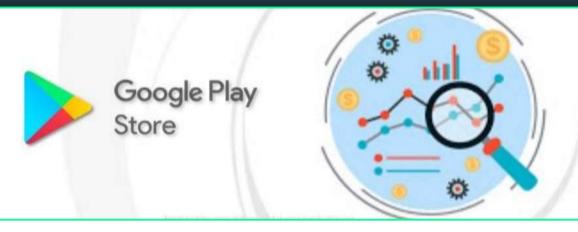# Capstone Project

Google Play Store Apps

Exploratory Data Analysis

BY POOVARASAN

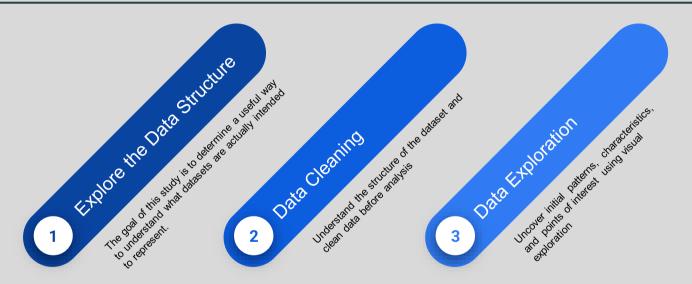shipfriend0368@gmail.com

Google Play Store

Why did I choose the google play store for analysis?

Google Play partners with developers to provide access to over 2 billion active monthly users. Distribution is free for developers who do not charge for their apps and games.

Companies may run beta focus groups, or app developers may receive feedback from testers and get certain amounts of reviews. We use this and some knowledge about the app to predict its success. Knowing the number of installs can help developers and business managers because they can predict the profit. This project's result may show the importance of reviews to apps in the market as it could be one of the determining factors for the number of installs.

OVERVIEW OF EXPLORATORY ANALYSIS

1 — Explore the Data Structure
The goal of this study is to determine a useful way to understand what datasets are actually intended to represent.

2 — Data Cleaning
Understand the structure of the dataset and clean data before analysis

3 — Data Exploration
Uncover initial patterns, characteristics, and points of interest using visual exploration

Explore the Structure of the Datasets

Play across different categories. We'll look for insights in the data to devise strategies to drive growth and retention. Let's take a look at the data, which consists of two files:

datasets/play_store_data.csv: This file contains all the details of the apps on Google Play. There are 13 features that describe a given app.

- **App:** Name of the app
- **Category:** Category of the app. Some examples are: ART_AND_DESIGN, FINANCE, COMICS, BEAUTY etc.
- **Rating:** The current average rating (out of 5) of the app on Google Play
- **Reviews:** Number of user reviews given on the app
- **Size:** Size of the app in MB (megabytes)
- **Installs:** Number of times the app was downloaded from Google Play
- **Type:** Whether the app is paid or free
- **Price:** Price of the app in US$
- **Content Rating:** A content rating (also known as maturity rating) rates the suitability of TV broadcasts, movies, comic books, or video games to its audience.To show which age group is suitable to view media and entertainment.

Continue…

Explore the Structure of the Datasets

- **Genres:** A category of artistic, musical, or literary composition characterized by a particular style, form, or content
- **Last Updated**: Date on which the app was last updated on Google Play
- **Current Ver**: Current Version means a version of the software that is currently being supported by its publisher.
- **Android Ver:** Android versions (codenames) are used to describe the various updates for the open source Android mobile operating system.

datasets/user_reviews.csv: This file contains a random sample of 100 user reviews for each app. The distribution of positive and negative reviews in each category has been pre-processed and passed through a sentiment analyzer.

- **App:** Name of the app on which the user review was provided. Matches the App column of the play_store_data.csv file
- **Translated Review**: The pre-processed user review text.
- **Sentiment:** Sentiment category of the user review - Positive, Negative or Neutral.
- **Sentiment Polarity:** Sentiment score of the user review. It lies between [-1,1]. A higher score denotes a more positive sentiment.

Continue…

Explore the Structure of the Datasets

```
[ ] # head of the play_store dataframe

    play_store.head()
```

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

```
[ ] # First 5 rows of Review Dataframe

    Df_review.head()
```

	App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
1	10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
2	10 Best Foods for You	NaN	NaN	NaN	NaN
3	10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000
4	10 Best Foods for You	Best idea us	Positive	1.00	0.300000

Data Cleaning

Basic Data Cleaning

Handling the Null values in dataset

Handling Duplicates

Quick Check for Outliers

Checks and Convert the Following columns for EDA analysis

Installs

Size

Price & Reviews

- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
- Some time when combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.
- There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Continue…

Data Cleaning

Handling the Null values in dataset

The first step is to detect the count of missing values in every column of the dataset. This will help us an idea about the distribution of missing values.

Handling Duplicates

Removed unwanted observations from your dataset, including duplicate observations or irrelevant observations.

Converted the few of columns for EDA analysis
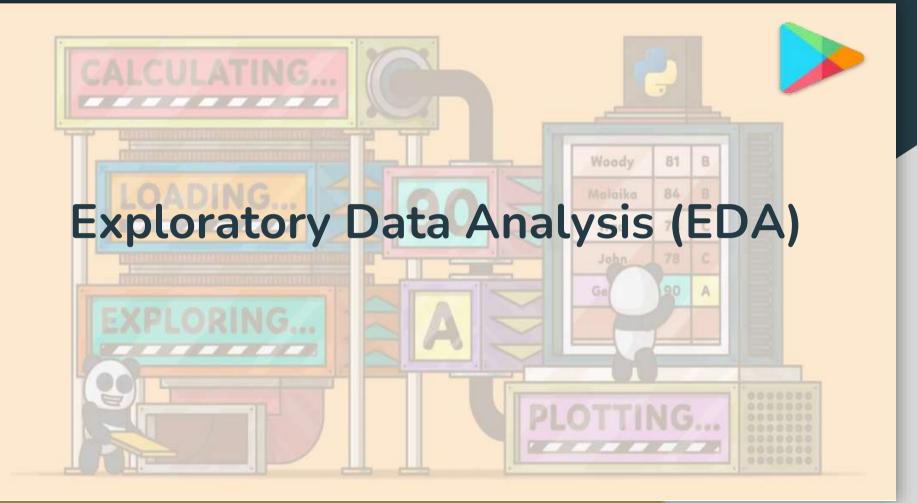
As some of feature column are non-numerical value so its converted the categorical variables the into numerical for ease analysis.
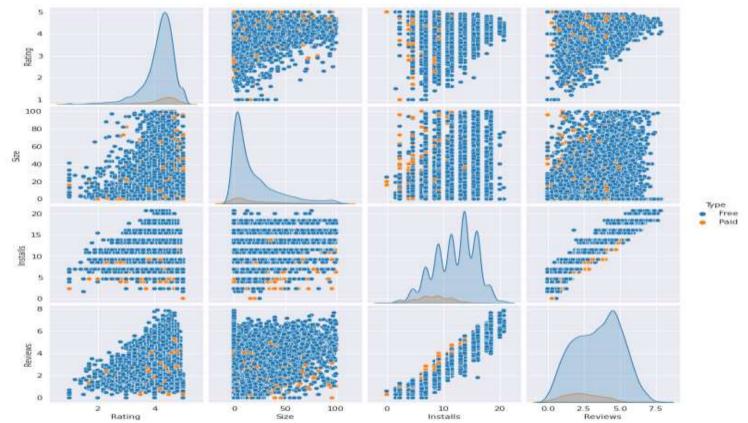
Quick Check for Outliers

On studying the dataset further, it was found that there was a data with some kind of weird anomaly. Let us find out the row in the data and purge it.
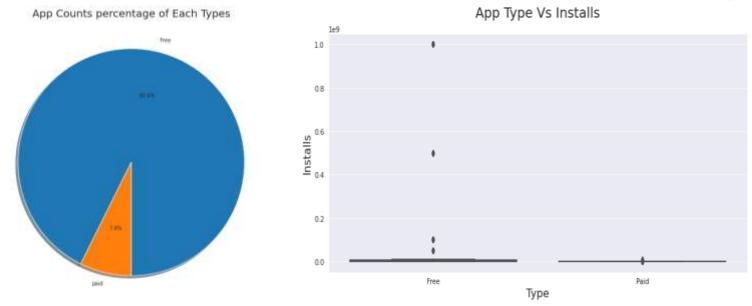
01
02
03
04

NAME	CITY	AGE	PY-SCORE
101	Xavier	Mexico City	41	88.0
102			28	79.0
			35	81.0
Yi	Shang		80.0
Roblin	Manchester		68.0
Amal	Cairo		61.0
Mori	Osaka		

Exploratory Data Analysis (EDA)

BIVARIATE ANALYSIS

1. Type that is appreciated by the majority of people?

App Counts percentage of Each Types
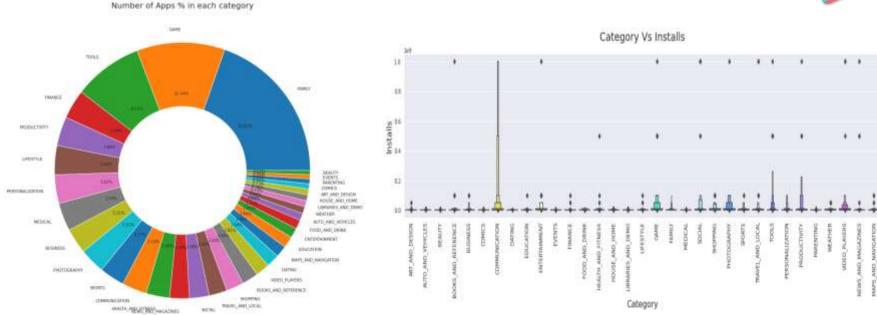
App Type Vs Installs

- The chart above shows that 92% of Play Store applications are free and 8% are on sale.
- Free apps are installed more frequently than expensive apps, as was predicted.

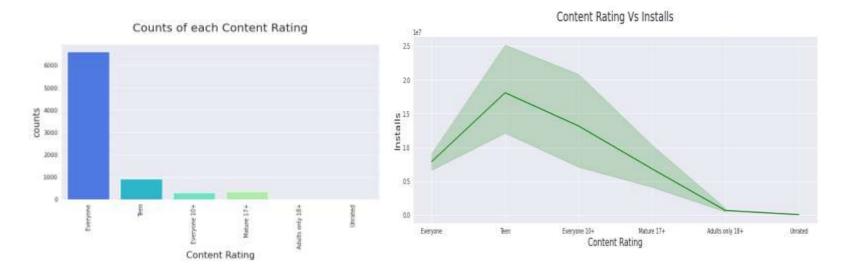## 2. How Category affects the installation ?

Number of Apps % in each category

Category Vs Installs

- Consequently, the dataset contains 33 categories in total. The results above indicate that the majority of apps in the Play Store fall into the FAMILY & GAME category, while the least number of apps go into the EVENTS & BEAUTY category.

- From this, it can be observed that the most often installed applications fall into the categories of communication, social, video players, entertainment, and productivity.

3. How content Rating influences Download ?

Counts of each Content Rating
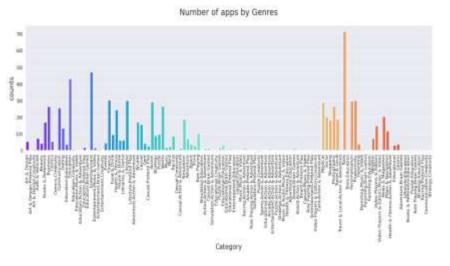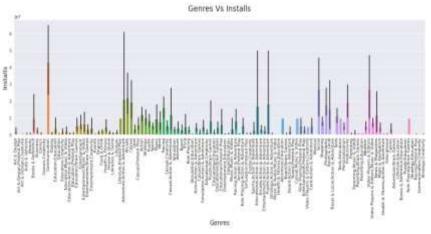

Content Rating Vs Installs

- A majority of the applications in the play store are available for everyone to use. Other applications have different age limits for using it.

- Installed are any age-appropriate apps other than those marked "Adults only 18+" and "Unrated." The 'Teen' and 'Everyone 10+' applications, on the other hand, have been downloaded more than 30 million times.

4. What kind of genres boost the downloads of apps?

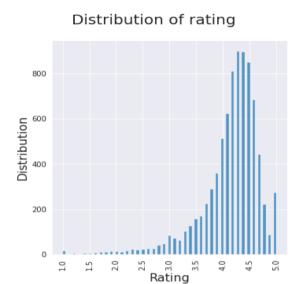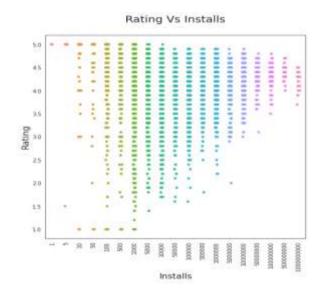Number of apps by Genres

Genres Vs Installs

- According to the results above, the vast majority of apps in the Google Play Store belong into the Tools, Entertainment and education category.

- Apps in the genre "Communication" do incredibly well. The following genres are also worth developing an app for: Social, Video Players & Editors, Arcade, Adventure; Action & Adventure, Casual, Productivity, Puzzle; Action & Adventure, Photography, Educational; Action & Adventure

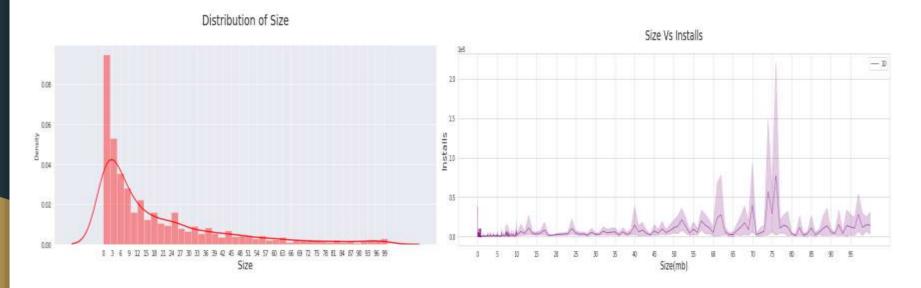# 5. Which rating has the most impact on download rates?

Distribution of rating

Rating Vs Installs

- The conclusion reached from this is that apps with ratings between 4.2 and 4.5 are downloaded more than 40 million times, while those with ratings between 3.9 and 4.2 as well as 4.6 are downloaded between 1 million and 2 million times. The number of installations for all the others is less than one million.

6. Which size of the app has received the most downloads?

Distribution of Size

Size Vs Installs

- More than 1 million people have downloaded apps between 72 and 76 Mb in size, followed by 0.5 to 1 million people who have downloaded apps between 61, 70, 97, and 98 Mb in size. Less than 0.5 million of the remaining apps have been downloaded, and others have never been downloaded at all.

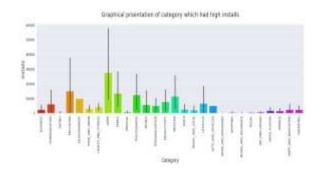# 7. Which Android version has received the most downloads from users?
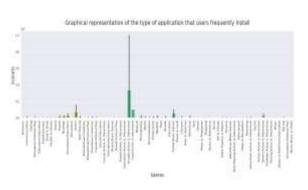

Android Ver Vs Installs

- More than 100 million apps have been downloaded for Android versions 4.1-7.1.1, followed by 33 million downloads for Android versions 7.1 and higher. Less than one million people download the remaining apps, and some people never download any at all.
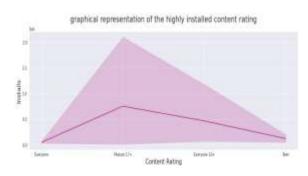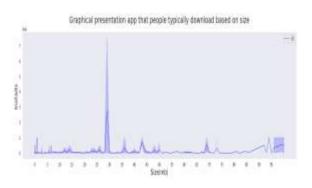
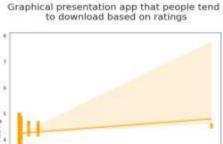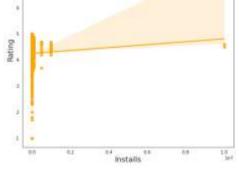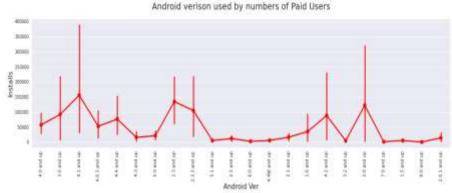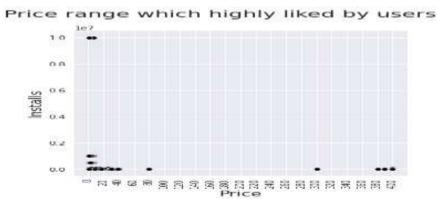# ED Analysis for Paid Apps

Results of a paid app analysis

Continue…

The study above shows:

- The majority of consumers spend money to download games, educational apps, and family apps.

- most paying users pay to download apps in the "Mature 17+"

- Arcade; Action & Adventure-themed apps are very popular.

- Ratings of 4.5 and 4.6 did exceptionally well in the market. The majority of premium customers only downloaded apps based on these two ratings.

- Nearly 2 million paying consumers downloaded apps with sizes between 27 and 30 megabytes, as shown in the following graph. These apps with these sizes performed excellently in stores.

- It is preferable to design apps with Android version 4.1 or higher because, according to the above graph, the majority of paying consumers download apps with that version of Android.

- Since the majority of paid customers pay between 1 to 13 Doller's to download an app, it is highly recommended to set the pricing of an app between those ranges. The graph above shows that at this rate, applications are downloaded more than 1 million times.

Android verison used by numbers of Paid Users

Price range which highly liked by users

Data Visualization for user_review Data frame
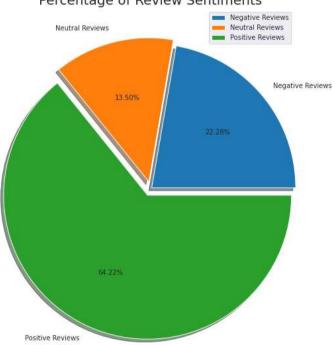
Results of a user_review analysis

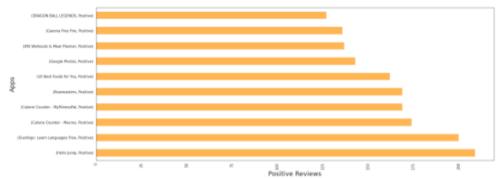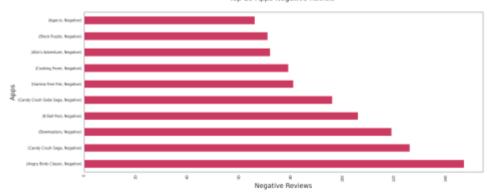Percentage of Review Sentiments

Top 10 Apps Positive Review

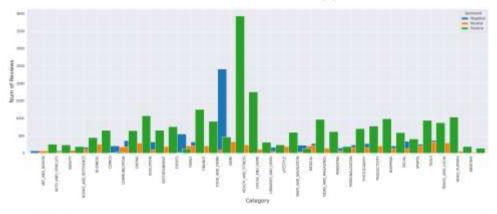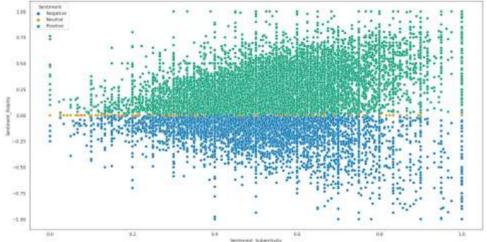Top 10 Apps Negative Review

Continue…

The study above shows

- Based on the above charts, 64% of reviews are positive and 22% are negative; the remaining reviews are regarded as neutral.

- According to the graph above, as sentiment polarity increases, sentiment subjectivity also rises and as subjectivity decreases, polarity also falls.

- According to the graph above, more than 11k individuals have submitted favorable reviews for games, and more than 7k people have done the same for family and health-and-fitness apps.

Aha!
My observations

What have you discovered after exploring and analyzing?

Size: App with the smallest size earned the highest ratings

Review: The highly reviewed applications has received numerous downloads from users. The majority (64%) of apps in the Google Play store received generally positive reviews.

Price: Since the majority of customers who purchase applications do so by paying between $1 and $3 to download them.

Conclusion

We have analyzed a variety of parameters as part of this study to analyze play store applications that would help Alma Better launch their apps successfully. In order to give them the best results from our study, we concentrated more on the problem statements and data cleaning in the initial phase.

Based on analysis, there is minimal information available regarding the play store datasets

- 92% of apps are available for free in play store.

- There are 33 categories in the dataset as a whole. The majority of play store apps fall under the FAMILY & GAME category.

- Most of the apps in the google play store are available to everyone.

- The genres of tools, entertainment, and education make up the great majority of apps in the google play store.

- The majority of apps in the google play store run on android 4.1 or later.

Continue…

I would advise anyone creating an app for the play store to follow these standards.

Genres

- Communication
- Social
- Video Players & Editors
- Arcade
- Adventure; Action & Adventure

Category

- COMMUNICATION
- SOCIAL
- VIDEO PLAYERS
- ENTERTAINMENT
- PRODUCTIVITY

Android versions

- 4.1-7.1.1

Size

- 70 to 76 (MB)

Content Rating

- Teen
- Everyone 10+

I would suggest anyone developing a paid app for the Play store to stick to these requirements.

Genres

- Arcade;Action & Adventure
- Card; Action & Adventure, Action
- Puzzle;Brain Games
- Sports; Action & Adventure

Category

- GAME
- EDUCATION
- FAMILY
- PHOTOGRAPHY
- WEATHER

Android versions

- 4.1 and up

Size

- 0 to 29 (MB)

Content Rating

- Mature 17+
- Everyone 10+

Price

- 1 to 13 Doller's

Thank You