PhonePe EDA Project - Revision Notes

Table of Contents

Project Overview

Objective: Perform comprehensive Exploratory Data Analysis (EDA) on PhonePe Pulse transaction data to understand digital payment trends across Indian states.

Business Context: PhonePe is one of India's leading digital payment platforms. Understanding transaction patterns helps in: - Market segmentation and targeting - Regional expansion strategies - Product development - Risk assessment

Technical Stack

Libraries Used:

import pandas as pd # Data manipulation and analysis import numpy as np # Numerical computations import plotly.express as px # Interactive visualizations import matplotlib.pyplot as plt # Static plots import seaborn as sns # Statistical visualizations

Why These Libraries?

- **Pandas**: Essential for data cleaning, filtering, grouping, and transformation - **Plotly Express**: Creates interactive, professional-looking charts - **NumPy**: Efficient numerical operations - **Matplotlib/Seaborn**: Alternative visualization options

Data Understanding

Dataset Structure:

- **Source**: phonepe_transaction.csv - **Key Columns**: - state: Indian states/UTs - year: Transaction year - quarter: Q1, Q2, Q3, Q4 - transaction_category: Payment type (P2P, Merchant, etc.) - transaction_count: Number of transactions - transaction_amount: Total transaction value

Data Characteristics:

- **Rows**: Multiple years of quarterly data - **Granularity**: State-wise, quarterly aggregated data - **Special Values**: 'All States' (aggregated row to exclude)

Key Analysis Steps

1. Data Loading & Initial Exploration

Load data df = pd.read_csv('phonepe_transaction.csv')

Basic exploration

df.shape # Dataset dimensions df.head() # First few rows df.describe() # Statistical summary df.isnull().sum() # Missing values check

2. Data Cleaning & Preprocessing

Critical step: Standardize column names df.columns = df.columns.str.lower()

Filter out aggregated data

df_filtered = df[df['state'] != 'All States']

3. State-wise Analysis

Top states by transaction amount top_states = df_filtered.groupby('state')['transaction_amount'].sum().sort_values(ascending=False)

Convert to crores for readability

top_states_crores = (top_states / 1e7).round(2)

4. Temporal Analysis

Quarterly trends quarterly_data = df_filtered.groupby('quarter')['transaction_amount'].sum()

5. Transaction Type Analysis

Category distribution category_counts = df_filtered.groupby('transaction_category')['transaction_count'].sum()

6. Advanced Metrics

Average Transaction Value (ATV) state_summary['average_transaction_value'] = (state_summary['transaction_amount'] / state_summary['transaction_count'])

7. Statistical Analysis

Correlation analysis correlation_matrix = correlation_data.corr()

Important Code Snippets

Data Filtering Pattern:

Always filter out 'All States' for state-wise analysis df_filtered = df[df['state'] != 'All States']

Groupby Aggregation Pattern:

Multi-column aggregation state_summary = df_filtered.groupby('state').agg({ 'transaction_amount': 'sum', 'transaction_count': 'sum' }).reset_index()

Visualization Pattern:

Interactive bar chart with Plotly fig = px.bar(data, x='column1', y='column2', title='Chart Title', labels={'column1': 'X Label', 'column2': 'Y Label'}) fig.show()

Unit Conversion:

Convert to crores (10 million) amount_crores = (amount / 1e7).round(2)

Key Findings & Insights

Geographic Insights:

- **Top States**: Karnataka, Maharashtra, Tamil Nadu dominate both volume and value - **ATV Variation**: Significant differences in average transaction values across states - **Market Concentration**: Few states contribute majority of transaction value

Temporal Patterns:

- **Growth Trend**: Consistent year-over-year growth in digital payments - **Seasonal Effects**: Q4 often shows higher transaction activity - **Quarterly Consistency**: Relatively stable patterns across quarters

Transaction Behavior:

- **Dominant Categories**: P2P payments and merchant transactions are primary use cases - **Volume vs Value**: Strong positive correlation between transaction count and amount - **Regional Preferences**: Different states show varying transaction type preferences

Statistical Relationships:

- **Strong Correlation**: Transaction amount ↔ Transaction count (r ≈ 0.9+) - **Moderate Correlation**: Amount/Count ↔ Average Transaction Value - **Business Implication**: Higher volume markets also tend to be higher value markets

Common Issues & Solutions

Issue 1: KeyError with Column Names

Problem: KeyError: 'State' or similar **Solution**: Convert all column names to lowercase for consistency

```
    df.columns = df.columns.str.lower()
```

Issue 2: Blank Visualizations

Problem: Charts appear empty **Solution**: Check data filtering and ensure correct column references

Issue 3: Large Number Display

Problem: Numbers like 1000000000 are hard to read **Solution**: Convert to appropriate units (crores, lakhs)

Issue 4: Overlapping Labels

Problem: State names overlap in charts **Solution**: Use log scale or limit to top N states

Interview Questions & Answers

Q1: "Walk me through your EDA process"

Answer: 1. Started with data loading and basic exploration (shape, head, describe) 2. Identified data quality issues and performed cleaning 3. Conducted univariate analysis (individual column distributions) 4. Performed bivariate analysis (relationships between variables) 5. Created visualizations to communicate insights 6. Calculated business metrics like ATV 7. Performed correlation analysis for statistical validation

Q2: "What challenges did you face in this project?"

Answer: - Data consistency issues with column naming (case sensitivity) - Handling aggregated 'All States' row that skewed analysis - Choosing appropriate visualization scales for large numerical ranges - Balancing between detailed analysis and clear communication

Q3: "What business insights can you derive?"

Answer: - Market prioritization: Focus on top-performing states for expansion - Product strategy: P2P and merchant payments are core use cases - Regional customization: Different states show different usage patterns - Growth opportunities: Significant variation in ATV suggests optimization potential

Q4: "How would you extend this analysis?"

Answer: - Time series forecasting for future transaction prediction - Cohort analysis to understand user behavior over time - Geographic visualization using choropleth maps - Customer segmentation analysis - Anomaly detection for fraud prevention

Extensions & Improvements

Machine Learning Applications:

1. **Time Series Forecasting**: Predict future transaction volumes using SARIMA/LSTM 2. **Clustering**: Group states by transaction behavior patterns 3. **Classification**: Predict transaction category based on other features 4. **Anomaly Detection**: Identify unusual transaction patterns

Advanced Analytics:

1. **Cohort Analysis**: Track user behavior over time 2. **Market Basket Analysis**: Understand transaction type combinations 3. **Geographic Analysis**: Choropleth maps and spatial statistics 4. **A/B Testing Framework**: Compare different regions or time periods

Technical Improvements:

1. **Data Pipeline**: Automate data loading and cleaning 2. **Interactive Dashboard**: Create real-time monitoring dashboard 3. **API Integration**: Connect to live PhonePe data sources 4. **Performance Optimization**: Handle larger datasets efficiently

Quick Reference Commands

Essential Pandas Operations:

Data exploration df.info() # Data types and memory usage df.nunique() # Unique values per column df.value_counts('column') # Frequency counts

Data manipulation

df.groupby('col').agg({'col2': ['sum', 'mean', 'count']}) df.sort_values('col', ascending=False) df.reset_index(drop=True)

Filtering

df[df['col'] > value] df[df['col'].isin(['val1', 'val2'])]

Essential Plotly Commands:

Bar chart px.bar(df, x='col1', y='col2', color='col3', title='Title')

Pie chart

px.pie(df, names='col1', values='col2', title='Title')

Correlation heatmap

px.imshow(corr_matrix, text_auto=True, color_continuous_scale='RdBu_r')

Summary

This EDA project demonstrates: - **Technical Skills**: Data manipulation, visualization, statistical analysis - **Business Acumen**: Translating data into actionable insights - **Problem-Solving**: Handling real-world data challenges - **Communication**: Presenting findings clearly and effectively

Key Takeaway: The project showcases the complete data science workflow from raw data to business insights, making it ideal for demonstrating analytical capabilities in interviews.

Last Updated: July 27, 2025 Project Repository: https://github.com/nish0753/phonepe_project