APRIL

CANDIDATE

PRESCREEN DS AI

HERU LEONARDO

2021

INTERVIEW PRESENTATION

TABLE OF CONTENTS

Our content today is divided into four parts. Each part will be described with examples.

Heru Leonardo

EXPERIENCE

03/2021-Now	AI Developer - Terra AI X AI4IMPACT Building Learning Platform to school in Indonesia

EDUCATION

06/2021-Now	Awardee FGA Scholarship Microsoft Fundamentals: Azure, Data, Artificial Intelligence, Power Platform- Microsoft
05/2021-08/2021	Dicoding Academy - Machine Learning Developer
08/2016-03/2021	Mechanical Engineer - Universitas Sumatera Utara

CERTIFICATION

IBM Data Science Professional Certificate

PUBLICATIONS

SIMULATION OF COLD-ENERGY STORAGE DISTRIBUTION TEMPERATURE USING PHASE-CHANGE MATERIAL

IJMME-IJENS Vol:21 No:01 | 2021

QUESTION 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.

HYPOTHESIS TESTING

Pre-production stage plays an important role in concept and project direction

PURPOSE

to look for correlation of each data

ASSUMPTIONS

if needed , we can make some assumptions to the data

Setting out future objectives and strategies for achieving them

How to Analyze

Analyze the correlation in each data with some statistical methods

Intrepret

Making summary of analysis

DATA TRANSFORM

Process

DATA PREPARATION

1

Data Sources
Missing Values
Grouping

DATA NORMALIZATION

2

Normalize Data

```
Shape of dataset before cleaning:  1926
Shape of dataset after cleaning:  1926
```

	a	b	c	d	e	f	g	h	i
0	1.51735	13.02	3.54	1.69	72.73	0.54	8.44	0.00	0.07
1	1.53125	10.73	0.00	2.10	69.81	0.58	13.30	3.15	0.28
2	1.52300	13.31	3.58	0.82	71.99	0.12	10.17	0.00	0.03
3	1.51768	12.56	3.52	1.43	73.15	0.57	8.54	0.00	0.00
4	1.51813	13.43	3.98	1.18	72.49	0.58	8.15	0.00	0.00

lets simply clear the dataset by dropping the rows that have null value

DATA CORRELATION

	a	b	c	d	e	f	g	h	i
a	1.000000	-0.191885	-0.122274	-0.407326	-0.542052	-0.289833	0.810403	-0.000386	0.143010
b	-0.191885	1.000000	-0.273732	0.156794	-0.069809	-0.266087	-0.275442	0.326603	-0.241346
c	-0.122274	-0.273732	1.000000	-0.481799	-0.165927	0.005396	-0.443750	-0.492262	0.083060
d	-0.407326	0.156794	-0.481799	1.000000	-0.005524	0.325958	-0.259592	0.479404	-0.074402
e	-0.542052	-0.069809	-0.165927	-0.005524	1.000000	-0.193331	-0.208732	-0.102151	-0.094201
f	-0.289833	-0.266087	0.005396	0.325958	-0.193331	1.000000	-0.317836	-0.042618	-0.007719
g	0.810403	-0.275442	-0.443750	-0.259592	-0.208732	-0.317836	1.000000	-0.112841	0.124968
h	-0.000386	0.326603	-0.492262	0.479404	-0.102151	-0.042618	-0.112841	1.000000	-0.058692
i	0.143010	-0.241346	0.083060	-0.074402	-0.094201	-0.007719	0.124968	-0.058692	1.000000

From the table, we can see that variable a has the best correlation with variable g

DATA SUMMARY

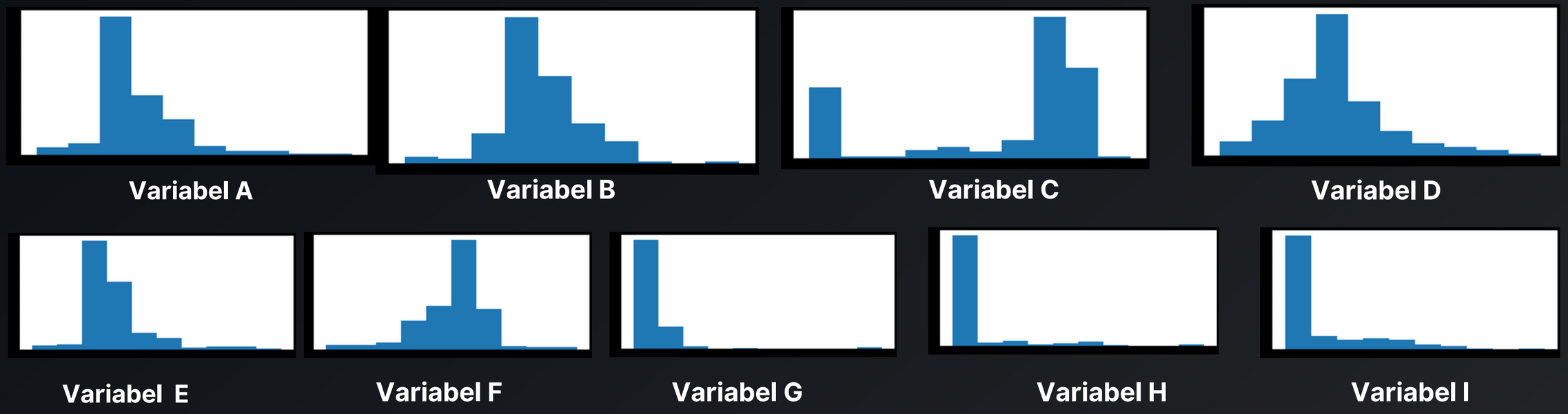
	a	b	c	d	e	f	g	h	i
count	214.000000	214.000000	214.000000	214.000000	214.000000	214.000000	214.000000	214.000000	214.000000
mean	1.518365	13.407850	2.684533	1.444907	72.650935	0.497056	8.956963	0.175047	0.057009
std	0.003037	0.816604	1.442408	0.499270	0.774546	0.652192	1.423153	0.497219	0.097439
min	1.511150	10.730000	0.000000	0.290000	69.810000	0.000000	5.430000	0.000000	0.000000
25%	1.516523	12.907500	2.115000	1.190000	72.280000	0.122500	8.240000	0.000000	0.000000
50%	1.517680	13.300000	3.480000	1.360000	72.790000	0.555000	8.600000	0.000000	0.000000
75%	1.519157	13.825000	3.600000	1.630000	73.087500	0.610000	9.172500	0.000000	0.100000
max	1.533930	17.380000	4.490000	3.500000	75.410000	6.210000	16.190000	3.150000	0.510000

NORMALITY TEST USING SHAPIRO-WILK TEST

tests If data is normally distributed

Assumption : Observations are identically distributed

Checking histogram indentify by one variable



From the overall histogram that being showed. It looks like it doesnt have a normal distribution.

Let's Clarify things with statistical method which is Shapiro-Wilk Test. Code in github

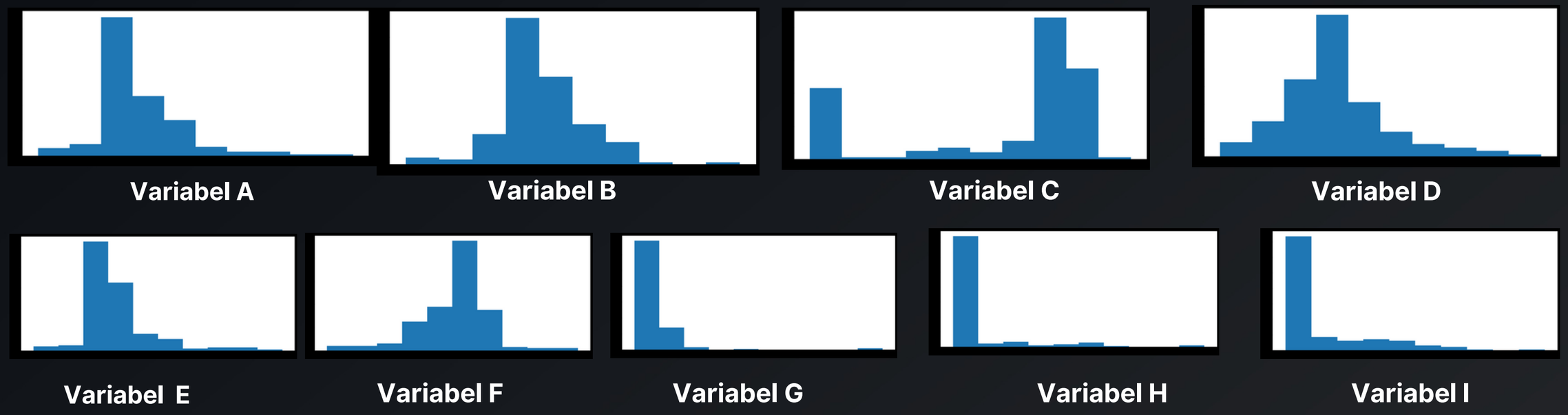
Okay From the test above, seems like our guess is true if it is not a normal distribution.

NORMALITY TEST USING K^2 NORMALITY TEST

tests If data is normally distributed

Assumption : Observations are identically distributed

Checking histogram indentify by one variable



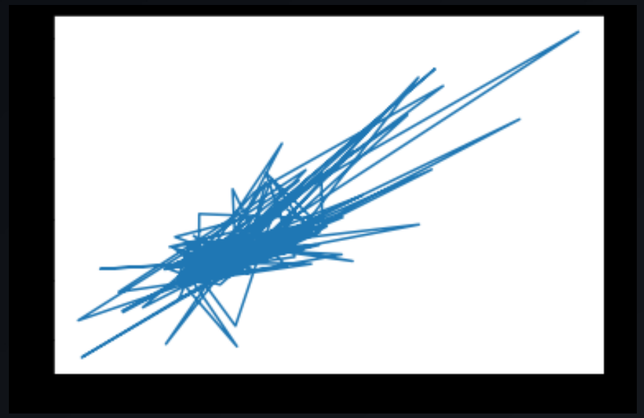
From the overall histogram that being showed. It looks like it doesnt have a normal distribution.

Let's Clarify things with statistical method which is K^2 Normality Test Test. Code in github

Okay From the test above, seems like our guess is true if it is not a normal distribution.

CORRELATION TEST - PEARSON AND SPEARMAN'S RANK CORRELATION

Line plot indentify correlation by two variable a and g



Spearman Rank Correlation

stat=0.704, p=0.000000
dependent samples

Pearson Correlation

stat=0.810, p=0.000000
dependent samples

The statistical test reports a strong positive correlation with a value of 0.704. The p-value is close to zero, which means that the likelihood of observing the data given that the samples are uncorrelated is very unlikely (e.g. 95% confidence) and that we can reject the null hypothesis that the samples are uncorrelated.

The p-value is close to zero (and printed as zero), as with the Spearman's test, meaning that we can confidently reject the null hypothesis that the samples are uncorrelated.

PARAMETIC TEST T-TEST AND ANOVA

correlation by two variable a and g

T-Test

ANOVA

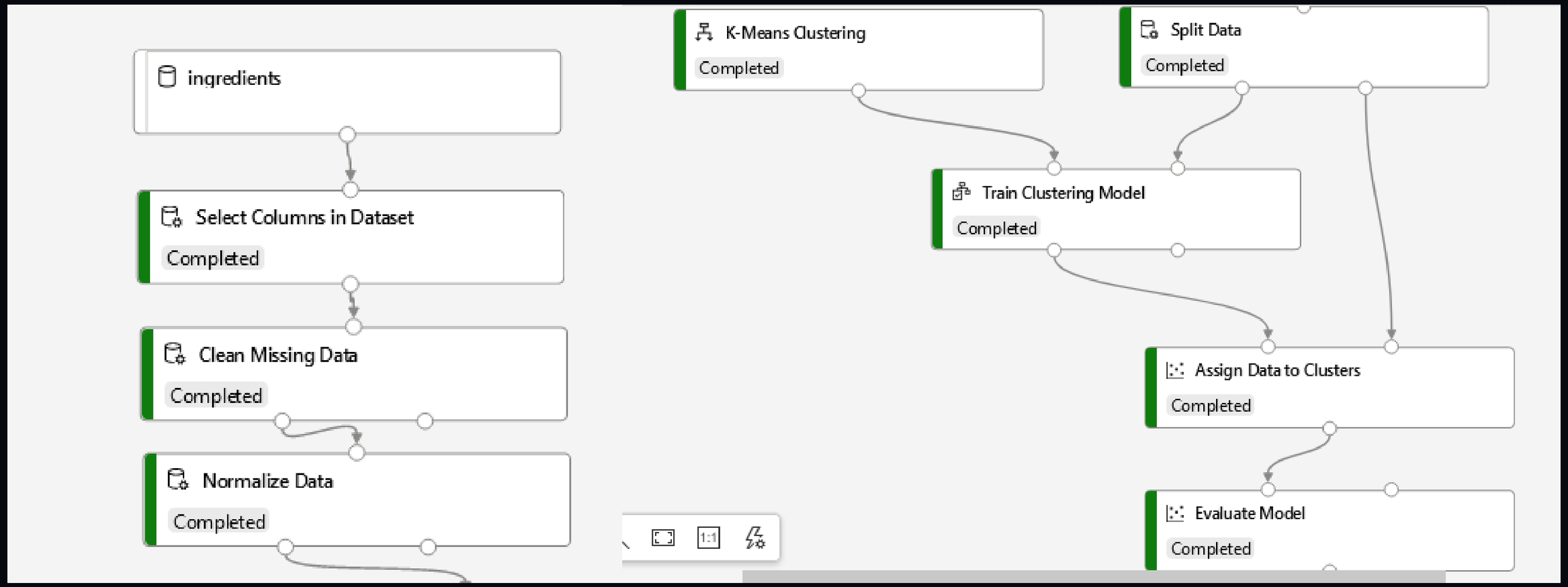
Tests whether the means of two or more independent samples are significantly different.

Statistics=-76.462, p=0.000
Different distributions (reject H0)

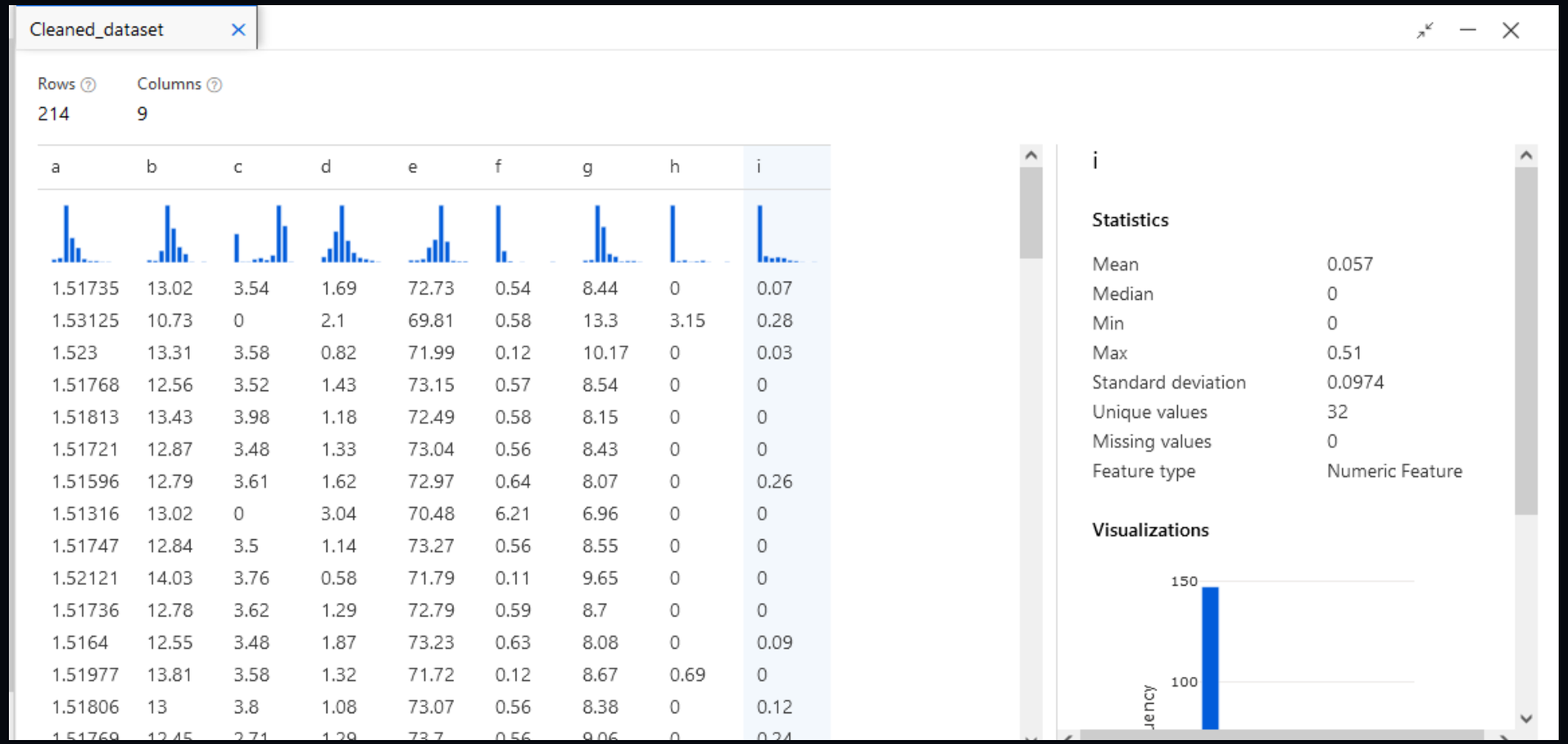
stat=-76.462, p=0.000
Different distributions of scores

The interpretation of the statistic finds that the sample means are different, with a significance of at least 5%.

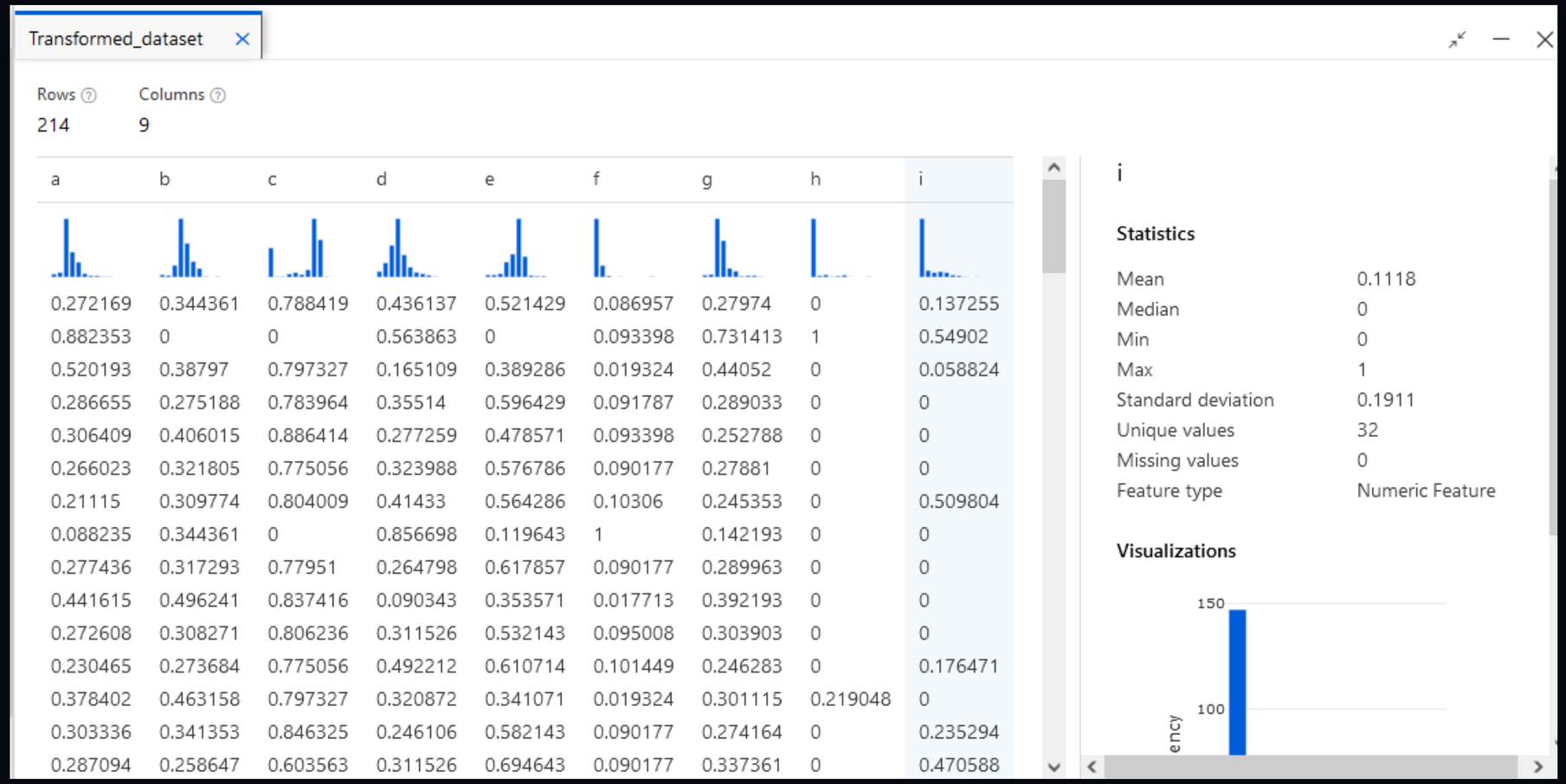
CLUSTERING USING AZURE ML



DATA CLEANING



NORMALIZE DATA



SPLIT DATA

Results_dataset1

Rows 150 Columns 9

a	b	c	d	e	f	g	h	i
0.272169	0.344361	0.788419	0.436137	0.521429	0.086957	0.27974	0	0.137255
0.882353	0	0	0.563863	0	0.093398	0.731413	1	0.54902
0.520193	0.38797	0.797327	0.165109	0.389286	0.019324	0.44052	0	0.058824
0.286655	0.275188	0.783964	0.35514	0.596429	0.091787	0.289033	0	0
0.306409	0.406015	0.886414	0.277259	0.478571	0.093398	0.252788	0	0
0.266023	0.321805	0.775056	0.323988	0.576786	0.090177	0.27881	0	0
0.21115	0.309774	0.804009	0.41433	0.564286	0.10306	0.245353	0	0.509804
0.088235	0.344361	0	0.856698	0.119643	1	0.142193	0	0
0.277436	0.317293	0.77951	0.264798	0.617857	0.090177	0.289963	0	0
0.441615	0.496241	0.837416	0.090343	0.353571	0.017713	0.392193	0	0
0.272608	0.308271	0.806236	0.311526	0.532143	0.095008	0.303903	0	0
0.230465	0.273684	0.775056	0.492212	0.610714	0.101449	0.246283	0	0.176471
0.378402	0.463158	0.797327	0.320872	0.341071	0.019324	0.301115	0.219048	0
0.303336	0.341353	0.846325	0.246106	0.582143	0.090177	0.274164	0	0.235294
0.287094	0.258647	0.603563	0.311526	0.694643	0.090177	0.337361	0	0.470588

Train Data

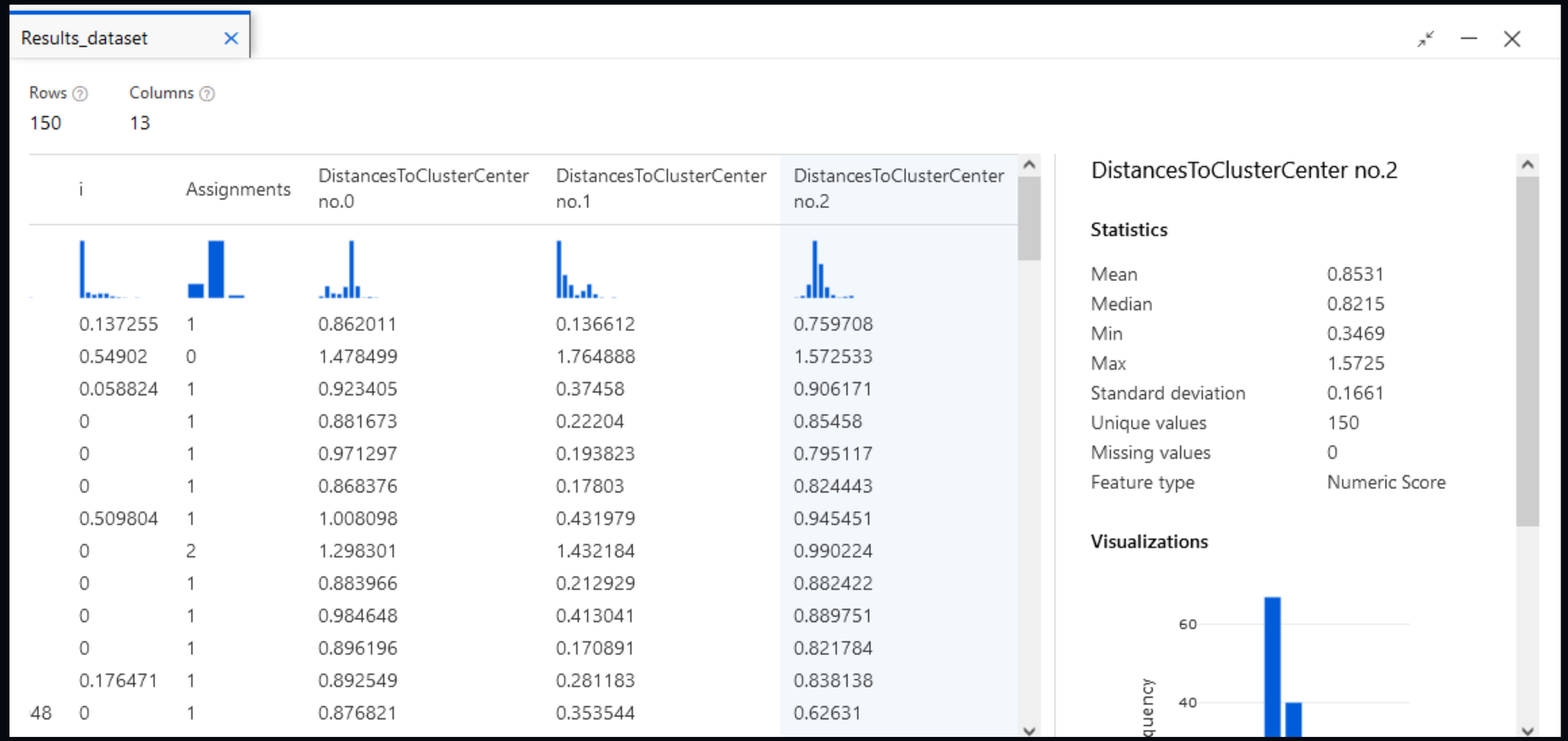
Results_dataset2

Rows 64 Columns 9

a	b	c	d	e	f	g	h	i
0.70676	0.461654	0.701559	0.115265	0.135714	0.012882	0.577138	0	0
0.320018	0.378947	0.837416	0.320872	0.4625	0.093398	0.277881	0	0
0.304214	0.406015	0.639198	0.280374	0.541071	0.088567	0.334572	0	0
0.526778	0.407519	0.743875	0.292835	0.458929	0.096618	0.315985	0	0
0.235733	0.425564	0.7951	0.367601	0.471429	0.10306	0.23513	0	0
0.226514	0.392481	0.7951	0.398754	0.546429	0.098229	0.228625	0	0
0.550483	0.407519	0	0.401869	0.430357	0.05153	0.6329	0	0
0.300702	0.44812	0.875278	0.389408	0.357143	0.086957	0.258364	0	0.294118
0.289728	0.318797	0.775056	0.292835	0.564286	0.098229	0.290892	0.028571	0.431373
0.213784	0.619549	0	0.65109	0.619643	0	0.30948	0.203175	0.176471
0.341089	0.410526	0.853007	0.302181	0.489286	0.091787	0.258364	0	0.27451
0.251975	0.291729	0.641425	0.442368	0.607143	0.117552	0.289033	0	0
0.043898	0.33985	0.772829	0.258567	0.566071	0.099839	0.271375	0	0.607843
0.279192	0.312782	0.7951	0.330218	0.573214	0.099839	0.29368	0	0
0.249342	0.57594	0	0.52648	0.621429	0	0.287175	0.498413	0.137255

Test Data

RESULT


Results_dataset
Rows
150
Columns
13
i
Assignments
DistancesToClusterCenter no.0
DistancesToClusterCenter no.1
DistancesToClusterCenter no.2
0.137255 1 0.862011 0.136612 0.759708
0.54902 0 1.478499 1.764888 1.572533
0.058824 1 0.923405 0.37458 0.906171
0 1 0.881673 0.22204 0.85458
0 1 0.971297 0.193823 0.795117
0 1 0.868376 0.17803 0.824443
0.509804 1 1.008098 0.431979 0.945451
0 2 1.298301 1.432184 0.990224
0 1 0.883966 0.212929 0.882422
0 1 0.984648 0.413041 0.889751
0 1 0.896196 0.170891 0.821784
0.176471 1 0.892549 0.281183 0.838138
48 0 1 0.876821 0.353544 0.62631
DistancesToClusterCenter no.2
Statistics
Mean 0.8531
Median 0.8215
Min 0.3469
Max 1.5725
Standard deviation 0.1661
Unique values 150
Missing values 0
Feature type Numeric Score
Visualizations
60
40
quency

EVALUATION

Evaluation_results

Rows 4 Columns 5

Result Description	Average Distance to Other Center	Average Distance to Cluster Center	Number of Points	Maximal Distance to Cluster Center
Evaluation For Cluster No.0	0.886307	0.584178	16	1.272154
Evaluation For Cluster No.1	0.819077	0.301809	47	0.767541
Evaluation For Cluster No.2	1.278354	0.986631	1	0.986631
Combined Evaluation	0.843061	0.383101	64	1.272154

FEATURES

	a	b	c	d	e	f	g	h	i
50	1.51593	13.09	3.59	1.52	73.10	0.67	7.83	0.0	0.00
30	1.51969	12.64	0.00	1.65	73.75	0.38	11.53	0.0	0.00
147	1.51888	14.99	0.78	1.74	72.50	0.00	9.95	0.0	0.00
163	1.51751	12.81	3.57	1.35	73.02	0.62	8.59	0.0	0.00
24	1.51514	14.01	2.68	3.50	69.89	1.68	5.87	2.2	0.00
132	1.51645	13.44	3.61	1.54	72.39	0.66	8.03	0.0	0.00
96	1.52247	14.86	2.20	2.06	70.26	0.76	9.76	0.0	0.00
40	1.51571	12.72	3.46	1.56	73.20	0.67	8.09	0.0	0.24
31	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.0	0.00
51	1.51829	13.24	3.90	1.41	72.33	0.55	8.31	0.0	0.10

Display a random sample of 10 observations (just the features)

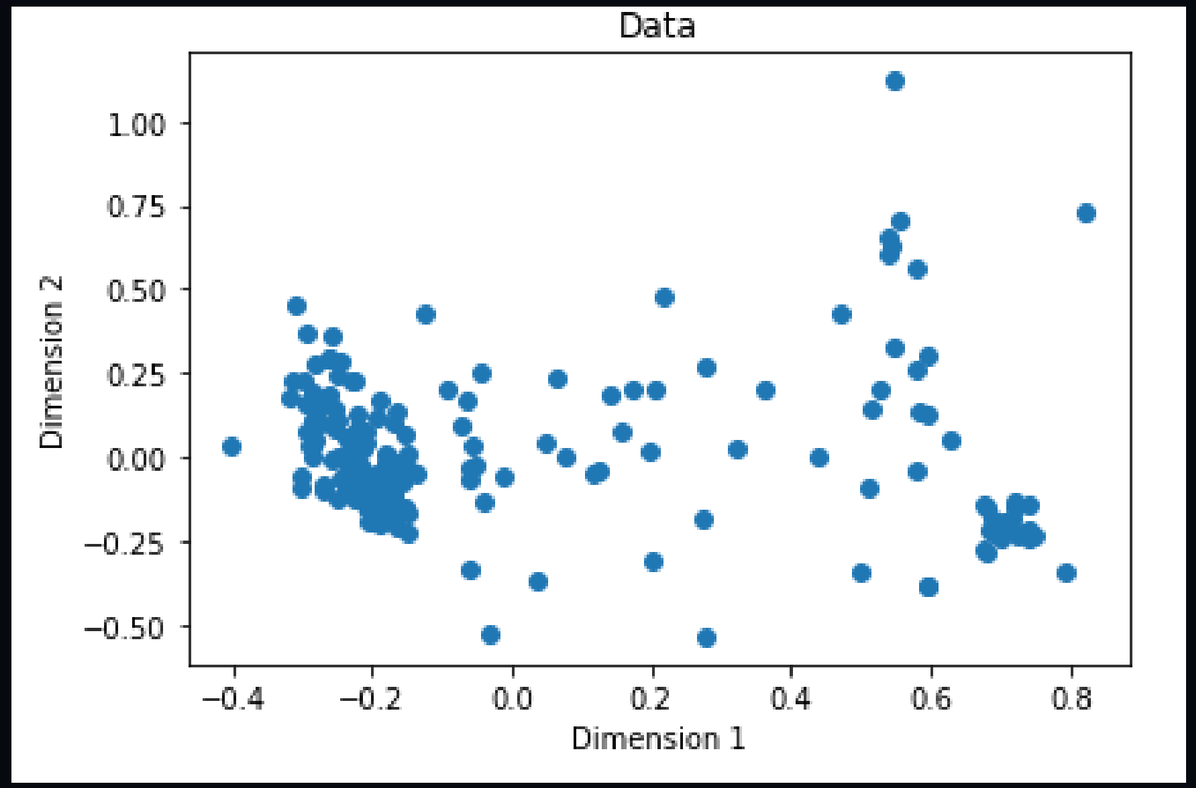
As you can see, the dataset contains nine data points (or features) for each instance (observation) of a seed. So you could interpret these as coordinates that describe each instance's location in nine-dimensional space.

PCA FEATURES

```
array([[-0.18478222, -0.06713815],
       [ 0.82254736,  0.73009536],
       [-0.22560457,  0.22594417],
       [-0.19360521, -0.09384108],
       [-0.3021048 , -0.08754554],
       [-0.19125623, -0.10250388],
       [-0.23834338,  0.06019619],
       [ 0.59743   , -0.37985552],
       [-0.20649451, -0.0789971 ],
       [-0.27601625,  0.1414611 ]])
```

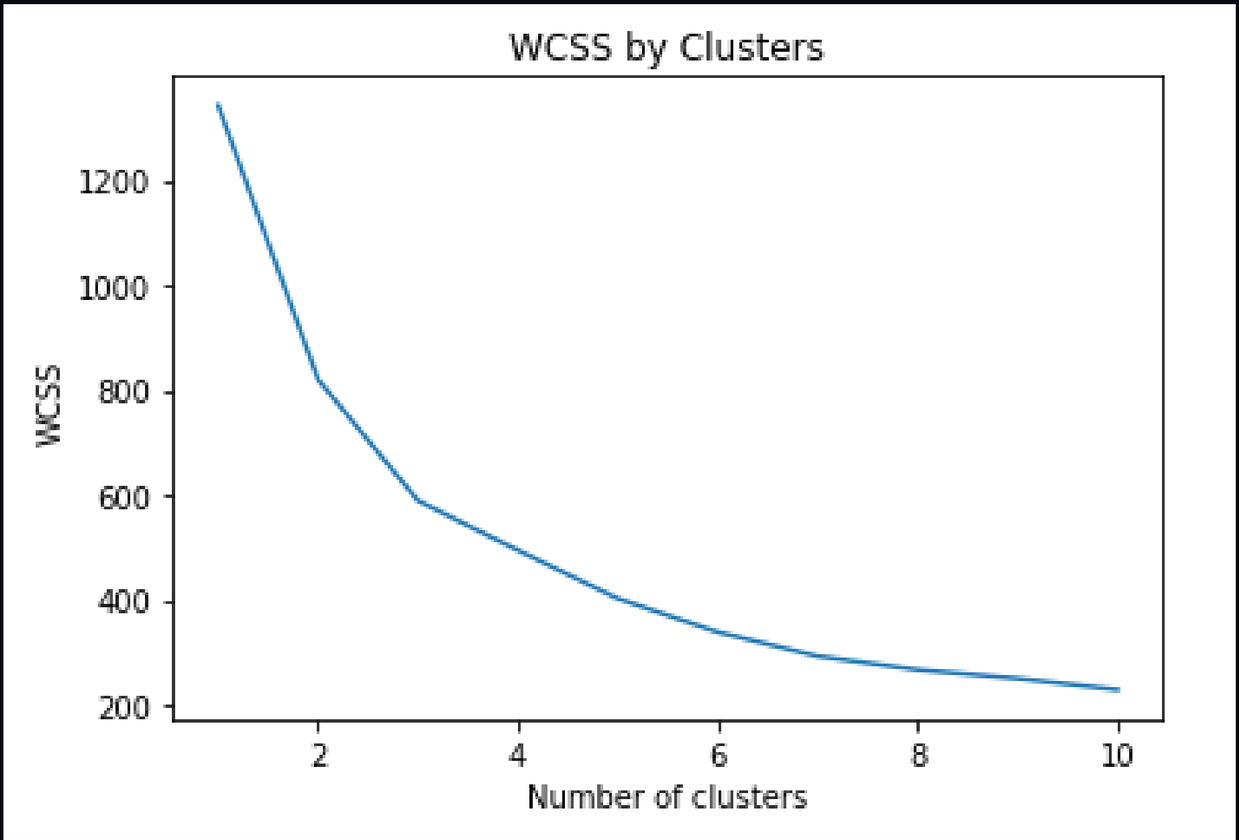
Now, of course nine-dimensional space is difficult to visualise in a three-dimensional world, or on a two-dimensional plot; so we'll take advantage of a mathematical technique called Principal Component Analysis (PCA) to analyze the relationships between the features and summarize each observation as coordinates for two principal components - in other words, we'll translate the nine-dimensional feature values into two-dimensional coordinates.

DATA VISUALIZATION



Hopefully you can see at least two, arguably three, reasonably distinct groups of data points; but here lies one of the fundamental problems with clustering - without known class labels, how do you know how many clusters to separate your data into?

WCSS



One way we can try to find out is to use a data sample to create a series of clustering models with an incrementing number of clusters, and measure how tightly the data points are grouped within each cluster. A metric often used to measure this tightness is the within cluster sum of squares (WCSS), with lower values meaning that the data points are closer. You can then plot the WCSS for each model.

QUESTION 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)

DATA CORRELATION

	SoilMoisture	Average_Temp	Min_Temp	Max_Temp	Precipitation	Working_days	HA_Harvested	FFB_Yield
SoilMoisture	1.000000	-0.649878	0.015839	-0.499936	0.552001	-0.057015	-0.326539	-0.003183
Average_Temp	-0.649878	1.000000	0.180396	0.761083	-0.369386	0.076321	0.446515	-0.005494
Min_Temp	0.015839	0.180396	1.000000	-0.124754	0.345944	0.068414	0.024396	0.103830
Max_Temp	-0.499936	0.761083	-0.124754	1.000000	-0.461117	-0.039112	0.314827	-0.071201
Precipitation	0.552001	-0.369386	0.345944	-0.461117	1.000000	0.127897	-0.265866	0.289604
Working_days	-0.057015	0.076321	0.068414	-0.039112	0.127897	1.000000	0.048876	0.116364
HA_Harvested	-0.326539	0.446515	0.024396	0.314827	-0.265866	0.048876	1.000000	-0.350222
FFB_Yield	-0.003183	-0.005494	0.103830	-0.071201	0.289604	0.116364	-0.350222	1.000000

From the table, we can see that FFB Yield has the best correlation with Precipitation, Min_Temp and Working_days
we can see that Precipitation has the best correlation with SoilMoisture, Min_Temp and Working_days

NORMALITY TEST

tests If data is normally distributed

Assumption : Observations are identically distributed

Checking histogram indentify by one variable

FFB Yield

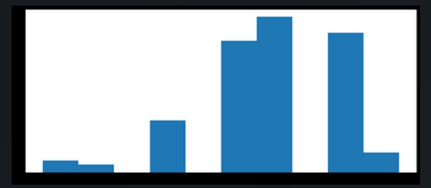
stat=0.98,
p=0.09
Normal
distribution

Precipitation

stat=0.98,
p=0.03
Not a normal
distribution

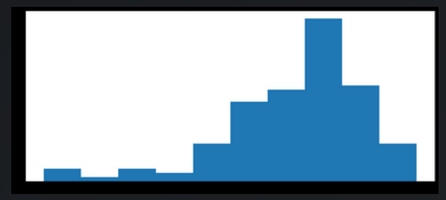
Soil Moisture

stat=0.98,
p=0.08
Normal
distribution



Working Days

stat=0.91,
p=0.0000003
Not a normal
distribution



Min Temp

stat=0.93,
p=0.0000063
Not a normal
distribution

DATA SUMMARY

	SoilMoisture	Average_Temp	Min_Temp	Max_Temp	Precipitation	Working_days	HA_Harvested	FFB_Yield
count	130.000000	130.000000	130.000000	130.000000	130.000000	130.000000	130.000000	130.000000
mean	527.646923	26.849918	21.379231	33.851538	188.980769	24.753846	793404.491565	1.602231
std	57.367844	0.651413	0.688971	1.079638	80.237210	1.239289	34440.893854	0.281751
min	380.700000	25.158065	18.900000	31.100000	2.000000	21.000000	683431.944400	1.080000
25%	488.625000	26.442285	21.000000	33.100000	140.300000	24.000000	768966.949100	1.390000
50%	538.300000	26.930645	21.500000	33.900000	182.150000	25.000000	790036.158050	1.585000
75%	571.025000	27.270726	21.800000	34.600000	226.100000	26.000000	821989.235250	1.807500
max	647.300000	28.580000	22.600000	36.000000	496.100000	27.000000	882254.225400	2.270000

THANK YOU

FOR WATCHING