Data Engineering
Mini Project
Report

On

INDIAN UNION
BUDGET ANALYSIS

By
Hrishit Madhavi
Hrishikesh Iyer
Rujuta Kulkarni
Ayush Patil

Under the guidance of
Dr. Prashant Lahane

Department of Computer Engineering and Technology
School of Computer Science and Engineering
Dr. Vishwanath Karad MIT World Peace University
AY 2024-25 Semester IV

CONTENTS

I.   Introduction

The Indian Union Budget is a critical framework that outlines the government's fiscal policies and financial allocations for the upcoming year. This project aims to analyze the Indian Union Budget data to identify trends and predict how various financial attributes influence one another. By employing statistical techniques, this analysis will enhance our understanding of the budget's impact on economic growth and public welfare.

II.   About Indian Union Budget Analysis

This project focuses on analyzing the Indian Union Budget to uncover insights into government expenditure, revenue generation, and fiscal policies. By examining historical budget data, we aim to understand the implications of budgetary allocations on different sectors of the economy. The analysis also considers the economic context surrounding budget decisions, such as global economic conditions and domestic challenges.

III. Scope

The analysis spans several years of the Indian Union Budget, providing a multi-year view of economic trends. The focus is on predicting changes in key metrics such as GDP per capita, population, HDI, and life expectancy. The study covers various sectors, including health, education, defense, and

infrastructure, offering a national perspective on the budget's impact.
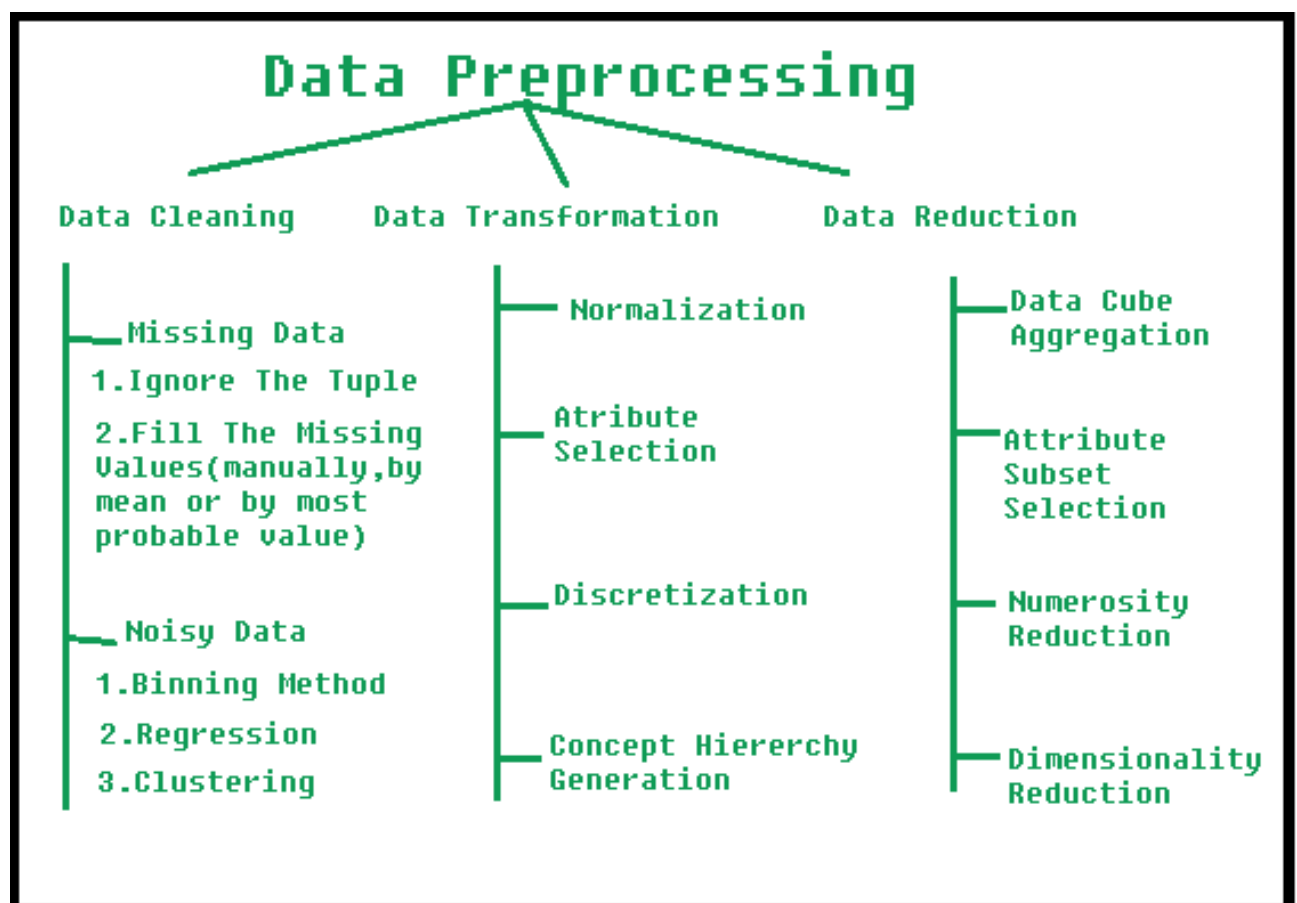
IV. Datasets

We have used the following datasets :
1. International Monetary Fund

V. Data Operations Performed

In this project, the data processing phase involved multiple stages, ensuring the datasets were prepared for accurate analysis and modelling. The key operations included:

Data Pre-processing

Data preprocessing is crucial for cleaning, transforming, and reducing the data to improve model performance. This process was divided into three main components:

1. **Data Cleaning**:

    ○ **Missing Data**:
      ▪ We handled missing data using two main approaches. In some cases, irrelevant records with too many missing values were removed (**ignoring the tuple**). For other cases, missing values were filled using the mean of the respective attribute, ensuring that the dataset remained balanced and complete.
    ○ **Noisy Data**:
      ▪ **Binning**: Continuous data was segmented into discrete intervals, where noisy values were replaced by a smoother set of values within those bins.
      ▪ **Regression**: Some of the outliers were smoothed using regression techniques to better fit the data trend.

2. **Data Transformation**:

    ○ **Normalisation**:
      ▪ We applied **Min-Max Normalisation**, scaling all numerical values between 0 and 1 to ensure that no single attribute dominated others due to magnitude differences. This normalisation was essential for accurate performance in the linear regression models.
      ▪ For some models, **Z-Score Normalisation** was applied, particularly when working with data that followed a normal distribution. This technique centred the data around a mean of 0 with a standard deviation of 1, ensuring equal contribution of all features.
    ○ **Attribute Selection**:

- Only relevant attributes were selected for the final analysis. For instance, in predicting GDP per capita, variables related to fiscal deficits and government spending were retained, while less relevant attributes were excluded.
  - **Discretisation**:
    - In cases where continuous data needed to be grouped, attributes were discretised into intervals. This was particularly useful for models that required categorical input data.

3. **Data Reduction**:

- **Dimensionality Reduction**:
  - High-dimensional data was reduced by selecting a subset of important attributes, ensuring the model was more efficient without sacrificing accuracy.
- **Attribute Subset Selection**:
  - By reducing the dataset to essential attributes, we improved the model's ability to focus on the variables most relevant to the analysis. Unnecessary features were discarded, speeding up the model training process.

Linear Regression Models

After data preprocessing, we built several linear regression models to analyse and predict how various attributes influenced one another:

- **Year vs. Population**: A linear regression model was applied to predict future population trends. The model showed consistent growth in population over time.

- **Year vs. GDP Per Capita**: This model predicted an increase in GDP per capita, reflecting positive economic growth over the years.
- **Year vs. HDI and Life Expectancy**: The model projected improvements in human development indicators, including life expectancy, based on historical data.
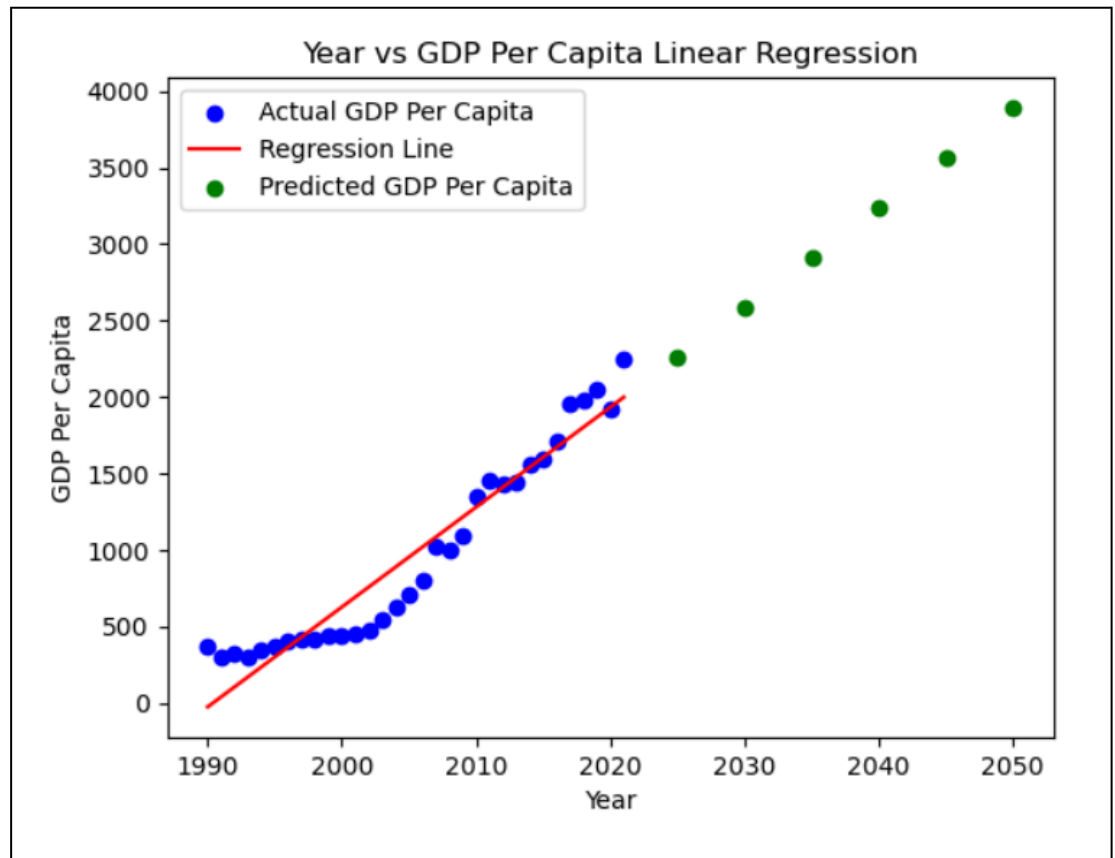
The results from the regression models were used to generate predictions, offering insights into how budgetary allocations affect population growth, economic health, and human development indicators.
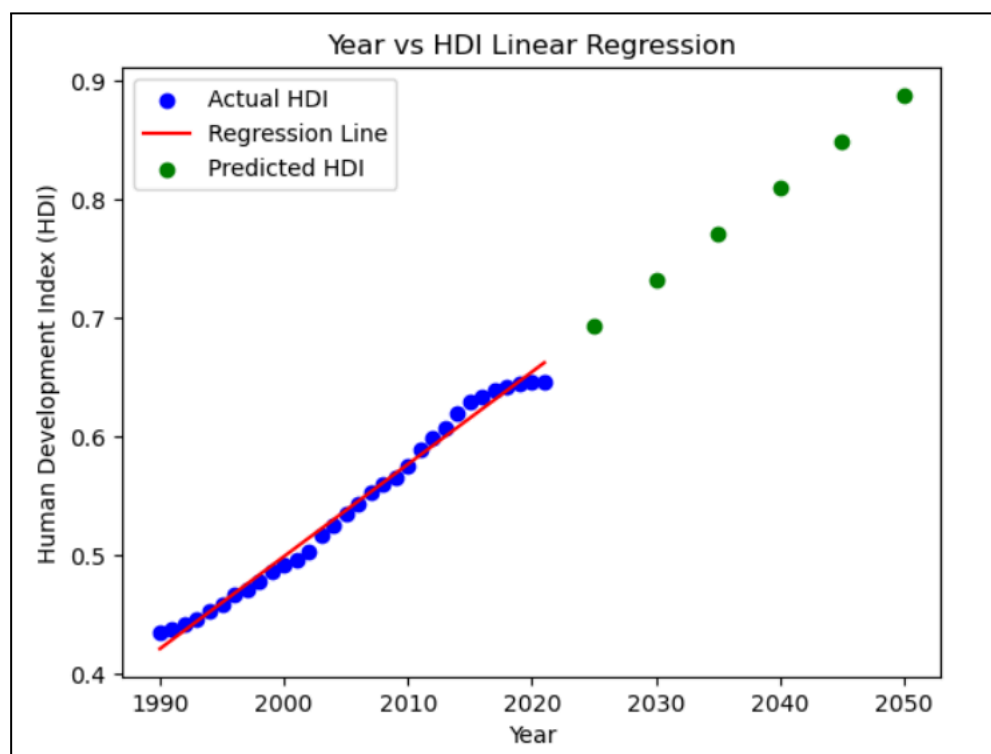
VI. Outcome and Visualisations
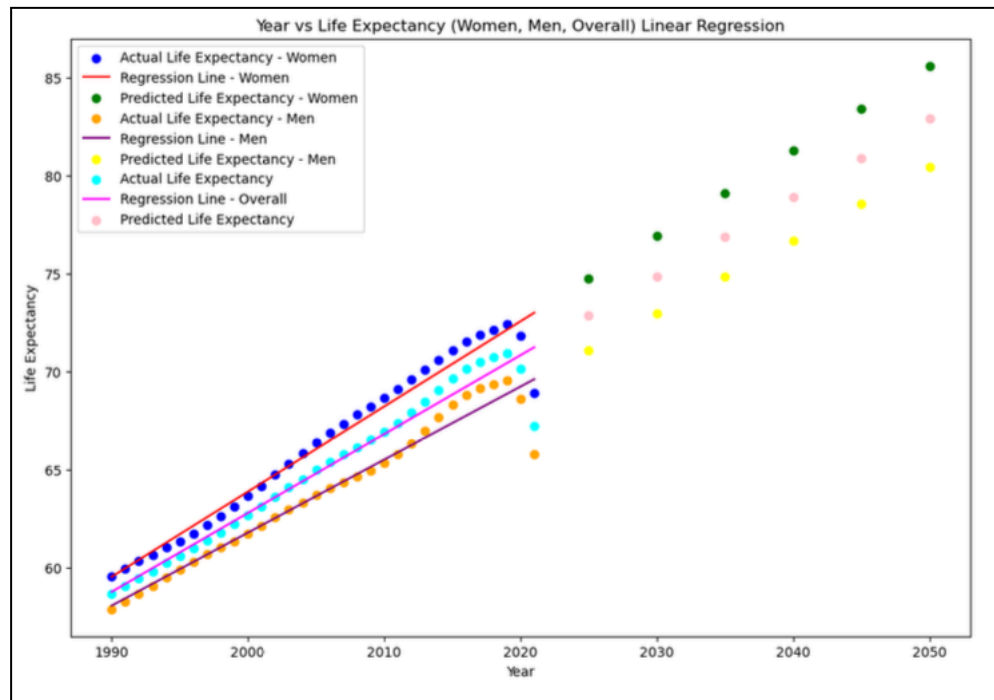
1. Year-Population

2. Year-GDP_per_Capita

Year vs GDP Per Capita Linear Regression
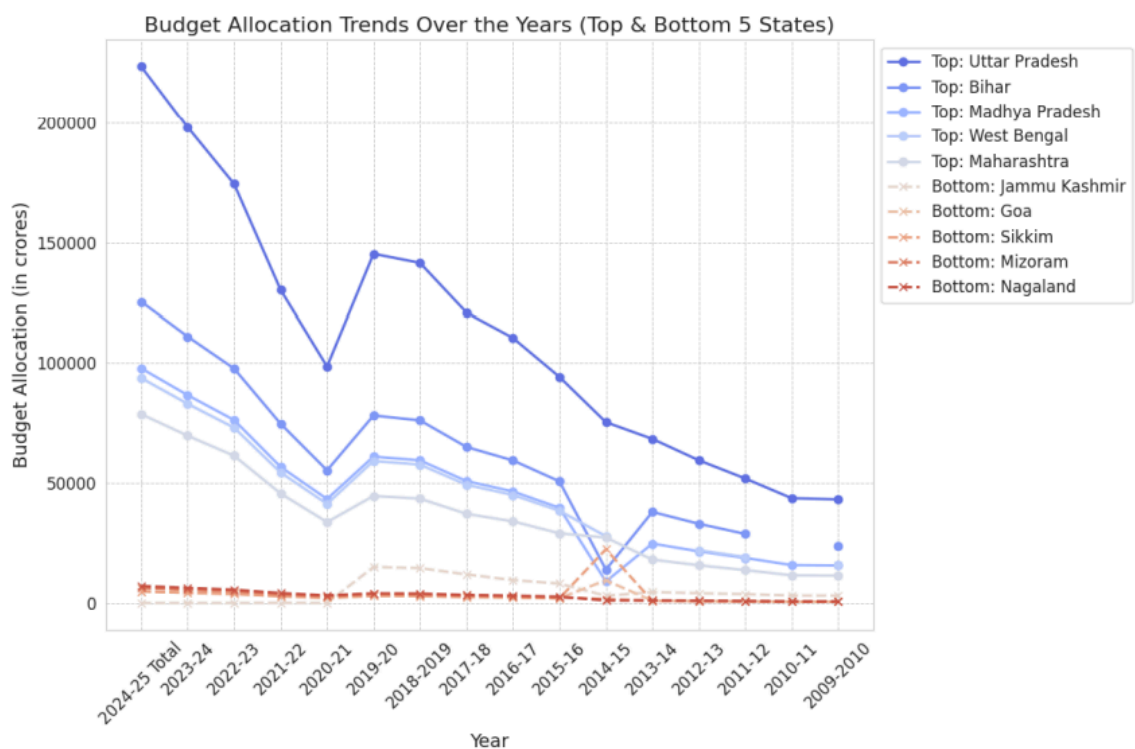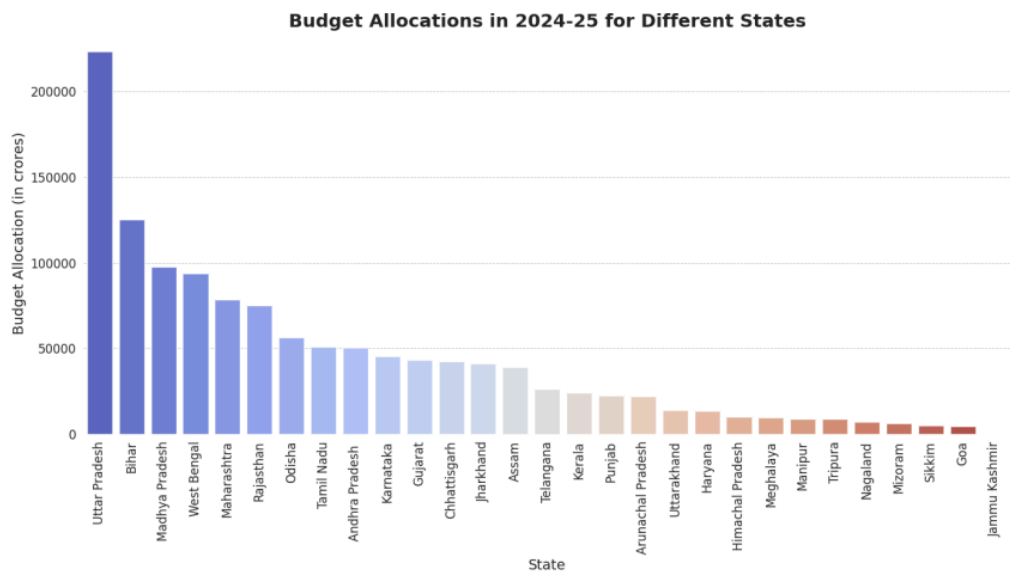
3. Year-HDI

Year vs HDI Linear Regression
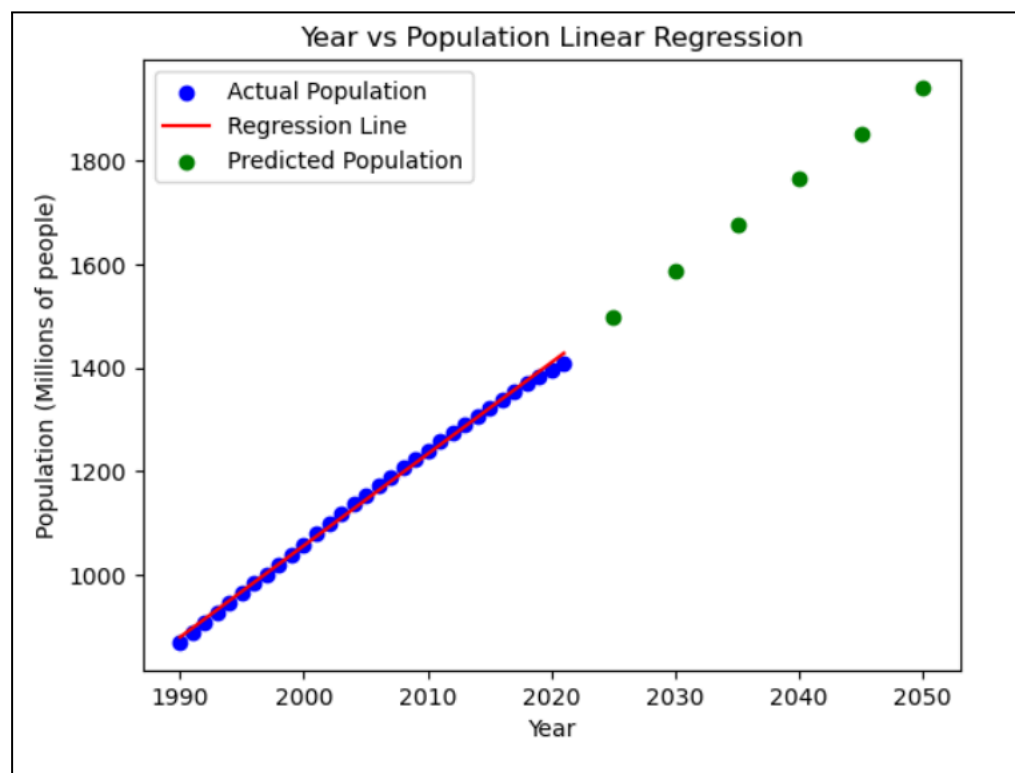
4. Year-Life_expectancy
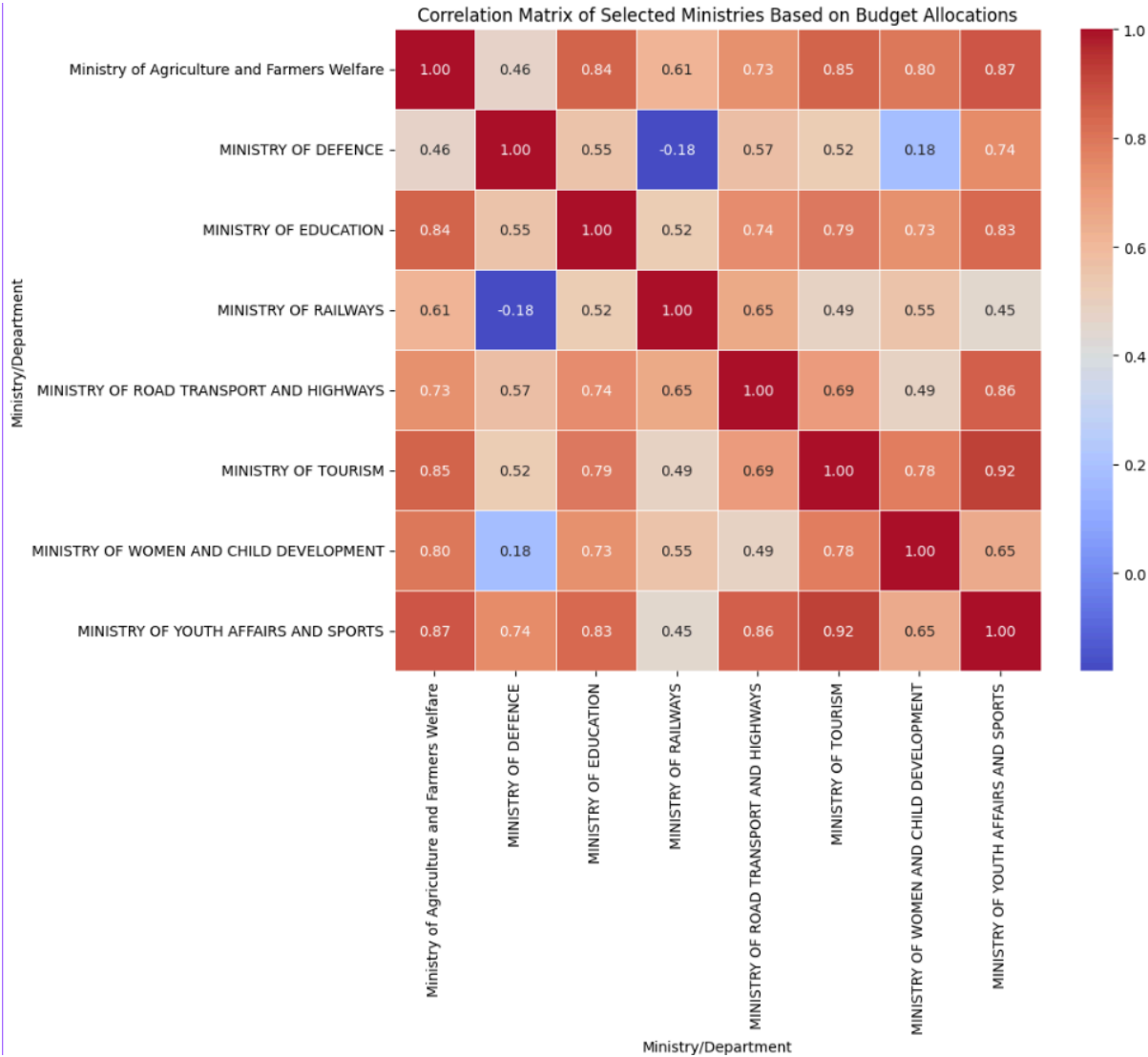

5. Budget Allocation Trends Over the Years (Top & Bottom 5 States)
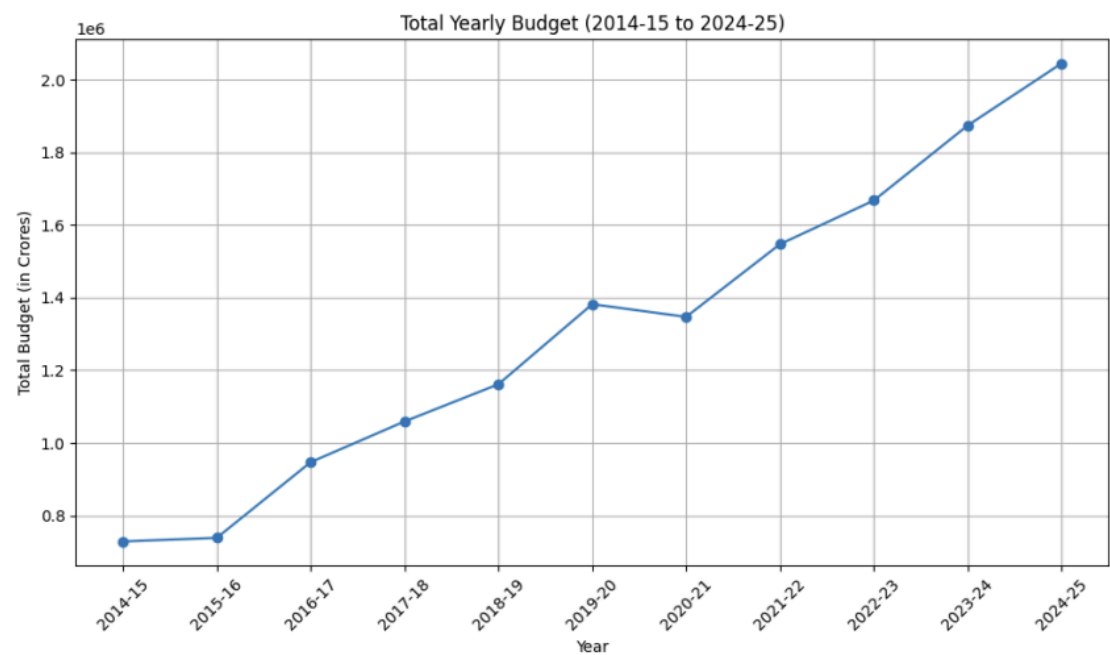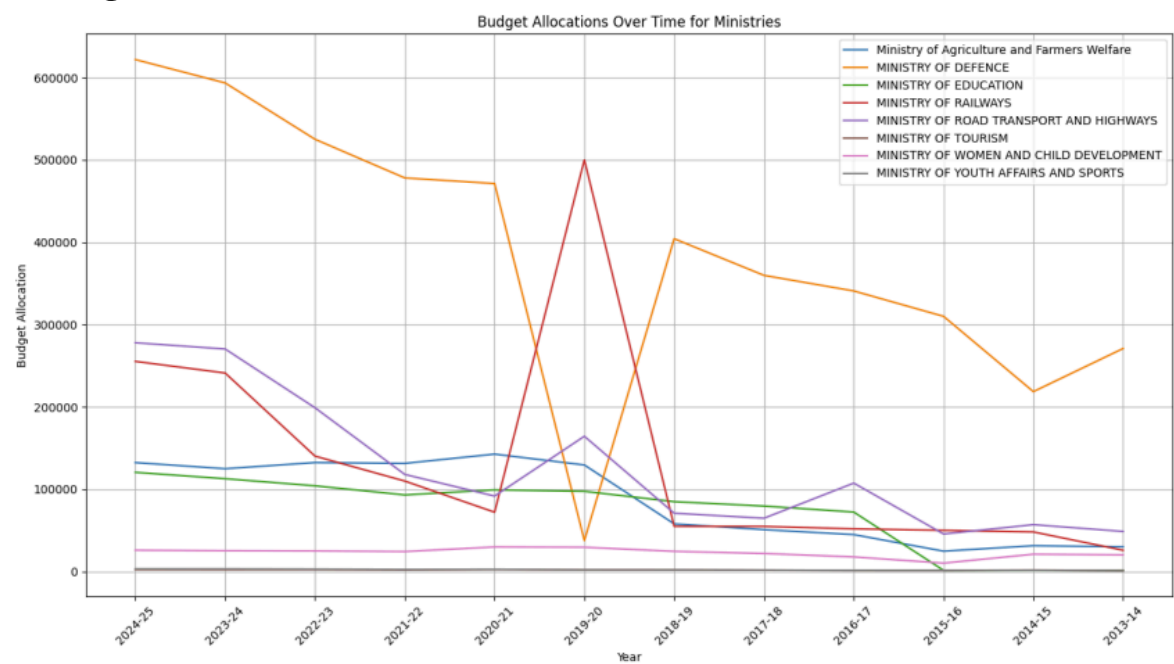
6. Budget Allocations in 2024-25 for Different States

Budget Allocations in 2024-25 for Different States

7. Correlation Heatmap of Budget Allocations Across Years

Correlation Matrix of Selected Ministries Based on Budget Allocations

| Ministry/Department	Ministry of Agriculture and Farmers Welfare	MINISTRY OF DEFENCE	MINISTRY OF EDUCATION	MINISTRY OF RAILWAYS	MINISTRY OF ROAD TRANSPORT AND HIGHWAYS	MINISTRY OF TOURISM	MINISTRY OF WOMEN AND CHILD DEVELOPMENT	MINISTRY OF YOUTH AFFAIRS AND SPORTS
Ministry of Agriculture and Farmers Welfare	1.00	0.46	0.84	0.61	0.73	0.85	0.80	0.87
MINISTRY OF DEFENCE	0.46	1.00	0.55	-0.18	0.57	0.52	0.18	0.74
MINISTRY OF EDUCATION	0.84	0.55	1.00	0.52	0.74	0.79	0.73	0.83
MINISTRY OF RAILWAYS	0.61	-0.18	0.52	1.00	0.65	0.49	0.55	0.45
MINISTRY OF ROAD TRANSPORT AND HIGHWAYS	0.73	0.57	0.74	0.65	1.00	0.69	0.49	0.86
MINISTRY OF TOURISM	0.85	0.52	0.79	0.49	0.69	1.00	0.78	0.92
MINISTRY OF WOMEN AND CHILD DEVELOPMENT	0.80	0.18	0.73	0.55	0.49	0.78	1.00	0.65
MINISTRY OF YOUTH AFFAIRS AND SPORTS	0.87	0.74	0.83	0.45	0.86	0.92	0.65	1.00

8. Total Yearly Budget (2014-15 to 2024-25)

9. Budget Allocations Over Time for Ministries

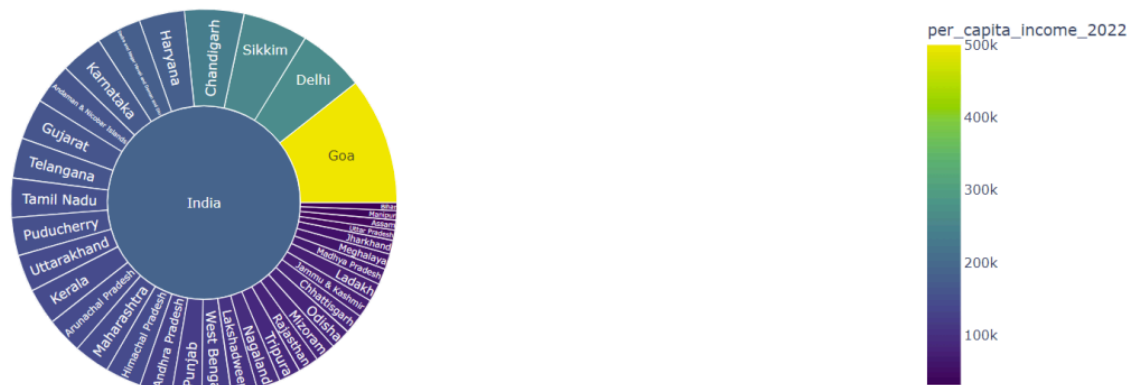# 10. Distribution of Budget Allocations by Ministry
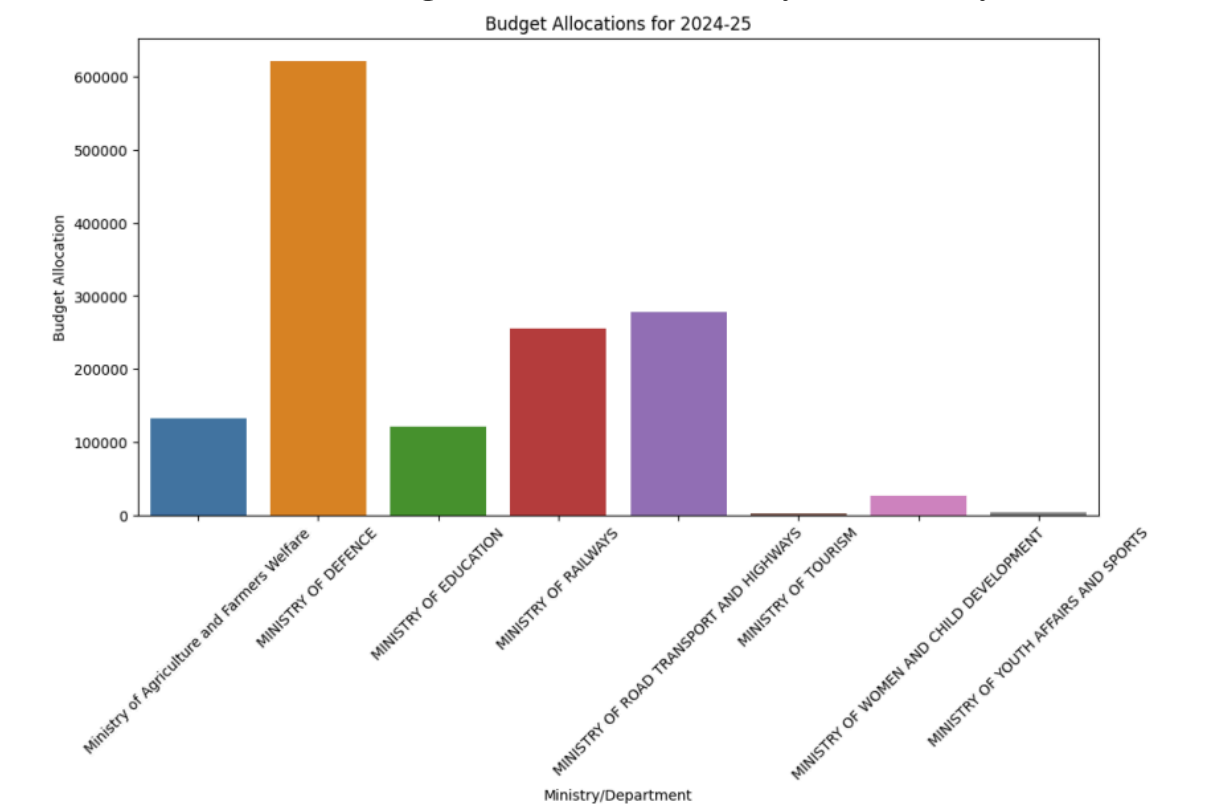
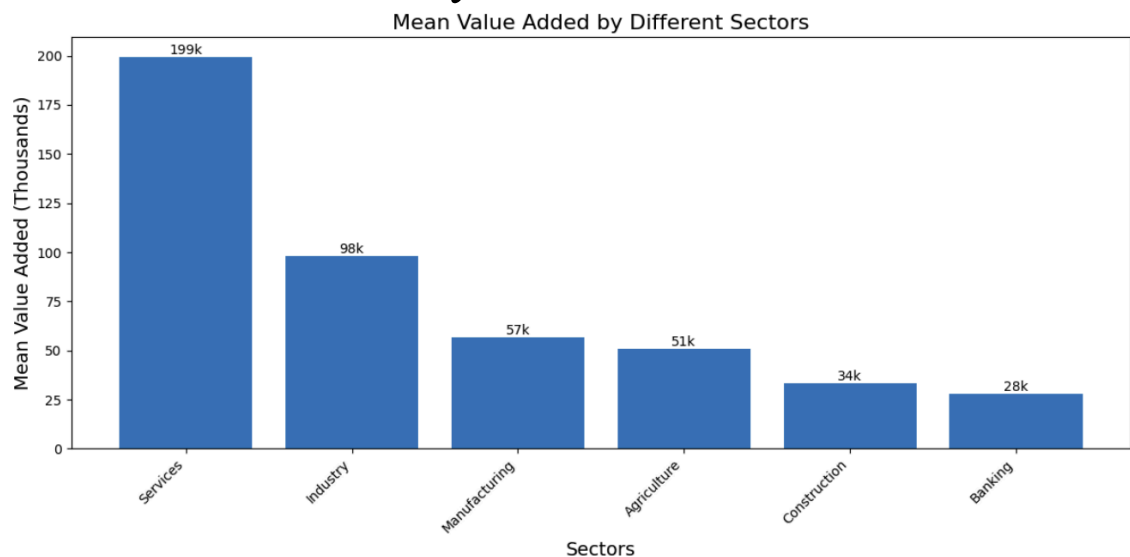# 11. Per Capita Income Trend Across States Over the Years
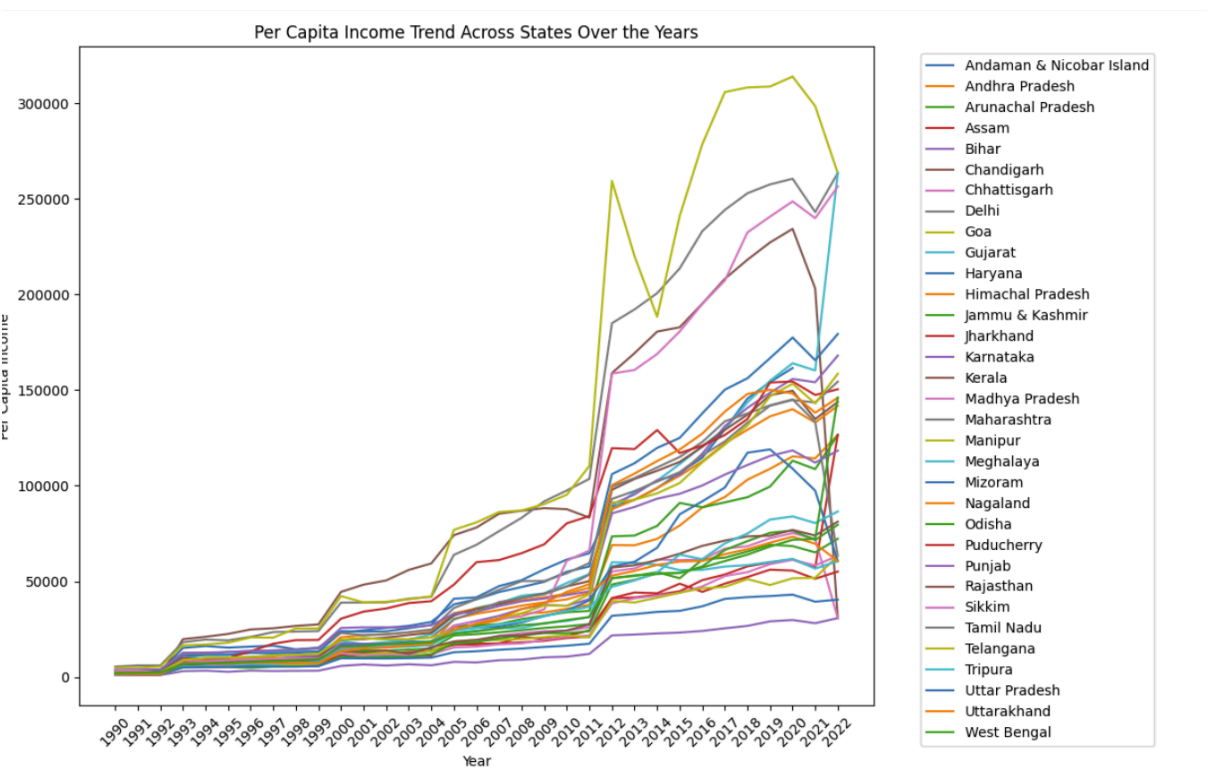
12. Per Capita Income of Indian States and UTs (2022)

Per Capita Income of Indian States and UTs (2022)

13. Mean Value Added by Different Sectors

Mean Value Added by Different Sectors

VII. Conclusion

The analysis of the Indian Union Budget using linear regression models reveals important trends in population growth, GDP, and human development. These findings underscore the significant impact of fiscal policies on economic indicators. The results can help policymakers understand future budgetary requirements and make informed decisions to promote balanced and sustainable economic growth.

IX. References

- https://kaggle.com/
- https://www.geeksforgeeks.org/courses/data-science-live/
- https://jupyter.org/
- https://www.rbi.org.in/
- https://www.indiabudget.gov.in/