**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Exploratory Data Analysis on Tax Report Dataset***

Submitted by:

Vatsal Rustagi

Registration No: 12307778

Course Code: INT375

Under the Guidance of

**Ms. Baljinder Kaur**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **Vatsal Rustagi** bearing Registration no **12307778** has completed **INT375** project titled, **“Exploratory Data Analysis on Tax Report Dataset”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Baljinder Kaur**

**Designation of the Supervisor**

**Assistant professor**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 11-04-2025

**DECLARATION**

I, Ekjot Singh student of **B. Tech** under CSE Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 11-04-2025 Signature

Registration No: **12307778**  **Vatsal Rustagi**

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to **Lovely Professional University** for providing me with the opportunity to work on this project titled **"Exploratory Data Analysis on Tax Report Dataset”.**

I would like to thank my respected faculty mentor, **Ms. Baljinder Kaur**, for her valuable guidance, support, and encouragement throughout the course of this project. Her insightful feedback and suggestions have played a crucial role in the successful completion of this work.

I am also thankful to the university's faculty and technical staff for creating a supportive and resourceful learning environment. Their teachings and assistance helped me apply theoretical knowledge to practical implementation.

Working on this Python project has not only enhanced my technical skills but also deepened my understanding of real-world data analysis and its applications.

**Vatsal Rustagi**  
**Reg No.: 12307778**  
**2nd Year, B.Tech (CSE)**  
**Lovely Professional University**

**TABLE OF CONTENT**

1. Introduction

2. Source of dataset

3. EDA process

4. Analysis on dataset (for each analysis)

1. Introduction
2. General Description
3. Specific Requirements, functions and formulas
4. Analysis results
5. Visualization

5. Conclusion

6. Future scope

7. References

**INTRODUCTION**

This report presents an Exploratory Data Analysis (EDA) of GST revenue data collected across various Indian states and union territories. The dataset includes revenue figures for Central GST (CGST), State GST (SGST), Integrated GST (IGST), and CESS over different months and financial years. The analysis aims to uncover patterns in tax collections, identify trends across time and geography, and offer insights into the fiscal performance of states under the GST regime.

Initial exploration focused on understanding the structure of the dataset, verifying data integrity, and assessing the completeness of records. Summary statistics helped to highlight central tendencies and variability in revenue collections, while visual tools such as time series plots, heatmaps, and bar charts allowed for a clearer view of temporal and regional patterns.

Key insights from the analysis reveal substantial variation in tax revenue contributions across states, with certain regions like Maharashtra and Gujarat showing consistently high collections. Seasonal patterns and fiscal year effects are also evident, suggesting potential cyclical behaviour in tax inflows. Moreover, the relationship between different tax components sheds light on the distribution of GST burden between central and state authorities.

This EDA serves as a foundational step in understanding India's GST implementation. It provides valuable insights for policymakers, economists, and administrators to evaluate tax performance, plan fiscal strategies, and identify areas for improvement in compliance and collection.

**Source of Dataset:** [**https://ndap.niti.gov.in/dataset/7102?filter\_id=3608**](https://ndap.niti.gov.in/dataset/7102?filter_id=3608)

**EDA PROCESS**

The Exploratory Data Analysis (EDA) process was conducted using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn. The primary goal was to understand the structure, patterns, and key insights within the GST revenue dataset across Indian states. The following steps outline the EDA process:

**1. Data Understanding**

* Loaded the dataset and reviewed its structure using df.info() and df.head().
* Reviewed column names, data types, and attributes such as state name, financial year, month, and revenue categories (CGST, SGST, IGST, CESS).

**2. Data Cleaning**

* Identified and handled missing values using df.isnull().sum() and replaced them with 0 using fillna().
* Removed duplicate records using df.duplicated().sum() and df.drop\_duplicates().
* Ensured consistency in column names and formatted data for easy analysis.

**3. Summary Statistics**

* Generated descriptive statistics using df.describe() to analyze central tendencies and variability.
* Evaluated the distribution of tax revenues across different GST categories.

**4. Univariate Analysis**

* Plotted bar charts and pie charts to examine revenue contributions of CGST, SGST, IGST, and CESS.
* Identified which tax components consistently generated the highest and lowest collections.

**5. Bivariate and Multivariate Analysis**

* Explored correlation between different tax types using correlation heatmaps.
* Compared monthly and yearly trends using line plots.
* Analysed revenue differences across states to identify top contributors.

**6. Trend Analysis**

* Created time series plots to visualize how GST collections evolved over months and years.
* Applied moving averages to smooth fluctuations and highlight long-term patterns.

**7. Forecasting**

* Conducted basic forecasting using non-ML methods such as linear trend extension and moving averages.
* Estimated future revenue trends based on historical patterns.

**8. Visualization**

* Used Seaborn and Matplotlib to create informative plots including line graphs, bar charts, pie charts, and heatmaps.
* Customized chart aesthetics (colors, titles, labels) for clear and impactful data storytelling.

**ANALYSIS ON DATA**

**Introduction:**

This report explores GST revenue data across Indian states, focusing on CGST, SGST, IGST, and CESS collections over time. The analysis reveals regional and seasonal trends, highlights top-contributing states, and uncovers relationships among tax components. These insights support informed fiscal planning and provide a foundation for deeper economic and policy analysis.

**General Description:**

The dataset contains monthly GST revenue data collected from various Indian states and union territories. It includes figures for four key tax components: Central GST (CGST), State GST (SGST), Integrated GST (IGST), and CESS. Each entry is associated with a specific month and financial year, along with corresponding calendar and coded date formats. The data provides a comprehensive view of tax collections at the state level, enabling analysis of revenue trends, regional comparisons, and tax policy impacts over time.

**Specific Requirements, Functions, and Formulas:**

### **Specific Requirements:**

* **Data Cleaning:** Handle missing values, remove duplicates, and standardize date formats.
* **Data Transformation:** Convert string-based dates to datetime objects, extract year and month, and ensure numeric types for tax columns.
* **Aggregation:** Summarize revenue by month, state, or tax type.
* **Visualization:** Create plots to identify trends, compare state-wise performance, and visualize seasonal patterns.

**Functions Used (Python - Pandas, NumPy, Matplotlib, Seaborn):**

**Loading & Exploring Data:** pd.read\_excel(), df.head(), df.info(), df.describe()

**Cleaning:** df.dropna(), df.duplicated(), df.fillna()

**Type Conversion:** pd.to\_datetime(df['srcMonth'])

df['Revenue'] = df['Revenue'].astype(float)

**Grouping & Aggregation:** df.groupby(['srcStateName'])['CGST'].sum()

df.groupby(['YearCode', 'Month'])[['CGST', 'SGST', 'IGST', 'CESS']].sum()

**Sorting & Filtering:** df.sort\_values(by='CGST', ascending=False)

df[df['YearCode'] == 2019]

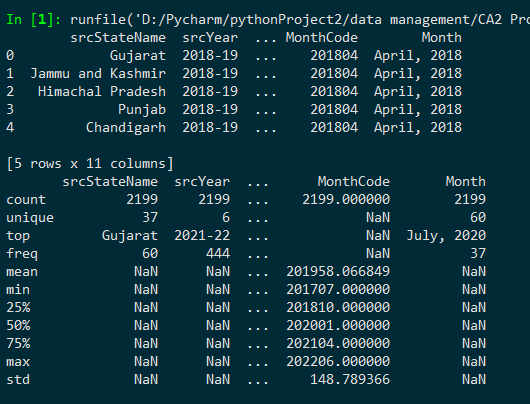
**Visualization (Matplotlib & Seaborn):** sns.barplot(), sns.lineplot(), plt.plot(), plt.xticks(rotation=45), sns.heatmap()

**ANALYSIS RESULTS**

**OBJECTIVES ACHIEVED:**

**Exploratory data analysis**

1. **GST revenue trend over years**
2. **Outlier detection in state-wise or monthly revenue**
3. **Bar plot showing distribution of GST components (CGST, SGST, IGST, CESS)**
4. **Pie chart showing percentage contribution of each GST component.**
5. **Individual GST component trends over time**
6. **Correlation between different GST components**
7. **Top 10 states/UTs with highest total GST collection**
8. **Top 10 months with highest GST revenue**
9. **Future GST revenue prediction using moving average.**
10. **Z-Test on GST revenue before and after 2020**

****

**A screenshot of a computer screen

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**VISUALIZATION**

**A green squares with white numbers

Description automatically generated**

**A group of graphs with numbers

Description automatically generated with medium confidence**

**A graph of a bar graph

Description automatically generated with medium confidence**

**A graph of a graph

Description automatically generated with medium confidence**

**A pie chart with numbers and a number of text

Description automatically generated with medium confidence**

**CONCLUSION**

The Exploratory Data Analysis conducted on the GST revenue dataset has provided valuable insights into the structure, trends, and distribution of tax collections across Indian states and union territories. By examining key components such as Central GST (CGST), State GST (SGST), Integrated GST (IGST), and CESS over various months and financial years, the analysis revealed meaningful patterns that can inform fiscal planning and policy decisions.

The analysis highlighted significant disparities in revenue contributions among states, with some regions consistently leading in collections while others lagged. This suggests varying levels of economic activity, compliance, and administrative efficiency. Temporal analysis further uncovered seasonal trends and year-on-year changes, pointing toward cyclical revenue behaviours likely influenced by policy changes, economic conditions, or external events.

Visualizations and summary statistics helped identify key relationships between different types of taxes and their respective growth trajectories. By calculating totals, averages, and percentage contributions, the analysis created a baseline understanding of how the GST system operates across different levels of governance.

Overall, this EDA serves as a strong foundation for more advanced analysis, such as predictive modelling, policy impact evaluation, or real-time revenue monitoring. It also demonstrates the importance of clean, well-structured data in driving insights that are not only descriptive but actionable. Going forward, integrating this dataset with broader economic indicators, and adopting advanced analytical tools can enhance its usefulness for government bodies, economists, and decision-makers alike.

**Future Scope**

**Predictive Modeling:**

Use machine learning techniques (like time series forecasting or regression models) to predict future GST revenue based on historical trends.

**Policy Impact Analysis:**

Analyze how changes in tax policies or rates affect revenue collection across states and over time.

**Comparative State Analysis:**

Develop dashboards or models to compare state-wise efficiency in tax collection, helping identify high- and low-performing regions.

**Integration with Economic Indicators:**

Combine GST data with other datasets (GDP, inflation, employment) to explore broader economic patterns and correlations.

**Anomaly Detection:**

Implement models to detect unusual spikes or drops in revenue, which could point to compliance issues or external economic factors.

**Interactive Visualizations:**

Build interactive tools or dashboards using Plotly, Tableau, or Power BI to allow stakeholders to explore trends more dynamically.

**Granular Revenue Segmentation:**

Expand the dataset to include sector-wise or industry-wise GST contributions to analyze how different economic sectors contribute to tax revenue.

**References**

Dataset: <https://ndap.niti.gov.in/dataset/7102?filter_id=3608>