Big Data Analytics

Assignment 1

1 Description

The housing dataset contains the prices and other attributes of almost 35,000 houses in the city of Melbourne. Your task is to perform an Exploratory Data Analysis on the dataset. This assignment is part of the continuous assessment and worth **30%** of your module grade.

2 Dataset

First download the dataset from Moodle's module page. To load the dataset, run the following commands:

```
# read the csv file
housing.dataset <- read.csv("<full path>/melbourne_data.csv")
```

To get started... see the structure of the data using: `str(housing.dataset)`. Below is some brief details of the dataset variables:

Rooms: Number of rooms
Price: Price in Australian dollars
Date: Date sold
Type: House type (h=house, u=unit/duplex, t=townhouse)
Distance: Distance from Central Business District in KMs
Regionname: General Region (West, North West, North, etc.)
Propertycount: Number of properties that exist in the suburb
Bathroom: Number of bathrooms
Car: Number of carspots
Landsize: Land size in Metres
BuildingArea: Building size in Metres
YearBuilt: Year the house was built

3 Task

Your task is to perform EDA and calculate the strength of relationships between the variables of the dataset. Consider below as a guideline:

1. Your first task is to clean the dataset and prepare it for analysis by e.g. removing/replacing NAs, outliers, and incorrect values. **(20 points)**
2. Begin your analysis with a summary of the variables (use basic statistical methods). Briefly describe your understanding. Prepare 4 plots: pie chart, bar chart, histogram, scatter plot. Each plot should display different variables. Each plot must have a title and meaningful labels. **(20 points)**
3. Focus your analysis on the price variable: **(20 points)**
   (a) Show the histogram of the price variable. Describe it briefly. Include summary statistics like mean, median, and variance.
   (b) Group houses by some price ranges (like low, medium, high, etc.) and summarise those groups separately.
   (c) Explore prices for different house types. You might want to use the `boxplot`.
   (d) How different attributes are correlated with the price? Which 3 variables are correlated the most with price?
4. List the frequencies of houses for various types. Create 2 scatter plots and colour the house price by landsize and type. **(10 points)**
5. In your Markdown document, you should use proper headings and commentary for each task. You can get up to **30 points** for clarity and quality of the report and the source code. You must show the final code blocks and the corresponding output in your knitted document.

Keep in mind the following...

- Acceptable file format: Knit your Markdown document in **pdf** output. Use the submission link on Moodle to upload your final pdf report.
- There will be a late submission penalty as per the College policy. I will be very strict on plagiarism. You may want to read Griffith's plagiarism policy. I will be awarding ZERO to both the Copyier and Copyee.