Explore Weather Trends

14th October, 2018

Project One

Udacity Nanodegree Program

Data Analyst Nanodegree

Introduction

In this project, we will analyze local and global temperature data and compare the temperature trends from few cities around the work to overall global temperature trends.

Tools Used

- SQL: SQL is a query language design for management of RDBMS. We used it to extract the data which will form as input to the analysis.
- Python3: Python is an interpreted high-level programming language for general-purpose programming which is widely popular in Data Science community.
- Anaconda: Anaconda is a package management tool for python. It helps in management of multiple environments present for various projects.
- pandas: pandas is an open source library providing high-performance, easy-to-use data structures(called Dataframes) and data analysis tools for the Python programming language.
- Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Solution

Step 1 : Getting the Data

First step was data collection. In the SQL schema there were 3 tables.
First step was identifying the number of cities that were present.

To check that we ran:

- select * from city_list;

It gave the number of cities as output.
For analysis on this project we choose 4 cities.

- Hyderabad , India
- Delhi, India
- Nagpur, India
- New York City, USA

After Identifying the cities the logical next step was extracting the data of the cities using SQL.

For this purpose we ran 4 scripts corresponding to each of the city.

- select * from city_data
  where city = 'Delhi'

- select * from city_data
  where city = 'Nagpur'

- select * from city_data
  where city = 'New York'

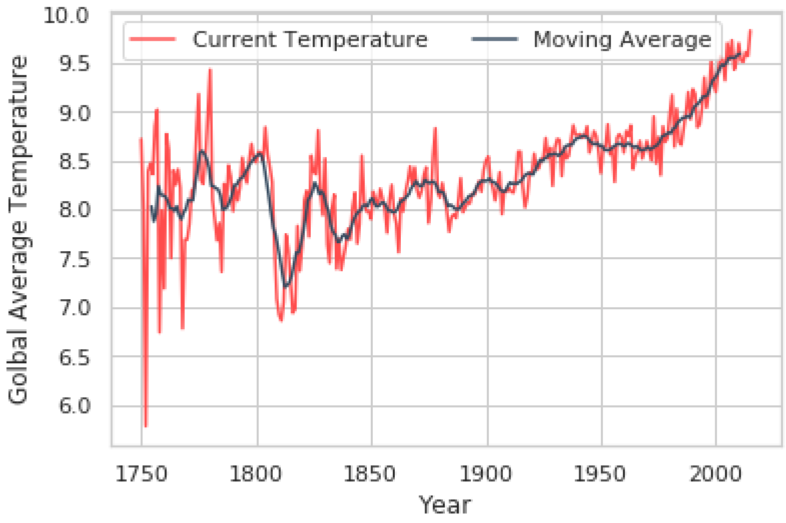
-
  ```
  select * from city_data
  where city = 'Hyderabad'
  AND
  country = 'India';
  ```

As, there are two Hyderabad in the dataset so country is also included in the query.
The output of this step was 5 csv files containing data of the aforementioned cities and global temperature.

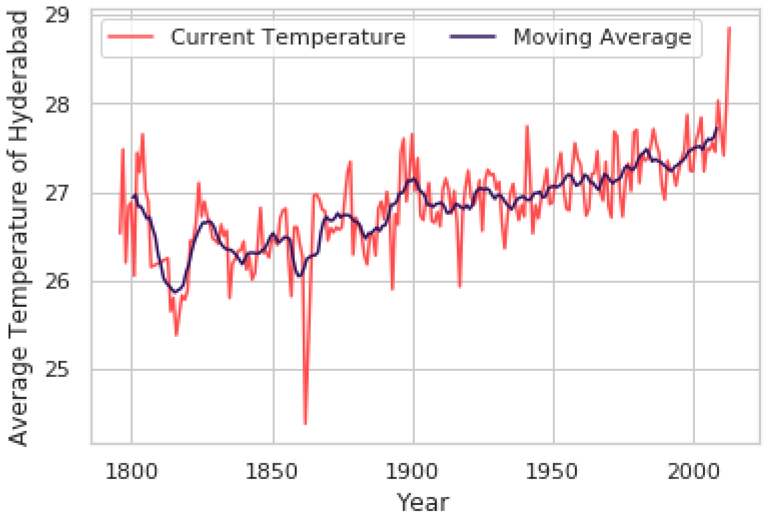
Step 2 : Loading and Preprocessing the data

The data was loaded into the pandas Dataframe using read_csv() function. We make 5 different data frames corresponding to each of city and the global trend. We then found the missing values present in the Dataframe (if any) Using **df.isna().sum()**. We then eliminated the missing values using interpolation, which used polynomial of order 1 to fill the data. The filling of missing values was done using the interpolate() function. We also deleted the columns name of city and country as they were of no use to us.

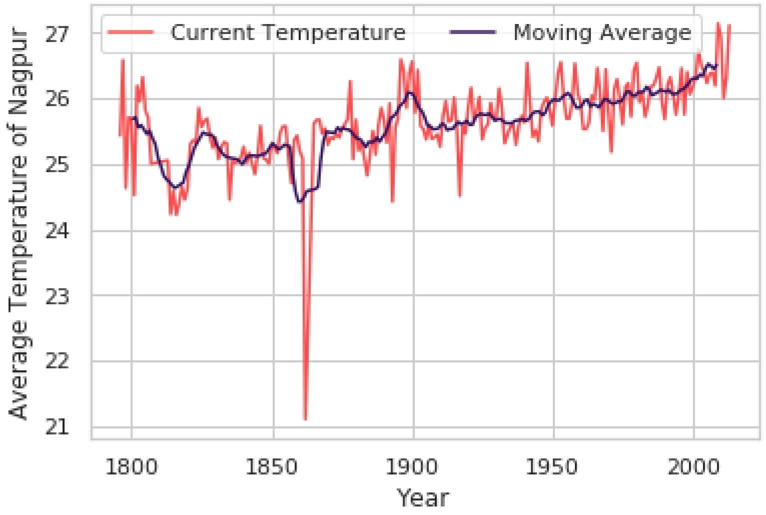
Step 3 : Plotting the simple line chart along with Moving Average.



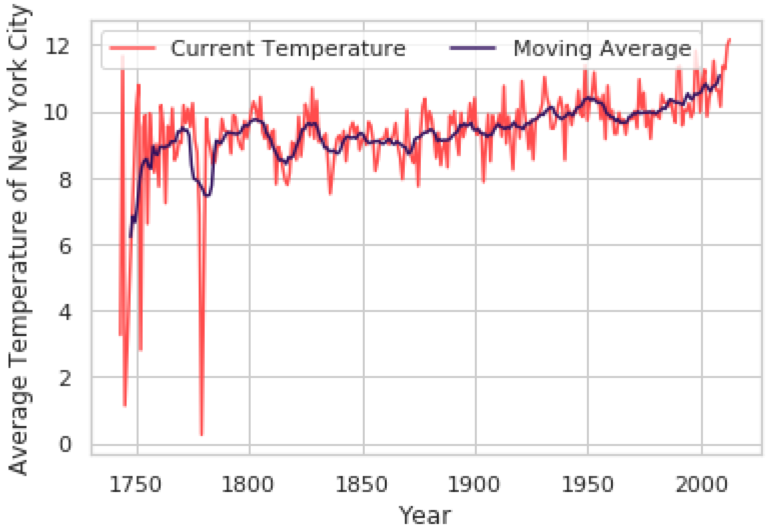
Global Average Temperature Vs Year



Average Temperature of Hyderabad Vs Year

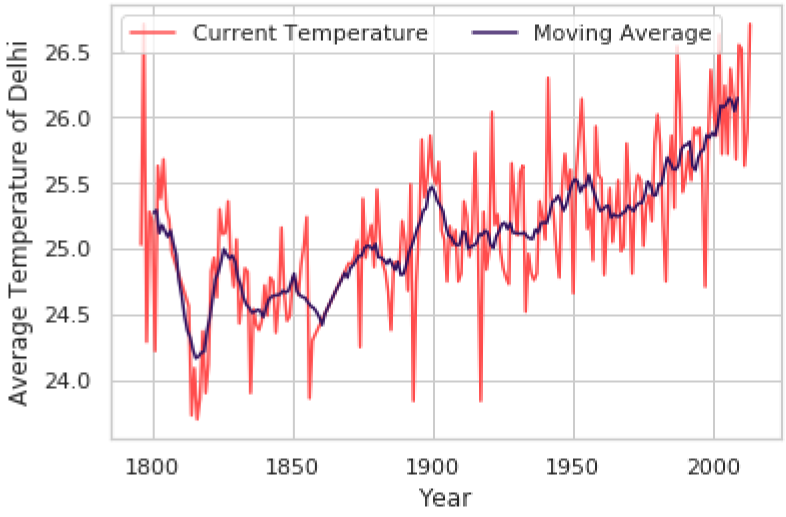


Average Temperature of Nagpur vs Year



Average Temperature of New York City vs Year

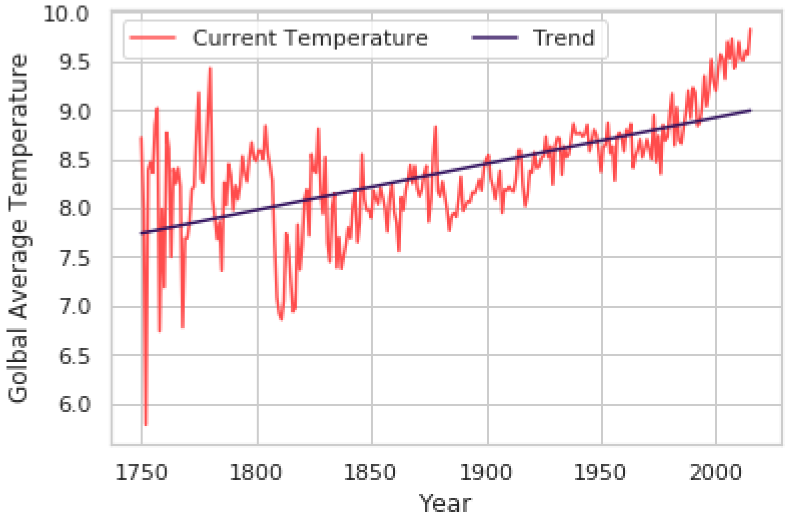
For plotting the moving average of the data we have used **rolling()** function with parameters 10 as size of moving window and mean as the smoothing function.

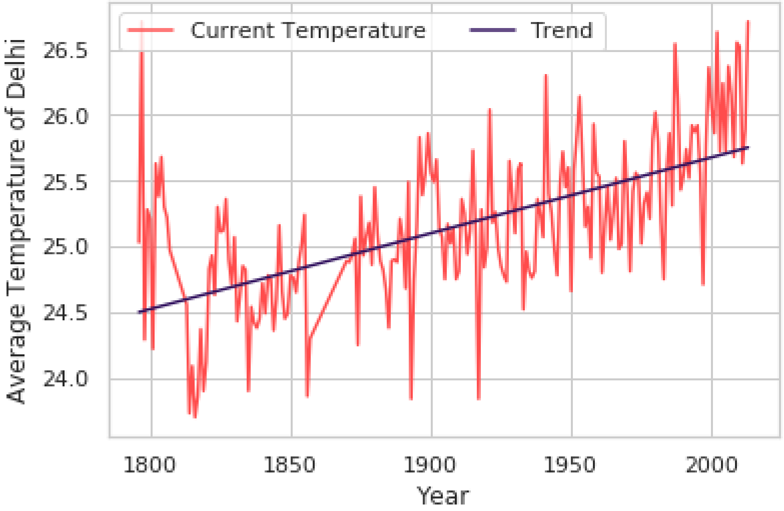


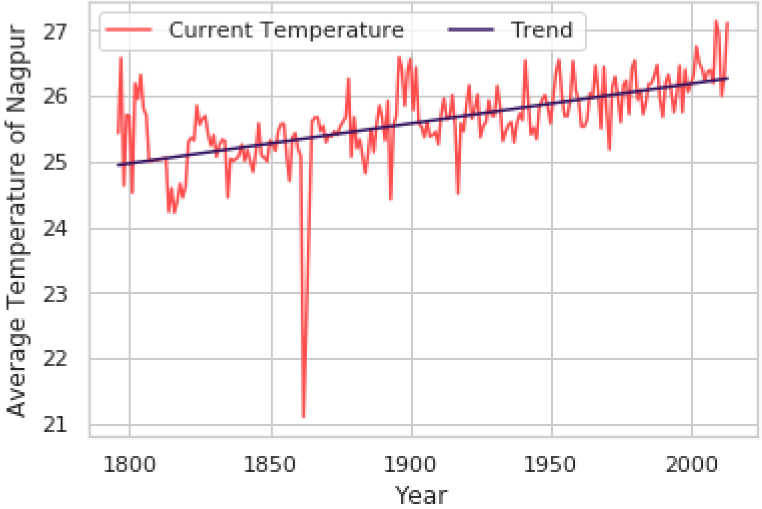
Average Temperature of Delhi vs Year

Step 4 : Making a trend line and calculating correlation coefficient.

We also made a trend line to see the growth of temperature wrt time.

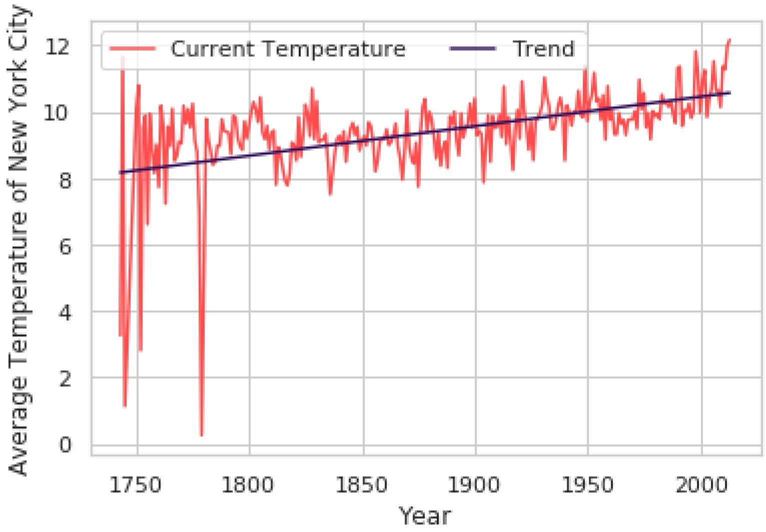






Correlation coefficient measure the correlation between two variables X and Y. For calculation of correlation coefficient we used the pandas library coerr() function. We calculated correlation of global temperature wrt cities we got the following result.

- Hyderabad : 0.7719087312896322
- Nagpur : 0.6840742641518918
- Delhi : 0.7437767384906534
- New York City : 0.712868290245507



Observations

- When we see the above plot of trend line of global temperature, we see that global temperature is increasing with time.
- Few data points vary a lot from the average trend. It might be possible that there was error in measuring the temperature.
- To counter the effect of the outliers we are using the moving averages which gives a more realistic outlook about the increase in the temperature.
- There also exist a correlation between global temperature and local temperature where Hyderabad closely followed the pattern in global temperature.

Source Code

https://github.com/abhiksark/UD-ND/blob/master/explore_weather_trends/trend_plot.ipynb

References

- https://pandas.pydata.org/pandas-docs/stable/
- https://seaborn.pydata.org/