Project #4 – Data Visualization Project
Big Data Analysis

Project Description

It is time for the Data Visualization Project. You will work with a large data set of your own choice and analyse the heck out of it. This project is a perfect opportunity to enrich your code portfolio with a proper Data Analysis & Visualization Project.

Project Goals

- Get comfortable analysing & visualizing big data sets
- Conduct the project from start to finish in a structured way
- Practice effectively communicating the value of your analysis

Project Requirements

The following requirements are **mandatory**:

- Collect a data set from the web (pro tip – Kaggle)
  - The data set must have a minimum of 10 000 rows and 10 columns
  - Briefly plan some of the questions you would like to explore with the data
  - You should merge (at least parts of) your main data with another data set from a different source
- Present your chosen data to us before finalizing your choice
- Continuously use Trello Board for project management
- Conduct a Data analysis that includes:
  - A descriptive analysis
  - A diagnostic analysis (building upon the descriptive analysis)
- Put the project on your GitHub

Technical Requirements

- Put the data into (one or more) Pandas DataFrame(s)
- Clean your DataFrame(s)
- Merge data
- Slice the DataFrame(s) in interesting ways
- Reshape the data (using groupby/pivot_table etc.) and perform agg. operations
- Create your own features/columns
- Use statistical tools in your analysis i.e.:
    - Classifying the data based on type and measurement level
    - Standard Deviations
    - Correlations
    - Boxplot with IQR
    - Distribution types
    - Etc…
- Visualize your findings!

Presentation

The presentation should take max 10 minutes

The slides should include the following (not necessarily in this order):
- Title of the project + Student name
- Clear description of the question(s) you were exploring
- Clear description of your data set
    - Visualize!
- Clear communication of your main finding(s)
    - Visualize!
- Challenges
- Learnings / highlights
- Improvements
    - If you were to do it again
    - If you were to spend more time on your current project

Schedule

The presentations will take place Monday afternoon next week and you should hand in the project (slides, links etc) before presenting.

Resources

[Kaggle (for Datasets)](#)

[4 types of data analytics](#)

[Data to Viz](#)

Gustavs slides!

Good Luck!!