The German International University
Faculty of Informatics and Computer Science
Dr. Nada Sharaf
Ali Salem

Data Engineering and Visualization, Winter Semester 2022
Project Milestone 1
Submission Date: Saturday 03/12/22

Project Overview:

The goal of the project is to go through the complete data engineering process to answer questions you have about the topic in the dataset below:

- https://www.kaggle.com/datasets/bartoszpieniak/poland-cars-for-sale-dataset
- You will acquire the data, design your visualizations, run exploratory and statistical analysis, preprocess the data as needed, and communicate the results.

It is critical to note that no extensions will be given for any of the milestone submission due dates. Projects submitted after any of the due dates will not be graded. If you anticipate any issues (e.g., due to travel plans) you need to send me an email at least one week in advance.

Any changes that you make to your GitHub repositories after the due date will be ignored. Please have all your work submitted and tested before the deadline.

Project Team (Up to 3 people per group)

In general, we do not anticipate that the grades for each group member will be different. However, we reserve the right to assign different grades to each group member based on your contributions on GitHub.

Note that your commits will be checked by the end of every milestone. Remember it is not by the number of commits, but by the actual contribution to the notebook.

If you fail to provide evidence through GitHub contribution you will lose 10 % or more of each milestoneâs grade depending on the severity of the case.

Milestone 1 : Cleaning and Visualization (20/11 -> 03/12)

In this milestone, you will be:

- Exploring your data
- Finding your reserach questions
  - The research questions are the questions that you will be targeting to answer througout the project.
- Cleaning your data
  - handling missing data
  - handling outliers
- Plotting your data

The grades will be given for:

- **Implementation:**
  - Is your EDA thorough?
  - Did you clean your data well?
    - Were missing values handled appropriately?
    - Were outliers detected and handled?
  - Are the plots relevant? (We strongly advise you to include as many relevant visualizations in your notebook as you can.)
  - Do the plots show non-trivial information?
  - Complexity of your research questions.
- **Descriptive markdown cells for your code cells.**. Your notebook is the place you describe and document the space of possibilities you explored at each step of your project. What visualizations did you use to look at your data in different ways? What are the different statistical methods you considered? Justify the decisions you made, and show any major changes to your ideas. How did you reach these conclusions?
- **Clean code** ex: descriptive variable names.
- **Comment your code when relevant.**
- Example of things that can lead to **deduction of marks**:
  - Over simplistic research questions.
  - Little contribution
  - Inappropriate choice of plots/labels
  - Messy code
  - Incorrect conclusions
  - Trivial Insights

Deliverables

You should submit

- Your submitted GitHub repo should include:
  - A jupyter/colab notebook
  - Equally important to your final results is how you got there! **Include a README.md file** that contains:
    - Overview of the used dataset.
    - Provide an overview of the project goals and the motivation for it.
    - Descriptive steps used for the work done in this milestone.
    - Data exploration questions.

Submission Instructions

Submission will be handled through GitHub Classroom. A further notice will be sent later about google classroom.