BUSS6002 Assignment 1

Due Date: Tuesday 16 April 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 Tuesday, 16 April 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 **Tuesday, 23 April 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 Students ID only in the report and do **NOT include your name**. The name of the report and code file must follow: **SID_BUSS6002_Assignment1_S12019.**
4. Your answers shall be provided as a word-processed report giving full explanation and interpretation of any results you obtain. Output without explanation will receive **zero** marks. You are required to also submit your 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 the functionality of your code segments. 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**.

Tasks

Suppose the year is 2010 and you are working as a Data Scientist for an investment firm. The firm is assessing locations for investing in housing redevelopment in the United States. The firm has selected Ames, Iowa as a candidate location. As a consequence, the firm would need to purchase existing houses, which would be demolished to make space for the development.

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

The Ames City Assessor's Office has been collecting data since 2006 on house sales and the characteristics of each house that was sold. You have been given access to a copy of original database "housing.db", which is an SQLite file. The Assessor's Office have also provided you with a data dictionary "housing_data_description.txt".

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;
```

You can download the dataset and detailed dataset description from the BUSS6002 Canvas site.

Question 1

To start your analysis, you wish to build a prototype model that 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.

To save you time, an experienced member of your team suggests to you that from their experience the above ground living area, basement size and the age of the house are most useful variables.

Perform EDA to determine which two of these features are most useful. Carefully explain your selection criteria and present the results to justify your choice.

Requirements:

a. To most accurately reflect the conditions under which the firm will purchase the houses you should limit your analysis to houses that are sold under normal conditions.
b. If you find any missing values in the relevant variables, then remove the affected observations.

Question 2

Suppose you are interested in using the above ground living area and basement size to estimate the price of a home.

a. Build a linear regression model WITHOUT an intercept term (MODEL1), write down the mathematical model and report the regression output.
b. Build a linear regression model WITH an intercept term (MODEL2), write down the mathematical model and report the regression output.
c. Compare the performance of the two models and explain the role and impact of the intercept term
d. Pick either MODEL1 or MODEL2 that you think is preferable and perform residual diagnostics to measure the goodness of fit. Report your findings.

Adhere to the same requirements from Question 1.

Question 3

The models you have built so far provide 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 possible through feature engineering and feature selection.

Instructions:
a. Your model should have a minimum adjusted R-Squared of 77%. If your modelling cannot achieve a adjusted R-Squared of 77%, report the best model you obtain.
b. Justify your choice of feature engineering strategies using domain knowledge or EDA and present your results.
c. Compare your new model with the preferable model in Question 2 with respect to Adjusted R-Squared. Explain why you should use Adjusted R-Squared here to compare the two models.
d. Provide analysis to justify why your new model is more reasonable.

Hint: Carefully read the data dictionary. In this dataset the “NA” (Not Applicable) code is represented by a NULL value in the database, which will be interpreted by Pandas as NaN. This means that the NaN values do not indicate whether the value is missing. For example, the “Garage Type” variable codes “No Garage” as “NA”, consequently it will be interpreted as NaN by Pandas.

If you wish to include such variables in your analysis, then you should recode NaN as a valid code. For example, to recode the Garage Type you could use:

```
df['Garage Type'] = df['Garage Type'].apply(lambda x: "NoGarage" if
pd.isnull(x) else x)
```

Question 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 data science project:

a. Provide a reflection of how you have utilized 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 you choose to answer the question.

a. 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

Questions	Marks
Question 1	20 marks
Question 2	20 marks
Question 3	40 marks
Question 4	10 marks
Report and Code Presentation	10 marks