# Explore Weather Trends

9th May 2019

Udacity - Data Analyst Nanodegree

Kushal Borkar

Term – 1, Project – 1,

Explore Weather Trends,

INDIA.

## Overview:

In this project, I have analyzed local and global temperature data and compare the temperature trends for the closest city from where I studied which is 'Hyderabad' to overall global temperature trends. I had been provided with a database on Udacity portal

The city I study or live in does not exist in the database so I used the closest city to my country.

## Objective:

The objective is to create a visualization and prepare a review describing the similarities and differences between global temperature trends and temperature trends in the nearest city to where I study.

## Tools used:

- **SQL Queries:** For acquiring the dataset from Udacity Portal.
- **Python:** For calculating the moving average and plotting a line chart.
- **ANACONDA - Jupyter Notebook:** For writing python code and making observations.

## Procedure:

### Step 1: Extraction of Data from Udacity Portal

SQL Queries were used to download (CSV) file that contains yearly average temperature of the City ‘Hyderabad’ and the global temperature.

```
/* Select the City from India */
```

```sql
select * from city_list where country = 'India';
select * from city_data where city = 'Hyderabad' and country = 'India';

/* Select Global Temperature */
select * from global_data;

/* Change the names of the columns in order to have distinct columns */
alter table city_data RENAME COLUMN avg_temp to local_avg_temp;
alter table global_data RENAME COLUMN avg_temp to global_avg_temp;

/* Join the two tables */
select
global_data.year,city_data.city,global_data.global_avg_temp,city_data.local
_avg_temp from global_data,city_data
WHERE(global_data.year = city_data.year) AND (city_data.city = 'Hyderabad'
AND city_data.country = 'India');
```

Thus, I have got an option of downloading the file as CSV format from Udacity Portal and downloaded as “results.csv”.

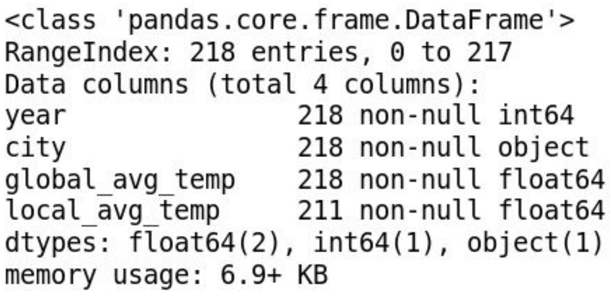
## Step 2: Handling Missing Data

I have used Python for this following work. To check whether the data has missing values or not, we need to check the following:

```python
# Using the necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Reading the data
data = pd.read_csv('results.csv')
data.info()
```

This is the result of the above code. And it clearly shows that ‘local_avg_temp’ has 7 missing values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 218 entries, 0 to 217
Data columns (total 4 columns):
year                  218 non-null int64
city                  218 non-null object
global_avg_temp       218 non-null float64
local_avg_temp        211 non-null float64
dtypes: float64(2), int64(1), object(1)
memory usage: 6.9+ KB
```

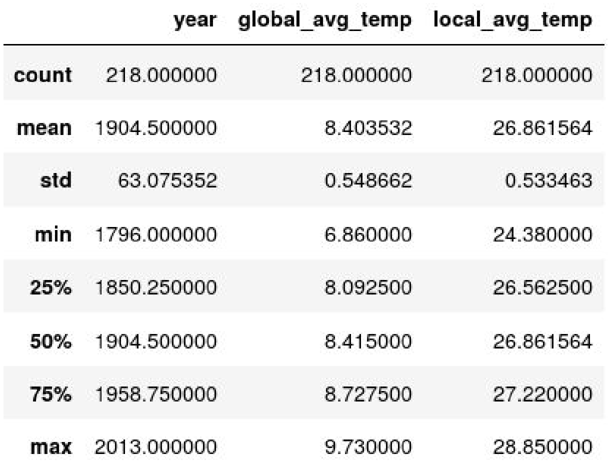
We have substituted missing values using the mean value of a feature when it is not available.

```
data['local_avg_temp'].fillna((data['local_avg_temp'].mean()),
inplace=True)
```

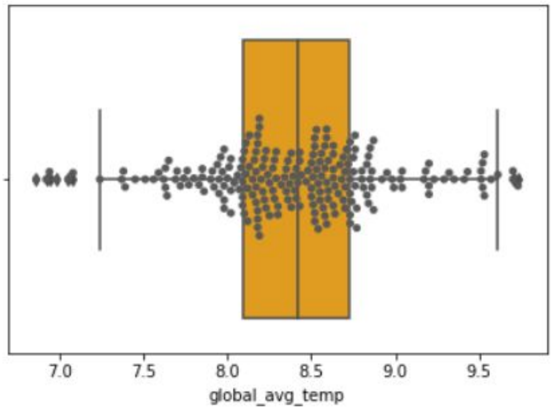
## Step 3: Dataset & Making Line Chart

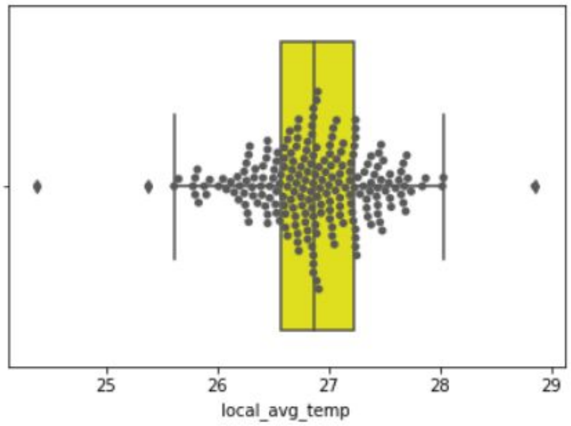
Before checking the Line charts, I would like to show details about the dataset. We can check the details about the dataset by using the following code and its corresponding output:

```
data.describe()
```



| | year | global_avg_temp | local_avg_temp |
|---|---|---|---|
| **count** | 218.000000 | 218.000000 | 218.000000 |
| **mean** | 1904.500000 | 8.403532 | 26.861564 |
| **std** | 63.075352 | 0.548662 | 0.533463 |
| **min** | 1796.000000 | 6.860000 | 24.380000 |
| **25%** | 1850.250000 | 8.092500 | 26.562500 |
| **50%** | 1904.500000 | 8.415000 | 26.861564 |
| **75%** | 1958.750000 | 8.727500 | 27.220000 |
| **max** | 2013.000000 | 9.730000 | 28.850000 |





The above two boxplot shows the distribution of the global average temperature and local average temperature.

An explicit function was written for moving average to check the trend. This helped us to make it easier to observe the trends when the line are shown in Charts.

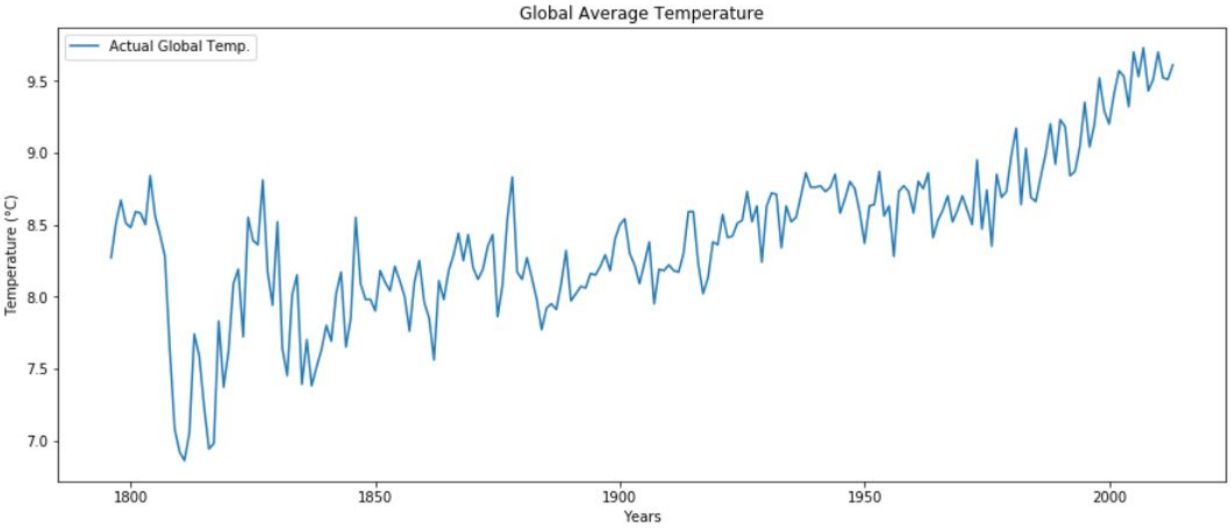
The moving average was programmed using an in-built function in ‘dataframe.rolling()’ function present in the Pandas library. The function is as follow:

```
def moving_avg(data_series,win):
    out = data_series.rolling(window = win, center = False).mean()
    return out
```

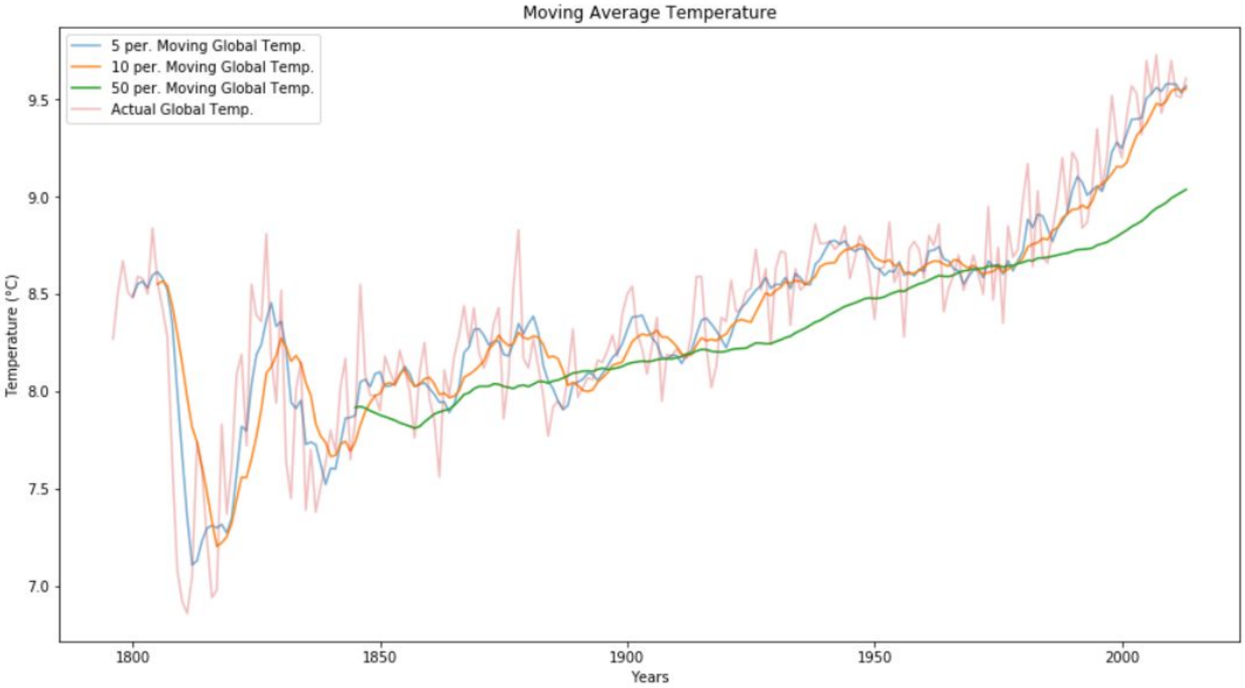
The above function enabled me to plot the graph showing line with the different moving average window. For example:

```
plt.plot(data['year'], moving_avg(data['global_avg_temp'], 50), label='50
per. Moving Global Temp.')
```

The graph of the actual global average temperature is as follow:

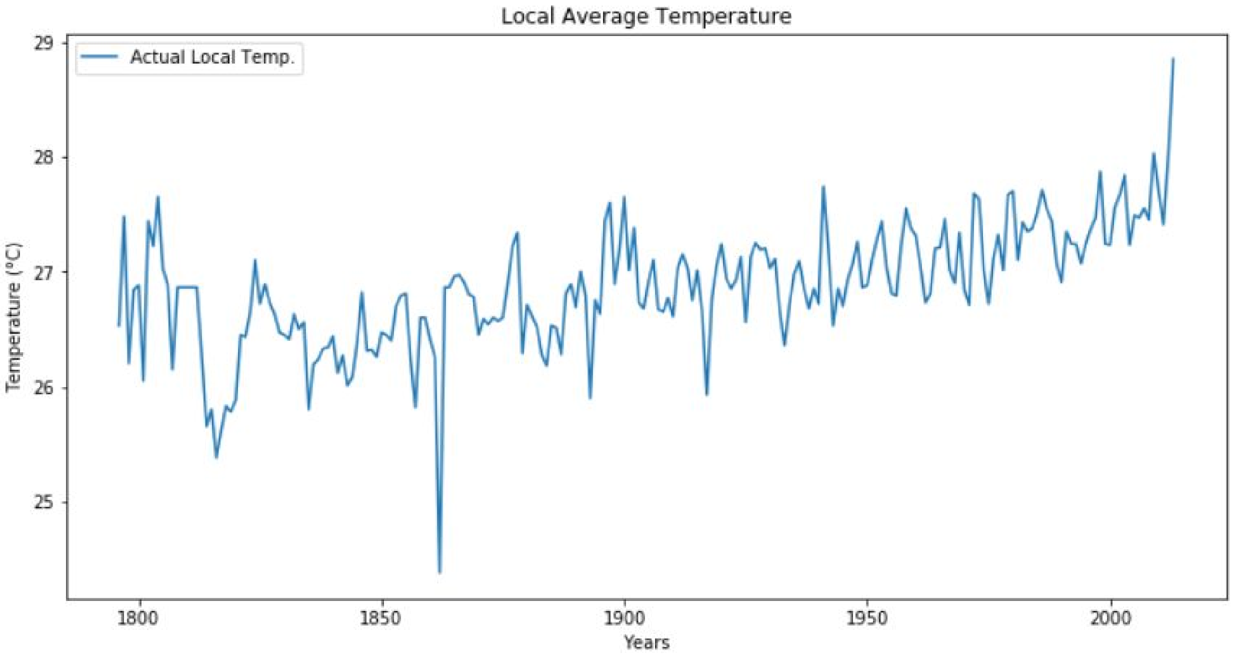


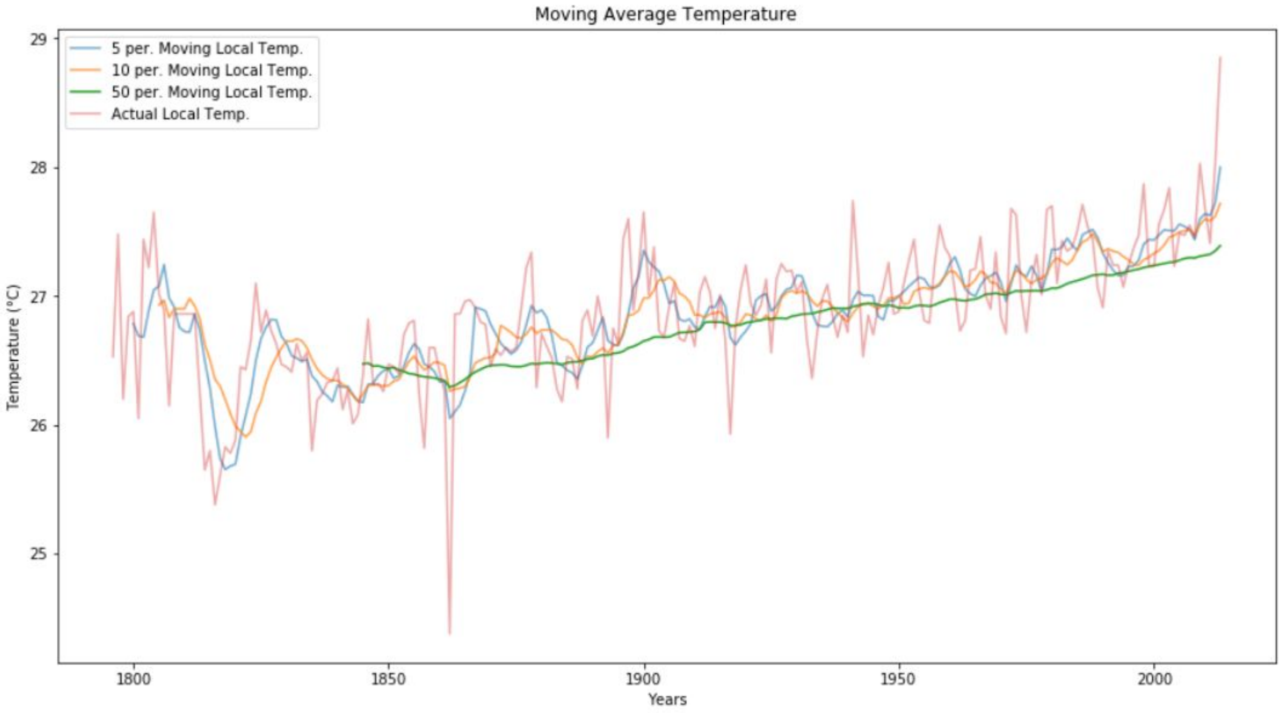
To observe the trend in the temperature, we plot the moving average and check the trend.



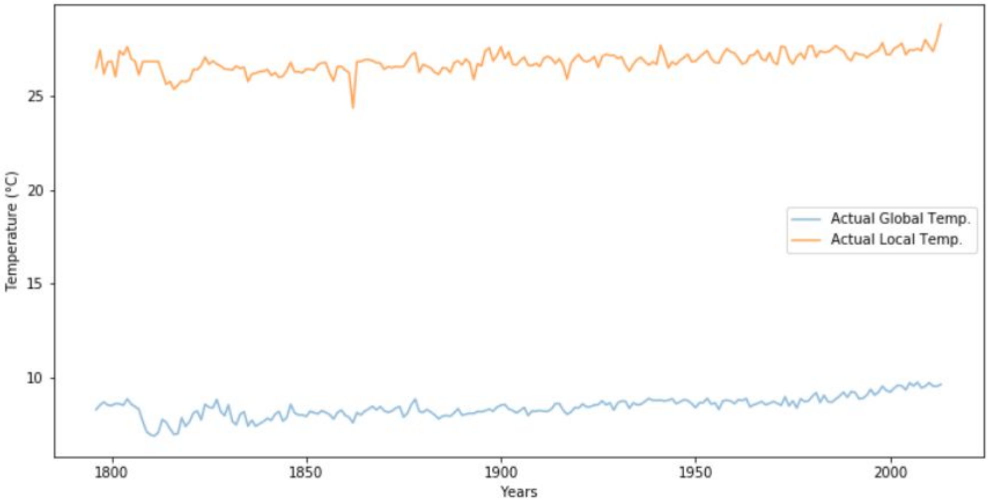
This shows that the global temperature is increasing.

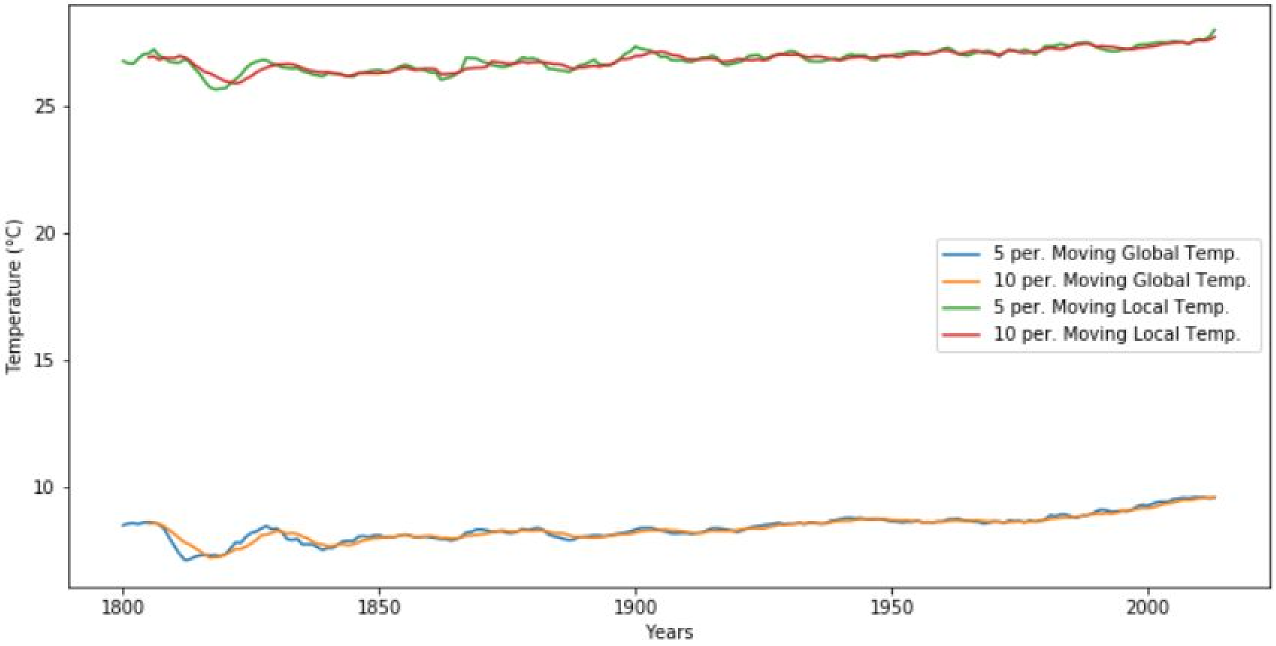
Similarly, the trend with the local average temperature as well as the moving average of the local temperature is given below.





I have separately analysed the global data in order to check and distinguish it from combined data of New Delhi and Global Average temperatures. Now I shall plot the combined graph along with its moving average.

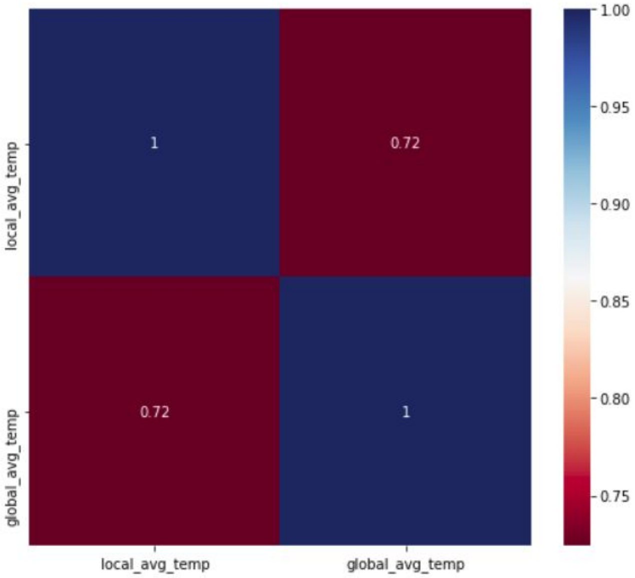




## Result: Observation

I have observed that, if I pick a short range for moving average, for example, 5 or 10, I will get an untidy line in the chart. Likewise, the range of the "Years" on x-axis turns out to be longer. Furthermore, if I utilize a bigger moving average window, for example, 50, I will get a moderately smooth line.

- The graph of Hyderabad versus Global Temperature: the Very enormous distinction(difference or gap) between the normal temperature of Hyderabad and that of the world.
- From both the moving average chart: I have observed that worldwide temp. and neighbourhood temp. is increasing constantly with years.
- Global Average Temperature varies between 6.86℃ to 9.73℃ whereas Hyderabad's Average temperature varies from 24.38℃ to 28.85℃.
- Examining the temperature in Hyderabad after 2010, we see a definite increment in the data. This is because of the temperature increasing by 0.1 or more each year. (This pattern is all the more significantly unmistakable in the Average Temperatures Graph than the Moving Averages Graph).
- Hyderabad and Global average temperature have a comparative sort of trend. Amid the early years, both the trends appear to have high points and low points then approx. around 1992 the moving average temperature begins to increment at a consistent rate.
    - This can be represented by the Heatmap(correlation between them which is positive):



The correlation between +0.72, which is a positive correlation, we can infer that both have a similar type of rises and falls.