

*Toronto*

**Module2: EDA of Rain Forecasting in Australia**

Dhairya Davea (002110382)

Parth Savaliaa (002982303)

Manan Sonia (002982645)

a *College of Professional Studies****,*** *Master of Professional Studies in Analytics.*

***Subject****: ALY6040*

Under the guidance of

**Prof. Kamran Hootan**

**Introduction:**

This data is from [Kaggle](https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package?datasetId=6012&sortBy=voteCount) with titles “Rain in Australia”. By gathering quantitative information about the atmosphere's current condition at a specific location and utilising meteorology to predict how the atmosphere will evolve, weather forecasts are created. To forecast whether or not it will rain tomorrow, use the Rain Dataset. About 10 years' worth of daily weather measurements from several Australia locales are included in the dataset. As of seeing this, this will be use for classification purposes which can be said as weather forecasting. Speaking of weather predictions, they might have an impact on daily activities as well as industries like the food sector, tourism, emergency healthcare, etc. There is a target variable that can be either "Yes" or "No”. Yes, if there was at least 1mm of rain that day. The variables in our dataset that are most likely to cause rain to fall are pressure, humidity, clouds, and sunlight.

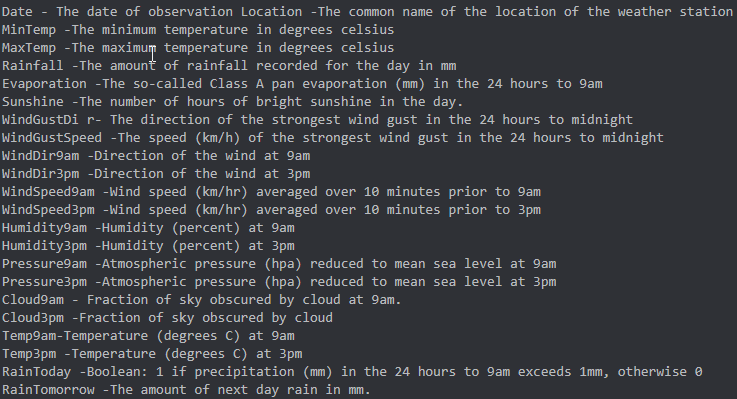
Problem statement for this dataset is as follows:

1. Using the given variables, approach to issue should be classification of rain.
2. Later developing the basic model, the issue will be up-scaling the model which has high accuracy than the previous and achieving advance classification via evaluating date and location variables.

Fig. 2 Information of Dataframe
These are the main 23 feature which eventually can cause a rain. Rain is a natural calamity and we have 145,460 samples which have some null values.

Due to less number of samples in this kind of research problem the performance of the modelcan be underfit.

**Fig. 1 Information of Data frame**



**Fig 2: Unit of Analysis**

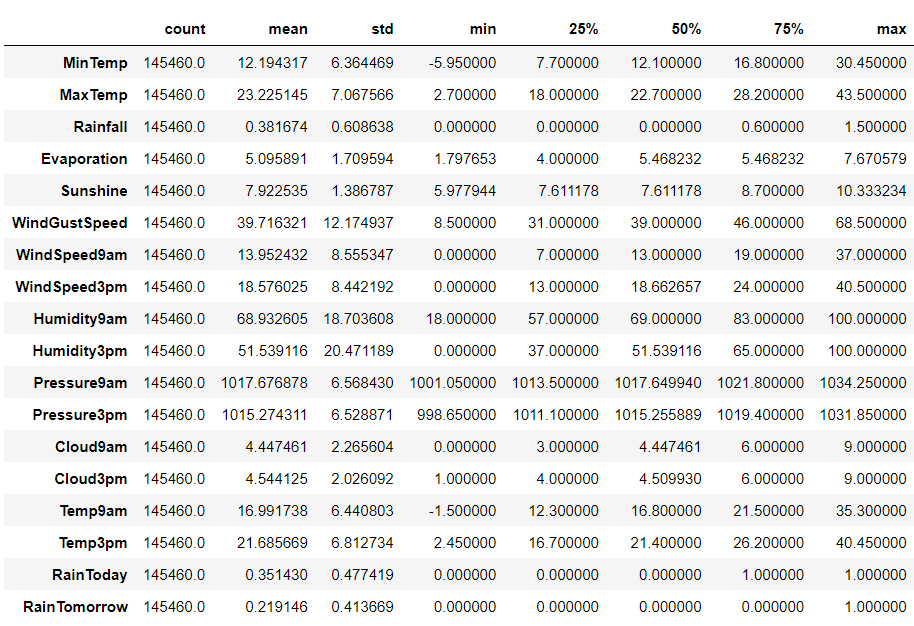
As illustrated in fig 2 (Unit of Analysis), we can see 24 hour format is reshaped from 9 am to 9 am for recording the sample. Another surprising fact is that the unit of cloud is fraction of sky covered by cloud, so we can assume that the fraction is in percentage.

As we can note in the Fig 3 (Description of Data) that difference in humidity at 9 am and 3 pm is significant as humidity reduce from 69 to 52 while both recorded maximum 100.

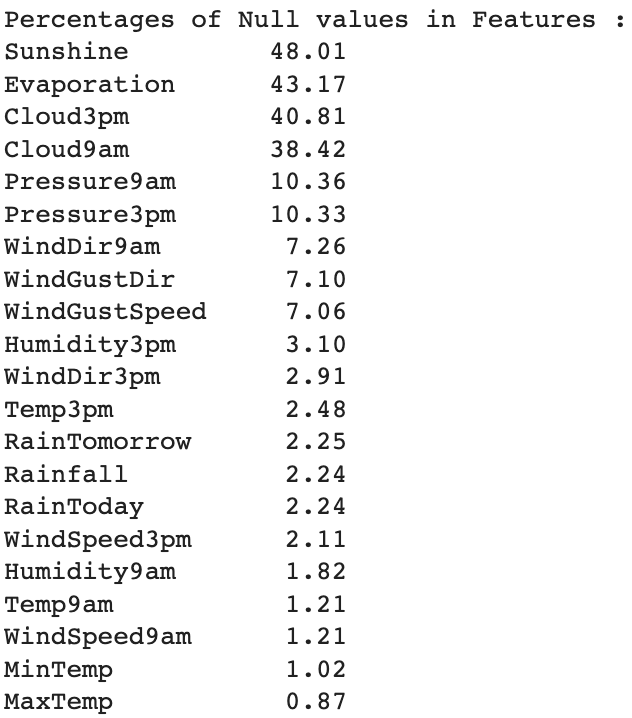
Whereas for pressure, where is slight difference for mean pressure at 9 am to 3pm.

It has recorded that maximum 9% fraction of sky was covered by cloud at specific time frame.

We can notice that there are some places where 371 mm rainfall has been register. And this amount of rain can cause flood like disaster.



**Fig.3 Description of Data**

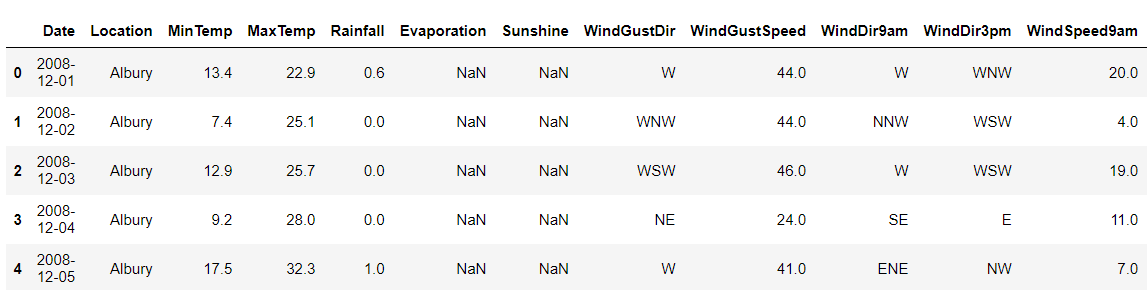
**Exploratory Data Analysis:**

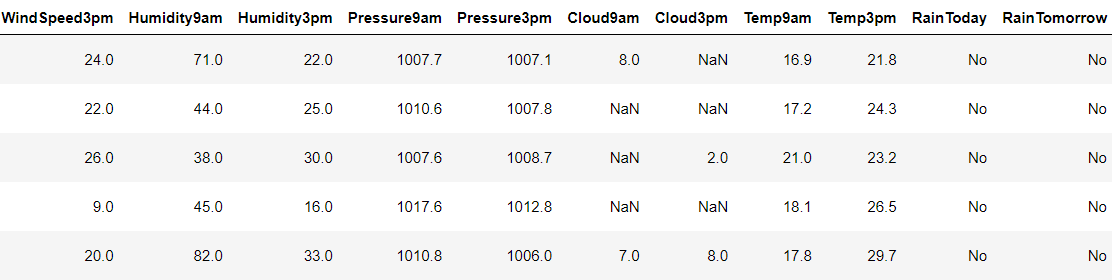
It is strongly advised to have a clean dataset in order to precede, for that same, we have just figuring out the percentage of null values in the research.

The existence of these features, which might have been absent at the time, could be the cause of these null values.

The least amount of percentage of nulls is in MinTemp and MaxTemp.

**Fig 4: Percentage of Null values in Features**

****

****

**Fig 5: First 5 rows of Data frame**

Fig 5 shows the first 5 rows that are present in the dataframe.

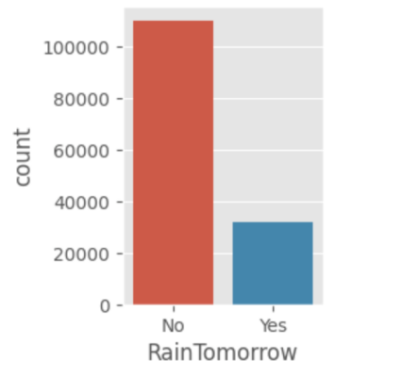
from we are focusing and aligning our resources to a target variables which is: RainTomorrow

Analyzing data of RainTomorrow we came up with amount of percentage of their values; which are:

No 75.839406

Yes 21.914616

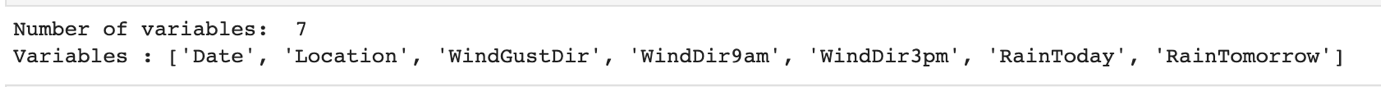
NaN represent no a number which is treated as null and we need to remove these samples containing nulls. It’s approx 2% so we can remove it from the data frame. Replacement of null value from our point of view should not be done as it is our target variable.

****Fig 6 (Count of each value) depicts the count of individual variable of the same attribute Number of count of NOs is relatively more than that of YES.

**Fig. 6 Count of each value**

**EDA for Categorical value:**

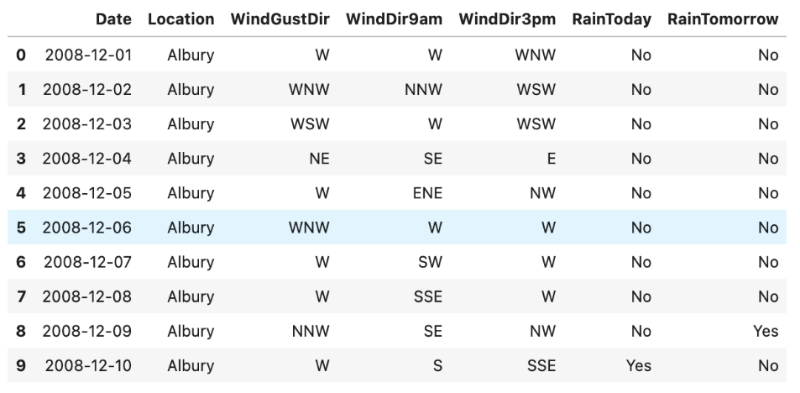
Now, EDA is performed on categorical vales of the dataset. Out of 23 total attribute we have 7 categorical values. Fig 6 represents the attribute in category. We also analysed distributions of WindDir3pm, WindDir9am and WindGustDir have shown same response as uniform distribution.



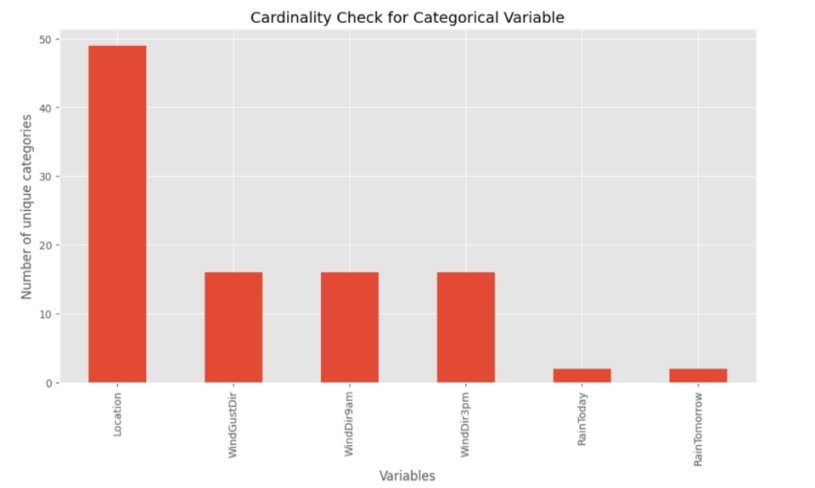
**Fig.7: Categorical Variable**

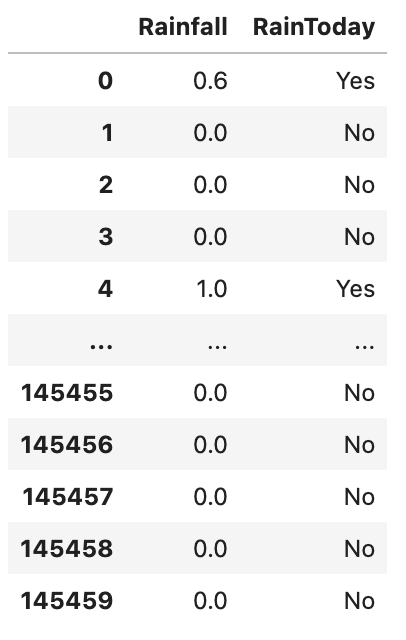
Table 1 displays the top 10 values of the categorical variables.

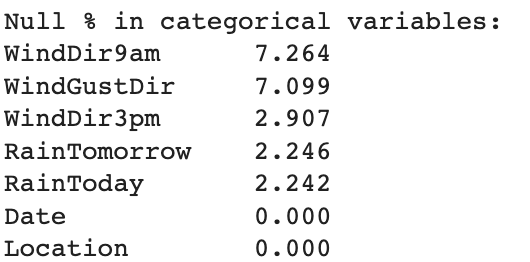
**Table 1: Top 10 values of Categorical variables**



**Cardinality**

The number of distinct values allocated to a dimension is referred to as its cardinality. A certain number of distinct values are specified for some dimensions. Cardinality checking is important because it specifies the relation between other categorical variables also, if we discover high cardinality in a dataset then it will provide an extremely large matrix which in result makes building of model extremely difficult or it will cause under-fitting. Depicting from the graph we can see that location has the highest count of unique categories which we will use for advance classification.

**Fig 8: Percentage of null values of categorical values Table 2: Cleaning of rainfall**



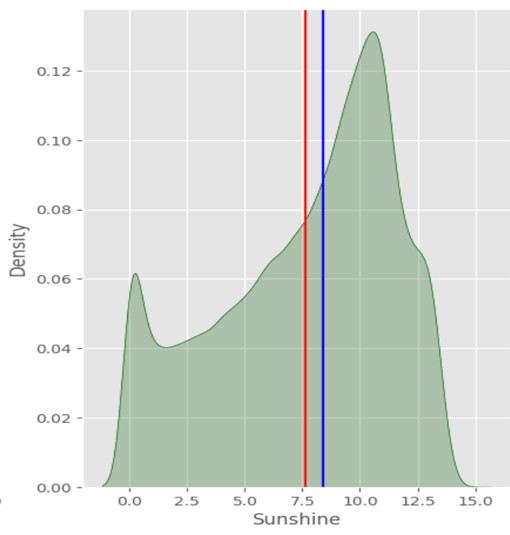
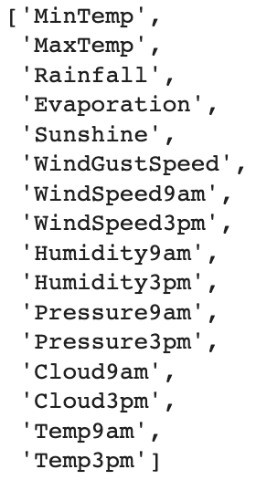
**Fig 9: Null percentage in categorical variable**

For cleaning Rainfall and RainToday of the categorical variables, we have to be sure that as we cannot just simply impute the data with mode and median only as it depends on the rainfall.

As per the procedure we corrected Null value to median of Rainfall and if the value of Rainfall is greater than 0.0 we are classifying RainToday as yes.

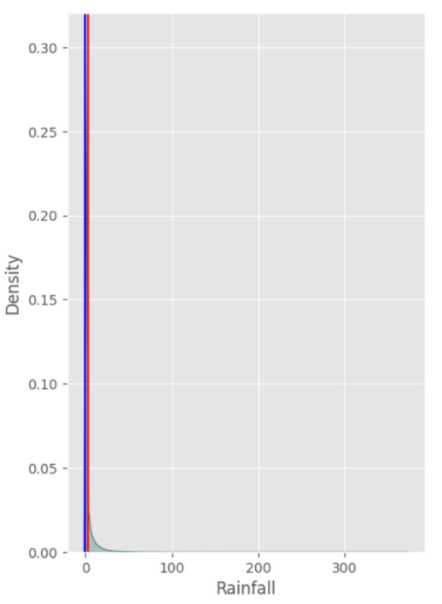
**EDA for Numerical Features:**

Firstly, we display name of the numerical data and thereby along with that we also defined each integer to a data type equal to float64.

**Fig 12: Distribution of Sunshine** **Fig 11: Numerical value**

Clean data is necessary which is input for our model. When it comes to imputation of missing values for numeric data, we need to check the distribution. This selection of method is determined on the basis of distribution of each variable.

****In Fig 12 (distribution of sunshine), the graph is unevenly distributed and weighted right side. And we can see that red line which is mean of the sunshine is present at the left side and hence we opt for mean to impute data.

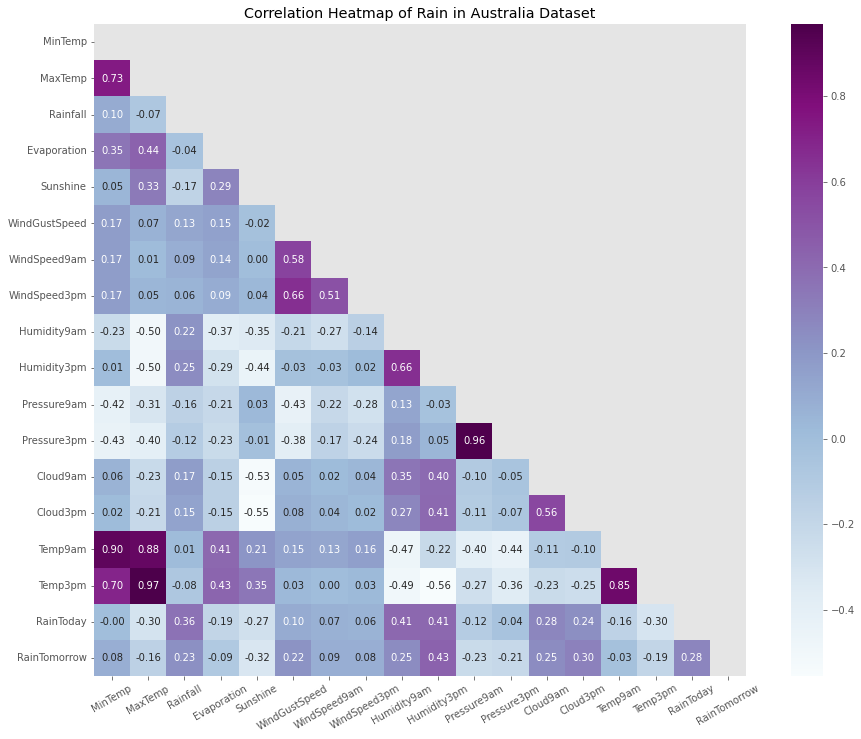
Here in Fig.13 (Distribution of Rainfall), we can see that data is having 0 for a lot of samples. Hence, for this mean and median both on the same line hence any of that is preferred to get imputed with. But replacing median is better choice as 1 mm of rainfall can affect RainToday variable and it should be classified as yes.

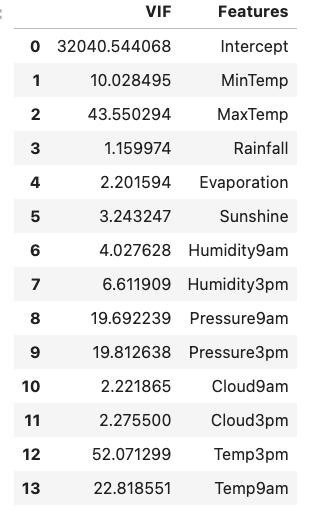
**Fig 13: Distribution of Rainfall**

**Collinearity and Feature Engineering**

Primarily, we need to factorise features into numeric. To add to that, we need to plot a heat map or correlation chart, to determine the collinearity between the features.

As these features are realistic there is hardly strong correlation between the variable. So we need to calculate the variation inflation factor (VIF) to identity multicollinearity between each variable. There are categorical variable with more unique value or cardinality, it can be ambiguous to understand for collinearity.



**Fig. 14 Heatmap**

From the Fig 14 , we can see different collinearity between the variable, There are variables which are having high positive collinearity and cause a rain such as Humidity, Temp and pressure and there are some negative such as sunshine and evaporation.

If value of VIF greater than 5 then it indicates higher potential correlation between the variables. This is not reliable for categorical data if that has cardinality.

VIF is generally used for regression models but for learning co linearity we can use it.

**Fig. 15 VIF factor**

**Conclusion**

In nutshell, after this EDA, we have enough evidence for feature engineering for our classification model. We cleaned a variable with mean, median and mode with use of distribution and logic. From the heat map we weren’t sure for which variable we should go for and there we can use correlation values as well as VIF to judge.

* We still haven’t figure out that how wind can cause a rain with help of other, so we left it as another goal for future scope.
* Considering a milestone, we are more tends to develop a model for classification and then go for higher scale like with the use of Date and Location variable and also the use of wind variables.

Note: We are bound for 1400 words are less for explaining all of the stuff. Rest is in notebook file. Attaching **GitHub** repo as reference.

**References**

[1] M. (2022, October 9). *GitHub - Mxnxn/Rain\_Forecasting\_Australia*. GitHub. Retrieved October 9, 2022, from <https://github.com/Mxnxn/Rain_Forecasting_Australia>

[2] Z., & posts by Zach, V. A. (2020, July 20). *How to Calculate VIF in Python - Statology*. Statology. Retrieved October 9, 2022, from <https://www.statology.org/how-to-calculate-vif-in-python/>