EC2011E – FOUNDATIONS OF DATA SCIENCE

Course Project Report

Names of the students: S. Manjusree

S. Sai Sri Harsha

Project Name: Analysis of Rainfall in Various Subdivisions of India over the period 1901-2021

Department of Electronics & Communication Engineering
NATIONAL INSTITUTE OF TECHNOLOGY CALICUT
Kozhikode - 673601, KERALA, INDIA

Table of contents

1) Description of Data:

1.1) Data source:

The dataset is collected from(Primary data source): https://www.data.gov.in/resource/rainfall-sub-division-and-its-departure-normal-monsoon-session-1901-2021
It consists of rainfall in various subdivisions of India over the period of 1901 – 2021. ChatGPT is used as the secondary source of data. The file is downloaded in CSV format and pandas is used to convert it into SQLite database format. The dataset has around 7 columns (Namely, **Subdivision, Jun, Jul, Aug, Sep, Jun – Sep**) and 4332 rows. The columns Jun, Jul, Aug, Sep indicate the amount of rainfall (in mm) in those respective months in a given year and over a given area. **Jun-Sep** column gives the cumulative sum of rainfall in a particular subdivision for a particular year. To convert the CSV file to SQL database and convert the database to 3rd Normal form, one of the seven columns is to be dropped (**Jun -Sep**). So, the database consists of only 6 columns and 4332 rows along with 2 tables. It is a very good collection for performing data preprocessing, data visualization and data exploration. The final database consists of **2 tables** and it is in **3rd normal form**. The first table consists of 6 attributes (namely, **subdivision, year, jun, jul, aug and sep**) and the second table consists of three attributes (**subdivision, year and jun – sep**). The attributes **SUBDIVISION and YEAR** combinedly act as composite key of the two tables.

1.2) Data attributes:

The following table describes about the attributes present in the data set.

SI. NO	ATTRIBTE	EXAMPLES
1	SUBDIVISION	Andhra Pradesh, Kerala etc.
2	YEAR	1901, 1902, 1910, 2021 etc
3	JUN	1091, 570, 250 etc.
4	JUL	1091, 570, 250 etc.
5	AUG	1091, 570, 250 etc.
6	SEP	1091, 570, 250 etc.
7	JUN - SEP	2000, 1500, 6300 etc.

2) Data preprocessing:

In this step, all the null and invalid values are eliminated. It is observed that the dataset taken consists of some negative values of rainfall data. As we know that negative values of rainfall doesnt make any sense, those values were filtered out and made equal to 0 to avoid inconsistencies. The final data is free of invalid values.

3) Data visualization:

The main purpose of this visualization is to study the rainfall information of different regions over the past 120 years and classify the regions based on the available data. The classification may not be accurate because the geopolitical factors and other factors that affect the classification are not considered. The following questions are to be answered here:

1) Classify the regions as drought regions and flood prone regions based on the mean rainfall over the past 120 years.

2) Plot the rainfall in a particular subdivision over 120 years and observe the anomalies in the data.

3) Compare the rainfall over different regions for a given year.

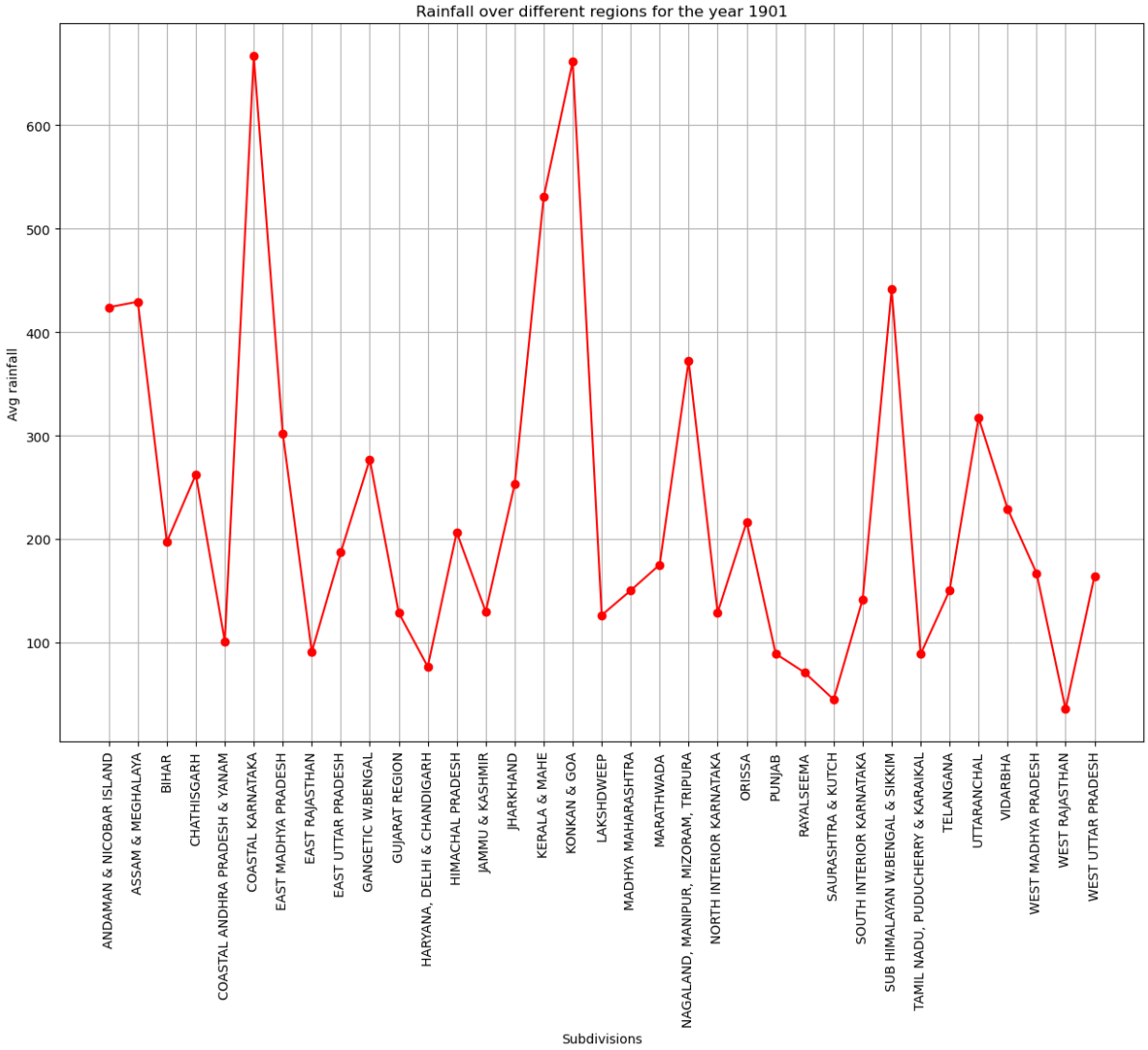
4) Ranking regions from drought prone to flood prone using the data collected over 120 years.

3.1) <u>Comparing the rainfall over different regions for a given year:</u>

In this step, we select a particular year and plot the graph between average rainfall in the given year and the subdivision. This is done to observe the rainfall pattern over all the regions for that given year. For instance, the following table gives the information about average rainfall in the year 1901.

subdivision	avg per year
ANDAMAN & NICOBAR ISLAND	424.075
ASSAM & MEGHALAYA	429.400
BIHAR	197.250
CHATHISGARH	262.300
COASTAL ANDHRA PRADESH & YANAM	101.100
COASTAL KARNATAKA	666.625
EAST MADHYA PRADESH	302.000
EAST RAJASTHAN	91.100
EAST UTTAR PRADESH	187.200
GANGETIC W.BENGAL	277.050
GUJARAT REGION	128.700
HARYANA, DELHI & CHANDIGARH	76.725
HIMACHAL PRADESH	206.475
JAMMU & KASHMIR	129.825
JHARKHAND	253.125
KERALA & MAHE	530.700
KONKAN & GOA	661.425
LAKSHDWEEP	126.425
MADHYA MAHARASHTRA	150.350
MARATHWADA	175.250
NAGALAND, MANIPUR, MIZORAM, TRIPURA	372.250
NORTH INTERIOR KARNATAKA	129.075
ORISSA	216.550
PUNJAB	89.200
RAYALSEEMA	71.000
SAURASHTRA & KUTCH	44.950
SOUTH INTERIOR KARNATAKA	141.650
SUB HIMALAYAN W.BENGAL & SIKKIM	441.350

Table showing the average rainfall in the year 1901 over different subdivisions.



Graph showing the Rainfall over different subdivisions for the year 1901.

It can be inferred from the graph that the regions of COASTAL ANDHRA PRADESH & YANAM and KONKAN & GOA recieved the heaviest rainfall and the regions SAURASHTRA AND KUTCH and WEST RAJASTHAN recieved the least amount of rainfall. Similar graphs can be plotted for the rest of the years and similar conclusions can be made using that data.

3.2) <u>Plotting the rainfall in a particular subdivision for 120 years and observing the anomalies:</u>

In this step, a particular region is selected and the rainfall data for this region is analysed for 120 years to find out the anomalies in the rainfall. In this step, we will be able to find in which years the given subregion experienced abnormally high or abnormally low rainfall. For demonstration purpose, ARUNACHAL PRADESH subdivision is selected.

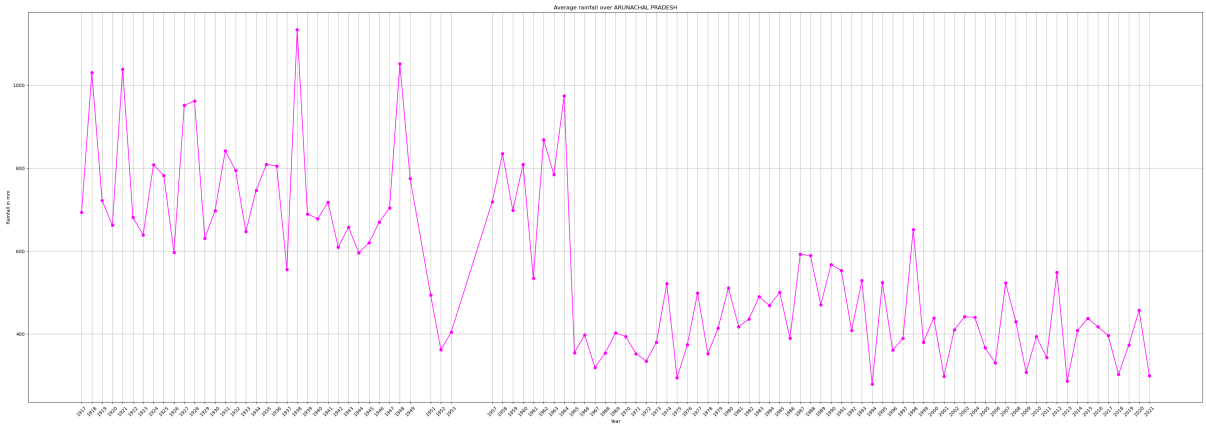


Figure showing the Plot of rainfall over arunachal pradesh over a span of 120 years.

From the above graph, the following conclusions can be drawn:

1) The rainfall in arunachal pradesh has drastically dropped in the year 1985 and remained in the same way after that year.

2) Arunachal pradesh experienced abnormally high rainfall during the year 1917.

Similar conclusions can be drawn from all the plotted graphs in this section.

3.3 and 3.4) <u>Classifying and rankng the regions from flood prone to drought prone:</u>

To classify the regions as flood prone, normal and drought prone, we need to use some **statistics**. We need a correct **measure of central tendency** to get the overall information accurately. For that purpose, **median** chosen as the required measure of central tendency.

1) **<u>Why median?:</u>**
The rainfall data is highly volatile in nature. It varies largely. So it is a highly skewed distribution. For highly skewed distribution, mean may give the accurate information as a single anomaly can highly disturb the value of mean. So median is preferred in such cases.

2) **<u>Statistical analysis:</u>**
A new data frame is created with the attributes **SUBDIVISION** and **median per region**. Median per region is calculated by using the data obtained in the **step 3.2** i.e A subdivision is selected and a table is formed containing **subdivision**, **year** and **avg_per_year**. From this table, median value is selected and is taken as the value of **median_per_region** in the new data frame.

After performing these operations, a new attribute called Classification is added and is given value, based on the median_per_region column. The following algorithm is used to assign a value for classsification:

1) If **median_per_region < 1st quartile**: Drought prone
2) If **median_per_region > 3rd quartile:** Flood prone
3) If **1st quartile < median_per_year < 3rd quartile:** Normal.

.The final graph obtained is as follows:

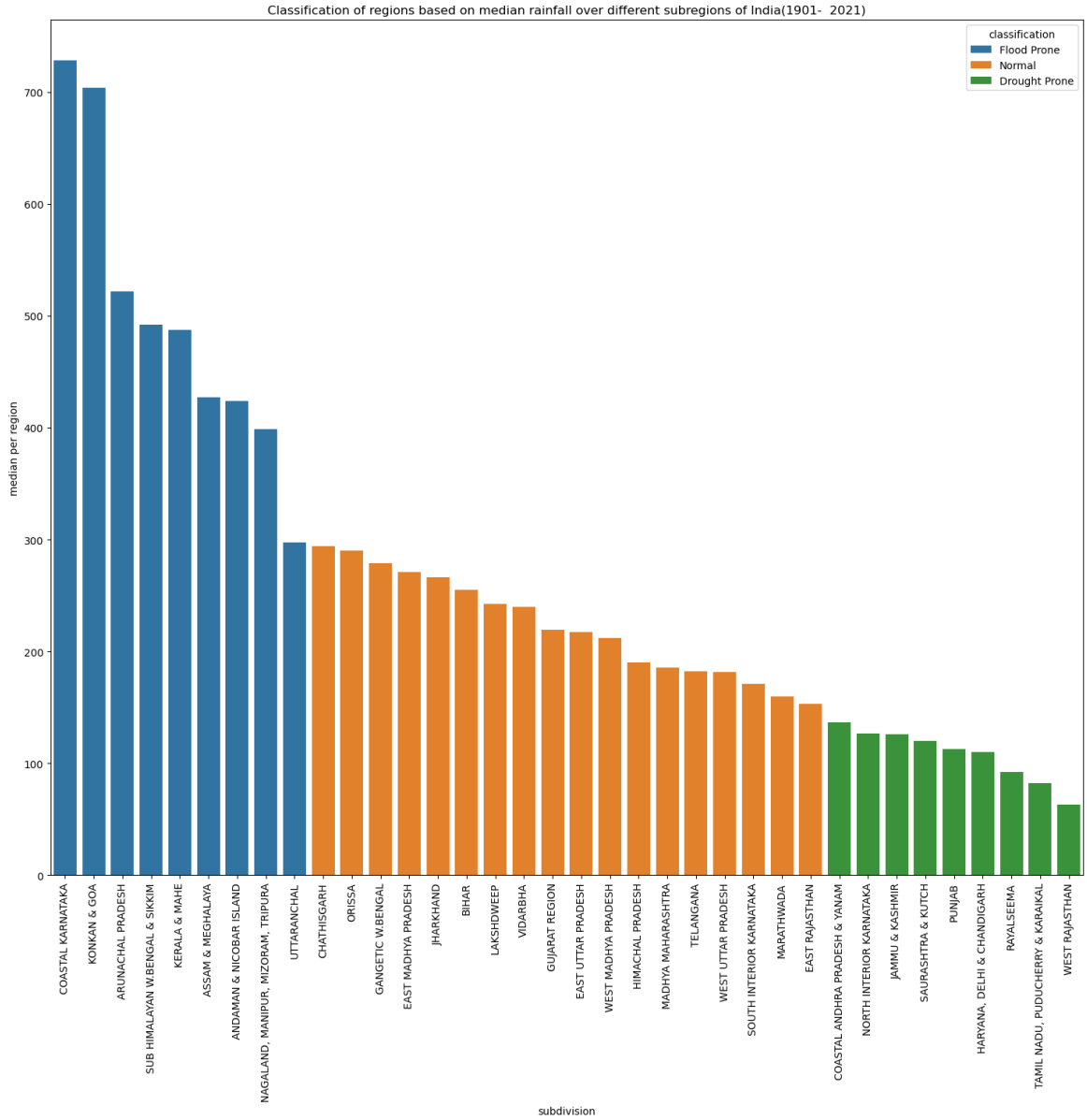


Figure shows the graph between median rainfall and subdivisions based on classification.

The classification column takes three values (Flood prone, normal and drought prone). Classification attribute is given as hue for the above bar chart.

4) Conclusion:

The following conclusions can be drawn from the above work:

1) 9 regions are flood prone regions with Coastal karnataka having the highest amount of rainfall. These places are not much suited for agriculture due to high amount of rainfall.

2) 18 subdivisions are classified as normal. These places are best suited for agriculture as they recieve sufficient amount of rainfall.
3) 9 Regions are classified as drought prone regions. These regions are not suitable for agriculture due to water scarcity.