Pricing of Diamonds

R Script write up

May 2023

1 Libraries

For this process of data cleaning four libraries are put in use and these include;

- tidyverse
- ggplot2
- readxl
- dplyr

2 Processes

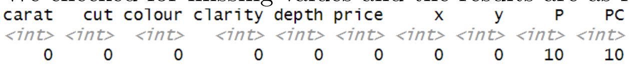
2.1 Analysing the data

This involves reading the data from the any data file provided ie.csv or excel. Further, we went a head to check for the data types of the different features in our data. using the "typeof" function.

Then we removed unwanted columns in our data. Initially, "ID" is not needed and we therefore remove it.

2.2 Checking for missing values

We checked for missing values and the results are as follows



```
carat   cut colour clarity depth price     x     y     P    PC
<int> <int>  <int>   <int> <int> <int> <int> <int> <int> <int>
    0     0      0       0     0     0     0     0    10    10
```

We realise that P and PC have missing values.

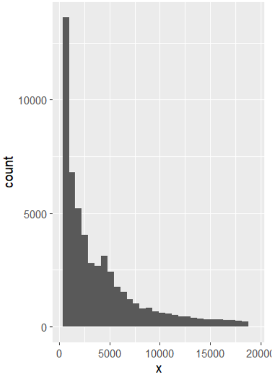
2.3 Grouping the data according to the different types

The different types of data we mean by this, is continuous and categorical data.

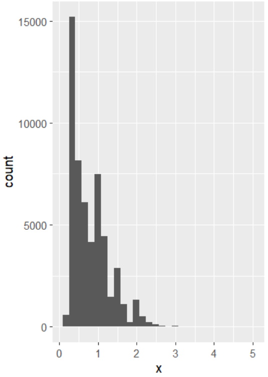
2.4 Plotting histograms for the continuous variables

This helps us to easily know the measure of central tendency that we shall be applying when performing imputation ie Mean or Median for normal distribution and skewed distribution respectively.

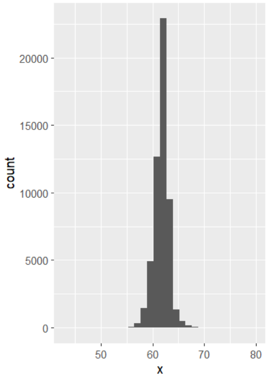
2.5 Price



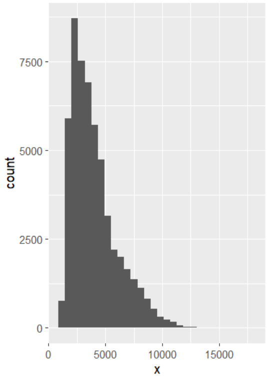
2.6 Carat



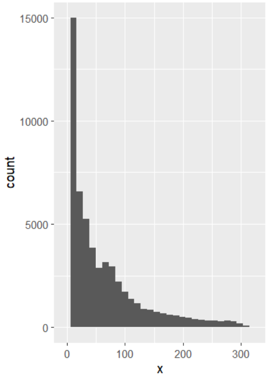
2.7 Depth



2.8 X



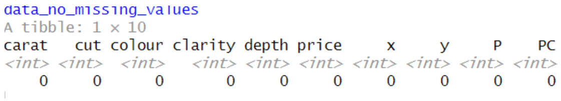
2.9 Y



2.10 Removing missing values

Since P and PC have type of character (categorical data), we are going to impute with the most occurring value (mode).

2.11 Checking to see if all missing values are out



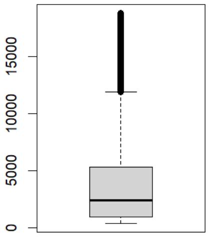
```
data_no_missing_values
A tibble: 1 × 10
carat   cut colour clarity depth price     x     y     P    PC
<int> <int>  <int>   <int> <int> <int> <int> <int> <int> <int>
    0     0      0       0     0     0     0     0     0     0
```

2.12 Plotting

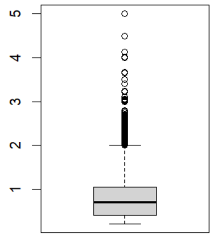
3 Checking for outliers.

We used box plots to visualize the outliers for the continuous features in our data And the results are ;

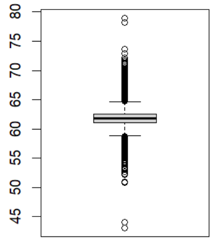
3.1 Price



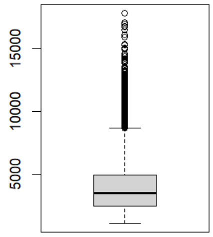
3.2 Carat



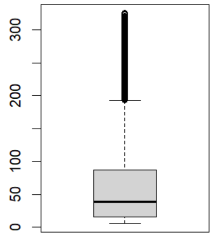
3.3 Depth



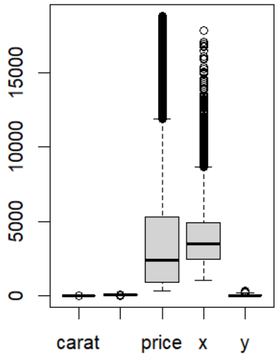
3.4 X



3.5 Y



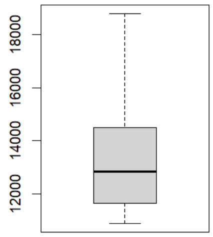
3.6 General plot for all continuous data



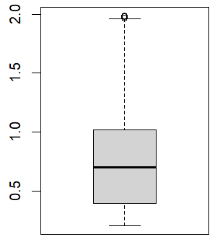
4 Removing the outliers

We used the inter quantile range IQR method to remove the outlier and then using box plots to visualize the data without outliers.

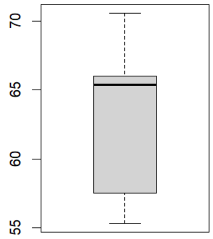
4.1 Price



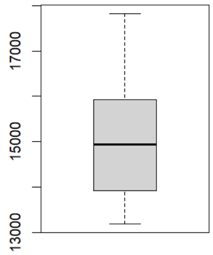
4.2 Carat



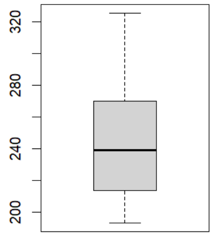
4.3 Depth



4.4 X



4.5 Y



5 Finding the relationship between the variable

This can be done through plotting of scatter plots to find the relationship between the categorical- categorical variables.
We can also use the t-test or the anova test to find the relationship between continuous - categorical variables
We can plot a frequency polygon to show the relationship between categorical - categorical features. [1]

[1] muganga charles