|  |  |
| --- | --- |
| A picture of a winding road and trees  AI Data Science Project - Statistics  February, 2025 | Abstract  This stage is a statistics compilation of the attributes that are present in the dataset. It has your usual measures of statistics with a correlation matrix, while association mining is also used for the categorical attributes.  Zhitan Wu  UTS Intern |

Table of Contents

[Code For Finding Statistics Measures For The Four US States 2](#_Toc189166042)

[Statistics Measures 3](#_Toc189166043)

[CA 3](#_Toc189166044)

[WA 3](#_Toc189166045)

[MI 3](#_Toc189166046)

[PA 3](#_Toc189166047)

[Correlation Matrix 3](#_Toc189166048)

[CA 3](#_Toc189166049)

[WA 3](#_Toc189166050)

[MI 3](#_Toc189166051)

[PA 3](#_Toc189166052)

[Association Mining 3](#_Toc189166053)

[Sequential Model From Hugging Face 3](#_Toc189166054)

# Code For Finding Statistics Measures For The Four US States

A screenshot of a computer program

AI-generated content may be incorrect.

To set up the environment for proper analysis to be conducted, we alter the display for it to show all the columns. We load in the dataset for us to perform our analysis on, by using the ‘pd.read\_csv’ function that takes a Excel csv file with all of its attributes and data records. We also set the display width to avoid truncation and for some of the numerical values to be represented as 6 decimal places instead of utilising scientific notation. Calculating the statistical metrics appropriately warrants the selection of only numerical attributes from each of the datasets that has variations in terms of their data. Any other numerical attributes which may represent categories, or attributes that have the same value are not considered at this step.

We select these attributes by using the ‘df[[]]’ functions. For specific statistic metrics, we use functions at the end. We then go on to use ‘.describe()’ for a summary of the main measures of spread in each attribute, but this does not cover all of them. We continue by using functions such as ‘.var()’ for variance, ‘.median()’ for median and ‘.mode().iloc[0]’ for arranging the data frame into a series and extracting one element for the mode in each attribute. We also physically calculate the interquartile range by deducing quartile 3 (approximately 75% of the data) and quartile 1 (approximately 25% of the data) and subtracting them from each other. Finally, we print and display them on the console at the end.

# Summary Statistical Measures (Numerical Attributes Only – 3.d.p)

## CA

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute** | IQR | Minimum Value | Maximum Value | Median | Mode | Range | Standard Deviation | Variance | Mean |
| PRCP | 0 | 0 | 6.09 | 0 | 0 | 6.09 | 0.323 | 0.104 | 0.082 |
| AWND | 2.124 | 0 | 17.22 | 4.661 | 4.92 | 17.22 | 1.909 | 3.643 | 4.536 |
| FMTM | 0 | 2 | 9999 | 1635 | 1635 | 9997 | 1269.705 | 1612152.033 | 1850.291 |
| PGTM | 105.131 | 0 | 2359 | 1561.708 | 1507 | 2359 | 231.274 | 53487.868 | 1544.134 |
| TMAX | 11 | 41 | 107 | 68.212 | 65 | 66 | 9.886 | 97.729 | 69.96 |
| TMIN | 11.521 | 19 | 68 | 42.775 | 50 | 49 | 7.120 | 50.701 | 43.019 |
| WDF2 | 82.172 | 10 | 360 | 208.656 | 140 | 350 | 63.668 | 4053.635 | 213.658 |
| WDF5 | 92.964 | 10 | 360 | 225.918 | 180 | 350 | 64.862 | 4207.072 | 226.21 |
| WSF2 | 4.986 | 0.9 | 40 | 14.1 | 14.1 | 39.1 | 4.159 | 17.299 | 13.728 |
| WSF5 | 6.592 | 4.9 | 53.9 | 17.9 | 17 | 49 | 5.605 | 31.412 | 17.93 |

## WA

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute** | IQR | Minimum Value | Maximum Value | Median | Mode | Range | Standard Deviation | Variance | Mean |
| PRCP | 0 | 0 | 1.66 | 0 | 0 | 1.66 | 0.091 | 0.008 | 0.029 |
| SNWD | 0 | 0 | 12 | 0 | 0 | 12 | 0.755 | 0.57 | 0.096 |
| AWND | 4.245 | 0.67 | 24.16 | 5.282 | 2.01 | 23.49 | 3.341 | 11.162 | 5.845 |
| TMAX | 22.333 | 10 | 105 | 42 | 35 | 95 | 18.858 | 355.637 | 46.83 |
| TMIN | 15.5 | -6.667 | 74 | 27.5 | 24 | 80.667 | 13.445 | 180.757 | 29.791 |
| WDF2 | 92.396 | 10 | 360 | 220 | 250 | 350 | 78.627 | 6182.181 | 210.971 |
| WDF5 | 91.714 | 10 | 360 | 214 | 190 | 350 | 76.293 | 5820.678 | 204.555 |
| WSF2 | 9.4 | 6 | 47 | 13.233 | 0.9 | 41 | 6.8 | 45.963 | 14.959 |
| WSF5 | 12.425 | 6.9 | 118.1 | 17 | 10.1 | 111.2 | 9.098 | 82.776 | 19.277 |

## MI

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute** | IQR | Minimum Value | Maximum Value | Median | Mode | Range | Standard Deviation | Variance | Mean |
| PRCP | 0.07 | 0 | 4.29 | 0 | 0 | 4.29 | 0.297 | 0.088 | 0.113 |
| SNOW | 0 | 0 | 14.9 | 0 | 0 | 14.9 | 0.918 | 0.843 | 0.22 |
| SNWD | 0 | 0 | 27.2 | 0 | 0 | 27.2 | 3.771 | 14.22 | 1.374 |
| WESD | 0.044 | 0 | 2.46 | 0 | 0 | 2.46 | 0.157 | 0.025 | 0.054 |
| AWND | 4.879 | 1.34 | 25.95 | 9.844 | 13.76 | 24.61 | 3.575 | 12.78 | 9.95 |
| FMTM | 620.986 | 3 | 9999 | 32 | 32 | 9996 | 1039.862 | 1081313.202 | 476.962 |
| PGTM | 1031.745 | 10 | 2356 | 1138.875 | 286.214 | 2346 | 610.534 | 372751.834 | 1055.37 |
| TMAX | 18 | -2 | 104 | 38 | 40 | 106 | 17.605 | 309.922 | 43.101 |
| TMIN | 14.5 | -32 | 77 | 26 | 21 | 109 | 16.114 | 259.655 | 26.796 |
| WDF2 | 89.449 | 10 | 360 | 225 | 215 | 350 | 74.565 | 5559.933 | 222.011 |
| WDF5 | 95 | 10 | 360 | 220 | 220 | 350 | 76.826 | 5902.205 | 218.599 |
| WSF2 | 7.44 | 6 | 40 | 19.9 | 26.55 | 34 | 5.524 | 30.517 | 20.195 |
| WSF5 | 11.4 | 6.9 | 57.9 | 26.8 | 35 | 51 | 8.095 | 65.542 | 27.515 |

## PA

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute** | IQR | Minimum Value | Maximum Value | Median | Mode | Range | Standard Deviation | Variance | Mean |
| PRCP | 0.05 | 0 | 6.59 | 0 | 0 | 6.59 | 0.329 | 0.108 | 0.116 |
| AWND | 3.58 | 0 | 22.82 | 4.03 | 2.91 | 22.82 | 3.044 | 9.269 | 4.822 |
| PGTM | 506 | 0 | 2359 | 1349 | 1353 | 2359 | 509.648 | 259741.229 | 1348.39 |
| TMAX | 31 | 11 | 100 | 66 | 83 | 89 | 18.54 | 343.713 | 64.049 |
| TMIN | 29 | -12 | 77 | 42 | 33 | 89 | 17.63 | 310.826 | 42.11 |
| WDF2 | 160 | 10 | 360 | 260 | 310 | 350 | 95.333 | 9088.353 | 227.556 |
| WDF5 | 150 | 10 | 360 | 250 | 310 | 350 | 94.963 | 9018.056 | 224.716 |
| WSF2 | 8 | 2 | 40.9 | 14.1 | 10.1 | 38.9 | 5.535 | 30.635 | 15.043 |
| WSF5 | 10.1 | 2 | 57.9 | 19.9 | 16.1 | 55.9 | 7.883 | 62.138 | 21.122 |

# 

# Correlation Matrix

## Code

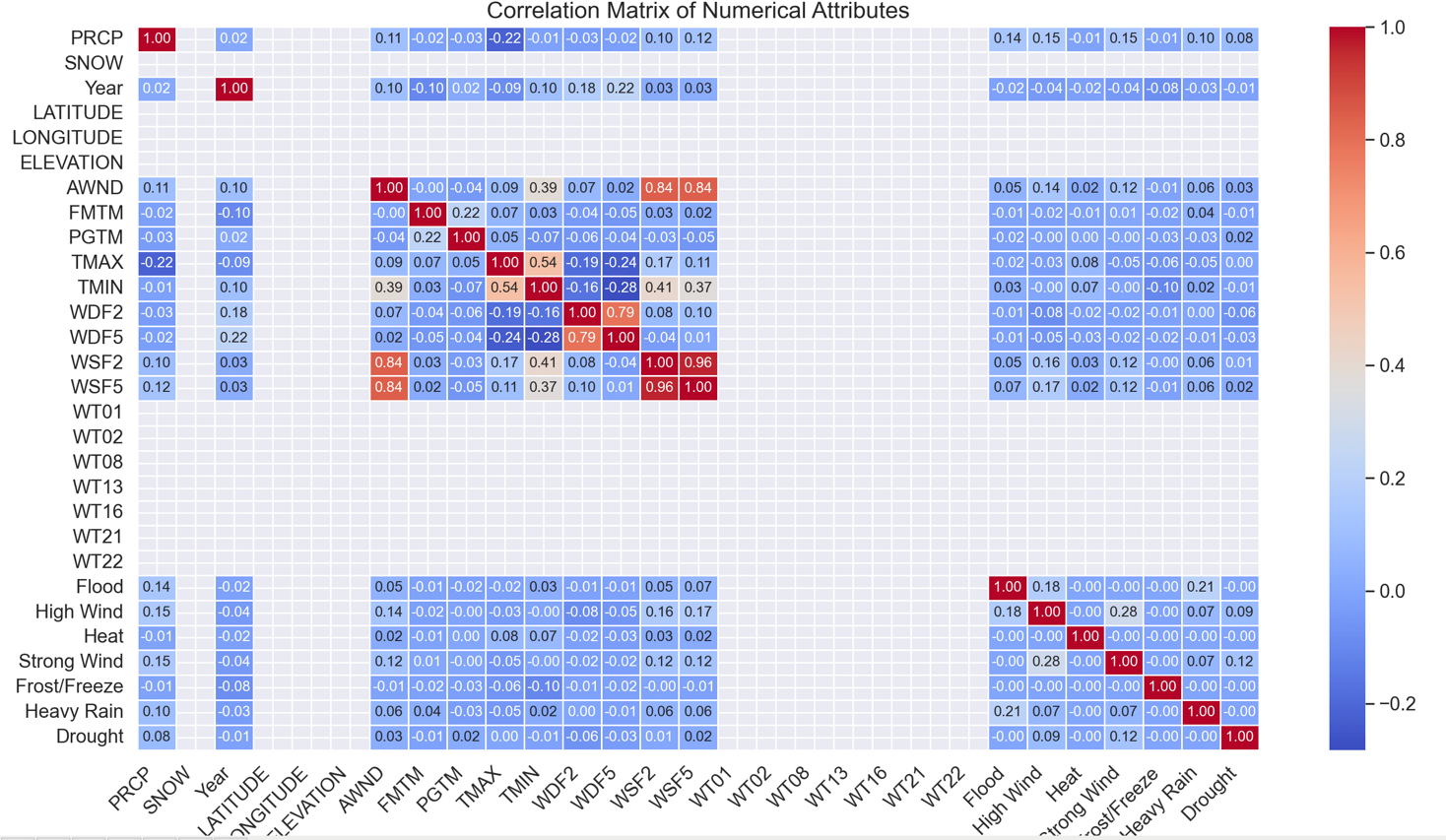
A screen shot of a computer program

AI-generated content may be incorrect.

## We first import the Python packages to help us generate a correlation matrix. The first package is pandas with a main role of selecting numerical columns from the dataset to form a correlation matrix on. Then the package ‘seaborn’ is used to create the heatmap, first by scaling the font of the labels up by 20%. Annotations are set to true (they will be visible and displayed on screen), the colour palette is set to ‘coolwarm’ from blue (low) to red (high), the correlation values are set to be displayed to 2 decimal places, the width of the lines is set to 1 while the size of the annotations in the heatmap cells has been set to 10.

Once all of this has been done, the x and y labels (which will be the attributes) are derived. The x labels are rotated to get the ones to be used as y labels. The final line of code, ‘plt.tight\_layout()’, ensures that everything can be generated in a compact form as a map and that the heatmap displays accurately with no problems before it actually appears on the screen. The result is also saved to a csv file for viewing later.

## CA



Key Observations:

* TMAX and TMIN has a strong positive relationship of 0.54, as expected, since temperatures would normally relate with each other.
* PRCP (Precipitation) negatively correlates with TMAX of -0.22, to suggest that as rainy weather kicks in, cooler temperatures are also brought along with it.
* The extreme weather event of flood and PRCP (approximately 0.14), the extreme weather event of high wind and PRCP (approximately 0.15), and the extreme weather event of strong wind and PRCP (approximately 0.15) all display moderate relationships.
* Wind related variables such as WSF2, WSF5 and AWND are positively correlated, with WSF2 and WSF5 at 0.84.
* There are attributes such as SNOW, LATITUDE, LONGITUDE and ELEVATION, alongside the waypoints of WT01, WT02 all the way to WT22, with no correlation marked. At first glance, this might be a result of missing data. However, the original processed dataset has these attributes as the same values (SNOW = 0, LATITUDE, LONGITUDE and ELEVATION as the same values, WT01 = WT02, etc., = 1), meaning that the variance is 0. Hence, there’s approximately no correlation.

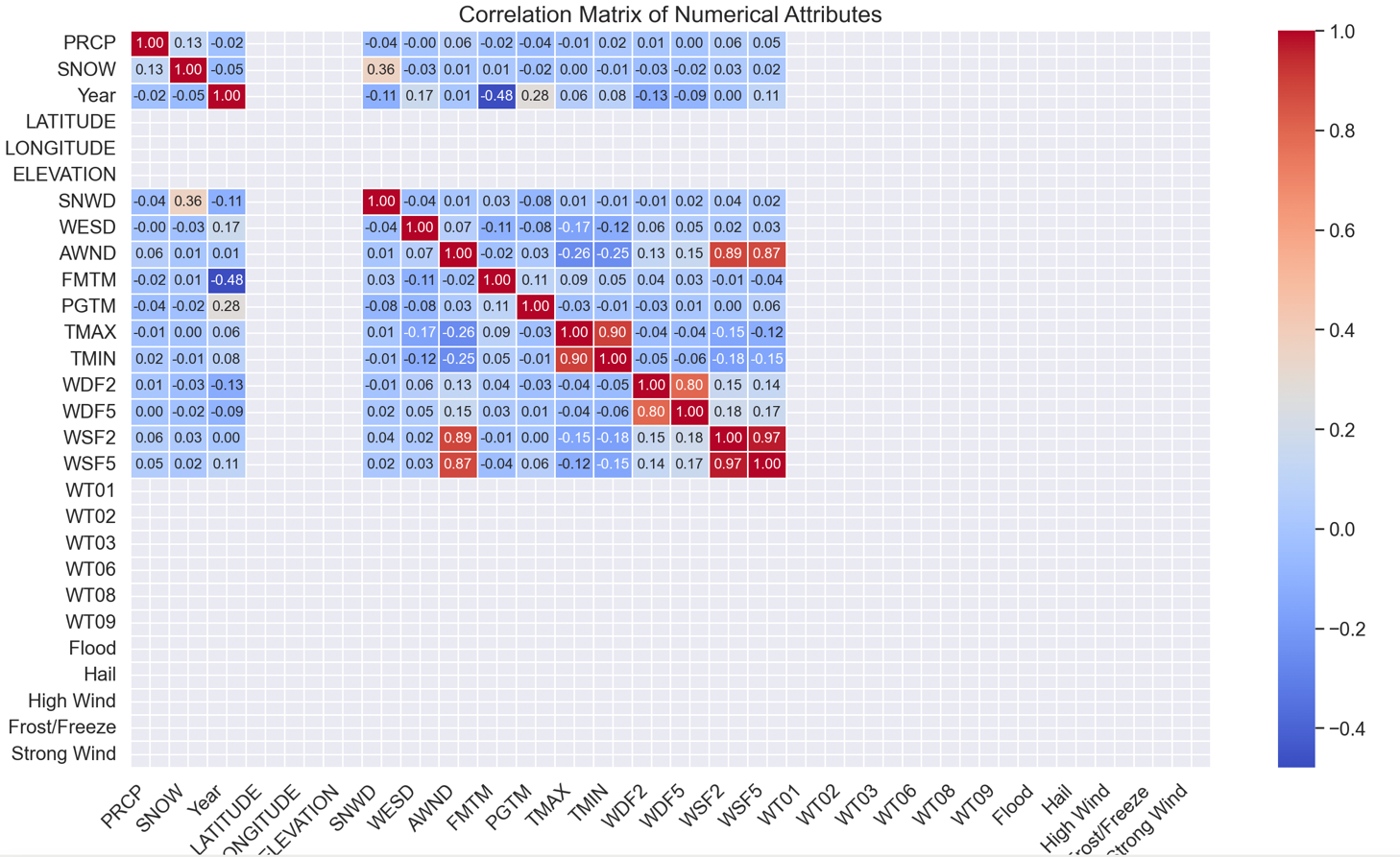
## WA



Key Observations:

* WDF2 and WDF5 has a moderate relationship with each other (0.59), and TMAX and TMIN have an extremely strong relationship (0.9), stronger than what we saw in CA but still expected.
* WESD values are all 0, meaning that variance is constantly near zero. Since it does not change much, it would not correlate with anything.
* WSF2 and WSF5 has an extremely strong relationship with AWND (0.85 and 0.88). If WSF5, WSF2, and AWND have high correlation values (e.g., >0.8), it indicates that locations or time periods with high short-term wind gusts also tend to have higher average wind speeds. This is expected since gusts and sustained winds are influenced by the same atmospheric conditions (e.g., storms, pressure systems).
* Wind direction (WDF2) varies depending on local geography, pressure systems, and obstacles (e.g., mountains, buildings), while average wind speed (AWND) depends more on overall weather conditions. If strong gusts come from random or shifting directions, they won’t always influence the average wind speed, leading to a low correlation (in this case of the heatmap, its -0.09). However, this does vary with the locale and the measuring methods implemented.
* Normal weather attributes such as PRCP, YEAR, SNOW and SNWD have weak relationships with the extreme weather event attributes (either weak positive or negative) because extreme weather events are sporadic and hence there’s not enough information to suggest a strong relationship. Floods depend on PRCP, but also on soil saturation, topography, and drainage. High winds may not always be linked to PRCP but instead to weather dynamics. This weakens the direct relationship between a single variable and extreme weather events. There are values for this because in the raw dataset, there are ‘1’s and they are not all 0s.

## MI



Key Observations:

* Extremely strong relationship between WSF2 and WSF5 with 0.97, almost perfectly aligned. It can mean that as the wind speed of WSF2 increases, WSF5 increases alongside it. It makes sense if it measures the wind speed at different times, but in this case, it will be at the same place since the data captures the wind speed at the airport of Michigan.
* The relationship between FMTM and Year has a moderate negative association (-0.48). This means that birth rates have decreased with time, but since this data measures an airport, it is unlikely that it would measure this. The most likely thing that it measures would be the migration patterns of people, which changes based on factors such as seasonal trends, economic opportunities and political stability.
* TMAX and WESD is -0.17, meaning that other factors likely contribute to wind speed and temperature changes more significantly rather than just solely those two attributes. In some meteorological scenarios, higher temperatures may be associated with lower wind speeds as regional patterns or seasonal variations.
* The relationship between SNOW and SNWD is 0.36, meaning that there is a notable tendency for snow depth to accumulate as more snow falls. However, it is not a perfectly strong or direct relationship, meaning that potentially there are other factors such as temperature and wind that determines it.
* AWND and TMAX or TMIN has a notable negative correlation, meaning that higher wind speeds seem to have a chilling effect on temperatures in general based on the dynamics of weather systems. Specifically at the airport, the strength is mild and not very strong. Though, the time of year and other specific weather conditions can further impact these results.

## PA

A screenshot of a graph

AI-generated content may be incorrect.

Key Observations:

* WDF2 and WDF5, with WSF2 and WSF5 have correlations of 0.28 and 0.29. This means that as the wind speed increases, wind direction tends to shift in a somewhat predictable manner even though the relationship is not very strong. It means that other factors likely influence wind variability. Wind direction fluctuates somewhat independently of wind speed.
* WSF2 and WSF5 have strong relationships with AWND (0.76 and 0.8). As short-term wind speeds increase, the average wind speed also increases in a consistent manner. It makes sense since AWND is likely calculated over a longer period such as daily or hourly, so higher short-term wind speeds contribute directly to increasing the overall average. Longer wind averaging periods of 5 minutes capture sustained wind speeds more effectively than very short-term (2-minute) measurements, which may be influenced by monetary gusts.
* PGTM has a bit of a relationship with TMAX and TMIN though little at 0.16 and 0.13. A slight positive correlation with **TMAX** could indicate that stronger peak wind gusts tend to occur on warmer days. The even weaker correlation with **TMIN** suggests that peak wind gust times are **not significantly influenced by minimum temperatures,** which typically occur during the night or early morning when atmospheric conditions are more stable.

# Association Mining/Rules

## Sample Code

A computer screen shot of a program

AI-generated content may be incorrect.

A computer screen shot of a program

AI-generated content may be incorrect.

The code:

* Imports the pandas Python package for dataFrame analysis and mlxtend for deciphering patterns using the Apriori algorithm for any association rules.
* Selects some of the attributes from a csv to be used for association mining and rules, which will not be the entire dataset since that will increase the computation time and memory usage.
* Converts any numerical attributes into 4 bins for the Apriori algorithm to work more efficiently: Low, Medium, High, Very High. Removes the original raw data after the binning process has succeeded.
* Properly fills in missing data with ‘Unknown’ by adding ‘Unknown’ as a category.
* Converts categorical variables in ‘df\_binned’ into binary columns (0 or 1) for the Apriori algorithm to work effectively while ensuring that the results are of an integer data type (preferred for association rule mining).
* Finds frequent itemsets with a minimum support threshold of 0.25, meaning that itemsets that appear in at least 25% of the transactions will be considered.
* Derives the association rules that are inherent in the dataset with at least 40% confidence and filters for the rules with at least 2 in the antecedents and consequents. Ensures that the rules are not sparse or have ‘Nan’ or ‘inf’ values.
* Alters the format when the rules are printed out by changing ‘frozen sets’ into lists for better readability, before compiling the association rules document and saving it as an Excel file on a device.
* Throws an error if no itemsets are found (minimum support threshold is too low).

Low, Medium, High, Very High classifications determine the 25%, 50%, 75% and 100% ranges of the given dataset, i.e., the dataset has been divided in 4 components.

## CA

Some Association Rules:

1. If WT01 or any of the proceeding waypoints occur such as WT02 and so on,
2. There is a 99.98% chance that the state is California.
3. TMAX is medium (around the 50% mark) with a 66% chance.
4. TMIN is medium (around the 50% mark) with a 48% chance.
5. TMIN is high (around the 75% mark) with a 45% chance.
6. PRCP is low (around the 25% mark) with a 99% chance.
7. If the TMIN is medium (around the 50% mark),
8. TMAX is medium (around the 50% mark) with a 82% chance.
9. If the TMAX is medium (around the 50% mark),
10. TMIN is medium (around the 50% mark) with a 60% chance.
11. If WT02 occurs, there is a 65% chance that both PRCP will be low and TMAX will be medium.
12. If WT02 and WT08 occurs, there is a 66% chance that both WT01 and TMAX will be medium.
13. If FMTM is low,
14. There is a 44% chance that AWND is low as well. If it is the reverse with AWND being low, then the chance of FMTM being low is 94%.
15. PGTM is high with a 91% chance.
16. TMAX is medium with a 66% chance.
17. If PGTM is high, there is a 43% chance for AWND to be low.
18. If TMAX is medium, AWND is low with a 46% chance.
19. If TMIN is high, AWND is medium with a 73% chance.
20. If WDF5 is high, WDF2 is high with a 73% chance.
21. If TMAX is medium and PGTM is high, AWND is medium with a 52% chance.
22. If TMIN is medium and PGTM is high, TMAX is medium with a 85% chance.

## WA

Some Association Rules:

1. If AWND is low,
2. TMAX is medium with a 51% chance.
3. TMIN is low with a 70% chance.
4. WDF5 is high with a 44% chance.
5. WSF2 is low with a 89% chance.
6. If TMAX is medium,
7. AWND is low with a 67% chance.
8. WDF2 is high with a 50% chance.
9. TMIN is medium with a 80% chance.
10. If TMIN is medium,
11. AWND is low with a 77% chance.
12. TMAX is medium with a 68% chance.
13. WDF2 is high with a 94% chance.
14. If WDF2 is high,
15. AWND is low with a 60% chance.
16. TMAX is medium with a 55% chance.
17. TMIN is medium with a 57% chance.
18. If WESD is medium,
19. AWND is low with a 67% chance.
20. WDF2 is high with a 47% chance.
21. WSF2 is low with a 65% chance.
22. If SNWD is low and WSF2 is low,
23. WDF2 is high with a 42% chance.
24. If WESD is medium and WSF2 is medium,
25. SNWD is low with a 98% chance.
26. SNWD is low and WSF5 is low with a 95% chance.
27. If WT01 or any other waypoint,
28. TMAX is medium with a 51% chance.
29. TMIN is medium with a 60% chance.
30. PRCP is low with a 99% chance.
31. If TMIN is medium,
32. TMAX is medium with a 68% chance.
33. PRCP is low with a 99% chance.
34. WT02 and PRCP is low with a 99% chance.
35. State is Washington and PRCP is low with a 99% chance.
36. If WT08 and WT01,
37. TMIN is medium and TMAX is medium with a 41% chance.
38. PRCP is low and TMIN is medium with a 60% chance.

## MI

Some Association Rules:

1. If WT01 or any subsequent waypoint,
2. TMAX is medium with a 68% chance.
3. TMIN is high with a 51% chance.
4. PRCP is low with a 98% chance.
5. If TMIN is medium,
6. TMAX is medium with a 81% chance.
7. PRCP is low with a 98% chance.
8. If PRCP is low,
9. TMAX is medium with a 68% chance.
10. TMIN is high with a 52% chance.
11. If WT01 and WT02,
12. TMAX is medium with a 68% chance.
13. TMIN is high with a 52% chance.
14. PRCP is low with a 98% chance.
15. If the state is Michigan and TMAX is medium,
16. TMIN is medium and WT03 with 41% chance.
17. If TMIN is medium and WT06,
18. TMAX is medium and PRCP is low with 80% chance.
19. If SNWD is low,
20. WESD is low with 98% chance.
21. WDF2 is high with 49% chance.
22. WSF5 is medium with 55% chance.
23. If AWND is medium,
24. SNWD is low with 91% chance.
25. If WESD is low,
26. TMAX is medium with 68% chance.
27. If PGTM is low and WESD is low,
28. SNWD is low with 88% chance.
29. If TMAX is medium and SNWD is low,
30. WESD is low with 99% chance.
31. If WDF5 is high and WESD is low,
32. SNWD is low with 91% chance.
33. If FMTM is low,
34. SNWD is low and WDF2 is high with 45% chance.

## PA

Some Association Rules:

1. If TMAX is medium,
2. PRCP is low with 100% confidence.
3. STATE is Pennsylvania and PRCP is low with 100% confidence.
4. If TMAX is high,
5. PRCP is low with 99% confidence.
6. If TMAX is very high,
7. PRCP is low with 99% confidence.
8. If PGTM is high,
9. AWND is low with 73% confidence.
10. If AWND is low,
11. PGTM is high with 54% confidence.
12. If WDF2 is very high,
13. AWND is low with 57% confidence.
14. If WSF2 is low,
15. AWND is low with 99% confidence.
16. If WT01 is medium,
17. AWND is low with 69% confidence.
18. PGTM is high with 51% confidence.
19. WDF2 is very high with 46% confidence.
20. WDF5 is very high with 43% confidence.
21. WSF2 is medium with 61% confidence.
22. WSF5 is medium with 59% confidence.
23. If WT01 is medium and AWND is low,
24. PGTM is high with 54% confidence.
25. If WSF2 is medium,
26. WSF5 is medium with 90% confidence.
27. If WT01 is medium and WSF5 is medium,
28. AWND is low and WSF2 is medium with 63% confidence.
29. If WT02 is medium and WSF5 is medium,
30. WSF2 is medium and WT08 is medium with 92% confidence.