Bicol University
College of Science
CSIT Department
AY 2023-2024

Detecting Anomalies in Philippine Weather using Temporal Weather Data

CS Elec 2 - Data Mining

2nd Semester

Submitted by

Paul Edward Balistoy

Lance Stephen Bronzal

Yna Gabrielle Foronda

Francis Maurice Miranda

Jonathan Pelon

BS Computer Science 3A

Submitted to

Jennifer Laraya-Llovido

INTRODUCTION

The atmosphere, or layer of air surrounding the Earth, is what we refer to as "weather.”. The weather has a variety of effects on humans, and daily variations can change our moods and perspectives on the outside world (Weather, n.d.). The planet's climate is changing, and with it, the frequency and intensity of extreme weather events such as record-breaking heat waves on land and in the ocean, torrential rains, severe floods, protracted droughts, intense wildfires, and widespread flooding during hurricanes (*Extreme Weather*, n.d.).

The Philippines is an archipelago nation located in the western Pacific Ocean that is no stranger to the drastic impacts of weather changes. The country has been particularly vulnerable to extreme weather events due to its geography and development. On average, the country experiences about 20 typhoons entering its waters each year, with eight to nine of them making landfall (How Is Climate Change Affecting the Philippines?, 2016). Climate change is believed to have played a major role in the typhoon’s rising frequency and intensity over the past ten years.

The Philippines’ climate is generally high, especially in landforms such as valleys and plains, averaging about 27°C throughout the year (World Bank Climate Change Knowledge Portal, n.d.). Additionally, the mean annual temperature of the Philippines based on the average of all weather stations in the Philippines, excluding Baguio, is 26.6°C. The coolest months usually fall in January, with a temperature of 25.5°C on average, while the warmest month falls in May, with a mean temperature of 28.3°C (PAGASA, n.d.).

However, despite the weather following a trend we have all been familiar with, these patterns have been undergoing various changes. Based on recent scientific assessments, the warming climate system occurring in the mid-20th century is most

likely due to man-made and artificial activities like fossil fuel burning and the urbanization of landmasses (PAGASA, n.d.). This warming has resulted in a significant number of environmental problems and will continue to persist in the future.

In this context, detecting weather anomalies using a temporal data set becomes crucial. Anomaly detection, or outlier detection, is a method of finding patterns or instances in a dataset deviating significantly from what is normal, or "mean behavior" (Tuychiev, 2023). The implications of these anomalies for the case of weather data may signify extreme events such as changes in climate patterns. However, Sania (n.d.) also states that it may also imply an error in data collection that could potentially result in inaccurate environmental science and policies.

The practical implications of anomaly detection contextualized to weather data are vast. For instance, Parimala (2024) states that its findings can be used to provide early warnings for earthquakes, tsunamis, hurricanes, and other natural disasters alike, enabling preparedness for disaster risk mitigation, reduction, and management. Furthermore, anomaly detection can help in classifying weather conditions for smart cities, keeping transportation systems running smoothly (Society, 2023).

The researchers are planning to use four algorithms for comparison. Namely, k-Nearest Neighbors (k-NN), One Class Support Vector Machine (OCSVM), Density-based spatial clustering of applications with noise (DBSCAN), Cluster-based Local Outlier Factor.

In this paper, the researchers will delve deeper into various methods and algorithms used for anomaly detection in weather data. This study's findings may provide beneficial insights on the weather patterns in the Philippines, potentially contributing to identifying and mitigating the impacts of extreme weather conditions in the country.

OBJECTIVES OF THE STUDY

This case study aims to explore anomalies in Philippine weather. Specifically, it aims to:

1. Collect and prepare the data for the modeling
2. Conduct the modeling for anomaly detection using the defined algorithms
3. Compare the results of the different algorithms using the following key performance measures:
    a. Visual Inspection / Domain Knowledge
    b. Cross-validation of results
    c. Density Estimation
4. Identify insights on the outliers of the weather data, such as:
    a. Percentage of outliers over total instances
    b. Months/Years with most number abnormal weather occurrences
    c. Identify any patterns/trends the outliers show

REVIEW OF RELATED LITERATURE

This chapter presents the related literature and studies contents wherein the existing knowledge about impact of weather anomalies, anomaly detection algorithms and characterization of weather anomalies were described. This chapter also contains the gaps bridged by the study.

A. Impact of Weather Anomalies

The economic impact of weather anomalies

The study of Gabriel Felbermayr et al. (2022) shifts their focus to the local level, moving away from varying country sizes or incomparable administrative entities. They use average annual night-time light emission as a proxy for economic activity. They pair this data with the Geological and Meteorological Events (GAME) database to create a balanced panel of grid-cells covering all land mass around the globe. This data set, named GAME-LIGHTS, covers 197 countries from 1992 to 2013 and is available online.

Their findings show a reduction in night-time lights following wind, low temperatures, and high precipitation anomalies. They also observe that these anomalies are localized phenomena, with neighboring cells benefiting from a weather shock and taking over some of the lost economic activity. This suggests

that effects measured at larger geographical units often obscure the true impact at the local level.

This study contributes to the literature by providing a publicly available data set, emphasizing the importance of studying weather events at a high spatial resolution, and demonstrating the significant negative impact of weather anomalies on the local economy. Their results hold policy relevance, indicating that extreme weather due to global climate change can significantly affect the economic geography of regions, countries, and continents.

Changes of extreme precipitation in the Philippines, projected from the CMIP6 multi-model ensemble

A study conducted by Hong et al. (2022) discussed the frequency of extreme precipitation in the Philippines. They applied Generalized Extreme Value distribution, Multivariate-bias correction technique, and PI-weighting method on the observations recorded by 53 stations and 24 CMIP6 models, and the obtained results of their study predicted an increase in the frequency of extreme precipitation in the country.

Given the results of their study, it is implied that preparedness for extreme weather be given more attention, and in turn supports this literature by

highlighting the importance of studying and detecting weather anomalies in the country, as well as the urgency of this literature, given the increasing frequency of extreme precipitation. The results of their study also presents an implication that this research be continuously improved and updated as the circumstance changes.

Investigating the Effect of Urbanization on Weather Using the Weather Research and Forecasting (WRF) Model: A Case of Metro Manila, Philippines

A study by Oliveros et al. (2019) delved into the impact of urbanization, particularly in Metro Manila, on various weather parameters utilizing the Weather Research and Forecasting (WRF) model. Meteorological data from the National Center for Environmental Prediction - Final (NCEP-FNL) grib1 dataset spanning from 2000 to 2010 were employed as inputs for the model. Through the Mann–Kendall trend test and Sen's slope estimator, trends in sensible heat flux, temperature, and rainfall were analyzed, shedding light on the influence of urbanization. Results indicate a significant difference in sensible heat flux between Metro Manila and surrounding rural areas, indicating the presence of an urban heat island (UHI) effect. Temperature differentials further underscore this phenomenon, with Metro Manila exhibiting higher values compared to adjacent regions. Moreover, Metro Manila experienced elevated levels of rainfall across all seasons when contrasted with proximal areas. These findings emphasize the

intricate relationship between urbanization and weather patterns, urging a closer examination of anomalies in Philippine weather.

The study's focus on detecting anomalies in Metro Manila's weather patterns due to urbanization serves as a pertinent reference for understanding broader anomalies in Philippine weather. Through employing advanced modeling techniques and rigorous analysis, the study highlights the nuanced impacts of urban development on key meteorological parameters. As such, this study contributes essential knowledge to the overarching goal of detecting and understanding anomalies in Philippine weather which would contribute valuable insights for future research and preparedness initiatives.

B. Anomaly Detection Algorithms

Multivariate weather anomaly detection using DBSCAN clustering algorithm

According to Wibisono et al. (2021), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method can be used to differentiate anomalous data from normal data groups in the case of using unlabeled data (unsupervised learning). They also added that multivariate type (using many variables) of anomaly detection provides better performance than univariate (using one variable) for more complex data handling.

For data pre-processing, they replaced unmeasured or no data with null, and also applied normalization. Using functions from the ‘dbscan’ package in R, they did several clustering/eps/minpts experiments to determine the final eps and minpts value to use for the DBSCAN algorithm. Eps is the maximum distance from a point to another to determine if they belong to the same cluster, while minpts is the minimum number of points to determine if a point belongs to the member of the cluster within the radius eps. The DBSCAN algorithm assigns a point to either the core points, border points, or noise points. The cluster combines the core points surrounded by border points and the noise points are the anomalies.

They used PCA to visualize the clusters formed and demonstrated that DBSCAN is capable of determining anomalous data, in which in the dataset they utilized is characterized by high humidity and low temperature.

A comparative study of Anomaly Detection Methods for Gross Error detection problems

A study by Dobos et al. (2023) compared 19 different Anomaly Detection Methods to detect Gross Errors on synthetic data with the aim of strengthening the potential beneficial use of the application of Data Mining methods on the chemical industry. Their study also conducted the comparison of these Anomaly

Detection methods twice; one using raw dataset and the other with data preparation performed on said dataset.

The results of their study shows that the One Class Support Vector Machine (OCSVM) performs with the best consistency among the different algorithms, and arguably the most versatile. The OCSVM method achieved the highest accuracy and F1 score among the 19 selected methods and has also shown to be more accurate as the training dataset increases in size while the accuracy of some decreases. It must be noted, however, that the reliability of this method is also undermined as it labeled the testing data with relatively more False Positives compared to the other methods.

Their study provides this research with a suggestion of which of the Machine Learning-based Anomaly Detection Methods are viable to use, and provides details of the characteristics and performances that further the understanding of these methods, thus providing a basis for deciding whether or not to use a certain method.

Deep Learning for Anomaly Detection in Spatio-Temporal Maharashtra Weather Data: A Novel Approach with Integrated Data Cleaning Techniques

In the paper of Kulkarni et al. (2024), they added the expectation maximization optimization technique in their data cleaning process to provide a

foundation for a more reliable anomaly detection and forecasting of weather conditions in Maharashtra. They made use of different algorithms such as One-Class SVM, Isolation Forest, LSTM Autoencoders, measuring their efficacy. As Maharashtra's economy heavily relies on agriculture, unpredictable weather patterns can badly affect it, that is why the paper's goal is to identify subtle patterns and trends in the meteorological data in both space and time dimensions.

The Expectation-Maximization (EM) is a statistical algorithm used for handling missing or null values in time-series data, starting with initialization, then alternating between expectation and maximization, and lastly, doing convergence check, implementing an iterative refinement throughout until the convergence is completed. To bypass the curse of dimensionality, they calculated for a correlation matrix and selected highly correlated features with a threshold of 0.5.

To measure effectiveness of algorithms they used accuracy, reconstruction error, and validation loss. In there comparative analysis of the applied models they discovered: (1) One-class SVM detected very less anomalies due to complex kernel function, (2) Isolation Forest has an unsupervised technique which may not be effective, (3) Autoencoders are sensitive to noisy data, not effective for detecting subtle anomalies, and low validation loss and mean reconstruction error, and (4) LSTM Autoencoders require careful hyperparameter

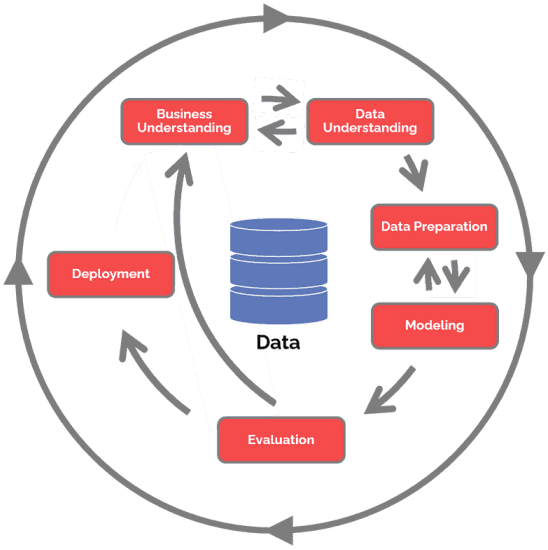
tuning and not suitable for real-time anomaly detection, shown good results with the validation loss and lower mean reconstruction error.

The paper recommends incorporating visualizations as it facilitates the interpretation of detected anomalies but also enhances the intuitive understanding of spatiotemporal dynamics of the dataset. The data from this research paper was concluded to possibly improve forecasting and support the building of climate resilience.

The aforementioned studies significantly contribute to understanding the impacts of weather anomalies and utilizing anomaly detection techniques. However, there are still gaps that still remain to be addressed particularly in developing countries like the Philippines to which this case study aims to help fill the gaps. There is still a lack of studies that focus specifically on weather anomalies in the country despite being prone to weather hazards, and this brings an opportunity for further research and development of solutions to address these challenges. This study aims to bridge the gap by developing and evaluating anomaly detection algorithms specifically optimized for temporal weather data analysis in the Philippine context and provide insights on the probable anomalous weather that we have already faced.

METHODOLOGY

The following figure shows the Cross-Industry Standard Process for Data Mining (CRISP-DM) Model, which serves as a basis for the data science process.



A. Business Understanding

For this case study, the objective is to identify the anomalous data in the weather dataset scraped from wundergound.com (Weather Underground). With how prone to extreme weather, weather changes, and natural disasters the Philippines is and how the livelihood depends on weather, it is crucial to take note of these anomalous events to better equip the community with policies, technology, procedures, etc. to mitigate or prepare for such events.

B. Data Understanding

To fully understand the Philippine weather dataset, the researchers accomplished this data dictionary for the raw dataset:

DATE	Date when observation was collected (YYYY-MM-DD)

TIME	Time of the day when observation was collected (HH:SS AM/PM)
TEMP	Temperature at given point (°F)
DP	Dew Point at given point (°F)
HUM	Relative Humidity at given point (%)
WND_DIR	Direction of the wind at a given point
WND_SPD	Speed of the wind at a given point (mph)
WND_GST	Gust of the wind at given point (mph)
PRES	Amount of pressure (in)
PRECIP	Amount of rainfall within the hour (in)
COND	Descriptive weather condition

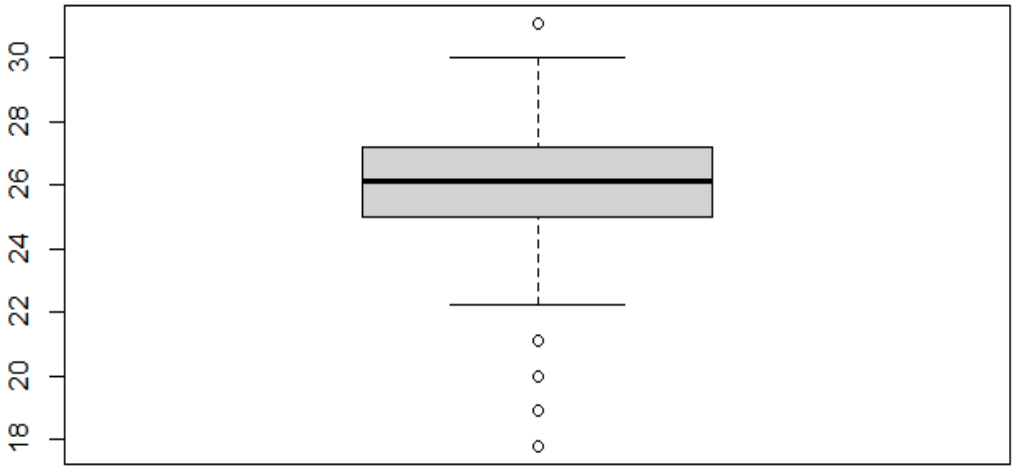
The researchers developed a python script to automatically obtain the hourly observations from "Pasay, Philippines Weather History" for each day from 2013-2023. The dataset has missing values during some days of 2020 (due to unavailability of the information on the website) which must be handled before training the model and incorrect values such as "0" values under temperature. The raw dataset includes all the attributes except the data and time in character type or string. Some variables are textual such as the wind direction and condition which may not be utilized for this case study. The original dataset has 97,421 instances and 11 columns.

Exploratory Data Analysis

Box Plots

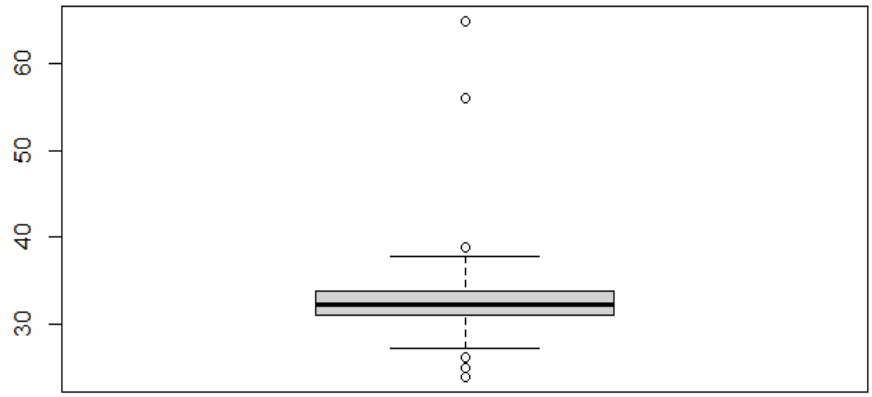
Each attribute of the merged data frame was displayed in the form of a box plot:

Minimum Temperature Box Plot



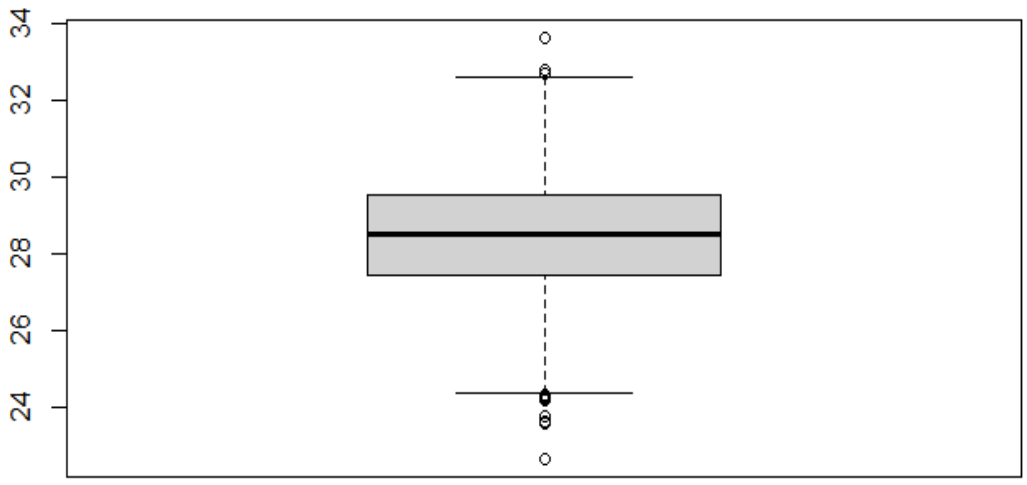
This box plot shows that the minimum temperature recorded is ranged between 22 to 30 degrees, without the outliers. The common minimum temperature range however is in between 25 to 27 degrees with the median sitting at 26 degrees. This means that the usual lowest temperature experienced in Pasay throughout the years is around 25 to 27 degrees with the most common being at 26.

Maximum Temperature Box Plot



The average maximum temperature recorded in Pasay is about in the range of 30 to 35 degrees. The highest minimum and maximum range, without considering the outliers, is between sub-30 to 39. There are however outliers which surpass this maximum in where the highest recorded temperature is above 60 degrees and the lowest being around 25 or lower.

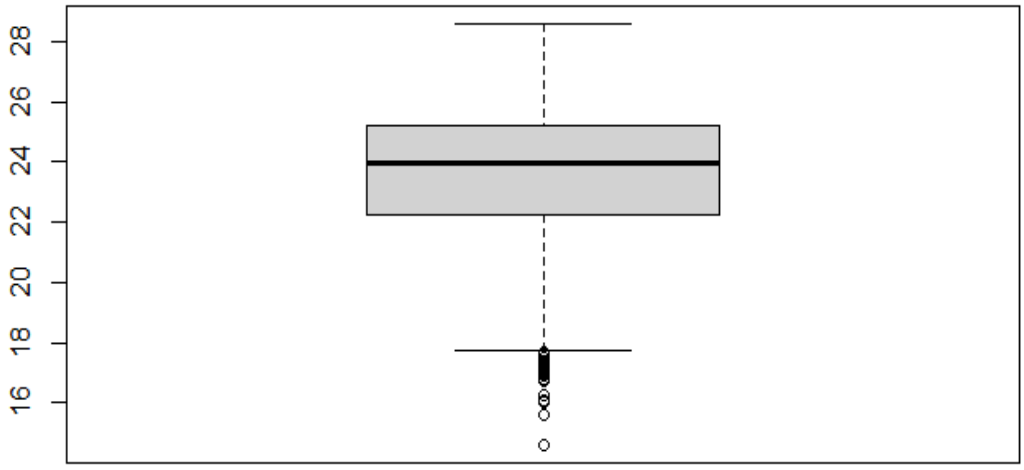
Average Temperature Box Plot



The average recorded temperature falls between 24 to 33 degrees with the vast majority of the records being between 27 to 29.5 degrees. Based on the

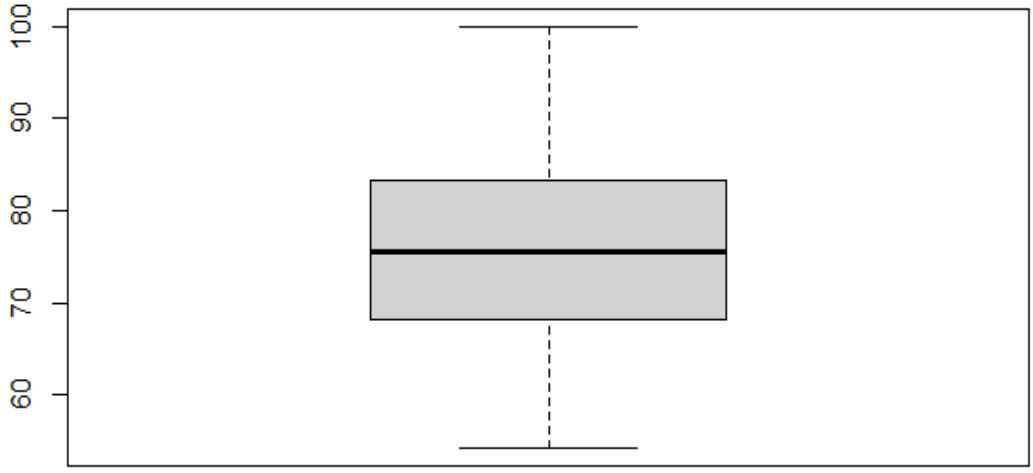
given data, Pasay usually experiences 29 degrees of average temperature in the past years.

Average Dew Point



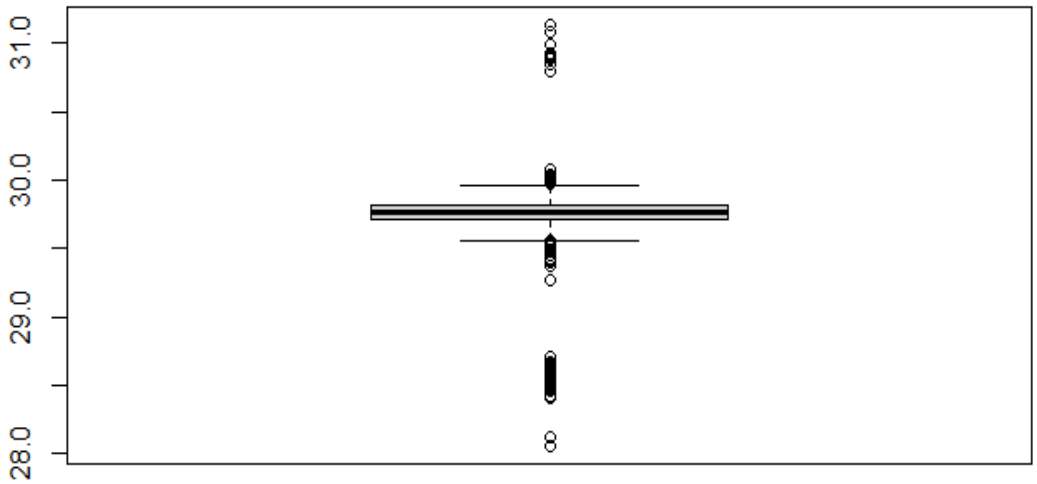
The overall average dew point falls into the range of 18 to 28 degrees with its outliers being below 18 degrees. The common dew point temperature value is in the 22 to 25 degrees range. The most common dew point is at 24 degrees.

Average Humidity



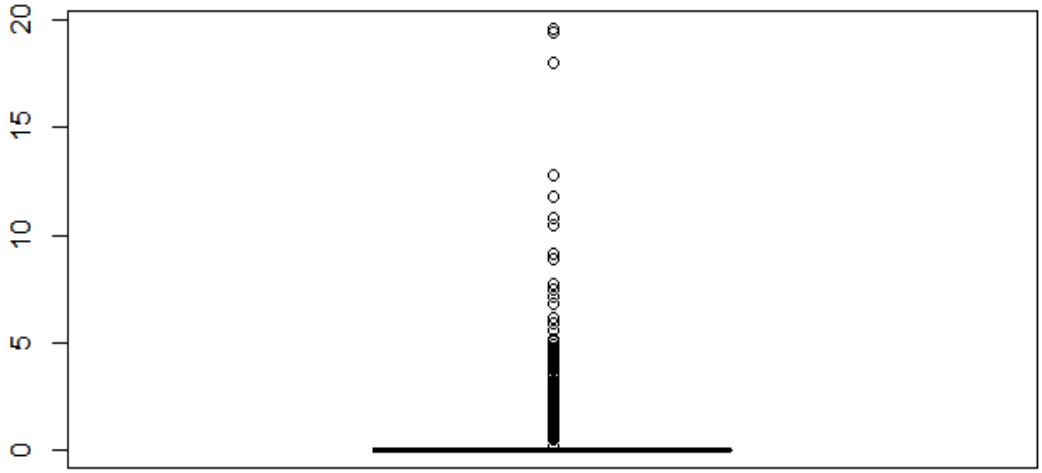
Unlike the other box plots, the average humidity doesn't have recorded outliers. This means the maximum value is at 100 percent and the minimum being below around 50 percent. The vast majority of recorded humidity values is between 65 to 85 percent with the median being at 75 percent, which indicates a rather humid average environment over the years.

Average Pressure



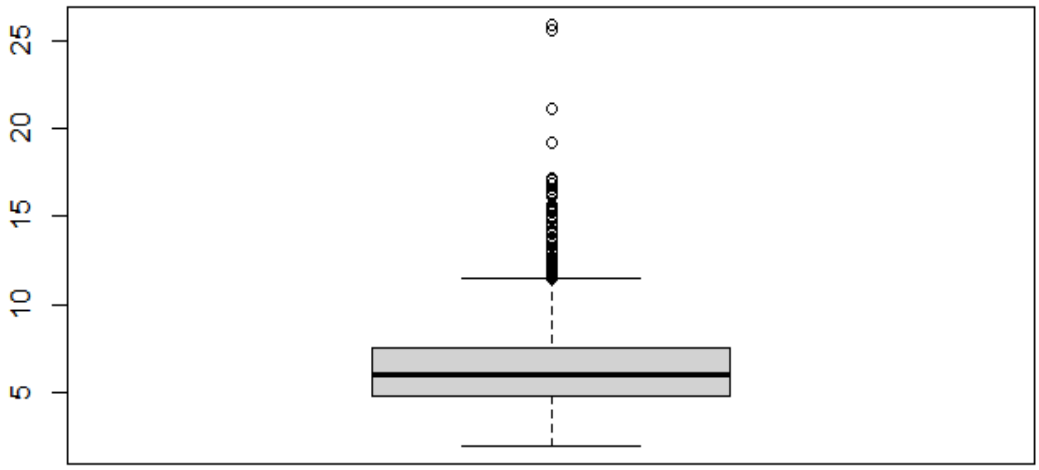
The average pressure box plot has a smaller interquartile range, falling in between 29.5 to 30.0 and with the rest of its values being outliers. This means that the common recorded pressure in Pasay over the past decade is at 29.5 to 29.7

Average Wind Gust



In the given result in the box plot, it can be seen that the most common value of recorded wind gusts is zero which makes the box plot being skewed close to zero. This indicates that Pasay rarely records any wind gust that is above 5 mph.

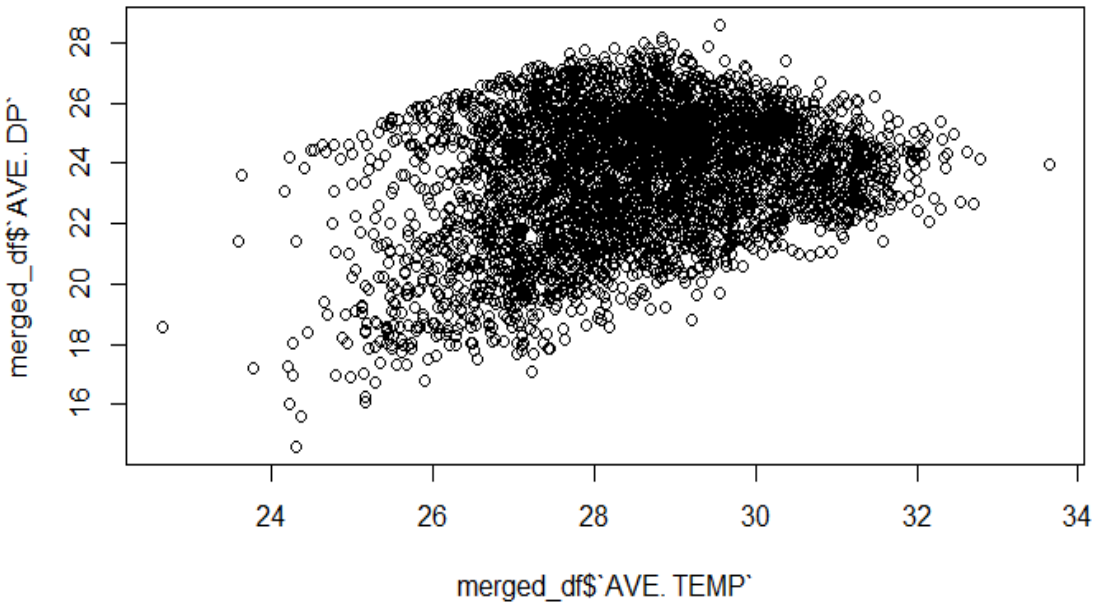
Average Wind Speed



The overall average wind speed in Pasay is observed to be rather slow, evident in the shown box plot in which the recorded range is between 0 to 12 mph. The common recorded speed however is in the range of 5 to 7 mph. There

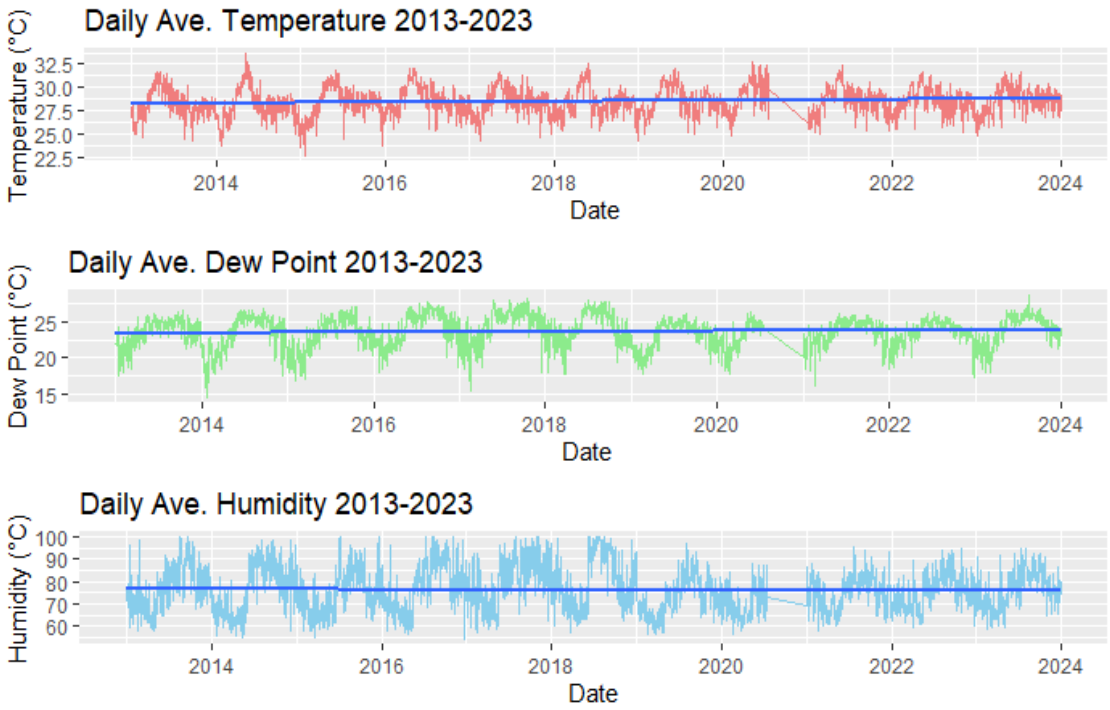
are however recorded cases where the wind speed is slightly above average which can be seen with the present outliers, with the highest being at 25 mph or above which can imply aberrant weather cases.

Scatter Plots



Based on the given scatter plot between the average dew point and average temperature, it can be observed that the two attributes follow a positive correlation as the point values of the dew point increases as the average temperature increases. With this, it can be inferred that the temperature experienced in Pasay is affected by its dew point and vice versa. Upon looking more on the data, there is a vast cluster of points in the middle which makes up around the range of 27 to 29 degrees of temperature with a variation of dew points with it peaking at around 28 degrees.

Time-Series Plots



The following time series plots can be further analyzed:

Daily Average Temperature

In the given data, it can be observed that the daily average temperature became less and less varied over the years with the most amount of variations can be seen during the 2014 to 2015 time period where it also peaked at around 32.5 and beyond and also has the recorded low point of 22.5 degrees. As the years go by, the average temperature can be seen to steady between 25 to 30 degrees with the most common daily temperature sitting at 27 to 27.5 degrees. It can also be noted that the following graph seems to follow a wave-like pattern which means the temperature changes based on seasons. This trend can be observed further when looking at each start of the year having a relatively low temperature and fluctuates as time goes by. Looking at the regression line, we see that the values as the dates go by, are steadily increasing.

Going into depth with the time-series plot of the daily average temperature, it can be observed that the time plot fluctuates regularly as the year changes and a norm that is seen is that the temperature for that year is usually at its highest around the start to middle. This can signify a hotter season which is usually around March to May. Beyond that, the temperature then dips at its average temperature around 27.5 degrees and goes lower as the year ends. Here, it can also be seen the highest recorded average temperature, being at above 32.5 and the lowest being slightly above 22.5 which is all recorded during mid-2014 to early 2015.

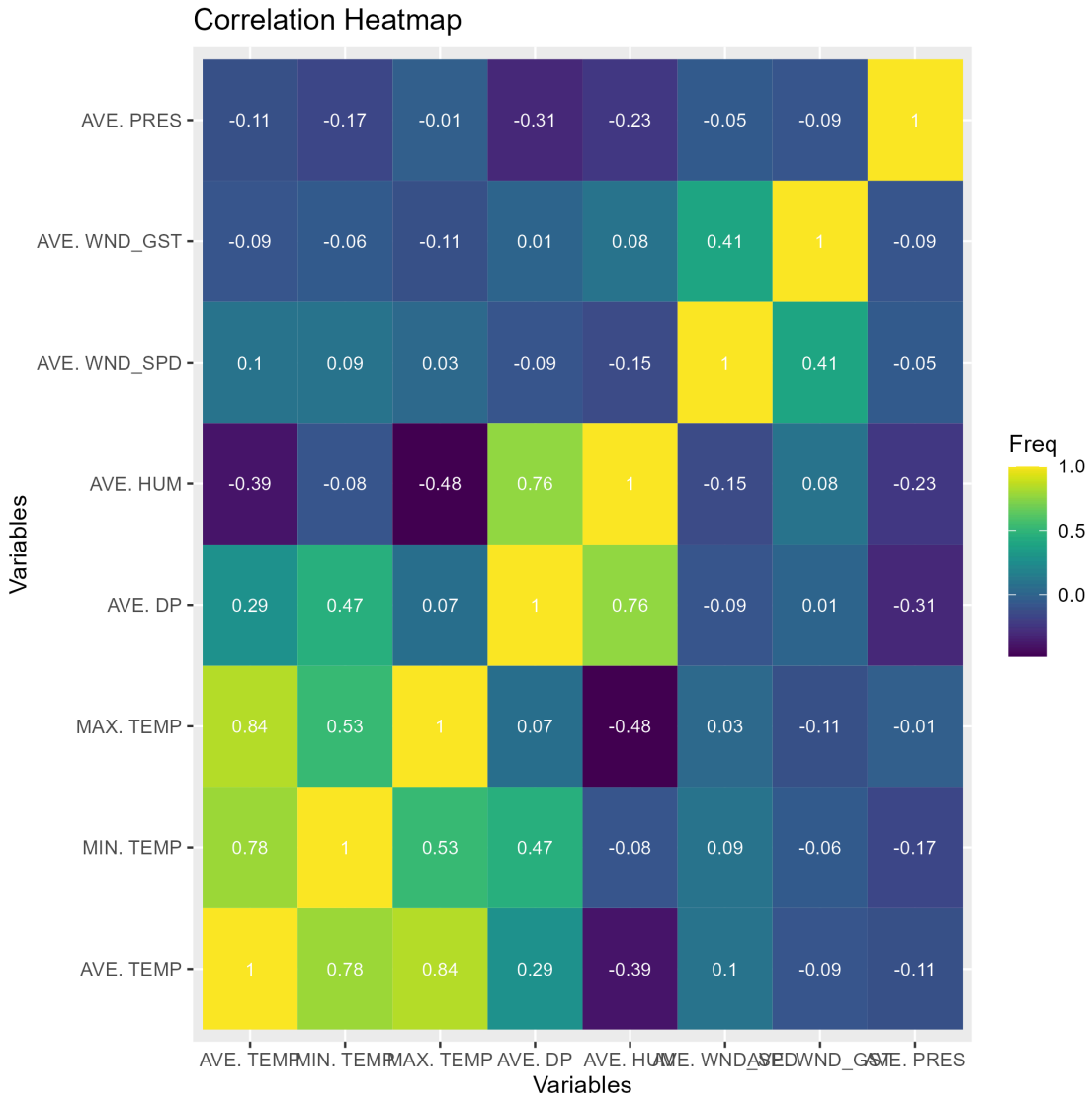
Daily Average Dew Point

The dew point has had a more consistent trend over the years with there being a noticeable wave-like pattern. At the start of the year the dew point is usually at its lowest point with the lowest recorded value being at 15 degrees which was in 2014. The highest recorded dew point was actually during 2023 to 2024 with the value going over 30 degrees. The lowest amount of fluctuations seen in the dew point data can be observed during 2021 to 2023 with the average dew point being set around the median point while also having a relatively low peak and dip.

Daily Average Humidity

It is observable in the past years that the average level is much higher as compared to recent years. There is also a foreseeable pattern in which the humidity level is at its highest during the middle of the given year and usually has its lowest point at the start of each year. Around mid-2018 is where the humidity levels of Pasay peaked, around 100 percent and the lowest being below 60 degrees during 2014. In the most recent years of 2021 to 2024 it can be observed that the humidity levels are less varied and comparatively has less fluctuations.

Correlation Heatmap



Upon looking at the heatmap we can view various correlations each attribute has to each other. Based on this, it is expected that the minimum and maximum temperature has a correlation to the average temperature. There is also a positive correlation with the average dewpoint and temperature with frequency values being close to 0.5, which is backed up by the results of the scatter plot. Similar to this, the average humidity and dew point also have a positive correlation as the nature of both deal with the moisture. It is also

noticeable that the average humidity and temperature has the darkest color, meaning they do not have any sort of correlation. This also applies to the average pressure with it being the next darkest. Among all of the attributes, the average pressure has the least amount of correlations with the other present attributes as all of its values are negative. This implies that the average pressure is an independent factor and is not affected with other environmental attributes.

C. Data Preparation

Data Scraping

The python script used to scrape the data makes use of packages such as selenium, webdriver_manager, and pandas. Selenium and webdriver_manager functions are used to automate interactions with the web browser, in this case, chrome. Given a range of dates with startDate and endDate, and the base URL for the website, the code accesses the URL for each day and parses the html to access the table with the hourly observations. It is then appended to the temps.csv file, created using the create_csv_file.py script.

Data Preprocessing

In preprocessing, the researchers made changes to the raw dataset to increase overall consistency, and handle noise present in the data.

1. Minor changes: Changing the column names to shortened versions, removed the units(°F, %, mph, in) appended to the values and as they were originally of chr type, they were also converted to numeric values (double) as it opens up to more options for preprocessing/visualization later on.
2. Handling null/incorrect values: The TEMP, DP, and HUMcolumns had at most 375 zero values which were considered as incorrect values and

converted to NA values. These instances were replaced by the mean value of their respective dates.

Ex: The computed mean value of Temperature for 2019-01-13 (without the 0-value row) was 81.3913. This value was used to replace the NA Temperature value at 2019-01-13 10:00:00

The WND_DIR column also had 342 null values but the case study will not be using the column so its null values were not handled.

3. Reducing data: As this case study is more concerned with daily observations rather than hourly, the average value of the numerical values for each day were computed and were merged into a new dataframe, along with the respective dates. This reduced the rows from 97,421 to 4,015.
4. Handling missing dates/rows: In the Date Completeness section of the Rmd file, the researchers checked for possible missing dates in the dataset and found out that it had two missing days: 2020-07-15 and 2020-07-16. Upon checking, there was no available data from the website which caused this mishap. These instances were replaced by the mean value of their respective months.

   Ex: The computed mean value of `AVE.TEMP` for 2020-07 was 85.67882. This value was used for the 2020-07-15 row.
5. Converting Fahrenheit to Celsius: As the Philippines uses mainly the metric system (*Memorandum Circular No. 1173, S. 1979*, 1979), Celsius included, the original Fahrenheit values of the dataset were converted to Celsius.
6. Dropping Columns: It was observed that the `AVE. PRECIP` column only had 0 values even if the COND column from the raw dataset had values of “rain”. As this suggests inconsistency and there is no way to determine values to replace it, this column was dropped.

The preprocessed dataset now contains 4017 instances and 11 columns with the following data dictionary:

DATE	Date when observation was collected (YYYY-MM-DD)
MIN. TEMP	Lowest temperature recorded during the whole day (°C)
MAX. TEMP	Highest temperature recorded during the whole day (°C)
AVE. TEMP	Average temperature for the whole day (°C)
AVE. DP	Average dew point for the whole day (°C)
AVE. HUM	Average relative humidity for the whole day (%)
AVE. PRES	Average temperature for the whole day (in.)
AVE. WND_GST	Average wind gust for the whole day (mph)
AVE. WND_SPD	Average wind speed for the whole day (mph)

Feature Selection

The researchers conducted multivariate methods for modeling, requiring careful selection of attributes or variables. While temperature and precipitation are commonly utilized indicators of extreme weather, the absence of precipitation data in the dataset necessitates the exploration of alternative options. The "AVE. HUM" column emerged as a promising candidate due to its negative correlation (-0.39) with "AVE. TEMP". Attributes highly correlated with "AVE. TEMP", such as "MIN. TEMP" and "MAX. TEMP", were deemed redundant for the second variable. Both temperature and humidity significantly contribute to weather-related risks. Elevated humidity exacerbates the severity of heatwaves by impeding the body's cooling mechanism through sweating. Consequently, hot and humid conditions pose greater risks compared to equally hot but dry environments (Steadman, 1979). Wibisono et al.'s research further validates this perspective, highlighting anomalous weather patterns characterized by high humidity and low temperature.

Therefore, the final choice for modeling attributes was "AVE. TEMP" and "AVE. HUM”.

D. Modeling

The researchers used five algorithms for comparison. Namely, k-Nearest Neighbors (k-NN), One Class Support Vector Machine (OCSVM), Density-based spatial clustering of applications with noise (DBSCAN), Local Outlier Factor, and Random Forest. Each member was assigned one of these listed algorithms to be implemented using R via RStudio. With that in mind, the model shall use the preprocessed data .csv file to gather the data with the columns “AVE. TEMP” and “AVE. HUM” used as the basis for the outlier detection. Given this, the different models are set to have various thresholds and parameters in order to get the best outcome which is compared by the assigned programmer to present in the final comparison. This is done in order to gauge the best possible performance of the anomaly detection algorithm and ensure that the results are consistent. Each of the models have a presented scatter plot in order to show the 2D visualization of the detected anomalies against the normal data points. Alongside this, additional visualization is provided in the form of histograms and density plots.

K-Nearest Neighbors Algorithm (k-NN)

The K-Nearest Neighbors (k-NN) algorithm was used for anomaly detection within the dataset due to its simplicity and effectiveness in identifying outliers based on their proximity to neighboring data points. To implement k-NN, essential packages such as kknn and FNN, which are crucial for k-NN computations and distance calculations, respectively, were installed. The dataset, which had been preprocessed and stored in a .csv file, was then loaded into a data frame for analysis, ensuring that it was in a format suitable for k-NN processing. Two specific columns, "AVE. TEMP" and "AVE. HUM", were selected

as they are deemed important indicators of anomalous weather conditions. Prior to training the model, parameters critical to the k-NN algorithm, such as the number of nearest neighbors (k) and the anomaly detection threshold, were predefined. For this study, k was set to 10, indicating that each data point's classification would be based on its ten nearest neighbors, while the threshold for anomaly detection was set to 1. The model was subsequently trained using the loaded dataset and specified parameters, leveraging the capabilities of the kknn package. During training, distances between instances were calculated utilizing the FNN package, and anomaly scores were generated using the apply() function. Instances exceeding the predefined threshold were flagged as anomalies, enabling their identification and cataloging for further analysis. This comprehensive methodology ensured that the k-NN algorithm was effectively employed for anomaly detection within our dataset, providing insights into potentially irregular data points indicative of anomalous weather conditions.

One Class Support Vector Machine (OCSVM)

One-Class Support Vector Machine (OCSVM) algorithm was used to detect anomalies within the dataset, chosen for its ability to identify outliers in new data based on a model trained with historical data. Initially, after loading the preprocessed dataset, essential R packages such as readr, ggplot2, dplyr, and e1071 were installed to support the analysis. To align with OCSVM's methodology, which relies on historical data for anomaly identification in new data, the dataset underwent trimming to encompass only the first 1000 rows for model training. This segmentation facilitated the differentiation between historical and new data, with the initial 1000 entries designated as historical data, and the subsequent 3000+ entries considered as potentially new data. Subsequently, the dataset was filtered to retain only the 'AVE.TEMP' and 'AVE.HUM' attributes, alongside the 'DATE', essential variables for weather anomaly detection. Following this step, a one-class Support Vector Machine (SVM) model was trained using the filtered dataset, enabling it to discern the unique characteristics

of the historical data. Finally, the entire dataset underwent prediction by the trained OCSVM model, simplifying the identification of anomalies based on deviations from learned patterns. This structured approach provided a systematic methodology for utilizing OCSVM in detecting anomalies within the dataset, offering valuable insights into potentially irregular data points indicative of anomalous weather occurrences.

Density-based spatial clustering of applications with noise (DBSCAN)

The Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was utilized to detect clusters of arbitrary shapes within the dataset, while accounting for noise and outliers. The implementation relied on the .dbscan package for DBSCAN clustering and ggplot2 and plotly for visualization, following a structured methodology. Initially, essential libraries were loaded, and the dataset was imported into a data frame. Subsequently, the variables 'AVE.TEMP' and 'AVE.HUM' were selected from the dataset for analysis. To ensure uniformity and compatibility with DBSCAN, the data underwent scaling. DBSCAN clustering was then performed, employing an epsilon value (eps) of 0.3 and a minimum points value (MinPts) of 10. The clustering results were organized into a clustered data frame, facilitating further analysis and interpretation. Notably, anomalies, representing noise or outliers, were identified based on their cluster assignments. Finally, data visualization techniques were employed to portray the clustered data effectively. Through plots generated using ggplot2 and plotly, the spatial distribution of clusters within the dataset was visually represented. This visualization aided in gaining insights into the composition and structure of the dataset, revealing patterns and clusters of interest.

Local Outlier Factor (LOF)

The Local Outlier Factor (LOF) algorithm was used to detect outliers within the dataset. Initially, the preprocessed dataset was loaded into a tibble or data frame and recognizing the significance of scaling data for density-based algorithms like LOF, features such as Average Temperature and Average Humidity were standardized using the scale() method. This ensured that both features contributed proportionally to calculations by maintaining consistent scales. Subsequently, LOF scores were computed using the lof() method from the Rlof package. Specifically, the 3rd and 5th columns of the scaled dataset, representing 'AVE. TEMP' and 'AVE. HUM' respectively, were employed for calculation, with a chosen value of k or MinPts set to 10, in accordance with recommendations by Breunig et al. (2000). An iterative process was utilized to determine the optimal threshold for identifying outliers. Initially set at 1.5, the threshold was refined to 1.4 after assessing the distribution of LOF scores through histograms and density plots. Instances with LOF scores surpassing the threshold were designated as outliers and cataloged in a separate data frame for further examination.

Random Forest (RF)

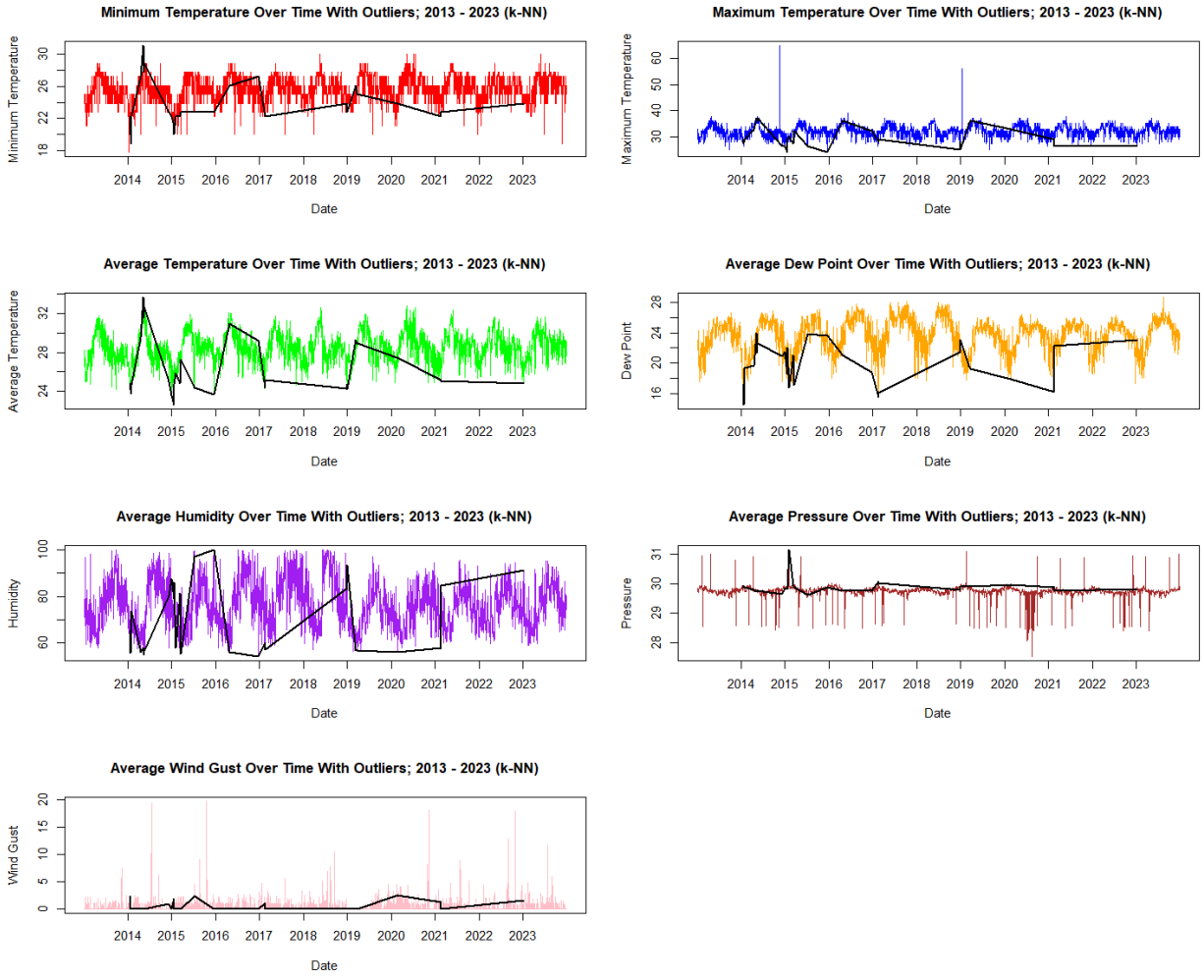
The Random Forest (RF) algorithm was used to detect anomalies within the dataset, aiming to visually represent these anomalies. Employing packages such as randomForest for modeling, dplyr for data manipulation, lubridate for date and time management, and plotly for interactive plotting, the researchers implemented a systematic approach to anomaly detection. Numeric columns underwent scaling using the scale() function to standardize the data range, a crucial step for machine learning algorithms sensitive to scale. Subsequently, the Random Forest model was trained with the proximity measure enabled, which assesses how often two points end up in the same leaf node, an important aspect for anomaly detection. The model focused primarily on variables such as 'Average Temperature' and 'Average Humidity'. Anomaly scores were computed using the proximity measure, with lower scores indicating potential anomalies. To

identify anomalies, a threshold was set at the 99th percentile of the anomaly scores, categorizing data points exceeding this threshold as anomalies. Additionally, data filtering by date was implemented by converting the 'DATE' column to a year format and filtering the data to include records starting from the year of the first detected anomaly. This comprehensive methodology enabled the researchers to effectively utilize Random Forest for anomaly detection and visualization within their dataset, providing insights into potentially irregular data points.

RESULTS AND DISCUSSION

Five anomaly detection algorithms were evaluated using different methods such as visual inspection, cross-validation and density estimation. Each algorithm has a unique turnout of anomalies detected. The visual inspection method aimed to see if the anomalies identified by each algorithm show distinctness between the plots of the anomalous and non-anomalous data. The OCSVM and Random Forest algorithm had anomalies present within the normal clusters but OCSVM had a more clustered, stratified and multitude of plots compared to the Random Forest. The k-NN, LOF, and DBSCAN however had a more reasonable spread on the plot as the anomalies were located from the normal plot points, indicating that these three were more accurate with identifying anomalies. The algorithms produced density peaks at different intervals of the average temperature although the k-NN algorithm produced a symmetric anomalous KDE graph, while the rest produced structures that were bimodal at least. Given their different average temperature intervals for each of the density peaks, the evaluation of the algorithms' Kernel Density Estimation alone could not identify which of the five algorithms performed the best. In the cross-validation of results of all the algorithms, the researchers compared the results to see if they have similarities between their results. The Random Forest did not identify any anomalies commonly with the other four algorithms. The OCSVM algorithm identified a large number of anomalies but only identified five common anomalies. The LOF, k-NN, and DBSCAN had the most common anomalies. The researchers computed a cross-validation score by adding the number for each no. of common anomalies and multiplying them by a weight (x1 for 2, x2 for 3, x3 for 4) as the more algorithms that have identified the anomaly, the more likely it is a real one. The last column consists of the cross-validation score from the computation for each algorithm. The k-NN algorithm had the highest score of 47 which indicated that its results are likely to be true anomalies as they were also identified by other algorithms. Given these results from the evaluation methods performed, the k-Nearest Neighbors (k-NN) algorithm emerged as the most effective algorithm.

The results of the chosen algorithm, k-NN, was then used to provide insights on the weather anomalies found within the years 2013 to 2023 in the Philippines (Pasay City). This study has identified thirty-three (33) anomalies out of 4,017 total data points, having a 0.82% outlier percentage that suggests that most of the weather in the Philippines are rather typical or normal. Most of these anomalies were found in the earlier years, twelve (12) of which were in the year 2014. The month of January had the most anomalies with a count of eleven (11), unsurprisingly, seven (7) of which were specifically in January 2014. These periods might be potential periods for increased vigilance or further investigation of possible extreme weather patterns.



Upon inspection of the graphs of the detected anomalies or outliers, it is observed that the year with the most number of detected anomalies (2014) also holds

the highest recorded temperatures in all three levels: minimum, average, and maximum. The runner-up year in terms of detected anomalies (2015), however, does not hold the second-highest record of temperature levels. This defeats the notion implying that the number of detected anomalies is directly correlated with the temperature levels. By extension, it also implies that using the average temperature feature alone as x-value for the models may have not produced the most accurate predictions.

On another note, it is observable that the three highest months in terms of detected anomalies are January, December, and February. And the distribution of detected anomalies further decreases as the years approach June. This suggests that weather anomalies are more likely to occur and be detected during the colder seasons than the hotter seasons.

CONCLUSION

With the evaluation and comparison of the results produced by the five anomaly detection algorithms, which are LOF, k-NN, OCSVM, DBSCAN, and RF, it is determined from the discussion of the results that the k-NN algorithm performed the best and possibly the most accurate among the five algorithms. Although the other algorithms were fairly competent for the task, it was evident in this case that OCSVM and Random Forest were relatively inaccurate when it came to anomaly detection. As with the other two, DBSCAN and LOF, proved to be viable options for this kind of data mining task however it is shown in this study that k-NN was slightly better when it came to results. The strength of k-NN lies in its ability to leverage the similarity of historical weather data points within the context of KDE-estimated density. By comparing a new data point to its k-nearest neighbors and their positions within the density distribution, k-NN can effectively identify significant deviations from typical weather patterns, highlighting the algorithm's optimal performance in anomaly detection.

In correspondence with the garnered results, the study was able to identify thirty-three (33) anomalies out of the 4,017 data points in the data set which spanned from 2013 to 2023 recorded in Pasay City. The total outlier percentage yielded a 0.82% anomaly rate which means that the overall data from the weather reports were normal and proved to be accurate. It was also discovered that the most amount of anomalies detected were in the year 2014, specifically in January with it containing twelve (12) of the anomalies out of 33. This result was in line with the initial inspection of the time-series plots in which most of the data spikes occurred during 2014. This means that during 2014 was when the weather conditions were abnormally higher than usual, making it an outlier from what the country usually experiences.

Overall, it can also be concluded that the k-Nearest Neighbors (k-NN) algorithm is the most effective method for detecting weather anomalies. It balances the identification of unique and common anomalies while maintaining high accuracy. The findings underscore k-NN's robustness and reliability in anomaly detection, providing

valuable insights into historical weather patterns in the Philippines. Given also its results, it can be concluded that the detected anomalies are either signs of abnormally high or low temperatures and humidity levels for the country in a given year.

RECOMMENDATIONS

This study is open for future revisions and assessments by future researchers tackling the same topic. It is recommended that further implementation and comparison of the five algorithms be performed to further understand and solidify insights acquired from this study. It is also recommended to find datasets that can provide true labels of anomalies for the algorithms to be evaluated more comprehensively using measures such as precision, accuracy, and more. If these labels are not available, it is recommended that the future researchers delve deeper in evaluating the results using domain knowledge like referencing publications, cross-referencing with real world weather events (e.g. typhoons, El Niño), or consulting subject matter experts. It is also recommended that in the case of performing the same analysis on areas with significant precipitation (e.g. rainy seasons), the future researchers be able to procure a dataset including the said attribute. Such future studies could pursue this advancement by utilizing a different dataset from the same discipline, or possibly a dataset from a different field of study altogether. Furthermore, it is also recommended that more anomaly detection algorithms be included in further comparisons to test against the best algorithm identified by this study.

REFERENCES

Dobos, D., Nguyen, T. T., Dang, T., Wilson, A. C., Corbett, H., McCall, J., & Stockton, P. (2023). A comparative study of anomaly detection methods for gross error detection problems. *Computers & Chemical Engineering*, *175*, 108263. https://doi.org/10.1016/j.compchemeng.2023.108263

Felbermayr, G., Grösch, J., Sanders, M., Schippers, V., & Steinwachs, T. - The economic impact of weather anomalies. ScienceDirect. https://doi.org/10.1016/j.worlddev.2021.105745

How is climate change affecting the Philippines? (2016, January 19). The Climate Reality Project. https://www.climaterealityproject.org/blog/how-climate-change-affecting-philippines

Hong, J., Agustin, W., Yoon, S., & Park, J. (2022). Changes of extreme precipitation in the Philippines, projected from the CMIP6 multi-model ensemble. Weather and Climate Extremes. Advance online publication. https://doi.org/10.1016/j.wace.2022.100480

Kulkarni, K. ., Mahale, Y. ., Khan, N. ., K., N. ., & Gite, S. . (2024). Deep Learning for Anomaly Detection in Spatio- Temporal Maharashtra Weather Data: A Novel Approach with Integrated Data Cleaning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 12(12s), 169–182. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/4502

Memorandum Circular No. 1173, s. 1979. (1979, November 12). Official Gazette. Retrieved April 3, 2024, from

https://www.officialgazette.gov.ph/1979/11/12/memorandum-circular-no-1173-s-1979/

Oliveros JM, Vallar EA, Galvez MCD. Investigating the Effect of Urbanization on Weather Using the Weather Research and Forecasting (WRF) Model: A Case of Metro Manila, Philippines. Environments. 2019; 6(2):10. https://doi.org/10.3390/environments6020010

PAGASA. (n.d.). https://www.pagasa.dost.gov.ph/information/climate-philippines

PAGASA. (n.d.). https://www.pagasa.dost.gov.ph/information/climate-change-in-the-philippines

Parimala, V. K. (2024). Introductory chapter: Anomaly Detection – Recent Advances, AI and ML Perspectives and Applications. In *Artificial intelligence*. https://doi.org/10.5772/intechopen.113968

Pierobon, G. (2023, September 5). K-Nearest Neighbors (KNN) for Anomaly Detection - Gabriel Pierobon - Medium. Medium. https://medium.com/@gabrielpierobon/k-nearest-neighbors-knn-for-anomaly-detection-d9bcc2d4f71a

Sania. (n.d.). *GitHub - sania111/Anomaly-Detection-in-Weather-Data-Using-LSTM-Autoencoder*. GitHub. https://github.com/sania111/Anomaly-Detection-in-Weather-Data-Using-LSTM-Autoencoder

Society, W. D. S. (2023, November 22). Stranger weather ahead: Detecting Anomalies in temporal weather data. *Medium*.

https://medium.com/@WDSS/stranger-weather-ahead-detecting-anomalies-in-temporal-weather-data-9630eae33ecf

Steadman, R. G. (1979). The assessment of sultriness. Part I: A temperature-humidity index based on human physiology and clothing science. Journal of Applied Meteorology, 18, 861–873.

Tuychiev, B. (2023, November 28). *A Comprehensive Introduction to anomaly Detection*. https://www.datacamp.com/tutorial/introduction-to-anomaly-detection

Wibisono, S. R., Anwar, M. T., Supriyanto, A., & Amin, I. H. A. (2021). Multivariate weather anomaly detection using DBSCAN clustering algorithm. *Journal of Physics: Conference Series*, *1869*(1), 012077. https://doi.org/10.1088/1742-6596/1869/1/012077

World Bank Climate Change Knowledge Portal. (n.d.). https://climateknowledgeportal.worldbank.org/country/philippines/climate-data-historical

APPENDIX A

List of Figures

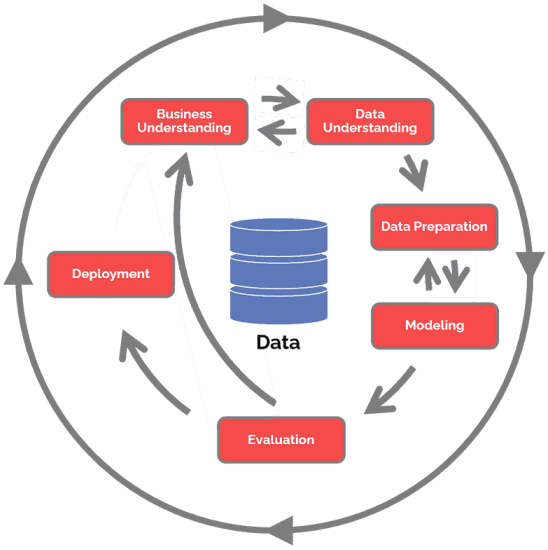


Figure 1. CRISP-DM Model

(dito muna yung dataset)

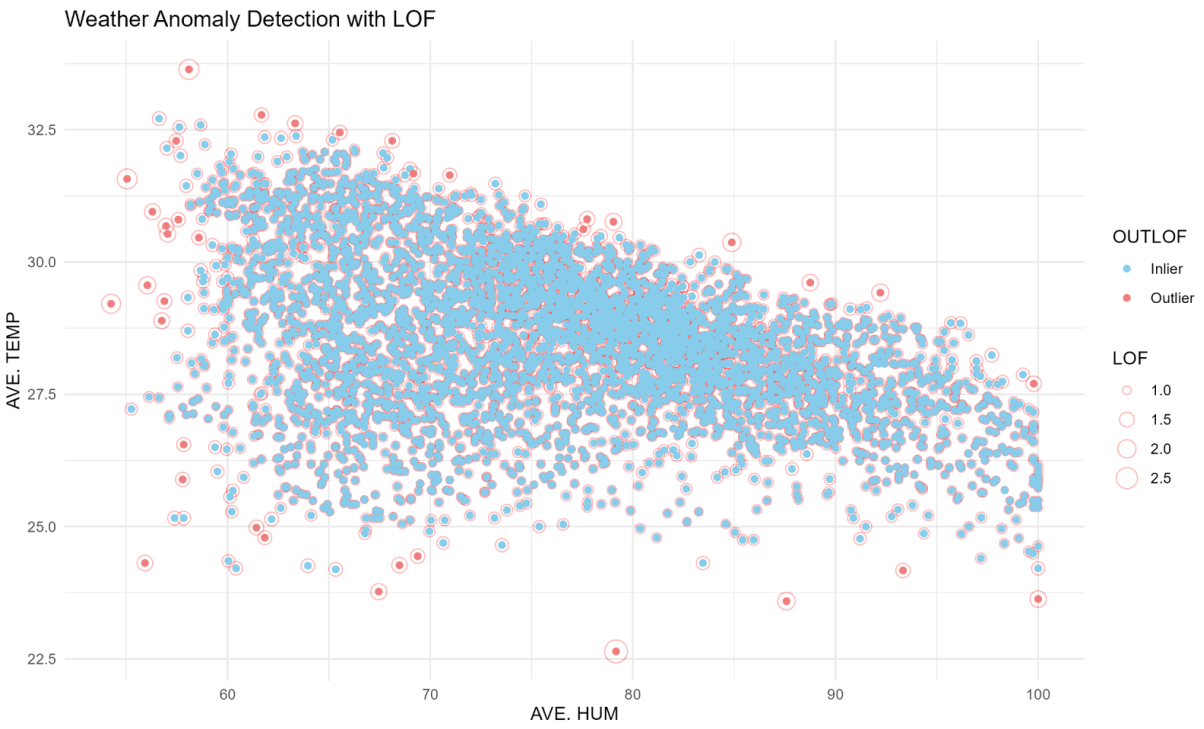


Figure 2. Visual inspection using Local Outlier Factor (LOF)

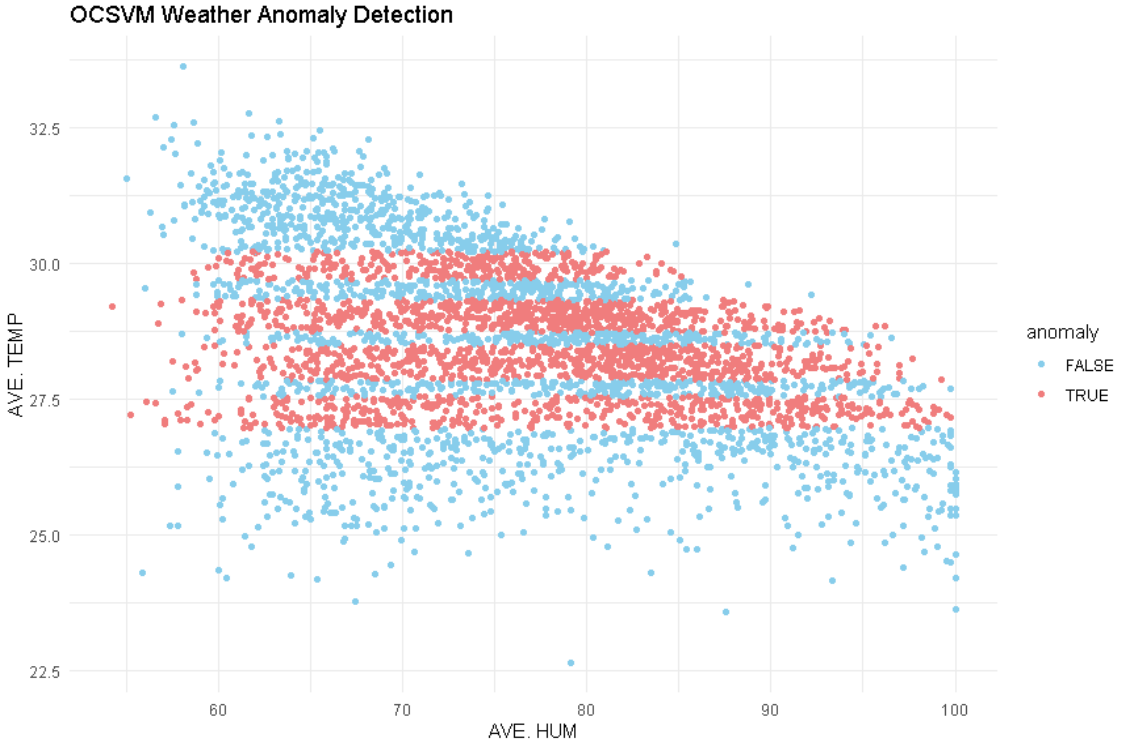


FIgure 3. Visual inspection using One Class Support Vector Machine (OCSVM)

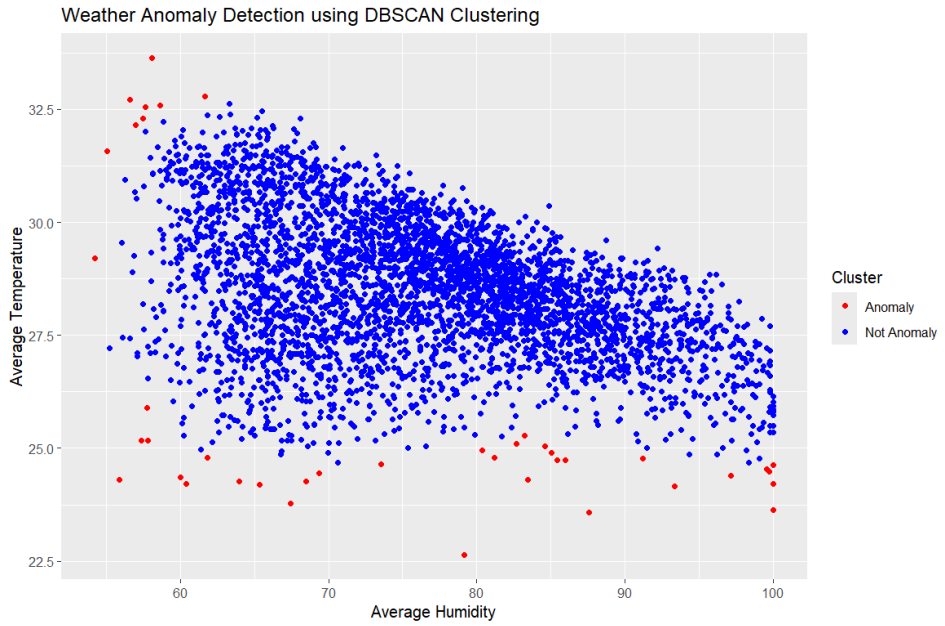


Figure 4. Visual inspection using Density-based Spatial Clustering of Applications with Noise (DBSCAN)

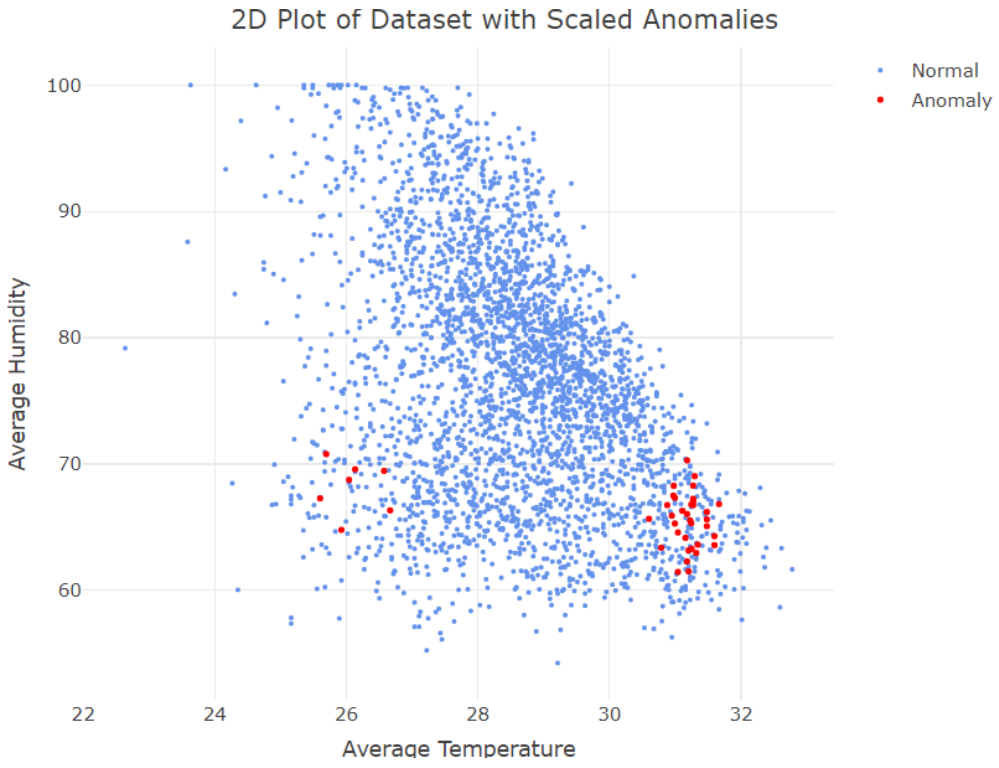


Figure 5. Visual Inspection using Random Forest

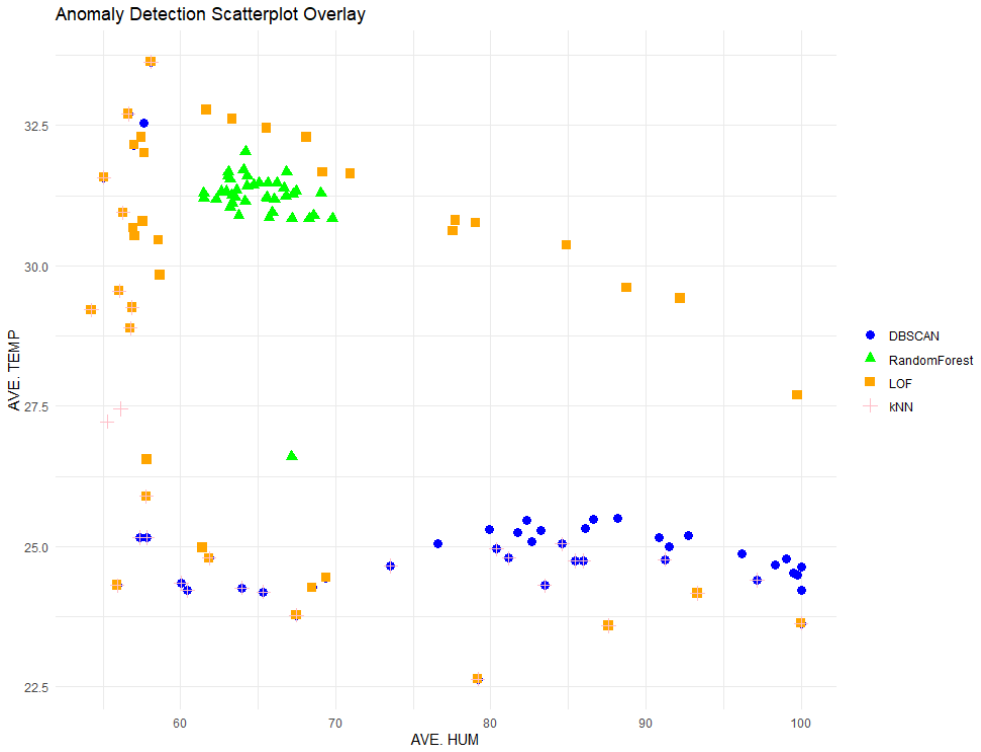


Figure 6. Combined Scatter plots of Algorithms

	Anomaly_Score	Z_Score	P_Value
1	0.42429714	1.010523591	0.3122445027
2	0.23111439	-0.171838399	0.8635645767
3	0.22740926	-0.194515419	0.8457723240
4	0.27199898	0.078392948	0.9375154827
5	0.20104318	-0.355887248	0.7219250061
6	0.19000737	-0.423431166	0.6719807099
7	0.13310832	-0.771677963	0.4403051856
8	0.26184319	0.016235146	0.9870467970
9	0.26894047	0.059673555	0.9524156346
10	0.21149812	-0.291898508	0.7703642243

Figure 7. Anomaly Scores, Z-Score and P-Value of k-NN results

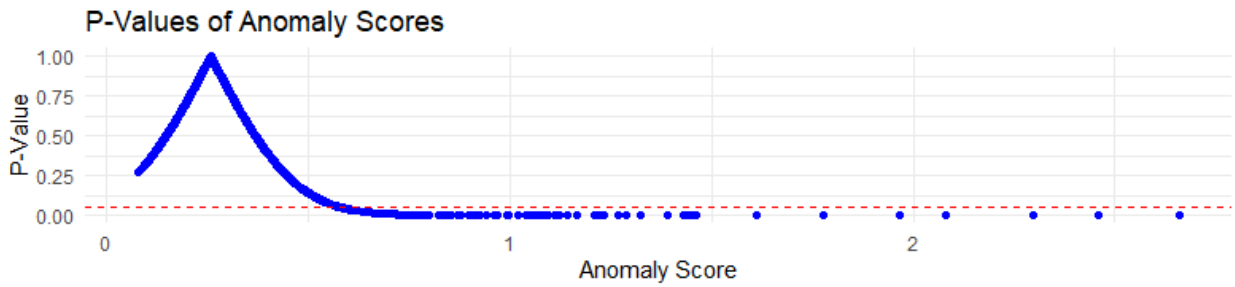


Figure 8. P-Value graph of k-NN

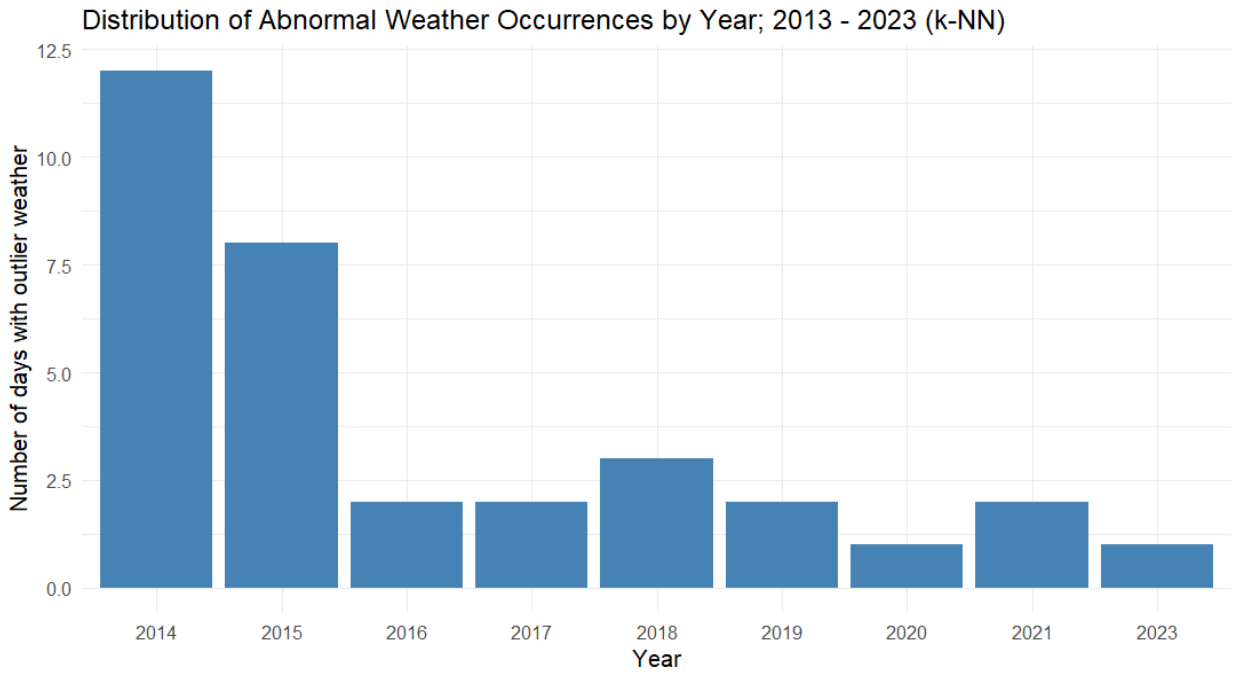


Figure 9. Histogram of outliers in k-NN based on year

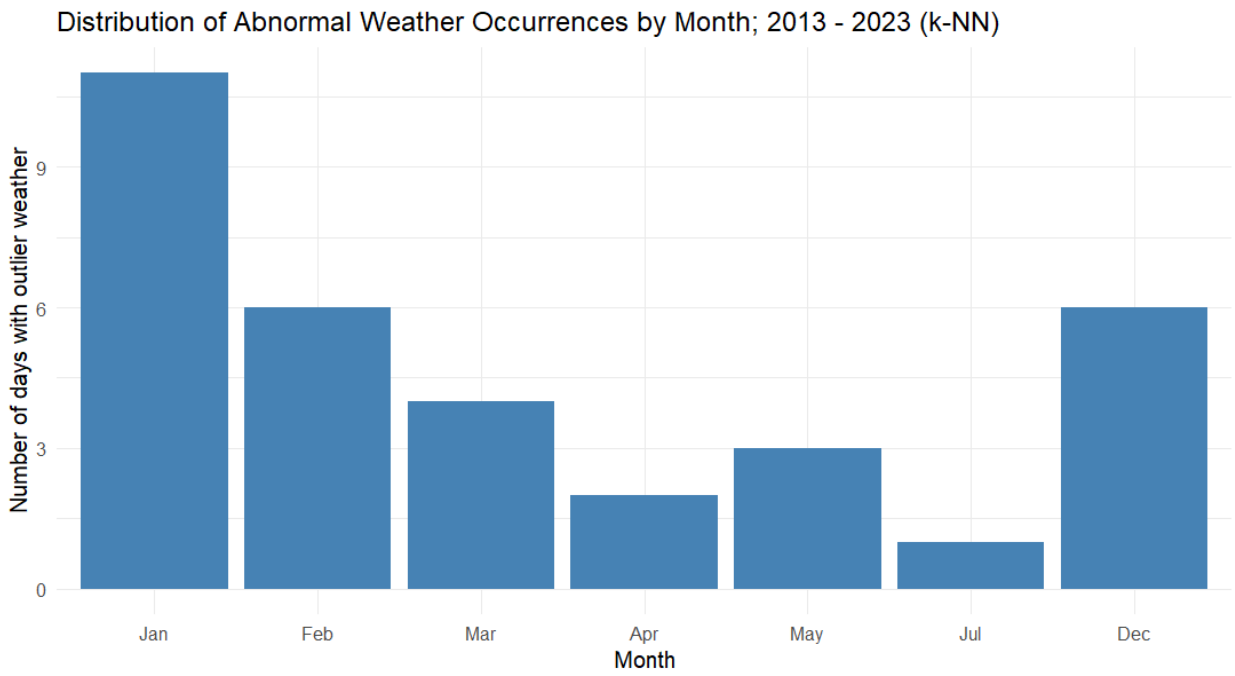


Figure 10. Histogram of outliers in k-NN based on months

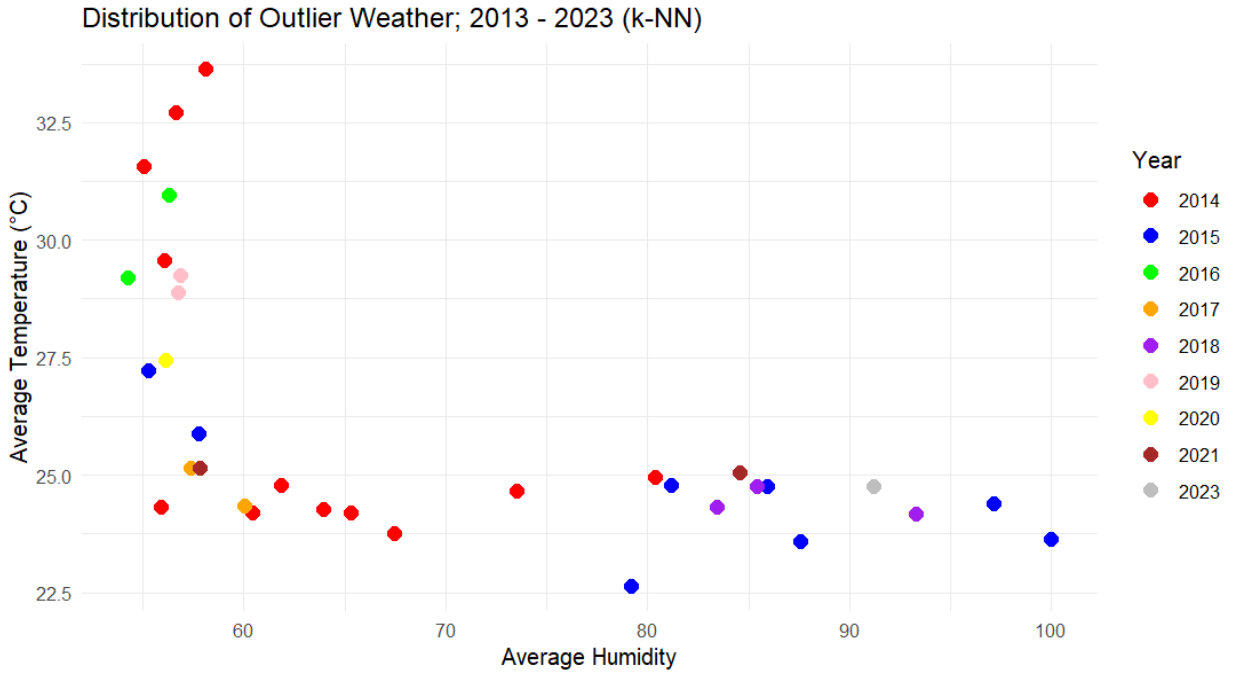


Figure 11.Plot data of detected weather outliers over the years

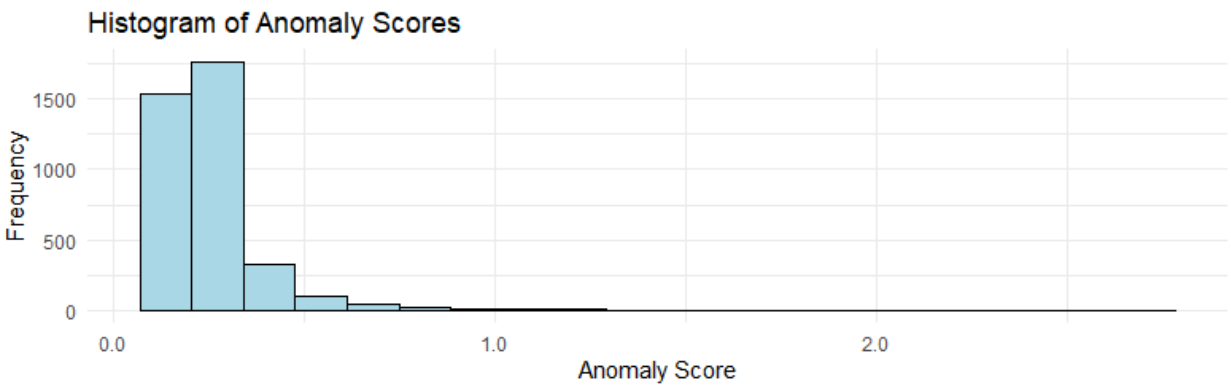


Figure 12. Histogram of Anomaly Score in k-NN

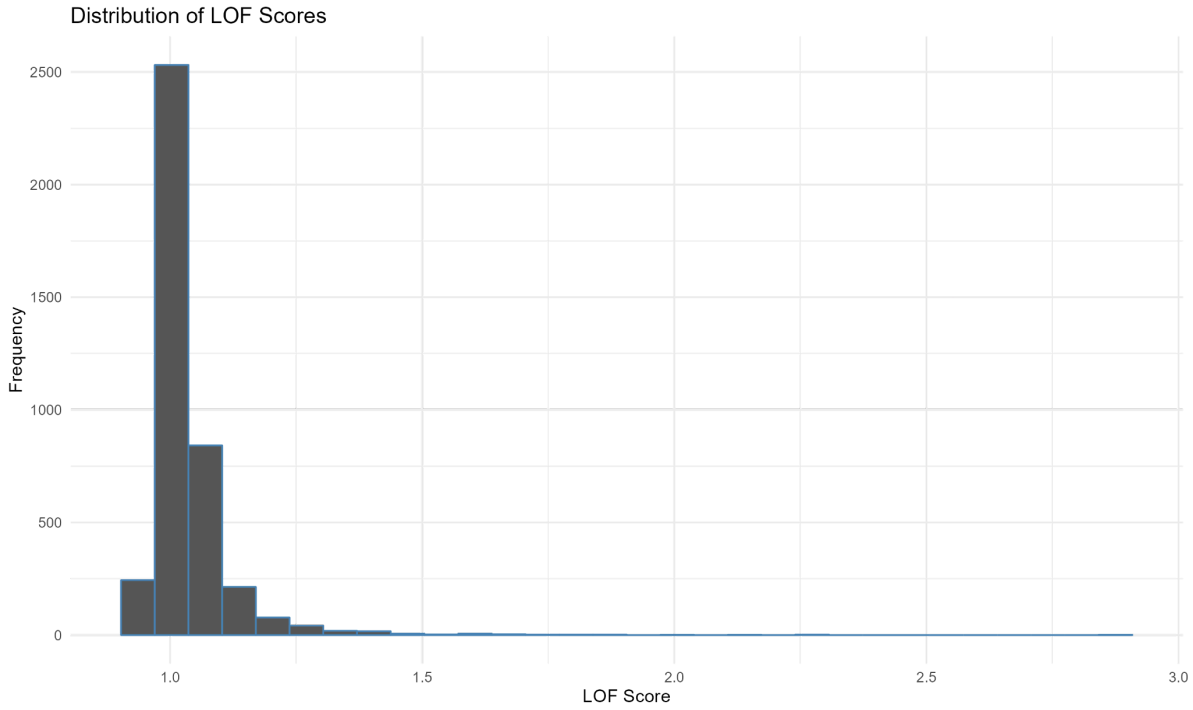


Figure 13. Histogram of LOF Scores based on Frequency

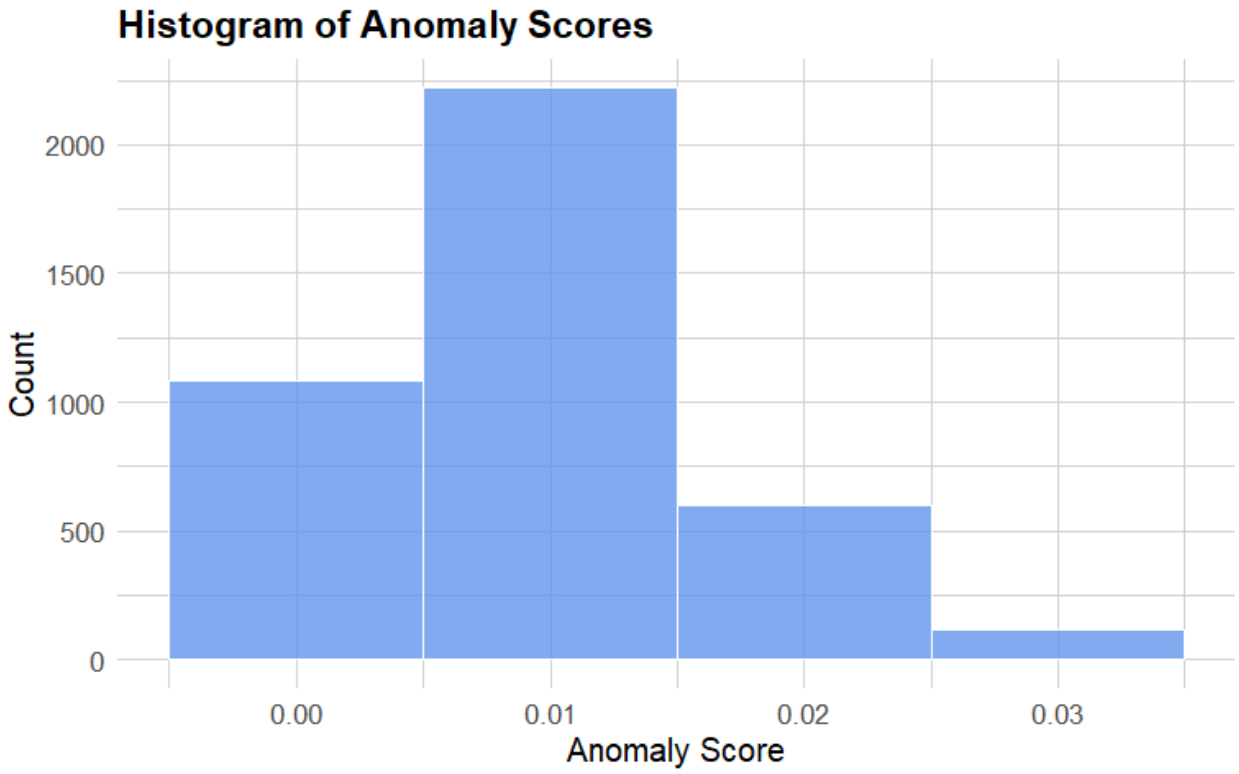


Figure 14. Histogram of Anomaly Scores of Random Forest

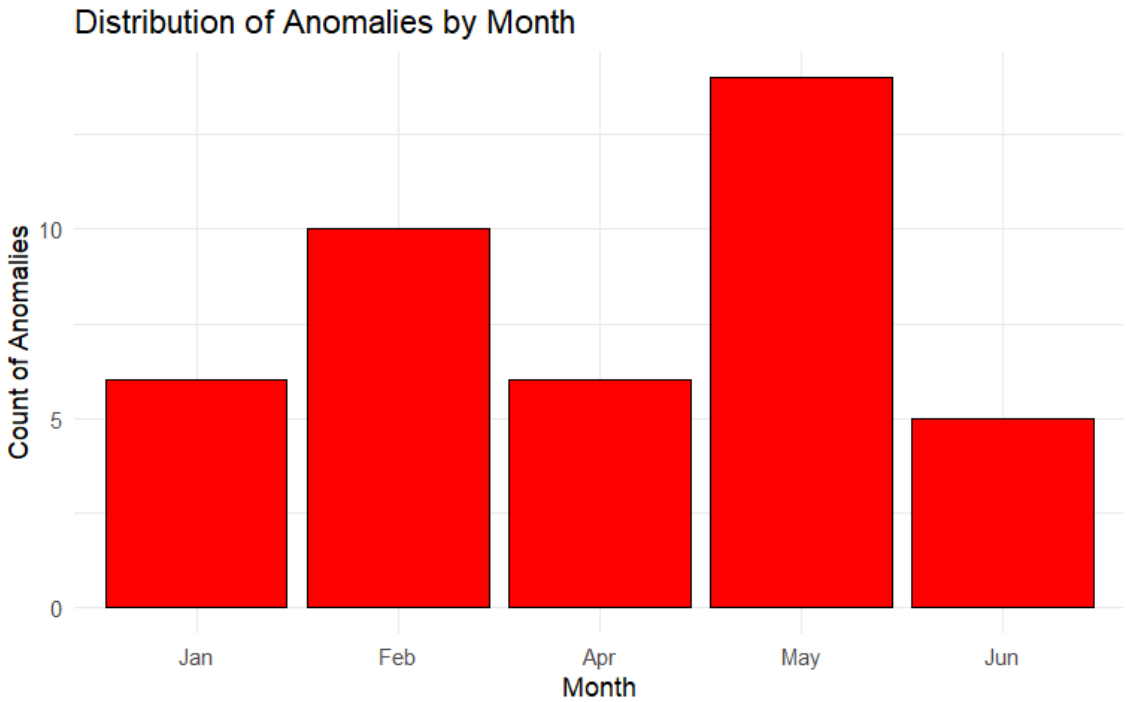


Figure 15. Histogram of Anomalies in Random Forest by Month

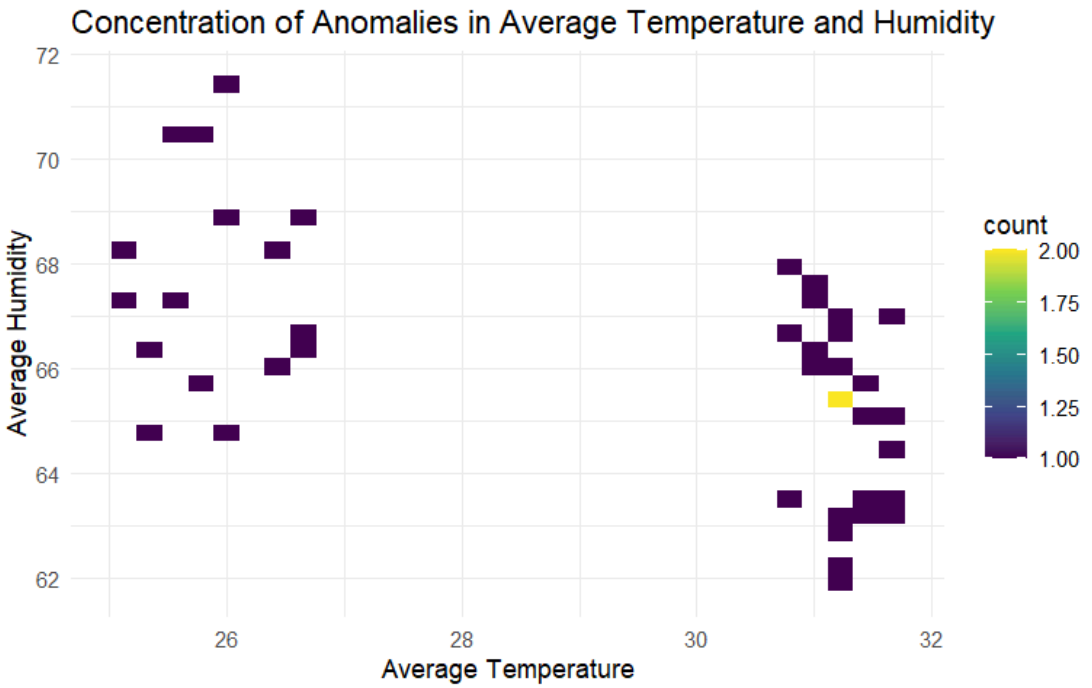


Figure 16. Correction Heatmap of Random Forest

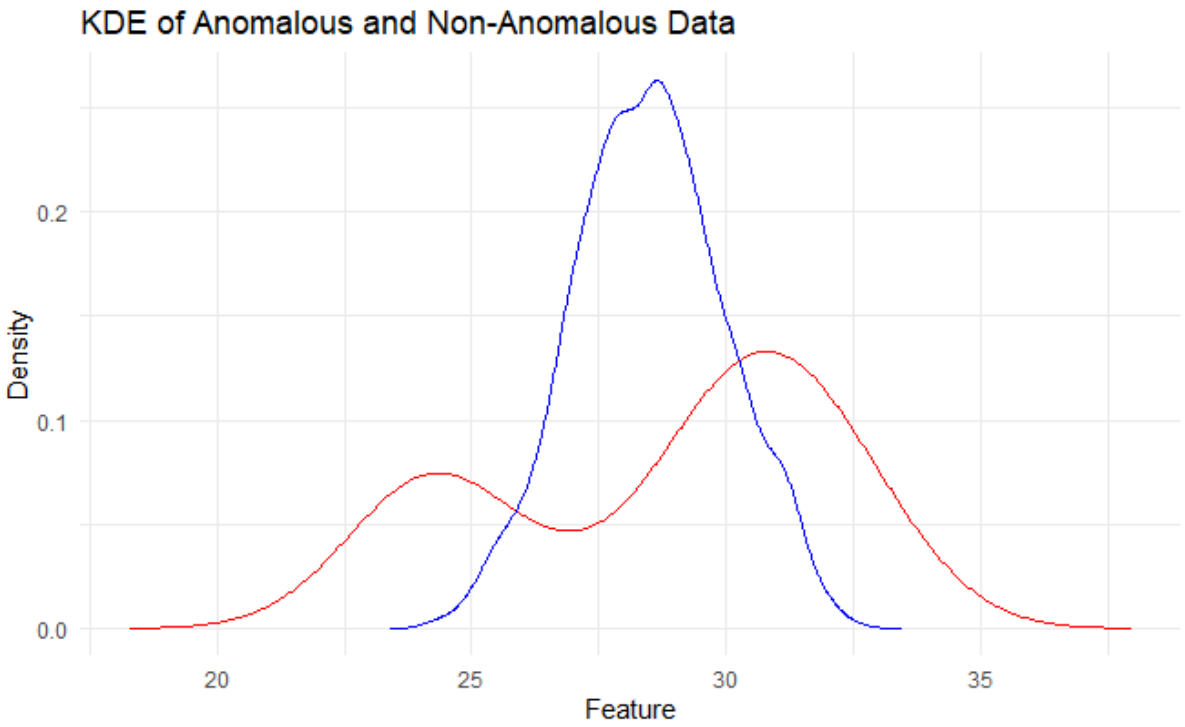


Figure 17. KDE of k-NN

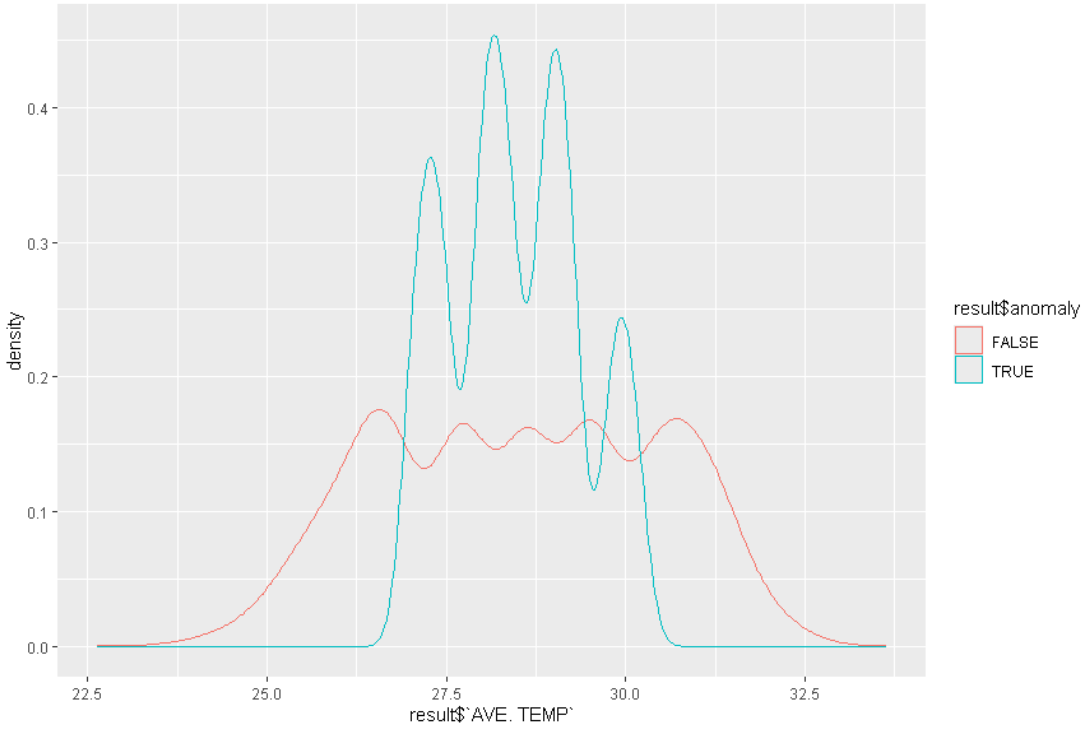


Figure 18. KDE of OCSVM

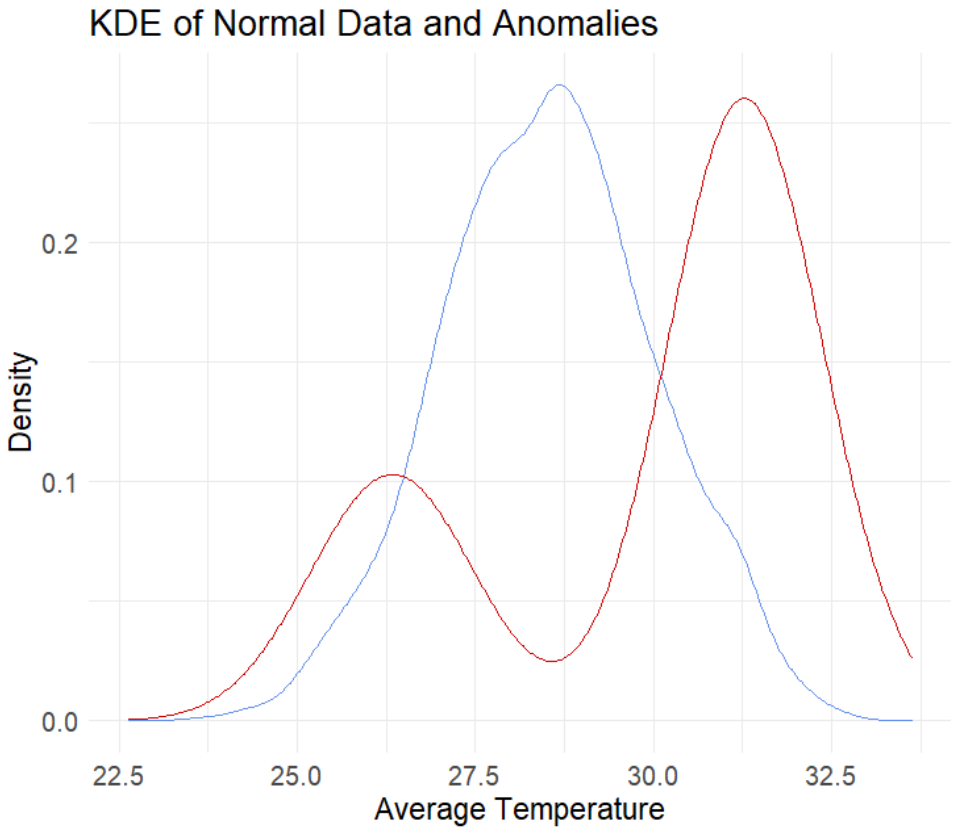


Figure 19. KDE of Random Forest

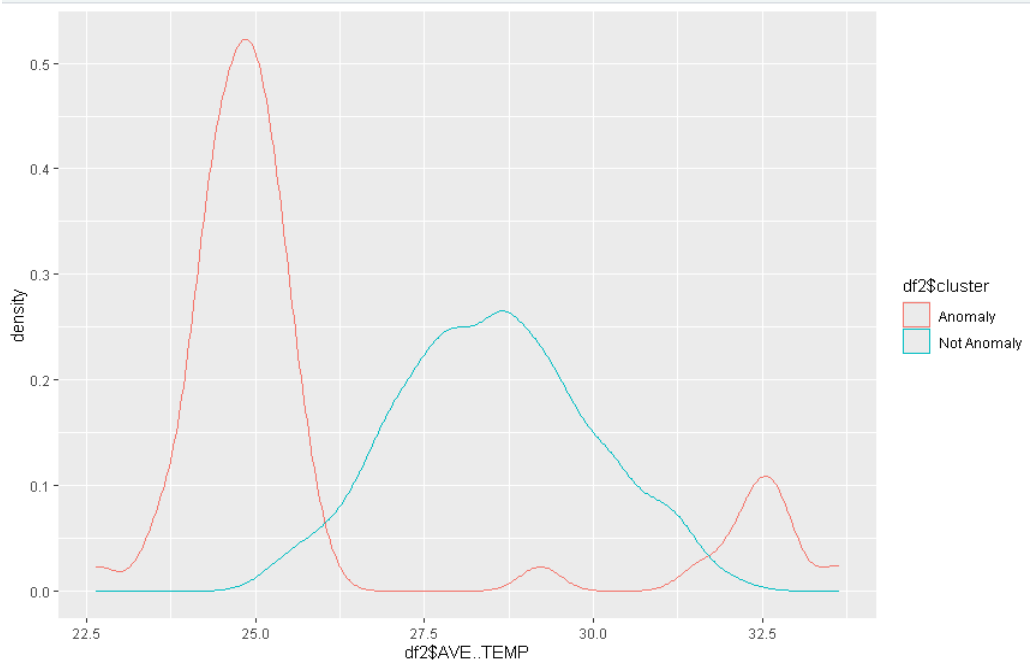


Figure 20. KDE of DBSCAN

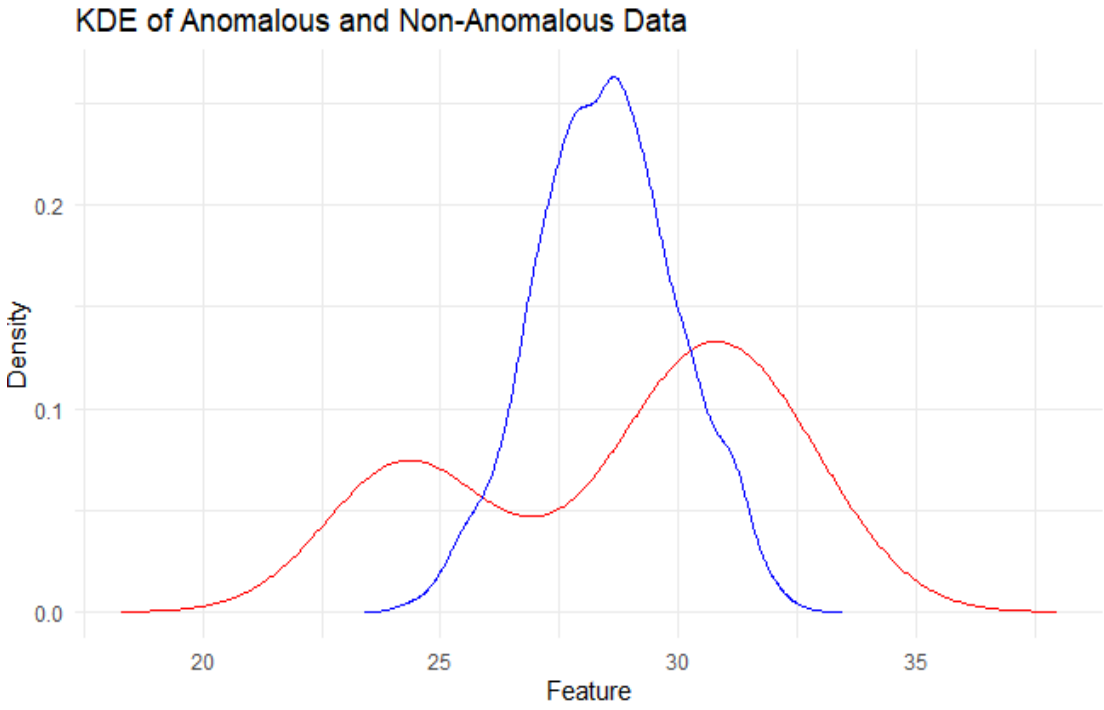


Figure 21. KDE of LOF

APPENDIX B

List of Tables

Matrix of Objectives, Methodology, and Expected Outputs

Objective Matrix		
Objectives	***Methodology***	***Expected Outputs***
Collect, prepare, and understand the data for the modeling.	- Business Understanding - Data Collection (web scraping from wundergound.com) - Data Pre-Processing - Exploratory Data Analysis - Finalization of dataset / Feature selection	- Raw dataset of web scraped data (csv) - Pre-processed dataset (csv) - Final features or dataset - EDA Visualizations and Interpretations
Conduct the modeling for anomaly detection using the defined algorithms	- Finalize algorithms to be utilized - Conduct the experiments/train the models using the different algorithms	- List of algorithms - Modeling specifications (parameters, algorithm descriptions, etc.) - Visualizations produced by the algorithms showing the anomalies detected
Compare the results of the different algorithms using key performance measures.	- Perform evaluation methods on each algorithm - Formulate a conclusion	- Comparative Analysis Table of the applied algorithms - Visualizations of the evaluations of each algorithm - Conclusion on best algorithm
Identify insights on the outliers of the weather data.	- Analysis of the visualizations	Presentation and discussion of the following information:

		- Percentage of outliers over total instances - Months with most number abnormal weather occurrences - Patterns/trends the outliers show - Tabulation of detected anomalies

No. of scores of cross-detection

No. of scores of cross-detection						
Algorithm Used	1	2	3	4	**Total Anomalies Detected**	**Cross-validation score:**
Random Forest	41	0	0	0	41	0
OCSVM	1984	2	2	1	1989	9
LOF	18	6	12	1	37	33
k-NN	0	20	12	1	33	47
DBSCAN	10	20	10	1	41	43
Total cross-detection	2053	24	12	1	2,141	

Summary of KDE Graph Attributes from the Five Anomaly Detection Algorithms

Model Used	**Anomalous KDE Graph Structure**	**Number of Density Peaks**	**Density of Highest Peak**	**Density of Lowest Peak**
K-NN	Symmetric	1	~0.26 (27.5 - 28.5)	~0.26 (27.5 - 28.5)
LOF	Bimodal	2	~0.13 (30 - 31)	~0.08 (24.5 - 25.5)
OCSVM	Multimodal	4	~0.45 (28.0)	~0.24 (29.5 - 30.0)

DBSCAN	Multimodal*	3	~0.52 (24.0 - 25.0)	~0.11 (32.5)
RF	Bimodal	2	~0.22 (31.25)	~0.1 (26.25)

* - DBSCAN's anomalous data KDE graph is multimodal with three density peaks, but earlier analysis only considered two.

Tabulation of detected anomalies with complete attributes (k-NN)

DATE	MIN..TEMP	MAX..TEMP	AVE.TEMP	AVE..DP	AVE.HUM	AVE..PRES	AVE..WND_GST	AVE..WND_SPD
2014-01-19	22.22	27.22	24.21	16.02	60.42	29.94	2.29	8.83
2014-01-20	21.11	27.22	24.19	17.25	65.33	29.91	0.00	5.38
2014-01-23	22.22	27.78	24.79	16.99	61.83	29.88	0.00	7.17
2014-01-24	22.22	27.78	24.26	16.94	63.96	29.88	2.00	8.92
2014-01-25	20.00	28.89	24.31	14.58	55.92	29.87	0.00	6.63
2014-01-26	18.89	27.78	23.77	17.20	67.46	29.87	0.00	5.17
2014-01-27	22.22	28.89	24.65	19.38	73.54	29.88	0.00	5.38
2014-04-15	27.22	32.78	29.56	19.69	56.04	29.78	0.00	7.00
2014-05-08	31.11	37.22	33.64	23.94	58.09	29.72	0.00	7.27
2014-05-13	27.78	36.11	31.57	21.39	55.04	29.75	0.00	7.17
2014-05-19	28.89	37.22	32.71	22.66	56.63	29.76	0.00	6.88
2014-12-08	22.78	26.11	24.96	20.98	80.36	29.64	0.96	11.68
2015-01-01	22.22	26.11	23.59	21.44	87.58	29.85	0.00	3.29
2015-01-18	21.11	23.89	22.64	18.54	79.17	29.81	1.88	8.71
2015-01-19	20.00	28.89	24.75	22.01	85.96	29.83	0.00	4.71
2015-02-02	22.22	30.00	25.89	16.76	57.78	31.14	0.00	7.74
2015-03-11	22.22	27.22	24.79	21.06	81.17	29.94	0.00	6.04
2015-03-17	22.78	32.22	27.22	17.11	55.25	29.86	0.00	8.00
2015-07-06	22.78	26.11	24.40	23.87	97.17	29.60	2.25	5.92
2015-12-19	22.78	23.89	23.63	23.63	100.00	29.88	0.00	4.23
2016-04-23	26.11	36.11	30.95	21.06	56.29	29.77	0.00	7.13
2016-12-25	27.22	32.22	29.21	18.80	54.25	29.75	0.00	10.17
2017-02-14	22.22	27.78	24.35	15.60	60.04	30.08	1.04	8.67
2017-02-15	22.22	28.89	25.16	16.06	57.38	30.02	0.00	9.38
2018-12-29	23.89	25.00	24.31	21.44	83.46	29.78	0.00	5.67

2018-12-30	22.78	27.78	24.75	21.99	85.42	29.84	0.00	3.92
2018-12-31	22.78	25.00	24.17	23.08	93.33	29.89	0.00	2.54
2019-03-14	26.11	35.00	29.26	19.63	56.88	29.89	0.00	8.58
2019-03-26	25.00	36.11	28.89	19.21	56.75	29.90	0.00	8.25
2020-02-21	23.89	32.78	27.45	17.85	56.13	29.97	2.50	9.79
2021-02-20	22.22	28.89	25.16	16.27	57.83	29.86	1.17	6.71
2021-02-22	22.78	26.11	25.05	22.27	84.58	29.75	0.00	3.04
2023-01-05	23.89	26.11	24.77	23.10	91.21	29.80	1.46	4.83