EXPLORING AGRICULTURE YIELD DYNAMICS

A Comprehensive Analysis of Environmental Factors and Crop Production

LAKSHYA SINGH (23MDT1019)

ANIKET KUMAR (23MST1007)

UPADHYE RUSHIKESH SUNIL (23MDT1059)

COURSE: PROBABILITY THEORY & DISTRIBUTIONS (LAB)

COURSE CODE: MAT5012

TABLE OF CONTENTS

Abstract

This study delves into the intricate dynamics of agricultural yield by examining a diverse dataset including crop yield, rainfall, pesticides, and temperature across various regions. Employing statistical techniques and visualizations, our study explores temporal trends, regional variations, and the impact of environmental factors on crop production. Through correlation analyses, we identify relationships between yield and climate variables. The entire analysis is done in the R programming language. Additionally, we employ the 'fitdistrplus' package in R to find the most suitable distribution for yield data. The results showcase the significance of understanding the interplay between meteorological conditions and agricultural output.

Introduction to dataset

The dataset employed in this analysis contains information on agricultural yield, including key variables such as geographical region, crop type, year, yield (in hectogram per hectare), average rainfall (in mm per year), pesticides usage (in tonnes), and average temperature (in degree celsius). Covering multiple countries, the dataset provides a comprehensive view of how these factors interact and influence crop production over time. The dataset's temporal scope spans from 1990, allowing for a thorough analysis of trends and patterns. With a focus on essential crops like Potatoes, Rice, Maize and Sorghum, the dataset facilitates a detailed understanding of the complexities inherent in agricultural systems.

Description of Attributes

The dataset used comprises of the following attributes:

```r
data = yield_df
#attribute names & types
names(data)
```

```
## [1] "...1"                          "Area"
## [3] "Item"                          "Year"
## [5] "hg/ha_yield"                   "average_rain_fall_mm_per_year"
## [7] "pesticides_tonnes"             "avg_temp"
```

Data types of the attributes are:

```
sapply(data, class)

##                          ...1                          Area
##                     "numeric"                   "character"
##                          Item                          Year
##                   "character"                     "numeric"
##                   hg/ha_yield average_rain_fall_mm_per_year
##                     "numeric"                     "numeric"
##             pesticides_tonnes                      avg_temp
##                     "numeric"                     "numeric"
```

Missing values in the dataset: (No missing values)

```
#check for missing values
any(is.na(data))

## [1] FALSE
```

Structure of the dataset

```
# Display the structure of your data
str(data)

## spc_tbl_ [28,242 × 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1                         : num [1:28242] 0 1 2 3 4 5 6 7 8 9 ...
##  $ 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] 121 121 121 121 121 121
121 121 121 121 ...
##  $ avg_temp                     : num [1:28242] 16.4 16.4 16.4 16.4 16.4
```

Displaying first few rows of the dataset:

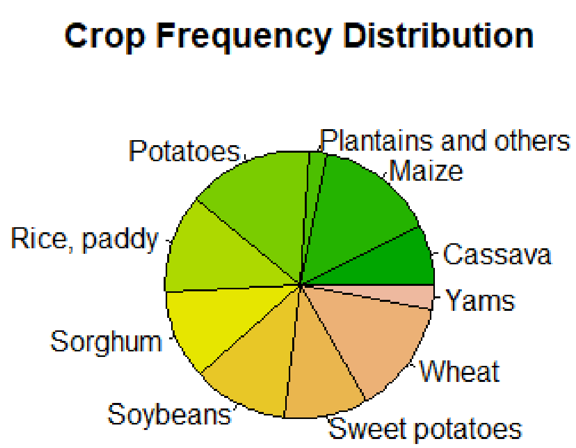
	...1	Area	Item	Year	hg/ha_yield	average_rain_fall_mm_per_year	pesticides_tonnes	avg_temp
1	0	Albania	Maize	1990	36613	1485	121.00	16.37
2	1	Albania	Potatoes	1990	66667	1485	121.00	16.37
3	2	Albania	Rice, paddy	1990	23333	1485	121.00	16.37
4	3	Albania	Sorghum	1990	12500	1485	121.00	16.37
5	4	Albania	Soybeans	1990	7000	1485	121.00	16.37
6	5	Albania	Wheat	1990	30197	1485	121.00	16.37
7	6	Albania	Maize	1991	29068	1485	121.00	15.36
8	7	Albania	Potatoes	1991	77818	1485	121.00	15.36

Crop Item Analysis

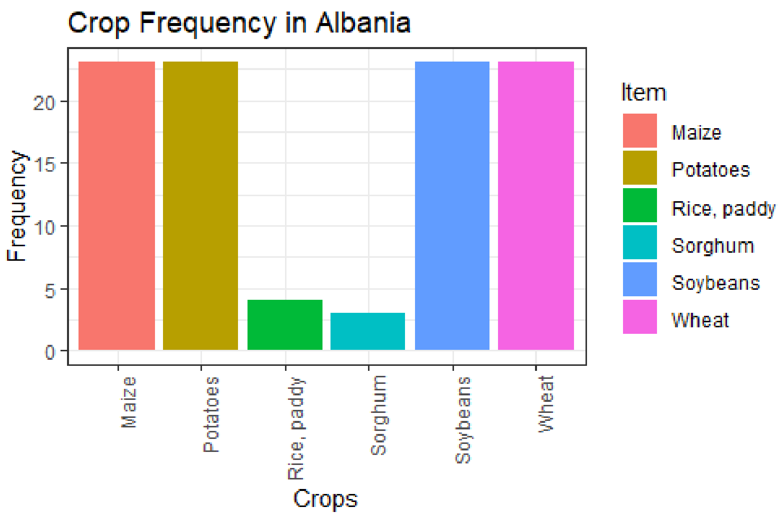
Frequency of the each crop grown in different regions of the world from 1990-2013 is given by:

Out of all crops, Potatoes are the most preferred crop around the world while Yams are the least.

```
# Analysing Item
table(data$Item)

##
##                Cassava                  Maize Plantains and others
##                   2045                   4121                  556
##               Potatoes            Rice, paddy              Sorghum
##                   4276                   3388                 3039
##               Soybeans         Sweet potatoes                Wheat
##                   3223                   2890                 3857
##                   Yams
##                    847
```

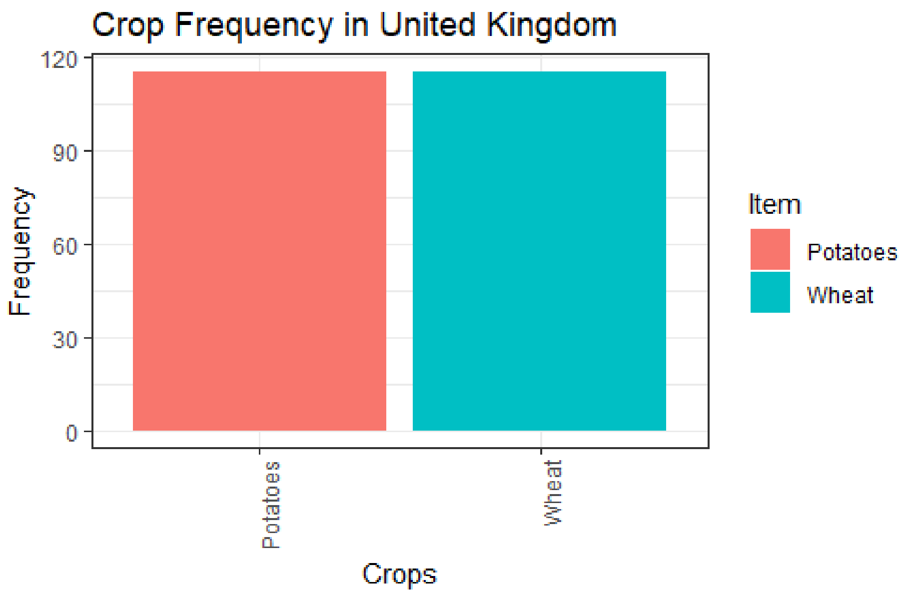


Following graph shows for a particular geographical area, the kinds of crops grown and the frequency of each crop item. For **Albania**, we see an equal preference for crops like Maize, Potatoes, Soybeans and Wheat while crops like Rice and Sorghum have an extremely low preference.

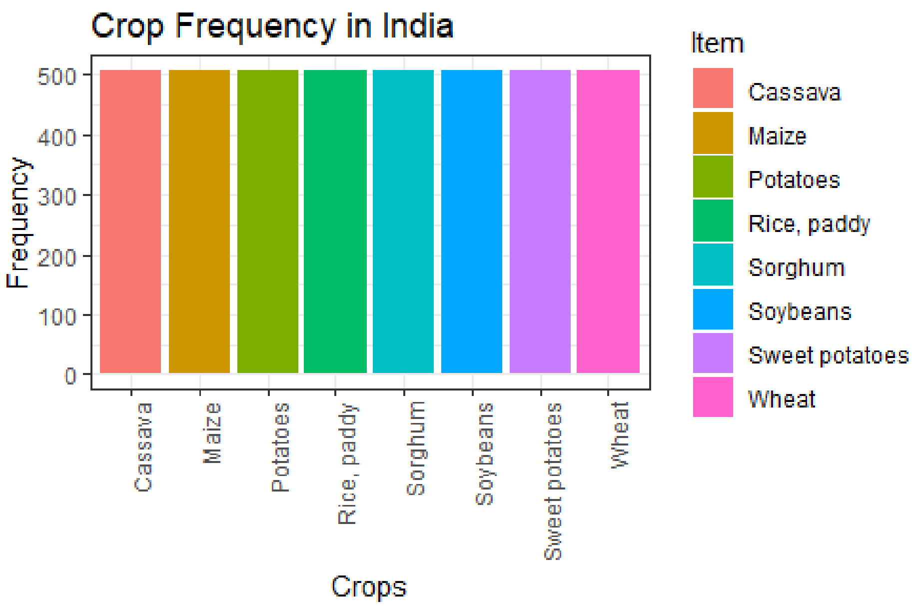


In a similar way we can have barplots for crop preference in different countries, such as United States, India, Botswana, etc

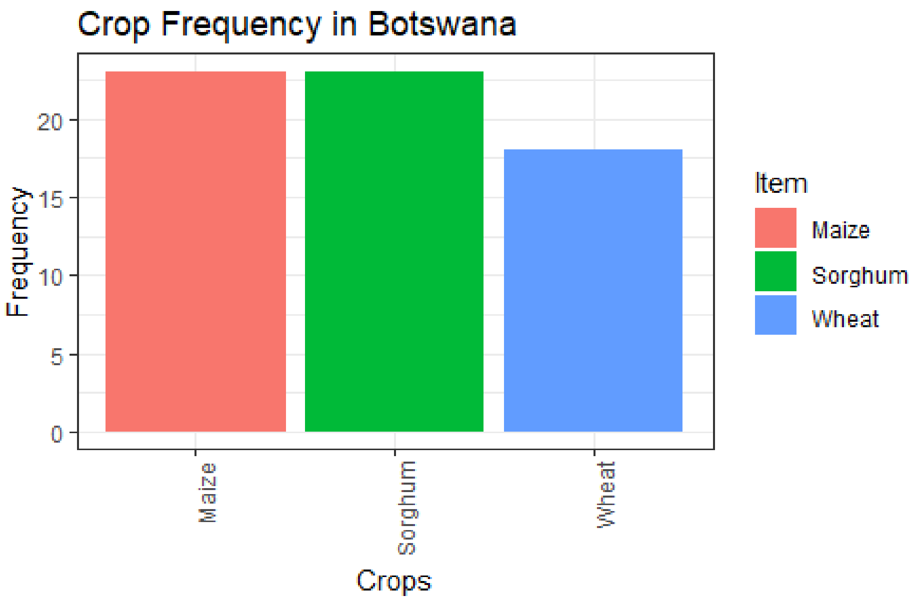
In **United Kingdom**, Potatoes and Wheat were the most eaten crops between 1990-2013.



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



In **Botswana**, only Maize, Sorghum and Wheat were grown between 1990-2013, out of which less preference was given to Wheat.

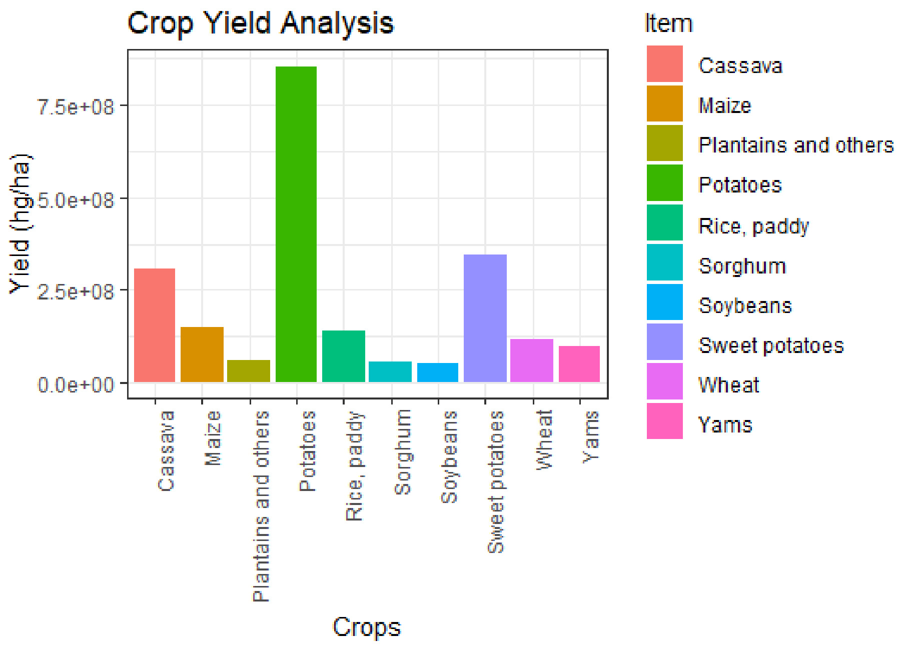


Crop Yield Analysis

Descriptive statistics of yield for different crop item:

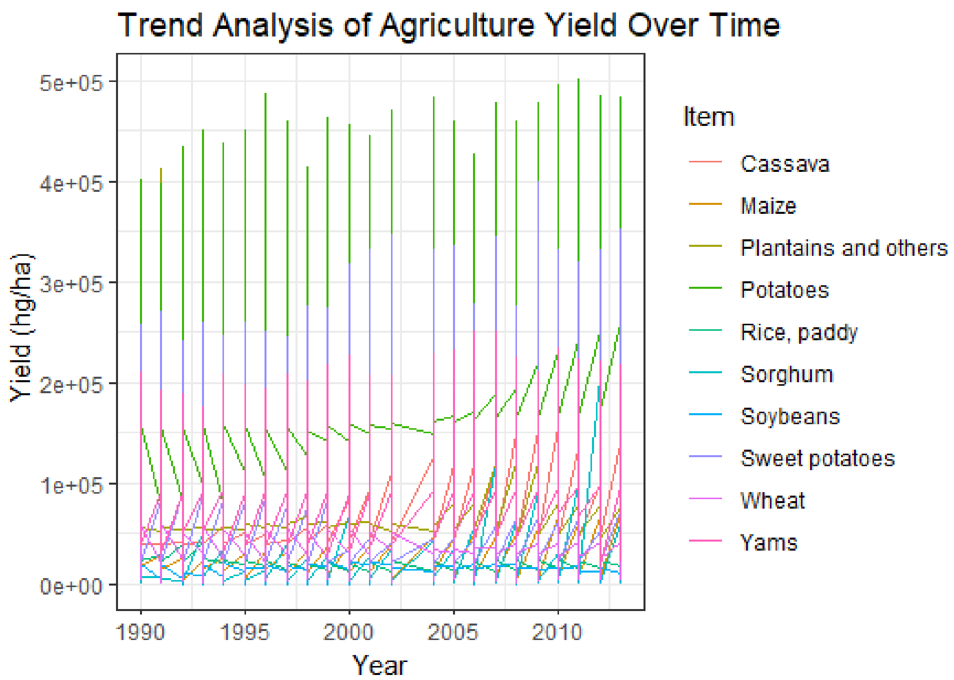
	Item	Mean_Yield	Median_Yield	Mode_Yield	Max_yield	Min_yield
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Cassava	150479.	128200	100000	385818	11778
2	Maize	36310.	25401	25000	207556	849
3	Plantains and others	106041.	89860.	100000	418505	21350
4	Potatoes	199802.	182271	169913	501412	8406
5	Rice, paddy	40730.	35878	32924	103895	2034
6	Sorghum	18636.	12885	7968	206000	578
7	Soybeans	16731.	15533	10000	41609	50
8	Sweet potatoes	119058.	99940	79663	400000	8799
9	Wheat	30116.	25497	21211	99387	1706
10	Yams	114140.	92593	92000	250000	11475

Following barplot compares yield of different crops grown in different areas of the world:

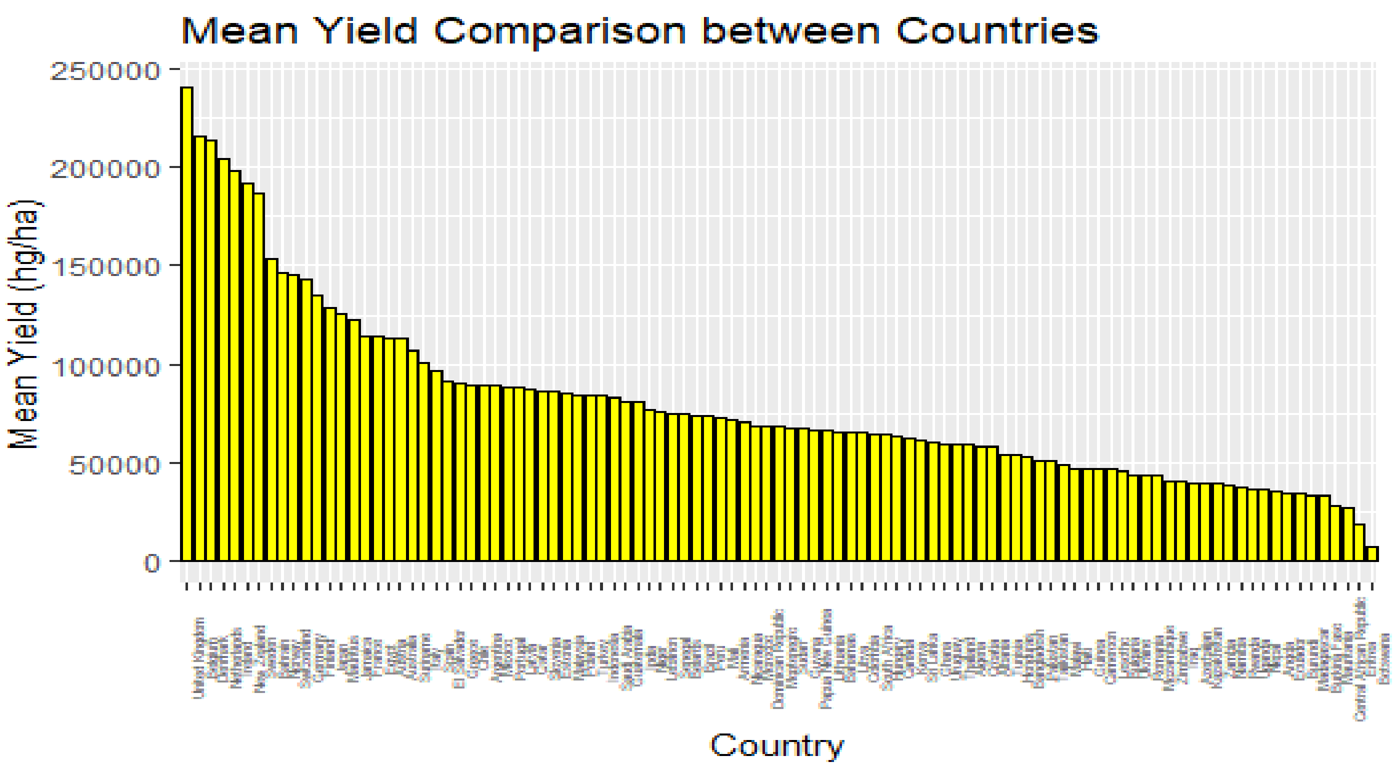


Above graph clearly tells that Potatoes have been the most preferred crop all around the world with a maximum yield of 501412 hg/ha. On the other hand we have crops like Plantains, Sorghum and Soybeans whose maximum yield stood at the same level.

Over the years, the trends of agricultural yield for different varieties of the crops have changes in the following way:

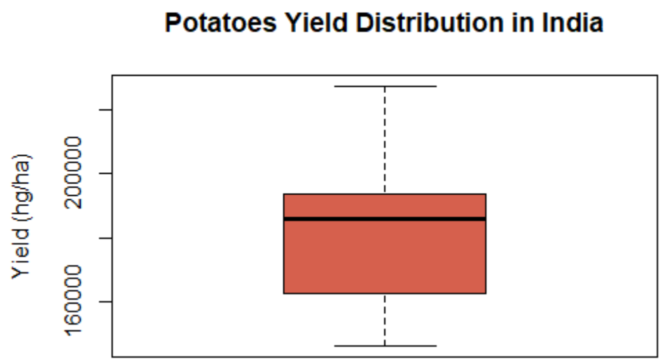


The mean yield comprising of all the crops for different areas can be shown through the following graph:

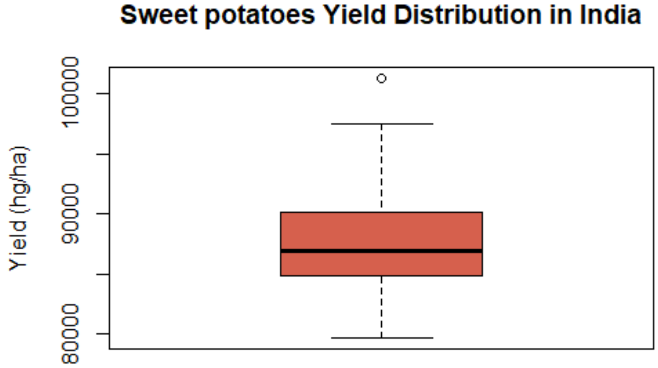


The graph shows that United Kingdom has the highest value of mean yield over the time period of 1990 to 2013 while Botswana has the lowest. The trend of the yield for United Kingdom and Botswana can compared using the following graphs:

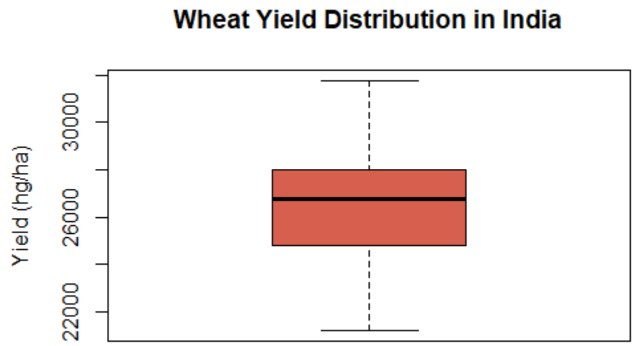
Outlier detection: The outlier values of net yield (in hectogram per hectare) for a particular crop for a country such as **India** can also be found out using a boxplot and applying Tukey's Method. For some of the crops outlier value of yield (in hectogram per hectare) is given by:




```
[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)
```




```
[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
[12] 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288
```

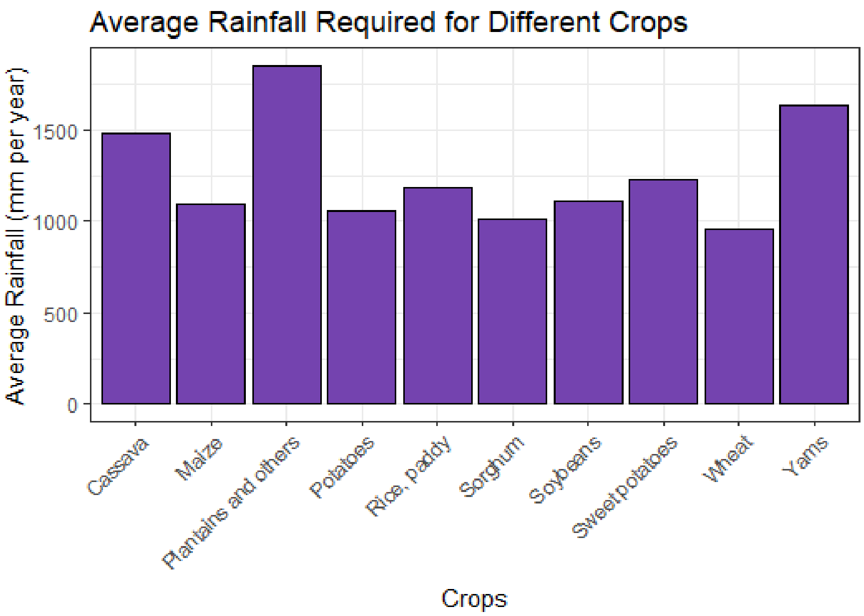


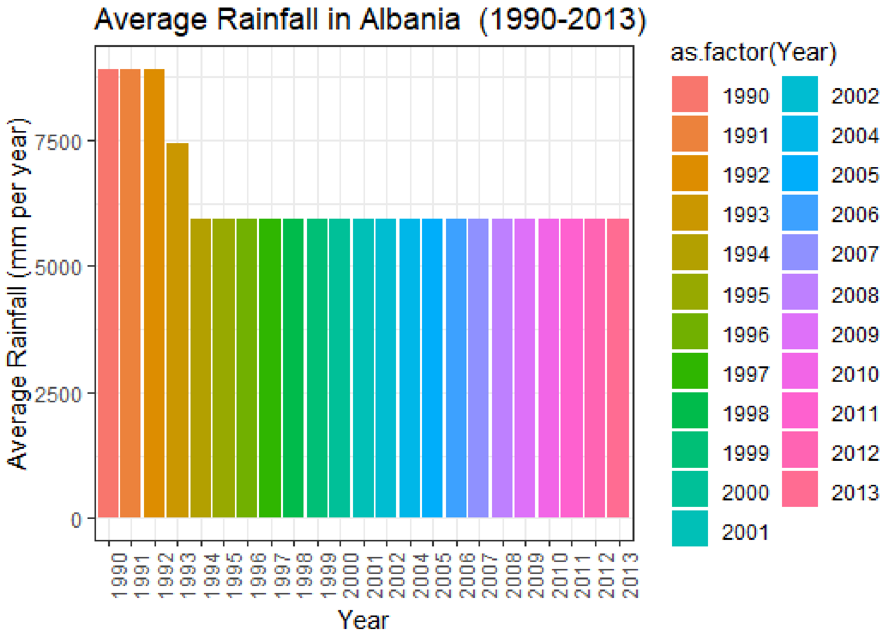

```
[1] "Summary Statistics for Wheat in India from 1990-2013:"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21211   24828   26789   26547   28022   31775
[1] "Outliers are:"
numeric(0)
```

Effect of Rainfall

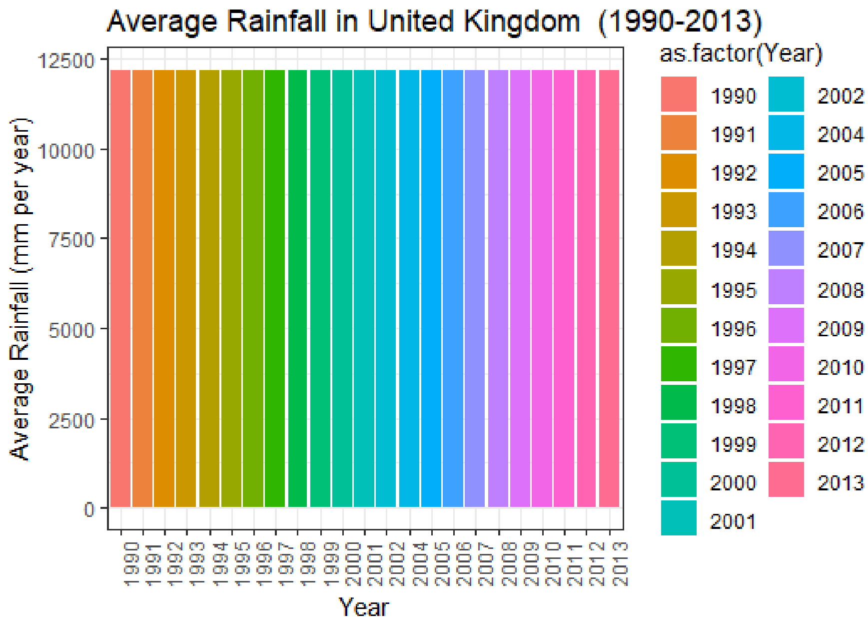
Different varieties of crops require different amount of rainfall every year which can be estimated using the following barchart:

Also the average amount of rainfall received by a particular area changes over a period of time. It can be clearly seen that Plantains and Yams require a huge amount of rainfall while crops like Maize, Potatoes, Sorghum, Wheat require almost the same amount of rainfall. In the following graphs, we have shown rainfall trends over a period of 1990-2013 for countries like Albania, United Kingdom and Botswana.

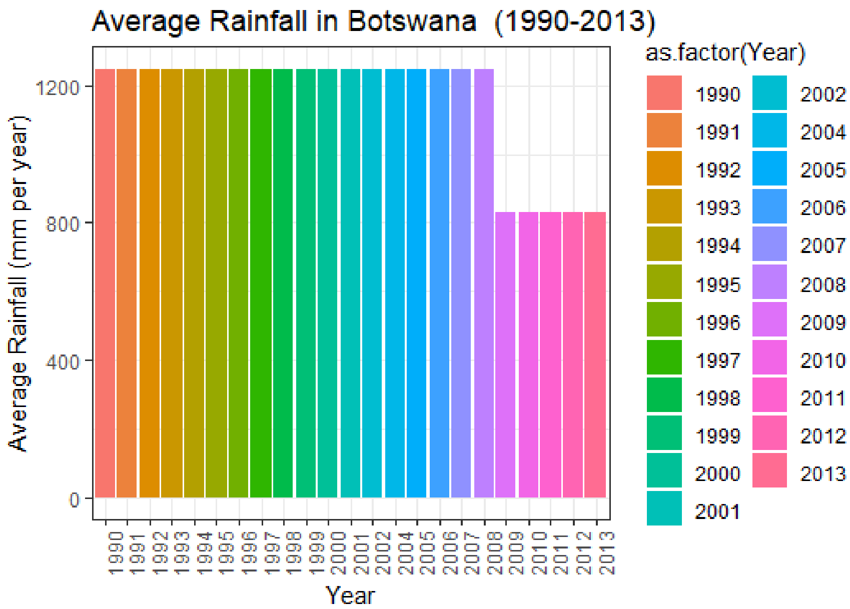




In **Albania,** we see a high amount of rainfall from 1990-1992 but a substantial fall in 1992. From 1994-2013, the average amount of rainfall stayed the same.

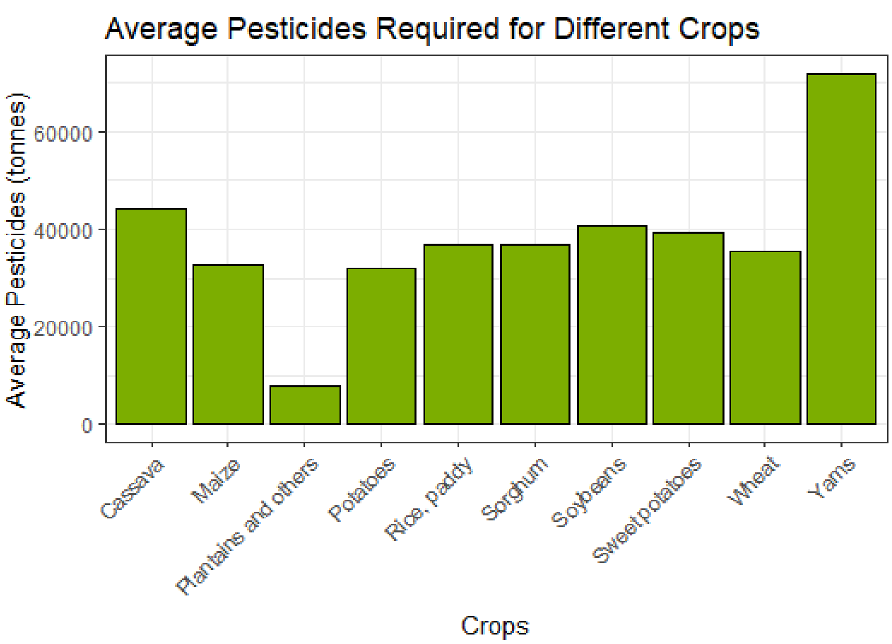


For all the years from 1990 -2013, the average amount of rainfall in **United Kingdom** stood at the same level.



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

Effect of Pesticides

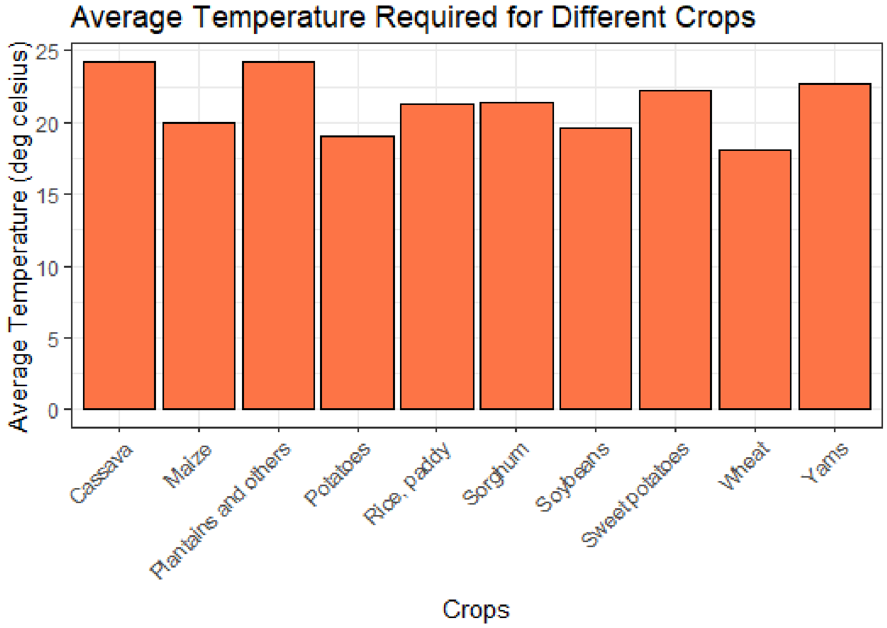


Different varieties of crops require different amount of pesticides for their growth which can be estimated using the following barchart:

Yams are the only crop which requires a huge amount of pesticides. This may be the reason why it is one of the least preferred crops all around the world as it is expensive to grow. Plantains on the other hand require the least amount of pesticides to grow.

Effect of Temperature

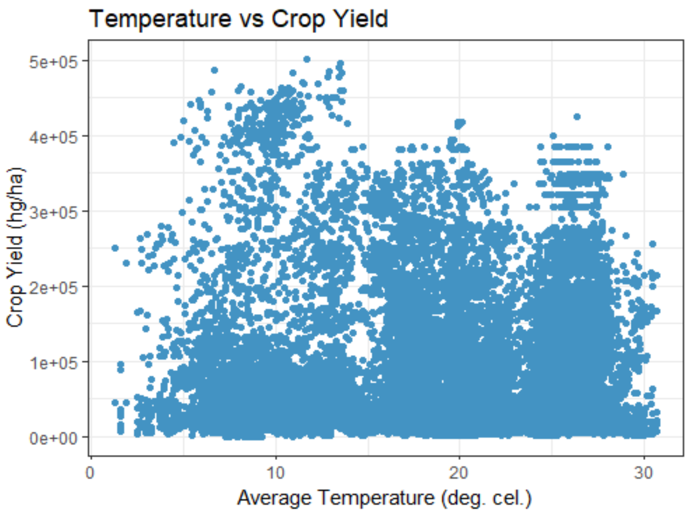
Different varieties of crops grow in different climatic conditions. The effect of temperature on crops can be estimated using the following barchart:

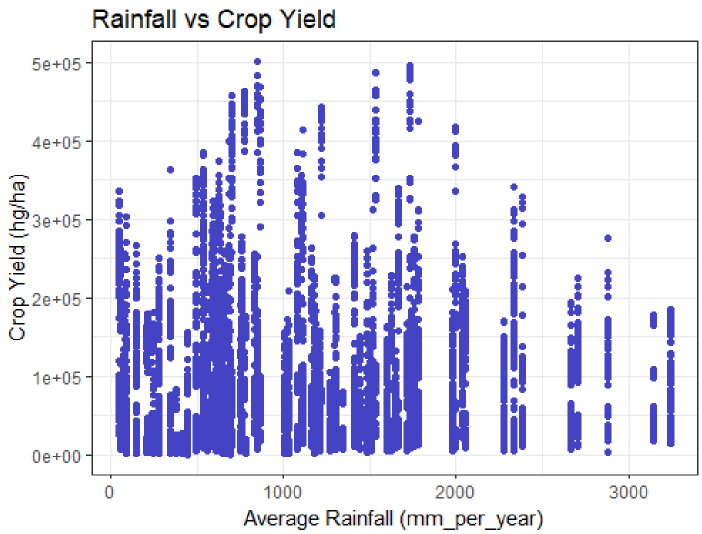


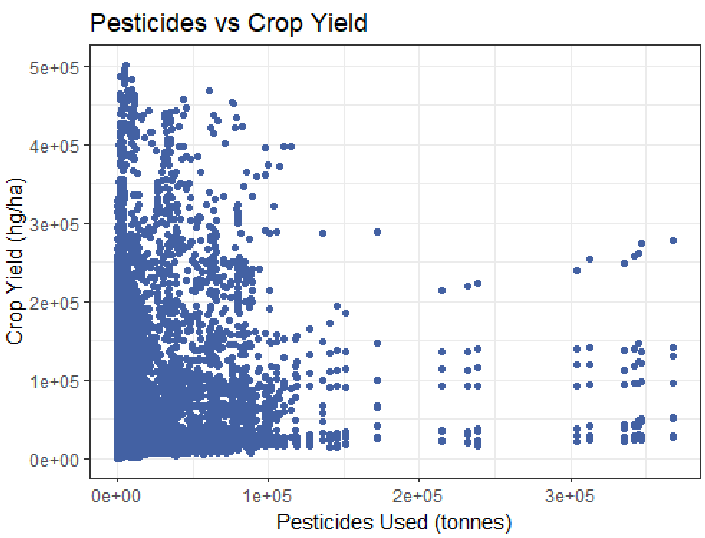
From this chart, we see that Cassava, Plantains, Sweet Potatoes and Yams grow in torrid areas.

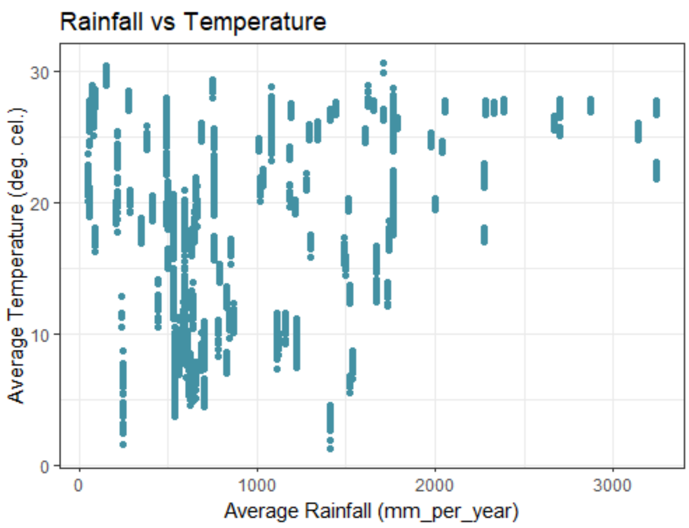
Correlation Analysis

Scatter plots to visualize the relationship of yield with temperature, rainfall and pesticides:







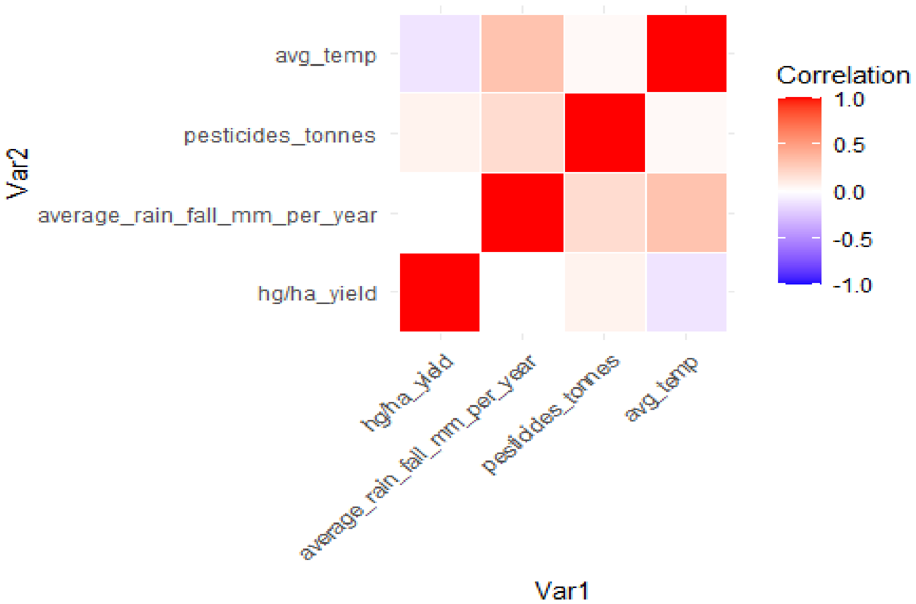


Correlation matrix for above relationships is given by:

```
correlation_matrix

##                                hg/ha_yield average_rain_fall_mm_per_year
## hg/ha_yield                   1.0000000000                  0.0009621545
## average_rain_fall_mm_per_year 0.0009621545                  1.0000000000
## pesticides_tonnes             0.0640850877                  0.1809836464
## avg_temp                     -0.1147769596                  0.3130395215
##                               pesticides_tonnes    avg_temp
## hg/ha_yield                          0.06408509 -0.11477696
## average_rain_fall_mm_per_year        0.18098365  0.31303952
## pesticides_tonnes                    1.00000000  0.03094611
## avg_temp                             0.03094611  1.00000000
```

The correlation matrix can be visualized into a heat map:

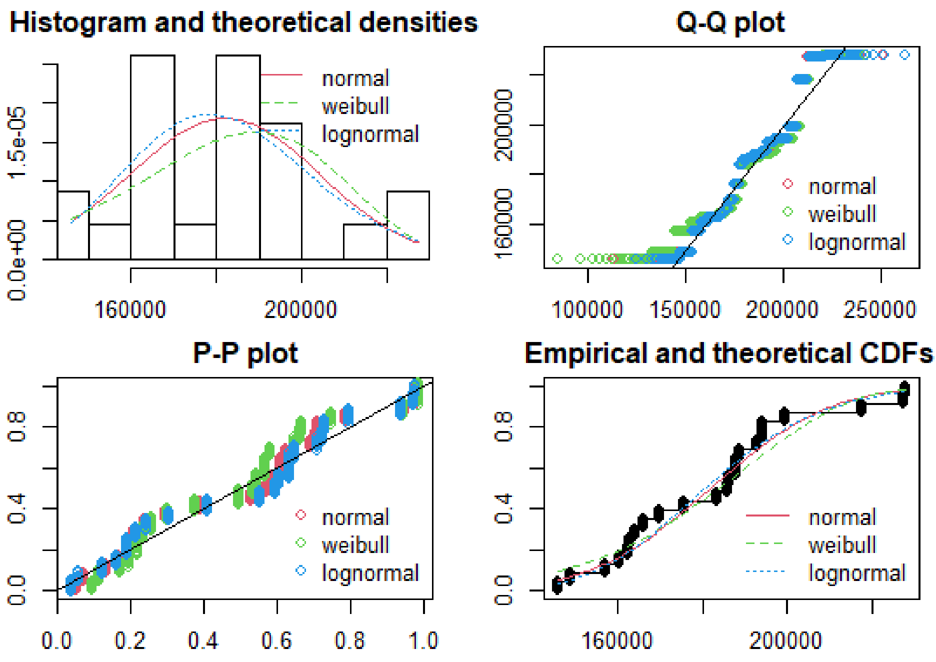
We see that yield is very less correlated with rainfall but more correlated with the amount of pesticides used. Also there is a negative correlation between yield and temperature. A strong positive correlation can also be seen between rainfall and temperature.



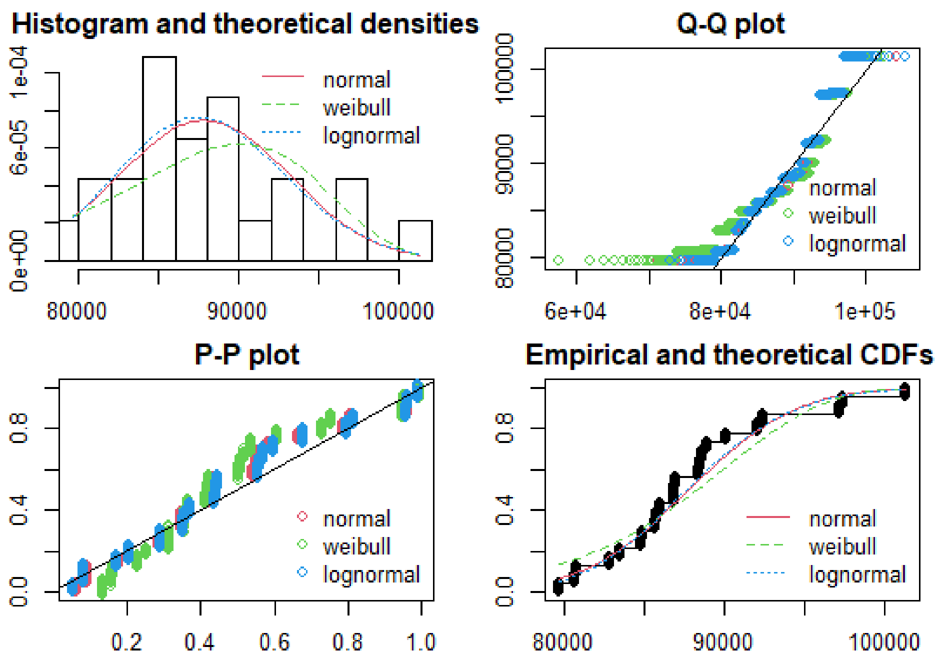
Fitting Distributions

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

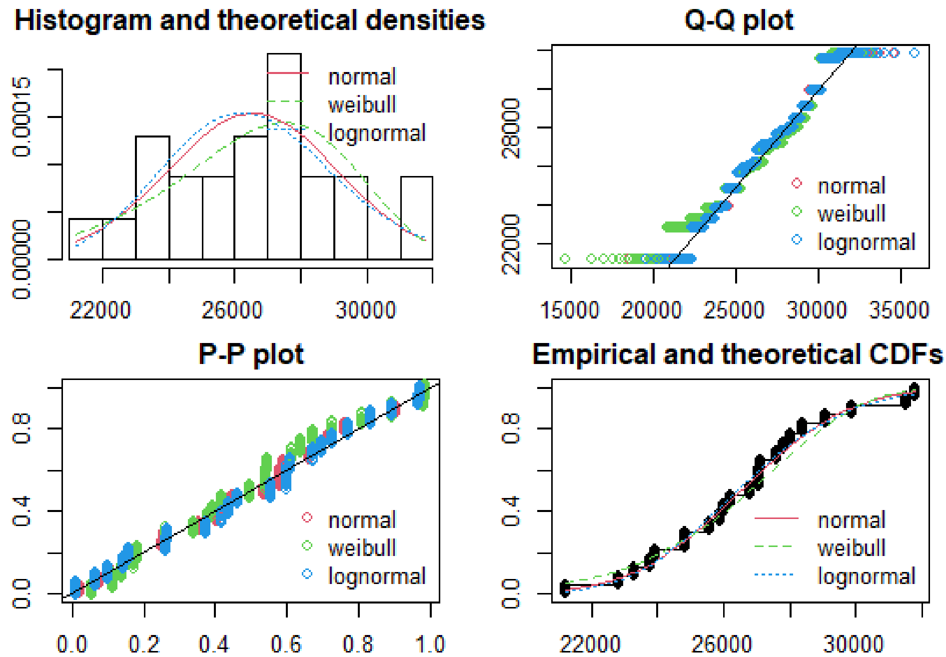
Distribution fitting for the net yield of Potatoes in India:



Distribution fitting for the net yield of Sweet potatoes in India:



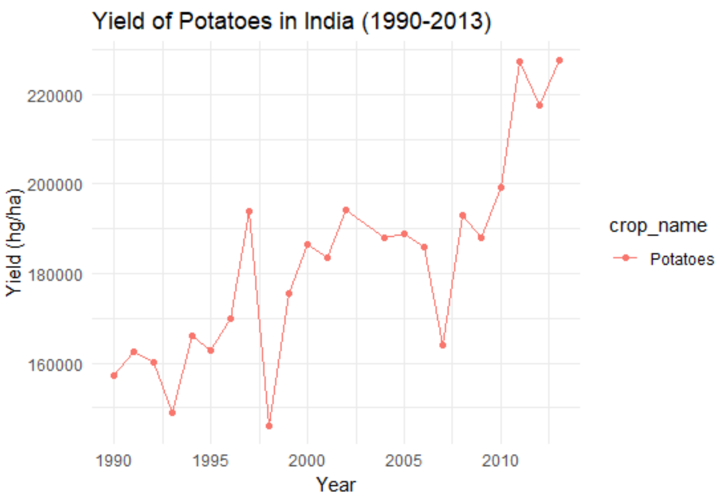
Distribution fitting for the net yield of Wheat in India:

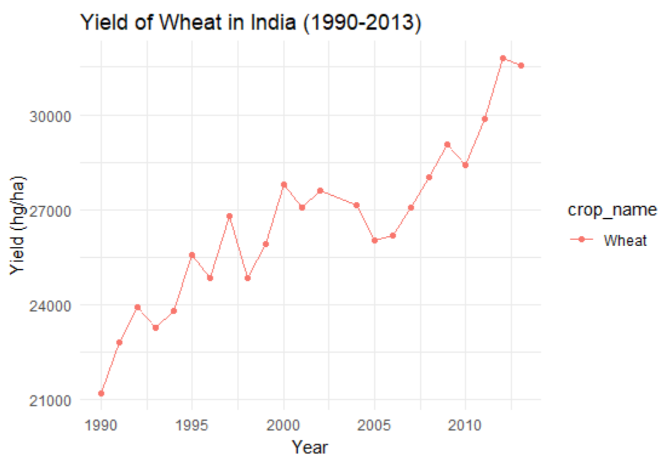


Conclusion

Analyzing the graph of crop yield over time, in relation to environmental factors like rainfall and temperature:

(can be done for any country and crop)





Here are some insights that can be gained:

1. Identification of Optimal Conditions: Examining periods of maximum yield and their corresponding environmental conditions helps identify optimal growing conditions for the specific crop. This information can guide decisions on the timing of planting and harvesting.

2. Identification of Critical Periods: Identifying critical periods during the crop growth cycle, where environmental conditions strongly impact yield, allows for targeted interventions. For example, if a particular stage of growth is sensitive to temperature variations, farmers can implement strategies to mitigate potential adverse effects during that period.

3. Risk Management: Recognizing years with lower yields and investigating the associated environmental conditions helps in assessing and managing risks. Farmers can implement risk mitigation strategies, such as diversifying crops or investing in resilient varieties, based on historical data.

4. Precision Agriculture: Insights from the analysis can inform the adoption of precision agriculture techniques. By leveraging data on historical yield variations and associated factors, farmers can implement precision irrigation, fertilization, and pest control strategies tailored to specific areas of their fields.

5. Crop Rotation and Diversification: Understanding the impact of environmental conditions on specific crops allows for strategic crop rotation and diversification. Farmers can choose crop combinations that complement each other and are resilient to different environmental challenges.

The fitting of distributions give us statistics such as:

AIC (**Akaike information criterion** is an estimator of prediction error and thereby relative quality of statistical models for a given set of data.),

BIC (**Bayesian information criterion** or **Schwarz information criterion** (also SIC, SBC, SBIC) is a criterion for model selection among a finite set of models; models with lower BIC are generally preferred.) and,

KS (**Kolmogorov–Smirnov test (K–S test or KS test)** is a nonparametric test of the equality of continuous (or discontinuous), one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-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.**

Summary of best fit for the net yield of Potatoes in India:

```
## [1] "For Normal Distribution:"
## Goodness-of-fit statistics
##                              1-mle-norm
## Kolmogorov-Smirnov statistic   0.117950
## Cramer-von Mises statistic     1.439504
## Anderson-Darling statistic     9.953863
##
## Goodness-of-fit criteria
##                                1-mle-norm
## Akaike's Information Criterion   11563.59
## 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-lnorm
## Kolmogorov-Smirnov statistic   0.1176472
## Cramer-von Mises statistic     1.3723255
## Anderson-Darling statistic     8.4590345
##
## Goodness-of-fit criteria
##                                1-mle-lnorm
## Akaike's Information Criterion    11544.19
## Bayesian Information Criterion    11552.64
```

From this 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.

Fitting a probability distribution for the yield of a particular crop in a country over a specific period of time is helpful for several reasons:

1. Risk Assessment: Probability distributions help assess the risk associated with different yield levels. By understanding the distribution of potential outcomes, farmers and policymakers can better assess and manage the risks associated with crop production, taking proactive measures to mitigate potential losses.

2. Decision Support: Fitted probability distributions serve as a basis for decision support. Farmers can use the distribution to make informed decisions on crop management strategies, resource allocation, and risk mitigation measures. For example, decisions related to crop insurance, planting schedules, and resource investments can be guided by the distribution of expected yields.

3. Supply Chain Management: Businesses and stakeholders in the agricultural supply chain can use probability distributions to anticipate and plan for variations in crop yield. This is crucial for supply chain optimization, pricing strategies, and overall market planning.

4. Early Warning Systems: Probability distributions can be incorporated into early warning systems for extreme events such as droughts or floods. By monitoring deviations from expected yield distributions, authorities can implement timely responses and support systems for affected regions.

In summary, fitting a probability distribution to crop yield data provides a robust framework for decision-making, risk assessment, and resource optimization in agriculture. It enhances the ability to anticipate and adapt to changing conditions, ultimately contributing to more resilient and sustainable agricultural practices.

References

- ☐ 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/#:~:text=It%20is%20usually%20expressed%20in,over%20a%20specified%20time%20period.
- ☐ https://www.intechopen.com/chapters/70658

Appendix (R Program)

AGRICULTURE DATASET

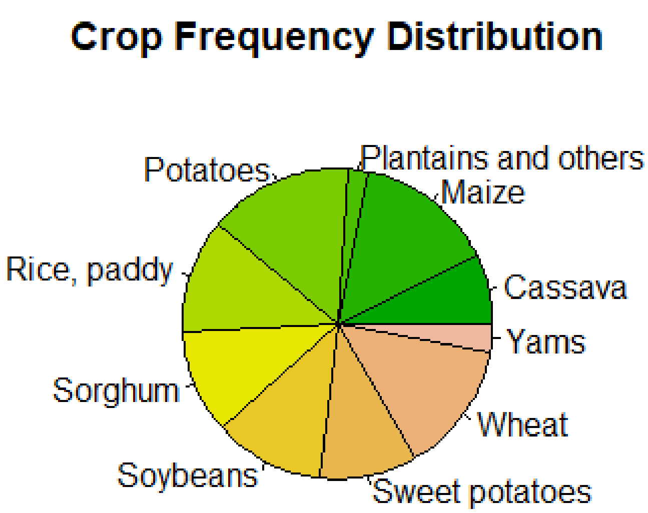
LAKSHYA SINGH, ANIKET KUMAR, UPADHYE RUSHIKESH SUNIL

2023-11-04

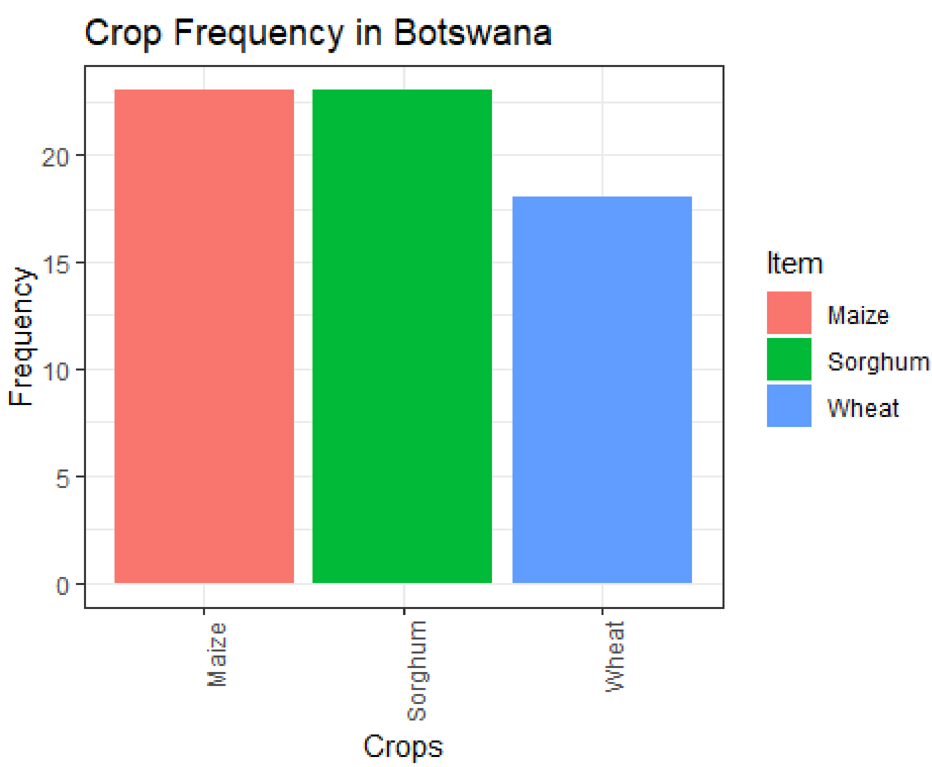
```
#Given dataset:

library(readr)
yield_df <- read_csv("C:/Users/Lakshya Singh/Downloads/yield_df.csv")

## New names:
## Rows: 28242 Columns: 8
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: ","
chr
## (2): Area, Item dbl (6): ...1, Year, hg/ha_yield,
## average_rain_fall_mm_per_year, pesticides_...
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this
message.
## • `` -> `...1`

data = yield_df
#attribute names & types
names(data)

## [1] "...1"                          "Area"
## [3] "Item"                          "Year"
## [5] "hg/ha_yield"                   "average_rain_fall_mm_per_year"
## [7] "pesticides_tonnes"             "avg_temp"

sapply(data, class)

##                          ...1                          Area
##                     "numeric"                   "character"
##                          Item                          Year
##                   "character"                     "numeric"
##                   hg/ha_yield average_rain_fall_mm_per_year
##                     "numeric"                     "numeric"
##             pesticides_tonnes                      avg_temp
##                     "numeric"                     "numeric"

#checking for missing values
any(is.na(data))
```

```
## [1] FALSE

# Displaying the structure of your data
str(data)

## spc_tbl_ [28,242 × 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1                         : num [1:28242] 0 1 2 3 4 5 6 7 8 9 ...
##  $ 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] 121 121 121 121 121 121
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>

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

## # A tibble: 6 × 8
##    ...1 Area  Item   Year `hg/ha_yield` average_rain_fall_mm…¹
pesticides_tonnes
##   <dbl> <chr> <chr> <dbl>         <dbl>                  <dbl>
<dbl>
## 1     0 Alba… Maize  1990         36613                   1485
121
## 2     1 Alba… Pota…  1990         66667                   1485
121
## 3     2 Alba… Rice…  1990         23333                   1485
121
## 4     3 Alba… Sorg…  1990         12500                   1485
121
## 5     4 Alba… Soyb…  1990          7000                   1485
```

```
121
## 6      5 Alba… Wheat  1990         30197                   1485
121
## # ℹ abbreviated name: ¹average_rain_fall_mm_per_year
## # ℹ 1 more variable: avg_temp <dbl>
library(pastecs)
#stat.desc(data)

library(ggplot2)


# Analysing Item
table(data$Item)
##
##                Cassava                 Maize Plantains and others
##                   2045                  4121                  556
##               Potatoes           Rice, paddy              Sorghum
##                   4276                  3388                 3039
##               Soybeans        Sweet potatoes                Wheat
##                   3223                  2890                 3857
##                   Yams
##                    847
pie(table(data$Item), main = "Crop Frequency Distribution", col =
terrain.colors(12))
```



```r
chosen_area = 'Botswana'

# 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))
```




```r
print('The crop which is grown in most of the countries is:')

## [1] "The crop which is grown in most of the countries is:"

which.max(table(data$Item))

## Potatoes
##        4

library(dplyr)

##
## Attaching package: 'dplyr'
##
```

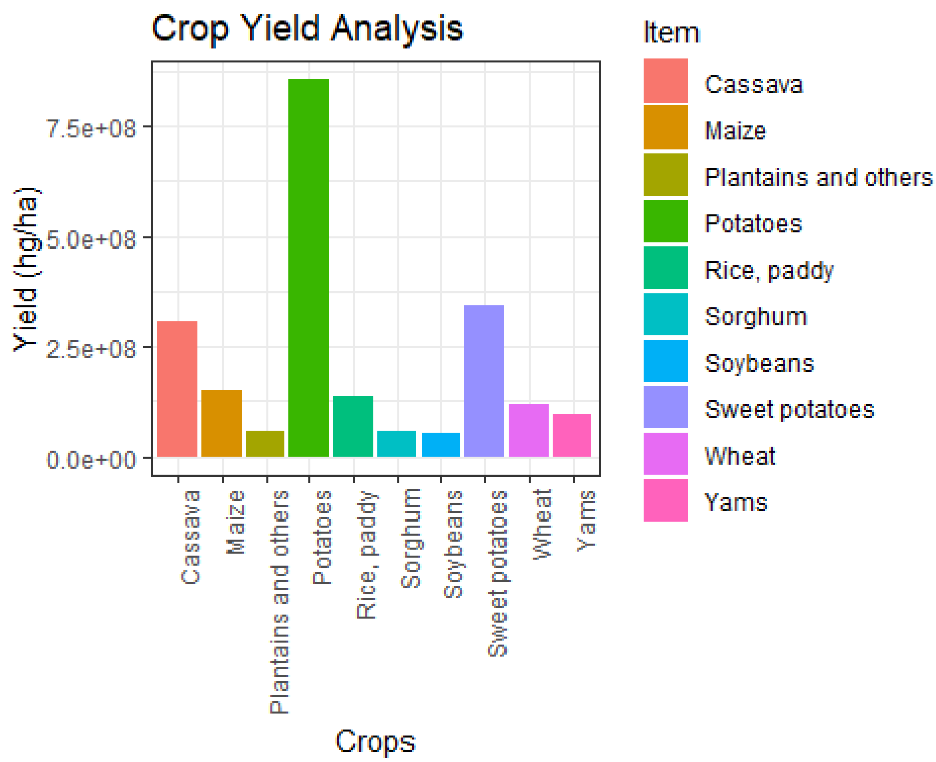
```
## The following objects are masked from 'package:pastecs':
##
##     first, last
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
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],  # Using modeest package for mode
    #Mode_Yield =
as.numeric(names(table(`hg/ha_yield`)[which.max(table(`hg/ha_yield`))]))
    Max_yield = max(`hg/ha_yield`),
    Min_yield = min(`hg/ha_yield`)
  )
print(crop_stats)
```

```
## # A tibble: 10 × 6
##    Item                 Mean_Yield Median_Yield Mode_Yield Max_yield
Min_yield
##    <chr>                     <dbl>        <dbl>      <dbl>     <dbl>
<dbl>
##  1 Cassava                 150479.       128200     100000    385818
11778
##  2 Maize                    36310.        25401      25000    207556
849
##  3 Plantains and others    106041.        89860.    100000    418505
21350
##  4 Potatoes                199802.       182271     169913    501412
8406
##  5 Rice, paddy              40730.        35878      32924    103895
2034
##  6 Sorghum                  18636.        12885       7968    206000
578
##  7 Soybeans                 16731.        15533      10000     41609
50
##  8 Sweet potatoes          119058.        99940      79663    400000
8799
```

```
##  9 Wheat                         30116.        25497        21211      99387
1706
## 10 Yams                         114140.        92593        92000     250000
11475
```

```r
# 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))
```

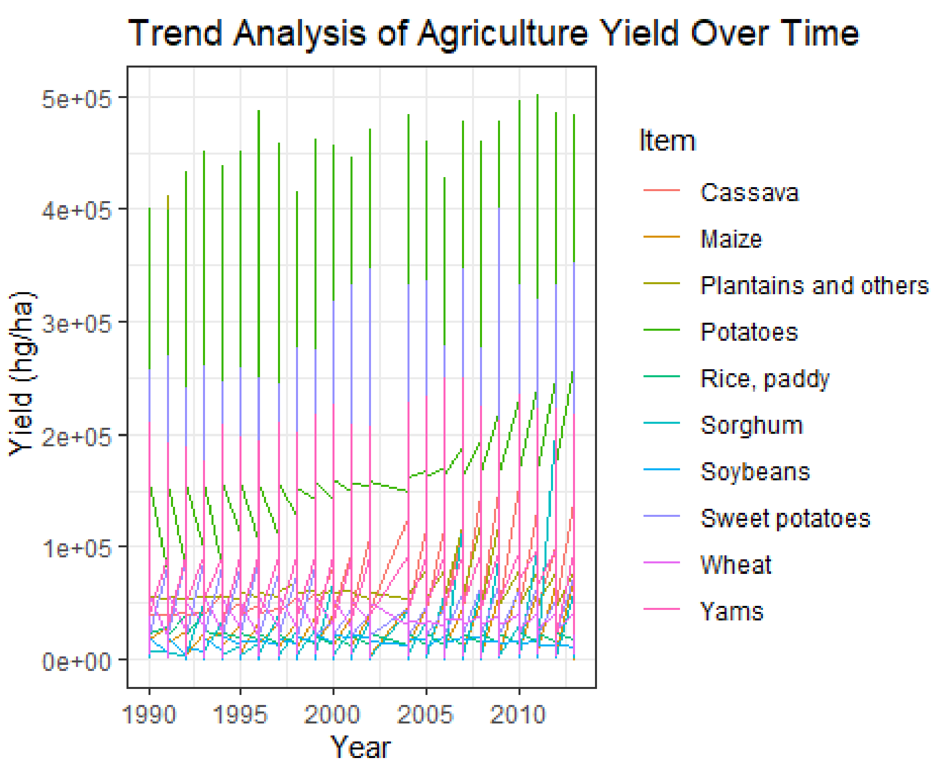



```r
#install.packages("tidyverse")
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.3     ✔ tibble    3.2.1
## ✔ purrr     1.0.1     ✔ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ tidyr::extract() masks pastecs::extract()
## ✖ dplyr::filter()  masks stats::filter()
```

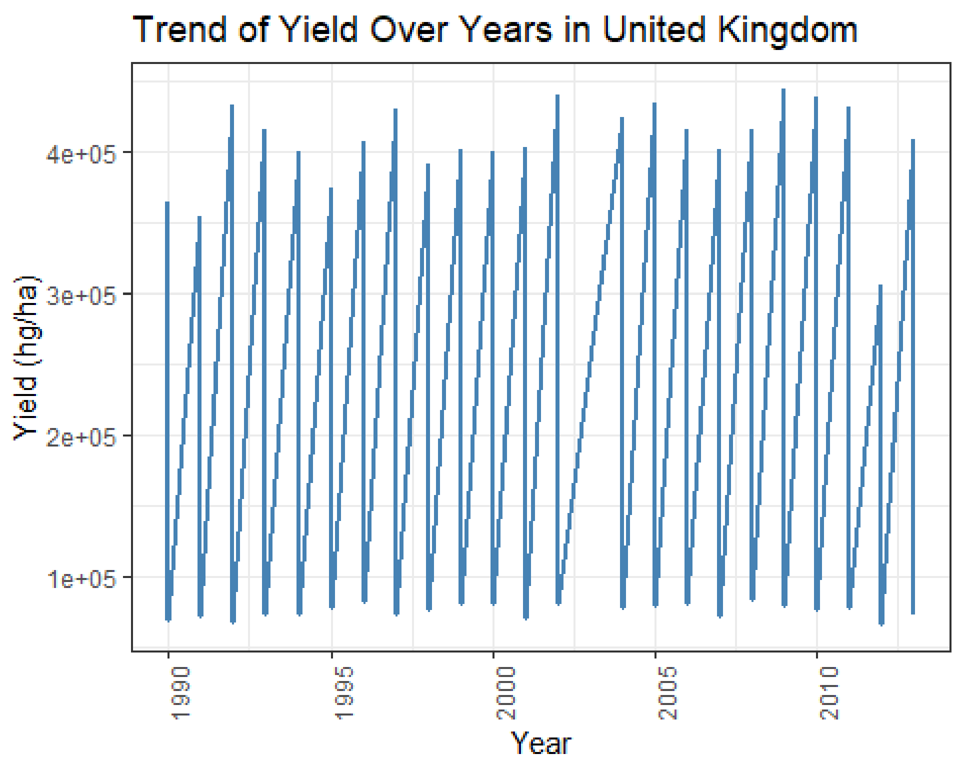
```
## ✖ dplyr::first()   masks pastecs::first()
## ✖ dplyr::lag()     masks stats::lag()
## ✖ dplyr::last()    masks pastecs::last()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
# 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()
```

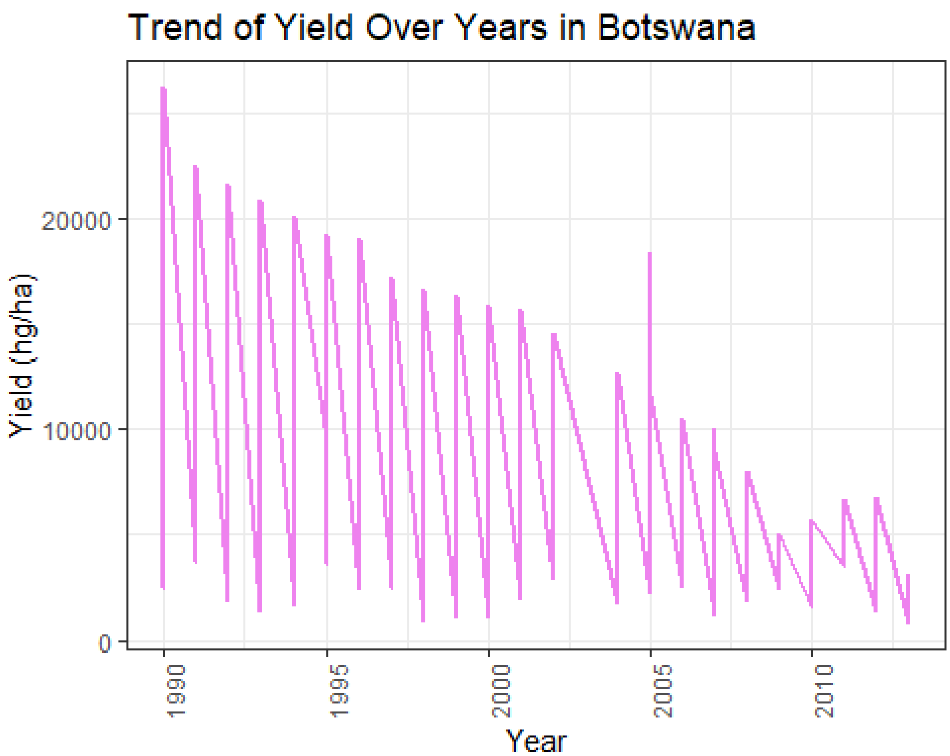



```
# Calculating mean yield for each country
country_yield = data %>%
  group_by(Area) %>%
  summarize(mean_yield = mean(`hg/ha_yield`))




#trends over the years for United Kingdom & Botswana
trend_plot1 = data %>%
  filter(Area == "United Kingdom") %>%
  ggplot(aes(x = Year, y = `hg/ha_yield`)) +
```

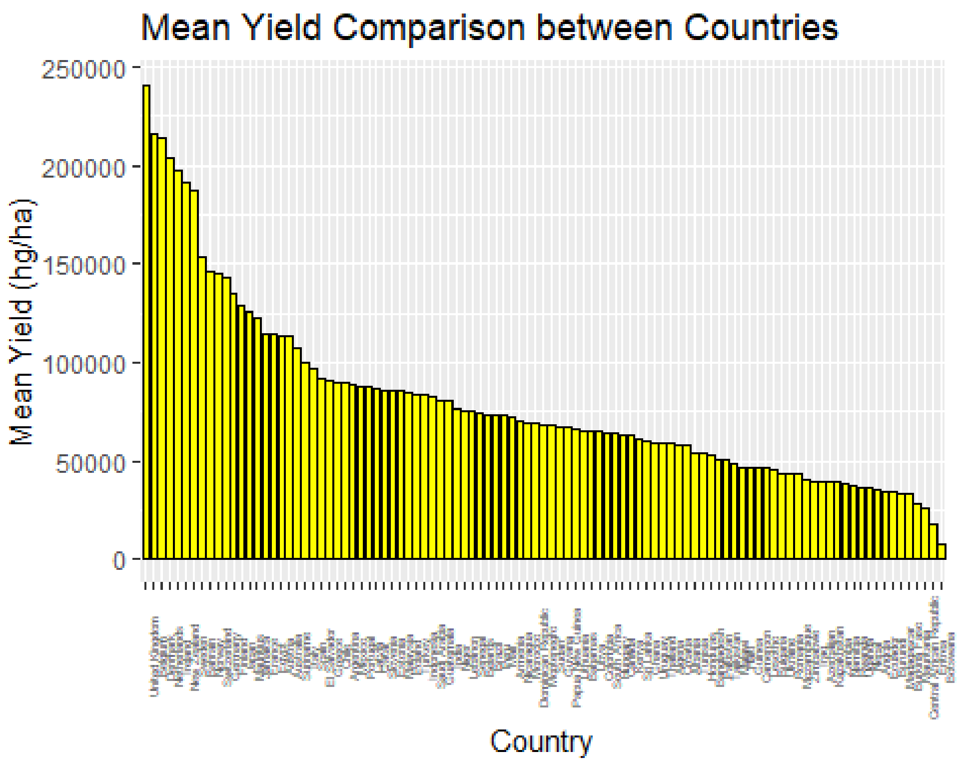
```
  geom_line(color = "steelblue", lwd = 1) +
  labs(title = "Trend of Yield Over Years in United Kingdom", x = "Year", y =
"Yield (hg/ha)") +
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

print(trend_plot1)
```




```
trend_plot2 = data %>%
  filter(Area == "Botswana") %>%
  ggplot(aes(x = Year, y = `hg/ha_yield`)) +
  geom_line(color = "violet", lwd = 1) +
  labs(title = "Trend of Yield Over Years in Botswana", x = "Year", y =
"Yield (hg/ha)") +
  theme_bw()+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

print(trend_plot2)
```

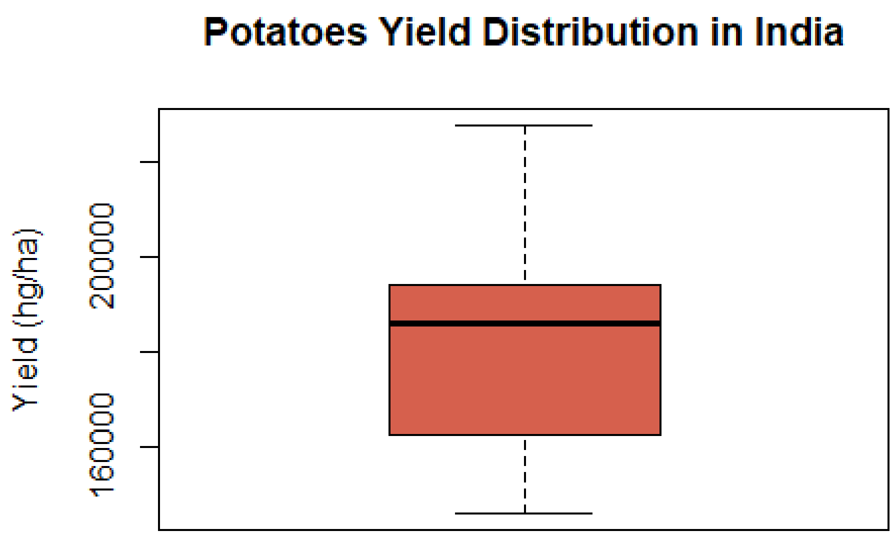



```
# Creating a bar plot to compare mean yield between countries
ggplot(country_yield, aes(x = reorder(Area, -mean_yield), y = mean_yield)) +
  geom_bar(stat = "identity", fill = "yellow",color = "black") +
  labs(title = "Mean Yield Comparison between Countries", x = "Country", y =
"Mean Yield (hg/ha)")+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

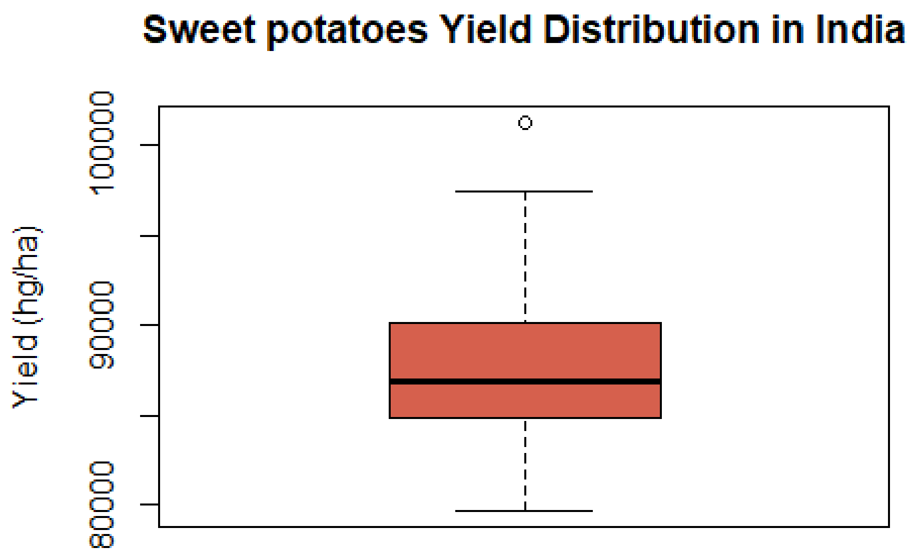



```
# Crop-wise analysis for a specific crop in a particular area from 1990-2013
crop_analysis = function(crop_name, area_name) {
  # Filtering data for the specific crop, area, and time period
  crop_data = subset(data, Item == crop_name & Area == area_name & Year >=
1990 & Year <= 2013)

  # 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)


}
```

```r
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
```



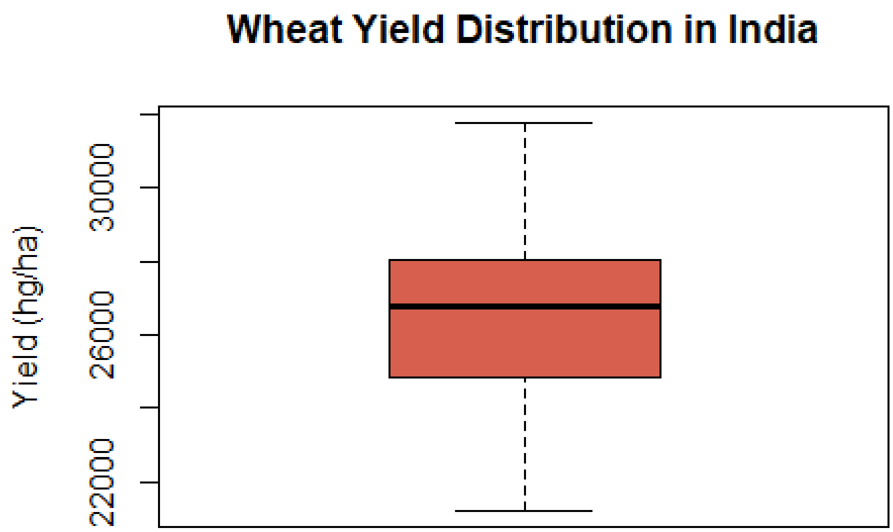

```r
## [1] "Outliers are:"
## numeric(0)

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
```



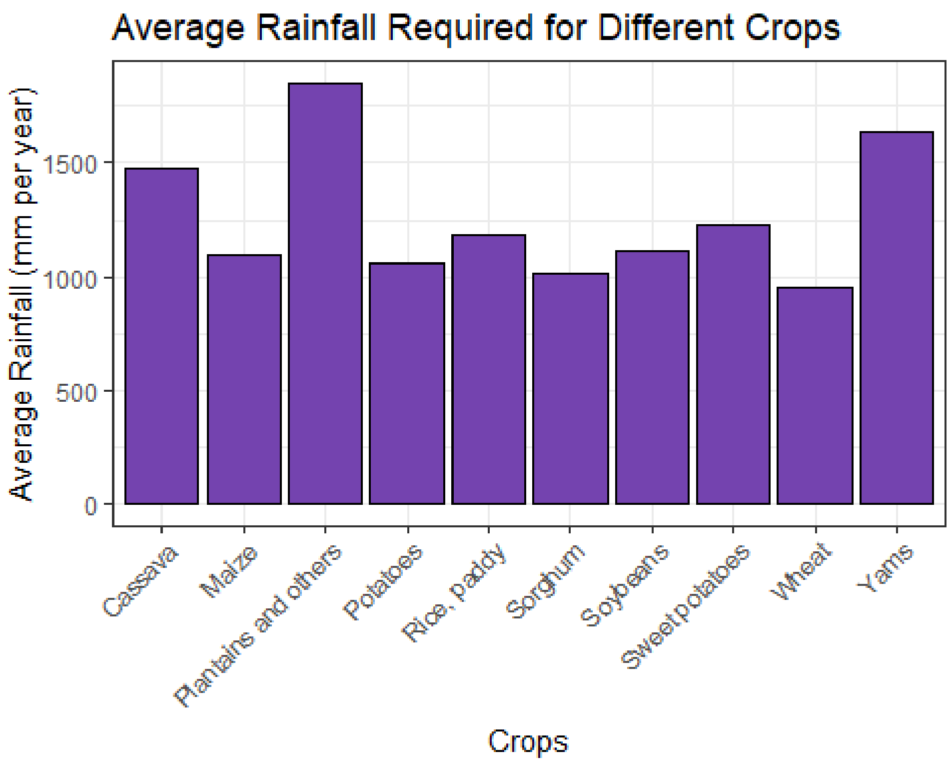

```
## [1] "Outliers are:"
##  [1] 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288
## [11] 101288 101288 101288 101288 101288 101288 101288 101288 101288 101288
## [21] 101288 101288
```

```r
crop_analysis("Wheat","India")
```

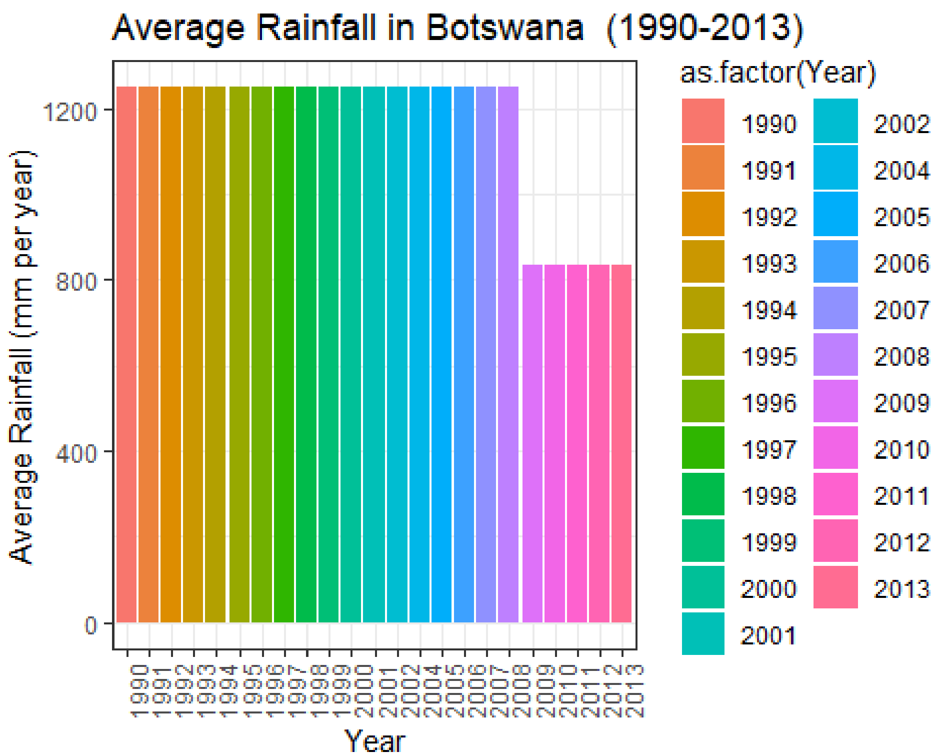
```
## [1] "Summary Statistics for Wheat in India from 1990-2013:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21211   24828   26789   26547   28022   31775
```



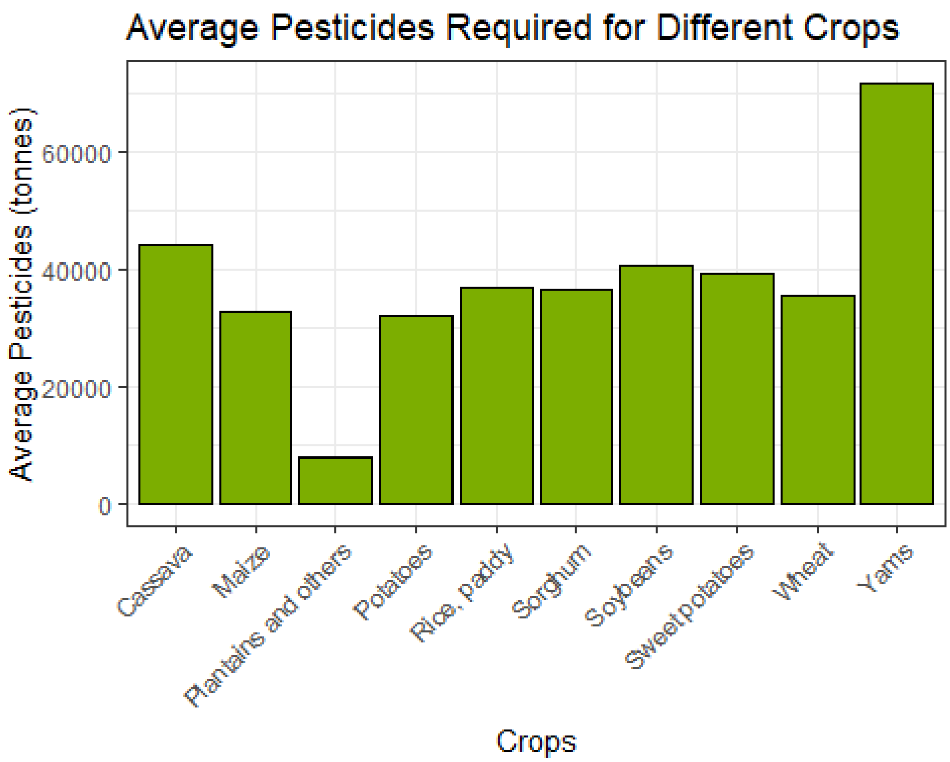

```
## [1] "Outliers are:"
## numeric(0)

#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))
```



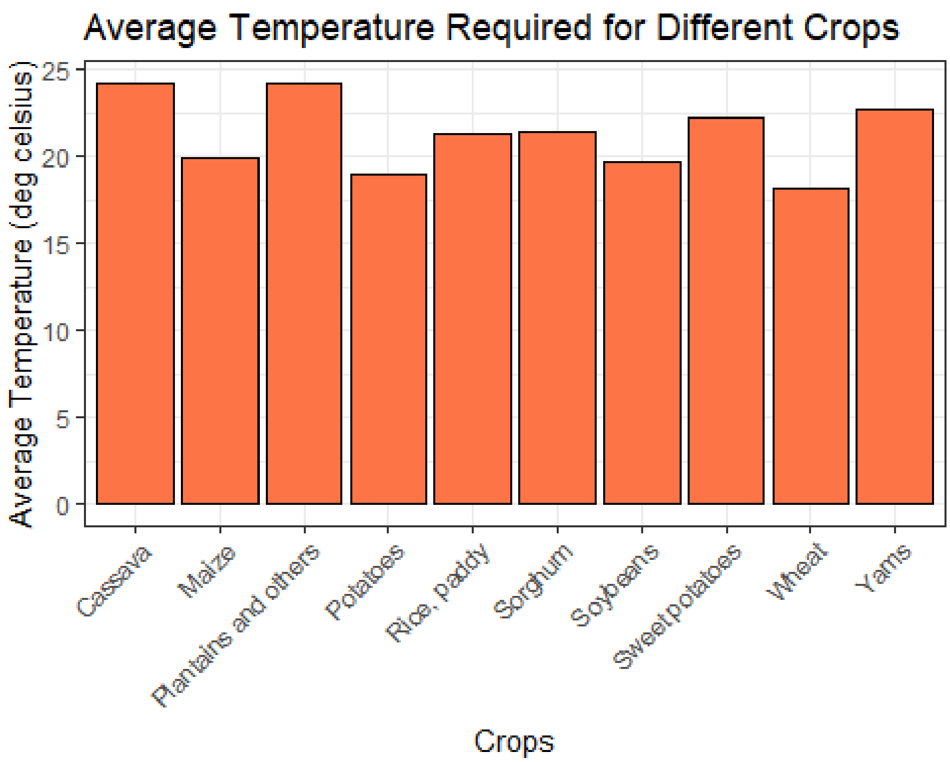

```r
# rainfall amount for a particular area
selected_country <- 'Botswana'

# 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))
```



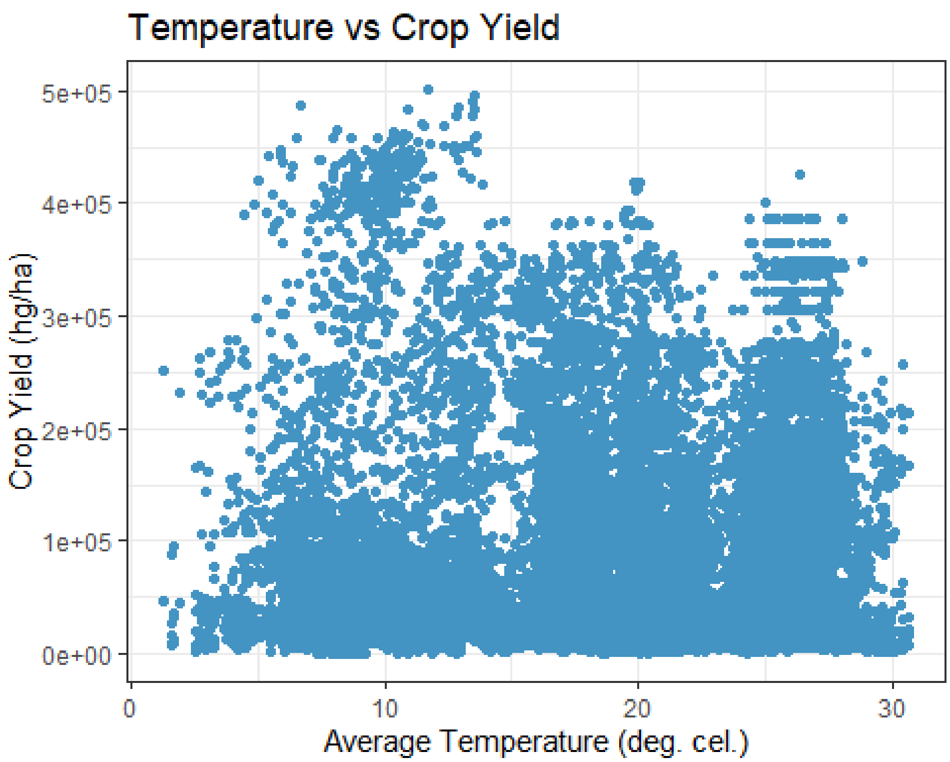

```
# effect of pesticide

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))
```



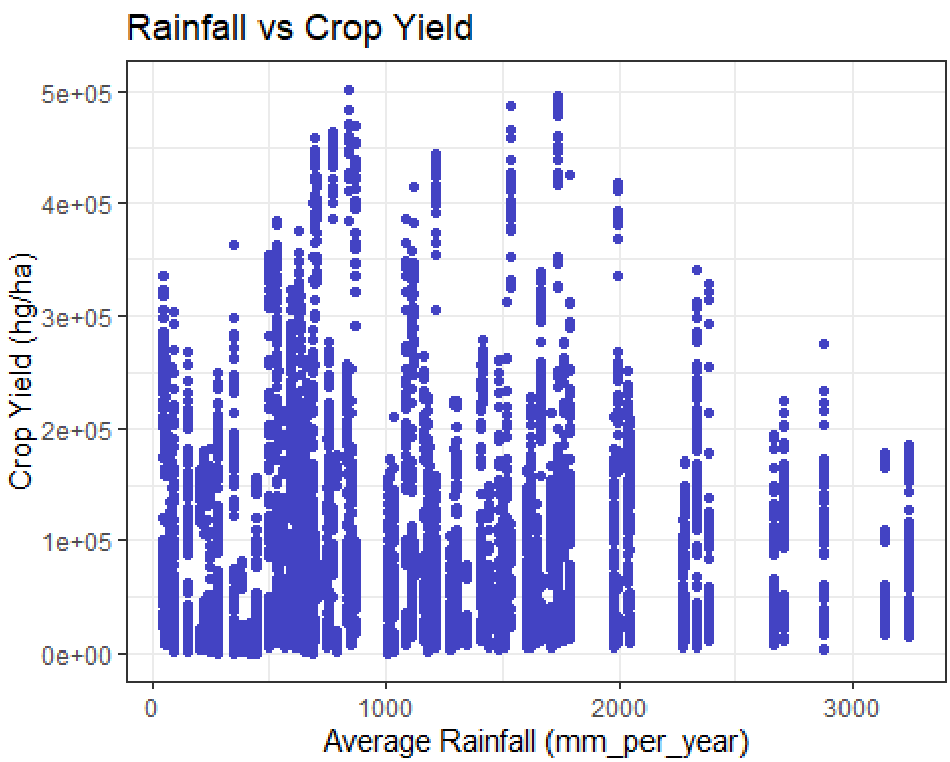

```r
# effect of temperature

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))
```



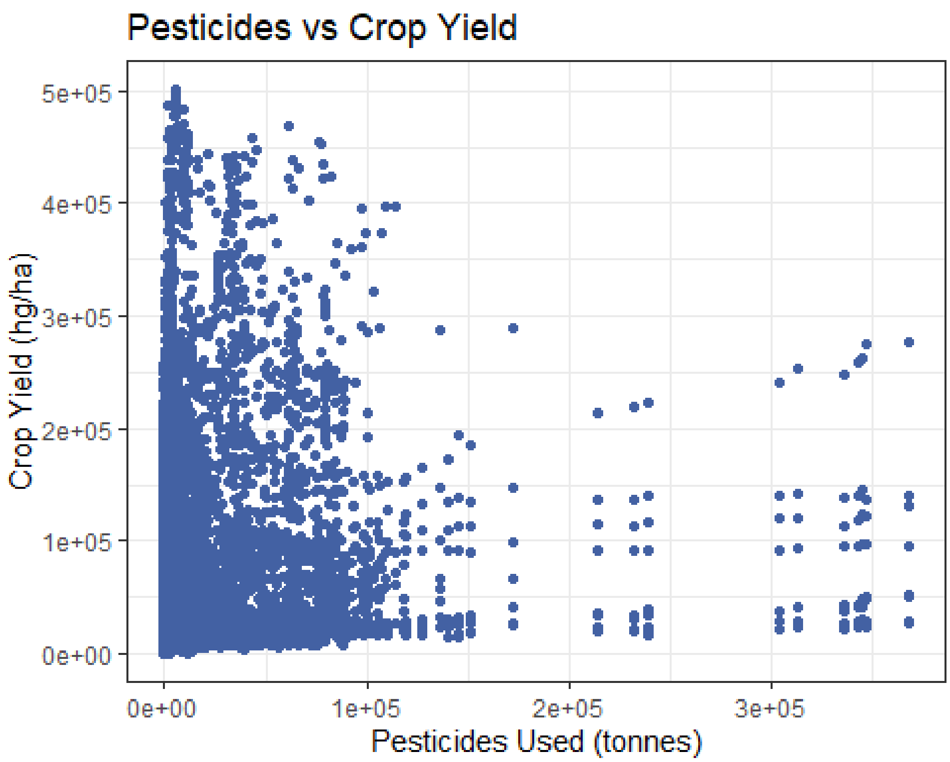

```
#correlation analysis

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



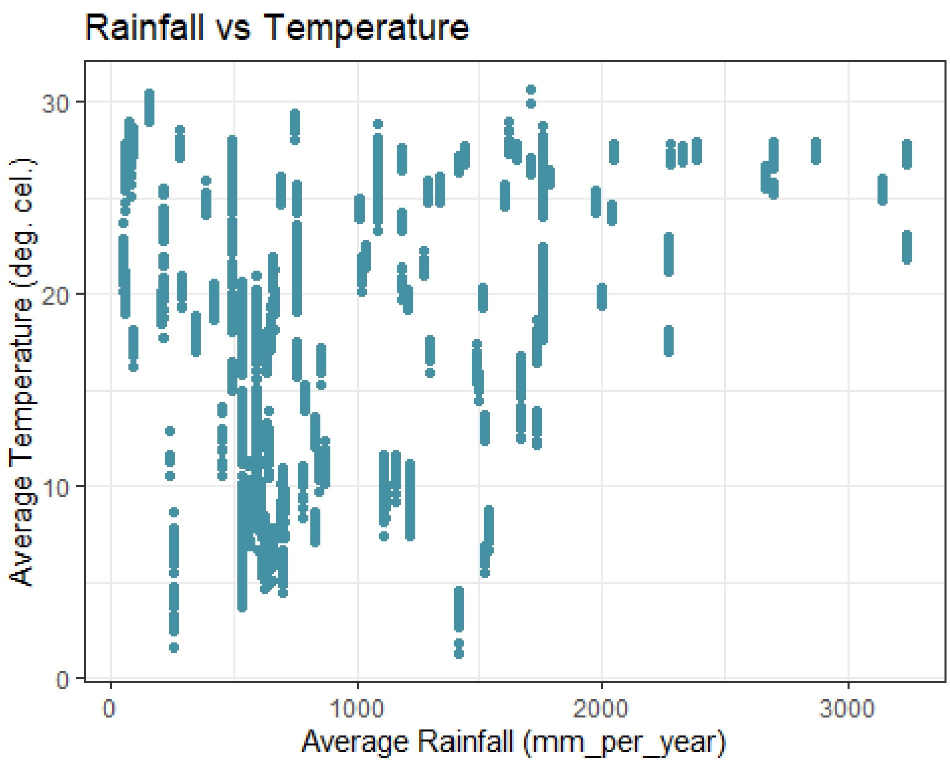

```
# Scatter plot to visualize the relationship between yield & rainfall
ggplot(data, aes(x = average_rain_fall_mm_per_year, y =`hg/ha_yield` )) +
  geom_point(col = "#4343C3") +
  labs(title = "Rainfall vs Crop Yield",
       x = "Average Rainfall (mm_per_year)",
       y = "Crop Yield (hg/ha)")+
  theme_bw()
```




```r
# 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()
```

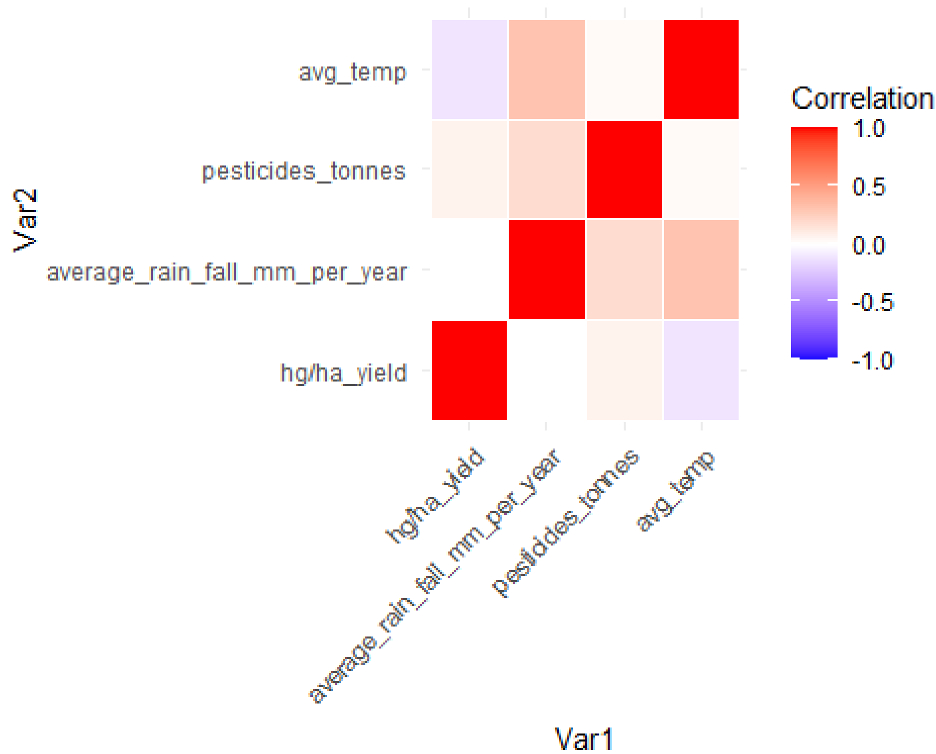



```
# Scatter plot to visualize the relationship between rainfall & temperature
ggplot(data, aes(x = average_rain_fall_mm_per_year, y = avg_temp)) +
  geom_point(col = "#4391A3") +
  labs(title = "Rainfall vs Temperature",
       x = "Average Rainfall (mm_per_year)",
       y = "Average Temperature (deg.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



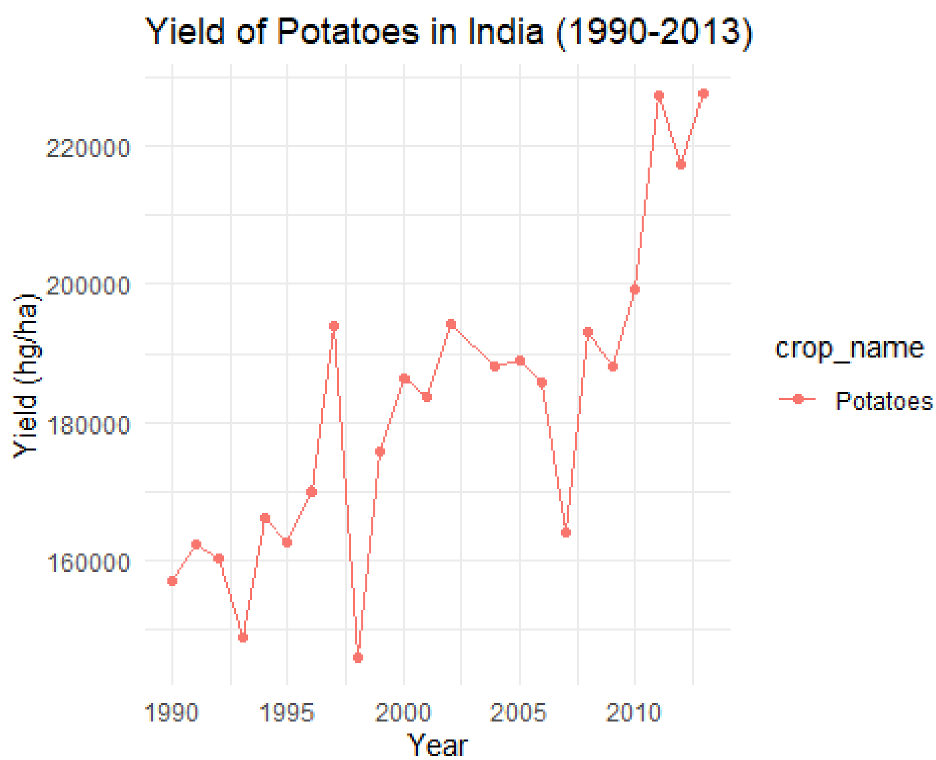

```
correlation_matrix = cor(data[, c(
"hg/ha_yield","average_rain_fall_mm_per_year", "pesticides_tonnes",
"avg_temp")])
correlation_matrix

##                                   hg/ha_yield average_rain_fall_mm_per_year
## hg/ha_yield                      1.0000000000                  0.0009621545
## average_rain_fall_mm_per_year  0.0009621545                  1.0000000000
## pesticides_tonnes              0.0640850877                  0.1809836464
## avg_temp                      -0.1147769596                  0.3130395215
##                               pesticides_tonnes    avg_temp
## hg/ha_yield                          0.06408509 -0.11477696
## average_rain_fall_mm_per_year        0.18098365  0.31303952
## pesticides_tonnes                    1.00000000  0.03094611
## avg_temp                             0.03094611  1.00000000

#install.packages("reshape2")
library(reshape2)

##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
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))
```

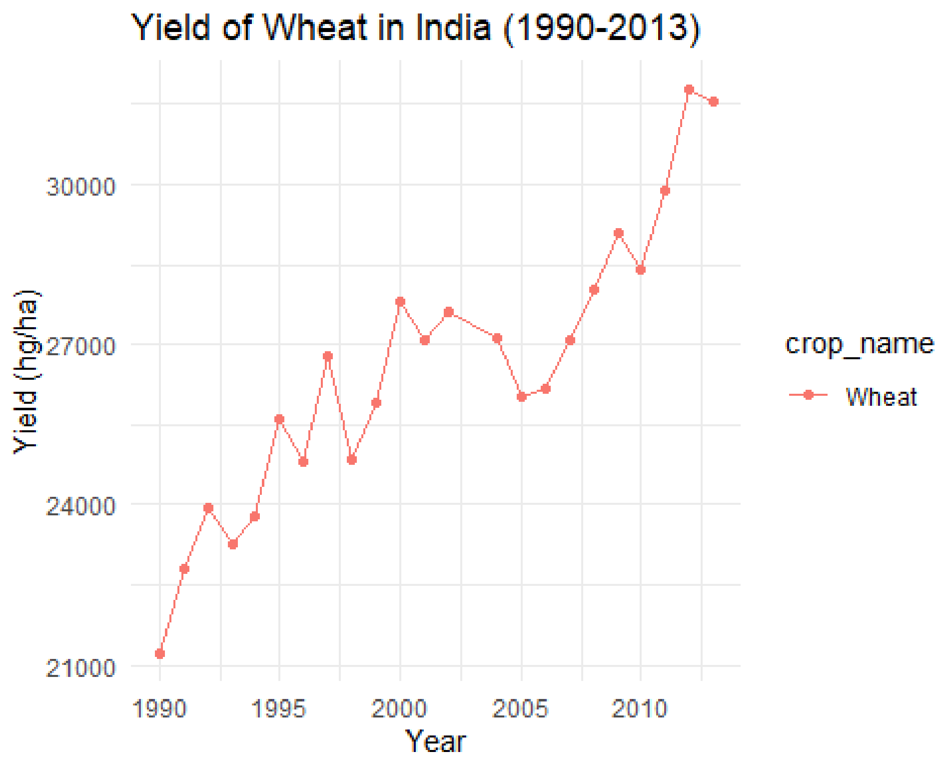



```r
#inferences
plot_yield_and_stats = function(crop_name, country_name) {
  crop_data = subset(data, Item == crop_name & Area == country_name & Year >=
1990 & Year <= 2013)

  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")

## [1] "Average Rainfall in 2013 : 1083 mm per year"
## [1] "Average Temperature in 2013 : 26.7518181818182 °C"
```

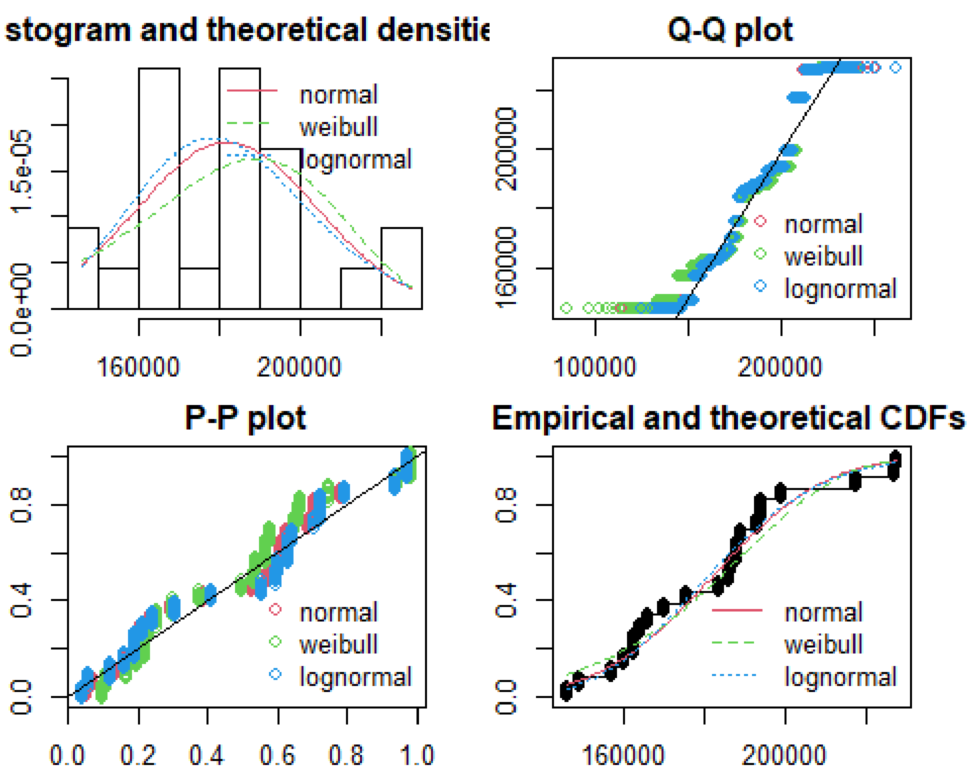



```
plot_yield_and_stats("Wheat", "India")

## [1] "Average Rainfall in 2012 : 1083 mm per year"
## [1] "Average Temperature in 2012 : 26.0172727272727 °C"
```



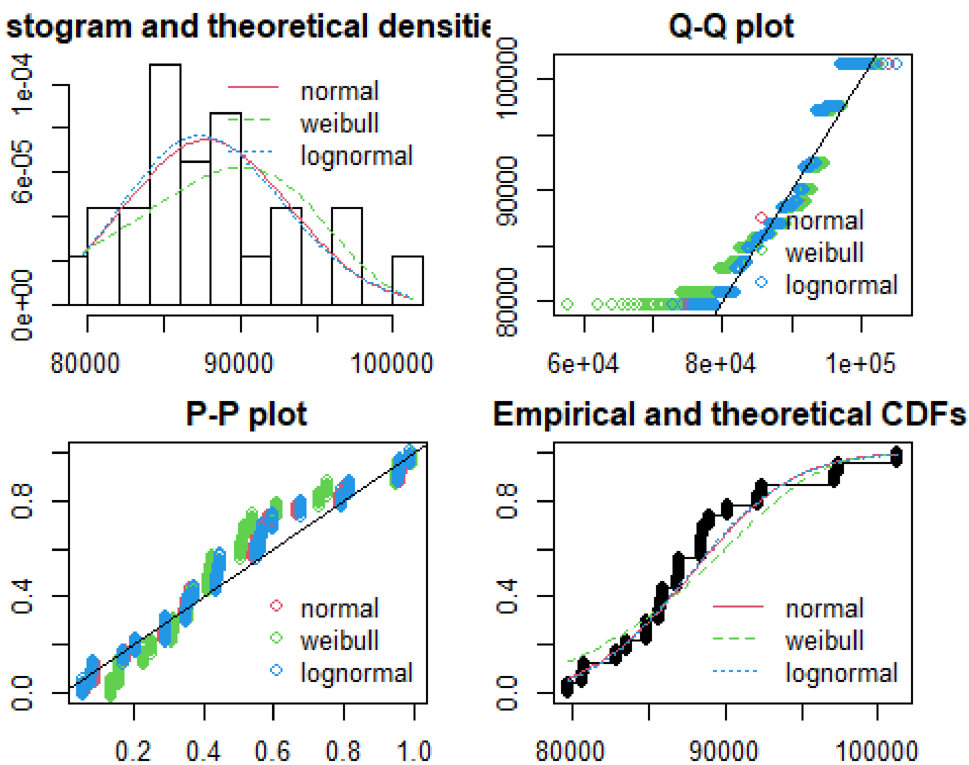

```
# fitting various distribution to the dataset for yield
#install.packages("fitdistrplus", "MASS", "survival")
library(fitdistrplus)

## Warning: package 'fitdistrplus' was built under R version 4.3.2

## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## Loading required package: survival

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)

}

#using above function
yield_fit("Potatoes","India")
```



```
## [1] "For Normal Distribution:"
## Goodness-of-fit statistics
##                              1-mle-norm
## Kolmogorov-Smirnov statistic   0.117950
## Cramer-von Mises statistic     1.439504
## Anderson-Darling statistic     9.953863
##
## Goodness-of-fit criteria
##                                1-mle-norm
## Akaike's Information Criterion   11563.59
## 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-lnorm
## Kolmogorov-Smirnov statistic   0.1176472
## Cramer-von Mises statistic     1.3723255
## Anderson-Darling statistic     8.4590345
##
## Goodness-of-fit criteria
##                                1-mle-lnorm
## Akaike's Information Criterion    11544.19
## Bayesian Information Criterion    11552.64
```

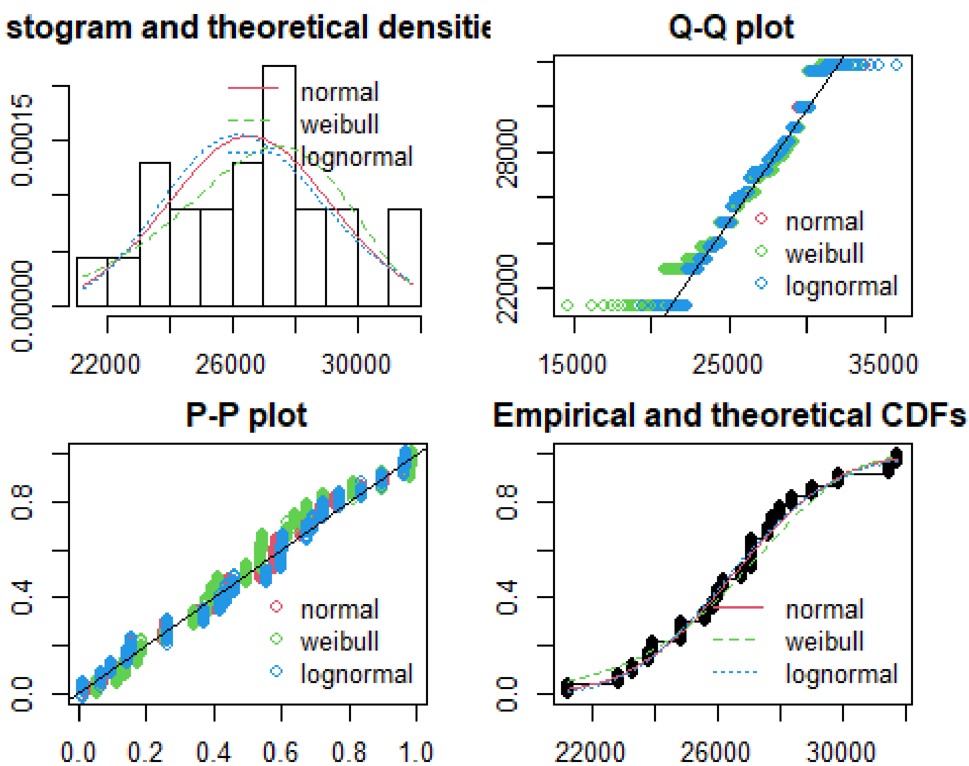
```r
yield_fit("Sweet potatoes",  "India")
```




```
## [1] "For Normal Distribution:"
## Goodness-of-fit statistics
##                               1-mle-norm
## Kolmogorov-Smirnov statistic   0.1560661
## Cramer-von Mises statistic     1.9152932
## Anderson-Darling statistic    12.0336060
##
## Goodness-of-fit criteria
##                                 1-mle-norm
## Akaike's Information Criterion    10123.97
## Bayesian Information Criterion    10132.42
## [1] "For Weibull Distribution:"
## Goodness-of-fit statistics
##                               1-mle-weibull
## Kolmogorov-Smirnov statistic      0.2019326
## Cramer-von Mises statistic        4.1979081
## Anderson-Darling statistic       24.5009109
##
## Goodness-of-fit criteria
##                                 1-mle-weibull
## Akaike's Information Criterion       10260.72
## Bayesian Information Criterion       10269.18
## [1] "For Lognormal Distribution:"
## Goodness-of-fit statistics
##                               1-mle-lnorm
## Kolmogorov-Smirnov statistic    0.1432406
## Cramer-von Mises statistic      1.5330153
```

```
## Anderson-Darling statistic     9.8294492
##
## Goodness-of-fit criteria
##                                1-mle-lnorm
## Akaike's Information Criterion    10102.07
## Bayesian Information Criterion    10110.52
```

```
yield_fit("Wheat",  "India")
```




```
## [1] "For Normal Distribution:"
## Goodness-of-fit statistics
##                                1-mle-norm
## Kolmogorov-Smirnov statistic  0.0669848
## Cramer-von Mises statistic    0.4324785
## Anderson-Darling statistic    3.3858694
##
## Goodness-of-fit criteria
##                                1-mle-norm
## Akaike's Information Criterion   9393.115
## Bayesian Information Criterion   9401.568
## [1] "For Weibull Distribution:"
## Goodness-of-fit statistics
##                                1-mle-weibull
## Kolmogorov-Smirnov statistic     0.1089176
## Cramer-von Mises statistic       1.0565093
## Anderson-Darling statistic       7.9292065
##
```

```
## Goodness-of-fit criteria
##                                1-mle-weibull
## Akaike's Information Criterion      9439.304
## Bayesian Information Criterion      9447.757
## [1] "For Lognormal Distribution:"
## Goodness-of-fit statistics
##                              1-mle-lnorm
## Kolmogorov-Smirnov statistic   0.0779731
## Cramer-von Mises statistic     0.4966465
## Anderson-Darling statistic     3.3768931
##
## Goodness-of-fit criteria
##                                1-mle-lnorm
## Akaike's Information Criterion    9393.506
## Bayesian Information Criterion    9401.959
```