



# **Data Analyst Course**

## **Exploring Weather Trends project**

Work by :

Nada Almutairi

- **Extract the data using SQL**

- the Global Data query

The screenshot shows a SQL query interface. On the left, under 'Input', there is a 'SCHEMA' section with a refresh icon and a list of tables: 'city\_data', 'city\_list', and 'global\_data', each with a dropdown arrow. The 'global\_data' table is selected. On the right, the SQL query is displayed in a text area:

```
1 SELECT *  
2 FROM global_data  
3
```

Below the query area, there is a green 'Success!' message and a blue 'EVALUATE' button. Below the input section, the 'Output' section shows '266 results' and a 'Download CSV' link. The output is a table with two columns: 'year' and 'avg\_temp'.

year	avg_temp
1750	8.72
1751	7.98
1752	5.78
1753	8.39
1754	8.47

- The Riyadh data query

The screenshot shows a SQL query interface. On the left, under 'Input', there is a 'SCHEMA' section with a refresh icon and a list of tables: 'city\_data', 'city\_list', and 'global\_data', each with a dropdown arrow. The 'city\_data' table is selected. On the right, the SQL query is displayed in a text area:

```
1 SELECT *  
2 FROM city_data  
3 WHERE city='Riyadh'
```

Below the query area, there is a green 'Success!' message and a blue 'EVALUATE' button. Below the input section, the 'Output' section shows '171 results' and a 'Download CSV' link. The output is a table with four columns: 'year', 'city', 'country', and 'avg\_temp'.

year	city	country	avg_temp
1843	Riyadh	Saudi Arabia	24.74
1844	Riyadh	Saudi Arabia	15.45
1845	Riyadh	Saudi Arabia	20.82
1846	Riyadh	Saudi Arabia	
1847	Riyadh	Saudi Arabia	
1848	Riyadh	Saudi Arabia	24.56
1849	Riyadh	Saudi Arabia	24.80

- **Preprocessing the data**

- used Excel to Prepare the two datasets

First, Riyadh Data has two NA value. Because the data is already small , I see the good way is replacement NA's with average . Also, I deleted City and country columns. They are useless for the comparison.

Moreover, the number of years in global data is larger than Riyadh data, it should be compatible to compare them. So I had to choose the same period which is from 1843 to 2013 .

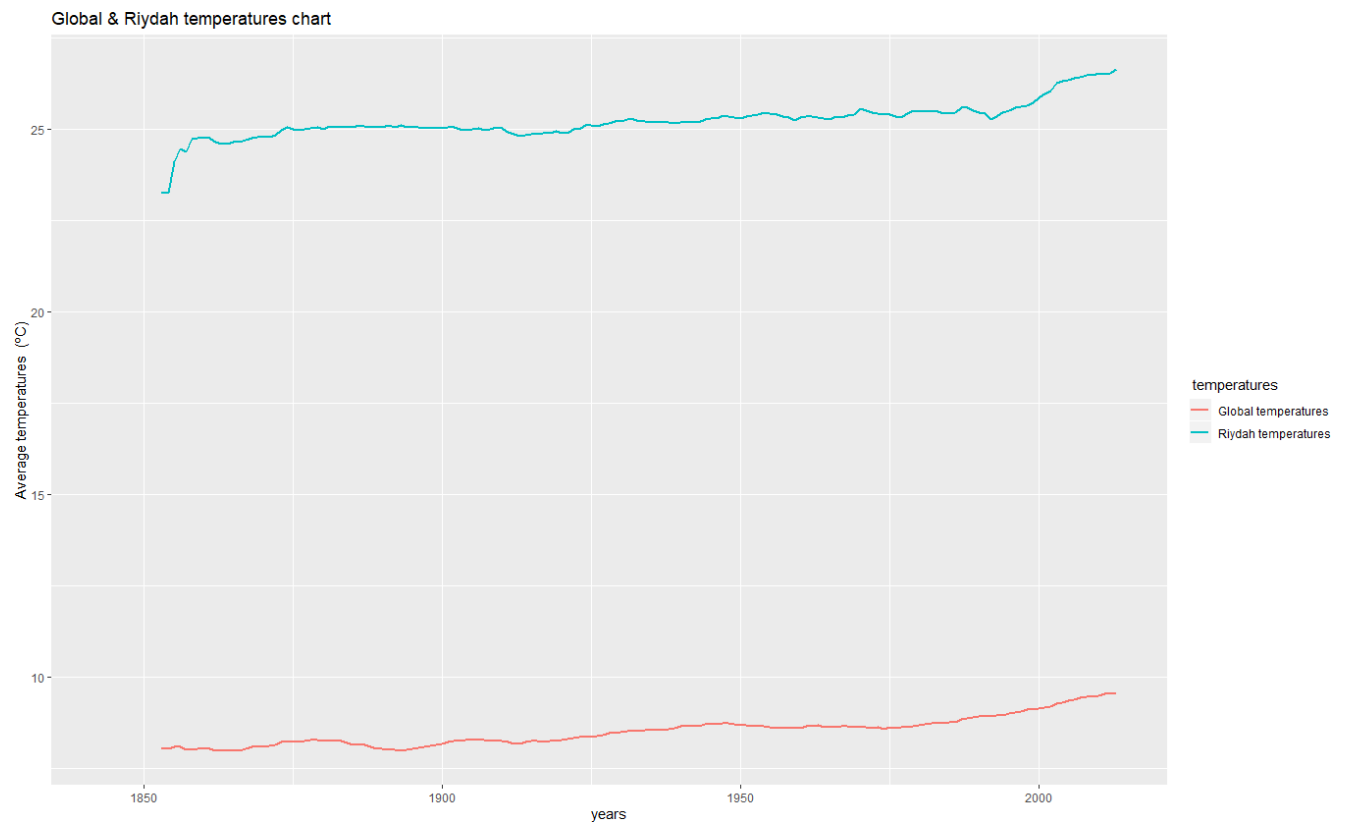
I find ten years is the best number to calculate the moving average with. Since I try the data distribution with five years, there was too much noise data. And, with 15 years the data distribution was totally different from original data. it may loss of details

	R	Q	P	O	N	M	L	K	J	I	H	G	F	E	D	C	B	A
																movAvg	avg temp	year
1																	24.74	1843
2																	15.45	1844
3																	20.82	1845
4																	25.21	1846
5																	21.21	1847
6																	24.56	1848
7																	24.8	1849
8																	24.34	1850
9																	25.03	1851
10																		1852
11																	=AVERAGE(B2:B11)	1852
12																	23.12	1853
13																	24.047	1854
14																	24.457	1855
15																	24.393	1856
16																	24.698	1857
17																	24.743	1858
18																	24.758	1859
19																	24.818	1860
20																	24.728	1861
21																	24.62	1862
22																	24.555	1863
23																	24.586	1864
24																	24.617	1865
25																	24.652	1866
26																	24.748	1867
27																	24.747	1868
28																	24.782	1869
29																	24.79	1870
30																	24.85	1871
31																	24.96	1872
32																	25.056	1873
33																	25.051	1874
34																	24.971	1875
35																	24.968	1876
36																	24.993	1877
37																	25.044	1878
38																	25.038	1879
39																	25.038	1880

- **Visualize the data with R**

- I choose it to visualize the Data with R .Since, I most familiar with it than other tools.

```
1 #read the datasets
2 riyData = read.csv("results Riydah.csv")
3 globData = read.csv("results global.csv")
4
5
6
7 #Moving Avg. from 1843 to 1852 is already calculated in Excel
8 #To avoid Warning messages from R I choose this subset
9 riyData = riyData [riyData$year >=1852,]
10 globData = globData [globData$year >=1852,]
11
12
13 # requier to visualize the datasets
14 library(ggplot2)
15
16
17 ggplot() +
18   geom_line(data=riyData, aes(year, movAvg , colour="red"), size=1 ) +
19   geom_line(data=globData, aes(year, movAvg , colour="blue"), size=1 ) +
20   labs(x='years' , y='Average temperatures (°C)', title='Global & Riydah temperatures chart') + #for title and axis names
21   scale_color_discrete(name=" temperatures" , labels = c("Global temperatures","Riydah temperatures")) #for lagned
22
23
```



- **Findings**

- Riyadh city temperature average is 25.19. While global average is 8.53. The difference between them is 16.66 degrees, it is clear that Riyadh is hotter than global.
- Riyadh Temperature from 1850 to almost 1875 has risen 2.5 degrees while global temperatures remain stable.
- In 2000 , the Riyadh has increased by two degrees then it gone up . While, The Global temperature increased by one degree.
- Also, In 1925, the Riyadh temperature has increased as global temperature is also increased. Which means there is a positives relationship between them. To make sure I find the Correlation coefficient between the average for both global & Riyadh, and the result is 0.77 . For moving average is 0.87 . Since it is rather close to 1, we can conclude that the global temperatures and Riyadh temperature are positively related.

Here is the code :

```
> cor(riyData$avg_temp, globData$avg_temp)
[1] 0.7763717
> #moving averages
> cor(riyData$movAvg, globData$movAvg)
[1] 0.8726988
```

## **References**

<http://www.rdocumentation.org/>

