Udacity Data Analyst Nano Degree Program

Mohanad Salem

January 2023

Project 1 Explore Weather Trends

Introduction:

In this project, we will analyze local and global temperature data and compare the temperature trends of New York to overall global temperature trends.

Goals:

1) Check for the country and city.
2) extract the city level data. Export to CSV.
3) extract global data. Export to CSV.

Tools:

1) SQL was used to extract all the necessary data from the database.
2) Python was used to open the csv files, manipulate the moving average and plot the line chart.

Practical:

STEP 1:

All data were selected from city_list table where the country and the city are 'United States' and 'New York' respectively, to make sure that both exist in the table.

Input

HISTORY MENU

SCHEMA

city_data

city_list

city

country

global_data

```
SELECT *
FROM city_list
WHERE country = 'United States' AND city = 'New York'
```

EVALUATE

Output 1 results

Download CSV

city	country
New York	United States

STEP 2:

In step 2 all data were selected from city_data where the country and the city are ‘United States’ and ‘New York’ respectively, to extract the necessary information we are looking for.

Note: 4 columns were found, but we only need the year and avg_temp columns.

Input

HISTORY MENU

SCHEMA

city_data

city_list

city

country

global_data

```
SELECT *
FROM city_data
WHERE country = 'United States' AND city = 'New York'
```

Success!

EVALUATE

Output 271 results

Download CSV

year	city	country	avg_temp
1743	New York	United States	3.26
1744	New York	United States	11.66
1745	New York	United States	1.13
1746	New York	United States	
1747	New York	United States	
1748	New York	United States	
1749	New York	United States	
1750	New York	United States	10.07

Step 3:

In step 3 global average temp, New York temp and the year were extracted after we joined the two tables together on the same years. Then I downloaded the CSV file.

Input

HISTORY MENU

SCHEMA

city_data

city_list

city

country

global_data

```
SELECT g.year, g.avg_temp global_temp , c.avg_temp Newyork_temp
    FROM global_data g
        JOIN city_data c
            ON c.year = g.year
                WHERE city = 'New York'
```

Success!

EVALUATE

Output 264 results

Download CSV

year	global_temp	newyork_temp
1750	8.72	10.07
1751	7.98	10.79
1752	5.78	2.81
1753	8.39	9.52
1754	8.47	9.88
1755	8.36	6.61
1756	8.85	9.94
1757	9.02	8.89

Step 4:

In this step I used python to plot the line chart. Using google sheets or excel would have been easier, but struggling with python now will help develop more skills in the future.

First, I had to import Pandas and Matplotlib libraries.

Second, I used Pandas to read the csv file from its PATH.

Third, I calculated 10 year moving average for both ‘newyork_temp’ and ‘`global_temp`‘ columns in the data frame using “.rolling(10).mean()“.

Fourth, I found some null results appearing in the data frame, so I had to use a function called dropna to eliminate all the null values.

Fifth, I specified the x-axis to be the year column and y-axis to be the temperature for both New York and the Global, using “.plot()”.

Sixth, the legends were adjusted to be "New York 10-year moving average temperature", "global 10-year moving average temperature" to make it more understandable to the audience.

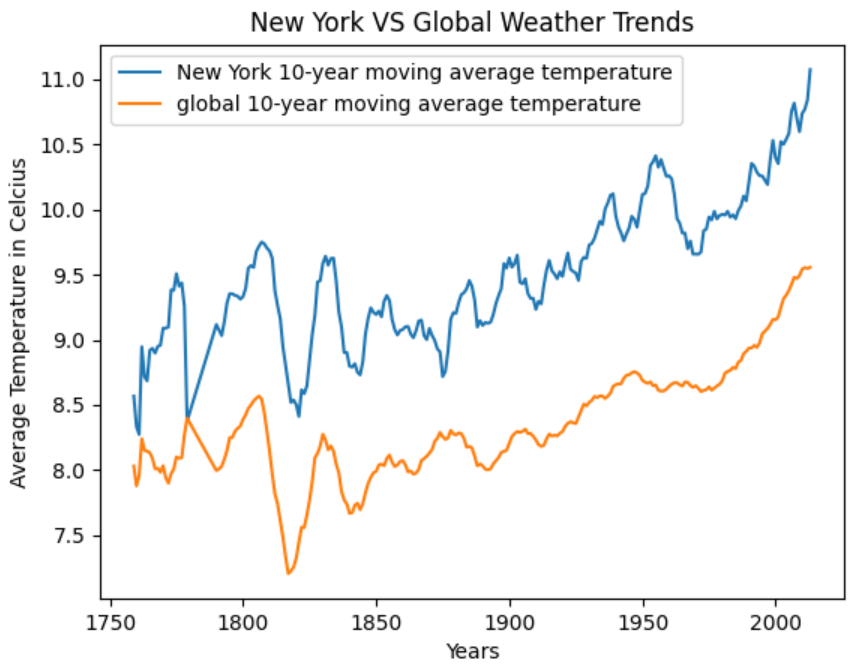
Seventh, I have labeled the x-axis, y-axis and have given a title to the chart using set_xlable(), set_ylable() and set_title() functions respectively.

Lastly, I have displayed the visualization of the plot to observe the line chart using ".show()" function in matplotlib.

```
Get Started    SELECT Untitled-1    project_1.py
D: > Python > Project_1 > project_1.py > ...
import pandas as pd
import matplotlib.pyplot as plt

# addding the csv file path
project1_csv_path= "D:\Python\Project_1/results.csv"

# reading the csv file
project1_file= pd.read_csv(project1_csv_path)
project1_file= project1_file.sort_values('year')

# calculating a moving average of 10 years
project1_file['NYC_moving_average_10']=project1_file['newyork_temp'].rolling(10).mean()
project1_file['global_moving_average_10']=project1_file['global_temp'].rolling(10).mean()

# eliminating the nulls
project1_file.dropna(inplace=True)

# selecting the x and y axis for the plot
ax = project1_file.plot(x='year', y=['NYC_moving_average_10', 'global_moving_average_10'])

# Adjusting the legend
plt.legend(["New York 10-year moving average temperature", "global 10-year moving average temperature"], loc ="upper left")

# adding xlable, ylabe and title for the plot
ax.set_xlabel('Years')
ax.set_ylabel('Average Temperature in Celsius')
ax.set_title('New York VS Global Weather Trends')

# here I am just displaying the info on the terminal to observe if there is any null values
project1_file.info()

# print the dataframe
print(project1_file)

# show the visualization of the dataframe
plt.show()
```

Observations:

1) New York temperature was always higher than the global temperature almost all the time.
2) New York temperature started at average below 9 degrees and ended above 11 degrees.
3) Global temperature started at an average below 8 degrees and ended at an average of 9.5 degrees.
4) The lowest temperature for the global tends to be at an average of 7.2 degrees, while the lowest temperature for New York city tends to be 8.2 degrees as an average.
5) The highest temperature for the global tends to be at an average of 11.1 degrees, while the highest temperature for New York city tends to be 9.5 degrees as an average.
6) Both global and New York are getting hotter over time.
7) Looks like the average differences between the global temperature and New York temperature are getting bigger over time.



Considerations:

1) x-axis has the label “years”.
2) y-axis has the label “Average Temperature in Celsius”.
3) The chart has a title.
4) Global and New York 10-year moving average temperature lines have 2 different colors.

References:

https://towardsdatascience.com/moving-averages-in-python-16170e20f6c

https://www.geeksforgeeks.org/how-to-calculate-moving-average-in-a-pandas-dataframe/