Data Exploration and Preparation

Introduction to Data Analytics (31250)

Aiden Abignano
13911211

1A. Initial Data Exploration

1A.1. Identify the attribute type of each attribute in your dataset

Attribute	**Attribute Type**	**Reason**
Date	Interval	- Ordered - Measured in fixed and equal units (days) - No zero point and different date systems (Gregorian/Julian) can affect the scale when using dates strictly instead of counting a fixed unit like days
Location	Nominal	- Distinct labels - No logical order
MinTemp	Interval	- Ordered - Measured in fixed and equal units (celsius) - No defined 0 point so product does not make sense - i.e. 80°F = 25°C; 40°F = 5°C So is it twice as hot or five times as hot?
MaxTemp	Interval	- Same as MinTemp
Rainfall	Ratio	- Ordered - Measured in fixed and equal units (millimetres) - Defined 0 point (can’t have negative rainfall) - i.e. 25mm = 1″; 50mm = 2″
Evaporation	Ratio	- Ordered - Measured in fixed and equal units (millimetres) - Defined 0 point - i.e. 25mm = 1″; 50mm = 2″
Sunshine	Ratio	- Ordered - Measured in fixed and equal units (hours) - Defined 0 point
WindGusDir	Nominal	- Distinct labels - Cannot be ordered
WindGusSpeed	Ratio	- Ordered - Measured in fixed and equal units (km/h) - Defined 0 point (can’t have negative speed)
Temp	Interval	- Same as MinTemp

Humidity	Ratio	- Ordered - Measured in fixed and equal units (percent) - Defined 0 point (0% represents no water in air) - Product works - i.e. 40% is twice as much water in the air as 20%
Cloud	Ordinal	- Ordered - Measured in fixed and equal units (Okta) - Defined 0 point (0 okta = no clouds)
WindDir	Nominal	- Same as WindGusDir
WindSpeed	Ratio	- Same as WindGusSpeed
Pressure	Ratio	- Ordered - Measured in fixed and equal units (hectopascals) - Defined 0 point - Product works - i.e. 100 hPa = 1.45 PSI; 200 hPa = 2.9 PSI
RainToday	Ordinal (binary)	- Measured in fixed and equal units (true/false) - Can be ordered (i.e. true = 1; false = 0) and mean value can be found
RainTomorrow	Ordinal (binary)	- Same as RainToday

1A.2. Identify the values of the summarising properties for the attributes

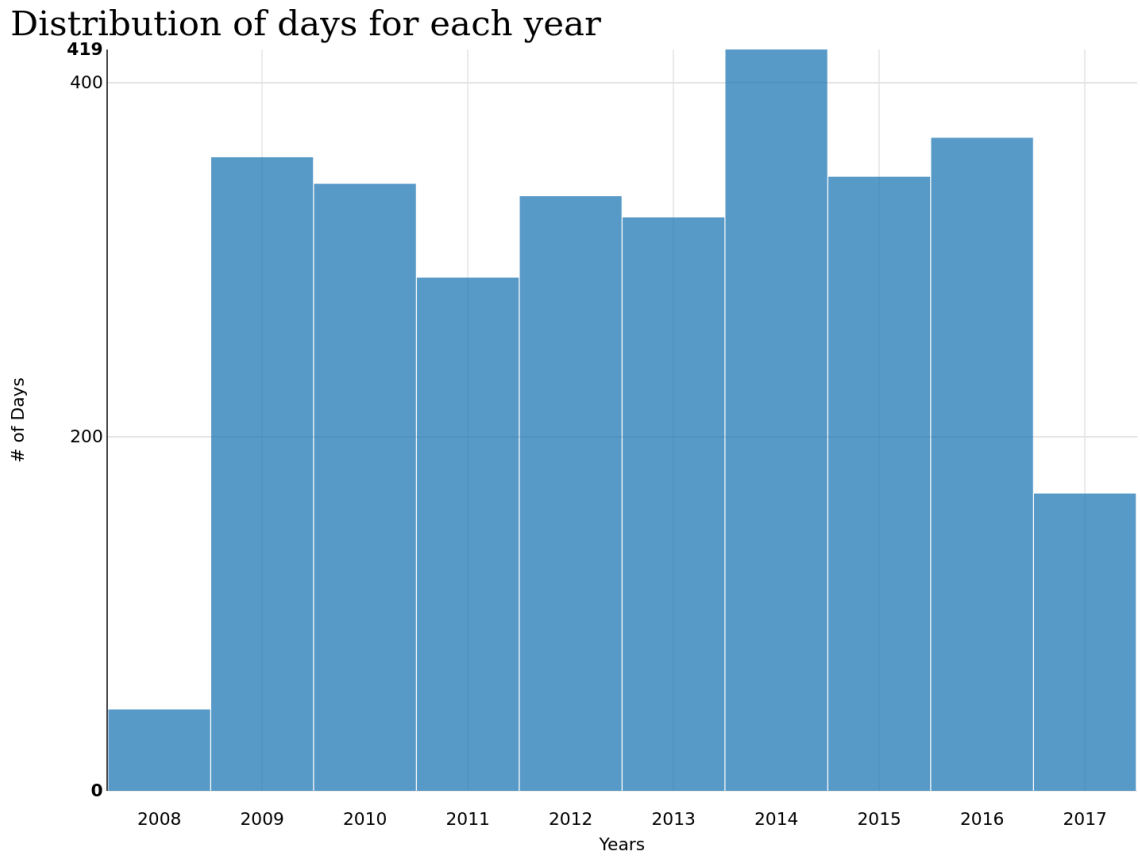
1A.2.1 Date

Data attribute has been split into a months and years attribute to TEST for any biases in the date sampling (i.e. too much data in months in the middle of the year could affect the average rainfall amount)

Years:

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
2008	2017	2012.762	2013	2014	6.371	2.524	-0.062

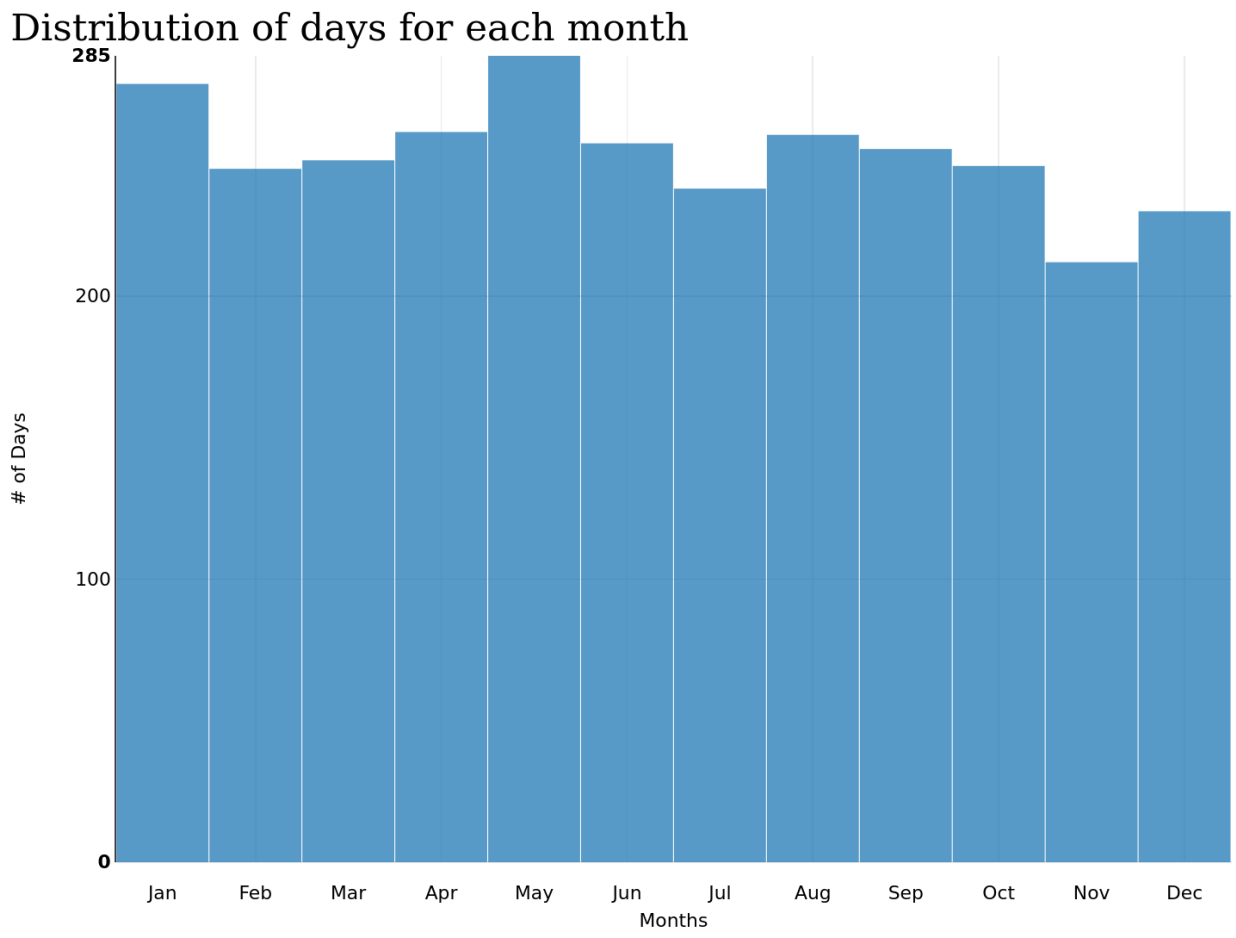
Histogram



Months:

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
1	12	6.344	6	5	11.684	3.418	0.047

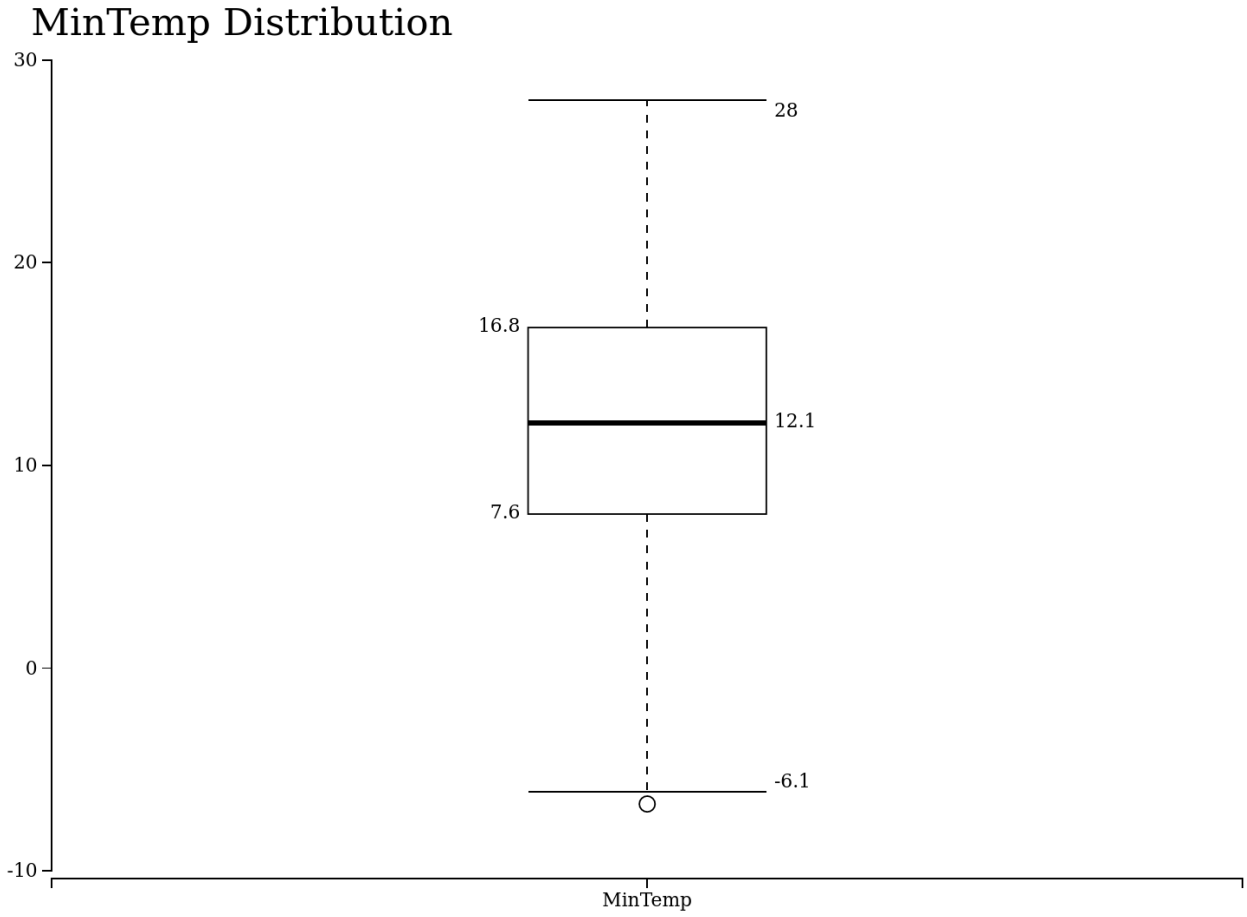
Histogram



1A.2.2 MinTemp

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
-6.7	28	12.152	12.1	13.1	41.085	6.41	-0.006

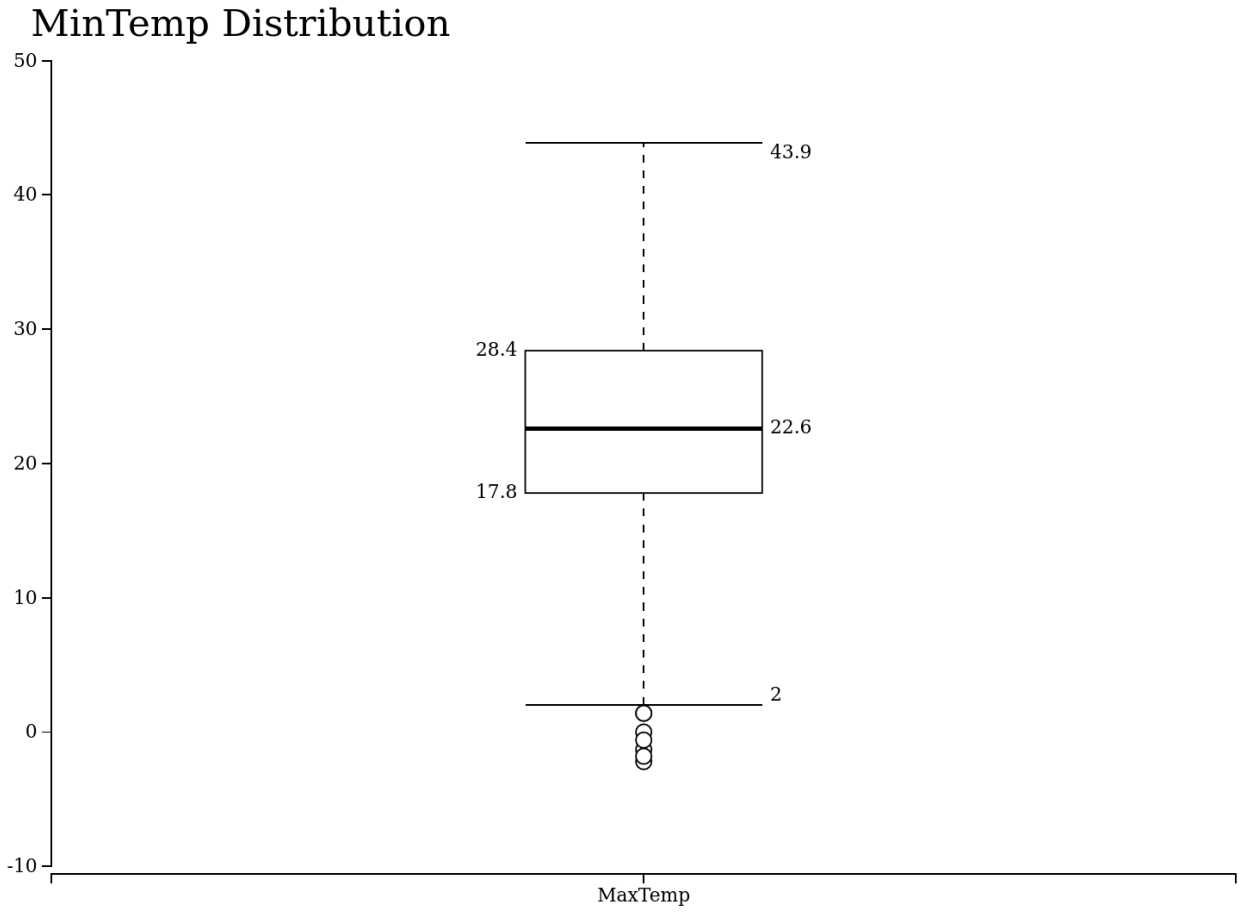
Box Plot



1A.2.3 MaxTemp

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
-2.2	43.9	23.204	22.6	16.3	50.862	7.132	0.177

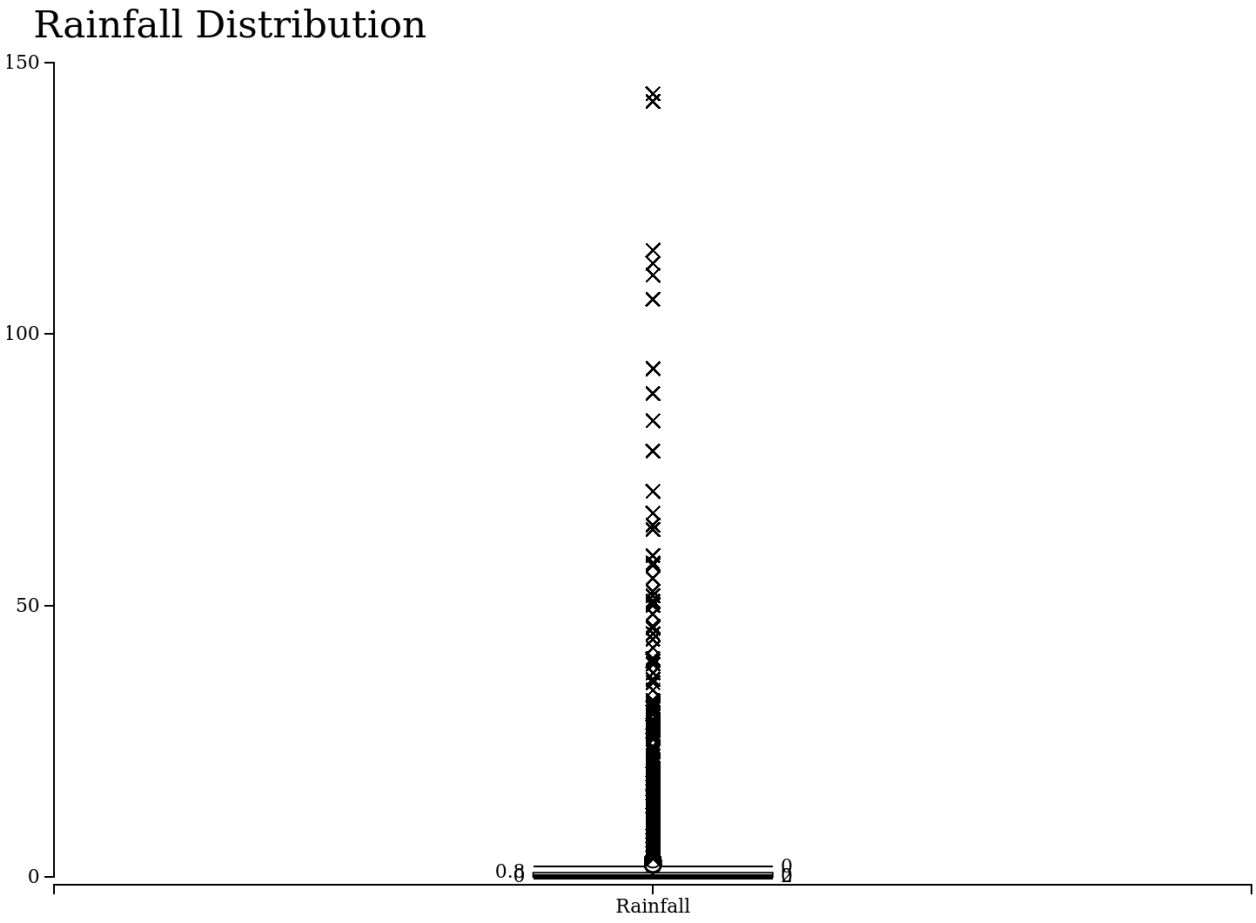
Box Plot



1A.2.4 Rainfall

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
0	144.2	2.492	0	0	83.608	9.144	7.878

Box Plot

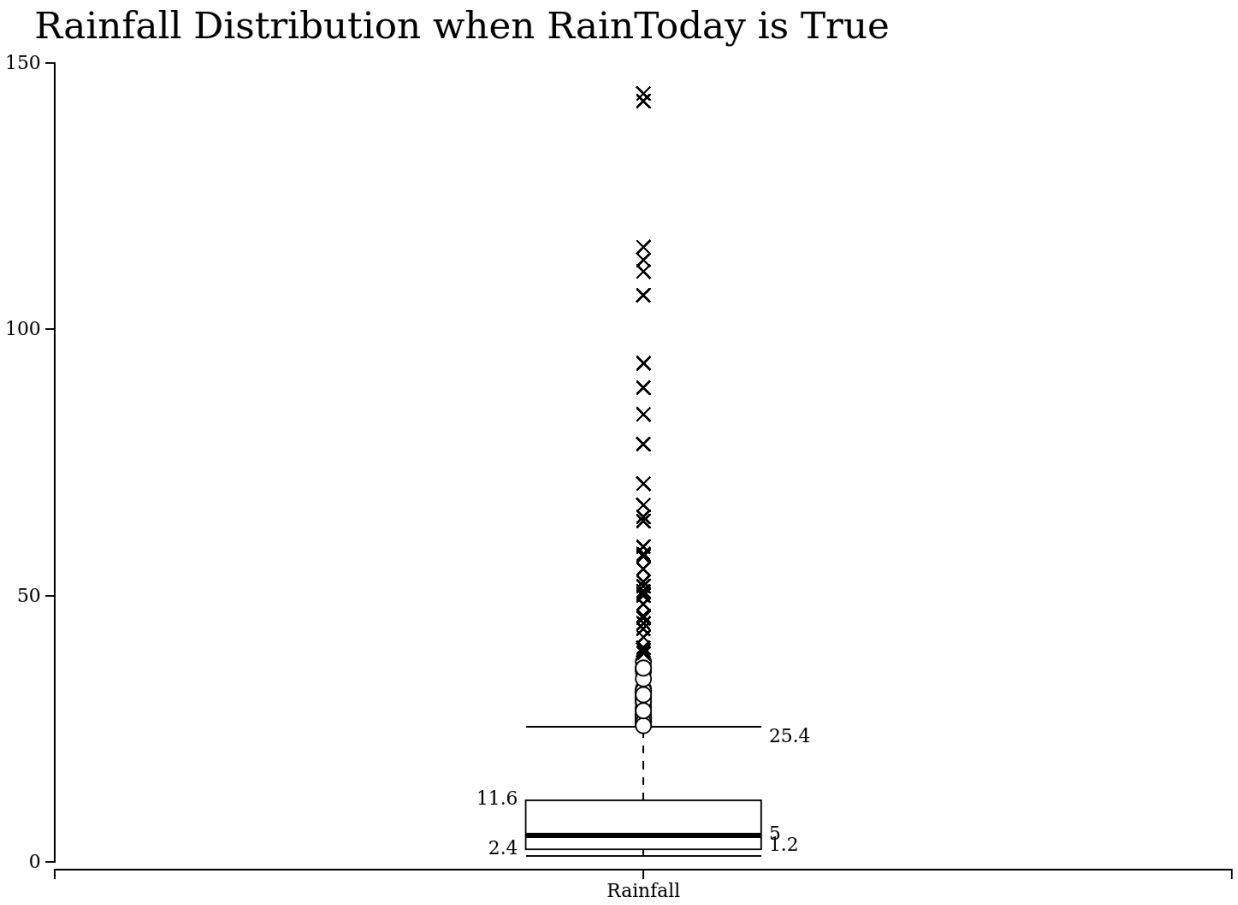


The data is obviously very skewed as a majority of dates, it did not rain leaving when it did rain to look like an outlier. For this reason, an extra summary has been added to more accurately reflect rainfall distribution as it only contains data when *RainToday* is *True*.

Rainfall (RainToday => True)

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
1.2	144.2	10.852	16.823	1.2	283.015	16.823	4.054

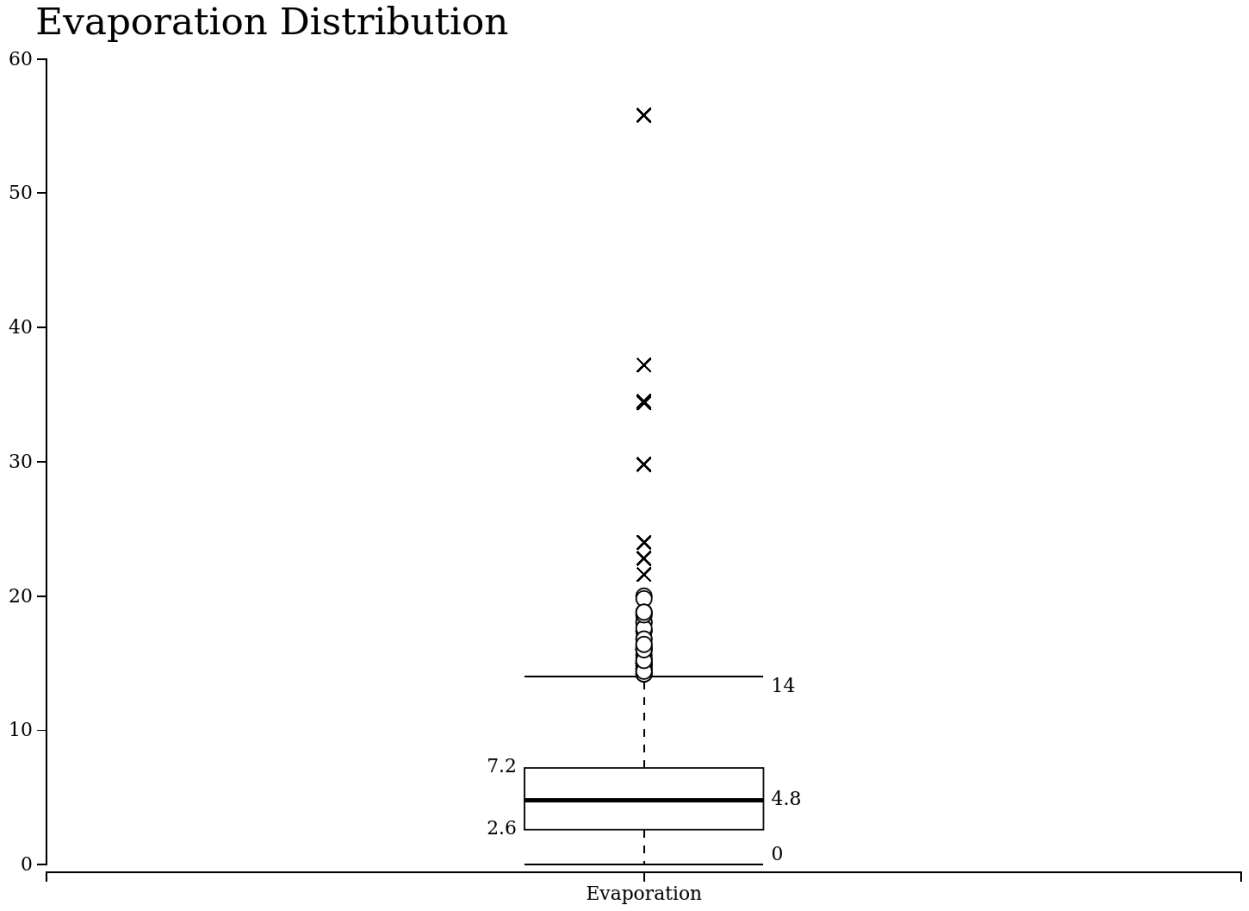
Box Plot



1A.2.5 Evaporation

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
0	55.8	5.451	4.8	4	15.885	3.986	2.876

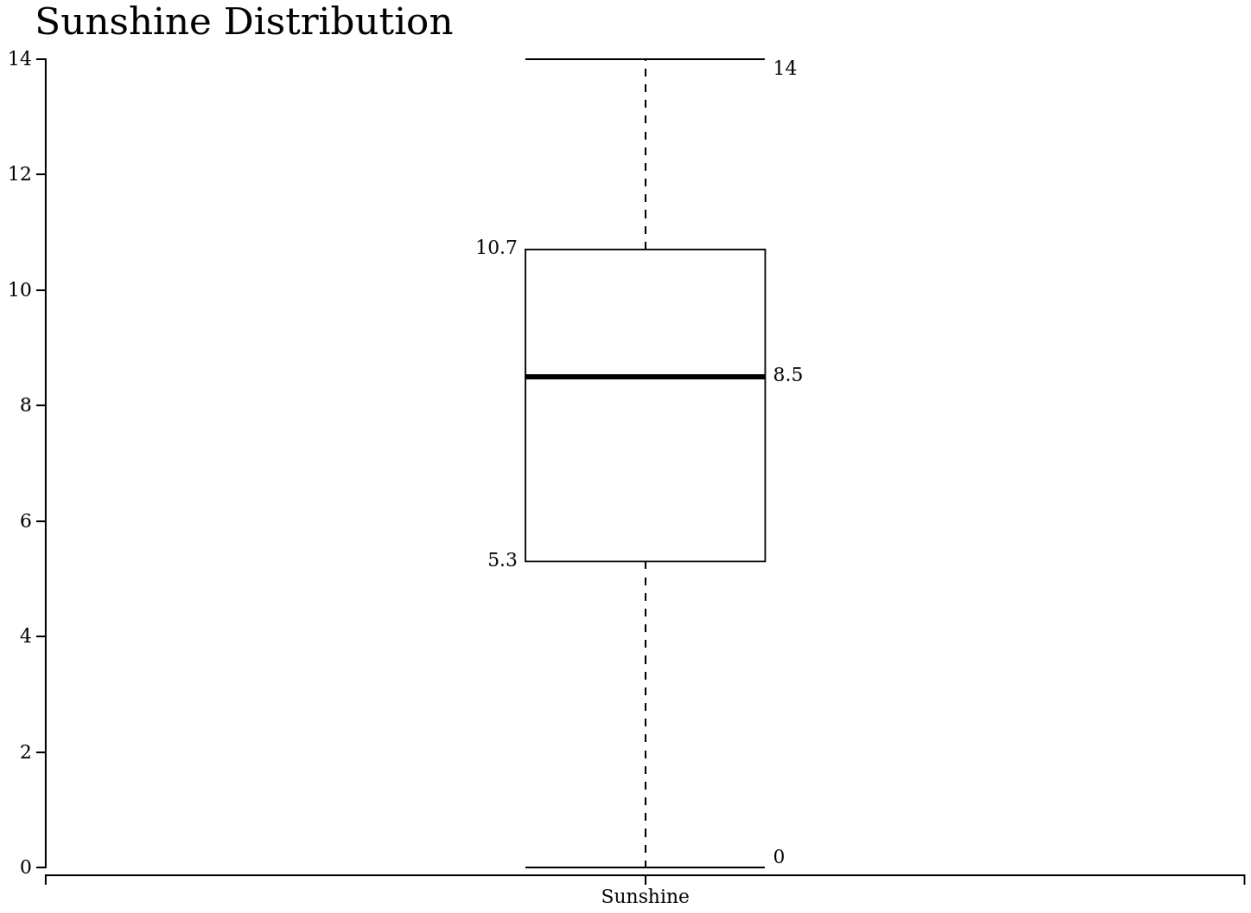
Box Plot



1A.2.6 Sunshine

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
0	14	7.751	8.5	0	13.842	3.721	-0.547

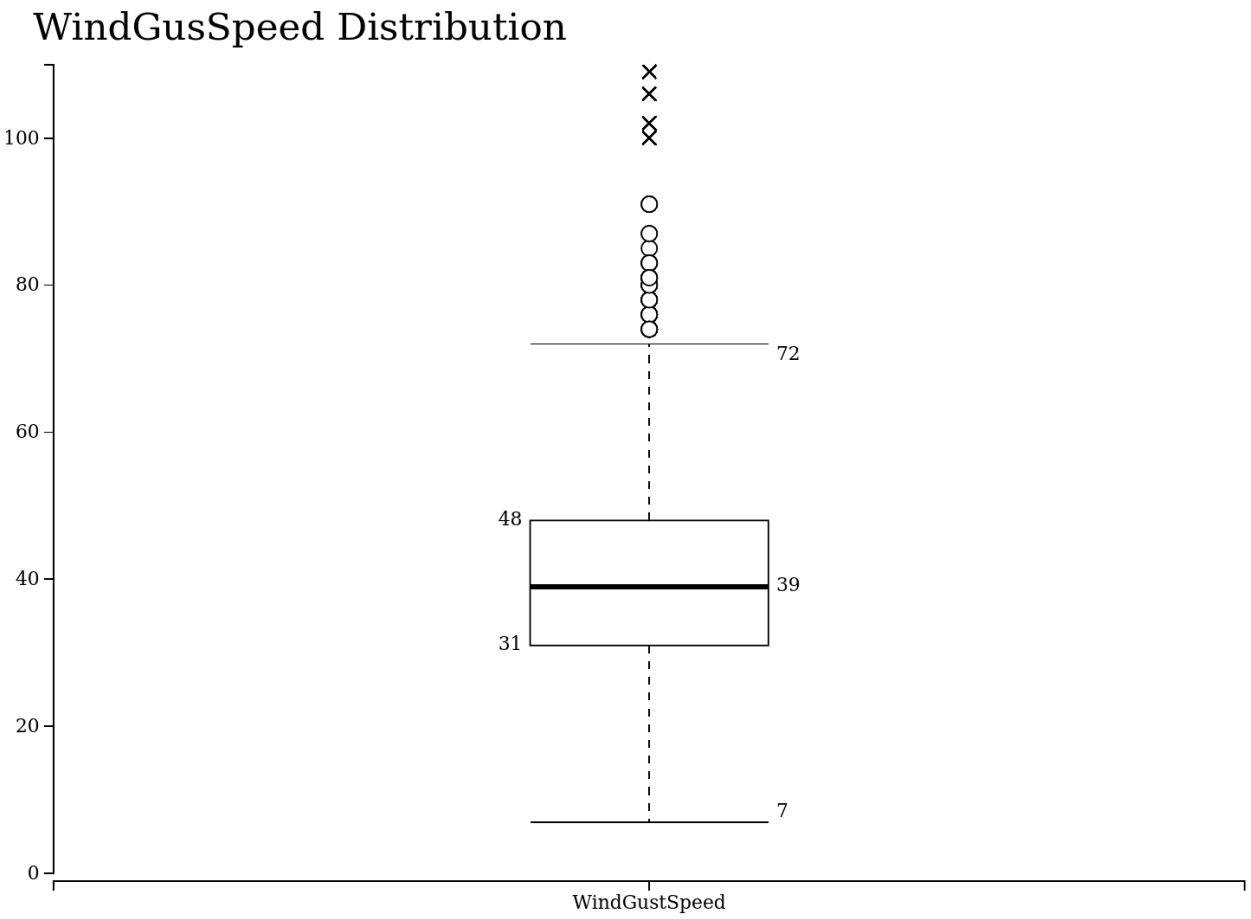
Box Plot



1A.2.7 WindGusSpeed

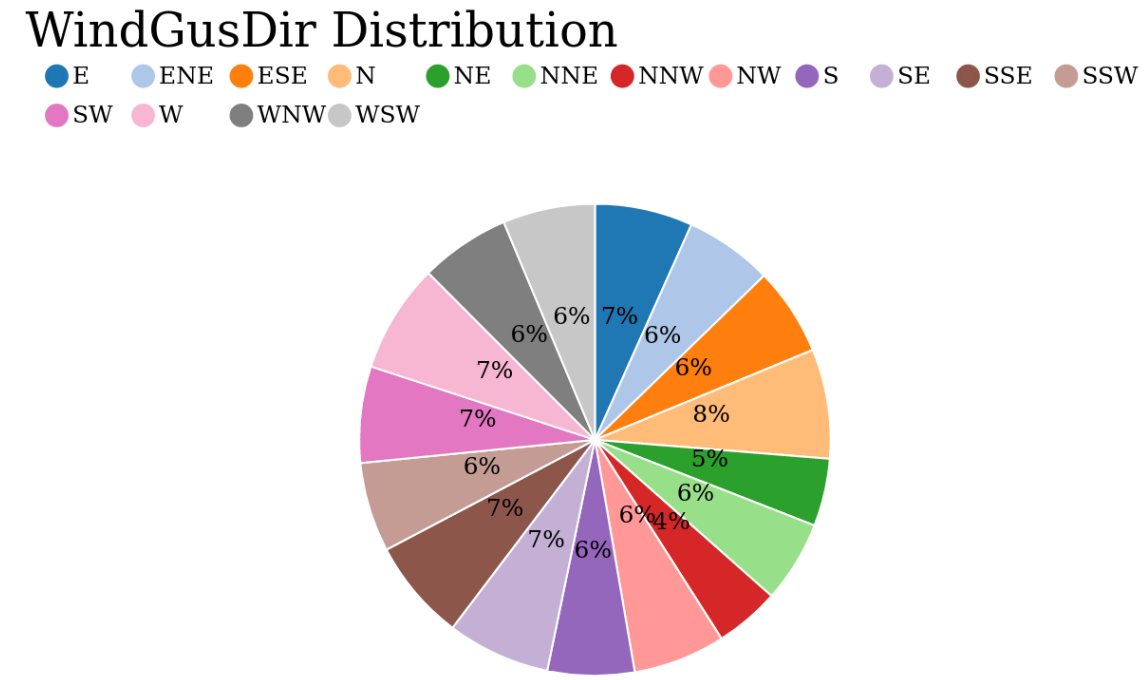
Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
7	109	40.16	39	39	178.702	13.368	0.83

Box Plot

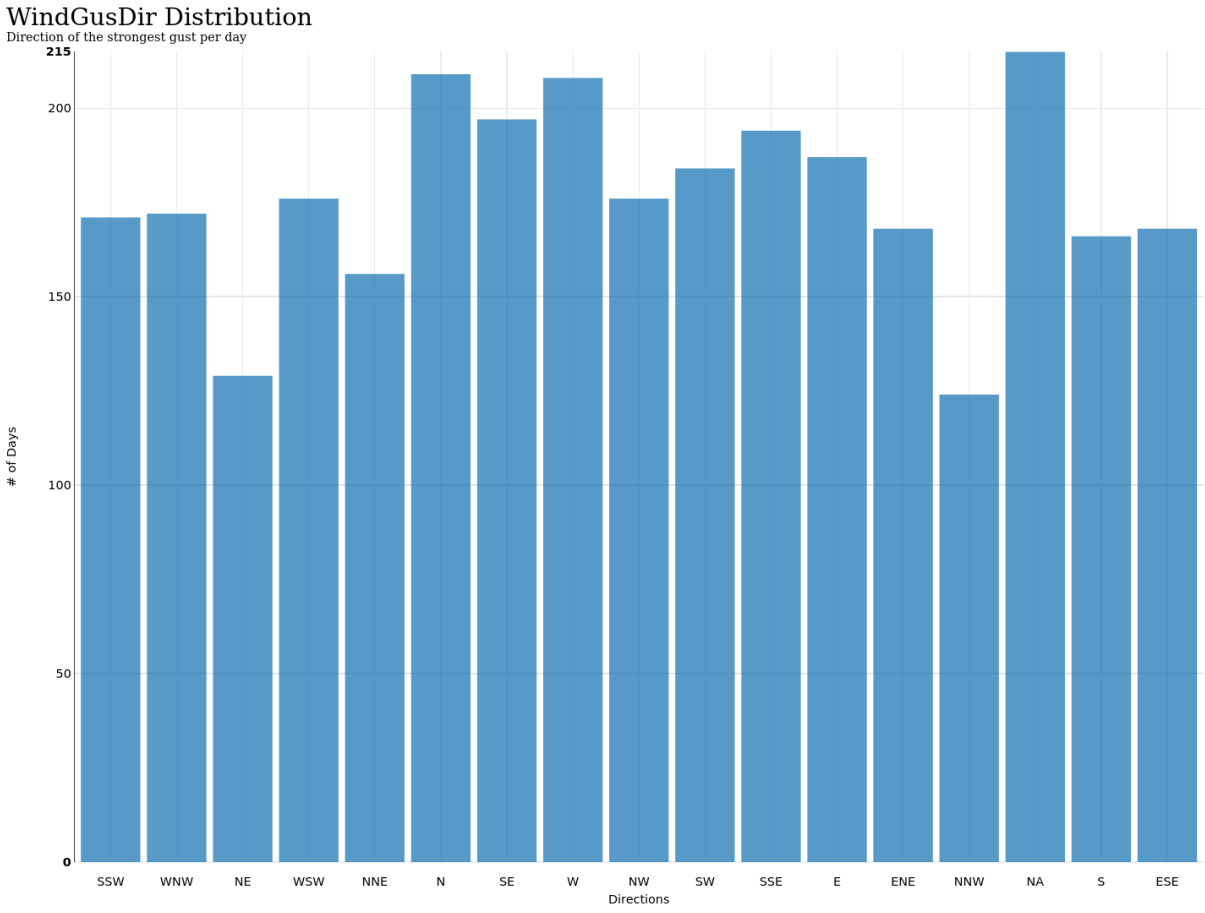


1A.2.8 WindGusDir

Pie Chart

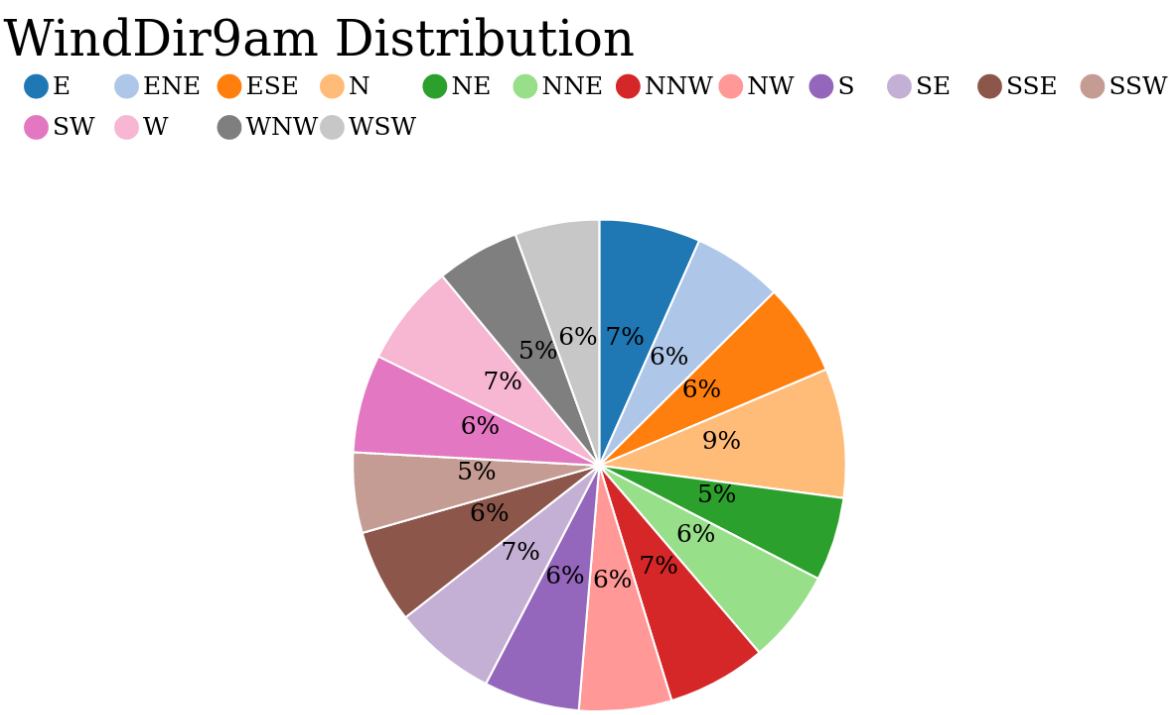


Bar Chart

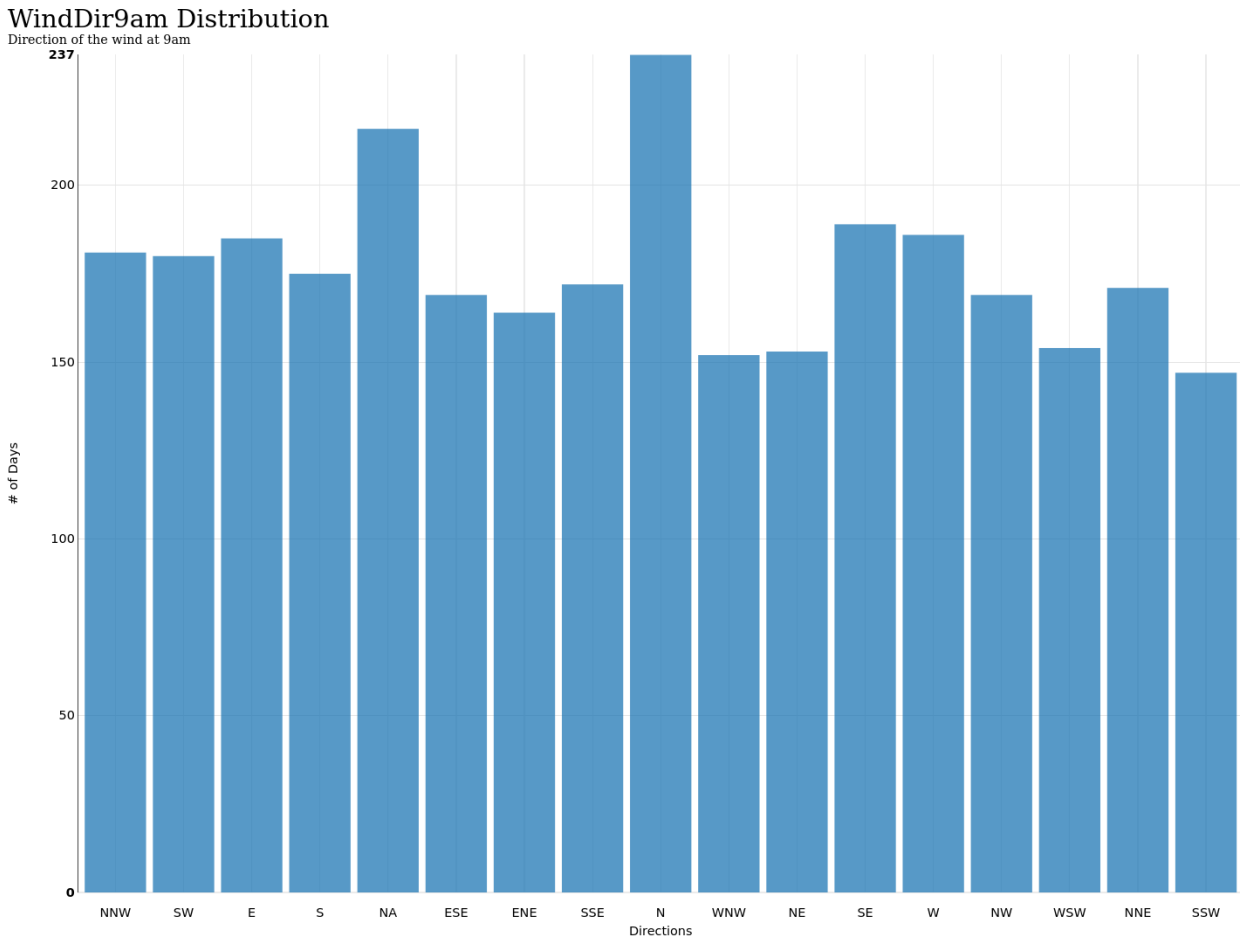


1A.2.9 WindDir9am

Pie Chart

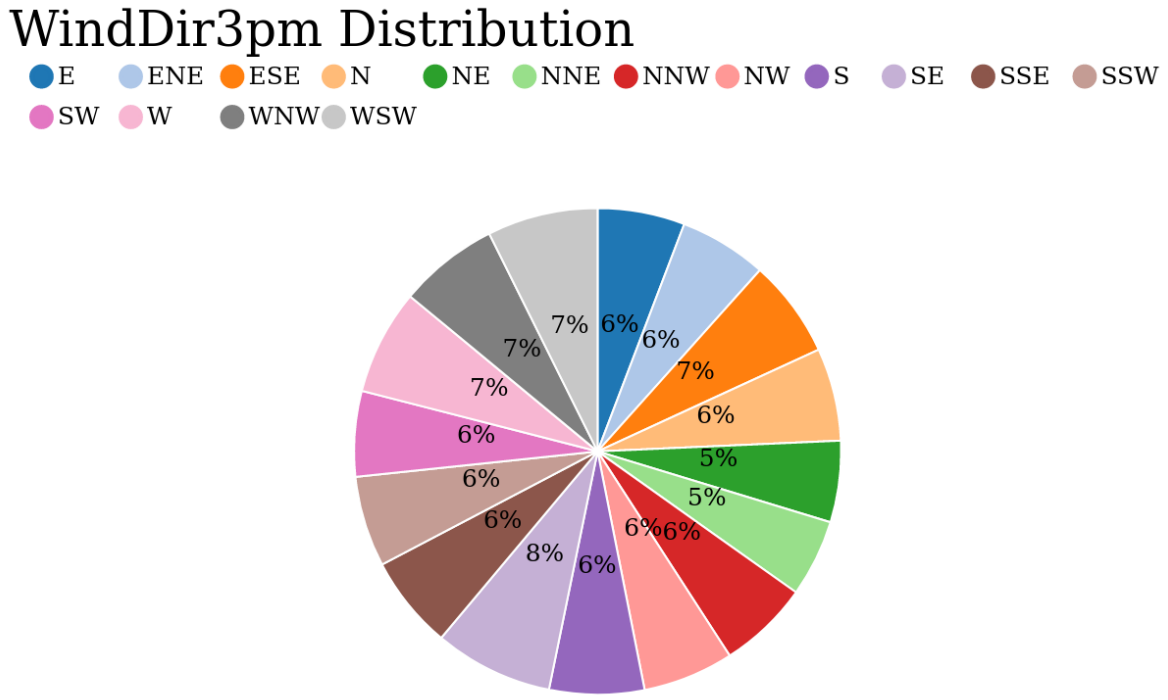


Bar Chart

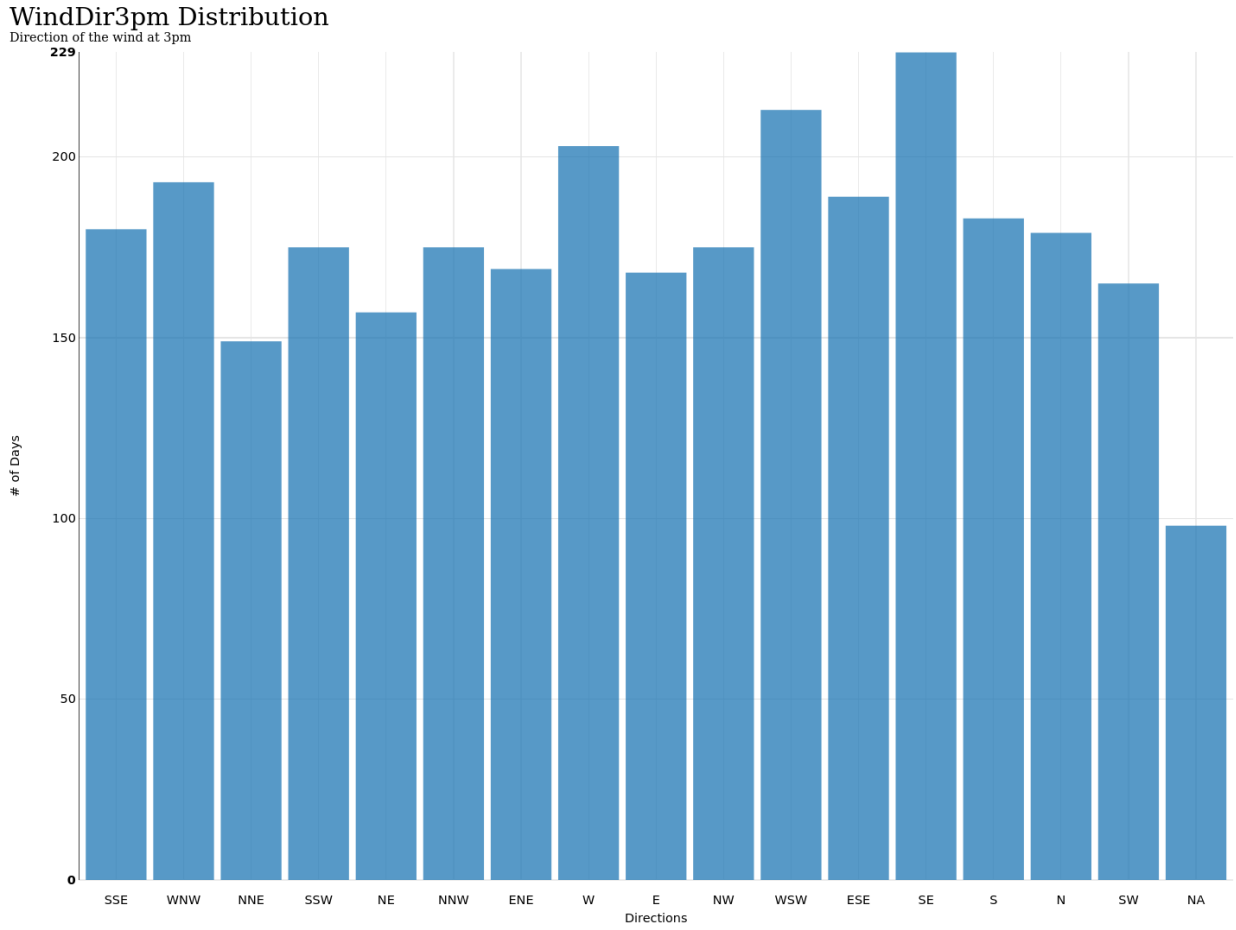


1A.2.10 WindDir3pm

Pie Chart



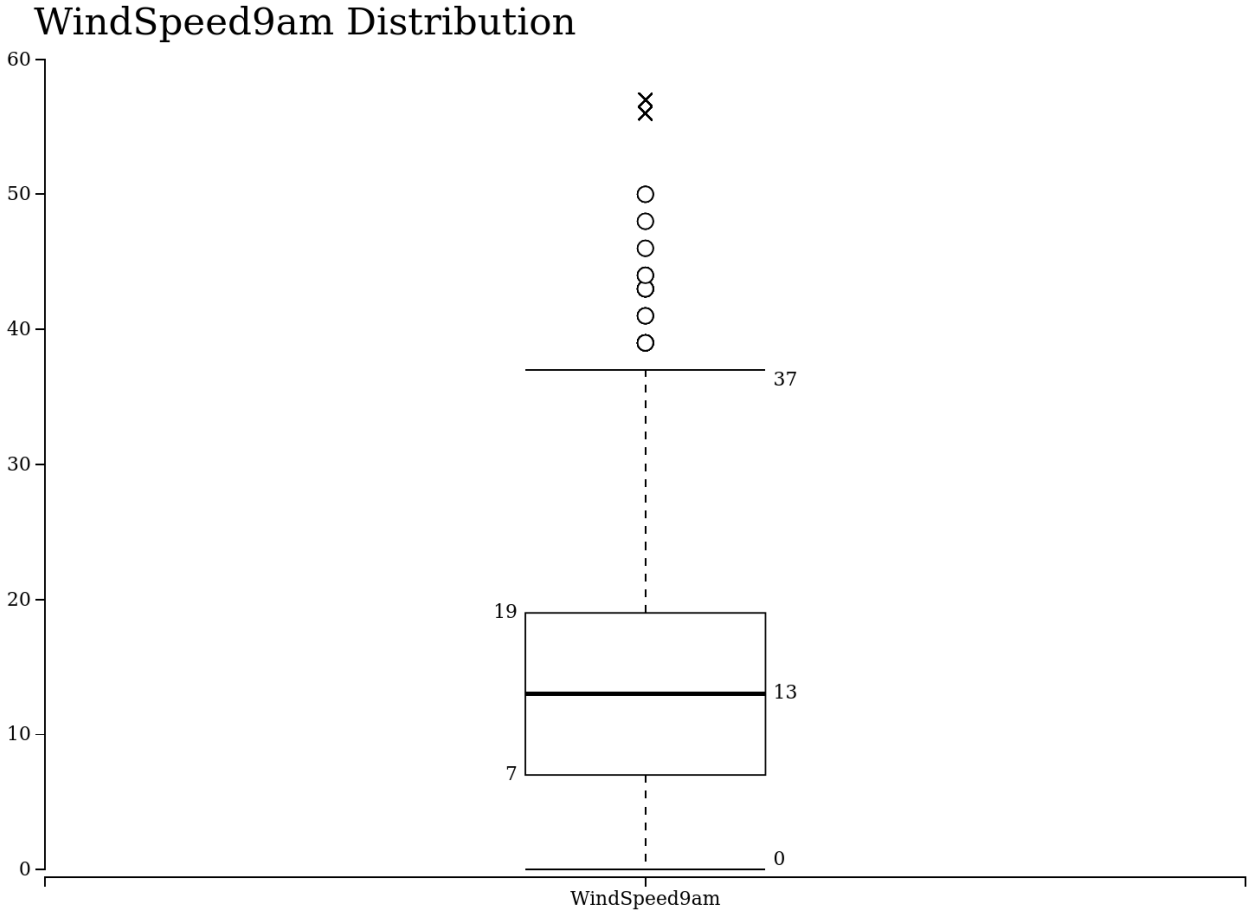
Bar Chart



1A.2.11 WindSpeed9am

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
0	57	14.144	13	9	77.213	8.787	0.702

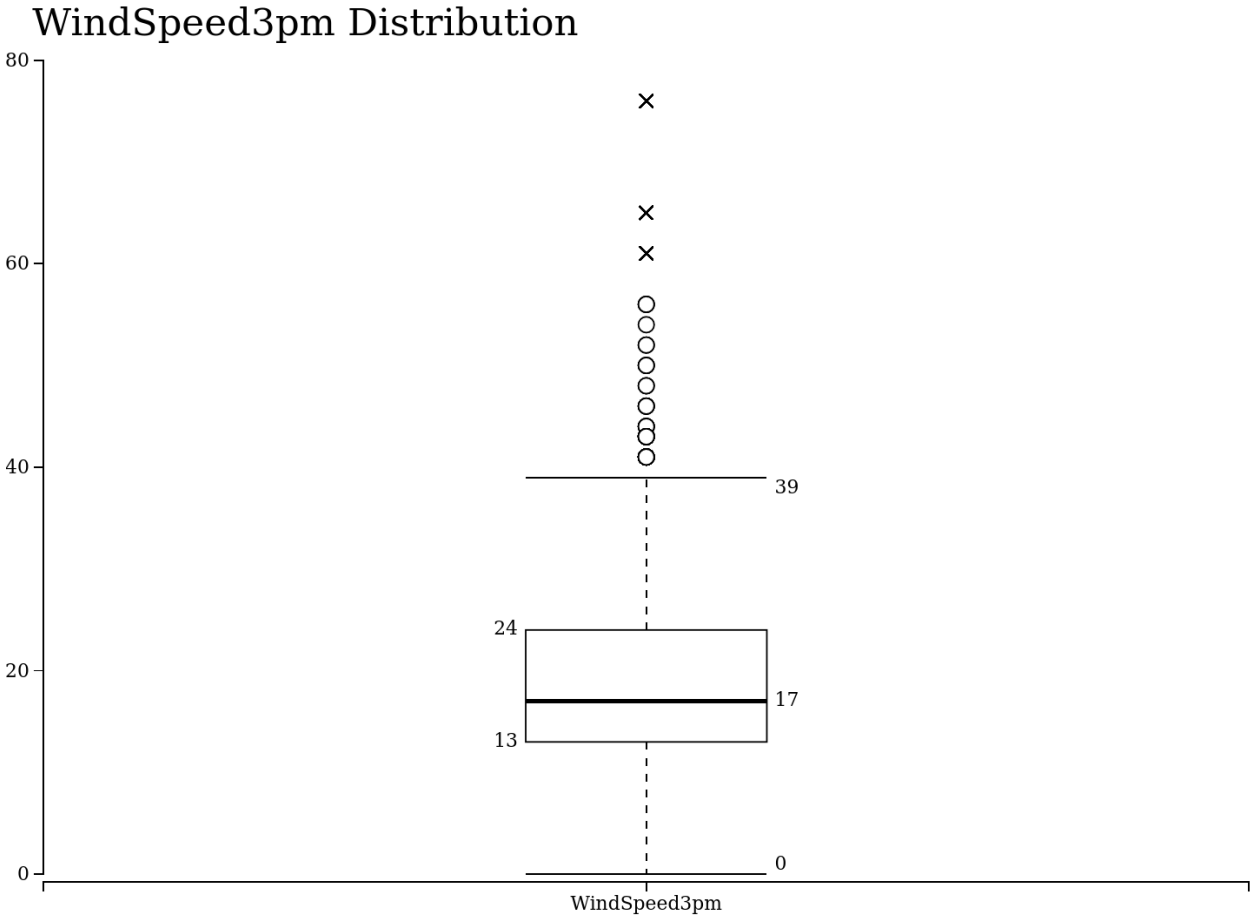
Box Plot



1A.2.12 WindSpeed3pm

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
0	76	18.548	17	17	76.633	8.754	0.716

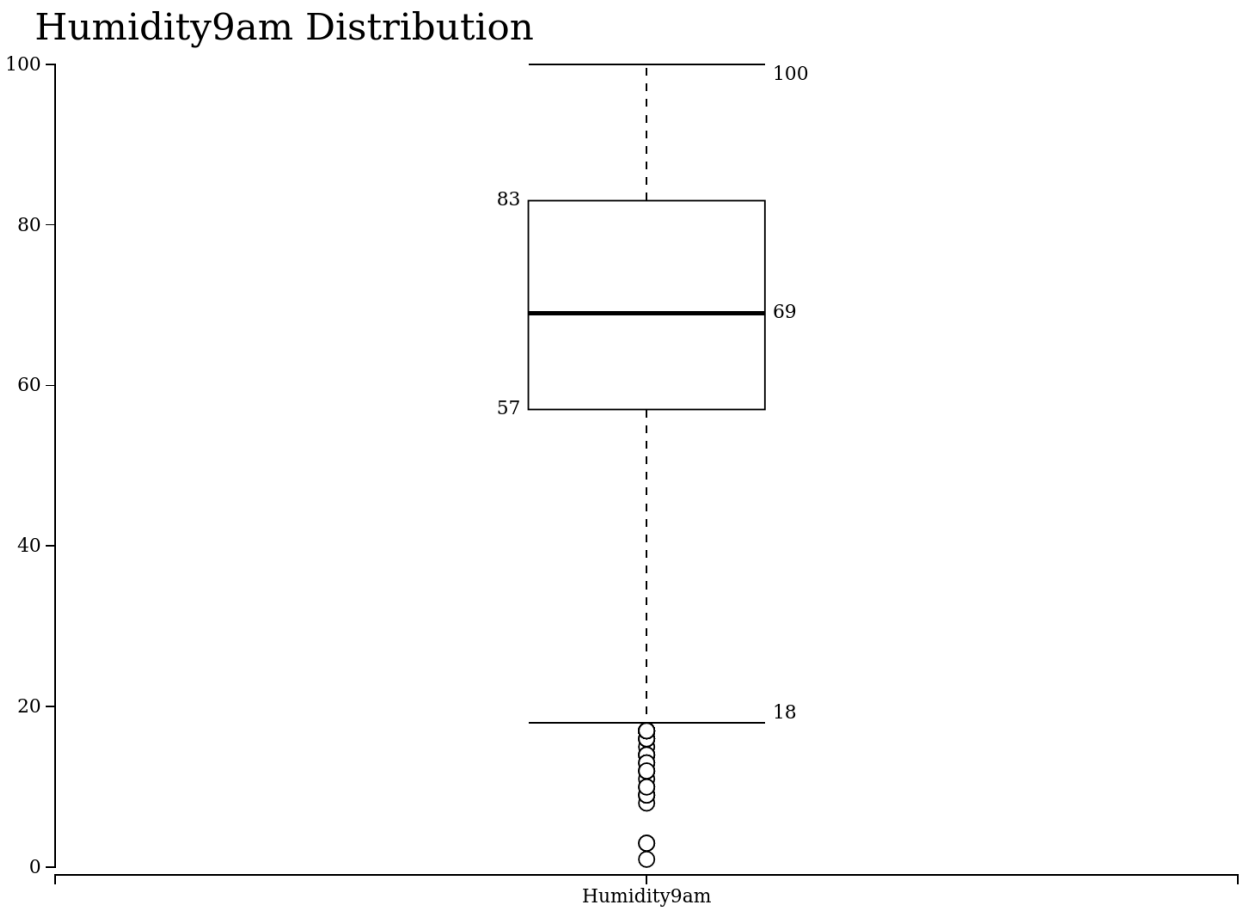
Box Plot



1A.2.13 Humidity9am

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
1	100	68.644	69	70	362.868	19.049	-0.485

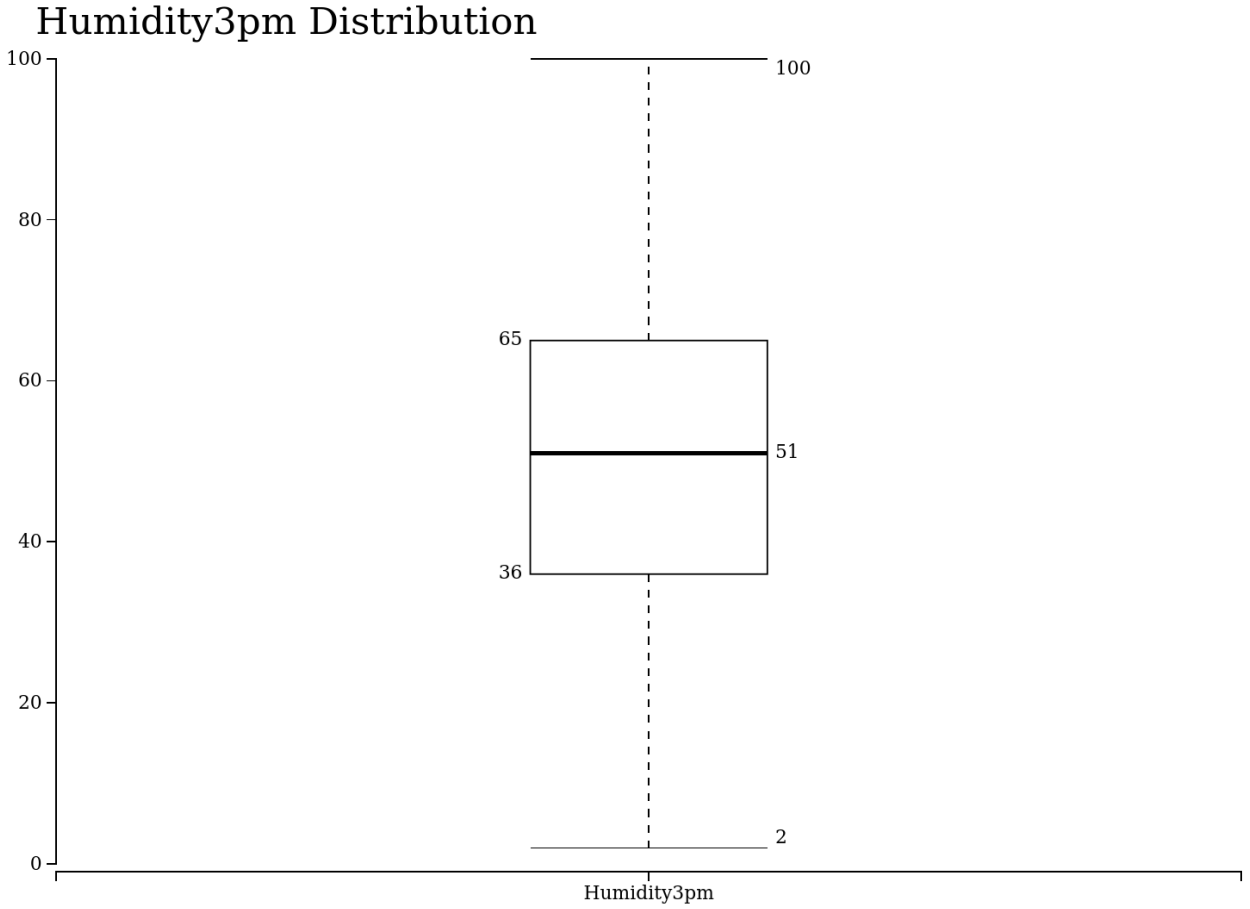
Box Plot



1A.2.14 Humidity3pm

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
2	100	51.278	51	54	432.208	20.79	0.057

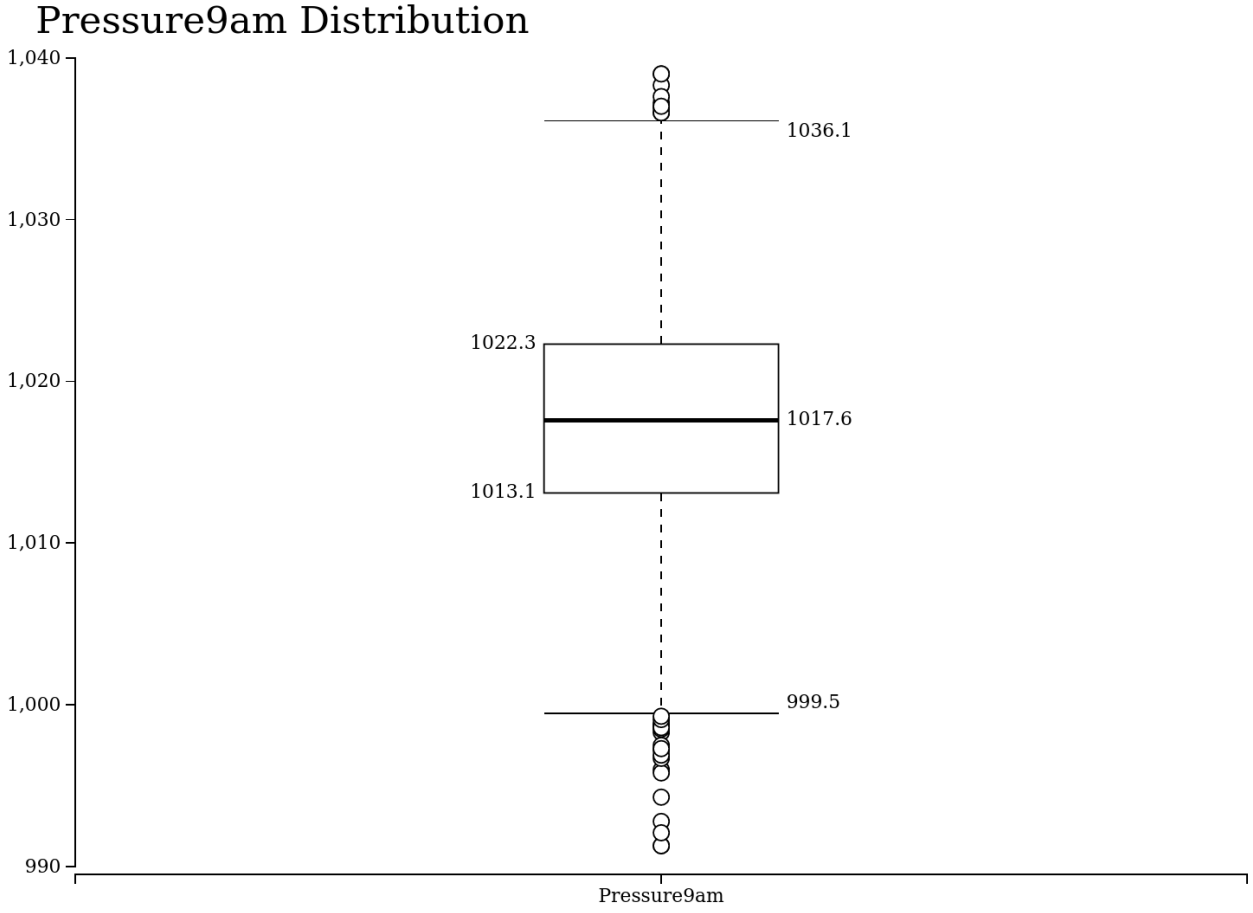
Box Plot



1A.2.15 Pressure9am

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
991.3	1039	1017.76	1017.6	1019.2	49.184	7.013	-0.028

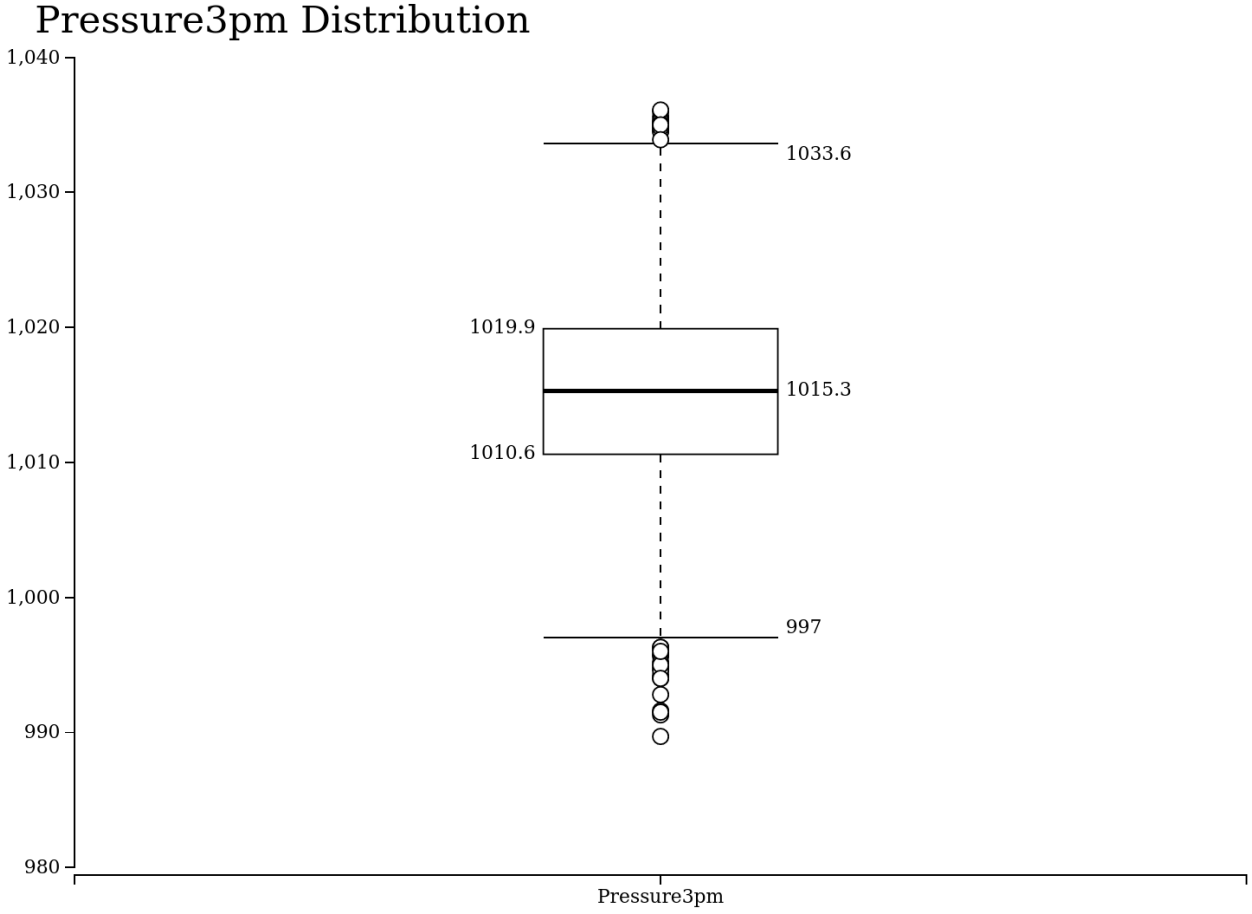
Box Plot



1A.2.16 Pressure3pm

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
989.7	1036.1	1015.342	1015.3	1013.8	48.755	6.982	-0.008

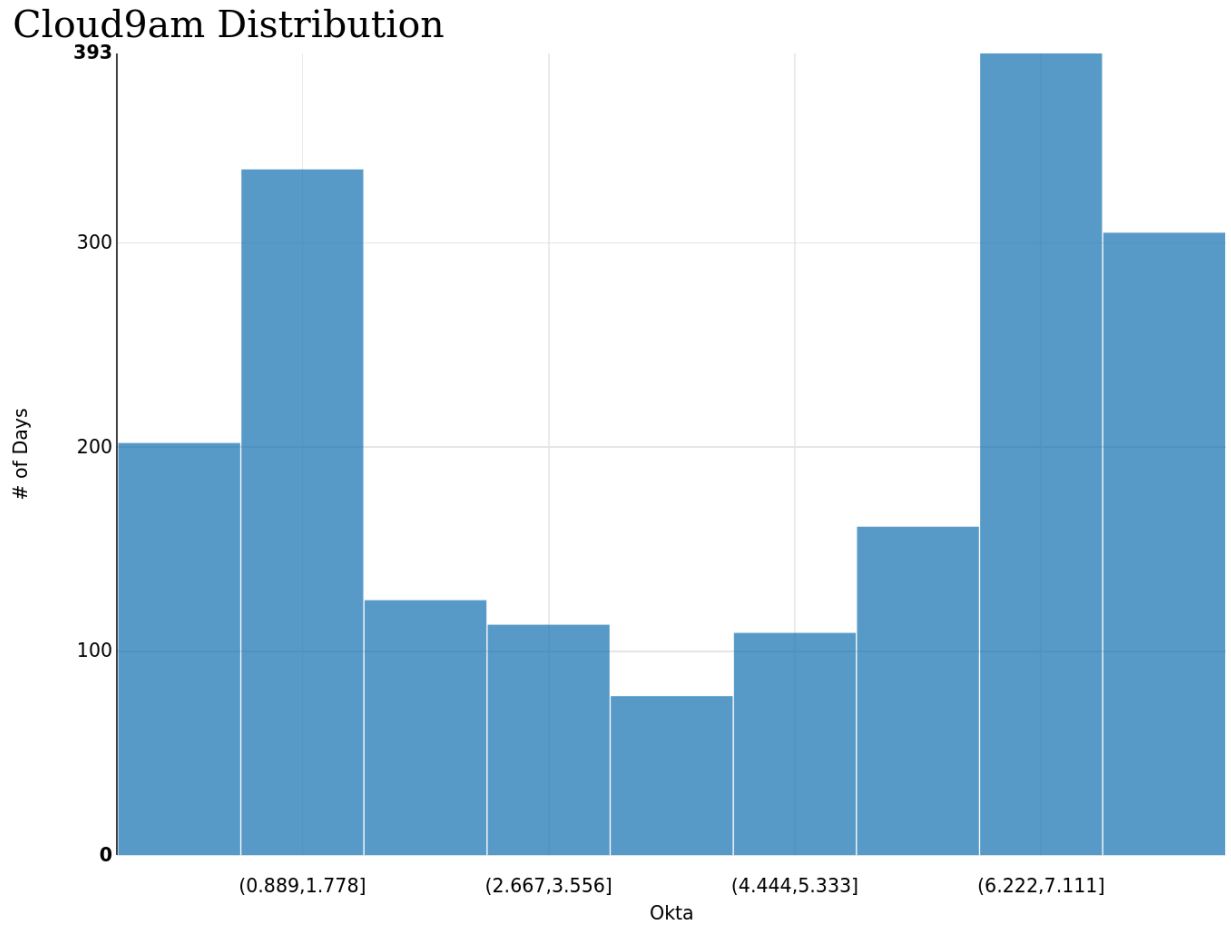
Box Plot



1A.2.17 Cloud9am

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
0	8	4.357	5	7	8.68	2.946	-0.185

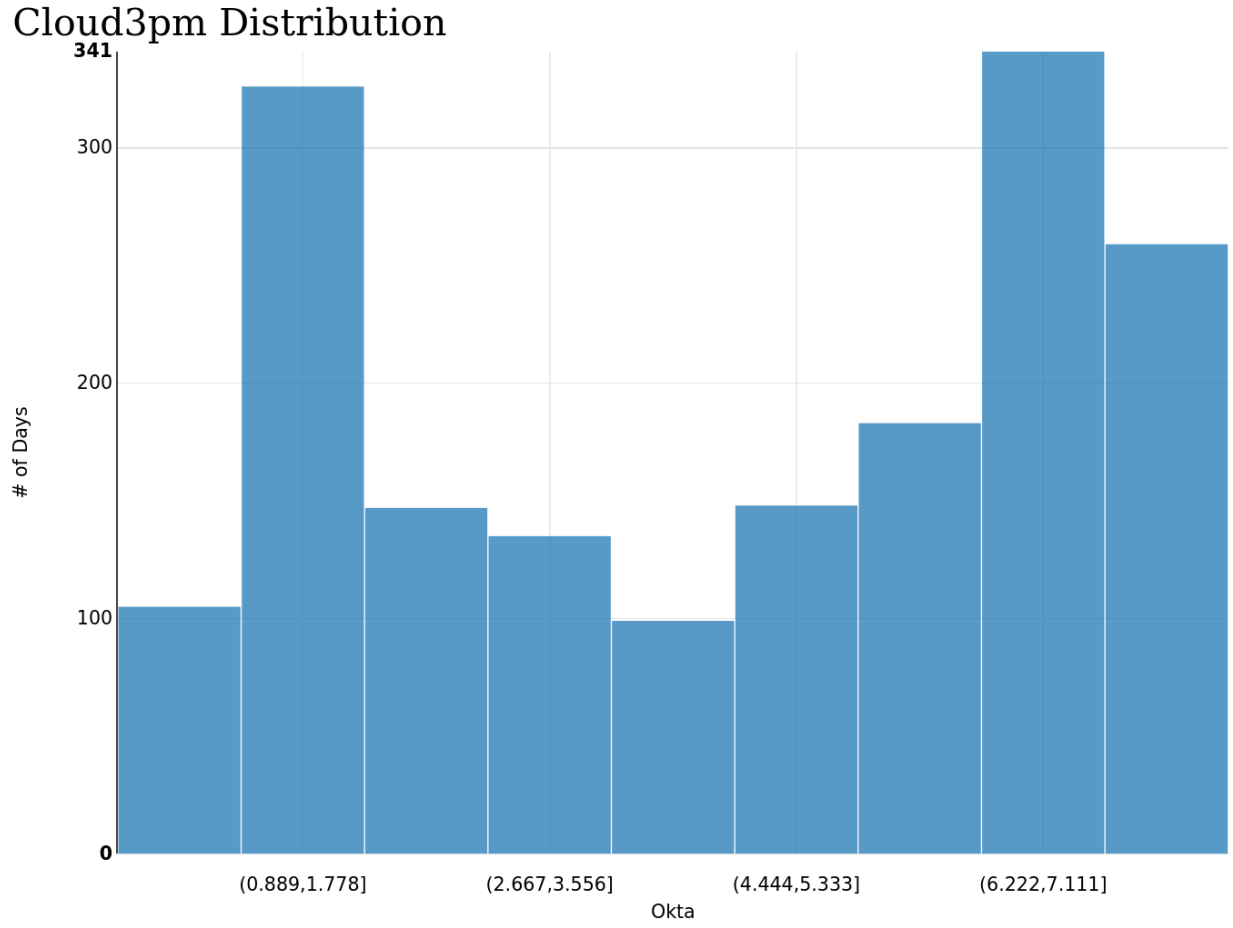
Histogram



1A.2.18 Cloud3pm

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
0	8	4.428	5	7	7.526	2.743	-0.178

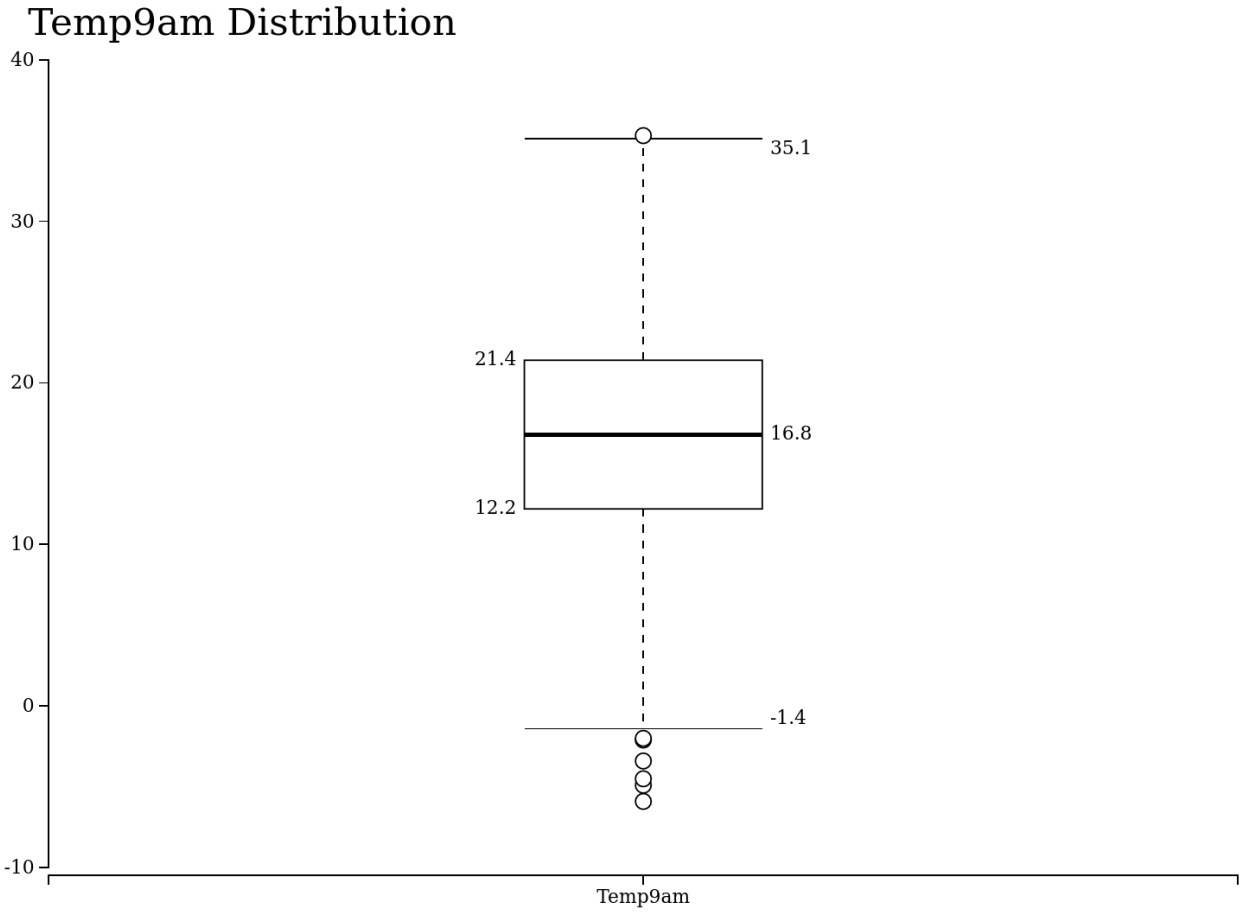
Histogram



1A.2.19 Temp9am

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
-5.9	35.3	16.96	16.8	18.5	42.402	6.512	0.056

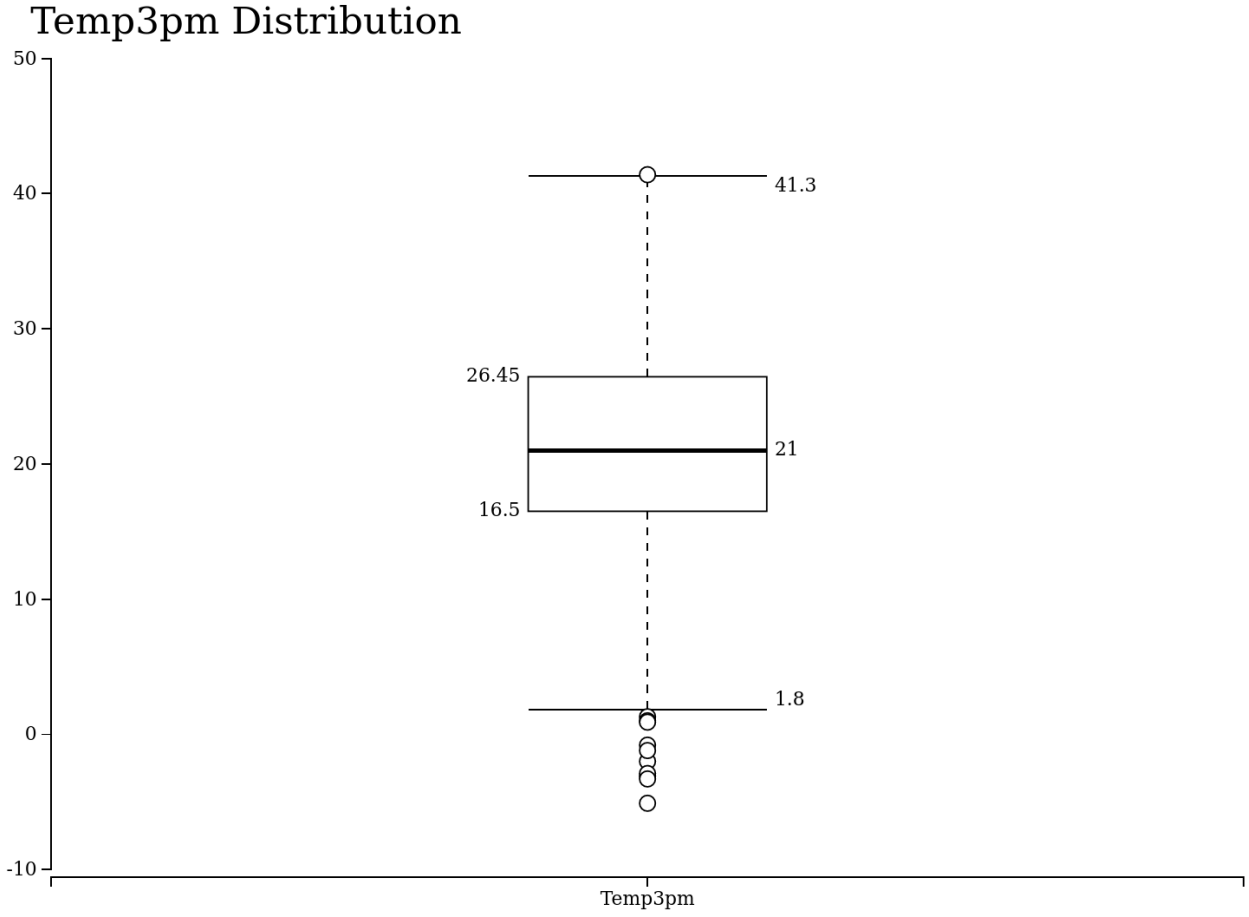
Box Plot



1A.2.20 Temp3pm

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
-5.1	41.4	21.657	21	18.5	48.406	6.957	0.158

Box Plot

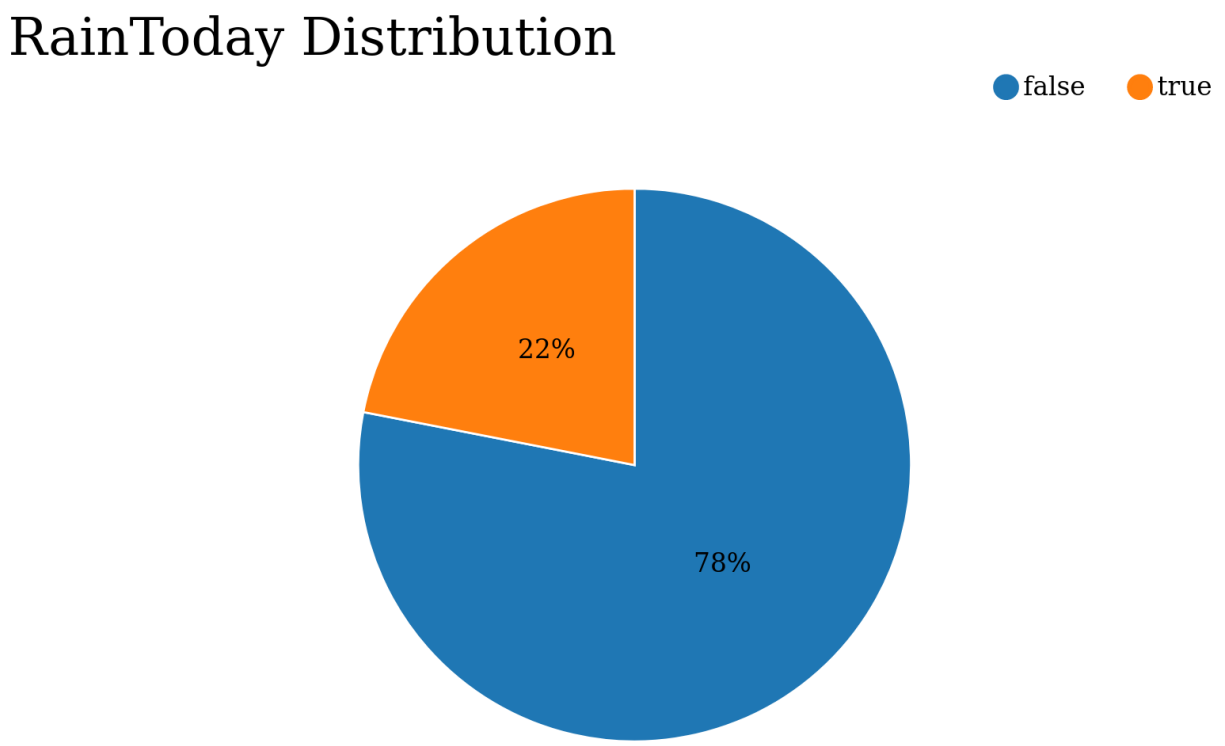


1A.2.21 RainToday

False = 0
True = 1

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
0	1	0.219	0	0	0.171	0.414	1.357

Pie Chart

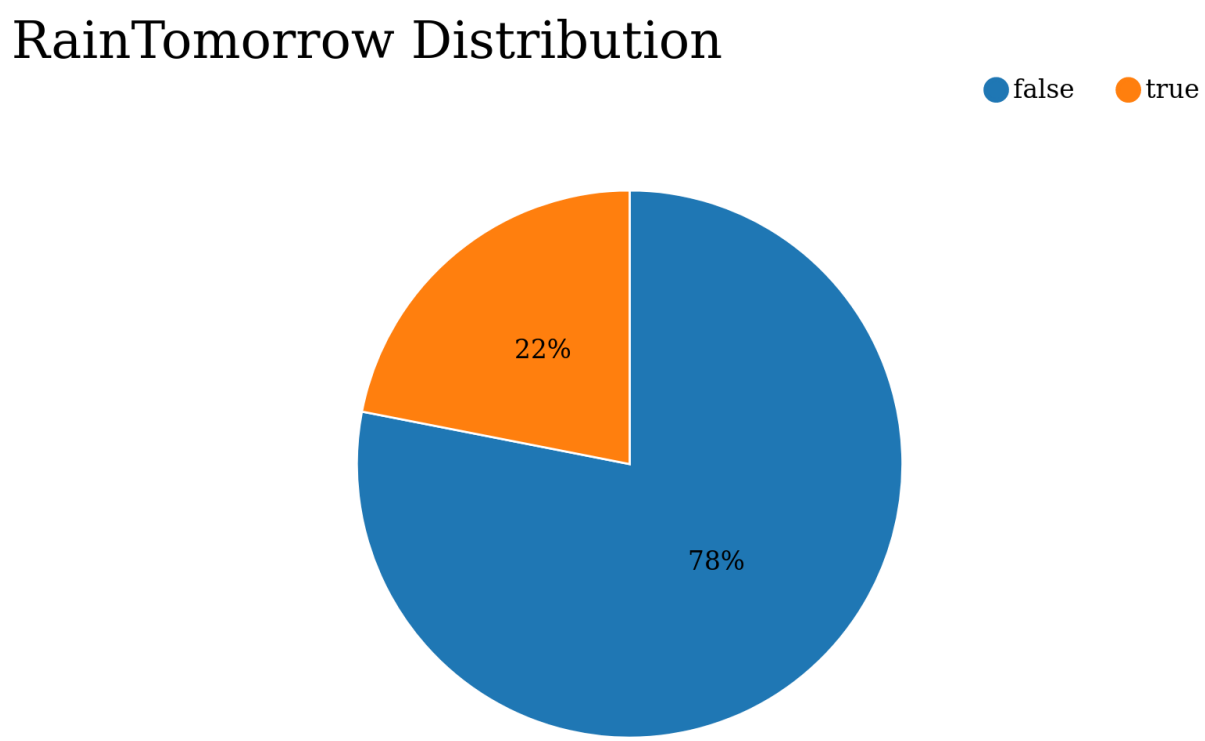


1A.2.22 RainTomorrow

False = 0
True = 1

Min	Max	Mean	Median	Mode	Variance	Std. Deviation	Skewness
0	1	0.215	0	0	0.169	0.411	1.386

Pie Chart



1A.3. Explore your dataset

1A.3.1 Clusters and correlated data

As the attributes are all different factors related to weather, there are many instances where the data is correlated in some fashion. For example the relationship between hours of sunshine per day and the amount of "Class A" pan evaporation in the 24 hours to 9am. This is shown in Fig ##

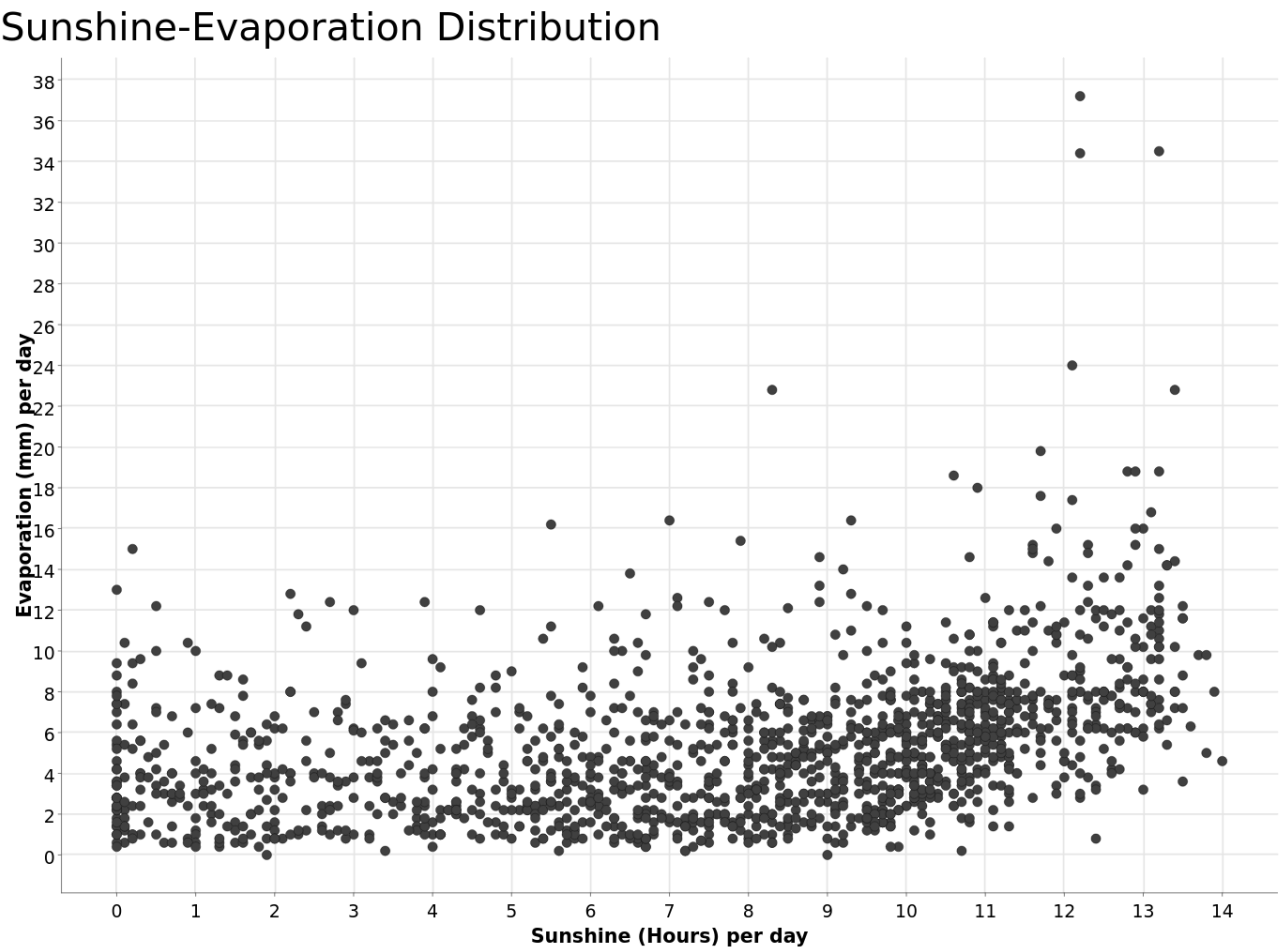
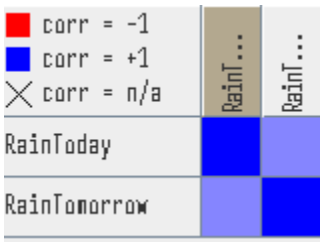


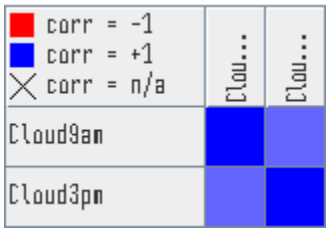
Fig ## of Sunshine-Evaporation scatter plot

This scatter plot depicts a positive relationship between Evaporation and Sunshine. Furthermore a linear correlation analysis states a correlation value of 0.3761. Also note the outliers in the top right of the graph far removed from the rest of the data cluster.

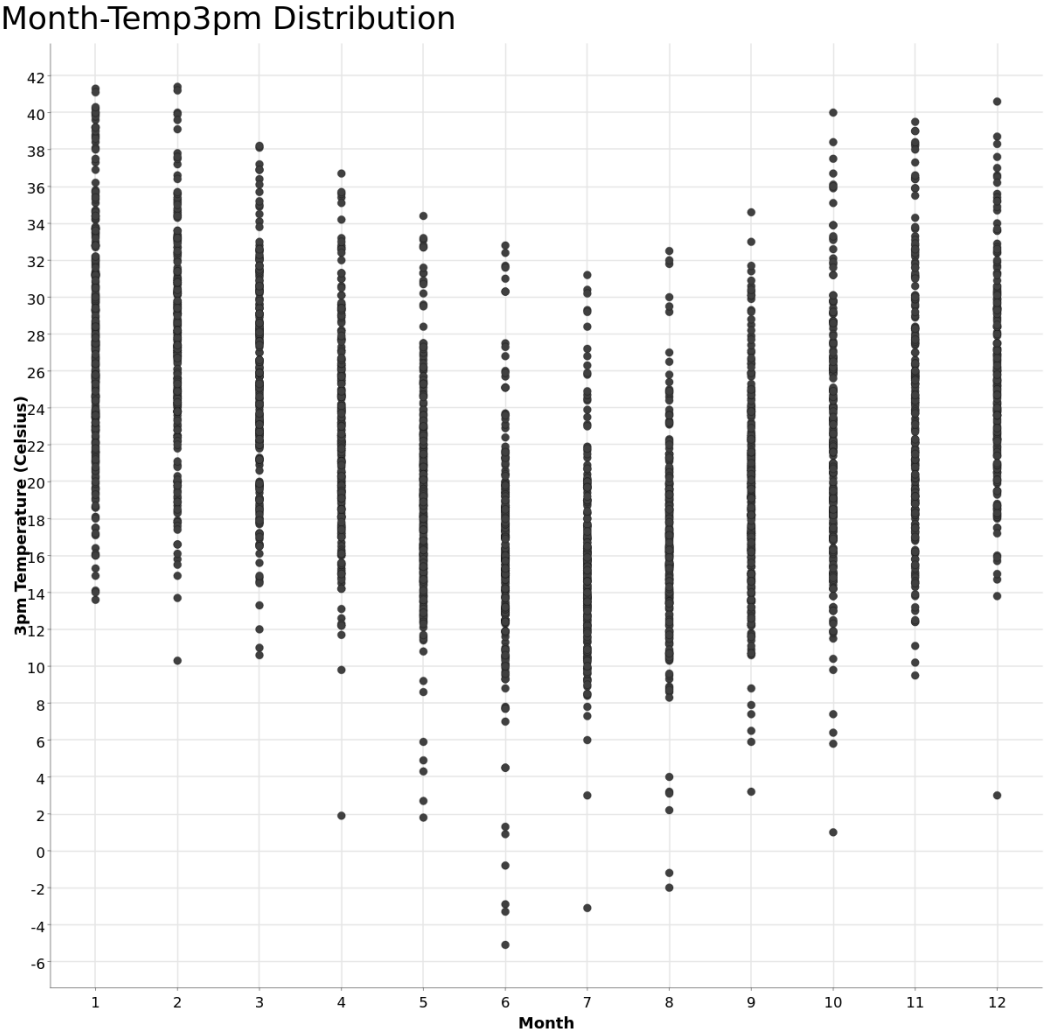
Moreover, using the *Linear Correlation* module in Knime also calculates a correlation value of 0.477 between RainToday and RainTomorrow showing a much stronger likelihood to rain in clusters over a period of several days, rather than being spread out throughout the year.



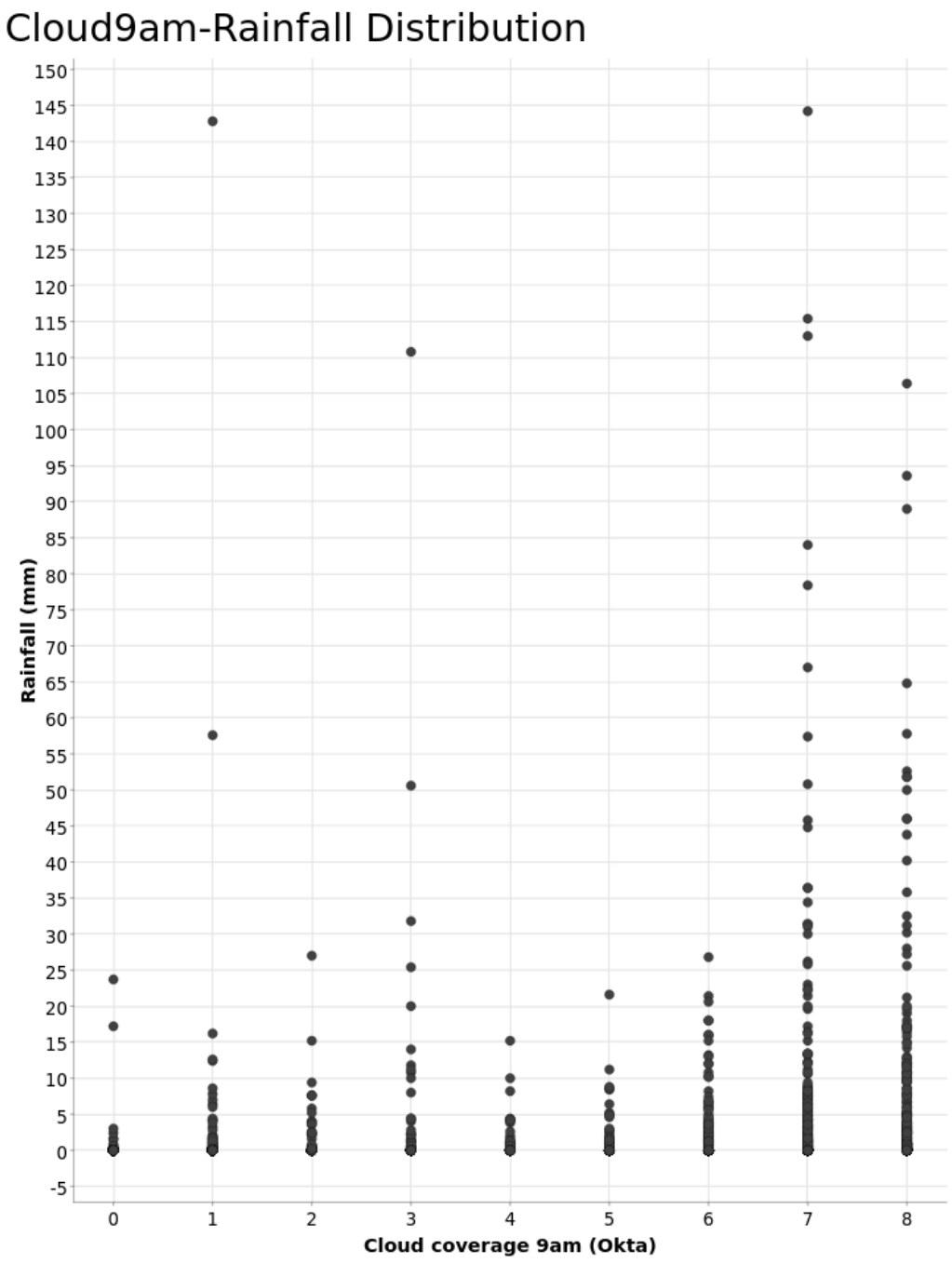
Similarly, Cloud9am and Cloud3pm have a very strong correlation value of 0.608 which means cloud coverage is more frequently consistent throughout the day rather than changing from the morning (9am) to the afternoon (3pm).



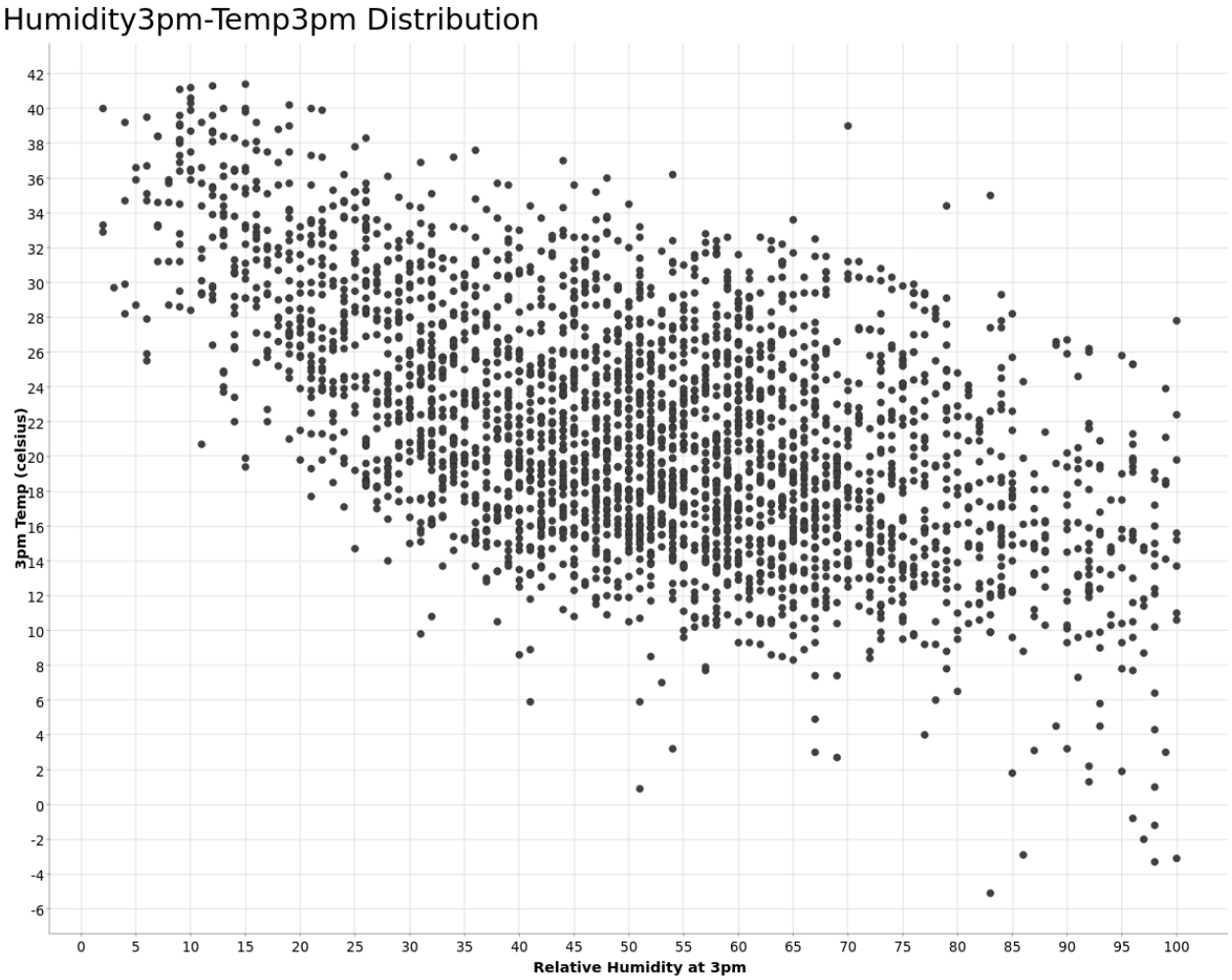
Here is a scatter plot portraying the relationship between the month and temperature using 3pm temperature



The data very strongly correlates to seasonal changes which, as the data was collected in the southern hemisphere, shows a decline in temperature during the day in the Winter months and a steady increase as the months transition to Summer.

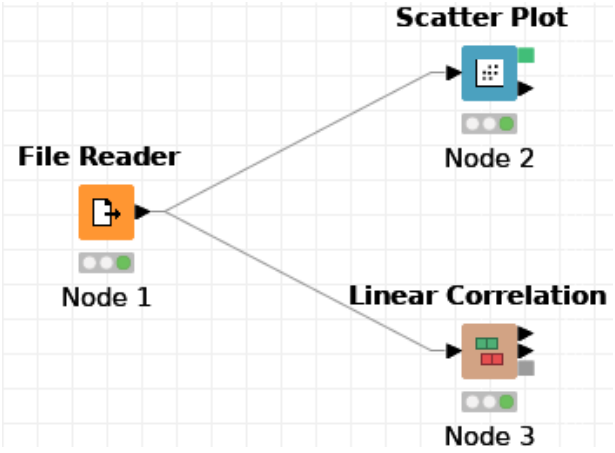


Scatter plot depicting a positive correlation between 9am cloud coverage and rainfall. Some outliers can also be seen such as the 144.2mm of rainfall with only 1 okta cloud coverage.



Scatter plot depicting a negative correlation between relative humidity at 3pm and 3pm temperature.

Charts have been made using Knime *Scatter Plot* module (for scatter plots) and *Linear Correlation* module (to calculate linear correlation value)



1A.3.2 Outliers

Some sections of the dataset also have a large amount of outliers. A prominent example is the Rainfall attribute as it measures the amount of rainfall each day with a maximum value of 144.2 but a median value of just 0. Only 22% of the data is tagged as having rained that day, only a portion of the data is actually measuring the amount of rainfall per day of rain. This makes the days of heavy rainfall especially problematic which can be seen in 1A.2.4 Box plot. Even after adjusting the data and only sampling rainfall where *RainToday* is true, heavy rain is still depicted as an outlier.

Both 9am and 3pm wind speed also has roughly a dozen outliers in their respective box plots. 9am wind speed has a median of 13, standard deviation of 8.787 but a max value of 57. Similarly, 3pm wind speed has a median 17, standard deviation of 8.754 and a max value of 76.

Finally evaporation has a positive skew with high end clusters as depicted in the box plot in 1A.2.5. Evaporation attribute has a median of 4.8, standard deviation of 3.986 and a max value of 55.8 with many other outliers hovering around high teens and low 20s. This has led to a skewness value of 2.876.

1B. Data Preprocessing

1B.1. Binning

1B.1.1. Equi-width binning

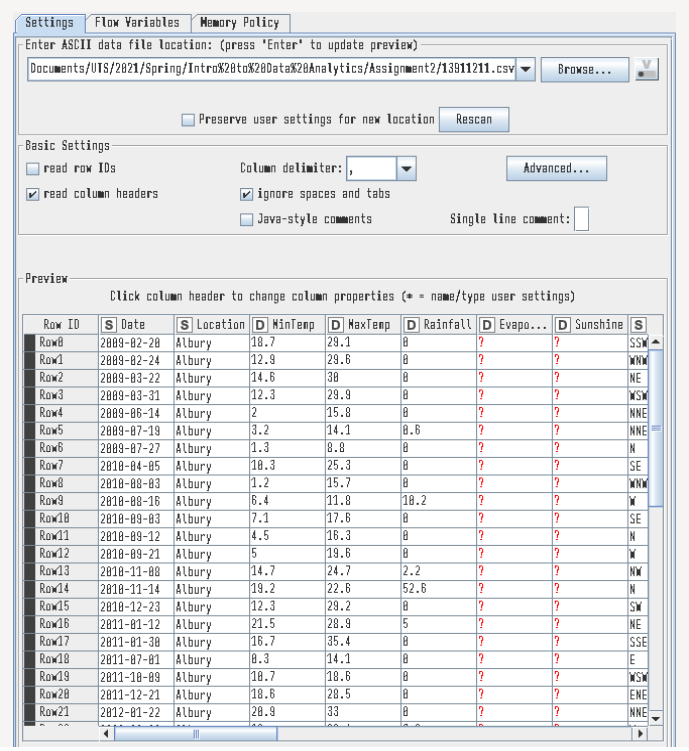
For equi-width binning i decided to use 12 bins with a width ~12

I chose this since the data is so sparse that, while more bins will help split up the positively skewed data, it will also split up what little data is in the remaining bins. Furthermore, reducing the number of bins will put too much data into smaller bins that removes visible trends.

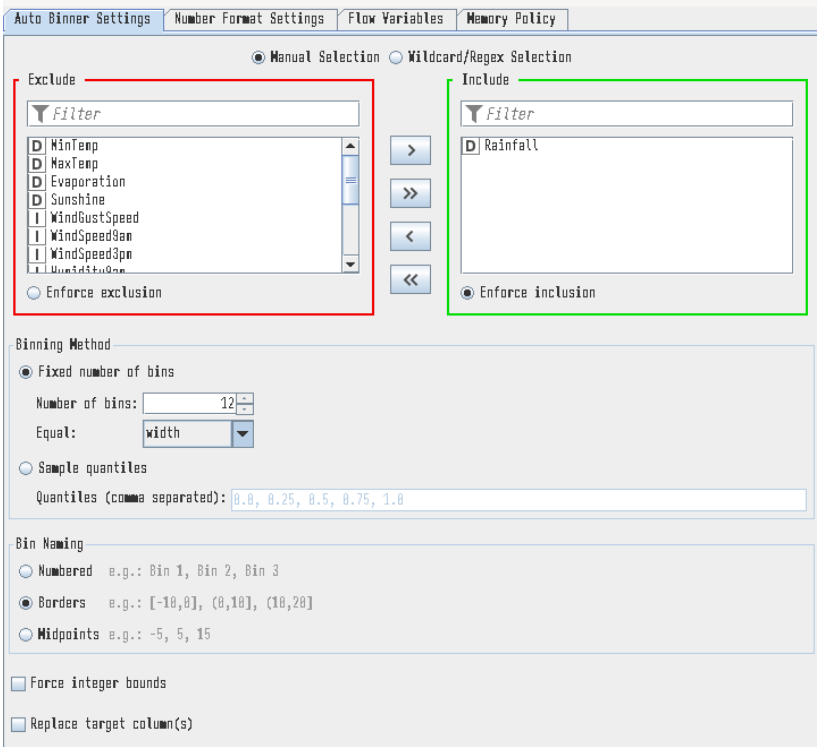
A nice side effect of having 12 bins is also with width 12 translates to 0.5mm of rainfall per hour on average for each bin since it is measured in 24 hours

I used Knime desktop to process the data:

1. Open the file in *File Reader* module



2. Connect *File Reader* to *Auto-Binner*



2.1. Select *Rainfall* column

2.2. Set *Fixed number of bins*

2.3. Set to 12 bins with equal width

2.4. Set *Bin Naming* to *Borders*

2.5. Untick *Replace target Column* to preserve the original data

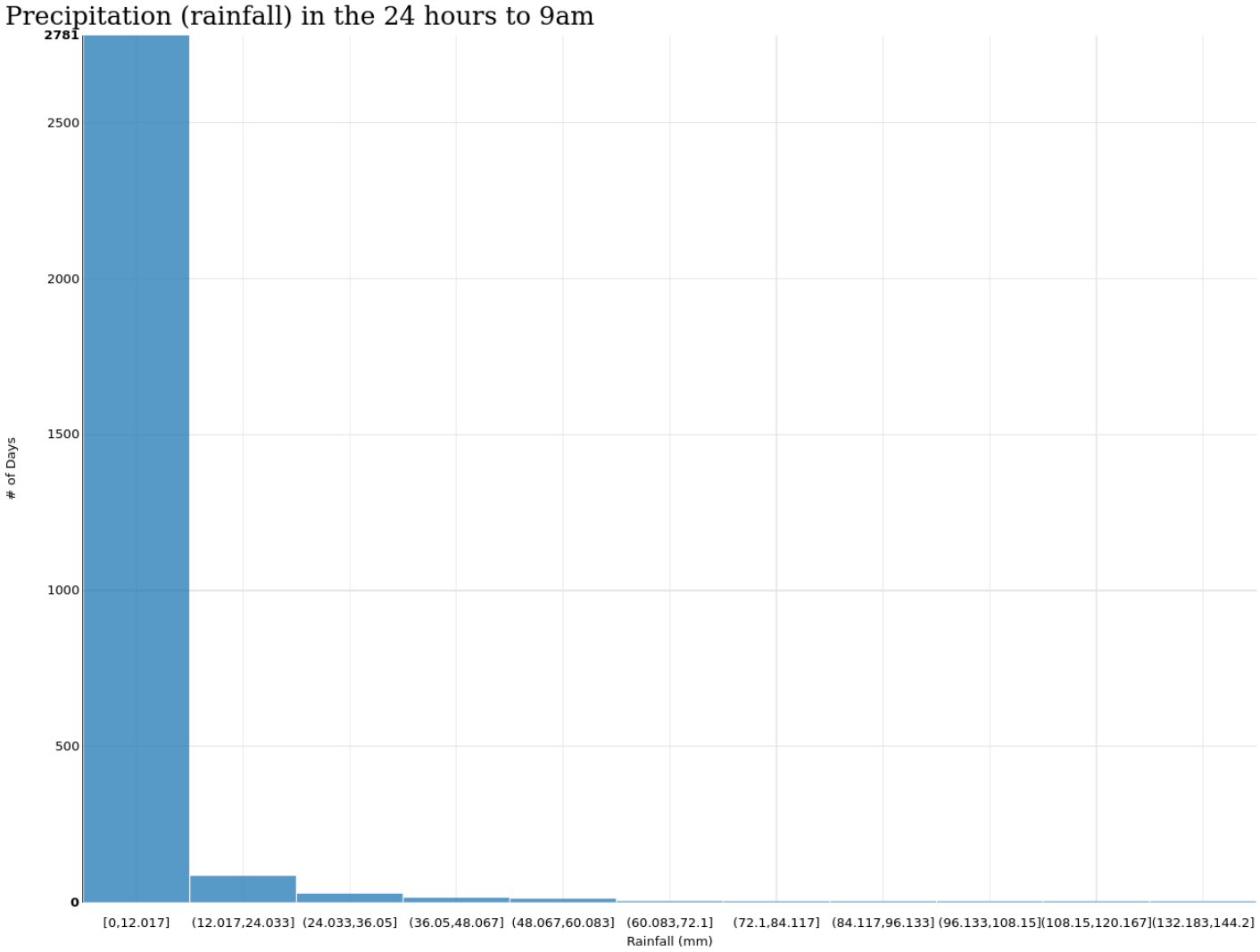


Fig. ## Histogram for 12 equi-width binned data to visualise the distribution

1B.1.2. Equi-depth binning

For equi-depth binning i decided to use 7 bins

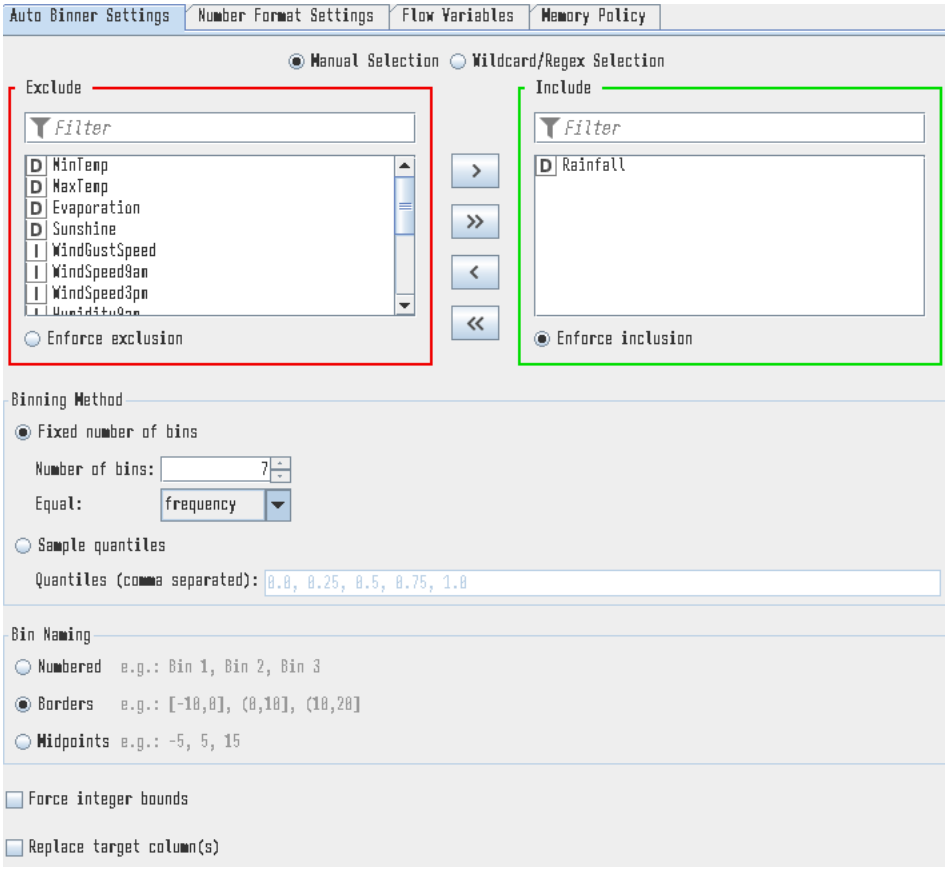
There are 3000 rows in the data table, however 65 rows are missing data (NA) and 1894 are 0 which means they will all be in one bin and cannot be separated. This leaves 1041 values left which will be split into 6 bins with ~173 occurrences, plus the first bin with the 1894 occurrences of 0.

Less than 7 bins will cause too much smoothing of the data on the high end and more bins will remove almost all smoothing on the low end as a vast majority of the data is between 0-1 and have too few occurrences per bin in relation to the first bin.

For equi-depth binning, I used the same setup for equi-width with some modified settings for auto-binner.

Overview

Auto-Binner settings



- Select *Rainfall* column

- Select *Fixed number of bins*

- Set to 7 bins with *equal frequency*

- Set *Bin Naming* to *Borders*

- Untick *Replace target Column* to preserve the original data

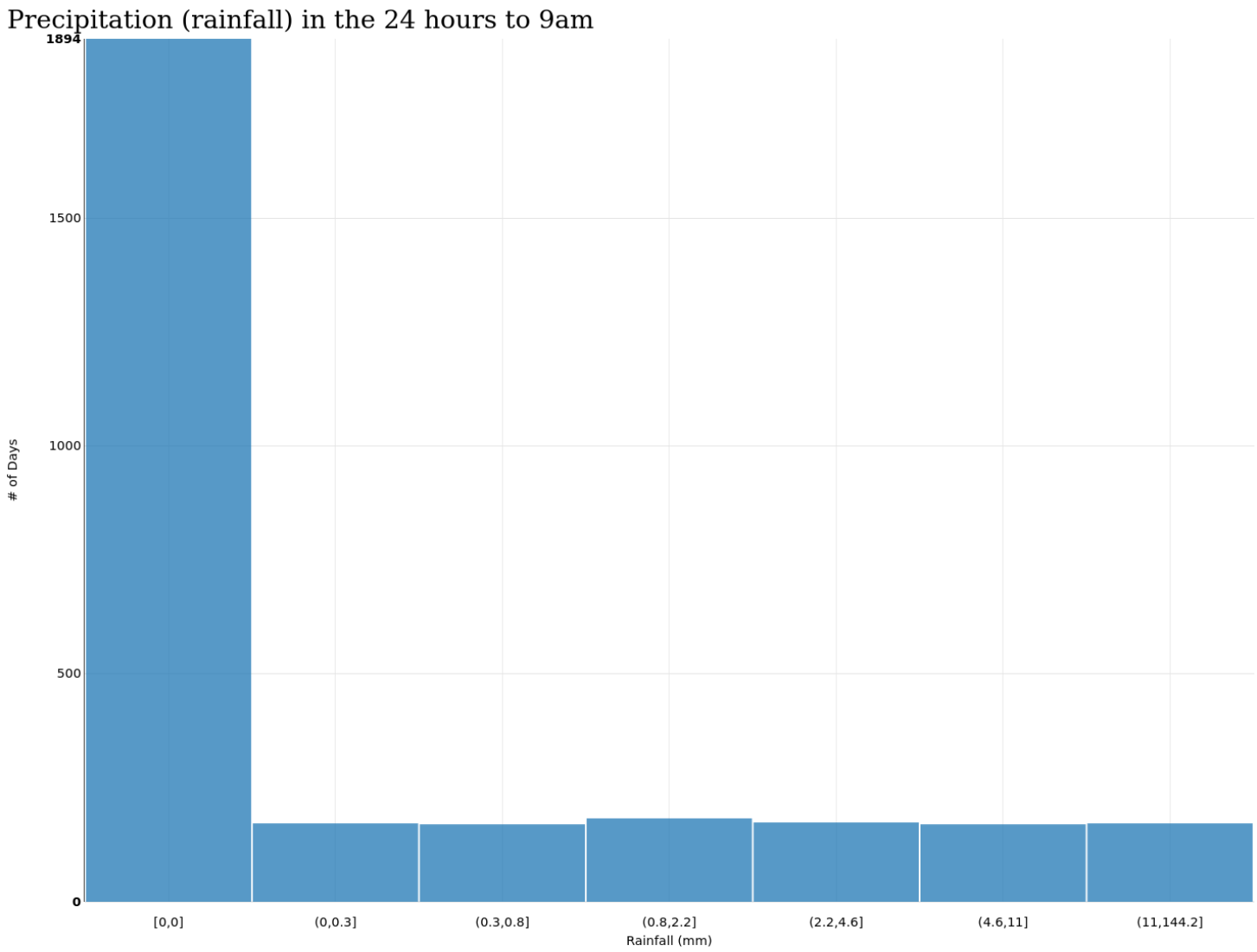


Fig. ## Histogram for 7 equi-depth binned data to visualise the distribution

1B.2. Normalisation

1B.2.1. Min-Max Normalisation

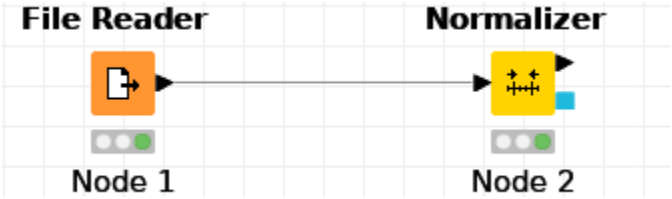
Normalisation is the process of taking values spanning a specific range, and representing them in a new, specified range.

Min-Max normalisation uses the formula shown in fig. ##

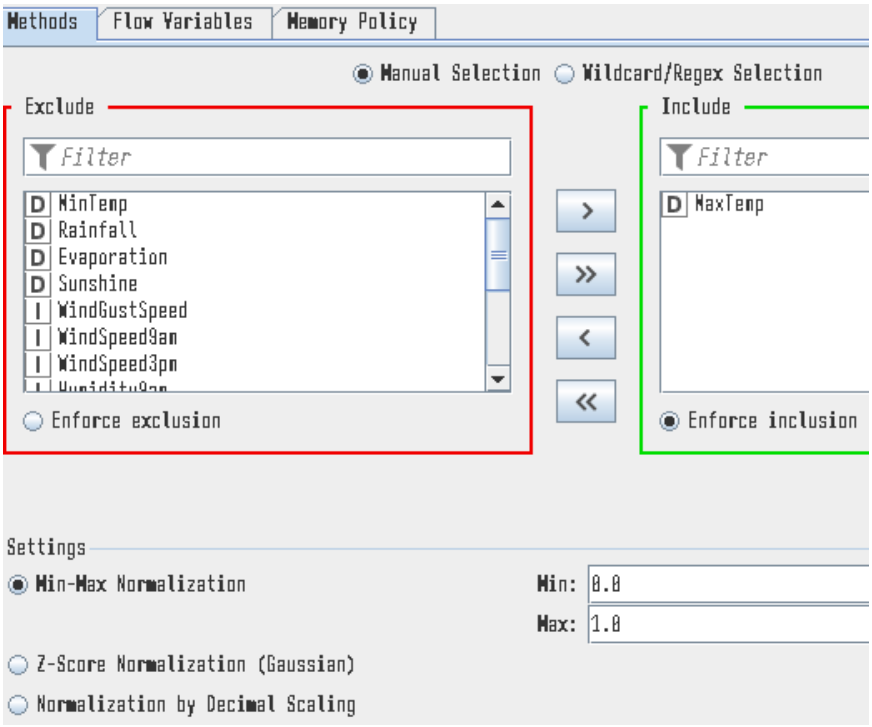
$$A' = \frac{A - Min}{Max - Min}(newMax - newMin) + newMin$$

Fig. ##

I used Knime to min-max normalise MaxTemp



Normalizer settings



- Include *MaxTemp*

- Tick *Enforce inclusion*

- Set *Min-Max Normalization* to Min=0 Max=1

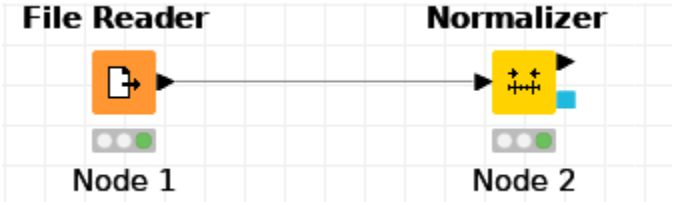
1B.2.2. Z-Score Normalisation

Z-Score normalisation is another formula used to normalise a data set. It uses the formula shown in fig. ##

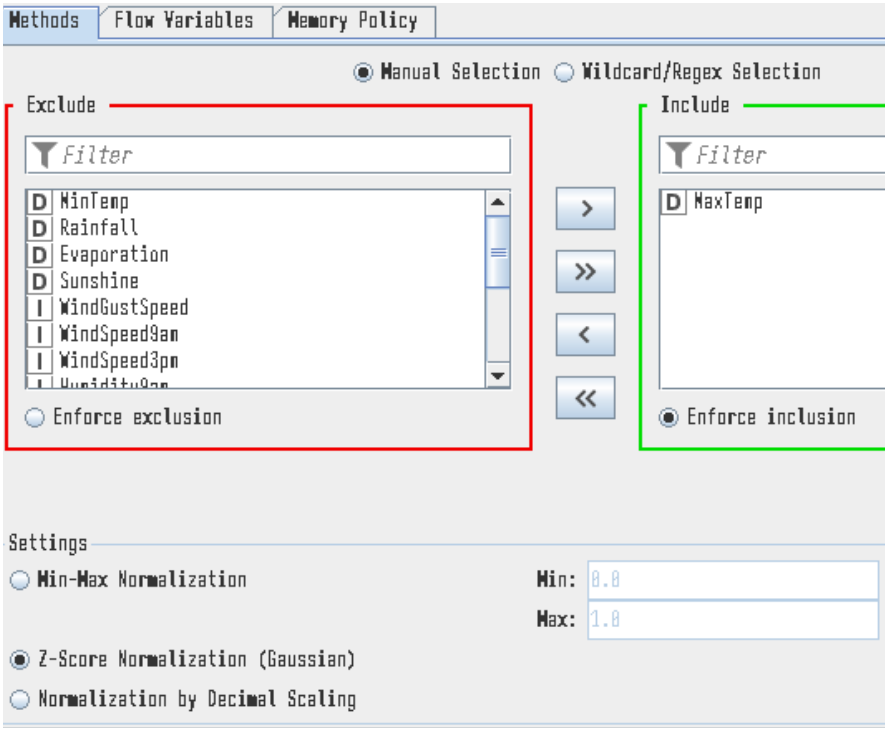
$$A' = \frac{A - Mean}{StandardDeviation}$$

Fig. ##

I used the same Knime setup for Z-Score normalisation with a modified *Normlizer* module



Normalizer settings



- Include *MaxTemp*

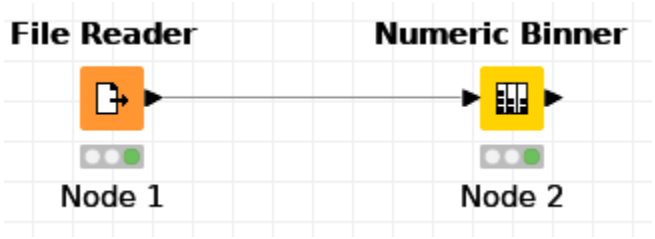
- Tick *Enforce inclusion*

- Set *Z-Score Normalization (Gaussian)*

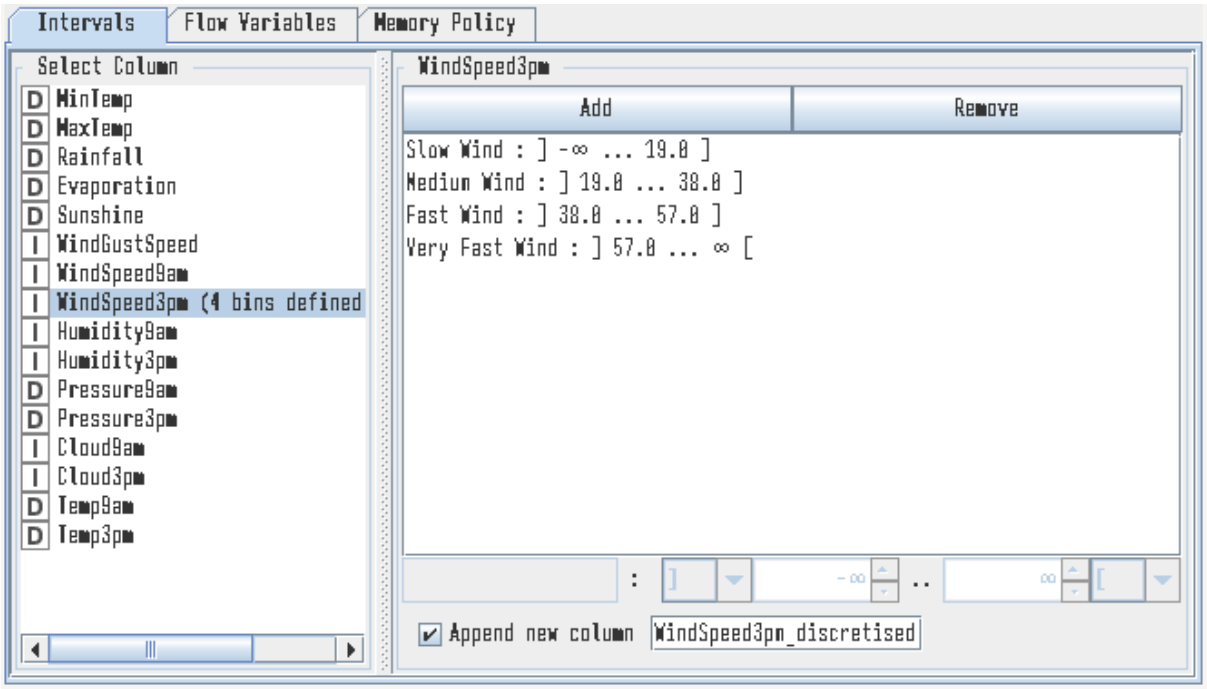
1B.3. Discretisation

Discretisation is used to divide a continuous data set into distinct categories

This can be done in Knime using the *Numeric Binner* module.



Numeric Binner settings

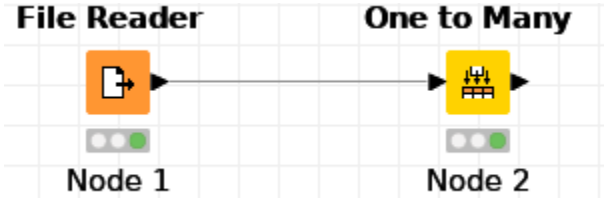


Four categories have been created for WinSpeed3pm each equally spaced by 19 from min (0) to max (76).

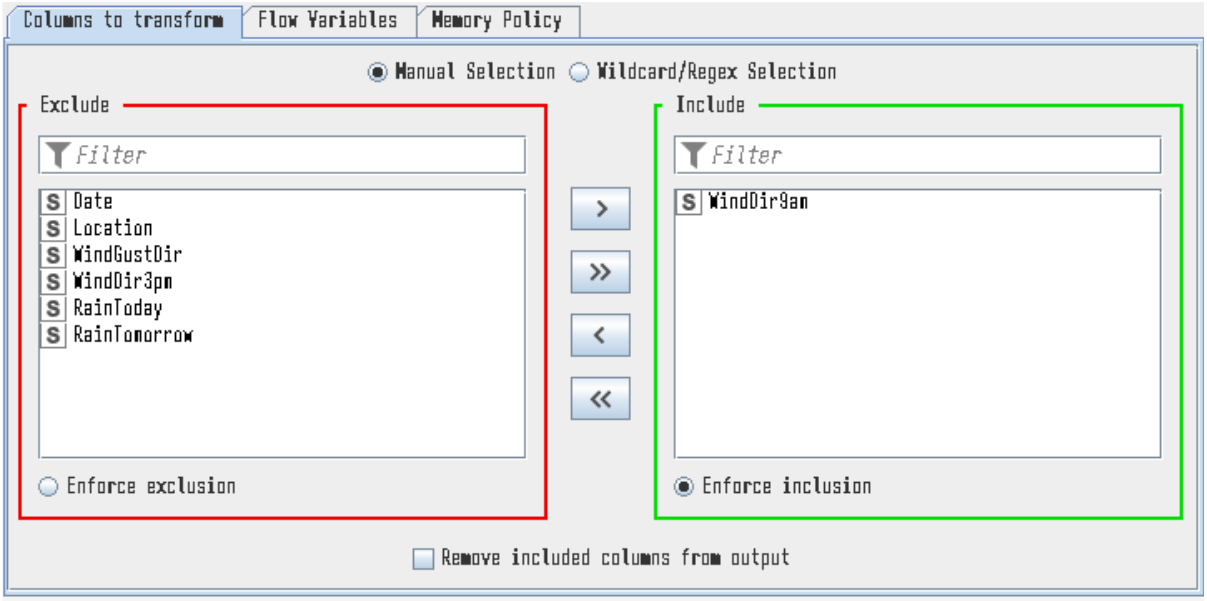
1B.4. Binarisation

Binarisation creates a binary representation of data by converting an attribute type with set categories into separate binary attributes for each category.

This can be done in Knime using the *One to Many* module.



One to Many settings



- Include *WindDir9am*
- Set *Enforce inclusion*
- Untick *Remove included columns from output* to preserve original data

1C. Summary

- Rainfall was the most sparse attribute with almost 2 thirds (1894/2935) being 0 and a majority of the remaining data still being below 0 with a max value of 144.2. This made it very hard to productively analyse. More preprocessing should be done to accurately examine the nature of rainfall in the data set to accommodate this and should be researched further.

- Although the data had little bias in the months sampled, there were biases in the years with only 46 samples for 2008 while up to 419 samples in 2014. There potentially may have been events in one of the sampled years which could skew the data such as a meteorological event. More sampling over more years will help to improve this and such data could be used to scrutinize the current dataset to probe for more accurate results.

- *RainToday* and *RainTomorrow* have a strong linear correlation value of 0.477 which does indicate days with reported rain takes place in clusters rather than in separate events.

- There is a strong correlation between each data's respective 9am morning and 3pm afternoon observations. More data points throughout the day such as 12am midnight could further help improve how data between adjacent days correlates instead of just between the morning and afternoon similar to *RainToday* and *RainTomorrow*.

- Evaporation data is heavily positively skewed with a skewness value of 2.876. While it only has a median of 4.8 and standard deviation of 3.986. The maximum measured value is 55.8 with four more values hovering low 30s/very high 20s and 24 more data points measured between 24 to 15.

- There does not seem to be any correlation between wind speed and any other attribute in the dataset including *WindGusSpeed* and *WindGusDir* (or *WindSpeed9am* and *WindDir9am*, etc) which indicates the wind direction may be truly random or correlated with data which has not been measured.

- Over the entirety of the dataset, it was gauged that 22% *RainToday* was true and therefore rained that day measured from 9am-9am.

- Both 9am and 3pm wind speed has a cluster of a dozen outliers in their datasets. 9am wind speed has a median of 13, standard deviation of 8.787 but a max value of 57. Similarly, 3pm wind speed has a median 17, standard deviation of 8.754 and a max value of 76. More research may be needed in the locations/dates these values have been observed to more accurately understand the relationship between such high wind speed and other weather data.