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, 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.

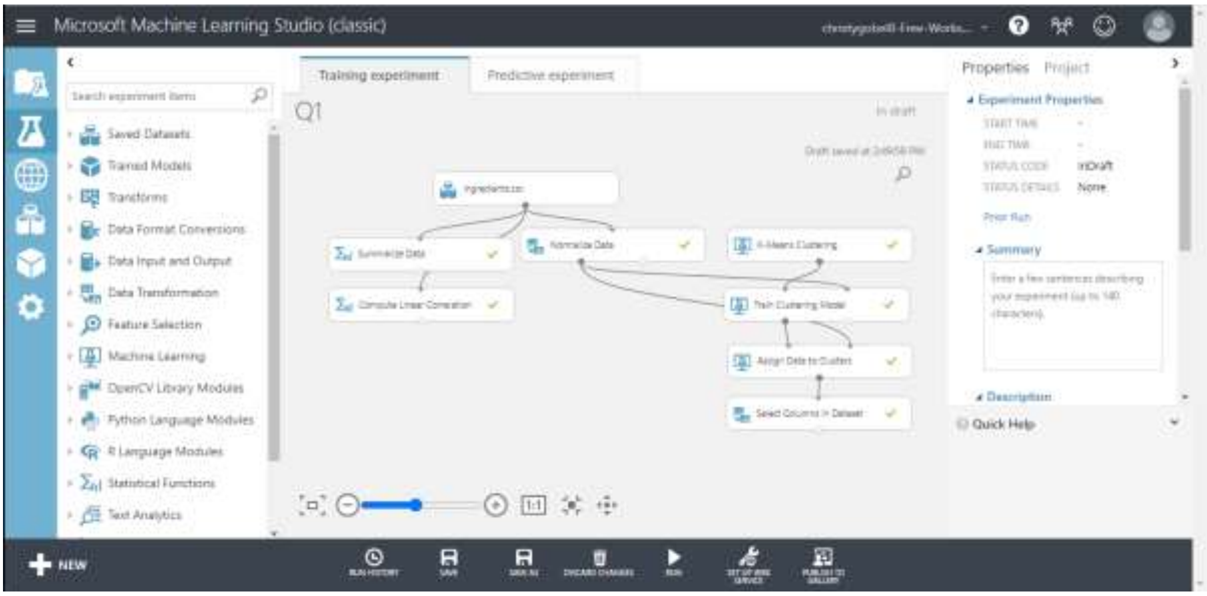
   Please assist the consultant in the area of statistical analysis by doing this;

   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.

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

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

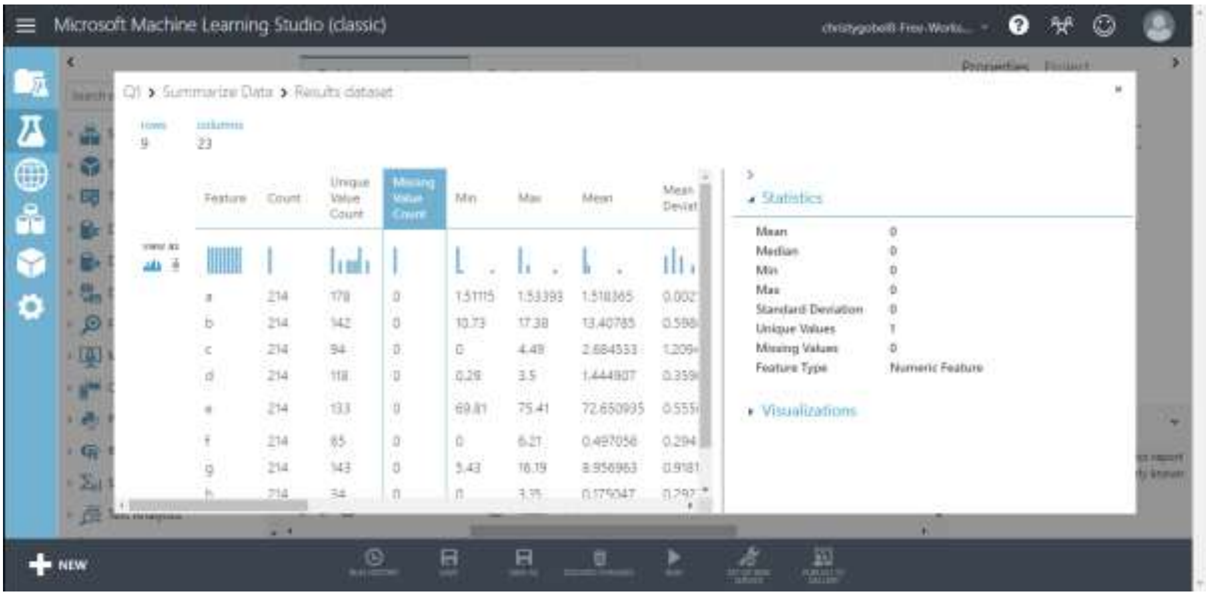
   (refer attachment : ingredients.csv)



STATUS CODE InDraft

STATUS DETAILS None

DESCRIPTIVE ANALYSIS AND CHECK MISSING VALUES


Microsoft Machine Learning Studio (classic)
christygobell-Free-Works...
Properties
Project
NEW

Q1 > Summarize Data > Results dataset

rows 9 columns 23

Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviat
a	214	178	0	1.51115	1.53393	1.518365	0.002
b	214	142	0	10.73	17.38	13.40785	0.598
c	214	94	0	0	4.49	2.684533	1.209
d	214	118	0	0.29	3.5	1.444907	0.359
e	214	133	0	69.81	75.41	72.650935	0.555
f	214	65	0	0	6.21	0.497056	0.294
g	214	143	0	5.43	16.19	8.956963	0.9181
h	214	34	0	0	3.15	0.175047	0.292

Statistics

Mean	0
Median	0
Min	0
Max	0
Standard Deviation	0
Unique Values	1
Missing Values	0
Feature Type	Numeric Feature

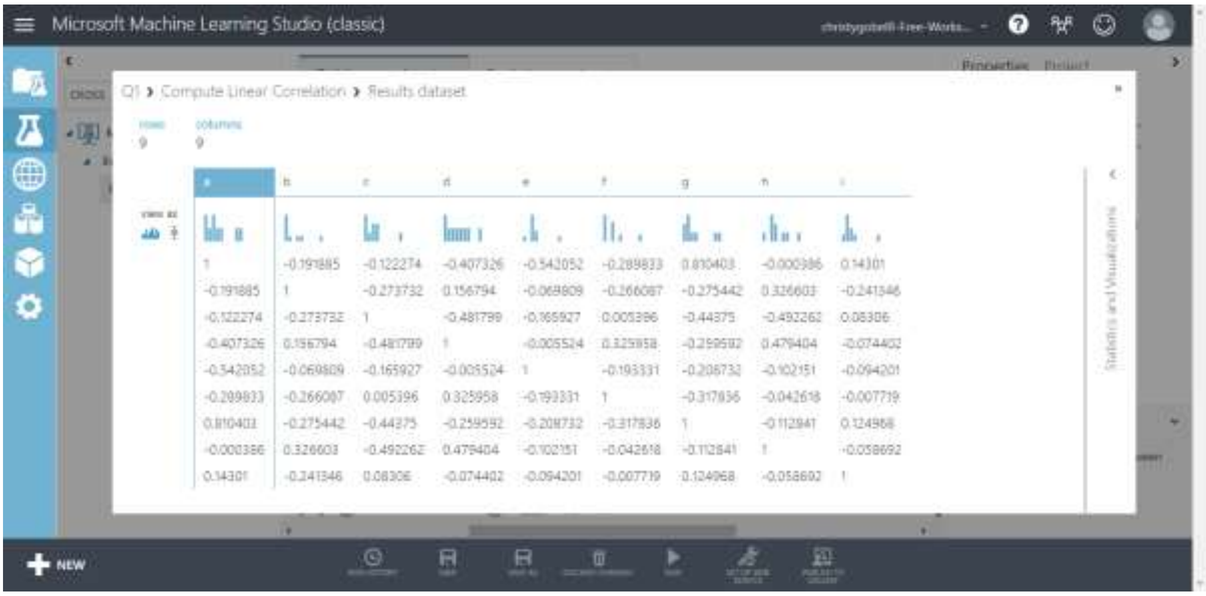
Visualizations

	a	b	c	d	e	f	g	h	i
Count	214	214	214	214	214	214	214	214	214
Missing Value Count	0	0	0	0	0	0	0	0	0
Min	1.51115	10.73	0	0.29	69.81	0	5.43	0	0
Max	1.53393	17.38	4.49	3.5	75.41	6.21	16.19	3.15	0.51
Mean	1.518365	13.40785	2.684533	1.444907	72.65094	0.497056	8.956963	0.175047	0.057009
Mean Deviation	0.002121	0.598898	1.209406	0.359052	0.555696	0.294363	0.918127	0.29237	0.07748
1st Quartile	1.516523	12.9075	2.115	1.19	72.28	0.1225	8.24	0	0
Median	1.51768	13.3	3.48	1.36	72.79	0.555	8.6	0	0
3rd Quartile	1.519157	13.825	3.6	1.63	73.0875	0.61	9.1725	0	0.1
Range	0.02278	6.65	4.49	3.21	5.6	6.21	10.76	3.15	0.51
Sample Variance	0.000009	0.666841	2.08054	0.24927	0.599921	0.425354	2.025366	0.247227	0.009494
Sample Standard Deviation	0.003037	0.816604	1.442408	0.49927	0.774546	0.652192	1.423153	0.497219	0.097439
Sample Skewness	1.625431	0.454181	-1.15256	0.90729	-0.73045	6.551648	2.047054	3.416425	1.754327
Sample Kurtosis	4.931737	3.052232	-0.41032	2.060569	2.967903	54.689699	6.681978	12.54108	2.662016

Summary :

- No missing values
- The values are close to the mean (low variance)

CORRELATION



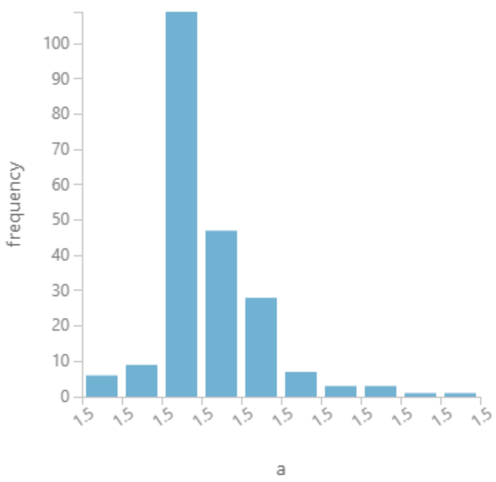
Q1 › Compute Linear Correlation › Results dataset

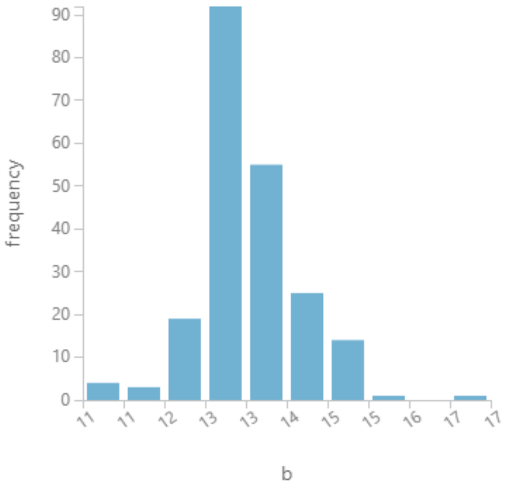
rows 9
columns 9

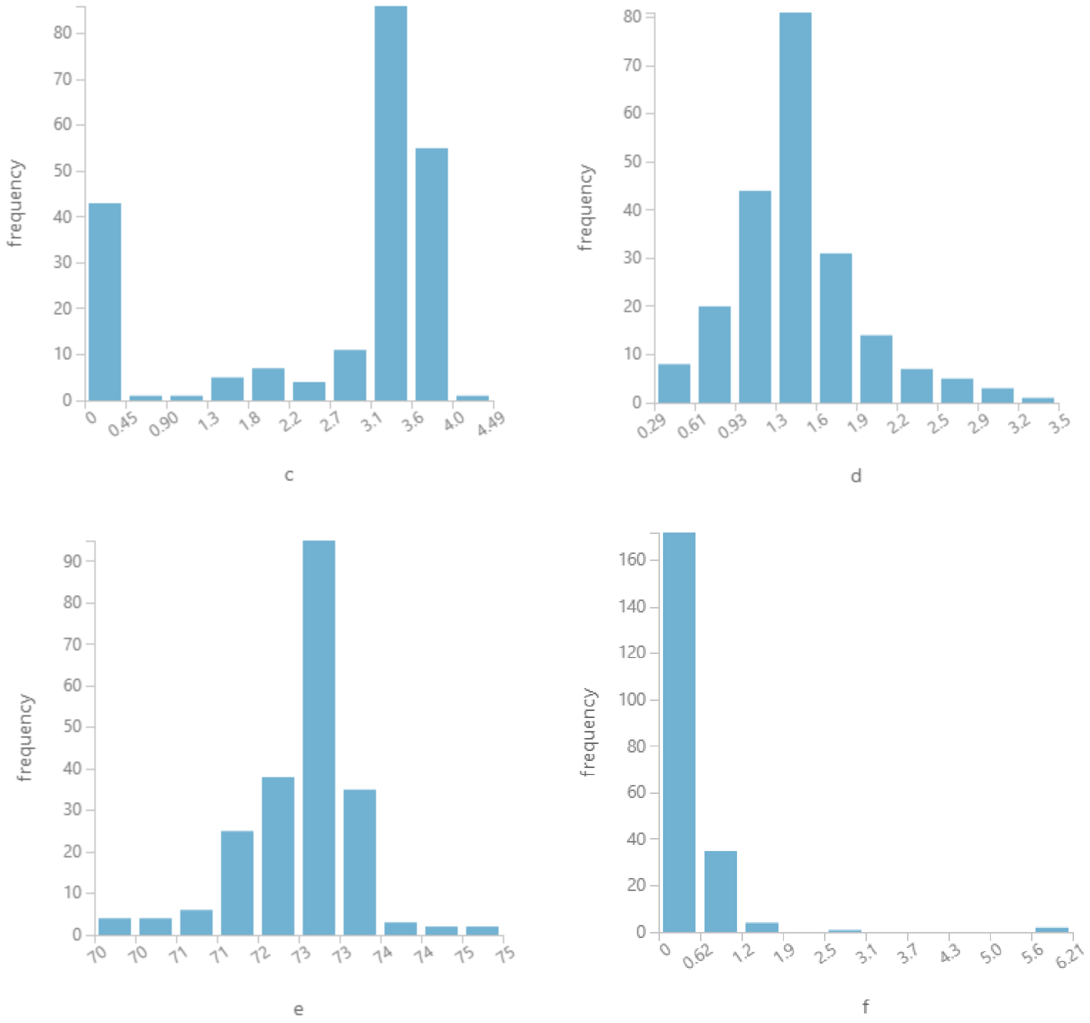
a	b	c	d	e	f	g	h	i
1	-0.191885	-0.122274	-0.407326	-0.542052	-0.289833	0.810403	-0.000386	0.14301
-0.191885	1	-0.273732	0.156794	-0.069809	-0.266087	-0.275442	0.326603	-0.241346
-0.122274	-0.273732	1	-0.481799	-0.165927	0.005396	-0.44375	-0.492262	0.08306
-0.407326	0.156794	-0.481799	1	-0.005524	0.325958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

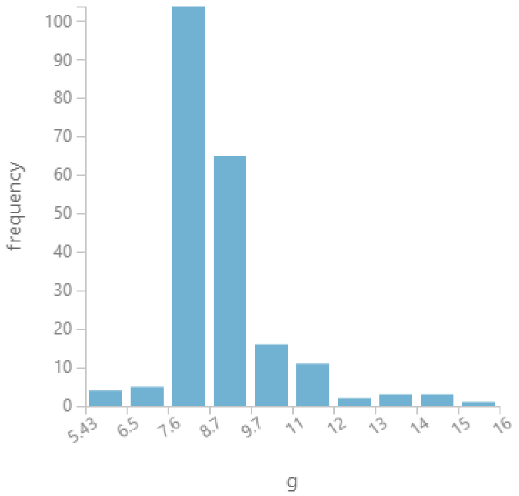
A correlation coefficient is a number between -1 and 1 that tells you the strength and direction of a relationship between variables. A correlation of 0 (uncorrelated variables) indicates no linear (straight line) relationship exists between the variables. A positive correlation close to +1 indicates a strong positive linear relationship. A correlation of 1 indicates a perfect linear relationship. A negative correlation close to −1 indicates a strong negative linear relationship. A correlation of −1 indicates a perfect inverse linear relationship.

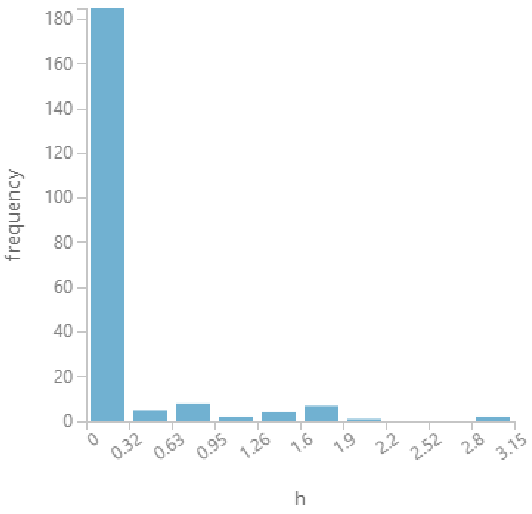
ANALYZE THE DATA DISTRIBUTION

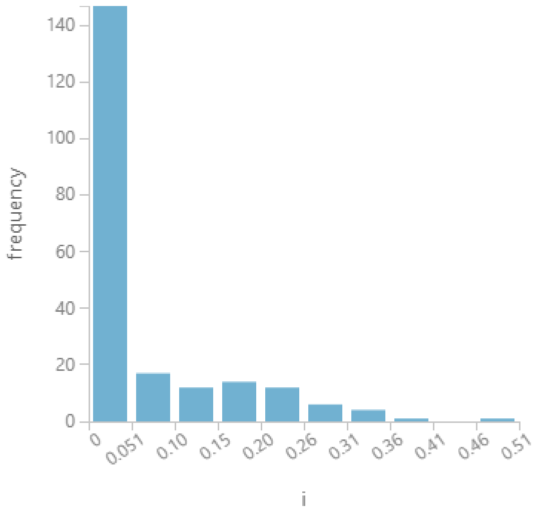





frequency
0
10
20
30
40
50
60
70
80
0
0.45
0.90
1.3
1.8
2.2
2.7
3.1
3.6
4.0
4.49
c
frequency
0
10
20
30
40
50
60
70
80
0.29
0.61
0.93
1.3
1.6
1.9
2.2
2.5
2.9
3.2
3.5
d
frequency
0
10
20
30
40
50
60
70
80
90
70
70
71
71
72
73
73
74
74
75
75
e
frequency
0
20
40
60
80
100
120
140
160
0
0.62
1.2
1.9
2.5
3.1
3.7
4.3
5.0
5.6
6.21
f





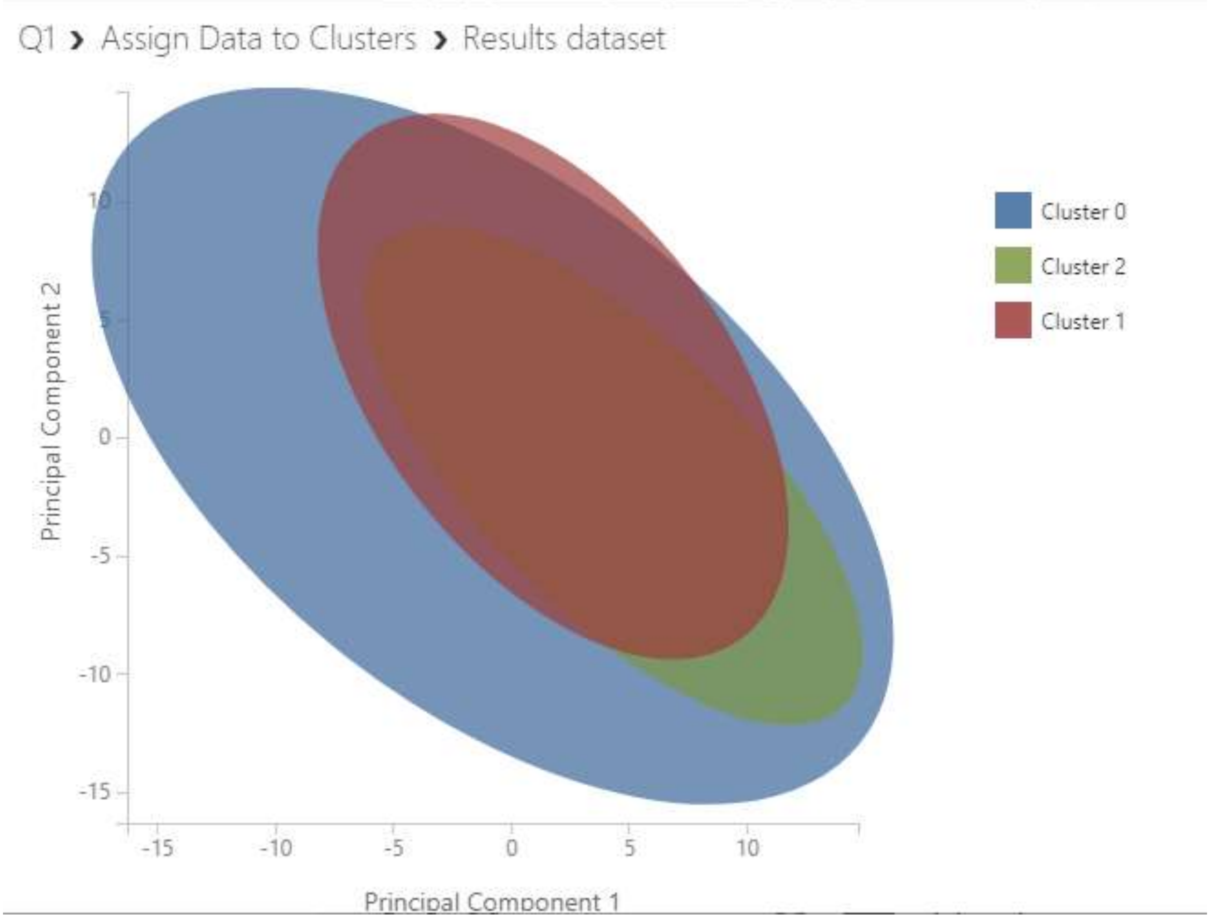


Normal distribution is a distribution that has most of the data in the center with decreasing amounts evenly distributed to the left and the right.

Summary:

Variable a, b, c, d, e, g are normally distributed.

K-MEANS CLUSTERING (UNSUPERVISED LEARNING)


Q1 › Assign Data to Clusters › Results dataset
Principal Component 2
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10
Principal Component 1
🟦 Cluster 0
🟩 Cluster 2
🟥 Cluster 1