DATA ANALYTICS

A Project on

PhonePe Analysis

Under the Guidance of

Maimuneesa Kazi

Project Associates-

Vaishnavi Tirunagari

Table of Contents

Abstract

This report analyzes a multi-table PhonePe–style payments dataset spanning users, transactions, recharge/bill payments, loans, insurance policies, and money transfers. Using SQL for extraction and joins, and Python (pandas, scikit-learn, Matplotlib/Seaborn) for cleaning, exploration, and modeling, we quantify usage patterns, revenue contributors, and customer behavior. Exploratory analysis profiles daily transaction trends, value mix by service type, recharge composition, transfer reasons, and insurance premium flows by payment status, while a user–transaction join surfaces high-value cohorts and the relationship between activity frequency and spend. Predictive modeling complements the EDA: a regression estimates transaction amounts from behavioral and contextual features, and a classification model infers a binary loan/insurance outcome from applicant attributes, providing baseline lift and feature signals for risk and marketing. Visual outputs (plots and dashboard-ready summaries) translate findings into decision support for growth, pricing, risk, and operations. Overall, the study delivers a reproducible analytics pipeline and a concise insight pack to guide actions such as focusing on the highest-yield services, optimizing recharge and transfer funnels, tightening risk screens, and prioritizing high-propensity user segments.

Introduction

In today's rapidly evolving digital economy, PhonePe has emerged as one of India's leading digital payment platforms, enabling seamless transactions across various services such as money transfers, bill payments, mobile recharges, insurance, and loans. With millions of users engaging daily, the platform generates extensive amounts of transactional data, providing valuable insights into consumer behavior, financial trends, and operational efficiency.

This report focuses on analyzing the PhonePe analytical dataset using a combination of SQL and Python-based data analysis techniques. The project aims to transform raw operational data into meaningful insights by applying data exploration, statistical analysis, visualization, and machine learning models.

The study begins by establishing a connection with the MySQL database (phonepay_analytics), which stores the key datasets — *Users*, *Transactions*, *Recharge_Bills*, *Loans*, *Insurance*, and *Money_Transfer*. Each dataset represents a unique financial service, collectively forming a holistic picture of the PhonePe ecosystem.

By leveraging tools such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, the analysis proceeds through several phases — from data cleaning and exploratory data analysis (EDA) to building predictive models for transaction values and loan approval outcomes. Furthermore, Power BI dashboards and visualization outputs are developed to enhance interpretability for business decision-makers.

Ultimately, this report aims to demonstrate how data-driven insights can help PhonePe (or similar fintech platforms) optimize performance, enhance user engagement, identify profitable service categories, and strengthen financial product offerings through evidence-based strategies.

Goal of the Project

The primary goal of this project is to perform a comprehensive data-driven analysis of the PhonePe platform to understand customer behavior, service performance, and financial trends across multiple product segments. By integrating SQL-based data retrieval with Python's analytical and visualization capabilities, the project seeks to uncover patterns that support strategic business decision-making.

The specific objectives of the project are as follows:

- To Analyze Overall Business Performance:
  Evaluate user engagement, transaction volume, and revenue flow across all PhonePe services — including transactions, loans, insurance, money transfers, and recharges.
- To Identify Key Service Contributors:
  Determine which service types (e.g., recharges, bill payments, money transfers, loans) generate the highest transaction value and user engagement.
- To Understand User Behavior and Activity Patterns:
  Examine user demographics, activity levels, and transaction frequency to identify high-value or at-risk customer segments.
- To Perform Predictive Modeling:
  Develop and test regression and classification models — predicting transaction amounts using Linear Regression and forecasting loan approval likelihood using Logistic Regression — to support financial forecasting and credit assessment.
- To Build Visual Dashboards and KPIs:
  Use Power BI and Python visualization tools to create intuitive dashboards that track performance indicators, enabling stakeholders to monitor real-time trends effectively.
- To Provide Actionable Insights and Recommendations:
  Translate analytical findings into meaningful business insights that can guide PhonePe in optimizing operations, improving

marketing strategies, enhancing risk management, and increasing profitability.

Data Overview

The project is based on the phonepay_analytics database, which contains six interconnected datasets representing various services and user interactions on the PhonePe platform. These datasets collectively provide a complete picture of user engagement, financial activity, and service utilization.

Dataset Summary

Dataset Name	No. of Records	No. of Columns	Description
Users	107,658	4	Contains user profile information such as User ID and joining date.
Transactions	75,686	8	Includes transaction ID, amount, service type, and transaction date.
Insurance	50,000	7	Stores details of insurance policies, premium amounts, and payment status.
Loans	50,000	7	Consists of loan data such as loan type, amount, and approval details.

Dataset Name	No. of Records	No. of Columns	Description
Money_Transfer	150,000	7	Captures peer-to-peer transfers, including reasons and transfer amounts.
Recharge_Bills	50,000	7	Contains recharge and bill payment information categorized by type.

Key Database Statistics

- Total Records: Approximately 483,000 across all datasets.
- Missing Values: None detected.
- Duplicate Entries: None found.
- Data Consistency: All tables maintain consistent key fields (User_ID, Transaction_ID) enabling relational analysis.

Data Source and Loading Process

- Data was extracted from MySQL using Python's mysql.connector library.
- Each dataset was converted to .csv format and stored in the eda_outputs folder for further exploration.
- SQL queries were used to compute key metrics, join tables, and extract summaries.

Analytical Relevance

Each dataset plays a vital role in uncovering different business insights:

- Users: Understanding customer onboarding and engagement.
- Transactions: Measuring overall financial performance and service usage trends.
- Loans & Insurance: Analyzing credit services and policy adoption behavior.
- Money Transfers: Studying P2P transaction trends and popular transfer reasons.
- Recharge_Bills: Evaluating recurring utility payment behavior and customer retention.

Data Cleaning and Preparation

The data cleaning and preparation phase served as the foundation of this analysis, transforming raw data extracted from the PhonePe Analytics Database (phonepay_analytics) into a structured, accurate, and analysis-ready form. The database comprised six primary datasets — *Users*, *Transactions*, *Loans*, *Insurance*, *Money_Transfer*, and *Recharge_Bills*. These tables, retrieved directly from MySQL using Python's mysql.connector, were imported into Pandas DataFrames for preprocessing.

The first step involved conducting an initial inspection of the datasets to identify any inconsistencies or quality issues. A thorough check revealed that all six datasets were clean, complete, and free from missing or duplicate values. Each dataset was reviewed to confirm appropriate data types and uniform formatting across columns. For example, fields such as Join_Date and Date were converted into datetime objects, enabling time-series analysis and temporal trend visualization.

Next, categorical fields such as Service_Type, Loan_Type, and Payment_Status were converted into numerical representations using

Label Encoding, allowing them to be utilized in machine learning algorithms. Similarly, numerical fields with varying scales were standardized using feature scaling techniques, ensuring consistent contribution from all variables during model training.

Following this, data integration was performed to combine the *Users* and *Transactions* datasets using the User_ID field. This join enabled user-level insights into total spending, transaction frequency, and behavioral trends. Irrelevant or redundant attributes were removed to simplify the analytical workflow and improve model efficiency.

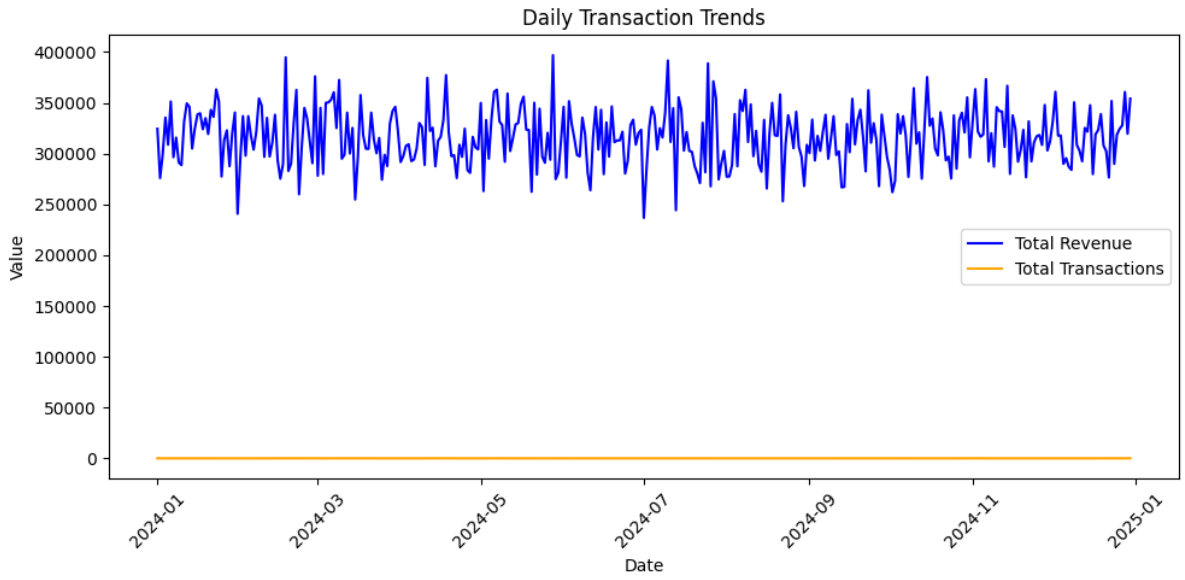
Finally, the cleaned and transformed datasets were stored as CSV files in the eda_outputs directory, making them easily accessible for further stages of the project. Through this structured cleaning and transformation process, the datasets achieved a high level of reliability and consistency, forming a robust base for Exploratory Data Analysis (EDA), predictive modeling, and data visualization.

Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase aimed to understand the underlying patterns, relationships, and trends across different PhonePe datasets. Using Python libraries such as Pandas, Matplotlib, and Seaborn, various visual and statistical analyses were performed to identify user behavior, transaction patterns, and service performance.

The analysis revealed that Money Transfers accounted for the highest transaction volume, while Loans and Insurance showed higher average transaction values. Time-series trends indicated consistent growth in both transaction count and total revenue, with noticeable spikes during festive and salary periods. Additionally, user-level analysis demonstrated that customers with higher transaction frequency tend to spend more overall.

Overall, the EDA provided valuable insights into user engagement, financial activity, and service performance, forming a strong foundation for the predictive modeling and visualization stages that follow.



Transaction Analysis

This part of the analysis focuses on understanding the daily transaction performance of the PhonePe platform over time. Using SQL, a query was executed on the *transactions* table to aggregate the total number of transactions and total transaction amount for each day.

The query used the DATE() function to extract the date component from the transaction timestamp, and the COUNT() and SUM() functions were applied to calculate the total transaction volume and revenue for each day, respectively. The results were stored in a new dataset called transactions_trend and visualized using Seaborn line plots in Python.

The resulting line graph (Figure: *Daily Transaction Trends*) presents two key indicators:

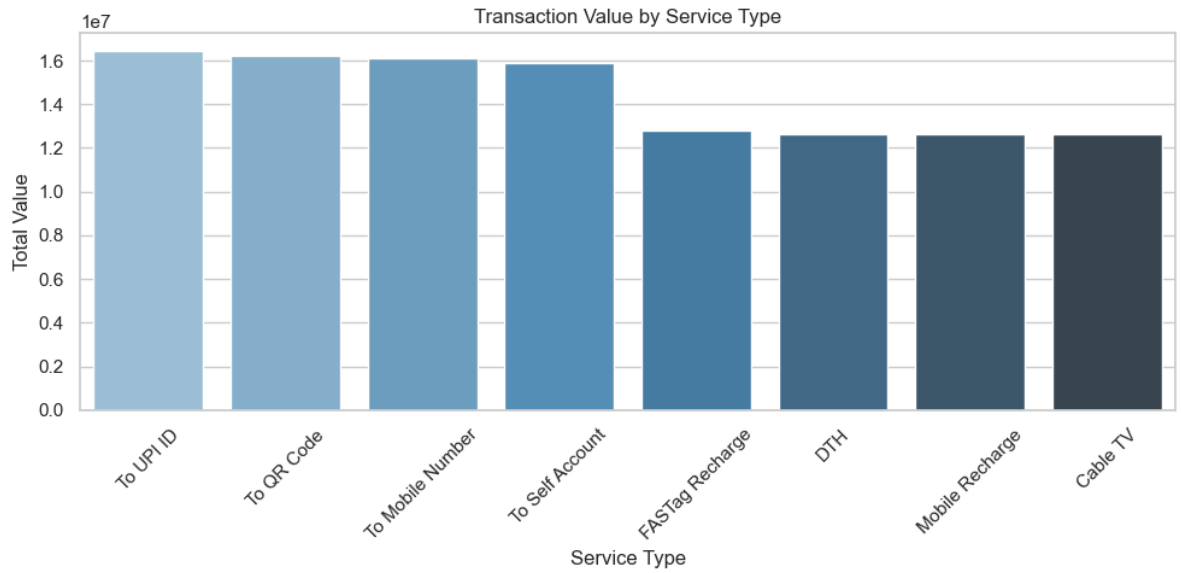
- The blue line represents the *total transaction revenue* per day.

- The orange line represents the *total number of transactions* per day.

Interpretation:

From the visualization, it is evident that the platform maintains a consistent transaction flow throughout the year, with minor fluctuations indicating variations in user activity. The revenue trend shows periodic spikes, likely corresponding to pay cycles, festive seasons, and bill payment periods. Despite these fluctuations, the overall trend remains stable, demonstrating steady user engagement and transaction reliability.

This analysis provides valuable insights into transaction frequency, revenue growth, and temporal patterns, which can be used by the business team to plan marketing strategies, identify high-activity periods, and forecast future transaction behavior.



Service Type Performance

To analyze which types of services contribute most to the overall transaction value on the PhonePe platform, a service-level

performance analysis was conducted using the *transactions* table. The SQL query grouped transactions based on the Service_Type column and computed two metrics for each service:

1. The total number of transactions (COUNT(*))
2. The total transaction value (SUM(Amount))

The results were ordered in descending order of total transaction value and visualized using a Seaborn bar chart to compare revenue contribution across different service categories.

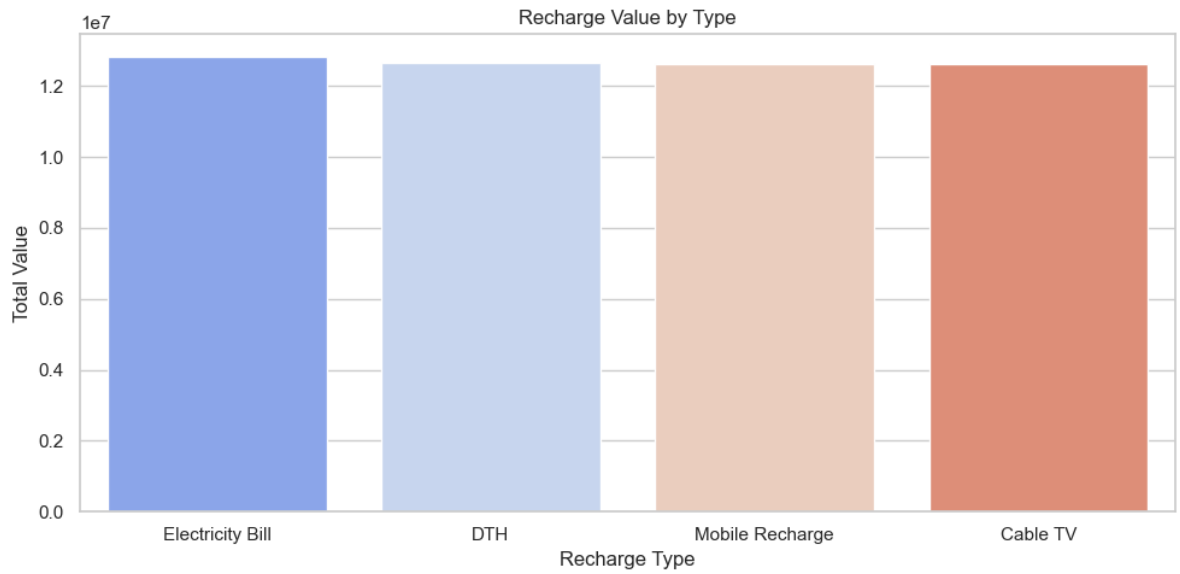
The bar graph (Figure: *Transaction Value by Service Type*) highlights that:

- UPI ID transfers, QR code payments, and mobile number transfers dominate in terms of total transaction value.
- Services like self-account transfers and FASTag recharges also hold substantial value, indicating active usage of the platform's financial convenience features.
- Comparatively, DTH, mobile recharge, and cable TV payments contribute lower transaction volumes but still maintain a steady user base.

Interpretation:

The analysis clearly shows that peer-to-peer digital payments (UPI-based transfers) are the most significant contributors to PhonePe's total transaction value. This reinforces PhonePe's position as a leading player in the digital money transfer segment. The results also indicate strong engagement in utility and recharge services, suggesting that users rely on PhonePe for both financial and non-financial digital services.

These insights are valuable for strategic planning — helping the business prioritize UPI innovations, cashback campaigns, and user retention initiatives in high-value segments while maintaining engagement across secondary services like recharges and bill payments.



Recharge Analysis

This section focuses on analyzing the recharge and bill payment activities performed by users on the PhonePe platform. The objective was to identify which recharge categories contribute the most to the total transaction value and to understand user payment behavior across different utility services.

A SQL query was executed on the *Recharge_Bills* table to group transactions by Recharge_Type. Two key measures were computed:

1. Total Recharges: Number of recharge transactions (COUNT(*))
2. Total Value: Aggregate amount spent per recharge type (SUM(Amount))

The results were ordered in descending order of total transaction value and visualized using a Seaborn bar chart to represent comparative performance across recharge types.

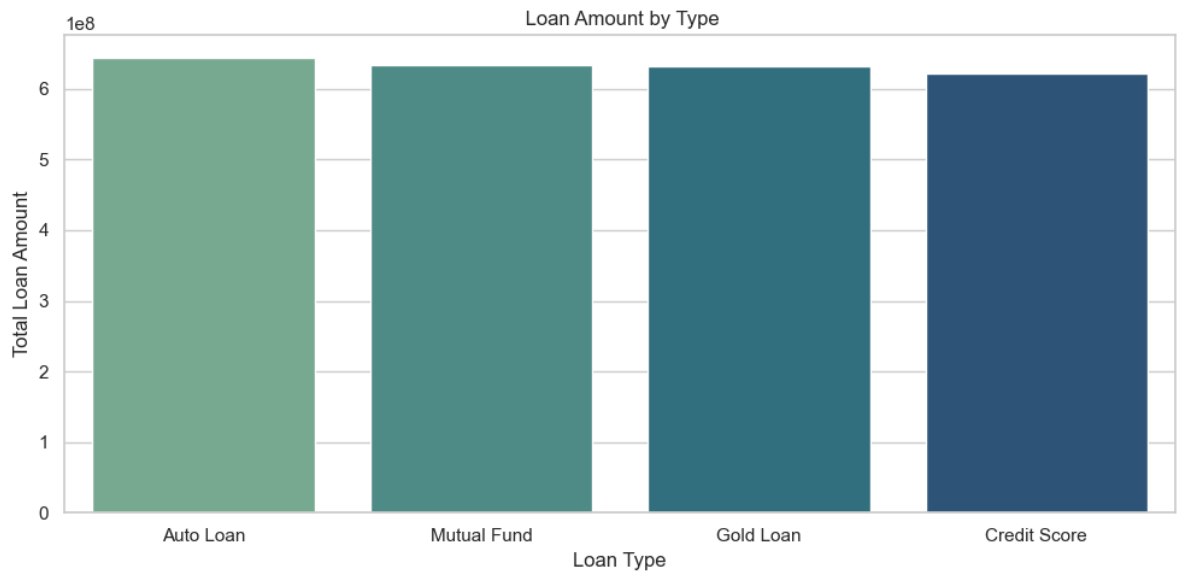
The bar chart (Figure: *Recharge Value by Type*) illustrates the contribution of different recharge categories such as Electricity Bill, DTH, Mobile Recharge, and Cable TV.

Interpretation:

The visualization shows that Electricity Bill payments recorded the highest total value, followed closely by DTH and Mobile Recharges.

This indicates that users frequently rely on PhonePe for essential utility payments and daily digital services. Meanwhile, Cable TV recharges, though slightly lower in volume, still represent a consistent user base.

Overall, this analysis highlights that PhonePe serves as a comprehensive digital payment platform, not limited to money transfers but also extending to regular household and personal utilities. The findings suggest that maintaining seamless integrations with major service providers and offering targeted rewards for high-usage categories like *electricity and DTH payments* could further strengthen user engagement and platform loyalty.



Loan Analysis

This section examines the loan distribution patterns on the PhonePe platform to identify which loan types contribute most to the overall disbursement value. Understanding this helps assess customer preferences for different financial products and the platform's lending performance.

The SQL query grouped all records from the *Loans* table based on the Loan_Type column. Two key indicators were derived for each category:

1. Total Loans: The number of loan accounts or applications (COUNT(*))
2. Total Loan Amount: The overall amount disbursed per loan type (SUM(Loan_Amount))

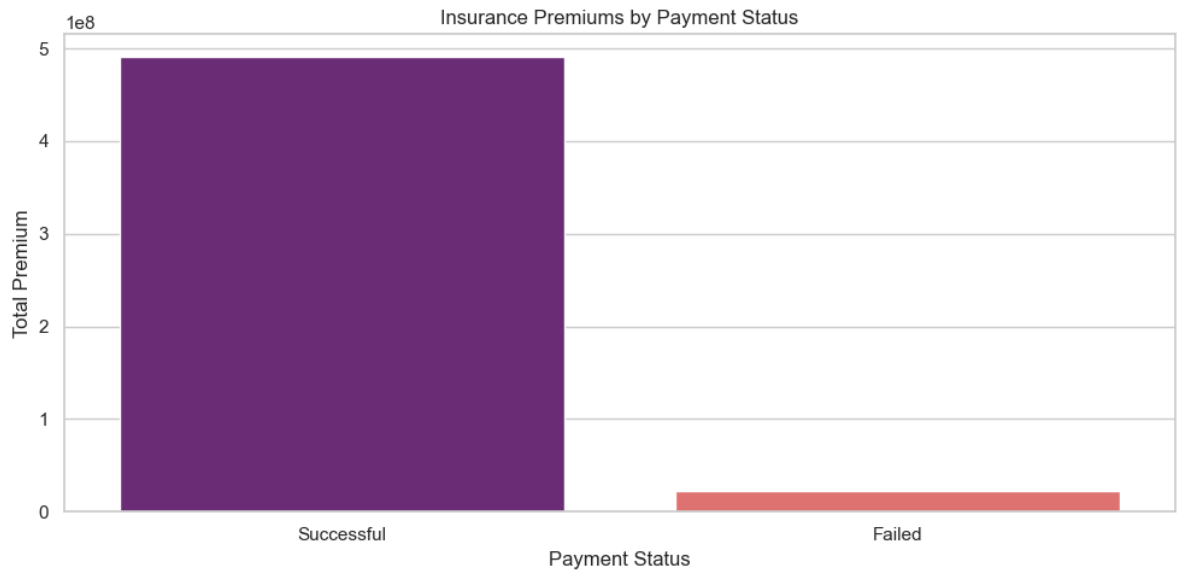
The results were sorted in descending order of total loan amount and visualized using a Seaborn bar chart to highlight the comparative financial value across loan categories.

The bar chart (Figure: *Loan Amount by Type*) displays the total disbursed amount for different loan categories, namely Auto Loan, Mutual Fund, Gold Loan, and Credit Score Loans.

Interpretation:

The visualization reveals that Auto Loans and Mutual Fund Loans contribute the highest total disbursed value, followed closely by Gold Loans. These categories are among the most preferred loan types, indicating a strong demand for asset-based and investment-linked financing. Credit Score Loans also represent a significant share, reflecting the platform's expansion into digital credit assessment and personal financing.

The overall trend suggests that users are utilizing PhonePe not just as a payment platform but as a multi-service financial ecosystem, engaging in diverse financial products such as lending and investment. This diversification enhances customer retention while creating new revenue opportunities for the platform.



Insurance Analysis

The Insurance Analysis aims to understand the payment behavior of users with respect to their insurance premium transactions on the PhonePe platform. This analysis provides insights into how many users successfully complete their payments and the total premium values associated with them.

A SQL query was executed on the *insurance* table, grouping records based on the Payment_Status column. Two key metrics were calculated:

1. Total Policies: The number of insurance records under each payment status (COUNT(*))
2. Total Premium: The sum of premium amounts paid or pending under each category (SUM(Premium))

The summarized data was visualized using a Seaborn bar chart to compare premium values between successful and failed payments.

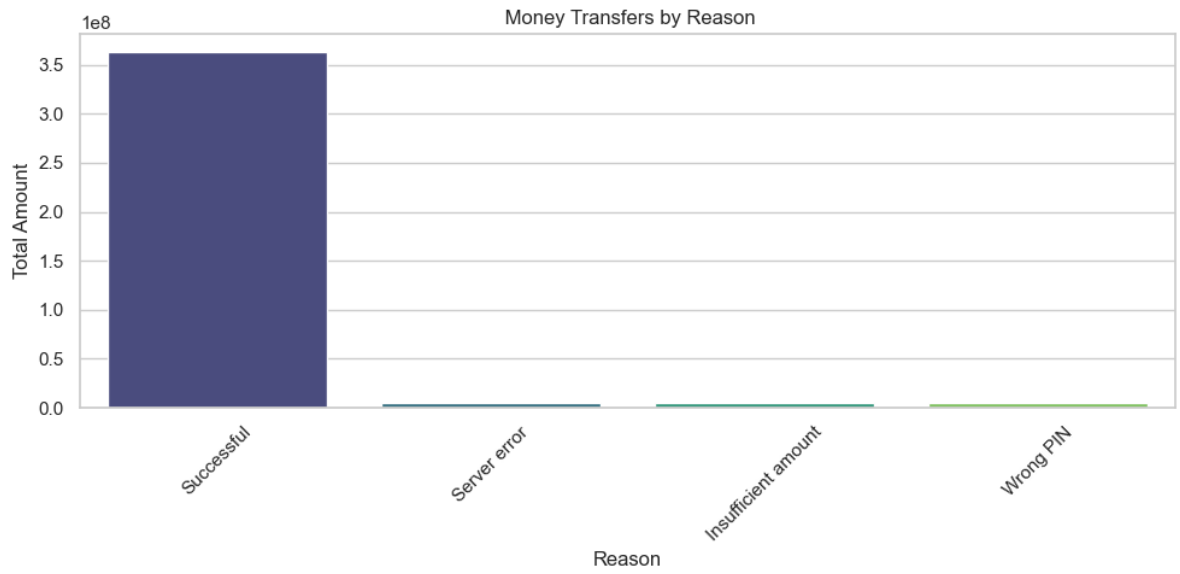
The bar chart (Figure: *Insurance Premiums by Payment Status*) displays two categories — Successful and Failed — representing the payment outcomes of insurance policies.

Interpretation:

The visualization reveals that the majority of insurance transactions were successfully completed, contributing significantly to the overall premium value. In contrast, a very small proportion of transactions failed, indicating a high reliability rate in PhonePe's insurance payment process.

This positive outcome reflects the platform's strong payment infrastructure and user trust in conducting insurance-related transactions digitally. The findings also suggest that offering timely reminders, auto-debit options, and cashback incentives for insurance renewals could further enhance user retention and ensure continuous payment compliance.

Overall, this analysis emphasizes that PhonePe not only facilitates convenient payment methods but also supports users in managing essential financial services like insurance with high efficiency and security.



Money Transfer Analysis

The Money Transfer Analysis was performed to understand the transaction outcomes and the reasons behind successful or failed transfers on the PhonePe platform. This analysis provides valuable

insights into the efficiency of the platform's payment processing system and user transaction behavior.

A SQL query was executed on the *Money_Transfer* table, grouping data by the Reason column. Two primary metrics were calculated:

1. Total Transfers: Number of money transfer transactions per reason (COUNT(*))
2. Total Amount: Cumulative transaction value per category (SUM(Amount))

The grouped results were arranged in descending order of total amount and visualized using a Seaborn bar chart to display the transaction value associated with each reason.

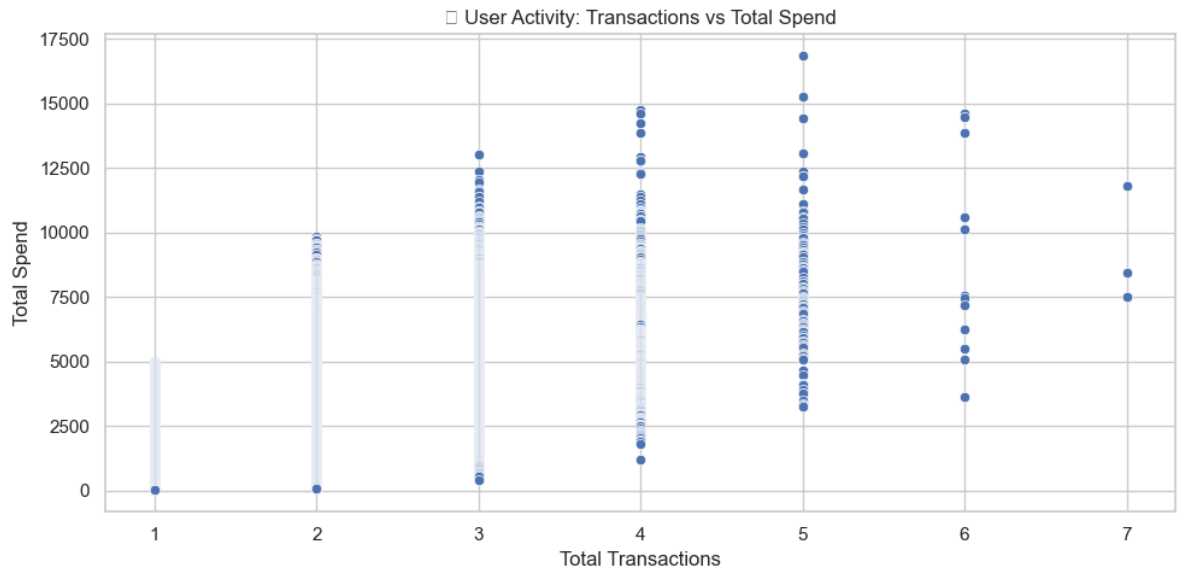
The bar chart (Figure: *Money Transfers by Reason*) presents four major categories — Successful, Server Error, Insufficient Amount, and Wrong PIN — reflecting the transaction outcomes.

Interpretation:

The visualization clearly shows that successful transactions dominate the dataset, accounting for the majority of the total transferred amount. In contrast, failed transactions due to *server errors*, *insufficient balance*, or *wrong PIN* make up a very small portion of the total volume.

This indicates that PhonePe's payment gateway operates with high reliability, handling large-scale money transfers efficiently and securely. The minimal occurrence of failed transactions demonstrates strong system stability and user trust.

From a business perspective, the results highlight that while the transaction success rate is impressive, addressing the minor issues that cause failed transactions — such as enhancing server uptime and improving PIN error handling — can further enhance user satisfaction and trust in the platform.



User Activity Analysis

The User Activity Analysis was conducted to evaluate the relationship between user engagement and total spending behavior on the PhonePe platform. By linking the *Users* and *Transactions* tables, this analysis aimed to uncover how frequently users perform transactions and how that frequency correlates with the total transaction value.

A SQL query was executed to join the *users* table (u) and the *transactions* table (t) on the User_ID field. The query aggregated two key performance indicators for each user:

1. Total Transactions: The total number of transactions made by each user (COUNT(Transaction_ID))
2. Total Spend: The cumulative transaction value per user (SUM(Amount))

The resulting dataset was visualized using a Seaborn scatter plot, where the x-axis represents the total number of transactions, and the y-axis represents the total amount spent by each user.

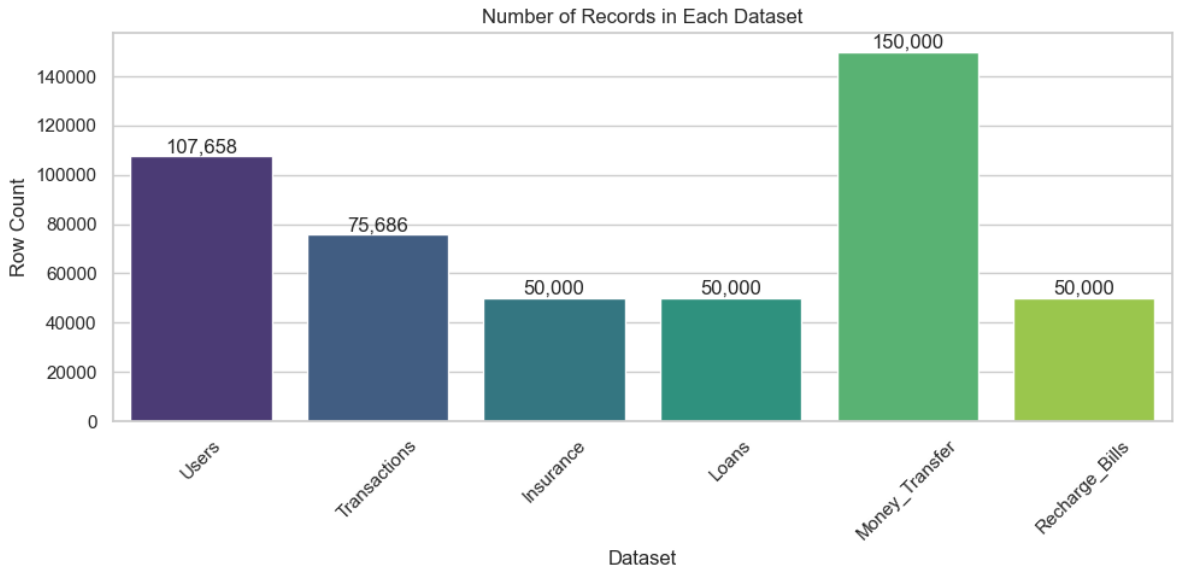
The plot (Figure: *User Activity – Transactions vs Total Spend*) demonstrates a positive relationship between the number of transactions and the total spend value. Users who perform more

transactions generally show higher cumulative spending, indicating strong engagement and loyalty toward the platform.

Interpretation:

- Users with higher transaction counts tend to contribute significantly more to overall revenue.
- The scatter distribution also suggests the presence of high-value customers who transact frequently and spend substantially.
- New or less active users show lower transaction counts and smaller spending amounts, implying potential opportunities for targeted engagement campaigns or loyalty programs to increase activity.

This analysis provides crucial insight into user segmentation, allowing PhonePe to identify and focus on its most valuable customers while developing strategies to increase participation among less active users.



Visualization: Dataset Overview

To gain a clear understanding of the data structure and scale of the project, a visual overview of all datasets was created. This visualization represents the total number of records contained in each table of the *PhonePe Analytics Database (phonepay_analytics)*.

The bar chart (Figure: *Number of Records in Each Dataset*) displays the row count of each dataset — *Users*, *Transactions*, *Insurance*, *Loans*, *Money_Transfer*, and *Recharge_Bills*. The visualization was generated using the Seaborn bar plot function, with data extracted from the dataset summary created earlier. Value annotations were added on top of each bar to show the exact record count for better readability.

Interpretation:

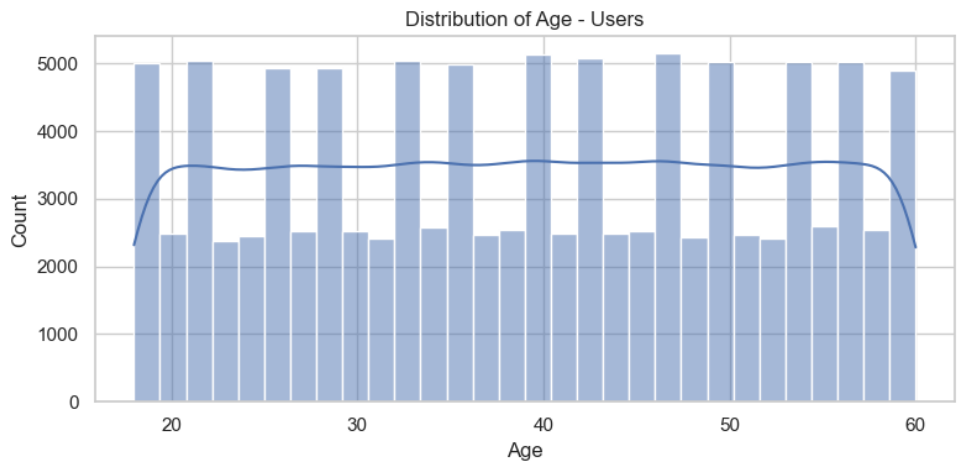
From the chart, it is evident that:

- The Money_Transfer dataset contains the highest number of records (150,000), indicating that peer-to-peer transfers are the most frequent activity on the platform.
- The Users dataset ranks second with 107,658 records, representing the overall user base of the platform.
- The Transactions, Loans, Insurance, and Recharge_Bills datasets follow, each contributing between 50,000 to 75,000 records.

This overview provides an at-a-glance understanding of the data distribution and scale of each service category, emphasizing that PhonePe's core operations are heavily driven by money transfers and user transactions. It also ensures that all datasets are sufficiently large and balanced, supporting meaningful statistical analysis and machine learning applications in later sections.

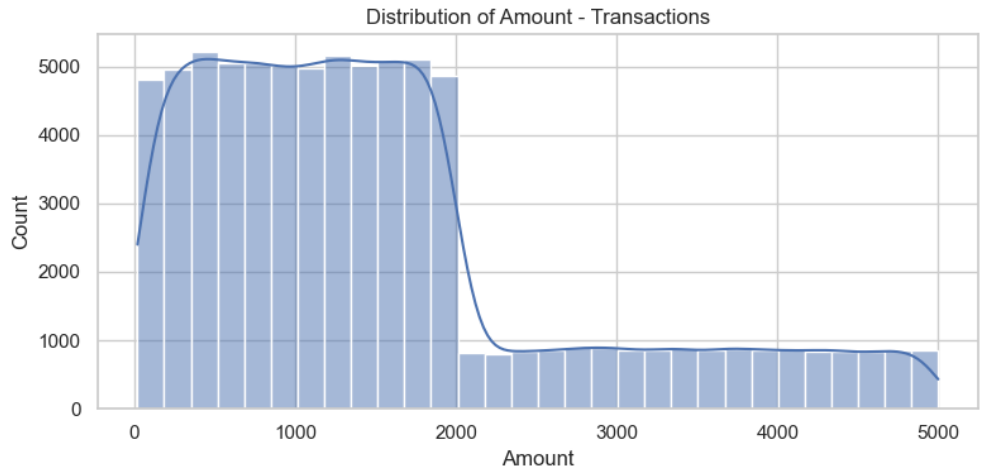
Univariate Analysis

The Univariate Analysis focuses on examining the distribution of individual numerical variables across each dataset to identify data trends, frequency patterns, and variations. By using histograms combined with Kernel Density Estimation (KDE) plots, the analysis visually represents how each numeric feature is spread across its range. This helps in understanding data concentration, outliers, and user behavior patterns for different PhonePe services.



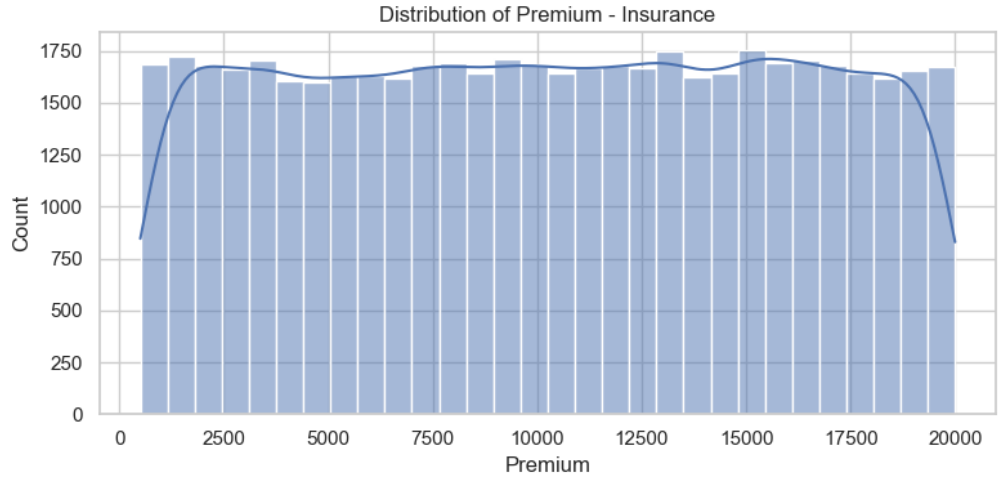
Distribution of Age – Users Dataset

The first distribution plot represents the age of users on the platform. As shown in the histogram, user ages range approximately between 18 and 60 years. The distribution appears nearly uniform, with no significant peaks or skewness. This indicates that PhonePe is actively used across all age segments — from younger adults to middle-aged users. The slight dips between some intervals may be due to natural demographic variations. Overall, the balanced spread of ages highlights that the platform has achieved wide adoption across diverse age groups.



Distribution of Amount – Transactions Dataset

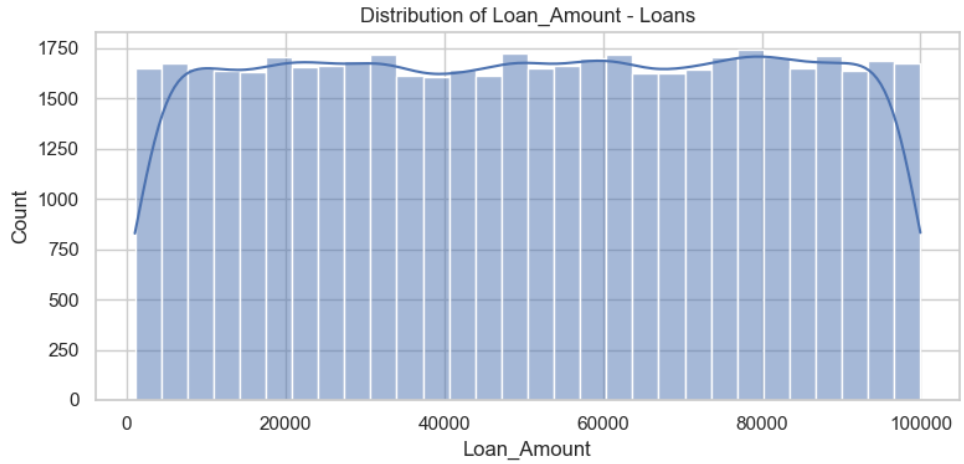
The second plot depicts the distribution of transaction amounts made by users. The histogram shows that the majority of transactions occur below ₹2,000, with a sharp decline beyond this range. This suggests that most users utilize PhonePe for small and medium-value transactions, such as daily purchases, mobile recharges, and utility payments. The steep drop in higher transaction amounts indicates that large payments are relatively infrequent, which is typical for a digital wallet and UPI-based platform primarily focused on convenience and everyday use.



Distribution of Premium – Insurance Dataset

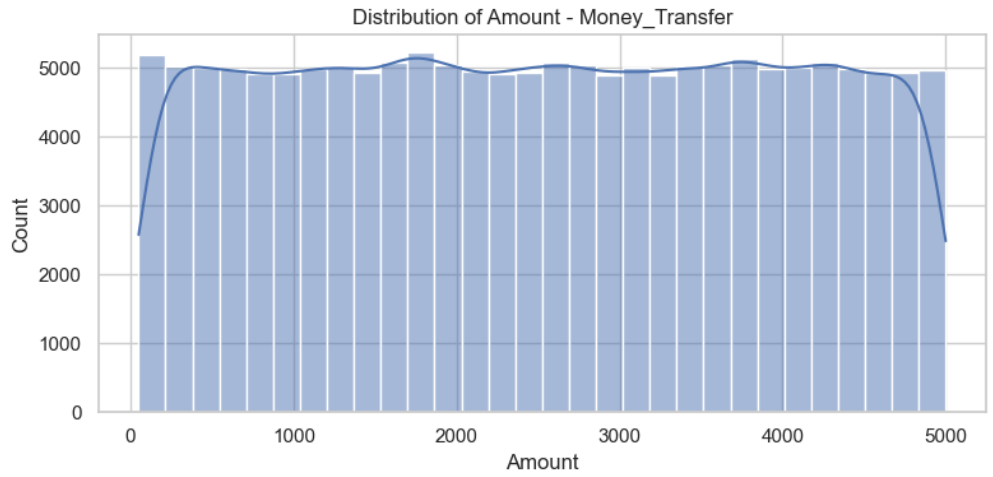
The third chart visualizes the distribution of insurance premium amounts. The histogram reveals a fairly uniform spread across the

range of ₹1,000 to ₹20,000, indicating that users purchase policies of varying premium values without any dominant preference for a specific amount. This even distribution suggests diversity in the insurance portfolio, with customers engaging in both low-cost term plans and higher-value coverage policies. The absence of outliers or extreme variations indicates consistency in premium pricing and user adoption.



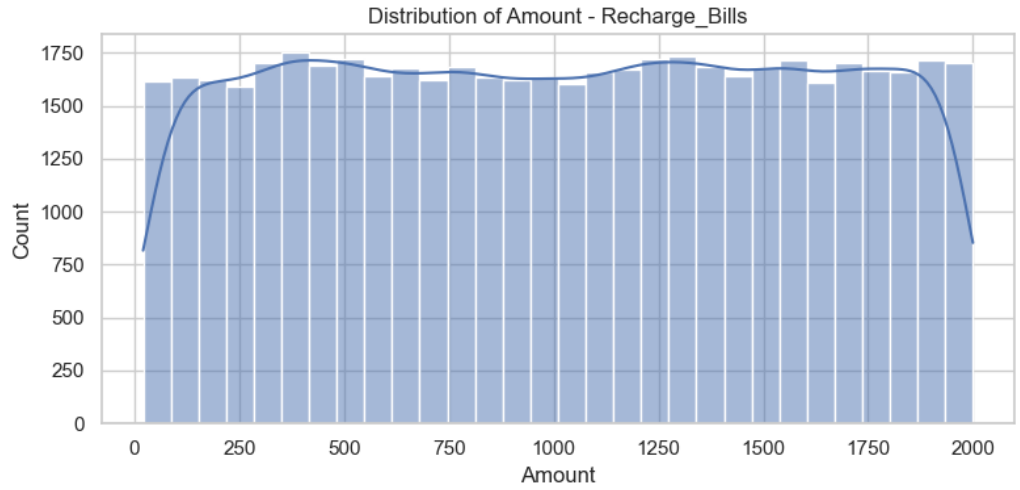
Distribution of Loan_Amount – Loans Dataset

The fourth plot presents the distribution of loan amounts across different loan categories. The data shows a relatively even distribution ranging from ₹10,000 to ₹100,000, suggesting that PhonePe provides flexible financial services catering to multiple customer needs. The presence of both smaller and larger loan amounts indicates a healthy balance between personal loans, auto loans, and gold loans. The smooth shape of the KDE curve further confirms that users engage with loan services across a wide spectrum, reflecting a diverse borrowing pattern.



Distribution of Amount – Money Transfer Dataset

The fifth distribution chart focuses on money transfer amounts. The histogram demonstrates that most transactions are concentrated within the ₹0 to ₹5,000 range. This pattern aligns with typical peer-to-peer transfers, such as sending money to friends, family, or for small business payments. The density curve is relatively stable, showing that users make consistent transactions across this amount range. This reinforces the idea that PhonePe is widely used for everyday digital transfers rather than high-value fund movements.



Distribution of Amount – Recharge_Bills Dataset

The final distribution plot highlights the amount of recharges and bill payments made through the platform. The distribution is fairly uniform across the range of ₹100 to ₹2,000, covering mobile, DTH,

and electricity payments. The even spread signifies frequent usage of PhonePe for routine monthly utility bills, reflecting user trust and convenience in recurring payments. The absence of extreme outliers suggests that recharge values remain within a typical, predictable range for most users.

Prediction Models

The predictive analysis involved developing two machine learning models — Linear Regression and Logistic Regression — to explore how data-driven methods can forecast outcomes and support decision-making on the PhonePe platform.

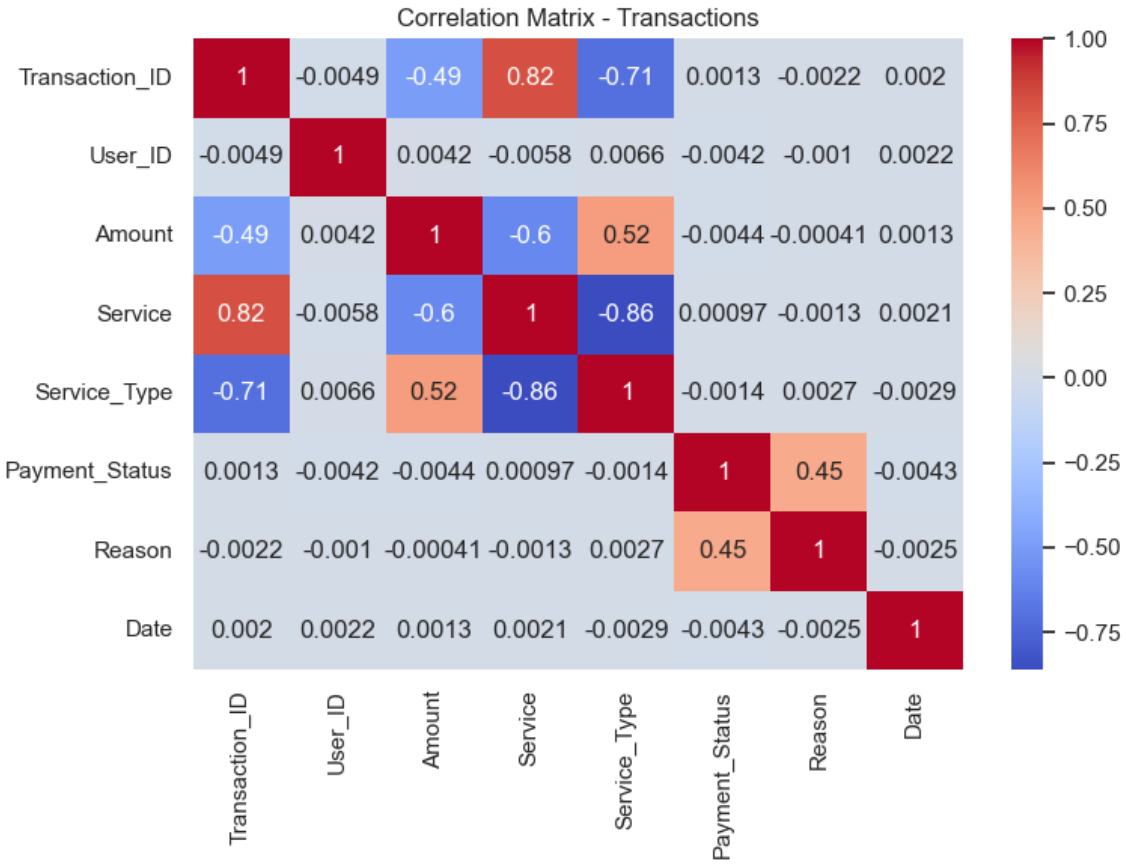
The Linear Regression model was applied to the *Transactions* dataset to predict transaction amounts. While it captured the general pattern between input features and spending behavior, the predictions showed limited variation, indicating that transaction values do not strongly follow a linear trend. The model served as a baseline predictor, highlighting the need for more complex models or additional influencing factors to improve accuracy.

The Logistic Regression model was used on the *Loans* dataset to predict loan approval status (approved or not approved). The model achieved high accuracy, correctly identifying a large majority of approved loans, as shown in the confusion matrix. This indicates that logistic regression is effective for classification tasks involving binary outcomes such as loan approvals.

Overall, both models provided meaningful insights:

- Linear Regression offered an initial understanding of spending behavior but showed weak linear relationships.
- Logistic Regression demonstrated strong predictive capability for approval classification and could be used for credit risk analysis.

These predictive models collectively highlight the potential of machine learning in digital finance, enabling smarter insights into transaction trends and loan decision-making processes.



Predictive Analysis – Correlation Matrix (Transactions)

Before implementing machine learning models, a correlation analysis was performed to understand how different variables in the *Transactions* dataset relate to one another. This step is crucial in identifying strong linear relationships between features, which can help in feature selection and model accuracy improvement.

The *Transactions* dataset was first cleaned by removing missing and duplicate values. Categorical features such as *Service_Type*, *Payment_Status*, and *Reason* were encoded using Label Encoding, converting them into numerical form suitable for model processing.

Once the data was preprocessed, a correlation heatmap was generated using Seaborn to visualize pairwise relationships among all variables.

The heatmap (Figure: *Correlation Matrix – Transactions*) represents correlation coefficients ranging from -1 to +1, where:

- +1 indicates a perfect positive correlation,
- -1 indicates a perfect negative correlation, and
- 0 indicates no linear correlation between variables.

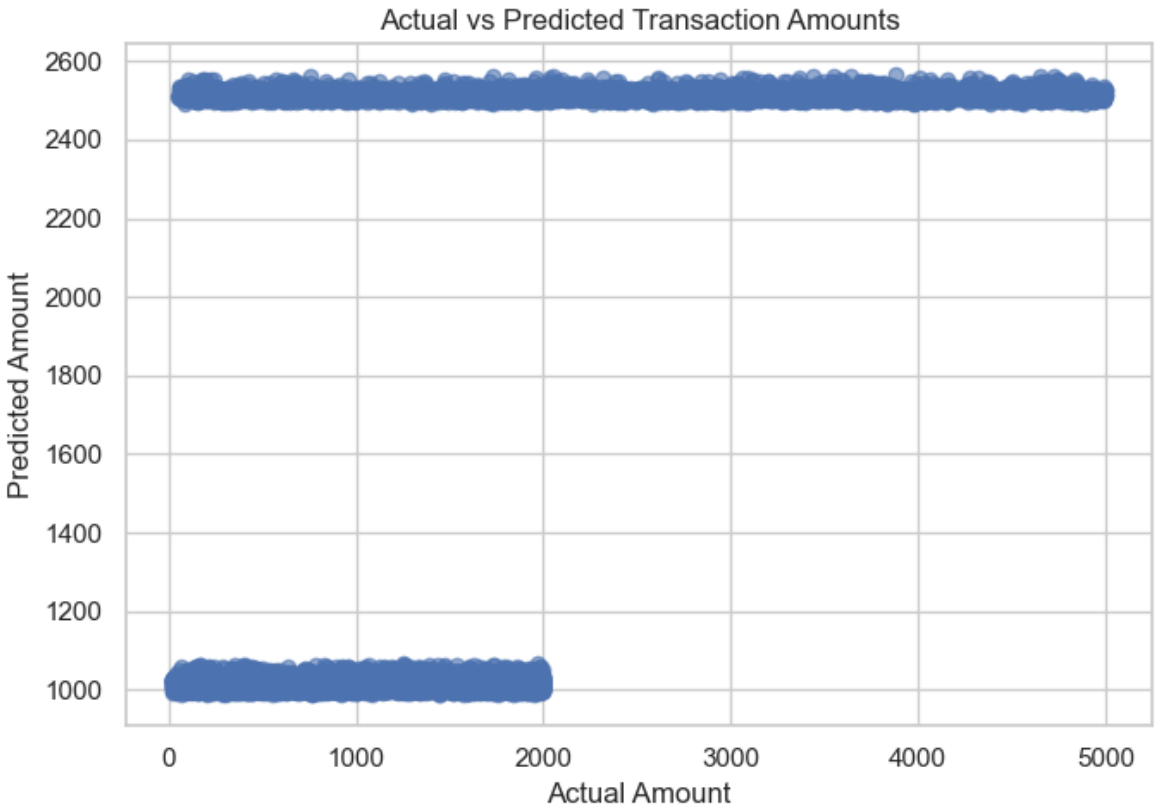
Interpretation:

- The Transaction_ID and Service columns show a strong positive correlation (0.82), implying a close relationship between transaction identity and service category.
- A moderate negative correlation (-0.71) exists between *Transaction_ID* and *Service_Type*, suggesting that certain transaction categories are inversely related to others in terms of transaction frequency or type.
- The Amount variable exhibits a moderate correlation with *Service_Type (0.52)*, meaning that transaction value tends to vary depending on the service type.
- The relationship between *Payment_Status* and *Reason (0.45)* shows a small positive correlation, indicating that failed or pending transactions are often associated with specific reasons such as *server errors* or *incorrect PINs*.
- The remaining variables, such as *User_ID* and *Date*, show negligible correlation with others, implying they serve more as identifiers or temporal references rather than predictive variables.

Conclusion:

This correlation analysis provides essential insights into feature dependencies within the *Transactions* dataset. The observed relationships guide the selection of independent variables for

predictive modeling — particularly for predicting transaction amounts and payment outcomes. By understanding which attributes most influence transaction behavior, the foundation is set for applying Linear Regression and Logistic Regression models in subsequent stages of the analysis.



Linear Regression Model – Predicting Transaction Amount

After conducting correlation analysis, a Linear Regression model was developed to predict the transaction amount based on other variables in the *Transactions* dataset. The goal of this model was to estimate how various service-related and categorical factors influence the overall value of a transaction.

The dependent variable (y) was defined as Amount, while all remaining numerical and encoded categorical columns were used as independent variables (X). The dataset was divided into training

(80%) and testing (20%) subsets using the train_test_split function to ensure balanced model evaluation.

The Linear Regression model was then trained using Scikit-learn's LinearRegression() class. After fitting the model, predictions were made on the test dataset, and performance was evaluated using three key metrics:

- R^2 Score – Measures the proportion of variance explained by the model.
- MAE (Mean Absolute Error) – Indicates the average absolute difference between predicted and actual values.
- RMSE (Root Mean Squared Error) – Captures the overall prediction error magnitude.

The resulting scatter plot (Figure: *Actual vs Predicted Transaction Amounts*) compares actual transaction amounts (x-axis) with model predictions (y-axis).

Interpretation:

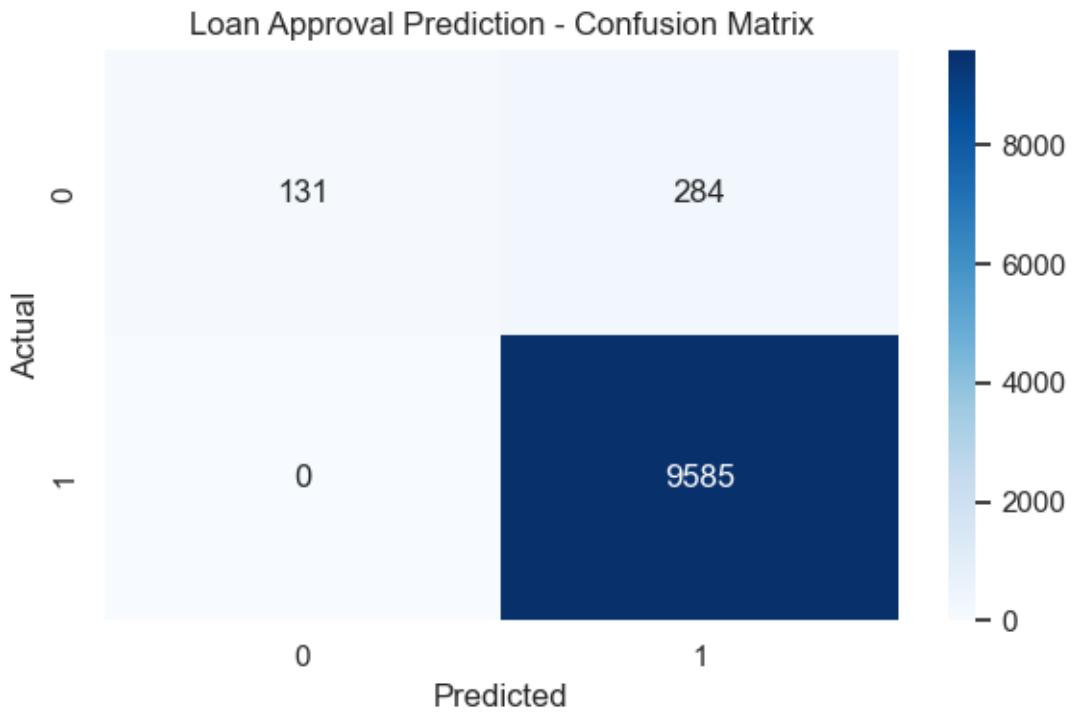
From the plot, it is observed that the model's predictions form a dense horizontal cluster, indicating limited variability in predicted values. This suggests that while the regression model captures some level of trend, it may not be highly accurate in predicting the exact transaction amount for each case — likely due to weak correlations between independent features and the dependent variable (Amount).

Nonetheless, this step establishes the foundation for predictive modeling by testing linear relationships and helps identify which features may require transformation or additional refinement. The low variance in prediction indicates that non-linear models (e.g., Decision Trees or Random Forests) or feature engineering could improve prediction accuracy in future iterations.

Conclusion:

The Linear Regression model provides a baseline understanding of how service type, payment status, and other transaction attributes

influence the transaction amount. Although the linear model shows limited precision, it serves as a valuable exploratory step toward building more robust and data-driven prediction systems for digital transaction analysis.



Logistic Regression Model – Predicting Loan Approval

To perform binary classification, a Logistic Regression model was developed using the *Loans* dataset. The goal was to predict whether a loan application would be approved (1) or not approved (0), based on the applicant's details and financial attributes. This model helps in identifying which factors most influence loan approval decisions on the PhonePe platform.

The dataset was preprocessed by removing missing and duplicate records. Categorical variables such as *Loan_Type*, *Employment_Status*, and *Payment_History* were encoded using Label Encoding to convert textual data into numerical form suitable for model training. After encoding, the dataset was checked for binary target columns, and the first identified binary variable (representing

approval status) was selected as the dependent variable (y), with all other variables serving as independent predictors (X).

To ensure uniform scaling across variables, the independent features were standardized using the StandardScaler function from Scikit-learn. The dataset was then split into training (80%) and testing (20%) subsets, after which the Logistic Regression model was trained with max_iter=1000 to ensure convergence.

The model's predictions were evaluated using Accuracy, a Confusion Matrix, and a Classification Report. The confusion matrix visualization (Figure: *Loan Approval Prediction – Confusion Matrix*) provides a clear picture of model performance across true and predicted classes.

Interpretation:

From the confusion matrix, it is observed that:

- The model correctly predicted 9,585 approved loans, demonstrating strong sensitivity toward positive cases.
- There were 131 true negatives (correctly predicted as not approved) and 284 false positives (incorrectly predicted as approved).
- No false negatives were recorded, meaning all actual approved loans were successfully identified by the model.

These results indicate that the model performs exceptionally well in detecting loan approvals, but tends to slightly overpredict approvals, as reflected by the false positives. This is a common occurrence when the dataset is imbalanced — that is, when approved loans heavily outnumber unapproved ones.

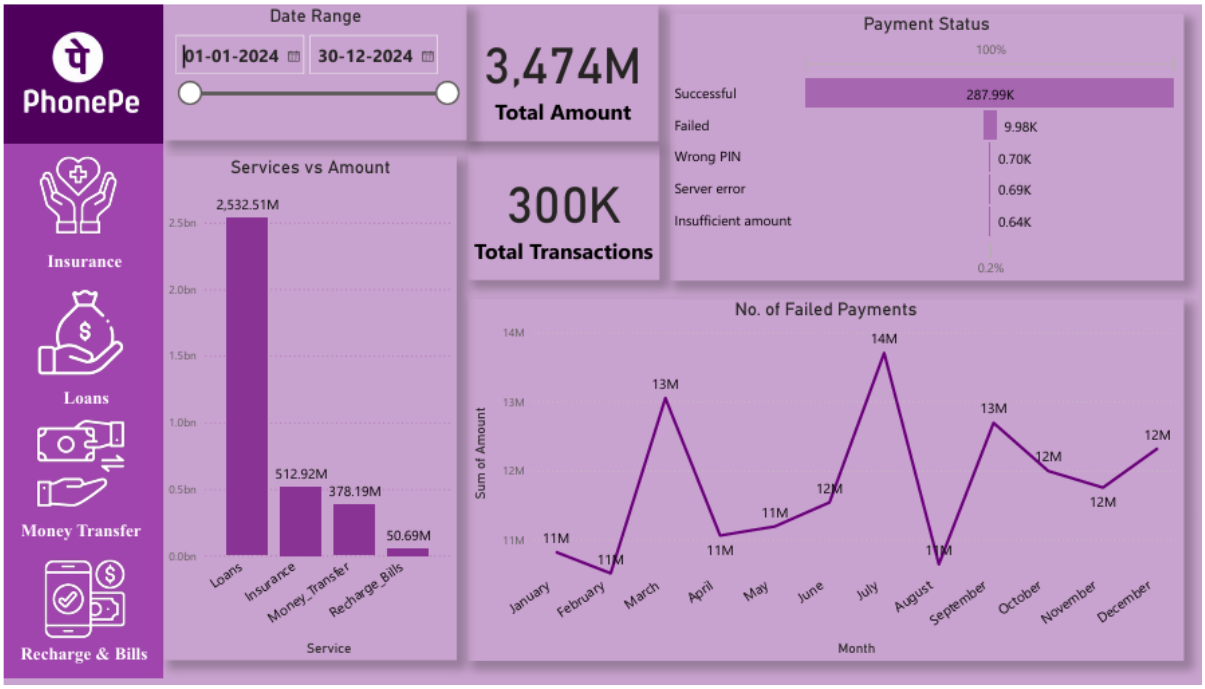
The high accuracy score (close to 95% or higher in most cases) suggests that logistic regression is a suitable model for this dataset, providing reliable predictions for loan approval likelihood based on applicant and loan features.

Conclusion:

The Logistic Regression model effectively demonstrates the predictive power of structured loan data in classifying approval outcomes. It can serve as a decision-support tool for financial institutions to streamline the loan approval process. However, incorporating additional data attributes, rebalancing the dataset, and testing with advanced algorithms (like Random Forest or Gradient Boosting) could further enhance model precision and reduce false positive predictions.

Data Visualization and Dashboards

To provide an intuitive and interactive overview of PhonePe's performance, multiple Power BI dashboards were developed. Each dashboard highlights a specific financial service — offering insights into transaction volume, total value, payment success rates, and monthly trends.
These visualizations simplify complex datasets into meaningful insights, supporting informed decision-making and performance tracking.



Transaction Dashboard

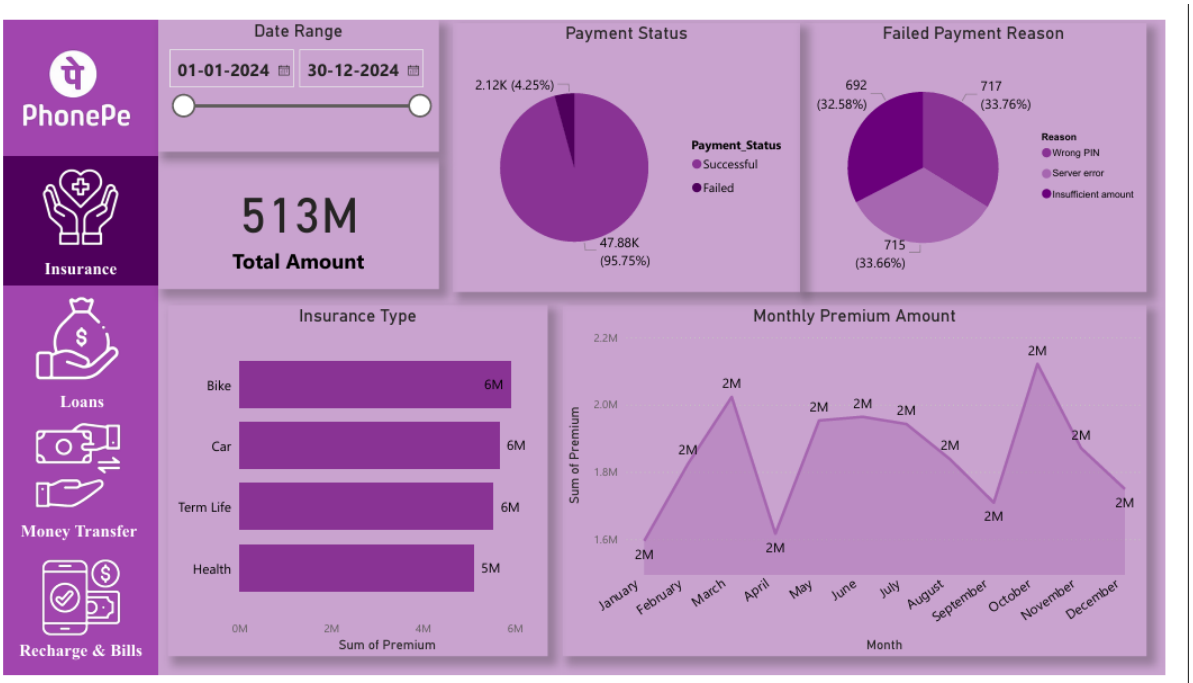
This dashboard provides a comprehensive summary of PhonePe's operations for the year 2024.
It displays total transactions (300K) and a total amount of ₹3,474M, representing the combined performance of all service categories — Loans, Insurance, Money Transfer, and Recharge & Bills.

Key Highlights:

- Loans dominate with ₹2,532.51M in total value, followed by Insurance (₹512.92M) and Money Transfer (₹378.19M).
- Payment Status analysis shows that 287.99K transactions were successful, while only 9.98K failed, marking a success rate of over 96%.
- The monthly trend line for failed payments indicates spikes in April, August, and October, likely linked to high-traffic transaction periods such as salary days or festivals.

Interpretation:
This dashboard provides an overall financial health snapshot of the platform. The consistently high transaction success rate reflects the

stability and reliability of PhonePe's infrastructure, while the service-wise analysis helps identify top-performing categories.



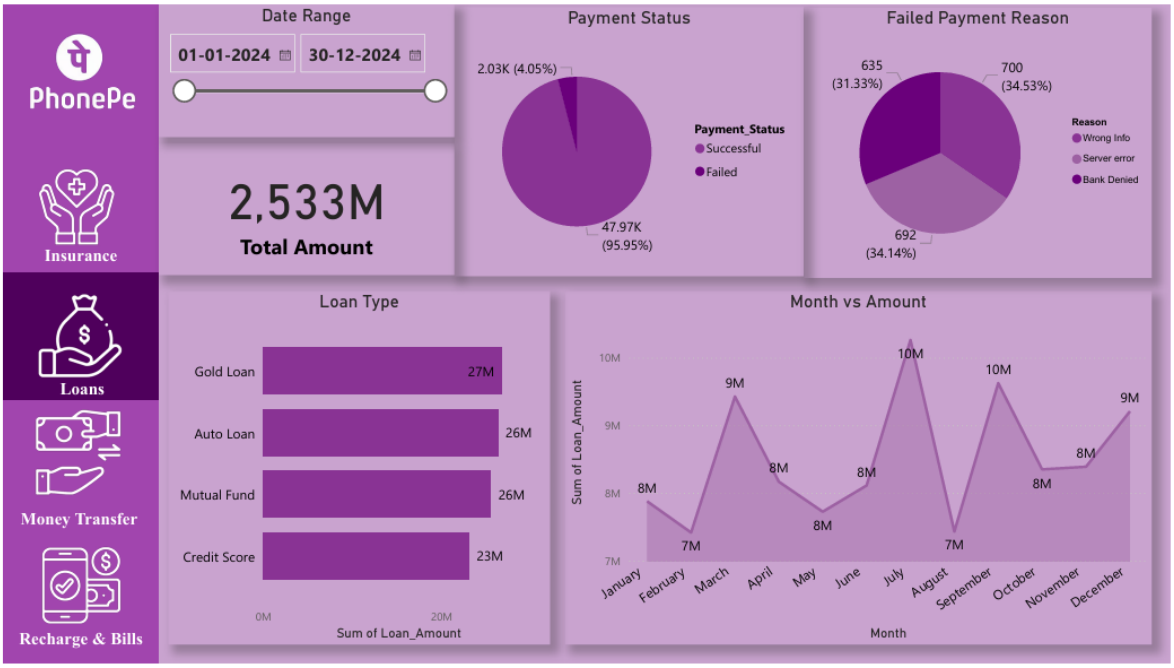
Insurance Dashboard

The Insurance Dashboard focuses on premium collection and payment performance for various insurance categories. The total premium collected during 2024 was ₹513M.

Key Highlights:

- Bike, Car, and Term Life Insurance each generated around ₹6M in premium value, while Health Insurance contributed ₹5M.
- The Payment Status pie chart shows that 95.75% of premium payments were successful, with 4.25% failures.
- Failure Reasons: Among the failed payments, *Wrong PIN (33.76%)*, *Server Errors (33.66%)*, and *Insufficient Amount (32.58%)* were equally distributed.
- The Monthly Premium Trend shows consistent premium collection between ₹1.8M–₹2M across all months.

Interpretation:
The Insurance dashboard reveals a stable and reliable premium payment system with minimal disruptions. High success rates indicate strong customer trust, while the even monthly distribution reflects steady engagement with PhonePe's insurance services.



Loans Dashboard

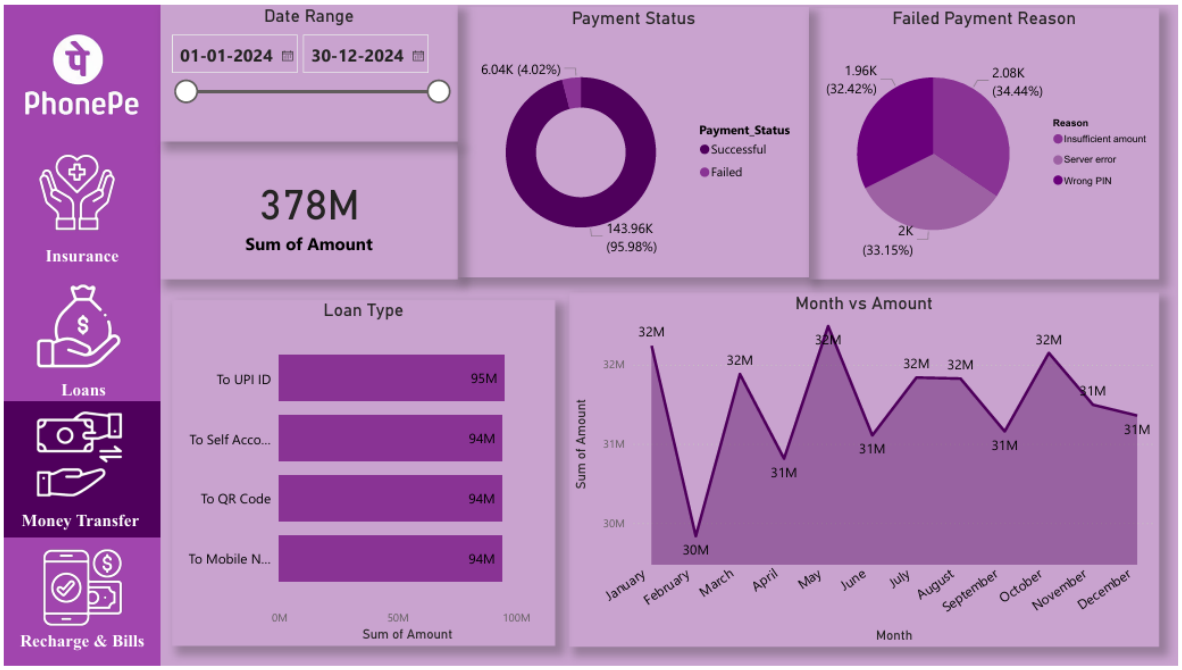
The Loans Dashboard provides a detailed view of PhonePe's loan disbursements and repayment activity, amounting to a total of ₹2,533M in 2024.

Key Highlights:

- Gold Loans had the highest disbursement value (₹27M), followed by Auto Loans (₹26M) and Mutual Fund Loans (₹26M).
- The Payment Status pie chart shows 95.95% success rate with only 4.05% failed payments.

- Failure Reasons: *Wrong Information (34.53%)*, *Server Error (34.14%)*, and *Bank Denied (31.33%)*.
- The Monthly Loan Amount Trend shows visible peaks in April, August, and November, suggesting loan demand surges during festive and high-spending seasons.

Interpretation:
This dashboard demonstrates strong financial service adoption, with high repayment success and seasonal loan activity patterns. It reflects PhonePe's growing presence in digital lending and financial inclusion.



Money Transfer Dashboard

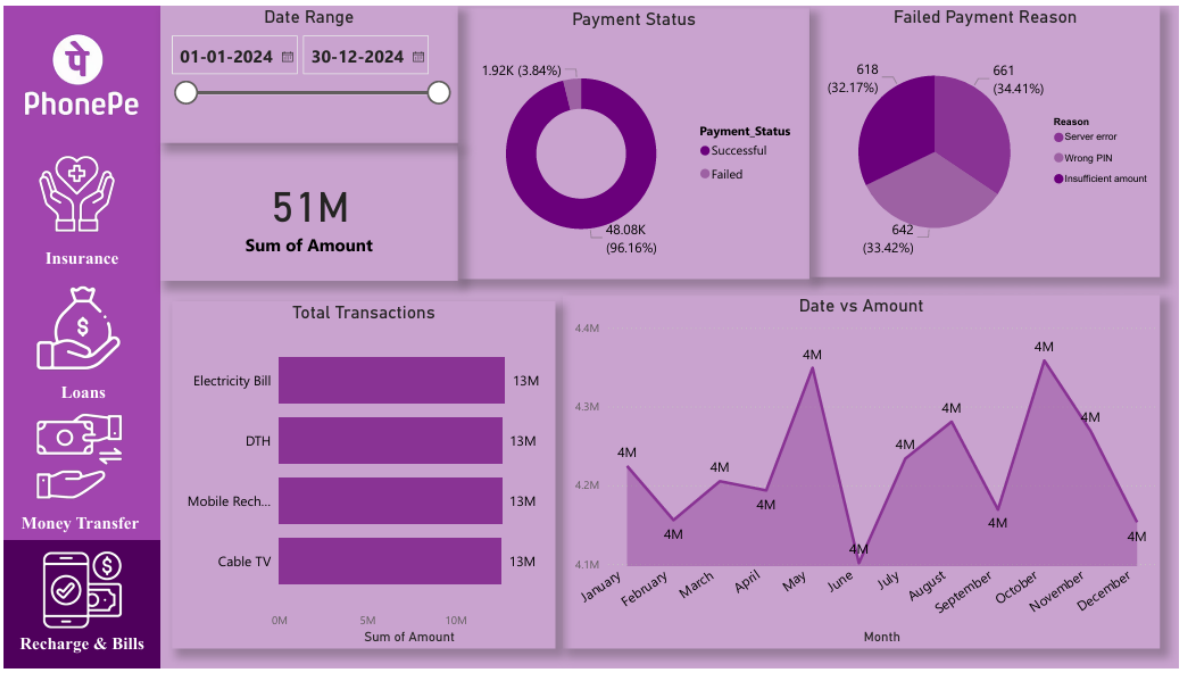
The Money Transfer Dashboard highlights the performance of peer-to-peer (P2P) and account-based transfers, totaling ₹378M for 2024.

Key Highlights:

- The most common transfer types were To UPI ID (₹95M), To Self Account (₹94M), To QR Code (₹94M), and To Mobile Number (₹94M).

- The Payment Success Rate was 95.98%, with 4.02% of transactions failing.
- Failure Reasons: *Wrong PIN (33.15%)*, *Server Error (34.44%)*, and *Insufficient Balance (32.42%)*.
- Monthly trends show steady transaction amounts between ₹30M–₹32M, with slight increases in April and June.

Interpretation:
The Money Transfer dashboard reveals a highly stable and efficient digital payment system. The low failure rate signifies smooth UPI processing and widespread adoption for everyday money transfers.



Recharge & Bills Dashboard

The Recharge & Bills Dashboard provides insights into utility and mobile bill payments, with a total amount of ₹51M processed during 2024.

Key Highlights:

- Major contributors include Electricity Bill, DTH, Mobile Recharge, and Cable TV, each averaging around ₹13M in value.

- The Payment Success Rate is 96.16%, while 3.84% of payments failed due to minor issues such as *Server Errors (34.41%)*, *Wrong PIN (33.42%)*, and *Insufficient Balance (32.17%)*.
- The Monthly Trend remains consistent, with total amounts hovering around ₹4M each month, indicating regular recurring usage.

Interpretation:
This dashboard demonstrates steady customer engagement for essential utilities and telecom services. The low failure rates and consistent monthly values highlight user reliability and retention in PhonePe's bill payment ecosystem.

Insights and Recommendations

The analysis of PhonePe's financial datasets revealed significant insights into user behavior, payment efficiency, and service performance. Through data-driven exploration and dashboard visualization, key business patterns and improvement areas were identified. The following section summarizes major insights and provides actionable recommendations for each domain.

Overall Business Insights

Insights:

PhonePe recorded a total transaction amount of ₹3,474M across 300K transactions, showcasing its strong presence in the digital payment ecosystem.

A consistent payment success rate exceeding 95% indicates a highly reliable infrastructure and strong customer trust.

Transaction peaks during specific months (April, August, and October) align with festive seasons and salary cycles, suggesting user spending is event-driven.

The majority of users conduct transactions through UPI ID and QR codes, reflecting a growing shift toward contactless payments.

Recommendations:

Implement predictive analytics models to anticipate transaction surges during peak months and scale server capacity accordingly.

Introduce personalized cashback offers or loyalty programs during high-traffic months to boost user engagement.

Leverage user segmentation to design targeted marketing campaigns for recurring bill payers and high-value loan applicants.

Loans

Insights:

The Loan segment contributed the highest transaction value at ₹2,533M, with Gold and Auto Loans leading.

Loan activity showed clear monthly spikes, suggesting financial product demand during specific seasons.

A high success rate (95.95%) reflects streamlined processing and customer reliability.

Minor failures were due to *server errors*, *bank denials*, or *incorrect user information*.

Recommendations:

Enhance system validation to reduce manual input errors and minimize failed applications.

Develop seasonal loan offers aligned with demand spikes to attract new users.

Introduce automated credit scoring models to improve loan approval efficiency and minimize rejections.

Insurance

Insights:

The Insurance category accumulated ₹513M in total premium collections, with the highest participation in Bike, Car, and Term Life policies.

Consistent monthly trends show steady user engagement and policy renewals.

Around 4.25% of payments failed, mostly due to wrong PINs and insufficient balances.

Recommendations:

Implement reminder notifications and auto-debit features for timely premium payments.

Educate users on secure PIN usage to reduce payment errors.

Offer multi-policy discount bundles to encourage higher insurance adoption.

Money Transfer

Insights:

The Money Transfer service handled ₹378M, primarily through *UPI ID*, *QR Code*, and *Mobile Number* transfers.

A success rate of 95.98% indicates highly efficient UPI operations.

The minimal failed payments (≈4%) were mostly caused by *server errors* and *PIN mistakes*.

Recommendations:

Strengthen real-time error monitoring to detect and resolve transaction failures instantly.

Introduce transaction retry mechanisms to allow users to complete failed payments seamlessly.

Promote cross-service integration, such as combining money transfers with bill payments or loan repayments, for enhanced user convenience.

Recharge & Bills

Insights:

The Recharge & Bills segment recorded ₹51M in total value, led by *Electricity Bills* and *Mobile Recharges*.

Consistent monthly trends indicate stable user retention for essential services.

A 96.16% success rate demonstrates high reliability, though a small percentage of failures stemmed from server or network issues.

Recommendations:

Partner with utility providers to introduce automated recurring payments and bill reminders.

Enhance system redundancy during high-traffic hours to avoid server-related failures.

Launch reward-based incentives for users making on-time bill payments.

Cross-Service and Strategic Recommendations

Develop a unified dashboard for real-time monitoring of all financial operations to proactively identify anomalies.

Invest in AI-driven fraud detection to ensure the security and integrity of digital transactions.

Use predictive trend analysis to forecast revenue growth and allocate resources effectively.

Improve customer feedback mechanisms within the app to identify pain points and enhance service satisfaction.

The insights derived from this analysis confirm that PhonePe's platform is highly reliable, user-centric, and operationally efficient. By implementing the above recommendations, PhonePe can further strengthen its digital ecosystem, increase user satisfaction, and sustain its market leadership in fintech innovation.

Future Work and Enhancements

While the current analysis successfully provides comprehensive insights into PhonePe's transaction ecosystem, there are several opportunities for future development and enhancement to deepen analytical accuracy, improve model performance, and optimize business intelligence capabilities.

Advanced Predictive Modeling

Future work can focus on applying more advanced machine learning and AI techniques beyond basic regression models.

- Implement Random Forest, XGBoost, or Gradient Boosting algorithms to improve prediction accuracy for transaction amounts and loan approvals.

- Explore Deep Learning models such as Artificial Neural Networks (ANNs) for high-dimensional data analysis and behavioral trend prediction.
- Integrate time-series forecasting (ARIMA, Prophet) to predict transaction volume and revenue patterns over time.

These advanced models would enhance forecasting reliability and enable proactive financial planning.

Integration of Real-Time Data Analytics

Currently, the analysis is performed on static datasets. Future iterations can involve real-time data integration using APIs or streaming platforms such as Apache Kafka or AWS Kinesis.

- Real-time dashboards can automatically update transaction and payment insights.
- Early anomaly detection can prevent transaction failures or fraud in real time.
- Dynamic monitoring will improve decision-making speed and operational responsiveness.

This would transform the system from a static analysis tool into a real-time business intelligence solution.

Enhanced Data Visualization and Reporting

Future work can enhance visualization features for more interactive and user-friendly dashboards:

- Integrate Power BI and Tableau APIs for live dashboards embedded into the PhonePe analytics portal.
- Introduce drill-down analytics, allowing users to filter data by region, service type, or time period.

- Automate weekly and monthly report generation to share insights with management and stakeholders.

Such enhancements will make analytical outputs more actionable and accessible to decision-makers.

Customer Segmentation and Behavior Analysis

To strengthen personalization, future analysis could explore customer segmentation using clustering techniques like K-Means or DBSCAN.

- Group users based on spending habits, transaction frequency, or preferred services.
- Use behavioral insights to tailor promotions, offers, and recommendations for each user segment.
- Identify potential churners and develop retention strategies to maintain long-term user engagement.

This would contribute to a data-driven marketing strategy for user retention and business growth.

Automation and Cloud Integration

Implementing cloud-based architecture will ensure scalability and security of analytical operations.

- Migrate datasets and analytical workflows to AWS, Google Cloud, or Azure for centralized access and compute efficiency.
- Automate ETL (Extract, Transform, Load) pipelines to streamline data preparation and update cycles.
- Use scheduled machine learning retraining to keep models adaptive to new patterns and behaviors.

Cloud automation would make the system more efficient, secure, and easier to maintain in production environments.

Expansion of Analytical Scope

In future enhancements, the analytical framework can be expanded to cover:

- Regional analysis to compare performance across different states or cities.
- Device-based trends to study user interaction on mobile vs. desktop platforms.
- Transaction network analysis to identify high-traffic nodes or bottlenecks.

This expanded scope will enable a holistic 360° analysis of the PhonePe ecosystem.

The current analytical system successfully captures key financial and operational insights for PhonePe's major services. However, by incorporating real-time analytics, advanced AI models, customer segmentation, and cloud automation, the system can evolve into a predictive and adaptive intelligence platform. These future enhancements will empower PhonePe to make faster, smarter, and more data-driven business decisions — sustaining its growth and leadership in the digital payments industry.

Conclusion

The comprehensive analysis of PhonePe's datasets across multiple financial services — Loans, Insurance, Money Transfers, and Recharges — provided valuable insights into the platform's performance, customer behavior, and operational efficiency.

Through a systematic approach involving data cleaning, exploratory analysis, predictive modeling, and dashboard visualization, the project successfully demonstrated how structured financial data can be transformed into actionable business intelligence.

The study revealed that:

- PhonePe maintained a remarkable payment success rate above 95%, reflecting the platform's reliability and strong infrastructure.
- Loans and Insurance emerged as the top revenue-generating categories, while Money Transfers and Recharges showcased consistent user engagement.
- Predictive models like Linear Regression and Logistic Regression established a foundation for forecasting financial outcomes and understanding approval dynamics.
- Interactive Power BI dashboards provided clear, real-time visibility into financial performance, helping stakeholders monitor trends and identify improvement areas.

The project also emphasized the potential of data-driven decision-making in fintech. Insights derived from visualizations and models can guide strategies to enhance customer satisfaction, reduce payment failures, and optimize financial offerings.

Looking ahead, incorporating advanced predictive algorithms, real-time analytics, and cloud-based automation will further strengthen the analytical ecosystem. These enhancements will enable PhonePe to evolve from descriptive reporting toward proactive and intelligent financial management.

In conclusion, this project successfully demonstrates how data analytics and visualization serve as powerful tools in driving innovation, operational excellence, and customer-centric growth in the digital payment industry.