II Seminarium Python Geo Data Science

Hel 6-10 maj 2019

wine_data.txt | Barensburg.csv | Alesund.csv | Hopen.csv | Hopen_NyAlesund_Barensburg.csv | iris_dane.csv | midwest_filter.csv | drinki.csv | oecd_bli_2015.csv

```
42702 "NOE00134778","NY ALESUND, NO","2018-06-03","0.0","0.0","1.8","2.7","0.4"
42703 "NOE00134778","NY ALESUND, NO","2018-06-04","0.0","0.0","3.3","4.9","1.8"
42704 "NOE00134778","NY ALESUND, NO","2018-06-05","0.0","0.0","2.8","4.8","1.9"
42705 "NOE00134778","NY ALESUND, NO","2018-06-06","0.0","0.0","2.8","4.1","1.3"
42706 "NOE00134778","NY ALESUND, NO","2018-06-07","0.0","0.0","2.3","5.1","-0.3"
42707 "NOE00134778","NY ALESUND, NO","2018-06-08","0.0","0.0","-0.1","0.8","-0.7"
42708 "NOE00134778","NY ALESUND, NO","2018-06-09","0.0","0.0","1.1","2.2","-0.9"
42709 "NOE00134778","NY ALESUND, NO","2018-06-10","0.0","0.0","2.1","2.6","1.6"
```

Normal text file | length : 4 893 718 | lines : 72 183 | Ln : 3 | Col : 21 | Sel : 0 | 0 | Unix (LF) | UTF-8

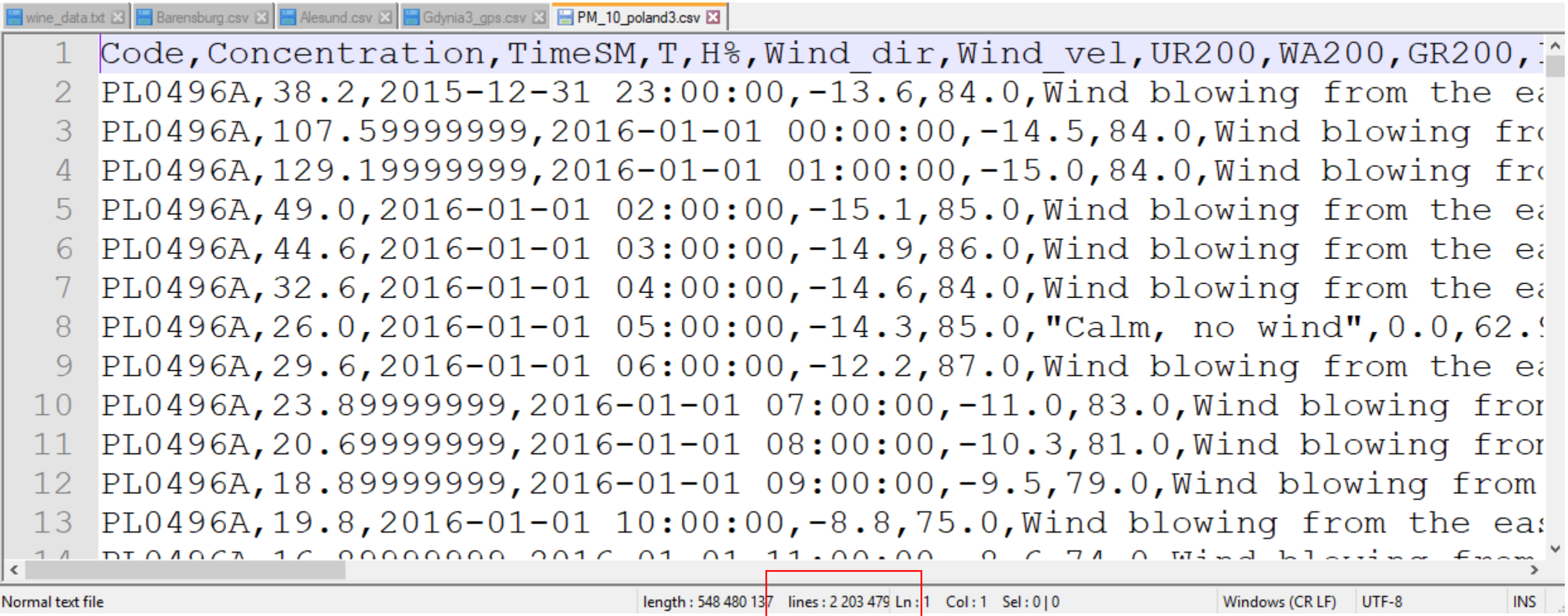
C:\JACEK2\Hel19_ESRI\dane\Gdynia3_gps.csv - Notepad++

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

wine_data.txt | Barensburg.csv | Alesund.csv | Gdynia3_gps.csv

```
1 ID,YY,XX
2 1,54.4651031,18.4819145
3 2,54.4666366,18.4797172
4 3,54.4679451,18.4777965
5 4,54.4707565,18.4738788
6 5,54.4748802,18.4719429
7 6,54.4766655,18.4722099
```

Normal text file | length : 5 888 903 | lines : 200 001



Dane tekstowe .csv

Zdjęcia satelitarne

DATA SCIENCE

DATA ENGINEERING

COMPUTATIONAL DATA SCIENCE

POZYSKIWANIE

PRZYGOTOWANIE

ANALIZA

RAPORTOWANIE

DZIAŁANIE

PLIKI TEKSTOWE

BAZY DANYCH

Web serwisy

No SQL storage

Eksploracja EDA

PREPROCESSING

Korelacja
Trend
Outliers
Statystyki

Wizualizacja

Histogramy
Boxplots
Line graph

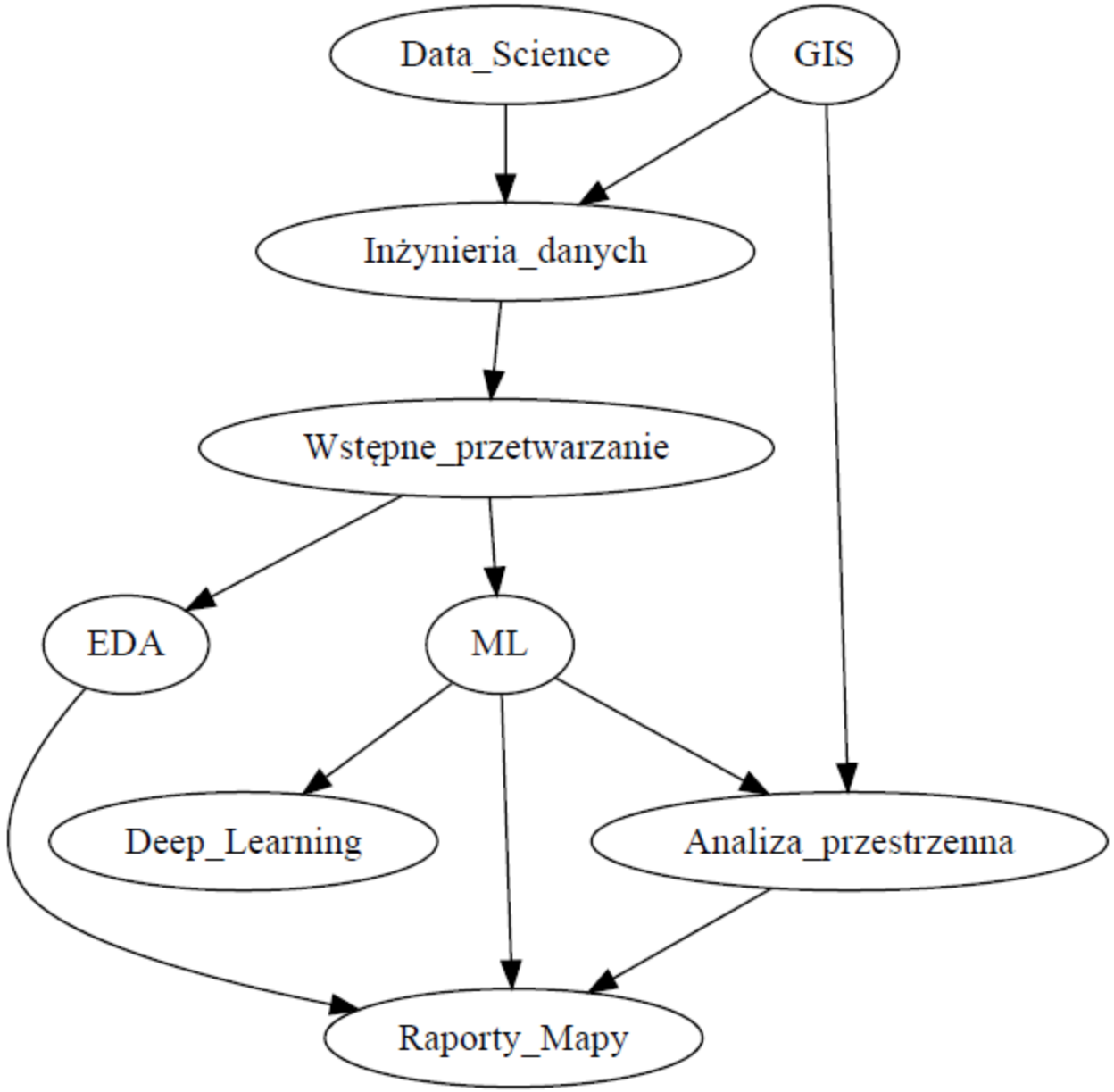
Czyszczenie

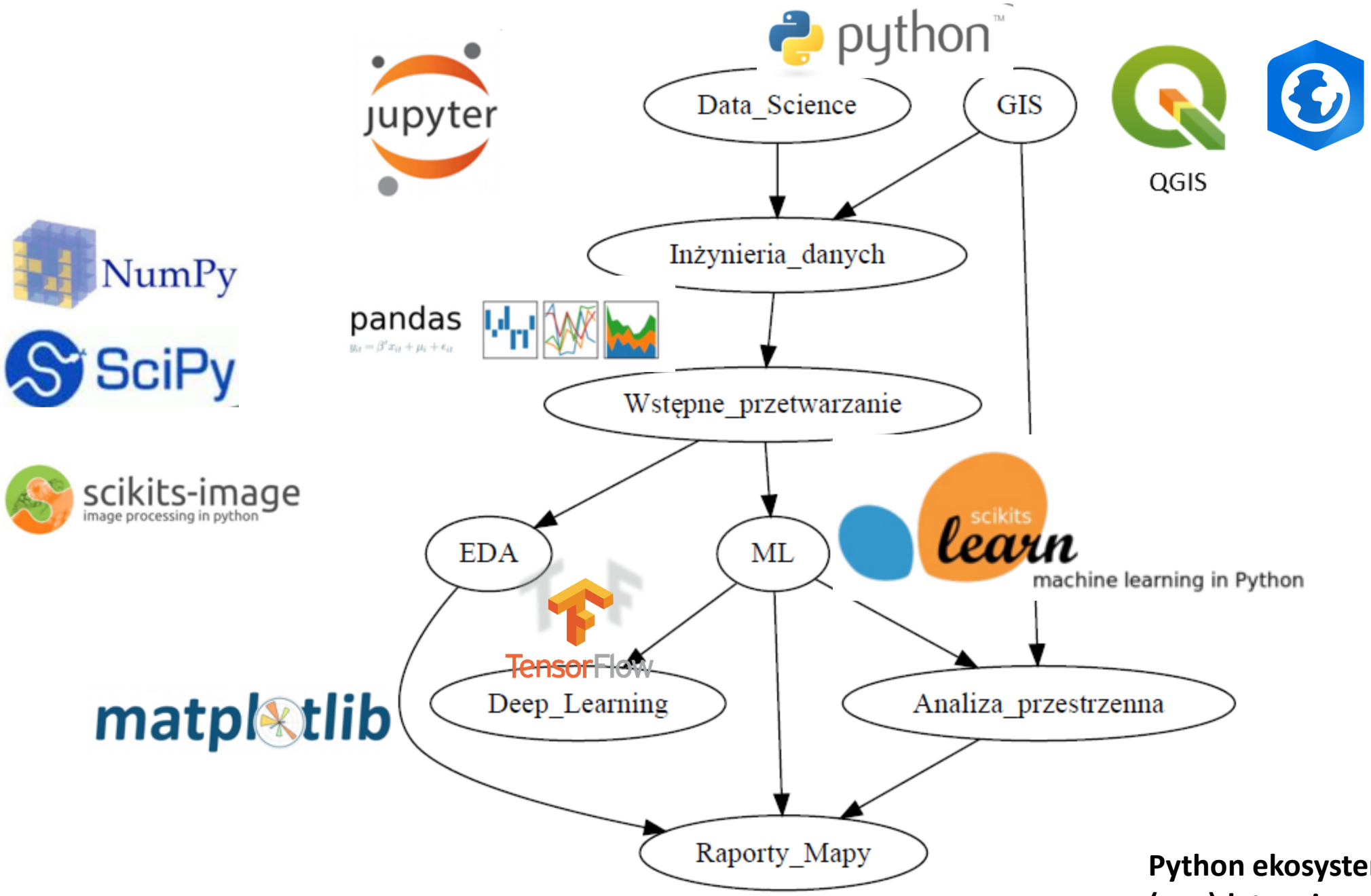
Duplicate
Inconsistent
Missing
Outlier

Transformacja

Scailing
Transformation
Feature selection
Dimensions PCA

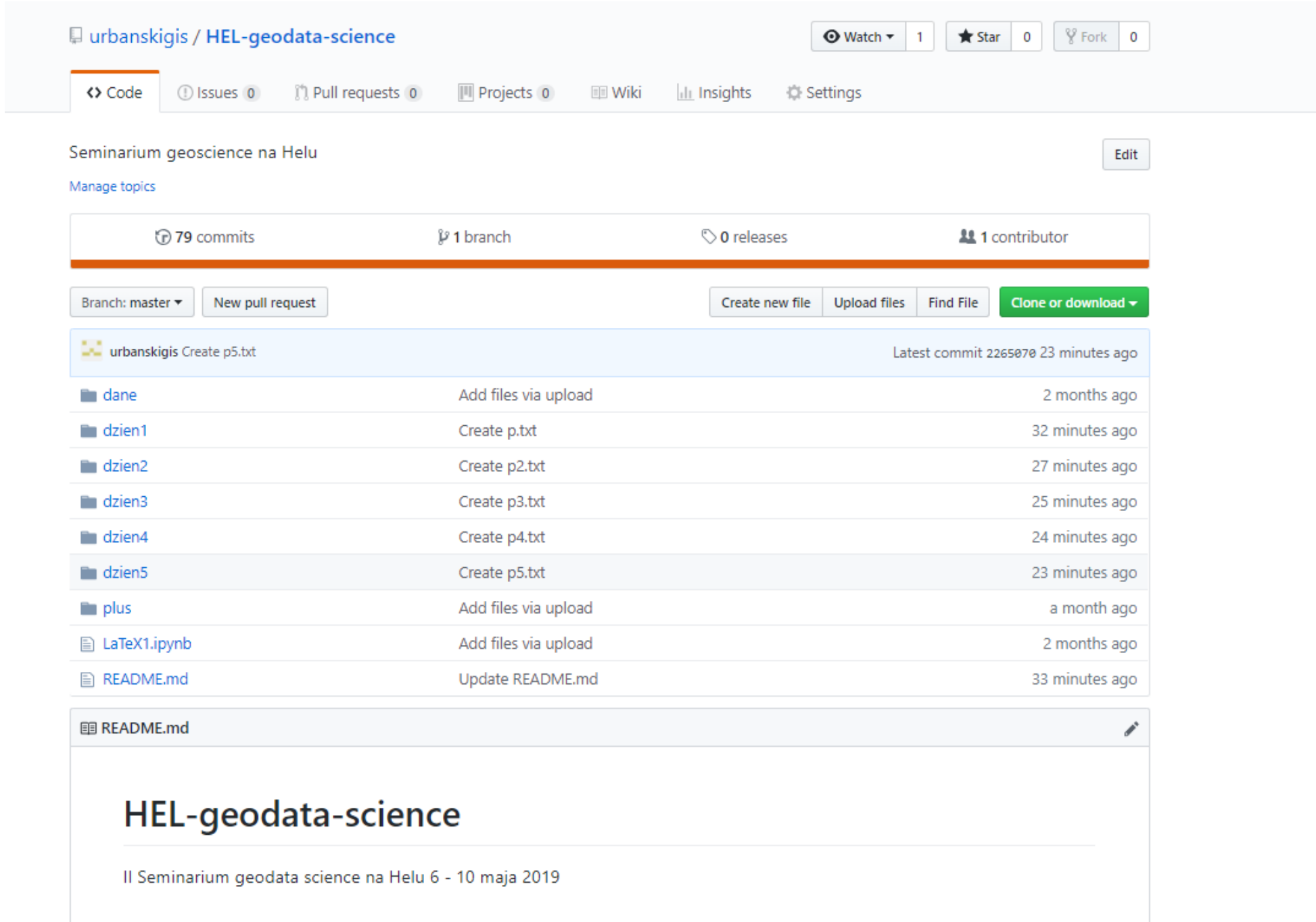
Classification
Regression
Clustering
Graph Analytics
……..


Data_Science
GIS
Inżynieria_danych
Wstępne_przetwarzanie
EDA
ML
Deep_Learning
Analiza_przestrzenna
Raporty_Mapy



Python ekosystem (środowisko) dla (geo)data science

https://github.com/urbanskigis/HEL-geodata-science



urbanskigis / HEL-geodata-science

Watch 1 | Star 0 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Seminarium geoscience na Helu

Edit

Manage topics

79 commits | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Create new file | Upload files | Find File | Clone or download

urbanskigis Create p5.txt — Latest commit 2265070 23 minutes ago

dane	Add files via upload	2 months ago
dzien1	Create p.txt	32 minutes ago
dzien2	Create p2.txt	27 minutes ago
dzien3	Create p3.txt	25 minutes ago
dzien4	Create p4.txt	24 minutes ago
dzien5	Create p5.txt	23 minutes ago
plus	Add files via upload	a month ago
LaTeX1.ipynb	Add files via upload	2 months ago
README.md	Update README.md	33 minutes ago

README.md

HEL-geodata-science

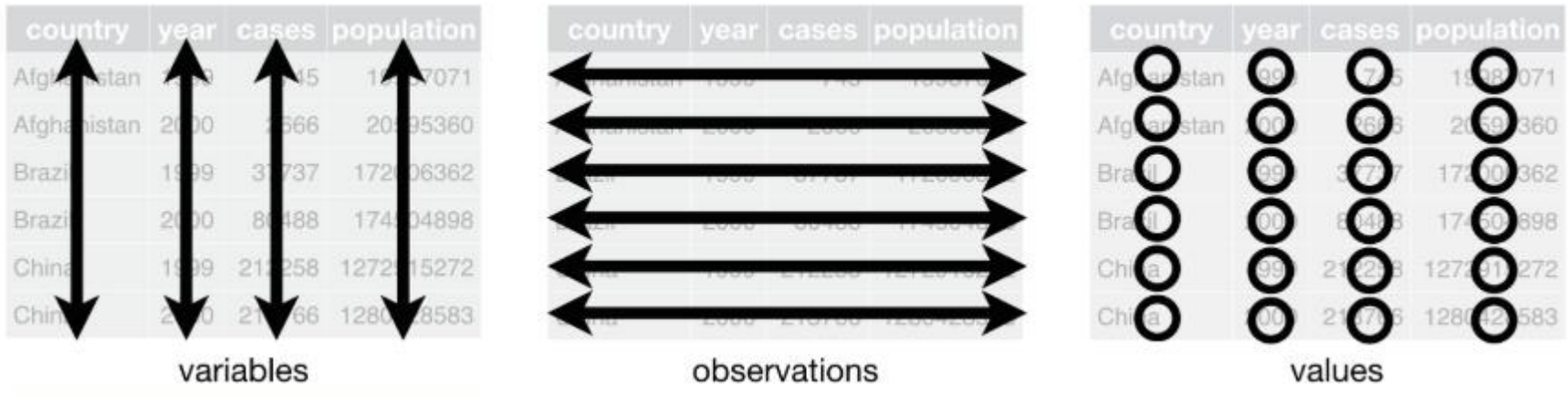
II Seminarium geodata science na Helu 6 - 10 maja 2019

Struktura danych - **Tidy data**

Tidy data to dane otrzymywane jako rezultat procesu zwanego **data tidying**. Jest to istotny etap czyszczenia danych podczas wstępnego opracowywania **big data** i jest istotnym praktycznym etapem data science. **Tidy data** posiadają strukturę zapewniającą ich łatwe przetwarzanie, modelowanie i wizualizację. **Tidy** zbiory danych (data sets) są zorganizowane tak, że każda zmienna tworzy kolumnę, a każda obserwacja tworzy wiersz.
Ta struktura danych tworzy standard dla **data cleaning** dając dostęp do wszystkich potrzebnych opcji.

Podsumowanie:

1. Każda zmienna, która jest mierzona (pozyskiwana) powinna być w jednej kolumnie.
2. Każda oddzielna obserwacja szeregu zmiennych powinna tworzyć oddzielny wiersz.
3. Dla każdego typu zmiennych powinna być oddzielna tabela.
4. Jeżeli posiadamy parę tablic, powinny zawierać kolumnę, która umożliwi ich łączenie.



https://towardsdatascience.com/

https://dataelixir.com/newsletters/

https://statquest.org/video-index/

https://www.youtube.com/watch?v=X3paOmcrTjQ

Dzień 1

Prezentacja – Data Science start

Wprowadzenie do jupyter notebook - film i ćwiczenia

Podstawowe operacje NumPy - trochę zadań
NumPy – praca z rastrami
NumPy algebra liniowa – parę zadań (jak to działa)

Matplotlib podstawy

Praca z czasem w Pythonie

Wykorzystanie składni LaTex a w Jupyter Notebook

Jupyter Notebook

NumPy

Matplotlib

Dzień 2

Pandas I
Pandas II

Wizualizacja – podstawowe rysunki w Pandas
Wizualizacja - Seaborn

Pandas w praktyce (praca samodzielna)

Ćwiczenia - pandas

Jupyter Notebook

Pandas

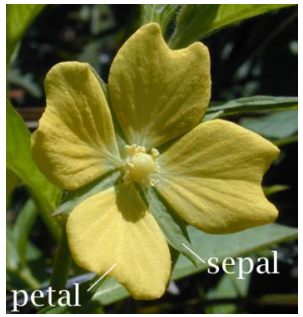
Seaborn

Dzień 3

Geopandas – plus zadania (mapy)

Machine Learning – prezentacja
Naive Bayes - prezentacja
Gaussian Naive Bayes – przykład (wino)

Gaussian Naive Bayes – zadanie (Iris flower data set)



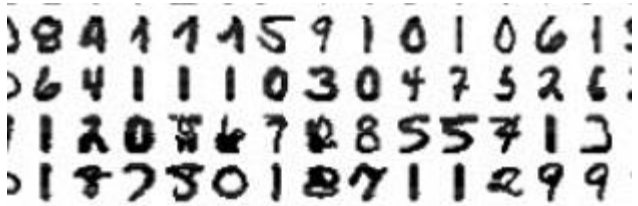
Ćwiczenia tworzenie rysunków (biblioteki Matplotlib, pandas, seaborn)

ML

Jupyter Notebook
Pandas
Geopandas
Naive Bayes
Seaborn
Matplotlib

Dzień 4

klasyfikacja



k – nearest neighbors classification (prezentacja)
k - nearest neighbors classification (przykład i zadania – MNIST, wino, storczyki)

Modified National Institute of Standards and Technology database

klasteryzacja

Hierarchiczna klasteryzacja (dendrogram)
Klasteryzacja – prezentacja
k – means clastering (przykład)
k – means clastering (zadanie-Iris) + t-SNE

t-Distributed Stochastic Neighbor Embedding”

regresja

Regresja – prezentacja
Regresja (przykład – Boston)
Regresja (zadanie – zanieczyszczenia)

Generowanie zbiorów testowych

Dzień 5

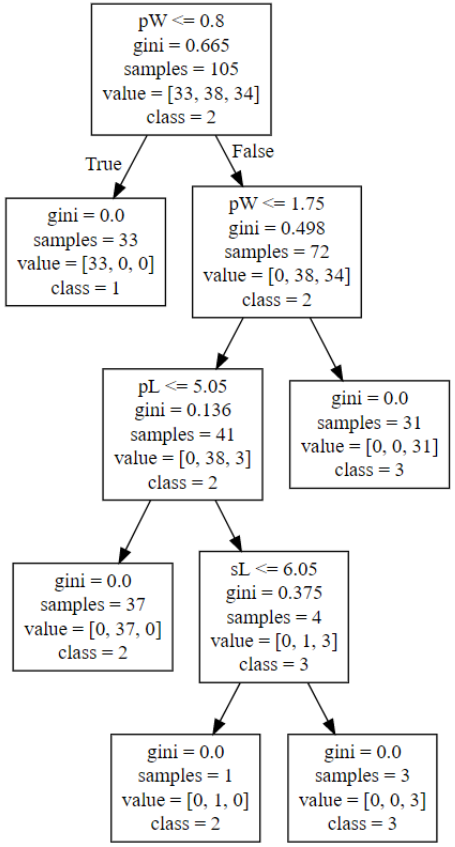
Decision Tree, Random Forest – prezentacja

Decision Tree Random Forest – przykład (wizualizacja drzew decyzyjnych Webgraphviz)

Logistic Tree Regression, SVM – prezentacja

Logistic Regression, SVM – przykład (Permutation Importance)

Zapisywanie modeli ML



Analiza danych czasowych – podstawy, I, II