Petrol Formulation Analysis

Scenario:

A customer informed their consultant that they have developed several formulations of petrol that gives different charactristics of burning pattern. The formulations are obtaining by adding varying levels of additives that, for example, prevent engine knocking, gum prevention, stability in storage, and etc. However, a third party certification organisation would like to verify if the formulations are significantly different, and request for both physical and statistical proof. Since the formulations are confidential information, they are not named in the dataset.

Please assist the consultant in the area of statistical analysis by doing this:

a. A descriptive analysis of the additives (columns named as "a" to "i"), which must include summaries of findings (parametric/non-parametric). Correlation and ANOVA, if applicable, is a must.

For determining relationships, scatter plots between 2 additives can be plotted. This is repeated for all additives. Intuitively, the scatter plots show that we are unable get a good fit by plotting a linear relationship due to the high number of outliers and the absence of a strong trend.

A Shapiro-Wilk normality test shows that all additives have a non-normal distribution. Therefore, a non-parametric test called the Spearman's Rank Correlation will be used to assess the relationship between two additives.

The evaluation metric for Spearman's Coefficient (adapted from Dancey and Reidy, 2004) is as follows:

Spearman R	Correlation	Observed pairs
>=0.7	Very strong positive relationship	"a" and "g"
0.4 to 0.69	Strong positive relationship	"b" and "h", "d" and "h"
0.3 to 0.39	Moderate positive relationship	none
0.2 to 0.29	Weak positive relationship	"c" and "f"
-0.19 to 0.19	No relationship	"a" and "b", "a" and "c", "a" and "h", "a" and "i", "b" and "c", "b" and "d", "b" and "g", "c" and "i", "d" and "e", "d" and "f", "d" and "i", "e" and "f", "e" and "h", "e" and "i", "f" and "i", "g" and "h", "g" and "i", "h" and "i",
-0.2 to -0.29	Weak negative relationship	"a" and "f", "b" and "e", "b" and "i", "c" and "g", "d" and "g", "e" and "g" , "f" and "h"
-0.3 to -0.39	Moderate negative relationship	"c" and "e"
-0.4 to -0.69	Strong negative relationship	"a" and "d", "a" and "e", "b" and "f", "c" and "d", "c" and "h", "f" and "g"
<=-0.7	Very strong negative relationship	none

A positive correlation indicates that one additive typically increases as the other additive increases in the formulation. A negative correlation indicates the opposite. No correlation means the two additives do not have any dependencies or influence over each other.

There is a 0.111 probability for any two additives combination to be positively correlated, 0.389 probability for them to be negatively correlated, while 0.5 probability of not being correlated at all.

Available at https://github.com/aplatyps/data_science_demo

The following is the summary of my analysis:

(i) "a" and "g" pair have very strong positive relationship.
(ii) "b" and "h" pair have strong positive relationship.
(iii) "a" and "d", "a" and "e", "b" and "f", "c" and "d", "c" and "h", "f" and "g" pairs have strong negative relationship.
(iv) There is only 1 very strong correlation.
(v) There are more negative relationships which are strong than positive ones.

b. A graphical analysis of the additives, including a distribution study.

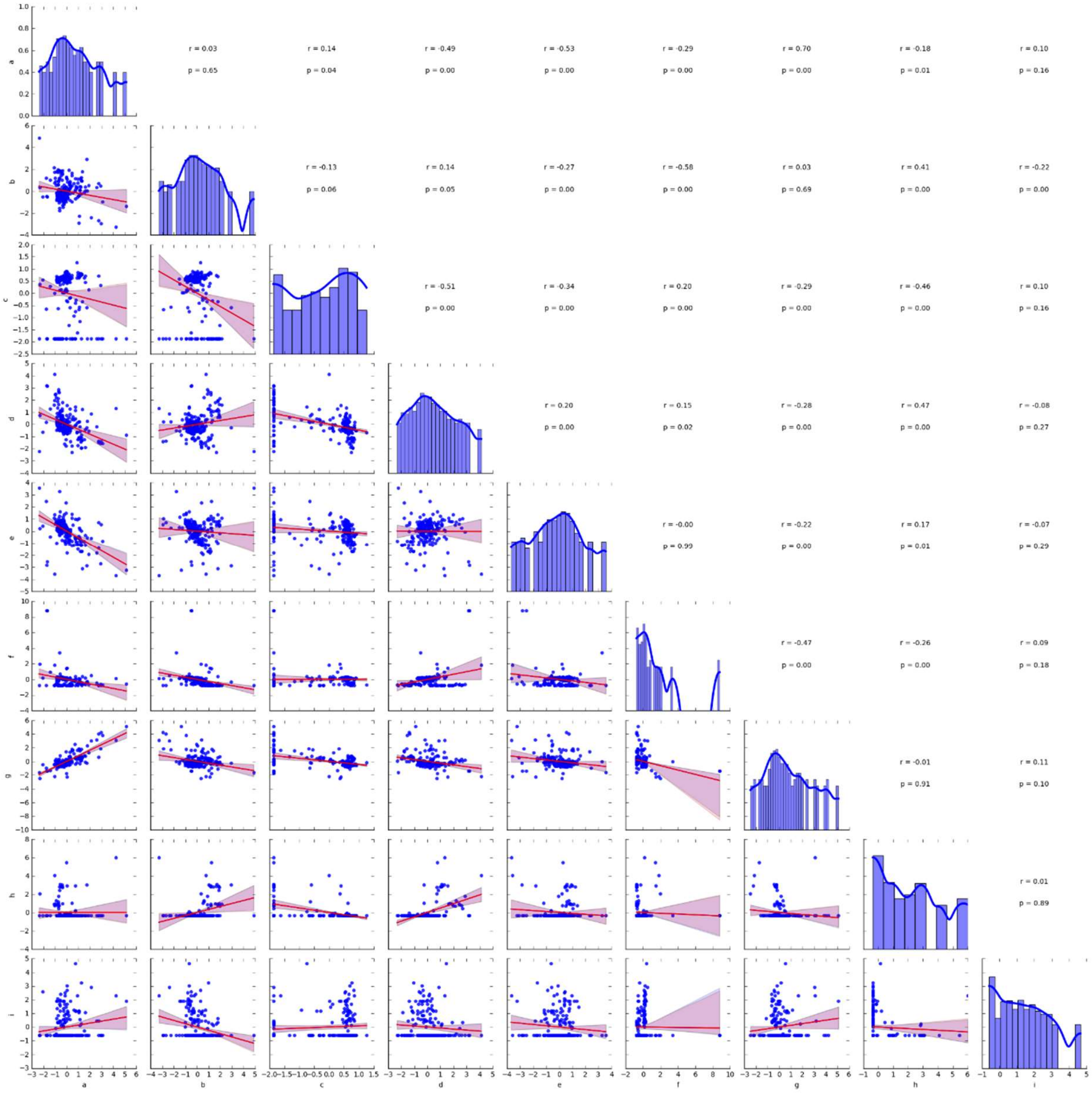

Figure. 1. The scatter-matrix plot of all additives and the pairs. Top-right shows the respective Spearman's ranks and p-values, bottom-left shows the respective scatter plots with Spearman's coefficient, while mid-diagonal shows the respective histograms. *The plot is logarithmically-scaled on the y-axis for better visualisation purpose.*

Available at https://github.com/aplatyps/data_science_demo

All additives “a”, “b”, “c”, “d”, “e”, “g”, “h”, “i” have non-normal distributions as statistically proven with Shapiro-Wilk normality test. This can also be observed from the histogram with kernel density estimation line plotted diagonally in Figure 1, where the distributions do not appear to be normal. A better visualisation can be found at the interactive .HTML report under Variables tag.

<profiling_report_for_petrol_formulation_additives.html>

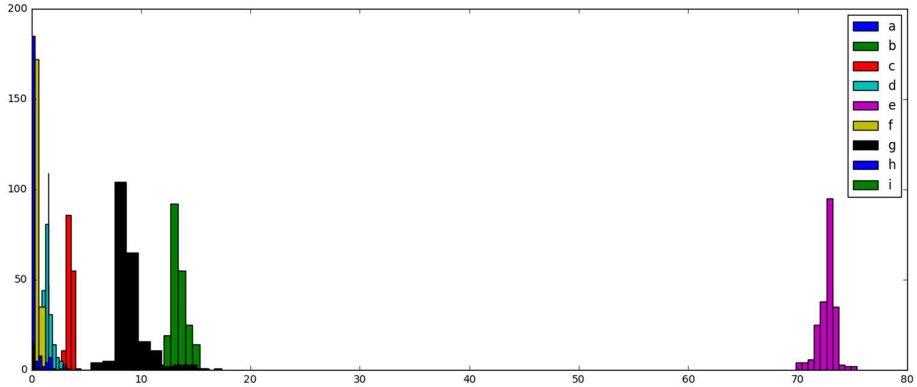

Figure. 2. The distribution of the entire dataset plotted on a histogram grouped by the additives to scale.

From Figure 2, we can observe that most of the additives are below a level of 20 with the exception of “e”. The top 3 ingredients are “e”, “b” and “g” which we naively assume that they are the main ingredients in the formulation while the rest of the ingredients are the key to the difference in burning patterns.

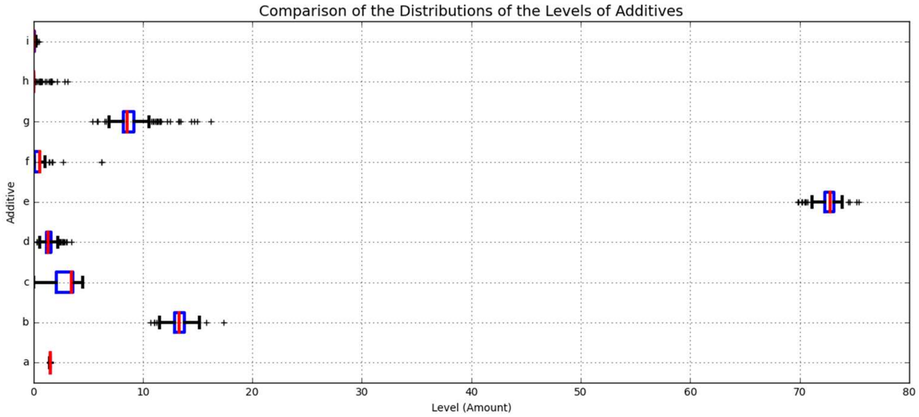

Figure. 3a. The distribution of the entire dataset plotted on a box-and-whisker plot grouped by the additives.

The box-and-whisker plot in Figure 3 provides a better view of the skewness of the distribution and the outliers. “a”, “d”, “g”, “h” and “i” are observed to be positively skewed, “c” and “e” are observed to be negatively skewed while “b” and “f” less skewed. Outliers can be observed for “b”, “d”, “f”, “g”, “h” and “i” where “h” is observed to have the highest number of outliers. Intuitively, we can assume

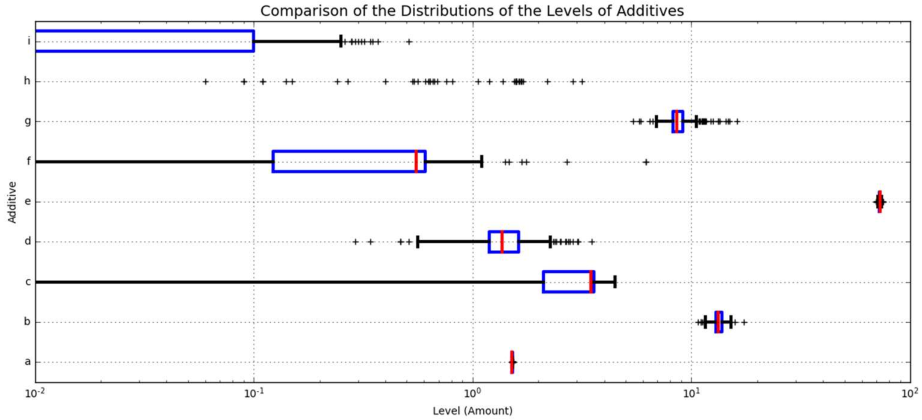

Figure. 3b. The distribution of the entire dataset plotted on a box-and-whisker plot grouped by the additives on logarithmic scale for better visualisation purpose.

the outliers to be erroneous in the formulation. However, since the outliers of "h" span across extremely large range, as seen in Figure 3b where the log scale shows relative rate in the level (amount) of additive. The presence of "h" in the formulation may be intended and possibly a key difference to the formulation. Hence, the outliers may not be anomaly at all. Another interesting observation from Figure 3 is that among the top 3 ingredients: "e", "b" and "g" that we have identified earlier, along with "a" have small sample variability. This could be an indication that the roles of "e", "b", "g" and "a" in petrol formulation are not too significant.

In all formulations, it is observed that:

- "a" is present in small amounts between 1.51115 to 1.53393.
- "b" is present in small amount between 10.73 to 17.38.
- "c" is either present in a formulation in varying levels up to 4.49 or not at all.
- "d" is present in very small amount between 0.29 to 3.5.
- "e" is present in extremely large amount between 69.81 to 75.41.
- "f" is largely present in small amount in most of the formulations or very rarely in small amount between 0 to 6.21.
- "g" is present in small amount between 5.43 to 16.19.
- "h" is largely not present in most of the formulations, very rarely in various very small amounts between 0 to 3.15.
- "i" is largely not present in most of the formulations, very rarely in various extremely small amounts between 0 to 0.51.

The following is the summary of my analysis:

(i) All additives have non-normal distributions, both statistically proven and observed by visual

(ii) Outliers in dataset which are often omitted in statistical modelling and machine learning modelling may not be anomaly at all, hypothesised with the example "h" (see Figure 3 explanation for details). Therefore, when performing clustering, model must be able to handle outliers because we could not remove them via pre-processing.

(iii) Additives with high value means they exist as a high percentage in the formulation, and vice versa. A naïve assumption of additives' by rank in petrol formulation ingredient in descending order would be: "e", "b", "g", "c", "d", "a", "f", "i", "h".

(iv) "e", "b", "g" and "a" may not play a significant role in distinguishing different petrol formulation.

c. A clustering test of your choice (unsupervised learning), to determine the distinctive number of formulations present in the dataset.

A distance-based solution K-Means is selected as is the most straightforward and easy to explain.

As K-Means is Euclidean distance-based, we must normalise our dataset first to ensure the measured distance will not be skewed.

From the summary of descriptive analysis, we understood that the dataset has features with relationships as described above. From a machine learning point of view, features with high correlations are typically dropped to reduce the dimensionality and complexity of data for the model to fit better. From the conclusion made in the descriptive and graphical analysis, the following strategy is derived for dimensionality reduction:

- Perform principal component analysis on "a" and "g" to reduce the components from 2 to 1
- Drop "a" and "g", include the new component into the dataset
- Drop "e" and "b"

By doing so, the dimension of dataset is reduced from 9 to 6.

To estimate the distinctive number of formulations present in the dataset, clustering methods can be used on the dataset. By determining the optimal number of clusters, this can be translated to the possible distinctive number of formulations.

The following strategy is used for determining the optimal number of clusters:

- Elbow method, minimise the sum of squared distance of samples (inertia)
- Silhouette method, maximise the separation distance between the clusters
- Validation with DBSCAN
- Validation with hierarchical clustering

K-Means is tested on a number of clusters from range of 2 to 20. From Figure 4, the elbow method indicates that the optimal cluster is 5, the silhouette method indicates that the optimal cluster is 4, DBSCAN indicates that the optimal cluster is between 3 to 4, hierarchical clustering indicates that the optimal cluster is 6. Taking the average of the 4 results, the number of optimal clusters would be 5.

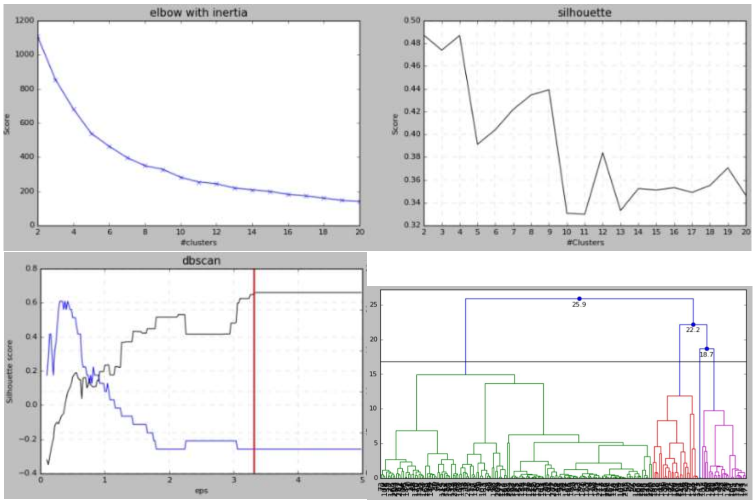

Figure. 4. Plots of various methods to check for optimal number of clusters.

K-Means model is fitted to the dataset with 5 clusters, the following is the result:

Cluster assigned to formulation:

1 4 1 1 1 1 0 3 1 1 1 1 1 0 0 0 0 4 0 1 1 1 2 1 2 1 1 1 4 1 4 1 2 1 1 1 1 1 1 1 0 4 4 1 4 1 2 1 2 2 1 1 1 0 0 1
1 2 1 2 1 1 1 1 0 1 1 0 0 2 0 2 0 0 0 1 1 1 1 1 1 1 0 1 2 1 1 2 0 1 1 1 1 0 1 0 1 1 0 2 1 4 1 2 1 1 1 1 0 4 1 1
0 1 2 2 1 1 1 4 0 4 0 1 0 1 4 1 1 1 1 1 1 2 1 2 0 1 1 0 1 2 1 1 4 1 4 1 1 0 4 1 1 1 1 1 4 0 0 2 0 1 0 1 2 2 1 1
1 4 1 2 1 0 1 0 2 1 4 4 0 2 0 1 1 1 4 1 1 1 1 1 2 1 2 1 4 0 1 1 1 4 0 0 1 1 1 1 1 0 0 1 0 3

Distinctive clusters count:

Cluster assigned	Count
0	43
1	120
2	27
3	2
4	22

The following is the summary of my unsupervised learning:

(i) K-Means is selected for clustering.
(ii) Dimensionality reduction done on dataset based on conclusion drawn from descriptive and graphical analysis.
(iii) Optimal parameter for K-Means through elbow and silhouette methods, as well as validating with DBSCAN and hierarchical clustering.
(iv) Distinctive number of formulations present in the dataset is 5.

=== END OF REPORT ===

Available at https://github.com/aplatyps/data_science_demo