Report template Coursework 1 Software Development for Data Science

This structure provides a **logical flow for your report** while leaving room for individual interpretation of each task.

Please note: you are free to deviate from this proposed report structure, as long as all the required content is in the report. Make sure that you check your work against the marking rubric to ensure that you have included all aspects listed within the rubric!

The report is expected to have a length of around **2000-3000 words**, *excluding* code cells and output. This is only a guide and there will be *no penalty* from having a lower or higher word count per se.

The report has to be generated from a Jupyter (Colab) Notebook, in which you must document all your code and findings in a well-organized manner, with clear comments and section headings.

TITLE

Student Name:
Student ID:
Date:

Abstract

1. Introduction

In this section, introduce the assignment's goals and outline your approach. Summarize the objective of analyzing the Cape Town Airbnb dataset.

2. Dataset Overview

Provide a brief description of the dataset. Mention the data source and the types of data it includes, such as listing information, pricing, availability, and host details.

3. Methodology

3.1 Data Loading and Initial Exploration

- Describe the steps taken to load and inspect the datasets.
- Mention any basic observations about the datasets from, such as their structure, column names, and the number of rows and columns.

3.2 Data Cleaning and Wrangling

Explain how you handled missing values, duplicates, or unusual observations with a clearly reasoned approach. Remember that data cleaning is an iterative process that involves

checking, correcting, transforming, and validating your data throughout your analysis workflow.

- Variable Selection: Summarize how you selected variables of interest and any columns you chose to drop due to missing data.
- Detecting and Handling Missing Data: Describe the strategy used to handle missing values and explain your reasoning (justify your choices).
- Duplicates and Data Types: Outline any steps taken to identify and remove duplicates and to ensure correct data types for each column.
- Text Standardization and Feature Creation: Briefly discuss any standardization of text fields or new columns created based on existing text data (if applicable).

3.3 Exploratory Data Analysis (EDA)

Provide an in-depth exploration of your selected variables with comprehensive statistics, well-constructed plots, and insightful interpretation of data relationships, including:

- Summary Statistics: Provide an overview of the statistical summaries of key numeric columns, highlighting aspects such as central tendency and spread.
- Distribution Analysis: Describe the distribution of main variables (e.g., price, review scores) and any outliers identified. Discuss the outliers in some detail. You may use visualisations here.
- Relationships Between Variables: Summarize key relationships between variables and illustrate these using visualisations and statistics (e.g., scatter/hex plots, box plots, correlation heatmaps).

4. Proposed Research Questions

Based on your data exploration, propose three research questions that can be explored with this dataset. For each question:

- List the Variables: Identify the relevant variables involved.
- Specify the Analysis Type: Describe the analysis approach you would use (e.g., correlation analysis, group comparison, predictive modeling).

Try to provide clear, well-defined questions that can provide valuable insights.

5. Findings and Conclusions

Summarize the main findings from your data cleaning, wrangling, and exploration exercise. Discuss any interesting trends, patterns, or anomalies you noticed that might provide insights for your proposed research questions. Note that you do **not** need to use formal statistical / ML modelling here, but instead you should explain the insights and trends you've observed. Conclude with suggestions for further analysis or steps that could be taken to deepen insights using this dataset.

7. References

List any sources or documentation referenced, including the Inside Airbnb project website and any Python libraries used.