**Data Exploration**

Author: Jose Manuel Alvarez Gonzalez

Email Id: [jmanuel.ag89@gmail.com](mailto:jmanuel.ag89@gmail.com)

Date: 20/4/2022

Contents

[Introduction 3](#_Toc132899550)

[Exploration, Transformation and Analysis 3](#_Toc132899551)

[1.- Data Set 3](#_Toc132899552)

[2.- Statistical Properties 4](#_Toc132899553)

[3.- Null Values 7](#_Toc132899554)

[4.- Logarithmic Transformation 9](#_Toc132899555)

[5.- Normalization 9](#_Toc132899556)

[6.- Transforming Categorical to Numerical 10](#_Toc132899557)

[7.- Binning 10](#_Toc132899558)

[8.- Feature Selection 10](#_Toc132899559)

[9.- Scattered Plot 11](#_Toc132899560)

[10.- Heat Map 12](#_Toc132899561)

[Conclusion 13](#_Toc132899562)

# Introduction

For this project, we are going to work with a dataset that contains daily weather observations for various Australian weather stations across 2019 and 2020. The data set has 25 features and 1 target variable which is if it rained that day or not.

The propose of this project is to explore the data set and prepare it to be used by a Predictive model. This report will document the stages of the process and the reasoning behind each decision taken.

# Exploration, Transformation and Analysis

The following part of the report is going to cover all the different steps taken during the stages of collecting the data, summarizing the data, identifying their distribution and properties, working with null data and noise, correcting any mistakes, transforming the features according to their data type, selecting the more suitable features and visualising to help identifying any insights found on the data.

## 1.- Data Set

For this part we want to understand the data set we have, we know that is 25 features and one target, from the original data we have a total of 11,848 samples. The first stage is to randomly select 70% of those samples, from that selection we end up with a total of 8,293 samples to work with.

To continue with the initial exploration, we label the data according to the data types they have.



Figure , Semantics and value type of the data

From this initial exploration we can identify a mixture of quantitative and qualitative data, at the same time we have nominal and ordinal values with some of it already on numeric form like “Cld 8th- 9:00 AM” but “Date” still on a non-numeric format, with the latest requiring further processing in following stages.

For the following stages, I changed the name of the features for simplicity.

## 2.- Statistical Properties

For the non-categorical data on an interval range, we also want to have an exploration of how they are distributed and the range of values they carry. We have the following table with the main statistics of the features:



Figure , Statistics for interval features

First is important to mention that the data currently has some null-values, these statistics should not be majorly affected by the way these values will be handle, we will talk more about it in the following chapter.

From the table we can identify that we have very different ranges of values and units from what the features are measuring. For example, we have four features measuring temperature going from a minimum of -5.9 to a max of 48.9, while “Rain\_mm” ranges from 0 to 254.8.

For the spread we have the standard deviation as this is easier to interpret than the variance, from the temperature features, we can see that all of them have a total of around 7.5 units for the sd. Meaning that 68% will be +- 7.5 units of their mean value. While “Rain\_mm” might have a big range, it also has a sd of 7.72, already giving hints for outliers.

At the same time, I can notice “AVG\_Rain\_mm” has only one value of 1.9, this is not a useful feature.

Next, we want to see the distribution of the features to find any normality around it or the lack of it.

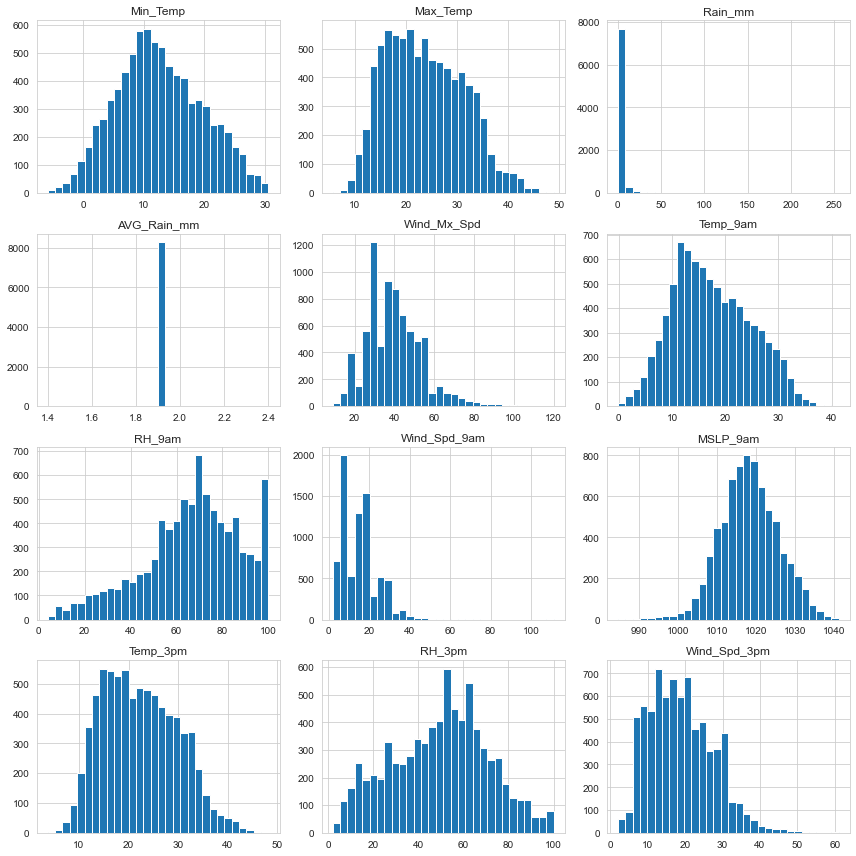


Figure , Histogram for 12 interval features

From figure 3, we can identify that most of the features are not normally distributed, this will require some work at trying to reduce the big ranges across the values to remove outliers. For the next part of the project, as we are working with not normally distributed data, we are going to look at the median for the measure of centrality, this will be essential at fixing null values.

For the non-interval features, we have figure 4, representing the total amount of unique values present on each one. Before producing this table, for the feature of “Month” we had two values for June (“Jun” and “June”) for the feature of “Station” we had various values for the "Ballarat Aerodrome", we had to convert them all to one.

For the feature of “Win\_Mx\_Tim” we are only keeping the hour, to reduce the total amount of possible results.



Figure , number of unique values for non-interval features

## 3.- Null Values

Most of the features in the data set had a “?” present, this changed the data type of the features to an object for most of them. We had to replace this to a numpy NaN value, then we were able to convert to numeric data type and perform statistic summaries. For the null values we have the following table:



Figure , total null values for each feature and proportion of total data

In figure 5, we can see the total amount of missing values per feature but most important we can see how much of that equals to the total amount of the feature. For “Cld\_9am” and “Cld\_3pm” with more than half of all values, we could include a model to predict the missing values, but at the same time we are using that feature for a future prediction. I decided to exclude those features from the final data set.

Then I looked at the median value of the interval features across the different States, we can see clear differences across the states for example the median Min\_Temp for ACT = 5.45 and NT = 20.6. I decided to also look across the stations and find if we have major differences, in this case is not as significant as per State.

To have a better idea of where the NaN values are located, I decided to investigate how many null values we have per Station. I noticed that is quiet even spread for most stations except for "Newcastle University {station 061390}" and "Albany {station 009500}", I created the following table for those stations:



Figure , Albany and Newcastle null values per feature

In figure 6, we can identify that Albany has a total of 8 features with 100% missing values and 9 features for Newcastle.   
  
I prefer to eliminate this stations from the dataset as we can’t fully rely on the quality of the stations, before doing that step, I looked at how much each of those stations represents of the total data and we have a combined of 7.1% from the two, I decided to exclude the stations of the next steps.

Now I replace the null values, first I grouped the data according to the State, we already saw how much impact it plays. For the interval features I will replace the null values with the median value for the State they belong, for the non-interval features, I will replace the value with the mode of the State they belong.

## 4.- Logarithmic Transformation

For the initial summaries and histograms for the interval features, we were able to identify some outliers on the data, at the same time high skewedness in some of the features. A way to try to eliminate it is by performing a Log transformation. Before performing a transformation, I collected the current number of outliers per feature before and after the Log transformation, we have the following:



Figure , Outliers before and after Log transformation

From the table we can see that it helped to eliminate outliers in the “RH\_3pm” feature but besides that feature we don’t have an improvement. I also looked at histograms, statistics and I couldn’t find any improvement.

For models where the range of values is an important part of the process like in a Linear Regression, is important that we adjust the features when we have different ranges like on this data set. For this particular case the Log transformation is not the best option, we will have to do a normalization for a better result.

Another problem for our Log transformation is that we have 0 and negative values in some of the features.

## 5.- Normalization

Based on the previous result, I decided to do a normalization for the interval features. For that I used the function Min Max Scaler, this creates a scale value from [0 to 1] from the minimum and maximum value. Doing this allows to have same units for all features.

## 6.- Transforming Categorical to Numerical

This part of the process must be carried on two steps, we first need to identify nominal and ordinal features.

For the nominal features, I decided to use the “LabelEncoder” function, I prefer to use this method at the moment as we don’t really know what type of model we are going to be using the data for. Using this encoder will allow us to easily go back and replace with a One-Hot encoding if required. At the same time, we have a total of 28 (minus two) stations this can lead to overfitting on the results if a particular station tends to be on a region with high amount of rain.

For the ordinal features, we first need to create a dictionary on the relevant order we require for each of the values. Once we have the order of the values, I used the function “OrdinalEncoder” to transform the data.

## 7.- Binning

For the categorisation of the interval features, I decided to convert the Temperature based features to categorical. To do the transformation, I first looked at the standard deviation of the features which is around 7.5, I want the ranges to be lower than that figure, I decided to go for intervals of 5c.

To calculate the number of bins for each feature, I followed the formula:

Bins = (Max Value – Min Value) / Width

Where width is 5c, for the function to allow this width between categories we use .cut(). For the propose of this project, I left the values on an numerical range starting from 1 for the lowest.

## 8.- Feature Selection

For this part of the process, I decided at using four different strategies to find what results we get:

* Correlation Analysis
* SelectKbest by chi2
* SlectKbest by mutual info classifier
* SelectFromModel, feature importance of a Decision Tree Model

From each of the four methods we have the following results of the features:



Figure , four different methods and results for feature selection

From the table we can identify that chi2 favours categorical features while Select From Model allow use to choose according to which model we are looking at work with. In general, we have relatively close results, with “Rain\_mm” and “RH\_9am” as some of the most important features while “Day” and “Year” scoring lower.

## 9.- Scattered Plot

Another method to choose features is by doing this visually, we created a paired scattered plot across all features with the Target “Rain” as a grouping factor. From the visual we could clearly identify some of the most relevant features, this method came closer to the Mutual Classifier function used previously. I took the top 3 and bottom 3 features of that model to create the following plot.



Figure , scattered plot with top and bottom three most important features

Starting with “Rain\_mm” we can clearly see how it clusters the data on all of the features from the days it rained and didn’t, relatively similar for both of the “RH” features, particularly when you look at both of them plot together. For the least important features we can see how the raining and not rainy days overlap on each other, and this is easy to think as we had rain across 2019 and 2020.

## 10.- Heat Map

For the visual representation of the correlation analysis, we have a heat map, here we can see how each feature, correlates to each other visually. For the scores, we have the same as on figure 8, with “RH\_9am” as the most correlated with a value of 0.41. Interestingly “Max\_Temp” is one on the highest inversely correlated, meaning that having a higher temperature reduces the chance of rain, correlation is with a value of -0.31.

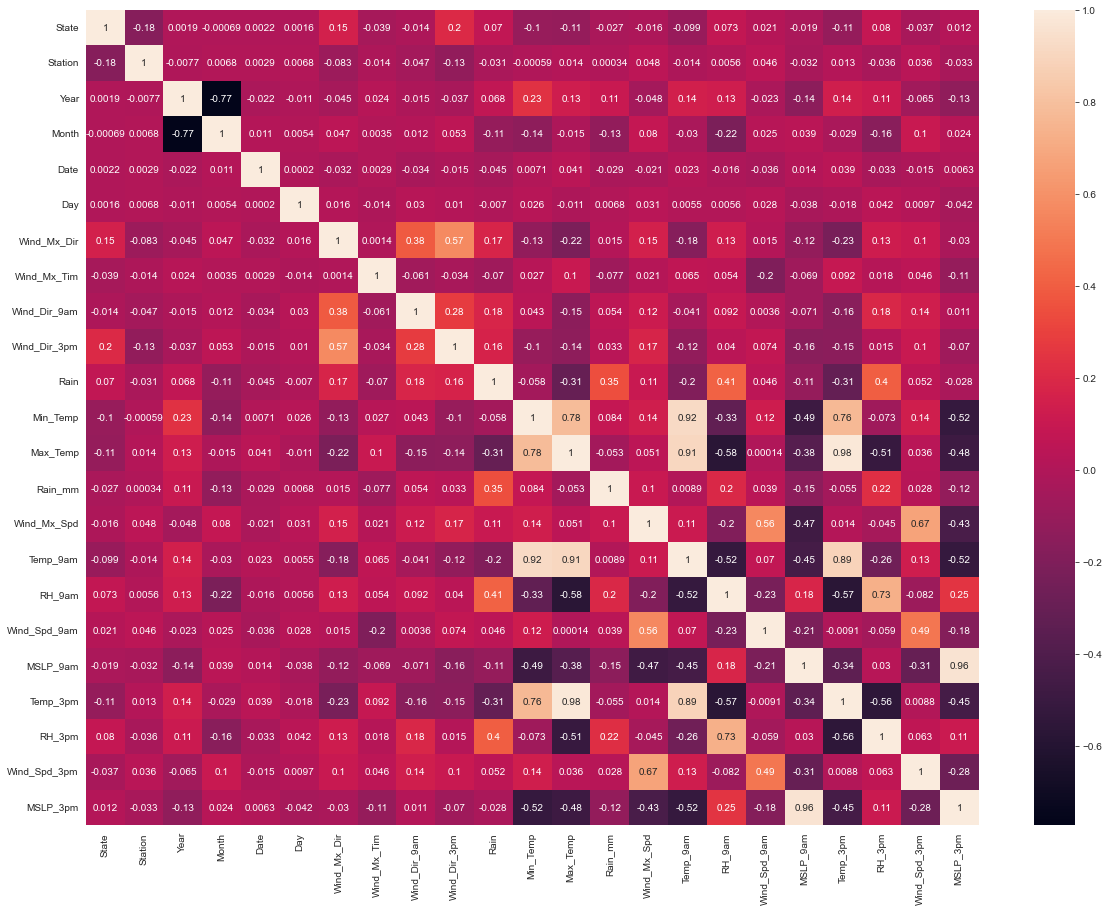


Figure , Heat map across all features

# Conclusion

In this project we were able to explore the data and discover the type of features we have, from quantitative to qualitative this later subdivided as a nominal or ordinal. From the start of the project, we had to identify missing values to properly assign correct data types to each feature. This step was essential to properly explore the distribution of the interval features.

From main statistics, we discover that the features had different units and ranges of values, also different level of spread across them with many outliers present. This made it clear that we had to normalize the data to allow the use of the different features across future predictive models.

Looking further at the missing values, we discover that the features “Cld\_9am” and “Cld\_3pm” had more than 50% of missing values in the entire data, I decided to remove those features from the data. At the same time, I removed “AVG\_Rain\_mm” as it only had one value, making it meaningless for a predictive model.

Looking at each of the stations, we discover that Albany and Newcastle had 8 and 9 features completely null making the values of those stations not reliable. I also removed those stations from the data, loosing 7.1% of samples but reducing majority of null values. For the remaining null values, we divided the data across states and input missing values with the median for interval features and the mode for the remaining.

The next step in the cleaning, normalized interval features and transformed non-numerical to numerical value on different stages for nominal and ordinal. With these transformations, we were able to use different functions to identify the most useful features from the data set: “Rain\_mm”, “RH\_9am” and “RH\_3pm”.