

Symbiosis Skills and Professional University

PROJECT REPORT On

Playstore Data Analysis



SUBMITTED BY :-

Ankita Karvande .

Ashutosh Vyawhare .

Sneha Bhosale .

REGISTERED BATCH: ML_08

UNDER THE GUIDANCE OF

TRAINER: AMRITA AJOTIKAR MA'AM

STUDENT DECLARATION AND ATTESTATION BY TRAINER

This is to declare that this report has been written by us. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. I aver that if any part of the report is found to be plagiarized, I shall take full responsibility for it.

Name Of Student

Signature

Ankita Karvande

Ashutosh Vyawhare

Sneha Bhosale

Signature of trainer

Amrita ma'am

CERTIFICATE

This is to certify that,

Ankita Karvande,

Ashutosh Vyawhare,

Sneha Bhosale,

Has completed and submitted the project entitled, “PLAYSTORE DATA ANALYSIS”, under the guidance of Amrita ma'am, to Symbiosis Skills and Professional University, Pune, Maharashtra, India, is a record of Bonafede project work carried out by them and is worthy of consideration for the completion of certificate course in “Machine Learning”.

Date:

Signature

Miss Amrita ma'am

Supervisor

ACKNOWLEDGEMENT

We are profoundly grateful to trainer Amrita ma'am for her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

We would like to express our deepest appreciation towards SYMBIOSIS CENTRE OF DISTANCE LEARNING, Pune. Prof. Baliram Sir whose invaluable guidance supported us in completing this project and also JP Morgan for funding and giving us this opportunity.

At last, we must express our sincere heartfelt gratitude to all staff members of Computer Science & Engineering Department who helped us directly or indirectly during this course of work.

Thanking you all!!!

Index

Contents

Abstract.....	6
1.Introduction	7
2. Methodology.....	9
3. Software Requirement.....	11
4 Project Implementation	12
5. Advantages.....	21
6. Limitations	21
7.Applications	21
8. Future Scope	22
9.Conclusion.....	22

Abstract

The PlayStore, as one of the largest digital marketplaces for Android applications, has become an invaluable platform for developers and users alike. This abstract provides an overview of a comprehensive data analysis and prediction study conducted on PlayStore data, aiming to gain insights into user behaviors, app performance, and future trends.

The study encompasses a vast dataset comprising app attributes, user reviews, download statistics, and other relevant information collected up to September 2021. Through rigorous data preprocessing, exploratory data analysis, and machine learning techniques, the following key findings and predictions have been derived:

The insights and predictions derived from this analysis have the potential to benefit various stakeholders, including app developers, marketers, and investors. By understanding user behavior, optimizing app performance, and staying ahead of market trends, stakeholders can make data-driven decisions to enhance their position in the highly competitive PlayStore ecosystem.

In conclusion, this PlayStore data analysis and prediction study offers a comprehensive understanding of user behavior, app performance, and emerging trends, enabling stakeholders to make informed decisions and succeed in the dynamic world of mobile applications.

1. Introduction

1.1 Motivation

A drive to learn new things.

To learn new programming language, platform.

Self satisfaction. It's a great feeling to achieve something.

A drive to create something useful for other people.

Career advancement. The skills and knowledge I gain will be useful for my future career.

Financial. It would be great to get some money out of all of the above.

1.2 Problem Statement

what factors influence their choices?

What trends can be identified in user behavior, and how can this knowledge be leveraged to improve app offerings?

1.3 Objective

Playstore-Analysis-Project The objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product.

The dataset is chosen from Kaggle. It is of 10k Play Store apps for analyzing the Android market.

1.4 Scope

Predictive models can be used to assess the potential success of a new app or update, guiding investment decisions.

1.5 Industrial use

Analyzing data from the Google Play Store can have several industrial applications, especially for businesses in the mobile app and software development industry.

Here are some ways in which the Play Store data analysis can be useful:

1. Market Research :

Businesses can use Play Store data to conduct market research. They can analyze user reviews, ratings, and download statistics to understand user preferences, identify trends, and discover gaps in the market. This information can guide product development and marketing strategies.

2. Competitor Analysis :

Companies can use Play Store data to analyze their competitors' apps.

3. App Performance Monitoring : Developers can monitor their own app's performance on the Play Store. They can track user ratings and reviews to identify issues or bugs and address them promptly.

4. Pricing Strategy : By examining how price changes affect download rates and revenue, businesses can optimize their pricing strategies for maximum profitability.

2. Methodology

2.1 Algorithms used :

- Linear regression
- Logistic Regression
- Random forest Classifier
- Decision Tree
- SVR
- KNN

2.2 Hypothesis Testing

We start by defining the null hypothesis (H_0) which states that there is no relation between the variables. An alternate hypothesis (H_1) would state that there is a significant relation between the two.

one-sample z-test

is to determine whether the mean of a single sample significantly differs from a known or hypothesized population mean, providing insights into whether observed differences are statistically meaningful.

The one-sample proportion test

assesses whether a sample proportion significantly differs from a hypothesized population proportion, helping determine if there is a statistically significant deviation in a categorical variable's distribution.

The two-sample paired t-test

evaluates whether there is a statistically significant difference between two related groups (e.g., before and after measurements), indicating whether an intervention or treatment had a significant effect on the paired observations.







ANOVA (Analysis of Variance) test

determines whether there are statistically significant differences in means across three or more groups, while the **pairwise Tukeyhsd test** identifies which specific group means differ significantly from one another, aiding in post-hoc group comparisons. In this article,

we will perform the test using a mathematical approach and then using Python's SciPy and Stat module.

If P value in respective test is less than alpha(default 0.05) then H0 is rejected.

2.3 Libraries used

-  **Pandas:** It provides fast, expressive, and flexible data structures to easily (and intuitively) work with structured (tabular, multidimensional, potentially heterogeneous)
-  **Numpy:** It has advanced math functions and a rudimentary scientific computing package. Numpy is a popular array – processing package of Python. It provides good support for different dimensional array objects as well as for matrices.
-  **Matplotlib:** Matplotlib helps with data analyzing, and is a numerical plotting library. Matplotlib can create such quality figures that are really good for publication. Figures you create with Matplotlib are available in hardcopy formats across different interactive platforms.
-  **Seaborn:** It provides a high-level interface for drawing attractive and informative statistical graphics.
-  **Sk-learn:** It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.
-  **Scipy:** The SciPy library, a collection of numerical algorithms and domainspecific toolboxes, including signal processing, optimization, statistics, and much more. Matplotlib, a mature and popular plotting package that provides publication-quality 2-D plotting, as well as rudimentary 3-D plotting

3. Software Requirement

3.1 System Requirements:

3.1.1 Database Requirement:

- Ms-Excel

3.1.2 Software Requirement:

- Jupyter Notebook
- Os Windows10
- Programming Language- Python.

3.1.3 Hardware Requirement:

- Any Device With Browser Support

4 Project Implementation

4.1 **Steps of Project:**

- Searching and deciding data
- Importing data into python
- Data cleaning
- Performing exploratory data analysis
- Hypothesis testing
- Feature engineering
- Implementation of algorithms
- Comparison of algorithms via accuracy

Collection of data:

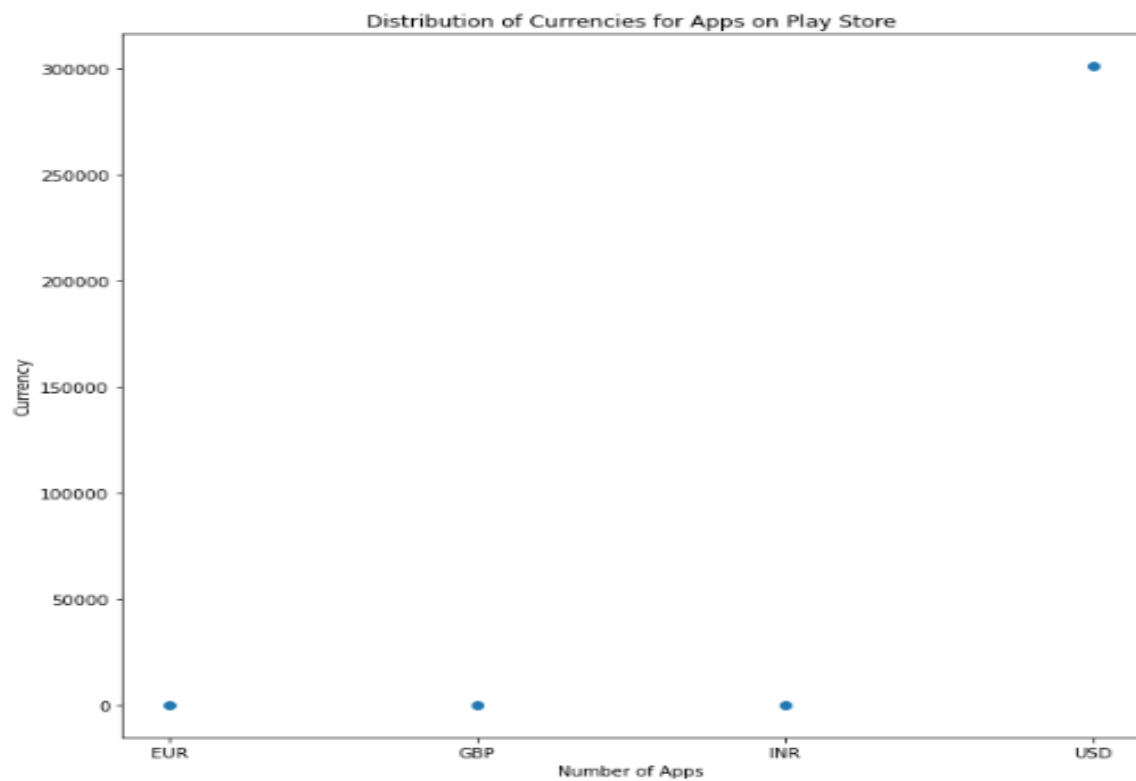
We chose our data from Kaggle site. The shape of our data was almost 450000 in the start but after cleaning and preprocessing it got reduced to around 400000.

Importing the data and cleaning the data:

All the coding is done in python programming language on jupyter notebook. The dataset was imported using pandas library. Head, tail and shape of data was checked. Columns of the dataset were checked. Index and info was checked.

Performing exploratory data analysis:

Before doing data exploration, we converted the categorical data set into numeric data set. After doing this, we performed data visualization techniques. Plotted graphs to find the spread of the data and to check measure of tendency and to find trends in our data. Groupby was also done check which gives more clear picture.

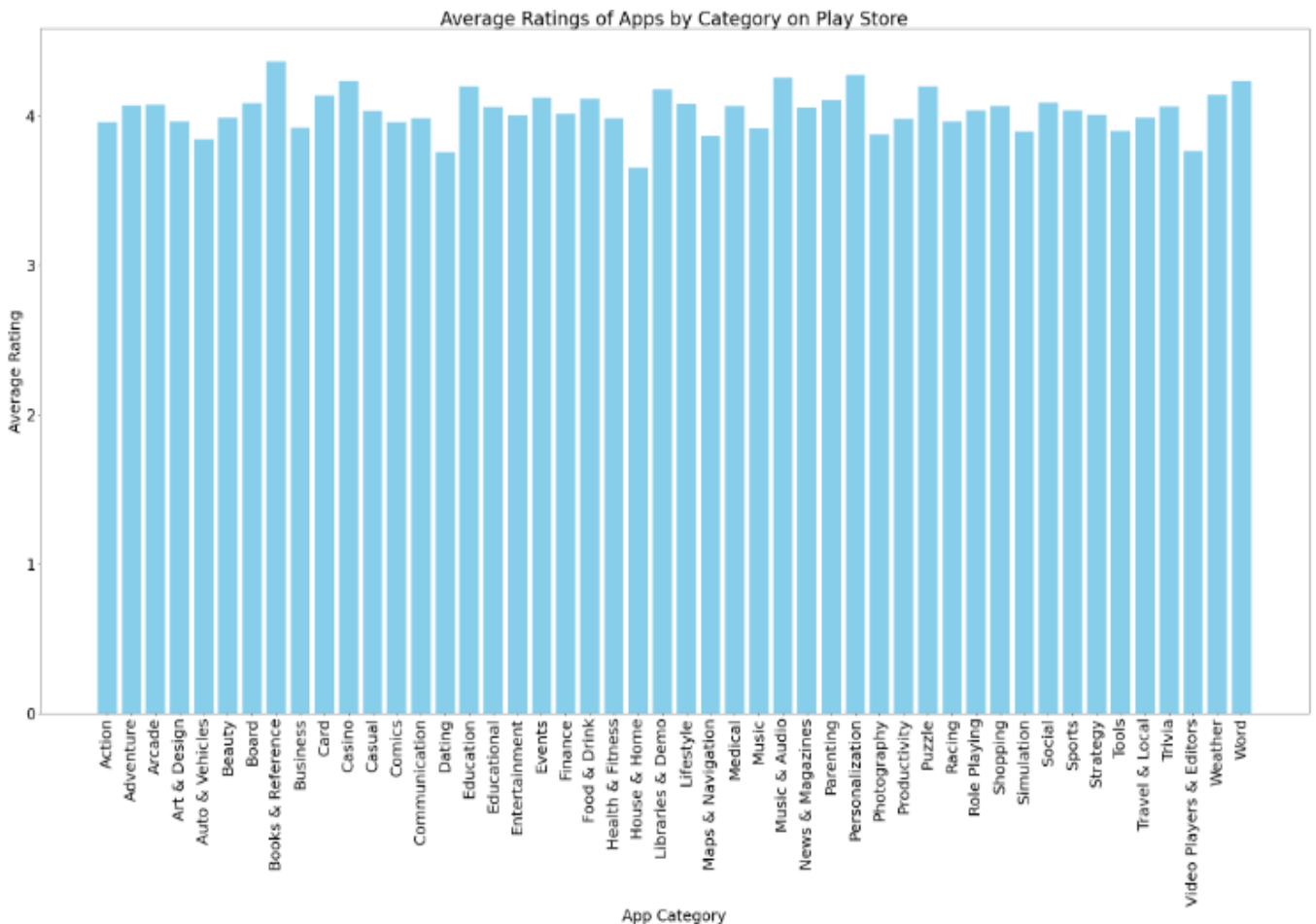


Here we have taken **Number of Apps** against **Currency**.

Inference :

From above graph we get **Currency Distribution** i.e. Which currencies are most common among apps on playstore. The longer the bar for a currency, the more apps use that currency.

we also get information about **Popular Currencies** which helps us to identify the most popular or widely accepted currencies within the app ecosystem on the Play Store.

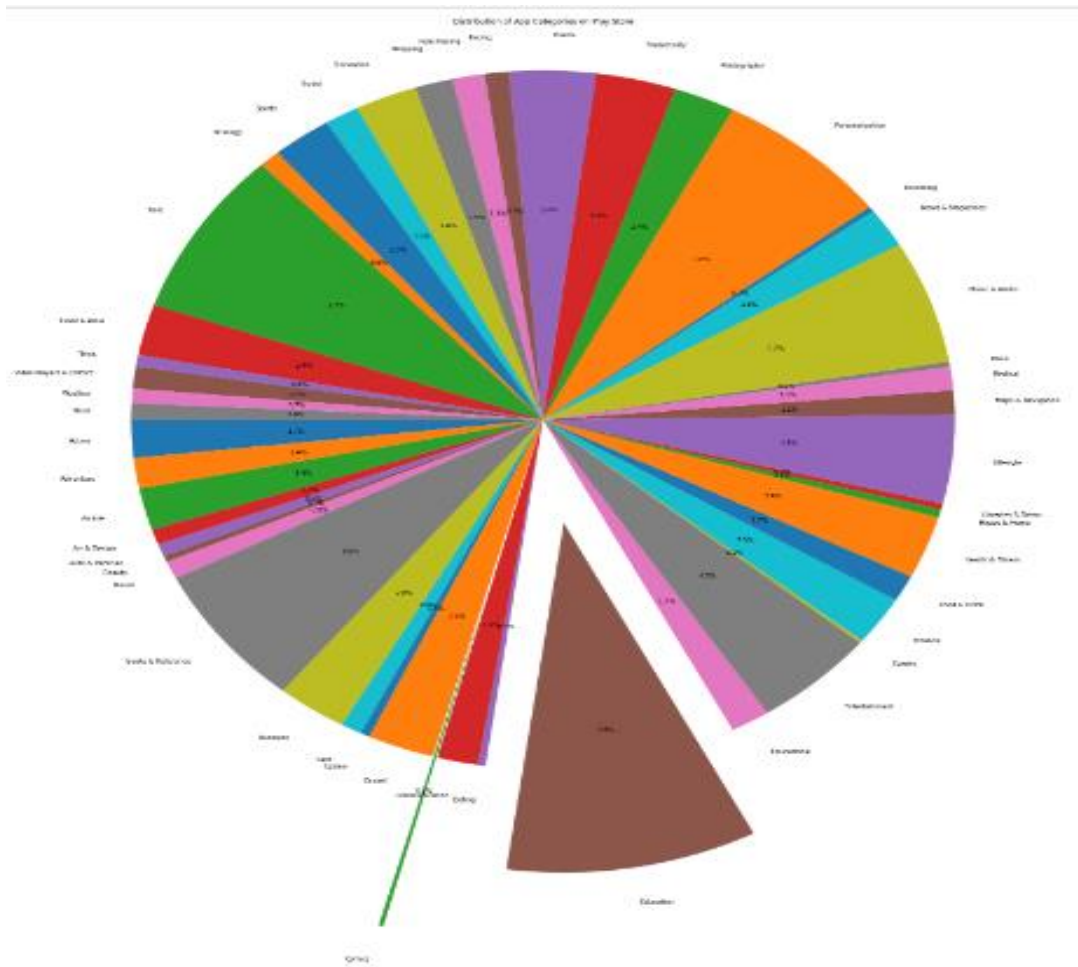


Here we have taken **App Category** against **Average rating**.

Inference :

from above graph we conclude **Category Variability** The plot shows the variability in average ratings across different app categories. Each bar represents a specific category, and the height of the bar represents the average rating within that category.

we can easily **compare the average ratings of different categories** by looking at the differences in bar heights.



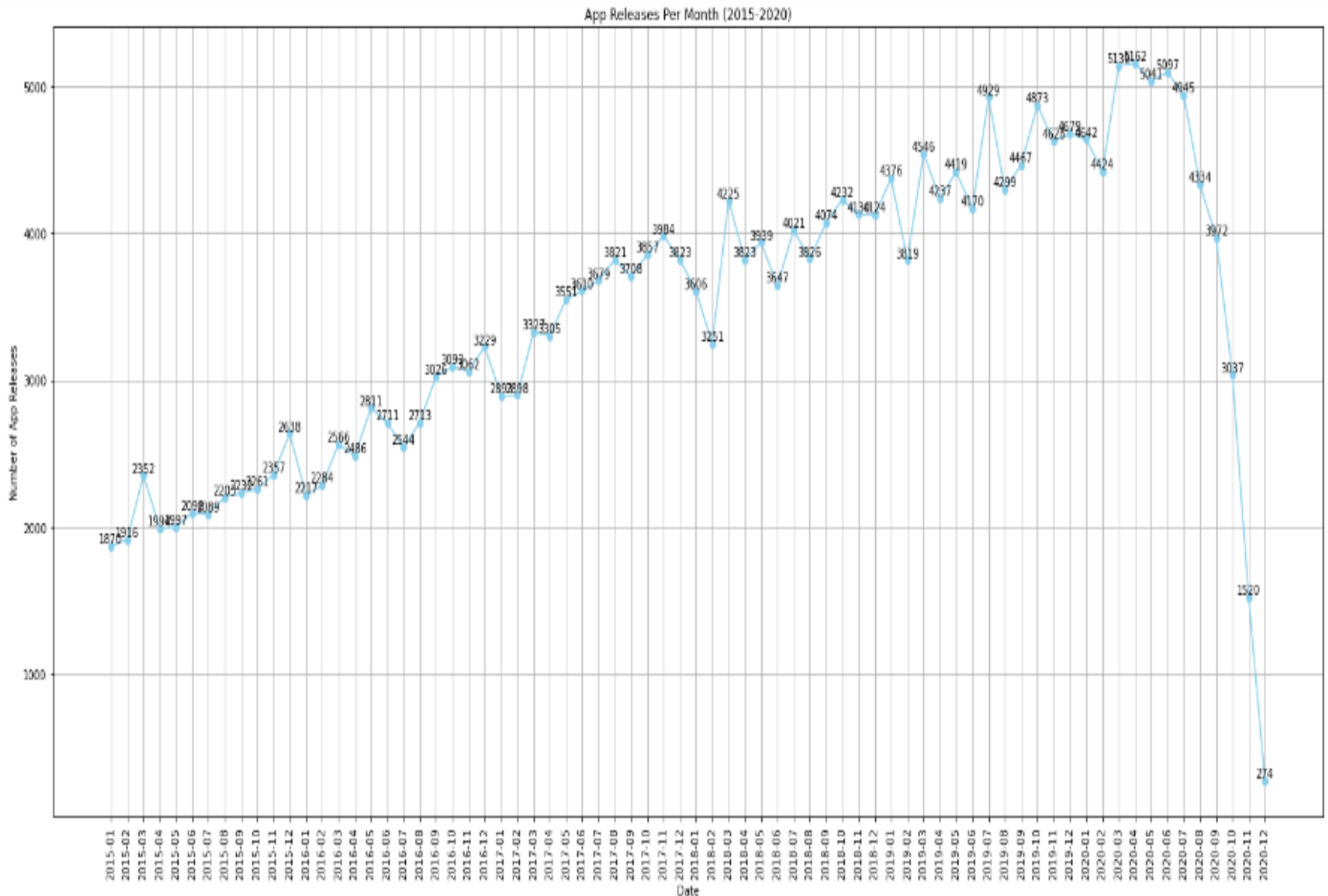
The pie chart illustrates the distribution of **app categories** available on the Play Store.

Inference :

Category Popularity i.e. The chart shows which categories of apps are more prevalent on the Play Store. Categories with larger slices have a higher number of apps, indicating their popularity among developers and users.

The size of each slice in the pie chart represents the proportion of apps in a specific category relative to the total number of apps in all categories combined. According to graph we get that **Education is Maximum Category** and **Comics is Minimum Category**

Playstore Data Analysis and Prediction



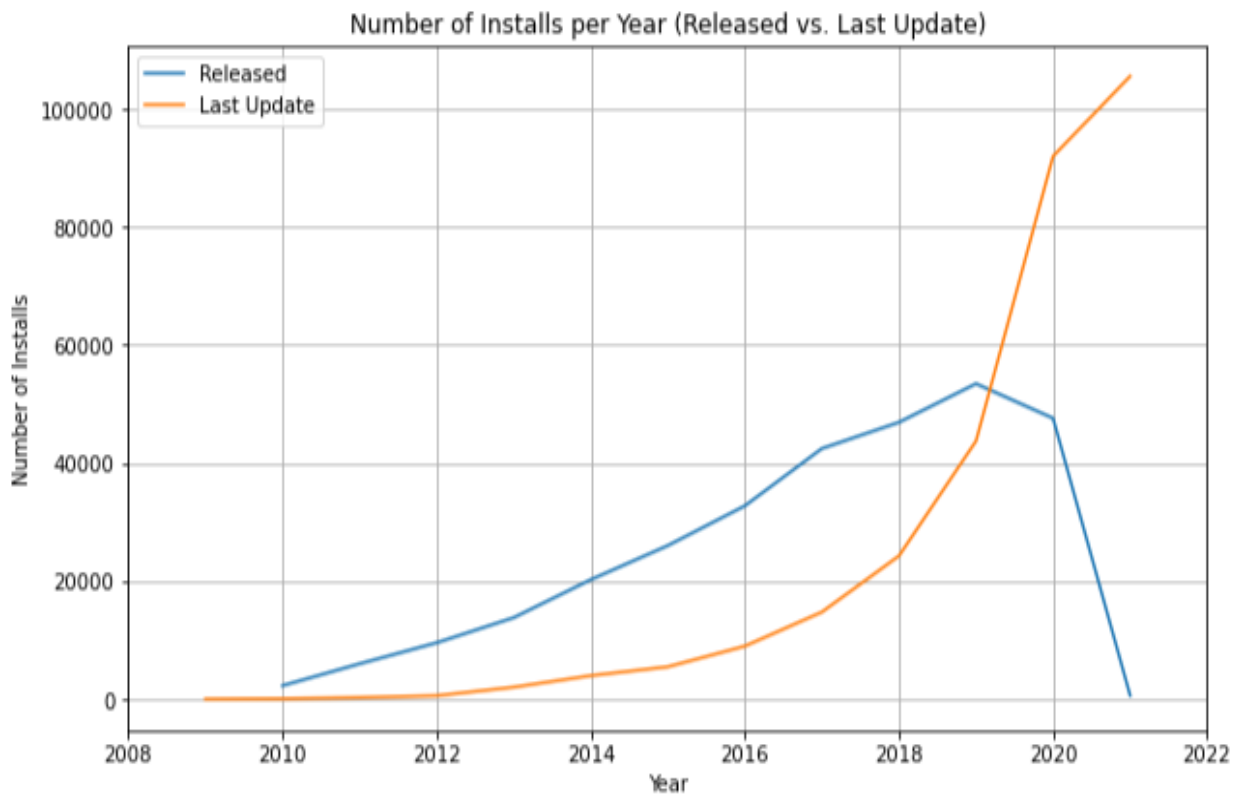
Here we have taken App released dates against No. apps released .

Inference :

In this graph annotated values give us **number of apps released in that specific Year/month.**

This graph allows us to see **how app releases are distributed over time.**

We also get **Monthly Trends** i.e. we can see if there are certain months that consistently have higher or lower numbers of app releases. we can observe that in 2020-12 there is huge drop in app releases



Here we have shown **Number of installs per year (Released vs Last update)**

Inference :

From above graph we got **Overall Growth** From 2008 to 2022, there has been a noticeable increase in the number of installs, as shown by the upward trend in both lines representing "**Installs based on Released**" and "**Installs based on Last Update.**"

we also get information about **peak years** There appear to be some peak years where the number of installs based on both release and last update is particularly high. These peak years may correspond to significant updates, marketing efforts, or other factors that attracted more users.

✚ **Hypothesis testing:**

We performed one sample z test to check average of sample and population of feature minimum installs.

Performed one proportion test to find positive reviews proportion to expected proportion

Performed two sample paired t-test to check if There is no significant difference between the measurements before and after an event or time period.

Performed Anova Test to Check if There is a significant difference between avg rating of the categories.

This helped us to analyze adata statistically and to find dependent and independent variables.

✚ **Feature engineering:**

After knowing which were the independent and dependent variables, feature engineering techniques were used for X and y. it is important step as the accuruacy will be dependent on X and y variable. It was try and error method because to get the highest accuracy we needed to check the dependency.

✚ **Implementation of algorithms:**

Before implementing the algorithms the data was splitted for train and test purpose. 80% of the data was for train and remaining 20% was for test. We also plotted heatmap and pairplot to get a better idea. Algorithms we used in this were Linear Regression, Naïve Bayes Algorithm, KNearest Algorithm, Decision Tree Algorithm, Random Forest Algorithm and Support Vector Machine. In all algorithms we plotted confusion matrix and overviewed classification report. On bases of y-test and y-pred we calculated accuracy. We also plotted AUC-ROC curve

✚ **Comparision of algorithms via accuracy:**

After implementing all the algorithms we created a dataframe of all accuracy with respect to the algorithms. So that we can clearly overlook which has highest accuracy. Mostly the accuracy were under 50% due to less number of data. But as the data will increase accuracy will also increase in upcoming days.

Regression algorithms Analysis:

Regression Analysis	
Algorithm Name	Mean Squared Error(MSE)
Linear Regression	0.358289069
Decision tree	0.393874278
Random forest	0.370034097
KNN Regresaor	0.415870479

From Regression Analysis ,we get that Linear Regression and Random Forest Regressor our best fit Regression Algorithms as they have less Mean Squared Error as compared to other regression Algorithms.

Classification algorithms Analysis

Logistic regression Classifier :

```
Accuracy: 0.942625
Classification Report:
              precision    recall  f1-score   support

     0           0.00        0.00        0.00        2295
     1           0.94        1.00        0.97       37705
```

Decision Tree Classifier :

```
Accuracy: 0.928425
Classification Report:
              precision    recall  f1-score   support

     0           0.38        0.38        0.38        2295
     1           0.96        0.96        0.96       37705
```

Random Forest Classifier:

Accuracy: 0.948875

Classification Report:

	precision	recall	f1-score	support
0	0.59	0.37	0.46	2295
1	0.96	0.98	0.97	37705

KNN Classifier :

Accuracy: 0.94085

Classification Report:

	precision	recall	f1-score	support
0	0.40	0.06	0.11	2295
1	0.95	0.99	0.97	37705

From Clasasification Analysis We conclude that our model has high accuracy and excellent performance for class 1 For Random forest Classifier.

5. Advantages

- The Google Play Store apps data analysis provides enough potential to drive apps making businesses to succeed.
- Actionable stats can be drawn for developers to work on and capture the Android market.
- Historical data can be used to make predictions about future trends and user behavior.
- By understanding which app features are most important to users and which are rarely used, businesses can prioritize development efforts and allocate resources more efficiently.
- Analyzing user reviews and feedback can help businesses understand user sentiments, pain points, and feature requests. This information can be used to enhance satisfaction.

6. Limitations

- Play Store data may not always be complete or up-to-date. Some apps may not provide comprehensive information, and data quality can vary.
- Ratings and reviews can be biased, and fake reviews can distort the analysis.
- External factors like changes in the mobile device market, economic conditions, or global events can influence app adoption and usage, making it challenging to isolate the effects of app-specific factors.

7. Applications

Commercial Applications

- To guide the development of new apps or updates to existing ones.
- Prioritize feature development based on user demand and app performance metrics.
- Identify markets with high growth potential based on user demand and trends.

User Point of View Applications

- Identify markets with high growth potential based on user demand and trends.
- Localize apps and marketing efforts to cater to different regions and languages.

8. Future Scope

- Understanding how users interact with apps, including the frequency of usage, duration, and user engagement metrics.
- Predicting user churn and identifying strategies for user retention.
- Identifying emerging trends in app categories and genres.
- Analyzing the impact of new technologies (e.g., AR/VR, AI) on app development.
- Analyzing app performance metrics (e.g., crash rates, loading times) and identifying areas for improvement.
- Benchmarking an app's performance against competitors.
- Identifying strengths and weaknesses relative to competitors.

9. Conclusion

- While there are many free apps on the Play Store, some premium (paid) apps also perform well in terms of downloads and ratings. Developers can consider both pricing models depending on their app's features and target audience.
- Some app categories experience seasonal fluctuations in popularity. For instance, fitness apps may see increased downloads at the beginning of the year due to New Year's resolutions. Developers can capitalize on these trends with targeted promotions.