International Institute of Information Technology, Bangalore

GEN - 511 Machine Learning Hackathon

RAINFALL PREDICTION

SHUBHAM KUMAR MT2019109
SHUBHAM LAMICHANE MT2019110
SUMIT RANA MT2019118

November 24, 2019

Contents

- **Introduction**
- **Dataset Description**
  Features
  Data Distribution
- **Data Visualization**
- **Model Building**
- **Results**
- **Conclusion**

Abstract

Accurate forecasting of rainfall has been one of the most important issues in hydrological research because early warnings of severe weather can help prevent casualties and damages caused by natural disasters, if timely and accurately forecasted. It can also help in knowing when to plant crops, when to build and when to prepare for drought and flood. To construct a predictive system for accurate rainfall, forecasting is one of the greatest challenges to researchers from diverse fields such as Weather Data Mining, Environmental Machine Learning, Operational Hydrology, and Statistical Forecasting. A common question in these problems is how one can analyse the past and use future prediction. The parameters that are required to predict rainfall are enormously complex and subtle even for a short term period. Scarcity or heavy, both rainfall effects rural and urban life to a great extent with the changing pattern of the climate. Unusual rainfall and long lasting rainy season is a great factor to take account into.

Introduction

Monsoon prediction is clearly of great importance for India. Two types of rainfall predictions can be done, they are :-

- Long term prediction of annual rainfall in the country
- Short term prediction of monthly rainfall in a subdivision or country

Indian meteorological department provides forecasting data required for project.In this project we have tried to make a short term prediction of rainfall both at the centre and the state level. The main motive of the project is to predict the amount of rainfall in a particular division or state well in advance. We predict the amount of rainfall using past data. The dataset has been taken from **data.gov.in**. It contains information about annual and monthly rainfall from 1901-2015.

We are trying to make two models here :

- One for an all India prediction of rainfall.
- Second for prediction of rainfall in the state of Karnataka.

Dataset Description

The Rainfall Prediction dataset has average rainfall from 1901-2015 for each district, for every month and contains the following set of features :-

1. Sub-Division - The dataset has 36 sub-divisions and for some of the sub-divisions, data is from 1950-2015. The subdivisions are made by the Indian Meteorological Department on the basis of size of the state and proximity to rain.Coastal areas like Karnataka, Kerala, Tamil Nadu are divided into many divisions. North eastern states which get a lot of rainfall are also subdivided. Other smaller states are combined.

2. Year - The year for which the data is collected.

3. Months(From Jan to Dec) - Average rainfall in each month measured in mm.

4. Annual - Annual rainfall measured in mm.

5. Cumulative Months (Jan-Feb, Mar-Apr-May, June-July-Aug-Sep, Oct-Nov-Dec) - Measure of rainfall in cumulative months.

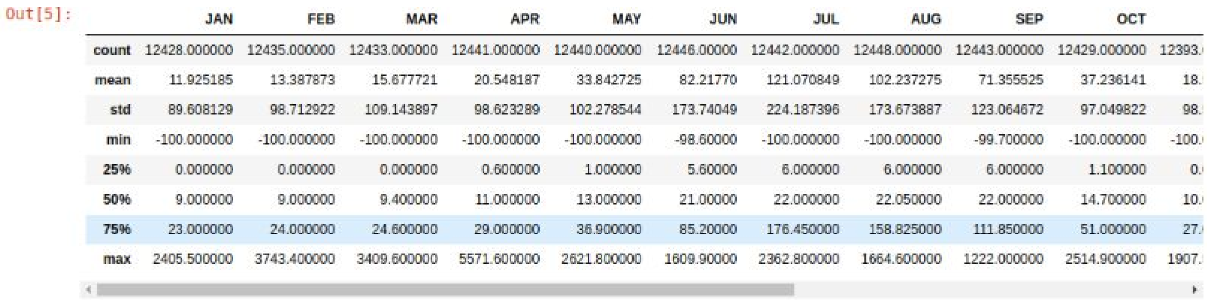
All the attributes has the sum of amount of rainfall in mm.

Methodology

- Converting data into the correct format to conduct experiments.
- Make a good analysis of data and observe variation in the patterns of rainfall.
- Finally, we try to predict the average rainfall by separating data into training and testing. We apply various statistical and machine learning approaches(*SVR*, Linear Regression etc) in prediction and make analysis over various approaches. By using various approaches, we try to minimize the error.

Data Visualization

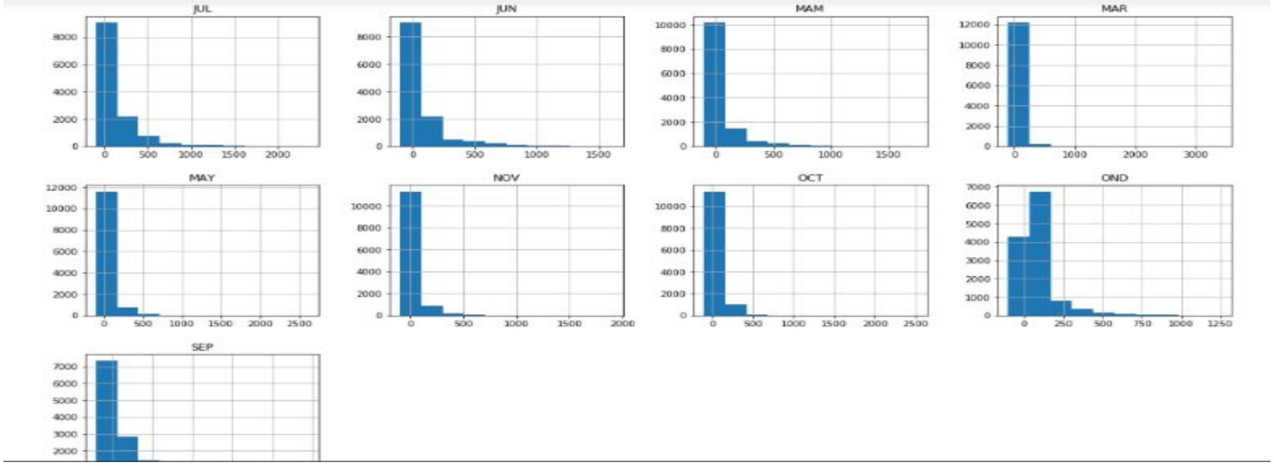
Description of the dataset reveals there are around 12k rows in the set. There are around 18 attributes which tells the variation of rainfall depending on the subdivision, year and month.



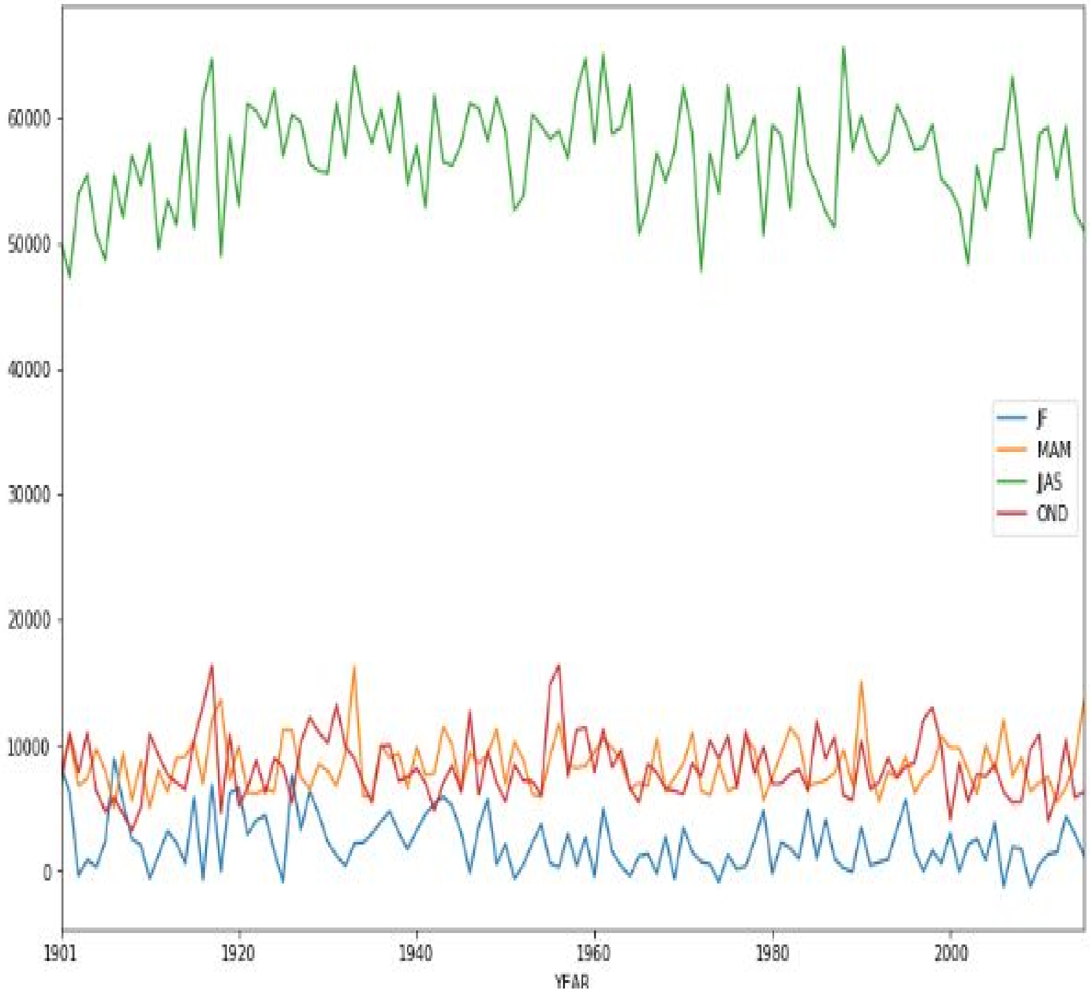
Out[5]:

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	
count	12428.000000	12435.000000	12433.000000	12441.000000	12440.000000	12446.00000	12442.000000	12448.000000	12443.000000	12429.000000	12393.
mean	11.925185	13.387873	15.677721	20.548187	33.842725	82.21770	121.070849	102.237275	71.355525	37.236141	18.
std	89.608129	98.712922	109.143897	98.623289	102.278544	173.74049	224.187396	173.673887	123.064672	97.049822	98.
min	-100.000000	-100.000000	-100.000000	-100.000000	-100.000000	-98.60000	-100.000000	-100.000000	-99.700000	-100.000000	-100.
25%	0.000000	0.000000	0.000000	0.600000	1.000000	5.60000	6.000000	6.000000	6.000000	1.100000	0.
50%	9.000000	9.000000	9.400000	11.000000	13.000000	21.00000	22.000000	22.050000	22.000000	14.700000	10.
75%	23.000000	24.000000	24.600000	29.000000	36.900000	85.20000	176.450000	158.825000	111.850000	51.000000	27.
max	2405.500000	3743.400000	3409.600000	5571.600000	2621.800000	1609.90000	2362.800000	1664.600000	1222.000000	2514.900000	1907.

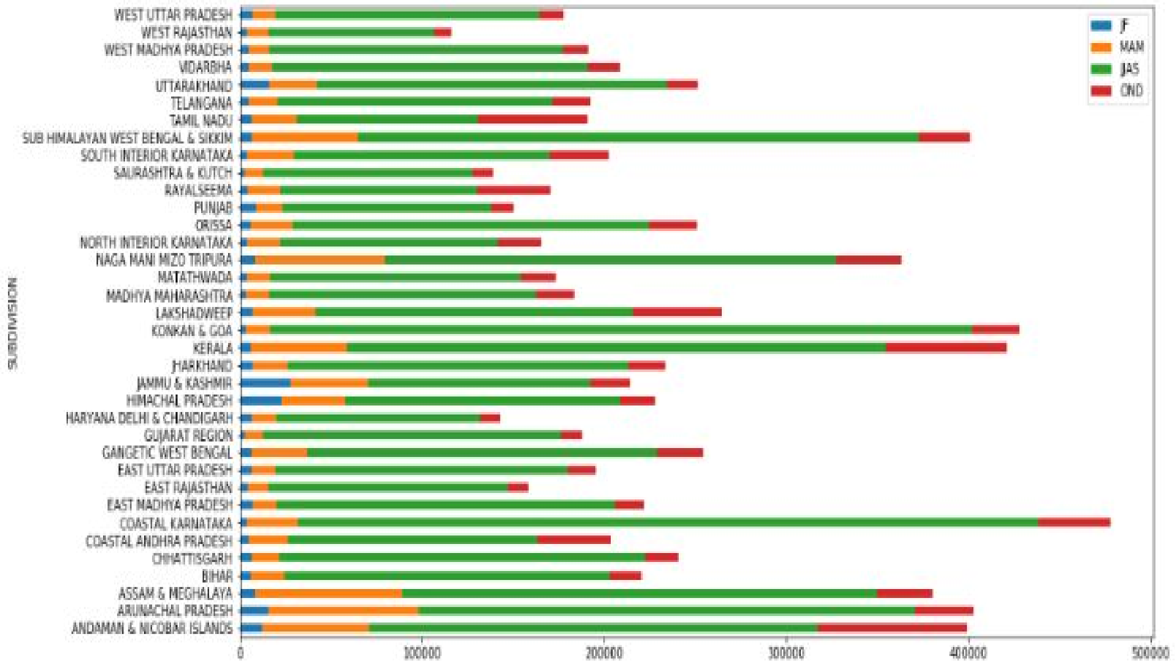
The histogram given below measures the average rainfall monthly and quarterly. It is quite clear that the months of June, July, August, September receives the highest amount of rainfall.The months of January, February is the driest in the country. There has been occasional showers in the months of November, December in some parts of the country.



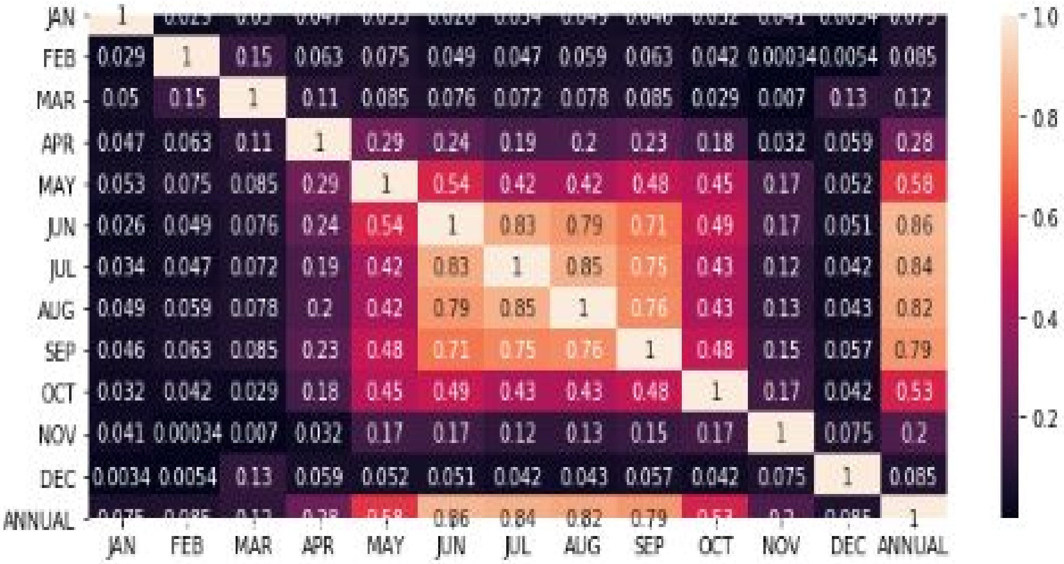
This plot gives us an idea about the amount of rainfall on a quarterly basis. It is again quite clear that throughout the years the month of June-July-August-September(shown in green) receives much more rainfall than any other quarter.This is considered to be the monsoon season in most parts of India.



This bar plot shows the distribution of rainfall in every subdivision in each quarter. As expected June-July-Aug-Sep (shown in green) are prominent for rain. In some North Eastern areas like Meghalaya and Assam, the month of March and April also witness considerable amount of rainfall. The coastal areas like Kerala, Karnataka; western ghats of Goa and Konkan region and North East experience highest rainfall in the country. Rajasthan, Delhi, Haryana are relatively dry.

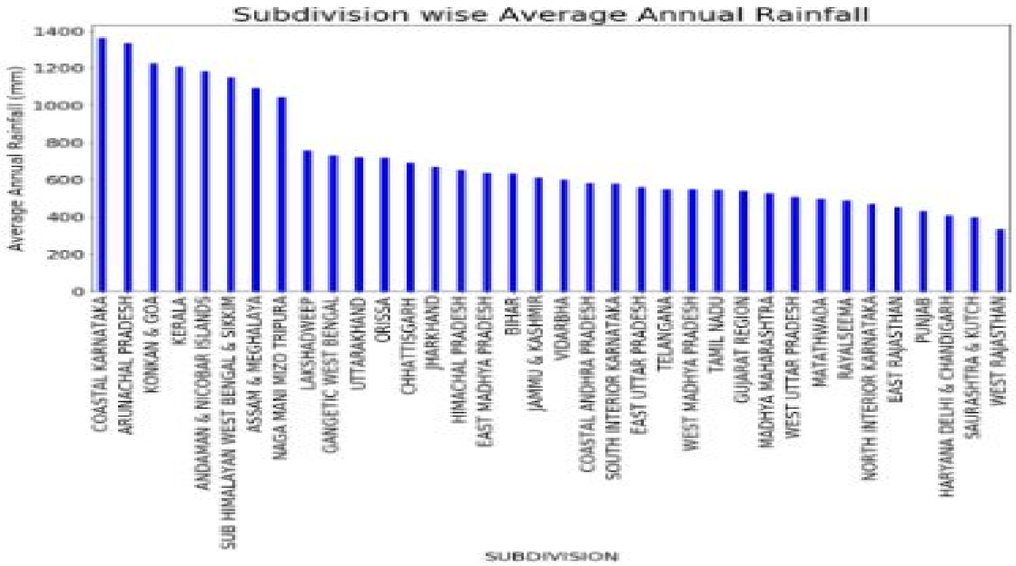


The correlation matrix below shows what all the other graphs have kind of predicted. The fact that the months of monsoon from June to September has the highest correlation with the annual rainfall is very obvious. Also months of May and October have strong correlation to the annual rainfall indicating some parts of the country receives rainfall in these months.

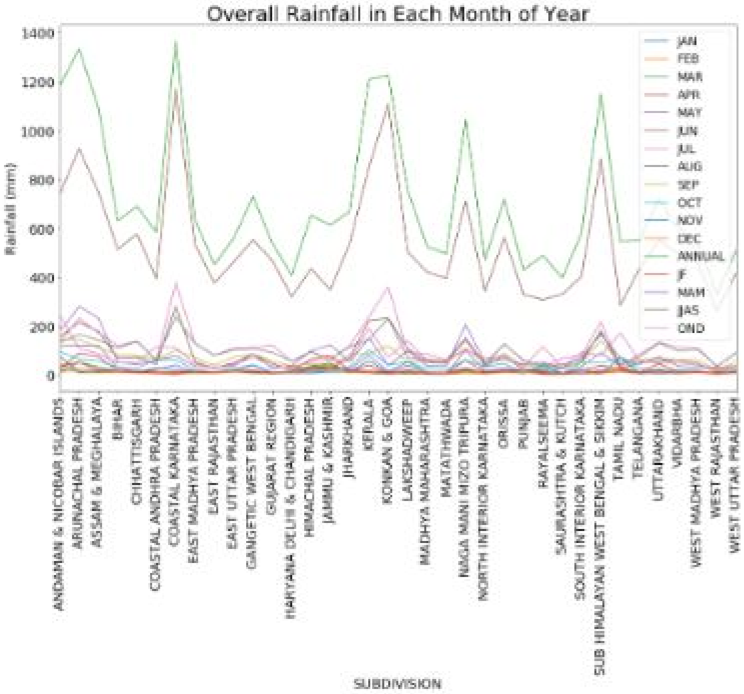


	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL
JAN	1	0.029	0.05	0.047	0.053	0.026	0.034	0.049	0.046	0.032	0.041	0.0034	0.075
FEB	0.029	1	0.15	0.063	0.075	0.049	0.047	0.059	0.063	0.042	0.00034	0.0054	0.085
MAR	0.05	0.15	1	0.11	0.085	0.076	0.072	0.078	0.085	0.029	0.007	0.13	0.12
APR	0.047	0.063	0.11	1	0.29	0.24	0.19	0.2	0.23	0.18	0.032	0.059	0.28
MAY	0.053	0.075	0.085	0.29	1	0.54	0.42	0.42	0.48	0.45	0.17	0.052	0.58
JUN	0.026	0.049	0.076	0.24	0.54	1	0.83	0.79	0.71	0.49	0.17	0.051	0.86
JUL	0.034	0.047	0.072	0.19	0.42	0.83	1	0.85	0.75	0.43	0.12	0.042	0.84
AUG	0.049	0.059	0.078	0.2	0.42	0.79	0.85	1	0.76	0.43	0.13	0.043	0.82
SEP	0.046	0.063	0.085	0.23	0.48	0.71	0.75	0.76	1	0.48	0.15	0.057	0.79
OCT	0.032	0.042	0.029	0.18	0.45	0.49	0.43	0.43	0.48	1	0.17	0.042	0.53
NOV	0.041	0.00034	0.007	0.032	0.17	0.17	0.12	0.13	0.15	0.17	1	0.075	0.2
DEC	0.0034	0.0054	0.13	0.059	0.052	0.051	0.042	0.043	0.057	0.042	0.075	1	0.085
ANNUAL	0.075	0.085	0.12	0.28	0.58	0.86	0.84	0.82	0.79	0.53	0.2	0.085	1

This graph indicates that Subdivisions with highest annual rainfall are "Arunachal Pradesh", "Coastal Karnataka" and "Konkan & Goa" with approximate annual rainfall of 3418mm, 3408mm and 2977mm respectively. On the other hand Subdivisions with lowest annual rainfall are "West Rajasthan", "Saurashtra & Kutch" and "Haryana Delhi & Chandigarh" with approximate annual rainfall of 292mm, 495mm and 530mm respectively.



This graph gives the Overall Rainfall in Each Month. The monsoon months are quite clearly ahead of other months, indicating most of the country receives rainfall during this season.



Model Building

From all the plots and graphs and the correlation matrix it is clear that the months of June, July, August & September has the highest bearing on the annual rainfall in the country. It is also quite revealing that few months have almost no effect on the amount of rainfall in the country. Considering these two cases we built our models.

Linear Regression

Linear Regression is a linear model which tries to find the best fit line between one dependent and other independent variables. We used this algorithm to build our first model to predict rainfall. Our prediction was basically predicting the amount of rainfall in the fourth month given the rainfall in the first three months. So we made our dataframe accordingly, iterating through each triplets of months and predicting the value for the fourth month.
We divided our dataset in 80:20 split and ran our model on it.

Support Vector Regression

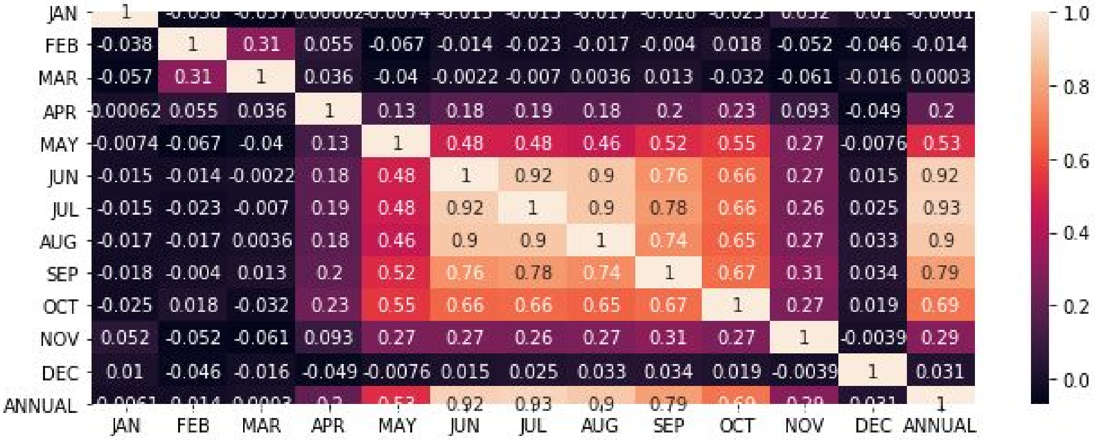
SVR uses the same basic ides as SVM, a classification algorithm, but applies it to predict real values rather than a class. SVR acknowledges the presence of non-linearity in the data and provides a proficient prediction model. SVR allows for non-linear fitting problems as well. We divided the dataset same as before and ran the model.

For error calculations we used Mean Absolute Difference(MAD) and Mean Square Error(MSE). MAD is a good statistic to use when analyzing the error of a single item.

Apart from doing the prediction on pan-India level, we also predicted the rainfall for the state of Karnataka. Karnataka is subdivided into 3 subdivisions mainly:

- Coastal Karnataka
- North Interior Karnataka
- South Interior Karnataka

Looking at the correlation matrix and the other plots for these 3 subdivisions, the monsoon months of June-September has even more prominence in this state.



	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL
JAN	1	-0.038	-0.057	0.00062	-0.0074	-0.015	-0.015	-0.017	-0.018	-0.025	0.052	0.01	-0.0061
FEB	-0.038	1	0.31	0.055	-0.067	-0.014	-0.023	-0.017	-0.004	0.018	-0.052	-0.046	-0.014
MAR	-0.057	0.31	1	0.036	-0.04	-0.0022	-0.007	0.0036	0.013	-0.032	-0.061	-0.016	0.0003
APR	0.00062	0.055	0.036	1	0.13	0.18	0.19	0.18	0.2	0.23	0.093	-0.049	0.2
MAY	-0.0074	-0.067	-0.04	0.13	1	0.48	0.48	0.46	0.52	0.55	0.27	-0.0076	0.53
JUN	-0.015	-0.014	-0.0022	0.18	0.48	1	0.92	0.9	0.76	0.66	0.27	0.015	0.92
JUL	-0.015	-0.023	-0.007	0.19	0.48	0.92	1	0.9	0.78	0.66	0.26	0.025	0.93
AUG	-0.017	-0.017	0.0036	0.18	0.46	0.9	0.9	1	0.74	0.65	0.27	0.033	0.9
SEP	-0.018	-0.004	0.013	0.2	0.52	0.76	0.78	0.74	1	0.67	0.31	0.034	0.79
OCT	-0.025	0.018	-0.032	0.23	0.55	0.66	0.66	0.65	0.67	1	0.27	0.019	0.69
NOV	0.052	-0.052	-0.061	0.093	0.27	0.27	0.26	0.27	0.31	0.27	1	-0.0039	0.29
DEC	0.01	-0.046	-0.016	-0.049	-0.0076	0.015	0.025	0.033	0.034	0.019	-0.0039	1	0.031
ANNUAL	-0.0061	-0.014	0.0003	0.2	0.53	0.92	0.93	0.9	0.79	0.69	0.29	0.031	1

In Fact the month of June and July pretty much determine the annual rainfall in Karnataka.

Result

For India

Linear Regression : MAD = 60.15
MSE = 60.79

SVR : MAD = 68.7
MSE = 68.34

For Karnataka

Linear Regression : MAD = 65.33
MSE = 72.56

SVR : MAD = 72.97
MSE = 85.37

Conclusion

We visualized the data through various plots and correlation matrices, arriving at a set of dominant features. Then we used those features and fed them into their respective models. We used two modes Linear Regression and SVR. We did our prediction on annual rainfall country wide and also in particular state of Karnataka.