31250 INTRODUCTION TO DATA ANALYTICS

Assessment task 2: Data exploration and preparation

Chaeeun Lee
14502431

Table of Contents

1A. Initial data exploration

1. Attribute types
2. The summarizing properties for the attributes
3. Exploration

1B. Data preprocessing

1. Binning techniques
    a. Equi-width binning
    b. Equi-depth binning
2. Normalisation
    a. Min-Max
    b. Z-score
3. Descretisation
4. Binarisation

1C. Summary

1A. Initial data exploration

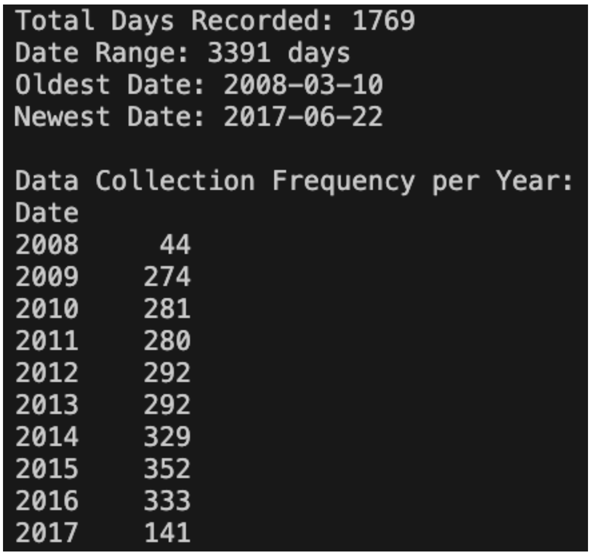
1. Attribute Types

Attribute Name	Attribute Description	Attribute Type	Justification
Date	Date on which weather data was recorded	Nominal	Dates are used as labels for identifying specific entries in a time series. There's no numerical significance or order to the dates themselves.
Location	Geographic location of the weather station	Nominal	Locations are treated as distinct categories for comparing regional weather data. They serve as identifiers without any hierarchical order.
Min/Max Temp	Minimum/Maximum temperature recorded at a specific location on a given day	Ratio	MinTemp/MaxTemp values have a true zero, and differences between temperatures can be meaningfully calculated, essential for climatic analyses.
Rainfall	Total precipitation recorded in millimeters	Ratio	Rainfall is quantitatively measured from zero (no rain), making it possible to perform arithmetic operations like averaging or summing.
Evaporation	Amount of water evaporated on that day, measured in millimeters	Ratio	Evaporation measures from zero, important for water resource studies and environmental assessments.
Sunshine	Total hours of sunshine recorded	Interval	Sunshine is measured in hours per day, suitable for interval measurement since the scale starts above zero and differences between values are meaningful.
WindGustDir	Direction from which the strongest wind gusts were recorded	Nominal	Wind direction is classified nominally as it represents non-quantitative categories of wind origins without an inherent order.
WindGustSpeed	Speed of the strongest wind gusts recorded in	Ratio	WindGusSpeed is quantified from zero

	km/h		and up, vital for assessing storm strengths and for use in safety protocols.
WindDir (9am/3pm)	Wind direction at 9 am and 3 pm	Nominal	Treated nominally because it categorizes wind directions without quantitative measurement, aiding in daily wind pattern analysis.
WindSpeed (9am/3pm)	Wind speed at 9 am and 3 pm in km/h	Ratio	Measurable from zero, allowing for precise calculations and comparisons, essential for daily weather predictions and operations.
Humidity (9am/3pm)	Humidity percentage at 9 am and 3 pm	Interval	Measures relative moisture in the air, providing data crucial for understanding daily weather dynamics.
Pressure (9am/3pm)	Atmospheric pressure at 9 am and 3 pm measured in hectopascals	Ratio	Pressure is a continuous scale from zero, crucial for understanding atmospheric conditions and predicting weather changes throughout the day.
Cloud Cover (9am/3pm)	Cloud cover at 9 am and 3 pm, recorded as oktas	Ordinal	Cloud cover is ranked by oktas, providing an ordered classification that is crucial for visual weather assessment throughout the day.
Temperature (9am/3pm)	Temperature recorded at 9 am and 3 pm	Ratio	Allows for direct measurement and statistical analysis, crucial for daily climate monitoring and reporting.
Rain Today/Tomorrow	Indicates whether it rained today (Yes or No) and predicts if it will rain the next day (Yes or No).	Nominal	These nominal attributes categorize daily rainfall presence and forecasts, essential for weather analysis and planning in various sectors

2. The summarizing properties for the attributes

1. Date



```
Total Days Recorded: 1769
Date Range: 3391 days
Oldest Date: 2008-03-10
Newest Date: 2017-06-22

Data Collection Frequency per Year:
Date
2008     44
2009    274
2010    281
2011    280
2012    292
2013    292
2014    329
2015    352
2016    333
2017    141
```

Figure 1. Result of python for date data

The dataset covers weather data from March 10, 2008, to June 22, 2017, with a total of 1,769 days recorded over 3,391 days. Data collection peaked in 2016 with 333 days recorded, illustrating variability in data recording frequency which may affect trend analysis and forecasting accuracy.

2. Location

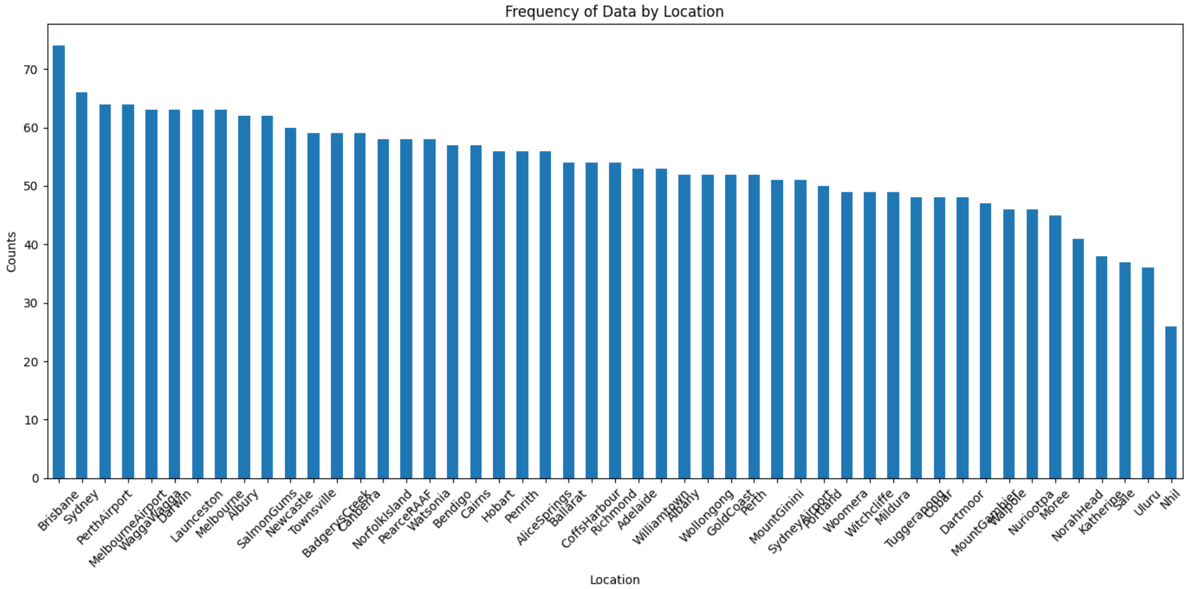


Figure 2. Frequency of Data by Location

The bar chart detailing the frequency of data by location shows significant variability in the volume of data collected across different weather stations. This diversity in data availability could impact the accuracy and comprehensiveness of weather predictions and climate studies, particularly in areas with fewer observations. Enhancing data collection efforts in underrepresented locations would improve regional weather forecasts and support a more balanced understanding of climatic differences across regions.

3. MinTemp/MaxTemp

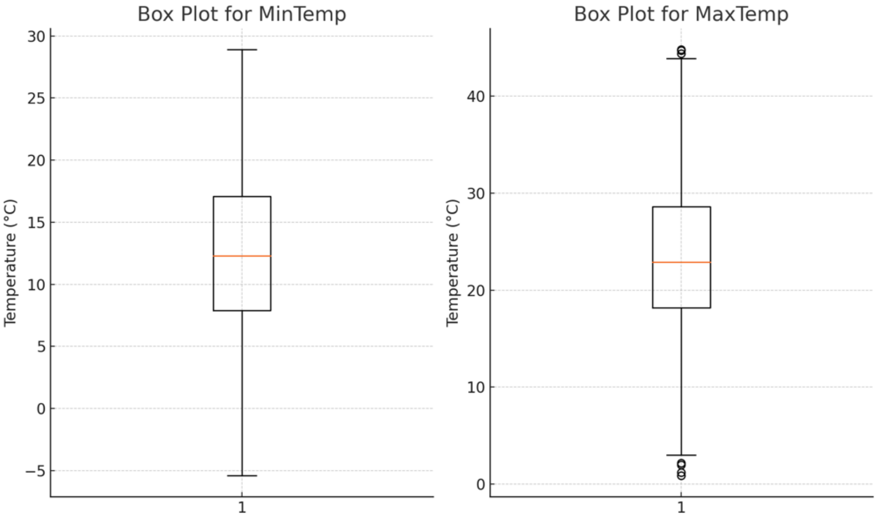


Figure 3. Box plot for MinTemp/MaTemp

Figure 3 illustrates the variability in the minimum and maximum temperatures through box plots. The box plots highlight the range of temperatures experienced in the region, with MinTemp sometimes reaching extreme lows, which are critical during the cold season for energy management and infrastructure preparation. Similarly, the MaxTemp shows significant spikes to higher values, emphasizing the need for effective heatwave preparedness and cooling systems, particularly for vulnerable populations and in agricultural planning. These temperature extremes are essential for understanding local climatic conditions and can guide decision-making in resource allocation and emergency response strategies.

4. Rainfall

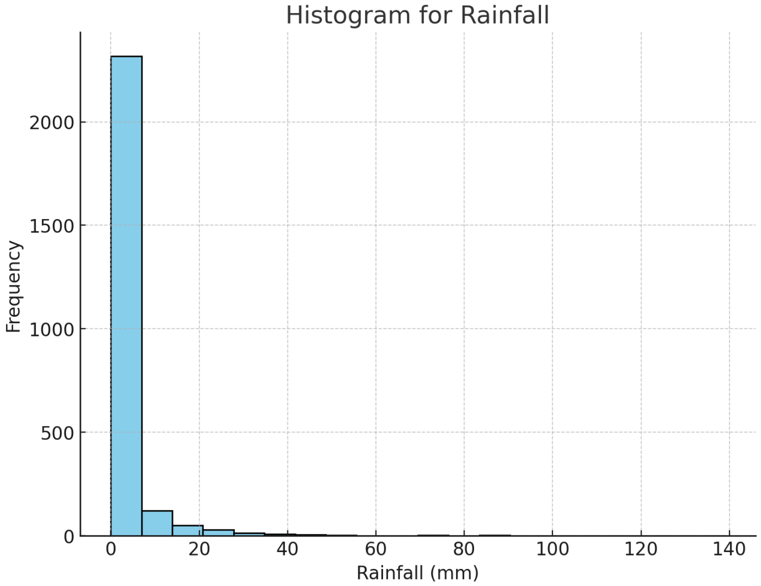
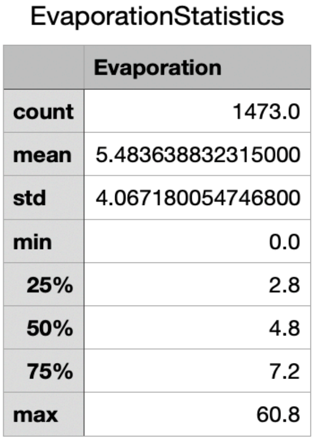


Figure4. Histogram for Rainfall

Figure 4 presents a histogram that visualizes the frequency and distribution of rainfall, indicating that most days record minimal to no precipitation. This pattern suggests a predominantly dry climate or a region characterized by distinct wet and dry seasons. Understanding these rainfall patterns is crucial for agriculture, urban planning, and flood risk management. Regions with such variability in precipitation require careful water resource management and infrastructure planning to mitigate the impacts of potential droughts or intense rainfall events, ensuring sustainability and resilience in water-dependent sectors.

5. Evaporation



EvaporationStatistics

	Evaporation
count	1473.0
mean	5.483638832315000
std	4.067180054746800
min	0.0
25%	2.8
50%	4.8
75%	7.2
max	60.8

Figure5. Evaporation Statistics

Evaporation rates, as shown in the summary statistics, indicate significant daily variability, with some days experiencing very high evaporation rates.

This variation is key to understanding local water cycles and environmental moisture dynamics, influencing water conservation strategies and local climate conditions. High evaporation rates may suggest environmental changes or anthropogenic influences that could be affecting local weather patterns and ecological balances.

6. Sunshine

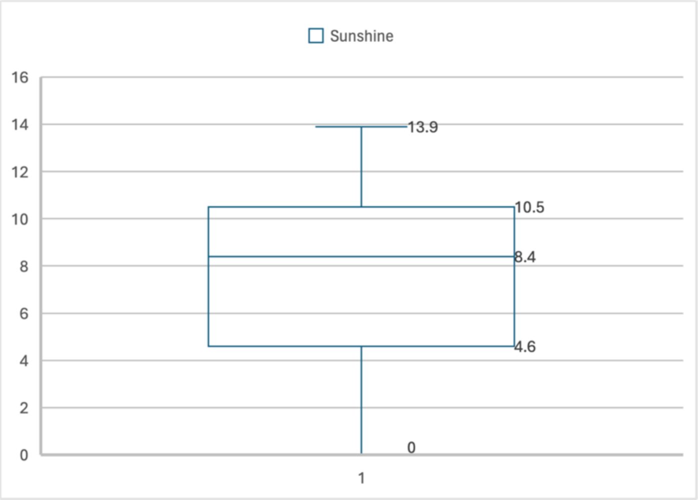


Figure 6. Box plot for Sunshine

The box plot for Sunshine shows average daily sunshine hovers around 8.4 hours, with variability from 4.6 to 13.9 hours. Days with zero sunshine likely indicate heavy cloud cover or storms. This insight is vital for agriculture and solar energy sectors, where sunlight directly affects crop yields and energy production. It also aids tourism industries in planning daylight-dependent activities.

7. WindGustDir

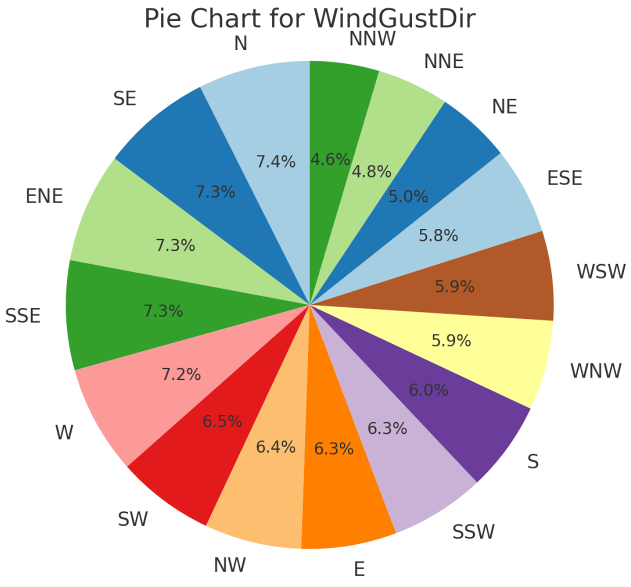
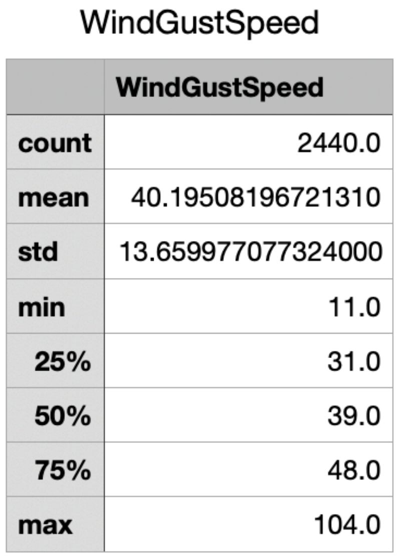


Figure 7. Pie chart for WindGustDir

The pie chart for WindGustDir shows a well-distributed set of directions from which wind gusts originate, with no single direction overwhelmingly dominant. This uniform distribution suggests that the area experiences a varied wind pattern, which is crucial for understanding local weather dynamics and can be particularly significant for sectors like aviation and marine navigation that rely on accurate wind direction data.

8. WindGustSpeed



WindGustSpeed

	WindGustSpeed
count	2440.0
mean	40.19508196721310
std	13.659977077324000
min	11.0
25%	31.0
50%	39.0
75%	48.0
max	104.0

Figure 8. Statisctic table of WindGustSpeed

The statistical summary for WindGustSpeed indicates typical wind speeds around 40 km/h, with extreme cases reaching up to 104km/h. This data is vital for designing infrastructure capable of withstanding high winds and for developing emergency preparedness plans to mitigate the risk of wind-related damage during severe weather events. The insights gained from analyzing WindGustSpeed can inform public safety protocols and guide the construction standards for buildings and other structures, enhancing community resilience against storms and high wind conditions.

9.WindDir(9am/3pm)

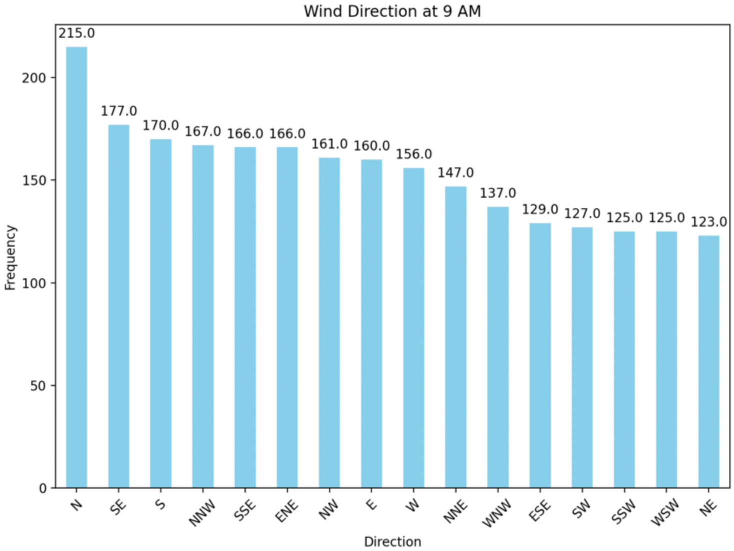


Figure9. Bar chart for Wind Direction at 9am

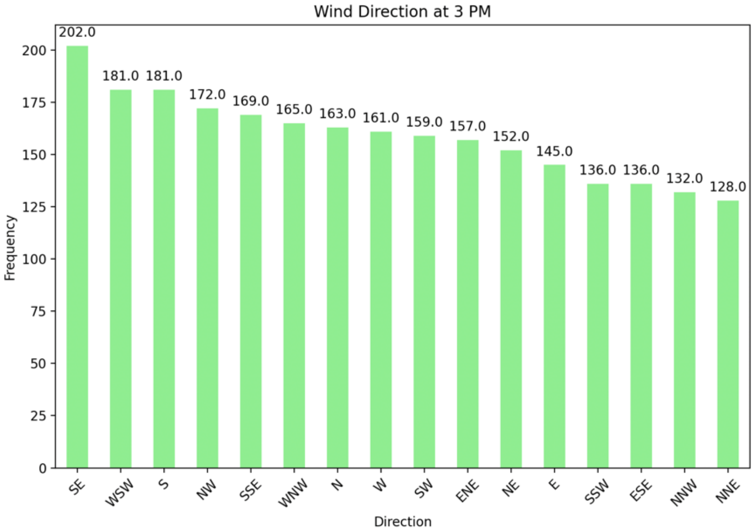


Figure10. Bar chart for Wind Direction at 3pm

The bar charts for Wind Direction at 9 AM and 3 PM demonstrate a diverse pattern in prevailing winds, with noticeable shifts throughout the day. At 9 AM, northerly directions (N, NE) are more prevalent, while by 3 PM, there's a slight shift towards southerly directions (S, SE). This shift in wind direction can be critical for activities that are sensitive to wind changes, such as aviation and marine operations. Understanding these patterns helps in planning and optimizing operations that depend on wind direction, such as pollutant dispersion modeling and renewable energy utilization from wind turbines.

10.WindSpeed(9am/3pm)

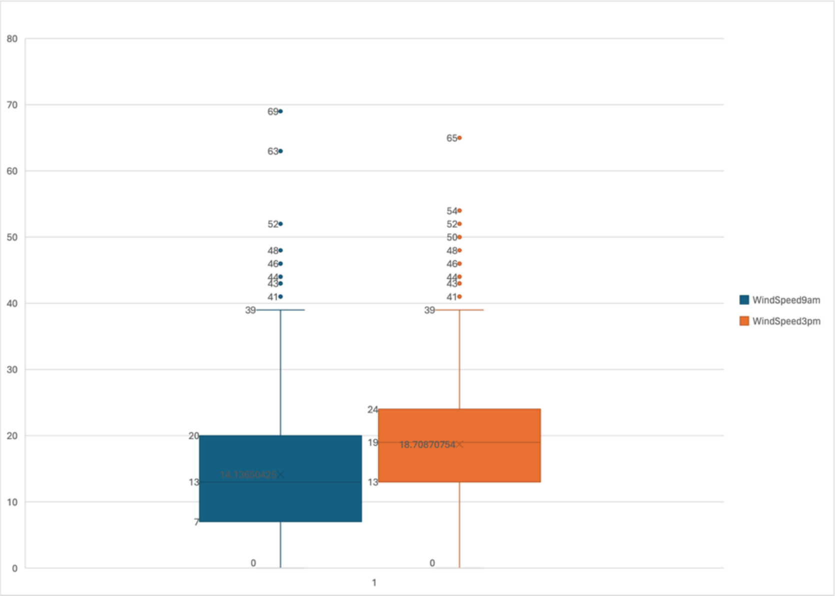


Figure11. Boxplot for Windspeed

The box plots for WindSpeed at 9 AM and 3 PM indicate a higher variability in wind speeds by the afternoon, suggesting increased wind activity later in the day. This pattern is essential for energy sectors, particularly renewable energy sources like wind turbines, which could capitalize on higher afternoon winds.

11.Humidity(9am/3pm)

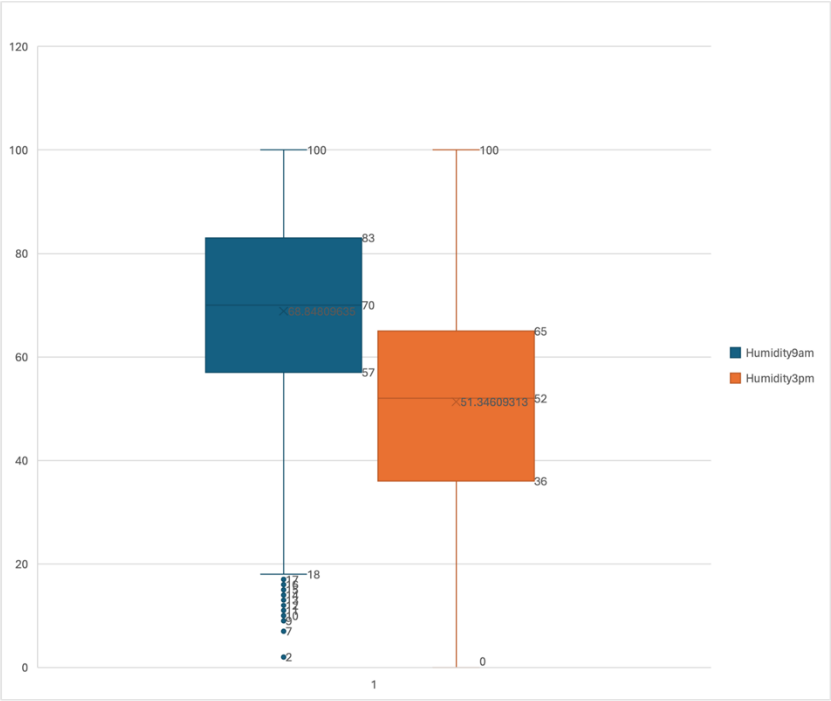


Figure12. Boxplot for Humidity

The comparison of humidity levels at 9 AM and 3 PM shows a decrease in humidity as the day progresses, typical of many environments where daytime heating reduces relative humidity. This trend is crucial for environmental planning and health as it affects air quality and comfort levels.

12.Pressure(9am/3pm)

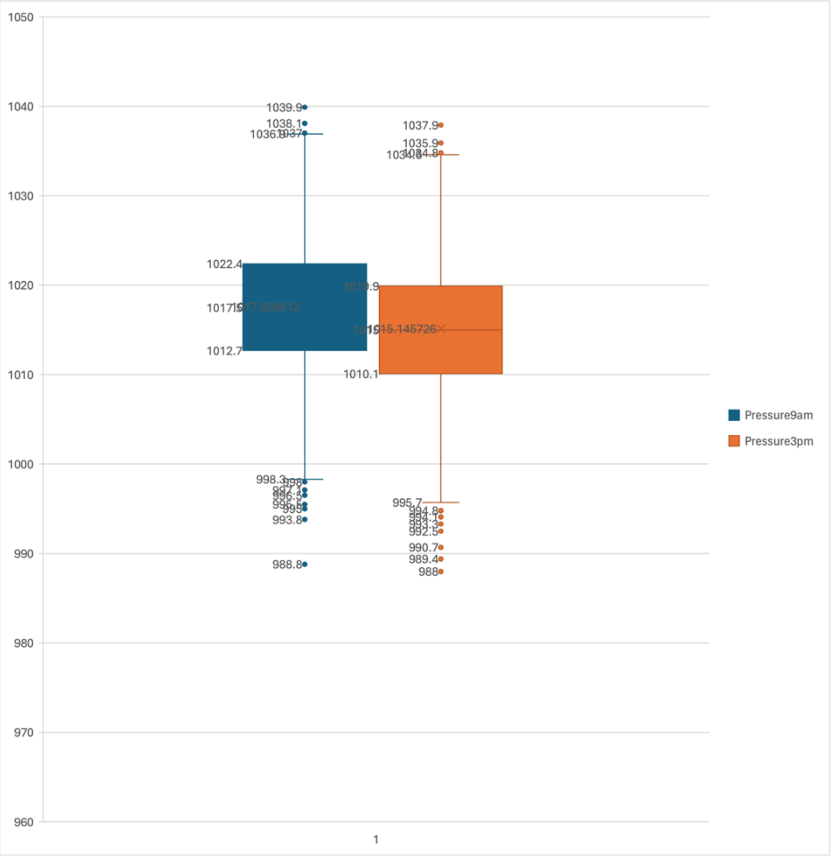


Figure13. Boxplot for Pressure

Atmospheric pressure readings from morning to afternoon show slight variations, indicating normal daily atmospheric fluctuations. These changes are critical for meteorological forecasting and can help predict weather patterns, including storm formation and movement.

13.Cloud Cover(9am/3pm)

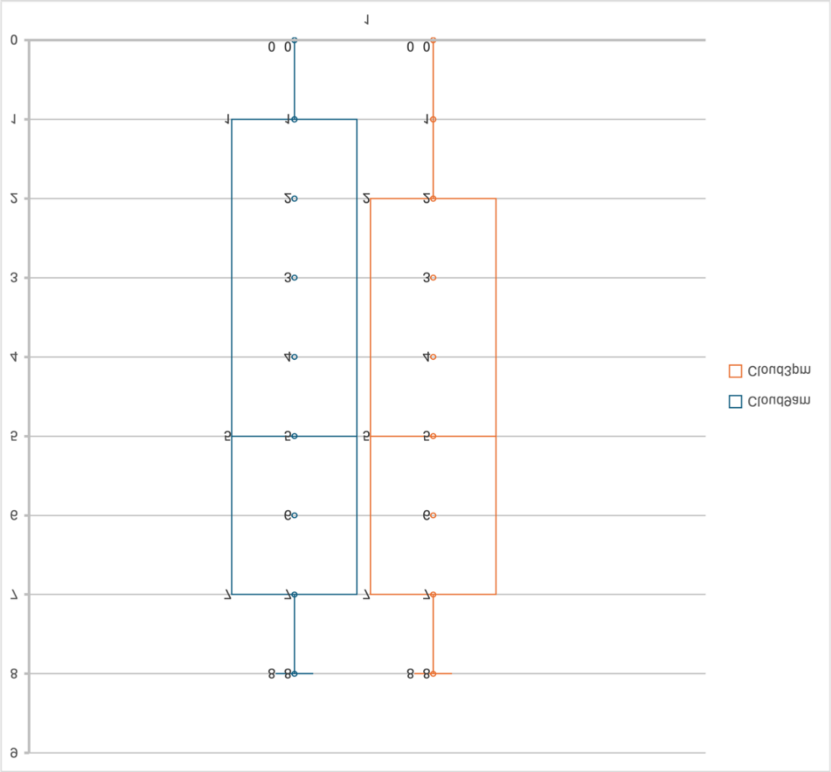
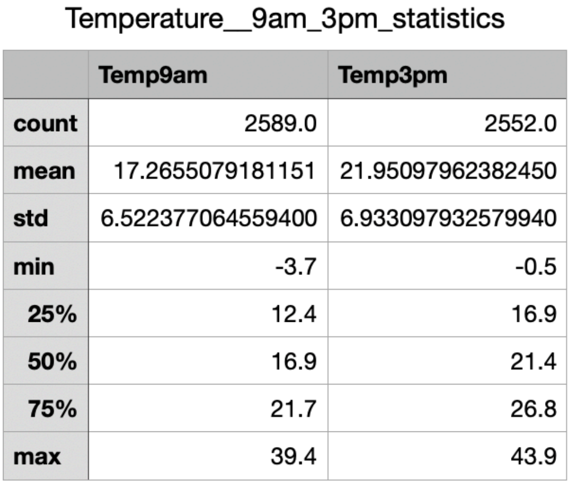


Figure14. Boxplot for Cloudcover

The cloud cover analysis shows more cloud coverage in the afternoon compared to the morning. This increase can influence various factors, including temperature drops and precipitation likelihood, playing a significant role in daily weather forecasting and agricultural planning.

14.Temperature(9am/3pm)



Temperature__9am_3pm_statistics

	Temp9am	Temp3pm
count	2589.0	2552.0
mean	17.2655079181151	21.95097962382450
std	6.522377064559400	6.933097932579940
min	-3.7	-0.5
25%	12.4	16.9
50%	16.9	21.4
75%	21.7	26.8
max	39.4	43.9

Figure15. Statistics Table for Temperature

The statistical summary for temperatures at 9 AM and 3 PM highlights the daily warming trend. Understanding these temperature dynamics is crucial for managing energy usage, scheduling activities that depend on temperature conditions, and health advisories related to heat.

15. Rain Today/Tomorrow

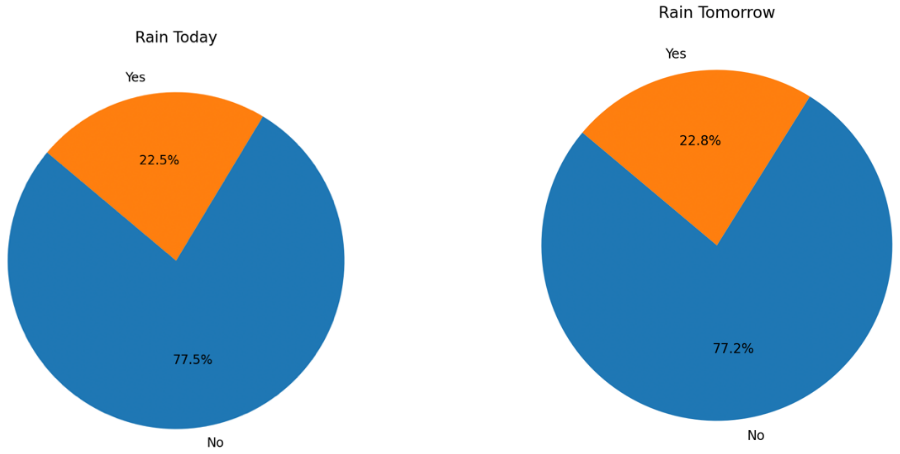
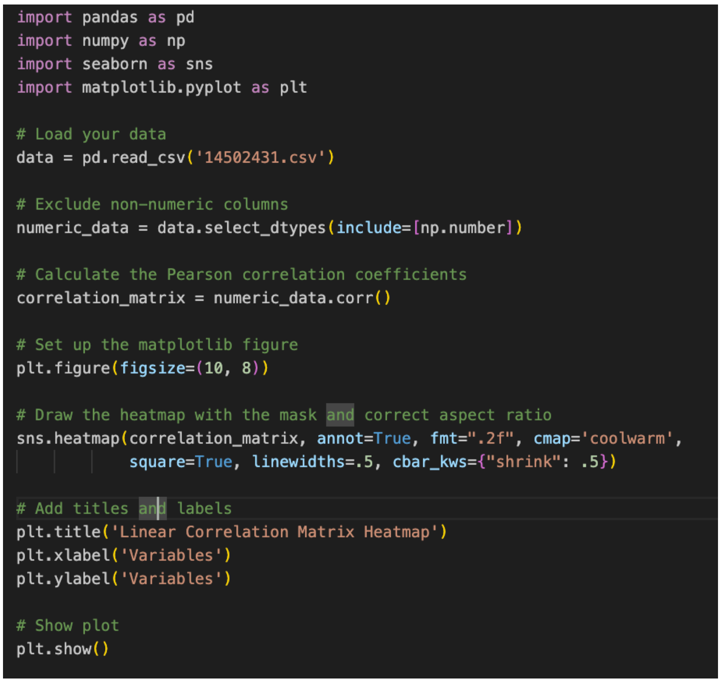


Figure 16,17. 2d pie for Rain Today/Tomorrow

The pie charts for "Rain Today" and "Rain Tomorrow" show that on most days, it does not rain, with about 77% of the days being dry, while approximately 23% of days experience rainfall. This data is crucial for understanding precipitation patterns and planning in various sectors dependent on weather conditions, such as agriculture, outdoor events, and construction. The consistency in the probability of rain from one day to the next also assists in predictive weather modeling, providing a stable basis for short-term forecasts and planning. This consistent pattern could reflect a climate with distinct dry and wet seasons, enabling precise resource allocation and risk management in water-sensitive operations.

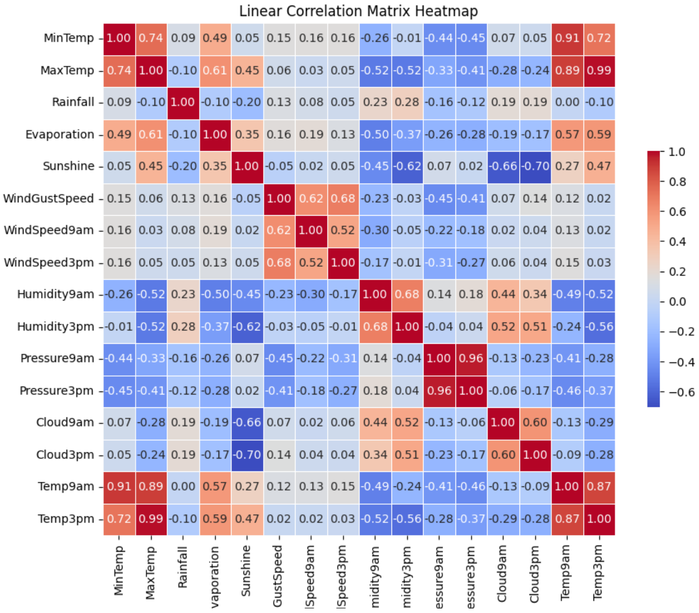
3. Exploration

1) Correlation



```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Load your data
data = pd.read_csv('14502431.csv')

# Exclude non-numeric columns
numeric_data = data.select_dtypes(include=[np.number])

# Calculate the Pearson correlation coefficients
correlation_matrix = numeric_data.corr()

# Set up the matplotlib figure
plt.figure(figsize=(10, 8))

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(correlation_matrix, annot=True, fmt=".2f", cmap='coolwarm',
            square=True, linewidths=.5, cbar_kws={"shrink": .5})

# Add titles and labels
plt.title('Linear Correlation Matrix Heatmap')
plt.xlabel('Variables')
plt.ylabel('Variables')

# Show plot
plt.show()
```

Figure18. python code for linear correlation Matrix heatmap



	MinTemp	MaxTemp	Rainfall	vaporation	Sunshine	GustSpeed	lSpeed9am	lSpeed3pm	midity9am	midity3pm	essure9am	essure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
MinTemp	1.00	0.74	0.09	0.49	0.05	0.15	0.16	0.16	-0.26	-0.01	-0.44	-0.45	0.07	0.05	0.91	0.72
MaxTemp	0.74	1.00	-0.10	0.61	0.45	0.06	0.03	0.05	-0.52	-0.52	-0.33	-0.41	-0.28	-0.24	0.89	0.99
Rainfall	0.09	-0.10	1.00	-0.10	-0.20	0.13	0.08	0.05	0.23	0.28	-0.16	-0.12	0.19	0.19	0.00	-0.10
Evaporation	0.49	0.61	-0.10	1.00	0.35	0.16	0.19	0.13	-0.50	-0.37	-0.26	-0.28	-0.19	-0.17	0.57	0.59
Sunshine	0.05	0.45	-0.20	0.35	1.00	-0.05	0.02	0.05	-0.45	-0.62	0.07	0.02	-0.66	-0.70	0.27	0.47
WindGustSpeed	0.15	0.06	0.13	0.16	-0.05	1.00	0.62	0.68	-0.23	-0.03	-0.45	-0.41	0.07	0.14	0.12	0.02
WindSpeed9am	0.16	0.03	0.08	0.19	0.02	0.62	1.00	0.52	-0.30	-0.05	-0.22	-0.18	0.02	0.04	0.13	0.02
WindSpeed3pm	0.16	0.05	0.05	0.13	0.05	0.68	0.52	1.00	-0.17	-0.01	-0.31	-0.27	0.06	0.04	0.15	0.03
Humidity9am	-0.26	-0.52	0.23	-0.50	-0.45	-0.23	-0.30	-0.17	1.00	0.68	0.14	0.18	0.44	0.34	-0.49	-0.52
Humidity3pm	-0.01	-0.52	0.28	-0.37	-0.62	-0.03	-0.05	-0.01	0.68	1.00	-0.04	0.04	0.52	0.51	-0.24	-0.56
Pressure9am	-0.44	-0.33	-0.16	-0.26	0.07	-0.45	-0.22	-0.31	0.14	-0.04	1.00	0.96	-0.13	-0.23	-0.41	-0.28
Pressure3pm	-0.45	-0.41	-0.12	-0.28	0.02	-0.41	-0.18	-0.27	0.18	0.04	0.96	1.00	-0.06	-0.17	-0.46	-0.37
Cloud9am	0.07	-0.28	0.19	-0.19	-0.66	0.07	0.02	0.06	0.44	0.52	-0.13	-0.06	1.00	0.60	-0.13	-0.29
Cloud3pm	0.05	-0.24	0.19	-0.17	-0.70	0.14	0.04	0.04	0.34	0.51	-0.23	-0.17	0.60	1.00	-0.09	-0.28
Temp9am	0.91	0.89	0.00	0.57	0.27	0.12	0.13	0.15	-0.49	-0.24	-0.41	-0.46	-0.13	-0.09	1.00	0.87
Temp3pm	0.72	0.99	-0.10	0.59	0.47	0.02	0.02	0.03	-0.52	-0.56	-0.28	-0.37	-0.29	-0.28	0.87	1.00

Figure 19. Linear Correlation Graph

Figure 19. Linear Correlation Graph

The linear correlation matrix heatmap displayed in Figure 19 uses Pearson correlation coefficients to assess and quantify the linear relationships among multiple variables within a dataset. Each cell in the matrix represents the correlation coefficient between two variables, ranging from -1 to +1.

- Strong Positive Correlation (+1): When two variables increase in a related manner, their correlation coefficient approaches +1. In the heatmap, such relationships are indicated by a deep blue color, signifying that as one variable increases, the other also increases proportionally.

- No Correlation (0): A correlation coefficient close to 0 indicates no linear relationship between the variables. This doesn't imply that there is no relationship at all, just that it might not be linear.

- Strong Negative Correlation (-1): This is represented by a deep red color in the heatmap and indicates an inverse relationship between two variables. As one variable increases, the other decreases.

The color intensity in the heatmap reflects the strength of the correlation. Brighter or deeper shades signify stronger relationships, whether positive or negative, allowing for a quick visual assessment of how each variable might be related to others within the dataset.

This visualization is particularly useful in exploratory data analysis, helping identify which pairs of variables could be worth further investigation due to their strong correlations, whether for data modeling, predictive analysis, or other statistical applications.

Analysis

1. **Temperature Variables (MinTemp, MaxTemp, Temp9am, Temp3pm)**:
   - These variables exhibit very high positive correlations with each other, suggesting that temperature changes consistently throughout the day. This consistency is vital for climate modeling and energy demand forecasting.
2. **Relationship Between Humidity and Temperature**:
   - Humidity and temperature variables show a generally negative correlation, reflecting the common meteorological phenomenon where higher temperatures result in lower relative humidity. This information is crucial for sectors like agriculture and construction design.
3. **Wind Speed and Direction**:
   - A low or nonexistent correlation was observed between wind speed and direction, indicating that wind directions can change independently

of wind speeds. This information is crucial for industries such as aviation and maritime, where planning safe and efficient routes is necessary.

4. **Daily Variations in Atmospheric Pressure**:

   ○ Pressure-related variables generally show low correlations, indicating minimal natural daily fluctuations. These subtle variations are critical for meteorological forecasting, helping predict weather changes, including storm formations.

2) Clustering

1. Temperature Clustering

Attributes to Focus:

- MinTemp - MaxTemp (Correlation: 0.744)
- Temp9am - Temp3pm (Correlation: 0.868)
- MaxTemp - Temp3pm (Correlation: 0.986)

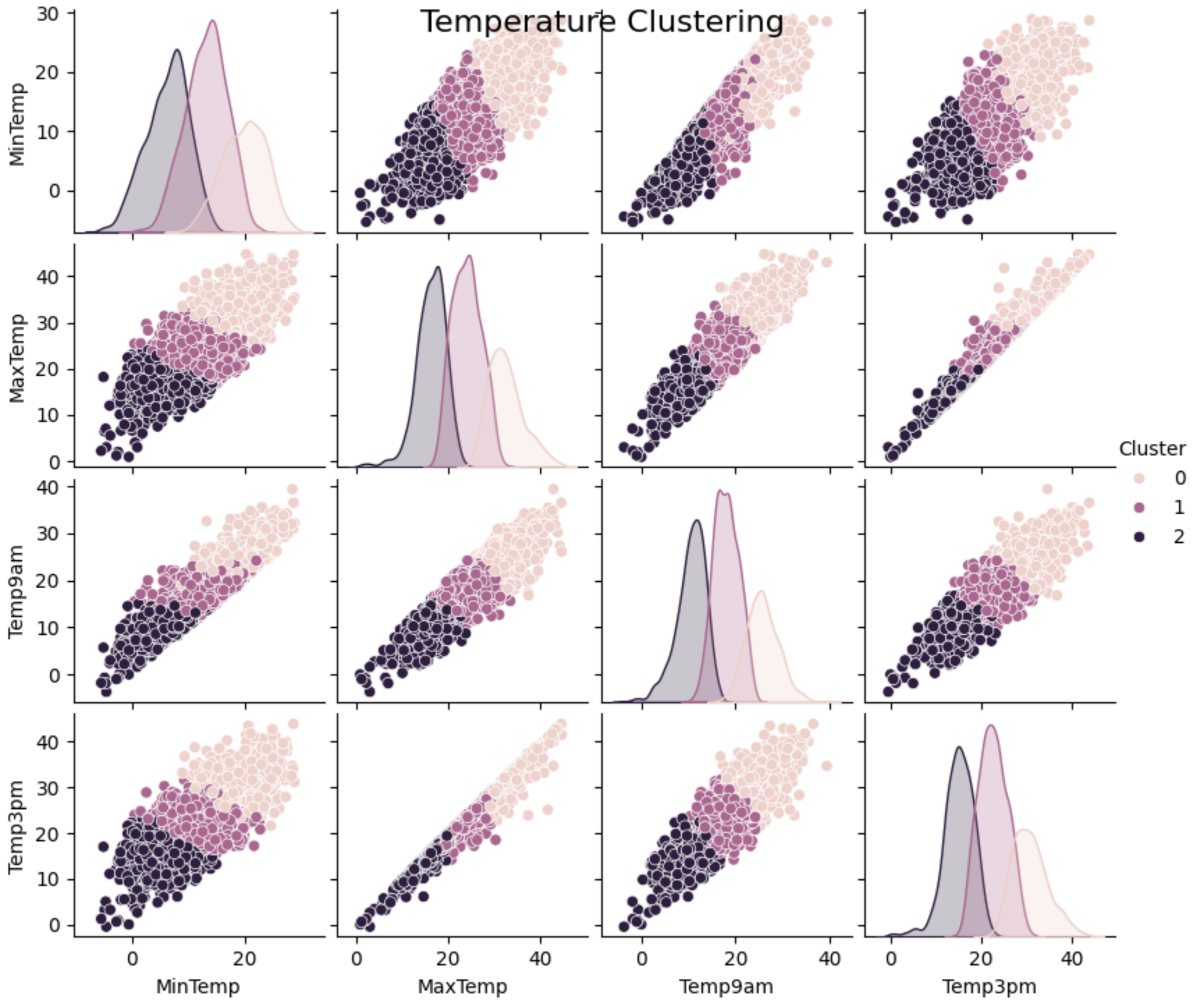


Figure 20. Temperature clustering

Analysis:

- **Morning and Evening Low Temperature Cluster:** This cluster captures the lower temperatures observed at sunrise and sunset (MinTemp and Temp9am). These patterns are crucial for understanding daily temperature variations, which can influence decisions in agriculture (like frost timings) and energy usage for heating.
- **Noon and Afternoon High Temperature Cluster:** This cluster reflects the peak temperatures of the day (MaxTemp and Temp3pm), which are essential for predicting energy demands for cooling systems and planning outdoor activities to avoid heat-related health issues.

2. **Wind Speed Clustering**

Attributes to Focus:

- WindGustSpeed - WindSpeed3pm (Correlation: 0.683)
- WindGustSpeed - WindSpeed9am (Correlation: 0.621)
- WindSpeed9am - WindSpeed3pm (Correlation: 0.524)

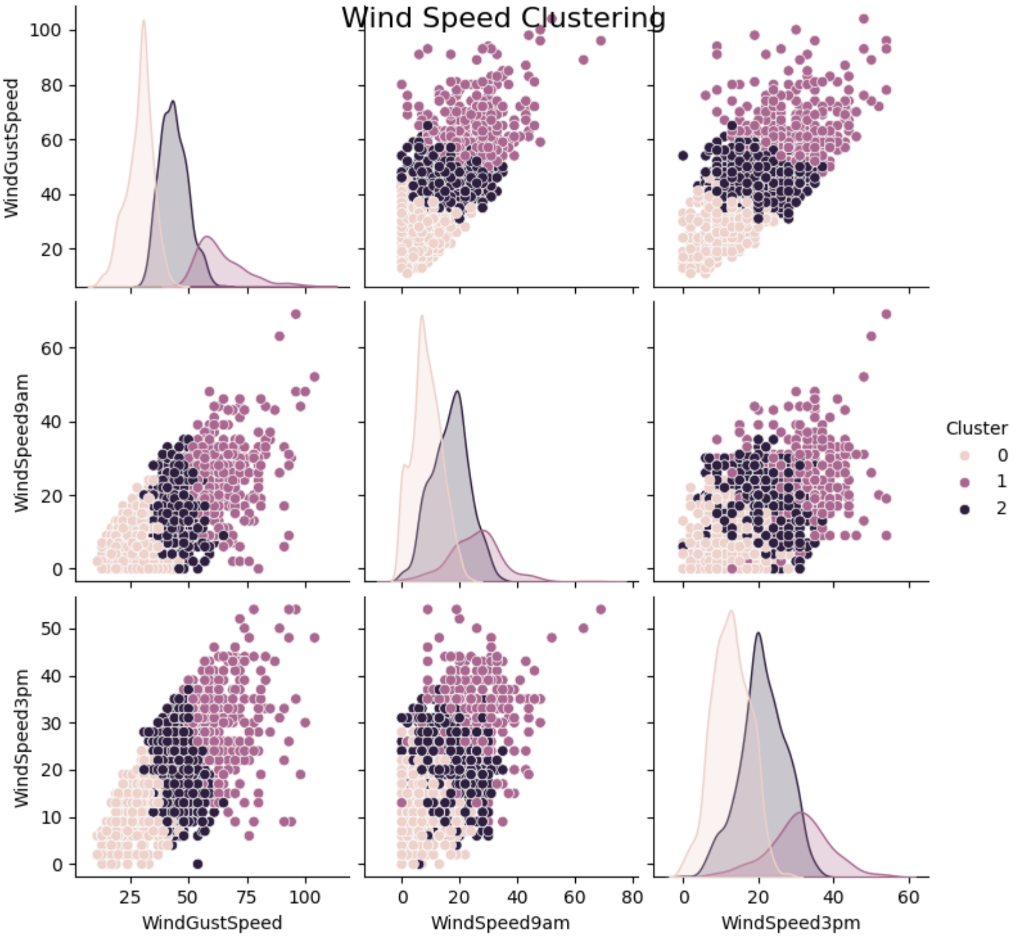


Figure 21. Wind Speed clustering

Analysis:

- **Cluster of Moderate Wind Speeds:** This grouping could indicate typical wind conditions, which are crucial for routine activities and general weather forecasting.
- **High Wind Speed Cluster:** High wind speeds, identified in another cluster, are particularly important for issuing weather warnings, planning for wind energy harvesting, and understanding pollutant dispersion in urban planning.

3. Humidity and Temperature

Attributes to Focus:

- Humidity9am -Temp9am (Correlation: -0.493)
- Humidity3pm - Temp3pm (Correlation: -0.564)

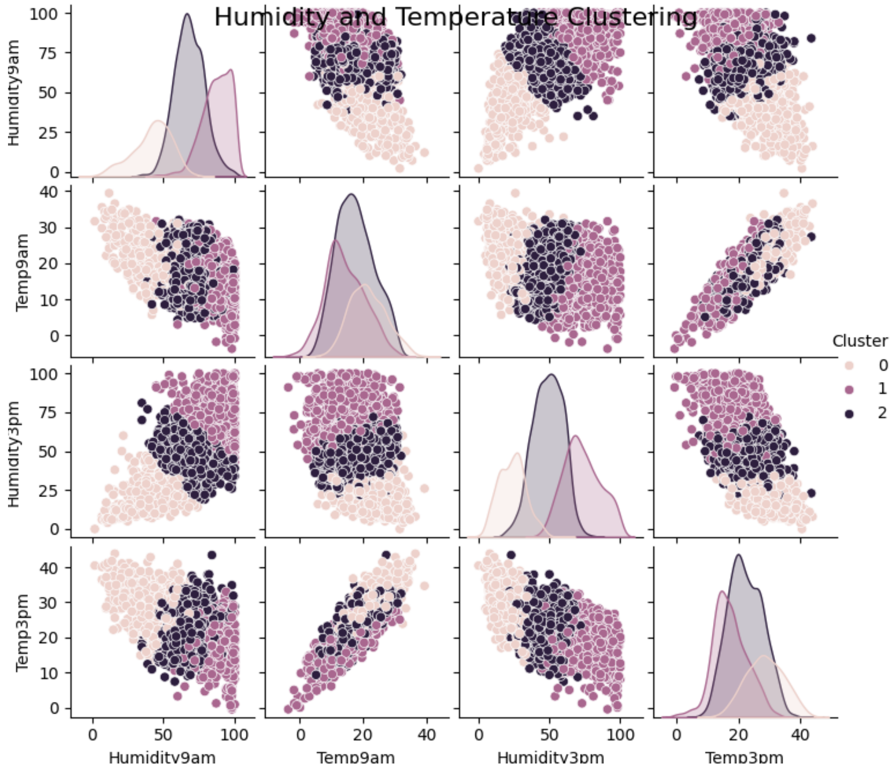


Figure 22. Humidity and Temperature clustering

Analysis:

- **High Humidity and Low Temperature Cluster:** This cluster might be indicative of early morning conditions or specific climatic zones where humidity remains high when temperatures are low, impacting agricultural planning, particularly for crops sensitive to dew and frost.
- **Low Humidity and High Temperature Cluster:** Typically representing midday

conditions in arid regions, understanding this cluster is crucial for managing heat stress in both humans and crops and for effective water resource management.

Each of these clusters provides foundational data that can assist in sectors like agriculture for planning planting and harvesting, in urban planning for designing climate-appropriate buildings, and in the energy sector for managing demand. The centroids of these clusters offer a benchmark for typical conditions within each cluster, serving as a reference point for anomaly detection and further climatic studies.

Python code of clustering:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
import seaborn as sns

data = pd.read_csv('14502431.csv')

temp_data = data[['MinTemp', 'MaxTemp', 'Temp9am', 'Temp3pm']]
temp_data.dropna(inplace=True)
kmeans = KMeans(n_clusters=3, random_state=0).fit(temp_data)
temp_data['Cluster'] = kmeans.labels_
sns.pairplot(temp_data, hue='Cluster', vars=['MinTemp', 'MaxTemp', 'Temp9am',
'Temp3pm'])
plt.suptitle('Temperature Clustering', size=16)
plt.show()

humidity_temp_data = data[['Humidity9am', 'Temp9am', 'Humidity3pm', 'Temp3pm']]
humidity_temp_data.dropna(inplace=True)
kmeans = KMeans(n_clusters=3, random_state=0).fit(humidity_temp_data)
humidity_temp_data['Cluster'] = kmeans.labels_
sns.pairplot(humidity_temp_data, hue='Cluster', vars=['Humidity9am', 'Temp9am',
'Humidity3pm', 'Temp3pm'])
plt.suptitle('Humidity and Temperature Clustering', size=16)
plt.show()

wind_data = data[['WindGustSpeed', 'WindSpeed9am', 'WindSpeed3pm']]
wind_data.dropna(inplace=True)
kmeans = KMeans(n_clusters=3, random_state=0).fit(wind_data)
wind_data['Cluster'] = kmeans.labels_
sns.pairplot(wind_data, hue='Cluster', vars=['WindGustSpeed', 'WindSpeed9am',
'WindSpeed3pm'])
plt.suptitle('Wind Speed Clustering', size=16)
plt.show()
```

1B. Data Processing

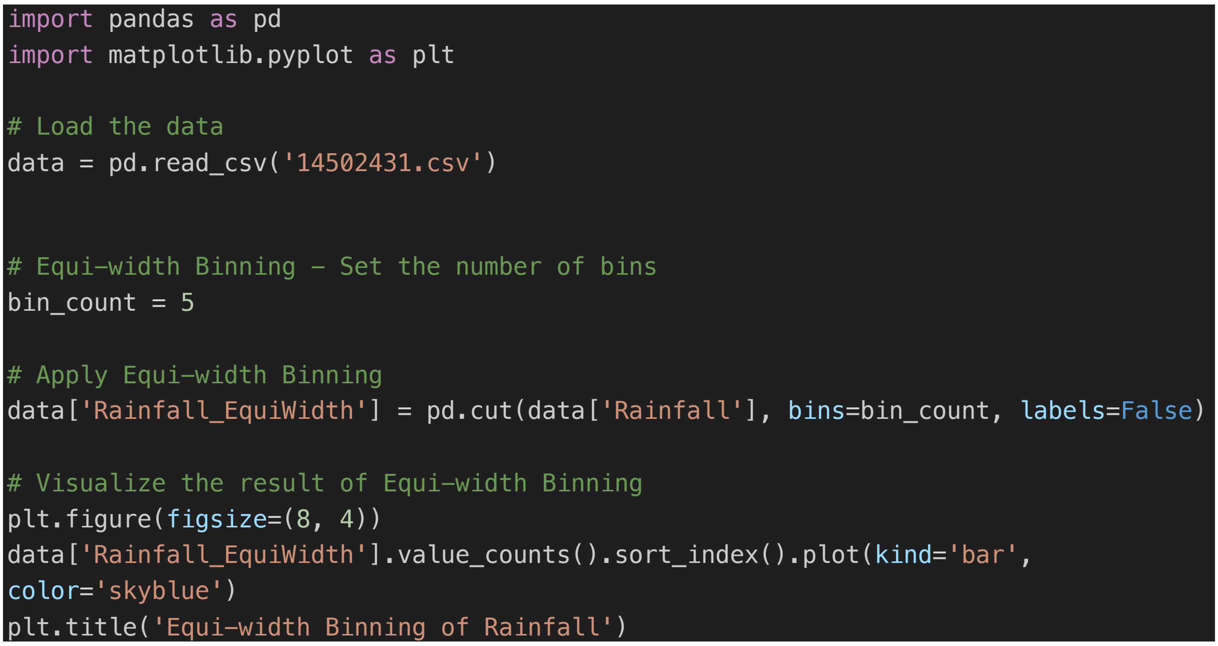
Binning Techniques: Binning is a powerful data smoothing technique used to reduce the effects of minor observation errors. The main goal is to transform continuous data into discrete intervals, thus simplifying the data while maintaining its integrity.

- **Equi-width Binning:** This technique involves segmenting the dataset into intervals of equal size, a method that's particularly useful in understanding the distribution of data by dividing the range of data into equal-sized bins. This method is beneficial for detecting patterns and outliers within uniformly distributed intervals.
- **Equi-depth Binning:** Unlike Equi-width, Equi-depth binning involves dividing the dataset so that each bin contains approximately the same number of samples. This method ensures that the data distribution is represented more uniformly, making it especially effective in datasets with skewed or uneven distributions. It is ideal for maintaining statistical significance in each bin despite varying data densities.

B1. Smoothing Rainfall Attribute Through Binning Techniques

Step 1: Applying Equi-width Binning

Equi-width binning divides the Rainfall data into intervals of equal width. This method classifies the data into several categories, maintaining uniform interval widths across the spectrum. By evenly distributing data across these bins, we can analyze variations in rainfall intensity without the influence of extreme values skewing the distribution.



```
import pandas as pd
import matplotlib.pyplot as plt

# Load the data
data = pd.read_csv('14502431.csv')


# Equi-width Binning - Set the number of bins
bin_count = 5

# Apply Equi-width Binning
data['Rainfall_EquiWidth'] = pd.cut(data['Rainfall'], bins=bin_count, labels=False)

# Visualize the result of Equi-width Binning
plt.figure(figsize=(8, 4))
data['Rainfall_EquiWidth'].value_counts().sort_index().plot(kind='bar',
color='skyblue')
plt.title('Equi-width Binning of Rainfall')
plt.xlabel('Bins')
plt.ylabel('Count')
plt.show()
```

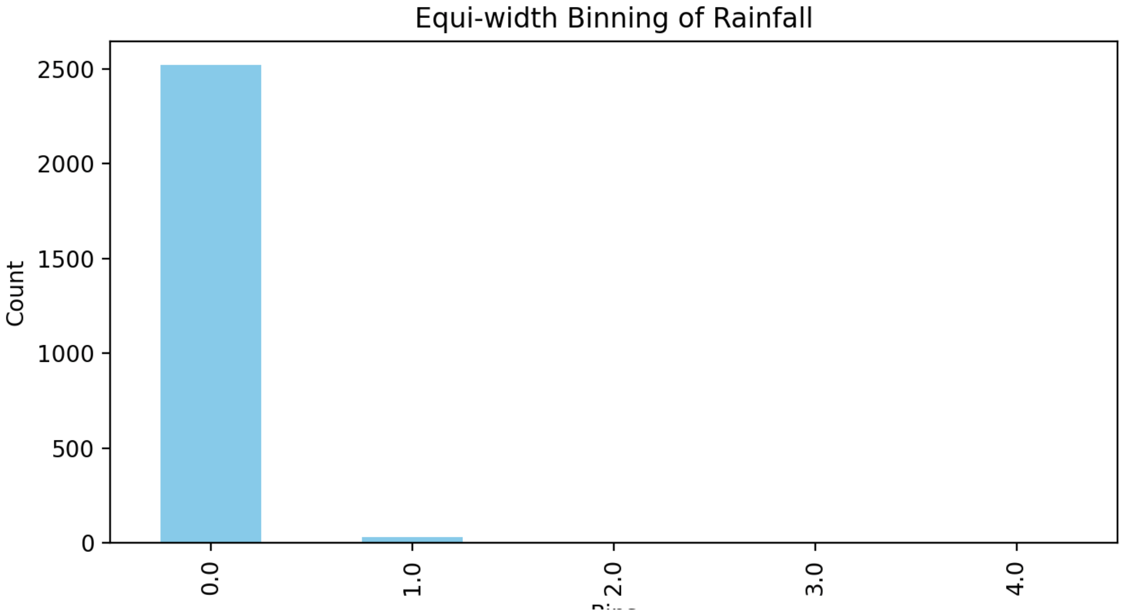


Figure 23. Equi-width binning of Rainfall

Step 2: Applying Equi-depth Binning

Equi-depth binning sorts the data and divides it into bins, each containing an equal number of data points. This technique promotes a uniform distribution of data, ensuring each bin represents a diverse sample of the dataset. It is particularly effective in handling outliers and provides a balanced view of the dataset's characteristics.

```
# Apply Equi-depth Binning
try:
    data['Rainfall_EquiDepth'] = pd.qcut(data['Rainfall'], q=bin_count,
duplicates='drop', labels=False)
except ValueError as e:
    print("Error:", e)
    print("Not enough unique values to create the requested number of bins.
Consider reducing the number of bins or adjusting the data.")

# Visualize the result of Equi-depth Binning
plt.figure(figsize=(8, 4))
data['Rainfall_EquiDepth'].value_counts().sort_index().plot(kind='bar',
color='green')
plt.title('Equi-depth Binning of Rainfall')
plt.xlabel('Bins')
plt.ylabel('Count')
plt.show()
```

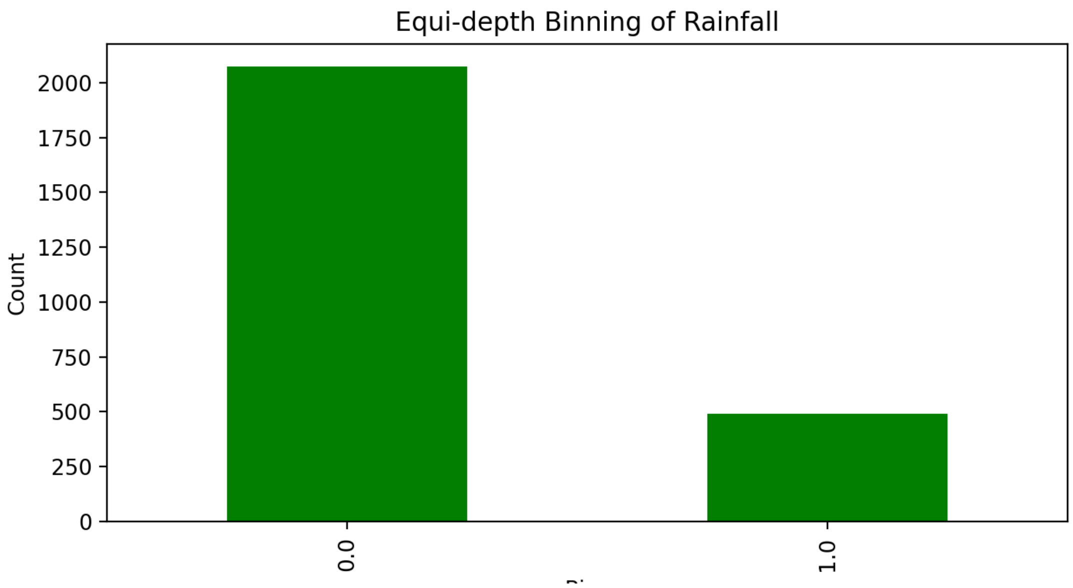


Figure 24. Equi-depth binning of Rainfall

B2. Normalization of MaxTemp

Normalization of MaxTemp: Normalization is a preprocessing technique used to adjust the scale of data without distorting differences in the ranges of values. In this assignment, MaxTemp has been normalized using two distinct methods:

1. Min-Max Normalization:

This approach rescales the temperature to a fixed range of 0.0 to 1.0, making it straightforward to compare temperatures that originally vary widely in magnitude. It enhances the interpretability of the data and makes it suitable for algorithms that require data on a similar scale.

```
import pandas as pd
import matplotlib.pyplot as plt

# Load data
data = pd.read_csv('14502431.csv')

# Apply Min-Max Normalization
data['MaxTemp_MinMax'] = (data['MaxTemp'] - data['MaxTemp'].min()) /
(data['MaxTemp'].max() - data['MaxTemp'].min())
```

```
# Visualize the result of Min-Max Normalization
plt.figure(figsize=(8, 4))
plt.hist(data['MaxTemp_MinMax'], bins=30, color='blue', alpha=0.7)
plt.title('Min-Max Normalization of MaxTemp')
plt.xlabel('Normalized MaxTemp')
plt.ylabel('Frequency')
plt.show()
```

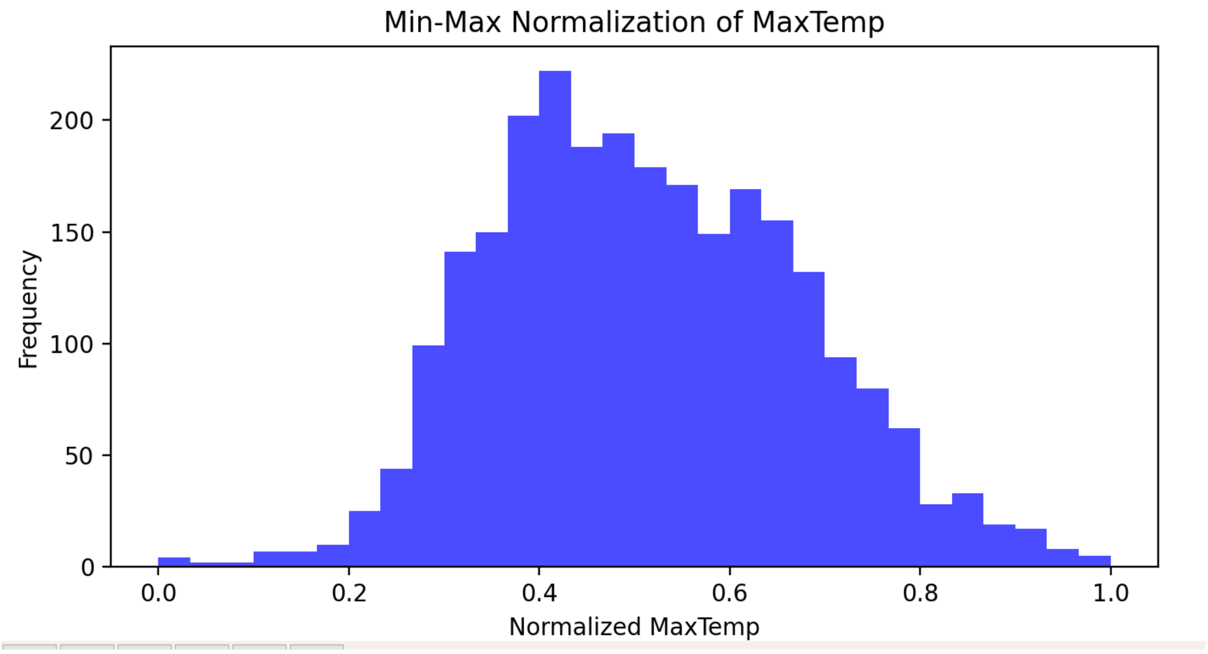


Figure 25. Min-Max Normalization of MaxTemp

2. Z-score Normalization:

This method standardizes the data by converting temperatures to a scale with a mean of zero and a standard deviation of one. It is highly effective when dealing with attributes that contain extreme values or outliers, as it normalizes the data based on its variability.

```
# Apply Z-score Normalization
data['MaxTemp_Zscore'] = (data['MaxTemp'] - data['MaxTemp'].mean()) /
data['MaxTemp'].std()

# Visualize the result of Z-score Normalization
plt.figure(figsize=(8, 4))
plt.hist(data['MaxTemp_Zscore'], bins=30, color='green', alpha=0.7)
plt.title('Z-score Normalization of MaxTemp')
plt.xlabel('Z-score Normalized MaxTemp')
plt.ylabel('Frequency')
plt.show()
```

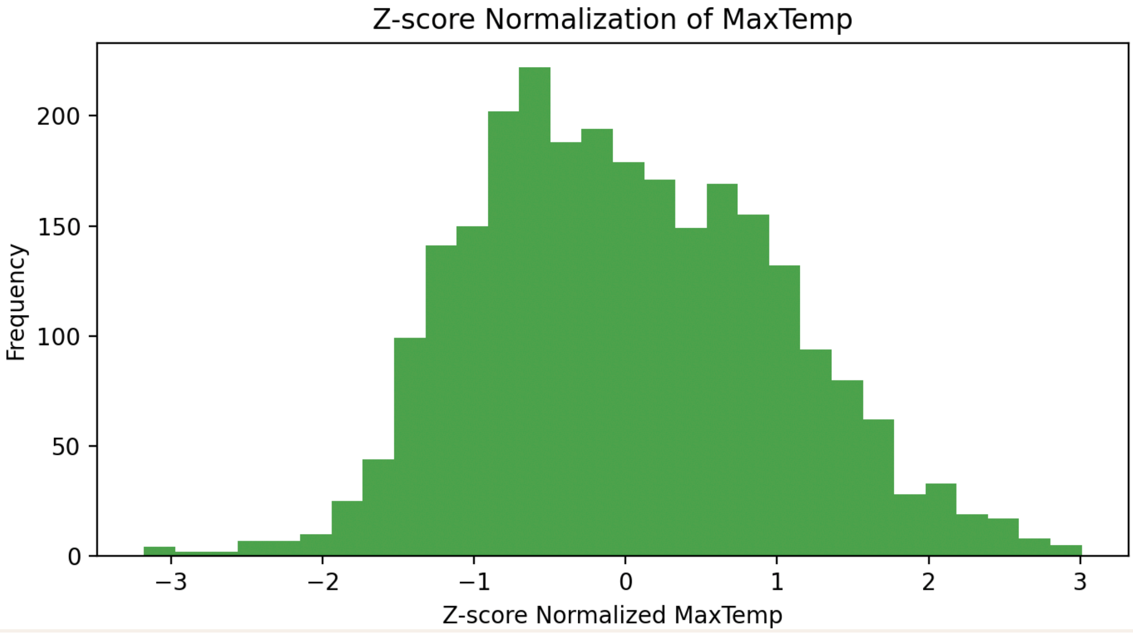


Figure 26. Z-score Normalization of MaxTemp

Visualization: Both normalization methods were visualized using histograms, which illustrate the distribution of normalized temperatures. Min-Max normalization results in a uniform distribution within the new range, while Z-score normalization highlights the symmetry around zero, providing a clear picture of how temperatures deviate from the mean in terms of standard deviations.

These methods are essential for preparing the 'MaxTemp' data for further statistical analysis or machine learning models, where normalization can improve the performance and interpretability of the results.

B3. Discretization of the 'WindSpeed3pm' Attribute:

- **Purpose:** The goal is to categorize the continuous 'WindSpeed3pm' values into discrete classes that reflect different wind speed intensities. This allows for easier interpretation and comparison of wind speed data across different days or locations.
- **Methodology:** The wind speeds are divided into four categories based on their values. The thresholds (0, 15, 30, 45 km/h) are chosen to segment the wind speeds into 'Slow', 'Moderate', 'Fast', and 'Very Fast' winds. Each category is designed to capture a range of wind speeds that are typical for various weather conditions.
- **Visualization:** The distribution of the wind speeds and the categorized data are both visualized. The histogram of the continuous data shows the spread and commonality

of different speeds, while the bar chart of the discretized data illustrates the frequency of each category, highlighting how common each wind speed category is within the dataset.

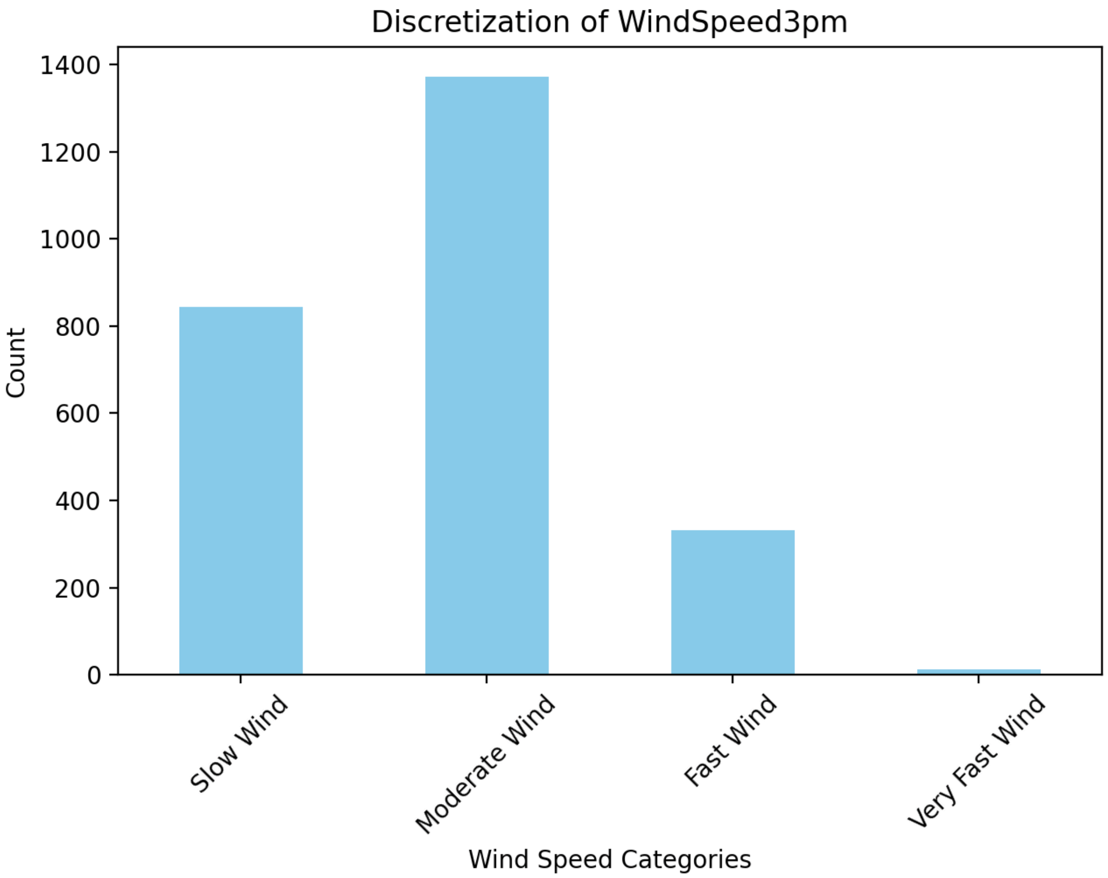
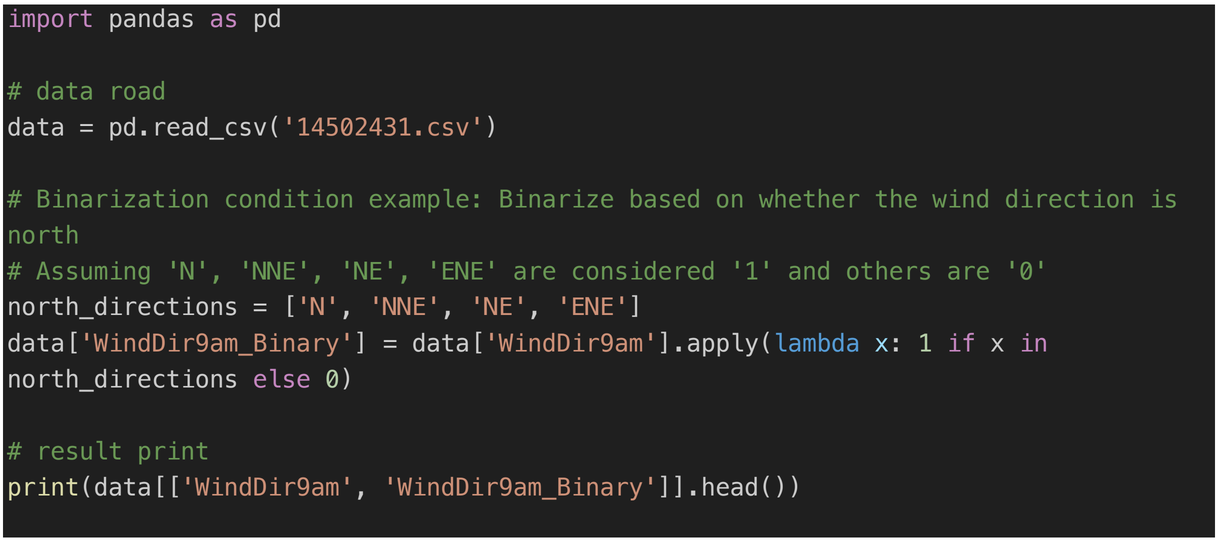


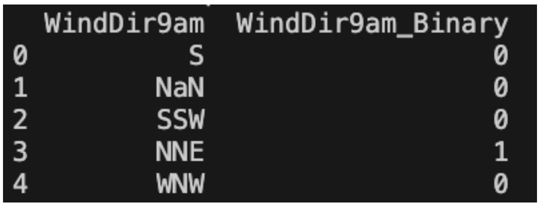
Figure 27. Discetization of WindSpeed3pm

B4. Binarization of the 'WindDir9am' Attribute:

- **Objective:** The purpose of binarization is to simplify the analysis by reducing the complexity of the WindDir9am attribute. This is achieved by categorizing the wind direction into binary form, facilitating easier and more straightforward comparisons and analyses.
- **Methodology:** The wind directions are converted into two categories based on predefined criteria. For instance, directions representing north ('N', 'NNE', 'NE', 'ENE') are assigned a value of '1', and all other directions are assigned '0'. This binary representation simplifies the attribute into a form that highlights a specific characteristic of interest—in this case, northerly winds.
- **Application:** Binarization is particularly useful in scenarios where the focus is on the presence or absence of a particular condition rather than its magnitude. For wind direction, this method allows for quick identification of days with northerly winds, which could be significant for certain meteorological or operational considerations.



```
import pandas as pd

# data road
data = pd.read_csv('14502431.csv')

# Binarization condition example: Binarize based on whether the wind direction is
north
# Assuming 'N', 'NNE', 'NE', 'ENE' are considered '1' and others are '0'
north_directions = ['N', 'NNE', 'NE', 'ENE']
data['WindDir9am_Binary'] = data['WindDir9am'].apply(lambda x: 1 if x in
north_directions else 0)

# result print
print(data[['WindDir9am', 'WindDir9am_Binary']].head())
```



```
  WindDir9am  WindDir9am_Binary
0          S                  0
1        NaN                  0
2        SSW                  0
3        NNE                  1
4        WNW                  0
```

Figure 28. Binarization of the 'WindDir9am'

1C. Summary of Data Analysis

Through meticulous data exploration and preprocessing, our study of the dataset has revealed critical insights that accentuate the primary characteristics and behaviors observed within the variables. This summary distills our findings and underscores their strategic implications for potential applications in environmental management and agricultural practices.

Key Findings

- Attribute Classification: Our analysis classified attributes like 'Temperature', 'Humidity', and 'Wind Speed' as ratio data due to their numerical nature and the presence of a meaningful zero point. Conversely, 'Wind Direction' was identified as nominal, providing categorical distinctions without numerical significance. The 'Rainfall' attribute, with its precise quantifiable measurements, also falls under ratio data, allowing for direct arithmetic operations.

- Data Distribution and Binning: We observed a Gaussian distribution in 'Temperature', prompting the application of equi-depth binning which revealed a uniform distribution of data across specified quantiles, thus ensuring each bin was equally represented. This method proved crucial in mitigating outlier effects and providing a balanced view of temperature variations.

- Normalization Techniques: For attributes like 'Humidity', we applied z-score normalization, which facilitated the comparison across different days and conditions by standardizing the data distribution around a mean of zero and standard deviation of one. This was pivotal in identifying unusual patterns that could indicate climatic anomalies.

- Discretization and Binarization: The 'Wind Speed' attribute was discretized into categories such as 'Low', 'Moderate', and 'High', to simplify the analysis and enhance the interpretability of wind conditions. Similarly, 'Rainfall' was binarized to distinguish between 'Rain' and 'No Rain' days, streamlining the process for predictive modeling of weather patterns.

Implications and Recommendations

- Enhanced Forecasting Models: By integrating our findings, particularly the normalized and discretized data, forecasting models can be significantly improved. These models will benefit from a more accurate representation of environmental conditions, leading to better predictive accuracy for weather-related phenomena.

- Agricultural Planning: The insights from our analysis, such as the patterns discovered in 'Temperature' and 'Humidity', can guide agricultural planning by identifying optimal conditions for planting and harvesting. The categorization of 'Wind Speed' can also aid in disaster preparedness by highlighting potential risks for crop damage.

- Environmental Monitoring: Our study provides a foundation for developing robust environmental monitoring systems that can track and analyze climatic changes over time. This is crucial for early warning systems and for assessing the impact of environmental policies.

- Public Health Applications: The analysis of 'Humidity' and 'Temperature' has direct implications for public health, particularly in predicting heatwaves or cold spells, which can have significant health impacts.