Name : M Afif Rizky A

Email : afifrizky933@gmail.com / 23521034@std.stei.itb.ac.id

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 addingvarying 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. Sincethe formulations are confidential information, they are not named in the dataset. Please assist the consultant in the area of statistical analysis by doing this

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, isa must

Answer:

Descriptive Analysis is the type of analysis of data that helps describe, show or summarize data points in a constructive way such that patterns might emerge that fulfill every condition of the data. For this analysis, I will use a few method:

- Univariate Analysis about how to describe and summarize a single variable such as:
  - Mean : average of all the items / variables in a dataset
  - Max : Maximum Value of all the items / variable
  - Min : minimum value of all the items / variable
  - Median : middle element of a sorted dataset. The dataset can be sorted in increasing or decreasing order
  - Variance : numerically how far the data points are from the mean. if variance is greater than mean, its mean the data point have more outlier and mean of the data is not trustable
  - Standard Deviation : numerically aims to find out how many values or the amount of data differ from the average or a measure of how far numbers lie from the average

    This table below show result of univariate analysis from the data:

	a	b	c	d	e	f	g	h	i
Mean	1.518	13.408	2.685	1.445	72.651	0.497	8.957	0.175	0.057
Max	1.534	17.38	4.49	3.5	75.41	6.21	16.19	3.15	0.51
Min	1.511	10.73	0	0.29	69.81	0	5.43	0	0
Median	1.518	13.3	3.48	1.36	72.79	0.555	8.6	0	0
Variance	0	0.6668	2.0805	0.2493	0.5999	0.4254	2.0254	0.2472	0.0095

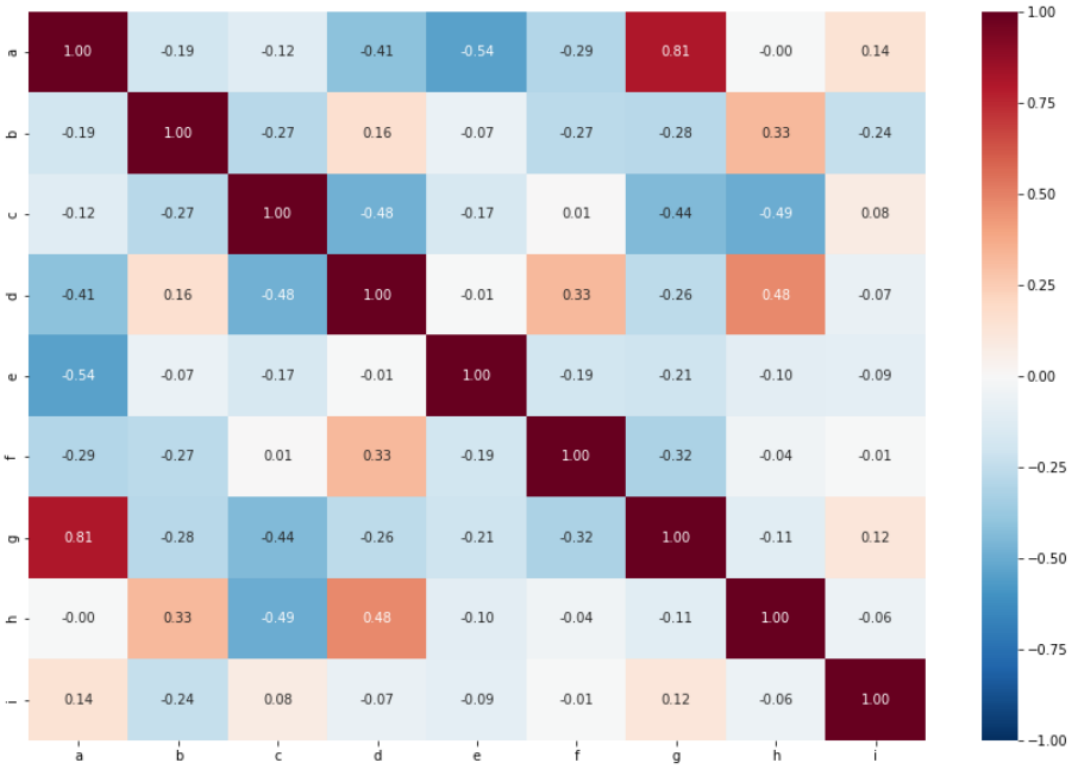
Standard Deviation	0.003	0.817	1.442	0.499	0.775	0.652	1.423	0.497	0.097
Count Values	214	214	214	214	214	214	214	214	214

Interpretation:

From table above, there are some insight that I can tell

1. The data contains no missing values
2. Variance and standard deviation of each data point is low, it's mean the data point is not to far from mean values, except column I and H, from this information, there is some outlier in this columns

- Multivariate Analysis ow to see statistical relationship among pair or multiple of variable
  - Parametric Testing : I will use Pearson Correlation to see relation ship between variables and Anova Testing to see population diffirence
    - Pearson Correlation



Interpretation:

From pearson table above, there are some insight:

1. There is at least one additive have slightly zero correlation with another additive
2. Additive I is the most non – correlation additive compared to another additive
3. Additive A is the most strong correlation additive compared to another additive

- Anova Testing

  ANOVA is a statistical test that assumes that the mean across 2 or more groups are equal

  H0 : All sample distributions are equal

  H1 : All sample distribution are not equal

  Code:

```
stat, p = f_oneway(data['a'], data['b'], data['c'], data['d'] , data['e'] , dat
a['f'], data['g'], data['h'], data['i'])
print('Statistics = %.3f, p = %.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
  print('Same distributions (fail to reject H0) P values is greater than Alpha'
)
else:
  print('Different distributions (reject H0) P values is less than Alpha')
```

  Result:

```
Statistics = 168331.969, p = 0.000
Different distributions (reject H0) P values is less than Alpha
```

- Non – parametric testing : I will use Mann Whitney U Test to see two independent variable is in have equal distribution or not

  - Mann Whitney U Test

    Mann-Whitney U test is a nonparametric statistical significance test for identifying whether two independent samples were drawn from a population with the same distribution

    H0 : All sample distributions are equal

    H1 : All sample distribution are not equal

    Code:

```
stat, p = mannwhitneyu(data['a'], data['g'])
print('Statistics = %.3f, p = %.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
  print('Same distributions (fail to reject H0) P values is greater than Alpha'
)
else:
```

```
print('Different distributions (reject H0) P values is less than Alpha')
```

Result:

```
Statistics = 0.000, p = 0.000
Different distributions (reject H0) P values is less than Alpha
```

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

Answer:

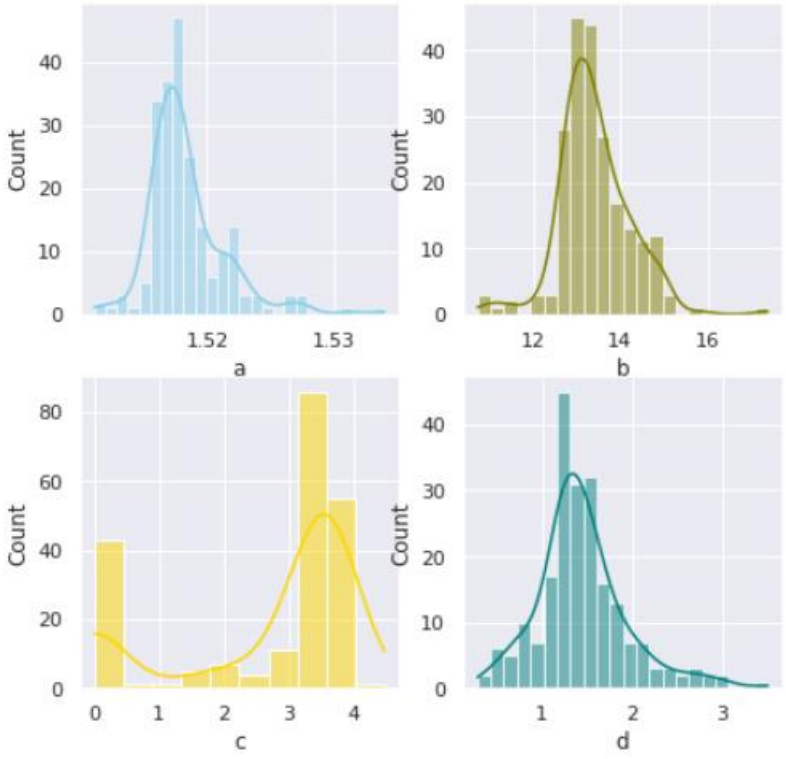
For graphical analysis and distribution study we use histogram and D'Agnostino's Method.

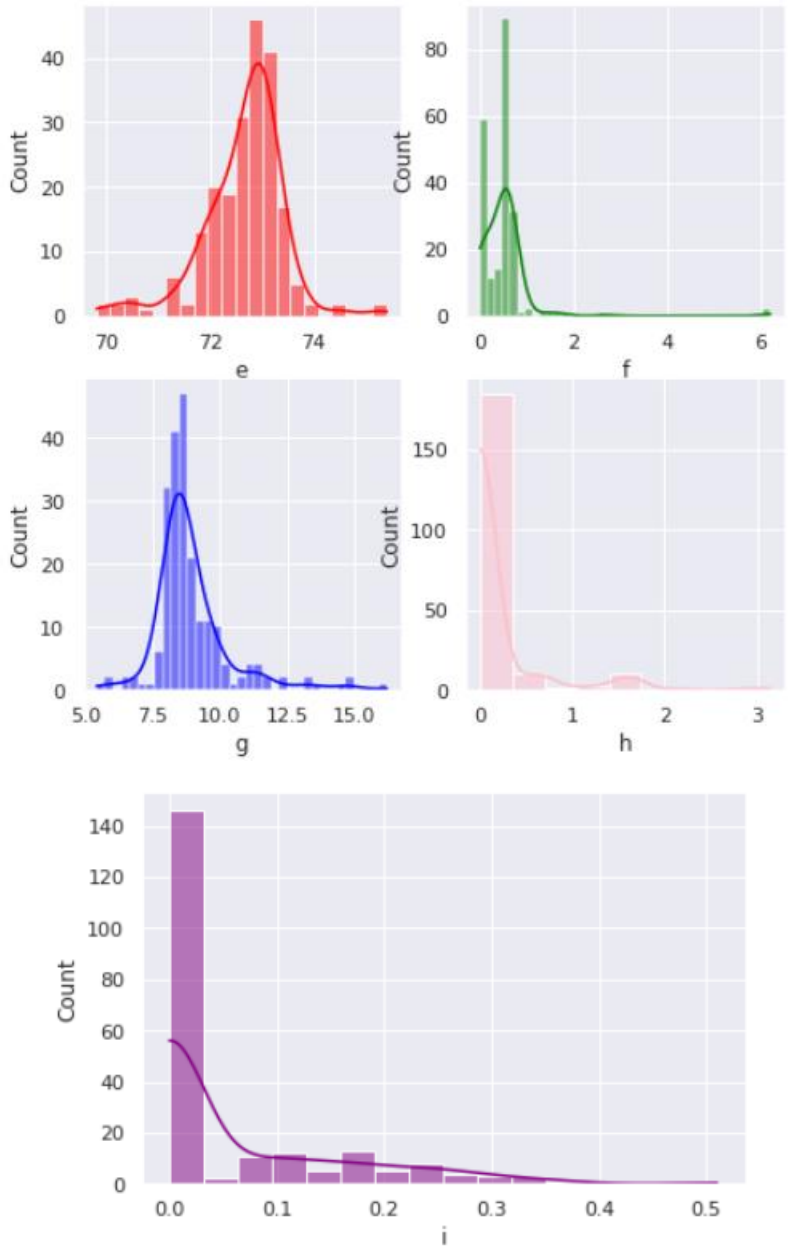
- Historgram Analysis

A histogram is an approximate representation of the distribution of numerical data

I will visualize every additive from the data to identify distribution waveform from the data

The result shown in figure below:





- D’Agnostino’s Method

  To interpret waveform from histogram, I will use D’Agnostino Method to indentify each column have gaussian / normal distribution or not

  Code:

```
stat, p = normaltest(data['a'])
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
  print('Sample looks Gaussian (fail to reject H0)')
```

```
else:
  print('Sample does not look Gaussian (reject H0)')
```

Results:

- Additive A

```
Statistics=84.358, p=0.000
Sample does not look Gaussian (reject H0)
```

- Additive B

```
Statistics=27.015, p=0.000
Sample does not look Gaussian (reject H0)
```

- Additive C

```
Statistics=35.885, p=0.000
Sample does not look Gaussian (reject H0)
```

- Additive D

```
Statistics=36.773, p=0.000
Sample does not look Gaussian (reject H0)
```

- Additive E

```
Statistics=35.873, p=0.000
Sample does not look Gaussian (reject H0)
```

- Additive F

```
Statistics=305.666, p=0.000
Sample does not look Gaussian (reject H0)
```

- Additive G

```
Statistics=109.473, p=0.000
Sample does not look Gaussian (reject H0)
```

- Additive H

```
Statistics=177.280, p=0.000
Sample does not look Gaussian (reject H0)
```

- Additive I

```
Statistics=76.867, p=0.000
Sample does not look Gaussian (reject H0)
```

Conclutions :

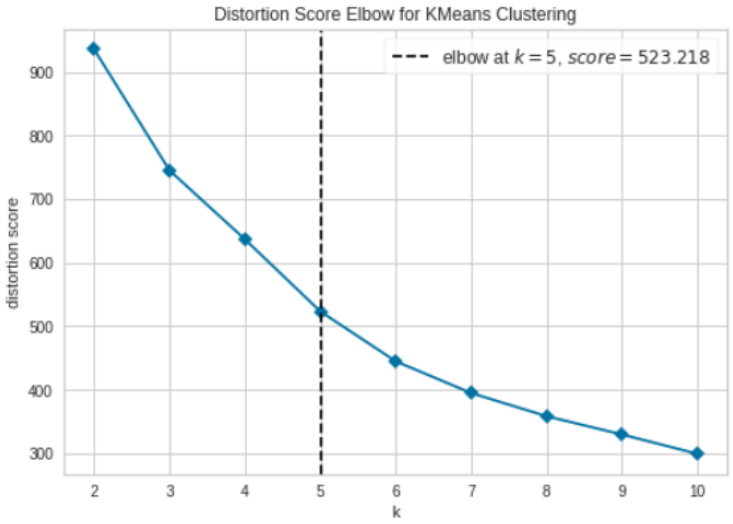
From Every waveform in histogram we have insight that data we use doesn't come from normal distribution but if we look back from costumer problem that we must prove that every additive in formulation is significantly different, the answer is yes. it's proved by correlation and ANOVA testing

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

Answer :

Based on K-Means Cluster, I Identify some insight such as :

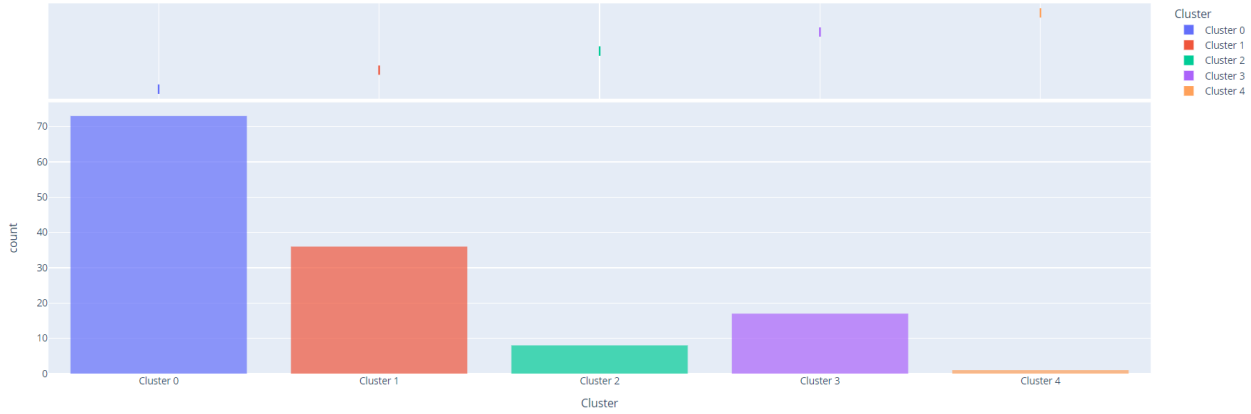
1. Optimum number of cluster is 5 based on elbow distortion score



2. Average Silhouette score based on K-Means cluster with 5 number of cluster is 0.31 which is every sample in this cluster is on or very close to decision boundaries

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.3158	42.9704	0.9738	0	0	0

3. First Cluster have distinctive number of formulations present in the dataset



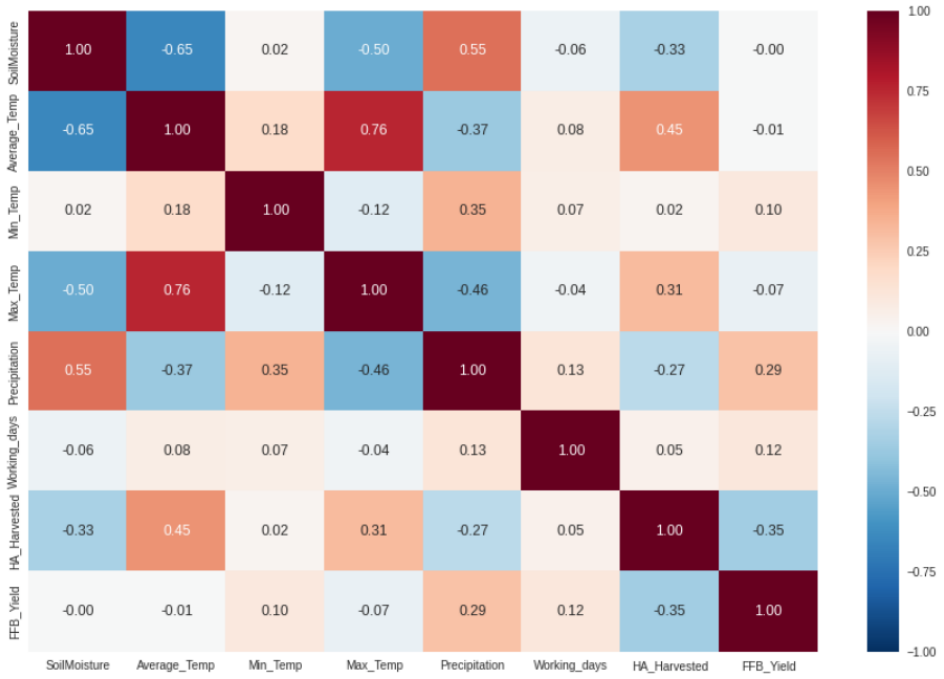
2. A team of plantation planners are concerned about the yield of oil palm trees, which seems to fluctuate. They have collected a set of data and needed help in analysing on how external factors influence fresh fruit bunch (FFB) yield. Some experts are of opinion that the flowering ofoil palm tree determines 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.

Answer:

There is some step to identify influencial feature to FFB yield, first by doing early statistics, such as

pearson correlation. Second, we can construct machine learning regression approach to identify what kind of feature that have influence to FFB yield. The result is shown below:

- Pearson Correlation



There are some insight that I can tell as a base knowladge to identify what is the influence feature to the FFB yield:

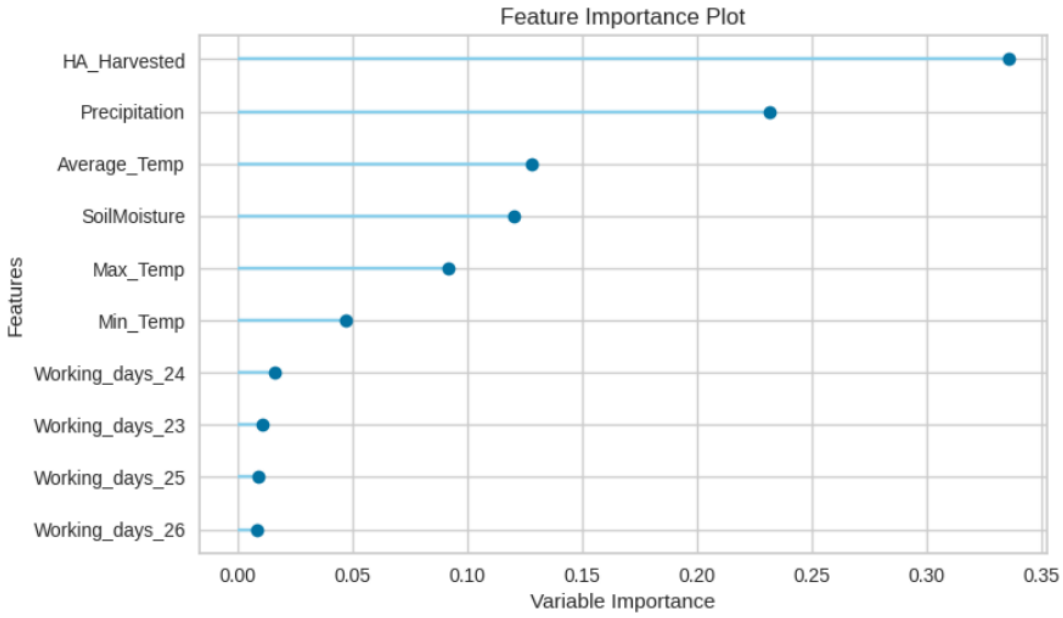
- Preciptation have positive correlation with FFB Yield (there is a chance that FFB Yield influenced by Preciptation)
- FFB yield is not correlated by temperature and soil moisture
- FFB yield slightly have positive correlation with working days (there is a chance that if we have more working days, more HA harvested, we got more FFB yield)

To prove this assumption we use machine learning regression approach and indentify what kind of external variable that influance FFB_yield. I will compare a few machine learning regression method to see what kind algorithm that give the best result

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	0.1861	0.0587	0.2358	0.1174	0.0914	0.1240	0.536
et	Extra Trees Regressor	0.1914	0.0591	0.2391	0.0994	0.0923	0.1260	0.377
ada	AdaBoost Regressor	0.1988	0.0652	0.2500	0.0167	0.0968	0.1331	0.075
lightgbm	Light Gradient Boosting Machine	0.2028	0.0650	0.2505	-0.0086	0.0967	0.1329	0.049
gbr	Gradient Boosting Regressor	0.1975	0.0647	0.2479	-0.0155	0.0964	0.1312	0.047
br	Bayesian Ridge	0.2207	0.0720	0.2634	-0.0252	0.1015	0.1459	0.025
knn	K Neighbors Regressor	0.2089	0.0715	0.2594	-0.0404	0.1015	0.1454	0.080
ridge	Ridge Regression	0.2189	0.0741	0.2655	-0.0748	0.1026	0.1458	0.025
lr	Linear Regression	0.2226	0.0765	0.2702	-0.1128	0.1051	0.1489	0.526
lar	Least Angle Regression	0.2232	0.0767	0.2705	-0.1138	0.1052	0.1494	0.032
huber	Huber Regressor	0.2234	0.0773	0.2720	-0.1249	0.1057	0.1501	0.047
lasso	Lasso Regression	0.2510	0.0876	0.2904	-0.1947	0.1125	0.1656	0.023
en	Elastic Net	0.2510	0.0876	0.2904	-0.1947	0.1125	0.1656	0.042
llar	Lasso Least Angle Regression	0.2510	0.0876	0.2904	-0.1947	0.1125	0.1656	0.027
dummy	Dummy Regressor	0.2510	0.0876	0.2904	-0.1947	0.1125	0.1656	0.010
omp	Orthogonal Matching Pursuit	0.2413	0.0888	0.2904	-0.2002	0.1123	0.1614	0.028
dt	Decision Tree Regressor	0.2320	0.0987	0.3006	-0.5115	0.1150	0.1521	0.026
par	Passive Aggressive Regressor	0.2850	0.1288	0.3537	-1.0466	0.1399	0.1929	0.030

From the table shown above, we get Random Forest reggresion model as the best model compared to other algorithm, this algorithm is best to handle noisy data because some of our independent variable from the data got abnormal distribution based on normality test

After construct machine learning model, the result of Features importance are show below:



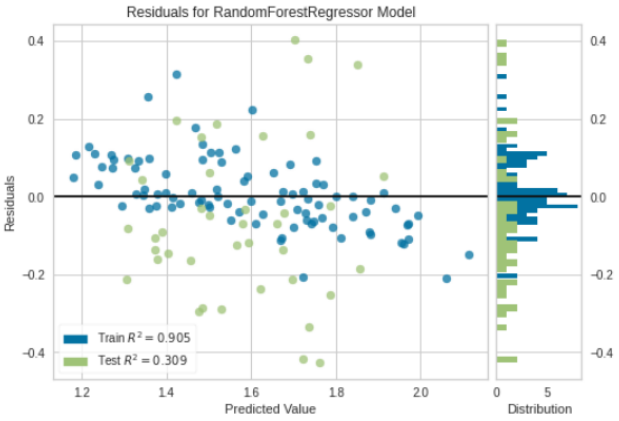
Conclution :

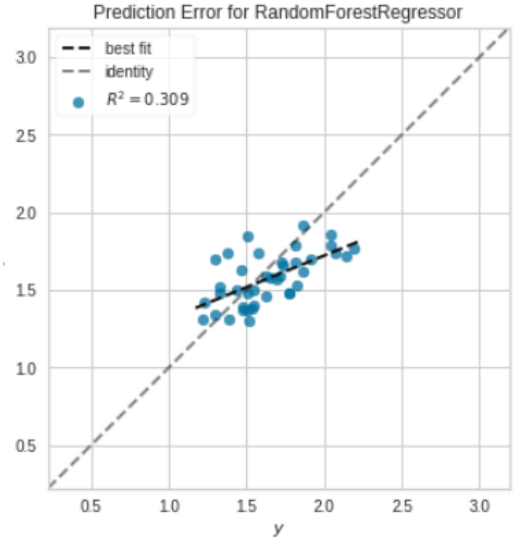
If we look at main feature that influance prediction value from the model are show below :

1. HA Harvested (this is correct, more HA Harvested is mode FFB_Yield)
2. Precipitation
3. Average temperature, Soil moisture, max and min temp (this is a contrary to the correlation table, if we look back to correlation table, which this feature have nearly zero correlation, but in fact these features enter into influential features)
4. FFB yield is not influenced by working days

Model Evaluation:

The regressor model still have to get major improvement because residual plot and error plot have major difference. This can happen because the sample data used is still quite small. For future improvement, more data is needed to get more robust model and give accurate result





3. Feed the following paragraph into your favourite data analytics tool, and answer the following:
    a. What is the probability of the word “data” occurring in each line ?
    Answer : `0.05625`
    b. What is the distribution of distinct word counts across all the lines ?

    Answer : `200`
    c. What is the probability of the word “analytics” occurring after the word “data” ?

    Answer : `0.011804384485666104`

Paragraph:

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. 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. 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. 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. 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. In addition, it can come from a mix of internal systems and external data sources. At a high level, data analytics methodologies include exploratory data analysis (EDA), which aims to findpatterns 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. 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. Tukey in his 1977 book Exploratory 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, themes and points of view