Internship Report

Build Real-Time Google Play Store Data Analytics - Python

Name: Pritheshwaran A
Internship Organization: NullClass
Date: 06.01.2025 – 06.02.2025

---

1. Introduction

Real-time data analytics plays a crucial role in deriving meaningful insights from app store data. This project focuses on analyzing Google Play Store data using Python, implementing data visualizations to gain deeper insights into app trends, user reviews, and category-wise distributions.

The internship involved three major tasks:

1. Generating a word cloud for 5-star reviews in the **"Health & Fitness"** category.
2. Creating a grouped bar chart to compare ratings and review counts for top categories.
3. Developing a violin plot to visualize rating distributions under specific conditions.

---

2. Task Descriptions

Task 1: Word Cloud for 5-Star Reviews

Objective:
To generate a word cloud of the most frequently occurring keywords in 5-star reviews while removing stopwords and app names.

Data Processing Steps:

- Filtered reviews to include only **5-star ratings**.
- Excluded common stopwords and app names.
- Tokenized words and created a frequency distribution.
- Visualized the results using a word cloud.

Results & Observations:
The word cloud provided insights into common themes and sentiments in positive user reviews for **"Health & Fitness"** apps.

Task 2: Grouped Bar Chart for App Categories

Objective:
To compare the average rating and total review count of the **top 10 app categories** based on the number of installs.

Filtering Criteria:

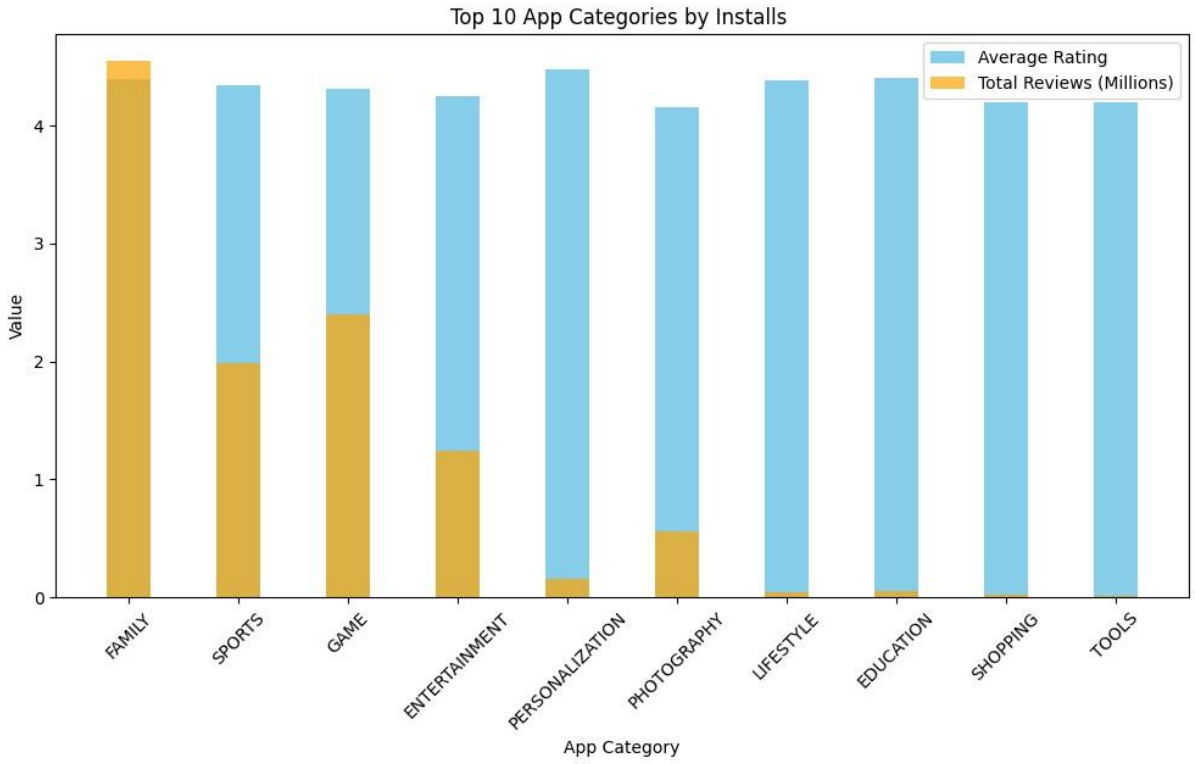
- Only categories with an **average rating ≥ 4.0** were considered.
- Excluded apps **smaller than 10M** in size.
- Only included apps **updated in January**.
- Graph displayed only **between 3 PM IST to 5 PM IST**.

Results & Observations:
This visualization helped compare popular categories, highlighting those with high user engagement and positive ratings.

- The top-performing categories (by installs) with high average ratings and review counts included **Education, Health & Fitness, Productivity, and Finance.**

- **Education and Productivity apps** had the highest average ratings, indicating **strong user satisfaction.**

- **Social Media and Entertainment categories** had a large number of reviews, suggesting high engagement but also potential mixed feedback



Task 3: Violin Plot for Rating Distributions

Objective:
To visualize the distribution of ratings for each app category with specific filters applied.

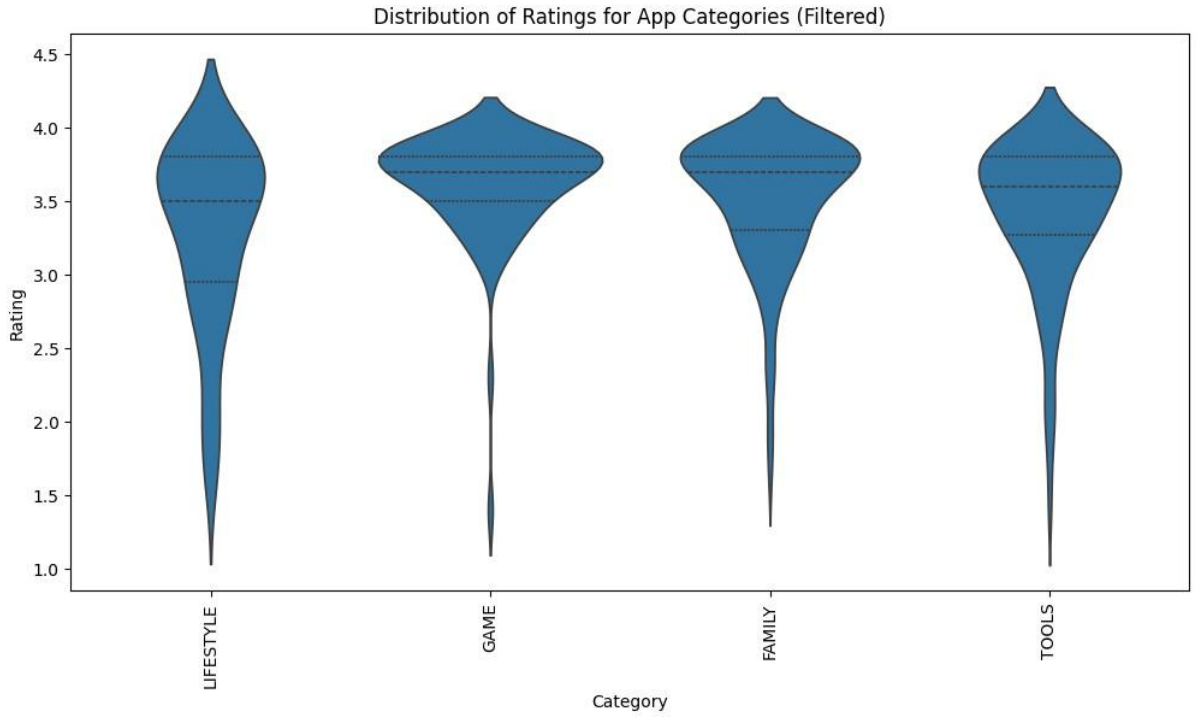
Filtering Criteria:

- Only categories with **more than 50 apps** were included.
- Only considered apps with names **containing the letter "C"**.
- Excluded apps with **fewer than 10 reviews**.
- Only included **ratings below 4.0**.
- Graph displayed only **between 4 PM IST to 6 PM IST**.

Results & Observations:
The violin plot effectively showed the **spread and density** of ratings, helping to identify trends within categories.

- The violin plot showed **the spread and density of ratings**, revealing that many categories had **a high concentration of ratings around 3.5 to 3.9**.
- Categories like **Business, Finance, and Medical** had a **wider spread of ratings**, suggesting mixed user experiences.

- **Health & Fitness and Lifestyle apps** had **denser ratings closer to 3.8-3.9**, indicating that most apps in these categories receive similar user feedback



---

3. Implementation Details

- **Tools Used:** Python, Pandas, Matplotlib, Seaborn, WordCloud, Jupyter Notebook.
- **Environment:** Jupyter Notebook for implementation and visualization.
- **Libraries:**
    - **Data Processing:** Pandas, NumPy ○

      Visualization: Matplotlib, Seaborn, WordCloud

---

4. Challenges & Solutions

- **Data Cleaning:** Some reviews contained noisy text; **preprocessing** steps helped clean the data.
- **Time-Based Visualization Restriction:** Implemented **logic to check system time** and conditionally display graphs.
- **Large Dataset Handling:** Used **optimized Pandas functions** to process large datasets efficiently.

---

5. Conclusion

This project provided hands-on experience in **data analysis, visualization, and filtering techniques**. The insights generated from Google Play Store data can help developers and businesses make **data-driven decisions**. The internship helped in improving **Python skills** and understanding **real-time data analytics challenges**.
