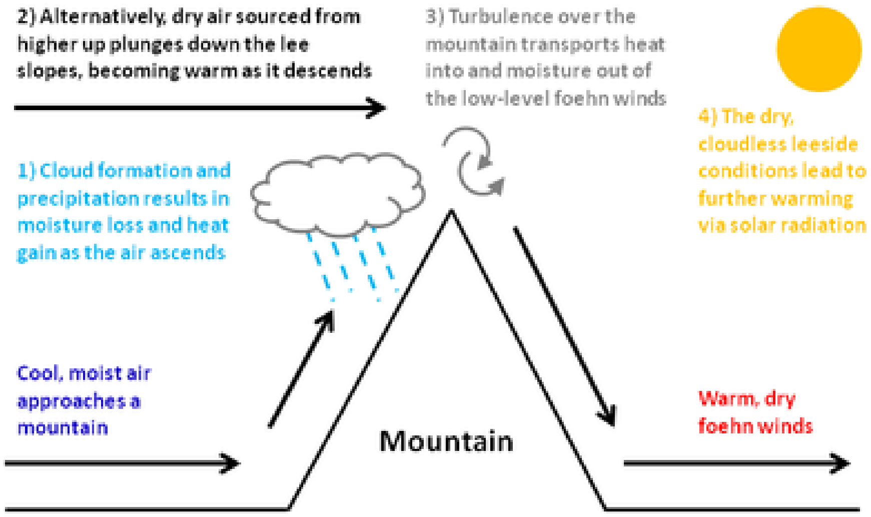
Calgary Temperature Model

Data 603: Statistical Modeling with Data
University of Calgary - Fall 2023
Group
Arie DeKraker, Claire McCallum, Victoria Nukiry

Introduction

When it comes to Alberta, it is uniquely positioned east of the Rockies in the North American interior to experience abrupt weather and temperature fluctuations caused by winds in the prairies. With the strong winds occurring at all times of the year, Chinooks are most noticeable in the winter with their namesake on full display, eating away at the snow. Known as a Foehn wind outside of Alberta, Chinooks occur one in three days in winter and “interrupt the cold artic air with extremely strong warming effects at the surface temperature level”. The rapid temperature shift is caused by the Foehn winds, which result from moist unstable western air that encounters the Rocky Mountain Range and is forced to ascend and cool.

As the Foehn winds descend into southwestern Alberta, temperatures rapidly change and local Calgarians enjoy clear skies and warm afternoons. See figure 1 for a visual that depicts some of the conditions that can cause a Foehn wind to form.


2) Alternatively, dry air sourced from higher up plunges down the lee slopes, becoming warm as it descends
3) Turbulence over the mountain transports heat into and moisture out of the low-level foehn winds
4) The dry, cloudless leeside conditions lead to further warming via solar radiation
1) Cloud formation and precipitation results in moisture loss and heat gain as the air ascends
Cool, moist air approaches a mountain
Mountain
Warm, dry foehn winds

The causes of Foehn winds are dependent on the mountain range, but can be additionally influenced by “meteorological conditions, such as the upstream wind speed, temperature, and humidity.” This project aims to evaluate the relationships between these variables and conditions, so that our model may be able to predict temperatures shifts.

The primary objective of this report is to create a multiple linear regression model that predicts Calgary temperature, with the aim that the ability to accurately forecast the temperature shifts will aid newcomers to learn how to acclimate to Calgary.

This topic is of interest to our team as we are all new-er residents of Calgary. We have each found that in acclimating to the environment, climate fluctuations can be unpredictable. This project will supplement our team’s growing familiarity with the weather in Calgary by modelling the predictability of the temperature.

Methodology

The objective of this project is to build a model for Calgary Temperature prediction and observe the relationships between temperature and the various variables. The model will reflect the material covered in Data 603: Statistical Modeling with Data class.

Dataset

The data sourced for this project is available for non-commercial use without extra permission. The data has been retrieved via weatherstats.ca data and is based on Environment and Climate Change Canada and data from the Citizen Weather Observer Program (CWOP). The data is collected from climate stations and records updates every 5 minutes. For a list of all Calgary Climate Stations, refer to appendix A. The data is summarized by day, except for average variables, which is average hourly record, summarized by day. Each row reflects the average of the variable recorded on the date listed in the column.

The data reflects Calgary weather records from January 2014 to September 2023. This is a population sample of the temperature records for the time represented. For the purpose of this project, we have selected a portion of the variables for analysis.

The data was cleaned by numbering the dates to allow for analysis of the time series data.

Variable Description

The variables that have been included in the scope of this project include the following. All variables are quantitative and continuous.

Dependent Variable

Average Hourly Temperature Data (quantitative): This is the average of all the hourly temperatures within the day and it is represented by ‘avg_hourly_temperature’ and measured in °C.

Independent Variables

Average Hourly Humidity (quantitative): Measurements of average daily humidity levels. It is represented by ‘avg_hourly_relative_humidity’ and measured in percentage of the present state of humidity compared to the maximum rate for that given temperature.

Average Hourly Dew Point (quantitative): The atmospheric temperature below which water droplets or dew is able to form. It is represented by ‘avg_hourly_dew_point’ and measured in °C

Average Hourly Wind Speed (quantitative): Information on the average wind speeds. It is represented by ‘avg_hourly_wind_speed’ and measured in km/h

Cloud Cover (quantitative): Data on the extent of cloud cover. It is represented by ‘avg_hourly_cloud_cover_8’ and is represented by Oktas which is a measurement for cloud coverage at any given location.

Barometric Pressure (quantitative): Readings of atmospheric pressure. It is represented by ‘avg_hourly_pressure_sea’ and measured in kPa.

Rain (quantitative): Measure of rainfall per day. It is represented by ‘rain’ and measured in millimeters (mm).

Snow (quantitative): Measure of precipitation type Snow in cm, column name is labeled as ‘snow’.

Date (quantitative): Measurement of dates, datatype datetime converted to ordered numerical data, represented as ‘date’.

Modelling Method

The method used to model the data is multiple linear regression to create a first order model as taught in Data 603.

To begin, we will create a full model, and then test for multicollinearity. One area of risk is the weather data may be correlated due to the fact that our variables may be related in ways outside our knowledge in meteorology. We will need to assess the coefficient estimates before and after removing any variables to assess the statistical significance of the variables. Once we remove highly correlated variables, we will apply an interval estimate for the coefficients to compare and then a stepwise analysis to build the recommended model. Once we have the recommended reduced model, we will create a t-test to evaluate the first order model against the reduced model.

Once we have determined the best model from the evaluation, we can test for interactions and higher order terms in the model and conduct an f-test for the significance of the terms. Once we have satisfied the model creation, we will test the following assumptions.

1. Linearity Assumption – Residual Plot of fitted values.
2. Independence Assumption – Residuals vs. Time for evidence of time-series correlation.
3. Equal Variance Assumption - Scale Location Plot, Breusch-Pagan test for Heteroscedasticity.
4. Normality Assumption – Shapiro Wilk Test, QQ plot, Histogram.
5. Multicollinearity – VIF Test
6. Outliers – Residuals & Leverage Plot, Cook's distance, Leverage Plot

If heteroscedasticity or non-normality is found, we will conduct a Box Cox transformation.

Project Methodology

Phase 1 – Project Preparation
Phase 2 – Rstudio Analysis
Phase 3 – Deliverables (Project Report, Project Presentation)

Team Member	Phase 1 Tasks	Phase 2 Tasks	Phase 3 Tasks
Arie	Checkpoint Deliverable	Model Selection/creation, variable T-tests, Interaction & Higher Order Term identification, Hypothesis statement & Anova Test, Linearity, Independence, Equal Variance, Normality, Multicollinearity tests, Outliers, Prediction.	Report: Methodology (Variable Description), code debugging and testing Results, Conclusion & Discussion
Claire	Terminology & Background, Checkpoint Deliverable		Report: Introduction, Methodology (Dataset, Modelling Method, Project Methodology), Results, Conclusion & Discussion
Victoria	Dataset Cleaning & Consolidation, Checkpoint Deliverable		Report: Model Creation, Code Writing and Testing, Conclusion & Discussion.

Analysis Results

Below are the results of our model testing in R.

Model Creation

To begin with creating our model, the first order model includes all variables as shown below.

First Order Model

$$\hat{Y}_{temp} = \beta_0 + \beta_1 \mathit{date} + \beta_2 \mathit{avg\_hourly\_dewpoint} + \beta_3 \mathit{avg\_hourly\_relative\_humidity} + \beta_4 \mathit{avg\_hourly\_wind\_speed} + \beta_5 \mathit{avg\_hourly\_pressure\_sea} + \beta_6 \mathit{avg\_hourly\_cloud\_cover\_8} + \beta_7 \mathit{snow} + \beta_8 \mathit{rain}$$

Reduced Model

After creating our full model, we wanted to see which independent variables would be best to use within our model, so we started looking at possible reduced models. When a summary is run on our first order model there are two variables that have a p-value higher than our alpha of 0.05, these being date (p-value= 0.193706) and avg_hourly_wind_speed (p-value= 0.106190). These variables are then dropped to create a reduced model. The summary was then rerun, on the reduced model, to see if any other variables would not be significant with the new model. From the resulting summary the only variable that returned a p-value higher than the alpha was avg_hourly_pressure_sea (p-value= 0.05858). This creates the second reduced model.

We then tested the two reduced models using an Anova test. From the resulting test we get a p-value of 0.05858 which is very close to our alpha value of 0.05. We also noticed that the adjusted p value is similar between the two models so to reduce the complexity of the model we chose to drop avg_hourly_pressure_sea, as well from our model. This gives us the reduced model provided below.

$$\hat{Y}_{temp} = \beta_0 + \beta_1 \mathit{avg\_hourly\_dewpoint} + \beta_2 \mathit{avg\_hourly\_relative\_humidity} + \beta_3 \mathit{avg\_hourly\_cloud\_cover\_8} + \beta_4 \mathit{snow} + \beta_5 \mathit{rain}$$

Interaction Model

$$i = \mathit{avg\_hourly\_relative\_humidity} * \mathit{avg\_hourly\_dew\_point},\ \mathit{avg\_hourly\_relative\_humidity} * \mathit{avg\_hourly\_cloud\_cover\_8},\ \mathit{avg\_hourly\_relative\_humidity} * \mathit{rain},\ \mathit{avg\_hourly\_relative\_humidity} * \mathit{snow},\ \mathit{avg\_hourly\_dew\_point} * \mathit{avg\_hourly\_cloud\_cover\_8},\ \mathit{avg\_hourly\_dew\_point} * \mathit{rain},\ \mathit{avg\_hourly\_dew\_point} * \mathit{snow},\ \mathit{avg\_hourly\_cloud\_cover\_8} * \mathit{rain},\ \mathit{avg\_hourly\_cloud\_cover\_8} * \mathit{snow},\ \mathit{rain} * \mathit{snow}$$

The next check was to see if there were any interaction terms of relevance that we wanted to include in our model. From the resulting summary on our interaction model there is one interaction term with a p-value higher than our alpha of 0.05, which is rain: snow (p-value=0.42060). The model was then refit, removing that term, and rerun to see if there were any other interaction terms/variables that would need to be dropped. When the summary is run again on the reduced interactive model there are a couple terms that no longer are significant to the model. Avg_hourly_dew_point*rain is found non-significant because it returns a p-value of 0.133559, avg_hourly_dew_point*snow is also shown not to be significant because it gives a p-value of 0.103987. Since both p-values are above 0.05, we can conclude they are not significant to the model. These two terms were then dropped to create the final interaction model represented below.

$$\hat{Y}_{temp} = \beta_0 + \beta_1 \mathit{avg\_hourly\_dewpoint} + \beta_2 \mathit{avg\_hourly\_relative\_humidity} + \beta_3 \mathit{avg\_hourly\_cloud\_cover\_8} + \beta_4 \mathit{snow} + \beta_5 \mathit{rain} + \beta_6 \mathit{avg\_hourly\_relative\_humidity} * \mathit{avg\_hourly\_cloud\_cover\_8} + \beta_7 \mathit{avg\_hourly\_relative\_humidity} * \mathit{avg\_hourly\_dew\_point} + \beta_8 \mathit{avg\_hourly\_relative\_humidity} * \mathit{rain} + \beta_9 \mathit{avg\_hourly\_relative\_humidity} * \mathit{snow} + \beta_{10} \mathit{avg\_hourly\_dew\_point} * \mathit{avg\_hourly\_cloud\_cover\_8} + \beta_{11} \mathit{avg\_hourly\_cloud\_cover\_8} * \mathit{rain} + \beta_{12} \mathit{avg\_hourly\_cloud\_cover\_8} * \mathit{snow}$$

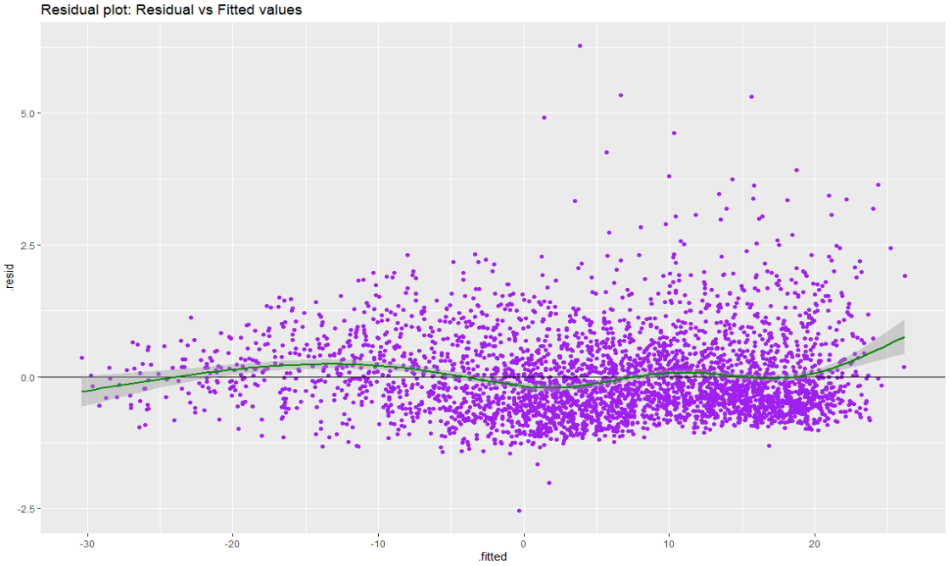
Refitted Model after Outlier Removal

After removing the influential outliers, a summary was reran on the interactive model. The results from the summary indicated that most of the interactive terms remained significant, but there was one term that lost its significance after the change. This being avg_hourly_dew_point*avg_hourly_cloud_cover_8, which returns a p-value of 0.012744. Since this is below our alpha value of 0.05, we can remove it as it is no longer significant to our model. Below represents the final model.

$$\hat{Y}_{temp} = \beta_0 + \beta_1 avg\_hourly\_dewpoint + \beta_2 avg\_hourly\_relative\_humidity + \beta_3 avg\_hourly\_cloud\_cover\_8 + \beta_4 snow + \beta_5 rain + \\ \beta_6 avg\_hourly\_relative\_humidity * avg\_hourly\_cloud\_cover\_8 \ + \beta_7 avg\_hourly\_relative\_humidity * avg\_hourly\_dew\_point \ + \\ \beta_8 avg\_hourly\_relative\_humidity * rain \ + \beta_9 avg\_hourly\_relative\_humidity * snow \ + \\ \beta_{10} avg\_hourly\_cloud\_cover\_8 * rain \ + \beta_{11} avg\_hourly\_cloud\_cover\_8 * snow$$

Linearity Assumption

To determine if our data met the linearity assumption, we plotted the residuals v the fitted values to determine if there is a pattern in the data.



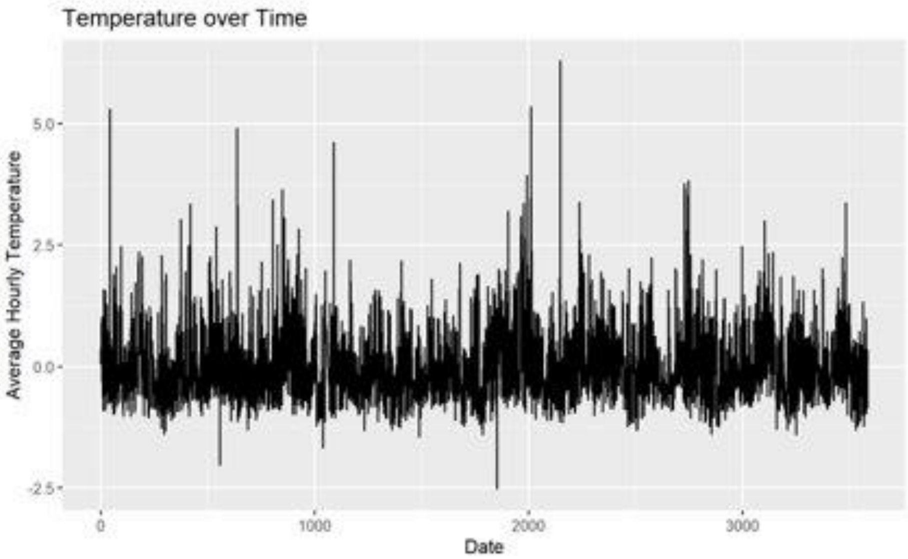
When observing our residual plot, at first, we interpreted the data as linear, upon closer look there are four concerns;

1. Residuals are not symmetrically distributed,
2. There is clustering on the right of the plot,
3. There appears to be a Cone pattern with the distribution increasing in volume and variance as the y-axis increases.
4. Heavy-tailed.

This leads us to believe that the data is not linear, but due to the trend line demonstrating a fairly linear depiction, we revert to our linear conclusion.

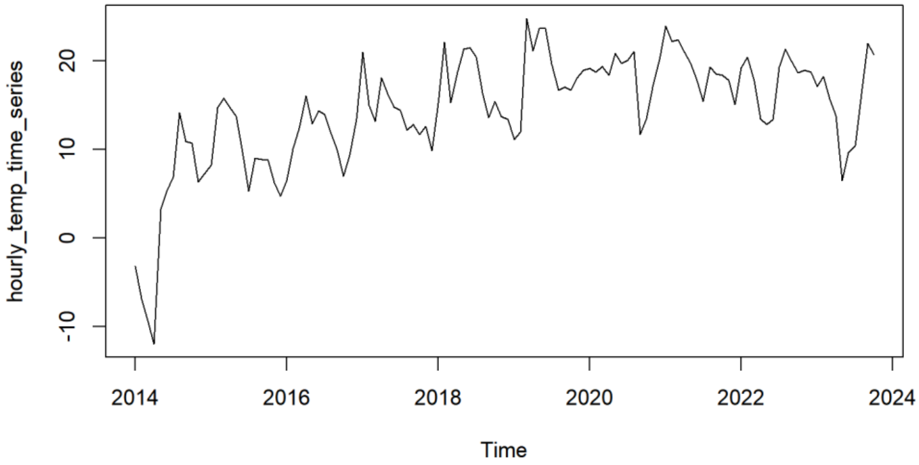
Independence Assumption

When reviewing the independence assumption, this is where we believed we would begin to encounter challenges as our data is time series and each weather observation is likely to be related due to the frequency and proximity in time to each observation in the population sample. Each record in the data reflects the averages from the date recorded, at a similar location, summarized by day. Meaning we look to identify any patterns in the data caused by the sequential nature of the weather data. To look at the independence assumption and determine if they were correlated, we plotted the regression residuals against time.



When observing the plot, we noted that the residuals depicted a subtle trend with deep valleys in an equally spaced cyclical pattern. We determined our data does have a serial correlation and it is not independent. The impact of this conclusion is that the confidence intervals and p-values will likely be incorrect, and we cannot trust further conclusions in our model, as the biases in our data will be undetectable.

And we thought about adding a graph that reflects the time series against the average hourly temperature.



To overcome the serial correlation in our data, this could be resolved with a more complex model that is able to address the biases in the time series weather data. Although this is outside the scope of this class, a potential weather forecasting model would be to use the Regression Learner application in MATLAB to normalize our data. This is explored using weather data by Gurwinder Singh & Harun in 'Advances in Computing and Data Sciences'.

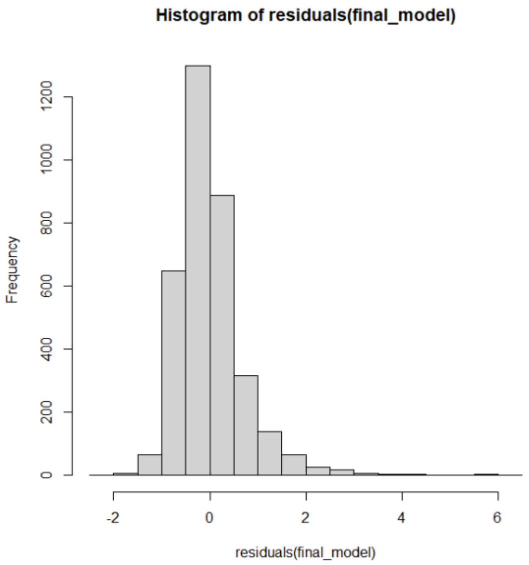
Equal Variance Assumption

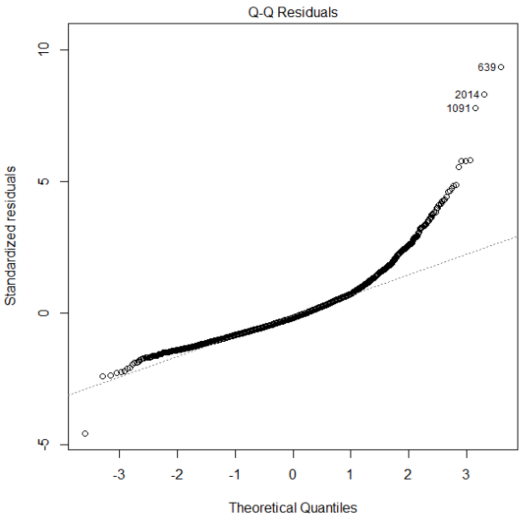
From the results of our Breusch-Pagan test we can see that we have a p-value below that of the alpha of 0.05. With these results we reject the null hypothesis that there is homoscedasticity shown within the model. With this we can conclude that the model expresses heteroscedasticity and we do not meet our equal variance assumption.

We can further prove this assumption by looking at the results of our scale-location plot. From the plot we see that the plots do not look to be randomly distributed. While there is a good amount that falls be the horizontal line, there is enough that deviate from the line to suggest that there is not homoscedasticity within the model.

Normality Assumption

To test the normality assumption below is a histogram plot of the residuals and a QQ plot. When interpreting the visuals, we note that the data is not normal. There is a noticeable spike of identical values is seen on the QQ plot showing that the residuals are heavy tailed, and right skewed. The histogram is not symmetrical and depicts the data as skewed. The next step is to attempt to normalize this data with a Box Cox transformation.

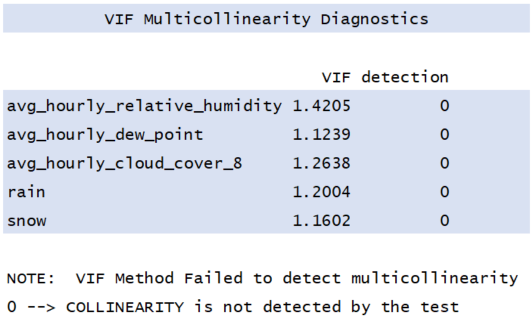




However, due to our dataset being temperatures, which are recorded in the negative, we cannot run a box cox transformation on our model. With this outcome we are only able to conclude with the assumption that our data does not express normality.

Multicollinearity Tests

Given the dataset, we anticipated there would be high correlation and the VIF test would detect multicollinearity among our variables. To our surprise, the VIF test did not detect multicollinearity.

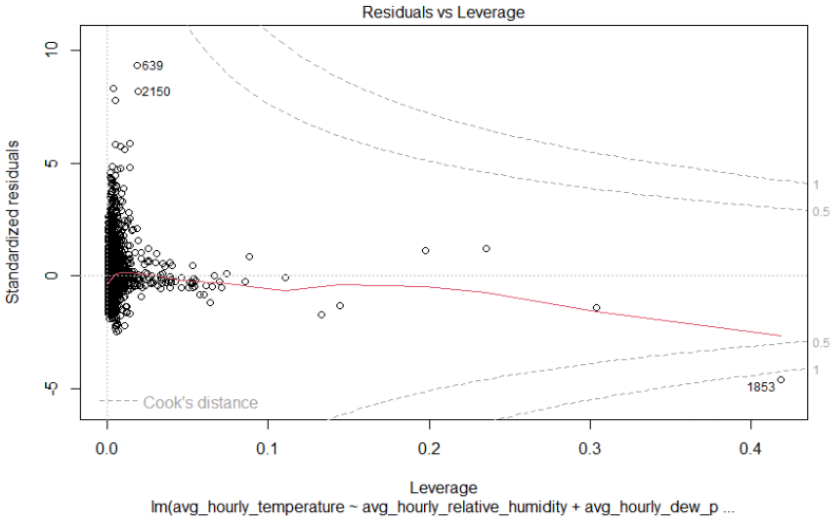


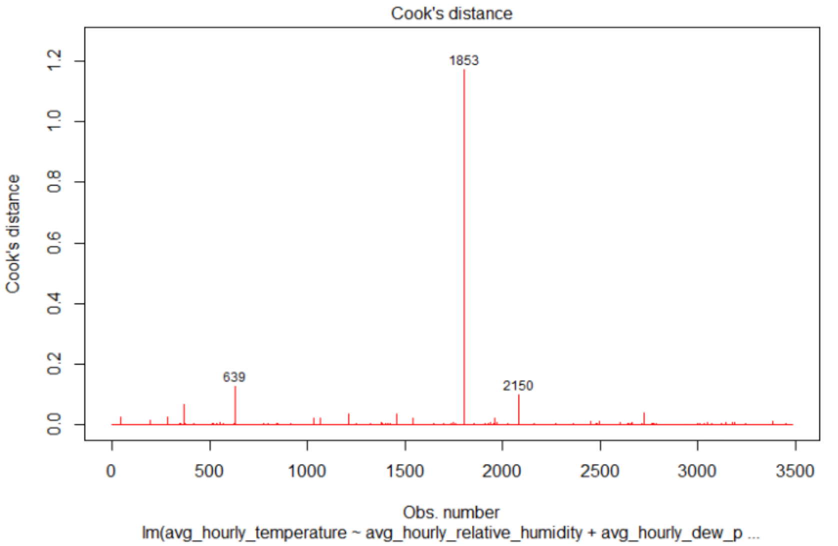
```
              VIF Multicollinearity Diagnostics

                                VIF detection
avg_hourly_relative_humidity 1.4205         0
avg_hourly_dew_point         1.1239         0
avg_hourly_cloud_cover_8     1.2638         0
rain                         1.2004         0
snow                         1.1602         0

NOTE:  VIF Method Failed to detect multicollinearity
0 --> COLLINEARITY is not detected by the test
```

This is explained as the undetectable bias in the non-independent data. When independent data are linearly correlated, this is the cause of multicollinearity. This multicollinearity can cause our coefficient estimates to become sensitive to minor changes in the model and weakens the statistical power of the regression model. Given our data is not independent we cannot measure the strength of the correlation in our variables.

Outliers

To resolve some of the assumptions, we moved to review our outliers and remove them from our dataset. We conducted a Residuals v Leverage Plot and Cook's distance.





In the residuals and leverage plot the point 1853 falls outside the cooks' distance and may be an influential point and this observation should be removed from the dataset, and the model needs to be refit. When plotting the Cooks distance, it shows that observation 1853 has an exceptionally large cooks' distance, and is above 1.0, and is larger than the values clustered at 0.25. This is an influential outlier. In reviewing the 3pn outliers, we have numerous outliers and remove them from the dataset.

Once the data is removed, the model has been refitted and tested through each assumption. The results have not improved and each of the assumptions have the same conclusions.

Conclusion and Discussion

Several steps were taken to create our final model. We initially started off with model fitting. We tried to find the best reduced, interactive, and final model within this step by employing the use of the summary command & inspecting p-values and running anova tests to see if variables/terms between two models were significant to the model or not. At the end of these steps, we got a model which is represented below:

$$\hat{Y}_{temp} = \beta_0 + \beta_1 avg\_hourly\_dewpoint + \beta_2 avg\_hourly\_relative\_humidity + \beta_3 avg\_hourly\_cloud\_cover\_8 + \beta_4 snow + \beta_5 rain + \beta_6 avg\_hourly\_relative\_humidity * avg\_hourly\_cloud\_cover\_8 + \beta_7 avg\_hourly\_relative\_humidity * avg\_hourly\_dew\_point + \beta_8 avg\_hourly\_relative\_humidity * rain + \beta_9 avg\_hourly\_relative\_humidity * snow + \beta_{10} avg\_hourly\_cloud\_cover\_8 * rain + \beta_{11} avg\_hourly\_cloud\_cover\_8 * snow$$

The next steps that we take were to address if the model met the assumptions. Several plots were created, and a few commands were run to test for these assumptions. The results from the analysis showed that the model failed the independence, normality, and equal variance assumption. We attempted to fix our normality, and equal variance assumption using a Box Cox transformation, but due to the temperature values of our dataset we are unable to run a normal Box Cox transformation, since the transformation can only take positive nonzero numbers. The only assumption that the model passes is the linearity assumption.

After the analysis, the best estimation model was found to be the one included below:

$$\hat{temperature} = 29.1439217 - 0.3472992 avg\_hourly\_relative\_humidity + 1.1496328 avg\_hourly\_dew\_point - 0.8068919 avg\_hourly\_cloud\_cover\_8 - 0.1362033 rain - 0.2245344 snow + 0.0131425 avg\_hourly\_relative\_humidity \times avg\_hourly\_cloud\_cover\_8 - 0.0015815 avg\_hourly\_relative\_humidity \times avg\_hourly\_dew\_point + 0.0018573 avg\_hourly\_relative\_humidity \times rain + 0.0047524 avg\_hourly\_relative\_humidity \times snow - 0.0173666 avg\_hourly\_cloud\_cover\_8 \times snow$$

we also tried removing the outliers from the data hoping this way could make better model so after removing the outliers we created a new model after reducing it but the new model had the same issues

So:

- Data is not independent
- Data is not normal and cannot perform a Box Cox transformation (as it has negative values).
- We have outliers.
- It is not equally variance.
- But we had the linearity assumption met

So, we cannot say we have a linear model

There are a few approaches we can take to further our analysis and understanding of weather patterns in Alberta. While the Box Cox transformation is limited to non-negative or zero numbers, there are some libraries that might be able to handle this condition. The Yeo-Johnson transformation might be of interest since it is described as a method to run Box Cox transformations without restrictions.

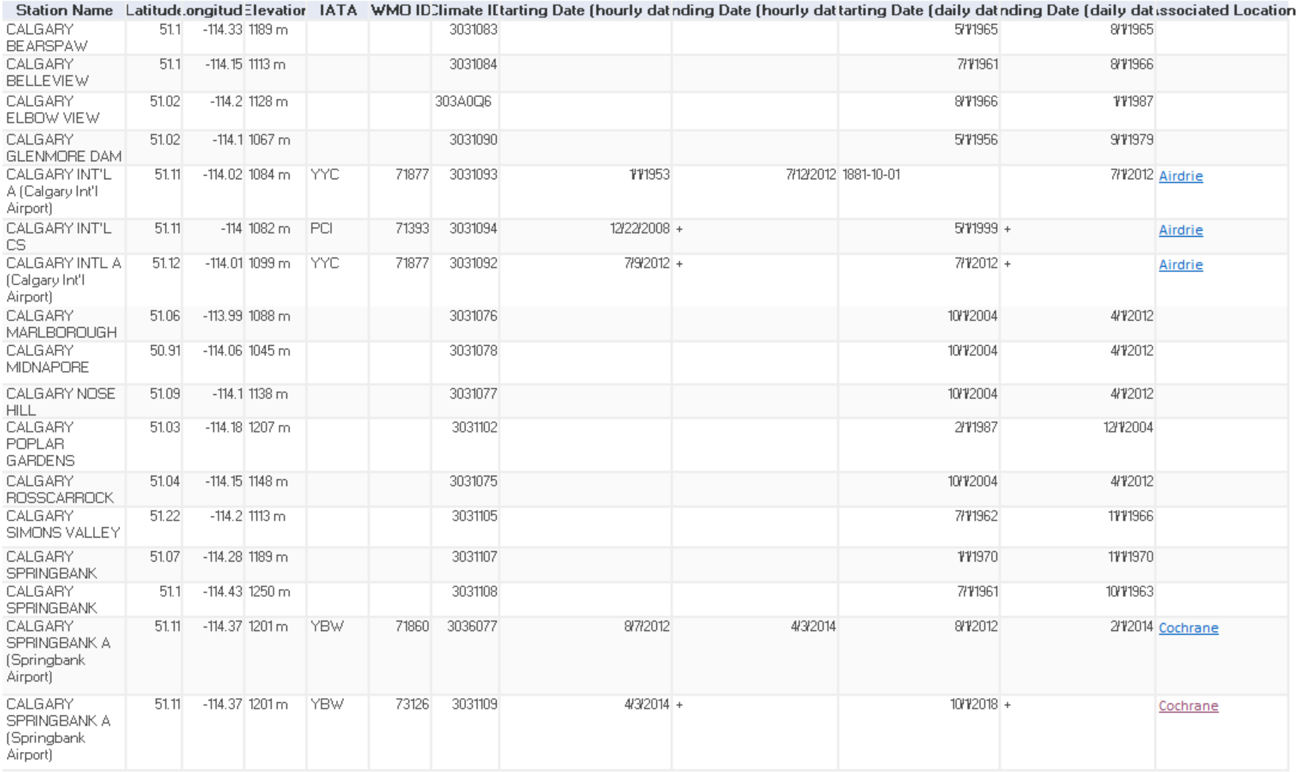
As the main purpose of our project is Forecasting so the main question is Predicting future values of a particular variable (temperature) based on historical data.

And for forecasting tasks, time series models like ARIMA, SARIMA, or even LSTM (a type of neural network) might be suitable.

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It is really simplified in terms of using it, yet this model is powerful.

Appendix

A – Calgary Climate Stations



Station Name	Latitud	ongitud	levatior	IATA	WMO ID	limate I	tarting Date (hourly dat	nding Date (hourly dat	tarting Date (daily dat	nding Date (daily dat	ssociated Location
CALGARY BEARSPAW	51.1	-114.33	1189 m			3031083			5/1/1965	8/1/1965	
CALGARY BELLEVIEW	51.1	-114.15	1113 m			3031084			7/1/1961	8/1/1966	
CALGARY ELBOW VIEW	51.02	-114.2	1128 m			303A0Q6			8/1/1966	1/1/1987	
CALGARY GLENMORE DAM	51.02	-114.1	1067 m			3031090			5/1/1956	9/1/1979	
CALGARY INT'L A (Calgary Int'l Airport)	51.11	-114.02	1084 m	YYC	71877	3031093	1/1/1953	7/12/2012	1881-10-01	7/1/2012	Airdrie
CALGARY INT'L CS	51.11	-114	1082 m	PCI	71393	3031094	12/22/2008	+	5/1/1999	+	Airdrie
CALGARY INTL A (Calgary Int'l Airport)	51.12	-114.01	1099 m	YYC	71877	3031092	7/9/2012	+	7/1/2012	+	Airdrie
CALGARY MARLBOROUGH	51.06	-113.99	1088 m			3031076			10/1/2004	4/1/2012	
CALGARY MIDNAPORE	50.91	-114.06	1045 m			3031078			10/1/2004	4/1/2012	
CALGARY NOSE HILL	51.09	-114.1	1138 m			3031077			10/1/2004	4/1/2012	
CALGARY POPLAR GARDENS	51.03	-114.18	1207 m			3031102			2/1/1987	12/1/2004	
CALGARY ROSSCARROCK	51.04	-114.15	1148 m			3031075			10/1/2004	4/1/2012	
CALGARY SIMONS VALLEY	51.22	-114.2	1113 m			3031105			7/1/1962	11/1/1966	
CALGARY SPRINGBANK	51.07	-114.28	1189 m			3031107			1/1/1970	11/1/1970	
CALGARY SPRINGBANK	51.1	-114.43	1250 m			3031108			7/1/1961	10/1/1963	
CALGARY SPRINGBANK A (Springbank Airport)	51.11	-114.37	1201 m	YBW	71860	3036077	8/7/2012	4/3/2014	8/1/2012	2/1/2014	Cochrane
CALGARY SPRINGBANK A (Springbank Airport)	51.11	-114.37	1201 m	YBW	73126	3031109	4/3/2014	+	10/1/2018	+	Cochrane

Citations

1. “Chinook Wind.” (2023) *Wikipedia*, Wikimedia Foundation, 30 Oct. 2023, en.wikipedia.org/wiki/Chinook_wind.
2. “Foehn Wind.” *Wikipedia*, Wikimedia Foundation, 22 Nov. 2023, en.wikipedia.org/wiki/Foehn_wind. Accessed 2023.
3. Nkemdirim, L.c. “Chinook.” *The Canadian Encyclopedia*, 6 Feb. 2006, www.thecanadianencyclopedia.ca/en/article/chinook. Accessed 2023.

4. "Weather Dashboard for Calgary." (2023) *Amateur Weather Statistics for Calgary, Alberta*, calgary.weatherstats.ca/. Accessed 2023.
5. Weisberg, Sanford. "Yeo-Johnson Power Transformations." *University of Minnesota*, 26 Oct. 2001, www.stat.umn.edu/arc/yjpower.pdf.
6. Wolfe, S.A. et al. "Foehn." (2023) *Foehn - an Overview | ScienceDirect Topics*, ScienceDirect, www.sciencedirect.com/topics/earth-and-planetary-sciences/foehn. Accessed 2023.
7. Pathak, Prabhat. "Building an ARIMA Model for Time Series Forecasting in Python." Analytics Vidhya, 29 Oct. 2020, https://www.analyticsvidhya.com/blog/2020/10/how-to-create-an-arima-model-for-time-series-forecasting-in-python/
8. Singh, Gurwinder, and Harun. "Multiple Linear Regression Based Analysis of Weather Data for Precipitation and Visibility Prediction." *SpringerLink*, Springer Nature Switzerland, 1 Jan. 1970, link.springer.com/chapter/10.1007/978-3-031-37940-6_6.