A PROJECT REPORT

Project Title	**PhonePe Transaction Insights**
Domain	**Finance/Payment Systems**

Submitted by

SAGHEER AHMED

Table of Contents

Overview

This repository contains a two-part PhonePe analytics project:

- **ETL (ETL.py)** - extracts PhonePe Pulse JSON files, transforms them into clean tabular form, and loads them into a MySQL database.
- **Dashboard (Dashboard.py)** - a Streamlit application that reads the MySQL tables and provides interactive analyses and visualizations across multiple business cases: transactions, insurance, user engagement, and geographic insights.

The dashboard targets business users, product managers, marketing, and growth teams to identify hotspots, measure penetration, and prioritize campaigns and product investments.

Project Structure

phonepe-pulse-project

```
├─ src
        ├─ ETL.py
        ├─ dashboard.py
├─ Docs
        ├─ Documentation.md
        ├─ presentation.pptx
├─ Pulse
        ├─ data/
                ├─ aggregated/
                ├─ map/
                └─ top/
├─ Sql
        └─ create_tables.sql
├─ README.md
├─ requirements.txt
```

Setup & Requirements

Core libraries :

- json

- pandas
- os
- streamlit
- plotly.express
- requests
- pymysql
- streamlit-option-menu

Database - MySQL

Install dependencies

- pip install -r requirements.txt

Run dashboard

- streamlit run Dashboard.py

Data Sources & Database Schema

The ETL script ingests raw JSON files from PhonePe Pulse dataset and creates the following MySQL tables (each table schema is created by ETL.py):

- agg_transaction : state-level aggregated transaction type metrics
- agg_insurance : state-level aggregated insurance metrics
- agg_user : state-level aggregated user (device/brand) metrics
- map_transaction : district-level transaction hover data
- map_insurance : district-level insurance hover data
- map_user : district-level user metrics (registered users, app opens)
- top_transaction : top pincodes for transactions
- top_insurance : top pincodes for insurance
- top_user : top pincodes for registered users
- top_district : district-level transaction top lists

Each table stores: States, Years, Quarter and relevant metric columns (transaction_Type, transaction_counts, transaction_amounts,transaction_percentage, registered users, app opens, District, pincodes).

ETL (ETL.py)

Purpose: Traverse the provided file tree of JSON Pulse data and materialize consistent tables in MySQL.

1. Connect to MySQL via pymysql.

2. Create database phonepe if not exists.
3. For each dataset path (aggregated, map, top folders), read JSON files recursively.
4. Parse JSON to collect rows into dictionaries then convert to Pandas DataFrames.
5. Normalize state names and create/insert into MySQL tables.
6. Commit bulk inserts using parameterized queries.

Notes: - File paths are currently absolute — update them to relative or configurable paths. - Ensure JSON schema stability; guard for missing keys.

Dashboard (Dashboard.py)

Purpose: Provide an interactive analytics interface using Streamlit and Plotly for the aggregated tables loaded into MySQL.

Main responsibilities: - Connect to MySQL, read tables into Pandas DataFrames. - Expose multiple business-case-driven visualization modules (ques1–ques5). - Provide map visualizations and a home page for quick KPIs.

The app fetches nine primary DataFrames and uses helper functions to build charts, maps, and ranked lists.

Helper Functions

Key helpers centralized in the script:

- safe_groupby(df, group_cols, agg_dict) — safe grouping & aggregation that returns empty DataFrame when input invalid.
- plot_bar(df, x, y, ...) — standardized Plotly bar chart wrapper.
- plot_line(...), plot_scatter(...) — wrappers for consistent Plotly line/scatter visuals.
- calc_penetration(df, group_cols, value_col, user_col) — computes penetration metric.
- calculate_year_growth(...) and calculate_year_growth1(...) — compute growth percentages (numeric and string-formatted %).

These help maintain consistent visuals and avoid duplicated code across business-case modules.

Business Cases (Modules)

Each (quesN) function implements a business case and renders interactive charts & tables.

Case 1 - Transaction Type Trends (ques1)

- Maps: Most used transaction type by amount and count (choropleth).

- State / Year / Quarter trends, distribution, and Top/Bottom 5 states.
- Useful for payment performance and category popularity insights.

Key Insights - Case 1: Transaction Type Trends

1. **Peer-to-Peer (P2P) Payments Dominate**
   - Across all states, peer-to-peer transfers are the most used transaction type, both by count and transaction value, accounting for 77.1% of total volume.

2. **Consistent Growth Over Time**
   - From 2018 to 2024, every transaction category (P2P, merchant payments, recharges, etc.) shows steady growth, with 2024 recording the highest transaction values across all types.
   - Within each year, Q4 consistently outperforms the other quarters.

3. **Category Distribution Highlights**
   - Overall contribution split:
     - P2P payments — 77.1%
     - Merchant payments — 18.9%
     - Recharge & Bill Payments — 3.86%
     - Financial Services — 0.41%
     - Others — 0.5%

This highlights the critical role of P2P while showing merchant payments as a fast-rising second.

4. **Top States by Transaction Amount (₹)**
   - Maharashtra (189,848 Cr), Karnataka (173,966 Cr), Uttar Pradesh (131,769 Cr), Tamil Nadu (122,072 Cr), Telangana (117,864 Cr) lead in transaction value, indicating urban & business-driven hubs as key markets.

5. **Top States by Transaction Count**
   - West Bengal (120M) tops in transaction count, followed by Maharashtra (116M), Karnataka (99M), Uttar Pradesh (92M), Andhra Pradesh (77M).
   - This shows a difference between volume (count) vs value (amount) trends, suggesting West Bengal has high frequency but relatively lower transaction amount size per transaction compared to Maharashtra and Karnataka.

Case 2 - Device Dominance & User Engagement (ques2)

- Engagement scoring per brand and state (Transaction_count × Transaction_percentage).
- Brand popularity, trends, and best brand per state (choropleth).
- User engagement metrics (Registered Users & App Opens) and engagement ratio maps.

Key Insights - Case 2: Device Dominance & User Engagement

1. **High Device Engagement States**

Engagement Score (Transaction Count × Transaction %):

- Maharashtra leads with the highest engagement score, followed by Karnataka.
- Indicates that these states not only generate high transaction volumes but also maintain strong brand-device linkages.

2. **Brand Dominance Across States**

- Xiaomi dominates in 32 states, with over 869M transactions, making it the most popular device brand across India.
- Vivo leads in 2 states (625M transactions), while Samsung leads in 1 state (672M transactions).
- Oppo and other brands trail significantly, reinforcing Xiaomi's nationwide dominance.

3. **Quarterly Brand Trends**

- From 2018 to 2022, Xiaomi consistently held the #1 spot in Q4 of every year, highlighting its sustained dominance and user retention over time.

4. **Brand Engagement Comparison**

Engagement scores rank as:

- Xiaomi (227M)
- Samsung (129M)
- Vivo (104M)

This confirms Xiaomi's double advantage, both in transaction count and engagement efficiency.

5. **User Engagement Ratios (App Opens per Registered User)**

- Top States: Meghalaya (174), Arunachal Pradesh (139), Mizoram (137), Ladakh (129), Andaman & Nicobar (92).
- Bottom States: Chandigarh (13), Delhi (14), Puducherry (17), Kerala (19), West Bengal (21).
- This reveals that smaller northeastern states have highly engaged users, while urban/metro regions like Delhi and Kerala show lower engagement per user despite high registered user counts.

Case 3 - Insurance Penetration & Trends (ques3)

- Insurance penetration choropleths (amount & count), hotspots, district/pincode rankings.
- Insurance vs user growth scatter plots (state/district/pincode).
- Penetration and growth analysis with year comparisons.

Key Insights - Case 3: Insurance Penetration & Trends

1. **State-Level Insurance Leaders**
   - By insurance amount, Karnataka (₹2,743M), Maharashtra (₹2,363M), and Uttar Pradesh (₹1,740M) lead.
   - By policy count, Karnataka again tops (1.95M), followed by Maharashtra (1.82M) and Tamil Nadu (1.22M).
   - This confirms Karnataka and Maharashtra as the core hubs for insurance adoption.
2. **Seasonal Growth Trend**
   - Quarter analysis shows Q4 consistently records the highest insurance transactions, followed by Q3.
   - This suggests year-end demand peaks — possibly due to financial year planning, tax-saving, or bonus-linked purchases.
3. **District & Pincode Hotspots**
   - Districts: Bengaluru Urban (₹1,491M, 1.1M policies), Pune, Thane, Rangareddy, and Chennai emerge as insurance hotspots.
   - Bottom districts like Hnahthial and Pherzawl show negligible penetration, indicating untapped markets.
   - Pin codes: 201301 (Noida, UP) is the top pincode with 3.03M policies in Q2 2024, making it a micro-level hotspot.
4. **Penetration Ratios (Insurance per Registered User)**
   - Top states: Andaman & Nicobar (11.53), Kerala (7.11), Goa (5.4), Lakshadweep (5.1).
   - Bottom states: Manipur (0.84), Nagaland (1.58), Odisha (1.64).
   - This shows strong insurance adoption in smaller states/UTs, while large states lag behind despite high user bases.
5. **Growth Leaders & Laggards (2018–2024)**
   - Top growth: Lakshadweep (+23,871%), South Garo Hills district (+1,252,065%), Pincode 794101 in Meghalaya (+86,906%).

- Bottom growth: Andhra Pradesh (only +854), Pherzawl district (-3.1%), and Pincode 682557 (-95.25%).
- This divergence shows rapid growth in small, emerging regions, while some mature states/districts face stagnation or decline.

Case 4 — Market Expansion (ques4)

- State-level transaction & user maps for expansion analysis.
- Penetration, growth%, and average usage (AppOpens / RegisteredUser) calculations.
- Top/bottom states for transactions, users and app opens.

Key Insights - Case 4: Market Expansion

1. **Top vs Bottom Transaction States**

- Transaction Amount Leaders (2018–2024): Telangana (₹41.6T), Karnataka (₹40.7T), Maharashtra (₹40.3T), Andhra Pradesh (₹34.7T), Uttar Pradesh (₹26.9T).
- Bottom States: Lakshadweep (₹1.6B), Mizoram (₹46.1B), Andaman & Nicobar (₹70.7B), Ladakh (₹88.9B), Sikkim (₹118.9B).
- This reflects strong digital adoption in southern states and limited penetration in small UTs/Northeast.

2. **User & Engagement Scale**

- Top Registered Users: Maharashtra (1.14B), Uttar Pradesh (942M), Karnataka (734M), Andhra Pradesh (557M), Rajasthan (556M).
- Top App Opens: Maharashtra (49.6B), Rajasthan (48.5B), Madhya Pradesh (39.7B), Karnataka (38.3B), Uttar Pradesh (33.2B).
- Bottom states like Lakshadweep (<200K users, ~5M app opens) show tiny market bases, ideal for targeted expansion pilots.

3. **State-Level Penetration (Transactions per User)**

- Leaders: Telangana (1.68M), Andhra Pradesh (1.31M), Karnataka (1.16M), Rajasthan (980K), Odisha (881K).
- Laggards: Lakshadweep (247K), Tripura (304K), Himachal (311K), Kerala (327K), Punjab (386K).
- Indicates deep penetration in South & East states, while Northern states have lower usage intensity.

4. **Growth Trends (2018–2024)**

- Fastest Growing States: Andaman & Nicobar (+16,158%), Ladakh (+14,111%).
- Slowest: Chandigarh (+1,370%), Manipur (+3,102%).

- Smaller regions like Andaman & Ladakh show explosive growth from low baselines, highlighting newly emerging digital markets.

5. **Average Usage (AppOpens / User)**

- Top Engagement States: Meghalaya (174), Arunachal Pradesh (139).
- Lowest: Chandigarh (13), Delhi (15).
- Shows that users in smaller Northeast states are highly engaged, while metro regions have lower per-user engagement despite large volumes.

Case 5 — User Engagement & Growth Strategy (ques5)

- Engagement Ratio (App Opens per Registered User) maps and trends.
- Loyalty index (App Opens vs Registered Users) for states and districts.
- Brand share pie chart and top registered-pincodes/districts analysis.

Key Insights — Case 5: User Engagement & Growth Strategy

1. **Engagement Ratio by State (AppOpens per Registered User):**

- India overall — Registered Users 8.86B, App Opens 402.29B, Engagement Ratio 45.38.
- Top states: Meghalaya (174.36), Arunachal Pradesh (138.85), Mizoram (136.57), Ladakh (129.49).
- Bottom states: Chandigarh (13.4), Delhi (14.93), Puducherry (~17), Kerala (~19), West Bengal (~21).

2. **Year-wise & Quarter-wise Trends:**

- Engagement ratio increased from 14.38 (2019) to 65.58 (2024).
- Q4 shows consistently higher engagement, while a dip occurred from Q3 2019 – Q1 2020 (onboarding of new users / market disruption).

3. **Registered User Growth:**

- From 46M in 2018 to 586.75M in 2024, showing a 12.7× increase.

4. **User Loyalty Index (AppOpens vs Users):**

- State-wise top: Rajasthan (87.28), Maharashtra (43.52).
- District-wise top: South West Khasi Hills (1783.16), Pakke Kessang (1599.93).
- Bottom districts: South East Delhi (13.8), Mumbai (13.71).

5. **Brand-wise Engagement Share:**

- Xiaomi (25.1%), Samsung (19.4%), Vivo (18.1%), Oppo (12.1%).

- Xiaomi is the dominant brand, but Samsung + Vivo together are close competitors.

Map & Home Page Visualizations

- The Map page (map()) provides a flexible map explorer that allows the user to:
    - Choose from multiple DataFrames (agg, map, top).
    - Select a column (e.g: Transaction_amount, Transaction_type, Brand).
    - Apply Year/Quarter filters and render state-level choropleths and top/bottom rankings.
- The Home page provides State, District and Pincode level filter and quick KPI cards, with tabbed bar charts at the selected Year and Quarter

---

Streamlit UI / Navigation

- Sidebar contains a logo and main menu
    - Home
    - Data Exploration
    - Business Cases
    - Map
- Each page uses Streamlit widgets (selectbox, tabs, columns) for interactive filters.
- GeoJSON for India states to render choropleth maps is used for state level.

Visualizations & Charts

Primary chart types used :

- **Choropleth (Plotly)** state-level geographical views.
- **Bar charts** (Plotly): ranking/top/bottom views.
- **Line charts** : time series for trends.
- **Scatter Plot**: correlation analyses (e.g., insurance vs users).
- **Pie/Donut Chart**: distribution (transaction type, brand share).

Design principles:

- Consistent color scales
- Use of tabs for side-by-side raw data & charts
- Empty-data handling with warnings.

Recommended Improvements

1. **Caching** : Cache GeoJSON and DB reads to speed up UI.
2. **Exports** : Add CSV/Excel/PDF download for filtered views.
3. **Performance** : Batch inserts and index commonly filtered columns in MySQL.
4. **Visualization** : Add district/pincode choropleths

How to run

1. Create a Python virtual environment.
2. Install packages from requirements.txt.
3. Ensure MySQL is running and ETL.py can connect.
4. Run ETL.py to populate the database
5. Run the dashboard:
    - streamlit run Dashboard.py
6. Use sidebar to explore pages:
    a. Home
    b. Data Exploration
    c. Business Cases
    d. Map.

Deliverables

- **ETL.py** - ETL script that loads PhonePe JSON files into MySQL.
- **Dashboard.py** - Streamlit application for visualization.
- This master documentation file.
- Additional deliverables: a README.md, sample SQL queries folder, and slide deck.