UDACITY- PROJECT 1: EXPLORE WEATHER TRENDS

PROJECT OVERWIEV

In this project I am analyzing and observing global weather data (temperatures) and data in selected cities (Ljubljana and Helsinki) from years 1849 to 2013. Data was retrieved from a Udacity database, using SQL queries and stored locally in csv files. In order to observe long term trends, I calculated 5-year moving averages for all datasets.

This report can also be found on GitHub <https://github.com/AndrejaCH/Weather_Trends>

DATA ANALYSIS PROCESS

**QUESTIONS**

* Is local city hotter or cooler on average compared to the global average?
* Has the difference been consistent over time?
* How do the changes in local city’s temperatures over time compare to the changes in the global average?
* What does the overall trend look like?
* Is the world getting hotter or cooler?
* What is the correlation coefficient?

**DATA WRANGLING**

Below are listed steps in data analysis wrangling process.

***Retrieving data form database using SQL queries and exporting data to csv file.***

|  |  |  |
| --- | --- | --- |
| **-- Select the city list** | **-- select data for specific cities** | **-- select all data form global\_data** |
| SELECT \* FROM city\_list | SELECT \* FROM city\_data WHERE city = 'Ljubljana'; | SELECT \* FROM global\_data |

***Data Wrangling in Excel***

1. Exported data from csv files to xlsx files, using copy sheet to new file in order to perform calculations and creating visuals.
2. Used `VLOOKUP` formula to gather data on the same worksheet.
3. Used `IFERROR` formula to handle missing values, replacing them with string “NULL” to avoid overwriting 0 (zeros).
4. Handling missing values:

If I want to compare the temperatures, I need data for all cities and global data. For this reason, I decided to filtered out missing values and keep only data that has values for all selected cities and global data. Filtered out missing data and copy-paste selected data to new worksheet in order to work on data without missing values (alternatively performing VLOOKUP formula on the data that has “the least data points” – in this case the null values are not transferred).

***Data Wrangling with Python in Jupyter Notebook***

1. Read in data from excel file.
2. Dropped unnecessary columns ‘city’ and ‘country’.
3. Dropped null values, using dropna() function.
4. Checked cleaned dataset, ensuring is ready for performing analysis.

**EDA & FEATURE ENGINEERING**

***What's the correlation coefficient?***

A great way to explore data, finding patterns and building intuitions is uncovering complex and unknown relationships between variables. Calculating and visualizing relationships between variables is a great way to get familiar with the data.

***Calculating correlation coefficient***

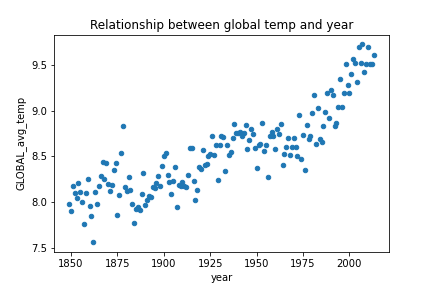
I calculated correlation coeficient between year & global temperature and correlation between global average temperatures and local city temperatures and observed what kind of correlation (if any) exists. Calculation for correlation coefficient can be calculated with Excel or Python.

* ***Calculation in excel: Using Data Analysis tool from Analysis tab***
* ***Calculation in Python: Using Pandas corr() function***

Correlation coefficient between year and global temperature is 0.86. This value indicates strong positive correlation, meaning with the year change there is an increase in the temperature.

Correlation coefficient between local city temperature and global temperature is 0.62. This value indicates positive correlation, but it is not a strong correlation.

***Visualizing Relationships***

****Graphs below represent correlation between variables in form of scatter plots and heat map.

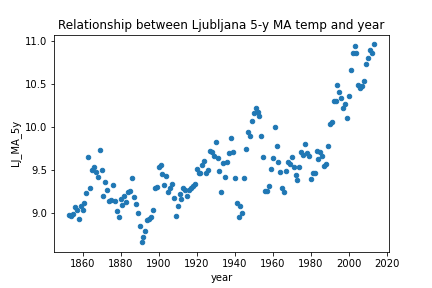
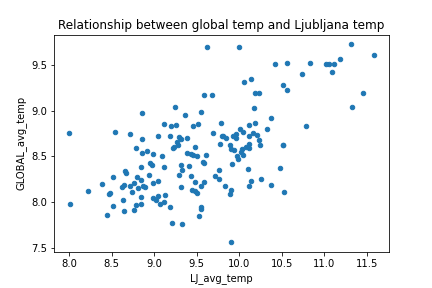
****

Figure 1: Relationship between Figure 2: Relationship between 5y Figure 3: Relationship between global

global temp. and year. moving average and year. Temp and city Ljubljana temp.

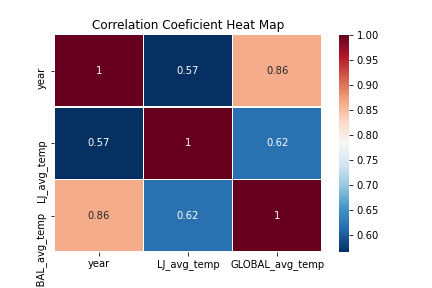
****

Figure 4: Correlation Coefficient Heat Map

**Calculating Moving average**

In general, the moving average smoothens the data and is mostly used with time series data to capture the short-term fluctuations while focusing on longer trends. Besides weather reports moving average is widely used in stock prices, gross domestic products, employment, etc. (1).

* ***Calculating 5-year moving average in Excel - steps***

Calculated average for first 5 years, using AVERAGE() function

Populate the formula for the entire column, by selecting the cell with the formula, moving the mouse to the bottom-left corner and double click.

Formula used:

*=AVERAGE(B2:B6)*

* ***Calculating 5-year moving average in Python – steps***

Formula for moving average for this project was created with the help of “datacamp website” (1). I used Pandas build in function “rolling window” and chain the function with mean() function. Parameter window was set to 5. Columns were selected with Pandas iloc method.

Formula used:

*df['LJ\_MA\_5y'] = df.iloc[:,1].rolling(window=5).mean()*

**OBSERVATIONS, ANSWERING QUESTIONS & DRAWING CONCLUSIONS**

* **Is local city hotter or colder on average compared to the global average?**
  + Observation from the chart:

From the line charts (Figure 5, Figure 6, Figure 7, Figure 8) I can observe that local city Ljubljana is hotter than global average, while local city Helsinki is cooler than global average. However, the temperatures are rising in both cities and so do global average temperatures.

* + Observation from descriptive statistics:

From descriptive statistics I can observe that Ljubljana is 1.02°C hotter and Helsinki is 4.10 °C cooler than global average.

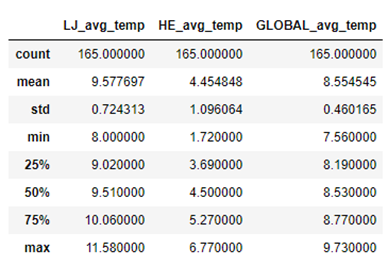


Figure 5: Descriptive Statistics for Average global Temperature

and Average Temperatures in Ljubljana and Helsinki.

* **How do changes in your city’s temperature over time compare to the changes in the global average?**

From the line charts (Figure 6, Figure 7, Figure 8, Figure 9) I can observe that temperature trend in local cities globally is similar. The fluctuation of temperatures in local cities is more obvious than global average temperatures.

* **What does the overall trend look like? Is the world getting hotter or cooler?**

From the line chart (Figure 6) and trend line for the global temperature average I can see that overall positive trend, meaning that temperatures are rising and the world’s getting hotter. The same observation we can see from line chart (Figure 7). City with cooler temperatures than global average is also getting hotter.

* **Has the trend been consistent over the last few hundred years? Has the difference been consistent over time?**

The temperatures are not consistent over time. Although, the overall trend is positive, I can observe fluctuation over the years. Observations from line charts (Figure 6, Figure 7, Figure 8, Figure 9). From scatter plots (Figure 1 and Figure 2) we can see almost exponential rise of the temperature in the last 25 years.

COMMUNICATION & VISUALIZATIONS

* **What were key considerations when deciding how to visualize the trends?**
  + Choosing the right chart type: line chart is best choice for continuous data, scatter plots are best to show correlation between two variables.
  + Ensuring that visualizations clearly communicate observations, without extra explanation.
  + Ensuring chart has all components (labels, title, legend).
  + Other visual components (colors, x and y axis intervals)

***VISUALS – IN EXCEL***

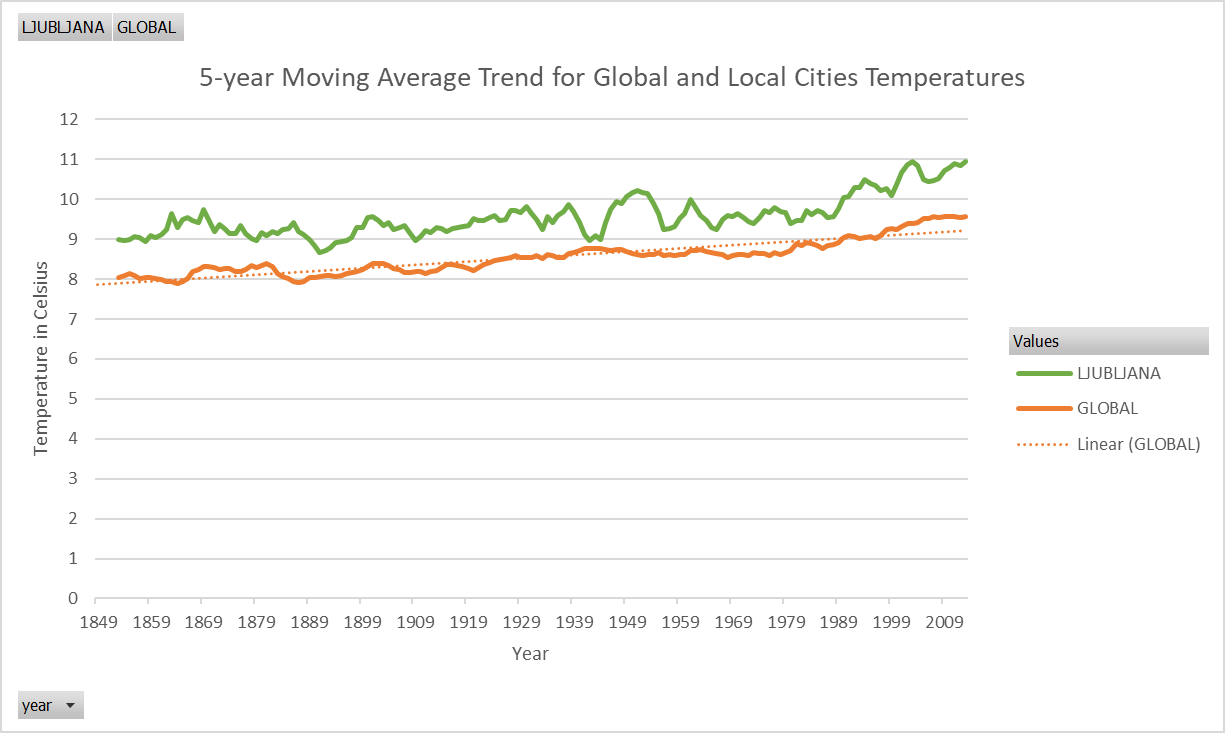


Figure 6: Line chart 5-year moving average Trend for Global and Local City Temperatures

Figure 7: Line chart 5-year moving average Trend for Global and Local Cities Temperatures

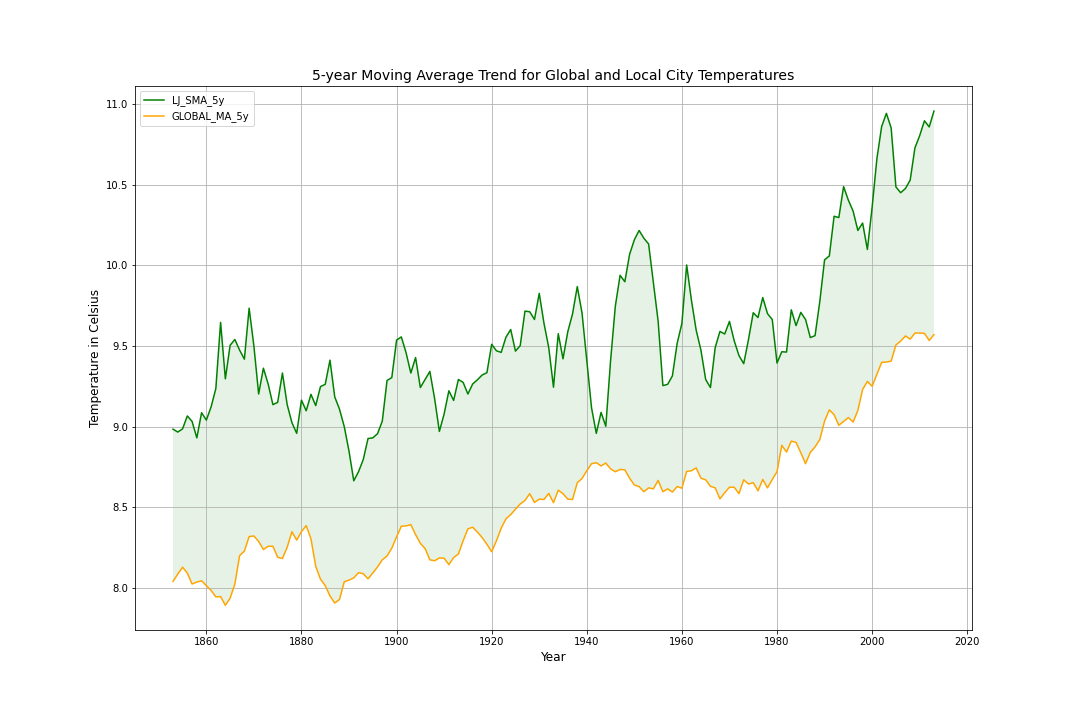
***VISUALS – IN PYTHON***

Figure 9: Line chart 5-year moving average Trend for Global and Local City Temperatures

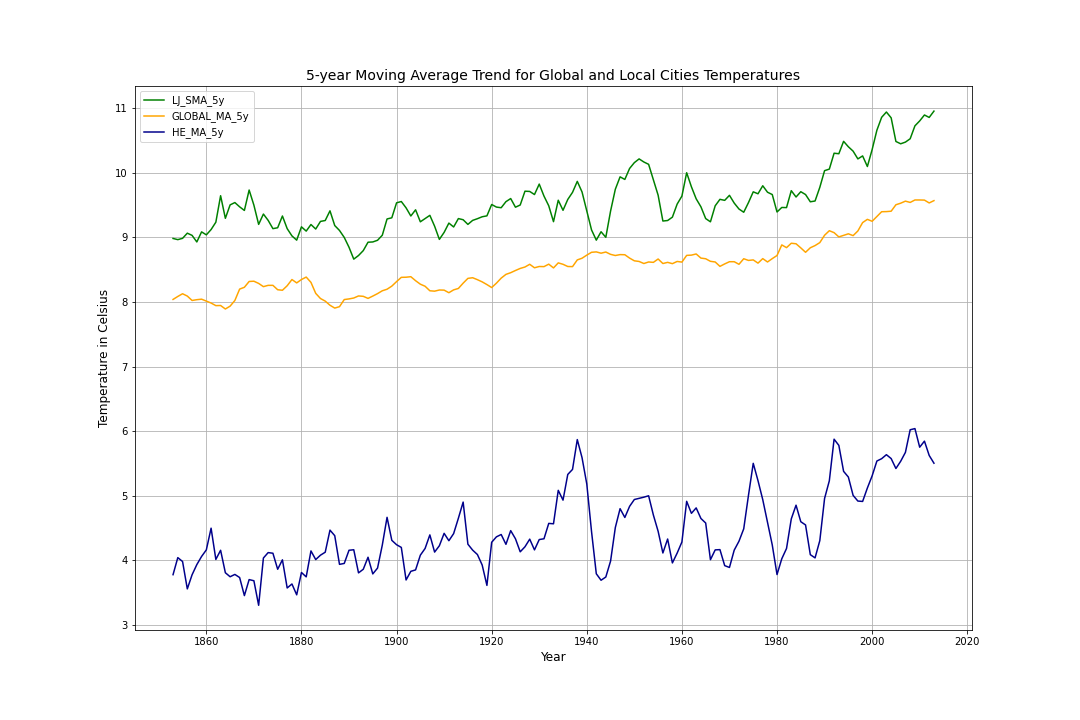


Figure 10: Line chart 5-year moving average Trend for Global and Local Cities Temperatures

REFERENCES

(1) DataCamp: <https://www.datacamp.com/community/tutorials/moving-averages-in-pandas>, Web 3/22/2021

(2) Statistics How To: <https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/>

(3) Towards Data Science: <https://towardsdatascience.com/pearson-coefficient-of-correlation-explained-369991d93404>

(4) Statistics by Jim: <https://statisticsbyjim.com/basics/correlations/>

(5) Data School: <https://dataschool.com/fundamentals-of-analysis/correlation-and-p-value/>

(6) Optimozely: <https://www.optimizely.com/optimization-glossary/statistical-significance/#:~:text=Statistical%20Significance%20Definition&text=A%20result%20of%20an%20experiment,a%20given%20statistical%20significance%20level.&text=It%20also%20means%20that%20there,that%20you%20could%20be%20wrong>.