Exploring Weather Trends Project

Steps taken:

- Data Extraction
- Data Wrangling
- Data Exploration
- Data Visualization

Data Extraction

Used SQL to extract data from database schema :

Query 1: Global Data

```
SELECT * FROM global_data
```

Query 2: San Francisco Data

```
SELECT year, avg_temp FROM city_data WHERE
country='United States' AND city='San
Francisco'
```

Query 3: San Diego Data

```
SELECT year, avg_temp FROM city_data WHERE
country='United States' AND city='San Diego'
```

Data Wrangling

I downloaded the 3 csv files in jupyter notebook and performed the rest of the process in that notebook.

```
import pandas as pd
import numpy as np
%matplotlib inline
df_sd = pd.read_csv('CSV/SanDiego.csv')
df_sf = pd.read_csv('CSV/SanFrancisco.csv')
df_g = pd.read_csv('CSV/global.csv')
```

I found that the global data starts from 1750 while the San Francisco and San Diego data start from 1849. To make sure the 3 data sources matched I merged them together using the pandas merge.

```
df = df_g.merge(
    df_sd.merge(df_sf, left_on='year', right_on='year'),
    left_on='year', right_on='year'
)
df.columns = ['Year', 'Global', 'San_Diego', 'San_Francisco']
```

Data Exploration

For a better visualization of data points I calculated the 10 year moving average for global temperature, San Francisco, and San Diego using pandas. I stored the 10 year moving averages in a second data frame called 'df_moving.'

```
columns = df.columns[1:]
df_moving = df.copy()
for col in columns :
    df_moving[col] = np.array([df.iloc[(i-19):(i+1)].mean()[col] for i in range(df.shape[0])])
```

Using the built-in pandas function I created a correlation matrix for the 3 different temperature averages.

```
df[df.columns[1:]].corr()
```

	Global	San_Diego	San_Francisco
Global	1.000000	0.505094	0.536038
San_Diego	0.505094	1.000000	0.854621
San_Francisco	0.536038	0.854621	1.000000

Correlation Coefficient Global vs. San Francisco = 0.536038
Correlation Coefficient Global vs. San Diego = 0.505094
Correlation Coefficient San Francisco vs. San Diego = 0.854621

Using the same built-in pandas function I created a correlation matrix for the 3 different temperature averages, but this time on the moving averages. I noticed the moving averages increased correlation.

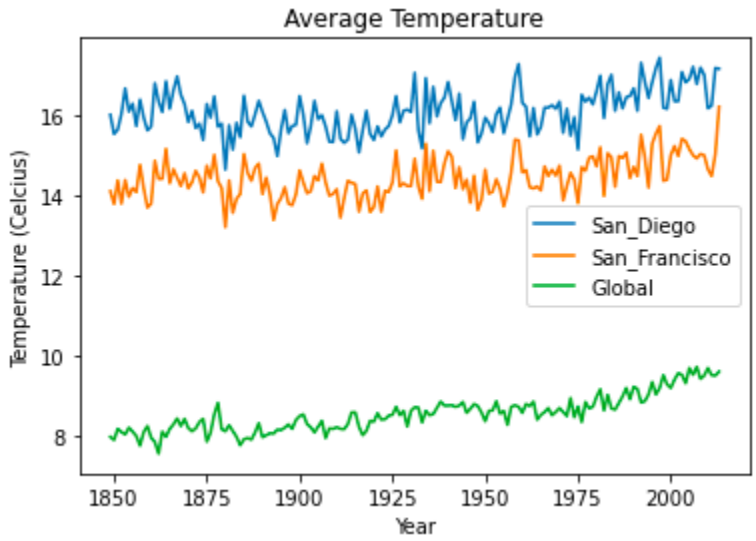
```
df_moving[df_moving.columns[1:]].corr()
```

	Global	San_Diego	San_Francisco
Global	1.000000	0.818355	0.861431
San_Diego	0.818355	1.000000	0.968903
San_Francisco	0.861431	0.968903	1.000000

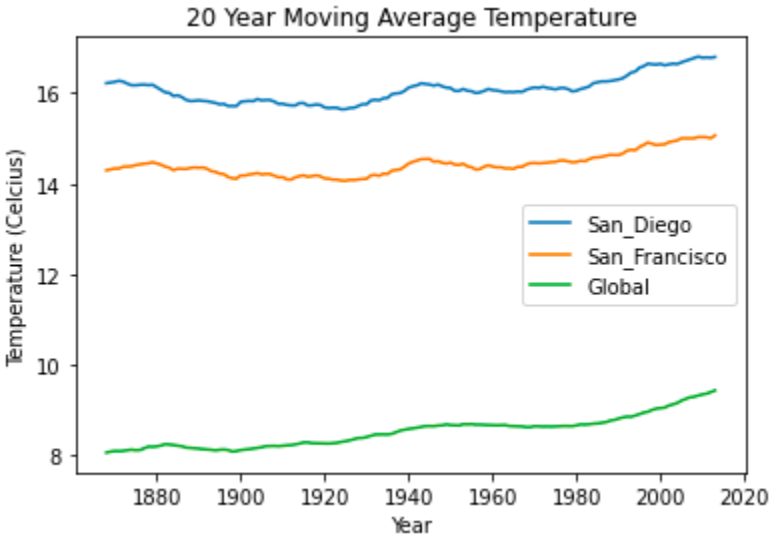
Correlation Coefficient Global vs. San Francisco = 0.818355
Correlation Coefficient Global vs. San Diego = 0.861431
Correlation Coefficient San Francisco vs. San Diego = 0.968903

Data Visualization

Using the built-in pandas dataframe plotting function, I plotted a line graph with the average temperature values on the Y-axis and the year on X-axis for Global, San Francisco and San Diego.



Since the trend is a little bit difficult to see, I did the 20 year moving average graph too, using the same method on 'df_moving.' I tried 5 yr and 10 yr, but both did not have enough smoothness.



Observations

- Global Warming is definitely confirmed by the data. Even in the regular chart (more so in the moving average one) there is a clear upward trend with global temperature. Both San Francisco and San Diego show a small upward trend, not really that visible on the regular chart but more-so on the moving average chart.
- As expected due to proximity of the two there was a huge correlation between San Francisco and San Diego. With a correlation coefficient of 0.85 for the regular graph and 0.97 for the moving graph. We can also see on the regular graph San Francisco and San Diego share a lot of the same dips and increases.
- Seeing from the line graph and later on when I inspected the data further and between 1940 and 1950, there was a small general dip in global temperatures. Similarly while its not as visible on the normal graph, we can see on the moving average graph that there seems to be a small downward trend in both San Francisco and San Diego from 1940 to 1950. Aside from that decade, for the most part global temperatures have consistently been on the rise.
- San Diego has consistently been warmer than San Francisco by around 2-3 degrees celsius and San Francisco has been around 6 degrees celsius above global temperatures.