**BLOG ARTICLE ON RAINFALL WEATHER FORECAST**

**Introduction**

Rainfall forecasts are crucial for safety, sustainability, and resource planning across various sectors. Accurate predictions help prepare for building projects, transportation, agriculture, aviation, and flood management. Predicting rainfall is one of the most difficult task of weather forecasting. By Machine learning techniques can uncover hidden patterns in historical data to improve rainfall predictions.

Weather forecasting involves predicting atmospheric conditions for a specific location and time. Meteorology uses quantitative data about the current state of the atmosphere to project future changes.

**Problem Definition**

We have a weather forecasting project focusing on rainfall in Australia.

The dataset contains 10 years of weather details for different Australian locations.

Features include location, month, year, wind direction, rain-related variables, temperature, evaporation, humidity, pressure, cloud cover, and more. By this features we have to avail two Project Goals.

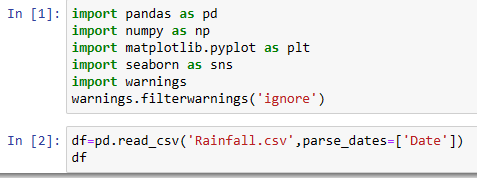
1. Design a predictive model to determine whether rain will occur tomorrow.

2. Create another model to predict the amount of rainfall (in millimeters).

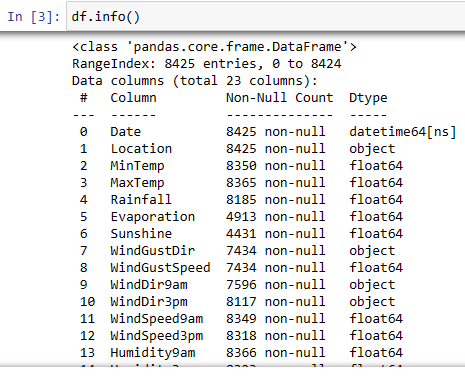
**Data Analysis**

Data analysis involves examining, cleaning, transforming, and interpreting data to discover patterns, trends, and meaningful information. It plays a crucial role in various fields, including business, science, research, and decision-making.

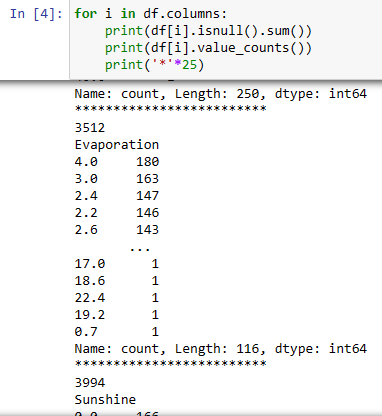
Python provides powerful libraries for data manipulation and analysis. Here we have Imported Some Needed Libraries.



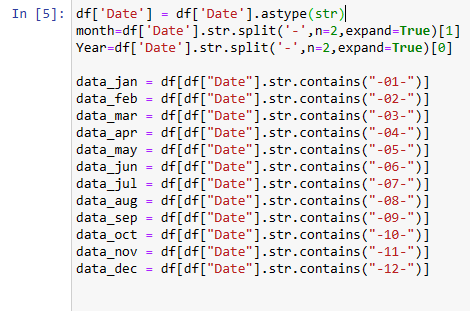
After importing Libraries and Extracting Dataset now we are preprocessing data, firstly we should check shape and info of dataset, so we can remove any unwanted column and many of things we can prepare by this DataFrame object.



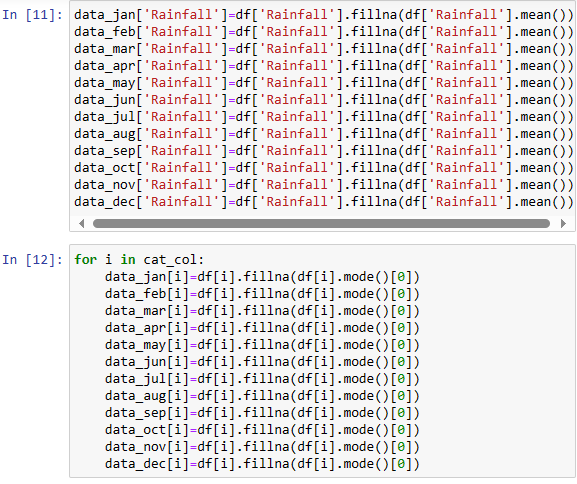
It is essential to identify any missing data (null values) in the dataset. We have used Pandas to check for null values column-wise (df. isnull().sum()) and the value\_counts function provides the count of unique values in a column. By examining the mode (most frequent value), we can determine a suitable value to fill null entries.

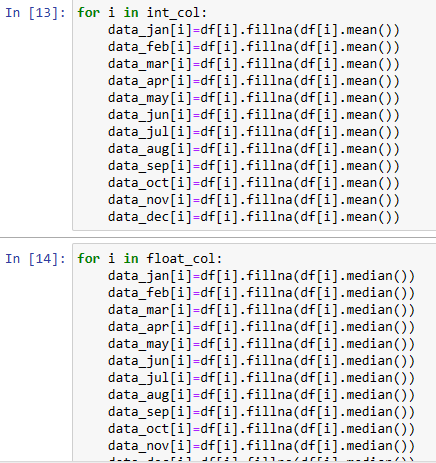


I have found null values in several features, including minTemp, maxTemp, Evaporation, Sunshine, Rainfall, WindGustDir, WindGustSpeed, Humidity, Pressure, Cloud, RainToday, and RainTomorrow.



In above Cell I have Divide data in month wise so We can better understand how rainfall occurs seasonally, like Rainfall occur in specific month, so as it is we can fill null wise month wise.

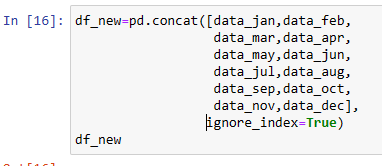


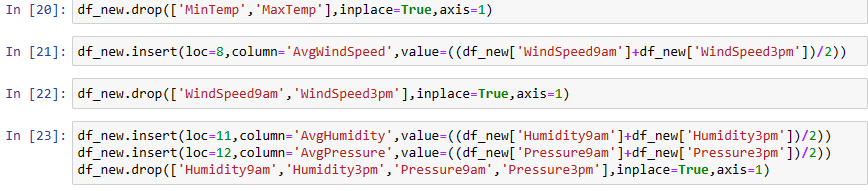


I have filled Sate wise null values by Mean, Median and Mode Techniques.

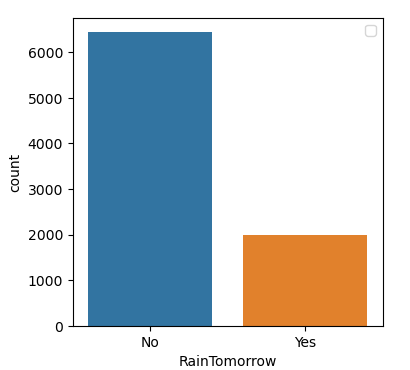
We filled state-wise null values using different techniques:

* Mean: Imputing with the average value of the feature.
* Median: Imputing with the middle value (50th percentile) of the feature.
* Mode: Imputing with the most frequent value.

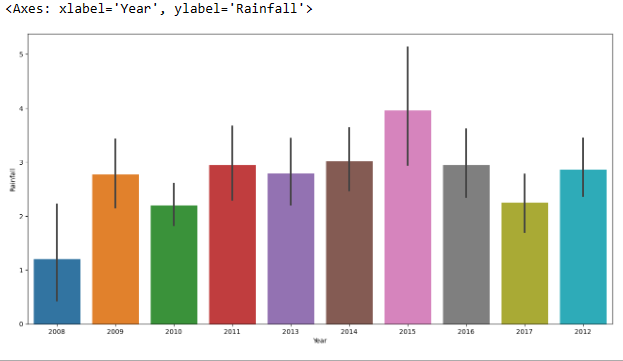




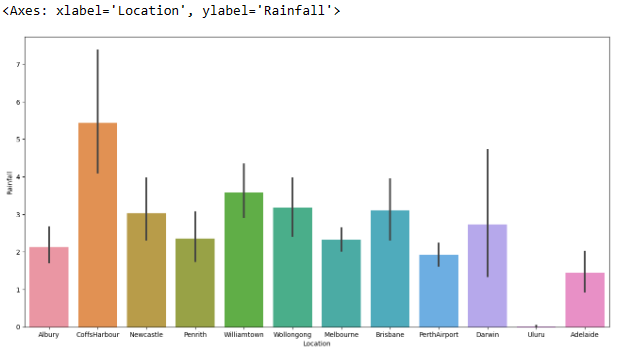
And Then we have Concate all month wise column in df DataFrame, and did some feature engineering like remove Date column after extracting Month and Year columns, and then we have drop minTemp and maxTemp after averaging both and inserting in new column by name Avg. Temp, as it is we have created AvgWindSpeed, AvgHumidity and AvgPressure column.

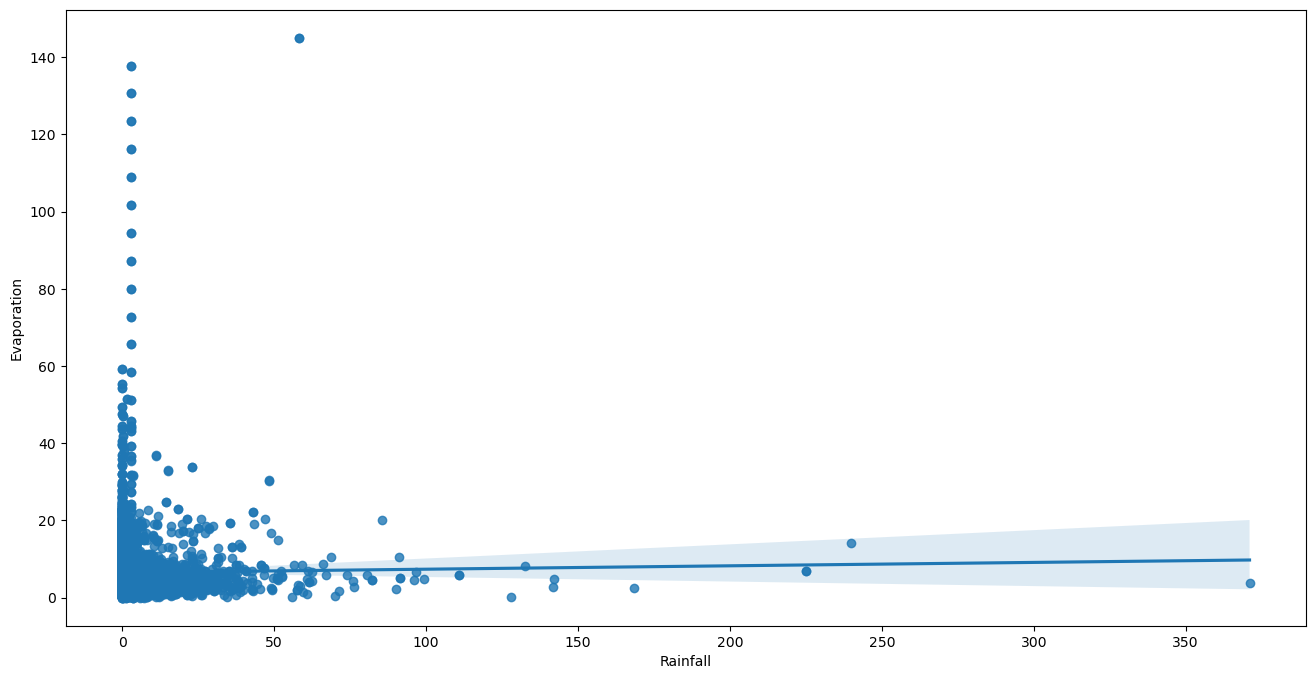


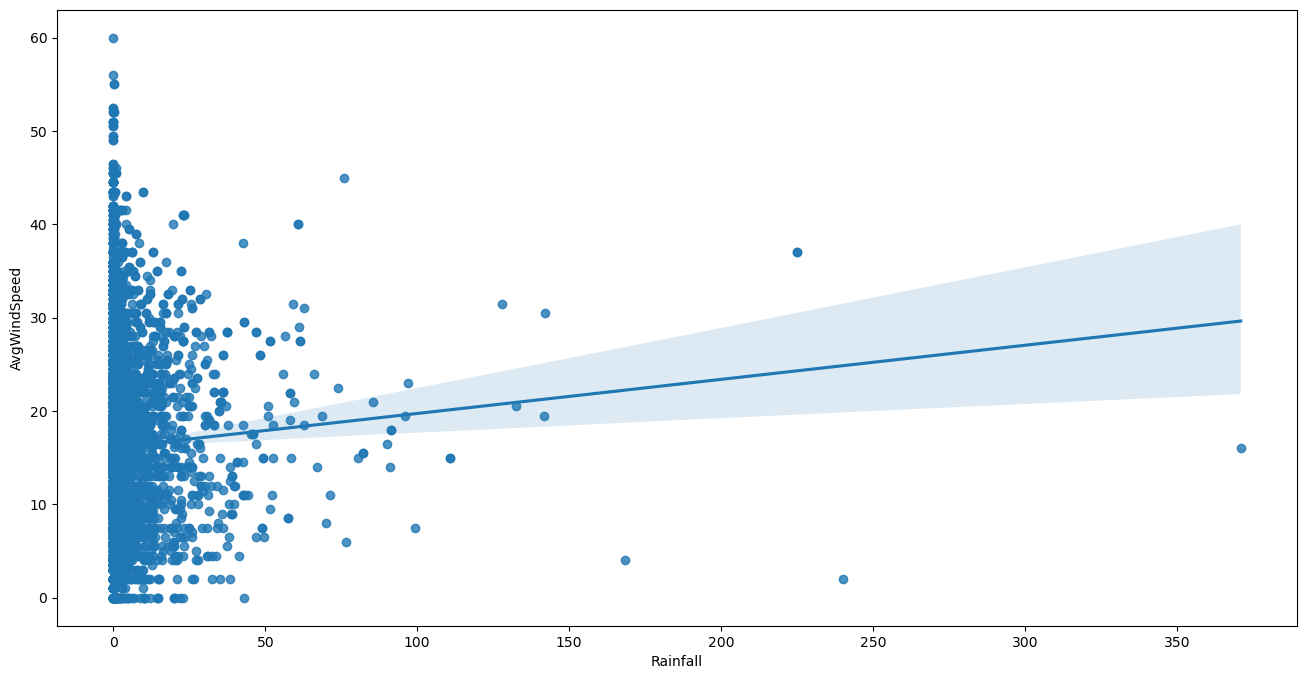
When we talk about Target Variable So There is huge difference between Categorical Variables 'Yes' and 'No', where 'No' variables count is nearly 7000 and 'Yes' Variables count is nearly 2000. So there is imbalance Data, so by SMOTE() Technique we will Balanced Data.

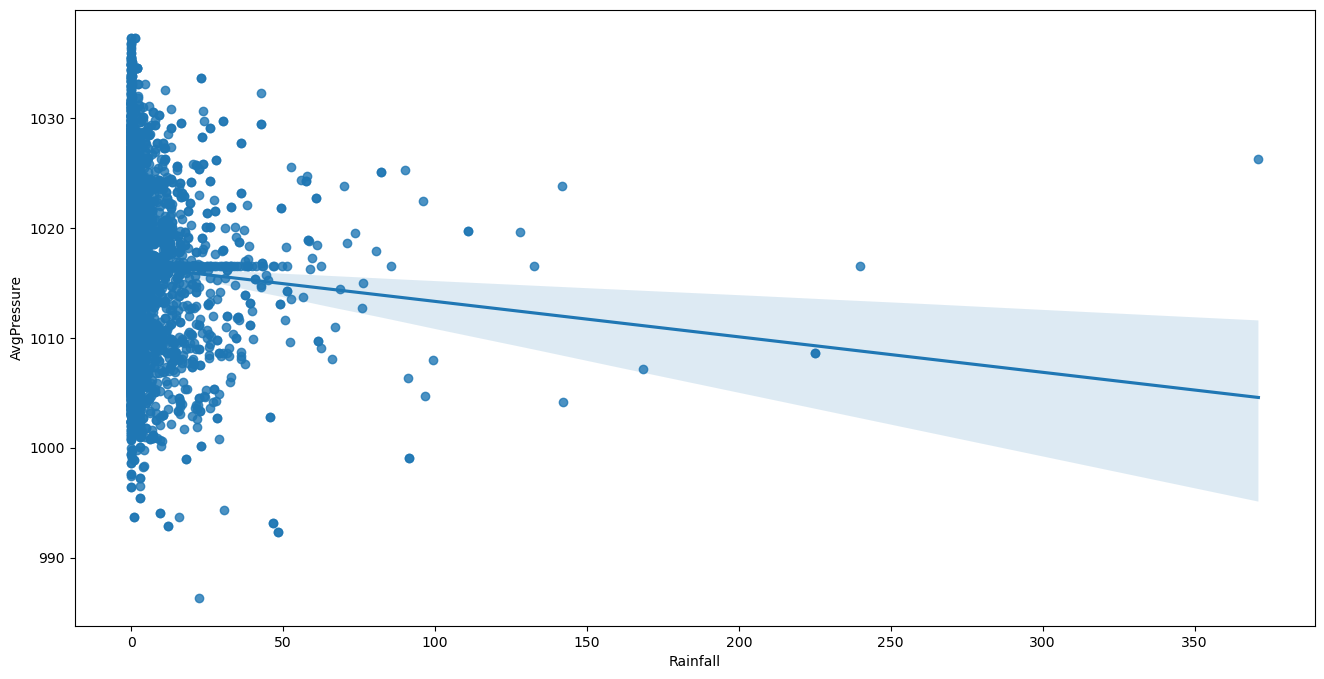


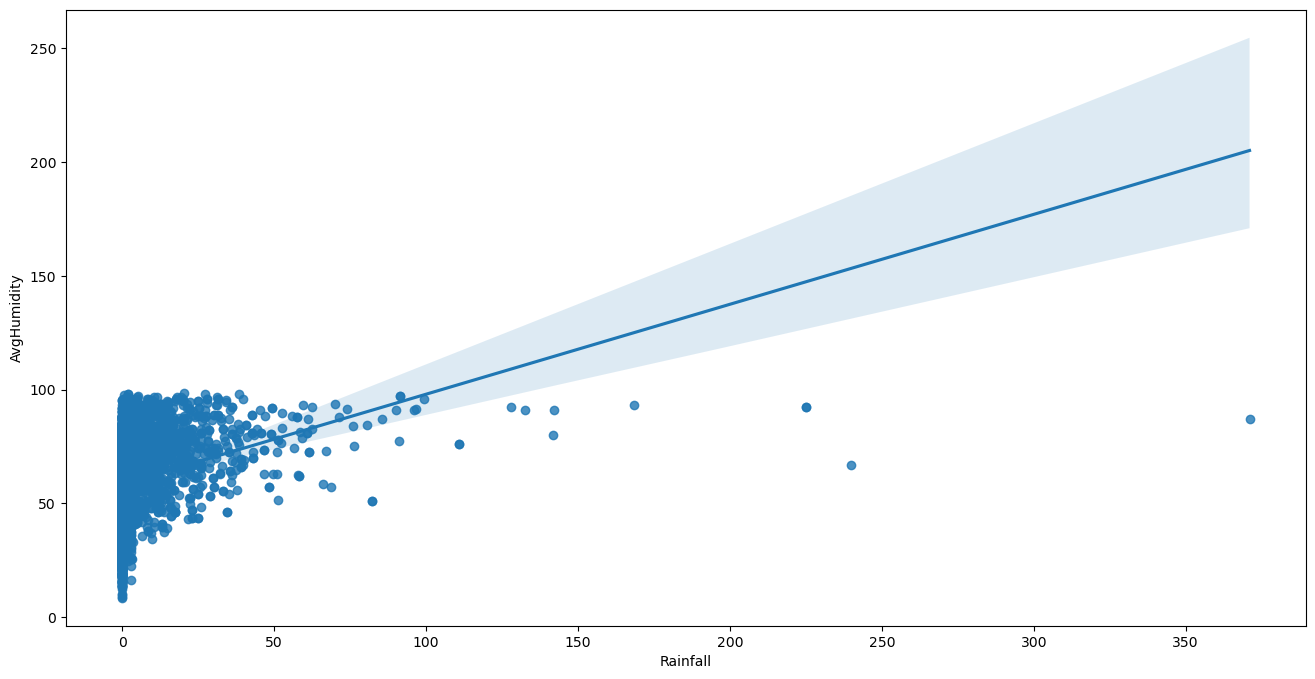
From above graph we can Predict that in CoffsHarbour City More Rainfall Occur, and in Uluru City There is almost no Rainfall. Remaining Cities have almost Same Rainfall during this Period.









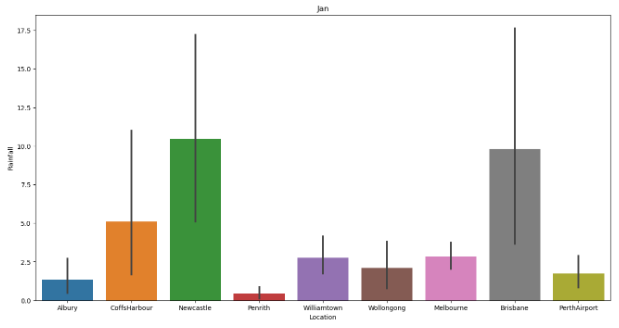


Above Graph indicates that When Avg. Temp of day is near to 20 then more number of days has record to Rainfall.

* When Evaporation is near to 0 There is more Rainfall is founded.
* As Sunshine Is Decreased Rainfall is increased.
* as Humidity is Increased Rainfall is Increased.
* As Avg. Pressure is decreased as Rainfall is Increased.

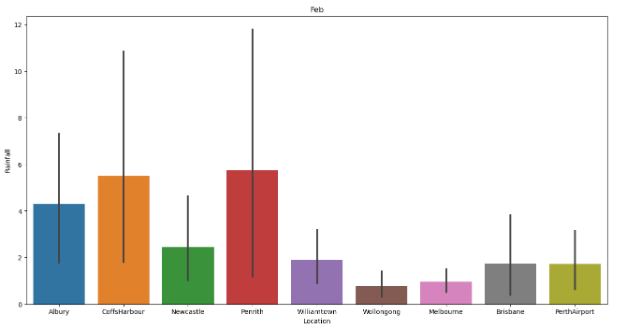
Rainfall Occurrence by Season in Different Locations:

Based on the graph, we can draw the following conclusions:



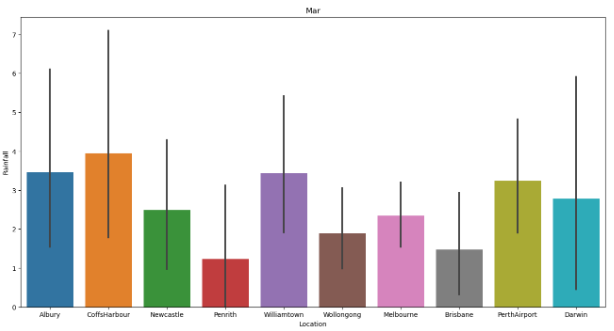
**January:**

* NewCastle and Brisbane experience higher rainfall.
* CoffsHarbour also receives some rainfall.



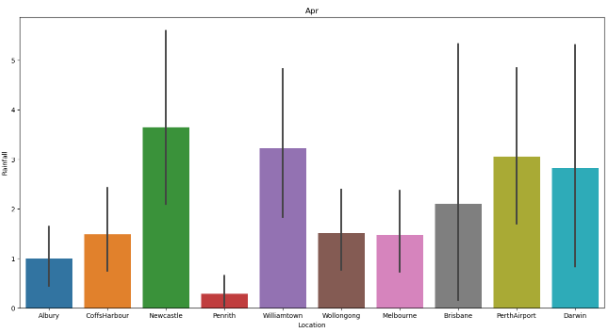
**February:**

* CoffsHarbour and Penrith have elevated rainfall levels.
* Albury also sees some rainfall.

****

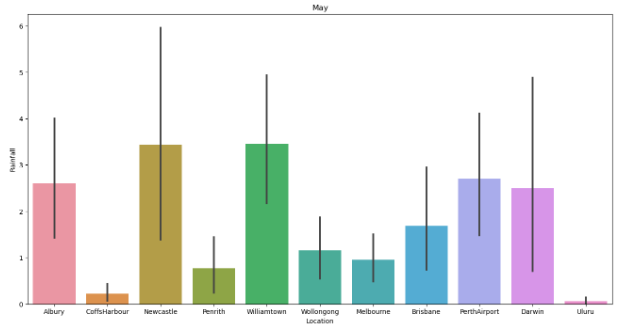
**March:**

* Albury, CoffsHarbour, Williamstown, and PerthAirport cities receive substantial rainfall. Other cities also experience rainfall, except for Uluru.



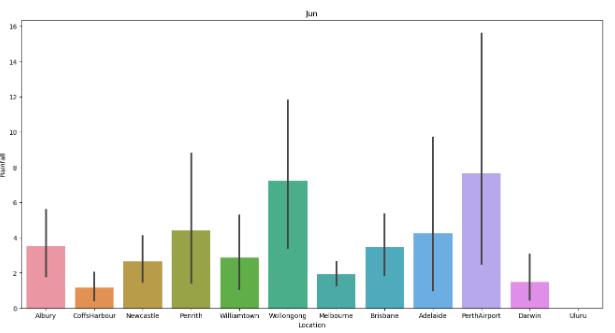
**April:**

* NewCastle, Williamtown, PerthAirport, and Darwin observe increased rainfall.



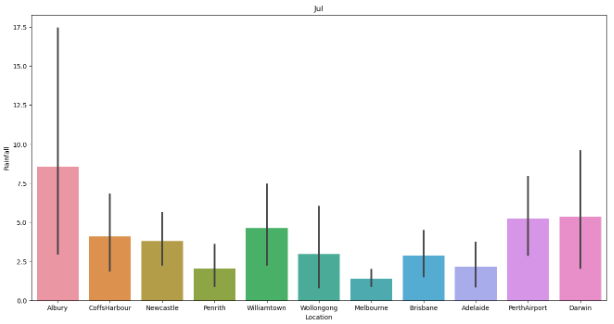
**May:**

* Albury, NewCastle, Williamtown, PerthAirport, and Darwin continue to have significant rainfall.



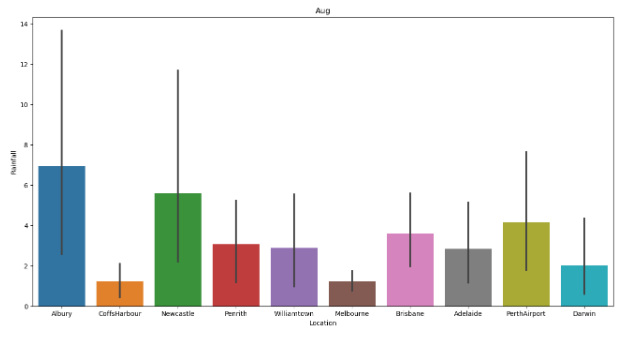
**June:**

* Wollongong and PerthAirport receive more rainfall during this month.



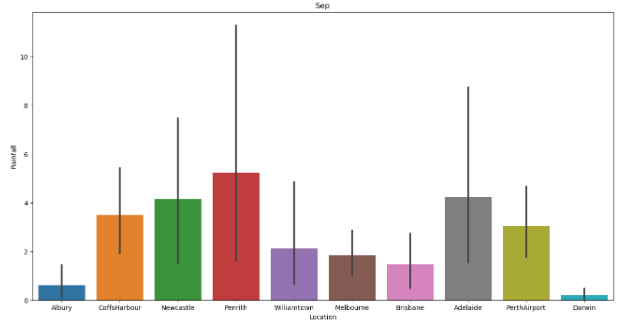
**July:**

* Albury remains a high-rainfall area.
* CoffsHarbour, Newcastle, Williamtown, Wollongong, Brisbane, PerthAirport, and Darwin also experience rainy conditions.



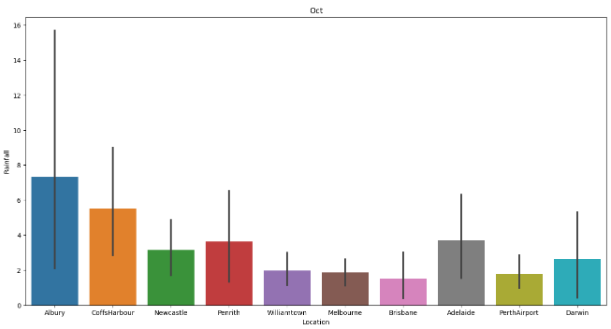
**August:**

* Albury and Newcastle still have substantial rainfall.
* Penrith, Williamtown, Brisbane, Adelaide, and PerthAirport also receive rain.



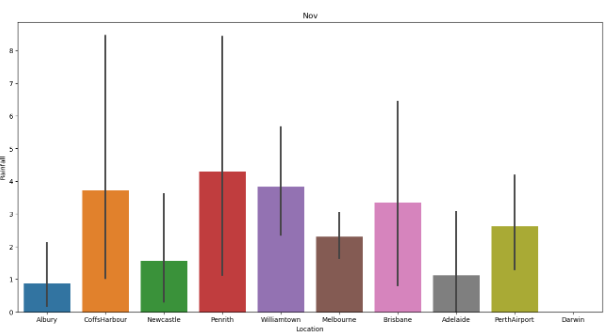
**September:**

* Penrith, Newcastle, Adelaide, PerthAirport, and CoffsHarbour see increased rainfall.



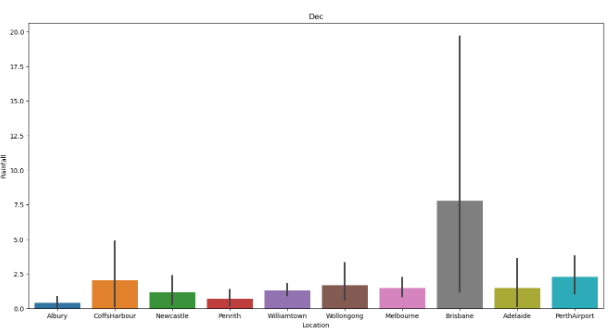
**October:**

* Albury, CoffsHarbour, Adelaide, Penrith, and Newcastle continue to have higher rainfall.



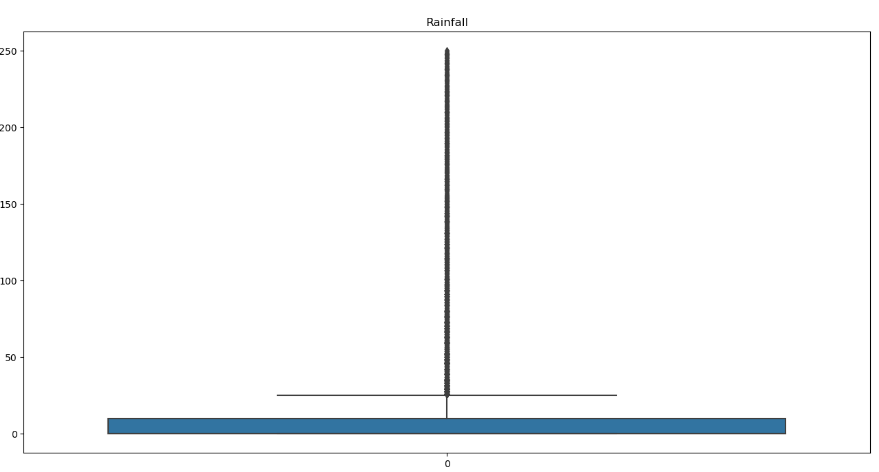
**November:**

* Penrith, Williamtown, CoffsHarbour, Brisbane, PerthAirport, and Melbourne receive more rainfall.

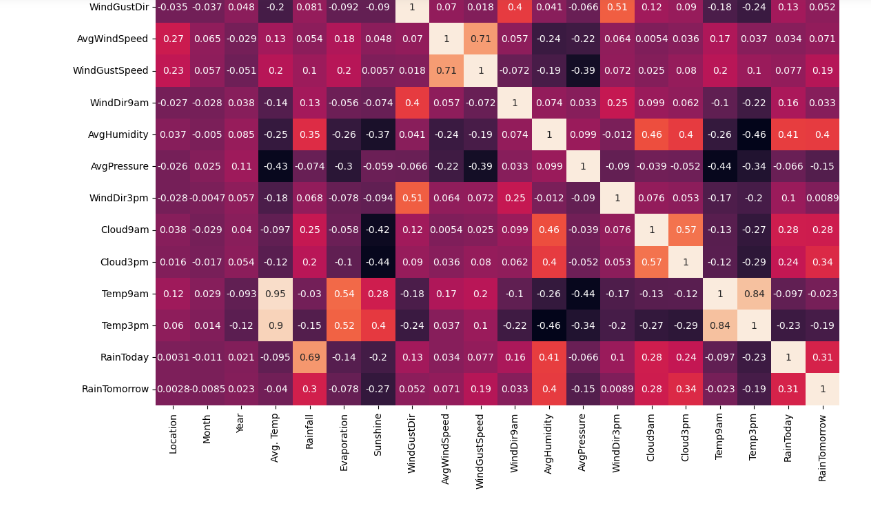


**December:**

* Brisbane experiences significant rainfall.



in above cell we had Check Outliers and skewness of data. so for removing outliers and skewness data should numeric, so for that we have encoded categorical data by Label Encoder Techniques. And we have found Positive Skewness in Rainfall and also found in boxplot. and we have fill null value by interpolate imputer techniques.



And then we have to check Correlation between Features by DataFarme object Corr (). by this we have found Correlation between Temp9am and Temp3pm. So because we have already create a column Avg. Temp from min Temp and MaxTemp, so that’s why we have to drop Temp9am and Temp3pm.Then we have also checked Multicolinearity by Variance\_inflation\_factor, but there is no multicolinearity found. And then we have checked Correlation of features, we have found correlation in between two columns and we had remove that columns.

**EDA CONCLUSION REMARK**

In EDA part we have checked null values in dataset and where is we have found null values there we have fill that null value by appropriate method. In few Features Like Evaporation, Sunshine, and Rainfall has more number of null values but that was appropriate columns for our prediction that's why we have filled them by interpolate imputer method. then we have analyzed Rainfall and weather aspects Extracted Month wise, and date column sorted into month and Year, and Remove Date column. and also we have Averaging Temp, Pressure, Humidity etc. and removed inappropriate columns. then we have done Encoding for Categorical data for further Operation. And at the end we had checked correlation between features, we have found and remove them.

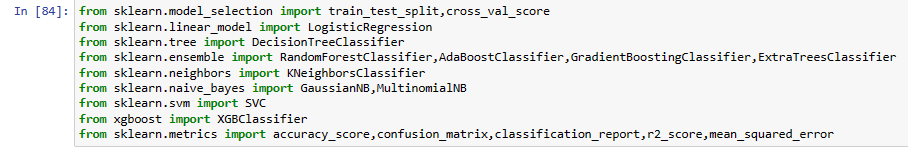
**Data Preprocessing Pipeline**

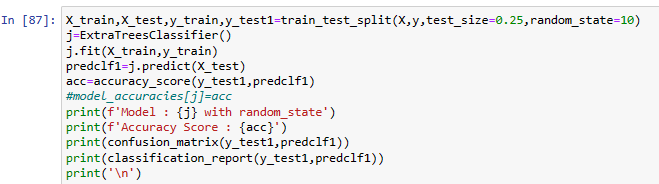
**Classification**

In this part first of all we have to separate X (Input or dependents) and y (output or Target or independent) variable for classification Problem and as it is we separate columns.

Building Machine Learning Models

To predict there is Rain Tomorrow or not which is classification problem, so for that from sklearn library appropriate module will import, and then finding best model to predict Rain for Tomorrow.





Model: ExtraTreesClassifier () with random\_state

Accuracy Score: 0.9328566987876904

[[1516 129]

[ 87 1485]]

precision recall f1-score support

0 0.95 0.92 0.93 1645

1 0.92 0.94 0.93 1572

accuracy 0.93 3217

macro avg 0.93 0.93 0.93 3217

weighted avg 0.93 0.93 0.93 3217



Cross-val score for LogisticRegression :0.7457318126836012

Cross-val score for DecisionTreeClassifier :0.7721527392600487

Cross-val score for RandomForestClassifier :0.8466805270071213

Cross-val score for AdaBoostClassifier :0.764696794479065

Cross-val score for GradientBoostingClassifier :0.7856770238574282

Cross-val score for ExtraTreesClassifier :0.8484696830731047

Cross-val score for KNeighborsClassifier :0.730114822105491

Cross-val score for GradientNB :0.7251361563958765

Cross-val score for MultinomialNB :0.6536437469563443

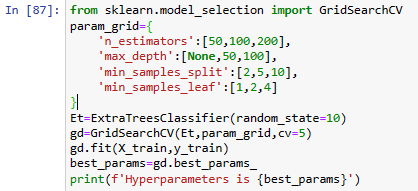
Cross-val score for SVC :0.7144180167508317

Cross-val score for XGBClassifier :0.7911195336545258

By creating dictionary of model and iterate over for loop to checked which model is best fit and which random state value is more suitable.

so we have found ExtraTreesClassifier has 93.28% accuracy at random state 10, and Precision, recall and f1-score accuracy found 93%. and cross-val score is 84%.

Hyper Parameter Tuning with ExtraTreesClassifier



Hyper parameters are {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 200}

These are the best parameters values got by using GridSearchCV for ExtraTreesClassifier.

R-Score is 0.7300443939147854

MSE is 0.06745414982903326

Model Score is 0.9325458501709667

Accuracy of best model is approximately same as we have before Hyper Parameter tuning which is best for perfect prediction.

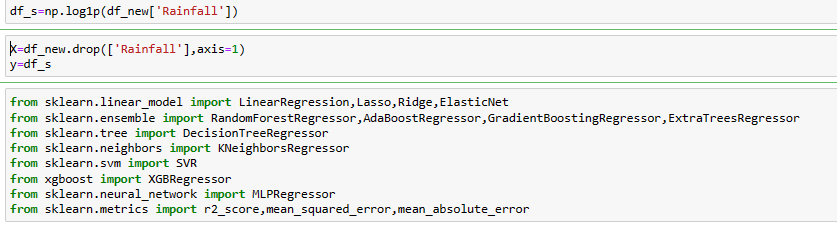


We have Saved model, and we can see that Our model Predicted very well to Actual value.

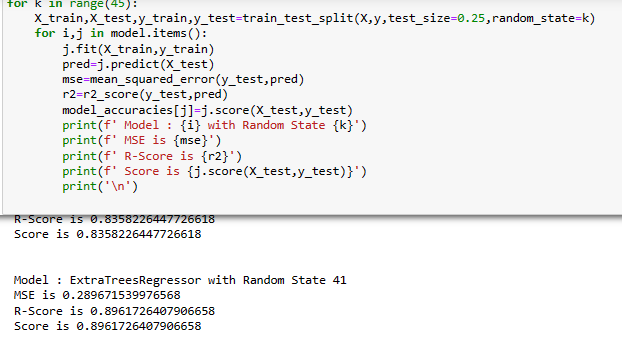
Now we have to solve 2nd Problem, to predict Rainfall in different location which is Regression Problem because Target variables in float type, so for that from sklearn library appropriate module will import, and then finding best model to predict Rainfall.

**Regression**

We had found skewness in rainfall which target variable so we have remove skewness by np.log1p which numpy object.



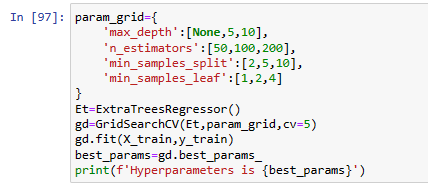
and then we have separate X(dependents) and y(independents) Variable.



ExtraTreesRegression is best model with best score is 89.61% at random state 41, and there is mse value is 0.28, now we do hyper parameter tuning for ExtraTreesRegression.

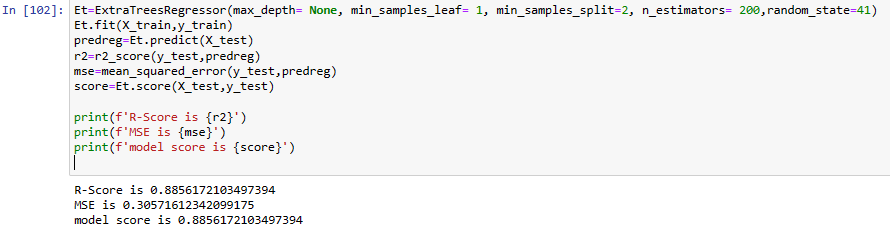
Best Model is ExtraTreesRegressor () with Score 0.8961726407906658

Hyperperameter Tuning with ExtraTreesRegressor



Hyper parameters are {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 200}

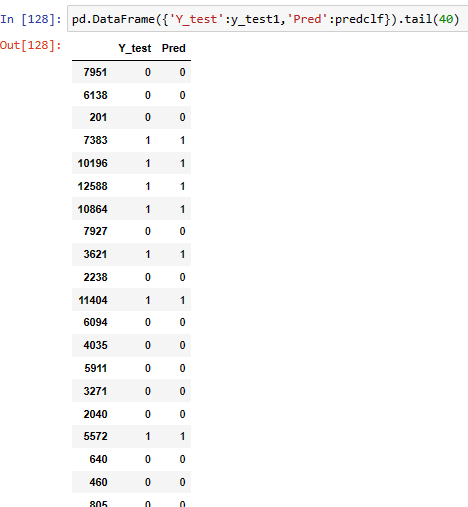
these are the best parameters got using GridSearchCV for ExtraTreesRegressor.



R-Score is 0.8856172103497394

MSE is 0.30571612342099175

model score is 0.8856172103497394



We can see that Our model predicts very well to actual value.

R-Score got is approximately same before and after tuning which is 88.68%, which is the best score to accurately predict Rainfall.

**Concluding Remark**

we have created model which is predict Rain will occur and how much Rainfall will be there. But as we all know, climate factors can change for a variety of reasons, by extracting and utilizing the hidden knowledge from prior meteorological data, data mining algorithms can predict rainfall. To achieve more accurate Result, we have to concentrate feature engineering and do focus on null values, skewness, correlation, collinearity.

we did Data Analysis, EDA and Feature Engineering carefully se we can have solved given problem by same dataset.

and by using hyper parameter we have improve our model accuracy and avoid Overfitting and get best model. we have to choose correct parameters accordingly classifier and regression model to get best accurate model.

Rainfall play crucial role in Agricultural work, aviation mission, and tourism sector many other aspect as well as the rest of the world. Our ML model can be predicting rainfall based on an analysis of a rainfall dataset. So that we can forecast rain in the coming year based on given features, which is extremely beneficial to farmers for agricultural purposes.