EC2011E – FOUNDATIONS OF DATA SCIENCE

Course Project Abstract

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

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. It is a very good collection for performing data preprocessing, data visualization and data exploration.

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 analysis that can be performed on the given dataset:

The following data analysis can be performed on the dataset.

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) Analyse how rainfall patterns have changed over years for a particular region.
5) Ranking regions from drought prone to flood prone using the data collected over 120 years.

3) Dataset on which analysis is to be performed:

RF_SUB_1901-2021.csv