# Exploring Weather Trends - Project

-Nirupama Puthur Venkataraman

# 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 'Chi%' AND country LIKE 'United%';
```

Query to extract city level data:

```
SELECT *
FROM city_data
WHERE city = 'Chicago' AND country = 'United States';
```

Query to extract global data:

```
SELECT *
  FROM global_data;
```

*Explanation: SQL queries were used to extract data in this phase.*

Phase 1- Obtain city of interest

View table structure:

```
SELECT *
FROM city_list
LIMIT 5;
```

Search for city of residence- Arlington Heights, USA

```
SELECT *
FROM city_list
WHERE city LIKE 'Arl%' AND country LIKE 'United%';
```

Results obtained for Arlington, not Arlington Heights; hence, altered query to search for Chicago city.

```
SELECT *
FROM city_list
WHERE city LIKE 'Chi%' AND country LIKE 'United%';
```

Result was positive, hence, city decided on Chicago in United States.

Phase 2: extract city data:
View table details:

```
SELECT *
FROM city_data
LIMIT 5;
```

```
SELECT *
FROM city_data
WHERE city = 'Chicago' AND country = 'United States';
```

Phase 3: Extract global data:
View database structure and details:

```
SELECT *
FROM global_data
LIMIT 5;
```

Query to extract global temperature 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. [Queries 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/nirupamaprv/Documents/nimmu/Nanodegree/WeatherProject')
getwd()

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

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

```
View(globalInfo)


# Create the data for the chart
world <- c(globalInfo$FiveYearMA)
chiTown<- c(cityInfo$FiveYearMA)


#assign colors- blue for global temperatures and red for Chicago
# as observed in CSV that global temperatures hotter than
Chicago temperatures
# following standard color conventions: red = hot, blue = cold
# Give the chart file a name.
png(file = "AvgTempGlobalvsChicago.jpg")

# Plot the bar chart.
plot(chiTown,type = "o",col = "red", xlab = "5 Year Moving
Averages", ylab = "Temperature",
    main = "Moving Average Temperatures")

lines(world, type = "o", col = "blue")
# Add a legend
legend('bottomright', legend=c("Chicago", "World"),
       lty=1,
       col=c("red", "blue"))
# Save the file.
dev.off()

# Create the data -10yrMA for the chart
world10 <- c(globalInfo$TenYearMA)
chiTown10<- c(cityInfo$TenYearMA)


#assign colors- blue for global temperatures and red for Chicago
# as observed in CSV that global temperatures hotter than
```

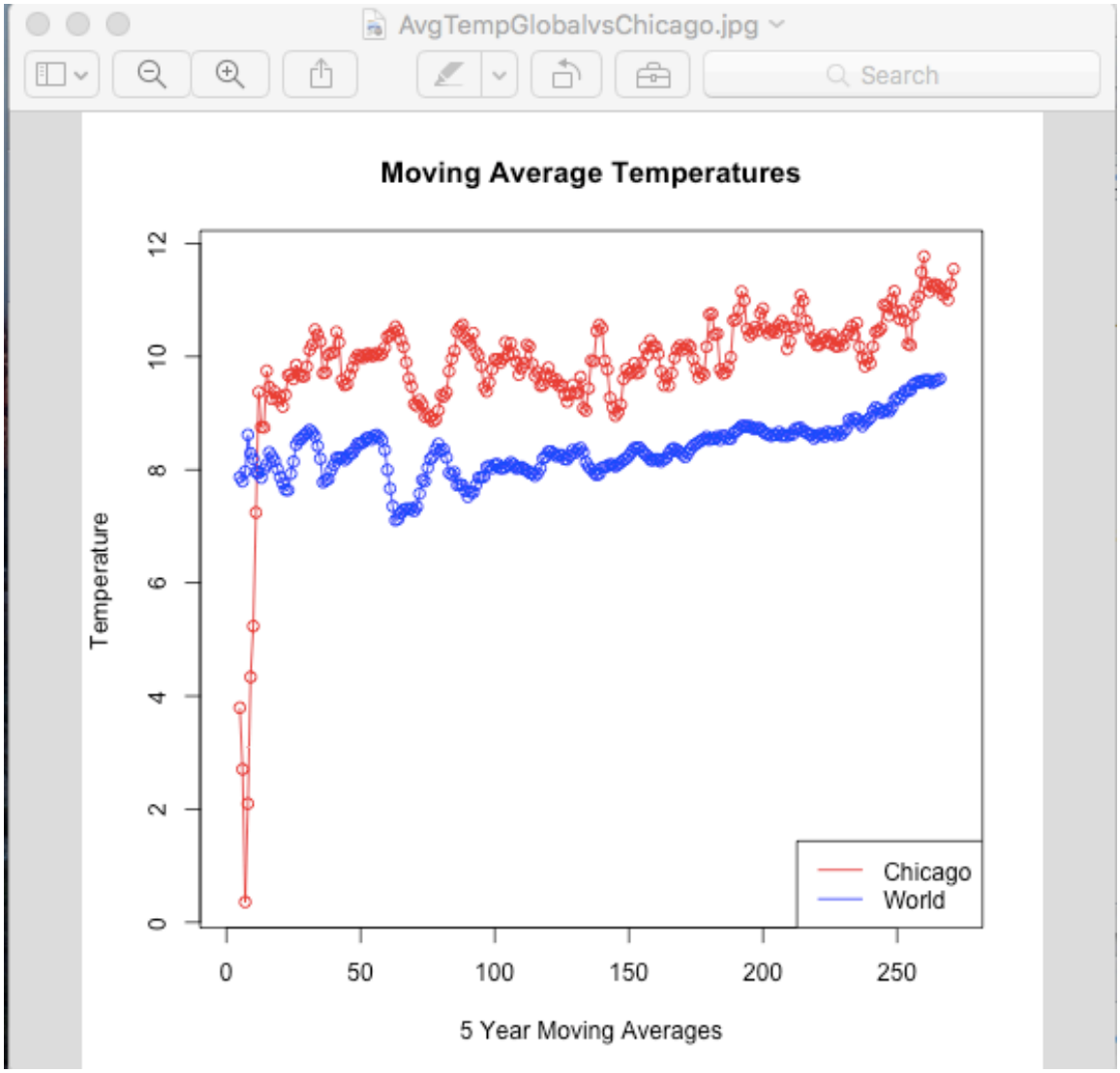
```
Chicago temperatures
# following standard color conventions: red = hot, blue = cold
# Give the chart file a name.
png(file = "AvgTempGlobalvsChicago10YrMALeg.jpg")

# Plot the bar chart.
plot(chiTown10,type = "o",col = "red", xlab = "10 Year Moving
Averages", ylab = "Temperature (Celsius)",
    main = "Moving Average Temperatures")
# Add a legend
legend('bottomright', legend=c("Chicago", "World"),
       lty=1,
       col=c("red", "blue"))
# Save the file.
dev.off()
```

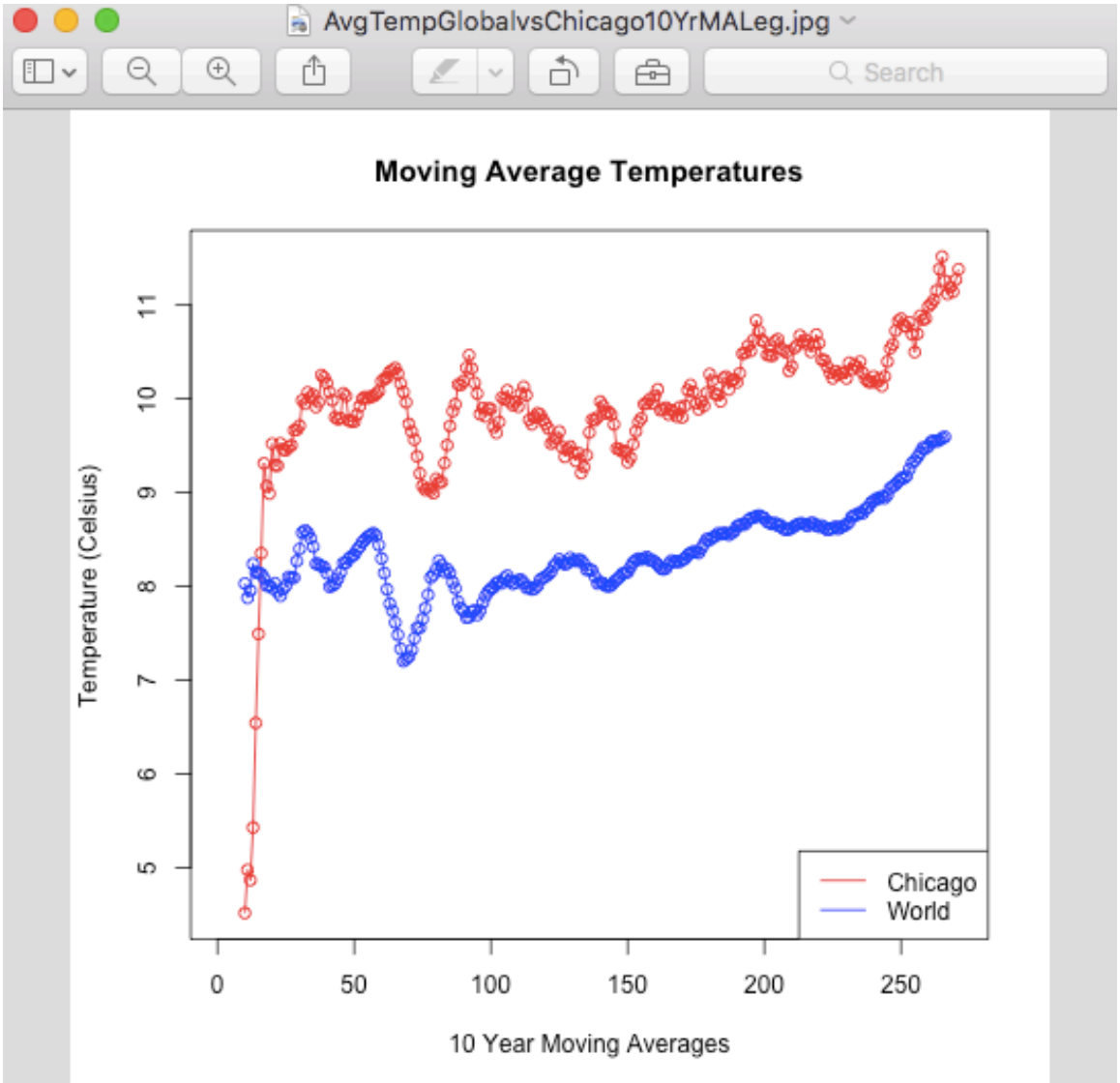
---

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 Chicago than the world average, Red color was assigned to Chicago 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:

1. Chicago city observed to have temperature greater than the global average. This is true in both plots - five and ten year Moving Averages.
2. Changes in city's temperature in comparison with global temperatures look related. In a year by year comparison, or Moving Average, when we perform point by point comparison, the highs and lows seem to coincide, but they data points appear to be moving in similar pattern.
3. Both Chicago 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 local (Chicago 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.

Key Considerations when deciding to visualize the trends:

- Line chart was specified in the question. Hence, it was plotted using function in R programming.
- In order to stay consistent with general accepted conventions, Chicago city [hotter than global average] was assigned color red while global data was assigned blue [for colder]. Legend is added to show which colored line is for which data.
- Data points were plotted in conjunction with the line plots to facilitate point-by-point comparison.
- Celsius was mentioned as the unit of temperature in Slack forum, and hence is used in the Y-axis label. X-axis is labeled to denote period-wise Moving Average: 5 or 10 years.
- Key criterion while plotting the graph was observing all the values. Hence, values from Chicago city dataset were considered as first dataset and global as second, since the plot function in reverse order cropped off end values unless scaled.
- 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:

- What's the correlation coefficient?

The correlation coefficient for global and Chicago

temperatures is 0.6691357 using Pearson's method. Using Kendall's method, we get: 0.442683 and using Spearman's method: 0.6154788.

For moving averages of five years, we get correlation coefficient as:

| | |
|---|---|
| Kendall's | 0.5822567 |
| Pearson's | 0.8054953 |
| Spearman's | 0.7784913 |

Thus, there seems to be a definite correlation.

In order to compute the correlation, the data from both datasets; global and Chicago 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 1750-2013.
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)

setwd('//Users/nirupamaprv/Documents/nimmu/Nanodegree/WeatherProject')
getwd()

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

```
#add temperature values into variables
globaltemp <- tempInfo$avg_temp_global #global temp
chitemp <- tempInfo$avg_temp_Chi       #Chicago city temp

#plot to visually view if relation exists between global and
local temperatures
ggplot(data = tempInfo, mapping = aes(x= globaltemp, y =
chitemp))+
  geom_point()

#computing correlation coefficients
?cor.test()
cor.test(globaltemp,chitemp)
cor.test(globaltemp,chitemp, method = 'pearson')
cor.test(globaltemp,chitemp, method = 'kendall')
cor.test(globaltemp,chitemp, method = 'spearman')

#compute correlation for five year Moving Averages
globaltemp_MA <- tempInfo$g_MA
chitemp_MA <- tempInfo$C_MA
cor.test(globaltemp_MA,chitemp_MA, method = 'kendall')
cor.test(globaltemp_MA,chitemp_MA, method = 'pearson')
cor.test(globaltemp_MA,chitemp_MA, method =
'spearman')
```