RAINFALL IN INDIA
IS603

FNU Naveen Kumar, Sudhanshu Jha, Rithvik Pabba, Srinija Dannamanane, Swathi Kotla
IS Department, University Of Maryland, Baltimore County

ABSTRACT

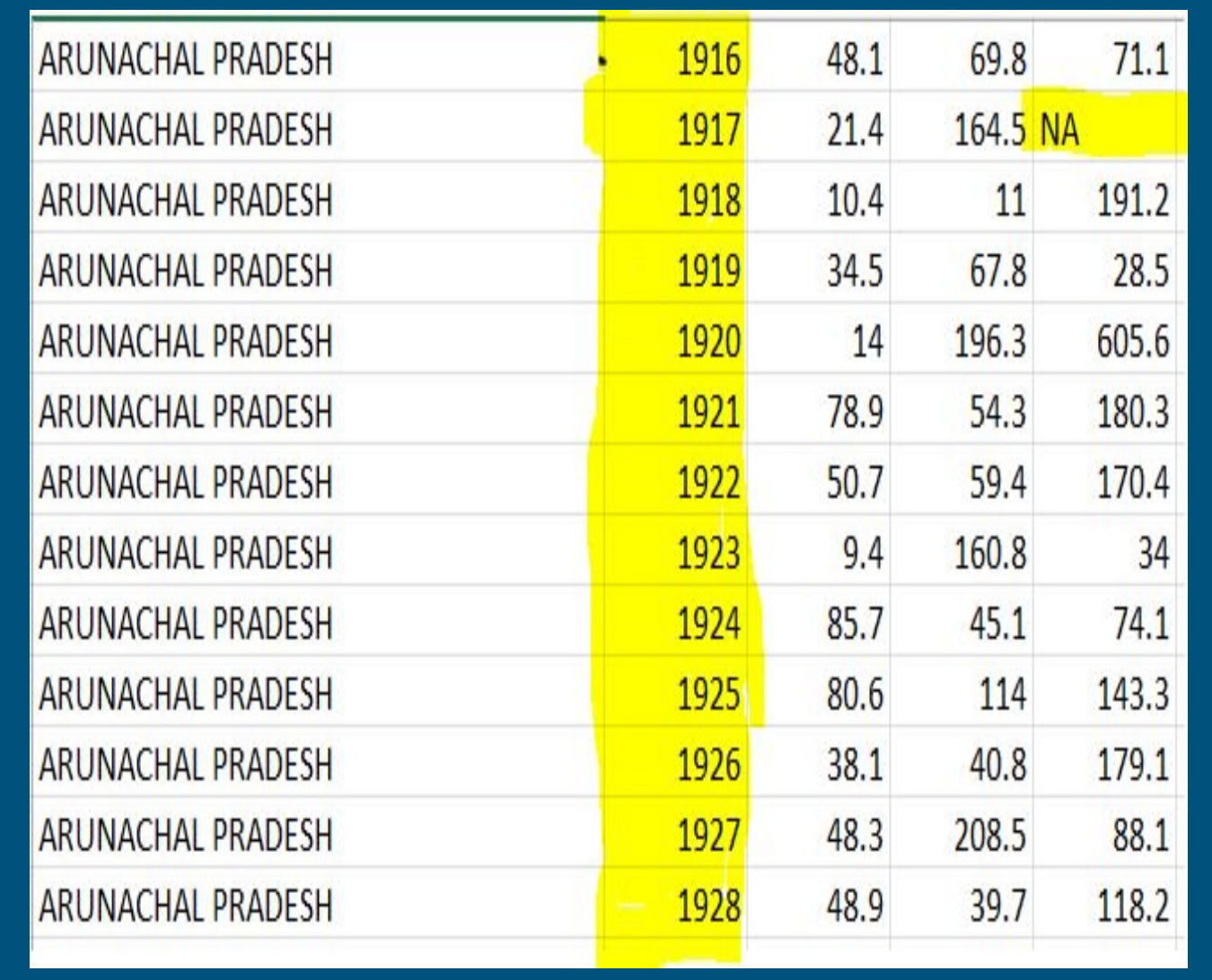
By reviewing enormous amounts of data, Data Science has grown increasingly popular in identifying hidden patterns, correlations, and other insights. As India is a country where agriculture is the only resource of income for most of population, in which rainfall plays a major role. By predicting the amount of rainfall, we can help farmers to know which kind of crops can be grown time of the year. “The Rainfall in India” dataset is used to analyze and visualize the rainfall in past 45 years (1970-2015).

INTRODUCTION

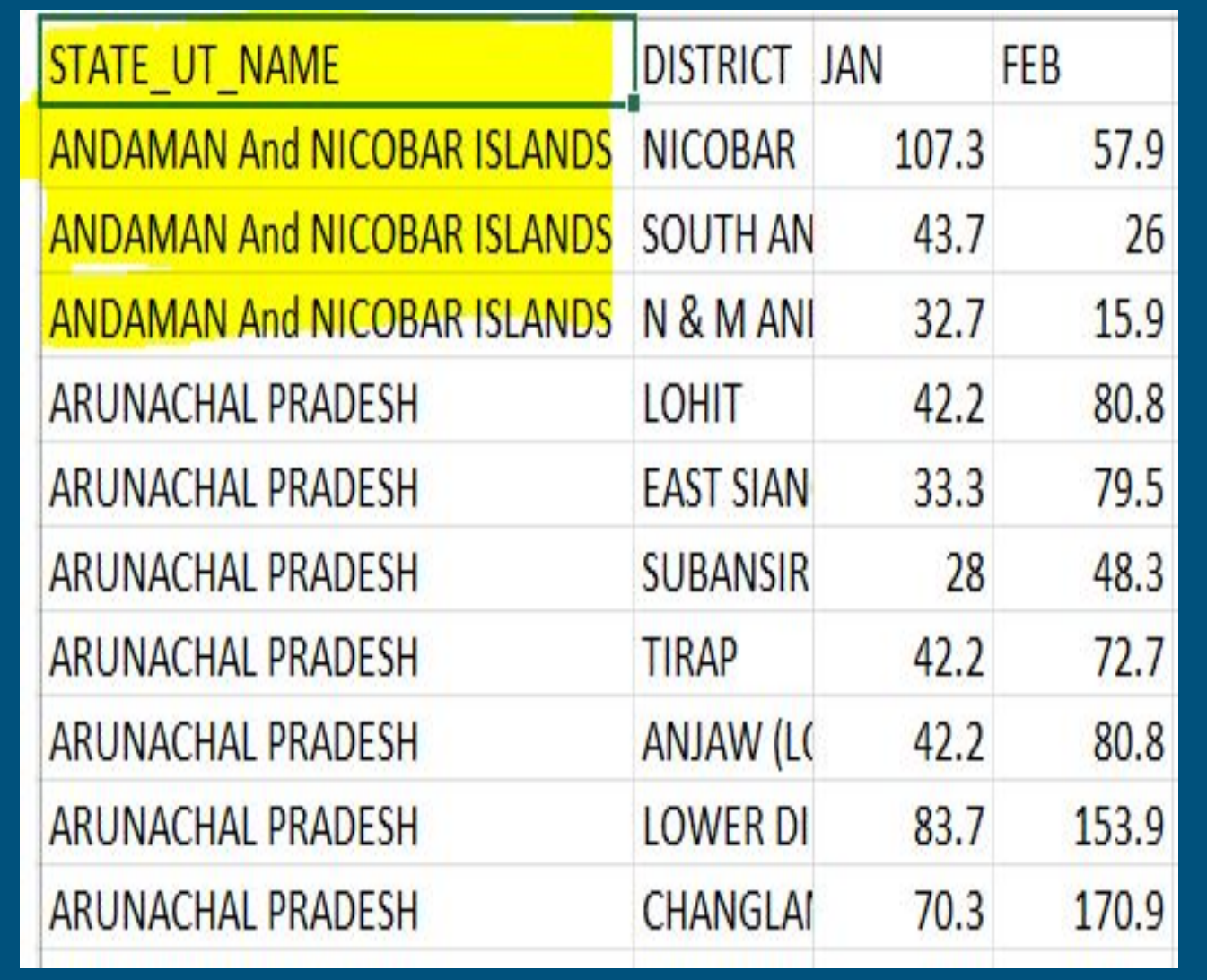
To perform the data analysis on the data set for rainfall in India prediction,We have taken two dataset one that distributes data of Indian rainfall with respect to states and districts. At first from each state some districts data were collected, and data cleansing is performed on the data set. Data visualization is used to show the rainfall prediction in different region based on district. Data-mining techniques like Classification, Clustering, Regression are used and tools like Weka for analysis and python for visualization is used.

DATA CLEANSING

In order to employ data mining methods, data cleansing requires transforming raw data into well-formed data sets. Raw data, is termed as incomplete and structured incorectly. Previuously we collected data for 115 years i.e., 1901-2015, and as there was some missing data in the data set such as inclusion of old region names we analysed and cleansed the data from 1971-2015.

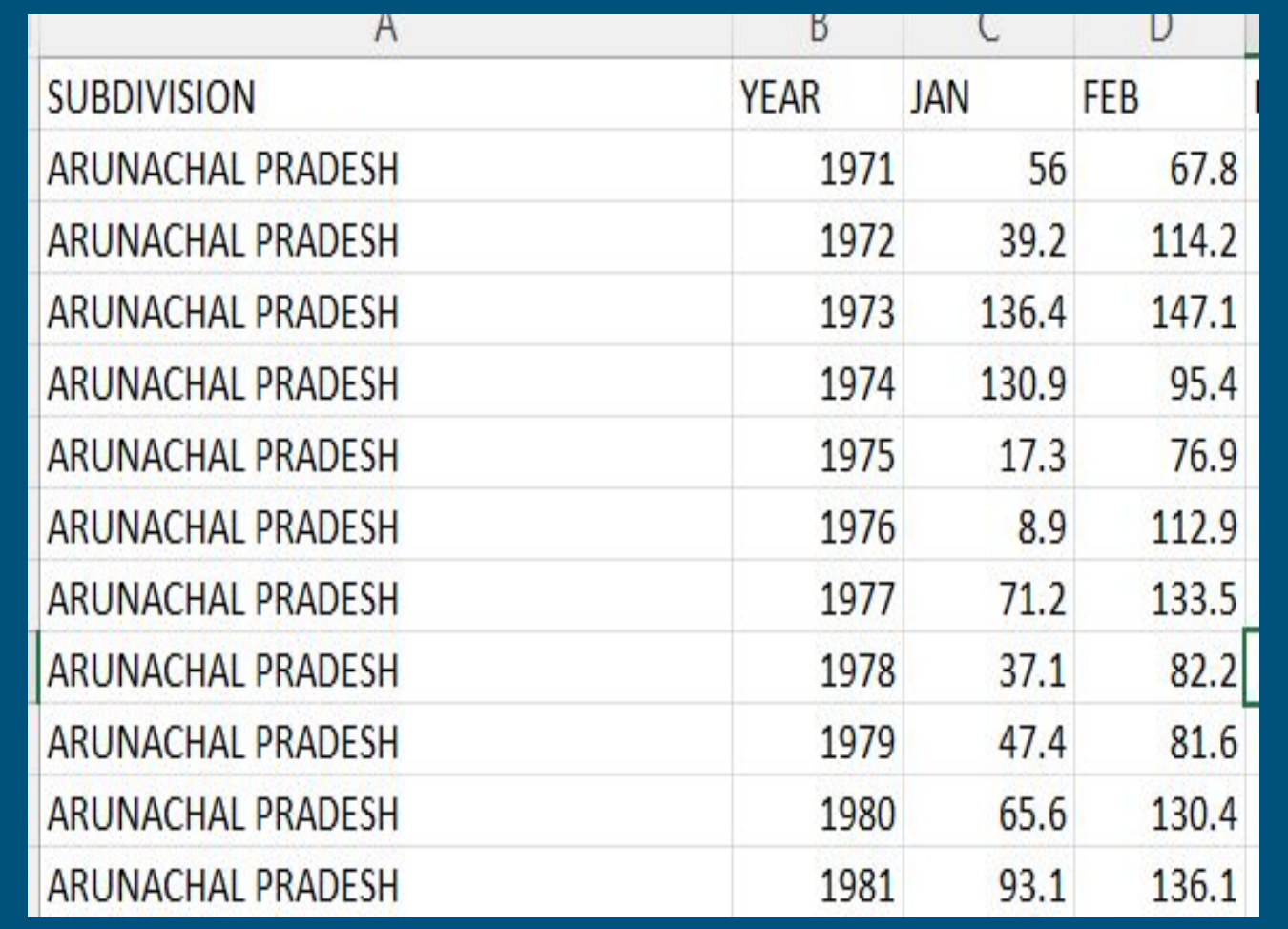


ARUNACHAL PRADESH	1916	48.1	69.8	71.1
ARUNACHAL PRADESH	1917	21.4	164.5	NA
ARUNACHAL PRADESH	1918	10.4	11	191.2
ARUNACHAL PRADESH	1919	34.5	67.8	28.5
ARUNACHAL PRADESH	1920	14	196.3	605.6
ARUNACHAL PRADESH	1921	78.9	54.3	180.3
ARUNACHAL PRADESH	1922	50.7	59.4	170.4
ARUNACHAL PRADESH	1923	9.4	160.8	34
ARUNACHAL PRADESH	1924	85.7	45.1	74.1
ARUNACHAL PRADESH	1925	80.6	114	143.3
ARUNACHAL PRADESH	1926	38.1	40.8	179.1
ARUNACHAL PRADESH	1927	48.3	208.5	88.1
ARUNACHAL PRADESH	1928	48.9	39.7	118.2

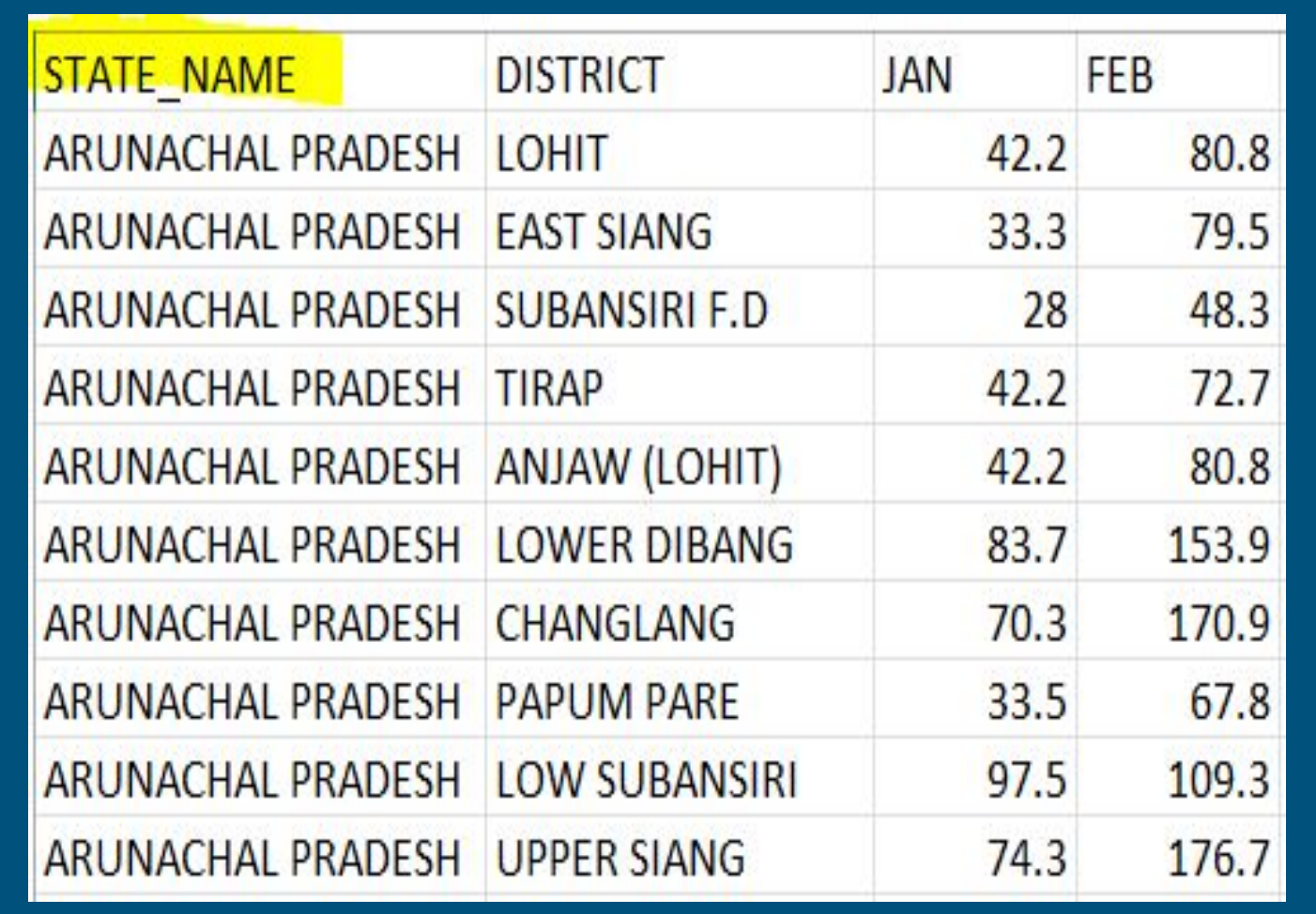


STATE_UT_NAME	DISTRICT	JAN	FEB
ANDAMAN And NICOBAR ISLANDS	NICOBAR	107.3	57.9
ANDAMAN And NICOBAR ISLANDS	SOUTH AN	43.7	26
ANDAMAN And NICOBAR ISLANDS	N & M ANI	32.7	15.9
ARUNACHAL PRADESH	LOHIT	42.2	80.8
ARUNACHAL PRADESH	EAST SIAN	33.3	79.5
ARUNACHAL PRADESH	SUBANSIR	28	48.3
ARUNACHAL PRADESH	TIRAP	42.2	72.7
ARUNACHAL PRADESH	ANJAW (L(	42.2	80.8
ARUNACHAL PRADESH	LOWER DI	83.7	153.9
ARUNACHAL PRADESH	CHANGLAI	70.3	170.9

BEFORE CLEANSING



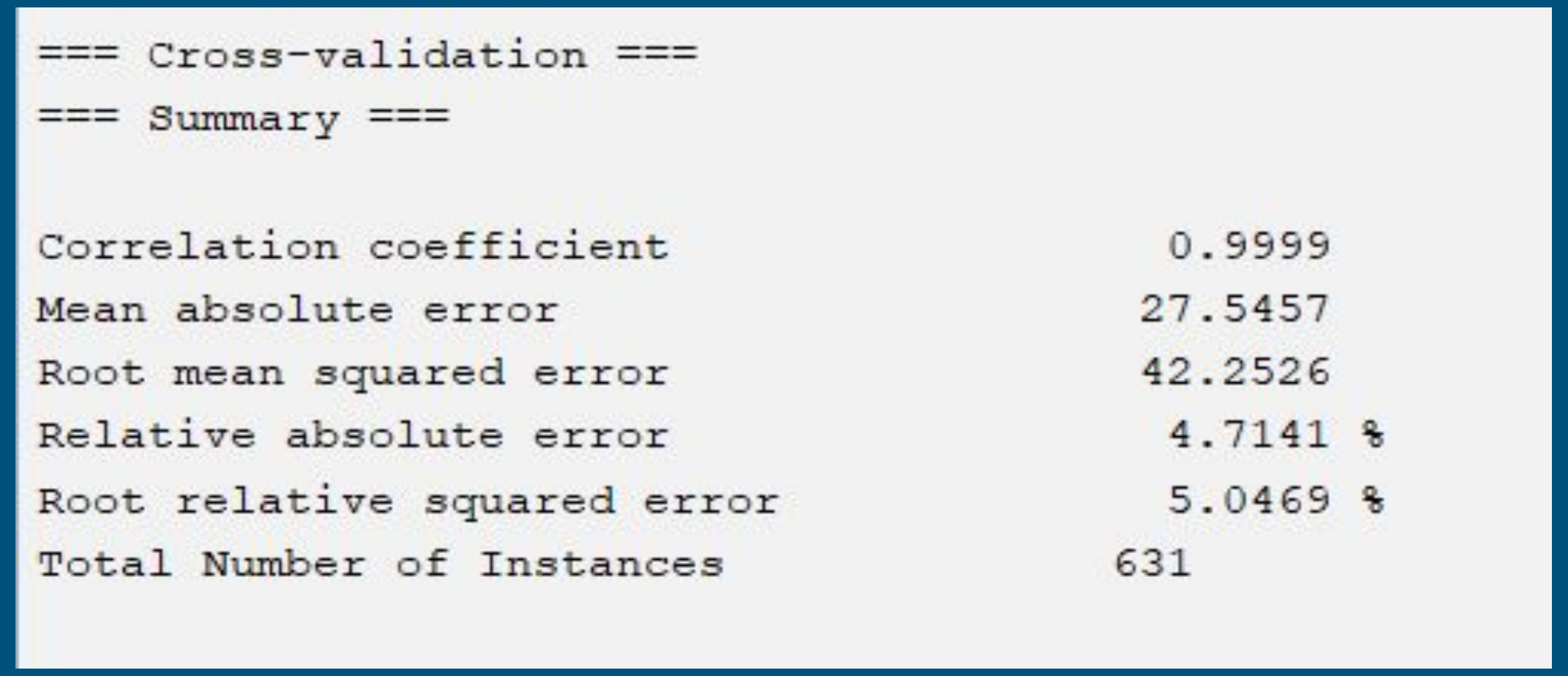
A	B	C	D
SUBDIVISION	YEAR	JAN	FEB
ARUNACHAL PRADESH	1971	56	67.8
ARUNACHAL PRADESH	1972	39.2	114.2
ARUNACHAL PRADESH	1973	136.4	147.1
ARUNACHAL PRADESH	1974	130.9	95.4
ARUNACHAL PRADESH	1975	17.3	76.9
ARUNACHAL PRADESH	1976	8.9	112.9
ARUNACHAL PRADESH	1977	71.2	133.5
ARUNACHAL PRADESH	1978	37.1	82.2
ARUNACHAL PRADESH	1979	47.4	81.6
ARUNACHAL PRADESH	1980	65.6	130.4
ARUNACHAL PRADESH	1981	93.1	136.1



STATE_NAME	DISTRICT	JAN	FEB
ARUNACHAL PRADESH	LOHIT	42.2	80.8
ARUNACHAL PRADESH	EAST SIANG	33.3	79.5
ARUNACHAL PRADESH	SUBANSIRI F.D	28	48.3
ARUNACHAL PRADESH	TIRAP	42.2	72.7
ARUNACHAL PRADESH	ANJAW (LOHIT)	42.2	80.8
ARUNACHAL PRADESH	LOWER DIBANG	83.7	153.9
ARUNACHAL PRADESH	CHANGLANG	70.3	170.9
ARUNACHAL PRADESH	PAPUM PARE	33.5	67.8
ARUNACHAL PRADESH	LOW SUBANSIRI	97.5	109.3
ARUNACHAL PRADESH	UPPER SIANG	74.3	176.7

AFTER CLEANSING

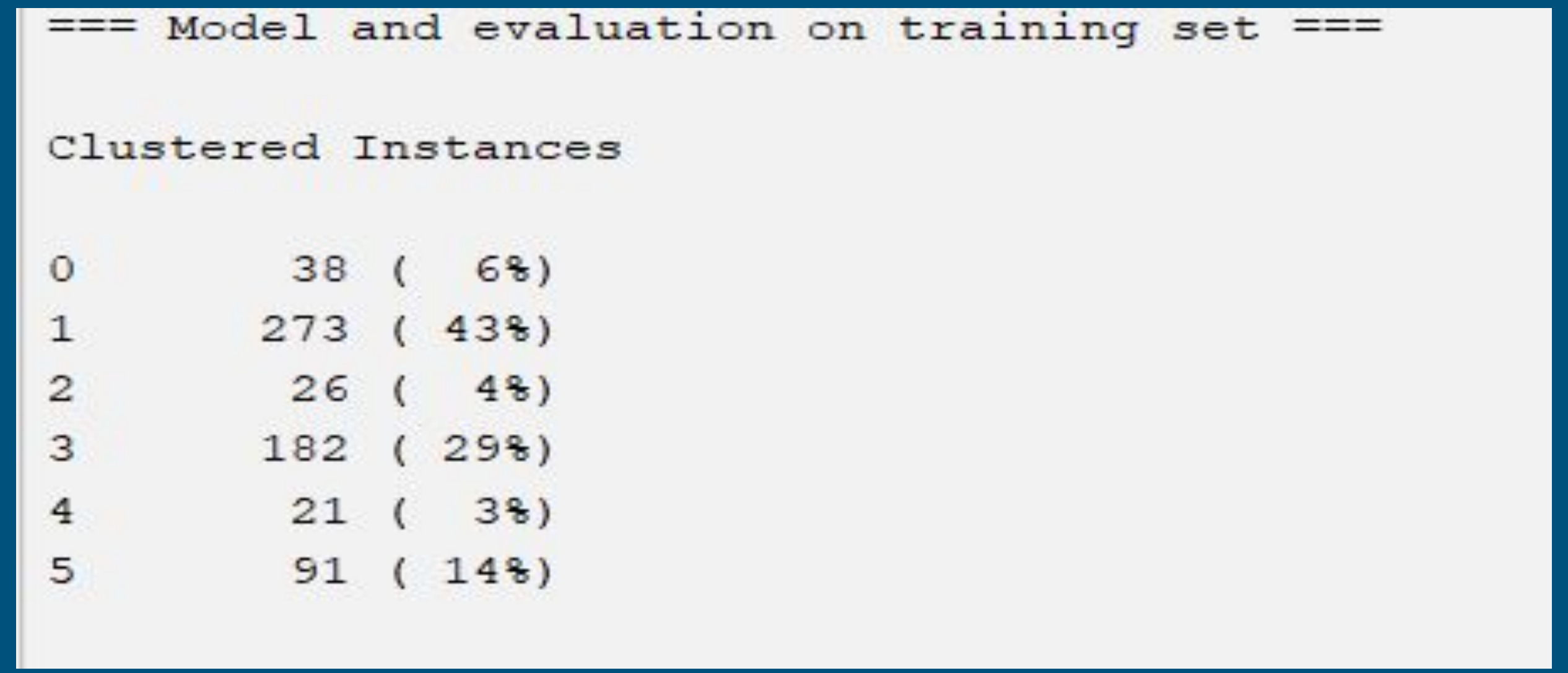
METHODLOGY



```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9999
Mean absolute error                     27.5457
Root mean squared error                 42.2526
Relative absolute error                  4.7141 %
Root relative squared error              5.0469 %
Total Number of Instances               631
```

REGRESSION CORELATION COEFFICIENT 0.99



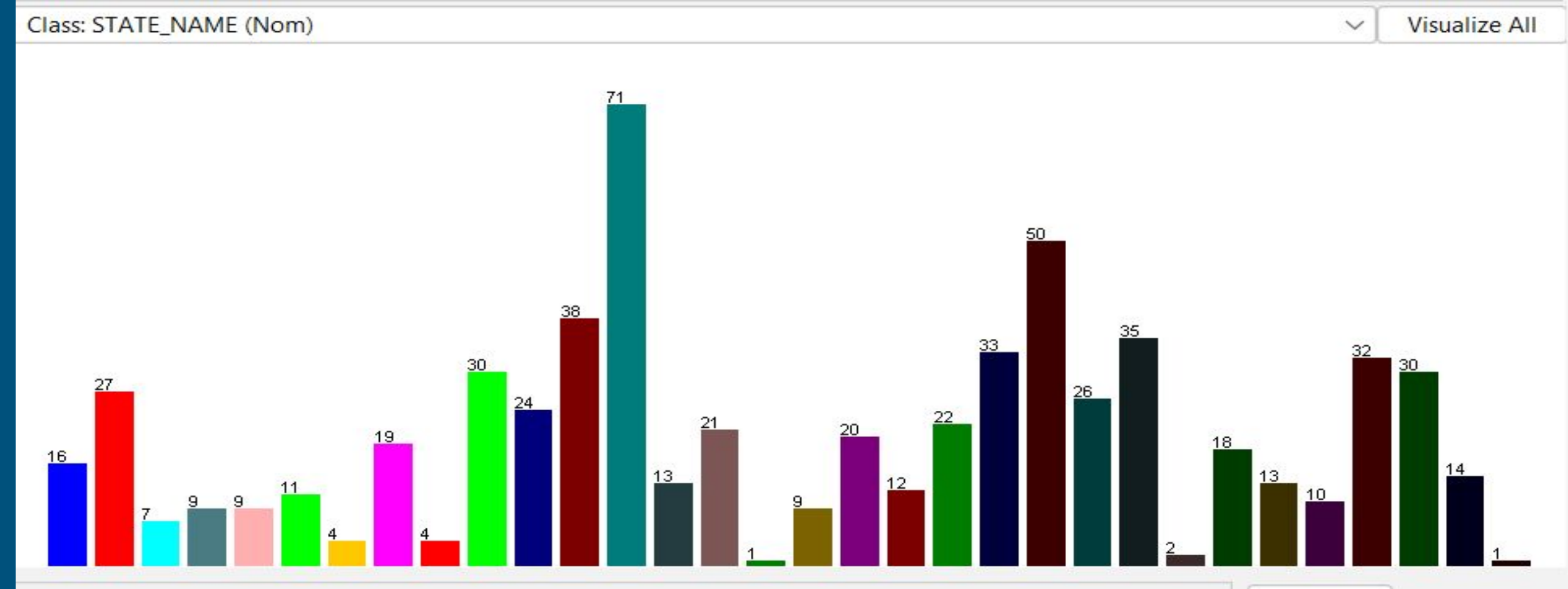
```
=== Model and evaluation on training set ===

Clustered Instances

0       38 (  6%)
1      273 ( 43%)
2       26 (  4%)
3      182 ( 29%)
4       21 (  3%)
5       91 ( 14%)
```

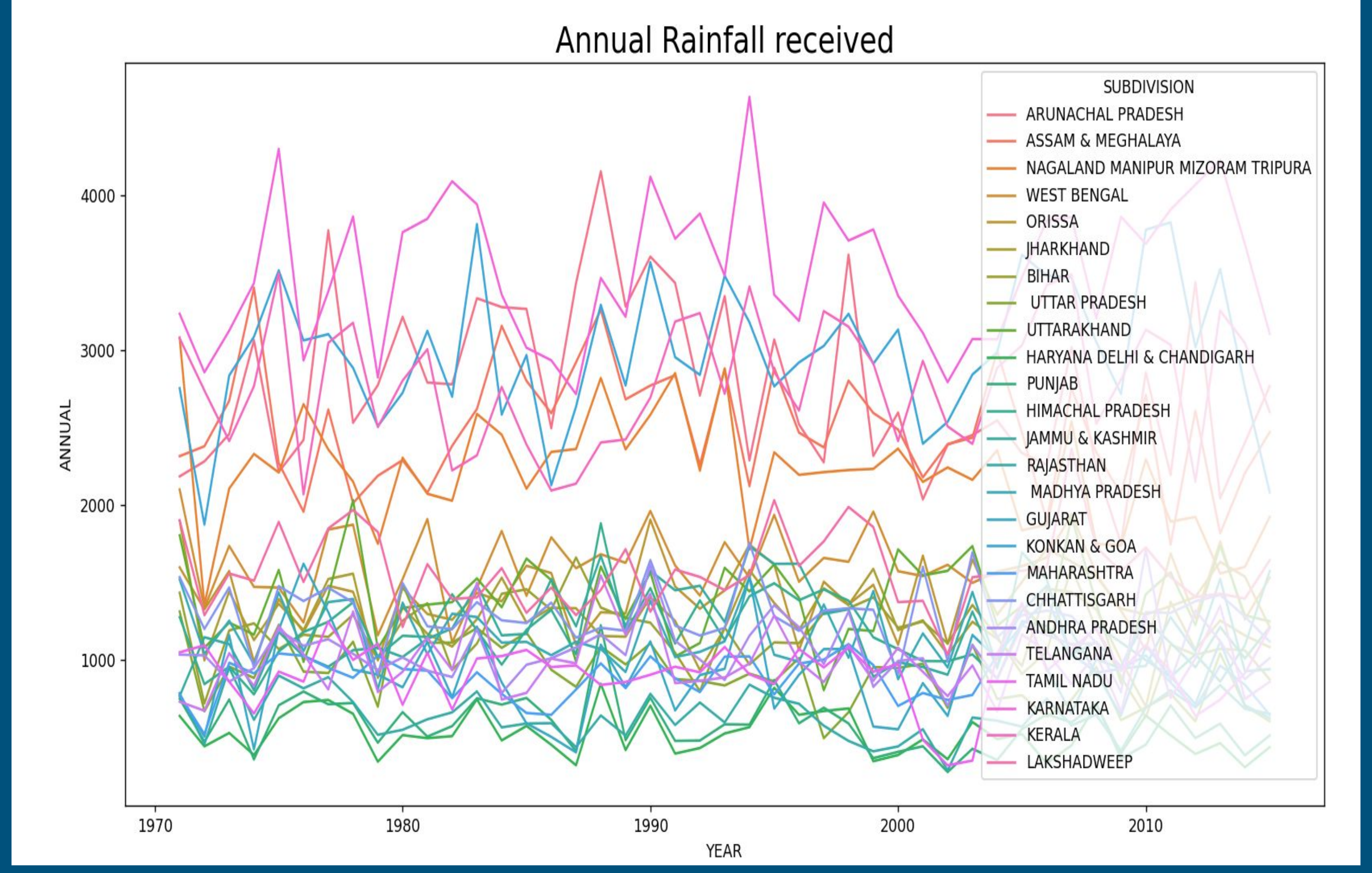
CLUSTERING WITH SIX CLUTER

RESULT

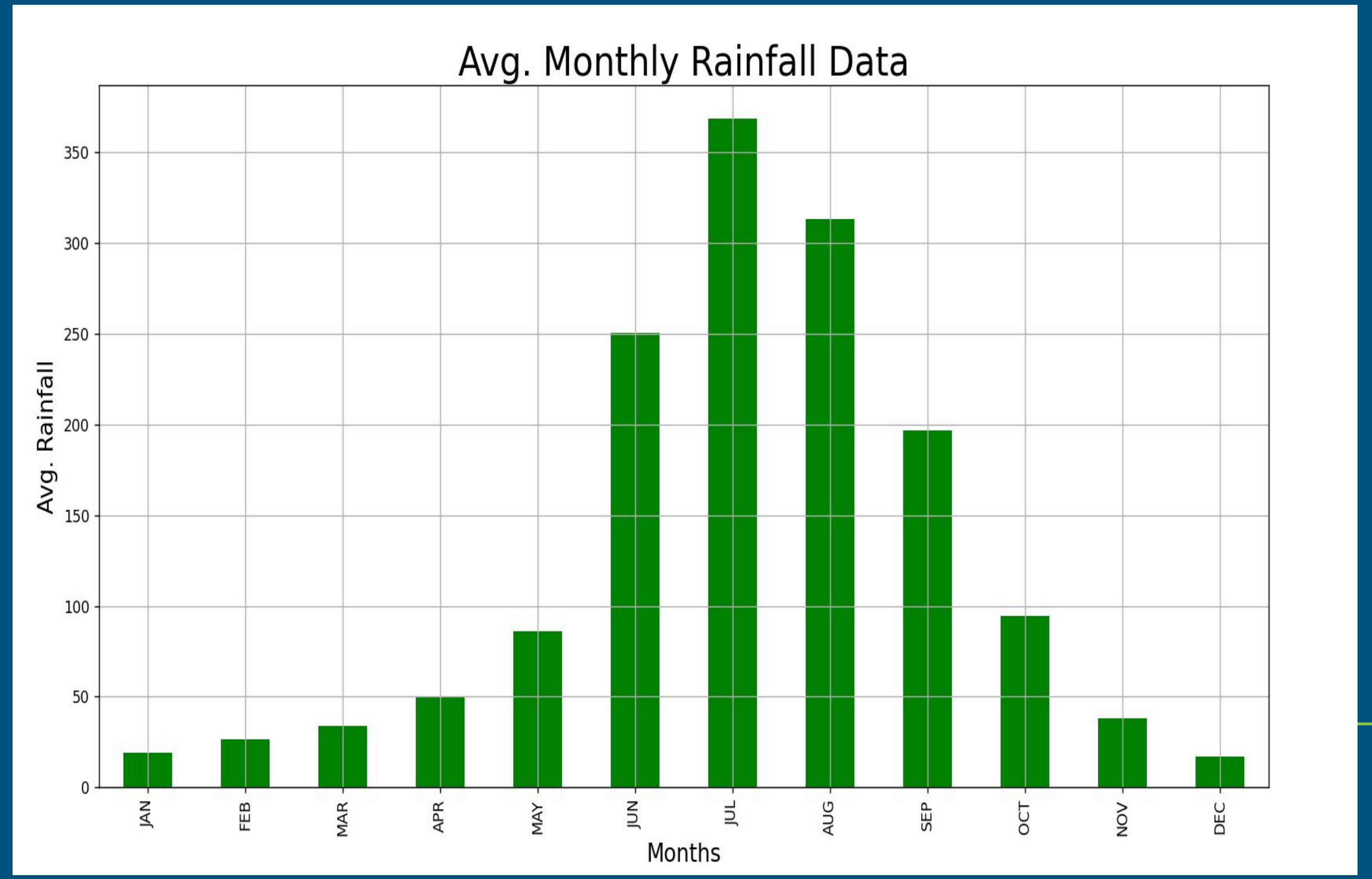


STATES CLUSTERED BY ITS DISTRICTS

DATA VISUALIZATION



YEARLY RAINFALL IN EACH STATE



AVERAGE RAINFALL IN EACH MONTH

CONCLUSION AND FUTURE WORK

The data mining tasks we executed on "Rainfall in India" revealed data that helps farmers estimate rain effectively using techniques such as regression and clustering. Currently machine learning is used in no. industries. As the data increases the complexity of that data will increase and for that we are using tools such as weka and python to better understand the data. In future we are planning to increase our work in Storm predictions and Crop prediction with the rainfall prediction.

Because farmers' survival is strongly based on rainfall, their suicide rates can be forecast using future data analysis.

REFERENCES

1) https://www.kaggle.com/rajanand/rainfall-in-india
2) Provost, F., & Fawcett, T. (2013). Data science for business: [what you need to know about data mining and data-analytic thinking]. Sebastopol, Calif.: O'Reilly