Список 02 – Домашнее Задание

Введение о науки данных и визуализации

Предмед: Введение в профессиональную деятельность

Преподаватель: Хольгер Эспинола Ривера

1. Data Analysis and Visualization. Take as template the jupyter notebook of lesson 03 called **lab03_ds** worked in laboratory of Data Science and Data Analysis, follow the next instructions bellow:

1) Select one dataset with tabular data structure, which contains numerical and categorical type of data (need both necessarily). Dataset need have minimum 500 rows and number of columns (features) min = 6 and max = 50. You can use your own datasets from your past projects, your company or take one of the public datasets in the most popular repositories like:

- UCI (https://archive.ics.uci.edu/datasets);

- Kaggle (https://www.kaggle.com/datasets)

- Google Dataset Search (https://datasetsearch.research.google.com).

Your dataset can have format of file like .csv, .xls, .json, .sql, .data, .txt ... etc;

2) In groups of 4 students, develop and publish your own data science project in Github repository, making the statistical analysis for numerical and categorical variables. Need have 1 folder which contains the dataset files and 1 jupyter notebook file **.ipynb**, which contains the python source code;

3) Define the Metadata, data collection and load as dataframe and take care the missing values;

4) Make data visualization for Numerical Variables: plot histogram of frequencies, bivariant charts, box plot, calculate the statistical metrics, and plot the correlation table of numerical features;

5) Make data visualization for Categorical Variables: plot bar chart for absolute frequencies for each instance using all categorical variables, plot the pie chart for relative frequencies for each instance using all categorical variables;

6) Make Statistical Analysis for Groups of Categorical Variables: choose 2 categorical variables of your dataset. Create groups using the instances of these 2 categories and count the number of samples for each of these combined instances. Plot the grouped absolute frequency analysis using bar chart and grouped relative frequency analysis using pie chart.

7) Using the software Grafana and the same dataset, create a dashboard containing 2 examples of statistical analysis for groups of categorical variables.