Project 1

Nikant Yadav

2024-05-27

This project uses R to measure and analyse climate change over the years using different data sets.

PART 1.1

Part 1.1 of the project involves analyzing the land-ocean temperature anomalies data over the period 1880-2016, taking the data from NASA's Goddard Institute for Space Studies website.

We begin with setting the working directory and importing the data file.

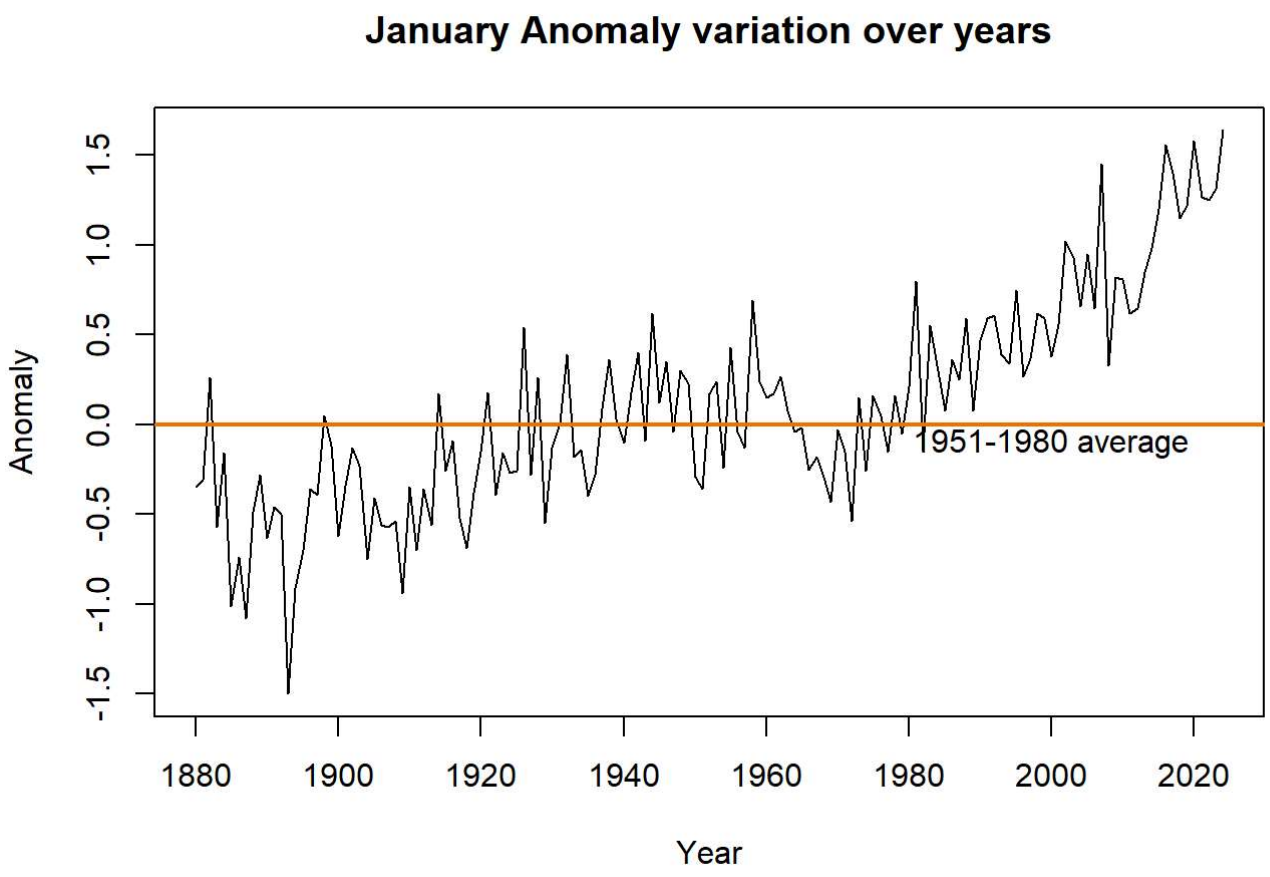
```
setwd("C:\\Users\\Nikant Yadav\\Desktop\\Internship Preparation\\ECO project\\Project 1\\Nikant
Yadav")
tempdata <- read.csv("NH.Ts+dSST.csv", skip = 1, na.strings = "***")
```

While running this code, replace the working directory with the directory that has your data file.

Now we choose the month January, and plot a line temperature with average temperature anomaly on the vertical axis and time on the horizontal axis.

```
plot( tempdata$Year, tempdata$Jan,
      xlab="Year", ylab="Anomaly",
      main="January Anomaly variation over years", type="l")

abline(h = 0, col = "darkorange2", lwd=2)
text(2000, -0.1, "1951-1980 average")
```



The average temperature during the period 1951-1980 is taken as the base temperature and all anomalies are measured from this base temperature.

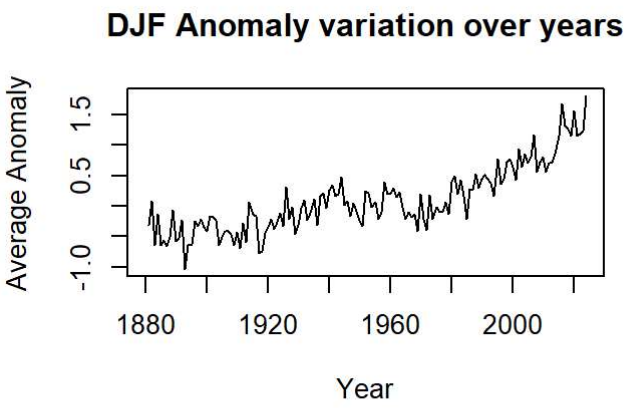
In order to make the navigation through data of different months, this 'column' vector is created.

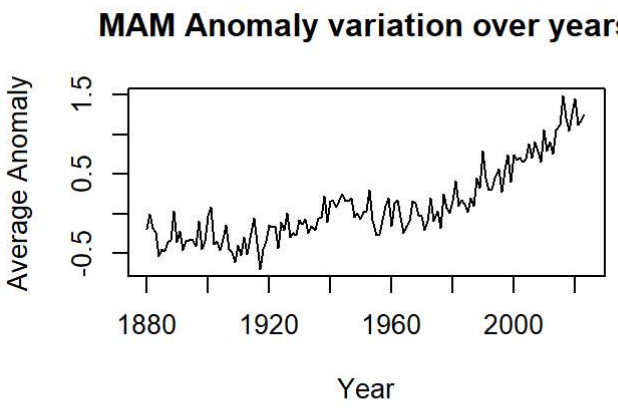
```
column = c("Jan", "Feb", "Mar", "Apr","May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec","J
D","DN" ,"DJF", "MAM","JJA","SON")
```

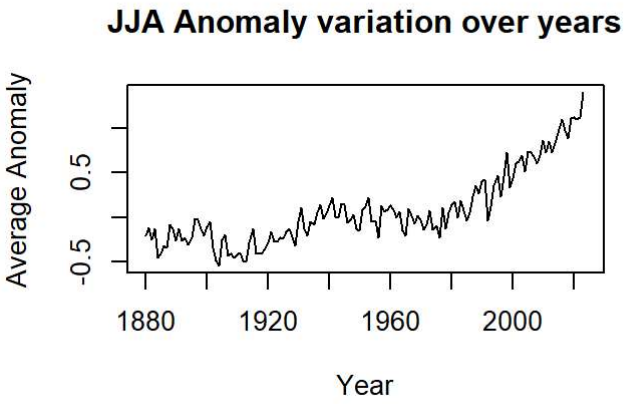
Now we plot the seasonal averages using average temperature anomaly for that season on the vertical axis and time on the horizontal axis.

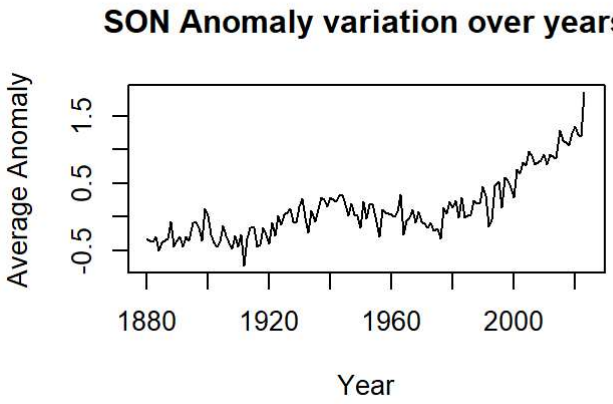
```
par(mfrow = c(2, 2))

for(i in 15:length(column)){
  yaxis = column[i]
  plot(tempdata$Year, tempdata[[yaxis]],
       xlab="Year", ylab="Average Anomaly",
       main=paste(column[i], "Anomaly variation over years"), type="l")
}
```







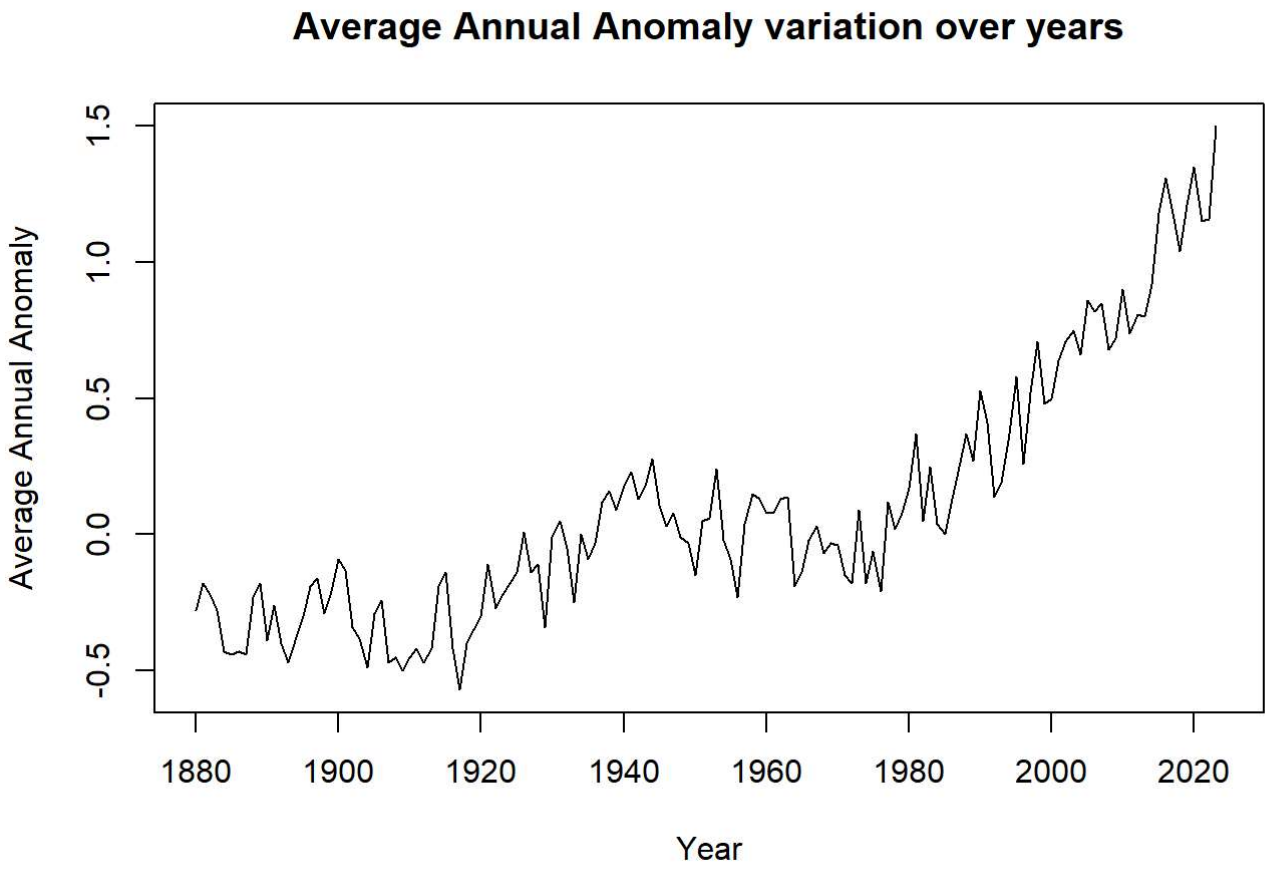


Setting the plotting area again to include one plot.

```
par(mfrow = c(1,1))
```

Plotting a line chart with annual average temperature anomaly on the vertical axis and time on the horizontal axis.

```
plot(tempdata$Year, tempdata$J.D,
     xlab="Year", ylab="Average Annual Anomaly",
     main= "Average Annual Anomaly variation over years", type="l")
```



According to the graphs, temperature anomalies are growing on the positive side of the base line as time goes on. This is virtually always the case, which explains why the three-month average anomalies and even the annual average temperature anomalies are growing over time.

For each time interval, discuss what we can learn about patterns in temperature over time that we might not be able to learn from the charts of other time intervals.

If we study, only month wise graphs/data we are missing out the factor of changing seasons on earth. If we study only seasons, then we are missing out the factor, that different part of the world has different seasons at different times. Also, the analysis of overall year gives a better view at the temperature anomalies. Along with that, monthly/season anomalies can help us analyze which seasons/months are contributing most to the global temperature rise, and thus can help us in identify the factors related to that.

PART 1.2

Now we start with part 1.2 of the project In this part, we analyse the temperature anomalies, by using frequency tables and histograms. For the frequency table, we use intervals of 0.05 each to create groups of anomaly temperatures.

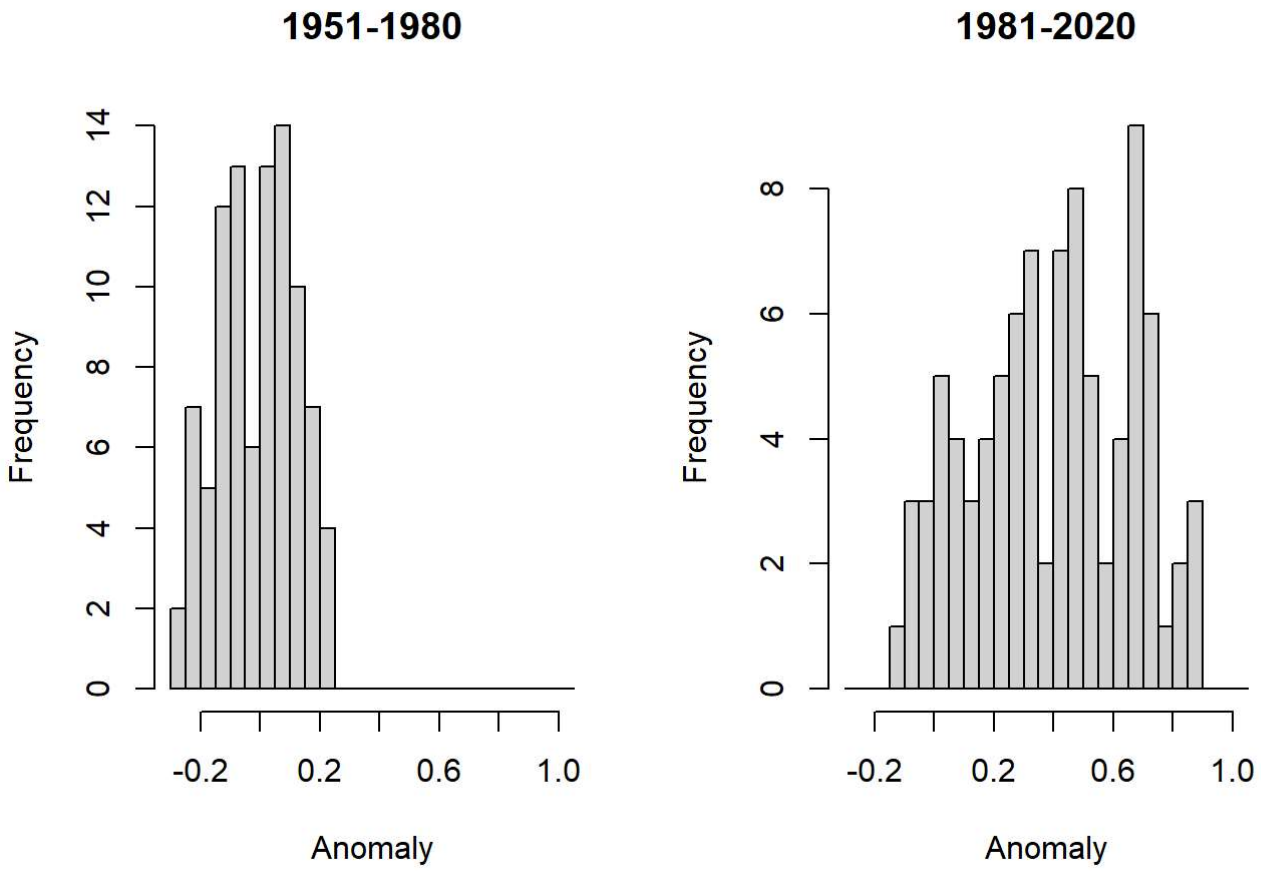
We are supposed to create frequency table and plot column charts for the years 1951–1980 and 1981-2020. The values in the first column range from −0.3 to 1.05, in intervals of 0.05.

Creating frequency table using my method!

```
rangeof_anomaly <- seq(-0.3, 1.05, 0.05)

data51_80 = array(tempdata[71:101, c("Jun", "Jul", "Aug")])
data81_20 = array(tempdata[102:131, c("Jun","Jul","Aug")])
vector_data51 <- unlist(data51_80)
vector_data81 <- unlist(data81_20)
freq1 <- table(cut(vector_data51, breaks=rangeof_anomaly))
freq2 <- table(cut(vector_data81, breaks=rangeof_anomaly))

par(mfrow = c(1, 2))
hist1 <- hist(vector_data51, breaks=rangeof_anomaly, plot=TRUE, xlab="Anomaly", main="1951-198
0")
hist2 <- hist(vector_data81, breaks=rangeof_anomaly, plot=TRUE, xlab="Anomaly", main="1981-202
0")
```



In this method, I first define a vector *(rangeof_anomaly)* that is a sequence from -0.3 to 1.05 with a gap of 0.05 at each interval.

Then the original dataframe, is divided into two parts corresponding to the year ranges initially assumed (1951-1980 and 1981-2020). The dataframe are converted to vectors, and then frequency tables are created using the table function.

Histograms for both temperature ranges are created using the hist function.

Creating frequency tables, using method given in the book.

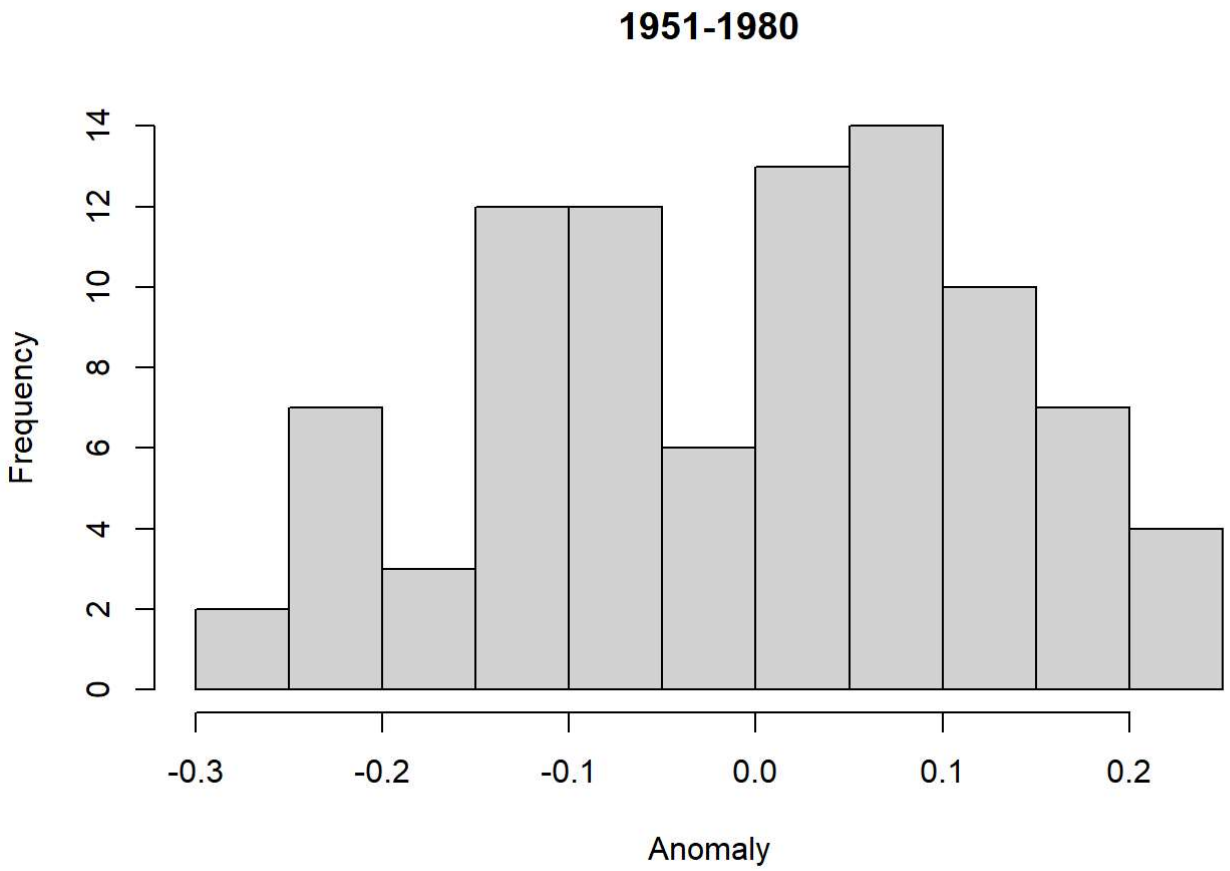
```
tempdata$Period <- factor(NA, levels = c("1921-1950", "1951-1980", "1981-2020"), ordered = TRUE)
```

A new column is added to the original dataframe using factor function on the levels 1921-1950, 1950-1980 and 1981-2020.

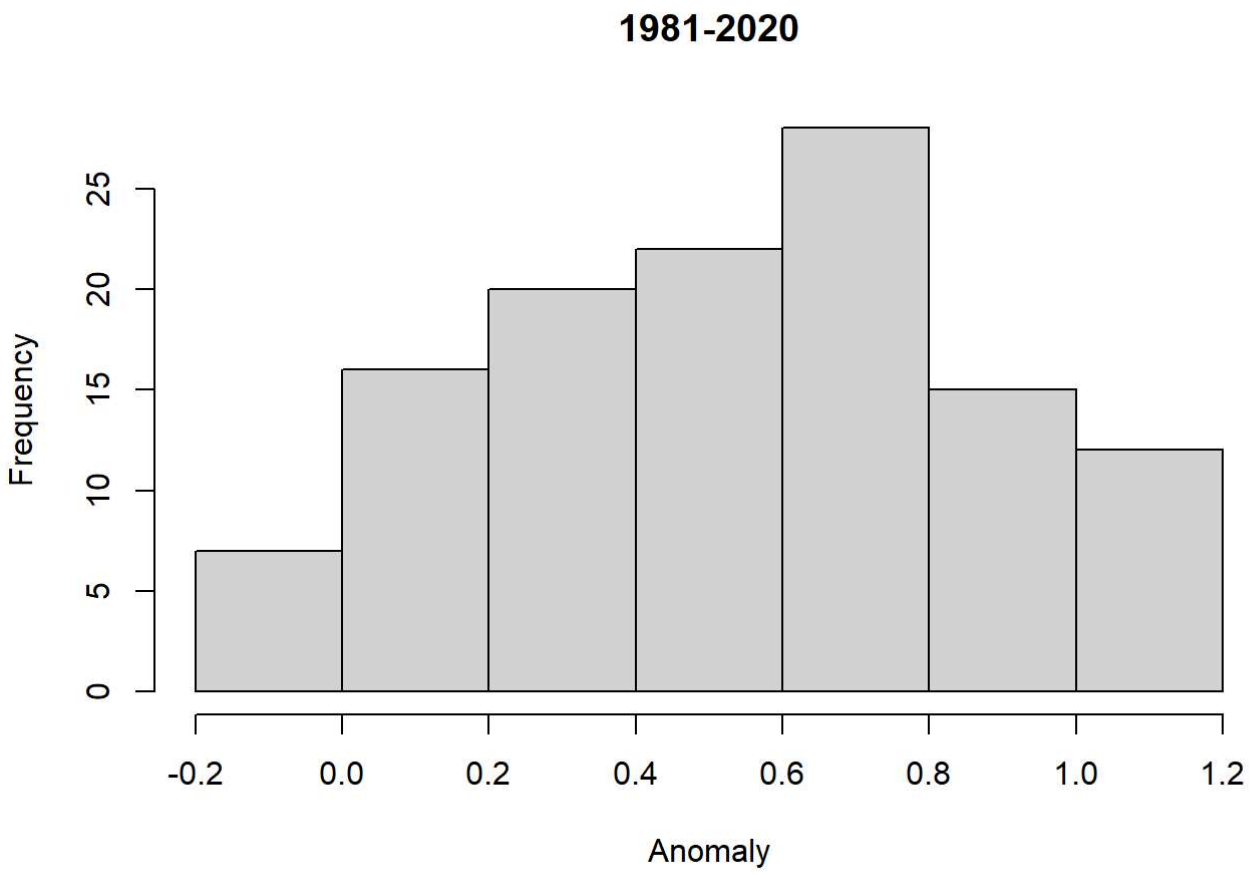
```
tempdata$Period[(tempdata$Year > 1920) & (tempdata$Year < 1951)] <- "1921-1950"
tempdata$Period[(tempdata$Year > 1950) & (tempdata$Year < 1981)] <- "1951-1980"
tempdata$Period[(tempdata$Year > 1980) & (tempdata$Year < 2021)] <- "1981-2020"
temp_summer <- c(tempdata$Jun, tempdata$Jul, tempdata$Aug)
temp_Period <- c(tempdata$Period, tempdata$Period, tempdata$Period)
temp_Period <- factor(temp_Period, levels = levels(tempdata$Period))
```

The dataframe is then filtered to only include the data in the ranges given in the above code. After that, a new dataframe *(temp_summer)* is created to include all the filter data into one data frame.Now we have one long variable *(temp_summer)*, with the monthly temperature anomalies for the three months. To make a variable showing the categories for the temp_summer variable, we use the c function again.After using the c function, we had to use the factor function again to tell R that our new variable *(temp_Period)* is a factor variable.

```
hist1 = hist(temp_summer[temp_Period == "1951-1980"], main = "1951-1980", xlab = "Anomaly", ylab
= "Frequency")
```




```
hist2 = hist(temp_summer[temp_Period == "1981-2020"], main = "1981-2020", xlab = "Anomaly", ylab
= "Frequency")
```



We use the hist function on the monthly temperature anomalies from the period '1951–1980'

Using both methods, we can easily observe, that in the period 1951-1980, the anomalies were more on the left side (closer to -0.3), but for the next period 1981-202, the anomalies shifted to the right side. This shows that there has been a temperature rise as time has progressed, in other words, an indication of global warming.

PART 1.2 (3)

In decile terms, temperatures in the 1st to 3rd decide are termed as 'cold' and temperatures in 7th to 10th decile are termed as 'hot'. We will now use two methods, to find the values that correspond to the 3rd and 7th decile across all months in 1951-1980

In descriptive statistics, the term "decile" refers to the nine values that split the population data into ten equal fragments such that each fragment is representative of 1/10th of the population. In other words, each successive decile corresponds to an increase of 10% points that the 1st decile or D1 has 10% of the observations below it. Then, the 2nd decile, or D2, has 20% of the observations below it, and so on.

For the first method, we use the definition of decile, and use the formula - D(x) = $\frac{(n+1)(x)}{10}$

where n is the total number of observations. In order to use this formula, we need to first arrange the given data in increasing order.

```
temp <-
temp_all_5180 <- subset(tempdata, (Year>=1951 & Year <= 1980))
temp51_80 <- unname(unlist(temp_all_5180[,2:13]))
```

We create a vector *(temp51_80)* containing the raw data from 1951-1980 without any headers etc.

```
n <- length(temp51_80)
for (j in 1:(n - 1)) {
  for (i in 1:(n - j)) {
    if (temp51_80[i + 1] < temp51_80[i]) {
      temp <- temp51_80[i]
      temp51_80[i] <- temp51_80[i + 1]
      temp51_80[i + 1] <- temp
    }
  }
}
```

Here we use bubble sort, to sort the data in increasing order and update into the existing vector *(temp51_80)*.

```
decile <- c()
for (i in 1:10) {
  decile[i] = i*(n+1)/10
}

dec3 <- temp51_80[floor(decile[3])] + (decile[3] - floor(decile[3])) * (temp51_80[ceiling(decile
[3])] - temp51_80[floor(decile[3])])

dec7 <- temp51_80[floor(decile[7])] + (decile[7] - floor(decile[7])) * (temp51_80[ceiling(decile
[7])] - temp51_80[floor(decile[7])])
```

Then, we calculate the D(x) for x ranging from 1 to 10, and then calculate the values corresponding to 3rd and 7th decile.

Next, we use the method that uses the quantile function of R, to simply calculate the percentiles. I find this method a little less informing for someone who does not have the proper knowledge of deciles, hence I prefer not to use this method.

```
perc <- quantile(temp51_80, c(0.3, 0.7))
p30 <- perc[1]
p70 <- perc[2]
```

PART 1.2 (4)

Here, we are supposed to count the number of anomalies, that are considered hot in 1981-2010, express this as a percentage of all the temperature observations in that period.

```
temp_all_8110 <- subset(tempdata, (Year>=1981 & Year <= 2010))
temp81_10 <- unname(unlist(temp_all_8110[,2:13]))
```

Here we again, create a vector *(temp81_10)* containing the raw data from the period 1981-2010 without headers etc.

```
n <- length(temp81_10)
for (j in 1:(n - 1)) {
  for (i in 1:(n - j)) {
    if (temp81_10[i + 1] < temp81_10[i]) {
      temp <- temp81_10[i]
      temp81_10[i] <- temp81_10[i + 1]
      temp81_10[i + 1] <- temp
    }
  }
}
```

Then, we use bubble sort to arrange this data in increasing order.

```
count <- 0

for (i in 1:n){
  if(temp81_10[i]>p70) {
    count <- count + 1
  }
}
percentage <- (count/n)*100
```

Next, we define a count variable initiating it with a value 0. We use a for loop, to increase the count by 1, every time an anomaly value is greater than the value corresponding to 7th decile or 70 percentile (considered as hot). Using this count value, we calculate the percentage.

We get the percentage value to be 84.72 %,

In 1951-1980, the hot temperatures are 30% of the overall anomalies, but in 1981-2010 the hot temperatures are 84.72% of the overall anomalies, hence the temperature is increasing more than before.

PART 1.2 (5)

Here, we are supposed to find the mean and variance separately for the following time periods: 1921–1950, 1951–1980, and 1981–2010.

This can be done using two methods. The first method is as follows -

```
temp_all_2150 <- subset(tempdata, (Year>=1921 & Year <= 1950))
season21_50 <- temp_all_2150[ , 16:19]
DJF21_50 <- unname(unlist(season21_50[1]))
MAM21_50 <-  unname(unlist(season21_50[2]))
```

Here, we first create/separate the dataset of which we need to find the mean and variance.

```
mean(DJF21_50)
```

```
## [1] -0.03366667
```

```
var(DJF21_50)
```

```
## [1] 0.0570723
```

Then we simply use the ‘mean’ and ‘var’ functions of R, to easily calculate the mean and variance for that period. This method is lengthy, because it requires creating differnet data set for different seasons of different time periods.

The second method resolves this issue

```
library(mosaic)
```

```
mean_DJF <- mean(~DJF|Period,data = tempdata)
mean_MAM <- mean(~MAM|Period,data = tempdata)
mean_JJA <- mean(~JJA|Period,data = tempdata)
mean_SON <- mean(~SON|Period,data = tempdata)
var_DJF <- var(~DJF|Period,data = tempdata)
var_MAM <- var(~MAM|Period,data = tempdata)
var_JJA <- var(~JJA|Period,data = tempdata)
var_SON <- var(~SON|Period,data = tempdata)
```

One way to calculate mean and variance is to use the mosaic package.We first install the mosaic package and load it into R with library(mosaic)

Using the data in tempdata (data = tempdata), we calculated the mean (mean) and variance (var) of variable ~DJF separately for (|) each value of Period. The mosaic package allows us to calculate the means/variances for each period all at once.

On comparing the variance of different seasons, for all seasons the variability of anomaly temperature has increased in the later period of time. Temperature in most seasons appears to be more variable in 1981–2010 compared to 1951–1980 or 1921–1950 (and the mean anomaly in each season has increased in each period for most seasons). The temperature anomalies in DJF have a larger variance than those in JJA. The variance in DJF is about three times larger than that in JJA, particularly until 1980. For the period 1981–2010, the JJA temperature anomalies start becoming more variable.

PART 1.3

We are now going to look at the relation between carbon emissions over time and temperature anomalies using scatter plots and ‘correlation’.

```
co2data <- read.csv("co2data3.csv", na.strings = "***")
```

We import the file containing the CO2 emmissions data in the project.

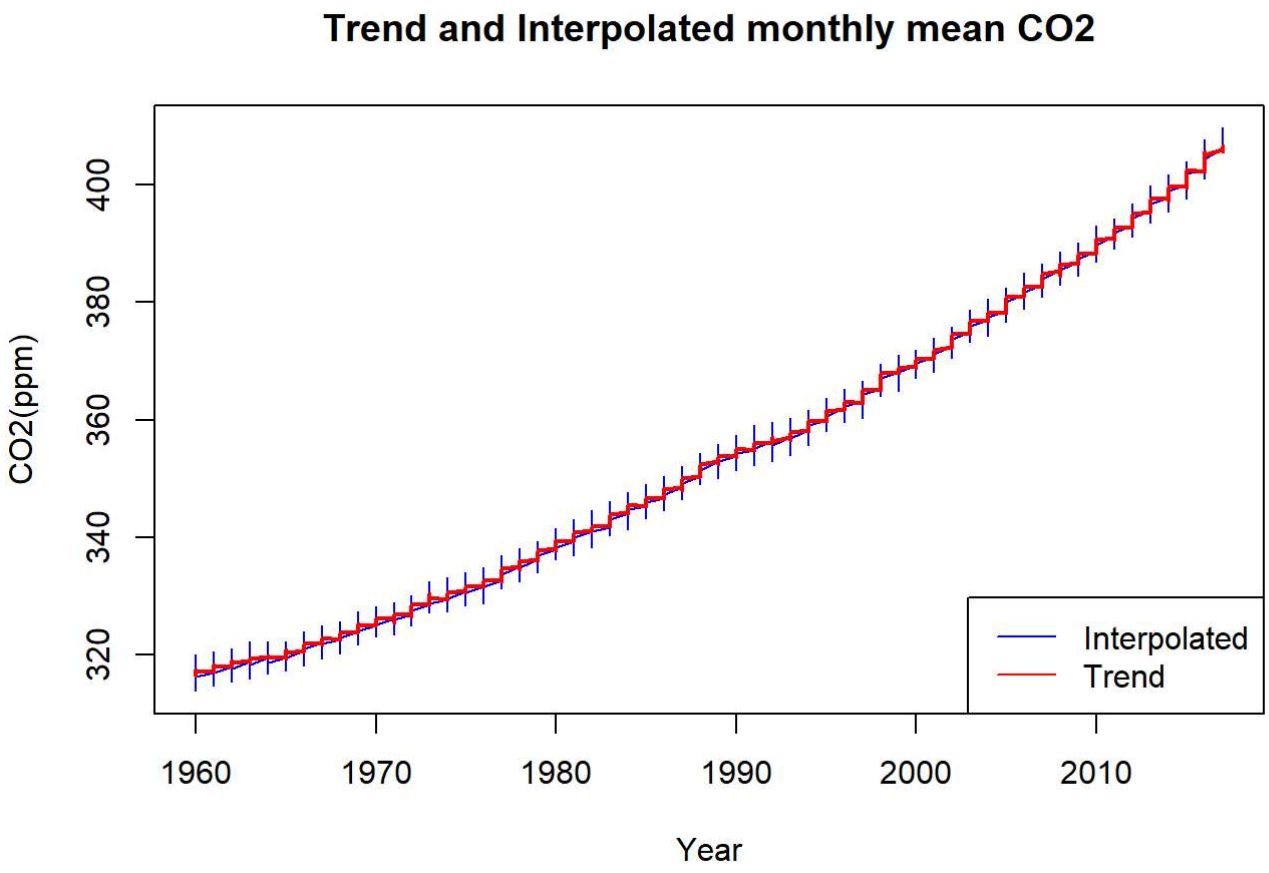
In the dataset, the trend mean mole fraction for each month is determined by removing the seasonal cycles. Trend values are linearly interpolated for missing values. The interpolated value is the sum of the average seasonal cycle value and the trend value.

Now we plot a line chart with interpolated and trend CO2 levels on the vertical axis and time (starting from January 1960) on the horizontal axis.

```
par(mfrow = c(1, 1))
plot(co2data$Year[23:length(co2data$Year)], co2data$Interpolated[23:length(co2data$Year)], type
="l",
     xlab="Year", ylab="CO2(ppm)", col="blue", main="Trend and Interpolated monthly mean CO2")

lines(co2data$Year[23:length(co2data$Year)], co2data$Trend[23:length(co2data$Year)], type="l",
      xlab="Year", ylab="CO2(ppm)", col = "red", lwd="2")

legend("bottomright", legend = c("Interpolated", "Trend"), col = c("blue", "red"), lty = 1)
```



The CO2 emmissions are increasing year by year fluctuating from a highest to a lowest value within a year.

We will now combine the CO2 data with the temperature data and then examine the relationship between these two variables visually using scatterplots, and statistically using the correlation coefficient.

We choose the month of Jan, add temperature data to it and do the analysis using scatter plots and Pearson correlation coefficient.

```
trend <- unlist(co2data[5])
month <- unlist(co2data[2])
jan <- c()
n = length(co2data$Year)
for(i in 1:n) {
  if(month[i] == 1) {
    jan <- append(jan,trend[i])
  }
}
```

Here, we first separate the trend and month values from the CO2 emmissions dataset in two different vectors.

We define an empty vector (jan) and using a for loop and if statement, update (jan) to include all trend values of the month January for all the years.

```
head(jan)
```

```
## Trend11 Trend23 Trend35 Trend47 Trend59 Trend71
##  315.70  316.51  317.03  318.06  318.91  319.67
```

The vector (jan) would look something like this.

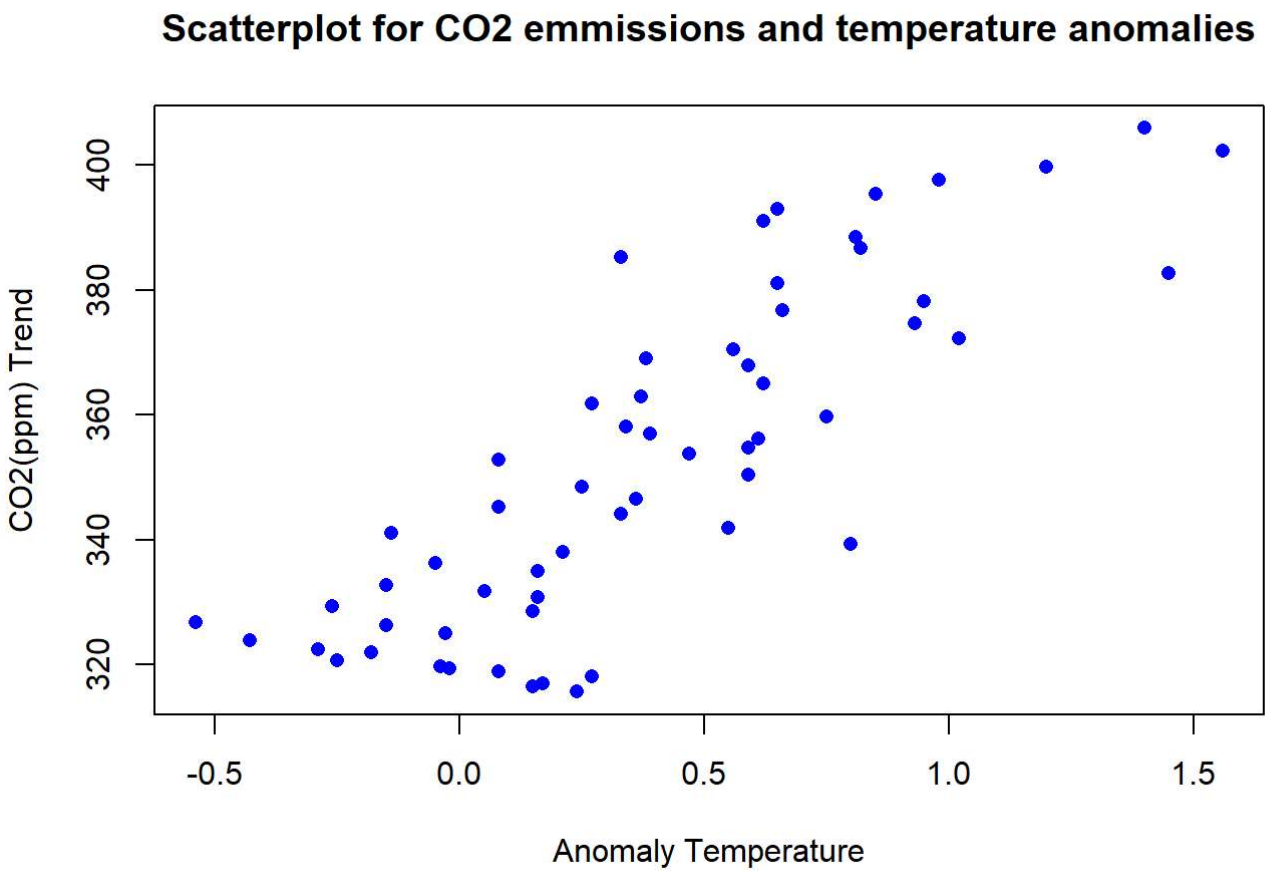
```
jan <- unname(jan)
zero1 <- rep(0,79)
zero2 <- rep(0,7)
jan <- append(zero1,jan)
jan <- append(jan, zero2)
tempdata$co2trendjan <- jan
```

Then we create two vectors containing only zeros, and append them before and after (jan) such that the length of this vector matches the length of tempdata. We create a new column in tempdata to include this newly updated vector (jan) such that it has trend values for only the period 1959-2017, and rest are 0.

This enables us to create a scatter plot easily in the following way -

```
filtered_data <- tempdata[tempdata$co2trendjan != 0, ]

plot(filtered_data$Jan, filtered_data$co2trendjan, xlab= "Anomaly Temperature", ylab="CO2(ppm) T
rend",
     pch = 16, col = "blue", main = "Scatterplot for CO2 emmissions and temperature anomalies")
```



Now we find the Pearson coefficient, using the 'cor' function of R

```
correlation = cor(filtered_data$Jan, filtered_data$co2trendjan,
                  method = "pearson")

print(correlation)
```

```
## [1] 0.829783
```

The **(Pearson)** correlation coefficient is 0.82, indicating a strong positive linear association between the two variables. When CO2 levels increase, temperatures increase.