
# Eksploracyjna analiza danych

## Wykład 3

October 14, 2021

Co-funded by the
Erasmus+ Programme
of the European Union

# Outline



Co-funded by the Erasmus+ Programme of the European Union

# Terminologia danych I

Po tej sekcji będziesz mógł...

- Opisać czym jest dana cecha (feature) i jaki jest jej związek z próbką
- Nazwać kilka alternatywnych terminów dla „feature”
- Podsumować, jak kategoryczna cecha różni się od cechy liczbowej

| | number | air_pressure_9am | air_temp_9am | avg_wind_direction_9am | avg_wind_speed_9am | max_wind_direction_9am |
|---|---|---|---|---|---|---|
| **0** | 0 | 918.060000 | 74.822000 | 271.100000 | 2.080354 | 295.400000 |
| **1** | 1 | 917.347688 | 71.403843 | 101.935179 | 2.443009 | 140.471548 |
| **2** | 2 | 923.040000 | 60.638000 | 51.000000 | 17.067852 | 63.700000 |
| **3** | 3 | 920.502751 | 70.138895 | 198.832133 | 4.337363 | 211.203341 |
| **4** | 4 | 921.160000 | 44.294000 | 277.800000 | 1.856660 | 136.500000 |

Figure: Samples (Próbki) i Variables (Zmienne) są prezentowane odpowiednio jako wiersze i kolumny tabeli

Co-funded by the Erasmus+ Programme of the European Union

# Terminologia danych II

Inne nazwy dla 'Sample' (Próbka)

- record (rekord)
- example (przykład)
- row (wiersz)
- instance (instancja)
- observation (obserwacja)

Inne nazwy dla 'Variable' (Zmienna)

- attribute (atrybut)
- field (pole)
- feature (cecha)
- column (kolumna)
- dimension (wymiar)

Co-funded by the Erasmus+ Programme of the European Union

# Terminologia danych III

## Typy danych

- Najczęściej używane
  - Numeric (Liczbowy)
  - Categorical (Kategoryjny)
- Inne
  - String (Ciąg znaków)
  - Date (Data)
  - ...

## Liczbowe zmienne

- Wartości są liczby
- Nazywany również „ilościowym”
- Przykłady: 1
  163.92
  $7 \times 10^5$

Co-funded by the Erasmus+ Programme of the European Union

# Terminologia danych IV

Przykłady zmiennych liczbowych

- Wysokość
- Wynik na egzaminie
- Liczba transakcji na godzinę
- Zmiana ceny akcji

Kategoryjne Zmienne

- Wartości to etykiety, nazwy lub kategorie
- Nazywany również „jakościowym” lub „nominalnym”

Co-funded by the Erasmus+ Programme of the European Union

# Terminologia danych V

| Kolor |
|---|
| Red |
| Silver |
| Blue |
| White |
| Black |

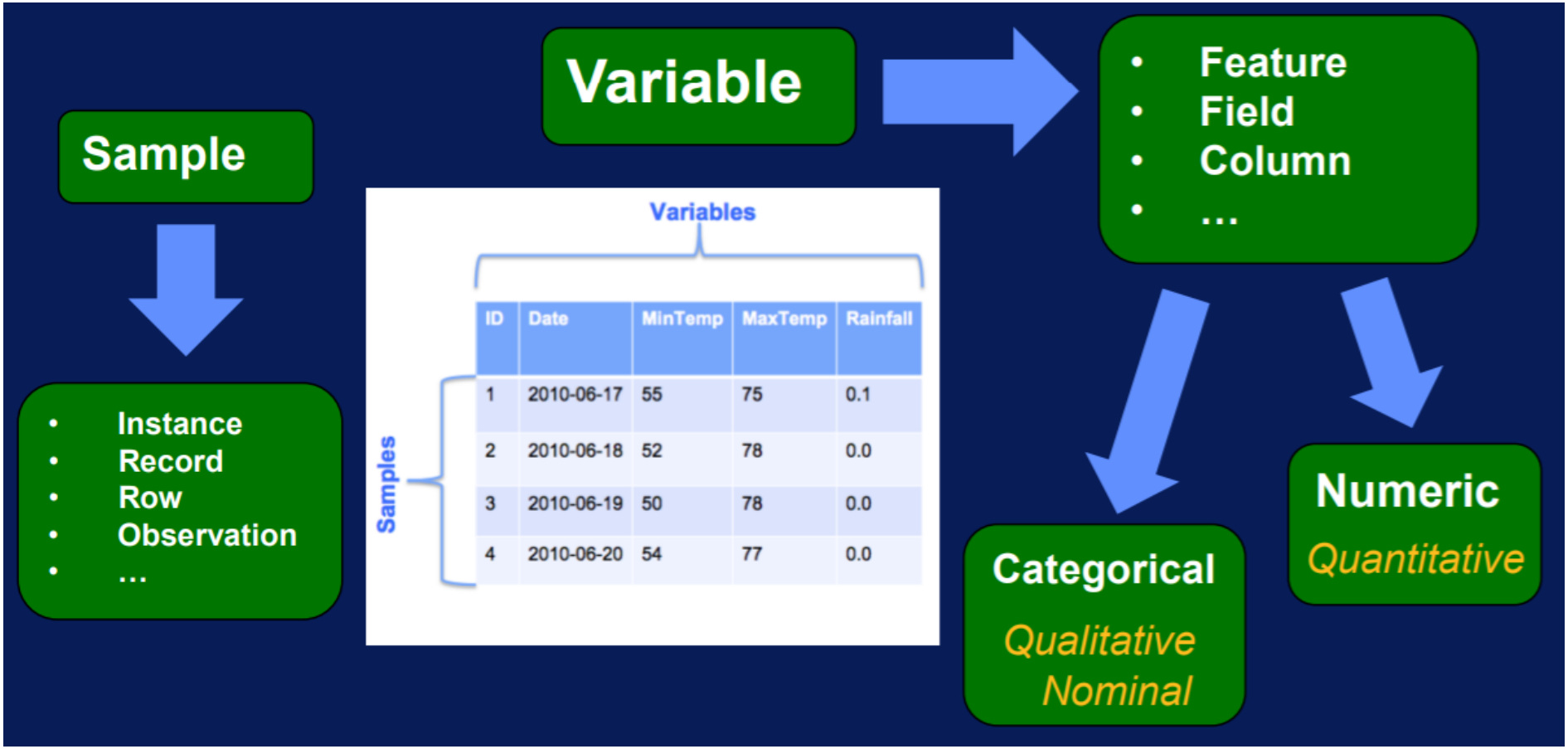
Table: Kolor jest zmienną kategoryczną; Wartości są etykietami

Przykłady zmiennych kategorialnych

- Płeć
- Stan cywilny
- Typ klienta
- Kategorie produktów
- Kolor elementu

Co-funded by the Erasmus+ Programme of the European Union

# Terminologia danych VI



# Eksploracja danych I

Po tej sekcji będziesz mógł...

- Wyjaśnić, dlaczego eksploracja danych jest konieczna
- Przedstawić cele eksploracji danych
- Wymienić kategorie technik eksploracji danych

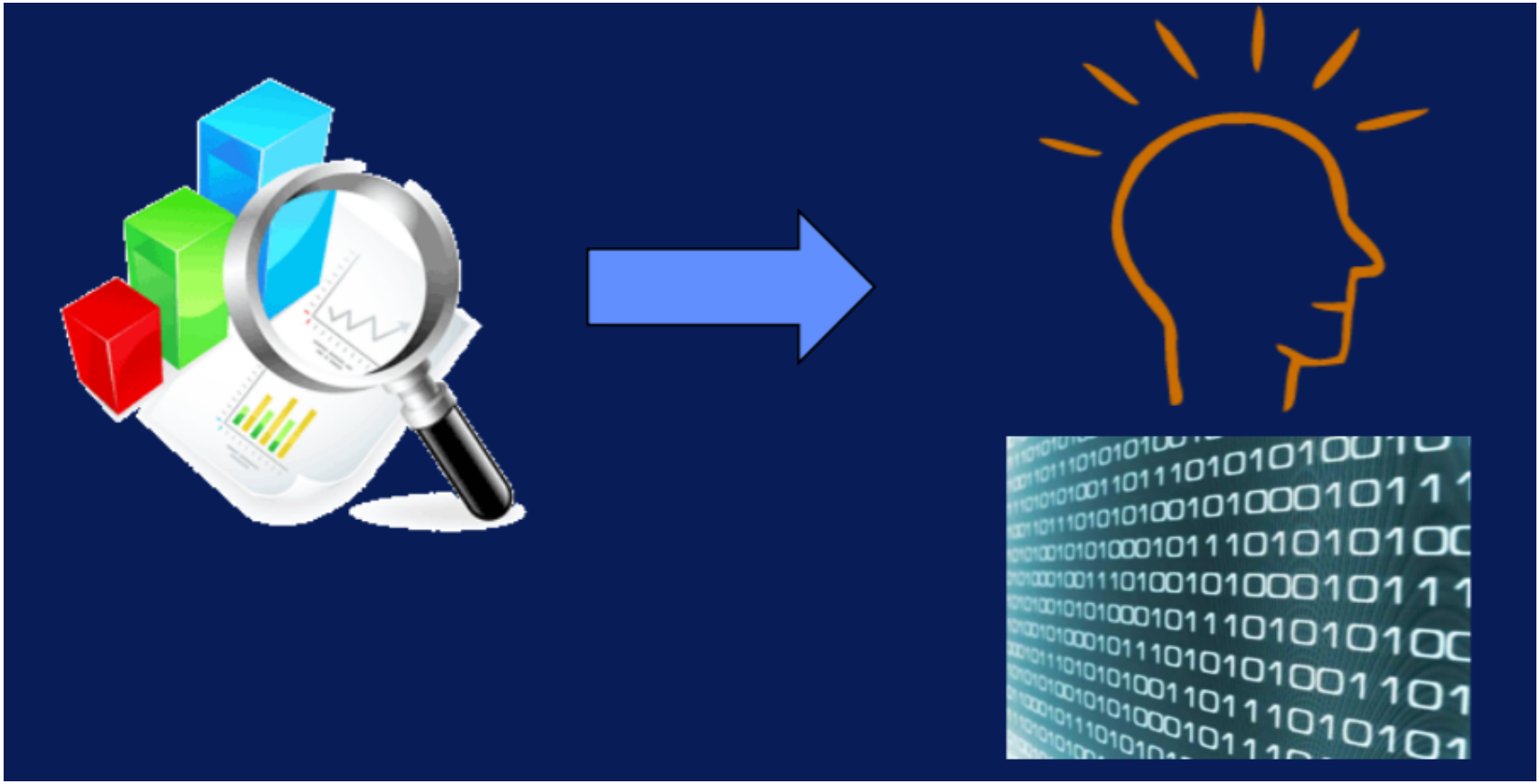
"Experts often possess more data than judgment"
(by Colin Powell)

## Eksploracja Danych

Gdy już zidentyfikujesz pytania, na które próbujesz uzyskać odpowiedź, i zdobędziesz pewne dane, możesz ulec pokusie, aby od razu zanurkować i zacząć budować modele i uzyskiwać odpowiedzi. Musisz jednak oprzeć się tej pokusie. Pierwszym krokiem powinna być eksploracyjna analiza danych.

## Dlaczego eksplorować dane?

Cel: Aby zrozumieć swoje dane

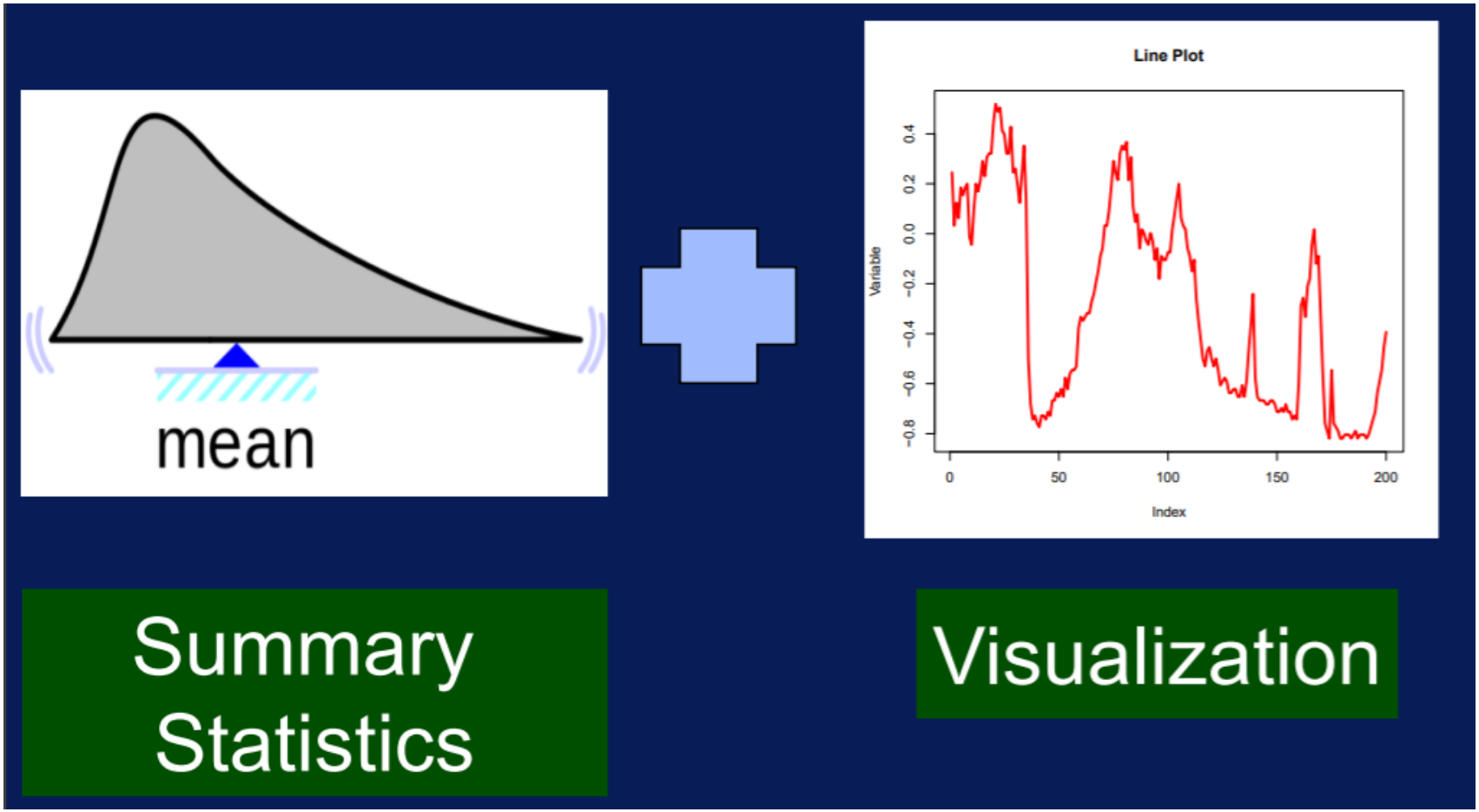


Eksploracyjna Analiza Danych (EDA)

Co-funded by the Erasmus+ Programme of the European Union

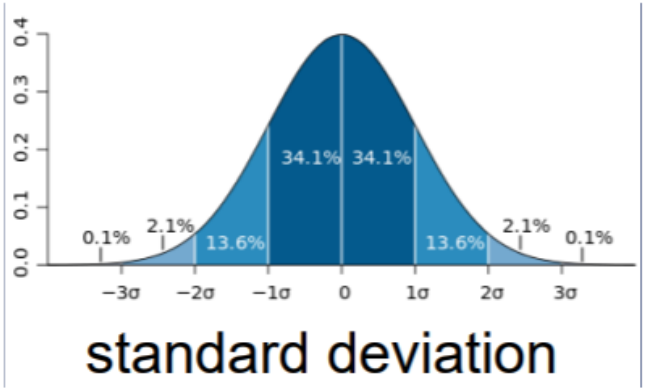
# Eksploracja danych III

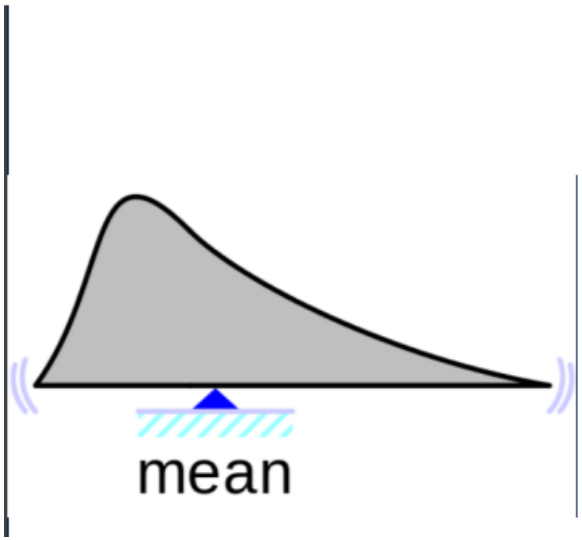
## Sposoby eksploracji danych

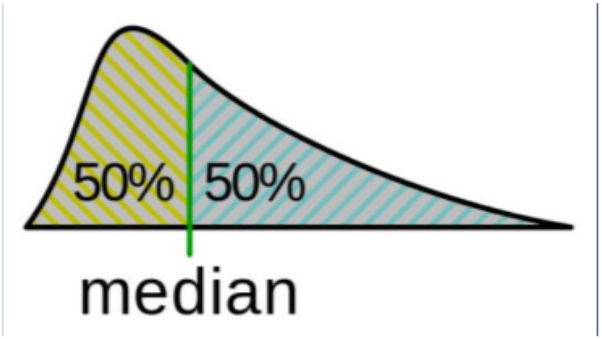


## Statystyki podsumowujące

- Informacje podsumowujące zbiór danych



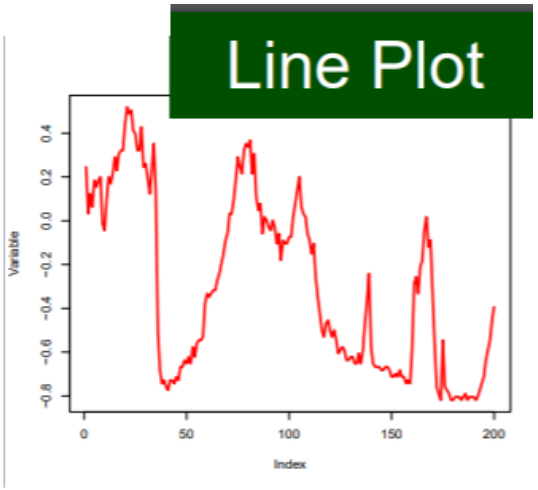




Co-funded by the Erasmus+ Programme of the European Union

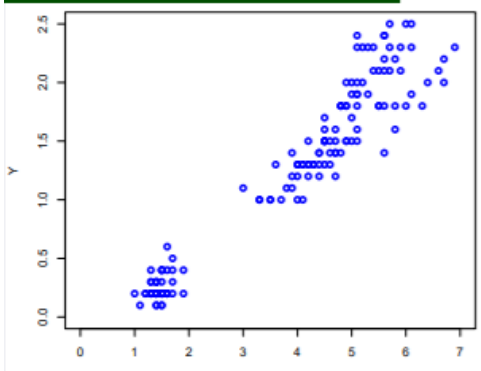
## Wizualizacja danych

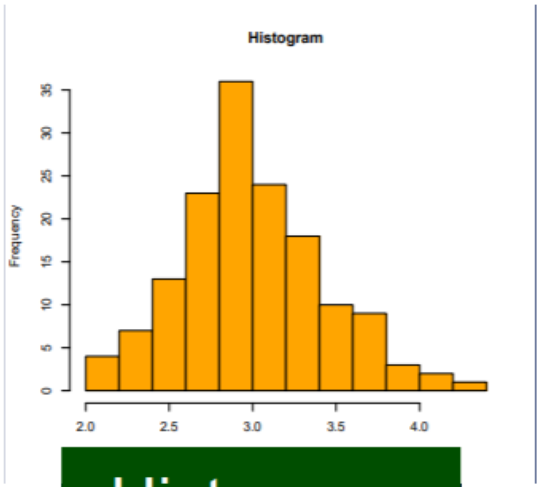
- Spójrz na dane graficznie

Line Plot




Scatter Plot



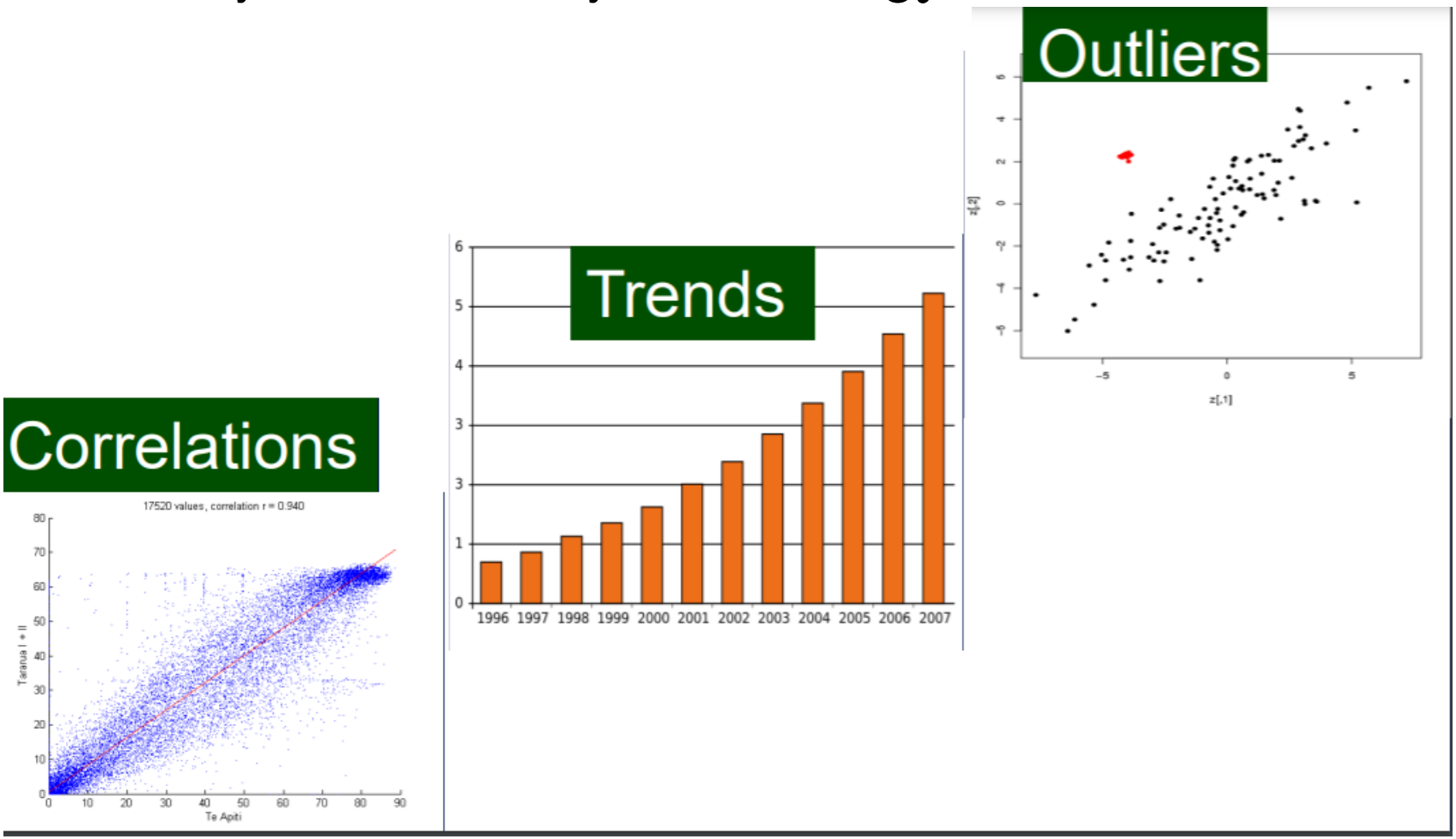




Histogram

Co-funded by the
Erasmus+ Programme
of the European Union

# Eksploracja danych VI

Kilka rzeczy, na które należy zwrócić uwagę

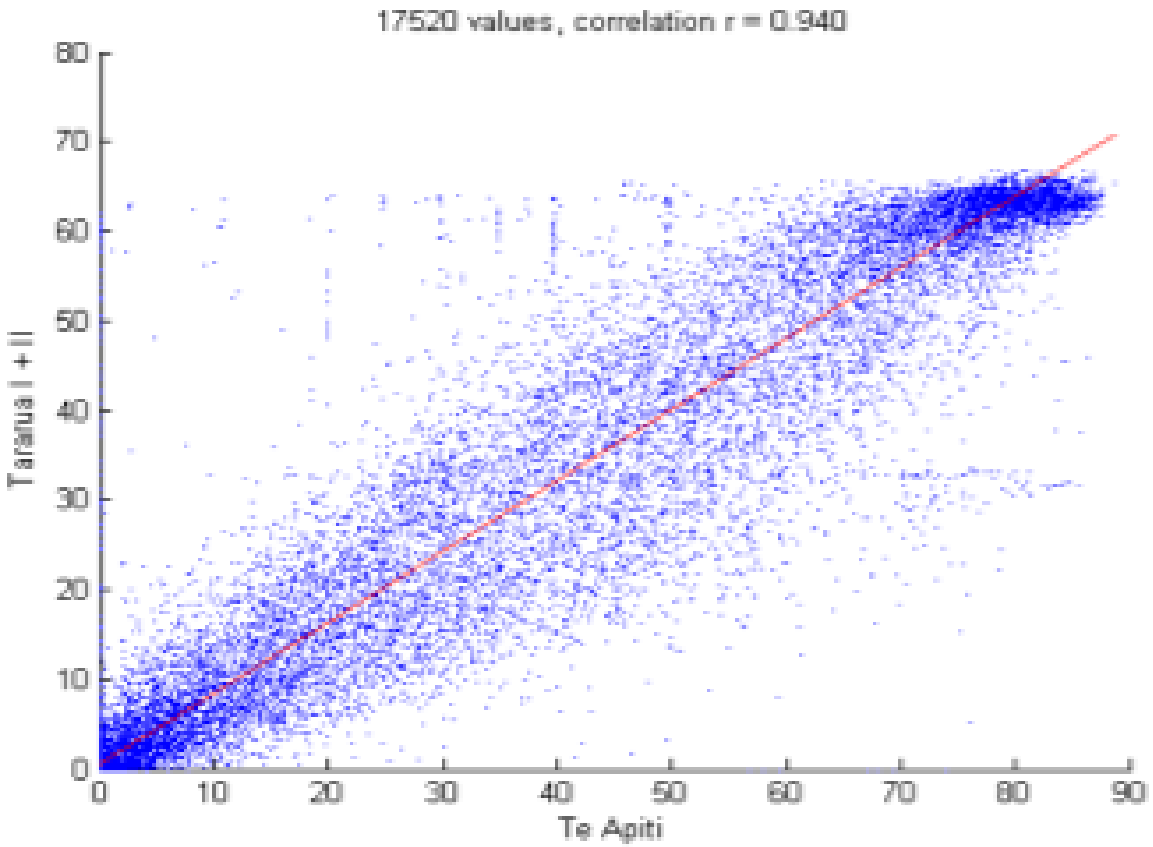



Co-funded by the Erasmus+ Programme of the European Union

# Eksploracja danych VII

## Korelacje

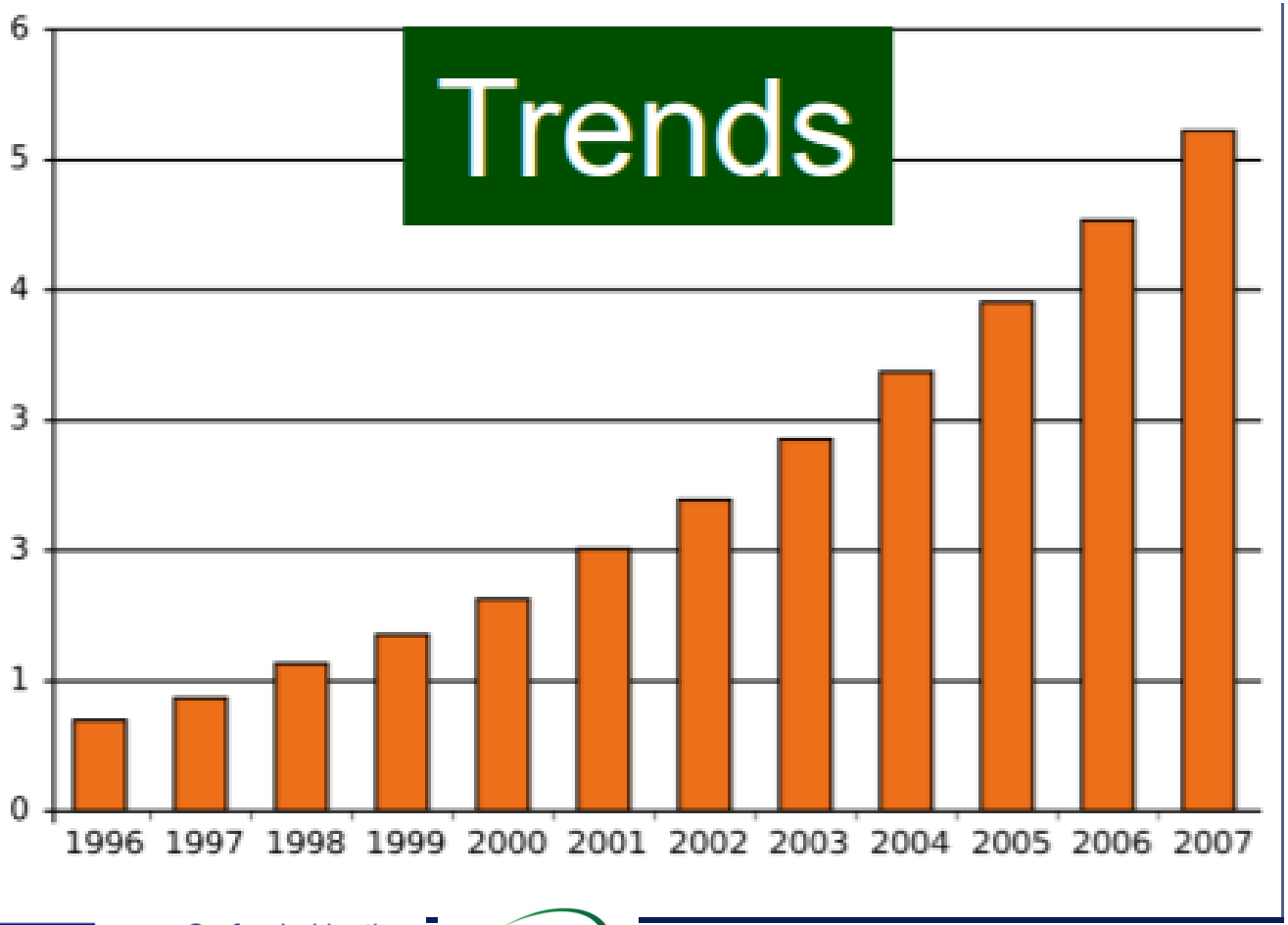
- Podają informacje o relacji między zmiennymi




Co-funded by the
Erasmus+ Programme
of the European Union

## Trendy

- Wskazują ogólną charakterystykę danych

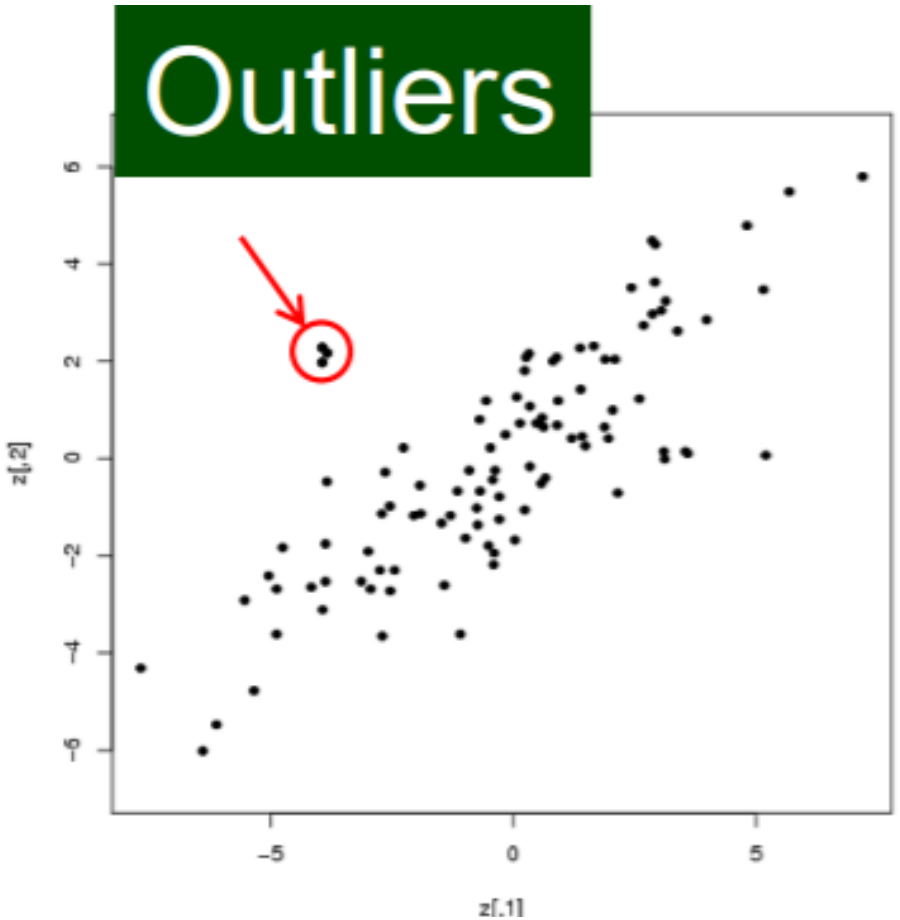



Co-funded by the
Erasmus+ Programme
of the European Union

# Eksploracja danych IX

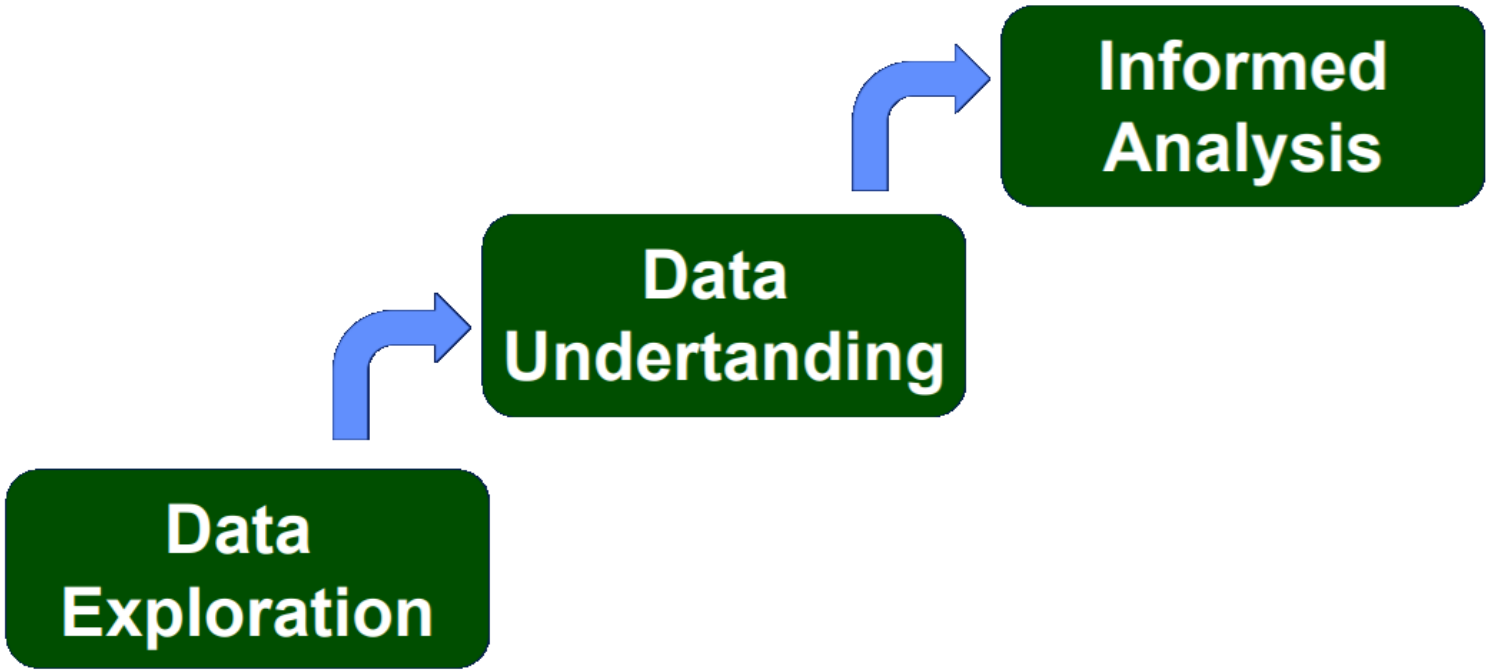
## (Outliers) Odstające

- Wskazują potencjalne problemy z danymi




Co-funded by the
Erasmus+ Programme
of the European Union

Eksploracja danych




Co-funded by the Erasmus+ Programme of the European Union

# Data Exploration through Summary Statistics I

Po tej sekcji będziesz mógł...

- Zdefiniować, czym jest statystyka podsumowująca
- Wymienić trzy popularne statystyki podsumowujące
- Wyjaśnić, w jaki sposób statystyki podsumowujące są przydatne w badaniu danych

## Definition

Co to są statystyki podsumowujące?

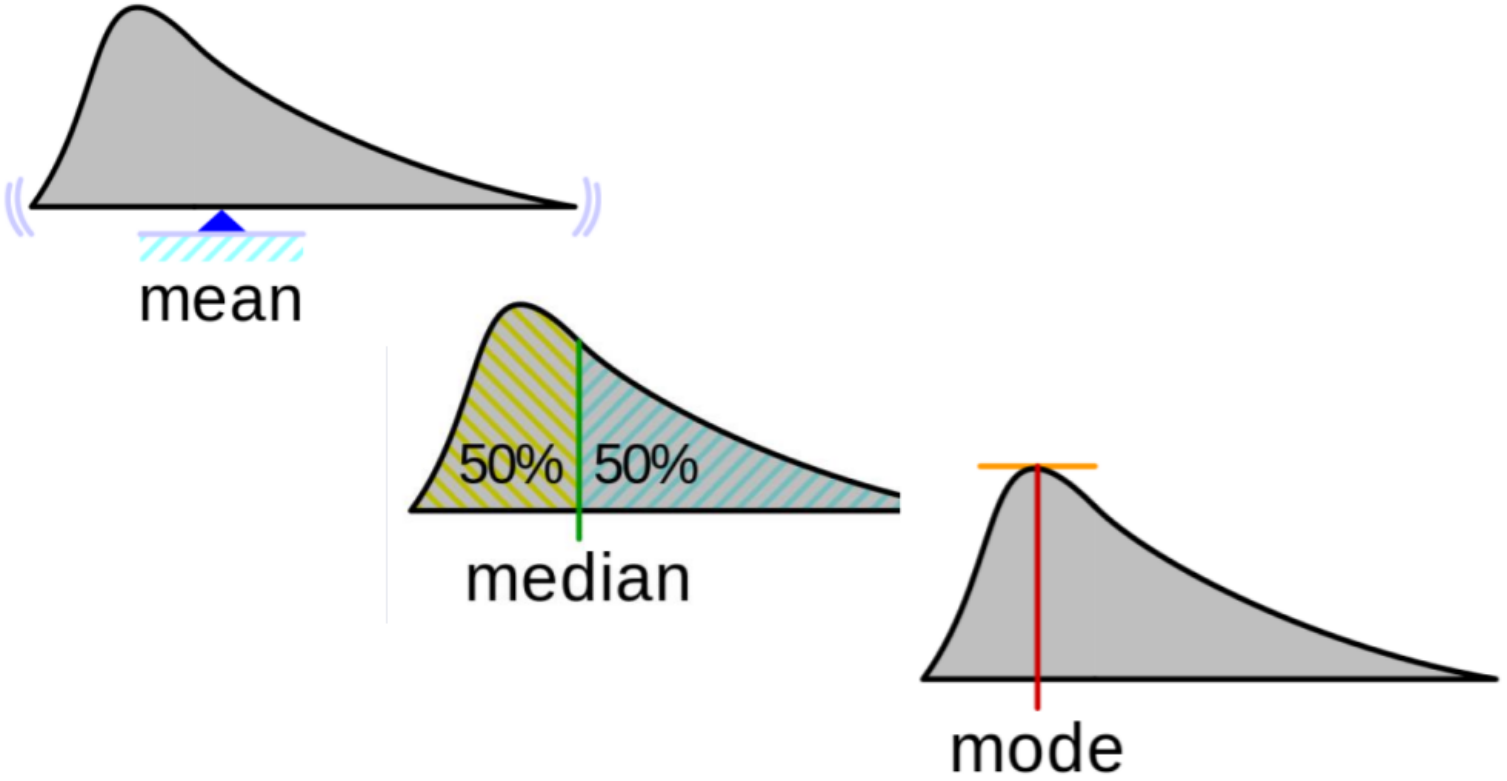
- Ilości podsumowujące i opisujące zbiór wartości danych
- Miary
  - Lokalizacja: średnia, mediana
  - Rozrzut: odchylenie standardowe
  - Kształt: skośność

Co-funded by the Erasmus+ Programme of the European Union

# Data Exploration through Summary Statistics II

## Miary lokalizacji

Opisują centralną lub typową wartość zbioru danych




Co-funded by the Erasmus+ Programme of the European Union

# Data Exploration through Summary Statistics III

## Miry lokalizacji - przykład

| Wiek | Wiek (posortowane) |
|---|---|
| 35 | 21 |
| 42 | 22 |
| 78 | 35 |
| 22 | 42 |
| 56 | 42 |
| 50 | 50 |
| 42 | 56 |
| 78 | 78 |
| 21 | 78 |
| 87 | 87 |

Mean = 51.1 Median = (42+50)/2 = 46 Mode = 42 & 78

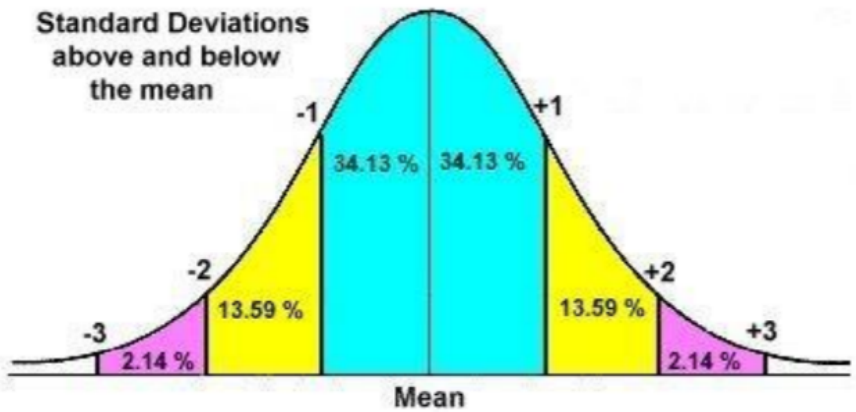
Co-funded by the Erasmus+ Programme of the European Union

# Data Exploration through Summary Statistics IV

## Miary rozprzestrzeniania się

Opisz, jak rozproszone lub zróżnicowane są dane

- minimum
- maximum
- range
- standard deviation
- variation




Co-funded by the Erasmus+ Programme of the European Union

# Data Exploration through Summary Statistics V

## Miary rozprzestrzeniania się – Przykład

Range = 87 – 21 = 66

Variance = 548.767

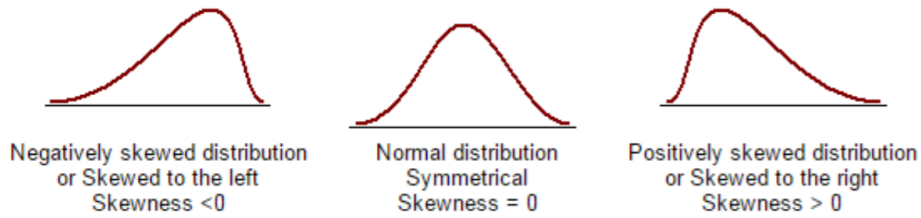
Standard deviation = 23.426

Co-funded by the Erasmus+ Programme of the European Union

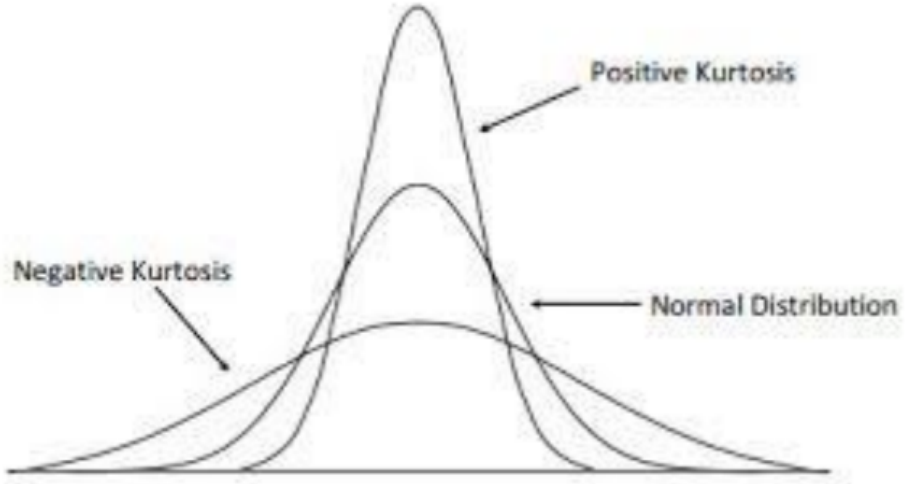
# Data Exploration through Summary Statistics VI

## Miary kształtu

skewness (skośność)




kurtosis (kurtoza)



# Data Exploration through Summary Statistics VII

## Miary kształtu - Przykład

| Age |
|---|
| 35 |
| 42 |
| 78 |
| 22 |
| 56 |
| 50 |
| 42 |
| 78 |
| 21 |
| 87 |

Skewness = 0.2995

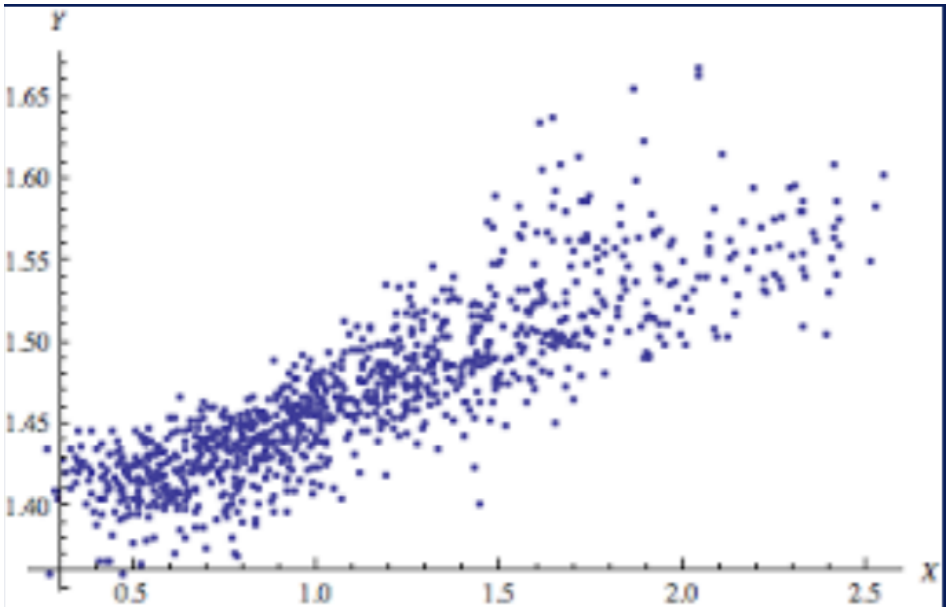
Kurtosis = -1.2028

Co-funded by the Erasmus+ Programme of the European Union

# Data Exploration through Summary Statistics VIII

## Miary zależności

Opisują zależność pomiędzy zmiennymi




korelacja

Co-funded by the
Erasmus+ Programme
of the European Union

# Data Exploration through Summary Statistics IX

## Miary zależności – Przykład

| Height | Weight |
|---|---|
| 180 | 68 |
| 153 | 70 |
| 204 | 84 |
| 133 | 44 |
| 208 | 81 |
| 142 | 53 |
| 122 | 40 |
| 168 | 50 |
| 175 | 64 |
| 200 | 72 |

Correlation = 0.8906

Co-funded by the Erasmus+ Programme of the European Union

# Data Exploration through Summary Statistics X

Statystyki dotyczące zmiennych kategorialnych

Opisują liczbę kategorii i częstotliwość każdej kategorii

| Color/Pet | White | Brown | Black | Orange | Total |
|---|---|---|---|---|---|
| Dog | 34 | 44 | 32 | 0 | 110 |
| Cat | 25 | 2 | 43 | 0 | 70 |
| Fish | 1 | 0 | 5 | 33 | 39 |
| **Total** | 60 | 46 | 80 | 33 | 219 |

contingency table (tablica kontyngencji)

Co-funded by the Erasmus+ Programme of the European Union

# Eksploracja danych poprzez wykresy I

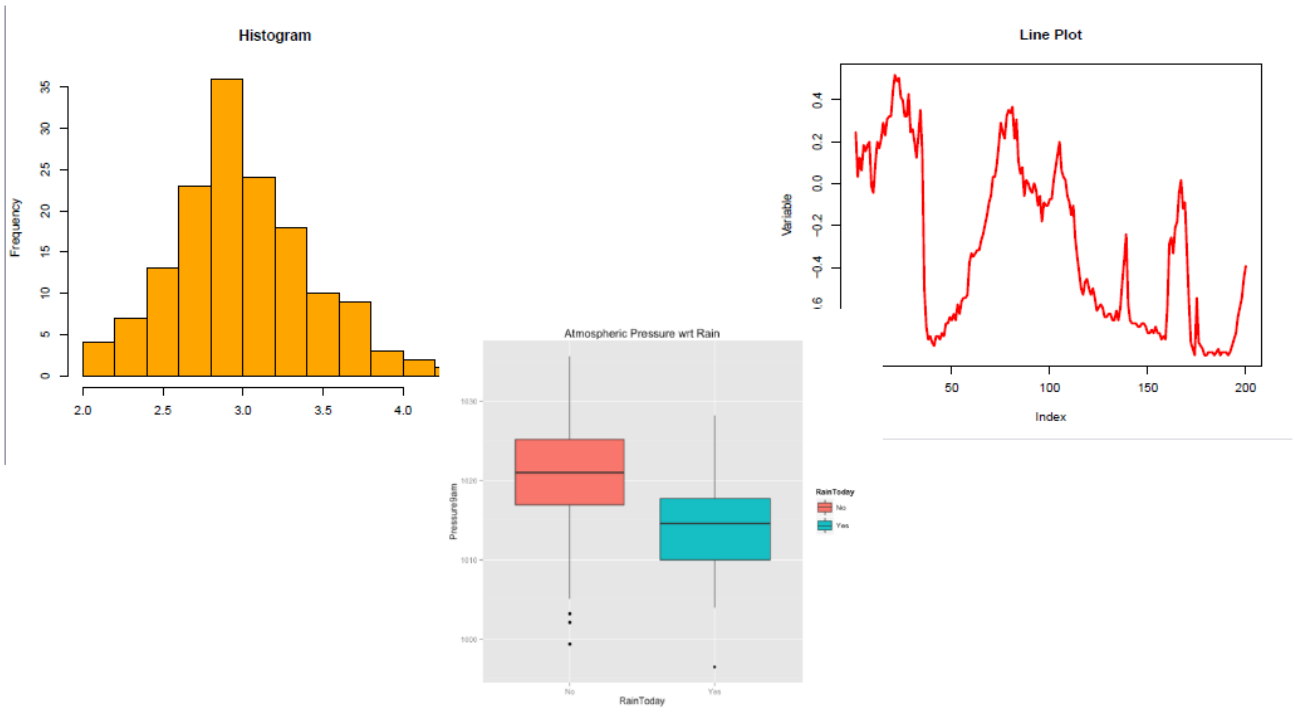
Po tej sekcji będziesz mógł ...

- omówić, w jaki sposób wykresy mogą być przydatne w badaniu danych
- opisać, jak wykorzystałbyś wykres punktowy
- podsumować, co pokazuje wykres pudełkowy

Co-funded by the
Erasmus+ Programme
of the European Union

# Eksploracja danych poprzez wykresy II

## Wizualizacja danych



# Eksploracja danych poprzez wykresy III

Typy wykresów

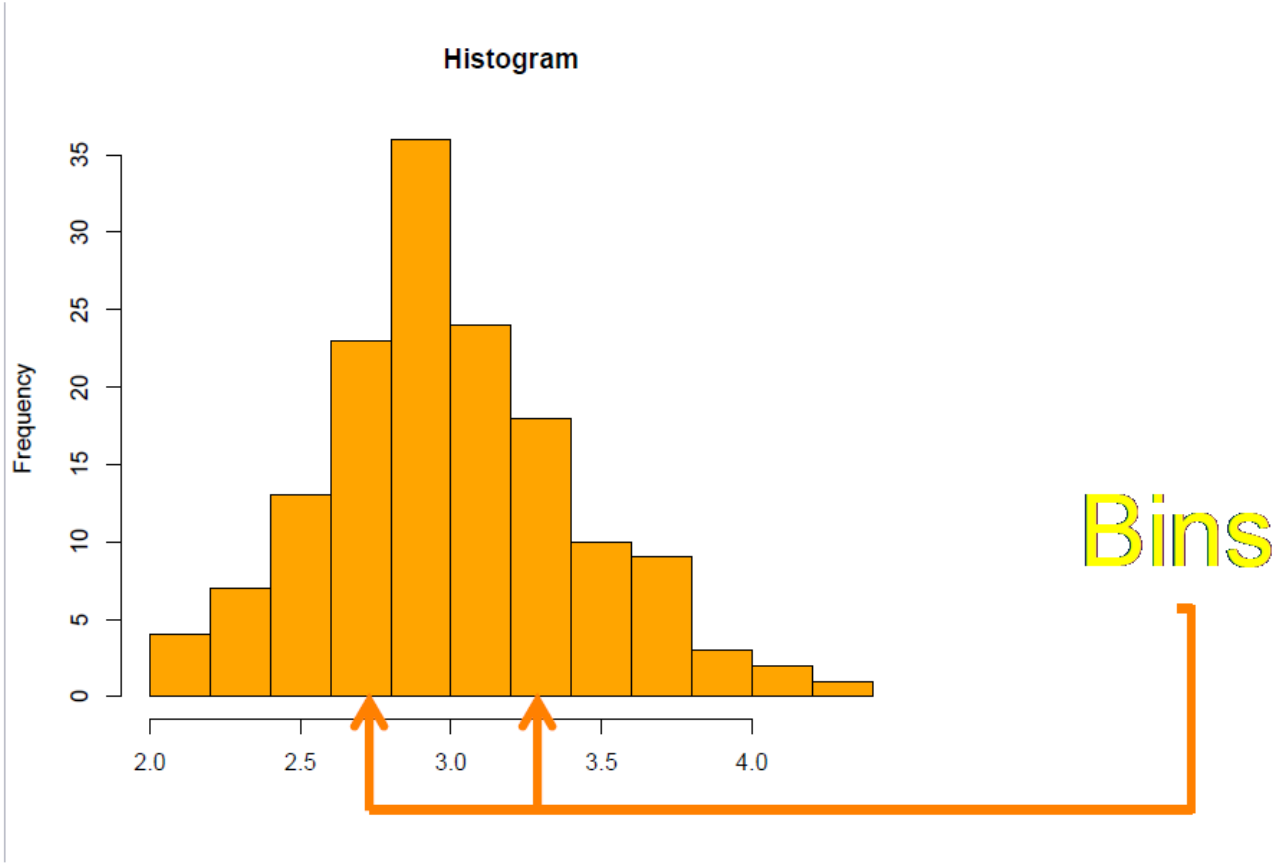
- Histogram
- Line plot
- Scatter plot (punktowy)
- Bar plot (słupkowy)
- Box plot (pudełkowy)
- inne

Co-funded by the
Erasmus+ Programme
of the European Union

# Eksploracja danych poprzez wykresy IV

## Histogram

pokazuje rozkład zmiennej numerycznej




Co-funded by the Erasmus+ Programme of the European Union

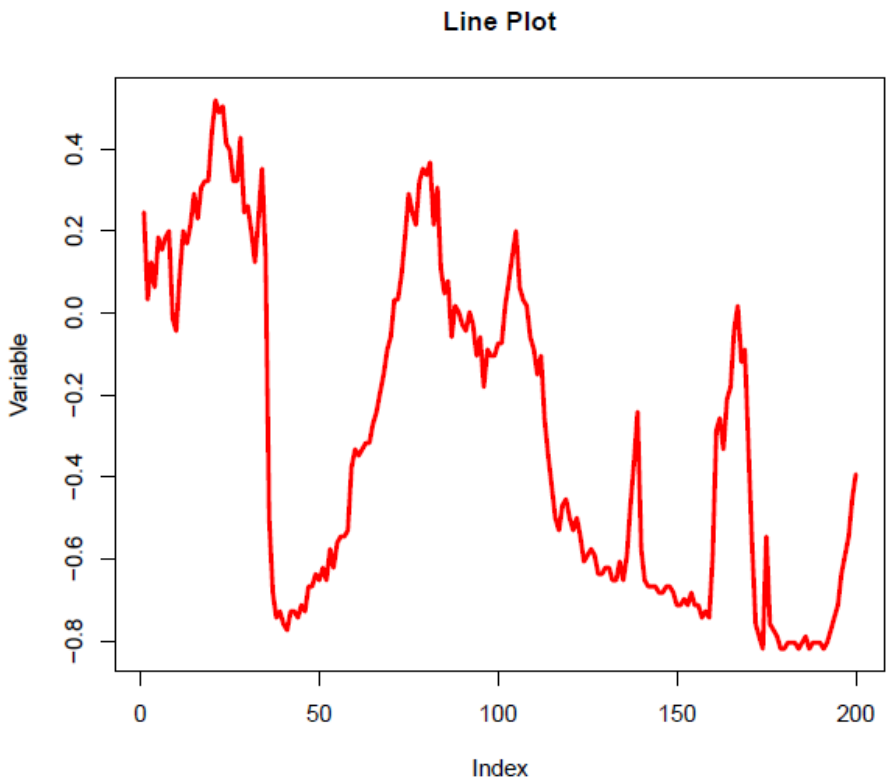
# Eksploracja danych poprzez wykresy V

## Co pokazuje histogram

# Eksploracja danych poprzez wykresy VI

## Wykres liniowy

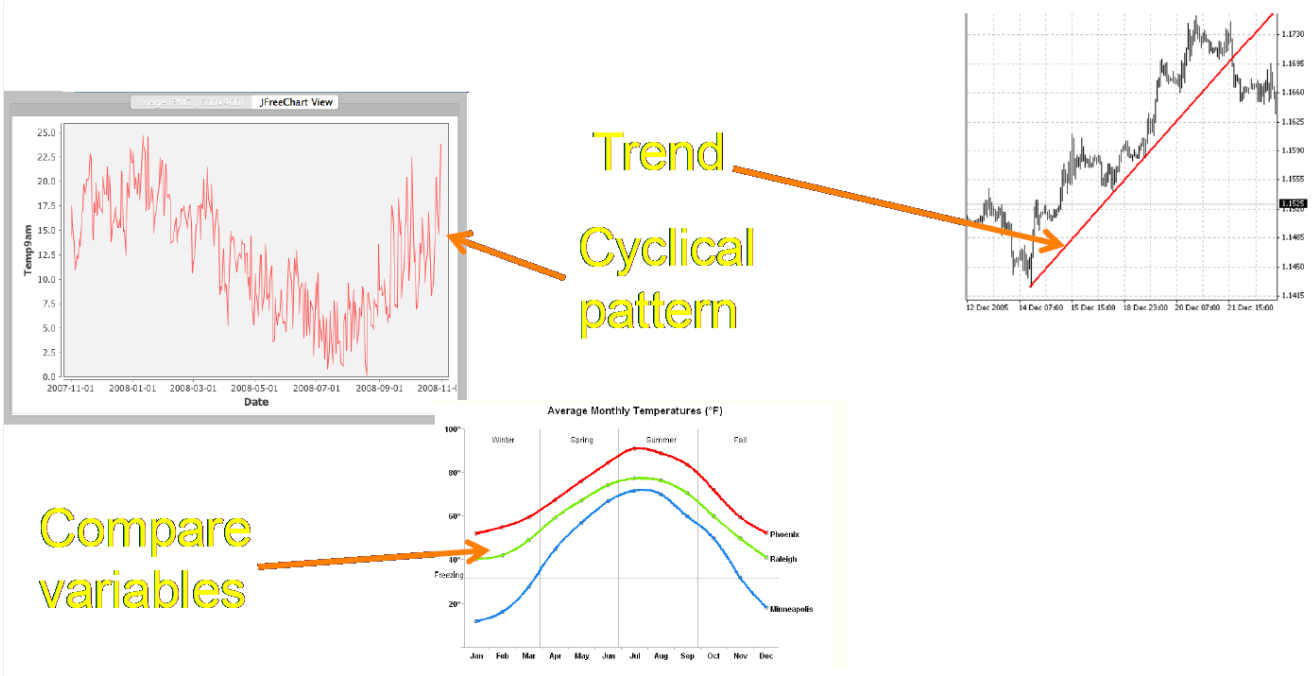
Pokazuje zmiany danych w czasie




Co-funded by the
Erasmus+ Programme
of the European Union

# Eksploracja danych poprzez wykresy VII

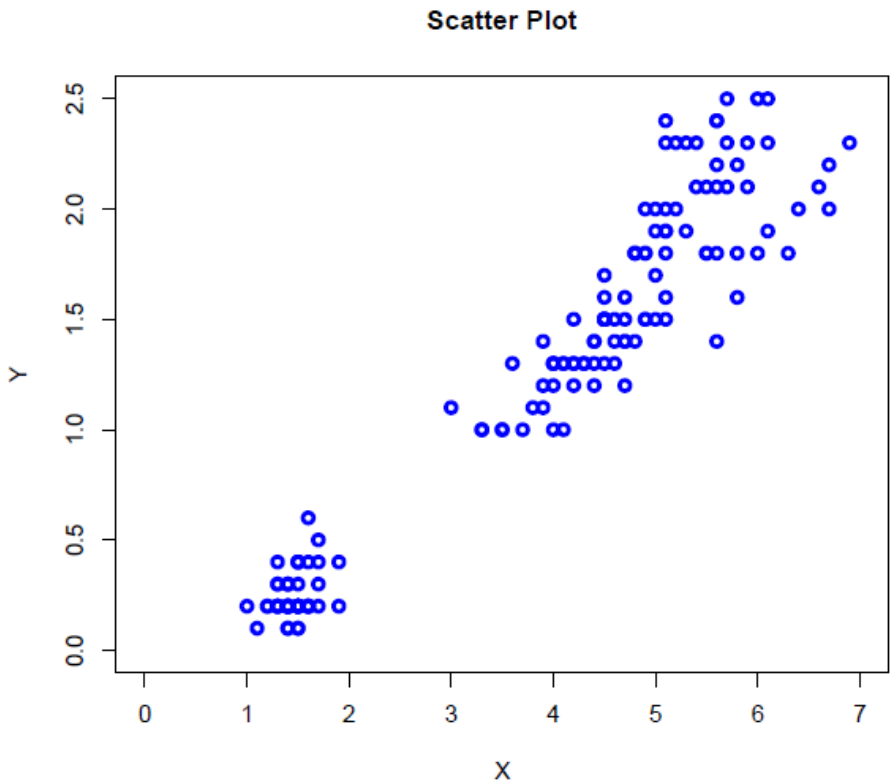
## Co pokazuje wykres liniowy



# Eksploracja danych poprzez wykresy VIII

## Wykres punktowy

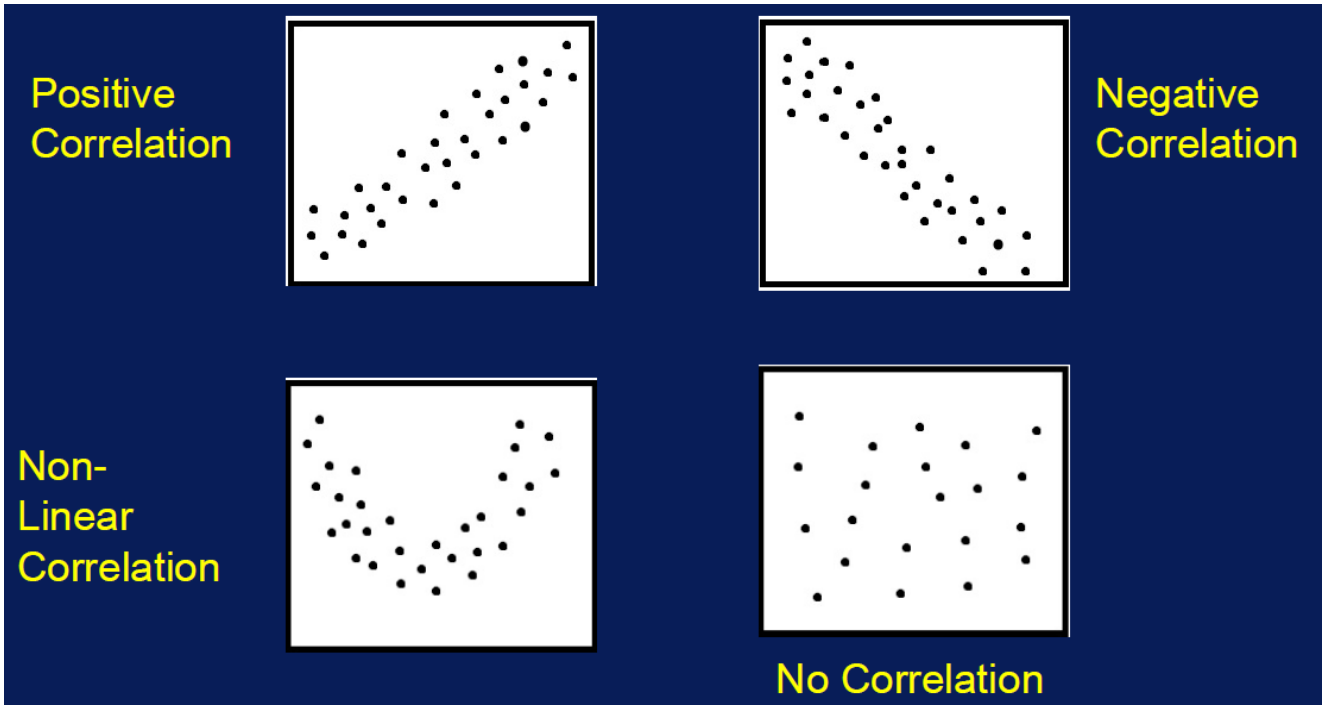
Pokazuje związek między dwiema zmiennymi




Co-funded by the Erasmus+ Programme of the European Union

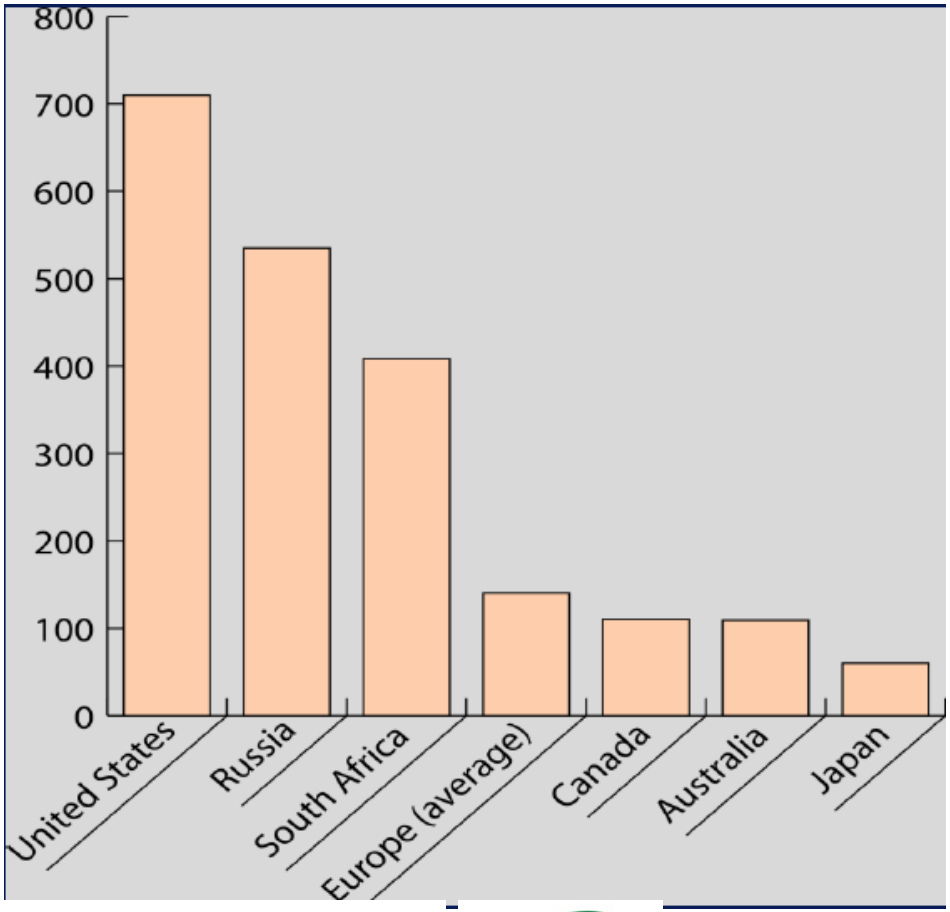
# Eksploracja danych poprzez wykresy IX

Co pokazuje wykres punktowy?



## Wykres słupkowy (Bar plot)

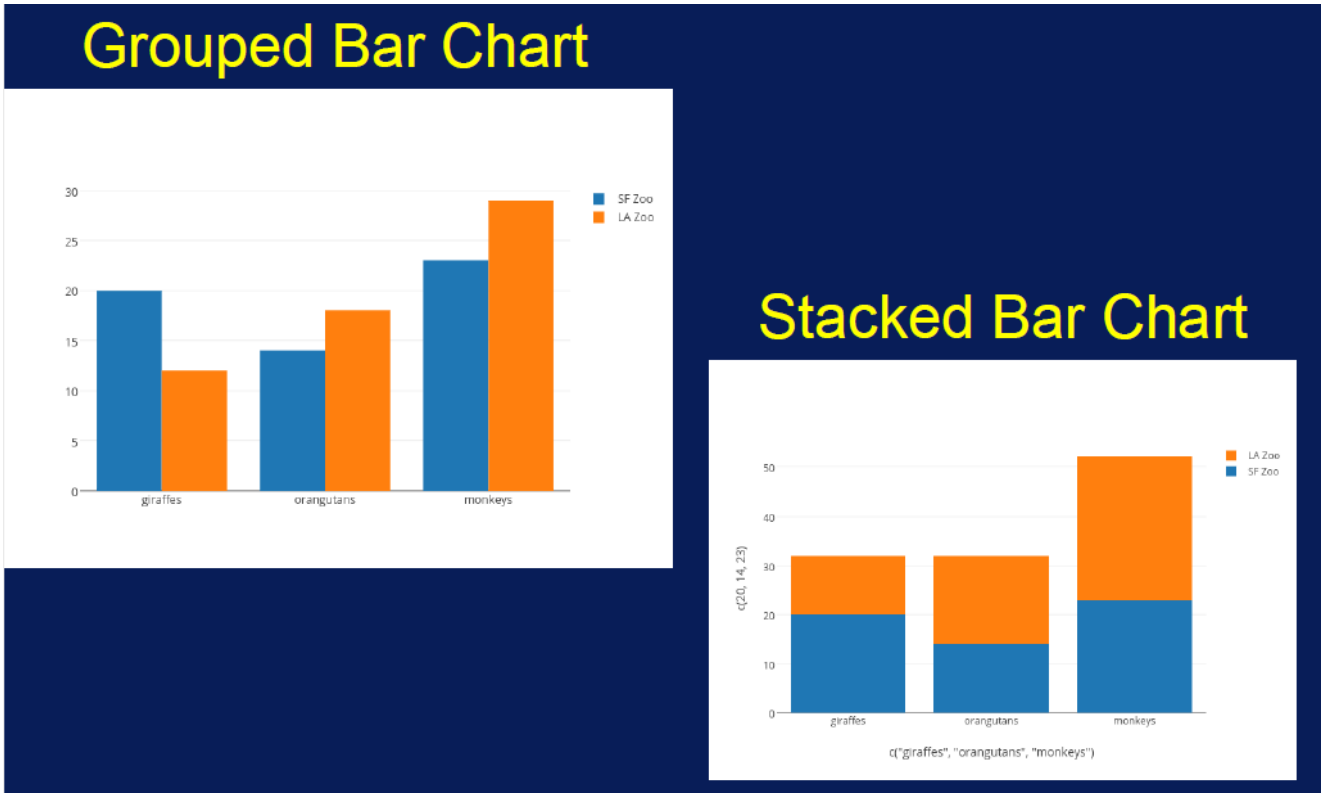
Pokazuje rozkład zmiennej kategorialnej




Co-funded by the Erasmus+ Programme of the European Union

# Eksploracja danych poprzez wykresy XI

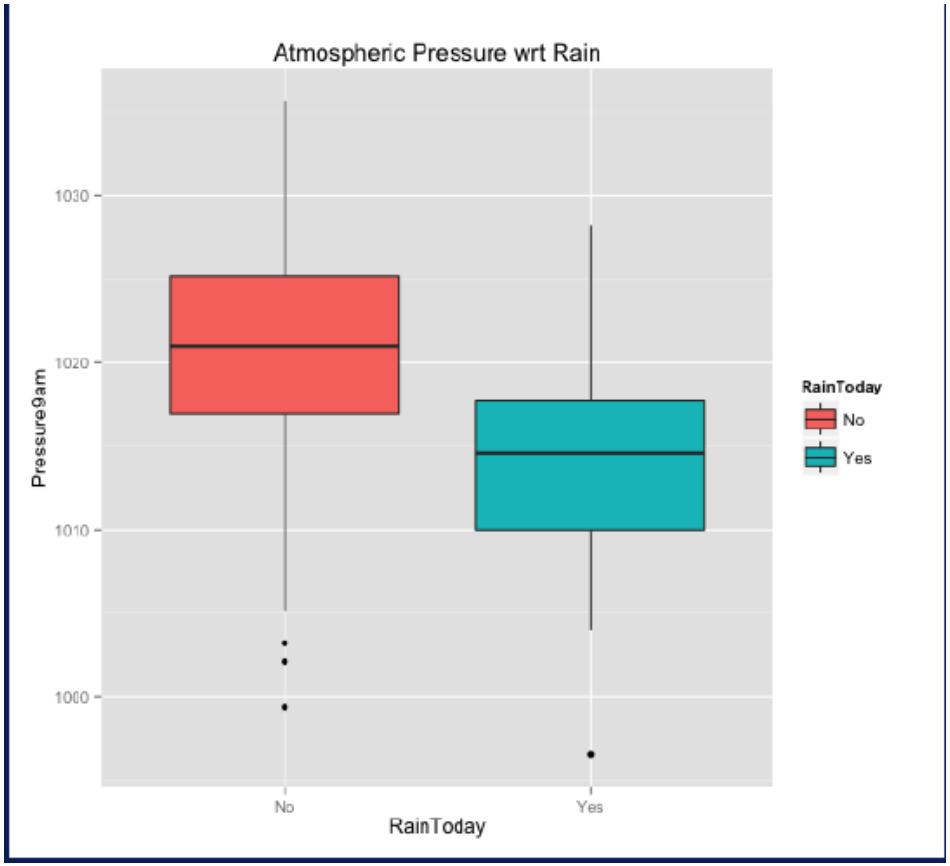
Co pokazuje wykres słupkowy



# Eksploracja danych poprzez wykresy XII

## Wykres pudełkowy (Box plot)

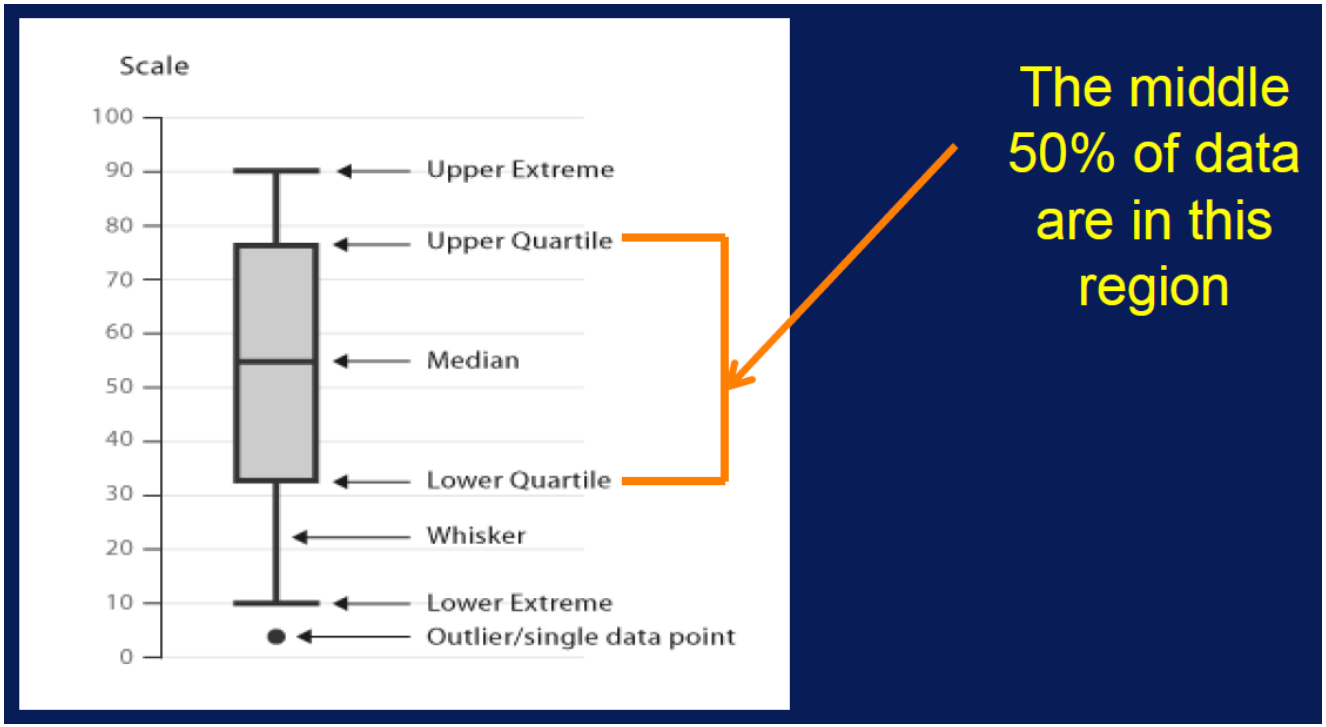
Porównuje rozkłady zmiennych




Co-funded by the Erasmus+ Programme of the European Union

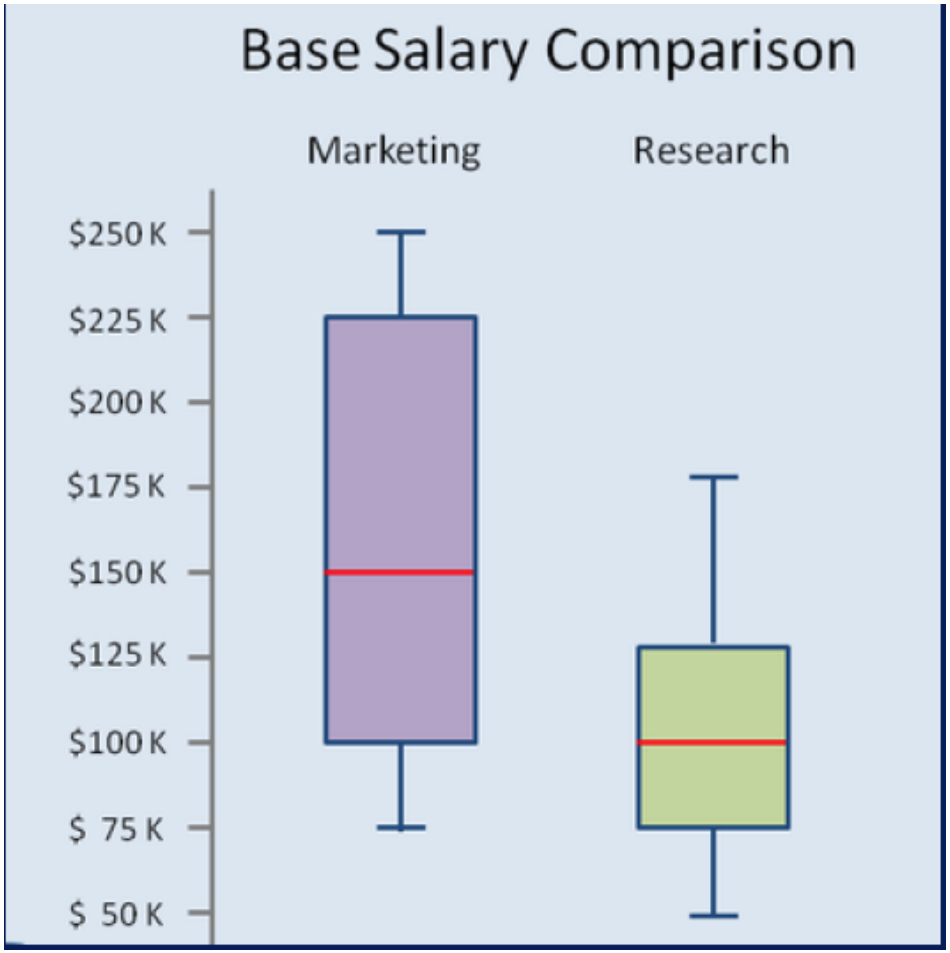
# Eksploracja danych poprzez wykresy XIII

## Składniki wykresu pudełkowego



# Eksploracja danych poprzez wykresy XIV

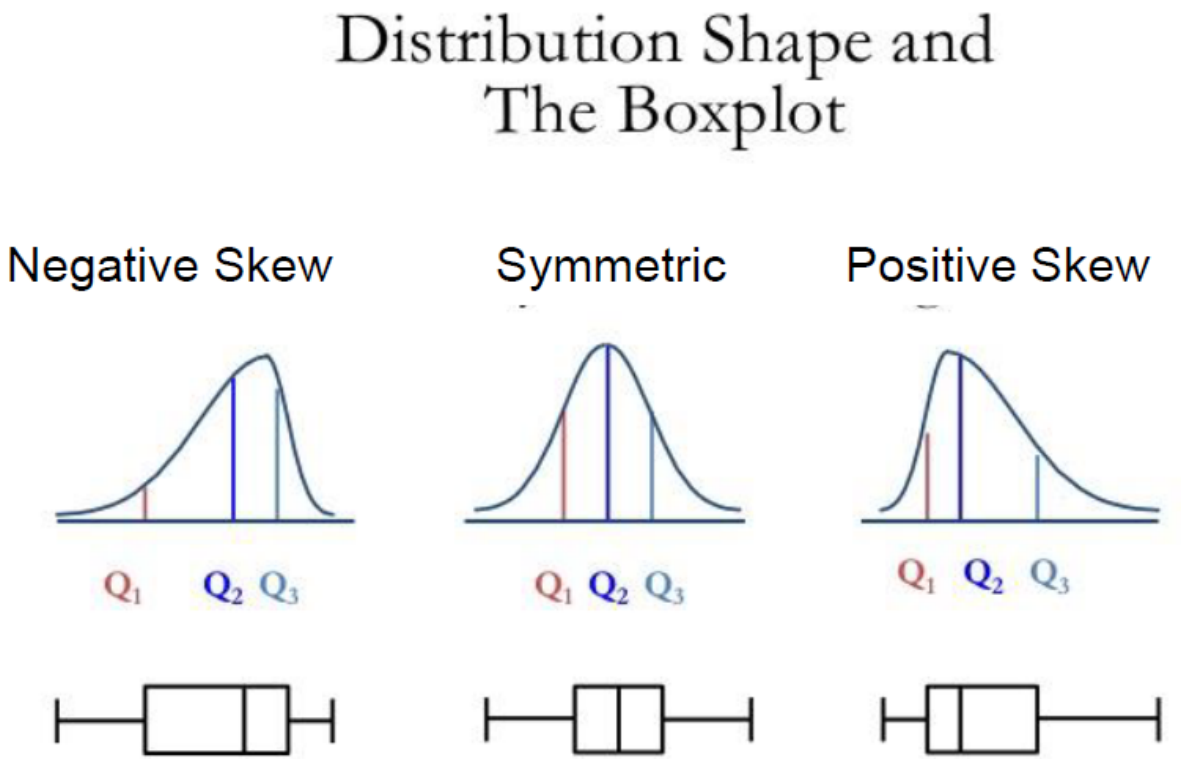
Co pokazuje wykres pudełkowy




Co-funded by the
Erasmus+ Programme
of the European Union

# Eksploracja danych poprzez wykresy XV

Co pokazuje wykres pudełkowy




Co-funded by the
Erasmus+ Programme
of the European Union

# Eksploracja danych poprzez wykresy XVI

Wizualizacja danych

- Zapewnia intuicyjny sposób przeglądania danych
- Powininna być używana ze statystykami podsumowującymi do eksploracji danych
- jest również przydatna do komunikowania wyników

Co-funded by the Erasmus+ Programme of the European Union

# Outline



Co-funded by the Erasmus+ Programme of the European Union

# Description of Daily Weather Dataset I

Description of Daily Weather Dataset

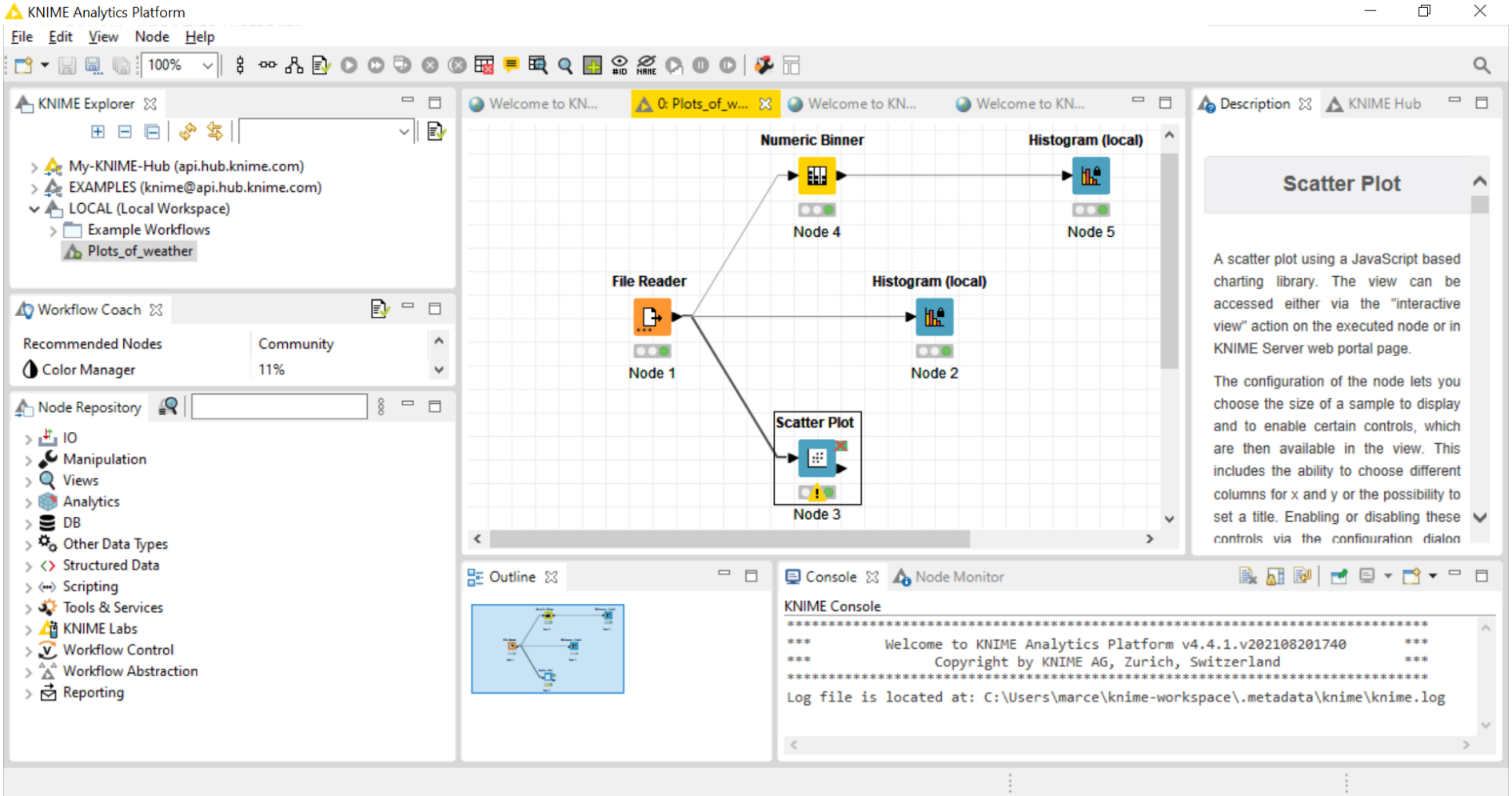
Co-funded by the
Erasmus+ Programme
of the European Union

# Outline



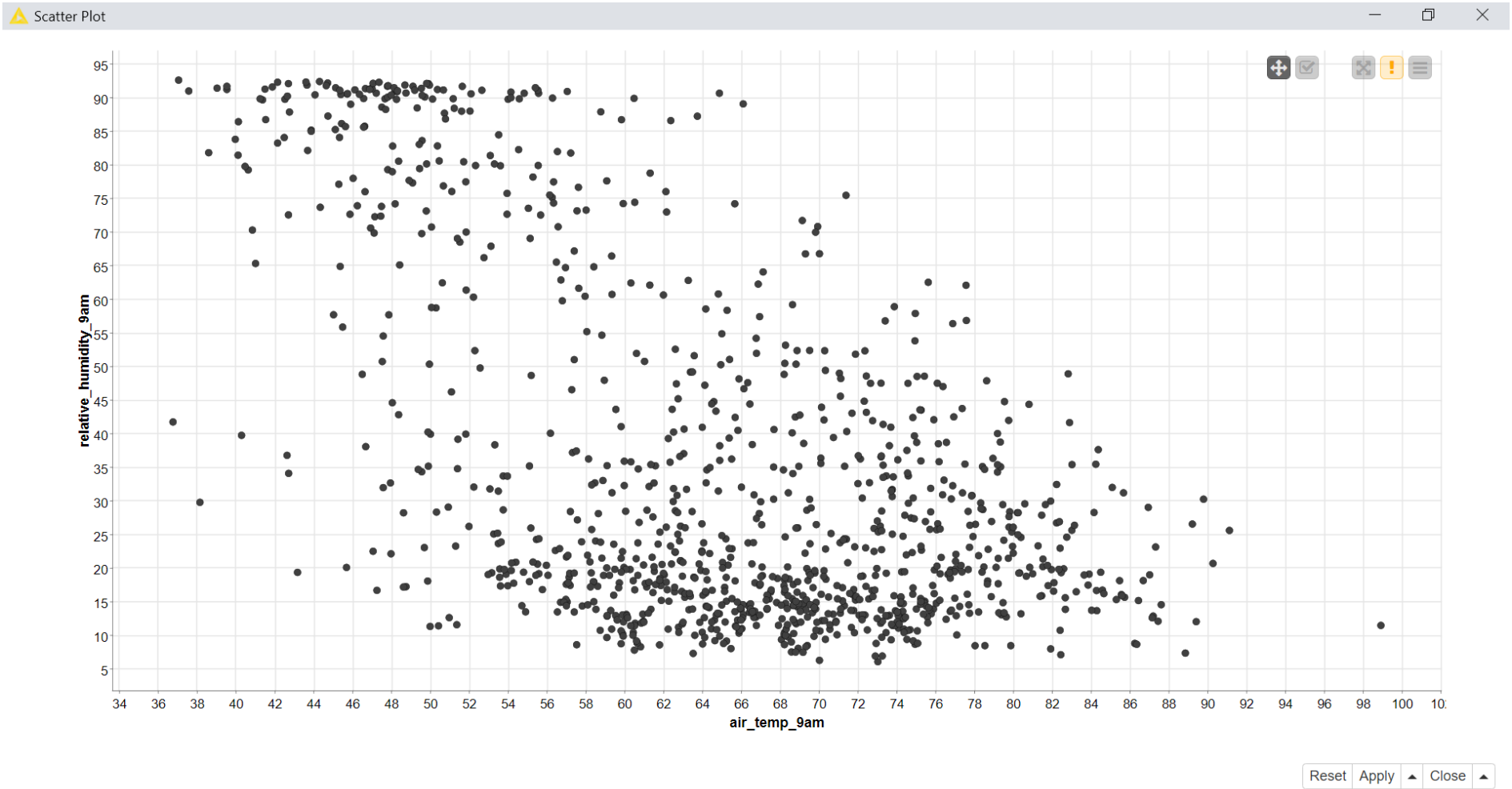
Co-funded by the Erasmus+ Programme of the European Union

# Exploring Data with KNIME Plots I



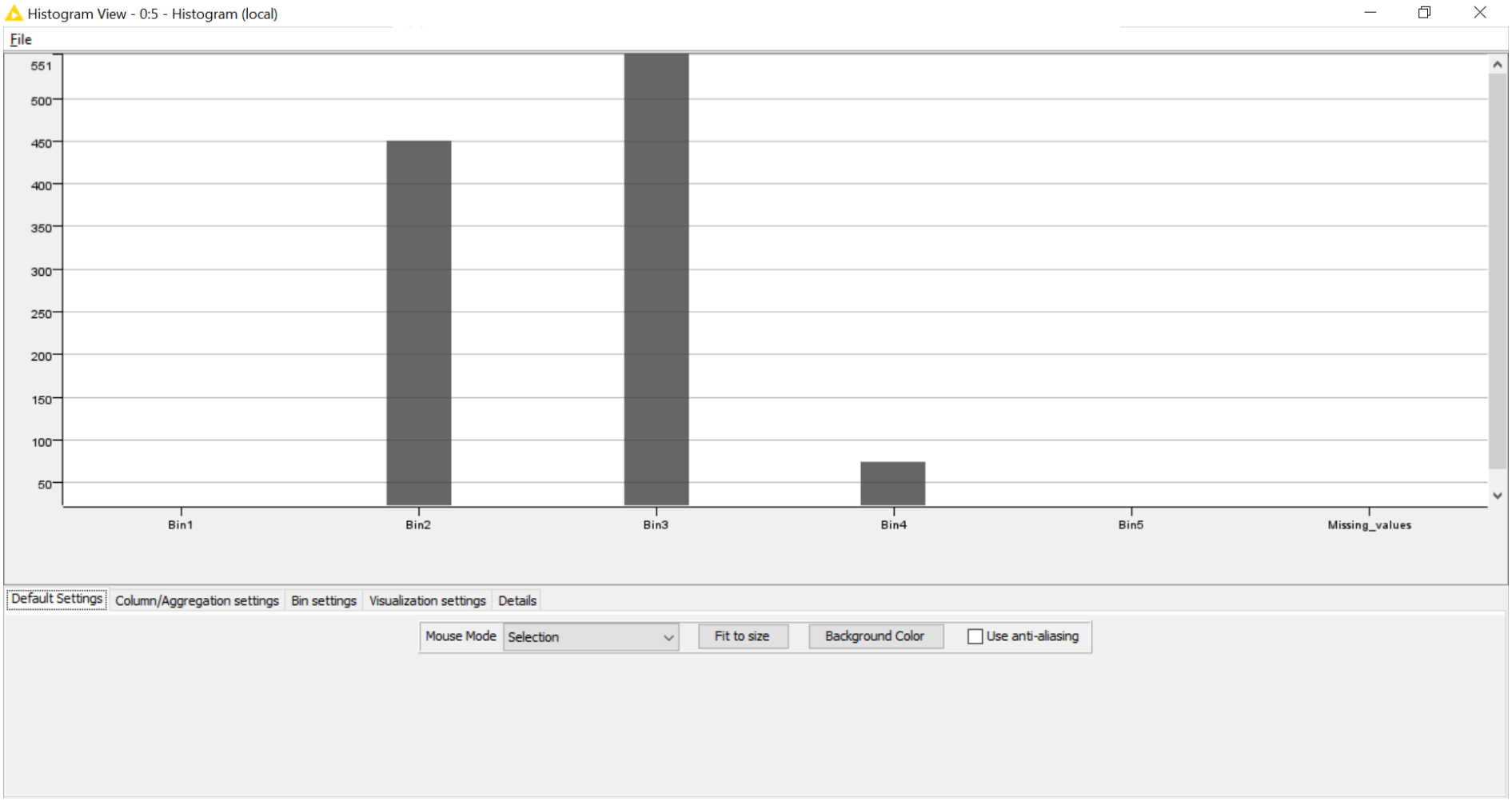
# Exploring Data with KNIME Plots II

## Scatter plot (interactive view)



# Exploring Data with KNIME Plots III

## Histogram



# Outline



Co-funded by the Erasmus+ Programme of the European Union

# Eksploracja danych w pandas

http://localhost:8889/lab/tree/OneDrive/Documents/PythonProjects/Tutorials/Jupyter_Polish_python_lessons-main/Wyklad3/Wyklad3_Exploratory_data_Analysis.ipynb

# Outline



Co-funded by the Erasmus+ Programme of the European Union

# Data Exploration in Spark I

In order to install pyspark you need to run:

```
! pip install pyspark

# Import PySpark
import pyspark
from pyspark.sql import SparkSession

#Create SparkSession
spark = SparkSession.builder.master("local[1]")
        .appName("SparkByExamples.com").getOrCreate()
sc=spark.sparkContext

from pyspark.sql import SQLContext

# creating sqlContext
[                    xt(sc)
```

# Data Exploration in Spark II

```
creating spark dataframe
df = sqlContext.read.load('file:///C:/Users/marce/OneDrive
/Documents/PythonProjects/Tutorials/DataExploration
/coursera-master/coursera-master/big-data-4
/daily_weather.csv',
format = 'com.databricks.spark.csv',
header = 'true', inferSchema='true')

# diplay the names of columns
df.columns

# display scheme of data (types of data and hierarchy)
df.printSchema()

# display summary statistics
df.describe().toPandas().transpose()
```

# Data Exploration in Spark III

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **summary** | count | mean | stddev | min | max |
| **number** | 1095 | 547.0 | 316.24357700987383 | 0 | 1094 |
| **air_pressure_9am** | 1092 | 918.8825513138094 | 3.184161180386833 | 907.9900000000024 | 929.3200000000012 |
| **air_temp_9am** | 1090 | 64.93300141287072 | 11.175514003175877 | 36.752000000000685 | 98.90599999999992 |
| **avg_wind_direction_9am** | 1091 | 142.2355107005759 | 69.13785928889189 | 15.500000000000046 | 343.4 |
| **avg_wind_speed_9am** | 1092 | 5.50828424225493 | 4.5528134655317185 | 0.69345139999974 | 23.554978199999763 |
| **max_wind_direction_9am** | 1092 | 148.95351796516923 | 67.23801294602953 | 28.89999999999991 | 312.19999999999993 |
| **max_wind_speed_9am** | 1091 | 7.019513529175272 | 5.598209170780958 | 1.1855782000000479 | 29.84077959999996 |
| **rain_accumulation_9am** | 1089 | 0.20307895225211126 | 1.5939521253574893 | 0.0 | 24.01999999999907 |
| **rain_duration_9am** | 1092 | 294.1080522756142 | 1598.0787786601481 | 0.0 | 17704.0 |
| **relative_humidity_9am** | 1095 | 34.24140205923536 | 25.472066802250055 | 6.090000000001012 | 92.6200000000002 |
| **relative_humidity_3pm** | 1095 | 35.34472714825898 | 22.524079453587273 | 5.3000000000006855 | 92.2500000000003 |

```
# summary statistics for a column
                                  sure_9am').show()
```

Co-funded by the Erasmus+ Programme of the European Union

# Data Exploration in Spark IV

```
# drop non-available value
df2 = df.na.drop(subset=['air_pressure_9am'])
# compute correlation
df2.stat.corr('rain_accumulation_9am','rain_duration_9am')
```

Co-funded by the
Erasmus+ Programme
of the European Union