Data Science Capstone Project

Google Play Store App Data Analysis

Assignment-2

Design Document

Written by: Ray Varghese

Professor: Dr. Monireh Reza

Table of contents

1. Introduction

The Google Play store is a global marketplace where users activate and install applications making millions of transactions each day. This data analysis and visualization gives an insight on the application characteristics or metrics- Rating, Installs, Revenue, Size, Reviews and User Sentiments which empowers the app developers, marketing, product managers and technology investors to make data driven decisions on understanding user expectations and focus on continuous improvement.

2. Scope
   The project covers the end-to-end data science lifecycle:

   - Data collection
   - Preprocessing
   - ETL (Extract, Transform, Load)
   - Modeling
   - Dashboard development
   - Target users are [stakeholders/business users/data scientists].

3. Mission

   To analyze user reviews, app features, and performance indicators across the Google Play Store ecosystem in order to uncover meaningful insights that support product enhancement, user satisfaction, and market positioning.

4. Objectives

   4.1 App Volume Analysis:

   Understand the distribution and availability of apps on the Google Play Store based on their characteristics.

   - Which app categories and genres have the highest number of available apps?
   - What proportion of apps are free versus paid across different categories?
   - How are apps distributed by content rating (e.g., Everyone, Teen, Mature)?

   4.2 App Size, Price, and Revenue Metrics:

Analyze the financial and structural aspects of apps to identify pricing trends and potential revenue patterns.

- What is the average price and size of paid apps across different genres?
- Which categories generate the highest estimated revenue from paid apps?
- Is there a relationship between app size and price?

4.3 App Installation and User Rating Behavior:

Assess how users interact with apps based on installation trends and user ratings.

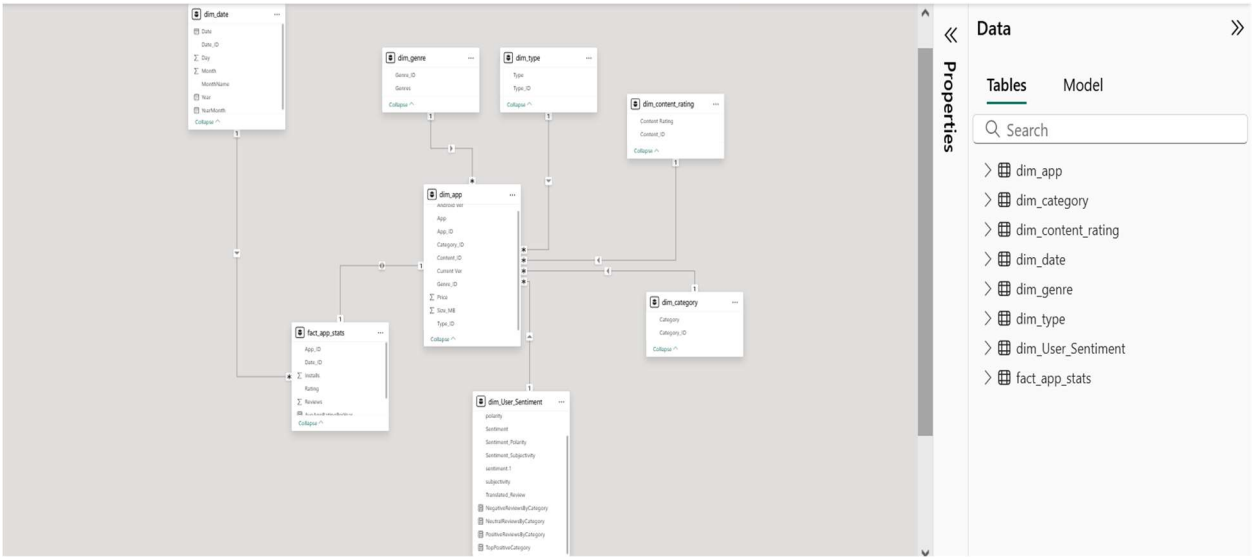
- Which categories and genres receive the highest number of installs?
- How do average user ratings vary by app category or type (free vs paid)?
- Is there a correlation between the number of installs and average user ratings?

4.4 User Sentiment Analysis:

Evaluate user reviews to identify overall satisfaction and concerns expressed by users.

- What is the overall sentiment distribution of user reviews (positive, neutral, negative)?
- Which app categories have the highest proportion of positive or negative reviews?
- What are the most common themes or keywords in negative reviews?

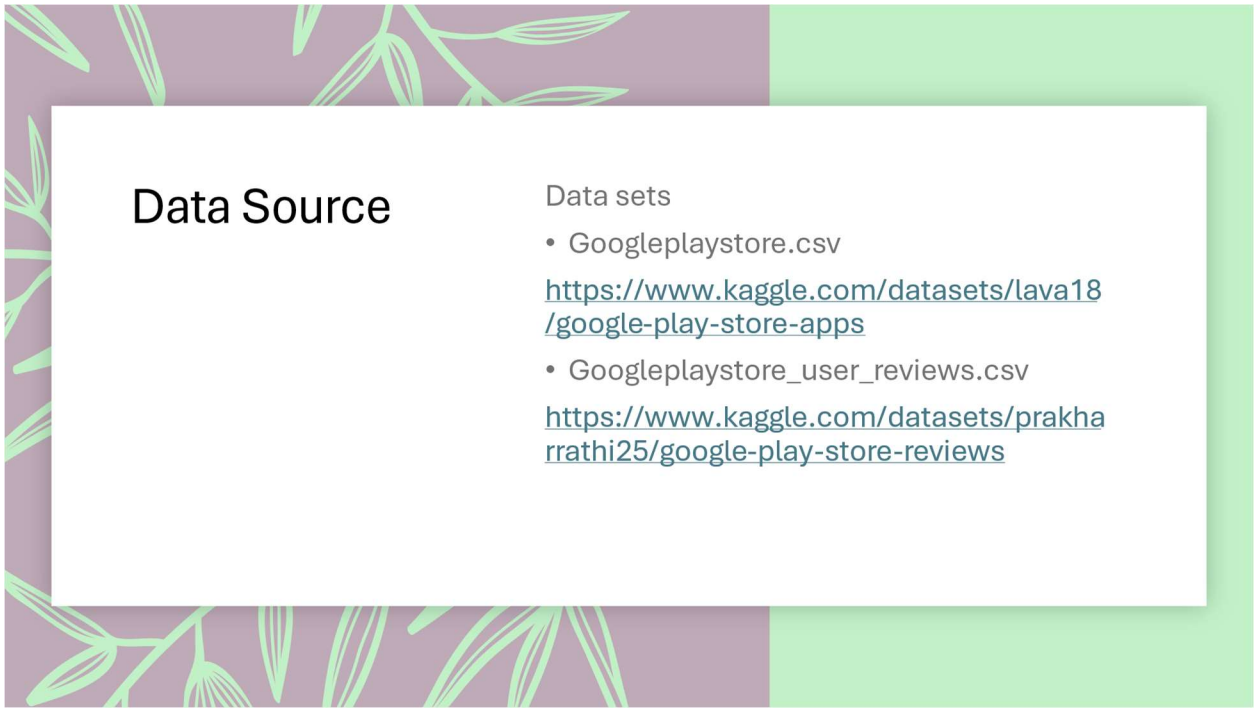
5. Entity Relationship Diagram



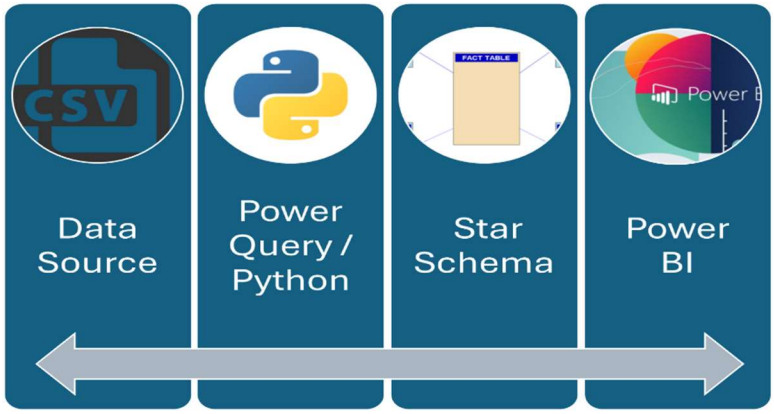
6. Data source: [e.g., Google Play Store App data & user reviews / Kaggle dataset)

File format: CSV, Excel

Frequency: Static or real-time ingestion



7. System Overview and Architecture



Ingests raw data in CSV format into Power BI,

Load and transform data for cleaning null values and outliers

Model it into a star schema with fact and dimension tables

Visualizes KPIs via interactive dashboards using DAX, Python, and Power Query.

8. Data Storage

   Power BI data model

   Intermediate storage (Excel/SQL Server/Azure Blob)

9. ETL – Power Query, Python

   Power Query: Used for loading, removing nulls, filtering columns, remove outliers

   Python: For advanced cleaning (e.g., sentiment analysis, text preprocessing)

10. Modelling & Visualization

    Power BI: Data Model (Star Schema)

    Visualizations include bar charts, line graphs, pie charts, slicers, KPI cards

11. System Components

    Data Ingestion Layer- Power M Query

    Data Cleaning & Transformation Layer- Use Python data pre-processing

    Data Model- Make an entity relationship model with fact and dimension tables

    Visualization Layer- Interactive dashboard with slicers and filters to drill up and down for the user interface.

12. Data Management

    Handling missing data- Remove columns with missing values, and remove null values. It also involves removal of duplicates from the dimension tables to maintain one to many relationships.

13. Data Modelling

    Star schema with fact and dimension tables

    Relationships based on primary and foreign keys

14. Schema Design

Fact Table

Table Name	Description	Key Attributes
fact_app_stats	Stores measurable app-related metrics for analysis	App_ID, Date_ID, Installs, Rating, Reviews, Revenue_Estimation

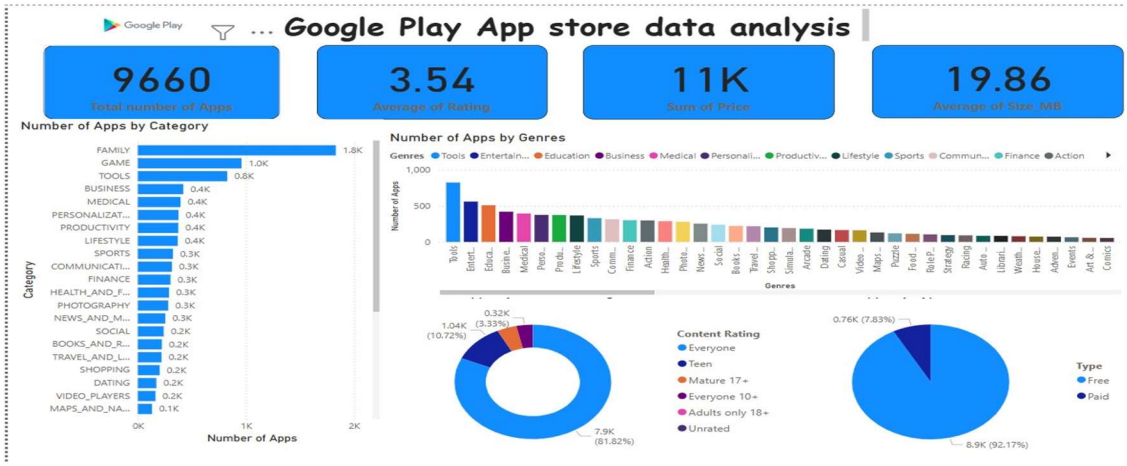
Dimension Tables

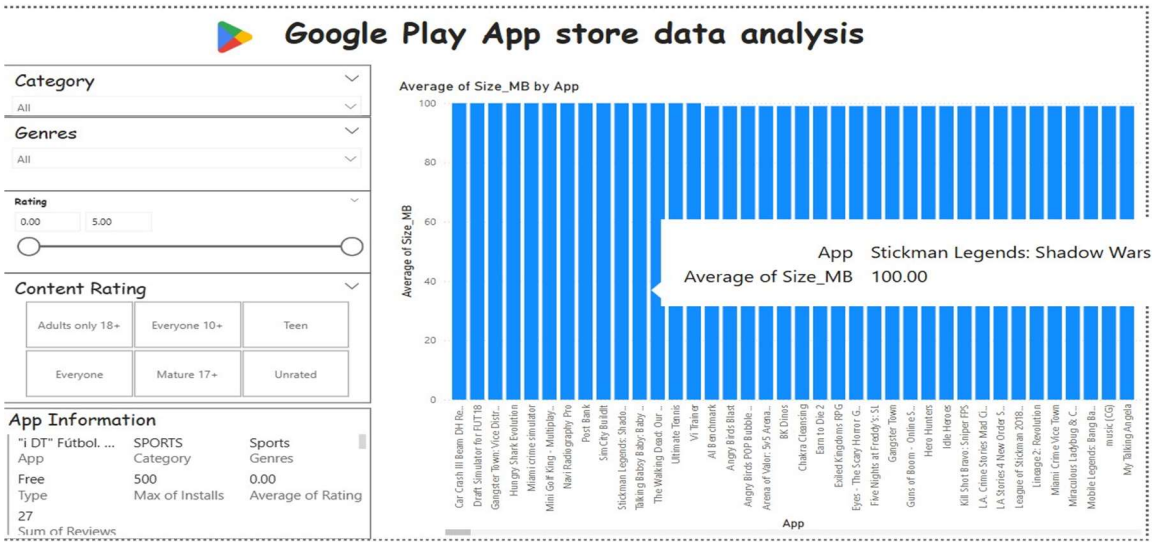
Table Name	Description	Key Attributes
dim_app	Stores detailed app metadata	App_ID, App Name, Category_ID, Content_ID, Genre_ID, Type_ID, Price, Size_MB
dim_category	Defines the classification of apps	Category_ID, Category Name (e.g., Family, Games, Tools)
dim_content_rating	Stores app age suitability and content restrictions	Content_ID, Content Rating (e.g., Everyone, Teen, Mature 17+)
dim_date	Stores time-based attributes to track trends over time	Date_ID, Day, Month, Year, MonthName, YearMonth
dim_genre	Categorizes apps based on genres	Genre_ID, Genre Name (e.g., Action, Productivity, Arcade)
dim_type	Defines app monetization models (Free vs Paid)	Type_ID, Type Name (Free, Paid)
dim_user_sentiment	Stores user reviews and sentiment analysis	Polarity, Sentiment Score, Sentiment Subjectivity, Translated Review, Review Counts by Category

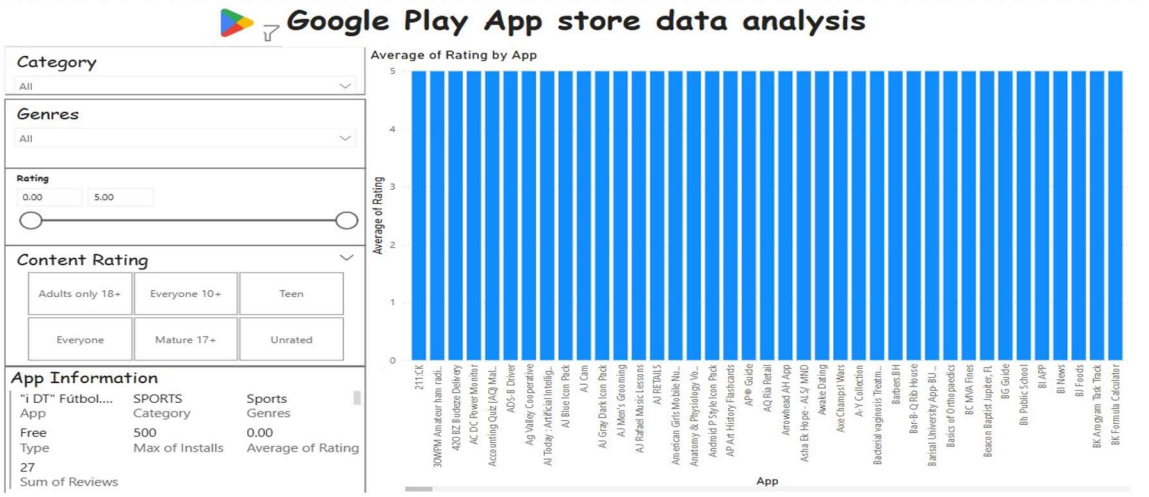
15. Visualization and User Interface- Dashboards

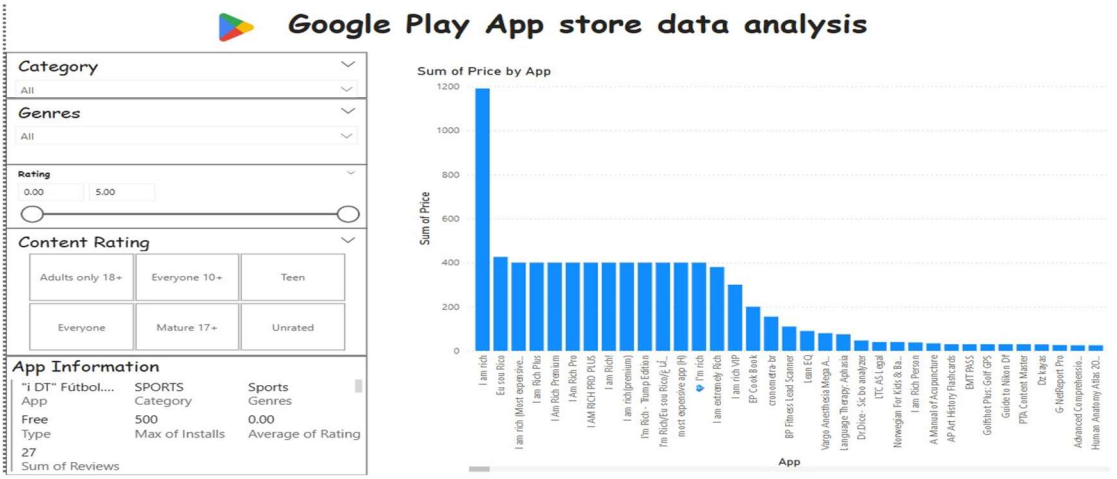
Clean, responsive layout in Power BI

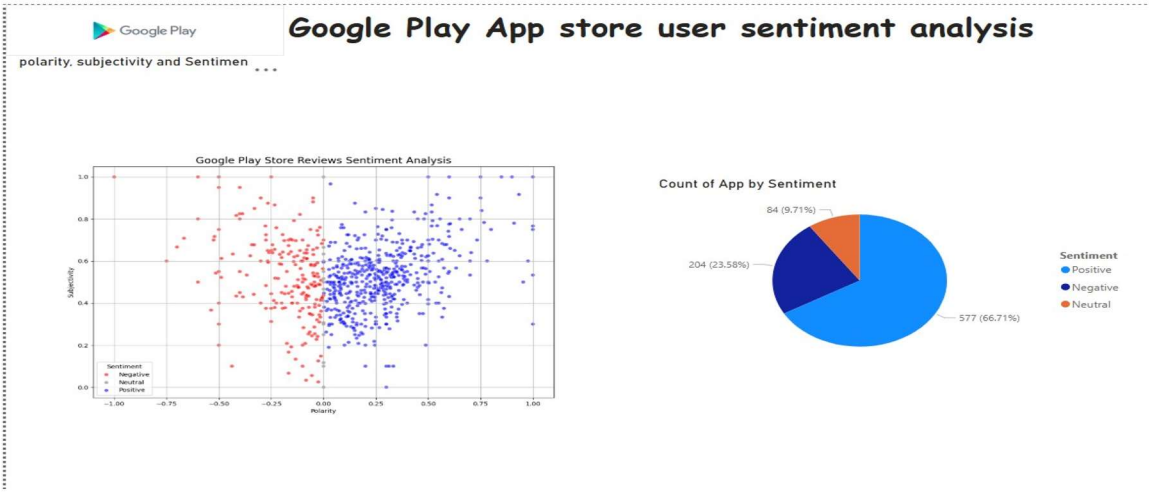
Clear navigation with bookmarks and drill-through

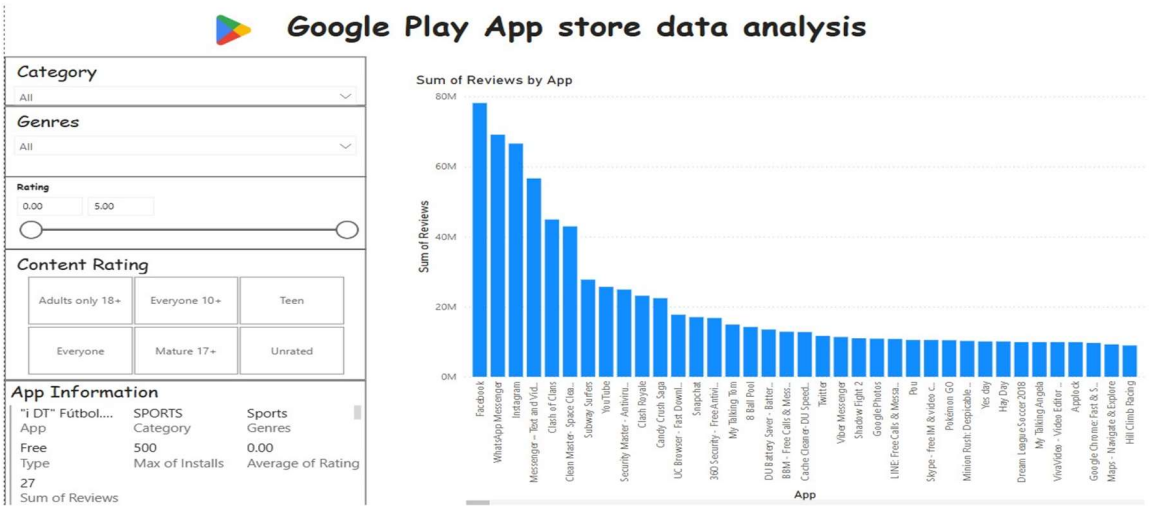



Google Play App store data analysis
Category
All
Genres
All
Rating
0.00
5.00
Content Rating
Adults only 18+
Everyone 10+
Teen
Everyone
Mature 17+
Unrated
App Information
"i DT" Fútbol. ...
SPORTS
Category
Sports
Genres
App
Free
Type
500
Max of Installs
0.00
Average of Rating
27
Sum of Reviews
Average of Size_MB by App
Average of Size_MB
App
Stickman Legends: Shadow Wars
Average of Size_MB
100.00


Google Play App store data analysis
Category
All
Genres
All
Rating
0.00
5.00
Content Rating
Adults only 18+
Everyone 10+
Teen
Everyone
Mature 17+
Unrated
App Information
"i DT" Fútbol....
SPORTS
Category
Sports
Genres
App
Free
Type
500
Max of Installs
0.00
Average of Rating
27
Sum of Reviews
Average of Rating by App
Average of Rating
App


Google Play App store data analysis
Category
All
Genres
All
Rating
0.00
5.00
Content Rating
Adults only 18+
Everyone 10+
Teen
Everyone
Mature 17+
Unrated
App Information
"i DT" Fútbol....
SPORTS
Category
Sports
Genres
App
Free
Type
500
Max of Installs
0.00
Average of Rating
27
Sum of Reviews
Sum of Price by App
Sum of Price
App





Slicers: for categories, genres, sentiment type

Filters: app size, rating range

Charts:

Bar Chart: Top categories by installs

Pie Chart: Sentiment distribution

Line Chart: Reviews over time

KPI Cards: Total installs, average rating

16. Features & Functionalities

Interactive dashboard

Dynamic filtering

Search functionality

Drill-down and tooltips

17. Technical Requirements

Software: Power BI Desktop, Python 3.x, Excel / SQL
Libraries: NumPy, Pandas, seaborn, matplotlib
Hardware: 8 GB+ RAM, 2GHz+ processor, 5GB free disk space

18. Milestones and Timeline

Milestone	**Date**
Dataset Identification	May 26, 2025
Data Cleaning and ETL	May 27-28, 2025
Star Schema Model development	May 29, 2025
Dashboard Development	May 30- June 02, 2025
Final Presentation	June 3, 2025

19. Conclusion

- This analysis uncovers key patterns in user sentiment and app behavior across the Google Play Store:
- Top App Categories: Games, Family, and Tools dominate the store by volume, indicating strong developer activity and user engagement.
- Leading Genres: Tools, Entertainment, and Education emerge as the most prevalent genres, pointing to a wide functional appeal across users.
- Installation Trends: Categories like Social, Communication, and Tools lead in total installs, reflecting high daily utility and user dependence.
- Ratings and App Size: Family and Games categories boast the highest average ratings and app sizes, suggesting content-rich experiences tailored for user engagement.
- Price Outlier: The app ‘I AM RICH’ stands out as the most expensive app on the store, exemplifying niche pricing strategies.
- Top Reviewed Categories: Social, Communication, Tools, and Games gather the largest number of user reviews, reflecting both high usage and feedback volume.
- Sentiment Insights: Sentiment analysis reveals that most user reviews are positive, signaling overall satisfaction with app quality and experience.
- These comprehensive insights enable stakeholders to: Prioritize feature development based on user sentiment and category popularity.
- Target high-performing app segments for investment or improvement.
- Optimize user experience and satisfaction using data-driven decisions