Nurul Fatiha binti Md Nor

1. A customer informed their consultant that they have developed several formulations of petrol that gives different characteristics of burning pattern. The formulations are obtaining by adding varying levels of additives that, for example, prevent engine knocking, gum prevention, stability in storage, 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.

Manual Descriptive Analysis

Original Dataset

	a	b	c	d	e	f	g	h	i
sum	324.9302	2869.28	574.49	309.21	15547.3	106.37	1916.79	37.46	12.2
mean	1.518365421	13.4079	2.68453	1.44491	72.6509	0.49706	8.95696	0.17505	0.05701
Varians	9.22254E-06	0.66684	2.08054	0.24927	0.59992	0.42535	2.02537	0.24723	0.00949

ANOVA Descriptive Analysis

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
a	214	324.9302	1.518365	9.22E-06
b	214	2869.28	13.40785	0.666841
c	214	574.49	2.684533	2.08054
d	214	309.21	1.444907	0.24927
e	214	15547.3	72.65093	0.599921
f	214	106.37	0.497056	0.425354
g	214	1916.79	8.956963	2.025366
h	214	37.46	0.175047	0.247227
i	214	12.2	0.057009	0.009494

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	943261.1	8	117907.6	168332	0	1.943226
Within Groups	1342.757	1917	0.700447			
Total	944603.8	1925				

This test is not applicable to run the ANOVA because it does not have replication, here some prove that show it is not applicable and being rejected.

On ANOVA step, the ***Count*** is referring on how many additives is used to differentiate the characteristics of burning pattern. So, for this finding, the total pattern is **214**.

The ***SUM*** column is referring to the total value of each additive. That's mean, the ANOVA calculated the sum of the value for each additive from ***a*** to ***i*** and the answer shows in the table.

The ***Average*** is referring to the mean of each additive, and same goes with ***Varian***. So, below is the Summary of the descriptive analysis.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
a	214	324.9302	1.518365	9.22E-06
b	214	2869.28	13.40785	0.666841
c	214	574.49	2.684533	2.08054
d	214	309.21	1.444907	0.24927
e	214	15547.3	72.65093	0.599921
f	214	106.37	0.497056	0.425354
g	214	1916.79	8.956963	2.025366
h	214	37.46	0.175047	0.247227
i	214	12.2	0.057009	0.009494

Then, below is the ANOVA table, which will give the sum of squares ***(SS),*** degrees if freedom ***(fd),*** mean squares ***(MS),*** statistic ***(F),*** the test ***(P-Value)*** and the critical value ***(F crit).***

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	943261.1	8	117907.6	168332	0	1.943226
Within Groups	1342.757	1917	0.700447			
Total	944603.8	1925				

Some of Squares (ss)

So, between group, ***some of squares (SS)*** which is ***943261.1*** saying that the ***SS*** is between the additives group **(from a to i).**

Then, the value of within group is based on; ***some of squares (SS)*** which is ***1342.757*** saying that the ***SS*** is between the count of each additive **(which is 214)** for all additive **(from a to i)** by using formula provided. After that, those additives (a to i) is added.

Then, by adding the value of Between Groups and Within Groups, we can get the Total which is **944603.8.**

Degree of Freedom (df)

For ***degree of freedom (df)*** for between group is using K-1 formula. So, in this case, 9-1, so the answer is **8**.

For ***degree of freedom (df)*** for within group is using K(n-1) formula.
So, it will be 9 (214-1) = **1917**

Then, by adding the value of Between Groups and Within Groups, we can get the Total which is **1925**.

Mean Square (MS)
We can get the value of between Groups of ***(MS)*** by dividing ***(SS) – (df).*** So, it will be, 943261.1 / 8 = **117907.6**.

Then, for the value of within Groups of ***(MS)***, same process, by dividing ***(SS) – (df).*** So, it will be, 1342.757 / 1917 = **0.700447.**

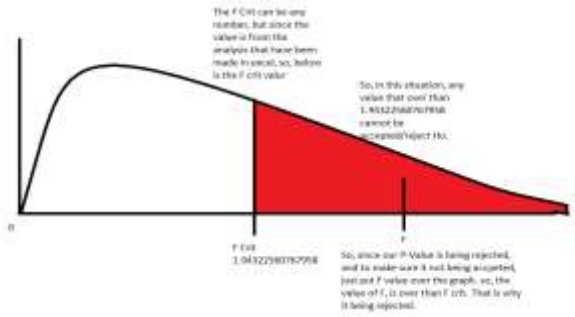
F Statistic (F)
So, the **F** is the **(MS)** between to distinguish 117907.6 and 0.700447. so, 117907.6 / 0.700447 = **168332**.

P-Value
This is the important result so that we know if the result is accepting the null hypothesis or not. But we need to remember the formula is Ho: $\mu\alpha = \mu\text{ь} = \mu c$ until i
So, because the P-Value is less than 0.05 which is 0.

Now, the alternative Hypothesis is **Hα; At Least One Inequality**

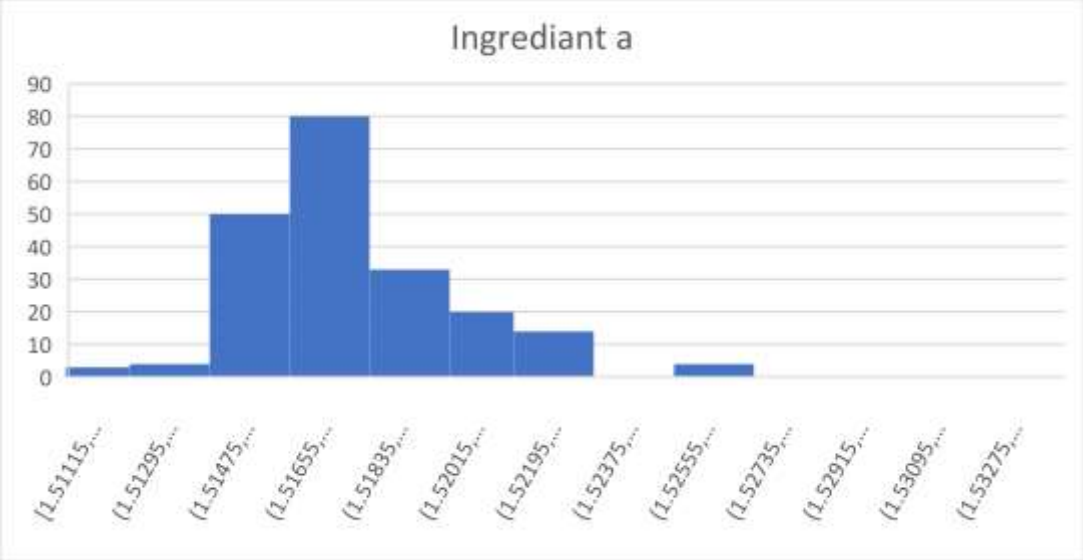
So, at least, the additives from **a until i** is different with the other. Of course, the assumption have been made when doing the ANOVA, which all the observation are coming form a normal distribution that this 9 additives have an equal variances except of "**e additive".** so, this **"e additive"** can be measured by using others test.
So, back to the P-value, the values show is less than alpha (0.05) which is I get **0** value. So, for this practice**, I cannot Accept the Ho.** This is because, to make the result acceptable, the value must not less than 0.05.

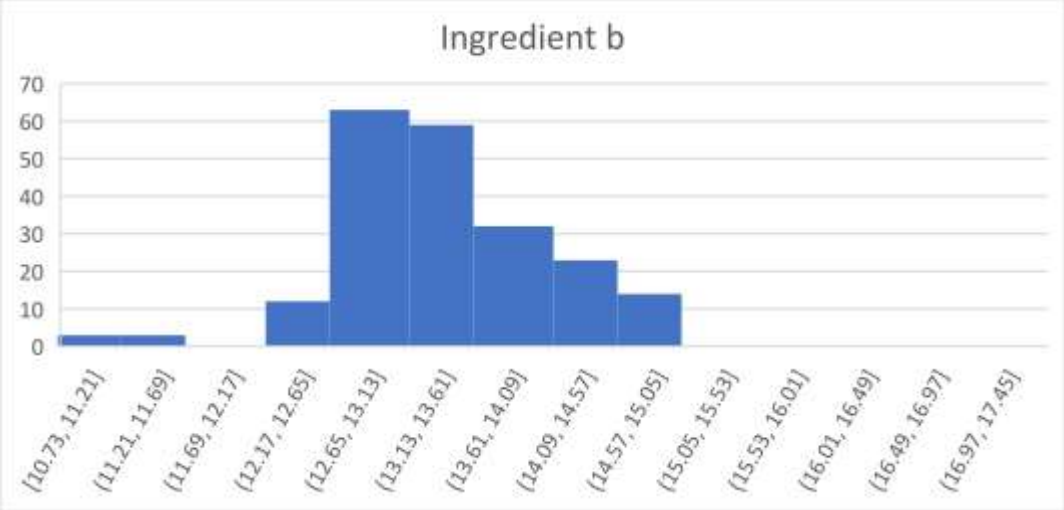
F Critical
For the F crit, the value, the example below.



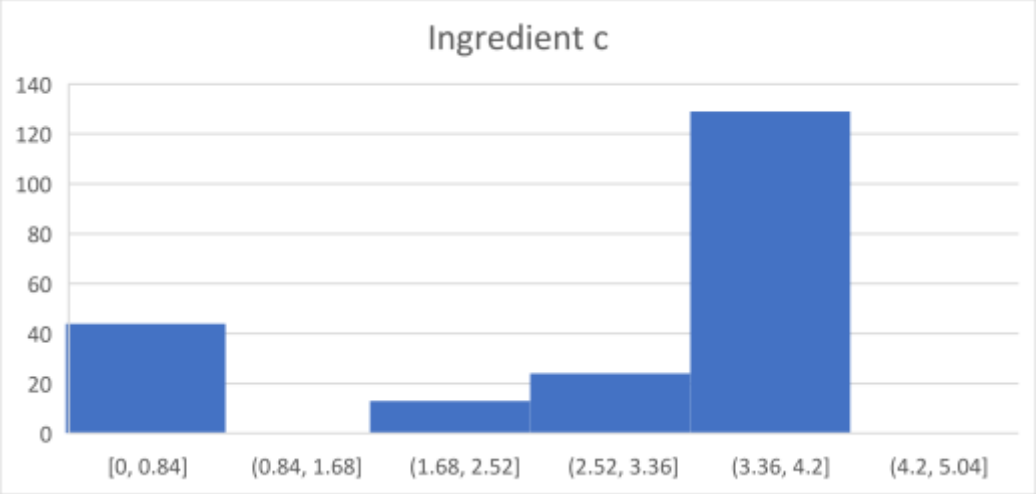
b. A graphical analysis of the additives, including a distribution study.
Answer:

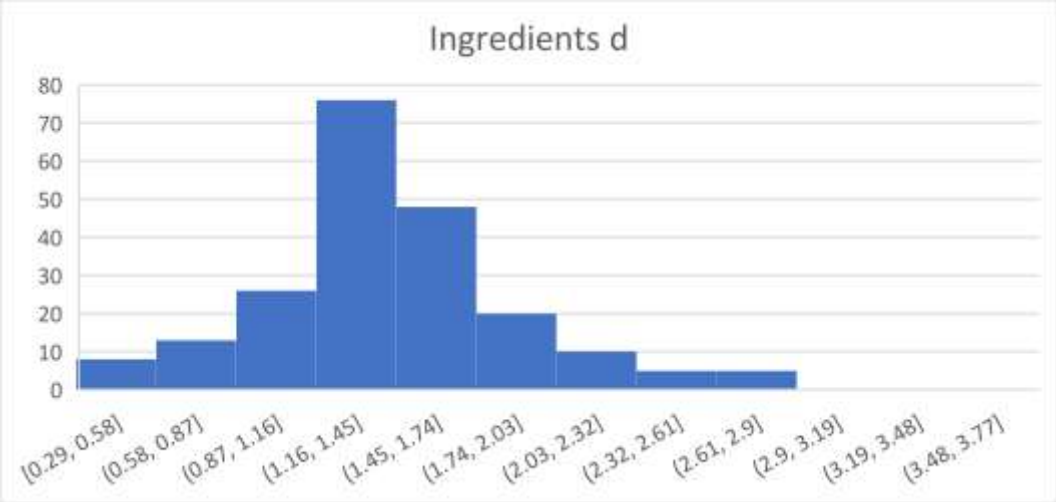
Normal distribution means that, the graph looks like bell shape, but not normal distribution is vice versa. So, in this case, the normal distribution is on ingredients **d.**

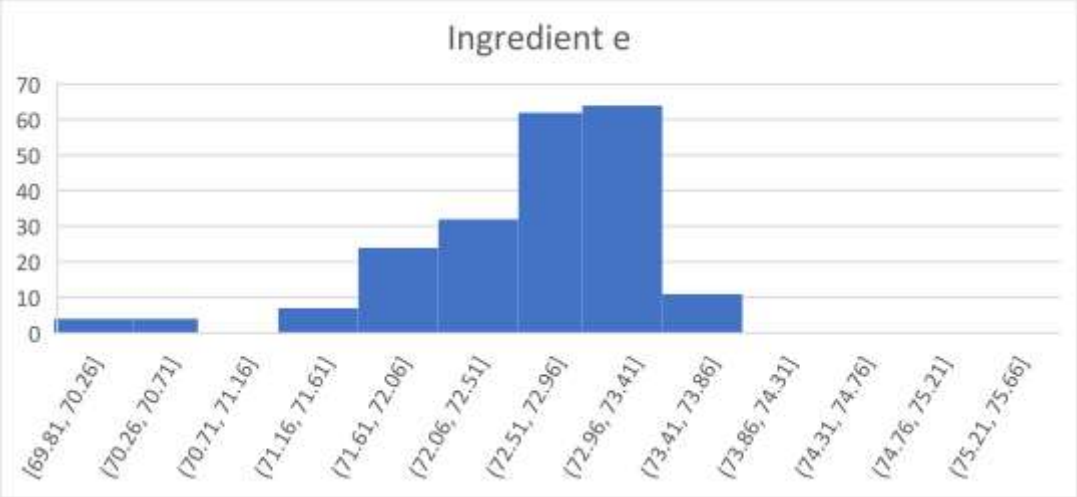




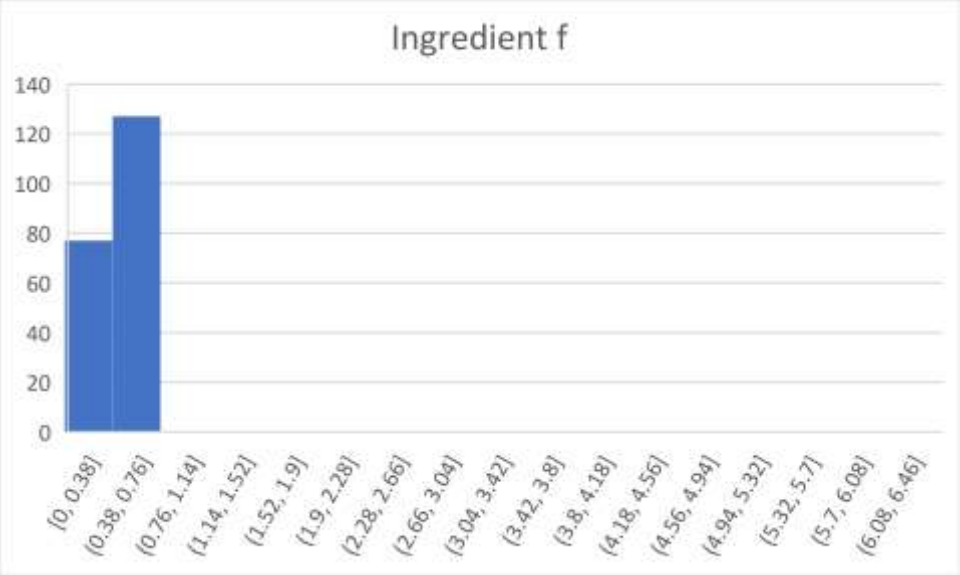
Nurul Fatiha binti Md Nor

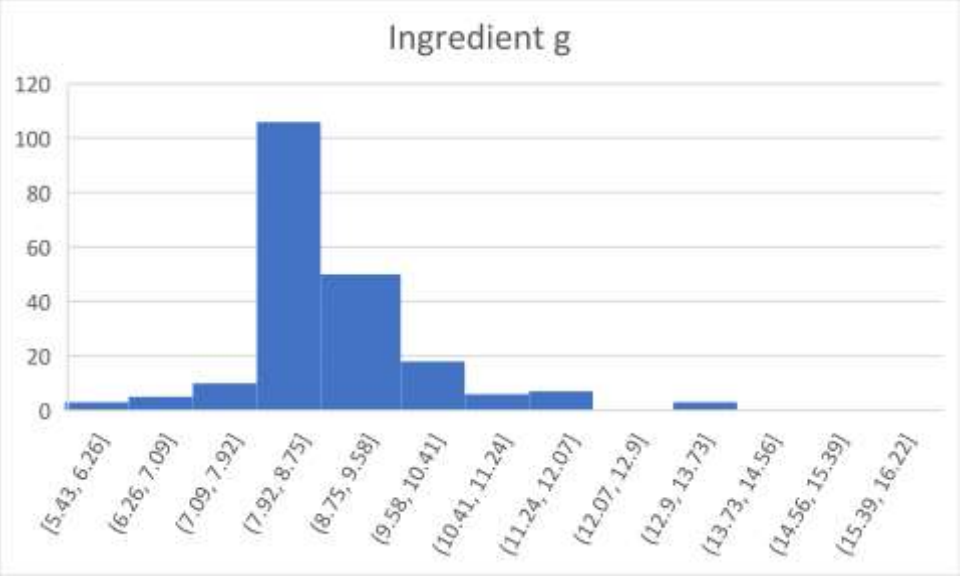


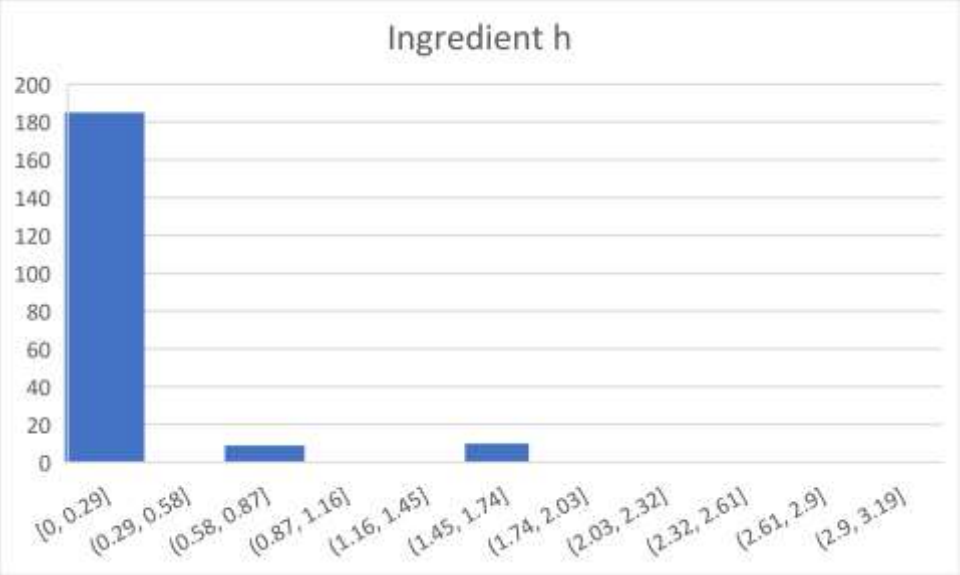


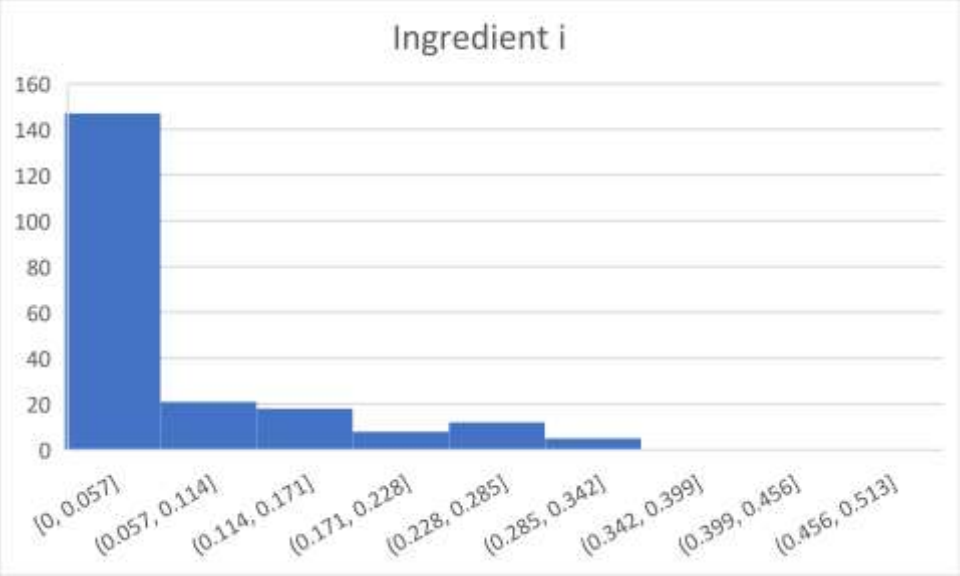


Nurul Fatiha binti Md Nor






Ingredient h
200
180
160
140
120
100
80
60
40
20
0
[0, 0.29]
(0.29, 0.58]
(0.58, 0.87]
(0.87, 1.16]
(1.16, 1.45]
(1.45, 1.74]
(1.74, 2.03]
(2.03, 2.32]
(2.32, 2.61]
(2.61, 2.9]
(2.9, 3.19]


Ingredient i
160
140
120
100
80
60
40
20
0
[0, 0.057]
(0.057, 0.114]
(0.114, 0.171]
(0.171, 0.228]
(0.228, 0.285]
(0.285, 0.342]
(0.342, 0.399]
(0.399, 0.456]
(0.456, 0.513]

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

The clustering (unsupervised learning) that have been chosen is K-mean clustering, which commonly used when we have un-label data for example data without define categories or groups.

K-means clustering algorithm computes the centroids and iterates until the optimal centroid is reached. In this practice, to get the result, I have performed 3 iterations to get "0" as a result for the iterations after minus the result of cluster from iteration 2 and 3. (result as given attachment)

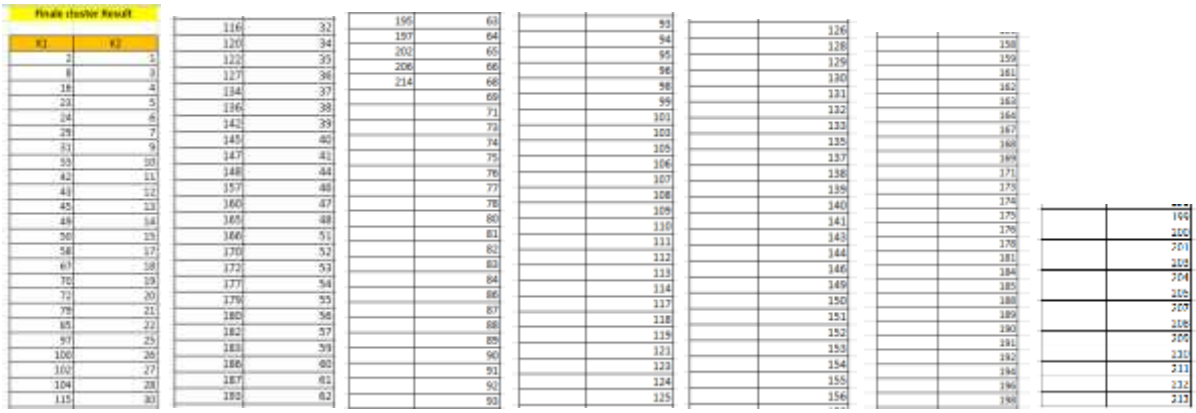
In this algorithm, the data points are assigned to a cluster in such a manner that the sim of the squared distance between the data points and centroid would be minimum.

How I am doing the K-means is by knowing how many clusters that we need. So, in this practice I am using only 2 cluster, for example K=2. Then, I must randomly guess the K cluster center location. So, I must choose randomly any center locations. Then, I need find each of data points to closest center and then, each center need to find the centroid of the points owns.

In this practice also, I am using single linkage method, where I have to choose the minimum/maximum value and decide which cluster are belong to which value. I put the low value as K2, and Highest value as K1.

I am not using any trial software since I have some technical problem with my computer. So, I do it using excel formula.

*So, as the prove, I attached together all of the needed data in excel sheet.



2. A team of plantation planners are concerned about the yield of oil palm trees, which seems to fluctuate. They have collected a set of data and needed help in analysing on how external factors influence fresh fruit bunch (FFB) yield. Some experts are of opinion that the flowering of oil palm tree determines the FFB yield, and are linked to the external factors. Perform the analysis, which requires some study on the background of oil palm tree physiology.

   (Refer attachment palm_ffb.csv)

   The answer: By using JASP software

   file:///C:/Users/nurul.fatiha.mdnor/Downloads/palm_ffb%20(1).html

Results

Descriptive Statistics

Descriptive Statistics

	SoilMoisture	Average_Temp	Min_Temp	Max_Temp	Precipitation	Working_days	HA_Harvested	FFB_Yield
Valid	130	130	130	130	130	130	130	130
Missing	0	0	0	0	0	0	0	0
Mean	527.647	26.850	21.379	33.852	188.981	24.754	793404.492	1.602
Std. Deviation	57.368	0.651	0.689	1.080	80.237	1.239	34440.894	0.282
Shapiro-Wilk	0.982	0.984	0.932	0.986	0.978	0.912	0.986	0.983
P-value of Shapiro-Wilk	0.084	0.137	< .001	0.203	0.037	< .001	0.192	0.096
Minimum	380.700	25.158	18.900	31.100	2.000	21.000	683431.944	1.080
Maximum	647.300	28.580	22.600	36.000	496.100	27.000	882254.225	2.270

Correlation

Correlation Table

Variable		SoilMoisture	Average_Temp	Min_Temp	Max_Temp	Precipitation	FFB_Yield	Working_days	HA_Harvested
1. SoilMoisture	Pearson's r	—							
	p-value	—							
	Upper 95% CI	—							
	Lower 95% CI	—							
	Spearman's rho	—							
	p-value	—							
	Upper 95% CI	—							
	Lower 95% CI	—							
2. Average_Temp	Pearson's r	-0.650***	—						
	p-value	< .001	—						
	Upper 95% CI	-0.538	—						
	Lower 95% CI	-0.739	—						
	Spearman's rho	-0.611***	—						
	p-value	< .001	—						
	Upper 95% CI		—						
	Lower 95% CI		—						
3. Min_Temp	Pearson's r	0.016	0.180*	—					
	p-value	0.858	0.040	—					
	Upper 95% CI	0.188	0.342	—					
	Lower 95% CI	-0.157	0.008	—					
	Spearman's rho	0.008	0.150	—					
	p-value	0.927	0.088	—					
	Upper 95% CI			—					
	Lower 95% CI			—					
4. Max_Temp	Pearson's r	-0.500***	0.761***	-0.125	—				
	p-value	< .001	< .001	0.157	—				
	Upper 95% CI	-0.359	0.825	0.048	—				
	Lower 95% CI	-0.619	0.678	-0.291	—				
	Spearman's rho	-0.466***	0.736***	-0.164	—				
	p-value	< .001	< .001	0.062	—				
	Upper 95% CI				—				
	Lower 95% CI				—				

5. Precipitation	Pearson's r	0.562***	-0.369***	0.346***	-0.461***	–			
	p-value	< .001	< .001	< .001	< .001	–			
	Upper 95% CI	0.661	-0.211	0.489	-0.314	–			
	Lower 95% CI	0.420	-0.509	0.185	-0.587	–			
	Spearman's rho	0.535***	-0.313***	0.368***	-0.427***	–			
	p-value	< .001	< .001	< .001	< .001	–			
	Upper 95% CI					–			
	Lower 95% CI					–			
6. FFB_Yield	Pearson's r	-0.003	-0.005	0.104	-0.071	0.290***	–		
	p-value	0.971	0.951	0.240	0.421	< .001	–		
	Upper 95% CI	0.169	0.167	0.271	0.102	0.440	–		
	Lower 95% CI	-0.175	-0.178	-0.070	-0.240	0.124	–		
	Spearman's rho	-0.054	-0.036	0.084	-0.111	0.312***	–		
	p-value	0.542	0.682	0.340	0.211	< .001	–		
	Upper 95% CI						–		
	Lower 95% CI						–		
7. Working_days	Pearson's r	-0.057	0.076	0.068	-0.039	0.128	0.116	–	
	p-value	0.519	0.388	0.439	0.659	0.147	0.187	–	
	Upper 95% CI	0.116	0.245	0.238	0.134	0.294	0.283	–	
	Lower 95% CI	-0.227	-0.097	-0.105	-0.210	-0.045	-0.057	–	
	Spearman's rho	-0.056	0.064	-0.011	-0.026	0.076	0.100	–	
	p-value	0.529	0.471	0.903	0.773	0.387	0.260	–	
	Upper 95% CI							–	
	Lower 95% CI							–	
8. HA_Harvested	Pearson's r	-0.327***	0.447***	0.024	0.315***	-0.266**	-0.350***	0.049	–
	p-value	< .001	< .001	0.783	< .001	0.002	< .001	0.581	–
	Upper 95% CI	-0.164	0.575	0.196	0.462	-0.098	-0.189	0.219	–
	Lower 95% CI	-0.472	0.297	-0.148	0.151	-0.419	-0.493	-0.124	–
	Spearman's rho	-0.347***	0.497***	-0.017	0.330***	-0.277***	-0.386***	0.045	–
	p-value	< .001	< .001	0.851	< .001	0.001	< .001	0.611	–
	Upper 95% CI								–
	Lower 95% CI								–

* p < .05, ** p < .01, *** p < .001

Assumption checks

Shapiro-Wilk Test for Multivariate Normality

Shapiro-Wilk	p
0.875	< .001

Linear Regression

Model Summary - FFB_Yield

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.405	0.164	0.151	0.260
H_1	0.500	0.250	0.214	0.250

Note. Null model includes Precipitation, HA_Harvested

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_0	Regression	1.681	2	0.841	12.475	< .001
	Residual	8.559	127	0.067		
	Total	10.240	129			
H_1	Regression	2.565	6	0.427	6.850	< .001
	Residual	7.676	123	0.062		
	Total	10.240	129			

Note. Null model includes Precipitation, HA_Harvested

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	3.370	0.564		5.975	< .001
	Precipitation	7.425e-4	2.955e-4	0.211	2.513	0.013
	HA_Harvested	-2.405e-6	6.884e-7	-0.294	-3.494	< .001
H_1	(Intercept)	3.055	1.458		2.096	0.038
	SoilMoisture	-0.001	5.681e-4	-0.232	-2.008	0.047
	Average_Temp	0.097	0.070	0.224	1.392	0.166
	Min_Temp	-0.027	0.038	-0.065	-0.692	0.491
	Max_Temp	-0.016	0.035	-0.061	-0.448	0.655
	Precipitation	0.001	3.713e-4	0.388	3.673	< .001
	HA_Harvested	-3.292e-6	7.222e-7	-0.402	-4.558	< .001

3. Feed the following paragraph into your favourite data analytics tool, and answer the following;

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

      the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics uses. In addition, it can come from a mix of internal systems and external data sources.

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

      exploratory data analysis (EDA), which aims to find patterns and relationships in data, and confirmatory data analysis (CDA), which applies statistical techniques to determine whether hypotheses about a data set are true or false.

   c. What is the probability of the word “analytics” occurring after the word “data”?
      Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves analysis of numerical data with quantifiable variables that can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio, and video, including common phrases, themes and points of view.