Explore Weather Trends Project

Outline

Collecting data

The data was collected from the provided database. I chose the city of Edinburgh as it is where I am currently studying. Two CSV files were downloaded from the database. The first file contained the data for the global average temperatures while the second file included specific data for the city of Edinburgh. SQL queries were used to extract data from the database. They are shown below:

Global data query:

```
SELECT * FROM global_data;
```

Edinburgh data query:

```
SELECT * FROM city_data
WHERE city = 'Edinburgh';
```

Text form:

SELECT * FROM global_data;

SELECT * FROM city_data
WHERE city = 'Edinburgh';

Preparing data using Python

I chose to use Jupyter notebooks to analyze my data. First, I imported and cleaned the data from each file. This involved setting the year column as an index and dropping useless columns. The following code was used to achieve this:

Global Data

```
global_data = pd.read_csv(r'C:\Users\ibabk\Documents\Udacity Data Analyst\Project 1\Global_Data.csv')
global_data.rename(columns={'year':'Year', 'avg_temp':'Global Average Temperature'}, inplace=True)
global_data.set_index('Year', inplace=True)
global_data
```

Edinburgh Data

```
edinburgh_data = pd.read_csv(r'C:\Users\ibabk\Documents\Udacity Data Analyst\Project 1\Edinburgh_Data.csv')
drop_columns = ['city', 'country']
edinburgh_data.drop(drop_columns, axis = 1, inplace=True)
edinburgh_data.rename(columns={'year':'Year', 'avg_temp':'Edinburgh Average Temperature'}, inplace=True)
edinburgh_data.set_index('Year', inplace=True)
edinburgh_data
```

Combined dataset:

```
dataset = edinburgh_data.join(global_data, how="inner")
dataset
```

Year	Edinburgh Average Temperature	Global Average Temperature
1750	8.41	8.72
1751	8.16	7.98
1752	4.78	5.78
1753	7.51	8.39
1754	7.42	8.47
...	...	...
2009	8.64	9.51
2010	7.43	9.70
2011	8.95	9.52
2012	8.08	9.51
2013	8.39	9.61

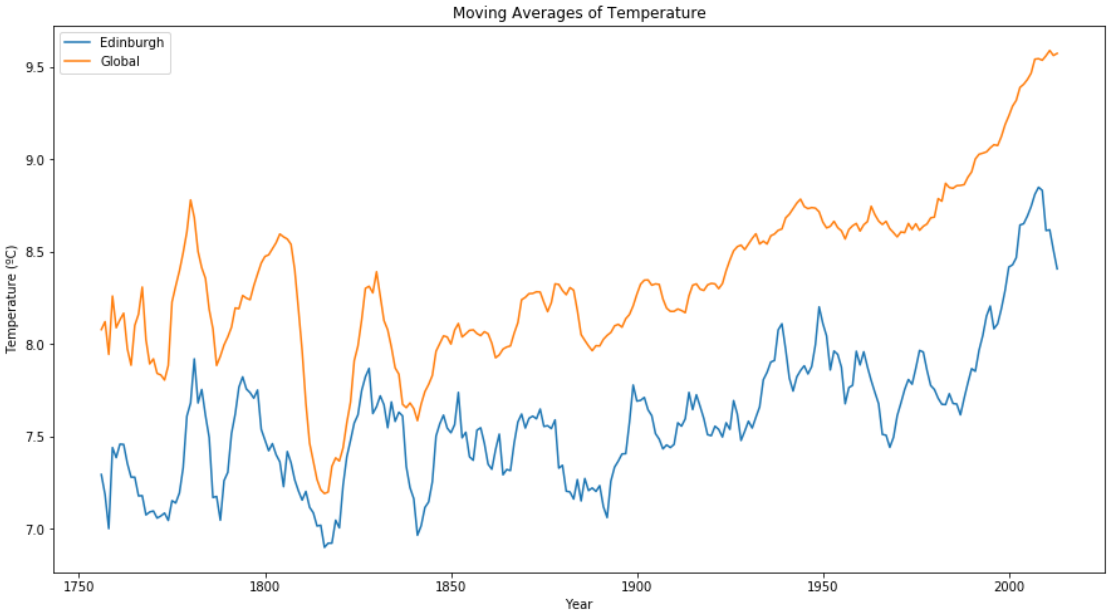
I created a separate data frame to store the moving averages of the temperatures. Then, the rolling method was used to calculate the moving average. The window was set to seven to calculate the moving average for each aggregated seven periods. The global and Edinburgh temperatures were stored in separate columns. The code is shown below:

```
moving_averages = pd.DataFrame()
moving_averages['Edinburgh'] = dataset['Edinburgh Average Temperature'].rolling(window=7).mean()
moving_averages['Global'] = dataset['Global Average Temperature'].rolling(window=7).mean()
moving_averages
```

Year	Edinburgh	Global
1750	NaN	NaN
1751	NaN	NaN
1752	NaN	NaN
1753	NaN	NaN
1754	NaN	NaN
...	...	...
2009	8.831429	9.535714
2010	8.614286	9.560000
2011	8.618571	9.588571
2012	8.508571	9.561429
2013	8.407143	9.572857

To plot the moving averages, I used matplotlib. I have set my own parameters and included the axis labels. The code and the output could be seen below:

```
moving_averages['Edinburgh'].plot(figsize = (15,8))
moving_averages['Global'].plot(figsize = (15,8))
plt.title('Moving Averages of Temperature')
plt.ylabel('Temperature (°C)')
plt.legend()
plt.show()
```



Adding more cities

I wanted to look at more than one city, and therefore I also added London's and Cardiff's average temperatures to the dataset and plotted it. The same SQL queries were used to extract the data from the database. However, to get the right data for the city of London, the country was specified in order to only get the data for London in the UK:

```
SELECT * FROM city_data
WHERE city = 'London' AND country = 'United
Kingdom';
```

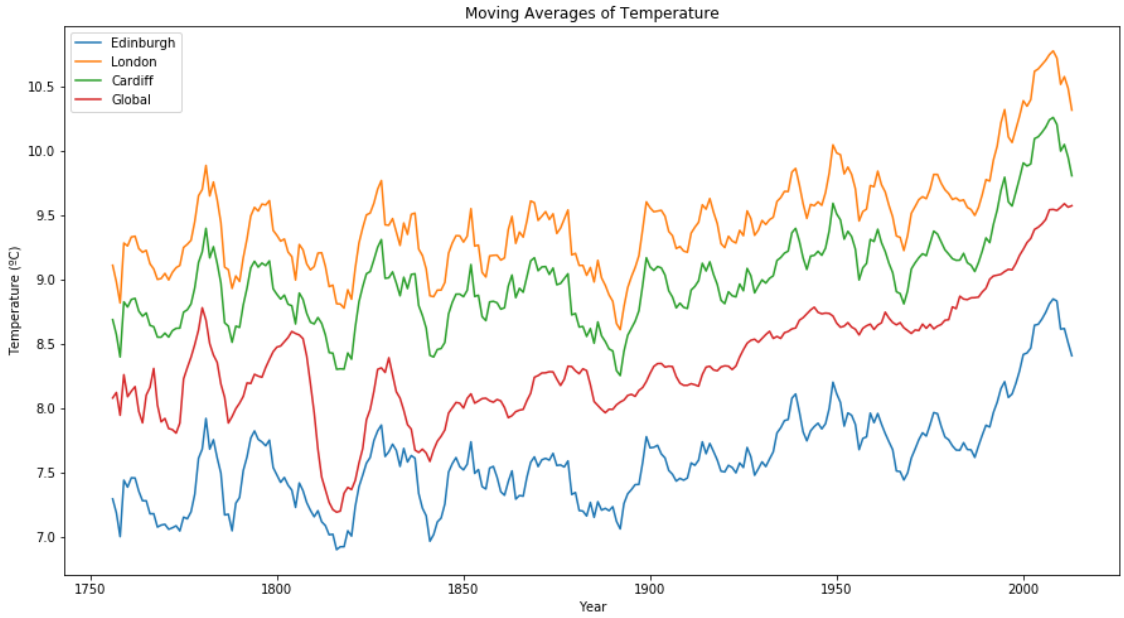
The same method to clean and aggregate the data was used. The code that is shown in the images above was used to perform this with slight changes in the variable names. Overall, the aggregated dataset data frame and moving averages data frame are shown below:

Year	Cardiff Average Temperature	London Average Temperature	Edinburgh Average Temperature	Global Average Temperature
1750	9.73	10.25	8.41	8.72
1751	9.51	9.99	8.16	7.98
1752	6.26	6.54	4.78	5.78
1753	8.97	9.42	7.51	8.39
1754	8.82	9.20	7.42	8.47
...	...	...	...	...
2009	9.90	10.52	8.64	9.51
2010	8.80	9.38	7.43	9.70
2011	10.63	11.19	8.95	9.52
2012	9.55	10.00	8.08	9.51
2013	9.57	9.91	8.39	9.61

Year	Edinburgh	Global	London	Cardiff
1750	NaN	NaN	NaN	NaN
1751	NaN	NaN	NaN	NaN
1752	NaN	NaN	NaN	NaN
1753	NaN	NaN	NaN	NaN
1754	NaN	NaN	NaN	NaN
...	...	...	...	...
2009	8.831429	9.535714	10.720000	10.204286
2010	8.614286	9.560000	10.515714	9.997143
2011	8.618571	9.588571	10.575714	10.048571
2012	8.508571	9.561429	10.485714	9.950000
2013	8.407143	9.572857	10.315714	9.805714

Data Visualisation

A line chart that shows the moving averages for the three cities and a global moving average is shown below:



The line chart above shows that the average temperatures for all cities and global average temperature seem to follow a similar trend. A general increase in average temperature from 1750 to 2013 could be observed. Similarly, the overall dataset appears to share the most notable sharp drop in average temperature around the 1820s. Nevertheless, this sharp drop seems to be the steepest for the global average temperature. It is possible that it is a consequence of missing data in that period. None of the specific data subsets ever interest or overlap each other. The average temperature in Edinburgh is constantly below the global average while the average temperatures in Cardiff and London are always above the global average, with the latter constantly being the highest average. Lastly, while it is true that all data shows a similar trend, the general path of average temperature in Edinburgh, London, and Cardiff are, indeed, very similar and have greater similarities

with each other than with the global data. The average temperature seems to be more volatile through the years analysed while all three cities experienced a significant decrease in average temperature in the 2010s while the global average temperature continued to increase in the same period.

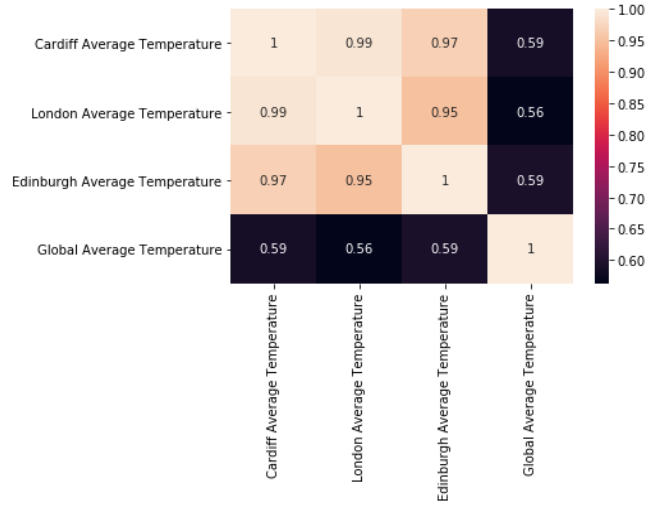
Correlation

As mentioned above, all average temperatures seem to be following a similar general trend. To investigate the relationship between the datasets further, I decided to produce a correlation matrix and a heatmap using seaborn. It is shown below:

```
dataset.corr()
```

	Cardiff Average Temperature	London Average Temperature	Edinburgh Average Temperature	Global Average Temperature
Cardiff Average Temperature	1.000000	0.991755	0.969152	0.588579
London Average Temperature	0.991755	1.000000	0.950873	0.563099
Edinburgh Average Temperature	0.969152	0.950873	1.000000	0.592681
Global Average Temperature	0.588579	0.563099	0.592681	1.000000

```
sn.heatmap(dataset.corr(), annot=True)
plt.show()
```



As can be seen from the matrix and the heatmap above, the correlation between average temperatures for the chosen cities and the global average is significantly weaker than between the cities themselves. Therefore, one could argue that the average temperature of the selected cities could not be estimated using the global average. In fact, the correlation between the cities is, indeed, very strong with all the correlation coefficients being in the high 0.9s. This is probably due to the similar geographical location of the cities as they are all within the UK. However, correlation does not mean causation and therefore, further analysis is required in order to establish whether it will be possible to use the average temperature of one of those cities to estimate the temperature of another. A regression could be used to further analyse this with variables such as the city's size, population, and urbanisation to be considered in the model.