Chennai Reservoir Rainfall Analytics

M. Rohith

Introduction:

Prediction of weather is one of the tedious jobs in the field of machine learning and data analytics. Machine Learning algorithms are mostly useful in predicting rainfall. Some of the major Machine Learning algorithms are ARIMA Model (Auto-Regressive Integrated Moving Average). In this report we will see the insights and prediction of Chennai rainfall.

Sections:

For analysis purpose, the entire report is categorized into four categories. They are

1. **Data processing and analysis in Google Colab.**
2. **Data visualization using Tableau**
3. **ARIMA Modeling using Google Colab**
4. **Auto AI – Time series experiment using IBM Watson studio.**

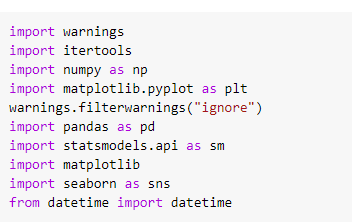
**Section 1 - Data processing and analysis in Google Colab**

**Data set collection:**

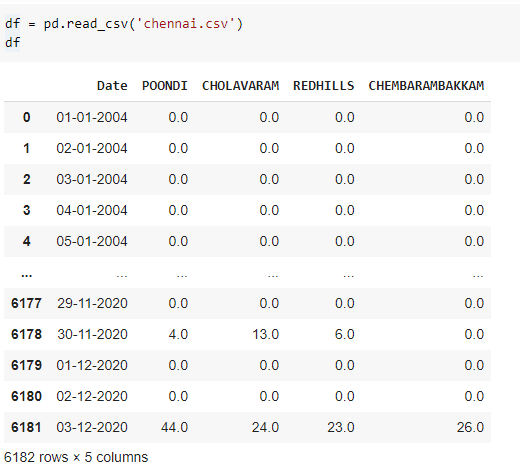
The data set is collected from government website and preprocessed using python language.

**Codes and output:**

**1. Importing the necessary libraries**

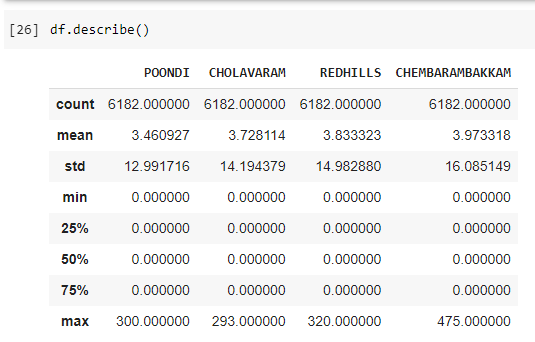


**2. Importing the dataset**



**Insights:** The dataset consists of rainfall measured in four different reservoirs of Chennai.

**3. Describe the dataset**

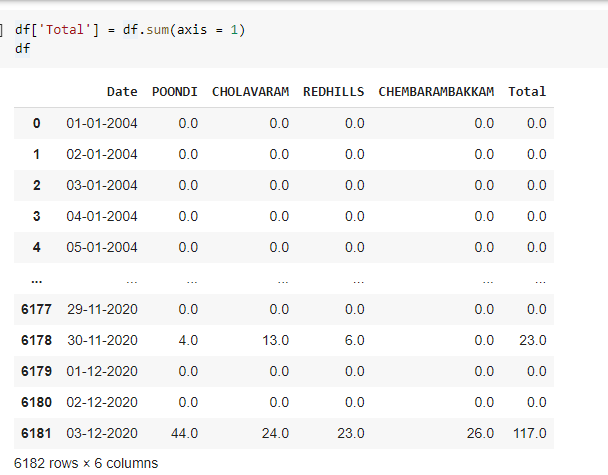


**Insights:** The dataset has a count of 6182 rows and 5 columns with minimum value as 0. The standard deviation lies between the intervals according to the reservoir.

**4. Creating a new column**

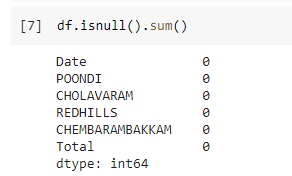
Now let us create a new column in the dataset. This Column is the sum of all the other four columns

The resulting dataset is as below.



**5. Check null values**

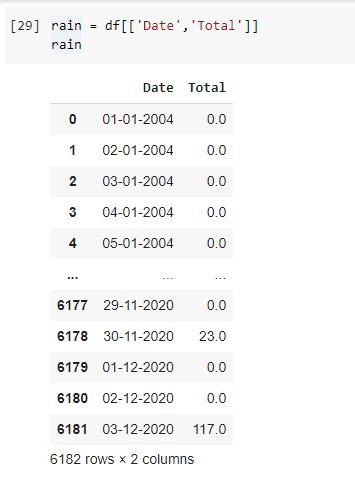
Let us check for null values



**Insight**: The dataset does not contain any null values.

**6. Sub setting the dataset**

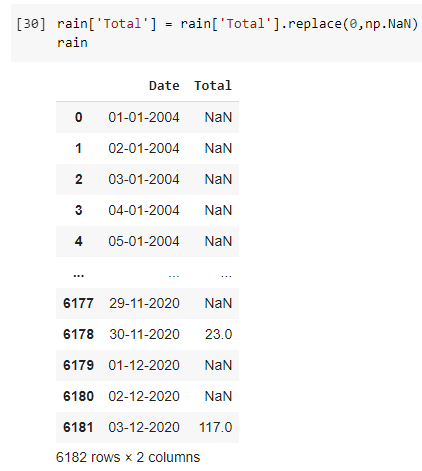
For easier analysis the dataset is made as a subset with Date and Total columns alone.



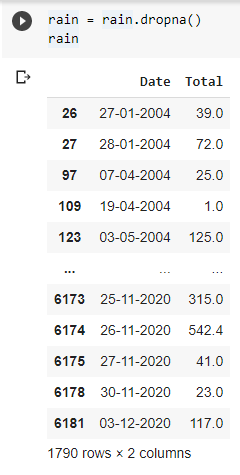
**7. Replacing values**

If we observe carefully, our dataset consists of 0 values which mean that there is no observation of rain on that particular date. Hence we have to replace it with NaN values for calculation.

The following result is obtained.



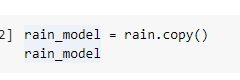
**8. Removing NaN values**



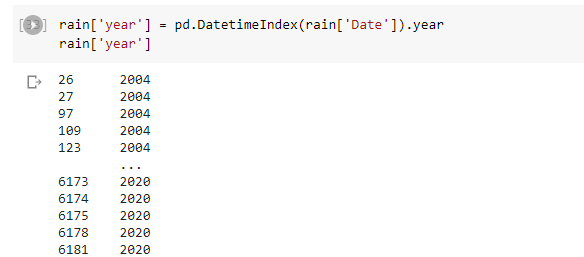
**Insights:** Once we have removed NaN values we get a total of 1790 rows. Hence we will use only this for our further analysis.

**9. Feature engineering**

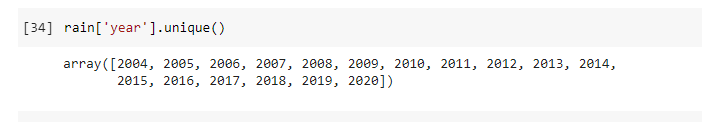
Let us copy the dataset without any moderation to a new variable.



Let us create a new date column with only year as a feature



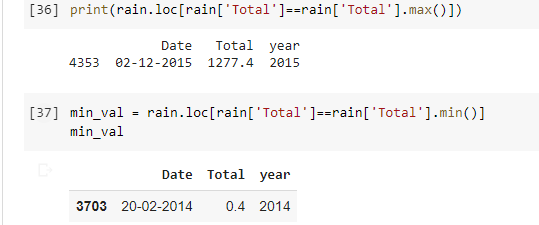
**10. Observation of unique values**



There are unique values from 2004 to 2020 without any intervention of the years

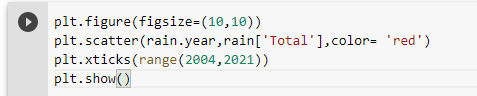
**11. Min value of rainfall and max value of rainfall**

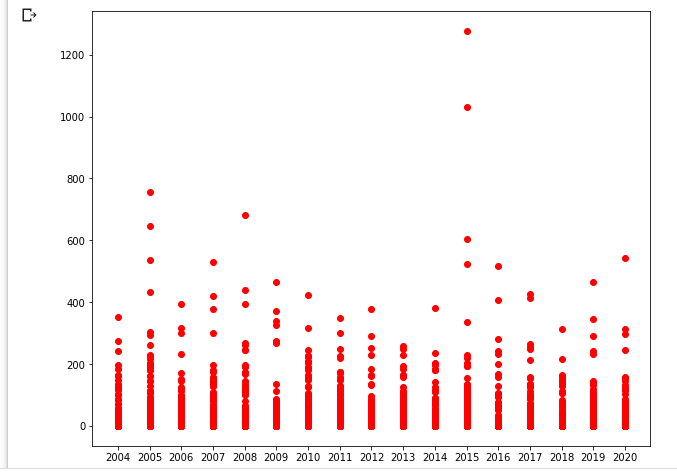
Let us find which date recorded the lowest and highest amount of rainfall.



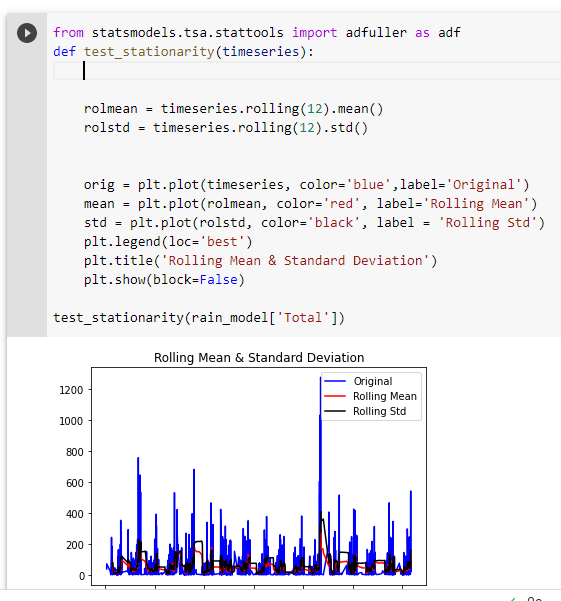
**Insights:** Minimum rainfall is recorded on 20-02-2014 and the maximum is recorded on 02-12-2015.

**12.** Individual dates analysis

****

****

**13. Time series of rolling mean and STD**

****

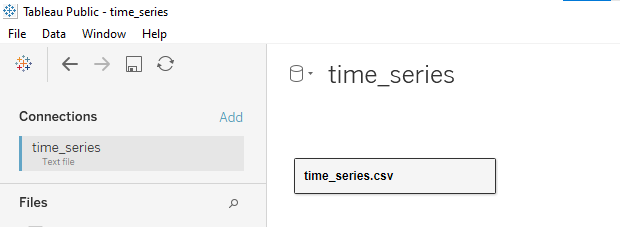
The mean value for the data is plotted and the average value is taken in this. The rolling mean is the average mean and the standard deviation is the std which rolls around the time-series analysis.

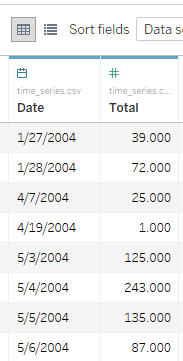
**Section 2 Data visualization using Tableau**

**Introduction:**

Now the dataset is processed. Let us have some visualization in Tableau for better understanding. Tableau provides us an interactive interface of visualization for better understanding. This will help us to indentify many possible insights.

**Loading the dataset**

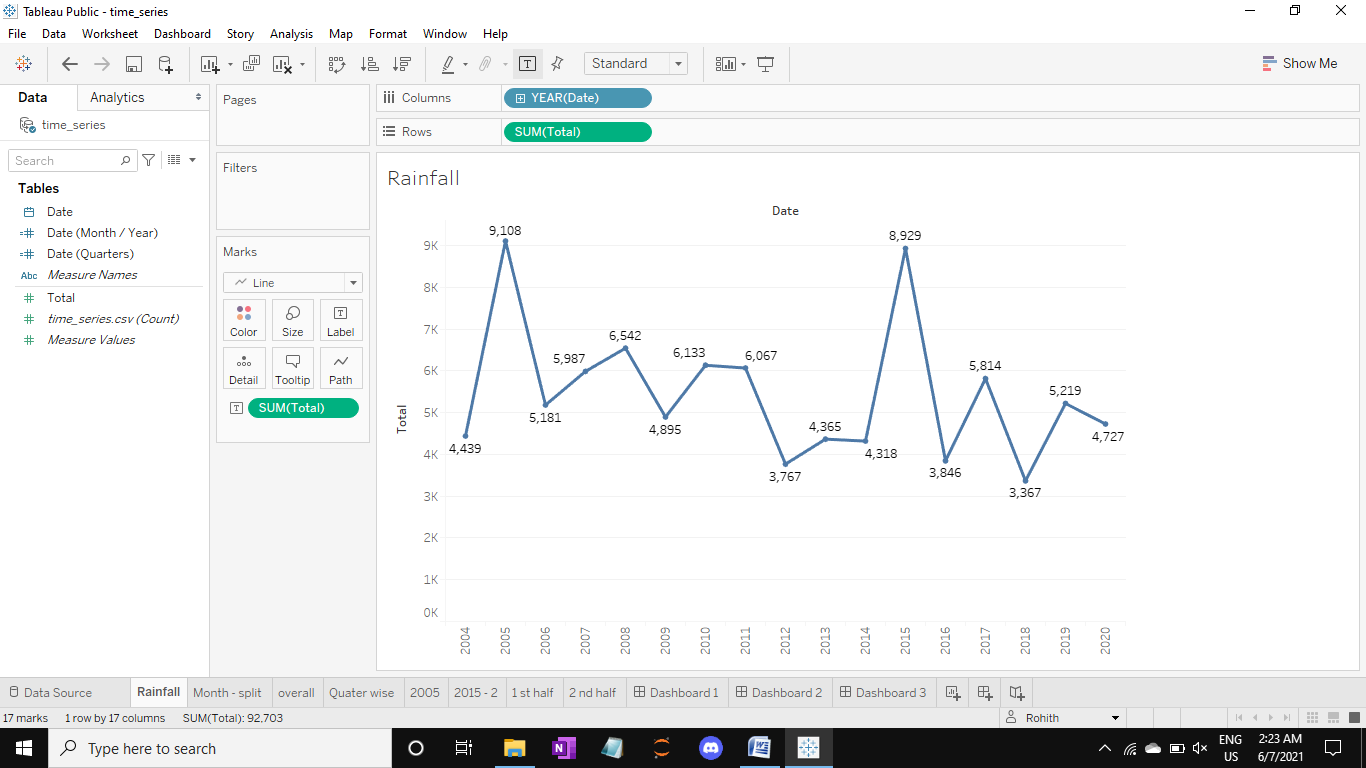




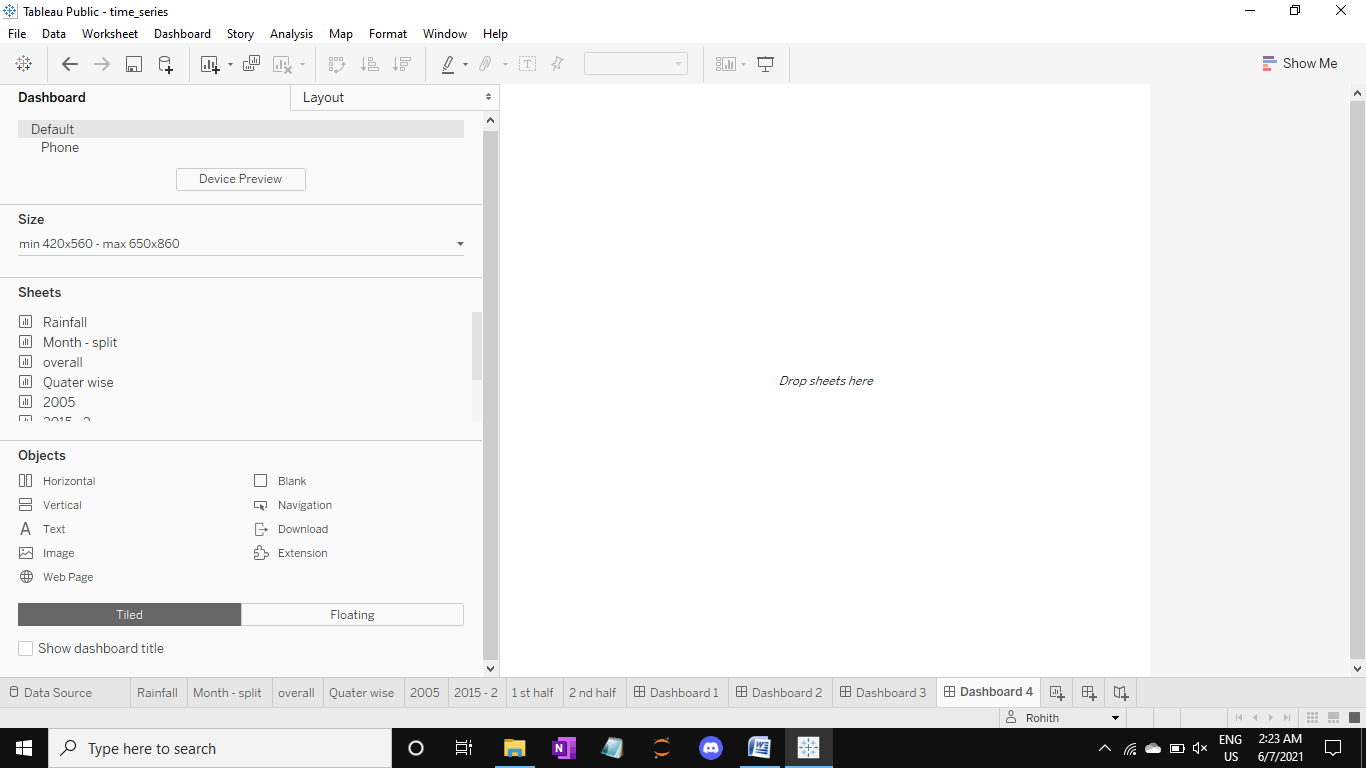
The dataset is successfully loaded in tableau public.

**Tableau interfaces**

**1. Worksheets**

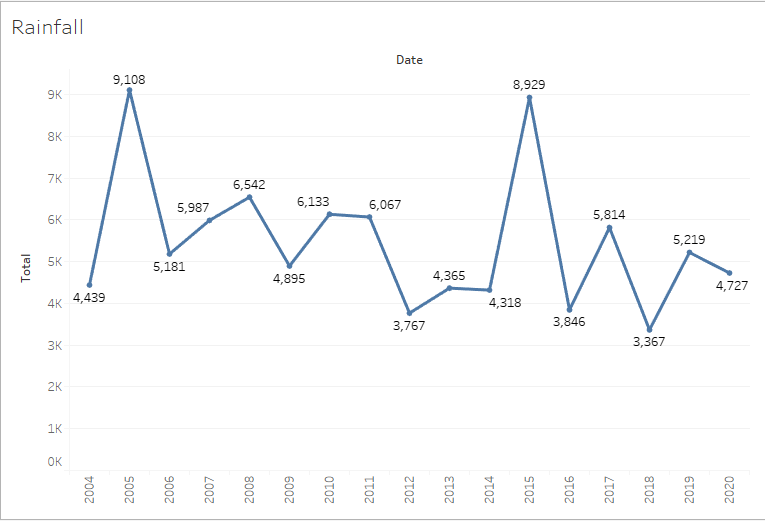
****

**2. Dashboard**

****

**Visualizations and insights**

1. Total rainfall



**Insights:**

1. The overall rainfall is aggregated and plotted for each year

2. The highest amount of rainfall is recorded during 2005 and the lowest was recorded during 2018.

3. The second highest was recorded during 2015 and the second lowest was recorded during 2012.

4. During the years from 2010 to 2014 there is a gradual decrease of rainfall.

5. From 2006 to 2008 there was a normal increase in the amount of rainfall.

**2. Month wise average rainfall**

****

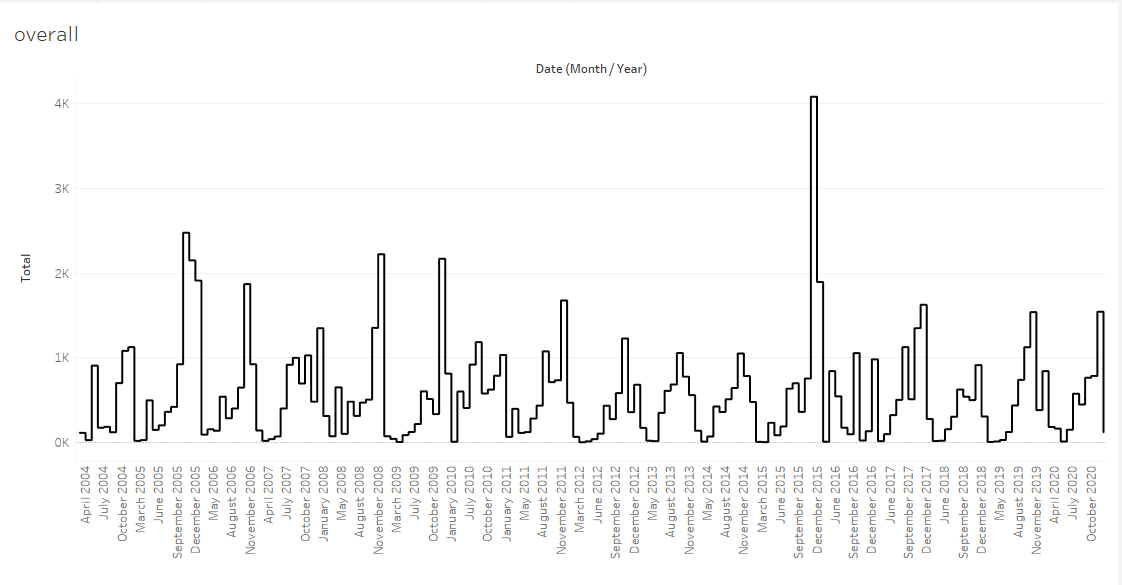
**Insights:**

1. Considering all the months, **November** month receives higher amount of rainfall.

2. The average amount of rainfall is less during **June.**

3. Chennai’s summer season lies between the months of March to June. Highest degree of temperature is recorded during May. But due to the southeast monsoon Chennai receives rainfall.

4. Due to cyclones and low pressure depressions greater amount of average rainfall is recorded during the months of October to December.

3. Overall rainfall segregation

**Insights**

1. **November 2015** received highest amount of rainfall.

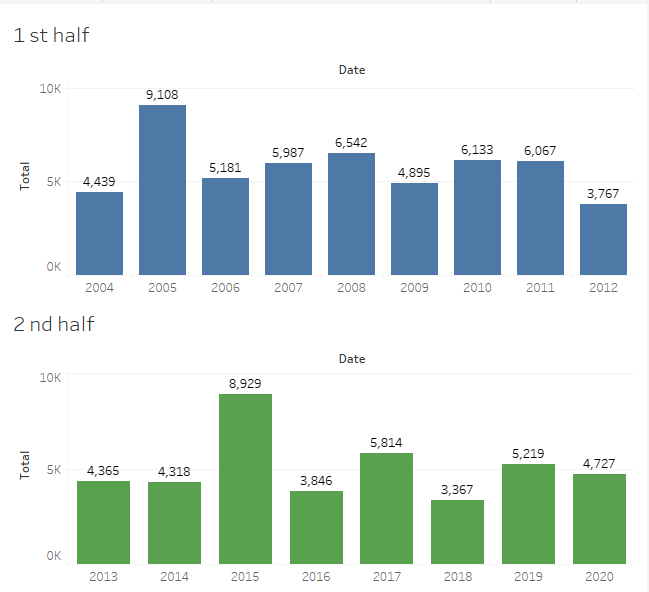
2. During all the years a steep amount of increase in rainfall is observed from September to December.

3. A pattern is observed in the dataset. That is a repetitive of events occurs.

4. A similar pattern is observed in consecutive years of 2013 and 2014.

For further analysis let us see whether the pattern is absolute in the recurrent years.

**4. Split up analysis**



**Insights:**

1. If we compare the data from 2004 – 2012 to 2013 – 2020 we can see a pattern in that.

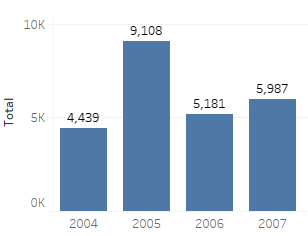
2. 2005 and 2015 records the highest amount of rainfall. This shows that once in 10 the probability of getting a heavy rain is high.

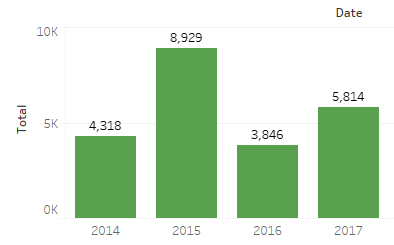
3. In addition to that hypothesis we can also see that before that selected year there is a less amount of rainfall which is same as 2004 and 2014.

4. The next thing is that 2006 and 2016 infers a less rainfall and in the consecutive years of both the rainfall slightly increases.

5. Hence forth, a thesis is formed that 4 years pattern is observed in a span of 10 years.

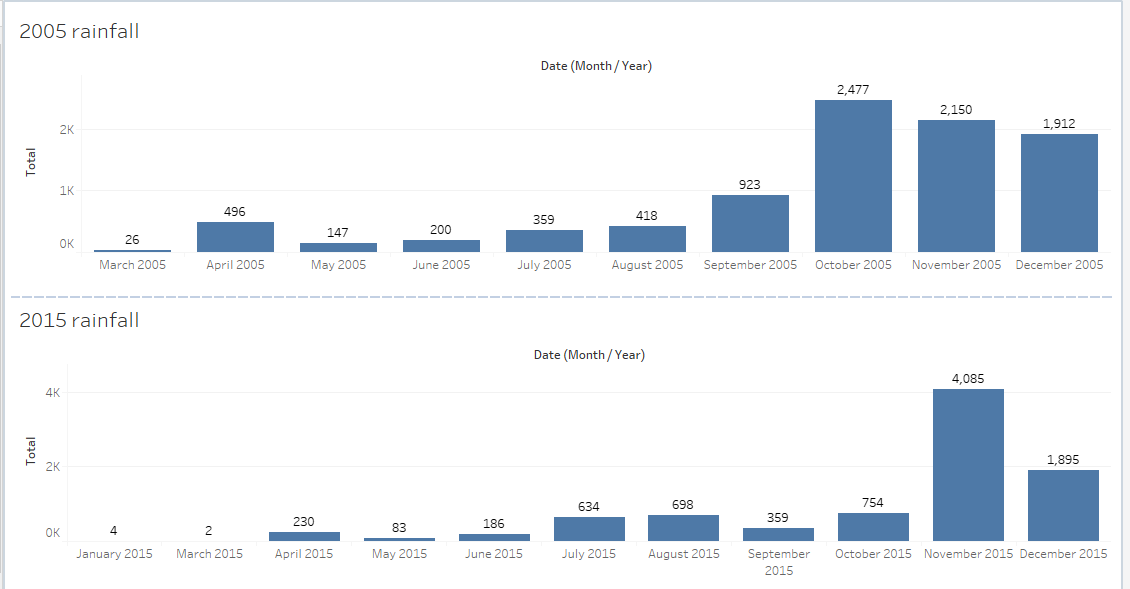
6. For further zooming and analyzing the four year trend is shown below.





In these two graphs, we can see a trend and a pattern.

**5. 2005 – 2015 individual analysis**



**Insights:**

1. In 2005, during the month of October, November and December there are high amount of rainfall.

2. The total amount of rainfall that we recorded during 2005 is recorded in a single month, i.e November 2015.

3. But eventually 2005 received highest amount of rainfall than 2015.

4. It is observed that the rainfall doesn’t suddenly come down but gradually gets down during the month of December.

**Section 3 ARIMA Modeling using Google Colab**

**ARIMA Model:**

ARIMA, short for ‘Auto Regressive Integrated Moving Average’ is actually a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Any ‘non-seasonal’ time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q

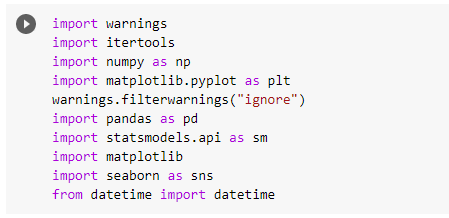
where,

p is the order of the AR term

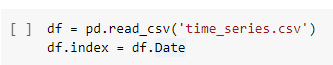
q is the order of the MA term

d is the number of differencing required to make the time series stationary.

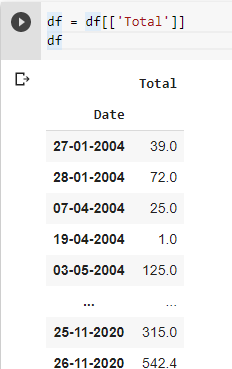
**Importing necessary libraries**



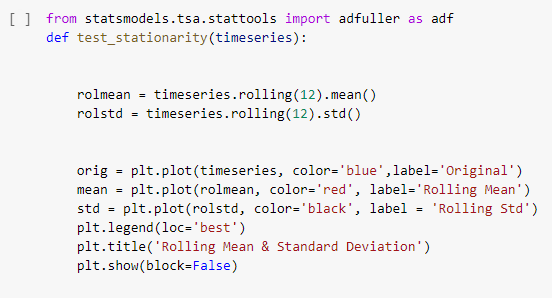
**Reading the dataset**

****

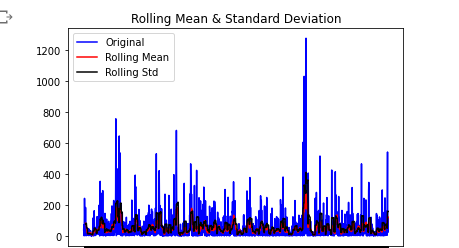
**Sub setting the dataset**

****

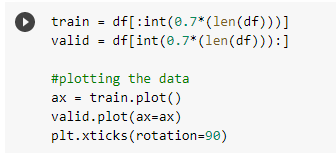
**Build a Time series model with rolling mean and STD**

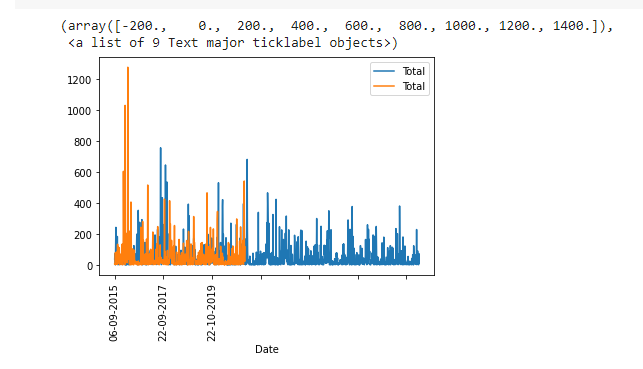
****

**Output**

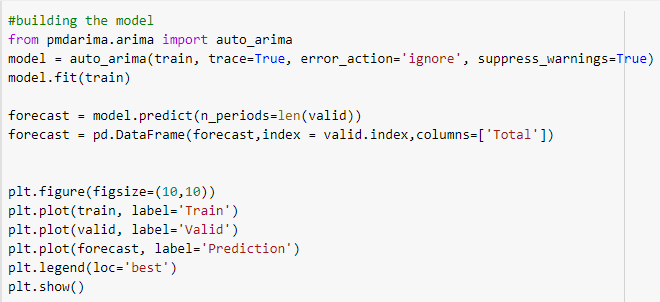
****

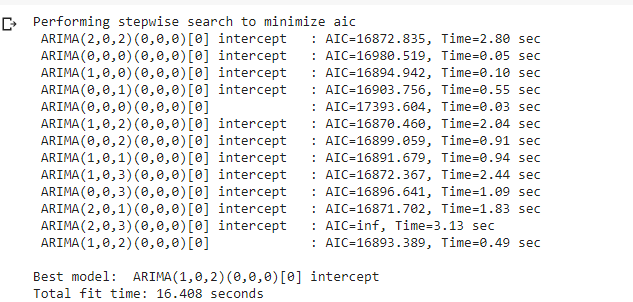
**Splitting the dataset into train and test**

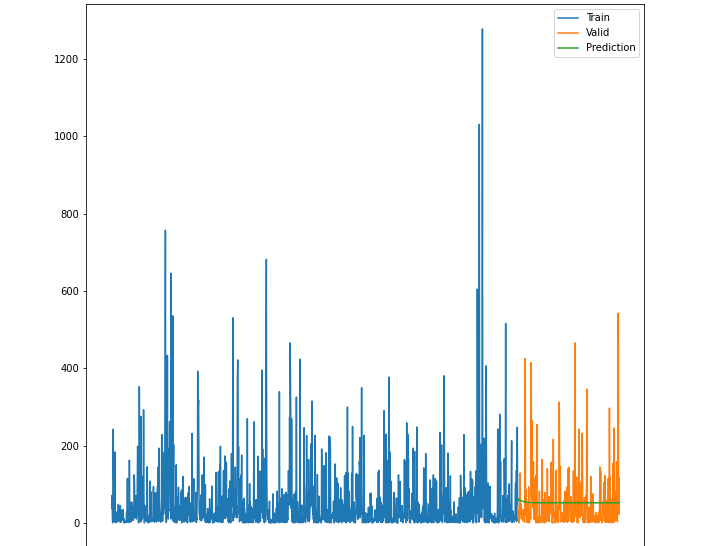
****

****

**Building the ARIMA Model**

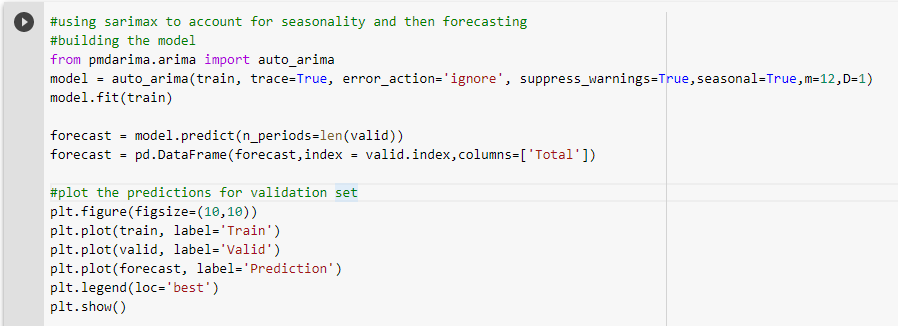
****

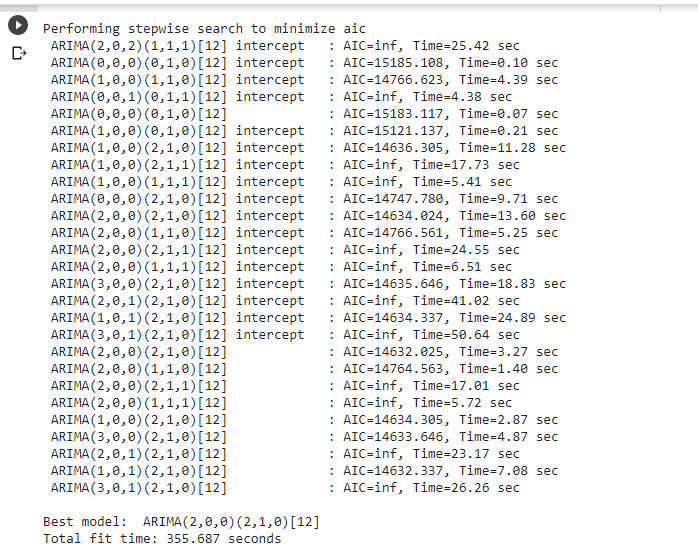
****

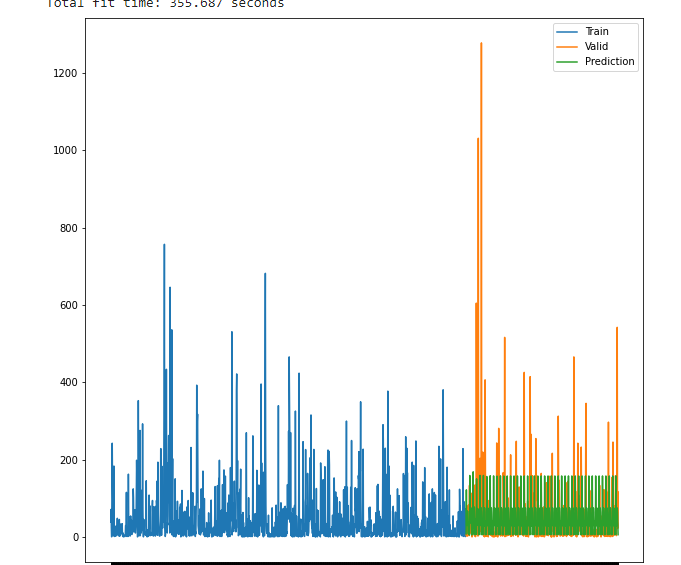
****

**We can see that the green line fits a bit but not accurately because of the training set. But the best fit it has in the graph.**

**Let check for seasonability and forecasting .**

****

****

****

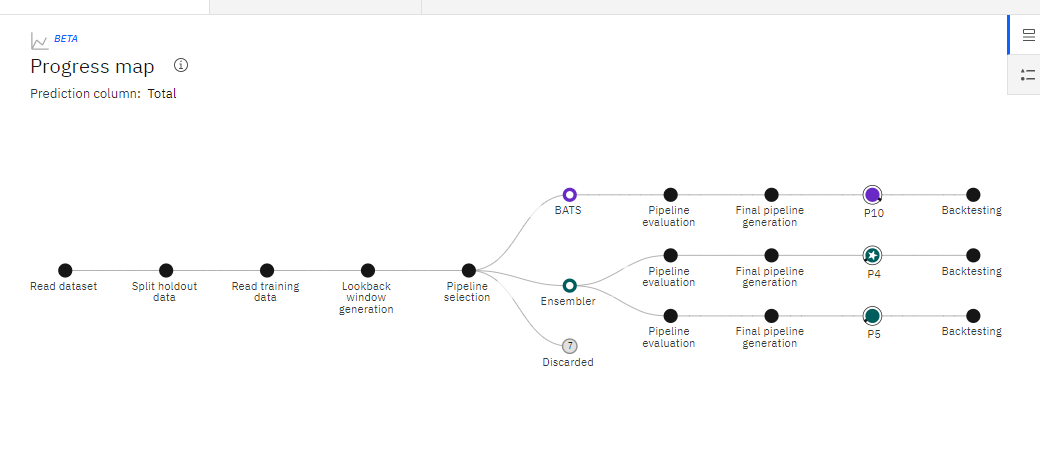
**The season ability model runs and gives us a god impression on our dataset.**

**Section 4 Auto AI – Time series experiment using IBM Watson studio.**

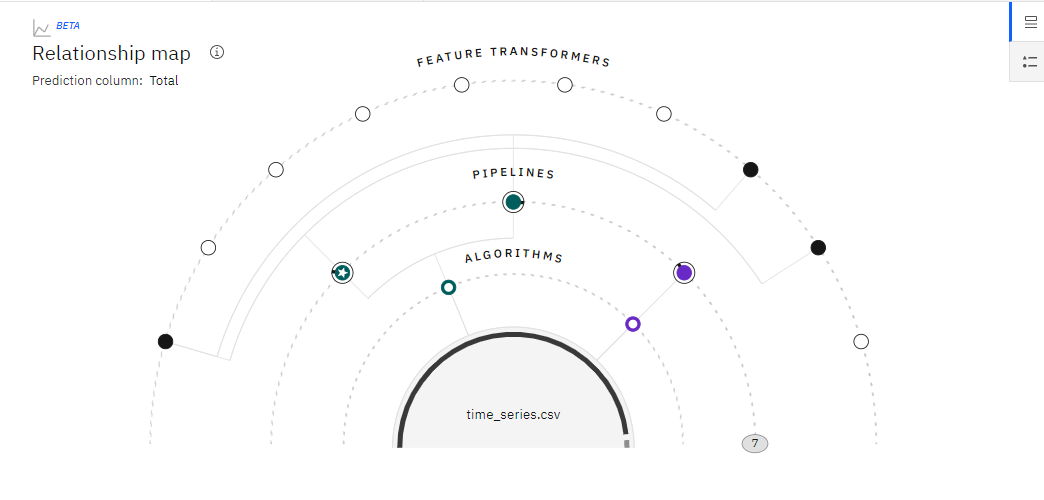
Though we have performed our analysis in coding and visualization method, let us run an auto AI experiment of time series analysis using IBM Watson studio.

The Auto AI experiment results are discussed below.

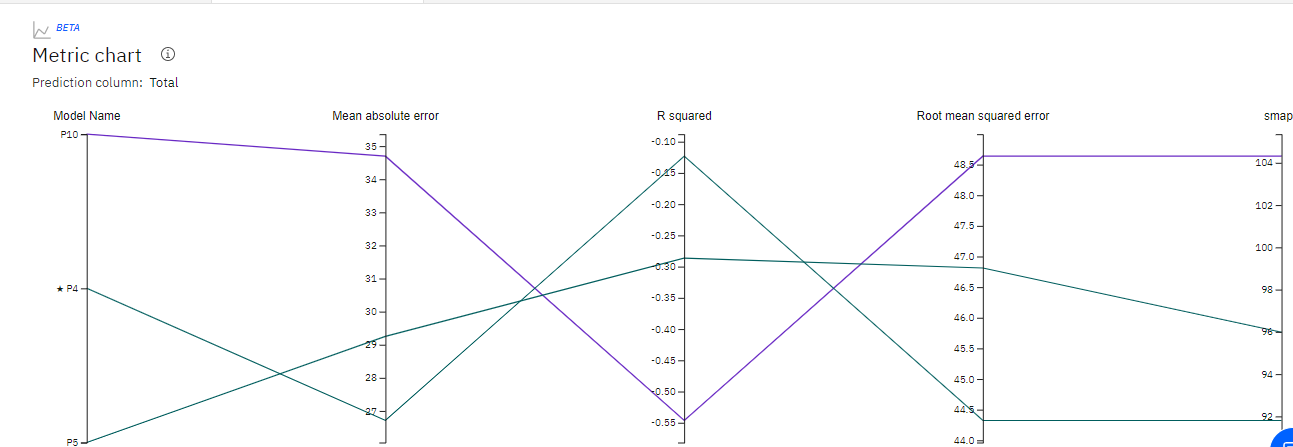
**Progress map for the entire pipeline**



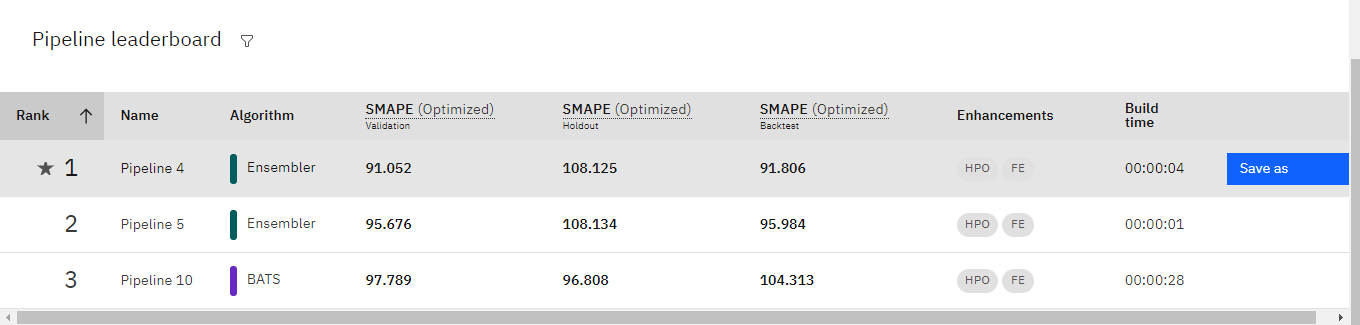
**Relationship map**



**Metric chart**



**Pipeline leader board**



This shows that ensemble algorithm has a better control over the dataset and holds out for a good accuracy.

**Evaluation metrics score**

