EXPLORING WEATHER TRENDS

STEP I: Extract the data from the database

The given database contains three tables with the following names:

1. city_data
2. city_list
3. global_data

In order to find the closet big city to where I live, it is important to check the presence of my country and the corresponding cities in the database. The following SQL query was executed to list all the cities corresponding to the country which in this case is '*India*' in the *city_list* table:

```
SELECT * FROM city_list WHERE country = 'India';
```

The result set gave a list of all the registered cities where country is India and we found the closet big city to where I live is *'Ranchi'*. Further, the columns of all the three tables were viewed via clicking on each table. This gave an idea as to which columns to extract from the table in the database. The below SQL query was executed to download the data from the city_data table:

```
SELECT year, avg_temp FROM city_data WHERE country='India' and city='Ranchi';
```

Along with Ranchi, I also downloaded the temperature data for two more cities which are '*Patna*' and *'Kanpur'* using the below mentioned query:

```
SELECT year, avg_temp FROM city_data WHERE country='India' and city='Patna';
SELECT year, avg_temp FROM city_data WHERE country='India' and city='Kanpur';
```

Finally, the global temperature data was downloaded using the following SQL query:

```
SELECT * FROM global_data;
```

Four files were downloaded with a .csv extension.

Note: All the temperatures mentioned in the downloaded files are in degree Celsius (ºC).

STEP II: Tools Used

- The four downloaded files were opened with **Excel**.
- Since, the starting year for the global temperature data is 1750 and the other three cities are 1796, the global temperature data were normalized by removing the extra data rows (i.e. from 1750 to 1795).
- A consolidated excel file was created with 9 columns namely,
  year, mv_global, mv_ranchi, mv_patna, mv_kanpur, avg_temp_global, avg_temp_ranchi, avg_temp_patna, avg_temp_kanpur
- 10 years Moving Averages were calculated for the mentioned cities and global data using the Excel formula:
  =AVERAGE(COLUMN_RANGE)
  Further, the formula was dragged down till last data point to calculate all the moving averages.

B11 =AVERAGE(F2:F11)

	A	B	C	D	E	F	G	H	I
1	year	mv_global	mv_ranchi	mv_patna	mv_kanpur	avg_temp_global	avg_temp_ranchi	avg_temp_patna	avg_temp_kanpur
2	1796					8.27	24.01	24.99	24.59
3	1797					8.51	25.22	26.49	26.21
4	1798					8.67	23.33	24.27	23.82
5	1799					8.51	24.28	25.25	24.85
6	1800					8.48	24.24	25.2	24.79
7	1801					8.59	23.24	24.19	23.74
8	1802					8.58	24.7	25.64	25.23
9	1803					8.5	24.46	25.4	24.98
10	1804					8.84	24.8	25.72	25.3
11	1805	8.551	24.263	25.245	24.84	8.56	24.35	25.3	24.89
12	1806	8.567	24.287	25.267	24.861	8.43	24.25	25.21	24.8
13	1807	8.544	24.131	25.087	24.679	8.28	23.66	24.69	24.39
14	1808	8.44	24.22	25.17777778	24.77444444	7.63			
15	1809	8.297	24.2125	25.16875	24.765	7.08			
16	1810	8.141	24.20857143	25.16428571	24.76142857	6.92			
17	1811	7.968	24.37	25.32666667	24.93166667	6.86			
18	1812	7.815	24.304	25.264	24.872	7.05			
19	1813	7.739	24.13	25.094	24.704	7.74	23.59	24.55	24.14
20	1814	7.614	23.742	24.71	24.308	7.59	22.86	23.8	23.32

Fig 1: Consolidated temperature data excel sheet with Moving Average formula and data

STEP III: Chart Plot

- To create a **line chart**, the independent variable which is year in this case is plotted against the x-axis and the dependent variable which are the temperature data points are plotted against the y-axis.
- The data points for all the three cities were plotted along with the global temperature data points. The chart is labelled accordingly which can be seen in the Fig 2 below:

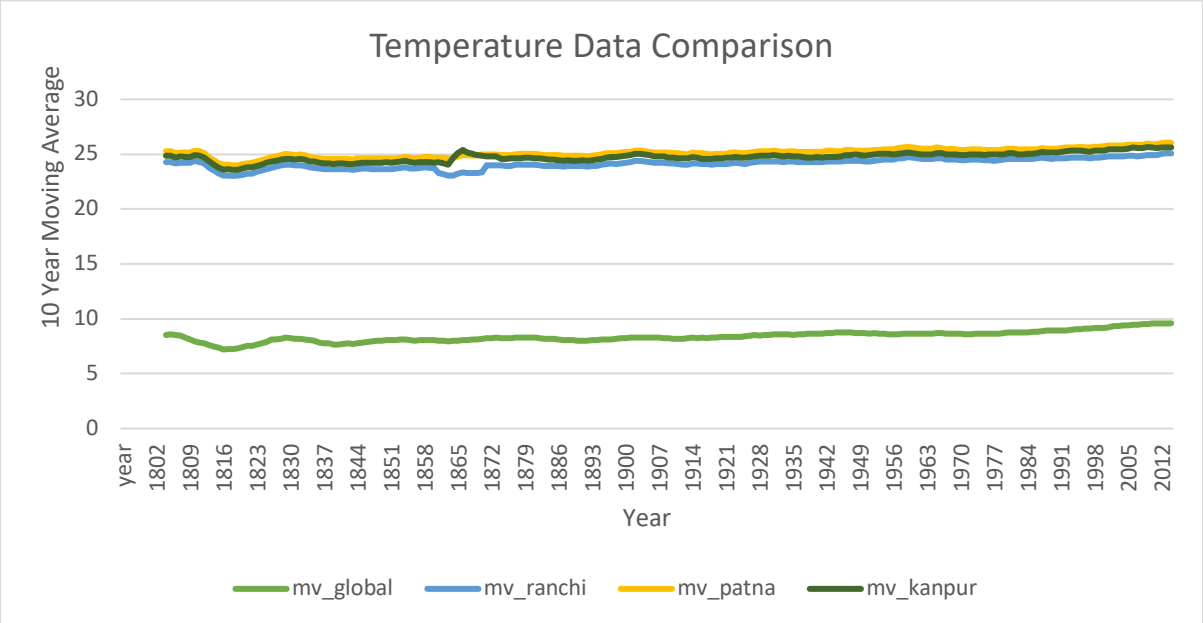

Fig 2: Global Temperature Data Comparison with Other Cities

STEP IV: Observation(s)

- According to the Line Cart above, the nearest largest big city to where I live i.e. ***'Ranchi'*** seems to be hotter than the global average temperature.
- If we analyze closely, Ranchi's temperature seems to be consistent throughout at around 25 degrees Celsius (ºC). Not only Ranchi, but the other two cities Patna and Kanpur have a consistent temperature throughout.
- Initially, during the year 1805-1810, the temperature of the nearest largest big city, Ranchi is around 24 and that the global average temperature was around 8.5 and both took a dip over the next 10-15 years resulting in a decline of approximately 1 ºC.
- The overall trend depicts a relation between the two but the nature of the relation is not confirmed. This can be verified via the correlation coefficient whose details are mentioned later in the report. It can also be noted that the global average trend line shows a slightly upward movement depicting gradual increase in the global average temperature over a period of time.

STEP V: Further Analysis

- In the above observation, based on the Line chart, we figured that there's a relationship between the global average and other cities temperature data. In order to find whether a relationship exists and if it does, what is the nature of relationship, we need to calculate the correlation coefficient.
- Again, we will use the Excel formula which is mentioned below to calculate the correlation coefficient:

  =CORREL(array1,array2)

- In order the use the data, we again need to normalize the data points and this time, we removed the missing rows which contains no data so as to make the dataset symmetric.

year	avg_temp_global	avg_temp_ranchi	year	avg_temp_global	avg_temp_patna	year	avg_temp_global	avg_temp_kanpur
1796	8.27	24.01	1796	8.27	24.99	1796	8.27	24.59
1797	8.51	25.22	1797	8.51	26.49	1797	8.51	26.21
1798	8.67	23.33	1798	8.67	24.27	1798	8.67	23.82
1799	8.51	24.28	1799	8.51	25.25	1799	8.51	24.85
1800	8.48	24.24	1800	8.48	25.2	1800	8.48	24.79
1801	8.59	23.24	1801	8.59	24.19	1801	8.59	23.74
1802	8.58	24.7	1802	8.58	25.64	1802	8.58	25.23
1803	8.5	24.46	1803	8.5	25.4	1803	8.5	24.98
1804	8.84	24.8	1804	8.84	25.72	1804	8.84	25.3
1805	8.56	24.35	1805	8.56	25.3	1805	8.56	24.89
1806	8.43	24.25	1806	8.43	25.21	1806	8.43	24.8
1807	8.28	23.66	1807	8.28	24.69	1807	8.28	24.39
1813	7.74	23.59	1813	7.74	24.55	1813	7.74	24.14
1814	7.59	22.86	1814	7.59	23.8	1814	7.59	23.32
1815	7.24	23.13	1815	7.24	24.08	1815	7.24	23.67
1816	6.94	22.86	1816	6.94	23.81	1816	6.94	23.33
1817	6.98	22.92	1817	6.98	23.87	1817	6.98	23.45
1818	7.83	22.99	1818	7.83	24	1818	7.83	23.8
1819	7.37	22.8	1819	7.37	23.74	1819	7.37	23.44
1820	7.62	23.07	1820	7.62	24.02	1820	7.62	23.67

Fig 3: Normalized data for 3 cities, namely, Ranchi, Patna and Kanpur to calculate correlation coefficient

The correlation coefficient for each of the cities are:

Cities	Correlation Coefficient
Ranchi	0.70943447
Patna	0.79373428
Kanpur	0.74571175

- The sign of correlation coefficient confirms the nature of the correlation which is positive and looking at the highest value, the correlation coefficient for Patna is 0.79 ~ 0.80 indicating a high positive relation between the temperature data of Patna and average global temperature.
- ***Hence, the line chart above, and the correlation coefficient values indicate that a linear relation exists between the global average and the cities data.***

Now, in order to find whether we can compute the average temperature in our city based on the average global temperature, we need to define a linear equation between the two, which can be obtained via the method of regression.

One of the easiest method to find the equation for the linear regression line is to plot a ***Scatter(X,Y) chart*** in the Excel and display the **equation and R-squared value (coefficient of determination)** using the trendline options in the chart itself. Below are the scatter charts for the 3 cities:

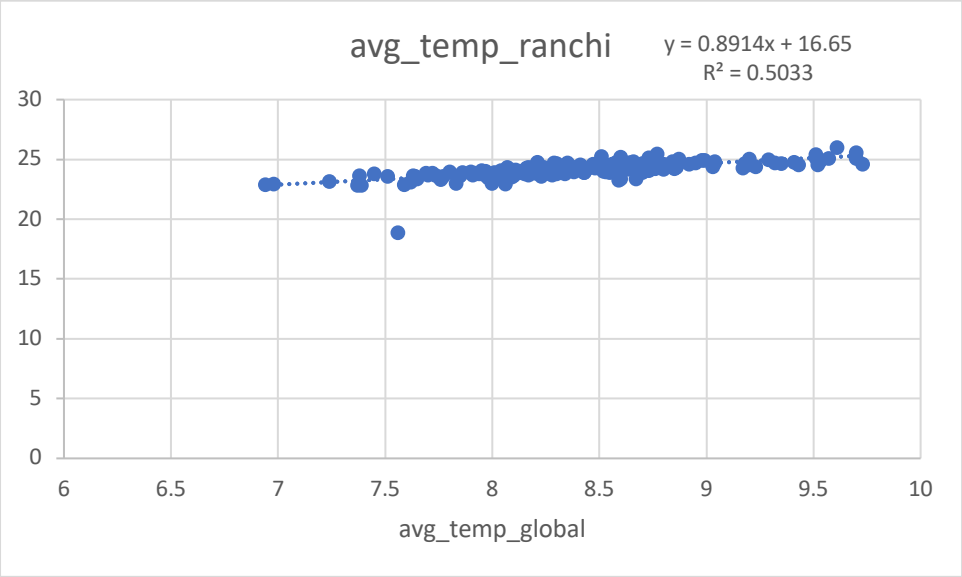

Fig 4: Scatter(X,Y) Plot chart for average temperature Ranchi

Here, the equation for the linear regression line for average temperature data for Ranchi is:

y = 0.8914x + 16.65

and that of coefficient of determination is,

R^2 = 0.5033

The above R^2 value implies that 50.33% of the total variation is explained by the least square regression line, and 49.67% of the variation in average temperature of Ranchi is explained by other factors.

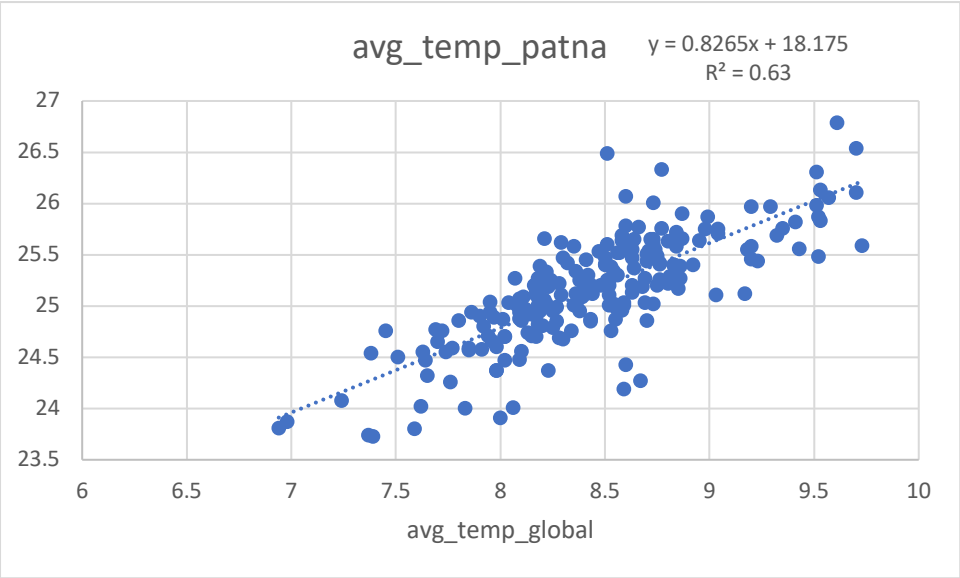

Fig 5: Scatter(X,Y) Plot chart for average temperature Patna

Here, the equation for the linear regression line for average temperature data for Patna is:

y = 0.8265x + 18.175

and that of coefficient of determination is,

R^2 = 0.63

The above R^2 value implies that 63% of the total variation is explained by the least square regression line, and 37% of the variation in average temperature of Patna is explained by other factors.

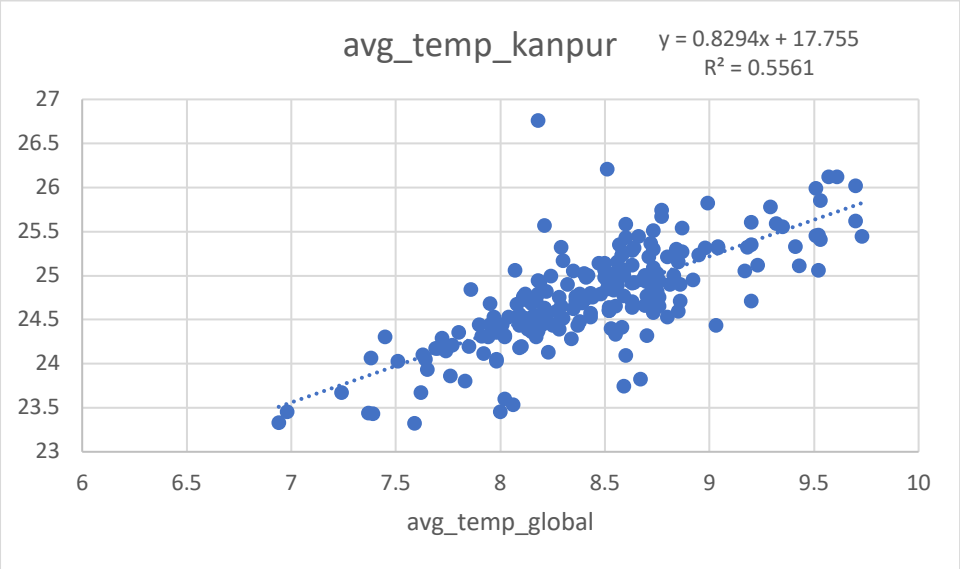

Fig 6: Scatter(X,Y) Plot chart for average temperature Kanpur

Here, the equation for the linear regression line for average temperature data for Kanpur is:

y = 0.8294x + 17.755

and that of coefficient of determination is,

R^2 = 0.5561

The above R^2 value implies that 55.61% of the total variation is explained by the least square regression line, and 44.39% of the variation in average temperature of Kanpur is explained by other factors.

Conclusion

Out of the 3 cities, Ranchi, Patna and Kanpur, the average temperature of all 3 can be computed based on the average global temperature. But the average temperature of Patna can be computed more accurately than Ranchi and Kanpur because of higher value of coefficient of determination.