COMP41680/COMP47670 Assignment 2

Summary:

The objective of this assignment is to collect historical house sale data from an online source and then perform a number of data analysis tasks on this data.

The three tasks listed below should be implemented in a single Jupyter Notebook (not script file). Your notebook should be clearly documented, using comments and Markdown cells to explain the code and interpret the results of your analysis.

Tasks:

1. Data Collection & Preparation

a) Scrape all of the house price data from the web page:

http://mlg.ucd.ie/modules/python/housing/

Parse the HTML to create a DataFrame representation of the data, including all of the descriptive features which represent house sales.

b) Perform an initial characterisation of the dataset to identify any data quality issues.

c) Apply appropriate preprocessing steps to address all of the data quality issues that were identified in Task 1(b).

2. Feature Associations & Regression

a) Analyse how house sale prices relate to the other numeric features in the dataset.

b) Analyse how house sale prices relate to each of the categorical features in the dataset.

c) Investigate the use of simple linear regression to predict house sale prices, based on each of the individual numeric features in the dataset. Which numeric feature appears to be most useful when predicting prices?

3. Classification

a) The price of a property is often said to be linked closely to its location, while different areas will have different types of housing stock.

Investigate whether it is possible to classify the location of a house, based on the other descriptive features in the house sale dataset. You can use any classification algorithm of your choice. You should evaluate the performance of the classifier using an appropriate strategy.

b) Experiment with applying the same classifier in combination with different subsets of descriptive features. Which feature(s) appear to be most useful

for classification?

Guidelines:

- The assignment should be completed individually. All submissions will be subject to plagiarism checking. Any evidence of plagiarism will result in a 0 grade.
- The grade awarded will depend on the complexity of the analysis and level of detail, i.e. data collection and preparation, analysis, interpretation etc.
- Submit your assignment via Brightspace. Your submission should be in the form of a single ZIP file containing your notebook file (.IPYNB).
- In your notebook please clearly state your student number.
- Late penalties will apply 1 week after the deadline (28th April).
  Penalties for late submissions:
  - 6-10 working days late: 2 grade point deduction, e.g. B to C+
  - Assignments will not be accepted later than 10 working days without Extenuating Circumstances formally approved by UCD.