Exploring Weather Trends – Project

- Mohit Patel

Extract the data :
Write a SQL query to extract the city level data. Export to CSV.
Write a SQL query to extract the global data. Export to CSV.
Solution:

Query to view details of city of residence:

```
SELECT *
FROM city_list
WHERE city LIKE 'Aha%' AND country LIKE 'IN%';
```

Query to extract city level data:

```
SELECT *
FROM city_data
WHERE city = 'Ahamedabad' AND country = 'INDIA';
```

Query to extract global data:

```
SELECT *
FROM global_data;
```

Open up the CSV :
First step was to download data based on filtering queries on city_data and global_data databases in CSV format in local computer. [as above]

Create a line chart:
Since moving averages are to be used, initial step was to calculate moving average for five and ten years. These were plotted in the previously downloaded CSV files using the formula:
=AVERAGE(Cell2:Cell6)
and
=AVERAGE(Cell2:Cell11)
respectively.
These values were saved in columns named ‘FIVEYEARMA’ and ‘TenYearMA’ in both CSV files: global data and city data for all the data points extracted. Care was taken to keep column names the same in both files.

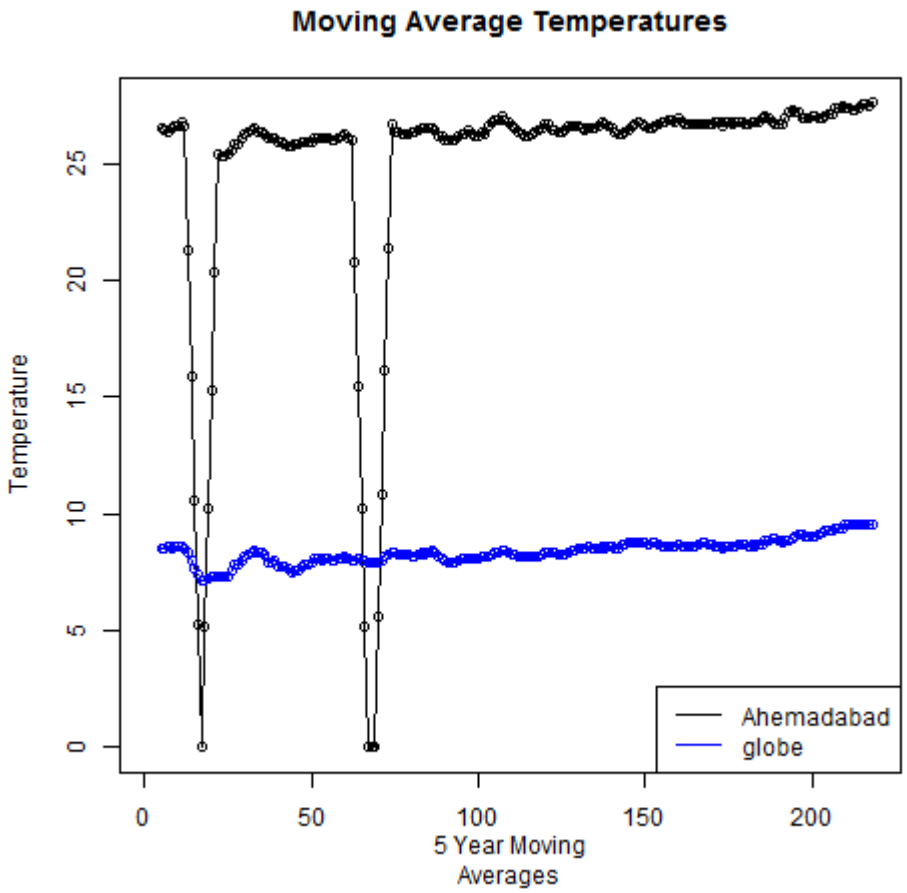
To create the line charts, R programming was used.
Below is the R program:

```
#Load Packages
install.packages('ggplot2')
library(ggplot2)
setwd('//Users/lenovo/Documents/Udacity/Project1/We
atherProject')
getwd()


#Read Ahamedabad city temperature averages
cityInfo <- read.csv('Ahamedabad.csv')
View(cityInfo)


#Read global temperature averages results
globalInfo <- read.csv('global.csv')
View(globalInfo)

# Create the data for the chart
world <- c(globalInfo$FIVEYEARMA)
Ahmcity<- c(cityInfo$fiveyearma)

#assign colors- blue for global temperatures andBlackfor Ahamedabad
# as observed in CSV that global temperatures hotter than
Ahamedabad temperatures
# following standard color conventions:Black= hot, blue = cold

# Give the chart file a name.
png(file = "AvgTempGlobalvsAhamedabad.jpg")

# Plot the bar chart.
plot(Ahmcity,type = "o",col = "black", xlab = "5 Year Moving
Averages", ylab = "Temperature",
main = "Moving Average Temperatures")
lines(world, type = "o", col = "blue")

# Add a legend
legend('bottomright', legend=c("Ahamedabad", "World"),
lty=1,
col=c("black", "blue"))
dev.off()
```

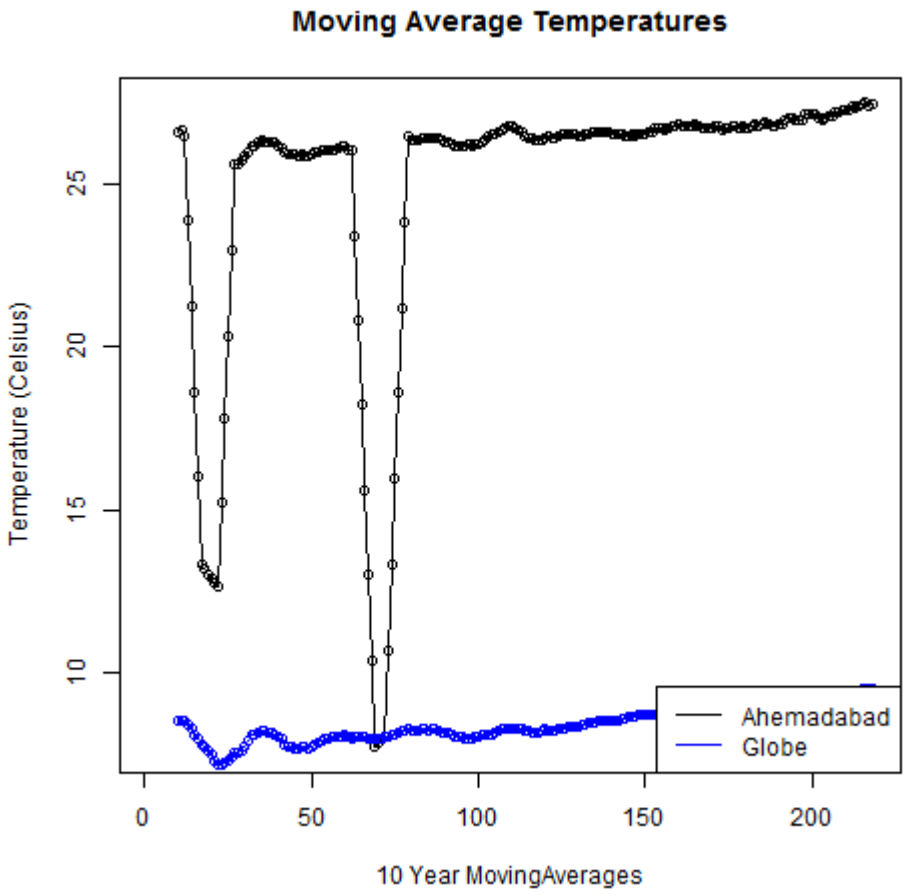
```
# Create the data -10yrMA for the chart

world10 <- c(globalInfo$TENYEARMA)
Ahmcity10<- c(cityInfo$tenyearma)

#assign colors- blue for global temperatures andBlackfor Ahamedabad

# as observed in CSV that global temperatures hotter than
Ahamedabad temperatures

# following standard color conventions: Black= high(hot), blue = low(cold)

# Give the chart file a name.
png(file = "AvgTempGlobalvsAhamedabad10YrMALeg.jpg")

# Plot the bar chart.

plot(Ahmcity10,type = "o",col = "black", xlab = "10 Year Moving
Averages", ylab = "Temperature (Celsius)",
main = "Moving Average Temperatures")

# Add a legend

legend('bottomright', legend=c("Ahemadabad", "World"),
lty=1,
col=c("black", "blue"))

# Save the file.
dev.off()
```

Explanation:
R programming was used to create the plot. The data frame values for moving averages are read into vector which are then plotted. Since, temperature values in CSV were observed to be higher for Ahmadabad than the world average, Black color was assigned to Ahmadabad and blue for the Global values. The convention is listed in the legend.

Line charts are shown here:
5 Year Moving Averages



10 Year Moving Averages



Observations of local v/s global temperature:

1. Ahemadabad city observed to have temperature greater than the global average. we can see in both plots - five and ten year Moving Averages.

2. Changes in city's temperature in comparison with global temperatures look very high. In a year by year comparison, or Moving Average, when we perform point by point comparison, but they data points appear to be moving in similar pattern.

3. Both Ahemadabad city and global temperatures show an upward trend. While lows and highs are clearly visible, the graph is clearly rising in both datasets. The trend has been consistent for the past few decades.

4. In the initial period, the Ahemadabad city data shows extreme jumps in temperature while the global temperatures rise slower. This point of disparity is worth looking into. While a lack of data is noted for a few years causing a lower average, year wise temperature deviations necessitate further scrutiny.

Considerations when deciding to visualize the trends:

i. Line chart was specified in the question. Hence, it was plotted using function in R programming.

ii. In order to stay consistent with general accepted conventions, Ahemadabad city [more hotter than global average] was assigned colour black while global data was assigned blue [for colder]. Legend is added to show which colored line is for which data.

iii. Data points were plotted in conjunction with the line plots to facilitate point-by-point comparison.

iv. Celsius was mentioned as the unit of temperature in Slack forum, and hence is used in the Y-axis label. X-axis is labelled to denote period-wise Moving Average: 5 or 10 years.

v. Key criterion while plotting the graph was observing all the values. Hence, values from Ahemadabad city dataset were considered as first dataset and global as second, since the plot function in reverse order cropped off end values unless scaled.

vi. Also, program creates and saves plot as png files, to enable reproducible results and analyses. This was a

secondary criterion while making the plots.
Additional Insights:

vii. What's the correlation coefficient?

The correlation coefficient for global and Ahemadabad temperatures is **0.4113259** using Pearson's method. Using Kendall's method, we get: **0.6178749** and using Spearman's method: **0.7939483**.

For moving averages of five years, we get correlation **coefficient** as:
Kendall's 0.4113259
Pearson's 0.6178749
Spearman's 0.7939483

Thus, there seems to be a definite correlation.
In order to compute the correlation, the data from both datasets; global and Ahemadabad city were added to a new CSV file. Matching was done along Year column and rows with no matching data were discarded. The final file contained values from years 1796-2013.
Moving averages for five years was also calculated using formula in this Excel workbook.
The correlation coefficients were calculated using R programming. Code is included below:

```
#Load Packages
install.packages('ggplot2')
library(ggplot2)

#Read temperatures for correlation coefficient into
dataframe
tempInfo <- read.csv('Globalvsahm.csv')
View(tempInfo)

#add temperature values into variables
globaltemp <- tempInfo$avg_temp_global

 #global temp
Ahmtemp <- tempInfo$avg_temp_Ahm

 #Ahemadabad city temp
```

```
#plot to visually view if relation exists between global and
local temperatures

ggplot(data = tempInfo, mapping = aes(x= globaltemp, y
=Ahmtemp))+geom_point()

#computing correlation coefficients
?cor.test()

cor.test(globaltemp,Ahmtemp)

#kendall
cor.test(globaltemp,Ahmtemp, method = 'pearson')

# pearson
cor.test(globaltemp,Ahmtemp, method = 'kendall')

# spearman
cor.test(globaltemp,Ahmtemp, method = 'spearman')

#compute correlation for five year Moving Averages

globaltemp_MA <- tempInfo$g_MA

Ahmtemp_MA <- tempInfo$C_MA
#kendall
cor.test(globaltemp_MA,Ahmtemp_MA, method = 'kendall')

# pearson
cor.test(globaltemp_MA,Ahmtemp_MA, method = 'pearson')

# Spearman
cor.test(globaltemp_MA, Ahmtemp_MA, method ='spearman')
```