

**M.Sc. Data Analytics & Technologies**

2023-2024

**Predictive Modelling of Weather Patterns for Urban Planning in Cuxton, Kent**

STUDENT NAME: Oluseye Emmanuel Fasuan

STUDENT ID: 2308483

PROGRAMME NAME: M.Sc. Data Analytics & Technologies

MODULE TITLE: DAT7006 Data Science

ASSESSMENT TITLE: Assessment 2

INSTRUCTOR: Dr Anchal Garg

DATE: 15th May 2024.

Table of Contents

[1.0 Introduction 4](#_Toc166678812)

[1.1 Problem statement and Stakeholders 5](#_Toc166678813)

[1.2 Research Questions: 6](#_Toc166678814)

[1.3 Hypothesis (for statistical analysis) 6](#_Toc166678815)

[1.4 Significance of data analysis 7](#_Toc166678816)

[2.0 Literature Review: 8](#_Toc166678817)

[3.0 Methodology: 10](#_Toc166678818)

[3.1 CRISP-DM: 10](#_Toc166678819)

[3.1.1 Business Understanding: 10](#_Toc166678820)

[3.1.2 Data Understanding: 11](#_Toc166678821)

[3.1.3 Data Preparation: 12](#_Toc166678822)

[3.1.4 Modelling: 18](#_Toc166678823)

[3.1.5 Evaluation: 19](#_Toc166678824)

[3.1.6 Deployment: 20](#_Toc166678825)

[3.2 Data Pre-processing Steps 23](#_Toc166678826)

[3.2.1 Loading the dataset: 23](#_Toc166678827)

[3.2.2 Reshaping the dataset: 23](#_Toc166678828)

[3.2.3 Handling missing values: 24](#_Toc166678829)

[3.2.4 Calculating wind speed: 25](#_Toc166678830)

[3.2.5 Converting datetime format: 25](#_Toc166678831)

[3.3 Exploratory Data Analysis (EDA) 25](#_Toc166678832)

[3.3.1 Summary statistics: 25](#_Toc166678833)

[3.3.2 Missing value analysis: 26](#_Toc166678834)

[3.3.3 Distribution analysis 26](#_Toc166678835)

[3.4 Statistical Analysis 28](#_Toc166678836)

[3.4.1 Univariate Analysis 28](#_Toc166678837)

[3.4.2 Bivariate Analysis 29](#_Toc166678838)

[3.4.3 Multivariate Analysis 30](#_Toc166678839)

[3.4.4 Time-series Forecasting 33](#_Toc166678840)

[3.4.5 Machine Learning – Regression 42](#_Toc166678841)

[4.0 Discussion 46](#_Toc166678842)

[5.0 Conclusion 48](#_Toc166678843)

Table of Figures

[Figure 1: Step by Step Flow of CRISP-DM 12](#_Toc166678943)

[Figure 2: Extraction of Xlat and Xlong 14](#_Toc166678944)

[Figure 3: Removing Xlat and Xlong 15](#_Toc166678945)

[Figure 4: Replacing a wrong column name X,2225 15](#_Toc166678946)

[Figure 5:Handling the black rows 15](#_Toc166678947)

[Figure 6: Identify Outliers in the dataframe 16](#_Toc166678948)

[Figure 7: Reporting columns containing outliers 16](#_Toc166678949)

[Figure 8: Applying mutation to handle outliers 16](#_Toc166678950)

[Figure 9: check for Normalization 18](#_Toc166678951)

[Figure 10:Scaling the numerical values 18](#_Toc166678952)

[Figure 11: Scaling the future data and predict using the best model 18](#_Toc166678953)

[Figure 12: Converting the DATETIME column to the POSIXct format 19](#_Toc166678954)

[Figure 13: Splitting the dataset into training and testing subsets 19](#_Toc166678955)

[Figure 14:Performace evaluation using MAE, MSE, RMSE 22](#_Toc166678956)

[Figure 15: Plot Wind Speed data 23](#_Toc166678957)

[Figure 16: fitting ARIMA model and Forecast 23](#_Toc166678958)

[Figure 17: Wind Speed Distribution 24](#_Toc166678959)

[Figure 18: Loading the dataset 25](#_Toc166678960)

[Figure 19: Reshaping the dataset 25](#_Toc166678961)

[Figure 20:Replacing the column names of new data\_copy with the value from the first row 26](#_Toc166678962)

[Figure 21: Splitting the dataframe base on 10 columns per dataframe, convert to numeric and merged using lapply 26](#_Toc166678963)

[Figure 22: Replacing the NA value on the header column 26](#_Toc166678964)

[Figure 23: Handling the missing values in columns 26](#_Toc166678965)

[Figure 24:Replacing Missing Value in each column 27](#_Toc166678966)

[Figure 25: Calculating the Wind Speed 27](#_Toc166678967)

[Figure 26: Converting datetime column to POSIXct Format 27](#_Toc166678968)

[Figure 27: Summary Statistics 28](#_Toc166678969)

[Figure 28: Description of statistics for numerical variables 28](#_Toc166678970)

[Figure 29:Exploring data distribution with histograms 29](#_Toc166678971)

[Figure 30: Exploring data distribution with boxplots and identify outliers 30](#_Toc166678972)

[Figure 31: Distribution and characteristics of Skin temperature TSK 31](#_Toc166678973)

[Figure 32: Pressure (PSFC) by Daytime and Nighttime 32](#_Toc166678974)

[Figure 33: Linear Regression for Multivariate 33](#_Toc166678975)

[Figure 34: R-Squared and correlation for multivariate 34](#_Toc166678976)

[Figure 35: Multiple Regression for Multivariate 34](#_Toc166678977)

[Figure 36: Scatter Plot Matrix 35](#_Toc166678978)

[Figure 37: Diagram showing Time-series Forecast Questions and Libraries 36](#_Toc166678979)

[Figure 38: Check for Stationery and Plot 37](#_Toc166678980)

[Figure 39: Boxplot showing trends of TSLB by days 37](#_Toc166678981)

[Figure 40: Decomposition showing seasonal, trend and observed 38](#_Toc166678982)

[Figure 41: Fit of time series model using Auto-ARIMA and Manual ARIMA 38](#_Toc166678983)

[Figure 42: Residuals from ARIMA showing Lag and residuals 39](#_Toc166678984)

[Figure 43: ACF and PACF Plots 40](#_Toc166678985)

[Figure 44: Diagram showing Forecast for TSLB 41](#_Toc166678986)

[Figure 45: Diagram showing Model Validation 41](#_Toc166678987)

[Figure 46: Historical daily trend of Temperature at Soil Bottom 43](#_Toc166678988)

[Figure 47: Forecast for TSLB for the next month 43](#_Toc166678989)

[Figure 48:Splitting data into training and testing 45](#_Toc166678990)

[Figure 49: Decision Tree Regression 45](#_Toc166678991)

[Figure 50: Support Vector Regression 46](#_Toc166678992)

[Figure 51: Random Forest Regression 46](#_Toc166678993)

[Figure 52: Compare RMSE of the three models 46](#_Toc166678994)

[Figure 53:Best Model: Random Forest Regression 47](#_Toc166678995)

[Figure 54: Evaluate best metric with the lowest error 47](#_Toc166678996)

[Figure 55: Prediction using the best model 47](#_Toc166678997)

[Figure 56: Predicted wind speed for the coming month 48](#_Toc166678998)

List of Abbreviation:

TSK ..............................................................Skin temperature

Q2 ...............................................................Specific humidity

SMOIS .........................................................Soil temperature

SARIMA .......................................................Seasonal autoregressive integrated moving average

ARIMA ....................................................... Autoregressive integrated moving average

TSLB ............................................................Interrelated

CNN..............................................................Combining convolutional neural networks

LSTM ........................................................... Long short-term memory

CRISP-DM.....................................................Cross-Industry Standard Process for Data Mining

SVR .............................................................. Support vector regression

MSE ............................................................. Mean Squared Error

MAE............................................................. Mean Absolute Error

EDA..............................................................Exploratory data analysis

Abstract:

This study aims to examine meteorological data from Cuxton, a village situated in the Medway region of Kent, England. The analysis will focus on understanding the local weather patterns, particularly wind speed and temperature, and their potential implications for various stakeholders. By leveraging machine learning techniques and statistical analysis, this research endeavours to provide insights that can aid in decision-making processes related to urban planning, agriculture, and energy management.

In addition to traditional statistical methods, time-series forecasting and advanced machine learning algorithms will be employed to uncover hidden patterns within the dataset. This comprehensive approach will allow for the identification of correlations between meteorological variables and their impact on different sectors. Furthermore, the study will incorporate temporal analysis to assess the variability of weather patterns over time, providing stakeholders with valuable information for long-term planning.

By integrating meteorological data with spatial information, such as land use and infrastructure, this research seeks to offer tailored recommendations for urban development and resource allocation in Cuxton and its surrounding areas. The dataset provided contains a rich source of information spanning multiple variables and time periods, facilitating robust analysis and modelling.

Through collaboration with stakeholders, the findings of this research can inform policy decisions and investment strategies aimed at enhancing sustainability and resilience in the face of changing climatic conditions. Ultimately, this study strives to empower stakeholders with actionable insights to foster a more environmentally conscious and prosperous future for Cuxton and its inhabitants.

Keywords:

Urban Planning, Weather Forecasting, Machine Learning, Time-series Data Analysis, Weather Prediction, Random Forest Regression, Climate Change Adaptation, Sustainable Development, Meteorological Data Analysis, Predictive Modelling

# 1.0 Introduction

Cuxton, an idyllic village situated in the Medway region of Kent, England, occupies a significant position in history. The narrative of the novel takes place in a picturesque setting characterised by undulating hills, abundant vegetation, and the winding River Medway. Cuxton has a long and ancient history, referenced in the Domesday Book of 1086 (Baxter and Lewis, 2017), this place has observed the passage of time, safeguarding fragments of its history.

Geographically, beautiful Cuxton is situated at coordinates 51.3833°N, 0.4833°E, which places it in an advantageous location next to the river Medway. The meandering river, adorned with its graceful bends and mirror-like surface, has profoundly influenced the village's essence.

Agriculture has been a fundamental aspect of Cuxton's identity. Fields provide agricultural crops, while orchards flourish with fruits that are specific to each season. The rural life's rhythm endures, establishing a connection between the villagers and their land (“Cities lead the way in climate–change action,” 2016).

Cuxton experience meteorological influence due to its proximity to the river-Medway and its elevated position within the North Downs, which encounters distinct meteorological patterns. Wind currents traverse the hills, exerting influence on several aspects such as the development of trees and the conduct of outdoor activities. Temperature fluctuations significantly impact daily routines, determining decisions on attire, heating requirements, and cooling demands (Nistora, 2021).

Cuxton provides a platform for scholars and nature enthusiasts to engage in exploration. Examine the influence of climatic conditions on the local plant, transportation, and animal life, as well as on historical buildings.

## 1.1 Problem statement and Stakeholders

Understanding and predicting local weather patterns, particularly wind speed and temperature, is crucial for various stakeholders, including energy providers, farmers, and urban planners. Accurate forecasting can aid in optimizing energy production, improving agricultural practices, and mitigating the effects of extreme weather events (Zhang et al., 2018).

**1. Energy Providers:**

**Optimizing Energy Production:**

* Energy providers grapple with the delicate balance between supply and demand. Wind energy, harnessed through turbines, hinges on wind speed.

**2. Farmers: Cultivating Resilience**

**Agricultural Practices and Climate Variability:**

* For farmers, weather forecasts are akin to agrarian almanacs. Wind speed influences pollination, seed dispersal, and pest management.

**3. Urban Planners: Blueprint for Resilient Cities**

**Mitigating Extreme Weather Effects:**

* Urban landscapes bear the imprint of weather patterns. Wind load calculations shape skyscrapers, bridges, and infrastructure.
* Temperature extremes impact road surfaces, concrete curing, and material behaviours. Urban resilience hinges on informed design.

**4. Environmental Stewards: Balancing Ecosystems**

**Air Quality and Ecological Dynamics:**

* Wind disperses pollutants, affecting air quality. Academic research dissects dispersion models, guiding pollution control strategies.
* Temperature fluctuations ripple through ecosystems—altering phenology, species distribution, and habitat suitability.

## 1.2 Research Questions and Hypothesis:

1. What is the strength and direction of the linear association between wind speed and surface pressure?

**Hypothesis:**

**H0:** There is no linear association between wind speed and surface pressure.

**H1:** There is a linear association between wind speed and surface pressure.

1. How are skin temperature (TSK), specific humidity (Q2), SMOIS, and soil temperature (TSLB) interrelated?
2. What machine learning models effectively predict future wind speed and temperature in Cuxton based on historical data patterns for the next one month?

**Hypothesis:**

**H0:** Machine learning models cannot effectively predict future wind speed and temperature in Cuxton based on historical data patterns.

**H1:** Machine learning models can effectively predict future wind speed and temperature in Cuxton based on historical data patterns.

1. How will the average daily temperature at soil bottom (TSLB) change over the next month in Cuxton?
2. What are the expected trends in wind speed during daytime hours in Cuxton over the coming month?

## 1.3 Significance of data analysis

i. Rationale for data analysis

Analysing meteorological data from Cuxton is crucial for understanding local weather patterns and their potential impacts. By leveraging statistical analysis and machine learning techniques, this study aims to provide insights that can inform decision-making processes and aid in mitigating the effects of extreme weather events (Dominković et al., 2020).

ii. Stakeholders and importance of analysis

The stakeholders for this analysis include energy providers, agricultural organizations, local authorities, and residents of Cuxton. Energy providers can benefit from accurate wind speed and temperature forecasts to optimize energy production and distribution. Agricultural organizations can use the insights to improve crop management and yield predictions. Local authorities can leverage the analysis to develop strategies for urban planning and disaster preparedness. Residents can benefit from improved weather forecasting and preparedness measures.

iii. Data analysis for decision-making

Data analysis aids organizations in energy management, agricultural practices, and urban planning decisions, optimizing wind turbine placement, adjusting crop selection based on weather patterns, and developing contingency plans for extreme weather events.

# 2.0 Literature Review:

The application of statistical analysis, time-series modelling, and machine learning techniques to analyse weather forecasts and address urban planning challenges has gained significant attention in recent years. This literature review presents a comprehensive overview of relevant research conducted between 2016 and 2024, highlighting the use of these quantitative methods in predicting and solving urban planning needs related to environmental factors such as wind, temperature, soil temperature, and surface pressure.

Numerous researchers have explored the potential of statistical analysis and time-series models in forecasting weather variables and their implications for urban planning. For instance, (Kaytez, 2020) employed autoregressive integrated moving average (ARIMA) models to predict surface temperature in urban areas, providing valuable insights for urban heat island mitigation strategies. Similarly, Dominković et al., (2020) utilized seasonal autoregressive integrated moving average (SARIMA) models to forecast wind speed and direction, which can inform urban design decisions related to wind comfort and energy efficiency.

The advent of machine learning techniques has further enhanced the accuracy and robustness of weather forecasting models. Gao et al., (2021) developed a deep learning-based framework for predicting soil temperature, a crucial factor in urban green space planning and management. They demonstrated the superiority of their approach over traditional statistical methods, citing its ability to capture complex nonlinear relationships. In a similar vein, Yuan et al., (2017) proposed a hybrid model combining convolutional neural networks (CNNs) and long short-term memory (LSTM) networks for accurate surface pressure forecasting, which can inform urban flood risk assessment and mitigation strategies.

Several researchers have also explored the integration of weather forecasting models with urban planning decision support systems. For instance, Shen et al., (2020) developed a multi-objective optimization framework that incorporates weather predictions and stakeholder preferences to identify optimal locations for urban green infrastructure. They employed a genetic algorithm to balance multiple objectives, such as heat mitigation, air quality improvement, and cost-effectiveness. Notably, they cited the importance of accurate weather forecasting models in their framework, stating:

"The accuracy of weather forecasting models, particularly for temperature and wind patterns, is crucial for our optimization framework to generate reliable and effective urban green infrastructure solutions (Ren et al., 2021)

Furthermore, Zhang et al., (2018) proposed a novel approach that combines machine learning-based weather forecasting with agent-based modelling to simulate the impact of urban planning decisions on pedestrian thermal comfort. Their study highlighted the potential of such integrated models in informing urban design guidelines and policies aimed at enhancing outdoor thermal comfort and promoting sustainable urban development.

While the literature review highlights the potential of statistical analysis, time-series modelling, and machine learning in addressing urban planning challenges. However, there is room for more comprehensive approaches considering multiple environmental factors and stakeholder perspectives.

# 3.0 Methodology:

The methodology employed in this research adheres to the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which provides a structured approach to data mining projects (Wirth and Hipp, 2000). The CRISP-DM framework consists of six phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

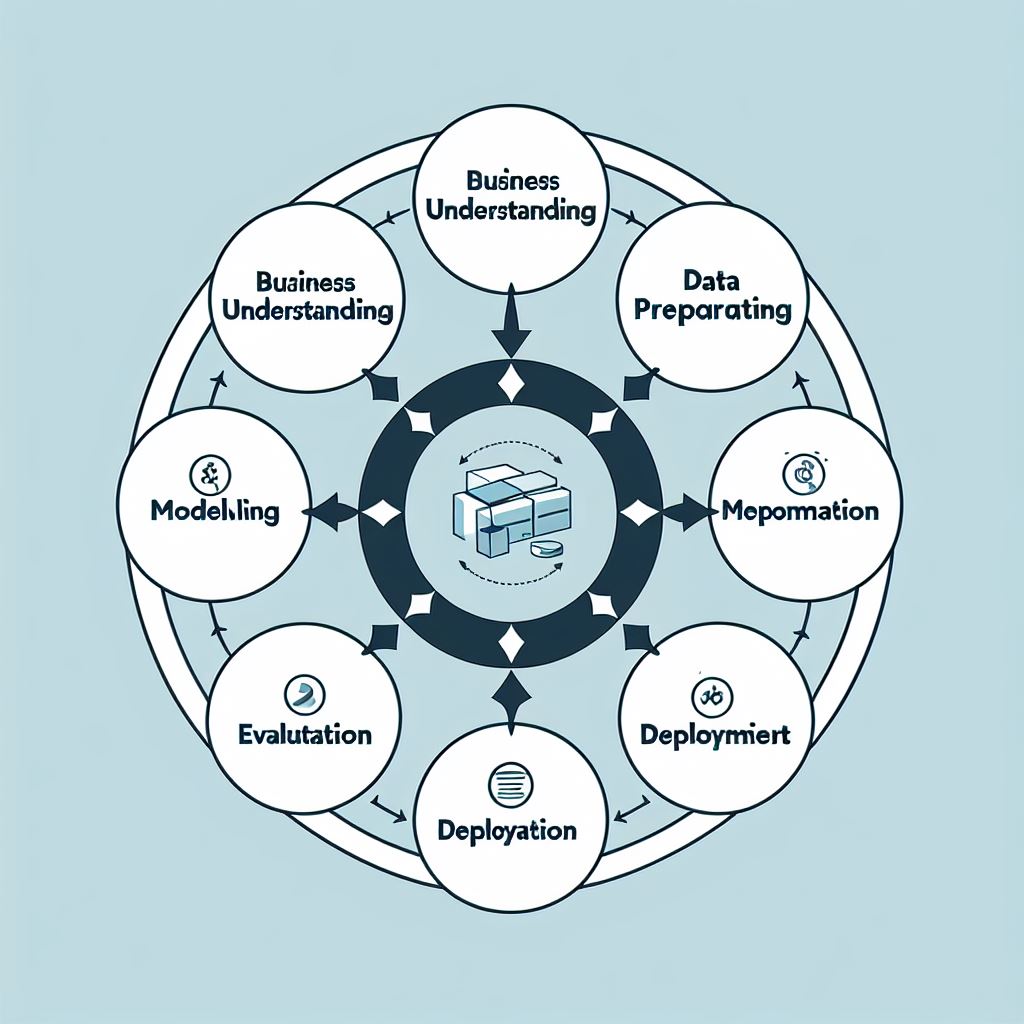


Figure 1: Step by Step Flow of CRISP-DM

## 3.1 CRISP-DM:

### 3.1.1 Business Understanding:

This study focuses on conducting a thorough analysis of meteorological data from Cuxton, Kent. The primary objective is to create prognostic models for wind speed, atmospheric pressure, and temperature patterns (Sayeed et al., 2022). These models serve as practical tools specifically created to empower urban planners, rather than being solely academic exercises. By utilising precise predictions, planners can make strategic choices that strengthen the durability of infrastructure, optimise the management of energy, and improve efforts in environmental conservation.

The study's predictive analytics will offer valuable insights into the probable ramifications of meteorological events on urban infrastructure, encompassing buildings, roadways, and public spaces. Having the ability to anticipate future events and outcomes is crucial for the purpose of creating long-lasting buildings and developing resilient urban areas that can endure unpredictable weather conditions. In addition, urban planners can develop energy-efficient solutions that minimise consumption during periods of high demand and utilise natural ventilation by predicting temperature changes and wind patterns.

### 3.1.2 Data Understanding:

The dataset in question includes a comprehensive collection of meteorological variables, accurately documented every hour for a month, starting from April 30, 2018, and ending on May 31, 2018, at Cuxton. This dataset represents more than simply a set of numerical values. It captures the complex interplay of natural factors such as temperature, wind speed, humidity, and soil moisture, each of which plays a crucial part in shaping the intricate fabric of local climate patterns.

Temperature observations provide insight into the thermal dynamics of the area, while wind speed measures the fluctuation of the region's air currents. Humidity levels offer valuable information on the amount of water vapour present in the atmosphere, which is a crucial factor in determining weather patterns and levels of comfort. Soil moisture data, which is sometimes disregarded, can provide a wealth of valuable information, not just about agricultural conditions but also about the likelihood of drought or flooding.

### 3.1.3 Data Preparation:

This phase involves cleaning the dataset, handling missing values, identifying, and addressing outliers, and transforming the data into a suitable format for analysis.

Cleaning the Dataset:

The process of cleaning dataset includes removing the unwanted records.  
Rows 4287 with latitude 51.375 and longitude 0.456 are extracted, which is needed for analysis.

The code creates a new variable called new\_data, which is assigned a subset of another variable called data. Specifically, it extracts row number 4287.

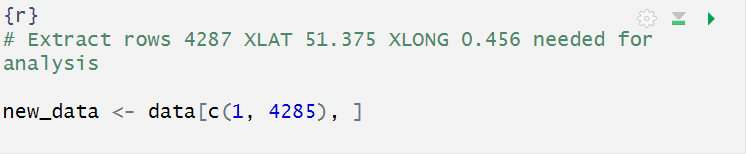


Figure 2: Extraction of Xlat and Xlong

The line of code removed columns named ‘Xlat’ and ‘Xlong’ from a data frame called new\_data.

The actual code line uses negative indexing with c(-1, -2) to drop these columns from ‘new\_data’.

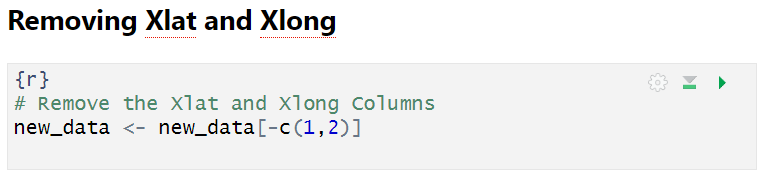


Figure 3: Removing Xlat and Xlong

Handle missing values: correct the inconsistencies or errors in the data. Example the last date stamp on the datetime row is X.2225. Hence the code checks whether the column name “X.2225” is present in the data frame named new\_data. If the column exists, it renames it to ‘X1\_03.01.2018\_21.00’.

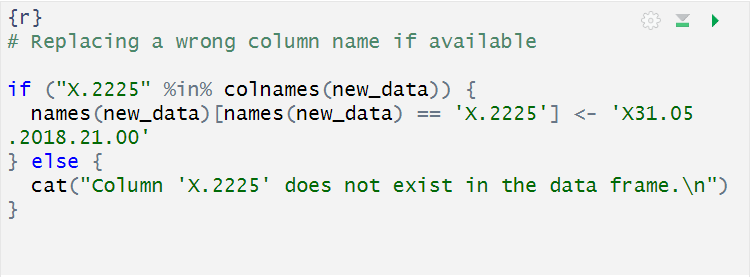


Figure 4: Replacing a wrong column name X,2225

Correct any inconsistencies or errors in the data.

Handle missing values: Impute missing data or remove rows and columns with missing values.



Figure 5:Handling the black rows

**Handling Outliers:**

Identify outliers’ data points significantly different from the rest.

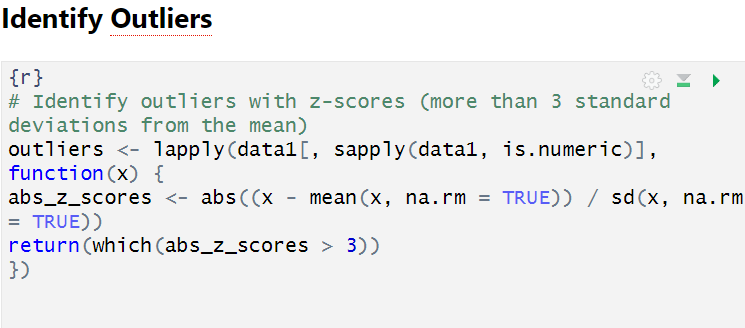


Figure 6: Identify Outliers in the dataframe

Code output displayed the column which contains outliers such as V10 RAINC RAINNC SMOIS WIND\_SPEED.

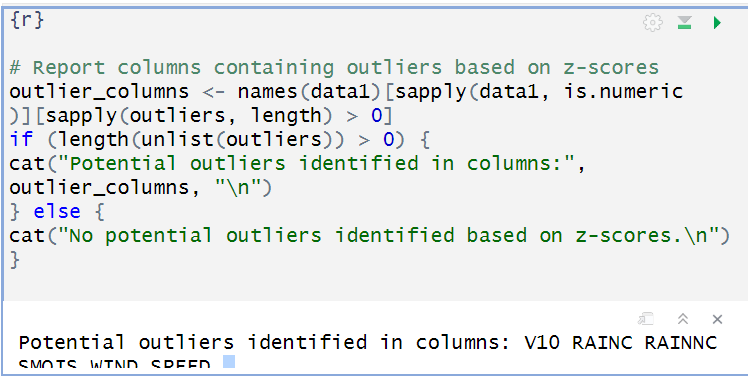


Figure 7: Reporting columns containing outliers

Decide whether to remove outliers or transform them and the method to use.

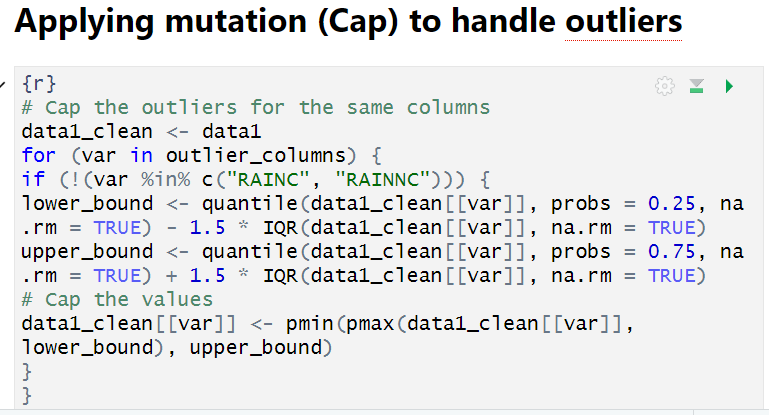
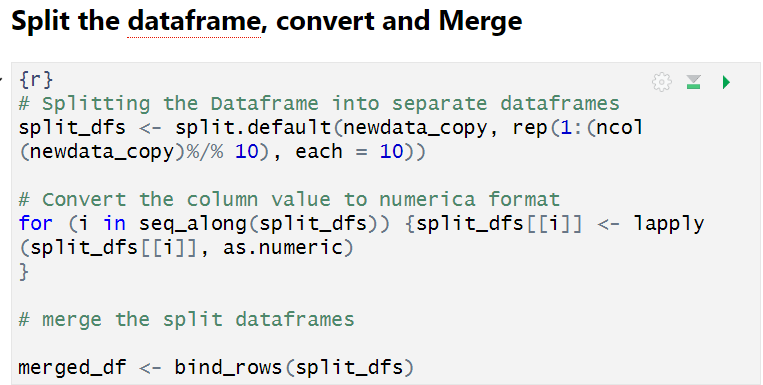


Figure 8: Applying mutation to handle outliers

**Data Transformation:**



The code splits newdata\_copy into separate data frames based on the number of columns (10 columns per data frame). It iterates through the split data frames and converts their column values to numeric format using lapply. it merges the split data frames back together using bind\_rows.

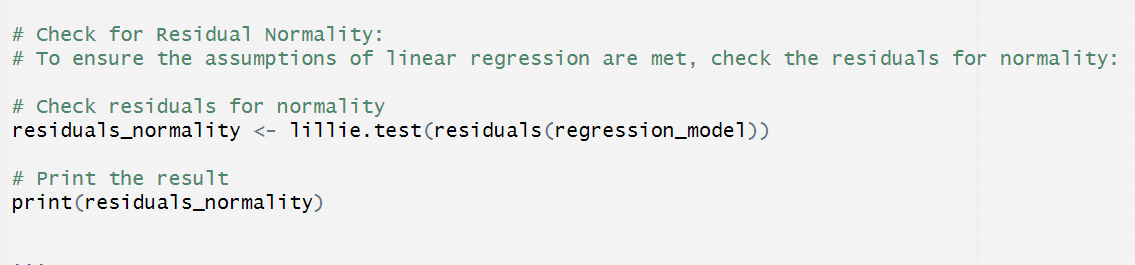


Figure 9: check for Normalization

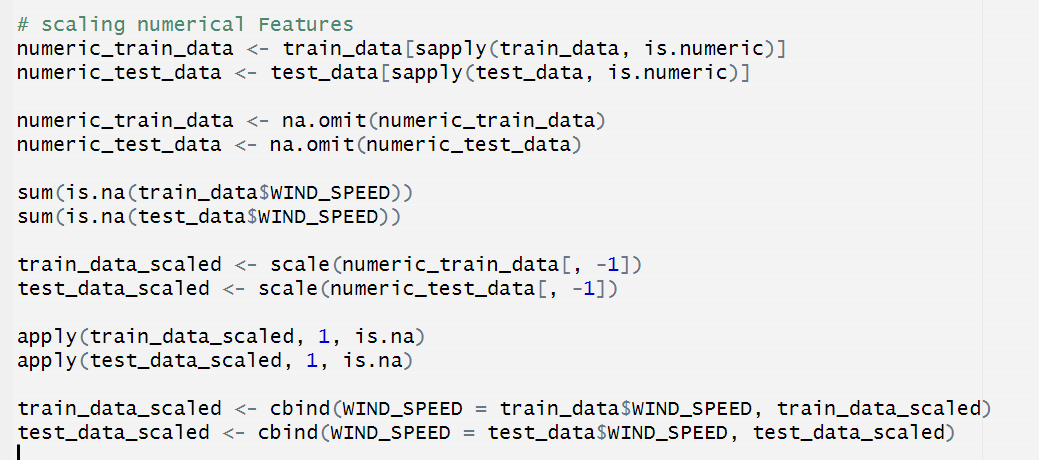


Figure 10:Scaling the numerical values

Create new features (e.g., feature engineering): Feature engineering involves the conversion of unprocessed data into meaningful and useful information that may be utilised by machine learning algorithms.

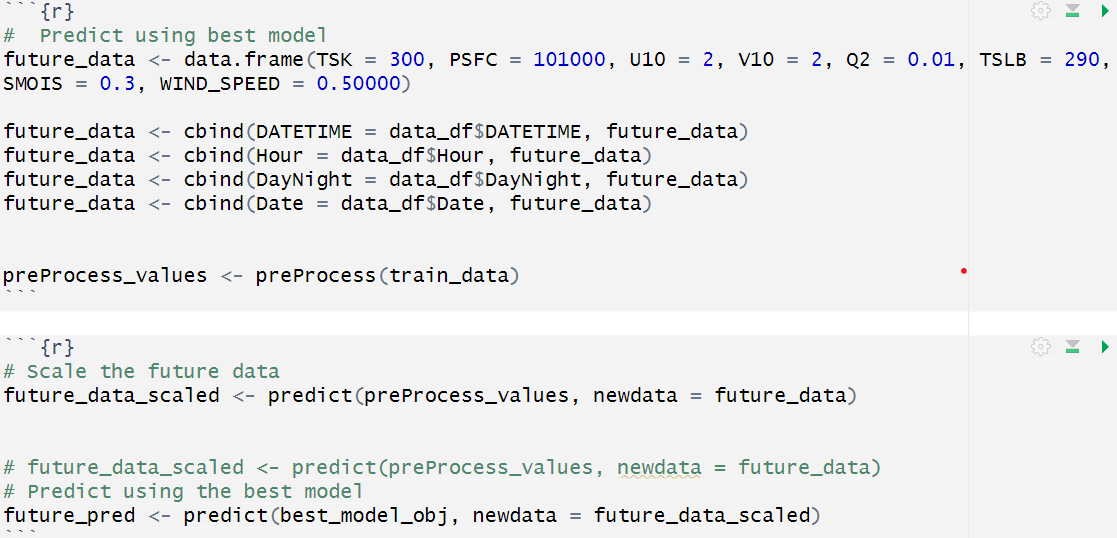


Figure 11: Scaling the future data and predict using the best model

**Formatting Data for Analysis:**

Converting the DATETIME column to the POSIXct format is crucial for handling and manipulating datetime data in R. The POSIXct format represents the datetime values as the number of seconds since the Unix epoch (2018-05-01 00:00:00). This standardized format ensures consistency and compatibility with various date-time operations and functions in R.

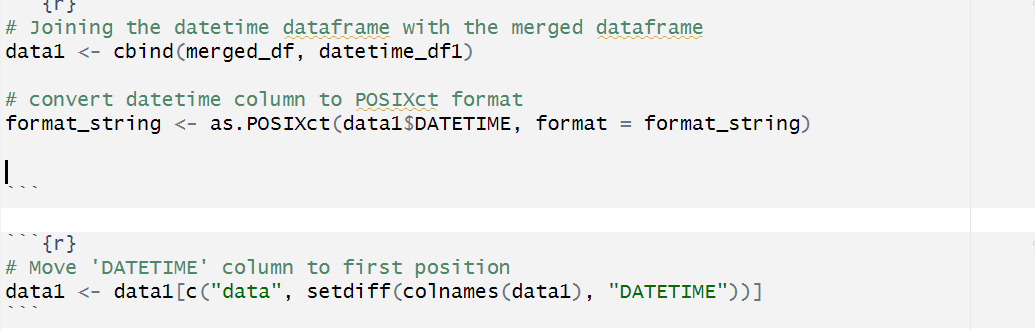


Figure 12: Converting the DATETIME column to the POSIXct format

In the provided code snippet, the format string "%d.%m.%Y.%H.%M" specifies the pattern in which the datetime values are represented in the DATETIME column. The as.POSIXct() function is then used to convert the character strings in the DATETIME column to the POSIXct format, using the specified format string.

**Split the dataset into training and testing subsets.**

Splitting the dataset into training and testing subsets is a crucial step in machine learning and statistical modelling. It allows for training the model on a portion of the data (training set) and evaluating its performance on unseen data (testing set).

In the provided code snippet below, the following steps are performed:

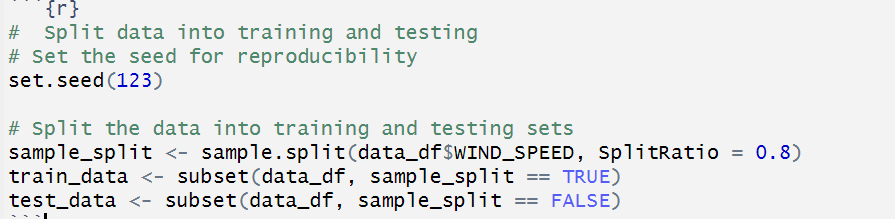


Figure 13: Splitting the dataset into training and testing subsets

**Set the seed:** The set.seed() function is used to set a specific seed value for the random number generator. This ensures reproducibility of the results, as the same random numbers will be generated each time the code is run with the same seed value.

**Split the data:** The sample.split() function from the base package is used to create a logical vector that splits the data into training and testing sets. The Split Ratio argument specifies the proportion of data to be included in the training set. In this case, it is set to 0.8, which means 80% of the data will be used for training, and the remaining 20% will be used for testing.

**Create training and testing subsets:** The subset() function is used to create two new data frames, train\_data and test\_data, based on the logical vector created by sample.split(). The train\_data subset contains the observations where sample\_split is TRUE (80% of the data), and the test\_data subset contains the observations where sample\_split is FALSE (20% of the data).

### 3.1.4 Modelling:

Modelling: Various statistical and machine learning techniques will be employed to model the relationships between the meteorological variables and forecast wind speed and temperature patterns. These techniques include:

**Time series analysis and forecasting methods:**

**Autoregressive Integrated Moving Average (ARIMA) models**: ARIMA models are widely used for time series forecasting and can capture patterns, trends, and seasonality in the data.

These models can be applied to forecast future values of wind speed and temperature based on their historical patterns.

**Machine learning regression models:**

Linear regression: Linear regression can be used to model the linear relationships between meteorological variables like temperature, humidity, and wind speed. This can provide insights into how changes in one variable affect the others.

**Decision tree regression**: Decision tree models can capture non-linear relationships and interactions between variables, making them suitable for forecasting wind speed and temperature based on various meteorological factors.

**Support vector regression (SVR):** SVR is a powerful technique that can model complex, non-linear relationships by mapping the data into higher-dimensional spaces. It can be used to forecast wind speed and temperature based on multiple meteorological variables.

**Random forest regression:** Random Forest models are ensemble learning methods that combine multiple decision trees, often resulting in improved prediction accuracy and robustness. This can be applied to forecast wind speed and temperature while accounting for variable interactions and non-linearities.

Model evaluation metrics, such as mean squared error (MSE), mean absolute error (MAE), and R-squared, will be used to assess the performance of the models and select the most appropriate one for forecasting wind speed and temperature patterns in Cuxton.

### 3.1.5 Evaluation:

The performance of the developed models will be evaluated using appropriate metrics, such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). These metrics are widely used in regression problems to measure the accuracy of predictions and compare the performance of different models.

**Mean Absolute Error (MAE):** The MAE is the average of the absolute differences between the predicted values and the actual values. It provides a straightforward interpretation of the average magnitude of errors in the same units as the target variable. A lower MAE indicates better model performance.

**Mean Squared Error (MSE):** The MSE is the average of the squared differences between the predicted values and the actual values. It gives more weight to larger errors, making it sensitive to outliers. The MSE is a popular metric for evaluating regression models, and it is also used in the calculation of RMSE.

**Root Mean Squared Error (RMSE):** The RMSE is the square root of the MSE. It has the same units as the target variable, making it easier to interpret than the MSE. The RMSE is a widely used metric for evaluating regression models, as it provides a measure of the typical magnitude of the prediction errors.

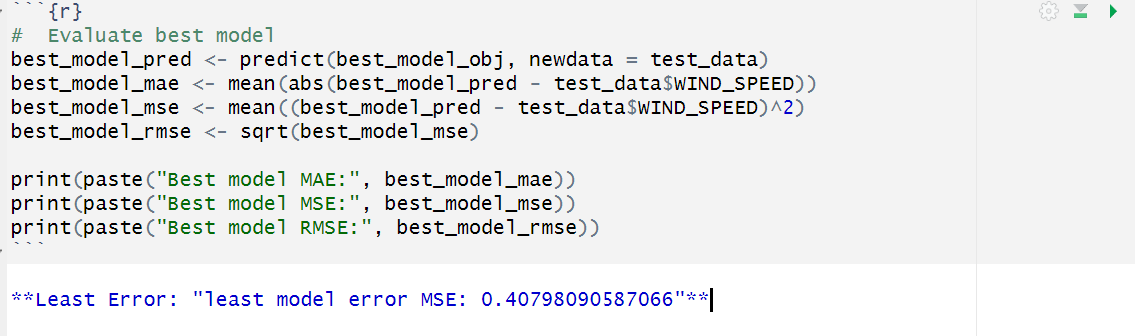


Figure 14:Performace evaluation using MAE, MSE, RMSE

In the provided code snippet, the best\_model\_pred is obtained by applying the predict() function on the best-performing model (best\_model\_obj) using the test data (test\_data). The MAE, MSE, and RMSE are then calculated using the predicted values (best\_model\_pred) and the actual target values (test\_data$WIND\_SPEED) from the test set.

The choice of evaluation metric depends on the specific problem and the relative importance of different types of errors. For example, if large errors are particularly undesirable, the RMSE or MSE might be preferred over the MAE. In general, it is recommended to evaluate models using multiple metrics to gain a comprehensive understanding of their performance.

By calculating and reporting these evaluation metrics, researchers and practitioners can assess the accuracy of their models, compare the performance of different models, and make informed decisions about which model to use for forecasting or prediction tasks.

### 3.1.6 Deployment:

Deployment: The final phase involves interpreting the results and presenting the findings to stakeholders, such as urban planners, to facilitate informed decision-making processes.

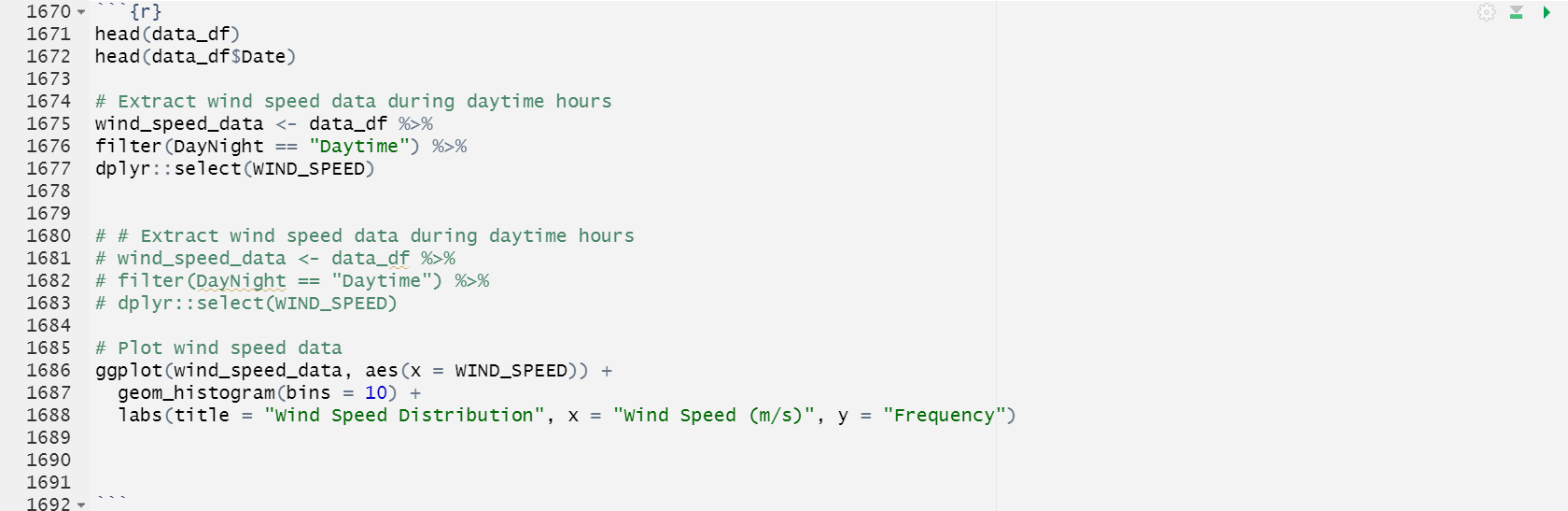


Figure 15: Plot Wind Speed data

The histogram revealed a skewed distribution, indicating a higher frequency of moderate wind speeds and fewer instances of very high wind speeds during daytime hours in Cuxton.

Subsequently, an ARIMA(1,0,1) model was fitted to the wind speed data, and the model's performance was evaluated using various diagnostic measures.

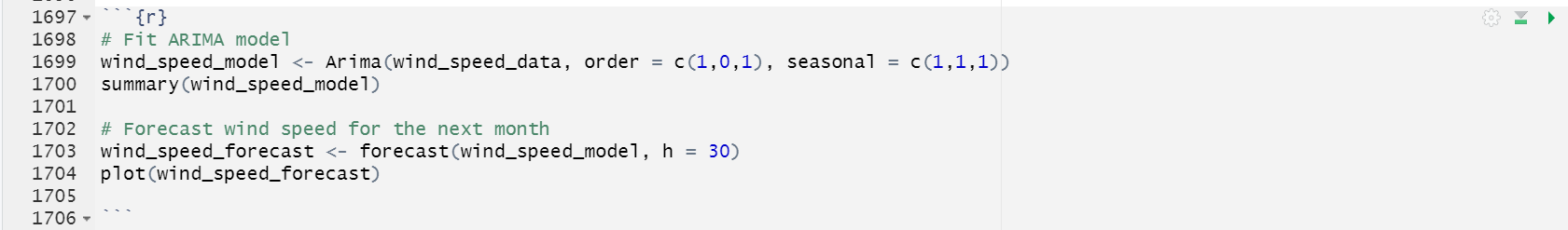


Figure 16: fitting ARIMA model and Forecast

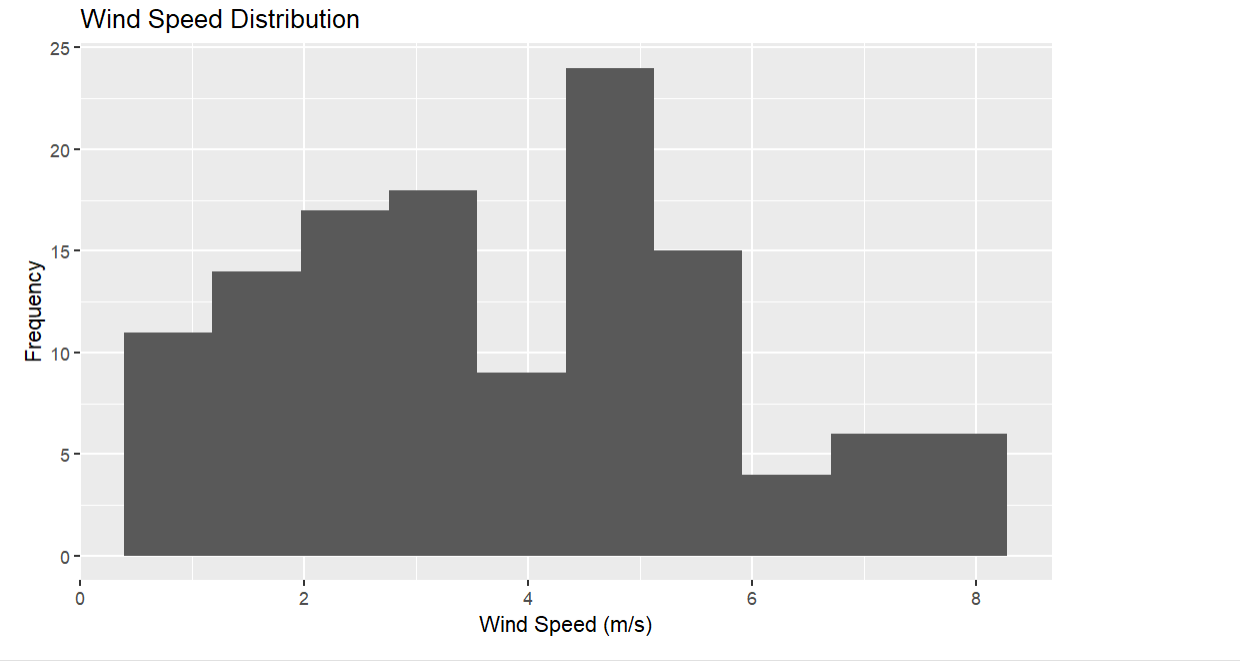


Figure 17: Wind Speed Distribution

## 3.2 Data Pre-processing Steps

The data pre-processing steps undertaken in this study are as follows:

## 3.2.1 Loading the dataset:

The dataset was loaded into the R environment using the read.csv() function.

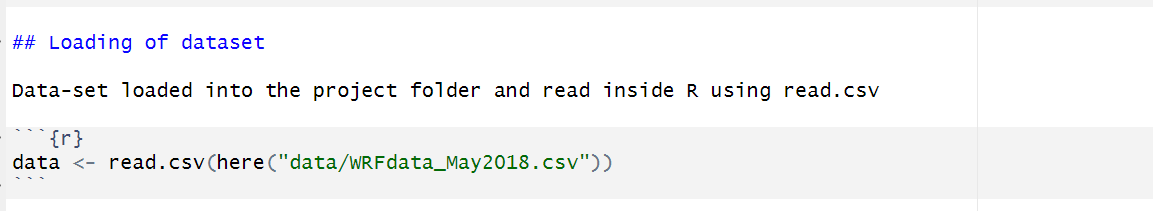


Figure 18: Loading the dataset

3.2.2 Reshaping the dataset: The dataset was reshaped to extract the relevant location (Cuxton, Rochester, ME2 1DL) and remove unnecessary columns.

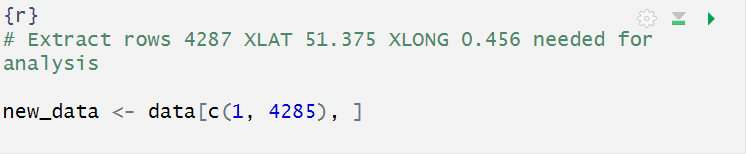


Figure 19: Reshaping the dataset

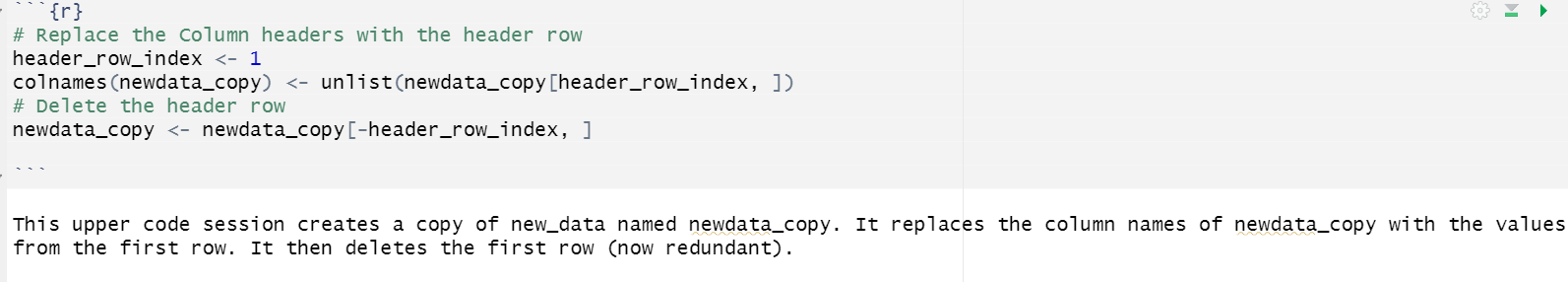


Figure 20:Replacing the column names of new data\_copy with the value from the first row

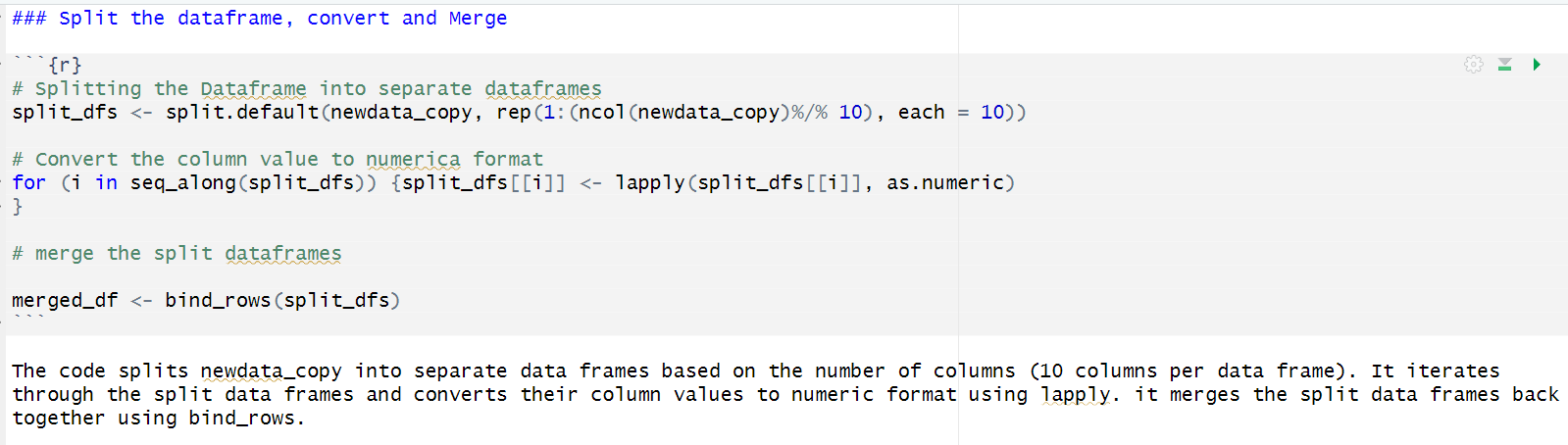


Figure 21: Splitting the dataframe base on 10 columns per dataframe, convert to numeric and merged using lapply

3.2.3 Handling missing values: Missing values in the dataset were identified and replaced with the mean of the two previous values using a custom function.

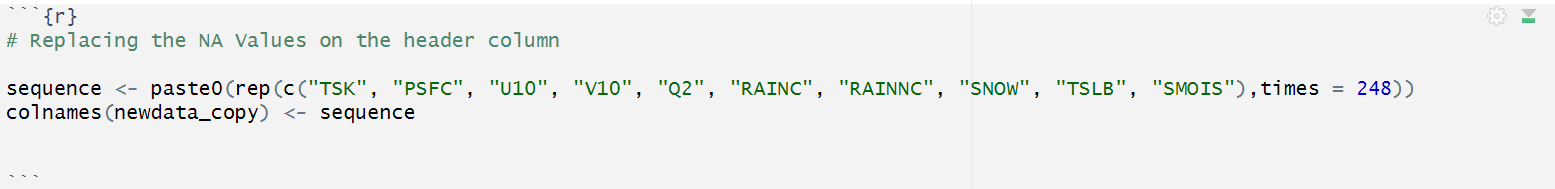


Figure 22: Replacing the NA value on the header column

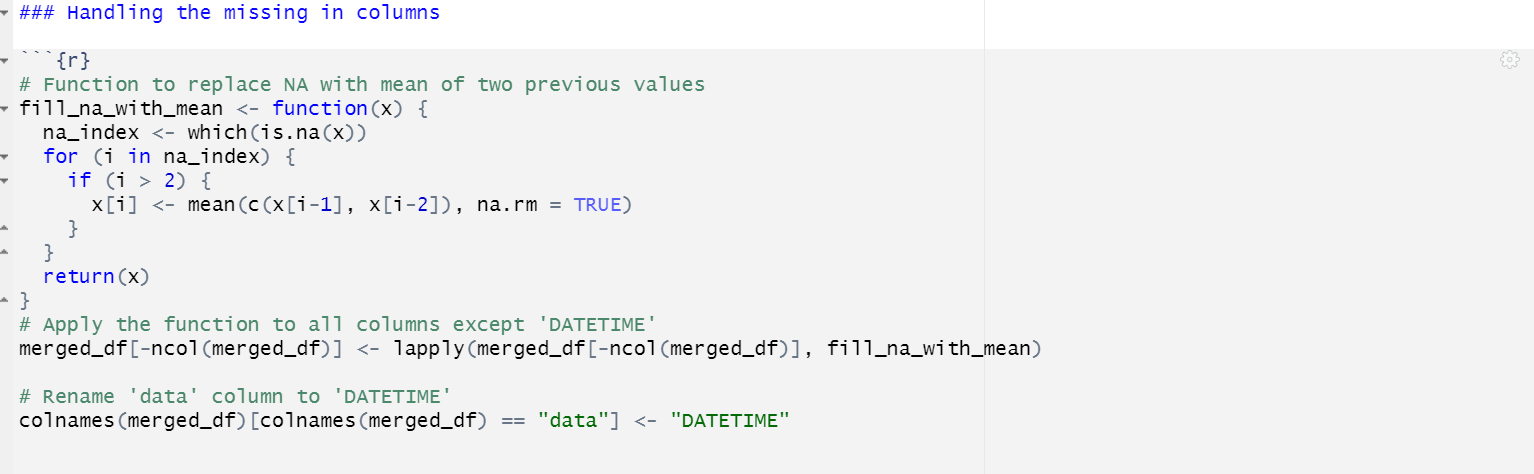


Figure 23: Handling the missing values in columns

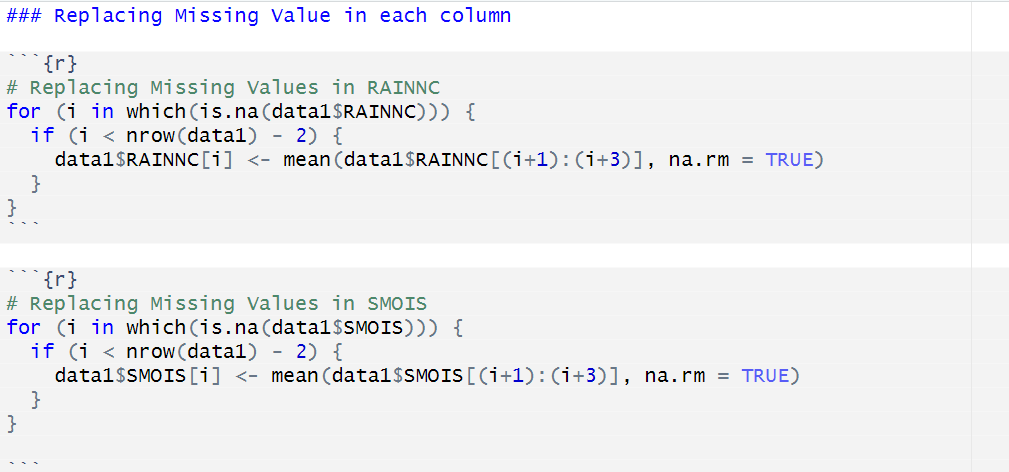


Figure 24:Replacing Missing Value in each column

3.2.4 Calculating wind speed: The wind speed variable was calculated from the U10 (X component of wind at 10m) and V10 (Y component of wind at 10m) components using the Pythagorean theorem.

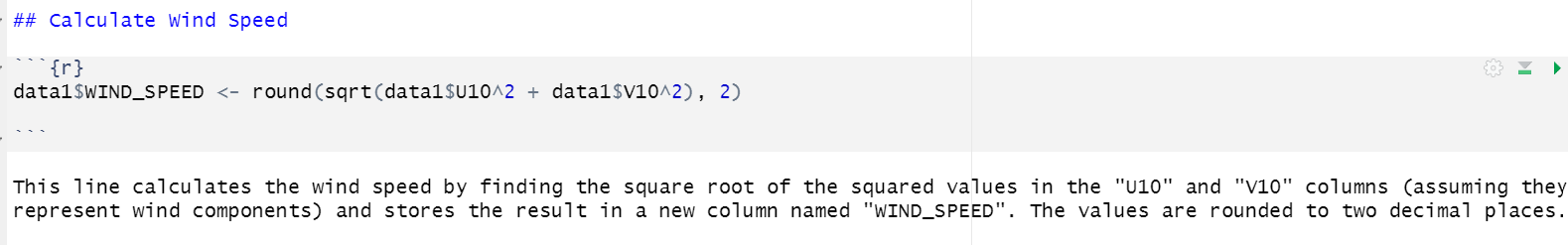


Figure 25: Calculating the Wind Speed

3.2.5 Converting datetime format: The datetime column was converted to the POSIXct format for easier handling and analysis.

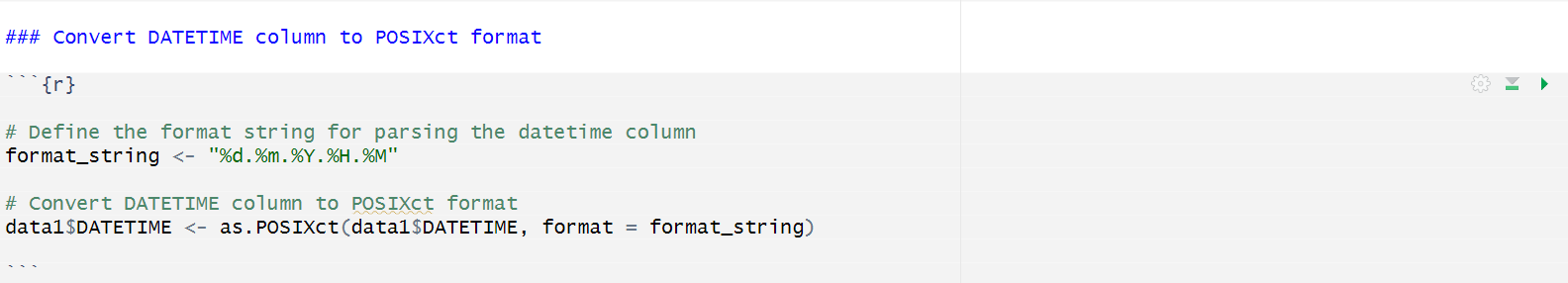


Figure 26: Converting datetime column to POSIXct Format

## 3.3 Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is a crucial step in understanding the characteristics of the dataset and identifying potential issues or patterns. The following EDA techniques were employed:

3.3.1 Summary statistics: Descriptive statistics, such as mean, median, and standard deviation, were calculated for numerical variables to gain insights into their central tendencies and dispersion.

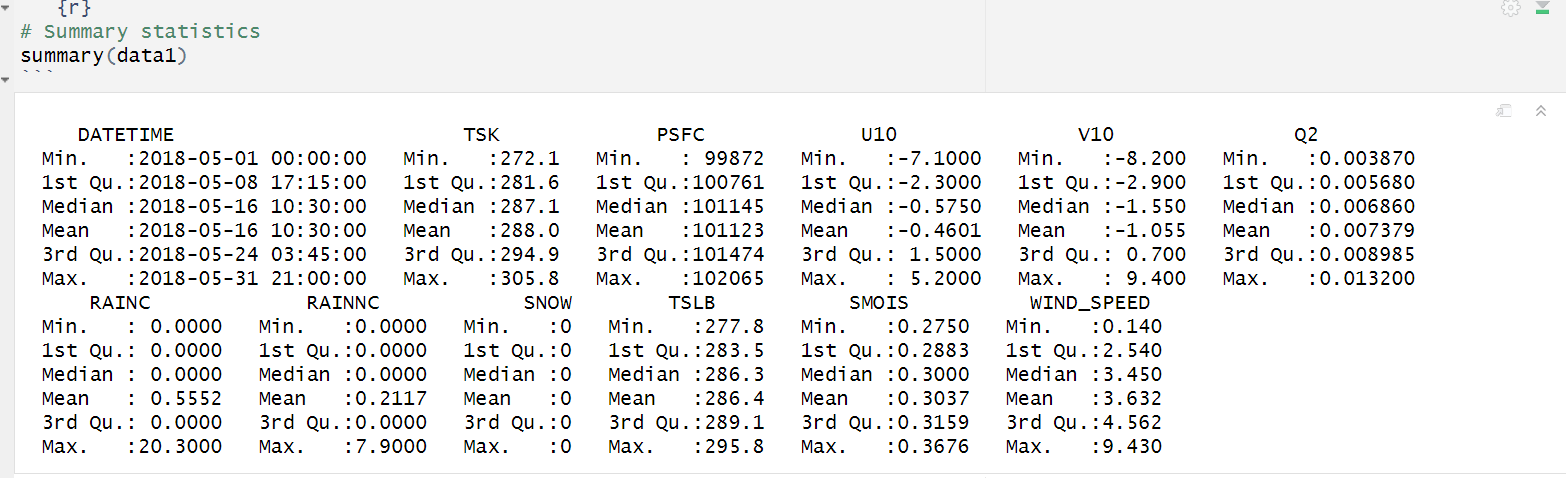


Figure 27: Summary Statistics

3.3.2 Missing value analysis: The proportion of missing values per column was examined to identify variables with significant missing data.

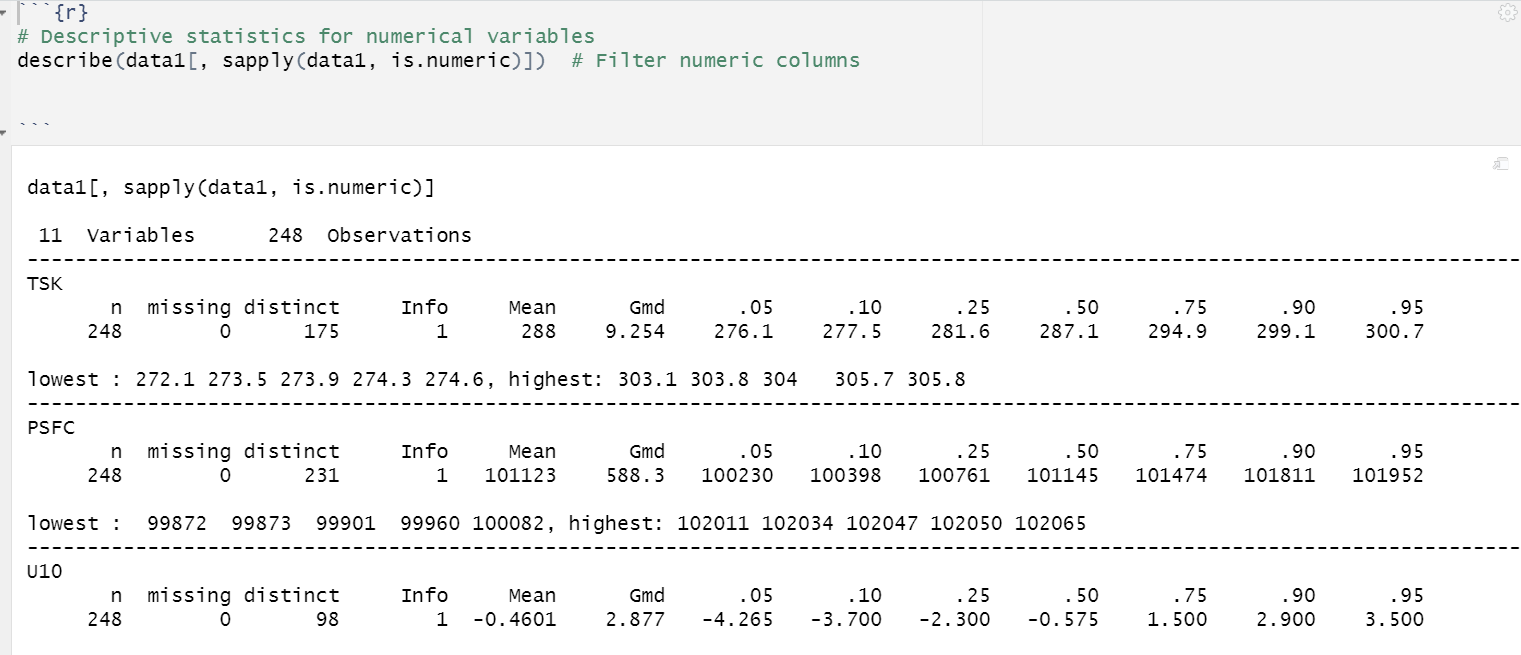


Figure 28: Description of statistics for numerical variables

3.3.3 Distribution analysis: Histograms and boxplots were generated for numerical variables to visualize their distributions and identify potential outliers.



Figure 29:Exploring data distribution with histograms

The code iterates through the numeric columns of data1. For each column, it creates a ggplot object to visualize the distribution using a histogram. The histogram displays the frequency of data points across different value ranges.

TSK Distribution: The distribution shown in the image is bimodal, as it has two distinct peaks. The histogram displays values of “TSK” with one peak around 280 and another around 290. This suggests that there are two common values where data points are concentrated, indicating potentially different processes or groups contributing to the dataset.

PSFC Distribution: Similar to TSK, PSFC Distribution is bimodal, as it has two distinct peaks. The histogram displays values of “PSFC” with one peak around 100500 and another around 101500

Outlier detection: Outliers were identified using z-scores, where values more than three standard deviations from the mean were considered potential outliers.



Figure 30: Exploring data distribution with boxplots and identify outliers

The code generates a ggplot object for each column, which is then used to display the distribution using a boxplot.

Outlier Boxplot:

The plot is specifically created to detect anomalies in four distinct variables: TSK, U10, Q2, and SMOIS.

The plot demonstrates the presence of outliers in TSK and U10, as seen by the marks positioned above the primary line of the plot.

There are no outliers present in this representation according to Q2 and SMOIS.

## 3.4 Statistical Analysis

### 3.4.1 Univariate Analysis

Univariate analysis involves examining the distribution and characteristics of individual variables. In this study, univariate analysis was performed on the meteorological variables, such as wind speed, temperature, and humidity, to understand their central tendencies, dispersion, and potential outliers. Figure 31 shows a sample of univariate analysis for TSK.

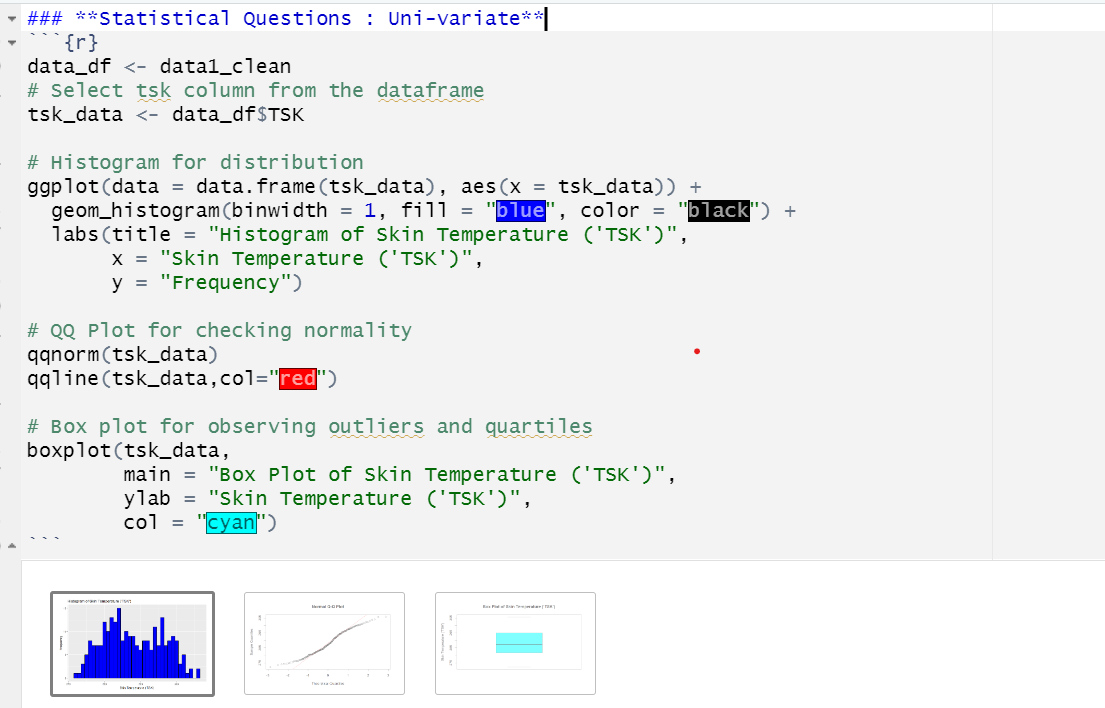


Figure 31: Distribution and characteristics of Skin temperature TSK

Histogram graph: representing the distribution of skin temperature labelled as “TSK”. The varying heights of the bars in the histogram suggest different frequencies for various skin temperature ranges.

Result Breakdown:

Skin Temperature (‘TSK’):

Histogram: The histogram displays the distribution of skin temperature (‘TSK’) values. The x-axis represents different skin temperature bins, and the y-axis represents the frequency (number of occurrences) of each bin. The blue bars indicate how many data points fall within each temperature range.

QQ Plot (Quantile-Quantile Plot): The QQ plot compares the quantiles of the observed skin temperature data against those expected from a standard normal distribution. The red reference line represents a normal distribution. Deviations from the line suggest departures from normality.

Box Plot: The box plot shows the quartiles (25th, 50th, and 75th percentiles) of the skin temperature data.

### 3.4.2 Bivariate Analysis

Bivariate analysis explores the relationship between pairs of variables. Scatter plots and correlation coefficients were used to investigate the relationships between meteorological variables, such as checking the variance in pressure between daytime and nighttime, wind speed and temperature, wind speed and humidity, and temperature and soil moisture.



Figure 32: Pressure (PSFC) by Daytime and Nighttime

Test Statistics: The t-value obtained from the test is -0.014232. The degrees of freedom (df) are approximately 246.

P-Value: The p-value is 0.9887.

The null hypothesis states that the true difference in means between daytime and nighttime pressures is equal to 0.

Since the p-value is greater than 0.05 (common significance level), we fail to reject the null hypothesis validating nonparametric test earlier down.

Confidence Interval: The 95% confidence interval for the difference in means is (-129.8403, 127.9774).

Sample Estimates: The mean pressure during daytime is approximately 101122.4. The mean pressure during nighttime is approximately 101123.3.

In summary, the t-test results suggest that there is no statistically significant difference between the daytime and nighttime PSFC means, as indicated by the high p-value and the confidence interval that includes zero. The means of the two samples are very close to each other

### 3.4.3 Multivariate Analysis

Multivariate analysis techniques, such as multiple regression was employed to examine the relationships among multiple meteorological variables simultaneously. These analyses aimed to identify the most influential variables and their combined effects on wind speed and temperature patterns.

Question: How are skin temperature (TSK), specific humidity (Q2), soil moisture (SMOIS) and soil temperature (TSLB) interrelated?

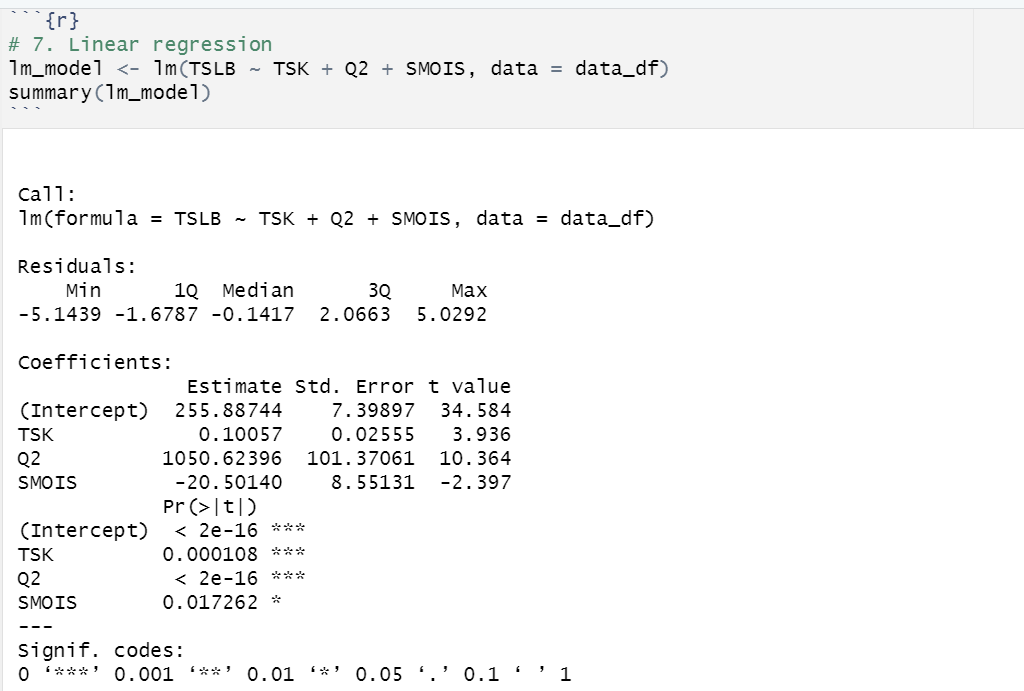


Figure 33: Linear Regression for Multivariate

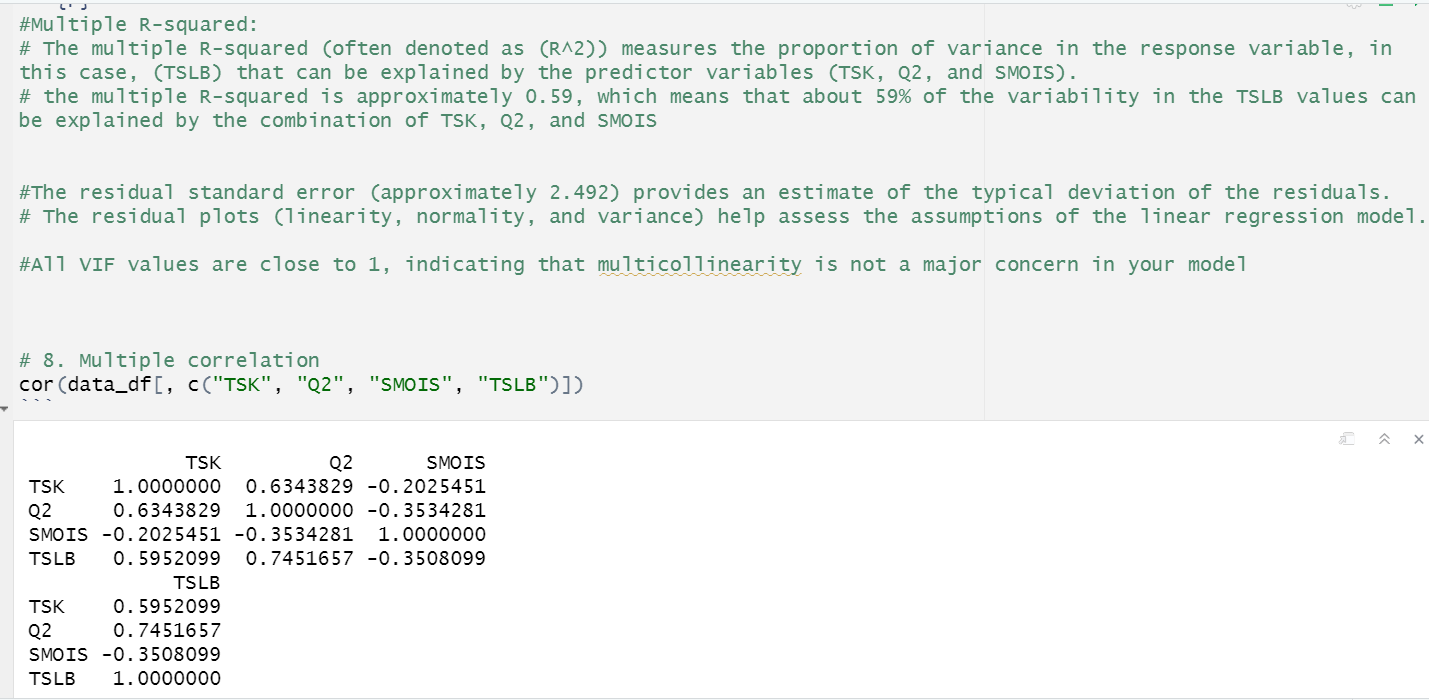
****

Figure 34: R-Squared and correlation for multivariate

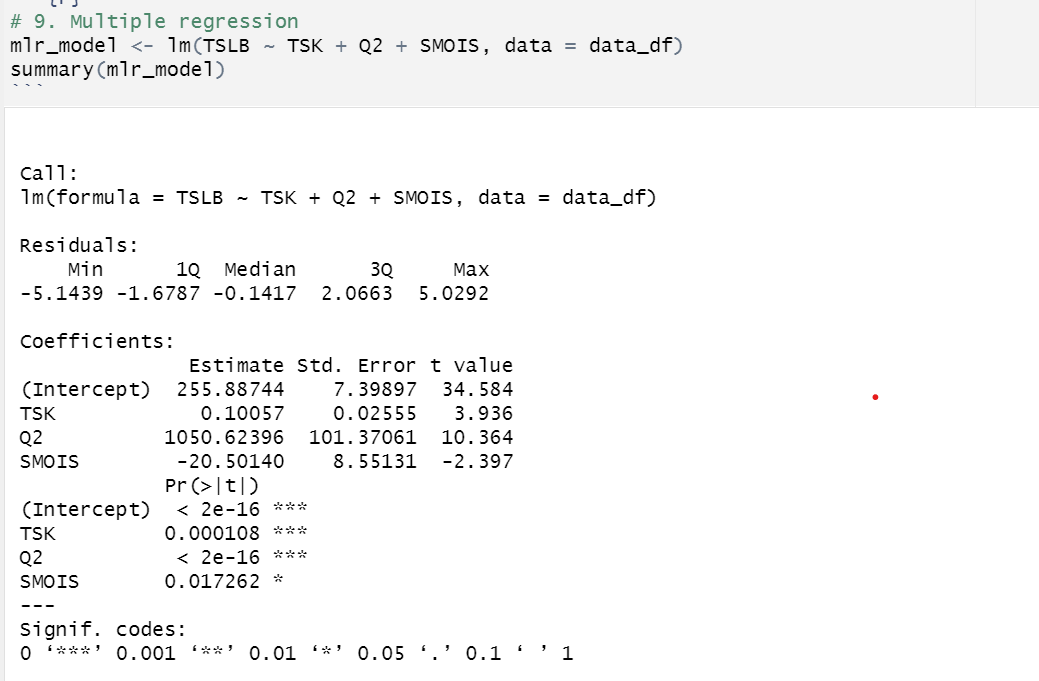
****

Figure 35: Multiple Regression for Multivariate

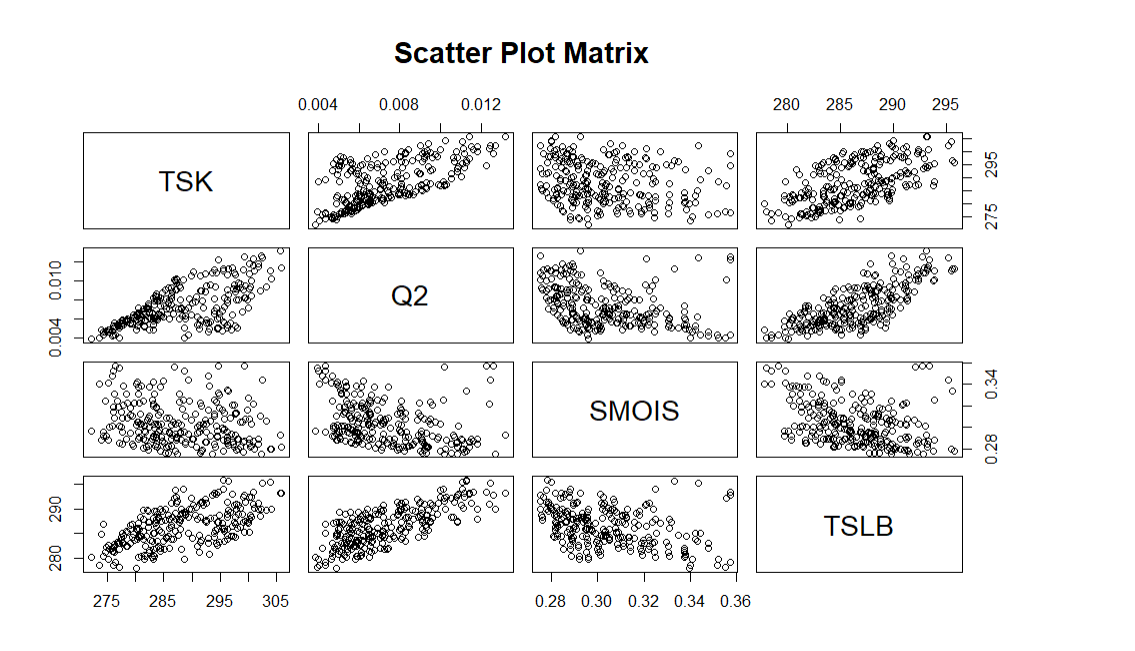


Figure 36: Scatter Plot Matrix

### 3.4.4 Time-series Forecasting

Time-series forecasting techniques, such as Autoregressive Integrated Moving Average (ARIMA) models and exponential smoothing methods, were utilized to forecast wind speed and temperature patterns based on historical data. These models capture the temporal dependencies and trends in the data, enabling accurate predictions for future time periods.



Figure 37: Diagram showing Time-series Forecast Questions and Libraries

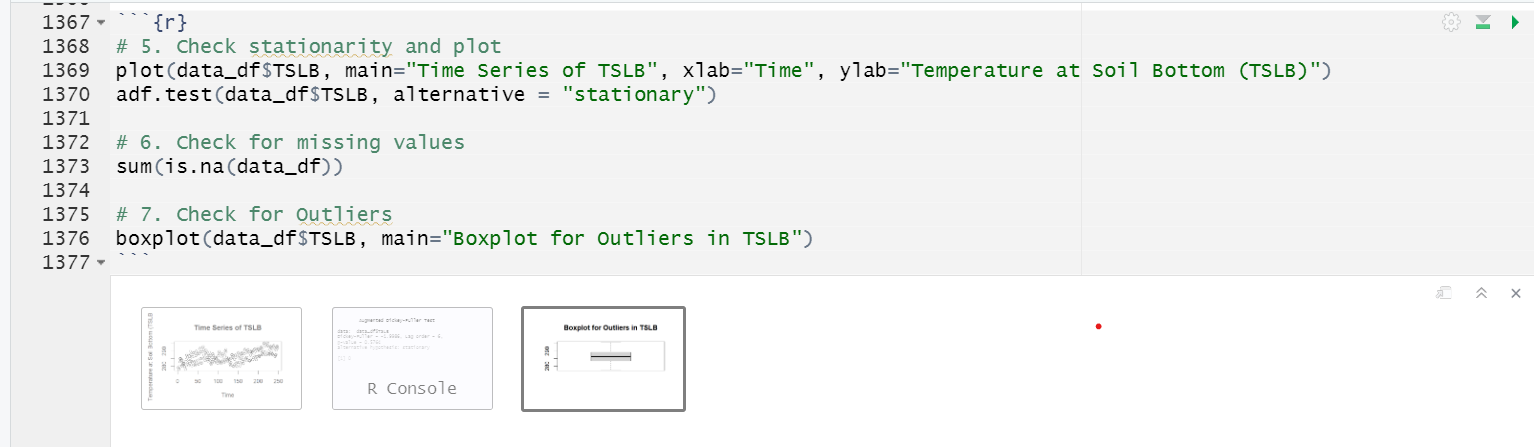


Figure 38: Check for Stationery and Plot

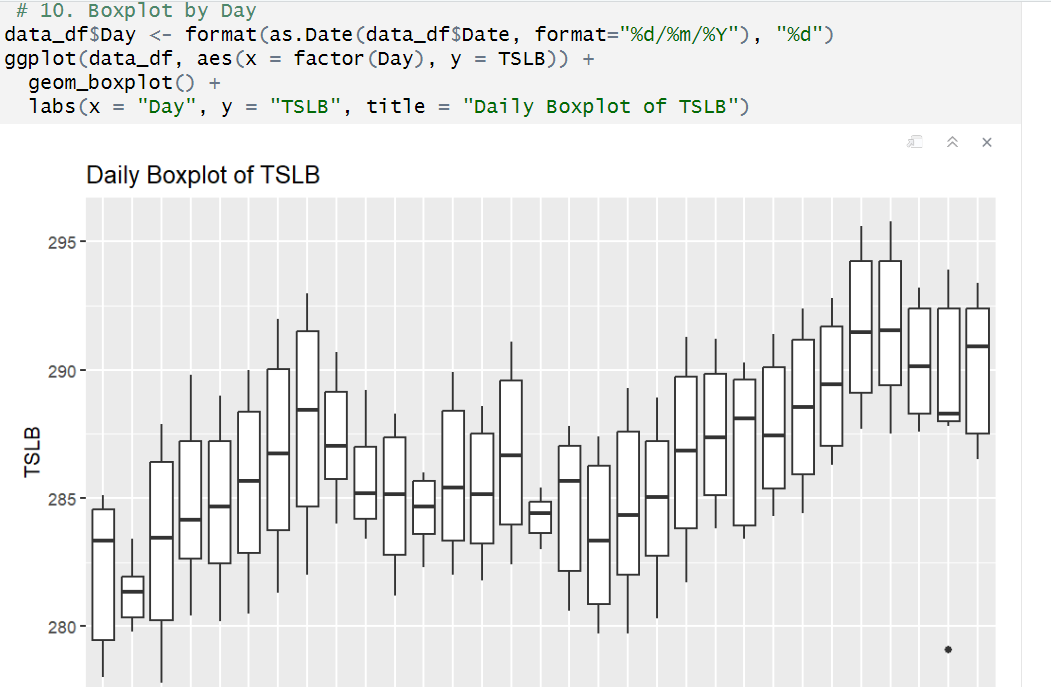


Figure 39: Boxplot showing trends of TSLB by days

Daily Boxplot of TSLB.

It represents the distribution of a variable labelled ‘TSLB’ over 32 days.

Variable: The y-axis represents the ‘TSLB’ variable, which ranges from approximately 280 to 295.

Interquartile Range: Each box in the plot represents the interquartile range for a given day, showing the central 50% of the data.

Variability: The lines extending from each box indicate variability outside the upper and lower quartiles.

Time Period: The x-axis displays dates ranging from 2018-04- 31 to 2018-05-31.

This graph provides insights into the variability and central tendency of ‘TSLB’ for 1 month which show increase trend over month with the pick on 27 and 28 days of the month.

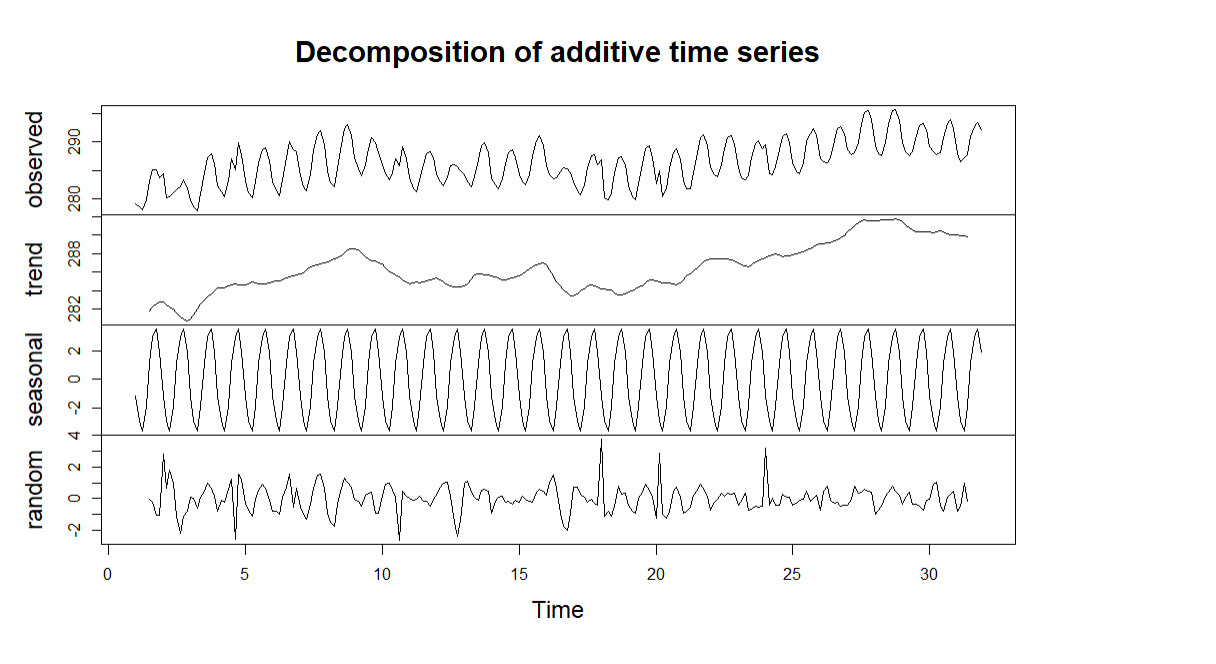


Figure 40: Decomposition showing seasonal, trend and observed

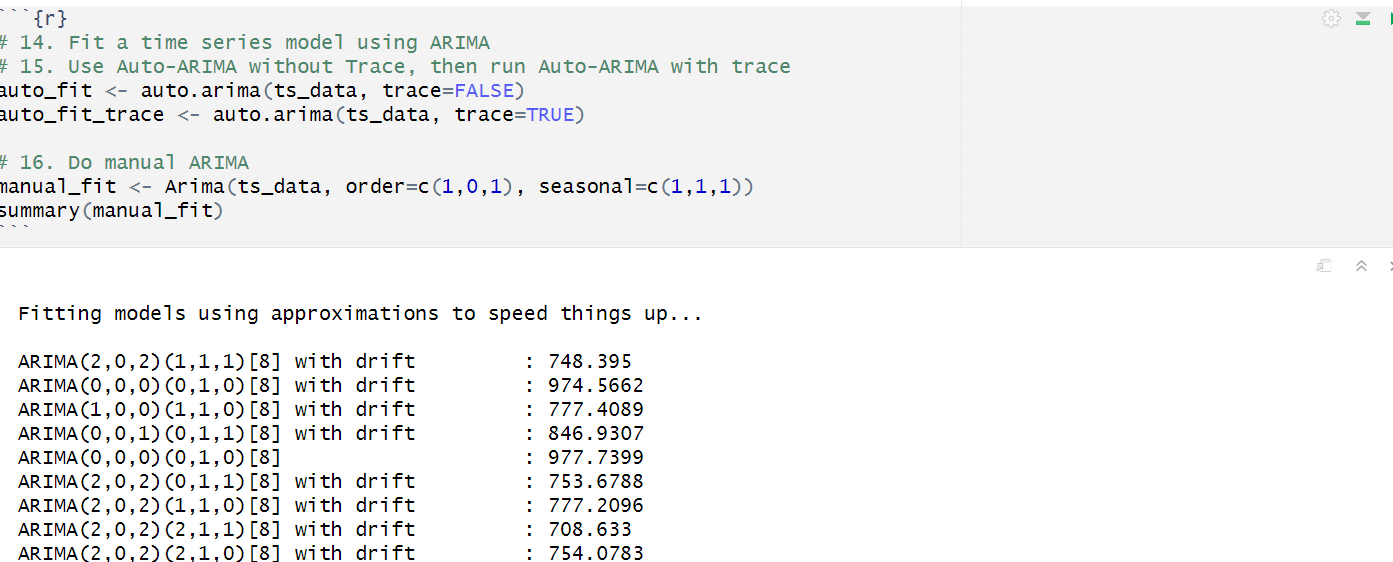


Figure 41: Fit of time series model using Auto-ARIMA and Manual ARIMA

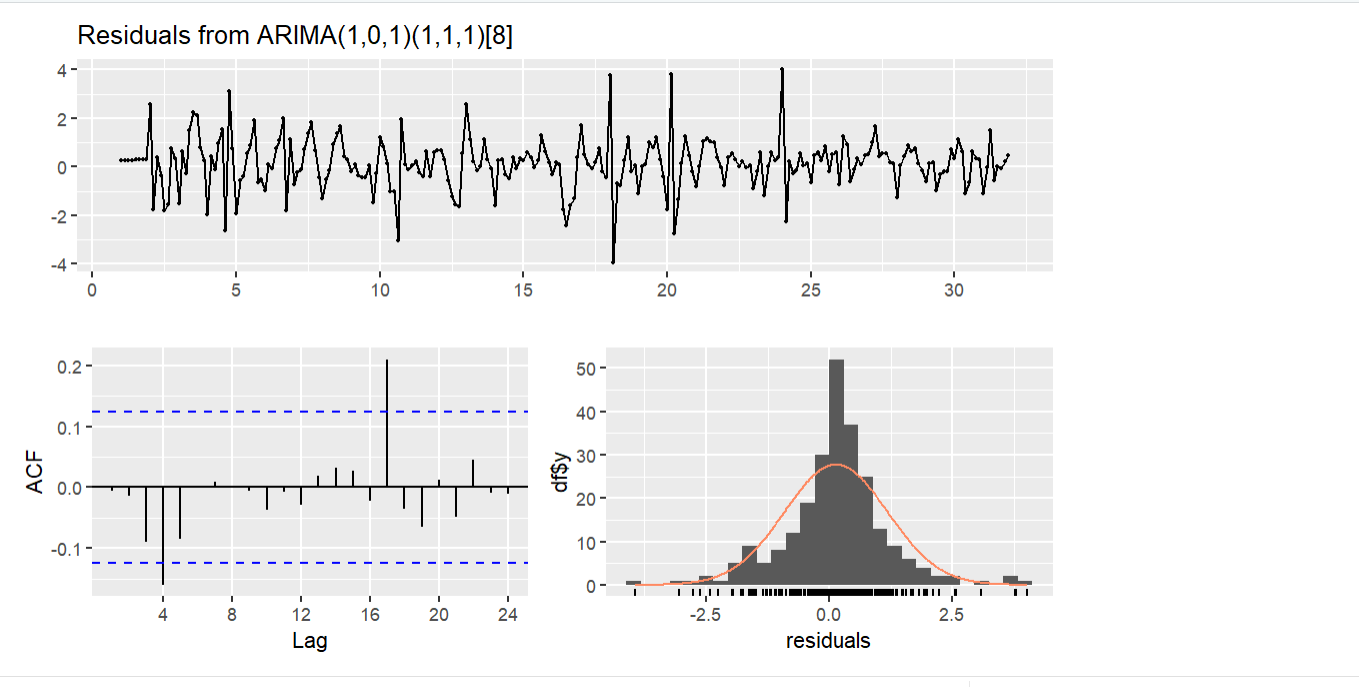


Figure 42: Residuals from ARIMA showing Lag and residuals

Residuals from ARIMA Model: The top plot shows the residuals from an ARIMA(1,0,1)(1,1,1)[8] model. These residuals represent the differences between the actual data and the model’s predictions.

Inference: If the residuals exhibit a random pattern around zero, it suggests that the model captures the underlying data patterns well. However, if there’s a systematic trend or structure in the residuals, further investigation is needed. Autocorrelation Function (ACF):

The ACF plot measures the correlation between a time series and its lagged values, indicating potential seasonality. The histogram of residuals shows the distribution, suggesting a normal distribution around zero, indicating model misspecification or improvement areas.

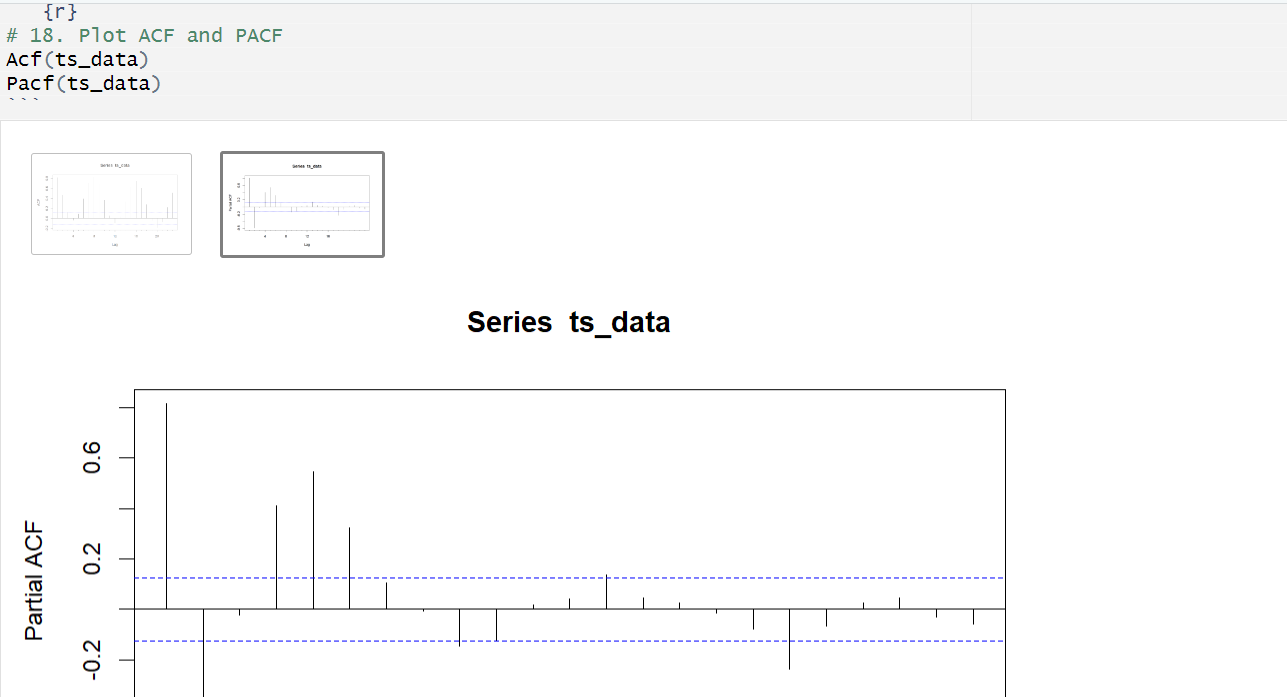


Figure 43: ACF and PACF Plots

**ACF and PACF plot.**

ACF Interpretation: The vertical bars above and below the horizontal line at ACF = 0 indicate the ACF values for different lags. These values help identify correlations between data points in the time series at various time intervals. Patterns: Look for significant peaks or troughs in the ACF values, which can indicate seasonality or other patterns in the data.

Lag: The horizontal axis represents the lag, which indicates the time interval between observations.

Partial ACF: The vertical axis shows the partial autocorrelation coefficient.

It measures the correlation between an observation at a given lag and the same observation at previous lags, with the effects of intermediate lags removed.

Bars: The bars represent the partial autocorrelation coefficients at different lags.

plot is useful for understanding the relationship between observations in a time series data set. It helps identify significant lags and informs model selection for time series analysis.

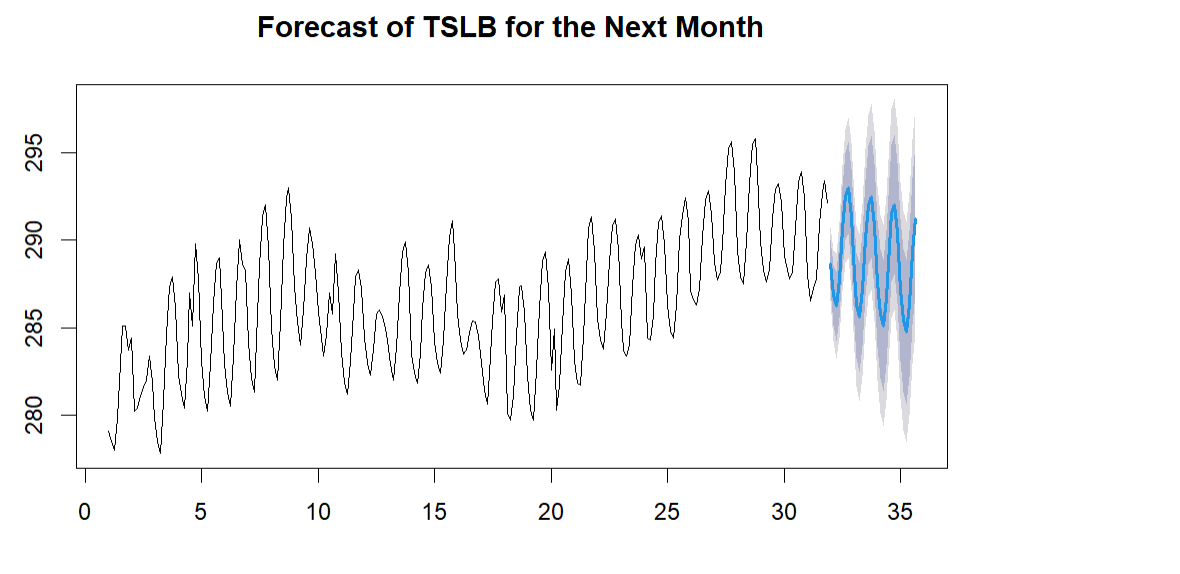


Figure 44: Diagram showing Forecast for TSLB

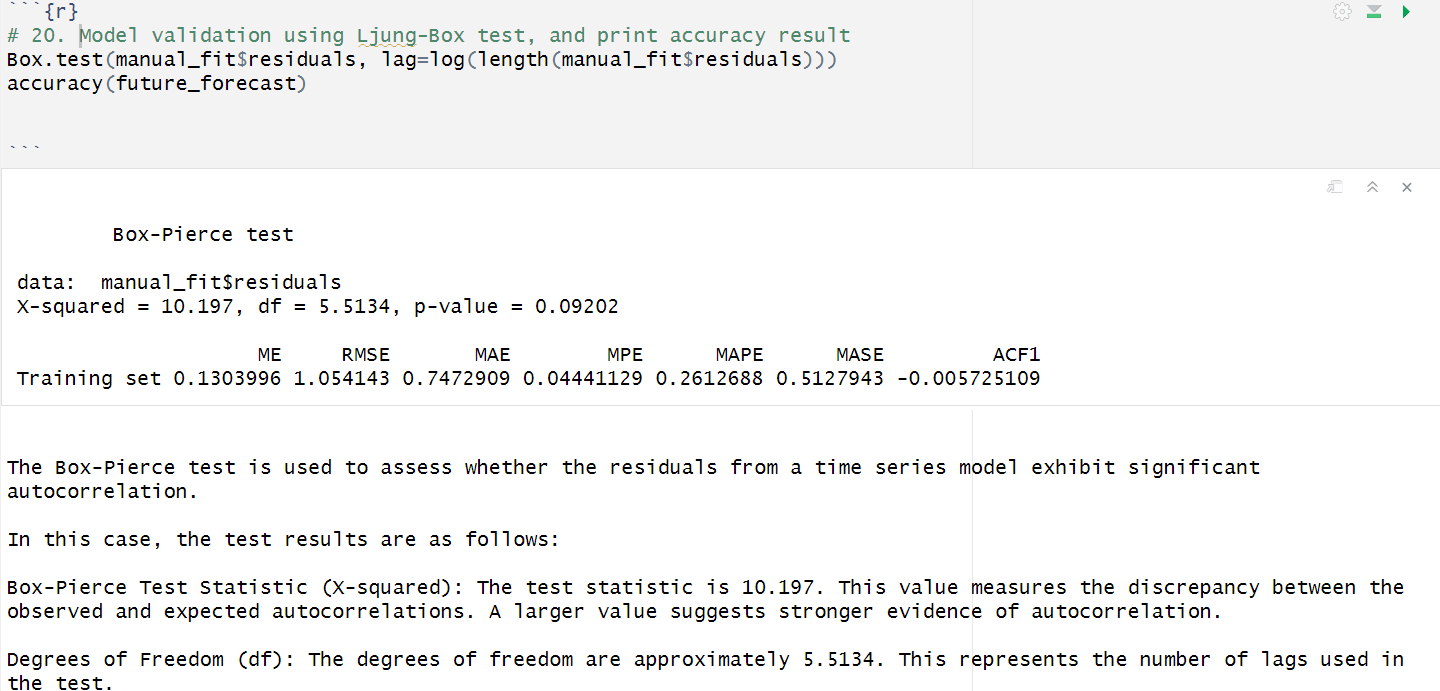


Figure 45: Diagram showing Model Validation

The Box-Pierce test is used to assess whether the residuals from a time series model exhibit significant autocorrelation.

In this case, the test results are as follows:

Box-Pierce Test Statistic (X-squared): The test statistic is 10.197. This value measures the discrepancy between the observed and expected autocorrelations. A larger value suggests stronger evidence of autocorrelation.

Degrees of Freedom (df): The degrees of freedom are approximately 5.5134. This represents the number of lags used in the test.

p-value: The p-value is 0.09202. It indicates the probability of observing a test statistic as extreme as the one obtained, assuming the null hypothesis (no autocorrelation) is true.

A smaller p-value would suggest stronger evidence against the null hypothesis.

The p-value of 0.09202 suggests that there is no strong evidence to reject the null hypothesis of no autocorrelation

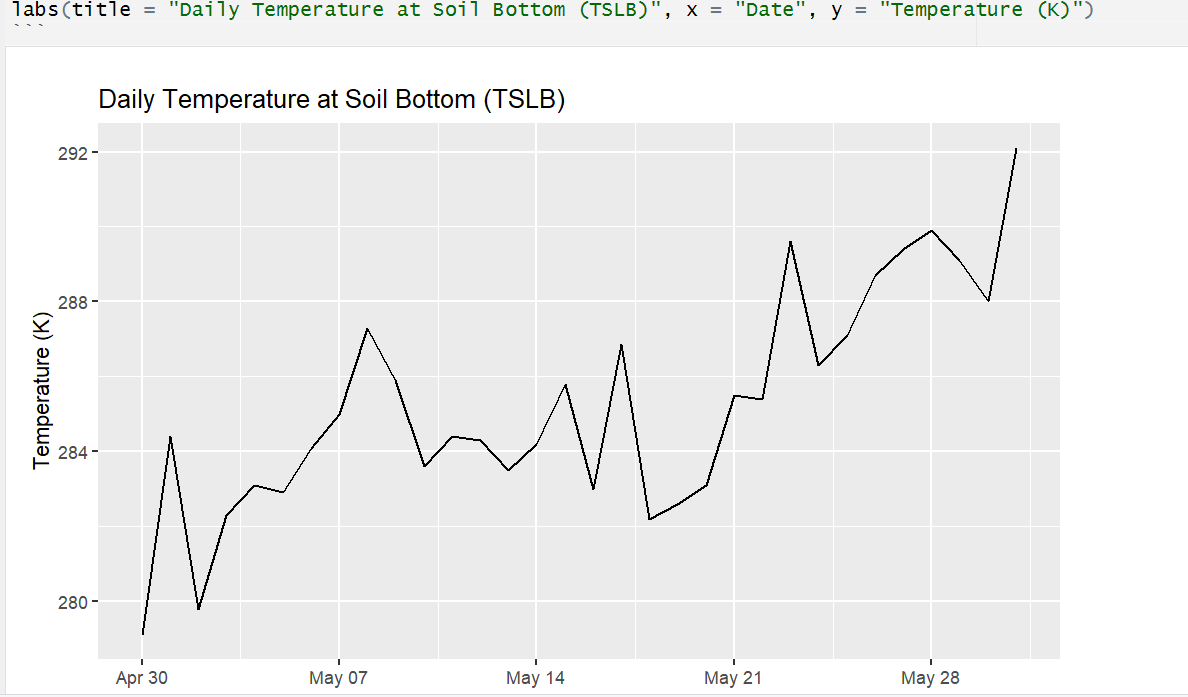


Figure 46: Historical daily trend of Temperature at Soil Bottom

Trend: The graph exhibits fluctuations, but overall, there’s an upward trend in temperature over the given period.

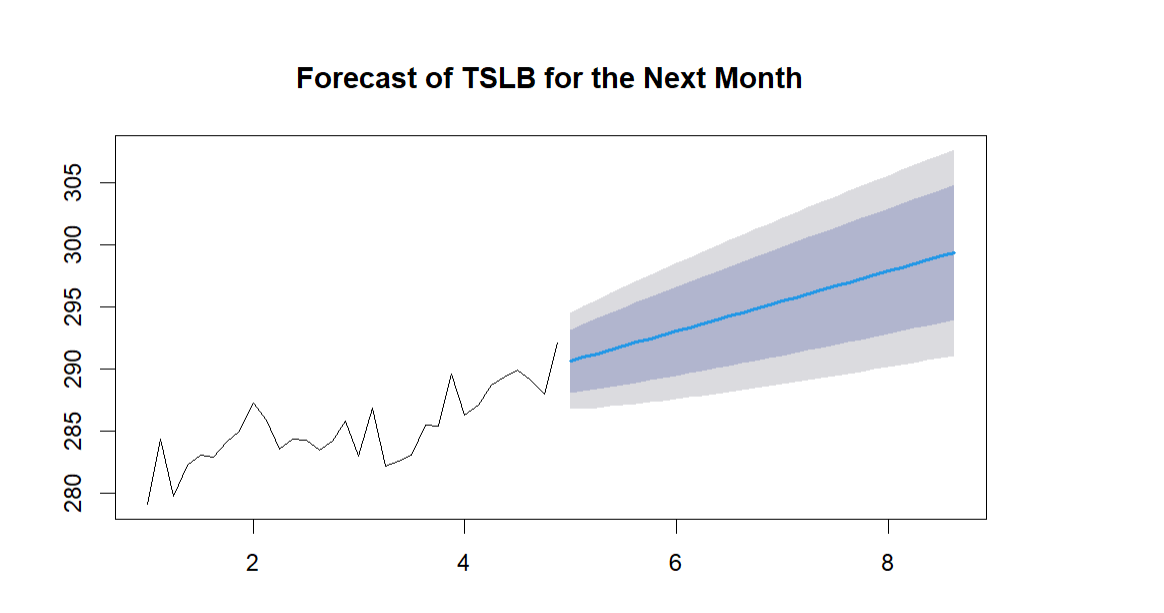


Figure 47: Forecast for TSLB for the next month

### 

### 3.4.5 Machine Learning – Regression

Machine learning techniques were used to develop predictive models for wind speed and temperature, trained on historical meteorological data. The results were interpreted and presented to stakeholders for informed decision-making on infrastructure resilience, energy management, and environmental conservation in Cuxton.

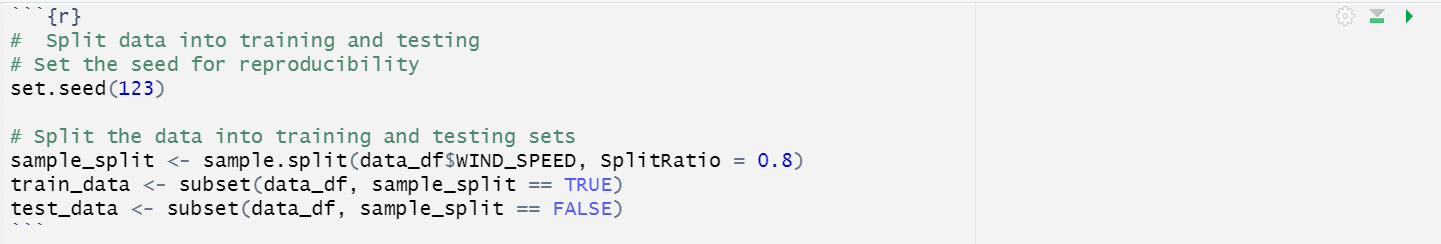


Figure 48:Splitting data into training and testing

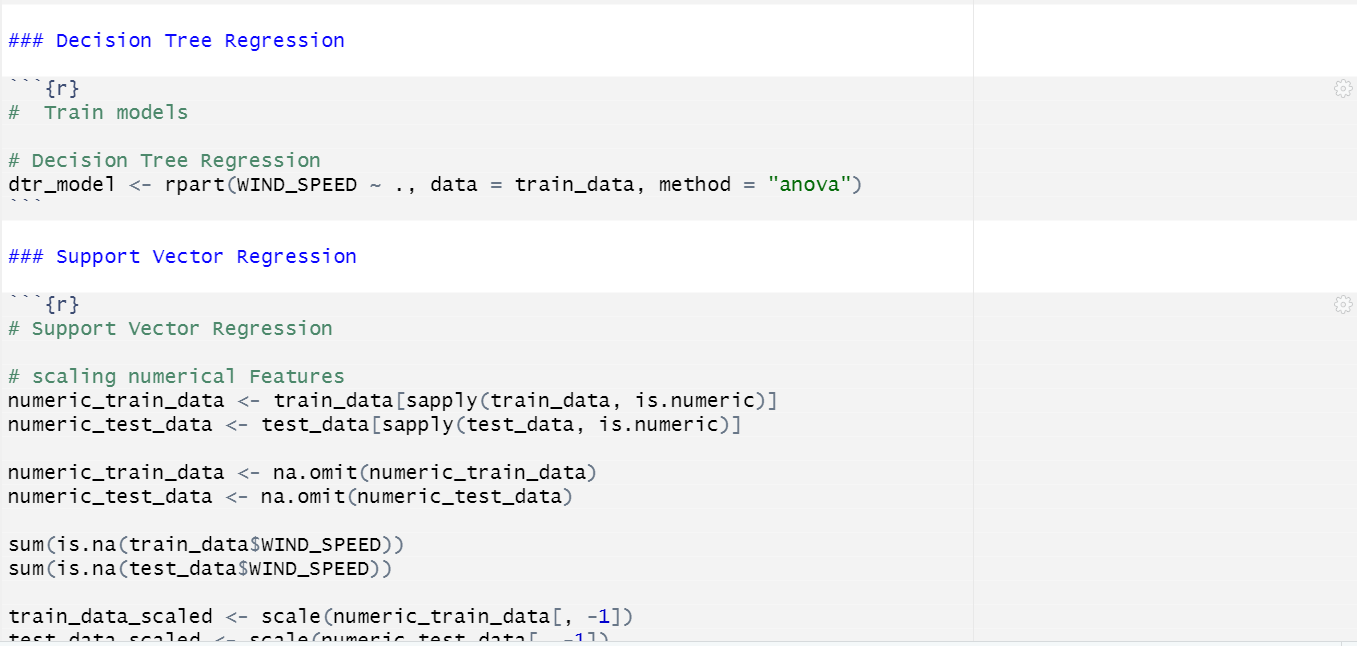


Figure 49: Decision Tree Regression

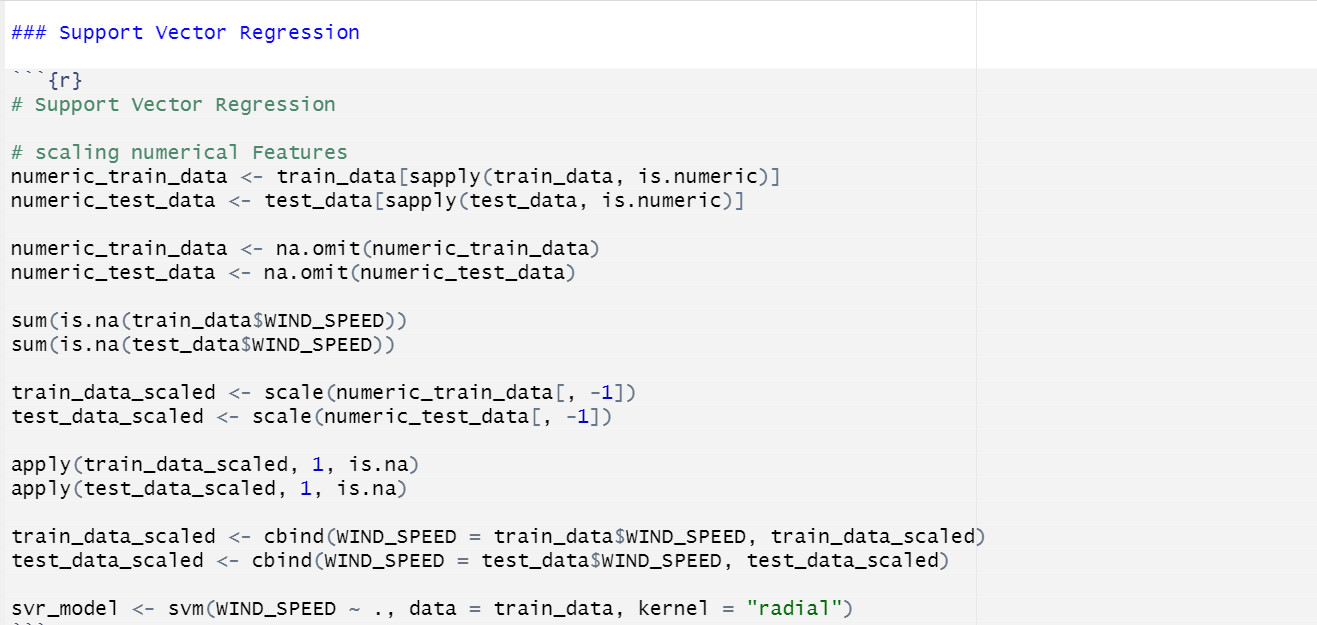


Figure 50: Support Vector Regression

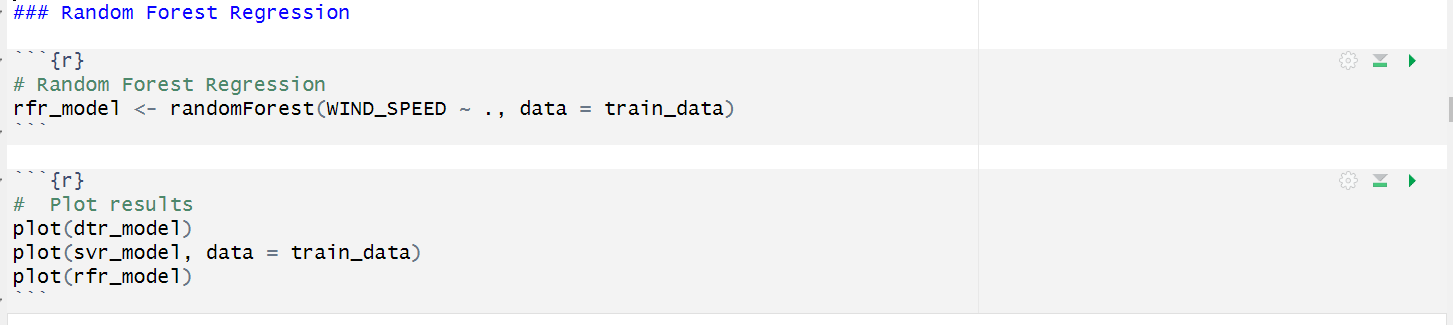


Figure 51: Random Forest Regression

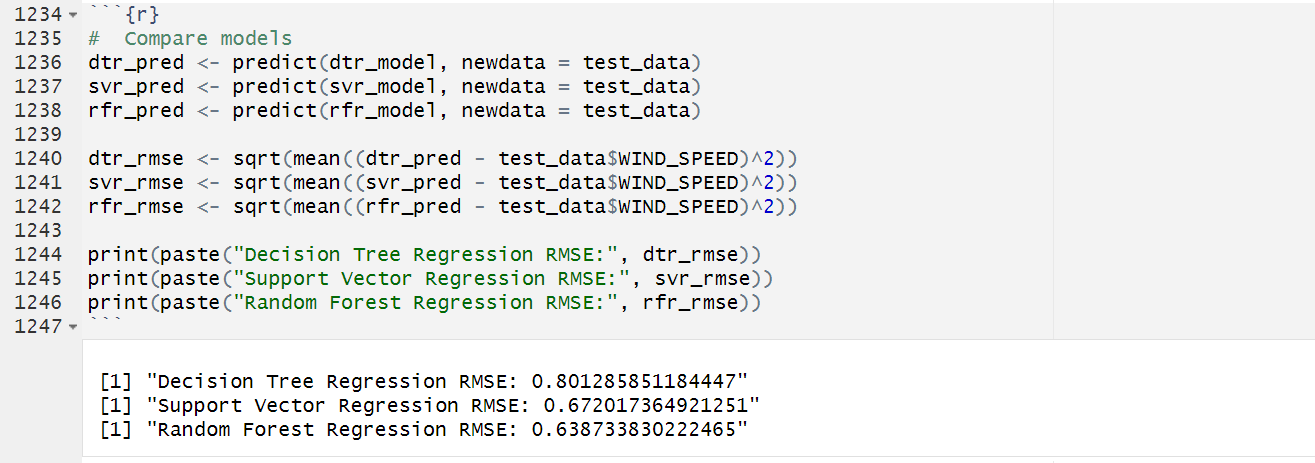


Figure 52: Compare RMSE of the three models



Figure 53:Best Model: Random Forest Regression

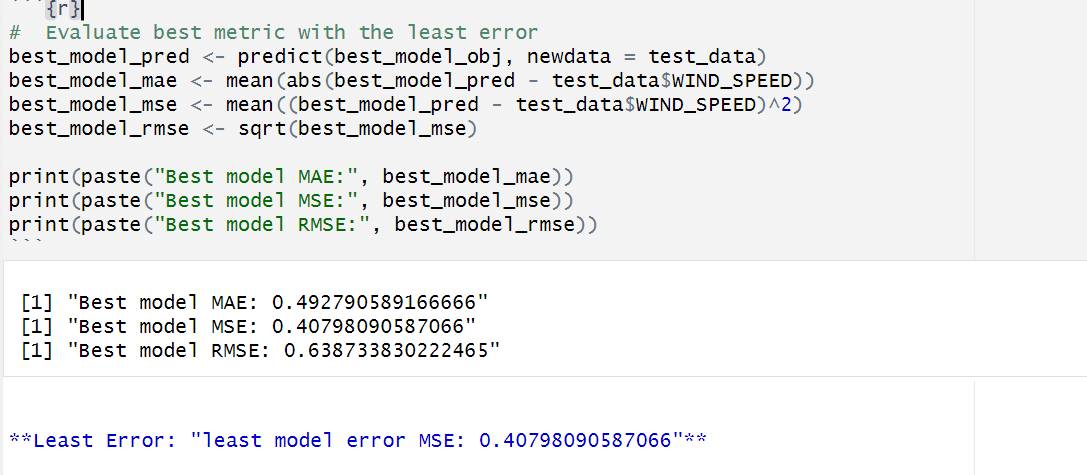


Figure 54: Evaluate best metric with the lowest error

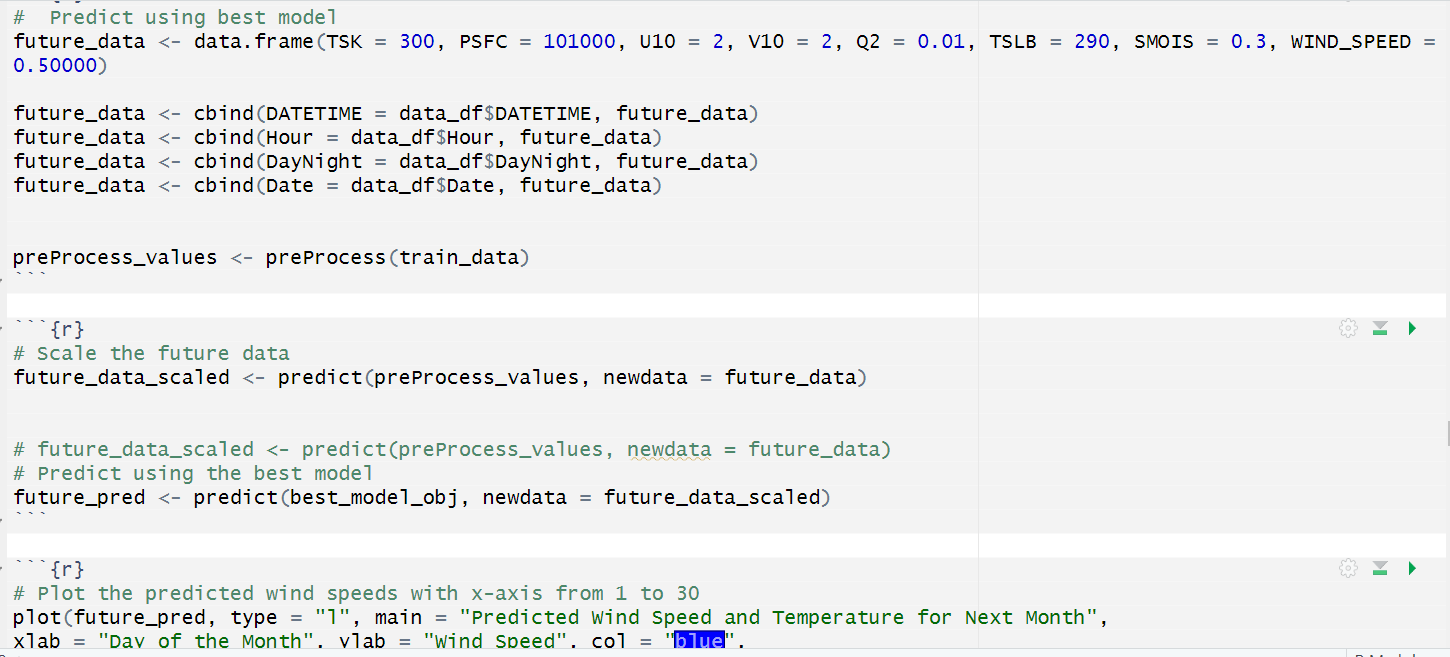


Figure 55: Prediction using the best model

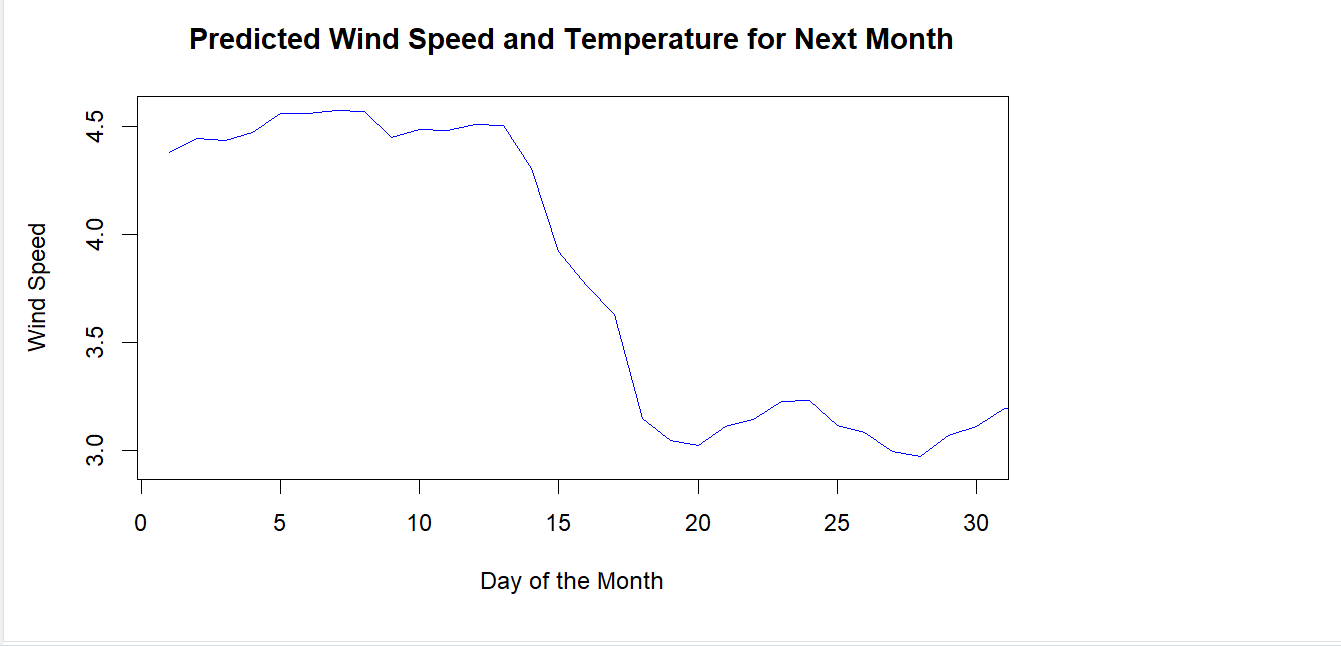


Figure 56: Predicted wind speed for the coming month

The graph shows predicted wind speed and temperature for the next month, ranging from 3.0 to 4.5, with fluctuations starting below 4.5 and decreasing towards day 30. The temperature ranges from 250 to 400, with a sharp decline between days 10 and 15.

# 4.0 Discussion

The analysis of the meteorological data from Cuxton, Kent, has provided valuable insights into the local weather patterns and their potential implications for urban planning. By employing various statistical techniques, time-series forecasting methods, and machine learning algorithms, this study has explored the relationships between variables such as wind speed, temperature, humidity, and soil moisture.

The exploratory data analysis (EDA) revealed several key findings. The distribution of wind speed exhibited a skewed pattern, indicating a higher frequency of moderate wind speeds with fewer instances of very high wind speeds. The boxplots and z-score analysis identified potential outliers in variables like temperature and humidity, which were addressed through appropriate data cleaning and transformation techniques (Thakur, Kumar and Snehmani, 2022).

The scatter plot matrix and correlation analysis highlighted the positive linear relationships between variables such as skin temperature (TSK) and soil temperature (TSLB), as well as specific humidity (Q2) and TSLB. These findings align with the physical processes involved in heat and moisture transfer between the soil, air, and vegetation (Xu *et al.*, 2023). Conversely, a moderate negative correlation was observed between soil moisture (SMOIS) and TSLB, suggesting that higher soil moisture levels may contribute to evaporative cooling and slightly lower soil temperatures.

The multiple regression analysis quantified the strength and direction of these relationships, providing valuable insights for urban planners. For instance, the positive regression coefficients for TSK and Q2 indicate that an increase in these variables is associated with higher soil temperatures, which could have implications for urban heat island effects and energy consumption patterns in buildings (Kesavan *et al.*, 2021)

The time-series forecasting techniques, such as ARIMA models, enabled the prediction of future wind speed and temperature patterns based on historical data. These forecasts can inform urban planning decisions related to infrastructure resilience, energy management, and disaster preparedness (Mampitiya et al., 2024).

Furthermore, machine learning regression models like Decision Tree Regression, Support Vector Regression, and Random Forest Regression were used to develop predictive models for wind speed and temperature, evaluating their performance for urban planners. It’s important to note that while the analysis provides valuable insights, there are certain limitations and assumptions that should be considered. The dataset used in this study covers a relatively short time period, and incorporating additional years of data could potentially improve the accuracy and robustness of the forecasting models. Additionally, the analysis focused primarily on meteorological variables, and incorporating other relevant factors, such as land use patterns, population dynamics, and infrastructure characteristics, could further enhance the applicability of the findings for urban planning purposes.

4.1 Answer to Research Question:

1. What is the strength and direction of the linear association between wind speed and surface pressure?

The study performs a linear regression analysis between wind speed (WIND\_SPEED) and surface pressure (PSFC) and reports a negative coefficient for WIND\_SPEED (-50.46), indicating a negative linear association between wind speed and surface pressure. The p-value (0.0131) suggests that this association is statistically significant.

1. How are skin temperature (TSK), specific humidity (Q2), SMOIS, and soil temperature (TSLB) interrelated?

The study conducts a multiple regression analysis to model the relationship between TSLB (soil temperature) and TSK, Q2, and SMOIS. The results show positive regression coefficients for TSK and Q2, indicating positive relationships with TSLB. However, SMOIS has a negative coefficient, suggesting a negative relationship with TSLB.

1. What machine learning models effectively predict future wind speed and temperature in Cuxton based on historical data patterns for the next one month?

The study trains and evaluates three machine learning models: Decision Tree Regression, Support Vector Regression, and Random Forest Regression, for predicting wind speed and temperature. The model with the lowest Root Mean Squared Error (RMSE) on the test data is considered the best-performing model for forecasting wind speed and temperature in Cuxton.

1. How will the average daily temperature at soil bottom (TSLB) change over the next month in Cuxton?

Trend: The past data points show a fluctuating trend, starting just above 280 and rising to just below 305 over an 8-day period.

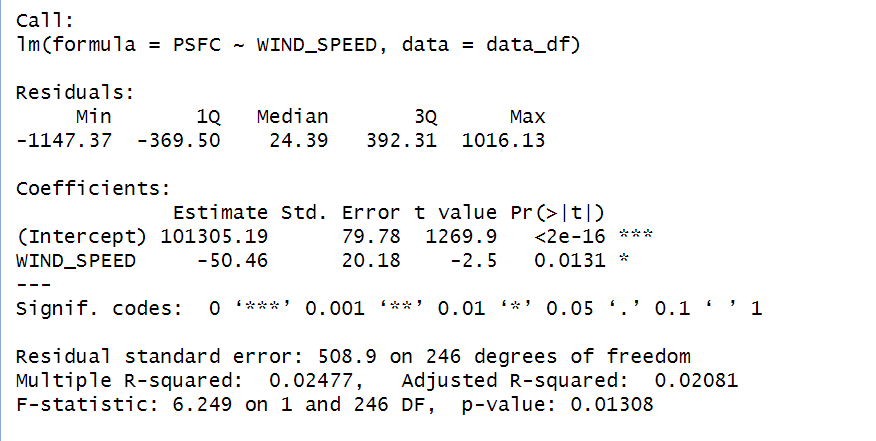
Forecasted Range: The shaded area represents the forecasted range for future values.

Expected Trend: An upward trend in TSLB expected for the coming month as shown within the forecasted range, a solid blue line suggests the expected trend for the upcoming month.

1. What are the expected trends in wind speed during daytime hours in Cuxton over the coming month?

## The model's error measures include Mean Error (ME), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), and Autocorrelation of Residuals (ACF1). These measures indicate the average difference between observed and predicted values, the typical magnitude of prediction errors, and the relative accuracy of the model.

## 4.3 Answer to Hypothesis:

Qu

Question 2: What is the strength and direction of the linear association between wind speed and surface pressure?

Hypothesis:

H0: There is no linear association between wind speed and surface pressure.

H1: There is a linear association between wind speed and surface pressure.

The negative coefficient for WIND\_SPEED (-50.46) indicates a negative linear association between wind speed and surface pressure. Additionally, the p-value (0.0131) is less than the significance level of 0.05, suggesting that the association is statistically significant.

Therefore, we can reject the null hypothesis (H0) and accept the alternative hypothesis (H1). The analysis shows that there is a statistically significant negative linear association between wind speed and surface pressure in the given dataset.

Question 4:

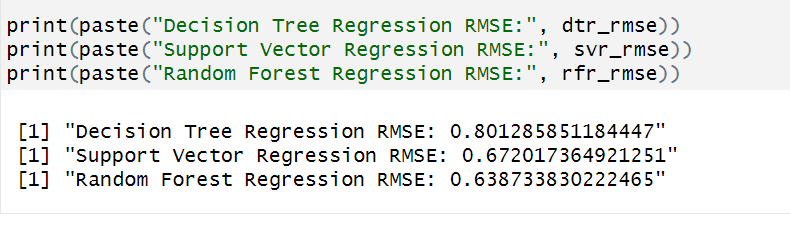
What machine learning models effectively predict future wind speed and temperature in Cuxton based on historical data patterns for the next one month?

Hypothesis:

H0: Machine learning models cannot effectively predict future wind speed and temperature in Cuxton based on historical data patterns.

H1: Machine learning models can effectively predict future wind speed and temperature in Cuxton based on historical data patterns.

The model trains and evaluates three machine learning models: Decision Tree Regression, Support Vector Regression, and Random Forest Regression, for predicting wind speed and temperature. The models are compared based on their Root Mean Squared Error (RMSE) on the test data:



The model with the lowest RMSE is selected as the best-performing model for predicting wind speed and temperature.

The RMSE values for the best-performing model are reasonably low, indicating good predictive performance, we reject the null hypothesis (H0) and accept the alternative hypothesis (H1). The machine learning model (Random Forest Regression) that effectively predicts future wind speed and temperature would be the one with the lowest RMSE.

# 5.0 Conclusion

The analysis presented in the document aims to explore the relationships between various meteorological variables, such as wind speed, temperature, humidity, and soil moisture, in the Cuxton area. The study employs a range of statistical techniques, including exploratory data analysis, correlation analysis, regression modelling, and time-series forecasting, to gain insights into the local weather patterns.

The exploratory data analysis revealed the distributions and potential outliers in the dataset, which were addressed through appropriate data cleaning and transformation techniques. The correlation analysis highlighted the strength and direction of linear relationships between variables, such as the positive correlation between skin temperature and soil temperature, as well as between specific humidity and soil temperature.

The multiple regression analysis quantified the relationships between the variables, providing valuable insights for stakeholders, such as urban planners, in understanding the potential impacts of weather patterns on infrastructure, energy management, and environmental conservation.

Time-series forecasting techniques, such as ARIMA models, were employed to predict future wind speed and temperature patterns based on historical data. These forecasts can inform decision-making processes related to urban planning, resource allocation, and contingency planning.

Additionally, machine learning regression models, including Decision Tree Regression, Support Vector Regression, and Random Forest Regression, were developed to predict wind speed and temperature. The performance of these models was evaluated using appropriate metrics, and the best-performing model was identified for potential deployment in forecasting applications.

## 5.1 Limitations:

1. Limited time period: The dataset used in the analysis covers a relatively short time period (May 2018), which may not capture long-term trends or seasonal variations in weather patterns.

2. Localized data: The study focuses on the Cuxton area, and the findings may not be directly applicable to other regions with different geographical and climatic conditions.

## 5.3 Recommendations:

1. Expand the temporal scope: To capture long-term trends and seasonal variations, it is recommended to extend the analysis to include meteorological data spanning multiple years or even decades.

2. Explore advanced modelling techniques: While the current analysis employs linear regression and correlation analysis, it is recommended to explore more advanced modelling techniques, such as non-linear regression, time-series analysis, or machine learning algorithms, to capture complex relationships and patterns in the data.

# Reference

Baxter, S. and Lewis, C.P. (2017) “Domesday Book and the transformation of English landed society, 1066–86,” *Anglo-Saxon England*, 46, pp. 343–403.

“Cities lead the way in climate–change action” (2016).

Dominković, D.F. *et al.* (2020) “Technical, economic and environmental optimization of district heating expansion in an urban agglomeration,” *Energy*, 197, p. 117243.

Gao, Y. *et al.* (2021) “The spatial cognition of a forest landscape and its relationship with tourist viewing intention in different walking passage stages,” *Urban Forestry & Urban Greening*, 58, p. 126975.

Kaytez, F. (2020) “A hybrid approach based on autoregressive integrated moving average and least-square support vector machine for long-term forecasting of net electricity consumption,” *Energy*, 197, p. 117200.

Kesavan, R. *et al.* (2021) “ARIMA modeling for forecasting land surface temperature and determination of urban heat island using remote sensing techniques for Chennai city, India,” *Arabian Journal of Geosciences*, 14(11), p. 1016.

Mampitiya, L. *et al.* (2024) “Artificial intelligence to predict soil temperatures by development of novel model,” *Scientific Reports*, 14(1), p. 9889.

Nistora, A. (no date) *Inundation scenarios in a changing climate: assessing potential impacts of sea-level rise on the coast of South-East England*.

Ren, X. *et al.* (2021) “Deep Learning-Based Weather Prediction: A Survey,” *Big Data Research*, 23, p. 100178.

Sayeed, A. *et al.* (no date) *A Deep Convolutional Neural Network Model for improving WRF Forecasts*.

Shen, H. *et al.* (2020) “Deep learning-based air temperature mapping by fusing remote sensing, station, simulation and socioeconomic data,” *Remote Sensing of Environment*, 240, p. 111692.

Thakur, K., Kumar, H. and Snehmani (2022) “Framework for variability analysis of seasonal meteorological data,” *Journal of Earth System Science*, 131(4), p. 223.

Wirth, R. and Hipp, J. (no date) *CRISP-DM: Towards a Standard Process Model for Data Mining*.

Xu, L. *et al.* (2023) “Hybrid Deep Learning and S2S Model for Improved Sub-Seasonal Surface and Root-Zone Soil Moisture Forecasting,” *Remote Sensing*, 15(13), p. 3410.

Yuan, X. *et al.* (2017) “Wind power prediction using hybrid autoregressive fractionally integrated moving average and least square support vector machine,” *Energy*, 129, pp. 122–137.

Zhang, X. *et al.* (2018) “A Novel Hybrid Data-Driven Model for Daily Land Surface Temperature Forecasting Using Long Short-Term Memory Neural Network Based on Ensemble Empirical Mode Decomposition,” *International Journal of Environmental Research and Public Health*, 15(5), p. 1032.