Toronto

Module2: EDA of Rain Forecasting in Australia

Dhairya Dave[a] (002110382)

Parth Savalia[a] (002982303)

Manan Soni[a] (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 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.

```
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           145460 non-null  object
 1   Location       145460 non-null  object
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object
 10  WindDir3pm     141232 non-null  object
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object
 22  RainTomorrow   142193 non-null  object
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

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 model can be underfit.

Fig. 1 Information of Data frame

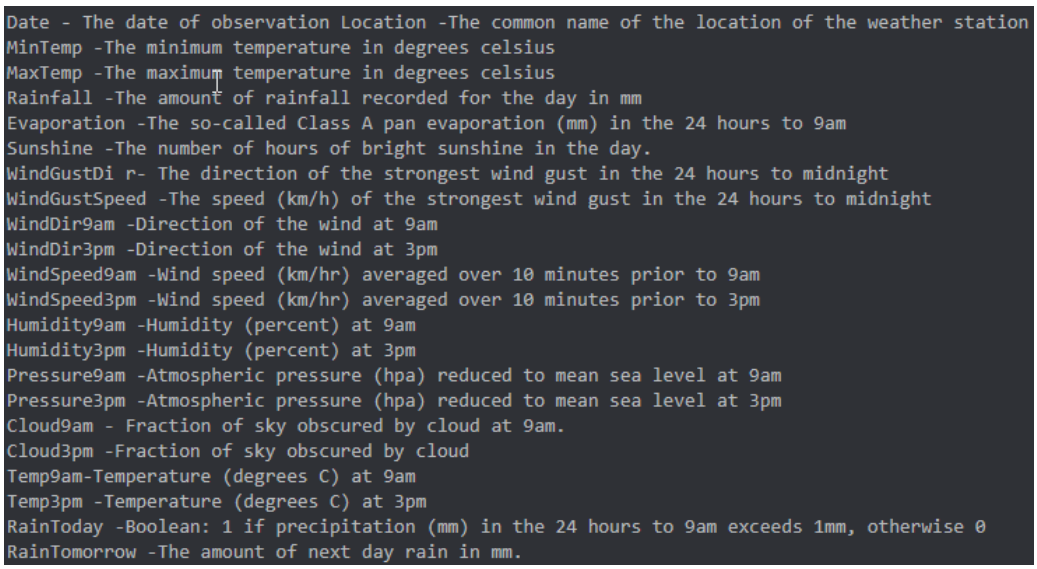


Fig 2: Unit of Analysis

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.

	count	mean	std	min	25%	50%	75%	max
MinTemp	145460.0	12.194317	6.364469	-5.950000	7.700000	12.100000	16.800000	30.450000
MaxTemp	145460.0	23.225145	7.067566	2.700000	18.000000	22.700000	28.200000	43.500000
Rainfall	145460.0	0.381674	0.608638	0.000000	0.000000	0.000000	0.600000	1.500000
Evaporation	145460.0	5.095891	1.709594	1.797653	4.000000	5.468232	5.468232	7.670579
Sunshine	145460.0	7.922535	1.386787	5.977944	7.611178	7.611178	8.700000	10.333234
WindGustSpeed	145460.0	39.716321	12.174937	8.500000	31.000000	39.000000	46.000000	68.500000
WindSpeed9am	145460.0	13.952432	8.555347	0.000000	7.000000	13.000000	19.000000	37.000000
WindSpeed3pm	145460.0	18.576025	8.442192	0.000000	13.000000	18.662657	24.000000	40.500000
Humidity9am	145460.0	68.932605	18.703608	18.000000	57.000000	69.000000	83.000000	100.000000
Humidity3pm	145460.0	51.539116	20.471189	0.000000	37.000000	51.539116	65.000000	100.000000
Pressure9am	145460.0	1017.676878	6.568430	1001.050000	1013.500000	1017.649940	1021.800000	1034.250000
Pressure3pm	145460.0	1015.274311	6.528871	998.650000	1011.100000	1015.255889	1019.400000	1031.850000
Cloud9am	145460.0	4.447461	2.265604	0.000000	3.000000	4.447461	6.000000	9.000000
Cloud3pm	145460.0	4.544125	2.026092	1.000000	4.000000	4.509930	6.000000	9.000000
Temp9am	145460.0	16.991738	6.440803	-1.500000	12.300000	16.800000	21.500000	35.300000
Temp3pm	145460.0	21.685669	6.812734	2.450000	16.700000	21.400000	26.200000	40.450000
RainToday	145460.0	0.351430	0.477419	0.000000	0.000000	0.000000	1.000000	1.000000
RainTomorrow	145460.0	0.219146	0.413669	0.000000	0.000000	0.000000	0.000000	1.000000

Fig.3 Description of Data

Exploratory Data Analysis:

```
Percentages of Null values in Features :
Sunshine          48.01
Evaporation       43.17
Cloud3pm          40.81
Cloud9am          38.42
Pressure9am       10.36
Pressure3pm       10.33
WindDir9am         7.26
WindGustDir        7.10
WindGustSpeed      7.06
Humidity3pm        3.10
WindDir3pm         2.91
Temp3pm            2.48
RainTomorrow       2.25
Rainfall           2.24
RainToday          2.24
WindSpeed3pm       2.11
Humidity9am        1.82
Temp9am            1.21
WindSpeed9am       1.21
MinTemp            1.02
MaxTemp            0.87
```

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

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	WindSpeed9am
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	WNW	20.0
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	WSW	4.0
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	WSW	19.0
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	E	11.0
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	NW	7.0

WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
24.0	71.0	22.0	1007.7	1007.1	8.0	NaN	16.9	21.8	No	No
22.0	44.0	25.0	1010.6	1007.8	NaN	NaN	17.2	24.3	No	No
26.0	38.0	30.0	1007.6	1008.7	NaN	2.0	21.0	23.2	No	No
9.0	45.0	16.0	1017.6	1012.8	NaN	NaN	18.1	26.5	No	No
20.0	82.0	33.0	1010.8	1006.0	7.0	8.0	17.8	29.7	No	No

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.

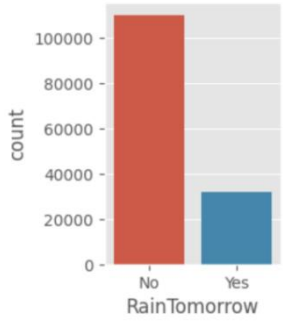


Fig 6 (Count of each value) depicts the count of individual variable of the same attribute Number of count of NOs is relatively more than that of YES.

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.

```
Number of variables:  7
Variables : ['Date', 'Location', 'WindGustDir', 'WindDir9am', 'WindDir3pm', 'RainToday', 'RainTomorrow']
```

Fig.7: Categorical Variable

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

Table 1: Top 10 values of Categorical variables

	Date	Location	WindGustDir	WindDir9am	WindDir3pm	RainToday	RainTomorrow
0	2008-12-01	Albury	W	W	WNW	No	No
1	2008-12-02	Albury	WNW	NNW	WSW	No	No
2	2008-12-03	Albury	WSW	W	WSW	No	No
3	2008-12-04	Albury	NE	SE	E	No	No
4	2008-12-05	Albury	W	ENE	NW	No	No
5	2008-12-06	Albury	WNW	W	W	No	No
6	2008-12-07	Albury	W	SW	W	No	No
7	2008-12-08	Albury	W	SSE	W	No	No
8	2008-12-09	Albury	NNW	SE	NW	No	Yes
9	2008-12-10	Albury	W	S	SSE	Yes	No

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.

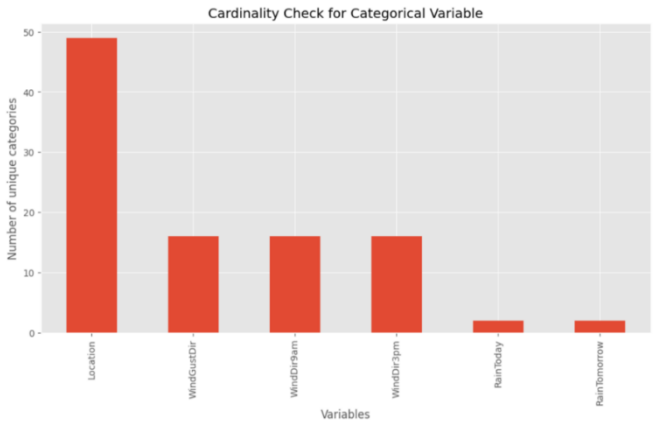


Fig 8: Percentage of null values of categorical values

```
Null % in categorical variables:
WindDir9am       7.264
WindGustDir      7.099
WindDir3pm       2.907
RainTomorrow     2.246
RainToday        2.242
Date             0.000
Location         0.000
```

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.

Table 2: Cleaning of rainfall

	Rainfall	RainToday
0	0.6	Yes
1	0.0	No
2	0.0	No
3	0.0	No
4	1.0	Yes
...	...	...
145455	0.0	No
145456	0.0	No
145457	0.0	No
145458	0.0	No
145459	0.0	No

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.

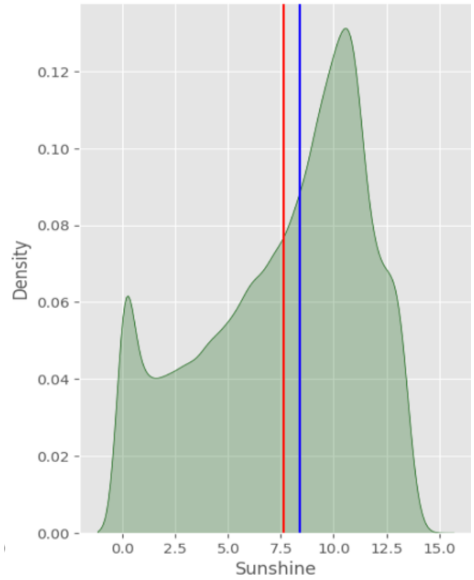


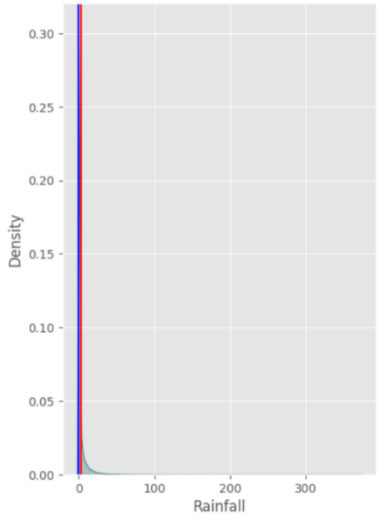
Fig 12: Distribution of Sunshine

```
['MinTemp',
 'MaxTemp',
 'Rainfall',
 'Evaporation',
 'Sunshine',
 'WindGustSpeed',
 'WindSpeed9am',
 'WindSpeed3pm',
 'Humidity9am',
 'Humidity3pm',
 'Pressure9am',
 'Pressure3pm',
 'Cloud9am',
 'Cloud3pm',
 'Temp9am',
 'Temp3pm']
```

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.

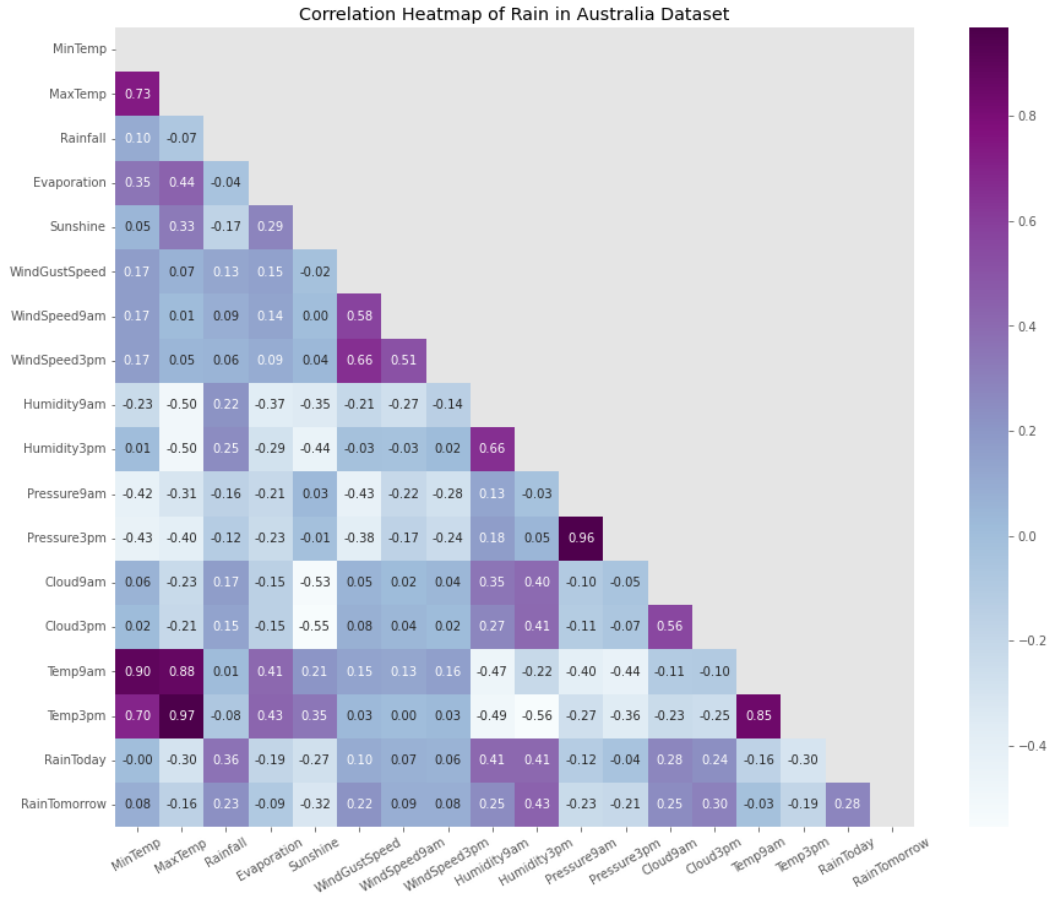


Fig. 14 Heatmap

	VIF	Features
0	32040.544068	Intercept
1	10.028495	MinTemp
2	43.550294	MaxTemp
3	1.159974	Rainfall
4	2.201594	Evaporation
5	3.243247	Sunshine
6	4.027628	Humidity9am
7	6.611909	Humidity3pm
8	19.692239	Pressure9am
9	19.812638	Pressure3pm
10	2.221865	Cloud9am
11	2.275500	Cloud3pm
12	52.071299	Temp3pm
13	22.818551	Temp9am

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/