LACHOO MEMORIAL COLLEGE OF SCIENCE & TECHNOLOGY, JODHPUR

Affiliated To:-

JAI NARAIN VYAS UNIVERSITY, JODHPUR

2025-2026

Project Title: Google Play Store Analytics Dashboard

Team Members : Manvendra Singh Chouhan (243501117)
Lakshya Gangwani (243501104)
Mahesh Choudhary (243501113)

Course Code : CCBCA321

Course Name : BCA-Python Programming

Semester/year : Semester III -2025

Submitted to : Ms. Sanjana Khatri

Abstract

The Google Play Store hosts millions of applications used by billions of users worldwide. Understanding user preferences, app performance, and category-wise growth is essential for developers, marketers, and businesses. This project presents an analytical study of the Google Play Store dataset consisting of 9,639 apps, including features like ratings, installs, reviews, app category, type, and content rating.

The dataset contained several inconsistencies such as missing values, duplicate entries, text-formatted numbers, and uncleaned fields. We used Python (Pandas, NumPy, Matplotlib, Seaborn) to perform extensive data cleaning and preprocessing, including handling missing values, data type conversions, outlier removal, encoding categorical values, and transforming columns like “Installs”, “Reviews”, and “Size”.

After cleaning, the dataset was imported into Power BI, where a multi-page dashboard—Overview, Visual Explorer, and Key Insights—was built. The dashboard provides interactive analytics, including category dominance, rating distribution, install patterns, and correlations between features.

The analysis revealed many important insights: the Tools category has the highest number of apps, but Games and Communication dominate the install count. Ratings are generally positive, with most apps rated between 3.5 and 4.5. Install trends show that popularity is influenced by marketing and category type more than rating alone.

This project demonstrates a complete end-to-end workflow: data acquisition → cleaning → transformation → visualization → insights, using the combined strengths of Python and Power BI.

Introduction

1. Background

Mobile applications have become an essential part of human life, and the Google Play Store is one of the largest application marketplaces in the world. With thousands of new apps being added each year, understanding app performance trends is necessary for:

- Developers (app improvement)
- Companies (product strategy)
- Marketers (user targeting)
- Researchers (trend analysis)

The dataset used in this project provides a detailed structure of apps across various categories.

2. Problem Statement

The raw dataset had several issues:

- Missing ratings and reviews
- Duplicate app entries
- Install counts stored as text with symbols (e.g., “1,000+”)
- App sizes in inconsistent formats (MB/KB)
- Presence of outliers
- Non-standard categories

This made direct analysis impossible without proper cleaning.

3. Objectives

- Clean and preprocess the dataset using Python
- Analyze app popularity, user ratings, install behavior, and categories
- Build an interactive dashboard using Power BI
- Derive insights about app trends and user engagement patterns
- Demonstrate a complete data analytics workflow

Dataset Description

The Google Play Store dataset contains **9,639 rows** and **13 columns**, including:

Key Columns Used

1. **App** – App name
2. **Category** – App category (Tools, Communication, Games, etc.)
3. **Rating** – User rating (1.0–5.0)
4. **Reviews** – Total user reviews
5. **Size** – App size (MB, KB, varies)
6. **Installs** – Number of installs
7. **Type** – Free or Paid
8. **Price** – Price (for paid apps)
9. **Content Rating** – Suitable for Everyone, Teen, 18+, etc.
10. **Genres** – Genre classification
11. **Last Updated** – Last updated date

Challenges Identified

- Missing values in Rating
- Installs formatted like “10,000+”
- Reviews formatted as strings
- Size values inconsistent (e.g., “12M”, “450k”, “Varies with device”)
- Duplicates for the same app
- Extreme values in Reviews and Installs

These issues required cleaning before further analysis.

Methodology (Python Data Cleaning)

Tools Used

- **Python 3**
- **Pandas**
- **NumPy**
- **Matplotlib**

1. Removing Duplicates

```
df = df.drop_duplicates(subset='App', keep='first')
```

2. Handling Missing Values

- Replaced missing Ratings with the median
- Removed rows with too many missing fields

```
df['Rating'].fillna(df['Rating'].median(), inplace=True)
```

3. Cleaning Installs Column

Converted “1,000+” → 1000

```
df['Installs'] = df['Installs'].str.replace('+','').str.replace(',','').astype(int)
```

4. Cleaning Reviews Column

Converted to integer:

```
df['Reviews'] = df['Reviews'].astype(int)
```

5. Cleaning App Size

Converted MB/KB to float:

```
def size_to_mb(x):
    if 'M' in x:
        return float(x.replace('M',''))
    elif 'k' in x:
        return float(x.replace('k',''))/1024
    else:
        return None
df['Size'] = df['Size'].apply(size_to_mb)
```

6. Handling Outliers

Used IQR method to remove extreme values.

Feature Transformation

Categorical Encoding

Converted columns like “Type” and “Category” into numerical values.

Rating Binning

Ratings grouped into:

- Low (1–2.5)
- Medium (2.5–3.5)
- High (3.5–5)

Scaling Install Values

Normalized install data for scatter plot accuracy.

Grouping by Categories

To create dashboards like:

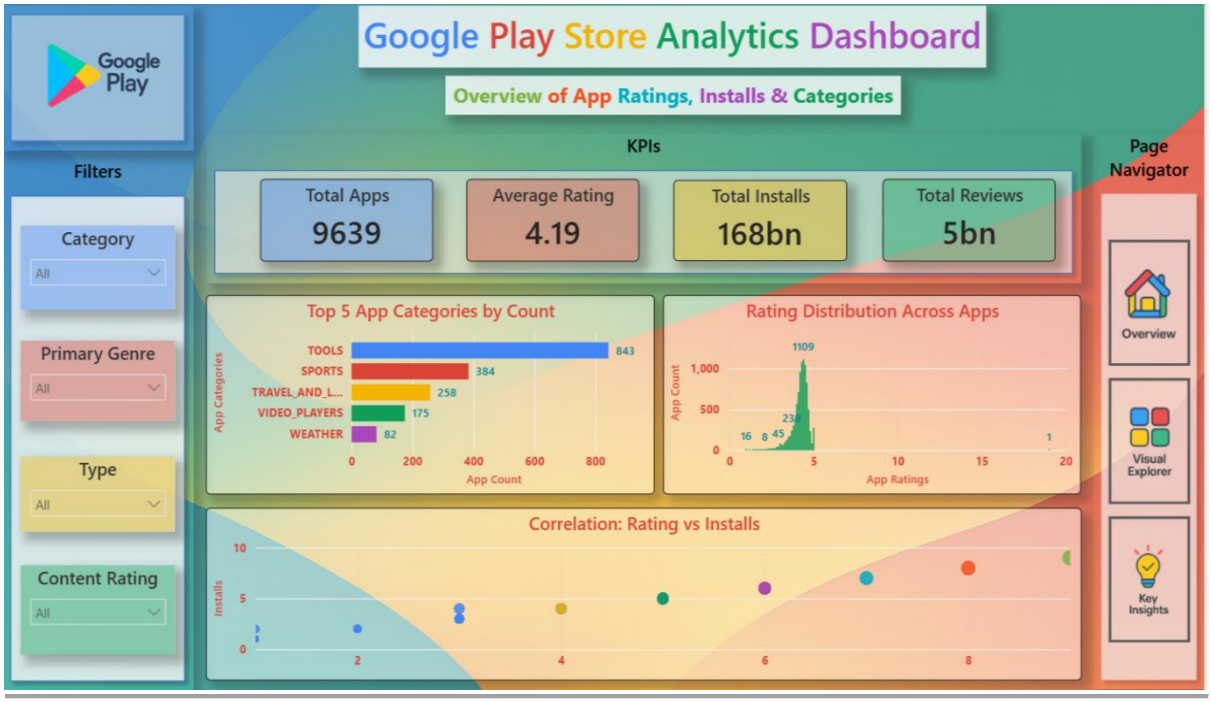
- Top Categories
- Installs by Category
- Rating Distribution

Visualization Tools

Power BI was used for:

- KPI Cards
- Treemaps
- Bar Charts
- Scatter Plots
- Filters & Slicers
- Page Navigation Buttons

Dashboard 1: Overview Page



KPIs Displayed

KPI	Value
Total Apps	9639
Average Rating	4.19
Total Installs	168 Billion
Total Reviews	5 Billion

Visuals

1. **Top 5 Categories by Count**

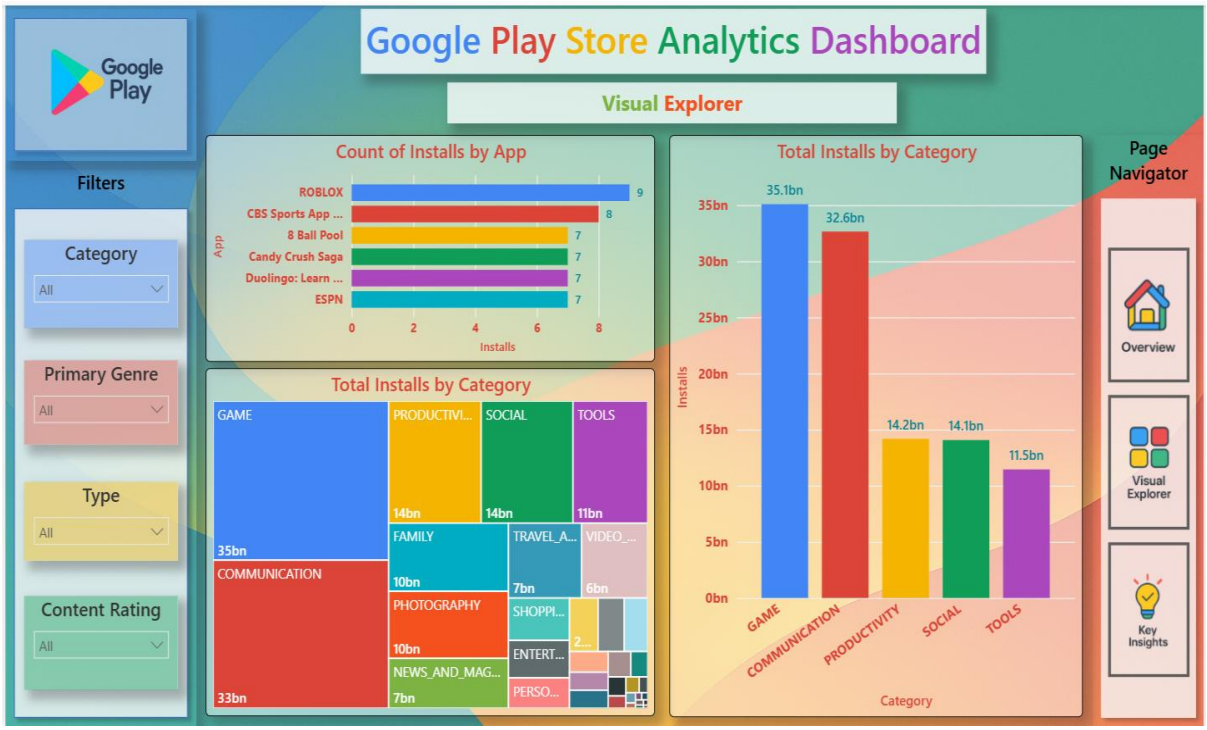
Tools	→	843 apps
Sports	→	384 apps
Travel & Local	→	258 apps
Video Players	→	175 apps

2. **Rating Distribution Across Apps**
   Most ratings fall between **3.5 and 4.5** indicating high user satisfaction.

3. **Correlation: Installs vs Ratings**
   Shows different patterns:

- High installs even with medium ratings
- Some highly rated apps have fewer installs

Dashboard 2: Visual Explorer



1. Count of Installs by App

Apps like **Roblox, ESPN, Candy Crush Saga** show the highest install numbers.

2. Total Installs by Category

Category	Installs
Game	35bn
Communication	33bn
Productivity	14bn
Social	14bn
Tools	11bn

3. Treemap – Category Share

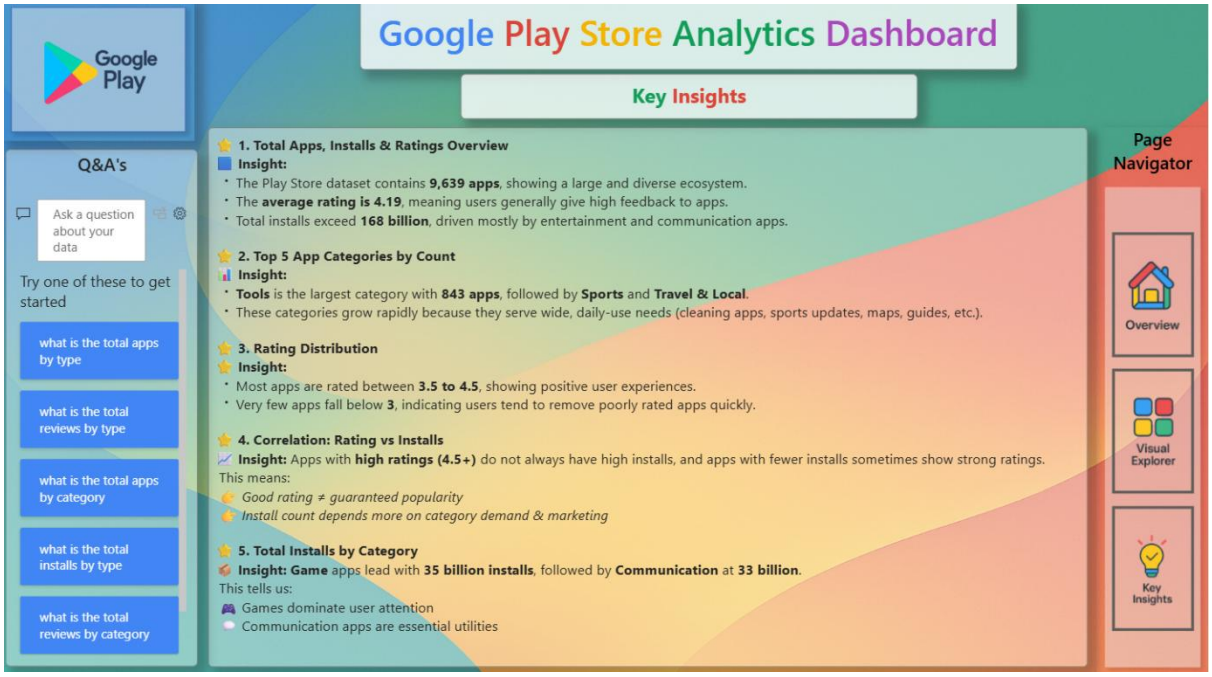
Games occupy the largest block, followed by Communication and Productivity.

4. Filters Added

- Category
- Primary Genre
- Type
- Content Rating

This allows users to explore data deeper.

Dashboard 3: Key Insights



Insight 1: Total Apps, Installs & Ratings

- Dataset contains **9,639 apps**
- Average user rating: **4.19**
- Total installs exceed **168 billion**

Meaning:
Android users are active and provide generally positive ratings.

Insight 2: Growth by Category

Tools have the largest number of apps, but Games dominate installs.

Insight 3: Rating Patterns

Most apps are rated above 3.5, indicating good quality across the Play Store.

Insight 4: Rating vs Install Correlation

High rating ≠ high installs
Install count depends more on:

- Marketing
- Category
- App age
- Popularity

Insight 5: Install Leader Categories

- Games – 35bn
- Communication – 33bn
- Productivity – 14bn

These categories are the most influential in the marketplace.

Conclusion, Future Scope & References

Conclusion

The Google Play Store dataset shows a diverse ecosystem where categories like Tools, Games, and Communication dominate. Ratings are generally positive, and installs vary widely depending on app type. Python helped clean and transform the dataset effectively, while Power BI provided impactful, interactive visualizations.
The project successfully demonstrates a complete data analytics pipeline.

Future Scope

- Sentiment analysis on user reviews
- Forecasting installs using ML models
- App success prediction
- Comparison with Apple App Store analytics
- Real-time analytics using APIs

References

- Google Play Store Dataset – Kaggle
- Pandas & NumPy Documentation
- Microsoft Power BI Documentation
- Python Official Documentation
- Academic papers on mobile app analytics

GitHub Repository Link

https://github.com/mansa022005/College-Project