**Q1 Report**

In this question, we are required to verify the difference in the formulations provided. Instead of starting from Q1a, I would approach the question starting from Q1b to understand the distribution of the additives’ loadings. By using the script “Q1b.py”, we can obtain the violin plots of the additives’ distributions and several key statistical metrics, as shown in the following figure.

Chart

Description automatically generated with medium confidence

Some of the additives, such as “a” did not deviate much with low standard deviation, suggesting that it shall be the same for different formulations. Generally, we can see that the kurtosis of the additives (except for “c”) are high, indicating that their distributions are concentrated close to the peak with heavy tails. Therefore, it can be said that the amount of the additives’ loadings tested are close to the central region.

With that respect, we can revert to Q1 to perform parametric/non-parametric description of the additives. In this context, I am not sure why a parametric/non-parametric test would be applicable since we are not performing hypothesis testing. Anyway, I tried to conduct the correlation analysis as well as the two-way ANOVA with linear model. The Pearson correlation heatmap is shown as following.

Chart, waterfall chart

Description automatically generated

We can see that most of the additives are not so strongly correlated or associated, except for the pair of “a” and “g”. This finding is also in agreement with our violin plots, where “a” and “g” have similar distributions. The additives are somehow weakly correlated. This is reasonable as the additives shall be independent of each other in order to study their influence of the performance of the petrol.

Subsequently, we tried to use each of the additive as the dependent variable and the remaining independent variables. I am not sure whether this is correct because normally when we perform ANOVA test, we will have a dependent variable given (for example, the performance of the petrol). The reports of the ANOVA test (using each variable and dependent variable in a for-loop) are stored in a “res” dictionary when running the script “Q1a.py”. The ANOVA test performed is incomplete as it did not consider the interactions among the additives.

A clustering method is used to determine the number of significantly distinct formulations present in the 214 samples. The clustering method used in this case is the conventional k-means algorithm, where we can explicitly determine the number of clusters. Since we do not know how many statistically distinct formulations are there, I employed the “elbow” and “silhouette” methods to select the optimum number for k. The results are shown as below.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

From these two figures, we can see that the optimum value of k is 3. The follow up investigation will be based on this k value.

Since we already know that there are probably three distinct clusters (or formulations) present in the sample, we can use the snake plot to visualise how the clusters (or formulations) differ from each other in terms of the mean values of the additives’ loadings. The snake plot is shown as below. Note that the mean values shown in this figure is normalised to the range of [0, 1].

Chart

Description automatically generated

As you can see from this figure, the three clusters do not vary much from each other in terms of “e” and “f”, where the three “snake” meet each other. Also, from the violin plots, we know that the deviation of “a” is very small, which suggest that it is the same for all the three clusters, despite the snake plot shows discrepancy due to the normalised value. Hence, from here, we can conclude that only the additives “b”, “c”, “d”, “g” and “i” are different across the three clusters.

The importance of each of the attributes (or additives) can be visualised using a relative importance heatmap. The relative importance is calculated based on the difference between the cluster’s mean to the population mean. It is shown as the following.

Chart, treemap chart

Description automatically generated

The darker the tone of the tiles (can be red or blue) indicates that the attribute is more important for the cluster. Therefore, we can see that “Cluster 1” is actually studying the effect of the low loading “h” additive. On the other hand, “Cluster 2” is studying the effect of high “h” loading with low “c” and “i” loading. Combining all these results, we can make a more informed conclusion such the 214 samples can be grouped into three clusters that are investigating the effect of varying loadings of the additives.