Exploration of global temperatures since the 1740s

Details of the dataset

I will be exploring earth surface temperature data available on this Kaggle page. The raw data is provided by Berkeley Earth. Specifically, I will focus on a file which lists average land temperature by country.

The data contains 577,462 observations of four variables, namely the date, the average temperature recorded in Celsius, the average temperature uncertainty (the half-width for the 95% confidence interval around the average) and the country. An example is shown in Table 1 below.

dt	AverageTemperature	AverageTemperatureUncertainty	Country
1855-05-01	25.544	1.171	Brazil

Table 1. A single row from the dataset

Using SQL I was able to confirm that were no duplicate rows. However, there were 32651 rows with missing temperature measurements. Since this was only approximately 5.7% of the data, I simply removed these rows. I also renamed the date variable from "dt" to "DateOfObs".

There was no need to separate this table into multiple tables (i.e. no need for normalization).

General statistics

After importing the data into SQL, I changed some of the data types before obtaining some general statistics. The date and temperature data were stored as strings so I converted DateOfObs to the DATE data type and the two temperatures variables to the DECIMAL data type.

The following observations were made:

1) The total number of records was observed to be 577,462. After data cleaning the new total was 544,811
2) There were 242 "countries" in total. Since 242 is more than the number of countries in the world, it means that some observations are for regions or places which are not actually countries
3) The highest number of observations for any country was 3166 and there were 49 countries that had this number of observations
4) Over all countries and dates, the average temperature was 17.2 °C
5) Over all countries and dates, the average temperature uncertainty was +/- 1.02 °C

6) The hottest places based on the mean average temperature over all dates were Dijibouti, Mali and Burkina Faso with average temperatures of 28.8, 28.4 and 28.0 °C respectively
7) The coldest places based on the mean average temperature over all dates were Greenland, Denmark and Svalbard with average temperatures of -18.6, -18.1 and -7.4 °C respectively
8) Kuwait had the single highest average temperature of 38.8 °C and Greenland had the single lowest average temperature of -37.7 °C. Although Greenland was also the coldest country overall, interestingly Kuwait was not in the top 3 hottest countries.
9) Based on average temperature uncertainty, Romania had the most reliable measurements (+/- 0.05 °C) and Denmark had the least reliable measurements (+/- 15.0 °C)
10) The earlier observation was for the 1st of November 1743 and the last observation was for the 1st of August 2013, so the time span for the observations was approximately 270 years

Analytics using R

Firstly, I checked the summary statistics of the average temperature and average temperature uncertainty, these are shown in Table 2.

Variable	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
Average temperature	-37.66	10.03	20.90	17.90	25.81	38.84
Average temperature uncertainty	0.05	0.32	0.57	1.02	1.20	15.00

Table 2. Summary statistics of average temperature and it's uncertainty

Next, I did box plots to get a visual representation of the data and to check for outliers. It was interesting to note that for average temperature the outliers were in the cold temperature region and for average temperature uncertainty the outliers were in the large uncertainty region.

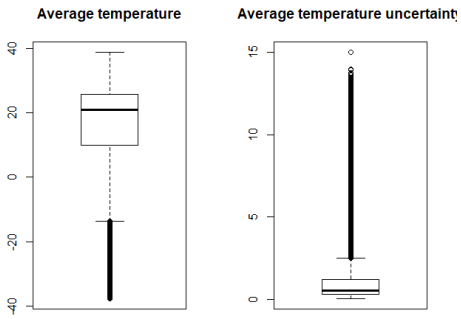

Figure 1. Box plots of numeric variables

Next, I queried the database from R to get the mean of the average temperature over all countries for countries which have observations starting in 1743 and ending in 2013. The reason for selecting these countries was to get the longest possible time span whilst keeping the same set of countries (adding different countries could distort the trend/conclusions). A scatter plot of the mean average temperature over time is shown in Figure 2.

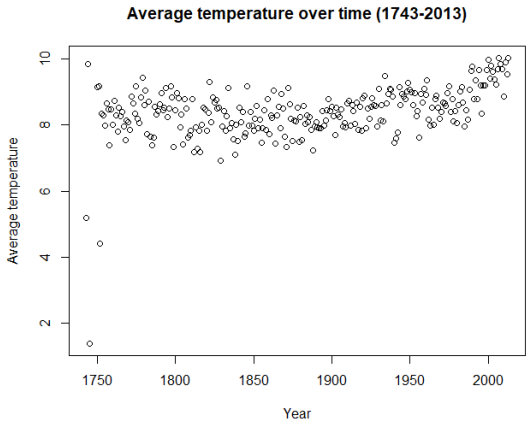

Figure 2. Variation of mean average temperature from 1743 till 2013

However, from Figure 2 we can see that there are some questionable points early on with very low values. Such large jumps in average global temperature do not make sense and are most likely due to measurement error, so instead I adjusted my data to start from the year 1753 instead. The scatter plot for this time period is given in Figure 3. The linear correlation between these two variables was also calculated and was observed to be 0.45.

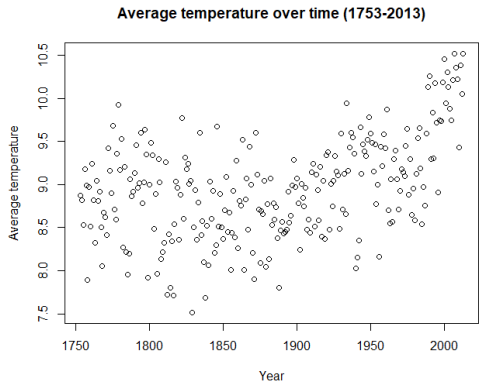

Figure 3. Variation of mean average temperature from 1753 till 2013

Based on both the scatter plot and correlation coefficient, it seemed the average temperature did have positive, linear correlation with time. So I proceeded to do a linear regression. The results are summarized in Figures 4 and 5.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.3456793  0.8173248   2.870  0.00444 **
tempsOverTime[, 1] 0.0035121  0.0004337   8.098  2.2e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5279 on 259 degrees of freedom
Multiple R-squared:  0.202,     Adjusted R-squared:  0.199
F-statistic: 65.58 on 1 and 259 DF,  p-value: 2.201e-14
```

Figure 4. Results of simple linear regression with average temperature as the response variable and time as the predictor variable

Firstly, based on the t/statistic and F-statistic (these statistic are equivalent for simple linear regression with one covariate) we can see that the regression is significant i.e. there is a relationship between average temperature and time. The model can be summarized as follows:

$$Average\ temperature = 2.34 + 0.0035 * Year$$

So over time the average temperatures globally are increasing. However, it is should be noted that the R-squared value is only 0.2 which means that only 20% of the variation in average temperature is explained by time.

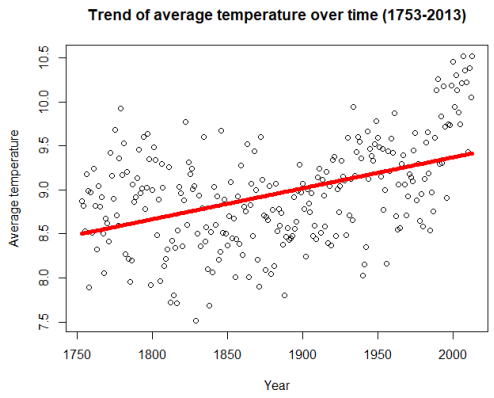

Figure 5. Average temperature trend over time based on regression results

Additionally, I thought it would be interesting to see how average temperature uncertainty has varied over time so I also extracted this information using a SQL query and then repeated the analysis for this variable. The regression was again observed to be significant and the final model was:

$$Average\ temperature\ uncertainty = 34.6 - \ 0.017 * Year$$

In other words, the reliability of measurements is increasing with time which makes sense since it depends on the sophistication of the equipment. This time the R-squared value of 0.82 was much higher (since linear correlation was much stronger, the correlation coefficient was -0.90). It is worth mentioning that based on the scatter plot the actual trend could be better estimated using a negative exponential relationship rather than a linear one.

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                34.5596353  0.9745226   35.46   <2e-16 ***
uncertaintyOverTime[, 1] -0.0174888  0.0005171  -33.82   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6294 on 259 degrees of freedom
Multiple R-squared:  0.8154,    Adjusted R-squared:  0.8147
F-statistic:  1144 on 1 and 259 DF,  p-value: < 2.2e-16
```

Figure 6. Results of simple linear regression with average temperature uncertainty as the response variable and time as the predictor variable

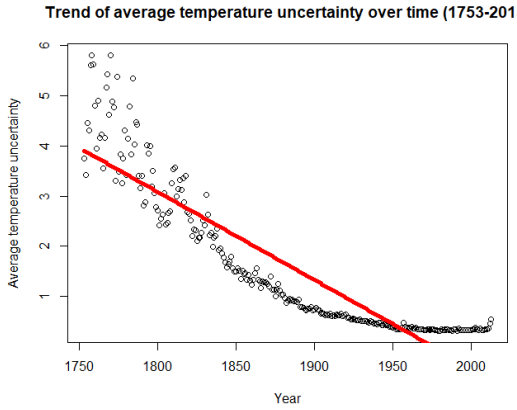

Figure 7. Average temperature uncertainty trend based on regression results

Visualization using Tableau

To visualize the data, I created a map of the countries with a diverging color scheme based on average temperature to show how the cold and hot countries are spread out around the globe. The visualization is interactive so you get a graph of the average temperature over time when you click on a given country. A screenshot is shown below in Figure 8.

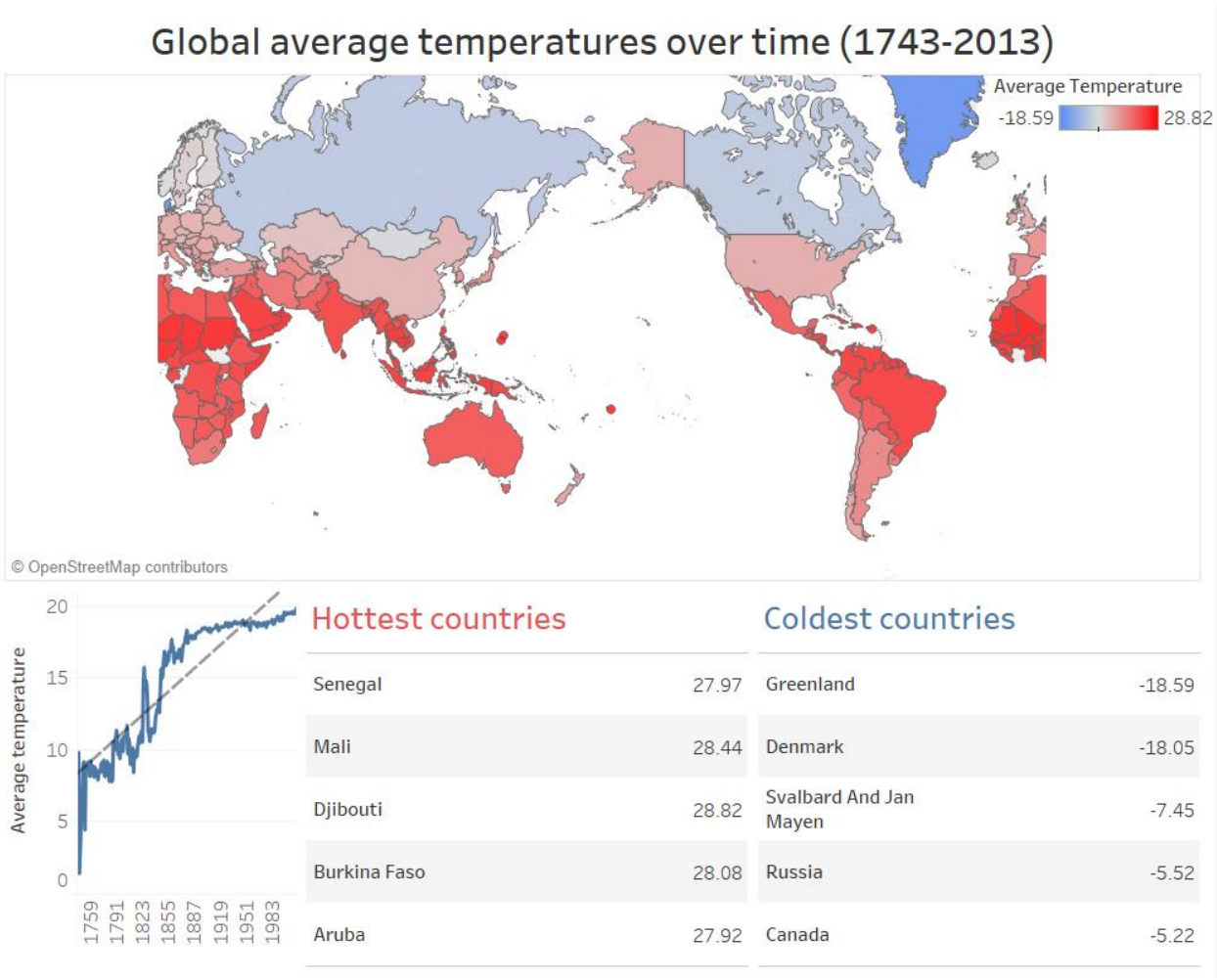

Figure 8. Visualization of earth temperature data

Summary

The regression results show that the average temperature globally is significantly and positively correlated with time. However, the low R squared value suggests the need for other predictors such as CO_2 emissions, world population etc. Moreover, keeping in mind that our earth is billions of years old, even 2.5 centuries is quite a short period of time to analyze.

Another finding is that temperature measurements are becoming more reliable but the rate of improvement seems to be slowing down which is quite intuitive.

Challenges faced

The main challenge during this assignment was the data cleaning. As indicated in the report, there were missing values to be dealt with, data type conversions were necessary and problematic outliers had to be removed.