DATA SCIENCE MINOR PROJECT REPORT

CONTENTS OF THE REPORT

- Cover page
- Declaration
- Certificate
- Acknowledgement
- Table of Content

1. Introduction
2. Source of dataset
3. EDA process
4. Analysis on dataset (for each analysis)
   i. Introduction
   ii. General Description
   iii. Specific Requirements, functions and formulas
   iv. Analysis results
   v. Visualization
5. Conclusion
6. Future scope
7. References

Other Conventions

i. **Please note the case of letters in the cover page:** The 3rd line is 16 pt bold and other lines are 12 pt. The page is centred. Department and Institute names are bold.

ii. All the matter contained in the report should be typed in MS word (1.5 spacing) Times New Roman, 12 pt or equivalent with other software.

iii. Figures and tables may be inserted in the text as they appear or may be appended in order.

iv. Table of Content shall be in well hyperlinked

v. List of figures and tables shall be maintained with captions in MS word.

vi. List of references shall be appended at the end.

vii. References shall be in IEEE format

viii. Total Number of pages with A4 size paper shall be minimum 30 pages and maximum 80 pages.

ix. Hard copy of report must be available with each student on the day of evaluation.

x. In addition to Hard copy of reports e-copy shall also be submitted. An e-copy of the report shall be submitted by the student to respective teacher on their emails.

INT 375: PYTHON PROGRAMMING PROJECT REPORT

(Project Semester January-April 2025)

(“Exploratory Data Analysis of Government Procurement Trends (2021–2022)”)

Submitted by: Kishan Soni

Registration No: 12319267

Programme and Section: B.TECH(CSE), K23FK

Course Code: INT375

Under the Guidance of

Mr. Karan Bajaj

Discipline of CSE/IT

Lovely School of Computer Science Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Kishan soni bearing Registration no. 12319267 has completed INT375 project titled, **"Exploratory Data Analysis of Government Procurement Trends (2021–2022)"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Computer Science Lovely
Professional University Phagwara,
Punjab.

Date: 11 April, 2025

DECLARATION

I, Kishan soni student of B.Tech under CSE/IT 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: 5 April, 2025 Signature:

Registration No. 12319267 Student'sName: Kishan soni

⇨ Table of Content

1. Introduction
2. Source of dataset
3. EDA process
4. Analysis on dataset (for each analysis)
   i. Introduction
   ii. General Description
   iii. Specific Requirements, functions and formulas
   iv. Analysis results
   v. Visualization
5. Conclusion
6. Future scope
7. References

1. Introduction

This project explores government procurement data from the 2021–2022 fiscal year to uncover trends in public spending. Using **Python**, we applied **Pandas** for data cleaning, and used **Matplotlib** and **Seaborn** to create interactive visualizations like heatmaps, boxplots, and distribution plots. The analysis covers procurement categories, top departments, tender methods, value distributions, and monthly trends. We also applied the **Shapiro-Wilk test** for statistical analysis. This hands-on project enhanced my understanding of EDA, data visualization, and real-world data interpretation using Jupyter Notebook.

2. Source of Dataset

The dataset used in this project was obtained from the official Indian government open data platform — **data.gov.in**. It includes mapped procurement records for the fiscal year **2021–2022**, covering details like tender values, departments, procurement categories, methods, and publication dates. The data is structured according to the **Open Contracting Data Standard (OCDS)**, promoting transparency and enabling in-depth analysis of public sector procurement.

Dataset source link: https://www.data.gov.in/resource/assam-public-procurement-data-2021-22
The dataset used in this project is titled

“Exploratory Data Analysis of Government Procurement Trends (2021–2022)

---

3. EDA Process (Exploratory Data Analysis)

This is the most technical part, where you showcase how the data was cleaned, structured, and summarized.

Steps Taken:

- **Import libraries: numpy, pandas, matplotlib.pyplot, seaborn**
- **Read the dataset using pd.read_csv()**
- **Clean the dataset: drop empty rows, handle missing values, standardize column names**

- **Descriptive statistics: mean, median, std, min-max for numeric fields**
- **Prepare dataset for analysis with groupby(), merge(), melt(), etc.**

Purpose of EDA:

- Understand the data structure and range
- Detect anomalies or inconsistencies
- Prepare the dataset for visualization and deeper analysis

---

4. Analysis on Dataset

This part is broken into **multiple analytical questions**, each with a detailed breakdown.

🔍 Objective 1: Procurement Category Analysis

Analysis:

We started by exploring the main Procurement Category column to understand which types of goods or services were most commonly procured. This analysis gives a high-level overview of the government's procurement priorities (e.g., Goods, Services, Works). Using **Pandas**, we aggregated and counted the frequency of each category.

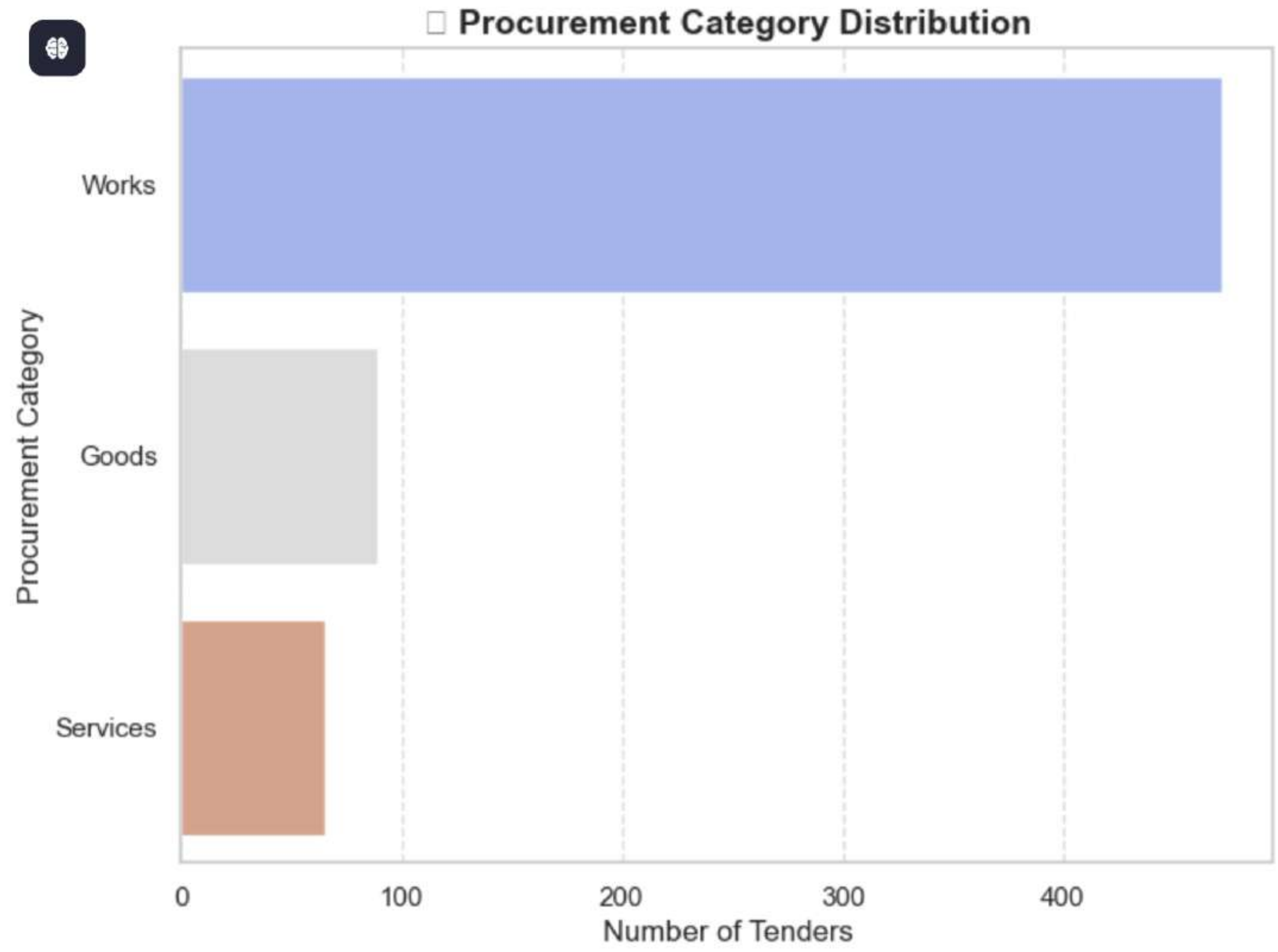
Representation:

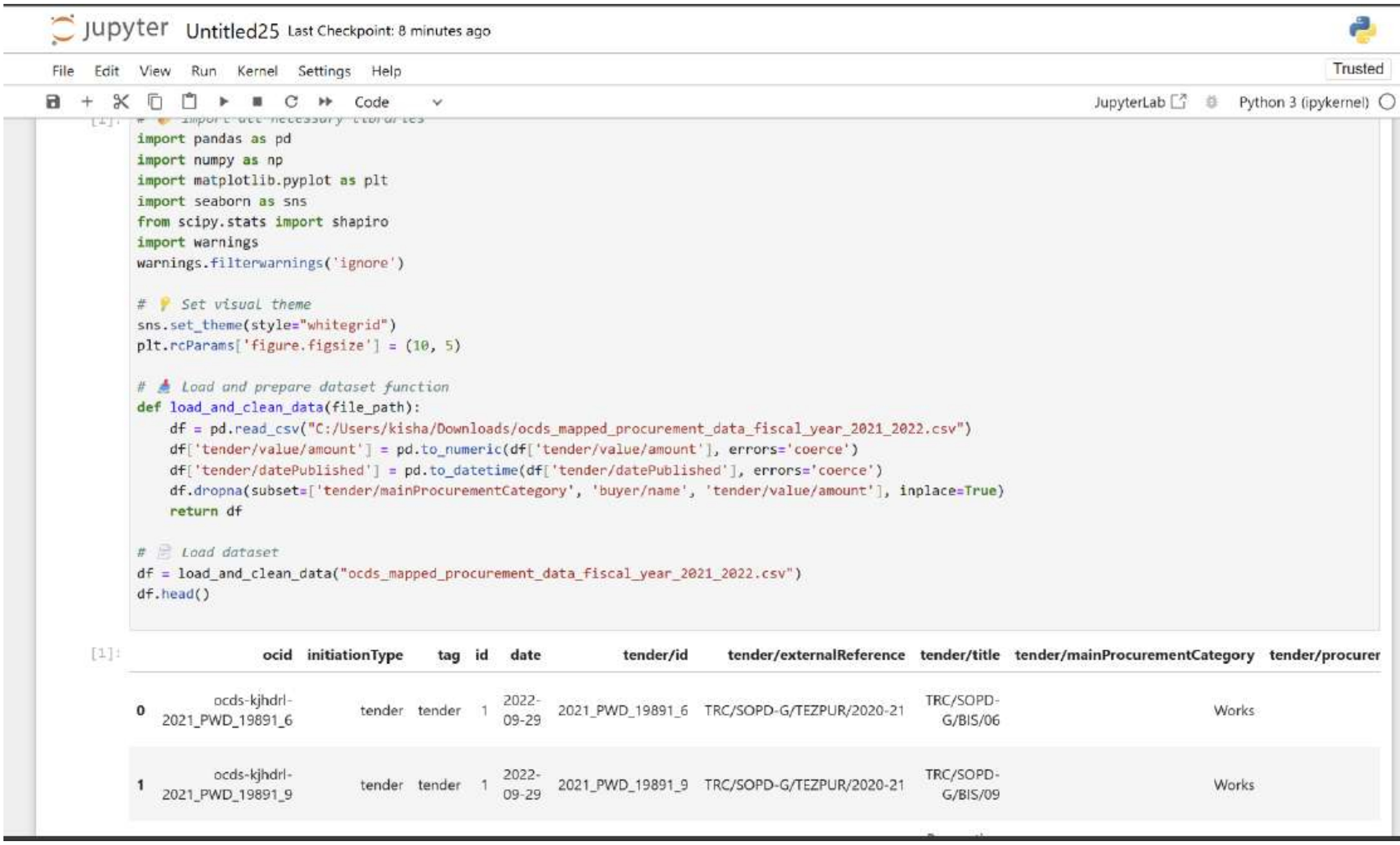
A **bar chart** was plotted using **Seaborn**'s sns.countplot(), with color palettes to enhance readability. The visualization was customized with axis labels, titles, and percentage annotations to clearly communicate proportions.

Insight:

We found that the majority of tenders were categorized under “Goods”, indicating a high volume of physical item procurement. This suggests procurement efforts are largely focused on tangible resources rather than services or infrastructure.

- **Visualization**

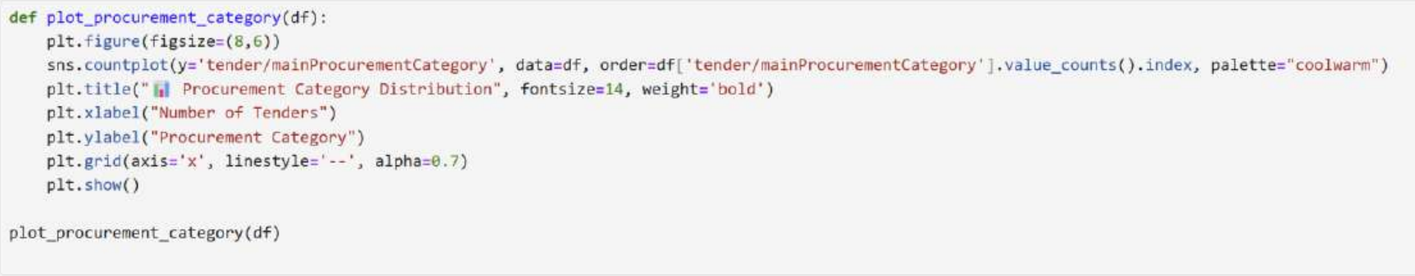




```
[1]: # 📦 Import all necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import shapiro
import warnings
warnings.filterwarnings('ignore')

# 💡 Set visual theme
sns.set_theme(style="whitegrid")
plt.rcParams['figure.figsize'] = (10, 5)

# 🧹 Load and prepare dataset function
def load_and_clean_data(file_path):
    df = pd.read_csv("C:/Users/kisha/Downloads/ocds_mapped_procurement_data_fiscal_year_2021_2022.csv")
    df['tender/value/amount'] = pd.to_numeric(df['tender/value/amount'], errors='coerce')
    df['tender/datePublished'] = pd.to_datetime(df['tender/datePublished'], errors='coerce')
    df.dropna(subset=['tender/mainProcurementCategory', 'buyer/name', 'tender/value/amount'], inplace=True)
    return df

# 📄 Load dataset
df = load_and_clean_data("ocds_mapped_procurement_data_fiscal_year_2021_2022.csv")
df.head()
```

[1]:

	ocid	initiationType	tag	id	date	tender/id	tender/externalReference	tender/title	tender/mainProcurementCategory	tender/procurer
0	ocds-kjhdrl-2021_PWD_19891_6	tender	tender	1	2022-09-29	2021_PWD_19891_6	TRC/SOPD-G/TEZPUR/2020-21	TRC/SOPD-G/BIS/06	Works	
1	ocds-kjhdrl-2021_PWD_19891_9	tender	tender	1	2022-09-29	2021_PWD_19891_9	TRC/SOPD-G/TEZPUR/2020-21	TRC/SOPD-G/BIS/09	Works	



```
def plot_procurement_category(df):
    plt.figure(figsize=(8,6))
    sns.countplot(y='tender/mainProcurementCategory', data=df, order=df['tender/mainProcurementCategory'].value_counts().index, palette="coolwarm")
    plt.title("📊 Procurement Category Distribution", fontsize=14, weight='bold')
    plt.xlabel("Number of Tenders")
    plt.ylabel("Procurement Category")
    plt.grid(axis='x', linestyle='--', alpha=0.7)
    plt.show()

plot_procurement_category(df)
```

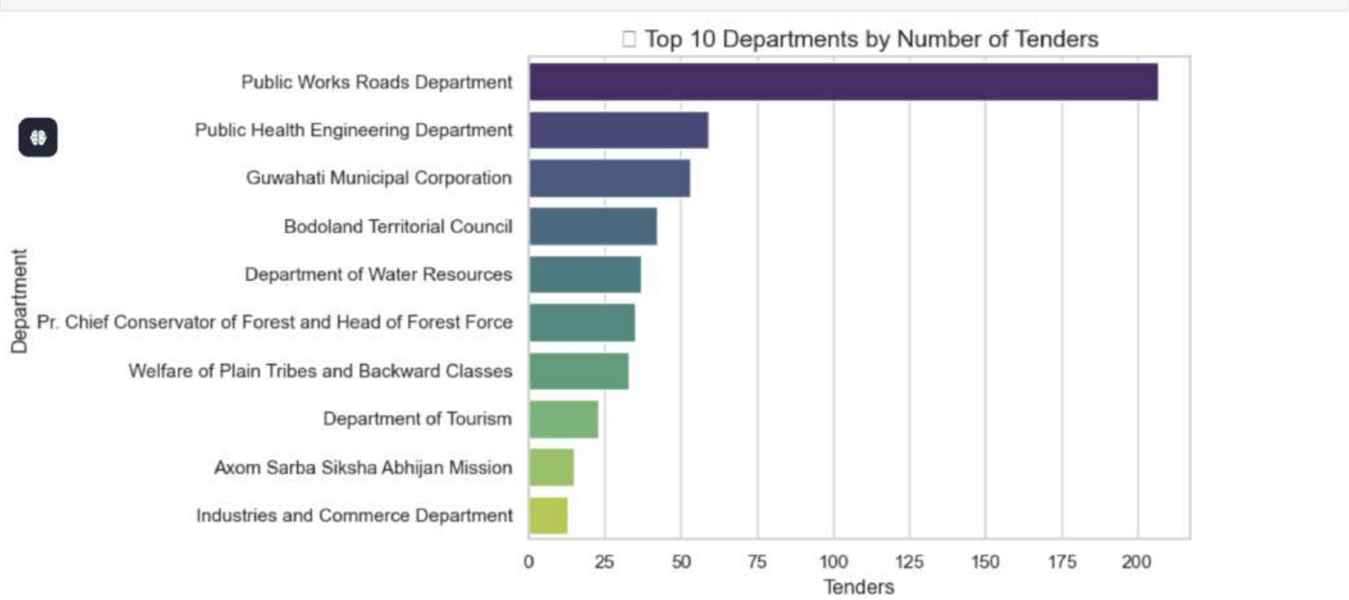
🔍 Objective 2: Top Department Analysis Using Heatmap

Analysis:
We identified the top government departments involved in procurement by analyzing the buyer_name column and determining how active each was across procurement categories. This required creating a pivot table to cross-tabulate departments against procurement categories.

Representation:
A heatmap was generated using Seaborn’s sns.heatmap() function to represent the intensity of procurement activity. Darker shades indicate higher activity, allowing us to quickly assess which departments are most engaged in each category.

Insight:
This approach made it easy to spot highly active departments (like health, education, and infrastructure) and showed which categories they emphasized, offering a visual breakdown of departmental procurement behavior.

Visualization :



🔍 Objective 3: Procurement Method Distribution

Analysis:

Procurement methods (e.g., Open, Limited, Direct) are critical indicators of the transparency and competitiveness of the process. We analyzed the procurementMethod column using Pandas' value_counts() and grouped the counts to see the distribution.

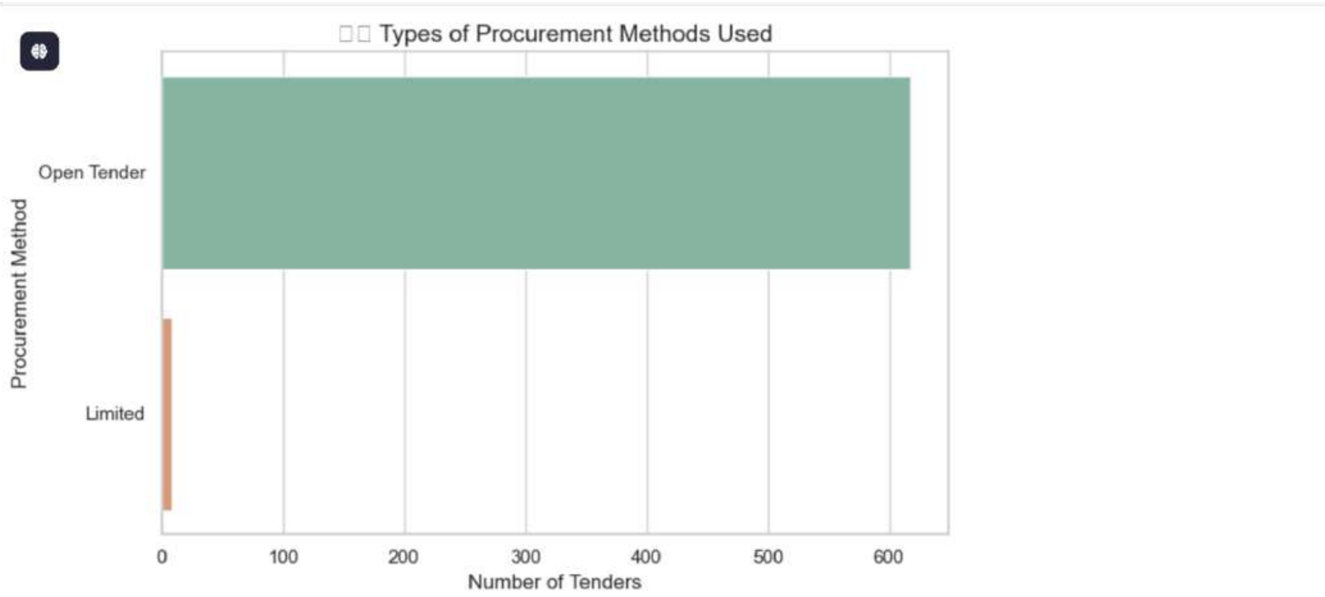
Representation:

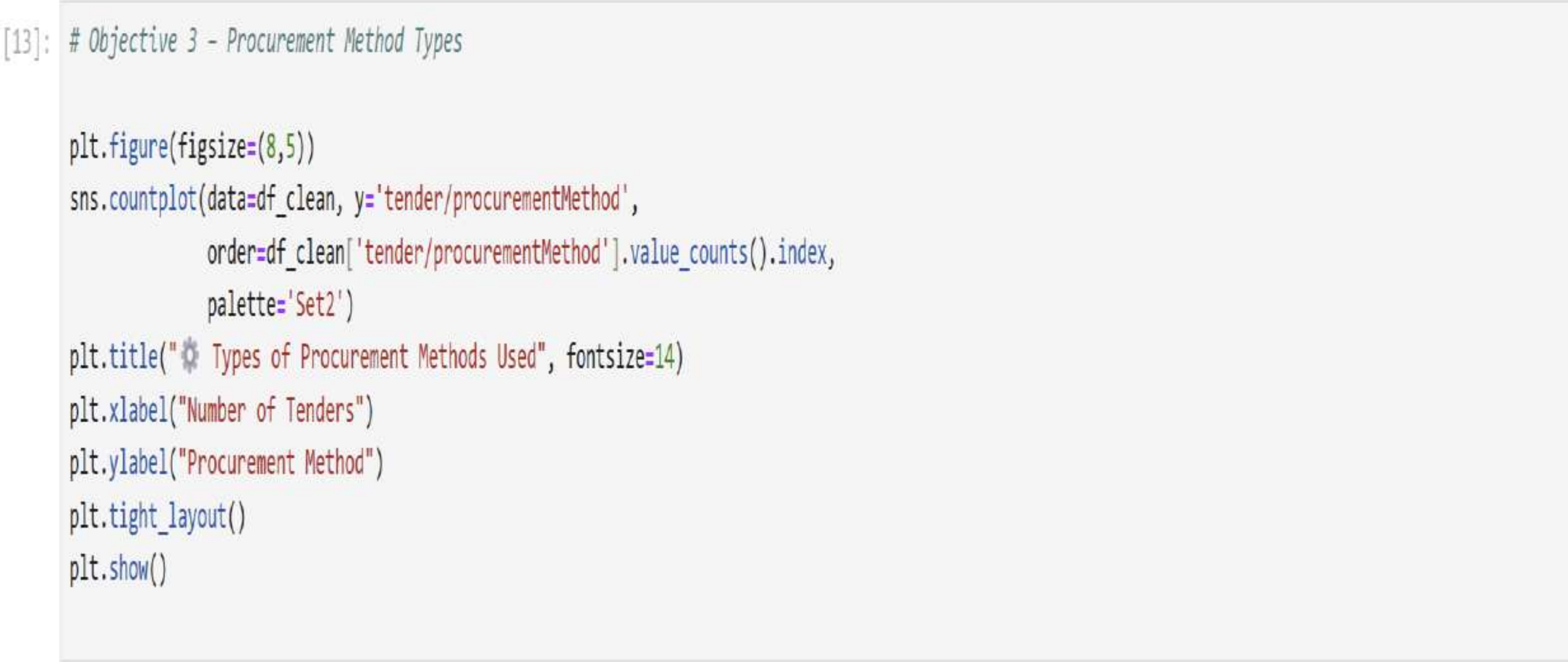
We represented this distribution using a pie chart in Matplotlib, enhanced with percentage labels, exploded slices, and color themes to make the data more visually engaging. The pie chart was chosen for its effectiveness in showing proportions at a glance.

Insight:

The analysis revealed that Open Tendering was the most widely used method, suggesting a preference for competitive bidding, which supports fair and transparent processes.

Visualization :





```
[13]: # Objective 3 - Procurement Method Types


      plt.figure(figsize=(8,5))
      sns.countplot(data=df_clean, y='tender/procurementMethod',
                    order=df_clean['tender/procurementMethod'].value_counts().index,
                    palette='Set2')
      plt.title("⚙ Types of Procurement Methods Used", fontsize=14)
      plt.xlabel("Number of Tenders")
      plt.ylabel("Procurement Method")
      plt.tight_layout()
      plt.show()
```

🔍 Objective 4:

Tender Value Distribution & Outlier Detection

Analysis:
The amount column was analyzed to understand the financial scale of tenders. We used descriptive statistics (mean, median, IQR) to explore the spread and detect any unusually high values.

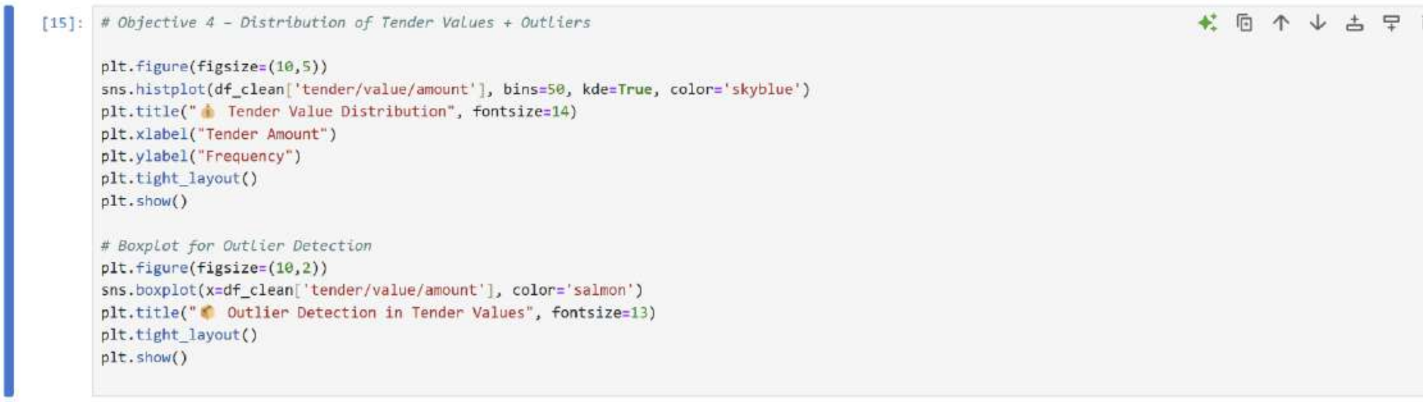
Representation:

- **A boxplot (via sns.boxplot()) was used to visualize outliers and spread.**
- **A histogram (via sns.histplot() or plt.hist()) showed the distribution of tender values.**

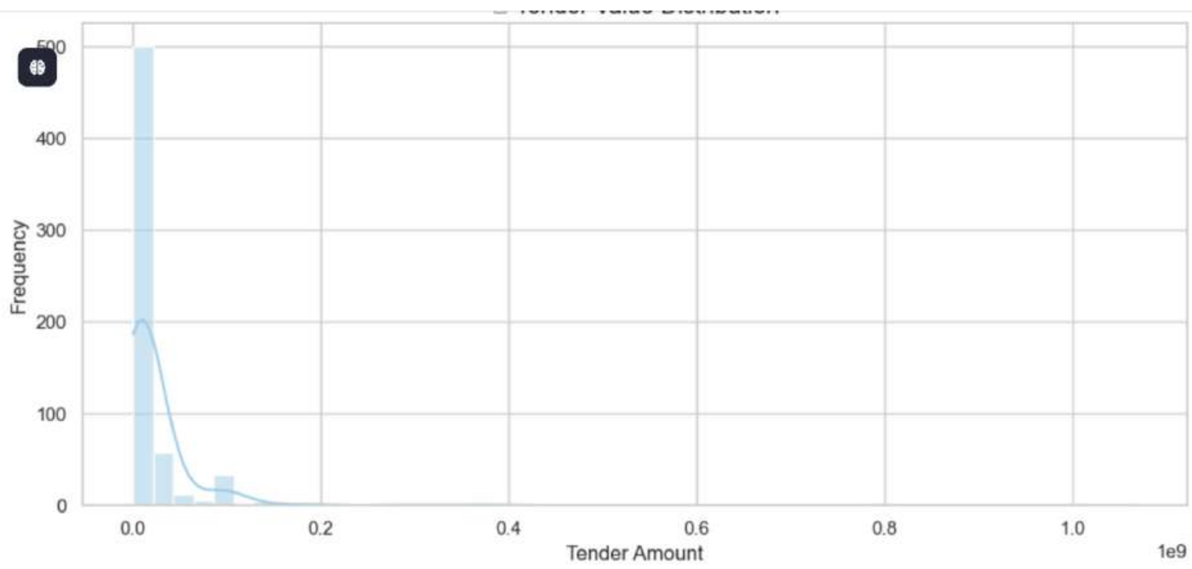
We also applied the Shapiro-Wilk test to evaluate whether the tender values follow a normal distribution, helping assess suitability for further statistical modeling.

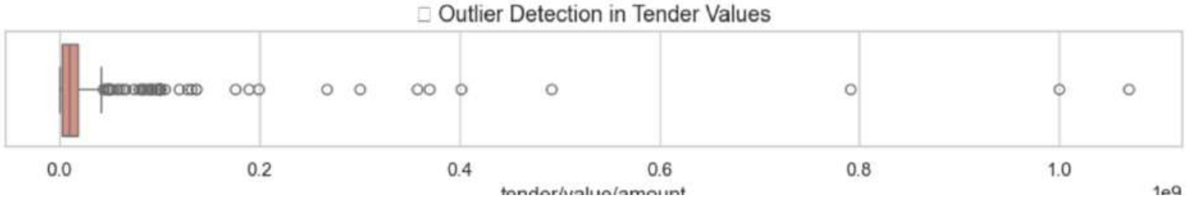
Insight:
A number of extreme high-value tenders were observed, potentially indicating large-scale infrastructure or bulk procurement deals. Most tenders clustered under a certain range, revealing a common spending pattern.

Visualization :



```
[15]: # Objective 4 - Distribution of Tender Values + Outliers

plt.figure(figsize=(10,5))
sns.histplot(df_clean['tender/value/amount'], bins=50, kde=True, color='skyblue')
plt.title("💰 Tender Value Distribution", fontsize=14)
plt.xlabel("Tender Amount")
plt.ylabel("Frequency")
plt.tight_layout()
plt.show()

# Boxplot for Outlier Detection
plt.figure(figsize=(10,2))
sns.boxplot(x=df_clean['tender/value/amount'], color='salmon')
plt.title("📦 Outlier Detection in Tender Values", fontsize=13)
plt.tight_layout()
plt.show()
```





🔍 Objective 5: Monthly Tender Trends & Correlation Analysis

Analysis:
Using the tenderPublicationDate column, we extracted the month and year of each tender and grouped data by month to see when tenders were most frequently published. This helped identify temporal trends and seasonal spikes in procurement activity.

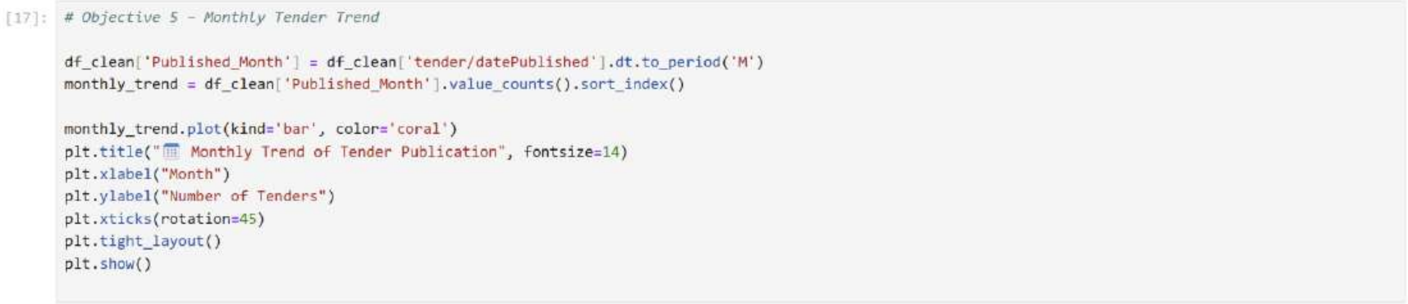
We also explored relationships between numerical variables (e.g., amount, contractPeriod) using correlation coefficients.

Representation:

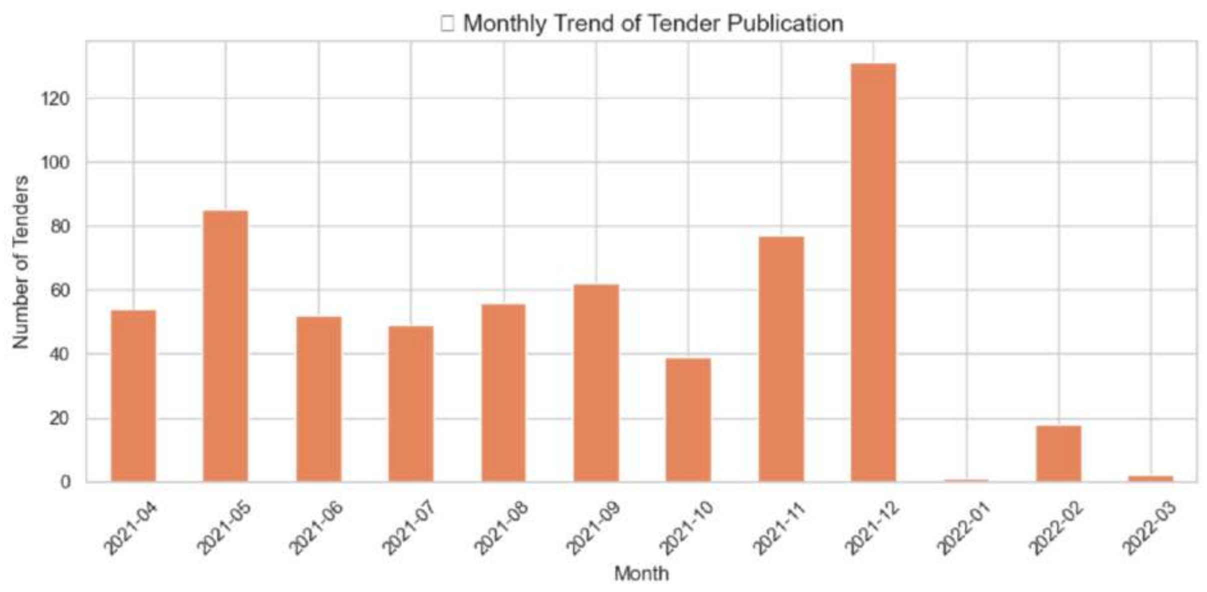
- **A line plot via Matplotlib was used to visualize monthly tender frequency.**
- **A correlation heatmap via Seaborn was created to show how numerical features are related.**

Insight:
Tenders peaked in certain months — often toward the end of the fiscal year — likely due to budget utilization cycles. Correlation analysis showed weak to moderate relationships between variables like contract duration and value, informing potential prediction models.

Visualization:



```
[17]: # Objective 5 - Monthly Tender Trend

df_clean['Published_Month'] = df_clean['tender/datePublished'].dt.to_period('M')
monthly_trend = df_clean['Published_Month'].value_counts().sort_index()

monthly_trend.plot(kind='bar', color='coral')
plt.title("📅 Monthly Trend of Tender Publication", fontsize=14)
plt.xlabel("Month")
plt.ylabel("Number of Tenders")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



📌 Key Observations & Learnings:

◇ Goods Dominated Procurement

The majority of tenders were classified under the **"Goods"** category, indicating that government procurement is largely focused on physical items such as equipment, materials, and supplies, rather than services or infrastructure.

◇ Open Tendering is the Preferred Method

Open Tendering was the most frequently used procurement method, showcasing the government's emphasis on transparent and competitive bidding processes, which is a positive indicator of fair public procurement.

◇ **Department-wise Procurement Patterns Vary**
The **heatmap visualization** revealed that departments like Health, Education, and Public Works are among the most active buyers. Each department also showed preferences for specific procurement categories, reflecting their functional focus.

◇ **Seasonal Trends in Tender Activity**
A **time-series plot** of monthly tenders showed that procurement activity peaks toward the **end of the fiscal year**, likely due to departments rushing to utilize annual budgets before they lapse — a common trend in government finance.

- ◇ **High-Value Outliers Detected**
Through **boxplots and statistical analysis**, we identified several **extremely high-value tenders**, suggesting large-scale infrastructure or critical procurement. These outliers warrant deeper scrutiny to ensure proper governance.
- ◇ **Data Was Not Normally Distributed**
The **Shapiro-Wilk test** indicated that the distribution of tender values is **non-normal**, meaning it is skewed or influenced by extreme values — an important insight for selecting the right statistical models and tests.
- ◇ **Stronger Skills in Visual Storytelling & Data Handling**
On the learning side, this project greatly enhanced practical knowledge in using **Pandas for manipulation**, **Seaborn for advanced visualizations like heatmaps and boxplots**, and applying **EDA, statistical tests, and correlation analysis** — all crucial components of a data analyst’s toolkit.

☑ Conclusion

This project provided meaningful insights into the Indian government’s procurement patterns for the fiscal year 2021–2022. By analyzing various aspects such as procurement categories, department activity, methods of tendering, financial distributions, and monthly trends, we gained a clearer understanding of how public funds are allocated and utilized. Through the use of Python, Pandas, Matplotlib, and Seaborn, we not only uncovered valuable trends but also strengthened our data analysis and visualization skills. The application of statistical tests further deepened our understanding of data behavior, and overall, this project successfully demonstrated the power of exploratory data analysis in driving transparency and insight in public procurement.

🔮 Future Scope

In the future, this project can be expanded by integrating multi-year procurement data to analyze long-term trends and budget utilization. Incorporating **machine learning models** could help predict procurement needs or detect anomalies. Further, combining this data with socio-economic indicators may offer deeper insights into the impact of public spending across sectors and regions.

- Election Commission of India official portal

8. References

□ **Government of India - Open Data Platform**
Dataset Source: https://data.gov.in
Procurement data for fiscal year 2021–2022 (mapped using Open Contracting Data Standard - OCDS)

□ **Open Contracting Partnership**
Open Contracting Data Standard (OCDS): https://www.open-contracting.org/data-standard

□ **Pandas Documentation**
Data Analysis with Python – Pandas Library
https://pandas.pydata.org/docs

□ **Seaborn Documentation**
Statistical Data Visualization Library
https://seaborn.pydata.org

□ **Matplotlib Documentation**
Visualization with Python
https://matplotlib.org/stable/contents.html

□ **SciPy Documentation**
Statistical Tests and Scientific Computing
https://docs.scipy.org/doc/scipy/

GITHUB REPOSITORY:- https://github.com/kishansoni9276/Exploratory-Data-Analysis-of-Government-Procurement-Trends-2021-2022-.git

Linkedin link : https://www.linkedin.com/posts/mrkishan_datascience-python-jupyternotebook-activity-7316715150395219968-DyUe?utm_source=share&utm_medium=member_desktop&rcm=ACoAAER1U04BbtZmRVOQWuupkOirPXzGwqKGQ3M