**Udacity‘s Data Analyst Nanodegree Project 1 – Exploring Weather Trends**

**Project Overview**

In this project I was tasked with exploring the global average temperature change compared to the average temperature change in the closest city I call home, Birmingham UK. The data starts in the mid 18th century and runs until 2013.

**Data Wrangling**

Before combining any tables I opened each table individually to get a quick feel for the data. I did using a simple SQL command:

SELECT \*  
FROM city\_data

SELECT \*

FROM city\_list

SELECT \*

FROM global\_data

At a glance, all the data looked well formatted.

I then searched to see if any duplicate values by using SELECT DISTINCT. In Birmingham UK’s case there were no duplicate rows of data.

SELECT c.city, clist.country, c.avg\_temp city\_temp, c.year, g.avg\_temp global\_temp

FROM city\_data c

JOIN global\_data g

ON g.year = c.year

JOIN city\_list clist

ON c.city = clist.city

WHERE c.city = 'Birmingham' AND clist.country = 'United Kingdom'

ORDER BY c.year;

This returned 528 rows of data

SELECT DISTINCT c.city, clist.country, c.avg\_temp city\_temp, c.year, g.avg\_temp global\_temp

FROM city\_data c

JOIN global\_data g

ON g.year = c.year

JOIN city\_list clist

ON c.city = clist.city

WHERE c.city = 'Birmingham' AND clist.country = 'United Kingdom'

ORDER BY c.year;

This also returned 528 rows of data.

I then searched to see if there were any missing values in the dataset.

SELECT DISTINCT c.city, clist.country, c.avg\_temp city\_temp, c.year, g.avg\_temp global\_temp

FROM city\_data c

JOIN global\_data g

ON g.year = c.year

JOIN city\_list clist

ON c.city = clist.city

WHERE c.city = 'Birmingham' AND clist.country = 'United Kingdom' AND

(c.city IS NULL

OR clist.country IS NULL

OR c.avg\_temp IS NULL

OR c.year IS NULL

OR g.avg\_temp IS NULL)

ORDER BY c.year;

Sure enough I found a missing entry in the cities temperature in 1780. I now had several options to consider.

I could leave it in and hope that the rolling average will lessen the impact of the missing value.

I could replace the missing value with the average surrounding summer temperatures as this value that was missing was the summer record.

I could remove the entire row of data.

I could remove the entire year of data.

I decided to exclude the entire year of data from my results. I choose this option because otherwise my rolling average would be skewed by having an ‘extra’ winter temperature without the corresponding summer value.

SELECT DISTINCT c.city, clist.country, c.avg\_temp city\_temp, c.year, g.avg\_temp global\_temp

FROM city\_data c

JOIN global\_data g

ON g.year = c.year

JOIN city\_list clist

ON c.city = clist.city

WHERE c.city = 'Birmingham' AND clist.country = 'United Kingdom'

AND c.year <> '1780'

ORDER BY c.year;

This returned 526 rows, exactly how many I was expecting and now have clean usable data.

**Data Extraction**

I downloaded the corresponding CSV file and opened it in Microsoft Excel. I did a quick visualisation and quickly decided that the data would be more presentable as a rolling average.

**Data Aggregation**

I calculated a rounded rolling average by using

=ROUND(AVERAGE(D2:D21),2)

And then dragging over the desired range.

I did this for both the city temperature and the global temperature.

**Data Visualisation**

Plotting both my rolling averages against year on a line graph gave the most information on how the overall temperatures, both globally and for Birmingham, were changing.

I also included a line to show the difference in temperature between both the city value and the global value.

**Observations**

The observations I can draw from the graph, and extra calculations, are the following:

The average temperature in Birmingham has been between 4.15 to 5.25 degrees hotter than the average global temperature. The average difference in temperatures is 4.67 degrees Celsius.

Both the average temperature of Birmingham and the average global temperature are increasing. This becomes more prominent in the 20th century.

The average temperatures at the start of the data set in the 18th century seem more volatile and inconsistent, but I imagine that has to do with the quality of the equipment available then.

Both the average global temperature and the average Birmingham temperature have a local minimum in the 1810’s. After researching, 1816 is known as the “Year Without a Summer” where temperatures plummeted across most of world due to a massive eruption of Mount Tambora in modern day Indonesia.

The rolling average for Birmingham’s temperature makes a significant drop around 1780 which leads me further to believe, that the missing data value could have been an equipment malfunction then an introduction of new, more accurate, equipment.

**Conclusion**

After the last mini ice age that ended in the 17th century both the global temperature and the temperature of my home city, Birmingham UK, have been maintained a similar difference.

Both the average temperature of Birmingham and the average global temperatures began to increase and has become more worrying toward the end of the 20th century and into the 21st century.

The Pearson Correlation Coefficient between Birmingham’s average temperature and the global average temperature is 0.875 which shows a strong relationship. I also imagine that the relationship between Birmingham’s temperature and the global temperature is linear.

The Pearson Correlation Coefficient between both local and global temperatures and the year is 0.762 which also shows a strong positive relationship. Meaning as the year increases, so does the temperature but this is probably not a linear relationship.

If I were to pursue this further, I would like to use an extended data set leading up to the present day. I would also like to investigate the relationship between global temperature with global population and Birmingham’s temperature and Birmingham’s population.