

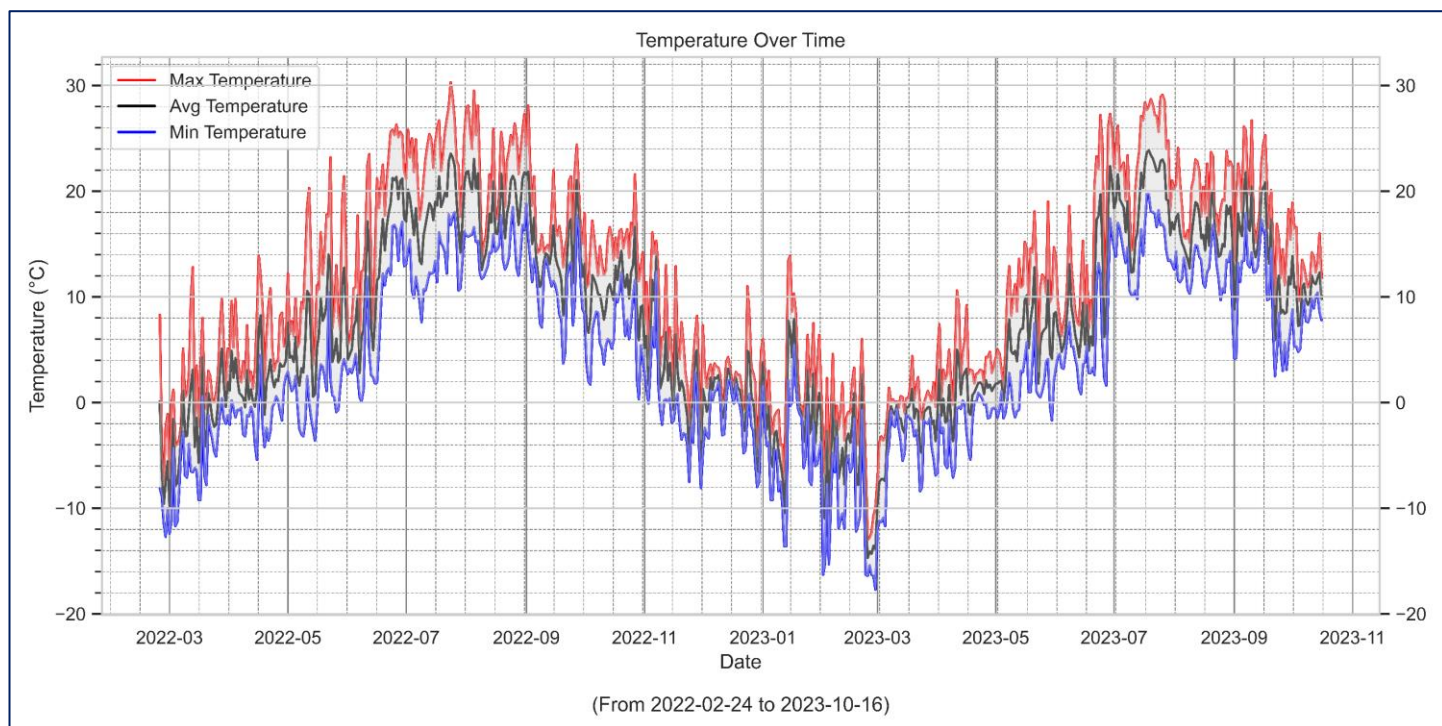
Overview of Dataset

I plan to analyze climate data for St. John's, which can be sourced from <https://stjohns.weatherstats.ca/download.html>. I've obtained 600 daily data points covering the period from February 24, 2022, to October 16, 2023. While I currently have 600 data points, I can retrieve additional data if necessary. For this project, we will primarily focus on temperature and wind speed features. If required, we may explore other features from the original dataset, which will be detailed as needed.

Attributes

- **max_temperature** and **min_temperature**: Maximum and Minimum Temperature (°C). **avg_temperature** is the average between the daily maximum and minimum temperatures, while **avg_hourly_temperature** is the average of all the hourly temperatures within the day.
- **max_wind_speed** and **min_wind_speed**: Maximum and Minimum Wind Speed (km/h)
- **precipitation**: Precipitation (mm): The amount of rain/snow/etc received. Snow is melted to create a water equivalent. 1cm snow is approx. 1mm of precipitation, but the exact amount depends on the snow density (e.g., corn snow vs light fluffy snow).

First Plot



In this time-series plot, we represent daily temperature trends. The blue curve shows minimum temperatures, the red curve indicates maximum temperatures, and the black line within the shaded area represents daily average temperatures. Notably, late July 2022 stands out with a broad range of high temperatures. We can see other trends from this plot, too.

Next Steps

Possible steps to consider include: **Descriptive statistics**: basic statistics for variables, such as mean, median, standard deviation (measure of data spread), and quartiles. **Extreme value analysis**: Identify and analyze extreme values in my data. For instance, the days with the highest and lowest temperatures, or those with the highest wind speeds. **Trends and Seasonality**: Examine temporal trends in my time-series data. I could calculate moving averages, identify seasonality, or look for patterns in temperature changes or wind speed variations over time. **Correlation analysis**: Explore relationships between different variables. For example, I could investigate whether wind speed is correlated with temperature. Use correlation matrix and heatmap. **Visualization**: I plan to use visualization techniques to complement my statistical analysis and data distribution assessments. For example, create time series plots, histograms, scatter plots and regression plots (analysis), box or violin plots to help convey the patterns and outliers in my data. For outliers, we can use Interquartile Range method and Standard Deviation method. **Compare** extreme events to historical averages to determine the extent to which they deviate from what is considered typical for a particular time. I can access 30-year historical data for my dataset through the data source. I may want to calculate how many days had temperatures above or below a certain threshold compared to historical averages. **Test** whether there's a statistically significant difference in temperature between different months, or test other hypotheses. **Try** different thresholds or criteria for what I consider extreme temperature events. For example, I might initially define an extreme temperature day as one with temperatures more than 5 degrees Celsius above the historical average. Then, I could explore how my results change when I use a threshold of 10 degrees Celsius or when I consider both high and low temperature extremes. **Future forecasting**: Consider forecasting future trends and compare them with new data available from the data source. I may want to use some common models including ARIMA (AutoRegressive Integrated Moving Average), Exponential Smoothing, or machine learning methods like LSTM (Long Short-Term Memory) for forecasting.