UCDPA Project Report

**Building a classification model to predict rainfall in Australia**

Date: 20 Dec 2021 Submitted by: Shoumik Goswami

## GitHub URL

<https://github.com/shoumikgoswami/UCDPA_shoumikgoswami/tree/main/Rainfall%20prediction>

## Abstract

This project has been done as a part of project submission for UCD Specialist Certificate in Data Analytics. The objective of the project is to Predict next-day rain by training classification models on the target variable RainTomorrow. The project uses an open weather dataset from Kaggle and dives deep into how the weather changes in Australia per location. After a comprehensive exploratory analysis of the data, a number of transformations and data imputations are done to ensure the data is fit for modeling. We also observed slight class imbalance in the dataset due to which the model accuracy gets impacted at the end of the analysis. The implementation uses functions, lists, label encoding, scaling and grid search to find the optimal model and parameters.

## Introduction

Climate AI is an interesting and trending topic currently where rainfall prediction is one of the most common use cases. There are different approaches to predict rainfall but often it’s a combination of a timeline-based approach supported by other weather factors such as humidity, wind, temperature etc. The dataset contains weather data of different locations in Australia and records rainfall information per day for 10 years. While the dataset size is good enough, there are complications with the data that needs to be dealt with to create a good model.

The approach taken in this project is to consider this problem statement as a binary classification problem where the goal is to predict whether it will rain the next day with respect to the information recorded for the current date. The answer could be Yes or No i.e., 2 classes.

The project also uses a CSV based approach of uploading data for data analysis. This has been done specifically to meet the course requirements of fetching data using multiple techniques.

## Dataset

## The data has been fetched from Kaggle - <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>

This dataset contains about 10 years of daily weather observations from many locations across Australia. RainTomorrow is the target variable to predict. It means - did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.

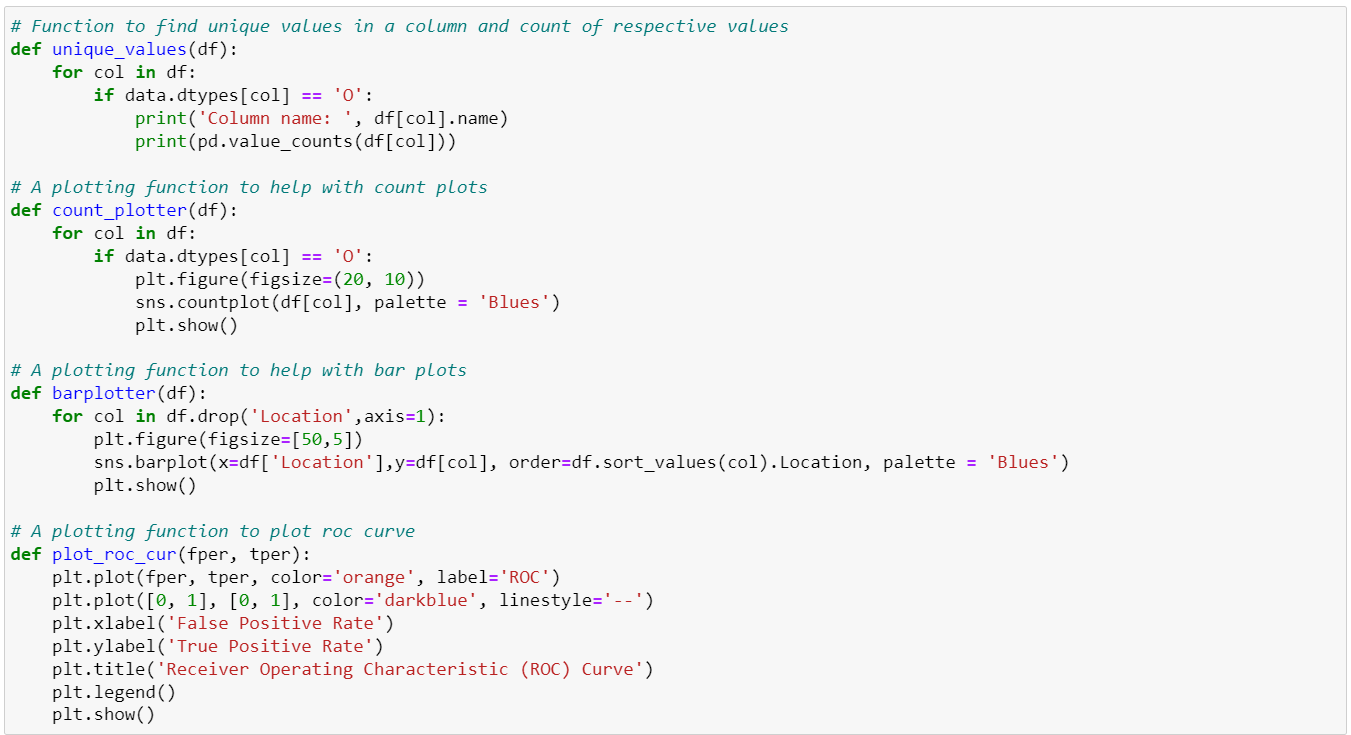
Source & Acknowledgements Observations were drawn from numerous weather stations.

The daily observations are available from <http://www.bom.gov.au/climate/data>. An example of latest weather observations in Canberra: <http://www.bom.gov.au/climate/dwo/IDCJDW2801.latest.shtml>  
Definitions adapted from <http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml>   
Data source: <http://www.bom.gov.au/climate/dwo/> and <http://www.bom.gov.au/climate/data>.  
Copyright Commonwealth of Australia 2010, Bureau of Meteorology.

## Implementation Process

The analysis has been performed in the below steps –

**Analysis pipeline - the OSEMN approach**

* Obtain the data  
  The data is obtained from Kaggle Rainfall prediction dataset. The dataset is downloaded as a CSV file and loaded into Jupyter notebook for the analysis. The dataset contains different types of data starting from datetime to location to weather attributes. This dataset contains about 10 years of daily weather observations from many locations across Australia. RainTomorrow is the target variable to predict. It means - did it rain the next day, Yes or No? This column is Yes if the rain for that day was 1mm or more.
* Scrubbing / Cleaning the data  
  Once the data is loaded, we check the information in the dataset. The dataset has 145460 entries, 1 entry per date per location and there are 23 columns, each column attributing to a specific weather information with a combination of date, geographical, categorical and numerical variables. There is missing information in the columns as well. We also check the percentage of missing information in the dataset, and columns Sunshine, Evaporation, Cloud 3pm and Cloud 9am have more than 30% of the data missing.   
    
  **Helper functions –** In order to make the analysis easier, we also create a number of helper functions -   
  
* Exploring / Visualizing our data  
  To understand the dataset we do comprehensive EDA. First, we check the unique values of all categorical columns in bulk. Then we run a couple of count plots to understand the distribution of data across each category. Since we have around 49 locations, we created a vertical line plot to check the distribution of entries per location.   
  After this, we try to understand the weather patterns over 10 years by creating trend lines against each of the weather attributes. We created trend plots for temperature, humidity, wind speed, pressure and cloud levels. We also observed how these attributes changed at different times of the day i.e., 9am and 3pm.   
  Then we created few bar plots to observe the average weather conditions per location, this helped identify various locations which are either very cold or very hot, very windy or very wet.   
  Finally, we created a correlation heatmap to see which all variables are correlated to one another and the strength and direction of correlation. This will help us identify attributes which we can drop for the model.
* Modeling the data  
  Before we create a model, we will prepare the data by following few steps.   
  First, we drop the attributes which have more than 30% null values - 'Sunshine', 'Evaporation', 'Cloud9am', 'Cloud3pm'. Additionally, we drop the Date column as we are more interested in doing a classification and not a timeseries model. We fetch the list of columns having only numerical values and categorical values respectively.   
  - The missing numerical values are replaced with the median of the respective attribute.   
  - The missing categorical values are replaced with the most frequently occurring value i.e., mode of the respective column.   
  Once the dataset is cleaned and missing values replaced, we convert the categorical values into numerical values by using label encoder. After this, we scale all the values based on the minimum and maximum values of the columns to normalize the data.   
  Now the data prep is complete and we separate the target variable i.e., RainTomorrow from the complete dataset. Then we separate the dataset into train and test dataset with 25% split.   
  We define 5 models for consideration i.e., 'AdaBoostClassifier', 'GradientBoostingClassifier', 'RandomForestClassifier', 'KNeighborsClassifier', 'LogisticRegressionClassifier'. We defined the parameters for each of these models and then use Grid Search Cross validation to identify the best model with optimal hyperparameters.   
  Based on the search, we find that Random Forest scores the highest in terms of accuracy and we proceed to creating the model using Random Forest algorithm. We fit the dataset to the model and do the predictions.
* iNterpreting the results  
  Once the model is created, we check the accuracy of the model and found it to be ~85%. Additionally, we check the ROC scores and classification report. The precision and recall scores are good for days predicting no rainfall. The scores are very low for days predicting rainfall, this is due to an imbalanced dataset. The F1-scores are good for days predicting no rainfall (0.91) while it is very low for days predicting rainfall (0.598). Based on this, the model will predict most days as days with no rainfall.

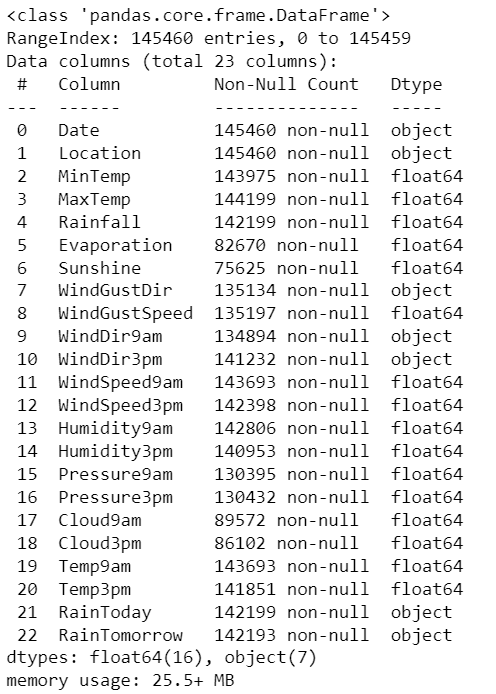
We also plot the ROC curve and looking at the ROC curve, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives.   
We also plot the confusion matrix and looking at the confusion matrix, the model will predict no rainfall days 75% of the time correctly however, it will only predict rainfall days 11% of the time. It will predict rainfall days as no rainfall days 33% of the time making them as incorrect predictions.  
Finally we save the model for further use.

**Environment set-up and loading dependencies**

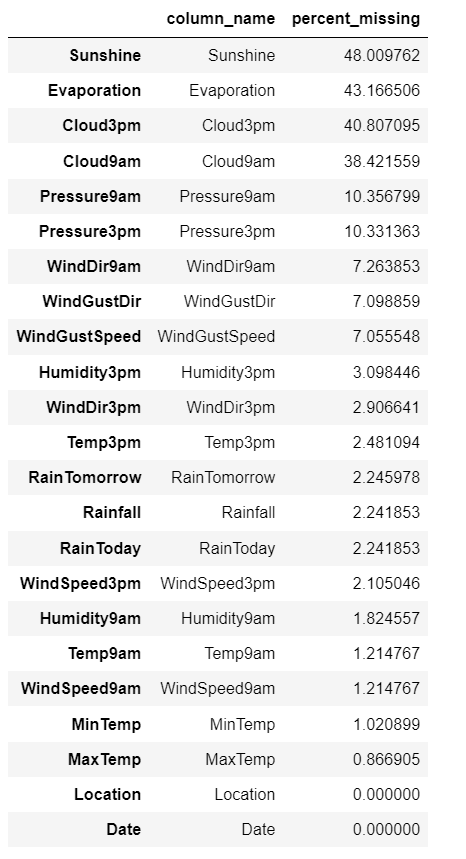
Jupyter notebook is used to do the analysis and Github is used to version the changes.

## Results

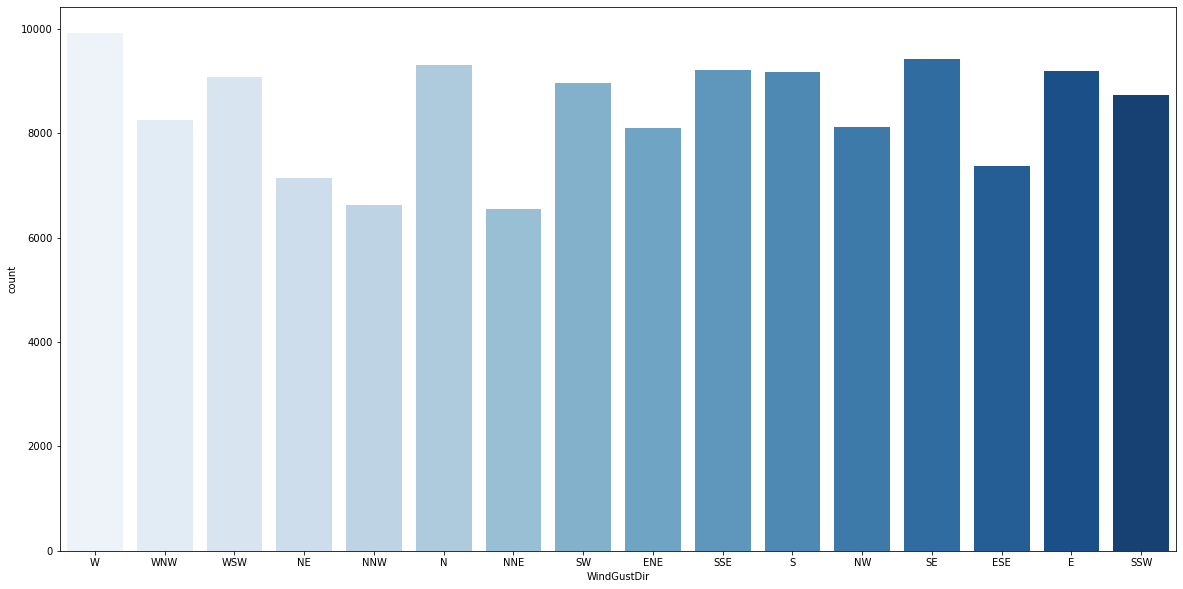
Below are the results based on the analysis

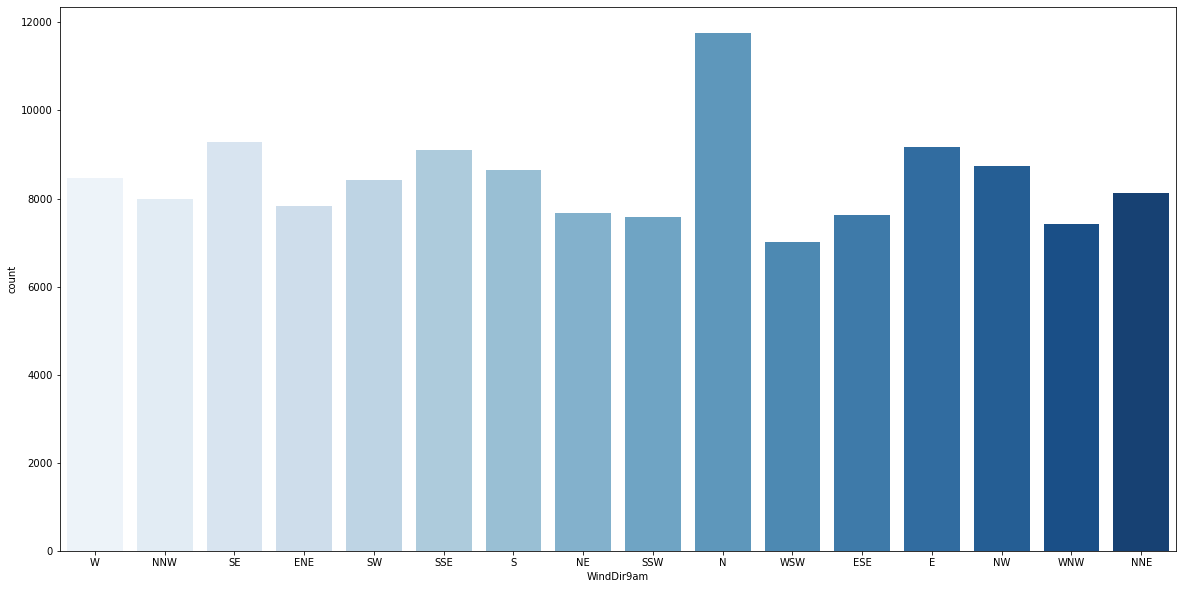


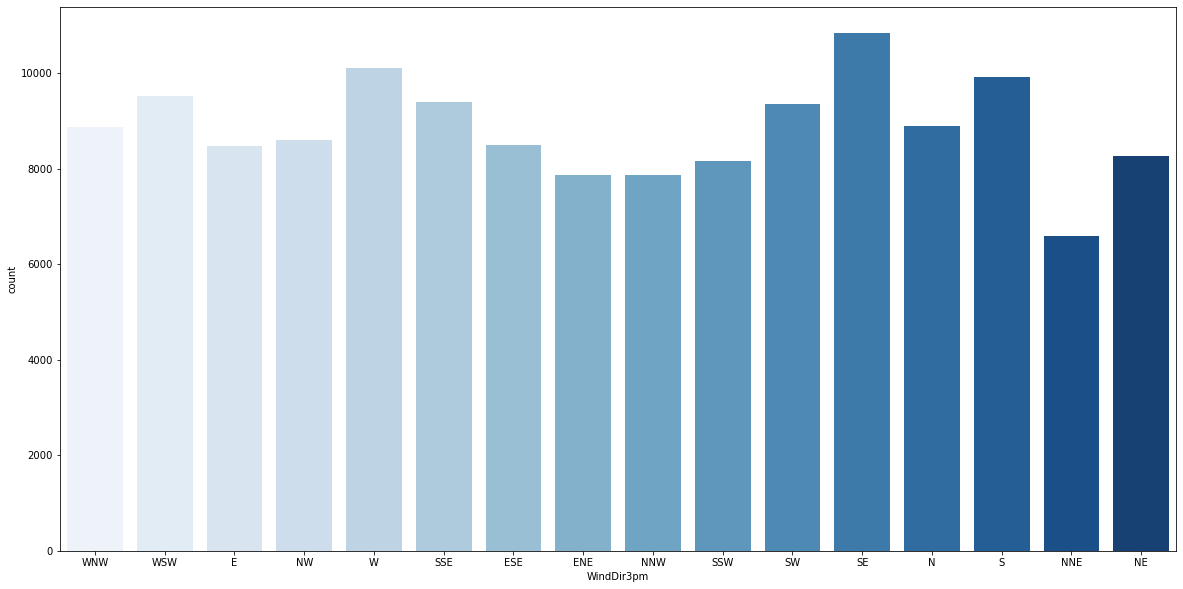
The dataset has 23 columns with a combination of date, geographical, categorical and numerical variables. There are missing information in the columns.

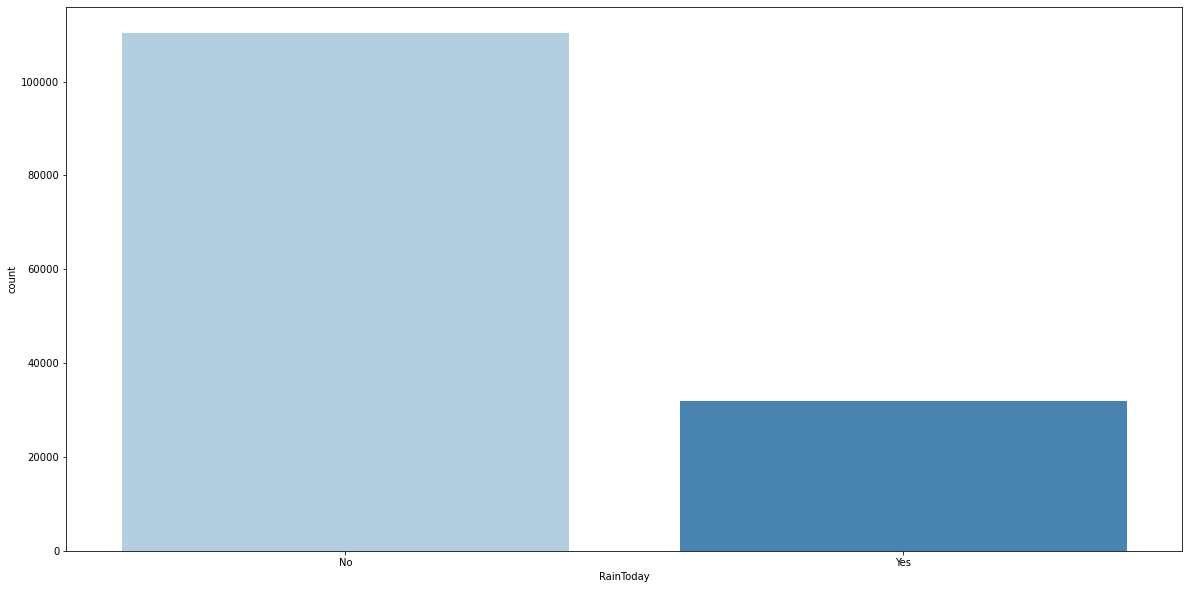


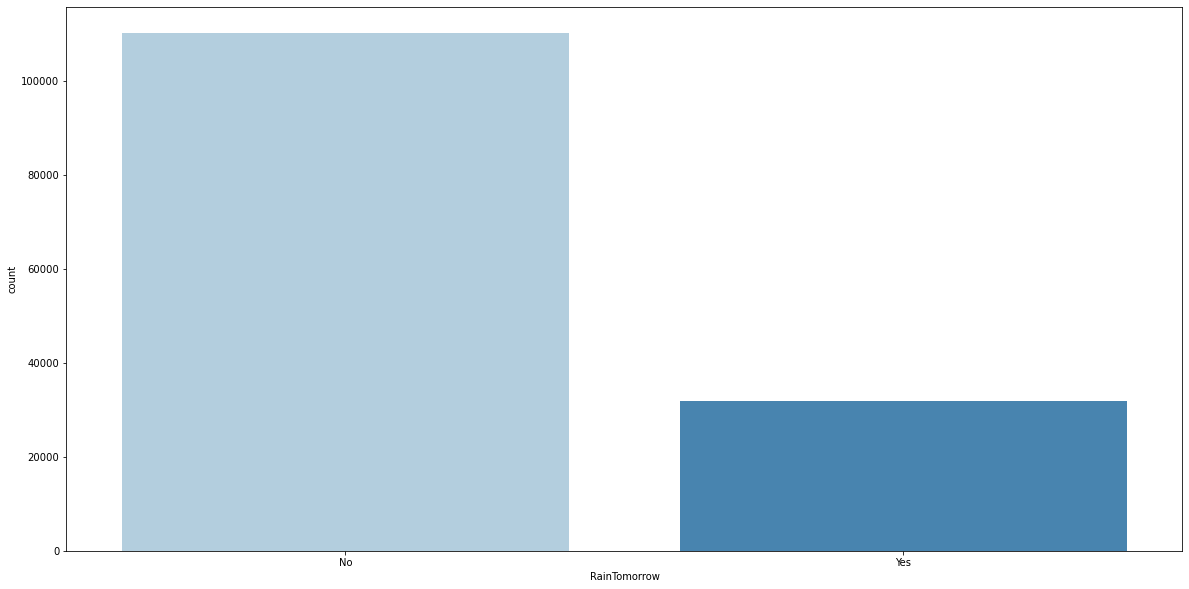
Sunshine, Evaporation have more than 40% of values missing in the data, followed by Cloud3pm and Cloud9am. We will drop these variables in the next steps as they have a lot of missing information.

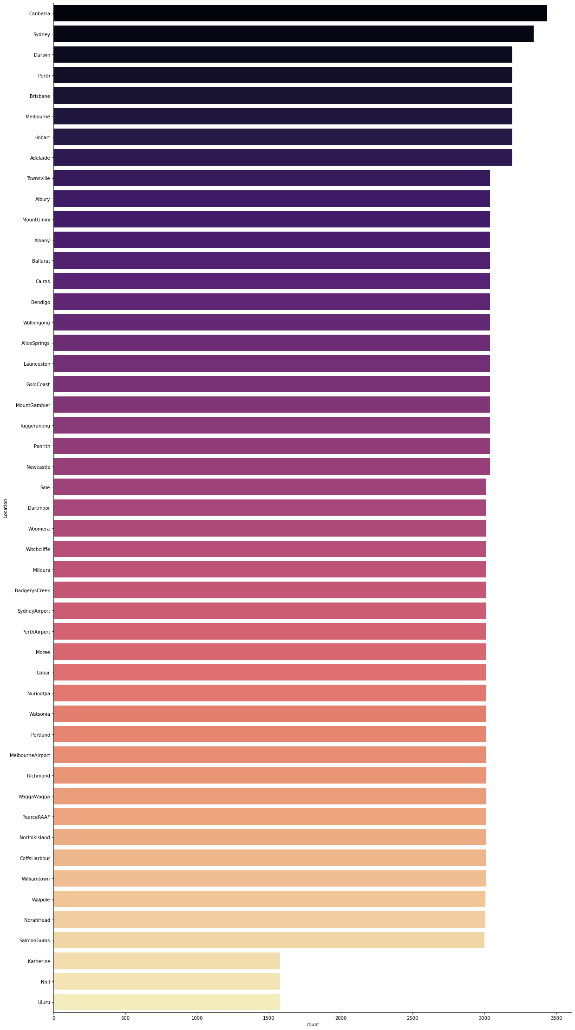
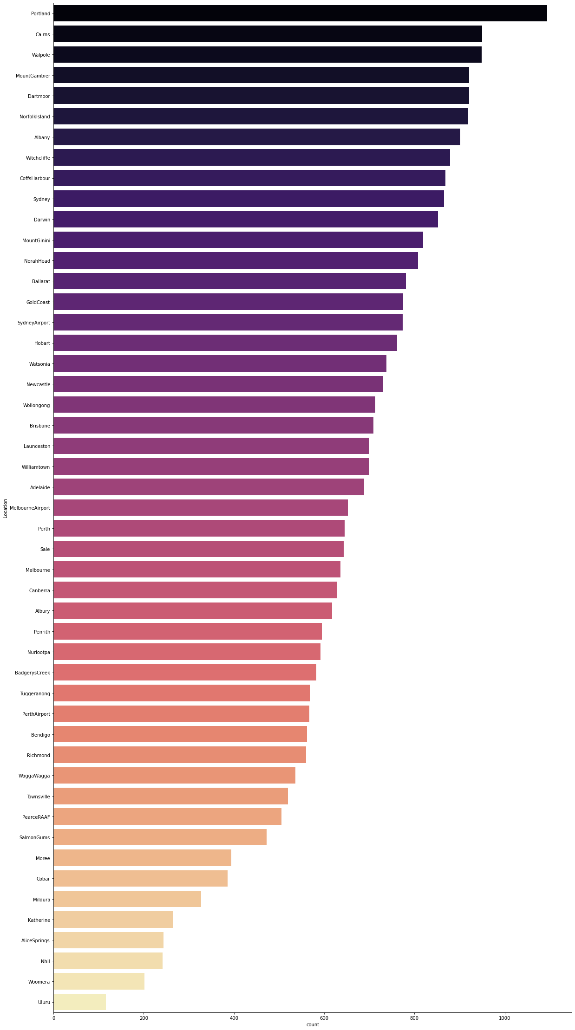


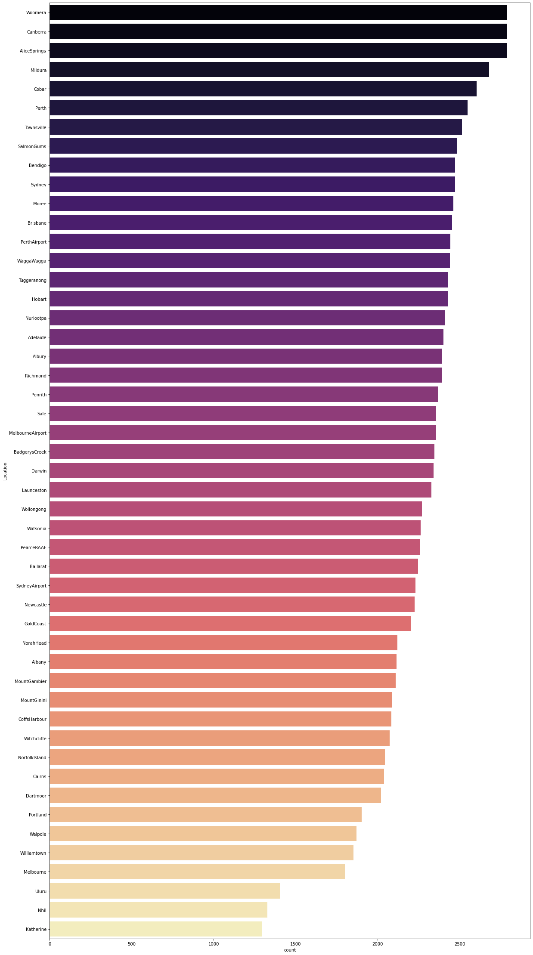






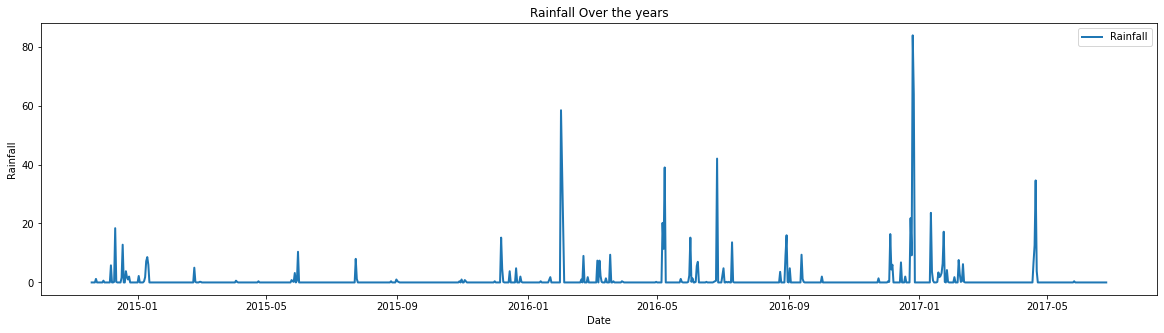


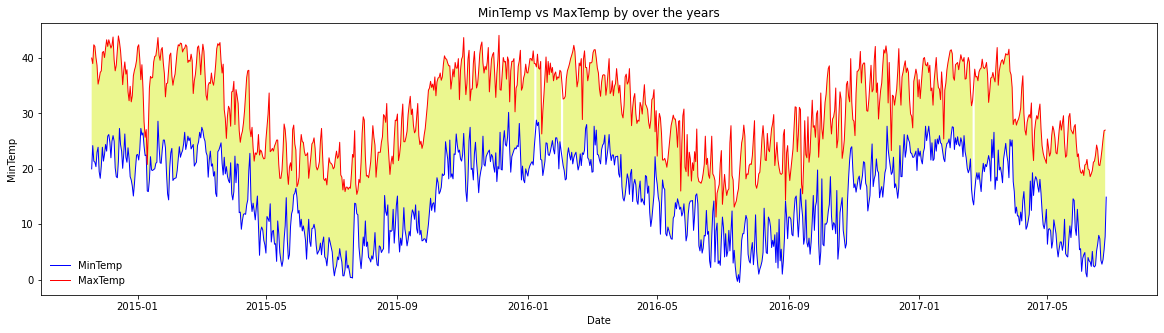


There are 49 unique locations in the dataset, top 5 locations are Canberra, Sydney, Darwin, Perth and Brisbane. Portland, Caims and Walpole receives the highest rainfall on a frequent basis. Woomera, Canberra, Alice Springs receive the lowest rainfall on a frequent basis. West and South are the directions where most of the wind gust direction data is.

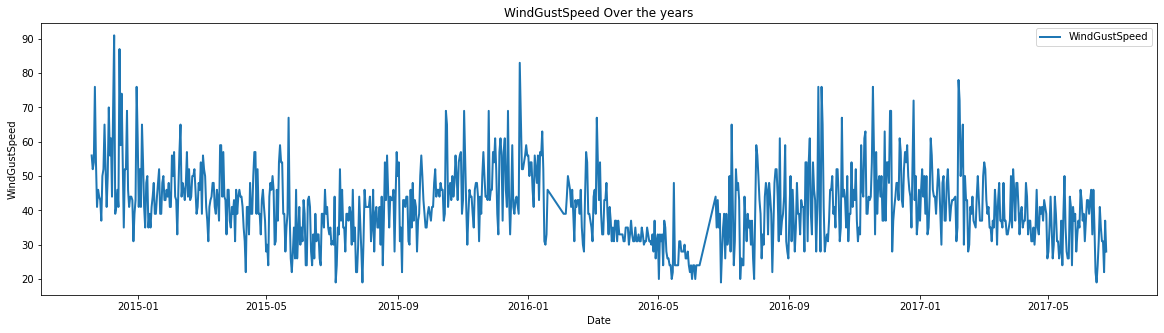
Wind direction at 9am is mostly from the North, while wind direction at 3pm changes towards South East. It seems the direction of wind changes from north to south during the day. Most of the time, there was no rain during the observation period (78%). The next day also observed lower rainfall.

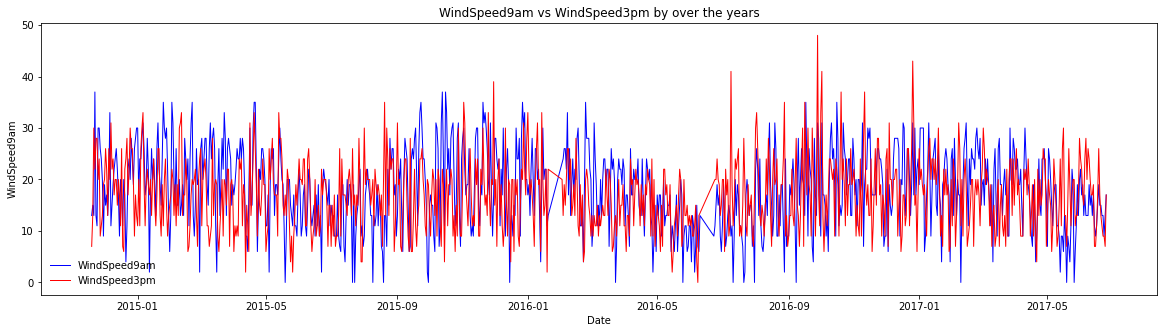


Australia receives maximum rainfall in the beginning of the year, particularly between January to May, highest in January. Frequency of rainfall has increased from 2015 to 2017, with the highest rainfall in January 2017. Minimal rainfall is experienced during the rest of the year.

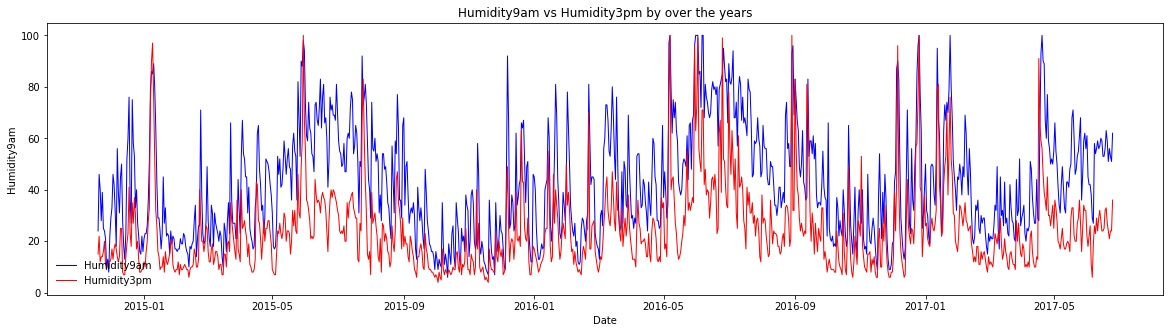


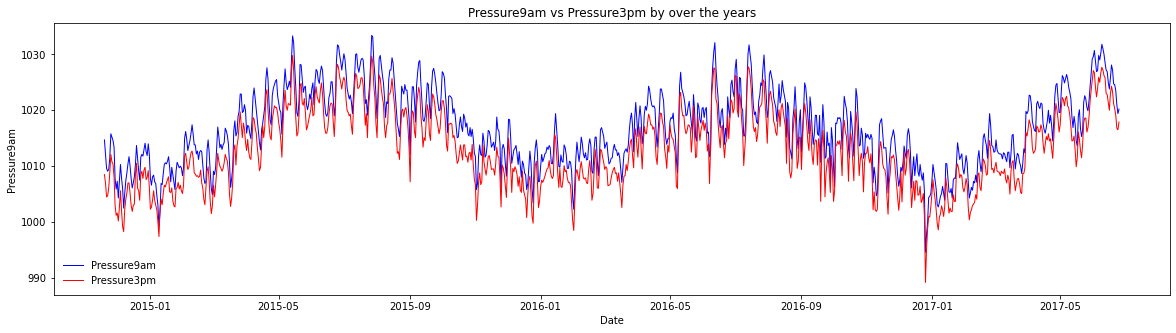
Temperature seems to go up in the first quarter of the year (Jan to March) and starts falling after April with the lowest in August. Australia experiences winters in the month of July to October and this can be seen from the plot. The max and min temperature has remained somewhat constant over the years without changing much.



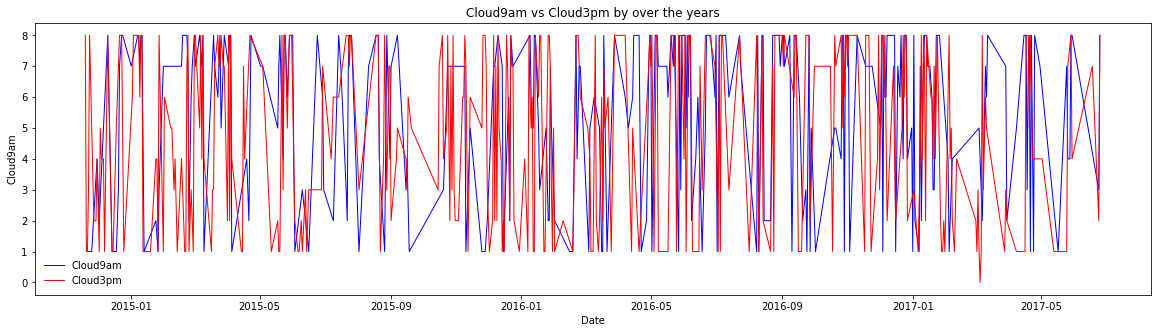


Australia experiences faster wind speeds in the summer periods with highest wind speeds in December-January period. Wind speeds usually come down during the winter season with lowest in June-July period. Overall wind speeds have declined over the observation period. Wind speeds are usually higher at 3pm compared to 9am. Afternoon wind speeds have increased over the observation period.

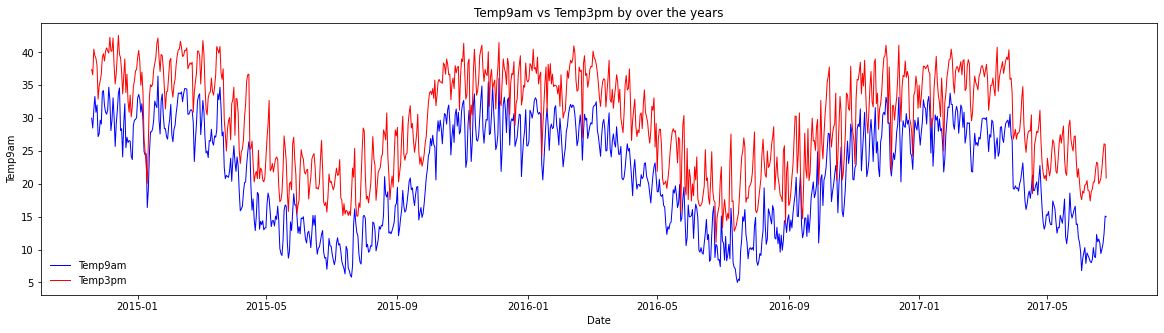




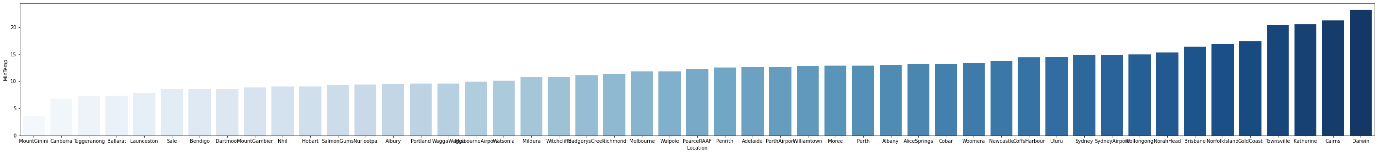
Humidity levels have remained constant over the observation period with 3pm humidity spiking in the winter months. Pressure levels have remained constant over the observation period with humidity spiking in the winter months.

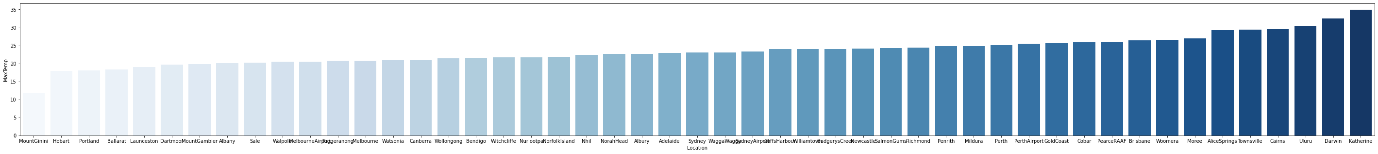


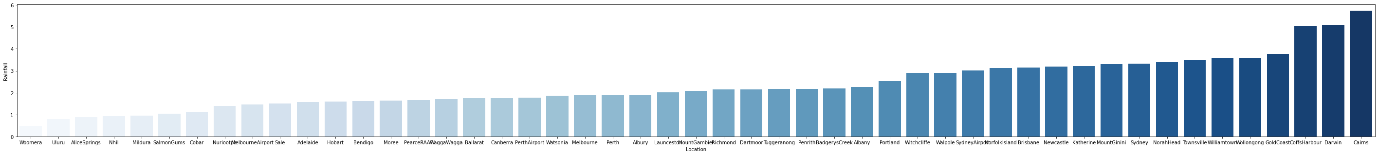
The Cloud information does not present any useful information based on the patterns, so no insights can be developed. As a result, we will drop the cloud attributes to improve the dataset.

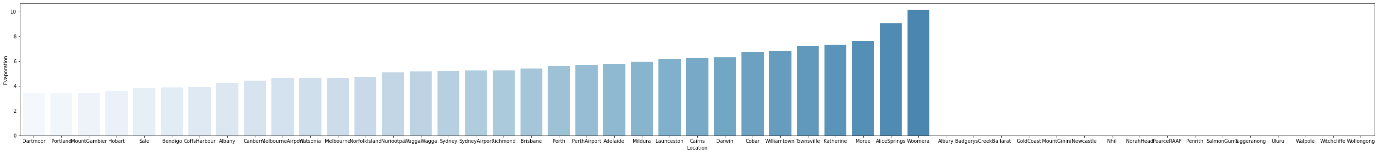


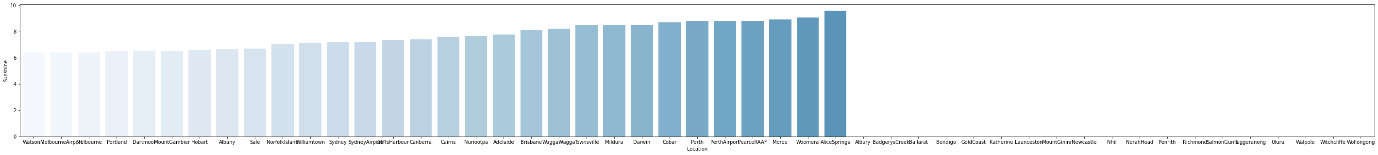
Temperature levels at 9am and 3pm have remained constant over the observation period. Afternoons are usually hotter, peaking in the summer months of Jan to Mar.

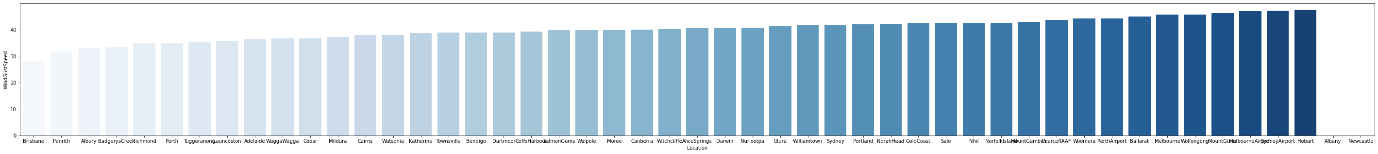


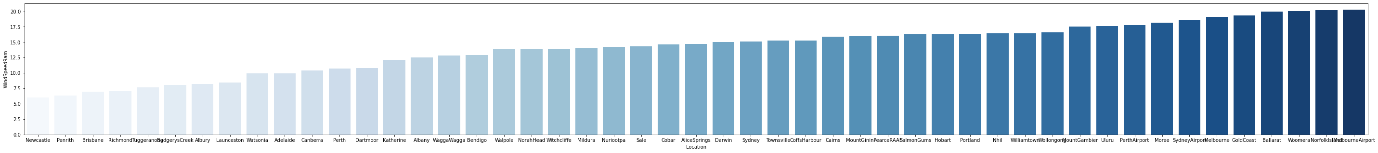


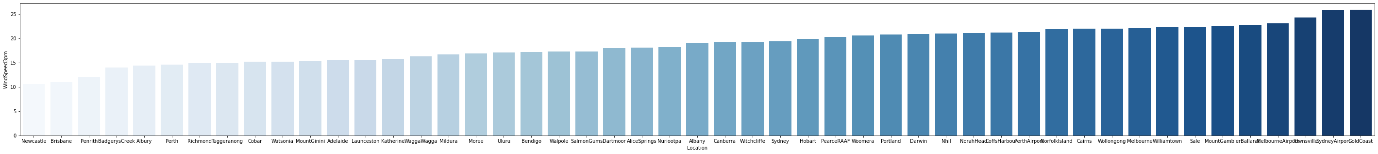


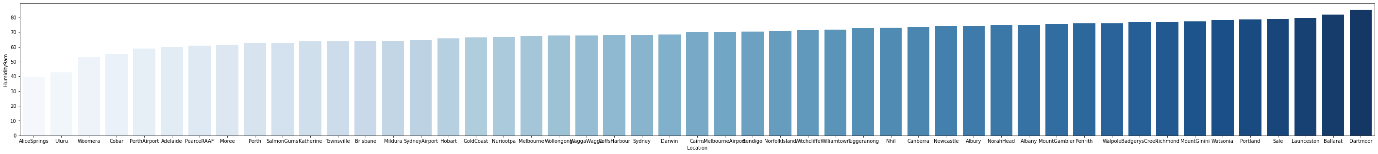


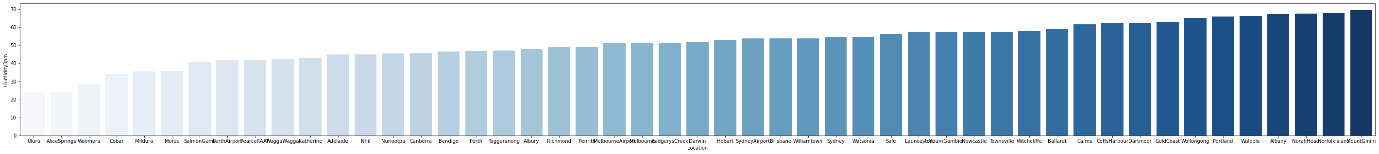


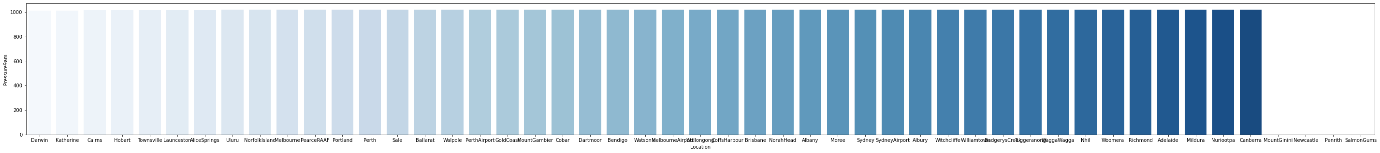


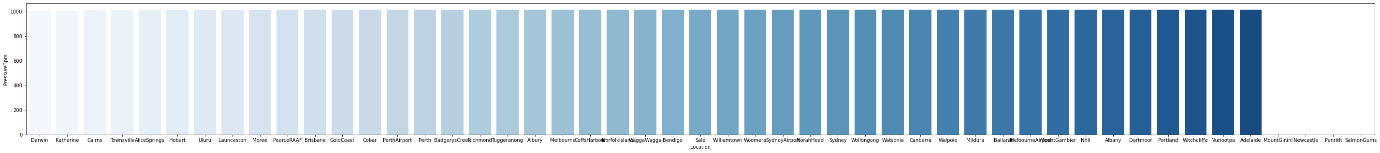


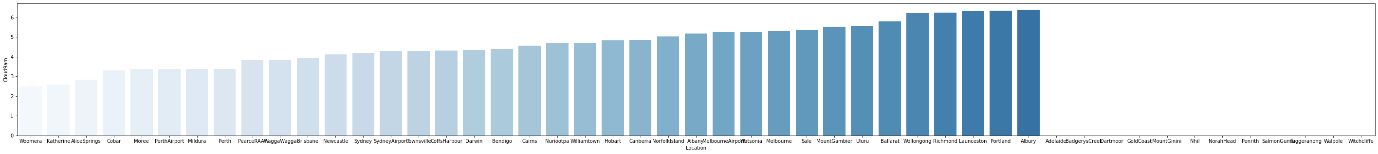


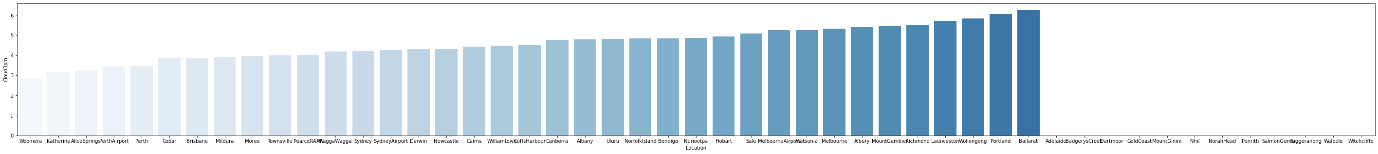


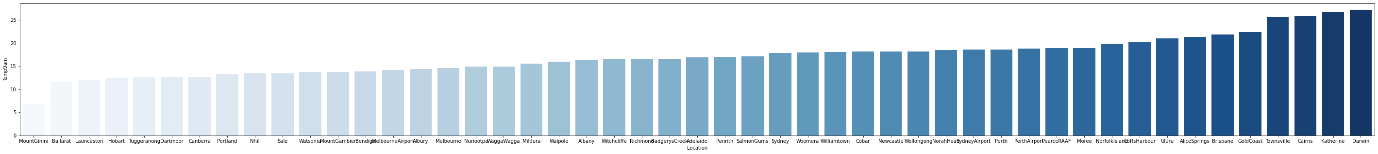


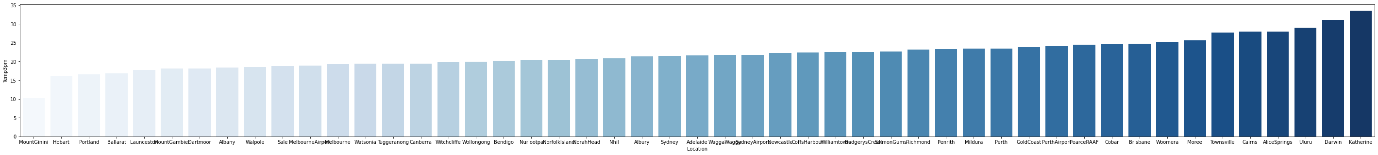








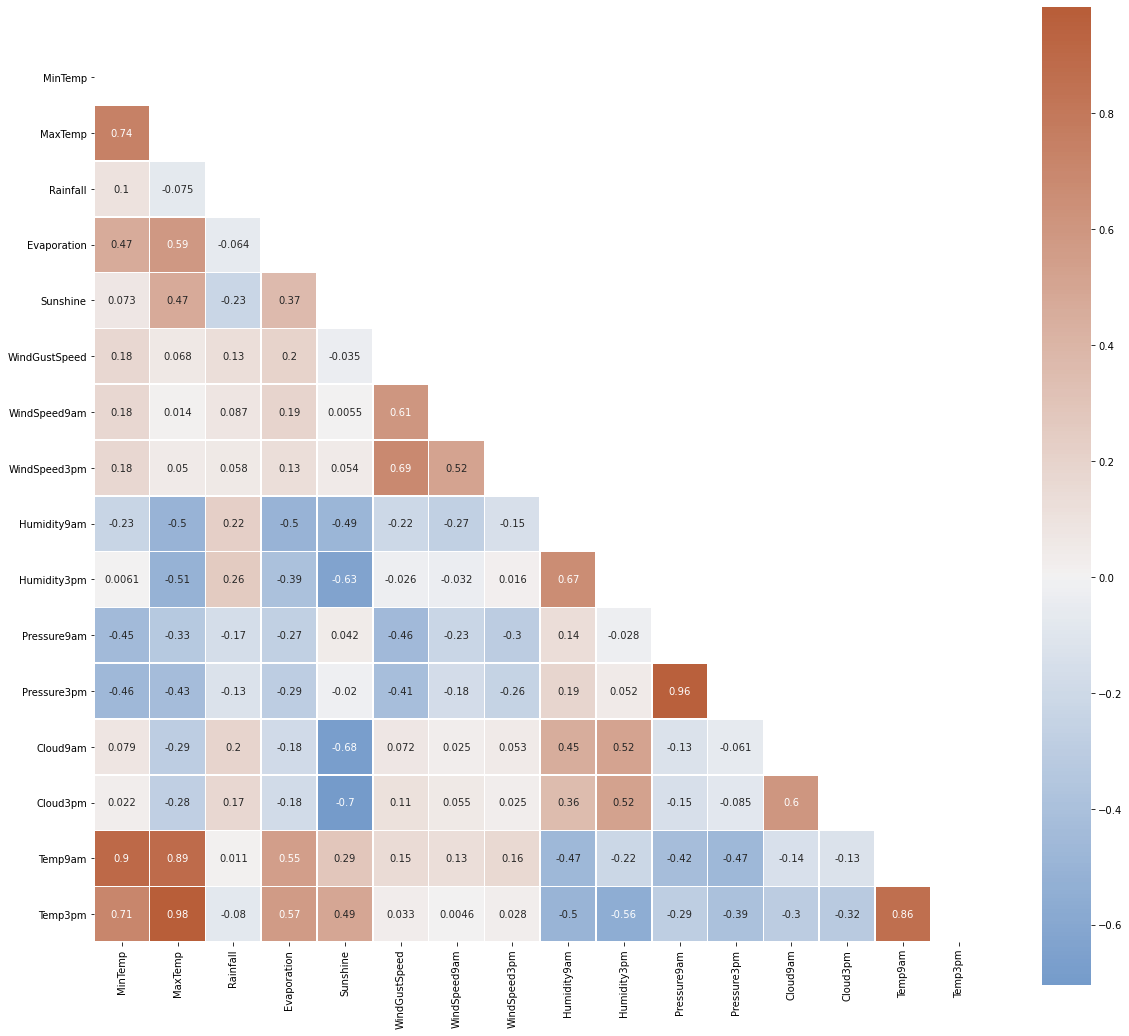




3 locations with least minimum temperature are Mount Ginini, Canberra and Tuggeranong. Highest minimum temperature are Katherine, Caims, Darwin. 3 locations with least maximum temperature are Mount Ginini, Hobart and Portland. Highest maximum temperature are Uluru, Darwin and Katherine.

Lowest rainfall experienced in Woomera, Uluru and Alice Springs. Highest rainfall experienced in Harbour, Darwin and Cairns. Lowest evaporation experienced in Dartmoor while highest in Woomera. Lowest sunshine levels are in Watson while highest in Alice Springs.

Wind speeds are lowest in Brisbane while highest in Hobart. Humidity is lowest in Alice Springs and Uluru while its highest in Dartmoor and Mount Ginini. Pressure levels are constant across all locations. Cloud levels are lowest in Woomera while highest in Portland, Albury and Ballarat.



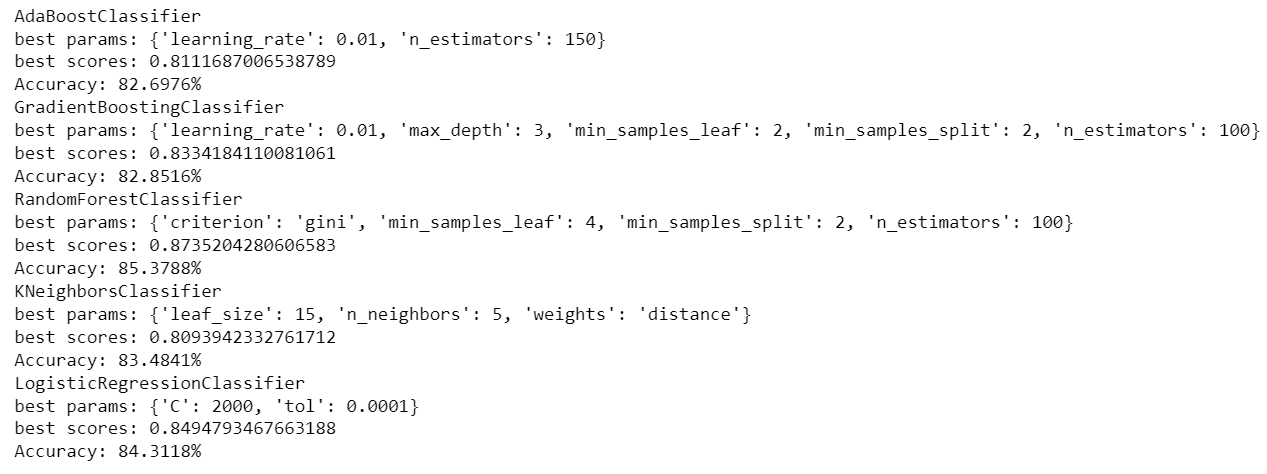
Max temperature is highly correlated to Temperature at 9am and Minimum temperature. Pressure at 9am is highly correlated to Pressure at 3pm, it seems the pressures impact each other at different time periods during the day.

Cloud and Sunshine are negatively correlated, which also makes sense. Wind and Pressure are also negatively correlated, although the correlation is weak. Humidity and Cloud levels are positively correlated, with a weak correlation.

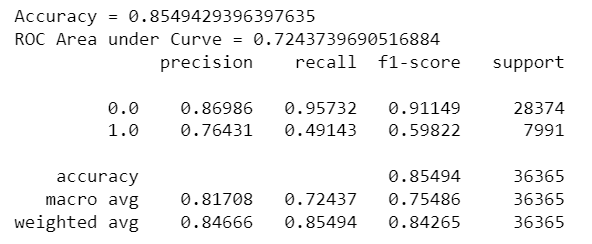
Windspeed at 9am and 3pm are positively correlated with overall Wind Gust speed. Evaporation is positively correlated with max temperature, as evaporation increases with increase in temperature.

**Modeling results**

Outcome of the Grid Search Cross Validation for 5 different models –



Based on the Grid Search, Random Forest seems to be best model out of the 5 with 85% accuracy and a model score of 0.875. The best hyper parameters for this are {'criterion': 'gini', 'min\_samples\_leaf': 4, 'min\_samples\_split': 2, 'n\_estimators': 100}

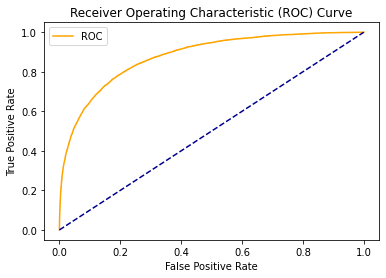


The Random forest based model has an accuracy of ~85%, this can be improved further with more tuning and feature engineering.

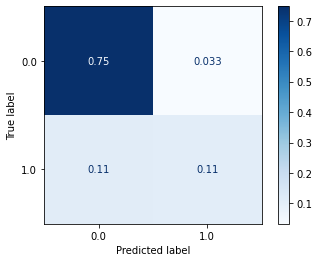
The precision and recall scores are good for days predicting no rainfall. The scores are very low for days predicting rainfall, this is due to an imbalanced dataset.

The F1-scores are good for days predicting no rainfall (0.91) while it is very low for days predicting rainfall (0.598).

Based on this, the model will predict most days as days with no rainfall.



Looking at the ROC curve, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives.



Looking at the confusion matrix, the model will predict no rainfall days 75% of the time correctly however, it will only predict rainfall days 11% of the time. It will predict rainfall days as no rainfall days 33% of the time making them as incorrect predictions.

The model is saved and provided for further reference.

## Insights

* There are 49 unique locations in the dataset, top 5 locations are Canberra, Sydney, Darwin, Perth and Brisbane. Portland, Caims and Walpole receives the highest rainfall on a frequent basis. Woomera, Canberra, Alice Springs receive the lowest rainfall on a frequent basis. West and South are the directions where most of the wind gust direction data is.
* Wind direction at 9am is mostly from the North, while wind direction at 3pm changes towards South East. It seems the direction of wind changes from north to south during the day. Most of the time, there was no rain during the observation period (78%). The next day also observed lower rainfall.
* Australia receives maximum rainfall in the beginning of the year, particularly between January to May, highest in January. Frequency of rainfall has increased from 2015 to 2017, with the highest rainfall in January 2017. Minimal rainfall is experienced during the rest of the year.
* Temperature seems to go up in the first quarter of the year (Jan to March) and starts falling after April with the lowest in August. Australia experiences winters in the month of July to October and this can be seen from the plot. The max and min temperature has remained somewhat constant over the years without changing much.
* Australia experiences faster wind speeds in the summer periods with highest wind speeds in December-January period. Wind speeds usually come down during the winter season with lowest in June-July period. Overall wind speeds have declined over the observation period. Wind speeds are usually higher at 3pm compared to 9am. Afternoon wind speeds have increased over the observation period.
* 3 locations with least minimum temperature are Mount Ginini, Canberra and Tuggeranong. Highest minimum temperature are Katherine, Caims, Darwin. 3 locations with least maximum temperature are Mount Ginini, Hobart and Portland. Highest maximum temperature are Uluru, Darwin and Katherine.
* Lowest rainfall experienced in Woomera, Uluru and Alice Springs. Highest rainfall experienced in Harbour, Darwin and Cairns. Lowest evaporation experienced in Dartmoor while highest in Woomera. Lowest sunshine levels are in Watson while highest in Alice Springs.
* Wind speeds are lowest in Brisbane while highest in Hobart. Humidity is lowest in Alice Springs and Uluru while its highest in Dartmoor and Mount Ginini. Pressure levels are constant across all locations. Cloud levels are lowest in Woomera while highest in Portland, Albury and Ballarat.
* Max temperature is highly correlated to Temperature at 9am and Minimum temperature. Pressure at 9am is highly correlated to Pressure at 3pm, it seems the pressures impact each other at different time periods during the day.
* Cloud and Sunshine are negatively correlated, which also makes sense. Wind and Pressure are also negatively correlated, although the correlation is weak. Humidity and Cloud levels are positively correlated, with a weak correlation.
* Windspeed at 9am and 3pm are positively correlated with overall Wind Gust speed. Evaporation is positively correlated with max temperature, as evaporation increases with increase in temperature.

## References

* <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>
* <https://scikit-learn.org/stable/supervised_learning.html>
* <https://www.kaggle.com/prashant111/extensive-analysis-eda-fe-modelling>
* <https://seaborn.pydata.org/generated/seaborn.heatmap.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html>
* <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html>