I Seminarium Python Geo Data Science

Hel 18-22 marzec 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

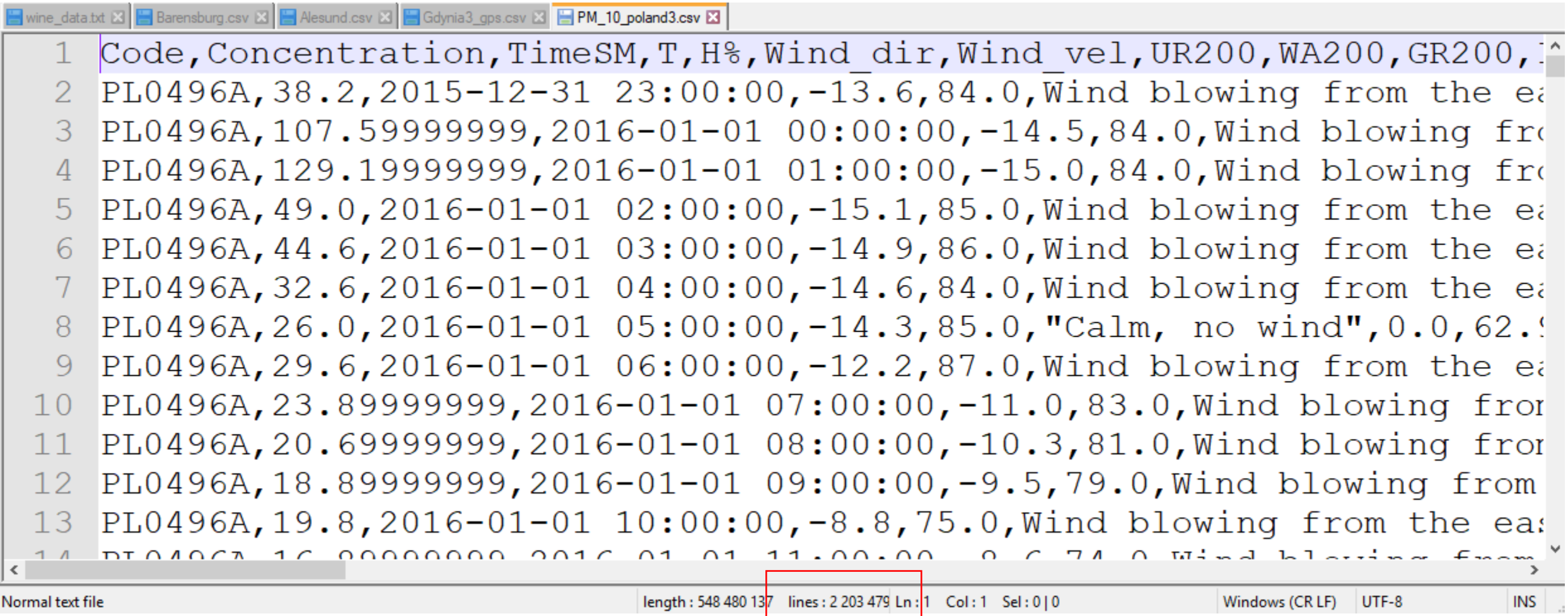
C:\JACEK2\Hel19_ESRI\dane\Gdynia3_gps.csv - Notepad++

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

DATA SCIENCE

DATA ENGINEERING

COMPUTATIONAL DATA SCIENCE

POZYSKIWANIE | **PRZYGOTOWANIE** | **ANALIZA** | **RAPORTOWANIE** | **DZIAŁANIE**

PLIKI TEKSTOWE

BAZY DANYCH

Web serwisy

No SQL storage

Eksploracja EDA

Korelacja
Trend
Outliers
Statystyki

Wizualizacja

Histogramy
Boxplots
Line graph

PREPROCESSING

Czyszczenie

Duplicate
Inconsistent
Missing
Outlier

Transformacja

Scailing
Transformation
Feature selection
Dimensions PCA

Classification
Regression
Clustering
Graph Analytics
……..

Python ekosystem (środowisko) dla (geo)data science

And many,
many more...

QGIS

urbanskigis/HEL-geodata-scienc

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

Apps | Wyborcza.pl - Najś... | Odebrane (6 475) - ... | Google | Centrum GIS Wydzi... | Course | DSE220x | ... | Oglądaj TV | Radio TOK FM - Inf... | Esri Training | Your ... | Other bookmarks

Search or jump to... / Pull requests Issues Marketplace Explore

urbanskigis / HEL-geodata-science

Watch 0 | Star 0 | Fork 0

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

Seminarium geoscience na Helu

Edit

Manage topics

6 commits | 1 branch | 0 releases | 1 contributor

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

urbanskigis Delete info.txt — Latest commit 99016bc a minute ago

dane	Delete info.txt	a minute ago
README.md	Update README.md	30 minutes ago
Wstęp do jupyter.ipynb	Add files via upload	3 hours ago

README.md

HEL-geodata-science

Seminarium geodata science na Helu 18-22 marca 2019

Struktura danych - **Tidy data**

Tidy data is the data obtained as a result of a process called data tidying. It is one of the important cleaning processes during big data processing and is a recognized step in the practice of data science. Tidy data sets have structure and working with them is easy; they're easy to manipulate, model and visualize. Tidy data sets are arranged such that each variable is a column and each observation (or *case*) is a row.[1][2]
Tidy data provide standards and concepts for data cleaning, and with tidy data there's no need to start from scratch and reinvent new methods for data cleaning.

Jeff Leek in his book *The Elements of Data Analytic Style* summarizes the characteristics of tidy data as the points:[3]

1. Each variable you measure should be in one column.
2. Each different observation of that variable should be in a different row.
3. There should be one table for each "kind" of variable.
4. If you have multiple tables, they should include a column in the table that allows them to be linked.