Internship Report:
Real-Time Google Play Store Data Analytics

1. Introduction

This report outlines the internship experience and project work completed during my time as a Data Analyst Intern. The focus of this project was to analyze Google Play Store data in real-time using Python and its libraries. The end goal was to create an interactive dashboard that highlights key metrics and trends in mobile application performance.

2. Background

With the rapid growth of mobile applications, understanding app performance based on user reviews, downloads, ratings, and category-based behavior is critical. Google Play Store offers a wide variety of apps, and analyzing this data provides valuable business insights. This project was aimed at building a real-time data analytics solution with advanced visualizations and NLP techniques.

3. Learning Objectives

- Gain hands-on experience in data cleaning, transformation, and analysis.
- Apply NLP for sentiment analysis of user reviews.
- Create meaningful, time-filtered interactive visualizations using Plotly.
- Develop a Python-based dashboard using Tkinter for desktop UI.
- Understand data storytelling and visualization best practices.

4. Activities and Tasks - Here are the main tasks performed during the internship:

Task 1: Word Cloud for 5-Star Reviews

- Filtered reviews from apps in the **Health & Fitness** category.
- Removed stopwords, app names, and non-informative tokens using nltk.
- Generated a word cloud to highlight commonly used positive terms in 5-star reviews.

Task 2: Dual-Axis Chart (Free vs Paid Apps)

- Compared **average installs and revenue** for **top 3 app categories**.
- Applied filters:
  - Installs > 10,000
  - Revenue > $10,000

    - Android version > 4.0
    - Size > 15MB
    - Content rating = “Everyone”
    - App name length ≤ 30 characters
- Time constraint: Visible only **between 1 PM – 2 PM IST**.

Task 3: Choropleth Map for Global Installs

- Displayed global installs using **Plotly choropleth map**.
- Filtered top 5 app categories (excluding those starting with A, C, G, S).
- Highlighted categories with installs > 1 million.
- Time constraint: Visible only **between 6 PM – 8 PM IST**.

Task 4: Time Series Chart (Category-Wise Growth)

- Showed install trends over time per app category.
- Applied translation for categories:
    - Beauty → Hindi
    - Business → Tamil
    - Dating → German
- Highlighted growth > 20% MoM.
- Filters:
    - Reviews > 500
    - App name doesn’t contain “S”
    - App name doesn’t start with X, Y, Z
    - App category starts with B, C, or E
- Time constraint: Visible only **between 6 PM – 9 PM IST**.

Task 5: Bubble Chart (Size vs Rating vs Installs)

- Plotted bubble chart showing relationship between:
    - App size (X-axis)
    - Rating (Y-axis)
    - Installs (bubble size)
- Filters:
    - Rating > 3.5
    - Reviews > 500
    - Sentiment subjectivity > 0.5
    - App name doesn’t contain “S”
    - Selected categories: Game (highlighted pink), Beauty, Business, Communication, Comics, Dating, Social, Entertainment, Event
- Category translations included as in Task 4.
- Time constraint: Visible only **between 5 PM – 7 PM IST**.

5. Skills and Competencies Developed

- **Python Programming**: Efficient handling of data pipelines and visualization logic.
- **Data Manipulation with Pandas**: Aggregation, filtering, merging, type conversions.
- **Plotly**: Advanced interactive visualizations including bubble charts, dual-axis, and choropleth maps.
- **NLP with NLTK**: Sentiment analysis, stopword removal, text normalization.
- **Time-based UI Logic**: Controlled visibility of charts based on time using datetime and tkinter.
- **UI Design**: Created a dashboard using Tkinter and HTML/CSS components.

6. Evidence

- **Dashboard Snapshots**: Screenshots showing the visualizations created.
- **GitHub Repository**: Contains source code, cleaned datasets, and visuals.

- ## 7. Challenges and Solutions

- **Data Inconsistencies**
  Solution : Applied cleaning steps like type conversion, null handling, and regex filters.
- **Time -based filtering in GUI**
  Solution : Used “datetime.now()” with “pytz.timezone('Asia/Kolkata')” for IST validation.
- **NLP performance on large datasets**
  Solution : Used optimized NLTK pipelines and filtered text length beforehand.
- **App name/category translation**
  Solution : Integrated googletrans library for real-time language translations.

8. Outcomes and Impact

- Created a **dashboard** for real-time analytics.
- Enabled stakeholders to **visually interpret trends**, reviews, and performance.
- Improved skills in **data science, visualization, and NLP** significantly.
- Contributed to a repository that can be further scaled or integrated with APIs.

9. Conclusion

This internship allowed me to explore the full cycle of data analytics from cleaning and NLP to visual storytelling. It improved my ability to derive insights from complex datasets and present them interactively. The project gave me a strong foundation in both analytical thinking and technical skills.