DSCI 510 Final Project Report

Description: For the final project, I have created a dataset in the form of CSV files that contain the weather data in 6 cities for the years 2018, 2019, and 2020. The 6 cities and corresponding weather stations are,

- Asia - New Delhi / Safdarjung
- Africa - Cairo International Airport
- North America - Los Angeles / Jefferson
- South America - Rio De Janeiro Aeroporto
- Europe - London Weather Centre
- Australia - Sydney - Observatory Hill

The file names are in the following format, "<city_name>_weather_data_<year>.csv". This dataset is taken from the weather API mentioned in the Data Sources section, representing 6 continents out of 7. Hence, it gives a good representation of the weather variation worldwide. Using these datasets, I have analyzed the variation of weather data over time and how global warming affects these weather trends.

Motivation: Global warming is one of the most discussed topics in the modern world. It has numerous dangerous effects on the planet, directly impacting the world's future. People are trying to identify the reasons for this and prevent it. Weather patterns all over the world have changed because of the causes of global warming. Therefore, I chose a dataset related to time series data of weather parameters worldwide and started analyzing them to see the actual variation of those parameters over time. These analyses of weather data over several years allowed me to identify trends in weather variation caused by the global warming effect. These analyses can be used to predict future weather conditions as well.

Dependencies:

The following Python libraries have been used in this project:

```
matplotlib==3.5.2
meteostat==1.6.5
pandas==1.4.2
seaborn==0.11.2
```

Installation:

To install the project requirements, run the following command in the Command Line Window:

```
pip install -r requirements.txt
```

Running the project:

Run the following command in the Command Line Window to run the project:

```
python code/main.py
```

Data Sources:

The primary data source is the "Meteostat API". There are three different interfaces to get daily weather data for past years from the API - JSON API, Python Library, and Bulk Data format. The endpoint for the Python library is-

"https://meteostat.p.rapidapi.com/point/daily"

An example link that can be used to fetch data is shown below. To access data from this endpoint, we need to subscribe to the API through the "RapidAPI" site to get an API key.

https://meteostat.p.rapidapi.com/point/daily?lat=43.6667&lon=-79.4&start=2020-01-01&end=2020-01-31&alt=184

To get data, we need to create a specific URL for each city by adding the latitude and longitude values for that corresponding city. These URL generations are already present in the code and are not to be given as input by the user. Using the "Meteosat" library, which is built for accessing the "Meteostat API" based on the "requests" library, I have accessed the CSV from the URL and created the data files. Each data file consists of 11 columns with 365 or 366 data records depending on leap years. In my dataset, only the year 2020 is a leap year. Therefore, there are 6576 data records in the dataset. The columns in the dataset are as follows:

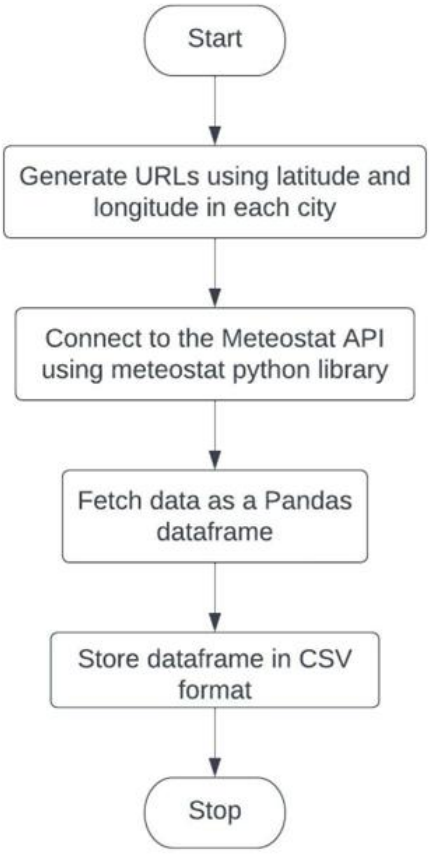
- time - The date string (YYYY-MM-DD) [String]
- tavg - The average air temperature in °C [Float]
- tmin - The minimum air temperature in °C [Float]
- tmax - The maximum air temperature in °C [Float]
- prcp - The daily precipitation total in mm [Float]
- snow - The maximum snow depth in mm [Integer]
- wdir - The average wind direction in degrees (°) [Integer]
- wspd - The average wind speed in km/h [Float]
- wpgt - The peak wind gust in km/h [Float]
- pres - The average sea-level air pressure in hPa [Float]
- tsun - The daily sunshine total in minutes (m) [Integer]

The first few rows of the Cairo_weather_data_2018 datafile are shown below:

	A	B	C	D	E	F	G	H	I	J	K	L
1	time	tavg	tmin	tmax	prcp	snow	wdir	wspd	wpgt	pres	tsun	
2	1/1/2018	14.8	11.8	19			232	20.9		1019.2		
3	1/2/2018	15.6	12	19			195	17.4		1020.6		
4	1/3/2018	16.4	13	20			197	17		1018.3		
5	1/4/2018	16	12.8	19			214	31.9		1014.2		
6	1/5/2018	16.9	15.1	19			240	33.5		1012.2		
7	1/6/2018	15.8	13	19			251	11.7		1019.9		
8	1/7/2018	15.7	10	21				15.3		1025.6		
9	1/8/2018	15.5	10	21.1			40	18.7		1027.8		
10	1/9/2018	15.8	10	21			48	18.1		1027		
11	1/10/2018	15.8	11	21			63	16.9		1021.7		
12	1/11/2018	18.9	12.9	24			202	19.8		1016.6		
13	1/12/2018	17.2	13	21				8.1				
14	1/13/2018	15.9	11	20			357	11.5		1016.9		
15	1/14/2018	16.5	13	20			229	19.6		1018.8		
16	1/15/2018	16.4	13.7	19.6			281	11.5		1025		
17	1/16/2018	15.2	12	19			24	11.6		1024.3		
18	1/17/2018	16.7	12	20				9.1		1018.7		
19	1/18/2018	17.5	13	22			212	33		1014		
20	1/19/2018	14.7	12.7	17			274	22.2		1017.6		
21	1/20/2018	14.8	10	19				9.5		1022.4		

Cairo_weather_data_2018

Flowchart for dataset generation:



Analysis Performed:

First, each city's maximum and minimum values in each year are calculated using the functions defined in the "run_analysis.py" file. The calculated data has been shown below:

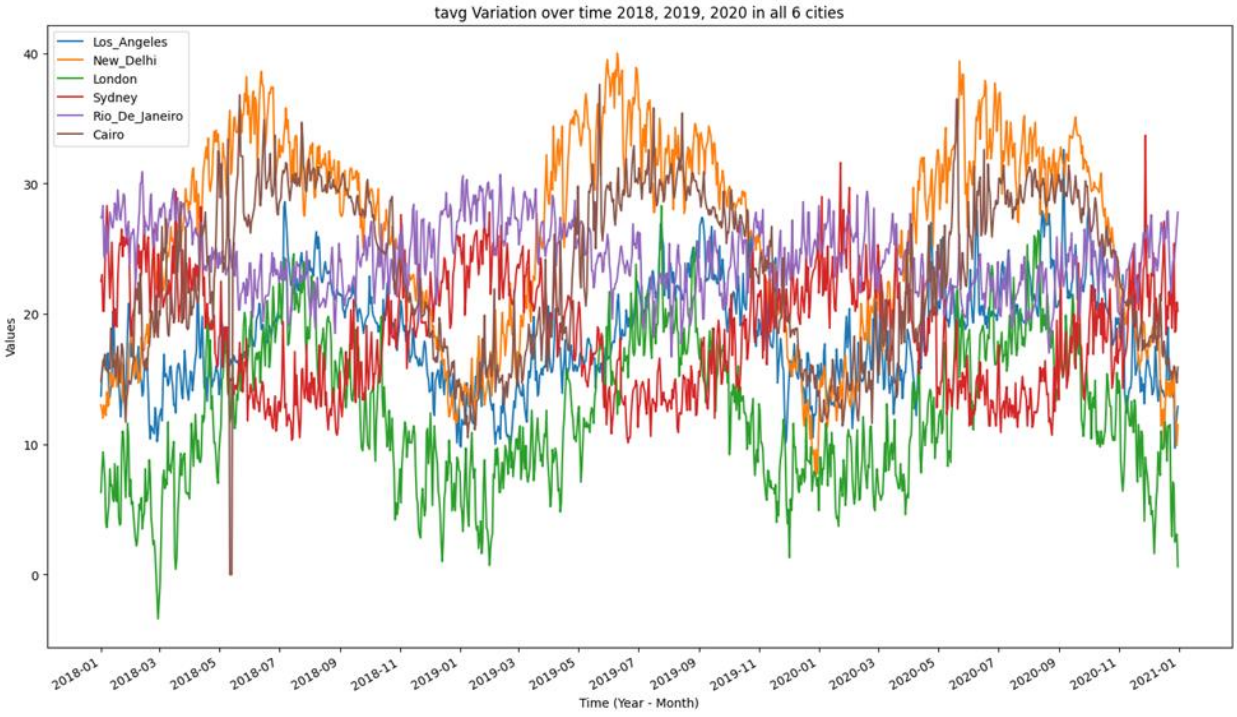
```
Maximum "tavg" values for the 6 cities in each year.

Maximum "tavg" in Los_Angeles in the year 2018 is 28.6
Maximum "tavg" in Los_Angeles in the year 2019 is 27.4
Maximum "tavg" in Los_Angeles in the year 2020 is 32.6

Maximum "tavg" in New_Delhi in the year 2018 is 38.6
Maximum "tavg" in New_Delhi in the year 2019 is 40.0
Maximum "tavg" in New_Delhi in the year 2020 is 39.4

Maximum "tavg" in London in the year 2018 is 24.6
Maximum "tavg" in London in the year 2019 is 28.3
Maximum "tavg" in London in the year 2020 is 26.4

Maximum "tavg" in Sydney in the year 2018 is 29.4
Maximum "tavg" in Sydney in the year 2019 is 27.8
Maximum "tavg" in Sydney in the year 2020 is 33.7

Maximum "tavg" in Rio_De_Janeiro in the year 2018 is 30.9
Maximum "tavg" in Rio_De_Janeiro in the year 2019 is 30.7
Maximum "tavg" in Rio_De_Janeiro in the year 2020 is 29.4

Maximum "tavg" in Cairo in the year 2018 is 36.8
Maximum "tavg" in Cairo in the year 2019 is 37.6
Maximum "tavg" in Cairo in the year 2020 is 36.5
```

In the analyses, the "tavg" parameter is considered out of the three temperature parameters since it has the median variation in temperature. The main reason to choose the "tavg" parameter as the primary statistical analyzing parameter is that global warming mainly affects temperature according to common factor considerations. As we can see in the above calculations, only the temperature in Los Angeles has increased continuously with time. Most other cities have the maximum "tavg" temperature for 2019, and 2018 and 2020 have fewer values than that.

In the minimum temperature calculation data, we can see a similar variation to the above-described scenario. Most cities had the same minimum temperature in 2019, and 2018 and 2020 had larger values than 2019.

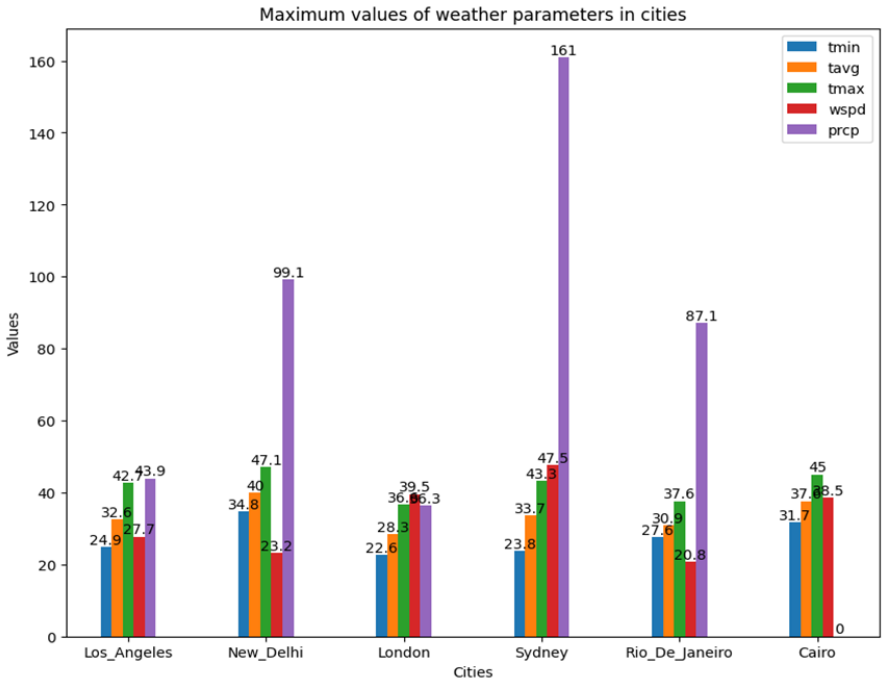
Using this data, we can only get an approximate idea about the effect of global warming since it represents point data. However, the following plots will help us identify the variation in weather parameters in a more descriptive manner:

```
Minimum "tavg" values for the 6 cities in each year.

Minimum "tavg" in Los_Angeles in the year 2018 is 10.1
Minimum "tavg" in Los_Angeles in the year 2019 is 9.4
Minimum "tavg" in Los_Angeles in the year 2020 is 9.7

Minimum "tavg" in New_Delhi in the year 2018 is 11.6
Minimum "tavg" in New_Delhi in the year 2019 is 7.7
Minimum "tavg" in New_Delhi in the year 2020 is 9.8

Minimum "tavg" in London in the year 2018 is -3.4
Minimum "tavg" in London in the year 2019 is 0.7
Minimum "tavg" in London in the year 2020 is 0.6

Minimum "tavg" in Sydney in the year 2018 is 10.3
Minimum "tavg" in Sydney in the year 2019 is 10.1
Minimum "tavg" in Sydney in the year 2020 is 10.7

Minimum "tavg" in Rio_De_Janeiro in the year 2018 is 18.5
Minimum "tavg" in Rio_De_Janeiro in the year 2019 is 16.7
Minimum "tavg" in Rio_De_Janeiro in the year 2020 is 17.0

Minimum "tavg" in Cairo in the year 2018 is 0.0
Minimum "tavg" in Cairo in the year 2019 is 10.5
Minimum "tavg" in Cairo in the year 2020 is 11.4
```



Looking at the above plot, we can't identify a massive variation in the "tavg" weather parameter. All of the cities have similar types of temperature variation over three years. It is like multiple bell-shaped variations over time. Other than that, if we look closely at the plot, we can identify a slight but increasing variation in each line in the plot. The bell-shaped curve variation is going higher with time. We can take it as an effect of global warming.

The plot below represents the all-time maximum values in each weather parameter in each city. This data representation gives little information about the global warming effect on the data. Still, it is helping us to identify the variations in the weather worldwide in a comparative way.

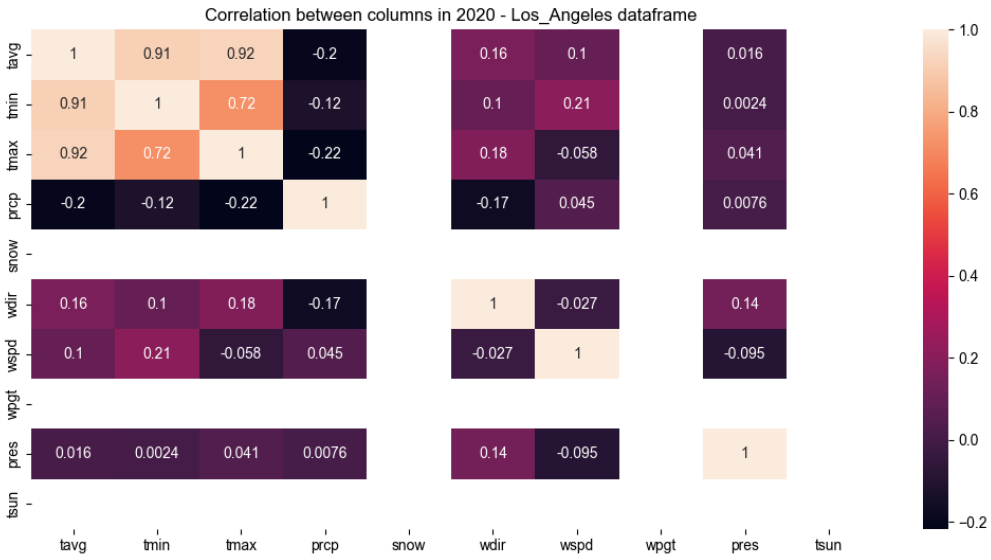


Other Analysis Performed:

1. **Heatmap**

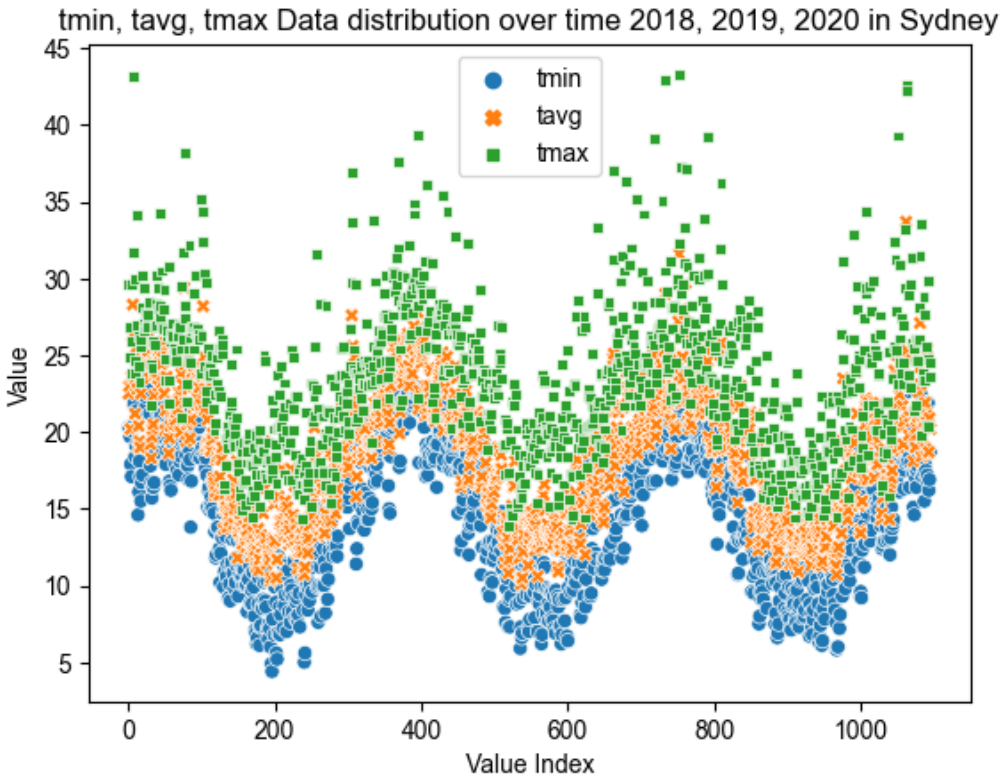
I decided to plot a heatmap to find the most correlated columns in the Los Angeles data frame for the year 2020 using all the weather paramters available.

The resultant heatmap plot displays the correlation coefficients (+1 to -1) of the columns in the data frame mentioned above. Using this, we can identify the numerical value representation of the correlation between each column. In addition, it helps to identify the most correlated columns.



2. Scatter Plot

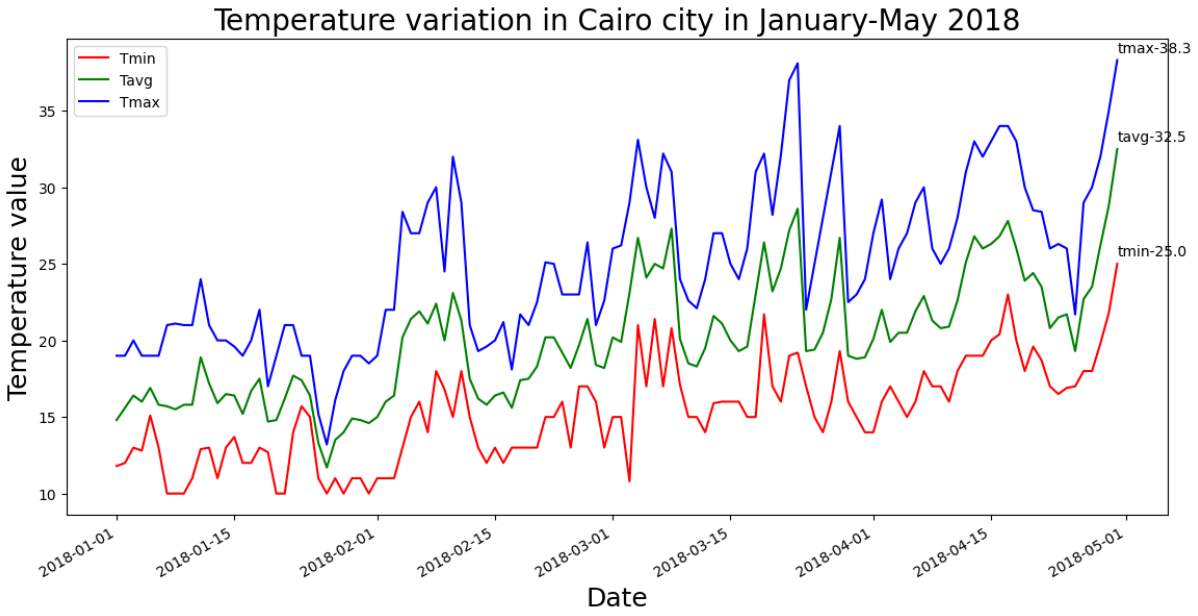
I created this plot for the continent of Australia, i.e., I analyzed the data available for Sydney. The result was a scatter plot showing the distribution of data points in 3 columns (tmin, tavg, and tmax) in a single plot for the three-year period. Using this, we can identify each region's density of data points.



3. Bonus Plot (Line Chart Simulation)

As an advanced plot, I created a line chart simulation, which shows the line plotting as a simulation with respect to time. Each data point is added to the plot separately, and a time interval is set between the two data points. The process is done by using a "for" loop. For this plot, I selected the data records from January 1st to April 30th, 2018, for Cairo city. Also, I added 3 line charts to show the variation of 'tmin', 'tavg', and 'tmax' values. Those three plots are shown in 3 different colors.

The specialty of this plot is that it shows past data as a real-time data stream. For example, the data I am using to plot was collected in 2018, but I can show it as a real-time data collection process with this plot.



Conclusion:

According to the above calculations and the data visualizations, we can identify a small effect on weather parameters due to the global warming effect. The above line plot shows this effect on the "tavg" parameter. This variation can be used as evidence on this point. Also, the minimum and maximum parameter calculations give insights into the facts.

As a future expansion, we can collect data for more years in more cities worldwide and use it for analysis. Also, we can use the other weather parameters except "tavg" for the processing, which will help identify the global warming effect on weather parameters more accurately.