CSCI 4146/6409 - Process of Data Science (Summer 2023)

Assignment 1: Data Exploration

This assignment aims to help students enhance their exploratory data analysis skills. Students will engage with a real-world dataset, learning to identify pertinent variables, evaluate data quality, produce summary statistics, and create visualizations. They will also derive initial insights and hypotheses. This hands-on experience will deepen students' understanding of the exploratory data analysis process, preparing them to apply these skills in practice.

Due Friday May 19, by 11:59 pm - see submission details below.

Dataset: U.S. Airbnb Open Data - https://www.kaggle.com/kritikseth/us-airbnb-open-data

In this assignment, you will use the U.S. Airbnb Open Data dataset. You will go through the first three stages of a data science project defined in CRISP-DM and derive some data insights.

1. [1.5] Business understanding
   a. Formulate a business problem that can be solved with the dataset. [0.25]
   b. For the business problem, propose 3 data science solutions and assess their feasibility. Select the final solution and explain your decision. [0.5]
   c. For the final solution, identify the prediction subject, its domain concepts, and sub-concepts (if there are any). Draw a hierarchical graph of the concepts. [0.25]
   d. For each domain concept, design descriptive features that best describe a concept using data from the dataset. Summarize the resulting ABT in a table with the following columns [0.5]:
      i. Feature Name
      ii. Domain Concept
      iii. Feature Description
      iv. Feature Type
      v. Data Type

2. [1] Data Exploration
   a. Build the data quality report of the resulting ABT. [0.5]
   b. Identify data quality issues and build the data quality plan. [0.5]

3. [0.75] Data Preparation
   a. Preprocess your data according to the data quality plan.

4. [1.75] Data Insights
    a. Build the correlation heatmap of the features in the ABT. Derive insights from it and relate it to the business problem being addressed. [0.75]
        i. What are the descriptive features that highly correlate with the target feature? Propose some hypotheses explaining the correlation. [0.5]
        ii. What are the domain concepts that highly correlate with each other? Propose some hypotheses explaining the correlation. [0.25]
        iii. Are there any features that are useless for a predictive model? [0.25]

Submission Details

Submissions should be made through Brightspace, adhering to the due date and time specified under the Assignments section. To prepare your assignment solution, make use of the provided assignment template notebook on Brightspace. The detailed requirements for writing and coding can be found within the evaluation rubric document available on the platform. Keep in mind that questions will be graded individually using letter grades, with their respective weights indicated in parentheses.

You may complete the assignment individually or with another individual. In the case of a pair submission, only one student should submit the assignment on Brightspace. Be aware that plagiarism detection tools will be employed to identify any instances of cheating or copying in both your code and the accompanying PDF.

Your submission should consist of a single Jupyter notebook as well as a PDF that contains the compiled results generated by the notebook. This PDF should include both the code and the results as part of the final printout. Name your files as follows:

- A1-<your_name1>-<your_name2>.ipynb and
- A1-<your_name1>-<your_name2>.pdf.

Failing to submit both files will result in a zero mark for both students.