**Question 2**

**Objective:**

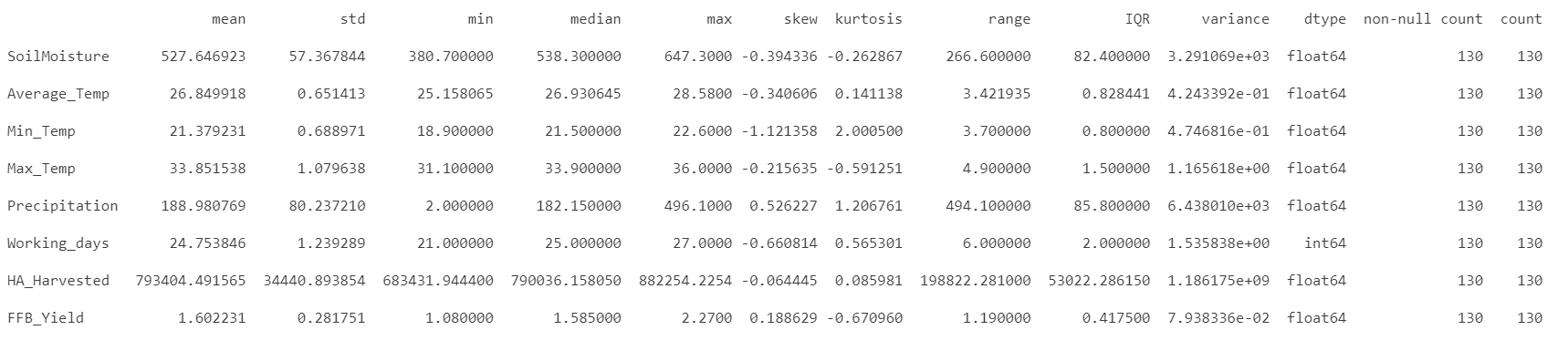
The primary objective of this analysis is to assess how specific external and operational factors affect the yield of oil palm trees, specifically the Fresh Fruit Bunch (FFB) yield. The data provided includes several variables that could influence yield, such as soil moisture, temperature, precipitation, and operational metrics like working days and harvested area. This study aims to:

1. Examine the influence of environmental factors (soil moisture, temperature, and precipitation) on the FFB yield. This involves determining the correlation and potential causation between these factors and the yield outcomes.
2. Analyse the impact of operational factors such as the number of working days and the area harvested on the yield. This will help understand how management practices contribute to yield variability.
3. Develop predictive models to forecast FFB yield based on the identified significant factors. These models will help in planning and optimizing operations to increase yield efficiency.
4. **Understanding the data**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Date | The month and year of the data entry. |
| SoilMoisture | Moisture level of the soil. |
| Average\_Temp, Min\_Temp, Max\_Temp | Average, minimum, and maximum temperatures respectively. |
| Precipitation | Amount of rainfall. |
| Working\_days | Number of working days in the month. |
| HA\_Harvested | Area of the plantation from where the harvest was done. |
| FFB\_Yield | Yield of Fresh Fruit Bunches. |

1. **Descriptive Analysis**

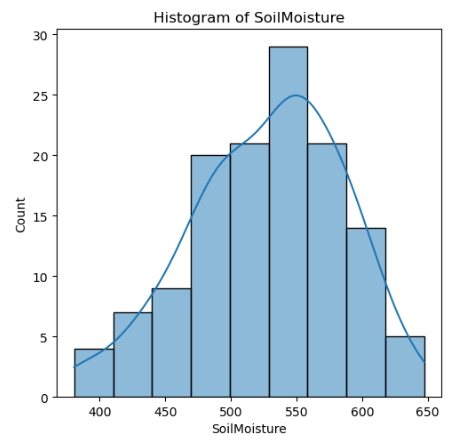
The dataset comprises several variables, each showing unique statistical characteristics, which are summarized as follows:



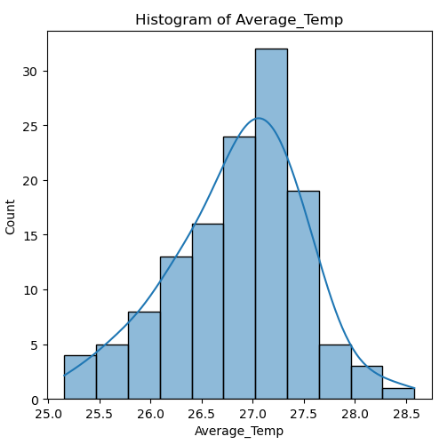
* **Skewness:**
  + All the variables show a mix of skewness direction, indicating diverse distribution shapes across different variable.
* **Variance and Standard Deviation:**
  + Soil Moisture and HA Harvested show high variance and standard deviation indicating significant spread in the data values around the mean.
  + Precipitation shows a very high standard deviation and variance, highlighting its significant variability.
* **Range and IQR:**
  + The range in Precipitation is extremely wide, going from 2 to 496, which emphasizes its variability across observations.
  + The IQR for variables like Soil Moisture and HA Harvested show notable spread, indicative of variability within the middle 50% of data points.
* **Overall Distribution:**
  + All variables are described for 130 non-null counts, suggesting complete data across all measured variables without missing values.
  + The data types are appropriate for the nature of each variable, with most being floating-point numbers, except Working Days, which is an integer.

1. **Histogram**

**3.1 Histogram of Soil Moisture**



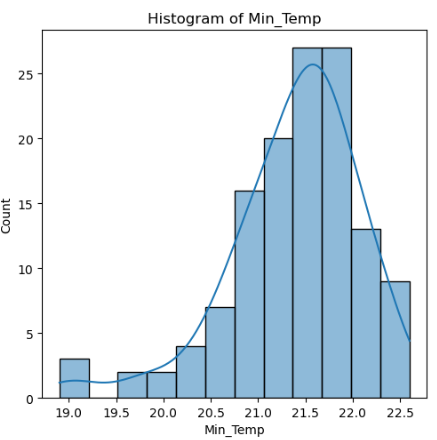
* **Histogram Shape**: Slightly skewed to the left, as indicated by a skewness value of -0.394, which suggests a tail with lower values extending to the left side of the distribution.
* **Distribution Characteristics**: The kurtosis value of -0.262 indicates that the distribution is relatively flat, lacking pronounced peaks or heavy tails compared to a normal distribution.
* **Central Tendency**: The median Soil Moisture value is around 538, situating near the center of the data's range, which aligns with the histogram's tallest bar.
* **Overall Observation**: The distribution of Soil Moisture is **mostly symmetrical around the median** but shows a **mild tendency for lower values** to extend beyond the general clustering of the data.
  1. **Histogram of Average Temp**



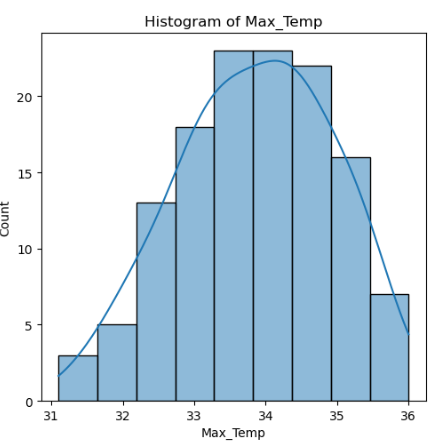
* **Histogram Shape**: The distribution appears mostly symmetrical around the central peak, which suggests a normal-like distribution. However, there is a slight skew to the left (-0.34), indicating a tail with a few more data points on the lower temperature side.
* **Distribution Characteristics:** The kurtosis value is 0.14, which suggests the peak of the distribution is slightly flatter than a normal distribution but not by a significant amount.

Overall, while not perfectly normal due to the slight skewness and mild deviation in kurtosis, the distribution of Average Temperature is **quite close to normal**. It shows characteristics mostly consistent with a normal distribution but with slight variations that prevent it from being perfectly normal.

* 1. **Histogram of Min Temp**



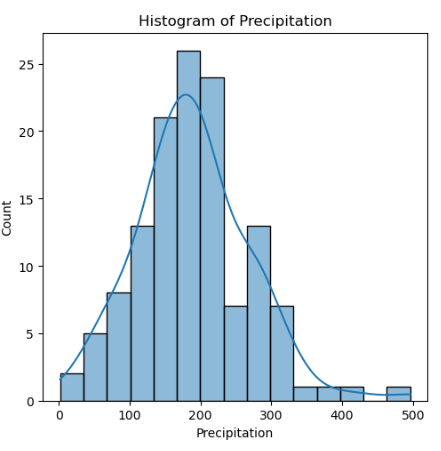
* **Histogram Shape**: Noticeably left-skewed (-1.12), with a significant tail extending toward lower temperatures.
* **Distribution Characteristics**: Kurtosis of 2, suggesting a peaked distribution with more values concentrated around the mean than expected in a normal distribution.
* **General Observation**: The distribution is concentrated with a **tendency toward higher values** and a pronounced tail in the lower range, indicating outliers or extreme low temperatures.
  1. **Histogram of Max\_Temp**



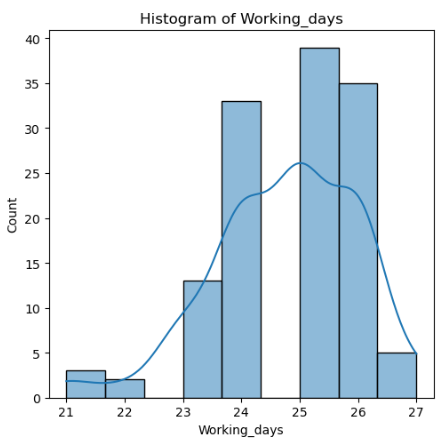
* **Histogram Shape:** Mildly left-skewed (-0.21), with a slight tail extending toward lower temperatures.
* **Distribution Characteristics**: The distribution appears slightly flat-topped, which corresponds to the kurtosis value of -0.59. This indicates the distribution is broader and flatter compared to a normal distribution, which would have more data points closer to the mean.
* **General Observation**: While the histogram does show a central peak that is characteristic of a normal distribution, the skewness and kurtosis values indicate that it deviates slightly towards lower temperatures and has a flatter peak than a normal distribution.

The Maximum Temperature distribution is **broadly normal but with a mild left skew and flatter peak**, suggesting a wider range of maximum temperatures than would be expected in a perfectly normal distribution.

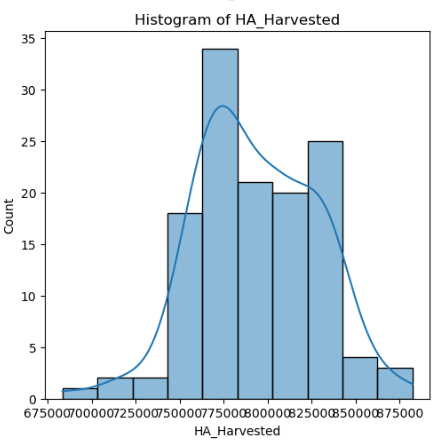
* 1. **Histogram of Precipitation**



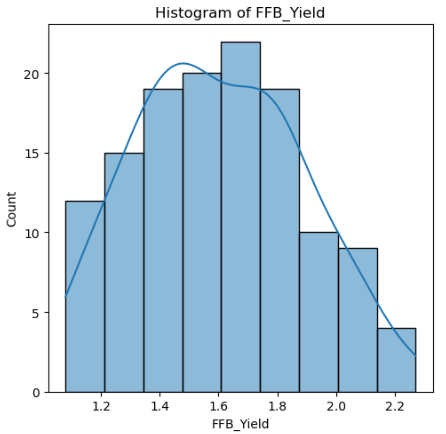
* **Histogram Shape**: Positively skewed (0.52), indicating more frequent lower values with a tail extending towards higher values.
* **Peak**: Kurtosis of 1.20, suggesting a more peaked distribution with more d**ata concentrated around the mean** than in a normal distribution.
* **Observation**: The data cluster around lower rainfall values, with fewer instances of very high rainfall.
* **Distribution Type**: **Non-normal distribution** due to positive skewness and higher kurtosis.
  1. **Histogram of Working\_days**



* **Histogram Shape**: Negatively skewed (-0.66), with data clustering towards the upper range and **a tail towards fewer days**.
* **Peak**: Kurtosis of 0.56, modestly sharper peak than normal.
* **Observation**: Most data points are concentrated around higher working days, with fewer instances of lower working days.
* **Distribution Type**: **Non-normal distribution** due to negative skewness and slight leptokurtosis.
  1. **Histogram of HA\_Harvested**



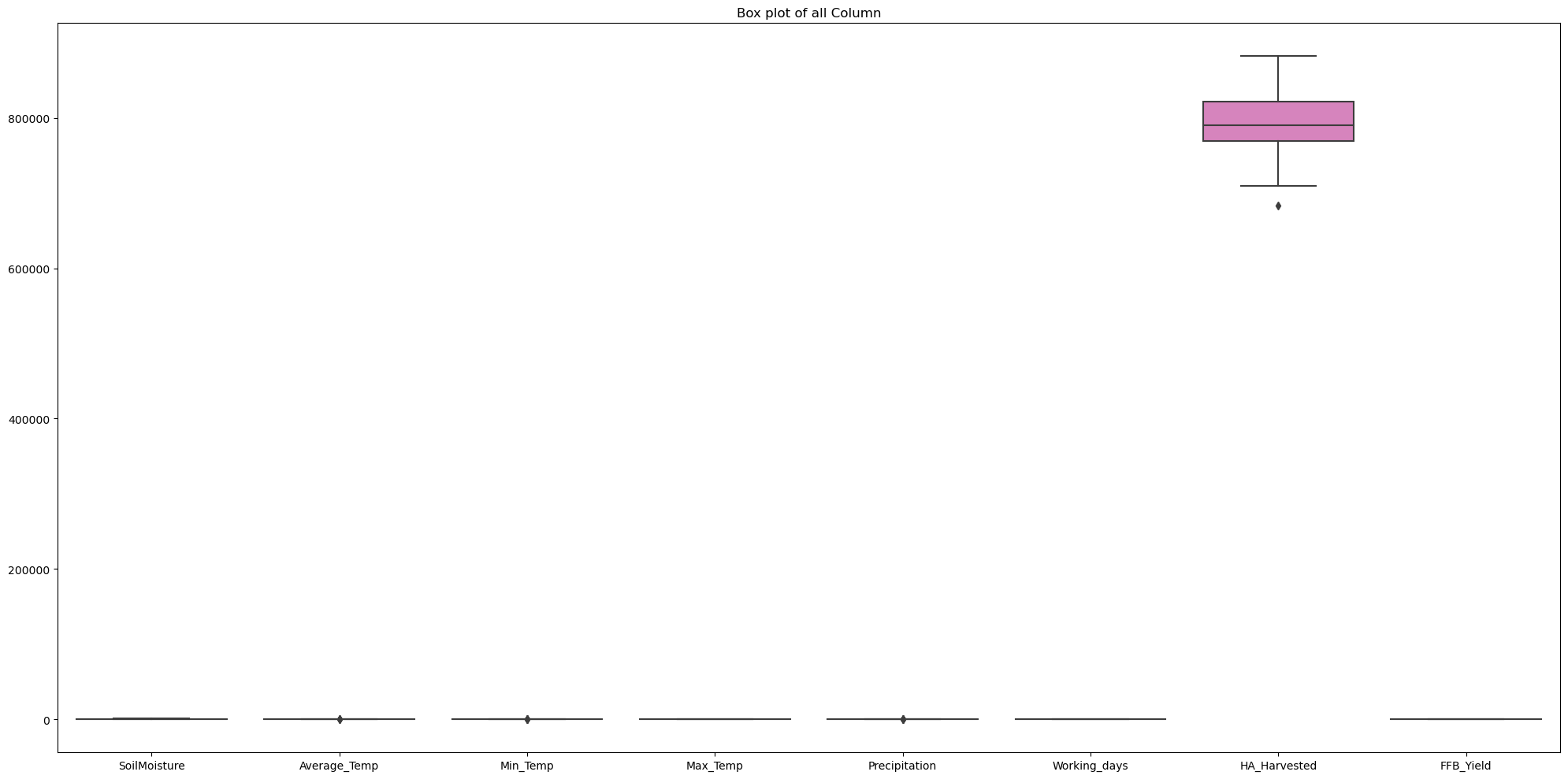
* **Histogram Shape**: Very slight negative skew (-0.06445), almost symmetric.
* **Peak**: Near normal kurtosis (0.0859), closely resembling a normal distribution.
* **Observation**: Fairly balanced distribution, with a mild tendency for more data points at lower values.
* **Distribution Type**: **Approximately normal distribution**, as both skewness and kurtosis are close to those of a normal distribution.
  1. **Histogram of FFB\_Yield**



* **Shape**: The histogram displays a slight positive skew (skewness = 0.188), suggesting that while most of the data are centered around the middle values, there are a few higher yield values that stretch the tail towards the right.
* **Peak**: Kurtosis of -0.67, which indicates a flatter peak than a normal distribution. This suggests that the data are more evenly spread across the range of yield values, rather than clustering tightly around the mean.
* **Overall Distribution**: The histogram shows that most FFB Yield values cluster around 1.6 to 1.8, with a gradual decline in frequency as yields increase beyond 1.8. The presence of a few higher values around 2.0 to 2.2 also highlights the variability in yield.
* **Distribution Type**: Non-normal distribution, primarily due to the slight positive skewness and the flatter peak, which deviates from the characteristics of a normal distribution.

**4.0 Exploratory Data Analysis**

**4.1 Box plot**

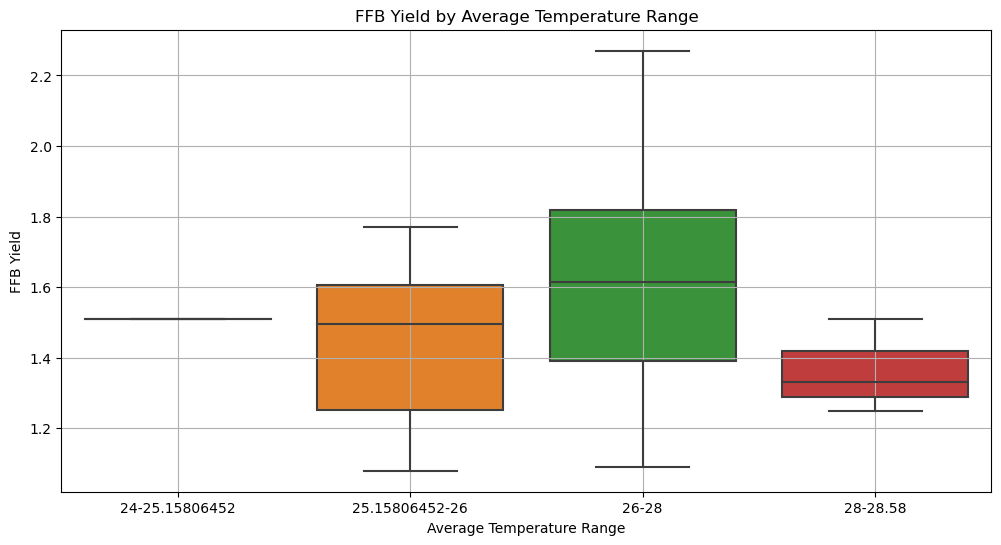


Temperature parameters (Soil Moisture, Average Temp, Min Temp, Max Temp) are represented as single points or very narrow boxes, indicating very low variability relative to the scale of the plot. This is due to their scale being significantly smaller compared to HA Harvested. The box plot highlights the challenge of visualizing data with vastly differing scales on a single plot. HA Harvested dominates due to its large scale, while other variables appear compressed.

**4.2 Box plot for all the variables**

|  |  |
| --- | --- |
|  | **Soil Moisture**: Shows moderate variability. The data is mostly clustered around the median. |
|  | **Average Temp**: Relatively stable with minimal variability and a couple of outliers indicating slightly higher temperatures. |
|  | **Min Temp**: With several outliers indicating much lower minimum temperatures. |
|  | **Max Temp**: Moderate variability with no outliers, data is fairly symmetrical around the median. |
|  | **Precipitation**: Shows considerable variability with several outliers on the higher end, indicating episodes of very high precipitation. |
|  | **Working Days**: Data is highly concentrated around the median with very little variability and no outlier. |
|  | **HA Harvested**: Also shows moderate variability with outliers indicating both significantly lower and higher values than most data. |
|  | **FFB Yield**: The data displays moderate variability around the median with a range of yields extending to both lower and higher values. |

**4.3 FFB Yield by Average Temperature Range**



The box plot shows the FFB yield across different average temperature ranges.

* **25.15°C Range (Orange):** The yield varies between approximately 1.4 to 1.8, with a median near 1.6.
* **26-28°C Range (Green):** The yield shows a slightly higher and more compact range, mostly between 1.5 and 1.7, with a median close to 1.6.
* **28-28.58°C Range (Red):** The yield decreases and shows less variation compared to other temperature ranges, mostly between 1.3 and 1.5, with a median around 1.4.

This suggests that there might be an **optimal temperature range (25.15-26°C) for higher FFB yield**, and as the temperature increases beyond this range, the yield tends to decrease.

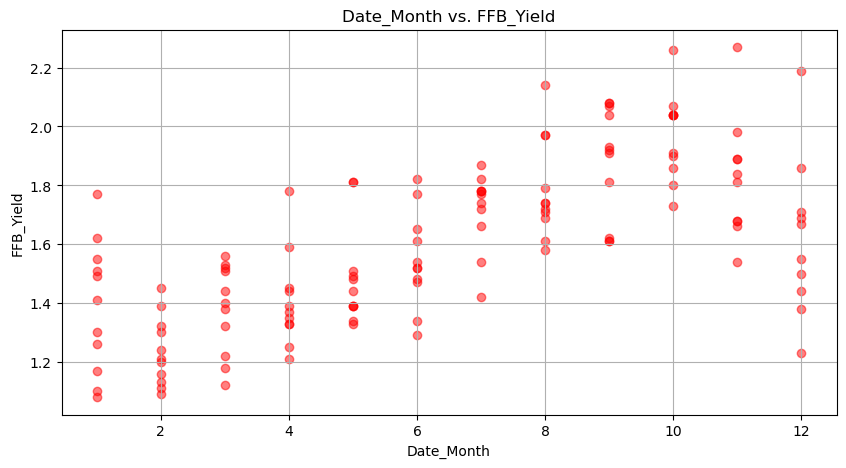
**4.4 Scatter Plot (Target variable vs Independent variable)**

|  |  |
| --- | --- |
|  | **FFB Yield vs. Soil Moisture:**  There is **no clear trend** indicating a consistent relationship between soil moisture levels and FFB yield. The data points are widely scattered across the range of soil moisture values. |
|  | **FFB Yield vs. Average Temperature:**  The plot shows a moderate scatter **without a clear linear relationship** between average temperature and FFB yield. Higher yields are not consistently associated with specific average temperatures. |
|  | **FFB Yield vs. Minimum Temperature:**  Similar to average temperature, there is **no evident pattern** that links minimum temperature with FFB yield. The data points are distributed broadly across the temperature range. |
|  | **FFB Yield vs. Maximum Temperature:**  This plot also **does not demonstrate a clear correlation** between maximum temperature and FFB yield. Yield fluctuates across the range of maximum temperatures. |
|  | **FFB Yield vs. Precipitation:**  Yields appear to be somewhat higher in the middle range of precipitation but decrease slightly at very low and very high precipitation levels. However, the **relationship is not strongly defined.** |
|  | **FFB Yield vs. Working Days:**  Yield is relatively stable across different numbers of working days, suggesting that the number of **working days might not significantly influence yield**, as the variation in yield is minimal regardless of the number of working days. |
|  | **FFB Yield vs. HA Harvested:**  The plot **does not show a clear trend** between the area harvested and FFB yield. While there are a few higher yields at mid-level areas harvested, the **overall distribution is quite scattered**. |

**4.5 Scatter plot (Temperature vs Precipitation)**

|  |  |
| --- | --- |
|  | **Min\_Temp vs. Precipitation:**  - This plot **does not show a clear linear relationship** between the minimum temperature and precipitation.  - The data points are quite spread out, indicating high variability in precipitation for similar values of minimum temperature.  - There is no apparent trend that higher or lower minimum temperatures are associated with more or less precipitation. |
|  | **Max\_Temp vs. Precipitation:**  - Similar to the Min\_Temp plot, this plot also **lacks a clear linear relationship**.  - The spread of the data points suggests that maximum temperature alone is not a strong predictor of precipitation.  - There might be a slight concentration of higher precipitation around the mid-range of maximum temperatures, but it's not strongly defined. |
|  | **Average\_Temp vs. Precipitation:**  - This plot, like the others, **doesn't show a definitive linear relationship**.  - There is variability across the range of average temperatures. |

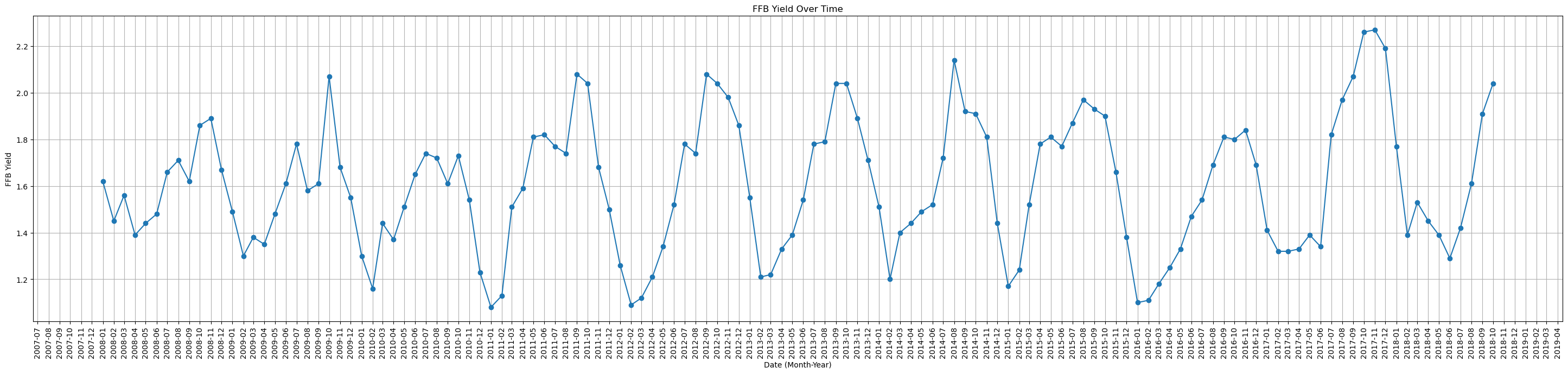
**4.6 Scatter plot (Month vs FFB Yield)**



The scatter plot shows the variable in FFB yield across different months of the year:

* There is a notable trend where FFB yield **increases towards the end of the year**, particularly from **August to November**. This upward trend in the data points suggests that yields tend to rise as the year progresses, possibly peaking during these months.
* The scatter plot indicates a seasonal pattern in FFB yield, with fluctuations observed throughout the year. These variations are likely influenced by seasonal weather changes or other environmental factors that impact oil palm productivity.

**4.7 Time Series plot of FFB Yield**

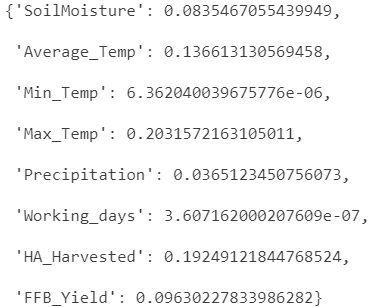


The time series plot of FFB yield from July 2007 to March 2019 clearly illustrates a cyclical pattern, indicating seasonal fluctuations in yield. Here are some refined observations based on the plot:

* **Cyclical Patterns**: The data shows consistent cycles where the FFB yield peaks, followed by a decrease to lower values, reinforcing the presence of seasonal influences on the yield. These cycles appear to recur annually.
* **Seasonal Peaks**: Generally, the peak yield occurs around August to November each year. This suggests that conditions during these months are favorable for the FFB production, possibly due to optimal weather conditions, plantation management practices aligned with this period, or biological cycles of the oil palm.
* **Annual Variability**: While the cyclical pattern is consistent, the amplitude of these cycles (i.e., the difference between peak and trough yields) varies from year to year. This variability could be influenced by additional factors such as varying rainfall, changes in cultivation practices, or other environmental conditions.
* **Low Points**: The lowest yields in the cycles often occur around the first quarter of the year, from January to March, suggesting less favorable conditions for FFB production during these months.

1. **Normality**

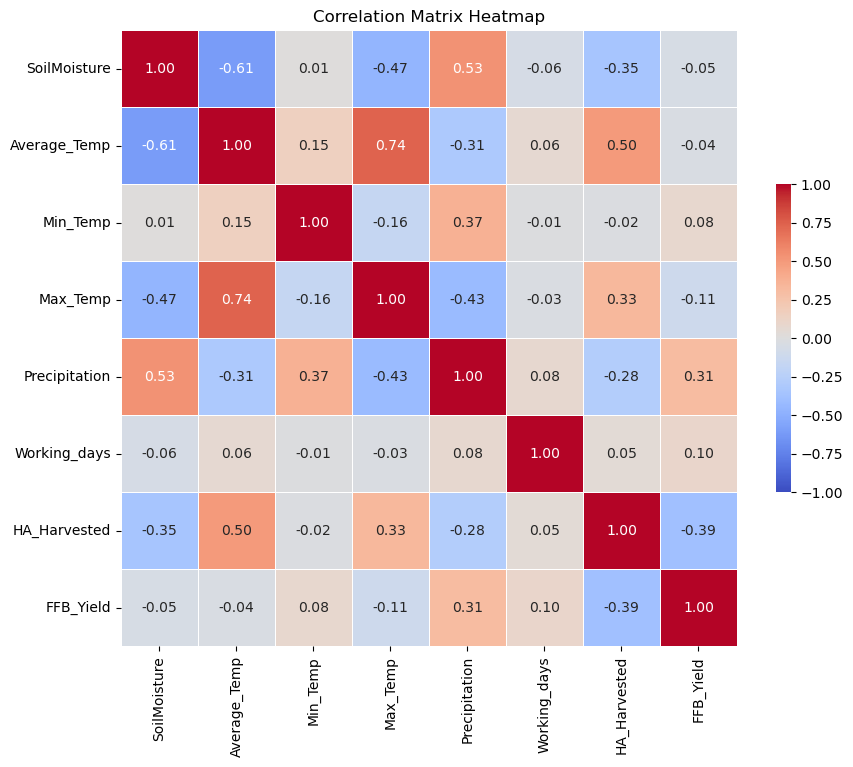
In this section, we will use Shapiro-Wilk test to check the data for each additive if is normal distributed.



* **Min\_Temp**: The p-values are very low (significantly less than 0.05), suggesting that these variables are **not normally distributed**.
* **SoilMoisture, Average\_Temp, Max\_Temp, HA\_Harvested, and FFB\_Yield**: The p-values are higher than 0.05, which indicates that you do not have sufficient evidence to reject the null hypothesis for these variables. They could be considered as **normally distributed**.
* **Precipitation and Working\_days**: The p-values are around 0.036 and extremely low, respectively, suggesting that the Precipitation variable is marginally not normal, and Working\_days is **not normally distributed**.

**5.1 Correlation Analysis**

Since some variables are not normally distributed according to the Shapiro-Wilk test, and to be more robust against the non-normality of your data, **Spearman’s rank correlation** would be a safer choice. It will provide you with insight into the monotonic relationships between variables without the requirement of normality.



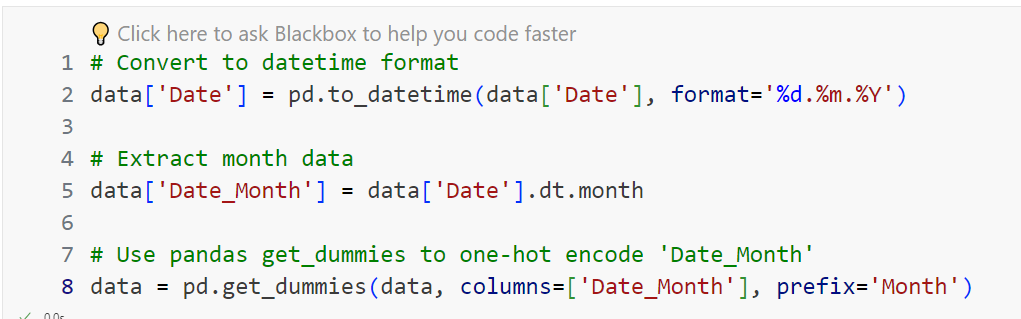
The correlation matrix heatmap shows the relationship between various environmental and operational factors and their impact on FFB yield. There are some key insights from the heatmap:

* **Soil Moisture and Temperature**:
  + Soil moisture has a strong negative correlation with average temperature (-0.61) and maximum temperature (-0.47). This suggests that higher temperatures might be associated with lower soil moisture levels.
* **Temperature Relationships:**
  + Average temperature and maximum temperature are strongly correlated (0.74), indicating that they tend to increase together.
  + Average temperature shows a notable negative correlation with precipitation (-0.31), implying that higher temperatures might occur during drier periods.
* **Precipitation and Soil Moisture**:
  + There is a moderate positive correlation between precipitation and soil moisture (0.53), suggesting that rainfall significantly increases soil moisture levels.
* **Yield Influences:**
  + FFB yield has a weak to negligible correlation with most variables, including soil moisture (-0.05), average temperature (-0.04), and working days (0.10).
  + The strongest correlation with yield is seen with precipitation (0.31), indicating that **higher precipitation may favorably affect FFB yield**.
* **Operational Factors**:
  + Harvested area (HA\_Harvested) shows a moderate positive correlation with average temperature (0.50) and a negative correlation with yield (-0.39), suggesting that larger harvested areas might not necessarily lead to increased yield.

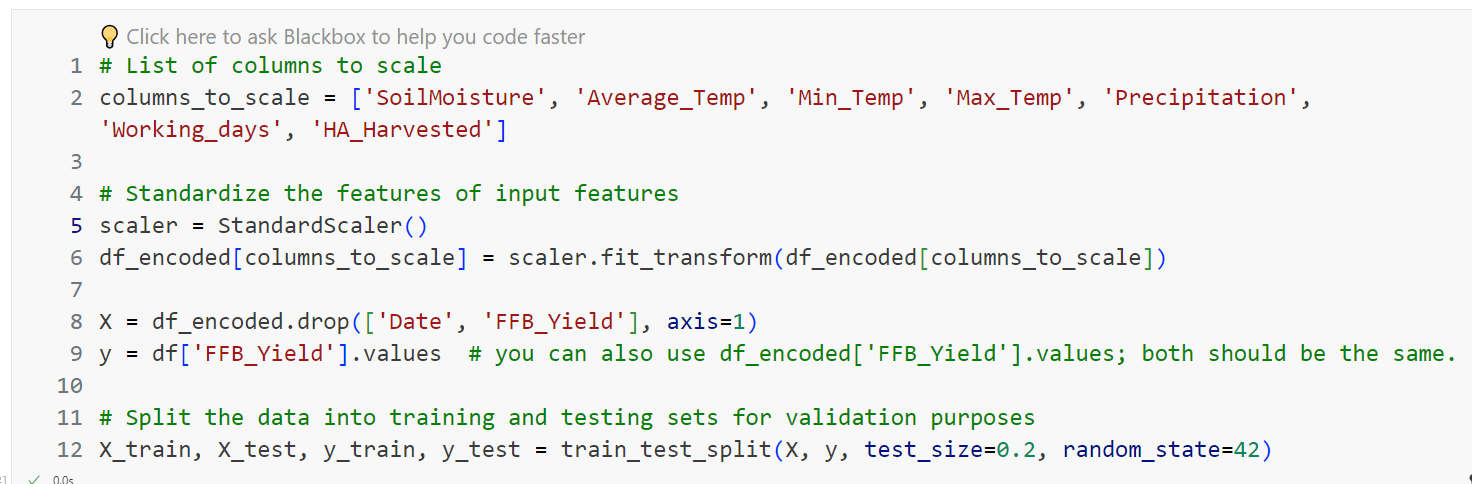
In summary, this heatmap highlights the intricate relationships between climate variables and operational factors, with temperature and precipitation showing significant interactions with other variables, and a moderate influence of precipitation on FFB yield.

1. **Model**

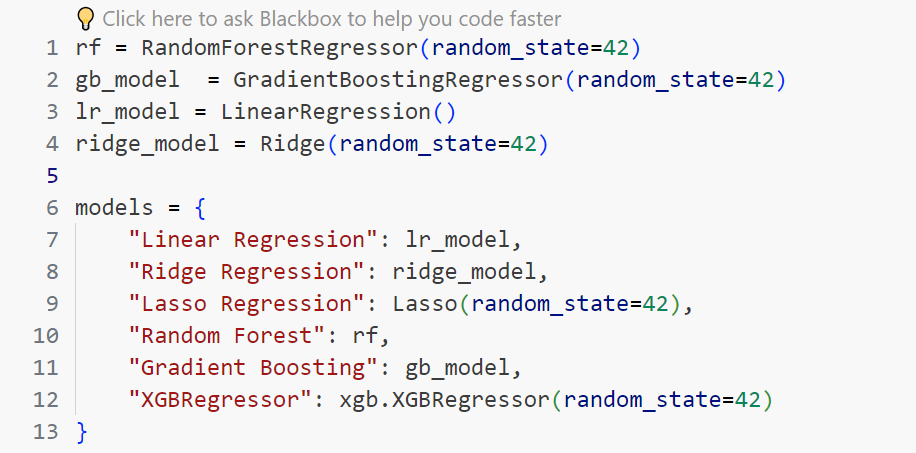
In the previous analysis, we observed that the data exhibits a seasonal pattern. Therefore, extracting the month as a categorical variable can significantly enhance our predictive model.



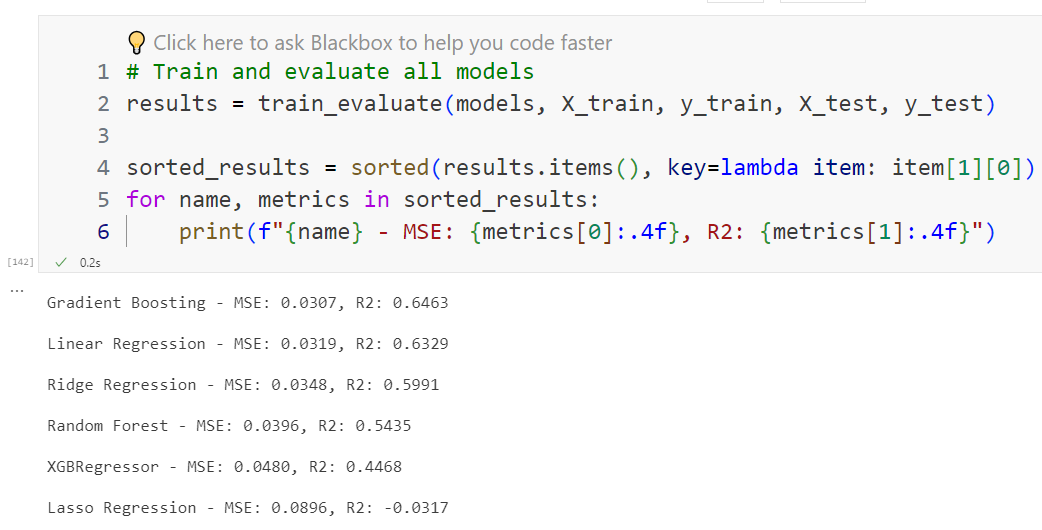
Before training the model, we will scale all variables, including 'SoilMoisture', 'Average\_Temp', 'Min\_Temp', 'Max\_Temp', 'Precipitation', 'Working\_days', and 'HA\_Harvested'. These variables, along with the one-hot encoded 'Month' data, will serve as predictors.



We plan to evaluate several base models to determine which performs best with our dataset:



* 1. **Base Model Result**



In the base model results:

* **Gradient Boosting and Linear Regression models** perform the best, with lower mean squared error (MSE) values indicating better predictive accuracy. They also have higher R-squared (R2) values, suggesting a better fit to the data.
* **Ridge Regression and Random Forest models** perform moderately, with slightly higher MSE and lower R2 values compared to Gradient Boosting and Linear Regression.
* **XGBRegressor model** performs relatively worse, with higher MSE and lower R2 values compared to the other models.
* **Lasso Regression model** performs the poorest, with the highest MSE and negative R2 value, indicating poor model performance.

Overall, **Gradient Boosting and Linear Regression models show promise as they have lower errors and higher explained variance compared to other models.**

* 1. **Hyperparameter**

After performing grid search, the best hyperparameters for each model are as follows:

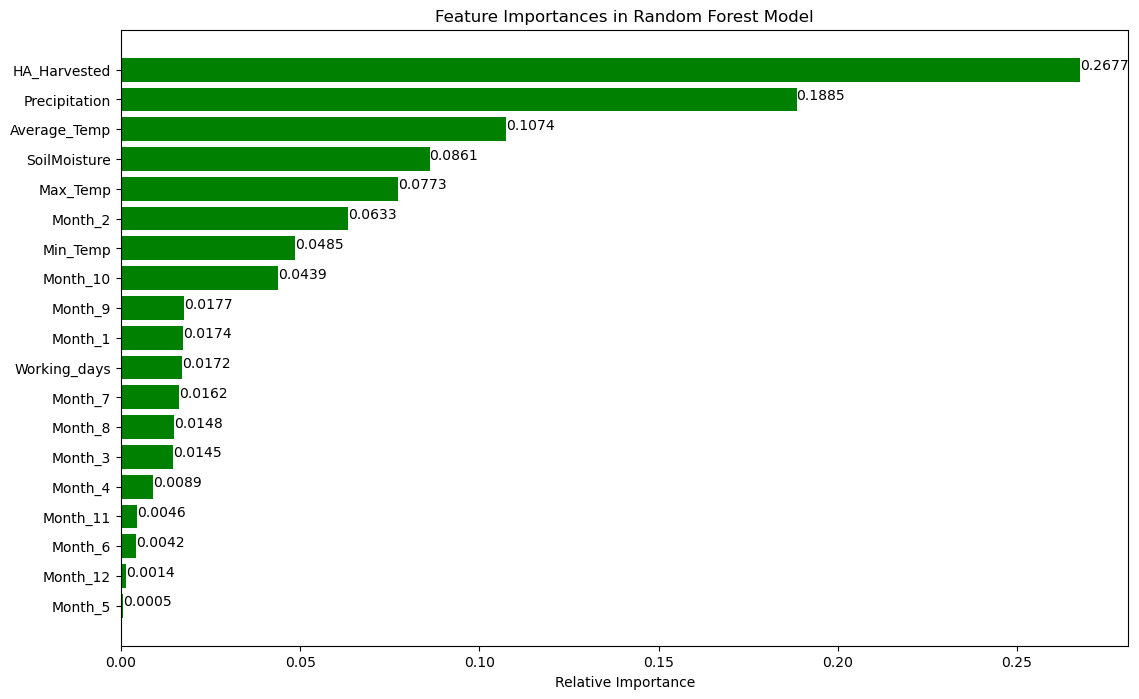
|  |  |
| --- | --- |
| Gradient Boosting Regressor | Fitting 5 folds for each of 27 candidates, totalling 135 fits  Best Hyperparameters: {'learning\_rate': 0.01, 'max\_depth': 5, 'n\_estimators': 200} |
| Linear Regression | Fitting 5 folds for each of 4 candidates, totalling 20 fits  Best Hyperparameters: {'fit\_intercept': False, 'normalize': True} |
| Ridge Regression | Fitting 5 folds for each of 12 candidates, totalling 60 fits  Best Hyperparameters: {'alpha': 0.1, 'fit\_intercept': True, 'normalize': True} |
| Random Forest | Fitting 5 folds for each of 27 candidates, totalling 135 fits  Best Hyperparameters: {'max\_depth': None, 'min\_samples\_split': 10, 'n\_estimators': 200} |

The table below compares the performance metrics (MSE and R-squared) before and after hyperparameter tuning for each model:

|  |  |  |
| --- | --- | --- |
| Model | Before | After |
| Gradient Boosting | MSE: 0.03, R2: 0.64 | MSE: 0.03, R2: 0.63 |
| Linear Regression | MSE: 0.03, R2: 0.63 | MSE: 0.03, R2: 0.63 |
| Ridge Regression | MSE: 0.03, R2: 0.59 | MSE: 0.03, R2: 0.61 |
| Random Forest | MSE: 0.03, R2: 0.54 | MSE: 0.03, R2: 0.53 |

Overall, while there were some minor fluctuations in performance metrics for certain models after hyperparameter tuning, there were no significant improvements.

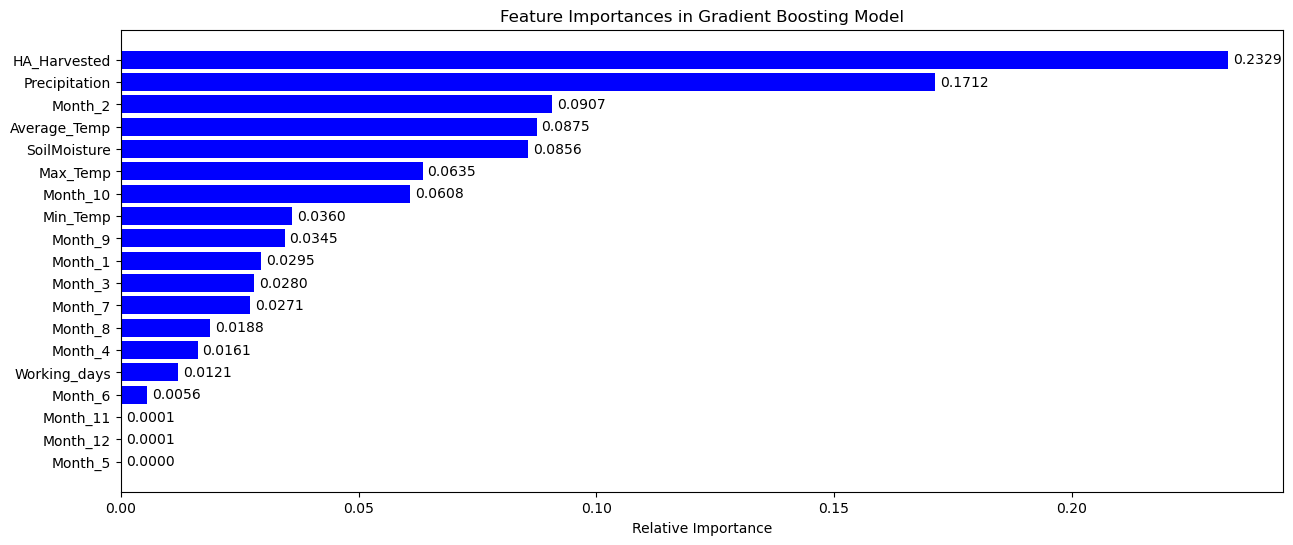
* 1. **Feature Importance**
     1. **Random Forest**



The feature importance graph from the Random Forest Model provides a clear visualization of the influence of various features on the model's predictions:

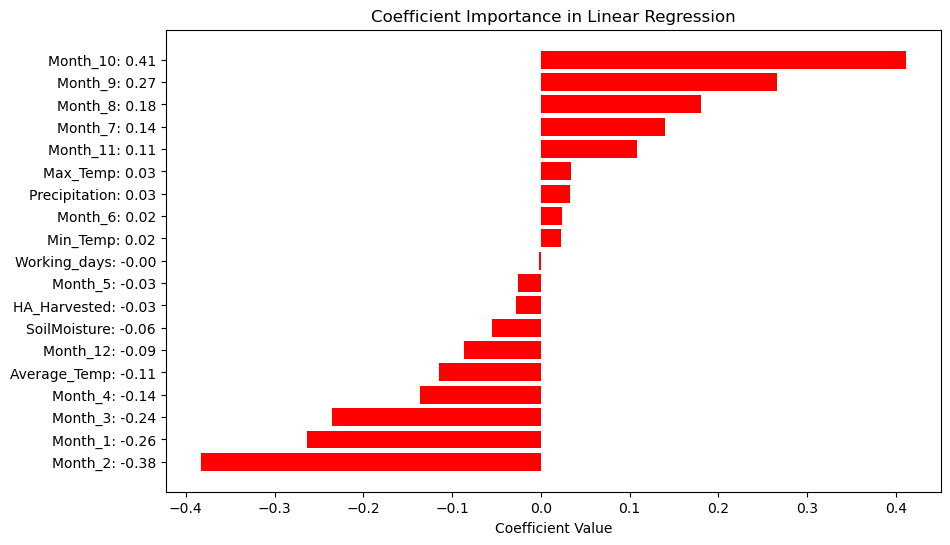
* HA\_Harvested: The most influential feature, contributing approximately 26.77% to the model's decisions. This indicates that the area harvested has the strongest impact on FFB yield.
* Precipitation: The second most important feature, with a relative importance of about 18.85%. This highlights the significant role of rainfall in affecting FFB yield.
* Average\_Temp: Shows a moderate influence with around 10.74% importance, suggesting that the average temperature also plays a key role in determining yield.
* SoilMoisture and Max\_Temp: Both features have a notable impact, with importances of 8.61% and 7.73% respectively, indicating their relevance to FFB yield.
* Month-Specific Influences: Several months show varying degrees of importance, with February (Month\_2) being the most significant among them, having a 6.33% impact. Other months like October (Month\_10) and September (Month\_9) also contribute but to a lesser extent.
* Less Influential Months: Some months, like May (Month\_5) and December (Month\_12), have minimal impact on the model's predictions.
* This distribution of feature importance suggests that both the environmental conditions and the time of year play critical roles in determining the FFB yield, with certain months and climatic factors being more influential than others.
  + 1. **Gradient Boosting Model**

The feature importance graph for the Gradient Boosting Model indicates the significance of various factors affecting FFB yield:



* HA\_Harvested is the most significant predictor, contributing about 23.29% to model predictions, highlighting the crucial role of the area harvested.
* Precipitation is the second most influential feature, with an importance of 17.12%, underscoring the impact of rainfall on yield.
* Month\_2 (February) stands out among the monthly data with 9.07% importance, indicating specific seasonal influences during this month.
* Average\_Temp and SoilMoisture also show substantial impacts, with 8.75% and 8.56% importance respectively, suggesting that temperature and soil moisture conditions significantly influence yield.
* Max\_Temp and Month\_10 (October) follow, indicating their relevance but with lesser impacts compared to other top variables.
* Less Influential Months: Several months like May (Month\_5), November (Month\_11), and December (Month\_12) show negligible importance, suggesting minimal impact on the yield predictions.
  + 1. **Linear Regression**

The coefficient importance graph from the Linear Regression model highlights how various factors affect the FFB yield:



**Positive Influences:**

* Month 10 (October): Shows the strongest positive influence with a coefficient of 0.41, suggesting that FFB yield tends to be significantly higher in October.
* Month 9 (September) and Month 8 (August): Also have notable positive impacts with coefficients of 0.27 and 0.18, respectively, indicating higher yields during these months.

**Negative Influences:**

* Month 2 (February): Exhibits the most substantial negative effect with a coefficient of -0.38, indicating a significant decrease in FFB yield during this month.
* Month 3 (March) and Month 1 (January): Follow closely with coefficients of -0.24 and -0.26, respectively, also suggesting lower yields.

**Minor Influences:**

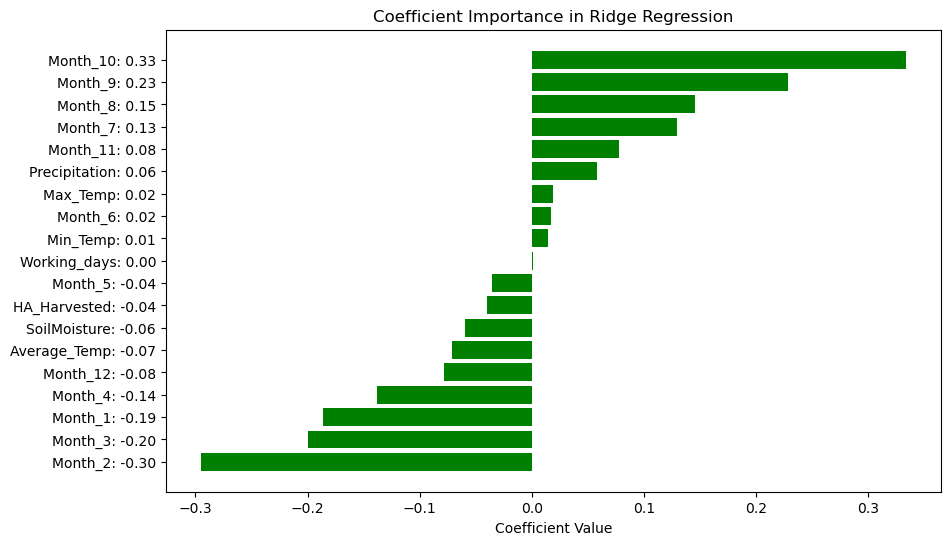
* Max Temp, Precipitation, and Min Temp: Have small positive coefficients (around 0.02 to 0.03), indicating slight positive impacts on yield.
* Soil Moisture and Average Temp: Show small negative impacts on yield, with coefficients of -0.06 and -0.11, respectively.

**Negligible or No Influence:**

* Working Days: Shows virtually no impact on yield with a coefficient close to zero.

This pattern of coefficients underscores the significant seasonal variation in FFB yield, with certain months clearly more favourable for yield than others. The variation in temperature and precipitation also plays a role but is less pronounced compared to the seasonal (monthly) effects.

* + 1. Ridge Regression



Similar as linear regression model, the pattern of coefficients again highlights the pronounced effect of seasonal variation on FFB yield, with certain months clearly more conducive to higher yields. The environmental factors like temperature and precipitation also contribute but are less impactful compared to the month-specific effects.

**6.4.5 Feature Importance summary**

The analysis of feature importance across Random Forest and Gradient Boosting models indicates that HA\_Harvested (Harvested Area), Precipitation, and Average Temperature are critical factors influencing FFB yield. These environmental and operational factors consistently show substantial impact across these models, underscoring their importance in predicting yield outcomes.

Conversely, Linear and Ridge Regression models highlight the significant influence of seasonal variations, with specific months like October, September, and August showing positive effects, and months like February, March, and January exhibiting negative impacts. This suggests a strong seasonal pattern in FFB yield, with certain times of the year being more favourable for yield than others.