Exploring Weather Trends

Data Analyst Nanodegree - Project 1

Jingjun He

Overview:

The aim of this project is to find out the temperature trends of the city I live (San Jose, USA) and the global temperature trends, and observe any similarities and differences between the two. Two methods were used to analyze local and global data: **Moving averages** and **Regression Analysis**. Instead of using yearly data, moving averages are calculated and visualized to make the overall trend more observable. Regression analysis is used to estimate the relationship between local and global temperatures.

Tools used:

SQL: extract data from the database
EXCEL: 1. Compute moving averages and create line charts
2. Calculate correlation coefficient and draw a scatter diagram
3. Use Data Analysis tool to run regression analysis

Extract the data using SQL:

First, I wrote the SQL query below to check whether the city I live, San Jose USA was in the city_list. The output returns San Jose, USA.

```
SELECT *
FROM city_list
WHERE city = 'San Jose';
```

So, now I can extract city level data by using WHERE clause to filter only those records that meet the criteria.

```
SELECT *
FROM city_data
WHERE city = 'San Jose' and country = 'United
States';
```

In order to make the city and country data more comparable and easier for analysis, I used JOIN clause to combine data from city_data and global_data tables and export the desired results into a

csv file. It is worth noting that since both tables have the same column name avg_temp, alias should be used.

```
SELECT c.year, c.avg_temp AS avg_city_temp,
g.avg_temp AS avg_global_temp
FROM city_data c
JOIN global_data g
ON c.year = g.year
WHERE city = 'San Jose' and country = 'United
States';
```

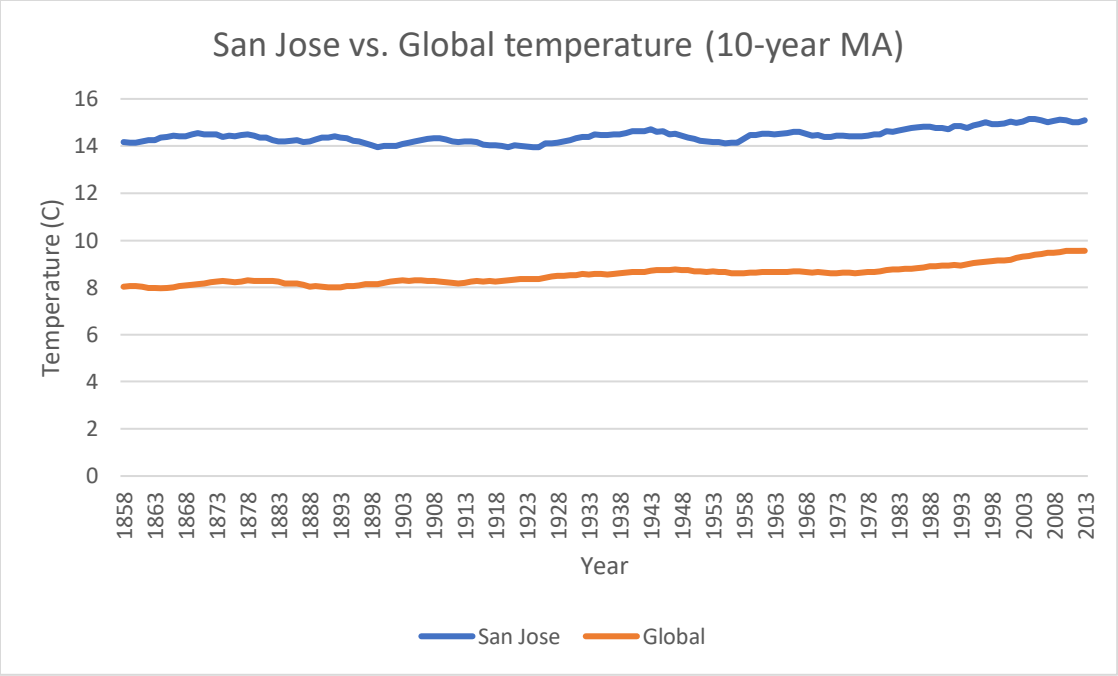
Analyze data using EXCEL: Moving Averages

1. **Compute moving averages**: I computed 10-year moving averages in order to smooth out data and make it easier to observe the long-term trends. The table below shows the EXCEL command to calculate 10-year moving averages.

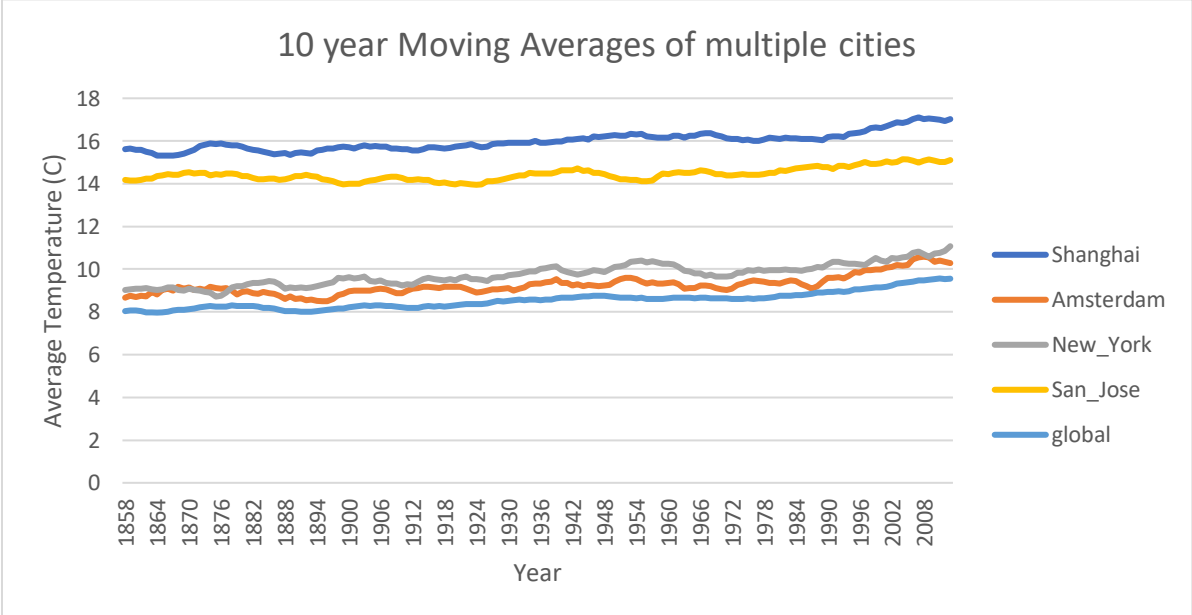
	A	B	C	F	G
1	year	avg_city_temp	avg_global_temp	10_yr_MA_city	10_yr_MA_global
2	1849	14.12	7.98		
3	1850	13.8	7.9		
4	1851	14.39	8.18		
5	1852	13.81	8.1		
6	1853	14.4	8.04		
7	1854	13.98	8.21		
8	1855	14.2	8.11		
9	1856	14.1	8		
10	1857	14.78	7.76		
11	1858	14.19	8.1	=AVERAGE(B2:B11)	
12	1859	13.71	8.25		
13	1860	13.81	7.96	14.137	8.071
14	1861	14.88	7.85	14.186	8.038

AVERAGE(number1, [number2], ...)

2. **Create a line chart**: plot the moving averages of city and global data. The line graph below illustrates the change in 10 year moving average temperature in San Jose and global relative to 1858 – 2013.



3. **Multiple cities and global temperatures**: The line graph below shows the 10-year moving averages of temperature in Shanghai, Amsterdam, New York, San Jose and global between 1858 and 2013.



Observations:

1. The temperature in San Jose is getting hotter. The average temperature increases from 14 ºC to 15 ºC approximately.

2. Global temperature is increasing, meaning the world is getting hotter. The average temperature rises from 8 ºC to 9.5 ºC approximately. This increasing trend has been consistent over the last hundred years.

3. The temperature in San Jose is hotter than the global average. The difference between San Jose average temperature and global average appears consistent over time.

4. The average temperature of San Jose and global both experienced gradual upward trend over the period. However, San Jose average showed more fluctuations while global average was relatively steady. The table below shows the minimum and maximum average temperature during the period.

	10 yr MA avg
Min city temp	13.951
Max city temp	15.148
Min global temp	7.968
Max global temp	9.556

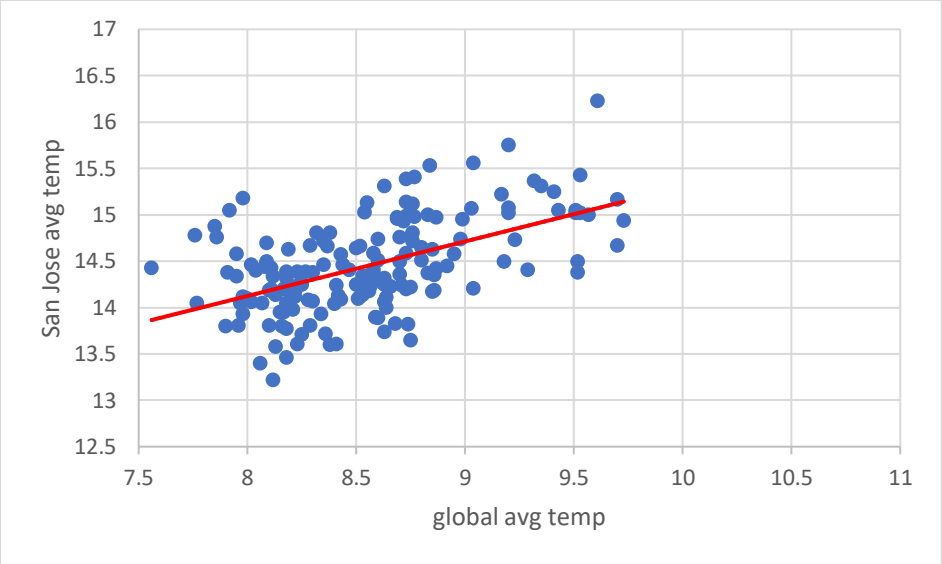
5. I plotted another line chart for multiple cities (Shanghai, Amsterdam, New York, San Jose) and global temperature using 10-year moving averages. The graph shows that temperature rises in all the cities. However, the temperature movement at city level has more volatility while the global has been more stable. This temperature rise, in a nutshell, is global warming.

Analyze data using EXCEL: Regression Model

1. **Correlation coefficient**: calculate correlation coefficient between San Jose and global average temperature using CORREL command in EXCEL. The correlation between San Jose average temperature and global average is 0.536, meaning the two variables are positively related and the strength of their relationship is moderate.

Correlation coefficient	0.536038143

2. **Scatter diagram:** create a scatter graph to quickly visualize the relationship between San Jose and global average temperature.



3. **Regression analysis**: use EXCEL Analysis Tool to build a linear regression equation. Simple linear regression models the relationship between San Jose average temperature (dependent variable) and global average temperature (independent variable) using a linear function.

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.536038				
R Square	0.287337				
Adjusted R Square	0.282965				
Standard Error	0.427526				
Observations	165				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12.01211802	12.01212	65.71957	1.17422E-13
Residual	163	29.79287955	0.182778		
Total	164	41.80499758			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	9.419591	0.621509182	15.156	6.12E-33	8.192344	10.64684	8.192344	10.64684
avg_global_temp	0.588131	0.072548257	8.106761	1.17E-13	0.444876	0.731387	0.444876	0.731387

San Jose avg temp = 0.5881 * global avg temp + 9.4196

With this estimated regression equation, I can estimate the average temperature in San Jose based on any given average global temperature. The **Significance F** value tells how reliable the results are. In this model, the significance F value is less than 5%, meaning

the model is fine. However, I also want to get an idea about how good this estimate is. So, I took another look at the R square value. **R square** is the coefficient of determination, which is used as an indicator of the goodness-of-fit for linear regression models. In this model, R square is 0.29 (rounded to 2 digits), which is relatively low. It means that 29% of the values fit the regression analysis model. In other word, only 29% of San Jose average temperature can be explained by global average temperature. So, this model can only provide a rough estimate.