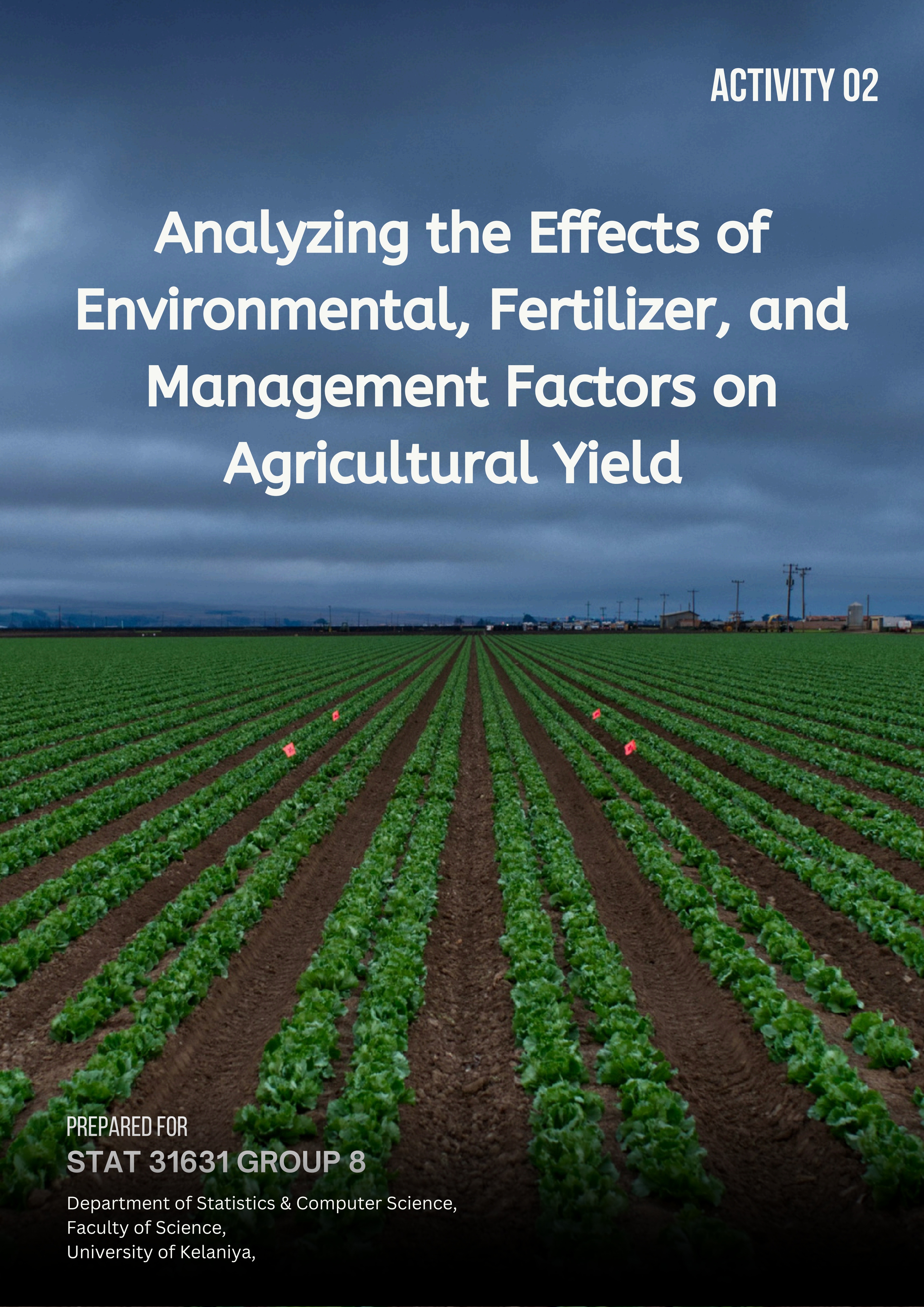


ACTIVITY 02

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

PREPARED FOR

STAT 31631 GROUP 8

Department of Statistics & Computer Science,
Faculty of Science,
University of Kelaniya,

AGENDA

6. METHODOLOGY → 03

7. DESCRIPTIVE ANALYSIS → 07

.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.

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.

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?

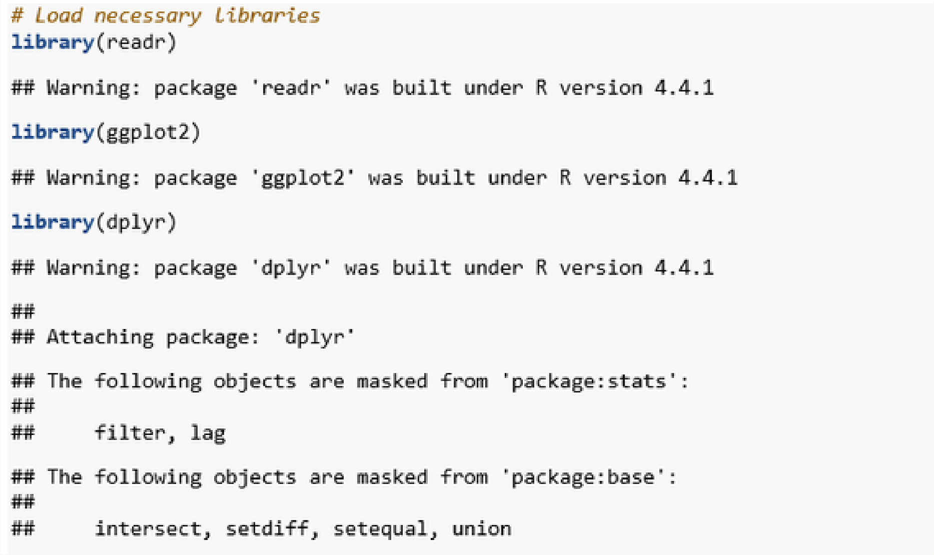
Conclusion

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

DESCRIPTIVE ANALYSIS

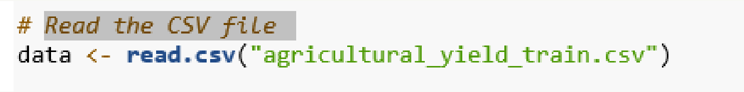
Libraries Loaded:

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
```

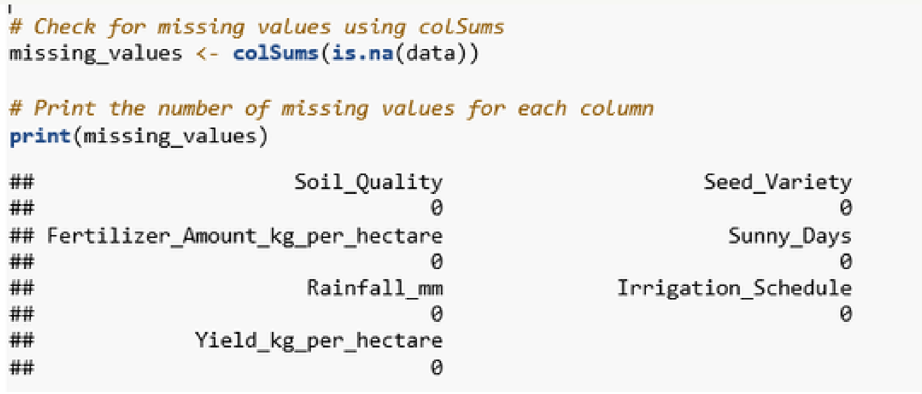
Read the CSV file



```
# 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.

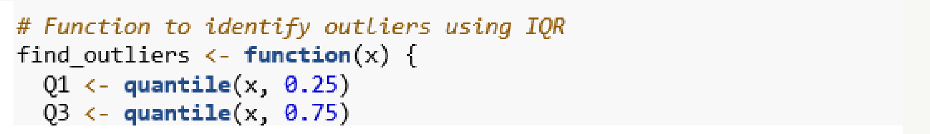
Outliers are identified for each numeric column, and the indices of all outliers are combined.

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

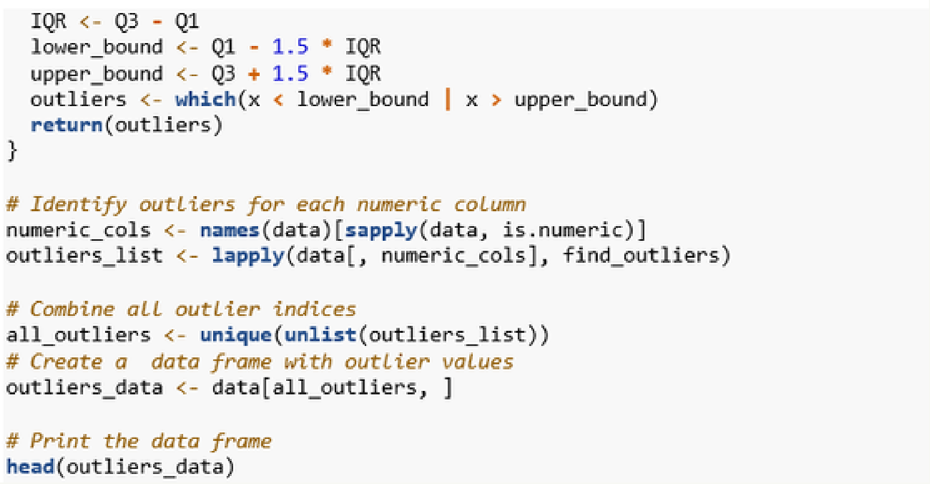
Outliers are removed from the original data to create a cleaned_data data frame.

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

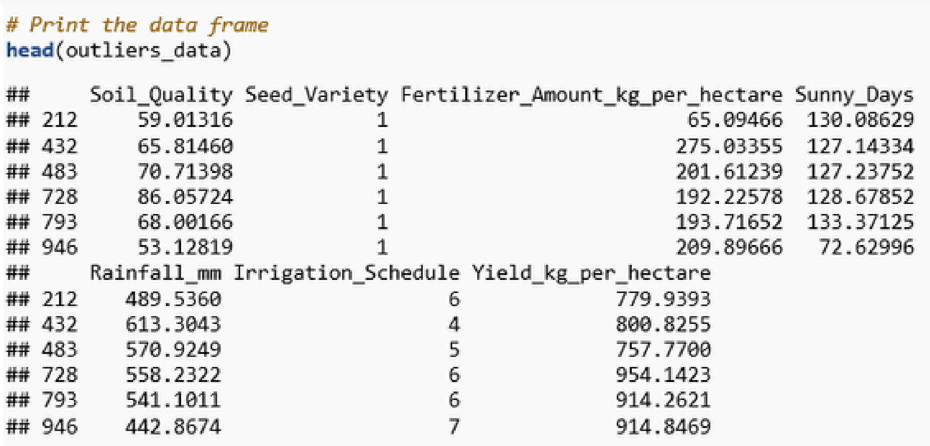
- Before: 16,000 rows
- After: 15,515 rows



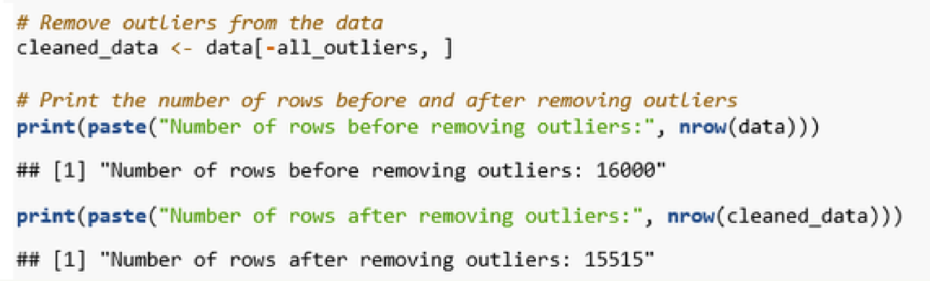
```
# Function to identify outliers using IQR
find_outliers <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
```



```
  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)
```



```
# 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
```

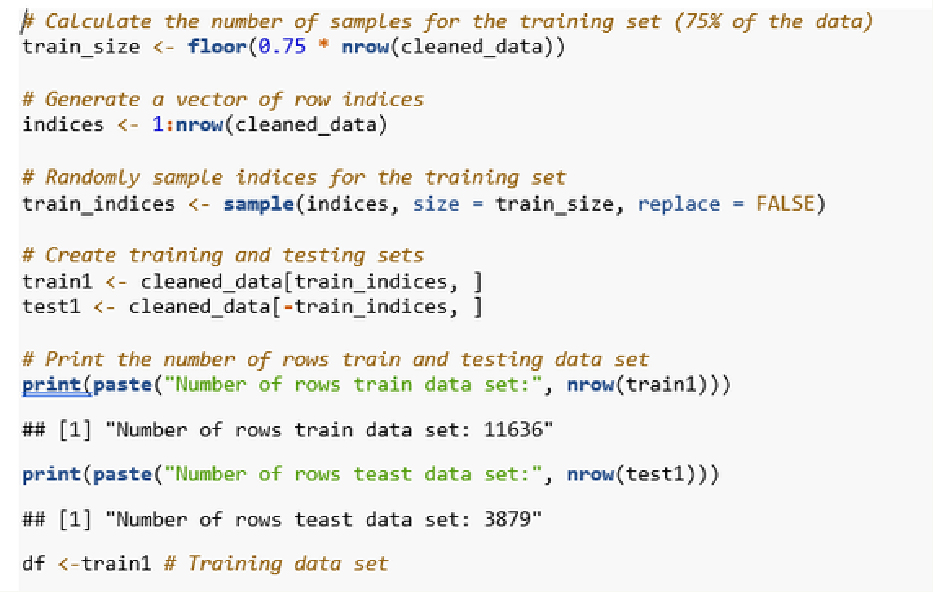


```
# Remove outliers from the data
cleaned_data <- data[-all_outliers, ]

# 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"

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:

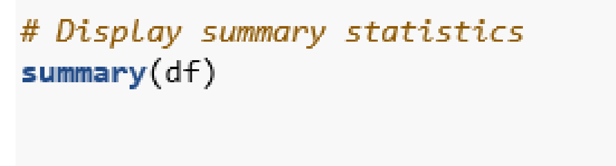
- Training set: 11,636 rows
- Testing set: 3,879 rows



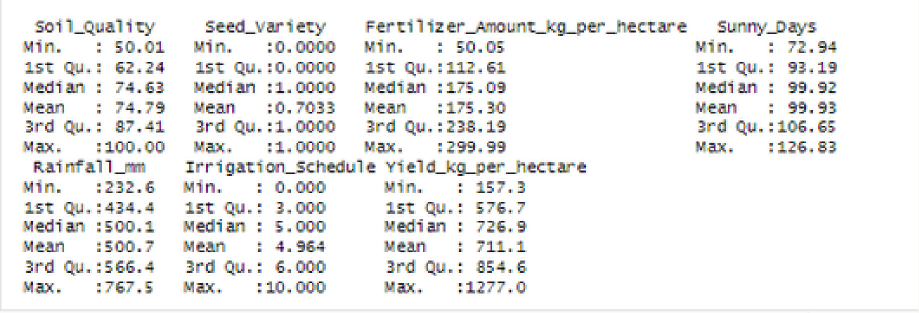
```
# Calculate the number of samples for the training set (75% of the data)
train_size <- floor(0.75 * nrow(cleaned_data))

# 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.



```
# 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.

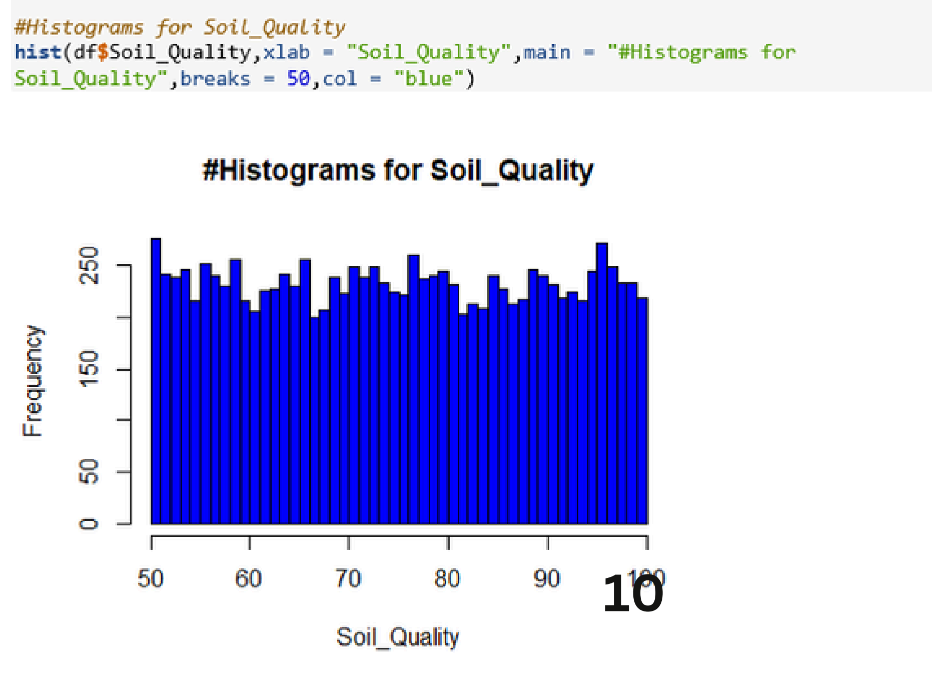
Rainfall ranges from 232.6 to 767.5 mm, with a median of 500.1mm. The mean is almost equal to the median, indicating a symmetric distribution. The interquartile range is 132 mm, suggesting substantial variability in rainfall.

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

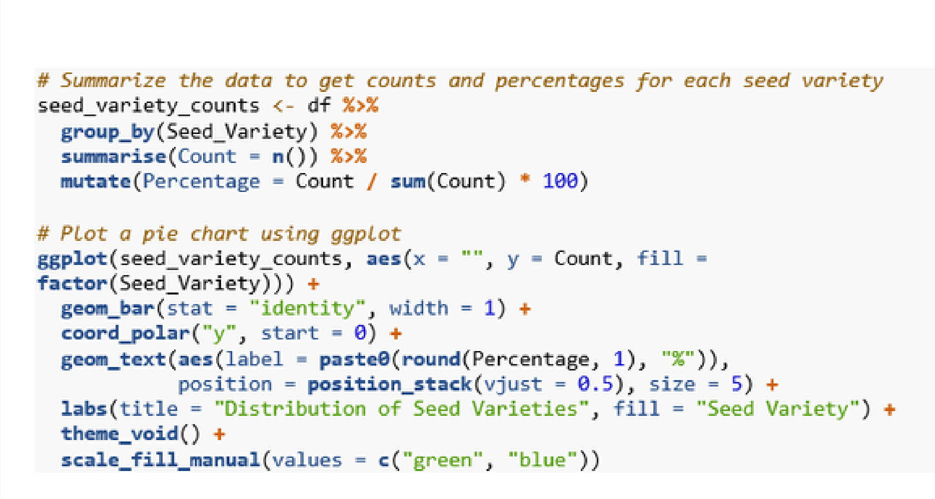
Data Visualization:

- 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

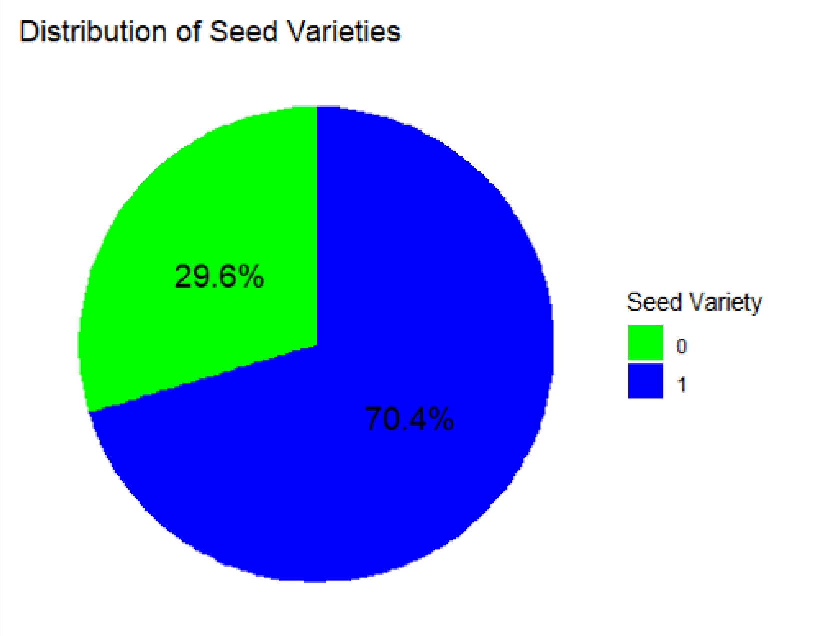


```
#Histograms for Soil_Quality
hist(df$Soil_Quality,xlab = "Soil_Quality",main = "#Histograms for
Soil_Quality",breaks = 50,col = "blue")
```

DESCRIPTIVE ANALYSIS

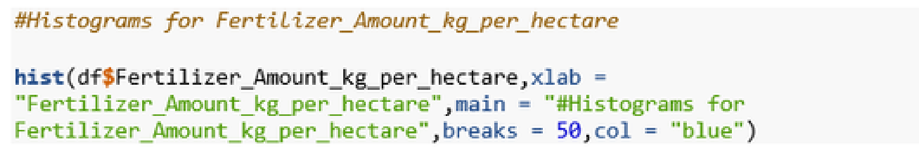


```
# Summarize the data to get counts and percentages for each seed variety
seed_variety_counts <- df %>%
  group_by(Seed_Variety) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = Count / sum(Count) * 100)

# 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"))
```

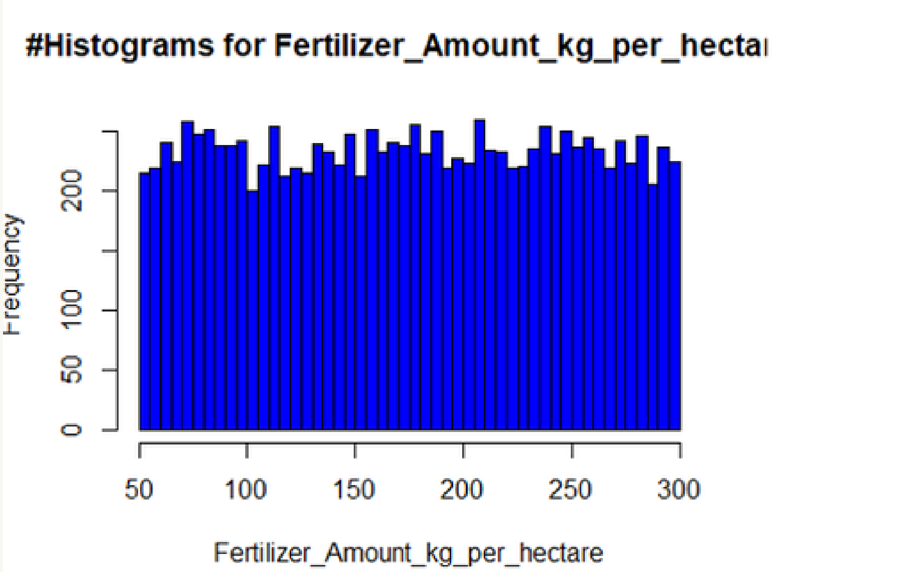


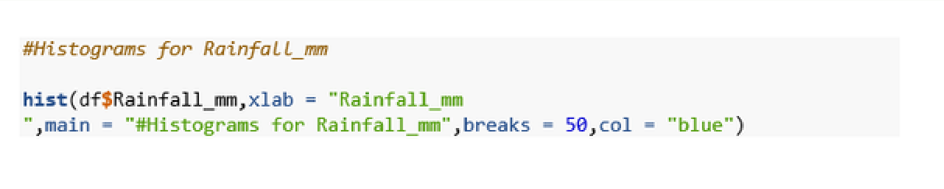
Interpretation:

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.

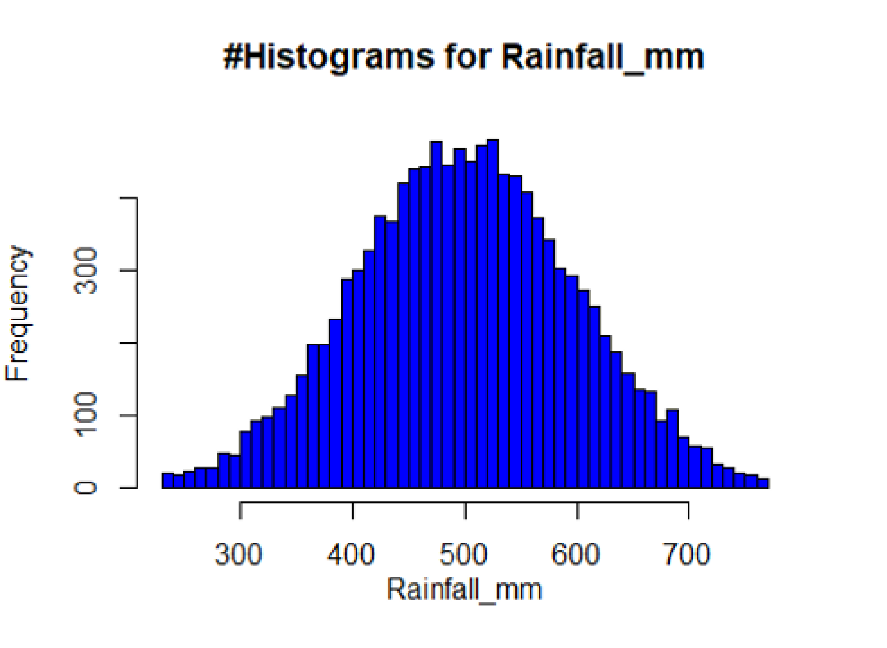


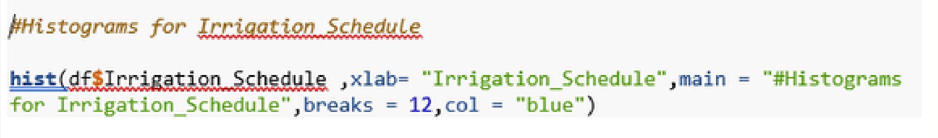
```
#Histograms for Fertilizer_Amount_kg_per_hectare

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")
```



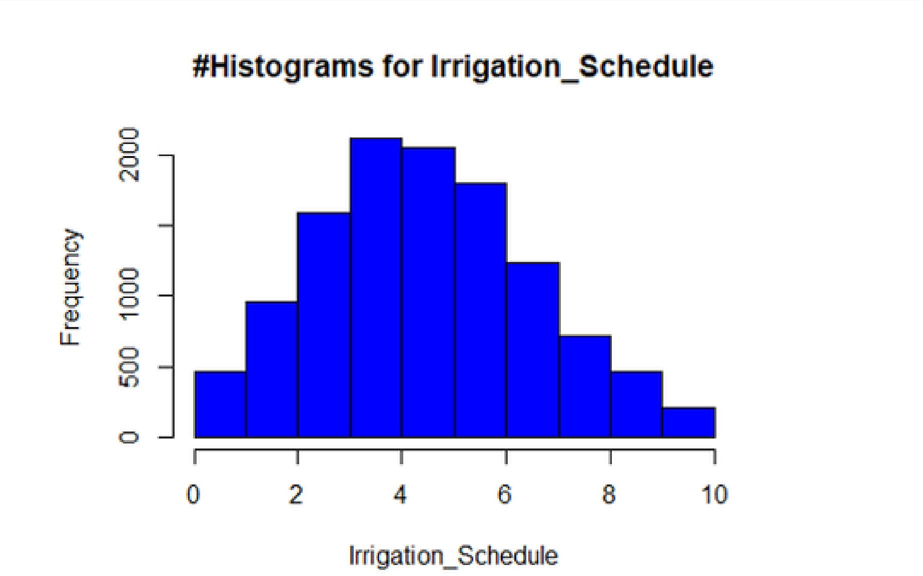


```
#Histograms for Rainfall_mm

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

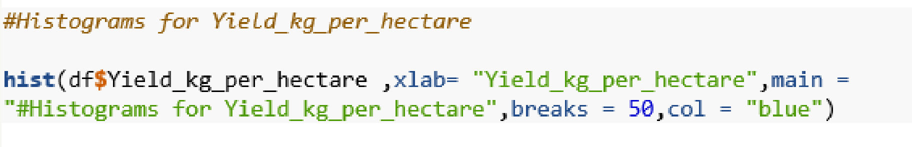




```
#Histograms for Irrigation_Schedule

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



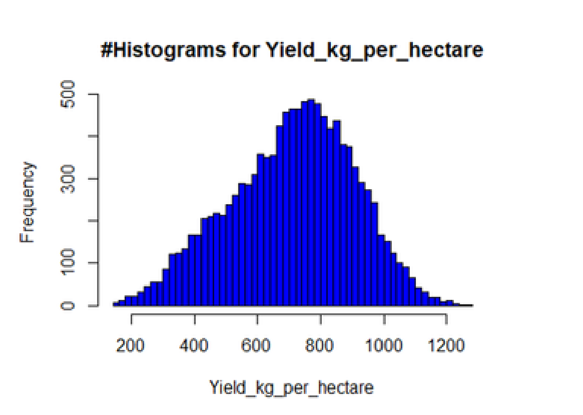
DESCRIPTIVE ANALYSIS



```
#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")
```

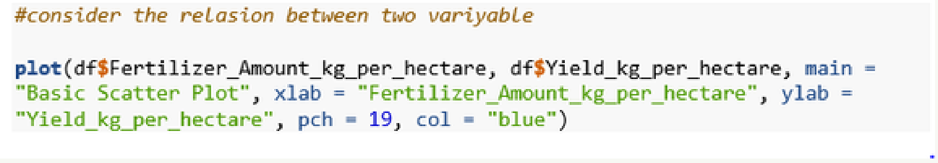
Interpretation

- **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.

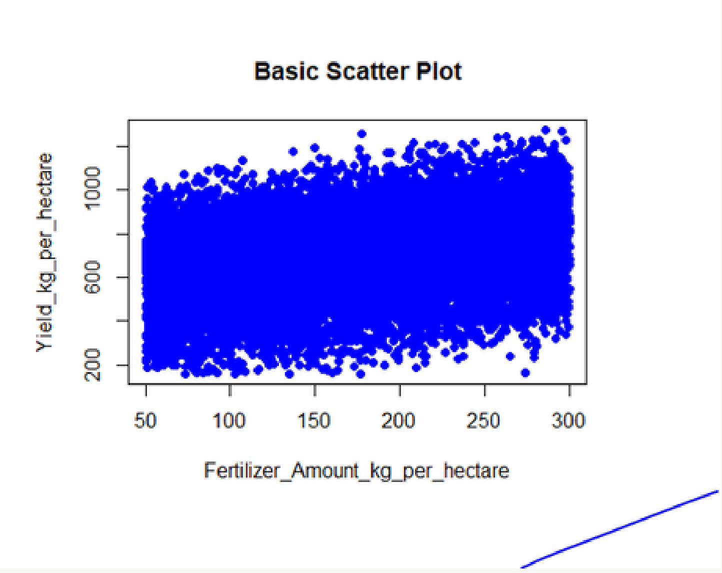


Scatter Plot:

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



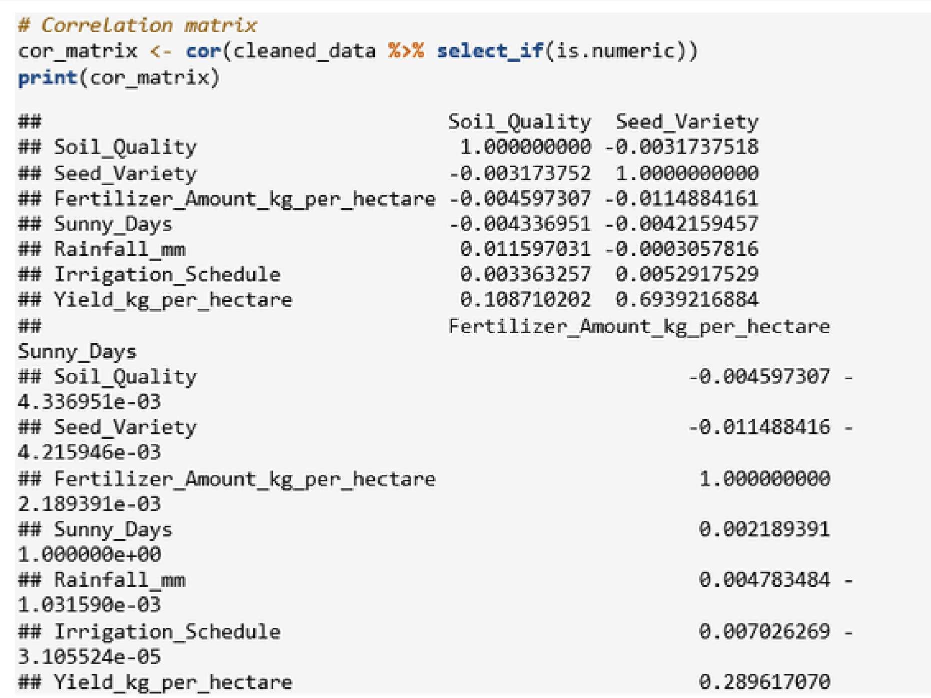
```
#consider the relasion between two variyable

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")
```



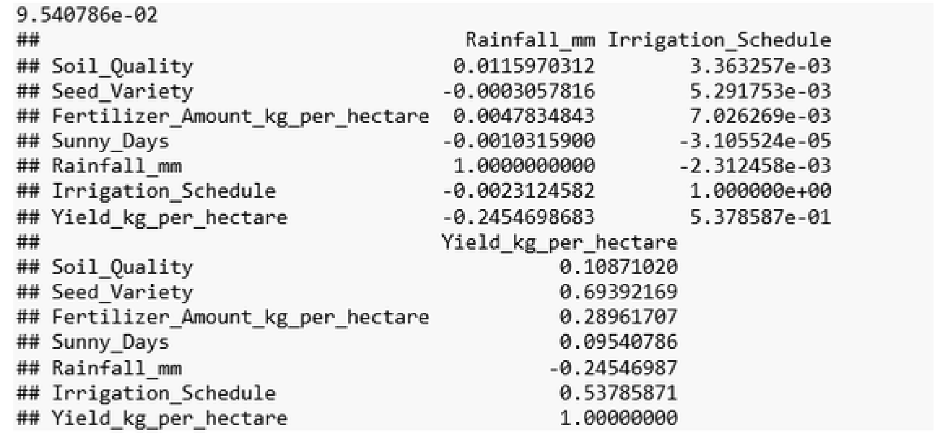
There is a positive relationship between fertilizer amount and crop yield. This suggests that applying more fertilizer generally results in higher yields.

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.

THE TEAM

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