Zurriyata Fatni

Email: zurriyata@gmail.com

**Analytics Evaluation**

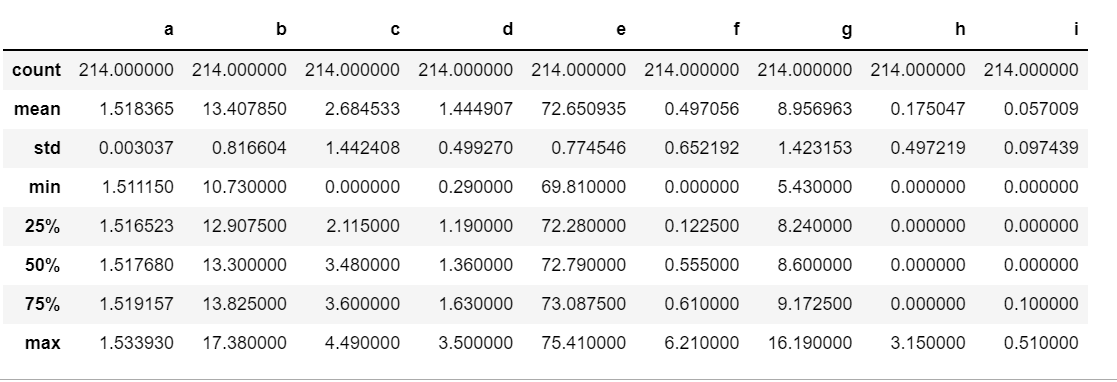
1. **A customer informed their consultant that they have developed several formulations of petrol that gives different characteristics of burning pattern. The formulations are obtaining by adding varying levels of additives that, for example, prevent engine knocking, gum prevention, stability in storage, 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;**

1. **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.**

**Explanation:**

Before we are doing to analysis of the data, we must check the data missing, data noise, or has different data types. We must cleaning the data and transform the data if we needed for analyze the data more easier. For this case, There are some method for descriptive analysis of data like mean, median, standard deviation, etc. I use Python for helping me to describe all data. Descriptive analysis for data additives will be explain below



From the image above, we can conclude that from each variable (additive) got a result:

1. Count of data

From the image above, we can see that rows of the data for each variables are same with 214 rows. There are no missing data

1. Mean

Mean is average of all value from the data. For each variable/additive, got a result of mean below:

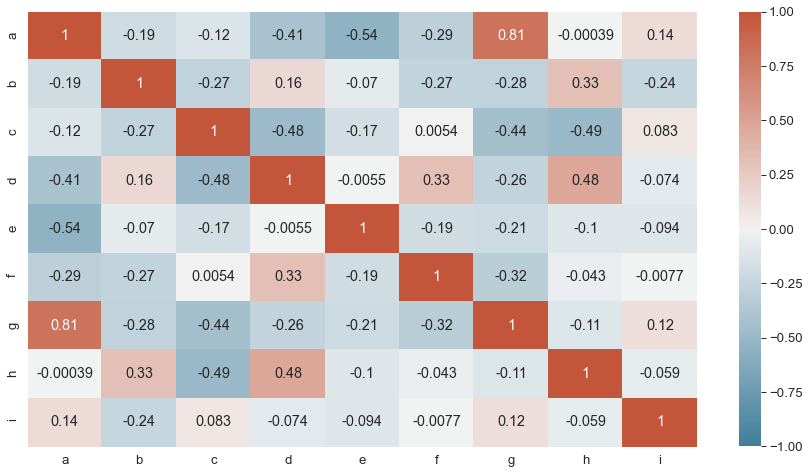
1. Additive a = 1.518265
2. Additive b = 13.407850
3. Additive c = 2.684533
4. Additive d = 1.444907
5. Additive e = 72.650935
6. Additive f = 0.49705
7. Additive g = 8.956963
8. Additive h = 0.175047
9. Additive i = 0.057009
10. Median

Median or quartile 2 is the middle of value from sorted data. From the image above, median value is representation of 50% . For each variables got a result below:

1. Additive a = 1.517680
2. Additive b = 13.3
3. Additive c = 3.48
4. Additive d = 1.36
5. Additive e = 72.79
6. Additive f = 0.555
7. Additive g = 8.6
8. Additive h = 0
9. Additive i = 0
10. Standard deviation (std)

Standard deviation is measure of difference of mean and each data. For each variables from the data above got a result below:

1. Additive a = 0.003037
2. Additive b = 0.816604
3. Additive c = 1.442408
4. Additive d = 0.499270
5. Additive e = 0.774546
6. Additive f = 0.652192
7. Additive g = 1.423153
8. Additive h = 0.497219
9. Additive i = 0.097439

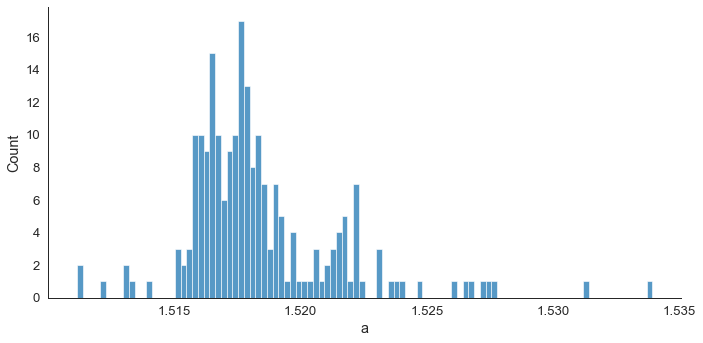
For the image below, this is a heatmap correlation matrix

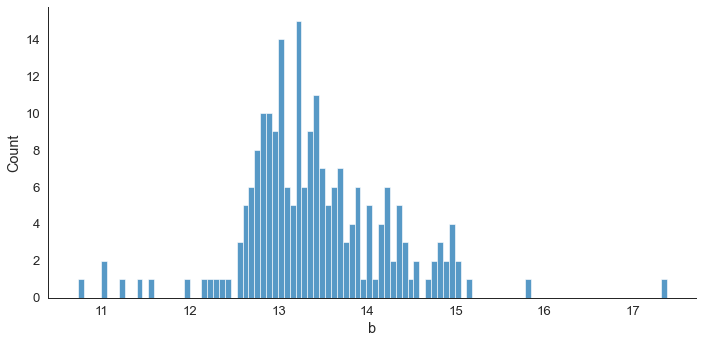
For knowing what the additive have a strong correlation or depend with some additive, we can using correlation matrix. From the image above, we can see that variables such as “a and i”, “a and g”, “b and h”, “b and d”, “c and I”, “c and f”, “d and h”, “d and f”, and “g and i” have a positive correlation.

Conclusion for descriptive analysis that there are some formulation not using all additive. We can see from the minimum value of some additive is 0 such as additive c, f, h, and i.

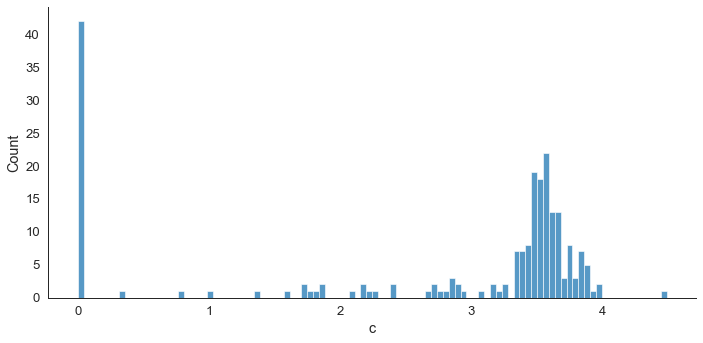
1. **A graphical analysis of the additives, including a distribution study**

Graphical analysis of the additives is a frequency of data distribution from each additive. The graph of each additive is shown below:

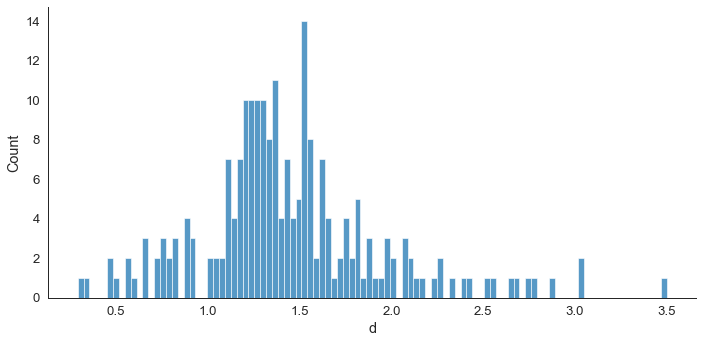
1. Distribution frequency of additive a
2. Distribution frequency of additive b



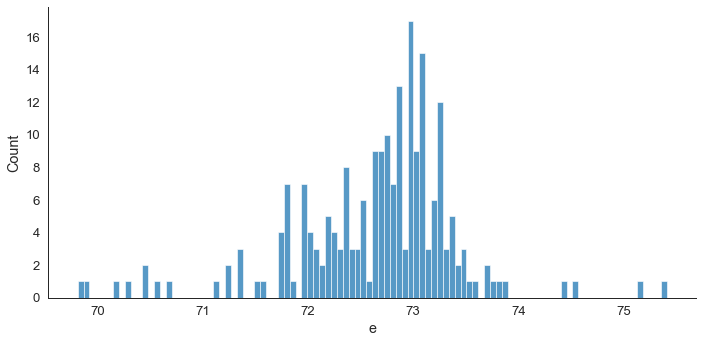
1. Distribution frequency of additive c



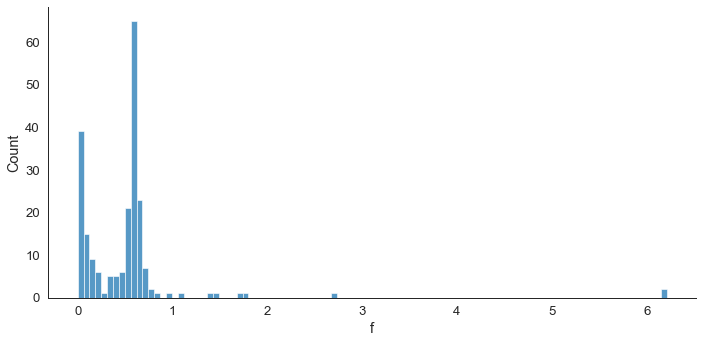
1. Distribution frequency of additive d



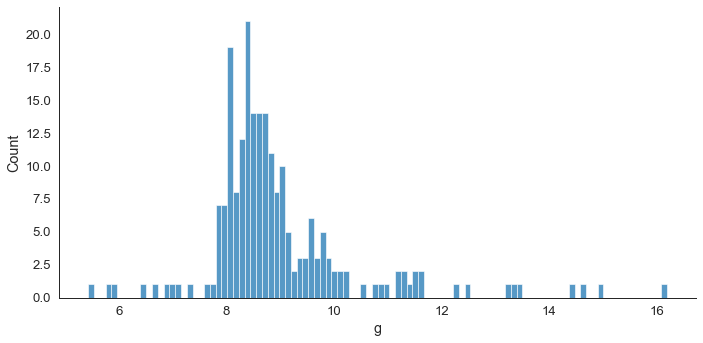
1. Distribution frequency of additive e



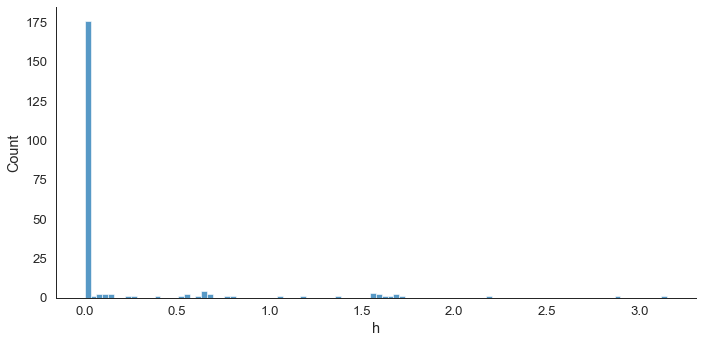
1. Distribution frequency of additive f



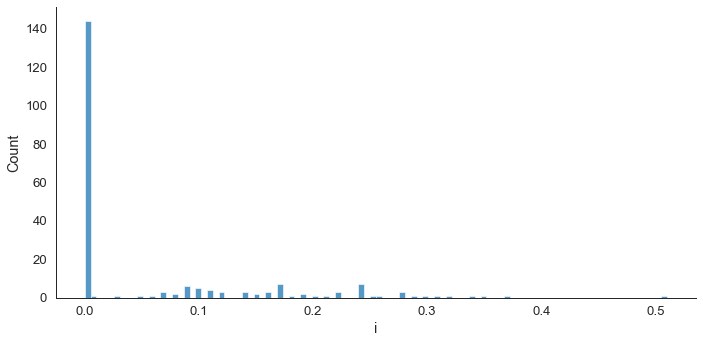
1. Distribution frequency of additive g



1. Distribution frequency of additive h



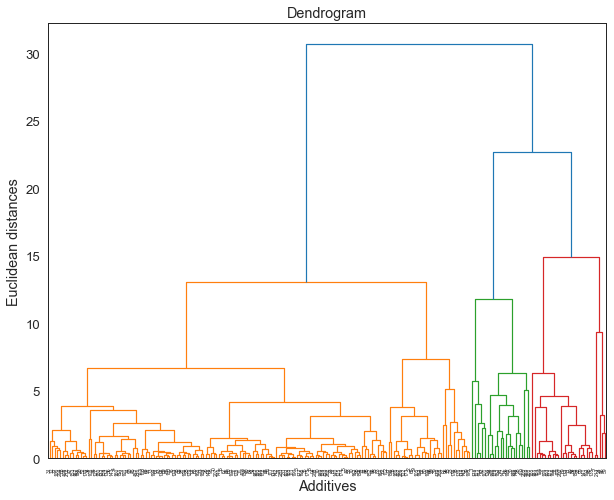
1. Distribution frequency of additive i



From the graph above, we can see that for additive c, f, h, i, the rows of additive it’s so many with 0 value

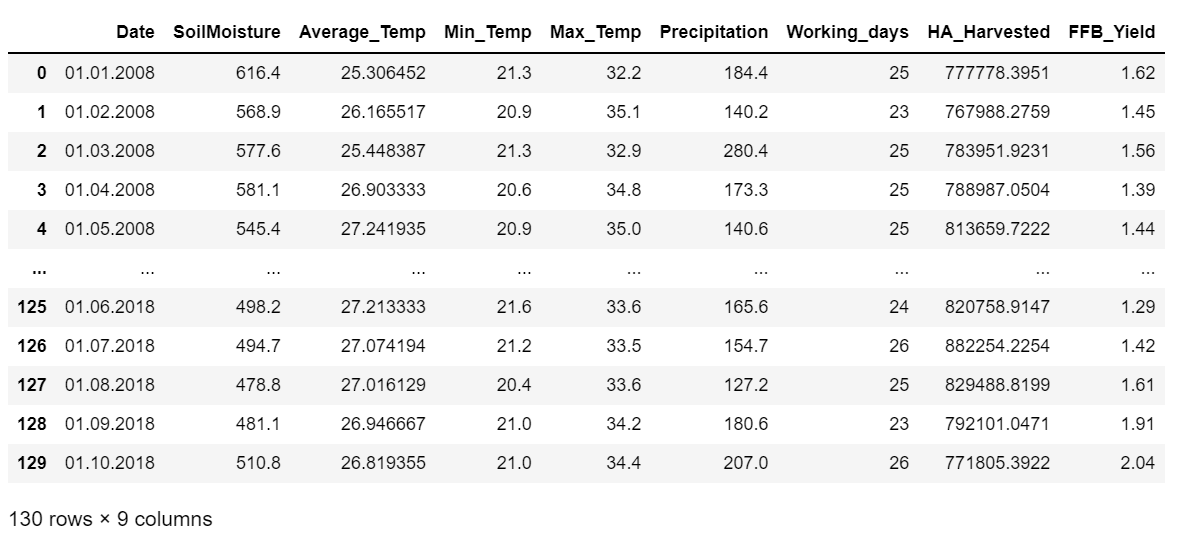
1. **A clustering test of your choice (unsupervised learning), to determine the distinctive number of formulations present in the dataset.**

To define number of cluster the data, I using hierarchial clustering to find number of cluster with using dendrogram. From this image, we can see that we got 3 cluster of the data. The graph is shown below:



1. **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.**

**Explanation:**

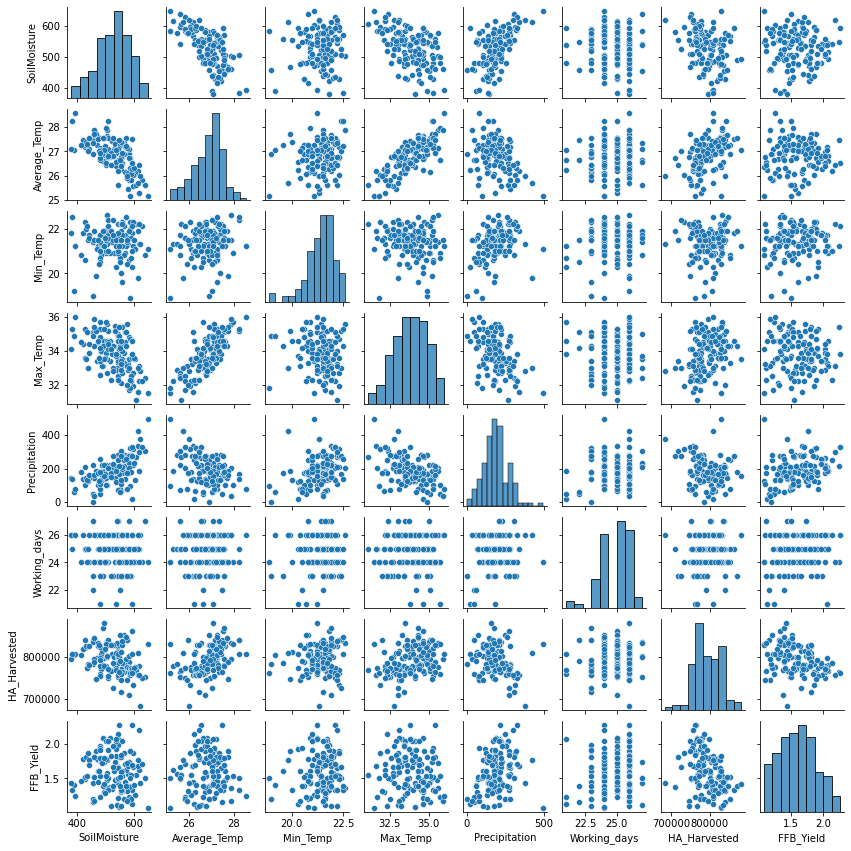
This is dataset of this case:

This dataset has some variables such as date, soil moisture, average temperature, minimum temperature, maximum temperature, precipitation, working days, HA harvested, and FFB yield. To analyze how external factors influence FFB (Fresh Fruit Bunch) yield, we can determine what variables influence FFB. For this case, FFB yield as a dependent variable of 6 other independent variables such as soil moisture, average temperature, minimum temperature, maximum temperature, precipitation, working days, HA harvested. Why do we not use date for an independent variable? Because the date is not an external factor of influence FFB yield, the date is only shown as a time series for this data.

There are some techniques to determine what variables are important to influence FFB yield. For this case, I used two techniques for solving. Such as:

1. ***Linearity assumption***

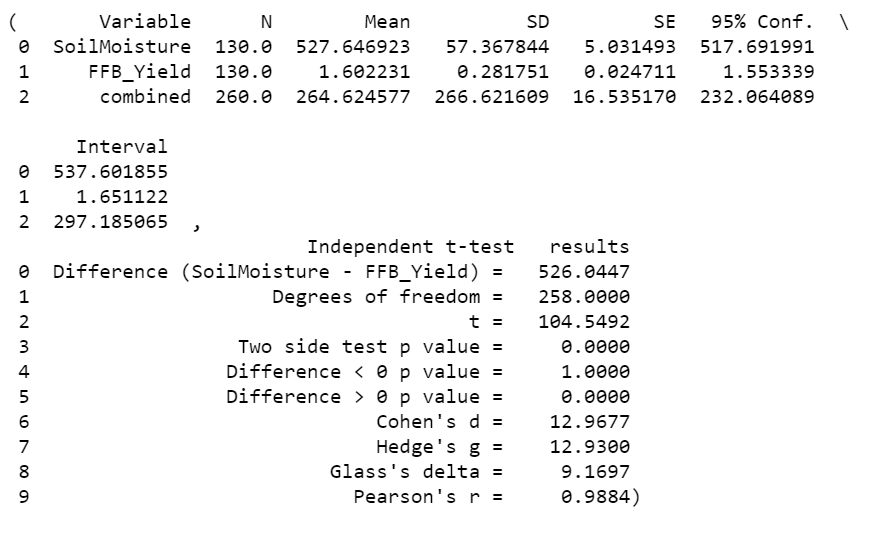
The linearity assumption is a graph to show what variables are linear to dependent variables. The variable is most important to influence independent with a positive linear. The linearity assumption is shown below:

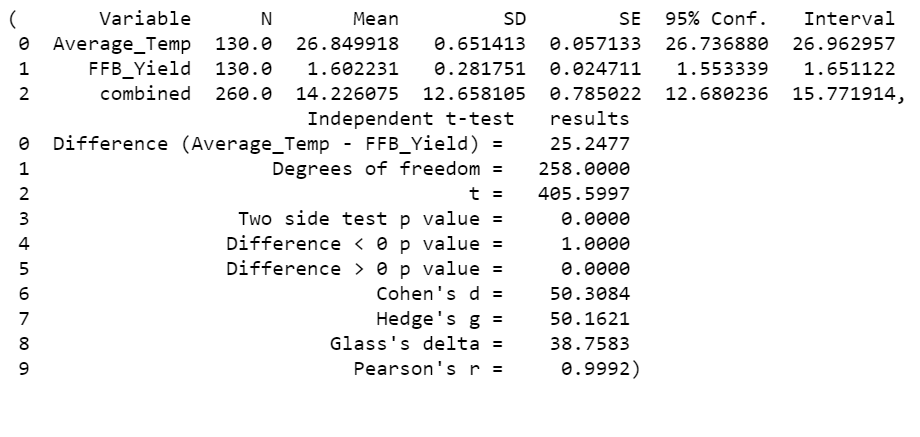


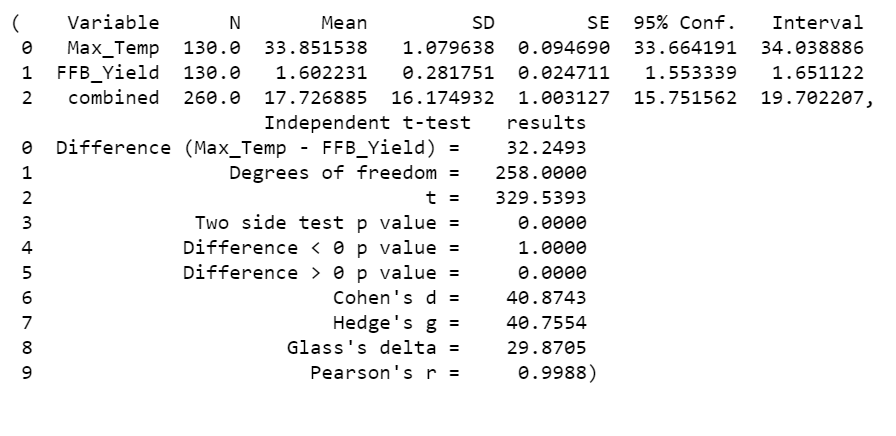
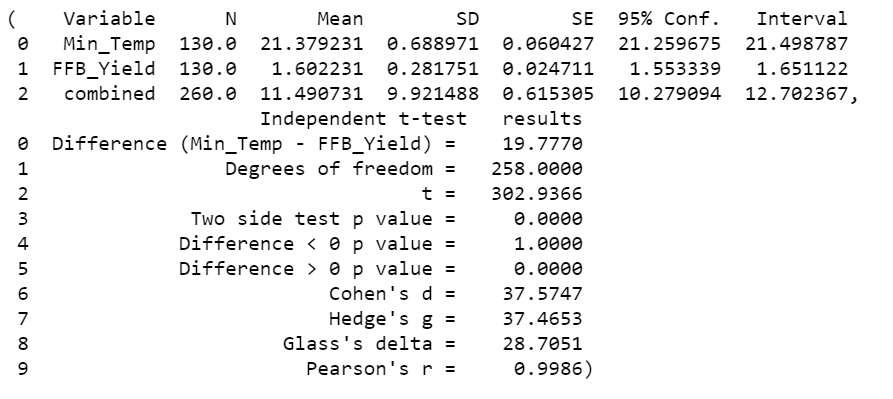
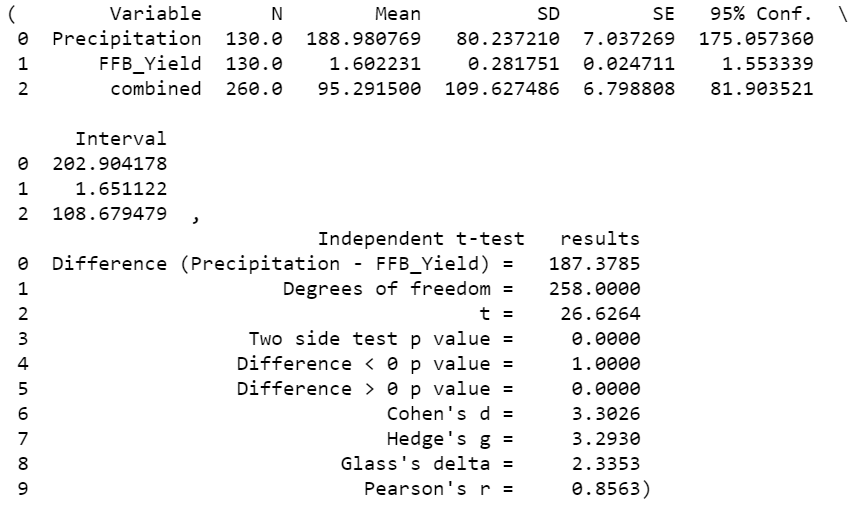
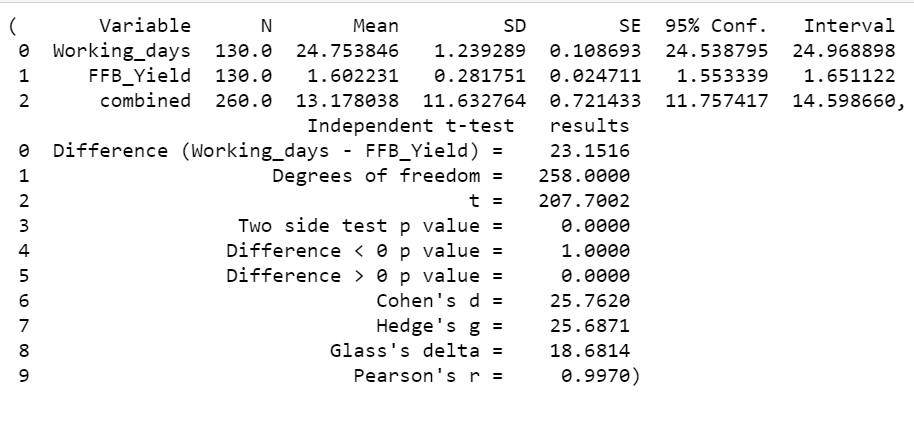
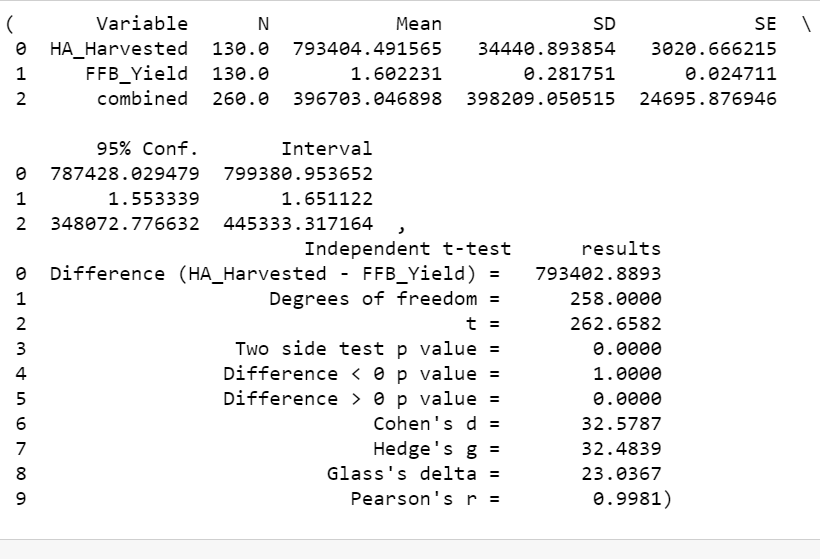
1. **T- test**

T-test is used for to known what the independet variables (X) such as soil moisture, average temperature, minimum temperature, maximum temperature, precipitation, working days, and HA harvested has a significant effect for dependent variable (Y), is FFB yield.

Based on result of t-test to t-table, for all independent variables is important to influence FFB yield because result of t-test is more higher than t-table. For see more result is shown below:

1. Soil moisture and FFB yield
2. Average temperature and FFB yield



1. Maximum temperature and FFB yield
2. Minimum temperature and FFB yield
3. Precipitation and FFB yield
4. Working days and FFB yield
5. HA harvested and FFB yield