**Aidil Fadilah**

[**aidilfadilah99@gmail.com**](mailto:aidilfadilah99@gmail.com)

**082384006087**

**INGREDIENT TEST**

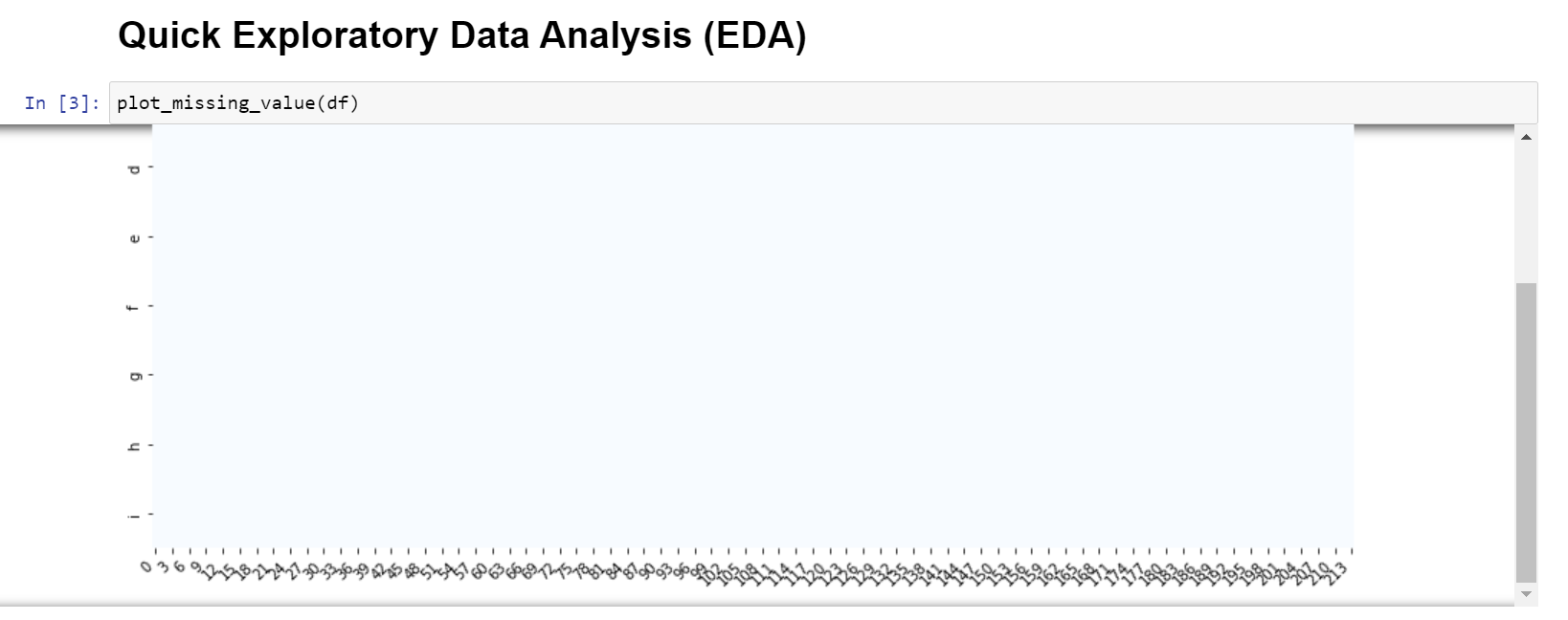
1. **A descriptive analysis of the additives (columns named as “a” to “i”), which must include summaries of findings (parametric/non-parametric). Correlation and ANOVA, if applicable, is a must. (1.a)**

The first step in analyzing a dataset is to check the contents of the dataset, which may contain empty data, noise, and other errors. This time I will use python with the help of jupyter notebook in analyzing the dataset which will be explained in the steps below.

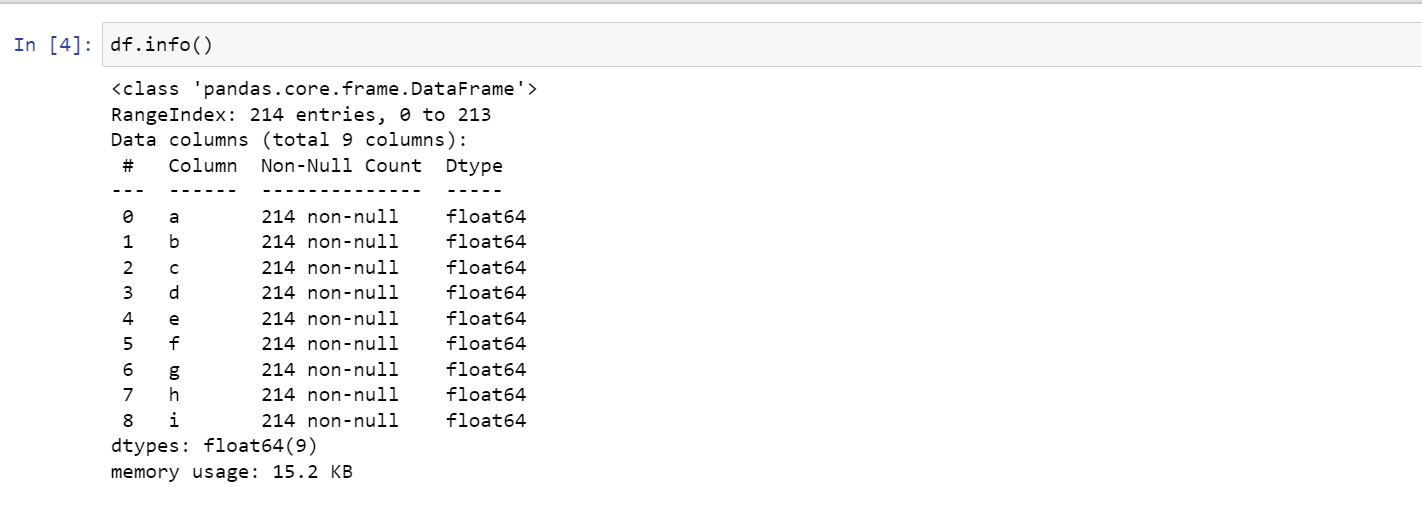
1. Load library and Read Dataset



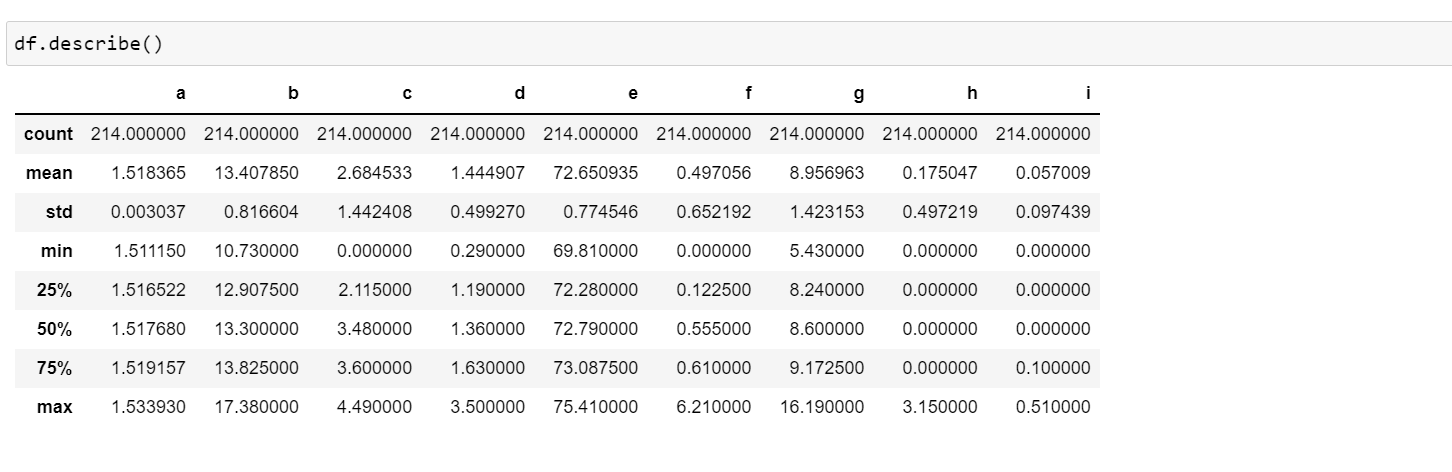
1. Quick Exploratory Data Analysis



1. Information Dataset



1. Describe Dataset



From the picture above, we can see that there are some addictive data that can be information in the dataset, namely:

1. Count

Information can be drawn that in the image there are columns or groups consisting of a - i and each row contains 214 rows of data.

1. Mean

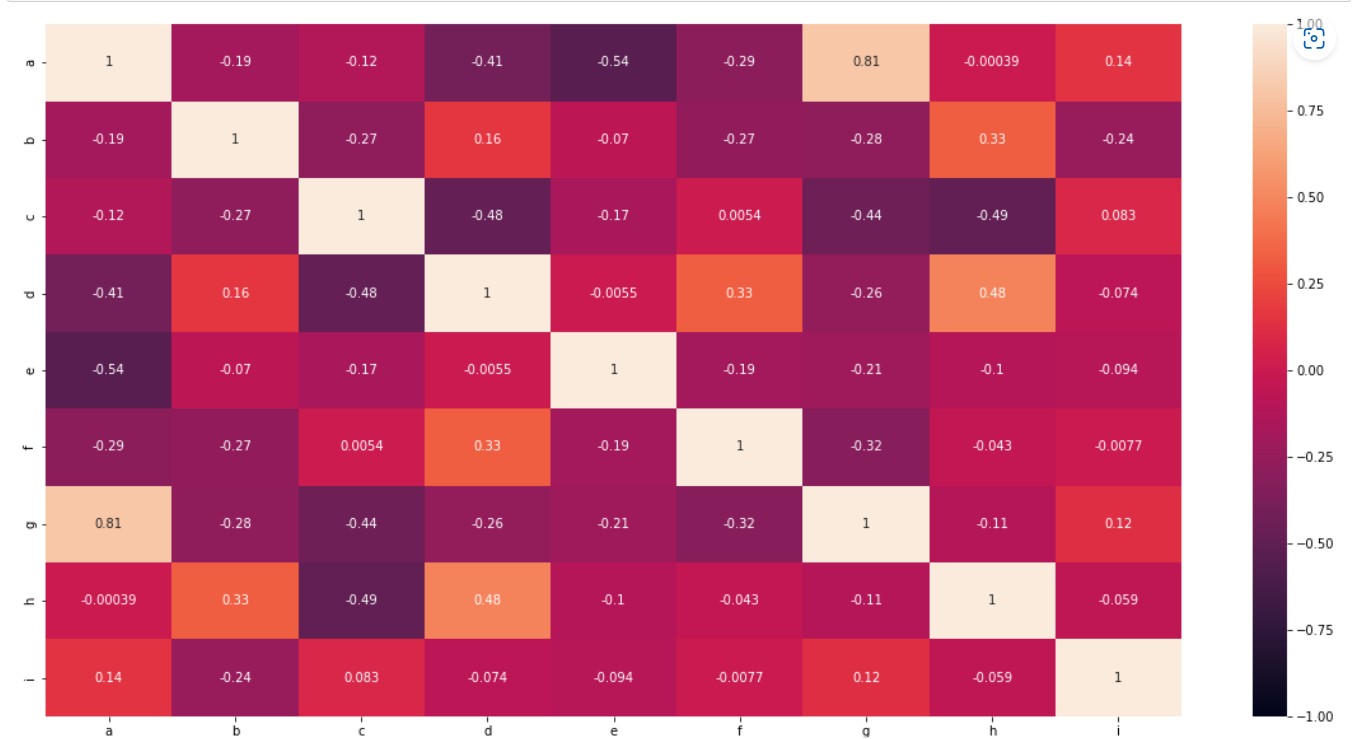
There is information on the average value of each variable a to i such as variable a = 1.518365 ... i = 0.057009.

1. Std

Also obtained information on the size of the average difference of each variable such as a = 0.003037... i = 0.097439

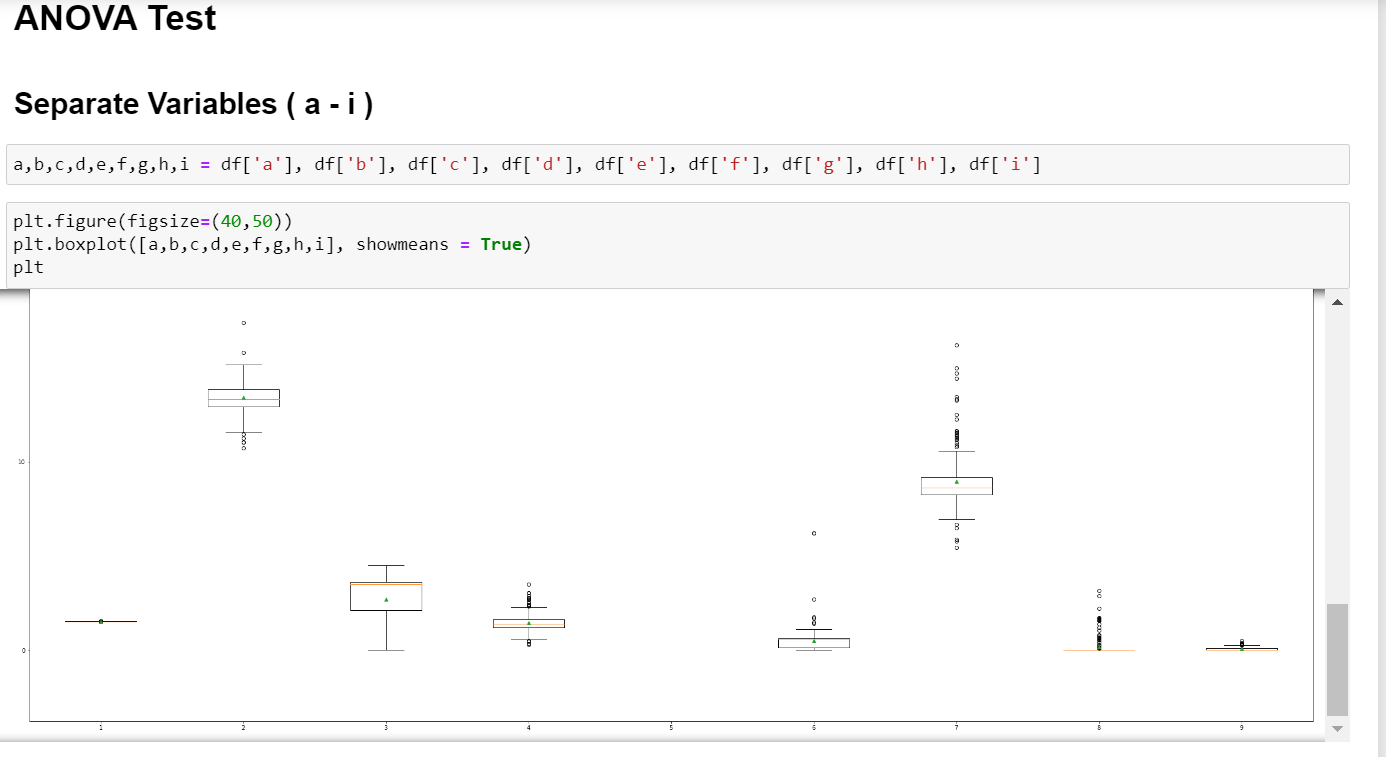
There is also information for the minimum, maximum, median values contained in the describe image. This image has a function to see what normal information we can see from the dataset.

1. Correlation



The picture above provides correlation information which can have a value ranging from -1 to 1 or the correlation value is strong to positive, it can be strong to negative. The image above uses a correlation matrix by using the corr() function which is already in the python library. Information from the matrix can be drawn that strong positive correlations such as a and g, a and i, b and d, b and h, c and f, c and i, d and f, d and h, g and i. But not all formulations use all variables depending on the value of the dependency.

1. ANOVA



In the picture above there is a data visualization using a boxplot by bringing up the average value, its function is to see whether each variable has the same or different average.



After visualizing the data, ANOVA testing will be carried out using the stats function from the Scipy library with statistical results and p value obtained. After that we make changes to the column into a categorical column so that the stats model can be used.

1. Result ANOVA

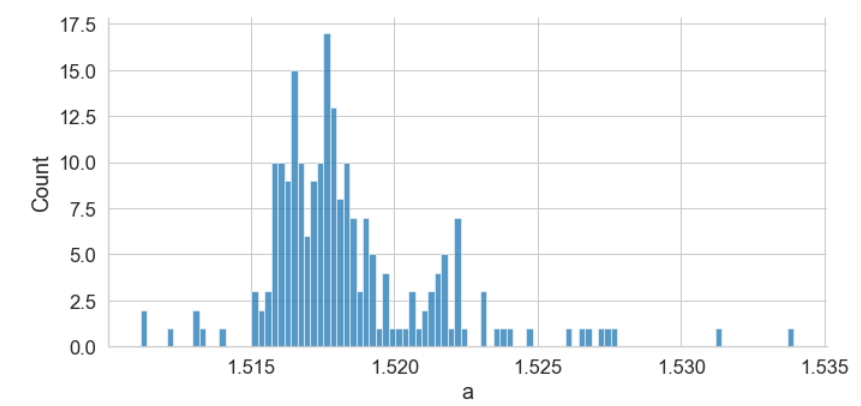


The results of the ANOVA test can be seen in the image above, there is information on the value of p value, df residual (within), de model (between), std, t value, and others. From this information, it can be seen whether the influence of each variable is significant or not.

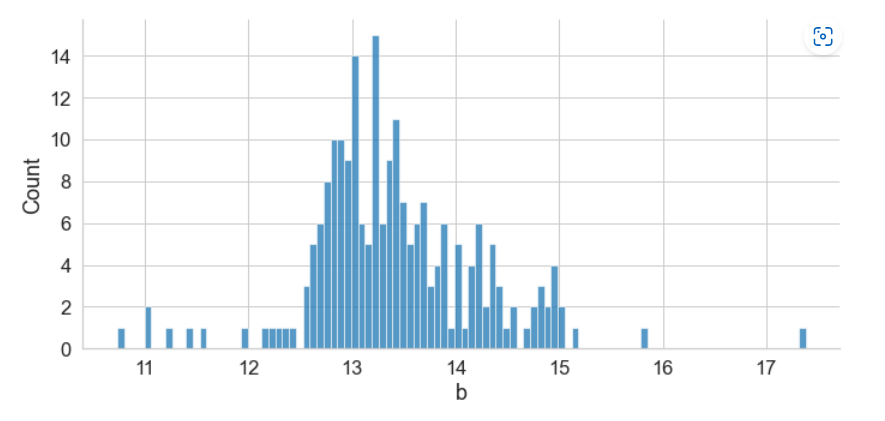
1. **A graphical analysis of the additives, including a distribution study. (1.b)**

Graphs of additives can be seen in the pictures below by showing the data distribution of each addictive.

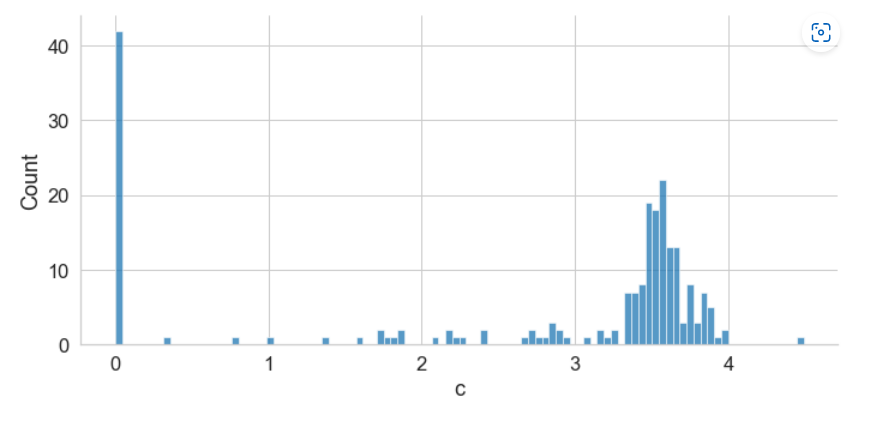
1. Distribution Addictive a



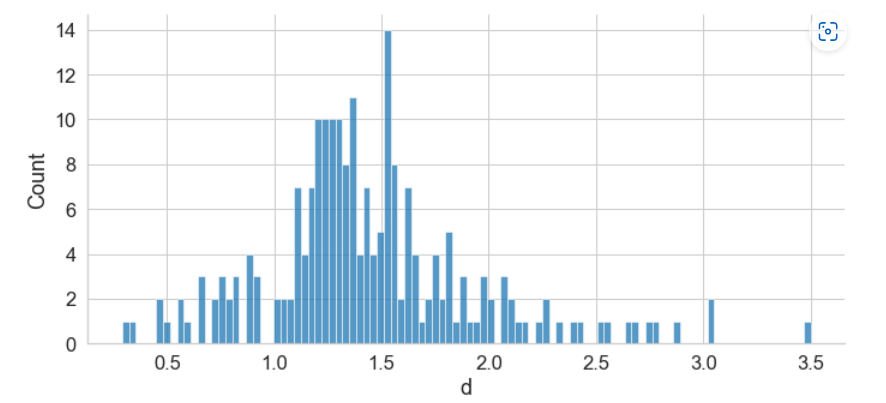
1. Distribution Addictive b



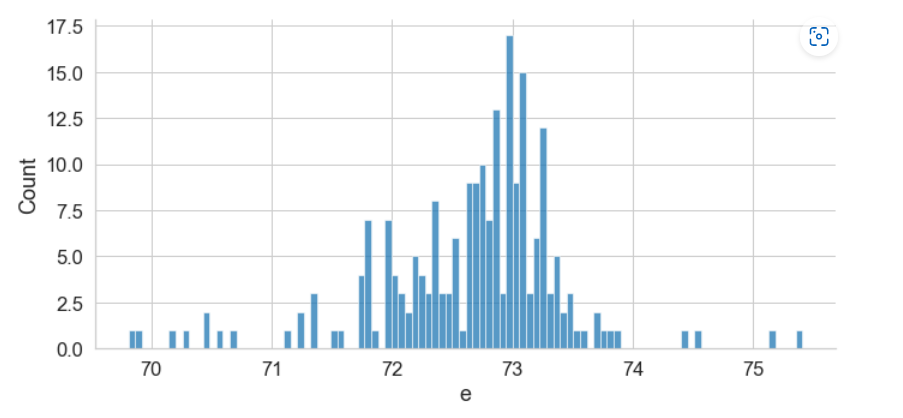
1. Distribution Addictive c



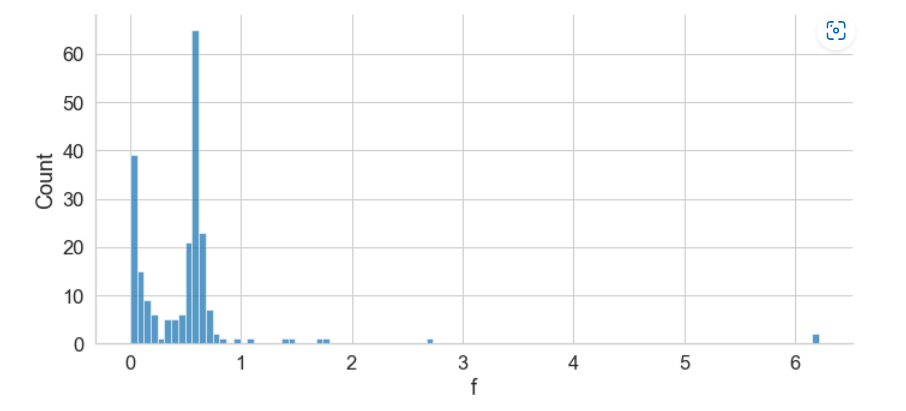
1. Distribution Addictive d



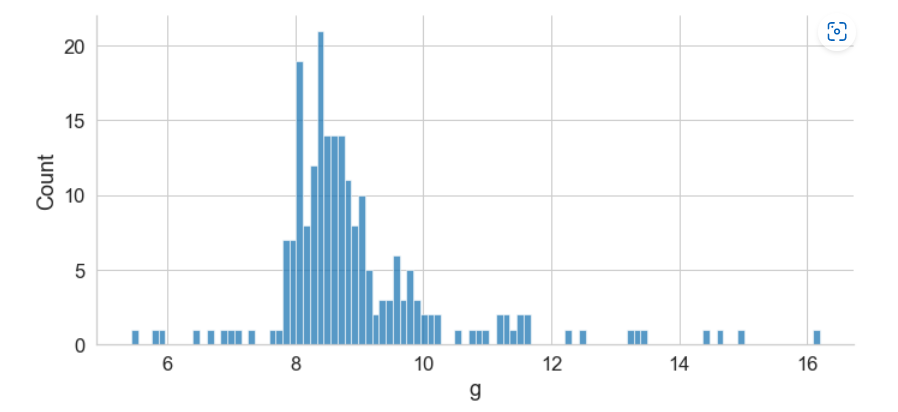
1. Distribution Addictive e



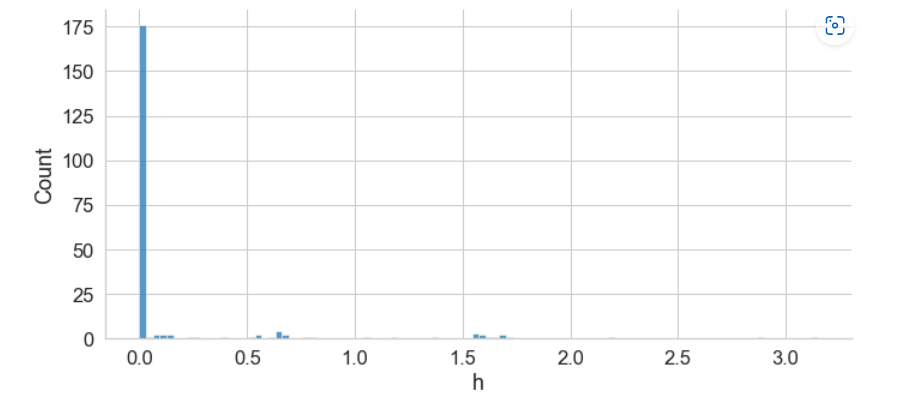
1. Distribution Addictive f



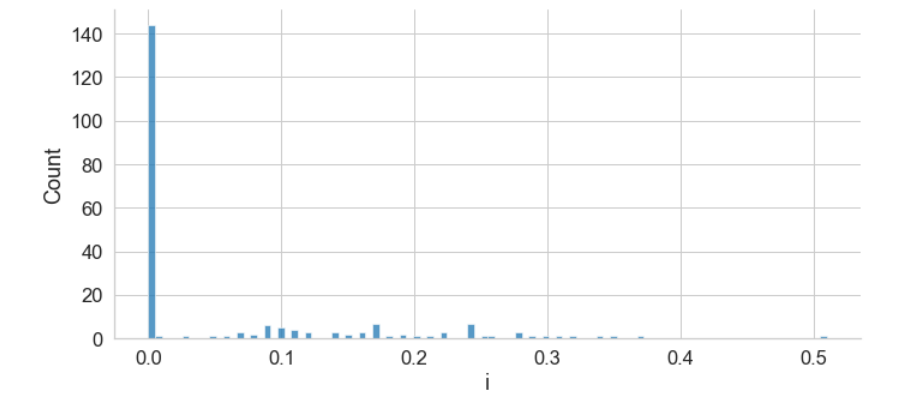
1. Distribution Addictive g



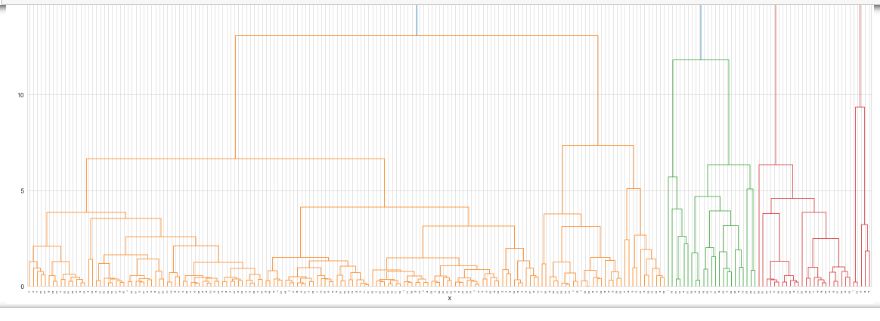
1. Distribution Addictive h



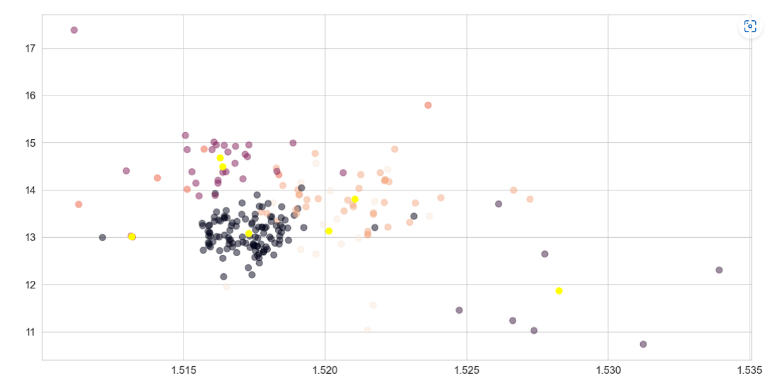
1. Distribution Addictive i



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

****

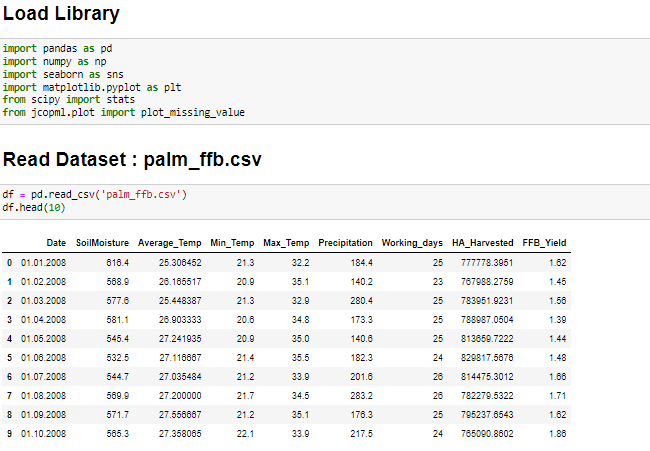
From the picture above, there are 3 clusters

****

Using k-means

**PALM\_FFB TEST**

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.**
2. Load Library and Read Dataset.

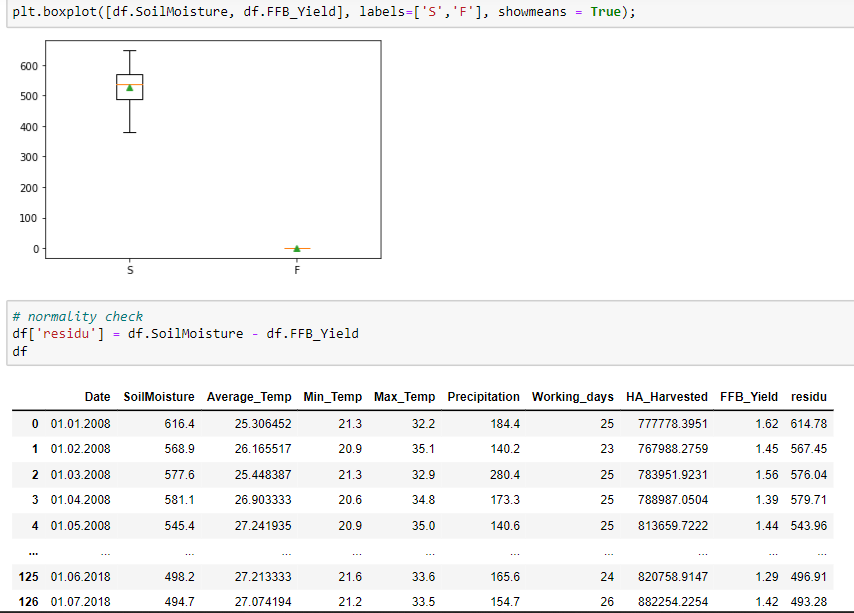


1. The next step is the same as the step in question number 1 so it will not be detailed again until the correlation.
2. Linear

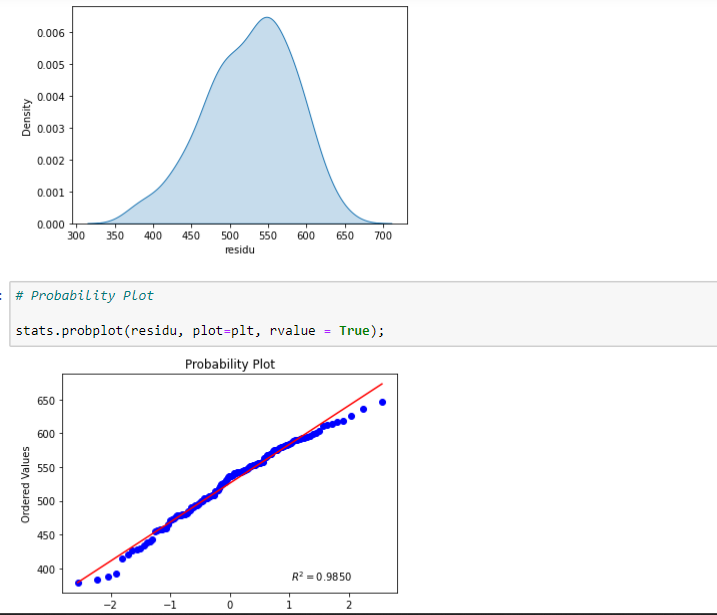
Then we can use plots to display graphs that can provide information on variables that are linear to the dependent. But it can be said that the assumptions of the variables can be seen in the picture below.

1. T – Test

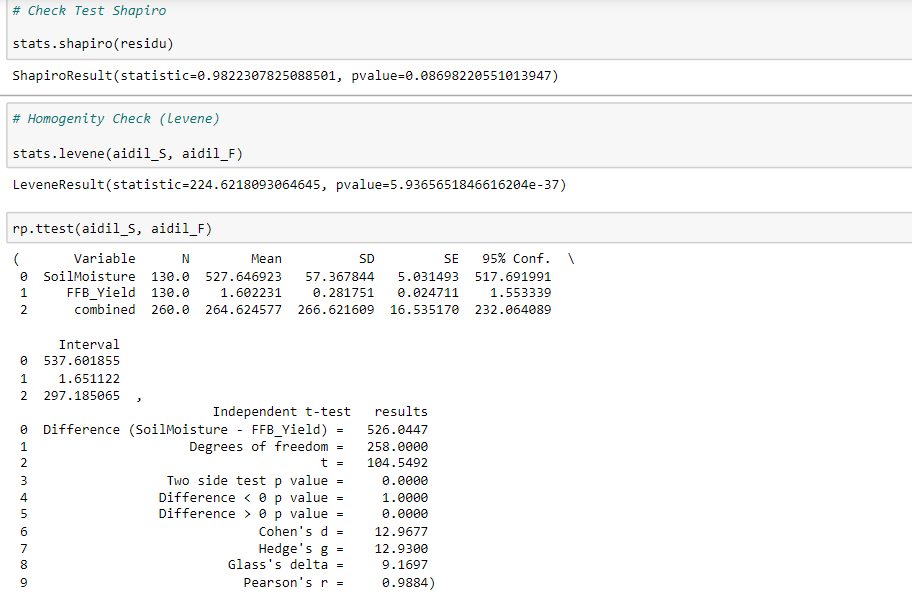
This step is to test from the FFB\_Yield column in other columns, to see the information that will be obtained, you can use t - test. Before doing the t-test, it is necessary to know that there are several stages to get to the t-test, such as checking that the dataset should not be too large, checking kernel density (plot like a bell), checking probability, checking Shapiro, and doing a homogeneous check. So if you want to do a t-test, these conditions must be met in order to be able to do it. below will show the image.



To see the boxplot and residuals of the variables to be t-tested.

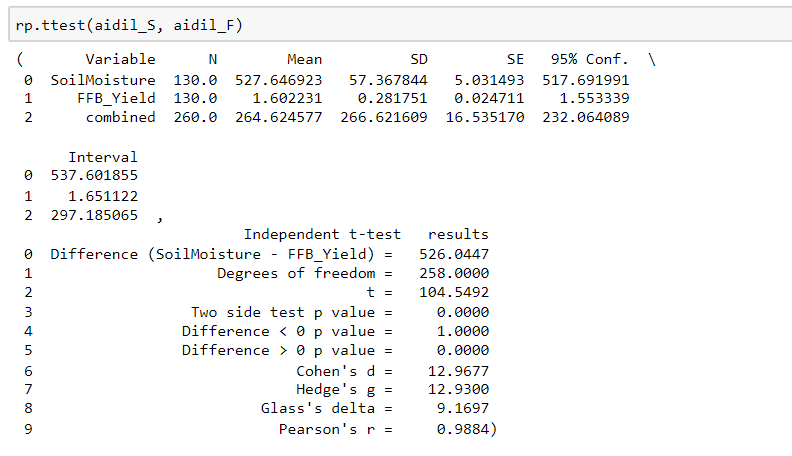


To see the kernel density and probability where the data has a value of 0.9850 from the red line means the data is very close to 1.

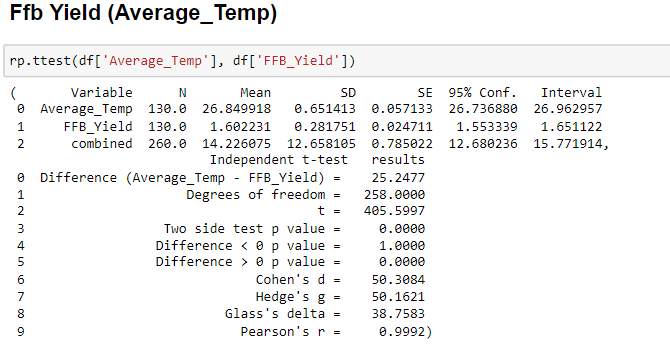


Check Shapiro and homogeneity and perform a t-test where it can be seen that there is some information obtained, such as N (number of data), Mean (average), SD (standard deviation), SE (standard error) , and more. The t-test was carried out to see whether FFB had an effect on other variables but the date variable was not used because this variable is generally a data series so it is only complementary information.

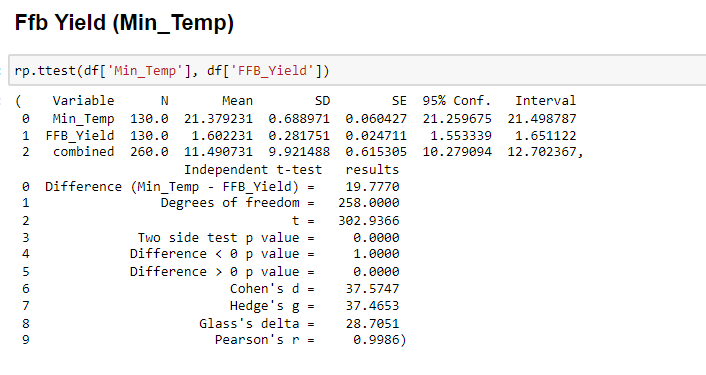
1. Soilmoisture



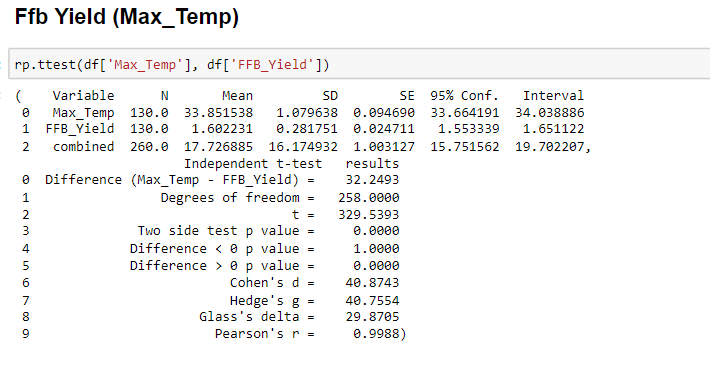
1. Average\_Temp



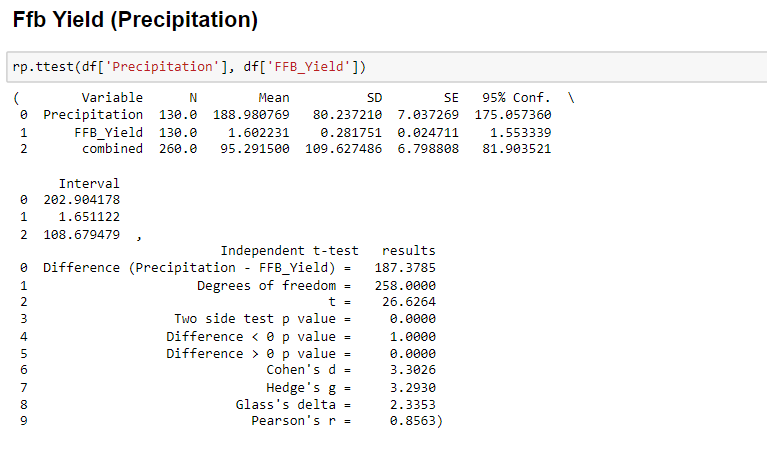
1. Min\_Temp



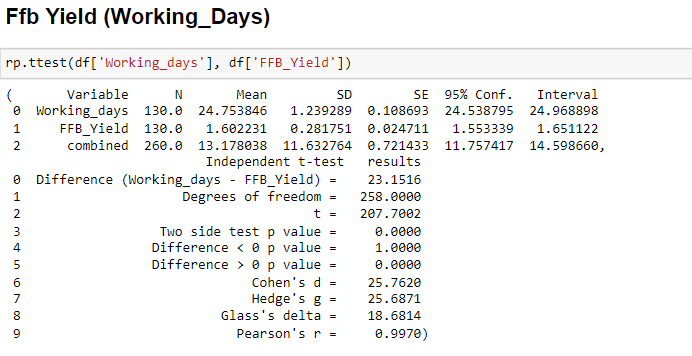
1. Max\_Temp



1. Precipitation



1. Working\_Days



1. HA\_Harvested

