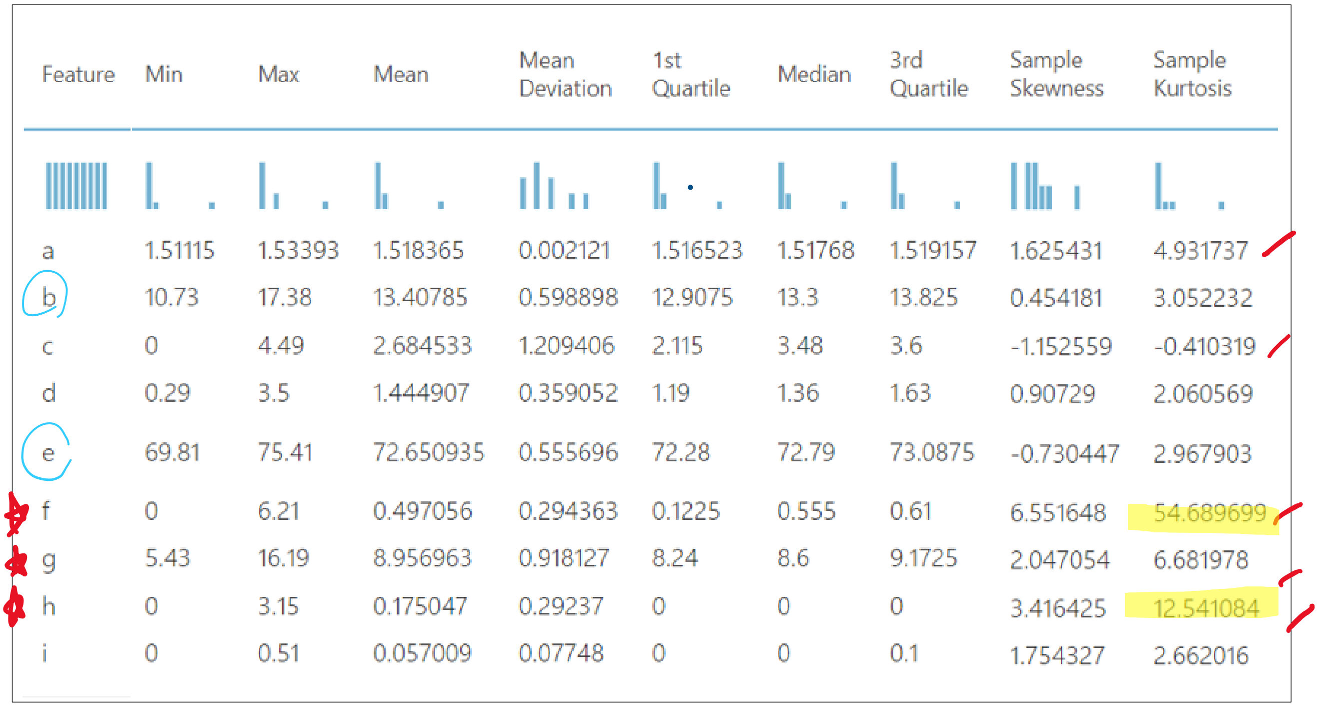
Question 1: Ingredients

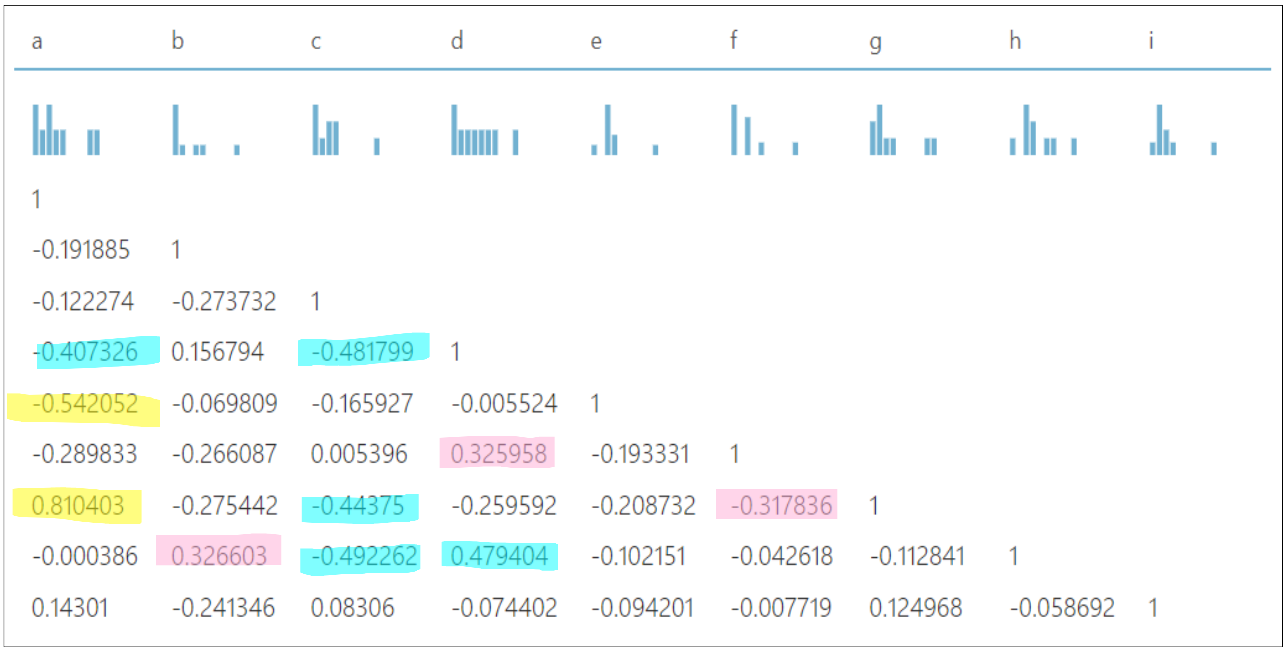
a. A descriptive analysis of the additives (columns named as “a” to “i”), which must include summaries of findings (parametric/non-parametric). Correlation and ANOVA, if applicable, is a must.

Summary statistics:



Feature	Min	Max	Mean	Mean Deviation	1st Quartile	Median	3rd Quartile	Sample Skewness	Sample Kurtosis
a	1.51115	1.53393	1.518365	0.002121	1.516523	1.51768	1.519157	1.625431	4.931737
b	10.73	17.38	13.40785	0.598898	12.9075	13.3	13.825	0.454181	3.052232
c	0	4.49	2.684533	1.209406	2.115	3.48	3.6	-1.152559	-0.410319
d	0.29	3.5	1.444907	0.359052	1.19	1.36	1.63	0.90729	2.060569
e	69.81	75.41	72.650935	0.555696	72.28	72.79	73.0875	-0.730447	2.967903
f	0	6.21	0.497056	0.294363	0.1225	0.555	0.61	6.551648	54.689699
g	5.43	16.19	8.956963	0.918127	8.24	8.6	9.1725	2.047054	6.681978
h	0	3.15	0.175047	0.29237	0	0	0	3.416425	12.541084
i	0	0.51	0.057009	0.07748	0	0	0.1	1.754327	2.662016

Correlation:



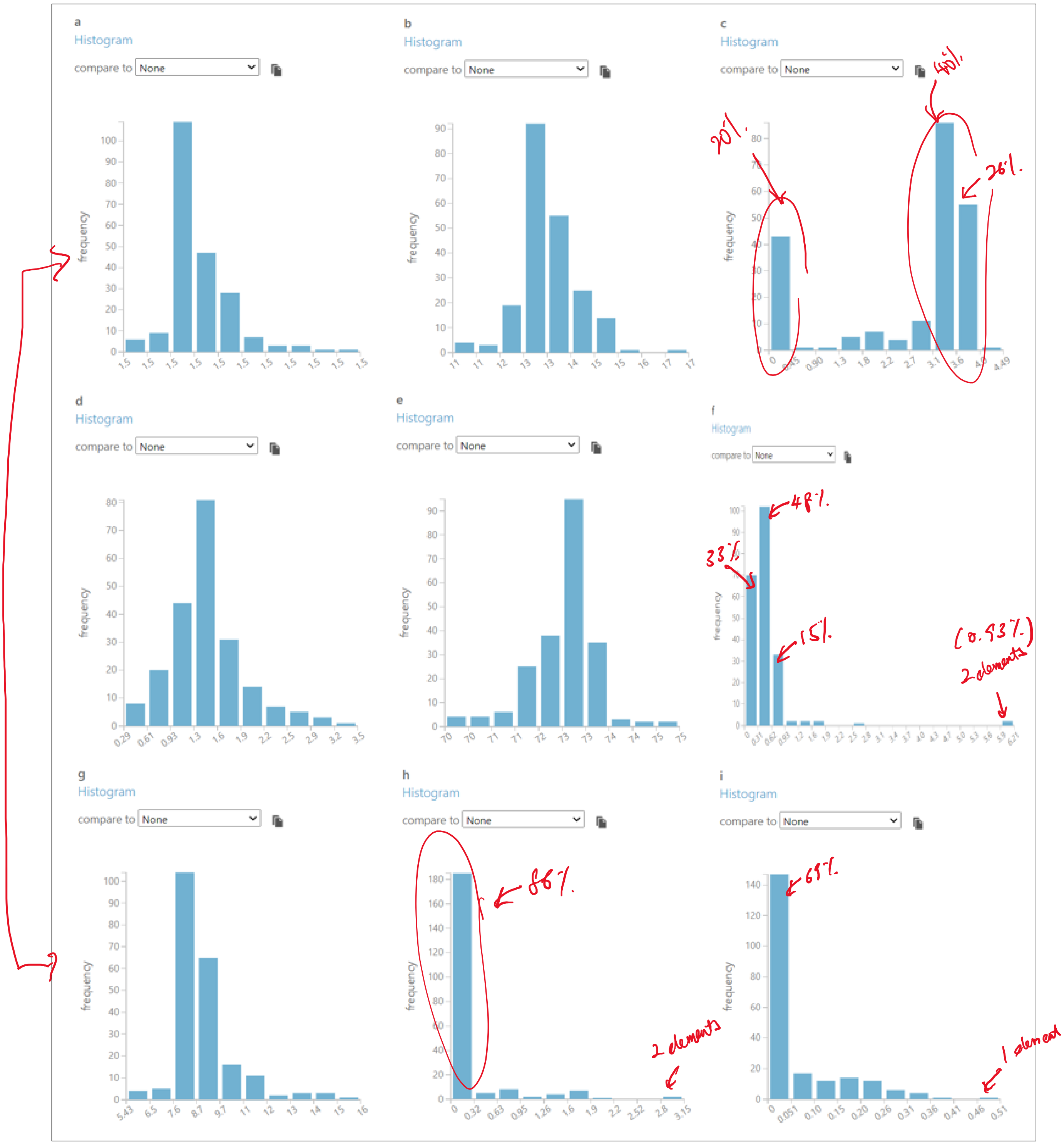
	a	b	c	d	e	f	g	h	i
a	1								
b	-0.191885	1							
c	-0.122274	-0.273732	1						
d	-0.407326	0.156794	-0.481799	1					
e	-0.542052	-0.069809	-0.165927	-0.005524	1				
f	-0.289833	-0.266087	0.005396	0.325958	-0.193331	1			
g	0.810403	-0.275442	-0.44375	-0.259592	-0.208732	-0.317836	1		
h	-0.000386	0.326603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

Findings:

Based on the details in summary statistics (minimum value, maximum value, mean, median, sample kurtosis), it is observed that ingredient e and b are the first and second major ingredients in petrol due to large volumes at the minimum value (69.81 and 10.73 respectively), whereby ingredient e occupies around 50% - 80% while ingredient b occupies around 12% - 13% of the overall full petrol formulation. Besides, the sample kurtoses of both ingredients are 2.97 and 3.05, which is close to 3 and indicate that the distributions are considered normal.

On the other hand, the distinguishing additives that influence the burning pattern of the petrol are ingredients f, h, g, a, c. As it is shown in the summary statistics figure that ingredient f and h may or may not be added to the formula (minimum values of 0), and these ingredients have the most significant sample kurtosis values, the statistics indicate that these ingredients make large differences to the petrol formulations. Whereas for ingredients g and a, both ingredients must be added (quite large volume for ingredient g), but the distinguishing effects are dependent on other ingredients. This statement is supported by the high correlation value between ingredient g and a which shows that both ingredients are interdependent. Besides, the volume of ingredient a is correlated to ingredient e, whereby when volume of e is increased by 1 unit, the volume of ingredient a is reduced by 54%. Being the most skewed variable, the volume of ingredient f to be added to the petrol formula is slightly dependent on the volumes of many other ingredients (correlations to most other ingredients are around 0.2-0.3), however is almost independent to the volume of ingredient h. In a contrary, the volume of ingredient h is highly dependent on the volumes of ingredient b, c, and d. It is also worth to point out that ingredient c is either not added at all, or added at high portion (min = 0, max = 4.49) (see distribution of ingredient c in part b), which is mostly dependent on whether ingredient h is added, hence have kurtosis value of -0.41. Based on the correlation between ingredient h and c, it is suspected that the addition of either one ingredient between h and c would give significantly different burning patterns.

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

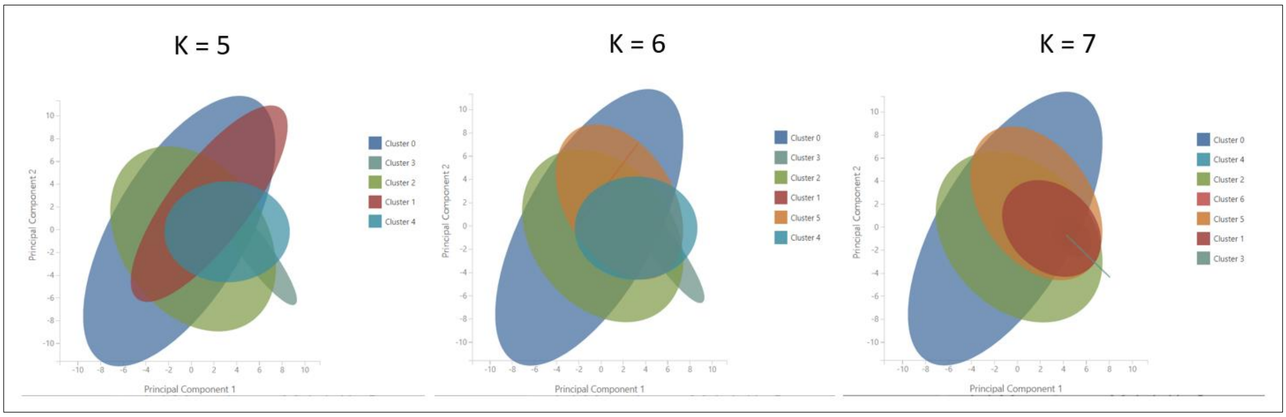


Findings:

Based on the distributions of each additive, it is obvious to see that ingredients b, d, and e are standard additives which have quite normal distributions, and the distributions of ingredients a and g are very identical due to high positive correlations. Besides, ingredient c is added around 3 – 4 units for 66% of the formulas while 86% of the formulas have minimal ingredient h. In relation to the correlation value between ingredients h and c, 66% of the formulations in ingredient c maybe also within the 86% of the formulations in ingredient h. Most of the formulations do not add ingredients h and i, yet the 31% of formulations with ingredient i and 13% with ingredient h are evenly distributed. Using 20 bins histogram, it is found that other than the 33% formulations without ingredient f, ingredient f is mostly added at small volume (48%) and 15% of the formulations are added ingredient f at volume greater than 3rd quartile (0.61), only a small proportion of the formulations (4%) have ingredient f added at extremely large volume (> 0.93 unit) which vary from the finding in part a based on the summary statistics and 10 bins histogram. On the other hand, the distributions of all ingredients f and h show that there are outliers in 2 formulations whereby these ingredients are added a lot to the formulas.

As a conclusion, based on the analyses in part a and b, the most distinguishing additives that affect the petrol burning effects are ingredient h, c, f, and g.

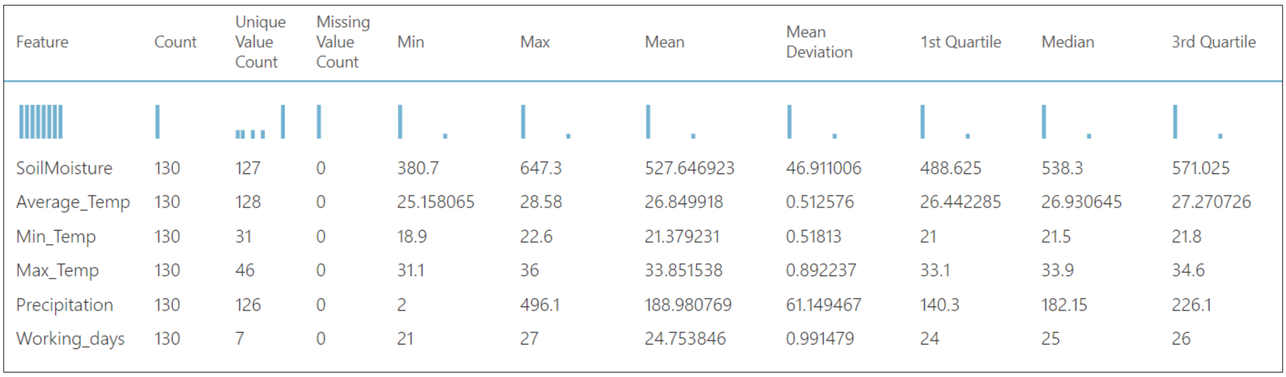
c. A clustering test of your choice (unsupervised learning), to determine the distinctive number of formulations present in the dataset.



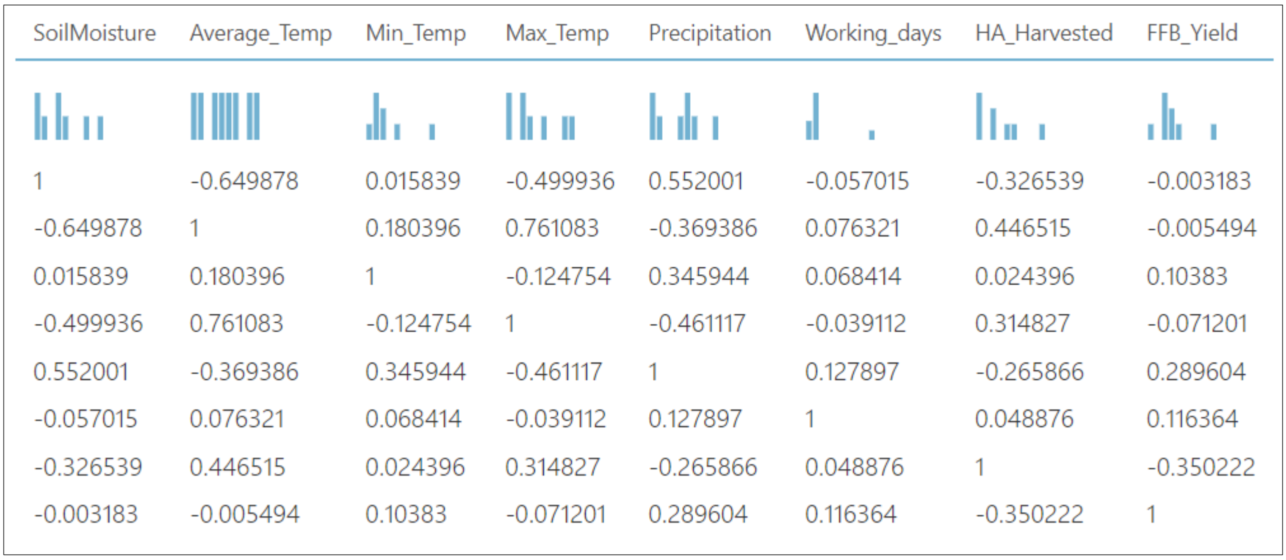
Findings:

Based on the analyses in part a and b, it is found that ingredient b, d, and e are the main ingredients in most formulations, and ingredient a is highly correlated to ingredient g and e. To find the distinctive number of formulations, the main ingredients that are not distinctive shall be excluded from the clustering test to enhance the distinguishing effect. Thus, only ingredients f, g, h, i, and c are included for clustering. Using K-Means ++ method and Euclidean metric, it is observed that K=5 would give the most distinctive clusters since when K = 6, cluster 4 is mainly like cluster 5, and when K = 7, cluster 1 and cluster 6 are similar to cluster 5. Therefore, it is concluded that there are 5 distinct formulations within the 214 petrol samples.

Question 2: Palm_ffb



Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation	1st Quartile	Median	3rd Quartile
SoilMoisture	130	127	0	380.7	647.3	527.646923	46.911006	488.625	538.3	571.025
Average_Temp	130	128	0	25.158065	28.58	26.849918	0.512576	26.442285	26.930645	27.270726
Min_Temp	130	31	0	18.9	22.6	21.379231	0.51813	21	21.5	21.8
Max_Temp	130	46	0	31.1	36	33.851538	0.892237	33.1	33.9	34.6
Precipitation	130	126	0	2	496.1	188.980769	61.149467	140.3	182.15	226.1
Working_days	130	7	0	21	27	24.753846	0.991479	24	25	26

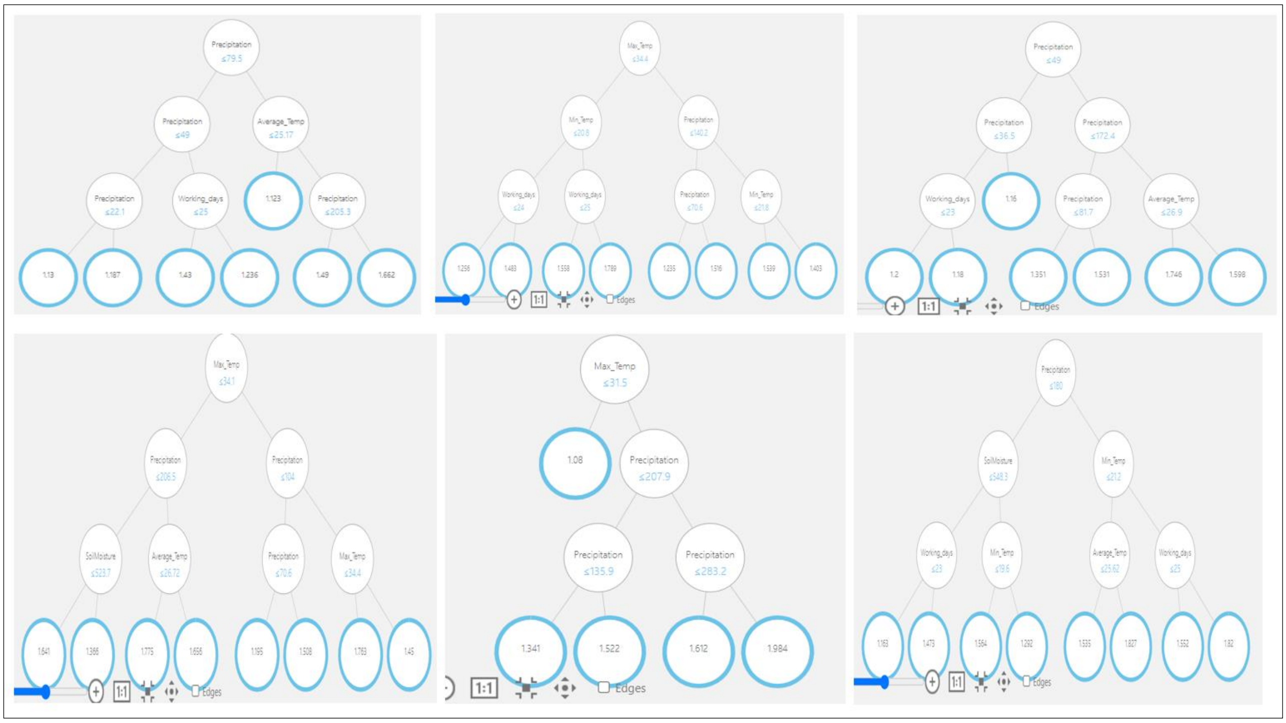


SoilMoisture	Average_Temp	Min_Temp	Max_Temp	Precipitation	Working_days	HA_Harvested	FFB_Yield
1	-0.649878	0.015839	-0.499936	0.552001	-0.057015	-0.326539	-0.003183
-0.649878	1	0.180396	0.761083	-0.369386	0.076321	0.446515	-0.005494
0.015839	0.180396	1	-0.124754	0.345944	0.068414	0.024396	0.10383
-0.499936	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

- FFB_Yield is most correlated to HA_Harvested, followed by Precipitation, Working_days, and Min_Temp, but FFB_Yield is derived from HA_Harvested, hence is not the influencing factor. Although HA_Harvested seems to be the most significant feature, the influence of this variable leads to an insight that “lowering harvesting area would improve FFB_Yield”, which actually reduces the harvest hence is not reasonable. Thus, this variable shall be excluded from the model.

- “Best possible FFB yields are obtained under optimal climatic conditions, with at least 2,000 mm of rainfall homogeneously distributed throughout the year corresponding to around 167 mm month−1. Also, minimum temperatures should be between 22 and 24 °C and maximum temperatures between 29 and 33 °C, while relative humidity should be greater than 85%.” (https://www.nature.com/articles/s41598-018-20298-0)

	MAE	RMSE
Linear Regression	0.226335	0.273042
Boosted Decision Tree	0.23342	0.283221
Decision Forest Regression	0.22541	0.272072
Neural network	0.249306	0.295633



Findings:

From the top 6 regression trees plotted, it is concluded that the main determinants of FFB_Yield are precipitation, followed by maximum, minimum, and average temperatures. The average temperature is seen to be majorly influenced by maximum and minimum temperatures, hence maximum and minimum temperatures are still the more significant factors. However, the models developed are not robust due to the low correlations between FFB_Yield and temperatures. This may also be due to the small dataset collection whereby the temperatures are the aggregated monthly. As the temperature in Malaysia does not fluctuate drastically, the aggregated temperatures do not form significant pattern to the FFB_Yield. Therefore, it is suggested that the data can be recorded at least 3 times per month to better capture the correlations between temperatures and FFB_Yield.

Question 3

This part of analysis is done in Google Colab.

Data preparation:

```
# import text and convert to lower case
Text = "As a term, data analytics predominantly refers to an assortment of applic
ations, from basic business intelligence (BI), reporting and online analytical pr
ocessing (OLAP) to various forms of advanced analytics. In that sense, it's simil
ar in nature to business analytics, another umbrella term for approaches to analy
zing data -
- with the difference that the latter is oriented to business uses, while data an
alytics has a broader focus. The expansive view of the term isn't universal, thou
gh: In some cases, people use data analytics specifically to mean advanced analyt
ics, treating BI as a separate category. Data analytics initiatives can help busi
nesses increase revenues, improve operational efficiency, optimize marketing camp
aigns and customer service efforts, respond more quickly to emerging market trend
s and gain a competitive edge over rivals -
- all with the ultimate goal of boosting business performance. Depending on the p
articular application, 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. At a high level, data analytics methodologies include expl
oratory data analysis (EDA), which aims to find patterns and relationships in dat
a, and confirmatory data analysis (CDA), which applies statistical techniques to
determine whether hypotheses about a data set are true or false. EDA is often com
pared to detective work, while CDA is akin to the work of a judge or jury during
a court trial -
- a distinction first drawn by statistician John W. Tukey in his 1977 book Explor
atory Data Analysis. 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, them
es and points of view. "
text = Text.lower()

# import needed libraries
import string
import nltk
nltk.download('punkt')
from collections import Counter
import pandas as pd
```

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

```
# to split text into lines with fullstop as the delimiter
lines = text.split(sep='.')

# to keep only lines with words
lines = lines[:-1]
lines

for line in lines:
  token_list = nltk.word_tokenize(line)
  word_list = []
  for word in token_list:
    if word.isalnum():
      word_list.append(word)
  total = len(word_list)
  cnt = 0
  for w in word_list:
    if w == 'data':
      cnt += 1
  prob = round(((cnt/total)*100),2)
  print("probability of 'data' appear in line\n'{}': ".format(line), prob, "%\n")
```

Output:

```
probability of 'data' appear in line
'as a term, data analytics predominantly refers to an assortment of applications,
from basic business intelligence (bi), reporting and online analytical processing
(olap) to various forms of advanced analytics':  3.45 %

probability of 'data' appear in line
' in that sense, it's similar in nature to business analytics, another umbrella
term for approaches to analyzing data -- with the difference that the latter is
oriented to business uses, while data analytics has a broader focus':  5.56 %

probability of 'data' appear in line
' the expansive view of the term isn't universal, though: in some cases, people
use data analytics specifically to mean advanced analytics, treating bi as a
separate category':  3.7 %

probability of 'data' appear in line
' data analytics initiatives can help businesses increase revenues, improve
operational efficiency, optimize marketing campaigns and customer service
efforts, respond more quickly to emerging market trends and gain a competitive
edge over rivals -- all with the ultimate goal of boosting business performance':
2.44 %

probability of 'data' appear in line
' depending on the particular application, the data that's analyzed can consist
of either historical records or new information that has been processed for real-
time analytics uses':  4.0 %

probability of 'data' appear in line
' in addition, it can come from a mix of internal systems and external data
sources':  6.67 %

probability of 'data' appear in line
```

```
' at a high level, data analytics methodologies include 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':  11.9 %

probability of 'data' appear in line
' eda is often compared to detective work, while cda is akin to the work of a
judge or jury during a court trial -- a distinction first drawn by statistician
john w':  0.0 %

probability of 'data' appear in line
' tukey in his 1977 book exploratory data analysis':  12.5 %

probability of 'data' appear in line
' data analytics can also be separated into quantitative data analysis and
qualitative data analysis':  21.43 %

probability of 'data' appear in line
' the former involves analysis of numerical data with quantifiable variables that
can be compared or measured statistically':  5.88 %

probability of 'data' appear in line
' 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':  3.57 %
```

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

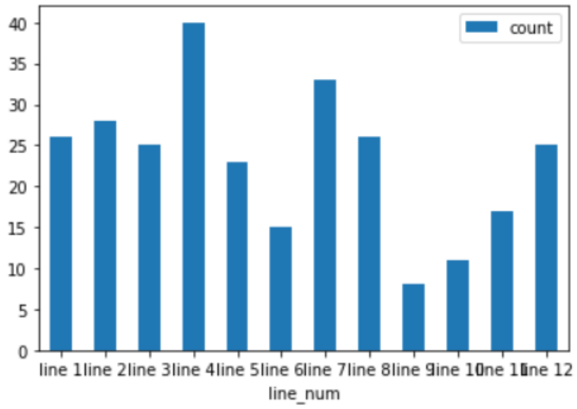
```python
# create list of line number
line_num = []
for i in range(len(lines)):
  i = i+1
  line_num.append("line {}".format(i))

# create dataframe with line number and lines in text
df = pd.DataFrame(line_num)
df.columns = ['line_num']
df['line'] = lines

# create a list of distinct word counts
count_list = []
for line in lines:
  token_list = nltk.word_tokenize(line)
  word_list = []
  for word in token_list:
    if word.isalnum() and word not in word_list:
      word_list.append(word)
  count = len(word_list)
  count_list.append(count)
  print("Distinct word counts in line\n'{}': \n".format(line), count, "\n")

# add number of distinct words in each line to the dataframe
df['count'] = count_list
```

Output:

```
Distinct word counts in line
'as a term, data analytics predominantly refers to an assortment of applications,
from basic business intelligence (bi), reporting and online analytical processing
(olap) to various forms of advanced analytics':
 26

Distinct word counts in line
' in that sense, it's similar in nature to business analytics, another umbrella
term for approaches to analyzing data -- with the difference that the latter is
oriented to business uses, while data analytics has a broader focus':
 28

Distinct word counts in line
' the expansive view of the term isn't universal, though: in some cases, people use
data analytics specifically to mean advanced analytics, treating bi as a separate
category':
 25

Distinct word counts in line
' data analytics initiatives can help businesses increase revenues, improve
operational efficiency, optimize marketing campaigns and customer service efforts,
respond more quickly to emerging market trends and gain a competitive edge over
rivals -- all with the ultimate goal of boosting business performance':
 40

Distinct word counts in line
' depending on the particular application, the data that's analyzed can consist of
either historical records or new information that has been processed for real-time
analytics uses':
 23

Distinct word counts in line
' in addition, it can come from a mix of internal systems and external data
sources':
 15

Distinct word counts in line
' at a high level, data analytics methodologies include 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':
 33

Distinct word counts in line
' eda is often compared to detective work, while cda is akin to the work of a judge
or jury during a court trial -- a distinction first drawn by statistician john w':
 26

Distinct word counts in line
' tukey in his 1977 book exploratory data analysis':
 8

Distinct word counts in line
' data analytics can also be separated into quantitative data analysis and
qualitative data analysis':
 11

Distinct word counts in line
' the former involves analysis of numerical data with quantifiable variables that
can be compared or measured statistically':
 17

Distinct word counts in line
' 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':
 25
```

```
# bar plot to show distribution of distinct words across all the lines
ax = df.plot.bar(x='line_num', y='count', rot=0)
```

Output:



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

```
# create list of tokens
word_list = []
for line in lines:
  token_list = nltk.word_tokenize(line)
  for word in token_list:
    if word.isalnum():
      word_list.append(word)

# create bigrams with tokens
bigram = list(nltk.bigrams(word_list))

# count the number of ('data', 'analytics') bigrams
cnt = 0
for grp in bigram:
  if grp[0] == 'data' and grp[1] == 'analytics':
    cnt += 1

# calculate the probability by divided to the total number of bigrams
print("probability when 'analytics' comes after 'data': ", round(((cnt/len(bigram
))*100),2), "%")
```

Output:

```
probability when 'analytics' comes after 'data':  1.92 %
```