**Project Report**

CMSC 6950 (Fall 2023)

**Mohammadali Mirmojarabian** (**Student #:** 202292543)

----------------------------------------------------------------------------------------------------------------------------------

# Description of Dataset

We analyzed climate data for St. John's, which can be sourced from <https://stjohns.weatherstats.ca/download.html>. We obtained 730 (two-year) daily data points covering the period from November 01, 2021, to October 31, 2023. For this project, we primarily focused on temperature, wind speed, and precipitation features.

## Attributes

All features of the data used in the project are shown in **Table 1**:

**Table 1**: Dataset Overview of Used Columns. Units are in red. Short names are used in this project for legibility.

|  |  |  |
| --- | --- | --- |
| Feature Name | Description | Short Name |
| date | date as index (yyyy-mm-dd) | date |
| max\_temperature | daily minimum temperature (°C) | Max Temp |
| min\_temperature | daily minimum temperature | Min Temp |
| avg\_temperature | average between the daily maximum and minimum temperatures | Avg Temp |
| avg\_hourly\_temperature | average of all the hourly temperatures within the day | Avg hr Temp |
| max\_wind\_speed | daily maximum wind speed (km/h) | Max Wind |
| min\_wind\_speed | daily minimum wind speed | Min Wind |
| avg\_hourly\_wind\_speed | average between the daily maximum and minimum wind speeds | Avg hr Wind |
| precipitation | amount of rain/snow/etc. received. 1cm snow ~ 1mm precipitation. The exact amount depends on snow density (mm) | Precip |
| avg\_hourly\_relative\_humidity | average of all the hourly relative humidities within the day (%) | Avg hr Humid |
| avg\_hourly\_pressure\_sea | average of all the hourly pressures within the day (kPa) | Avg hr Press |
| avg\_hourly\_visibility | average of all the hourly visibilities within the day (m) | Avg hr Visib |

# Methodology

In this project, we did the analysis operations through the code notebook code\_project.ipynb. We have defined all the functions in functions.py and have tested the computational ones (the first two) sufficiently by pytest in test\_functions.py. The used dataset is in the data folder. The daily.csv file contains daily climate data for the mentioned period, and normal\_daily.csv file contains the 30-year historical values. Our produced results (plots) are in the figures folder. The requirements.txt file also mentions the appropriate library versions used in our project.

As our methodology, we discuss how we **presented** (visualized) our data to understand its characteristics. We then introduce different approaches we used to detect **extreme values**. And then, we go through the different ways we used to explore **trends** existing in our data. After the Methodology section, we will show and discuss the results. To save space and be concise, we plotted most of our figures for a selected number of columns as it can be done for the others by calling the same functions on the favorite columns: selected columns: avg\_hourly\_temperature, avg\_hourly\_wind\_speed, precipitation.

## Initial Data Presentation (Visualization)

In this section, our methodology's aim was to get to know our data by plotting them in a series of time-series plots as well as histogram probability plots. We chose to present our temperature group data in one time-series plot. We can have the plots for the other features as well, if needed.

We calculated descriptive statistics of our data as well as their probability distributions. The histograms were plotted using plot\_histograms\_density function. In that function, we used hist and density to plot the distribution of our columns probabilities.

## Extreme Values

In this part, our approach was to first have the boxplots of our columns. In this way, we can clearly understand the distribution and outliers of our data. We showed for our selected columns in the Results section.

Then, we used Interquartile Range (IQR) method to replace the outliers with the IQR limits. Then we will see both unmodified and modified versions of our data. To remove outliers, we developed remove\_outliers\_IQR function. In this function, the computational iqr\_limits function is used for calculating the IQR limits (its test function is available in test\_functions.py as test\_iqr\_limits):

Equation :

Equation :

Equation :

Q1 and Q3 are lower and upper quartiles; we can use the IQR method of identifying outliers to set up a "fence" outside of Q­1 and Q3. Any values that fall outside of this fence are considered outliers. To build this fence, we take 1.5 (as a rule of thumb, but can vary) times the IQR and then subtract this value from Q1 and add this value to Q3.

As our (temperature) data follows a cyclical pattern and experiences regular bumps based on seasonal variations, the assumption of a normal distribution may not hold. So, we focused on using the comparison with historical values method. We have 30-year historical (normal) values for our maximum and minimum data as well as precipitation. We illustrated specifically for max\_temperature and min\_temperature. In this approach, we developed a computational function called historical\_extremes (its test function available in test\_functions.py as test\_historical\_extremes) to pinpoint those data points placed outside of a threshold up and down from their corresponding historical values. We analyzed its sensitivity for values of 5 and 10 and have compared the plots. For statistics, we surveyed the descriptive statics of the extreme values candidates: count, mean, std, min, 25%, 50%, 75%, max. We then could answer many questions, such as:

What is the frequency, range, and variability of these extreme values? Are extreme values associated with particular times or conditions? And some more questions.

## Trends

In this part, we did three works: Correlation Heatmap, Moving Average over features, Average Monthly Precipitation Barplot across years.

For the Correlation Heatmap, we surveyed relationship between every two features from all columns. We created a mask to remove the extra (repeated) part from plot. Red color (corresponding to +1) in plot shows a perfect positive relationship, while blue (corresponding to -1) shows a perfect negative relationship. We used .corr() on our data frame to get correlation matrix; then used sns.heatmap to visualize the map.

In the Moving Average method, we have a function called plot\_moving\_average, through which you can visualize a moving average time series over the original column data with a favorite window size (10, 30 days, etc.). We used df.rolling() and then .mean() to calculate the moving averages.

And the last way, to figure out more pattern behind the scene, we sum up the total precipitation in each month. Then we averaged for all the same months as a group. Then, we showed results in a bar plot.

In the next section, we will show the results and discuss and analyze the findings.

# Results

## Initial Data Presentation

In **Figure 1**, We have plotted max\_temperature, min\_temperature, avg\_temperature as red, blue, and black lines, respectively, over date in one time-series plot. As you can see, the average temperature has been placed between the maximum and minimum temperatures. You can see the time series all over the two-year (2021-11-01 to 2023-10-31) period. We can have the time series plots for the other features as well but we showed for temperatures as samples.

In **Figure 2**, we have probability values distributions for the selected columns of avg\_hourly\_temperature, avg\_hourly\_wind\_speed, precipitation. In this plot, we have both density and normalized histogram for each feature. As you can see, we have a normal distribution for avg\_hourly\_wind\_speed, but this is not the case for other shown distributions. Histograms are a good choice for almost all data types to assess the distributions. When combined with probability density plots, histograms can help us recognize the type of our distribution. Density plot is a smoothed version of the histogram and is used in the same concept. Further, we have calculated descriptive statistics of our data to see the centrality and spread measurements in the form of numbers.

**Table 2**: Descriptive Statistics of Dataset Columns. Short names and units used are from Table 1. (25%, 50%, and 75% are lower, mid, and upper quartiles)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Measure  Feature** | **Max Temp** | **Avg hr Temp** | **Avg Temp** | **Min Temp** | **Avg hr Humid** | **Max Wind** | **Avg hr Wind** | **Min Wind** | **Avg hr Press** | **Avg hr Visib** | **Precip** |
| count | 730 | 730 | 730 | 730 | 730 | 730 | 730 | 730 | 730 | 730 | 730 |
| mean | 10.14 | 6.21 | 6.35 | 2.57 | 84.84 | 38.18 | 24.63 | 11.55 | 101.34 | 18899.05 | 4.35 |
| std | 9.27 | 8.35 | 8.57 | 8.21 | 10.60 | 13.05 | 9.39 | 8.10 | 0.97 | 6603.22 | 8.82 |
| min | -13 | -15.07 | -14.7 | -17.7 | 46.6 | 12 | 7.29 | 1 | 98.04 | 250 | 0 |
| 25% | 2.7 | -0.19 | -0.17 | -3 | 78.2 | 28 | 17.55 | 5 | 100.79 | 16321.85 | 0 |
| 50% | 9.45 | 5.06 | 5.15 | 1.35 | 86.1 | 36 | 23.65 | 9 | 101.48 | 21910.4 | 0.4 |
| 75% | 17.15 | 12.98 | 13.3 | 10.05 | 93.68 | 46 | 30.03 | 16 | 102 | 24100 | 4.4 |
| max | 30.3 | 24.33 | 23.85 | 19.7 | 100 | 83 | 60.61 | 49 | 104.09 | 24100 | 80.6 |

A graph with red and blue lines

Description automatically generated

**Figure 1**: Time-series illustration for max\_temperature, min\_temperature, and avg\_temperature, represented with red, blue, and black lines, respectively. Each square is a monthly span. The whole period is two years, from 2021-11-01 to 2023-10-31. Maximum and minimum temperature range is observable.

A screenshot of a graph

Description automatically generated

**Figure 2**: Probability density and normalized histogram distributions illustration. The plot is shown for avg\_hourly\_temperature, avg\_hourly\_wind\_speed, and precipitation features. Numerical descriptive statistics are available in **Table 2**.

## Extreme Values

First, we have shown the boxplots for selected columns in Figure 3. Features may or may not have outliers based on the 1.5 IQR rule. It shows the distribution of the features as well. For precipitation, it shows a right-skewed distribution with many outliers. Rarely does one feature show a normal distribution. You can compare histograms and boxplots together.

A diagram of a graph

Description automatically generated with medium confidence

**Figure 3**: Boxplots for selected columns: avg\_hourly\_temperature, avg\_hourly\_wind\_speed, and precipitation. The points outside the 1.5 × interquartile range are outliers. Outliers based on IQR method are enormous for precipitation. For avg\_hourly\_temperature, we do not see any outliers.

A graph of a graph of a graph

Description automatically generated with medium confidence

**Figure 4**: Illustration of Modified and Original Editions of Selected Columns. The used scale in (scale × IQR) is 1.5. Orange shows parts of the original time series replaced by upper or lower IQR method limits. We see a massive correction for precipitation. Outlier ratios are 0, 1.23, and 12.19% of the data points in the above features, respectively.

In **Figure 4**, parts of the original time series of the selected columns are replaced by IQR limits, showing the orange version. The IQR method for detecting outliers may not be suitable for all columns as the distributions show different behaviors, and also, some features, like temperature, show a strong cyclic pattern over time; So it's not appropriate to use a constant scale in IQR method to detect outliers in such plots. From **Figure 3** and **Figure 4**, we notice that we should use a more reliable method for extreme values detection.

A graph of a graph

Description automatically generated with medium confidence

**Figure 5**: Illustration of Extreme Values Detection Based On Historical Values. Extreme values (yellow and cyan color points) are detected respectively for max\_temperature and min\_temperature values (red and blue time series). Historical time series for both high and low temperatures are the two comparatively parallel black lines. The sensitivity values tested for threshold are 5 and 10 days above/below a historical line for each of the temperature lines separately. We see extreme values mostly around pits and peaks in historical trend lines.

In **Figure 5**, you can see we used historical values for temperatures and have defined a distance threshold so that values outside that range up and down from historical values are considered extreme values. The advantage of this method is that the historical values are changeable over time, and we detect extreme values suitable for each time slot. By tuning the threshold and sensitizing the result to extreme values, we can achieve a reliable way of detecting extreme values.

Figure 5 shows that a threshold of 5 days seems suitable for detecting extreme high and low temperature values. We can see there is extreme values around pits and peaks of historical lines. Around the second pit, we see outliers for both 5 and 10-day thresholds. Where we have outliers, it's for both low and high temperatures. We might say extreme values are associated with times when overall temperatures are too high or low (cold and hot seasons). We can see the descriptive statistics for the detected extreme values in **Table 3**.

**Table 3**: Extreme Values Descriptive Statistics. S is sensitivity, which is done for threshold days of more than 5 or 10 days. Short names are used for feature names.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measure  Feature** | **Max Temp**  **(S > 5 days)** | **Max Temp**  **(S > 10 days)** | **Min Temp**  **(S > 5 days)** | **Min Temp**  **(S > 10 days)** |
| count | 205 | 15 | 138 | 5 |
| mean | 13.18 | 7.21 | 4.19 | 4.12 |
| std | 10.83 | 12.67 | 10.98 | 13.54 |
| min | -13 | -13 | -17.7 | -17.7 |
| 25% | 6.4 | -0.35 | -1.58 | 2.7 |
| 50% | 13.4 | 11.9 | 3.9 | 5.5 |
| 75% | 23.5 | 13.80 | 14.68 | 12.4 |
| max | 30.3 | 23.2 | 19.7 | 17.7 |

## Trends

A screenshot of a graph

Description automatically generated

**Figure 6**: Correlation Heatmap. This shows the relationship between every two features from all columns. The repetitive part of the map has been removed. The red color (corresponding to +1) in the plot shows a perfect positive relationship, while blue (corresponding to -1) shows a perfect negative relationship. Short names have been used for feature names.

In **Figure 6**, we have a correlation heatmap. There are good relationships among the temperature group (min, max, avg). This is the case for wind speeds. There is an intermediate relationship between pressure and wind speed. There is a good relationship between humidity and visibility. There is an intermediate relationship between visibility/humidity and precipitation.

A graph of a graph

Description automatically generated with medium confidence

**Figure 7**: Moving Average Illustration For Trend Extraction. Moving average (blue series) has been done with a window of 10 days on selected features original data (gray series). The time span is from 2021-11-01 to 2023-10-31.

In **Figure 7**, we have used the moving average technique to lessen the fluctuations and the trends over time. We can see that there is a cyclic trend for temperature. So we expect up and down bumps in average temperature over time. And this is the case as well for maximum and minimum temperatures as there is a perfect relationship between temperature group members (max, min, average). We can see there has been a sharp decrease in temperature around March of 2023. The pattern for other features is not that cyclical but somewhat fluctuating. We can use moving average plots as a means for detecting outliers as well.

In **Figure 8**, we want to see what month, on average, has the most total precipitation during its days. As you can see, January is the month you can expect to see the most total precipitation during the month or maybe experience precipitation during most of its days. After January, we see an overall decline in total precipitation and then see a starting increase from August.

A graph of blue bars

Description automatically generated

**Figure 8**: Average Monthly Precipitation Bar Plot. Each bar for each specific month shows the average of total of daily precipitations (rain, snow equivalent, etc) during that month. The precipitation unit is mm.