Capstone Project - 1

Play Store App Review Analysis

By: Pooja Dalwani
Cohort Hardeol

Roadmap

- Introduction
- Defining Problem Statement
- Data Summary
- Data Cleaning
- Exploration and Visualization
- Inferences and Conclusion

Introduction

- **Importance of apps**
  - An app for every utility
  - Their importance can’t be overstated
- **Benefits to a business**
  - Greater reach due to smartphones
  - Loyalty and increased customer engagement
- **Challenges and Opportunities**
  - Huge supply: 3.48 Mn apps, increased competition
  - Learn and leverage from existing apps; increase customer satisfaction and success
- **Google Play Store dataset**
  - Apps and features
  - User Reviews

Defining Problem Statement

As we know that, there is no shortage when it comes to availability of apps. But one wishes to have the best app for some utility. The challenge is to create such an app, in the current competitive environment, and leverage from existing data. Our aim is to discover the factors/features on which the success (Installs) of an app depends.

Data Summary (for App and Features dataset)

- App: Name of the app
- Category: The category a particular app belongs to, e.g., “Game”, “Travel and Local”, “Dating” etc.
- Rating: An average of user rating out of 5 given at that time when this data was extracted (as opposed to average of a life time).
- Reviews: Count of number of reviews received by an app.
- Size: The amount of space an app will occupy in your device storage.
- Installs (Independent variable): Number of times an app has been downloaded since the time of its launch.
- Type: Whether the app is free or a paid one.

Data Summary (Cont.)

- Price: The amount you have to pay if you purchase the app.
- Content Rating: An indication of that age-group, for which the content of an app is suitable. E.g., many a times some apps/content are rated as “X”. Meaning that, people below 17 years of age are not allowed or advised to watch that particular content.
- Genre: Similar to category
- Last Updated: The date when new additions/features were introduced in the app.
- Current Version: Version of the app being used.
- Android Version: That minimum version on your android device, which is required for an app to function.

Data Summary (for User Reviews dataset)

- App: Name of the app
- Translated Review: Various textual reviews given by users.
- Sentiment: Indication of overall emotion of a particular user review. It can be identified either as positive, negative, or neutral.
- Sentiment Polarity: Polarity is simply a numeric representation of the overall sentiment. It gives us the degree of an emotion. This number can lie between -1 and 1. "-" sign representing a negative emotion and "+" a positive emotion.

Data Summary (Cont.)

- Sentiment Subjectivity: It gives us a measure of the subjectivity or objectivity of a particular review/statement. A statement will be considered as subjective, when it is more of a personal opinion (which can differ from individual to individual), and it can be considered objective when the review is more of a fact rather than an opinion (it cannot change from individual to individual). This number can lie anywhere between 0 and 1. Less than 0.5 meaning it is more of a fact, and greater than 0.5 meaning it is more of an opinion.

Data Cleaning

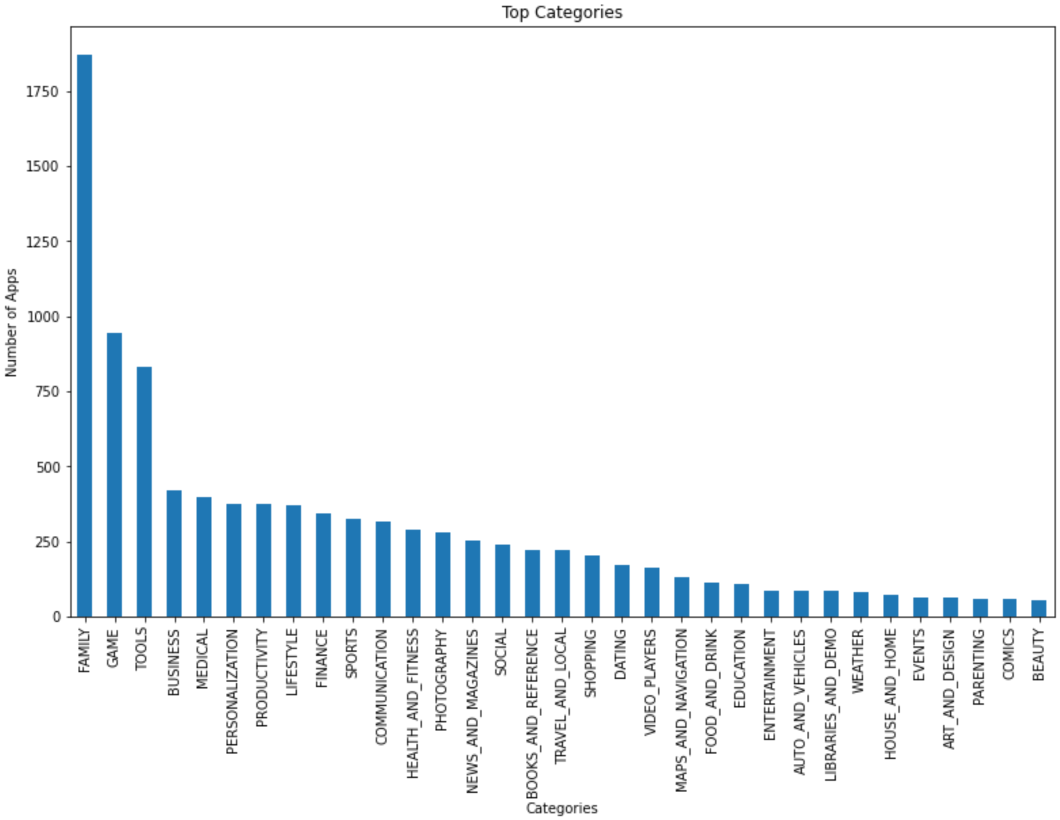
The first and most crucial step without which one cannot hope to proceed. Before diving straight away into the analysis part, it is crucial to make the dataset analysis-ready first. In this step, I have done the following:

1. Removing unnecessary rows and duplicates
2. Checking and converting the datatype of variables
3. Treatment of null values
4. Ensuring one row for each app

Exploration and Visualization

The next step involves basic exploration of the dataset, its variables; both dependent and independent, and the relation between them. In the next few slides, the visualization of the same has been covered.

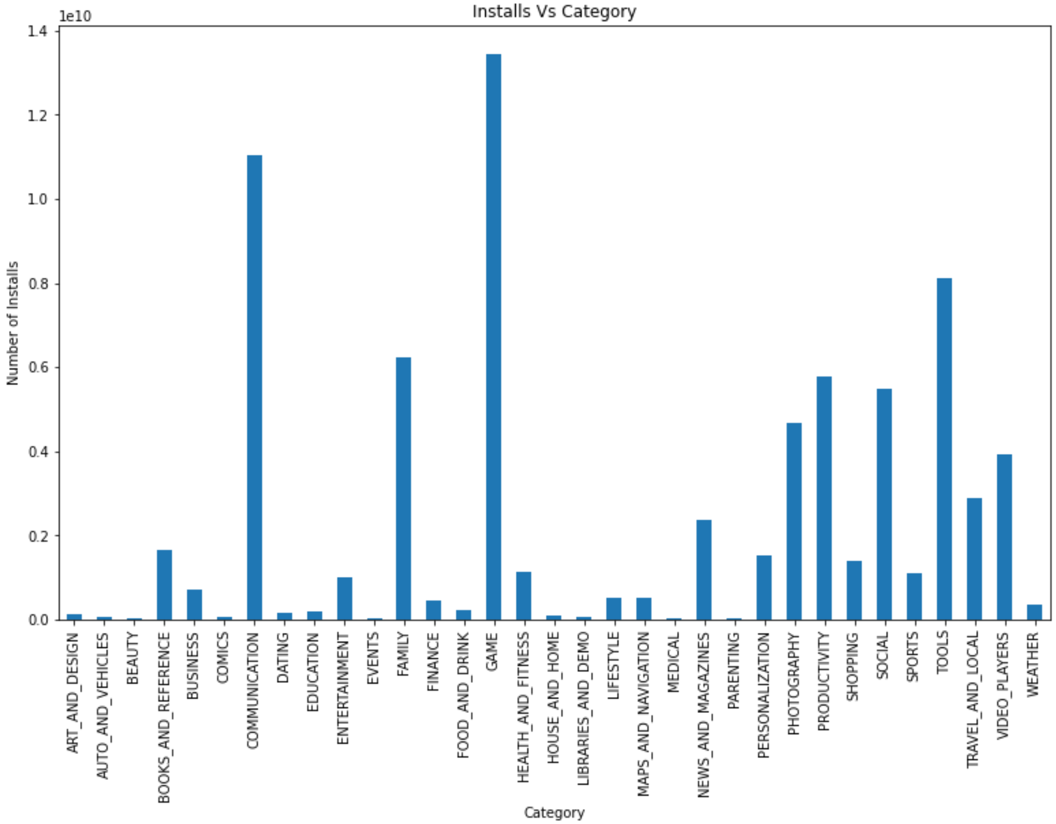
Exploration and Visualization (Cont.)



- Maximum apps offered by: Family, Game, and Tools

- Least apps offered by: Parenting, Comics, and Beauty

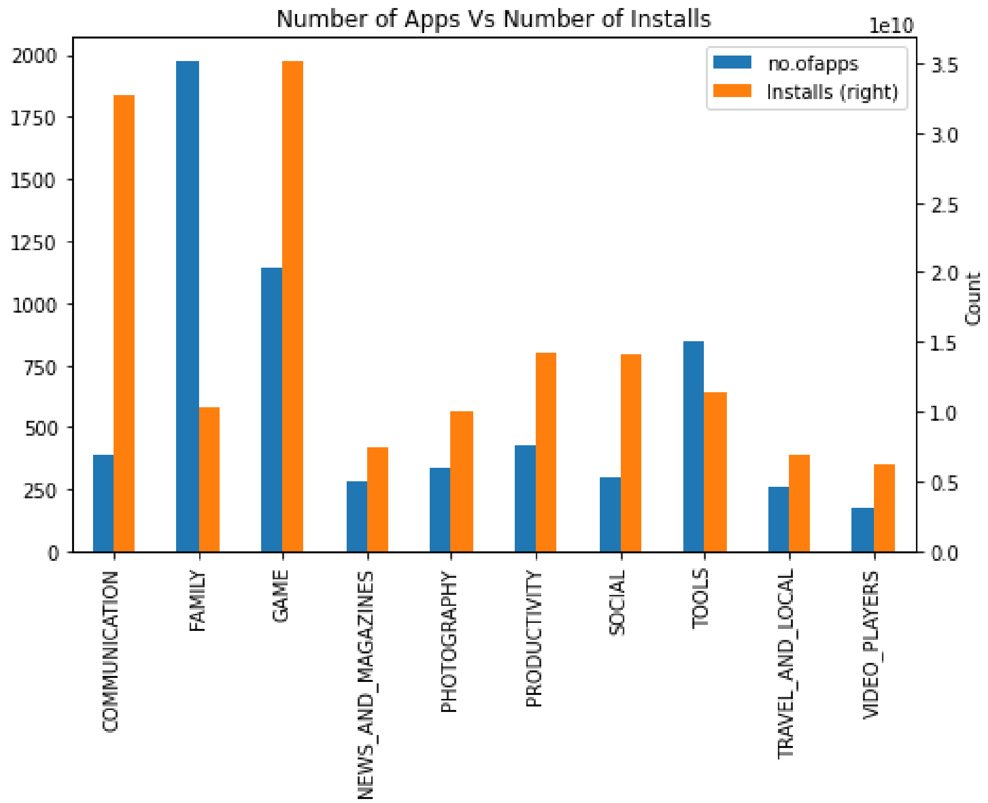
Exploration and Visualization (Cont.)



- ➢ Most installed categories: Game, Communication, and Tools

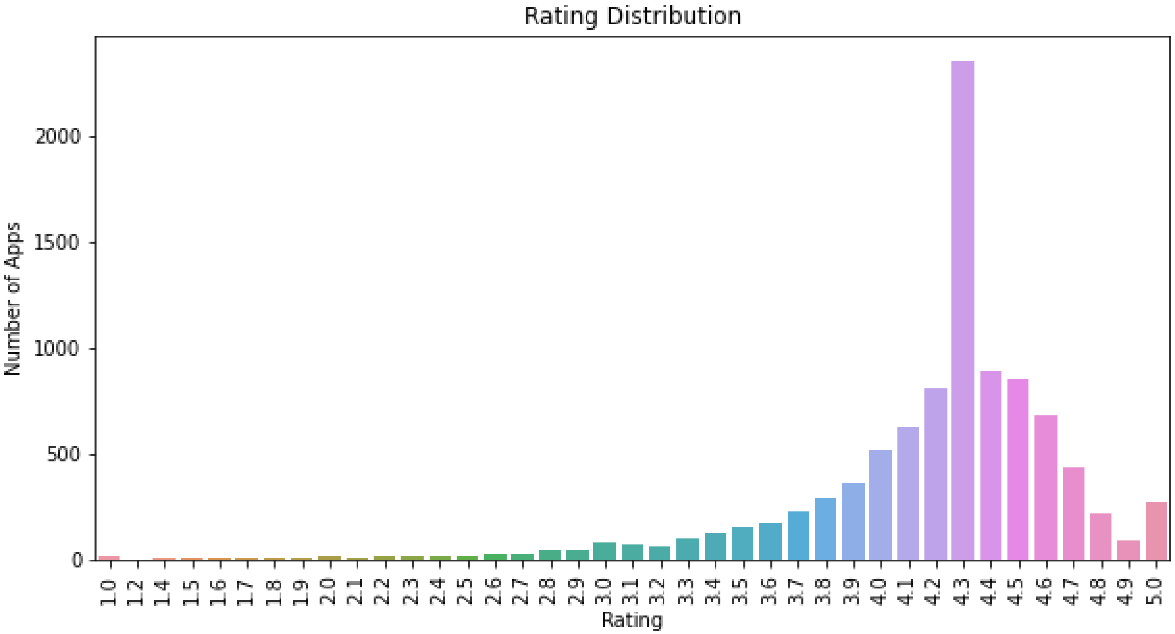
- ➢ Least installed categories: Medical, Parenting, Events, and Beauty

Exploration and Visualization (Cont.)



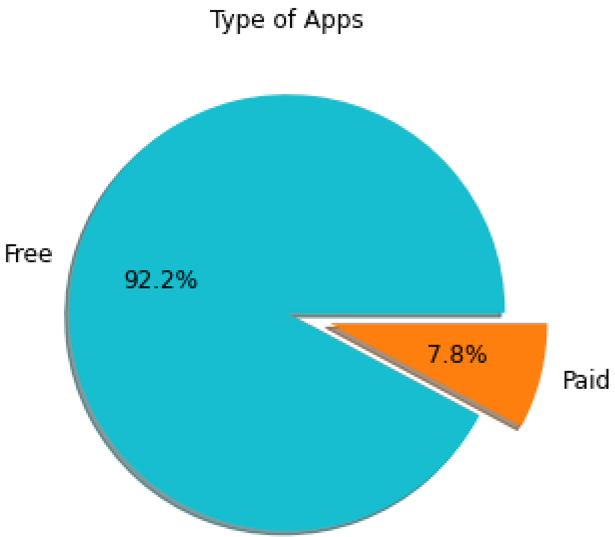
- ➢ A better way to compare both
- ➢ Although Family has many apps available, installs are very less compared to it. On the other hand, Communication offers less number of apps and installs are very high in comparison to it.

Exploration and Visualization (Cont.)



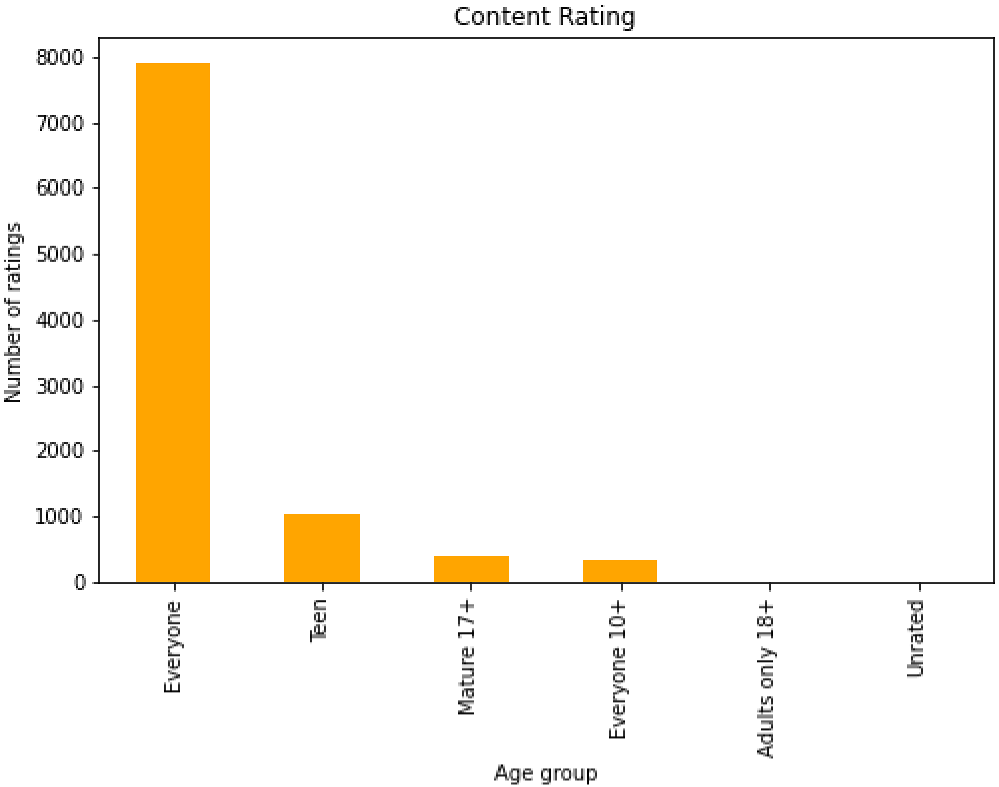
- Most of the apps lie between 4 and 5, and majority on 4.3

Exploration and Visualization (Cont.)



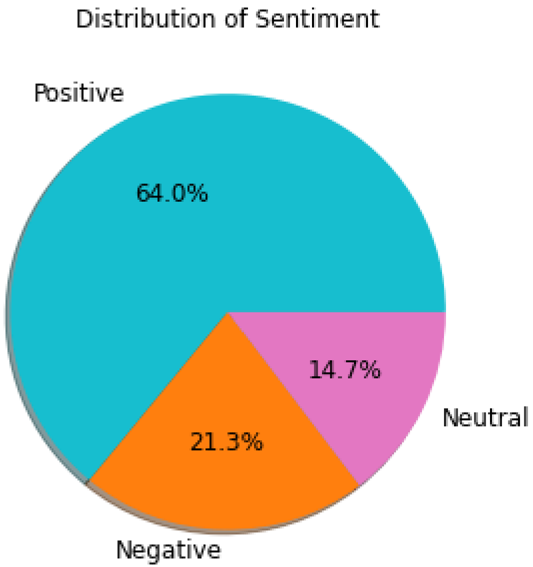
- Approximately 92% of the are free of nature, and 8% paid.

Exploration and Visualization (Cont.)



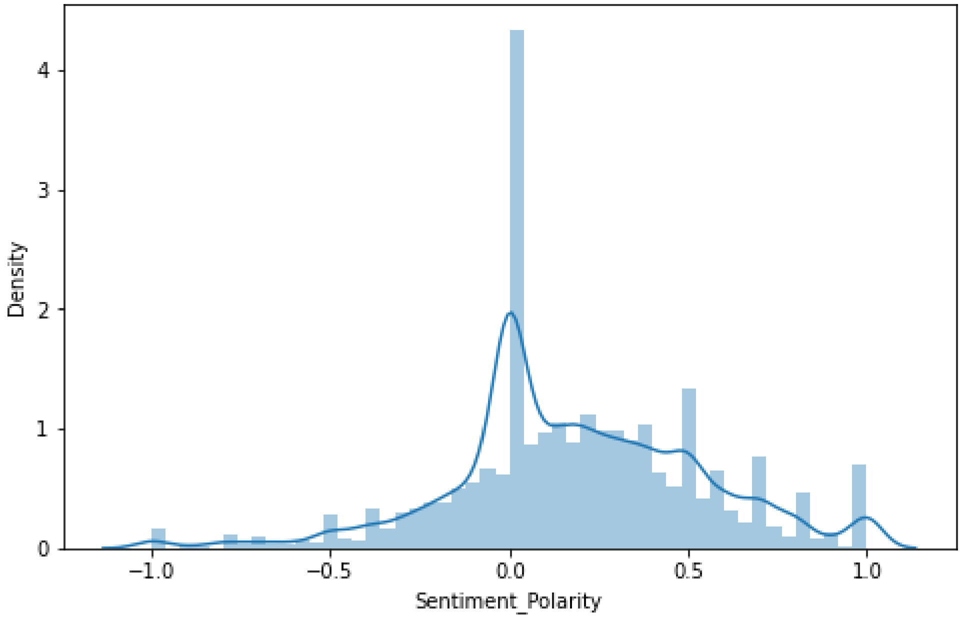
- Majority of the apps have been rated for Everyone, which gives a scope of huge market. Where as few apps are restricted for Under 17, and some specifically cater to teenage group.

Exploration and Visualization (Cont.)



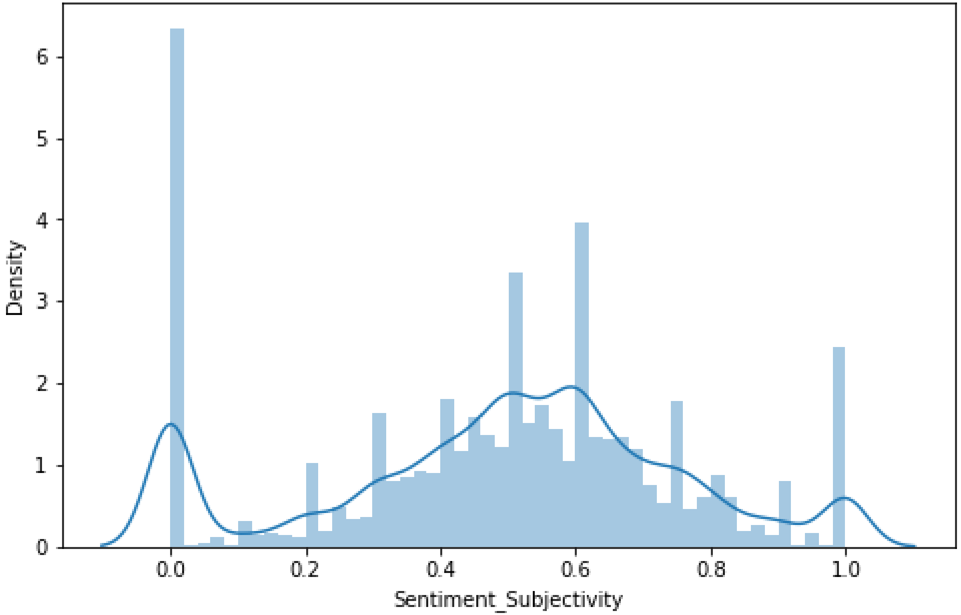
- Out of all the reviews, maximum are positive, and about 36% negative and neutral combined

Exploration and Visualization (Cont.)



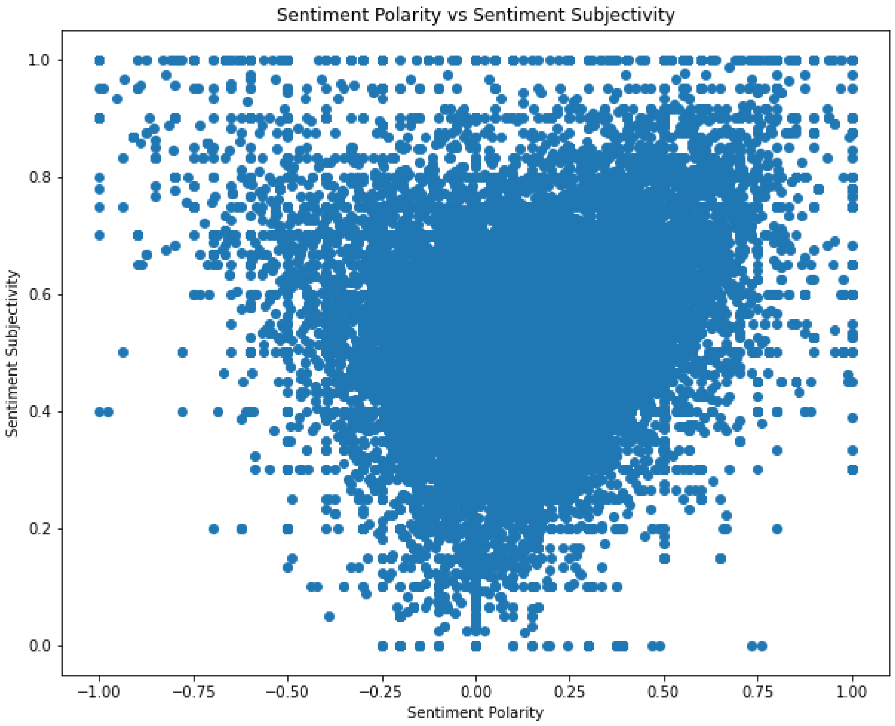
- As we know that maximum reviews are positive, the frequency is higher between 0 and 1.

Exploration and Visualization (Cont.)



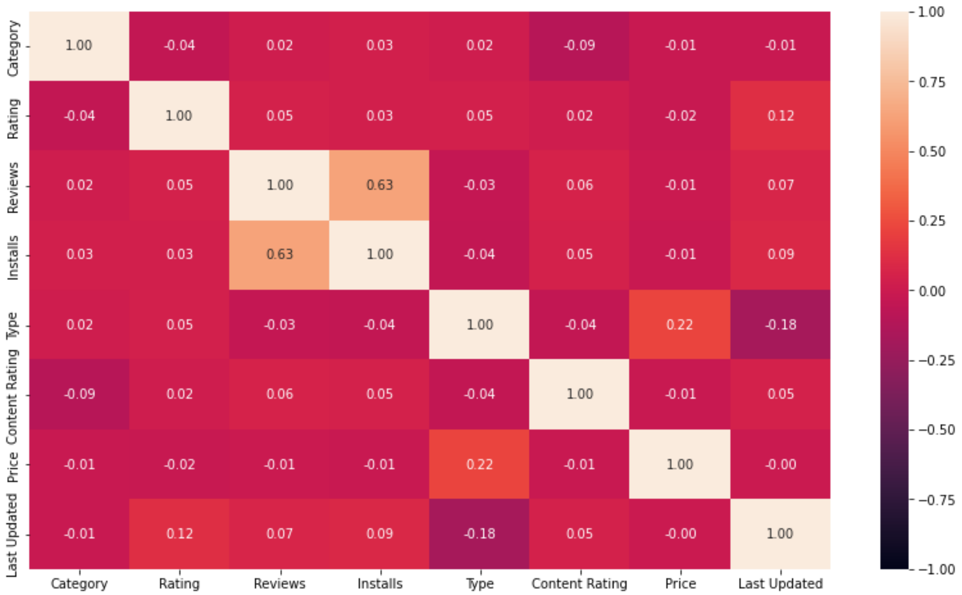
- Distribution of Sentiment Subjectivity. It is hard to tell which side is bigger from this figure.

Exploration and Visualization (Cont.)



- ➢ Here we can see that Sentiment Subjectivity and Sentiment Polarity are not proportional. We have more amount of positive reviews that are personal opinions.

Exploration and Visualization (Cont.)



- As we can see, unlike we had thought, correlation does not much depend on factors like size, rating, type etc. The maximum correlation it has, is with Reviews, as they go hand in hand.

Inferences and Conclusion

- Population / market size plays an important role as seen in as seen in slide #13. Despite having limited apps under “Communication” it has amazing number of installs due to its market size. It is useful across all age groups and segments.
- Installs does not have much correlation with these features like Size, Rating, Type, Price etc.

Challenges

- Time consuming data cleaning
- Indefinite features

THANK YOU