Weather prediction in Australia

Project report

Submitted to

DataScientest, France

to obtain the Certificate of Data Scientist

By

Abhishek Tiwari

Claudia Wiese

Shiksha Ajmera

DS-February 2024

21/03/2024

Contents - Report 1: exploration, data visualization and data pre-processing report

Rain prediction in Australia project

Report 1: exploration, data visualization and data pre-processing report

1. Introduction to the project

1.1 Overview

The following project aims at analysing the weather data in Australia from 2009 to 2017. We want to predict the incidence of rain, the temperature and wind speed. The dataset used contains 10 years of daily weather data from several locations across Australia.

1.2 Context

- Context of the project's integration into your business – The project could integrate the weather predictions into various business sectors, spanning from large-scale operations such as energy production from solar and wind sources to smaller-scale enterprises like food catering businesses and transportation services. Predictions derived from this project, including weather and traffic forecasts, hold significant economic implications. Rainfall and temperature is particularly interesting for the agricultural sector as well. Moreover, if wind prediction is included, it can impact flight services, contributing to both economic and safety considerations.

- From a technical point of view – Handling missing data (NaNs) and addressing dataset imbalance are crucial technical challenges in this project. The dataset contains numerous missing values and exhibits imbalanced distributions, posing significant obstacles to accurate analysis and prediction.

- From an economic point of view – As outlined in the context, the predictions generated by this project can influence the economic landscape by affecting both large and small businesses. Consequently, the accuracy and reliability of these predictions play a vital role in shaping economic decisions and strategies.

- From a scientific point of view – Weather prediction is a complex scientific endeavor influenced by various factors such as geography, terrain, temperature, humidity, and vegetation. Understanding and predicting such intricate phenomena is of scientific interest and significance. Especially rainfall and the temperature is interesting to predict because Australia is known for extensive droughts and extreme heat waves.

- General Info on the Climate in Australia – Due to its large size, Australia has a wide variety of climates. The largest part of the continent has a semi-arid

or desert-like climate. The southeast and southwest have more temperate climate and the northern part tropical climate. Most of our data comes from the temperate zones (See Annex, Figure 5, Map with locations).

1.3 Objectives

- What are the main objectives to be achieved? Describe in a few lines – The primary objective of the project is to utilize weather information to predict rain incidence, and secondarily the temperature and wind speed for the following day, and possibly the next week and month for the Australian locations in the dataset.

- For each member of the group, specify the level of expertise around the problem addressed? – Abhishek- Moderate experience, in dealing with weather data from his master's thesis with weather data as one of the focal points for a non-data science project. Claudia and Shiksha – no previous experience/expertise.

- Have you contacted business experts to refine the problem and the underlying models? If yes, detail the contribution of these interactions. – No.

- (Are you aware of a similar project within your company, or in your entourage? What is its progress? How has it helped you in the realization of your project? How does your project contribute to improving it?). – No.

2. Understanding and manipulation of data

2.1 Framework

- Which set(s) of data(s) did you use to achieve the objectives of your project?

  Rain in Australia: Weather in Australia, dataset for 10 years from 2007 to 2017 from numerous Australian weather stations.

- Are these data freely available? If not, who owns the data?

  The data is freely available at Kaggle under the following link: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

- Describe the volume of your dataset? – 145460 rows x 23 columns

- Description of columns in the dataset

Date	Date of observation
Location	The common name of the location of the weather station

MinTemp	The minimum temperature in degrees celsius
MaxTemp	The maximum temperature in degrees celsius
Rainfall	Amount of rainfall
Evaporation	The so-called Class A pan evaporation (mm) in the 24 hours to 9am
Sunshine	The number of hours of bright sunshine in the day.
WindGustDir	The direction of the strongest wind gust in the 24 hours to midnight
WindGustSpeed	The speed (km/h) of the strongest wind gust in the 24 hours to midnight
WindDir9am	Direction of the wind at 9am
WindSpeed3pm	Wind speed (km/hr) averaged over 10 minutes prior to 3pm
Humidity9am	Humidity (percent) at 9am
Humidity3pm	Humidity (percent) at 3pm
Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9am
Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 3pm
Cloud9am	Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
Cloud3pm	Fraction of sky obscured by cloud at 3pm. (measured in the same way as Cloud9am)
Temp9am	Temperature (degrees C) at 9am
Temp3pm	Temperature (degrees C) at 3pm
RainToday	Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
RainTomorrow	Boolean: 1 if yesterday's precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0

2.2 Relevance

- Which variables seem most relevant to you with regard to your objectives?

Based on correlation data the following variables, we found that he following variables are relevant:

- RainToday,
- RainTomorrow,
- Rainfall
- WindGustSpeed,
- WindSpeed9am,
- WindSpeed3pm,
- MinTemp,
- Max Temp
- Temp9am
- Temp3pm
- Humidity9am,
- Humidity3pm,
- Cloud9am,
- Cloud3pm,
- Pressure9am
- Presure3pm
- Sunshine
- Evaporation

For more details and explanations see Annex, Figure 3, Heatmap.

We also think the Date and Location columns are very important for our analysis because with the Date we can observe trends and seasonality over time.

We also keep Wind direction data. By transforming it, as explained in the pre-processing section, we think it could potentially be interesting for predictions of our target variables.

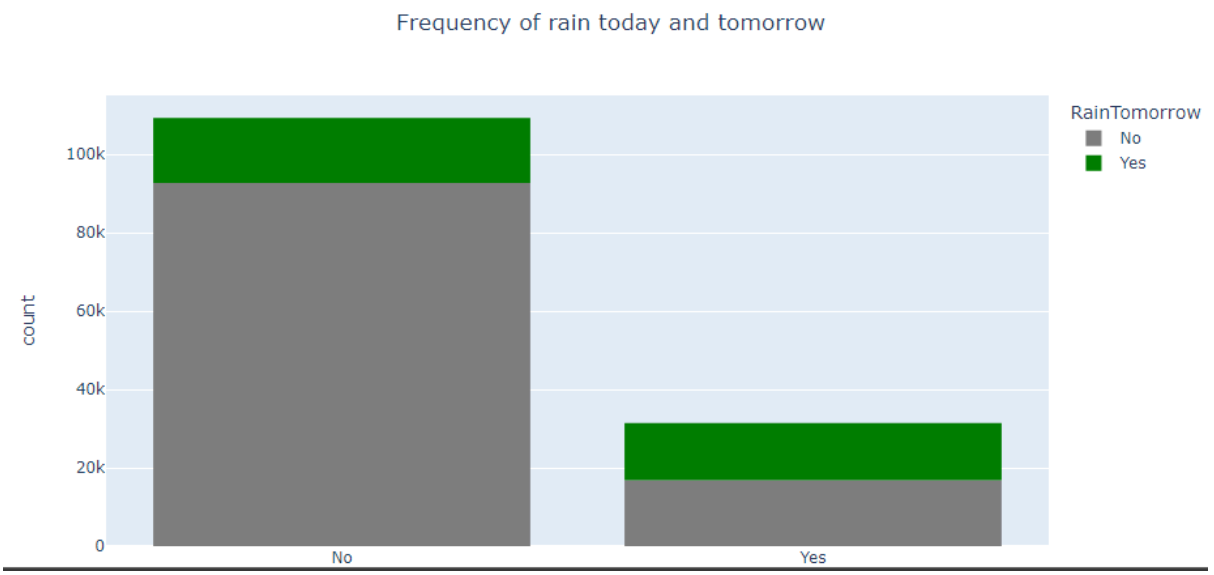
- What is the target variable?

Rain tomorrow is the primary target variable. It is a binary variable taking the values 0 if no rain and 1 if there is rain. If time permits we would also like to predict the Temperature (Mean temperature - created after preprocessing), and Wind speed (created after preprocessing) along with Rain for next week and month. Temperature and Wind Speed are both continuous variables.

- What features of your dataset can you highlight?

The data is special in several ways. First of all we have a time dimension, as we have daily weather data for 10 years and secondly we have data from 49 weather stations all across Australia (See Annex, Figure 5 Map with locations).

We have an imbalanced dataset when it comes to rainfall data (See Figure below)



Most of the distributions of our variables are not normal. Concerning our target variables, we can observe that the temperature variables (TempMax, TempMin, Temp9am, Temp3pm) seem to be distributed almost normally. Rainfall on the other hand is heavily skewed to the right and the Wind Speed variables (WindGustSpeed, WindSpeed9am, WindSpeed3am) are also slightly skewed to the right (See Annex, Figure 4 Distributions of variables).

We also have a lot of missing data for certain columns and for certain years as well as locations some data is non-existent, i.e. has not been collected (see limitations).

- Are you limited by some of your data?

  We have a lot of NaNs, i.e missing data. 4 columns really stick out: Sunshine, Evaporation, Cloud3am and Cloud9am (see Annex, Figure 1 Missing Values)

  The missing data is not missing at random, as it is missing for specific locations which limits us with respect to our strategies of filling the NaN values because most of the classic strategies such as deleting data listwise, using Mean/Mode/Median imputation assume that data is missing at random.

  Another problem is that for some years, data has simply not been collected for certain locations. So the data is not missing but none existing.

2.3 Pre-processing and feature engineering

- Did you have to clean and process the data? If yes, describe your treatment process

  Yes, we did have to clean the dataset with the following process:

Transforming Date Column, Adding Month and Year Columns

We had to transform the Date column into a DateTime column and then extract the Year as well as the Month from the column for further analysis. Extracting the month is useful to check for seasonality which is common when dealing with weather data.

Transforming RainToday and RainTomorrow Columns

We had to transform RainToday and RainTomorrow because initially these were object columns with the categories 'Yes' and 'No'. But to use them for our classification analysis the values have to be integers, 1 for 'Yes' and 0 for 'No'.

Transforming Wind Direction Columns

For better computation wind direction data at 9 am and 3 pm was removed, and the wind direction of the highest wind speed of the day was kept while changing its values from the "compass rose" to the "azimuthal system".

Adding Coordinates Columns

For our location data we added the coordinates, one column for the latitude and one column for the longitude. The coordinates can be useful for our KNN analysis.

Dropping data for 2007 and 2008

When doing our data audit we have seen that there is a lot of monthly data that is non existent, i.e. not been collected for 2007 and 2008. Additionally 2007 and 2008 do not have all locations and also fall into the Australia Millennium Drought Period in the 2000s, so we will drop these years from the dataset.

Dropping Locations

We drop all data of the location Uluru as it has only data for very few years and since it is in the middle of the desert while most of the other locations are close to the seaside we consider it as a location with extreme weather.

We also drop data for the locations Sydney, Melbourne and Perth because for these locations we have Sydney Airport, Melbourne Airport, Perth Airport data and in our data audit we checked that these locations which are basically the same locations have less missing values.

Since the missing data is very location dependent, i.e. not random at all we decided to drop the locations with the most missing data, i.e. if we dropped all NaN's these locations would no longer exist in the dataframe. This leaves us with data for 23 locations.

Dropping Outliers

We used the IQR method (interquartile method) to identify outliers. The interquartile (IQR) is the difference between Q1 and Q3 (the first quartile at 25% and the last quartile at 75% respectively). To detect outliers, all data points which fall below (Q1 – 1.5 * IQR) or above (Q3 + 1.5 *IQR) are considered as outliers. All the outliers seem to be realistic values, except some values for humidity.

We will drop all values where the humidity is zero because the theoretical humidity can not be zero. The other outliers are kept to ensure keeping the realistic and true data to avoid unnecessary bias.

Feature Reduction

We performed feature reduction by aggregating highly correlated columns, namely pressure, temperature, cloud and humidity, which were sampled at distinct times of the day (9am and 3pm). To address multicollinearity and streamline the dataset, we retained the mean values derived from both time points, ensuring a more parsimonious representation of the data while preserving the essential information encapsulated by these variables.

We computed the temperature fluctuation throughout the day by determining the difference between the minimum and maximum temperature values.
We also excluded the wind speed at distinct times of the day (9am and 3pm) to reduce the redundancy and to simplify the analysis.

Imputations

After getting rid of the locations with the most missing NaNs, we have less location dependent missing values and decided to deal with the remaining NaNs by using KNN imputation. KNN imputation or K-Nearest Neighbor Imputation works by finding the nearest neighbors to the missing observation and then imputing it with the K-nearest neighbor. It can be used when data is missing at random. The advantage of this imputation method is that it is easy to implement and doesn't make any assumptions about the data.

Over / Undersampling

To balance the data out we will certainly do some over or undersampling during the modelling process to make better predictions on rain data. We will compare different over and undersampling methods to see what works better in our case. We might start with SMOTE (Synthetic Minority Over-sampling Technique), which is an oversampling technique used to address class imbalance in datasets. SMOTE works by generating synthetic examples of the minority hence balancing the class distribution.

- Did you have to carry out normalization/standardization type transformations of your data? If yes, why?

Yes, because the normalization ensures that all features have the same scale, and we have many variables on different magnitudes than others, and it will prevent larger magnitude variables from dominating.

- Are you considering dimension reduction techniques in the modeling part? If yes, why?

  As mentioned previously in the feature reduction section, we are doing feature engineering. Briefly, this includes Feature Extraction, in which we are taking the mean of multiple columns and creating a new column; e.g., creating a new column of cloud from the mean of cloud9am and cloud3pm. For a complete list of features added and reduced please refer to the feature reduction section above.

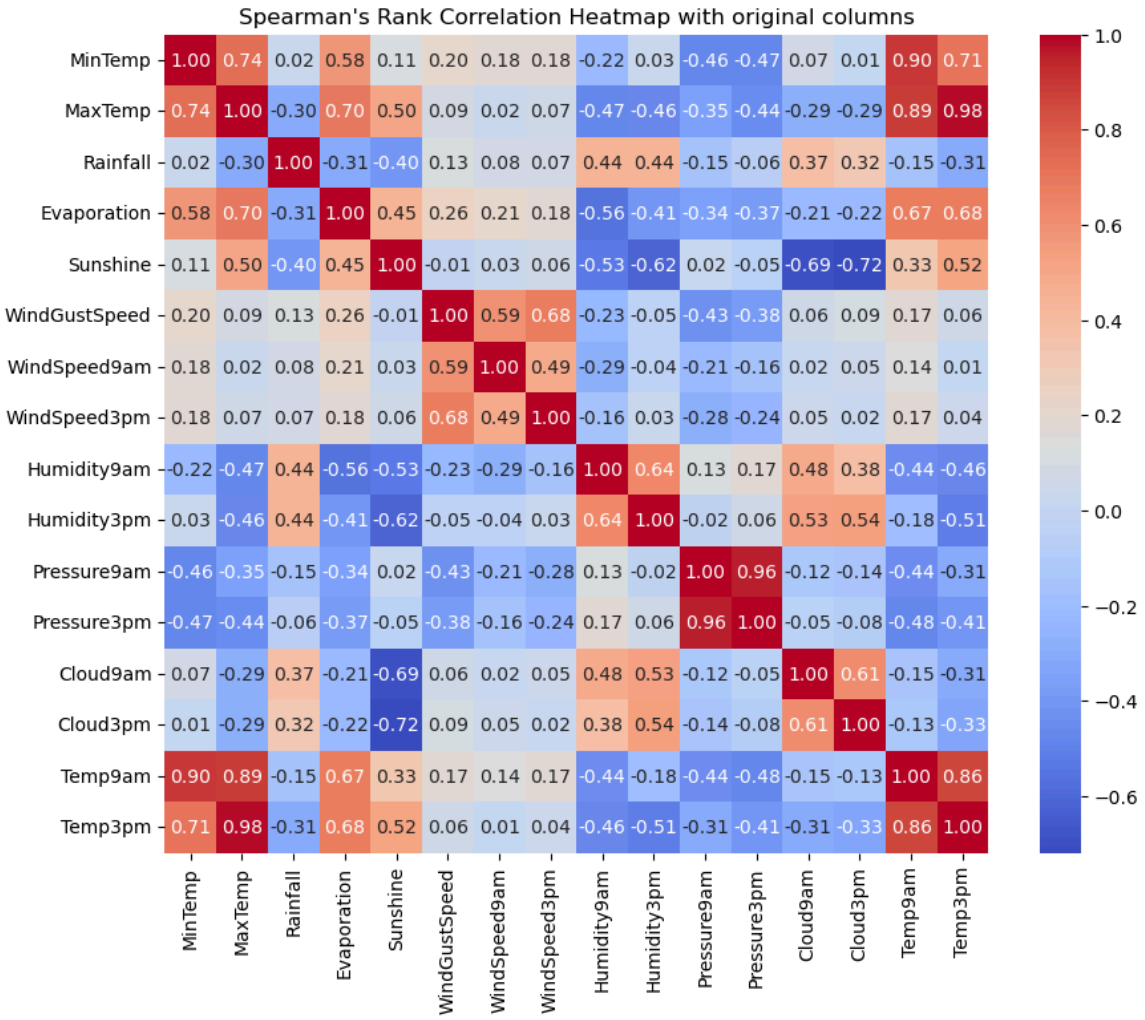
  After reduction, we get the following columns: Date, Year, Month, Location, Rainfall, Evaporation, Sunshine, WindGustSpeed, Latitude, Longitude, WindGustDir_angle, Cloud, Pressure, Humidity, Temp, temp_fluctuation, RainToday, and RainTomorrow.

2.4 Visualizations and Statistics

- Have you identified relationships between different variables? Between explanatory variables? and between your explanatory variables and the target(s)?

  Several variables including explanatory variables are correlated to each other. For exact correlation between all the target and feature variables we have attached a correlation heatmap with detailed explanations below (Figure 1, Heatmap).

Figure 1 - Heatmap showing the correlation between different features and target variables



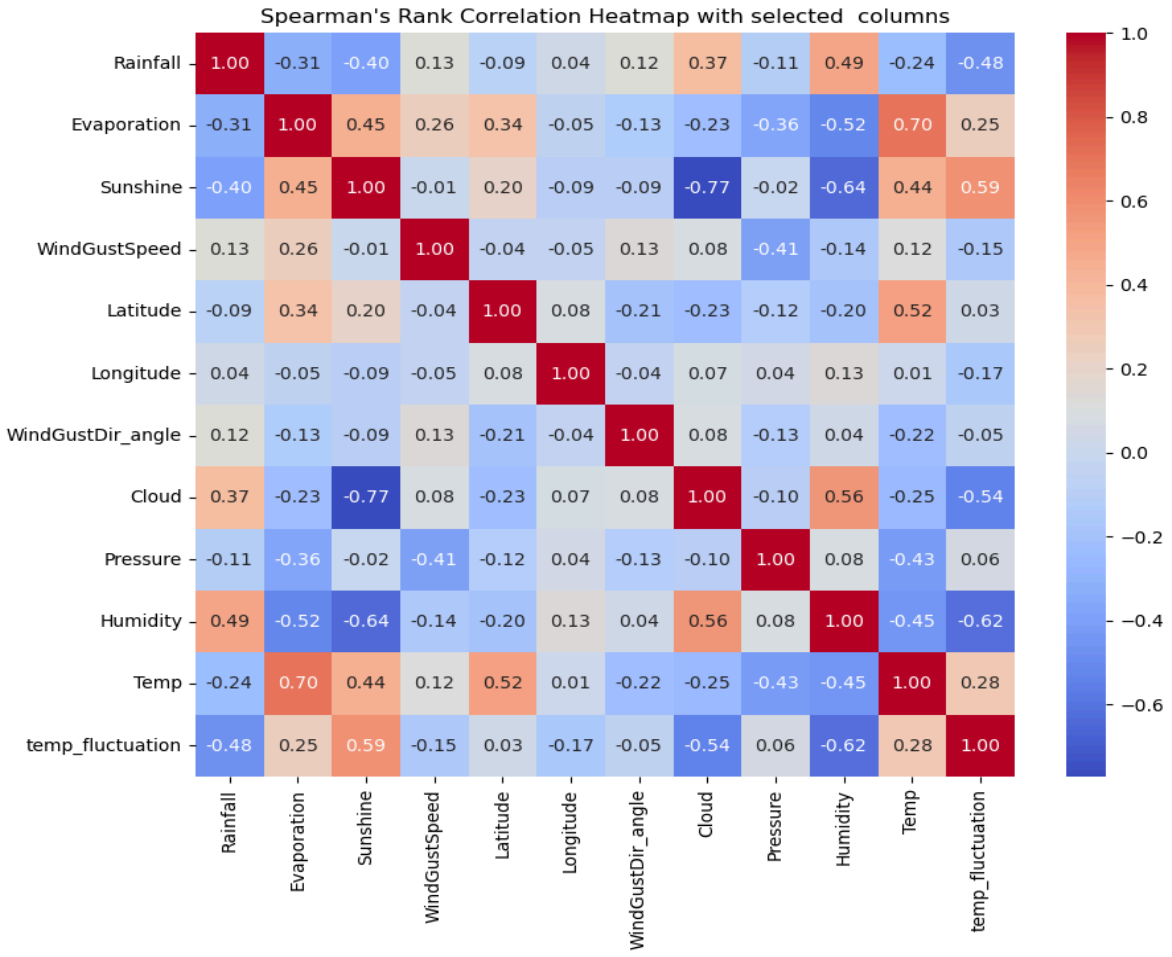
We tested the Spearman's rank correlation of different features with Rainfall on which the target RainTomorrow is based. Even without any feature engineering we can already see that humidity has a significant and the strongest positive correlation with the Rainfall. Cloud data seems to have a positive correlation with Rainfall as well and finally Sunshine also seems to have a quite strong negative correlation with Rainfall. These observed relationships are not very surprising as it seems quite logical that on rainy days there are more clouds and less sunshine.

When it comes to Temperature, it seems to be most strongly correlated with Evaporation, followed by Sunshine (both have positive correlations) and the humidity variables as well as pressure variables (all negative correlations)

Finally wind speed seems to be quite related to the pressure variables (negative correlation)

The heatmap also clearly shows the high correlation within the different temperature variables, the cloud variables, pressure variables and wind speed variables. That is why we fused these variables to avoid multicollinearity as explained in the data engineering section.

Here is the heatmap showing correlations after feature engineering:



As we can see the correlations we observed before are pretty much still the same.

Most interestingly, we can see a quite strong correlation between temperature and Latitude which probably can be explained by the different climate zones in Australia as explained in the context section above. These different climate zones go hand in hand with different temperatures.

- Describe the distribution of these data, distribution, outliers.. (pre/post processing if necessary)

We have identified 3 outliers in humidity (value 0) which is an implausible value, hence dropped. Other outliers are kept as they seem plausible data points with imbalance in the dataset as the identified reason for categorizing them as outliers, so they are kept and used in the modelling.

For more details on the distribution of the data see Annex, Figure 4, Distributions of variables.

- Present the statistical analysis used to confirm the information present on the graphs.

  The graphs for our analysis can all be seen in the annex. For the correlation of all the variables (See Annex, Figure 3 - Heatmap) we tested them statistically using Spearman correlation and Mann-whitneyU test. They were all significant with p-values very close to 0.

- Draw conclusions from the elements noted above allowing them to project themselves into the modeling part

  As explained before the biggest challenge for this data set was to handle the missing values. The Sunshine variable by far exceeds all other variables with 48% missing data, followed by Evaporation with 43% and the Cloud data with around 40%. We handled the missing values in the preprocessing step using the knn method. In this step we have also engineered some new features and dropped the non-correlated features to clean up our data.

  We are not able to make predictions of rain, temperature and wind speed data for all locations present in the original dataset because we had to drop a lot of locations since data was either missing or simply non existent for these locations.

  Another major challenge is the imbalance in our dataset. This bar plot shows the imbalance of the data set when it comes to rain data. There are a lot more dry days than rainy days – before modelling we would need to use oversampling or undersampling methods.

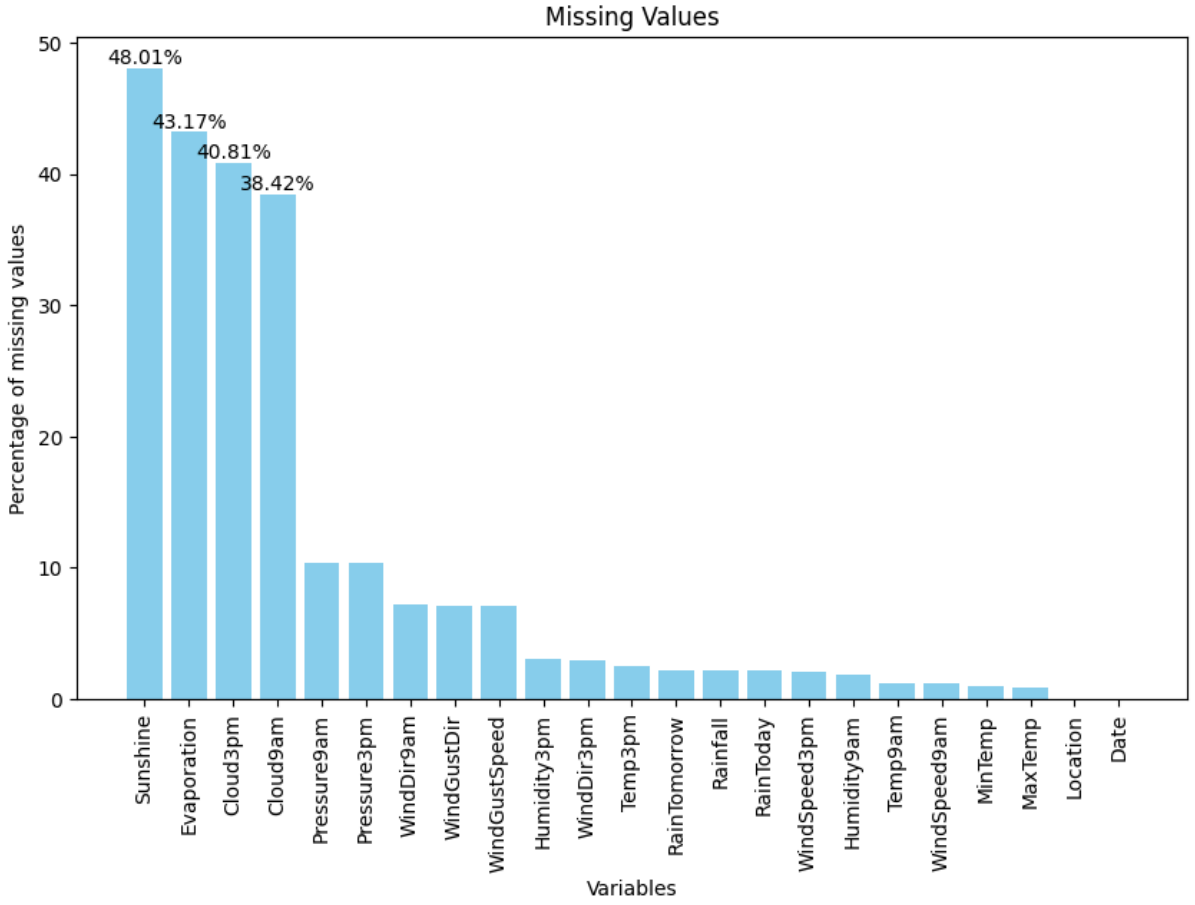
  For the modelling itself, we will predict our target RainTomorrow with a classification model such as XGBoost as it is a binary variable. Here the challenge will be to handle the imbalance of the dataset by using over/under sampling strategies.

  If we have time to work on temperature and wind predictions, we will likely use some advanced regression models.

  We also have to take into consideration the time dimension of the data and handle this by using models specific for time series.

2.5 Annex

Figure 1 - Missing Values



The Sunshine variable by far exceeds all other variables with 48% missing data, followed by Evaporation with 43% and the Cloud data with around 40%.

Figure 2 - Imbalanced Rain Data

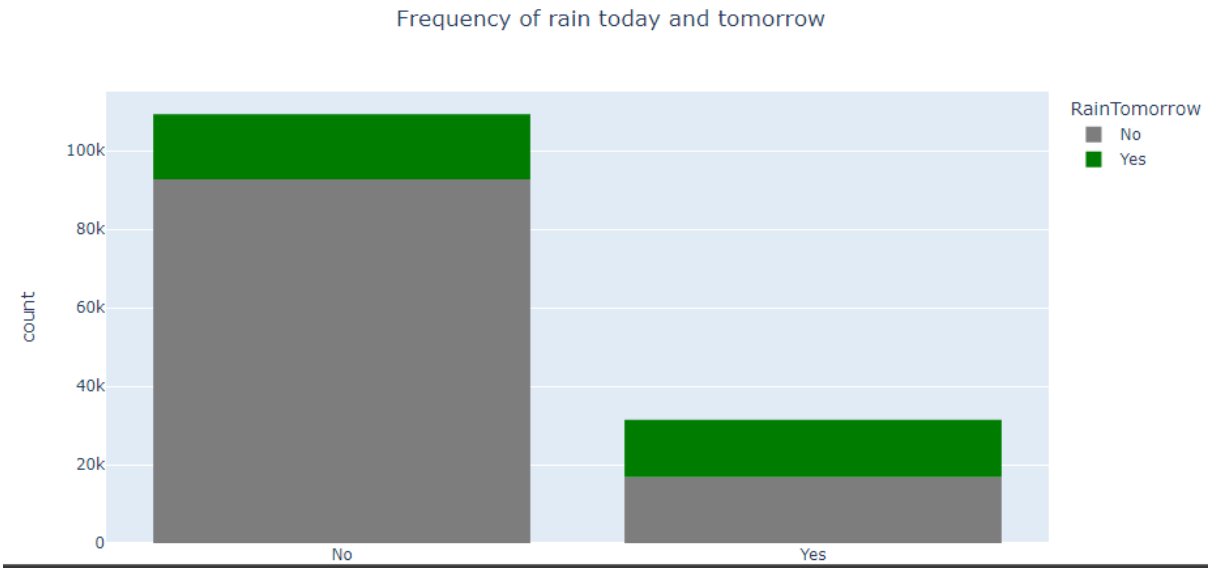
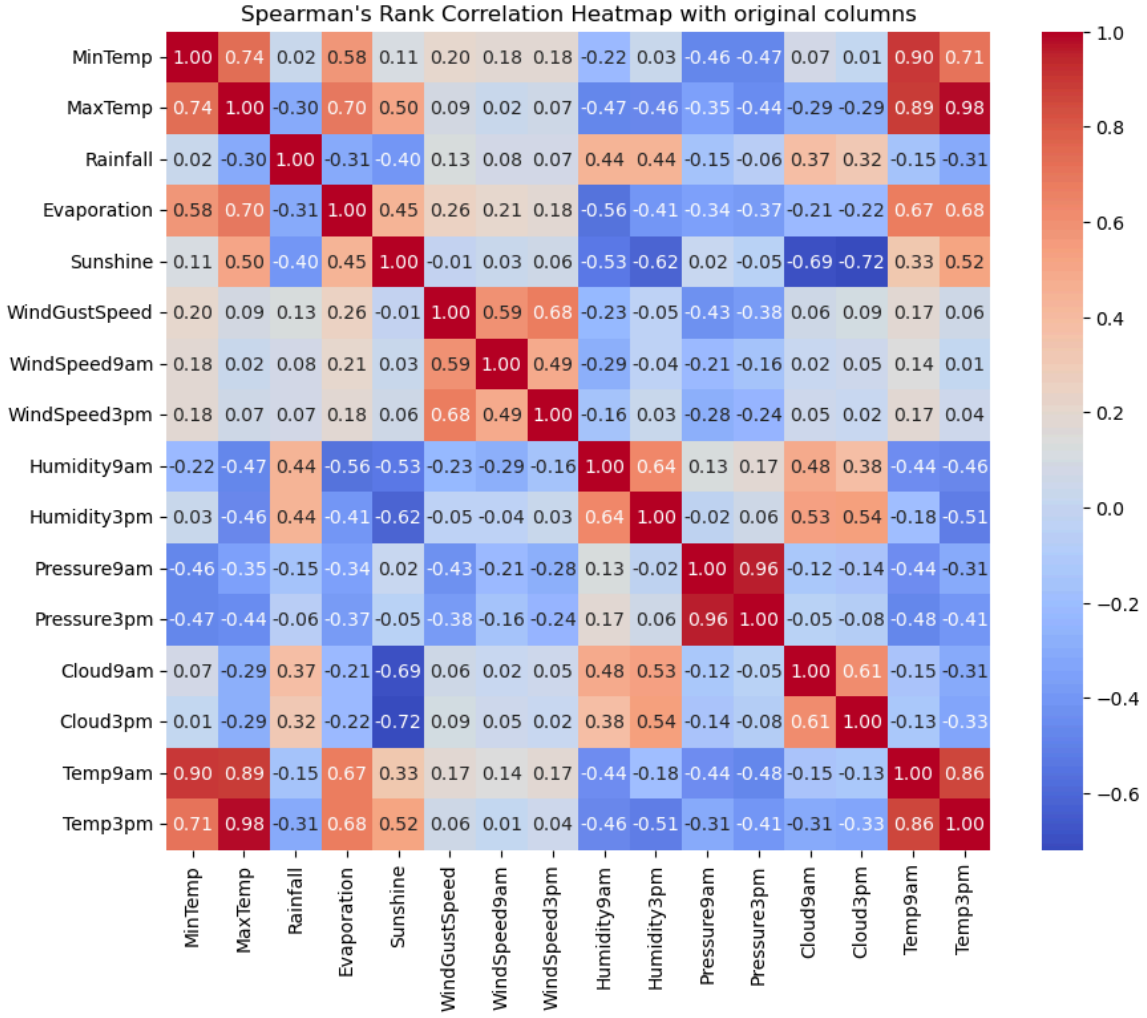


Figure 3 - Heat Map



Here is the heatmap showing correlations after feature engineering:

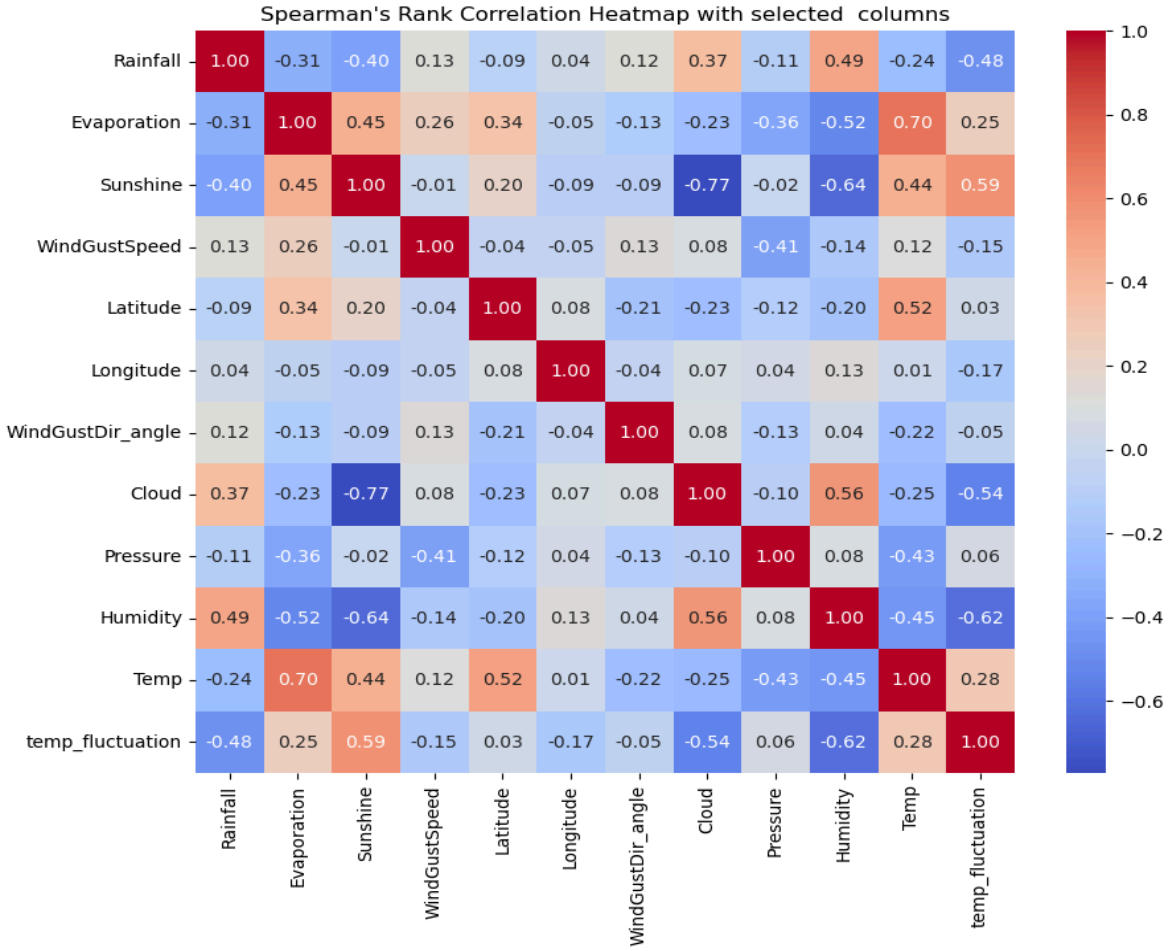
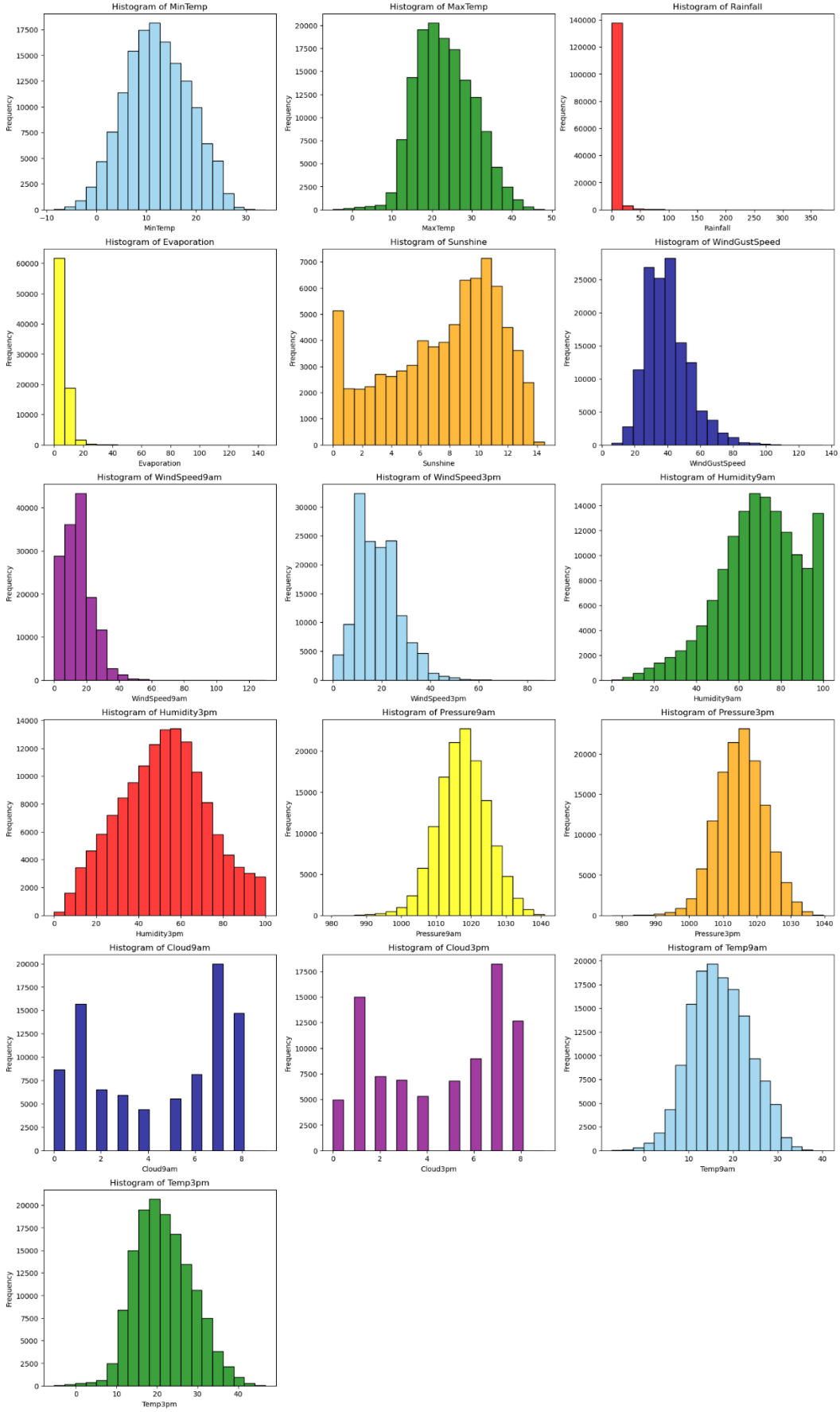


Figure 4 - Distributions of variables



From these graphs we can conclude that most of the data is not normally distributed and hence we will be using non-parametric statistical tests to check for correlations between the variables and any other statistical tests which would be needed later in the study.

Figure 5 - Map with locations

The geographical distribution of the data collection sites is shown above. We can see that most of the locations are in the Southeast. Since the locations are mostly clustered together and have neighbours, it is therefore possible to use the KNN method to handle the missing values.