CIS 3200 – GHCND Weather Dataset Term Paper

Team 2 – Francisco Cortes, Kevin Danao, Kevin De La Torre, Riker Santivong

Department of Information Systems, California State University

Los Angeles

Abstract: The temperature has been recorded by various institutions since the 19th century. The goal of this report is to visualize and analyze data to help identify changes and rising temperatures. It will utilize data collected from 1764 to 1894 in intervals of 10 years and visualize minimum, maximum, and average temperatures in a single location.

1. Introduction

The Global Historical Climatology Network (GHCN) is a database by the National Oceanic and Atmospheric Administration (NOAA), consisting of climate data from sites all over the globe. It contains records of over hundreds of thousands of sites with records in some of them going back to more than 175 years. Furthermore, it updates daily while ensuring that the data is accurate, as the weather continues to change around the world. In this project, my team will analyze the historical record to determine temperature changes over decades to discover shifting weather patterns. This project provides weather data from 1764 to 1894 from a total of seven sites, which allows users to see the weather patterns over the course of 130 years, every 10 years. Using the data, we can model weather patterns based upon the patterns between 1764 to 1894 and compare it to actual weather patterns up to today to note differences and other discrepancies.

2. Related Work

Similar studies of temperature trends were performed by big industries such as NASA and the National Oceanic and Atmospheric Administration, the NOAA. In NASA's study, they observed global weather trends that dates to the 1880's. Keep in mind that during the 1880's there was very few observations and research to the limited weather stations and technology at the time. NASAs studied the geographical trend by taking measurements of the average temperature ratings throughout the globe in a given time frame. Like our studies, the studies that these companies performed also took the average temperature and concluded that the "global temperature on Earth has increased by a little more than 1° Celsius." (NASA) This checks out since our study on Milan, Italy also had an upwards trend when it came to the average temperature. On top of that they also included temperature anomalies which have been recorded between 1880 and 2019, something that we felt that our group should have included during our presentation. It was not until 1980's where our study we took studies every 10 years.

The NOAA also conducted a similar study in global weather conditions, even providing a future trend for the upcoming years. According to the NOAA's study, "By 2020, models project that global surface temperature will be more than 0.5°C (0.9°F) warmer than the 1986-2005 average." In their research they also provided visuals and graphs such as geological maps which shows the trends. Similar to NASA's study, they've been recording records since the 1880's and have noticed slow upwards trend in temperature. The study also ranked the warmest years since its study.

Both studies from the two mentioned sources have conducted studies related to ours, the study of weather trends throughout the years. In the case of NASA, they took their average trends every 30 years where our study we took studies every 10 years. Another similarity from their studies was the average measurement from the temperatures. Unfortunately, due to file size and dataset limitations we were not able to complete our study based on every part of the world, making our study different from NASA's and NOAA's study.

3. Background/Existing Work:

Kaggle offers many different data sets that are related to weather patterns, but some were smaller in comparison to the one our team chose. The data set we used was analyzed and filtered for our use. The data set comprised of 257 individual CSV files, each one representing a different year beginning from 1763. Each of the CSV files individual contained only seven columns while they all varied in the number of rows, they contained anywhere from 800 to over 34,000,000 different entries. The dataset was comprised of detailed recordings of the given weather recorded in degrees Celsius: locations, highs and lows for the day, the precipitation, snow (if any. This allowed us to analyze the data and separate it by months that coincide in certain seasons.

This dataset also included a key attribute that was the location code of the weather station. With that location code we were able to determine the city and country the data was taken from. Due to certain restrictions and limitations with Kibana, along with the

large amounts of data, we were only able to use datasets in ten-year intervals. As well as incomplete collection of data for certain locations we decided it was best to focus on one single location, Milan, Italy.

4. Our Work:

In the beginning of our work, we were deciding on what to run our analysis on. What can our group possibly do with all this weather information and how can we run an analysis? With some time and thinking we decided to run an analysis which took the maximum, minimum, and average temperatures and compare them throughout various years. However, analyzing every part of the world proved to be too much, with and without our limitations that we will discuss momentarily. We then decided to focus on one location – Milan, Italy. From then, we took the years from 1764 all the way from 1894 and split them every 10 years. This decision was ultimately made due to software and service limitations.

As mentioned earlier, our dataset had limitations. The dataset contained weather information such as snow, precipitation, highs, and lows of temperatures, and so on. The problem lied with the consistency between the years and file sizes. One example of inconsistency was with the regions our dataset featured. Each file dataset contained geographical measurements from different regions around the world which were notated as region codes, ITE00100554 representing Milan, making it hard to decode every region these measurements were taking place. However, despite being given these region codes, there were some years where the region codes did not exist. In other words, some regions had missing years while others contained all the years when it came to capturing weather data. Eventually we decided to stick with Milan, Italy since it was the only region that was in common with the years we studied.

Our next problem was the dataset size and upload limitations of Kibana. When downloading the dataset, the compressed file size was 14GB. Uncompressing the data set turned out to be 93GB worth of weather information, containing over 253 CSV files! These files ranged from 24kb to 1.3GB, each year being larger and larger. Due to upload limitations of Kibana, only allowing uploads of file sizes 100MB or less, we were at a stand-still on what years we should conduct our study on. We were not able to perform any type of data analysis with the past 10 years of weather history due to file size limitations. We were not able to analyze each year due to the inconsistent data. Eventually, we decided to conduct a data analysis from 1764 to 1894, with intervals of every 10 years, since these years are the largest in file size without breaking upload limitations.

After running into these problems and coming up with solutions to solve them, our next step was to come up with visualizations to represent the trends throughout the decades. We used both Kibana and Azure ML to come up with different bar graphs, line charts, and tree maps depicted in our analysis.

4.1 Analysis:

Figure 1 Maximum Temperatures between 1764 and 1794

This figure shows the maximum temperature taken from Milan, Italy in 1764, 1774, 1784, and 1794. Keep in mind, these are temperatures taken from spring, which are between March and May. The bar chart shows the temperatures in Celsius and Fahrenheit.

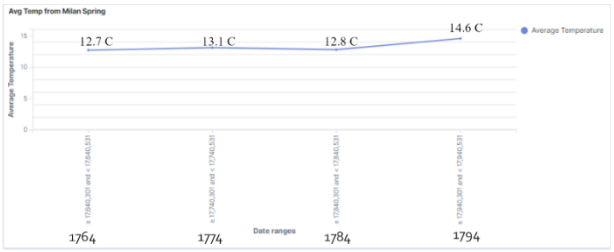
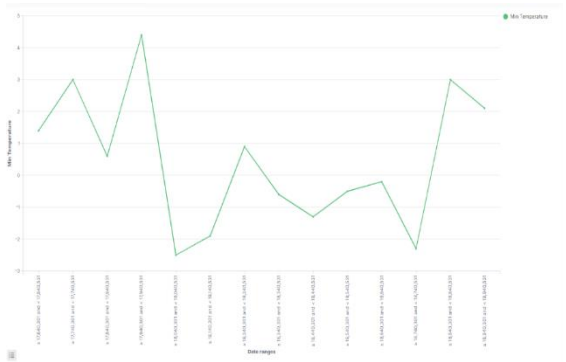


Figure 2 Average Temperatures between 1764 and 1794

This figure shows the same data but in a line graph instead of a vertical bar graph. This is measuring the average temperature in Celsius from the same year range. Notice how by 1794 the average temperature in Milan raised at least 2 degrees!

Once the datasets were visualized from our selected years, we moved on to using Azure ML and continued our analysis. Because we could not tamper with the dataset while creating these visuals, we had to filter out each visual, vertical bar and line graph, to only show data from Milan, Italy. This was done by applying the country code of Milan, ITE00100554, into the filter feature in Kibana. We also had to filter out the dates to range between March and May, the Spring months. In order for our team to achieve that we had to make a range filter as well and apply it. Our dataset did not provide actual dates, but instead numbers. For example, March 1st of 1764 was represented by 03011764. Using these numbers, we ranged it to their respective dates to

project our maximum, minimum, and average spring temperatures from Milan.



This line chart (Figure 4) shows us the lowest temperatures for each year that we uploaded, for a total of fourteen different years. We can see that in between those ten years there was a drop in temperature and that drop in temperature stayed prevalent for almost seventy years causing them to be lower for those years in between. After that, it shoots right back up to similar temperatures to what it was almost seventy to eighty years prior.

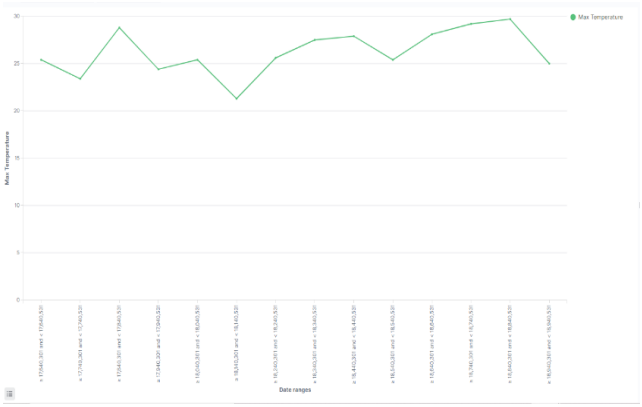


Figure 4 Line Chart Highest Temps

In the graph above (Figure 5) it shows us what the highest recorded temperature was in Milan for those fourteen chosen years. We can see that the high temperatures are relatively close to one another in the 140-year span, not much in terms of discrepancies in it rising than the norm or going lower. Only one singular year where the temperature was lower than what seems to be the average, with the temperature rising by a few degrees.

Using the data, our team sought to create a model that would accurately fit the data using Microsoft Machine Learning. Furthermore, my team decided to use two different models, score the model, and then compare the results of the two models using the evaluate model function. The two models we decided to use were Bayesian Linear Regression and the Decision Forest Regression. However, my team quickly encountered a major issue when trying to run the model. Due to the large datasets, we used, especially the 1884 and 1894 datasets, the process seemingly went on. As a result, my team decided to use the data without the 1884 and 1894 datasets first.

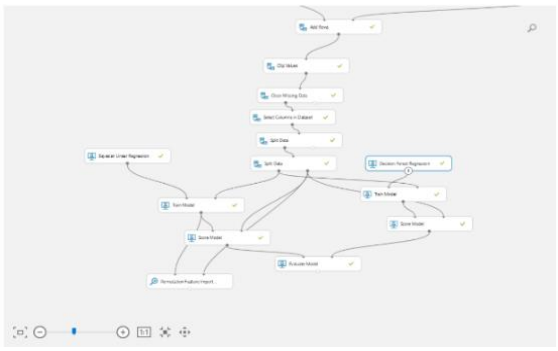


Figure 5 First Tree Model created in Azure ML

Without the large datasets, Azure ML program. loaded quickly, and my team were able to create and evaluate the models within less than 5 minutes.

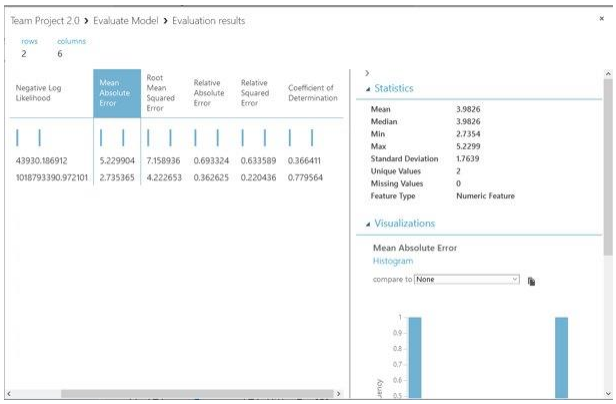


Figure 6 First Evaluating Model. Bayesian Linear regression is on the top and the Decision Forest Regression is at the bottom

The first result of the evaluate model function showed that the Decision Forest Regression model was more accurate in predicting the data, which we expected. Linear regression is generally less accurate in predicting data.

However, my team still wanted to predict the models with all the data available. In the end my team and I decided to include the 1884 and 1894 datasets in another iteration of the model and allow the process as much time as it needed to complete. It took more than 24 hours to complete the modeling, and its results were surprising to us.

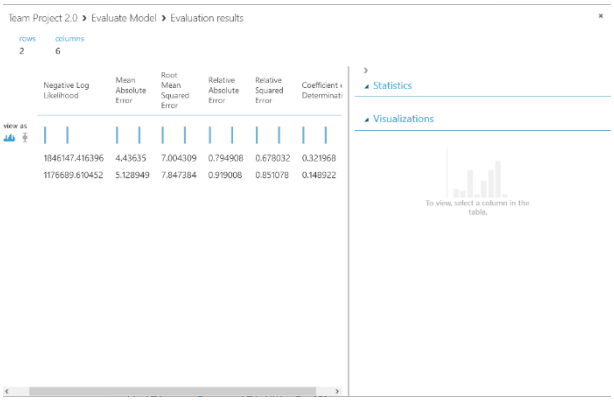


Figure 6 Second Evaluating Model. Note that the results of this one is significantly different from the model without 1884 and 1894.

However, when we include the 1884 and 1894 datasets it shows that the Decision Tree Regression has significantly dropped in accuracy. My team is not completely sure why there is a significant discrepancy between the two models. However, my team's theory to why there is such a discrepancy is that 1884 and 1894 datasets add many new sites that collects the temperatures and weather of their region. The data collected from these new sites are much different because of their locations in different climates. As a result, the data from the new sites confuses the Decision Tree Regression model, as it tries to predict the previous data with the data from these new sites and largely fails.

Due to the time it takes to run these models my team was unable to have the time to rectify this issue. However, if we were to run this model again, my team would exclude the new sites in the 1884 and 1894 datasets by using the split module. My team believes by excluding the data from the new sites and only using data from the old sites the new model should be corrected. Thus, the output from the evaluate model module should lead to similar results that we found in the original model.

5. Conclusion:

In conclusion, we used a multitude of data provided by the National Oceanic and Atmospheric Administration (NOAA). The data included weather information from years 1763-2019, but we only used 14 years in intervals of 10 years due to size constraints. The recorded information contained countless information such as Meteorological Station ID, Date, Max Temperatures / Min Temperatures, and Temperature values. From this dataset, we were able to analyze it using the following programs: Elasticsearch, Kibana, and Azure Machine Learning. We used Kibana Machine Learning to upload the 14-years files. From there, we used to "Kibana Visualize" to started visualizing our dataset by using Vertical bar charts, horizontal bar charts, Map charts, and line charts. Overall, we found the data fascinating; Max temperatures having minimal variation while the Min Temperatures have a lot of variation. After that, we adopt two Machine Learning algorithms to build models in Azure Machine Learning to predict values from the temperature column. The two models we used are Bayesian Linear Regression and Decision Forest Regression. Lastly, we ran and compared the models' results, which lead to Decision Forest Regression having the best accuracy.

6. References:

Global historical Climate Network daily - Description. (n.d.). Retrieved May 14, 2021, from https://www.ncdc.noaa.gov/ghcn-daily-description

NASA. (n.d.). *World of Change: Global Temperatures*. NASA. https://earthobservatory.nasa.gov/world-of-change/global-temperatures.

Climate Change: Global Temperature: NOAA Climate.gov. Climate Change: Global Temperature | NOAA Climate.gov. (2021, March 15). https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature#:~:text=According%20to%20NOAA's%202020%20Annual,more%20than%20twice%20that%20rate.