School of Computing, Engineering and Built Environment

Software Development for Data Science

Module Code: MMI226822

Coursework 1

Issue date: 1 November 2023

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

Attention is drawn to the university regulations on plagiarism. Whilst discussion of the coursework between individual students is encouraged, the actual work has to be undertaken individually. Collusion may result in a zero mark being recorded for the coursework for all concerned and may result in further action being taken.

Exploratory Data Analysis – House Prices Dataset

1. Introduction

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.

In this coursework you will use Google Colab to generate a Jupyter notebook for exploring a given dataset. You will need to ingest, clean, transform (wrangle), summarise and visualise the data.

2. Dataset

The dataset is a house prices dataset that describes the sale of individual residential properties from 2006 to 2010 in a small town in the USA. Please note that this data set contains some mock-up data.

For this Coursework you will use the Python Data Science methods covered in class to ingest the data, clean and transform the data where needed, perform any other pre-processing required on the data, and explore the data using statistics and visualisation. You will need to detail and appropriately document your processing steps and any meaningful information you can extract from the data.

Note that this Coursework does not need to include formal modelling (regression, Machine Learning etc.). Instead, the focus of this coursework is on understanding, cleaning and wrangling the data, and exploring the data using summary statistics and data visualisations.

The data are included as 2 text files in the Coursework 1 folder on GCULearn. A data dictionary is also provided which described the dataset and the variables included.

Please read the data dictionary carefully, including the 'Special Notes' section at the end which gives useful information on outliers and how to handle these!

3. Instructions

Here's what you will need to do:

1. **Download** the datafiles and data dictionary from GCU Learn/Coursework 1 folder and upload these to your Google Drive.
2. Using a Google Colab Notebook, employ the **Python** libraries that we have covered in class to **ingest, understand, clean, transform, summarise and visualise the data**. The notebook must provide a clear demonstration of the steps you have undertaken to perform exploratory data analysis on the dataset, detailing both the methods required and the **justification** for undertaking each step.

3. Regarding the Notebook in Google Colab, 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.
   - **Data ingestion**: use code cells to load the dataset into a DataFrame (e.g., import pandas as pd and pd.read_csv('your_dataset.csv')).
   - **Data Overview:** use code and markdown cells to provide an initial overview of the dataset (e.g., **df.head()**).
   - **Data Cleaning and Preprocessing:** identify and handle missing values, duplicates, and outliers if present in the data. Use markdown cells to explain the rationale behind each cleaning step.
   - **Exploratory Data Analysis (EDA)**: use basic (summary) statistics and visualisations to explore the data. Use markdown cells to explain insights gained.
   - 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.
   - Please include **references** e.g. background reading on house price data analysis or sources (online, books etc.) that helped you write your code.
   - **Exporting the Report:** Go to File -> Print - save as PDF.
   - Ensure that your code cells are **well-commented** for clarity!

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

4. Final deliverables

The Coursework submission should be in the format of a **Jupyter notebook** and a generated **pdf report**.

Coursework reports should be submitted to GCULearn via Turnitin no later than **Wednesday 6th of December 2023 23.59**.

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 1.

Criteria	Excellent	Good	Acceptable	Substandard	Insufficient
General introduction [/5%]	Cohesive and detailed, well-written introduction and background to the data, with appropriate references. Introduction places the purpose of the work in context.	Introduction and background for the data is mostly complete and contains clear, relevant information. Some information is missing or not cohesive.	General introduction and background to data is provided with several omissions and/or explanations are unclear.	Limited introduction and background of the dataset.	Not shown or not sufficient (eg. no Introduction provided or not suitable).
Data processing choices with justification, and explanations of observations [35%]	Data processing skills are demonstrated in a thorough manner and include some theoretical components. Appropriate choice of data processing steps with justification and explanation of observations made. Clear explanation of how missing values, duplicates and outliers (if present) were handled.	Appropriate choice of data processing steps with justification and explanation of observations made, with some further clarity/justification required. Missing values, duplicates and outliers are considered.	Choice of data processing steps for the most part appropriate but with limited justification and explanation of observations made. Some further clarity/justification required. Missing values and outliers are considered to an extent.	Evidence of minimal data processing is shown, but the choices are not justfied and lack detail.	Not shown or not sufficient.
Exploratory Data Analysis [35%], incl. summary statistics and data visualisations	Clear presentation of summary statistics and data visualisation with explanatiomns. Compelling and well-formatted plots . Creativity is demonstrated.	Adequate presentation of summary statistics and visualisations with evidence of formatting, but with some further improvements or clarity required.	Acceptable summary statisics and/or data visualisations presented, with some errors or further clarity required.	Minimal summary statisics and/or data visualisations presented, with errors.	Not shown or not sufficient.
Conclusions [5%]	Well-written and detailed conclusion provided, including a summary of key steps and some discussion around implications and/or recommendations for further analysis.	Conclusion provided with sufficient detail, with some further detail/clarity required.	General conclusion provided, with some omissions and lacking detail.	Limited conclusions shown, or the conclusions do not follow from the report.	Not shown or not sufficient (eg. no Conclusion section is provided, no conclusions drawn).
Python Code (10%]	Clearly formatted and commented code which works correctly. Code is commented and reproducible.	Code is provided for all steps and works well, with commenting. A few errors or further clarity required.	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 missing in places, or code contains lots of errors. Code is not commented and/or errors in coding style.	No evidence of Python code, or code not working.
Presentation of information [/10%]	All information clearly presented. Formatting and structure of the document is professional and includes headings and clear paragaphs. References provided.	Information presented in all cases with some improvements in clarity required. Well-structured and formatted report incl headings and paragraphs. References provided.	Information presented for most of the material, with some omissions; formatting and structure mostly fine incl headings and paragraphs.	Some attempt at formatting the report is evident, but with several errors or omissions in the presentation of information . Narrative is hard to follow, paragraphs are poorly organized.	Not sufficient. No evidence of formatting. Narrative is unclear, paragraphs are poorly organized; or no evidence that report was created using Jupyter Notebook.

Figure 1. Coursework 1 Marking Rubric.