Big Data Analytics

Assignment 2

1 Description

The dataset for this assignment contains the prices and other attributes of 50,000 diamonds. A pre-processed and cleaned version of the dataset is made available on Moodle for download. Your task is to perform hypothesis testing, regression and classification on the dataset. Submit an R Markdown report summarising your findings together with the source code. Check Moodle for the deadline. This assignment is part of the continuous assessment and worth **30%** of your 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. You might be familiar with the dataset already, 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 Tasks

Complete the following tasks.

3.1 Task A

Before you start statistical analysis, you have to define hypotheses, which will be tested. You should state at least 2 different hypotheses, each to test different data (so not all hypotheses should be checking the same statement just on different variables). Remember that there are different types of tests and you should use as many as you can (given if they are valid and make sense). Your ultimate goal is to report some findings. You should also prove that these findings are statistically correct. Take the below points as hints but do not limit yourself to these:

- Look at different plots you have created during exploratory analysis. What conclusions can be drawn based on these? These could become your hypotheses.
- If you focus on one attribute, what is your intuition about the distribution that could explain such results? You can check and measure how well the data fits some distribution.
- For each valid hypothesis test you will get 15 marks. This section consists of **30 marks** in total.

  Remember that data analysis is not only about finding and proving hypotheses but also about summarising data and communicating it. It is not a failure if you do not get "significant" results, you still have to

report that. If your analysis makes sense (e.g. it is valid from the statistical point of view), there is no such thing as a bad result. Present your analysis in the form of a report. Each hypothesis should be described, you should state what you want to prove. If you are claiming that groups have different characteristics, first show these on plots and comment on them. Report should be written in a way that a person without prior knowledge of the data is able to follow it.

3.2 Task B

- Divide the dataset into training and test data. Use 75/25 split.
- Perform Linear Regression with Multiple Variables to predict the diamond price.
- Report adjusted R squared (on training data). Use RMSE and correlation to report the prediction accuracy of the test data.
- Normalize the data and repeat the process of performing Linear Regression with Multiple Variables on normalized data to predict the diamond price.
- Highlight the difference in prediction accuracy of both models.
- Write your findings in this section. Each valid iteration Linear regression, will get you 15 marks. This section consists of **30 marks** in total.

3.3 Task C

- Divide the dataset into training and test data. Use 80/20 split.
- Use kNN to classify diamond cuts into appropriate types based on their features.
- Use C5.0 to classify diamond cuts into appropriate types based on their features.
- Use ANN (hidden=5) to classify diamond cuts into appropriate types based on their features.
- Compare the (best) performance of each classifier.
- Write your findings in this section. Each valid classification technique, will get you 10 marks. This section consists of **30 marks** in total.

Keep in mind the following...

- You can also get up to **10 points** for clarity and quality of the report and the source code.
- Acceptable file formats: R Markdown document (.Rmd) and pdf. Zip both files together and submit.
- Your R Markdown document must compile correctly into html and pdf formats.
- Do not submit work thatffs not your own. Do not let others copy work that is your own. Both Copyier and Copyee will get ZERO marks.