Rainfall Prediction in India

Mentor : Dr.Sreeja SR

project Code : B21SSR02

Topic : Rainfall Prediction

Department of Computer Science and Engineering
Indian Institute of Information Technology Sri City

Rainfall Prediction

Group-20:
(Team Members)

G.Sai Divya(S20180010059)

G.Sai Mani(S20180010057)

R.Bhavani(S20180010024)

Outline of Presentation:

- Problem Statement
- Challenges/Motivation
- Literature Survey
- Dataset Description
- Missing value methods
- Visualization
- Machine learning techniques
- Work done upto know
- Future Plan
- References

Problem Statement:

- To Study the seasonal variations of rainfall and Annual Average Rainfall Variation of years range from 1901 to 2015 with month wise distribution of rainfall in mm for all major Indian states and Union Territories.
- Predict the Rainfall for upcoming years in India based on dataset we have .
- Display the predictions through website.

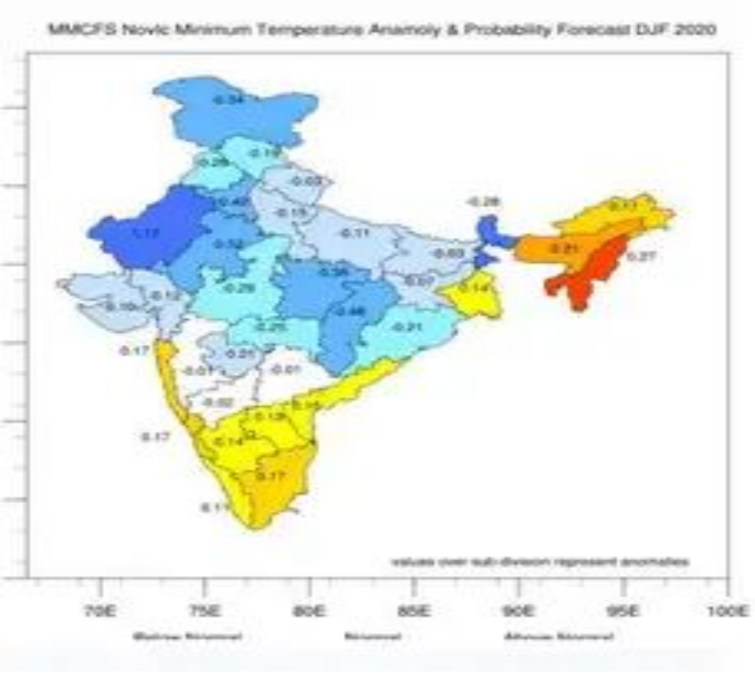
Challenges/Motivation:

- Rainfall is a key part of hydrological cycle and alteration of its pattern directly affect the water resources. The changing pattern of rainfall in consequence of climate change is now concerning issues to water resource managers and hydrologists.

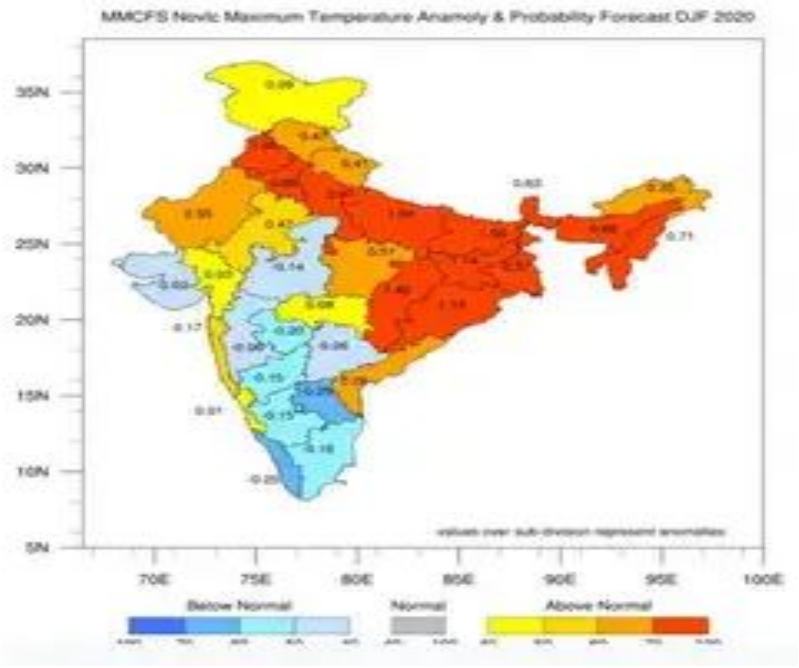
- The drought and flood like hazardous events can be occurred frequently because of the extreme changes of rainfall trend. the amount of soil moisture for crop production is totally determined by the amount of rainfall.

Challenges/Motivation:

- The monsoon rainfall plays a vital role for agriculture in India. Hence, the research on the climate change or most specifically on the changes of rainfall occurrences and its allocation are the most significant way for sustainable water resource management.



MMCFS Novic Minimum Temperature Anamoly & Probability Forecast DJF 2020
values over sub-division represent anomalies
70E
75E
80E
85E
90E
95E
100E
Below Normal
Normal
Above Normal



MMCFS Novic Maximum Temperature Anamoly & Probability Forecast DJF 2020
35N
30N
25N
20N
15N
10N
5N
values over sub-division represent anomalies
70E
75E
80E
85E
90E
95E
100E
Below Normal
Normal
Above Normal

Literature Survey:

Authors	**Published on, year**	**Technique**	**Dataset**	**Advantage**
Bushra Praveen et al., IIT INDORE [1]	Scientific Reports JUNE 2020	Artificial Neural Network-Multilayer Perceptron (ANN-MLP)	Indian Rainfall from 1901-2015	analyzes and forecasts the long-term Spatio-temporal changes
Nikhil Oswal., University of Ottawa,canada [2]	arXiv:1910.13827 October 2019	Linear classier, Tree based, Distance based, Rule based and Ensemble for original, undersampled, oversampled dataset	https://www.kaggle.com/jsphyg/weather-dataset-rattle-package	comparative study , concentrating on three aspects: modeling inputs, modeling methods, and pre-processing techniques.
Emilcy Hernandez et al., Coventry University UK [3]	International Conference on Hybrid Artificial Intelligence Systems. (2016)	Autoencoder and MLP, MSE RMSE errors	meteorological station located in a central area of Manizales, Colombia. (almost 30yrs)	Autoencoder- feature treatment in time series. MLP-prediction task

Literature Survey:

Authors	Published on, year	Technique	Dataset	Advantage
Pritpal Singh et al., Tezpur University [4]	*Stochastic Environmental Research and Risk Assessment* 13 Feb 2013 @Springer	ANN using BPNN with MLFF	time series data for 140 years (1871– 2010)	a model for prediction of ISMR on monthly and seasonal time scales, its applicability for advance prediction of the seasonal rainfall amounts.
Kumar Abhishek et al., NIT Patna [5]	Control and System Graduate Research Colloquium (ICSGRC), 2012 IEEE	BPA, LRN, CBP with different adaptive learning function, training functions	Data from 1960 to 2010.	The multilayered artificial neural network with learning by back-propagation algorithm configuration is the most common in use, due to of its ease in training

Dataset Description:

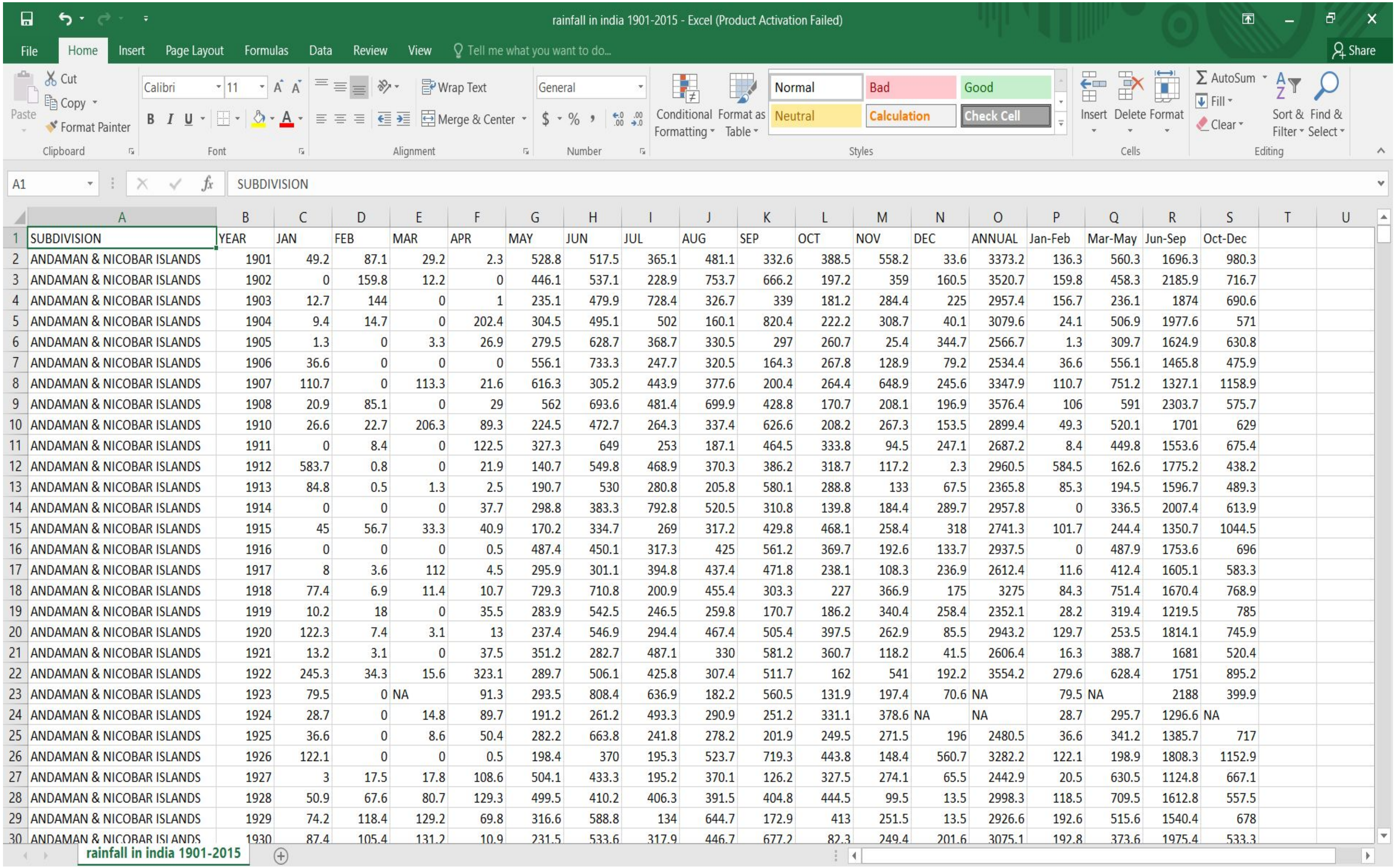
The data set consists of the measurement of rainfall from year 1901-2015 for each state.

- Data consists of 19 attributes (individual months, annual, and seasonal months) for 36 sub divisions.
  [subdivision, year, jan, feb, mar, apr, may, june, july, aug, sep, oct, nov, dec, annual, jan-feb, mar-may , jun-sep, oct-dec]

- The attributes are amount of rainfall measured in mm

Simple glimpse of data:

Dataset Description:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
2	ANDAMAN & NICOBAR ISLANDS	1901	49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2	136.3	560.3	1696.3	980.3
3	ANDAMAN & NICOBAR ISLANDS	1902	0	159.8	12.2	0	446.1	537.1	228.9	753.7	666.2	197.2	359	160.5	3520.7	159.8	458.3	2185.9	716.7
4	ANDAMAN & NICOBAR ISLANDS	1903	12.7	144	0	1	235.1	479.9	728.4	326.7	339	181.2	284.4	225	2957.4	156.7	236.1	1874	690.6
5	ANDAMAN & NICOBAR ISLANDS	1904	9.4	14.7	0	202.4	304.5	495.1	502	160.1	820.4	222.2	308.7	40.1	3079.6	24.1	506.9	1977.6	571
6	ANDAMAN & NICOBAR ISLANDS	1905	1.3	0	3.3	26.9	279.5	628.7	368.7	330.5	297	260.7	25.4	344.7	2566.7	1.3	309.7	1624.9	630.8
7	ANDAMAN & NICOBAR ISLANDS	1906	36.6	0	0	0	556.1	733.3	247.7	320.5	164.3	267.8	128.9	79.2	2534.4	36.6	556.1	1465.8	475.9
8	ANDAMAN & NICOBAR ISLANDS	1907	110.7	0	113.3	21.6	616.3	305.2	443.9	377.6	200.4	264.4	648.9	245.6	3347.9	110.7	751.2	1327.1	1158.9
9	ANDAMAN & NICOBAR ISLANDS	1908	20.9	85.1	0	29	562	693.6	481.4	699.9	428.8	170.7	208.1	196.9	3576.4	106	591	2303.7	575.7
10	ANDAMAN & NICOBAR ISLANDS	1910	26.6	22.7	206.3	89.3	224.5	472.7	264.3	337.4	626.6	208.2	267.3	153.5	2899.4	49.3	520.1	1701	629
11	ANDAMAN & NICOBAR ISLANDS	1911	0	8.4	0	122.5	327.3	649	253	187.1	464.5	333.8	94.5	247.1	2687.2	8.4	449.8	1553.6	675.4
12	ANDAMAN & NICOBAR ISLANDS	1912	583.7	0.8	0	21.9	140.7	549.8	468.9	370.3	386.2	318.7	117.2	2.3	2960.5	584.5	162.6	1775.2	438.2
13	ANDAMAN & NICOBAR ISLANDS	1913	84.8	0.5	1.3	2.5	190.7	530	280.8	205.8	580.1	288.8	133	67.5	2365.8	85.3	194.5	1596.7	489.3
14	ANDAMAN & NICOBAR ISLANDS	1914	0	0	0	37.7	298.8	383.3	792.8	520.5	310.8	139.8	184.4	289.7	2957.8	0	336.5	2007.4	613.9
15	ANDAMAN & NICOBAR ISLANDS	1915	45	56.7	33.3	40.9	170.2	334.7	269	317.2	429.8	468.1	258.4	318	2741.3	101.7	244.4	1350.7	1044.5
16	ANDAMAN & NICOBAR ISLANDS	1916	0	0	0	0.5	487.4	450.1	317.3	425	561.2	369.7	192.6	133.7	2937.5	0	487.9	1753.6	696
17	ANDAMAN & NICOBAR ISLANDS	1917	8	3.6	112	4.5	295.9	301.1	394.8	437.4	471.8	238.1	108.3	236.9	2612.4	11.6	412.4	1605.1	583.3
18	ANDAMAN & NICOBAR ISLANDS	1918	77.4	6.9	11.4	10.7	729.3	710.8	200.9	455.4	303.3	227	366.9	175	3275	84.3	751.4	1670.4	768.9
19	ANDAMAN & NICOBAR ISLANDS	1919	10.2	18	0	35.5	283.9	542.5	246.5	259.8	170.7	186.2	340.4	258.4	2352.1	28.2	319.4	1219.5	785
20	ANDAMAN & NICOBAR ISLANDS	1920	122.3	7.4	3.1	13	237.4	546.9	294.4	467.4	505.4	397.5	262.9	85.5	2943.2	129.7	253.5	1814.1	745.9
21	ANDAMAN & NICOBAR ISLANDS	1921	13.2	3.1	0	37.5	351.2	282.7	487.1	330	581.2	360.7	118.2	41.5	2606.4	16.3	388.7	1681	520.4
22	ANDAMAN & NICOBAR ISLANDS	1922	245.3	34.3	15.6	323.1	289.7	506.1	425.8	307.4	511.7	162	541	192.2	3554.2	279.6	628.4	1751	895.2
23	ANDAMAN & NICOBAR ISLANDS	1923	79.5	0	NA	91.3	293.5	808.4	636.9	182.2	560.5	131.9	197.4	70.6	NA	79.5	NA	2188	399.9
24	ANDAMAN & NICOBAR ISLANDS	1924	28.7	0	14.8	89.7	191.2	261.2	493.3	290.9	251.2	331.1	378.6	NA	NA	28.7	295.7	1296.6	NA
25	ANDAMAN & NICOBAR ISLANDS	1925	36.6	0	8.6	50.4	282.2	663.8	241.8	278.2	201.9	249.5	271.5	196	2480.5	36.6	341.2	1385.7	717
26	ANDAMAN & NICOBAR ISLANDS	1926	122.1	0	0	0.5	198.4	370	195.3	523.7	719.3	443.8	148.4	560.7	3282.2	122.1	198.9	1808.3	1152.9
27	ANDAMAN & NICOBAR ISLANDS	1927	3	17.5	17.8	108.6	504.1	433.3	195.2	370.1	126.2	327.5	274.1	65.5	2442.9	20.5	630.5	1124.8	667.1
28	ANDAMAN & NICOBAR ISLANDS	1928	50.9	67.6	80.7	129.3	499.5	410.2	406.3	391.5	404.8	444.5	99.5	13.5	2998.3	118.5	709.5	1612.8	557.5
29	ANDAMAN & NICOBAR ISLANDS	1929	74.2	118.4	129.2	69.8	316.6	588.8	134	644.7	172.9	413	251.5	13.5	2926.6	192.6	515.6	1540.4	678
30	ANDAMAN & NICOBAR ISLANDS	1930	87.4	105.4	131.2	10.9	231.5	533.6	317.9	446.7	677.2	82.3	249.4	201.6	3075.1	192.8	373.6	1975.4	533.3

Missing Values methods:

Ignore the missing values :

- ❖ Missing data under 10% for an individual case or observation can generally be ignored
- ❖ **Drop the missing values:**

Dropping a variable

- ❖ The number of missing values in a feature is very high in a data, then that feature should be left out of the analysis. [6]

Imputation by Mean/Mode/Median

- ❖ Imputation is the process of substituting the missing data by some statistical methods. [2]
- ❖ If the missing values in a column or feature are numerical, the values can be imputed by the mean of the complete cases of the variable. [2]

Missing Values methods:

Regression Methods

- ❖ The variables with missing values are treated as dependent variables and variables with complete cases are taken as predictors or independent variables. [7]
- ❖ The independent variables are used to fit a linear equation for the observed values of the dependent variable. This equation is then used to predict values for the missing data points. [7]

K-Nearest Neighbour Imputation (KNN)

- ❖ This method uses k-nearest neighbour algorithms to estimate and replace missing data.
- ❖ The k-neighbours are chosen using some distance measure and their average is used as an imputation estimate. [8]
- ❖ This could be used for estimating both qualitative attributes (The most frequent value among the k nearest neighbours) and quantitative attributes (the mean of the k nearest neighbours). [8]

Applying Missing Values method:

None and NaN are used for indicating missing or null values.

To facilitate this convention, there are several useful methods. They are:

- ❖ **isnull():** Generate a boolean mask indicating missing values
- ❖ **notnull():** Opposite of isnull()
- ❖ **dropna():** Return a filtered version of the data
- ❖ **fillna():** Return a copy of the data with missing values filled or imputed

So first we will find total missing values using sum:

dataset.isnull().sum()

Then we apply missing values method using fillna method

Again we check the missing value sum to confirm that no missing values in our dataset

Missing values for the dataset :

- **Mean:** Mean (commonly known as average) is equal to the sum of all values in the column divided by the number of values present in the column. In Excel, you can use the AVERAGE() function to compute the mean.

  dataset = dataset.fillna(dataset.mean())

- **pad():** This method is similar to the dataframe.fillna(). method and it fills NA/NaN values using the ffill() method.It returns the object with missing values filled or none if replace=True.

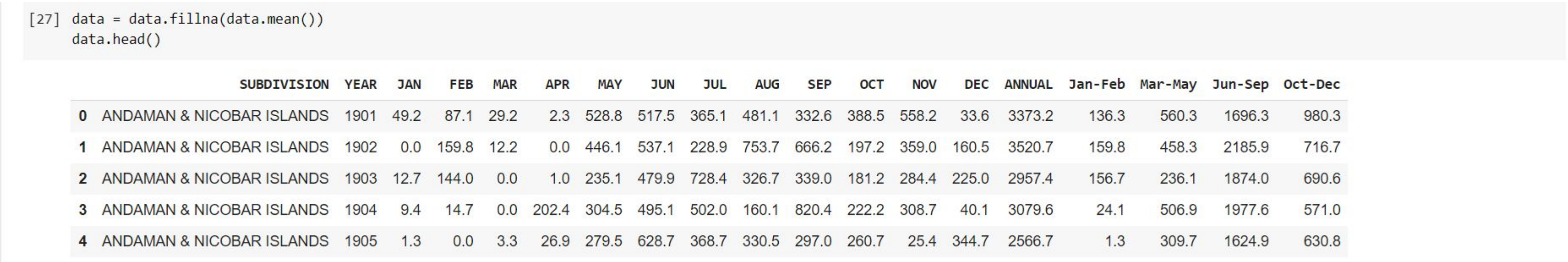
  dataset = dataset.fillna(method ='pad')

- **bfill():** is used to backward fill the missing values in the dataset. It will backward fill the NaN values that are present in the pandas dataframe.

  dataset = dataset.fillna(method ='bfill')

Missing values for the dataset:

Before applying mean method	After applying mean method
SUBDIVISION 0	SUBDIVISION 0
YEAR 0	YEAR 0
JAN 4	JAN 0
FEB 3	FEB 0
MAR 6	MAR 0
APR 4	APR 0
MAY 3	MAY 0
JUN 5	JUN 0
JUL 7	JUL 0
AUG 4	AUG 0
SEP 6	SEP 0
OCT 7	OCT 0
NOV 11	NOV 0
DEC 10	DEC 0
ANNUAL 26	ANNUAL 0
Jan-Feb 6	Jan-Feb 0
Mar-May 9	Mar-May 0
Jun-Sep 10	Jun-Sep 0
Oct-Dec 13	Oct-Dec 0
dtype: int64	dtype: int64

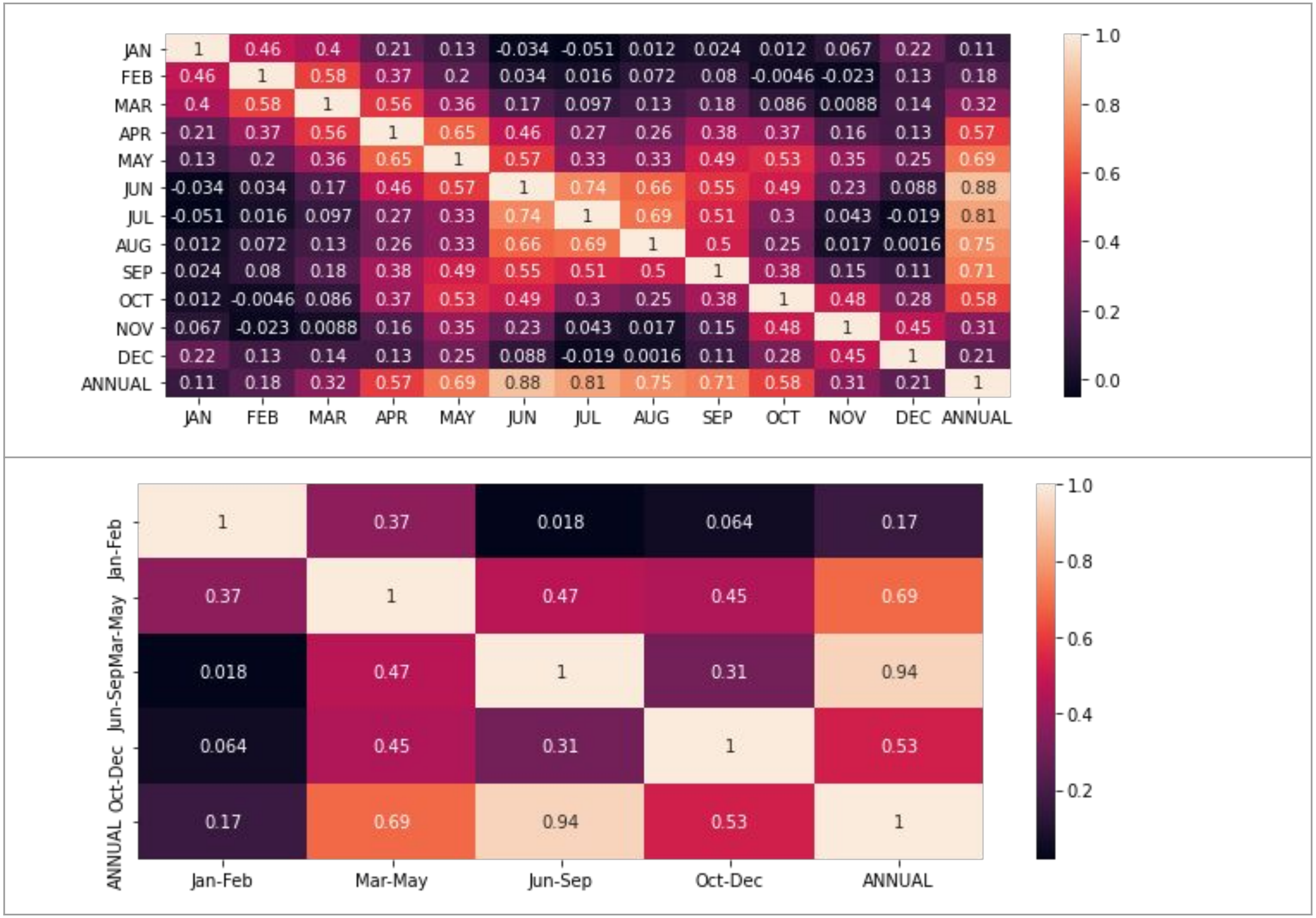


```
[27] data = data.fillna(data.mean())
     data.head()
```

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
0	ANDAMAN & NICOBAR ISLANDS	1901	49.2	87.1	29.2	2.3	528.8	517.5	365.1	481.1	332.6	388.5	558.2	33.6	3373.2	136.3	560.3	1696.3	980.3
1	ANDAMAN & NICOBAR ISLANDS	1902	0.0	159.8	12.2	0.0	446.1	537.1	228.9	753.7	666.2	197.2	359.0	160.5	3520.7	159.8	458.3	2185.9	716.7
2	ANDAMAN & NICOBAR ISLANDS	1903	12.7	144.0	0.0	1.0	235.1	479.9	728.4	326.7	339.0	181.2	284.4	225.0	2957.4	156.7	236.1	1874.0	690.6
3	ANDAMAN & NICOBAR ISLANDS	1904	9.4	14.7	0.0	202.4	304.5	495.1	502.0	160.1	820.4	222.2	308.7	40.1	3079.6	24.1	506.9	1977.6	571.0
4	ANDAMAN & NICOBAR ISLANDS	1905	1.3	0.0	3.3	26.9	279.5	628.7	368.7	330.5	297.0	260.7	25.4	344.7	2566.7	1.3	309.7	1624.9	630.8

Visualization :

1. HeatMap:



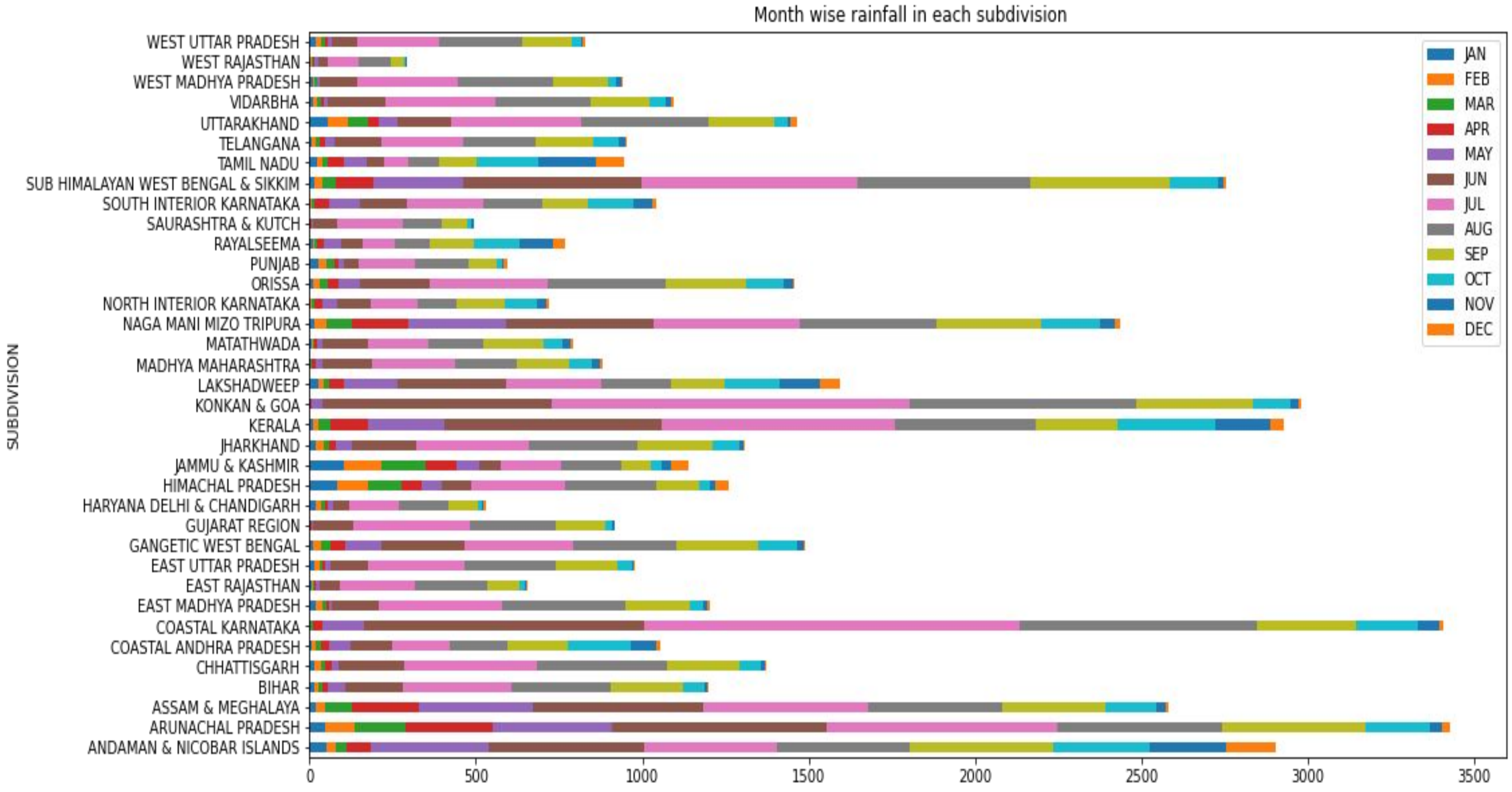
	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL
JAN	1	0.46	0.4	0.21	0.13	-0.034	-0.051	0.012	0.024	0.012	0.067	0.22	0.11
FEB	0.46	1	0.58	0.37	0.2	0.034	0.016	0.072	0.08	-0.0046	-0.023	0.13	0.18
MAR	0.4	0.58	1	0.56	0.36	0.17	0.097	0.13	0.18	0.086	0.0088	0.14	0.32
APR	0.21	0.37	0.56	1	0.65	0.46	0.27	0.26	0.38	0.37	0.16	0.13	0.57
MAY	0.13	0.2	0.36	0.65	1	0.57	0.33	0.33	0.49	0.53	0.35	0.25	0.69
JUN	-0.034	0.034	0.17	0.46	0.57	1	0.74	0.66	0.55	0.49	0.23	0.088	0.88
JUL	-0.051	0.016	0.097	0.27	0.33	0.74	1	0.69	0.51	0.3	0.043	-0.019	0.81
AUG	0.012	0.072	0.13	0.26	0.33	0.66	0.69	1	0.5	0.25	0.017	0.0016	0.75
SEP	0.024	0.08	0.18	0.38	0.49	0.55	0.51	0.5	1	0.38	0.15	0.11	0.71
OCT	0.012	-0.0046	0.086	0.37	0.53	0.49	0.3	0.25	0.38	1	0.48	0.28	0.58
NOV	0.067	-0.023	0.0088	0.16	0.35	0.23	0.043	0.017	0.15	0.48	1	0.45	0.31
DEC	0.22	0.13	0.14	0.13	0.25	0.088	-0.019	0.0016	0.11	0.28	0.45	1	0.21
ANNUAL	0.11	0.18	0.32	0.57	0.69	0.88	0.81	0.75	0.71	0.58	0.31	0.21	1

	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec	ANNUAL
Jan-Feb	1	0.37	0.018	0.064	0.17
Mar-May	0.37	1	0.47	0.45	0.69
Jun-Sep	0.018	0.47	1	0.31	0.94
Oct-Dec	0.064	0.45	0.31	1	0.53
ANNUAL	0.17	0.69	0.94	0.53	1

Correlation matrix

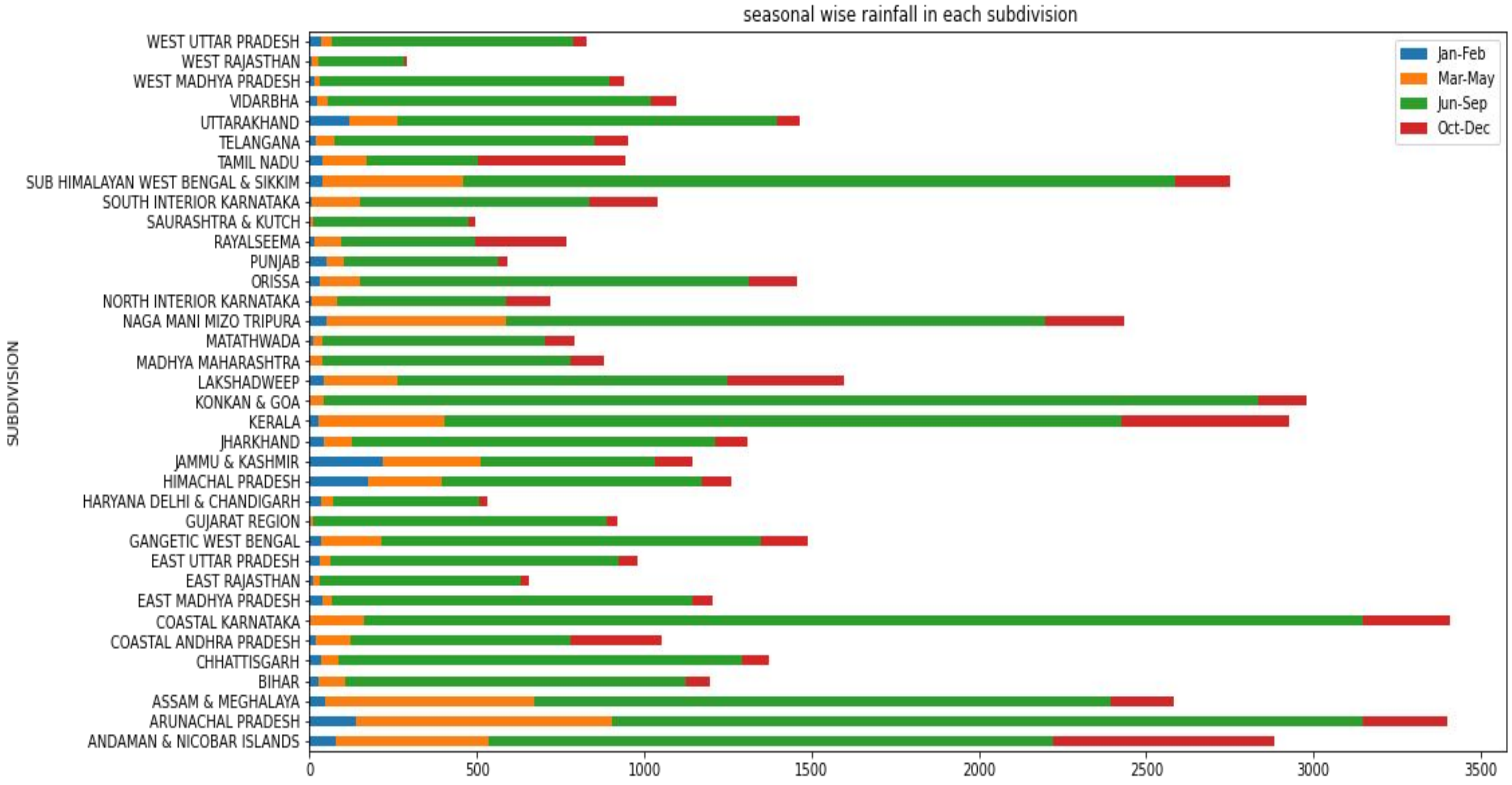
Visualization :

2. BarPlot:



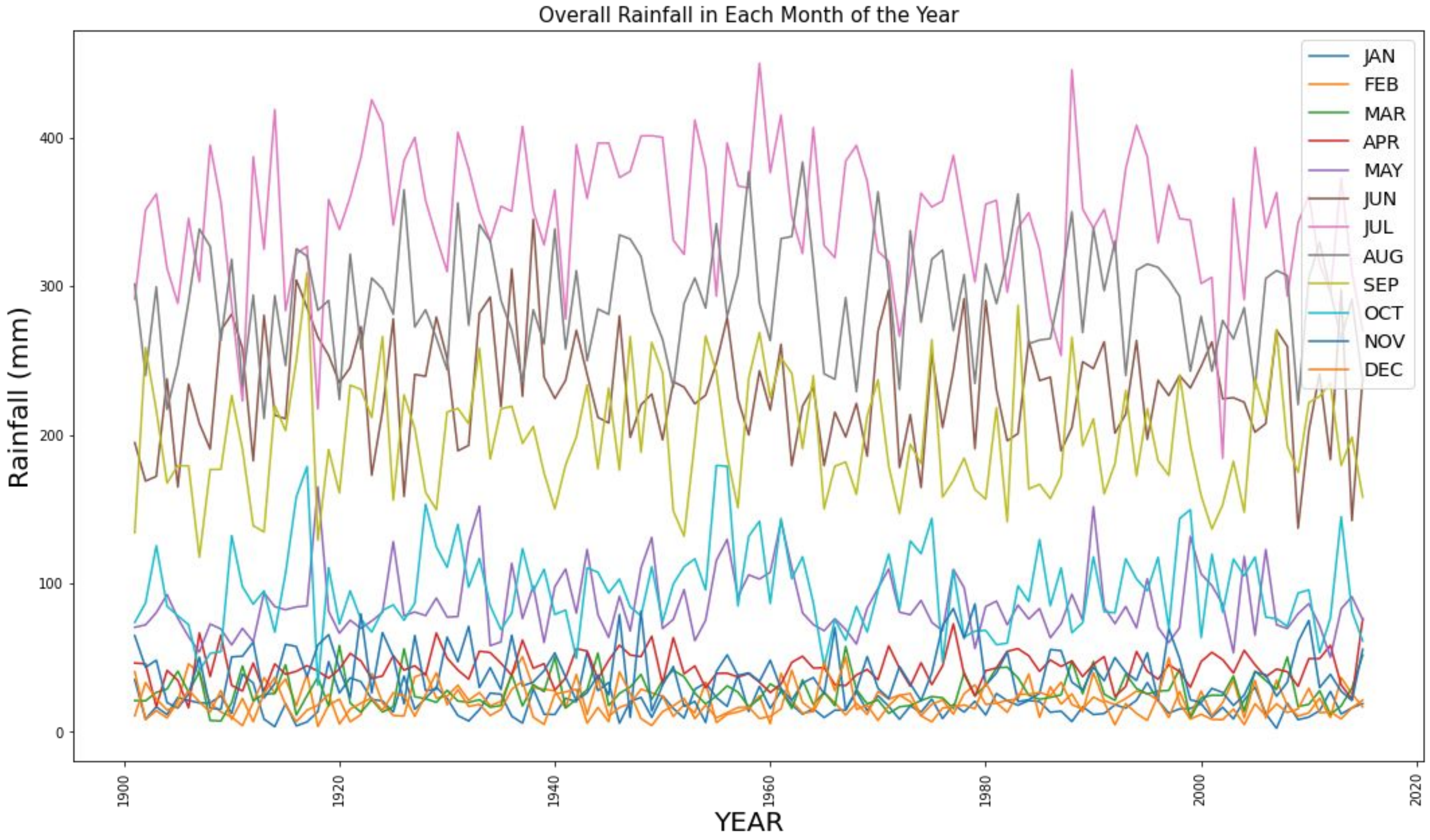
Visualization :

2. BarPlot:



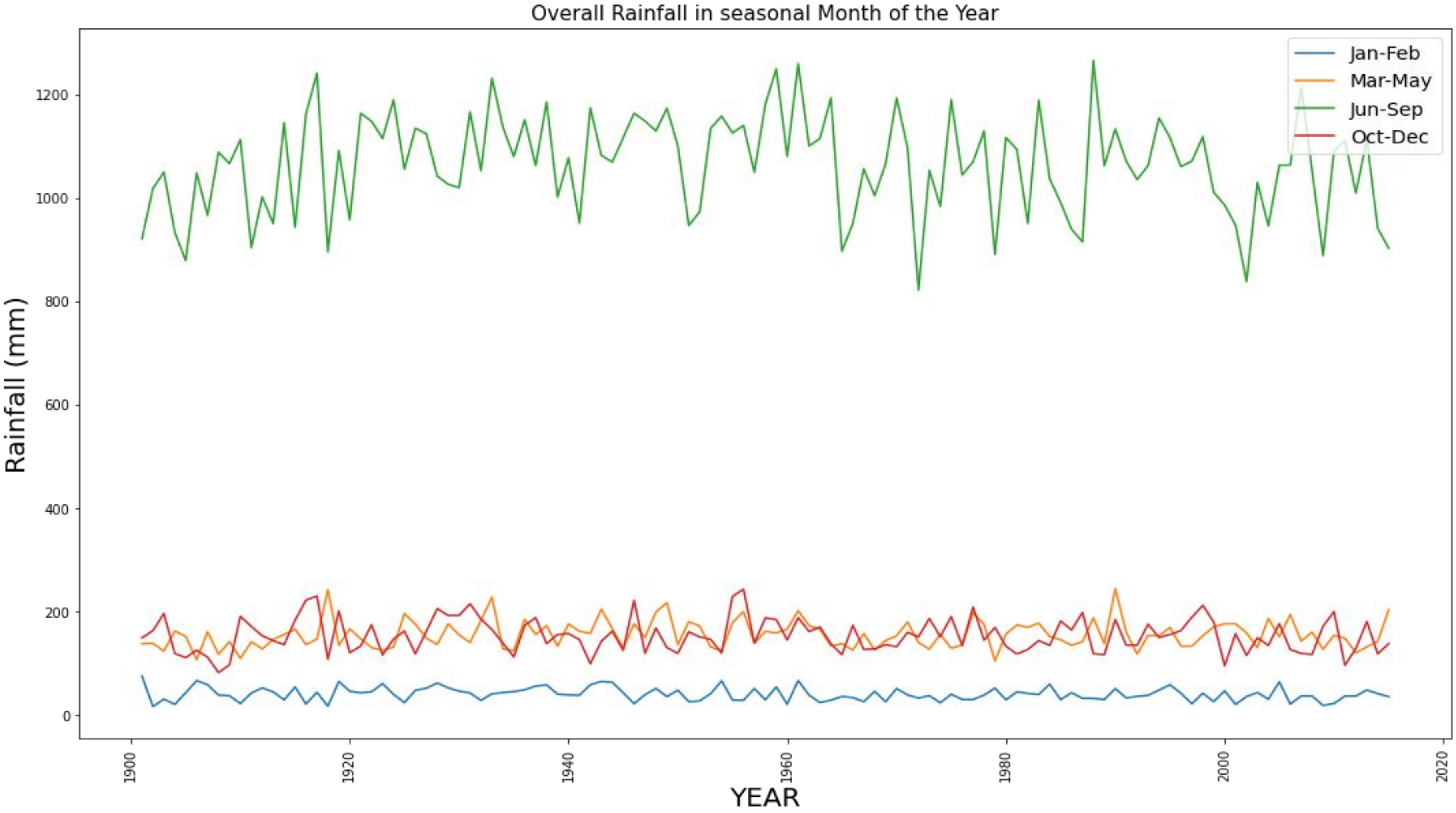
Visualization :

3. Line Plot:



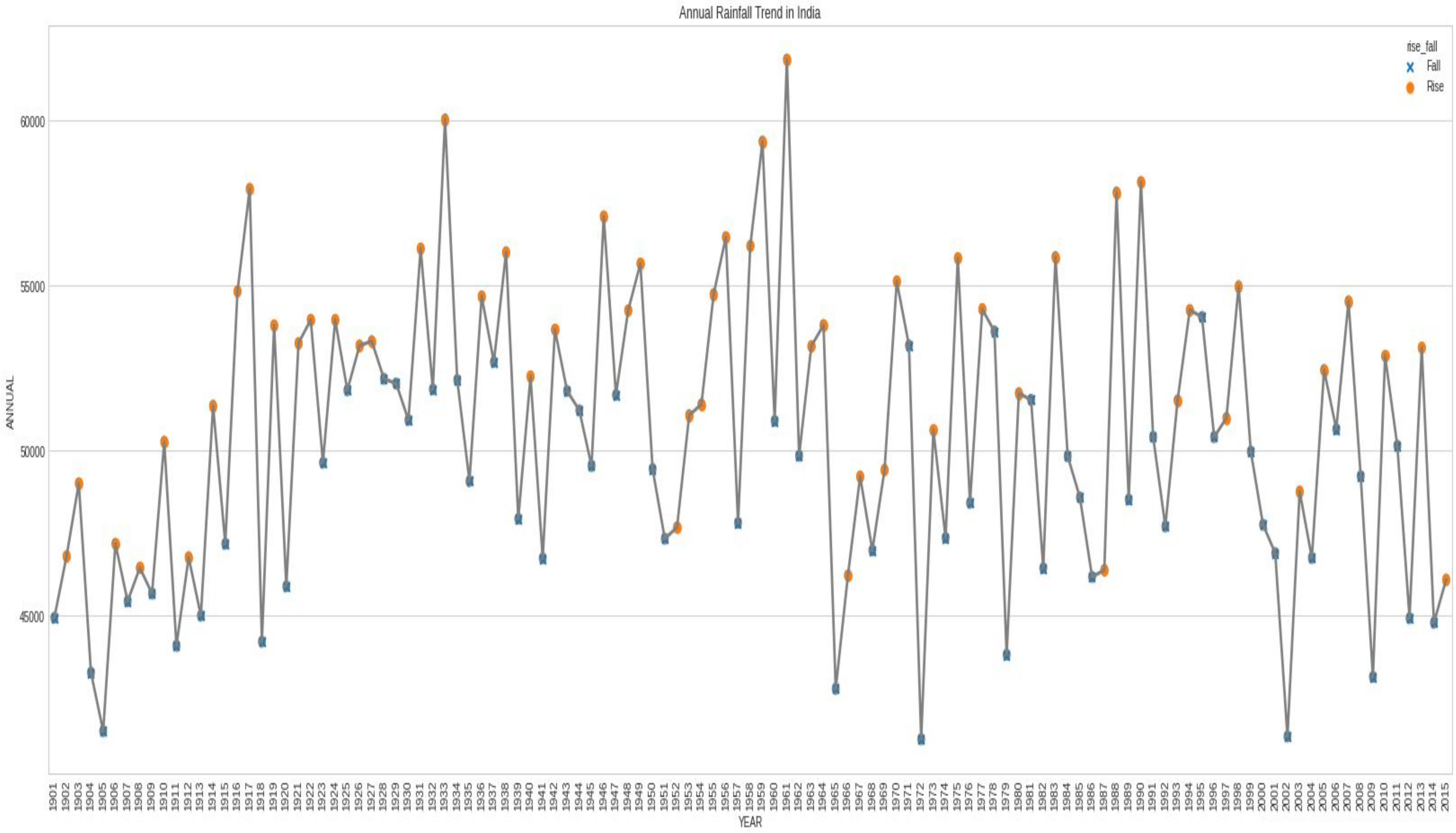
Visualization :

3. Line Plot:



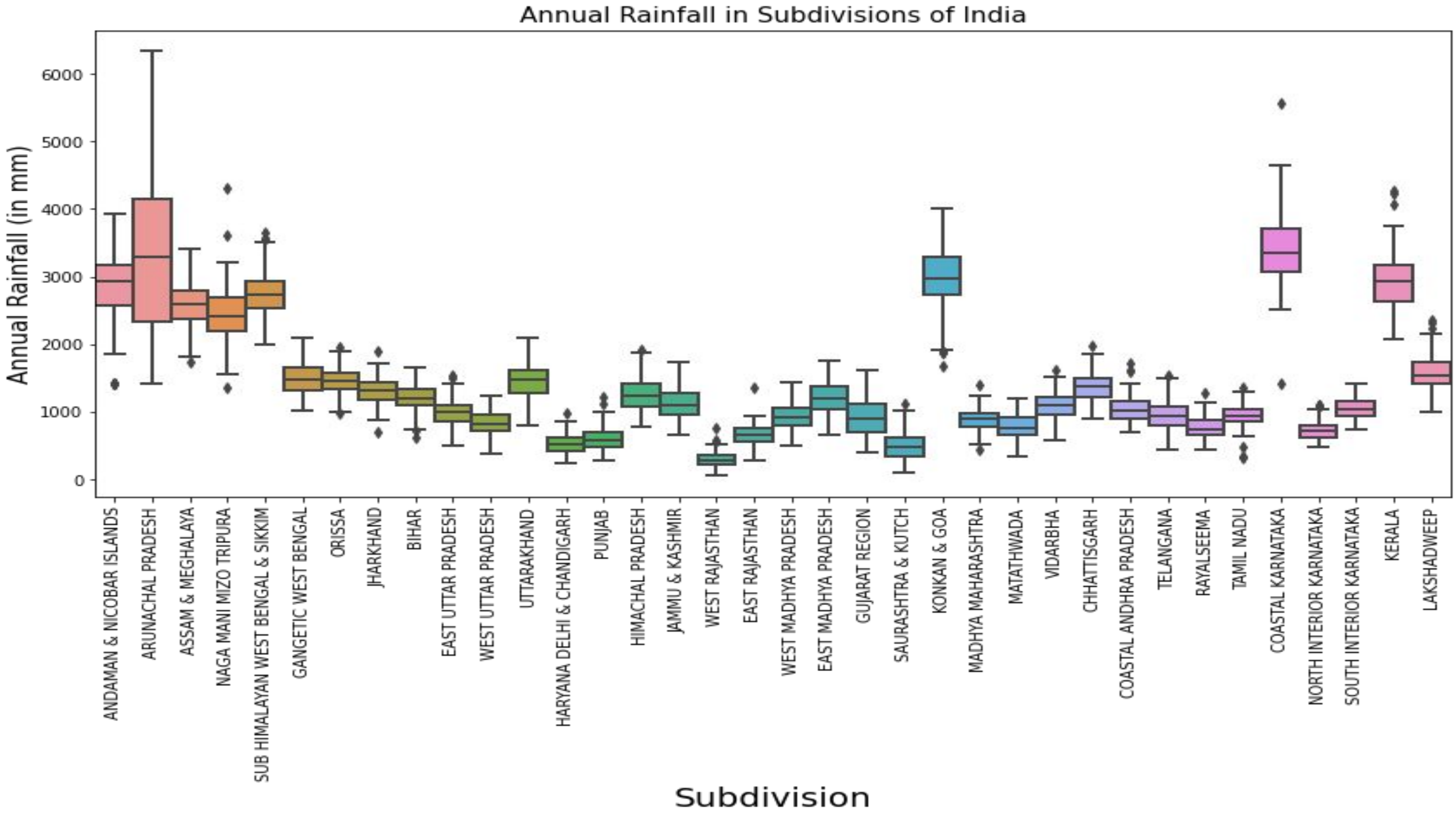
Visualization :

3. Line Plot:



Visualization :

4. Box Plot:



Applying Machine Learning Techniques:

Process:

Step1: Import the rainfall data set.csv file.

Step2: Fill the missing values with mean value of the data.

Step 3: The data is divided into training set (80%) and testing set (20%).

Step 4: Apply machine learning or deep learning algorithms and calculate the errors like Mean Absolute Error, Mean Squared Error.

Step 5: Choose the best Model among all models.

The models we used are: Elastic net, Lasso, Ridge, SVR, Random forest regression,Neural network model.

For prediction we formatted data in the way, given the rainfall in the last three months we try to predict the rainfall in the next consecutive month.

Applying Machine Learning Techniques:

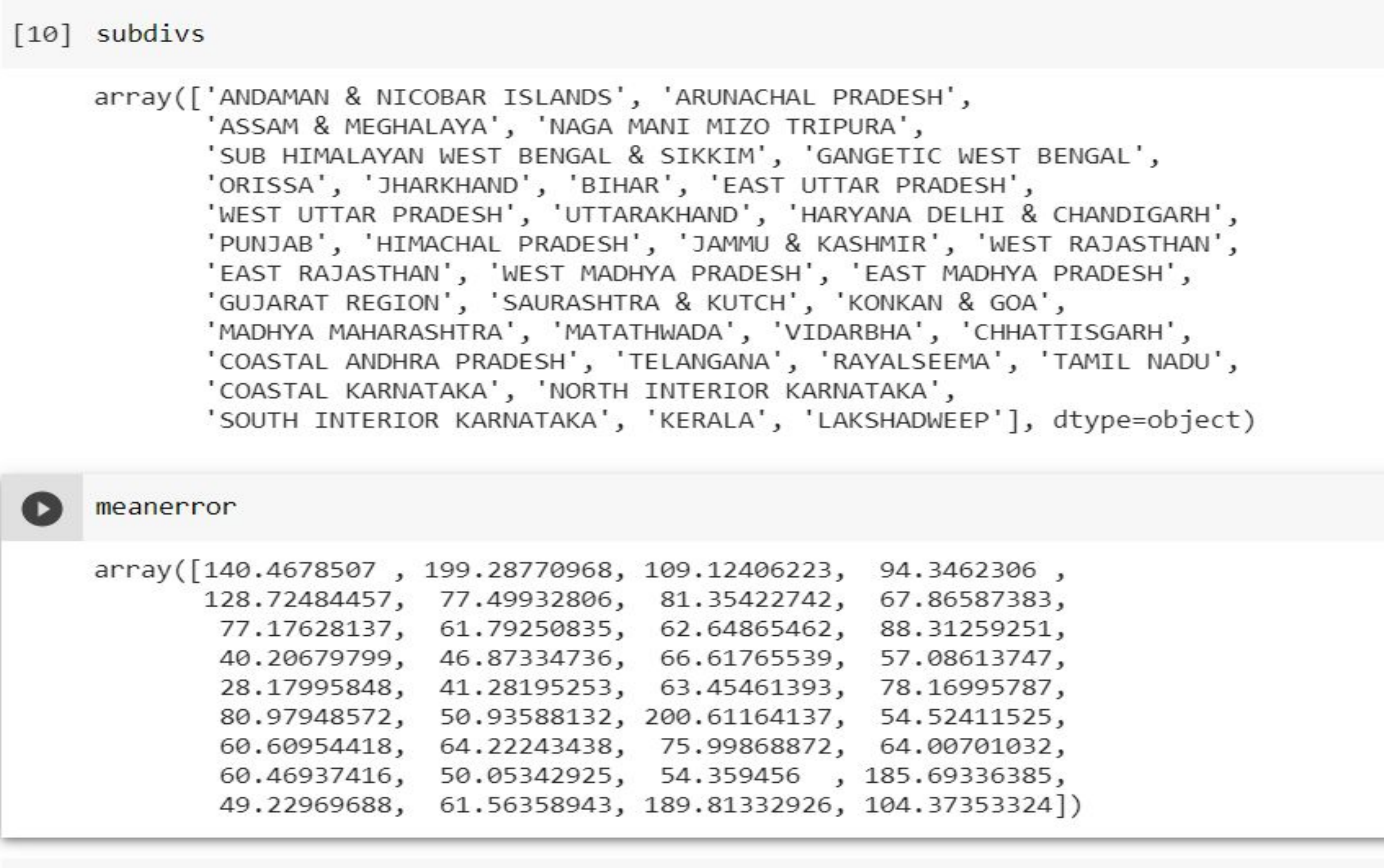
Model	Mean Absolute Error	Mean Squared Error
Elastic Net Regression	**93.3734**	**142.5711**
Ridge Regression	**93.9519**	**141.8863**
Lasso Regression	**93.9520**	**141.8863**
SVR	**125.1810**	**202.5503**
Random Forest Regressor	**91.7085**	**145.7345**
Neural Network	**85.1830**	**135.8926**

We Observe that all the linear models got almost same MAE and MSE.

Neural Network got less errors among all other methods , second better algorithm was Random Forest Regression.

So we want to proceed using neural network model for upcoming work.

Applying Machine Learning Techniques:



```
[10]  subdivs

      array(['ANDAMAN & NICOBAR ISLANDS', 'ARUNACHAL PRADESH',
             'ASSAM & MEGHALAYA', 'NAGA MANI MIZO TRIPURA',
             'SUB HIMALAYAN WEST BENGAL & SIKKIM', 'GANGETIC WEST BENGAL',
             'ORISSA', 'JHARKHAND', 'BIHAR', 'EAST UTTAR PRADESH',
             'WEST UTTAR PRADESH', 'UTTARAKHAND', 'HARYANA DELHI & CHANDIGARH',
             'PUNJAB', 'HIMACHAL PRADESH', 'JAMMU & KASHMIR', 'WEST RAJASTHAN',
             'EAST RAJASTHAN', 'WEST MADHYA PRADESH', 'EAST MADHYA PRADESH',
             'GUJARAT REGION', 'SAURASHTRA & KUTCH', 'KONKAN & GOA',
             'MADHYA MAHARASHTRA', 'MATATHWADA', 'VIDARBHA', 'CHHATTISGARH',
             'COASTAL ANDHRA PRADESH', 'TELANGANA', 'RAYALSEEMA', 'TAMIL NADU',
             'COASTAL KARNATAKA', 'NORTH INTERIOR KARNATAKA',
             'SOUTH INTERIOR KARNATAKA', 'KERALA', 'LAKSHADWEEP'], dtype=object)
```

```
      meanerror

      array([140.4678507 , 199.28770968, 109.12406223,  94.3462306 ,
             128.72484457,  77.49932806,  81.35422742,  67.86587383,
              77.17628137,  61.79250835,  62.64865462,  88.31259251,
              40.20679799,  46.87334736,  66.61765539,  57.08613747,
              28.17995848,  41.28195253,  63.45461393,  78.16995787,
              80.97948572,  50.93588132, 200.61164137,  54.52411525,
              60.60954418,  64.22243438,  75.99868872,  64.00701032,
              60.46937416,  50.05342925,  54.359456  , 185.69336385,
              49.22969688,  61.56358943, 189.81332926, 104.37353324])
```

Work done upto now/Future plan:

Work done upto now:

- Importing data
- Applying missing value method
- Visualization (Exploratory Data Analysis)
- Predicted the best model using MAE, MSE.

Future plan:

Based on the model we try to predict future values, by changing the data every time like appending the result at the end, as we are taking the data as an array.

Based on the future values we want to do an application

References:

1. Praveen, Bushra, et al. "Analyzing trend and forecasting of rainfall changes in India using non-parametrical and machine learning approaches." *Scientific reports* 10.1 (2020): 1-21.

2. Oswal, Nikhil. "Predicting rainfall using machine learning techniques." *arXiv preprint arXiv:1910.13827* (2019).

3. Hernández, Emilcy, et al. "Rainfall prediction: A deep learning approach." *International Conference on Hybrid Artificial Intelligence Systems*. Springer, Cham, 2016.

4. Singh, Pritpal, and Bhogeswar Borah. "Indian summer monsoon rainfall prediction using artificial neural network." *Stochastic environmental research and risk assessment* 27.7 (2013): 1585-1599.

5. Abhishek, Kumar, et al. "A rainfall prediction model using artificial neural network." *2012 IEEE Control and System Graduate Research Colloquium*. IEEE, 2012.

References:

6. Kang, Hyun. "The prevention and handling of the missing data." *Korean journal of anesthesiology* 64.5 (2013): 402.

7. Mostafa, Samih M. "Imputing missing values using cumulative linear regression." *CAAI Transactions on Intelligence Technology* 4.3 (2019): 182-200.

8. De Silva, Hiroshi, and A. Shehan Perera. "Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data." *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2016.

Thank you!

ANY QUESTIONS?