GOOGLE PLAY STORE ANALYSIS

INTRODUCTION

Google Play Store is a vast platform that offers millions of applications for users around the world. Tw datasets have been chosen for the same. One dataset contains information regarding apps rating, genres, categories, reviews count, installs count, etc. Another dataset contains information regarding the reviews and sentiments. Google Play Store data analysis is all about digging into the details of apps—like their ratings, reviews, download numbers, and categories—to uncover meaningful insights. By looking at this data, we can spot trends, patterns, and connections that help developers and businesses make smarter decisions. Essentially, it's about understanding what works, what doesn't, and what users really want, so that developers can keep improving their apps and making them more appealing to their audience.

BACKGROUND

With over 2.8 million apps available in the Google Play Store, each app's success depends not only on its functionality but also on how well it meets user needs and how effectively it can be discovered in the crowded marketplace. Here's why analysing Google Play Store data is so essential:

1. **Increase in Mobile apps usage**: As the mobile phones have become an important part of everyone's lives, number of installs of apps have increased by a huge amount and allows the users to interact with the apps, which in turn helps the developers to know the user preferences, behavioural patterns, and market dynamics.
2. **The Role of Data in Competitive Advantage**: By analysing data such as ratings, reviews, and download numbers, developers can identify what features or designs are working well and what areas need improvement. This data-driven approach can be the difference between a successful app and one that fails to stand out.
3. **User Feedback and Sentiment**: User reviews and ratings are direct reflections of an app's performance and user satisfaction. Analysing these reviews and sentiments leads the developers to know about the dissatisfying features according to users.
4. **Trends and Patterns:** The mobile app market changes fast, with new trends, features, and user behaviours popping up all the time. By analysing data from different app categories and regions, developers and analysts can spot these emerging trends—like the rise of fitness or gaming apps—and jump on them early. This gives them a chance to stay ahead of the curve, create apps that align with what people want, and tap into new opportunities before they become mainstream.
5. **Monetization and Business Models**: Google Play Store data can also provide insights into different monetization strategies. Developers can compare free vs. paid app models, in-app purchases, ads, and subscription-based services. By understanding which business model works best for their app, developers can better optimize their revenue.
6. **Global Reach:** As google play store is a global platform, data analysis helps the developers understand regional preferences and behaviour. This helps in tailoring apps for specific markets, optimizing user acquisition, and improving overall user retention across different geographies.

LEARNING OBJECTIVES

The objectives involve:

- Learn about the different categories of apps available in the Google Play Store.
- Recognize the importance of metrics like app ratings, user reviews, download numbers, and category rankings for performance evaluation.
- Understand how the data analysis helps to get to know information about user preferences, apps defaults and how to improve the apps.
- Learn about emerging trends and patterns.
- Understand the factors driving trends, such as user behaviour shifts or technological advancements.
- Gain skills in analysing user reviews to identify sentiment (positive, negative, neutral).
- Learn how to analyse app pricing models (free, freemium, paid, etc.) and in-app monetization strategies.

ACTIVITIES AND TASKS

Initially, all the necessary libraries like pandas, numpy , plotly express, nltk, etc are imported. After both the datasets have been imported namely apps_df and reviews_df. The data is cleaned after dropping the null values and converting the installs columns to numeric by removing commas and +, converting price column to numeric after removing $, converting the size into M. Sentiments Intensity Analyser of Reviews was performed by importing SentimentIntensityAnalyzer. After merging both the datasets, merged_df is achieved. HTML path was defined and after the tasks have been performed, web page has been created for the same.

TASK 1

The first task was to visualize the sentiment distribution (positive, neutral, negative) of user reviews using a stacked bar chart, segmented by rating groups (e.g.,1-2 stars, 3-4 stars, 4-5 stars). Include only apps with more than 1,000 reviews and group by the top 5 categories. After filtering out the apps having more than 1000 reviews, the top 5 categories found were games, family, health and fitness, dating and travel and local. After categorising ratings into different segments, Sentiment distribution df is achieved after grouping the sentiments of merged_df on the basis of rating groups. The bar is plotted using plotly and the html is saved as plot. The insight which can be forecasted is that the Sentiments in reviews show a mix of positive and negative feedback, with a slight lean towards positive sentiments. The apps having rating 2 and 3 stars didn't have any reviews.

TASK 2

The second task is to create a dual-axis chart comparing the average installs and revenue for free vs. paid apps within the top 3 app categories. Apply filters to exclude apps with fewer than 10,000 installs and revenue below $10,000 and android version should be more than 4.0 as well as size should be more than 15M and content rating should be Everyone and app name should not have more than 30 characters including space and special character .this graph should work only between 1 PM IST to 2 PM IST apart from that time we should not show this graph in dashboard itself. The first

step was to replace 'and up' and 'Varies with device', strip spaces, and convert to float safely for the column version and the data was filtered out after following all the conditions mentioned. The top categories achieved were family, game and sports. The apps_df is grouped on the basis of Category and Type. The library pytz is imported and datetime is also imported to make sure that the graph is visible only during the mentioned time. Then the figure is plotted, traces were added and the layout was updated. The chart is only available between 1 PM and 2 PM IST.

TASK 3

The last task was to generate a heatmap to show the correlation matrix between installs, ratings, and review counts. Filter the data to include only apps that have been updated within the last year and have at least 100,000 installs and reviews count should be more than 1k and genres name should not be starting with characters A, F, E, G, I, K. This graph should work only between 2 PM IST to 4 PM IST apart from that time we should not show this graph in dashboard itself. In this case the latest year where the apps were updated was 2018. S0 the last year will be considered as 2018. The data(apps_df) is filtered out after satisfying the mentioned conditions. Correlation matrix was formed on the basis of installs, rating and review counts. The next thing performed was to restrict display to between 2 PM IST and 4 PM IST by using datetime.now().strftime('%H:%M'). Plotly was used to visualise this correlation matrix.

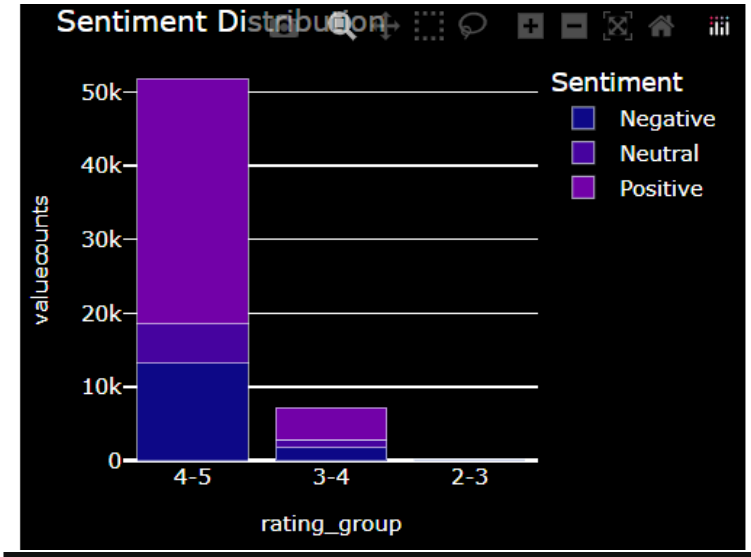
SKILLS AND COMPETENCIES

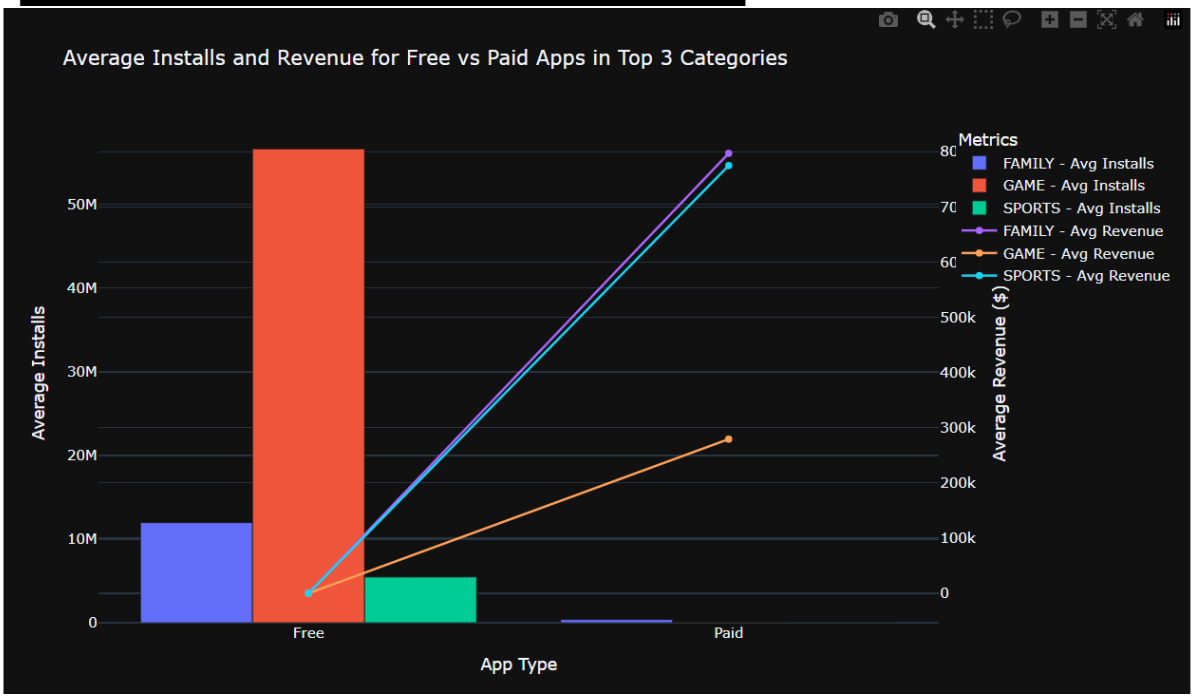
Following skills and competencies are achieved:

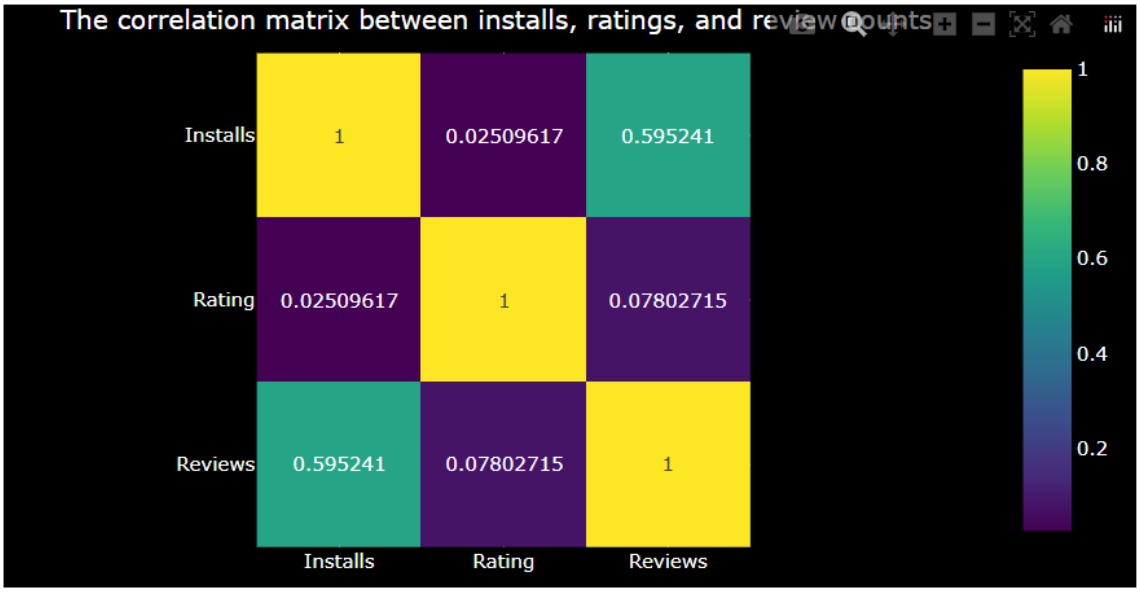
- Proficiency in cleaning and preparing raw data, handling missing or inconsistent data, and ensuring the dataset is ready for analysis.
- Knowledge of basic statistical methods to summarize data, such as calculating averages, medians, and standard deviations, to interpret app performance.
- Knowledge of basic statistical methods to summarize data, such as calculating averages, medians, and standard deviations, to interpret app performance using Sentiment Intensity Analyzer.
- Proficiency in using tools like Plotly to create visual dashboards that help stakeholders understand trends and key performance indicators.
- Competence in analyzing and comparing different monetization strategies (e.g., in-app purchases, ads, subscriptions) and understanding their impact on app revenue.

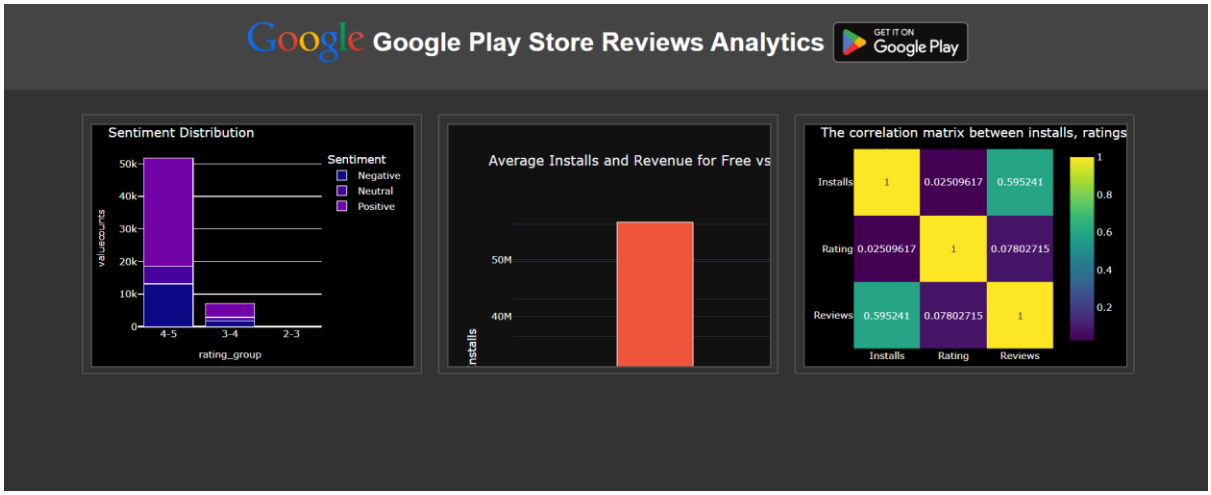
FEEDBACK AND EVIDENCE

The images of the visuals are mentioned below as evidence.


Sentiment Distribution
Sentiment
Negative
Neutral
Positive
valuecounts
50k
40k
30k
20k
10k
0
4-5
3-4
2-3
rating_group


Average Installs and Revenue for Free vs Paid Apps in Top 3 Categories
Metrics
FAMILY - Avg Installs
GAME - Avg Installs
SPORTS - Avg Installs
FAMILY - Avg Revenue
GAME - Avg Revenue
SPORTS - Avg Revenue
Average Installs
50M
40M
30M
20M
10M
0
Average Revenue ($)
80
70
60
500k
400k
300k
200k
100k
0
Free
Paid
App Type


The correlation matrix between installs, ratings, and review counts
Installs
Rating
Reviews
1
0.02509617
0.595241
0.02509617
1
0.07802715
0.595241
0.07802715
1
Installs
Rating
Reviews
1
0.8
0.6
0.4
0.2



CHALLENGES AND SOLUTIONS

1. **Handling Missing or Incomplete Data**

   Challenge: The Google Play Store dataset contains missing or incomplete entries, such as missing ratings, app categories, or reviews, etc.

   Solution:

   - **Data Cleaning:** Python libraries like **Pandas** are used to handle missing data by either fill missing values with a placeholder (e.g., 0, "Unknown"), drop rows with missing data, etc.
   - **Imputation:** There are cases where numerical data is missing and values are imputed or put using the mean or median of the column.

2. **Data Size and Performance**

   Challenge**:** Google Play Store datasets are large, and working with massive datasets may lead to performance issues.

   Solution:

   - **Efficient Data Processing:** The efficient data structures like **Pandas DataFrame** are used for manipulation and processing.
   - **Filtering Data:** Unnecessary columns and rows are filtered out (e.g., filtering by category or rating) before processing.

3. **Data Visualization**

   Challenge: Presenting the data in an insightful and visually engaging manner is often tricky because of having a large number of apps with multiple features.

   Solution:

   - Python library mainly **Plotly** is used for visualizing the data. Various charts, such as bar charts, dual axis chart heatmaps are created in the tasks to understand the distribution of ratings, number of installs, app categories, etc.

4. **Data Inconsistencies**

   Challenge: The dataset has inconsistent data, such as incorrect formats for fields like "Installs" (contain symbols like "+" or ","), or size that go beyond the valid range.

   Solution:

   - **Data Standardization:** Use of regular expressions or string manipulation techniques to clean data fields are made. (e.g., converting '+' to ' ')

5. **Analyzing Reviews & Sentiment Analysis**

   Challenge: Analysing user reviews can be challenging due to the unstructured text data.

   Solution:

   - **Text Preprocessing:** The reviews have been cleaned by removing stop words, punctuation, and special characters using **NLTK**
   - **Sentiment Analysis:** Sentiment analysis is done on reviews using pre-trained models like **VADER** to classify reviews as positive, negative, or neutral.

OUTCOMES AND IMPACT

1. App Performance Insights

- **Outcome:** By analysing data like ratings, reviews, and install numbers, developers can understand how well their apps are performing in the market.
- **Impact:** Developers can pinpoint areas for improvement—whether it's UI/UX, functionality, or bugs. This helps in prioritizing features or fixes, improving app quality, and boosting user satisfaction. Apps with better ratings and user reviews are more likely to attract new users, leading to higher downloads and revenue.

2. Market Trends and Competitive Analysis

- **Outcome:** Data analysis can reveal emerging trends within specific app categories, such as gaming, education, or health and fitness.
- **Impact:** By understanding these trends, businesses can adapt to market demands, create products that align with consumer interests, or identify gaps in the market. This gives companies a competitive edge and helps them stay ahead in a fast-paced app ecosystem.

3. User Behaviour Understanding

- **Outcome:** Sentiment analysis of user reviews and ratings provides insights into user sentiments, likes, dislikes, and suggestions for improvement.
- **Impact:** Developers and marketers can tailor their strategies based on actual user feedback, improving customer retention and engagement. This also helps in targeting specific user segments for personalized marketing or feature development.

4. Business Decision-Making

- **Outcome:** By analysing install trends, revenue estimates, and ratings, businesses can make data-driven decisions about marketing budgets, promotional strategies, or app updates.
- **Impact:** Organizations can focus resources on the most profitable or high-potential apps. This leads to more effective marketing campaigns, resource allocation, and ultimately increased profitability.

5. Monetization Strategy Refinement

- **Outcome:** By analyzing the relationship between app ratings, installs, and in-app purchases or ad revenue, businesses can gauge the effectiveness of their monetization strategies.
- **Impact:** If certain features are correlated with higher revenue or engagement, businesses can optimize those features. Whether it's through in-app purchases, subscriptions, or ads, this data helps businesses maximize their revenue potential.

6. User Acquisition and Retention

- **Outcome:** Analyzing the number of installs, app ratings, and review sentiment can help predict how likely users are to continue using the app after download.
- **Impact:** By identifying what factors lead to better retention (e.g., updates, customer support), companies can refine their acquisition and retention strategies. Apps that keep users engaged longer tend to see more success over time, resulting in a sustainable growth trajectory.

7. Improved Product Development

- **Outcome:** Data-driven insights allow companies to continuously improve their product by identifying pain points and opportunities based on user feedback and app performance metrics.
- **Impact:** Iterating on product features with real, actionable data leads to a more refined product that aligns better with users' needs, improving both customer satisfaction and brand loyalty.

8. Forecasting and Predictive Analytics

- **Outcome:** Analysing trends over time, such as growth in installs or ratings, can help forecast future app performance or market behaviour.
- **Impact:** Forecasting empowers businesses to make proactive decisions, such as scaling server infrastructure, increasing marketing budgets, or preparing for a seasonal surge. Predictive analytics also allows companies to anticipate challenges and plan accordingly.

CONCLUSION

In conclusion, tackling the challenges of Google Play Store data analysis can be rewarding, and by applying the right solutions, we can extract valuable insights that reveal trends, patterns, and user behaviour. The process often involves several stages: gathering the data, cleaning it up, exploring it, and then analysing it—sometimes requiring a few rounds of fine-tuning to get things just right.

The key is to stay adaptable, whether we're working with missing data, cleaning up outliers, or diving into visualizations. While it might take some time and effort, once we've navigated these hurdles, the insights we uncover can be extremely valuable, whether we're identifying app trends, understanding user preferences, or making data-driven decisions.

In the long run, data analysis of Google Play Store data equips businesses with a deeper understanding of their users, competitors, and the app ecosystem at large. The insights gathered can drive smarter decisions, enhance customer satisfaction, and lead to more successful apps. It's not just about collecting data—it's about turning that data into actionable strategies that can shape a more profitable and sustainable future for developers, marketers, and companies as a whole.

By continuously analysing and adapting based on these insights, businesses can stay competitive, drive innovation, and ensure they're meeting the evolving needs of the app market.