Rainfall Pattern Analysis and Prediction

Dhruv Dharmendra Shah
Computer Science and Information Systems
BITS Pilani Hyderabad Campus
Hyderabad,Telangana, India
dhruv23899@gmail.com

Ashish Gupta
Computer Science and Information Systems
BITS Pilani Hyderabad Campus
Hyderabad,India
ashishggupta0201@gmail.com

Vedant Goyal
Computer Science and Information Systems
BITS Pilani Hyderabad Campus
Hyderabad, Telangana, India
vedant2719@gmail.com

Vaibhav Yadav
Computer Science and Information Systems
BITS Pilani Hyderabad Campus
Hyderabad,Telangana, India
yadav.vaibhav46.vy@gmail.com

Abstract— **The purpose of this project is to mine rainfall data for Indian states using the dataset as provided on data.gov.in/. The project is divided into two phases. The first phase consists of data cleaning, data pre-processing and preliminary data visualization. The second phase will involve the application of different data mining algorithms and gaining insights from it subsequently. The project also aims at predicting future rainfall trends and provide useful insights in rainfall distribution of Indian Subcontinent. This can further help in policy making and planning by the government.**

I. Introduction

Data Mining has a great importance in today's world. It plays an important role in defining and creating solutions ranging from gaining important business insights and helping in making future decisions related to government as well as non-government organizations.
Prediction of future rainfall trends which is an essential data mining task has a significant impact on decision making.
The insights gained from these data mining tasks can be used to plan the crop sowing periods as well as identify the appropriate harvesting time. It can also help in predicting drought patterns in different regions of the country and prepare redressal policies to cope up with the same.

II. Dataset Description

The data consists of the rainfall pattern from 1901 to 2017 for the Indian subcontinent as well as its individual regions. All the rainfall measures are in millimetres(mm). Rainfall data is available in monthly as well as quarterly manner. Data is available for many regions with a few missing values. Appropriate data pre-processing techniques can be applied to help mitigate these missing data points. There are approximately 17 columns and 4190 rows in the Indian Subcontinent dataset. There are multiple datasets for regional locations which can be merged and processed upon for clustering and other essential data mining tasks.

The attributes in the data represent monthly, annual as well as quarterly rainfall data for the Indian subcontinent. The objects here represent various states with respect to different years (1901-2016).

III. Data Cleaning

Data cleaning is the process of detecting and correcting inaccurate records from a dataset.

Two data cleaning techniques were performed.

- Handling missing data: The dataset consisted of multiple rows having cells with missing values. This problem was overcome by using measures of central tendency like mean and median . Mean and median for each column grouped by regions were computed and were substituted in the place of the missing values. The basic assumption here is that the missing values won't differ much from the measures of central tendency. Otherwise complete rows would have to be ignored leading to data loss.
- Noise Reduction: This technique is usually performed to deal with fluctuations in data. In this case smoothening by bin medians was performed. The technique involves dividing each column into intervals. Median for each interval was computed and the data points lying in those intervals were replaced by the computed median. This leads to smoothening of data and hence removal of unnecessary fluctuations in data.

III.Data preprocessing techinques

Following Data pre-processing Techniques have been performed.

A. Normalization

Normalization is used to scale the data of an attribute so that it falls in a smaller range. It is generally useful for classification algorithms.
In this case, we require Normalization for running different data mining algorithms in future such as Clustering etc.

Two Normalization techniques were applied:

- Min-Max Normalization: In this technique of data normalization, linear transformation is performed on the original data. Minimum and maximum value from data is fetched and each

value is replaced according to the following formula.

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)}(\text{new\_max}(A) - \text{new\_min}(A)) + \text{new\_min}(A)$$

- Z-score normalization: In this technique, values are normalized based on mean and standard deviation of the data. The formula used is:

$$v' = \frac{v - \overline{A}}{\sigma_A}$$

B. Aggregation

- State-wise Aggregation: The data for all the years (1901-2017) was combined and aggregated for each of the states. The mean of rainfall data across the years was captured in month wise, annual and quarter wise manner for each state. This aggregated data will prove to be useful while comparing rainfall among states based on historical as well as recent data.
- Decade-wise aggregation: The data for all the years was aggregated and combined into decades for each of the states. The mean of rainfall data across years in each decade was captured in month wise, annual and quarter wise manner for each state. This aggregated data will prove to be useful while comparing rainfall data among states as well as the trends in rainfall across decades for each state.

C. Discretization

Discretization is the process of putting values into bins so that there are a limited number of possible states. Two types of discretization have been performed:

- Equal Interval Discretization: All the rainfall data across the years and months was combined and minimum and maximum values were computed. The whole dataset was divided into 3 intervals (low, medium and high) based on min and max values. All the datapoints were dumped into their respective intervals based on their values.
- Equal frequency discretization: All the rainfall data across the years and months was combined and 3 intervals (low, medium and high) were computed such that each interval consists of equal number of data points. All the datapoints were dumped into their respective intervals based on their values.

The above discretization techniques were used for the data visualization to compare states based on the frequency of low, medium and high rainfall months across the years for each state.

D. Sampling

This technique is usually performed when the number of data points are more than what is computationally feasible. While sampling, efforts are made so that the obtained sample is representative of the actual data.
In this case, stratified sampling was performed based on states of India. For each state, sampling with replacement was performed. With-replacement sampling ensures that the probability of selecting any data point remains same while sampling is performed.

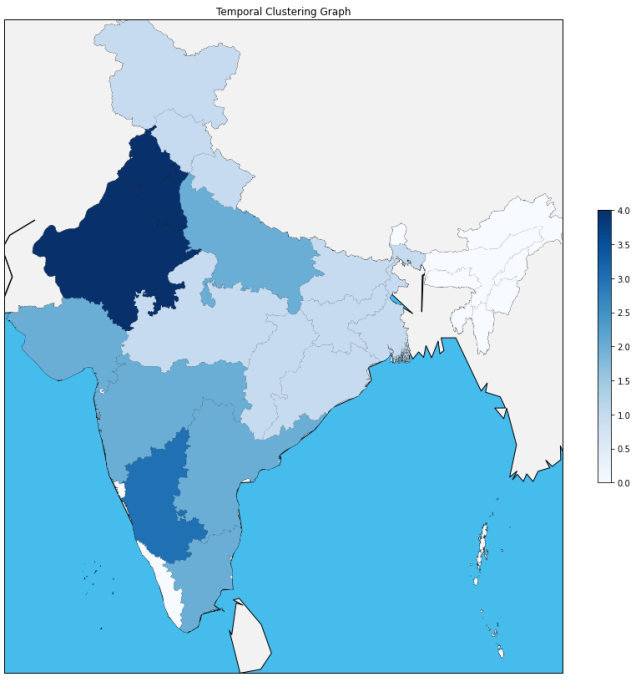
IV. Clustering

Clustering is the task of dividing the data points into a number of groups such that data points in the same group are more similar to other data points in the same group and dissimilar to the data points in other groups. Two kinds of clustering were done in our project.

A. Temporal Clustering using K-Means

It divides the states into 5 clusters based on the rainfall from 1901-2017. Temporal Clustering using K-Means refers to the factorization of multiple time series into a set of non-overlapping segments that belong to k temporal clusters. The most important thing for doing K-means clustering is to know the distance between the data points. So for the case of temporal clustering we defined the temporal distance between two states as the square root of the sum of squares of difference of rainfall between those two states for each year which is nothing but Euclidean distance between two states. Before having a look at the algorithm let us define the mean distance between a state and a cluster which is nothing but the mean of the temporal distance between the selected state and all the states that are present in that cluster. Now let us have a look at the K-Means algorithm. First, we select K=5 that is we want to divide all the states into five cluster. Let us assign a state to each of the clusters randomly. So now each cluster has one state. Now for each state calculate the mean distance between the state and all the five clusters and add the state to the cluster having the least mean distance. Repeat the above step till there is no change in all the clusters for 2 consecutive iterations.

After performing the above algorithm, we get five independent clusters. Now when we plot the clusters in the India map having each cluster with different color, we get the following results.



In the above map the clusters are represented as integers. Cluster_0 is represented with 0. Cluster1 is represented with 1 and so on.

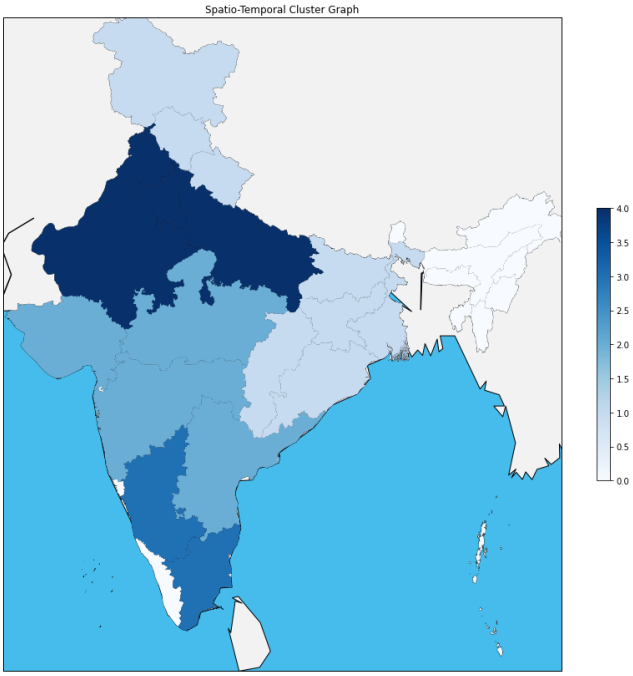
From the above map we can say that the states that are near to each other fall in the same cluster that means that the rainfall occurs in the similar way in the states that are adjacent to each other. So the government can plan easily as to the regions where drought may appear and the regions where there be more than normal rainfall and can plan transfer of food items accordingly.

B. Spatio-Temporal Clustering

Spatio-temporal clustering using K-Means is a process of grouping objects into K clusters based on their spatial and temporal similarity. For the case of spatio-temporal clustering we defined the temporal distance between two states as the square root of sum of squares of difference of rainfall between those two states for each year which is nothing but euclidean distance between two states. For spatial distance we use the Haversine formula which calculates the shortest distance between two points on a sphere using their latitudes and longitudes measured along the surface.

Now we define spatio-temporal distance as the sum of temporal distance and the Haversine distance. Let us define the mean distance between a state and a cluster which is nothing but the mean of the Spatio-temporal distance between the selected state and all the states that are present in that cluster. Now let us have a look at the algorithm. First, we select K=5 that is we want to divide all the states into five cluster. Let us assign a state to each of the clusters randomly. So now each cluster has one state.
Now for each state calculate the mean distance between the state and all the five clusters and add the state to the cluster having the least mean distance. Repeat the above step till there is no change in all the clusters for 2 consecutive iterations.

After performing the above algorithm we get five independent clusters. Now when we plot the clusters in the India map having each cluster with different colour we get the following results.



In the above map the clusters are represented as integers. Cluster_0 is represented with 0. Cluster1 is represented with 1 and so on.

From the above map we can say that the states that are near to each other falls in the same cluster that means that the rainfall occurs in the similar way in the states that are adjacent to each other. Since the adjacent areas are in same cluster that means that these regions may have similar rainfall distribution in the coming years so there can be simplification in the crop plantation and we can also easily classify the regions where there are the chances of drought.

C. *Outlier detection*

Outlier detection (also known as anomaly detection) is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies.
We took the result of Spatio-temporal clustering using K-Means and temporal clustering using K-Means and performed outlier detection.

1.) Outlier detection from the result of Spatio-temporal clustering using K-Means: -
For each cluster in the result calculate the mean of all the data points(states) in the cluster and name it as the mean point. Now calculate the spatio-temporal distance (defined in Spatio-temporal clustering using K-Means) between all the data points(states) in the cluster and the mean point. The data point(state) that has the largest spatio-temporal distance is termed as the outlier. After completion of the above algorithm we get five outliers.

2.) Outlier detection from the result of temporal clustering using K-Means: -
For each cluster in the result calculate the mean of all the data points(states) in the cluster and name it as the mean point. Now calculate the temporal distance (defined in temporal clustering using K-Means) between all the data points(states) in the cluster and the mean point. The data point(state) that has the largest temporal distance is termed as the outlier. After completion of the above algorithm we get five outliers.

D. Correlation

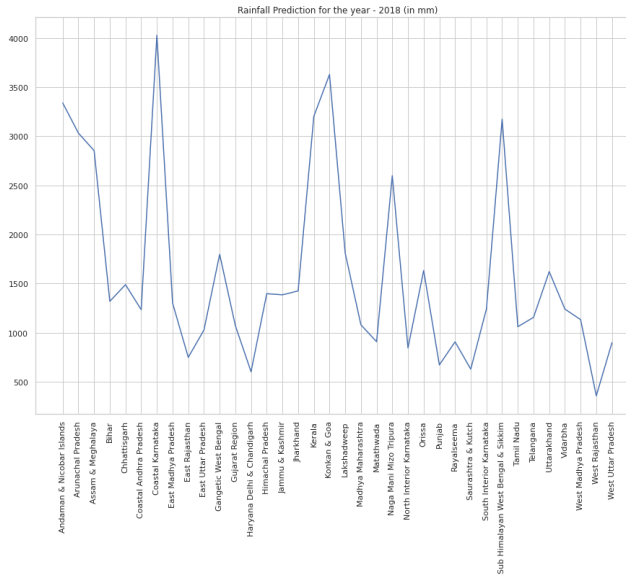
Pearson correlation was used to determine correlation between the states. The Pearson coefficient is a type of correlation coefficient that measures the strength of the association between two continuous variables.
Pearson coefficients range from +1 to -1, with +1 representing a positive correlation, -1 representing a negative correlation, and 0 representing no relationship. The Pearson Correlation formula was used to find the temporal correlation between all the pair of states in India.

V. Rainfall Prediction

A. Next Year Rainfall Prediction

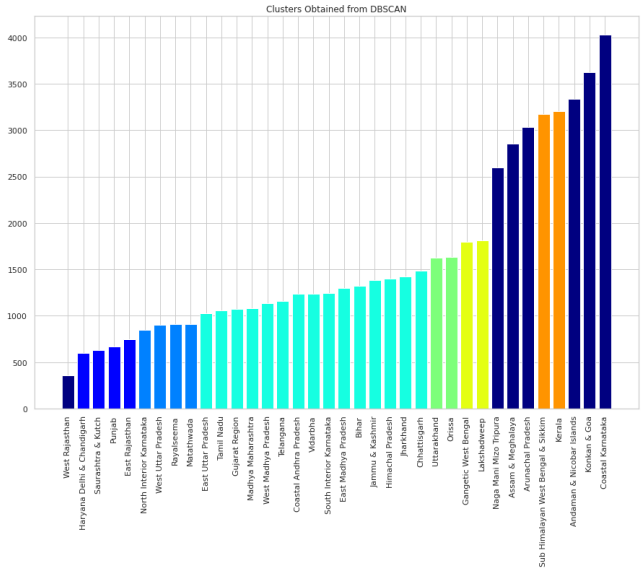
The prediction of rainfall data for the upcoming year holds significant importance as many policy measures at the Government level depend on it. This prediction also has significance in preparing for future Disaster Management challenges.

The dataset consists of the Annual rainfall data of all states in India from the year 1901 to 2017. Linear Regression Algorithm was applied on the above data to get accurate predictions of rainfall amount for the year 2018. For each of the states, the Algorithm was trained on data from the year 1901 to 2017 for 20000 epochs each. The below plot summarizes the amount of rainfall(in mm) predicted for all states for the year 2018.



One important use of predicted rainfall data is to identify flood prone and drought prone areas, to preplan Disaster Management and subsequently reduce damage to life and property. In this case clustering techniques could prove to be useful in classifying states as flood prone or drought prone, based on the predicted data

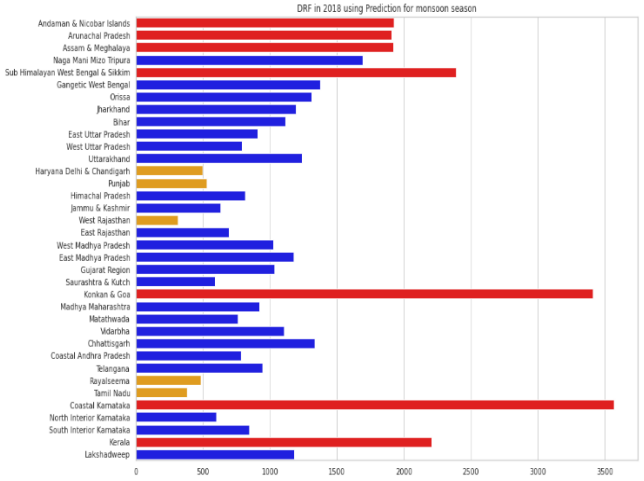
A popular Clustering technique, named DBSCAN was applied on the “Rainfall Prediction for 2018” data obtained above by Linear Regression Analysis.



The above plot depicts the clusters obtained by applying the DBSCAN clustering technique on the predicted data. The states have been sorted based on the predicted annual rainfall amount for 2018. The parameters chosen for the DBSCAN were MinPts=2 and Radius=95mm of rainfall.
The color of the bar for each of the state depicts the cluster to which it belongs. Although the states at the lower end of rainfall amount had higher density as compared to states at higher end of rainfall amount, the DBSCAN technique gave satisfactory results. From the plot, first two clusters from the left can be identified as drought prone areas and the first three clusters from the right end can be identified as flood prone areas. The above clustering can be validated from the fact that states like Rajasthan and Kutch(Gujarat) have been traditionally drought prone areas and states like Kerala and Coastal Karnataka have been flood prone areas.

B. Next Year Monsoon Rainfall Prediction

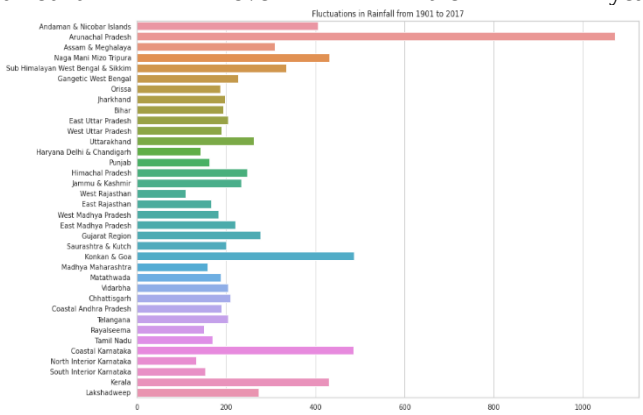
The above two plots pertain to rainfall data for year as a whole. But prediction of rainfall data specific to rainfall months, namely June, July, August and September (JJAS) could prove to be more useful for farmers and predicting Crop Shortages.
Hence Linear Regression Algorithm was applied on rainfall data for the year 1901 to 2017, but limited to rainy months of JJAS. The below plot summarizes the predicted rainfall in Monsoon Season for 2018.



Along with depicting the amount of rainfall the color of the graphs show the low, medium and high rainfall areas. The orange bars depict low rainfall areas, blue bars indicate medium rain areas and red bars indicate high rainfall areas. It can be observed that high monsoons states are also indicated as flood prone areas in the Annual plots. This leads to an important insight that among all twelve months, the June, July, August and September(JJAS) months are of most significance for most of the states. This can be attributed to the fact that the arrival and departure of Monsoon in India occurs during the JJAS months.
The code also includes data of the principal crop for each of the states gathered from expert knowledge. It indicates possible crop shortages in states which are classified as flood or drought prone.

C. *Fluctuations in Rainfall from 1901 to 2017*

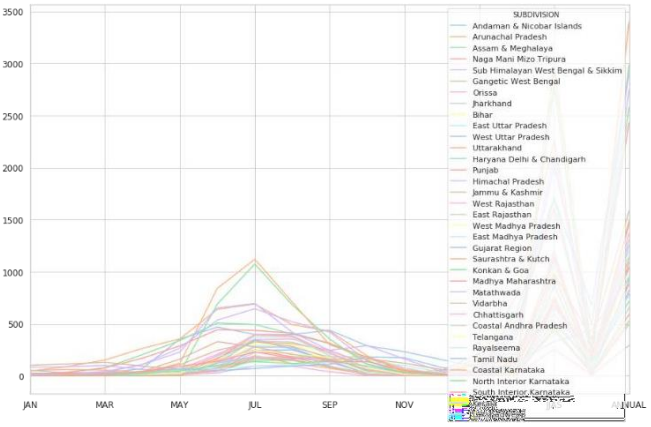
The below plot shows the fluctuations in amount of rainfall over the years in each of the states. The fluctuations have been calculated by evaluating the standard deviation of rainfall amount over the years.



It can be observed from the above plot that states like Arunachal Pradesh, Goa and Coastal Karnataka have highest fluctuations in rainfall which can be attributed to high rainfall patterns in these states.

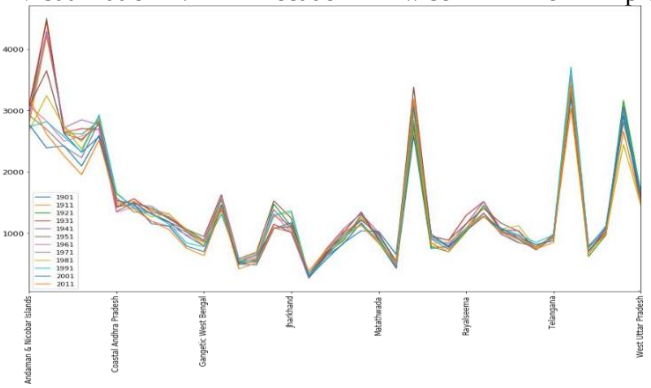
VI. VISUALIZATION

Visualization is the use of graphics to create visual images which helps in the understanding of complex representations of data. Visualization-1: Line Plot



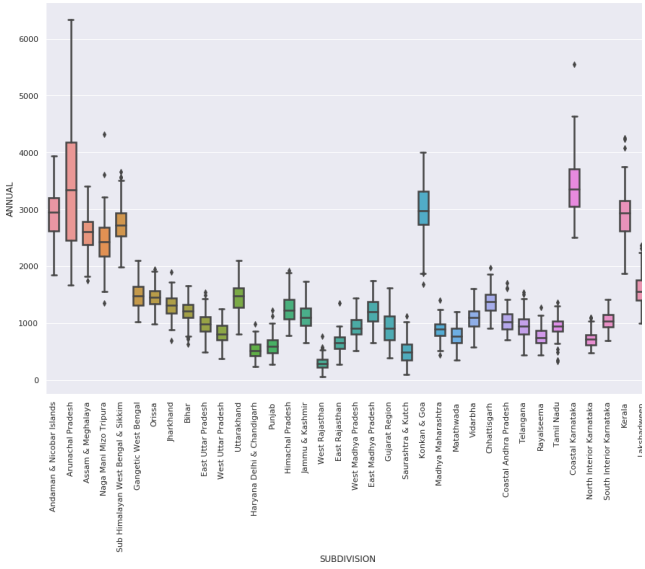
The above plot indicates the rainfall pattern of states over the year. As the plot indicates, there is a surge in rainfall across the country between July and September. (Guhathakurta & Rajeevan, 2008) It also shows that Karnataka and the north east region (West Bengal, Assam etc.) receives high amount of rainfall compared to other states.

Visualization-2: Decade wise line plot



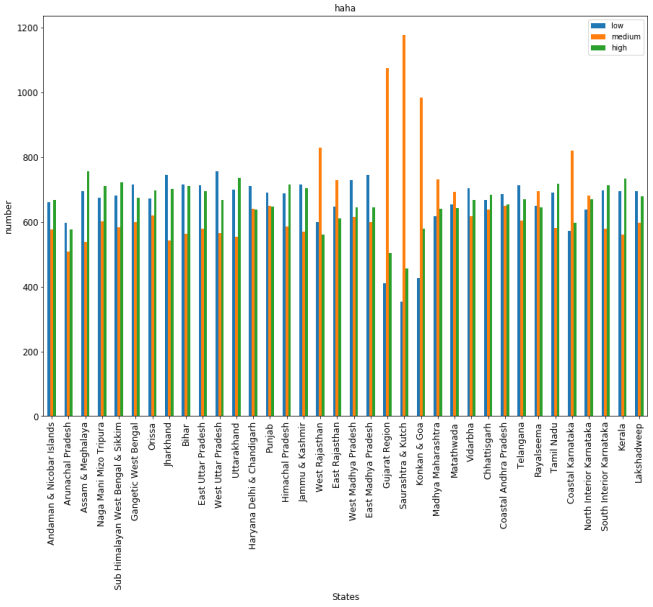
The above plot shows the variation in rainfall for different states across the decades from 1901-2017. As the plot shows, the variation in rainfall across decades does not vary much, particularly for the inland states. We can notice variation for the islands like Andaman and Nicobar Islands as they lie in ocean away from the mainland.

Visualization-3: box plot



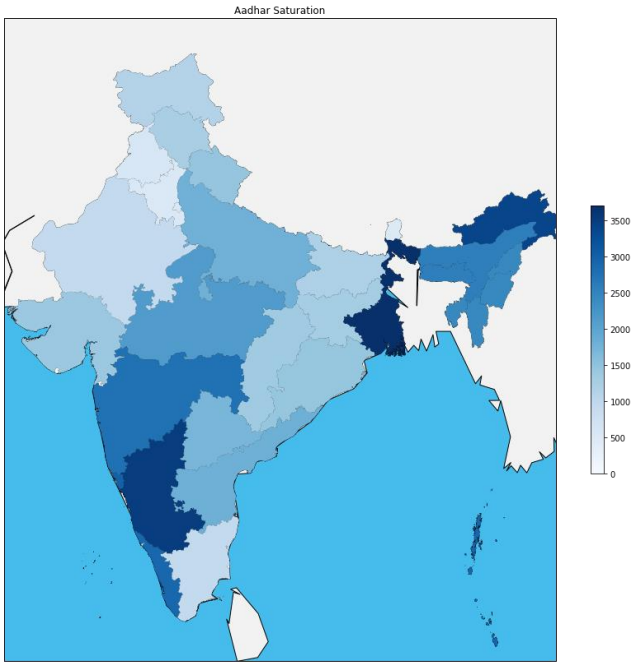
The above plot shows the box plot diagram for all the states. The box plot diagram for Arunachal Pradesh indicates significant spread of rainfall around the median. The box plots located above the normal range indicates high amount of rainfall for those particular states.

Visualization-4: Bar Plot + Histogram (Low, Medium, High)



The above plot shows the frequency of monthly rainfall data points classified as low, medium and high for all states.We can notice that for desert areas like Rajasthan, the frequency of monthly rainfall being classified as low is more than being classified as medium or high. For South-Western states and North-east regions, we see that frequency of monthly rainfall being classified as high is more than being classified as medium or low. For other states there is no significant difference among the frequency. (Kumar, Jain, & Singh, 2010)

Visualization-5: India spatial graph

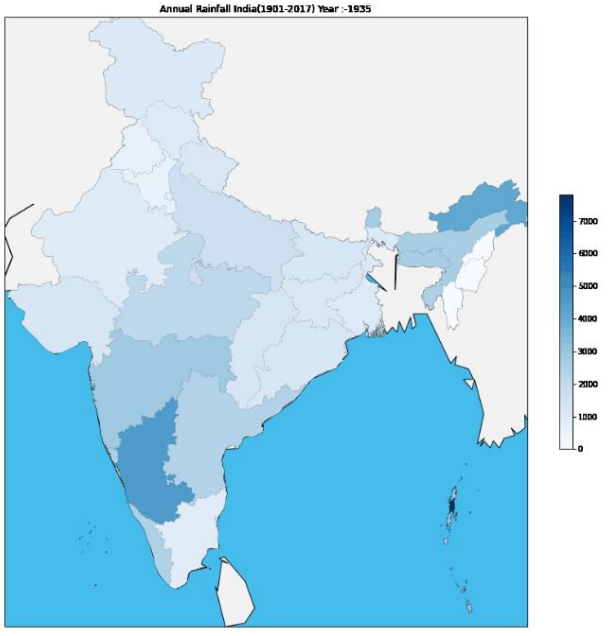


The above plot depicts the annual mean rainfall across years (1901-2017) for different states. The colour intensity highlights the amount of rainfall for different states or regions.

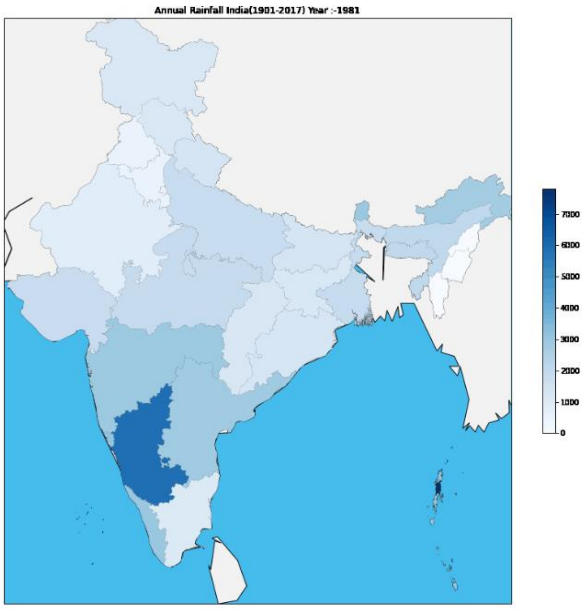
The plot clearly shows that the amount of rainfall in south western states like Karnataka, Kerala, Maharashtra and north eastern regions have significantly higher amount of rainfall . The north western states like Rajasthan, Punjab and Gujrat are drier than other regions. This data supplements the fact that the north western states have desert regions. Also, islands like Lakshadweep and Andaman and Nicobar Islands have high amount of rainfall as they lie deep in the ocean away from the mainland. The path of the southwest monsoons can also be identified to an extent from the graph as the direction from southwest to northeast includes states having higher rainfall than their adjoining regions. (Guhathakurta & Rajeevan, 2008)

Visualization-6: GIF to plot the Annual rainfall of all the states between 1901-2017

Snip1:- 1935



Snip2 :- 1981



The above snips are the visualizations of the annual rainfall in India for the year 1935 and the year 1981 in the GIF. The dark blue color indicates more rainfall and as the colour fades to white the amount of rainfall decreases for the state. From the GIF we can say that all the states have almost similar amounts of rainfall each year. The other trend that we get is that the amount of annual rainfall decreases as we go from the southern region to the northern region for almost every year that tells us the impact of the south-west monsoon on India. We can tell that the south western states like Karnataka, Kerala, Maharashtra dominates the overall rainfall in India in almost every year and the north western states like Rajasthan and Gujarat have the least rainfall in almost every year. The rainfall is evenly distributed in middle India every year.

References

[1] Guhathakurta, P., & Rajeevan, M. (2008). Trends in the rainfall pattern over India. International Journal of Climatology: A Journal of the Royal Meteorological Society, 1453--1469.

[2] Kumar, V., Jain, S. K., & Singh, Y. (2010). Analysis of long-term rainfall trends in India. *Hydrological Sciences Journal--Journal des Sciences Hydrologiques*, 484-496.

[3] Github Repo Link: https://github.com/ashish0201/Rainfall-Pattern-Analysis-and-Prediction