PROJECT I

EXPLORE WEATHER TRENDS

APRIL 18, 2020

DATA ANALYST NANODEGREE PROGRAM - UDACITY

Kolin Schunck
Hamburg, Germany

PROJECT OBJECTIVES, TOOLS & KEY CONSIDERATIONS

The project "Explore Weather Trends" aims at analyzing the development of global average temperatures vis-á-vis with local average temperatures of Hamburg (Germany). The results are visualized and subsequently interpreted to derive implications of the comparison. The word "global" and "worldwide" are used interchangeably in this paper.

The **OBJECTIVES** are:

(1) to **extract data** from the database provided by Udacity using SQL queries and export to a CSV file.
(2) to **visualize** the data, specifically, comparing the 10-year moving average of temperatures between Hamburg (Germany) with the worldwide average.
(3) to **discuss** and **interpret** the findings.

In order to meet the objectives, the following **TOOLS** have been used:

(1) **SQL** for data extraction from the database.
(2) **EXCEL** for storing and saving the extracted data.
(3) **TABLEAU** for calculating the moving average of the temperatures and turning the data into a visualization format.

The following **KEY CONSIDERATIONS** are taken before data visualization:

(1) **Color:** both lines are shown with a different color. I chose green for local data and a light grey for global data. Further, values for local data were considered first, and the line moved before the global data line.

(2) **Axis:** I labeled the Y-axis in Celsius degrees and formatted the starting point to zero in order not to manipulate the view and validity of the data. Further, the X-axis only shows the years, but no title, which I did intentionally, to reduce distraction and focus on only the relevant message I want to communicate with the chart.

(3) **Title:** the title incorporates the most crucial information of the chart, namely: 10-year moving average temperatures, the local name of the city (highlighted in green), and compared with global data (highlighted in light grey). The highlighted names in the title correspond to the lines in the chart, intending to avoid double annotations in the chart, reduce distractions and leave out a legend.

(4) **Tool:** the use of Tableau enabled adding interactive elements to the graph for a better user experience. For example: the user gets more detailed information for each year by hovering over the chart or clicking on the lines.

(5) **Calculations:** I calculated the 10-year moving average to account for smoother lines, which enables the user to see a trend. Further, I calculated the Pearson correlation coefficient to assess whether or not there is a positive correlation between the data and how strong the relationship is.

OBJECTIVE I: DATA EXTRACTION USING SQL QUERIES

1. **Step:** To see if the city of Hamburg in Germany is listed in the database

   SQL Query:

```
SELECT *
FROM city_list
WHERE city LIKE 'Ham%' AND country LIKE 'Ger%'
```

2. **Step:** Combine the two tables (city_data, global_data) to create one dataset

   Explanation: Both data tables (city_data, global_data) contain the same column 'avg_temp'. I changed this column to 'avg_temp_city' for city_data and 'avg_temp_global' for global_data to have distinct columns.

   SQL Query:

```
ALTER TABLE city_data RENAME avg_temp to avg_temp_city
ALTER TABLE global_data RENAME avg_temp to avg_temp_global
```

3. **Step:** Query to extract city level and global data in one file

   SQL Query:

```
SELECT global_data.year, global_data.avg_temp_global,
city_data.avg_temp_city
FROM global_data JOIN city_data
ON global_data.year = city_data.year
WHERE city LIKE 'Hamburg' AND country LIKE 'Germany'
```

The three-step approach enabled me to create three columns with the data of temperature per year for Hamburg (Germany) and worldwide. This approach helped to eliminate one step, namely merging two datasets after downloading the files. The file is saved in the format of .xls to be used for data analysis.

OBJECTIVE II: VISUALIZE THE DATA USING TABLEAU

The dataset was uploaded to Tableau. First, I had to manually adjust the field "year" to a date field, as Tableau classified it as a measure (general number). After correctly adjusting the field, I moved to a worksheet and dragged "Year" to columns and the respective measures "avg_temp_city" and "avg_temp_global" to the rows shelf. Second, I chose a dual axis for "avg_temp_global" to display the result in a single line chart. I synchronized the axis and included zero on the y-axis as the starting point. Third, I calculated the moving average for ten years. I decided on the ten-year moving average given the length of the time, covering >200 years. The function to calculate the 10-year moving average is as follows:

10-year moving average calculation:

for Hamburg: **Window_AVG(Sum([avg_temp_city]), -10, 0)**

for the World: **Window_AVG(Sum([avg_temp_global]), -10, 0)**

Lastly, I formatted the axis (x and y), title, and legend. I intentionally did not include a legend and X-axis description because of obviousness and reduction of complexity.

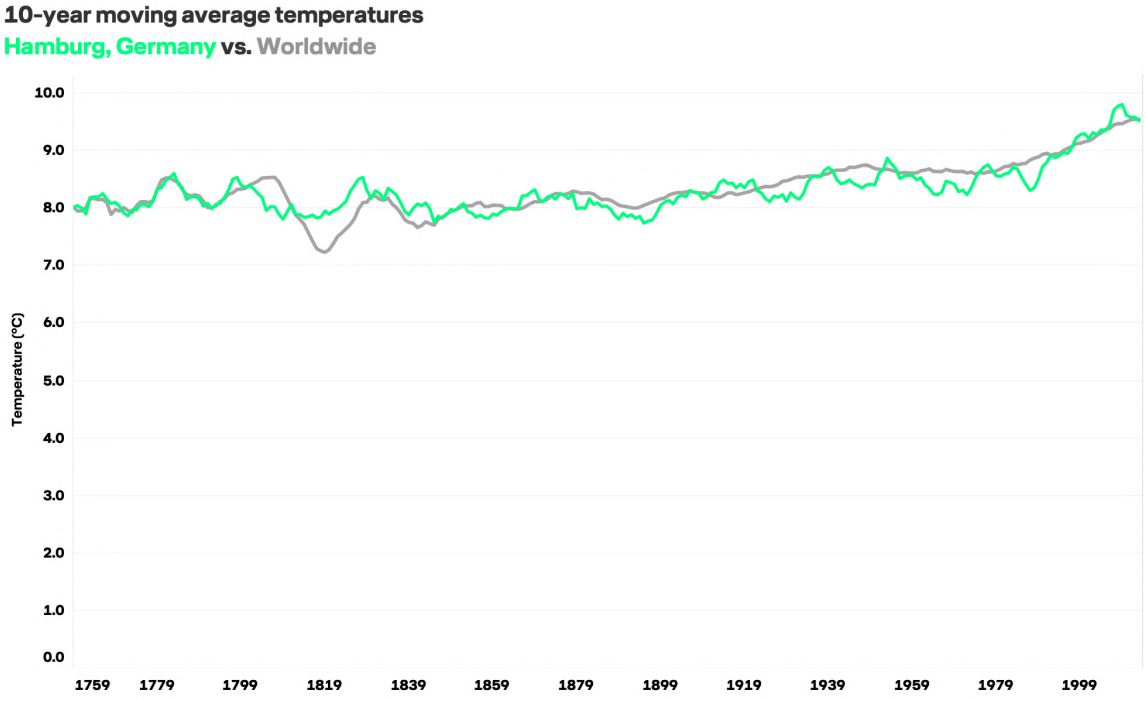

Figure 1: 10-year moving average temperature development comparison Hamburg vs. worldwide

Furthermore, I was curious about the average temperature development of my favorite cities in the world. Hence, I repeated the SQL query steps for the additional cities, as outlined on page 4. Following, I transferred the data to Tableau for visualization purposes and followed the steps as described on the previous page 5.

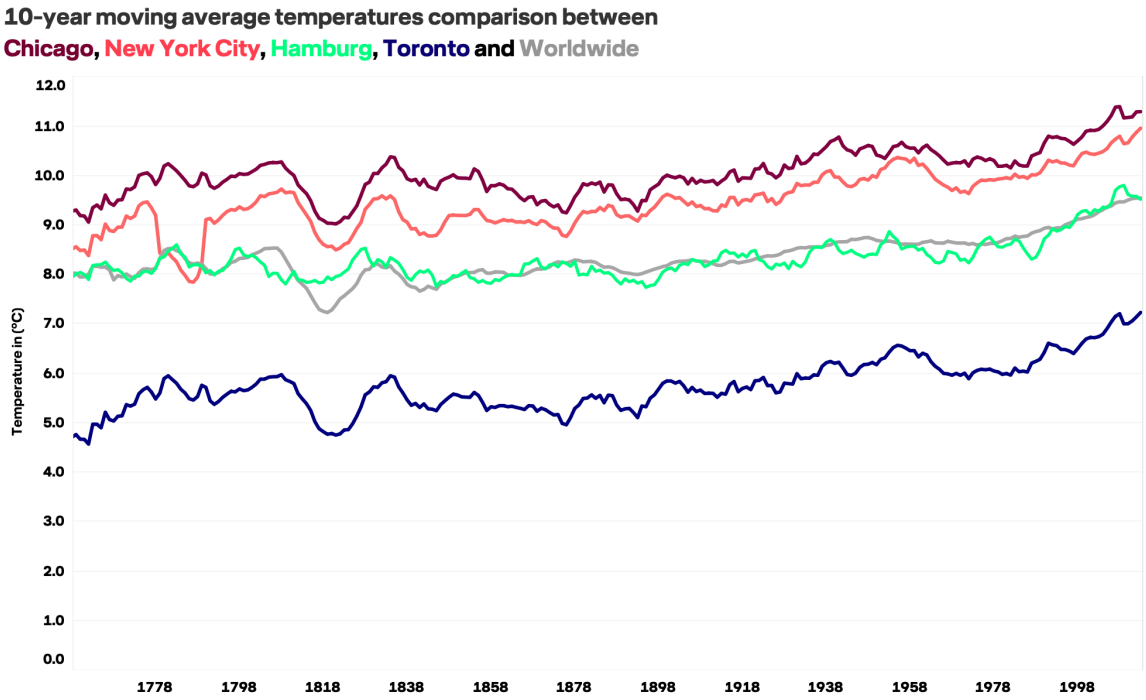

Figure 2: 10-year moving average temperature development comparison across four major cities vs. worldwide

I further calculated the correlation coefficient using Tableau. For that purpose, I chose "Trend Lines" in Tableau and viewed the "Describe Trend Model". The "Describe Trend Model" shows the R-squared value. Following, I created a calculated field using the CORR function to calculate the Pearson correlation coefficient. The Pearson correlation coefficient measures the linear relationship between two variables. Results range from -1 to +1, where 1 indicates an exact positive linear relationship, 0 no linear relationship, and -1 an exact negative relationship. The results are shown in the below table.

City	R-squared	CORR-Formula	Correlation coefficient	Interpretation
New York City	0.317434	CORR([avg_temp_city_NYC], [avg_temp_global])	**.5634**	Strong positive relationship
Chicago	0.447743	CORR([avg_temp_city_CHI], [avg_temp_global])	**.6691**	Strong positive relationship
Hamburg	0.262425	CORR([avg_temp_city_HAM], [avg_temp_global])	**.5123**	Strong positive relationship
Toronto	0.484679	CORR([avg_temp_city_YYZ], [avg_temp_global])	**.6962**	Strong positive relationship

Table 1: Pearson correlation coefficient

OBJECTIVE III: INTERPRETATION OF DATA VISUALIZATION

The core-work of an analyst is not only to turn data into readable and "fancy" visualizations but to provide an interpretation of the data analysis. Hence, the following will provide observations that can be drawn from the analysis.

1. **Hamburg vs. Worldwide**

   Hamburg follows a very similar temperature development trend compared with the worldwide average. The only significant difference lies between the years 1800 and 1818, where the worldwide average temperature declines more than the one in Hamburg. There is an upward trend since 1987 in average temperature for both "locations", which are consistent for the past few decades.

2. **The continuous increase in temperature since 1900**

   Prior to 1900, the average temperature in any of the locations analyzed remained relatively stable with some peaks/lows in individual years, but no continuous increase can be seen. But as of 1900, a steady rise in average temperature can be seen across all cities. The city with the highest relative increase is Toronto with a rise of 122% with a correlation coefficient (Corr) of .6962, followed by Hamburg of 117% (Corr of .5123), New York City by 115% (Corr of .5634) and Chicago by 113% (Corr .6691) between 1900 until 2013. The worldwide temperature average increased by 117%.

3. **The hottest and coldest place are not far away from each other**

   Chicago is on average, the warmest place in our sample, whereas Toronto is the coldest place. Surprisingly, these two cities are not far away from each other geographically but seem to differ significantly in terms of temperature. In 2013, the difference in average temperature between these two cities was a staggering 3.9 degree Celsius.

4. **New York City's unusual temperature drop between 1778 and 1780**

   What instantly catches the eye is the sharp drop in temperature starting in 1778 of New York City. This surprising trend is worth looking into the data in more detail. It turns out that the data for the years 1778, 1779 and 1780 are most likely not very

reliable due to two aspects: (1) lack of data for the year 1980, and (2) unusual low temperature for the years 1778 and 1779 (down as far as to 0.25 degree Celsius). These data inconsistencies cause a lower average, year-wise temperature deviations necessitate further scrutiny.

CONCLUSION

Every city looked at in this data set showed an increase in temperatures. The most significant changes I saw were from 1987 until 2013. Hamburg, located in the cold to moderate environment, closely matched the worldwide temperature averages, whereas the other cities in hotter and colder cities such as Chicago, New York City, and Toronto were higher/ significantly lower. Correlation coefficient numbers all showed a strong positive linear relationship and ranged from .5123 to .6962.

SOURCES

SQL Rename Column Sytnax
Available at: 1keydata.com/sql/alter-table-name-column.html

SQL Sytnax Overview
Available at: w3schools.com/sql/sql_alter.asp

Introduction to SQL:
Available at: coursera.org/lecture/data-analytics-business/1-introduction-to-sql-mhzio

Pearson correlation in Tableau:
Available at: kb.tableau.com/articles/howto/finding-the-pearson-correlation