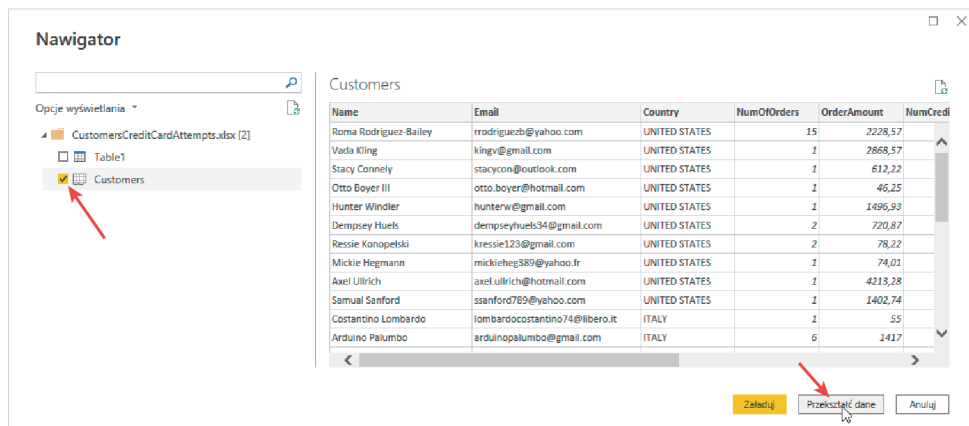
Rysunek 6.6. Wybierz arkusz Customers i kliknij Przekształć dane



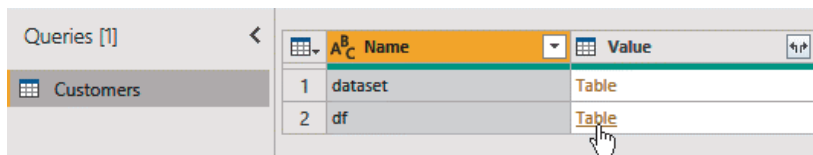
Rysunek 6.7. Wybieranie zestawu danych df w wyniku transformacji za pomocą skryptu w Pythonie



Rysunek 6.8. Przekształcony zestaw danych w wyniku działania skryptu Pythona



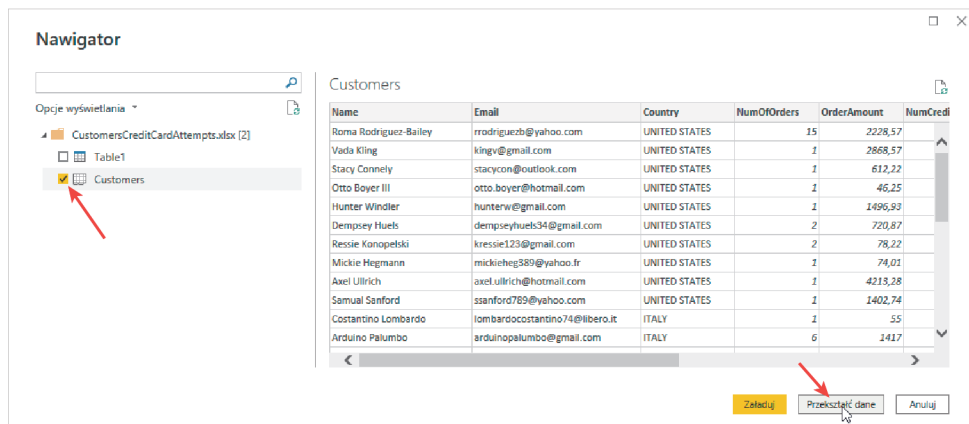
Rysunek 6.9 Wybór arkusza Customers i kliknięcie przycisku Przekształć dane



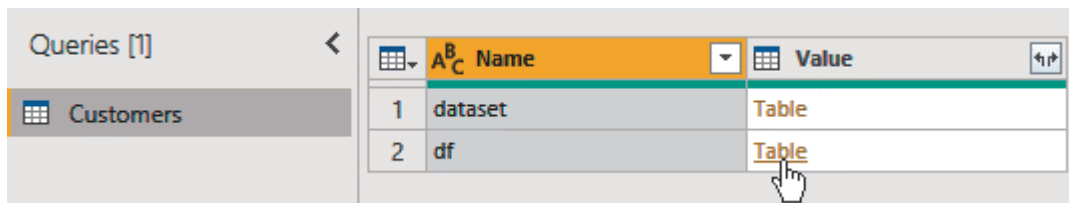
Rysunek 6.10. Wybieranie zestawu danych df jako wyniku przekształcenia za pomocą skryptu języka R

A ^B C Name	A ^B C Email	A ^B C Country	A ^B C Notes
1 Kathy Schumm	evelena46@gmail.com	UNITED STATES	21 It seems the customer has another account with the name Jahir Rolfson and email kiyoshi35@hegmann.com
2 Ms. Phyllis Miller DVM	rick.pfeffer@kemmer.com	UNITED STATES	3
3 Tillie Barrows	shantell81@yahoo.com	UNITED STATES	2
4 Mrs. Suzie Howell MD	pleas27@yahoo.com	UNITED STATES	1

Rysunek 6.11. Przekształcony zestaw danych w wyniku działania skryptu języka R



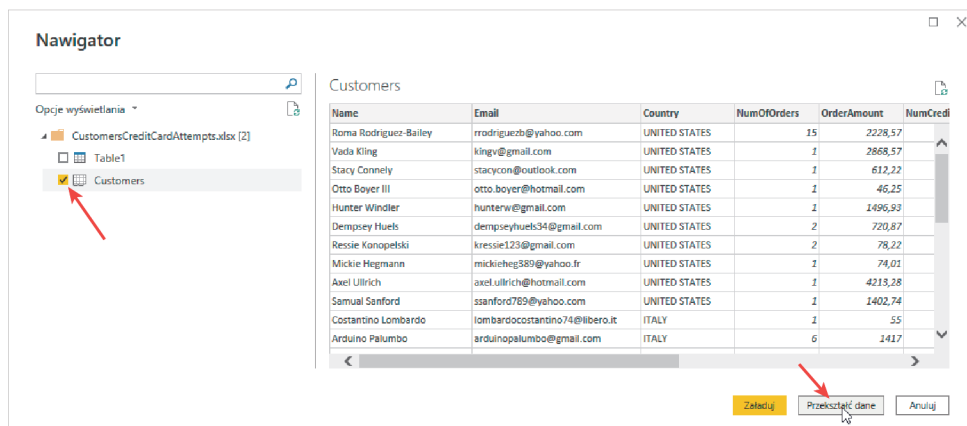
Rysunek 6.12. Wybierz arkusz Customers i kliknij Przekształć dane



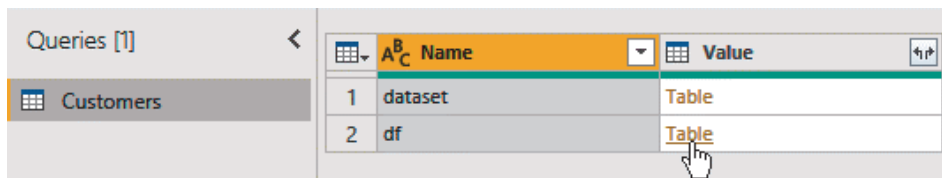
Rysunek 6.13. Wybór zbioru danych df w wyniku przekształcenia skryptu Python

A ^B Name	A ^B Email	A ^B Country	A ^B Notes
1 Sheena Santana	michaellivingston@example.com	UNITED STATES	21 It seems the customer has the name Maureen Wheeler and email oconnorrichard@example.com
2 Melinda Gray	danielle43@example.org	UNITED STATES	3 nan
3 Cleveland Connelly	andrewjackson@example.org	UNITED STATES	2 nan
4 Tiffany...	kimberly...@example.com	UNITED STATES	1 nan
9 Stephanie Daniels	stluc4@example.org	UNITED STATES	1 nan
10 Hailey Marshall	davidbrown@example.org	UNITED STATES	6 The customer used a different email address. They are christinatran@example.com and william47@example.org
11 Alfio Migliaccio	serenasofa@example.com	ITALY	1 nan
12 Dott. Galasso Grimaldi	bettoniflavia@example.com	ITALY	7 nan
13 Filippa Pedrazzini	cugiaadele@example.com	ITALY	1 nan

Rysunek 6.14. Przekształcony zestaw danych jako wynik działania skryptu Python



Rysunek 6.15. Wybierz arkusz Customers i kliknij przycisk Przekształć dane



Rysunek 6.16. Wybieranie zestawu danych df jako wyniku przekształcenia za pomocą skryptu języka R

A ^B Name	A ^B Email	A ^B Country	A ^B Notes
1 Kathy Schumm	evelena46@gmail.com	UNITED STATES	21 It seems the customer has another account with the name Jahir Rolfson and email kiyoshi35@hegmann.com
2 Ms. Phyllis Miller DVM	rick.pfeffer@kemmer.com	UNITED STATES	3 nan
3 Tillie Barrows	shantell81@yahoo.com	UNITED STATES	2 nan
4 Mrs. Suzie Howell MD	pleas27@yahoo.com	UNITED STATES	1 nan

Rysunek 6.17. Przekształcony zestaw danych w wyniku działania skryptu języka R

Rozdział 7. Zapisywanie danych z usługi Power BI do źródeł zewnętrznych

```
1 Col1,Col2,Col3CRLF
2 A,23,3.5CRLF
3 B,27,4.8CRLF
```

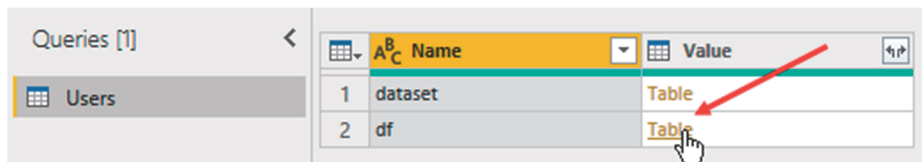
Rysunek 7.1. Przykład zawartości pliku CSV

```
1 Col1,Col2,Col3CRLF
2 A,23,3.5CRLF
3 B,27,4.8CRLF
4 "C,D",18,2.1CRLF
5 "E"F",19,2.5CRLF
6 "G"CRLF
7 "H",21,3.1CRLF
```

Rysunek 7.2. Przykład wartości CSV zawierających przecinki, cudzysłowy lub znaki podziału wiersza

	Col1	Col2	Col3
0	A	23	3.5
1	B	27	4.8
2	C,D	18	2.1
3	E"F	19	2.5
4	G\r\nH	21	3.1

Rysunek 7.3. Wynik ładowania zawartości przykładowego pliku CSV do obiektu DataFrame modułu pandas



Rysunek 7.4. Wybieranie zbioru danych df w wyniku transformacji za pomocą skryptu Pythona

1 ² 3	UserId	A ^B C Email	A ^B C BannedDate
1	1	@example.com	05/29/2018
2	4	firstname-lastname@example.com	06/07/2019
3	6	email@example-one.com	11/06/2017
4	7	email@example.co.jp	012/05/2018
5	8	email@example.name	06/7/2019
6	10	email@example.com	9/04/2017
7	11	email@subdomain.example.com	11/22/2018
8	12	email@123.123.123.123	07/31/17
9	14	firstname.lastname@example.com	11/22/18
10	15	firstname+lastname@example.com	05/16/2018
11	17	1234567890@example.com	05/16/2018
12	18	email@example.museum	03/26/2018
13	19	email@[123.123.123.123]	06/31/2017

Rysunek 7.5. Tabela zawierająca wyłącznie poprawne adresy e-mail

```
> data_df
# A tibble: 5 x 3
  col1      col2 col3
  <chr>    <dbl> <dbl>
1 "A"      23    3.5
2 "B"      27    4.8
3 "C,D"    18    2.1
4 "E\F"    19    2.5
5 "G\r\nH" 21    3.1
```

Rysunek 7.6. Funkcja read_csv poprawnie zaimportowała znaki \r\n

Queries [1]	A ^B C Name	Value
Users	1 dataset	Table
	2 df	Table

Rysunek 7.7. Wybór zbioru danych df w wyniku przekształcenia za pomocą skryptu języka R

	123 Userid	A ^B C Email	BannedDate	123 isInvalidDateFromRegex
1	1	@example.com	2018-05-29	1
2	2	example1@example.com/exampl...	2019-06-07	1
3	3	example33@example.com.	2018-02-05	1
4	4	firstname-lastname@example.com	2019-06-07	1
5	6	email@example-one.com	2017-11-06	1
6	8	email@example.name	2019-06-07	1
7	9	example	2018-10-22	1
8	10	email@example.com	2017-09-04	1
9	11	email@subdomain.example.com	2018-11-22	1
10	12	email@123.123.123.123	2017-07-31	1
11	14	firstname.lastname@example.com	2018-11-22	1
12	15	firstname+lastname@example.com	2018-05-16	1
13	16	example@example.c	2019-06-07	1
14	17	1234567890@example.com	2018-05-16	1
15	18	email@example.museum	2018-03-26	1

Rysunek 7.8. Tabela zawierająca tylko poprawne daty

data			
	Col1	Col2	Col3
0	A	23	3.5
1	B	27	4.8
2	C,D	18	2.1
3	E"F	19	2.5
4	G\nH	21	3.1

Rysunek 7.9. Wynik ładowania zawartości przykładowego pliku example.xlsx do obiektu DataFrame modułu pandas

	A	B	C
1	Col1	Col2	Col3
2	A	23	3.5
3	B	27	4.8
4	C,D	18	2.1
5	E"F	19	2.5
6	G H	21	3.1

Rysunek 7.10. Zawartość pliku Excel utworzonego za pomocą funkcji pandas

	A	B	C
1	Col1	Col2	Col3
2	A	23	3.5
3	B	27	4.8
4	C,D	18	2.1
5	E" F	19	2.5
6	G		
7	H	21	3.1

My data

Rysunek 7.11. Zawartość została zapisana w nazwanym arkuszu

123 Userid	A ^B Email	A ^B BannedDate	123 isEmailValidFromRegex	123 isValidDateFromRegex
1	_____@example.com	05/29/2018	1	1
2	example1@example.com/example2@exampl...	06/07/2019	1	1
3	example33@example.com.	02/05/2018	1	1
4	firstname-lastname@example.com	06/07/2019	1	1
5	example@example.com --> check	02/29/18	0	0
6	email@example-one.com	11/06/2017	1	1
7	email@example.co.jp	012/05/2018	1	0
8	email@example.name	06/7/2019	1	1
9	example	10/22/2018	0	1
10	email@example.com	9/04/2017	1	1
11	email@subdomain.example.com	11/22/2018	1	1
12	email@123.123.123.123	07/31/17	1	1
13	_.example.com	04/24/018	0	0
14	firstname.lastname@example.com	11/22/18	1	1
15	firstname+lastname@example.com	05/16/2018	1	1
16	example@example.c	6/7/2019	1	1
17	1234567890@example.com	05/16/2018	1	1
18	email@example.museum	03/26/2018	1	1
19	email@[123.123.123.123]	06/31/2017	0	0

Rysunek 7.12. Przekształcone dane zawierają zarówno flagi dla prawidłowych dat, jak i adresów e-mail

123 Userid	A ^B Email	BannedDate	123 isEmailValidFromRegex	123 isValidDateFromRegex
1	_____@example.com	2018-05-29	1	1
4	firstname-lastname@example.com	2019-06-07	1	1
6	email@example-one.com	2017-11-06	1	1
8	email@example.name	2019-06-07	1	1
10	email@example.com	2017-09-04	1	1
11	email@subdomain.example.com	2018-11-22	1	1
12	email@123.123.123.123	2017-07-31	1	1
14	firstname.lastname@example.com	2018-11-22	1	1
15	firstname+lastname@example.com	2018-05-16	1	1
17	1234567890@example.com	2018-05-16	1	1
18	email@example.museum	2018-03-26	1	1

Rysunek 7.13. Wyniki zawierają wiersze z prawidłowymi adresami e-mail i datami

```
# A tibble: 5 x 3
  Col1      Col2      Col3
  <chr>    <dbl>    <dbl>
1 "A"      23      3.5
2 "B"      27      4.8
3 "C,D"    18      2.1
4 "E\"F"    19      2.5
5 "G\r\nH" 21      3.1
```

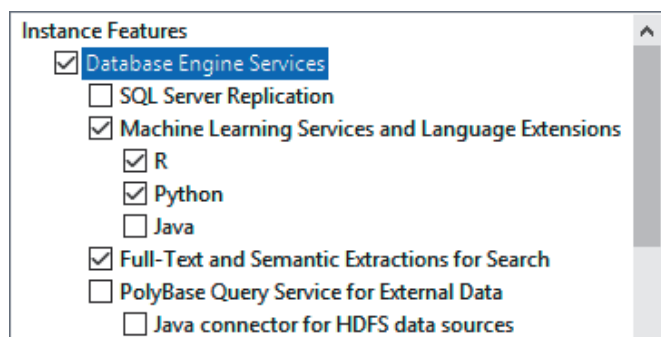
Rysunek 7.14. Odczyt danych programu Excel za pomocą funkcji read_xlsx

123 Userid	123 Email	123 BannedDate	123 isEmailValidFromRegex	123 isDateValidFromRegex
1	@example.com	05/29/2018	1	1
2	example1@example.com/example2@exampl...	06/07/2019	1	0
3	example33@example.com.	02/05/2018		0
4	firstname-lastname@example.com	06/07/2019	1	1
5	example@example.com -> check	02/29/18	0	0
6	email@example-one.com	11/06/2017	1	1
7	email@example.co.jp	012/05/2018	1	0
8	email@example.name	06/7/2019	1	1
9	example	10/22/2018	0	1
10	email@example.com	9/04/2017	1	1
11	email@subdomain.example.com	11/22/2018	1	1
12	email@123.123.123.123	07/31/17	1	1
13	@example.com	04/24/018	0	0
14	firstname.lastname@example.com	11/22/18	1	1
15	firstname+lastname@example.com	05/16/2018	1	1
16	example@example.c	6/7/2019	1	0
17	1234567890@example.com	05/16/2018	1	1
18	email@example.museum	03/26/2018	1	1
19	email@[123.123.123.123]	06/31/2017	0	0

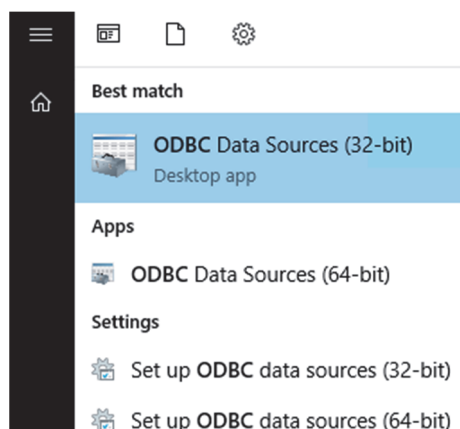
Rysunek 7.15. Przekształcone dane zawierają flagi dla dat i prawidłowych adresów e-mail

123 Userid	123 Email	123 BannedDate	123 isEmailValidFromRegex	123 isDateValidFromRegex
1	@example.com	05/29/2018	1	1
4	firstname-lastname@example.com	06/07/2019	1	1
6	email@example-one.com	11/06/2017	1	1
8	email@example.name	06/7/2019	1	1
10	email@example.com	9/04/2017	1	1
11	email@subdomain.example.com	11/22/2018	1	1
12	email@123.123.123.123	07/31/17	1	1
14	firstname.lastname@example.com	11/22/18	1	1
15	firstname+lastname@example.com	05/16/2018	1	1
17	1234567890@example.com	05/16/2018	1	1
18	email@example.museum	03/26/2018	1	1

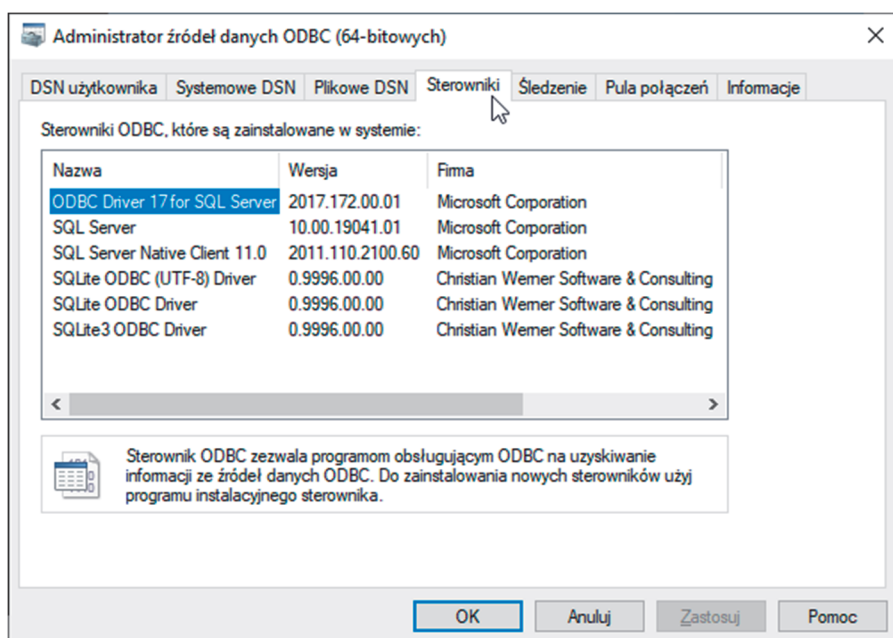
Rysunek 7.16. Dane wyjściowe zawierają wiersze z prawidłowymi datami i adresami e-mail



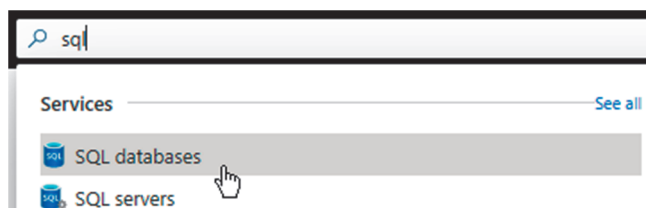
Rysunek 7.17. Sugerowane funkcje egzemplarza usługi instancji



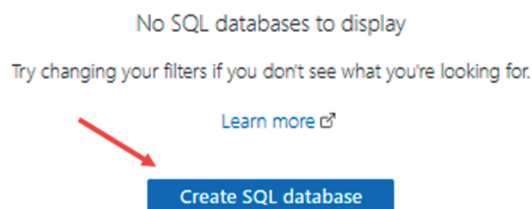
Rysunek 7.18. Narzędzia konfiguracji źródeł danych ODBC w systemie Windows



Rysunek 7.19. Poprawnie zainstalowany sterownik ODBC 17 dla programu SQL Server



Rysunek 7.20. Wybieranie baz danych SQL w witrynie Azure Portal



Rysunek 7.21. Kliknij przycisk Create SQL database

Database details

Enter required settings for this database, including picking a logical server and configuring the compute and storage resources

Database name * ✓

Server * ⓘ ▼
[Create new](#)

Rysunek 7.22. Wprowadzanie nazwy bazy danych i nowego serwera

Want to use SQL elastic pool? * ⓘ ☐ Yes ☒ No

Compute + storage * ⓘ **General Purpose**
Gen5, 2 vCores, 32 GB storage, zone redundant disabled
[Configure database](#)

Rysunek 7.23. Konfigurowanie usługi Compute

Basics Networking Security **Additional settings** Tags Review + create

Customize additional configuration parameters including collation & sample data.

Data source

Start with a blank database, restore from a backup or select sample data to populate your new database.

Use existing data * ☐ None ☐ Backup ☒ Sample

AdventureWorksLT will be created as the sample database.

Rysunek 7.24. Wybór instalacji przykładowej bazy danych

[Copy](#) [Restore](#) [Export](#) [Set server firewall](#) [Delete](#)

1. [^ Essentials](#) [Set server firewall](#)

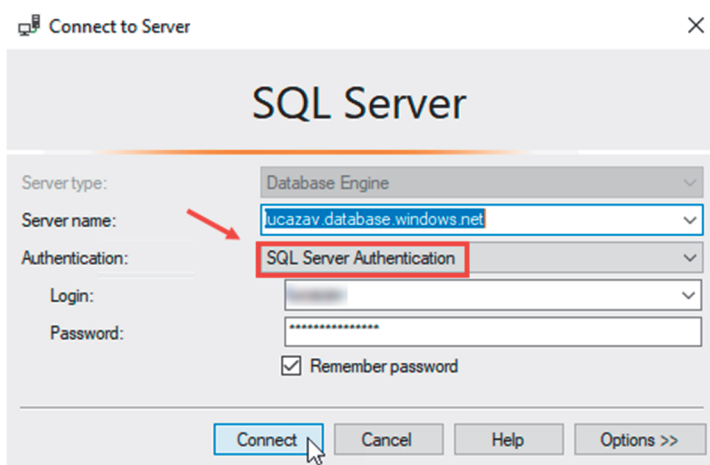
Rysunek 7.25. Ustawianie reguł w zaporze firewall serwera Azure SQL Server

Client IP address

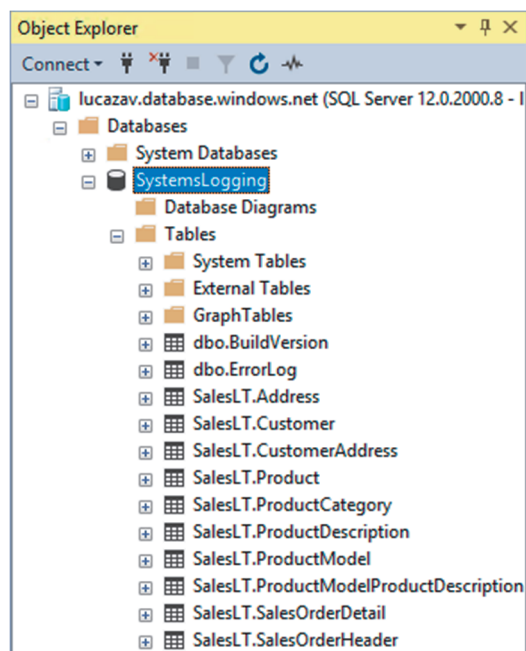
Rule name	Start IP	End IP
<input type="text"/>	<input type="text"/>	<input type="text"/> ...

No firewall rules configured.

Rysunek 7.26. Kopiowanie bieżącego adresu IP i korzystanie z niego w regule



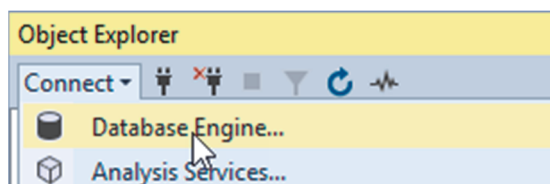
Rysunek 7.27. Nawiązywanie połączenia z usługą Azure SQL Database za pomocą programu SSMS



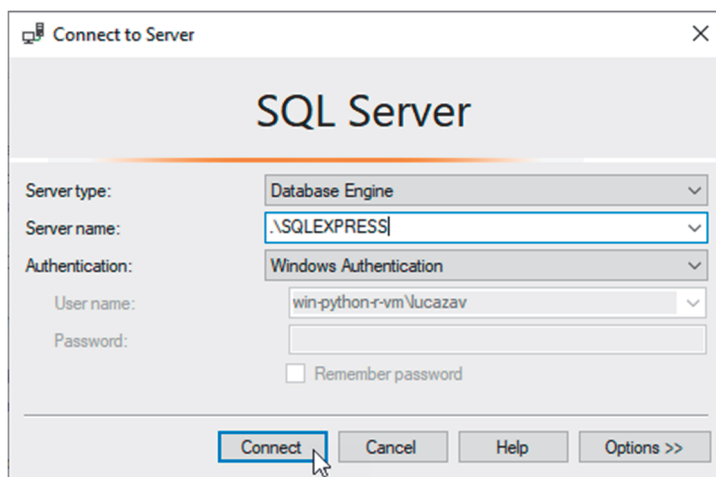
Rysunek 7.28. Połączenie z usługą Azure SQL Database

database_id	name
0	1 master
1	5 SystemsLogging

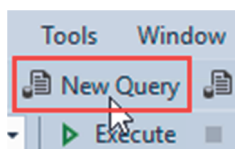
Rysunek 7.29. Wynik zapytania w usłudze Azure SQL Database



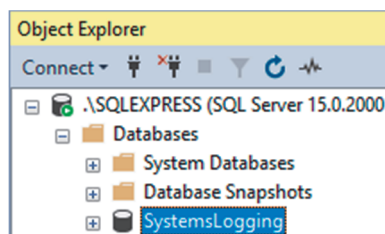
Rysunek 7.30. Nawiązywanie połączenia z silnikiem bazy danych w programie SSMS



Rysunek 7.31. Nawiązywanie połączenia z egzemplarzem bazy danych SQLExpress



Rysunek 7.32. Otwieranie nowego edytora zapytań



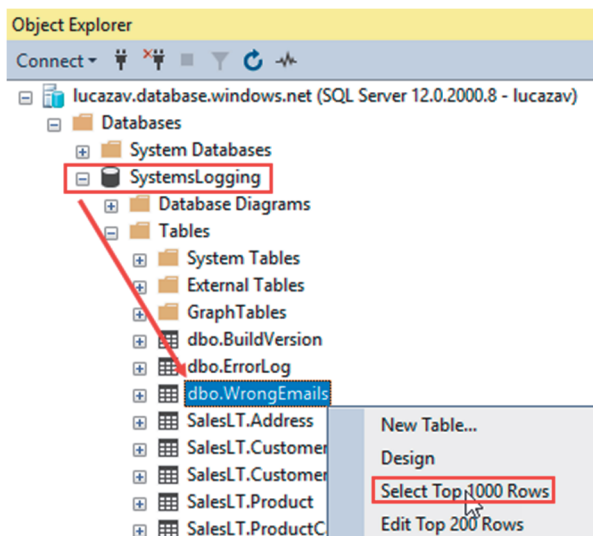
Rysunek 7.33. Nowa baza danych SystemsLogging w egzemplarzu serwera SQLExpress

	UserId	Email
0	202	a0@adventure-works.com
1	29943	a0@adventure-works.com
2	345	abigail0@adventure-works.com
3	29792	abigail0@adventure-works.com
4	511	abraham0@adventure-works.com

Rysunek 7.34. Zawartość tabeli WrongEmails

	1 ² 3 UserId	A ^B C Email	A ^B C BannedDate	1 ² 3 isEmailValidFromRegex
1		1 @example.com	05/29/2018	1
2		4 firstname-lastname@example.com	06/07/2019	1
3		6 email@example-one.com	11/06/2017	1
4		7 email@example.co.jp	012/05/2018	1
5		8 email@example.name	06/7/2019	1
6		10 email@example.com	9/04/2017	1
7		11 email@subdomain.example.com	11/22/2018	1

Rysunek 7.35. Tabela zawierająca wyłącznie prawidłowe adresy e-mail



Rysunek 7.36. Wykonywanie zapytania do tabeli WrongEmails w programie SSMS

SQLQuery1.sql - luc...ging ()

```

/***** Script for SelectTopNRows command fr
SELECT TOP (1000) [UserId]
,[Email]
FROM [dbo].[WrongEmails]

```

100 %

Results Messages

	UserId	Email
1	2	example1@example.com/example2@example.com
2	3	example33@example.com.
3	5	example@example.com --> check
4	9	example
5	13	..@example.com
6	16	example@example.c

Rysunek 7.37. Zawartość tabeli WrongEmails wyświetlona w programie SSMS

SystemsLogging - dbo@win-python-r-vm\SQLEXPRESS Microsoft SQL Server

- master
- msdb
- SystemsLogging
 - dbo
 - WrongEmails
 - UserId : int
 - Email : nvarchar
 - INFORMATION_SCHEMA
 - sys
 - tempdb

Rysunek 7.38. Eksplorator obiektów RStudio dla nawiązanego połączenia

```

> head(data)
  database_id      name
1           1      master
2           5 SystemsLogging
> |

```

Rysunek 7.39. Wynik zapytania w usłudze Azure SQL Database

```

> head(df)
  UserId      Email
1      2 example1@example.com/example2@example.com
2      3 example33@example.com.
3      5 example@example.com --> check
4      9 example
5     13 ..@example.com
6     16 example@example.c
> |

```

Rysunek 7.40. Zawartość tabeli WrongEmails

1 2 3	UserId	A B C Email	A B C BannedDate	1 2 3 isEmailValidFromRegex
1	1	@example.com	05/29/2018	1
2	4	firstname-lastname@example.com	06/07/2019	1
3	6	email@example-one.com	11/06/2017	1
4	7	email@example.co.jp	012/05/2018	1
5	8	email@example.name	06/7/2019	1
6	10	email@example.com	9/04/2017	1
7	11	email@subdomain.example.com	11/22/2018	1

Rysunek 7.41. Tabela zawierająca tylko prawidłowe adresy e-mail

SQLQuery1.sql - luc...ging (lucazav (82))*

```

/***** Script for SelectTopNRows command fr
SELECT TOP (1000) [UserId]
, [Email]
FROM [dbo].[WrongEmails]

```

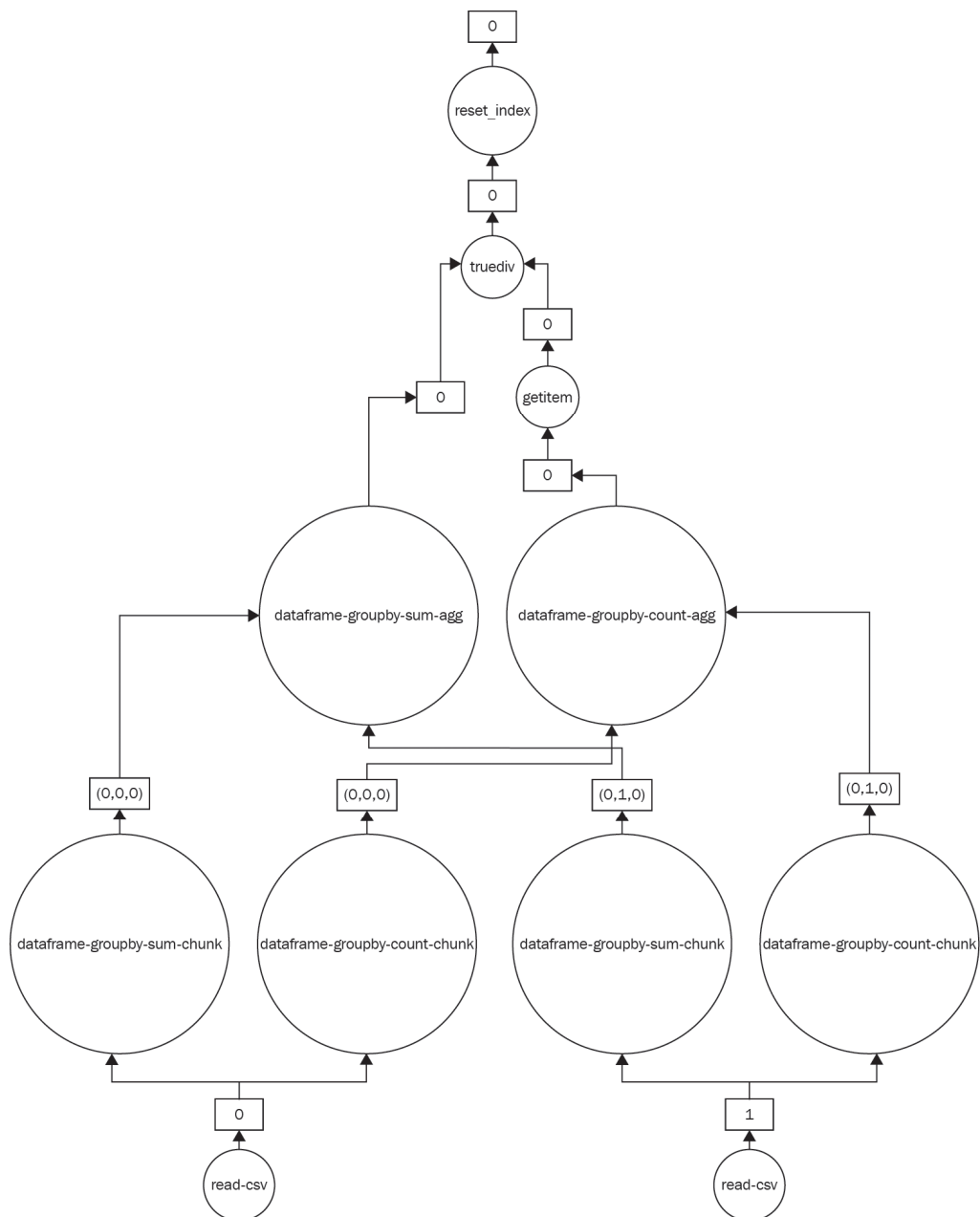
100 %

Results Messages

	UserId	Email
1	2	example1@example.com/example2@example.com
2	3	example33@example.com.
3	5	example@example.com -> check
4	9	example
5	13	.@example.com
6	16	example@example.c

Rysunek 7.42. Zawartość tabeli WrongEmails w SSMS

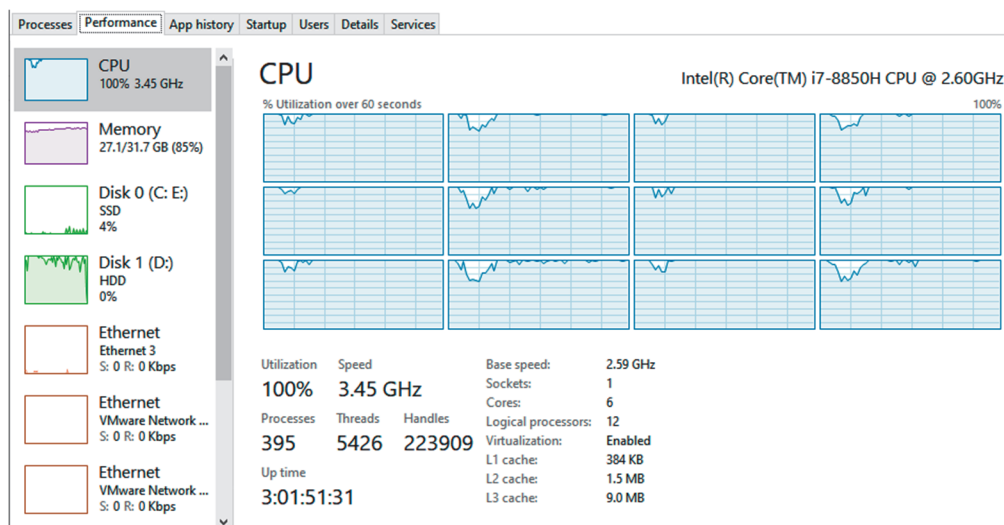
Rozdział 8. Ładowanie do usługi Power BI zbiorów danych przekraczających dostępną pamięć RAM



Rysunek 8.1. Wykres wykonywanych zadań

```
[7] with ProgressBar():...  
[#####] | 100% Completed | 3min 43.7s
```

Rysunek 8.2. Pasek postępu w Visual Studio Code



Rysunek 8.3. Obliczenia równoległe wyświetlane w Menedżerze zadań

```
[8] mean_dep_delay_df.head(10)
```

	YEAR	MONTH	DAY_OF_MONTH	ORIGIN	DEP_DELAY
0	1987	10	1	ABE	1.600000
1	1987	10	1	ABQ	2.494253
2	1987	10	1	AGS	2.000000
3	1987	10	1	ALB	5.843750
4	1987	10	1	AMA	0.086957
5	1987	10	1	ANC	1.142857
6	1987	10	1	APF	0.000000
7	1987	10	1	ATL	4.051604
8	1987	10	1	ATW	0.833333
9	1987	10	1	AUS	2.341463

Rysunek 8.4. Pierwsze 10 wierszy obiektu DataFrame pakietu pandas

Navigator

Display Options

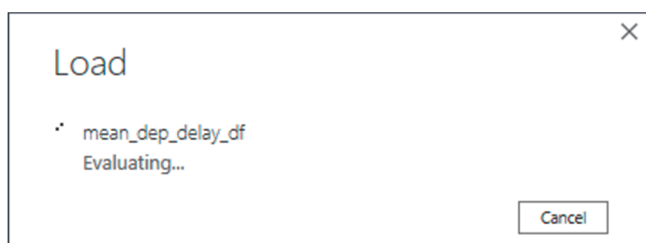
Python [1]

☒ mean_dep_delay_df

mean_dep_delay_df

YEAR	MONTH	DAY_OF_MONTH	ORIGIN	DEP_DELAY
1987	10	1	ABE	1.6
1987	10	1	ABQ	2.494252874
1987	10	1	AGS	2
1987	10	1	ALB	5.84375

Rysunek 8.5. Ramka danych mean_dep_delay_df poprawnie załadowana w usłudze Power BI



Rysunek 8.6. Ostatnia faza ładowania danych do usługi Power BI

mean_dep_delay_df.csv	
1	YEAR, MONTH, DAY_OF_MONTH, ORIGIN, DEP_DELAY
2	1987, 10, 1, ABE, 1.6
3	1987, 10, 1, ABQ, 2.4942528735632186
4	1987, 10, 1, AGS, 2.0
5	1987, 10, 1, ALB, 5.84375

Rysunek 8.7. Zawartość pliku CSV utworzonego za pomocą skryptu języka Python w usłudze Power BI

```
> head(mean_dep_delay_df, 10)
# A tibble: 10 x 7
# Groups:   YEAR, MONTH, DAY_OF_MONTH [1]
  YEAR MONTH DAY_OF_MONTH ORIGIN avg_delay
  <int> <int>    <int> <chr>    <dbl>
1  1987    10         1 ABE      1.6
2  1987    10         1 ABQ      2.49
3  1987    10         1 ACV      14
4  1987    10         1 AGS       2
5  1987    10         1 ALB      5.84
6  1987    10         1 ALO       0
7  1987    10         1 AMA     0.0870
8  1987    10         1 ANC      1.14
9  1987    10         1 APF       0
10 1987    10         1 ATL      4.02
>
```

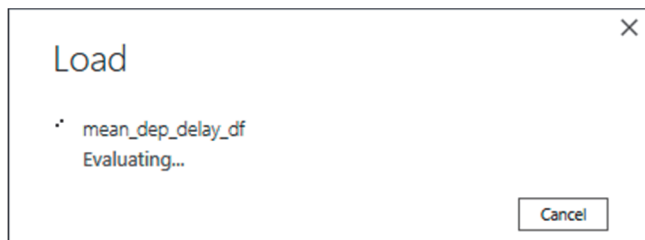
Rysunek 8.8. Pierwsze wiersze obiektu tibble z danymi średnich opóźnień

Navigator

mean_dep_delay_df
Preview downloaded on Thursday, April 1, 2021

YEAR	MONTH	DAY_OF_MONTH	ORIGIN	avg_delay
1987	10	1	ABE	1.6
1987	10	1	ABQ	2.494252874
1987	10	1	ACV	14
1987	10	1	AGS	2

Rysunek 8.9. Ramka danych mean_dep_delay_df załadowana w usłudze Power BI



Rysunek 8.10. Ostatnia faza ładowania danych do usługi Power BI

```
1 YEAR,MONTH, DAY_OF_MONTH,ORIGIN, avg_delay
2 1987,10,1,ABE,1.6
3 1987,10,1,ABQ,2.4942528735632186
4 1987,10,1,ACV,14
5 1987,10,1,AGS,2
6 1987,10,1,ALB,5.84375
```

Rysunek 8.11. Zawartość pliku CSV utworzonego za pomocą skryptu języka R w usłudze Power BI

Rozdział 9. Wywoływanie zewnętrznych interfejsów API w celu wzbogacania danych

 Bing maps | Dev Center

My account ▾

Data sources ▾

Announcements

Contacts & Info

My keys

Key created successfully.



Click [here](#) to create a new key.

Click [here](#) to download complete list of keys.

View Specific Key:

Enter key to search...



Application name	Key details	Enable Preview for All Keys 
geocoding-test	<p>Key: Show key</p> <p>Application Uri:</p> <p>Key type: Basic / Dev/Test</p> <p>Created date: 05/07/2021</p> <p>Expiration date: None</p> <p>Key Status: Enabled</p> <p>Security Enabled: No</p>	<p>Update</p> <p>Copy key</p> <p>Usage Report</p> <p>Enable Security</p> <p>Enable Preview </p>

Rysunek 9.1. Potwierdzenie utworzenia klucza w serwisie Bing Maps

JSON	Raw Data	Headers
Save	Copy	Collapse All Expand All
▼ brandLogoUri:	▼ copyright:	▼ resourceSets:
authenticationResultCode:	"ValidCredentials"	estimatedTotal: 1
	"http://dev.virtualearth.net/Branding/Logo_powered_by.png"	▼ resources:
	"Copyright © 2021 Microsoft and its suppliers. All rights reserved. This API cannot be accessed and the content and any results may not be used, reproduced or transmitted in any manner without express written permission from Microsoft Corporation."	▼ 0:
		__type: "Location:http://schemas.microsoft.com/search/local/ws/rest/v1"
		▼ bbox:
		0: 47.63550528242932
		1: -122.13598214165333
		2: 47.643230717570674
		3: -122.12069585834666
		name: "1 Microsoft Way, Redmond, WA 98052"

Rysunek 9.2. Pierwsze geokodowanie przy użyciu interfejsu API Bing Maps Locations z wykorzystaniem przeglądarki

[bazowy_ur1]query=[adres]?key=[KLUCZ_UWIERZYTELNIANIA]

Rysunek 9.3. Format adresu URL żądania GET do interfejsu API Bing Maps Locations



Rysunek 9.4. Struktura wizualna odpowiedzi interfejsu API Bing Maps Locations

	full_address	lat_true	lon_true	numResources	formattedAddress	latitude	longitude	text	status	url
0	200 K St NE, Washington DC, 20002	38.903155	-77.003274	1	200 K St NE, Washington, DC 20002	38.903161	-77.003053	{"authent"...	OK	http://dev.virtualearth...
1	200 K St North East, Washington DC, 20002	38.903155	-77.003274	1	200 K St NE, Washington, DC 20002	38.903161	-77.003053	{"authent"...	OK	http://dev.virtualearth...
2	200 K St Northeast, DC	38.903155	-77.003274	1	200 K St NE, Washington, DC 20002	38.903161	-77.003053	{"authent"...	OK	http://dev.virtualearth...

Rysunek 9.5. Zawartość ramki DataFrame z danymi geokodowania

```
> tbl_enriched
# A tibble: 120 x 11
  full_address lat_true lon_true numOfResources formattedAddress lat lng statusDesc statusCode text url
  <chr> <dbl> <dbl> <int> <chr> <dbl> <dbl> <chr> <int> <chr> <chr>
1 200 K St NE, Was~ 38.9 -77.0 1 200 K St NE, Washin~ 38.9 -77.0 OK 200 {"\authentcat~ http://dev
2 200 K St North E~ 38.9 -77.0 1 200 K St NE, Washin~ 38.9 -77.0 OK 200 {"\authentcat~ http://dev
3 200 K St Northea~ 38.9 -77.0 1 200 K St NE, Washin~ 38.9 -77.0 OK 200 {"\authentcat~ http://dev
4 500 L'enfant Pla~ 38.9 -77.0 1 500 L'enfant Plaza S~ 38.9 -77.0 OK 200 {"\authentcat~ http://dev
5 500 Lenfant Plaz~ 38.9 -77.0 1 500 L'Enfant Plaza ~ 38.9 -77.0 OK 200 {"\authentcat~ http://dev
6 500 Lenfant Plaz~ 38.9 -77.0 1 500 L'Enfant Plaza ~ 38.9 -77.0 OK 200 {"\authentcat~ http://dev
7 2197 Plumleigh D~ 37.5 -122. 1 2197 Plumleigh Dr, ~ 37.5 -122. OK 200 {"\authentcat~ http://dev
8 2197 Plumleigh D~ 37.5 -122. 1 2197 Plumleigh Dr, ~ 37.5 -122. OK 200 {"\authentcat~ http://dev
9 2197 Plumleigh D~ 37.5 -122. 1 2197 Plumleigh Dr, ~ 37.5 -122. OK 200 {"\authentcat~ http://dev
10 5034 Curtis St, ~ 37.5 -122. 1 5034 Curtis St, Fre~ 37.5 -122. OK 200 {"\authentcat~ http://dev
# ... with 110 more rows
> |
```

Rysunek 9.6. Zawartość obiektu DataFrame wzbogacona danymi geokodowania

Navigator

Display Options

Python [1]

☒ enriched_df

full_address	formattedAddress
200 K St NE, Washington DC, 20002	200 K St NE, Washington, DC 20002
200 K St North East, Washington DC, 20002	200 K St NE, Washington, DC 20002
200 K St Northeast, DC	200 K St NE, Washington, DC 20002
500 L'enfant Plaza SW, Washington DC, 200024	500 Lenfant Plaza SW, Washington,
500 Lenfant Plaza SW, Washington DC, 200024	500 Lenfant Plaza SW, Washington,

Rysunek 9.7. Obiekt DataFrame enriched_df załadowany w usłudze Power BI

Navigator

Display Options

R [3]

☐ tbl

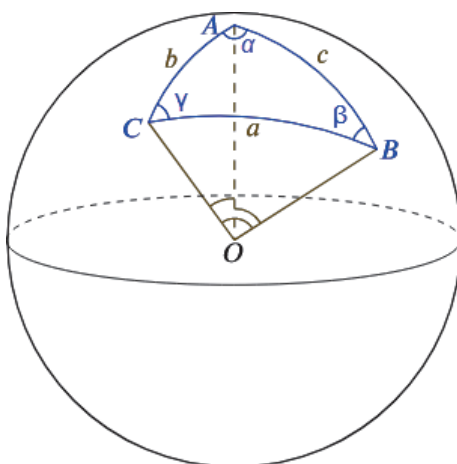
☒ tbl_enriched

☐ tbl_orig

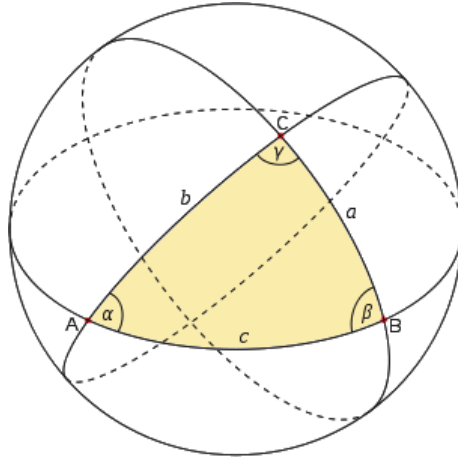
full_address	formattedAddress
200 K St NE, Washington DC, 20002	200 K St NE, Washington, DC 20002
200 K St North East, Washington DC, 20002	200 K St NE, Washington, DC 20002
200 K St Northeast, DC	200 K St NE, Washington, DC 20002
500 L'enfant Plaza SW, Washington DC, 200024	500 Lenfant Plaza SW, Washington,
500 Lenfant Plaza SW, Washington DC, 200024	500 Lenfant Plaza SW, Washington,

Rysunek 9.8. Obiekt DataFrame tbl_enriched załadowany w usłudze Power BI

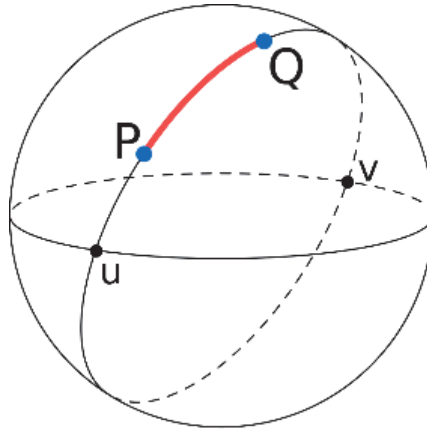
Rozdział 10. Obliczanie kolumn przy użyciu złożonych algorytmów



Rysunek 10.1. Trójkąt sferyczny



Rysunek 10.2. Wielkie okręgi, które generują trójkąt sferyczny



Rysunek 10.3. Odległość wielkiego okręgu między punktami P i Q

$$d = 2R \arcsin \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos \phi_1 \cos \phi_2 \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right)$$

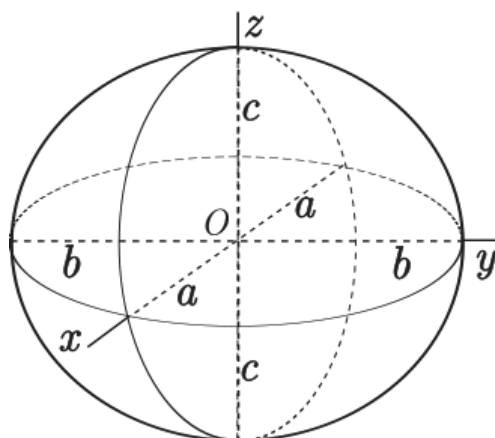
Rysunek 10.4. Wzór na odległość między dwoma punktami wynikający z prawa cosinusów

$$\text{hav } \theta = \frac{1 - \cos \theta}{2} = \sin^2 \left(\frac{\theta}{2} \right)$$

Rysunek 10.5. Definicja funkcji haversine

$$d = R \text{hav}^{-1}(\text{hav}(\phi_2 - \phi_1) + \cos \phi_1 \cos \phi_2 \text{hav}(\lambda_2 - \lambda_1))$$

Rysunek 10.6. Definicja wzoru odległości Haversinesa



Rysunek 10.7. Reprezentacja elipsoidalna

name	haversineDistanceFromJFK	karneyDistanceFromJFK	haversineDistanceFromLGA	karneyDistanceFromLGA
Element New York Times Square West	22103.493557	22130.681538	10314.345715	10338.546883
Sanctuary Hotel New York Times Square	21696.697737	21720.930350	9541.202663	9563.793968
The Jane	22309.739387	22345.798923	12316.656741	12342.187319
Hotel Olcott	22651.190545	22669.169997	8832.336480	8854.870363
Fairfield Inn & Suites New York Manhattan/Time...	22156.605785	22184.919308	10568.505016	10593.031275

Rysunek 10.8. Dodano odległości od hoteli do lotnisk JFK i LGA obliczone metodami Haversine i Karney

Nawigator

Opcje wyświetlania ▾

☒ hotels-ny.xlsx [1]

☒ Sheet 1

Sheet 1

objectId	latitude	longitude	name
E4tVCEHyZd	40,75584	-73,99178	Element New York Times Square West
L1kE2gwRDA	40,7585	-73,98317	Sanctuary Hotel New York Times Square
KyHuNG0abX	40,738224	-74,009476	The Jane
BKp7Egd1m	40,77704	-73,97749	Hotel Olcott

Rysunek 10.9. Wybieranie arkusza Sheet 1

	1.2 haversineDistanceFromJFK	1.2 karneyDistanceFromJFK	1.2 haversineDistanceFromLGA	1.2 karneyDistanceFromLGA
1	22103.49356	22130.68154	10314.34571	10338.54688
2	21696.69774	21720.93035	9541.202663	9563.793968
3	22309.73939	22345.79892	12316.65674	12342.18732
4	22651.19055	22669.17	8832.33648	8854.870363

Rysunek 10.10. Wynik transformacji za pomocą skryptu Pythona

```
# A tibble: 100 x 16
  objectId latitude longitude name type address stars phone website state city pl haversineDistanceFromLGA karneyDistanceFromLGA
  <chr> <dbl> <dbl> <chr> <chr> <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <dbl> <dbl>
1 E4tVCEHyZd 40.8 -74.0 Element New York Times Square We 3 1844~ http:// New ~ New ~ <dbl> 22103. 22131.
2 L1kEZgwRDA 40.8 -74.0 Sanctuary Hotel New York Times Sq 4 1646~ http:// New ~ New ~ <dbl> 21697. 21721.
3 KyHuNGOabX 40.7 -74.0 The Jane 0 NA NA New ~ New ~ <dbl> 22310. 22346.
4 BKp7Egdj1m 40.8 -74.0 Hotel Olcott 2 NA NA New ~ New ~ <dbl> 22651. 22669.
5 BvCu5xD~ 40.8 -74.0 Fairview Hotel 38 E 4~ 3 1646~ https:// New ~ New ~ <dbl> 22157. 22185.
6 22toTiz~ 40.7 -74.0 Soho Hotel 27 W 7~ 3 1646~ http:// New ~ New ~ <dbl> 20247. 20284.
7 Nch0nza~ 40.8 -74.0 AmeriStar Hotel 230 We~ 3 1646~ http:// New ~ New ~ <dbl> 22065. 22088.
8 Z2oXs3w~ 40.7 -74.0 Lafa Hotel 38 E 4~ 3 +1 2~ http:// New ~ New ~ <dbl> 20440. 20474.
9 NhhvKa1~ 40.7 -74.0 The Jane 0 NA NA New ~ New ~ <dbl> 19875. 19913.
10 16C4Hxx~ 40.8 -74.0 Nati Hotel 315 W ~ 0 NA NA New ~ New ~ <dbl> 22342. 22364.
# ... with 90 more rows, and 2 more variables: haversineDistanceFromLGA <dbl>, karneyDistanceFromLGA <dbl>
>
```

Rysunek 10.11. Dane hoteli wzbogacone informacjami o odległościach

Navigator

Opcje wyświetlania ▾

hotels-ny.xlsx [1]

Sheet 1

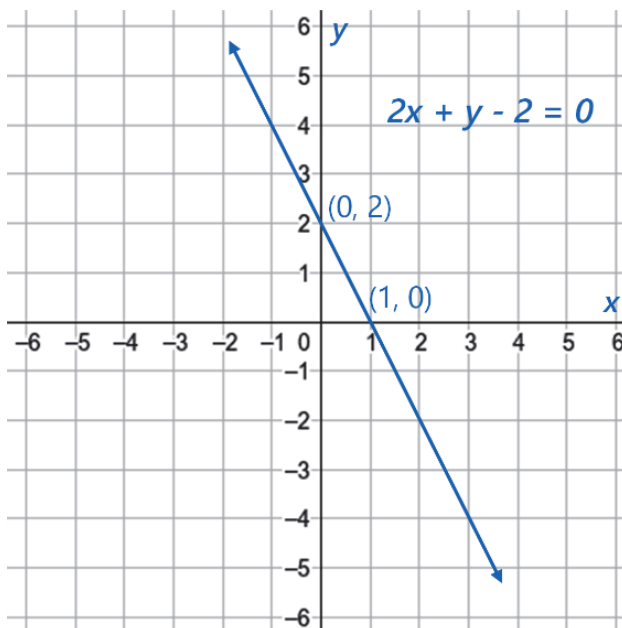
Sheet 1

objectId	latitude	longitude	name
E4tVCEHyZd	40.75584	-73.99178	Element New York Times Square We
L1kEZgwRDA	40.7585	-73.98317	Sanctuary Hotel New York Times Sq
KyHuNGOabX	40.738224	-74.009476	The Jane
BKp7Egdj1m	40.77704	-73.97749	Hotel Olcott

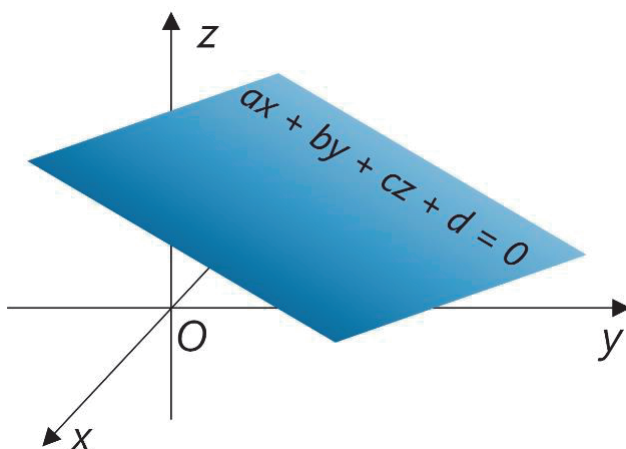
Rysunek 10.12. Wybieranie arkusza Sheet 1

	1.2 haversineDistanceFromJFK ▾	1.2 karneyDistanceFromJFK ▾	1.2 haversineDistanceFromLGA ▾	1.2 karneyDistanceFromLGA ▾
1	22103.49356	22130.68154	10314.34571	10338.54688
2	21696.69774	21720.93035	9541.202663	9563.793968
3	22309.73939	22345.79892	12316.65674	12342.18732
4	22651.19055	22669.17	8832.33648	8854.870363

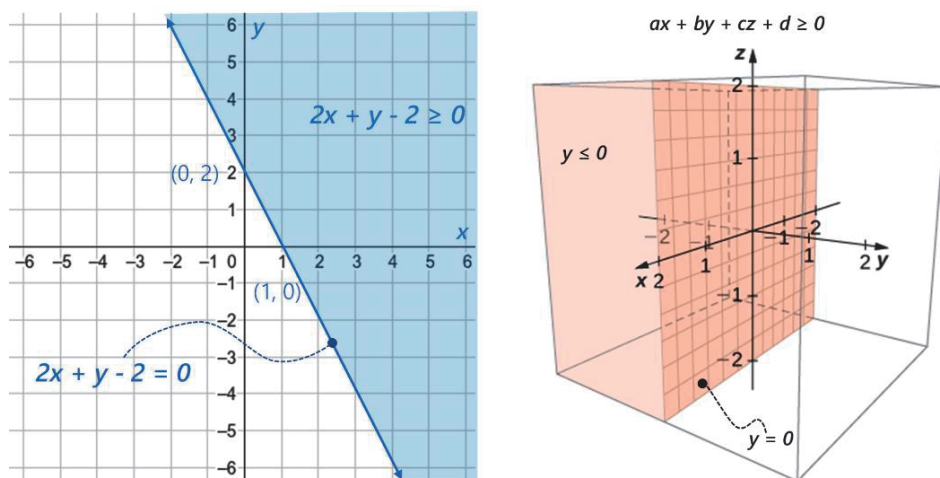
Rysunek 10.13. Wynik transformacji wykonanej za pomocą skryptu Pythona



Rysunek 10.14. Reprezentacja równania liniowego $2x + y = 2$



Rysunek 10.15. Reprezentacja ogólnego równania liniowego w postaci $ax + by + cz + d = 0$

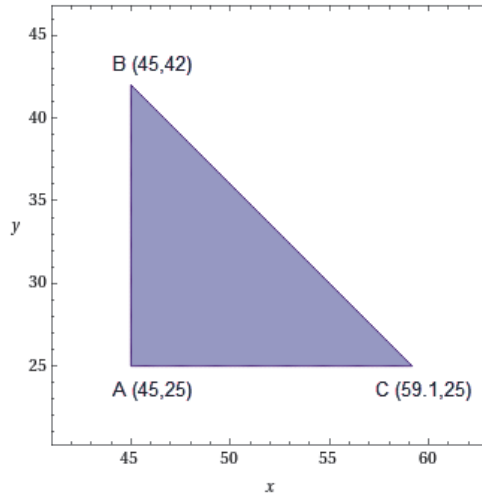


Rysunek 10.16. Reprezentacja ogólnych nierówności liniowych z dwiema lub trzema zmiennymi

Input interpretation:

plot $30x + 25y \leq 2400 \wedge x \geq 45 \wedge y \geq 25$

Inequality plot:



Rysunek 10.17. Obszar wykonalności narysowany w serwisie WolframAlpha

warehouse_name	product_qty
Magazyn ITA	50000
Magazyn DEU	30000
Magazyn JPN	40000
Magazyn USA	55000

Rysunek 10.18. Liczby sztuk produktu dostępnych w magazynach

country_name	product_qty
Włochy	40000
Francja	15000
Niemcy	25000
Japonia	45000
Chiny	25000
USA	25000

Rysunek 10.19. Liczby sztuk produktu zamawianych przez poszczególne kraje

warehouse_name	country_name	shipping_cost
Magazyn ITA	Włochy	8
Magazyn ITA	Francja	18
Magazyn ITA	Niemcy	14
Magazyn ITA	Japonia	40
Magazyn ITA	Chiny	40
Magazyn ITA	USA	25
Magazyn DEU	Włochy	12
Magazyn DEU	Francja	10
Magazyn DEU	Niemcy	8
Magazyn DEU	Japonia	18
Magazyn DEU	Chiny	40
Magazyn DEU	USA	18
Magazyn JPN	Włochy	34
Magazyn JPN	Francja	32
Magazyn JPN	Niemcy	30
Magazyn JPN	Japonia	10
Magazyn JPN	Chiny	33
Magazyn JPN	USA	35
Magazyn USA	Włochy	25
Magazyn USA	Francja	20
Magazyn USA	Niemcy	18
Magazyn USA	Japonia	35
Magazyn USA	Chiny	30
Magazyn USA	USA	10

Rysunek 10.20. Koszty wysyłki produktu z magazynów do krajów zamawiających

Definitions	Włochy (j=1)	Francja (j=2)	Niemcy (j=3)	Japonia (j=4)	Chiny (j=5)	USA (j=6)
Magazyn ITA (i=1)	x_{11}, C_{11}	x_{12}, C_{12}	x_{13}, C_{13}	x_{14}, C_{14}	x_{15}, C_{15}	x_{16}, C_{16}
Magazyn DEU (i=2)	x_{21}, C_{21}	x_{22}, C_{22}	x_{23}, C_{23}	x_{24}, C_{24}	x_{25}, C_{25}	x_{26}, C_{26}
Magazyn JPN (i=3)	x_{31}, C_{31}	x_{32}, C_{32}	x_{33}, C_{33}	x_{34}, C_{34}	x_{35}, C_{35}	x_{36}, C_{36}
Magazyn USA (i=4)	x_{41}, C_{41}	x_{42}, C_{42}	x_{43}, C_{43}	x_{44}, C_{44}	x_{45}, C_{45}	x_{46}, C_{46}

Rysunek 10.21. Definicja macierzy liczby sztuk wysyłanych produktów i związanych z tym kosztów

	Włochy	Francja	Niemcy	Japonia	Chiny	USA
Magazyn ITA	40000.0	0.0	10000.0	0.0	0.0	0.0
Magazyn DEU	0.0	10000.0	15000.0	5000.0	0.0	0.0
Magazyn JPN	0.0	0.0	0.0	40000.0	0.0	0.0
Magazyn USA	0.0	5000.0	0.0	0.0	25000.0	25000.0

Rysunek 10.22. Liczby sztuk wysłanych produktów zgodnie ze znalezionym rozwiązaniem

	liczba
Magazyn ITA	50000.0
Magazyn DEU	30000.0
Magazyn JPN	40000.0
Magazyn USA	55000.0

Rysunek 10.23. Łączne liczby sztuk produktów wysłane z poszczególnych magazynów

Nawigator

Opcje wyświetlania ▾

RetailData.xlsx [8]

☒ CountryDemand
☒ ShippingCost
☐ Table5
☒ WarehouseSupply
☐ Country Demand
☐ Problem
☐ Shipping Cost
☐ Warehouse Supply

WarehouseSupply
Data pobrania podglądu: piątek

warehouse_name	product_qty
Magazyn ITA	50000
Magazyn DEU	30000
Magazyn JPN	40000
Magazyn USA	55000

Rysunek 10.24. Wybieranie trzech arkuszy ze skoroszytu Excel

Zapytania [3]

CountryDemand
ShippingCost
WarehouseSupply

fx

Table.TransformColumnTypes(CountryDemand_Table,{

country_name	product_qty
1 Włochy	40000
2 Francja	15000
3 Niemcy	25000
4 Japonia	45000
5 Chiny	25000
6 USA	25000

Ustawienia zapytania

WŁAŚCIWOŚCI
Nazwa
CountryDemand
Wszystkie właściwości

ZASTOSOWANE KROKI
Źródło
Nawigacja
Zmieniono typ

Rysunek 10.25. Zapytanie CountryDemand wraz ze stosem zastosowanych kroków

Scal zapytania ▾

Scal zapytania
Scal zapytania jako nowe
Połącz

Analiza tekstu
Przetwarzanie obrazów
Machine Learning

Scal to zapytanie z innym zapytaniem w tym pliku.

Rysunek 10.26. Scalanie innego zapytania z zapytaniem Scalanie1

Opcje

GLOBALNE

Ładowanie danych
 Edytor Power Query
 DirectQuery
 Obsługa skryptów języka R
 Wyświetlanie wyników
 Zabezpieczenia
Prywatność

Poziomy prywatności

- ☐ Zawsze łącz dane zgodnie z ustawieniami poziomu prywatności dla każdego źródła
- ☐ Połącz dane zgodnie z ustawieniami poziomu prywatności dla każdego pliku
- ☒ **Zawsze ignoruj ustawienia poziomu prywatności**

To ustawienie może spowodować ujawnienie poufnych informacji nieupoważnionej osobie.

Rysunek 10.27. Opcje w ustawieniach poziomów prywatności

	A^B_C Name	Value
1	dataset	Table
2	result_df	Table


Rysunek 10.28. Zaznacz tabelę result_df










	A^B_C warehouse_name	A^B_C country_name	A^B_C shipped_qty	A^B_C cost
1	Magazyn ITA	Włochy	40000.0	320000.0
2	Magazyn ITA	Francja	0.0	0.0
3	Magazyn ITA	Niemcy	10000.0	140000.0
4	Magazyn ITA	Japonia	0.0	0.0
5	Magazyn ITA	Chiny	0.0	0.0
6	Magazyn ITA	USA	0.0	0.0

Rysunek 10.29. Zawartość tabeli result_df

Nawigator

Wyszukiwanie:

Opcje wyświetlania 

-  RetailData.xlsx [8]
 - ☒  CountryDemand
 - ☒  ShippingCost
 - ☐  Table5
 - ☒  WarehouseSupply
 - ☐  Country Demand
 - ☐  Problem
 - ☐  Shipping Cost
 - ☐  Warehouse Supply

WarehouseSupply

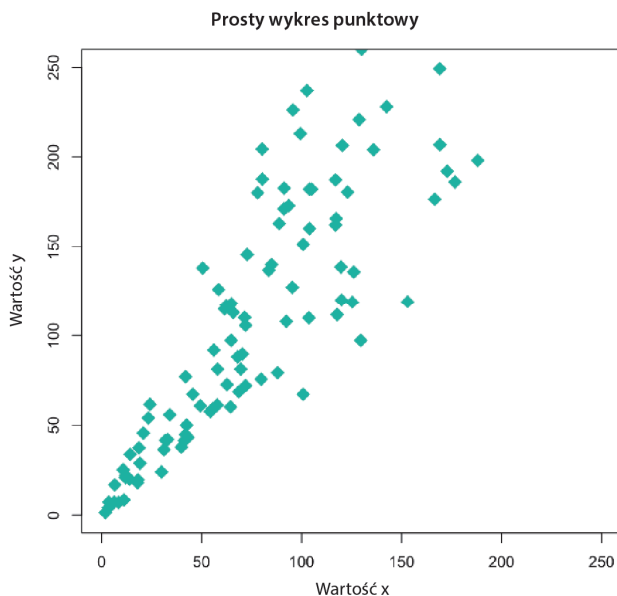
warehouse_name	product_qty
Magazyn ITA	50000
Magazyn DEU	30000
Magazyn JPN	40000
Magazyn USA	55000

Rysunek 10.30. Wybieranie trzech arkuszy skoroszytu Excela

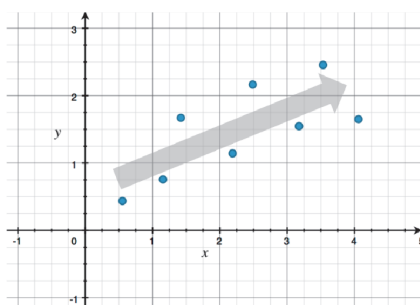
	A ^B _C warehouse_name	A ^B _C country_name	1.2 shipped_qty	1.2 cost
1	Magazyn ITA	Włochy	40000	320000
2	Magazyn ITA	Francja	0	0
3	Magazyn ITA	Niemcy	10000	140000
4	Magazyn ITA	Japonia	0	0
5	Magazyn ITA	Chiny	0	0
6	Magazyn ITA	USA	0	0
7	Magazyn DEU	Włochy	0	0

Rysunek 10.31. Zawartość tabeli WarehouseSupply

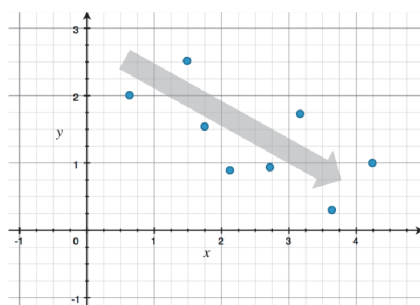
Rozdział 11. Dodawanie statystyk: powiązania



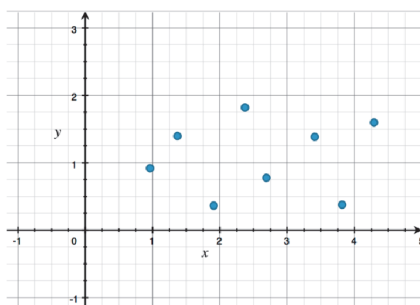
Rysunek 11.1. Prosty wykres punktowy



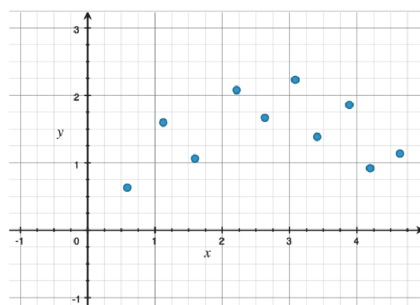
Kierunek dodatni



Kierunek ujemny

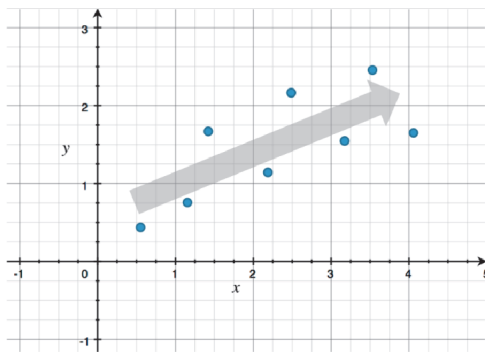


Niezdefiniowany
brak powiązania

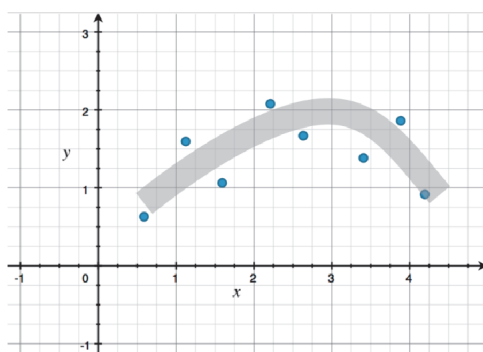


Niezdefiniowany
rosnący i malejący

Rysunek 11.2. Typy kierunków powiązań

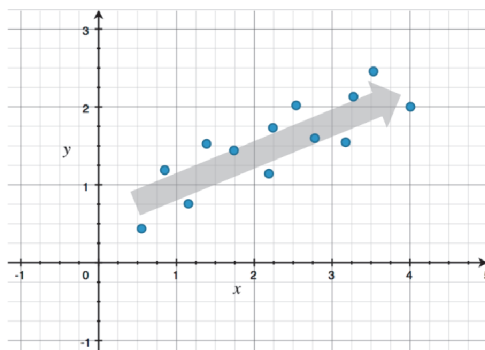


Liniowe

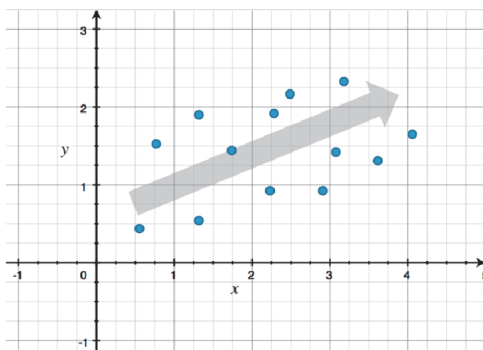


Krzywoliniowe
nieliniowe

Rysunek 11.3. Formy powiązania



Powiązanie silne

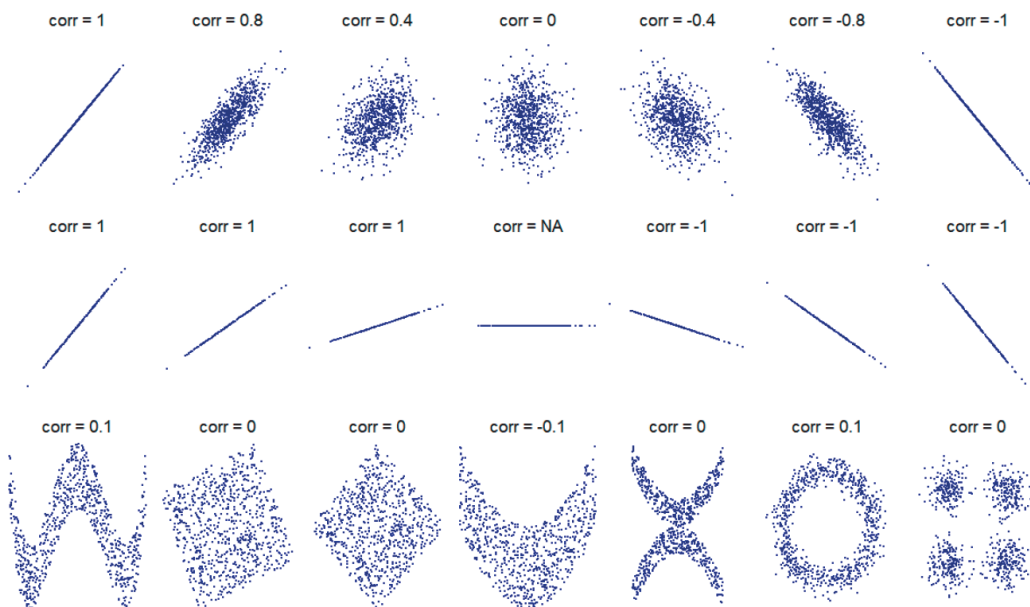


Powiązanie słabe

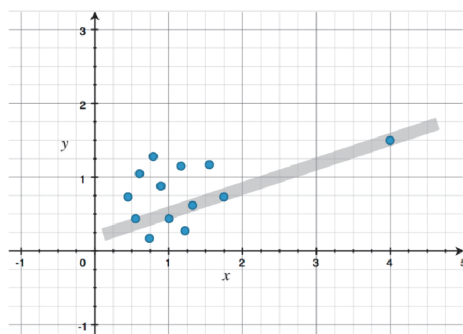
Rysunek 11.4. Siła powiązania

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

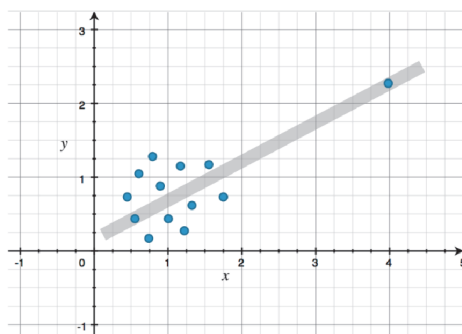
Rysunek 11.5. Wzór na współczynnik korelacji Pearsona



Rysunek 11.6. Korelacja Pearsona obliczona na podstawie rozkładów Boigelota

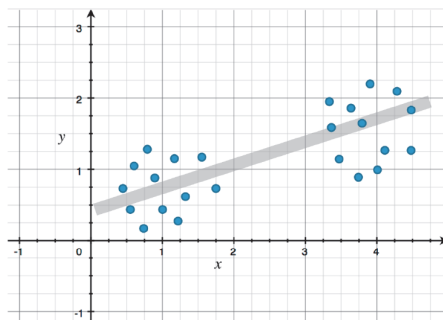


Nieprawidłowa korelacja



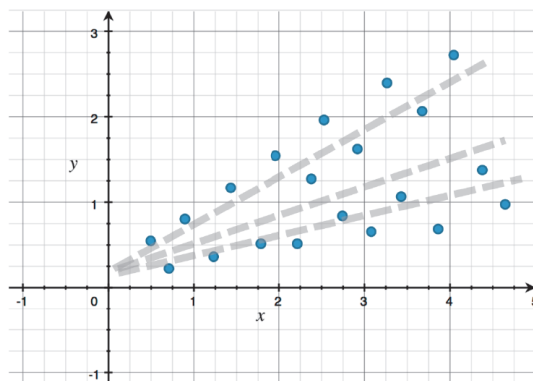
Nieprawidłowa korelacja

Rysunek 11.7. Nieprawidłowe wartości korelacji z powodu wartości odstających



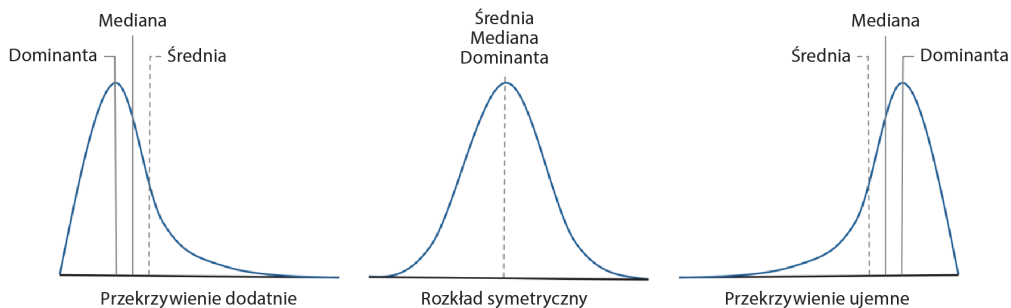
Nieprawidłowa wartość korelacji spowodowana klastrami

Rysunek 11.8. Nieprawidłowa wartość korelacji spowodowana klastrami



Nieprawidłowa wartość korelacji
z powodu heteroskedastyczności

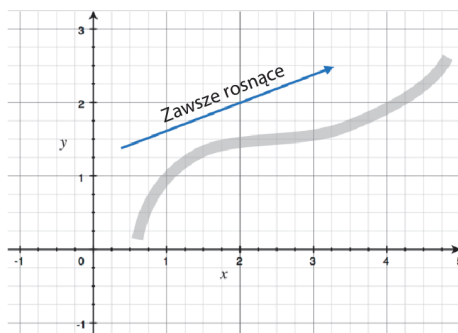
Rysunek 11.9. Nieprawidłowa wartość korelacji z powodu heteroskedastyczności



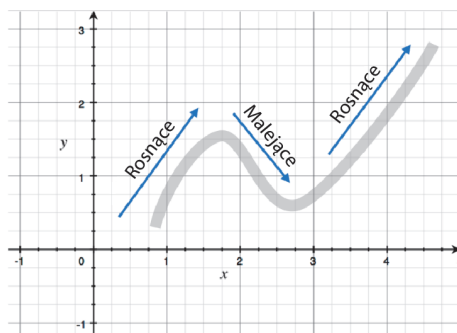
Rysunek 11.10. Typy przekrzywienia rozkładu

$$\rho = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

Rysunek 11.11. Wzór na współczynnik korelacji rang Spearmana

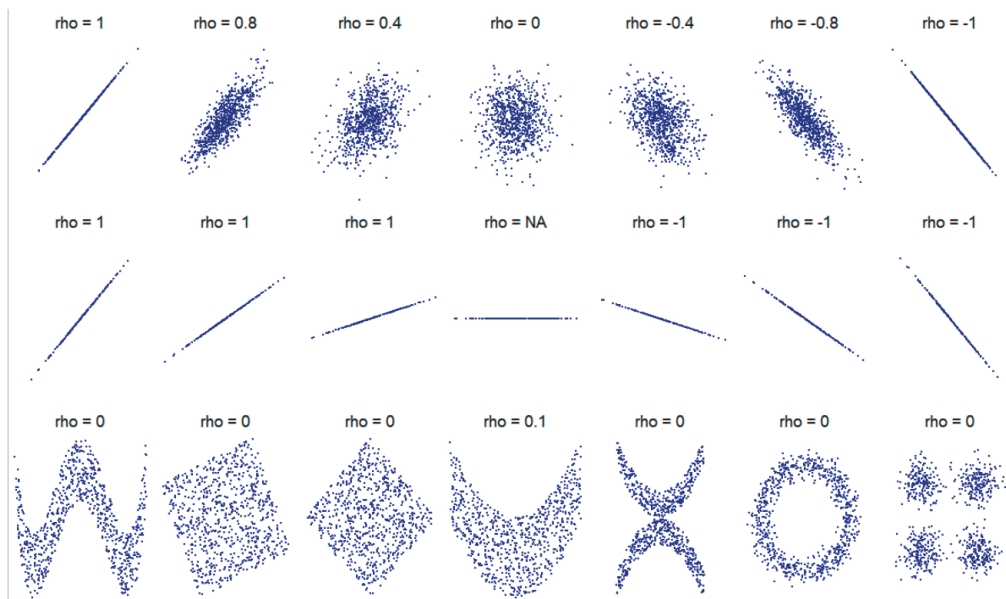


Powiązanie monotoniczne

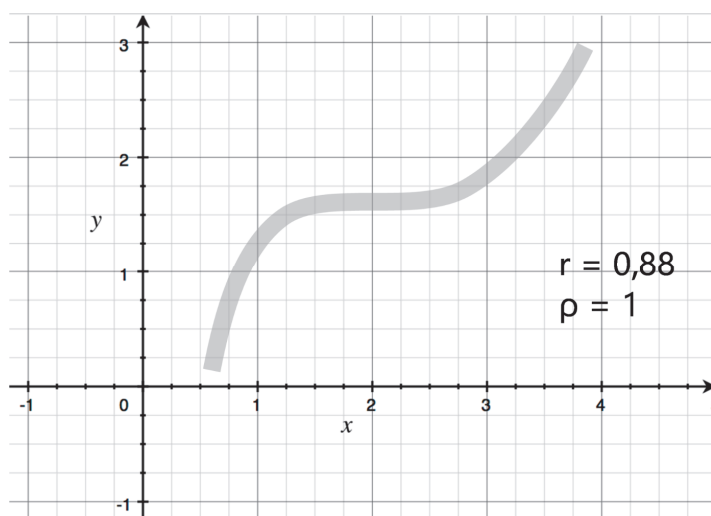


Powiązanie niemonotoniczne

Rysunek 11.12. Powiązania monotoniczne i niemonotoniczne



Rysunek 11.13. Korelacja rang Spearmana obliczona na podstawie rozkładów Boigelota



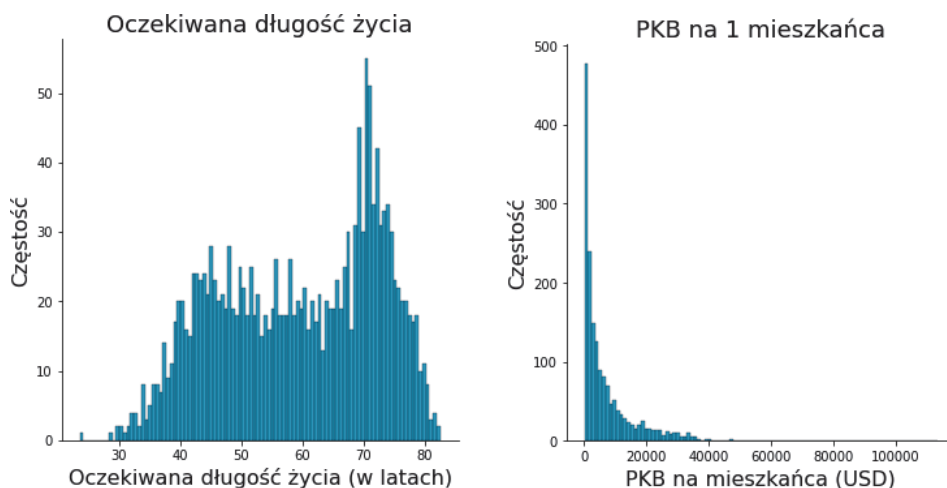
Rysunek 11.14. Korelacja Pearsona i Spearmana dla nieliniowego powiązania monotonicznego

$$\tau = \frac{2(n_c - n_d)}{n(n - 1)}$$

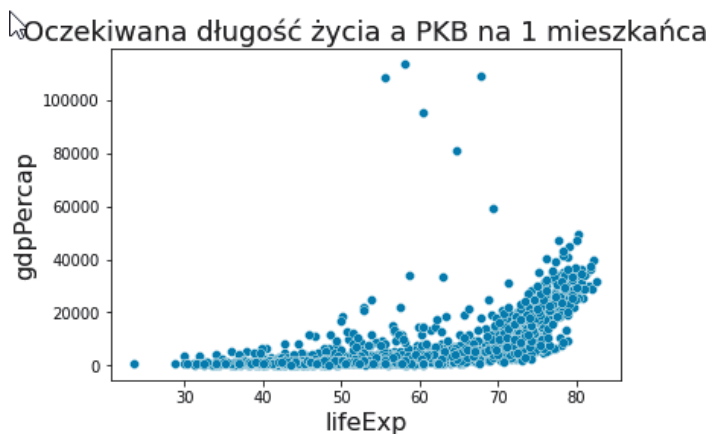
Rysunek 11.15. Wzór na obliczanie współczynnika korelacji Kendalla

	country	year	pop	continent	lifeExp	gdpPercap
0	Afghanistan	1952	8425333.0	Asia	28.801	779.445314
1	Afghanistan	1957	9240934.0	Asia	30.332	820.853030
2	Afghanistan	1962	10267083.0	Asia	31.997	853.100710
3	Afghanistan	1967	11537966.0	Asia	34.020	836.197138
4	Afghanistan	1972	13079460.0	Asia	36.088	739.981106

Rysunek 11.16. Próbkę zbioru danych o PKB i oczekiwanej długości życia



Rysunek 11.17. Rozkład zmiennych średniej długości życia i PKB na 1 mieszkańca



Rysunek 11.18. Wykres punktowy zależności pomiędzy oczekiwaną długością życia a wysokością PKB na 1 mieszkańca

Pearson		
	lifeExp	gdpPercap
lifeExp	1.000000	0.583706
gdpPercap	0.583706	1.000000

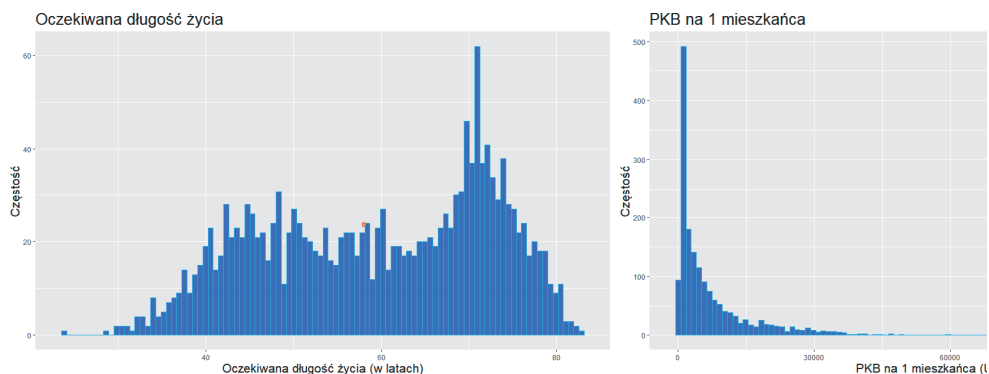
Spearman		
	lifeExp	gdpPercap
lifeExp	1.000000	0.826471
gdpPercap	0.826471	1.000000

Kendall		
	lifeExp	gdpPercap
lifeExp	1.000000	0.636911
gdpPercap	0.636911	1.000000

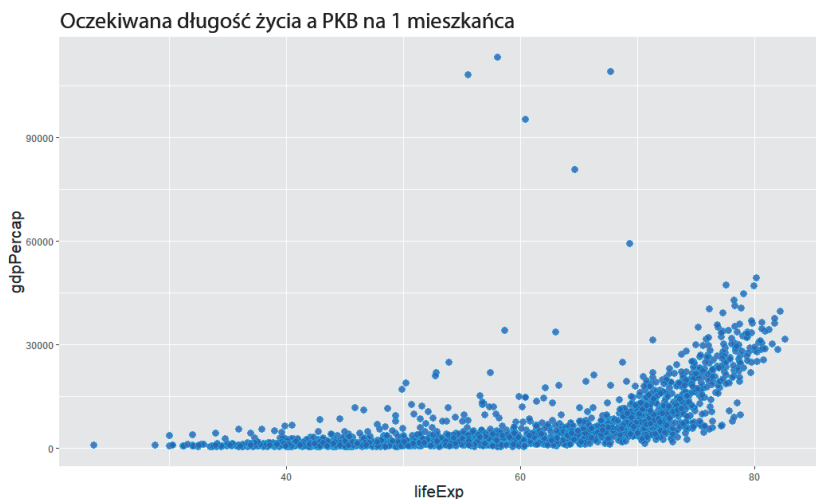
Rysunek 11.19. Korelacje między oczekiwaną długością życia a wysokością PKB na jednego mieszkańca

```
# A tibble: 1,704 x 6
  country    year    pop continent lifeExp gdpPercap
  <chr>      <dbl> <dbl> <chr>      <dbl> <dbl>
1 Afghanistan 1952  8425333 Asia      28.8   779.
2 Afghanistan 1957  9240934 Asia      30.3   821.
3 Afghanistan 1962 10267083 Asia      32.0   853.
4 Afghanistan 1967 11537966 Asia      34.0   836.
5 Afghanistan 1972 13079460 Asia      36.1   740.
6 Afghanistan 1977 14880372 Asia      38.4   786.
7 Afghanistan 1982 12881816 Asia      39.9   978.
8 Afghanistan 1987 13867957 Asia      40.8   852.
9 Afghanistan 1992 16317921 Asia      41.7   649.
10 Afghanistan 1997 22227415 Asia      41.8   635.
# ... with 1,694 more rows
> |
```

Rysunek 11.20. Pierwsze wiersze obiektu tibble zawierające dane populacji



Rysunek 11.21. Wykresy rozkładu oczekiwanej długości życia i PKB na 1 mieszkańca



Rysunek 11.22. Wykres punktowy zależności między oczekiwaną długością życia a wysokością PKB na 1 mieszkańca

Correlation method: 'pearson'
Missing treated using: 'pairwise.
complete.obs'

```
# A tibble: 2 x 3
  rowname lifeExp gdpPercap
<chr>    <dbl>    <dbl>
1 lifeExp NA      0.584
2 gdpPercap 0.584 NA
```

Correlation method: 'spearman'
Missing treated using: 'pairwise.
complete.obs'

```
# A tibble: 2 x 3
  rowname lifeExp gdpPercap
<chr>    <dbl>    <dbl>
1 lifeExp NA      0.826
2 gdpPercap 0.826 NA
```

Correlation method: 'kendall'
Missing treated using: 'pairwise.
complete.obs'

```
# A tibble: 2 x 3
  rowname lifeExp gdpPercap
<chr>    <dbl>    <dbl>
1 lifeExp NA      0.637
2 gdpPercap 0.637 NA
```

Rysunek 11.23. Współczynniki korelacji w obiektach tibble

Pobierz dane

internet X

Wszystkie

Inne

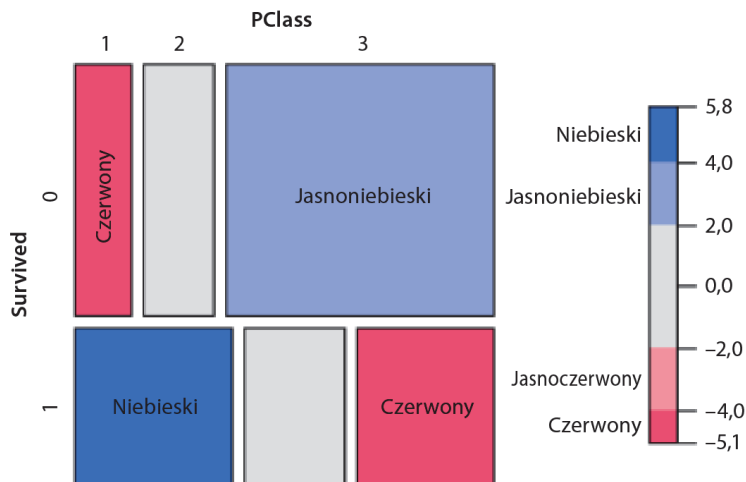
Wszystkie

Internet

Importuj dane ze strony sieci Web.

Rysunek 11.24. Pobieranie danych z internetu

Przeżycie a klasa pasażera



Rysunek 11.25. Wykres mozaikowy dla zmiennych Survived i PClass

	Pclass			Razem
	1	2	3	
0	80 133 14,6% 37%	97 113 17,7% 52,7%	372 303 67,8% 75,8%	549 549 100% 61,6%
1	136 83 39,8% 63%	87 71 25,4% 47,3%	119 188 34,8% 24,2%	342 342 100% 38,4%
Razem	216 216 24,2% 100%	184 184 20,7% 100%	491 491 55,1% 100%	891 891 100% 100%

$\chi^2=102,889$ $df=2$ Współczynnik V Cramera=0,340 $p=0,000$

Liczba zaobserwowana

Liczba oczekiwana

% wśród ocalałych/nieocalałych

% wśród pasażerów klasy

Rysunek 11.26. Tabela kontyngencji dla zmiennych Survived i PClass

$$V = \phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

Rysunek 11.27. Wzór na współczynnik V Craméra

Miara wielkości efektu	Zakres współczynników V Cramér
Mały	[0 - 0,3)
Średni	[0,3 - 0,5)
Duży	[0,5 - 1)

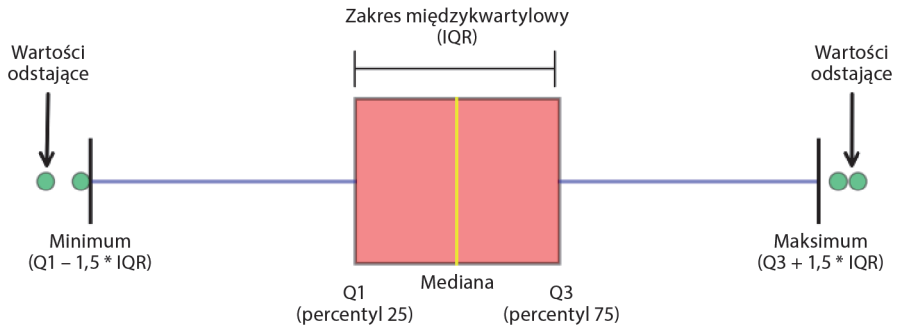
Rysunek 11.28. Zakresy miar efektów współczynnika V Craméra

JestOszustem	Hobby
Nie jest oszustem	Koszykówka
Nie jest oszustem	Pianino
Nie jest oszustem	Koszykówka
Oszust	Szachy
Oszust	Szachy
Oszust	Siłownia

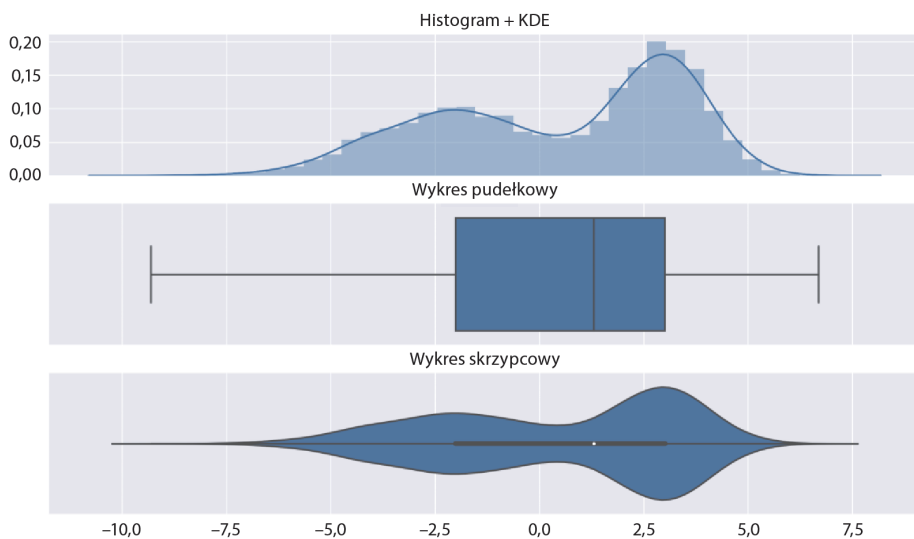
Rysunek 11.29. Przykładowy zestaw danych zmiennych kategoriycznych

$$U(X|Y) = \frac{H(X) - H(X|Y)}{H(X)}$$

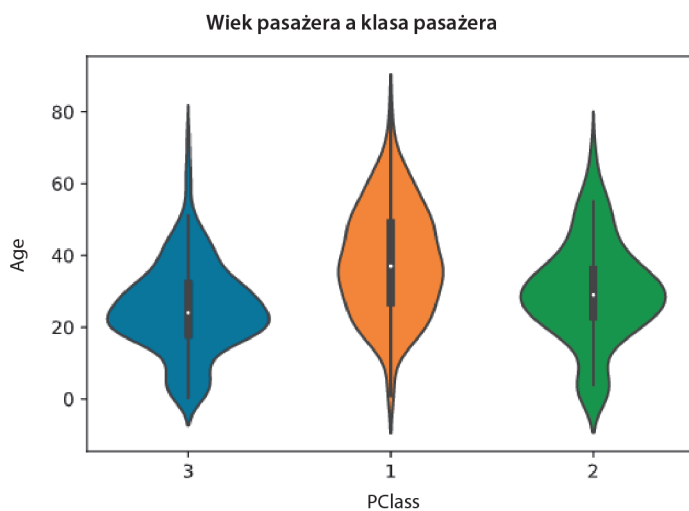
Rysunek 11.30. Wzór na współczynnik niepewności Theila



Rysunek 11.31. Graficzne objaśnienie wykresu pudełkowego



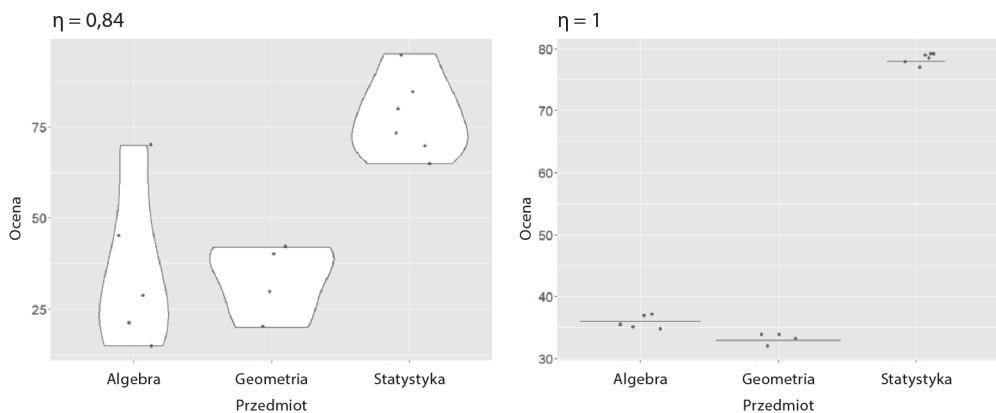
Rysunek 11.32. Graficzne objaśnienie wykresu skrzypcowego



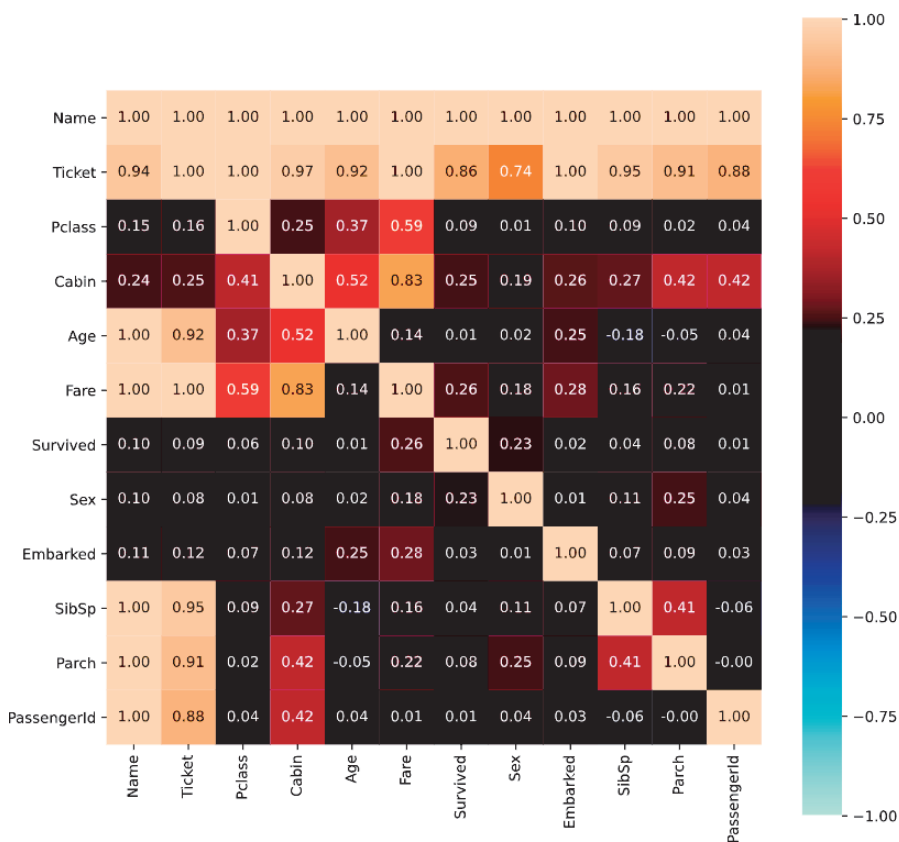
Rysunek 11.33. Graficzne objaśnienie wykresu skrzypcowego

$$\eta^2 = \frac{\sum_x n_x (\bar{y}_x - \bar{y})^2}{\sum_{x,i} (y_{xi} - \bar{y})^2} = \frac{\sigma_{\bar{y}}^2}{\sigma_y^2}$$

Rysunek 11.34. Wzór na współczynnik korelacji



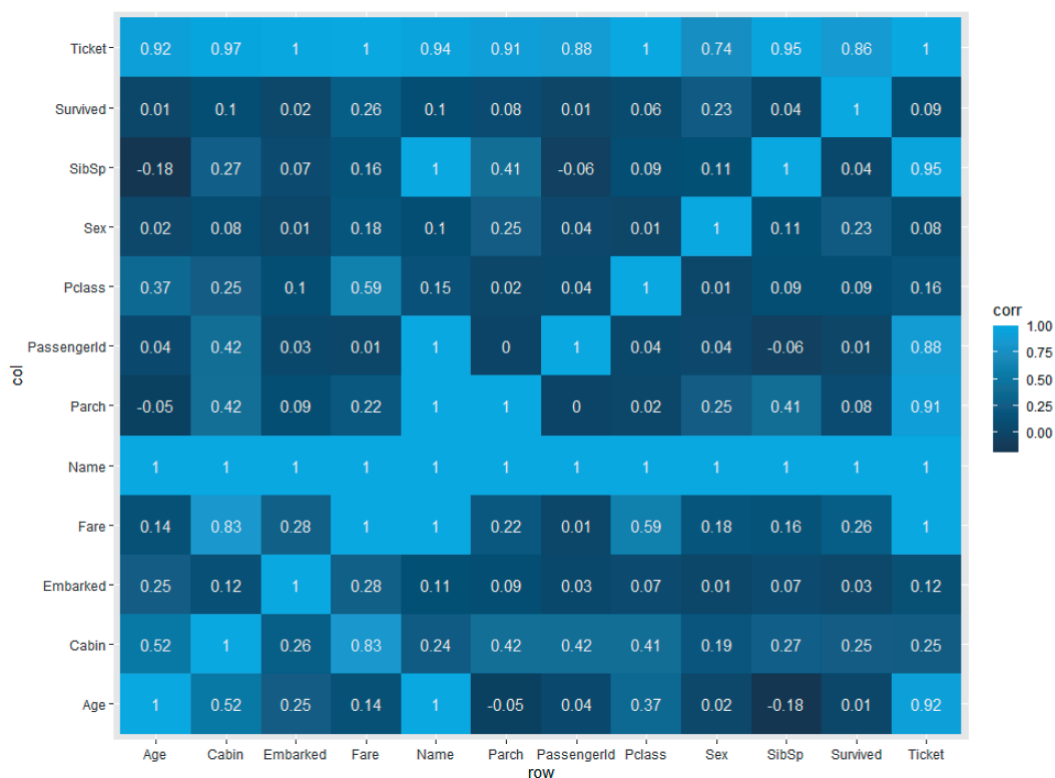
Rysunek 11.35. Różnice w η po zmianie rozkładu ocen z przedmiotów



Rysunek 11.36. Mapa termiczna korelacji

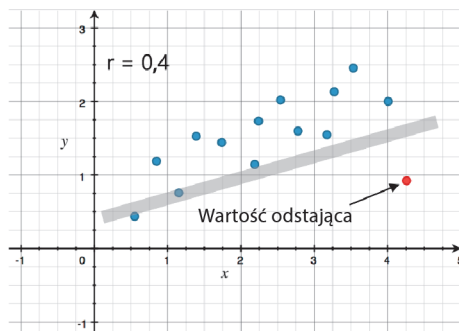

```
# A tibble: 144 x 3
  row   col      corr
  <chr> <chr>   <dbl>
1 Age   Age      1
2 Age   Cabin    0.524
3 Age   Embarked  0.249
4 Age   Fare     0.136
5 Age   Name      1
6 Age   Parch    -0.0488
7 Age   PassengerId 0.0381
8 Age   Pclass    0.366
9 Age   Sex      0.0250
10 Age  SibSp    -0.185
# ... with 134 more rows
```

Rysunek 11.37. Obiekt tibble korelacji dla danych dotyczących katastrofy Titanica

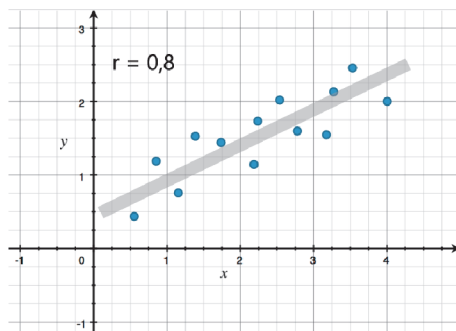


Rysunek 11.38. Mapa termiczna obiektu tibble korelacji

Rozdział 12. Dodawanie statystyk: wartości odstające i wartości brakujące

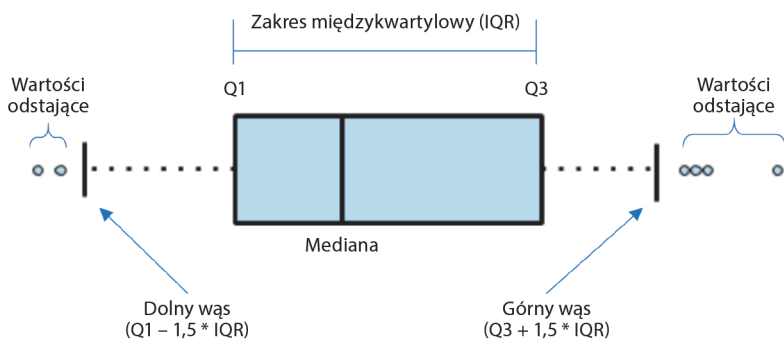


Korelacja W OBECNOŚCI wartości odstającej

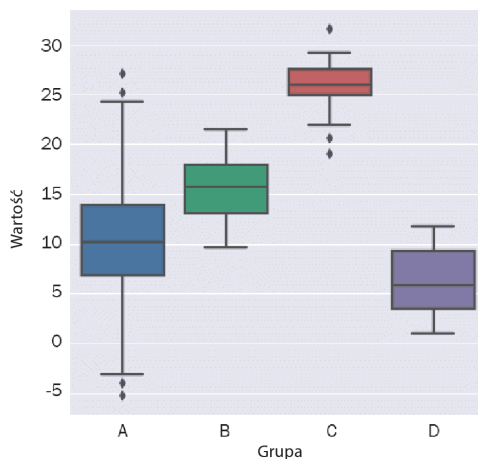


Korelacja BEZ OBECNOŚCI wartości odstającej

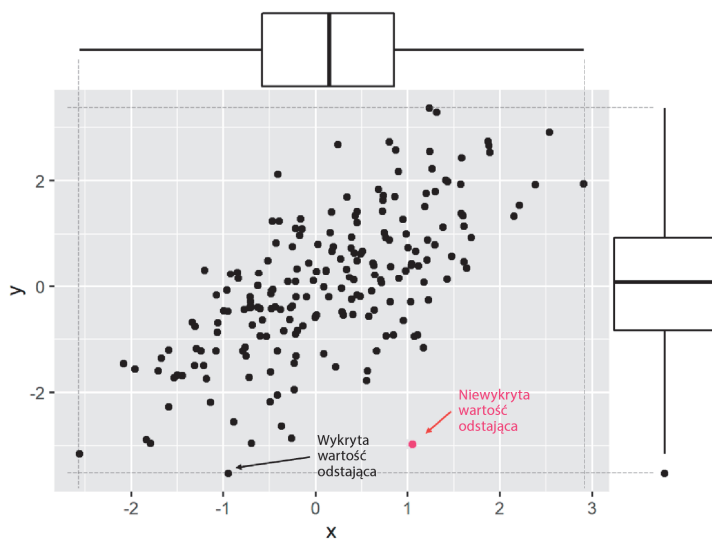
Rysunek 12.1. Wykresy korelacji liniowej w przypadku obecności w zestawie danych wartości odstającej i bez niej



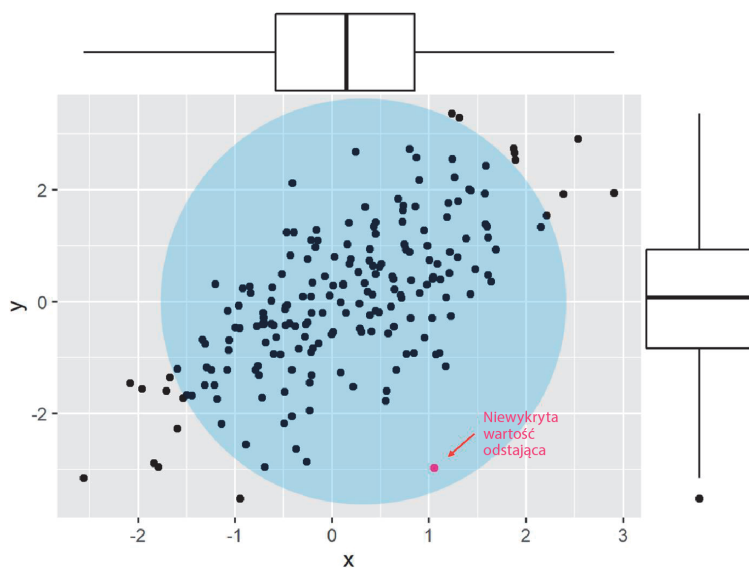
Rysunek 12.2. Główne cechy wykresu pudełkowego



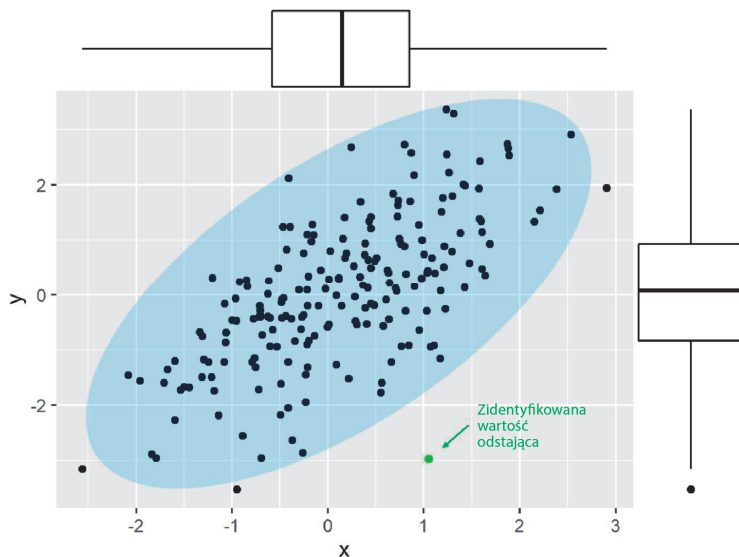
Rysunek 12.3. Zmienne liczbowe a zmienne kateryczne



Rysunek 12.4. Wykres punktowy dla dwóch zmiennych liczbowych



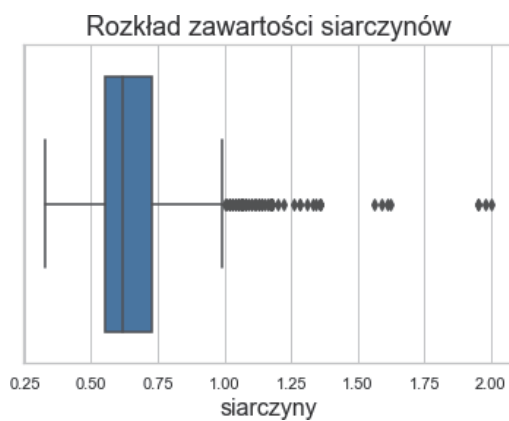
Rysunek 12.5. Odległość euklidesowa od środka rozkładu



Rysunek 12.6. Odległość Mahalanobisa od środka rozkładu

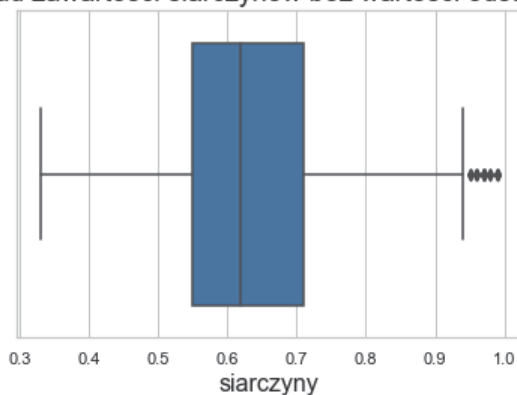
$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T \mathbf{S}^{-1} (\vec{x} - \vec{\mu})}$$

Rysunek 12.7. Wzór na odległość Mahalanobisa



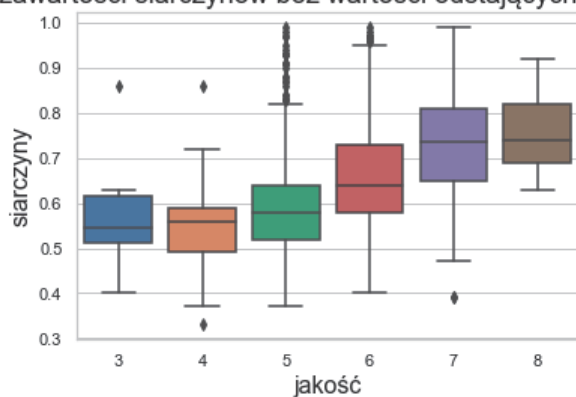
Rysunek 12.8. Wykres pudełkowy dla zmiennej sulphates

Rozkład zawartości siarczynów bez wartości odstających

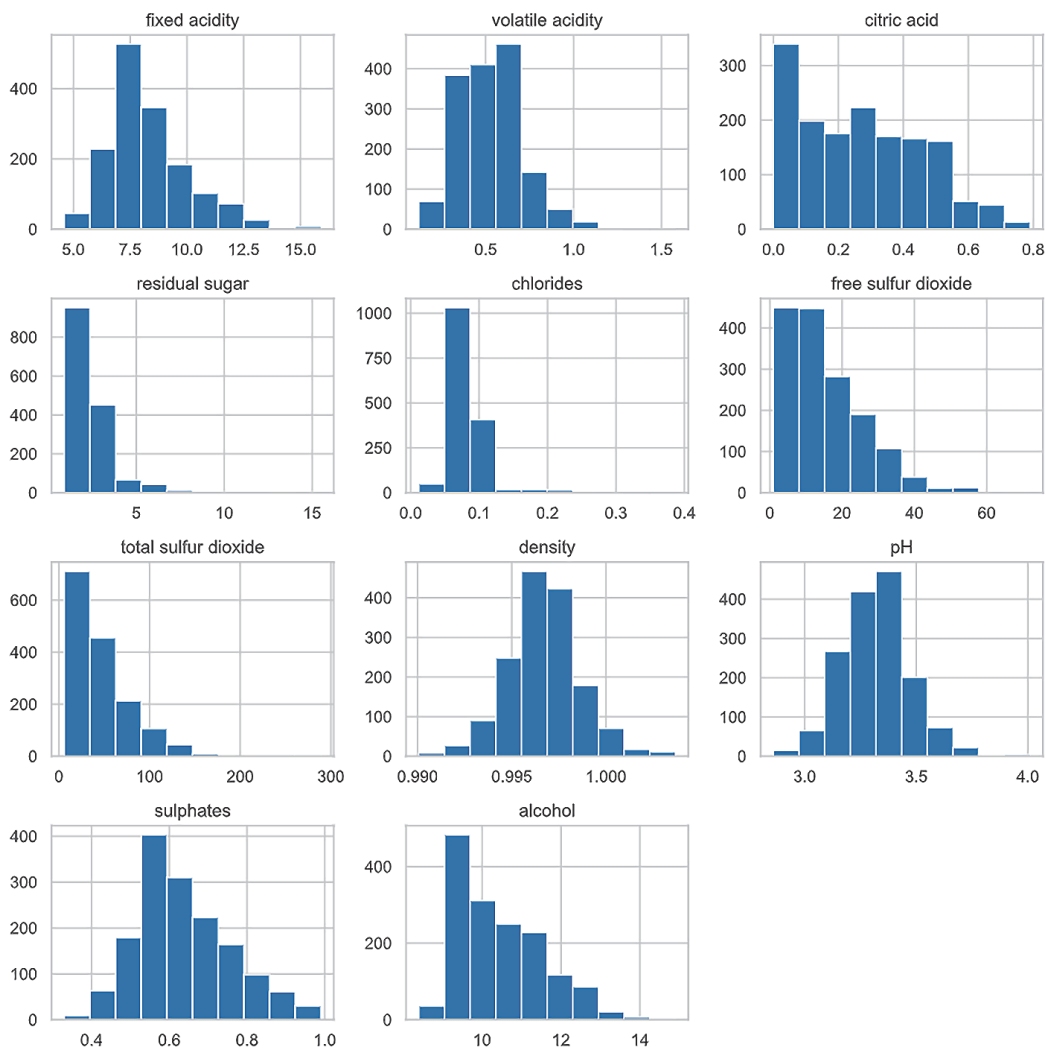


Rysunek 12.9. Wykres pudełkowy zawartości siarczynów po usunięciu wartości odstających

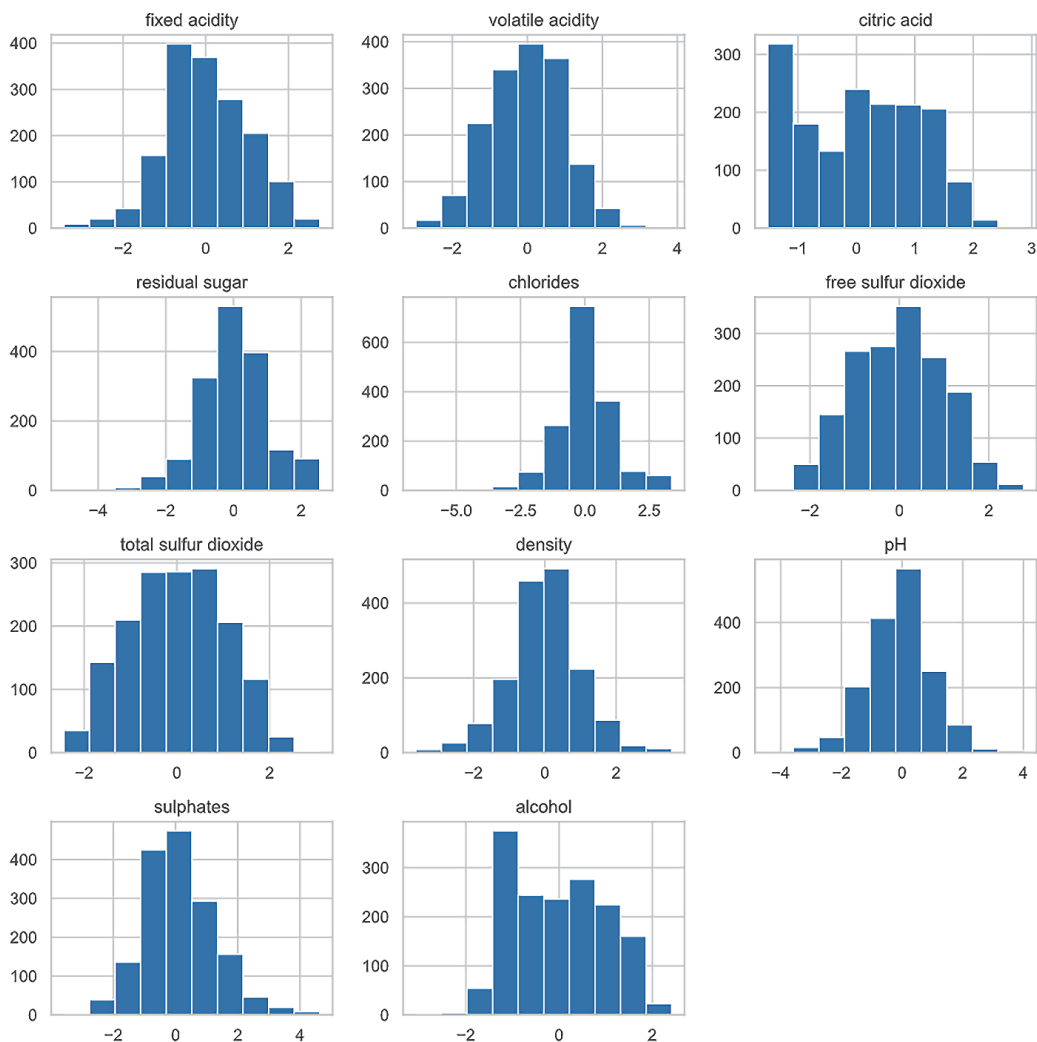
Rozkład zawartości siarczynów bez wartości odstających według jakości



Rysunek 12.10. Wykresy pudełkowe zawartości siarczynów dla każdej wartości jakości



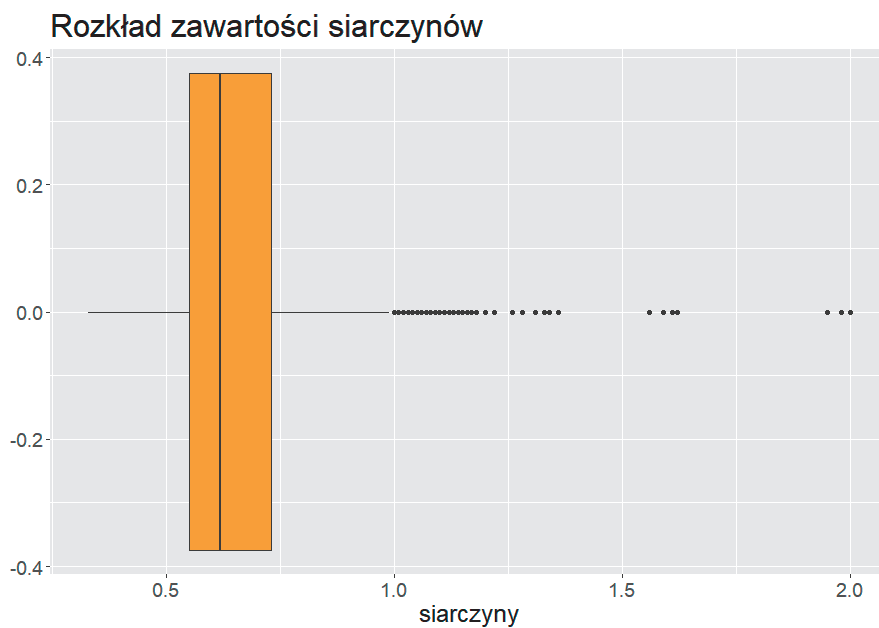
Rysunek 12.11. Histogramy wszystkich zmiennych jakościowych wina bez wartości odstających



Rysunek 12.12. Histogramy wszystkich przekształconych zmiennych jakości wina

	fixed acidity		is_mahalanobis_outlier	mahalanobis_outlier_proba
13	7.8		True	0.999987
14	8.9		True	0.996728
15	8.9	:	True	0.997628
17	8.1	2	True	1.000000
19	7.9	.2	True	1.000000
...
1558	6.9		True	0.999994
1570	6.4	se	True	0.999999
1574	5.6		True	1.000000
1589	6.6	-	True	0.985707
1598	6.0		True	0.989600

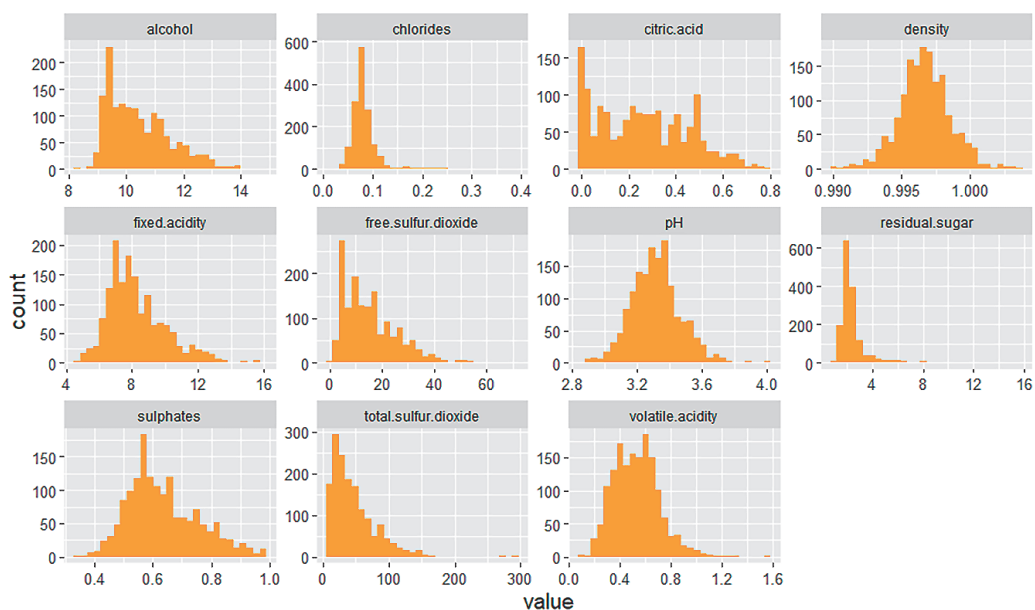
Rysunek 12.13. Informacje o wartościach odstających pokazane w ramce danych



Rysunek 12.14. Wykres pudełkowy zmiennej sulphates



Rysunek 12.15. Wykres pudełkowy zmiennej sulphates po usunięciu wartości odstających

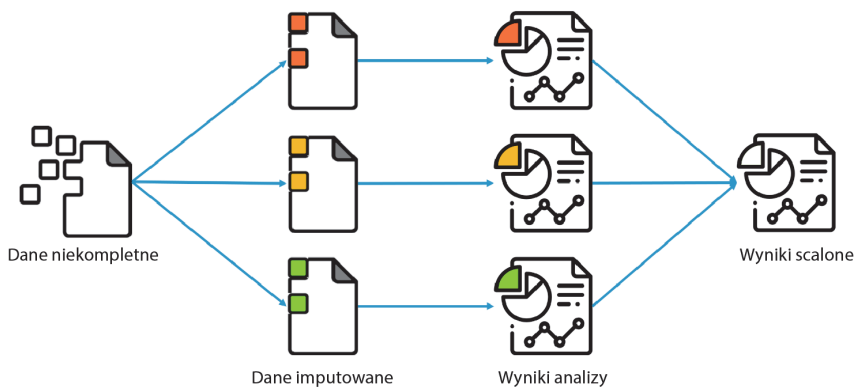


Rysunek 12.16. Histogramy dla każdej zmiennej liczbowej

```
# A tibble: 182 x
```

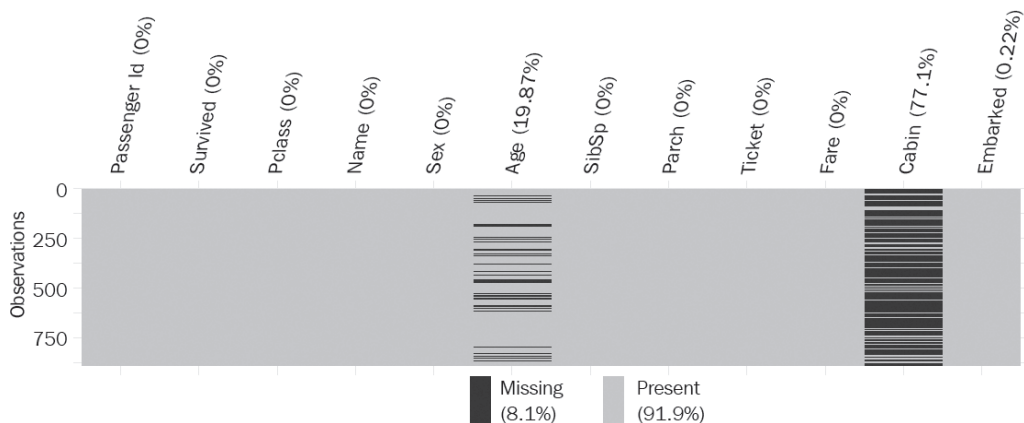
	fixed.acidity	is_mahalanobis_outlier	mahalanobis_outlier_proba
	<dbl>	<lgl>	<dbl>
1	0.313	TRUE	0.997
2	0.313	TRUE	0.998
3	-0.970	TRUE	0.982
4	-2.54	TRUE	1.00
5	-2.04	TRUE	1.00
6	-0.515	TRUE	1.00
7	-3.18	TRUE	1.00
8	-0.376	TRUE	0.998
9	-2.14	TRUE	1.00
10	-0.376	TRUE	0.982

Rysunek 12.17. Wynikowy obiekt tibble zawierający informacje o wielowymiarowych wartościach odstających



Rysunek 12.18. Proces imputacji wielokrotnej

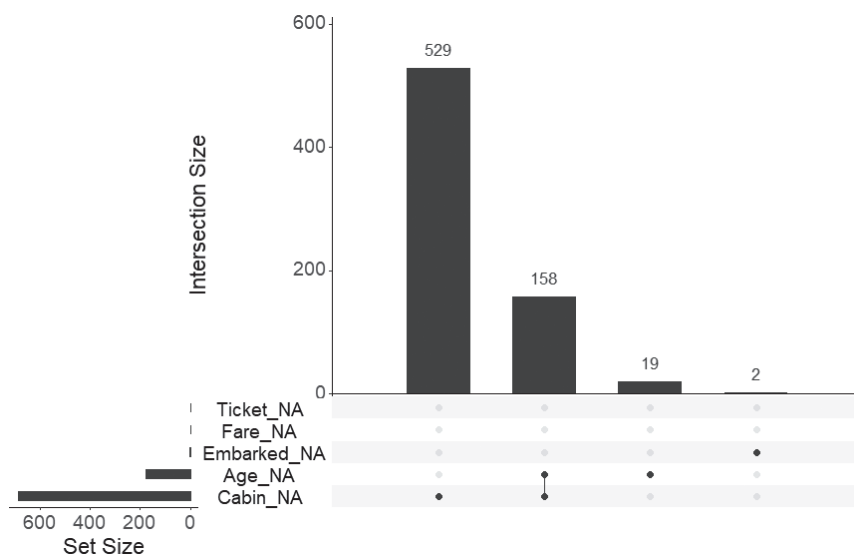
Oto najczęstsze implementacje imputacji wielokrotnej:



Rysunek 12.19. Wykres brakujących wartości w całym zestawie danych

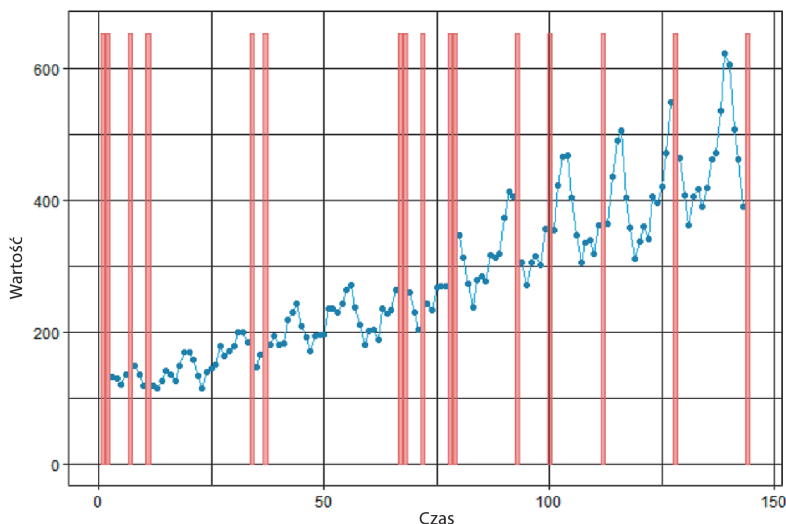
```
> miss_var_summary(tbl)
# A tibble: 12 x 3
  variable    n_miss pct_miss
  <chr>      <int>   <dbl>
1 Cabin         687    77.1
2 Age          177    19.9
3 Embarked        2     0.224
4 PassengerId     0     0
5 Survived        0     0
6 Pclass         0     0
7 Name           0     0
8 Sex            0     0
9 SibSp          0     0
10 Parch         0     0
11 Ticket         0     0
12 Fare          0     0
```

Rysunek 12.20. Podsumowanie brakujących wartości



Rysunek 12.21. Wykres typu UpSet brakujących wartości w zestawie danych

Rozkład brakujących wartości
Szeregi czasowe z wyróżnionymi brakującymi obszarami



Rysunek 12.22. Wykrywanie brakujących wartości w szeregach czasowych

```
> # Usuwanie typu listwise
> cor( tbl_num, method = 'pearson', use = 'complete.obs' )
```

	PassengerId	Survived	Pclass	Age	Sibsp	Parch	Fare
PassengerId	1.00000000	0.02934016	-0.03534911	0.03684720	-0.08239772	-0.01161741	0.00959178
Survived	0.02934016	1.00000000	-0.35965268	-0.07722109	-0.01735836	0.09331701	0.26818862
Pclass	-0.03534911	-0.35965268	1.00000000	-0.36922602	0.06724737	0.02568307	-0.55418247
Age	0.03684720	-0.07722109	-0.36922602	1.00000000	-0.30824676	-0.18911926	0.09606669
Sibsp	-0.08239772	-0.01735836	0.06724737	-0.30824676	1.00000000	0.38381986	0.13832879
Parch	-0.01161741	0.09331701	0.02568307	-0.18911926	0.38381986	1.00000000	0.20511888
Fare	0.00959178	0.26818862	-0.55418247	0.09606669	0.13832879	0.20511888	1.00000000

```
> # Usuwanie typu pairwise
> cor( tbl_num, method = 'pearson', use = 'pairwise.complete.obs'
+ )
```

	PassengerId	Survived	Pclass	Age	Sibsp	Parch	Fare
PassengerId	1.00000000	-0.00500666	-0.03514399	0.03684720	-0.05752683	-0.00165201	0.01265822
Survived	-0.00500666	1.00000000	-0.33848104	-0.07722109	-0.03532250	0.08162940	0.25730652
Pclass	-0.03514399	-0.33848104	1.00000000	-0.36922602	0.08308136	0.01844267	-0.54949962
Age	0.03684719	-0.07722109	-0.36922602	1.00000000	-0.30824676	-0.18911926	0.09606669
Sibsp	-0.05752683	-0.03532249	0.08308136	-0.30824676	1.00000000	0.41483770	0.15965104
Parch	-0.00165201	0.08162940	0.01844267	-0.18911926	0.41483770	1.00000000	0.21622494
Fare	0.01265821	0.25730652	-0.54949962	0.09606669	0.15965104	0.21622494	1.00000000

Rysunek 12.24. Macierz korelacji obliczona przy użyciu usuwania typu listwise i typu pairwise

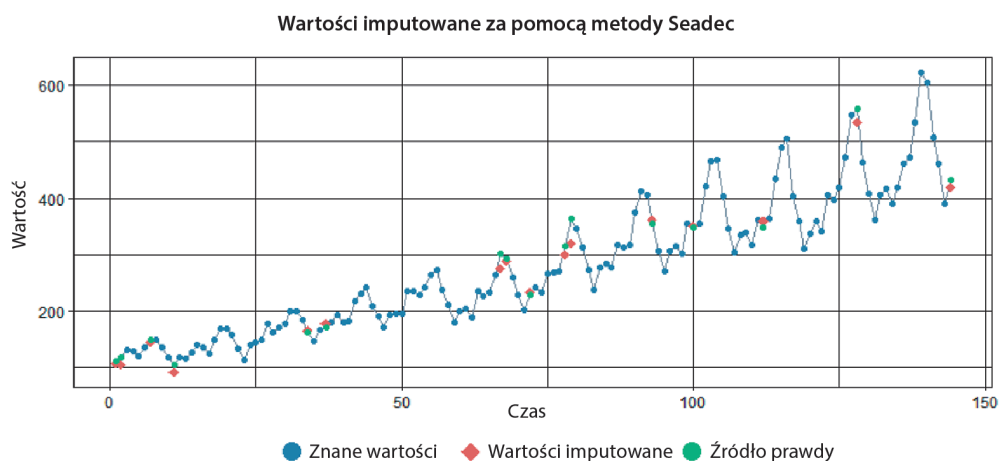
```
# A tibble: 30 x 11
  variable1 variable2      r      rse fisher_r fisher_rse fmi      t      p lower95 upper95
  <chr>      <chr>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Age      Fare      0.0947 0.0342 0.0950 0.0345 0.0561 2.75 5.91e- 3 0.0273 0.161
2 Age      Parch    -0.193 0.0339 -0.195 0.0352 0.0952 -5.54 3.07e- 8 -0.258 -0.125
3 Age      PassengerId 0.0397 0.0419 0.0398 0.0420 0.398 0.947 3.43e- 1 -0.0425 0.121
4 Age      Pclass    -0.381 0.0316 -0.401 0.0370 0.189 -10.8 2.01e-27 -0.441 -0.317
5 Age      Sibsp     -0.311 0.0314 -0.322 0.0348 0.0704 -9.26 2.06e-20 -0.371 -0.248
6 Fare      Age      0.0947 0.0342 0.0950 0.0345 0.0561 2.75 5.91e- 3 0.0273 0.161
7 Fare      Parch    0.216 0.0320 0.220 0.0336 0      6.55 5.88e-11 0.153 0.278
8 Fare      PassengerId 0.0127 0.0336 0.0127 0.0336 0      0.377 7.06e- 1 -0.0531 0.0783
9 Fare      Pclass    -0.549 0.0234 -0.618 0.0336 0     -18.4 1.18e-75 -0.594 -0.502
10 Fare     Sibsp     0.160 0.0327 0.161 0.0336 0      4.80 1.60e- 6 0.0950 0.223
```

... with 20 more rows

Rysunek 12.25. Wnioskowanie statystyczne dla korelacji w przypadku wielu imputowanych zbiorów danych

```
# A tibble: 10 x 6
  strategy      ME      RMSE      MAE      MPE      MAPE
  <chr>      <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 locf       1.11    11.8     3.10    0.133    1.22
2 nocb       1.20    11.6     2.92    0.0728   1.05
3 ewma_1     1.17     8.66     2.20    0.135    0.849
4 ewma_2     1.27     9.87     2.48    0.134    0.944
5 ewma_3     1.30    11.3     2.75    0.129    1.03
6 ewma_6     1.33    12.5     3.10    0.131    1.14
7 ewma_9     1.34    12.6     3.13    0.130    1.15
8 linear     1.16     8.58     2.21    0.107    0.845
9 spline     0.261    5.71     1.58   -0.0913   0.714
10 seadec     0.900    5.47     1.36    0.365    0.547
```

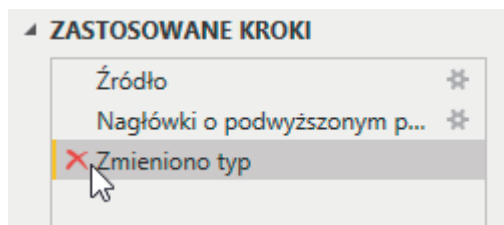
Rysunek 12.26. Metryki błędów dla wartości przypisanych w szeregach czasowych



Rysunek 12.27. Reprezentacja wartości imputowanych w szeregach czasowych

	AB_C variable1	AB_C variable2	1.2 r	1.2 rse
1	Age	Fare	0.093679086	0.034107542
2	Age	Parch	-0.197494964	0.033906333
3	Age	PassengerId	0.03778199	0.03500428
4	Age	Pclass	-0.376451518	0.02990363
5	Age	SibSp	-0.308517666	0.03137899
6	Fare	Age	0.093679086	0.034107542

Rysunek 12.28. Tabela korelacji obliczona techniką imputacji wielowymiarowej

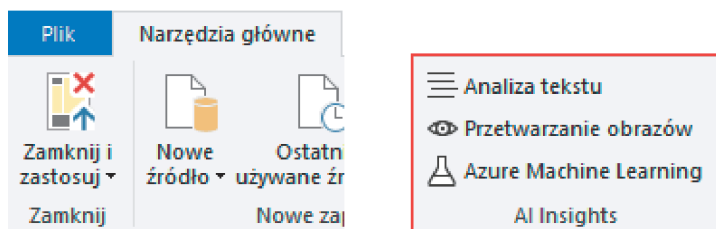


Rysunek 12.29. Usuwanie kroku „Zmieniono typ”

	date	1.3 value	1.2 locf	1.2 nocb	1.2 ewma_1
1	1949-01-01	null	132	132	131
2	1949-02-01	null	132	132	131
3	1949-03-01	132	132	132	132
4	1949-04-01	129	129	129	129
5	1949-05-01	121	121	121	121
6	1949-06-01	135	135	135	135

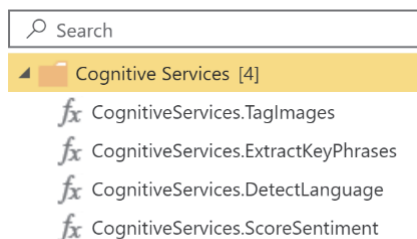
Rysunek 12.30. Tabela z przypisanymi szeregami czasowymi

Rozdział 13. Korzystanie z uczenia maszynowego bez licencji Premium lub Embedded



Rysunek 13.1. Zakładka AI Insights w programie Power BI Desktop

AI insights



Rysunek 13.2. Funkcje usług Cognitive Services w Power BI

Azure Machine Learning Models

Azure Machine Learning Models...

fx

AzureML.my-diabetes-model

AGE

AGE

SEX

SEX

BMI

BMI

BP

BP

Rysunek 13.3. Funkcje usługi Azure Machine Learning w Power BI

Choose a model type

Classification

✓/✗

Binary Prediction

Determine the likelihood of a specific outcome being achieved.

General Classification

Identify the category or class an entity belongs to.

Regression

Regression

Estimate a numeric value

Forecasting

Forecasting

Estimate values and trends based on historical data.

Rysunek 13.4. Funkcje AutoML dla przepływów danych w usłudze Power BI

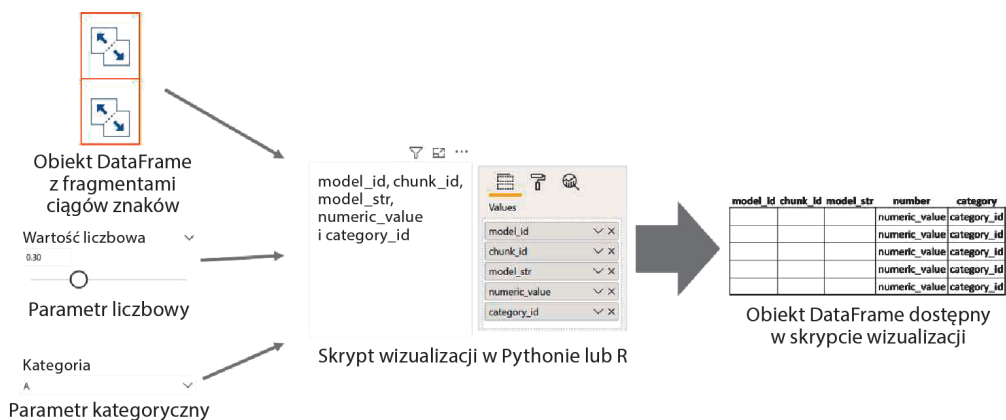
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8136	0.8519	0.7158	0.8075	0.7543	0.6051	0.6122	0.0080
gbc	Gradient Boosting Classifier	0.8034	0.8454	0.6864	0.8013	0.7340	0.5802	0.5892	0.0150
rf	Random Forest Classifier	0.7881	0.8538	0.7371	0.7399	0.7348	0.5589	0.5628	0.1070
lr	Logistic Regression	0.7814	0.8339	0.6739	0.7575	0.7077	0.5352	0.5422	0.0150
et	Extra Trees Classifier	0.7780	0.8360	0.7114	0.7305	0.7181	0.5353	0.5380	0.1000
ada	Ada Boost Classifier	0.7763	0.8175	0.6859	0.7417	0.7079	0.5277	0.5328	0.0160
lda	Linear Discriminant Analysis	0.7763	0.8329	0.6699	0.7481	0.7023	0.5250	0.5308	0.0030
ridge	Ridge Classifier	0.7746	0.0000	0.6658	0.7471	0.6998	0.5213	0.5269	0.0040
nb	Naive Bayes	0.7576	0.7977	0.6866	0.7050	0.6921	0.4929	0.4964	0.0040
dt	Decision Tree Classifier	0.7424	0.7297	0.6696	0.6889	0.6751	0.4623	0.4661	0.0040
qda	Quadratic Discriminant Analysis	0.7237	0.7409	0.5953	0.7102	0.6067	0.4081	0.4269	0.0040
knn	K Neighbors Classifier	0.6898	0.7214	0.5460	0.6320	0.5848	0.3393	0.3425	0.0260
svm	SVM - Linear Kernel	0.6898	0.0000	0.7136	0.6083	0.6385	0.3757	0.3954	0.0040

Rysunek 13.5. Wydajność wszystkich modeli

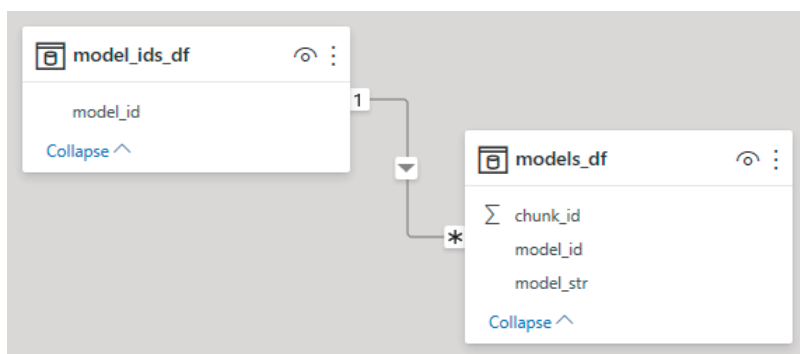
137

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Survived	Label	Score
0	2.0	1.0	32.4	0.0	0.0	13.0000	2.0	1.0	1.0	0.7764
1	3.0	1.0	28.0	0.0	0.0	9.5000	2.0	0.0	0.0	0.9100
2	3.0	0.0	27.0	0.0	0.0	7.9250	2.0	1.0	1.0	0.8400
3	1.0	0.0	58.0	0.0	1.0	153.4625	2.0	1.0	1.0	0.9900
4	3.0	1.0	28.0	1.0	0.0	15.8500	2.0	0.0	0.0	0.9900
...
839	3.0	0.0	29.0	1.0	1.0	10.4625	2.0	0.0	0.0	0.9300
840	3.0	0.0	17.2	0.0	0.0	7.2292	0.0	1.0	1.0	0.9700
841	1.0	1.0	34.0	0.0	0.0	26.5500	2.0	1.0	1.0	0.9600
842	2.0	1.0	19.0	0.0	0.0	10.5000	2.0	1.0	0.0	0.5717
843	3.0	0.0	21.0	0.0	0.0	7.7500	1.0	0.0	0.0	0.7700

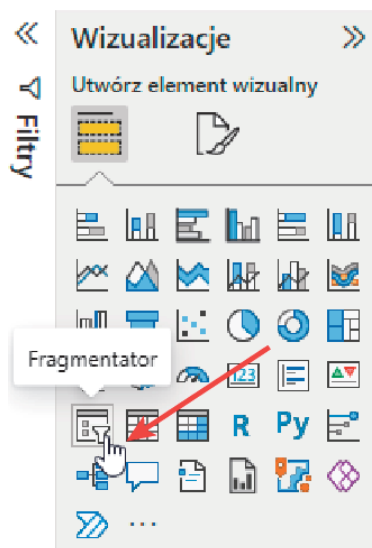
Rysunek 13.6. Prognozy dla zbioru danych testowych



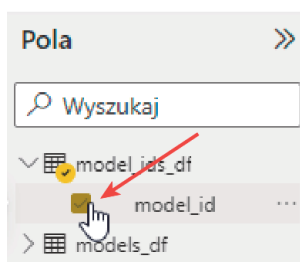
Rysunek 13.7. Deserializacja zawartości pliku PKL do wizualizacji Pythona



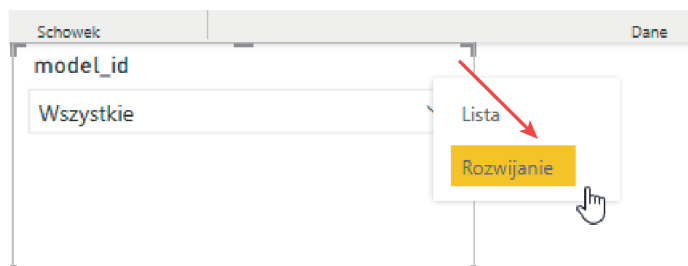
Rysunek 13.8. Utworzona automatycznie relacja między tabelami modelu



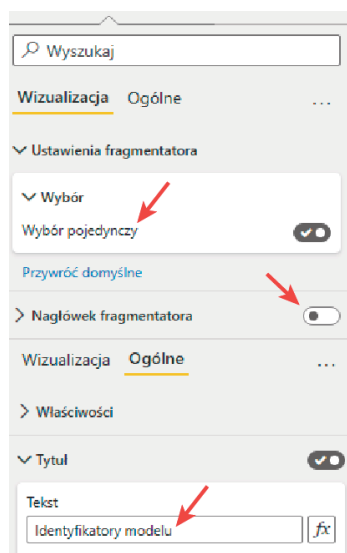
Rysunek 13.9. Zaznaczanie wizualizacji fragmentatora



Rysunek 13.10. Kliknij pole `model_id`, aby wyświetlić je we fragmentatorze



Rysunek 13.11. Wybieranie typu fragmentatora Rozwijanie



Rysunek 13.12. Ustawianie opcji fragmentatora

Parametr What-if

Nazwa

Typ danych

Minimum

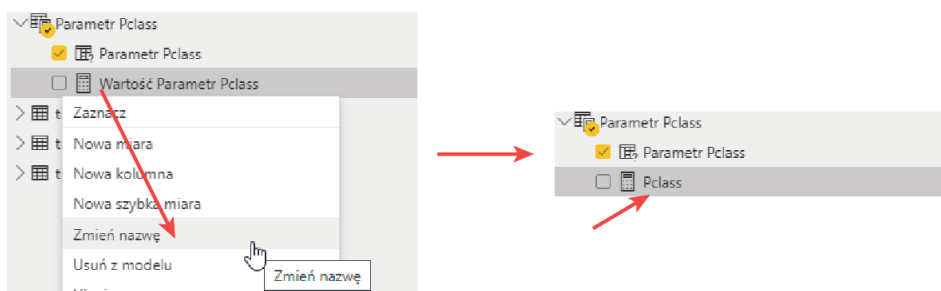
Maksimum

Zwiększ

Domyślny

☒ Dodaj fragmentator do tej strony

Rysunek 13.13. Dodawanie parametru what-if dla zmiennej Pclass



Rysunek 13.14. Zmiana nazwy wartości parametru Pclass

Tworzenie tabeli

	Sex	SexLabel	+
1	0	Kobieta	
2	1	Mężczyzna	
3			
+			

Nazwa:

Rysunek 13.15. Ręczne wprowadzanie danych do tabeli Sex

Klasa pasażera

☐

Płeć

Rysunek 13.16. Nowy rozwijany fragmentator do wyboru płci

Parametr Age

☐

☐

Rysunek 13.17. Zmiana nazwy wartości parametru zmiennej Age

Parametr SibSp

☒

☐

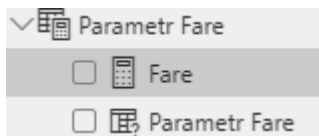
Rysunek 13.18. Zmiana nazwy wartości parametru SibSp

Parametr Parch

☒

☐

Rysunek 13.19. Zmiana nazwy wartości parametru Parch



Rysunek 13.20. Zmiana nazwy wartości parametru Fare

Tworzenie tabeli

	Embarked	EmbarkedLabel	+
1	0	Cherbourg	
2	1	Queenstown	
3	2	Southampton	
4			
+			

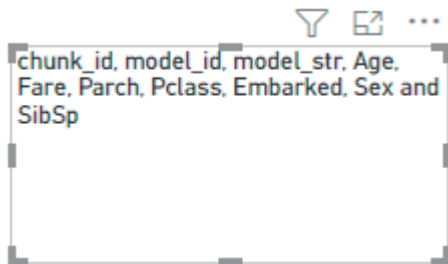
Nazwa:

Załaduj

Edytuj

Anuluj

Rysunek 13.21. Ręczne wprowadzanie danych dla Embarked



Rysunek 13.22. Wybrane nazwy miar widoczne w wizualizacji Pythona

Klasa pasażera

2

Identyfikator modelu

model01

Płeć

Kobieta

Wiek

25

Rodzeństwo(współmałżonek) na pokładzie

1

Rodzice (dzieci) na pokładzie

1

Taryfa

255

Port zaokrętowania

Queenstown

Progniza

Survived = 1(prob = 0.9422)

Rysunek 13.23. Kompletny raport symulacji prognoz dla modelu Titanica

Microsoft Azure Machine Learning

Home

Author

Notebooks

Automated ML

Designer

Assets

Datasets

Experiments

Modules

Pipelines

Models

Endpoints

Manage

Compute

Environments (preview)

Datastores

Data Labeling

Linked Services

Home

Welcome to the Azure Machine Learning Studio

Create new

Notebooks

Automated ML

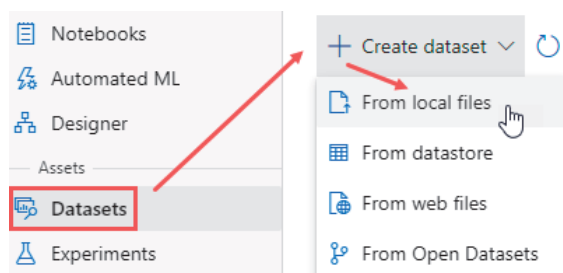
Designer

Recent resources

Runs

Run	Run ID	Experiment	Status	Submitted time	Submitted by	Run type
Run 1...	a4e676fd-5baa-43...	mlops-wo...	Completed	May 26, 2021 4:21 PM	Luca Zavarella	Pipeline
Run 1...	36d1f73b-bdcd-4b...	training-p...	Completed	May 26, 2021 9:30 AM	Service Princ...	Pipeline
Run 1...	3533a45d-e96d-40...	training-p...	Failed	May 25, 2021 3:55 PM	Service Princ...	Pipeline

Rysunek 13.24. Portal usługi Azure ML Studio



Rysunek 13.25. Tworzenie nowego zestawu danych w usłudze Azure ML

Basic info

Name * Dataset version

titanic-imputed 1

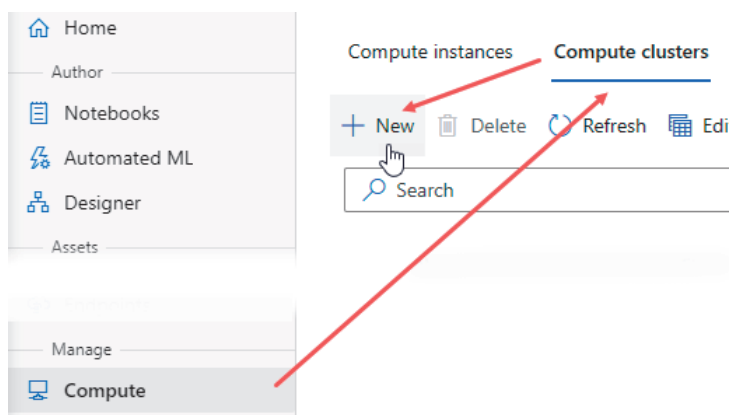
Dataset type * ⓘ

Tabular

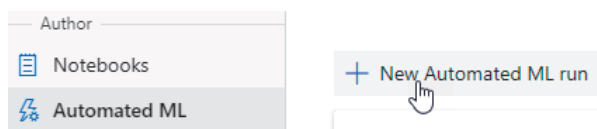
Description

Dataset description

Rysunek 13.26. Wybieranie nazwy i typu zestawu danych



Rysunek 13.27. Tworzenie nowego klastra obliczeniowego



Rysunek 13.28. Tworzenie nowego eksperymentu AutoML

Configure run

Select from existing experiments or create a new experiment

Dataset

titanic-imputed ([View dataset](#))

Experiment name *

☐ Select existing ☒ Create new

New experiment name

titanic

Target column * [i](#)

Survived (Integer)

Select compute cluster * [i](#)


cluster

[Create a new compute](#) [Refresh compute](#)

Rysunek 13.29. Konfigurowanie przebiegu AutoML


Select task type


Select the machine learning task type for the experiment. To fine tune the experiment, choose additional configuration or featurization settings.

 **Classification**
To predict one of several categories in the target column.
yes/no, blue, red, green.



☐ Enable deep learning [i](#)

 **Regression**
To predict continuous numeric values

 **Time series forecasting**
To predict values based on time

[View additional configuration settings](#) [View featurization settings](#)

Rysunek 13.30. Konfigurowanie typu zadania AutoML

Run 1 ▶ Running

[Refresh](#)
[Generate notebook](#)

[Details](#)
[Data guardrails](#)
[Models](#)

Properties

Status

▶ Running ▼

Setting up the run

Submitting run to compute

Rysunek 13.31. Uruchomiony eksperyment AutoML

Run 1 ✔ Completed

[Refresh](#)
[Generate notebook](#)
[Cancel](#)

[Details](#)
[Data guardrails](#)
[Models](#)
[Outputs + logs](#)
[Child runs](#)
[Snapshot](#)

[Deploy](#)
[Download](#)
[Explain model](#)
[Refresh](#)
[Edit columns](#)
[Reset view](#)

Showing 1-25 of 58 models

Algorithm name	Explained	AUC weighted ↓
VotingEnsemble	View explanation	0.89354
StandardScalerWrapper, XGBoostClassifier		0.89243
StackEnsemble		0.89230

Rysunek 13.32. Znalezione przez AutoML potoki o najlepszej wydajności

Run 63 ✔ Completed

[Refresh](#)
[Deploy](#)
[Download](#)
[Explain model](#)

[Details](#)
[Model](#)
[Explanations \(preview\)](#)
[Metrics](#)

Model summary

Algorithm name

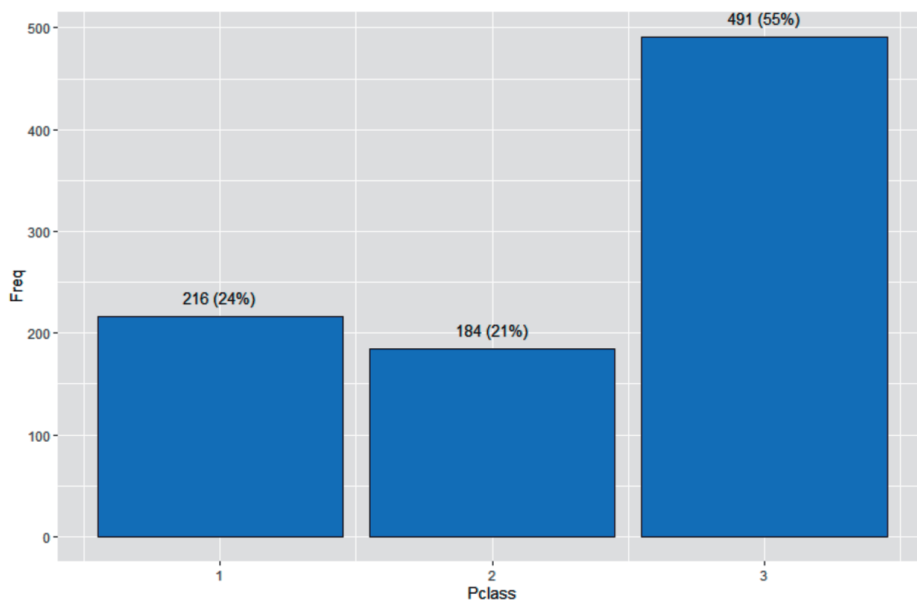
VotingEnsemble

Rysunek 13.33. Wdrażanie do usługi sieciowej najlepszego modelu

AB comment_sentiment	1.2 overall_positive_score	1.2 overall_neutral_score	1.2 overall_negative_score
mixed	0.26	0.01	0.73
negative	0.03	0.05	0.92
mixed	0.5	0.04	0.46
mixed	0.44	0.03	0.53
neutral	0.01	0.99	0

Rysunek 13.34. Dodatkowe kolumny analizy wydźwięku

Rozdział 14. Eksploracyjna analiza danych



Rysunek 14.1. Wykres słupkowy dla zmiennej Pclass

attribute	value
all_missing_columns	0.00
columns	11.00
complete_rows	183.00
continuous_columns	6.00
discrete_columns	5.00
duplicated_rows	0.00
memory_usage	184,456.00
rows	891.00
total_missing_values	866.00
total_observations	9,801.00
Total	196,219.00

Rysunek 14.2. Podstawowa tabela informacyjna w pierwszej fazie

Podsumowanie dla zbioru danych

Variable	Stats Values	Unique Valid	Freq of Valid	Valid	Missing
Age	Mean (sd) : 29.7 (14.5) min < med < max: 0.4 < 28 < 80 IQR (CV) : 17.9 (0.5)	88 distinct values	88 distinct values	714 (80.1%)	177 (19.9%)
Cabin	1. A10 2. A14 3. A16 4. A19 5. A20 6. A23 7. A24 8. A26	147 distinct values	1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%)	204 (22.9%) 687 (77.1%)	

Statystyki opisowe dla zmiennych liczbowych

variable	mean	sd	min	q1	med	q3	max	mad	iqr	cv	skewness	kurtosis
Age	29.70	14.53	0.42	20.00	28.00	38.00	80.00	13.34	17.88	0.49	0.39	0.16
Fare	32.20	49.69	0.00	7.90	14.45	31.00	512.33	10.24	23.09	1.54	4.77	33.12
Parch	0.38	0.81	0.00	0.00	0.00	0.00	6.00	0.00	0.00	2.11	2.74	9.69
Pclass	2.31	0.84	1.00	2.00	3.00	3.00	3.00	0.00	1.00	0.36	-0.63	-1.28
SibSp	0.52	1.10	0.00	0.00	0.00	1.00	8.00	0.00	1.00	2.11	3.68	17.73
Survived	0.38	0.49	0.00	0.00	0.00	1.00	1.00	0.00	1.00	1.27	0.48	-1.77

Podstawowe informacje o zbiorze danych

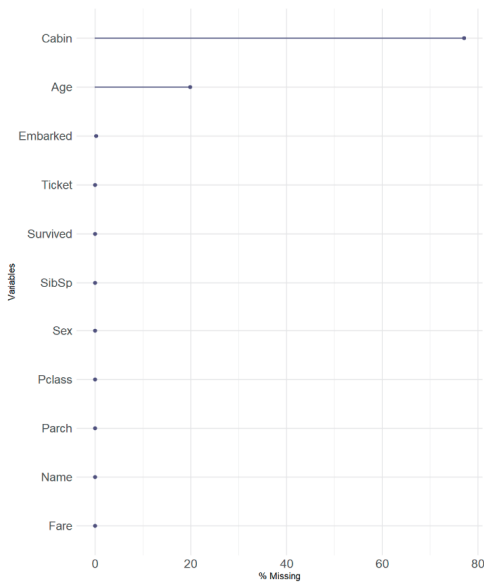
attribute	value
all_missing_columns	0.00
columns	11.00
complete_rows	183.00
continuous_columns	6.00
discrete_columns	5.00
duplicated_rows	0.00
memory_usage	184 456.00
rows	891.00
total_missing_values	866.00
total_observations	9 801.00

Próbka zestawu danych

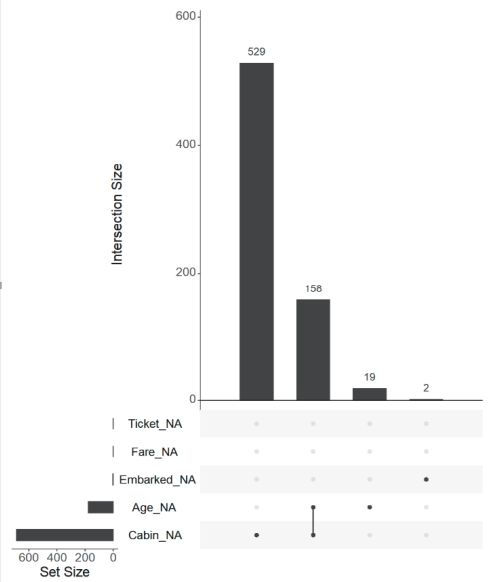
Name	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
Ahlin, Mrs. Johan (Johanna Persdotter Larsson)	3	female	40.00	1	0	7546	9.48		S
Allen, Mr. William Henry	3	male	35.00	0	0	373450	8.05		S
Andersson, Miss. Erna Alexandra	3	female	17.00	4	2	3101281	7.93		S
Andersson, Mr. Anders Johan	3	male	39.00	1	5	347082	31.28		S
Andersson, Mr. Paul Edwin	3	male	20.00	0	0	347466	7.85		S
Arnold-Franchi, Mrs. Josef (Josefine Franchi)	3	female	18.00	1	0	349237	17.80		S
Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	3	female	38.00	1	5	347077	31.39		S
Backstrom, Mrs. Karl Alfred (Maria Mathilda Gustafsson)	3	female	33.00	3	0	3101278	15.85		S
Beesley, Mr. Lawrence	2	male	34.00	0	0	248698	13.00	D56	S
Bing, Mr. Lee	3	male	32.00	0	0	1601	56.50		S
Bonnell, Miss. Elizabeth	1	female	58.00	0	0	113783	26.55	C103	S
Braund, Mr. Owen Harris	3	male	22.00	1	0	A/5 21171	7.25		S
Caldwell, Master. Alden Gates	2	male	0.83	0	2	248738	29.00		S
Cann, Mr. Ernest Charles	3	male	21.00	0	0	A/5. 2152	8.05		S
Carrau, Mr. Francisco M	1	male	28.00	0	0	113059	47.10		S
Celotti, Mr. Francesco	3	male	24.00	0	0	343275	8.05		S
Suma	240		2 142.33	73	44		2 951.76		

Rysunek 14.3. Strona podsumowania raportu EDA

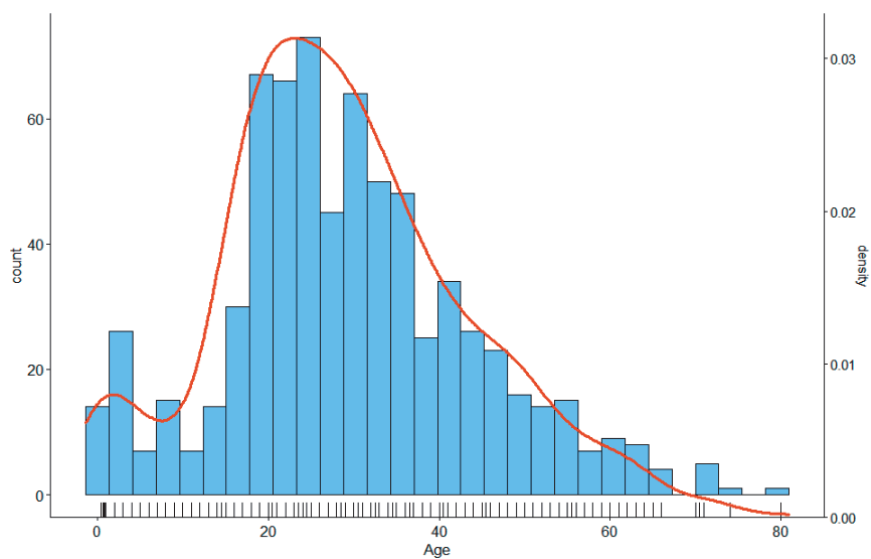
Procent brakujących wartości



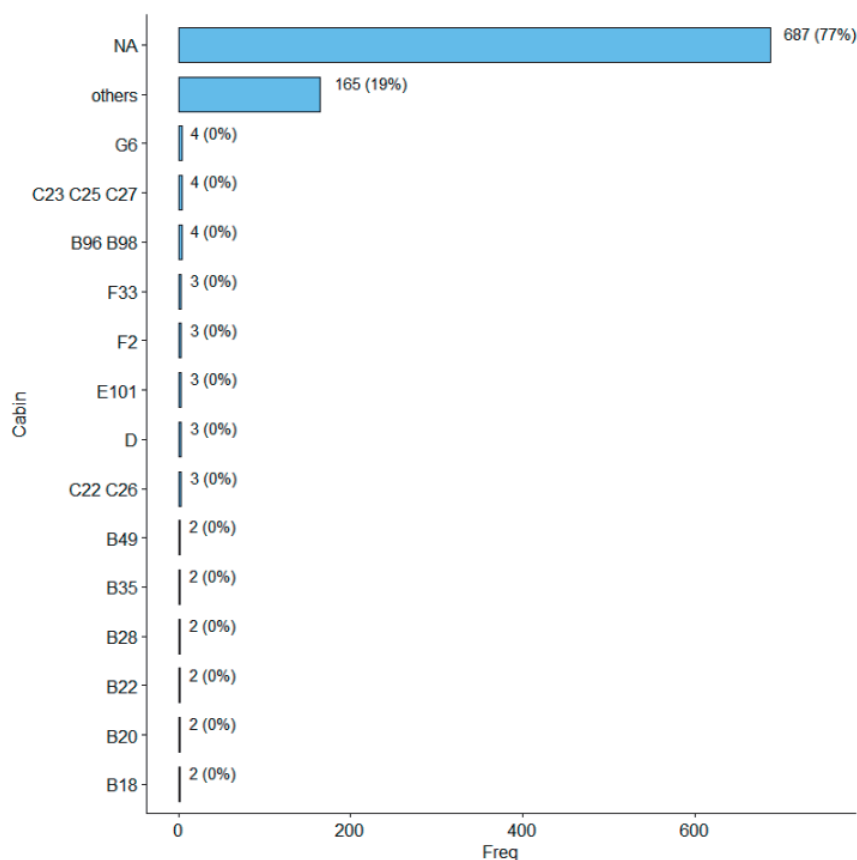
Zmienne brakujące - razem



Rysunek 14.4. Strona analizy brakujących wartości w raporcie EDA



Rysunek 14.5. Histogram i wykres częstości dla zmiennej Age



Rysunek 14.6. Wykres słupkowy dla zmiennej Cabin

numeric_df

numeric_col_name	transf_type
Survived	standard
Pclass	standard
Age	standard
SibSp	standard
Parch	standard
Fare	standard
Survived	yeo-johnson
Pclass	yeo-johnson
Age	yeo-johnson
SibSp	yeo-johnson
Parch	yeo-johnson
Fare	yeo-johnson

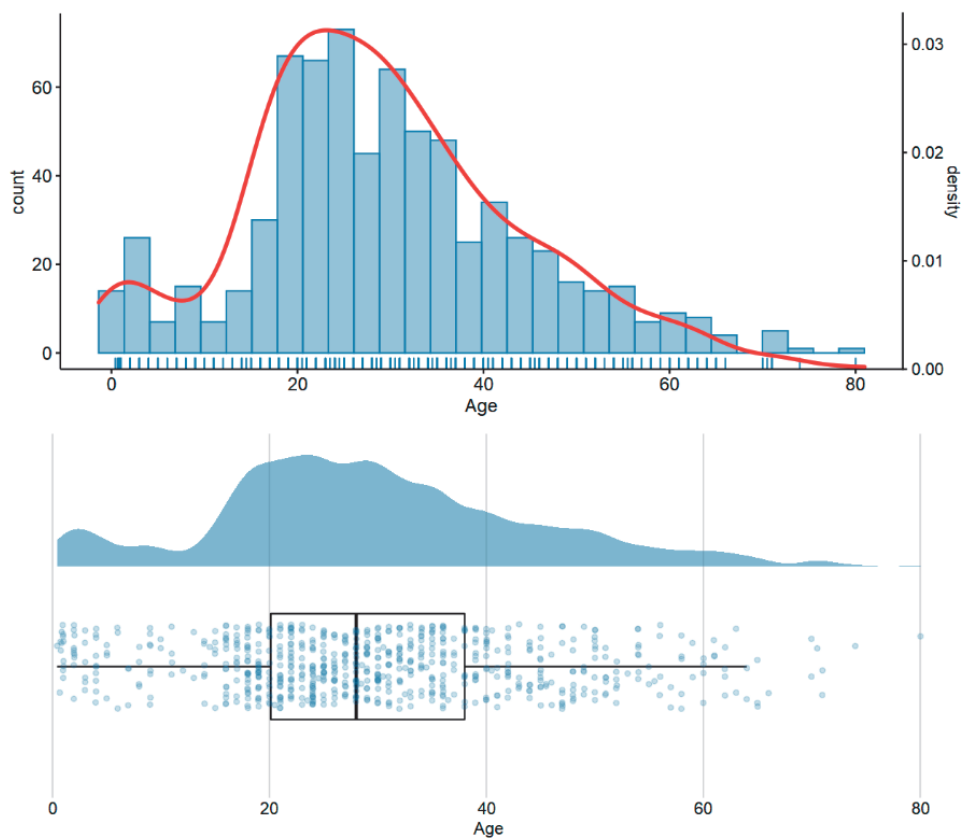
Rysunek 14.7. Zawartość tabeli numeric_df

Zmienne liczbowe

Age

Przekształcenia

standard



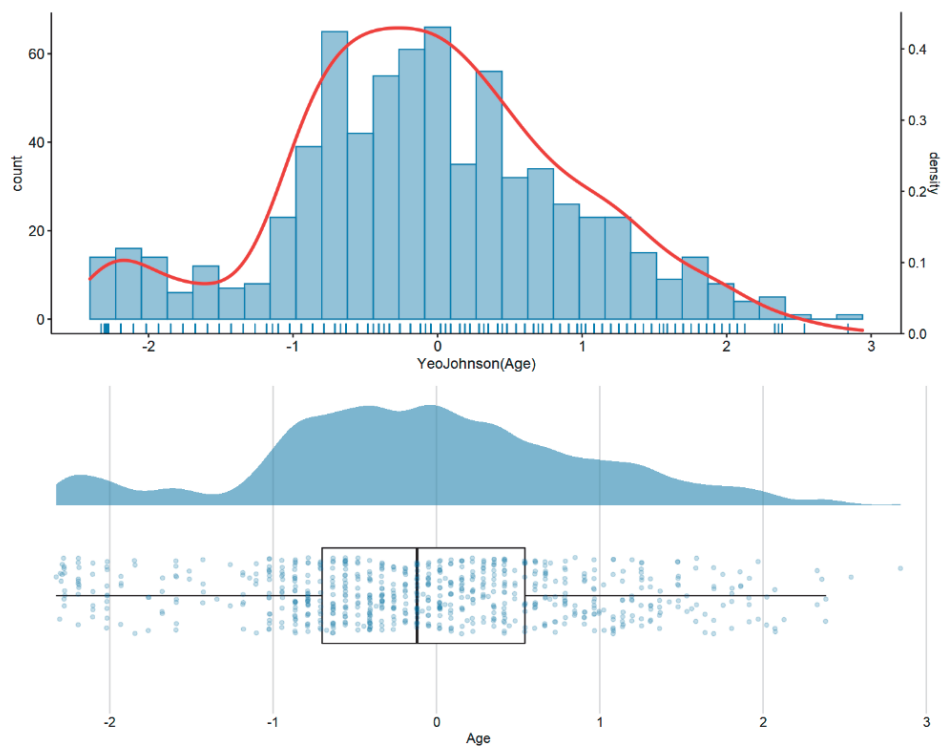
Rysunek 14.8 Jednowymiarowa analiza zmiennej Age

Zmienne liczbowe

Przekształcenia

Age

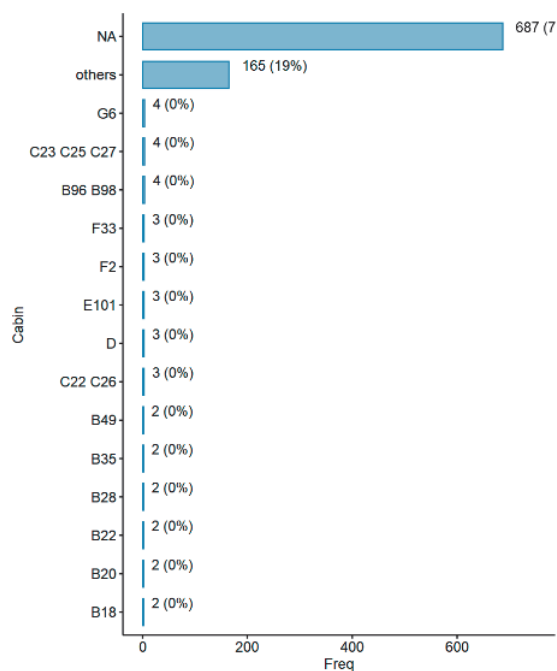
yeo-johnson



Rysunek 14.9. Jednowymiarowa analiza zmiennej Age przekształconej z wykorzystaniem przekształcenia yeo-johnsona

Zmienne kateryczne

Cabin



Rysunek 14.10. Jednowymiarowa analiza zmiennej Cabin

Zmienne liczbowe

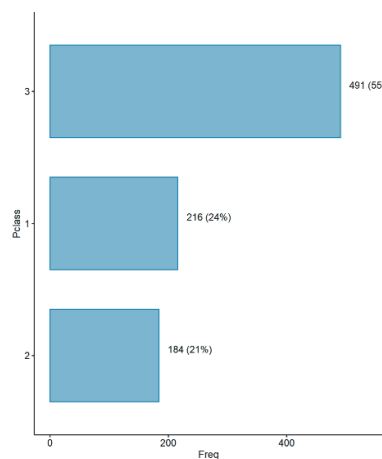
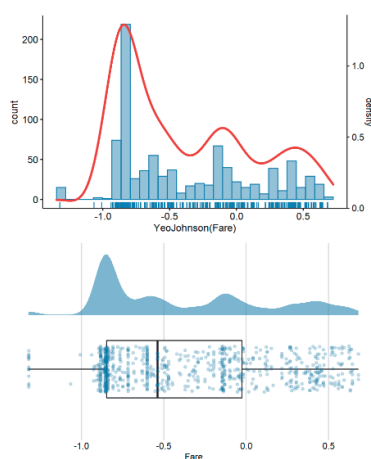
Fare

Przekształcenia

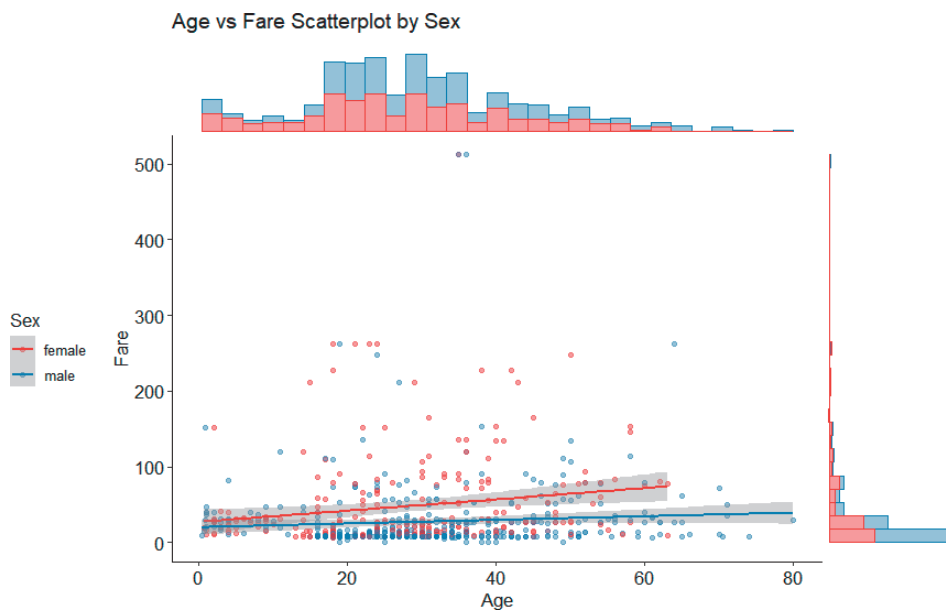
yeo-johnson

Zmienne kateryczne

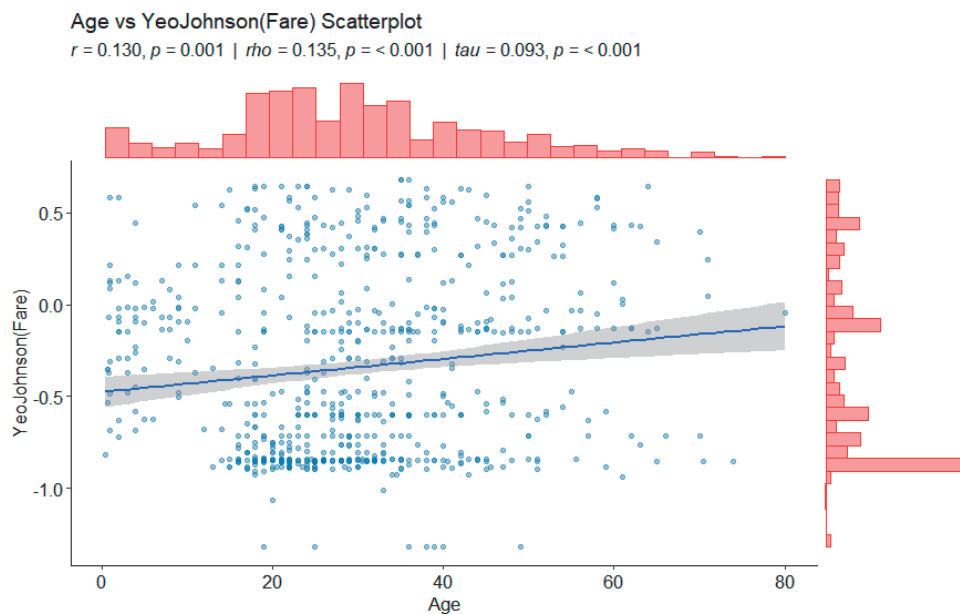
Pclass



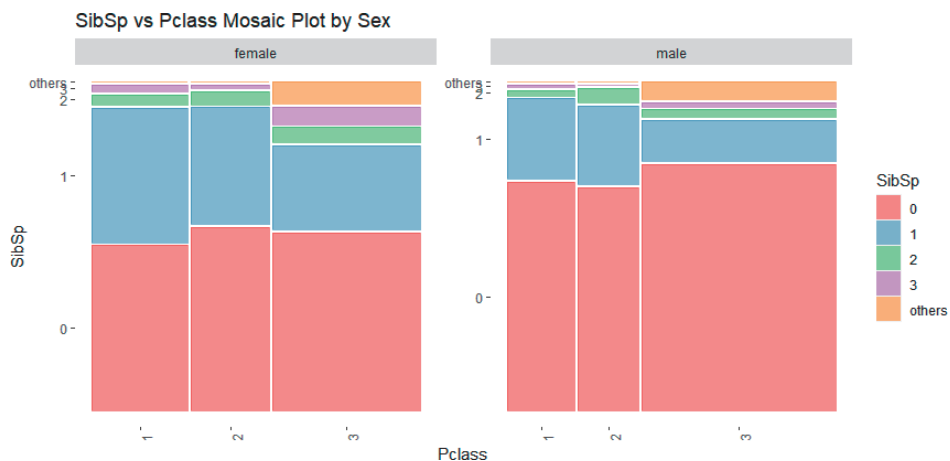
Rysunek 14.11. Strona raportu EDA dla analizy jednowymiarowej



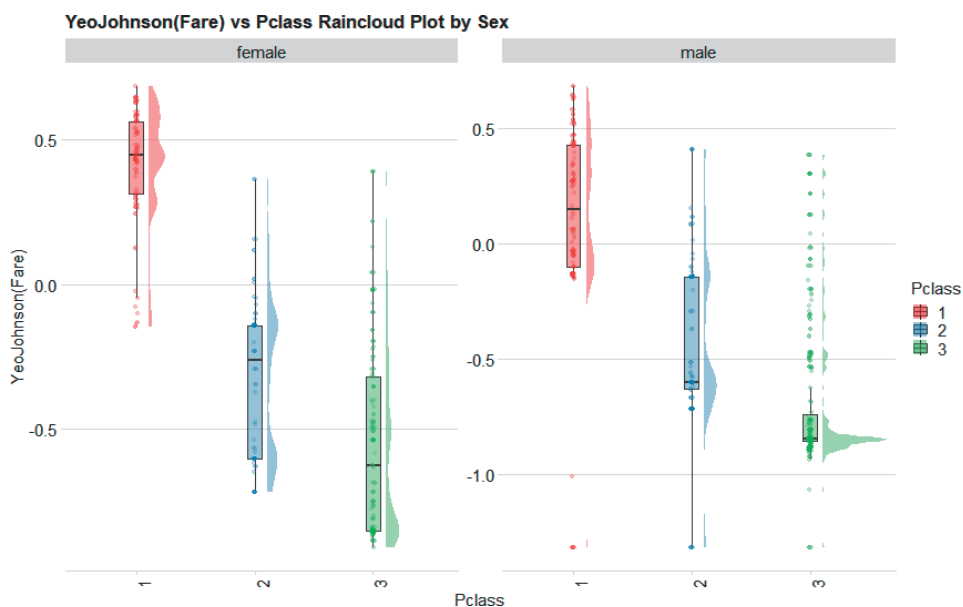
Rysunek 14.12. Wykres punktowy z histogramami krańcowymi dla zmiennych Age i Fare według zmiennej Sex



Rysunek 14.13. Wykres punktowy dla kategorii zmiennych Age i Fare (przekształcony), bez grupowania



Rysunek 14.14. Wykres mozaikowy dla zmiennych Pclass i SibSp pogrupowany według zmiennej Sex



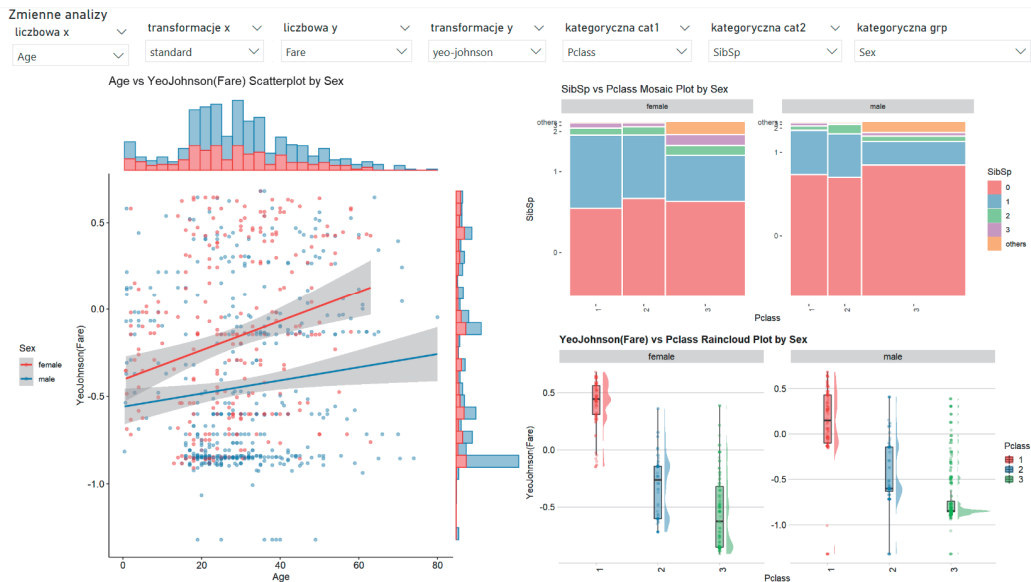
Rysunek 14.15. Wykresy chmury deszczowej zmiennych Fare i Pclass pogrupowane według zmiennej Sex

A^B_c x	A^B_c x_transf_type	A^B_c y	A^B_c y_transf_type	A^B_c cat1	A^B_c cat2	A^B_c BFP
Survived	standard	Survived	standard	Survived	Survived	Survived
Pclass	standard	Survived	standard	Survived	Survived	Survived
Age	standard	Survived	standard	Survived	Survived	Survived
SibSp	standard	Survived	standard	Survived	Survived	Survived
Parch	standard	Survived	standard	Survived	Survived	Survived
Fare	standard	Survived	standard	Survived	Survived	Survived

Rysunek 14.16. Zawartość tabeli multivariate_df

Zmienne analizy
 liczbowe x: Age, transformacje x: standard, liczbowe y: Fare, transformacje y: Wszystkie, kategoryczne cat1: Cabin, kategoryczne cat2: Cabin, kategoryczna grp: Sex

Rysunek 14.17. Fragmentatory na górze strony analizy wielowymiarowej



Rysunek 14.18. Strona raportu EDA dla analizy wielowymiarowej



Rysunek 14.19. Mapa termiczna korelacji zbioru danych

corr_tbl

row	col	numeric_corr_type	categorical_corr_type	corr
Age	Age	pearson	theil	1
Age	Age	pearson	cramer	1
Age	Age	spearman	theil	1
Age	Age	spearman	cramer	1
Age	Age	kendall	theil	1
Age	Age	kendall	cramer	1
Age	Cabin	pearson	theil	0.524193719
Age	Cabin	pearson	cramer	0.524193719
Age	Cabin	spearman	theil	0.524193719

Rysunek 14.20. Zawartość tabeli corr_tbl

Zmienne analizy

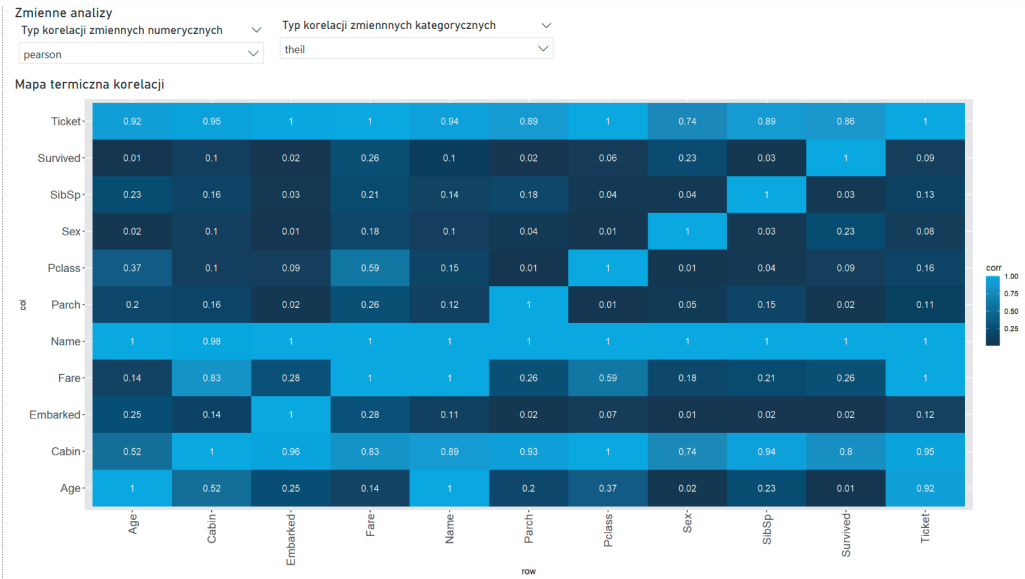
Typ korelacji zmiennych numerycznych

pearson

Typ korelacji zmiennych kategoriycznych

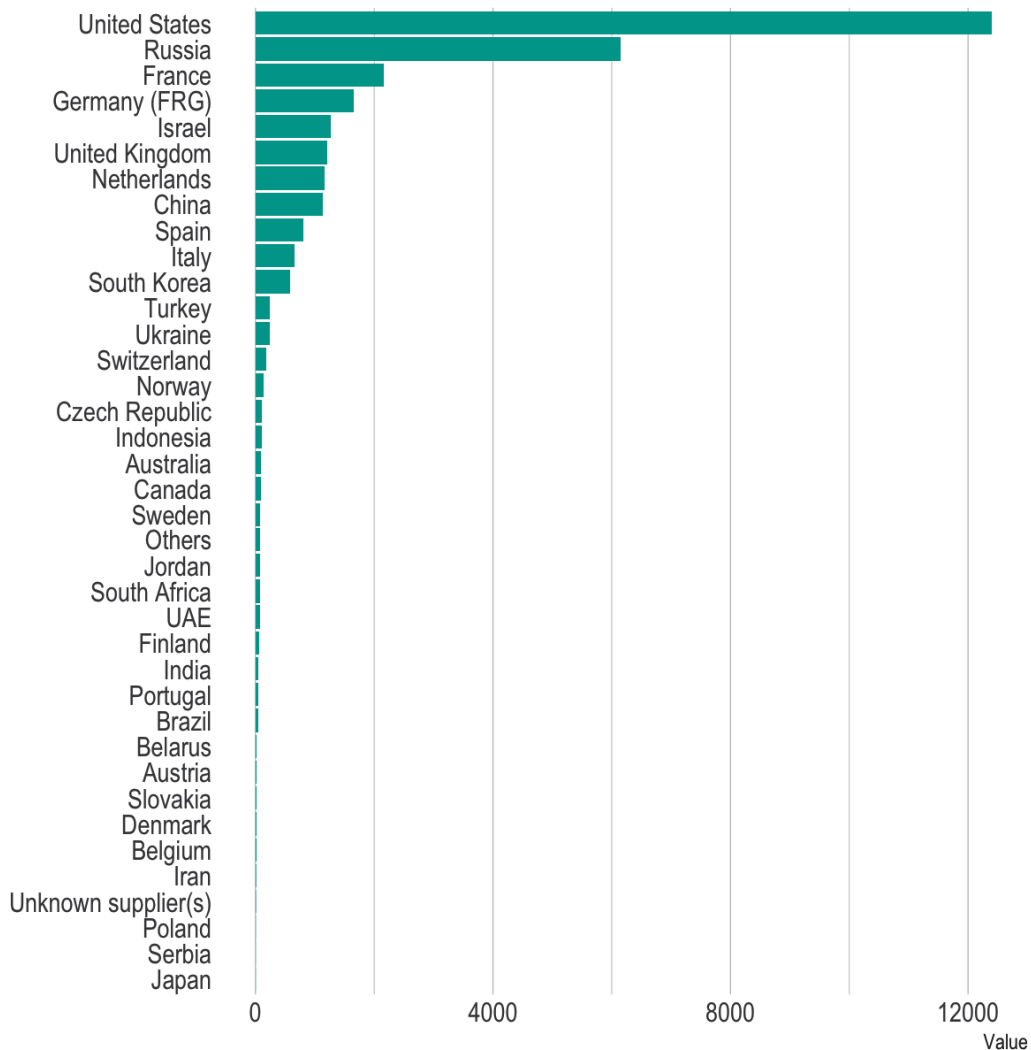
theil

Rysunek 14.21. Fragmentary strony Powiązania u góry strony

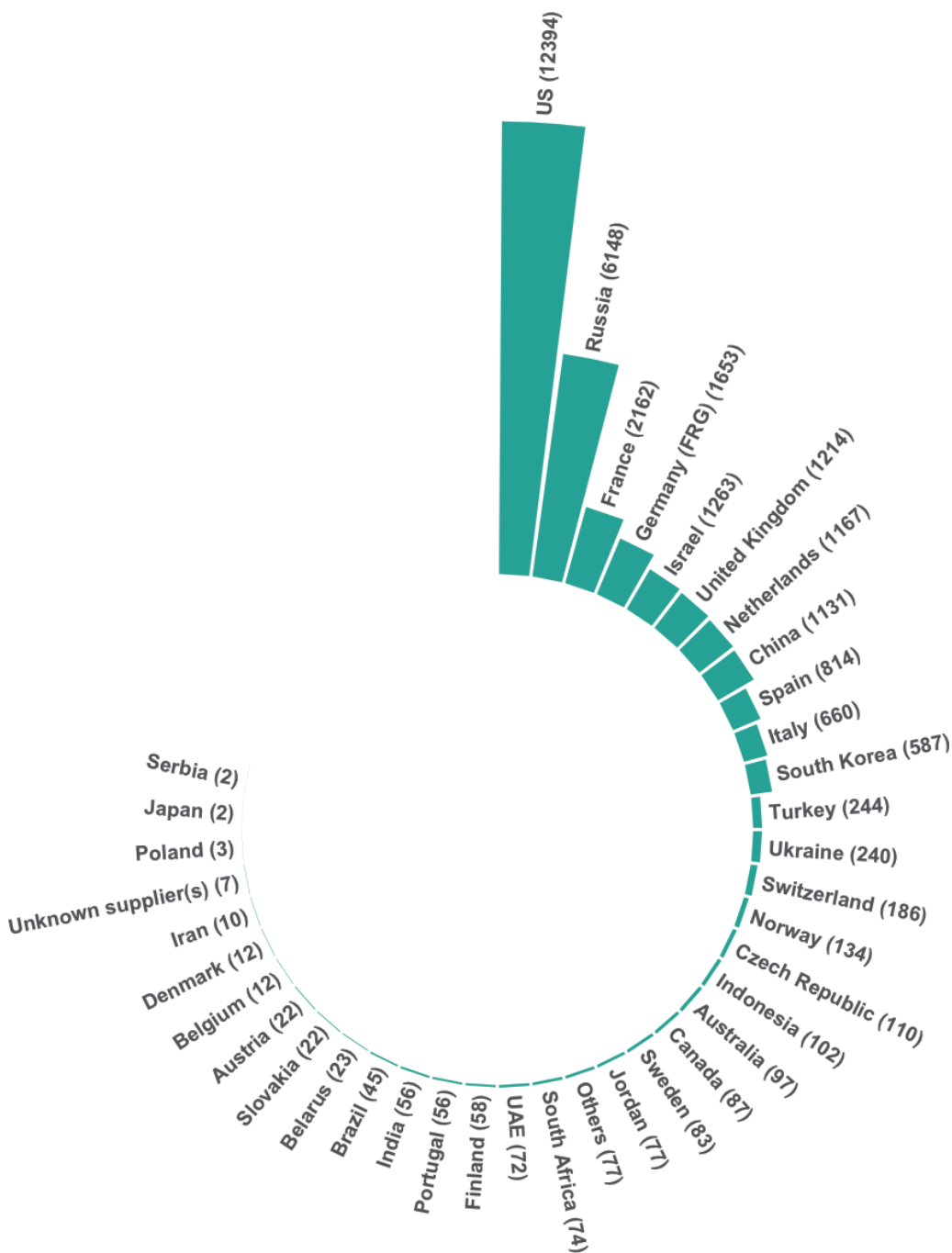


Rysunek 14.22. Strona analizy powiązań raportu EDA

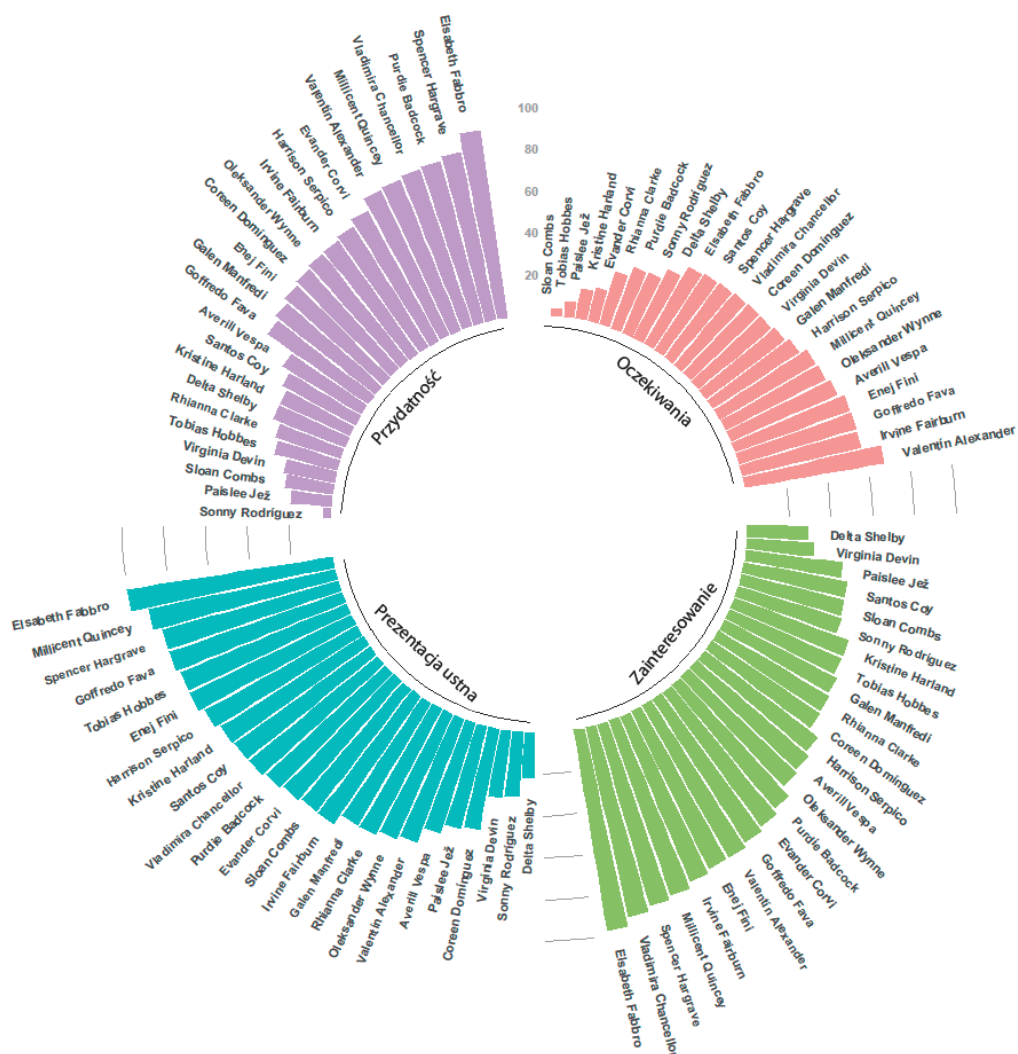
Rozdział 15. Zaawansowane wizualizacje



Rysunek 15.1. Wykres słupkowy światowych eksporterów broni



Rysunek 15.2. Okrągły wykres słupkowy światowych eksporterów broni



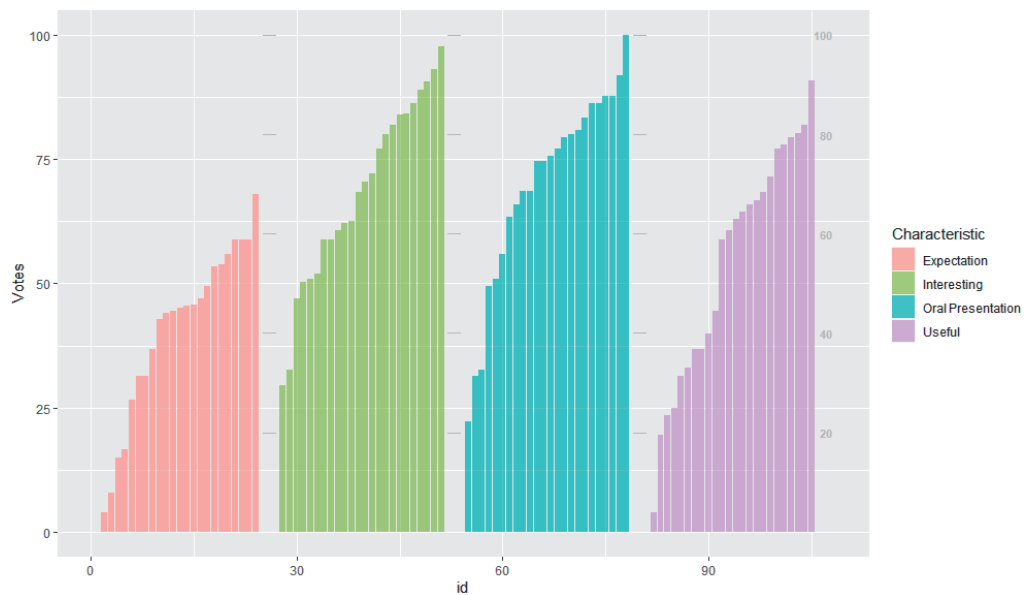
Rysunek 15.3. Kołowy wykres słupkowy pogrupowany według ocen prelegentów

```
# A tibble: 72 x 3
  SpeakerName Characteristic votes
  <chr>      <chr>      <dbl>
1 Sloan Combs Expectation 3.93
2 Tobias Hobbes Expectation 4
3 Paislee Jez Expectation 4.07
4 Kristine Harland Expectation 4.2
5 Evander Corvi Expectation 4.23
6 Rhianna Clarke Expectation 4.41
7 Purdie Badcock Expectation 4.5
8 Sonny Rodriguez Expectation 4.5
9 Delta Shelby Expectation 4.6
10 Elisabeth Fabbro Expectation 4.71
# ... with 62 more rows
```

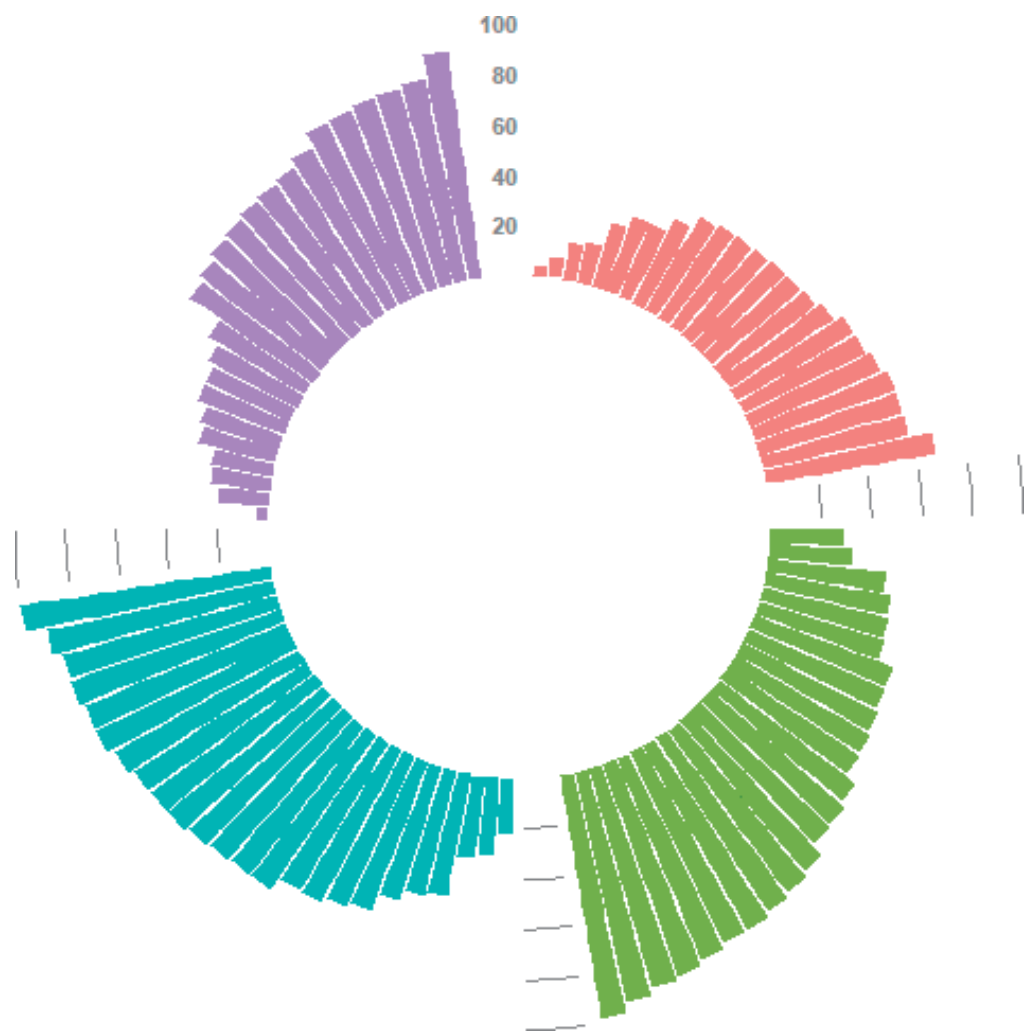
Rysunek 15.4. Obiekt tibble z ocenami prelegentów

	SpeakerName	Characteristic	votes
1	NA	Expectation	NA
2	NA	Expectation	NA
3	NA	Expectation	NA
4	NA	Interesting	NA
5	NA	Interesting	NA
6	NA	Interesting	NA
7	NA	OralPresentation	NA
8	NA	OralPresentation	NA
9	NA	OralPresentation	NA
10	NA	Useful	NA
11	NA	Useful	NA
12	NA	Useful	NA

Rysunek 15.5. Puste kolumny ramki danych



Rysunek 15.6. Pierwsza, robocza wersja wykresu słupkowego

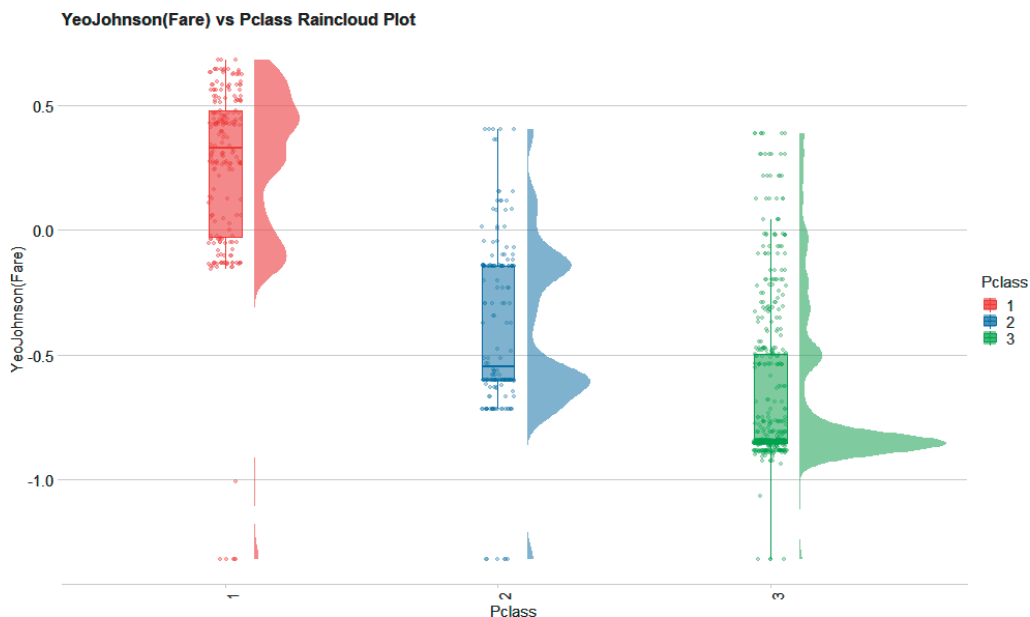


Rysunek 15.7. Pierwsza wersja kołowego wykresu słupkowego

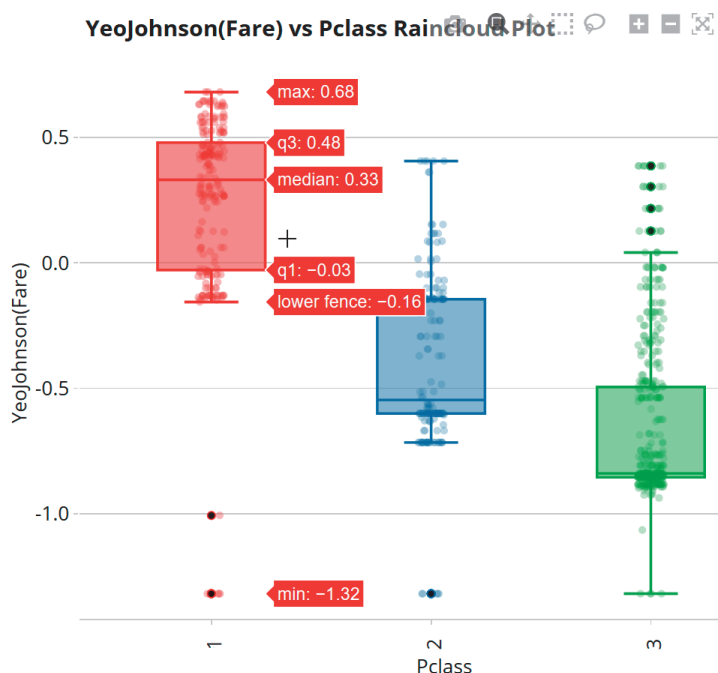


Rysunek 15.8. Kołowy wykres słupkowy filtrowany przez fragmentator w usłudze Power BI

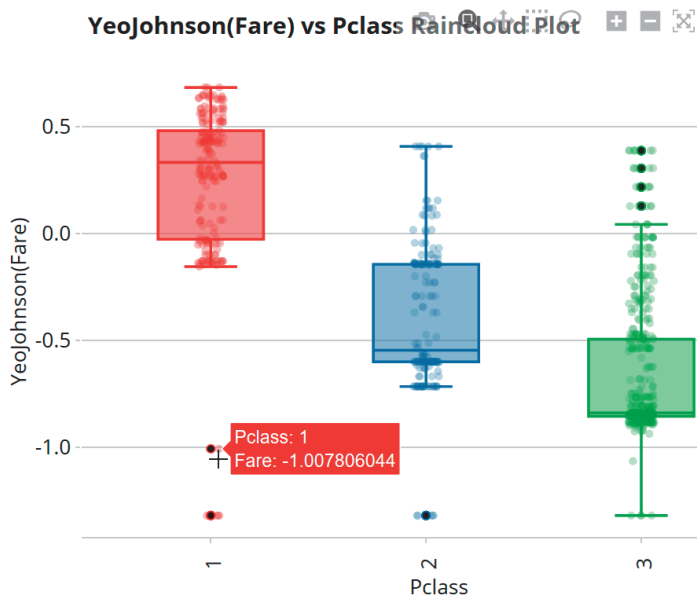
Rozdział 16. Interaktywne niestandardowe wizualizacje języka R



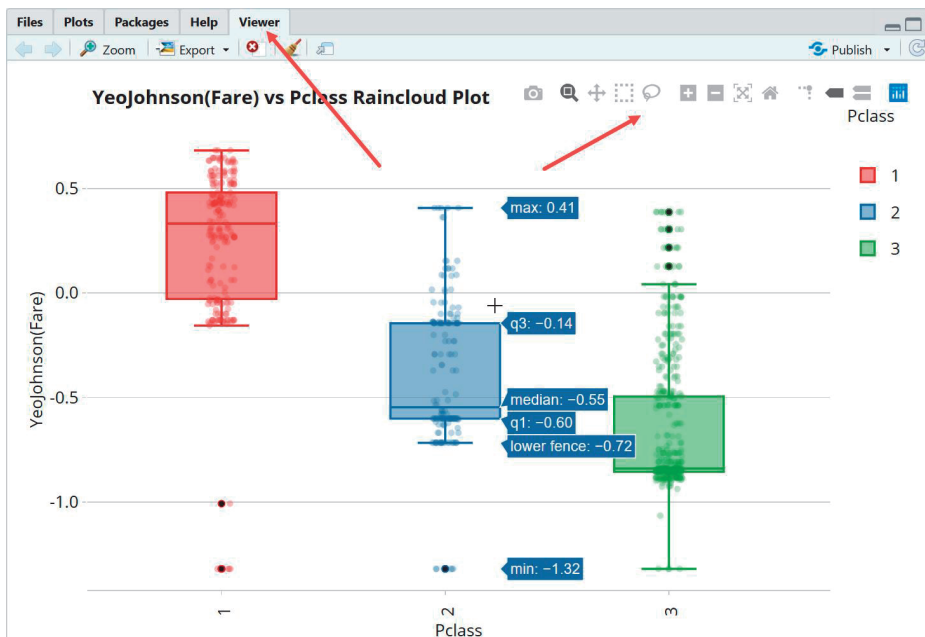
Rysunek 16.1. Wykres chmury deszczowej dla zmiennych Fare (przekształconej) i zmiennej Pclass



Rysunek 16.2. Główne etykiety pokazane na wykresie Fare (przekształcona) dla pierwszej klasy



Rysunek 16.3. Wartości zmiennej Fare (przekształconej) i zmiennej Pclass dla wyróżnionego elementu odstającego



Rysunek 16.4. Wynik zastosowania funkcji ggplotly() do wykresu chmury deszczowej

```

(base) PS C:\Users\LucaZavarella> pbiviz

+syys+/
oms/+osyhdhyso/
ym/      /+oshddhys+/
ym/      /+oyhddhyo+/
ym/      /osyhdho
ym/      yddy      sm+
ym/      shho /mmm/ om+
ym/      / oys/ +mmm /mmm/ om+
oso ommmh +mmm /mmm/ om+
ymmy smmmh +mmm /mmm/ om+
ymmy smmmh +mmm /mmm/ om+
ymmy smmmh +mmm /mmm/ om+
+dmd+ smmmh +mmm /mmm/ om+
      /hmdo +mmm /mmm/ /so+/ym/
      /dmh /mmm/ /osyhhy/
      // dmd
      ++

PowerBI Custom Visual Tool

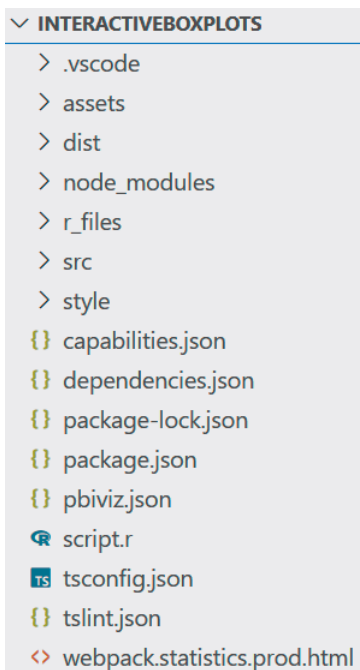
Usage: pbiviz [options] [command]

Options:
-V, --version      output the version number
--install-cert     Creates and installs localhost certificate
-h, --help         output usage information

Commands:
new [name]         Create a new visual
info              Display info about the current visual
start            Start the current visual
package          Package the current visual into a pbiviz file
update [version]  Updates the api definitions and schemas in the
current visual. Changes the version if specified
help [cmd]        display help for [cmd]
(base) PS C:\Users\LucaZavarella>

```

Rysunek 16.5. Poprawnie zainstalowane narzędzia pbiviz



Rysunek 16.6. Widok zawartości folderu `interactiveboxplots` w programie VS Code



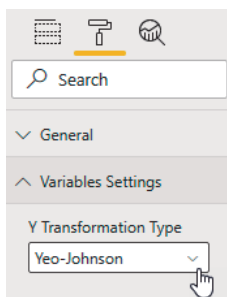
Rysunek 16.7. Edycja zawartości pliku `pbiviz.json` w programie VS Code

<pre> 1 { 2 "dataRoles": [3 { 4- "displayName": "Values", 5 6- "kind": "GroupingOrMeasure", 7- "name": "Values" 8] 9 } </pre>	<pre> 1 { 2 "dataRoles": [3 { 4+ "displayName": "X Split Variable", 5+ "description": "Integer or categorical variable used on x 6+ axis to split boxplots", 7+ 8+ "kind": "GroupingOrMeasure", 9+ "name": "x" 10+ }, 11+ { 12+ "displayName": "Y Quantitative Variable", 13+ "description": "Numeric variable (integer or double) used on 14+ y axis, the distribution of which is represented by 15+ boxplots", 16+ "kind": "GroupingOrMeasure", 17+ "name": "y" 18+ }, 19+ { 20+ "displayName": "Grouping Variable", 21+ "description": "Integer or categorical variable used to 22+ group data in facets", 23+ "kind": "GroupingOrMeasure", 24+ "name": "grp" 25+ } 26] 27 } </pre>
--	--

Rysunek 16.8. Edycja sekcji dataRoles pliku capabilities.json

<pre> "dataViewMappings": [{ "scriptResult": { "dataInput": { "table": { "rows": { "select": [{ "for": { "in": "Values" } }] } } } } }], "script": { "scriptProviderDefault": "R", "scriptOutputType": "html", "source": { "objectName": "rcv_script", "propertyName": "source" }, "provider": { "objectName": "rcv_script", "propertyName": "provider" } } </pre>	<pre> "dataViewMappings": [{ "scriptResult": { "dataInput": { "table": { "rows": { "select": [{ "for": { "in": "x" } }, { "for": { "in": "y" } }, { "for": { "in": "grp" } }] } } } } }], "script": { "scriptProviderDefault": "R", "scriptOutputType": "html", "source": { "objectName": "rcv_script", "propertyName": "source" }, "provider": { "objectName": "rcv_script", "propertyName": "provider" } } </pre>
--	---

Rysunek 16.9. Edycja węzła dataViewMappings w pliku capabilities.json



Rysunek 16.10. Niestandardowe parametry w panelu Formatuj element wizualny

<pre> 43 "objects": { 44 "rcv_script": { 45 "properties": { 46 "provider": { 47 "type": { 48 "text": true 49 } 50 }, 51 "source": { 52 "type": { 53 "scripting": { 54 "source": true 55 } 56 } 57 } 58 } </pre>	<pre> 66 "objects": { 67 "rcv_script": { 68 "properties": { 69 "provider": { 70 "type": { 71 "text": true 72 } 73 }, 74 "source": { 75 "type": { 76 "scripting": { 77 "source": true 78 } 79 } 80 } 81 } 82 }, 83 "settings_variable_params": { 84 "displayName": "Variables Settings", 85 "properties": { 86 "y_transf_name": { 87 "displayName": "Y Transformation Type", 88 "description": "Type of transformation to be + applied to the variable Y. The type 'Standard' + indicates that no transformation is applied", 89 "type": { 90 "enumeration": [91 { 92 "displayName": "Standard", 93 "value": "standard" 94 }, 95 { 96 "displayName": "Yeo-Johnson", 97 "value": "yeo-johnson" 98 } 99] 100 } 101 } 102 } 103 } 104 }, 105 "suppressDefaultTitle": true 106 } </pre>
---	--

Rysunek 16.11. Edycja obszaru objects pliku capabilities.json

<pre> 27 "use strict"; 28 29 import { dataViewObjectsParser } from "powerbi-visuals-utils-dataviewutils"; 30 import DataViewObjectsParser = dataViewObjectsParser.DataViewObjectsParser; 31 32 export class VisualSettings extends DataViewObjectsParser { 33- public rcv_script: rcv_scriptSettings = 34- new rcv_scriptSettings(); 35 } 36- export class rcv_scriptSettings { 37- // undefined 38- public provider // undefined 39- public source } </pre>	<pre> 27 "use strict"; 28 29 import { dataViewObjectsParser } from "powerbi-visuals-utils-dataviewutils"; 30 import DataViewObjectsParser = dataViewObjectsParser. DataViewObjectsParser; 31 32 export class VisualSettings extends DataViewObjectsParser { 33+ public settings_variable_params: settings_variable_params = 34+ new settings_variable_params(); 35 } 36+ export class settings_variable_params { 37+ public y_transf_name: string = "standard"; 38+ } </pre>
---	--

Rysunek 16.12. Edycja pliku settings.ts


```

[DONE] Compiled successfully in 2802ms 6:19:25 PM

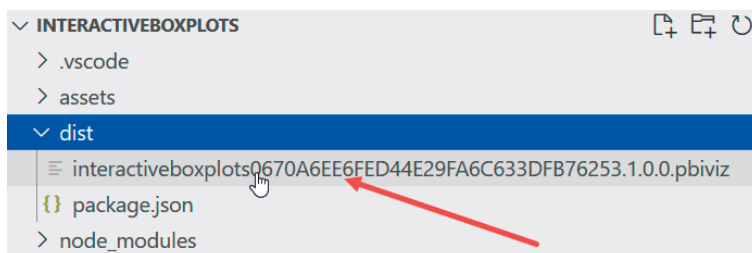
warn Please, make sure that the visual source code matches to requirements of certification:

info Visual must use API v3.8 and above
info The project repository must:
info Include package.json and package-lock.json;
info Not include node_modules folder
info Run npm install expect no errors
info Run pbiviz package expect no errors
info The compiled package of the Custom Visual should match submitted package.
info npm audit command must not return any alerts with high or moderate level.
info The project must include Tslint from Microsoft with no overridden configuration, and this command should
n't return any tslint errors.
info https://www.npmjs.com/package/tslint-microsoft-contrib
info Ensure no arbitrary/dynamic code is run (bad: eval(), unsafe use of setTimeout(), requestAnimationFrame(
), setInterval(some function with user input).. running user input/data etc.)
info Ensure DOM is manipulated safely (bad: innerHTML, D3.html(<some user/data input>), unsanitized user input
t/data directly added to DOM, etc.)
info Ensure no js errors/exceptions in browser console for any input data. As test dataset please use this sa
mple report

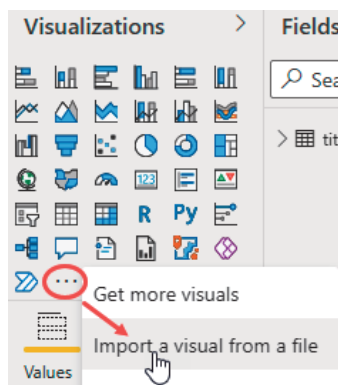
info Full description of certification requirements you can find in documentation:
info https://docs.microsoft.com/en-us/power-bi/power-bi-custom-visuals-certified#certification-requirements
(base) PS C:\Users\LucaZavarella\Power-BI-Custom-Visuals\interactiveboxplots>

```

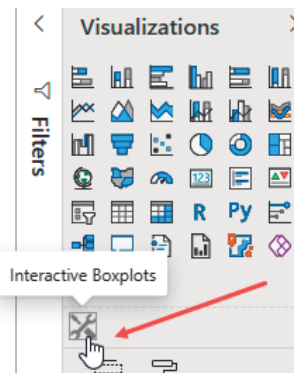
Rysunek 16.13. Pomyślna kompilacja niestandardowej wizualizacji



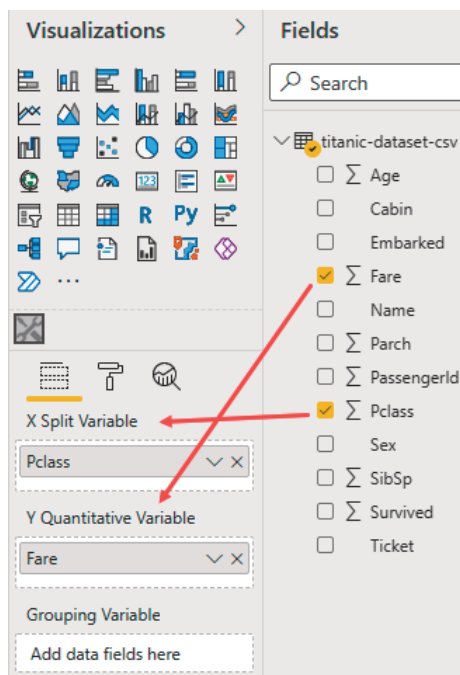
Rysunek 16.14. Skompilowany pakiet .pbviz



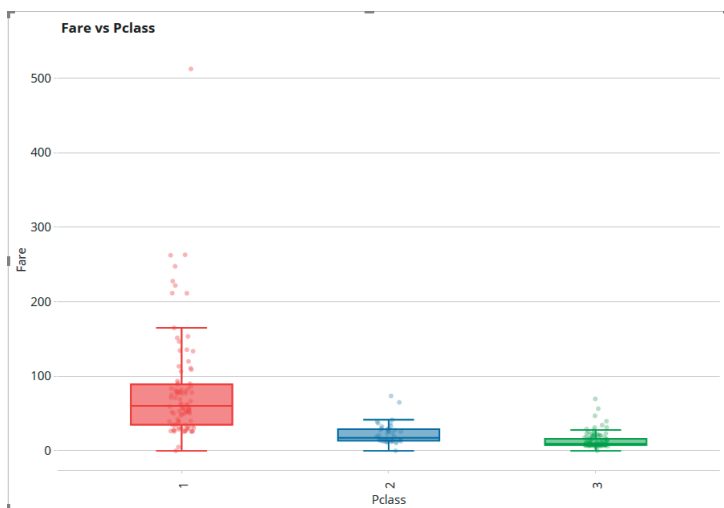
Rysunek 16.15. Importowanie niestandardowej wizualizacji z pliku



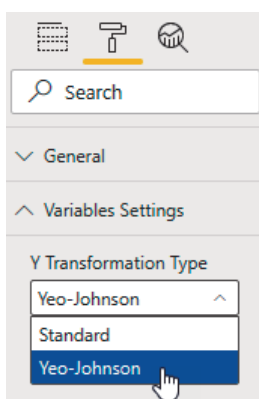
Rysunek 16.16. Importowanie niestandardowej wizualizacji z pliku



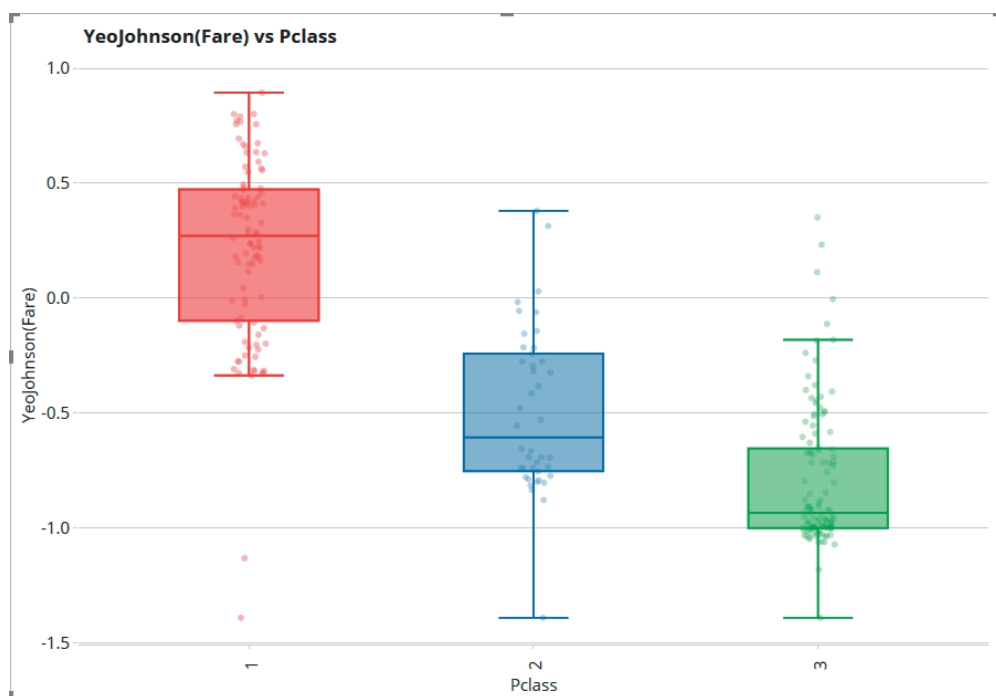
Rysunek 16.17. Zaznacz pola Pclass i Fare jako zmienne X i Y



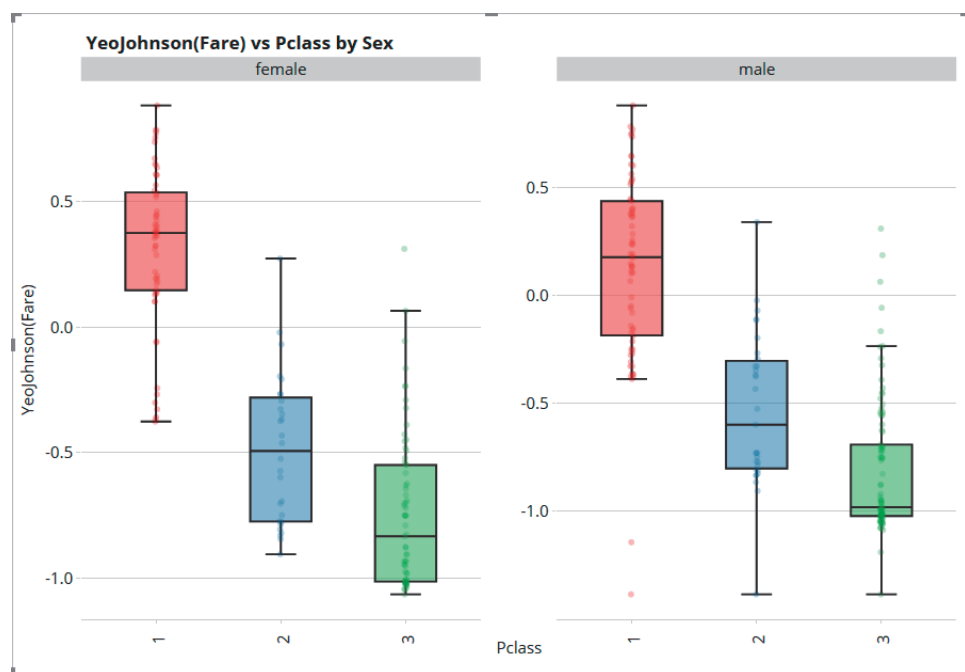
Rysunek 16.18. Niestandardowa wizualizacja przedstawiająca wykresy pudełkowe dla zmiennych Fare i Pclass



Rysunek 16.19. Wybierz Yeo-Johnsona jako typ transformacji Y



Rysunek 16.20. Wizualizacja niestandardowa przedstawiająca wykresy pudełkowe dla zmiennej Fare (przekształconej) w porównaniu ze zmienną Pclass



Rysunek 16.21. Niestandardowa wizualizacja przedstawiająca wykresy pudełkowe dla zmiennych Fare (przekształconej) i Pclass z podziałem na grupy według zmiennej Sex