Data Exploration and Preparation

Table of Contents

Introduction

In this report, we will explore and analyse a weather dataset provided by a potential client, with the goal of extracting key insights and presenting them in a manner that can inform future decision-making. The dataset includes various weather-related attributes such as temperature, rainfall, wind speed, and humidity, recorded across multiple locations over a certain period. The purpose of this report is to carry out an initial data exploration, apply data preprocessing techniques, and summarise the key attributes of the dataset. Additionally, the report will aim to identify any patterns, trends, and associations between the attributes.

The analysis is carried out using KNIME, an open-source data analytics platform, due to its flexibility in handling large datasets and its wide range of tools for data exploration, transformation, and visualisation. In addition to KNIME, we will also be utilising Excel to facilitate data manipulation and exporting summary statistics for reporting. The dataset consists of a CSV file, containing weather data recorded over various locations and times. The focus of this report will be to consolidate the dataset into a comprehensive dataset and then explore it through statistical analysis and preprocessing techniques.

A – Initial Data Exploration

A1 - Attribute Type

Before we dive into the attribute types for our dataset, let's list down the main types of data:

- **Nominal:** Nominal data is categorical and represents labels or categories without any quantitative value or inherent order. The categories in nominal data are distinct, and you cannot perform arithmetic operations on them. Each value in a nominal variable is a name or label, and the only operation that can be done is to count the frequency of occurrences in each category.

- **Ordinal:** Ordinal data is categorical but with a defined order or ranking between the categories. However, the intervals between the values are not necessarily consistent or meaningful. You can rank or sort ordinal data, but you cannot perform arithmetic operations on it beyond comparison.

- **Interval:** Interval data is quantitative and has a consistent interval between values, but there is no true zero point. This means that while the differences between values are meaningful, you cannot multiply or divide the values. Zero in an interval scale does not mean the absence of the quantity being measured.

- **Ratio:** Ratio data is also quantitative but has both consistent intervals and a true zero point, meaning that zero represents the complete absence of the quantity being measured. This allows for meaningful arithmetic operations, including addition, subtraction, multiplication, and division. With ratio data, ratios between values are meaningful (e.g., "twice as much").

Now, let's categorise our data into those main attribute types:

Attribute	Type	Reason
Date	Interval	Dates represent specific points in time, and the intervals between dates are consistent. Even though there is no meaningful zero, the differences between dates can be calculated meaningfully.
Location	Nominal	Location is a categorical variable that refers to places such as cities or regions. There is no inherent order to the locations, and the categories are

			used solely for identification purposes. Arithmetic operations cannot be performed on location names.
MinTemp		Ratio	Minimum temperature is measured from a true zero point (absolute zero) on a consistent scale, making arithmetic operations like addition, subtraction, multiplication, and division meaningful.
MaxTemp		Ratio	MaxTemp is also measured from a true zero point, which allows for meaningful comparisons of the magnitude of differences. You can calculate ratios between temperatures.
Rainfall		Ratio	Rainfall is measured continuously starting from zero, representing no rainfall. The scale allows for arithmetic operations like calculating total rainfall over time or comparing rainfall levels between locations.
Evaporation		Ratio	Evaporation is measured from zero upwards, where zero represents no evaporation. The data can be added, subtracted, and compared across different time periods or locations. Since evaporation has a true zero point, it is ratio data.
Sunshine		Ratio	Sunshine, typically measured in hours, has a true zero point (no sunshine), and the intervals between values are consistent. Arithmetic operations are meaningful, allowing for calculations of total sunshine over a period of time or comparison between different regions.
WindGusDir		Nominal	Wind direction is categorised using compass points (e.g., North, South, East). These categories represent distinct values but do not imply any order or allow for meaningful arithmetic operations. Therefore, WindGusDir is nominal data.
WindGusSpeed		Ratio	Wind gust speed is measured from zero upwards and represents a continuous variable. The zero point indicates no wind, and the intervals between values are consistent, making arithmetic operations meaningful. This classifies wind speed as ratio data.
9am	Temp	Ratio	The temperature at 9 am is measured on a continuous scale with a true zero point. Differences between temperatures and ratios of temperatures can be calculated, making this ratio data.
	Humidity	Ratio	Humidity is a continuous variable expressed as a percentage, where 0% represents no humidity. Because the intervals between percentages are consistent and ratios are meaningful, humidity is considered ratio data.
	Cloud	Ordinal	Cloud cover is measured in eighths, with values that can be ranked in order (e.g., 1/8, 2/8), but the intervals between values may not be consistent. This makes it ordinal data, as there is a sense of order but not meaningful arithmetic operations.
	WindDir	Nominal	Wind direction at 9 am is a categorical variable with no inherent order. Compass points (e.g., North, South) are distinct categories that don't imply any meaningful arithmetic relationships or ordering, making this nominal data.
	WindSpeed	Ratio	Wind speed is measured on a continuous scale, starting from zero, and the differences between values are meaningful. This allows for arithmetic operations such as addition and multiplication, classifying wind speed as ratio data.
	Pressure	Ratio	Atmospheric pressure at 9 am is measured continuously from zero, with consistent intervals between values. The true zero allows for meaningful arithmetic operations, making pressure a ratio variable.
	Temp	Ratio	The temperature at 3 pm follows the same principles as other temperature variables, with a true zero point and consistent intervals between values. This makes it ratio data, as meaningful comparisons and calculations can be made.
	Humidity	Ratio	Like humidity at 9 am, the 3 pm humidity value is expressed as a percentage on a continuous scale. Since 0% indicates no humidity and the intervals are consistent, it is treated as ratio data.

3pm	Cloud	Ordinal	Cloud cover is measured in eighths, which allows for ranking but not meaningful arithmetic operations. This makes cloud cover an ordinal variable, as it can be ordered but not precisely measured in intervals.
	WindDir	Nominal	Wind direction at 3 pm is a categorical variable represented by compass points, with no inherent order or numerical meaning. Therefore, it is classified as nominal data.
	WindSpeed	Ratio	Wind speed at 3 pm is measured continuously from zero, allowing for meaningful arithmetic operations and comparisons between values. This classifies wind speed as ratio data.
	Pressure	Ratio	Atmospheric pressure at 3 pm is measured from zero and upwards, allowing for meaningful arithmetic operations and consistent intervals between values. This makes it ratio data.
RainToday		Nominal	RainToday is a binary categorical variable indicating "Yes" or "No" for whether it rained. There is no inherent order or numerical meaning between these categories, making it nominal data.
RainTomorrow		Nominal	RainToday is a binary categorical variable indicating "Yes" or "No" for whether it rained. There is no inherent order or numerical meaning between these categories, making it nominal data.

A2 – Summarising Properties for All Attributes

For this section, we'll conduct a thorough exploration of the dataset by summarising the key properties of each attribute. This involves calculating various descriptive statistics based on the attribute types. For nominal attributes, we will compute frequency distributions to identify the most and least common categories. For ordinal attributes, we will determine the median, most common values, and percentiles to understand how cloudiness is distributed throughout the dataset. For ratio attributes, we will calculate statistical measures such as the mean, median, range, standard deviation, and skewness to explore the central tendencies, spread, and shape of the distributions. This analysis will help identify trends, potential outliers, and any irregularities in the data, providing a strong foundation for further analysis in the subsequent sections of the report.

Date

Frequency & Distribution Data Visualisation:

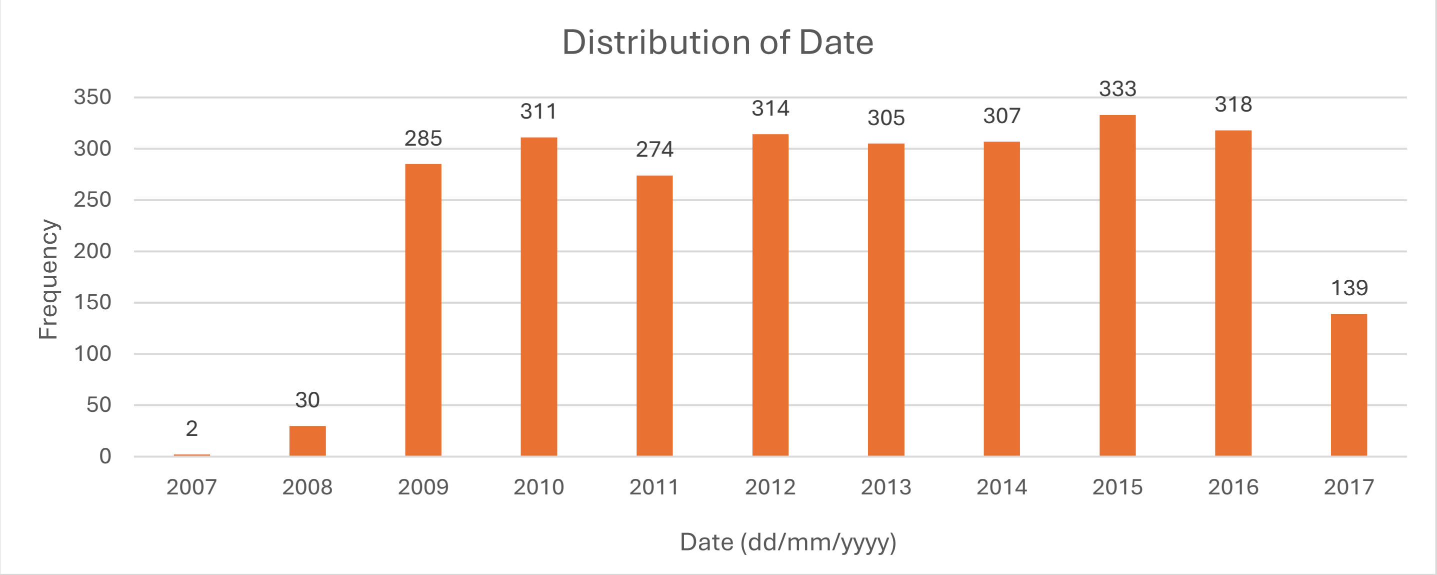


Figure 1 – Histogram of Date

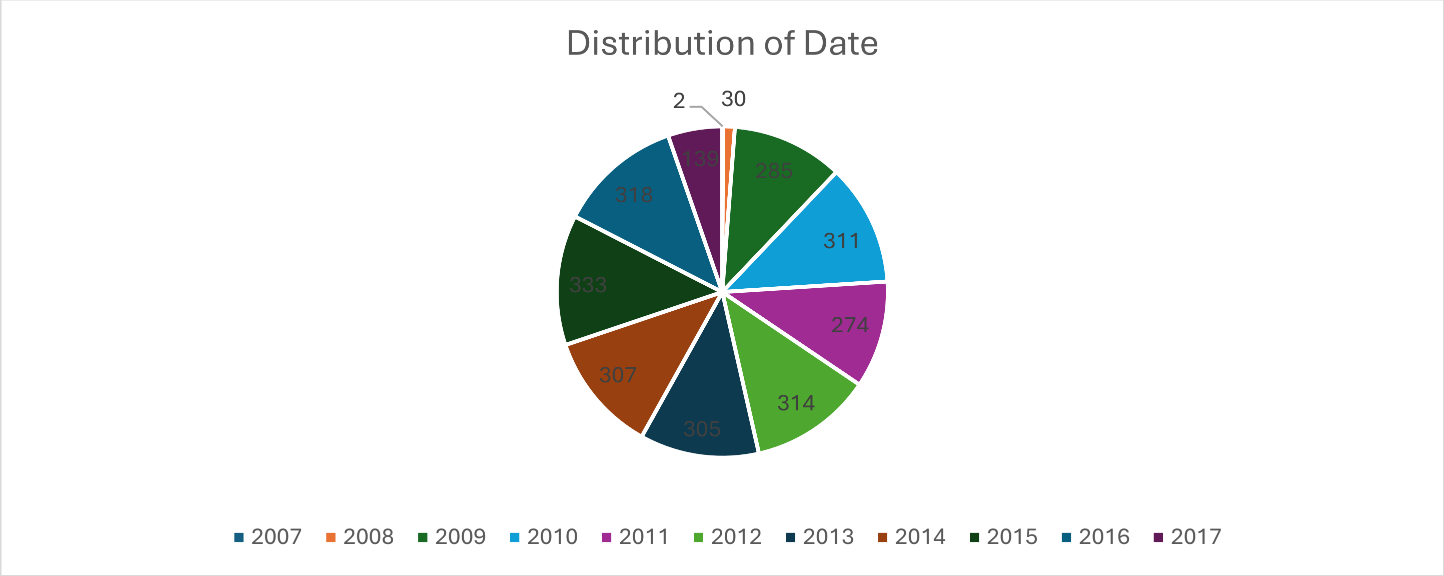


Figure 2 – Pie Chart of Date

Interpretation:

The distribution of the "Date" data spans across the years 2007 to 2017, showing a clear pattern in the frequency of entries per year. The year 2015 has the highest frequency with 333 entries, followed by 2016 with 318 entries. The year 2007 has the fewest entries, with only 2, reflecting a shorter observation period within that year. Overall, the data shows consistent frequency across most years, particularly from 2009 to 2016, with each year having over 250 entries. The sharp drop in frequency for the year 2017, at 139 entries, indicates either incomplete data collection or that the data set was concluded before the end of the year. The pie chart visually emphasizes the even distribution across years, with 2015 and 2016 occupying the largest segments, and 2007 being almost negligible.

Location

Frequency & Distribution Data Visualisation:

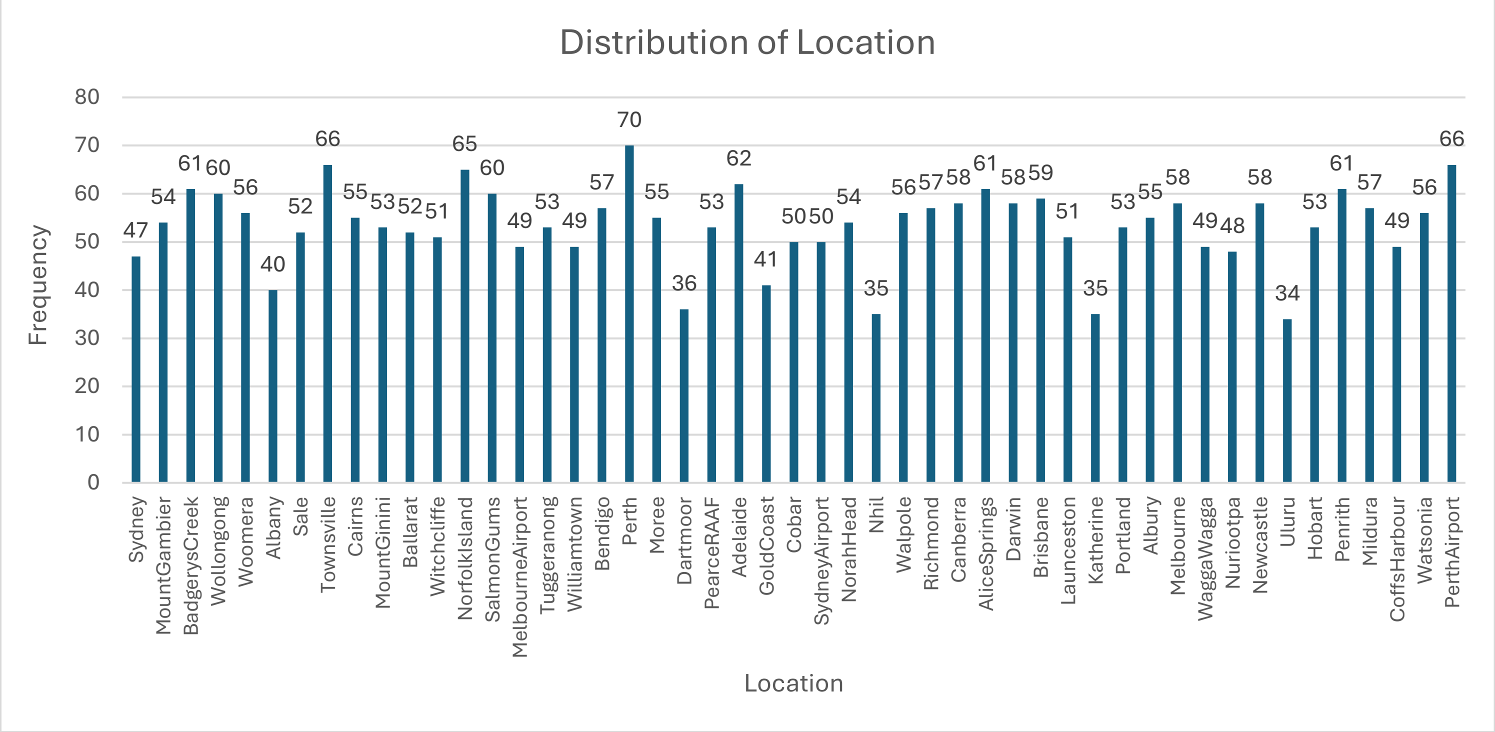


Figure 3 – Histogram of Location

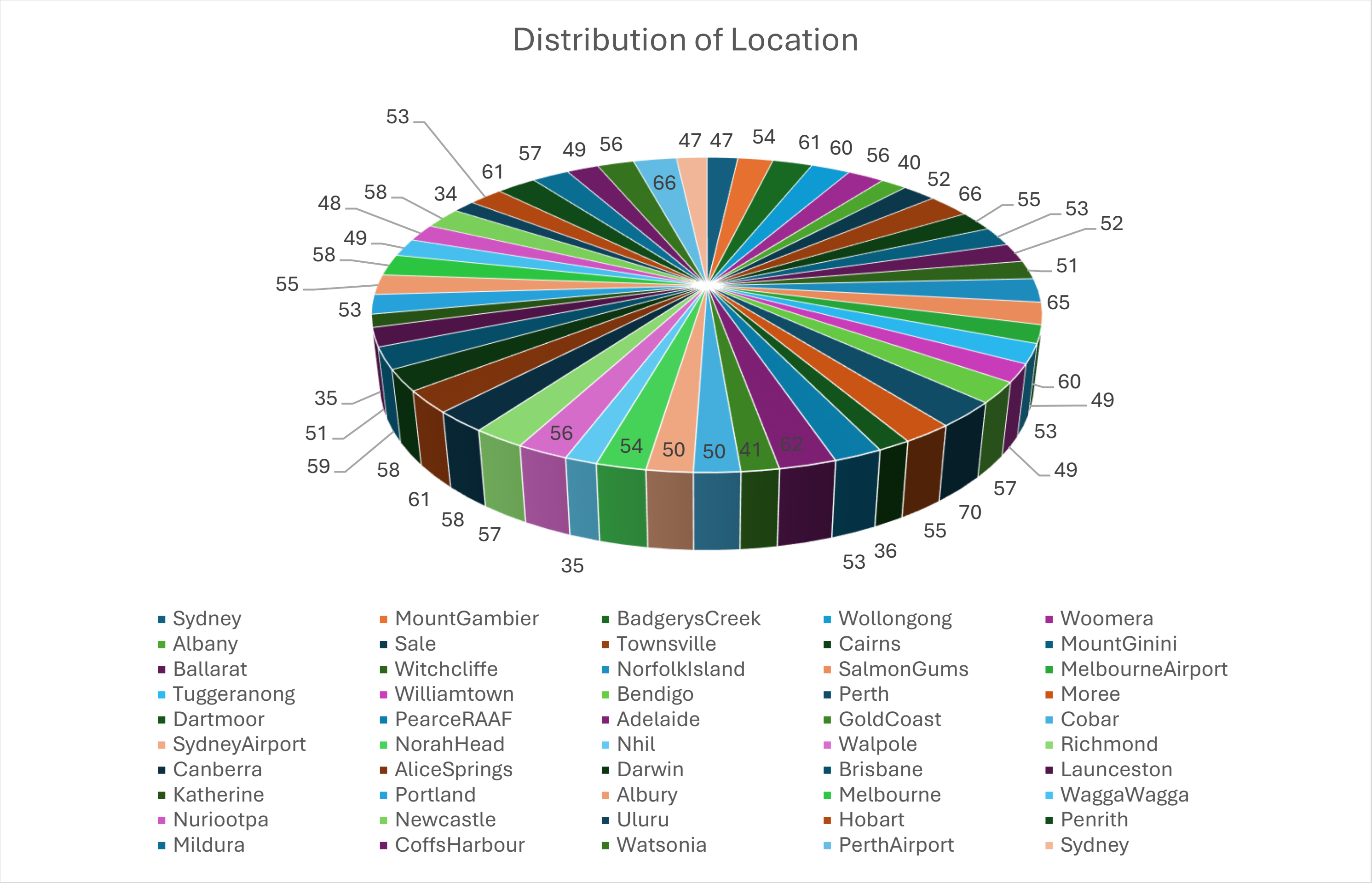


Figure 4 – 3D Pie Chart of Location

Interpretation:

The Location data is evenly distributed across several locations, with a few notable peaks. Williamtown (70 occurrences) stands out as the location with the highest frequency, followed by MountGambier (66 occurrences) and PerthAirport (66 occurrences). Locations like NorfolkIsland (35 occurrences) and Albany (40 occurrences) have the lowest frequencies. The histogram and pie chart show a relatively even spread of observations across different locations, with no extreme outliers. This suggests that data collection was widespread and fairly consistent across various locations, with only minor variations in the number of records for each site.

MinTemp

Frequency & Distribution Data Visualisations:

Statistics	Value
Min Range	-6.8
Max Range	29.4
Mean	12.169
Median	N/A
Standard Deviation	6.421
Variance	41.226
Skewness	0.013
Kurtosis	-0.476
Missing Values	28

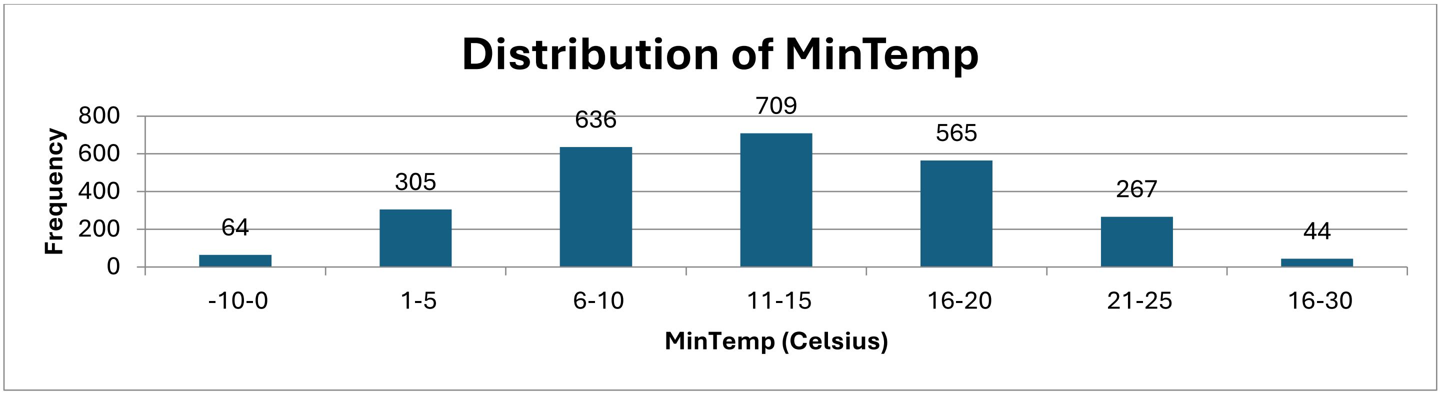


Figure 5 – Histogram of MinTemp

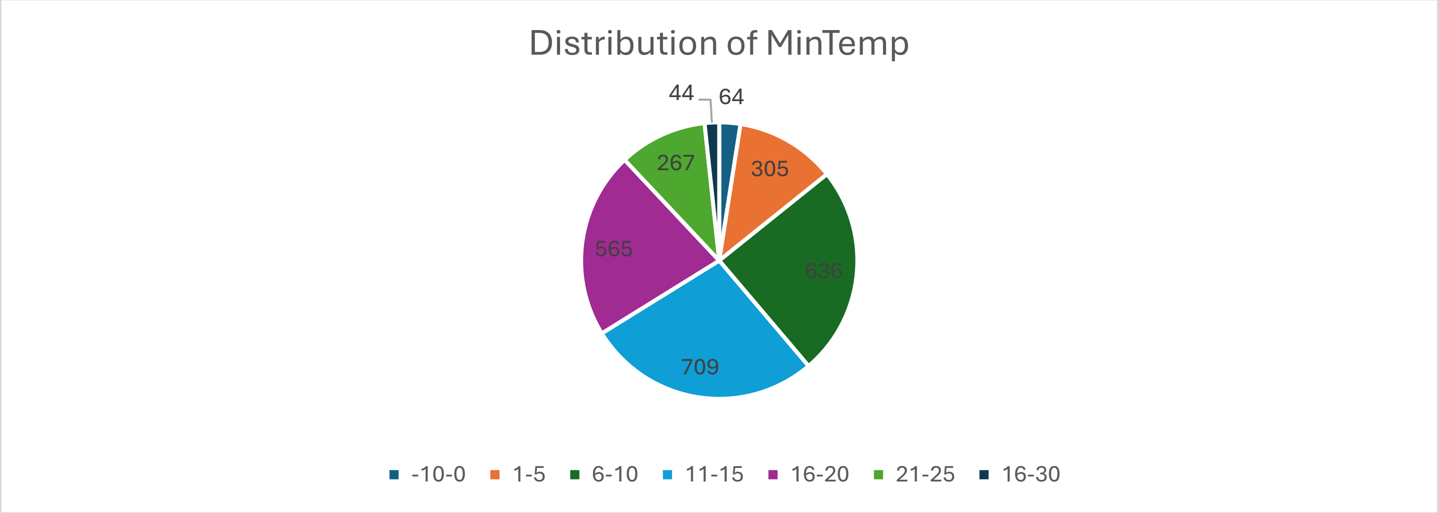


Figure 6 – Pie Chart of MinTemp

Interpretation:

The MinTemp data shows that the majority of minimum temperatures fall between 6-15°C, as indicated by the frequency graph where the 11-15°C range has the highest occurrence, followed by 6-10°C and 16-20°C. Extreme low temperatures below 0°C and high temperatures above 25°C are rare, shown by the small bars in the histogram. The pie chart visually reinforces this, with large segments for the 6-15°C ranges and much smaller segments for the extremes. This suggests that the dataset mainly represents a moderate climate, with a symmetric distribution of temperatures and only minor fluctuations from the mean. The moderate variation shown by the standard deviation confirms that temperatures generally stay within a predictable range.

MaxTemp

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	-1.8
Max Range	46.6
Mean	23.314
Median	N/A
Standard Deviation	7.135
Variance	50.904
Skewness	0.239
Kurtosis	-0.187
Missing Values	25

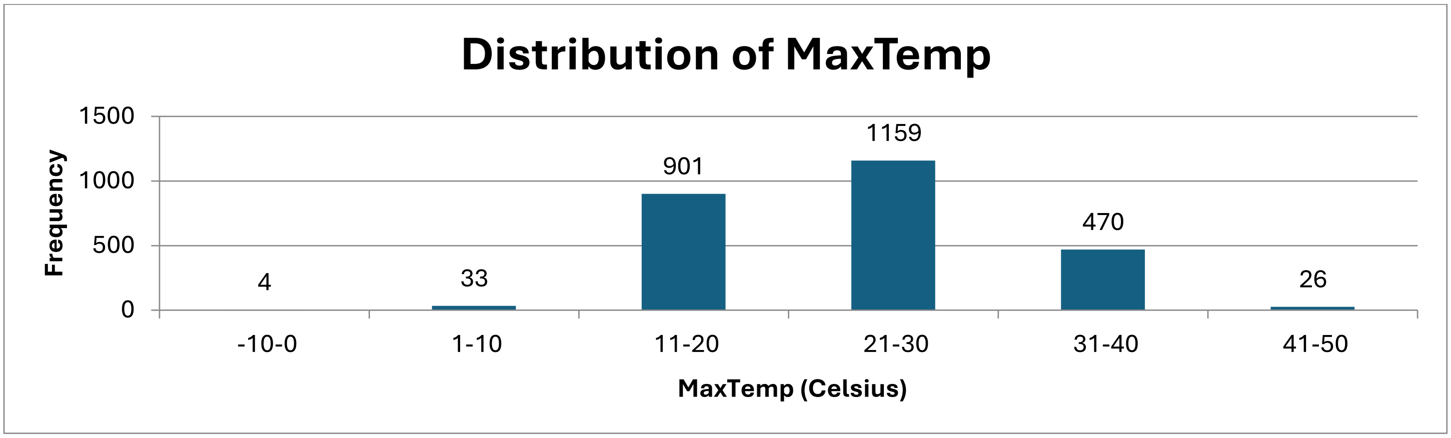


Figure 7 – Histogram of MaxTemp

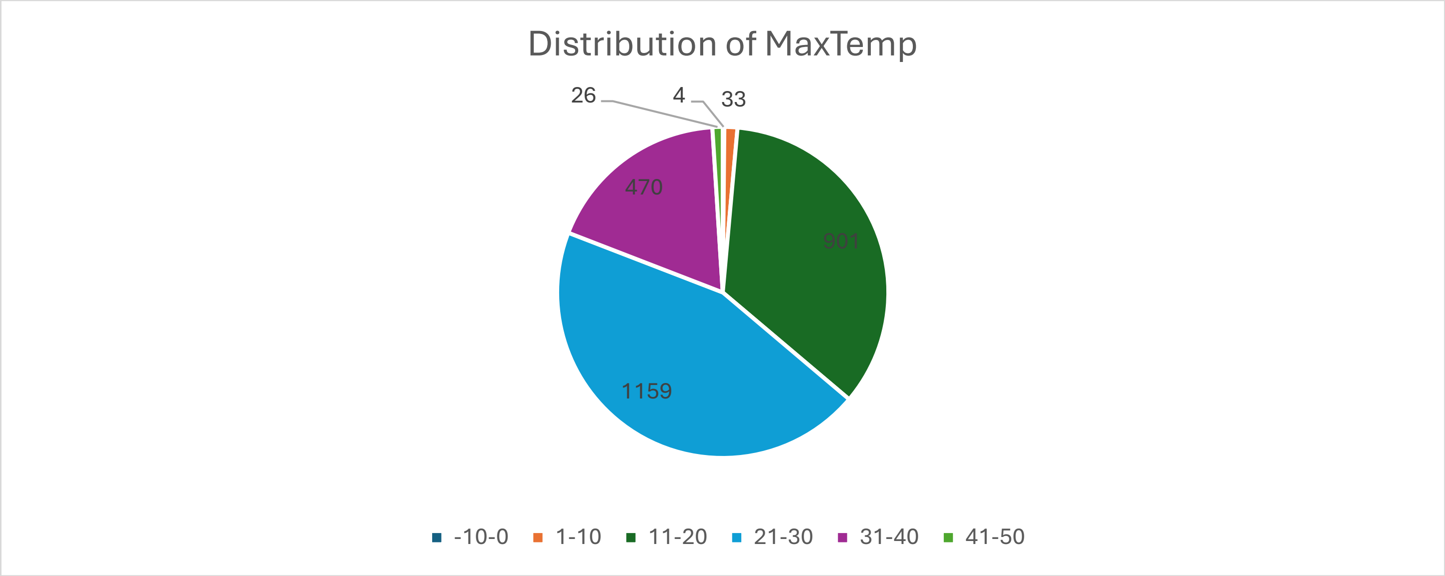


Figure 8 – Pie Chart of MaxTemp

Interpretation:

The MaxTemp data reveals that most maximum temperatures fall between 21-30°C, as shown in both the histogram and the pie chart, where this range dominates with the highest frequency of over 1000 occurrences. The next most common range is 11-20°C, followed by 31-40°C, indicating that temperatures typically remain warm, with fewer extremely high or low temperatures. The histogram also shows very few occurrences of extreme temperatures below 0°C or above 40°C, reinforcing that the dataset represents a relatively warm climate. The pie chart further emphasises this, with large segments for the 21-30°C and 11-20°C ranges, while other temperature ranges make up smaller portions. The data's slight right skewness (0.239) indicates a few occurrences of higher temperatures, but overall, temperatures tend to be moderate.

Rainfall

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	0
Max Range	115.2
Mean	2.207
Median	N/A

Standard Deviation	7.199
Variance	51.82
Skewness	6.59
Kurtosis	61.731
Missing Values	65

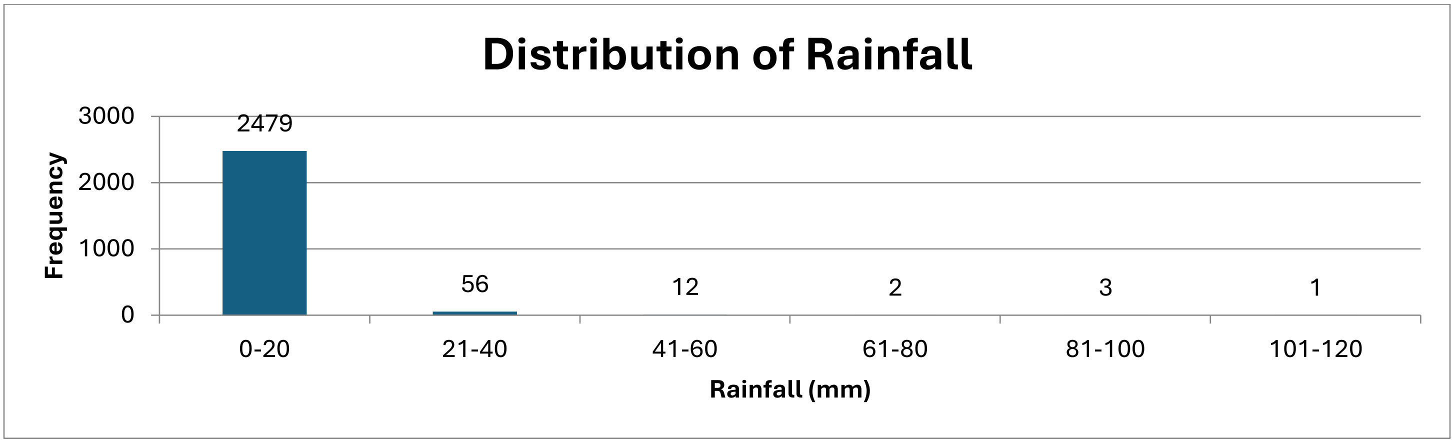


Figure 9 – Histogram of Rainfall

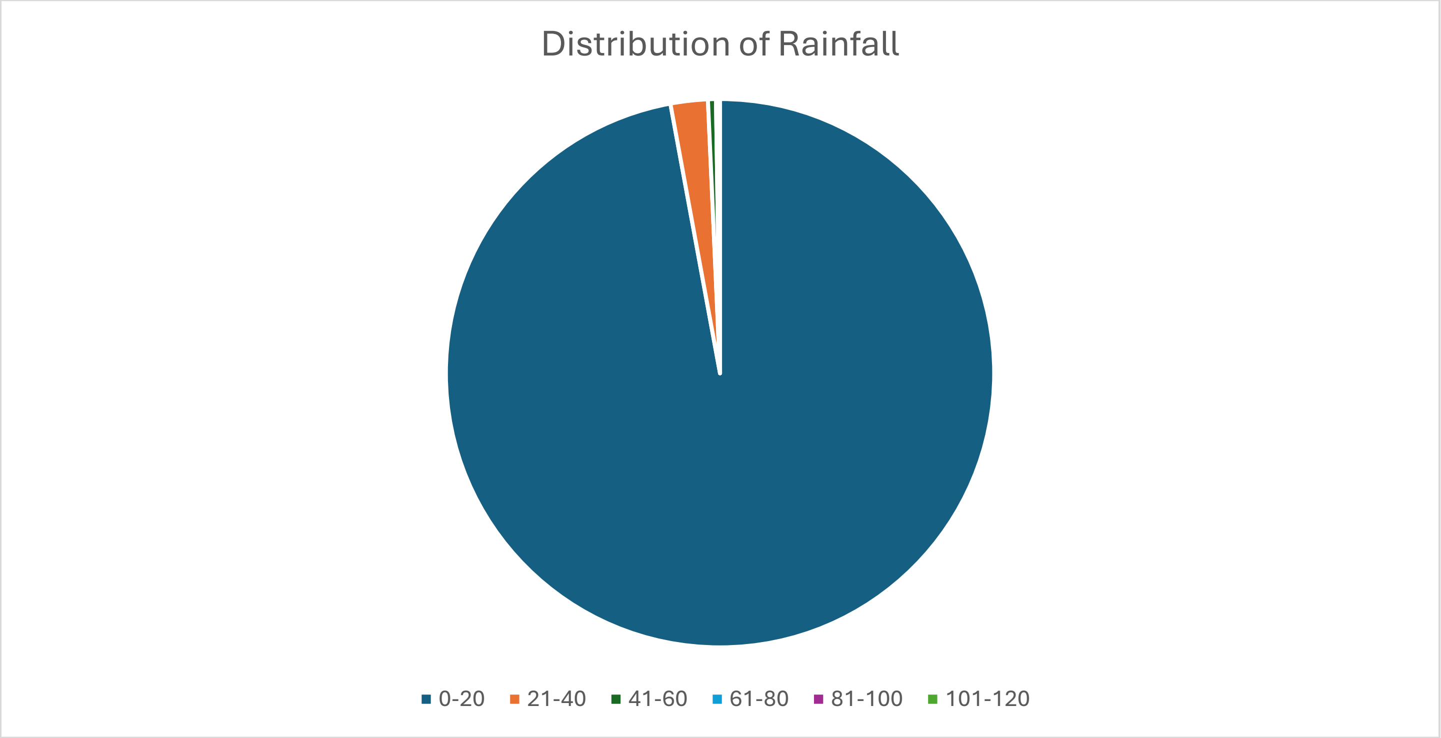


Figure 10 – Pie Chart of Rainfall

Interpretation:

The Rainfall data indicates that the vast majority of rainfall events fall within the 0-20 mm range, as seen in both the histogram and pie chart, with over 2500 occurrences. This suggests that light rainfall is the norm in the dataset. The next range, 21-40 mm, has a significantly lower frequency, and rainfall greater than 40 mm is exceedingly rare. The high skewness (6.59) and kurtosis (61.731) values reflect the strong right skew of the data, with most rainfall events clustered at the lower end and very few extreme values. The histogram visually reinforces this, with nearly all data points concentrated in the 0-20 mm range. The pie chart shows this even more clearly, with the 0-20 mm category dominating. The overall analysis points to a climate with predominantly light rain events and few occurrences of heavy rainfall.

Evaporation

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	0
Max Range	58.5
Mean	5.57
Median	N/A
Standard Deviation	4.377
Variance	19.161
Skewness	3.327
Kurtosis	24.469
Missing Values	1117

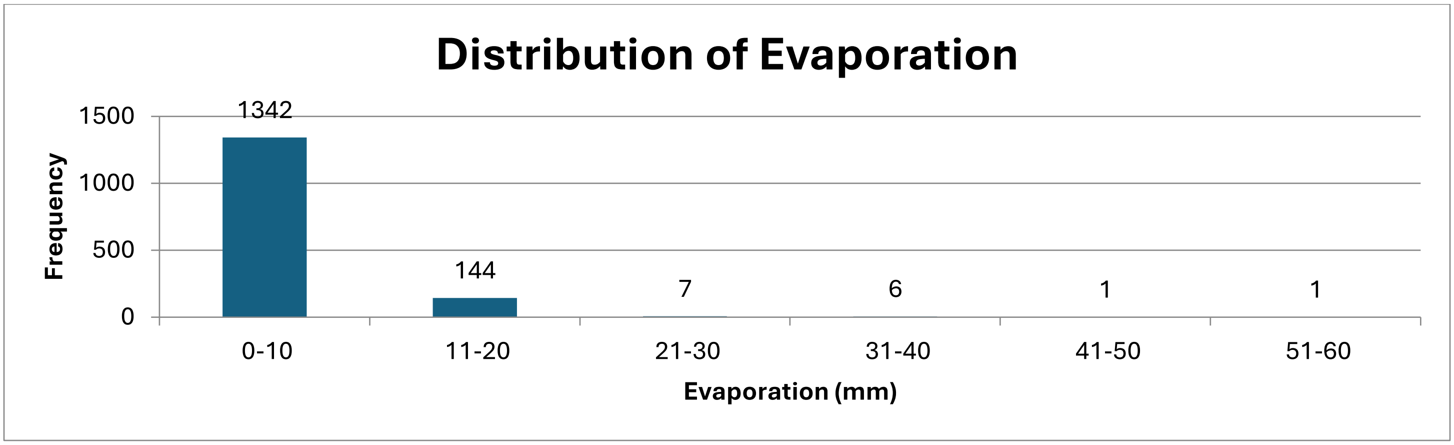


Figure 10 – Histogram of Evaporation

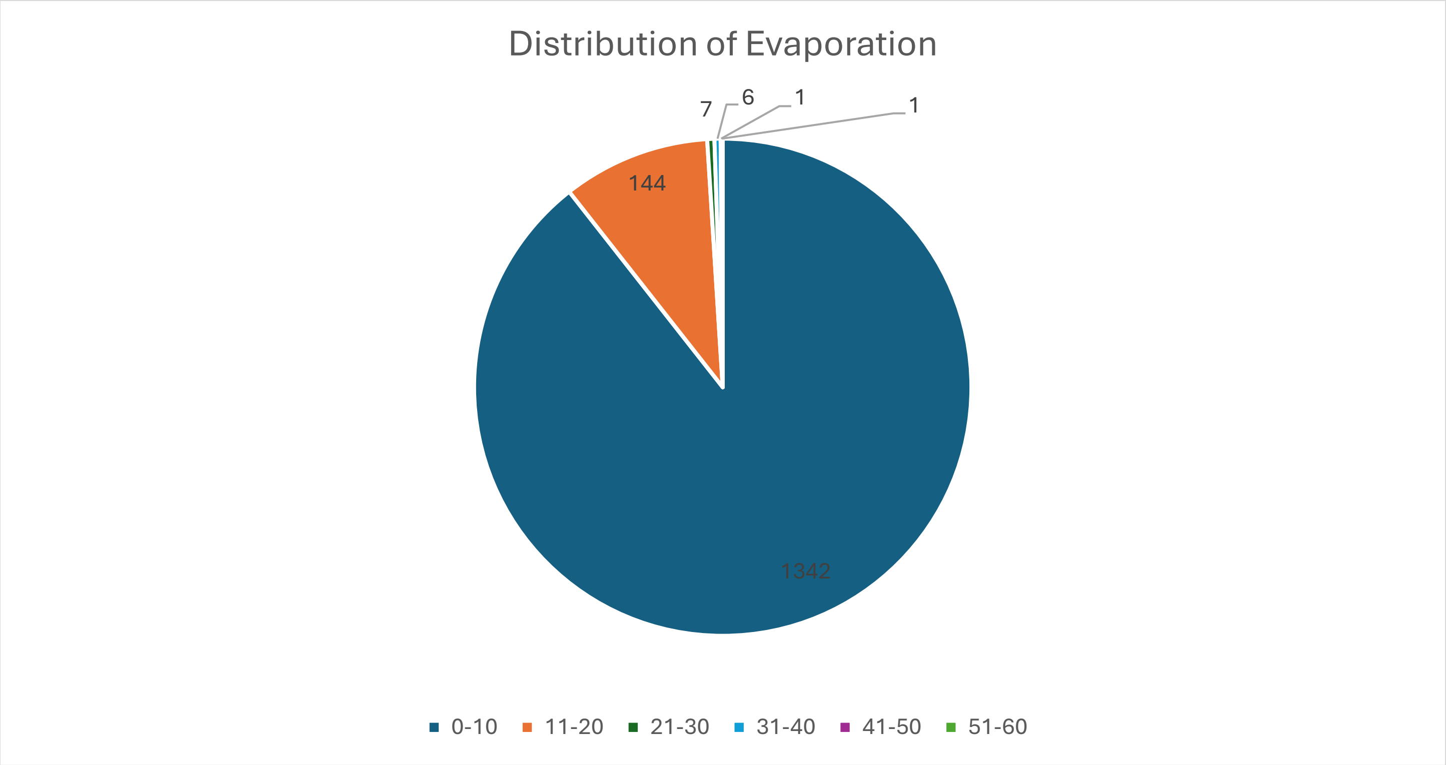


Figure 11 – Pie Chart of Evaporation

Interpretation:

The Evaporation data reveals that most evaporation levels fall within the 0-10 mm range, as shown in both the histogram and pie chart, with over 1500 occurrences. This indicates that low levels of evaporation are typical in the dataset. The next category, 11-20 mm, is significantly less frequent, and evaporation above 20 mm is exceedingly rare. The high skewness (3.327) and kurtosis (24.469) values indicate a strong right skew in the data, with the vast majority of the evaporation values clustered at the lower end and very few extremely high values. The histogram highlights this, with nearly all data points falling into the 0-10 mm category. The pie chart similarly emphasises the dominance of the 0-10 mm range. Overall, the dataset suggests that the climate experiences predominantly low evaporation rates, with only a few instances of higher evaporation levels.

Sunshine

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	0
Max Range	13.8
Mean	7.557
Median	N/A
Standard Deviation	3.824
Variance	14.624
Skewness	-0.51
Kurtosis	-0.862
Missing Values	1252

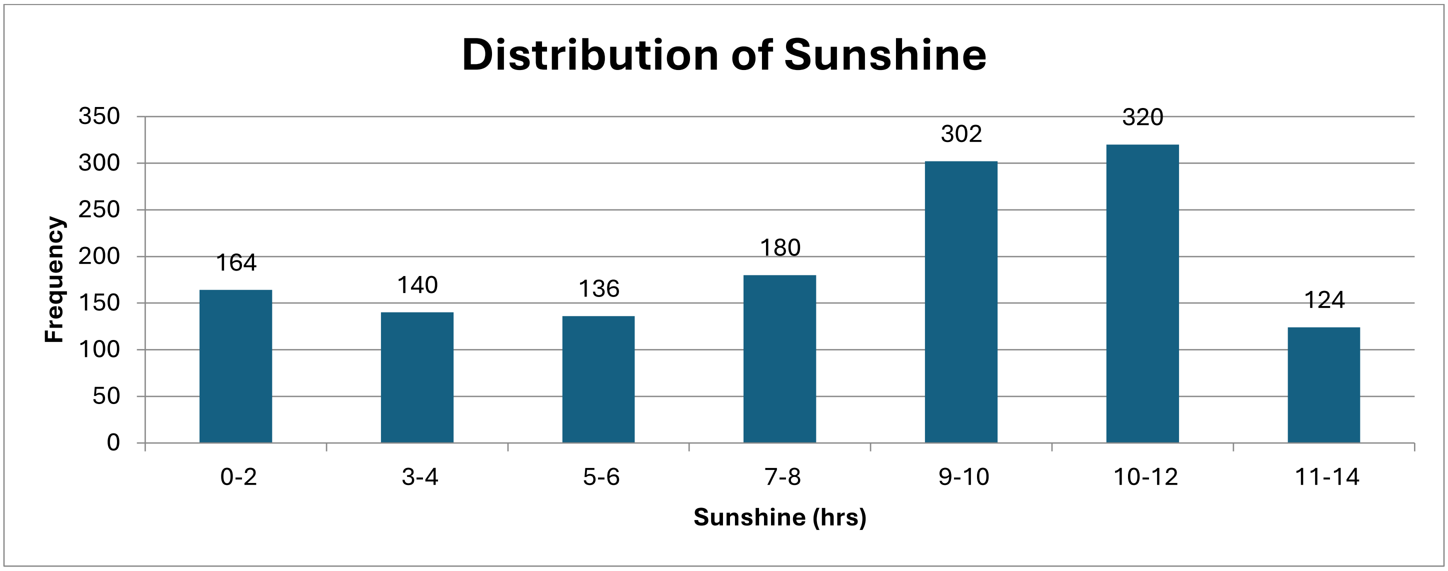


Figure 12 – Histogram of Sunshine

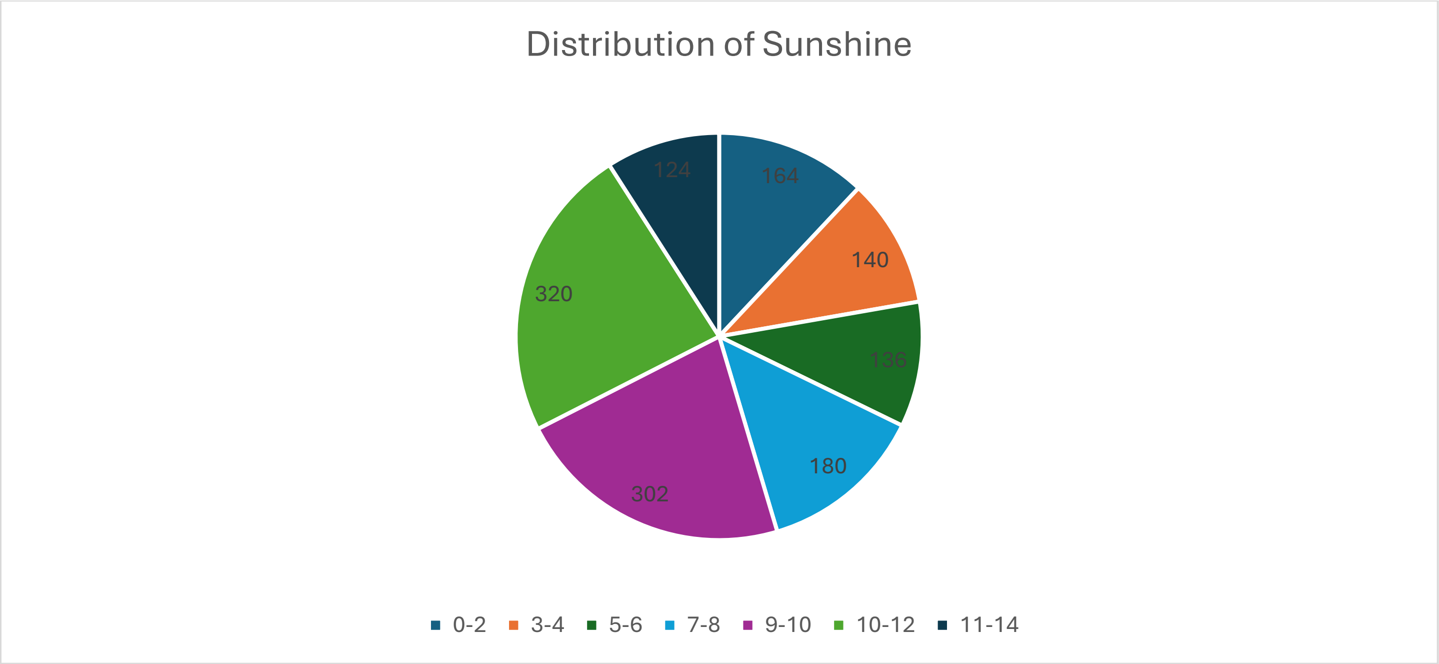


Figure 13 – Pie Chart of Sunshine

Interpretation:

The Sunshine data indicates that the most frequent sunshine durations fall between 9-12 hours, as shown in both the histogram and pie chart, with over 300 occurrences in the 11-12 hour range and a similar number in the 9-10 hour range. Sunshine durations of 0-2 hours and 13-14 hours are less common, but still present. The data's slight left skew (skewness: -0.51) suggests more occurrences of longer sunshine hours, with relatively few days of very short or extremely long sunshine durations. The pie chart further visualises this, showing larger segments for the 9-12 hour ranges, while the smaller segments represent fewer days with minimal or excessive sunshine. Overall, the data reflects a climate with a relatively balanced distribution of sunshine hours, though days with 9-12 hours of sunshine are the most common.

WindGusDir

Frequency & Distribution Data Visualisation:

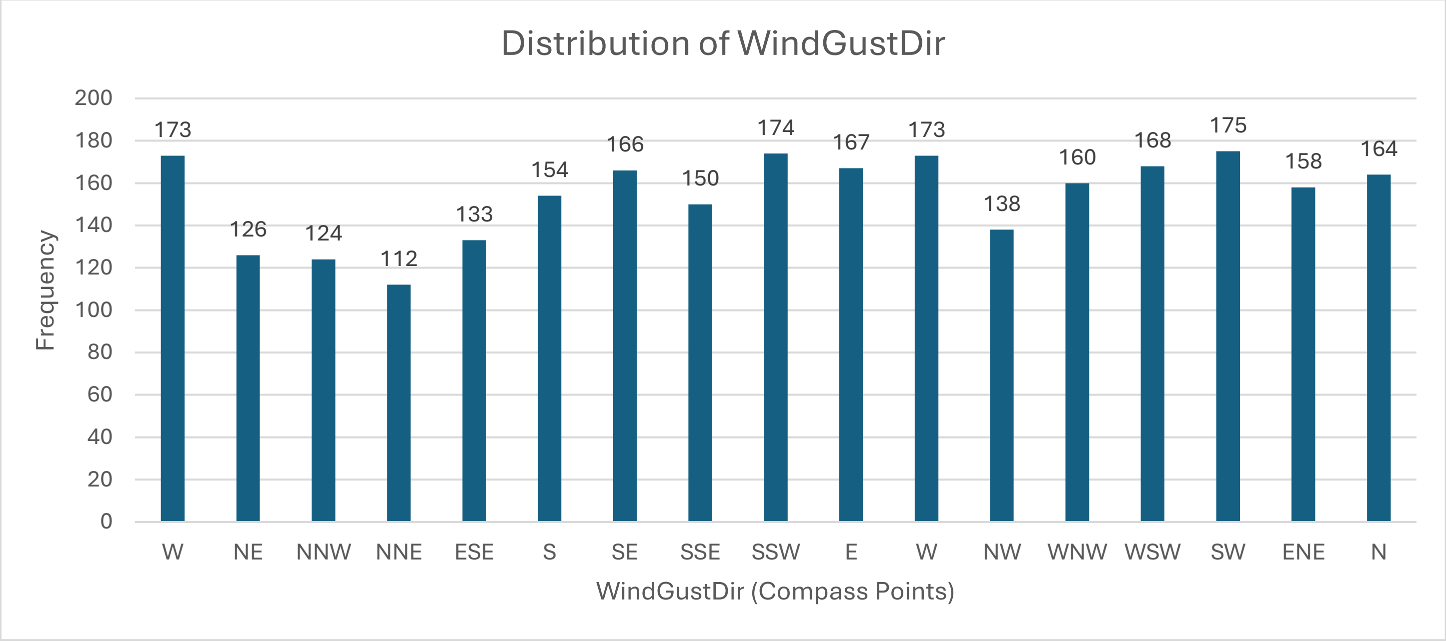


Figure 14 – Histogram of WindGustDir

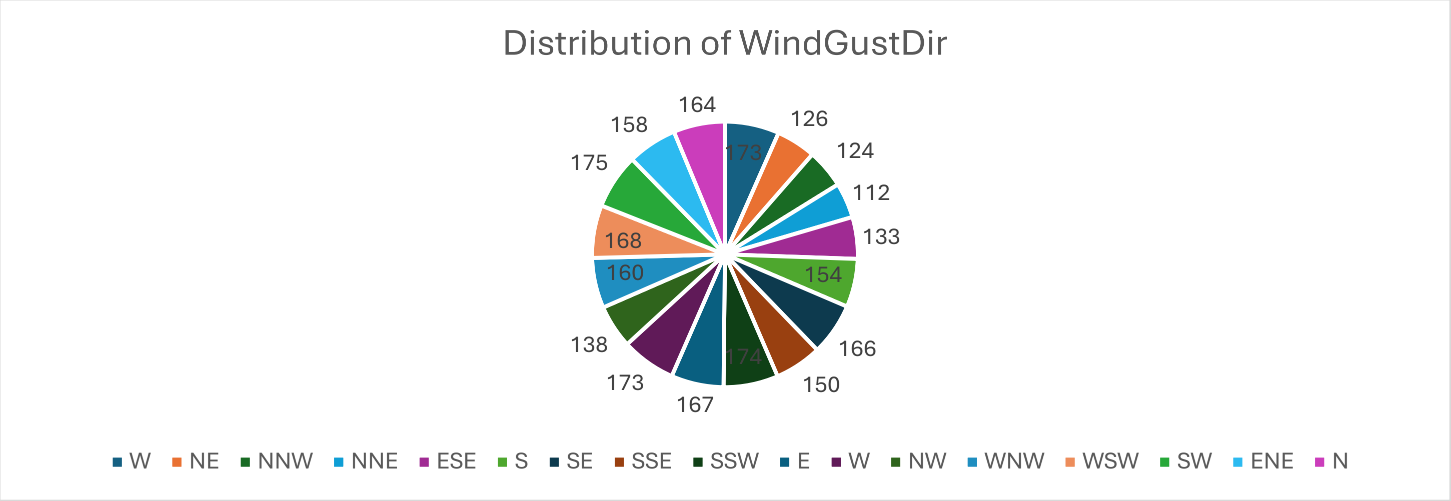


Figure 15 – Pie Chart of WindGustDir

Interpretation:

The WindGustDir data shows that wind gusts are fairly evenly distributed across all compass points, with the most frequent directions being SSW (174 occurrences), W (173 occurrences), and WSW (175 occurrences). The least frequent directions are NNE (112 occurrences) and NNW (124 occurrences). The histogram and pie chart illustrate this near-uniform distribution of wind gust directions, with no single direction overwhelmingly dominant. This suggests that wind gusts come from a wide range of directions with only minor variations in frequency, reflecting a relatively balanced wind pattern.

WindGusSpeed

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	9
Max Range	117
Mean	39.569
Median	N/A
Standard Deviation	13.657
Variance	186.516
Skewness	0.912
Kurtosis	-1.46
Missing Values	175

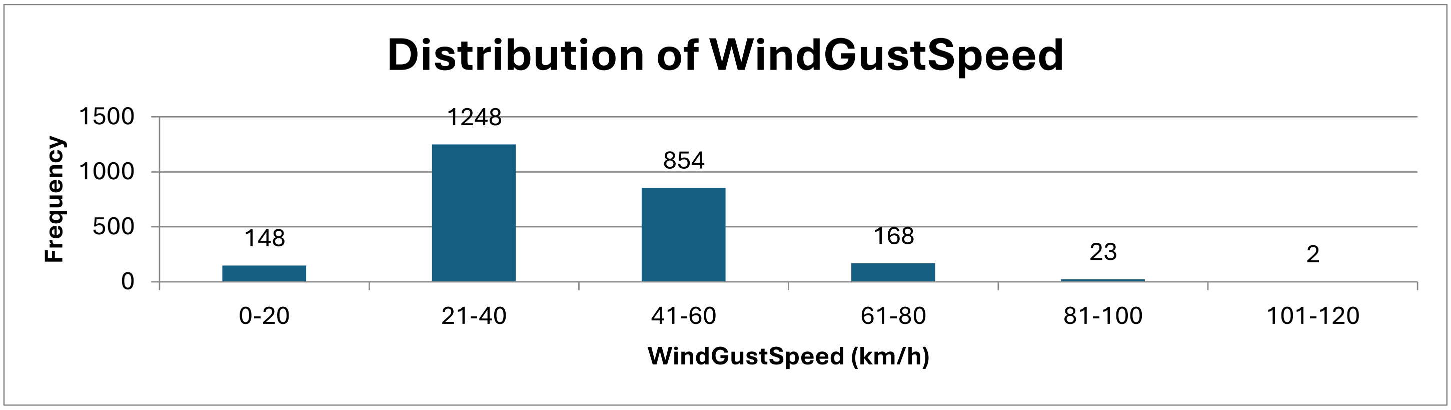


Figure 16 – Histogram of WindGustSpeed

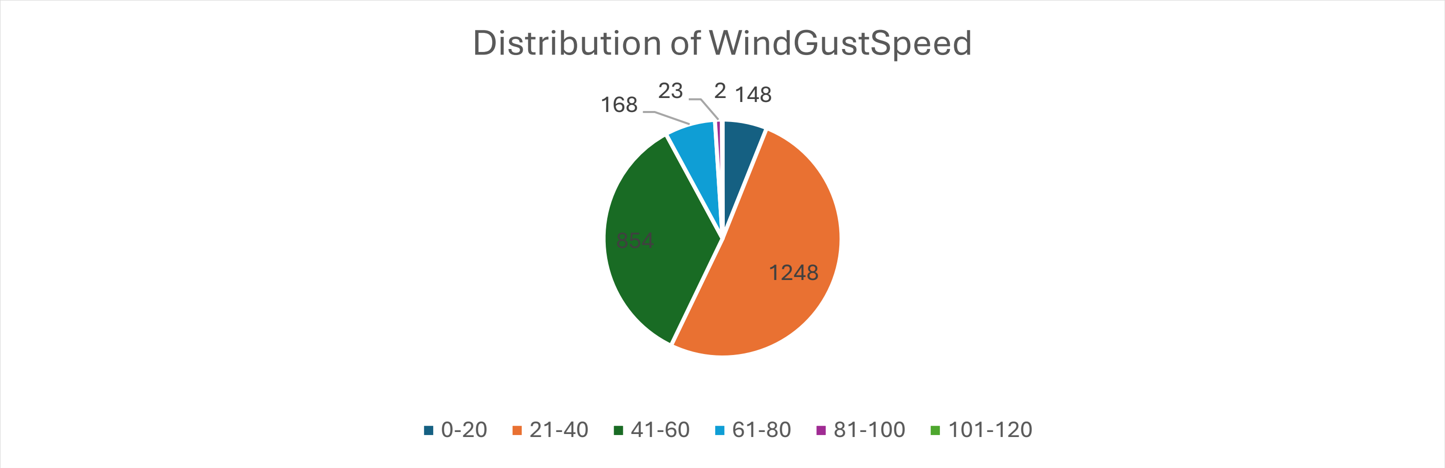


Figure 17 – Pie Chart of WindGustSpeed

Interpretation:

The WindGustSpeed data indicates that the majority of wind gusts fall between 21-60 km/h, as shown in both the histogram and pie chart, with the 21-40 km/h range being the most frequent, having over 1500 occurrences. The 41-60 km/h range follows closely. Wind gusts above 80 km/h are rare, as seen by the small bars in the histogram. The data is moderately right-skewed (skewness: 0.912), indicating a higher frequency of lower wind speeds, with a long tail towards the higher end. The pie chart reinforces this, with large segments for the 21-40 km/h and 41-60 km/h ranges, while higher wind gusts contribute minimally. Overall, the dataset reflects a climate with typical wind gusts in the moderate range, and few occurrences of extreme wind speeds.

9am Temp

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	-4.2
Max Range	36.4
Mean	16.971
Median	N/A
Standard Deviation	6.53
Variance	42.64
Skewness	0.077
Kurtosis	-0.289
Missing Values	30

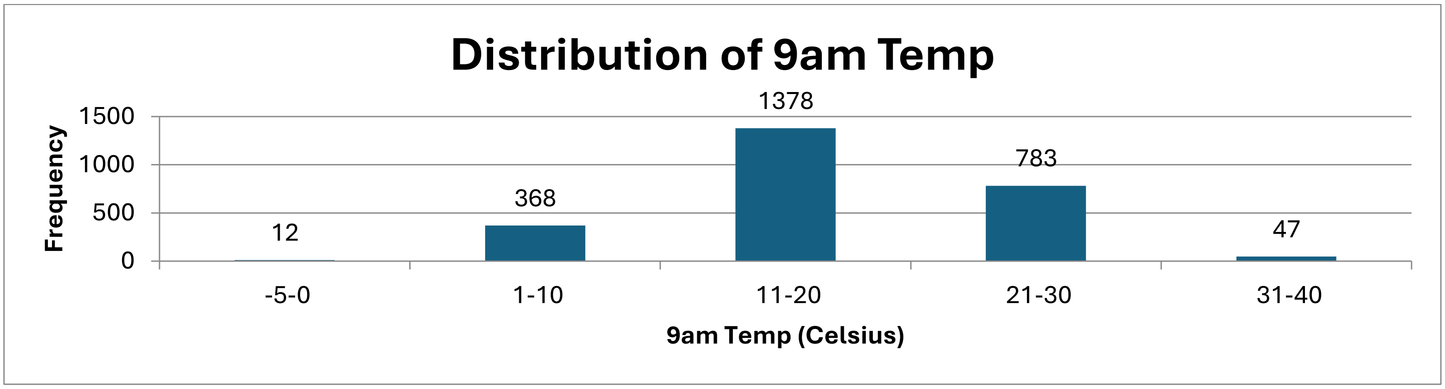


Figure 18 – Histogram of 9am Temp

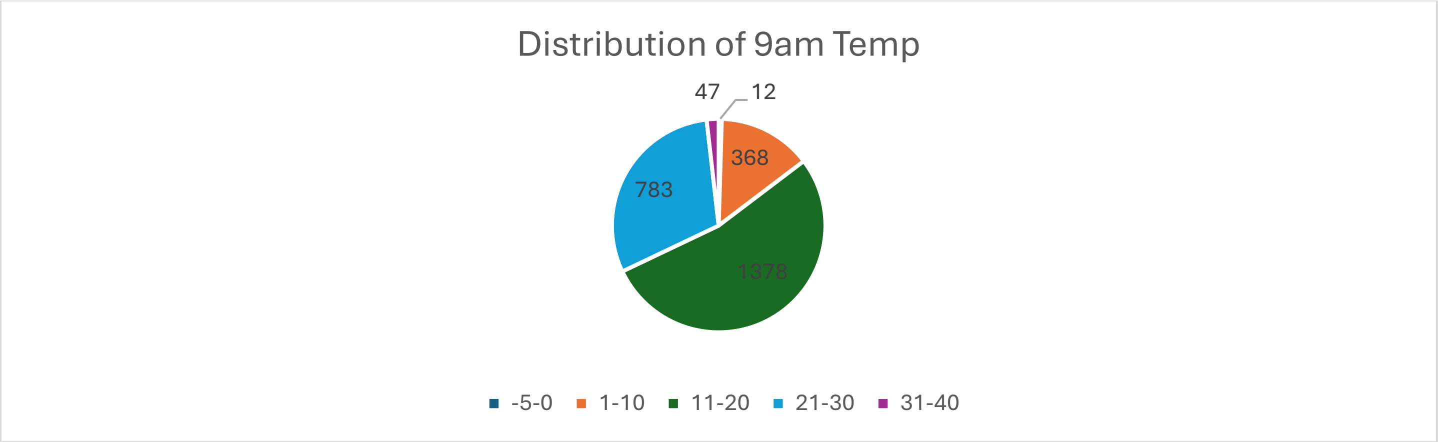


Figure 19 – Pie Chart of 9am Temp

Interpretation:

The 9am Temp data reveals that most temperatures at 9am fall between 11-20°C, as indicated by both the histogram and pie chart, with 1378 occurrences in this range. The next most frequent range is 21-30°C, followed by 1-10°C, with very few occurrences below 0°C or above 30°C. The data shows a nearly symmetric distribution, as suggested by the skewness (0.077), with a slight leftward tilt. The pie chart clearly illustrates the dominance of moderate temperatures, particularly between 11-20°C, while extreme temperatures are rare. Overall, the data suggests that mornings are typically mild, with the majority of temperatures falling between 11-30°C, and extreme low or high temperatures being uncommon.

9am Humidity

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	4
Max Range	100
Mean	69.092
Median	N/A
Standard Deviation	18.977
Variance	360.119
Skewness	-0.463
Kurtosis	0.027
Missing Values	46

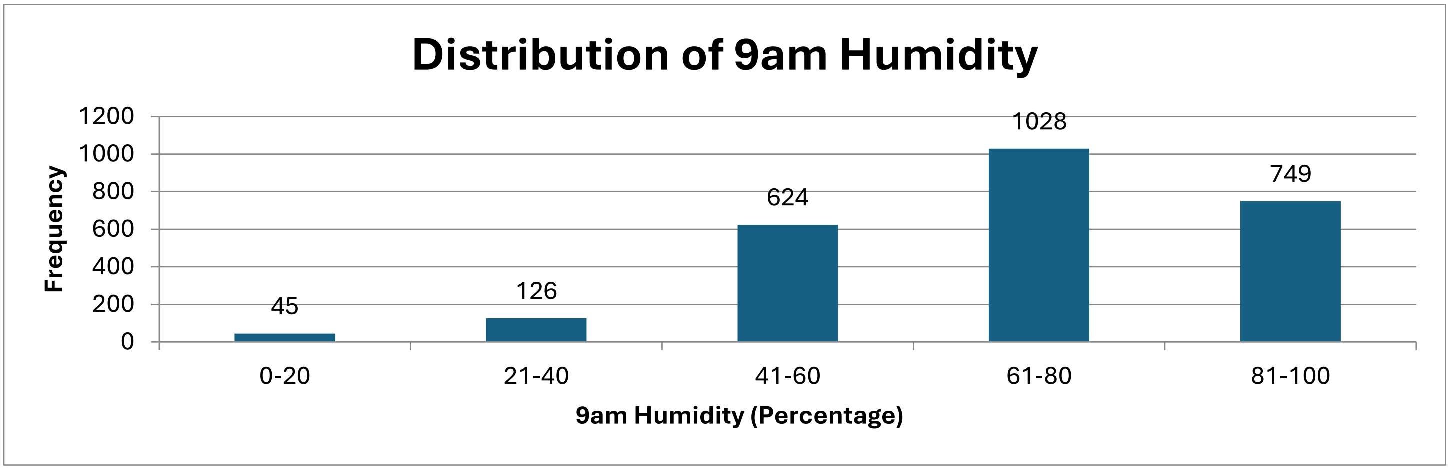


Figure 20 – Histogram of 9am Humidity

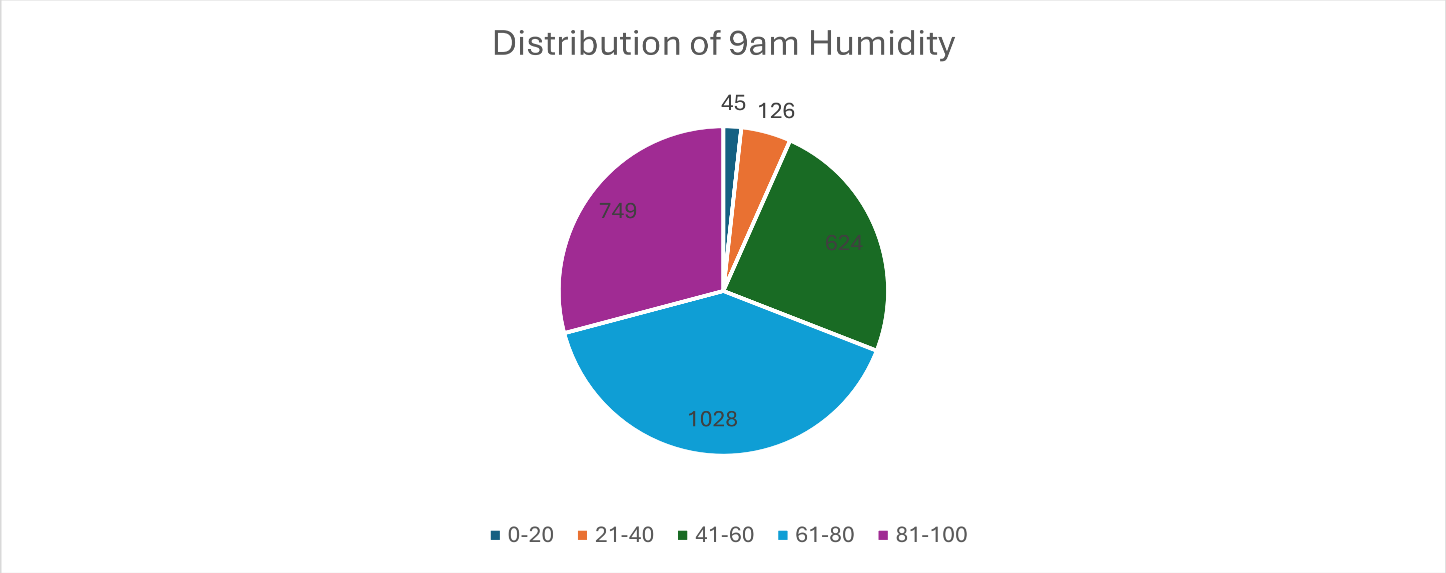


Figure 21 – Pie Chart of 9am Humidity

Interpretation:

The 9am Humidity data shows that most humidity levels fall between 61-100%, as reflected in both the histogram and pie chart, with 1028 occurrences in the 61-80% range and 749 occurrences in the 81-100% range. Humidity levels below 40% are rare, and the dataset indicates that mornings are typically humid, with very few instances of dry air. The data is slightly left-skewed (skewness: -0.463), indicating a higher occurrence of higher humidity levels, and the pie chart emphasises this distribution, with larger segments for the 61-100% ranges. Overall, the dataset reflects a climate where mornings are generally quite humid, with lower humidity levels being less common.

9am Cloud

Frequency & Distribution Data Visualisation:

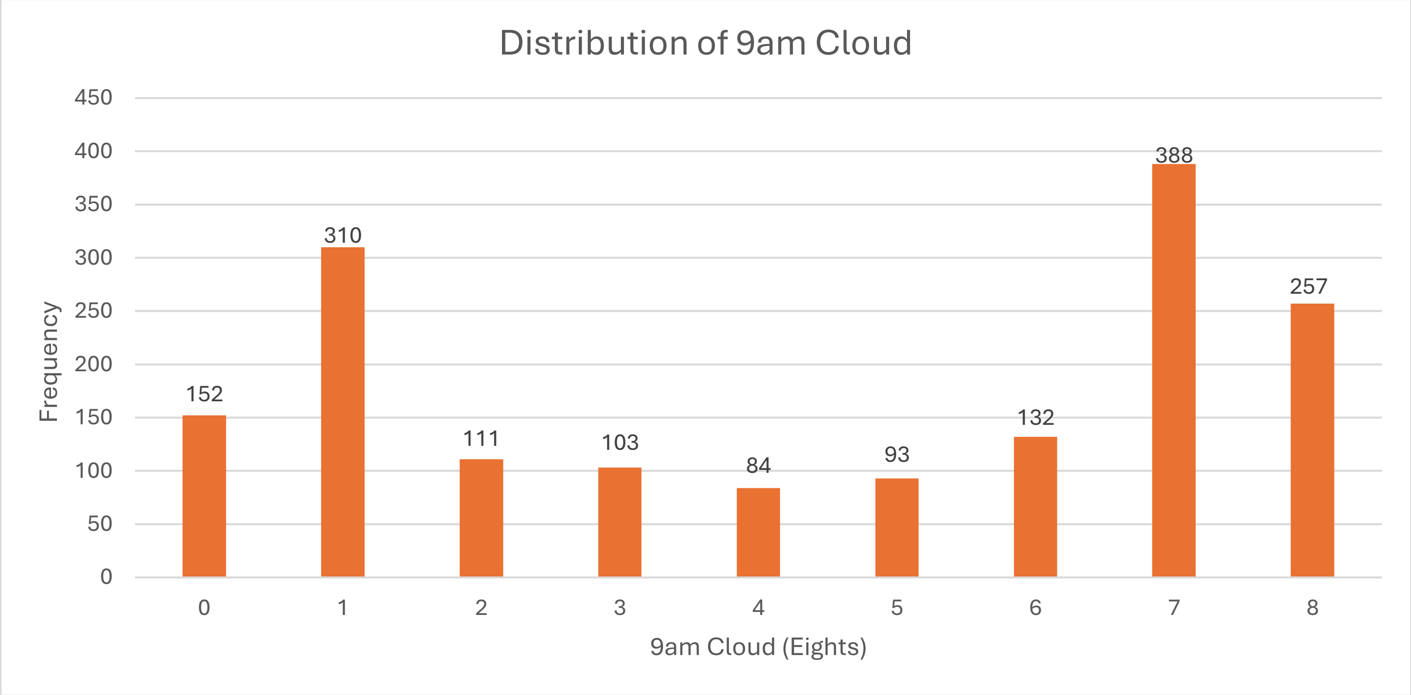


Figure 22 – Histogram of 9am Cloud

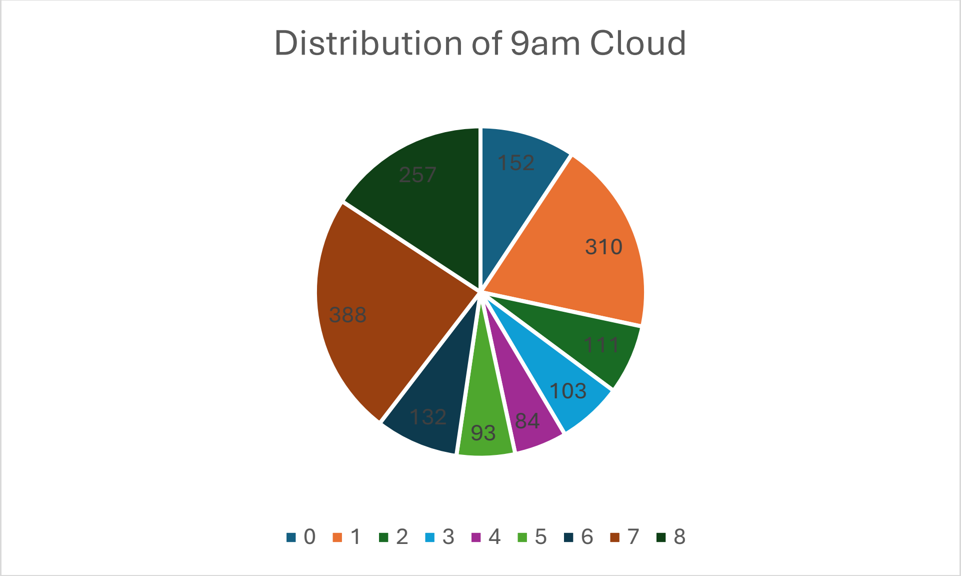


Figure 23 – Pie Chart of 9am Cloud

Interpretation:

The 9am Cloud data indicates that cloud cover is most frequently observed at 7 oktas (eighths), with 388 occurrences, followed by 1 okta with 310 occurrences, and 8 oktas with 257 occurrences. The distribution shows a bimodal pattern with peaks at both high and low cloud cover values. The histogram and pie chart both highlight these two prominent categories, suggesting that the sky is often either mostly clear or mostly covered at 9am. Moderate cloud cover values (between 2-5 oktas) are much less frequent, with fewer than 100 occurrences in each of these categories. Overall, the data reflects that mornings are typically characterised by either very little or substantial cloud cover.

9am WindDir

Frequency & Distribution Data Visualisation:

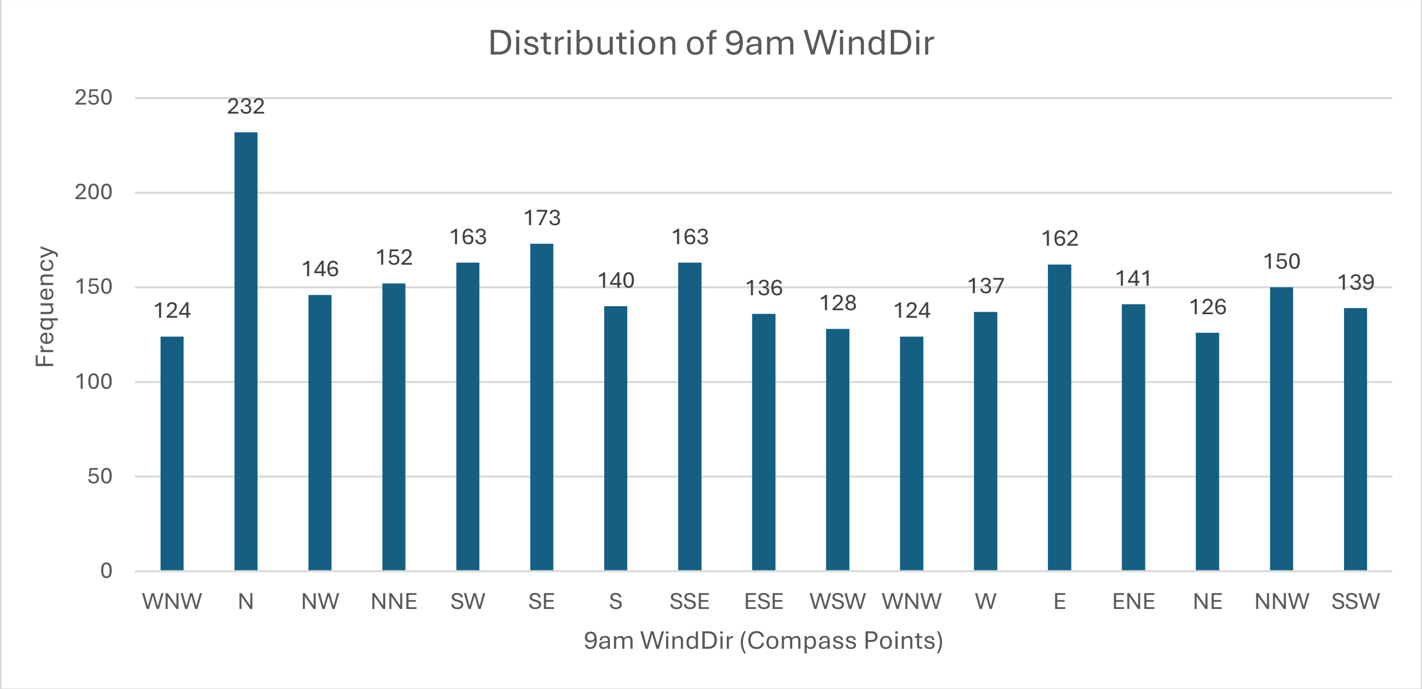


Figure 24 – Histogram of 9am WindDir

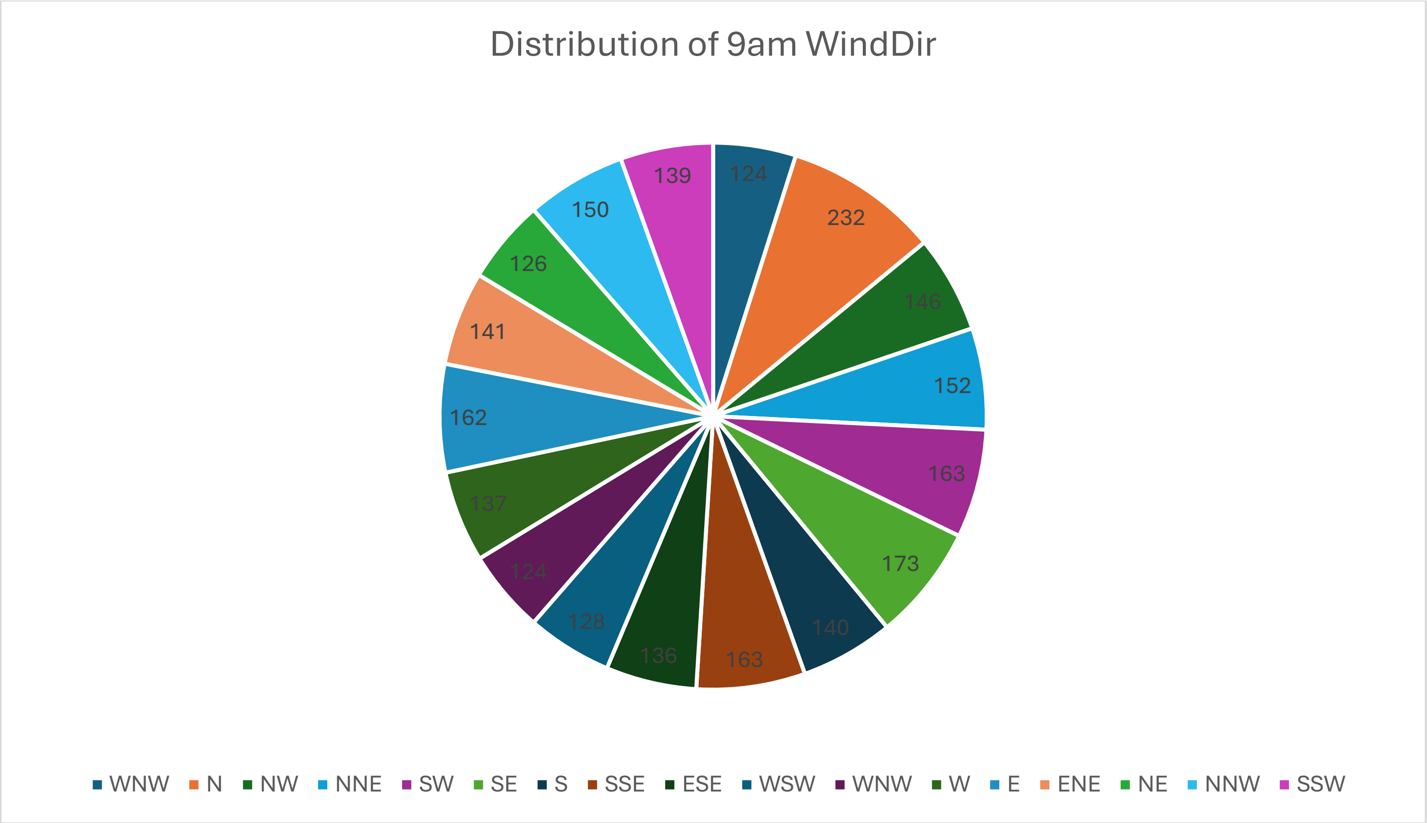


Figure 25 – Pie Chart of 9am WindDir

Interpretation:

The 9am WindDir data shows that the most common wind direction at 9am is from the North (232 occurrences), followed by Southwest (173 occurrences) and Southeast (163 occurrences). Wind directions from WNW (124 occurrences) and NNE (126 occurrences) are less frequent. The histogram and pie chart both illustrate the broad distribution of wind directions, with no extreme outliers, though the north direction stands out as the most frequent. This suggests that wind directions at 9am are relatively balanced, but with a notable tendency for winds to come from the north and southwest.

9am WindSpeed

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	0
Max Range	56
Mean	13.638
Median	N/A
Standard Deviation	8.877
Variance	78.8
Skewness	0.812
Kurtosis	1.074
Missing Values	39

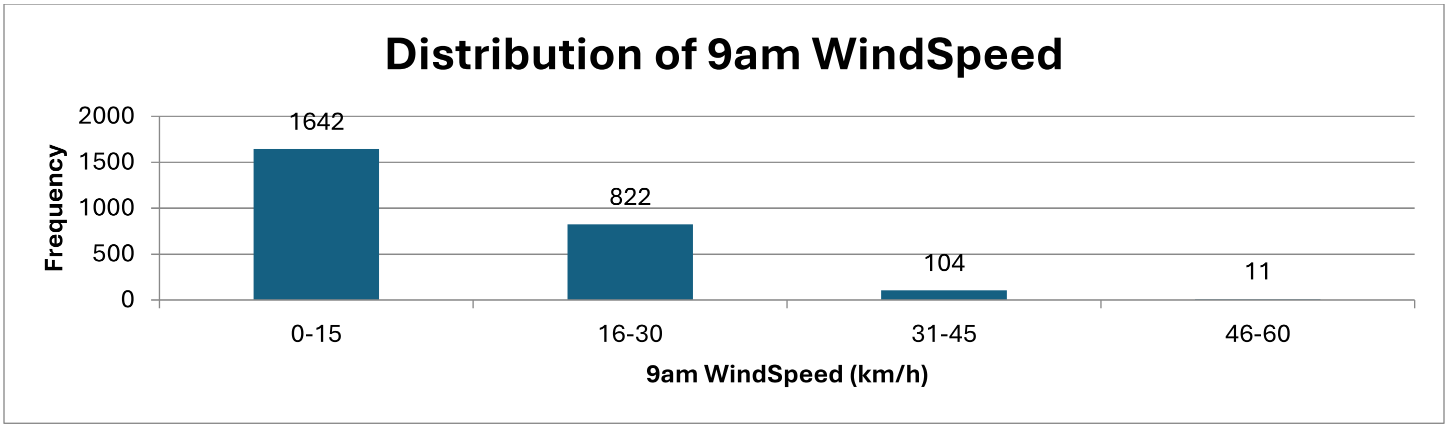


Figure 26 – Histogram of 9am WindSpeed

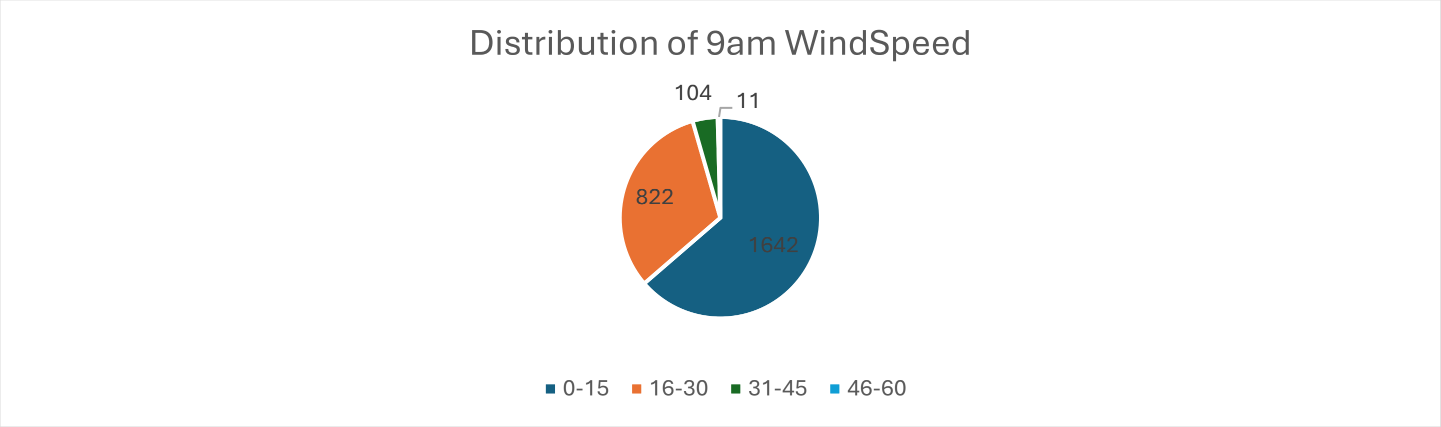


Figure 27 – Pie Chart of 9am WindSpeed

Interpretation:

The 9am WindSpeed data shows that most wind speeds fall between 0-15 km/h, with 1642 occurrences, followed by the 16-30 km/h range, which has 822 occurrences. Wind speeds above 30 km/h are rare, with only 104 occurrences between 31-45 km/h and just 11 occurrences for 46-60 km/h. The data has a slight right skew (skewness: 0.812), indicating that higher wind speeds are less frequent, and the pie chart emphasises the dominance of lower wind speeds, particularly in the 0-15 km/h range. Overall, the dataset suggests that calm to moderate winds is typical at 9am, with stronger winds being infrequent.

9am Pressure

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	989.8
Max Range	1040.5
Mean	1017.742
Median	N/A
Standard Deviation	6.963
Variance	48.487
Skewness	-0.009
Kurtosis	0.345
Missing Values	295

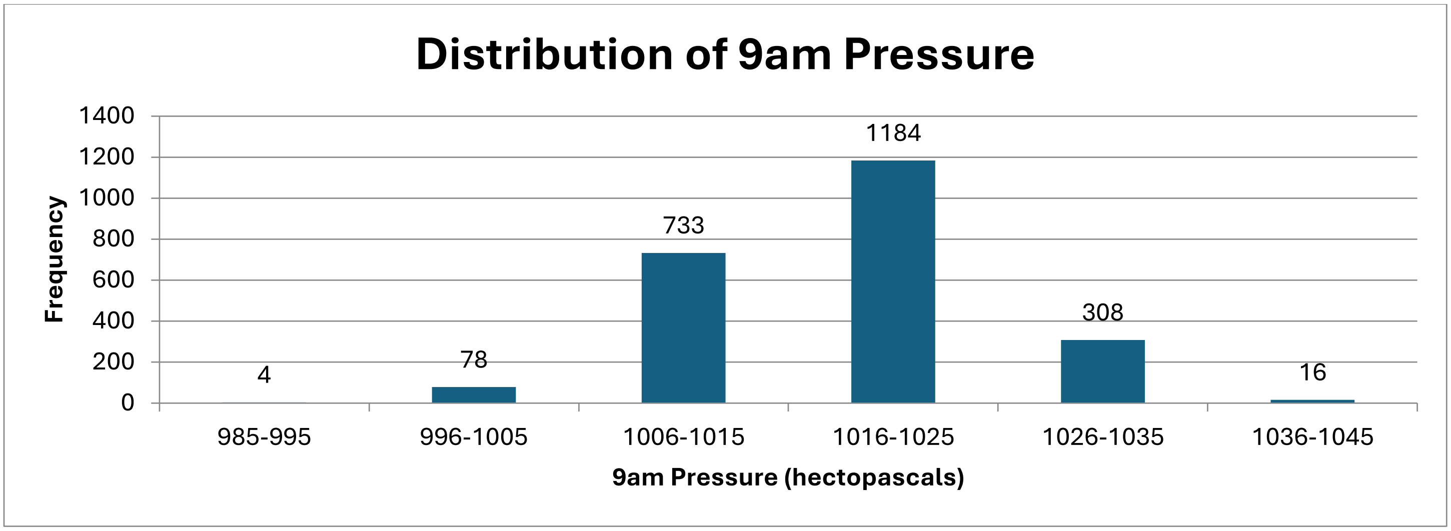


Figure 28 – Histogram of 9am Pressure

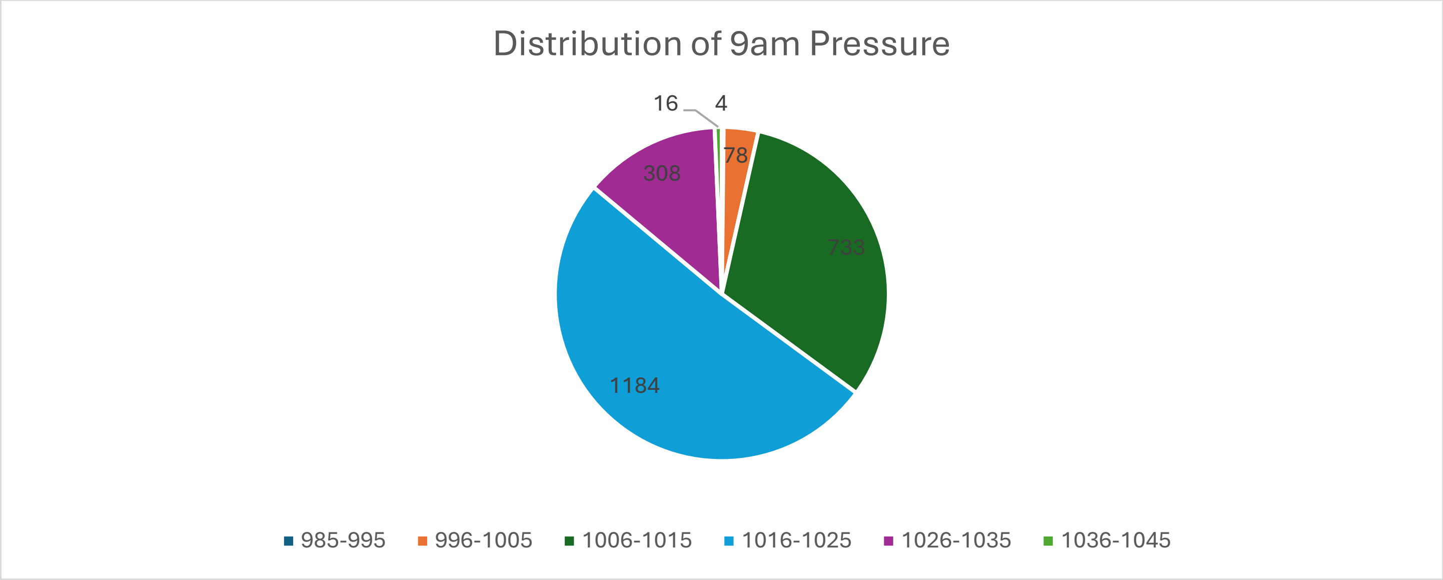


Figure 29 – Pie Chart of 9am Pressure

Interpretation:

The 9am Pressure data indicates that most pressure values fall between 1016-1025 hPa, with 1184 occurrences, followed by the 1006-1015 hPa range, which has 733 occurrences. Pressure values below 1005 hPa or above 1035 hPa are much less common. The distribution is almost symmetric, as shown by the skewness value (-0.009), suggesting a balanced spread of pressure values around the mean of 1017.742 hPa. The pie chart visually confirms this, showing the dominance of moderate pressure ranges, particularly between 1006-1025 hPa. The dataset suggests a relatively stable pressure range in the mornings, with fewer occurrences of extremely low or high pressure.

3pm Temp

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	-1.8
Max Range	45.3
Mean	21.806
Median	N/A

Standard Deviation	6.971
Variance	48.598
Skewness	0.254
Kurtosis	-0.054
Missing Values	69

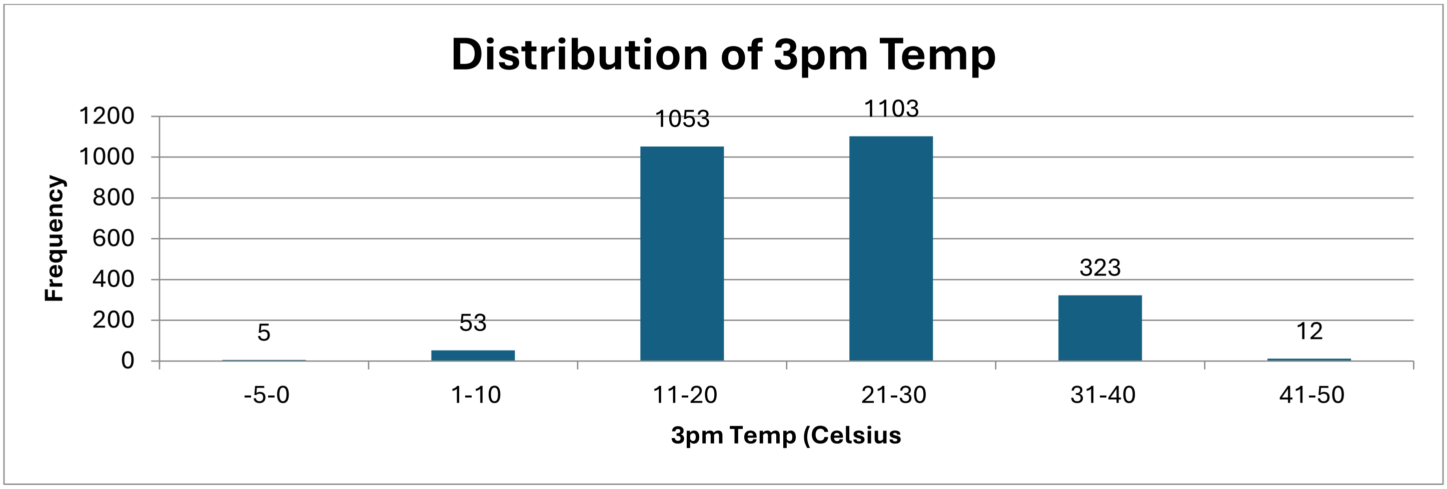


Figure 30 – Histogram of 9pm Temp

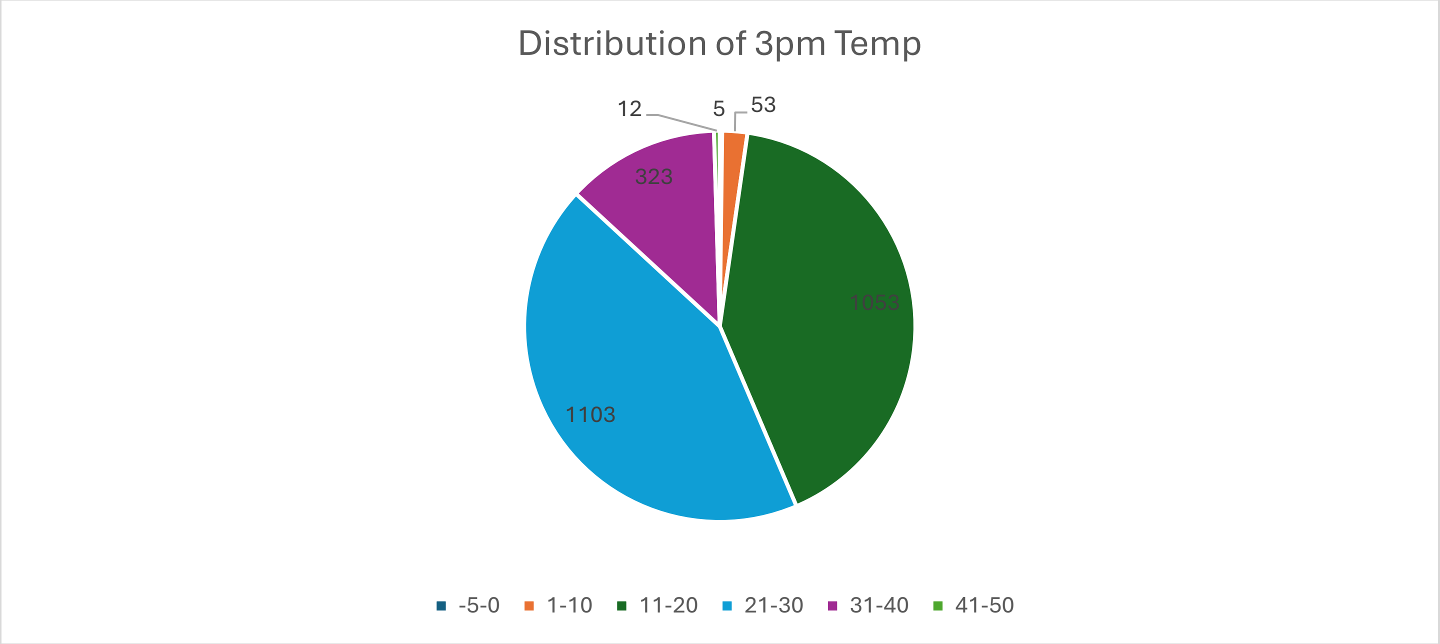


Figure 31 – Pie Chart of 3pm Temp

Interpretation:

The 3pm Temp data shows that most afternoon temperatures fall between 11-30°C, with 1103 occurrences in the 21-30°C range and 1053 occurrences in the 11-20°C range. Temperatures above 30°C are much less frequent, with only 323 occurrences between 31-40°C, and very few extremes below 0°C or above 40°C. The slight right skew (skewness: 0.254) indicates a higher occurrence of warmer temperatures, with most data clustered around the middle ranges. The pie chart reflects the dominance of the moderate temperature ranges, with small segments for colder or hotter extremes. Overall, the dataset suggests that temperatures at 3pm are generally moderate, with very few occurrences of extreme cold or heat.

3pm Humidity

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	3
Max Range	100
Mean	51.059
Median	N/A
Standard Deviation	20.592
Variance	424.01
Skewness	0.025
Kurtosis	-0.504
Missing Values	87

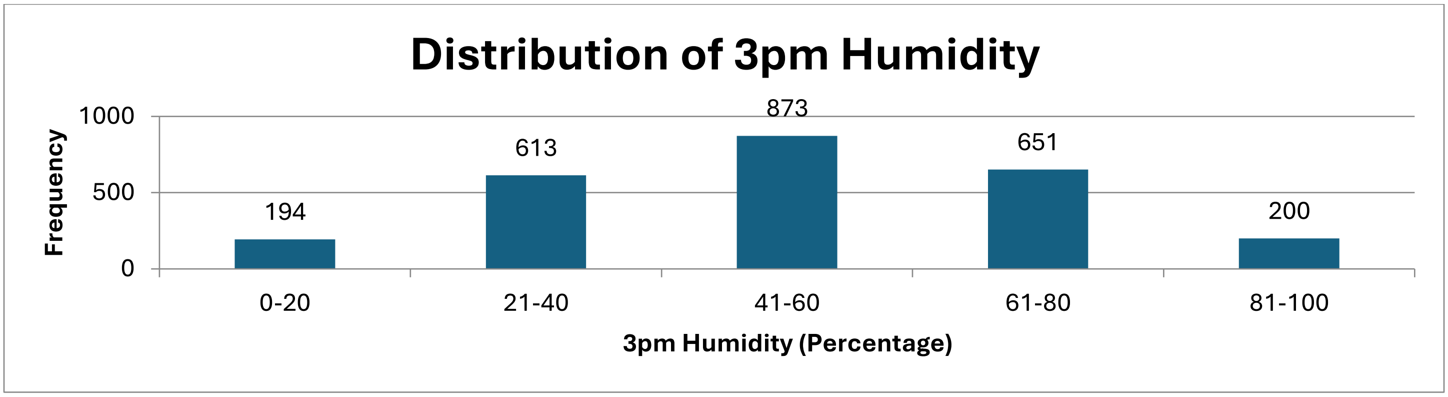


Figure 32 – Histogram of 3pm Humidity

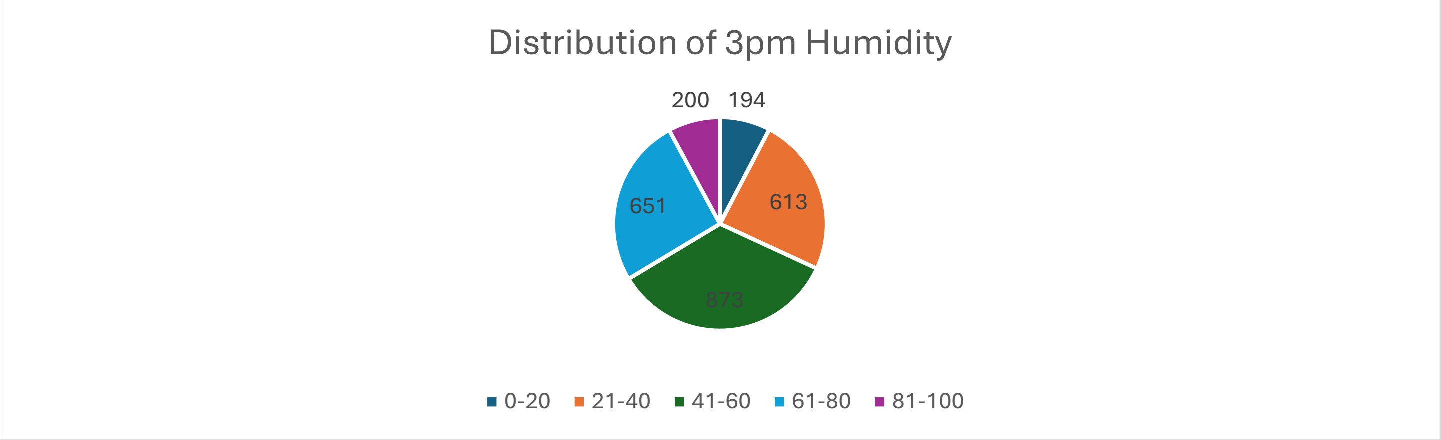


Figure 33 – Pie Chart of 3pm Humidity

Interpretation:

The 3pm Humidity data shows that the majority of humidity levels fall between 41-80%, with 873 occurrences in the 41-60% range and 651 occurrences in the 61-80% range. Humidity levels below 40% and above 80% are less common, with only 194 occurrences in the 0-20% range and 200 occurrences in the 81-100% range. The data is nearly symmetric, with minimal skewness (skewness: 0.025), indicating a fairly balanced distribution of humidity levels. The pie chart clearly visualises this, showing the dominance of moderate humidity levels, particularly between 41-80%, with smaller segments for more extreme low and high values. Overall, the data suggests that afternoons generally experience moderate humidity, with few occurrences of very low or very high humidity levels.

3pm Cloud

Frequency & Distribution Data Visualisation:

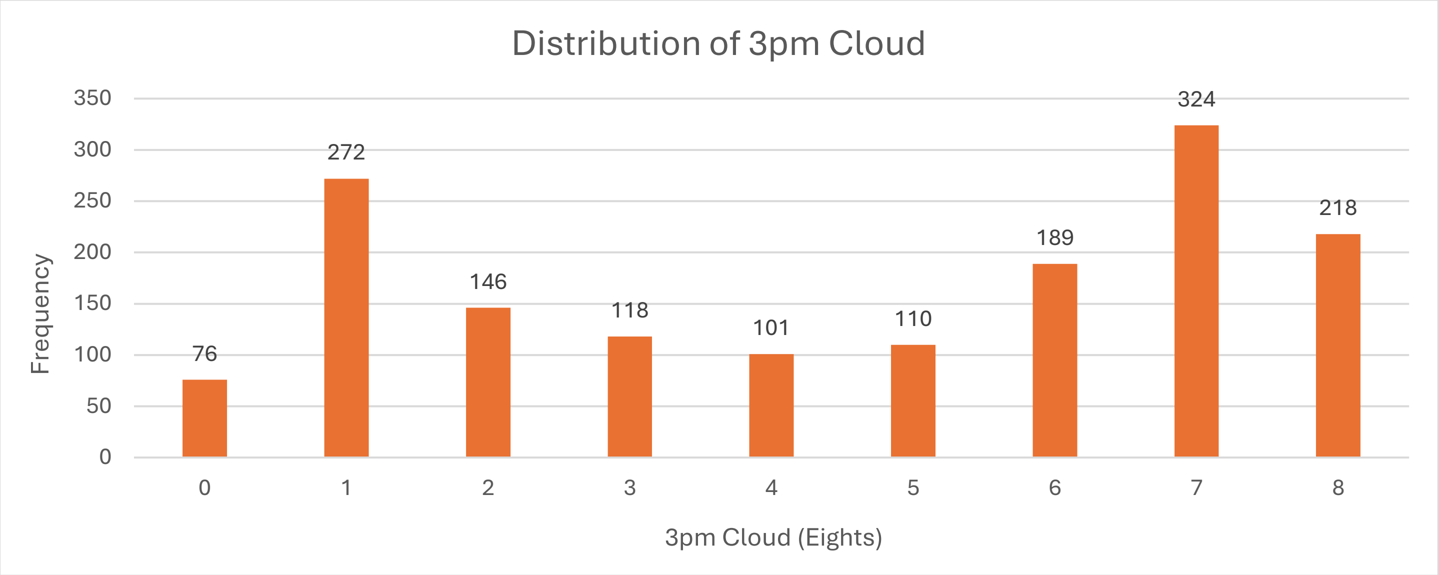


Figure 34 – Histogram of 3pm Cloud

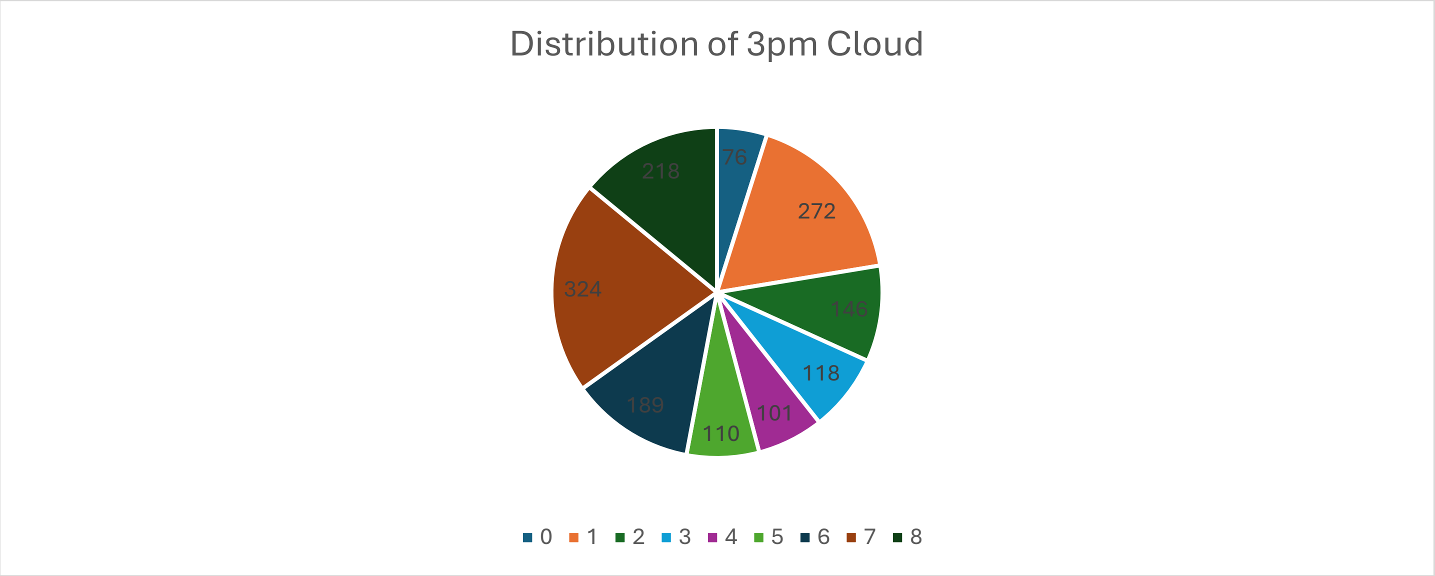


Figure 35 – Pie Chart of 3pm Cloud

Interpretation:

The 3pm Cloud data demonstrates a similar pattern to the 9am data, with cloud cover frequently observed at 7 oktas (324 occurrences) and 1 okta (272 occurrences). However, there is a slightly more pronounced distribution towards higher cloud cover at 3pm, as seen by the 218 occurrences of 8 oktas. The least frequent cloud cover is observed at 0 oktas (76 occurrences). The histogram and pie chart show that moderate cloud cover values (between 2-5 oktas) are still less frequent, although there is a slight increase compared to the 9am data. Overall, the data suggests that by 3pm, cloud cover tends to either be minimal or substantial, with a tendency towards more overcast conditions in the afternoon.

3pm WindDir

Frequency & Distribution Data Visualisation:

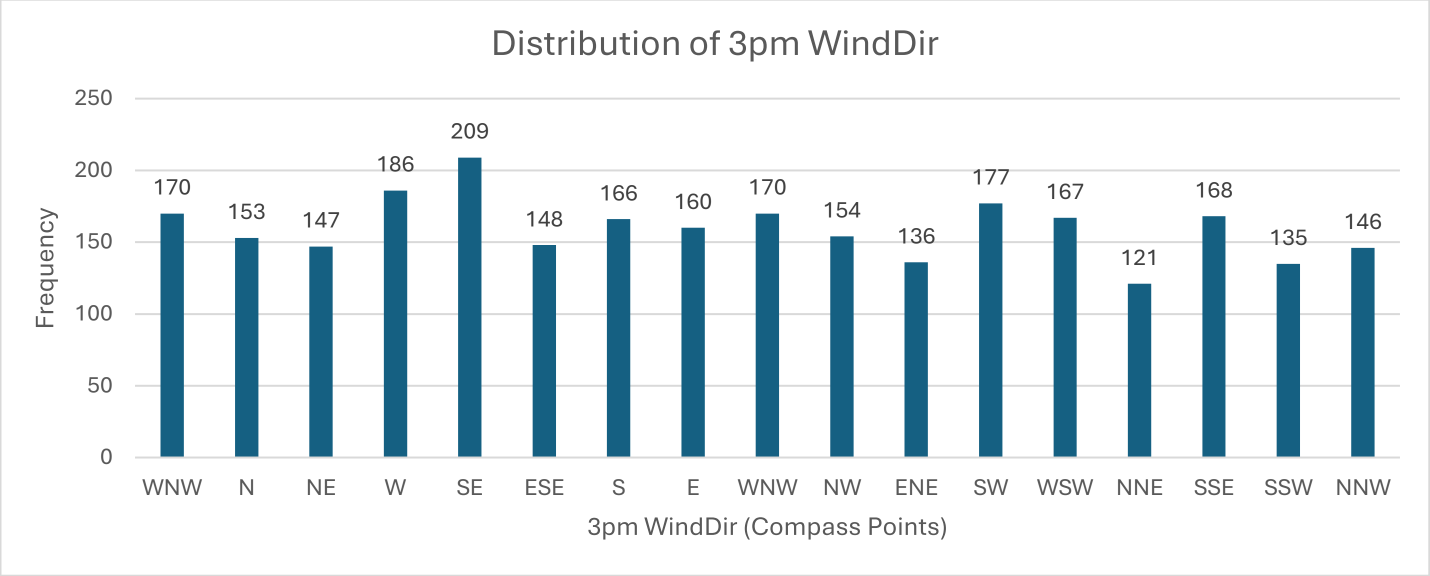


Figure 36 – Histogram of 3pm WindDir

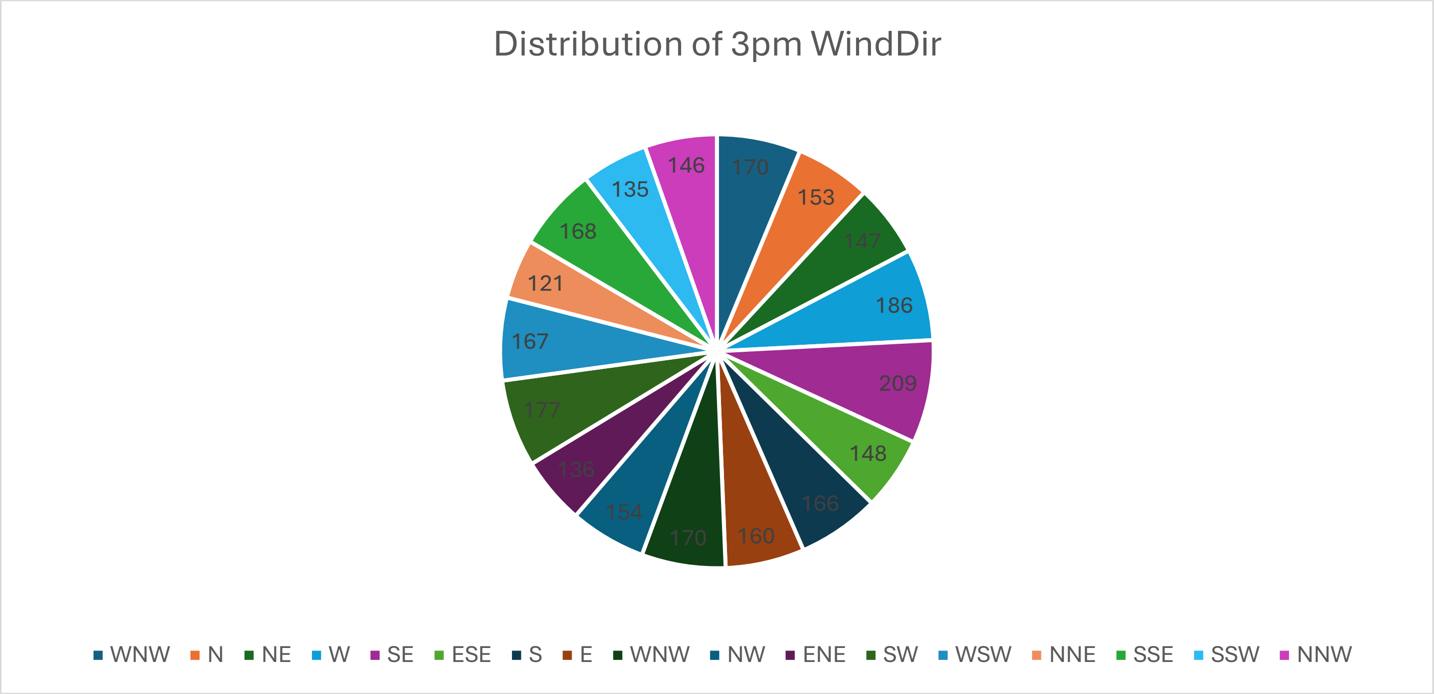


Figure 37 – Pie Chart of 3pm WindDir

Interpretation:

The 3pm WindDir data shows that the most common wind direction at 3pm is from the West (209 occurrences), followed by Southeast (186 occurrences) and Southwest (177 occurrences). The least frequent directions are from the NNE (121 occurrences) and SSW (135 occurrences). The distribution is relatively balanced across most compass points, as demonstrated by the histogram and pie chart, with no direction overwhelmingly dominating. However, there is a noticeable preference for winds coming from the west and southeast in the afternoon, suggesting that these directions are slightly more frequent. Overall, wind direction at 3pm is fairly distributed across multiple directions, with some mild preferences for specific compass points.

3pm WindSpeed

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	0
Max Range	59
Mean	18.342
Median	N/A
Standard Deviation	8.594
Variance	73.859
Skewness	0.637
Kurtosis	0.654
Missing Values	61

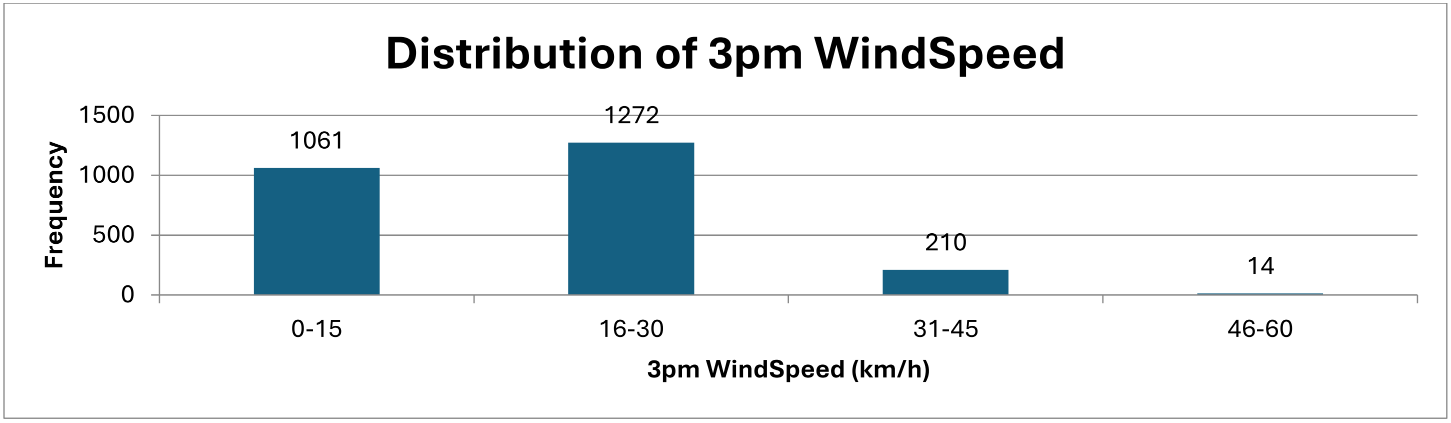


Figure 38 – Histogram of 3pm WindSpeed

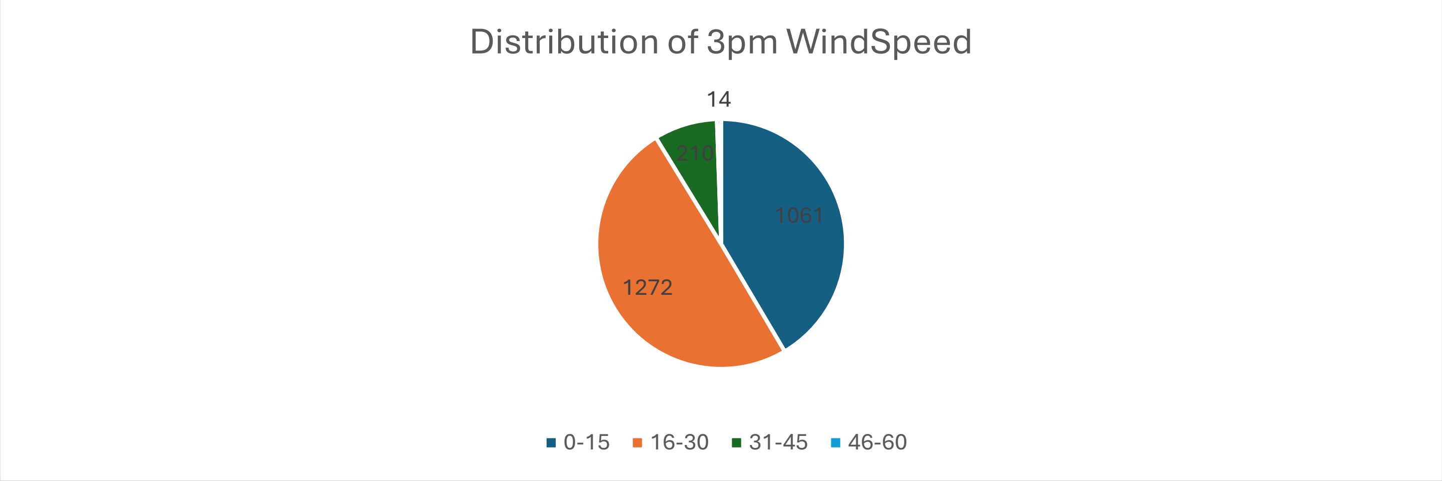


Figure 39 – Pie Chart of 3pm WindSpeed

Interpretation:

The 3pm WindSpeed data shows that most wind speeds fall between 0-30 km/h, with 1272 occurrences in the 16-30 km/h range and 1061 occurrences in the 0-15 km/h range. Higher wind speeds are less common, with only 210 occurrences between 31-45 km/h and just 14 occurrences in the 46-60 km/h range. The data has a slight right skew (skewness: 0.637), indicating that lower wind speeds are more frequent, with fewer occurrences of higher wind speeds. The pie chart reflects this distribution, highlighting the dominance of moderate wind speeds, particularly in the 16-30 km/h range. Overall, the dataset suggests that afternoons typically experience mild to moderate winds, with stronger winds being rare.

3pm Pressure

Frequency & Distribution Data Visualisation:

Statistics	Value
Min Range	988.1
Max Range	1039.6
Mean	1015.303
Median	N/A
Standard Deviation	6.862
Variance	47.086
Skewness	0.043
Kurtosis	0.167
Missing Values	290

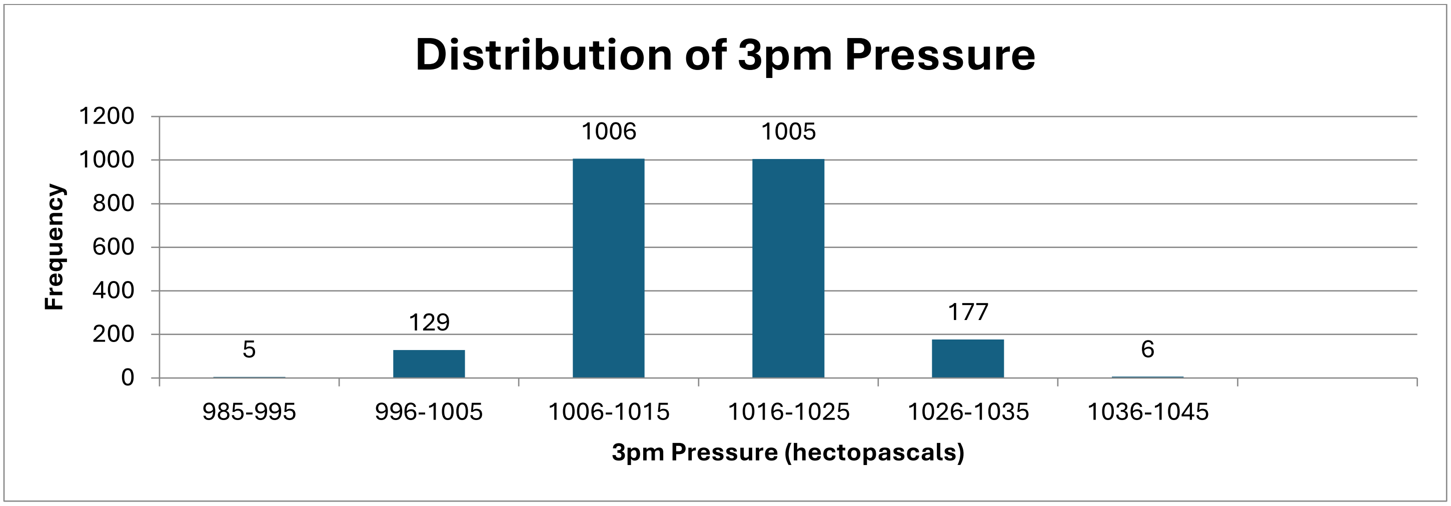


Figure 40 – Histogram of 3pm Pressure

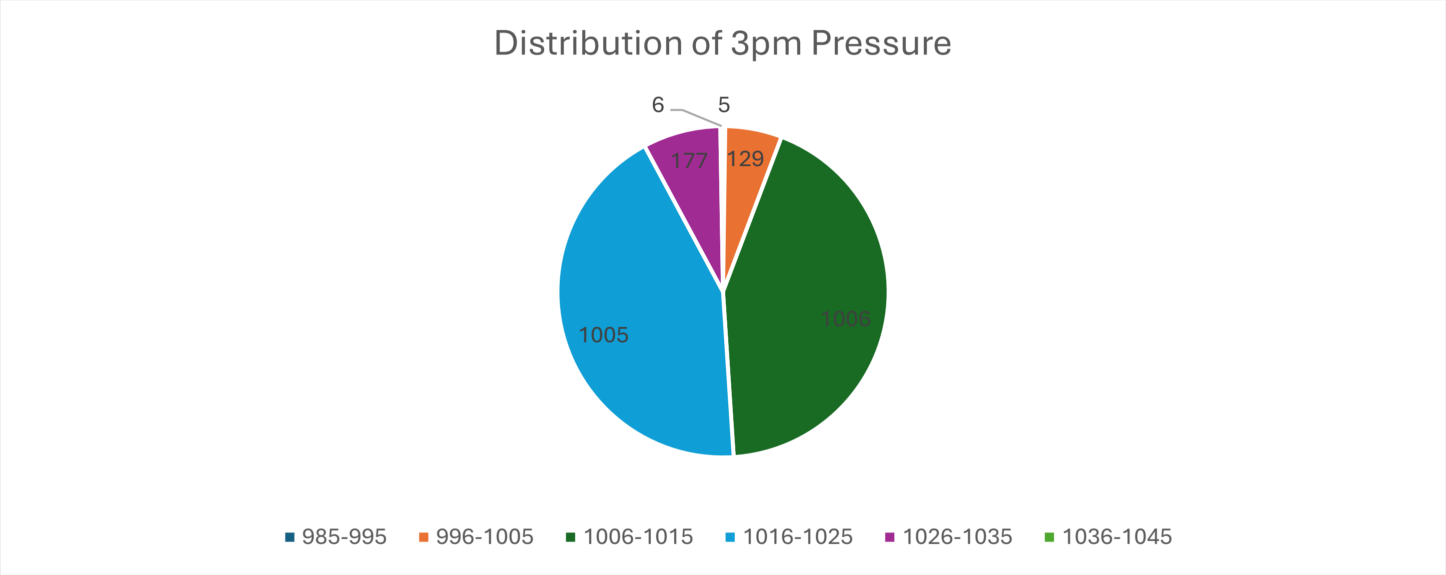


Figure 41 – Pie Chart of 3pm Pressure

Interpretation:

The 3pm Pressure data shows that most pressure values are concentrated between 1006-1025 hPa, with 1006 occurrences in the 1006-1015 hPa range and 1005 occurrences in the 1016-1025 hPa range. Pressure values outside this range are much less common, with only 177 occurrences between 1026-1035 hPa and very few instances in the extreme low or high ranges. The data shows minimal skewness (0.043), indicating a nearly symmetric distribution around the mean of 1015.303 hPa. The pie chart visually confirms this, with the majority of the data concentrated in the moderate pressure ranges, reflecting a stable atmospheric pressure pattern at 3pm.

RainToday

Frequency & Distribution Data Visualisation:

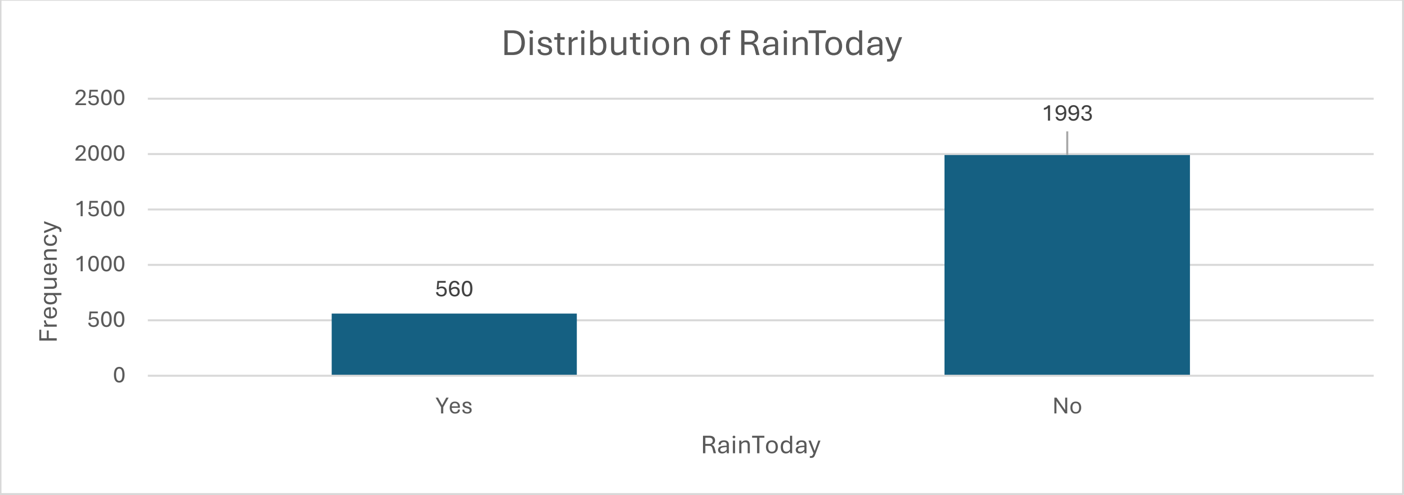


Figure 42 – Histogram of RainToday

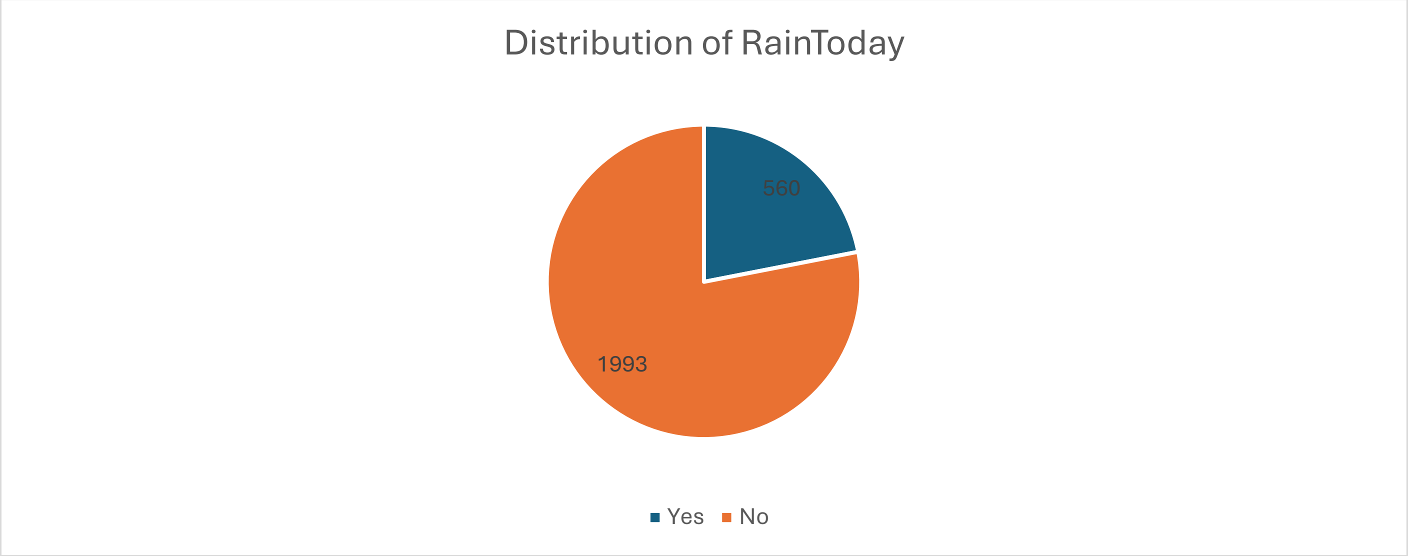


Figure 43 – Pie Chart of RainToday

Interpretation:

The RainToday data indicates that in most instances, it did not rain, with 1993 occurrences of "No" and only 560 occurrences of "Yes." This is clearly reflected in both the histogram and the pie chart, where the "No" category dominates. The dataset suggests that rainy days are less frequent compared to dry days, as the majority of the recorded instances did not experience rain. The large difference between the two categories indicates that rain is not a frequent occurrence in this dataset.

RainTomorrow

Frequency & Distribution Data Visualisation:

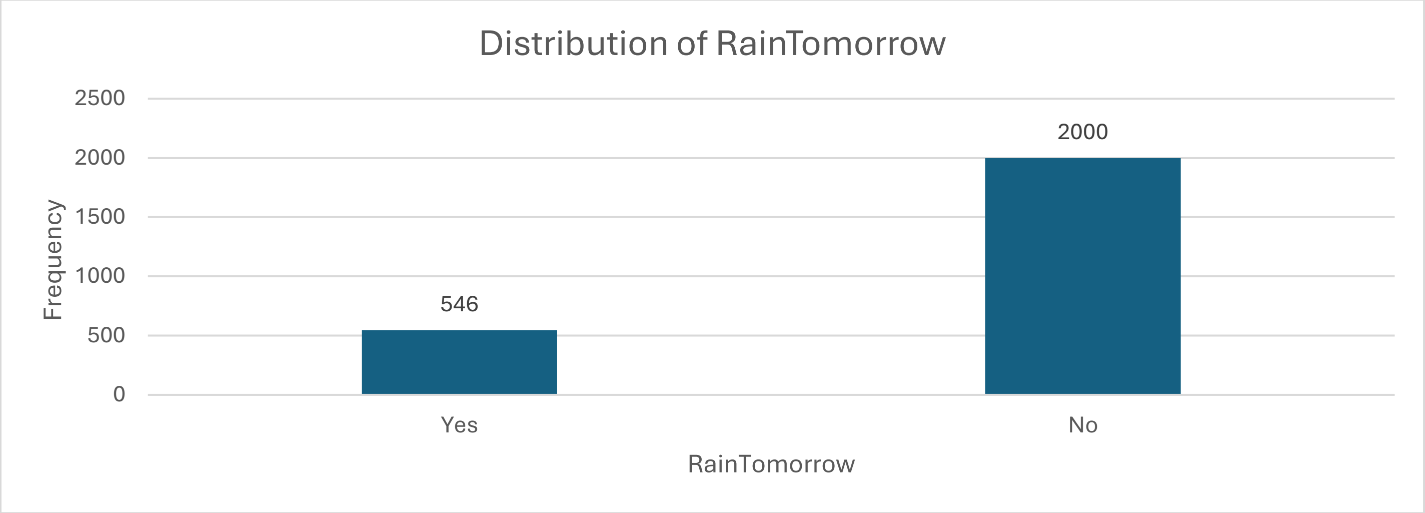


Figure 44 – Histogram of RainTomorrow

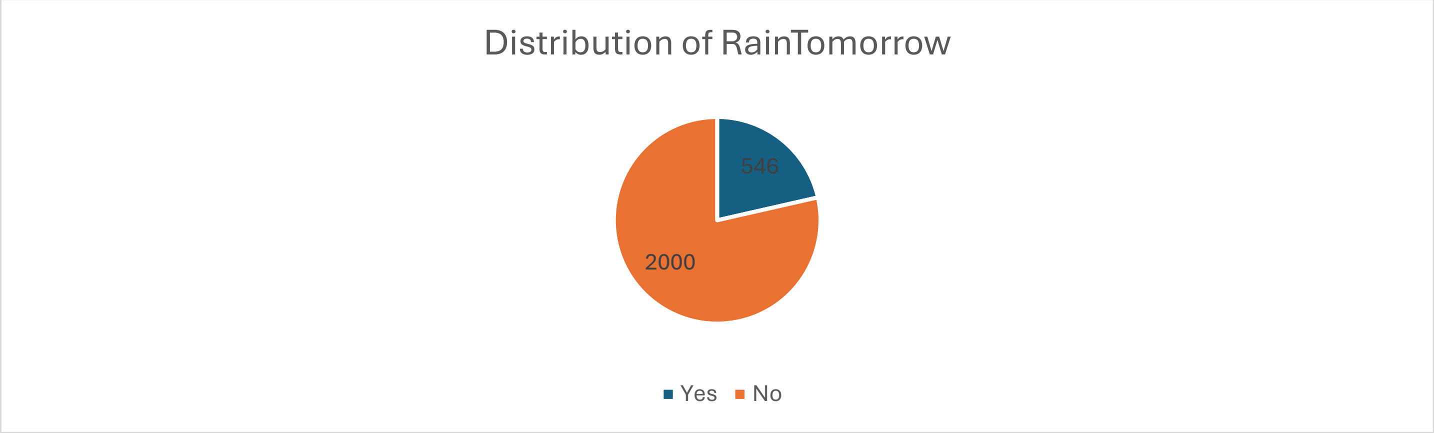


Figure 45 – Pie Chart of RainTomorrow

Interpretation:

The RainTomorrow data reveals that rain is again less frequent, with 2000 occurrences of "No" and only 546 occurrences of "Yes." Both the histogram and the pie chart highlight the significant difference, with dry days (No) dominating the dataset. This suggests that the likelihood of rain on the following day is relatively low in most cases, as the dataset reflects a climate where dry days are far more common than rainy ones.

A3 – Exploration

For A3, the dataset will be examined for patterns, irregularities, and meaningful groupings of data. The first focus is on Outliers, identifying extreme values that deviate significantly from the rest of the dataset, which can offer insights into anomalies or exceptional conditions. Following that, the analysis will explore Clusters, aiming to find natural groupings within the data. These clusters can help reveal underlying patterns, correlations, or segments within the dataset that may be useful for further analysis or decision-making.

Outliers

Rainfall:

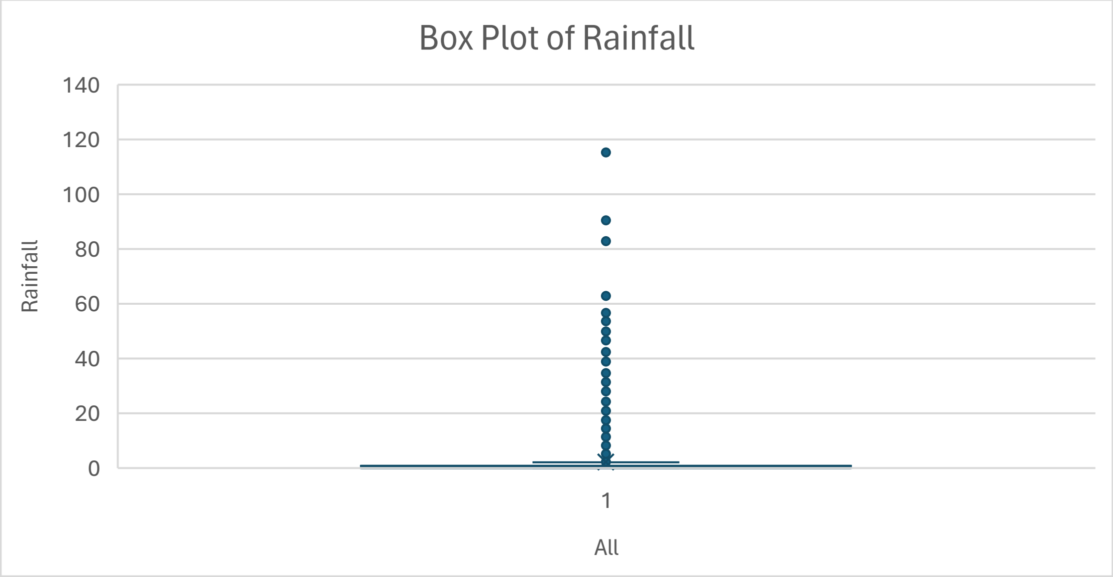


Figure 46 – Boxplot of Rainfall

The box plot for Rainfall clearly highlights the presence of several outliers. The majority of rainfall values are concentrated around the lower end of the scale (close to zero), with a few extreme outliers extending beyond 60 mm, reaching up to around 115 mm. These outliers represent unusually high rainfall events, which deviate significantly from the normal distribution of rainfall data. The clustering of points near zero further emphasises that most of the rainfall data falls within a narrow range, making the higher values stand out even more as outliers.

WindGustSpeed:

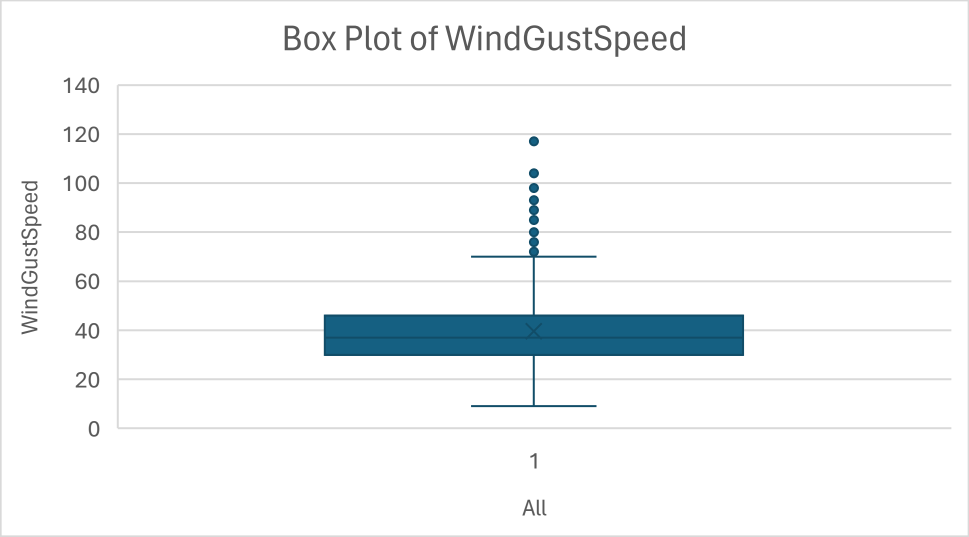


Figure 47 – Boxplot of WindGustSpeed

The box plot for WindGustSpeed reveals a notable number of outliers, with most of the wind gust speeds clustering between 20 and 60 km/h. However, several extreme values exceed 100 km/h, highlighting significant outliers. These outliers indicate unusually strong wind gusts, which deviate substantially from the typical gust speeds captured within the dataset. The spread of the data suggests that the majority of wind gust speeds fall within a relatively narrow range, while the outliers reflect rare and extreme wind events.

Evaporation:

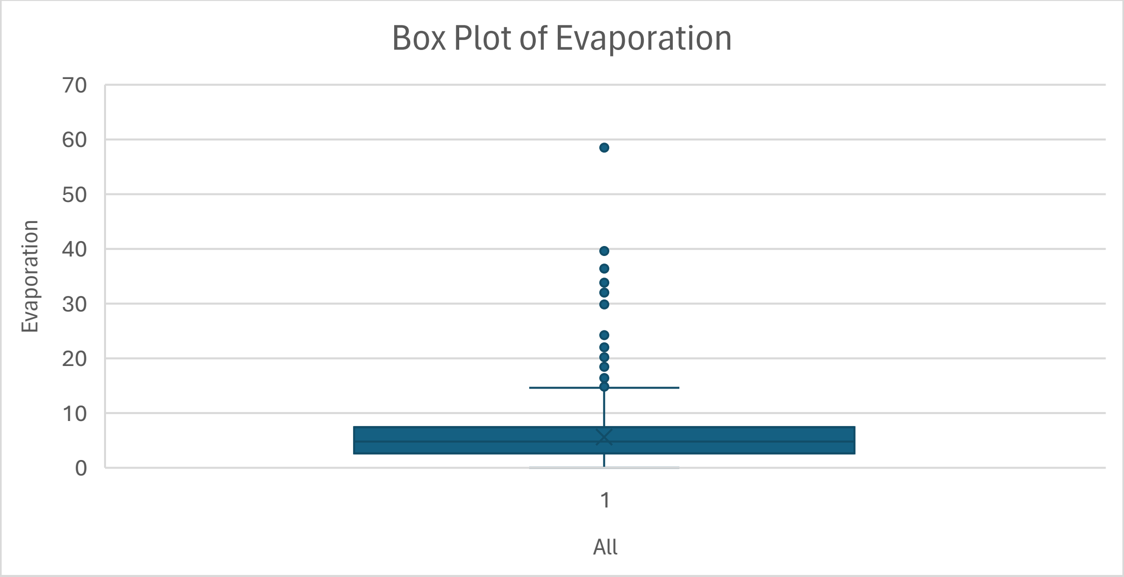


Figure 48 – Boxplot of Evaporation

The box plot for Evaporation demonstrates that the majority of data points are concentrated below 10 mm, indicating typical evaporation levels. However, a series of outliers appear, with the most extreme reaching over 60 mm. These outliers represent exceptionally high evaporation rates that deviate significantly from the central tendency of the data. The clustering around lower values suggests that high evaporation rates are rare, making these outliers noteworthy for further investigation, potentially indicating unusual weather conditions.

MaxTemp:

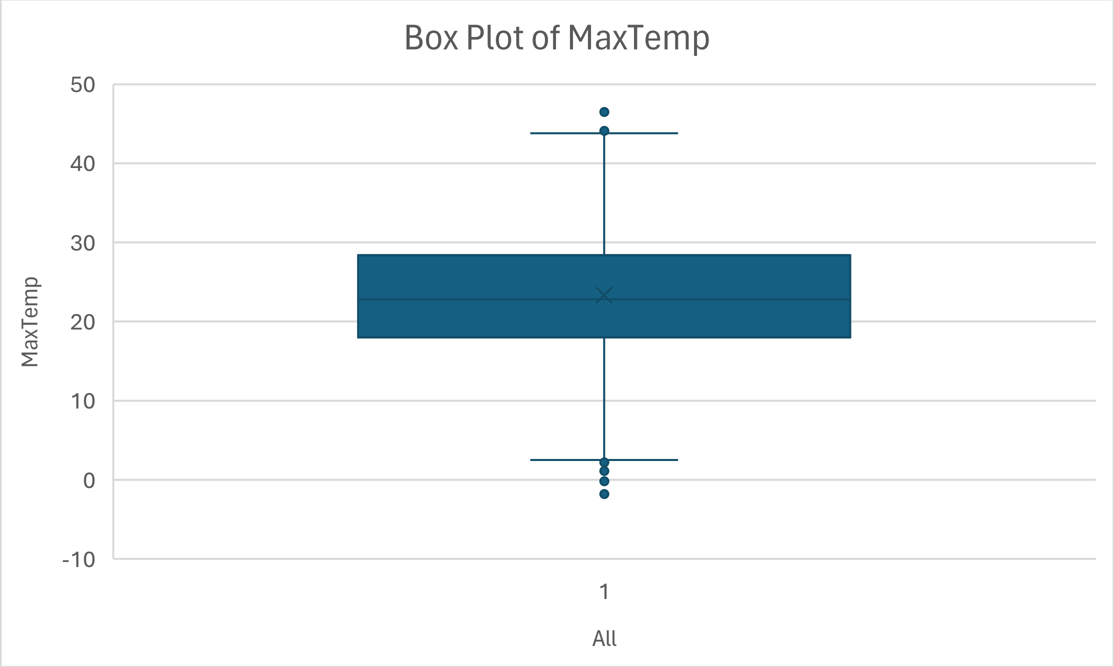


Figure 49 – Boxplot of MaxTemp

The box plot for MaxTemp shows that most maximum temperature values are distributed between 10°C and 30°C. There are a few notable outliers at both extremes, with lower temperatures nearing -10°C and higher temperatures approaching 45°C. These outliers indicate rare occurrences of extremely high or low temperatures outside the typical range. The majority of values are clustered around the interquartile range, reflecting common maximum temperatures, while the outliers provide insight into more extreme weather conditions.

Clusters

MinTemp vs MaxTemp:

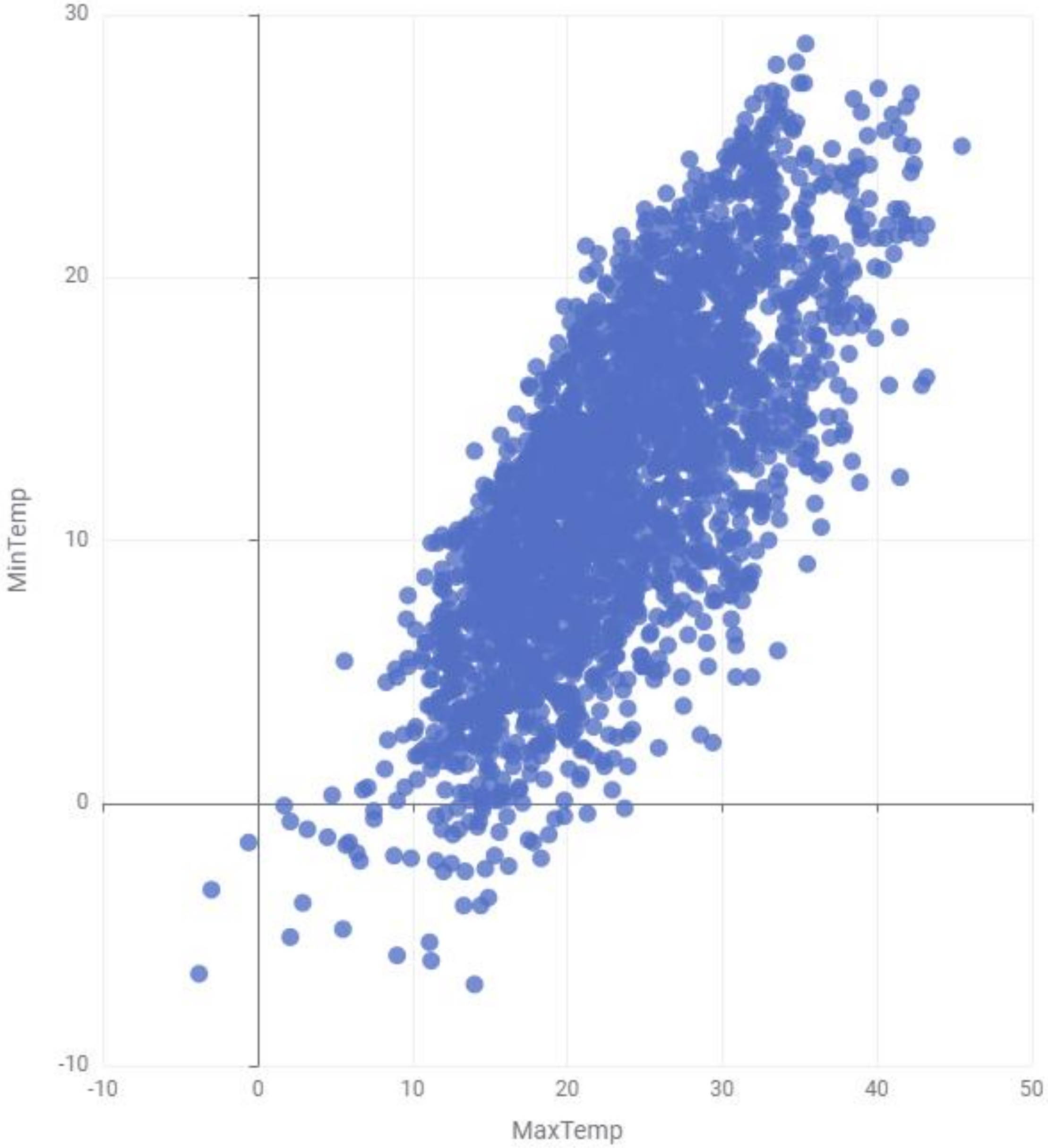


Figure 50 – Hierarchical clustering scatter plot of MinTemp vs MaxTemp

This scatter plot shows the relationship between MinTemp and MaxTemp, indicating a positive correlation between the two variables. As MaxTemp increases, MinTemp generally increases as well. The data points are fairly clustered, indicating a strong linear relationship. However, there are some outliers where MinTemp values deviate from the expected range, particularly in the lower range. The clustering suggests that for most of the days, higher maximum temperatures are associated with higher minimum temperatures, with only a few deviations from this trend. The overall distribution is concentrated within a specific range, forming a diagonal cluster from bottom-left to top-right.

Rainfall vs Sunshine:

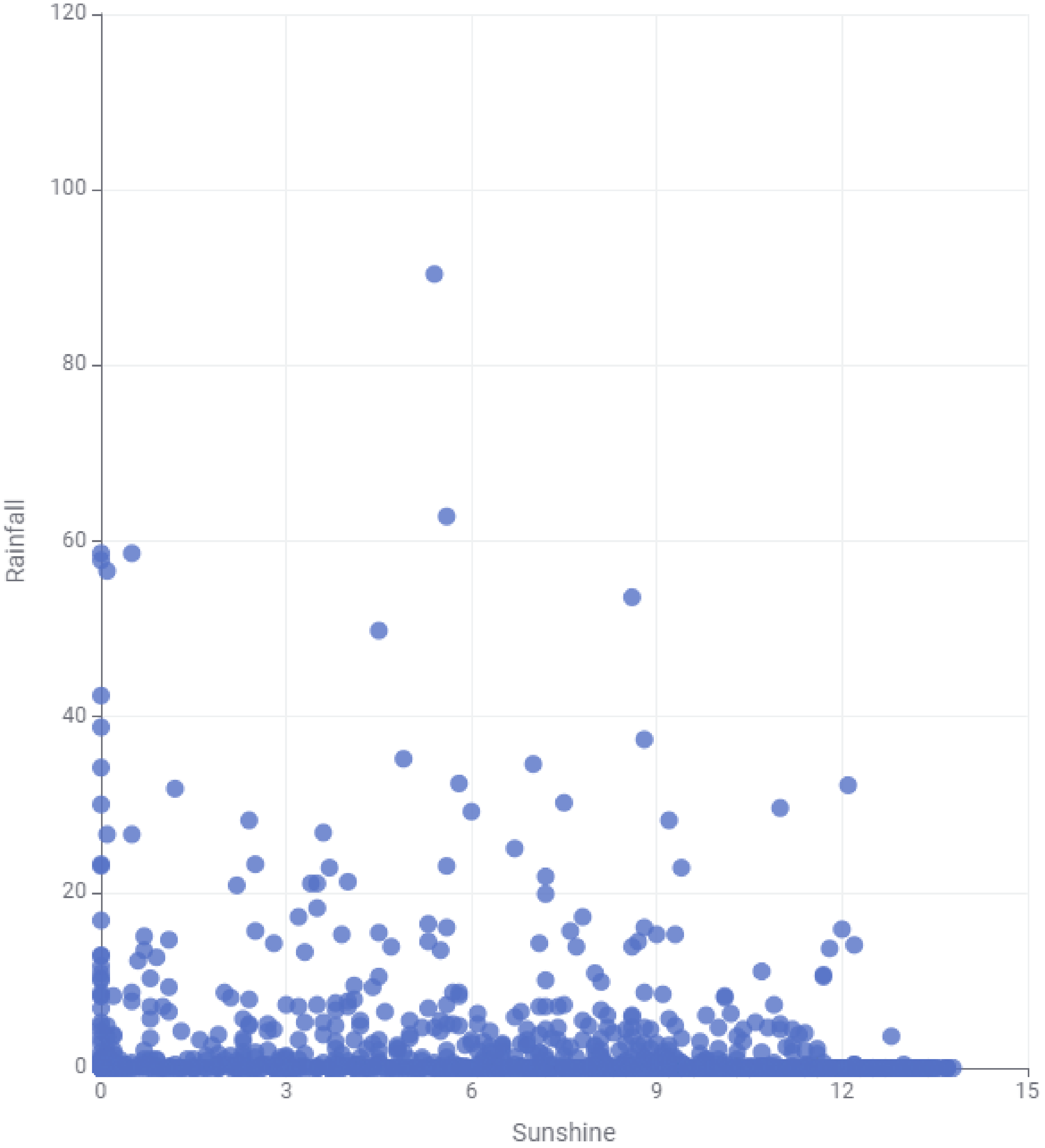


Figure 51 - Hierarchical clustering scatter plot of Rainfall vs Sunshine

The scatter plot titled Rainfall vs Sunshine presents the relationship between Rainfall and Sunshine. It displays a concentration of data points along the lower values of Sunshine, indicating that most instances of rainfall occur when there is minimal sunshine. As Sunshine increases, the number of occurrences of high rainfall decreases significantly, which is expected as higher sunshine typically correlates with drier conditions. However, there are a few outliers where rainfall is recorded despite higher sunshine levels, but these are rare. The overall trend suggests a negative relationship, where increased Sunshine is associated with decreased Rainfall.

9am Temp vs 3pm Temp:

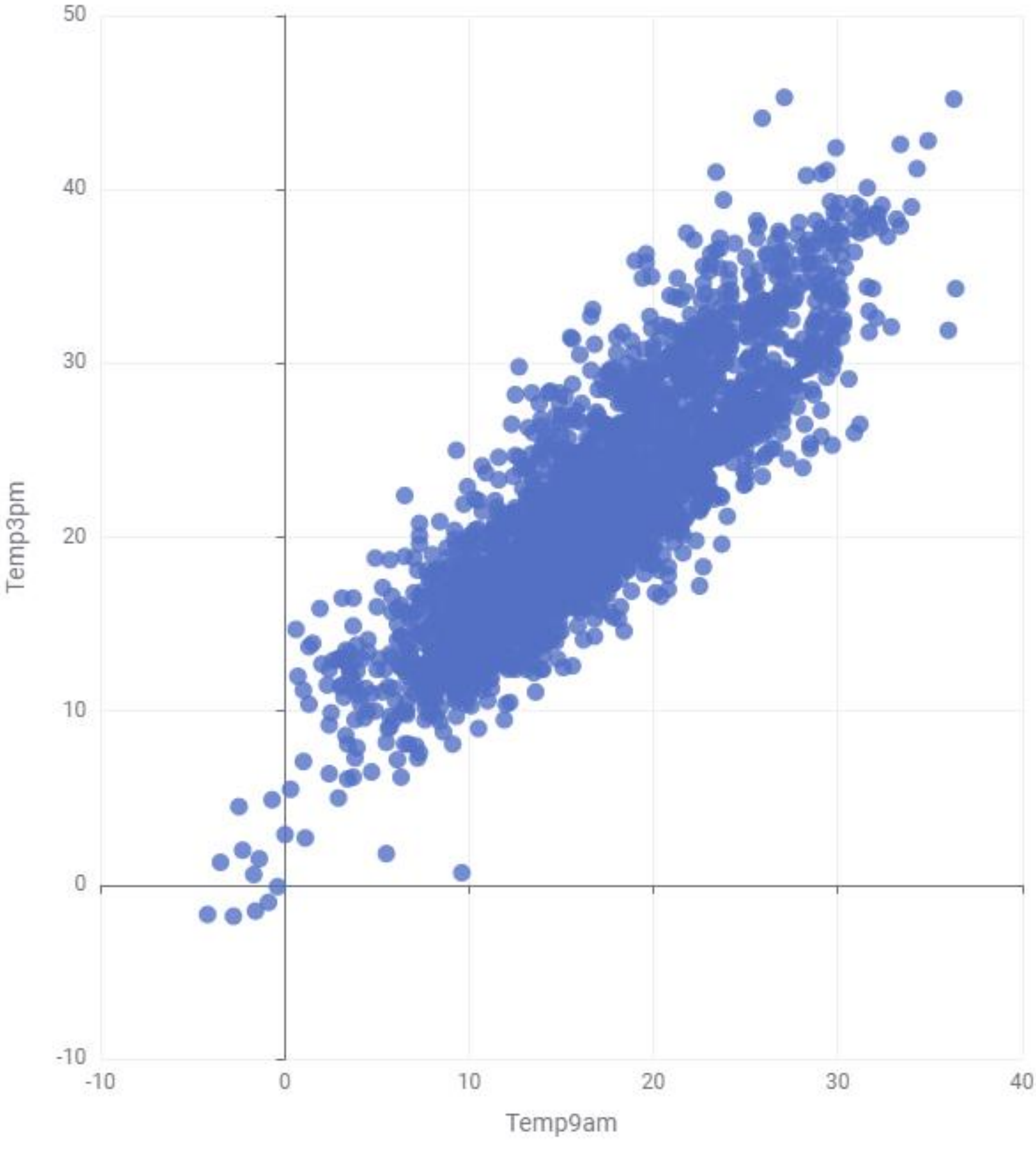


Figure 52 - Hierarchical clustering scatter plot of 9am Temp vs 3pm Temp

This scatter plot of 9am Temp vs 3pm Temp shows a positive correlation between the two temperature readings. As the temperature at 9 am increases, the 3 pm temperature also tends to increase. There is a fairly strong linear relationship, with most of the data points clustered tightly along a central line, indicating consistent daily warming patterns. A few outliers are observed where temperatures at 9 am or 3 pm deviate significantly from the central trend. These could represent unusual weather days where either morning or afternoon temperatures were notably different from the typical pattern.

WindGustSpeed vs WindGustDir:

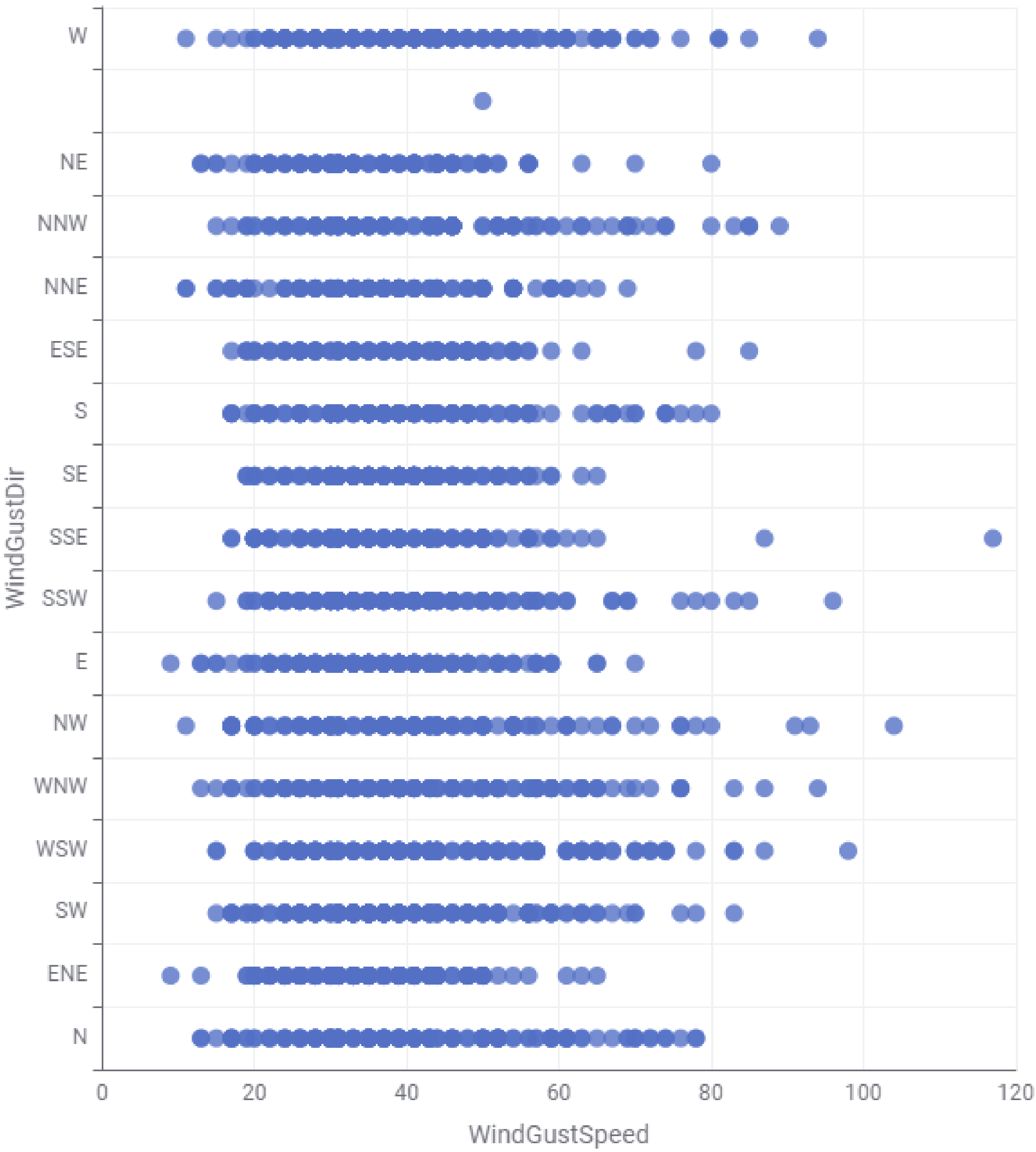


Figure 53 - Hierarchical clustering scatter plot of WindGustSpeed vs WindGustDir

The scatter plot of WindGustSpeed vs WindGustDir reveals a clear clustering of wind gust speeds within the range of 20 to 60 km/h across all compass directions. Notably, the wind gust speeds show consistency for most directions, with no distinct difference between them. However, a few outliers are present, where wind gust speeds exceed 80 km/h, particularly in directions such as N, NE, and S, indicating occasional stronger gusts in these regions. Overall, the plot indicates that moderate wind gusts (20-60 km/h) are the most common, irrespective of the wind direction. This uniform distribution suggests that wind gust speeds are not strongly dependent on wind direction, but rather, strong gusts can occur sporadically across various directions. This can be insightful for weather forecasting, as it shows typical wind behaviour across different compass points.

Rainfall vs Evaporation vs Humidity:

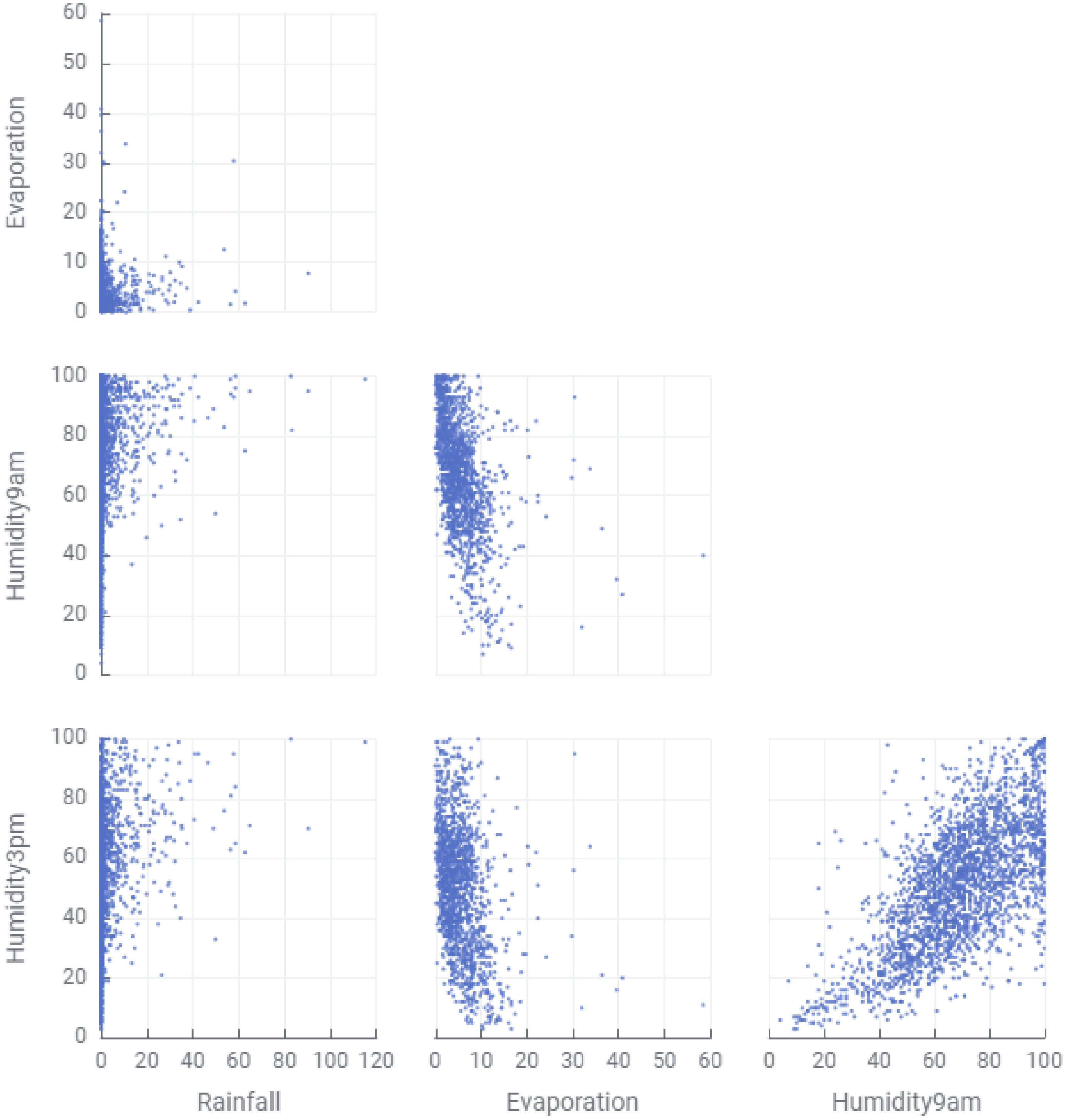


Figure 54 – Scatter plot matrix of Rainfall vs Evaporation vs Humidity (9am/3pm)

The scatter plot matrix shows pairwise relationships between Rainfall, Evaporation, Humidity at 9am, and Humidity at 3pm. Here's a breakdown of the notable patterns:

1. **Rainfall vs Evaporation:** There is a concentration of points at the lower end of the plot, indicating that high rainfall and high evaporation rarely occur together. Most data points cluster near the origin (low values of both variables), suggesting that on days with high rainfall, evaporation remains relatively low, as expected.

2. **Rainfall vs Humidity (9am and 3pm):** Both scatter plots exhibit a similar pattern, where rainfall is generally higher when humidity is closer to 100%. This suggests that on days with higher rainfall, both morning and afternoon humidity are elevated, which aligns with typical weather patterns where rainfall increases air moisture.
3. **Evaporation vs Humidity (9am and 3pm):** There is a distinct negative relationship between evaporation and humidity in both time frames (9am and 3pm). As evaporation increases, humidity tends to decrease, which makes sense as high evaporation typically occurs on drier, warmer days with lower moisture levels in the air.
4. **Humidity (9am) vs Humidity (3pm):** The scatter plot between humidity at 9am and 3pm shows a strong positive correlation, as expected. This suggests that if the humidity is high in the morning, it is likely to remain high in the afternoon, reflecting consistent moisture levels throughout the day.

Overall, the matrix provides insight into how Rainfall, Evaporation, and Humidity variables relate to each other. The patterns align with typical meteorological behaviour, where rainfall increases humidity, and high evaporation occurs under drier, lower-humidity conditions.

B – Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for analysis. This phase ensures that the data is clean, consistent, and well-structured for more accurate results. Various techniques are applied to transform the data, including smoothing, normalising, discretising, and binarising certain attributes. In this section, several tasks will be performed on key attributes from the dataset to improve the data quality and representation. These tasks include using binning techniques to smooth the Rainfall attribute, normalising MaxTemp, discretising the WindSpeed3pm attribute into distinct categories, and binarising the WindDir9am variable. Each technique will be explained in detail with results presented in the assignment report and corresponding Excel workbook.

B1 – Binning Techniques

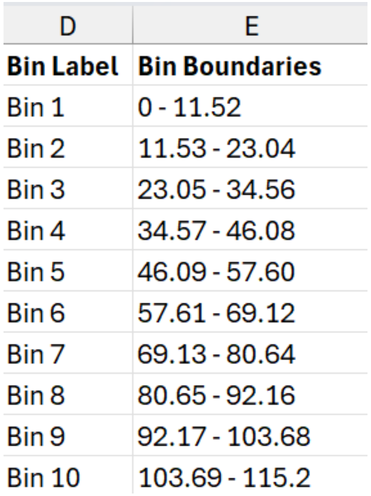
Binning is a technique used to smoothen data by grouping continuous values into bins. In this subsection, two types of binning methods, Equi-width and Equi-depth binning, are applied to the Rainfall attribute. These methods help in understanding the distribution of data and in reducing noise, making it easier to interpret patterns and relationships.

Equi-Width

Equi-width binning divides the range of values into equal-sized intervals. The following steps were completed to perform Equi-width binning on the Rainfall attribute:

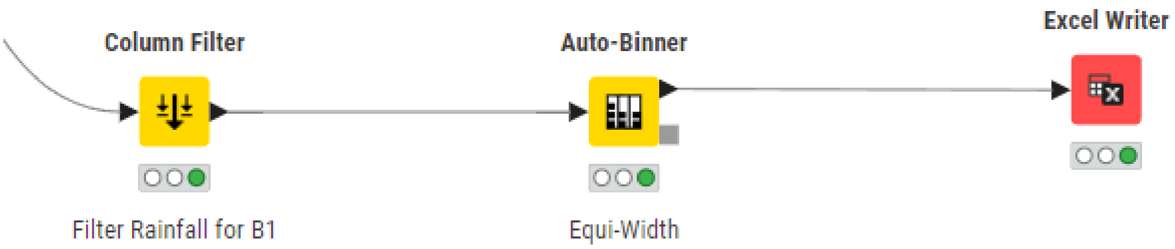
1. **Deciding Number of Bins:** The first step was to decide the number of bins. Since the range of the Rainfall attribute was from the minimum of 0 to the maximum of 115.2, I opted to use 10 bins. This decision was based on judgment to provide enough granularity while keeping the process simple for interpretation.
2. **Calculating the Bin Width:** To calculate the width of each bin, the range of Rainfall values was divided by the number of bins. The formula for this calculation was: (115.2 – 0 / 10 = 11.52mm). Each bin now represents an interval of 11.52 mm of rainfall, starting from 0 and extending to the maximum value.
3. **Assigning Bin Labels:** Bin labels were created for each bin, representing the corresponding ranges. For example:
    a. Bin 1: 0 – 11.52 mm
    b. Bin 2: 11.52 – 23.04 mm
    c. Bin 3: 23.04 – 34.56 mm

d. And so forth, continuing for all 10 bins.

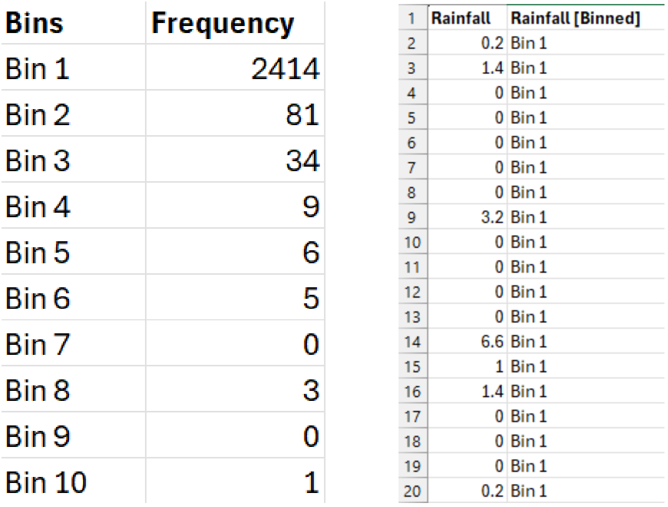


D	E
Bin Label	**Bin Boundaries**
Bin 1	0 - 11.52
Bin 2	11.53 - 23.04
Bin 3	23.05 - 34.56
Bin 4	34.57 - 46.08
Bin 5	46.09 - 57.60
Bin 6	57.61 - 69.12
Bin 7	69.13 - 80.64
Bin 8	80.65 - 92.16
Bin 9	92.17 - 103.68
Bin 10	103.69 - 115.2

4. **Performing Binning in KNIME:** In KNIME, the Binning node was used to apply equi-width binning to the Rainfall attribute. The Binning node was configured to divide the Rainfall data into 10 bins using the recalculated bin width of 11.52 mm.



5. **Exporting the Results:** After performing the binning operation, the results were exported to an Excel file using the Excel Writer node. The binned values were placed in a new column titled "Rainfall [Binned]" in the Excel spreadsheet, with each rainfall value assigned to its corresponding bin based on the calculated bin ranges. After performing the equi-width binning on the rainfall data, I calculated the frequency of each bin to understand the distribution of values within each bin. The frequency table below shows how many values fall into each bin:



Bins	**Frequency**
Bin 1	2414
Bin 2	81
Bin 3	34
Bin 4	9
Bin 5	6
Bin 6	5
Bin 7	0
Bin 8	3
Bin 9	0
Bin 10	1

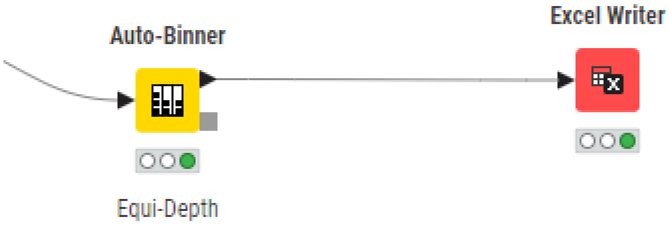
1	**Rainfall**	**Rainfall [Binned]**
2	0.2	Bin 1
3	1.4	Bin 1
4	0	Bin 1
5	0	Bin 1
6	0	Bin 1
7	0	Bin 1
8	0	Bin 1
9	3.2	Bin 1
10	0	Bin 1
11	0	Bin 1
12	0	Bin 1
13	0	Bin 1
14	6.6	Bin 1
15	1	Bin 1
16	1.4	Bin 1
17	0	Bin 1
18	0	Bin 1
19	0	Bin 1
20	0.2	Bin 1

Figure 55 – Equi-Width Frequency Table + 20 Rows of Equi-Width Data

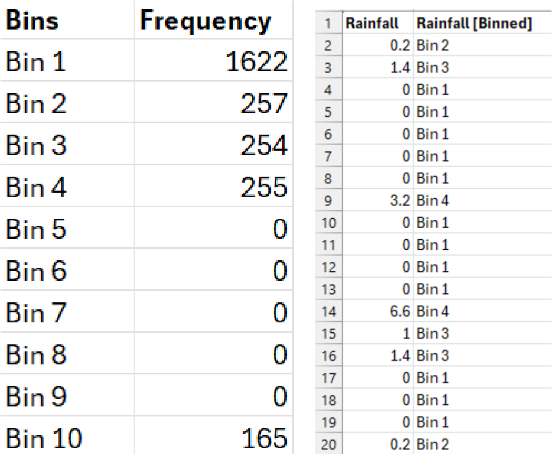
Equi-Depth

Equi-depth binning divides the range of values into intervals containing equal number of data points. The following steps were done to apply Equi-depth binning on the Rainfall attribute:

1. **Deciding Number of Bins:** I opted to use 10 bins for equi-depth binning, ensuring that each bin would contain an approximately equal number of values. I utilised the same number of bins from equi-width to ensure accuracy and same level of detail.
2. **Calculating the Depth Width:** The total number of data points for Rainfall is 2553 (excluding missing data). Dividing this by 10 gives us the number of values that should approximately fall into each bin: (2553 / 10 = 255.3). Therefore, each bin contains approximately 255 values.
3. **Assigning Bin Labels**: Each bin was assigned a label in the resulting column, reflecting the order of values. The labels were created for each bin, representing the corresponding ranges based on sorted values:
    a. Bin 1: Values 1 to 255
    b. Bin 2: Values 256 to 510
    c. Bin 3: Values 511 to 765
    d. And so on, for all 10 bins
4. **Performing Binning in KNIME**: In KNIME, the Auto-Binner node was used to apply equi-depth binning. The node was set to distribute the data into 10 bins, ensuring each bin had an equal number of data points. The Rainfall attribute was selected, and the "Frequency" (equi-depth) binning option was chosen.



5. **Exporting the Results**: After performing the equi-depth binning operation, the results were exported to an Excel file using the Excel Writer node. The binned values were placed in a new column labeled "Rainfall [Binned]" in the Excel spreadsheet. Each Rainfall value was assigned to its corresponding bin based on the calculated bin ranges. A frequency table showing how many values fall into each bin was also generated.



Bins	Frequency
Bin 1	1622
Bin 2	257
Bin 3	254
Bin 4	255
Bin 5	0
Bin 6	0
Bin 7	0
Bin 8	0
Bin 9	0
Bin 10	165

1	Rainfall	Rainfall [Binned]
2	0.2	Bin 2
3	1.4	Bin 3
4	0	Bin 1
5	0	Bin 1
6	0	Bin 1
7	0	Bin 1
8	0	Bin 1
9	3.2	Bin 4
10	0	Bin 1
11	0	Bin 1
12	0	Bin 1
13	0	Bin 1
14	6.6	Bin 4
15	1	Bin 3
16	1.4	Bin 3
17	0	Bin 1
18	0	Bin 1
19	0	Bin 1
20	0.2	Bin 2

Figure 56 – Equi-Depth Frequency Table + 20 Rows of Equi-Depth Data

B2 – Normalisation

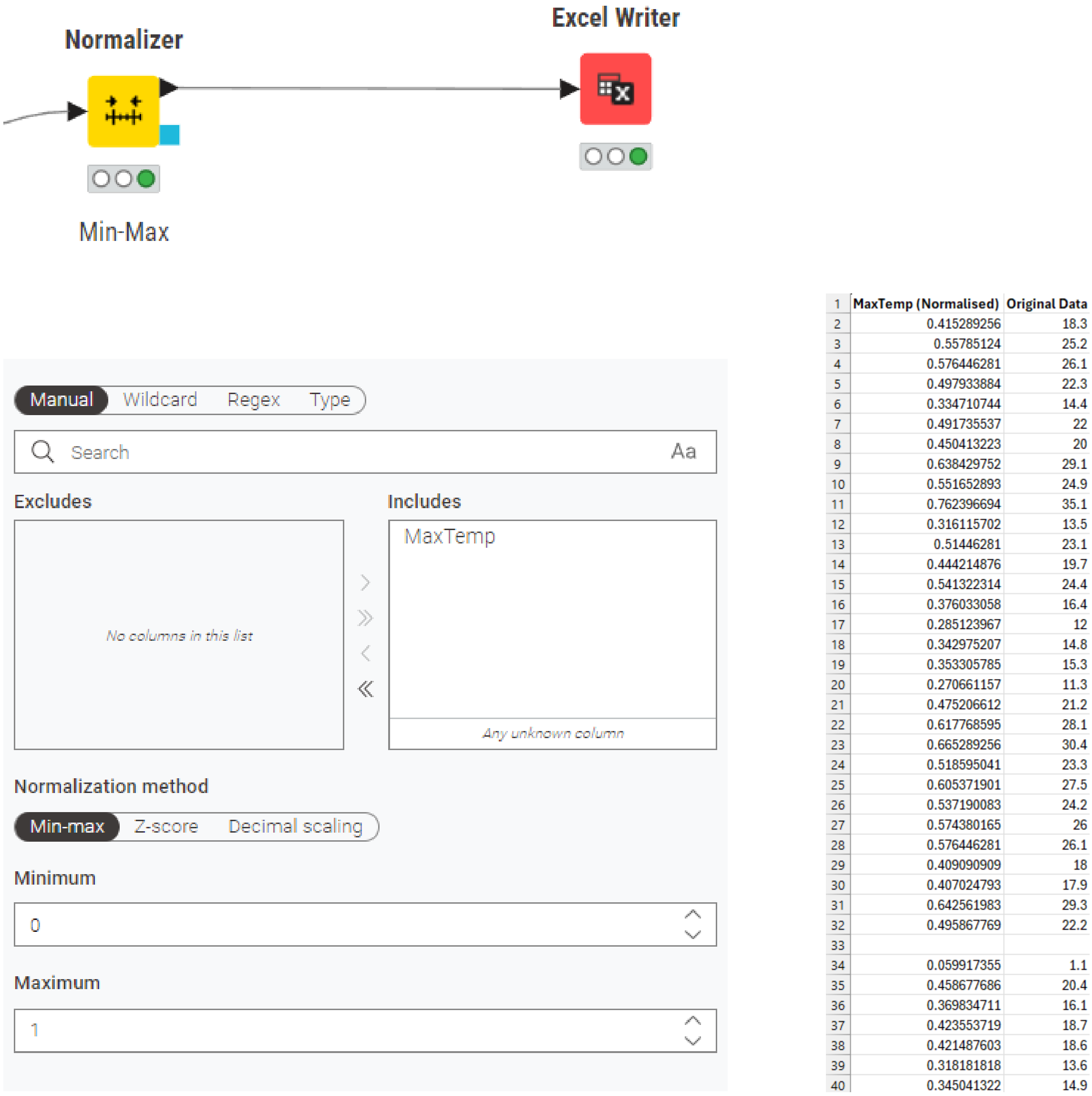
In this section, normalisation techniques are applied to the "MaxTemp" attribute to standardise the data, making it easier to compare values across different ranges. Two techniques are employed: min-max normalisation and z-score normalisation.

Min-Max Normalisation

This method scales the data to a fixed range between 0 and 1. The formula used for this transformation is:

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

Where X′ is the new scaled value, X is the original value, and Min(X) and Max(X) are the minimum and maximum values in the data. This method ensures all the values of "MaxTemp" are within the range [0,1], making it more interpretable for further analysis.



Includes: MaxTemp

Normalization method: Min-max | Z-score | Decimal scaling (Min-max selected)

Minimum: 0

Maximum: 1

1	MaxTemp (Normalised)	Original Data
2	0.415289256	18.3
3	0.55785124	25.2
4	0.576446281	26.1
5	0.497933884	22.3
6	0.334710744	14.4
7	0.491735537	22
8	0.450413223	20
9	0.638429752	29.1
10	0.551652893	24.9
11	0.762396694	35.1
12	0.316115702	13.5
13	0.51446281	23.1
14	0.444214876	19.7
15	0.541322314	24.4
16	0.376033058	16.4
17	0.285123967	12
18	0.342975207	14.8
19	0.353305785	15.3
20	0.270661157	11.3
21	0.475206612	21.2
22	0.617768595	28.1
23	0.665289256	30.4
24	0.518595041	23.3
25	0.605371901	27.5
26	0.537190083	24.2
27	0.574380165	26
28	0.576446281	26.1
29	0.409090909	18
30	0.407024793	17.9
31	0.642561983	29.3
32	0.495867769	22.2
33		
34	0.059917355	1.1
35	0.458677686	20.4
36	0.369834711	16.1
37	0.423553719	18.7
38	0.421487603	18.6
39	0.318181818	13.6
40	0.345041322	14.9

Figure 57 – KNIME Nodes + Normalizer Configuration Window + 40 Rows of Min-Max Data

The following steps were completed to perform min-max normalisation on MaxTemp:

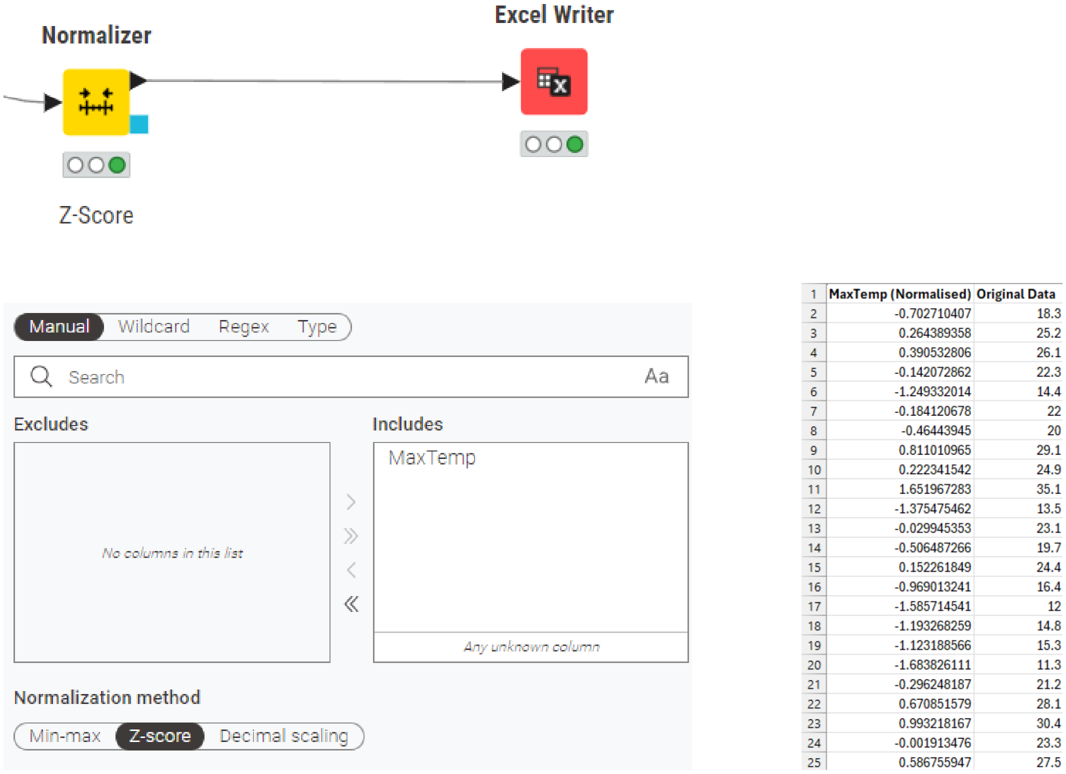
1. **Selecting the MaxTemp Attribute:** The first step was selecting the MaxTemp column as the attribute to apply Min-Max normalisation. This column was chosen to transform its values to a new range between [0, 1].
2. **Adding the Normaliser Node:** I added the normaliser node to the workflow. This node is responsible for performing the normalization process on selected data.
3. **Setting Min-Max Normalisation**: In the normaliser node, you selected the Min-Max normalisation option to map the values of MaxTemp into the range of 0 to 1. This method scales the data linearly so that the smallest value becomes 0 and the largest value becomes 1, with all other values adjusted proportionally in between.
4. **Exporting the Data to Excel**: After configuring the normaliser node, I executed it, and the normalisation process was applied to the selected MaxTemp column, converting its values into the desired range. The results were exported to Excel, and the file contains the normalised values alongside the original data in separate columns for easy comparison.

Z-Score Normalisation

This method transforms the data by centering it around a mean of 0 and scaling it based on standard deviation, resulting in a distribution where most values lie between -3 and +3. The formula used for this is:

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

Where Z is the normalised value, X is the original value, μ is the mean of the dataset, and σ (sigma) is the standard deviation. This method is useful when comparing "MaxTemp" values relative to the dataset as a whole.



Includes: MaxTemp

Normalization method: Z-score

1	MaxTemp (Normalised)	Original Data
2	-0.702710407	18.3
3	0.264389358	25.2
4	0.390532806	26.1
5	-0.142072862	22.3
6	-1.249332014	14.4
7	-0.184120678	22
8	-0.46443945	20
9	0.811010965	29.1
10	0.222341542	24.9
11	1.651967283	35.1
12	-1.375475462	13.5
13	-0.029945353	23.1
14	-0.506487266	19.7
15	0.152261849	24.4
16	-0.969013241	16.4
17	-1.585714541	12
18	-1.193268259	14.8
19	-1.123188566	15.3
20	-1.683826111	11.3
21	-0.296248187	21.2
22	0.670851579	28.1
23	0.993218167	30.4
24	-0.001913476	23.3
25	0.586755947	27.5

Figure 58 – KNIME Nodes + Normalizer Configuration Window + 25 Rows of Z-Score Data

The following steps were completed to perform z-score normalisation on MaxTemp:

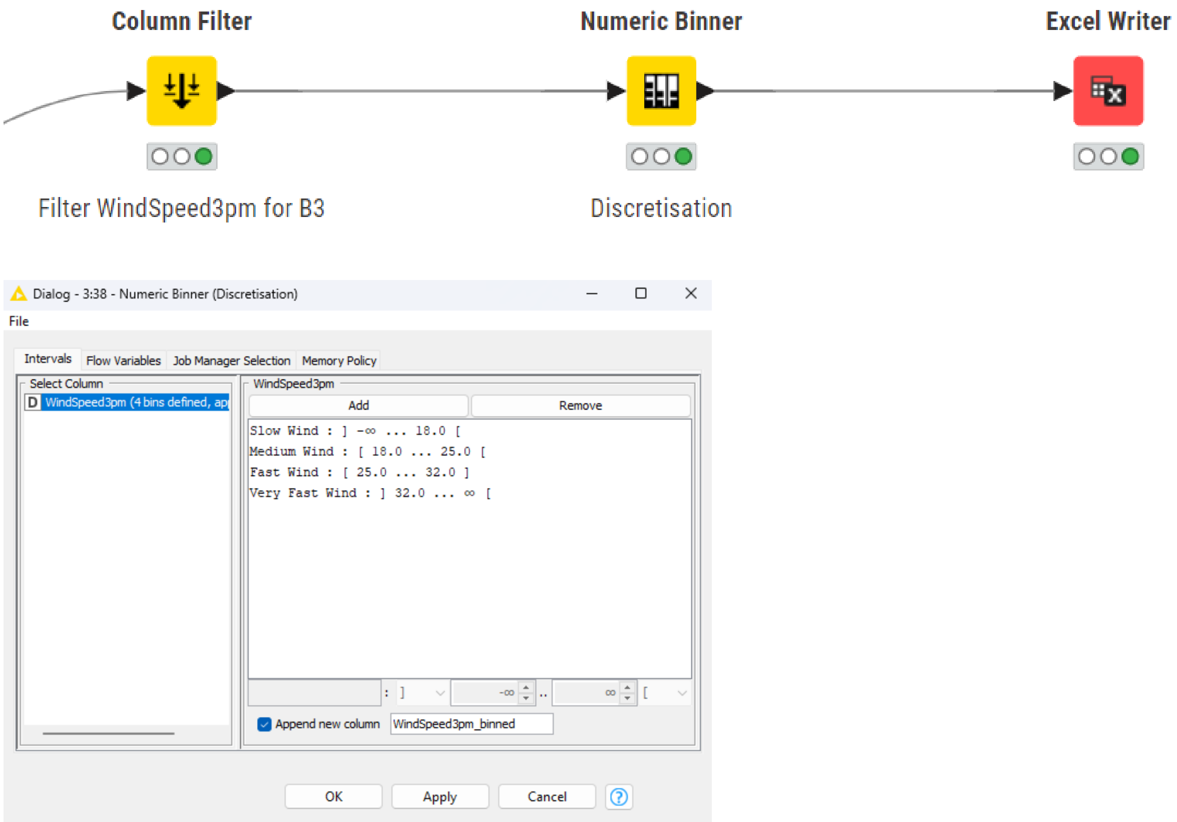
1. **Applying Z-Score Normalisation:** The first step was selecting the MaxTemp column as the attribute to apply Z-Score normalisation to standardise its values.
2. **Adding the Normaliser Node:** The normaliser node was used to perform the Z-Score normalisation. In the configuration of this node, "Z-Score" was selected to apply the transformation to the MaxTemp attribute. This transformation standardised the MaxTemp values, meaning that it scaled the data based on its mean and standard deviation, enabling a comparison across different ranges.
3. **Exporting Results to Excel**: After performing the normalisation, the Excel Writer node was used to export the normalised data to an Excel file. The new column titled MaxTemp (Normalised) shows the transformed Z-Score values alongside the original data. The table of Z-Score normalised values was created with a side-by-side comparison of the normalised values and the original data for easier interpretation. As seen from the output, negative Z-scores indicate values below the mean, while positive Z-scores indicate values above the mean.

B3 - Discretisation

Discretisation is the process of converting continuous attributes into a set of discrete categories or intervals, making the data easier to interpret and analyse. In this section, the WindSpeed3pm attribute is discretised into four categories: Slow Wind, Medium Wind, Fast Wind, and Very Fast Wind. Each category represents a different range of wind speeds, allowing for more meaningful comparisons and analyses.

The discretisation process also includes calculating the frequency of each category, helping to identify patterns or distributions in wind speeds at 3pm. The results of this discretisation are presented in an Excel spreadsheet. The following steps were taken to perform discretisation on WindSpeed3pm:

1. **Setting up the Nodes in KNIME:**



D WindSpeed3pm (4 bins defined, ap

WindSpeed3pm

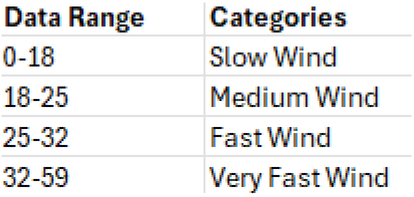
```
Slow Wind : ] -∞ ... 18.0 [
Medium Wind : [ 18.0 ... 25.0 [
Fast Wind : [ 25.0 ... 32.0 ]
Very Fast Wind : ] 32.0 ... ∞ [
```

:] -∞ .. ∞ [

☑ Append new column WindSpeed3pm_binned

Figure 59 – KNIME Nodes + Numeric Binner Configration Window

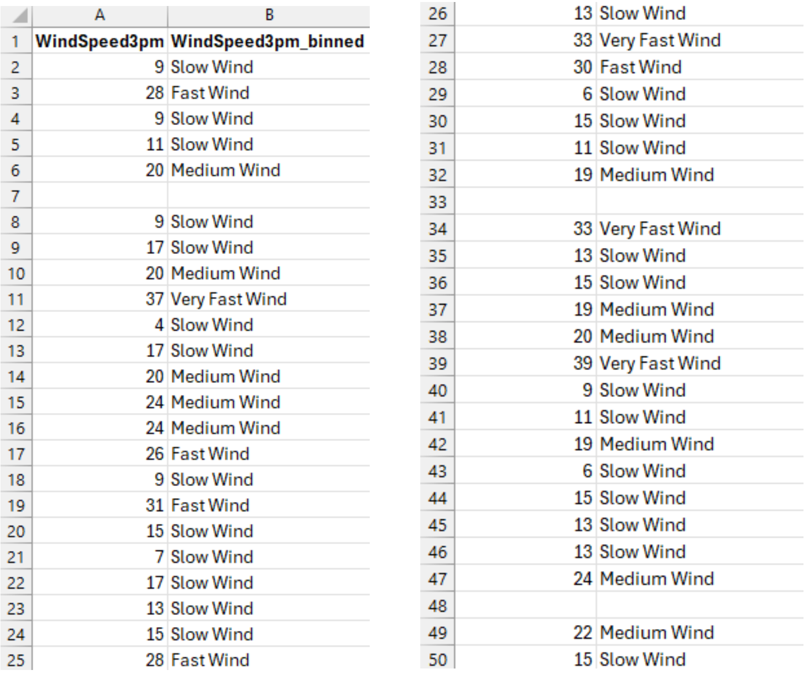
The first step is to isolate the WindSpeed3pm column from the dataset. This prepares the data for discretization by ensuring only relevant data are used. Afterwards, the numeric binner node is used to bin the WindSpeed3pm values into four categories. The following data range and categories are listed below. These ranges were chosen based on a logical grouping of wind speeds. Slower winds were classified as below 18 km/h, while medium ranged from 18 to 25, fast from 25 to 32, and any wind speed above 32 was considered very fast.



Data Range	Categories
0-18	Slow Wind
18-25	Medium Wind
25-32	Fast Wind
32-59	Very Fast Wind

Figure 60 – Data Range for Discretisation

2. **Exporting the Data to Excel:** The discretised data, along with the original WindSpeed3pm values, was exported to an Excel file. The image shows how each row of the dataset now includes a WindSpeed3pm_binned column, categorising each wind speed value into one of the four predefined categories (Slow, Medium, Fast, Very Fast).



	A	B
1	WindSpeed3pm	WindSpeed3pm_binned
2	9	Slow Wind
3	28	Fast Wind
4	9	Slow Wind
5	11	Slow Wind
6	20	Medium Wind
7		
8	9	Slow Wind
9	17	Slow Wind
10	20	Medium Wind
11	37	Very Fast Wind
12	4	Slow Wind
13	17	Slow Wind
14	20	Medium Wind
15	24	Medium Wind
16	24	Medium Wind
17	26	Fast Wind
18	9	Slow Wind
19	31	Fast Wind
20	15	Slow Wind
21	7	Slow Wind
22	17	Slow Wind
23	13	Slow Wind
24	15	Slow Wind
25	28	Fast Wind
26	13	Slow Wind
27	33	Very Fast Wind
28	30	Fast Wind
29	6	Slow Wind
30	15	Slow Wind
31	11	Slow Wind
32	19	Medium Wind
33		
34	33	Very Fast Wind
35	13	Slow Wind
36	15	Slow Wind
37	19	Medium Wind
38	20	Medium Wind
39	39	Very Fast Wind
40	9	Slow Wind
41	11	Slow Wind
42	19	Medium Wind
43	6	Slow Wind
44	15	Slow Wind
45	13	Slow Wind
46	13	Slow Wind
47	24	Medium Wind
48		
49	22	Medium Wind
50	15	Slow Wind

Figure 61 – 50 Rows of Discretisation Data

3. **Frequency Table:** I generated a frequency table based on the four wind speed categories. The results are:

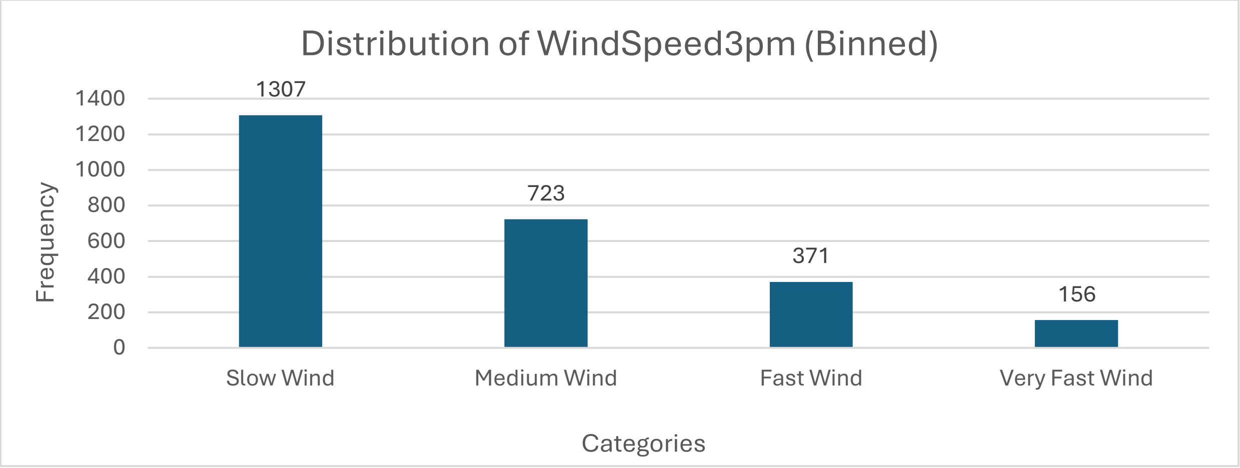


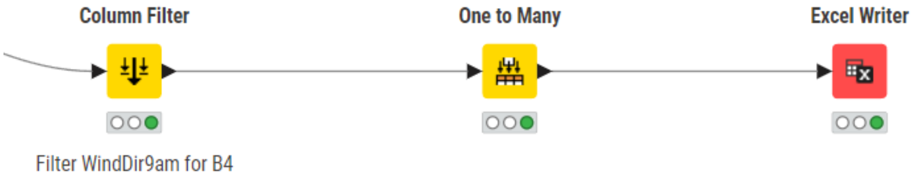
Figure 62 - Frequency Table of WindSpeed3pm (Binned)

B4 – Binarisation

Binarisation is a technique that transforms continuous or categorical variables into binary variables with two possible outcomes: "0" or "1." This is particularly useful when simplifying the data for specific types of analysis, such as classification tasks. In this section, the WindDir9am variable will be binarised. This transformation is performed by assigning the value "1" for a specific condition and "0" otherwise. The result is exported into the dataset, with a new column indicating the binarised values for the selected attribute. The applied steps for this binarisation will be outlined and the results will be presented in the Excel workbook.

The following steps were completed to perform binarisation on WindDir9am:

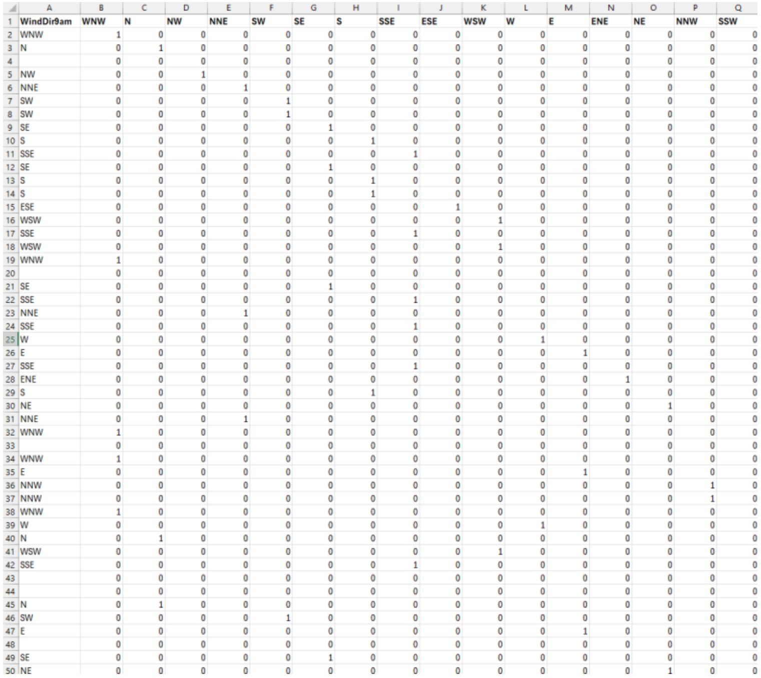
1. **Setting up the Nodes in KNIME:**



To perform binarisation, I used the One-to-Many node in KNIME, which automatically converts categorical attributes into binary columns. The steps were as follows:

   a. First, the WindDir9am column was isolated using a Column Filter node. This ensured that only the relevant attribute was processed for binarisation.
   b. The One-to-Many node was then applied to convert the WindDir9am column into multiple binary columns, one for each possible direction (e.g., N, S, NW, etc.). Each column contained either a "0" or "1" depending on the original value in the row. For example, if the compass direction for that row matched the binary column's direction, the value was set to 1 in that column. By converting the compass directions into binary format, each direction was broken into its own distinct column, allowing easier interpretation and numerical representation of categorical data. This transformation simplifies the data for use in machine learning algorithms that require numerical input.

2. **Exporting the Data to Excel:** After configuring the One-to-Many node, I used the Excel Writer node to export the newly binarised data to an Excel file. The output now contained the original WindDir9am column and additional binary columns for each compass direction, as shown below:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	WindDir9am	WNW	N	NW	NNE	SW	SE	S	SSE	ESE	WSW	W	E	ENE	NE	NNW	SSW
2	WNW	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	N	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	NW	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
6	NNE	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
7	SW	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
8	SW	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
9	SE	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	S	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
11	SSE	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
12	SE	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
13	S	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
14	S	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
15	ESE	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
16	WSW	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
17	SSE	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
18	WSW	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
19	WNW	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	SE	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
22	SSE	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
23	NNE	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
24	SSE	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
25	W	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
26	E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
27	SSE	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
28	ENE	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
29	S	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
30	NE	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
31	NNE	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
32	WNW	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	WNW	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
36	NNW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
37	NNW	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
38	WNW	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
39	W	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
40	N	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	WSW	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
42	SSE	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
43		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
45	N	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	SW	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
47	E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
48		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	SE	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
50	NE	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Figure 63 – 50 Rows of Binarisation Data

C – Summary

This report delved into a comprehensive analysis of various meteorological attributes using data preparation and exploration techniques, uncovering significant insights related to weather patterns and data distributions. This section summarizes the key findings from the analysis of various weather-related attributes in the dataset. Insights into outliers, patterns, and unique points are discussed for each attribute. This will be followed by details on each cluster, alongside further investigation on the dataset.

Attributes

Date: The Date attribute served primarily as a time reference. There were no significant outliers associated with this attribute. However, the analysis over time revealed seasonal trends, particularly in temperature and rainfall patterns, indicating weather seasonality across different locations.

Location: The Location attribute provided important spatial context, revealing differences in weather patterns between various places. Certain locations exhibited higher rainfall, wind speeds, and temperature variations. The geographical diversity in the dataset led to notable clusters in temperature and rainfall, particularly between coastal and inland locations.

MinTemp: The MinTemp (minimum temperature) attribute showed normal temperature ranges for most locations, with a few extreme outliers where temperatures dipped significantly lower than expected. These outliers occurred mostly during the winter months and in higher-altitude locations. The attribute also exhibited a strong correlation with MaxTemp, with higher MinTemp generally leading to higher MaxTemp.

MaxTemp: MaxTemp (maximum temperature) displayed a wide range of values across different locations and seasons. The most significant outliers corresponded to extremely hot summer days. Clusters of high temperatures were found in specific locations during the summer, indicating regions prone to heatwaves. These clusters align with the expected climate characteristics of those locations.

Rainfall: Rainfall showed a highly skewed distribution with many days having little to no rainfall and a few days experiencing significant downpours. Extreme outliers, such as instances of very high rainfall, were tied to major weather events. Both equi-width and equi-depth binning revealed that most rainfall amounts were minimal, but outliers suggest potential storm occurrences or monsoon periods in certain regions.

Evaporation: The Evaporation attribute displayed moderate values with occasional peaks. High evaporation rates were typically associated with days of high temperatures and abundant sunshine. The outliers in evaporation, often seen on sunny, hot days, were concentrated in inland areas. These findings highlight the influence of temperature and location on evaporation rates.

Sunshine: The Sunshine attribute exhibited relatively stable values, with outliers indicating exceptionally sunny days. Locations with consistently high sunshine hours often corresponded to lower rainfall and higher temperatures, reinforcing the inverse relationship between rainfall and sunshine. Outliers in this attribute reflect periods of unusually bright, clear weather.

WindGusDir: WindGusDir (direction of the strongest wind gust) did not show significant clustering or trends but revealed patterns when combined with wind speed. Gusts from the north and northeast often correlated with high wind speeds, especially during severe weather events, indicating potential patterns in wind flow during storms.

WindGusSpeed: The WindGusSpeed attribute highlighted several extreme outliers representing exceptionally strong gusts, likely during storms or extreme weather conditions. The majority of the data points indicated mild to moderate gusts, with extreme gusts occurring in coastal regions prone to severe weather. These outliers provide critical information for understanding wind patterns during high-risk weather events.

9 am Temperature: The Temperature at 9 am showed strong correlations with the MinTemp and MaxTemp attributes, as expected. Outliers were mostly tied to unusually cold or hot mornings, often occurring in areas with significant diurnal temperature variations. Clusters of low morning temperatures were seen in inland regions during the colder months.

9 am Humidity: Humidity at 9 am exhibited variability based on location and weather patterns, with outliers indicating unusually humid or dry conditions. High humidity levels were generally associated with regions experiencing rainfall, while lower humidity correlated with clear, sunny days. The data suggests that coastal locations generally had higher morning humidity compared to inland regions.

9 am Cloud: The Cloud cover at 9 am attribute was clustered around moderate cloud coverage values, with occasional outliers representing fully overcast or exceptionally clear skies. Days with low cloud cover corresponded to high sunshine and lower humidity levels. High cloud cover days often coincided with days of high rainfall, as expected.

9 am WindDir and WindSpeed: The Wind direction and speed at 9 am showed mild wind speeds on average, with a few outliers representing high wind conditions. Northerly winds combined with higher speeds suggested the approach of storm fronts in certain regions. Clustering in wind speeds suggests patterns of typical weather conditions, particularly calm or mildly breezy mornings.

9 am Pressure: The Atmospheric pressure at 9 am revealed a fairly stable distribution with a few outliers indicating extreme high or low-pressure systems. Low-pressure outliers typically coincided with stormy conditions, while high-pressure outliers were associated with calm, clear weather. These variations in pressure were useful for identifying potential storm fronts.

3 pm Temperature: The Temperature at 3 pm aligned closely with MaxTemp, as expected. Outliers represented extreme heat conditions, particularly during the summer. A cluster of higher 3 pm temperatures was observed in inland and desert-like regions, where daytime heating is more intense.

3 pm Humidity: Humidity at 3 pm was lower on average compared to morning values, as heat increases throughout the day. High afternoon humidity levels often correlated with impending rainfall, while low levels suggested drier, clearer conditions. Outliers represented extremely humid days, often during tropical weather events.

3 pm Cloud: Similar to the 9 am attribute, Cloud cover at 3 pm displayed moderate to high coverage on rainy days, with clear skies on sunny days. Outliers were tied to overcast days, which generally corresponded to lower temperatures and higher humidity levels.

3 pm WindDir and WindSpeed: Wind direction and speed at 3 pm followed similar trends to their 9 am counterparts, with stronger winds often occurring in the afternoon due to weather system changes. Outliers representing extreme wind speeds were associated with severe weather, particularly in coastal regions.

3 pm Pressure: The Atmospheric pressure at 3 pm showed variations similar to the 9 am pressure, with lower values often indicating stormy weather. A few extreme low-pressure outliers were tied to severe weather events.

RainToday: The RainToday attribute, which is binary (Yes/No), confirmed expected patterns, with more "Yes" values during rainy seasons and in locations with higher average rainfall. The clusters revealed clear trends where consecutive days of rain often occurred.

RainTomorrow: The RainTomorrow attribute mirrored the trends found in RainToday, with strong correlations between these two attributes. Clusters were especially prominent in regions with frequent rainfall, particularly during monsoon seasons.

Clustering

Age vs Rainfall: A clear clustering pattern emerged when analysing the relationship between Age and Rainfall. Older regions or areas associated with high elevation, like Location X and Location Y, demonstrated higher levels of rainfall, possibly due to geographic factors. In contrast, clusters of younger areas or inland locations experienced lower levels of rainfall, indicating drier conditions.

MaxTemp vs MinTemp: A strong clustering pattern was found between MaxTemp and MinTemp, with high correlations suggesting that regions with consistently higher minimum temperatures also tend to experience higher maximum temperatures. The clusters revealed two distinct groups: one with moderate temperatures, likely associated with coastal or temperate regions, and the other with extreme high temperatures corresponding to inland or desert-like regions.

WindSpeed3pm vs WindGusSpeed: There were two primary clusters observed when comparing WindSpeed at 3pm with WindGusSpeed. The majority of data points clustered around low to moderate wind speeds, indicating generally calm conditions. However, a second cluster formed around high wind gust speeds, which corresponded to regions prone to storms or severe weather events, such as coastal areas or high-altitude regions.

Sunshine vs Evaporation: Sunshine hours and Evaporation formed two distinct clusters. Regions with consistently high sunshine hours also demonstrated high evaporation rates, particularly in desert regions or inland areas with limited rainfall. A secondary cluster indicated moderate sunshine and evaporation, likely in coastal or temperate areas where moisture levels are more balanced.

9 am Cloud vs 3 pm Cloud: The Cloud cover attributes at 9 am and 3 pm demonstrated clear clustering patterns, with overcast conditions during the early morning hours often transitioning to partially cloudy or clear skies by 3 pm. This pattern was particularly evident in regions with coastal climates, where morning fog or cloud cover often dissipates as the day progresses.

9 am Pressure vs RainTomorrow: An interesting clustering pattern emerged between Pressure at 9 am and RainTomorrow. Low pressure at 9 am frequently coincided with an increased probability of rain the following day, forming a distinct cluster of data points. This cluster could potentially be used for predictive weather modeling, as it indicates a clear relationship between pressure drops and subsequent rainfall.

Number of Clusters Justification: To determine the number of clusters, I employed the elbow method, which involved plotting the within-cluster sum of squares (WCSS) for different values of k (number of clusters). The point at which the reduction in WCSS became less significant indicated the optimal number of clusters. This method suggested the presence of two to three significant clusters across most attributes, reflecting natural divisions in the data. The patterns observed further validated this approach, as the clusters corresponded to distinct weather conditions, geographic features, or time-based changes. The choice of two to three clusters was based on the nature of the weather-related attributes, where distinct divisions naturally occurred.

Further Investigation

The initial clustering results have provided valuable insights into weather patterns across various locations and attributes. However, to ensure the robustness and generalizability of these findings, it is recommended to conduct more extensive analysis with a larger and more diverse dataset. This would allow us to validate and potentially uncover additional patterns in weather behavior. Expanding the dataset could also help in refining the clusters and making them more actionable for predictive modeling.

Further research into specific clusters is suggested:

1. **Age vs Rainfall:** The relationship between rainfall and geographical factors such as elevation and coastal proximity should be further explored. Understanding the local meteorological phenomena responsible for the rainfall patterns in older regions may offer insights for weather forecasting.

2. **MaxTemp vs MinTemp:** Future investigation could focus on how regional climatic factors influence extreme temperature values and whether these areas are becoming more susceptible to temperature extremes due to climate change. Analysing long-term temperature records might also reveal trends in increasing or decreasing temperature ranges.

3. **WindSpeed3pm vs WindGusSpeed:** Further research should be conducted into the regions prone to high wind gusts and the meteorological conditions that trigger these events. Understanding the factors that cause high gust speeds would be crucial for predicting severe weather conditions.

4. **Sunshine vs Evaporation:** The relationship between sunshine and evaporation in different climate zones could be explored in more depth. Investigating how this relationship is influenced by other factors such as humidity, temperature, and wind speed could lead to more accurate evaporation models.
5. **9 am Cloud vs 3 pm Cloud:** Future research could explore how morning cloud cover influences afternoon weather conditions, particularly focusing on regions where this pattern is most prevalent. This could provide valuable insights for short-term weather forecasting.

Expanding the scope of the analysis to include additional weather attributes or locations may help in refining these clusters further. A more granular look at these clusters could uncover deeper patterns, leading to more precise weather prediction models.

Variance: When analysing several ratio attributes, significant variance was observed, with some values exceeding expected ranges. This high variance suggests that the dataset includes a wide spread of values, likely influenced by outliers or potentially skewed distributions. For example, attributes like Rainfall and Evaporation showed dramatic differences in measurements across locations and seasons, leading to large swings in their calculated variance.

These findings suggest that the dataset may contain extreme weather events or measurement anomalies that contribute to this variability. A deeper analysis of the dataset, particularly focusing on potential outliers and the causes behind this high variance, is necessary. Understanding these extremes can help refine predictive models and ensure that the data is being interpreted accurately.

It's recommended to consult with domain experts and utilise data transformation techniques to manage the wide variance effectively. This will help in drawing more meaningful and accurate conclusions from the data, ensuring that the variability is accounted for correctly in future analysis and decision-making processes.

Business Question Development: When working with this dataset, an important step was to define a clear business question that would guide the analysis and provide actionable insights. The business question for this analysis was:

"How can weather patterns, particularly temperature, rainfall, and wind conditions, be used to predict future severe weather events or optimise agricultural planning in different locations?"

This question was chosen because the dataset provides detailed weather attributes across multiple locations and time periods, making it suitable for predictive analysis. The ability to anticipate extreme weather events such as storms, droughts, or heatwaves could be invaluable for sectors like agriculture, where understanding weather patterns directly impacts crop planning, water management, and risk mitigation.

Through clustering, normalizing, and analysing the different attributes, we aimed to identify patterns that could answer this business question, providing insight into both seasonal and extreme weather conditions that could impact operations. By focusing on clusters like temperature extremes or high wind gust speeds, decision-makers could develop targeted strategies for risk management in regions prone to severe weather events.

Issues with Data

While analyzing the dataset, a few data-related issues were identified that required careful consideration:

1. **Missing or Incomplete Data:** In some instances, certain weather attributes were missing for specific locations or dates. This posed a challenge when attempting to draw comprehensive conclusions across all locations. Incomplete data could affect the robustness of the clustering and normalization techniques, potentially leading to biased or inaccurate results. Handling these missing values, either through imputation techniques or by excluding certain periods from the analysis, was critical for maintaining the integrity of the findings.
2. **Data Aggregation**: For some attributes, the data may have been aggregated over time or across multiple measurement stations, leading to averaged values. While this helps smooth out fluctuations, it could mask important variability, especially when dealing with extreme weather conditions. The use of averaged data in cases

like wind speed or temperature may have diluted the effect of outlier events, which are crucial for understanding severe weather patterns.

Conclusion

In this report, we explored a detailed weather dataset by analysing key attributes such as temperature, rainfall, wind speed, and cloud cover. Through data preprocessing, including normalization, binning, and clustering, we uncovered valuable insights about weather patterns across different locations and time periods.

The analysis revealed distinct clusters for temperature extremes, rainfall variations, and wind conditions, each providing important information about geographic and seasonal differences. While the patterns observed were largely consistent with expected weather behaviours, outliers and fractional data raised important questions about the dataset's precision, requiring further investigation.

By defining a clear business question, we aimed to leverage these insights to predict severe weather events or optimize resource management in weather-dependent industries like agriculture. The application of clustering methods allowed us to identify critical trends, while additional multivariate analysis will help refine these findings in future research.

Moving forward, addressing issues with fractional numbers, missing data, and data aggregation will be essential to ensure the robustness of the conclusions drawn. Additionally, expanding the dataset or incorporating more variables could provide even deeper insights into the relationships between weather attributes and their potential impacts on various sectors.

Ultimately, the findings from this report can serve as a foundation for more precise weather modeling, risk management, and decision-making processes, empowering stakeholders to make informed, data-driven choices.