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

Data Engineering and Visualization, Winter Semester 2022
Project Milestone 2
Submission Date: Saturday 31/12/22

In this milestone you will perform

- Data Integration:
  - You will be integrating 3 datasets airlines.csv , airports.csv and flights.csv The integration should result in a new dataset (the original one should stay as is since we will not use all columns for the integration). The link for the dataset is https://www.kaggle.com/datasets/usdot/flight-delays/versions/1?resource=download
  - The newly added data should be included in at least one research question per individual in the group.
- Handling the outliers (if there are any outliers in the dataset).
- Feature Engineering
  - You will be creating at least two new features that will add value to your analysis.
  - Each feature should be used in at least one research question (either an existing research question or coming up with a new one)
- Analysis
  - You should include the answers to all your research questions in the notebook. Answers should be provided as visualizations with a markdown cell explaining the insights we gain from the graph.

The grades will be given for:

- **Implementation:**
  - 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.
  - Implementation:
    - Is the data integrated properly?
    - Do the newly engineered features contribute to your insights?
    - Did you provide all the required research questions?
    - Did you provide visualizatins answering your research questions clearly?

- Example of things that can lead to **deduction of marks**:
  - Unuseful features
  - Wrong integration
  - Research questions with no answers
  - Inappropriate choice of plots/labels/title
  - 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:
    * Description of the newly added dataset.
    * Description of the features you added.

Submission Instructions

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