# Hypotheses, Regression, Clustering and Predictions

*Prateek Chauhan*

*18 December 2016*

# Basic summary

The diamonds dataset chips with ggplot2 and contains the prices and specs for more than 50 thousand diamonds collected from diamondse.info between the years of 2008 to 2014. It consists of 53940 observations of 10 variables. Below we can see the basic structure of the data set.

```
str(diamonds)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    53940 obs. of  10 variables:
##  $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

---

# Hypotheses Testing

Before diving into this report, first it has to be mentioned, what Hypotheses Testing is. A **hypotheses test** is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population. A **hypotheses test** examines two opposing hypotheses about a population: the *null* hypotheses and the *alternative* hypotheses.

There are various kinds of hypotheses tests available on different statistical measures, such as **mean**, **median**, **variance** etc.

## 1. Z Test

A Z-test can determine if there is a statistically significance difference between a sample mean and a population mean with known population standard deviation.

The term **population** in statistics includes all members of a defined group that we are studying or collecting information on for data driven decisions. A part of the population is called **sample**.

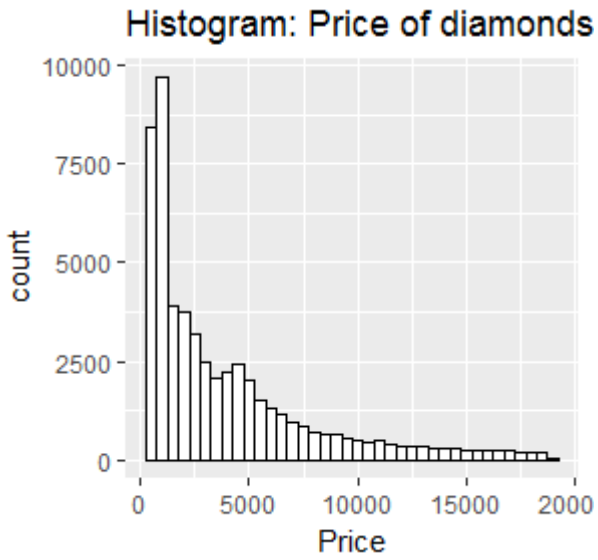
The Z-test uses z-score to determine the probability that the sample mean is drawn randomly from a known population. If the test fails, the conclusion is that random sampling is likely to have produced this. If the test rejects the null hypothesis, then the sampling is likely to be a result of non-random sampling.

Here we are going to perform a one-mean z-test for which we will use one-tail hypothesis test. The null hypotheses will be that there is no difference between the sample mean and the population mean. The alternate hypotheses will test to see if the sample mean is greater.

- $H_0 : \bar{x} = \mu$
- $H_A : \bar{x} > \mu$

Let's have a look at a plot, a histogram for the variable 'price' in the diamond dataset.

Plot:




Now, here two sets of sample from the data set are created: one is entirely random and the other is heavily weighted towards higher prices of diamonds. The null hypothesis will be that in both the sets, there's no difference between the sample mean and the population mean. The alternative hypotheses will be that the sample mean is greater than the population mean. Here are two sets of an n=300 sample and R code to construct them.

```
#unbiassed
price_sample <- sample(diamonds.price, size = 300)
sample_mean <- mean(price_sample)

#biased - higher price

cut <- 1:53940
weights <- cut^.6
sorted_price <- sort(diamonds.price)
price_sample_biased <- sample(sorted_price, size = 300, prob = weights)
```

```
11 sample_mean_biased <- mean(price_sample_biased)
```

The population mean is 3932.8, the mean of first unbiased sample is 4258.57 and the mean of biased group is 4695.50. Both samples are higher than the population mean, but are both significantly higher than the mean? To figure this out, z-stats is calculated and is compared with the critical value using the equation:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

```
1  > #unbiased
2  > #z-stat caclulation
3  > sample_mean
4  [1] 4258.87
5  > z<- (sample_mean - pop_mean)/(pop_sd/sqrt(300))
6  > z
7  [1] 1.415676
8
9  #z-stat
10 > sample_mean_biased
11 [1] 4695.493
12 > z<-(sample_mean_biased - pop_mean)/(pop_sd/sqrt(300))
13 > z
14 [1] 3.311333
```

The critical value for a one-tailed z-test at 95% confidence interval is 1.645. We can determine the sample for unbiased random sample is not significantly different, but the higher price biased sample is significantly different. This is because the z-stat for the unbiased sample is less than the critical value, while the higher price biased is higher than the critical value.

```
1 > #Calculating the p-value
2 > pnorm(z)
```

The p-value for the unbiased sample is 0.1426 or there is a 14.26% chance that the result obtained was due to random chance, while for the biased sample, the p-value is 0.9999 or 99.99% chance of being a result of random sample selection. We have taken the level of significance as 0.05; therefore in both the cases the null hypothesis is retained. But, for the biased case, as the p-value is extremely high, it means that even though we cannot reject the null hypothesis, the result is not significant. A high p-value can't be used as evidence that the null is actually true.

As we have taken sample from the entire population, every time we run the code, the values might change. Therefore, a csv file is attached to view the mean values of biased and unbiased groups, from which the calculations have been performed.

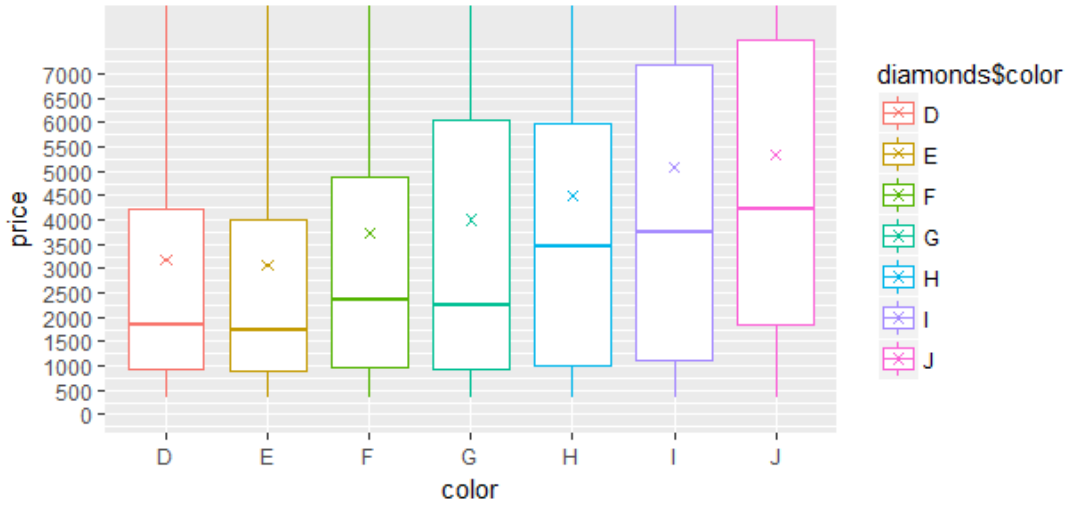
sample_out.csv

---

## 2. 2-sample T-Test

A T-Test is used to determine whether the means of two groups are equal to each other. The null hypothesis in this case is that the two means are equal, and alternative hypothesis is that they are not. In the diamond dataset, we have a categorical variable 'color'. So if we plot this variable against a quantitative variable, let's say, price, we get the following plot:

Plot:




A 2-sample T-test, checks whether the difference between the means of two independent populations is equal to target value or not. In the above plot, on first glance we may observe that the means of color type 'F' and 'G' may be equal i.e. the difference between the means of price of categories F and G, for the qualitative variable color may be 0. This becomes our null hypotheses. Alternative hypotheses will be that the difference is not equal to zero. To verify this we perform a 2-sample T-test.

```
1  diamonds.cut.F <- subset(diamonds,diamonds$color == "F")
2  diamonds.cut.G <- subset(diamonds,diamonds$color == "G")
3  t.test(diamonds.cut.F$price,diamonds.cut.G$price)
4
5          Welch Two Sample t-test
6
```

```
7   data:  diamonds.cut.F$price and diamonds.cut.G$price
8   t = -5.0453, df = 20623, p-value = 4.567e-07
9   alternative hypothesis: true difference in means is not equal to 0
10  95 percent confidence interval:
11   -380.7945 -167.7041
12  sample estimates:
13  mean of x mean of y
14   3724.886  3999.136
```

So, T-test has been performed on mean prices of the color categories F and G. Data has been subset using the first two lines of code.

The result of T-test can be interpreted in following ways:

1. t = -5.0453: T-tests are based on t-values, which are an example of test statistics. A test statistics is a standardized value that is calculated from sample data during hypotheses tests. The procedure that creates the test statistics compares the data to what is expected under the null hypotheses. T-value of 0, indicates that the sample results exactly equal null hypotheses. As the difference between the sample data and null hypotheses increases, the absolute value of t-value increases. Just a t-value cannot tell us anything. A larger context is needed to place the t-values before interpretation.
2. df = 20623: degree of freedom (df) are the amount of information the data provides which can be spent to estimate the values of unknown population parameters, and calculate the variability of these estimates.
3. p-value = 4.567e-07: This is one of the most important outcomes of a statistical testing, as it helps us to accept or reject a null hypotheses. As mentioned in our test result also, the confidence level is 95%, which makes the level of significance to be 0.05. Now, if our p-value comes out to be greater than the level of significance we accept our null hypotheses, but if it is smaller than the level of significance, we reject it. Here the p-value is 4.567e-0.7 which is extremely small than value 0.05, therefore we reject our null hypotheses and accept our alternative hypotheses, which states that the difference in means is not equal to 0 i.e. the mean price of color type ‘F’ and ‘G’ is not same. Thus concluding our hypotheses.

Note: In the result, the T-test description that we got was that it was a Welsh T-test. Looking at the documentation of t-test, we can see that there is an argument called ‘var.equal’, which gives us the option to tell the test that the variance of the two samples that we have taken for testing is same or different. For Welsh test, the value of var.equal = FALSE. If we make this value TRUE, we will get the classical t-test.

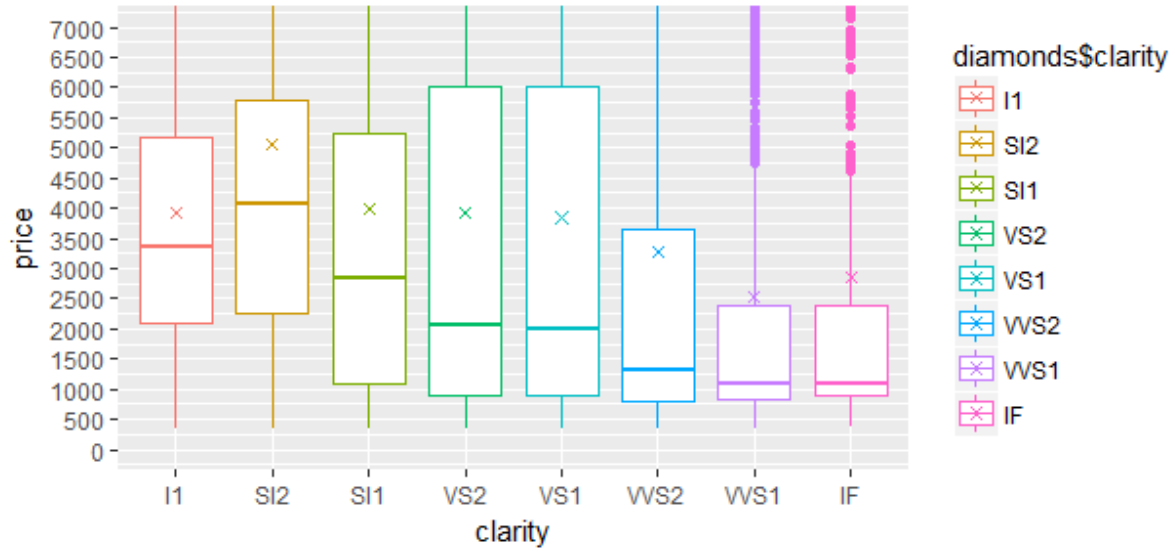
---

# 3. Paired T-Test

A paired t-test is used to compare two population means where we have two samples in which, observations in one sample can be paired with observations in the other sample. So here, we create a plot between a categorical variable ‘clarity’ and price. We take two categories from the ‘clarity’ variable VS1 and VS2, and can presume by observing the graph that the mean price of both these categories is same. So the null hypotheses becomes that the difference between the means of price

of categories VS1 and VS2, for the qualitative variable clarity may be 0. Alternative hypotheses becomes that this difference is not equal to zero. The plot is:

Plot:




To perform the paired t-test, we can sample our data with a size of 8000, so that the length of the variables, upon which the paired t-test is going to be performed, is the same.

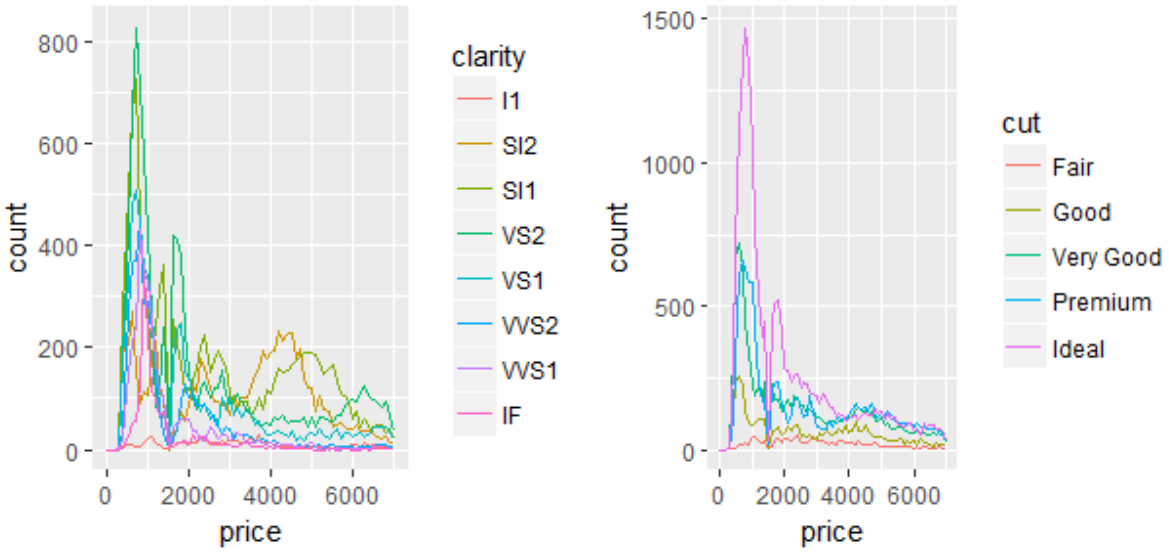
```
1  t.test(diamonds.clarity.VS1_price,diamonds.clarity.VS2_price,
2  +       paired = TRUE)
3
4                 Paired t-test
5
6  data:  diamonds.clarity.VS1_price and diamonds.clarity.VS2_price
7  t = -0.69369, df = 7999, p-value = 0.4879
8  alternative hypothesis: true difference in means is not equal to 0
9  95 percent confidence interval:
10  -168.31868   80.32818
11 sample estimates:
12 mean of the differences
13            -43.99525
```

Looking at the result of the test, we have two interesting outcomes; one is the p-value and second is the mean of the differences. In a confidence interval of 95%, the p-value comes out to be 0.4879 which is greater than the level of significance (0.05). Therefore we cannot reject our null hypotheses and cannot conclude that these two groups differ in mean. Also, a negative value in the mean of differences can mean that the sample mean is smaller than our hypothesized mean.

# 4. Chi-squared Test

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of other. A chi-squared test is any statistical hypotheses test wherein the sampling distribution of the test statistics is a chi-squared distribution when the null hypotheses is true. Basically, chi-squared test checks the dependency of two categorical variables on each other and also provides a measure of goodness-of-fit.

To show the distribution of two categorical variables, we can plot them with a quantitative variable. Here we are going to plot clarity with price and cut with price.




Now to check the dependency of cut variable and clarity variable with each other, we perform the chi-squared test. In order to establish that 2 categorical variables are dependent, the chi-squared statistic should be above a certain cut-off. This cut-off increases as the number of classes within the variable increases. Here the null hypothesis is that the two variables are independent. The alternative hypothesis is that the two variables are related.

```
1  chisq.test(diamonds$clarity,diamonds$cut)
2
3                  Pearson's Chi-squared test
4  data:  diamonds$clarity and diamonds$cut
5  X-squared = 4391.4, df = 28, p-value < 2.2e-16
```

Here, we have got p-value as 2.2e-16, which means it is less than our level of significance. Now, if p-value is less than level of significance, we reject our null hypotheses. This concludes that alternative hypothesis is correct which means that the two categorical variable clarity and cut are dependent on each other.

When we perform a chi-squared test, apart from the general attributes, we get four more attributes namely:

1. observed
2. expected
3. residual

4. stdres

The chi-squared test is based on a test statistics that measures the divergence of the observed data from the values that would be expected under the null hypotheses of no association. This requires calculation of the expected value based on the data. The standardized residual is a measure of the strength of the difference between observed and expected values.

---

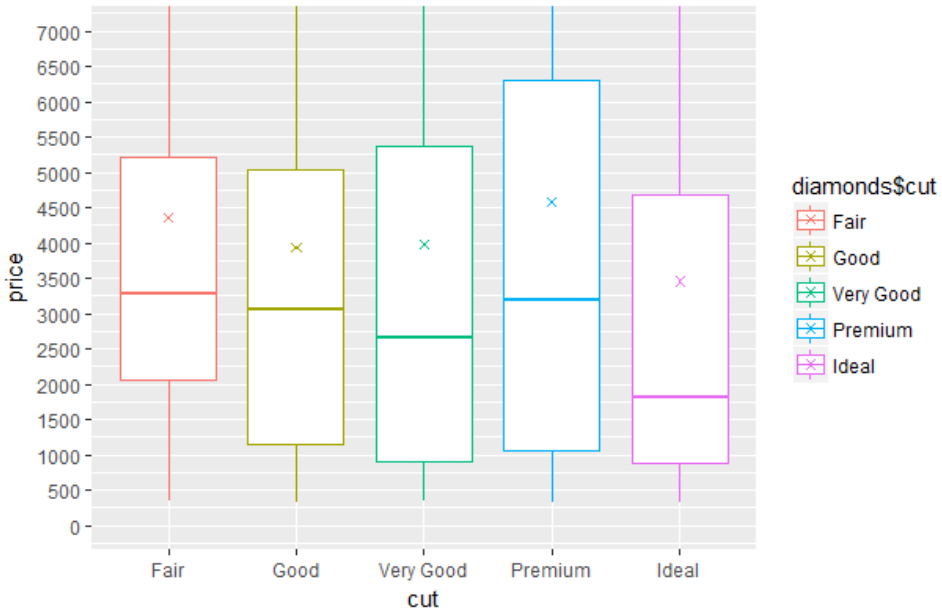
# 5.1. ANOVA Test

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the difference among group means and their associated procedures (such as "variations" among and between groups). The ANOVA test that will be performed is called one-way ANOVA which compares the means between the interested groups and determine whether any of those means are statistically significantly different from each other. Specifically, it tests the null hypotheses:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

where $\mu$ = group mean and k = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypotheses, which is that there are at least two group means that are statistically significantly different from each other.

So we take the categorical variable 'cut' and create a boxplot against price. The mean has been marked with cross.




First we do the ANOVA test, and look at the result the test has produced.

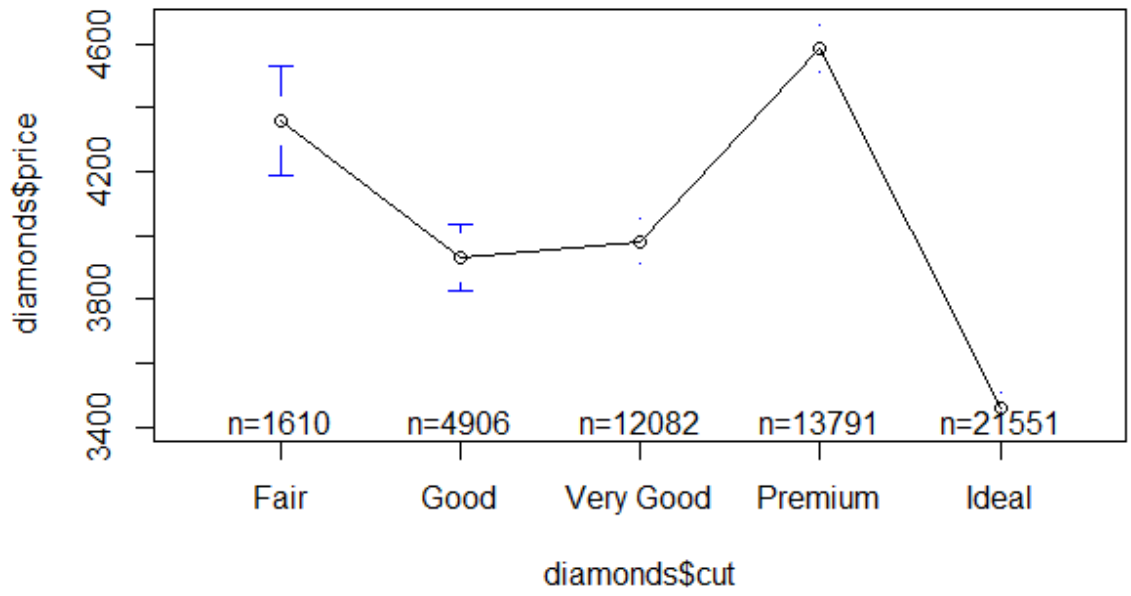
```
1  > summary(aov.test)
2                Df    Sum Sq    Mean Sq  F value Pr(>F)
3  diamonds$cut   4 1.104e+10 2.760e+09   175.7 <2e-16 ***
```

```
4  Residuals    53935 8.474e+11 1.571e+07
5  ---
6  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here, focusing on the p-value that we got (Pr), which comes out to be 2e-16. Now, as the p-value is less than the level of significance, we can reject our null hypothesis. Therefore, we can conclude that the difference between some of the means is statistically significant. F-value is much larger and p-value is much smaller, it means that the variation of price means among different category of cuts is much larger than the variation of mean prices among the categories of cuts.

Now we can create an interval plot to display the mean and confidence interval of each group. The interval plot shows the following:

1. Each dot represents a sample mean
2. Each interval is a 95% confidence interval for the mean of a group. We can be 95% confident that the category’s mean is within the confidence interval of that category.




By doing the ANOVA test we came to know that there are at least 2 categories in cut variable whose difference in means is not equal to zero. Still, we want to know that which those two variables are, or are there more than 2 variables whose mean difference is not zero. So to do this, we do another test called Tukey’s HSD.

## 5.2. Tukey’s Honest Significance Difference Test

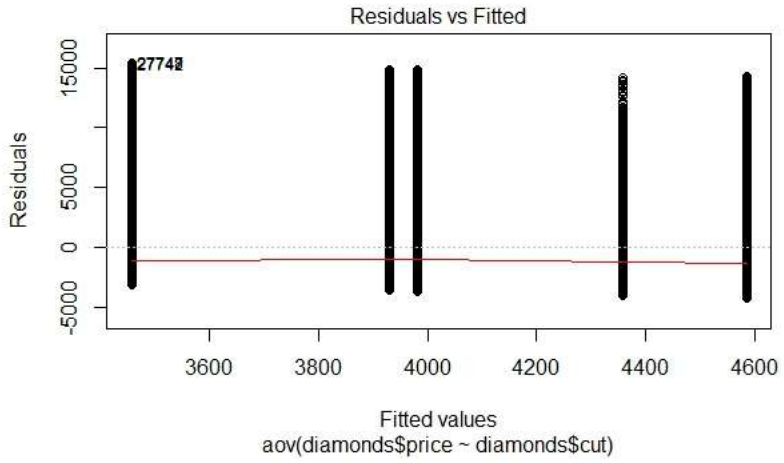
Also known as post-hoc analysis, it is a single step multiple comparison procedure and statistical test. Here, we are trying to find out what are those two categories, because of which we rejected ANOVA test’s null hypotheses.

```
> TukeyHSD(aov.test)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = diamonds$price ~ diamonds$cut)

$`diamonds$cut`
```

| | diff | lwr | upr | p adj |
|---|---|---|---|---|
| Good-Fair | -429.89331 | -740.44880 | -119.3378 | 0.0014980 |
| Very Good-Fair | -376.99787 | -663.86215 | -90.1336 | 0.0031094 |
| Premium-Fair | 225.49994 | -59.26664 | 510.2665 | 0.1950425 |
| Ideal-Fair | -901.21579 | -1180.57139 | -621.8602 | 0.0000000 |
| Very Good-Good | 52.89544 | -130.15186 | 235.9427 | 0.9341158 |
| Premium-Good | 655.39325 | 475.65120 | 835.1353 | 0.0000000 |
| Ideal-Good | -471.32248 | -642.36268 | -300.2823 | 0.0000000 |
| Premium-Very Good | 602.49781 | 467.76249 | 737.2331 | 0.0000000 |
| Ideal-Very Good | -524.21792 | -647.10467 | -401.3312 | 0.0000000 |
| Ideal-Premium | -1126.71573 | -1244.62267 | -1008.8088 | 0.0000000 |

As we can see, post-hoc test gives us the difference in means of each category in the variable that we want to test. Here, it has given the difference in mean price of every category in cut.

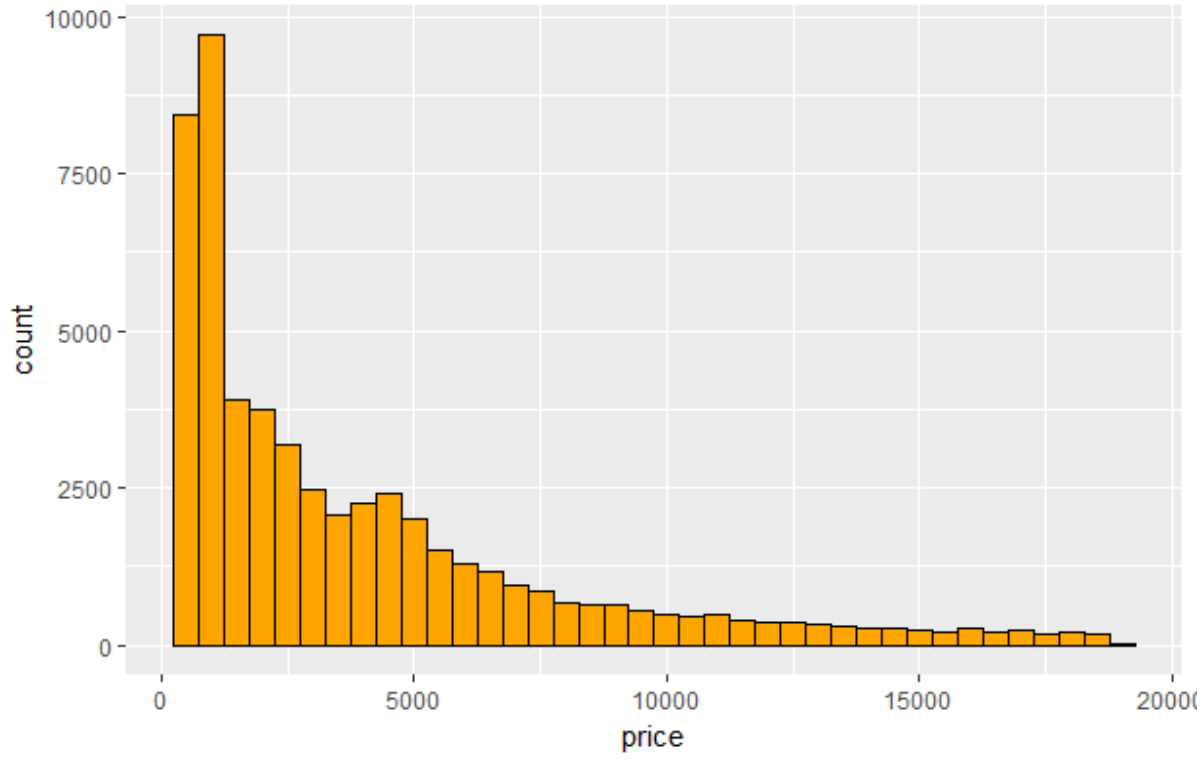
We can also the residual versus fit plot to verify the assumption that the residuals are randomly distributed and have constant variance.



# Variable Distributions

By definition, the distribution of a statistical data set (or a population) is a listing or a function showing all the possible values (or intervals) of the data and how often they occur. When a distribution of categorical data is organized, you see the number or percentage of individual in each group.

So, let's start by analyzing price variable. If we create a histogram of price, we get the following:



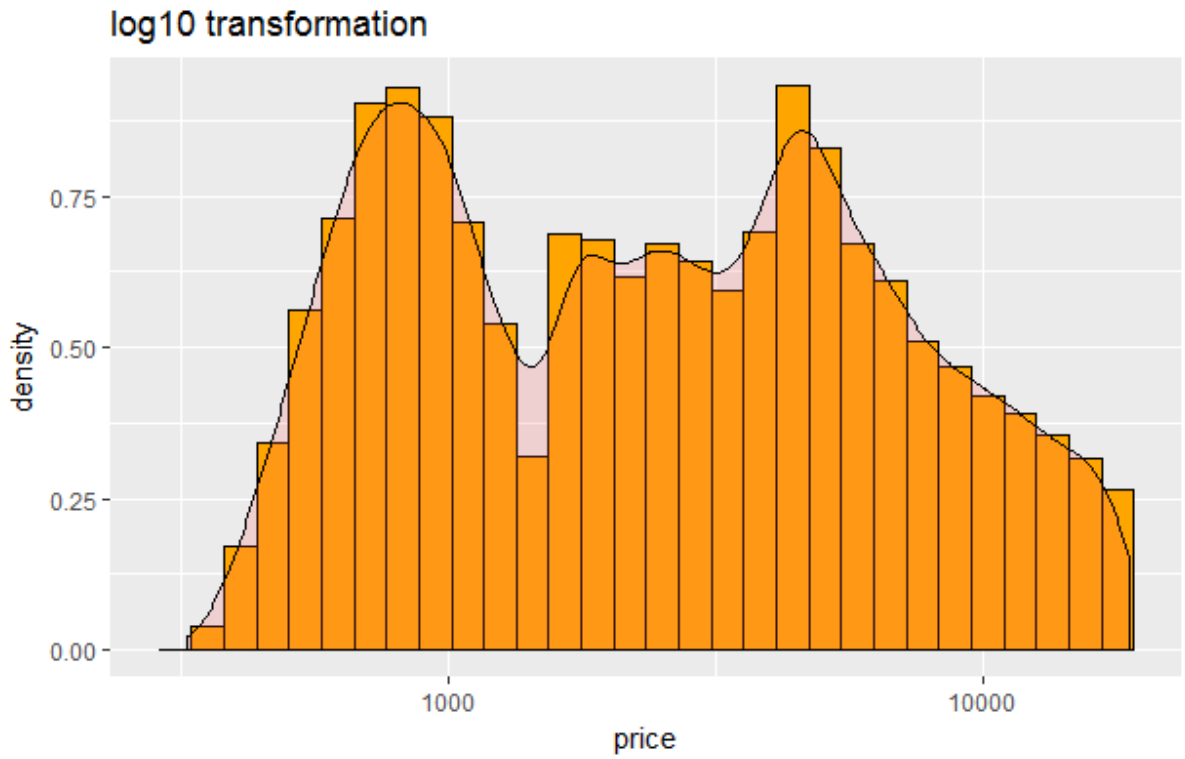

The histogram that we get here is heavily right-skewed. For a right skewed distribution, the mean is typically greater than the median. We can see it:

```
1  > mean(diamonds$price)
2  [1] 3932.8
3  > median(diamonds$price)
4  [1] 2401
```

Using the summary command, we can see that the median is closer to the first quartile than the third quartile.

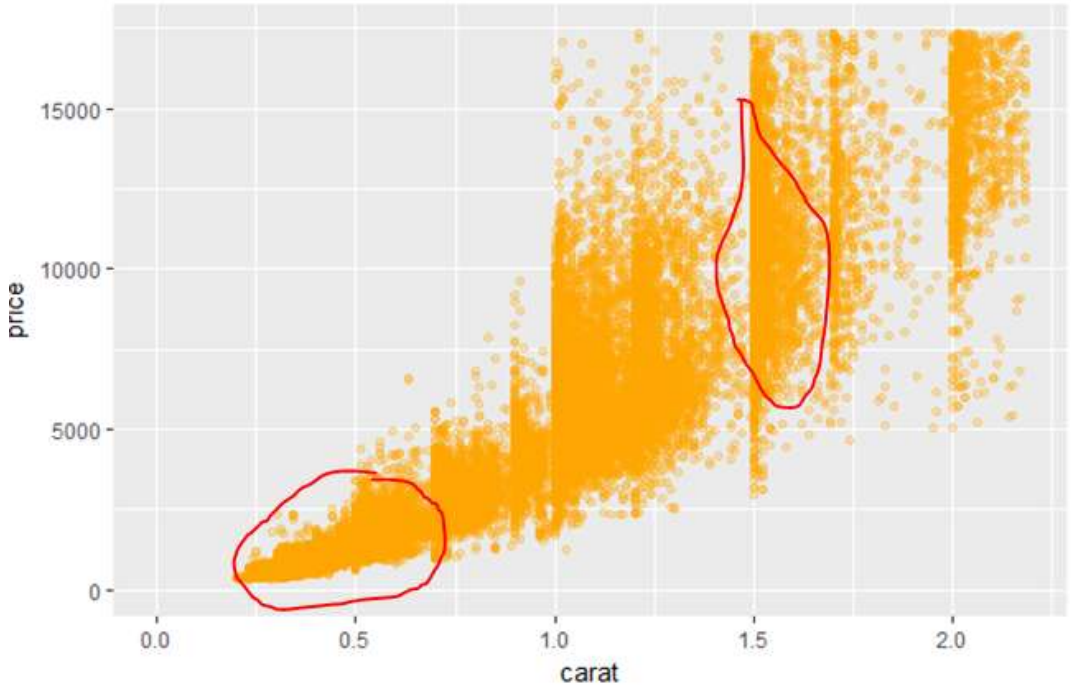
```
1  > summary(diamonds$price)
2     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3      326     950    2401    3933    5324   18820
```

Now we can apply some transformation on price, to see whether its distribution changes or not.



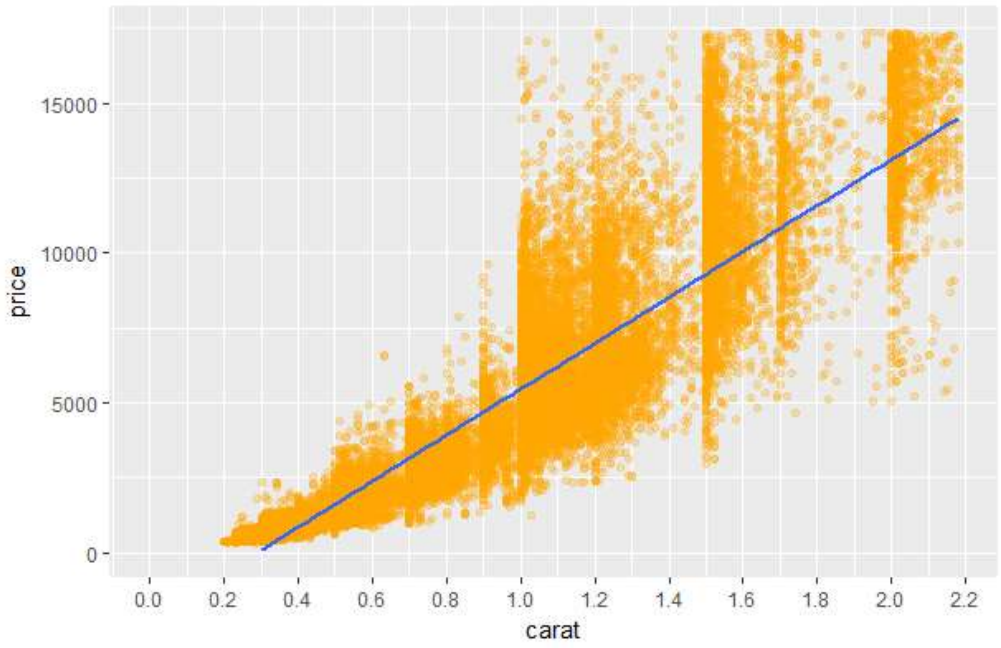

We have transformed the x-axis using log10 transformation and have also added density plot in it. What we can see is, the y axis is now showing density instead of counts. Also this distribution is called as Bimodal distribution. A bimodal distribution is a continuous probability distribution with two different modes. These appear as distinct peaks in probability density function, as shown in the figure above. Each peak is a **local maximum** since they represent the highest values relative to the data points immediately surrounding them. The valley between these peaks is called a **local minimum**. Bimodal distributions come in all shapes and sizes, but they all have the same thing in common; each graph has two distinctive maxima with a relative minimum between them.

Now if we create a scatter plot between carat and price, we might see areas having high density, and because of this reason, the price variable is bimodal distribution.



The two areas, highlighted in red, are the cause of the two spikes in the distribution of price variable. This can be due to the fact that, people having less money tend to go for diamonds whose carat value is lower and people who have more money tend to go for diamonds whose carat value is higher. This can also be seen in the plot, where the higher density areas have been divided by carat value of 1.

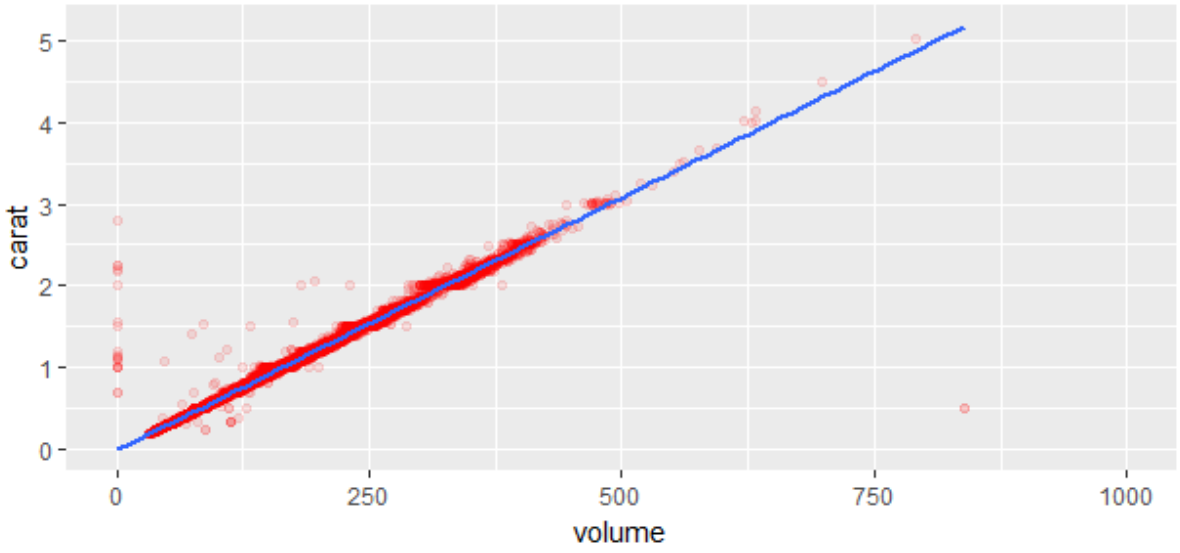
Now, if we want to see the regression between these two variables, we can do that by adding a regression line or line of best-fit. So adding the regression line, the graph becomes:




The regression line increases with increase in values of carat and price, i.e. positive linear regression.

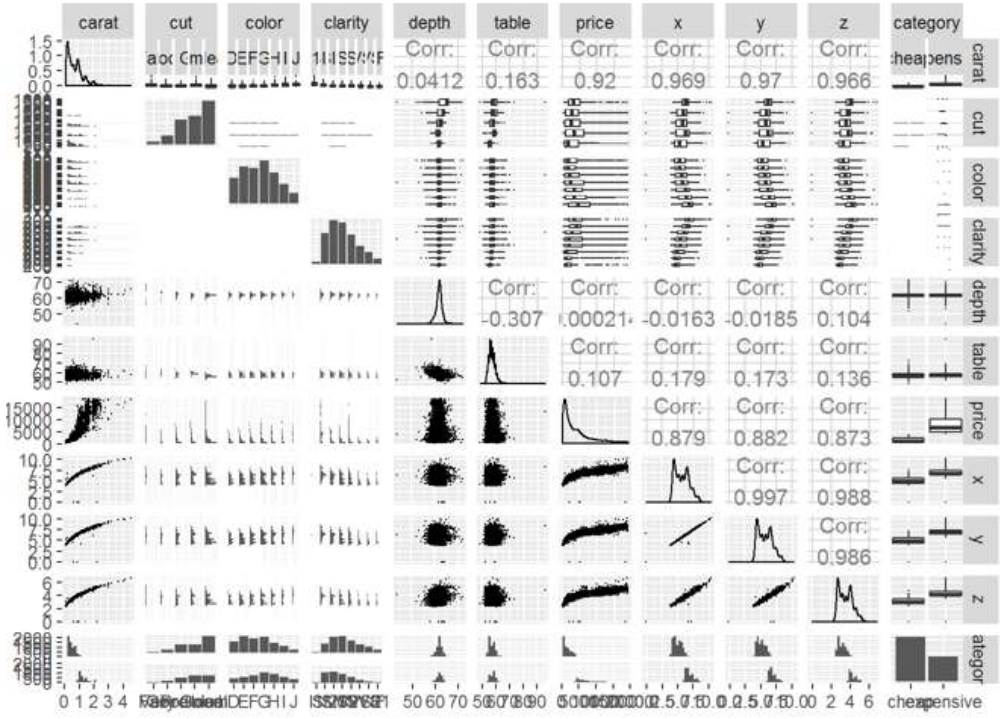
We have three variables in diamond dataset, namely: x, y and z. We can calculate the volume of diamonds using these variables and test out the regression between other variables based on physical characteristics.

We can check the regression between the volume of a diamond and the carat variable, by plotting a regression line through them. The graph turns out to be:



We can see that the linear regression line superimposes the points of scatter plot between carat and volume variables. This means that they might be strongly correlated. After performing the correlation test, we know that these two values are highly correlated with each other as their coefficient of correlation is 0.97, which is almost tending to 1 i.e. a perfect straight positive regression line.

Now to show the correlation between each variable of the diamond dataset, we can create a correlation matrix, which can be seen below.




# Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than those in other groups (cluster).

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them.

K-means clustering is the most popular partitioning method. It requires the analyst to specify the number of clusters to be extracted. A plot of within group sum of squares by numbers of clusters extracted can help determine the appropriate numbers of clusters. The analyst looks for a bend in the plot similar to a bend in “elbow”.

We use the kmeans function on our clean dataset with the k-, i.e. the number of clusters in a data set.

We have different elements of the cluster output.

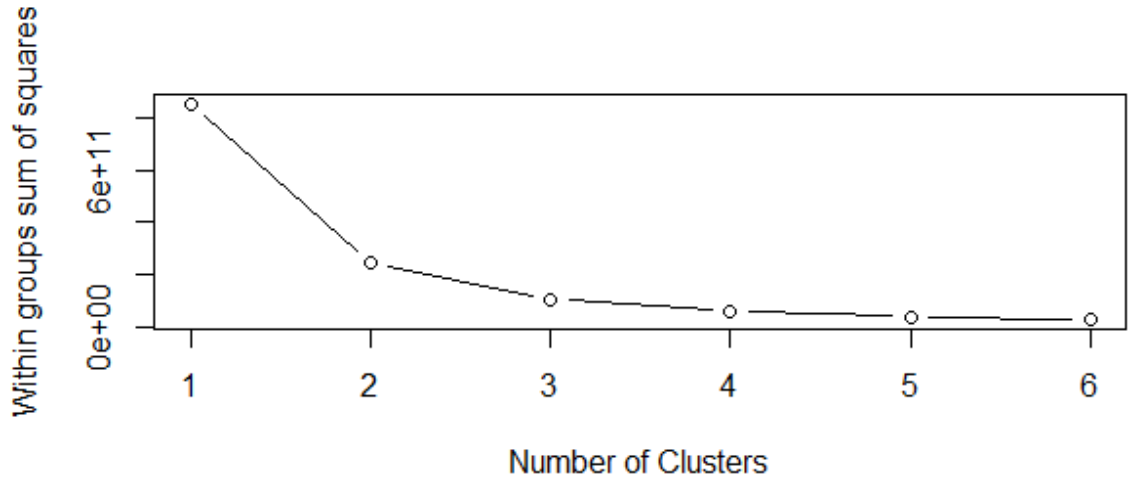
```
1  > attributes(k.means.fit) #all the elements of cluster output
```

| | |
|---|---|
| 2 | $names |
| 3 | [1] "cluster" "centers" "totss" "withinss" "tot.withinss" |
| 4 | [6] "betweenss" "size" "iter" "ifault" |
| 5 | |
| 6 | $class |
| 7 | [1] "kmeans" |

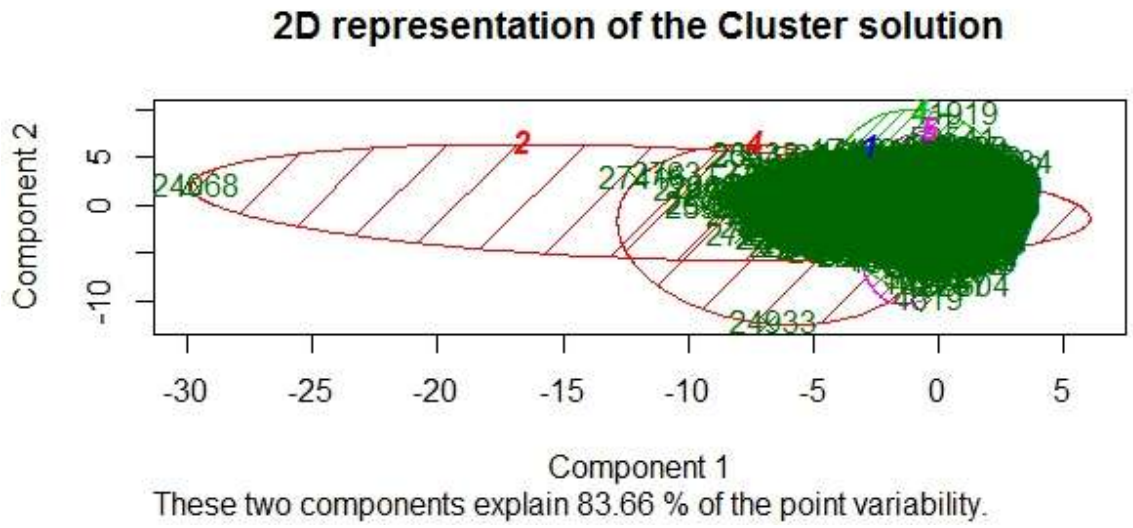
For example:

| | |
|---|---|
| 1 | > k.means.fit$size #cluster size |
| 2 | [1] 9874 44066 |
| 3 | > k.means.fit$centers #centroids |
| 4 | diamonds.carat diamonds.depth diamonds.table diamonds.price diamonds.volume |
| 5 | 1 1.5308173 61.69245 57.79571 11052.805 248.5876 |
| 6 | 2 0.6337217 61.76217 57.38133 2337.399 103.2434 |

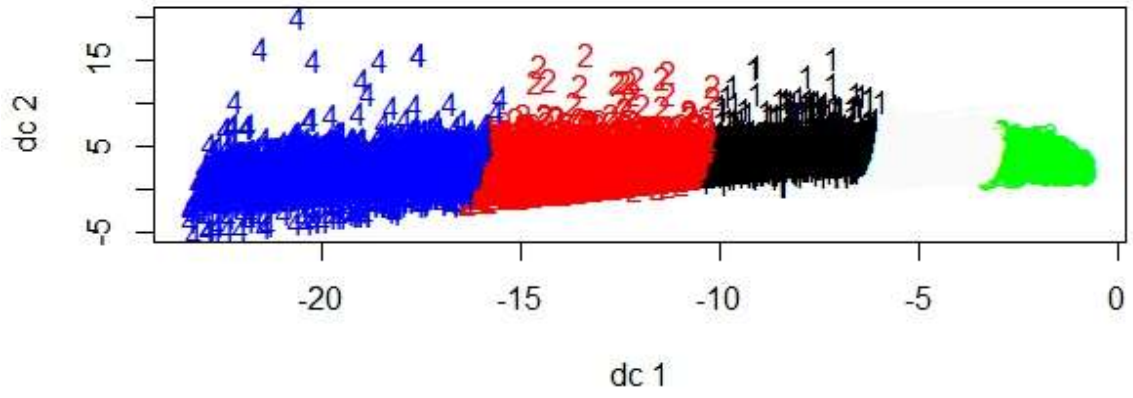
The k value, as mentioned before can be found from this elbow curve:



Now taking the value of k as 5, a Bivariate Cluster Plot (of a partitioning object) has been created, which draws a 2-dimensional clustering plot on the current graphic device




Another plot, called a discriminant projection plot, is created to distinguish given classes by ten available projection methods.




In order to evaluate the clustering performance we build a confusion matrix:

| | 0.2 | 0.21 | 0.22 | 0.23 | 0.24 | 0.25 | 0.26 | 0.27 | 0.28 | 0.29 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 12 | 9 | 5 | 293 | 254 | 212 | 253 | 233 | 198 | 130 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

This is a confusion matrix between the carat variable and the number of clusters. The values in the cells represent the occurrence of carat values in each cluster.

# Conclusion

Thus, we have seen in the report how hypotheses are formed and tested according to the variables of the data set, the significance of p-value and confidence intervals for statistical interference. We have also seen how regression works and how it can be predicted from coefficient of correlation, how smoothening lines are drawn over a plot and what is there significance. We also saw clustering done via k-mean algorithm and how data can be grouped into different clusters.

Note:

1. The matrix for correlation between every variable of diamond dataset is taken from assignment 1.
2. For some hypotheses, the values are taken via sampling, so at each compilation p-values may vary, but still description has been written about it, at its best.
3. Hierarchical clustering was attempted but could not be done due to Memory performance problems of the system. Same with model based clustering. Therefore, only k-mean clustering has been presented.
4. References are being given in a separate text file.
5. If the csv file cannot be opened, It has been included in the folder uploaded to moodle.
6. The plot for post-hoc test (Tukey's HSD) hasn't been included in the report, but is there in the R code with explaination.