C O V E N T R Y

U N I V E R S I T Y

Faculty of Engineering, Environment and Computing

School of Computing, Mathematics and Data Science

MSc. Data Science

7150CEM

Individual Research Project

Temperature Forecasting in Canada’s Capital

Author: Norhaan Saadawi

SID: 14538007

1st Supervisor: Dr. Maha Moustafa

2nd Supervisor: Dr. Omid Chatrabgoun

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science

Academic Year: 2024/25

Declaration of Originality

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: Norhaan Saadawi Date: 2024 12 10

Please complete all fields.

First Name:	Norhaan
Last Name:	Saadawi
Student ID number	14538007
Ethics Application Number	P182047
1st Supervisor Name	Dr. Maha Moustafa
2nd Supervisor Name	Dr. Omid Chatrabgoun

This form must be completed, scanned and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

Abstract

This project investigated machine learning models autoregressive integrated moving average (ARIMA), seasonal autoregressive integrated moving average with exogenous variables SARIMAX, and long short-term memory (LSTM) ability to forecast hourly temperatures for the Canadian winter season. This project used historical weather data covering the years 2019 to 2023 from the Ottawa CDA RCS weather station to fit and test the models. Evaluation metrics that were used to evaluate performance included mean-squared error (MSE) and mean absolute error (MAE). The findings revealed that LSTM performed the best, achieving an MSE of 0.987 and an MAE of 0.654. Overall, the research conducted in this project supports the use of machine learning for long term seasonal temperature prediction to aid with enhancing climate resilience.

Table of Contents

List of Figures

List of Tables

List of abbreviations

Acronym	Definition
ACF	Autocorrelation Function
AIC	Alkaline Information Criterion
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
CCS	Carbon Capture and Storage
CNN	Convolutional Neural Network
GDP	Gross Domestic Product
GHG	Greenhouse Gas
K-NN	K- Nearest Neighbors
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MSE	Mean Squared Error
PACF	Partial Autocorrelation Function
RMSE	Root Mean Squared Error
SARFIMA	Seasonal Autoregressive Fractionally Integrated Moving Average
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Autoregressive Integrated Moving Average with Exogenous Variables
SVM	Support Vector Machines

1 Introduction

Temperature prediction using machine learning to assist with understanding climate trends has become a popular topic over the past few years (Fister et al, 2023). Temperature prediction is an important area to study as it has many different applications across many different fields (Fister et al, 2023). Some fields that temperature prediction can apply to include energy management, agriculture, urban planning, and health and safety (Fister et al, 2023; Tyler et al, 2024). As temperature affects many areas in our lives including our food sources, environment, our health and our future, it is important that we understand how it is changing to prepare adequately for the future (NCAS, 2023; Tyler et al, 2024).

Climate change refers to the long term pattern of weather conditions (UN, 2024b). Climate change is currently a rapidly growing issue that comes with many consequences that different populations may not be prepared to handle (NCAS, 2023). One significant effect that has come from climate change is the constant rise in long term average temperatures (Fister et al, 2023). However, this not only affects long-term temperature patterns, but it also affects the short-term. This causes extreme changes from normal temperature levels during the short-term which can include increased heat waves or intense droughts (Fister et al, 2023). These extreme events have global consequences and the better prepared we are for them, the easier the future will be (NCAS, 2023).

Machine learning is a field in computer science that deals with using data and algorithms to allow artificial intelligence to learn the same way humans learn (IBM, 2024). Machine learning works in three parts (IBM, 2024). The first is the decision process, where input data is provided to the algorithm, producing an estimate regarding a pattern in the data (IBM, 2024). The second is error function, where the model is evaluated (IBM, 2024). Last, is the model optimization process where parameters are adjusted to see if the model can fit better to the training data (IBM, 2024). If the model can fit better then discrepancies will be reduced between the model predictions and the test data enhancing accuracy (IBM, 2024). This stage can be repeated until an accuracy threshold is met (IBM, 2024).

It has been found in the literature that climate change is primarily controlled by changes in temperature and precipitation, thus research that aims to assess climate change is often based on analyzing temperature or precipitation time series (Khediri, 2022). Air temperature data when analyzed for seasonal climate predictions however does come with challenges, due to the long term predictions involved (Fister et al, 2023). However, much of the literature has found that approaching the data with machine learning could address the associated challenges (Fister et al, 2023). To address this theory, this project will explore and evaluate the performance of 3 different machine learning models to predict hourly air temperature for one winter season ahead.

1.1 Background

To assist with climate change and its effects, temperature prediction can play a large role in promoting climate resilience across many areas such as economic planning, energy management, health and safety, and urban planning. Below are some examples of how climate resilience can be achieved using temperature prediction within these areas:

1.1.1 Economic Planning

Temperature prediction is important in economic planning in helping individuals, businesses, and governments in foreseeing different challenges and making knowledgeable decisions to reduce the possible consequences. Temperature prediction can help in many different industries that fall under economic planning with some examples being the agriculture industry, tourism industry, and the aviation industry.

In agriculture, farmers depend on accurate forecasts for planning planting and harvesting schedules (EPA, 2024b). Accurate forecasts also help farmers know the type of farming method that will work best whether it is indoor farming or outdoor farming for the expected forecast.
This is important in regions that may experience periods of extreme heat or extreme frost suddenly which can ruin crops (UN, 2024).Temperature prediction that can provide long term forecasts can assist farmers in assessing the type of crops they can farm and the type of farming method that will be used for the upcoming season (EPA, 2024b). This early planning can help farmers account for

potential crop shortages that may occur due to extreme weather events (EPA, 2024b). By knowing about potential crop shortages ahead of time, markets, businesses, and governments that depend on certain crops can prepare for international imports if needed without having last minute shocks (EPA, 2024b). Additionally, knowing about crop shortages ahead of time can also allow markets to input household limits on certain crop purchases ahead of time or use crop-safe chemicals to extend freshness to reduce the impact the shortage may have on the society. Businesses, markets, and governments can also identify times where high demand crops can grow best ahead of time and plan for high production to help with international exports and to use in freezing if needed for the next season (EPA, 2024b). Overall, it is estimated that worldwide economic losses from food and water shocks due to extreme weather events are expected to reach USD 771 billion over the next five years thus it is essential that we try to explore different solutions to reduce the consequences from extreme weather events (Tyler et al, 2024).

In tourism, temperature prediction can help reduce health risks and costs from extreme weather events and help businesses prepare for shifting tourism seasons. In Greece, travel and tourism account for 15% of gross domestic product (GDP), however extreme heat is affecting the tourism industry and has caused significant disruptions (Shine, 2023). Rhodes island suffered from wildfires which led to over 2000 tourist evacuations, while the Acropolis in Athens, the top tourist attraction, had to temporarily close due to temperatures reaching 45°C causing a heatwave (Shine, 2023; Smith, 2023). In response to this extreme heat wave, 30,000 water bottles were distributed by noon, first aid rescue workers were sent to help collapsed and fainted individuals, and shading areas were built at the attraction entrance (Smith, 2023). As these extreme heat events become more common, tourists will start to change their travel plans (Edenido, 2023; Scott & Lemieux, 2010). Traditional beach vacations during the summer seasons may change to vacations to cooler areas during the summer seasons (Edenido, 2023; Scott & Lemieux, 2010).This can shift tourism patterns and provide sudden economic shifts in both colder areas and warm areas that the areas may not be prepared for (Edenido, 2023; Scott & Lemieux, 2010; World Travel and Tourism Council, 2022). These difficult events highlight the importance of using temperature prediction for long term planning as well as a need for climate adaptation strategies. Long term temperature prediction can help areas expect rising temperatures during certain seasons plan timely to explore different measures to address the climate's impacts.

Measures can include promoting off season tourism or developing alternative attractions to maintain the same number of visitors without climate or economic related consequences (Scott & Lemieux, 2010; World Travel and Tourism Council, 2022). Examples of alternative attractions that can be offered in warm areas can include indoor activities, nighttime events, or water-based activities, however they would need business planning in advance which temperature prediction can help in. Temperature prediction can also help identify extreme forecasts which can be made to advise tourists ahead of time to avoid these time periods to ensure safe vacationing.

In the aviation industry, extreme weather events can significantly impact different operations. In Las Vegas, temperatures have reached 46°C which led to airlines reducing passengers, restricting baggage, and limiting fuel to ensure takeoff safety (Helmore, 2023). These aviation adjustments along with extreme changes in climate in areas dependent on travel and tourism show how temperature prediction can help reform operational strategies to minimize delays and maintain safety.

Overall, temperature prediction can help individuals, businesses, governments, and policymakers plan infrastructure and resource management to aid in economic stability and appropriately respond to the challenges of the changing climate. In conclusion, the involvement of temperature prediction in decision making in economic planning can enhance climate resilience and create sustainable systems that can respond to climate variability.

1.1.2 Energy Management

Temperature prediction is important in the energy management sector as changes in the climate introduce many different challenges concerning energy transmission, consumption, and production (University of Cambridge, 2014). The increase in global temperatures along with more frequent and severe extreme weather events has a direct impact on energy systems (University of Cambridge, 2014). This affects fossil fuel supply, hydropower and thermal generation, and transmission (University of Cambridge, 2014). Thermal power plants currently provide about 80% of worldwide electricity and are made to work in different weather conditions, however they have started to face challenges in thermal conversions due to the changing temperatures (University of Cambridge, 2014). Along with this, energy demand is increasing worldwide which increases greenhouse gas (GHG) emissions which further worsens climate challenges (University of Cambridge, 2014).
To address climate challenges, reducing GHG emissions is essential (University of Cambridge,

2014). This can be done by reducing emissions coming from fossil fuel extraction and conversion, using lower carbon fuels, improving transmission and distribution system efficiency, and using renewable energy or nuclear power (University of Cambridge, 2014). A carbon capture and storage (CCS) model can also be implemented to further help with reducing energy demand and reducing emissions (University of Cambridge, 2014). Overall, these strategies along with temperature prediction can increase the resilience of energy systems when responding to climate changes and still provide reliable and high quality energy, reducing environmental damage.

1.1.3 Health and Safety

Changes in climate have many impacts on health and safety but can be broadly categorized into two concerns. The first being existing health problems being worsened, and the second being new health challenges being created (EPA, 2024c). As extreme weather events become more frequent and intense, it is estimated that between 2030 and 2050 about 250,000 additional deaths will occur yearly due to the extreme weather event consequences (EPA, 2024c; WHO, 2023).

Existing health problems that have been increased by changes in the climate include respiratory and heart diseases, weather-related illnesses, mental health issues, and vector-borne diseases (EPA, 2024c; UN, 2024a). It is important to understand the changes in the climate that are occurring in order to anticipate future risks for known health problems. Temperature prediction can assist health practitioners and communities prepare for the impacts of extreme weather for different health problems. For example, when a heatwave occurs it can cause worsened cardiovascular or respiratory disease, heatstroke, mortality, or heat exhaustion (WHO, 2024). By knowing the days likely to have heat waves, healthcare groups can prevent unnecessary hospital admissions and prepare resources in advance (WHO, 2024). This is extremely important as the population continues to age and become more vulnerable to the effects of extreme weather events (WHO, 2023).

Regarding the creation of new health challenges, invasive species including pests and pathogens that cannot survive in their previous habitats due to changes in weather are migrating to different areas (USGS, 2024). This can lead to disruptions in the food supply chain and cause different diseases to spread in areas not previously seen before (UN, 2024a). This may cause sudden shocks or challenges that are not expected by scientists which will need further research to address. Additionally, it has also been found that 65% of insects could face extinction in the next century due to changes in temperature, possibly leading to significant effects on ecosystems and public health (NASA, 2022). Temperature prediction can help in this area by guiding scientists to know which regions to conduct additional research for and possibly the type of species before the effects of the extreme weather are too late to reverse.

Overall, temperature prediction can help health systems include climate resilience in health infrastructure and policies, ensure vulnerable populations are protected, and create targeted interventions to reduce some of the consequences ahead of time (WHO, 2023). By knowing the temperature trends and possible impacts ahead of time, health groups can aim to promote tailored health guidelines and respond appropriately to the health threats that may come.

1.1.4 Urban Planning

Temperature prediction in urban planning can be very useful in guiding urban planners to know which designs are better across different regions to withstand the climate changes coming ahead. The urban heat island effect, for example, occurs when structures absorb and re-emit the sun's heat causing certain urbanized areas to be hotter than other areas (EPA, 2024d). This is harmful to society as it increases air pollution, heat-related illnesses and deaths, worsens naturally occurring heat waves, and contributes to higher daytime temperatures and reduced nighttime cooling (EPA, 2024a). Along with climate changes, the consequences from urban heat islands will continue to worsen thus addressing them is necessary (EPA, 2024a; Shalaby & Aboelnaga, 2017). Climate-adaptive architecture that focuses on enhancing airflow, integrating cooling features, and includes sustainable construction such as green roofs can help reduce the effects of climate change on society (Argonne, 2024; Shalaby & Aboelnaga, 2017). However, construction projects take time, thus including temperature prediction into the planning and risk and vulnerability assessments can assist in building design (Shalaby & Aboelnaga, 2017).

Conventional urban systems often rely on using concrete and asphalt in their designs (Shalaby & Aboelnaga, 2017). This can increase risks of flooding during intense rainfall (Shalaby & Aboelnaga, 2017). Along with the increased frequency and intensity of extreme weather events in the changing climate, these structures face high vulnerability to infrastructure damage and urban flooding (Shalaby & Aboelnaga, 2017). Along with temperature prediction, constructing porous surfaces, better stormwater management systems, updated flood-risk maps, and designing raised streets and buildings can help with flood resilience (European Environment Agency, 2024; Shalaby & Aboelnaga, 2017).

Overall, temperature prediction can assist urban planners with building resilient infrastructure and guide them on the regions in need of resilient infrastructure and the approximate time frames needed to reduce the consequences that could arise from the climate changes (Shalaby & Aboelnaga, 2017).

1.2 Project Overview

This project is organized in 12 Chapters. The organization starts with Chapter 1 covering the introduction, Chapter 2 covering the literature review, Chapter 3 covering the methodology, Chapter 4 covering the requirements, and Chapter 5 covering the dataset. Once the reader understands the reasoning behind the project and the theory, we dive into Chapter 6 to explain the design and go in depth with explanations of the project application including pre-processing techniques and machine learning modelling. Chapter 7 introduces the reader to the techniques, Chapter 8 covers the implementation, and Chapter 9 covers the results. This brings the project to the concluding chapters going over project management in Chapter 10, the critical appraisal in Chapter 11, and lastly conclusions in Chapter 12.

1.2.1 Aims

This project aims to accurately predict hourly air temperature for an upcoming winter season using the previous 4 winter seasons weather data. This project will compare different machine learning models, autoregressive integrated moving average (ARIMA), seasonal autoregressive integrated moving average with exogenous variables (SARIMAX), and long short-term memory (LSTM) to determine which has the best performance. The aims will be achieved by using the objectives to guide the project. The research questions will determine whether the aims have been met or further work needs to be done in the area of air temperature prediction using machine learning to assess climate trends.

1.2.2 Objectives

The objectives for this project are presented below:

- To understand the relationships between different variables included when recording weather data
- To visualize and identify patterns in temperature and variables highly correlated with temperature
- To develop and deploy ARIMA, SARIMAX, and LSTM models to perform hourly air temperature prediction
- To compare the different prediction models to determine which performs the best based on different evaluation metrics

1.2.3 Research Questions

The research questions that this project aims to answer are presented below:

- How accurately can different prediction models forecast temperature with minimal historical data provided and focusing on only one season of the year?

- Can any trends or patterns be found within temperature data covering only one season of the year over 5 years to show signs of climate change within Ottawa, Canada?

1.2.4 Purpose and Gap Analysis

The purpose of choosing to analyze and predict temperature for the Canadian winter season is due to the many challenges that arise during this season (Howes, 2024). A better understanding regarding the patterns that this season follows and the ability to predict the approximate temperatures for the next winter season can help Canadians to better prepare for the season in many different ways. This model can help identify extreme temperatures ahead of time which can allow for better preparations for icy conditions, snowstorms, frostbite warnings, energy demand and distribution, road safety, and frost protection for certain crops (Omstead, 2024). Additionally, winter in Canada is a very important season for wildlife and with the warming climate, this is causing many issues for the different species (Howes, 2024). Some animals need the snow to hunt, some are more mobile in it and rely on it to keep them warm, some have their fur adapted to the white snow color, some need snow for hibernation, and some need ice on lakes or rivers for migration (Howes, 2024). Thus, it is important that with the warming climate, temperature prediction would be a good tool to use to understand how the weather is changing and prepare for how the next winter season will display itself.

This research also addresses the gap in seasonal forecasting. It is often hard to predict seasonal temperatures as there are many factors that can impact weather changes in between seasons that may need to be accounted for (Fister et al, 2023). This research will test if seasonal winter data can be predicted with only winter data involved in model training.

2 Literature Review

Conventional weather forecasts are based on collecting current weather station data, giving the data to numerical weather prediction models, then providing the numerical output to meteorologists for refining and production of forecasts to the public (Met Office, 2024; Price et al, 2024). Forecast updates are then provided at different intervals based on the region and how far they differ from the original forecast (Met Office, 2024). In Canada, forecast updates are provided three times a day if there are changes, and the accuracy of the weather forecast from day 1 to day 7 ranges from 92% for the first day to around 60% on the 7th day (Government of Canada, 2023). This reduction in accuracy by the 7th day provides an opportunity for researchers to experiment with machine learning models to determine whether machine learning models can predict temperature for longer timeframes and for greater accuracy across all timestamp predictions. The literature provides much research done using different prediction models to predict temperature with many valuable and positive results. Some of the related literature is reviewed below:

Research done by Mung and Phyu aimed to evaluate different deep learning techniques' ability to predict one month's weather conditions (Mung & Phyu, 2023). Weather conditions that the researchers focused on for the predictions included minimum and maximum temperature, humidity, and wind speed (Mung & Phyu, 2023). The different deep learning techniques evaluated in this study include convolutional neural networks (CNN), LSTM, and an ensemble of CNN and LSTM (Mung & Phyu, 2023). The models were trained using daily weather data covering 20 years (Mung & Phyu, 2023). The models were compared for performance based on root mean squared error (RMSE), and the results found that the ensemble of CNN and LSTM had the best performance when predicting for each weather condition compared to the individual models (Mung & Phyu, 2023). This research supports the use of deep learning techniques for temperature prediction.

Research done by Chen and colleagues aimed to evaluate the seasonal autoregressive integrated moving average (SARIMA) model to predict monthly mean temperature in Nanjing China (Chen et al, 2018). The researchers used 35 years of data to train the model and the aim was to predict monthly mean temperature for 3 years ahead (Chen et al, 2018). The model accuracy was tested based on the mean squared error (MSE) (Chen et al, 2018). The first year's MSE was 0.84, the second year was 0.89 and the third year was 0.94 (Chen et al, 2018). Overall, the increase of 0.05 for each year was not significant and the MSE'S are low enough to accept the results (Chen et al, 2018). This study supports the time series SARIMA model being used in temperature forecasting and provides opportunity to explore other time series models for temperature prediction.

Research done by Azari and colleagues aimed to compare six machine learning models for predicting daily average temperature (Azari et al, 2022). The researchers used data from the Memphis International Airport Weather Station covering 1980 to 2021 (Azari et al, 2022). Features within the dataset included relative humidity, 1-hour precipitation, wind speed, barometric sea pressure, and dew point (Azari et al, 2022). The models that were compared included four simple models and two ensemble models (Azari et al, 2022). The simple models were K-nearest neighbors (KNN), linear regression, artificial neural network (ANN), and support vector machines (SVM) (Azari et al, 2022). The ensemble models were random forest and adaptive boosting. The models were trained on data from 1980 to 2014 and tested using data from 2015 to 2021 (Azari et al, 2022). This study evaluated the models with 7 different evaluation metrics (Azari et al, 2022). The evaluation metrics used were relative standard error (RSR), Nash Sutcliffe efficiency (NSE) , mean absolute error (MAE), RMSE, R-squared, index of agreement, and percentage bias (Azari et al, 2022). The results found that all the models performed very well and had accurate results, however ANN performed the best (Azari et al, 2022). This study supports the use of deep learning and machine learning for temperature prediction. The study however, missed the opportunity to investigate time series models for prediction, additional deep learning models for prediction, and only looked at daily average temperature (Azari et al, 2022).

Anjali and researchers aimed to compare the performance machine learning models, SVM, multiple linear regression, and ANN for predicting temperature for data collected in Central Kerala (Anjali et al, 2019). To carry out the work, the researchers chose to use variables humidity, pressure, wind speed, and wind direction to build the models due to their high correlation coefficients with the target variable temperature (Anjali et al, 2019). The data collected for the study included combining hourly, daily, and monthly data from different sources between the dates of 2007 and 2015 from 3 districts in Central Kerala (Anjali et al, 2019). The evaluation metrics used included mean error (ME), MAE, and RMSE (Anjali et al, 2019). The researchers found that multiple regression and SVM that are regression based worked best (Anjali et al, 2019). The researchers proposed that future work in the field be done using hybrid approaches, and deep neural networks for accurate and advanced weather prediction models (Anjali et al, 2019).

Research done by Li and colleagues aimed to evaluate the performance of different statistical and machine learning techniques for long term daily temperature downscaling in Ontario, Canada (Li et al, 2020). The techniques investigated included SVM and LSTM for machine learning and multiple linear regression and arithmetic ensemble mean for the statistical techniques (Li et al, 2020). The evaluation metrics chosen to evaluate performance included R-squared, RMSE, and ratios of RMSE to standard deviation (Li et al, 2020). The data was collected from 12 different weather stations across the province of Ontario covering the years from 1980 to 1989 (Li et al, 2020). The models were trained on data from 1980 to 1986 and tested on data from 1987 to 1989 (Li et al, 2020). The researchers found that overall, all the models performed well, however the machine learning methods performed slightly better than the statistical methods and both methods struggled with predicting extreme values for temperatures below -10 and above 20 degrees Celsius (Li et al, 2020). This study supports the theory that machine learning based weather prediction can achieve better performance than conventional statistical methods.

Research done by Khedhiri aimed to evaluate the performance of seasonal autoregressive fractionally integrated moving average (SARFIMA) and LSTM for monthly minimum and maximum temperature prediction in Canada (Khediri, 2022). The data was collected from weather stations in 4 different areas in Canada covering the western, eastern, and central regions and the data was kept separate as modelling and forecasting were done separately for each area (Khediri, 2022).
The data collected ranged from 1840 to 2016 with 90% used as training and 10% used for testing (Khediri, 2022). The evaluation metric used was RMSE and it was found that LSTM performed better than SARFIMA within each of the regions (Khediri, 2022). The SARFIMA model did also however perform well to be used in further work with additional improvements (Khediri, 2022). This study supports the use of deep learning and time series models for temperature prediction (Khediri, 2022).

Overall, the results from the literature provide good evidence to test the theory of seasonal temperature prediction. The literature provides support for machine learning and deep learning techniques to accurately predict temperature. The literature also provides opportunities to investigate temperature prediction for seasonal durations and hourly predictions.

3 Methodology

The methodology followed for this project was the sample, explore, modify, model, and assess (SEMMA) methodology (Binus University, 2021). This methodology was chosen because it allowed for critical analysis and assessments of the outcomes at each stage independently before moving onto the next stage. This helped a lot in determining if the data was appropriate for the types of analysis that were going to be done during the model stage. Furthermore, if the outcomes were not appropriate it was easy to go back to the previous stage to implement more techniques to better prepare the data for the next analysis. This methodology involves a lot of feedback during each stage which helps to ensure the project is going according to plan. It also helps in knowing which stages to work on again if something is missing rather than starting from the beginning and repeating everything. This methodology was very useful in guiding the project and keeping in mind the research aims, objectives and questions at all times. The step by step framework was very clear and helped ensure no task in the project was skipped. The step by step framework also helped for time management by allocating time periods for each stage. The flexibility in the model stage was also appreciated as there were 3 models being compared. Lastly, this methodology is very transparent and allows reproducibility as each stage is separated and the tasks in each stage are documented (Binus University, 2021). Below is a brief summary of the steps taken during each stage as well as a visual representation of the SEMMA process in figure 1.

Sample
This stage included choosing the dataset, downloading it, ensuring it can be opened properly on the computer software being used, and cleaning the data.

Explore

A significant part of the project took place in the explore stage to uncover trends and patterns. This helped in understanding the data and uncovering different techniques that can work in the modify and model phases. The explore stage also helped greatly in ensuring the data quality was sufficient for the project and that missing data was taken care of early into the process. Techniques done in this phase included time series plots, average distribution plots, seasonality decomposition, correlation analysis, summary statistics, stationarity checks, autocorrelation and partial autocorrelation plots.

Modify

This stage included splitting the data into a training and a test set. This stage also included scaling the data for the LSTM model. This stage was helpful in ensuring the data was ready to be input into the models and all actions that could improve model performance were taken.

Model

This stage included creating the 3 models, ARIMA, SARIMAX and LSTM and identifying the parameters to be used for each.

Assess

Lastly, this stage included assessing each of the models using different evaluation metrics and plots showing the predicted values and the actual values. This phase was helpful in determining the validity and reliability of the models and determining which can be deployed in future works.

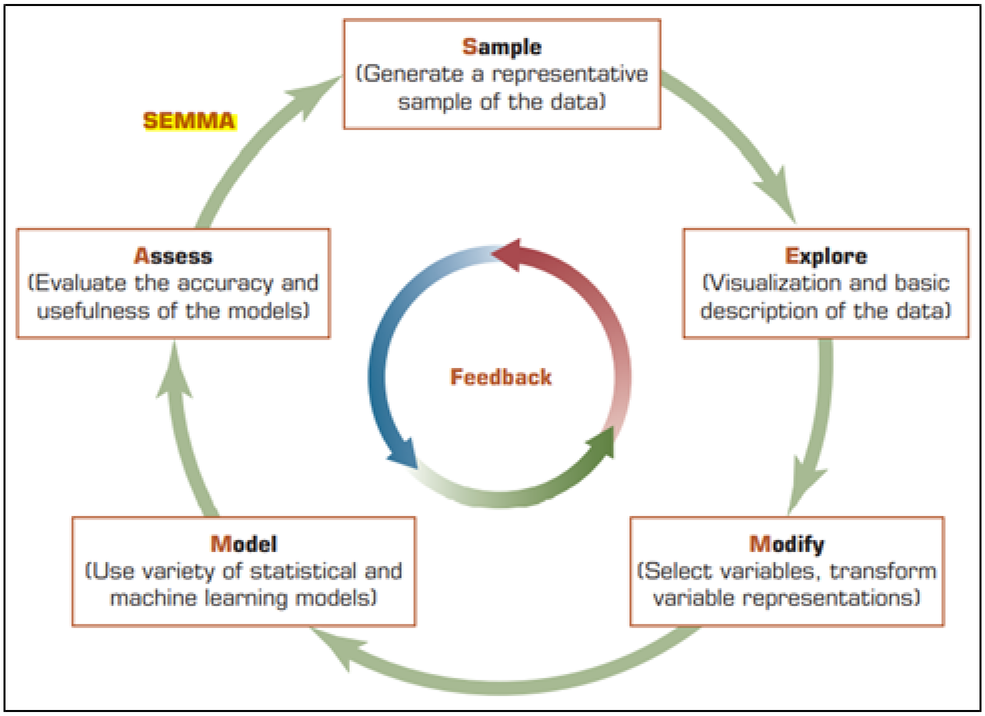

Figure 1-SEMMA Methodology Process (Sharda et al, 2018)

4 Requirements

Requirements in this project included system, functional, and non-functional which are addressed below:

4.1 System Requirements

The system requirements for this project are included in table 1.

Table 1- System Requirements

System type	X64 - Based laptop
OS Name	MacOS Monterey 12.7.4
Processor	1.8GHz dual-core Intel Core i5
Physical Memory	8 GB

4.2 Functional Requirements

The functional requirements within this project are listed below:

I. Joining all the separate files together
II. Setting a date and time index for the data
III. Specifying the exact timeframe for the winter season and filtering the dataset based on the dates
IV. Checking for missing values, filling missing values that can be filled, deleting ones that cannot be filled
V. Identifying if there are observations for each timestamp within the winter season
VI. Performing exploratory data analysis
VII. Performing stationarity testing
VIII. Performing seasonal decomposition
IX. Performing autocorrelation and partial autocorrelation plots
X. Splitting the data into training and test sets
XI. Testing for optimal parameters for ARIMA and SARIMAX
XII. Fitting and testing the ARIMA and SARIMAX models
XIII. Performing normalization on the test data for LSTM
XIV. Creating, fitting, and testing the LSTM model
XV. Visualizing prediction results and calculating evaluation metrics

4.3 Non-functional Requirements

I. The only non-functional requirement needed for this project is that the laptop used and the platform used has the capacity to process the code.

5 Dataset

This research project used a combination of historical weather data recorded by the hour for the 2019 to 2023 winter seasons in Ottawa, Canada. The data has been collected by the Ottawa CDA RCS meteorological station operator by Environment and Climate Change Canada - Meteorological Service of Canada (Government of Canada, 2024a; Government of Canada, 2024b). The station operator is situated at 45°23'00.000" N latitude and 75°43'00.000" W longitude at an elevation of 79.20m (Government of Canada, 2024b). The location can be seen in figure 2 within the city of Ottawa. The historical weather data was spread across 20 different CSV files and was combined into one dataset with each observation identified uniquely by year, month, day, and hour. The dataset once combined contained 14,454 observations which can be found in appendix C and the different variables found in the dataset can be seen in table 2.

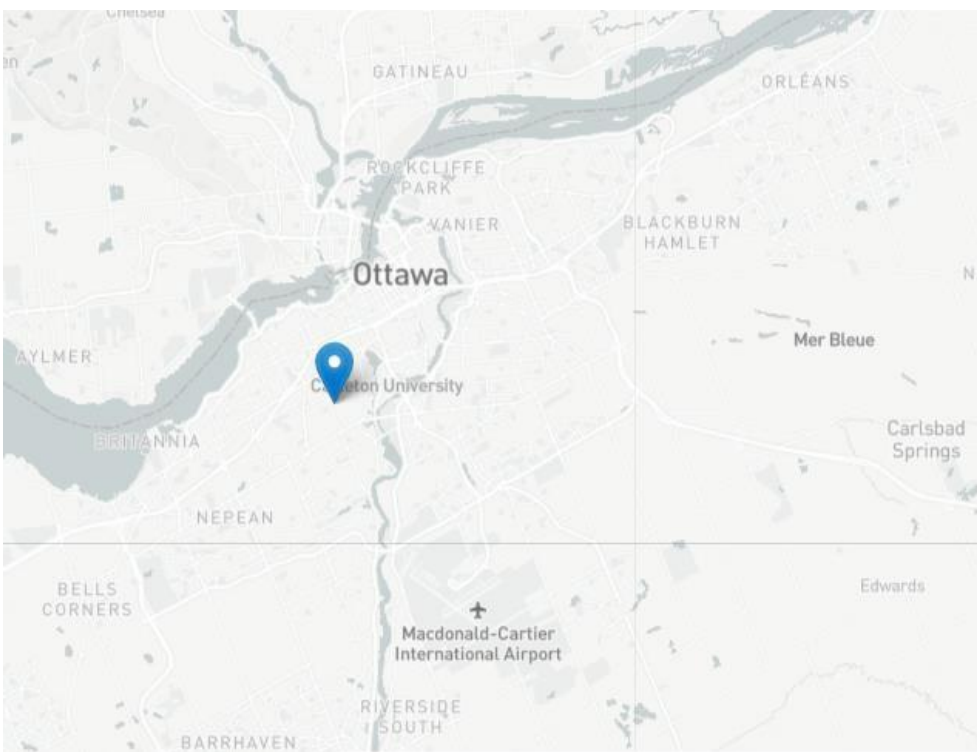

Figure 2-Location of the Ottawa CDA RCS Meteorological Station Operator (Ottawa Weatherstats, n.d)

Table 2-. Dataset Variable Descriptions (Government of Canada, 2024a)

Field	Description	Additional Information
Longitude (x)	Longitude coordinates in degrees	
Latitude (y)	Latitude coordinates in degrees	
Station Name	Name of the meteorological station	
Climate ID	Identifier for the site where the observations are taken	
Date/Time (UTC)	Date and time the observation was taken	
Year	Year the observation was taken	
Month	Month the observation was taken	
Day	Day the observation was taken	
Time (UTC)	Time the observation was taken in coordinated universal time	
Temperature (°C)	Air temperature in degrees Celsius (°C)	Blank = unobserved
Temperature Flag	Identifies data quality or if any specific conditions affected the data	Blank = valid E = estimated M = missing NA = Not available
Dew Point Temperature (°C)	A measure of the air humidity in degrees Celsius (°C)	Blank = unobserved
Dew Point Temperature Flag	Identifies data quality or if any specific conditions affected the data	Blank = valid E = estimated M = missing NA = Not available

Relative Humidity (%)	The ratio of the amount of water vapor the air contains compared to the maximum amount it can hold at the observed temperature	Blank = unobserved
Relative Humidity Flag	Identifies data quality or if any specific conditions affected the data	Blank = valid E = estimated M = missing NA = Not available
Precipitation Amount (mm)	The vertical depth in millimeter (mm) of water that reaches the ground during the stated period	Blank = unobserved
Precipitation Amount Flag	Identifies data quality or if any specific conditions affected the data	E = estimated Blank = valid NA = not available M = missing
Wind Direction (10s deg)	The direction (geographic or true) the wind blows from. It is the average direction during the two minute period ending at the observation time calculated in tens of degrees (10's deg)	Blank = unobserved
Wind Direction Flag	Identifies data quality or if any specific conditions affected the data	Blank = valid E = estimated M = missing NA = Not available
Wind Speed (km/h)	The speed of air movement in kilometers per hour (km/h). It is the average pace during the one, two, or ten minute period ending at the observation time	Blank = unobserved
Wind Speed Flag	Identifies data quality or if any specific conditions affected the data	Blank = valid E = estimated M = missing NA = Not available
Visibility (km)	The distance in kilometers (km) where suitable sized objects can be seen and identified	Blank = unobserved
Visibility Flag	Identifies data quality or if any specific conditions affected the data	Blank = valid E = estimated M = missing NA = Not available

Station Pressure (kPa)	The force per unit area applied by the atmosphere due to the mass of air in a vertical column from the elevation of the top of the weather station operator to the top of the atmosphere. This is measured in kilopascal (kPa)	Blank = unobserved
Station Pressure Flag	Identifies data quality or if any specific conditions affected the data	Blank = valid E = estimated M = missing NA = Not available
Humidex	An indication of how humid or hot the weather feels to the average person. Hourly values are shown when the temperature is 20C or greater and the humidex value is at least 1 degree greater than the temperature. The humidex formula is: Humidex = (air temperature) + h Where: h = (0.5555)*(e - 10.0); e = vapor pressure in hPa (mbar), given by:	Blank = unobserved
	e = 6.11 * exp[5417.7530 * ((1/273.15) - (1/dewpoint))] exp = 2.71828	
Humidex Flag	Identifies data quality or if any specific conditions affected the data	Blank = valid E = estimated M = missing NA = Not available

Wind Chill	An indication of how cold the weather feels to the average person. Two formulas are used. The first is used when the temperature is ≤ 0 °C and the wind speed is ≥ 5 km/h. The second is used when the temperature is ≤ 0 °C and the wind speed is > 0 km/h but < 5 km/h. The Wind Chill formulas used by Environment and Climate Change Canada are: 1. W = 13.12 + 0.6215 × T - 11.37 × V0.16 + 0.3965 x T × V0.16 2. W = T + [(-1.59+0.1345 × T)/5] × V Where: W is the wind chill index, based on the Celsius temperature scale T is the temperature in degrees Celsius (°C) V is the wind speed at 10 meters elevation, in kilometers per hour (km/h).	Blank = unobserved
Wind Chill Flag	Identifies data quality or if any specific conditions affected the data	Blank = valid E = estimated M = missing NA = Not available
Weather	The state of the atmosphere at the observation time	E = estimated Blank = unobserved NA = not available M = missing

6 Design

The design of this project is divided into 5 sections based on the sections followed in the SEMMA methodology. Each section's explanations and reasoning are explained below.

Section 1: Data Loading and Cleaning

This is one of the most important sections in the project design as it is responsible for ensuring the data is cleaned and processed accurately for analysis. This section not only included implementing different techniques on the data to prepare it, but this section also included an intensive literature search to apply accurate calculations for missing values. In this section, data was collected from the Government of Canada for hourly temperatures covering 2019 to 2023 months January to March, and December for each year. However, due to the focus of this project being on seasonal temperature prediction of the winter season, a subset of the data needed to be created to accurately represent that population. This required an intensive search to be done on the Government of Canada website to determine the exact dates of the winter season for accurate filtering. This information was found within the National Research Council of Canada sector and was applied to the data to create the sample and start data cleaning. The exact winter season dates are shown in table 3 for reference. The next priority was checking for missing values in the dataset. Once determined that there are missing values, an extensive search was done in the climate data glossary on the Government of Canada site to determine which values can be filled based on mathematical equations and which would need to be done using the Pandas interpolate method. Functions were created in Python to fill the missing values in the columns that depended on mathematical equations and columns with less than 10% of missing data were filled according to Pandas interpolate method. Columns having more than 30% of data missing were completely removed and columns having between 10% and 20% of missing data had the rows removed that contained the missing data. Lastly, columns that included constant information and did not have an effect on the analysis were removed.

Table 3-Winter Start and End Times in Canada (UTC) (NRC, 2024)

Year	Winter Start	Winter End
2019	December 22 04:14	March 20 21:58
2020	December 21 10:02	March 20 03:49
2021	December 21 15:59	March 20 09:37
2022	December 21 21:48	March 20 15:33
2023	December 22 03:27	March 20 21:24

Section 2: Exploratory Data analysis

Exploratory data analysis was done to view the relationships among the variables including the target variable. This analysis also provided insights on the descriptive statistics of the data and data distribution. This area also identified the trend and seasonality components of the time series data.

Section 3: Model Creation

This section focused on implementing different transformations on the data, using different methods to choose parameters for the data and splitting the data. The same transformations and methods to choose parameters were used for the ARIMA, and SARIMAX models. In order to create the LSTM, model a sequence length needed to be chosen. Although some observations were removed when pre-processing so not all daily data consists of all the 24 observations, 99.98% of the training data remained so 24 was still chosen for the sequence length. Seasonality of 24 was also used for SARIMAX. The LSTM model design included 2 LSTM layers and one dense layer. The first LSTM layer consisted of 50 units and returned sequences, the second included 50 units and returned one output, and the dense layer to produce one final output. The hyperparameters were experimented for and the final values chosen were 10 for epochs and 32 for the batch size.

Section 4: Model Training

This section focused on training and testing the model’s with the chosen parameters.

Section 5: Model Evaluation

In this section, each model was evaluated based on performance using MSE and MAE evaluation metrics.

7. Techniques

7.1 ARIMA

The ARIMA model is a machine learning technique that is used for time series analysis and forecasting (Kim, 2024). The univariate model is characterized by three parameters, p, d, and q (Kim, 2024). The p parameter represents the autoregressive portion of the model and captures the relationship between the current observation and the previous observations (Kim, 2024). The d parameter represents the integrated portion of the model and assists in making the data stationary if not already stationary (Kim, 2024). This parameter tells us the degree of differencing needed to make the data stationary (Kim, 2024). Lastly, the q parameter represents the moving average (Kim, 2024). This parameter tells us the size of the moving average window and helps in smoothing noise in the data to allow for better analysis (Kim, 2024). The model's structure is shown in figure 3.

$$\text{ARIMA}\ (p, d, q)$$

Figure 3-ARIMA Structure (Penn State Eberly College of Science, 2024)

7.2 SARIMAX

The SARIMAX model is an extension of the ARIMA model that accounts for seasonality and exogenous factors. The exogenous factors included are any factors that can influence the time series. The model’s structure is seen in figure 4. The model's structure is the same as the SARIMA model, however within the mathematical equation it includes exogenous factors making it different from the SARIMA model. This multivariate model is characterized by parameters p,d,q,P,D,Q and S (Kim, 2024). The p parameter represents the autoregressive portion of the model and captures the relationship between the current observation and the previous observations (Kim, 2024). The d parameter represents the integrated portion of the model and assists in making the data stationary if not already stationary (Kim, 2024). This parameter tells us the degree of differencing needed to make the data stationary (Kim, 2024). Lastly, the q parameter represents the moving average (Kim, 2024). This parameter tells us the size of the moving average window and helps in smoothing noise in the data to allow for better analysis (Kim, 2024).The parameters P,D,Q and S, however, are all seasonal parameters (Kim, 2024). P represents the seasonal lag order, D represents the seasonal degree of differencing, and Q represents the seasonal size of the moving average window (Kim, 2024). Lastly, S represents the time span of the repeating seasonal period (Kim, 2024).

$$\text{ARIMA } (p, d, q) \times (P, D, Q)S$$

Figure 4-SARIMAX Structure (Penn State Eberly College of Science, 2024)

7.3 LSTM

LSTM is a type of recurrent neural network used in the subset of machine learning, deep learning (Hamad, 2023). This model works by using feedback connections to determine time related dependencies within sequences of data (Hamad, 2023). This model uses memory cells to keep, store, update, and retrieve information over long sequences (Hamad, 2023). Memory cells have three main components which include an input gate, a forget gate, and an output gate (Hamad, 2023). These gates work by assisting in flow regulation of information going in or out of the cells (Hamad, 2023). The input gate controls how much current data is stored within the cell

(Hamad, 2023). The current provided data along with earlier hidden state data is taken in by the cell and a value between 0 and 1 are output for each element (Hamad, 2023). The hidden state is the model's output at a specific time based on the cell's long term memory (Hamad, 2023). The forget gate determines which information to keep or dispose of within the cell (Hamad, 2023).
Again, current provided data along with earlier hidden state data is taken in by the cell and a value between 0 and 1 are output for each element (Hamad, 2023). For this gate however, a value of 0 means the information is disposed of and a value of 1 means it is kept (Hamad, 2023). The last gate is the output gate that determines how much of the content within the memory cell should be used to determine the hidden state data (Hamad, 2023). Again, current provided data along with earlier hidden state data is taken in by the cell and a value between 0 and 1 are output for each element to be used to make predictions (Hamad, 2023).

7.4 Auto_arima Model Selection

The auto_arima function is a technique imported from the pmdarima library in Python to provide the best parameters for the ARIMA model (SKtime, 2024). This function also supports SARIMA and SARIMAX parameter fitting (SKtime, 2024). This function works by first conducting differencing tests to determine the order of differencing, then tries different combinations for the autoregressive and moving average components of the model (SKtime, 2024). If the model is enabled to find parameters for SARIMA or SARIMAX, seasonal differencing is done next followed by experimenting to find the optimal seasonal autoregressive and moving average components (SKtime, 2024). The model uses a user specified information criterion to guide itself when choosing the best model (SKtime, 2024).

7.5 Augmented Dickey-Fuller Test

The augmented dickey-fuller (ADF) test is used to check the stationarity of the data (Musbah & El-Hawary, 2019). Stationarity in data refers to the mean, covariance, and autocorrelation being constant over time (Musbah & El-Hawary, 2019). In order for the data to be modelled and forecasted accurately it needs to be stationary (Santra, 2023). If the data is not stationary there will be too much room for unpredictability and the model may not work properly, reducing performance (Santra, 2023). The results of the test identify whether the null hypothesis is accepted or rejected (Santra, 2023). The null hypothesis is that data is not stationary while the alternative hypothesis that the data is stationary (Santra, 2023). Based on the test statistic value, if less than the critical value, or if the p-value associated with the test is less than the specified significance level, the time series will be considered stationary (Santra, 2023). The test statistic formula is shown in figure 5 with SE being standard error and $\hat{\beta}_i$ being the estimated coefficient (Santra, 2023).

$$t_{\hat{\beta}_i} = \frac{\hat{\beta}_\iota}{SE(\hat{\beta}_\iota)}$$

Figure 5-ADF Test Statistic Formula (Santra, 2023)

7.6 Feature Scaling

This preprocessing technique works by transforming feature values to a similar scale (Bhandari, 2024). This assists the user as it ensures all the features contribute equally within the chosen model (Bhandari, 2024). Furthermore, this method aids in promoting improved model performance, convergence, and preventing bias arising from the different unscaled values (Bhandari, 2024). The main feature scaling methods include normalization and standardization (Bhandari, 2024). Each is explained below with normalization being the chosen technique used for this project.

Normalization

This technique is also known as the Min-Max scaling method (Bhandari, 2024). This method scales the values and places them between the range of 0 and 1 (Bhandari, 2024). The formula for this technique is shown in figure 6 where Xmax refers to the maximum value and Xmin refers to the minimum value within the chosen feature data column (Bhandari, 2024).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Figure 6-Normalization Mathematical Formula (Bhandari, 2024)

Standardization

This technique scales the values and centers them around the mean and gives a unit standard deviation (Bhandari, 2024). The formula for this technique is shown in figure 7 where μ represents the mean and σ represents the standard deviation for the values in the chosen feature data column (Bhandari, 2024).

$$X' = \frac{X - \mu}{\sigma}$$

Figure 7-Standardization Mathematical Formula (Bhandari, 2024)

7.7 Mean Squared Error

Mean squared error is an evaluation metric that calculates the average squared difference between the actual and predicted values (Rajawat et al, 2022). The equation is shown in figure 8 with n being the total number of data observations, Yi being the actual value, and $\hat{Y}i$ being the predicted value (Rajawat et al, 2022).

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{Y_i} \right)^2$$

Figure 8- MSE Mathematical Formula (Rajawat et al, 2022)

7.8 Mean Absolute Error

Mean absolute error is an evaluation metric that calculates the average absolute difference between the actual and predicted values (Rajawat et al, 2022). The equation is shown in figure 9 with n being the total number of data observations, Yi being the actual value and $\hat{Yi}$ being the predicted value (Rajawat et al, 2022).

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \widehat{Y_i} \right|$$

Figure 9-MAE Mathematical Formula (Rajawat et al, 2022)

7.9 Autocorrelation function

The autocorrelation function (ACF) plot shows the correlation time series data has with itself at different lags (Ahmed, 2023). If the correlation coefficient is 1, there is a strong positive relationship, while a correlation coefficient of -1 means a strong negative relationship (Ahmed, 2023). This plot can also be used to identify the autoregressive model order (Ahmed, 2023). The lags that are included in the autoregressive model will have spikes on the ACF plot and the number of lags that are included in the model represent the order of the model (Ahmed, 2023).

7.10 Partial autocorrelation function

The partial autocorrelation function (PACF) plot shows the correlation time series data has with itself at different lags after the effects of the previous lags have been removed (Ahmed, 2023). This plot can be used to identify the moving average model order (Ahmed, 2023). The lags that are included in the moving average model will have spikes on the PACF plot and the number of lags that are included in the model represent the order of the model (Ahmed, 2023).

7.11 Seasonal decomposition

This is a technique that splits the time series data into trend, seasonality, and residual components (Malkari, 2023). This technique is used in time series analysis to help analyse the trends and seasonality within the data to provide guidance when preparing parameters for time series predictive models (Malkari, 2023).

8. Implementation

8.1 Platform, languages, and tools

The platform used for this project was Spyder 5.4.3 accessed through Anaconda Navigator 2.5.2. The programming language used was Python. Tools that were used for data preprocessing included NumPy and Pandas. Tools that were used for data visualization included Matplotlib and Seaborn. Tools that were used for the ARIMA and SARIMAX preparation, assessment, and modelling included Statsmodels, Seaborn, Pmdarima, Matplotlib, Pandas and Scikit-learn. Tools that were used for LSTM preparation, assessment, and modelling included Scikit-learn, TensorFlow, Matplotlib, NumPy, and Pandas. This combination of tools allowed for smooth progress during the project and efficiency in the model development and evaluation stages throughout the project. The code can be found in appendix D.

8.2 Prototype Development

Initially, the testing started on ARIMA, SARIMA, and LSTM. However, based on the SARIMA results, insights from the autocorrelation and partial autocorrelation plots, and correlation matrix during the exploratory data analysis, it was identified that the data has a weak seasonality and some exogenous variables in the dataset have high correlation with the target variable. This provided an opening to experiment with the SARIMAX model which is an extension of the SARIMA model that includes exogenous variables when producing predictions. SARIMAX improved the results which made sense as it most likely focused greatly on the relationships between the exogenous variables and the target variable. Another initial plan was to predict temperatures for two winter seasons ahead. However, when splitting the data, it resulted in a 60% training to 40% test split which could be improved to provide higher accuracy and efficiency of the model which is a goal in the project. The split was then adjusted to predict for one winter season ahead which split the data into an 82% training to an 18% testing set. This adjustment improved the results for all the models further placing an importance on data distribution and splitting optimization in model training.

8.3 Current Status and Deployment

The models currently work to predict hourly temperatures for the winter season ahead. The models were trained on all hourly winter data except for 28 timestamps as they did not meet the requirements to be included in the dataset. The specific dates and hours that were not included in training are shown in figure 10. As the winter season cut off dates for most of the years from 2019 to 2022 were December 21st and March 20th at different hours, it makes sense that most of the data on these dates were not included in training as seen in figure 10. The models were still able to train on the remaining 99.98% of the winter season data, thus supporting their effective deployment.

```
missing data for: Index(['02-11-04', '02-11-05', '02-24-16',
'02-24-17', '03-09-19', '03-10-19',
       '03-10-20', '03-10-21', '03-10-22', '03-10-23',
'03-11-17', '03-19-21',
       '03-19-22', '03-20-13', '03-20-15', '03-20-16',
'03-20-17', '03-20-18',
       '03-20-19', '03-20-20', '03-20-21', '12-21-11',
'12-21-12', '12-21-13',
       '12-21-14', '12-21-15', '12-21-18', '12-21-19'],
      dtype='object')
```

Figure 10 - Dates not Included in the Training Set in ‘Month,Day,Hour’ Format

9. Results

The analysis started with an exploratory data analysis to identify the variables highly correlated with temperature shown in figure 11. Following this, visualizations for each of the highly correlated variables and temperature were done to further identify patterns in the data which can be shown in figures 12 to 23. Next, stationarity testing was done to ensure the data can be modelled accurately without unexpected seasonality interference for ARIMA and SARIMAX. Next, seasonal decomposition was done to view the components of the time series data. Following, was the autocorrelation and partial autocorrelation plots to aid with parameter identification for ARIMA and SARIMAX. Once this was all completed, model selection was done using the auto_arima function being based on the alkaline information criterion (AIC) showing the selected models shown in table 4. Next, model fitting and testing was done for ARIMA, and SARIMAX with results shown in figure 27. LSTM was last to be fit and modelled, with results shown in figure 28. Table 5 provides a summary of the evaluation metrics for each model.

9.1 Exploratory Data Analysis

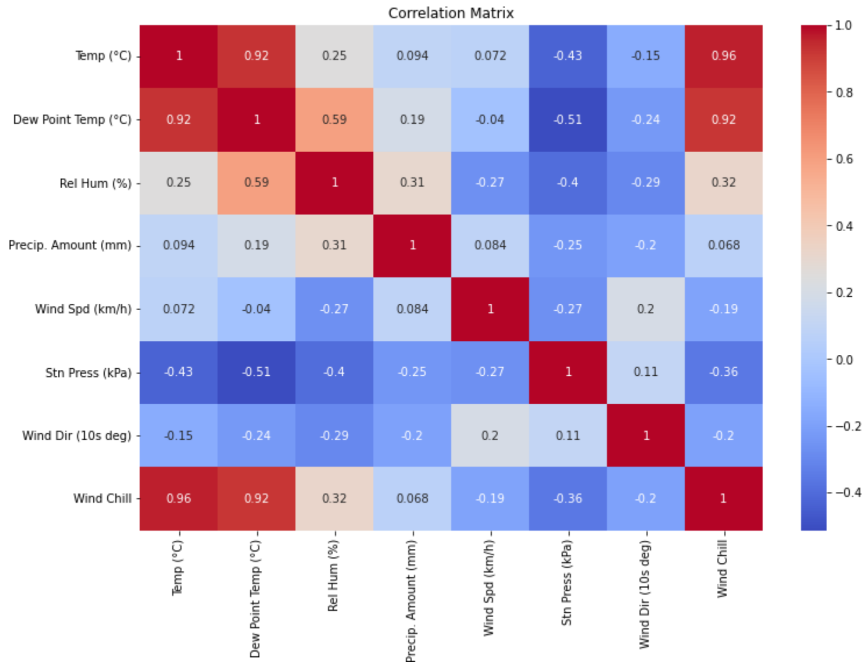

Figure 11 - Correlation Matrix Analysis

Figure 11 shows that temperature has high positive correlations with dew point temperature and wind chill. Dew point temperature has a high positive correlation with wind chill, a moderate positive correlation with relative humidity, and a moderate negative correlation with station pressure. All other correlations are weak. The findings identify dew point temperature, and wind chill for having high correlations with temperature to be selected for further analysis.

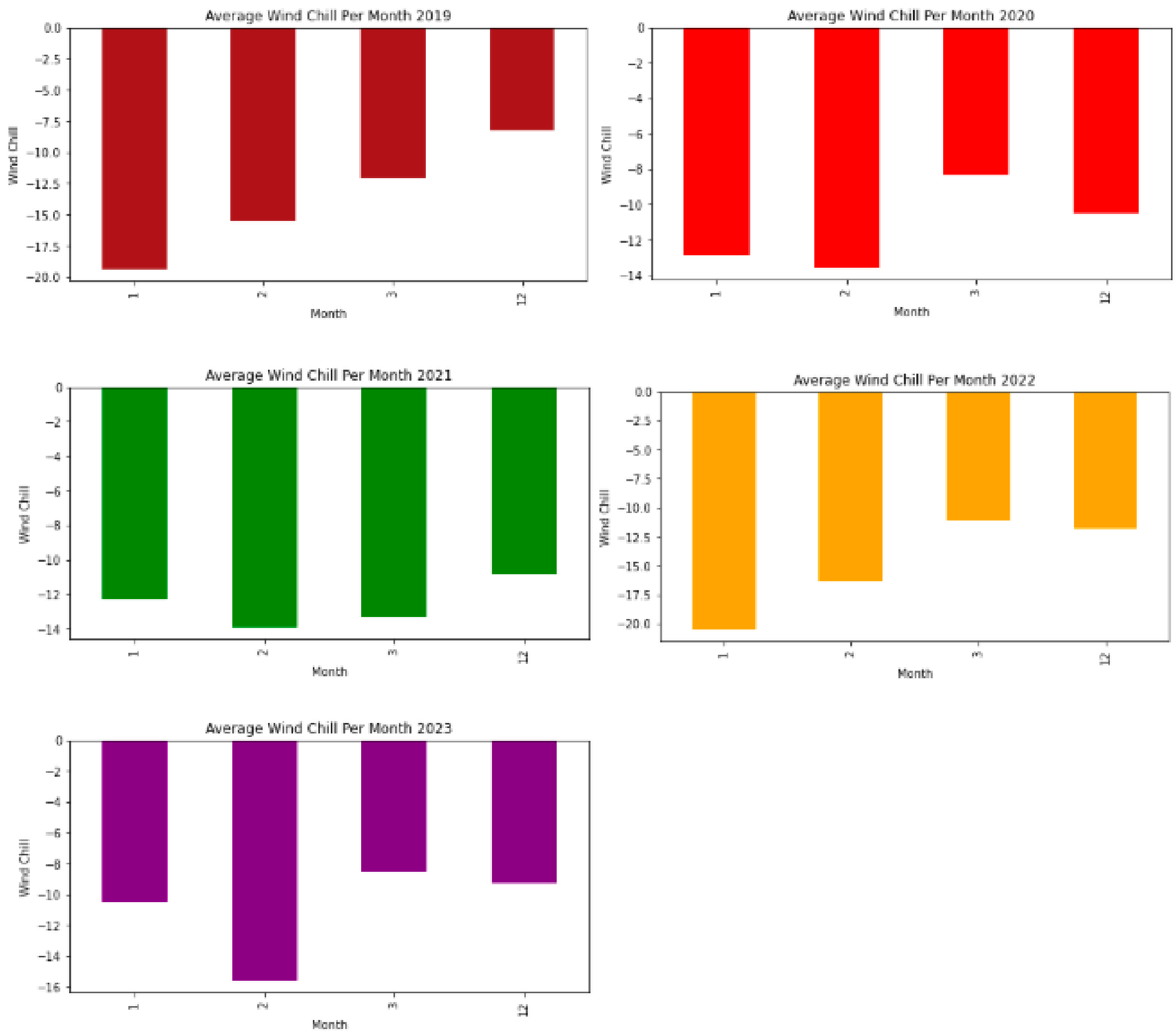

Figure 12 - Average Wind Chill Per Month 2019 to 2023

In figure 12, for years 2019 and 2022, the month that experienced the coldest average wind chill was January. For years 2020, 2021, and 2023, the month that experienced the average coldest wind chill was February. For years 2019 and 2021, the month that experienced the warmest average wind chill was December. For years 2020, 2022, and 2023, the month that experienced the warmest average wind chill was March. Overall, the month that had the average coldest wind chill was February 2022, with wind chill reaching -20. Conversely, the month with the warmest average wind chill was December 2019 reaching -8.2 wind chill. No overall pattern can be seen over the years.

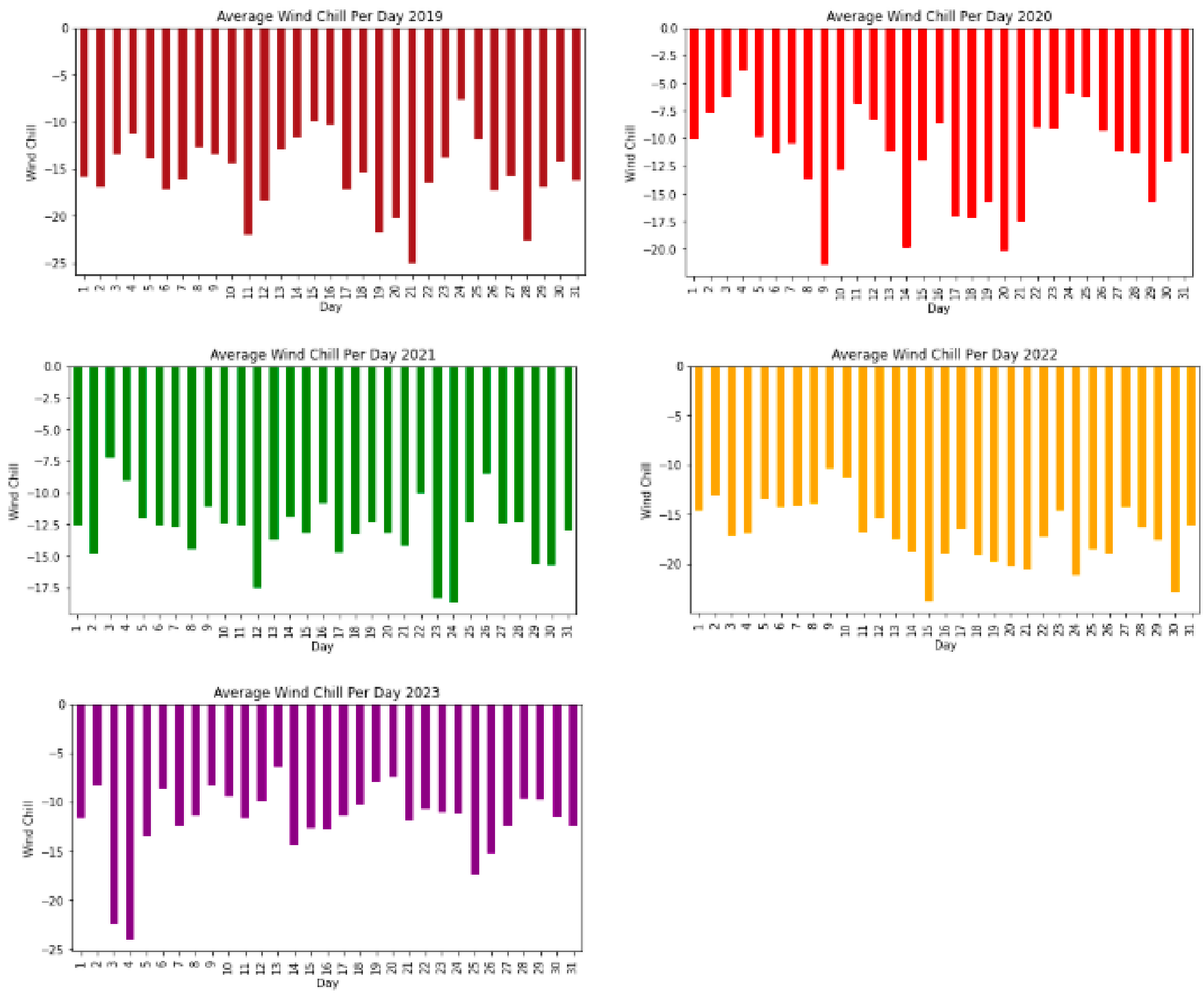

Figure 13 - Average Wind Chill Per Day 2019 to 2023

It can be seen in figure 13 for 2019 that the coldest average wind chill per day was on the 21st day of each of the months in the winter season. It can also be seen that the warmest average wind chill per day in 2019 was on the 24th day of each of the months in the winter season. For 2020 the coldest average wind chill per day was on the 9th day of each of the months in the winter season. The warmest average wind chill per day in 2020 was on the 4th day of each of the months in the winter season. For 2021 the coldest average wind chill per day was on the 24th day of each of the months in the winter season. The warmest average wind chill per day in 2021 was on the 3rd day of each of the months in the winter season. For 2022 the coldest average wind chill per day was on the 15th day of each of the months in the winter season. The warmest average wind chill per day in 2022 was on the 9th day of each of the months in the winter season. Lastly, for 2023 the coldest average wind chill per day was on the 4th day of each of the months in the winter season. The warmest average wind chill per day in 2023 was on the 13th day of each of the months in the winter season. No overall pattern can be seen over the years.

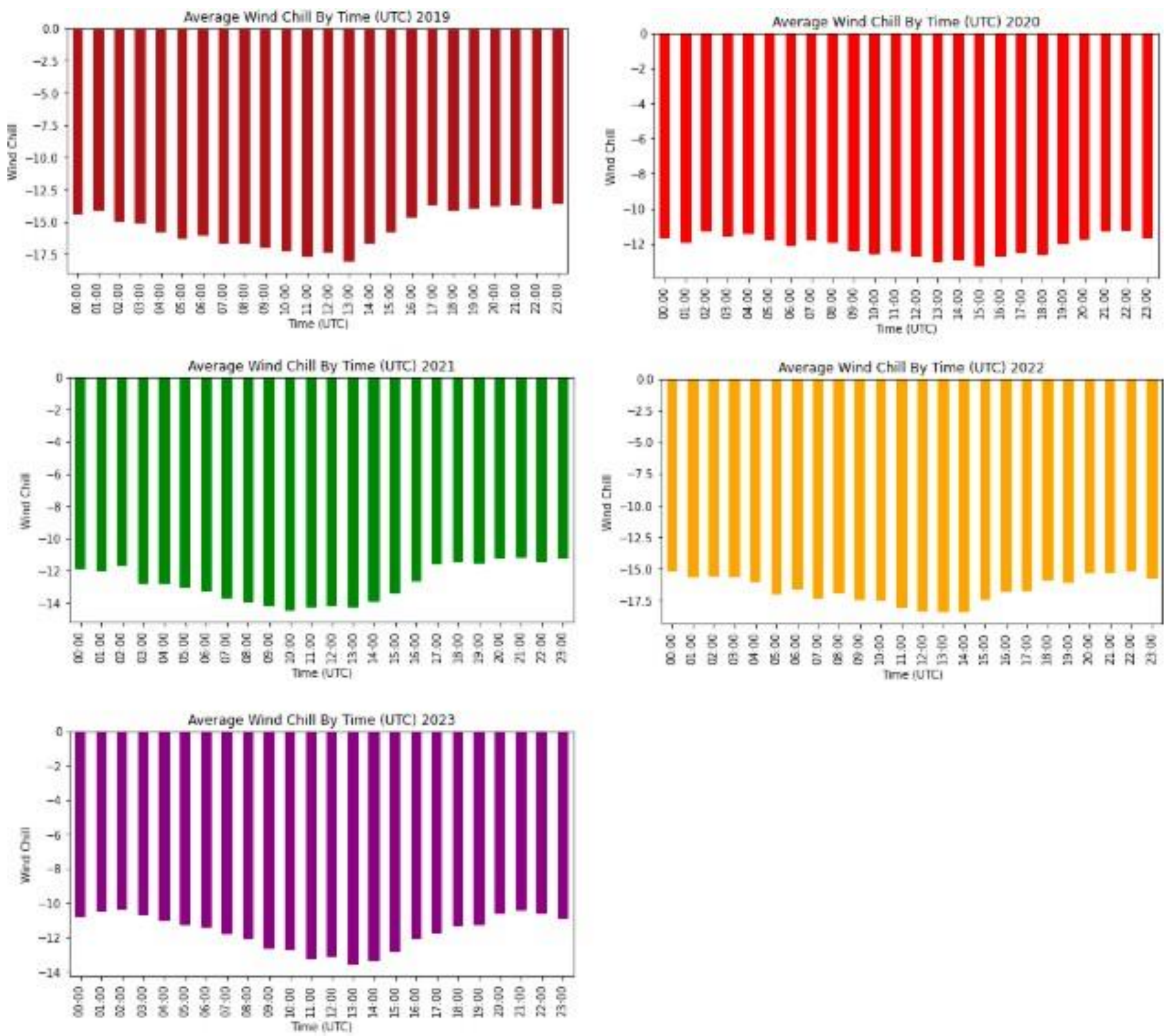

Figure 14 - Average Wind Chill By Time 2019 to 2023

In figure 14 it can be seen that in 2019, 2022, and 2023, the average coldest wind chill by time was at 13:00 for each day of the winter season. For 2020 the average coldest wind chill by time was 15:00 for each day of the winter season. For 2021 average coldest wind chill by time was at 10:00 for each day of the winter season. The warmest average warmest wind chill by time in 2019 was at 23:00 for each day of the winter season. For 2020, and 2022, the average warmest wind chill by time was at 22:00 for each day of the winter season. For 2021, the average warmest wind chill by time was at

21:00 for each day of the winter season. Lastly, 2023 average warmest wind chill by time was at 02:00 for each day of the winter season. An overall pattern can be seen in the hourly wind chill graphs where wind chill is high around the early mornings and decreases throughout the day to have the lowest wind chill during the noon hours then increases steadily to another high during the evenings to late night to repeat the cycle.

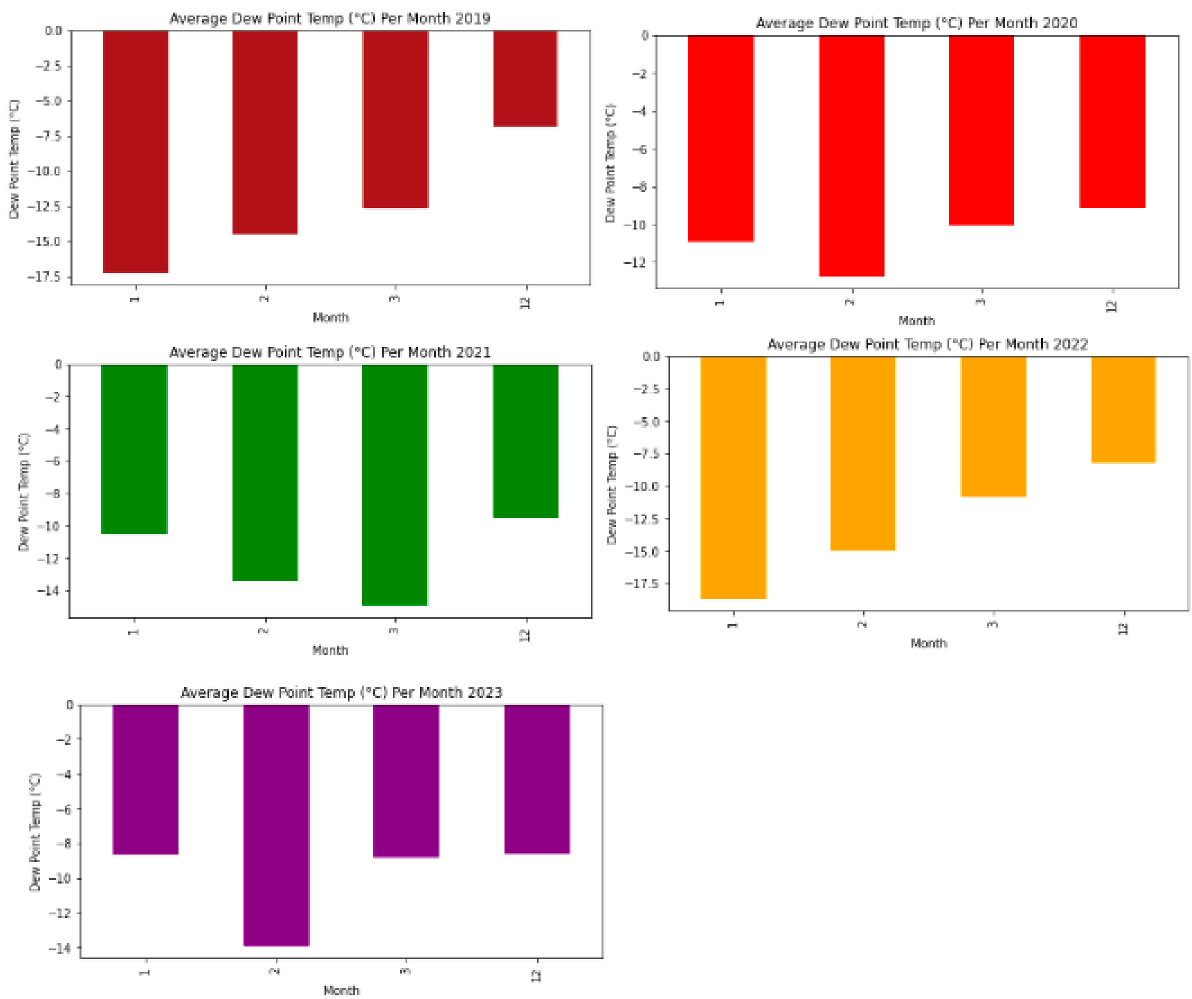

Figure 15 - Average Dew Point Temperature Per Month 2019 to 2023

In figure 15 it can be seen that in 2019, and 2022 the month that experienced the average coldest dew point temperature was January. For 2020 and 2023 the month that experienced the average coldest dew point temperature was February. In 2021, the month that experienced the average coldest dew point temperature was March. It can be seen that January 2022 had the coldest average dew point temperature reaching -18.67 ℃. The month with the warmest average dew point temperature was

December 2019 reaching -6.83 °C. No overall pattern can be seen over the years, except that December in every year has the warmest average dew point which makes sense as it is the first month of the winter season.

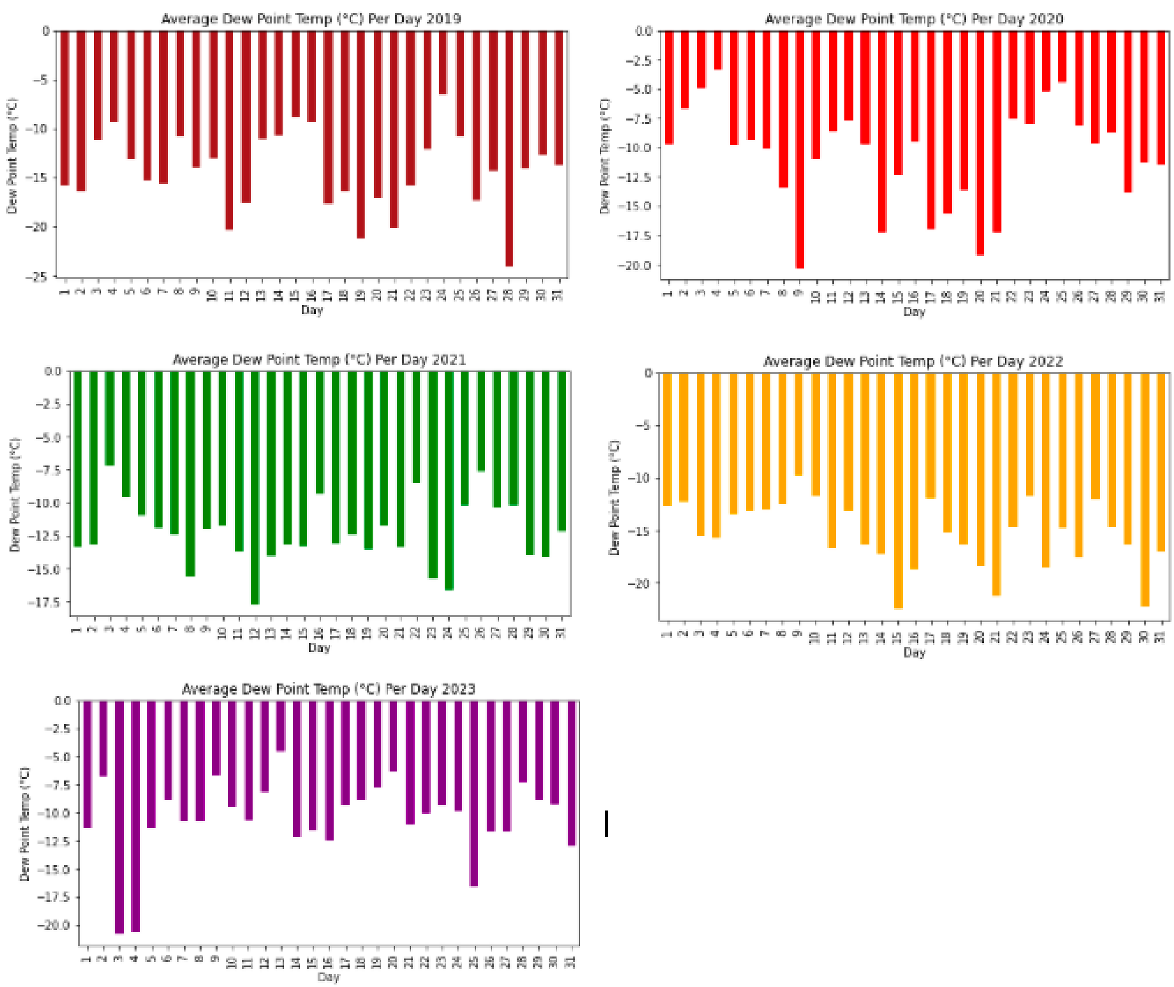

Figure 16 - Average Dew Point Temperature Per Day 2019 to 2023

In figure 16, the coldest average dew point temperature per day in 2019 was on the 28th day of each of the months in the winter season. The warmest average dew point temperature per day in 2019 was on the 24th day of each of the months in the winter season. The coldest average dew point temperature per day in 2020 was on the 9th day of each of the months in the winter season. The warmest average dew point temperature per day in 2020 was on the 4th day of each of the months in the winter season. The coldest average dew point temperature per day in 2021 was on the 12th day of each of the months in the winter season. The warmest average dew point temperature per day in 2021 was on the 3rd day of each of the months in the winter season.

The coldest average dew point temperature per day in 2022 was on the 15th day of each of the months in the winter season. The warmest average dew point temperature per day in 2022 was on the 9th day of each of the months in the winter season. The coldest average dew point temperature per day in 2023 was on the 3rd day of each of the months in the winter season. The warmest average dew point temperature per day in 2023 was on the 13th day of each of the months in the winter season. Overall, no patterns can be found in the average dew point temperature per der day over the years.

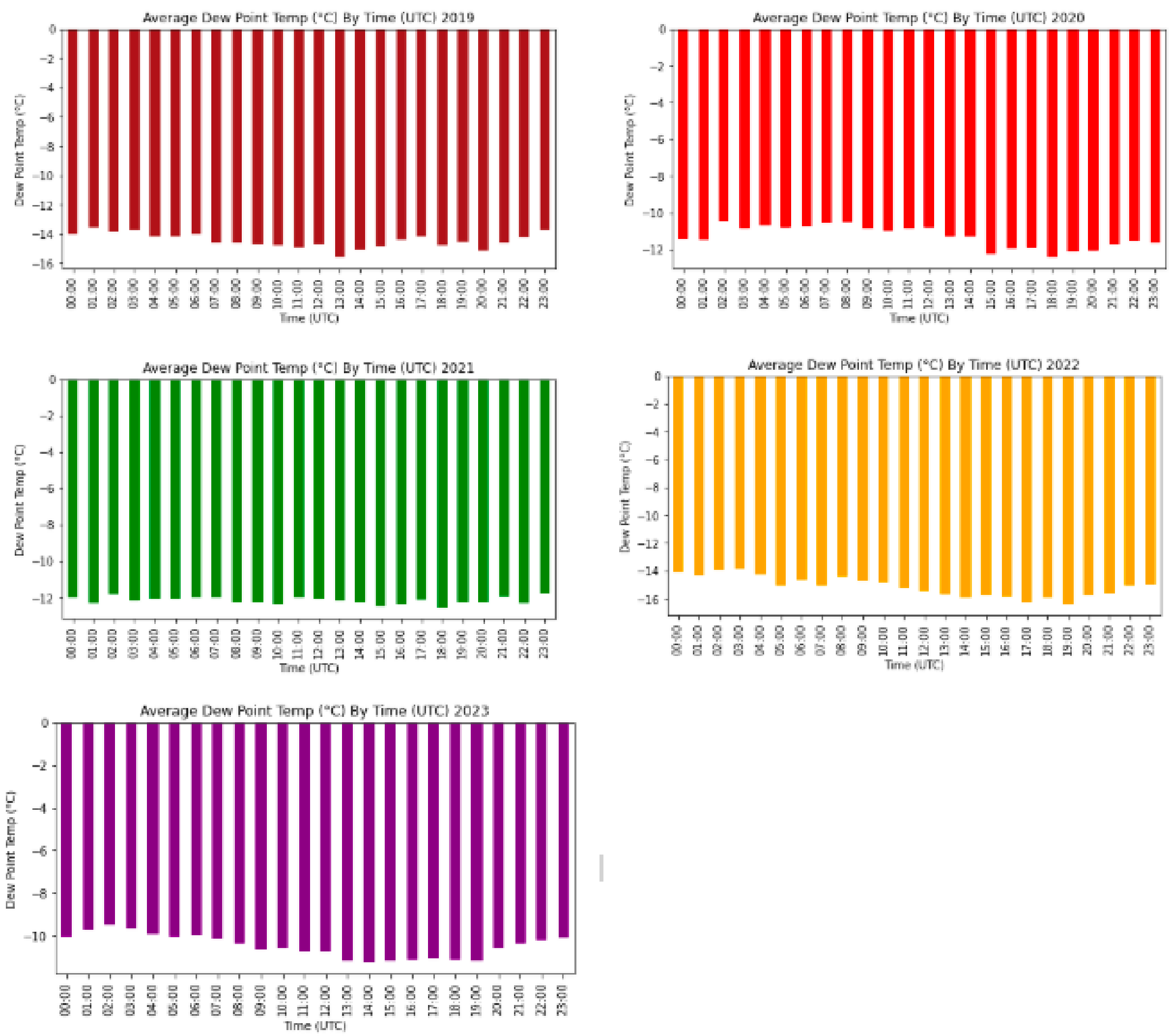

Figure 17 - Average Dew Point Temperature by Time 2019 to 2023

In figure 17, the average coldest dew point temperature by time in 2019 was at 13:00 for each day of the winter season. The average warmest dew point temperature by time in 2019 was at 23:00 for each day of the winter season. The average coldest dew point temperature by time in 2020 was at 18:00 for each day of the winter season. The average warmest dew point temperature by time in 2020 was at 02:00 for each day of the winter season. The average coldest dew point temperature by time in 2021 was at 18:00 for each day of the winter season. The average warmest dew point temperature by time in 2021 was at 23:00 for each day of the winter season. The average coldest dew point temperature by time in 2022 was at 19:00 for each day of the winter season. The average warmest dew point temperature by time in 2022 was at 03:00 for each day of the winter season. The average coldest dew point temperature by time in 2023 was at 14:00 for each day of the winter season. The average warmest dew point temperature by time in 2023 was at 02:00 for each day of the winter season. A small trend can be seen that the average dew point temperatures in 2020, 2022, and 2023 are coldest during the late afternoons to evenings.

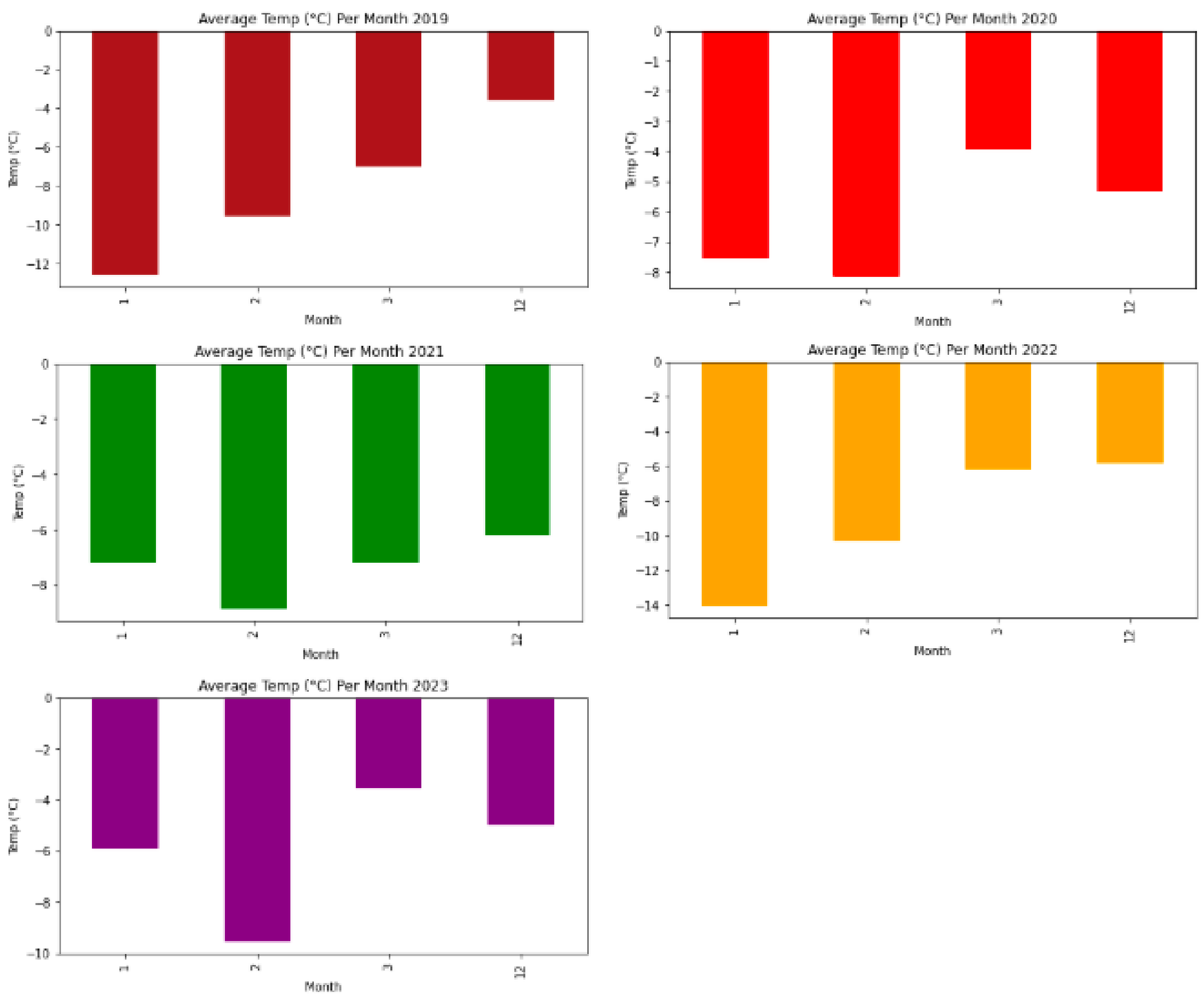

Figure 18 - Average Temperature Per Month 2019 to 2023

In figure 18, for 2019, and 2022 the month that experienced the coldest average temperature was

January. For 2020, 2021, and 2023, the month that experienced the coldest average temperature was February. For 2019, 2021, and 2022, the month with the average warmest temperature was December. For 2020 and 2023, the month with the average warmest temperature was March. It can be seen that January 2022 had the coldest average temperature reaching -14.04 ℃. The month with the warmest average temperature was March 2023 reaching -3.57 ℃. Overall, no clear pattern can be seen showing the warming effects of climate change.

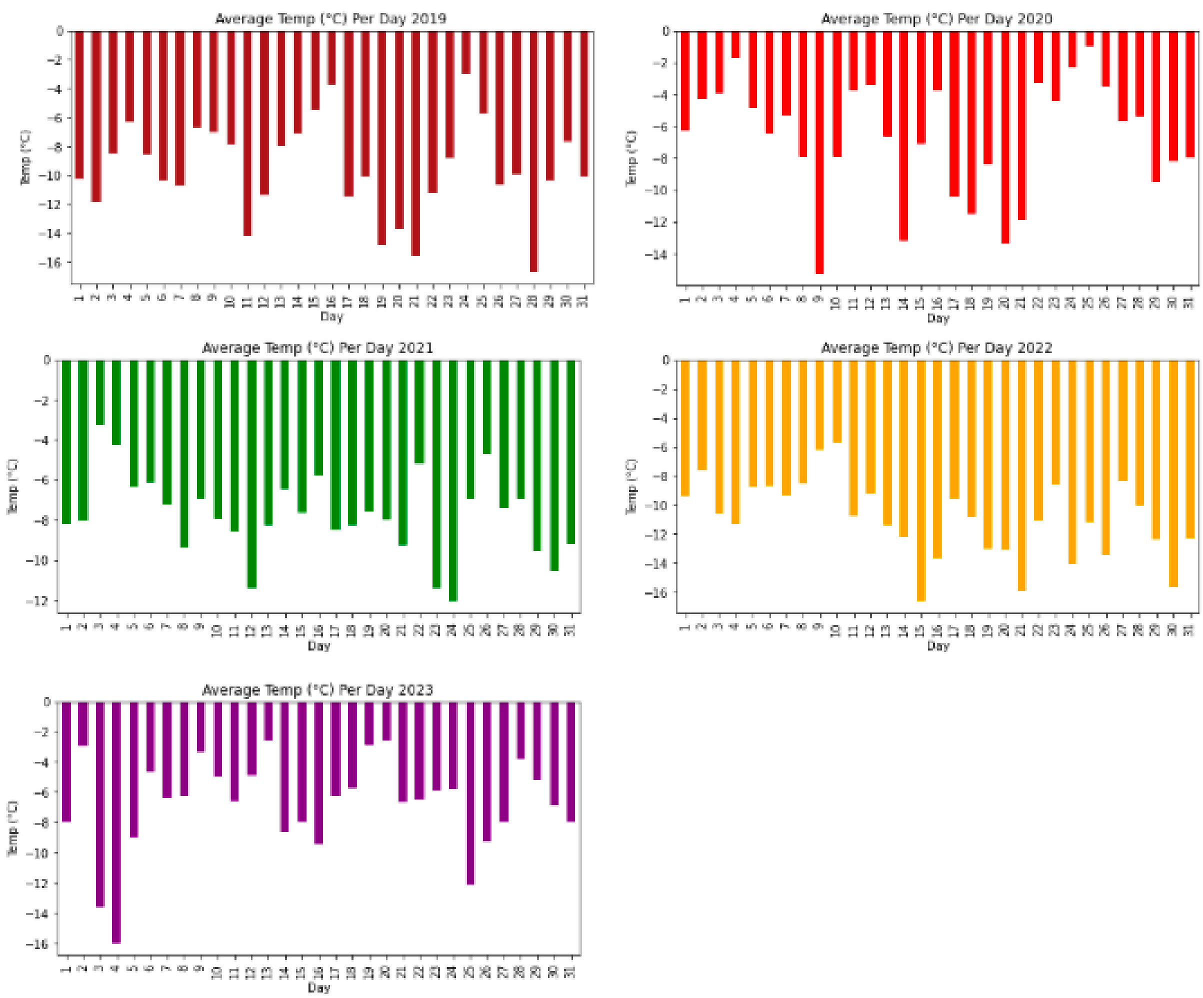

Figure 19 - Average Temperature Per Day 2019 to 2023

In figure 19, the coldest average temperature per day in 2019 was on the 28th day of each of the months in the winter season. The warmest average temperature per day in 2019 was on the 24th day of each of the months in the winter season. The coldest average temperature per day in 2020 was on the 9th day of each of the months in the winter season. The warmest average temperature per day in

2020 was on the 25th day of each of the months in the winter season. The coldest average temperature per day in 2021 was on the 24th day of each of the months in the winter season. The warmest average temperature per day in 2021 was on the 3rd day of each of the months in the winter season. The coldest average temperature per day in 2022 was on the 15th day of each of the months in the winter season. The warmest average temperature per day in 2022 was on the 10th day of each of the months in the winter season. The coldest average temperature per day in 2023 was on the 4th day of each of the months in the winter season. The warmest average temperature per day in 2023 was on the 13th day of each of the months in the winter season. Overall, no pattern or trend can be found in average temperature per day over the years 2019 to 2023.

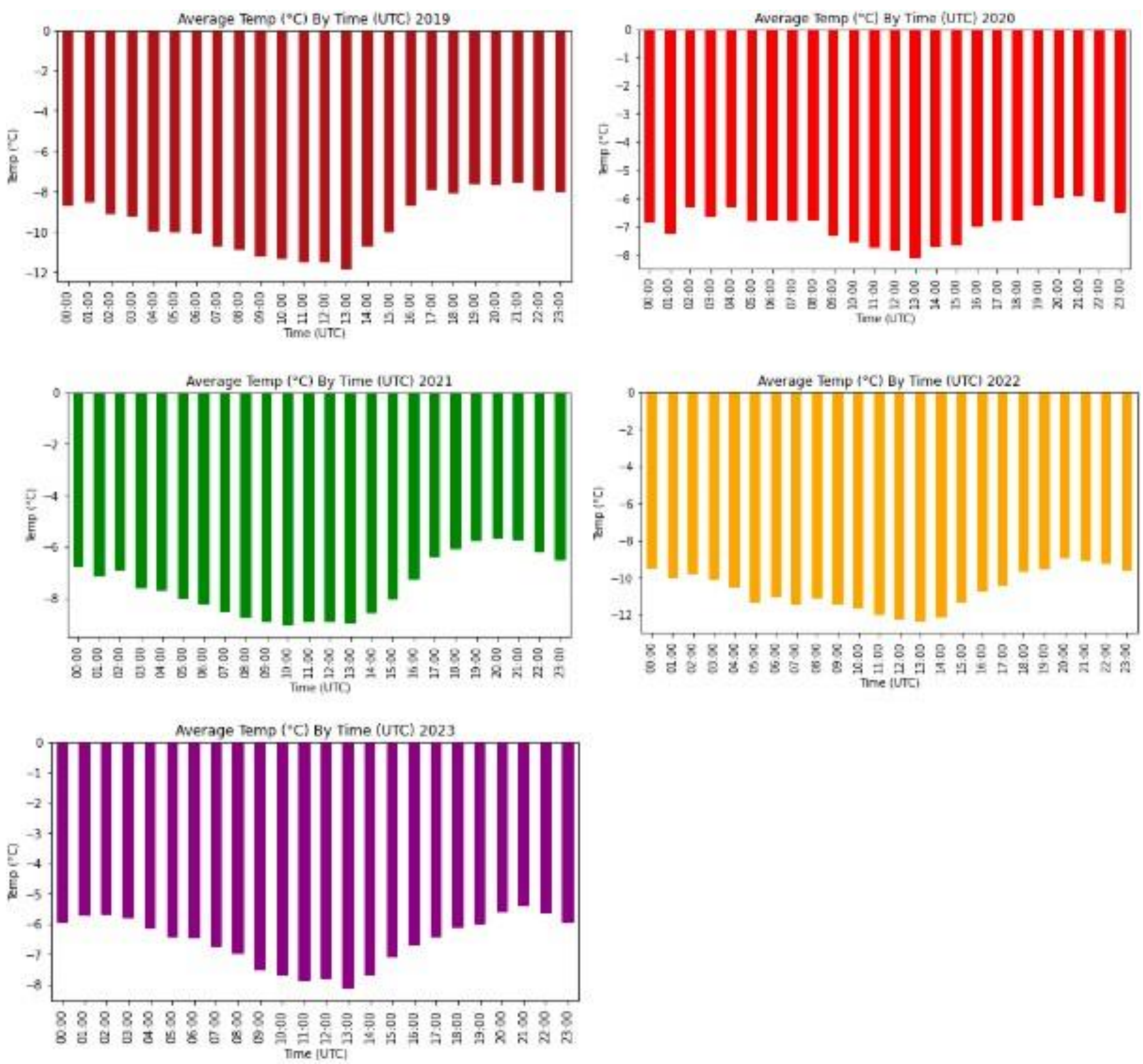

Figure 20 - Average Temperature By Time 2019 to 2023

In figure 20, for 2019, 2020, 2022, and 2023, the average coldest temperature by time in was at 13:00 for each day of the winter season. For 2021, the average coldest temperature by time in was at 10:00 for each day of the winter season. For 2019, 2020, and 2023, the average warmest temperature by time was at 21:00 for each day of the winter season. For 2021 and 2022, the average warmest temperature by time was at 20:00 for each day of the winter season. A pattern can be seen in the hourly temperature graphs where temperature is high around the early mornings and decreases throughout the day to have the lowest temperature during the noon hours then increases steadily to another high during the evenings to late night to repeat the cycle. This pattern in hourly temperature follows the same pattern found in hourly wind chill. This provides evidence for a seasonal temperature pattern within each 24 hour cycle.

Overall, when comparing the wind chill, dew point temperature, and temperature plots from figures 12 to 20, it can be seen that for averages by time, wind chill and temperature follow the same patterns of being high in the mornings, decreasing throughout the day and increasing again during the late evenings and night to repeat the cycle. Regarding average daily values, wind chill, dew point temperature, and temperature follow the same patterns, however the values for wind chill and dew point temperature are more closely related to each other than temperature. Regarding monthly average values, wind chill, dew point temperature, and temperature all follow the same patterns but the values of wind chill and dew point temperature are closer than with temperature. In 2019 they all have an increasing pattern from January to December, in 2020 the same random trend from January to March, in 2022 the same increasing pattern from January to March, and 2023 the same decreasing pattern from January to February.

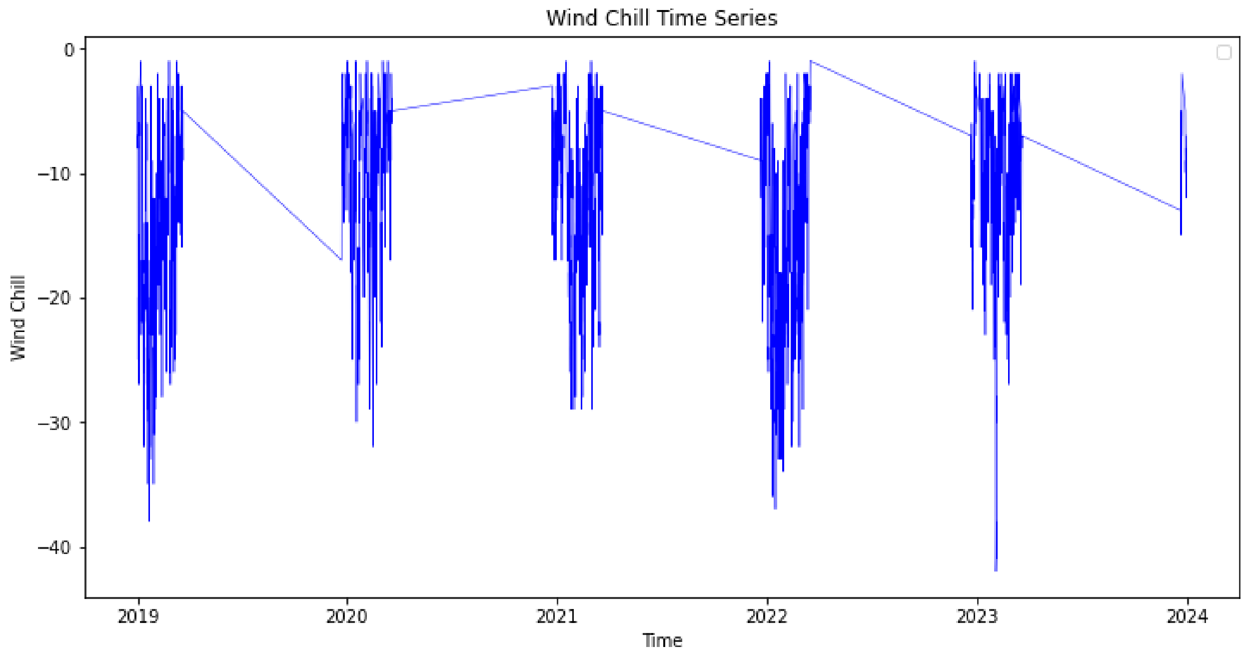

Figure 21 - Wind Chill Time Series 2019-2023

The wind chill time series from 2019 to 2023 can be seen in figure 21. It can be seen that 2023 experienced a minimum windchill value that was not previously seen in any years and that 2022 had many very low wind chill values compared to the other years. Overall, no seasonality or trend can be seen over the years.

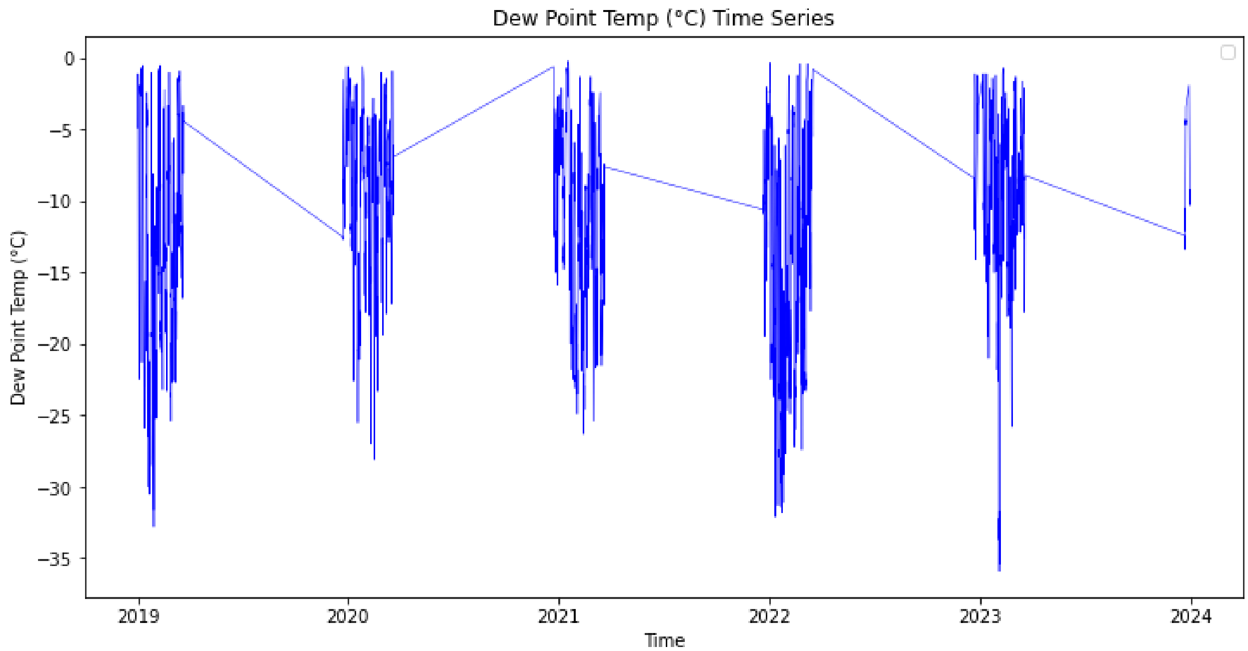

Figure 22 - Dew Point Temperature Time Series 2019-2023

The dew point temperature time series from 2019 to 2023 can be seen in figure 22. It can be seen that 2023 experienced a minimum dew point temperature value that was not previously seen in any years and that 2022 had many very low dew point temperature values compared to the other years. Overall, no seasonality or trend can be seen over the years.

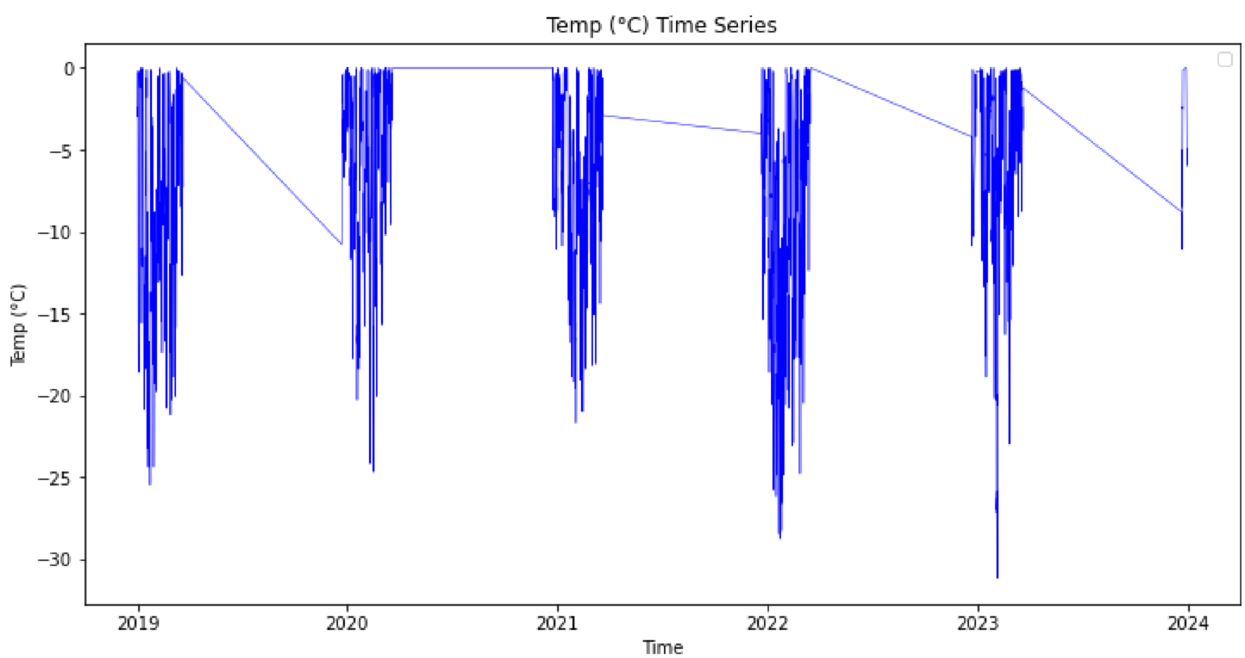

Figure 23 - Temperature Time Series 2019-2023

The temperature time series from 2019 to 2023 can be seen in figure 23. It can be seen that 2023 experienced a minimum temperature value that was not previously seen in any years and that 2022 had many very low temperature values compared to the other years. It can also be seen that each winter the maximum temperature is about within the same range, hovering under 0 showing very little to no effect from the warmings of climate change. The 2023 winter can be seen as the warmest over the years. Overall, no seasonality or trend can be seen over the years.

The time series plots in figures 21 to 23 show that wind chill, dew point temperature, and temperature all look very similar following the same patterns. The values of wind chill and dew point temperature are more similar than to temperature. Temperature has a more even data structure near the maximum temperatures over every winter season compared to wind chill and dew point temperature. The documented analysis and figures 12-23 support the objectives of understanding the relationships between variables included when recording weather data and visualizing and identifying patterns in temperature and variables highly correlated with it. One of the research questions has also been answered that there are no clear trends or patterns found within the data to show signs of climate warming.

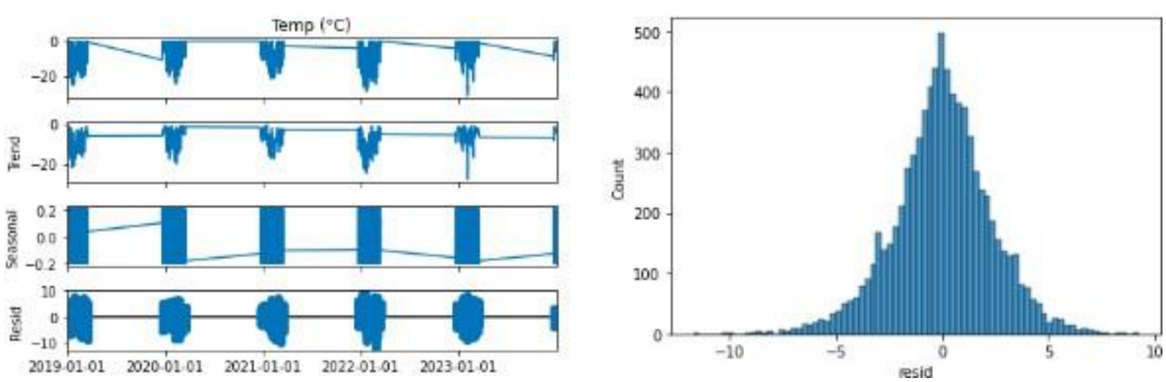

Figure 24 - Seasonal Decomposition 2019 to 2023 On The Left & Residuals Histoplot On the Right

The seasonal decomposition plot in figure 24 for the 4 years shows there is seasonality and almost no trend. The residuals cannot be seen clearly and were plotted in their histogram plot showing they are normally distributed indicating no pattern is left after removing trend and seasonality from the data. Due to the difficulty to see the exact period that is followed for seasonality, seasonal decomposition for each year covering January to March was done for an in depth analysis. The results can be seen in figure 25. December was not included in these figures as the winter period included in the data for each year for this month had many missing values and covers less than 10 days so it would not help produce a clear conclusion on seasonality.

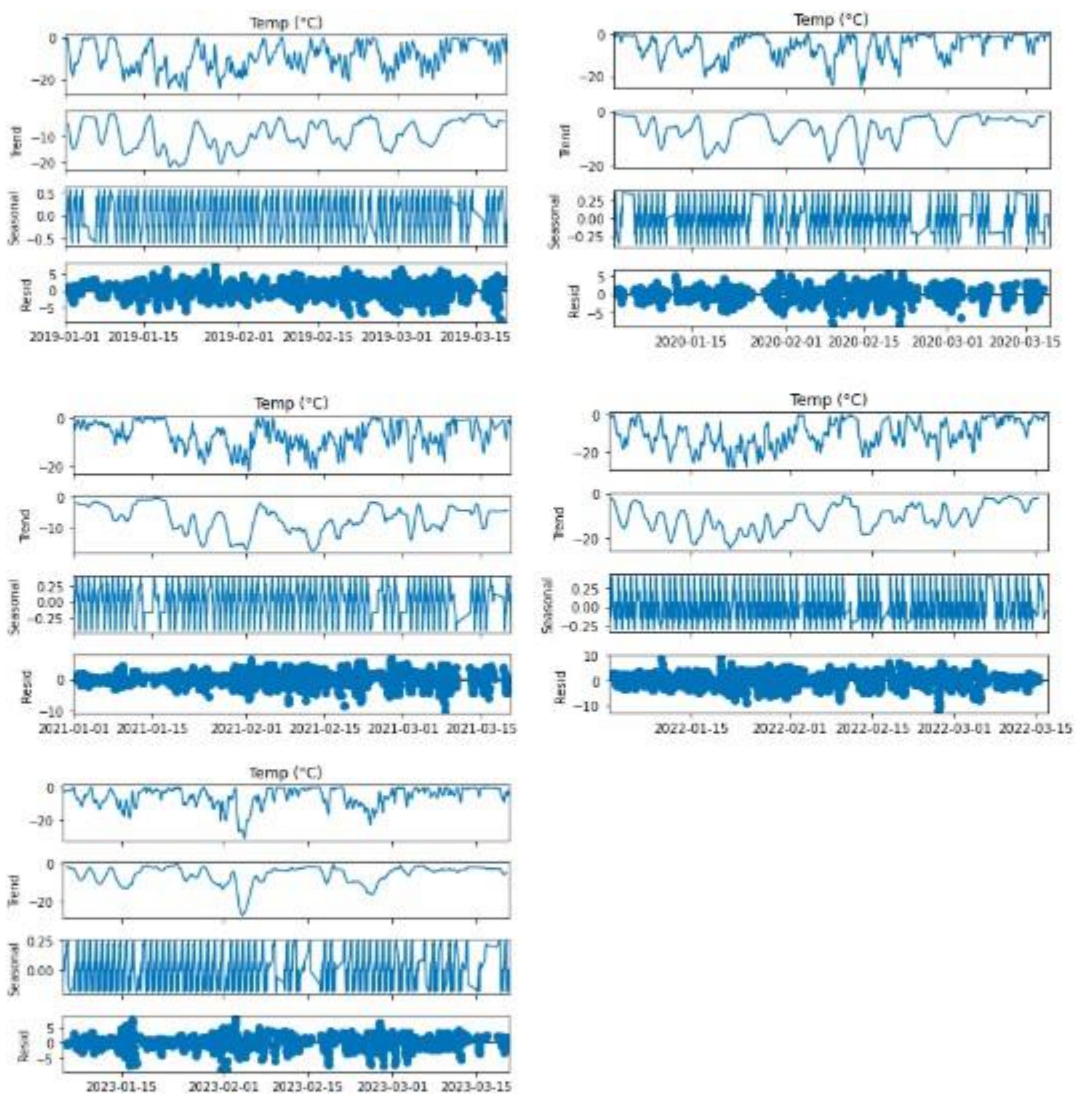

Figure 25 - Seasonal Decomposition 2019, 2020, 2021, 2022, 2023

In figure 25, clear 24 hour seasonality can be seen, all residuals follow random distribution, and little trend is seen. This supports using 24 hour seasonality in SARIMAX and continuing to the next phase.

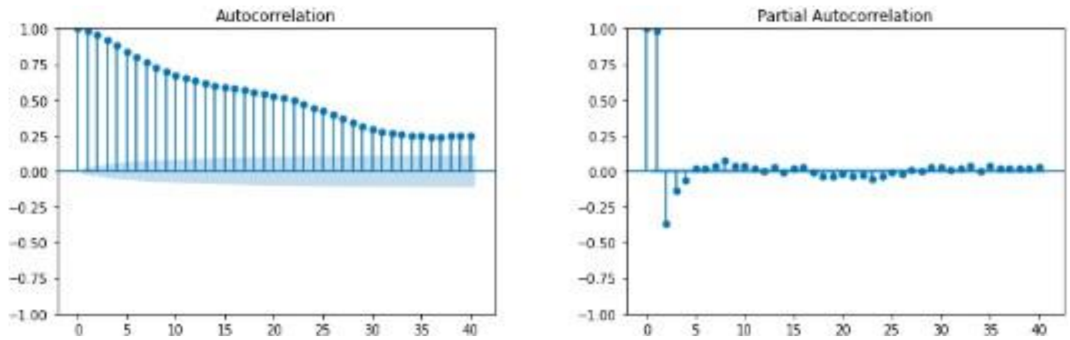

Figure 26 - Autocorrelation Plot On The Left and Partial Autocorrelation Plot on the Right

The autocorrelation plot in figure 26 shows a gradual decay in the lags indicating an autoregressive model. The moving average cannot be determined from this plot. There are no recurring peaks at specific lags indicating very weak seasonality patterns in the data. This makes sense as the seasonality discovered is daily which is very small to be seen over 4 years of data. The partial autocorrelation plot has spikes for the first 3 lags indicating an autoregressive term of 3 to be used in the SARIMA modelling. The lags cut off suddenly indicating the model follows an autoregressive model. The moving average cannot be determined from this plot, thus leading to using the auto_arima model selection method to confirm the best parameters to be used for the model.

9.2 Model Output

Table 4- Model Selection

Model	Best Model	AIC
ARIMA	ARIMA (3,0,5)	17957.587
SARIMAX	SARIMAX (3,0,5)(0,0,0)(24)	17957.587

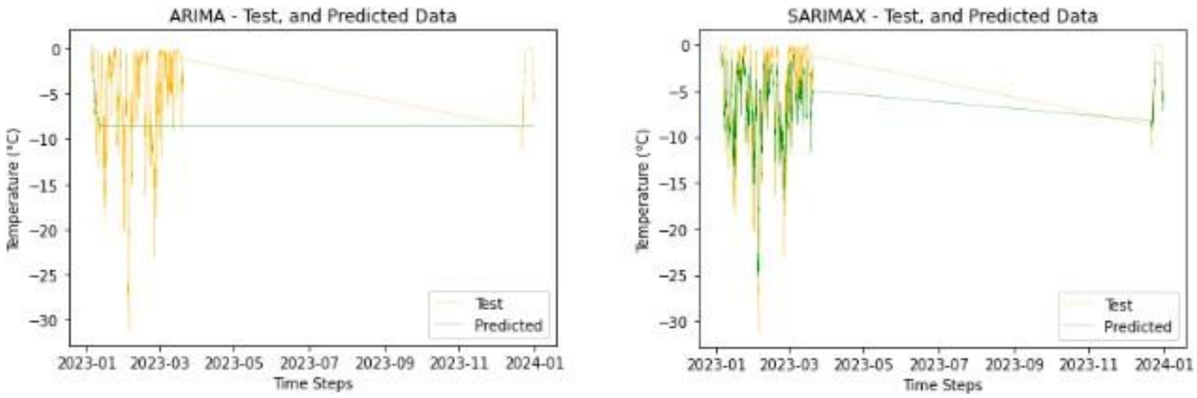

Figure 27 - ARIMA and SARIMAX Results

The ARIMA and SARIMAX models identified in table 4 have results shown in figure 27. It can be seen that ARIMA did not perform well when predicting the hourly temperature for the winter season ahead. SARIMAX however has performed well and has many overlapping areas with the test set and mimics the test structure well. ARIMA had an MSE of 34.332, and MAE of 4.752 shown in table 5. These values indicate that there were large errors and the data is spread very far away from the mean. The ARIMA model did not fit the data well. The SARIMAX model had an MSE of 3.841, and MAE of 1.578 shown in table 5. These values indicate that SARIMAX fit the data well. The SARIMAX model had very little error but can be improved.

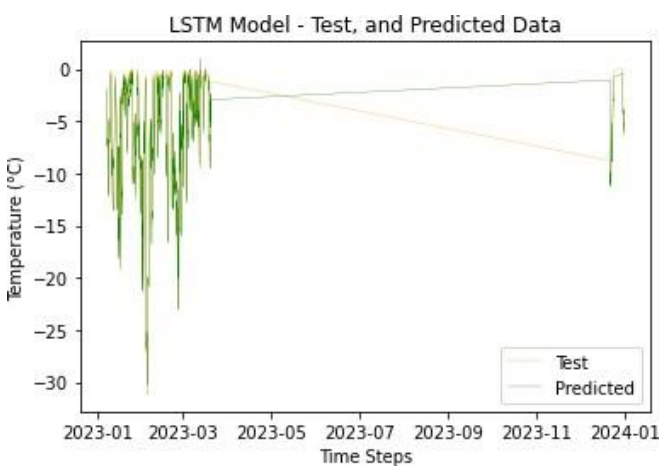

Figure 28 - LSTM Results

The LSTM model performed extremely well as shown in figure 28, the data mimics the test structure very well. This model had an MSE of 0.987, and MAE of 0.654, shown in table 5. These values indicate the predictions had very little error when compared to the test values.

Table 5 Models and Evaluation Metrics

Model	MSE	MAE
SARIMAX	3.841	1.578
LSTM	0.987	0.654
ARIMA	34.332	4.752

Overall, the LSTM and SARIMAX model performed very well for predicting hourly winter data one year ahead. The ARIMA model did not perform well and the exclusion of exogenous factors or seasonality may be the reason for this. The model that performed the best was LSTM. This may be due to the fact that LSTM works by using memory and sees that hourly temperatures have close relationships with the temperatures of the previous hours when put together in a sequence. This analysis has completed the objectives of developing and deploying ARIMNA, SARIMAX, and LSTM for hourly prediction and identifying which performs best. The second research question has also been answered that LSTM has very high accuracy, followed by SARIMAX with moderate accuracy and ARIMA with almost no accuracy.

10 Project Management

10.1 Project Schedule

In order to finish my project on time, I set daily reminders on my phone and on the calendar application on my laptop. These helped me track my project status and ensure I have enough time to complete the project within the allocated time frame.

My original plan began in October where I selected the topic artificial intelligence safety and was planning to investigate the worldwide opinions regarding the topic. A detailed project plan is shown in Appendix A table 7. However, on November 8th, I was informed that my original project needed to be changed due to issues with the dataset. I did an extensive search that day to find a new topic that I was interested in and that would have a large amount of data to work with. By November 13th, I found a new topic and proposed the idea to my supervisor. My supervisor was happy with the topic and excited to work alongside me on the project journey. On November 14th I submitted a detailed proposal going over the project specifications and received approval from my supervisor to submit an ethics application for project approval. Due to the limited timeframe to complete the project, I developed a new plan that same day which can be seen in Appendix A table 8. That same day my supervisor had also agreed to increase our biweekly meetings to weekly sessions to provide me with any additional guidance or feedback needed in the process. Meeting records can be found in appendix B.

For the new project idea, I initially aimed to forecast hourly temperatures two years ahead. However, this required merging 60 CSV files which was not optimal for the available time frame. To adjust, I reviewed the literature to see if I could target a specific season within Canada that comes with challenges due to climate change and would produce valuable results that can benefit Canadians. The results of my literature review found that recently winter climate analysis has been of high interest in Canada due to many wildlife issues arising from climate change during that season. This became the new focus of my project. By narrowing the scope and remaining on the same topic I reduced the file conversion to 20 files and gave myself more time to pre-process the data and focus on project quality. Another challenge experienced during the project was missing value handling. Due to the data being hourly temperature, it was very important to have minimal missing values so that the models can produce accurate predictions. Therefore, this included an extensive search within the glossary of the dataset attributes to determine how values were recorded or calculated. The values that could be calculated were calculated for, the remaining data if less than 10% was missing it was interpolated based on time, and the remaining data that could not be filled for was removed. Due to this, not every timestamp was included in the years so I had to adjust the predictions from 2 winter seasons ahead to one winter season ahead for the model's to have sufficient data to work. The data preprocessing took longer than the allocated time frame and extended into week 11 which was not a part of the project plan in Appendix A table 8. However, with these challenges I was still able to meet my goal of completing a project draft for Week 12 and ensuring the project maintained its high quality within the available time frame.

10.2 Risk Management

Table 6- Risk Management

Risk	Consequences	Mitigation Plan	Did it Happen?
Using inaccurate data in the analysis	Analysis could show incorrect trends or patterns, predictions may have bias	Ensuring when downloading each separate dataset that each is in UTC time and corresponds to the months being studied	No
Certain project sections taking longer than the approximated timeframe	Readjusting the project plan, unexpected changes, increased stress	Giving myself more time than allocated for each task in case any adjustments need to be made, reassessing the project plan regularly	Yes
Coding difficulties, different information from different sources	Could lead to longer work hours, conducting same	Double check techniques with many sources, include	Yes

	analysis more than once, rewriting and adjusting code many times	YouTube video tutorials in learning, use textbooks instead of online blogs or articles that are not peer reviewed or academically published, confirm with professors in the digital literacy center or statistics center	
Need for high computational processing power	Could lead to long code processing time, system crashes, sudden changes in the project plan	Splitting tasks for the system to follow, use the University computers, use online coding platforms such as Google Colab, adjust the code if possible to follow different steps	Yes

10.3 Quality Management

To ensure high quality of the project I set standards regarding the ways I addressed missing values within the data. I set the criteria being if less than 10% of the column observations are missing I would fill based on Pandas interpolate method, if more than 30% was missing the column was completely removed, and if between 10% and 20% were missing I would remove the rows containing missing values. This ensured that missing values would not greatly affect the results of the predictions and that the data still contained high accuracy. Additionally, a literature search was done of the Government of Canada sites to determine which observations could be manually calculated for. This resulted in being able to calculate wind chill values which was important to keep in the data as it is highly correlated with temperature. Another literature search was done of the Government of Canada sites to determine the exact winter season timestamps which helped maintain seasonal accuracy. Lastly, the evaluation metrics used in the project were based on the positive results of the literature review as MSE and MAE, were used in many of the literature reviews to evaluate model performance.

10.4 Social, Ethical, & Legal Considerations

While completing the project and creating the prediction models no social, ethical or legal issues arose. The data used to fit the models is free, publicly available and does not contain any sensitive information. Additionally, any sources that have been used in this project are cited and can be found in the bibliography and references section. Ethical certificate is included in appendix E. Social, ethical, and legal considerations regarding model deployment have been looked into and are included below.

Social Considerations

Implementing a model for temperature prediction includes many social considerations that need to be considered before the model is deployed for public use. A social consideration that needs to be considered is the possibility of unemployment among forecasters, meteorologists, and scientists. Incorporating temperature prediction models into workplaces that analyze weather data can cause challenges between humans and artificial intelligence. Especially if the models are depended on completely and human intervention or evaluations of the model's training and production methods are not done. Challenges may also arise if the model does not provide transparency to how the decisions are made which can cause humans that work with the models to have little trust in them. A balance would need to be incorporated into workplaces among humans and artificial intelligence in order to keep judgment and competence of the predicted temperatures. Another social issue to consider is that the data provided to the model's may be under-represented in certain regions. This can occur if temperature data is not consistently collected or limited due to resource limitations. This may lead to biased data going into the training models which would have a negative effect on the accuracy of the predicted temperatures. This could worsen inequities and lead to environmental injustice.

Ethical Considerations

Regarding ethical considerations, it is important to keep in mind that the model once deployed should be available for use in all regions and not specific to use in certain more developed regions. It is also important to consider the risks that the model may have if predictions are incorrect. As incorrect predictions could lead to misallocating resources or under-representing weather risks leaving people at risk of possible extreme weather consequences. Lastly, before the model is implemented into society, risks the predictions may have on animal wellbeing and the other living organisms in the ecosystem need to be addressed.

Legal Considerations

The prediction model once implemented would need to follow legal standards in each country it is deployed in. For example, if deployed to use in the United States, the governing organizations guidelines that the model would need to follow include the American Meteorological Society and the National Science Foundation’s artificial intelligence framework. Each country has different governing organizations and guidelines so this is important to keep in mind during model deployment. Another legal consideration is accountability for incorrect predictions if they cause financial loss or other harmful consequences to individuals, businesses, or governments. It is essential that along with deployment to include education about the model that it is not to be relied on completely and one's own judgement should be used alongside the model.

11 Critical Appraisal

The main successes of the project were getting accurate temperature predictions using the SARIMAX and LSTM models. Although only 4 years of historical data was used for training the models, they are still able to produce accurate predictions. A limitation in the project is that the training data was missing 28 timestamps. This amounted to about 0.02% of data missing; however, if the data were included for those timestamps, it is possible accuracy could have been improved, leading to better model performance. An unexpected result was the failure of the ARIMA model to predict temperatures. This most likely was due to the lack of seasonality and exogenous variables in the model training. Another unexpected finding were the minimal patterns found in the wind chill, dew point temperature, and temperature hourly, daily, monthly, and yearly plots. An expected finding in the project is not being able to see clear trends or patterns of climate warming as the changes may be extremely small, about 1 degree Celsius or less. Overall, studying the winter season temperature predictions has been a very interesting topic and different to what is found in the literature. This project gave me a deeper understanding of both the theory and application of time series analysis and deep learning recurrent neural network models. I am happy with this as I have had experience with the traditional machine learning techniques including linear regression, SVM, K-NN, but have not had experience with time series or deep learning prior to this project.

12 Conclusions

12.1 Achievements

The aim to accurately predict hourly air temperature for an upcoming winter season has been achieved as well as the aim to compare ARIMA, SARIMAX, and LSTM prediction models to identify which has the best performance. The objectives of the project were all met. Relationships between different variables were analyzed by the correlation matrix, visualization of patterns and trends within temperature and its highly correlated variables were done using hourly, daily, monthly, and yearly plots. Development and deployment of the different prediction models were done and evaluations for the comparisons were carried out by 3 evaluation metrics. The research questions have also been answered. LSTM can provide high accuracy, SARIMAX moderate accuracy, and ARIMA unfortunately close to no accuracy. Unfortunately, no patterns or trends were found to see the changes of climate change, and this may be due to not using enough historical data. ARIMA not performing well was an unexpected result but identifies that seasonality and exogenous variables play important roles in predictions. Overall, the project went well, the main problem was the limited timeframe affecting the depth of the project.

12.2 Future Work

For future work, I would like to explore more time series models and neural networks. I would also like to use more historical data to view the changes of climate change as the data used may not have been enough showing no changes. I would also like to see if I could forecast temperature for other seasons such as summer where there is no wind chill. This would be interesting as wind chill is a highly correlated variable with temperature in the winter data so without it in summer makes me wonder if the forecast would work or not. I would also like to look at further prediction timeframes.

12.3 Student Reflections

This project came with many achievements, both personal and professional. My ability to effectively manage my time under limited timeframes and use creative thinking to work around the challenges faced during this project to deliver a high quality final year project to showcase my skills gained in the master's course is a large personal achievement for myself. Another personal achievement was keeping calm and following my project plan even when the deadline was fast approaching. This reminded me that I was motivated each step of the way and that no task is too difficult to handle. Prior to this project, I had moderate experience with coding for machine learning, however during this project I immersed myself into learning many new techniques. I gained more skills in pre-processing data, creating visualizations, fitting and training models. This professional achievement will help me a lot when completing projects in my future work and when doing application exams for certain jobs. Another professional achievement is that I was able to achieve all the aims, objectives, and research questions set up in this project. This shows my ability to complete all tasks assigned to myself. Overall, I am very happy with the results of the project and am very happy that 2 out of the three models were accurate in predicting temperature for the winter season ahead. To conclude, beyond this project, I am interested in using the LSTM model in my personal life as I am a very adventurous person and have a strong hobby for travelling and believe this model can benefit me when making travel decisions.

Bibliography & References

Ahmed, I. (2023, May 31). What are ACF and PACF plots in time series analysis?. Medium. https://ilyasbinsalih.medium.com/what-are-acf-and-pacf-plots-in-time-series-analysis-cb586b119c5d

Anjali, T., Chandini, K., Anoop, K., & Lajish, V. L. (2019, July). Temperature prediction using machine learning approaches. In 2019 2nd International conference on intelligent computing, instrumentation and control technologies (ICICICT) (Vol. 1, pp. 1264-1268). IEEE.

Argonne. (2024, July 23). Advanced hyper-local air temperature prediction in urban environments. Argonne National Laboratory. https://www.anl.gov/evs/article/advanced-hyperlocal-air-temperature-prediction-in-urban-environments#:~:text=Second%2C%20being%20able%20to%20estimate,airflow%20and%20cool%20the%20space

Azari, B., Hassan, K., Pierce, J., & Ebrahimi, S. (2022). Evaluation of machine learning methods application in temperature prediction. Environ Eng, 8(1), 1-12.

Bhandari, A. (2024, November 29). Feature scaling: Engineering, Normalization, and standardization (updated 2025). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Binus University. (2021, September 30). Data Mining Semma. School of Information Systems. https://sis.binus.ac.id/2021/09/30/data-mining-semma/

Chen, P., Niu, A., Liu, D., Jiang, W., & Ma, B. (2018, August). Time series forecasting of temperatures using SARIMA: An example from Nanjing. In IOP conference series: materials science and engineering (Vol. 394, p. 052024). IOP Publishing.

Edenido, N. (2023, June 5). Travel to Europe is really hot, but some seek cooler destinations: Travel Weekly. Travel Weekly. https://www.travelweekly.com/Travel-News/Tour-Operators/Travelers-seek-cooler-destinations-this-summer

EPA. (2024a, August 20). Heat Island Impacts. https://www.epa.gov/heatislands/heat-island-impacts#:~:text=Heat%20islands%20contribute%20to%20higher,and%20non%2Dfatal%20heat%20stroke

EPA. (2024b, September 23). Climate Change Impacts on Agriculture and Food Supply. https://www.epa.gov/climateimpacts/climate-change-impacts-agriculture-and-food-supply

EPA. (2024c, October 22). Climate Change and Human Health. https://www.epa.gov/climateimpacts/climate-change-and-human-health

EPA. (2024d, October 30). Heat Island Effect. https://www.epa.gov/heatislands#:~:text=Structures%20such%20as%20buildings%2C%20roads,temperatures%20relative%20to%20outlying%20areas

European Environment Agency. (2024, May 17). Floating or elevated roads. Climate Adapt. https://climate-adapt.eea.europa.eu/en/metadata/adaptation-options/floating-or-elevated-roads#:~:text=Elevated%20roads%20are%20effective%20against,situated%20higher%20than%20water%20levels

Fister, D., Pérez-Aracil, J., Peláez-Rodríguez, C., Del Ser, J., & Salcedo-Sanz, S. (2023). Accurate long-term air temperature prediction with machine learning models and data reduction techniques. Applied Soft Computing, 136, 110118.

Government of Canada. (2023, June 12). Government of Canada. Weather tools: frequently asked questions - Canada.ca. https://www.canada.ca/en/environment-climate-change/services/weather-general-tools-resources/frequently-asked-questions.html

Government of Canada. (2024a, October 1). Glossary. Climate Weather Glossary. https://climate.weather.gc.ca/glossary_e.html#stnOper

Government of Canada. (2024b, October 1). Hourly Data Report for February 01, 2020. Climate. https://climate.weather.gc.ca/climate_data/hourly_data_e.html?timeframe=1&hlyRange=2000-10-19%7C2024-12-08&dlyRange=2000-10-19%7C2024-12-08&mlyRange=2003-04-01%7C2006-12-01&StationID=30578&Prov=ON&urlExtension=_e.html&searchType=stnProx&optLimit=specDate&StartYear=1840&EndYear=2016&selRowPerPage=25&Line=0&txtRadius=25&optProxType=navLink&time=UTC&Year=2020&Month=2&Day=1#

Hamad, R. (2023, December 11). What is LSTM? introduction to long short-term memory. Medium. https://medium.com/@rebeen.jaff/what-is-lstm-introduction-to-long-short-term-memory-66bd3855b9ce

Helmore, E. (2023, July 23). Extreme heat forces US airlines to limit passengers and fuel loads. The Guardian. https://www.theguardian.com/us-news/2023/jul/23/us-airlines-heat-delays-limit-passengers-fuel

Howes, N. (2024, November 3). Climate change effects on canadian wildlife show how vital winter is. The Weather Network. https://www.theweathernetwork.com/en/news/climate/impacts/climate-change-effects-on-canadian-wildlife-shows-how-vital-winter-is

IBM. (2024, October 28). What is machine learning (ML)? https://www.ibm.com/topics/machine-learning

Jonatasv. (2024, February 26). Metrics evaluation: MSE, RMSE, Mae and mape. Medium. https://medium.com/@jonatasv/metrics-evaluation-mse-rmse-mae-and-mape-317cab85a26b

Khedhiri, S. (2022). Comparison of SARFIMA and LSTM methods to model and to forecast Canadian temperature. Regional Statistics, 12(02), 177-194.

Kim, J. (2024, April 3). Exploring the Seasonal Arima (Sarima) model for forecasting: Differences from Arima. Medium. https://medium.com/@jihyun.kim423/exploring-the-seasonal-arima-sarima-model-for-forecasting-differences-from-arima-e30c3488e5f6

Li, X., Li, Z., Huang, W., & Zhou, P. (2020). Performance of statistical and machine learning ensembles for daily temperature downscaling. Theoretical and Applied Climatology, 140(1), 571-588.

Malkari, N. (2023, April 16). Seasonal decomposition. Medium. https://medium.com/@nikhilmalkari18/seasonal-decomposition-425a2d7490e8

Melanie. (2024, March 14). SARIMAX model: What is it? how can it be applied to time series?. DataScientest. https://datascientest.com/en/sarimax-model-what-is-it-how-can-it-be-applied-to-time-series

Met Office. (2024). How weather forecasts are created. https://www.metoffice.gov.uk/weather/learn-about/how-forecasts-are-made

Mung, P. S., & Phyu, S. (2023, February). Time Series Weather Data Forecasting Using Deep Learning. In 2023 IEEE Conference on Computer Applications (ICCA) (pp. 254-259). IEEE.

Musbah, H., & El-Hawary, M. (2019, May). SARIMA model forecasting of short-term electrical load data augmented by fast fourier transform seasonality detection. In 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE) (pp. 1-4). IEEE.

NASA. (2022, November 10). Climate change can put more insects at risk for extinction – climate change: Vital signs of the planet. https://climate.nasa.gov/news/3231/climate-change-can-put-more-insects-at-risk-for-extinction/#:~:text=New%20NASA%20research%20models%20are,extinct%20over%20the%20next%20century

NCAS. (2023). Why is climate important? https://ncas.ac.uk/learn/why-is-climate-important/#:~:text=Studying%20the%20climate%20helps%20us,are%20threatened%20by%20climate%20change

NRC. (2024, November 21). When do the seasons start?https://nrc.canada.ca/en/certifications-evaluations-standards/canadas-official-time/3-when-do-seasons-start

Ottawa Weatherstats. (n.d.). About Ottawa (kanata - orléans). Amateur Weather Statistics for Ottawa (Kanata - Orléans), Ontario. https://ottawa.weatherstats.ca/about.html

Omstead, J. (2024, November 27). After warmest on record, winter looks to "salvage its reputation": Weather network. CTVNews. https://www.ctvnews.ca/climate-and-environment/after-warmest-on-record-winter-looks-to-salvage-its-reputation-weather-network-1.7124904

Penn State Eberly College of Science. (2024). 4.1 Seasonal ARIMA models. STAT 510 Applied Time Series Analysis. https://online.stat.psu.edu/stat510/lesson/4/4.1

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., ... & Willson, M. (2024). Probabilistic weather forecasting with machine learning. Nature, 1-7.

Rajawat, A. S., Mohammed, O., Shaw, R. N., & Ghosh, A. (2022). Renewable energy system for industrial internet of things model using fusion-AI. In Applications of AI and IOT in Renewable Energy (pp. 107-128). Academic Press.

Santra, R. (2023, May 12). Tests for stationarity in Time Series - dickey fuller test & augmented Dickey Fuller(ADF) test. Medium. https://medium.com/@ritusantra/tests-for-stationarity-in-time-series-dickey-fuller-test-augmented-dickey-fuller-adf-test-d2e92e214360

Scott, D., & Lemieux, C. (2010). Weather and climate information for tourism. Procedia Environmental Sciences, 1, 146-183.

Shalaby, H., & Aboelnaga, S. (2017, November). Climate change impacts on urban planning in the cities. In 1st International Conference on Towards a Better Quality of Life.

Sharda, R., Delen, D., Turban, E., Aronson, J. E., Liang, T.-P., & King, D. (2018). Business Intelligence, analytics, and Data Science: A Managerial Perspective. Pearson.

Shine, I. (2023, August 14). Rising global temperatures are already affecting the tourism industry - here's how. World Economic Forum. https://www.weforum.org/stories/2023/08/temperatures-tourism-climate-impact/#:~:text=Greece%20–%20where%20travel%20and%20tourism,temperatures%20reached%2045°C.

SKtime. (2024). AutoARIMA. AutoARIMA - sktime documentation. https://www.sktime.net/en/latest/api_reference/auto_generated/sktime.forecasting.arima.AutoARIMA.html

Smith, H. (2023, July 15). Acropolis closes to protect tourists as Greece faces unprecedented heatwave. The Guardian. https://www.theguardian.com/world/2023/jul/15/acropolis-greece-closed-heatwave-48c

Tyler, A., Mcdonald, M., Rydge, J., Kobel, Y., & Chan, W.-S. (2024, February 20). AI: Weather forecasting - HSBC global research. HSBC Global Research. https://www.research.hsbc.com/C/1/1/320/9xhMRr6

UN. (2021). Climate resilience *. CLIMATE ACTION PATHWAY. https://unfccc.int/sites/default/files/resource/Resilience_Vision%26Summary.pdf

UN. (2024a). Causes and effects of climate change. United Nations. https://www.un.org/en/climatechange/science/causes-effects-climate-change#:~:text=As%20temperatures%20rise%2C%20more%20moisture,waters%20at%20the%20ocean%20surface

UN. (2024b). What is climate change?. United Nations. https://www.un.org/en/climatechange/what-is-climatechange#:~:text=Climate%20change%20refers%20to%20long,activity%20or%20large%20volcan ic%20eruptions.

University of Cambridge. (2014). Climate change - implications for the energy sector. World Energy Council. https://www.worldenergy.org/assets/images/imported/2014/06/Climate-Change-Implications-for-the-Energy-Sector-Summary-from-IPCC-AR5-2014-Full-report.pdf

USGS. (2024). How does climate change affect the challenge of invasive species?. How does climate change affect the challenge of invasive species? | U.S. Geological Survey. https://www.usgs.gov/faqs/how-does-climate-change-affect-challenge-invasive-species

WHO. (2023, October 12). Climate change. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/climate-change-and-health

WHO. (2024, May 28). Heat and health. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/climate-change-heat-and-health

World Travel and Tourism Council. (2022). Prospects for growth. TRAVEL & TOURISM IN THE CARIBBEAN. https://wttc.org/Portals/0/Documents/Reports/2022/Travel-and-tourism-in-the-caribbean.pdf

Appendix A Project Plans

Table 7- Original Project Plan

Task	Week and Time Commitment	Necessary Resources	Dependency
Literature review	Week 1	Google Scholar PubMed ScienceDirect Locate at Lanchester Library, Coventry University	N/A
Collecting the Data	Week 2	Online	Previous Task
Cleaning and Preparing the Data	Weeks 2-3	Python Excel CSV	Previous Task
Data Analysis, Visualization and further feature selection and feature engineering	Weeks 4-6	Tableau Python	Previous Task
Feedback on write up	Week 5	Meeting with Supervisor	Previous Task
Literature review; Machine learning models and related works	Week 6	Google Scholar PubMed ScienceDirect Locate at Lanchester Library, Coventry University	N/A
Choosing the Machine Learning Models for comparison	Week 7	Google Scholar PubMed ScienceDirect Locate at Lanchester Library, Coventry University	Previous Task
Training and evaluating the models	Week 7-9	Python	Previous Task

Parameter Tuning, retraining and evaluating the models	Week 7-9	Python	Previous Task
Feedback on write up	Week 9	Meeting with Supervisor	Previous Task
Conclusions and ethical, legal, and social considerations	Week 10	Google Scholar PubMed ScienceDirect Locate at Lanchester Library, Coventry University	Previous Task
Feedback on write up	Week 11	Meeting with Supervisor	Previous Task
Submission	Week 13	Aula	Previous Task

Table 8- Revised Project Plan

Task	Week and Time Commitment	Necessary Resources	Dependency
Literature review	Week 10	Google Scholar PubMed ScienceDirect Locate at Lanchester Library, Coventry University	N/A
Collecting the Data	Week 10	Government of Canada site	Previous Task
Cleaning and Preparing the Data	Weeks 10	Python Excel CSV	Previous Task
Data Analysis, Visualization and further feature selection and feature engineering	Week 11	Tableau Python	Previous Task
Feedback on write up	Week 11	Meeting with Supervisor	Previous Task

Training and evaluating the models	Week 11	Python	Previous Task
Parameter Tuning, retraining and evaluating the models	Week 11	Python	Previous Task
Conclusions and ethical, legal, and social considerations	Week 12	Google Scholar PubMed ScienceDirect Locate at Lanchester Library, Coventry University	Previous Task
Feedback on write up	Week 12	Meeting with Supervisor	Previous Task
Submission	Week 13	Aula	Previous Task

Appendix B Meeting Records

Meeting Minutes

Meeting 1

Date: October 11 2024

Time: 11:00am – 11:30am

Notes:

- Introduction to the module
- Supervisor talked about expectations
- Went over the project overview and what is needed for the project proposal
- Set dates for bi-weekly meetings

Meeting 2

Date: October 17 2024

Time: 10:00am – 10:30am

Notes:

- Identified topic: Predicting drug related deaths in Connecticut U.S
- Reviewed dataset and dissertation proposal feedback with supervisor
- Got approval to submit ethics application

Meeting 3

Date: October 24 2024

Time: 10:00am – 10:30am

Notes:

- Dataset was not appropriate for the timeframe or machine learning, so a new topic was proposed by me
- New topic was investigating worldwide opinions regarding artificial intelligence safety
- Went over the dissertation proposal and dataset with supervisor and got approval to submit another ethics application

Meeting 4

Date: November 7 2024

Time: 10:00am – 10:30am

Notes:

- Supervisor informed me that my dataset does not meet the requirements for the dissertation project and the dataset needs to be changed or the topic needs to be changed

Meeting 5

Date: November 8 2024

Time: 1pm – 1:30pm

Notes:

- I went over the new topic idea of temperature forecasting with the supervisor and got approval to send a project proposal and ethics application for the project

Meeting 6

Date: November 14 2024

Time: 10:00am – 10:30am

Notes:

- Supervisor approved the topic
- We spoke about the goal of the week and I informed the supervisor I would work on the coding for the project which included pre-processing, modelling, and testing

Meeting 7

Date: November 28 2024

Time: 10:00am – 10:30am

Notes:

- We went over the progress of the project, possibility of getting an extension, and ethics application changes
- Supervisor asked for the draft to be shared for feedback

Appendix C Dataset

All the hourly data was retrieved from:
https://climate.weather.gc.ca/historical_data/search_historic_data_e.html?hlyRange=2000-10-19%7C2024-11-06&dlyRange=2000-10-19%7C2024-11-06&mlyRange=2003-04-01%7C2006-12-01&urlExtension=_e.html&searchType=stnProv&optLimit=yearRange&StartYear=2000&EndYear=2024&selRowPerPage=25&Line=279&lstProvince=ON&timeframe=1&time=LST&Year=2022&Month=4&Day=23

The combined data is too large to attach to the project. It is uploaded to Microsoft OneDrive as a public Excel file:
https://livecoventryac-my.sharepoint.com/:x:/g/personal/saadawin_uni_coventry_ac_uk/EdAfXkbg2qpBpfXm9TqR1V4B7_2t30GIXSunzMjrTwMViA?rtime=meqNEEsa3Ug

Appendix D Code

The code is too large to be attached to the project and when run on Google Colab it used up all the available RAM and crashed and did not produce the model selection and output results. The code has been highlighted by a syntax highlighter and shared through Microsoft OneDrive in a public document:
https://livecoventryac-
my.sharepoint.com/:w:/g/personal/saadawin_uni_coventry_ac_uk/EY3D7Ojd5nVMtVhoNIJsRwABRzu2jOKDl9HI9Nz7QRRZeQ?rtime=0lS55Eoa3Ug

Appendix E Certificate of Ethics Approval

Certificate of Ethical Approval

Applicant:	Norhaan Saadawi
Project Title:	Time Series Temperature Forecasting in Canada's Capital

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:	05 Dec 2024
Project Reference Number:	P182047