# NCS\_Assessment

Cheng CHEN - NCS Data Specialist Assessment

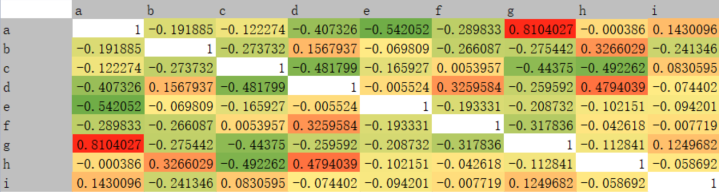
**Q1**

**Q1-a**

It is suitable for correlation analysis, but not for ANOVA analysis because there are no groups for doing ANOVA.

It can do the ANOVA analysis after grouping by clustering, however it does not make sense to do ANOVA after clustering.

correlation matrix of the ingredient of additives is as follows:



According to the correlation matrix, it shows high and positive correlation between ingredient a and g, and relatively high correlation among 1) b and h; 2) d, f, h.

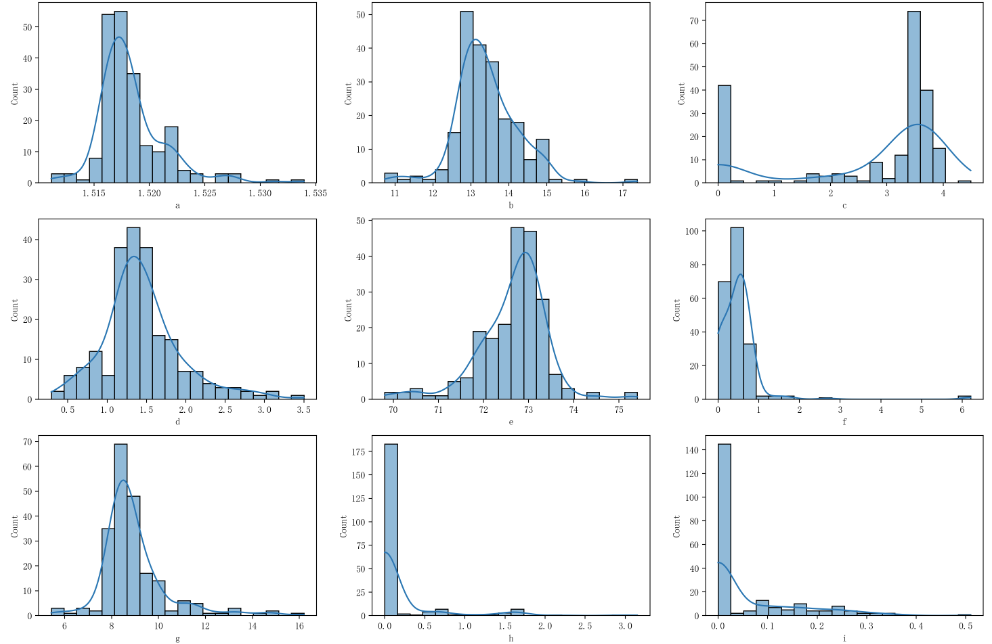
Meanwhile, it demonstrates negative relationship between ingredient: 1) a, d, e; 2) c, d, g, h; 3) f, g.

In other words, the ingredient a and g are always added together in the formulations. It is common that the ingredient mixed together with relatively high correlation. Conversely, the additives are not commonly added together with negative relationship stated above.

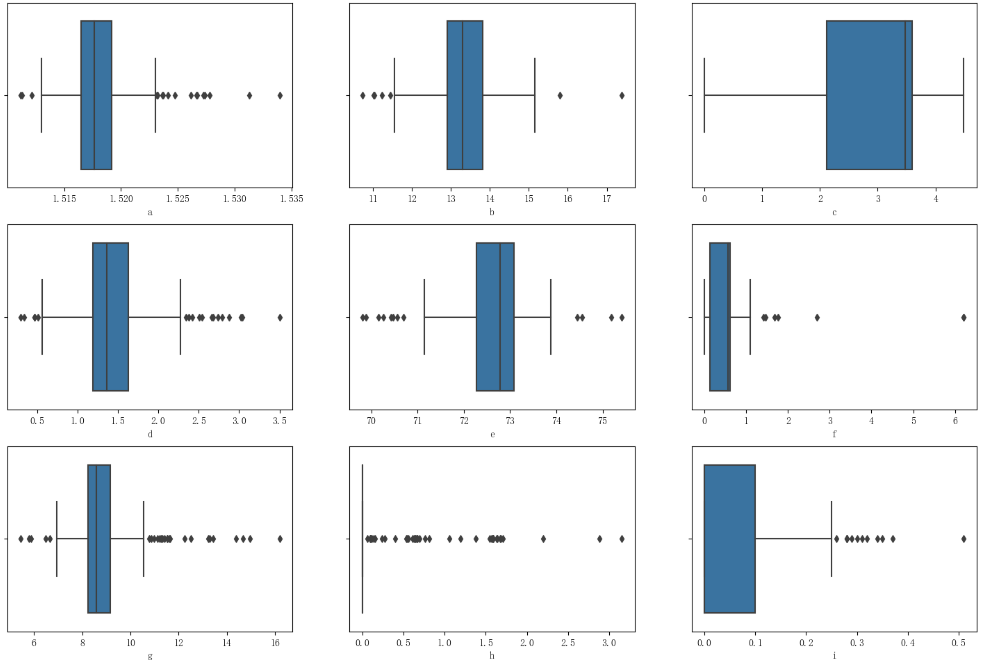
**Q1-b**

The distribution of each additive is shown in **Graph 1**, and the box-plot of each additive is shown in **Graph 2**.

Graph 1 and 2 indicate that most of the formulations are without or with less additive f, h and i. For the additive c, it shows two extremes -- some formulations without or with less additive c, but others formulations with high proportion of it. Finally, additives a, b, d, e, and g are the main ingredients for most of the formulation.



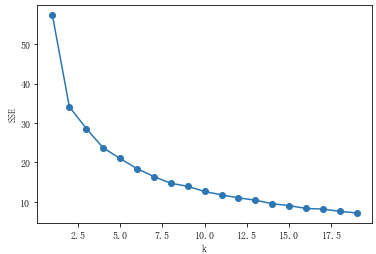
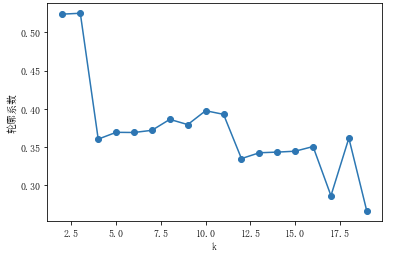
**Graph 1: Distribution of Additives**



**Graph 2: Box-plot of Additives**

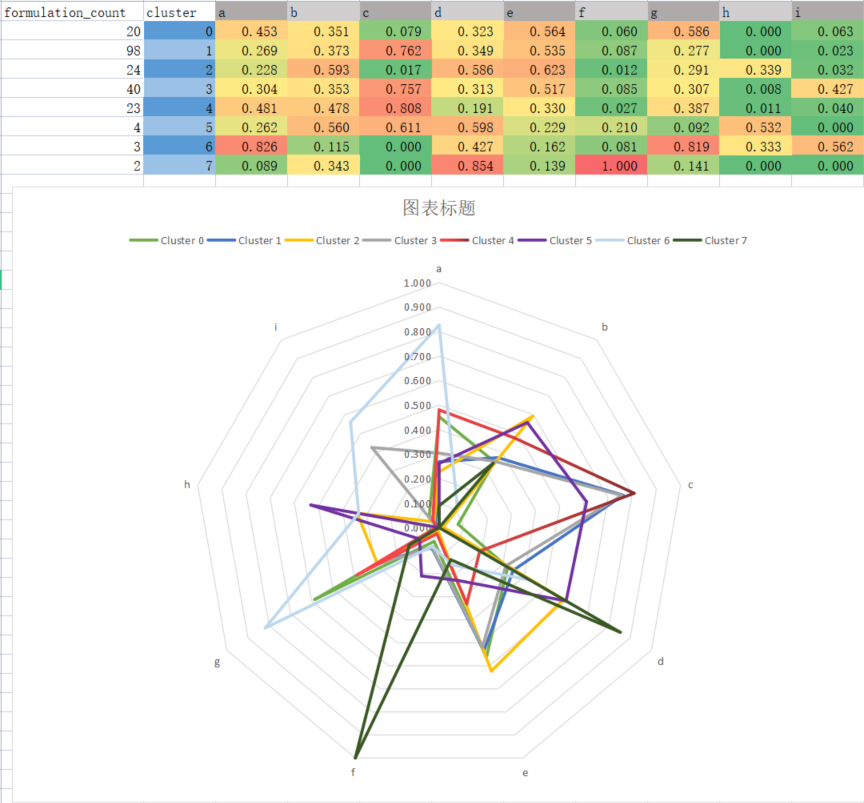
**Q1-c**

Based on elbow method and Silhouette Coefficient (shown in **Graph 3**), the optima k of KMeans is k=8. Thus, the distinctive number of formulations present in the dataset is **8**.

****

**Graph 3: K Selection**

**Graph 4** shows the means of each additives (after max-min standardization) and the radar map of each cluster of formulations. It states that each cluster contains different proportions among 9 additives. For example, cluster 0 and cluster 2 (with similar proportion of additive c, e, f, and i) are significantly different on additive a, b, d, g, h.

****

**Graph 4: Clusters’ Additives Means and Radar Map**

**Q2**

**Feature Engineering**

Firstly, HA\_Harvested is dividend by 1000 to make it keep the same scale with other features. Temperature range is also created which is defined by Max\_Temp - Min\_Temp, indicating the variance of temperature of current month. Finally, The Date is converted to dummy variable of months.

Moreover, some aggregation features are generated:

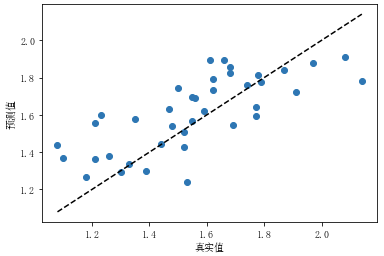
|  |  |  |  |
| --- | --- | --- | --- |
| Feature | New Feature Name | Aggregation mth | Statistics |
| SoilMoisture | SoilMoisture\_avg\_{mth}m\_agg  SoilMoisture\_std\_{mth}m\_agg | 3, 6, 9, 12 | Means, STD |
| Average\_Temp | Average\_Temp\_avg\_{mth}m\_agg  Average\_Temp\_std\_{mth}m\_agg | 3, 6, 9, 12 | Means, STD |
| Min\_Temp | Min\_Temp\_std\_{mth}m\_agg | 3, 6, 9, 12 | STD |
| Max\_Temp | Max\_Temp\_std\_{mth}m\_agg | 3, 6, 9, 12 | STD |
| Precipitation | Precipitation\_avg\_{mth}m\_agg  Precipitation\_std\_{mth}m\_agg | 3, 6, 9, 12 | Means, STD |
| HA\_Harvested | HA\_Harvested\_adj\_avg\_{mth}m\_agg  HA\_Harvested\_adj\_std\_{mth}m\_agg | 3, 6, 9, 12 | Means, STD |
| temp\_range | temp\_range\_avg\_{mth}m\_agg  temp\_range\_std\_{mth}m\_agg | 3, 6, 9, 12 | Means, STD |

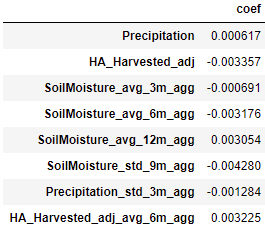
The features above calculate the statistics of original factor within recent x (3, 6, 9, 12) months, implying the lagging effects of the weather on the FFB.

**Modeling**

The model input (X) 68 features into the linear regression model. However, it will cause problem if these features are thrown into the trainer. Thus, we first run the Lasso regression the do the features selection, and the parameter is set as Lasso(alpha = 1, precompute = True, tol = 1e-7). After the features selection by Lasso regression, the linear regression is run with the selected features. The samples split by 70% for training and 30% of testing. Because of aggregation features with 12 months, I tried exclude the first 12 rows of data and compared the results without significantly different. Thus the model input kept the first 12 rows of data where the aggregation features does not aggregate the whole period.

The R square is 0.5 and MSE is 0.03 of this model. The results are shown in following graphs.





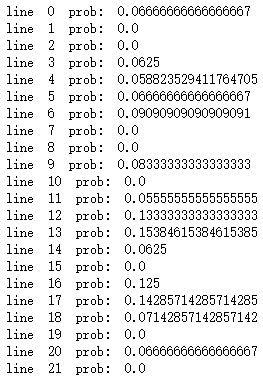
**Analysis**

The model indicates that the FFB is positively affected by the recent precipitation, last 12 months average of soil moisture, and last 6 months average of HA\_Harvested. At the same time, it shows negative effects of recent HA\_Harvested, average soil moisture within last 3/6 month, the variance of soil moisture in last 9 months and precipitation in last 3 months.

**Q3**

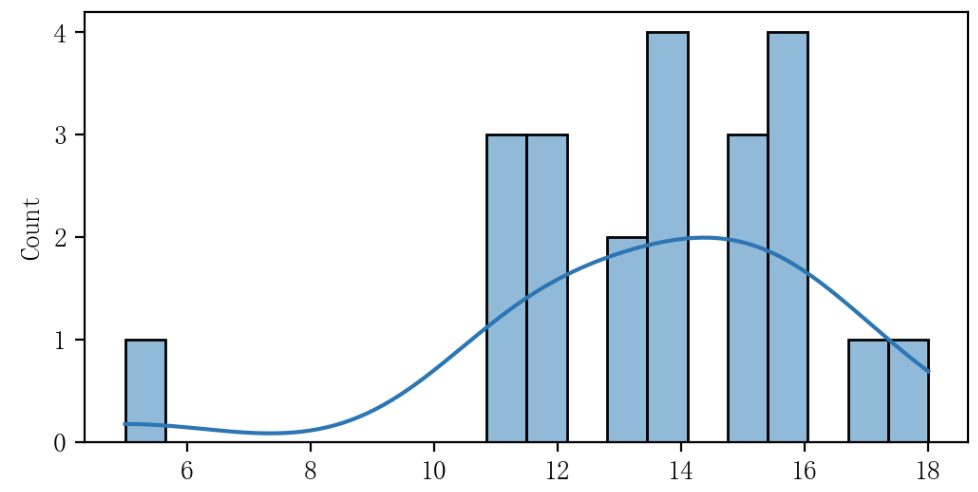
**Q3-a**

1. What is the probability of the word “data” occurring in each line ?



**Q3-b**

1. What is the distribution of distinct word counts across all the lines ?



**Q3-c**

1. What is the probability of the word “analytics” occurring after the word “data” ?

Answer: It is 1/3

