

# EXPLORATORY DATA ANALYSIS OF AN AGRICULTURAL DATASET IN R

*Name: Rhian Manwani*

*SRN: PES1UG21CS486*

*Course: Programming with R*

*Course Code: UE22CS222A*



**PES**  
**UNIVERSITY**

## TABLE OF CONTENTS

<b>Section</b>	<b>Page Number</b>
Abstract	3
Introduction to Dataset	3
Description of Attributes	3
Structure of the Dataset	5
Crop Item Analysis	6
Crop Yield Analysis	9
Effect of Rainfall	14
Effect of Pesticides	16
Effect of Temperature	19
Correlation Analysis	21
Fitting Distributions	23
Conclusion	25
Findings	28
References	29
Appendix (R Program)	30

## Abstract

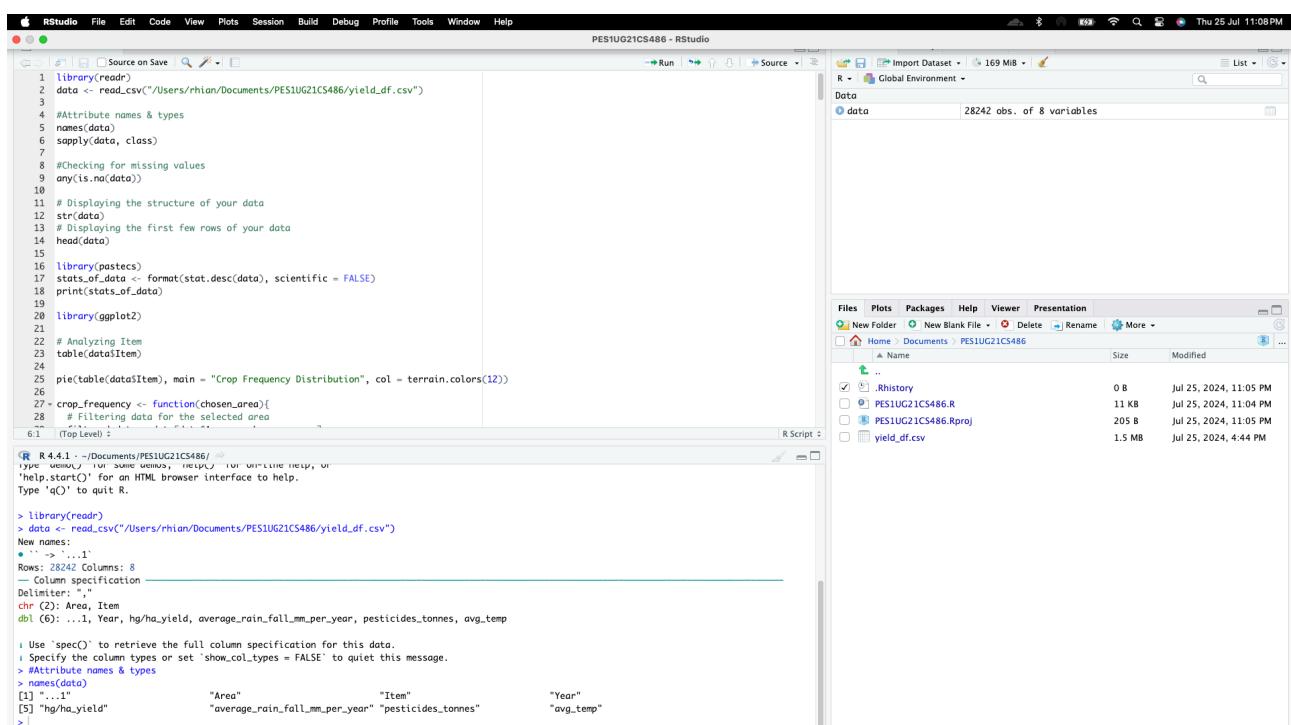
In order to better understand the intricate relationships affecting agricultural productivity, a large dataset comprising information on temperature, pesticide use, rainfall, and crop yield from different locations is analyzed in this study. The study looks at regional variations, temporal trends, and the effects of environmental factors on crop productivity using statistical tools and visualizations. Analysis of correlations is done to find connections between climate factors and yield. The R programming language is used for the entire analysis, and the 'fitdistrplus' tool is used to find the distribution that best fits the yield data. The results emphasize how crucial it is to comprehend how agricultural productivity and climatic conditions are related.

## Introduction to Dataset

Key variables like geographic location, crop type, year, yield (in hectograms per hectare), average annual rainfall (in millimetres), pesticide consumption (in tonnes), and average temperature (in degrees Celsius) are all included in the dataset used in this analysis. The dataset, which spans several nations, offers a thorough understanding of how these variables interact and affect agricultural productivity across time. The historical scope of the dataset extends until 1990, making a detailed examination of trends and patterns possible. With an emphasis on staples like potatoes, rice, maize, and sorghum, the dataset makes it easier to comprehend in depth the intricacies present in agricultural systems.

## Description of Attributes

The dataset comprises of the following attributes:



The screenshot shows the RStudio interface with the following details:

- Code Editor:** Displays R script code for reading a CSV file, checking for missing values, displaying data structure, and creating a pie chart of crop frequency.
- Environment Pane:** Shows the 'data' object with 28242 observations and 8 variables.
- File Browser:** Shows the project structure with files like .Rhistory, PES1UG21CS486.R, PES1UG21CS486.Rproj, and yield\_df.csv.

```

library(readr)
data <- read.csv("~/Users/rhian/Documents/PES1UG21CS486/yield_df.csv")
#Attribute names & types
names(data)
sapply(data, class)
#Checking for missing values
any(is.na(data))
# Displaying the structure of your data
str(data)
# Displaying the first few rows of your data
head(data)
#Displaying the first few rows of your data
stats_of_data <- format(stat.desc(data), scientific = FALSE)
print(stats_of_data)
library(ggplot2)
# Analyzing Item
table(data$item)
# Filtering data for the selected area
crop_frequency <- function(chosen_area){
  # Filtering data for the selected area
}
pie(table(data$item), main = "Crop Frequency Distribution", col = terrain.colors(12))
# Filtering data for the selected area
# Use 'spec()' to retrieve the full column specification for this data.
# Specify the column types or set 'show_col_types = FALSE' to quiet this message.
# Attribute names & types
names(data)
[1] ".," "Area"           "Item"          "Year"          "avg_temp"
[5] "hg/hyield"        "average_rain_fall_mm_per_year" "pesticides_tonnes"
>

```

Following is the data type of these attributes:

The screenshot shows the RStudio interface with the following details:

- Code Editor:** Displays R code for reading a CSV file, displaying its structure, and printing the first few rows.
- Global Environment:** Shows the variable `data` with 28242 observations and 8 variables.
- File Explorer:** Shows the project directory structure including files like `.Rhistory`, `PES1UG21CS486.R`, `PES1UG21CS486.Rproj`, and `yield_df.csv`.
- Console:** Displays the output of the R code, including the structure of the `data` frame and its columns.

## Check for missing values:

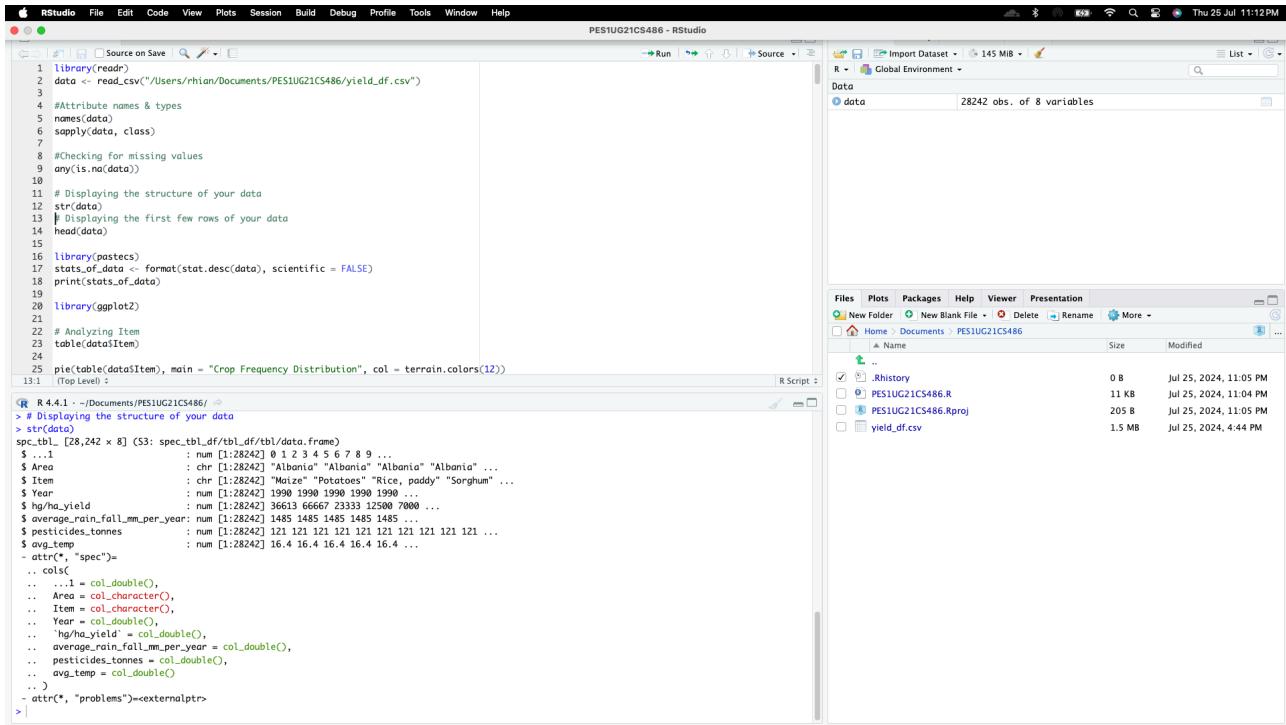
The screenshot shows the RStudio interface with the following details:

- Code Editor:** Displays the same R code as the previous screenshot, including the check for missing values.
- Global Environment:** Shows the variable `data` with 28242 observations and 8 variables.
- File Explorer:** Shows the project directory structure including files like `.Rhistory`, `PES1UG21CS486.R`, `PES1UG21CS486.Rproj`, and `yield_df.csv`.
- Console:** Displays the output of the R code, including the structure of the `data` frame and its columns, followed by the result of the `any(is.na(data))` command which returns `[1] FALSE`.

No missing values found in the dataset.

# Structure of Dataset

Following is the structure of the dataset



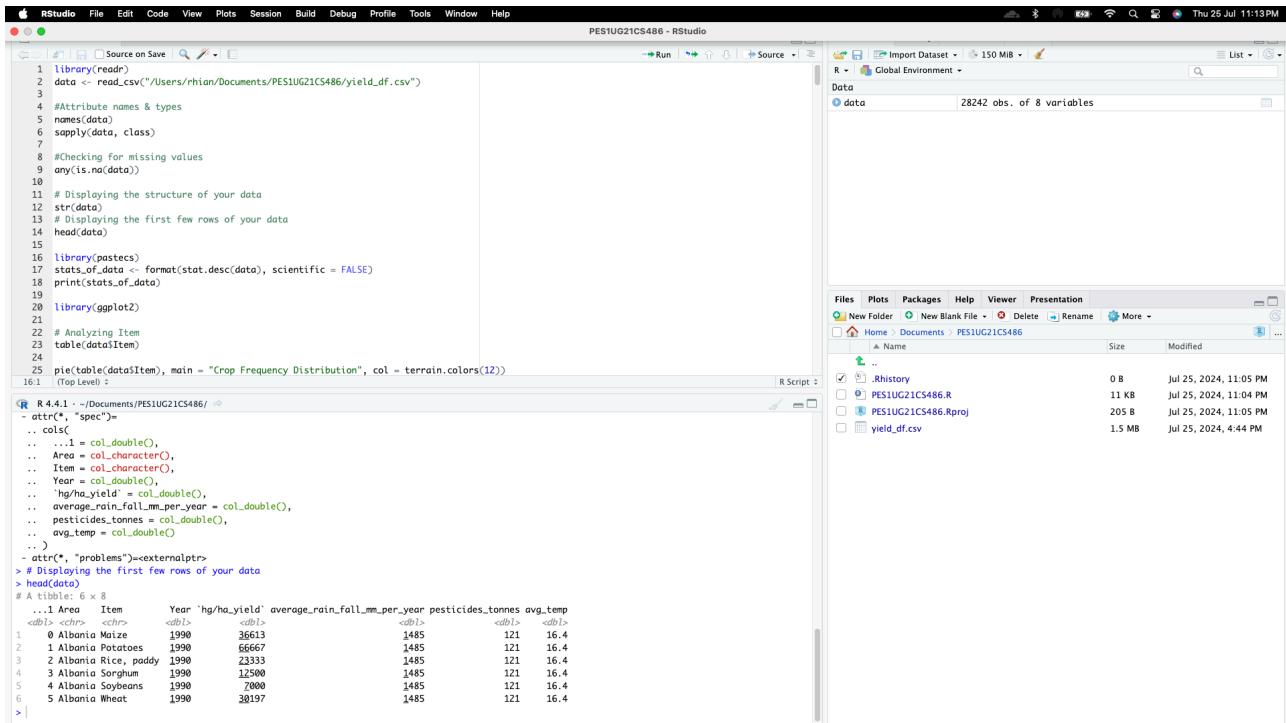
```

library(readr)
data <- read_csv("/Users/rhian/Documents/PES1UG21CS486/yield_df.csv")
# Attribute names & types
names(data)
sapply(data, class)
# Checking for missing values
any(is.na(data))
# Displaying the structure of your data
str(data)
# Displaying the first few rows of your data
head(data)
library(ggplot2)
stats_of_data <- format(stat.desc(data), scientific = FALSE)
print(stats_of_data)
library(ggplot2)
# Analyzing Item
table(data$item)
pie(table(data$item), main = "Crop Frequency Distribution", col = terrain.colors(12))

```

R 4.4.1 - ~/Documents/PES1UG21CS486.R  
> # Displaying the structure of your data  
> str(data)  
R Dataframe: data  
 28242 obs. of 8 variables  
 \$ Area : chr [1:28242] "Albania" "Albania" "Albania" "Albania" ...  
 \$ Item : chr [1:28242] "Maize" "Potatoes" "Rice, paddy" "Sorghum" ...  
 \$ Year : num [1:28242] 1990 1990 1990 1990 1990 ...  
 \$ hg\_ha\_yield : num [1:28242] 36613 66667 23333 12500 7000 ...  
 \$ average\_rain\_fall\_mm\_per\_year : num [1:28242] 1485 1485 1485 1485 1485 ...  
 \$ pesticides\_tonnes : num [1:28242] 121 121 121 121 121 ...  
 \$ avg\_temp : num [1:28242] 16.4 16.4 16.4 16.4 16.4 ...  
- attr(\*, "spec")<-  
 cols<-  
... .1 : col\_double(),  
... Area = col\_character(),  
... Item = col\_character(),  
... Year = col\_double(),  
... hg\_ha\_yield = col\_double(),  
... average\_rain\_fall\_mm\_per\_year = col\_double(),  
... pesticides\_tonnes = col\_double(),  
... avg\_temp = col\_double()  
... )  
- attr(\*, "problems")=<externalptr>

The first few rows of the dataset:



```

library(readr)
data <- read_csv("/Users/rhian/Documents/PES1UG21CS486/yield_df.csv")
# Attribute names & types
names(data)
sapply(data, class)
# Checking for missing values
any(is.na(data))
# Displaying the structure of your data
str(data)
# Displaying the first few rows of your data
head(data)
library(ggplot2)
stats_of_data <- format(stat.desc(data), scientific = FALSE)
print(stats_of_data)
library(ggplot2)
# Analyzing Item
table(data$item)
pie(table(data$item), main = "Crop Frequency Distribution", col = terrain.colors(12))

```

R 4.4.1 - ~/Documents/PES1UG21CS486.R  
> # Displaying the first few rows of your data  
> head(data)  
# A tibble: 6 × 8  
#> Area Item Year `hg\_ha\_yield` average\_rain\_fall\_mm\_per\_year pesticides\_tonnes avg\_temp  
#> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 1 Albania Maize 1990 36613 1485 121 16.4  
2 1 Albania Potatoes 1990 66667 1485 121 16.4  
3 2 Albania Rice, paddy 1990 23333 1485 121 16.4  
4 3 Albania Sorghum 1990 12500 1485 121 16.4  
5 4 Albania Soybeans 1990 2000 1485 121 16.4  
6 5 Albania Wheat 1990 38197 1485 121 16.4

Following is a statistical description of the dataset:

```

1 library(readr)
2 data <- read_csv("~/Users/rhian/Documents/PES1UG21CS486/yield_df.csv")
3
4 #Attribute names & types
5 names(data)
6 sapply(data, class)
7
8 #Checking for missing values
9 any(is.na(data))
10
11 # Displaying the structure of your data
12 str(data)
13 # Displaying the first few rows of your data
14 head(data)
15
16 library(pastecs)
17 stats_of_data <- format(stat.desc(data), scientific = FALSE)
18 print(stats_of_data)
19
20 library(ggplot2)
21
22 # Analyzing Item
23 table(data$item)
24
25 pie(table(data$item), main = "Crop Frequency Distribution", col = terrain.colors(12))

```

(Top Level) :

```

R 4.4.1 - ~/Documents/PES1UG21CS486/ 
2 1 Albania Potatoes 1990 66667 1485 121 16.4
3 2 Albania Rice, paddy 1990 23333 1485 121 16.4
4 3 Albania Sorghum 1990 12500 1485 121 16.4
5 4 Albania Soybeans 1990 2000 1485 121 16.4
6 5 Albania Wheat 1990 38197 1485 121 16.4
> library(pastecs)
> stats_of_data <- format(stat.desc(data), scientific = FALSE)
> print(stats_of_data)
..._1 Area Item Year hg/ha_yield average_rain_fall_mm_per_year pesticides_tonnes avg_temp
nbr.val 28242.0000000 NA NA 28242.0000000 28242.0000000 28242.0000000
nbr.null 1.0000000 NA NA 0.0000000 0.0000000 0.0000000
nbr.na 0.0000000 NA NA 0.0000000 0.0000000 0.0000000
min 1990.0000000 NA NA 1990.0000000 50.0000000 51.0000000
max 28241.0000000 NA NA 2013.0000000 501412.0000000 3240.0000000
range 28241.0000000 NA NA 2000.0000000 501362.0000000 3189.0000000
sum 398791161.0000000 NA NA 56527614.000000000 2176140205.0000000 32451639.0000000 1047126073.679950 580164.860000000
median 14120.5000000 NA NA 2001.544295730 38295.0000000 1083.0000000 17529.4400000 20.54262658
mean 14120.5000000 NA NA 0.041962248 501.532047 4.223726 356.783776 0.03755976
SE.mean 48.5137438 NA NA 0.041962248 501.532047 8.278705 699.313322 0.07361892
CI..mean.0.95 95.0892659 NA NA 0.082248020 990.686836 7.278705 699.313322 0.07361892
var 66469000.5000000 NA NA 49.729368154 7217626074.874197 503833.288216 3595055858.510375 39.84198576
std.dev 8152.9074875 NA NA 7.051905285 84956.612897 709.812150 59958.784665 6.31205084
coef.var 0.5773809 NA NA 0.003523232 1.102569 0.617735 1.617146 0.30726601
> |

```

## Crop Item Analysis

Frequency of the each crop grown in different regions of the world from 1990-2013 is given by:

```

6 sapply(data, class)
7
8 #Checking for missing values
9 any(is.na(data))
10
11 # Displaying the structure of your data
12 str(data)
13 # Displaying the first few rows of your data
14 head(data)
15
16 library(pastecs)
17 stats_of_data <- format(stat.desc(data), scientific = FALSE)
18 print(stats_of_data)
19
20 library(ggplot2)
21
22 # Analyzing Item
23 table(data$item)
24
25 pie(table(data$item), main = "Crop Frequency Distribution", col = terrain.colors(12))
26
27 crop_frequency <- function(chosen_area){
28   # Filtering data for the selected area
29   filtered_data = data[data$Area == chosen_area, ]
30 }
31
32 crop_frequency(chosen_area) :

```

```

R 4.4.1 - ~/Documents/PES1UG21CS486/ 
2 1 Albania Potatoes 1990 66667 1485 121 16.4
3 2 Albania Rice, paddy 1990 23333 1485 121 16.4
4 3 Albania Sorghum 1990 12500 1485 121 16.4
5 4 Albania Soybeans 1990 2000 1485 121 16.4
6 5 Albania Wheat 1990 38197 1485 121 16.4
> library(ggplot2)
> # Analyzing Item
> table(data$item)

```

Cassava	Maize	Plantains and others	Potatoes	Rice, paddy	Sorghum
3845	4131	556	4276	3388	3039
Soybeans	Sweet potatoes		Wheat	Yams	
3223	2890	3857	847		

```

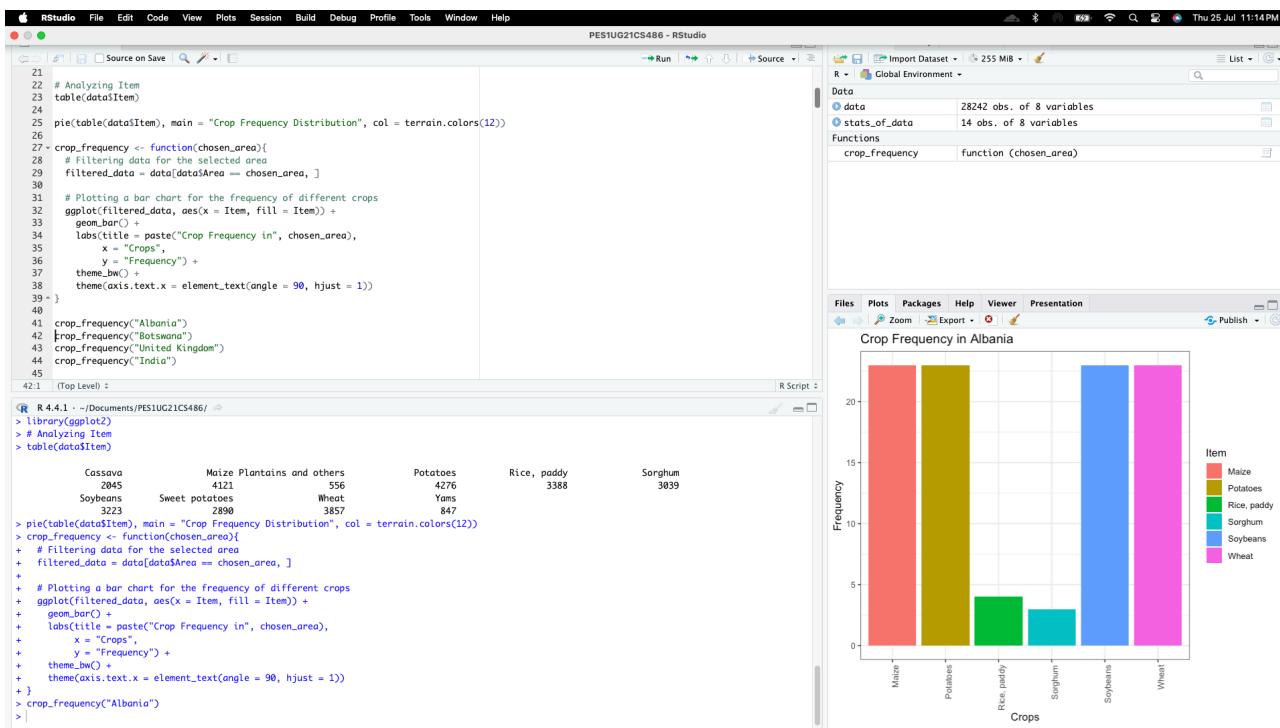
> pie(table(data$item), main = "Crop Frequency Distribution", col = terrain.colors(12))
> |

```

**Crop Frequency Distribution**

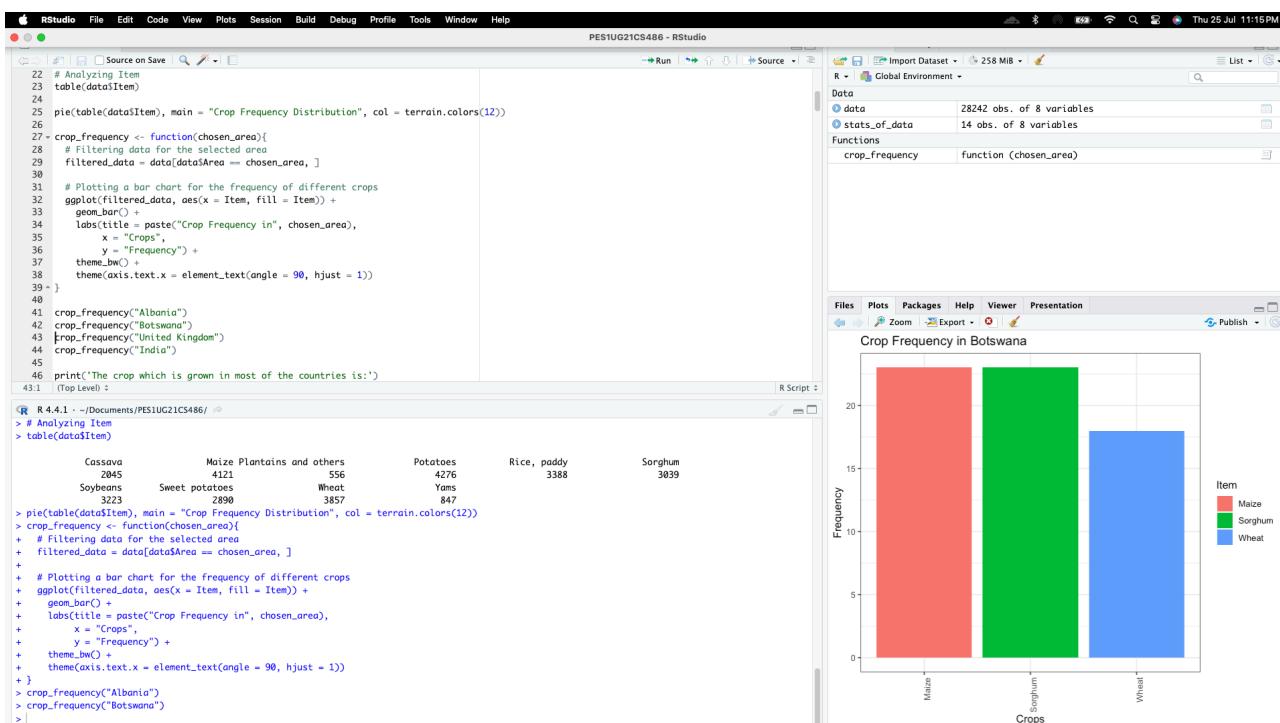
On the right side of the image we can see a pie chart of the crop frequency distribution. Out of all crops, crops like Potatoes, Maize and Wheat are the most preferred crop around the world while Yams, Plantains and others are the least preferred.

Following graph shows for a particular geographical area, the kinds of crops grown and the frequency of each crop item. For example in Albania, we see an equal preference for crops like Maize, Potatoes, Soybeans and Wheat while crops like Rice and Sorghum have an extremely low preference.

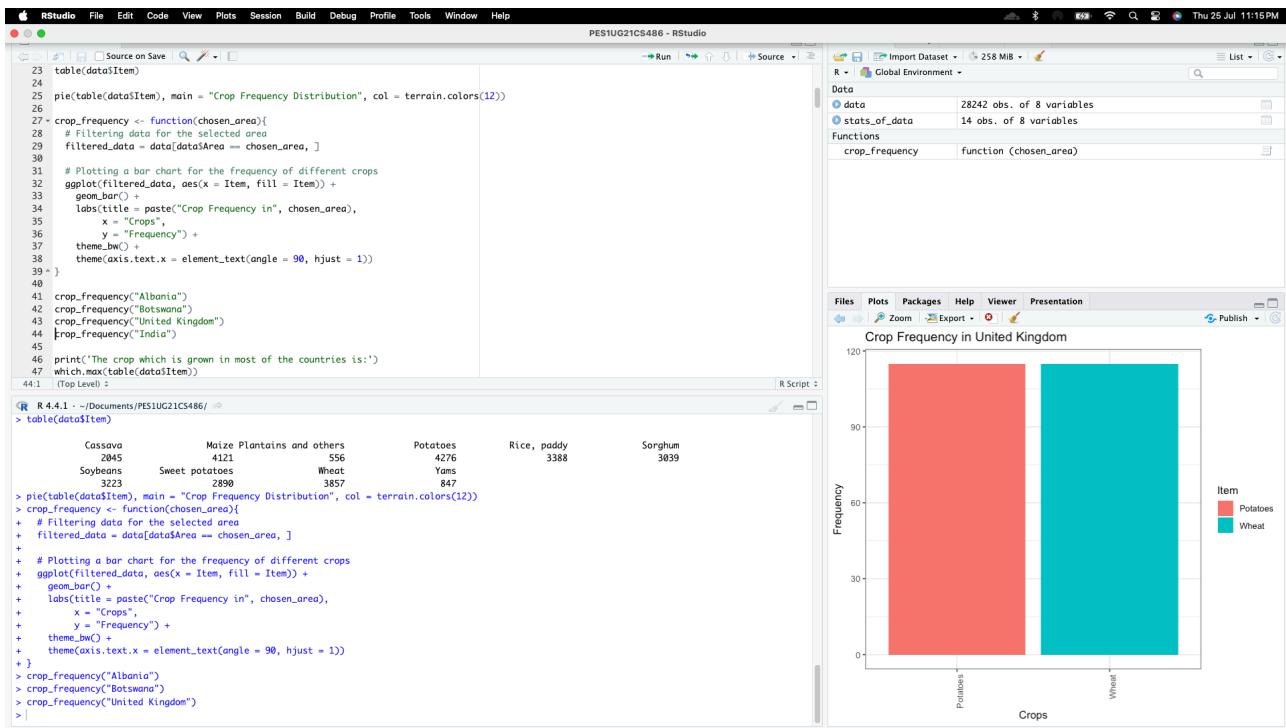


In a similar way we can have barplots for crop preference in different countries, such as United Kingdom, India, Botswana, etc

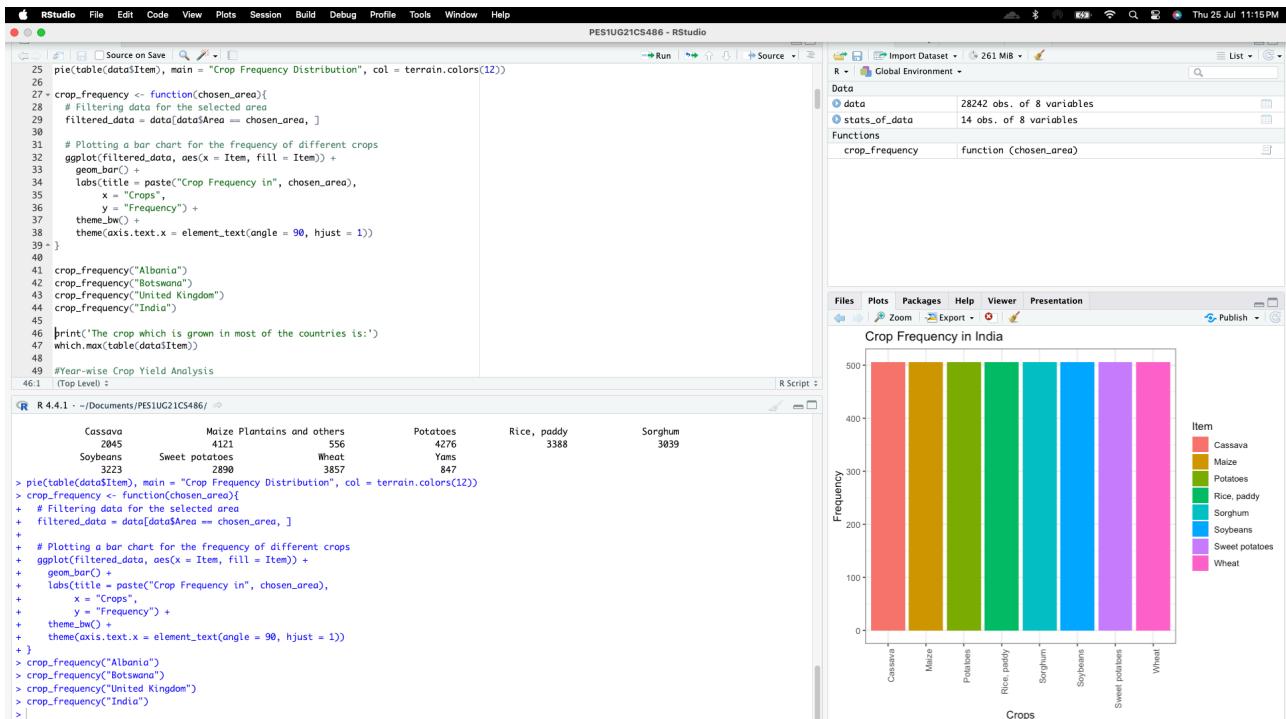
*On exploring further it was observed that in Botswana, only Maize, Sorghum and Wheat were grown between 1990-2013, out of which less preference was given to Wheat.*



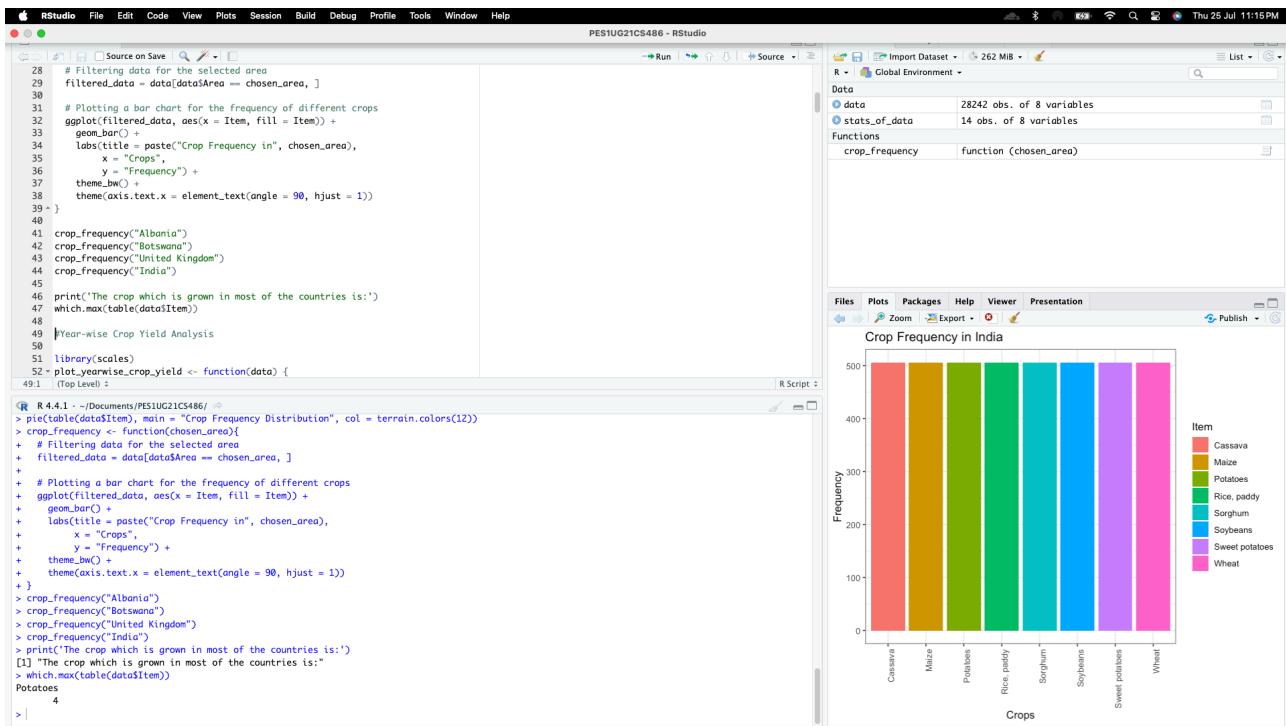
In United Kingdom, Potatoes and Wheat were the most eaten crops between 1990-2013.



In India, an equal preference is given to crops like Cassava, Maize, Potatoes, Rice, Sorghum, Soybeans, Sweet Potatoes and Wheat.

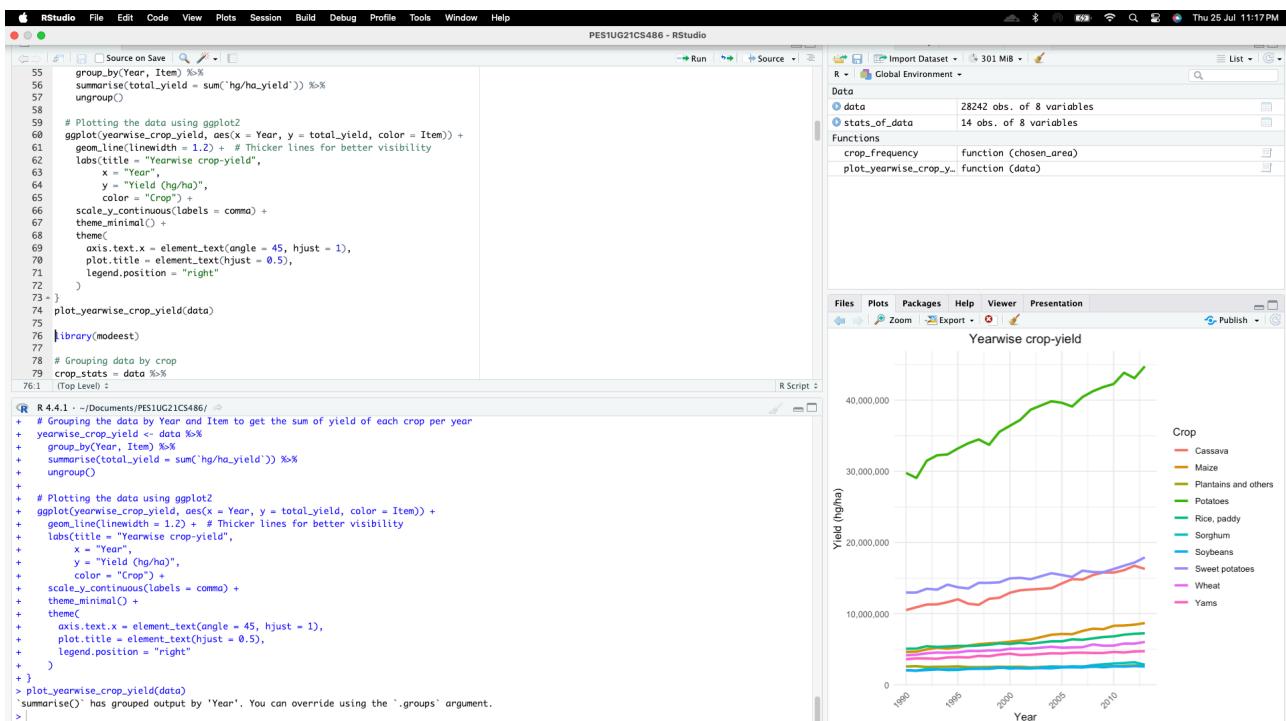


We can also conclude from the dataset that the most grown crop in all the countries is Potato.

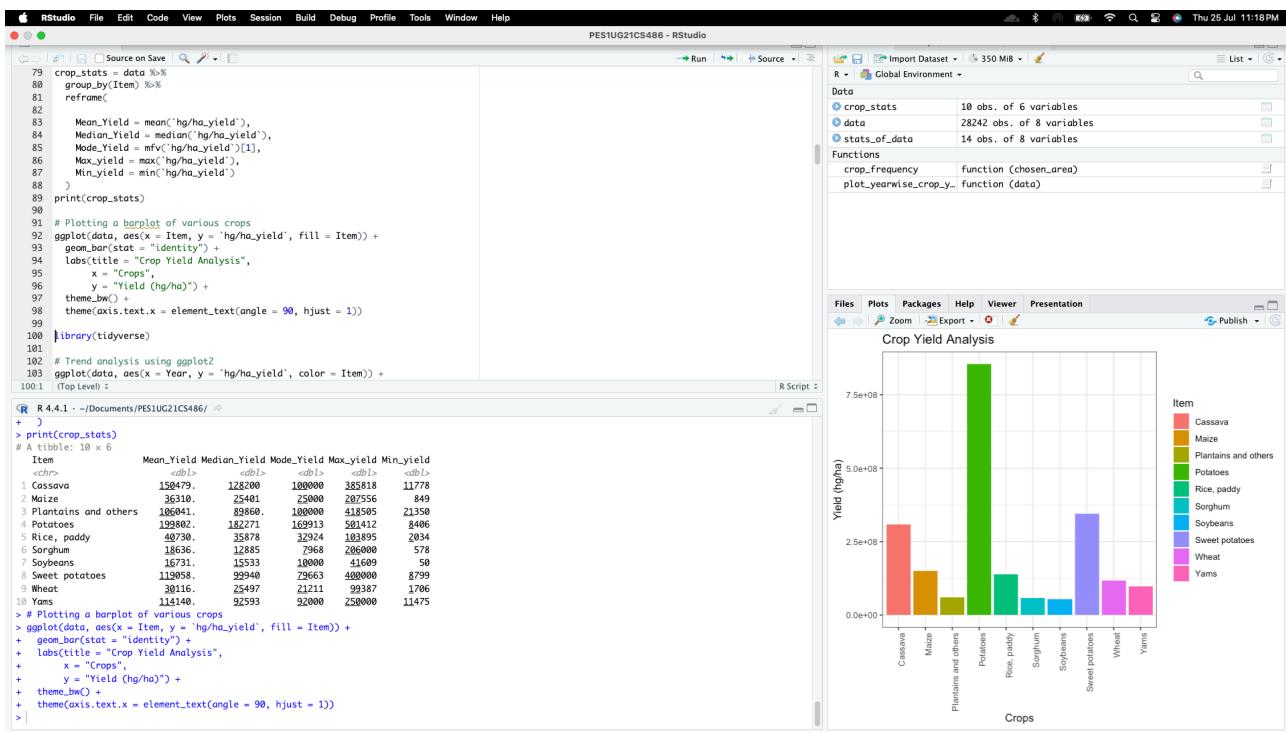


## Crop Yield Analysis

Looking closely at the year wise crop-yield we notice that potatoes are the highest by a big margin. We can also observe that sweet potatoes and cassava yields have gotten closer from 1990 to 2013. Crops like Soybean and Sorghum have the lowest yield.

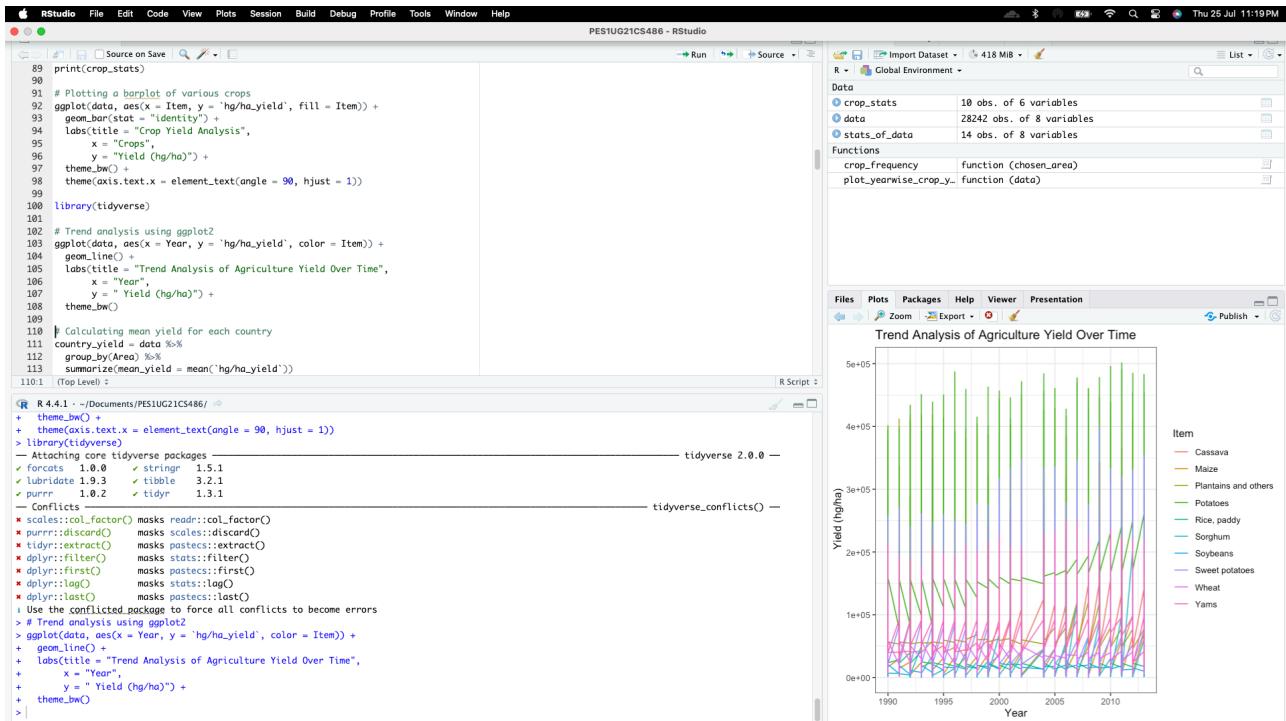


Descriptive statistics of yield for different crop item and the adjacent barplot compares yield of different crops grown in different areas of the world:

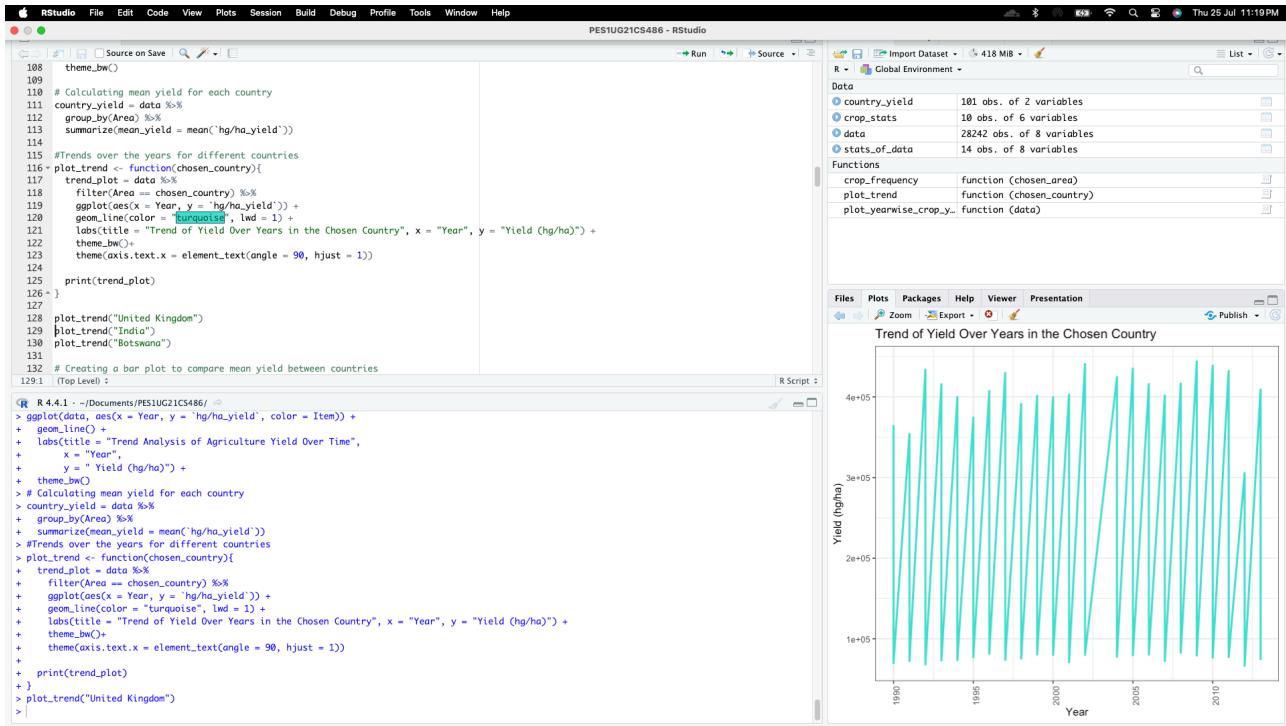


Clearly visible from the above graph, potatoes have been the most popular crop globally, with a maximum yield of 501412 hg/ha. Conversely, there are crops such as soybeans, sorghum, and plantains whose maximum output was constant.

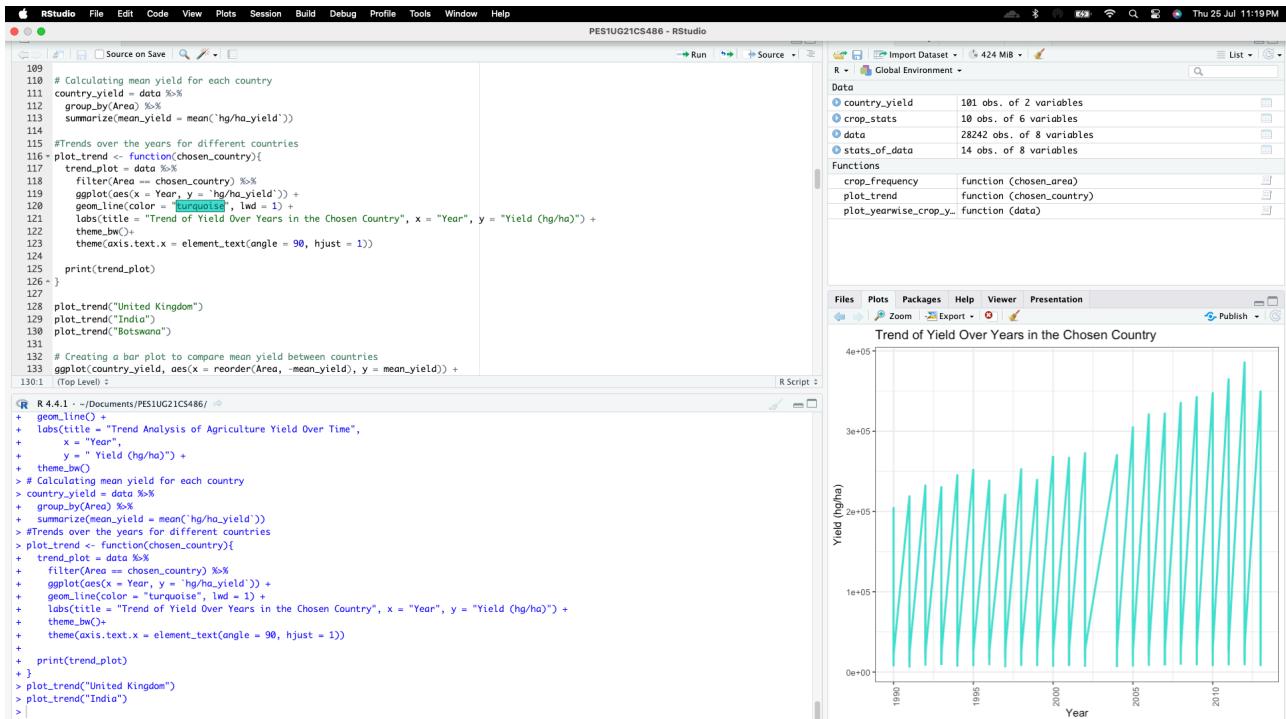
The agricultural yield trends for various crop varieties have changed throughout time in the following ways:



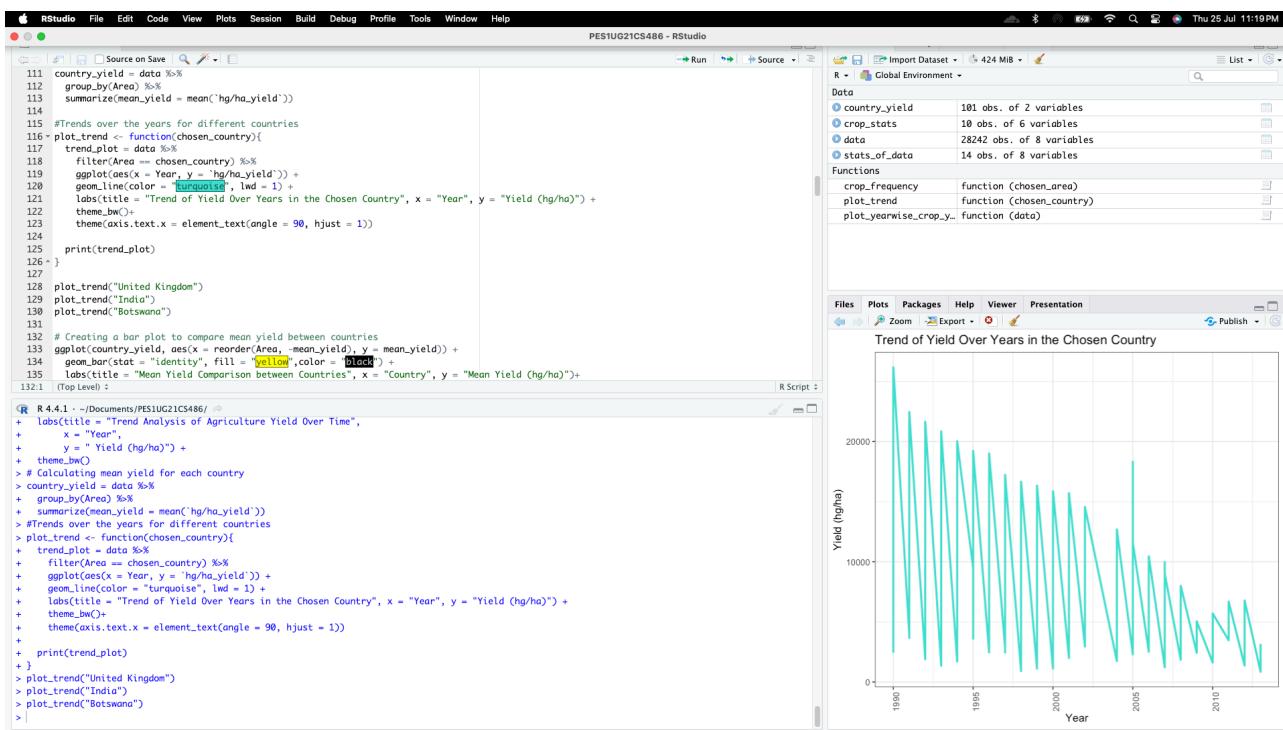
The following plot shows the trends over the years for different countries. This graph helps in visualizing how agricultural yield has evolved over time in the specified country, highlighting both long-term trends and year-to-year variations. The following plot is of the United Kingdoms and the graph shows that United Kingdom has the highest value of mean yield over the time period of 1990 to 2013.



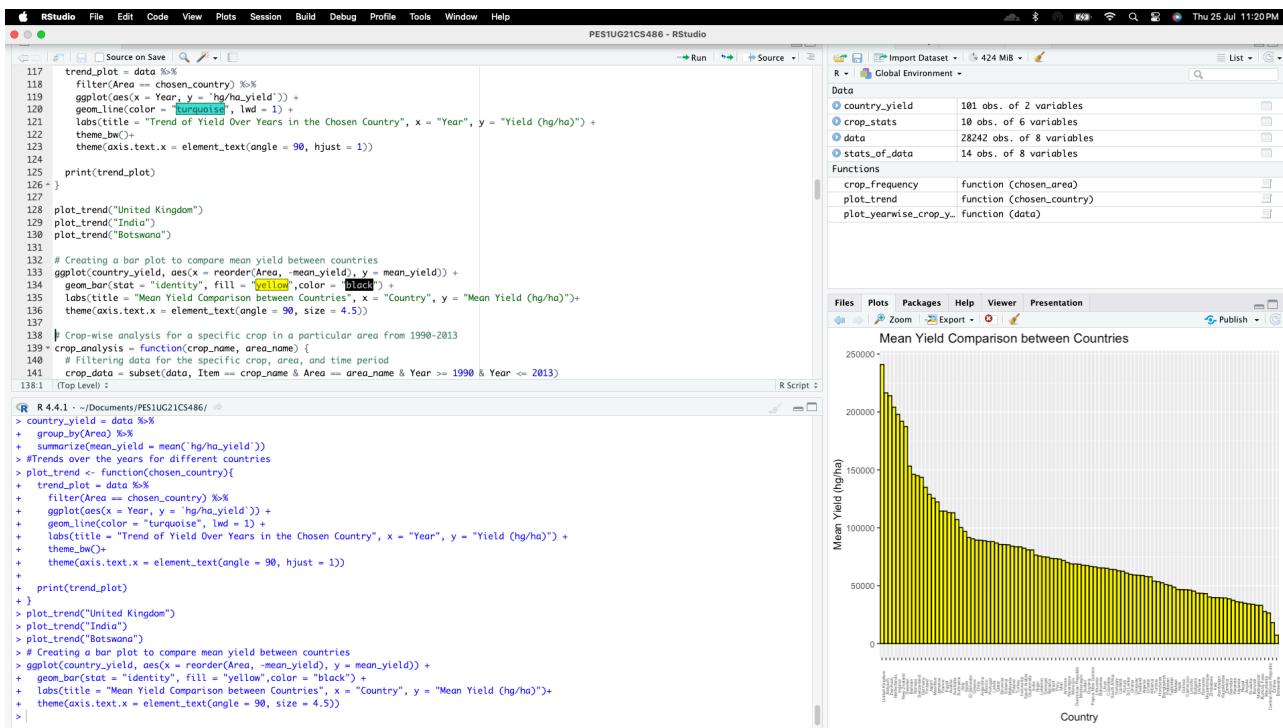
We can also observe that from 2005, the mean yield in India has constantly been rising emphasising on the fact that agriculture in India is advancing due to various factors.



From, the following graph, we can observe that in Botswana the crop yield has significantly decreased from 1990 to 2013.



The following graph illustrates the average yield for all crops grown in various areas:



## Outlier Detection:

It is also possible to determine the net yield outlier values (in hectograms per hectare) for a certain crop in a nation like India by utilising a boxplot.

The yield outlier value (in hectograms per hectare) for a few crops is provided by:

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help PES1UG21CS486 - RStudio

```

138 # Crop-wise analysis for a specific crop in a particular area from 1990-2013
139 crop_analysis <- function(crop_name, area_name) {
140   # Filtering data for the specific crop, area, and time period
141   crop_data <- subset(data, Item == crop_name & Area == area_name & Year >= 1990 & Year <= 2013)
142 
143   # Displaying summary statistics
144   print(paste("Summary Statistics for", crop_name, "in", area_name, "from 1990-2013:"))
145   print(summary(crop_data$hg/ha_yield))
146 
147   # Creating a boxplot
148   boxplot(crop_data$hg/ha_yield, main = paste(crop_name, "Yield Distribution in", area_name), ylab = "Yield (hg/ha)", col = "#D66040")
149 
150   # Identification of outliers using the Tukey method
151   outliers <- boxplot.stats(crop_data$hg/ha_yield)$out
152 
153   # Displaying outliers
154   print("Outliers are:")
155   print(outliers)
156 }

157 crop_analysis("Potatoes", "India")
158 crop_analysis("Sweet potatoes", "India")
159 crop_analysis("Wheat", "India")
160 
161 #Effect of rainfall
162 #Effect of rainfall
163 
```

(Top Level) :

R 4.4.1 - ~/Documents/PES1UG21CS486/

```

+ # Filtering data for the specific crop, area, and time period
+ crop_data <- subset(data, Item == crop_name & Area == area_name & Year >= 1990 & Year <= 2013)
+
# Displaying summary statistics
+ print(paste("Summary Statistics for", crop_name, "in", area_name, "from 1990-2013:"))
+ print(summary(crop_data$hg/ha_yield))
+
# Creating a boxplot
+ boxplot(crop_data$hg/ha_yield, main = paste(crop_name, "Yield Distribution in", area_name), ylab = "Yield (hg/ha)", col = "#D66040")
+
# Identification of outliers using the Tukey method
+ outliers <- boxplot.stats(crop_data$hg/ha_yield)$out
+
# Displaying outliers
+ print("Outliers are:")
+ print(outliers)
+
} 
```

> crop\_analysis("Potatoes", "India")

[1] "Summary Statistics for Potatoes in India from 1990-2013:"

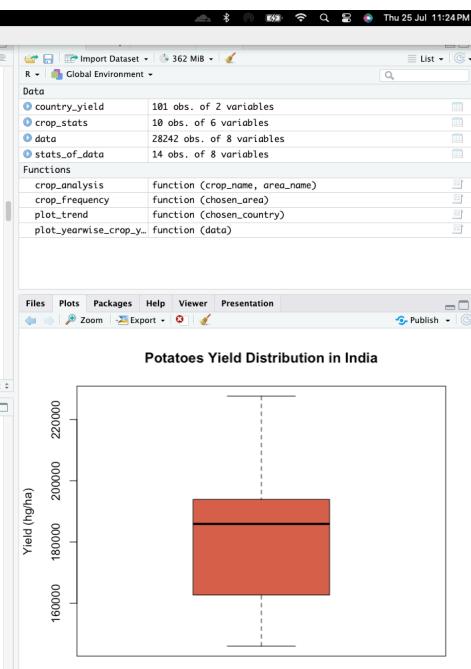
Min. 1st Qu. Median Mean 3rd Qu. Max.

146020 162720 185920 182060 193913 227606

[1] "Outliers are:"

numeric(0)

>



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help PES1UG21CS486 - RStudio

```

139 crop_analysis <- function(crop_name, area_name) {
140   # Filtering data for the specific crop, area, and time period
141   crop_data <- subset(data, Item == crop_name & Area == area_name & Year >= 1990 & Year <= 2013)
142 
143   # Displaying summary statistics
144   print(paste("Summary Statistics for", crop_name, "in", area_name, "from 1990-2013:"))
145   print(summary(crop_data$hg/ha_yield))
146 
147   # Creating a boxplot
148   boxplot(crop_data$hg/ha_yield, main = paste(crop_name, "Yield Distribution in", area_name), ylab = "Yield (hg/ha)", col = "#D66040")
149 
150   # Identification of outliers using the Tukey method
151   outliers <- boxplot.stats(crop_data$hg/ha_yield)$out
152 
153   # Displaying outliers
154   print("Outliers are:")
155   print(outliers)
156 }

157 crop_analysis("Potatoes", "India")
158 crop_analysis("Sweet potatoes", "India")
159 crop_analysis("Wheat", "India")
160 
161 #Effect of rainfall
162 #Effect of rainfall
163 
```

(Top Level) :

R 4.4.1 - ~/Documents/PES1UG21CS486/

```

+ # Creating a boxplot
+ boxplot(crop_data$hg/ha_yield, main = paste(crop_name, "Yield Distribution in", area_name), ylab = "Yield (hg/ha)", col = "#D66040")
+
# Identification of outliers using the Tukey method
+ outliers <- boxplot.stats(crop_data$hg/ha_yield)$out
+
# Displaying outliers
+ print("Outliers are:")
+ print(outliers)
+
} 
```

> crop\_analysis("Potatoes", "India")

[1] "Summary Statistics for Potatoes in India from 1990-2013:"

Min. 1st Qu. Median Mean 3rd Qu. Max.

146020 162720 185920 182060 193913 227606

[1] "Outliers are:"

numeric(0)

> crop\_analysis("Sweet potatoes", "India")

[1] "Summary Statistics for Sweet potatoes in India from 1990-2013:"

Min. 1st Qu. Median Mean 3rd Qu. Max.

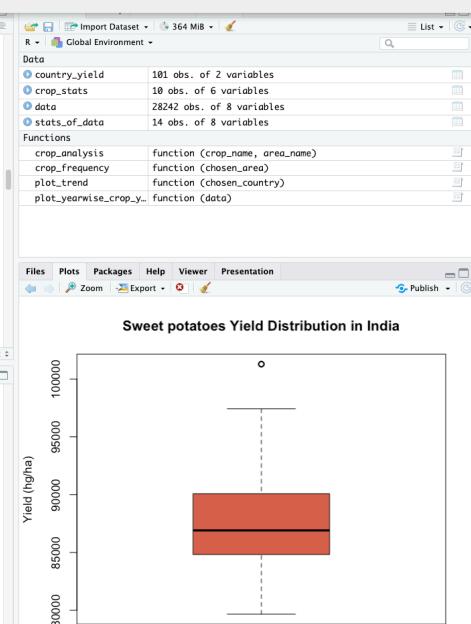
79663 84823 86907 87825 90080 101288

[1] "Outliers are:"

[1] 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288

[2] 101288 101288

>



The screenshot shows an RStudio interface with a code editor and a plot viewer. The code editor contains R script for crop analysis, including a boxplot for wheat yield. The plot viewer displays a boxplot titled 'Wheat Yield Distribution in India' with the y-axis labeled 'Yield (kg/ha)' ranging from 22000 to 32000. The plot shows a median around 27000, a box spanning approximately 25000-28500, and several outliers above 30000.

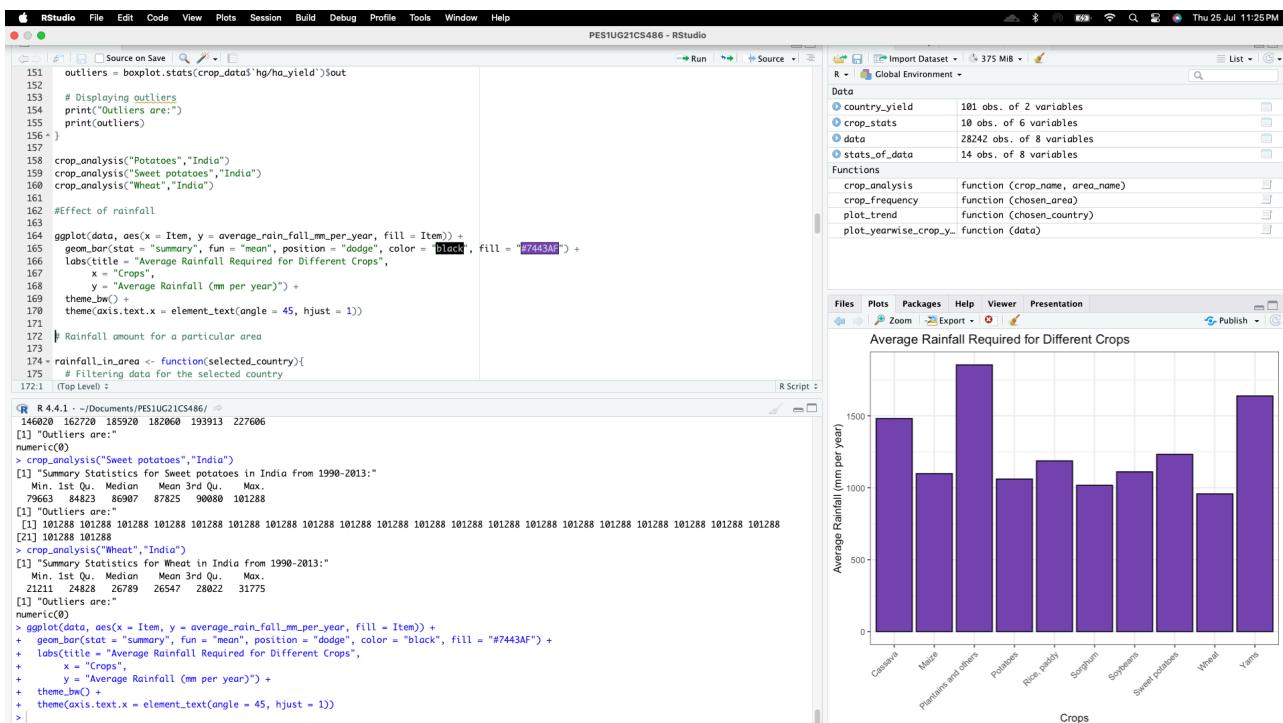
```

141 crop_data = subset(data, Item == crop_name & Area == area_name & Year >= 1990 & Year <= 2013)
142
143 # Displaying summary statistics
144 print(paste("Summary Statistics for", crop_name, "in", area_name, "from 1990-2013:"))
145 print(summary(crop_data$hg/ha_yield))
146
147 # Creating a boxplot
148 boxplot(crop_data$hg/ha_yield, main = paste(crop_name, "Yield Distribution in", area_name), ylab = "Yield (kg/ha)", col = "#E69138")
149
150 # Identification of outliers using the Tukey method
151 outliers = boxplot.stats(crop_data$hg/ha_yield)$out
152
153 # Displaying outliers
154 print("Outliers are:")
155 print(outliers)
156
157
158 crop_analysis("Potatoes", "India")
159 crop_analysis("Sweet potatoes", "India")
160 crop_analysis("Wheat", "India")
161
162 #Effect of rainfall
163
164 ggplot(data, aes(x = Item, y = average_rain_fall_mm_per_year, fill = Item)) +
165 geom_bar(stat = "summary", fun = "mean", position = "dodge", color = "black", fill = "#E69138") +
166 labs(title = "Average Rainfall Required for Different Crops",
167 x = "Crops",
168 y = "Average Rainfall (mm per year)") +
169 theme_bw() +
170 theme(axes.text.x = element_text(angle = 45, hjust = 1))
171
172 # Rainfall amount for a particular area
173
174 rainfall_in_area <- function(selected_country){
175 # Filtering data for the selected country
176
177 (Top Level):

```

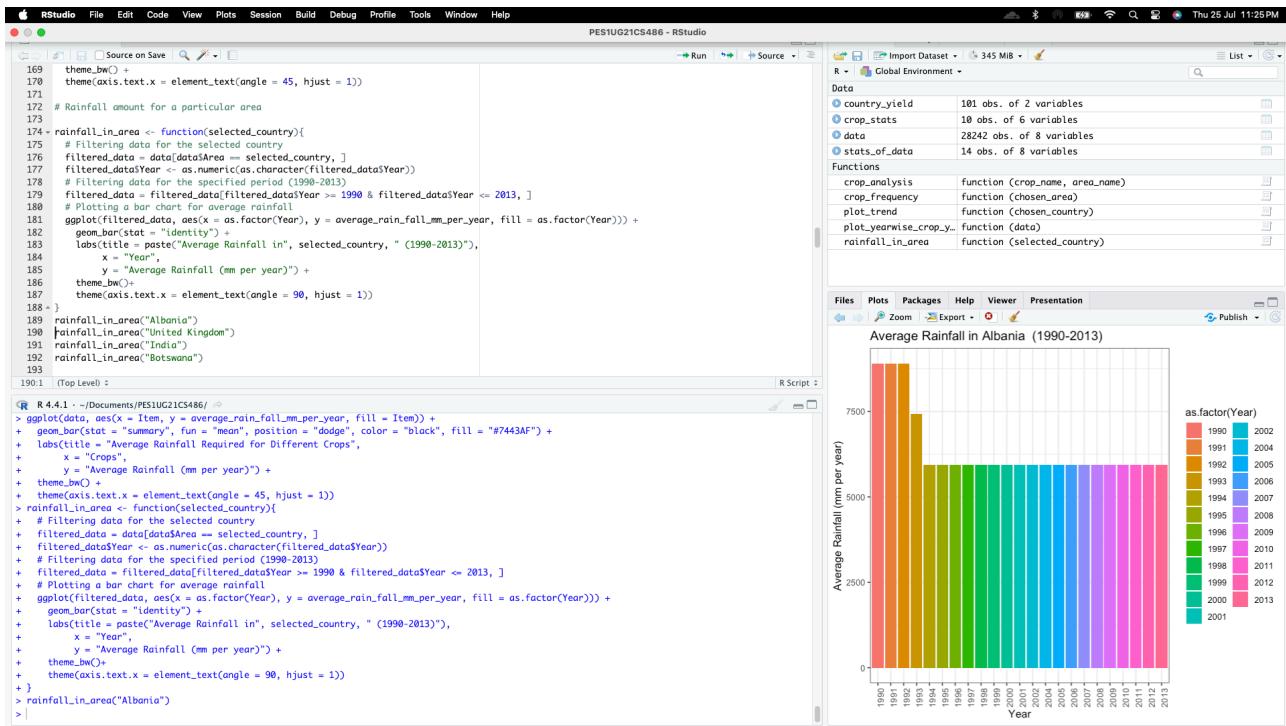
## Effect of Rainfall

The quantity of rainfall required annually by various crop kinds can be determined using the bar chart below:

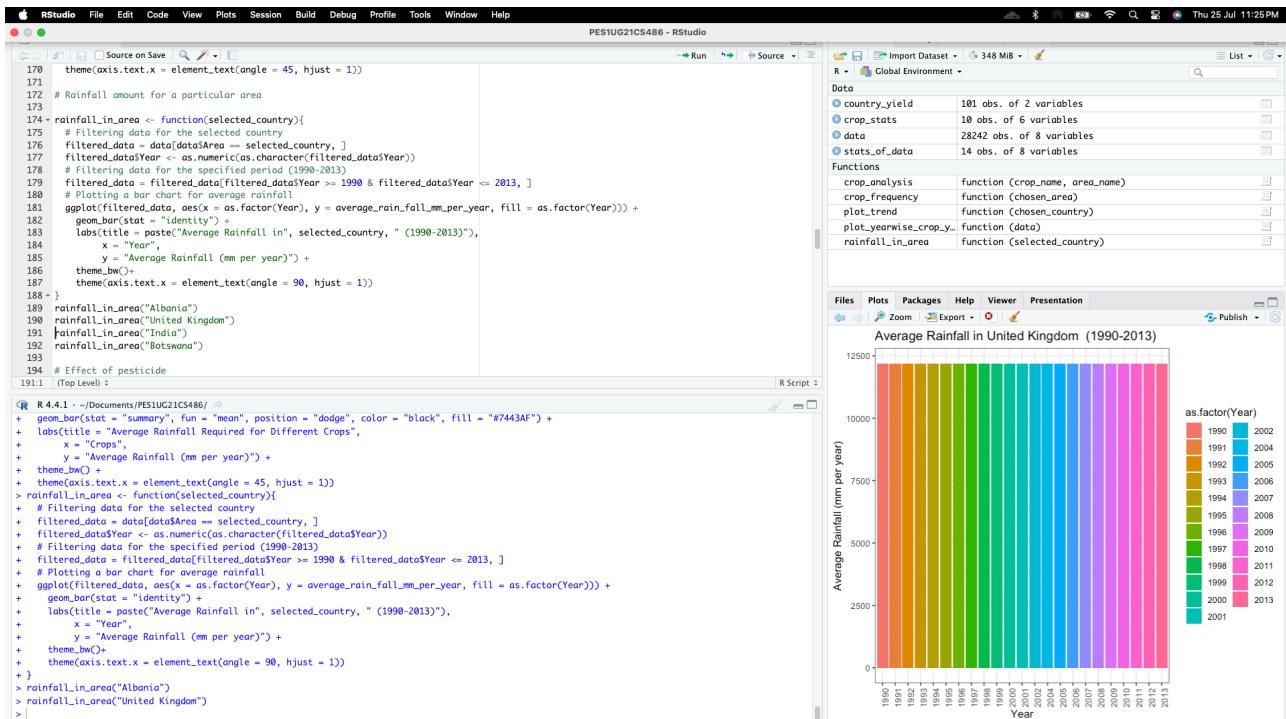


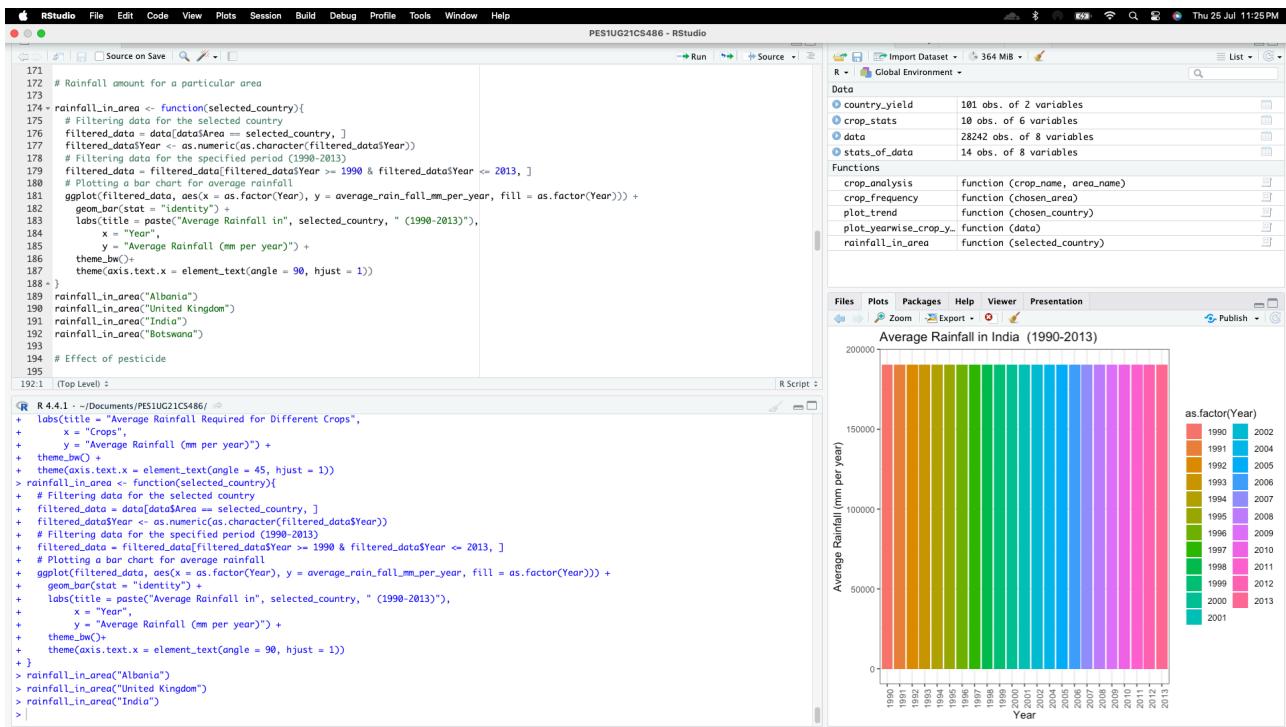
Additionally, throughout time, the average amount of rainfall in a given area varies. It is evident that whereas crops like maize, potatoes, sorghum, and wheat require nearly the same quantity of rainfall, plants and yellows require a significant amount of it. The graphs also illustrates patterns in rainfall for nations including Albania, the United Kingdom, and Botswana from 1990 to 2013.

In Albania, we see a high amount of rainfall from 1990-1992 but a substantial fall in 1992. From 1994-2013, the average amount of rainfall stayed the same.

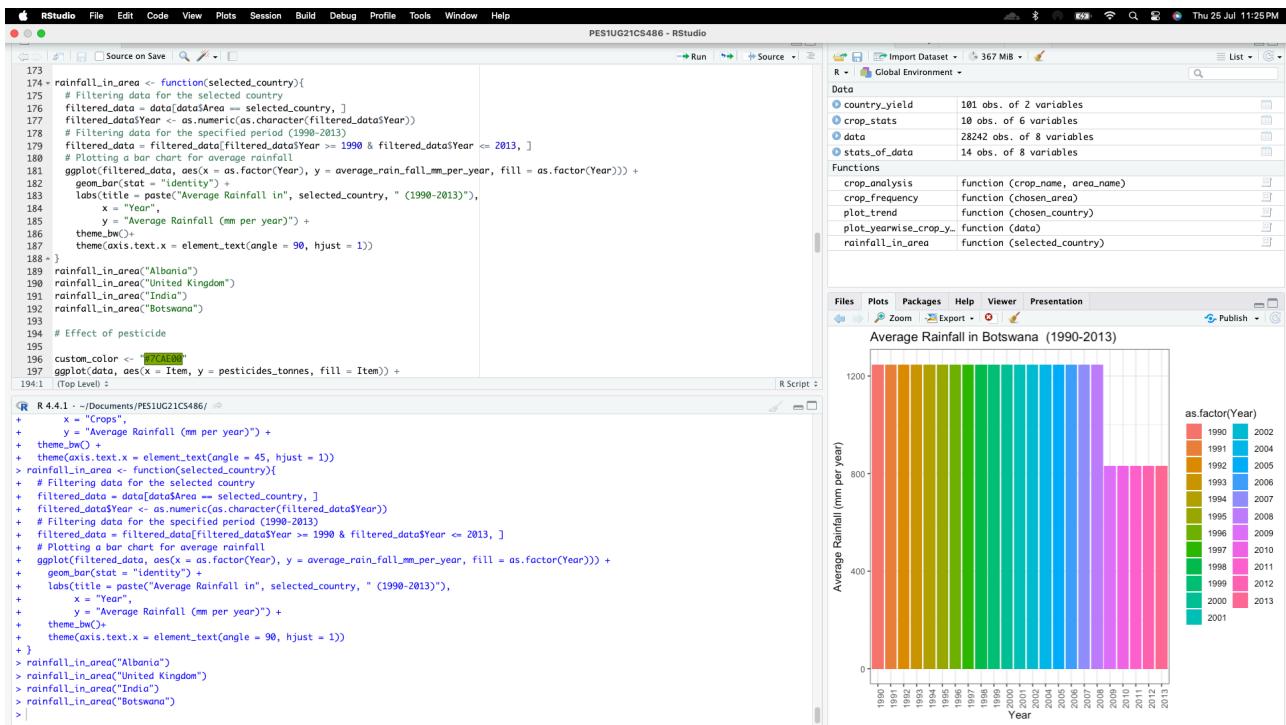


For all the years from 1990 -2013, the average amount of rainfall in regions like The United Kingdom and India stood at the same level.



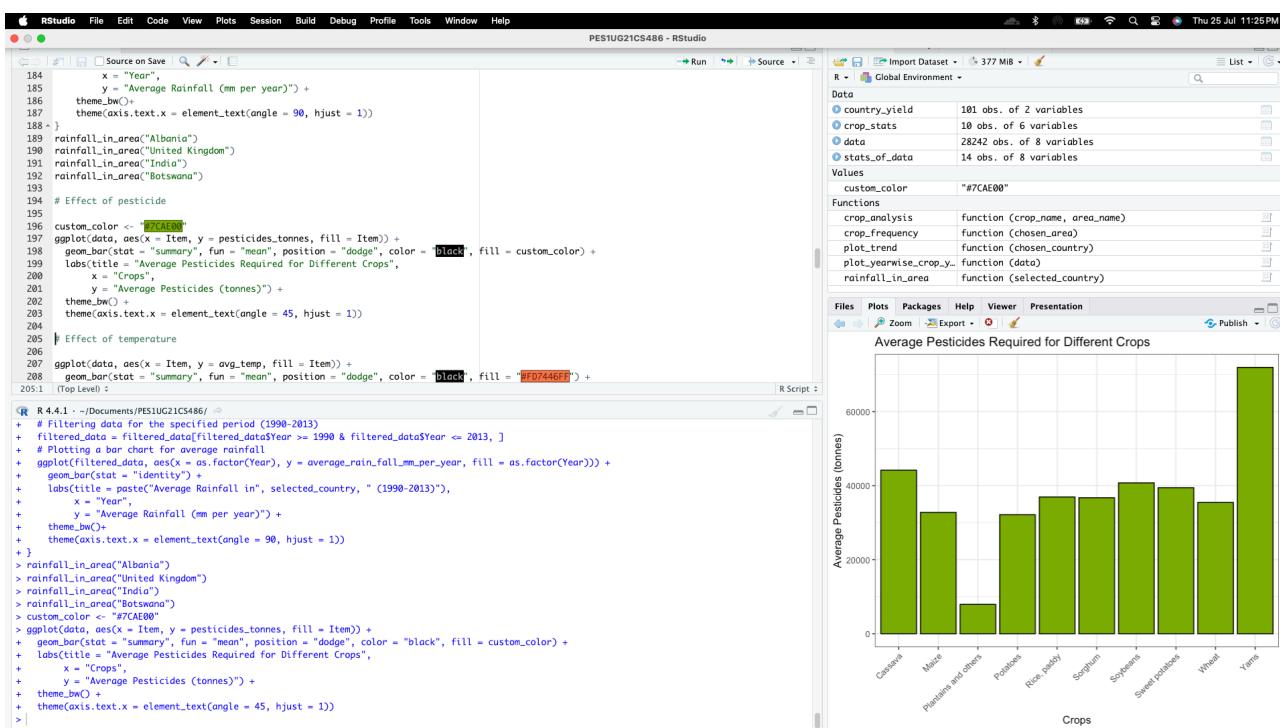


In Botswana there was a sudden drop in rainfall pattern after 2008 and remained the same till 2013.



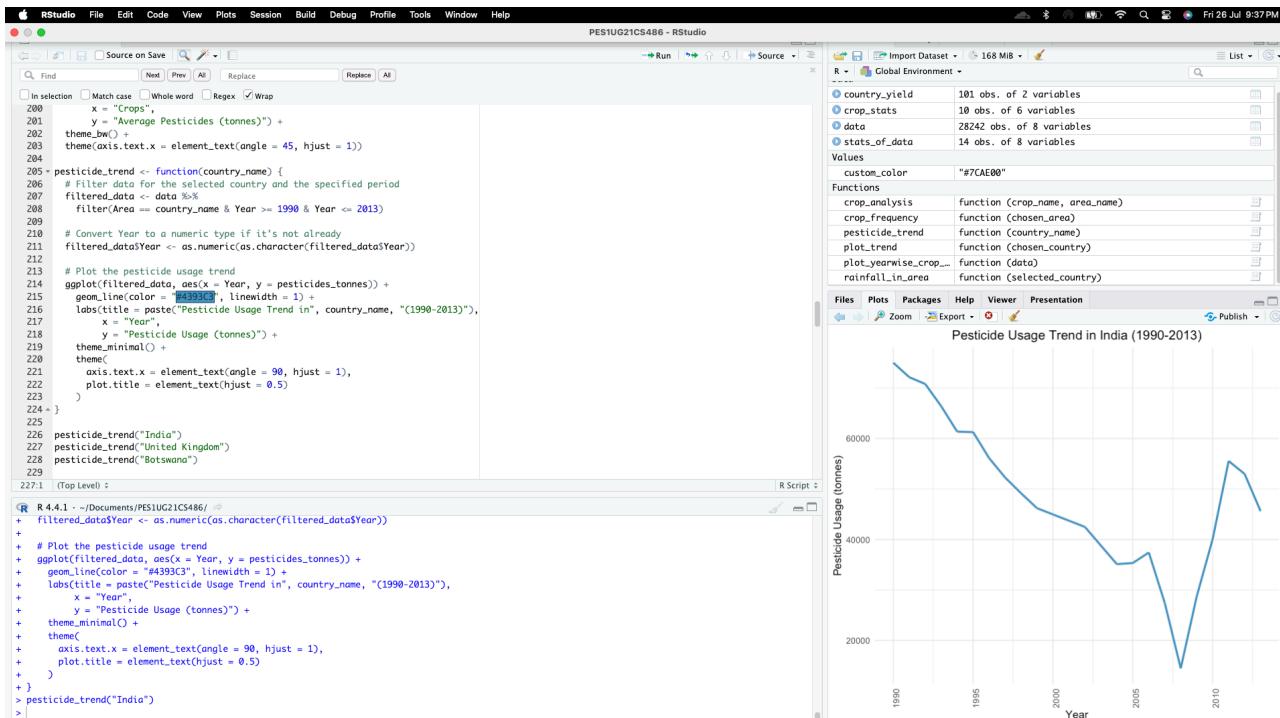
## Effect of Pesticides

The following bar chart can be used to estimate the amount of pesticides that different crop kinds require for growth:

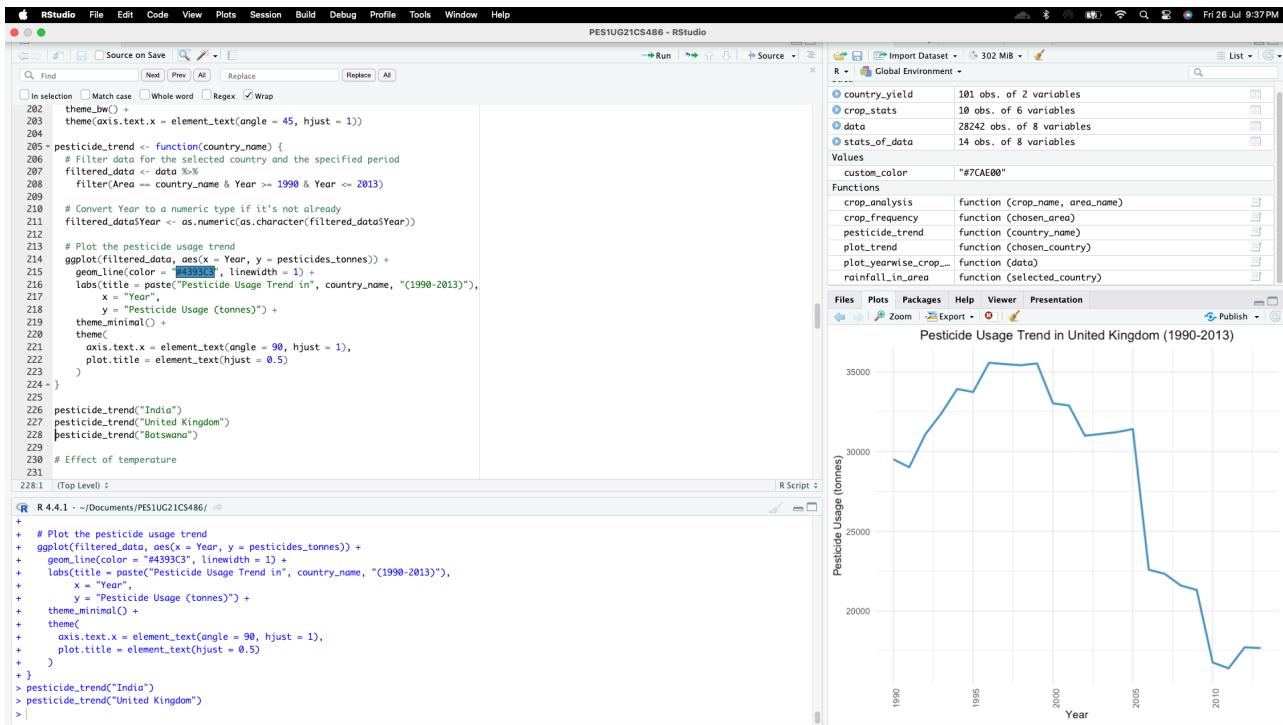


Some observations we can make are that the only crop that needs an excessive number of insecticides is Yams. Given how costly to raise, this could be the reason it is among the least popular crops worldwide. Conversely, the least amount of pesticides is needed for plantains to grow.

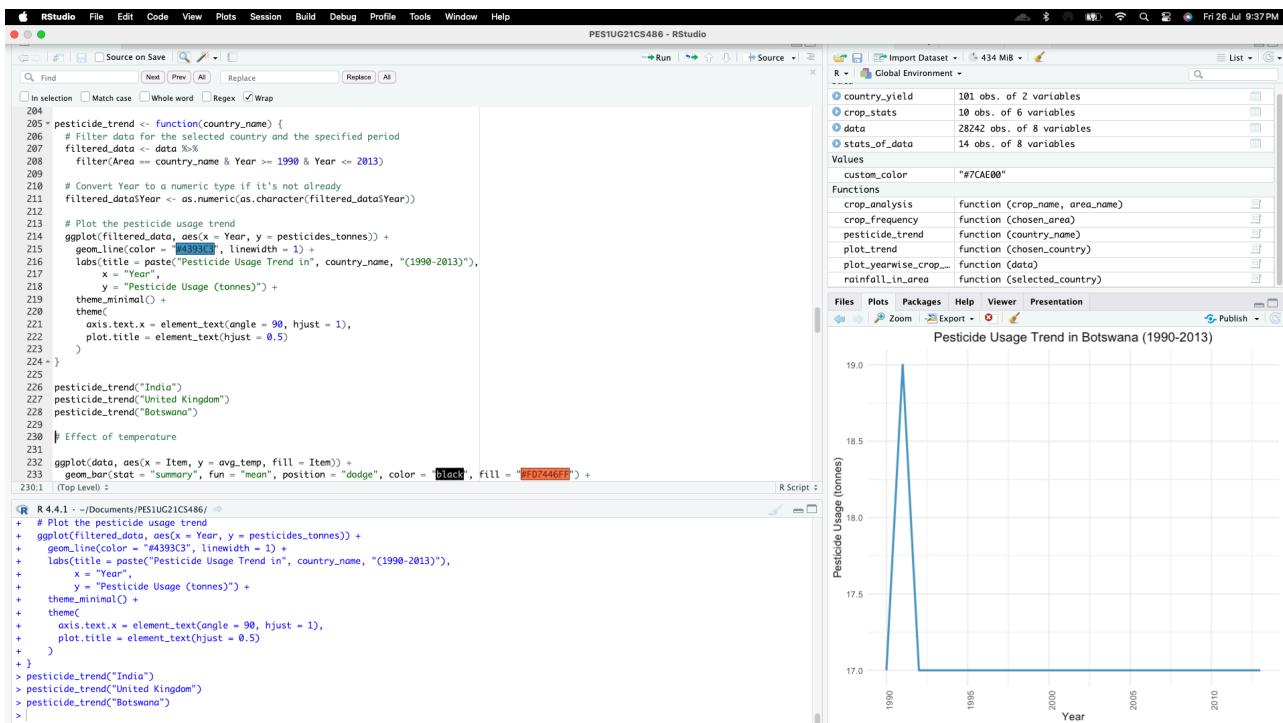
To study the effect of pesticides further, we plot the trends of pesticides over the years in operate countries to study how changes in pesticide usage have affected the yield of the crop. This experiment were conducted on United Kingdom which has the highest mean crop yield , India and Botswana which has the lowest mean yield.



From the above graph we can see that the use of pesticides in India from 2007 onwards has been increased which is an indication that it has been a factor in the increasing mean yield as we saw from our analysis earlier.



Conversely, we also saw a big fall in the mean yield of the United Kingdom post 2010 which further tells us that decreased pesticide usage in the United Kingdom had a big role to play.

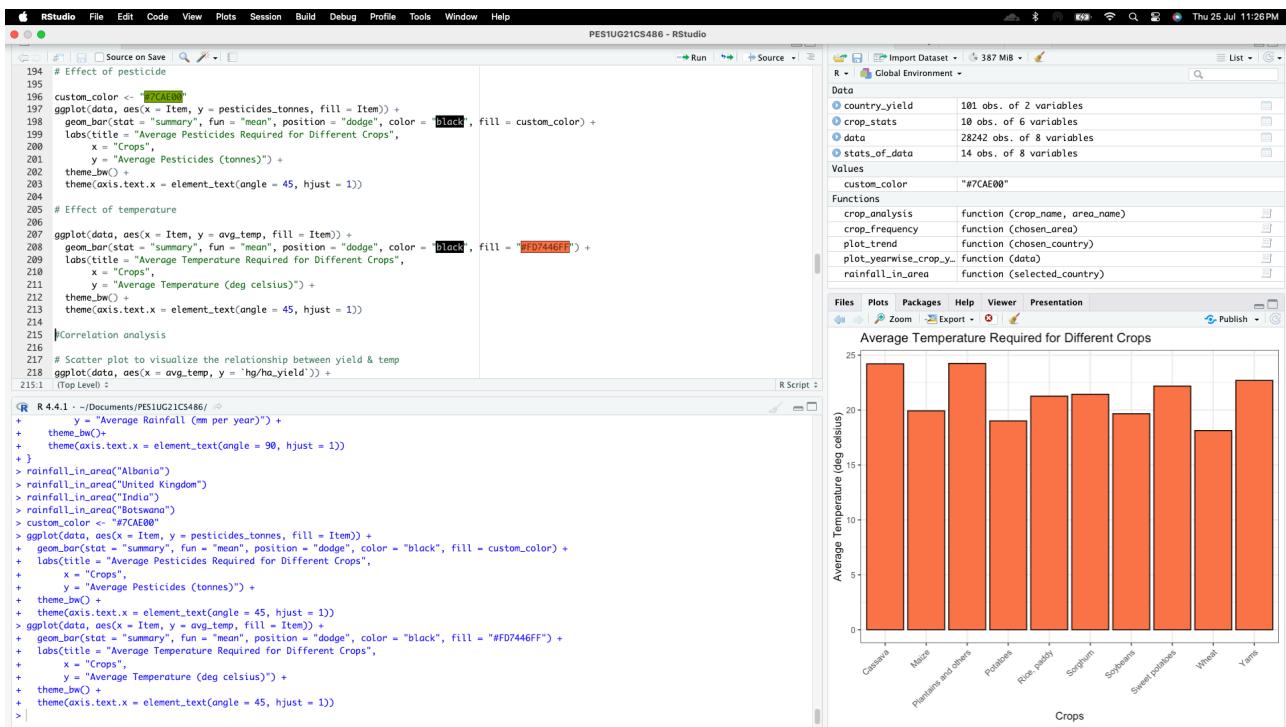


We also saw Botswana having the lowest mean yield and the graph of pesticide usage further confirms our hypothesis that the amount of pesticides used is directly affects the mean yield of a country.

Note: It was interesting to compare the amount of pesticide used in Botswana with any other country, however it could be a human error of recoding incorrect values in the dataset.

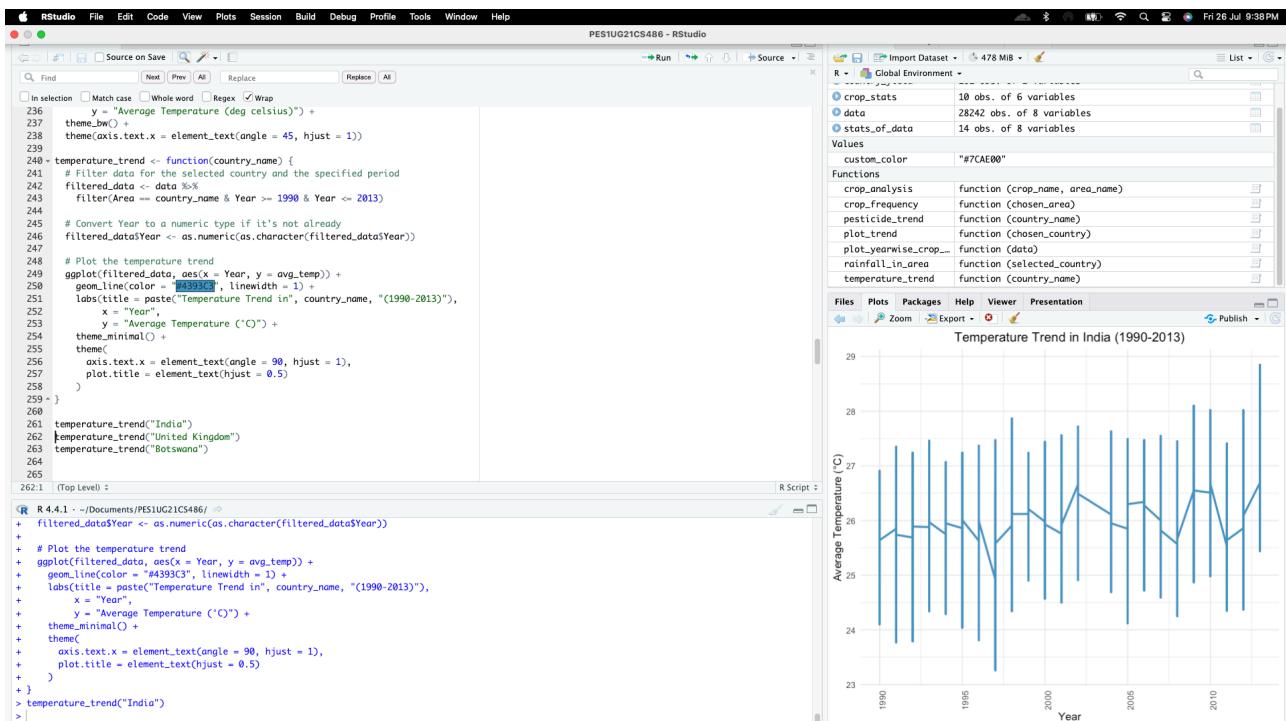
## Effect of Temperature

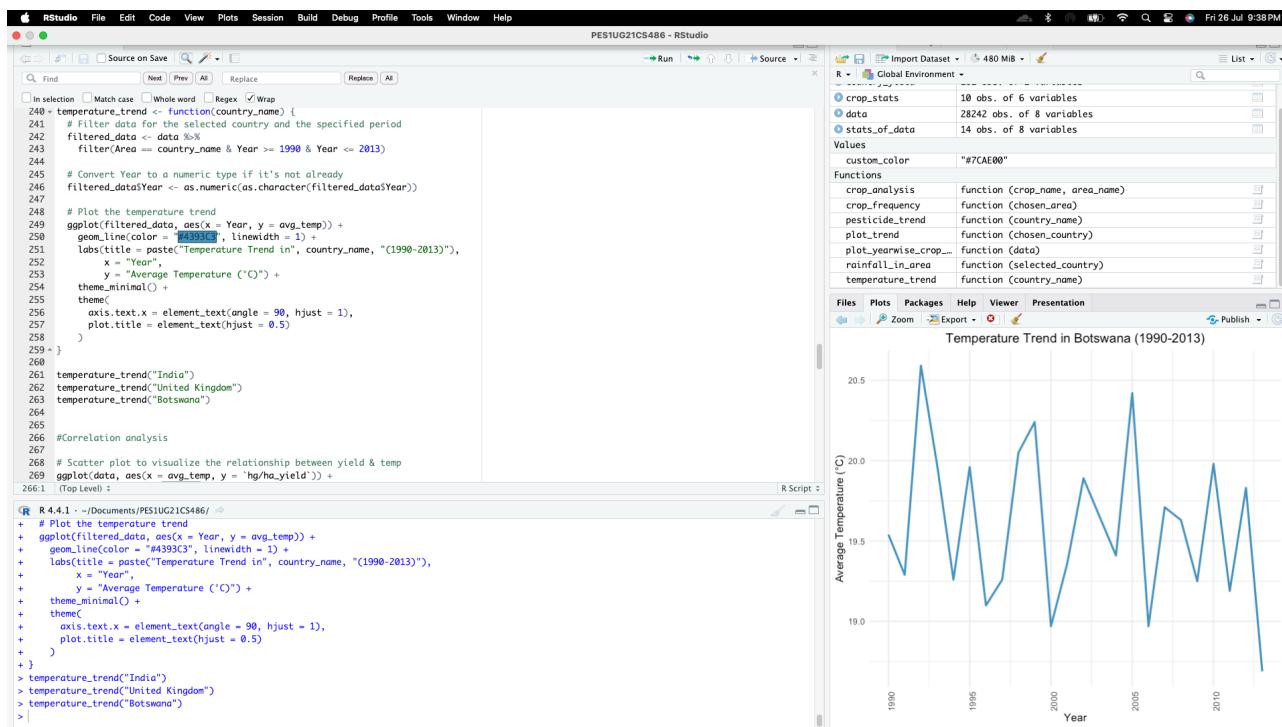
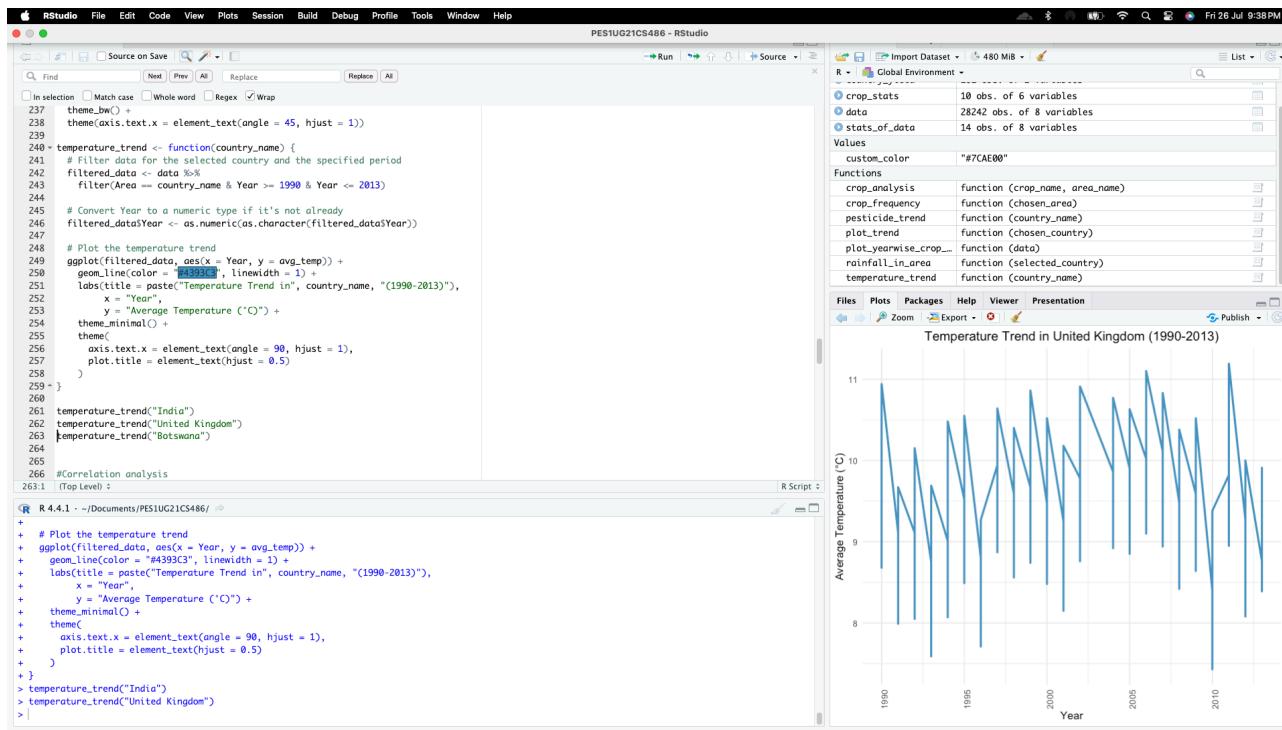
Crop varieties vary in the climates in which they grow. The following bar chart can be used to estimate how temperature affects crops:



The above figure shows that yams, plantains, cassava, and sweet potatoes flourish in arid regions whereas crops like Maize, Potatoes, Soybeans and Wheat flourish in more favourable weather conditions and hence could attribute to their higher yield statistics.

On doing a trend analysis similar to pesticide trends, I want to see if there have been any unfavourable weather conditions that lead to an impact on yield of a country. Here also I have plotted the trends for United Kingdom, India and Botswana

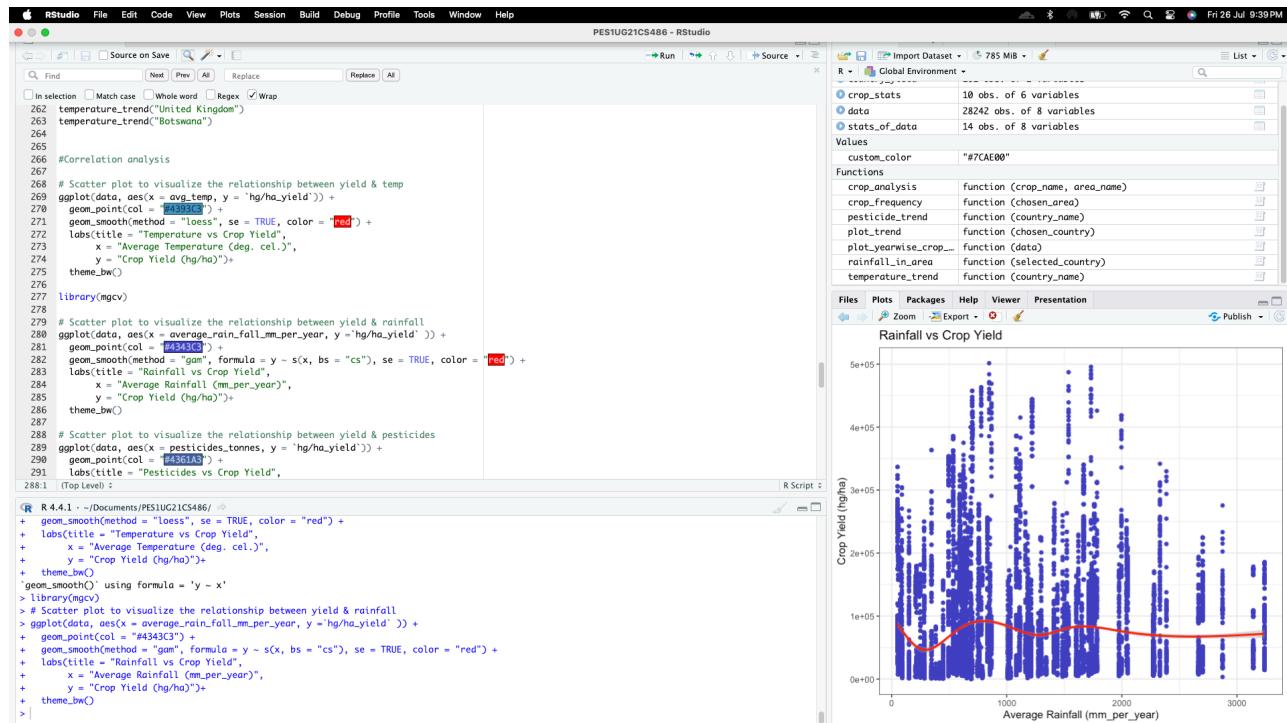
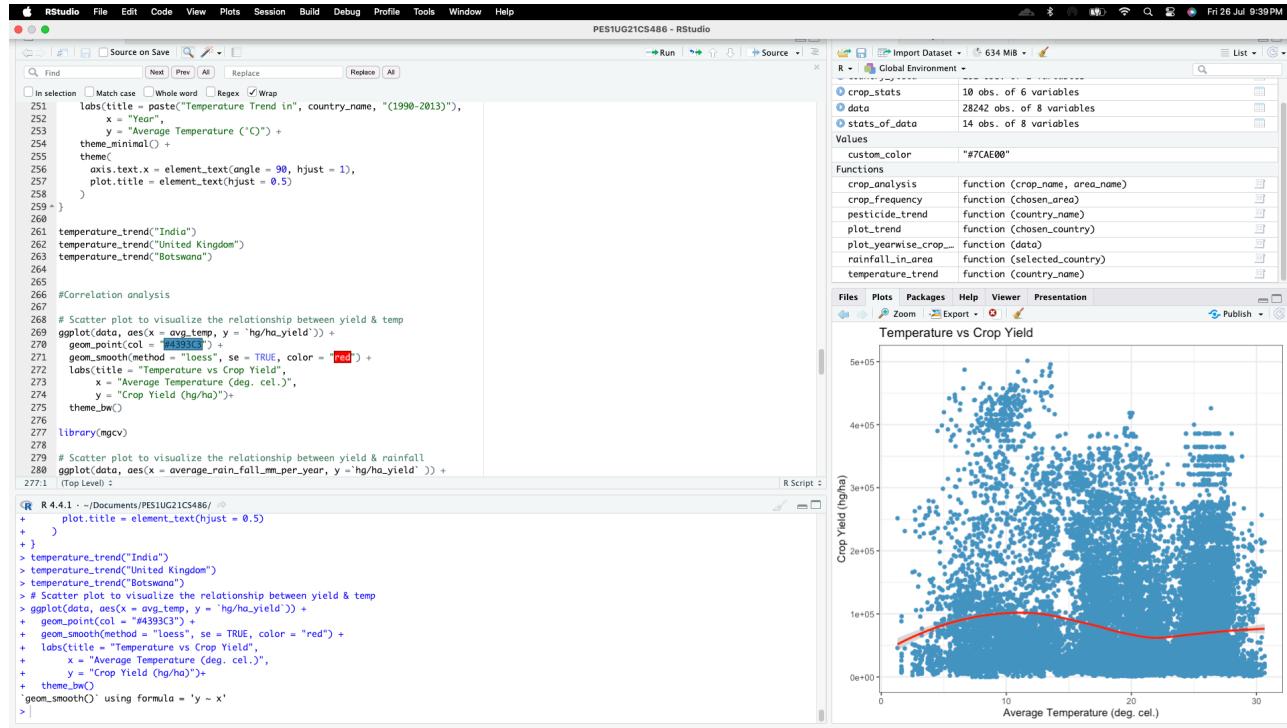


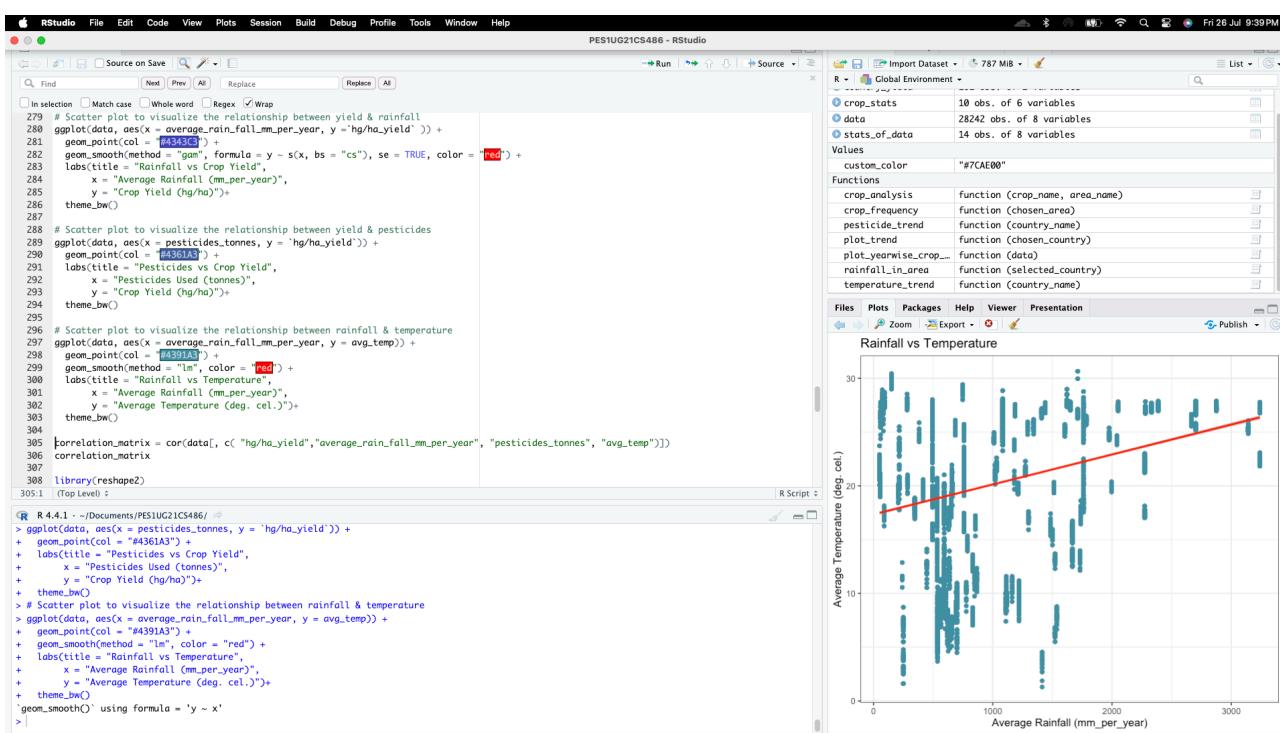
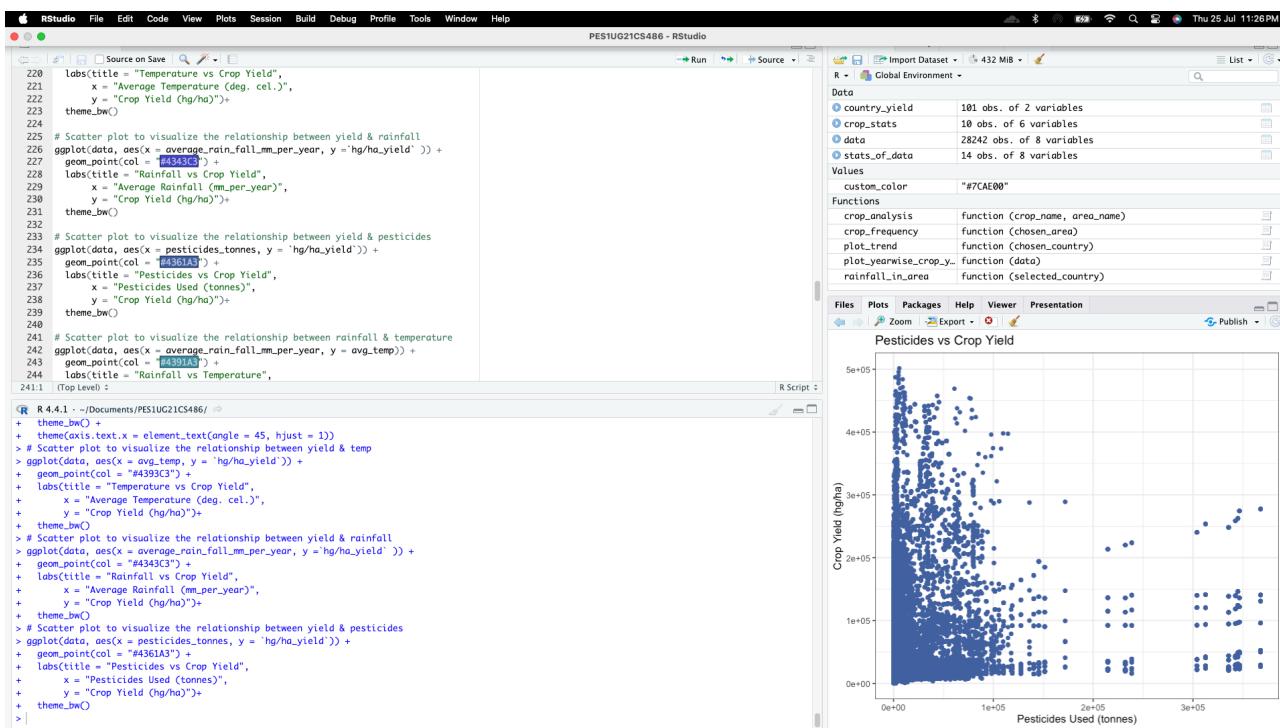


From these graphs, we can observe something very interesting. While countries like India and United Kingdom experience rising temperatures Botswana on the other hand has seen lowering temperatures. This makes me to believe that there is a bit of a negative correlation between temperature and mean yield. However temperature is not the sole factor in deciding the yield.

# Correlation Analysis

Using scatter plots, we can see how temperature, precipitation, and pesticide use affect yield:





We can now make a correlation Matrix for above relationships and is presented below along with a heat map visualisation for the same:

The screenshot shows the RStudio interface. On the left, the R console displays R code and its output, including a correlation matrix for variables like avg\_temp, pesticides\_tonnes, average\_rain\_fall\_mm\_per\_year, hg/ha\_yield, and avg\_temp. On the right, a correlation matrix heatmap is displayed, showing a strong positive correlation between avg\_temp and average\_rain\_fall\_mm\_per\_year, and a negative correlation between hg/ha\_yield and avg\_temp.

```

239 theme_bw()
240
241 # Scatter plot to visualize the relationship between rainfall & temperature
242 ggplot(data, aes(x = average_rain_fall_mm_per_year, y = avg_temp)) +
243   geom_point(col = "#E91E63") +
244   labs(title = "Rainfall vs Temperature",
245       x = "Average Rainfall (mm_per_year)",
246       y = "Average Temperature (deg. cel.)")+
247   theme_bw()
248
249 correlation_matrix = cor(data[, c("hg/ha_yield", "average_rain_fall_mm_per_year", "pesticides_tonnes", "avg_temp")])
250 correlation_matrix
251
252 library(reshape2)
253
254 ggplot(data = melt(correlation_matrix), aes(Var1, Var2, fill = value)) +
255   geom_tile(color = "white") +
256   scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1), space = "Lab", name="Correlation") +
257   theme_minimal() +
258   theme(axes.text.x = element_text(angle = 45, hjust = 1))
259
260 #Inferences
261 plot_yield_and_stats = function(crop_name, country_name) {
262   crop_data = subset(data, Item == crop_name & Area == country_name & Year >= 1990 & Year <= 2013)
263   max_yield_year = crop_data$Year[which.max(crop_data$hg/ha_yield")]
264 }
265
266 (Top Level) :

```

R 4.4.1 - ~/Documents/PES1UG21CS486/

```

+ x = "Average Rainfall (mm_per_year)",
+ y = "Average Temperature (deg. cel.)"
+ theme_bw()
> correlation_matrix = cor(data[, c("hg/ha_yield", "average_rain_fall_mm_per_year", "pesticides_tonnes", "avg_temp")])
> correlation_matrix

```

	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
hg/ha_yield	1.0000000000	0.0009621545	0.06408508	-0.11477699
average_rain_fall_mm_per_year	0.0009621545	1.0000000000	0.18098365	0.31303952
pesticides_tonnes	0.0640850877	0.1809836464	1.00000000	0.03094611
avg_temp	-0.1147769596	0.3130395215	0.03094611	1.00000000

```

> library(reshape2)

Attaching package: 'reshape2'

The following object is masked from 'package:tidyverse':
  smoths

> ggplot(data = melt(correlation_matrix), aes(Var1, Var2, fill = value)) +
+   geom_tile(color = "white") +
+   scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1), space = "Lab", name="Correlation") +
+   theme_minimal() +
+   theme(axes.text.x = element_text(angle = 45, hjust = 1))
> |

```

We see that yield is very less correlated with rainfall but more correlated with the amount of pesticides used. Also there is a negative correlation between yield and temperature. A strong positive correlation can also be seen between rainfall and temperature.

## Additional Statistics: Fitting Distributions

Now we can attempt to fit a few distributions to the yield of certain crops produced in India using the “fitdistrplus” package.

The screenshot shows the RStudio interface. On the left, the R console displays R code and its output, including goodness-of-fit statistics for normal, Weibull, and Lognormal distributions for potato, sweet potato, and wheat yields. On the right, four plots are shown: a histogram and theoretical densities plot, a Q-Q plot, a P-P plot, and an empirical and theoretical CDFs plot, all comparing normal, Weibull, and Lognormal distributions.

```

312 # View the summary
313 print("For Normal Distribution:")
314 print(best_fit1)
315 print("For Weibull Distribution:")
316 print(best_fit2)
317 print("For Lognormal Distribution:")
318 print(best_fit3)
319 }
320
321 yield_fit("Potatoes", "India")
322 yield_fit("Sweet potatoes", "India")
323 yield_fit("Wheat", "India")
324
325 (Top Level) :

R 4.4.1 - ~/Documents/PES1UG21CS486/
> yield_fit("Potatoes", "India")
[1] "For Normal Distribution:"
Goodness-of-fit statistics
  1-mle-normal
Kolmogorov-Smirnov statistic: 0.117950
Cramer-von Mises statistic: 1.439504
Anderson-Darling statistic: 9.953863

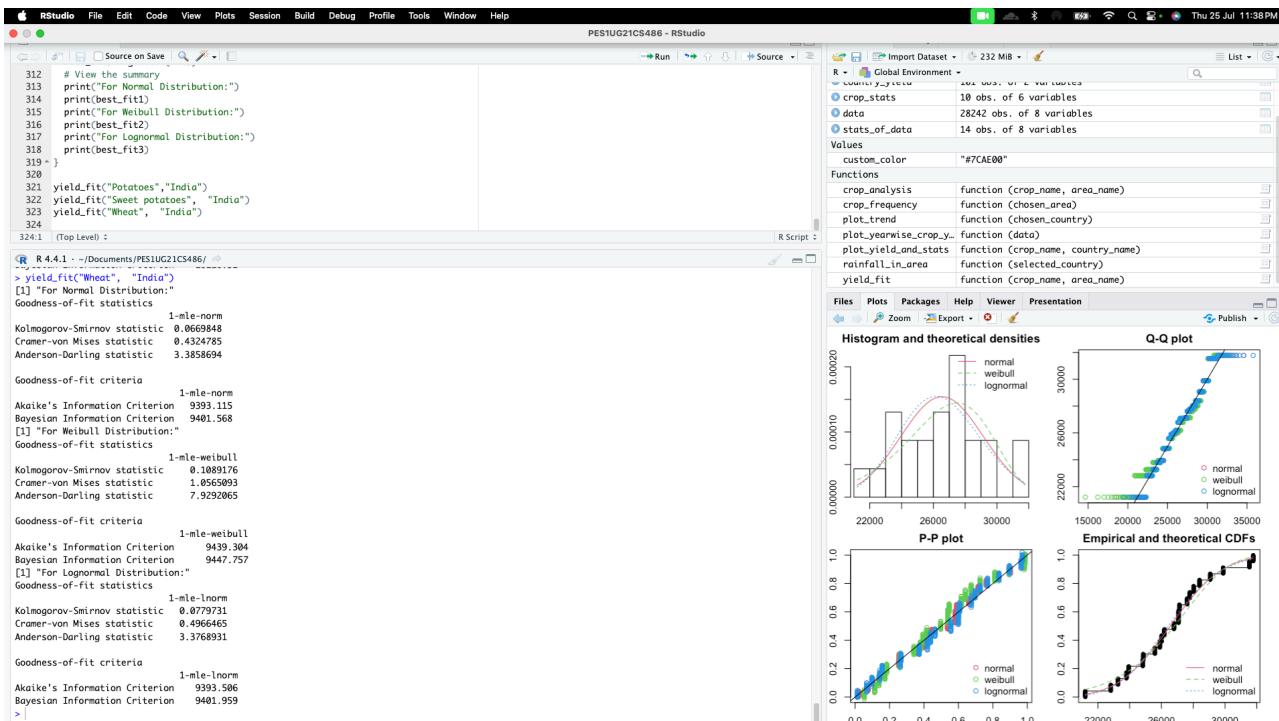
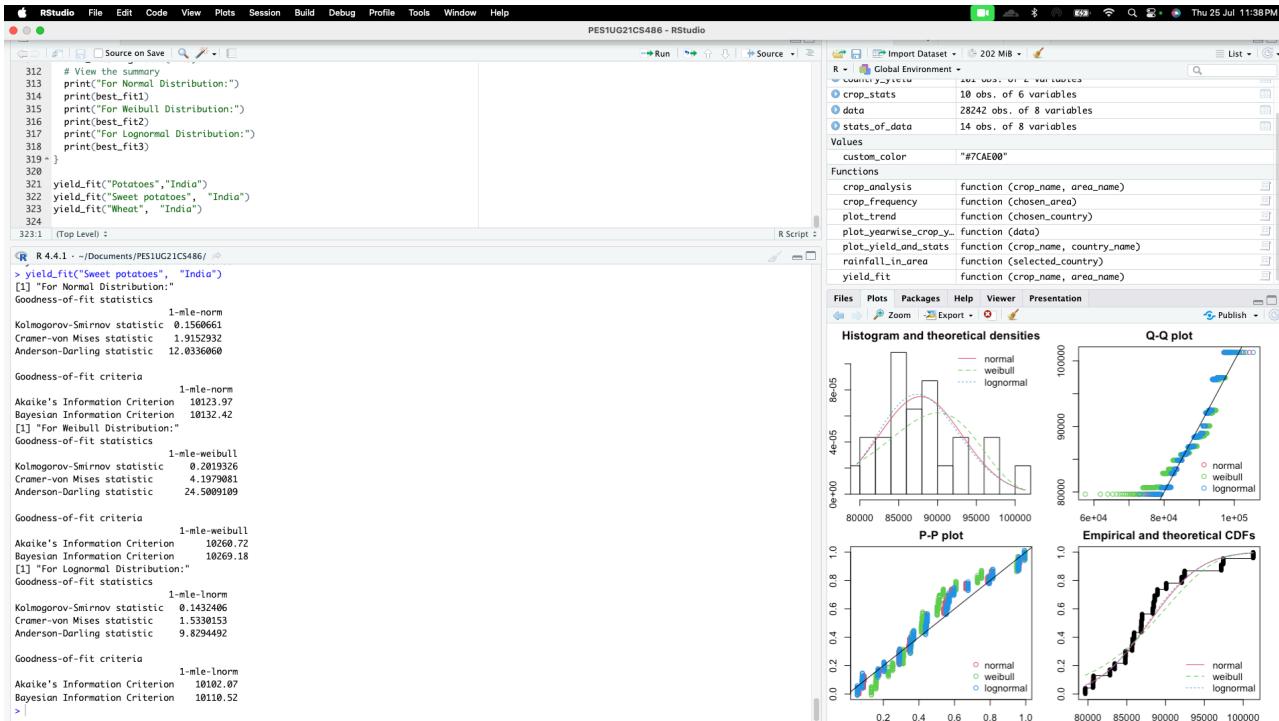
Goodness-of-fit criteria
  1-mle-normal
Akaike's Information Criterion: 11561.59
Bayesian Information Criterion: 11572.04
[1] "For Weibull Distribution:"
Goodness-of-fit statistics
  1-mle-weibull
Kolmogorov-Smirnov statistic: 0.1606313
Cramer-von Mises statistic: 2.1322611
Anderson-Darling statistic: 15.0668609

Goodness-of-fit criteria
  1-mle-weibull
Akaike's Information Criterion: 11627.95
Bayesian Information Criterion: 11636.40
[1] "For Lognormal Distribution:"
Goodness-of-fit statistics
  1-mle-lognorm
Kolmogorov-Smirnov statistic: 0.1176472
Cramer-von Mises statistic: 1.3723255
Anderson-Darling statistic: 8.4590345

Goodness-of-fit criteria
  1-mle-lognorm
Akaike's Information Criterion: 11544.19
Bayesian Information Criterion: 11552.64
> |

```

From the above information we can conclude that lognormal distribution fits the best for the yield of Potatoes in India. (low KS, AIC and BIC). In a similar fashion we can do so for other crops and countries too.



- Statistics like the AIC (Akaike information criterion), which estimates prediction error and, consequently, the relative quality of statistical models for a given set of data, are obtained through the fitting of distributions. Among a finite collection of models, the Bayesian information criterion (BIC) or Schwarz information criterion (also known as SIC, SBC, or SBIC) is a criterion for model selection; models with lower BIC are typically chosen.) Furthermore, the KS (Kolmogorov–Smirnov test, also known as the K–S test or KS test) is a nonparametric test of the

equality of continuous (or discontinuous), one-dimensional probability distributions. It can be used to compare two samples (the two-sample K-S test) or one sample with a reference probability distribution (the one-sample K-S test).

- Lower AIC and BIC values indicate a better fit. The KS statistic measures the maximum difference between the empirical distribution function of your data and the theoretical distribution. Smaller KS values suggest a better fit.

## Benefits of Fitting Distributions

It is beneficial to fit a probability distribution for a country's crop yield over a certain time period for a number of reasons.

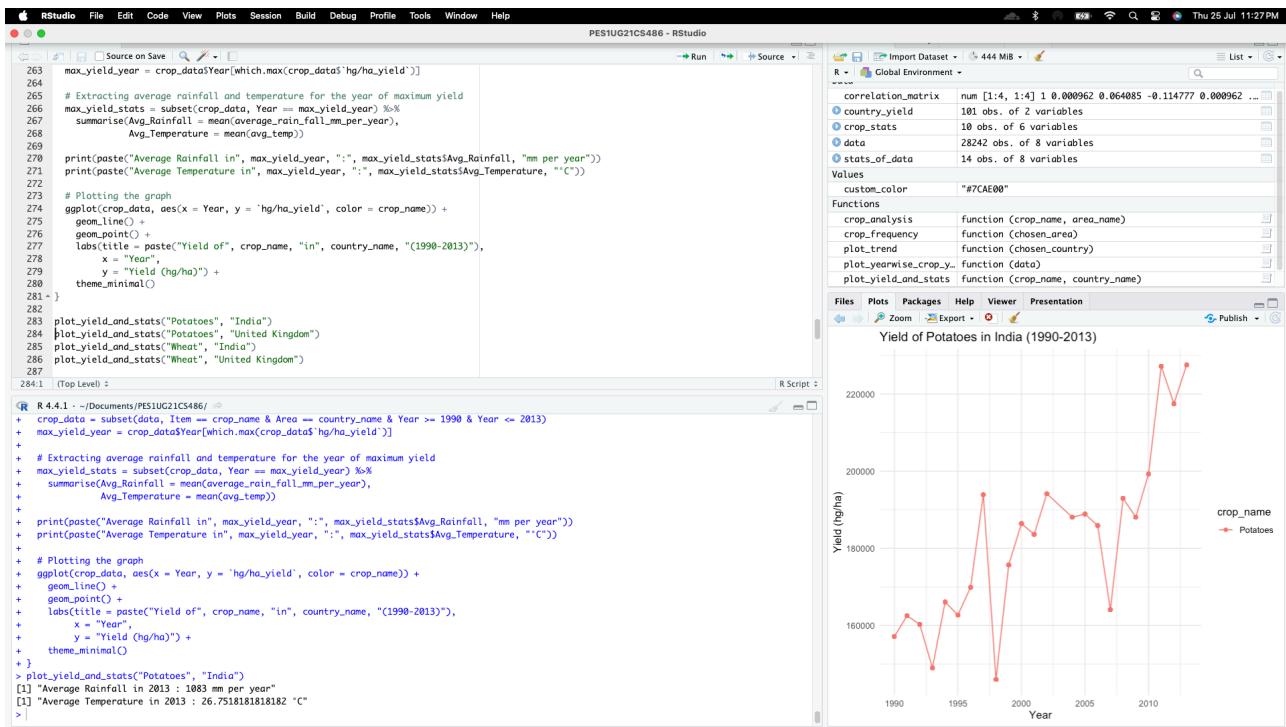
- 1) **Risk Assessment:** The risk connected to various yield levels can be evaluated with the use of probability distributions. Farmers and policymakers can more effectively evaluate and manage the risks involved in crop production by taking preemptive steps to reduce potential losses by knowing the distribution of possible outcomes.
- 2) **Decision Support:** The foundation for decision support is provided by fitted probability distributions. The distribution can help farmers make well-informed choices on how to allocate resources, manage their crops, and reduce risks. For example, the distribution of predicted yields can help determine decisions about planting schedules, crop insurance, and resource investments.
- 3) **Supply Chain Management:** Organisations and parties involved in the farming supply chain can forecast and prepare for changes in crop yield by using probability distributions. Pricing tactics, supply chain optimisation, and general market strategy all depend on this.
- 4) **Early Warning Systems:** For catastrophic occurrences like droughts or floods, early warning systems might include probability distributions. Authorities are able to promptly respond to and provide support systems for impacted regions by keeping an eye on deviations from planned yield distributions.

To summarise, the process of fitting a probability distribution to data on crop yields offers a strong foundation for agricultural decision-making, risk evaluation, and resource optimisation. It improves the capacity to foresee and adjust to changing circumstances, which eventually helps to create agricultural methods that are more robust and sustainable.

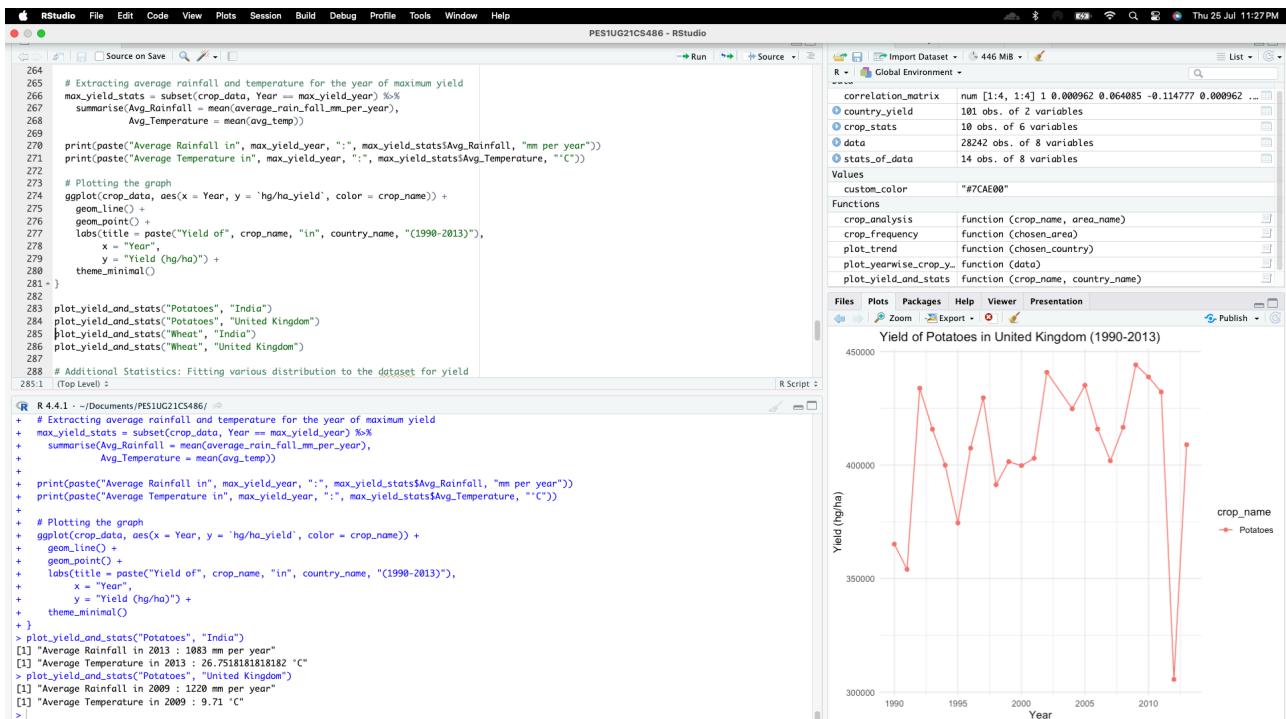
## Conclusion

Examining the agricultural yield graph over time in connection to meteorological variables like temperature and rainfall which applicable to any crop and nation:

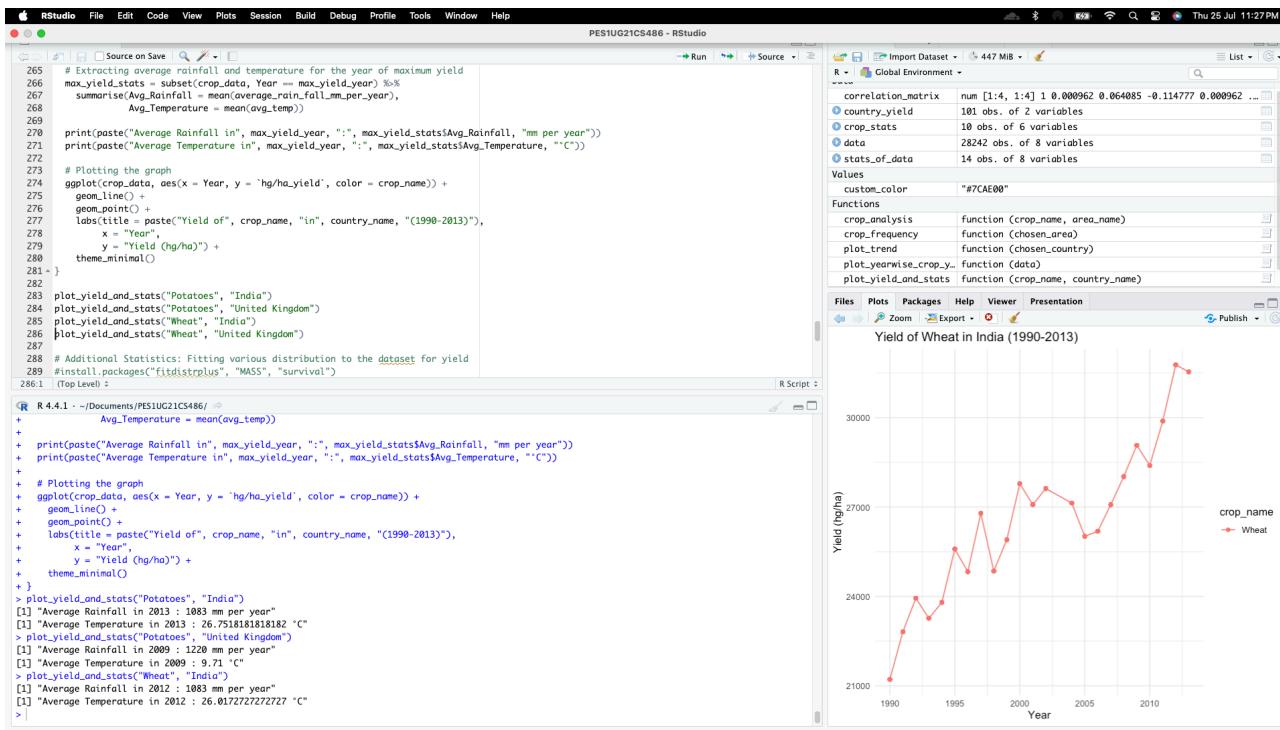
From the following graph we can conclude that following the trends seen everywhere in the world, India's yield of potatoes has increased significantly from 1990 to 2013 and it remains to be the highest yield crop in India.



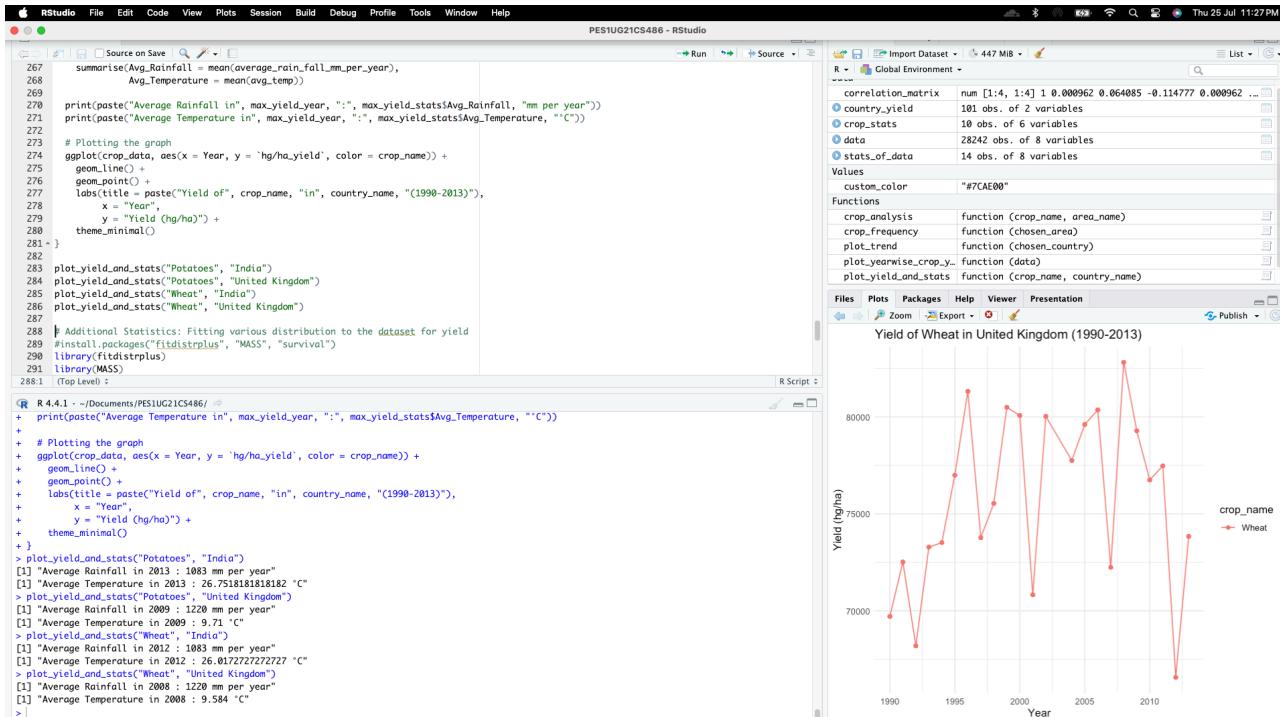
Analysing the trends of yield of potatoes in United Kingdom we see a sudden drop in 2012 which shows that in 2012 the yields in potatoes took a massive hit which could have been a result of several factors like bad harvest season, poor quality of pesticides or unfavourable weather conditions.



We can also see from the below grasp that yields of other crops, in this case Wheat has also significantly increased in India depicting good agricultural practices, favourable weather conditions etc.



Following graph showing Wheat yield also confirms our hypothesis of unfavourable conditions in United Kingdom in the year 2012 which also affected the yield of wheat significantly. We see a slight rise in 2013 just like the potatoes yield, however we can still see a significant decline in yield in United Kingdom from the 1990's.



## Findings

### 1. Effect of Pesticides:

- The analysis shows that increased use of pesticides in India from 2007 onwards has been associated with a rise in mean crop yield.
- A significant decrease in the mean yield of the United Kingdom post-2010 correlates with reduced pesticide usage.
- Botswana, with the lowest mean yield, also shows minimal pesticide usage, supporting the hypothesis that pesticide use directly affects crop yield.

### 2. Effect of Temperature:

- The trend analysis indicates that countries like India and the United Kingdom have experienced rising temperatures, while Botswana has seen a decrease in temperatures.
- There is a negative correlation between temperature and mean yield, suggesting that higher temperatures are generally associated with lower yields. However, temperature is not the sole factor influencing yield.

### 3. Effect of Rainfall:

- Rainfall patterns were analyzed to determine their impact on crop yield. In Albania, there was high rainfall from 1990-1992 but a significant drop in 1992. From 1994-2013, rainfall levels remained stable.
- The United Kingdom and India experienced consistent rainfall levels from 1990-2013, whereas Botswana saw a sudden drop in rainfall after 2008, which remained consistent until 2013.
- The average rainfall required for different crops was analyzed, showing varying needs. The analysis highlighted how rainfall impacts crop yield in different regions.

### 4. Outlier Detection:

- Outlier detection was conducted to identify outliers in crop yield data for different countries and crops. This method helps in understanding the anomalies and ensuring the accuracy of data analysis.
- For example, in the crop yield analysis for India, outliers were identified in the yield distributions of sweet potatoes whereas no outliers for potatoes, and wheat from 1990-2013

### 5. Correlation Analysis:

- A correlation matrix and heat map reveal that yield has a low correlation with rainfall but a stronger correlation with pesticide usage.
- There is a negative correlation between yield and temperature, and a strong positive correlation between rainfall and temperature.

### 6. Fitting Distributions:

- The analysis fitted various statistical distributions to crop yield data, finding that the lognormal distribution fits best for potato yields in India. This method can be extended to other crops and countries to support agricultural decision-making and risk assessment.

Here are some insights that can be gained:

- 1) **Identification of Optimal Conditions:** To determine the ideal growth conditions for a given crop, it is helpful to look at the times of maximum yield and the environmental variables that go along with it. The timing of planting and harvesting can be determined using this information.
- 2) **Identification of Critical Periods:** Targeted interventions are made possible by pinpointing the times during the crop growth cycle when environmental factors have a significant impact on production. For instance, farmers can put plans in place to lessen any negative effects during a growth stage that is susceptible to temperature fluctuations.
- 3) **Risk management:** Identifying years with reduced yields and looking at the related environmental factors aid in risk assessment and management. Based on past performance, farmers might employ risk-reduction techniques including crop diversification and resilient variety purchases.
- 4) **Precision Agriculture:** The application of precision agriculture methods can be guided by the analysis's conclusions. Farmers are able to apply precision irrigation, fertilisation, and pest control tactics that are customised to certain regions of their fields by utilising data on past production changes and related variables.
- 5) **Crop Rotation and Diversification:** Strategic crop rotation and diversification are made possible by an understanding of how environmental factors affect certain crops. Crop combinations that compliment one another and are resistant to various environmental stresses might be chosen by farmers.

## References

- Dataset -> [https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset?select=yield\\_df.csv](https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset?select=yield_df.csv)
- <https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset/discussion/436426>
- <https://www.kaggle.com/code/noujoudgabed/eda-crop-yield>
- <https://eos.com/blog/crop-yield-increase>
- <https://www.intechopen.com/chapters/70658>

## Appendix (R Program)

```

library(readr)
data <- read_csv("/Users/rhian/Documents/PES1UG21CS486/
yield_df.csv")

#Attribute names & types
names(data)
sapply(data, class)

#Checking for missing values
any(is.na(data))

# Displaying the structure of your data
str(data)
# Displaying the first few rows of your data
head(data)

library(pastecs)
stats_of_data <- format(stat.desc(data), scientific = FALSE)
print(stats_of_data)

library(ggplot2)

# Analyzing Item
table(data$Item)

pie(table(data$Item), main = "Crop Frequency Distribution", col =
terrain.colors(12))

crop_frequency <- function(chosen_area){
  # Filtering data for the selected area
  filtered_data = data[data$Area == chosen_area, ]

  # Plotting a bar chart for the frequency of different crops
  ggplot(filtered_data, aes(x = Item, fill = Item)) +
    geom_bar() +
    labs(title = paste("Crop Frequency in", chosen_area),
        x = "Crops",
        y = "Frequency") +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
}

crop_frequency("Albania")
crop_frequency("Botswana")
crop_frequency("United Kingdom")
crop_frequency("India")

print('The crop which is grown in most of the countries is:')
which.max(table(data$Item))

```

```
#Year-wise Crop Yield Analysis
library(dplyr)
library(scales)
plot_yearwise_crop_yield <- function(data) {
  # Grouping the data by Year and Item to get the sum of yield of
  each crop per year
  yearwise_crop_yield <- data %>%
    group_by(Year, Item) %>%
    summarise(total_yield = sum(`hg/ha_yield`)) %>%
    ungroup()

  # Plotting the data using ggplot2
  ggplot(yearwise_crop_yield, aes(x = Year, y = total_yield, color
= Item)) +
    geom_line(linewidth = 1.2) + # Thicker lines for better
  visibility
    labs(title = "Yearwise crop-yield",
        x = "Year",
        y = "Yield (hg/ha)",
        color = "Crop") +
    scale_y_continuous(labels = comma) +
    theme_minimal() +
    theme(
      axis.text.x = element_text(angle = 45, hjust = 1),
      plot.title = element_text(hjust = 0.5),
      legend.position = "right"
    )
}
plot_yearwise_crop_yield(data)

library(modeest)

# Grouping data by crop
crop_stats = data %>%
  group_by(Item) %>%
  reframe(
    Mean_Yield = mean(`hg/ha_yield`),
    Median_Yield = median(`hg/ha_yield`),
    Mode_Yield = mfv(`hg/ha_yield`)[1],
    Max_yield = max(`hg/ha_yield`),
    Min_yield = min(`hg/ha_yield`)
  )
print(crop_stats)

# Plotting a barplot of various crops
ggplot(data, aes(x = Item, y = `hg/ha_yield`, fill = Item)) +
  geom_bar(stat = "identity") +
  labs(title = "Crop Yield Analysis",
       x = "Crops",
       y = "Yield (hg/ha)") +
  theme_bw()
```

```

  theme(axis.text.x = element_text(angle = 90, hjust = 1))

library(tidyverse)

# Trend analysis using ggplot2
ggplot(data, aes(x = Year, y = `hg/ha_yield`, color = Item)) +
  geom_line() +
  labs(title = "Trend Analysis of Agriculture Yield Over Time",
       x = "Year",
       y = "Yield (hg/ha)") +
  theme_bw()

# Calculating mean yield for each country
country_yield = data %>%
  group_by(Area) %>%
  summarize(mean_yield = mean(`hg/ha_yield`))

# Trends over the years for different countries
plot_trend <- function(chosen_country){
  trend_plot = data %>%
    filter(Area == chosen_country) %>%
    ggplot(aes(x = Year, y = `hg/ha_yield`)) +
    geom_line(color = "turquoise", lwd = 1) +
    labs(title = "Trend of Yield Over Years in the Chosen Country",
         x = "Year", y = "Yield (hg/ha)") +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))

  print(trend_plot)
}

plot_trend("United Kingdom")
plot_trend("India")
plot_trend("Botswana")

# Creating a bar plot to compare mean yield between countries
ggplot(country_yield, aes(x = reorder(Area, -mean_yield), y =
mean_yield)) +
  geom_bar(stat = "identity", fill = "yellow", color = "black") +
  labs(title = "Mean Yield Comparison between Countries", x =
"Country", y = "Mean Yield (hg/ha)") +
  theme(axis.text.x = element_text(angle = 90, size = 4.5))

# Crop-wise analysis for a specific crop in a particular area from 1990-2013
crop_analysis = function(crop_name, area_name) {
  # Filtering data for the specific crop, area, and time period
  crop_data = subset(data, Item == crop_name & Area == area_name &
Year >= 1990 & Year <= 2013)

  # Displaying summary statistics
}

```

```

  print(paste("Summary Statistics for", crop_name, "in",
area_name, "from 1990-2013"))
  print(summary(crop_data$`hg/ha_yield`))

# Creating a boxplot
  boxplot(crop_data$`hg/ha_yield`, main = paste(crop_name, "Yield
Distribution in", area_name), ylab = "Yield (hg/ha)", col =
"#D6604D")

# Identification of outliers using the Tukey method
  outliers = boxplot.stats(crop_data$`hg/ha_yield`)$.out

# Displaying outliers
  print("Outliers are:")
  print(outliers)
}

crop_analysis("Potatoes","India")
crop_analysis("Sweet potatoes","India")
crop_analysis("Wheat","India")

#Effect of rainfall

ggplot(data, aes(x = Item, y = average_rain_fall_mm_per_year, fill
= Item)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge",
color = "black", fill = "#7443AF") +
  labs(title = "Average Rainfall Required for Different Crops",
      x = "Crops",
      y = "Average Rainfall (mm per year)") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Rainfall amount for a particular area

rainfall_in_area <- function(selected_country){
  # Filtering data for the selected country
  filtered_data = data[data$Area == selected_country, ]
  filtered_data$Year <-
as.numeric(as.character(filtered_data$Year))
  # Filtering data for the specified period (1990-2013)
  filtered_data = filtered_data[filtered_data$Year >= 1990 &
filtered_data$Year <= 2013, ]
  # Plotting a bar chart for average rainfall
  ggplot(filtered_data, aes(x = as.factor(Year), y =
average_rain_fall_mm_per_year, fill = as.factor(Year))) +
    geom_bar(stat = "identity") +
    labs(title = paste("Average Rainfall in", selected_country, "
(1990-2013")),
        x = "Year",
        y = "Average Rainfall (mm per year)") +
    theme_bw()+
}

```

```

    theme(axis.text.x = element_text(angle = 90, hjust = 1))
}
rainfall_in_area("Albania")
rainfall_in_area("United Kingdom")
rainfall_in_area("India")
rainfall_in_area("Botswana")

# Effect of pesticide

custom_color <- "#7CAE00"
ggplot(data, aes(x = Item, y = pesticides_tonnes, fill = Item)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge",
color = "black", fill = custom_color) +
  labs(title = "Average Pesticides Required for Different Crops",
    x = "Crops",
    y = "Average Pesticides (tonnes)") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

pesticide_trend <- function(country_name) {
  # Filter data for the selected country and the specified period
  filtered_data <- data %>%
    filter(Area == country_name & Year >= 1990 & Year <= 2013)

  # Convert Year to a numeric type if it's not already
  filtered_data$Year <- as.numeric(as.character(filtered_data$Year))

  # Plot the pesticide usage trend
  ggplot(filtered_data, aes(x = Year, y = pesticides_tonnes)) +
    geom_line(color = "#4393C3", linewidth = 1) +
    labs(title = paste("Pesticide Usage Trend in", country_name,
"(1990-2013")),
      x = "Year",
      y = "Pesticide Usage (tonnes)") +
    theme_minimal() +
    theme(
      axis.text.x = element_text(angle = 90, hjust = 1),
      plot.title = element_text(hjust = 0.5)
    )
}

pesticide_trend("India")
pesticide_trend("United Kingdom")
pesticide_trend("Botswana")

# Effect of temperature

ggplot(data, aes(x = Item, y = avg_temp, fill = Item)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge",
color = "black", fill = "#FD7446FF") +
  labs(title = "Average Temperature Required for Different Crops",

```

```

      x = "Crops",
      y = "Average Temperature (deg celsius)") +
theme_bw() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))

temperature_trend <- function(country_name) {
  # Filter data for the selected country and the specified period
  filtered_data <- data %>%
    filter(Area == country_name & Year >= 1990 & Year <= 2013)

  # Convert Year to a numeric type if it's not already
  filtered_data$Year <- as.numeric(as.character(filtered_data$Year))

  # Plot the temperature trend
  ggplot(filtered_data, aes(x = Year, y = avg_temp)) +
    geom_line(color = "#4393C3", linewidth = 1) +
    labs(title = paste("Temperature Trend in", country_name,
  "(1990-2013")),
        x = "Year",
        y = "Average Temperature (°C)") +
    theme_minimal() +
    theme(
      axis.text.x = element_text(angle = 90, hjust = 1),
      plot.title = element_text(hjust = 0.5)
    )
}

temperature_trend("India")
temperature_trend("United Kingdom")
temperature_trend("Botswana")

```

### #Correlation analysis

```

# Scatter plot to visualize the relationship between yield & temp
ggplot(data, aes(x = avg_temp, y = `hg/ha_yield`)) +
  geom_point(col = "#4393C3") +
  geom_smooth(method = "loess", se = TRUE, color = "red") +
  labs(title = "Temperature vs Crop Yield",
       x = "Average Temperature (deg. cel.)",
       y = "Crop Yield (hg/ha)") +
  theme_bw()

```

```
library(mgcv)
```

```

# Scatter plot to visualize the relationship between yield &
rainfall
ggplot(data, aes(x = average_rain_fall_mm_per_year, y = `hg/
ha_yield` )) +
  geom_point(col = "#4343C3") +

```

```

geom_smooth(method = "gam", formula = y ~ s(x, bs = "cs"), se =
TRUE, color = "red") +
  labs(title = "Rainfall vs Crop Yield",
       x = "Average Rainfall (mm_per_year)",
       y = "Crop Yield (hg/ha)")+
  theme_bw()

# Scatter plot to visualize the relationship between yield &
pesticides
ggplot(data, aes(x = pesticides_tonnes, y = `hg/ha_yield`)) +
  geom_point(col = "#4361A3") +
  labs(title = "Pesticides vs Crop Yield",
       x = "Pesticides Used (tonnes)",
       y = "Crop Yield (hg/ha)")+
  theme_bw()

# Scatter plot to visualize the relationship between rainfall &
temperature
ggplot(data, aes(x = average_rain_fall_mm_per_year, y = avg_temp)) +
  geom_point(col = "#4391A3") +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Rainfall vs Temperature",
       x = "Average Rainfall (mm_per_year)",
       y = "Average Temperature (deg. cel.)")+
  theme_bw()

correlation_matrix = cor(data[, c( "hg/
ha_yield", "average_rain_fall_mm_per_year", "pesticides_tonnes",
"avg_temp")])
correlation_matrix

library(reshape2)

ggplot(data = melt(correlation_matrix), aes(Var1, Var2, fill =
value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
midpoint = 0, limit = c(-1, 1), space = "Lab", name="Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#Inferences
plot_yield_and_stats = function(crop_name, country_name) {
  crop_data = subset(data, Item == crop_name & Area ==
country_name & Year >= 1990 & Year <= 2013)
  max_yield_year = crop_data$Year[which.max(crop_data$hg/
ha_yield)]  

  # Extracting average rainfall and temperature for the year of
maximum yield
}

```

```

max_yield_stats = subset(crop_data, Year == max_yield_year) %>%
  summarise(Avg_Rainfall = mean(average_rain_fall_mm_per_year),
            Avg_Temperature = mean(avg_temp))

print(paste("Average Rainfall in", max_yield_year, ":", max_yield_stats$Avg_Rainfall, "mm per year"))
print(paste("Average Temperature in", max_yield_year, ":", max_yield_stats$Avg_Temperature, "°C"))

# Plotting the graph
ggplot(crop_data, aes(x = Year, y = `hg/ha_yield`, color = crop_name)) +
  geom_line() +
  geom_point() +
  labs(title = paste("Yield of", crop_name, "in", country_name,
                     "(1990-2013")),
       x = "Year",
       y = "Yield (hg/ha)") +
  theme_minimal()
}

plot_yield_and_stats("Potatoes", "India")
plot_yield_and_stats("Potatoes", "United Kingdom")
plot_yield_and_stats("Wheat", "India")
plot_yield_and_stats("Wheat", "United Kingdom")
plot_yield_and_stats("Wheat", "Botswana")

# Additional Statistics: Fitting various distribution to the dataset for yield
#install.packages("fitdistrplus", "MASS", "survival")
library(fitdistrplus)
library(MASS)
library(survival)

yield_fit <- function(crop_name, area_name) {
  crop_data = subset(data, Item == crop_name & Area == area_name & Year >= 1990 & Year <= 2013)
  datatoplot = crop_data$`hg/ha_yield`
  # Fit Poisson distribution (or other discrete distribution) and compare
  fit1 = fitdist(datatoplot, "norm")
  fit2 = fitdist(datatoplot, "weibull")
  fit3 = fitdist(datatoplot, "lnorm")
  par(mfrow = c(2,2))
  plot.legend = c("normal", "weibull", "lognormal")
  par(mar = c(2,2,2,2))
  denscomp(list(fit1, fit2, fit3), legendtext = plot.legend)
  qqcomp(list(fit1, fit2, fit3), legendtext = plot.legend)
  ppcomp(list(fit1, fit2, fit3), legendtext = plot.legend)
  cdfcomp(list(fit1, fit2, fit3), legendtext = plot.legend)
  # Summary of the best fit
  best_fit1 = gofstat(fit1)
}

```

```
best_fit2 = gofstat(fit2)
best_fit3 = gofstat(fit3)
# View the summary
print("For Normal Distribution:")
print(best_fit1)
print("For Weibull Distribution:")
print(best_fit2)
print("For Lognormal Distribution:")
print(best_fit3)
}

yield_fit("Potatoes", "India")
yield_fit("Sweet potatoes", "India")
yield_fit("Wheat", "India")
```