https://www.projectmates.eu/

MATES ED2MIT

Education and Training for Data Driven Maritime Industry

Data Science and Analytics Foundation

Practice 1 – Preparing your working environment

Yuri Demchenko MATES Project

University of Amsterdam

Maritime Alliance for fostering the European Blue economy through a Marine Technology Skilling Strategy

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

Outline

- Recommended tools for working with Data Analytics tasks
- Installing and configuring Anaconda and Jupyter Notebook
- Configuring your Python working environment
  - Changing location of your working directory
- Installing additional Python libraries

- Setting up RapidMinerEnvironment

- Link to existing tutorials and datasets

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

Recommended Learning/working/development Tools

- Working with Jupyter notebooks
  - Anaconda
  - Jupyter Notebook
- Direct working with Python with Command Line Interface (CLI)
  - Python 3.4+
- Visual Studio Code
  - Can work with variety programming languages
  - Can work with python and Jupyter Notebooks
- Rapid Miner Studio

Specialised and Advanced Tools

Pandas Profiling https://github.com/pandas-profiling/pandas-profiling

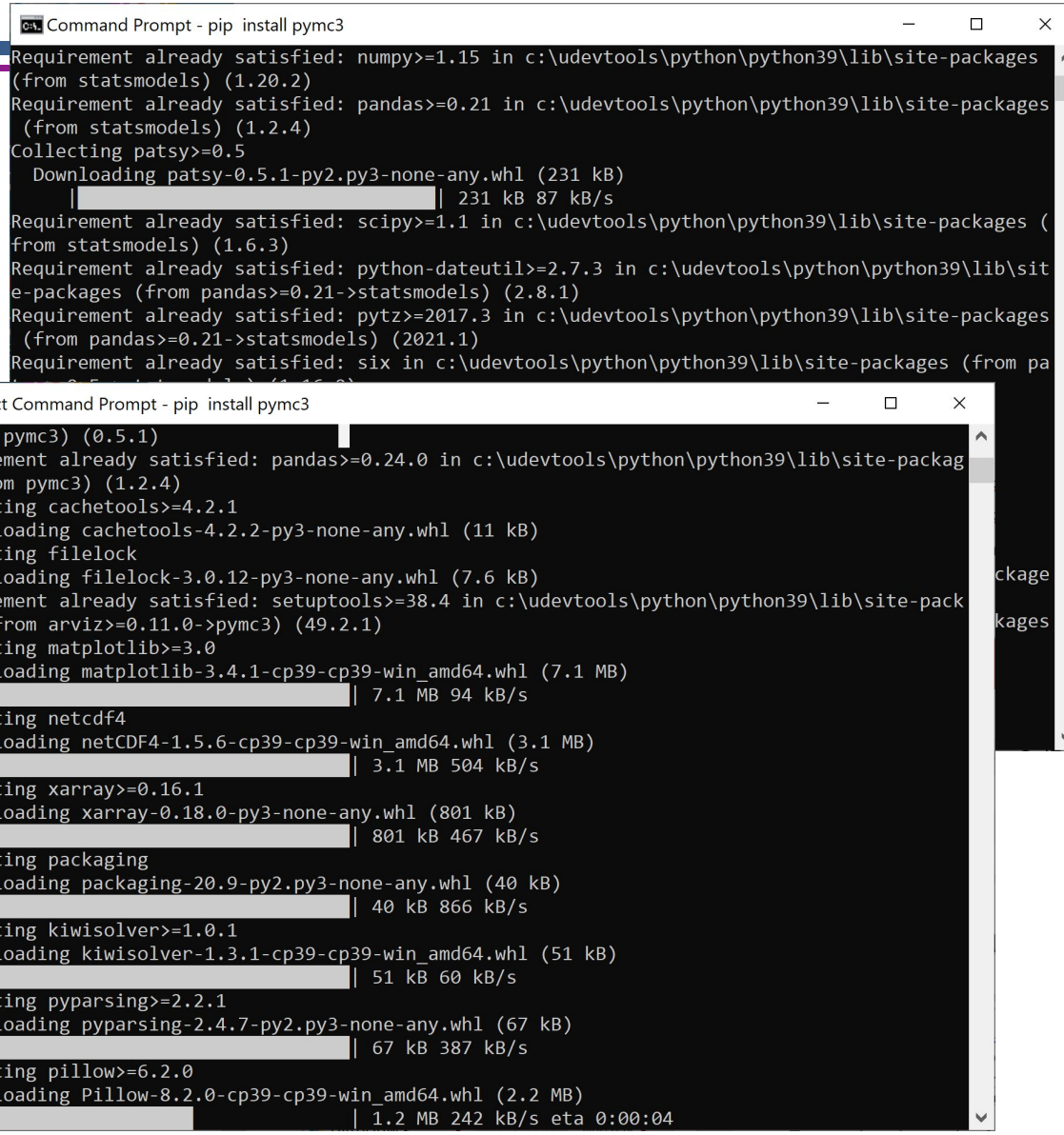
- Generates profile reports from a pandas DataFrame.
  - Extends *pandas df.describe()* function for effective exploratory data analysis with df.profile_report() for quick data analysis.
- For each column the following statistics are presented in an interactive HTML report:
  - Type inference: detect the types of columns in a dataframe.
  - Essentials: type, unique values, missing values
  - Quantile statistics like minimum value, Q1, median, Q3, maximum, range, interquartile range
  - Descriptive statistics like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
  - Most frequent values
  - Histogram
  - Correlations highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices
  - Missing values matrix, count, heatmap and dendrogram of missing values
  - Text analysis learn about categories (Uppercase, Space), scripts (Latin, Cyrillic) and blocks (ASCII) of text data.
  - File and Image analysis extract file characteristic and EXIF information.

Good set of examples

- Census Income (US Adult Census data relating income)
- NASA Meteorites (comprehensive set of meteorite landings)
- Titanic (the "Wonderwall" of datasets)
- NZA (open data from the Dutch Healthcare Authority)
- Stata Auto (1978 Automobile data)
- Vektis (Vektis Dutch Healthcare data)
- Colors (a simple colors dataset)
- UCI Bank Dataset (banking marketing dataset)
- RDW (RDW, the Dutch DMV's vehicle registration 10 million rows, 71 features)

Installing Python and updating main packages

- Required software
  - Install Python 3.6 and newer
  - Install the standard scientific Python stack: Jupyter, NumPy, SciPy, Matplotli
  - For statistics: pandas, statsmodels, PYMC3
- Checking if packages Available
  - Import numpy – *or if not available* - pip install numpy
  - Import scipy
  - Import pandas
  - Import statsmodels
  - Import pymc3
- Update Python installation module [python installation directory]python.exe -m pip install --upgrade pip
  - **sudo pip install ansible**



Downloading and Installing Exercise files and datasets

- Create the directory for all your tutorials or courses
  - For Jupyter Notebook default location is Desktop (you don't to change Jupyter configuration)
  - Default recommended location for exercises and datasets with Download exercise files and datasets and place them in separate directories
  - You can use the same directories structure as in the provided zipped exercise files
  - Optionally, you can create and maintain the separate directory for all datasets

Change location of the working directory in Anaconda

- **How to change Jupyter notebook start up folder in Anaconda**
- [ref] https://www.planetofbits.com/python/change-jupyter-notebook-startup-folder-anaconda/
- In Anaconda Navigator open a command prompt window via Environment > base (root) > Open Terminal.
  - Type the command ***jupyter notebook –generate-config*** in the command window and press *Enter*.
  - This will create a file with the name ***jupyter_notebook_config.py*** in the location *C:\Users\YOUR_USERNAME\.jupyter*
- Go to the folder location *C:\Users\YOUR_USERNAME\.jupyter* and open the file named, ***jupyter_notebook_config.py*** in any text editor.
  - Find the key string, ***#c.NotebookApp.notebook_dir***
  - Uncomment the key string by deleting the # sign and *in single quotes* type the location of your custom startup folder and save the changes.
- Restart Anaconda and start Jupyter Notebook from the Anaconda Navigator