2019S2 BUSS6002 Assignment 1

Due Date: Friday 27 Sep 2019

Value: 15% of the total mark

Instructions

1. **Required Submission Items:**
    1. **ONE** written report (PDF format). submitted via Canvas.
        - Assignments > Report Submission (Assignment 1)
    2. **ONE** Jupyter Notebook .ipynb submitted via Canvas.
        - Assignments > Upload Your Code File (Assignment 1)
2. The assignment is due at **12:00pm (noon) on Friday, 27 Sep 2019**. The late penalty for the assignment is 5% of the assigned mark per day, starting after 12:00pm on the due date. The closing date **Friday, 4 Oct 2019, 12:00pm (noon)** is the last date on which an assessment will be accepted for marking.
3. As per anonymous marking policy, please include your Student ID only in the report and do **NOT include your name**. The name of the report and code file must follow: **SID_BUSS6002_Assignment1.** Failing to name your submitted files correctly would incur a penalty.
4. Your answers should be provided as a final report giving full explanation and interpretation of any results you obtain. Output without explanation will receive **zero** marks. You are required to also submit code that can reproduce your reported results, as reproducibility is a key component to data science. Not submitting your code will lead to a loss of 50% of the assignment mark.
5. Be warned that plagiarism between individuals is always obvious to the markers of the assignment and can be easily detected by Turnitin.
6. Presentation of the assignment is part of the assignment. There will be 10 marks for the presentation of your report and code submission.
7. The report should be **NOT more than 10 pages** including text, figures, tables, small sections of inserted code etc. Think about the best and most structured way to present your work, summarise the procedures implemented, support your results/findings and prove the originality of your work. You will provide your code as a separate submission to the report; however, you may insert small sections of your code into the report when necessary.
8. Your code submission has no length limit, however marks are assigned for code presentation, so make your code as concise as possible and add comments when necessary to explain your logic and the purpose of each code segment. Make sure to remove any unnecessary code and ensure that your code can be run without error.
9. Numbers with decimals should be reported to the **third-decimal** point.

Project Description and Dataset

Suppose you are working as a Data Scientist for a real estate investment firm. The firm is assessing locations for investing in housing redevelopment in the United States. For this purpose, the firm has identified several potential locations in Seattle to purchase existing houses, which would be demolished to make space for the redevelopment.

In order to estimate the costs involved the firm needs to know the current market value of the houses that it needs to purchase. You are working on a project that aims to build a model to estimate the house prices.

Seattle's Department of Assessments has been collecting data since 2014 on house sale prices and the characteristics of each house that was sold. You have been given access to a copy of original database "house.db", which is an SQLite file, as well as a data dictionary file "house_dict.txt". You can download the dataset and detailed dataset description from the BUSS6002 Canvas site.

Hint: To list all tables in the database you can use the following query

```
SELECT name FROM sqlite_master WHERE type='table' ORDER BY name;
```

Task 1

To start your analysis, you wish to perform a thorough EDA to help you better understand the given datasets. The results you obtain in this task will be used to inform your modelling choice.

Requirements:

a. Check and deal with any missing data (if any) in the given dataset.
b. Look for and remove any potential outliers (if any) that would possibly affect your modelling. Justify your answer.
c. Visualise the relationships between explanatory variables and the target variable through appropriate plotting. Report your analysis and findings.

Task 2

Suppose now you want to build a prototype model to predict house sale prices, which will be demonstrated to a wider team. Therefore, it needs to be easily understood by non-experts, meaning that you can only use a few variables in your model as a starting point.

In order to make informed decisions on your modelling choices, you need to answer the following questions:

a. Suppose you would like to build a linear regression model to predict house sale prices, do you wish to include an intercept term in your model? Carefully explain your answer.
b. Do you think multicollinearity could be a potential problem on the given dataset? Use your understanding of variables to justify your answer and verify your

hypothesis using appropriate numeric measures. Explain your decisions to proceed based on your findings.

c. If you wish to use only three variables to predict house sale prices, which three variables would you choose? Carefully justify your choice and explain your selection criterion.
d. Build a linear regression model using the three variables you have chosen (Use original, i.e. not engineered, variables for this task). Report and interpret your regression results.
e. Perform residual diagnostics to measure the goodness of fit. Report your findings.

Task 3

The model you have built so far provides an approximate estimate of house prices. However, to accurately estimate the costs of the redevelopment plan you must be able to estimate house prices as accurately as possible.

Your goal is now to improve your model as much as you can through feature engineering and feature selection. You may consider all variables and apply appropriate transformation to the variables as necessary.

Requirements:

a. Your model should have a minimum adjusted R-Squared of 75%. If your modelling cannot achieve an adjusted R-Squared of 75%, report the best model you can obtain.
b. Justify your choice of feature engineering strategies using EDA and present your results.
c. Compare your new model with the model you have built in Task 2 with respect to Adjusted R-Squared. Explain why you should use Adjusted R-Squared here to compare the two models.
d. Provide residual analysis to justify why your new model is more reasonable.

Task 4

Suppose you have finished your analysis, now you need to report to your manager and reflect on what you have experimented with in your project:

a. Provide a reflection of how you have utilised the data science process model to arrive at modeling and model evaluation based on how you answered the previous three questions. Choose only one process model (CRISP-DM or Snail Shell) to answer this question. Explain how each part of the questions aligns with the different phases of the process model.
b. The firm is also considering redevelopment projects in other locations. Comment on whether the model you have built can or cannot be applied in other locations. Justify your answer.

Marking Outline

Tasks	Marks
Task 1	20 marks
Task 2	30 marks
Task 3	30 marks
Task 4	10 marks
Report and Code Presentation	10 marks