Moduł 10
Eksplorowane dane. Eksploracja w rzeczywistych przypadkach użycia

iBigWorld:
Innovations for Big Data in a Real World

UBB team

iBigWorld:2020-1-PL01-KA203-082197

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

Rozważane przypadki użycia w modułach10-12

	Przypadek użycia (link do zasobu)	10	11	12
1	COVID-19 (dane numeryczne, etykiety) (https://github.com/covid19datahub/R)	x		x
2	Historyczne dzienne dane pogodowe 2020 (https://www.kaggle.com/vishalvjoseph/weather-dataset-for-covid19-predictions)	x		x
3	Globalna wymiana danych zdrowotnych (http://ghdx.healthdata.org/) (dane numeryczne, etykiety)	x	x	x
4	Czujniki zanieczyszczenia powietrza (dane geolokalizacyjne)	x		x

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

Use cases under consideration throughout topics 10-12

	Use case (link to dataset)	10	11	12
5	Segmentacja naczyń (zdjęcia) (https://www.idiap.ch/software/bob/docs/bob/bob.db.drive/stable/index.html)		x	
6	Numery domu - Google Street View (SVHN) danych(https://github.com/aditya9211/SVHN-CNN) (obrazy)		x	
7	Fizjonet (sygnały) (physionet.org)		x	
8	Przetwarzanie języka naturalnego (teksty)	x	x	

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

Hub danych (COVID Data Hub)

- Pobieranie danych poprzez URL

```
from pyspark import SparkFiles
spark.sparkContext.addFile('https://storage.covid19datahub.io/level/1.csv')
df = spark.read.csv(SparkFiles.get("1.csv"), header=True)
```


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

COVID Data Hub

- Definiowanie struktury danych za pomocą schematów modułu Spark

```
data = spark.read.csv(
'data/1.csv',
sep=',',
header=True,
)
data.printSchema()
```

```
root
 |-- id: string (nullable = true)
 |-- date: string (nullable = true)
 |-- confirmed: string (nullable = true)
 |-- deaths: string (nullable = true)
 |-- recovered: string (nullable = true)
```

iBigWorld:2020-1-PL01-KA203-082197

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

COVID Data Hub

- Strukturyzowanie danych za pomocą schematów Spark

```
from pyspark.sql.types import *
data_schema = [
StructField('confirmed', IntegerType(), True),
StructField('people_vaccinated', IntegerType(), True),
StructField('economic_support_index', DoubleType(), True),
StructField('iso_currency', StringType(), True),
]
final_struc = StructType(fields = data_schema)
data = spark.read.csv('data/1.csv', final_struc)
```

iBigWorld:2020-1-PL01-KA203-082197

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

COVID Data Hub

- Do strukturyzacji danych oraz ich kontroli można zastosować następujące funkcje:
  - dtypes
  - show
  - head
  - first
  - take, description, columns, count, different, printSchema

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

COVID Data Hub

- Przetwarzanie danych danych
- Moduł Spark SQL umożliwia wykorzystanie następujących operacji:
  - Select
  - Filter
  - Between
  - When
  - Like
  - GroupBy
  - Aggregate

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

COVID Data Hub

- Pozyskiwanie danych

```
data.select(['confirmed', 'people_vaccinated',
'iso_currency'])
.groupBy('iso_currency')
.mean()\
.show()
```

```
+-----------+-------------+--------------------+
|iso_currency|avg(confirmed)|avg(people_vaccinated)|
+-----------+-------------+--------------------+
| 467| null| null|
| 675| 1.962488E7| null|
| 296| 3357425.0| null|
| 1090| null| null|
| 1572| null| null|
```

iBigWorld:2020-1-PL01-KA203-082197

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

COVID Data Hub

Radzenie sobie z brakującymi wartościami

- Usuń wiersze
- Zastąp średnią / medianą (małe zestawy danych liczbowych)
- Zastąp najczęściej występującą wartością w kolumnie (atrybuty kategoryczne) -> może to wprowadzić błąd w danych
- Wymień z pomocą KNN

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

COVID Data Hub

Radzenie sobie z brakującymi wartościami (**missing values**)

```
# usuwanie wierszy
data.na.drop()
# zamiana na średnią
data.na.fill(data.select(f.mean(data['confirmed'])).collect()[0][0])
# zastąp nową wartością
data.na.replace(old_value, new_vallue)
```


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

COVID Data Hub

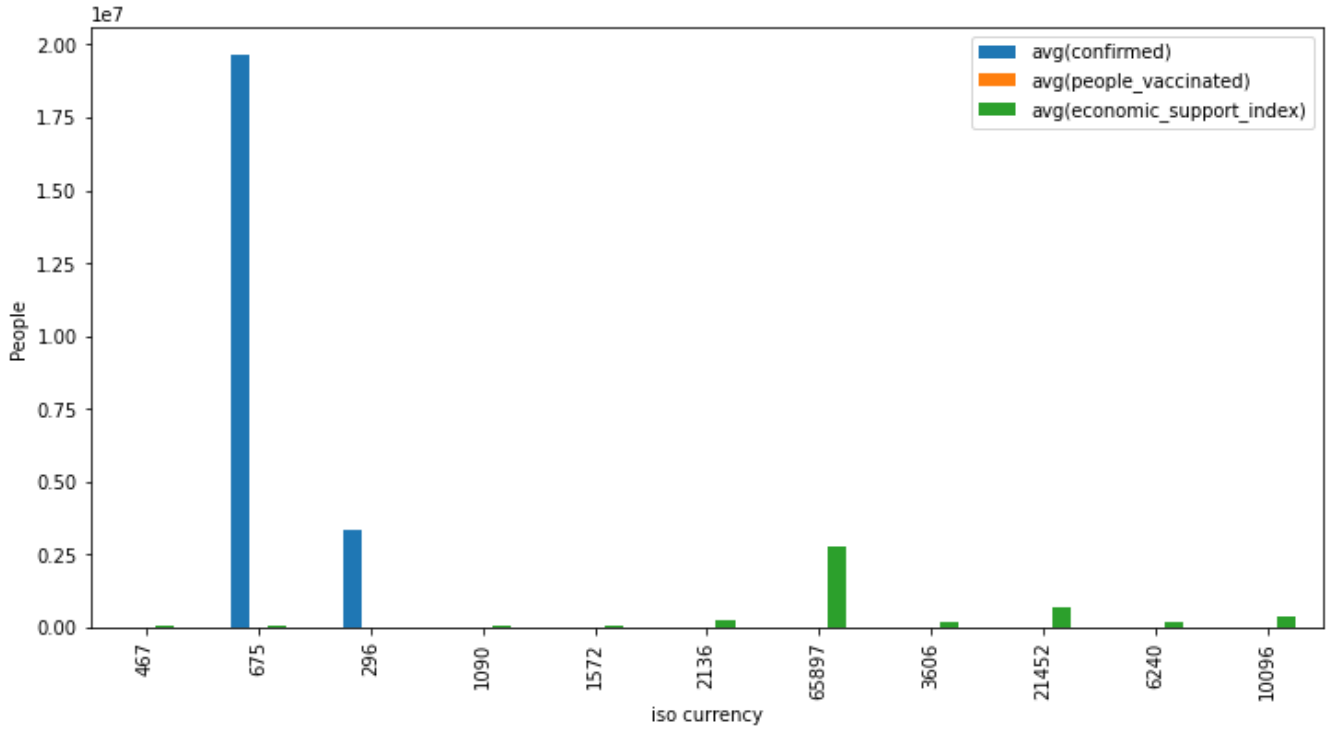
- Wizualizacja danych

```
from matplotlib import pyplot as plt
currency_df = data.select(['iso_currency',
'confirmed',
'people_vaccinated',
'economic_support_index']
)
.groupBy('iso_currency')
.mean()
.toPandas()
ind = list(range(12))
ind.pop(6)
currency_df.iloc[ind ,:].plot(kind='bar', x='iso_currency',
y=currency_df.columns.tolist()[1:],figsize=(12, 6), ylabel='People')
```


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

COVID Data Hub

- Wizualizacja danych - przykład



iBigWorld:2020-1-PL01-KA203-082197

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

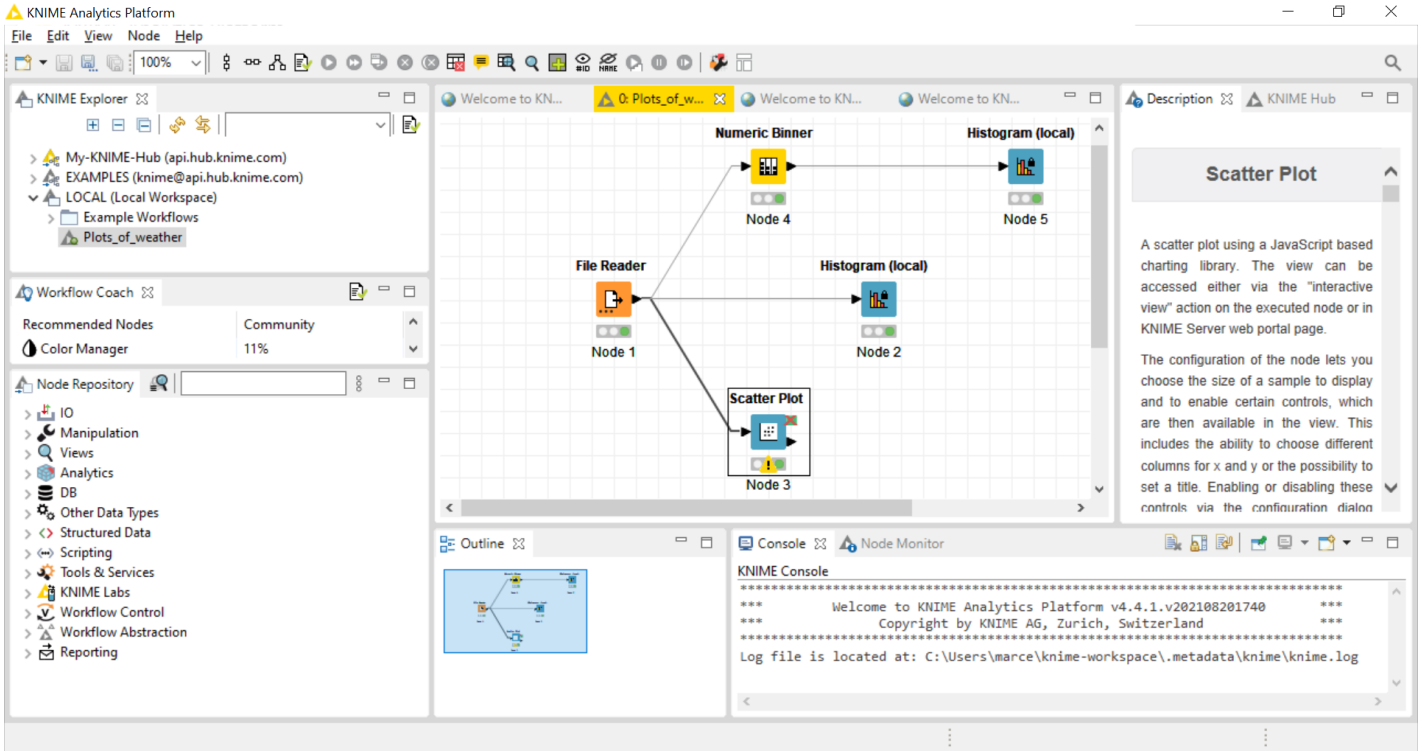
Dzienny zbiór danych o pogodzie

	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

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

Dzienny zbiór danych o pogodzie

- Eksploracja danych w KNIME

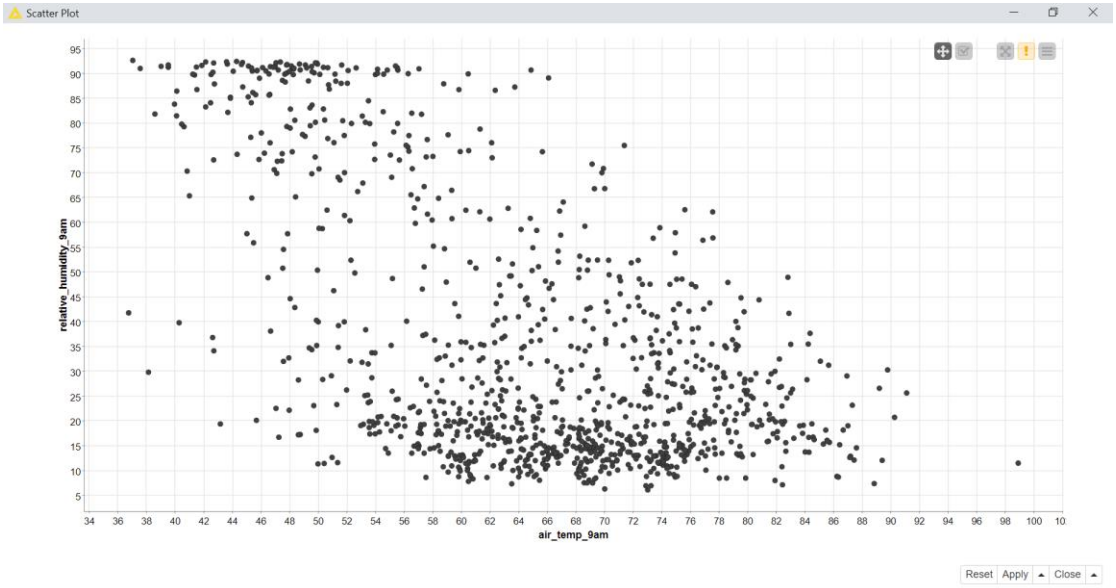


iBigWorld:2020-1-PL01-KA203-082197

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

Dzienny zbiór danych o pogodzie

- Eksploracja danych za pomocą KNIME
- Wykres punktowy

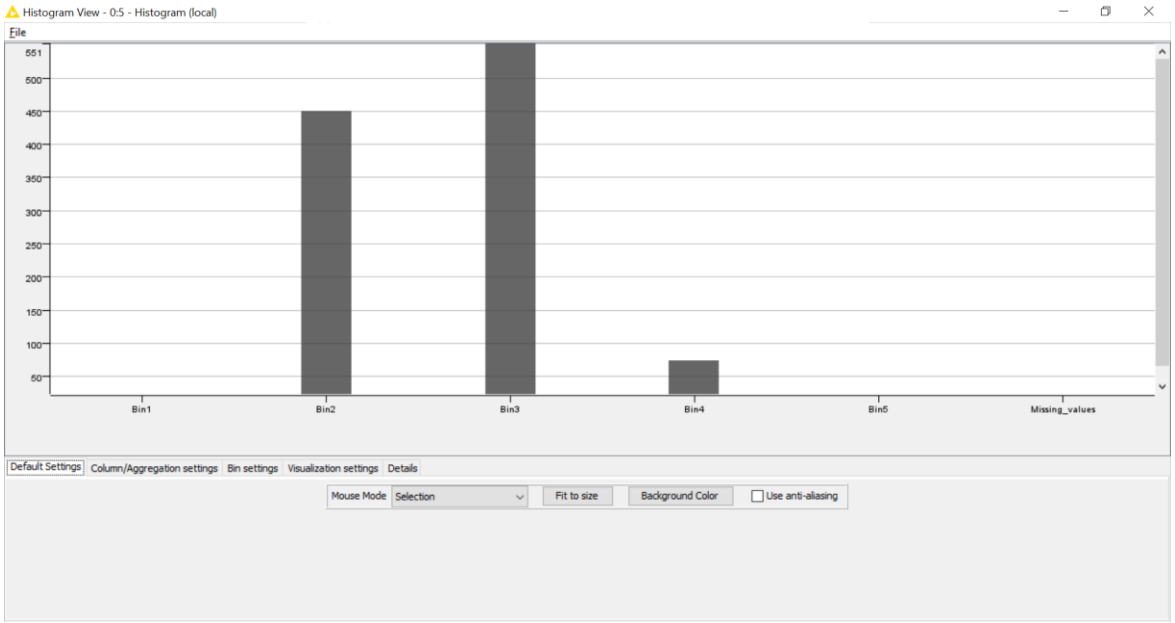


iBigWorld:2020-1-PL01-KA203-082197

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

Dzienny zbiór danych o pogodzie

- Eksploracja danych za pomocą KNIME
- Histogram



iBigWorld:2020-1-PL01-KA203-082197

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