NullClass Internship Project Report On

Google Play Store Data Analysis

Submitted By : Chaitanya Kalebere

Internship Role – Data Analytics

Institute/Organization: NullClass

Project Details:

-Project Title: Learn to Build Realtime Google Play Store Data Analytics

-Tools & Technologies: Python, Jupyter Notebook, Pandas, NumPy, Plotly, Scikit-learn, NLTK

-Datasets Used: Play Store Data.csv and User Reviews.csv

1. Introduction

This project, developed as part of the NullClass Internship, focuses on analysing Google Play Store application metadata and user reviews. The goal is to uncover insights into app performance, user sentiment, and category-level trends using Python, Jupyter Notebook, and visualization libraries.

The datasets used are:

- **Play Store Data.csv**: App metadata including category, rating, installs, size, and last updated date.
- **User Reviews.csv**: Translated user reviews with sentiment polarity and subjectivity scores.

2. Objectives

- Clean and preprocess large datasets for reliable analysis.
- Merge app metadata with user reviews to connect installs, ratings, and sentiment.
- Generate interactive visualizations (bar charts, bubble charts, stacked area charts, choropleths).
- Apply category translations (Beauty → Hindi, Business → Tamil, Dating → German).
- Implement time-based restrictions so certain charts only display during specified IST windows.
- Document the workflow for reproducibility and clarity.

3. Methodology

3.1 Data Cleaning

- Converted Reviews, Installs, and Size columns to numeric values.
- Handled missing values using dropna and fillna.
- Normalized installs by removing commas and plus signs.
- Converted Last Updated into datetime format for time-series analysis.

3.2 Data Merging

- Joined Play Store Data.csv with User Reviews.csv on the App column.
- Integrated sentiment subjectivity and polarity into the metadata for richer insights.

3.3 Filtering Rules

Applied filters across tasks, such as:

- Minimum reviews (e.g., >500 or >1000).
- App names not starting with certain letters or not containing “S”.
- Categories restricted to specific sets (e.g., Entertainment, Beauty, Business, etc.).
- Ratings thresholds (e.g., >3.5 or ≥4.2).
- Size constraints (e.g., 20–80 MB).

3.4 Visualizations

- **Time Series Line Chart**: Trend of installs over time, shaded for >20% growth.
- **Bubble Chart**: App size vs rating, bubble size = installs, sentiment subjectivity filter applied.
- **Stacked Area Chart**: Cumulative installs by category, highlighting >25% growth months.
- **Grouped Bar Chart / Choropleth**: Category-level comparisons with translations applied.

3.5 Category Translations

- Beauty → **सौंदर्य** (Hindi)
- Business → **வணிகம்** (Tamil)
- Dating → **Dating (Deutsch)** (German)
- Travel & Local → **Voyage et Local** (French)
- Productivity → **Productividad** (Spanish)
- Photography → **写真撮影** (Japanese)

3.6 Time Restrictions

Charts are displayed only during specific IST windows:

- Line Chart: 6–9 PM IST
- Bubble Chart: 5–7 PM IST
- Stacked Area Chart: 4–6 PM IST

5. Tools & Libraries

- **Python**: Core programming language.
- **Pandas, NumPy**: Data cleaning and manipulation.
- **Plotly Express**: Interactive visualizations.
- **Scikit-learn**: Regression modeling (Random Forest).
- **NLTK (VADER)**: Sentiment analysis.
- **Jupyter Notebook**: Development and documentation environment.

6 Conclusion:

This project successfully demonstrates how large-scale app metadata and user reviews can be transformed into actionable insights through structured preprocessing, filtering, and visualization. The internship tasks were implemented with reproducible workflows, multilingual support, and time-based dashboard restrictions, ensuring both technical rigor and user-focused design.