App Data Analysis: Extracting and Visualizing Insights

Arushi Bajpai, Meenakshi Kanungo

B.Sc. Economics, Semester-III, Loreto College

Project Guide / Mentor Name:

Diptendu Dutta

Period of Internship: 14th Jan 2025 - 30th April 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

Table of Contents

Abstract

This project showcases how we used DAX Studio, Excel, and VegaLite to transform raw data into insights. We extracted the dataset from a Power BI file, cleaned it using Power Query in Excel, and reconnected it to DAX for querying. Finally, we visualized the data using VegaLite to reveal key patterns and trends.

Introduction

This project focuses on the end-to-end data analysis process, beginning with data extraction from a Power BI file using DAX Studio and ending with data visualization through VegaLite and Python (in Google Colab). With the increasing use of interactive dashboards in business intelligence, this project highlights how underlying data can be programmatically accessed, cleaned, and analyzed. After extracting the embedded data model using DAX Studio, the dataset was saved in CSV format and cleaned using Microsoft Excel's Power Query Editor. We then conducted further data analysis and visualized patterns using Python in Google Colab and VegaLite for clarity and flexibility. The objective is to demonstrate an efficient pipeline from raw data to insights, integrating widely-used tools across the analytics lifecycle.

Project Objective

This project involved extracting the full dataset embedded in a Power BI file using DAX Studio, followed by cleaning and formatting the data in Excel with Power Query for structured analysis. Exploratory data analysis was conducted using Python in Google Colab to uncover patterns and insights. Key trends and distributions were visualized using both VegaLite and Python libraries. The overall goal was to build a reusable and transparent data workflow that supports effective and insightful analysis.

Methodology

The dataset used in this project was provided by one of our mentors and sourced from the following link: https://www.kaggle.com/code/faisaljanjua0555/eda-google-play-store-apps/input .We followed a structured, step-by-step process to extract, clean, and analyze the data effectively. Initially, we used DAX Studio to extract data from a Power BI (.pbix) file and exported it as a CSV for easier handling. The CSV file was then loaded into Microsoft Excel, where we utilized Power Query to perform data cleaning tasks. This included removing unnecessary columns, filtering out irrelevant rows, and correcting data types to ensure consistency and accuracy. After cleaning, we reconnected the refined data back to DAX Studio

to run targeted DAX queries for deeper extraction and inspection. Finally, for data analysis and visualization, we used VegaLite to generate clear and insightful visualizations such as bar charts and dot plots. These visualizations helped us identify patterns, trends, and key insights in the data. This structured workflow—from data extraction to visualization—ensured a smooth, accurate, and efficient analytical process.

Data Analysis and Results

We cleaned and transformed the data in Excel, then analyzed it in Google Colab using Python. Visualizations were created using both VegaLite and Python libraries to explore trends, distributions, and category-wise comparisons. Key findings included performance differences across departments and skewed data distributions.

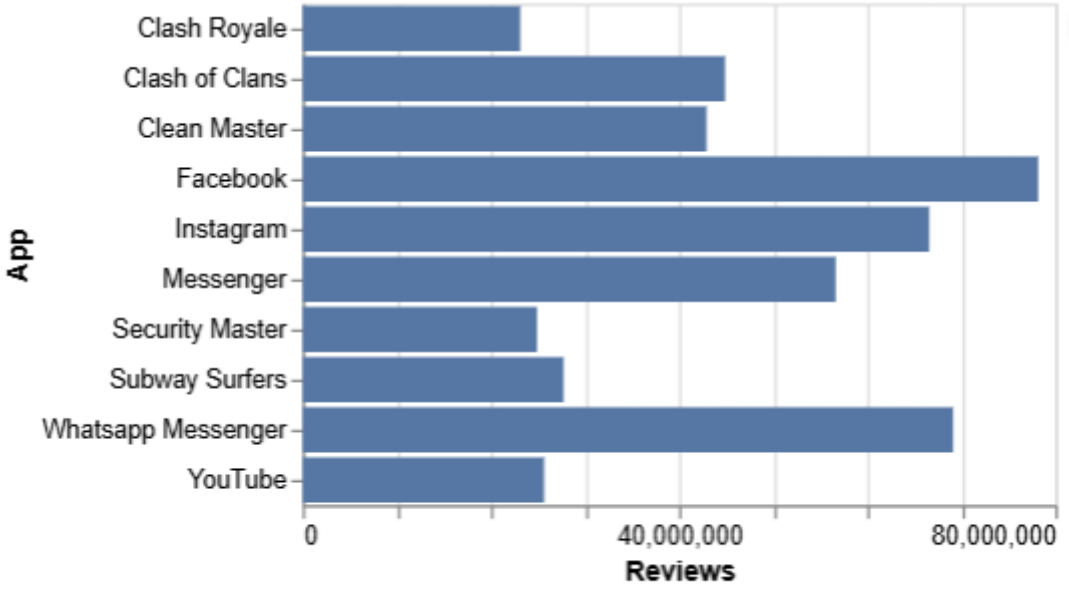

Figure 1: Top 10 apps by number of reviews

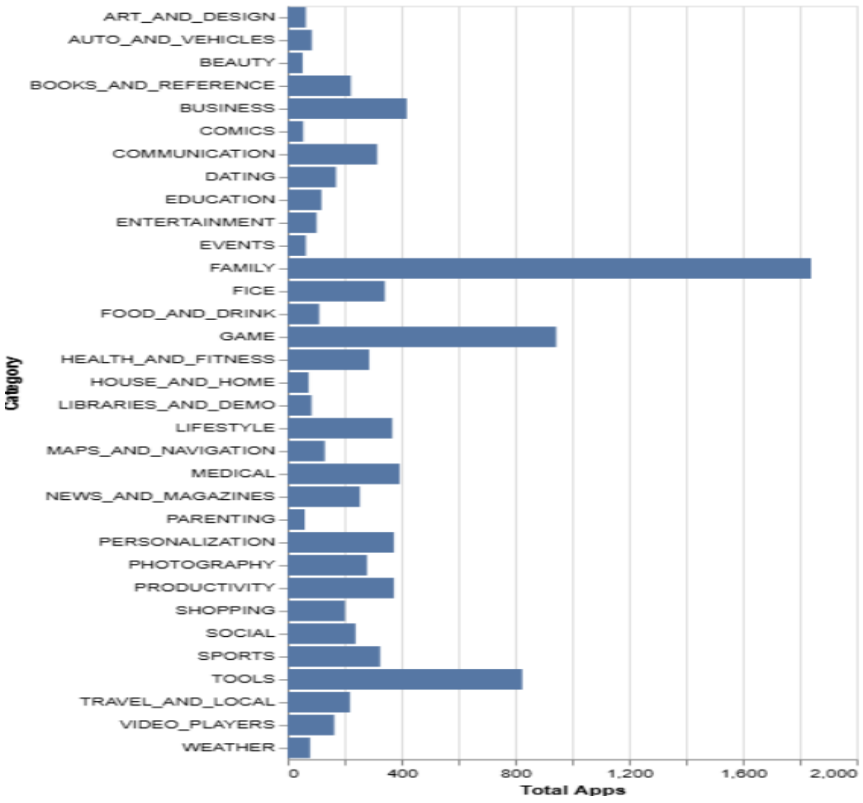

Figure 2: Total apps for each app category

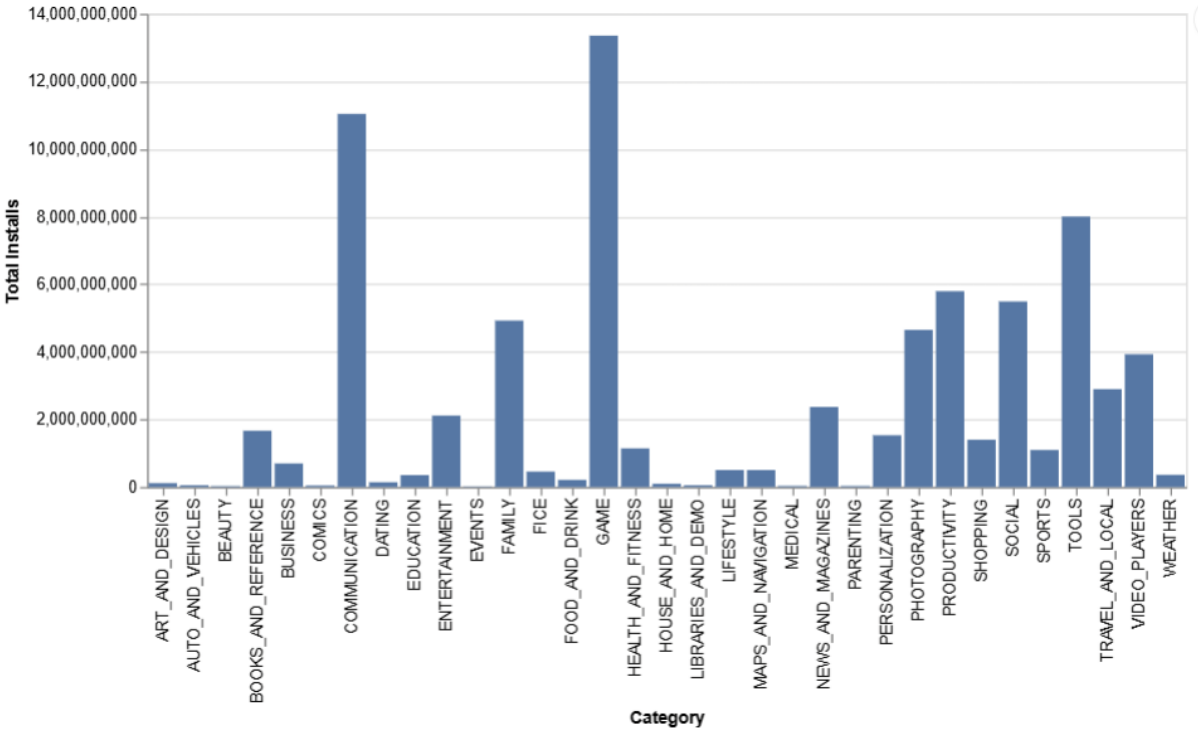

Figure 3: Average Reviews for each app category

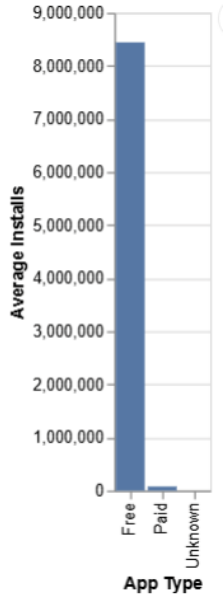

Figure 4: Average installs for each app type

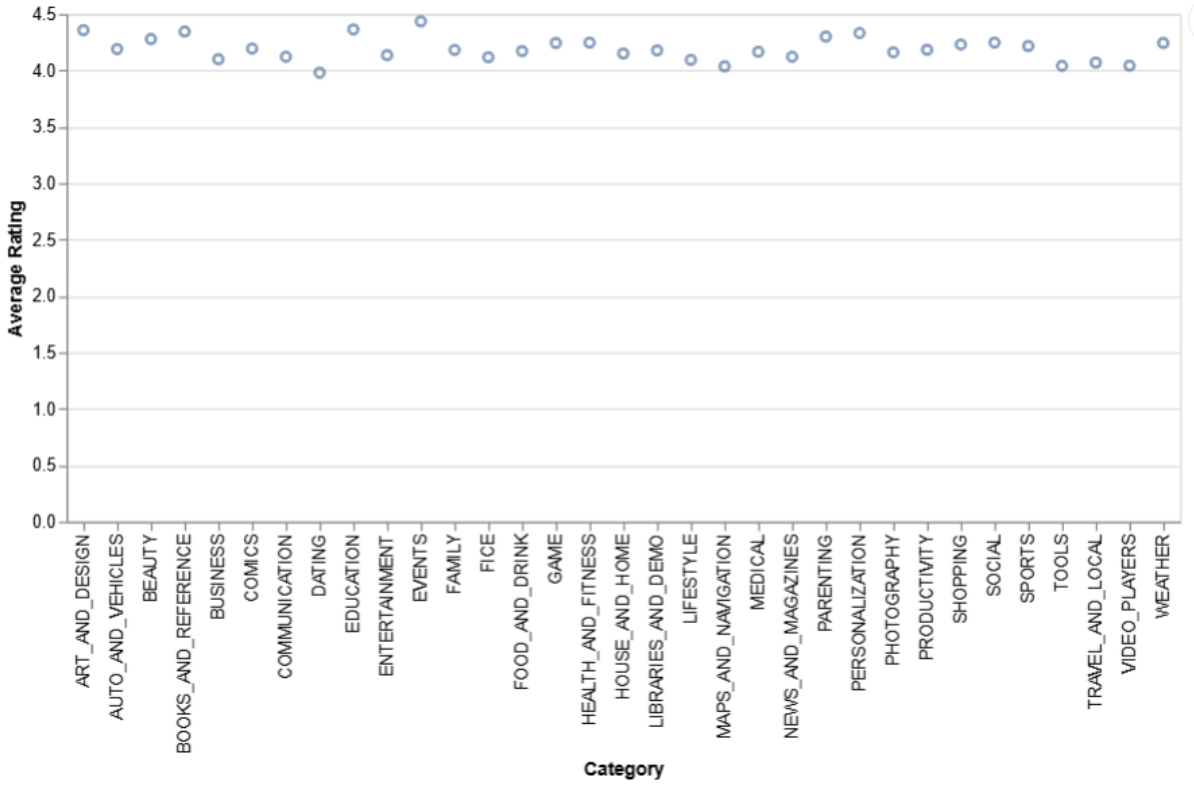

Figure 5: Average rating for each app category

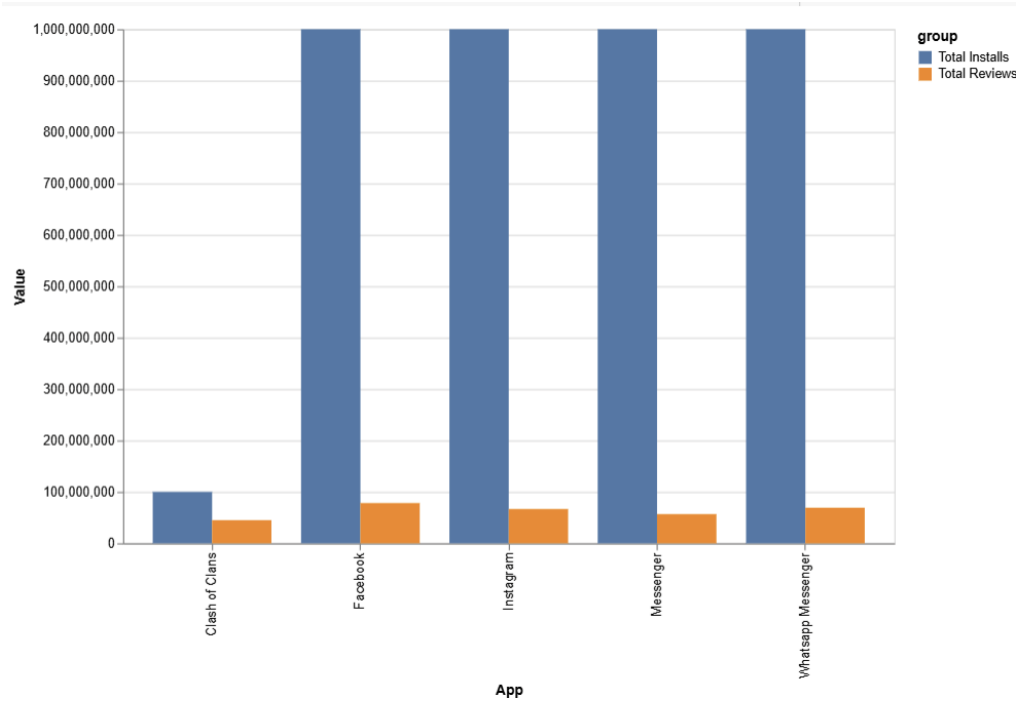

Figure 6: Total installs vs total reviews for the top 5 apps with the most installs

Conclusion

This project successfully demonstrated an end-to-end data analysis workflow by integrating tools like DAX Studio, Excel Power Query, Python, and VegaLite. By extracting embedded datasets from a Power BI file, cleaning them in Excel, and analyzing them in Python using Google Colab, the project uncovered key insights such as performance differences across departments and skewed distributions in app data. Visualizations created with Python libraries and VegaLite further clarified trends like category-wise review counts and installation averages. These findings confirm the effectiveness of a simple, transparent workflow for deriving insights from raw data. For future work, incorporating real-time or larger datasets and automating parts of the workflow (e.g., scheduled extraction or live dashboards) could enhance the scalability and practicality of the approach.

Appendices

- https://www.kaggle.com/code/faisaljanjua0555/eda-google-play-store-apps/input
- https://docs.streamlit.io/develop/api-reference/charts/st.vega_lite_chart
- https://altair-viz.github.io/user_guide/data.html
- https://vega.github.io/vega-lite/examples/#repeat--concatenation
- https://docs.streamlit.io/develop/api-reference/charts/st.vega_lite_chart
- https://colab.research.google.com/drive/16KVRA0vP4lKQVmdGt8OC78YECRI42m2-?usp=sharing