Project - IE6200 - Sec 7 - Group 10

Kush Adhvaryu, Hansika Karkera, Abhishek Shetty

Project - IE6200 - Sec 7 - Group 10

Kush Adhvaryu, Hansika Karkera, Abhishek Shetty

11/27/2019

1. Objective:

Rainfall is one of the primary inputs for various engineering designs. Statistical analysis of the rainfall data is important to facilitate decision making in cropping pattern, construction of roads, urban and rural engineering. Using the data recorded for Rainfall in India from 1901 to 2017 we plot various visualizations to draw suitable conclusions about the rainfall pattern. With the help of these visualizations, we go on to test various hypothesis and predict the annual rainfall using linear regression.

2. Data Description:

The rainfall data is obtained from https://data.gov.in/ which is an open government data platform. It presents details on the amount of rainfall received (in mm) by various Subdivisons in India for the period of 1901 - 2017.

Below is the description of the variables in the dataset:

1) SUBDIVISION: Represents the 36 subdivisions of India.

2) Year: The period for which rainfall is recorded.

3) The next 12 columns (JAN, FEB-NOV, DEC) represent the amount of rainfall received in the 12 months in a year.

4) ANNUAL: The annual rainfall received by a subdivision in a particular year.

5) JF: The amount of rainfall received in January & February.

6) MAM: The amount of rainfall received in March, April & May.

7) JJAS: The amount of rainfall received in June, July & August.

8) OND: The amount of rainfall received in October, November & December.

3. Data Visualization:

In this section we draw visualization plots to establish varoius patterns in the rainfall throughout the country.

```
library(magrittr)
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.
2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()

library(ggthemes)

rain <- read.csv("D:/NEU/Probs & Stats/rainfall-in-india/rainfall-in-india/ra
in.csv", na.strings="", stringsAsFactors=FALSE)

rain_update <- rain %>%
  group_by(YEAR) %>%
  summarise(average = mean(ANNUAL, na.rm = TRUE))

ggplot(rain_update, aes(x = YEAR, y = average)) +
   geom_line() +
   geom_point(size = 3) +
   geom_hline((aes(yintercept = mean(average))), color = "red", linetype = "t
wodash") +
   theme_economist_white() +
   labs(title = "Evolution of Annual Rainfall in India",
       x = "Year", y = "Annual Rainfall (in mm.)")
```

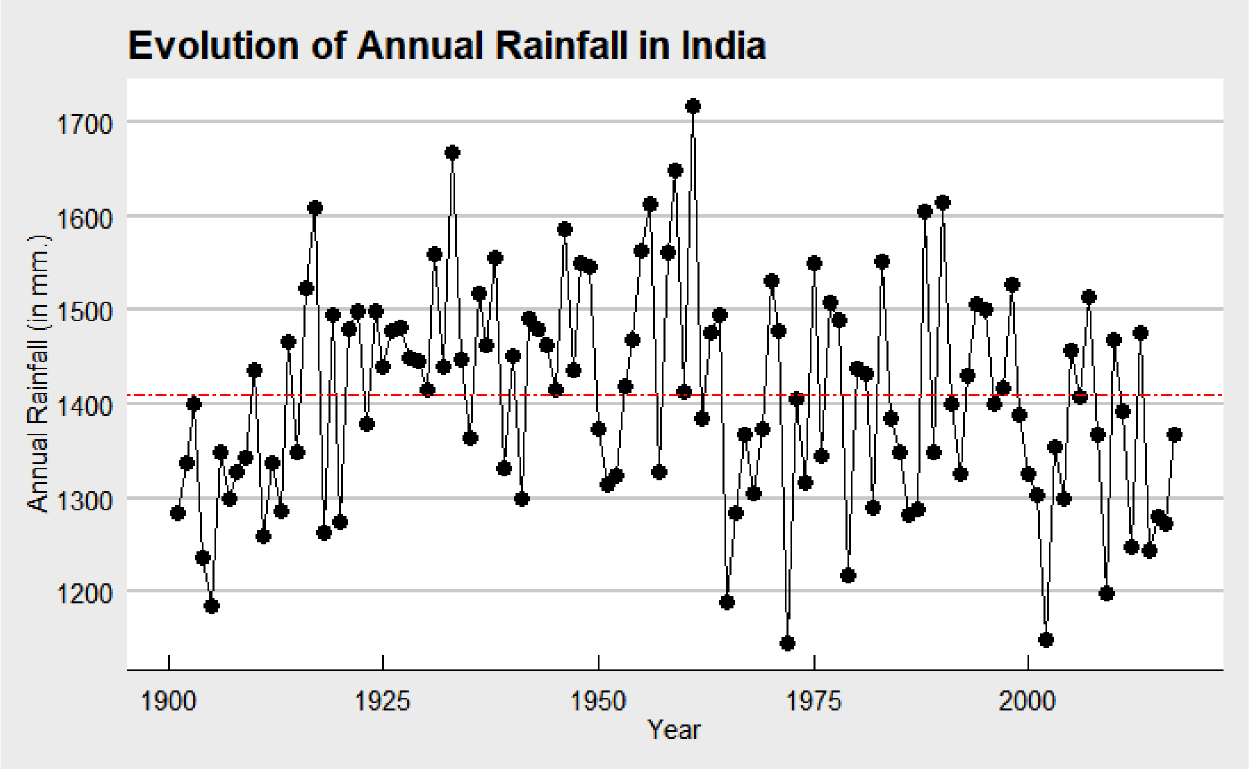

Fig. 1

From Fig. 1 we can see that the highest average rainfall in India was recorded in the year 1961 and the lowest average rainfall was recorded in the year 1972. The red line is the average of the annual rainfall in India throughout 1901 - 2017. It looks like after 1960s there has been a slight dip in the rainfall in India.

```
a <- rain %>%
  group_by(YEAR)%>%
  summarise(average = mean(JF, na.rm = TRUE))
a$seasonal <- "January - February"
b <- rain %>%
  group_by(YEAR)%>%
  summarise(average = mean(MAM, na.rm = TRUE))
b$seasonal <- "March - May"
c <- rain %>%
  group_by(YEAR)%>%
  summarise(average = mean(JJAS, na.rm = TRUE))
c$seasonal <- "June - September"
d <- rain %>%
  group_by(YEAR)%>%
  summarise(average = mean(OND, na.rm = TRUE))
d$seasonal <- "October - December"

e <- rbind(a, b, c, d)
e$order <- c(1:468)
```

```r
ggplot(e, aes(x = YEAR, y = average, group = seasonal, color = reorder(season
al, order))) +
  geom_line() +
  geom_point() +
  labs(title = "Evolution of Average Seasonal Rainfall in India", x = "Year",
y = "Average Seasonal Rainfall (in mm.)", color = "Seasons") +
  theme_bw()
```

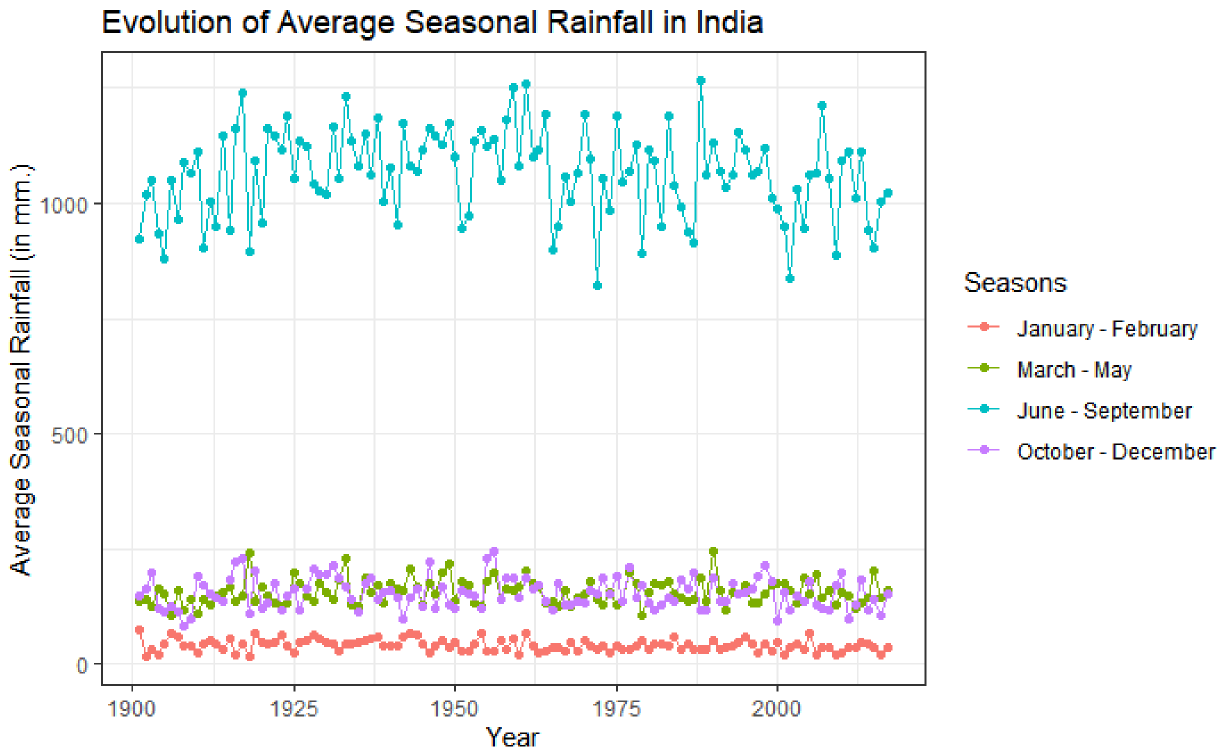

Fig. 2

The above Fig. 2 plots a seasonal rainfall, where we can observe that highest rainfall is received between June - September which is the monsoon season in India where the mean annual rainfall ranges from 800mm - 1500mm. March - May is the season before the monsoon and October - December marks the return of monsoon, thus we can see scanty rainfall during these periods. January - February is the late winters, when mostly there is no rainfall but due to the climatic changes, we can see showers below 50mm from time to time.

```r
f <- rain %>%
  group_by(SUBDIVISION)%>%
  summarise(average = mean(JF, na.rm = TRUE))
f$seasonal <- "January - February"
g <- rain %>%
  group_by(SUBDIVISION)%>%
  summarise(average = mean(MAM, na.rm = TRUE))
g$seasonal <- "March - May"
h <- rain %>%
  group_by(SUBDIVISION)%>%
  summarise(average = mean(JJAS, na.rm = TRUE))
h$seasonal <- "June - September"
i <- rain %>%
  group_by(SUBDIVISION)%>%
  summarise(average = mean(OND, na.rm = TRUE))
i$seasonal <- "October - December"

j <- rbind(f, g, h, i)
j$order <- c(1:144)

ggplot(j, aes(x = SUBDIVISION, y = average, fill = reorder(seasonal, order)))
+
  geom_bar(stat = "identity") +
  theme_bw() +
  labs(title = "Rainfall in Subdivisions of India", x = "", y = "Average Annu
al Rainfall (in mm.)", fill = "Seasons") +
  coord_flip()
```

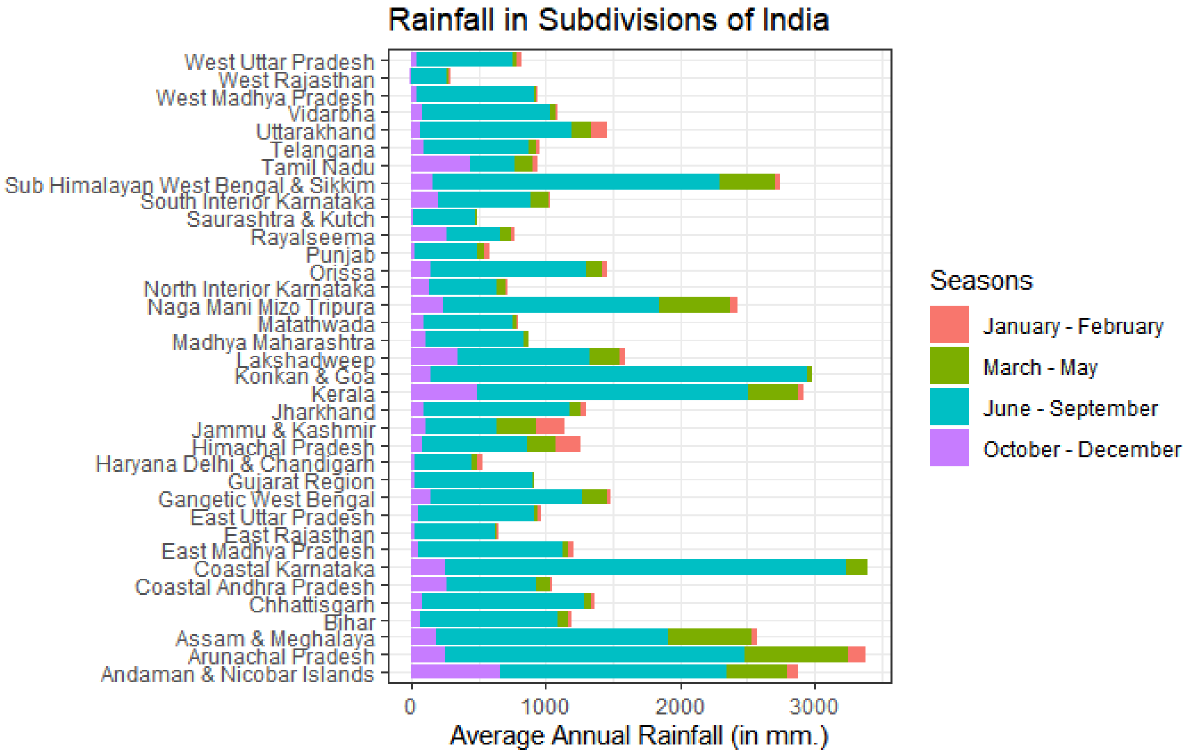

Fig. 3

Fig. 3 represents the Average annual rainfall received by the various sub-divisions in India throughout the seasons. Like Fig. 2 it portrays that most of the rainfall is received in the months of June - September which is the monsoon season and there are scanty showers throughout the years. Also, some of the sub-divisions namely, Coastal Karnataka, Arunachal Pradesh, Konkan Goa and Kerala receive highest rainfall throughout all the years. The sub-divisions like Rajasthan, Gujarat, Haryana and Punjab receives low rainfall. Interesting observation here is that Punjab and Haryana have high agricultural output despite low rainfall.

```r
ggplot(rain, aes(x = SUBDIVISION, y = ANNUAL)) +
  geom_boxplot() +
  theme_bw() +
  labs(title = "Annual Rainfall in Subdivisons of India", x = "Subdivision",
y = "Annual Rainfall (in mm.)") +
  coord_flip()
```

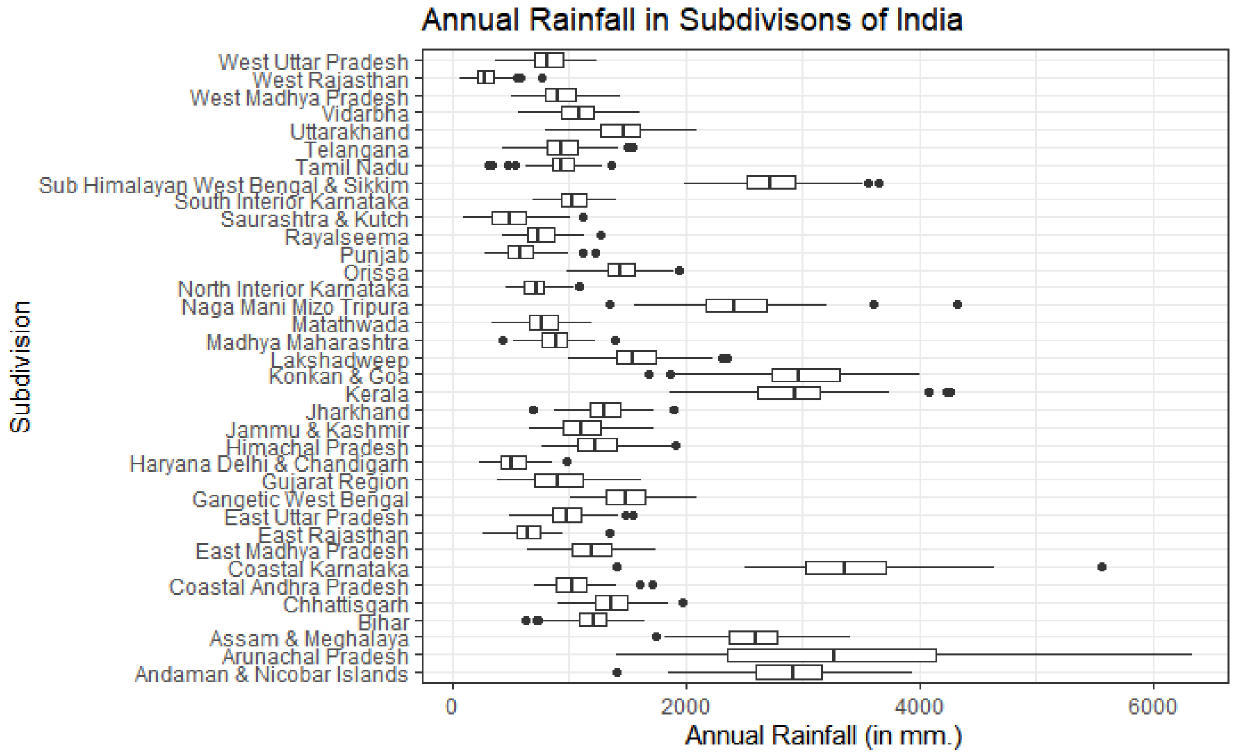

Fig. 4

The box plot shown in Fig. 4 is used to analyze the difference between the maiximum and minimum rainfall for all the Indian subdivisions. We can observe that Arunachal Pradesh has a wider range of annual rainfall, while West Rajasthan has the lowest range. Also, the mean annual rainfall for West Rajasthan is the lowest and the highest for Coastal Karnataka.

```
l <- rain %>%
  group_by(SUBDIVISION)%>%
  summarise(average = mean(ANNUAL, na.rm = TRUE))

ggplot(l, aes(x = reorder(SUBDIVISION, average), y = average)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  labs(title = "Subdivision wise Average Annual Rainfall", x = "Subdivision",
y = "Average Annual Rainfall (in mm.)") +
  coord_flip()
```

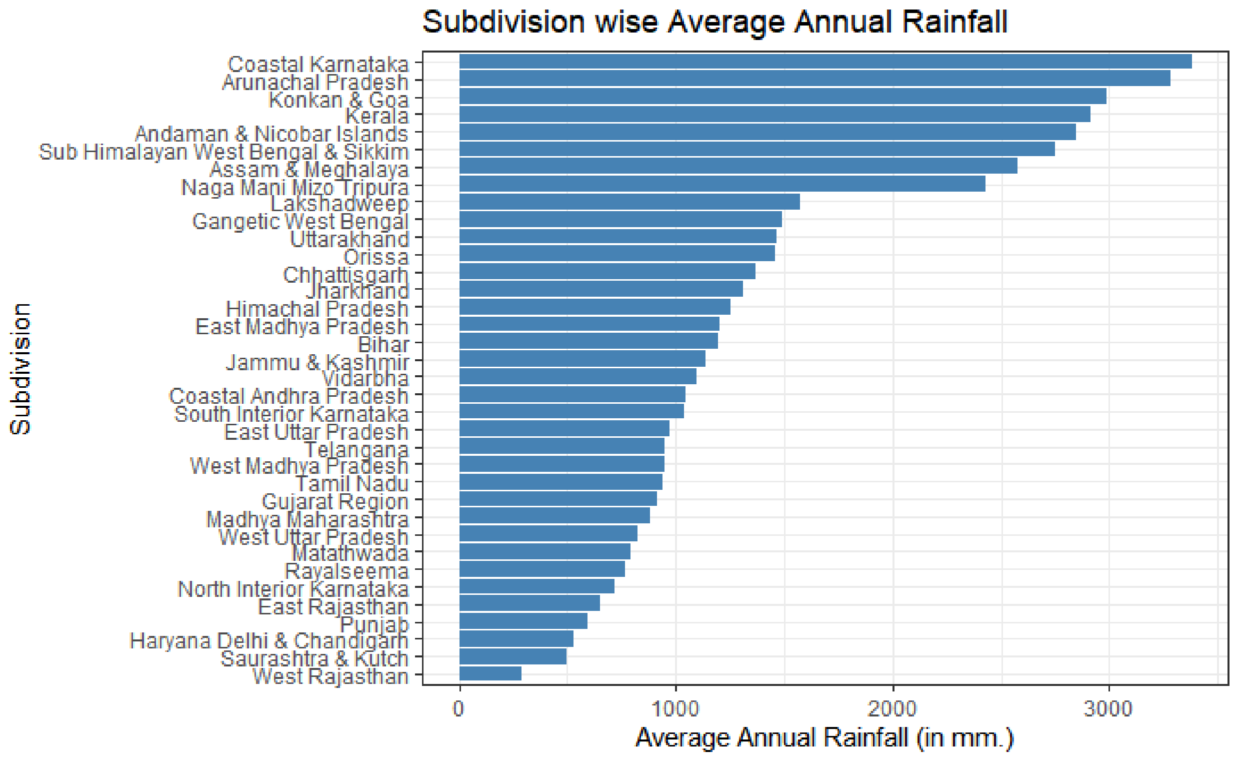

Fig. 5

Fig. 5 above is a bar plot that arranges all the subdivisions in the descending order depending upon their mean annual rainfall throughout all these years. We can see that Coastal Karnataka is the wettest region while West Rajasthan is the driest region.

```r
m <- rain %>%
  summarise(average = mean(JAN, na.rm = TRUE))
m$month <- "Jan"
n <- rain %>%
  summarise(average = mean(FEB, na.rm = TRUE))
n$month <- "Feb"
o <- rain %>%
  summarise(average = mean(MAR, na.rm = TRUE))
o$month <- "Mar"
p <- rain %>%
  summarise(average = mean(APR, na.rm = TRUE))
p$month <- "Apr"
q <- rain %>%
  summarise(average = mean(MAY, na.rm = TRUE))
q$month <- "May"
r <- rain %>%
  summarise(average = mean(JUN, na.rm = TRUE))
r$month <- "Jun"
s <- rain %>%
  summarise(average = mean(JUL, na.rm = TRUE))
s$month <- "Jul"
t <- rain %>%
  summarise(average = mean(AUG, na.rm = TRUE))
t$month <- "Aug"
u <- rain %>%
  summarise(average = mean(SEP, na.rm = TRUE))
u$month <- "Sep"
v <- rain %>%
  summarise(average = mean(OCT, na.rm = TRUE))
v$month <- "Oct"
w <- rain %>%
  summarise(average = mean(NOV, na.rm = TRUE))
w$month <- "Nov"
x <- rain %>%
  summarise(average = mean(DEC, na.rm = TRUE))
x$month <- "Dec"

y <- rbind(m, n, o, p, q, r, s, t, u, v, w, x)
y$order <- c(1:12)

ggplot(y, aes(x = reorder(month, order), y = average)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_bw() +
  labs(title = "Monthly Rainfall in India", x = "Month", y = "Average Monthly
Rainfall (in mm.)")
```

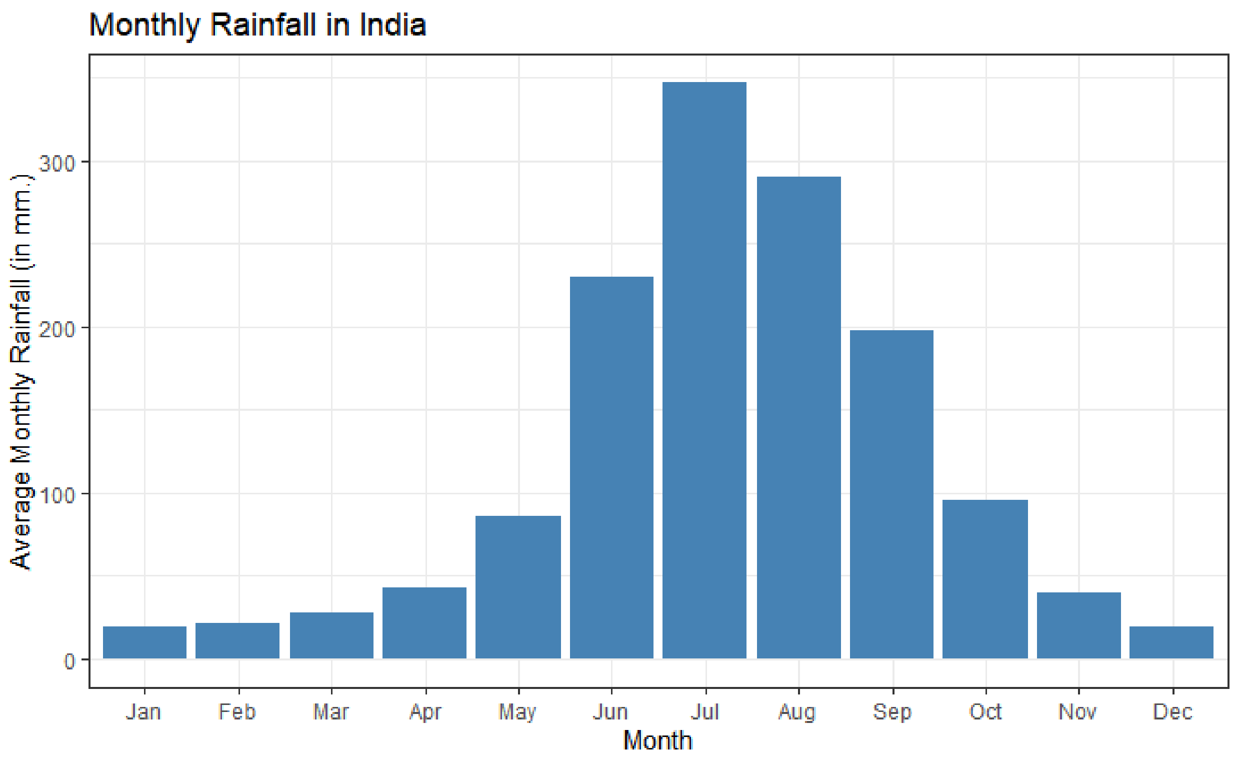

Fig. 6

From Fig. 6 we observe that a maximum of 347 mm. of average rainfall is received in the month of July, hence making it the peak of the monsoon season in India.

```
library(PerformanceAnalytics)

## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend

rain_data <- rain[, c(15, 16, 17, 18, 19)]
chart.Correlation(rain_data, histogram = TRUE, pch = 19)
```

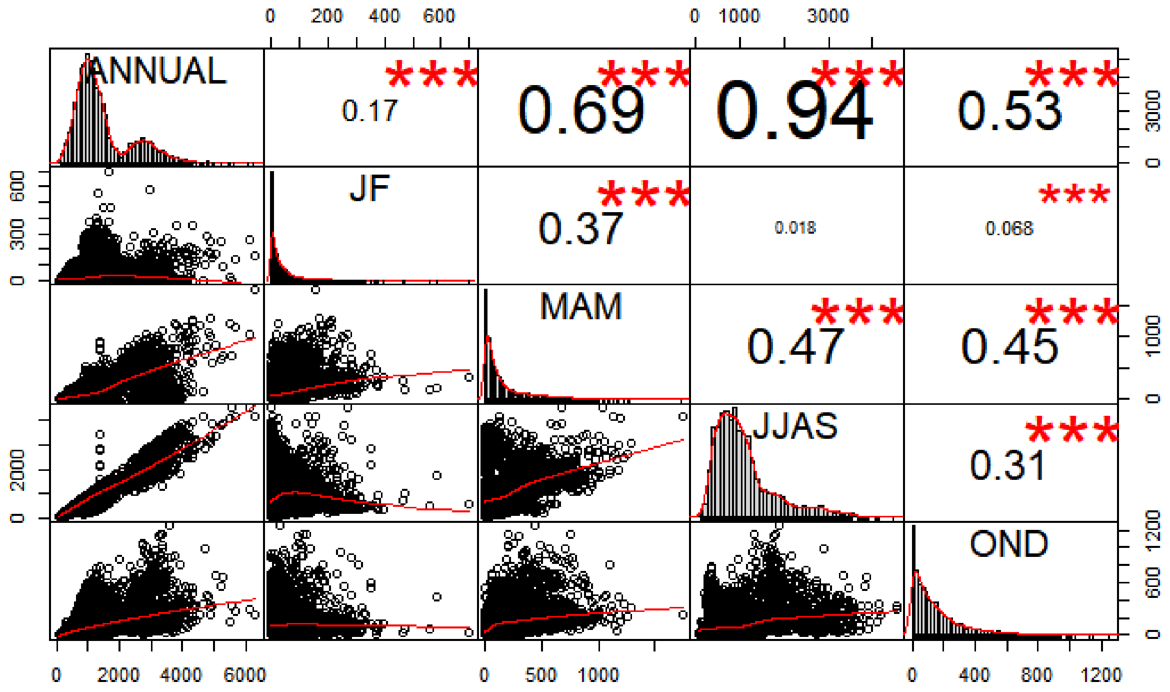

Fig. 7

Fig 7. plots the correlation between the four seasonal divisions and annual rainfall. The distribution of each variable is shown on the diagonal and below it are the bivariate scatter plots with a fitted line. The numbers above the diagonal are the values of the correlation plus the significance level as stars. Each significance level represents a p-value. Symbol (“***”, "**", "*”, “.”, "") represents p-value of 0, 0.001, 0.01, 0.05, 0.01, 1. We can see that the period of June - September has the highest correlation coefficient of 0.94, followed by March - May, October - December and finally January - February.

4. Statistical Analysis

Before performing any statistical analysis, we check if the annual rainfall data is normal using the descdist function.

```
library(fitdistrplus)

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: survival

## Loading required package: npsurv

## Loading required package: lsei

descdist(rain$ANNUAL)
```

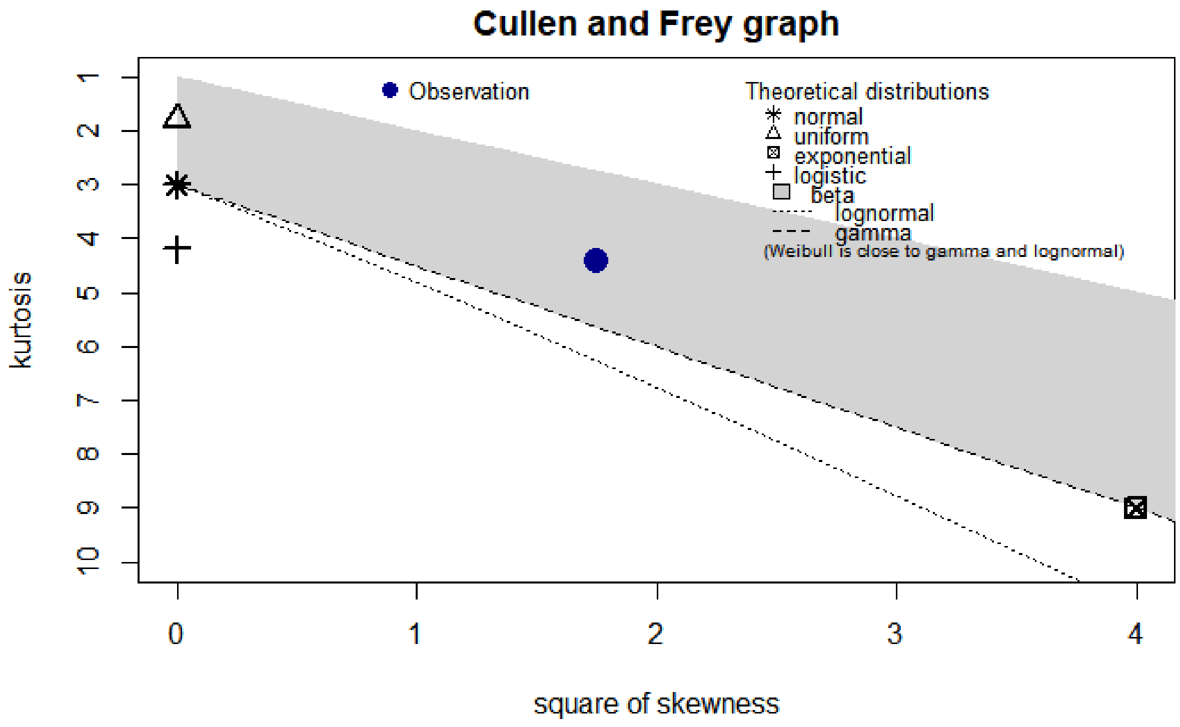

Fig. 8

```
## summary statistics
## ------
## min:  62.3   max:  6331.1
## median:  1124.15
## mean:  1409.449
## estimated sd:  899.7926
## estimated skewness:  1.322162
## estimated kurtosis:  4.410366
```

In Fig.8 we can clearly see that the observation point is not near to the normal distribution. Since there is only one possible value for skewness and kurtosis for a normal distribution it is represented as a point in the Cullen Frey Graph. Hence, the annual rainfall data is not normal.

```
rain$logannual <- log(rain$ANNUAL)
descdist(rain$logannual)
```

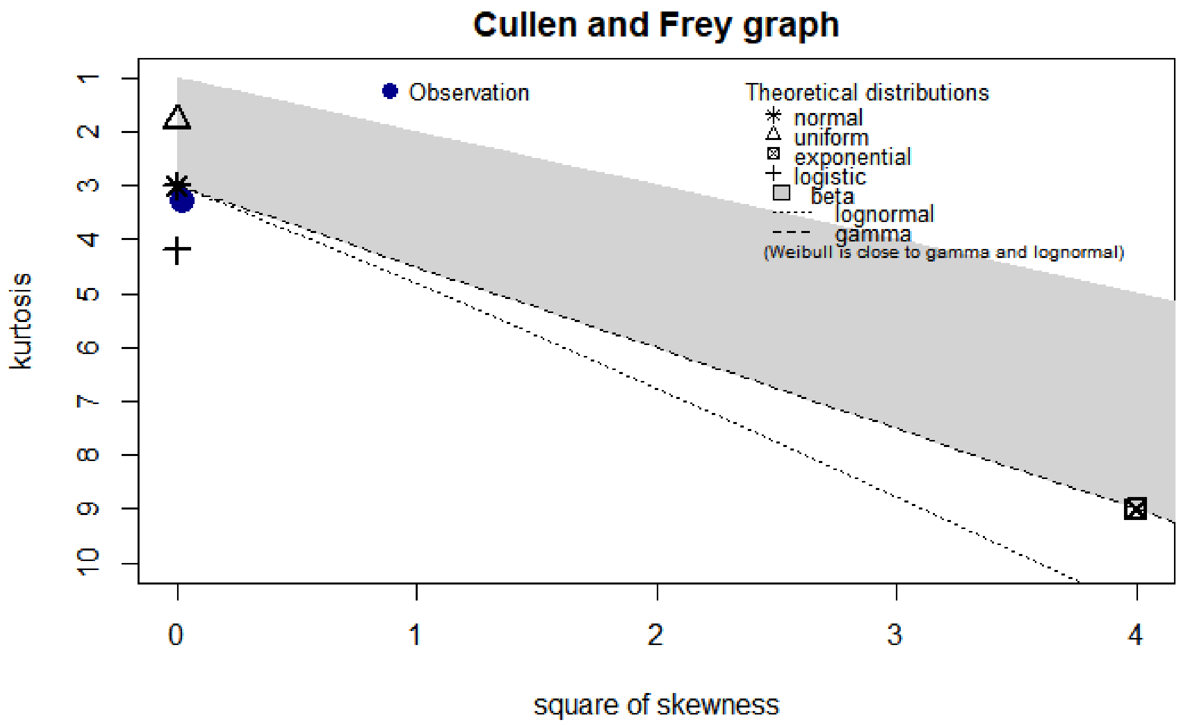

Fig. 9

```
## summary statistics
## ------
## min:  4.131961   max:  8.753229
## median:  7.024782
## mean:  7.062806
## estimated sd:  0.6234069
## estimated skewness:  -0.1658738
## estimated kurtosis:  3.308764
```

To make the annual rainfall data normal we perform log transformation on the entire column and save it in a new column named logannual. Using the Cullen and Frey graph, as seen in Fig. 9, we can confirm that the loganuual column is normal. For further statistical analysis we will be using the normalized annual rainfall data.

From the visualizations, we can see that Arunachal Pradesh is one of the states which recieves the highest rainfall.

Hypothesis test 1 : We want to test if the mean annual rainfall of Arunachal Pradesh is greater than the mean annual rainfall for all the subdivisions.

Hypothesis :

Ho : x > Mean annual rainfall

H1 : x < Mean annual rainfall

Population : The mean annual rainfall for all the subdivisions through 1901 - 2017. Sample : The mean annual rainfall for Arunachal Pradesh through 1901 - 2017.(x) Since n is greater than 30, we will apply Z - test in this case.

```
rain_ap <- rain[(rain$SUBDIVISION == "Arunachal Pradesh"), ]

mean_ap <- mean(rain_ap$logannual)                              # sample mean
n <- length(rain_ap$logannual)                                  # sample size
mu0 <- mean(rain$logannual)                                     # hypothesized va
lue
sigma <- sd(rain$logannual)                                     # population stan
dard deviation

z <- (mean_ap - mu0) / (sigma / sqrt(n))
z                                                               # test statistic

## [1] 15.49198

# We then compute the critical value at .05 significance level.

alpha <- .05
z.alpha <- qnorm(1 - alpha)
-z.alpha                                                        # critical value

## [1] -1.644854
```

This being a left-tailed test with critcal value of -1.645 Z does not lie in the rejection region. Hence,we fail to reject the null hypothesis. We can say that the mean annual rainfall for Arunachal Pradesh is greater than the mean annual rainfall for all the sub-divisions.

From the visualizations, we can see that Haryana, Delhi & Chandigarh is one of the driest subdivision with very less rainfall.

Hypothesis test 2 : We want to test if the mean annual rainfall of Haryana Delhi & Chandigarh is greater than the mean annual rainfall for all the subdivisions.

Hypothesis :

Ho : x > Mean annual rainfall

H1 : x < Mean annual rainfall

Population : The mean annual rainfall for all the subdivisions through 1901-2017. Sample : The mean annual rainfall for Haryana Delhi & Chandigarh through 1901-2017.(x) Since n is greater than 30, we will apply Z - test in this case.

```
rain_hdc <- rain[(rain$SUBDIVISION == "Haryana Delhi & Chandigarh"), ]

mean_hdc <- mean(rain_hdc$logannual)          # sample mean
n <- length(rain_hdc$logannual)               # sample size
mu0 <- mean(rain$logannual)                   # hypothesized value
sigma <- sd(rain$logannual)                   # population standard deviation

z <- (mean_hdc - mu0) / (sigma / sqrt(n))
z                                             # test statistic

## [1] -14.37828

# We then compute the critical value at .05 significance level.

alpha <- .05
z.alpha <- qnorm(1 - alpha)
-z.alpha                                      # critical value

## [1] -1.644854
```

This being a left - tailed test with critcal value of -1.645, Z lies in the rejection region. Hence, we reject the null hypothesis. We can say the mean annual rainfall for Haryana, Delhi & Chandigarh is less than the mean annual rainfall for all the sub-divisions.

Hypothesis test 3: We want to test which subdivision between Telangana and West Madhya Pradesh receives the greater amount of mean annual rainfall till 2017.

Hypothesis :

Ho : x1-x2 > 0

H1 : x1-x2 < 0

Sample : The mean annual rainfall for Telangana through 1901 - 2017.(x1) & mean annual rainfall for West Madhya Pradesh through 1901 - 2017. Since n is greater than 30, we will apply Z - test in this case.

```
rain_t <- rain[(rain$SUBDIVISION == "Telangana"), ]
rain_mp <- rain[(rain$SUBDIVISION == "West Madhya Pradesh"), ]
```

```
mean_t <- mean(rain_t$logannual)               # Mean of first sample
mean_mp <- mean(rain_mp$logannual)             # Mean of second sample
var_t <- var(rain_t$logannual)                 # Variance of first sample
var_mp <- var(rain_mp$logannual)               # Variance of second sample
n_t <- length(rain_t$logannual)                # Sample size of first sample
n_mp <- length(rain_mp$logannual)              # Sample size of second sample

z <- ((mean_t) - (mean_mp) - 0) / (sqrt((var_t / n_t) + (var_mp / n_mp)))
z                                              # test statistic

## [1] 0.03333038

# We then compute the critical value at .05 significance level.

alpha <- .05
z.alpha <- qnorm(1 - alpha)
-z.alpha                                       # critical value

## [1] -1.644854
```

Since Z-score does not lie in the rejection region, we fail to reject the hypothesis. Thus, we can state that the annual rainfall in Telangana is greater than the annual rainfall in West Madhya Pradesh.

5. Advanced Statistical Analysis

Linear Regression

Linear regression is one of the simplest and most common supervised algorithms that data scientists use for predictive modeling. In R studio we will be using lm, abline and predict functions to perform linear regression. lm() is used to fit linear models and abline() is used to plot the regression line over the current plot. While the predict function predicts the result using the linear model formed in lm.

Here, we go ahead and check the amount of annual rainfall that Arunachal Pradesh will receive in the year 2020.

```
library(graphics)
library(stats)

# Scatter plot between year and annual rainfall for Arunachal Pradesh

{plot(rain_ap$YEAR, rain_ap$logannual, xlab = "YEAR", ylab = "ANNUAL")
fit_1 <- lm(logannual ~ YEAR, data = rain_ap)
summary(fit_1)
abline(fit_1)}
```

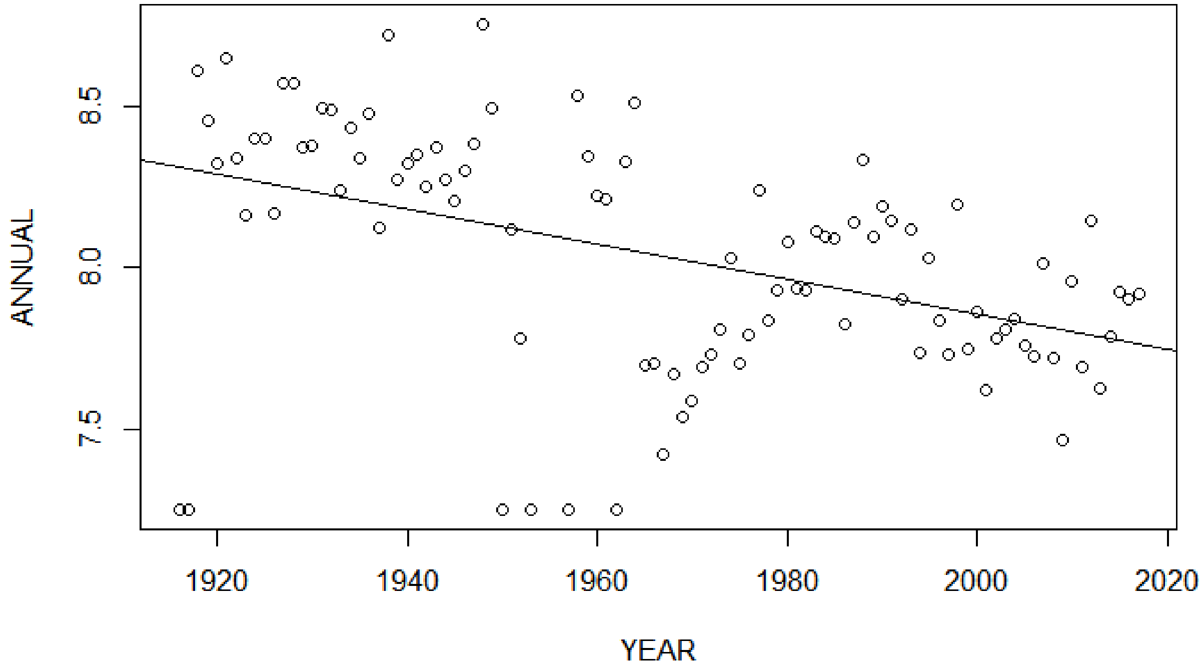

Fig. 10

```
# Predict Annual rainfall in Arunachal Pradesh in year 2020

p <- predict(fit_1, data.frame(YEAR = 2020))
exp(p)

##        1
## 2307.635
```

As per the above result, Arunachal Pradesh will receive 2307.635mm of rainfall in the year 2020.

Secondly, we predict the amount of annual rainfall that Haryana, Delhi & Chandigarh will receive in the year 2020.

```
# Scatter plot between year and annual rainfall for Haryana Delhi & Chandigar
h

{plot(rain_hdc$YEAR, rain_hdc$logannual, xlab = "YEAR", ylab = "ANNUAL")
fit_2 <- lm(logannual ~ YEAR, data = rain_hdc)
summary(fit_2)
abline(fit_2)}
```

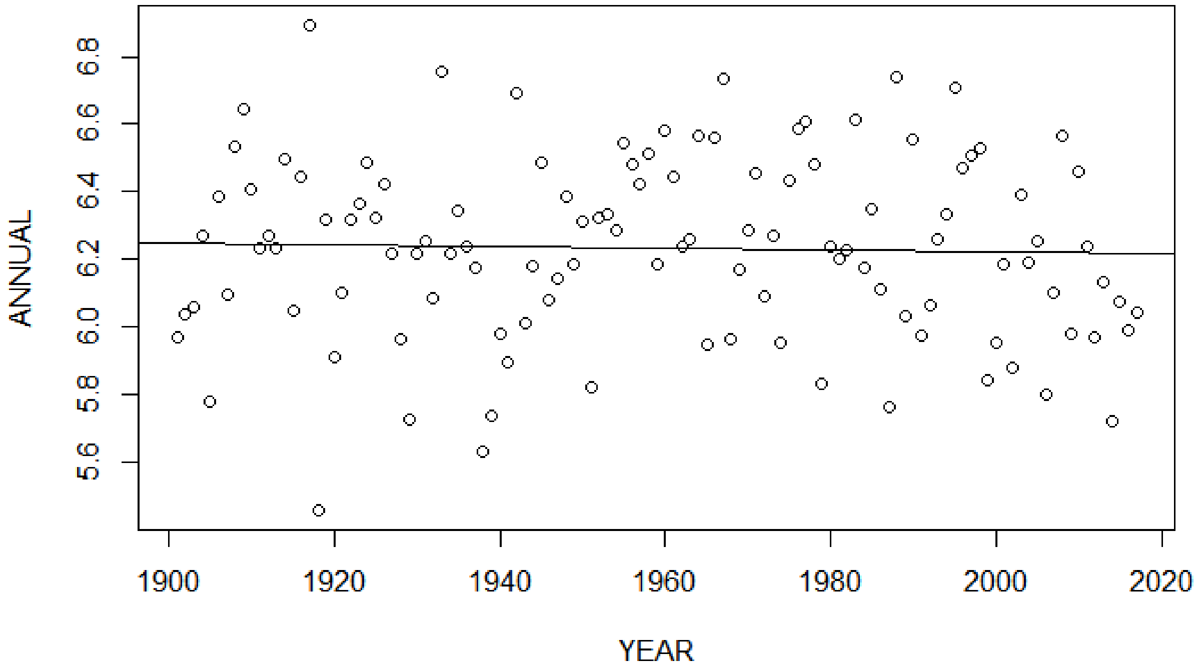

Fig. 11

```
# Predict Annual rainfall in Arunachal Pradesh in year 2020
q <- predict(fit_2, data.frame(YEAR = 2020))
exp(q)

##       1
## 501.893
```

As per the above result Haryana, Delhi & Chandigarh will receive 501.893 mm. of rainfall in the year 2020.