HackStat 2018 - Round 1
Case study- Diamonds dataset

Institute: Sri Lanka Institute of Information and Technology

Group name: Geek Gods

Group Members:

- S.M.A.M. Manchanayake (madhawa242@gmail.com)
- Gowshalini Rajalingam (gowshalinirajalingam@gmail.com)
- Y. N. Senanayake (yohanneranga40@gmail.com)
- L.P.J.Perera (pjayendra101@gmail.com)

Questions

1. **Identify a research question/s and describe the objective/s.**

Research question:-

Predicting the category (High, Medium, Low) or the type of the price with relevant to the attributes of a diamond.

Attributes-→ carat, depth, table, x, y, z, cut, color, clarity

Describe the objectives:-

- pre-processing data
    a) change the data types into correct data types, to get an accurate prediction
        - In the given CSV file price is in Integer data type, change it into float
    b) Handle missing values.
        - In this data set there is no missing values.
    c) Assigning discrete values to categorical data (Ordinal variables)
        - Cut → (Fair=1, Good=2, Very Good=3, Premium=4, Ideal=5)
        - Clarity→ ("I1"=1, "SI2"=2, "SI1"=3, "VS2"=4, "VS1"=5, "VVS2"=6, ("VVS1"=7, "IF"=8)
        - Color→ ("J"=1, "I"=2, "H"=3, "G"=4, "F"=5, "E"=6, "D",7)

- Divide the data set into training data set and test data set
- check using descriptive statistical approach where the model could be parametric or non-parametric
    - Check for distributions of each attribute (both categorical numeric) to check weather that it is normally distributed. (Using normal test, displots, histograms, Box plots, Heat map)

      Most of the attributes are not normally distributed, there are ordinal variables. And there are outliers (Using Box plots) also, because of all these facts we have to go for a non-parametric approach.
- Predict the test set
    - The test data is predicted using Logistic Regression.

2. **Which statistical analysis (and/or modelling) technique/s do you suggest?**

This figure shows GLM (Generalized Linear model)

Model	Random	Link	Systematic
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

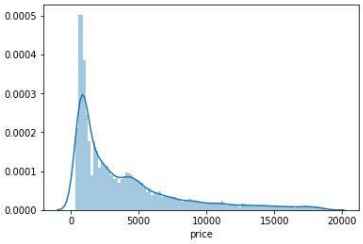
- We have proved that the data are not normally distributed.
- The data set is not time series data or not related with time. So, random variable is not Poisson distribution.
- As there are more than 2 Random variable values (not binary values) the random variable is not binomial distribution.
- As there are more than 2 Random variable values (not binary values) the random variable is Multinomial.
- Therefore, the model is Multinomial Response (Multinomial Logistic Regression).

3. **Interpret your findings to support your objective/s.**

Descriptive Analysis

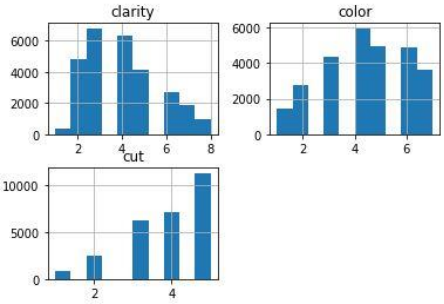
I. Check the distributions

- check dependent (Price) variable



Price variable is positively skewed (left skewed) so not normally distributed.

- check for independent variables
  - i. Using histograms (cut, color, clarity)
    - ❖ Checking for Categorical variables (cut, color, clarity)



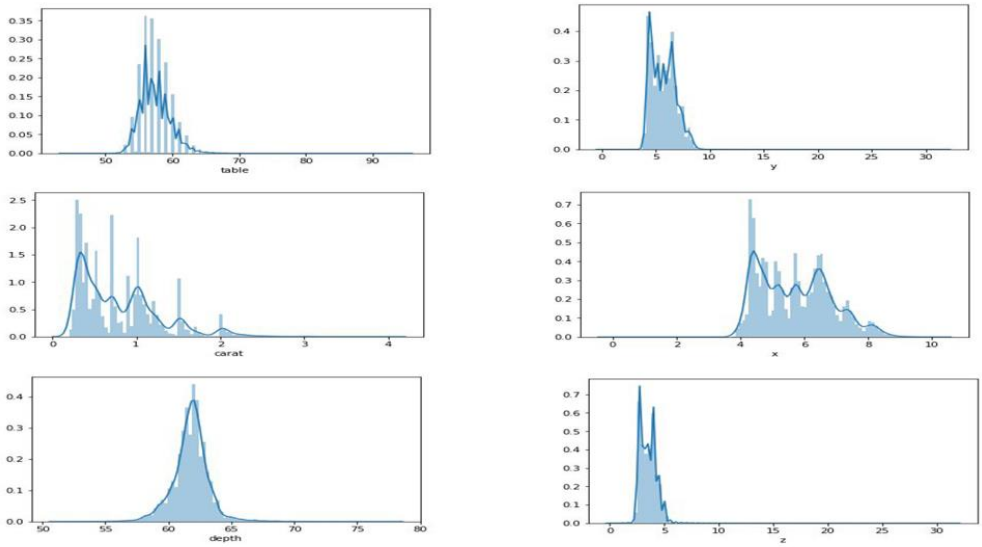
Clarity, color and cut are not normally skewed

Clarity--> left skewed (positive skewed)

Color and cut ---> right skewed (negative skewed)

- ❖ checking for numerical (carat, depth, table, x, y, z)

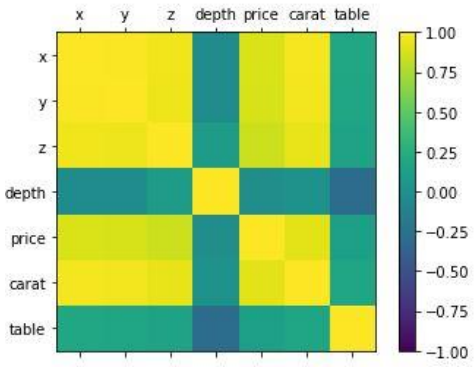
Distributions for carat, depth, table, x, y and z



As we can see from the above graphs most of the variables are not normally distributed, so we have used a non-parametric approach. As we can see for the depth variable it seems it has a normal distribution when we

look at the graph, to ensure we have applied normal test and from the test it says depth also not normally distributed.

II. To check the importance of the variables with the dependent variable and each other we have plot a correlation table and a correlation matrix using pivot.



	x	y	z	depth	price	carat	table
x	1.000000	0.988568	0.954324	-0.023719	0.884686	0.975607	0.196747
y	0.988568	1.000000	0.946774	-0.027050	0.877081	0.964718	0.187377
z	0.954324	0.946774	1.000000	0.095539	0.846108	0.937111	0.149699
depth	-0.023719	-0.027050	0.095539	1.000000	-0.012142	0.029175	-0.294196
price	0.884686	0.877081	0.846108	-0.012142	1.000000	0.920970	0.129027
carat	0.975607	0.964718	0.937111	0.029175	0.920970	1.000000	0.184032
table	0.196747	0.187377	0.149699	-0.294196	0.129027	0.184032	1.000000

x, y, z and the carat variables are highly related with the dependent variable (price). Since they have correlation values closer to the 1, and depth and table variables are not related to the price, but when we model the depth and table with other variables there significance also has improved.

Predictive Analysis

- Divide greater than 0.8 percent of data to training set and rest to test set
- Converted the continuous dependent variable to categorical variable by binning.

  Bin width =6165.
  Number of bins =3
  Range: 325-6491 => Low price
  6492-12657 => Medium price
  12658-18824 => High price
- Predict the test set using multinomial logistic regression. The accuracy is 94.06%

```
In [195]: accuracy_score(Ytestdf, yhat)
     ...:
Out[195]: 0.9406168924762179
```

4. **What are the limitations of your study?**
- The Random variable price has been categorized as Low, medium, high values as our model is using a logistic model.
- If we can increase our training dataset we can improve the our prediction more.

References:

- Choosing Between a Nonparametric Test and a Parametric Test
- *From <http://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test>*
- Introduction to Generalized Linear Models
- *From <https://onlinecourses.science.psu.edu/stat504/node/216/>*