

# Project 3: Google Play store Apps

## Introduction

The Google Play Store is a vital ecosystem for Android applications, offering millions of apps across various categories.

This project analyses a dataset of Google Play Store apps to uncover valuable insights into app ratings, user engagement, and revenue opportunities. The study employs Python-based exploratory data analysis (EDA), sentiment analysis, and machine learning techniques to provide actionable recommendations for app developers and businesses.

## Objective

The project aims to:

- Analyse app characteristics such as ratings, reviews, installs, and categories.
- Identify trends, patterns, and factors contributing to app success.
- Identify trends, patterns, and factors contributing to app success.
- Provide actionable recommendations to enhance user satisfaction and revenue.

## Methodology

The project followed a systematic approach:

- **Data Loading and Exploration:** The dataset was loaded, and its structure was examined to identify key features.
  - **Data Cleaning and Preprocessing:** Addressed missing values, standardized data formats, and handled inconsistencies.
  - **Exploratory Data Analysis (EDA):** Visualized data distributions and relationships between variables.
  - **Revenue and Market Position Analysis:** Assessed profitability and popularity across categories.
  - **Sentiment Analysis:** Analysed user reviews for sentiment polarity and subjectivity.
- Machine Learning Models:
- Linear Regression: Predicted app ratings based on features like reviews and installs
  - Decision Tree Classifier: Classified apps as free or paid based on app attribute

# Analysis and Results

## Data Cleaning and Preprocessing:

- Missing Values:
  - Rating: 13.6% missing values were replaced with the mean rating (4.19).
  - Columns like `Type` and `Content Rating` with few missing values were dropped.
- Data Transformation:
  - Converted `Reviews`, `Installs`, and `Price` to numeric formats.
  - Size column values (e.g., '19M') were transformed into KB for consistency.
- Handling Duplicates:
  - Removed 483 duplicate entries, reducing the dataset to 10,346 rows.

## Exploratory Data Analysis (EDA):

The dataset was visualized and analysed to identify trends and patterns:

- Rating Distribution: Most apps have ratings between 4.0 and 4.5, indicating high user satisfaction.
- Top Categories: The 'Family' category has the highest number of apps, followed by 'Games'.
- Price Analysis: Paid apps with moderate prices (below \$10) received higher user ratings than those with extreme prices.
- Installs and Reviews: A strong positive correlation was observed between the number of installs and reviews, particularly in popular categories.

## Revenue and Market Position Analysis:

- Revenue Analysis: Paid apps in the 'Games' and 'Productivity' categories generated the most revenue.
- Market Position: Categories with high installs and ratings, like 'Games' and 'Family', demonstrated strong market presence.

## Sentiment Analysis:

User reviews were analyzed for sentiment polarity and subjectivity:

- Positive Sentiments: Positively correlated with higher app ratings, indicating user satisfaction.
- Category Trend: 'Games' and 'Family' categories received the highest positive feedback.
- Negative Sentiments: More frequent in 'Business' and 'Tools' categories, highlighting areas for improvement.

## Machine Learning Analysis:

Two predictive models were implemented:

- **Linear Regression:**
  - Predicted app ratings with a Root Mean Square Error (RMSE) of 0.49.
  - Significant predictors: Installs, Reviews, and App Size.
- **Decision Tree Classifier:**
  - Classified apps as Free or Paid with an accuracy of 100%.
  - Features like Price and Installs showed high predictive power.

## Key Insights

- **User Engagement:** Categories like 'Games' and 'Family' dominate in installs and reviews, indicating high user engagement.
- **Sentiment Correlation:** Positive sentiment correlates with higher app ratings.
- **Pricing Strategies:** Moderately priced paid apps tend to receive higher user ratings.
- **Revenue Generation:** Categories like 'Productivity' and 'Games' drive significant revenue.

## Recommendations

1. Focus on improving user experience in categories with high installs but low ratings, such as 'Business' and 'Tools'.
2. Leverage positive user sentiment for targeted marketing campaigns to increase app downloads.
3. Optimize pricing strategies for paid apps to balance affordability and revenue.
4. Regularly analyse user reviews and sentiment trends to identify issues early and maintain user satisfaction.

## Conclusion

This comprehensive analysis of Google Play Store apps highlights critical factors contributing to app success.

By leveraging insights from data analysis and machine learning, app developers and businesses can enhance user engagement, optimize pricing strategies, and maximize revenue in a competitive app marketplace.

## **Tools & Technologies**

- Programming Language: Python
- Libraries Used: Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn.
- Platform: Jupyter Notebook.
- Dataset Source: Data given by Unified Mentor Team
- Dataset Features: The dataset consists of 13 columns.