
School of Computing, Engineering and Built Environment

# Software Development for Data Science

## Module Code: MMI226822

### Coursework 1

Issue date: 31 October 2024

This coursework comprises 50% of the overall mark for the module.

# Exploratory Data Analysis – Airbnb Cape Town dataset

Figure 1. Cape Town waterfront. Image source: https://www.waterfront.co.za/business/permits/faq/

## 1. Introduction

This assignment will guide you through essential data cleaning, wrangling, and exploratory data analysis (EDA) on an Airbnb dataset of Cape Town, the capital city of South Africa (Figure 1). Your task is to transform and analyze the data, then develop three research questions that could be answered using it. The goal of this coursework is for you to demonstrate proficiency in the techniques we have covered in this class (and beyond if you like) using Python and apply them to a novel dataset in a meaningful way.

## 2. Dataset

The data are made available by *Inside Airbnb*, which is an independent, community-driven project that compiles and analyzes publicly available data on Airbnb listings. Airbnb data typically includes information about listings (such as location, pricing, availability, and host details), providing insights into market trends, housing availability, and neighborhood dynamics.

You can find more information on Inside Airbnb here: https://insideairbnb.com/about/

A **Data Dictionary** can be found on the above website and is available in the Coursework 1 folder on GCU Learn as well. Here you will find information on the variables in your datasets. Please note that this document also contains information on variables not included in your datasets.

There are 2 raw data files in the Coursework 1 folder on GCU Learn, which you should download: *airbnb1.csv* and *airbnb2.csv.*

## 3. Assignment Instructions

**Objective**: This assignment will guide you through essential data cleaning, wrangling, and exploratory data analysis (EDA) on a Cape Town Airbnb dataset. Your task is to clean, wrangle and explore the data, and develop three research questions that could be answered using it.

**Here is what you need to do:**

**Step 1 - Reading and Understanding the Data**

- ***Load*** in the Airbnb datasets into Google Colab. ***Check*** the first few rows, the shape of the dataset, and column names. Make sure to check the Data Dictionary to find out more information about the column variables in the data.

- ***Merge*** the two datasets into one dataset.
- Provide a ***summary*** of the dataset. This can include things like data types, unique values, and the number of missing values for each column.

**Step 2 - Data Cleaning and Wrangling**

- ***Select*** a number of variables of interest (aim for 20-30 variables) including numeric and string variables. The other variables can be dropped.
    - **Hint**: Drop columns with excessive missing data.
- Identify and handle ***missing values***.
    - **Hint**: Choose the method(s) to handle missing values, and justify your choice. You can choose to analyse all available data, drop rows with missing values, imputing (replacing or interpolating) missing values, or another strategy.
- Check for ***duplicate rows***. Remove duplicates if found.
- Ensure all columns have appropriate ***data types***, and convert columns to their correct types if necessary.
- ***Standardise text columns*** if applicable (e.g., amenities), and/or ***generate new columns*** based on the presence or absence of specific strings or patterns in existing columns.

**Step 3 - Data Exploration**

- Generate ***summary statistics*** for numeric columns, including mean, median, minimum, maximum, and standard deviation.
- ***Plot*** the distribution of key variables (for example, price, review_scores_rating, etc.) to understand their spread and identify potential outliers.
- ***Explore*** the relationship between relevant columns:
    - Create plots to understand relationships between variables (for eample, scatterplots, line charts, box plots etc.)
    - Check for correlations between numeric variables.

**Step 4 - Research Questions**

- Based on your exploration, propose **three research questions** that could be addressed with this dataset. Each question should aim to address and interpret relationships between variables.
- **For each question**, include:
    - The variables that would be analyzed to address the question.
    - The type of analysis you would conduct to address it (for example, correlation analysis, group comparisons, time-series analysis, clustering analysis, predictive modeling etc.).

## 4. Final deliverables

The Coursework submission should be in the format of:

1) a **Jupyter notebook**
2) a **pdf report** generated from the Jupyter notebook.

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.

In the Jupyter (Colab) Notebook, you must document all your code and findings in a well-organized Jupyter Notebook, with clear comments and section headings.

- The notebook must provide a clear demonstration of the steps you have undertaken, detailing both the methods required and the **justification** for undertaking each step.
- At the end of your Notebook, clearly state your **three research questions** with brief explanations on how each question could be investigated using the dataset.

Please take note of the following:

- **Sections and Headings.** Use markdown headings (**#**, **##**, **###**, etc.) to structure the report by creating different sections. For example: "Data Overview," "Data Cleaning," "Exploratory Data Analysis," etc. Add subheadings as needed to organise content.
- Include a **Title** and a brief **Introduction/background** section.
- Don't forget the **Conclusion** at the end! The conclusion should summarise the work done and any findings or insights you have gained. You can also discuss any challenges faced and potential next steps.
- You should include one or more **references**, for example the source of the data or any sources (online, books etc.), a useful article on predictors of good reviews, or any useful sources that helped you write your code.
- **Exporting the Report:** Go to File -> Print - save as PDF.
- Ensure that your code cells are **well-commented!** Comments within code cells are preceded by a hash tag (#).

By following the above steps, you can create a well-structured report from a Jupyter Notebook, combining code and explanations effectively.

Coursework reports should be submitted to GCULearn via Turnitin no later than **Thursday 5 December 2024 17:00 BST**.

## 5. Marking criteria

Coursework 1 is worth 50% of the Software Development for Data Science module assessment. A rubric for Coursework 1 is provided in Figure 2.

Note that this Coursework does not need to include formal modelling (regression, Machine Learning etc.). You will not receive marks for including regression, classification, clustering or other types of Machine Learning. Instead, the focus of this coursework is on understanding, cleaning and wrangling the data, and exploring the data using summary statistics and data visualisations.

| Criteria | Excellent | Good | Satisfactory (50%) | Insufficient | Clear fail |
|---|---|---|---|---|---|
| General Intro and Conclusion [/10%] | Well-written and cohesive introduction, which clearly defines the scope of the work. Information on the data is provided and contextualised. Well-written an cohesive conclusion with good amount of detail. | Good introduction including some background to the dataset and mostly clear, relevant information. Conclusion provided with sufficient detail. | Basic introduction that outlines the work but lacks depth, with several omissions and/or explanations are unclear. Basic conclusions drawn. | Minimal introduction and conclusion that lacks content and clarity. | No introduction and conclusion provided or irrelevant content. |
| Research Questions [15%] | Three insightful, well-defined questions with excellent articulation of variables and analysis types. | Three relevant questions provided. Clear connection to dataset variables and analysis methods. | Three research questions provided with some connection to the data but lacking in detail (e.g. regarding variables or analysis methods). Or only two research questions provided but with sufficient level of detail. | Less than three research questions provided, and/or research questions are vague without clear relevance to the dataset. | No research questions provided. |
| Data Cleaning and Wrangling [30%] | Through data cleaning is demonstrated. Effective variable selection, justifiable handling of missing values and outliers. Duplicates, data types and text standardisation are considered and explained in detail with some theoretical component. | Reasonably thorough data cleaning done and relevant variables selected. Choice of data processing steps with justification and explanation of observations made, with some further clarity/justification required. Relevant variables selected, missing values handled, and duplicates removed. Outliers are considered to an extent. | Basic cleaning done and some relevant variables selected and missing values addressed to some extent.Choice of data processing steps for the most part appropriate but with limited justification and explanation of observations made. | Some cleaning performed, but significant issues persist (e.g., excessive missing data). | No cleaning performed or evidence of very minimal data processing is shown, without justification. |
| Data Exploration [25%] | In-depth exploration with comprehensive statistics, well-constructed plots, and insightful interpretation of data relationships. Some creativity is demonstrated. | Good exploration with detailed statistics, insightful plots, and analysis of relationships. Some further clarity or detail required. | Satisfactory exploration with statistics, plots, and analysis of relationships but with some errors or lacking clarity. | Basic exploration with some summary statistics; few plots presented. | No exploration or analysis conducted. |
| Code (10%] | Excellent code quality; well-organized, fully commented, and includes effective use of functions. | Well-structured code with clear comments and logical flow. | Code is presented for all the main sections in the report and works for the most part with some errors present. Code is mostly commented. | Code runs with errors; minimal use of comments or documentation. | Code is ot show in the report, or code is unreadable or fails to execute. |
| Presentation [/10%] | Highly organized, clear, and professional presentation with hardly any noticeable errors. Formatting and structure of the document is professional and includes headings and clear paragraphs. Reference(s) provided. | Clear organization and professional presentation with minor errors. Information presented in all cases with some improvements in clarity required. Reference(s) provided. | Basic organization; some clarity but lacks professional presentation. | Some organization, but clarity issues detract from understanding. Narrative is unclear. | Presentation is disorganized and unclear. Narrative is hard to follow. Or no evidence that report was created using Jupyter Notebook. |

Figure 2. Coursework 1 Marking Rubric.