DataDisca
machine-learning algorithms

Module 2: Analysis and Visualisation Basics

The data scientist has to be comfortable in data extraction, transformation and loading. Then visualisations help the data scientist solve many important problems. There are numerous examples where the problems are solved with the right statistics and visualisations even before applying any specialised data science algorithm. To familiarise with data analysis and visualisation basics, we study numpy, pandas, seaborn, plotly and folium OpenStreetMap in this section.

Data Access:
Please check the data folder in the share.

Powerpoint Template:
Please check the slide template folder in the share.

To Study

Please search study materials on the Internet. You can also use books or videos. Order up to 3 materials for each category as you think useful. The best one should come first. Share your jupyter notebooks or python codes with us. We use your codes to assess and guide you if there are components that need more attention. You do not have to tidy up your study code.

1. Study numpy using a tutorial found on the internet
    a. What is the link to the tutorial/ what is the name of the book? You can list up to 3
    b. What are the most useful features you learnt?
    c. Find a cheat sheet and write the link below.
    d. Master the following
        array, zeros, arange, linespace, eye, empty, reshape, resize, append, insert, concatenate

2. Study pandas using a tutorial found on the internet
    a. What is the link to the tutorial/ what is the name of the book? You can list up to 3
    b. What are the most useful features you learnt?
    c. Find a cheat sheet and write the link below.
    d. Master the following
        i. Read and write data frames from csv, Excel and json files
        ii. Create a data frame from numpy arrays
        iii. Create a data frame from dictionaries

iv. Append and delete rows
v. Data read/write with iloc and loc
vi. Data subsetting and sampling
vii. Rename columns
viii. Add and delete columns
ix. Change column data types
x. Sort by columns
xi. Sort by index
xii. Add delete indexes
xiii. Add delete multi-level indexes
xiv. Muti indexes: add, remove
xv. Multi-level column names, add, remove, rename
xvi. map, apply, applymap and lamda functions
xvii. idxmin, idxmax
xviii. Pivot tables
xix. iteritems, iterrows
xx. groupy
xxi. Replace, fillna, dropna
xxii. merge
xxiii. Join
xxiv. concat
xxv. value_counts, unique
xxvi. Resampling
xxvii. Pipelines

3. Study seaborn using a tutorial found on the internet
    a. What is the link to the tutorial/ what is the name of the book? You can list up to 3
    b. What are the most useful features you learnt?
    c. Find a cheat sheet and write the link below.
    d. Master
        i. Bar charts
        ii. Line and scatter plots
        iii. Pie charts

4. Study Plotly using a tutorial found on the internet
    a. What is the link to the tutorial/ what is the name of the book? You can list up to 3
    b. What are the most useful features you learnt?
    c. Find a cheat sheet and write the link below.
    d. Master
        i. Generating interactive HTML plots
            1. Line and scatter plots
                a. Single y-axis

   b. Multiple y-axes
   c. n x m plots per page
2. Bar Charts

5. Study folium using a tutorial found on the internet
   a. What is the link to the tutorial/ what is the name of the book? You can list up to 3
   b. What are the most useful features you learnt?
   c. Find a cheat sheet and write the link below.

Application

This is about exploratory data analysis of product_a.csv.

1. Import product_a.csv dataset into python pandas data frame df_product_a. The first column is the index.
2. Convert date_w field to a suitable datetime data type
3. Values of the year column do not match with the values of the date_w column. Correct the values of the year column.
4. Create df_stats with the following details from df_product_a
   Columns: filed_name, minimum, maximum, mean, standard deviation, variance, mode, median, 10th, 20th .. 90th percentiles, 1st, 2nd and 3rd quartiles, interquartile distance, skewness and kurtosis.
5. Theory: Discuss the relationships between the fields of df_stats. For example, 2nd quartile and the median are the same.
6. Discuss how the columns of df_stats are useful in data analysis.
7. Analyse data based on your discussion and explain the results. What are the notable features of the dataset?
8. Create a Pearson correlation matrix (it is a square matrix) between all the possible fields. What are the conclusions you make?
9. Create a Spearman's Rank correlation matrix (it is a square matrix) between all the possible fields. What are the conclusions you make?
10. Create a seaborne pairplot for df_product_a. What are the conclusions you can make using the analysis sofar
11. Using Plotly, draw weekly and monthly time-series graphs of the numeric fields. Explain the results.
12. Draw year based location and type bar charts using Plotly. Discuss your results.
13. Compare and contrast the prices of each type, each location and {location and type} combination. Visualise the results using suitable plots.
14. Visualise data on a folium map. The locations should have markers with a colour range based on the mean values of bags_t. Tooltips should show the total values of bags_t

and total values of bag_t for each type. When markers are clicked, the average values of all numeric fields should be shown.

How to present the results of your analysis

We expect commercial quality presentations.

1. Share your Jupyter notebook page(s) or python code with us. **Please tidy up this work. This has to be a professional-quality code.**
2. Create an attractive presentation. You can use Microsoft PowerPoint or any alternative commercial or open-source products. Use DataDisca template when possible. Extend the template for different slide formats that you would like to have. Otherwise, create a similar DataDisca template for your presentation software. Random presentation themes are not accepted.
3. Present using zoom or skype. We also want to see your code and the original results over zoom or skype.