PhonePe Pulse Data Analysis and Visualization Report

Name: Pavan Kumar Dirisala
Project Title: PhonePe Pulse Data Analysis and Visualization
Database: MySQL
Data Sources: PhonePe Pulse JSON Data
Frontend Tool: Streamlit

---

1. Problem Statement

With the surge in digital payment adoption, platforms like PhonePe generate massive volumes of transactional and user engagement data. Understanding and utilizing this data is essential to drive growth, improve customer experience, and stay competitive. This project focuses on extracting, transforming, analyzing, and visualizing PhonePe's publicly available transaction data from its Pulse GitHub repository. The goal is to derive actionable insights related to user behavior, transaction trends, and digital insurance adoption across different regions and quarters in India.

---

2. Business Objective

This project aims to support strategic business decisions using data-driven insights by:

- Identifying top-performing regions based on transaction volume and value.
- Analyzing insurance penetration across states.
- Understanding device usage patterns for better app optimization.
- Mapping user registrations to evaluate adoption rates.
- Creating a Streamlit dashboard for interactive, real-time data exploration.
- Recommending region-specific strategies for marketing, expansion, and product development.

---

3. Project Summary

The PhonePe Pulse Data Analysis and Visualization project integrates **data engineering, analytics, and visualization**into a streamlined workflow that supports informed decision-making.

✓ **Key Stages**

A. Data Extraction and Transformation

- Downloaded and parsed JSON files from PhonePe Pulse GitHub repository.
- Transformed nested JSON structures into tabular formats using Python.
- Created CSVs such as `aggregated_transaction.csv`, `aggregated_insurance.csv`, and `map_user.csv` etc..

Example Code Snippet:

```
import os, json
import pandas as pd

path = 'data/aggregated/transaction/country/india/state/'
data = []
for state in os.listdir(path):
    for year in os.listdir(os.path.join(path, state)):
        for file in os.listdir(os.path.join(path, state, year)):
            with open(os.path.join(path, state, year, file)) as f:
                json_data = json.load(f)
            # Extract and process data here...
df = pd.DataFrame(data)
df.to_csv("aggregated_transaction.csv", index=False)
```

B. Data Storage

- Loaded CSV files into a **MySQL** database using **SQLAlchemy**.

Loading CSV to MySQL:

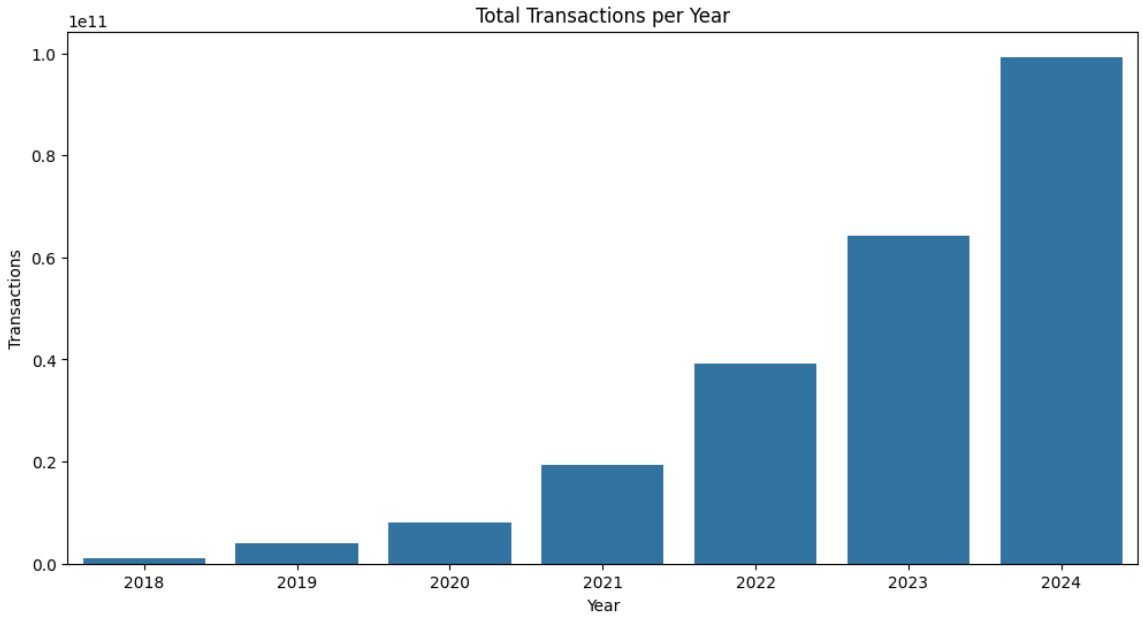
```
from sqlalchemy import create_engine
import urllib.parse
import pandas as pd

password = urllib.parse.quote("Pavan@123")
engine =
create_engine(f"mysql+mysqlconnector://root:{password}@localhost/phonep
e_db")
df = pd.read_csv("aggregated_transaction.csv")
df.to_sql("aggregated_transaction", con=engine, if_exists='replace',
index=False)
```

---

4. Data Analysis and Insights

Using SQL + Python (Pandas, Seaborn, Matplotlib), the following key insights were uncovered:

A. Total Transactions per Year



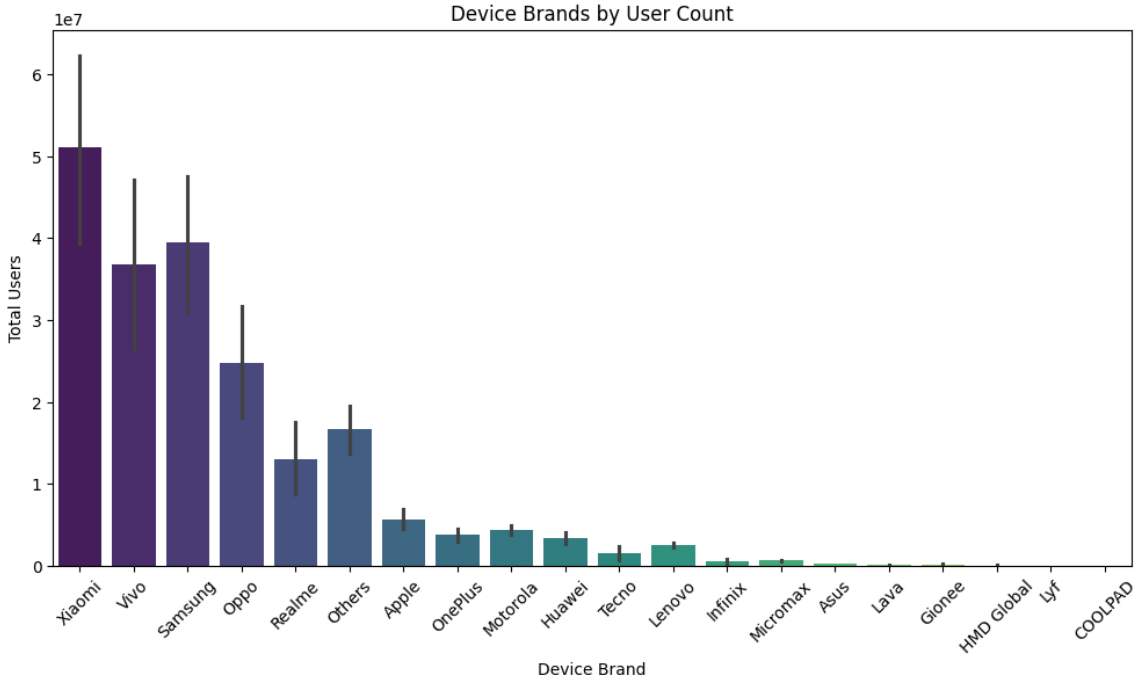
📊 **Insight**:

- From **2018 to 2024**, total transactions have seen **consistent exponential growth** year-over-year.
- The volume has increased from below ₹10 billion in 2018 to nearly ₹100 billion in 2024.
- This indicates **rapid digital adoption**, user trust, and wider availability of digital payment infrastructure across the country.

✅ **Implications**:

- The company must **scale its backend infrastructure**, cloud resources, and fraud detection capabilities.
- Opportunity to **launch new financial services** (e.g., micro-loans, wealth management).
- **Regulatory compliance and cybersecurity** should be prioritized to maintain trust.

B. Device Brands by User Count



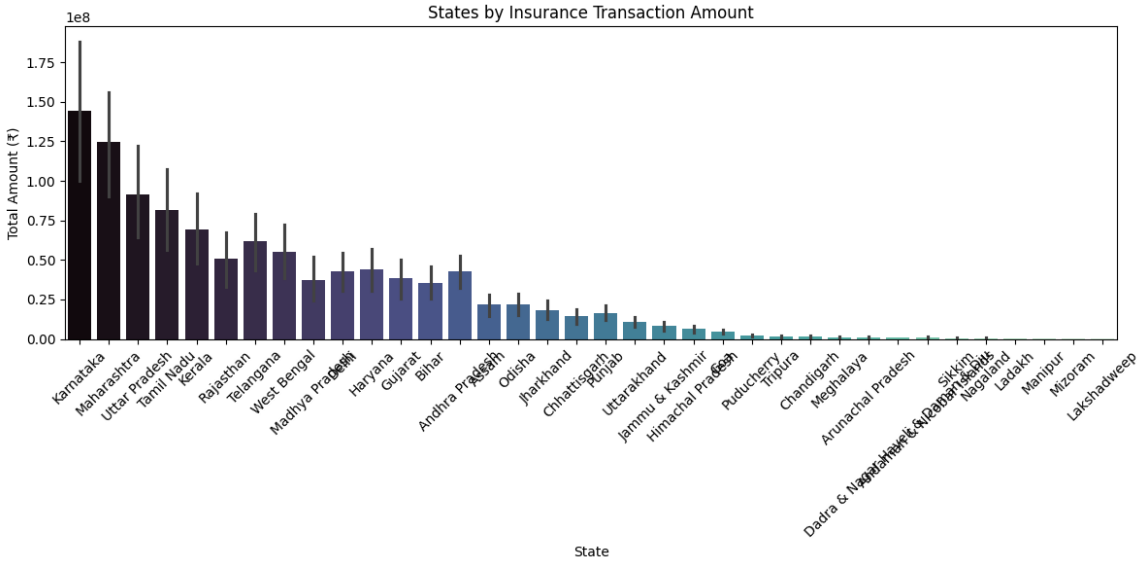
📱 **Insight**:

- **Xiaomi**, **Vivo**, and **Samsung** dominate in terms of user base, reflecting a strong presence of **mid-range Android phones**.
- Brands like **OPPO**, **Realme**, and **Apple** follow.

✅ **Implications**:

- Focus on optimizing app performance for **top Android brands**.
- Create **targeted promotions or onboarding tutorials** for common device brands.
- Develop **premium financial tools** for Apple users.

C. States by Insurance Transaction Amount



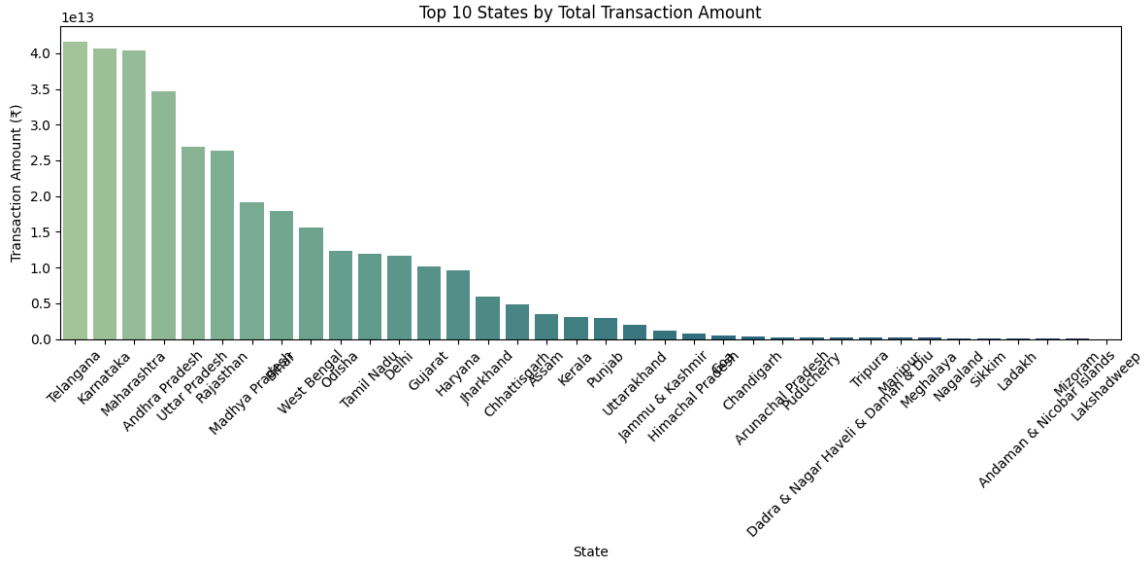
🏥 **Insight**:

- States like **Karnataka**, **Maharashtra**, and **Uttar Pradesh** are leading in digital insurance transactions.
- Urban and rural states both show adoption.

✅ **Implications**:

- Push **localized insurance campaigns** in top-performing regions.
- Offer **vernacular language support** for rural users.
- Collaborate with **regional insurers** for better outreach.

D. States by Total Transaction Amount



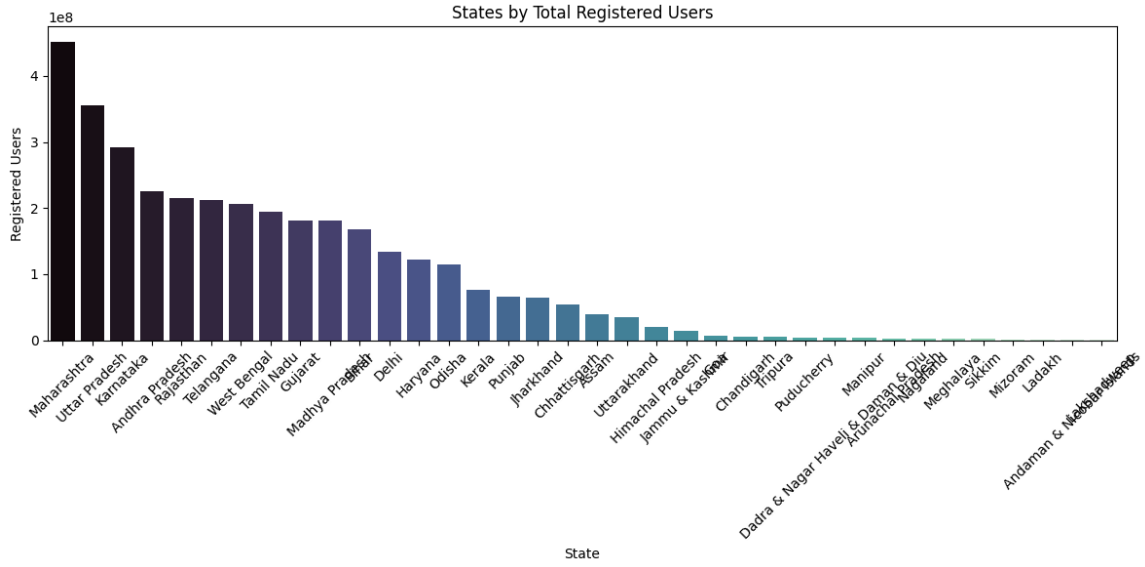
💸 Insight:

- **Telangana**, **Karnataka**, and **Maharashtra** contribute the most in transaction volume.
- These states are tech hubs and indicate high digital penetration.

✅ Implications:

- Launch **loyalty or cashback programs** in these regions.
- Build stronger **merchant networks and app adoption drives** in Tier-2 cities.

E. States by Total Registered Users



👥 Insight:

- **Maharashtra**, **Uttar Pradesh**, and **Karnataka** lead in total user registrations.
- App awareness is high in urban and semi-urban India.

✅ Implications:

- Leverage this base to roll out **beta features** or **personal finance tools**.
- Run **retargeting campaigns** for inactive users in these states.

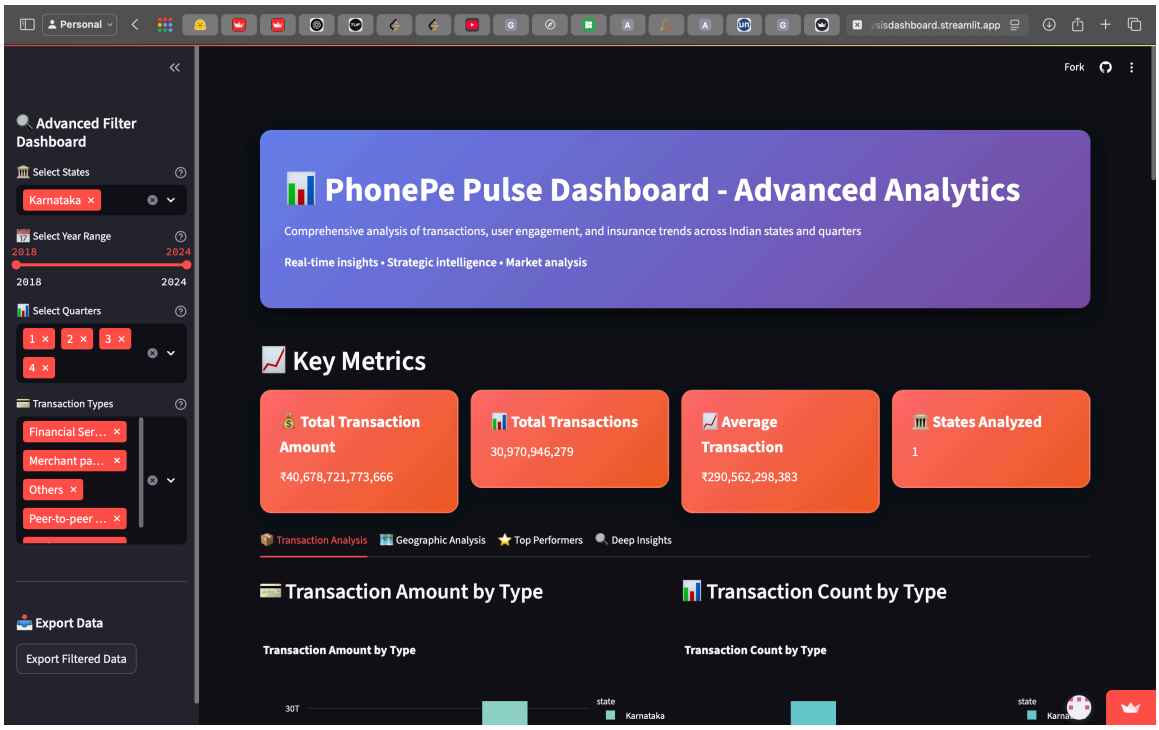
---

5. Streamlit Dashboard

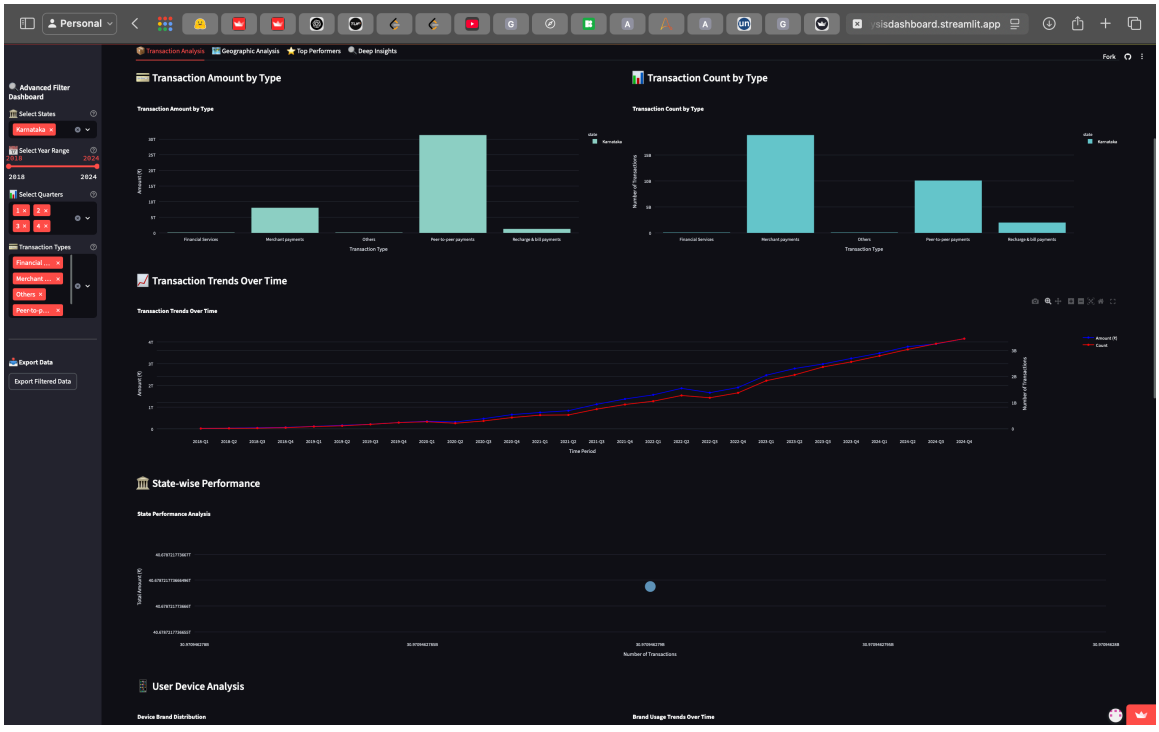
A dynamic dashboard was built with:

- **Filters:** State, Year, Quarter , Transaction Type
- **Visuals:** Bar charts, pie charts using Plotly & Seaborn
- **Interactivity:** Download CSVs or JSONs of filtered results
- **Deployment:** Streamlit for seamless web access

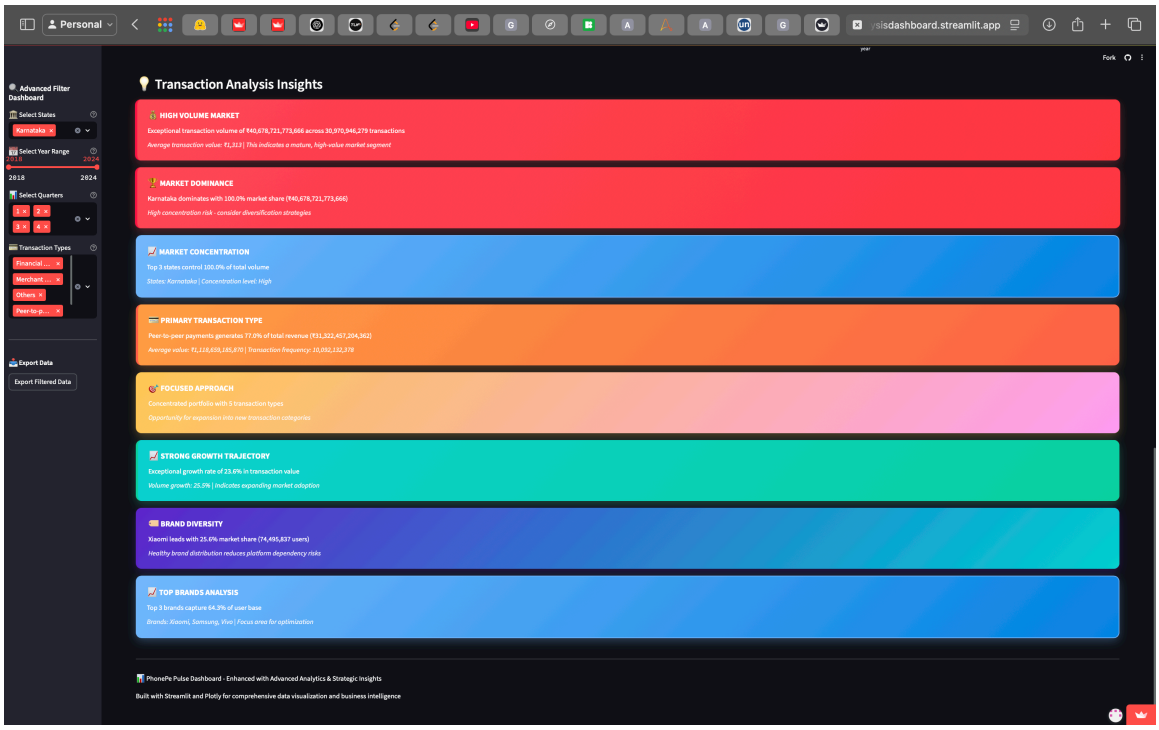
Dashboard Details:



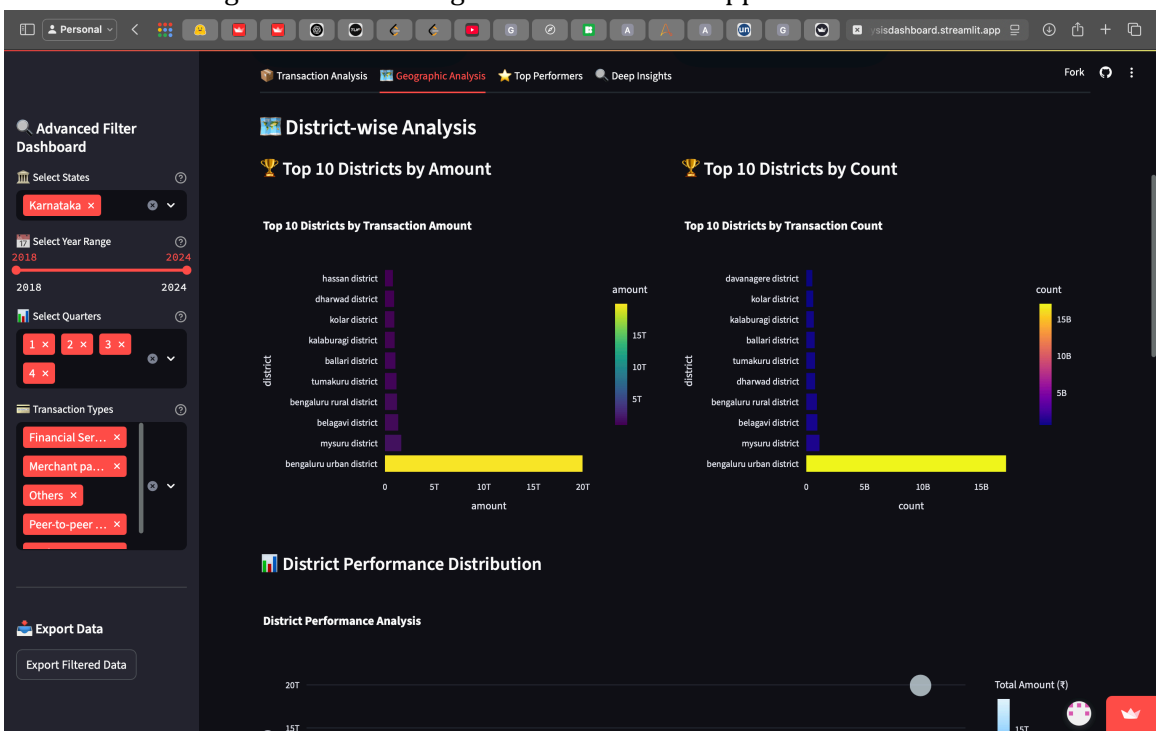
This screenshot is the home page of the Dashboard



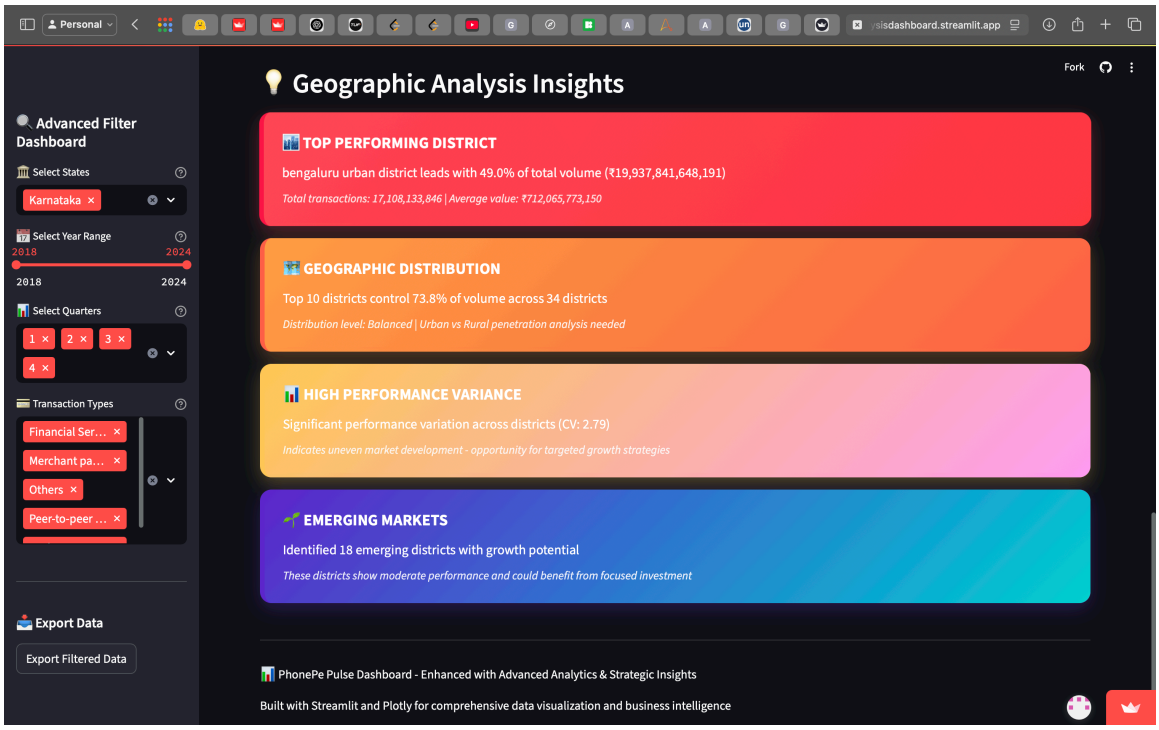
This is the aggregated files analysis



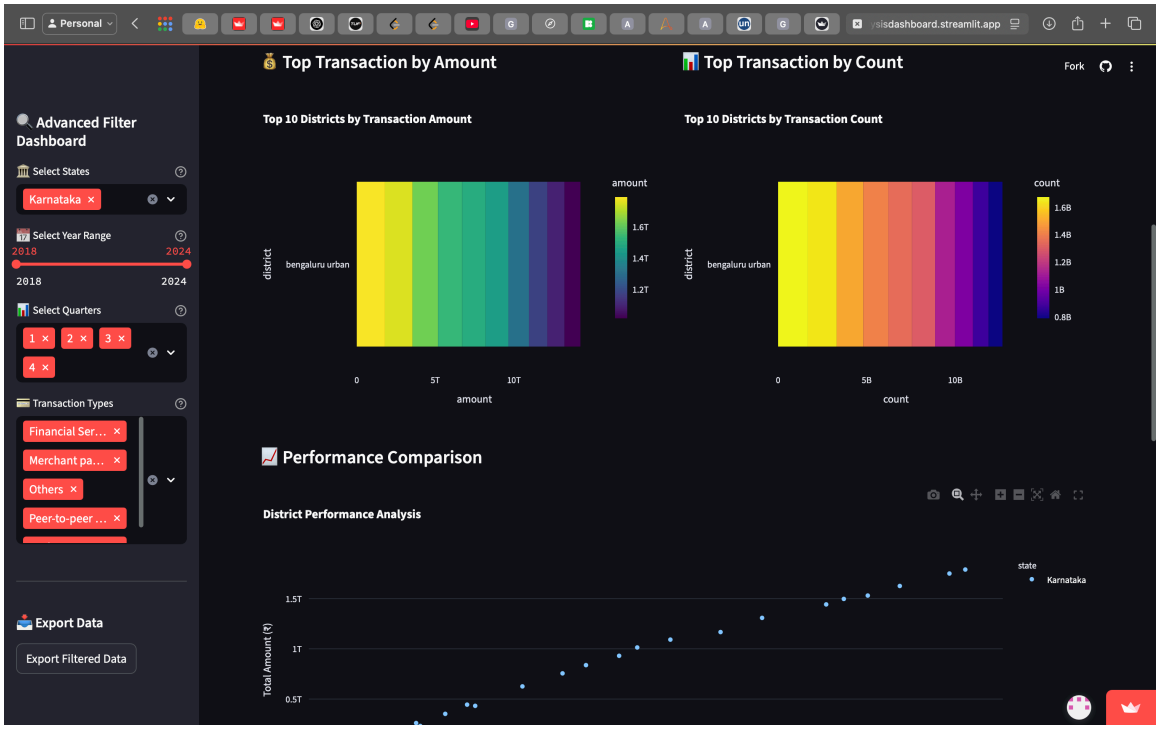
These are the insights that we have got for the filters we applied.

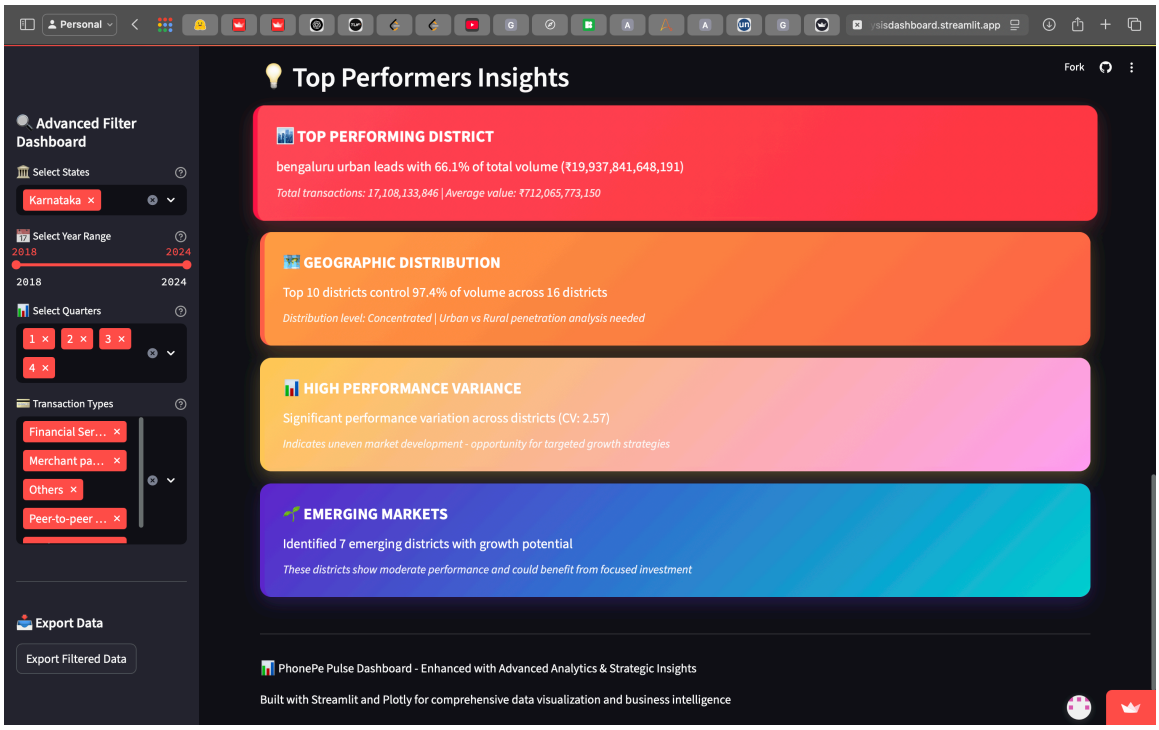


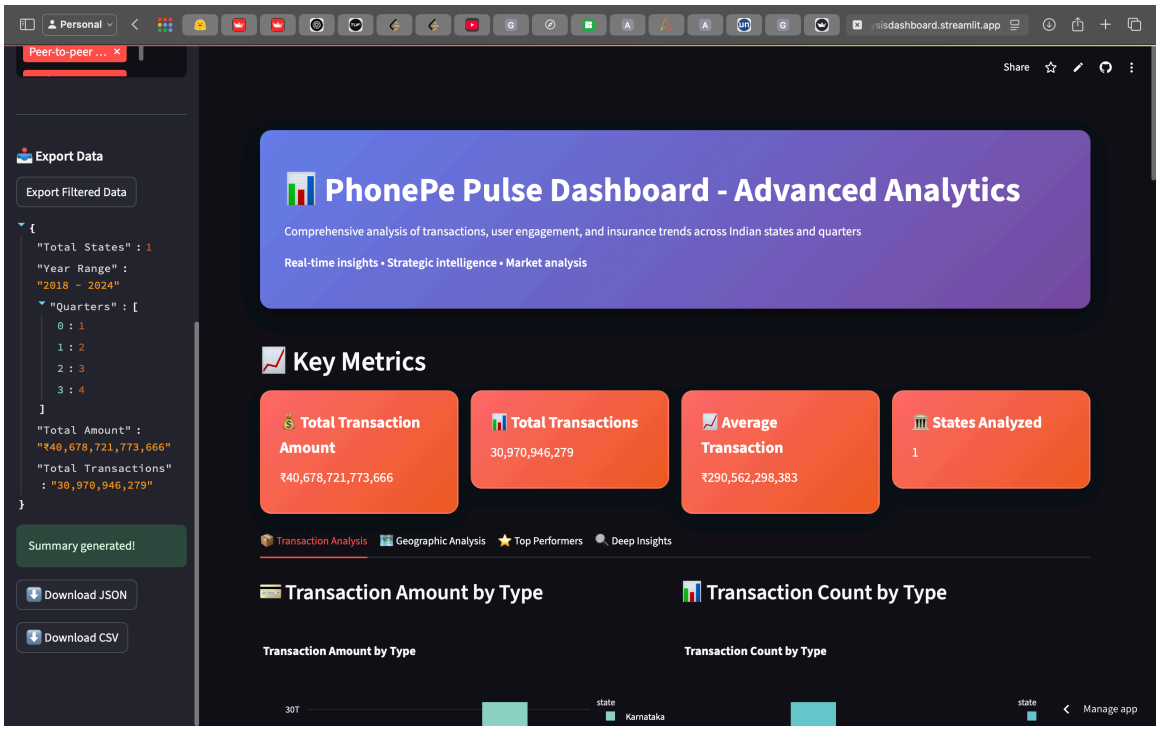
This is the map files analysis



These are the insights after analysing the map data to our filters.







---

6. Technologies Used

- **Language:** Python
- **Database:** MySQL

- **Libraries:** Pandas, Seaborn, Matplotlib, SQLAlchemy, Streamlit
- **Tools:** GitHub, Streamlit, MySQL Workbench

---

7. Project Deliverables

- Python scripts for extraction and transformation
- SQL schema and analytical queries
- CSV datasets and MySQL tables
- Streamlit dashboard codebase
- Final documented report (this)

---

8. Future Enhancements

- Add **geospatial maps** using Folium
- Include **district-level and PIN-code trends**
- Integrate **ML for predictive analytics**
- Deploy via **Railway, Hugging Face Spaces, or AWS**
- Add **user segmentation and personalization**

---

9. Conclusion

The primary objective of this project was to analyze and visualize digital transaction data from the PhonePe Pulse repository. To begin with, I downloaded the entire JSON dataset and stored it locally. Using Python's `json` library, I parsed and extracted the relevant data points from the nested JSON structure. This structured data was then transformed into tabular format using the `pandas` library and exported as multiple CSV files for ease of use and scalability.

Next, I imported these CSV files into a MySQL database to facilitate efficient querying and relational data operations. From this database, I identified and analyzed five real-world business use cases, such as state-wise transaction trends, top-performing districts, device usage distribution, and quarterly growth analysis. These use cases were selected to provide actionable insights into digital payment adoption and patterns across various regions in India.

Initially, I considered deploying the database on a cloud-based remote server to enable dynamic queries and real-time dashboard interactions. However, due to pricing constraints and limited resource availability, I opted to proceed with the CSV files for data visualization.

The final phase involved building an interactive and user-friendly dashboard using Streamlit. The dashboard includes multiple filtering options, interactive charts (using Plotly), and tabular summaries to allow users to explore the data intuitively. This approach ensured a cost-effective, yet functional and insightful presentation of the analysis, successfully meeting the project's goals of data extraction, transformation, storage, analysis, and visualization.