Exploratory Data Analysis – Ames House Price Dataset

Figure 1 – Ames Iowa Main Street

In this coursework you will utilize Google Colab to generate a Jupyter notebook for analysing the Ames Housing dataset. The dataset describes the sale of individual residential property from 2006 to 2010 in Ames (a small town in Iowa, USA). It was compiled by Dean De Cock who wrote a paper on this (which you should read). More details of this dataset are described in the paper: *Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.*[1]

This CW will require you to read in the data, perform any pre-processing required on the data, and utilise the Python Data Science Software Ecosystem to undertake and Exploratory Data Analysis of the House prices detailing and appropriately documenting any information you can extract from the data.

The dataset is included as a text file within the Coursework section on GCULearn.

Jupyter Notebook

This notebook should provide a discussion and demonstration of the steps you have undertaken to perform exploratory data analysis on the dataset detailing both the software required and the reasoning for undertaking each step.

You should discuss:

- The formatting of the data
- Data cleaning
- Exploratory Data Analysis techniques to describe the data
- Data visualization.

[1] De Cock, D., 2011. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. Journal of Statistics Education, 19(3).

Coursework reports should be submitted to GCULearn via Turnitin no later than **23.59 on Monday 25th July 2022.**

Please note that only one submission can be made so please make sure you are happy with your notebook before you submit.

A rubric for this component of the coursework is provided below:

Rubric	Not attempted / insufficient (0-10%)	Basic (10-40%)	Moderate (40-60%)	Good (60-80%)	Excellent (80-100%)
Presentation of information [/20%]	Not shown or not suitable	Some explanations of the the exercise undertaken.	Explanations for most of the material, with some omissions.	Information presented in all cases with improvements in clarity required.	All information clearly presented.
Details of Python experiments [/40%]	Not shown or not suitable	Some EDA code is presented, but limited funcationality or not functional.	Use of EDA tools in Python demonstrated, limited data descriptions.	Use of Python Data Science ecosystem software to describe the dataset employing EDA analysis with some reporting/ visualization.	Appropriate use of Python Data Science ecosystem software to fully describe the dataset inlcuding robust analysis and appropriate data visualizaiton. Code fully documented.
Data processing choices and explanations of observations within the notebook [/40%]	Not shown or not suitable	Marginal theoretical components introduced.	Some components introduced, yet details required are missing.	All required explanations with some further clarity required.	Appropriate choices with detailed explanations, where required all observations are back up with approriate theory, including robuts analysis.

Figure 2 - Coursework 1 Report RUBRIC