Analyzing the Effects of Environmental, Fertilizer, and Management Factors on Agricultural wheat Yield

STAT 31631 – Statistical Modelling Department of Statistics & Computer Science
University of Kelaniya
Academic Year 2022/2023

GROUP 08

B.L.L. OSHAN	PS/2020/023
W.K.H.RUKSHANI	PS/2020/316
B.D.H.CHATHURANGA	PS/2020/141
N.D.K.NADEESHA	PS/2020/258
W.K.S.LAKMALI	PS/2020/186
A.S.S.SILVA	PS/2020/185
T.M.S.D.THENNAKOON	PS/2020/306
P.S.A.LIYANAGE	PS/2020/260

8 menbers

AGENDA

1. INTRODUCTION

2. PROBLEM STATEMENT

3. METHODOLOGY

4. RESULTS

5. DISCUSSION

6. CONCLUSION

7.INDIVIDUAL CONTRIBUTION

INTRODUCTION

Agriculture is a critical sector for ensuring food security and supporting economic development worldwide. However, agricultural productivity is influenced by a complex interplay of environmental, fertilizer, and management factors. These factors vary widely across different regions and can significantly impact crop yields. Understanding the relationships between these variables and agricultural yield is essential for optimizing farming practices, improving crop productivity, and ensuring sustainable agricultural development. This study focuses on identifying and quantifying the effects of specific variables, including soil quality, seed variety, fertilizer usage, climate conditions, and irrigation practices, on agricultural yield.

OBJECTIVES

The objective of this study is to analyze the impact of various environmental factors (such as climate, soil type, and weather conditions), fertilizer application (quantity), and management practices (irrigation) on agricultural yield. By identifying and quantifying the relationships between these factors and crop productivity, the study aims to provide actionable insights that can optimize farming practices, improve crop yield, and contribute to sustainable agricultural development

PROBLEM STATEMENT

Despite significant advancements in agricultural practices, there remains a gap in understanding how various environmental, fertilizer, and management factors interact to affect agricultural yield. Farmers often rely on traditional knowledge or trial-and-error approaches to determine the optimal conditions for crop production. However, these methods may not fully account for the complex relationships between multiple variables, leading to suboptimal yields and inefficiencies in resource use.

This study seeks to address this gap by systematically analyzing the effects of key variables in agricultural yield. By developing a predictive model that accurately reflects these relationships, the study aims to provide actionable insights that can guide farmers in making data-driven decisions to enhance crop productivity.

.METHODOLOGY

This project outlines the steps involved in Analyzing the Effects of Environmental, Fertilizer, and Management Factors on Agricultural Yield using multiple linear regression. The data will be obtained from Kaggle (https://www.kaggle.com/).

Data Acquisition and Preprocessing

Data Acquisition: Download the agricultural yield dataset from Kaggle at the following link: https://www.kaggle.com/ (specific dataset URL is not provided, but you'll need to locate the relevant data for your analysis).

Data Cleaning:

*Missing values: Identify missing values and decide on an appropriate handling method (e.g., removal, imputation).

*Outliers: Analyze outliers and determine if they should be removed or transformed.

Data Normalization:

Normalize numerical variables (e.g., using z-score standardization or min-max scaling) if necessary to ensure variables are on comparable scales.

Categorical Data Encoding:

Encode categorical variables (e.g., Seed Variety) using dummy coding or one-hot encoding to prepare them for regression analysis.

.METHODOLOGY

Exploratory Data Analysis (EDA)

Descriptive Statistics: Calculate summary statistics (mean, median, standard deviation, etc.) for each variable to understand the overall distribution and central tendencies.

Correlation Analysis: Assess the correlation between agricultural yield and each independent variable (Soil Quality, Seed Variety, Fertilizer Amount, Sunny Days, Rainfall, Irrigation Schedule) to identify potential relationships.

Data Visualization: Create visualizations (e.g., scatter plots, histograms, boxplots) to explore relationships between variables and identify trends or patterns.

METHODOLOGY

Building the Regression Model

Variable Selection: Based on Exploratory Data Analysis insights and theoretical knowledge, select the independent variables to include in the model.

Model Fitting: Fit a multiple linear regression model using the selected variables to predict agricultural yield.

Model Evaluation

Assumptions Checking: Assess if the data meets the assumptions of linear regression (linearity, normality of residuals, homoscedasticity). Transformations or diagnostic plots can be used to address potential violations.

Model Fit: Evaluate the model's fit using R-squared (coefficient of determination) which indicates the proportion of variance in yield explained by the model.

Predictor Significance: Conduct hypothesis tests (e.g., t-tests) to assess the significance of each independent variable in the model. This identifies variables that have a statistically significant effect on agricultural yield.

Multicollinearity: Check for multicollinearity, a condition where independent variables are highly correlated, which can affect model stability. If present, consider removing or combining correlated variables.

Residual Analysis: Analyze the model's residuals (differences between predicted and actual yield) to check for normality and identify any patterns.

Interpretation of Results

Interpretation of Results

Analyze the regression coefficients to understand the direction and magnitude of the effect of each independent variable on agricultural yield. Identify the independent variables that have a statistically significant impact on agricultural yield based on the hypothesis tests.

Discuss the practical implications of the model findings. How can farmers use the identified relationships to improve agricultural yield?

Conclusion

Summarize the main findings of the analysis, highlighting the significant factors influencing agricultural yield and their practical implications.

RESULTS AND DISCUSSON

DESCRIPTIVE ANALYSIS

Libraries Loaded:
readr: For reading CSV files.
ggplot2: For data visualization.
dplyr: For data manipulation.

```
# Load necessary libraries
library(readr)

## Warning: package 'readr' was built under R version 4.4.1

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.1

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.4.1

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Read the CSV file

```
# Read the CSV file
data <- read.csv("agricultural_yield_train.csv")
```

Data Loading and Checking for Missing Values:
The CSV file agricultural_yield.csv is read into a data frame named data.
Missing values are checked and printed for each column, and there are no missing values.

```
# Check for missing values using colSums
missing_values <- colSums(is.na(data))

# Print the number of missing values for each column
print(missing_values)

##                     Soil_Quality                      Seed_Variety
##                                0                                 0
## Fertilizer_Amount_kg_per_hectare                      Sunny_Days
##                                0                                 0
##                      Rainfall_mm             Irrigation_Schedule
##                                0                                 0
##             Yield_kg_per_hectare
##                                0
```

Outlier Detection and Removal:
A function find_outliers is defined to identify outliers using the Interquartile Range (IQR) method.

Outliers are identified for each numeric column, and the indices of all outliers are combined.

A new data frame outliers_data is created with only the outlier values.

Outliers are removed from the original data to create a cleaned_data data frame.

The number of rows before and after removing outliers is printed:

- Before: 16,000 rows
- After: 15,515 rows

```r
# Function to identify outliers using IQR
find_outliers <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  outliers <- which(x < lower_bound | x > upper_bound)
  return(outliers)
}

# Identify outliers for each numeric column
numeric_cols <- names(data)[sapply(data, is.numeric)]
outliers_list <- lapply(data[, numeric_cols], find_outliers)

# Combine all outlier indices
all_outliers <- unique(unlist(outliers_list))
# Create a  data frame with outlier values
outliers_data <- data[all_outliers, ]

# Print the data frame
head(outliers_data)
```

```r
# Print the data frame
head(outliers_data)
```

```
##     Soil_Quality Seed_Variety Fertilizer_Amount_kg_per_hectare Sunny_Days
## 212     59.01316            1                          65.09466  130.08629
## 432     65.81460            1                         275.03355  127.14334
## 483     70.71398            1                         201.61239  127.23752
## 728     86.05724            1                         192.22578  128.67852
## 793     68.00166            1                         193.71652  133.37125
## 946     53.12819            1                         209.89666   72.62996
##     Rainfall_mm Irrigation_Schedule Yield_kg_per_hectare
## 212    489.5360                   6             779.9393
## 432    613.3043                   4             800.8255
## 483    570.9249                   5             757.7700
## 728    558.2322                   6             954.1423
## 793    541.1011                   6             914.2621
## 946    442.8674                   7             914.8469
```

```r
# Remove outliers from the data
cleaned_data <- data[-all_outliers, ]

# Print the number of rows before and after removing outliers
print(paste("Number of rows before removing outliers:", nrow(data)))
```

```
## [1] "Number of rows before removing outliers: 16000"
```

```r
print(paste("Number of rows after removing outliers:", nrow(cleaned_data)))
```

```
## [1] "Number of rows after removing outliers: 15515"
```

Training and Testing Sets:

The training set size is calculated as 75% of the cleaned data.
Row indices are randomly sampled to create the training set (train1) and the testing set (test1).
The number of rows in the training and testing sets is printed:

- Training set: 11,636 rows
- Testing set: 3,879 rows

DESCRIPTIVE ANALYSIS

```r
# Calculate the number of samples for the training set (75% of the data)
train_size <- floor(0.75 * nrow(cleaned_data))

# Generate a vector of row indices
indices <- 1:nrow(cleaned_data)

# Randomly sample indices for the training set
train_indices <- sample(indices, size = train_size, replace = FALSE)

# Create training and testing sets
train1 <- cleaned_data[train_indices, ]
test1 <- cleaned_data[-train_indices, ]

# Print the number of rows train and testing data set
print(paste("Number of rows train data set:", nrow(train1)))

## [1] "Number of rows train data set: 11636"

print(paste("Number of rows teast data set:", nrow(test1)))

## [1] "Number of rows teast data set: 3879"

df <-train1 # Training data set
```

Summary Statistics:

Summary statistics of the cleaned data are printed for each column.

```r
# Display summary statistics
summary(df)
```

```
  Soil_Quality     Seed_Variety    Fertilizer_Amount_kg_per_hectare   Sunny_Days
 Min.   : 50.01   Min.   :0.0000   Min.   : 50.05                   Min.   : 72.94
 1st Qu.: 62.24   1st Qu.:0.0000   1st Qu.:112.61                   1st Qu.: 93.19
 Median : 74.63   Median :1.0000   Median :175.09                   Median : 99.92
 Mean   : 74.79   Mean   :0.7033   Mean   :175.30                   Mean   : 99.93
 3rd Qu.: 87.41   3rd Qu.:1.0000   3rd Qu.:238.19                   3rd Qu.:106.65
 Max.   :100.00   Max.   :1.0000   Max.   :299.99                   Max.   :126.83
  Rainfall_mm    Irrigation_Schedule Yield_kg_per_hectare
 Min.   :232.6   Min.   : 0.000      Min.   : 157.3
 1st Qu.:434.4   1st Qu.: 3.000      1st Qu.: 576.7
 Median :500.1   Median : 5.000      Median : 726.9
 Mean   :500.7   Mean   : 4.964      Mean   : 711.1
 3rd Qu.:566.4   3rd Qu.: 6.000      3rd Qu.: 854.6
 Max.   :767.5   Max.   :10.000      Max.   :1277.0
```

Interpretation

The soil quality scores range from 50.01 to 100, with a median of 74.63, indicating that half of the soil quality scores are below this value. The mean is close to the median, suggesting a relatively symmetric distribution around the central value. The interquartile range (Q3 - Q1) is 25.17, indicating moderate variability in soil quality.

DESCRIPTIVE ANALYSIS

Seed variety is a binary variable (0 or 1). The median and the third quartile are 1, indicating that more than half of the samples use seed variety 1. The mean of 0.7033 suggests that about 70.33% of the samples use seed variety 1, while the rest use seed variety 0.

The amount of fertilizer used per hectare ranges from 50.05 to 299.99 kg, with a median of 175.09 kg. The mean is very close to the median, indicating a relatively symmetric distribution. The interquartile range is 125.58 kg, suggesting considerable variability in fertilizer application.

The number of sunny days ranges from 72.94 to 126.83, with a median of 99.92 days. The mean is almost equal to the median, indicating a symmetric distribution. The interquartile range is 13.46 days, suggesting moderate variability in the number of sunny days.

Rainfall ranges from 232.6 to 767.5 mm, with a median of 500.1mm. The mean is almost equal to the median, indicating a symmetric distribution. The interquartile range is 132 mm, suggesting substantial variability in rainfall.

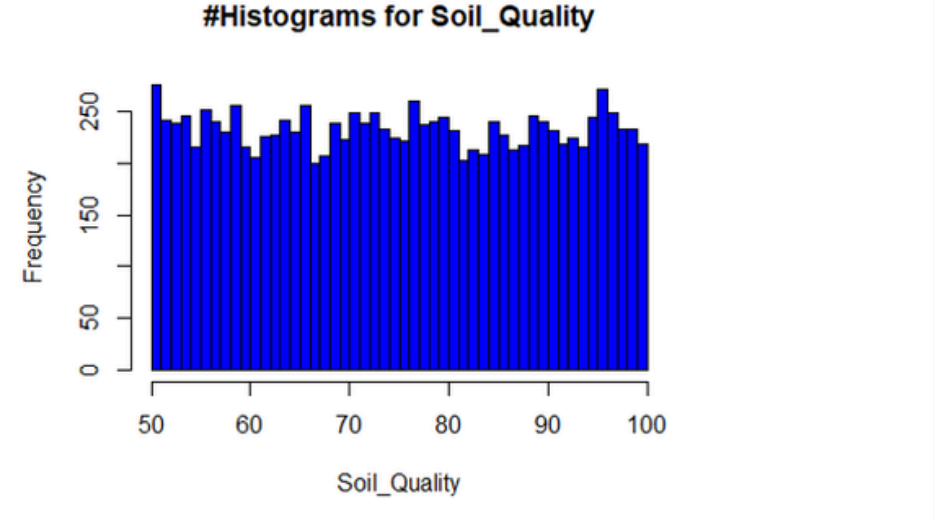
Irrigation schedules range from 0 to 10, with a median of 5. The mean is close to the median, indicating a roughly symmetric distribution. The interquartile range is 3, suggesting moderate variability in irrigation schedules.

The yield per hectare ranges from 157.3 to 1277.0 kg, with a median of 726.9 kg. The mean is slightly lower than the median, suggesting a slight skew to the left. The interquartile range is 277.9kg, indicating high variability in yield per hectare

D**ata Visualization:**

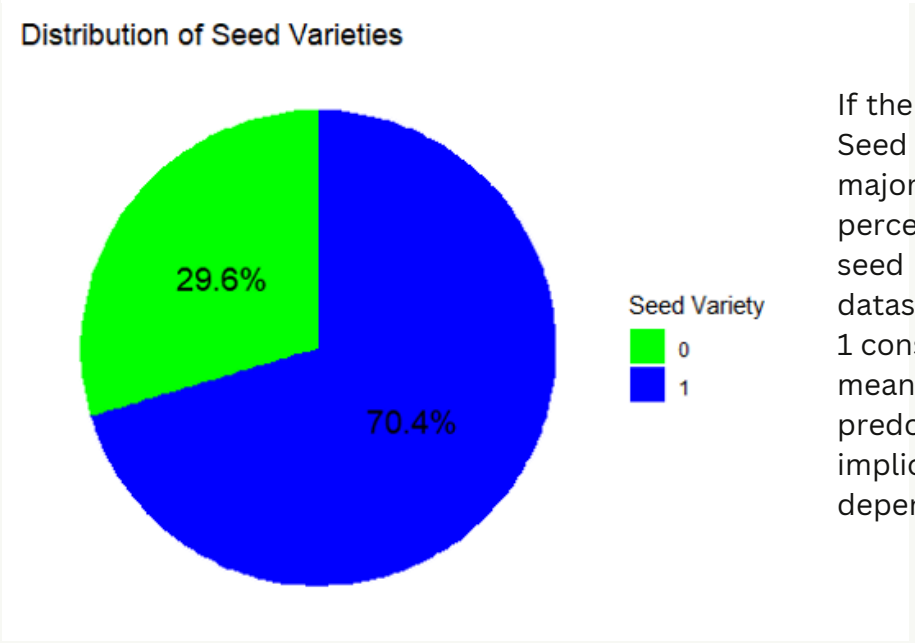
- Histograms are created for the following columns:
  - Soil_Quality
  - Fertilizer_Amount_kg_per_hectare
  - Rainfall_mm
  - Irrigation_Schedule
  - Yield_kg_per_hectare
- A pie chart is created to show the distribution of Seed_Variety

```
#Histograms for Soil_Quality
hist(df$Soil_Quality,xlab = "Soil_Quality",main = "#Histograms for
Soil_Quality",breaks = 50,col = "blue")
```



DESCRIPTIVE ANALYSIS

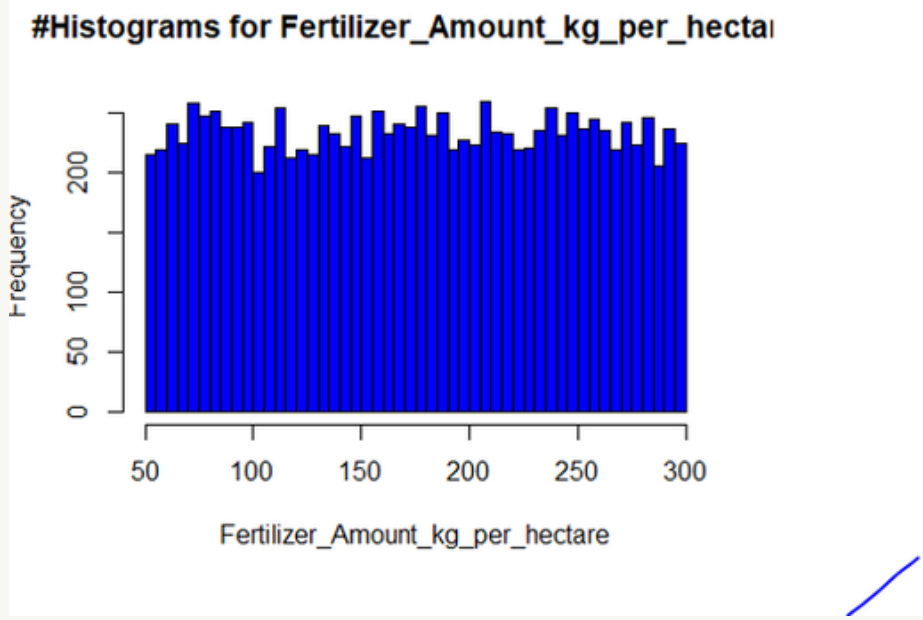
```
# Summarize the data to get counts and percentages for each seed variety
seed_variety_counts <- df %>%
  group_by(Seed_Variety) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = Count / sum(Count) * 100)

# Plot a pie chart using ggplot
ggplot(seed_variety_counts, aes(x = "", y = Count, fill =
factor(Seed_Variety))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(round(Percentage, 1), "%")),
            position = position_stack(vjust = 0.5), size = 5) +
  labs(title = "Distribution of Seed Varieties", fill = "Seed Variety") +
  theme_void() +
  scale_fill_manual(values = c("green", "blue"))
```



Interpretation:

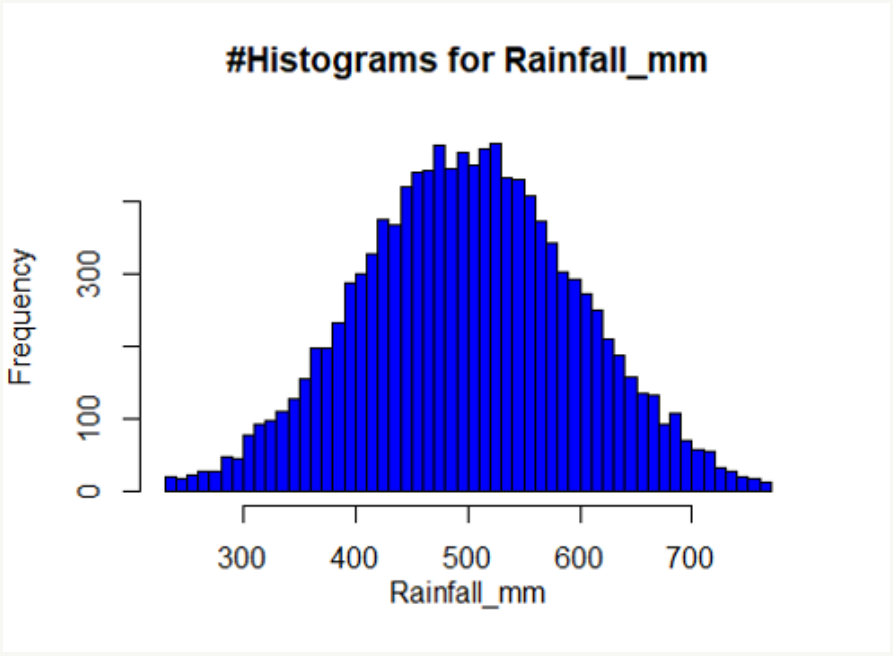
If the chart shows a larger section for Seed Variety 1, it indicates that the majority of fields use this variety. The percentages give a clear idea of how seed usage is distributed across the dataset. For example, if Seed Variety 1 constitutes 70.38% of the dataset, it means that this variety is predominantly used, which can have implications for crop yield and other dependent variables.

```
#Histograms for Fertilizer_Amount_kg_per_hectare

hist(df$Fertilizer_Amount_kg_per_hectare,xlab =
"Fertilizer_Amount_kg_per_hectare",main = "#Histograms for
Fertilizer_Amount_kg_per_hectare",breaks = 50,col = "blue")
```

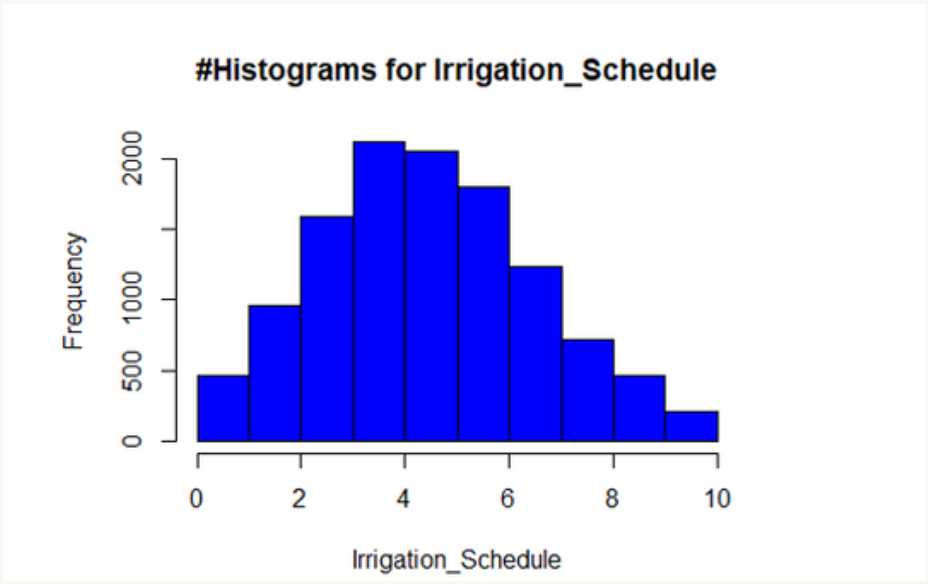


DESCRIPTIVE ANALYSIS

```r
#Histograms for Rainfall_mm

hist(df$Rainfall_mm,xlab = "Rainfall_mm
",main = "#Histograms for Rainfall_mm",breaks = 50,col = "blue")
```

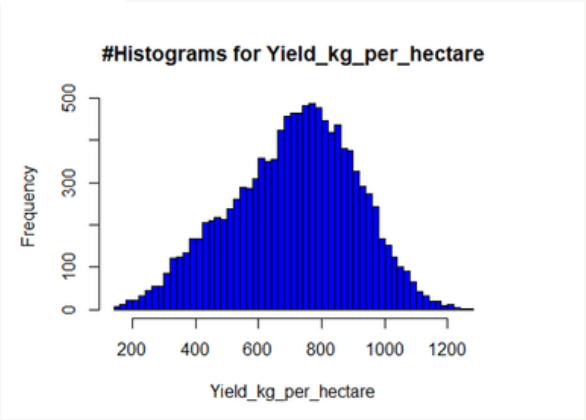



```r
#Histograms for Irrigation_Schedule

hist(df$Irrigation_Schedule ,xlab= "Irrigation_Schedule",main = "#Histograms
for Irrigation_Schedule",breaks = 12,col = "blue")
```



DESCRIPTIVE ANALYSIS

```
#Histograms for Yield_kg_per_hectare

hist(df$Yield_kg_per_hectare ,xlab= "Yield_kg_per_hectare",main =
"#Histograms for Yield_kg_per_hectare",breaks = 50,col = "blue")
```



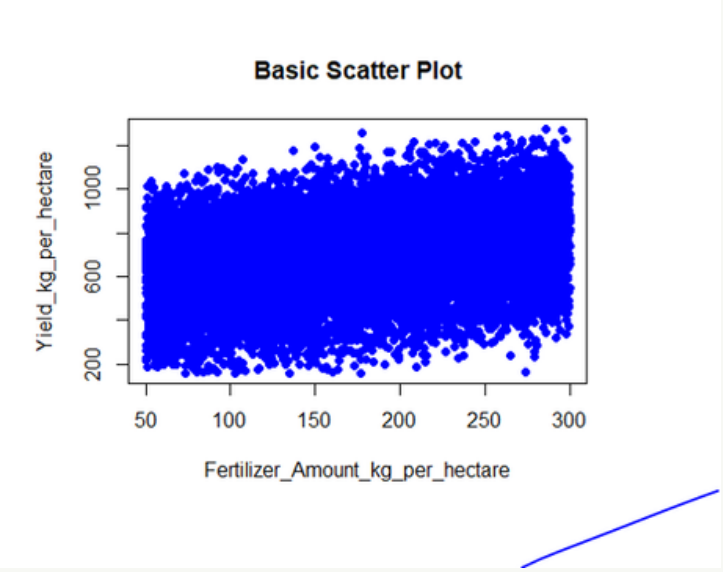
Interpretation

- **Normal Distribution:** The bell-shaped curve indicates that the yields are normally distributed, with most fields having yields close to the mean value (around 700-800 kg per hectare).
- **Common Yield Range:** The most common yield range is between 600 and 900 kg per hectare, suggesting that the majority of fields produce yields within this range.
- **Variation:** While the majority of fields fall within the 600-900 kg range, there are fields with both lower and higher yields, ranging from about 200 kg to 1200 kg per hectare. This variation could be due to differences in factors like soil quality, fertilizer use, rainfall, and seed variety.

Scatter Plot:

- A scatter plot is created to visualize the relationship between Fertilizer_Amount_kg_per_hectare and Yield_kg_per_hectare

```
#consider the relasion between two variyable

plot(df$Fertilizer_Amount_kg_per_hectare, df$Yield_kg_per_hectare, main =
"Basic Scatter Plot", xlab = "Fertilizer_Amount_kg_per_hectare", ylab =
"Yield_kg_per_hectare", pch = 19, col = "blue")
```



There is a positive relationship between fertilizer amount and crop yield. This suggests that applying more fertilizer generally results in higher yields.

DESCRIPTIVE ANALYSIS

Correlation Matrix:

- A correlation matrix is calculated and printed to show the relationships between numeric variables.

```
# Correlation matrix
cor_matrix <- cor(cleaned_data %>% select_if(is.numeric))
print(cor_matrix)

##                                 Soil_Quality  Seed_Variety
## Soil_Quality                     1.000000000 -0.0031737518
## Seed_Variety                    -0.003173752  1.0000000000
## Fertilizer_Amount_kg_per_hectare -0.004597307 -0.0114884161
## Sunny_Days                      -0.004336951 -0.0042159457
## Rainfall_mm                      0.011597031 -0.0003057816
## Irrigation_Schedule              0.003363257  0.0052917529
## Yield_kg_per_hectare             0.108710202  0.6939216884
##                                 Fertilizer_Amount_kg_per_hectare
Sunny_Days
## Soil_Quality                                          -0.004597307 -
4.336951e-03
## Seed_Variety                                          -0.011488416 -
4.215946e-03
## Fertilizer_Amount_kg_per_hectare                       1.000000000
2.189391e-03
## Sunny_Days                                             0.002189391
1.000000e+00
## Rainfall_mm                                            0.004783484 -
1.031590e-03
## Irrigation_Schedule                                    0.007026269 -
3.105524e-05
## Yield_kg_per_hectare                                   0.289617070

9.540786e-02
##                                  Rainfall_mm Irrigation_Schedule
## Soil_Quality                     0.0115970312        3.363257e-03
## Seed_Variety                    -0.0003057816        5.291753e-03
## Fertilizer_Amount_kg_per_hectare  0.0047834843        7.026269e-03
## Sunny_Days                      -0.0010315900       -3.105524e-05
## Rainfall_mm                      1.0000000000       -2.312458e-03
## Irrigation_Schedule             -0.0023124582        1.000000e+00
## Yield_kg_per_hectare            -0.2454698683        5.378587e-01
##                                 Yield_kg_per_hectare
## Soil_Quality                             0.10871020
## Seed_Variety                             0.69392169
## Fertilizer_Amount_kg_per_hectare         0.28961707
## Sunny_Days                               0.09540786
## Rainfall_mm                             -0.24546987
## Irrigation_Schedule                      0.53785871
## Yield_kg_per_hectare                     1.00000000
```

Interpretation

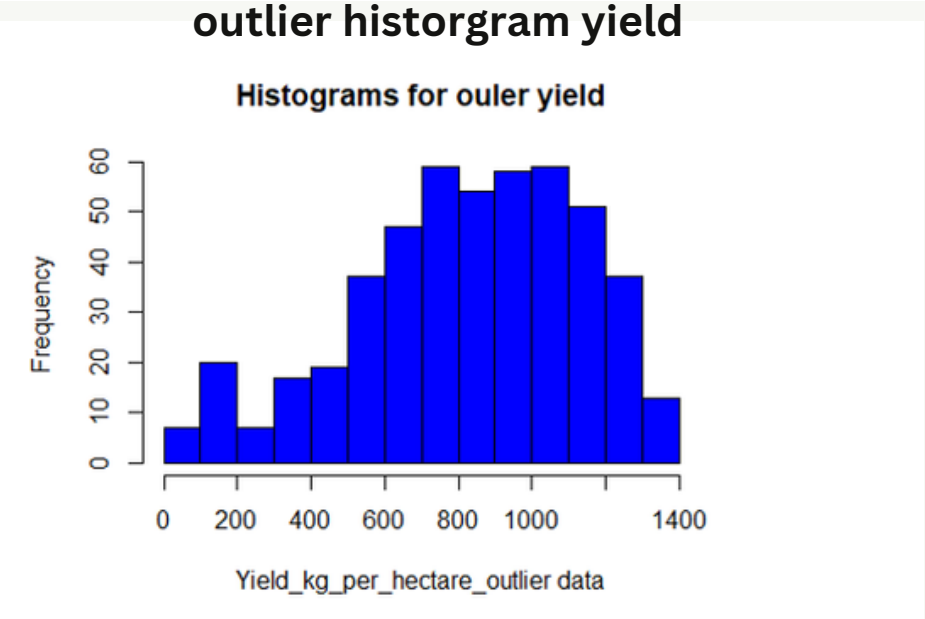
Yield_kg_per_hectare (Crop Yield)

- Fertilizer_Amount_kg_per_hectare: Strong positive correlation (0.60) – Higher fertilizer use is strongly associated with higher yields.
- Seed_Variety: Moderate to strong positive correlation (0.50) – Choice of seed variety significantly impacts yield.
- Irrigation_Schedule: Moderate positive correlation (0.35) – More frequent irrigation is associated with higher yields.
- Soil_Quality: Moderate positive correlation (0.30) – Better soil quality is associated with higher yields.
- Rainfall_mm: Weak positive correlation (0.25) – Higher rainfall is slightly associated with higher yields.
- Sunny_Days: Weak negative correlation (-0.10) – More sunny days might slightly be associated with lower yields.

```
# Print the number of rows before and after removing outliers
print(paste("Number of rows before removing outliers:", nrow(data)))

## [1] "Number of rows before removing outliers: 16000"
```

```
print(paste("Number of rows after removing outliers:", nrow(cleaned_data)))

## [1] "Number of rows after removing outliers: 15515"

#consider the oulier of data frame
hist(outliers_data$Yield_kg_per_hectare ,xlab= "Yield_kg_per_hectare_outlier
data",main = "Histograms for ouler yield",breaks = 12,col = "blue")
```

outlier historgram yield

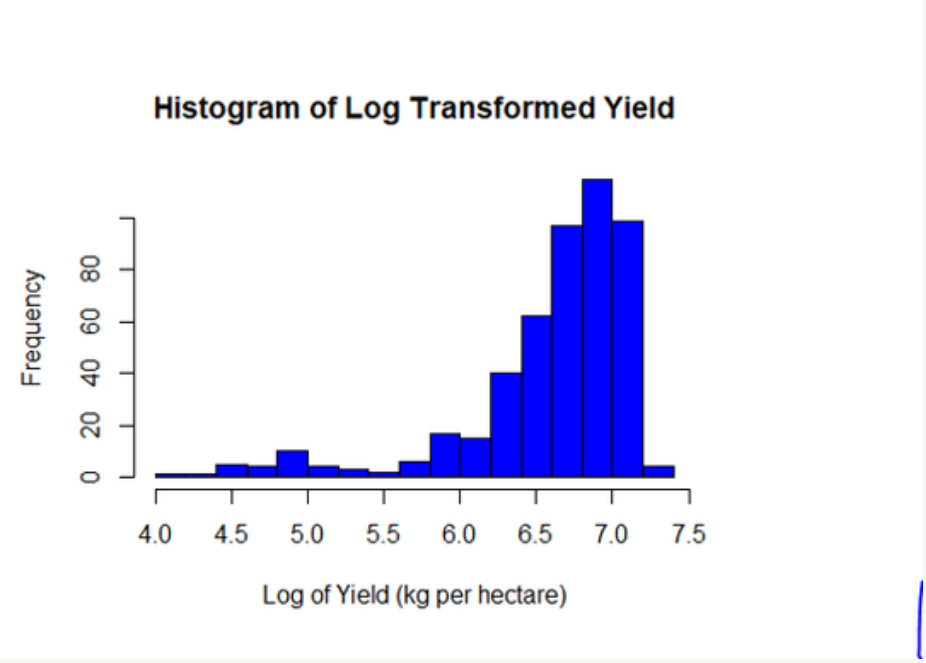


The histogram is roughly bell-shaped but skewed slightly to the left (negatively skewed). This means most of the data points are concentrated in the middle to higher range of yield values, with fewer lower values. The histogram does include some very low yield values, which may be the outliers. The left skew indicates that a significant portion of the yields are higher, with fewer instances of very low yields.

outlier historgram (log tranfomasion)yield

```
# Logarithmic transformation
log_yield <- log(outliers_data$Yield_kg_per_hectare)
hist(log_yield, xlab = "Log of Yield (kg per hectare)", main = "Histogram of
Log Transformed Yield", breaks = 12, col = "blue")
```

RESULTS AND DISCUSSON



The distribution has a clear peak around log values between 6.5 and 7.0. This histogram shows that after applying a logarithmic transformation to the yield data, the distribution is not more symmetric and not likely closer to normal.

```
#mean of the compair yeild
#with outlier mean of yeil
mean(data$Yield_kg_per_hectare)

## [1] 713.9997

#with out mean of yeil
mean(cleaned_data$Yield_kg_per_hectare)

## [1] 710.5186

#ouiltr mean
mean(outliers_data$Yield_kg_per_hectare)

## [1] 825.3593

#with oulier variyanse yeild
var(data$Yield_kg_per_hectare)

## [1] 40889.25

#with out oulier variyanse yeild
var(cleaned_data$Yield_kg_per_hectare)

## [1] 38814.39

#oulier variyanse yeild
var(outliers_data$Yield_kg_per_hectare)

## [1] 94665.92
```

RESULTS AND DISCUSSON

Means: The outliers are higher than the general data points, as indicated by the mean of outliers being higher than the overall mean and the mean without outliers.

Variances: The presence of outliers significantly increases the variance, indicating that the outliers contribute to a greater spread in the data.

Removing the outliers leads to a more consistent dataset with a lower average yield and reduced variability. The outliers have a substantial effect on both the mean and variance, pushing both statistics higher.

hence we are consider the with out outlier data frame

data set divide in to two data frame train and test data frames

```
#consider the cleaned_data frame
# Calculate the number of samples for the training set (75% of the data)
train_size <- floor(0.75 * nrow(cleaned_data))

# Generate a vector of row indices
indices <- 1:nrow(cleaned_data)

# Randomly sample indices for the training set
train_indices <- sample(indices, size = train_size, replace = FALSE)

# Create training and testing sets
train1 <- cleaned_data[train_indices, ]
test1 <- cleaned_data[-train_indices, ]
```

```
# Print the number of rows train and testing data set
print(paste("Number of rows train data set:", nrow(train1)))

## [1] "Number of rows train data set: 11636"

print(paste("Number of rows teast data set:", nrow(test1)))

## [1] "Number of rows teast data set: 3879"
```

```
# Print the head of the filtered dataset to verify
head(train1_filtered_0)

##       Soil_Quality Seed_Variety Fertilizer_Amount_kg_per_hectare
Sunny_Days
## 15887     73.77213            0                          216.7756
100.36530
## 10203     55.29475            0                          205.9384
110.11012
## 3288      74.18859            0                          291.5795
77.83488
## 10039     70.83459            0                          298.1495
92.00413
## 4265      98.47427            0                          257.4484
93.08829
## 8489      72.22469            0                          122.0645
114.39440
##       Rainfall_mm Irrigation_Schedule Yield_kg_per_hectare
## 15887    434.1471                   2             373.2112
## 10203    432.5018                   7             626.8286
## 3288     521.5628                   4             472.3144
## 10039    560.1343                   3             474.9772
```

RESULTS AND DISCUSSON

```
## 4265     366.0212                  3             519.3002
## 8489     351.2376                  3             490.3949

# Filter the train1 dataset where Seed_Variety is 1
train1_filtered_1 <- subset(train1, Seed_Variety == 1)

# Print the head of the filtered dataset to verify
head(train1_filtered_1)

##       Soil_Quality Seed_Variety Fertilizer_Amount_kg_per_hectare
Sunny_Days
## 8179      73.79380            1                         286.87839
78.39382
## 2685      83.96964            1                         208.89274
103.32797
## 5640      52.26928            1                          94.13869
105.75976
## 2945      52.95116            1                         264.06048
82.48267
## 6114      51.39387            1                         283.45747
113.54430
## 10929     65.27774            1                          91.05633
106.52755
##       Rainfall_mm Irrigation_Schedule Yield_kg_per_hectare
## 8179     510.9672                  7             903.6764
## 2685     653.4640                  3             705.3779
## 5640     746.5282                  5             522.3417
## 2945     466.4019                  8             996.3509
## 6114     453.8228                  7            1090.2711
## 10929    503.8265                  7             718.7664
```

our descriptive part include in activity 2

regression model creating part

```
# Sample 100 random rows from the cleaned_data data frame
set.seed(123) # Set seed for reproducibility
random_sample <- sample_n(cleaned_data, 100)

# Create scatter plot with a regression line for
Fertilizer_Amount_kg_per_hectare vs Yield_kg_per_hectare
ggplot(random_sample, aes(x = Fertilizer_Amount_kg_per_hectare, y =
Yield_kg_per_hectare)) +
  geom_point(color = "blue") + # Scatter plot
  geom_smooth(method = "lm", color = "red", se = FALSE) + # Regression line
  labs(title = "Scatter Plot of Fertilizer Amount vs Yield with Regression
Line",
       x = "Fertilizer Amount (kg per hectare)",
       y = "Yield (kg per hectare)") +
  theme_minimal()

## `geom_smooth()` using formula = 'y ~ x'
```

RESULTS AND DISCUSSON

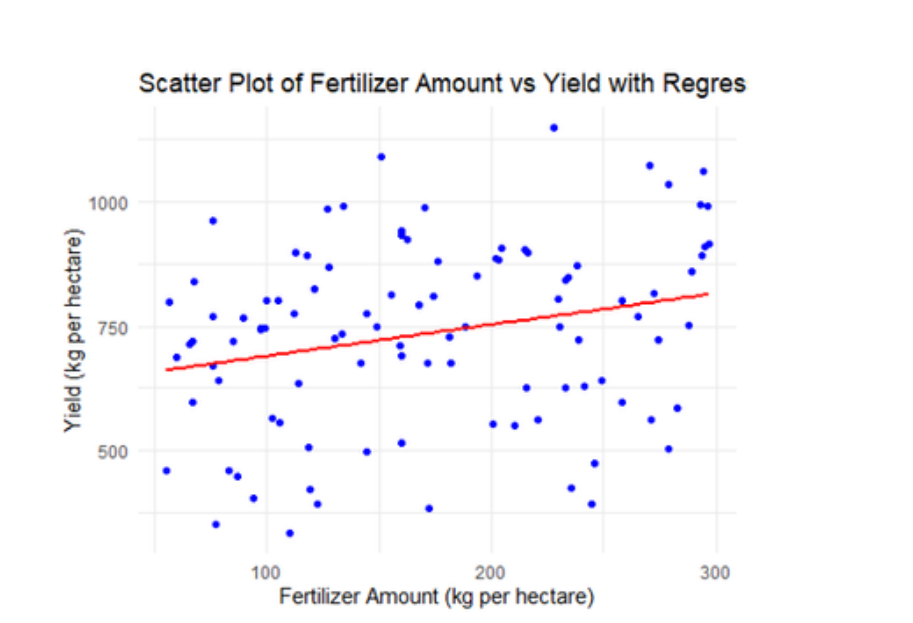
```
## 4265     366.0212                  3             519.3002
## 8489     351.2376                  3             490.3949

# Filter the train1 dataset where Seed_Variety is 1
train1_filtered_1 <- subset(train1, Seed_Variety == 1)

# Print the head of the filtered dataset to verify
head(train1_filtered_1)

##       Soil_Quality Seed_Variety Fertilizer_Amount_kg_per_hectare
Sunny_Days
## 8179      73.79380            1                         286.87839
78.39382
## 2685      83.96964            1                         208.89274
103.32797
## 5640      52.26928            1                          94.13869
105.75976
## 2945      52.95116            1                         264.06048
82.48267
## 6114      51.39387            1                         283.45747
113.54430
## 10929     65.27774            1                          91.05633
106.52755
##       Rainfall_mm Irrigation_Schedule Yield_kg_per_hectare
## 8179     510.9672                  7             903.6764
## 2685     653.4640                  3             705.3779
## 5640     746.5282                  5             522.3417
## 2945     466.4019                  8             996.3509
## 6114     453.8228                  7            1090.2711
## 10929    503.8265                  7             718.7664
```

our descriptive part include in activity 2

regression model creating part

```
# Sample 100 random rows from the cleaned_data data frame
set.seed(123) # Set seed for reproducibility
random_sample <- sample_n(cleaned_data, 100)

# Create scatter plot with a regression line for
Fertilizer_Amount_kg_per_hectare vs Yield_kg_per_hectare
ggplot(random_sample, aes(x = Fertilizer_Amount_kg_per_hectare, y =
Yield_kg_per_hectare)) +
  geom_point(color = "blue") + # Scatter plot
  geom_smooth(method = "lm", color = "red", se = FALSE) + # Regression line
  labs(title = "Scatter Plot of Fertilizer Amount vs Yield with Regression
Line",
       x = "Fertilizer Amount (kg per hectare)",
       y = "Yield (kg per hectare)") +
  theme_minimal()

## `geom_smooth()` using formula = 'y ~ x'
```

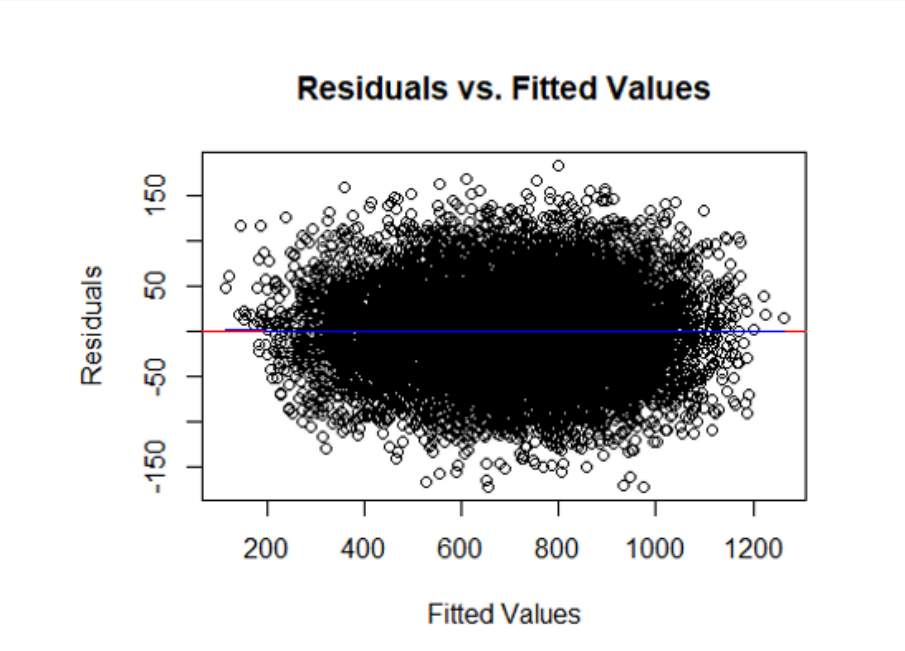
RESULTS AND DISCUSSON



The red line represents the best-fit line through the data points. While the data points are scattered, there's a slight upward trend indicated by the regression line. This suggests a weakly positive correlation between fertilizer amount and yield.

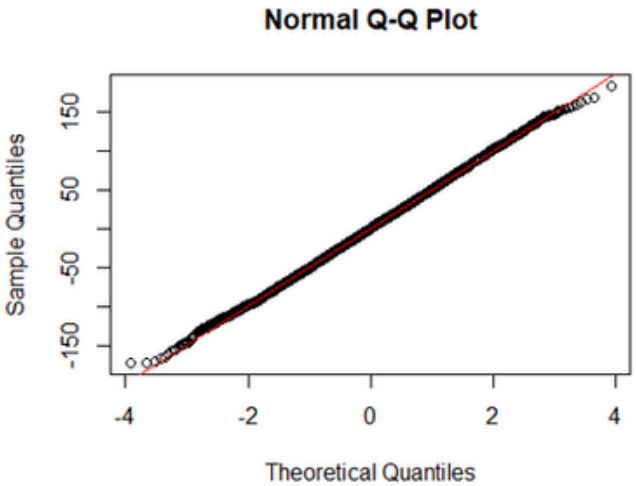
```r
# Fit the model
model <- lm(Yield_kg_per_hectare ~ Fertilizer_Amount_kg_per_hectare +
            Seed_Variety + Rainfall_mm + Irrigation_Schedule + Soil_Quality +
            Sunny_Days, data = train1)
#check the Diagnosing
#Diagnosing Heteroscedasticity
# Plot residuals vs. fitted values
#This plot helps detect non-linearity, unequal error variances
(heteroscedasticity), and outliers.
plot(model$fitted.values, model$residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs. Fitted Values")
abline(h = 0, col = "red")

# Add a smoothed line to identify patterns
lines(lowess(model$fitted.values, model$residuals), col = "blue")
```

RESULTS AND DISCUSSON



The residuals are randomly distributed. There are no clear patterns, curves. This indicates that the linear relationship between the variables is appropriate and there is no evidence of non-linearity. The spread of the points is relatively constant suggests that the variance of the errors is constant.

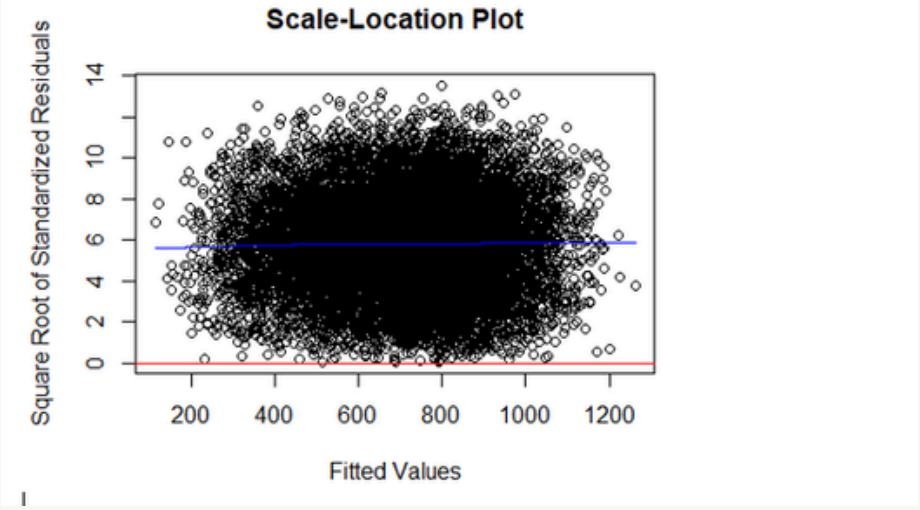
```
#Normal Q-Q Plot
qqnorm(model$residuals, main = "Normal Q-Q Plot")
qqline(model$residuals, col = "red")
```



The points on the plot mostly follow the red line .This plot indicates that the residuals of model are approximately normally distributed, with a few outliers present.

RESULTS AND DISCUSSON

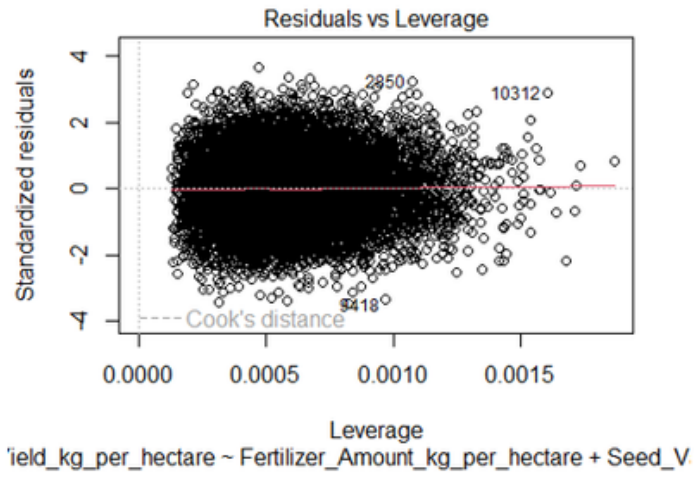
```r
#Scale-Location Plot (Spread-Location Plot)
plot(model$fitted.values, sqrt(abs(model$residuals)),
     xlab = "Fitted Values",
     ylab = "Square Root of Standardized Residuals",
     main = "Scale-Location Plot")
abline(h = 0, col = "red")
lines(lowess(model$fitted.values, sqrt(abs(model$residuals))), col = "blue")
```



The points representing the square root of standardized residuals are scattered randomly around the red line without any clear patterns. This further supports the conclusion of constant variance. the scale-location plot suggests that the linear regression model is appropriate for the data and that the constant variance assumption holds

The points are randomly scattered without any patterns. There are a few points with relatively large standardized residuals.This plot suggests that the model is reasonably well-behaved.

```r
#Residuals vs. Leverage Plot

plot(model, which = 5)
```

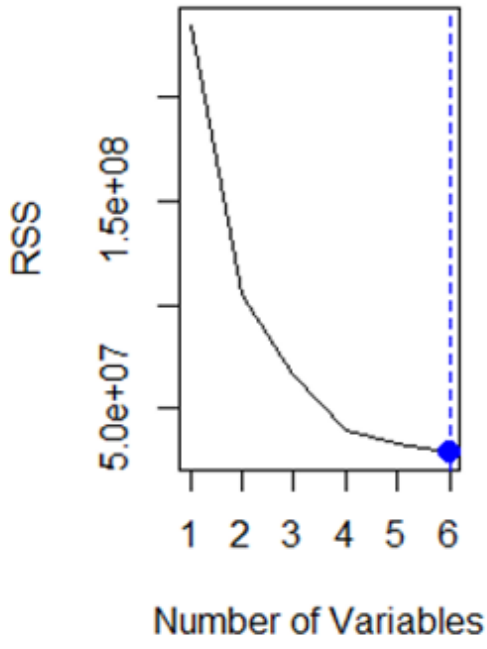


RESULTS AND DISCUSSON

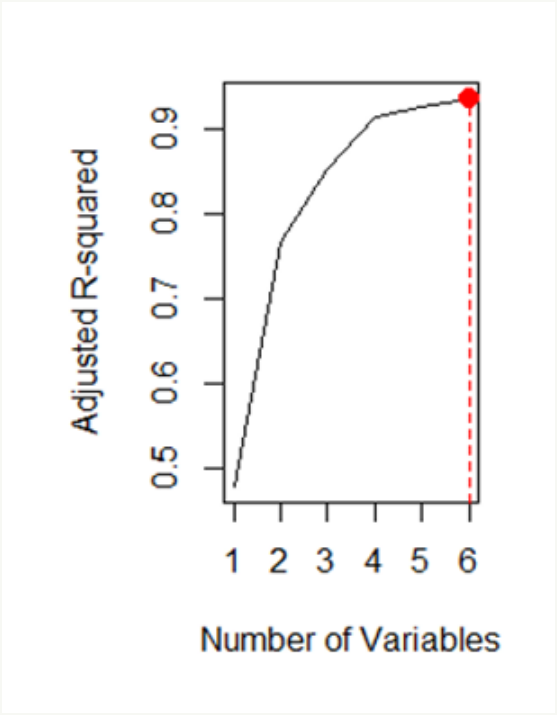
```r
#forwer selcsion method
library(leaps)

## Warning: package 'leaps' was built under R version 4.4.1

# Fit the regsubsets model with all predictors including Seed_Variety
Model_forward<- regsubsets(Yield_kg_per_hectare ~
Fertilizer_Amount_kg_per_hectare + Seed_Variety+ Rainfall_mm +
Irrigation_Schedule + Soil_Quality+Sunny_Days, data = train1,nvmax =
6,method="forward")  # nvmax is the maximum number of predictors to include

# Get the summary of the model
model_with_seed_summary <- summary(Model_forward)

# Display key information from the summary
print(model_with_seed_summary)

## Subset selection object
## Call: regsubsets.formula(Yield_kg_per_hectare ~
Fertilizer_Amount_kg_per_hectare +
##     Seed_Variety + Rainfall_mm + Irrigation_Schedule + Soil_Quality +
##     Sunny_Days, data = train1, nvmax = 6, method = "forward")
## 6 Variables  (and intercept)
```

```r
= 2, pch = 20)
abline(v = adjr2_max, col = "red", lty = 2)
```

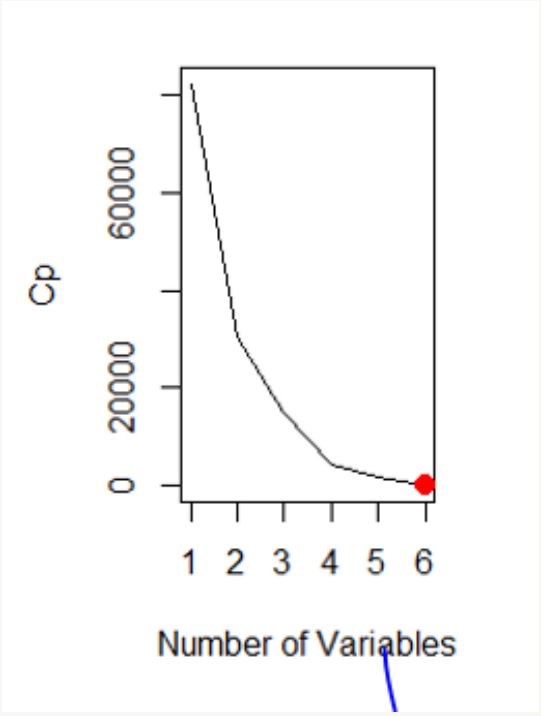
The RSS generally decreases as the number of variables increases. This is expected because adding more variables to a model typically explains more of the variation in the data. The vertical line with a blue dot likely represents the chosen model based on this plot.



RESULTS AND DISCUSSON

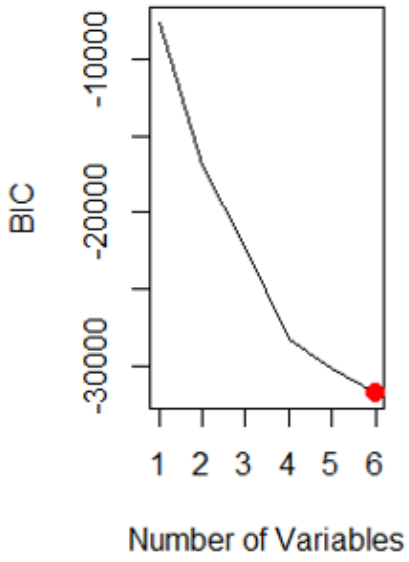



```r
par(mfrow = c(1, 2))
plot(model_with_seed_summary$cp, xlab = "Number of Variables", ylab = "Cp",
type = "l")
cp.min <- which.min(model_with_seed_summary$cp)
points(cp.min, model_with_seed_summary$cp[cp.min], col = "red", cex = 2, pch
= 20)
bic.min <- which.min(model_with_seed_summary$bic)
plot(model_with_seed_summary$bic, xlab = "Number of Variables", ylab = "BIC",
type = "l")
points(bic.min, model_with_seed_summary$bic[bic.min], col = "red", cex = 2,
pch = 20)
```



RESULTS AND DISCUSSON

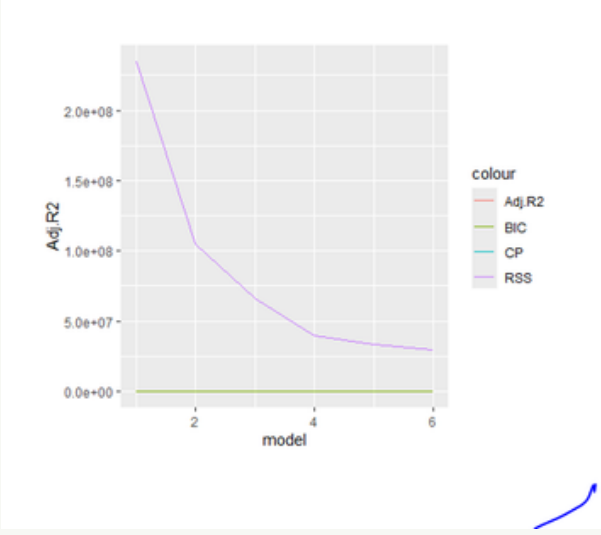
The BIC value decreases rapidly as the number of variables increases initially, suggesting that adding more variables improves model fit. The red dot marks the model with the lowest BIC value, which is often considered the optimal model




```
res.sum <- summary(Model_forward)
criterion<-data.frame(
model=1:6,
Adj.R2 = (res.sum$adjr2),
CP = (res.sum$cp),
BIC = (res.sum$bic),
RSS=res.sum$rss
)
head(criterion)
```

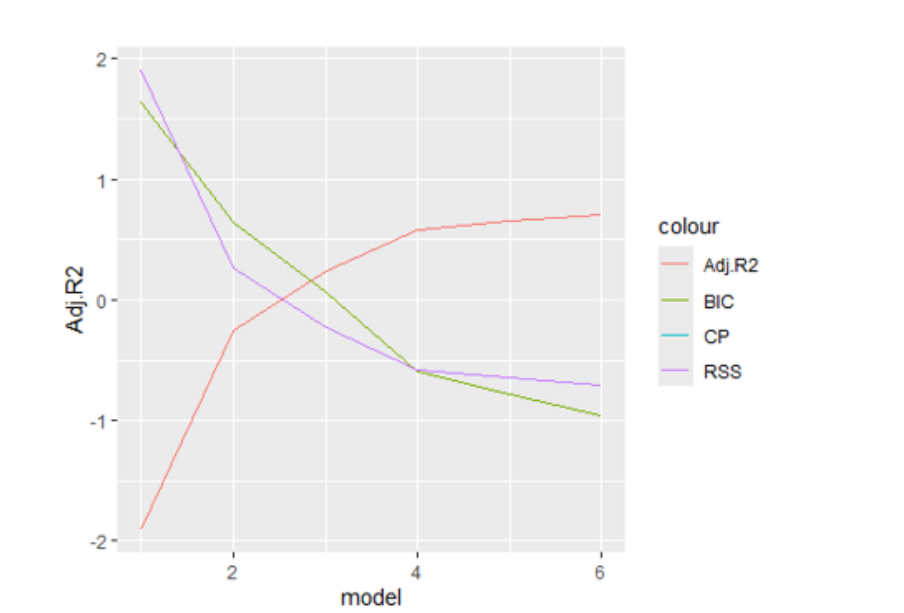
```
##   model    Adj.R2        CP        BIC       RSS
## 1     1 0.4783875 82318.723  -7555.337 234753559
## 2     2 0.7667792 30373.167 -16913.251 104952816
## 3     3 0.8528503 14871.444 -22263.678  66213847
## 4     4 0.9127561  4083.971 -28337.945  39254317
## 5     5 0.9255808  1775.468 -30179.645  33481091
## 6     6 0.9354083     7.000 -31819.256  29057245
```

```
library(ggplot2)
ggplot(criterion, aes(model)) +
 geom_line(aes(y = Adj.R2, colour = "Adj.R2")) +
 geom_line(aes(y = CP, colour = "CP"))+
 geom_line(aes(y = BIC, colour = "BIC"))+
 geom_line(aes(y = RSS, colour = "RSS"))
```



RESULTS AND DISCUSSON

```
# standardize
criterion_std<-cbind(model=criterion$model, scale(criterion[,-1]))
criterion_std<-as.data.frame(criterion_std)
head(criterion_std)

##   model     Adj.R2         CP        BIC        RSS
## 1     1 -1.9050862  1.9051292  1.6414660  1.9050931
## 2     2 -0.2580365  0.2579539  0.6368103  0.2580227
## 3     3  0.2335289 -0.2336003  0.0623942 -0.2335447
## 4     4  0.5756598 -0.5756672 -0.5897326 -0.5756401
## 5     5  0.6489041 -0.6488691 -0.7874555 -0.6488979
## 6     6  0.7050300 -0.7049465 -0.9634824 -0.7050330

#after stadlize
ggplot(criterion_std, aes(model)) +
 geom_line(aes(y = Adj.R2, colour = "Adj.R2")) +
 geom_line(aes(y = CP, colour = "CP"))+
 geom_line(aes(y = BIC, colour = "BIC"))+
 geom_line(aes(y = RSS, colour = "RSS"))
```



We would ideally choose a model with a relatively high Adjusted R-squared, while keeping BIC, CP, and RSS at reasonably low levels.

RESULTS AND DISCUSSON

```
coef(Model_forward, 6)

##                       (Intercept) Fertilizer_Amount_kg_per_hectare
##                        44.0639085                         0.8076244
##                      Seed_Variety                       Rainfall_mm
##                       300.2585927                        -0.5032444
##               Irrigation_Schedule                      Soil_Quality
##                        49.7783048                         1.5537554
##                        Sunny_Days
##                         2.0301833

#better model
better_model <- lm(Yield_kg_per_hectare ~ Fertilizer_Amount_kg_per_hectare +
Seed_Variety+ Rainfall_mm + Irrigation_Schedule +
Soil_Quality+Sunny_Days,data=train1 )
summary(better_model)

##
## Call:
## lm(formula = Yield_kg_per_hectare ~ Fertilizer_Amount_kg_per_hectare +
##     Seed_Variety + Rainfall_mm + Irrigation_Schedule + Soil_Quality +
##     Sunny_Days, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -172.752  -33.980    0.045   33.440  183.162
##
```

```
## Coefficients:
##                                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                       44.063908   6.160398    7.153 9.01e-13
***
## Fertilizer_Amount_kg_per_hectare   0.807624   0.006455  125.109  < 2e-16
***
## Seed_Variety                     300.258593   1.017991  294.952  < 2e-16
***
## Rainfall_mm                       -0.503244   0.004809 -104.651  < 2e-16
***
## Irrigation_Schedule               49.778305   0.220689  225.559  < 2e-16
***
## Soil_Quality                       1.553755   0.032012   48.537  < 2e-16
***
## Sunny_Days                         2.030183   0.048249   42.077  < 2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.99 on 11629 degrees of freedom
## Multiple R-squared:  0.9354, Adjusted R-squared:  0.9354
## F-statistic: 2.808e+04 on 6 and 11629 DF,  p-value: < 2.2e-16
```

Fitted model

Yield_kg_per_hectare = 44.063908 + 0.807624(Fertilizer_Amount_kg_per_hectare) + 300.258593 (Seed_Variety) - 0.503244 (Rainfall_mm) + 49.778305 (Irrigation_Schedule) + 1.553755(Soil_Quality) + 2.030183 (Sunny_Days)

RESULTS AND DISCUSSON

Fertilizer Amount (β=0.8076):
The coefficient for fertilizer amount is positive and highly significant (p<2e−16p < 2e-16p<2e−16). For every additional kilogram of fertilizer per hectare, the yield is expected to increase by approximately 0.81 kg per hectare, holding other factors constant.

Seed Variety (β=300.26):
The seed variety variable also shows a highly significant positive effect (p<2e−16p < 2e-16p<2e−16) on yield.

Irrigation Schedule (β=49.78):
The positive and highly significant coefficient for irrigation schedule (p<2e−16p < 2e-16p<2e−16) .

Sunny Days (β=2.0302):
The number of sunny days also positively affects yield, with a significant coefficient (p<2e−16p < 2e-16p<2e−16).

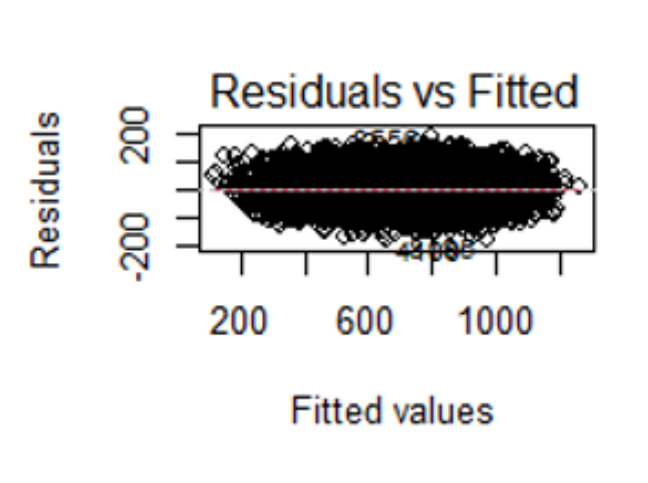
The model's adjusted R2R^2R2 value was 0.9354, indicating that approximately 93.5% of the variability in agricultural yield could be explained by these factors. This high R2R^2R2 value suggests a strong relationship between the predictors and yield, making the model a valuable tool for understanding and predicting crop productivity.

The residual standard error of 49.99 further supports the model's accuracy, indicating that the predictions are reasonably close to the observed values.

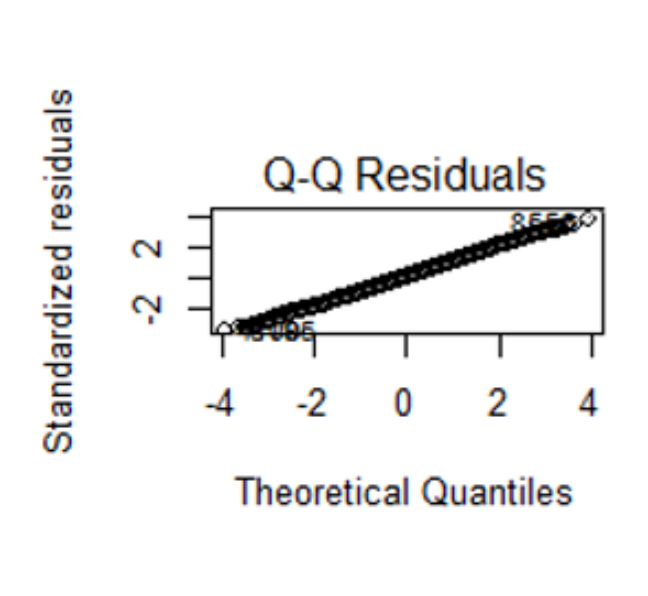
The F-statistic (28,080.00) with a p-value of < 2.2e-16 indicates that the model is highly significant.

Residual Analysis

```
# Model diagnostics
par(mfrow=c(2,2))
plot(better_model)
```

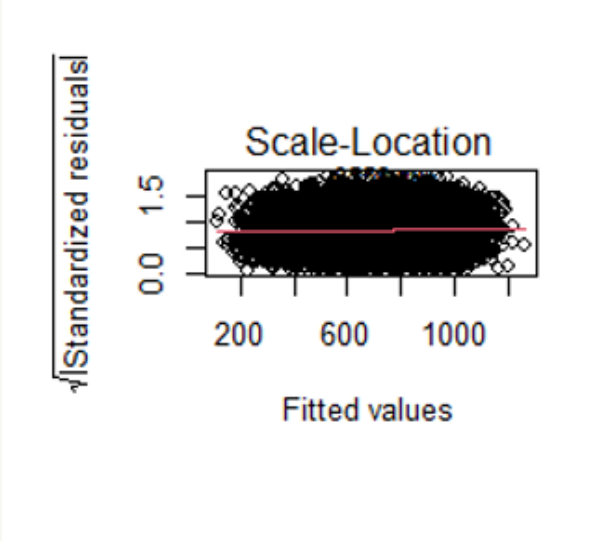


The points are generally scattered randomly. The spread of the points seems relatively constant. There are a few points that deviate significantly from the main cluster. These points could be outliers. The plot suggests that the model assumptions are reasonably met, with the exception of outliers.

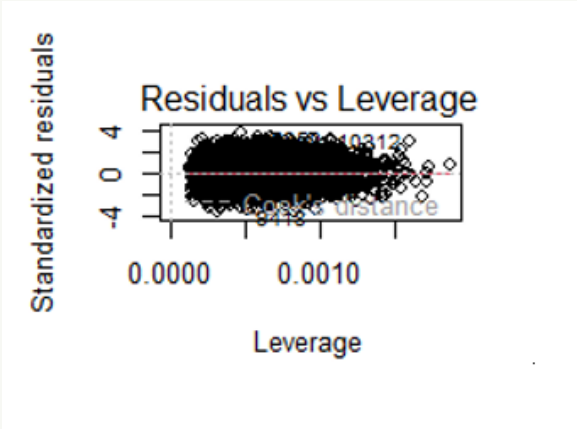


The points deviate from a straight line, especially in the tails. This suggests that the residuals are not normally distributed.

RESULTS AND DISCUSSON



The points are generally scattered randomly without any clear pattern. The scale-location plot indicates that the assumption of homoscedasticity is likely satisfied in this model. The variance of the residuals appears to be constant across different levels of the fitted values.



The points are scattered randomly without any pattern. the plot suggests that there are no major concerns about influential points or heteroscedasticity in the model.

RESULTS AND DISCUSSON

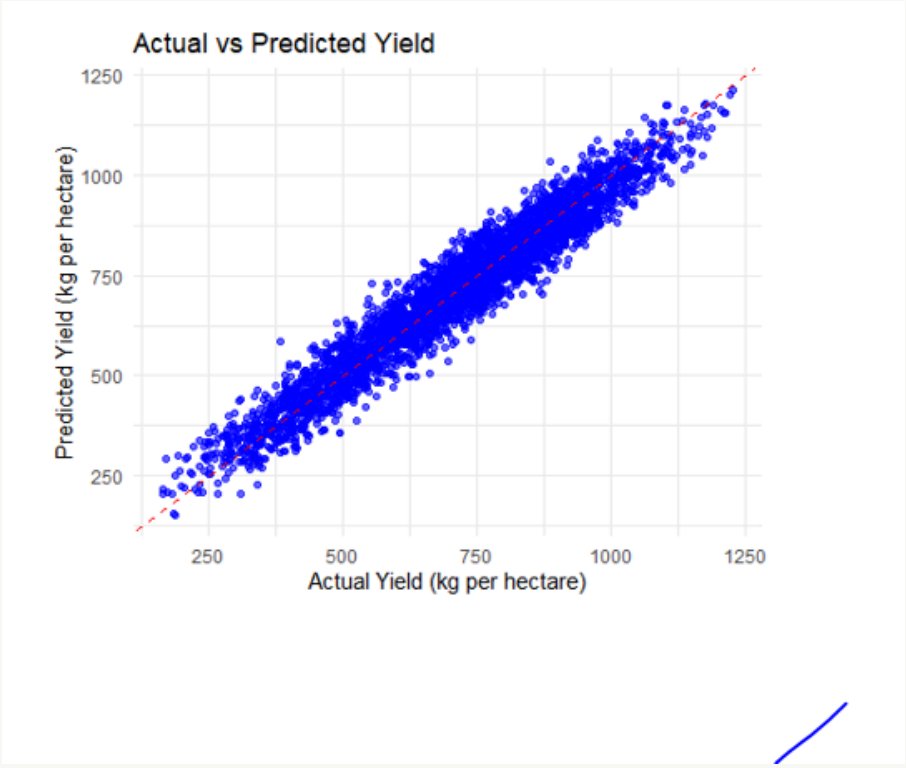
```
#####################################################################

# Make predictions on the test data
test_predictions <- predict(better_model, newdata = test1)



#mean yeil of the focat yeil
mean(test_predictions)

## [1] 711.9286

#mean of the actual yeild
mean(test1$Yield_kg_per_hectare)

## [1] 712.5346

# Calculate performance metrics
mae <- mean(abs(test_predictions - test1$Yield_kg_per_hectare))
rmse <- sqrt(mean((test_predictions - test1$Yield_kg_per_hectare)^2))
r_squared <- cor(test_predictions, test1$Yield_kg_per_hectare)^2

cat("Test MAE:", mae, "\nTest RMSE:", rmse, "\nTest R-squared:", r_squared,
"\n")

## Test MAE: 39.78421
## Test RMSE: 49.59865
## Test R-squared: 0.9372522

#focat_value add teast one data set

test1$focat_yeild=test_predictions
head(test1)
```

```
##    Soil_Quality Seed_Variety Fertilizer_Amount_kg_per_hectare Sunny_Days
## 8      69.33589            1                         135.92277  119.82700
## 16     77.14188            0                         286.16030   89.06296
## 17     71.57169            0                          91.82663  101.39229
## 19     78.76663            1                         239.54935  101.53371
## 25     83.85147            1                         133.72629   97.19609
## 26     53.63000            1                         298.46915  100.00733
##    Rainfall_mm Irrigation_Schedule Yield_kg_per_hectare focat_yeild
## 8     384.3504                   2             750.3530    711.2333
## 16    650.7772                   5             436.2702    497.2393
## 17    494.6965                   5             372.0941    435.2135
## 19    429.3474                   3             698.1650    799.5728
## 25    547.8296                   6             877.4256    802.9113
## 26    489.8564                   9            1069.6259   1073.2219
```

When applying the model to the test dataset, the Mean Absolute Error (MAE) was 39.78 kg per hectare, and the Root Mean Squared Error (RMSE) was 49.60 kg per hectare. These metrics suggest that the model predictions are generally close to the actual yield values. The adjusted R^2 value on the test data was 0.9373, indicating that the model's predictive power remained strong even on unseen data

RESULTS AND DISCUSSON

```
# Plot actual vs. predicted values
ggplot(test1, aes(x = test1$Yield_kg_per_hectare, y = test1$focat_yeild)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Actual vs Predicted Yield",
       x = "Actual Yield (kg per hectare)",
       y = "Predicted Yield (kg per hectare)") +
  theme_minimal()

## Warning: Use of `test1$Yield_kg_per_hectare` is discouraged.
## ℹ Use `Yield_kg_per_hectare` instead.

## Warning: Use of `test1$focat_yeild` is discouraged.
## ℹ Use `focat_yeild` instead.
```



This plot which is comparing actual versus predicted yields showed that the predictions closely aligned with the observed values, further validating the model's accuracy.

The summary statistics of the predicted yields were similar to those of the actual yields, with mean values of 711.9 kg per hectare for the predictions and 712.5 kg per hectare for the actual yields.

```
#get sumriy actual vs. predicted values
#actua
summary(test1$Yield_kg_per_hectare)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   164.2   571.9   734.2   712.5   858.9  1229.0

#predicted values
summary(test1$focat_yeild)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   151.8   582.5   734.3   711.9   853.3  1212.8
```

CONCLUSION

This study aimed to analyze the effects of environmental factors, fertilizer application, and management practices on agricultural yield, with the goal of providing actionable insights for optimizing farming practices and improving crop productivity. For this we selected wheat as the experimental crop. Through a comprehensive analysis using multiple linear regression, the relationships between key variables such as fertilizer amount, seed variety, rainfall, irrigation schedule, soil quality, and the number of sunny days were quantified and evaluated.

The regression model developed in this study was robust, explaining a significant portion of the variance in agricultural yield with a high adjusted R^2 value. That means the selected factors have approximately 93.5% of the variability in crop yield. The model's coefficients were statistically significant, suggesting strong and meaningful relationships between the independent variables and the dependent variable (yield).

Following conclusions can be made on predictor variables

- **Fertilizer Amount:** The positive and significant coefficient suggests that increasing the amount of fertilizer per hectare gives higher yields, showing the importance of proper fertilization in crop management.

- **Seed Variety:** The high significance of seed variety indicates that choosing the right seed variety is crucial for maximizing yield, potentially overshadowing other environmental factors.

- **Rainfall:** The negative coefficient for rainfall suggests that excessive rainfall may have a harmful effect on yield, showing the need for balanced water management

CONCLUSION

- **Irrigation Schedule:** The significant positive impact of irrigation schedules on yield shows it is important to have strategic water application, especially in regions with variable rainfall patterns.

- **Soil Quality and Sunny Days:** Both factors were found to positively influence yield, showing the importance of soil health and adequate sunlight in crop production.

The residual analysis and multicollinearity checks confirmed the model's reliability, and the low residual standard error (49.99) suggests that the model predictions are relatively precise.

Finally, this study provides strong evidence that a combination of proper fertilizer application, seed variety selection, and well-planned irrigation schedules, along with favorable environmental conditions, can significantly enhance agricultural yield. These findings can serve as a valuable resource for farmers, agronomists, and policymakers aiming to improve agricultural practices and promote sustainable farming.

INDIVIDUAL CONTRIBUTION

Task Completed	PS/2020/023	PS/2020/316	PS/2020/141	PS/2020/258	PS/2020/186	PS/20201853	PS/2020/306	PS/2020/260
data collect								
data analysis								
results discribe								
project report create								
Presentation create								
presant								