Big Data Analytics

Assignment 1

1 Description

The dataset for this assignment contains the prices and other attributes of 50,000 diamonds. Your task is to perform an Exploratory Data Analysis on the dataset. Submit an R Markdown report summarising your findings together with the source code. Check Moodle for deadline. This assignment is part of the continuous assessment and worth **30%** of the module grade.

2 Dataset

First download the dataset from Moodle. As the dataset contains 50K records, generating the plots may take a few moments. One way is to start with a small sample and carry out analysis, for example, you can pick 10,000 observations (without replacement) using the function: `sample`. Run the following code to do so:

```
s <- sample(nrow(diamonds.dataset), size=10000, replace = FALSE, prob = NULL)
diamonds.subset <- diamonds.dataset[s, ]
```

The above piece of code creates a new dataset named: `diamonds.subset` containing 10,000 observations from diamond dataset. You can use the sampled dataset (`diamonds.subset`) first to write and test your code. And then use the full dataset for completing the task given below. REMEMBER! You must report your findings on the full dataset. In your final report, there is no need to include your findings on the sampled dataset. When you load the dataset, you will find the following variables in the dataset:

carat: weight of diamond (0.2 to 5.01)
cut: quality of the Cut (Fair, Good, Very Good, Premium)
color: diamond color from D (Best) to J (Worst)

clarity: a measurement of how clear the diamond is from I1 (Worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (Best)
table: width of top of the diamond
x: length in mm
y: width in mm
z: depth in mm
depth: total depth percentage = 2*z/(x+y)
price: price in US dollars

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 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 (do not use `price` variable now). 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 diamonds by some price ranges (like low, medium, high, etc.) and summarise those groups separately.
   (c) Explore prices for different cut 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 diamonds for various cuts and clarity levels. Create 2 scatter plots and colour the diamonds price by clarity and cuts. **(10 points)**
5. Now focus your analysis on the carat, depth, table and dimensions (x, y , z) variables: **(20 points)**

(a) Compute a `volume` variable from x, y, z - add it to the dataset. Plot it against the price. Describe your findings.

(b) Are the carat and volume attributes correlated? Is that a strong relationship? Draw a plot with regression line.

(c) Explore the relationships between table and depth variables.

(d) Now explore relationships between table and rest of other variables. Compute correlations and describe your findings.

6. In your Markdown document, you should use proper headings and commentary for each task. You can get up to **10 points** for clarity and quality of the report and the source code.

Keep in mind the following...

- Acceptable file formats: R Markdown document (.Rmd) and its pdf output. Submit these two files as a zipped folder.
- 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. If found, I will be awarding ZERO to both the Copyier and Copyee.