Project 1 - Exploring Weather Trends

Importing and Cleaning Data

SQL Import Query:

Since we will be comparing the annual averge temperatures of Bangalore with the average global temperature, we should look at the the years for which both values are available. Hence, we should perform an inner join between the two tables on 'year'.

```
SELECT A.YEAR,
A.AVG_TEMP AS GLOBAL_TEMP,
B.AVG_TEMP AS BLORE_TEMP
FROM
(
  SELECT DISTINCT YEAR, AVG_TEMP FROM GLOBAL_DATA
) A
INNER JOIN
(
  SELECT DISTINCT YEAR, AVG_TEMP FROM CITY_DATA WHERE CITY = 'Bangalore'
) B
ON A.YEAR = B.YEAR
ORDER BY YEAR;
```

```
# Importing relevant libraries
import pandas as pd
import plotly.graph_objects as go
import warnings
warnings.filterwarnings('ignore')
from scipy.stats import pearsonr,spearmanr

#!pip uninstall pandas-profiling
#!pip install pandas-profiling

from pandas_profiling import ProfileReport
import matplotlib.pyplot as plt
import ipywidgets as widgets
import seaborn as sns


#Importing Data
weather_data = pd.read_csv('/content/drive/My Drive/Udacity Courses/Data Analyst/weather_data.csv')
```

Data Profile Checks

```
profile_report = ProfileReport(weather_data)
display(profile_report.to_widgets())
```

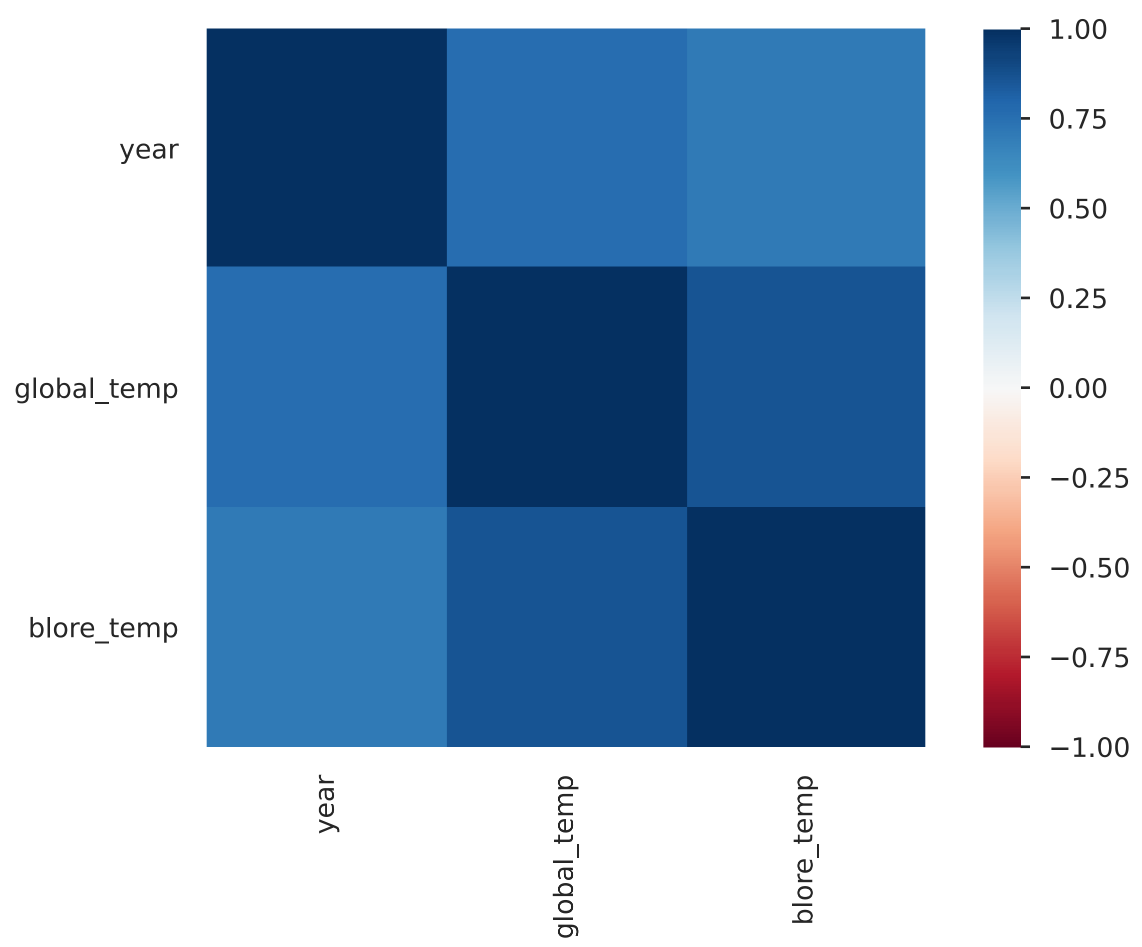
Summarize dataset: 100% 17/17 [01:41<00:00, 5.96s/it, Completed]

Generate report structure: 100% 1/1 [01:36<00:00, 96.86s/it]

Overview | Variables | Interactions | Correlations | Missing values | Sample

Toggle correlation descriptions

Pearson's r | Spearman's ρ | Kendall's τ | Phik (φk)



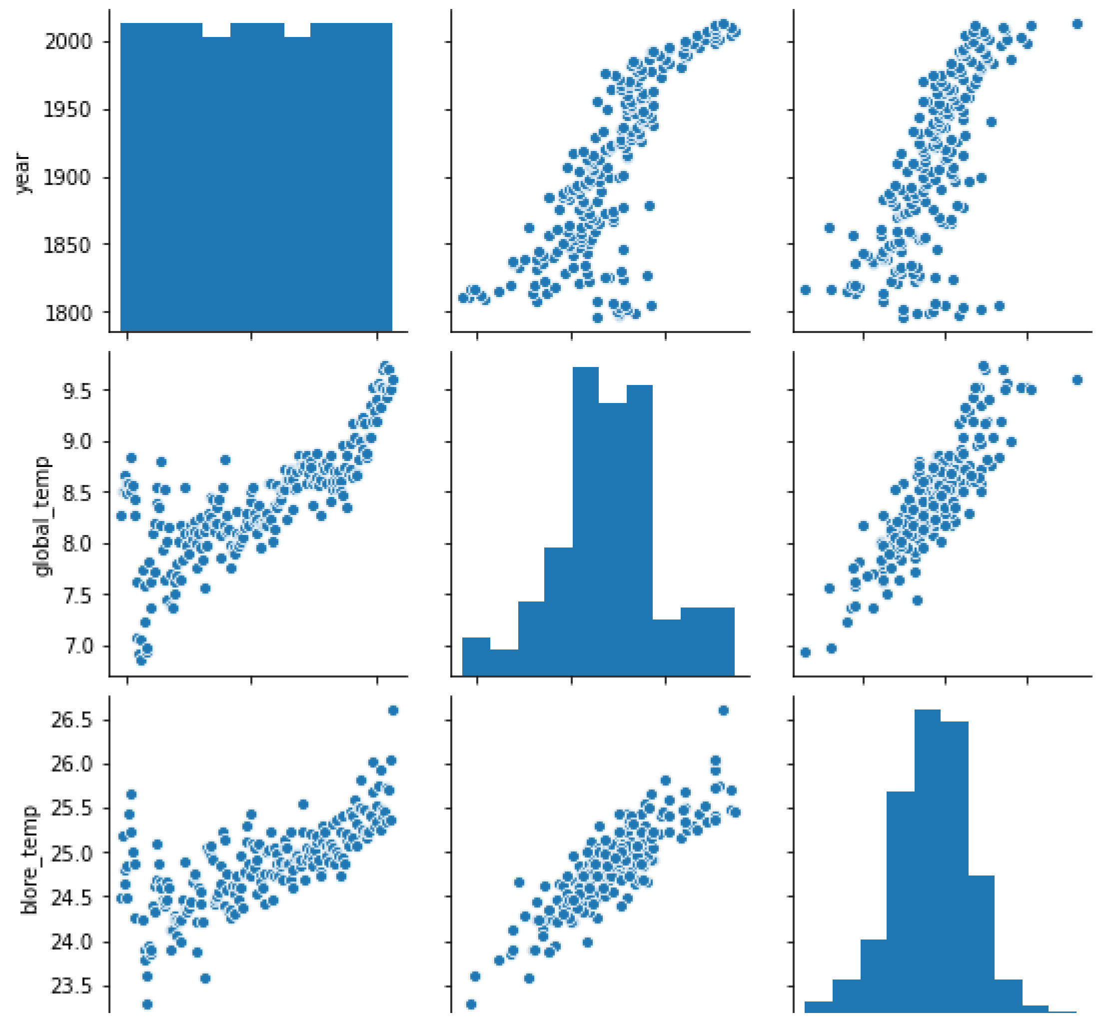
Report generated with [pandas-profiling](https://github.com/pandas-profiling/pandas-profiling).

None

Scatter Plot

```
display(sns.pairplot(weather_data))
```

<seaborn.axisgrid.PairGrid at 0x7fba9c4954a8>



Observations

Missing Data : We see that the average temperature for Bangalore is missing for 7 records/years in the data. We should explore ways to handle the missing data before calculating moving averages, since they will skew the averages.

Distribution (Scatter Plots) :

- *Global average annual temperatue across the years*:

We see that there is a linear relationship between year and global average temperature (barring a few outliers), which suggests that the average global temperature has had an increasing trend over the years.

- *Bangalore average annual temperature across the years* :

We see a linear relationship between year and local average temperature sugessting an increasing trend similar to the global temperature trends.

- *Global average annual tempaerature vs Bangalore average annual temprature*

There is a clear linear relationship between global annual tmeperature and local annual temperature based on the scatter plot. We can investigate this relationship further by looking into the correlation between the two metrics.

Correlation Grid

Given the linear relatinship between global and local temperature, based on the correlation grid, we can see a high positive correlation between global temperature and local temperature. This suggests that global temperature trends could possibly be a good indicator of local trends. We can investigate the strength of this linear relationship once we iron out the outliers using moving averages.

Handling Missing Data

Options:

1. ***Drop the rows with missing temperature and consider only overlapping years***: Since this impacts only ~3% of the records, we can choose to drop these records and go ahead and compare the moving average trends. We have 211 years of overlapping data across the two metrics (1796-2013 minus the 7 dropped years). This allows for sufficient obsbervations for comparison using moving averages while handling null years.
2. *Fill the nulls*: Since we do not have sufficient background on how weather trends are aggregated, the decision to fill the nulls will be based on unsubstantiated assumptions, which is not preferable.
3. *Filter data to after 1900*: By doing so, we limit the analysis time period to 1900 - 2013

```
weather_data_clean = weather_data.dropna(inplace=False)
```

Implemeting Moving Averages

Moving Average Calculation

```
count = None
weather_data_clean['global_rolling'] = weather_data_clean.global_temp.rolling(5).mean()
weather_data_clean['blore_rolling'] = weather_data_clean.blore_temp.rolling(5).mean()
output = widgets.Output()
int_slider = widgets.IntSlider(
    value=5,
    min=5,
    max=100,
    step=5,
    description='Step Size:',
    disabled=False,
    continuous_update=False,
    orientation='horizontal',
    readout=True,
    readout_format='d'
)
int_slider.style.handle_color = 'lightblue'
def ins_slider_handler(change):
  output.clear_output()
  weather_data_clean['global_rolling'] = weather_data_clean.global_temp.rolling(change.new).mean()
  weather_data_clean['blore_rolling'] = weather_data_clean.blore_temp.rolling(change.new).mean()
  global count
  count = change.new
  with output:
    display(weather_data_clean[['blore_temp','global_temp','blore_rolling','global_rolling']].describe())
int_slider.observe(ins_slider_handler, names = 'value')
```

```
display(int_slider)
```

Step Size: 15

Picking the moving average step size

Descriptive Statistics

Bangalore Temperature: Bangalore's temperature ranges from 23.3°C - 26.6°C with a mean tmeperature of 24.9°C. We see that the standard deviation from mean is around ~0.48°C which the suggests that most of the temperature values are clustered close to the mean.

Global Temperature: The global temperature ranges from 6.9°C - 9.7°C with a mean temperature of 8.4°C. We see that the standard deviation of ~0.51°C which similar to the Bangalore trend suggests points clustered close to the mean.

Moving Average Period Selection

After trying different values (5,10,15,20,25) of the moving average period using the Integer Slider above, we can see the that when we use 15 years, we are able to smoothen the variation without losing the key trends in the data.

```
display(output)
```

	blore_temp	global_temp	blore_rolling	global_rolling
count	211.000000	211.000000	197.000000	197.000000
mean	24.853081	8.437630	24.821218	8.395953
std	0.485181	0.516615	0.356848	0.410864
min	23.300000	6.940000	24.186667	7.736000
25%	24.530000	8.120000	24.539333	8.072667
50%	24.880000	8.430000	24.862667	8.269333
75%	25.165000	8.730000	25.033333	8.652667
max	26.610000	9.730000	25.616667	9.504000

```
# Line Plot Function
def line_plot1(df,name1,y1,title, num,name2=None,y2=None):
  fig, ax = plt.subplots()

  ax.plot(df['year'],df[y1])
  if num == 2:
    ax.plot(df['year'],df[y2])
    ax.legend([name1,name2])
  ax.set_title(title)
  ax.xaxis.set_label_text('Year')
  ax.yaxis.set_label_text('Average Temperature')
  plt.show()
```

```
weather_data_clean.to_csv('/content/drive/My Drive/Udacity Courses/Data Analyst/output.csv')
```

Longitudinal Trends of Moving Averages

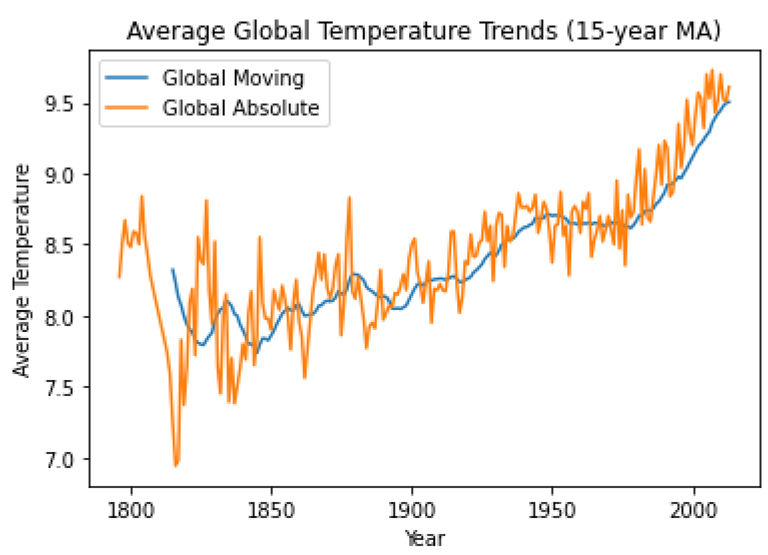
Global Trend

- As was noted earlier, we observe a general increasing trend in the average global temperatures with a steep rise starting in the 1980s through 2013.
- Let's explore the latest three 50-year windows to understand the rate of increase in temperature:

1. Window 1 (1861 - 1911): The average temperature increases by 0.22°C in 50 years, thereby increasing at a rate of **0.4%**.
2. Window 2 (1912 - 1962): The average temperature increases by 0.40°C units in 50 years, thereby increasing at a rate of **0.8%**.
3. Window 3 (1963 - 2013): The average temperature increases by 0.85°C units in 50 years, thereby increasing at a rate of **1.7%**.

The rate of increase in the latest window, Window 3 is 2.1 times Window 2 and 3.8 times Window 1. This is a clear indication of an escalation in rapidity with which the average global temperature has increased the latest 50 years.

```
display(line_plot1(weather_data_clean,"Global Moving",'global_rolling', 'Average Global Temperature Trends ('+str(count)+'-year MA)',2,"Global Absolute","global_temp"))
```



None

Local Trend

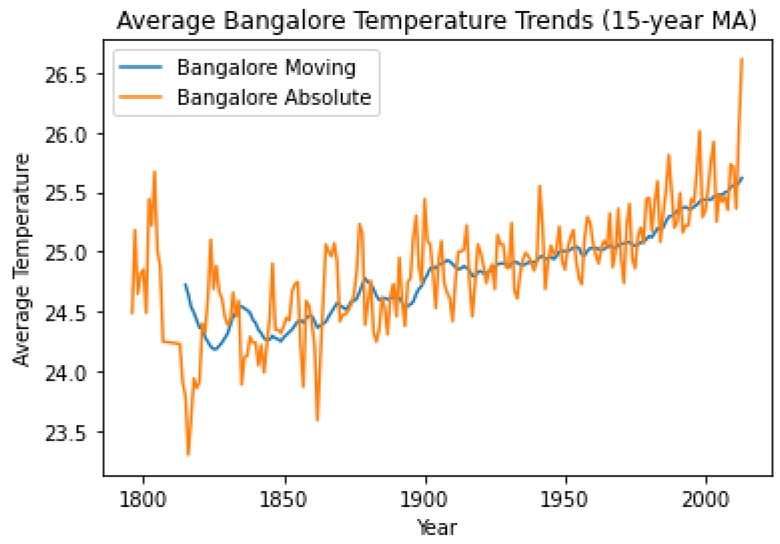
- As was noted earlier, we observe a general increasing trend in the average local temperature in Bangalore.
- Just like we did with the Global trend, let's explore the latest three 50-year windows to understand the rate of increase in temperature:

1. Window 1 (1861 - 1911): The average temperature increases by 0.45°C in 50 years, thereby increasing at a rate of **0.9%**.
2. Window 2 (1912 - 1962): The average temperature increases by 0.17°C in 50 years, thereby increasing at a rate of **0.3%**.
3. Window 3 (1963 - 2013): The average temperature increases by 0.61°C in 50 years, thereby increasing at a rate of **1.2%**.

The rate of increase in the latest window, Window 3 is 3.5 times Window 2 and 1.3 times Window 1. This suggests that the latest window has witnessed the highest rate of increase in local temperature in the 150 year timeframe.

```
display(line_plot1(weather_data_clean,"Bangalore Moving",'blore_rolling', 'Average Bangalore Temperature Trends ('+str(count)+'-year MA)',2,"Bangalore Absolute",'blore_temp'))
```

```
display(line_plot1(weather_data_clean,"Bangalore Moving",'blore_rolling', 'Average Bangalore Temperature Trends ('+str(count)+'-year MA)',2,"Bangalore Absolute",'blore_temp'))
```



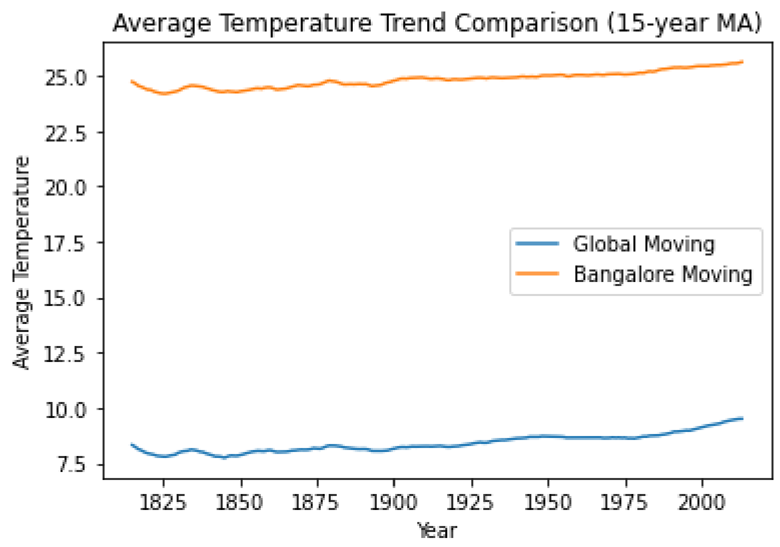
None

Global vs Local Trends

Longitudinal Trend

- We see that the average temperature in Bangalore is around the 25°C mark whereas in globally, it is close to the 9°C mark. Due to this difference in scale, the line graph does not provide a clear comparison of the temperature trends.
- However, it does confirm what we had noted using the scatter plot, which is that we see a linear relationship between the two metrics. We see similar increasing trends both locally and globally across the years.
- We can also imply from the rate of increase observations that for both metrics, the latest 50-year window shows the highest rate of increase in average temperature.

```
display(line_plot1(weather_data_clean,"Global Moving",'global_rolling', 'Average Temperature Trend Comparison ('+str(count)+'-year MA)',2,"Bangalore Moving", 'blore_rolling'))
```



None

Correlation

To better understand the strength of linear relationship between the global and local trends, we must look at the correlation coefficient.

The below output indicates that the two metrics are highly postively correlated which means that global trends are largely indicative of local trends.

```
# Correlation between Global and Local average annual temperatures
corr_abs, _ = pearsonr(weather_data_clean['global_temp'],weather_data_clean['blore_temp'])
print("Correlation between Global and Local average annual temperatures: %.2f"%(corr_abs))

# Correlation between Global and Local 25-year moving average temperatures
corr_rolling, _ = pearsonr(weather_data_clean[weather_data_clean['global_rolling'].notna()]['global_rolling'],weather_data_clean[weather_data_clean['blore_rolling'].notna()]['blore_rolling'])
print("\nCorrelation between Global and Local 25-year moving average temperatures: %.2f"%(corr_rolling))
```

```
Correlation between Global and Local average annual temperatures: 0.86

Correlation between Global and Local 25-year moving average temperatures: 0.97
```

Next Steps:

Further, we can look into the predictive ability of each of the metrics for the other, i.e can we predict local trends using global trends or vice versa by doing a Linear Regression Analysis.