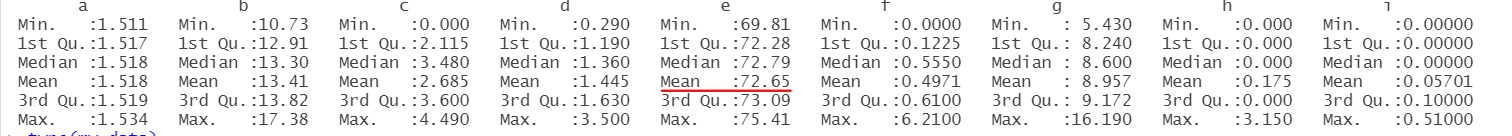
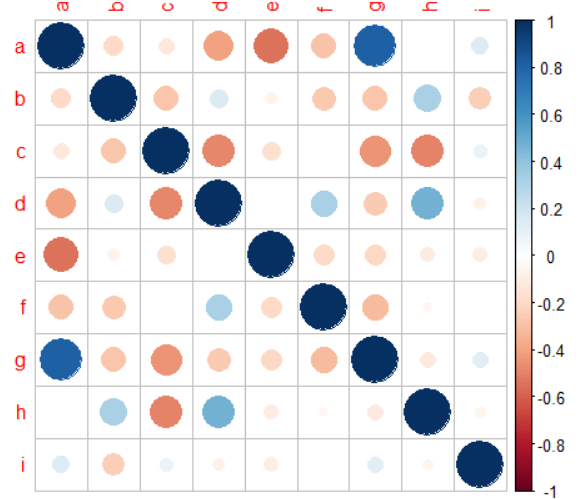
**1a)** After reading the csv file provided the first step taken to perform the analysis was to understand the data given. To do that I first begin by having a peak at the data and it can be seen that data consist of float values. Furthermore, it is crucial to check for any null values present and upon checking it can be seen that the data does not contain any null values. By diving deeper into the descriptive statistics of the data it can be seen that additive e has relatively high mean value compared to other column.

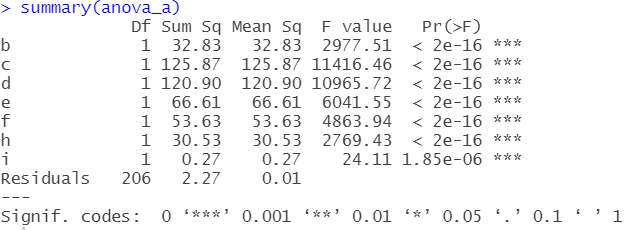


To observe the relationship between the additives I will be plotting a correlation matrix .



From our correlation matrix it can be seen that additives a and g has the highest positive correlation between them. Meanwhile other additive shown to have very little no correlation. It is self-explanatory on why many weak relationship are shown as each additive are added to formulations of petrol to observe their performance. Thus, it shall not be depending on other additives.

Furthermore, as per the requirement parametric ANOVA test will be performed. However, the correlation matrix illustrates poor relationship between the additives and ANOVA test uses a dependant variable. Thus, based on the correlation matrix I have used additives a and g as a dependent variable. The summary of result is shown below.



* F-test result of additives c and d are relatively large which indicates the variation associated with the independent variable is real and it is not by a chance. On the other hand, F-test value for additive i is significantly low which indicated variation associated with independent variable might be due to a chance.
* All of the p-values are < 0.05 which explains that it is likely for additives **b,c,d,e,f,h,i** have significant effect on additives **a and g.**

A picture containing table

Description automatically generated**1b)** To understand the distribution of the additives we will be using boxplot and histogram.

By looking at the histogram we can observe that all of the additives are right skewed except for additive “e” where it is left skewed. We can also observe the difference between the centre of the group for different additives. While some distribution for additives “b,c,g,e” more spread out, the distribution for additive “a,d,f,h” are almost in same centre which indicates the consistent level of additives added to the formulation for those 4 additives.

To further investigate the distribution with boxplot some of the interesting fact we can observe are the number of outliers and the deviation.

A picture containing table

Description automatically generatedAdditives “a” and “i” has less deviation with their low standard deviation values 0.003036864 and 0.0974387 respectively. Furthermore, it can be seen that additives “g” and “h” has many outliers compared to other additives. Lastly, it can be seen that the median value for majority of the additives added to the formulation are similar except for “b,c,g,e”

Thus, by looking at the two distribution and similarity pattern between the histogram and boxplot we can conclude that level of additive added to the formulation for the test is close to central region.

***Note: The above visualisation is available in the GitHub repository for further viewing.***

**1c)** Unsupervised learning clustering test will be used to determine the number of distinctive number present in our dataset. For this I will be using k-Means partitional clustering approach where k will represent the number of cluster. First of all before I start the model implementation I will use the silhouette method to determine what is the optimum number of k (cluster) to be used by calculating how well each data point sits within its cluster.

**Pre-Processing steps:**

* Min-Max normalisation : This will ensure large values attributes wont exert greater influence on the clustering.

Chart, line chart

Description automatically generated

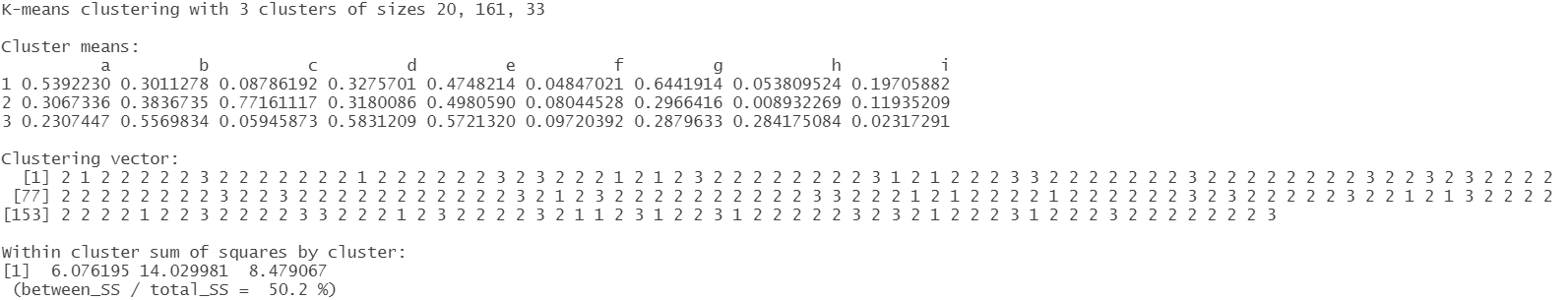
Chart, line chart

Description automatically generatedAverage silhouette score for 2-10 clusters were calculated and plotted as above. As we can see, the optimum number of clusters to be used is 3 clusters. Alternatively, this can also be done using elbow method as well as shown below.

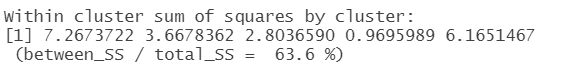
Chart, line chart

Description automatically generatedUpon getting the recommended number of cluster the k-means clustering algorithm was implemented. Moreover, mean value of each additives within each cluster is calculated to visualize the result of the clustering.

Some of the things we can observe from the graph are that for additives “e” and “f” mean values in the clusters are close to each other. On the other side values in the mean clusters are different and spread out for across the clusters. Moreover, for additives “c”, “d”, “g” mean values in 2 of the clusters are close to one another while it is different from remaining one cluster. To sum up we can say that additives “a” , “b”, “c”, “d”, “g”, “h” and “i” have different mean values across all 3 clusters.



The classification accuracy obtained after dividing between\_SS (Between Sum of Square) by total\_SS (Total Sum of Square) for 3 clusters is 50.2% which seems to be low.



To increase the classification accuracy, we can simply increase the number of clusters which increases the dividing between\_SS (Between Sum of Square). For instance, when we increase the number of clusters to 5. The accuracy has now increased to 63.6% which is which is an increased of 13.4%.

Some of the extra post-processing steps we can perform are:

* Eliminate small clusters that may represent outliers
* Merge clusters that are ‘close’ and that have relatively low SSE.

Partitional clustering was performed in this report which is easy to implement and scales to large dataset. However, the downside of partitional clustering is that the centroids size can be dragged by the outliers. Alternatively we can also use hierarchical clustering which allows us to easily calculate minimum , maximum and group average distance between the centroids.