Discovering Temperature Trends in New Brunswick: Implications for Planning and Policy

Authors: Elijah Kojo Danso Appau, Mike Mico, Andrew O'Connor

CS4403 Data Mining: Final Report

Dr. Jake van der Laan

April 14, 2023

Introduction/Background

The goal of this project is to explore the effect of geography on the temperature of New Brunswick, and how it has changed over time. This was done by analyzing a publicly available weather dataset (Government of Canada, 2023) showing the daily temperature readings from 33 weather stations throughout New Brunswick. We downloaded the data from January 2016 - March 2023.

We then grouped this data based on how close the temperatures were to each other. This was done using a technique called K-Means Clustering.

K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters. K-means clustering tries to group similar kinds of items in form of clusters. It finds the similarity between the items and groups them into the clusters.

Motivation

To find relevant climate insights for New Brunswick, we looked at some related studies that had been conducted at the federal level or for other countries. This gave us the inspiration for this topic. We were looking for information that would be helpful in making decisions about everything from public policies to where to go on vacation.

Data Exploration

The dataset from 2016 to 2023 consisted of 82514 rows and 36 columns of data where each row represents a daily report from a single weather station and each column represents a recorded feature. To better visualize the data over time, we grouped the data by year and month and removed any null values, which reduced the number of rows to 2320. We also eliminated the features that we deemed unnecessary for our study. This left us with 11 features to work with:

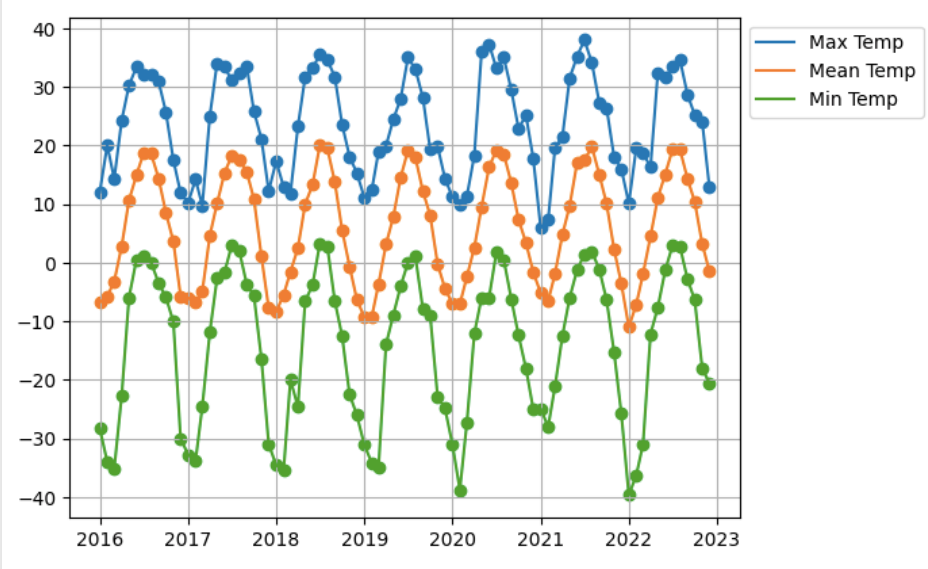
- x: longitudinal location of weather station
- y: latitudinal location of weather station
- STATION_NAME: station name
- CLIMATE_IDENTIFIER: Unique weather station identifier.
- MEAN_TEMPERATURE: Average temperature recorded on that day
- MIN_TEMPERATURE: minimum temperature recorded on that day
- MAX_TEMPERATURE: maximum temperature recorded on that day
- LOCAL_DATE: Time Stamp of the date
- LOCAL_YEAR: Year recorded
- LOCAL_MONTH: month recorded
- LOCAL_DAY: day recorded

Some of our station ID data was recorded as strings which was affecting our results. We fixed this by converting those values from strings to floats.

Many of the features outside of temperature were missing data. For example, out of the 82514 rows, both TOTAL_RAIN and TOTAL_SNOW were null in approximately 60500 of those rows. This being over 73% of the data, removing these rows would be unhelpful.

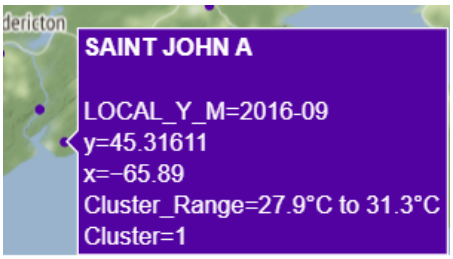
However, there was a TOTAL_PRECIPITATION column that was a sum of the TOTAL_RAIN and TOTAL_SNOW columns, and it was only null in 12804 rows (15%). There were several cases where the TOTAL_PRECIPITATION was zero while the TOTAL_RAIN and TOTAL_SNOW columns were null. In these cases, the rain and snow columns could be safely set to zero, which reduced the number of nulls to only about 33000. This was still 44% of the data. Because of this missing data, we largely omitted these other features.

After cleaning our data, we were able to create line charts that showed useful trends in the temperature recordings. One of them (shown below) displays the changes in temperature over the course of multiple years on a month-by-month level.



We were also able to display our data on an interactive map of New Brunswick to better visualize our results. By using the slider on the bottom, you can observe the change in temperature clusters over the course of each month, with each cluster representing a certain temperature range.

The temperature range represented by each cluster is not immediately apparent when looking at the chart. To remedy this, we found the minimum and maximum temperature for each cluster. We then created a new column containing the temperature range as a string of text and had this label display when hovering over a data point.



Key Idea

To demonstrate the correlation between geography and temperature, naturally we know that areas in the north are more likely to be cold compared to other directions, but we wanted to test this hypothesis and see these changes throughout the year as temperature varies based on so many factors. We also want to be able to predict temperature with any given year month and longitude and latitudes.

Efforts and Results

Initial Plan

Our first plan was to try to find a connection between regional temperature, energy usage, and COVID-19 rates in New Brunswick. We believed there may be a causal link between people needing to stay at home and energy usage, and wanted to explore if there was also a correlation between energy usage and temperature. However, the project team was unable to find any household energy usage data that was separated by region, nor could we find any energy usage data from later than 2019 (i.e., before any COVID-19 data became available).

Predictive Models

The weather data had both many data points over time and a high number of features, so we decided to focus our efforts on that one dataset.

Our first attempt at exploring the weather data involved creating a predictive model to guess the maximum temperature recorded at a given point and time. We later decided that finding the temperature regions using K-means clustering would be a more interesting challenge to explore, but near the project deadline we attempted to use the regressor model again to collect information for this report.

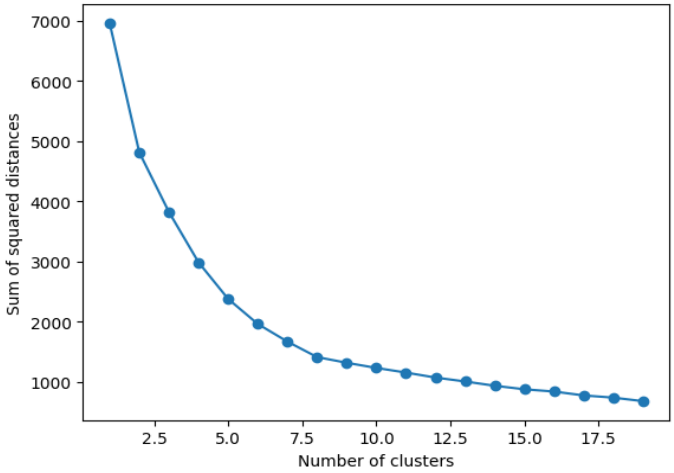
Using only location and date as the input features and max temperature as the target feature, we found it to be very successful on test data using an 80-20 train-test split on a random forest model. Our accuracy score was 98.4% at guessing the temperature based solely on these features. Given additional time, this may have been an interesting path to explore.

Temperature Clustering

The K Means Clustering grouped the location based on temperature where every group had a maximum and minimum temperature which changed over the course of months.

We used the K-Means Python library from Scikit-Learn (scikit-learn, 2023) to perform the clustering. To prepare the data for clustering, we first grouped it by station name, location, year, and month. The MAX_TEMPERATURE feature was aggregated to get the maximum temperature for each month for each station. The pandas DataFrame containing the data to be clustered was called X.

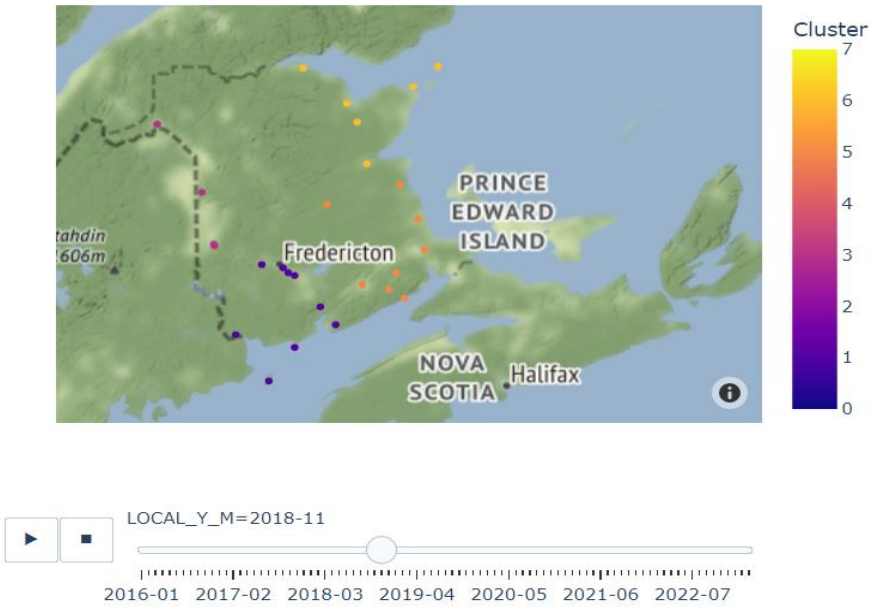
We used elbow analysis to gauge an appropriate number of clusters. This involved fitting the data multiple times using a different number of clusters, and seeing which point in the graph had the most notable “elbow-shaped” bend. This plot is shown below:



Based on this plot, we decided that eight clusters would make for a good amount.

To get more useful clustering information, we used the scale function from Sci-Kit Learn's preprocessing library on X before fitting the K-Means model to it. This was done to normalize the data, which is expected to improve accuracy.

We took two broad approaches to clustering the data. In one approach, we set X to contain the longitude, latitude, and temperature features. This gave clearly defined boundaries around each region:



However, each cluster did not have a clearly defined temperature boundary. The maximum and minimum temperatures for each cluster are shown below. For example, cluster 4 ranges from 18.4 to 37.2 degrees, and cluster 5 ranges from 20 to 38 degrees.

Cluster	Min	Max
0	20.3	37.3
1	-0.3	20.0
2	19.4	36.0
3	0.2	18.1
4	18.4	37.2
5	20.0	38.0
6	-6.1	19.2
7	1.5	20.2

Including the latitude and longitude in the features resulted in closer stations getting clustered together regardless of whether they share temperature characteristics. This clustering may not be as useful for determining what temperature regions exist in New Brunswick, so we repeated the clustering while only using temperature as the feature to be clustered. This resulted in the following chart:

This gave the following temperature ranges:

Cluster	Min	Max
0	24.1	27.7
1	-6.1	6.8
2	11.5	15.6
3	31.3	38.0
4	6.9	11.4
5	27.8	31.2
6	15.7	19.9
7	20.0	24.0

It can be seen that there are no overlaps in temperature. In general, we found that the northern regions were often cooler than the south.

Conclusions

We were able to explore the effect of geography on temperature, and with the help of k-means clustering we identified regions that had similar temperature range at a given time of the year. We also found that our predictive model using a random forest regressor was able to predict minimum, maximum, and mean temperature on a given date and location with an accuracy of close to 98%. We believe with this accomplishment it will make it possible for both the government and individuals to make day-to-day decisions based on the weather, implement policies that are temperature-based and help one select the best vacation spot during the summer.

Bibliography

Government of Canada. 2023. Adjusted and Homogenized Canadian Climate Data. *Climate Change Canada.* [Online] March 2023. [Cited: March 23, 2023.] https://climate-change.canada.ca/climate-data/#/adjusted-station-data.

scikit-learn. 2023. sklearn.cluster.KMeans. *scikit-learn.* [Online] scikit-learn, February 20, 2023. [Cited: April 14, 2023.] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

Statistics Canada. 2022. Census Profile, 2021 Census of Population. *Statistics Canada.* [Online] Statistics Canada, February 9, 2022. [Cited: March 7, 2023.] https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/details/page.cfm?Lang=E&SearchText=New%20Brunswick&DGUIDlist=2021A000213&GENDERlist=1,2,3&STATISTIClist=1,4&HEADERlist=0.

—. 2022. Household energy consumption, by household income, Canada and provinces. *Statistics Canada.* [Online] May 2, 2022. [Cited: March 7, 2023.] https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2510006201.

—. 2023. Preliminary dataset on confirmed cases of COVID-19, Public Health Agency of Canada. *Statistics Canada.* [Online] March 17, 2023. https://www150.statcan.gc.ca/n1/en/catalogue/13260003.