Python Final Project
predict will it rain tomorrow

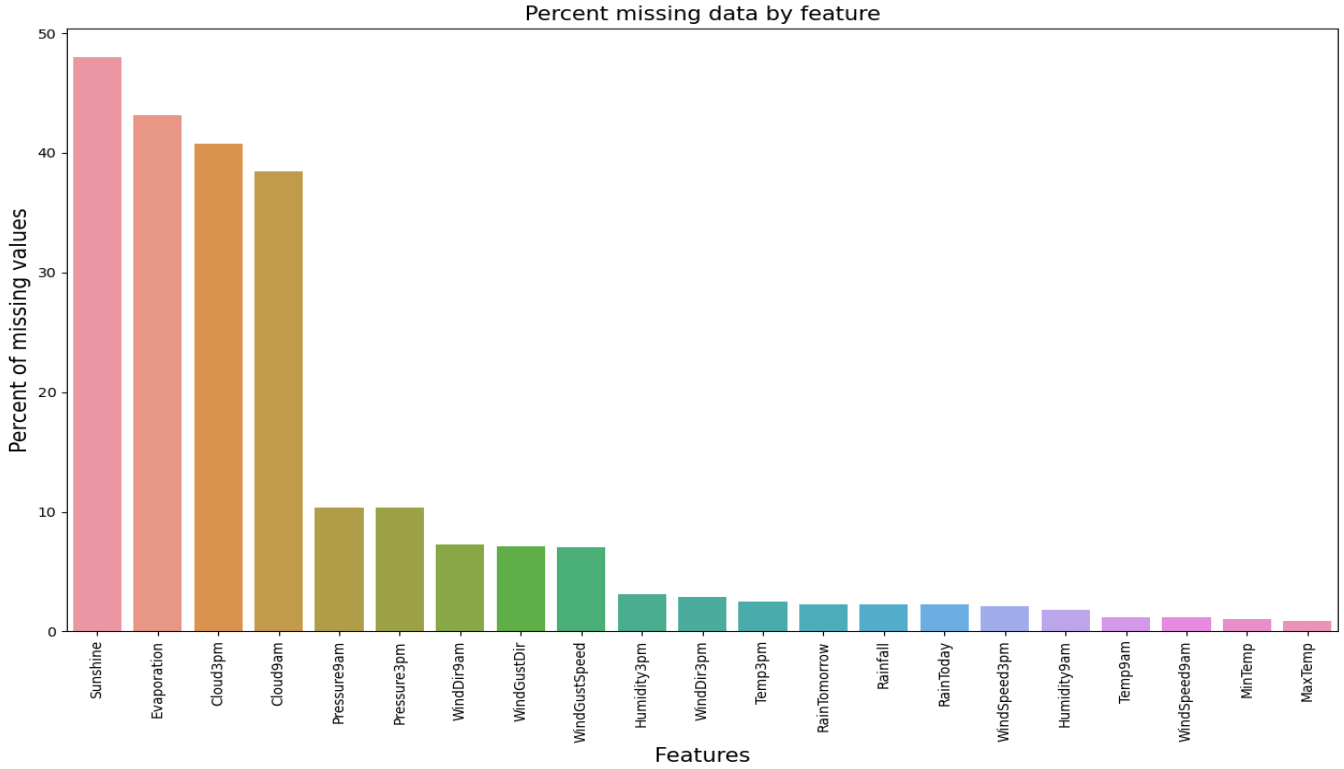
Danit Noa Yechezkel 203964036

1.introduction:

The data provided includes dates from 2008 to late 2017.
This dataset includes 145,460 observations from 49 different points across Australia.
From 23 variables describing various factors. One of the columns is the categorical target class column, 6 categorical and 16 quantitative variables.
Variables include information such as wind speed, humidity, temperature, cloud cover and more. The memory usage of the full file is 25.5 MB.

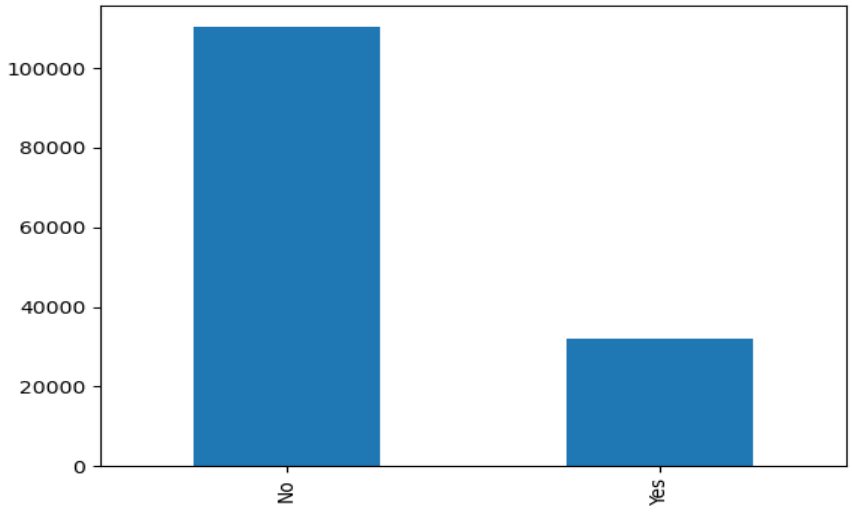
The total data missing can range widely from 1,261 "MaxTemp" values to 69,835 missing values in the Sunshine column. Evaporation and Sunshine suffer greatly from missing data with 43% and 48% of values missing respectively.
Most variables are missing a number of values, so we'll need to sanitize the data. Phase two of this project will be preparing the data for analysis by eliminating or filling those missing values, then identifying potentially strong predictors of the target class "RainTomorrow" variable.

	Number of NaN	Number of NaN in %
Date	0	0.0
Location	0	0.0
MinTemp	1485	1.0
MaxTemp	1261	0.9
Rainfall	3261	2.2
Evaporation	62790	43.2
Sunshine	69835	48.0
WindGustDir	10326	7.1
WindGustSpeed	10263	7.1
WindDir9am	10566	7.3
WindDir3pm	4228	2.9
WindSpeed9am	1767	1.2
WindSpeed3pm	3062	2.1
Humidity9am	2654	1.8
Humidity3pm	4507	3.1
Pressure9am	15065	10.4
Pressure3pm	15028	10.3
Cloud9am	55888	38.4
Cloud3pm	59358	40.8
Temp9am	1767	1.2
Temp3pm	3609	2.5
RainToday	3261	2.2
RainTomorrow	3267	2.2


Percent missing data by feature
Percent of missing values
50
40
30
20
10
0
Sunshine
Evaporation
Cloud3pm
Cloud9am
Pressure9am
Pressure3pm
WindDir9am
WindGustDir
WindGustSpeed
Humidity3pm
WindDir3pm
Temp3pm
RainTomorrow
Rainfall
RainToday
WindSpeed3pm
Humidity9am
Temp9am
WindSpeed9am
MinTemp
MaxTemp
Features

2.Initial Data Analysis

The bar chart below shows that the data is highly imbalanced. We have more than 100,000 observations with 'No' and only about 30,00 with 'yes'. In order to avoid biasing the prediction we will proceed with sampling the dataset.



First I checked if there were duplicate rows in the data in order to remove them, but there were no such rows in the dataset.
Second, we will delete the rows that are missing more than 10 values, because such a row could be considered not trustworthy and therefore, I have deleted 3609 rows from the dataset. Now this dataset includes 141,851 rows.
I also separated the date column into year, month and day.
For the categorical columns - I replaced the values in them with numerical values.
In columns where up to 10 percent of values were missing, and the columns were of a numeric type, the missing data was filled with the average in the column.
In columns where up to 10 percent of the values were missing, and the columns were of the category type, they were completed by the value that appeared most in this column (the mode).

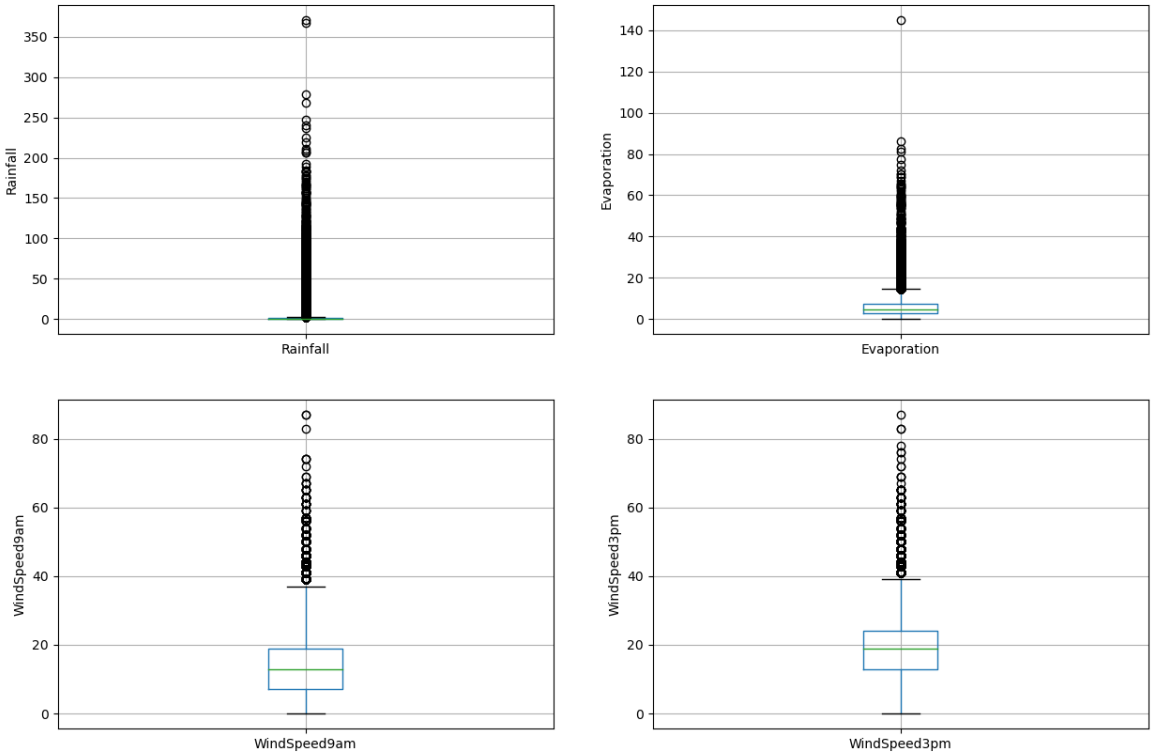
On closer inspection, we can see that the Rainfall, Evaporation, WindSpeed9am and WindSpeed3pm columns may contain outliers.

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	\
count	141621.0	141636.0	139803.0	81820.0	75033.0	134200.0	
mean	12.0	23.0	2.0	5.0	8.0	40.0	
std	6.0	7.0	8.0	4.0	4.0	14.0	
min	-8.0	-5.0	0.0	0.0	0.0	6.0	
25%	8.0	18.0	0.0	3.0	5.0	31.0	
50%	12.0	23.0	0.0	5.0	8.0	39.0	
75%	17.0	28.0	1.0	7.0	11.0	48.0	
max	34.0	48.0	371.0	145.0	14.0	135.0	

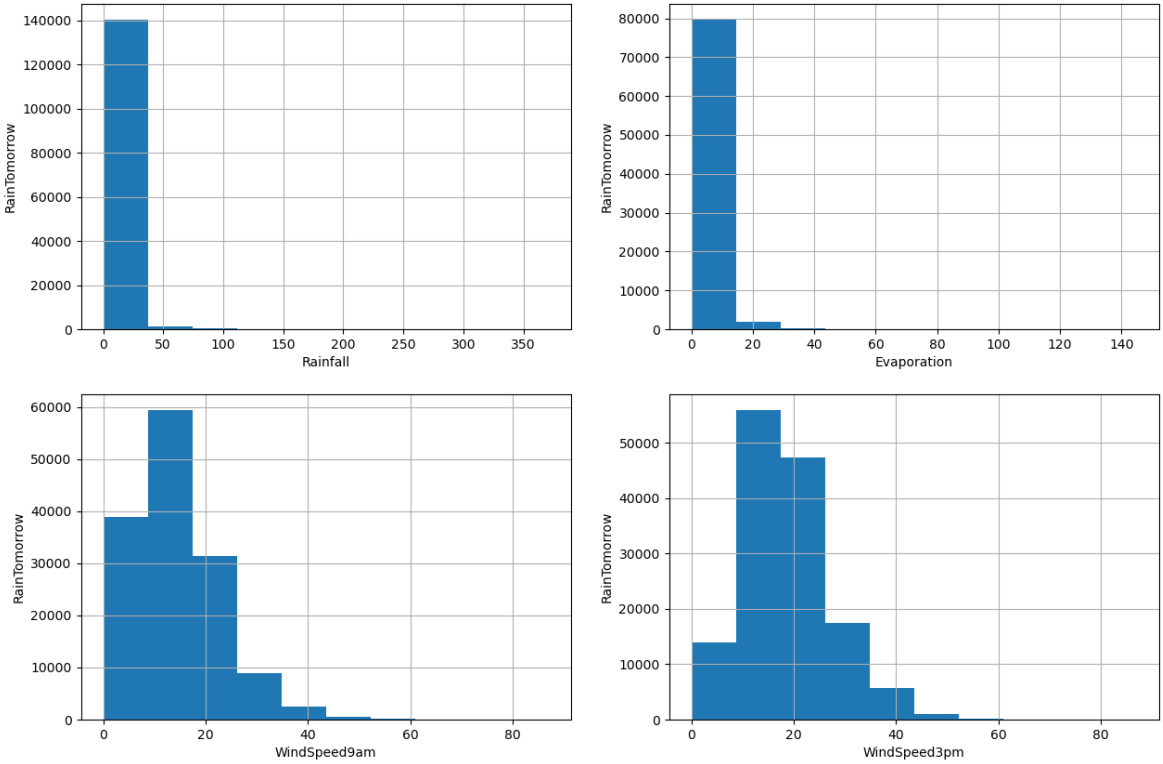
	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	\
count	141591.0	141188.0	140759.0	139793.0	129769.0	
mean	14.0	19.0	69.0	52.0	1018.0	
std	9.0	9.0	19.0	21.0	8.0	
min	0.0	0.0	0.0	0.0	586.0	
25%	7.0	13.0	57.0	37.0	1013.0	
50%	13.0	19.0	70.0	52.0	1018.0	
75%	19.0	24.0	83.0	66.0	1022.0	
max	87.0	87.0	100.0	100.0	1041.0	

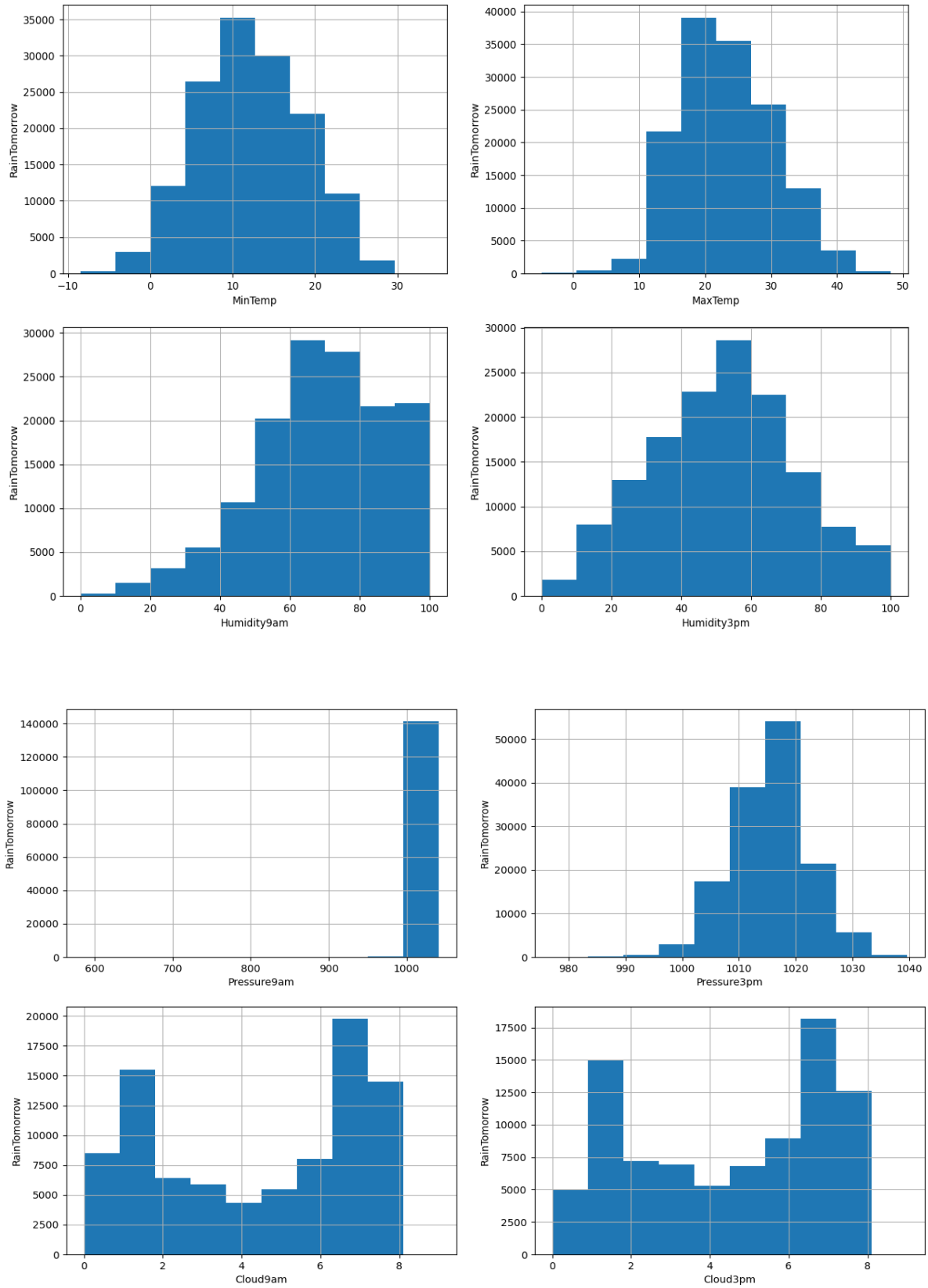
	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
count	129794.0	88451.0	85946.0	141556.0	140621.0
mean	1015.0	4.0	5.0	17.0	22.0
std	7.0	3.0	3.0	6.0	7.0
min	977.0	0.0	0.0	-7.0	-5.0
25%	1010.0	1.0	2.0	12.0	17.0
50%	1015.0	5.0	5.0	17.0	21.0
75%	1020.0	7.0	7.0	22.0	26.0
max	1040.0	9.0	9.0	40.0	47.0

The boxplots below confirm that there are many outliers in these variables.



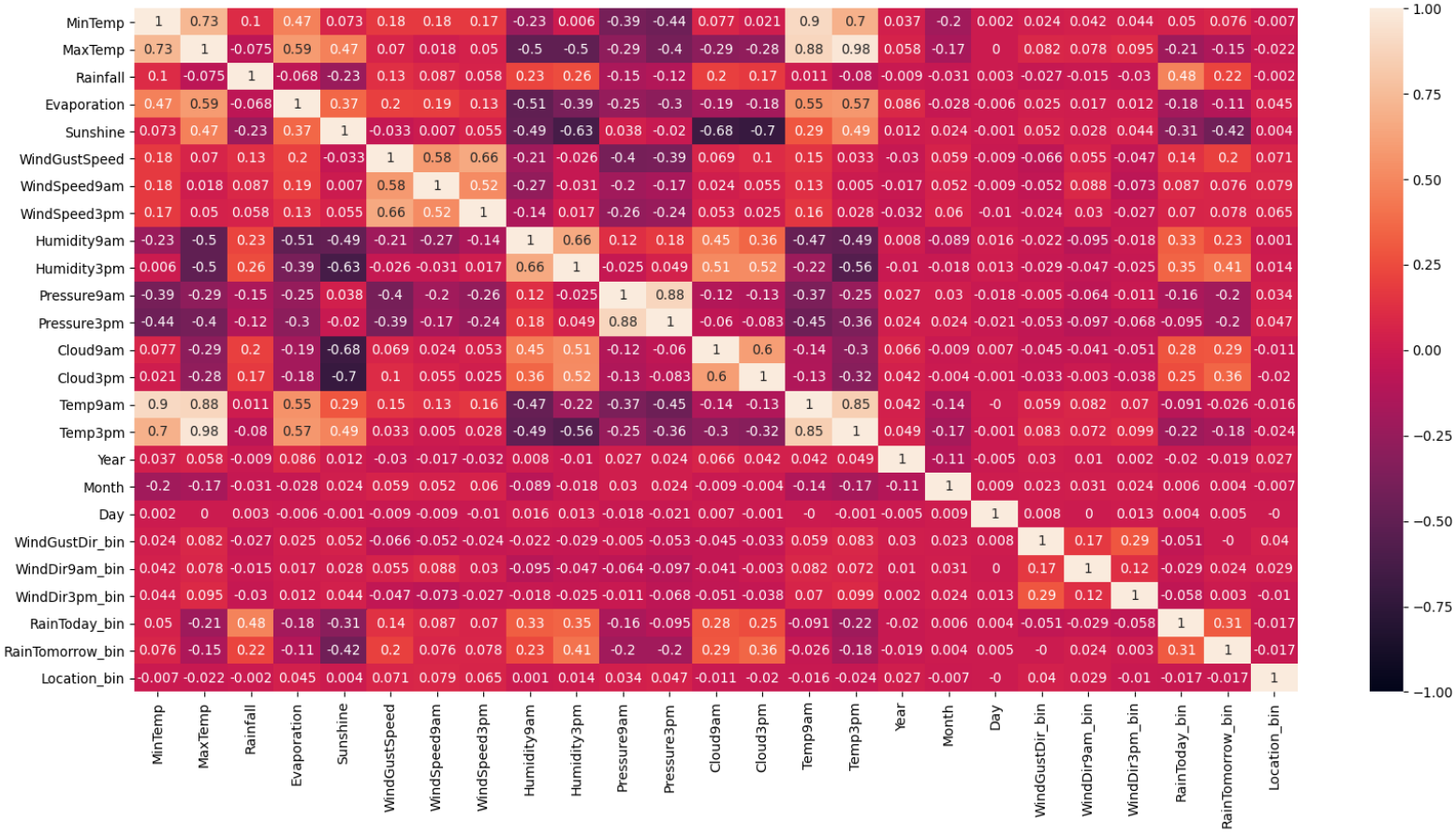
It can be concluded that the histograms (below) are biased and therefore we will need to delete abnormal values in order to perform better scattering





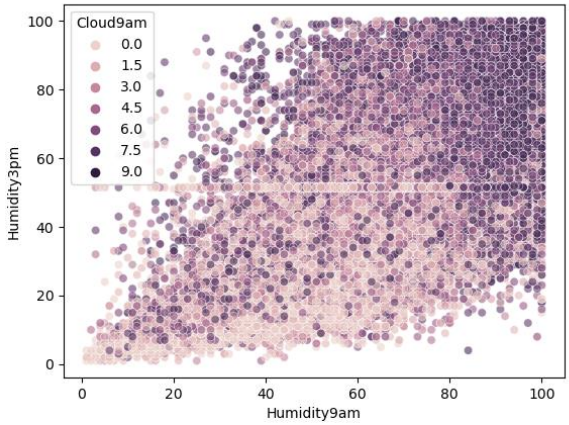
The blatant outlier values in each column amount to nine rows and are therefore deleted.

Another check of exceptions found that in most columns the value zero is a valid value, but in columns Humidity9am and Humidity3pm the value is not valid, there are only 4 such rows in total so I decided to delete them.
The last treatment left to perform, is for the four columns where there is close to 50 percent of missing values.
We will notice which columns have a high correlation, and we will complete the mising values by direct relationship with these columns.

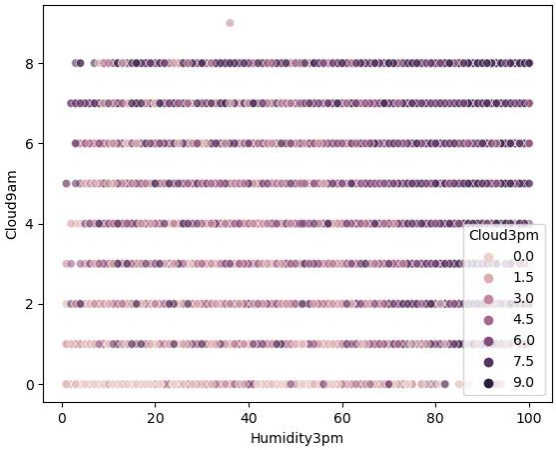


The evaporation column was with a deficiency of over 40 percent, but its value range is very small (as seen in the box plots above) so the missing values were completed with an average.

The Cloud9am column is missing about 40 percent of entries.
This column is highly correlated with my target class so I decided not to delete the rows or make an average - but to supplement it by relativity with other columns that are correlated with it. In this column the high correlations were with Humidity9am, Humidity3pm and looking at the scatter plot below it is easy to see the relationship. Therefor the values we filled using the Humidity9am and Humidity3pm values.

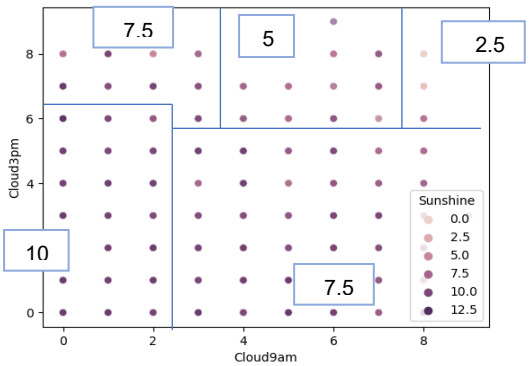


The Cloud3pm column is missing about 40 percent of entries.
This column is highly correlated with my target so I decided not to delete the rows or make an average - but to supplement it by relativity with other columns that are correlated with it. In this column the high correlations were with Cloud9am, Humidity3pm.



The Sunshine column is missing about 40 percent of entries.
This column is highly correlated with my target so I decided not to delete the rows or make an average - but to supplement it by relativity with other columns that are correlated with it. In this column the high correlations were with Cloud9am, Cloud3pm.
Here is an example of dividing the Sunshine column by the values placed in Cloud9am, Cloud3pm.

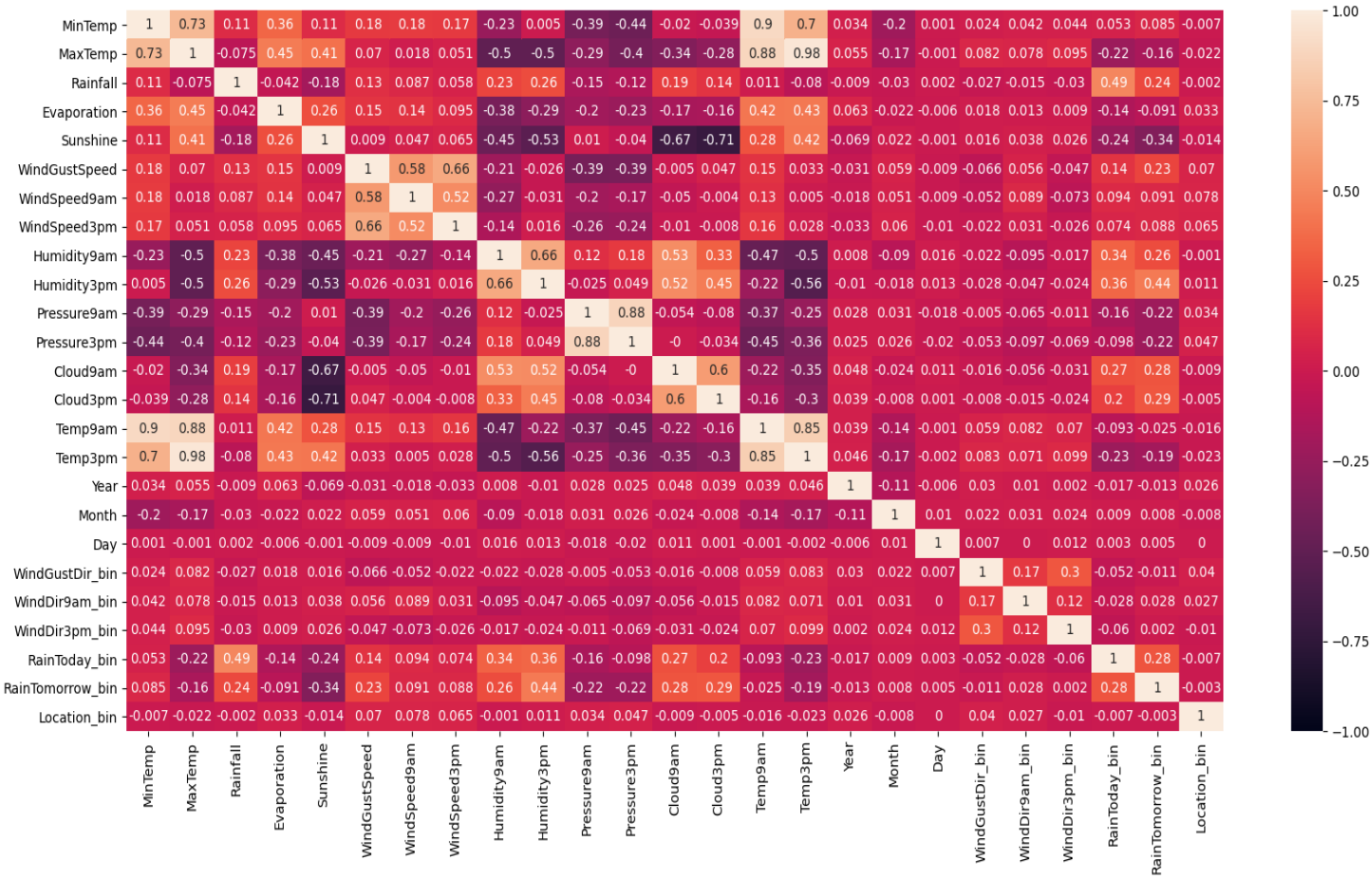
Below is an example of dividing the Sunshine column by the completed values, for example when Cloud3pm is greater than zero but less than six and also Cloud9am is greater than zero but less than two The missing value in the Sunshine column was filled as ten.



I decided at this point not to remove any columns at all in order to allow the algorithms in the next section to decide what the best attributes are, so I decided to delete only rows based on the data shown above.

3.Exploratory Data Analysis

Now the dataset has 139,792 rows and the new correlation map is:



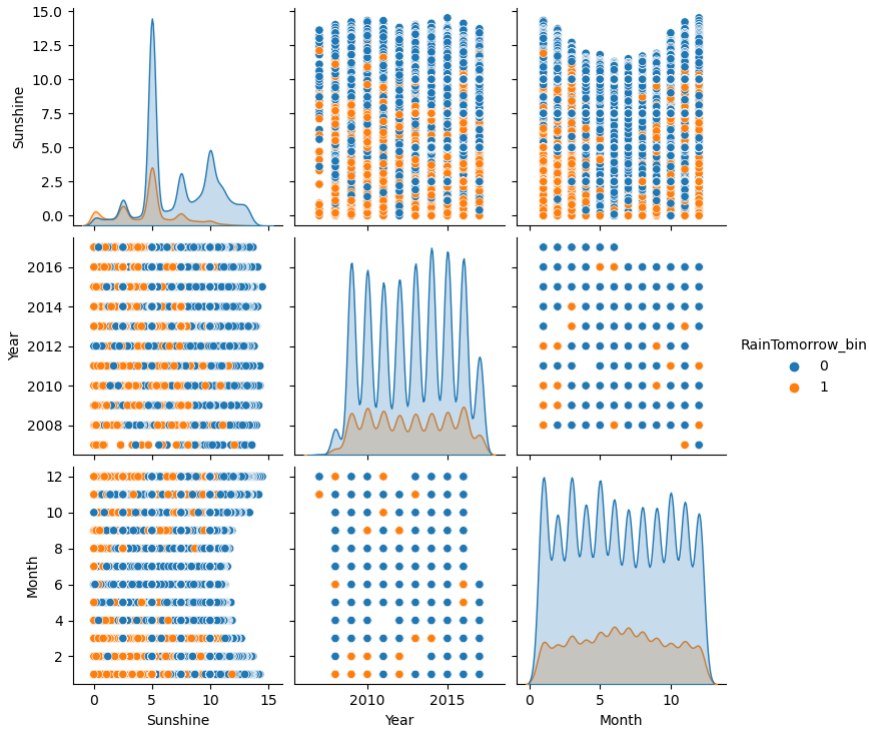
It is easy to see that the columns with the high correlation with the target class are: ‘Sunshine’, ‘rainToday’, both of the ‘Cloud’ columns, both of the ‘Humidity’ columns, both of the ‘Pressure’ columns, ‘windGustSpeed’ and ‘Rainfall’.

Now, knowing the correlations, I will try to create combinations on these column to try and maximize the correlation even more. This will hopefully enable a good prediction outcome of the question: “will it rain tomorrow?”.

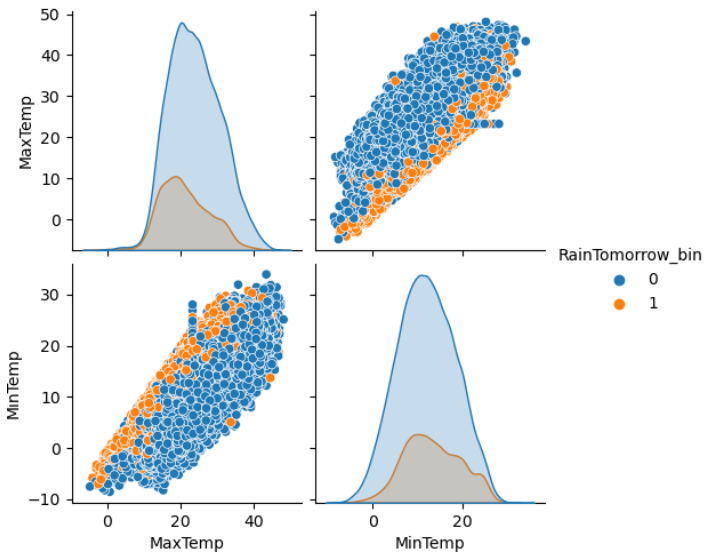
The plot below shows the correlation between the ‘Month’, ‘Year’ and ‘Sunshine’ columns.

We can infer from this that the summer months are roughly 10 (October) to roughly 3 (march) and in these months there is a better chance of rain and more sunshine hours in a day.

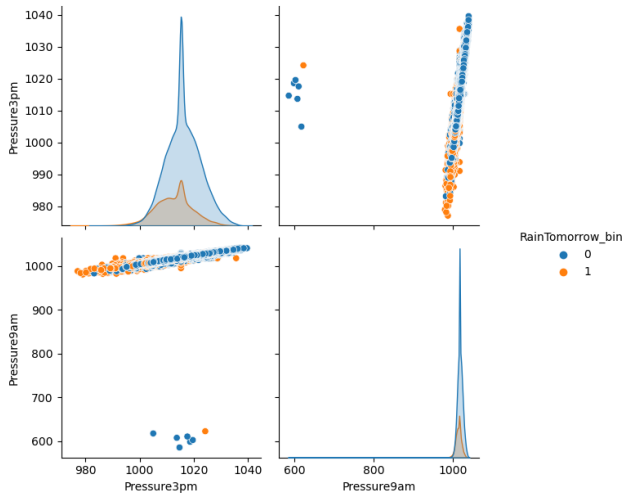
The winter months, from 4 (April) to 9 (September) have less sunshine and also less rain.



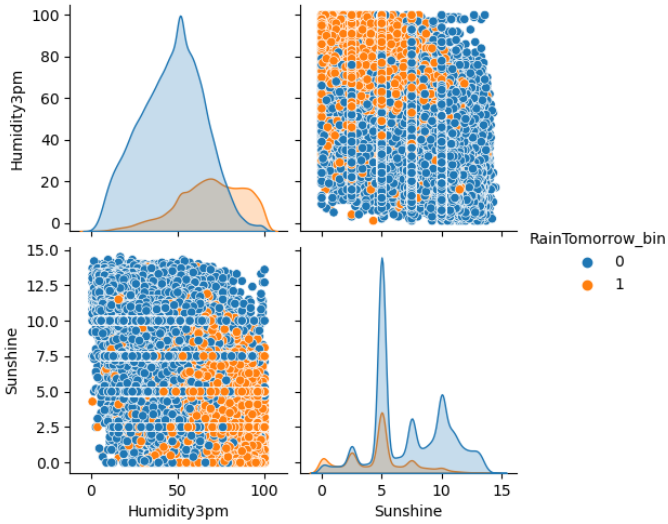
From the next plot below we can learn that there is a strong correlation between the 'MaxTemp' and the 'MinTemp', also we can see that it is more likely to rain on the days where the temperature drop between max and min were rather small (the two temperatures were in a linear correlation with each other).



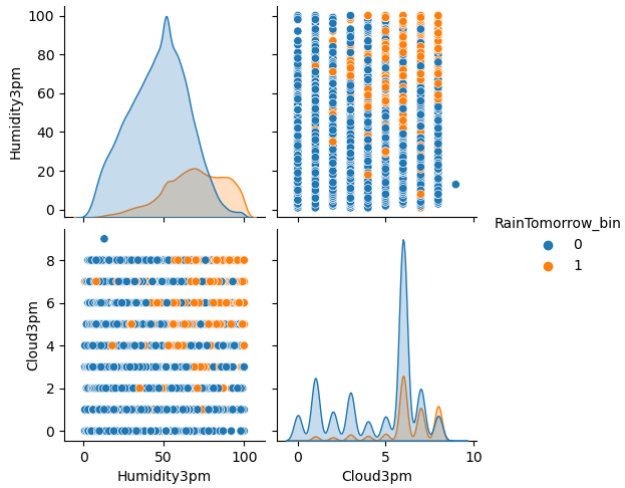
The Pressure plot tell us that when the morning pressure is lower than 700 and when the afternoon pressure is lower than 990 then the chance of rain gets significantly higher.



The Humidity and Sunshine plot shows that when the humidity is high (more than 60) and the sunshine is low (less than 7) then the chance of rain is rather high.

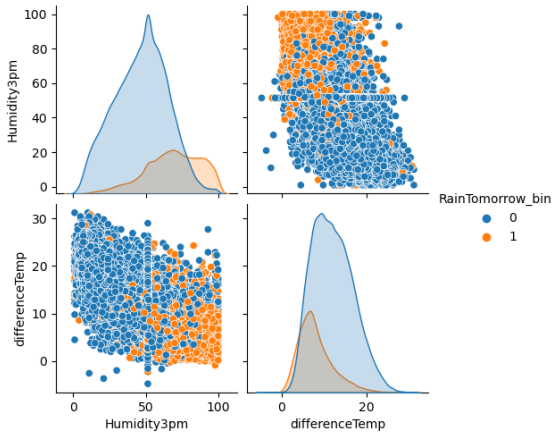


The Humidity and Cloud plot shows that the cloudier the sky and also the more humid the air, the chance of rain gets higher.



The next plot shows the correlation between the ‘Humidity3pm’ and a column that I have created by calculating the difference between the ‘MaxTemp’ and the ‘MinTemp’.

We can learn from this plot that when the difference between the two temps is smaller (as also mentioned above) then the chance of rain goes up. And also, the chance of rain increases with the increase of humidity.



In general, I have noticed that the greatest impacts on the question of “if it will rain tomorrow” are the difference between the min and max temperatures, the air pressure and the air humidity. Additionally it is more likely that it will rain sunshine is low, cloudiness is high and interestingly, the actual temperature is not relevant, just that there is no major drop or increase, so that it is just a likely that it would rain if the temperature is very low (10 degrees) as it is when the temperature is high (30 degrees) as long as the lowest and the highest are rather similar.

4.Classification Model

I shall explore two classification models in this chapter, the gaussian naïve bayes and the tree classifier.

I shall discuss the GNB first:

It is a two-feature GNB model and for this reason I would like to combine columns in order to get as much data into the two features as possible so that the learnability of the algorithm is the best possible.

The first column I have chosen for the GNB model is a combination of the 4 columns: ‘rainToday’, ‘Humidity3pm’, ‘Pressure3pm’, and both of the ‘Cloud’ columns.

The second column is also a combination of a few features: ‘windGustSpeed’, ‘Humidity9am’ and ‘Rainfall’.

The accuracy score of these column as the basis of the GNB model is as shown:

```
'Cloud_RainHum3_Pressure3pm' and 'WindGustSpeed_Humidity9_Rainfall'
accuracy score: 0.8094
```

Now I will move on to the tree classifier:

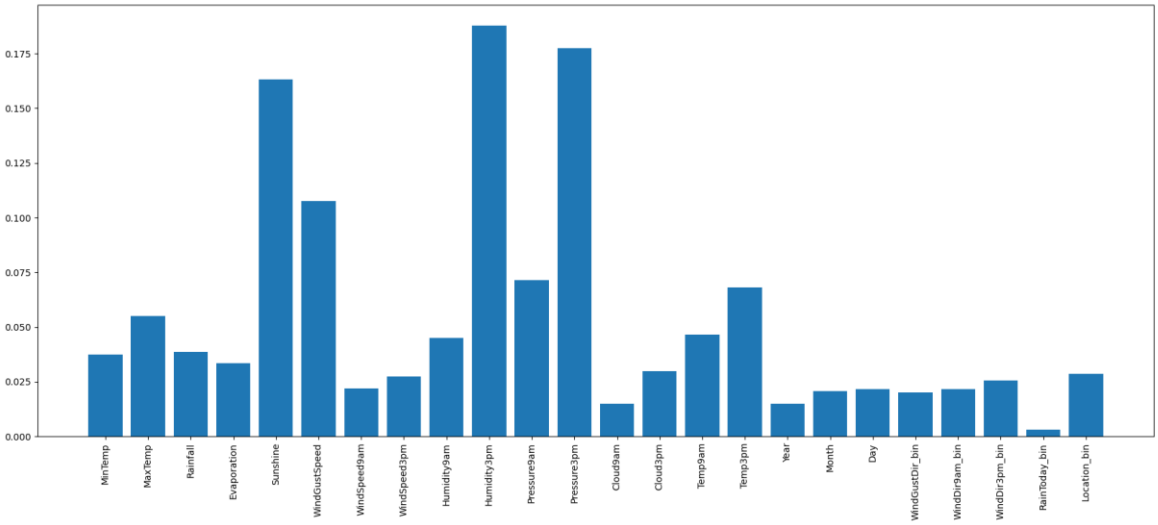
First I will start with the original data for a baseline tree. This provided me with a tree that stands at 80% accuracy:

```
              precision    recall  f1-score   support

           1       0.87      0.86      0.87      8781
           2       0.54      0.56      0.55      2503

    accuracy                           0.80     11284
   macro avg       0.71      0.71      0.71     11284
weighted avg       0.80      0.80      0.80     11284
```

The permutation importance of the original data tree is as shown:



And the most important features are:

Sunshine, windGustspeed, humidity3pm, pressure3pm, pressure9am, temp3pm.

Now I will create the tree from the data that I have cleaned and have added columns to.

	precision	recall	f1-score	support
0	0.86	0.85	0.86	21784
1	0.49	0.52	0.51	6157
accuracy			0.78	27941
macro avg	0.68	0.69	0.68	27941
weighted avg	0.78	0.78	0.78	27941

Unfortunately, it is easy to tell that I was not able to improve the learnability of the model to improve its prediction, the reasons for this might be that the amount of missing data was exceptional and I did not erase it all as I have done with the original data prior to the creation of the tree above. This amount of missing data (more than half of the rows contained at least one missing value) might have had a very bad impact when I tried to “fix” instead of deleting it outright and it is possible that I have skewed the learnability.

After the creation of the tree itself there was an increase in its success.

	precision	recall	f1-score	support
0	0.87	0.94	0.90	21784
1	0.69	0.49	0.57	6157
accuracy			0.84	27941
macro avg	0.78	0.71	0.74	27941
weighted avg	0.83	0.84	0.83	27941

Summary

The findings from this dataset have led to the considerations of these facts:

First: check the highest correlated features: temperature, humidity, pressure.

During this project I have encountered many difficulties, most of which occurred during the “cleaning” phase.

Having a dataset with so many features was also a challenge because I truly tried not to remove any columns and I was left with so many that it was hard to keep track of everything.

The target class column had missing data as well and this led me to think what would be the best practice for this situation. I ended up erasing these rows because I felt that they were needed for the learnability of the model algorithms.

Another major issue I had trouble with was the amount of missing values in the dataset. More than half of the rows had missing values and some of the columns were missing over 40% of their values. Not being sure what to do with the missing values, I hade some replaced with the average of the column and some replaced with values using correlated features.

The ‘Location’ column which was categorical, Had 50 different values and I was confused how I should take this into consideration. I decided in the end to leave the column be and not use it because my target question was whether it would rain in Australia and not in a specific place in the continent.

And lastly, seeing that the accuracy score was not improving even with checking multiple of features and creating new ones it felt very disappointing.

In conclusion, with this being the first time I am dealing with a project such as this I feel as though I have done well in studying the data provided and fell that I have reached a good outcome.

This project has improved my experience and knowledge in both data analysis and python and also, of course, meteorology.

Video:

https://drive.google.com/file/d/1ZaxSadf4NQiaErI1aq0KthnZNIBzmSUu/view?usp=sharing