Questions 1

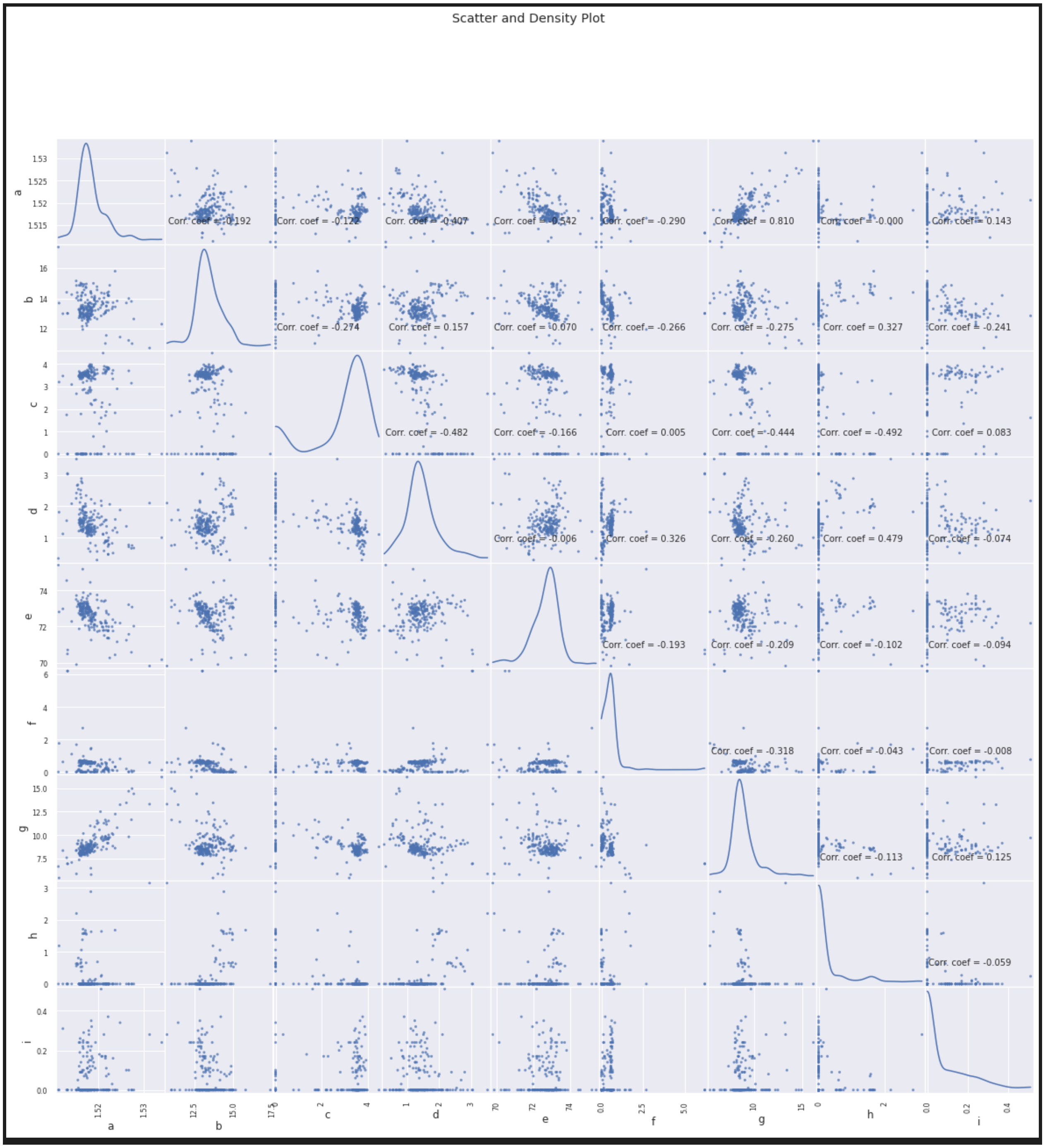
* All the features are in numerical format.
* There are no missing values.
* Mean and median are quite close to each other.
*There are no such outliners
* Not variance that much except for two columns c and g.

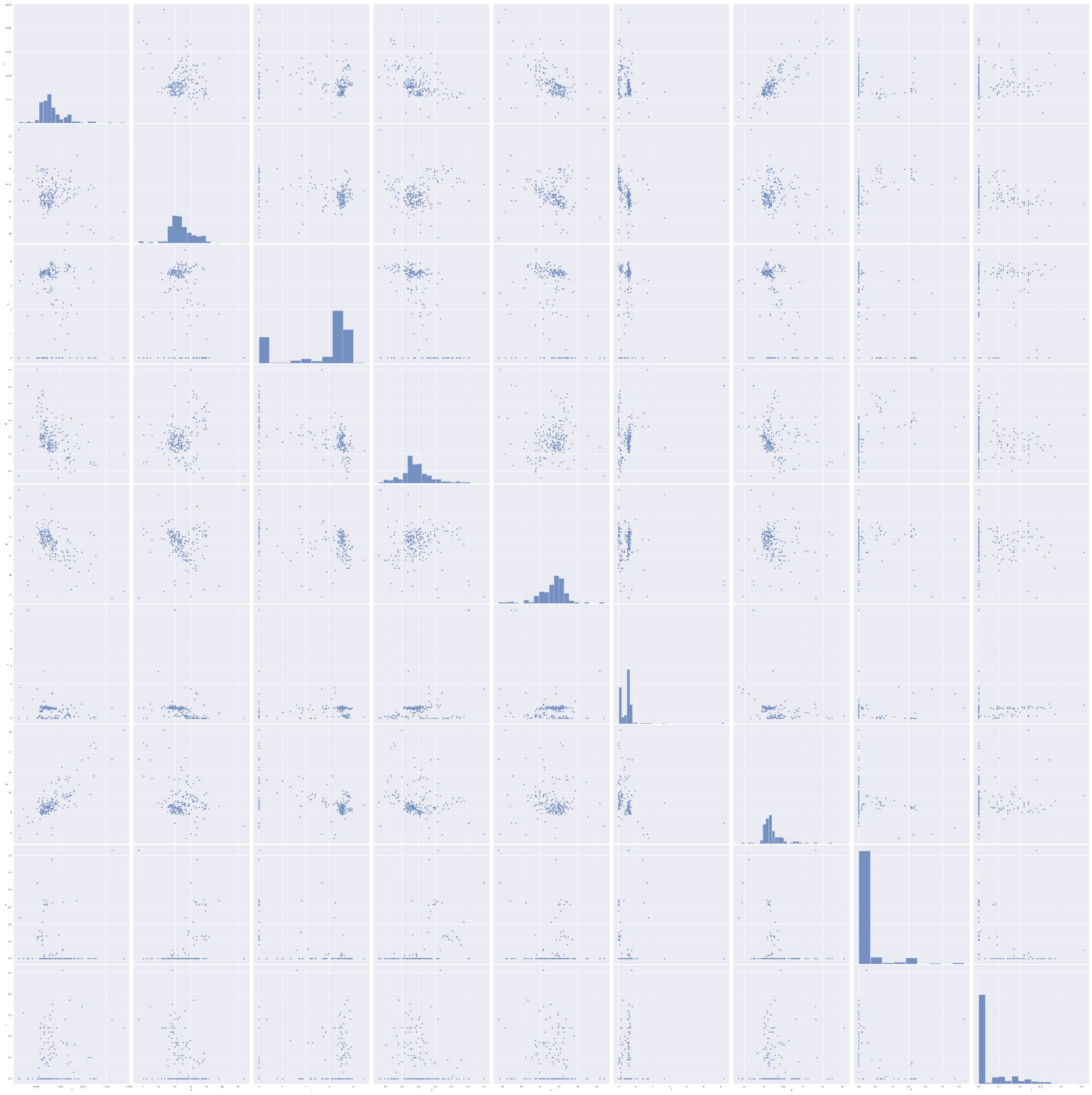
By the correlation method “Pearson”, All the features are normally distributed except for f, h, and i.

strongly positive correlation with

* a & g
* b & h
* d & h
* d & f

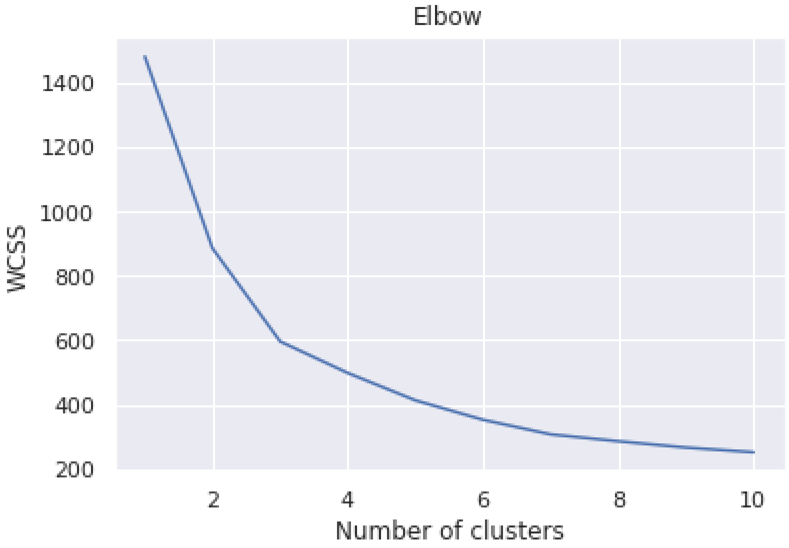
Below graphs are Distribution of Scatter and density plots


Scatter and Density Plot
Corr. coef = -0.192
Corr. coef = -0.122
Corr. coef = -0.407
Corr. coef = -0.542
Corr. coef = -0.290
Corr. coef = 0.810
Corr. coef = -0.000
Corr. coef = 0.143
Corr. coef = -0.274
Corr. coef = 0.157
Corr. coef = -0.070
Corr. coef = -0.266
Corr. coef = -0.275
Corr. coef = 0.327
Corr. coef = -0.241
Corr. coef = -0.482
Corr. coef = -0.166
Corr. coef = 0.005
Corr. coef = -0.444
Corr. coef = -0.492
Corr. coef = 0.083
Corr. coef = -0.006
Corr. coef = 0.326
Corr. coef = -0.260
Corr. coef = 0.479
Corr. coef = -0.074
Corr. coef = -0.193
Corr. coef = -0.209
Corr. coef = -0.102
Corr. coef = -0.094
Corr. coef = -0.318
Corr. coef = -0.043
Corr. coef = -0.008
Corr. coef = -0.113
Corr. coef = 0.125
Corr. coef = -0.059
a
1.53
1.525
1.52
1.515
b
16
14
12
c
4
3
2
1
0
d
3
2
1
e
74
72
70
f
6
4
2
0
g
15.0
12.5
10.0
7.5
h
3
2
1
0
i
0.4
0.2
0.0
1.52
1.53
a
12.5
15.0
17.5
b
0
2
4
c
1
2
3
d
70
72
74
e
0.0
2.5
5.0
f
10
15
g
0
2
h
0.0
0.2
0.4
i

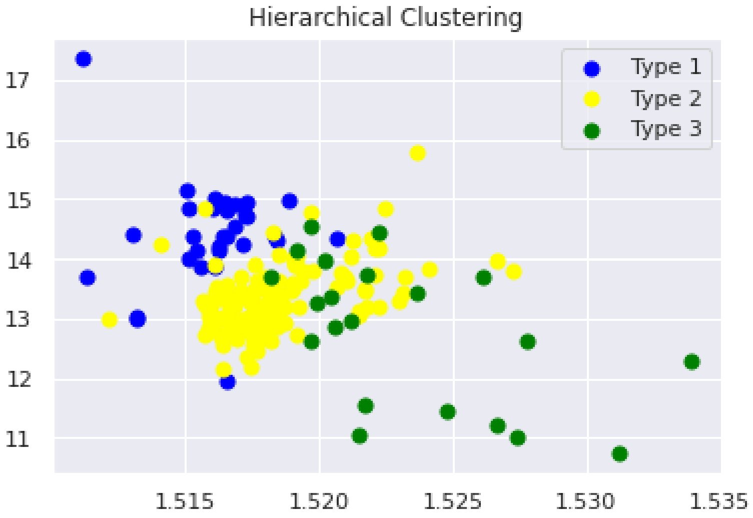


K-means clustering

In below graph, “Elbow” method mentions we should choose 3 clusters



- Below image is hierarchical clustering clusters



Questions 2 (Oil palm data)

As a pairwise correlation, against FFB_yield, one Vs all correlation has been used to study all external factors correlation. So, We can reject the hypothesis that the two variables are not correlated if the p-value is below 0.05, generally. So we could mention, that there is a significant correlation between all the variables against FFB_yield.

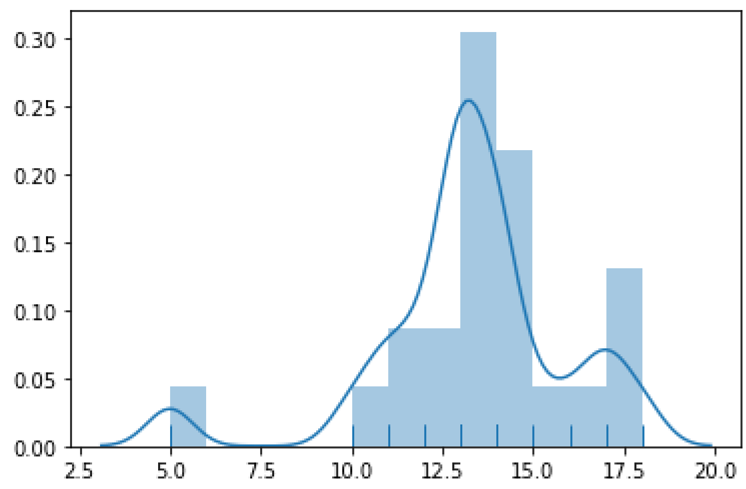
FFB_Yield has
•weak positive correlation with Max_Temp and Min_Temp
•strong negative correlation with HA_Harvested
•strong positive correlation with Precipitation
•strong moderate correlation with Working days.

- Soil Moisture is negatively associated with FFB_Yield.
- Temperature variables show no significant association with FFB_Yield.

Questions 3(NLP- paragraph)

- Probability of “data” appearing in every line id 0.782608695652174
- The number of times “data analytics” appear together is 6
- The number of times only “analytics” appear in the complete text is 10

Below the graph is the distribution of the distinct word counts in every line -



The last question of Question3 is the Probability of “analytics” appearing after “data”.

- Its Probability is **0.6**