Data Analyst Nanodegree Program
Project Report: Exploring Weather Trends

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

OUTLINE

1. Purpose of the study

In this project, the temperature trends in global and in Istanbul were examined and compared so as to gain a knowledge from the given data.

2. The data

The data was extracted by using SQL queries as given in table 2.1.

Table 2.1: Used SQL codes to extract given datasets.

Data	SQL Code
city_data (Istanbul)	SELECT year,avg_temp FROM city_data WHERE city='Istanbul' ORDER BY year;
city_data (Oslo)	SELECT year,avg_temp FROM city_data WHERE city=‘Oslo' ORDER BY year;
city_data (Sydney)	SELECT year,avg_temp FROM city_data WHERE city=‘Sydney' ORDER BY year;
city_data (Rio de Janeiro)	SELECT year,avg_temp FROM city_data WHERE city=‘Rio De Janeiro' ORDER BY year;
city_data (Toronto)	SELECT year,avg_temp FROM city_data WHERE city=‘Toronto' ORDER BY year;
city_list	SELECT * FROM city_list
global_data	SELECT * FROM global_data

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: uuresin@ford.com.tr
Report Date: 2019/07/17

Table 2.2: The size of the given datasets.

Data	Number of Variables	Number of Observations
city_level (Istanbul)	2	271
city_level (Oslo)	2	271
city_level (Sydney)	2	173
city_level (Rio De Janeiro)	2	182
city_level (Toronto)	2	271
city_list	2	345
global_data	2	266

Datasets were downloaded in csv format in a local drive. The size of the datasets are given in table 2.2. It can be noticed from table 2.2, number of observations vary according to city. The reason behind this is that some of the average temperature values are missing. Normally, there are some methods to handle missing data (such as mean/median imputation). However, this is not the subject of this project, missing values are ignored and no imputation method is used. Further details are given in next section.

3. The method and tools

Since the ultimate purpose of this study is to gain knowledge and a conclusion based on the data, visualization and fundamental statistics were used. All datasets were processed in R environment.

4. Analysis

In the given datasets, the ‘city_list’ contains a list of cities and countries in the database. In this study, ‘Istanbul’ was chosen in order to compare with the global temperature trends (and also other 4 cities are also selected but the details are mentioned in next pages). The ‘global_data’ contains the average global temperatures (ºC) by year.

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: uuresin@ford.com.tr
Report Date: 2019/07/17

The 'city_data' contains the average temperatures (ºC) for each city by year. Instead of entire 'city_data', city_level subdata (Istanbul, Oslo etc.) were used.

In figure 4.1, the average global temperature (ºC) by year is shown.

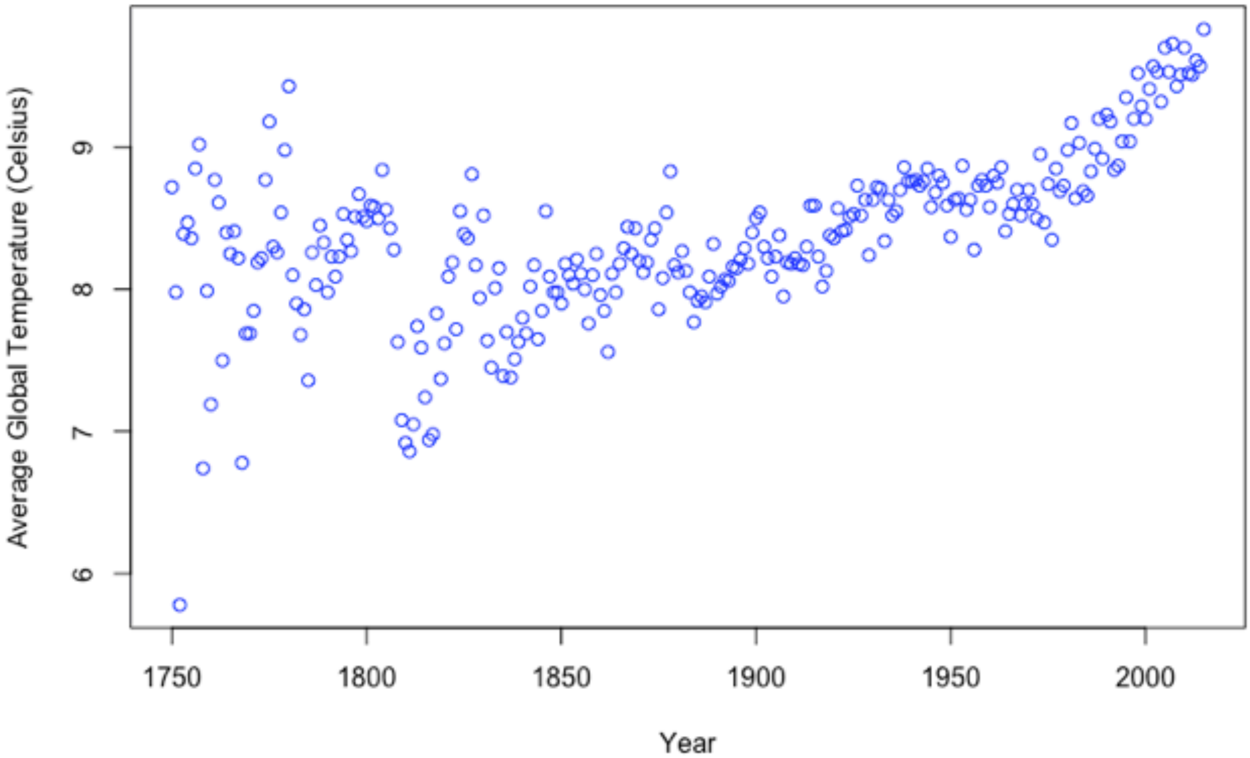

Figure 4.1: Average temperature(ºC) in global over years

It can be visually inferred that a trend of increment in average global temperatures has started in 1850 and this increment has gained an acceleration since 1950.

In figure 4.2, a visual comparison (scatter plot) of the global average temperatures and the average temperatures of the city Istanbul is given. In figure 4.2, it can be clearly seen that the average temperature of Istanbul is much higher than the global average temperature. Table 4.1 presents a descriptive summary of the average temperatures.

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

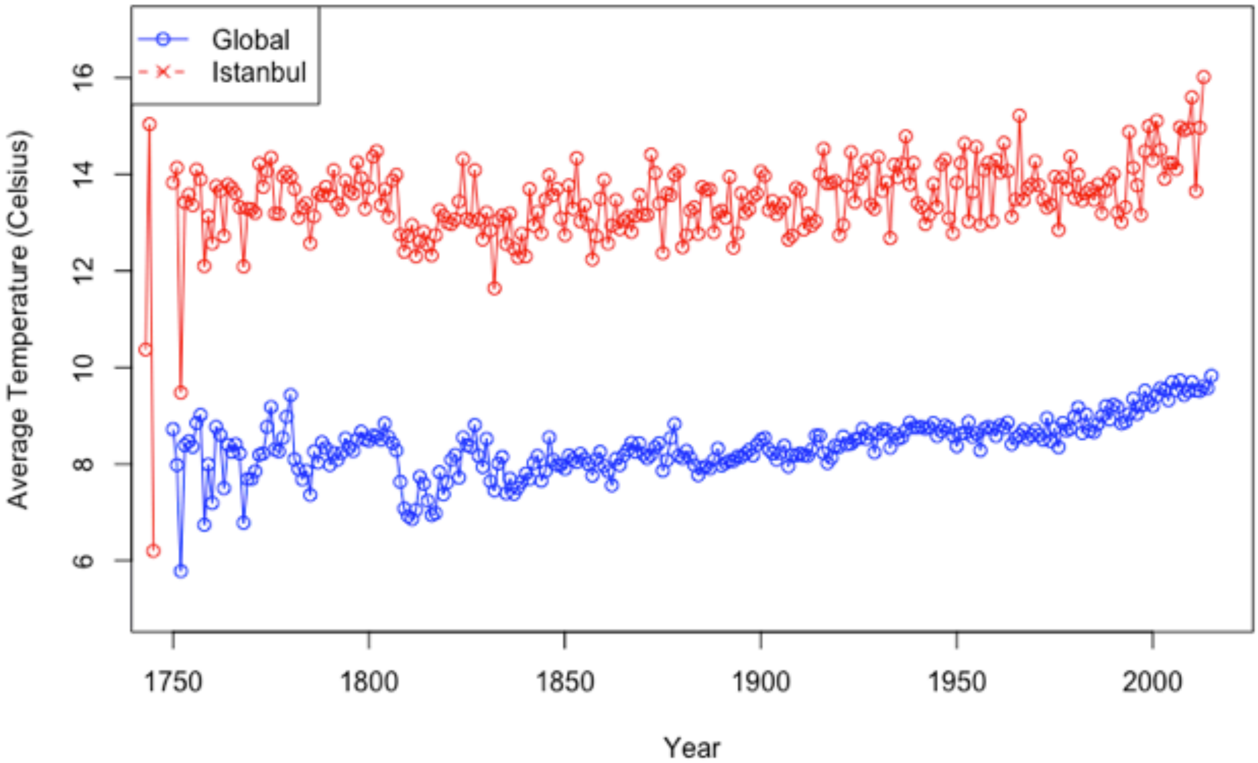

Figure 4.2: Average temperatures(ºC) in global and in Istanbul over years

Moving average is a calculation for analyzing data -especially time series data- by forming a series of averages (means) of different subsets of the complete dataset. As a simple explanation, the moving average is the replacement of each point for the ordered data points by replacing the average of the n points preceding it. Since the first n-1 number is not preceded by n points (fewer than n), those points can be omitted or the average of the number of points can be written regardless number of the previous data point for those points. In this study, the second approach is used here. In this study, moving averages are calculated by using R (codes are given in 'Appendices' section).

Moving averages by the step sizes 10 years and 20 years reflected onto the scatter plot and shown in the figure 4.3 and figure 4.4 respectively.

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

Table 4.1: The average temperatures(ºC) in global and in Istanbul.
(*Note that average temperatures for Istanbul are available up to 2013)

	Min Average Temperature (ºC)			Max Average Temperature (ºC)		
	[1750-1963]	[1964-2013*]	%Change	[1750-1963]	[1964-2013*]	%Change
Global	5.78	8.35	44.46	9.43	9.73	3.08
Istanbul	6.2	12.85	51.75	15.04	16.02	0.98

As shown in the table 4.1, minimum average temperatures has been raised globally and in Istanbul in last 50 years. Comparing the changes in the minimums of the average temperatures and the maximums of the average temperatures is a naive approach due to the fact that the variance and the distribution of the average temperature values can vary.

Also, as shown in figure 4.2, both trends are fluctuating. The meaning behind the fluctuation can be explain several physical phenomena. For a robust conclusion, moving average approach was used.

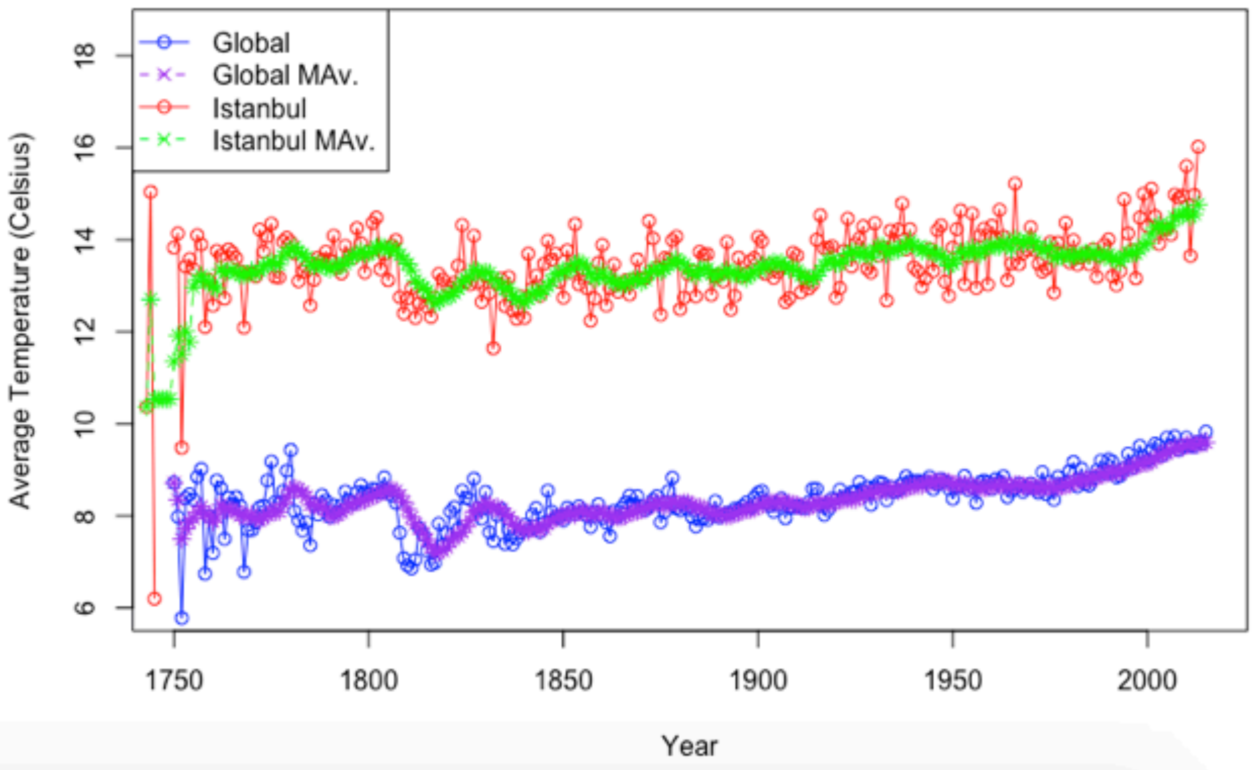

Figure 4.3: Average temperatures(ºC) in global and in Istanbul over years and moving averages with 10 years step size.

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: uuresin@ford.com.tr
Report Date: 2019/07/17

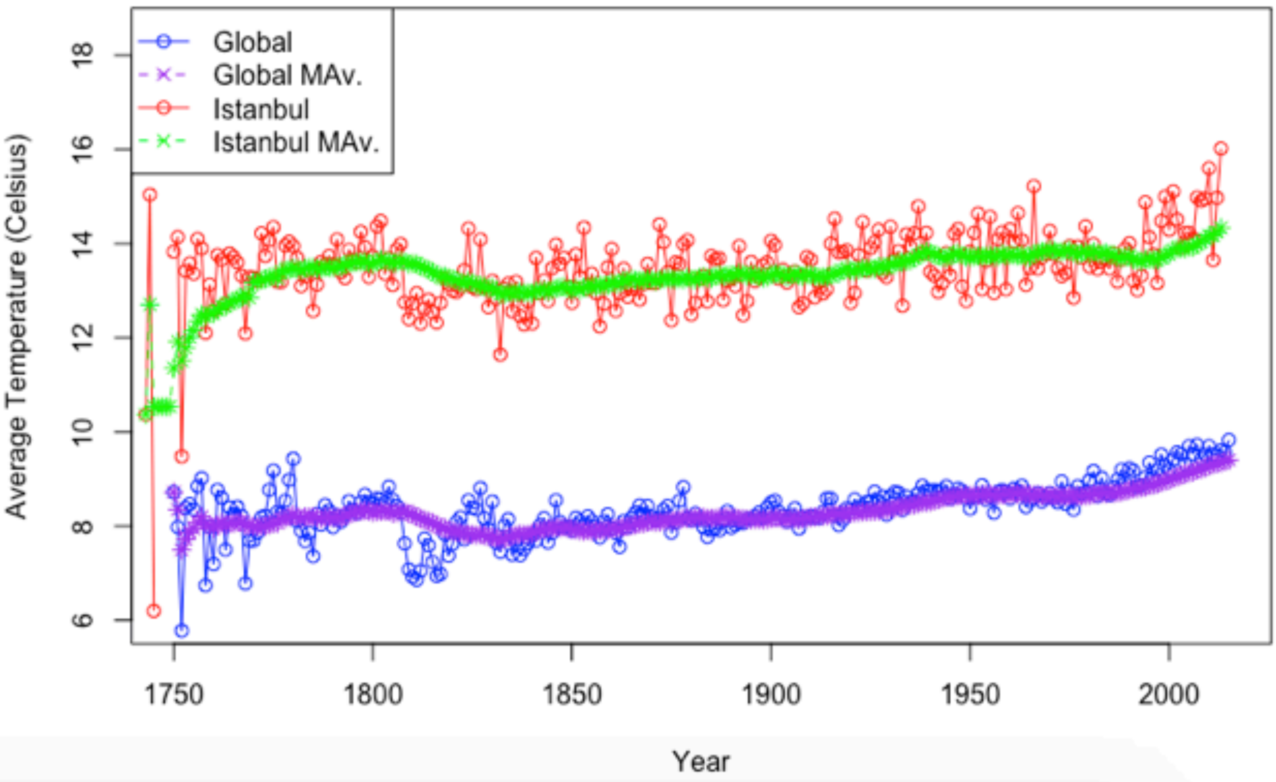

Figure 4.4: Average temperatures(ºC) in global and in Istanbul over years and moving averages with 20 years step size.

From figure 4.3 and figure 4.4, it can be clearly inferred that there has been an increase in global average temperatures, especially since 1950. The same upward trend is observed in Istanbul, and the increase has gained momentum especially since the beginning of 2000.

In this study, Istanbul was initially chosen randomly. Istanbul has a special geographical position because it is close to the equator line and is surrounded by seas. Therefore, four other cities that differ from each other in terms of geography are also examined.

Moving averages of temperatures of 5 cities (the cities are shown in figure 4.5) in different geographic locations and global temperature values are calculated and shown together in figure 4.6.

In figure 4.6, although average temperatures levels vary from city to city, it can be seen that average temperature trends are similar and tend to increase especially in recent years.

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

Figure 4.5: Selected 5 geographically different cities in this study (1:Istanbul, 2:Oslo, 3:Sydney, 4:Rio de Janeiro, 5:Toronto)

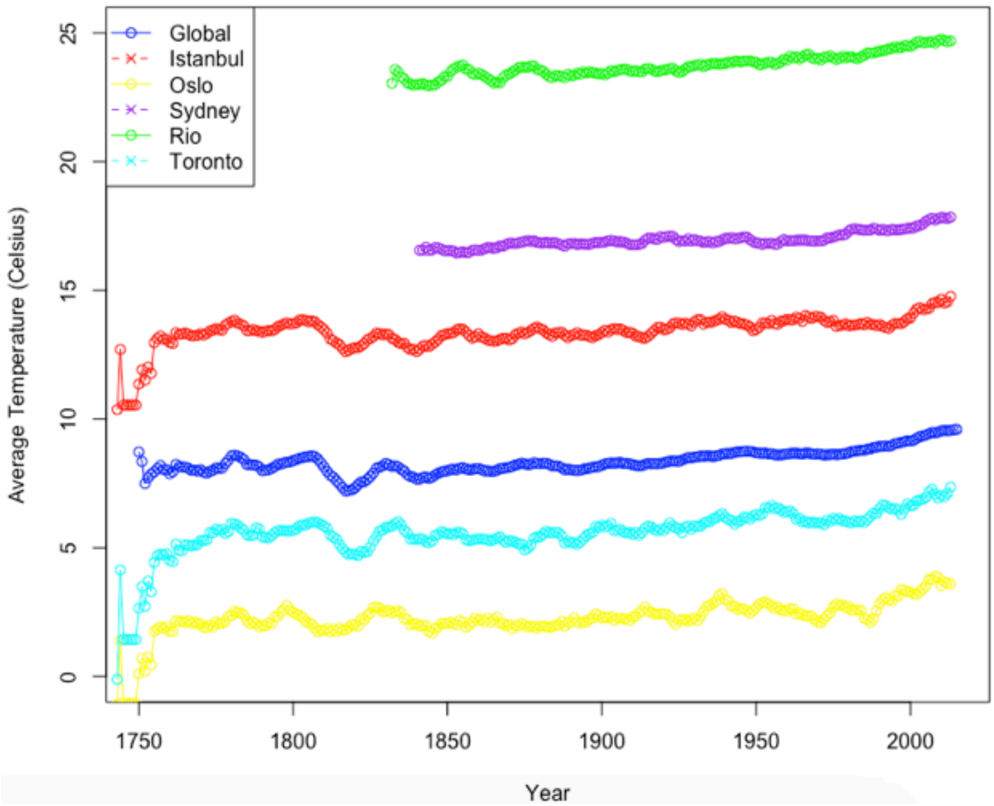

Figure 4.6: Average temperatures(ºC) in global and in 5 cities over years calculated as moving averages with 10 years step size.

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: uuresin@ford.com.tr
Report Date: 2019/07/17

To understand how the average temperature behaviors in Istanbul behave similar to the global temperature averages, a scatter diagram that contains global average temperature values against average temperature values for istanbul in the years between 1750 and 2013 is plotted and shown in figure 4.7.

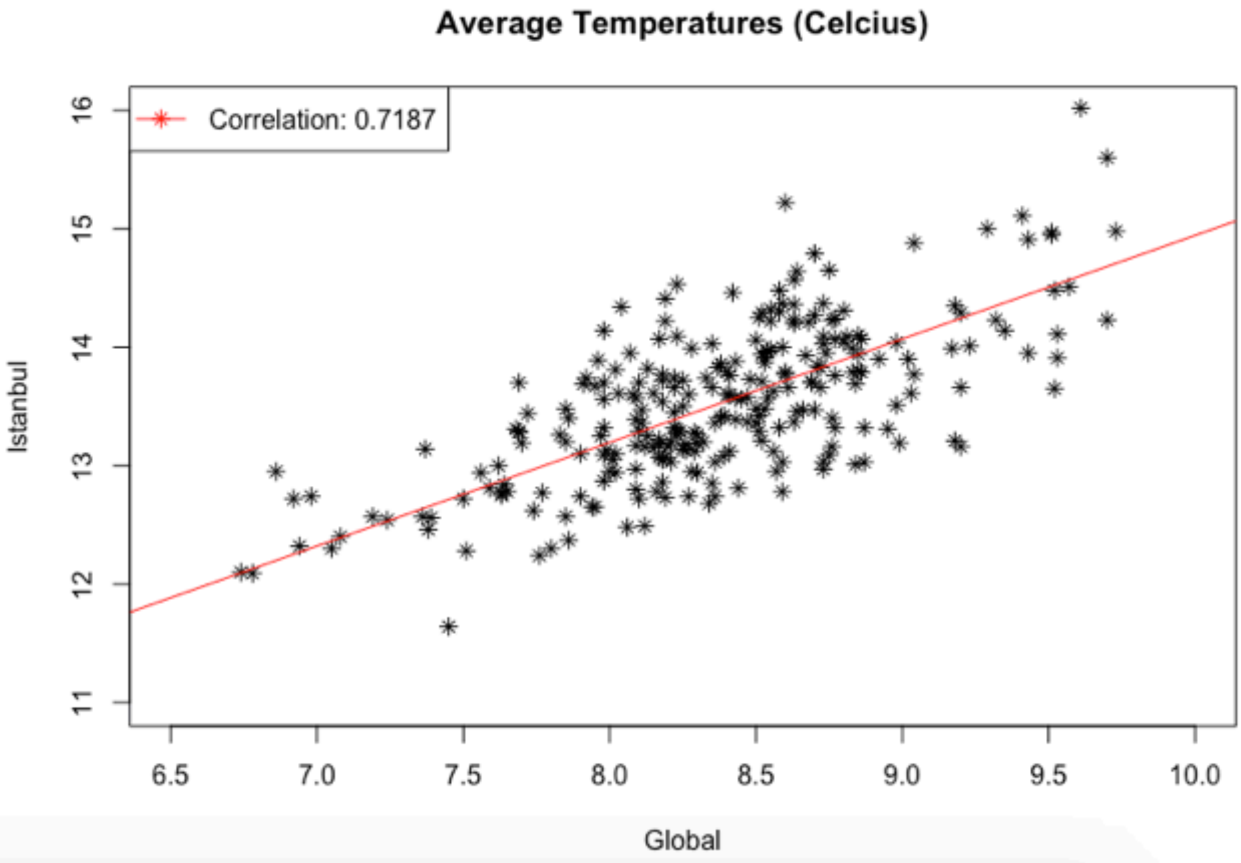

Figure 4.7: Average temperatures(ºC) values in global and in Istanbul in the period of time 1750-2013.

Also, a regression line is fitted to the diagram and a strong positive correlation was found with a correlation coefficient 0.7187. It can also be inferred that the average temperature in Istanbul (or other cities) can be predicted approximately based on the global average temperature. Due to the fact that trends are similar, a linear regression model can be used for prediction. Since, the purpose of this study is not building a prediction model, it’s not added to this report.

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

5. Conclusion

5 selected cities (Istanbul, Oslo, Sydney, Rio de Janeiro, Toronto) show similar temperature trends inline with global temperature levels. Even if there has been fluctuations over time, average temperatures are tend to increase over time especially last couple decades. This can be interpret as a result of global warming but this is just an intuition. Any other phenomena can be the root cause!

As a conclusion, global average temperatures increases and the this trend become more significant in the recent years.

6. References

Analyst (Uğur ÜRESİN, PhD. student in the field of data analytics in Istanbul Technical University) benefited from his own experience and knowledge.

Thus, no reference is given.

7. Appendices

Analyst shares his R codes in next pages that he wrote for this study.

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

```
#setting the working directory
setwd("/Users/ugur/Desktop/mycodes/myprojects/training/Udacity_DAND/project1/")

#data import
data_global = read.csv("data/global_data.csv")
data_city = read.csv("data/city_data.csv")
data_citynames = read.csv("data/city_list.csv")
data_istanbul = read.csv("data/istanbul.csv")
data_oslo = read.csv("data/oslo.csv")
data_sydney = read.csv("data/sydney.csv")
data_rio = read.csv("data/rio.csv")
data_toronto = read.csv("data/toronto.csv")

#initials
describe(data_global)
describe(data_city)
describe(data_citynames)
#
summary(data_global)
summary(data_city)
summary(data_citynames)

#scatter plot - average global temperature
(plot_avglobal <- plot(data_global, type="p", col="blue",
     xlab="Year",
     ylab="Average Global Temperature (Celsius)", col.lab="black"))
#
#scatter plot - average global temperature
(plot_avglobal2 <- plot(data_global[150:175,], type="p", col="blue",
                        xlab="Year",
                                   ylab="Average  Global  Temperature  (Celsius)",
col.lab="black"))

##comparison of the average temperatures: global vs. istanbul
#setting axis limits
max_x1 <- round(max(max(data_istanbul[,2], na.rm = TRUE),
                    max(data_global$avg_temp, na.rm = TRUE)))+1
min_x1 <- round(min(min(data_istanbul[,2], na.rm = TRUE),
                    min(data_global$avg_temp, na.rm = TRUE)))-1
#plotting
(plot_avglobal <- plot(data_global, type="o", col="blue",
                       xlab="Year", ylab="Average Temperature (Celsius)",
                       col.lab="black", ylim=c(min_x1,max_x1),
                       pch=1, lty=1))
```

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

```
lines(data_istanbul, type="o", col="red",pch=1, lty=1)
legend('topleft',c("Global","Istanbul"),    col=c("blue","red"),    pch=c(1,4),
lty=c(1,2))

#max-min average values
min(data_global$avg_temp[1:214], na.rm=TRUE)
max(data_global$avg_temp[1:214], na.rm=TRUE)
min(data_istanbul[1:221,2], na.rm=TRUE)
max(data_istanbul[1:221,2], na.rm=TRUE)
#
min(data_global[215:264,]$avg_temp, na.rm=T)
max(data_global[215:264,]$avg_temp, na.rm=T)
min(data_istanbul[222:271,2], na.rm=T)
max(data_istanbul[222:271,2], na.rm=T)

#moving average function
#REFERENCE for this function is given below!!!
#http://www.cookbook-r.com/Manipulating_data/Calculating_a_moving_average/)
mav <- function(x, n=10, centered=FALSE) {
  if (centered) {
    before <- floor  ((n-1)/2)
    after  <- ceiling((n-1)/2)
  } else {
    before <- n-1
    after  <- 0
  }
  # Track the sum and count of number of non-NA items
  s     <- rep(0, length(x))
  count <- rep(0, length(x))
  # Add the centered data
  new <- x
  # Add to count list wherever there isn't a
  count <- count + !is.na(new)
  # Now replace NA_s with 0_s and add to total
  new[is.na(new)] <- 0
  s <- s + new
  # Add the data from before
  i <- 1
  while (i <= before) {
    # This is the vector with offset values to add
    new   <- c(rep(NA, i), x[1:(length(x)-i)])
    count <- count + !is.na(new)
    new[is.na(new)] <- 0
    s <- s + new
```

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

```
    i <- i+1
  }
  # Add the data from after
  i <- 1
  while (i <= after) {
    # This is the vector with offset values to add
    new   <- c(x[(i+1):length(x)], rep(NA, i))
    count <- count + !is.na(new)
    new[is.na(new)] <- 0
    s <- s + new
    i <- i+1
  }
  s/count# return sum divided by count
}
#end of the mav function

#moving averages of the data
mav_global <- cbind(data_global$year, mav(data_global$avg_temp))
mav_istanbul <- cbind(data_istanbul[,1], mav(data_istanbul[,2]))

#plotting of moving averages
max_x1 <- round(max(max(data_istanbul[,2], na.rm = TRUE),
                    max(data_global$avg_temp, na.rm = TRUE)))+2.5
min_x1 <- 6
#
(plot_avglobal <- plot(data_global, type="o", col="blue",
                       xlab="Year", ylab="Average Temperature (Celsius)",
                       col.lab="black", ylim=c(min_x1,max_x1),
                       pch=1, lty=1))
lines(mav_global, type="o", col="purple",pch=8, lty=2)
lines(data_istanbul, type="o", col="red",pch=1, lty=1)
lines(mav_istanbul, type="o", col="green",pch=8, lty=2)
legend('topleft',c("Global","Global MAv.","Istanbul","Istanbul MAv."),
       col=c("blue","purple","red","green"), pch=c(1,4), lty=c(1,2))
#
limits <- cbind(max(data_oslo[,2], na.rm=T),
           min(data_oslo[,2], na.rm=T),
           max(data_sydney[,2], na.rm=T),
           min(data_sydney[,2], na.rm=T),
           max(data_rio[,2], na.rm=T),
           min(data_rio[,2], na.rm=T),
           max(data_toronto[,2], na.rm=T),
           min(data_toronto[,2], na.rm=T))
#
#
```

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

```
mav_oslo <- cbind(data_oslo[,1], mav(data_oslo[,2]))
mav_sydney <- cbind(data_sydney[,1], mav(data_sydney[,2]))
mav_rio <- cbind(data_rio[,1], mav(data_rio[,2]))
mav_toronto <- cbind(data_toronto[,1], mav(data_toronto[,2]))

#plotting of moving averages: 5 cities & global
max_x1 <- round(max(limits))
min_x1 <- round(min(limits))+6
#
plot(mav_global, type="o", col="blue",
     xlab="Year", ylab="Average Temperature (Celsius)",
     col.lab="black", ylim=c(min_x1,max_x1),
     pch=1, lty=1)
lines(mav_istanbul, type="o", col="red",pch=1, lty=1)
lines(mav_oslo, type="o", col="yellow",pch=1, lty=1)
lines(mav_sydney, type="o", col="purple",pch=1, lty=1)
lines(mav_rio, type="o", col="green",pch=1, lty=1)
lines(mav_toronto, type="o", col="cyan",pch=1, lty=1)
legend('topleft',c("Global","Istanbul","Oslo","Sydney","Rio","Toronto"),
              col=c("blue","red","yellow","purple","green","cyan"),  pch=c(1,4),
lty=c(1,2))
#
m=10
plot(tail(mav_global,m), type="o", col="blue",
     xlab="Year", ylab="Average Temperature (Celsius)",
     col.lab="black", ylim=c(min_x1,max_x1),
     pch=1, lty=1)
lines(tail(mav_istanbul,m), type="o", col="red",pch=1, lty=1)
lines(tail(mav_oslo,m), type="o", col="yellow",pch=1, lty=1)
lines(tail(mav_sydney,m), type="o", col="purple",pch=1, lty=1)
lines(tail(mav_rio,m), type="o", col="green",pch=1, lty=1)
lines(tail(mav_toronto,m), type="o", col="cyan",pch=1, lty=1)
legend('topleft',c("Global","Istanbul","Oslo","Sydney","Rio","Toronto"),
              col=c("blue","red","yellow","purple","green","cyan"),  pch=c(1,4),
lty=c(1,2))

#correlations
data_glo_ist                                                                  <-
as.data.frame(cbind(data_global[1:264,2],data_istanbul[8:271,2]))
cor(data_global[1:264,2],data_istanbul[8:271,2])
plot(data_glo_ist, xlim=c(6.5,10), ylim=c(11,16), pch=8,
     xlab="Global", ylab="Istanbul", main="Average Temperatures (Celcius)")
abline(lm(V2~., data=data_glo_ist),col="red")
legend('topleft',c("Correlation: 0.7187"),
       col=c("red"), pch=8, lty=1)
```

Prepared by: Uğur ÜRESİN
(PhD. Student in Data Analytics, ITU)
E-mail: **uuresin@ford.com.tr**
Report Date: 2019/07/17

```
#end of the code
```