PRE-SCREEN – NUR ADNIN BINTI MOHD NASIR

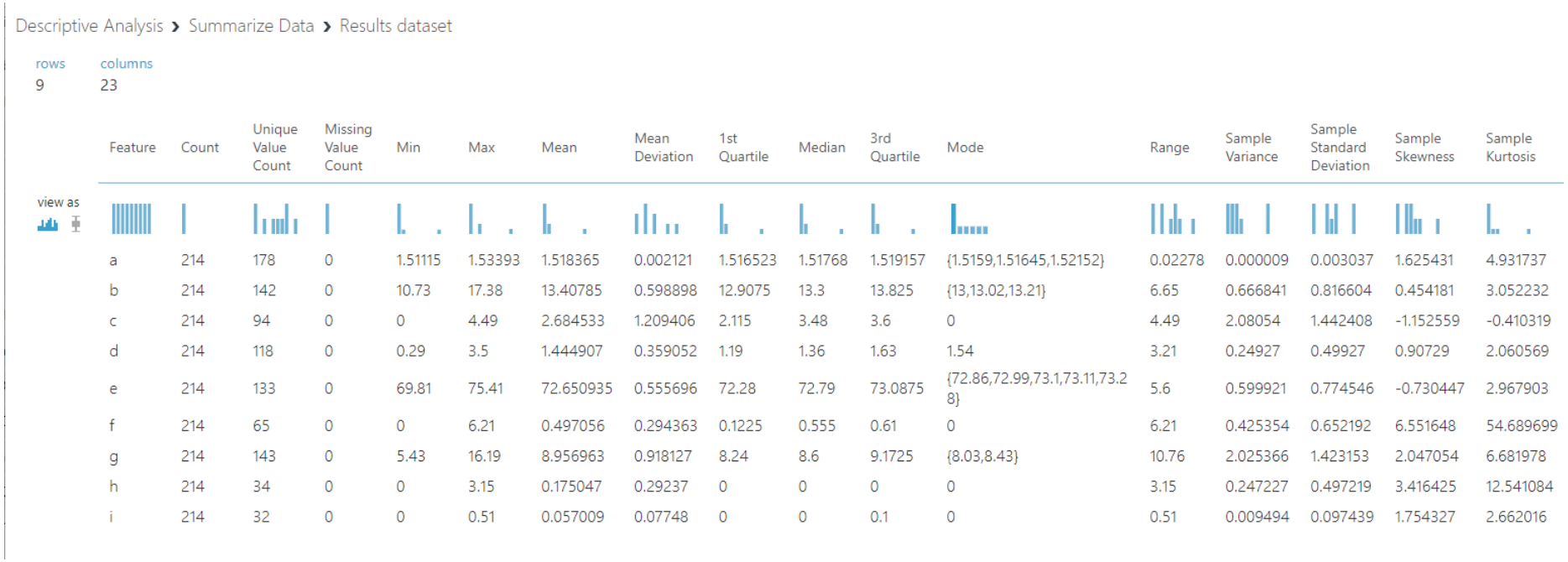
Platform : Azure Machine Learning

Question 1 : a) Descriptive analysis of additives

Link : https://gallery.cortanaintelligence.com/Experiment/Descriptive-Analysis

1. Use Summarize Data module to generate descriptive analysis for each of the additives.

Output :



Descriptive Analysis › Summarize Data › Results dataset

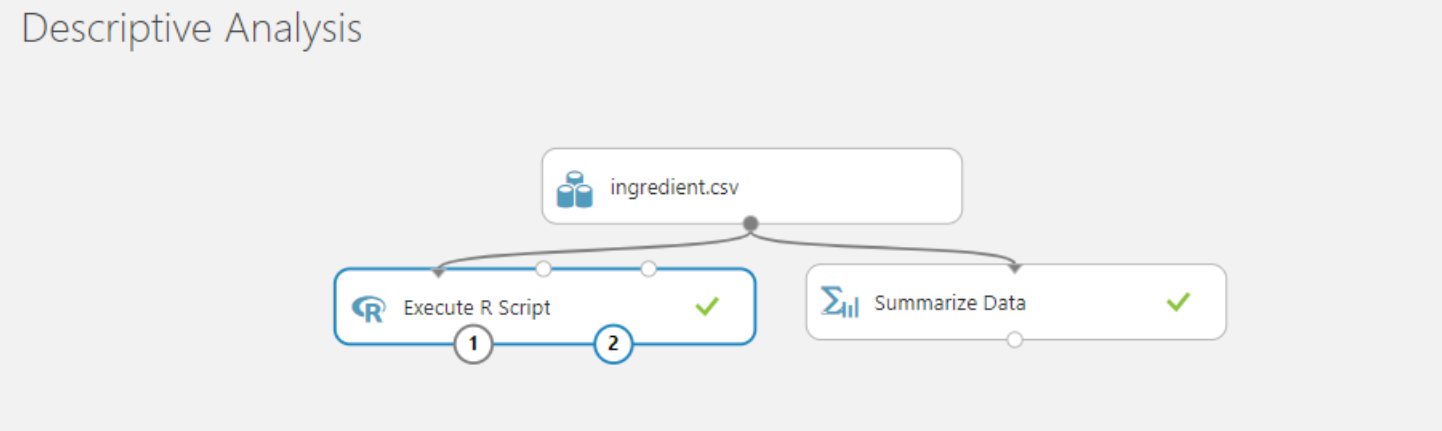
rows 9 columns 23

Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation	1st Quartile	Median	3rd Quartile	Mode	Range	Sample Variance	Sample Standard Deviation	Sample Skewness	Sample Kurtosis
view as																
a	214	178	0	1.51115	1.53393	1.518365	0.002121	1.516523	1.51768	1.519157	{1.5159,1.51645,1.52152}	0.02278	0.000009	0.003037	1.625431	4.931737
b	214	142	0	10.73	17.38	13.40785	0.598898	12.9075	13.3	13.825	{13,13.02,13.21}	6.65	0.666841	0.816604	0.454181	3.052232
c	214	94	0	0	4.49	2.684533	1.209406	2.115	3.48	3.6	0	4.49	2.08054	1.442408	-1.152559	-0.410319
d	214	118	0	0.29	3.5	1.444907	0.359052	1.19	1.36	1.63	1.54	3.21	0.24927	0.49927	0.90729	2.060569
e	214	133	0	69.81	75.41	72.650935	0.555696	72.28	72.79	73.0875	{72.86,72.99,73.1,73.11,73.28}	5.6	0.599921	0.774546	-0.730447	2.967903
f	214	65	0	0	6.21	0.497056	0.294363	0.1225	0.555	0.61	0	6.21	0.425354	0.652192	6.551648	54.689699
g	214	143	0	5.43	16.19	8.956963	0.918127	8.24	8.6	9.1725	{8.03,8.43}	10.76	2.025366	1.423153	2.047054	6.681978
h	214	34	0	0	3.15	0.175047	0.29237	0	0	0	0	3.15	0.247227	0.497219	3.416425	12.541084
i	214	32	0	0	0.51	0.057009	0.07748	0	0	0.1	0	0.51	0.009494	0.097439	1.754327	2.662016

Findings :

There is no missing value in the dataset. Additives (a) has the lowest variance while additives (c) have the highest variance.

2. Using R module to generate basic statistic summaries and boxplot for each additive. Using summary() and boxplot() function.



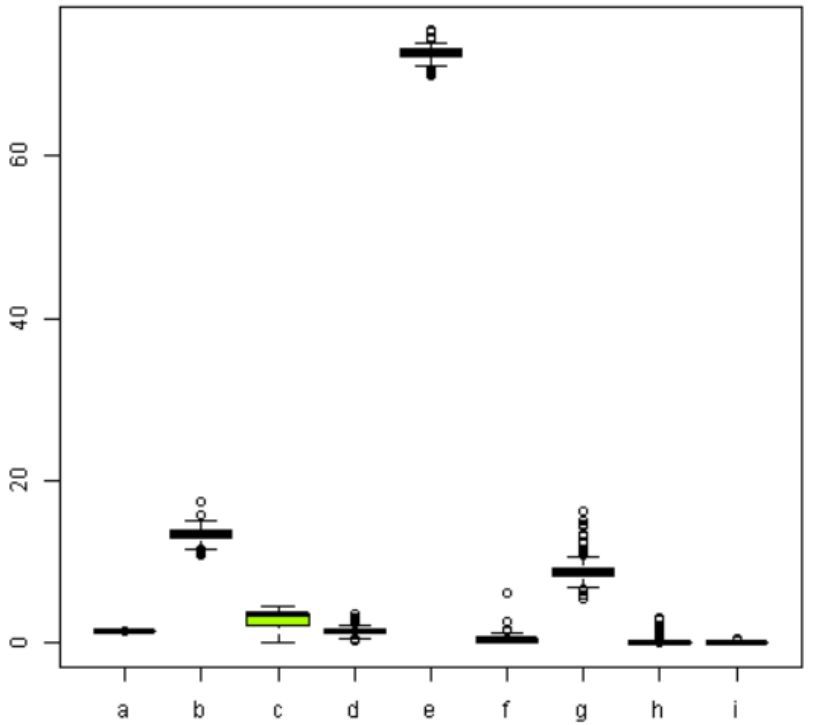
R Script:

R Script

```r
# Map 1-based optional input ports to variables
dataset1 <- maml.mapInputPort(1) # class: data.frame


# Contents of optional Zip port are in ./src/
# source("src/yourfile.R");
# load("src/yourData.rdata");

# Sample operation
summary(dataset1);


# You'll see this output in the R Device port.
# It'll have your stdout, stderr and PNG graphics device(s).
boxplot(dataset1, col = rainbow(ncol(dataset1)));

# Select data.frame to be sent to the output Dataset port
maml.mapOutputPort("dataset1");
```

Output :

```
       a               b               c               d
Min.   :1.511   Min.   :10.73   Min.   :0.000   Min.   :0.290
1st Qu.:1.517   1st Qu.:12.91   1st Qu.:2.115   1st Qu.:1.190
Median :1.518   Median :13.30   Median :3.480   Median :1.360
Mean   :1.518   Mean   :13.41   Mean   :2.685   Mean   :1.445
3rd Qu.:1.519   3rd Qu.:13.82   3rd Qu.:3.600   3rd Qu.:1.630
Max.   :1.534   Max.   :17.38   Max.   :4.490   Max.   :3.500
       e               f               g               h
Min.   :69.81   Min.   :0.0000   Min.   : 5.430   Min.   :0.000
1st Qu.:72.28   1st Qu.:0.1225   1st Qu.: 8.240   1st Qu.:0.000
Median :72.79   Median :0.5550   Median : 8.600   Median :0.000
Mean   :72.65   Mean   :0.4971   Mean   : 8.957   Mean   :0.175
3rd Qu.:73.09   3rd Qu.:0.6100   3rd Qu.: 9.172   3rd Qu.:0.000
Max.   :75.41   Max.   :6.2100   Max.   :16.190   Max.   :3.150
       i
Min.   :0.00000
1st Qu.:0.00000
Median :0.00000
Mean   :0.05701
3rd Qu.:0.10000
Max.   :0.51000
```



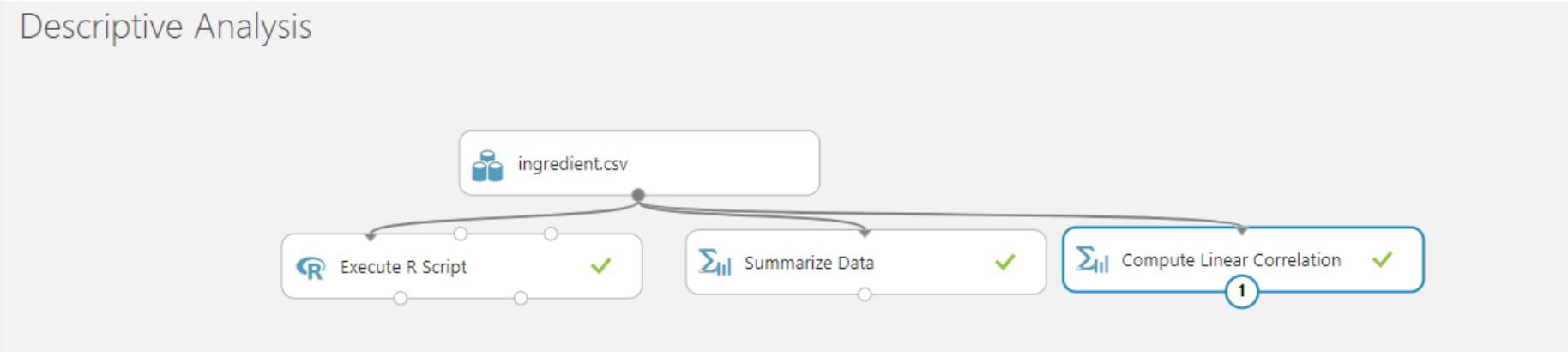
Findings :

Additive (e) is significantly different from the others and is above all of the other additives.

Additives (a), (d), (f), (h) and (i) are around the same level and is considered the lowest.

Almost all of the additives have outliers except additive (c). The content of additive (c) across all formulations are almost consistent.

3. Using Compute Linear Correlation module to see if there is any correlation between all the additives.



Output :

Descriptive Analysis › Compute Linear Correlation › Results dataset

rows 9 columns 9

view as

	a	b	c	d	e	f	g	h	i
a	1	-0.191885	-0.122274	-0.407326	-0.542052	-0.289833	0.810403	-0.000386	0.14301
b	-0.191885	1	-0.273732	0.156794	-0.069809	-0.266087	-0.275442	0.326603	-0.241346
c	-0.122274	-0.273732	1	-0.481799	-0.165927	0.005396	-0.44375	-0.492262	0.08306
d	-0.407326	0.156794	-0.481799	1	-0.005524	0.325958	-0.259592	0.479404	-0.074402
e	-0.542052	-0.069809	-0.165927	-0.005524	1	-0.193331	-0.208732	-0.102151	-0.094201
f	-0.289833	-0.266087	0.005396	0.325958	-0.193331	1	-0.317836	-0.042618	-0.007719
g	0.810403	-0.275442	-0.44375	-0.259592	-0.208732	-0.317836	1	-0.112841	0.124968
h	-0.000386	0.326603	-0.492262	0.479404	-0.102151	-0.042618	-0.112841	1	-0.058692
i	0.14301	-0.241346	0.08306	-0.074402	-0.094201	-0.007719	0.124968	-0.058692	1

Findings :

Additive (a) and (g) have high correlation.

Additive (a) and (e) are moderately correlated.

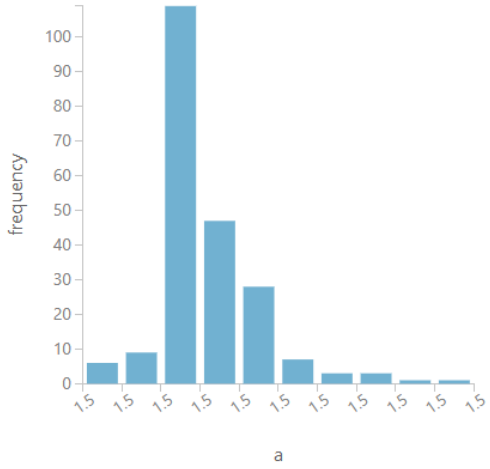
Additive (a) has little to no correlation to additive (h).

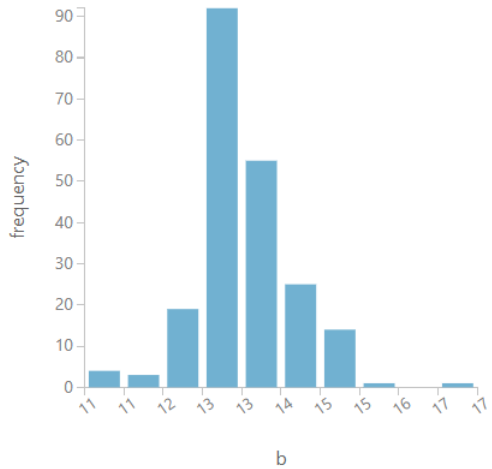
Additive (b) and (h) have low correlation.

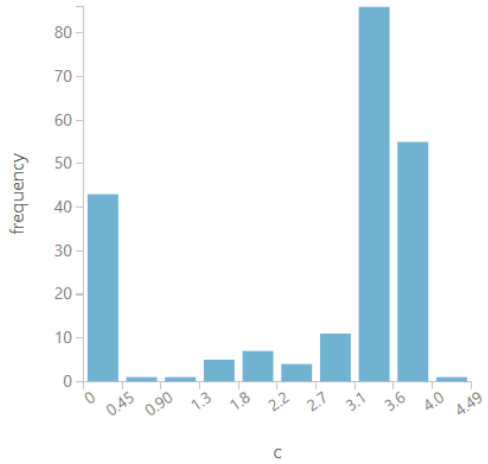
Additive (c) and (d) have low correlation.

Question 1 : b) Graphical analysis of additives and distribution study.

1. Histogram of each additives :



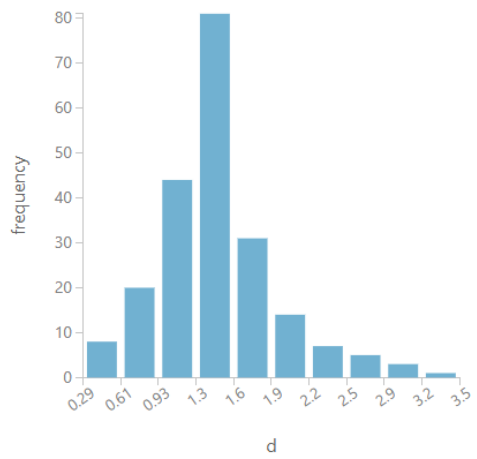


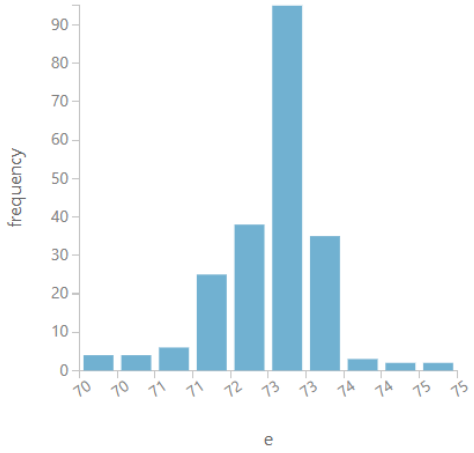


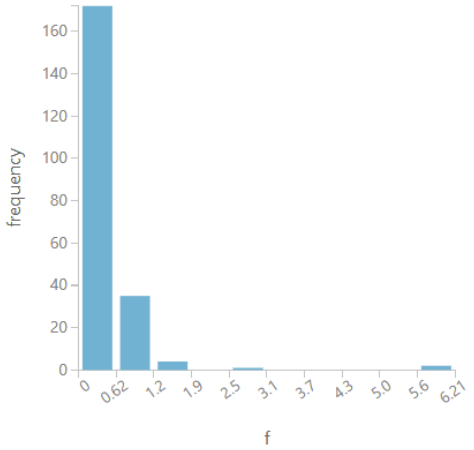
Findings :

Additive (a) and (b) have a right-skewed distribution.

Additive (c) have a bimodal distribution.



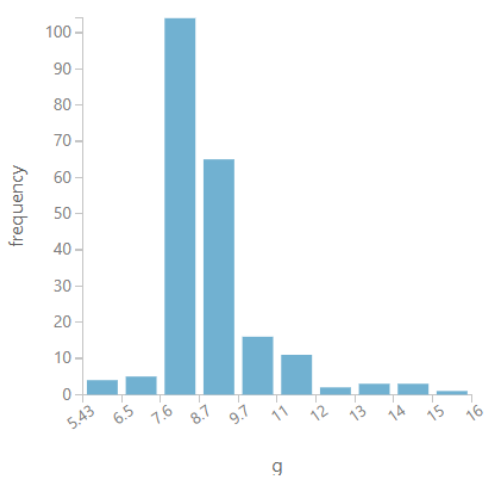


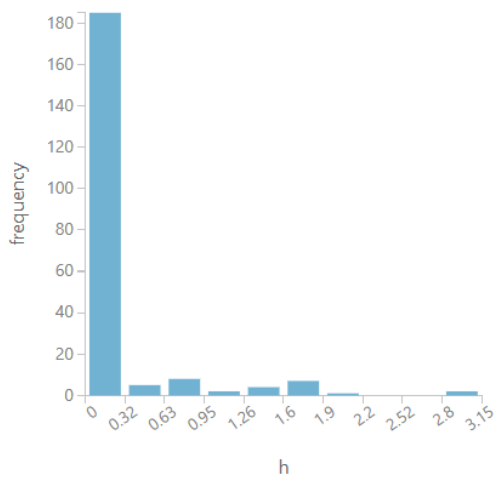


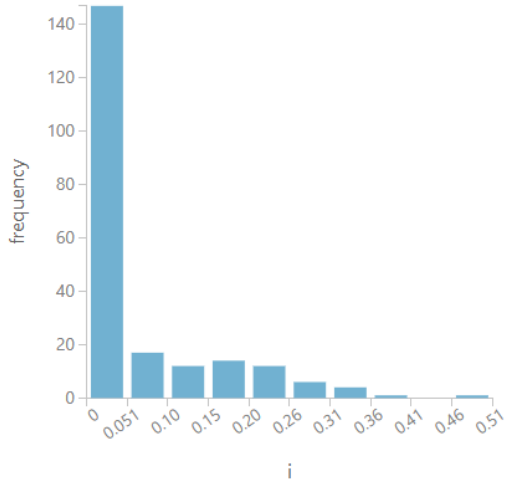
Findings :

Additive (d) and (f) are right-skewed.

Additive (e) is left-skewed.







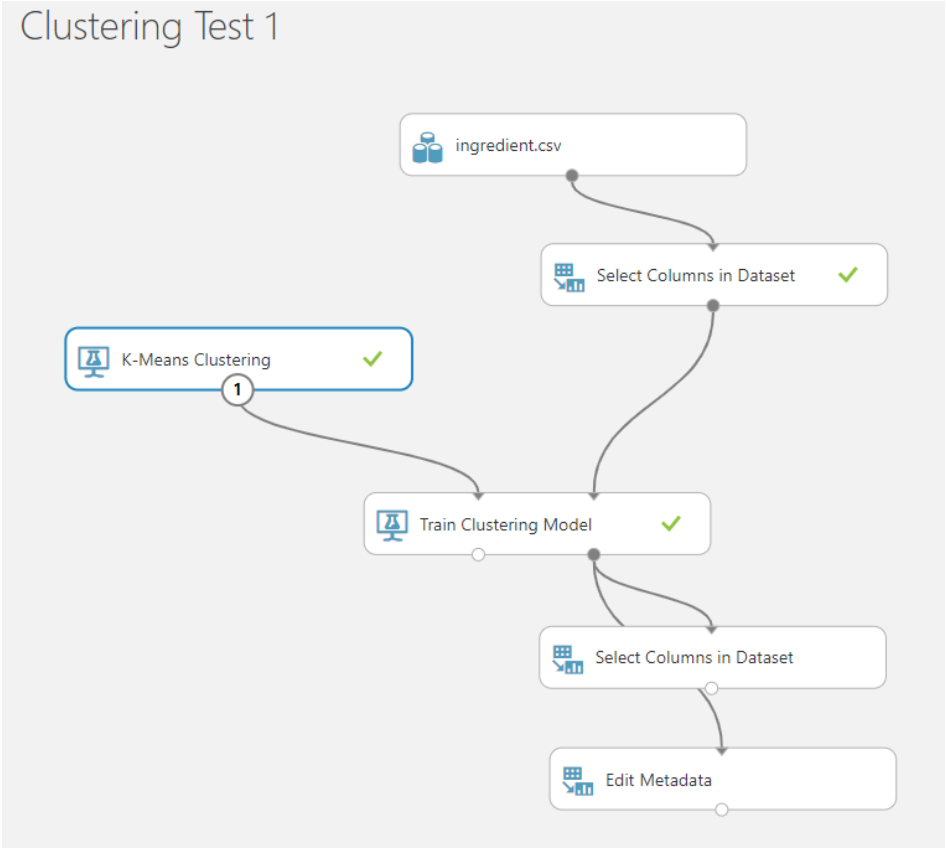
Findings :

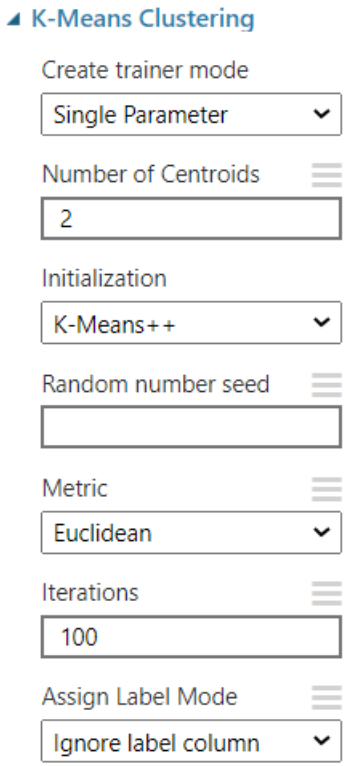
Additive (g) , (h) and (i) are right-skewed.

Question 1 : c) Clustering test to determine the distinctive number of formulations.

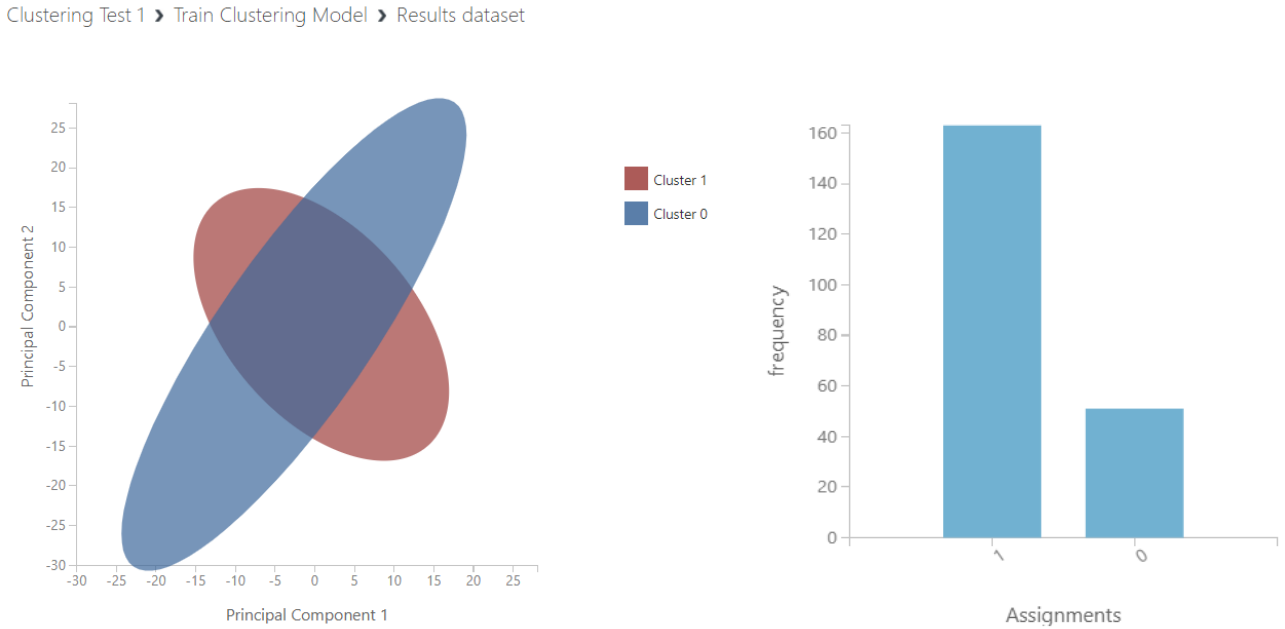
Link : https://gallery.cortanaintelligence.com/Experiment/Clustering-Test-1

1. Using K Means Clustering module to determine the number of clusters in the dataset.





Output :



Findings :

Cluster 1 contains 160 observations while Cluster 0 contains 54 observations.

Distance to Cluster No.0

Statistics

Mean	3.6183
Median	3.8684
Min	1.1981
Max	7.1205
Standard Deviation	0.9971
Unique Values	213
Missing Values	0
Feature Type	Numeric Score

Distance to Cluster No.1

Statistics

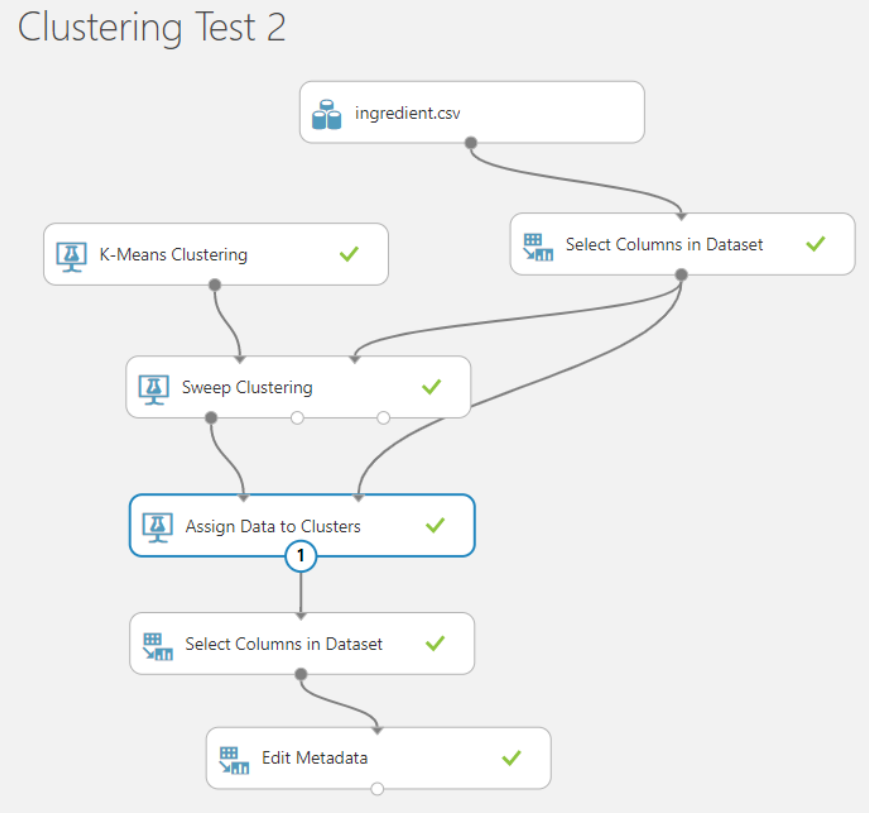
Mean	1.9343
Median	1.0018
Min	0.1709
Max	8.768
Standard Deviation	1.8201
Unique Values	213
Missing Values	0
Feature Type	Numeric Score

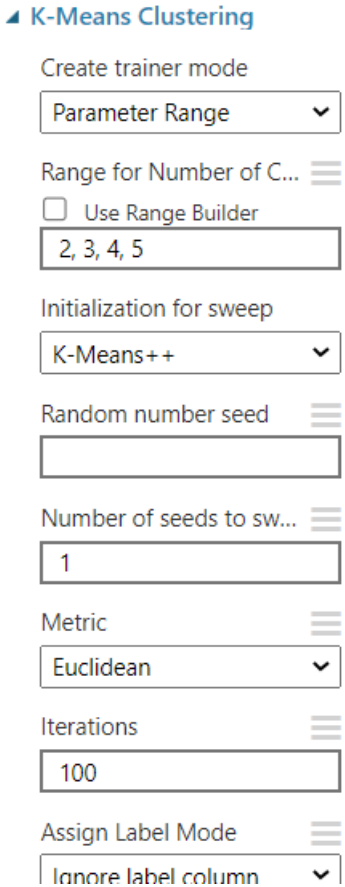
Findings :

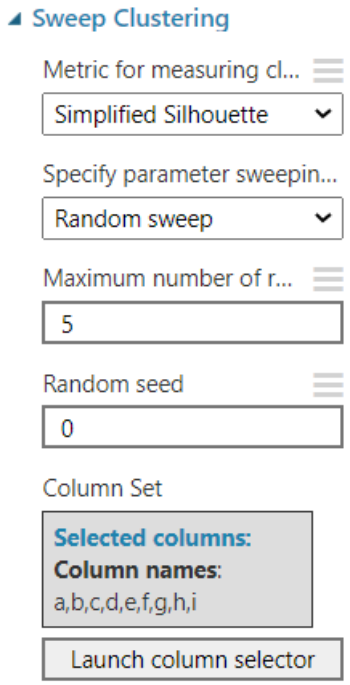
The mean distance from centroid is lowest in Cluster 1 and highest in Cluster 0.

Link : https://gallery.cortanaintelligence.com/Experiment/Clustering-Test-2

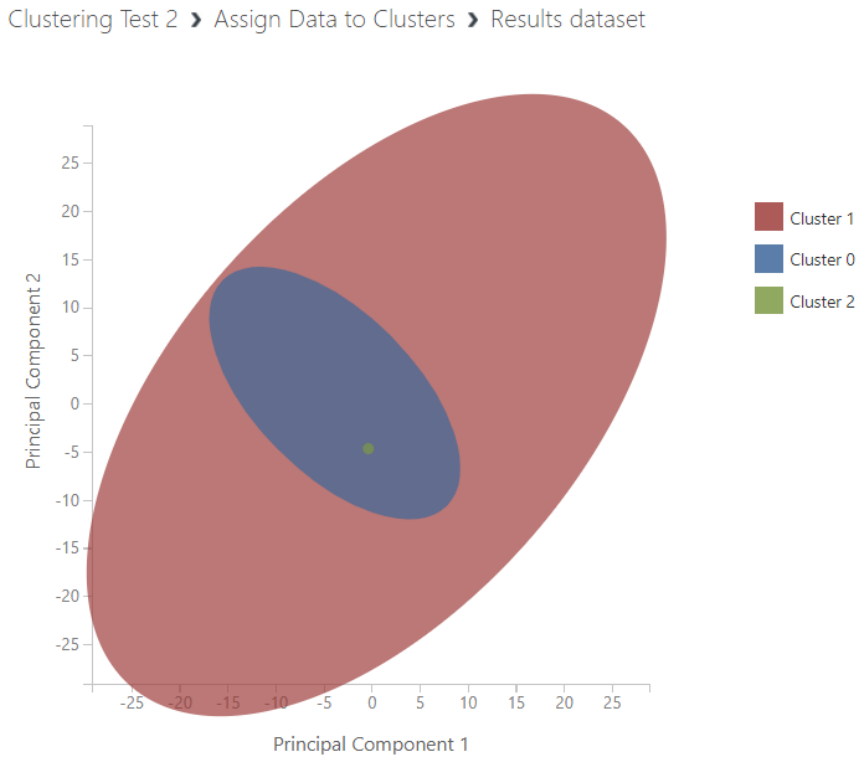
2. Using Sweep Clustering module to find the optimum number of clusters in the dataset.







Output :



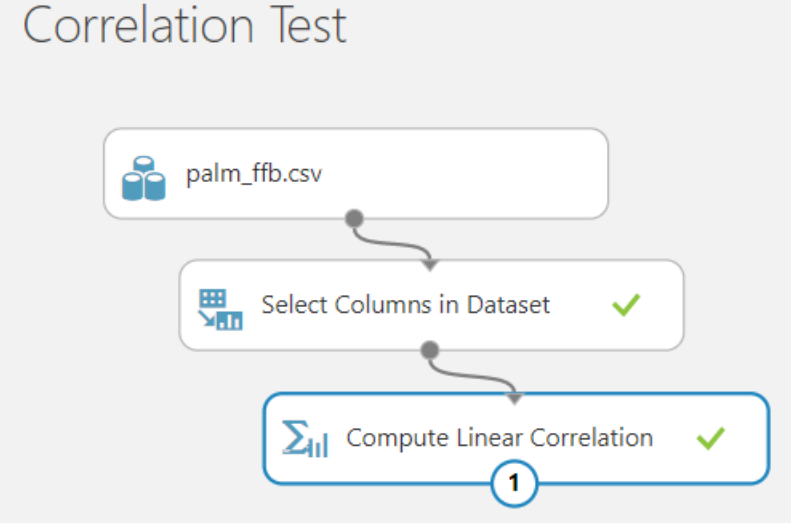
Findings :

The optimal number of clustering is 3.

Question 2 : Analyzing on how external factors effects the fresh fruit bunch (FFB) yield.

Link : https://gallery.cortanaintelligence.com/Experiment/Correlation-Test

1. Using Compute Linear Correlation module to see if there is any correlation between the external factors.



Output :

SoilMoisture	Average_Temp	Min_Temp	Max_Temp	Precipitation	Working_days	HA_Harvested	FFB_Yield
1	-0.649878	0.015839	-0.499936	0.552001	-0.057015	-0.326539	-0.003183
-0.649878	1	0.180396	0.761083	-0.369386	0.076321	0.446515	-0.005494
0.015839	0.180396	1	-0.124754	0.345944	0.068414	0.024396	0.10383
-0.499936	0.761083	-0.124754	1	-0.461117	-0.039112	0.314827	-0.071201
0.552001	-0.369386	0.345944	-0.461117	1	0.127897	-0.265866	0.289604
-0.057015	0.076321	0.068414	-0.039112	0.127897	1	0.048876	0.116364
-0.326539	0.446515	0.024396	0.314827	-0.265866	0.048876	1	-0.350222
-0.003183	-0.005494	0.10383	-0.071201	0.289604	0.116364	-0.350222	1

Findings :
SoilMoisture and Average_Temp are moderately correlated. (negative)
SoilMoisture and Precipitation are moderately correlated. (positive)
Average_Temp and HA_Harvested are weakly correlated. (positive)
HA_Harvested and FFB_Yield are weakly correlated. (negative)

From the correlation analysis, the external factors that would affect FFB_Yield the most would be :
HA_Harvested, Average_Temp, Percipitation and SoilMoisture.