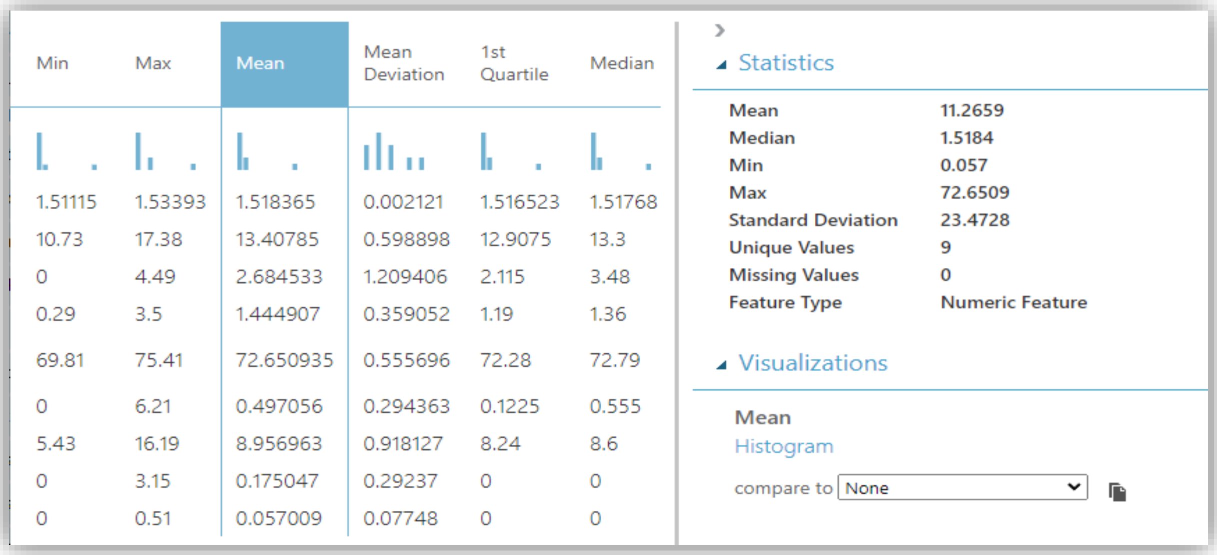
PreScreen Q1

a)



Min	Max	Mean	Mean Deviation	1st Quartile	Median
1.51115	1.53393	1.518365	0.002121	1.516523	1.51768
10.73	17.38	13.40785	0.598898	12.9075	13.3
0	4.49	2.684533	1.209406	2.115	3.48
0.29	3.5	1.444907	0.359052	1.19	1.36
69.81	75.41	72.650935	0.555696	72.28	72.79
0	6.21	0.497056	0.294363	0.1225	0.555
5.43	16.19	8.956963	0.918127	8.24	8.6
0	3.15	0.175047	0.29237	0	0
0	0.51	0.057009	0.07748	0	0

Statistics

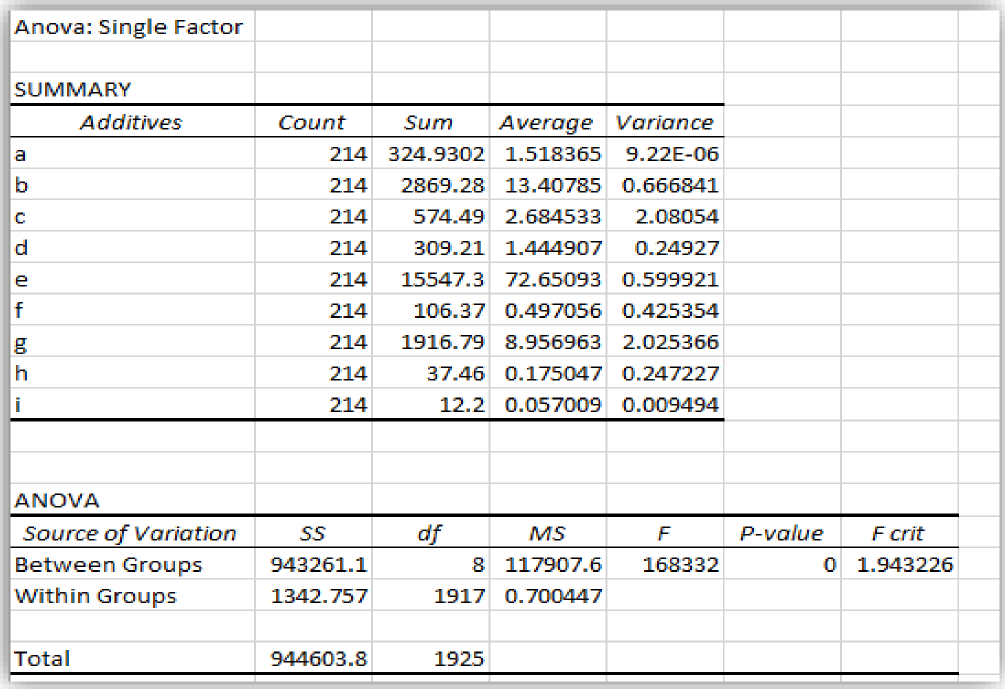
Mean	11.2659
Median	1.5184
Min	0.057
Max	72.6509
Standard Deviation	23.4728
Unique Values	9
Missing Values	0
Feature Type	Numeric Feature

Visualizations

Mean
Histogram

compare to None

As shown in the figure above, the mean of each additive significantly varies with from around 0.05 to 72. From the value stated, we can see that the formulations are significantly different but with several groups as there are very high readings and some do not differ from each other outstandingly. This conclusion will be further supported with the one-way ANOVA.



Anova: Single Factor

SUMMARY

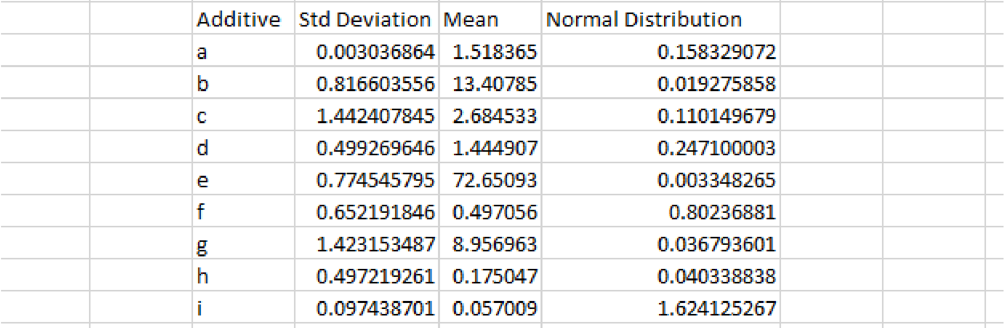
Additives	Count	Sum	Average	Variance
a	214	324.9302	1.518365	9.22E-06
b	214	2869.28	13.40785	0.666841
c	214	574.49	2.684533	2.08054
d	214	309.21	1.444907	0.24927
e	214	15547.3	72.65093	0.599921
f	214	106.37	0.497056	0.425354
g	214	1916.79	8.956963	2.025366
h	214	37.46	0.175047	0.247227
i	214	12.2	0.057009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

ANOVA

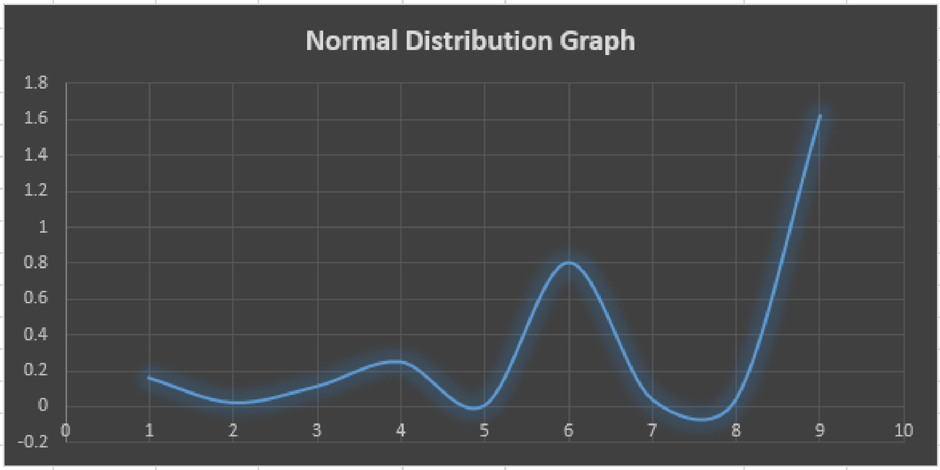
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	943261.1	8	117907.6	168332	0	1.943226
Within Groups	1342.757	1917	0.700447			
Total	944603.8	1925				

By applying one-way ANOVA to the data, we can observe that the P-value is less than the usual significance level that is 0.05. This helps us to reject the null hypothesis and can conclude that the additives as significant differences.

b)

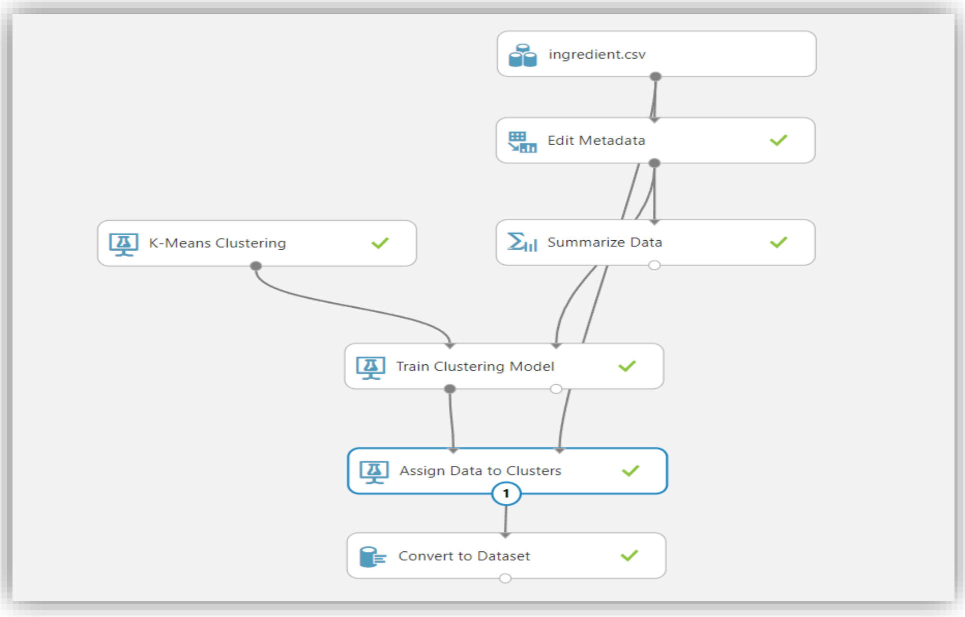


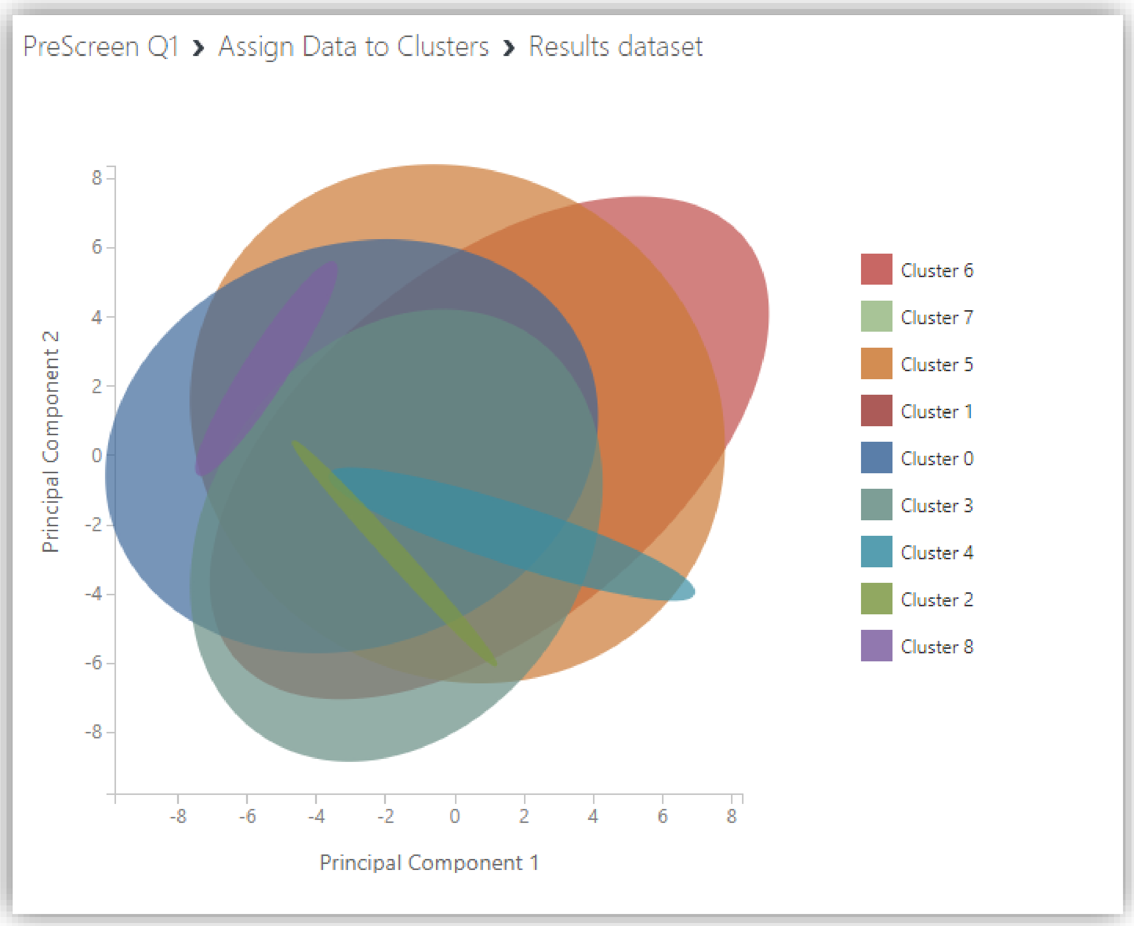
Additive	Std Deviation	Mean	Normal Distribution
a	0.003036864	1.518365	0.158329072
b	0.816603556	13.40785	0.019275858
c	1.442407845	2.684533	0.110149679
d	0.499269646	1.444907	0.247100003
e	0.774545795	72.65093	0.003348265
f	0.652191846	0.497056	0.80236881
g	1.423153487	8.956963	0.036793601
h	0.497219261	0.175047	0.040338838
i	0.097438701	0.057009	1.624125267



In the graph above, number represents the additives starting from 1 as “a" and 9 as “i”. We can see that the value are not normally distributed with some has significantly higher value than the other. The normal bell shaped curve graph is not present here.

c)





By producing clusters using the dataset, we can observe that there are only 8 clusters in the figure above compared to the 9 different additives. However all the clusters' centroid are far from each other. This strengthen the claim that the formulations of additives are significantly different from each other.

PreScreen Q2

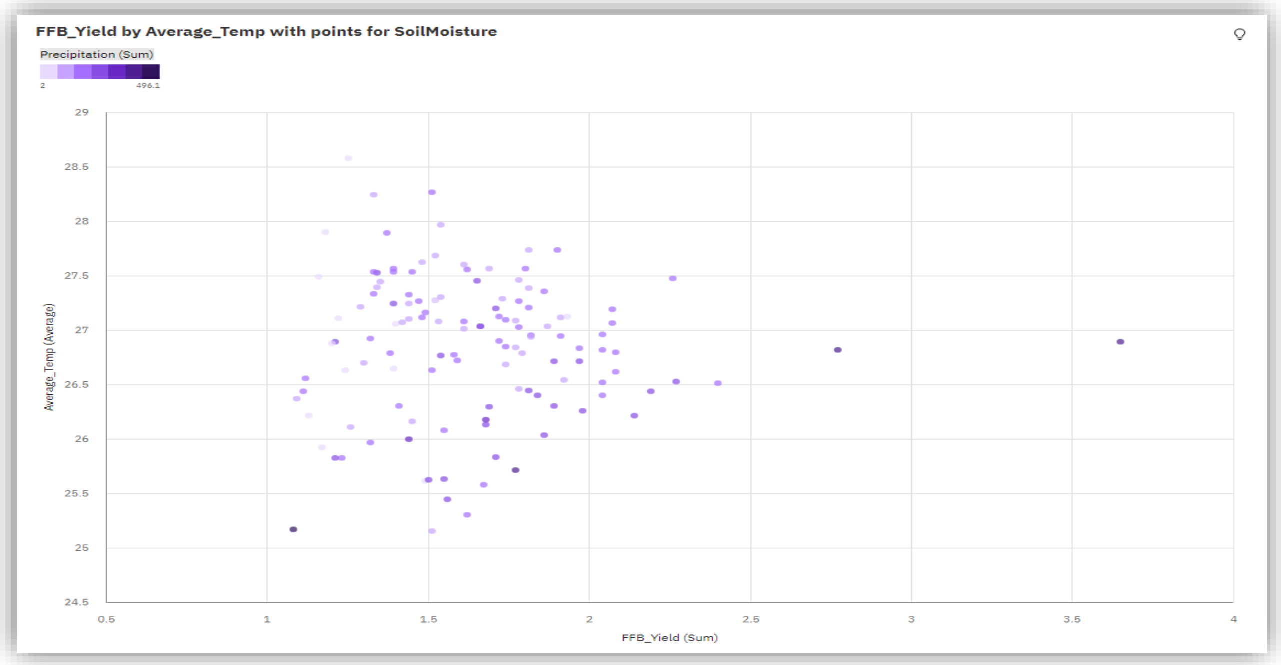
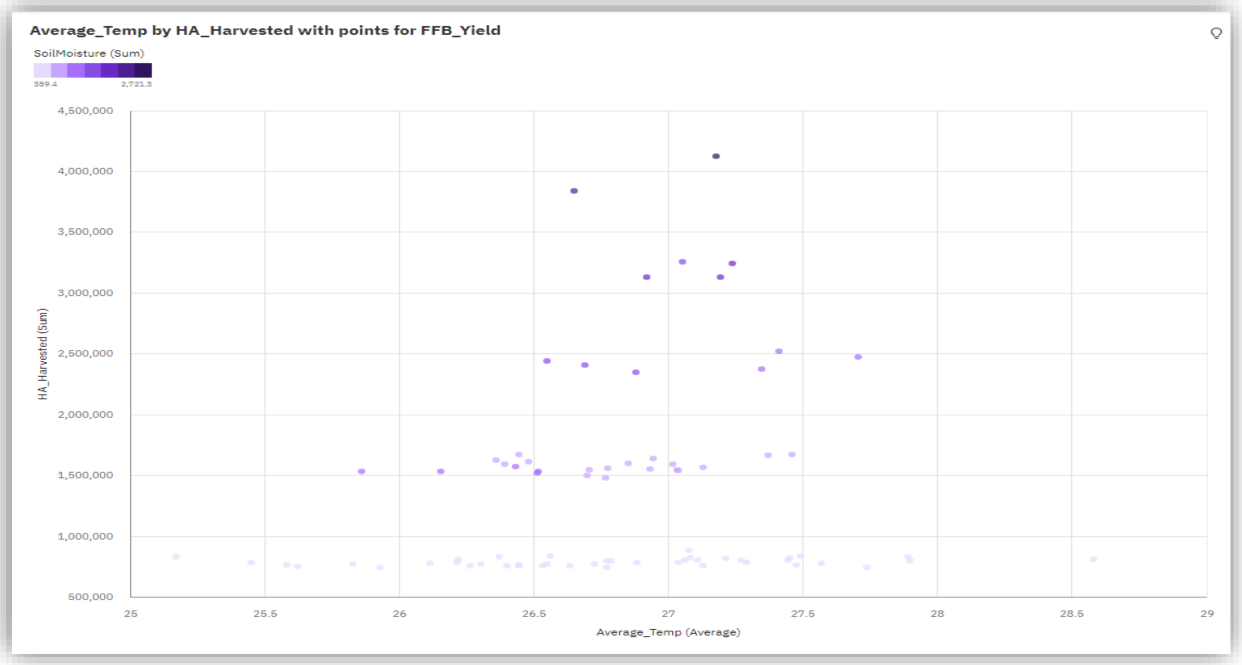
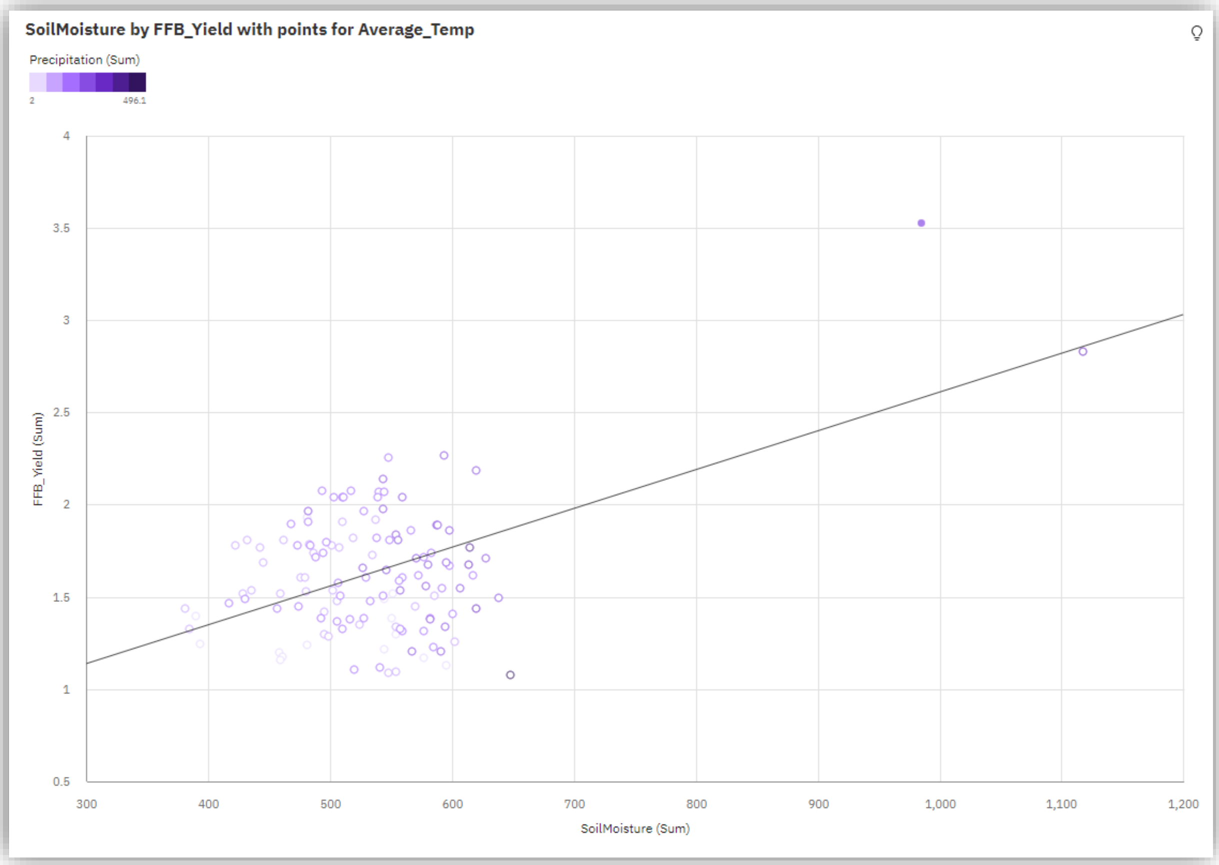


Figure above shows the relationship between the sum of FFB_Yield with Average Temperature by using Soil Moisture as the points and precipitation as color. We can observe that the higher number of FFB_Yield sum are concentrated around 26 to 27.5 degrees celcius. Those temperature too are where the Soil Moisture is suitable for the crops.



The figure potrays the relationship between average temperature and sum of HA_Harvested with FFB_Yield as points and Soil Moisture as color. Just as the sum of FFB_Yield, we can see that most of the crops are Harvested in big amount on the average temperature between 25 to 27.5 degrees celcius.



The scatter plot above shows the correlation between sum of Soil Moisture and FFB_Yield with Average Temperature as points. We can observe that the FBB_Yield is unusually high when the Soil Moisture is 984.3 and it differs from the trendline with 99% confidence level making the data for that point is not reliable. Most of the points are scattered between 400 and 650 of the Soil Moisture sum and are not far from the trendline. FFB_Yield and SoilMoisture have a weak positive linear association.

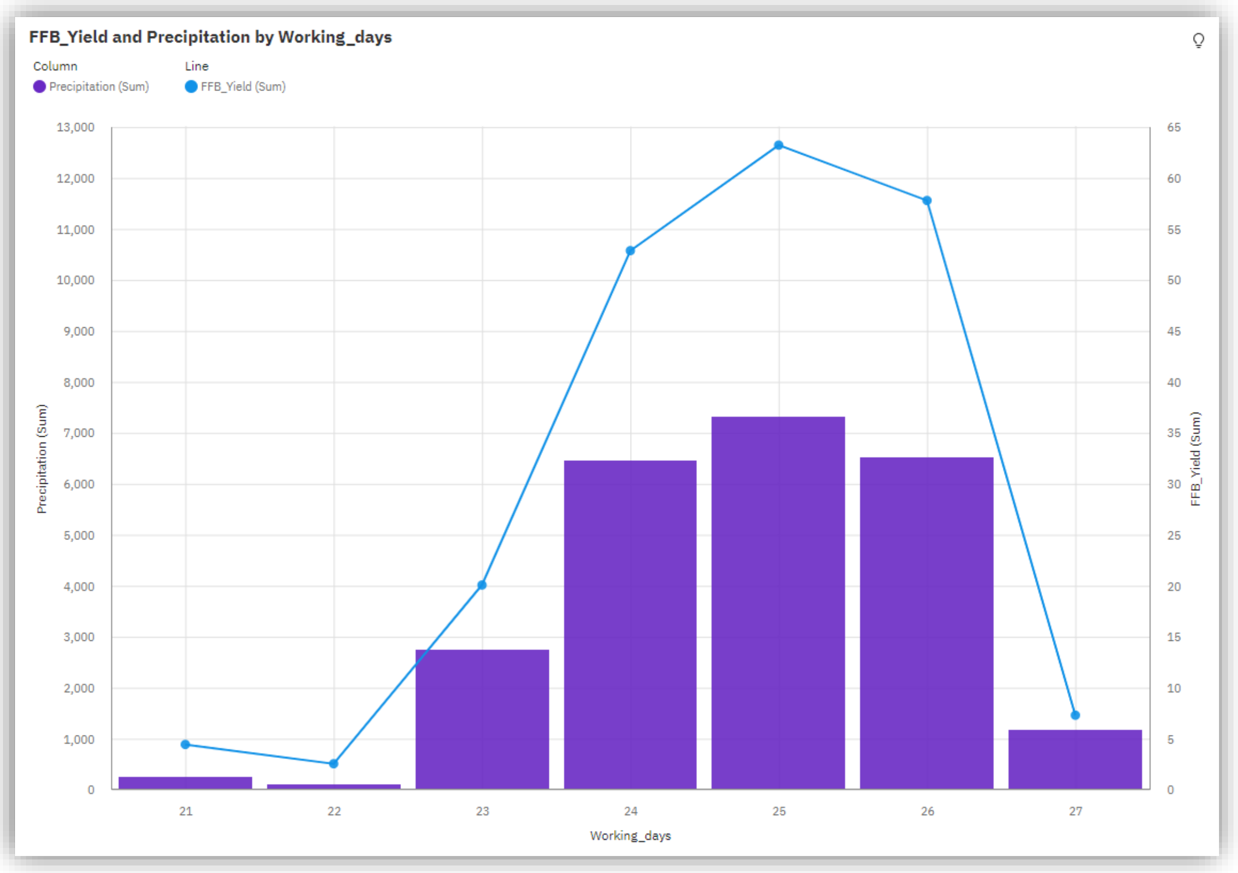
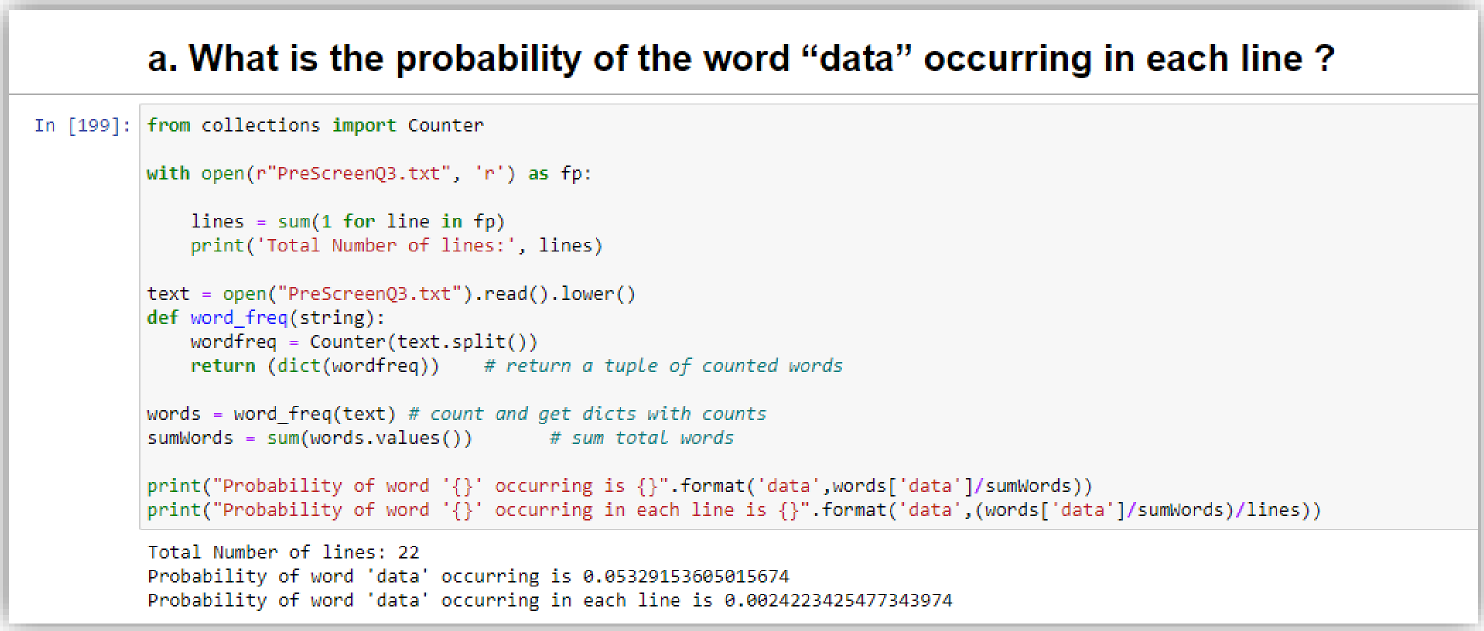


Figure above shows the relationship between the FFB_Yield with Precipitation and Working_days. For FFB_Yield, the most significant values of Working_days are 25, 26, and 24, whose respective FFB_Yield values add up to 173.9, or 83.5 % of the total. The FFB_Yield ranges from 2.55, when Working_days is 22, to 63.23, when Working_days is 25. This also shows that the FFB_Yield is significantly high when the Working_days is 25. We can see that production is good when the farmers are working for 25 days.

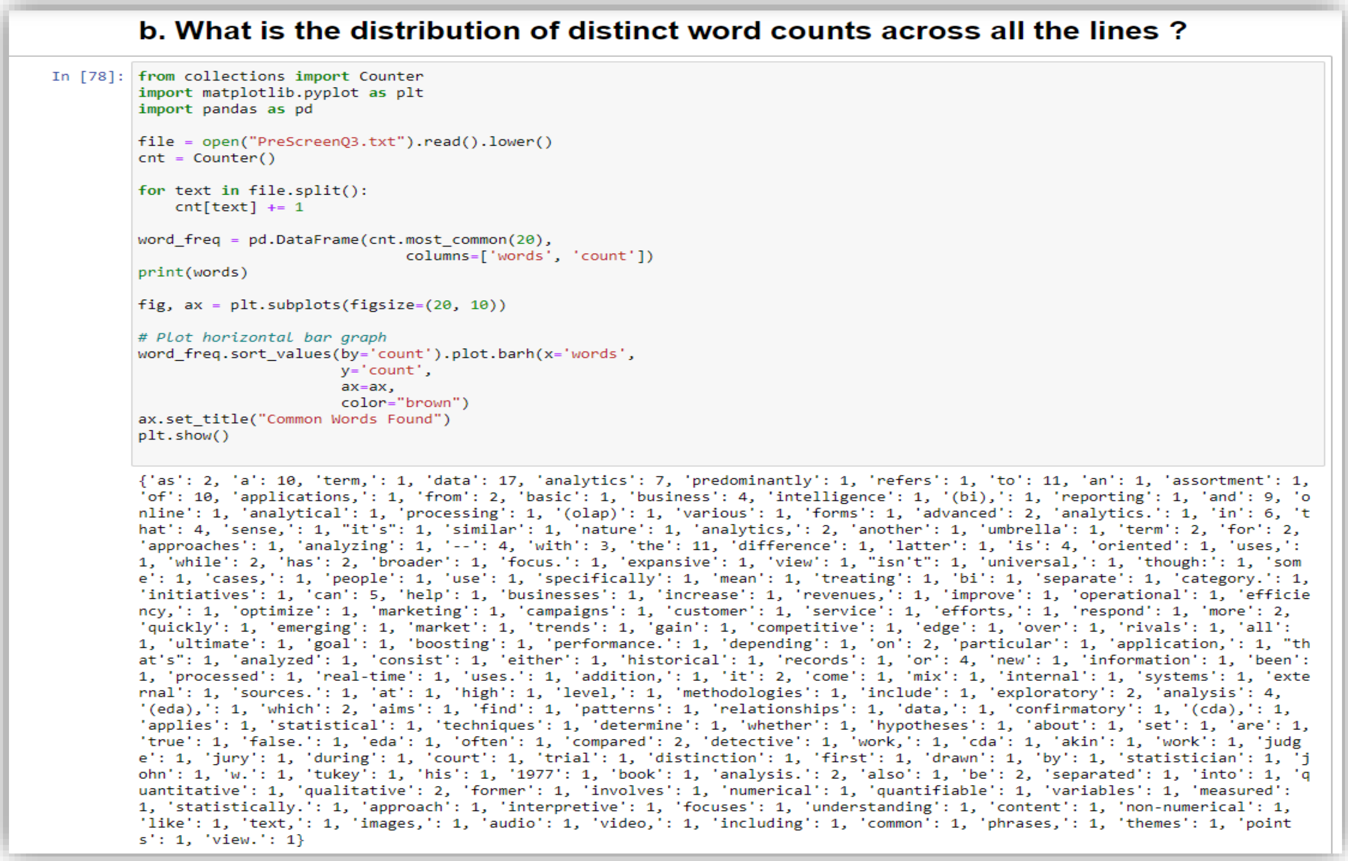
PreScreen Q3



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

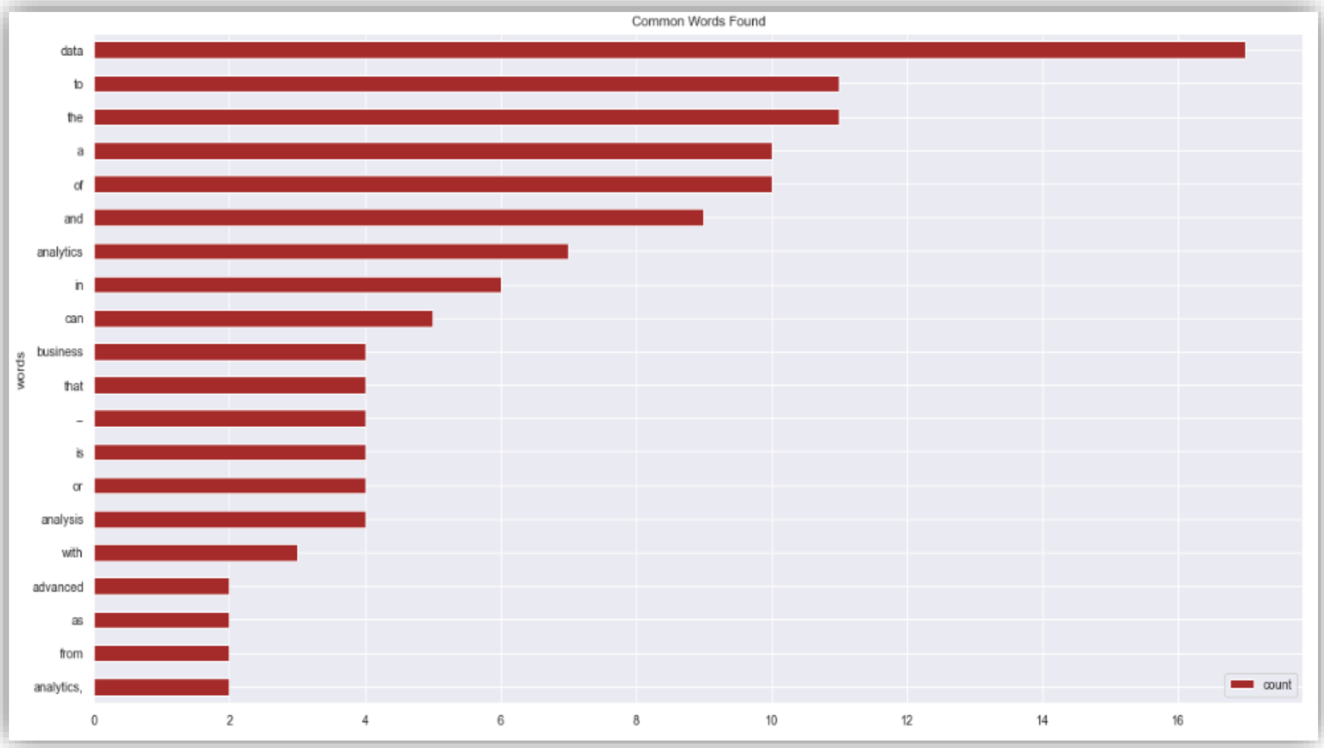
```
In [199]: from collections import Counter

          with open(r"PreScreenQ3.txt", 'r') as fp:

              lines = sum(1 for line in fp)
              print('Total Number of lines:', lines)

          text = open("PreScreenQ3.txt").read().lower()
          def word_freq(string):
              wordfreq = Counter(text.split())
              return (dict(wordfreq))    # return a tuple of counted words

          words = word_freq(text) # count and get dicts with counts
          sumWords = sum(words.values())        # sum total words

          print("Probability of word '{}' occurring is {}".format('data',words['data']/sumWords))
          print("Probability of word '{}' occurring in each line is {}".format('data',(words['data']/sumWords)/lines))

          Total Number of lines: 22
          Probability of word 'data' occurring is 0.05329153605015674
          Probability of word 'data' occurring in each line is 0.0024223425477343974
```

In order to get the probability of the word “data” occurring in each line, we have to get the total number of line in the text which is 22. After that we need to get the probability of the word data occurring in the whole text. Finally we divide the value of probability and total number of lines to get the probability of word “data” occurring in each line.

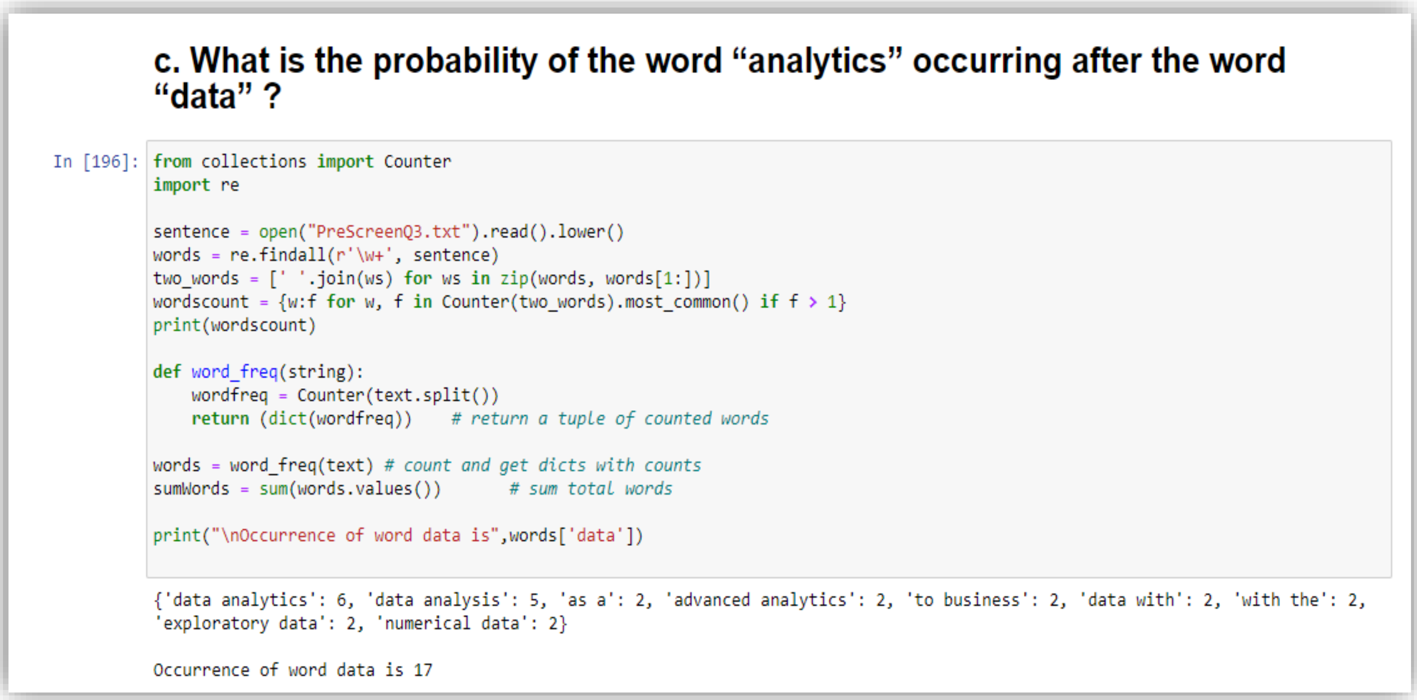


b. What is the distribution of distinct word counts across all the lines ?

```
In [78]: from collections import Counter
         import matplotlib.pyplot as plt
         import pandas as pd

         file = open("PreScreenQ3.txt").read().lower()
         cnt = Counter()

         for text in file.split():
             cnt[text] += 1

         word_freq = pd.DataFrame(cnt.most_common(20),
                                  columns=['words', 'count'])
         print(words)

         fig, ax = plt.subplots(figsize=(20, 10))

         # Plot horizontal bar graph
         word_freq.sort_values(by='count').plot.barh(x='words',
                               y='count',
                               ax=ax,
                               color="brown")
         ax.set_title("Common Words Found")
         plt.show()

         {'as': 2, 'a': 10, 'term,': 1, 'data': 17, 'analytics': 7, 'predominantly': 1, 'refers': 1, 'to': 11, 'an': 1, 'assortment': 1,
         'of': 10, 'applications,': 1, 'from': 2, 'basic': 1, 'business': 4, 'intelligence': 1, '(bi),': 1, 'reporting': 1, 'and': 9, 'o
         nline': 1, 'analytical': 1, 'processing': 1, '(olap)': 1, 'various': 1, 'forms': 1, 'advanced': 2, 'analytics.': 1, 'in': 6, 't
         hat': 4, 'sense,': 1, "it's": 1, 'similar': 1, 'nature': 1, 'analytics,': 2, 'another': 1, 'umbrella': 1, 'term': 2, 'for': 2,
         'approaches': 1, 'analyzing': 1, '--': 4, 'with': 3, 'the': 11, 'difference': 1, 'latter': 1, 'is': 4, 'oriented': 1, 'uses,':
         1, 'while': 2, 'has': 2, 'broader': 1, 'focus.': 1, 'expansive': 1, 'view': 1, "isn't": 1, 'universal,': 1, 'though:': 1, 'som
         e': 1, 'cases,': 1, 'people': 1, 'use': 1, 'specifically': 1, 'mean': 1, 'treating': 1, 'bi': 1, 'separate': 1, 'category.': 1,
         'initiatives': 1, 'can': 5, 'help': 1, 'businesses': 1, 'increase': 1, 'revenues,': 1, 'improve': 1, 'operational': 1, 'efficie
         ncy,': 1, 'optimize': 1, 'marketing': 1, 'campaigns': 1, 'customer': 1, 'service': 1, 'efforts,': 1, 'respond': 1, 'more': 2,
         'quickly': 1, 'emerging': 1, 'market': 1, 'trends': 1, 'gain': 1, 'competitive': 1, 'edge': 1, 'over': 1, 'rivals': 1, 'all':
         1, 'ultimate': 1, 'goal': 1, 'boosting': 1, 'performance.': 1, 'depending': 1, 'on': 2, 'particular': 1, 'application,': 1, "th
         at's": 1, 'analyzed': 1, 'consist': 1, 'either': 1, 'historical': 1, 'records': 1, 'or': 4, 'new': 1, 'information': 1, 'been':
         1, 'processed': 1, 'real-time': 1, 'uses.': 1, 'addition,': 1, 'it': 2, 'come': 1, 'mix': 1, 'internal': 1, 'systems': 1, 'exte
         rnal': 1, 'sources.': 1, 'at': 1, 'high': 1, 'level,': 1, 'methodologies': 1, 'include': 1, 'exploratory': 2, 'analysis': 4,
         '(eda),': 1, 'which': 2, 'aims': 1, 'find': 1, 'patterns': 1, 'relationships': 1, 'data,': 1, 'confirmatory': 1, '(cda),': 1,
         'applies': 1, 'statistical': 1, 'techniques': 1, 'determine': 1, 'whether': 1, 'hypotheses': 1, 'about': 1, 'set': 1, 'are': 1,
         'true': 1, 'false.': 1, 'eda': 1, 'often': 1, 'compared': 2, 'detective': 1, 'work,': 1, 'cda': 1, 'akin': 1, 'work': 1, 'judg
         e': 1, 'jury': 1, 'during': 1, 'court': 1, 'trial': 1, 'distinction': 1, 'first': 1, 'drawn': 1, 'by': 1, 'statistician': 1, 'j
         ohn': 1, 'w.': 1, 'tukey': 1, 'his': 1, '1977': 1, 'book': 1, 'analysis.': 2, 'also': 1, 'be': 2, 'separated': 1, 'into': 1, 'q
         uantitative': 1, 'qualitative': 2, 'former': 1, 'involves': 1, 'numerical': 1, 'quantifiable': 1, 'variables': 1, 'measured':
         1, 'statistically.': 1, 'approach': 1, 'interpretive': 1, 'focuses': 1, 'understanding': 1, 'content': 1, 'non-numerical': 1,
         'like': 1, 'text,': 1, 'images,': 1, 'audio': 1, 'video,': 1, 'including': 1, 'common': 1, 'phrases,': 1, 'themes': 1, 'point
         s': 1, 'view.': 1}
```



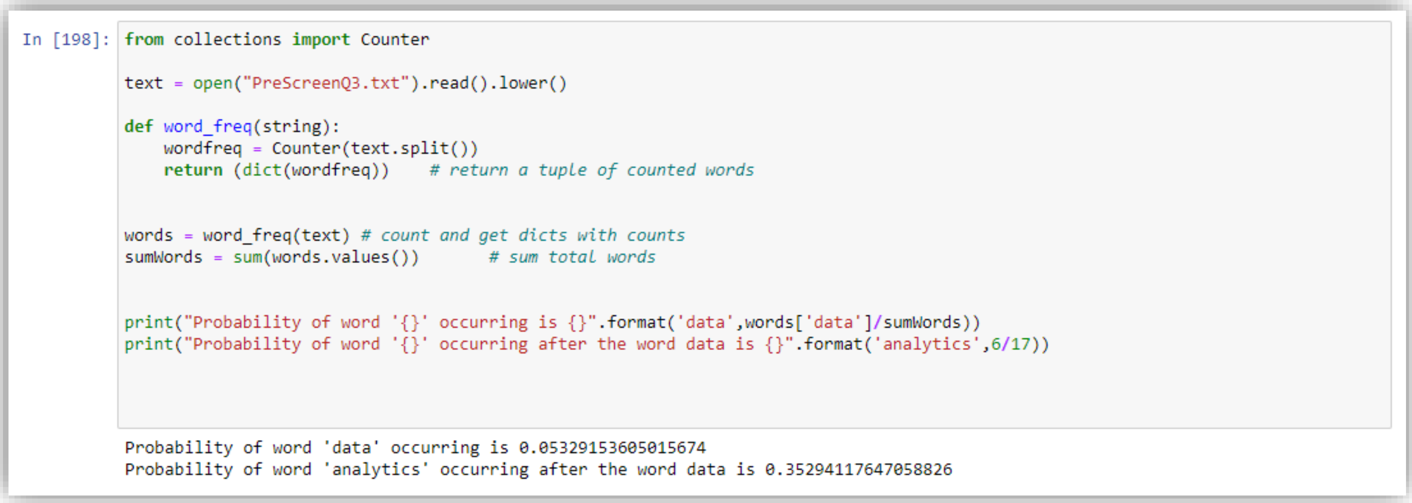
To get the distribution of distinct word in the whole text, we get all the words in the text and the frequency of how many the words are present in the whole text. After that we create a horizontal bar graph with the count as x-axis and words as y-axis. From that, we can see which word has the highest frequency with its distribution.



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

```
In [196]: from collections import Counter
          import re

          sentence = open("PreScreenQ3.txt").read().lower()
          words = re.findall(r'\w+', sentence)
          two_words = [' '.join(ws) for ws in zip(words, words[1:])]
          wordscount = {w:f for w, f in Counter(two_words).most_common() if f > 1}
          print(wordscount)

          def word_freq(string):
              wordfreq = Counter(text.split())
              return (dict(wordfreq))    # return a tuple of counted words

          words = word_freq(text) # count and get dicts with counts
          sumWords = sum(words.values())       # sum total words

          print("\nOccurrence of word data is",words['data'])
```

```
{'data analytics': 6, 'data analysis': 5, 'as a': 2, 'advanced analytics': 2, 'to business': 2, 'data with': 2, 'with the': 2,
'exploratory data': 2, 'numerical data': 2}

Occurrence of word data is 17
```



```python
from collections import Counter

text = open("PreScreenQ3.txt").read().lower()

def word_freq(string):
    wordfreq = Counter(text.split())
    return (dict(wordfreq))     # return a tuple of counted words


words = word_freq(text) # count and get dicts with counts
sumWords = sum(words.values())       # sum total words


print("Probability of word '{}' occurring is {}".format('data',words['data']/sumWords))
print("Probability of word '{}' occurring after the word data is {}".format('analytics',6/17))
```

```
Probability of word 'data' occurring is 0.05329153605015674
Probability of word 'analytics' occurring after the word data is 0.35294117647058826
```

To get the probability of word analytics to occur after data to form data analytics, we have to calculate the occurrence of word data and data analytics in the whole text. After that, we have to divide the number of data analytics occurrence with the word of data occurrence to get the probability.