## **Dataset understanding**:

The raw dataset contains 10,841 records and 13 attributes. Our target variable is the attribute ‘Rating’, which varies from 0 to 5, 5 being the highest possible rating. The dataset is a combination of numerical and string attributes.

## **Dataset cleaning and pre-processing**:

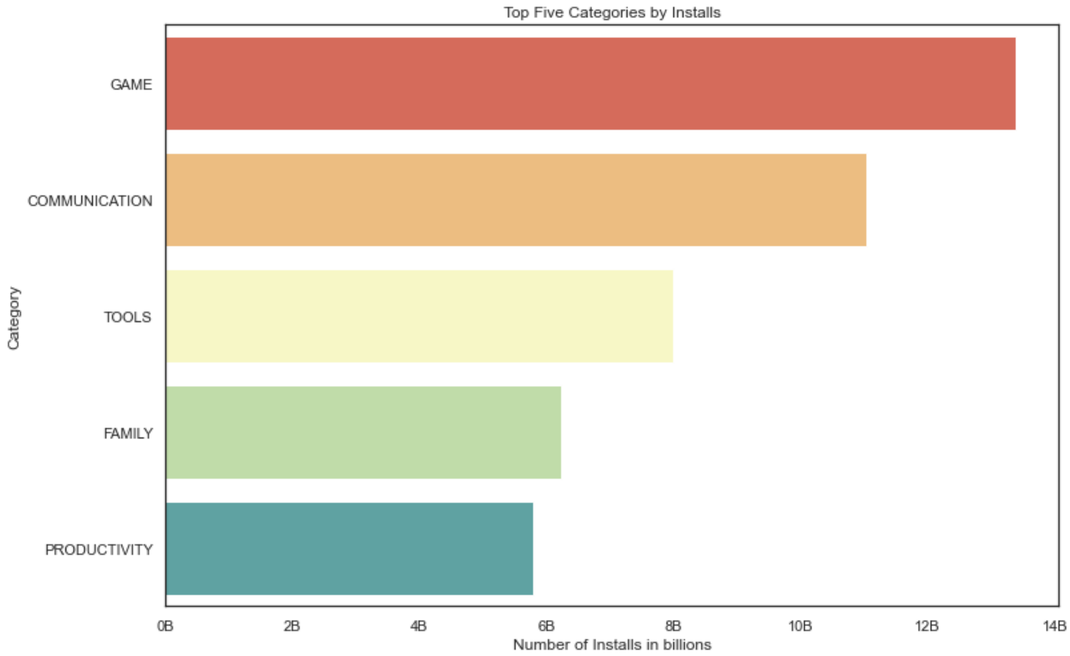
The steps we took to clean and pre-process data:

1. Listed out all column names and renamed some of them (added underscores) to enhance readability and for ease of use.
2. Found the number of nulls in each column. ‘Rating’ column had the greatest number of nulls (1474) followed by a few other columns like ‘Current_Version’, ‘Android_Version’, ‘Type’, etc. Since this is unwanted data, we dropped all records with null values. The number of records dropped to 9360.
3. The ‘Price’ attribute had a ‘$’ sign and comma separators in between the digits. This makes it difficult to treat price as a numerical value and perform calculations and analysis. So, we removed these characters from the ‘Price’ feature.
4. Similarly, the ‘Installs’ feature also had unnecessary characters like ‘+’ and ‘,’ which we removed.
5. Next, we checked the dtype of each feature and found that all but ‘Rating’ was listed as an object. We changed ‘Price’ to ‘float’ and ‘Installs’ to ‘int’ data type to make them easier to work with.

Our final cleaned dataset consists of 9360 records and 13 columns.

## Dataset Exploration and Visualization:

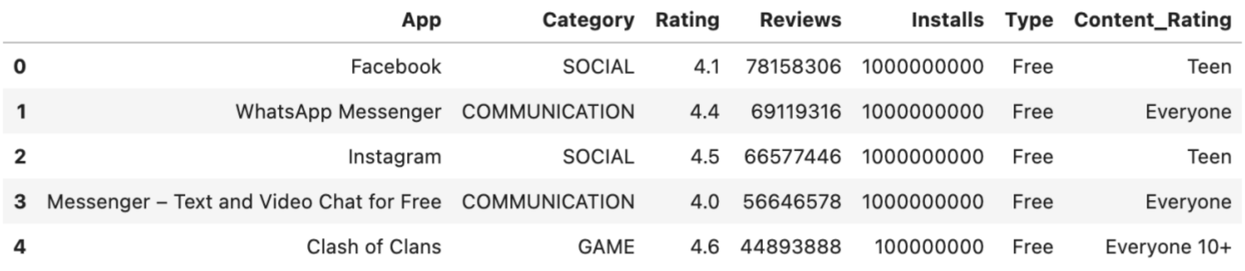
## 1. Top 5 App Categories based on Number of Installations



*Observations:*

- From the chart we can see that **most apps downloaded** on Playstore are either **Games or Communication** based.
- This is a great representation because entertainment is top priority and communication services like Whatsapp and Messenger are essential in today's world.

## 2. Top 5 Highest Rated Apps based on Number of Reviews

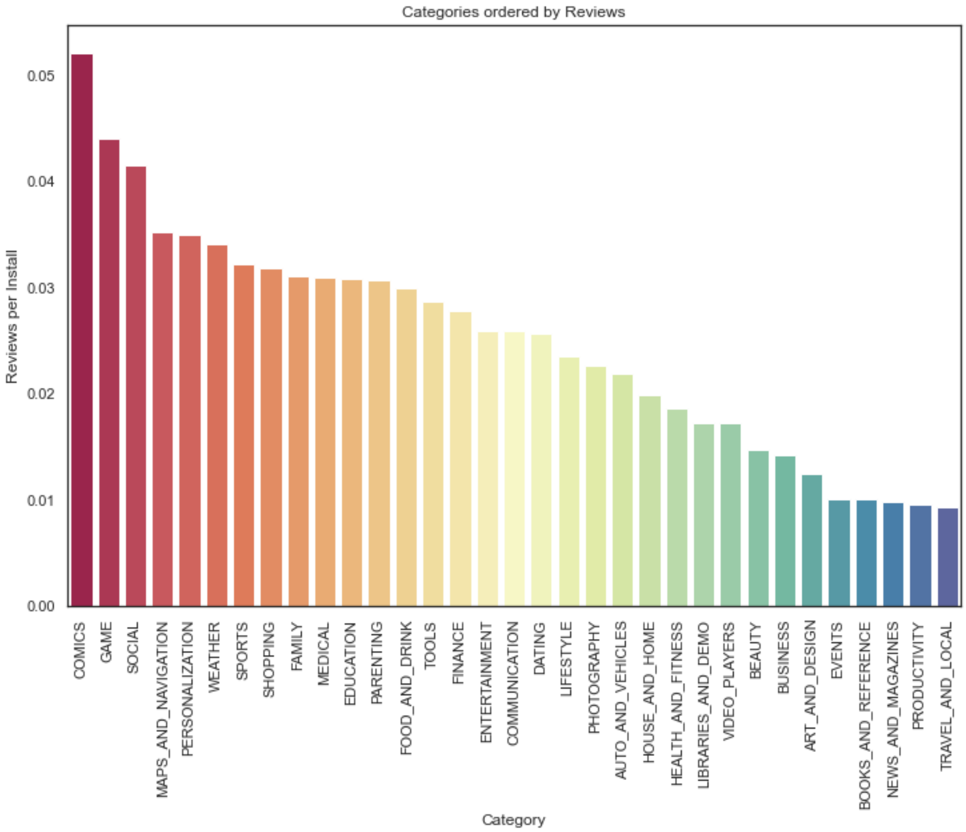


| | App | Category | Rating | Reviews | Installs | Type | Content_Rating |
|---|---|---|---|---|---|---|---|
| **0** | Facebook | SOCIAL | 4.1 | 78158306 | 1000000000 | Free | Teen |
| **1** | WhatsApp Messenger | COMMUNICATION | 4.4 | 69119316 | 1000000000 | Free | Everyone |
| **2** | Instagram | SOCIAL | 4.5 | 66577446 | 1000000000 | Free | Teen |
| **3** | Messenger – Text and Video Chat for Free | COMMUNICATION | 4.0 | 56646578 | 1000000000 | Free | Everyone |
| **4** | Clash of Clans | GAME | 4.6 | 44893888 | 100000000 | Free | Everyone 10+ |

*Observations:*

- The top 4 apps with **most number of reviews and highest rating** all belong to Meta - they are all **social media apps or communication services**.
- **Clash of Clans** looks to be one of the **most popular games** on Playstore.

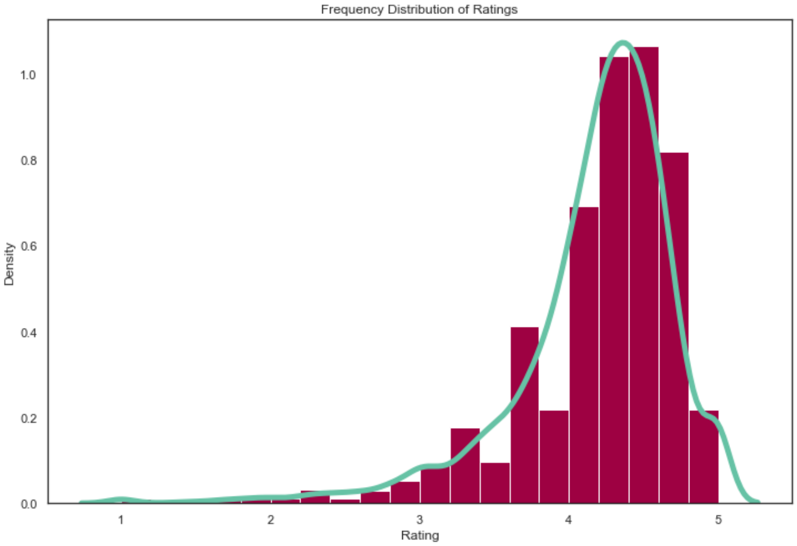
## 3. App Categories ordered by Number of Reviews



*Observations:*

- We calculated a new field **'Reviews per install'** to use for this chart so as to not skew the results (some records have high number of installs and hence more reviews).
- **Comics** is the **most reviewed category** followed by **games and social media**.
- The **least reviewed categories** are **books, news, productivity and travel**.

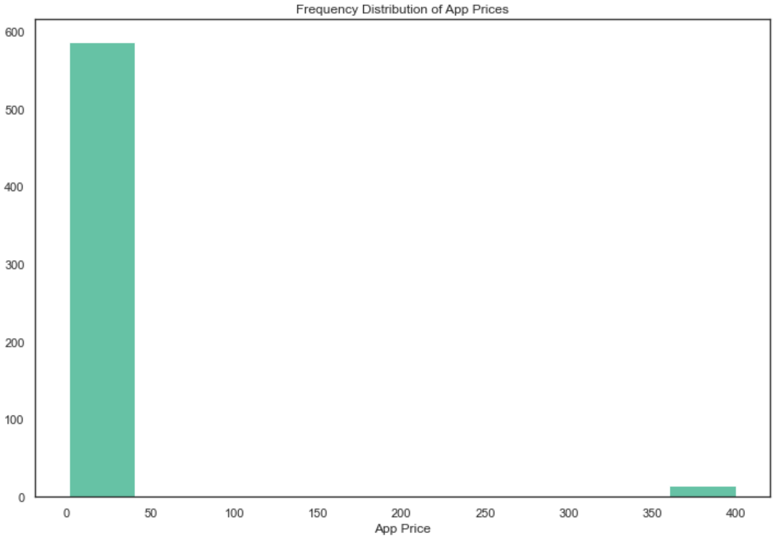
## 4. a) Frequency Distribution of Ratings on Playstore



*Observations:*

- The **most common rating** on playstore is between **4.5 and 4.6** as per this dataset.
- The distribution tells us that most ratings **range between 4 and 5**.
- **Ratings below 3 are low** in number which implies that people are generally content with apps.
- **Ratings around 1.5** are practically non-existent.

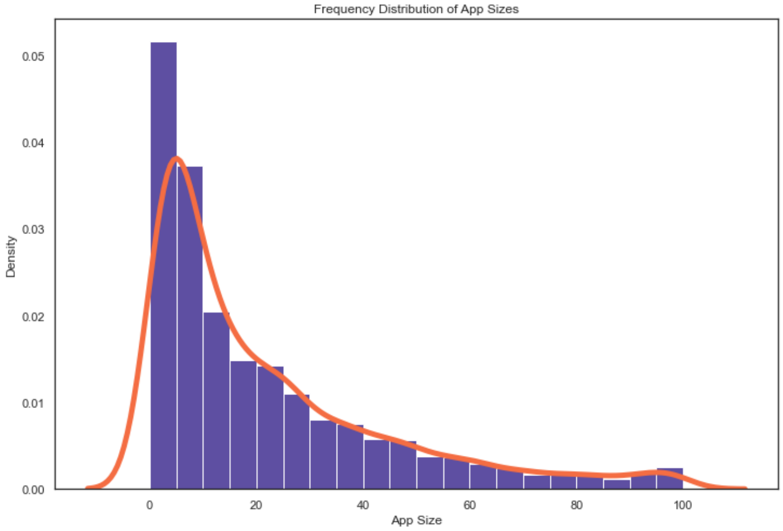
## b) Frequency Distribution of App Prices



*Observations:*

- Of the paid apps, most apps have their prices in the range **1-45 USD**.
- About **14 apps** have a price of **>350 USD and <400 USD**.

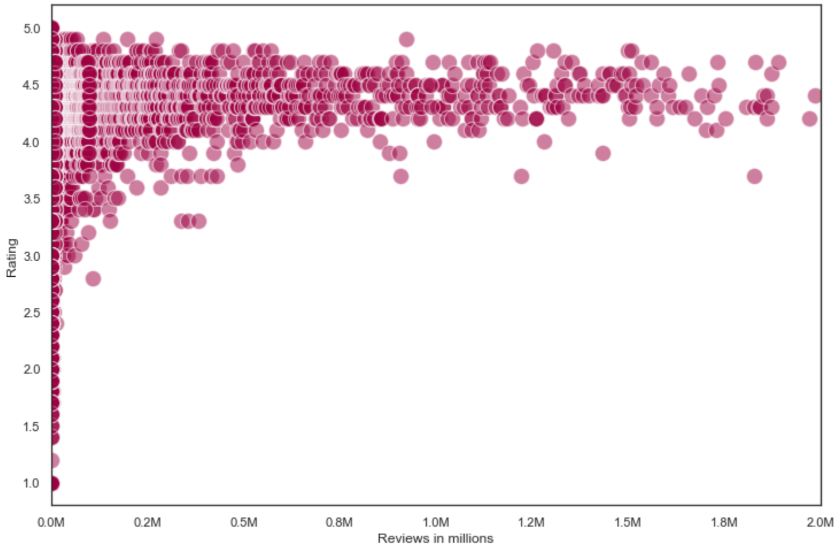
## c) Frequency Distribution of App Sizes



*Observations:*

- Of the 8000+ apps in the dataset, **~4263 apps** have a size of **<20 MB**.
- About **14 apps** have a size of **100 MB** which is the maximum in this dataset.

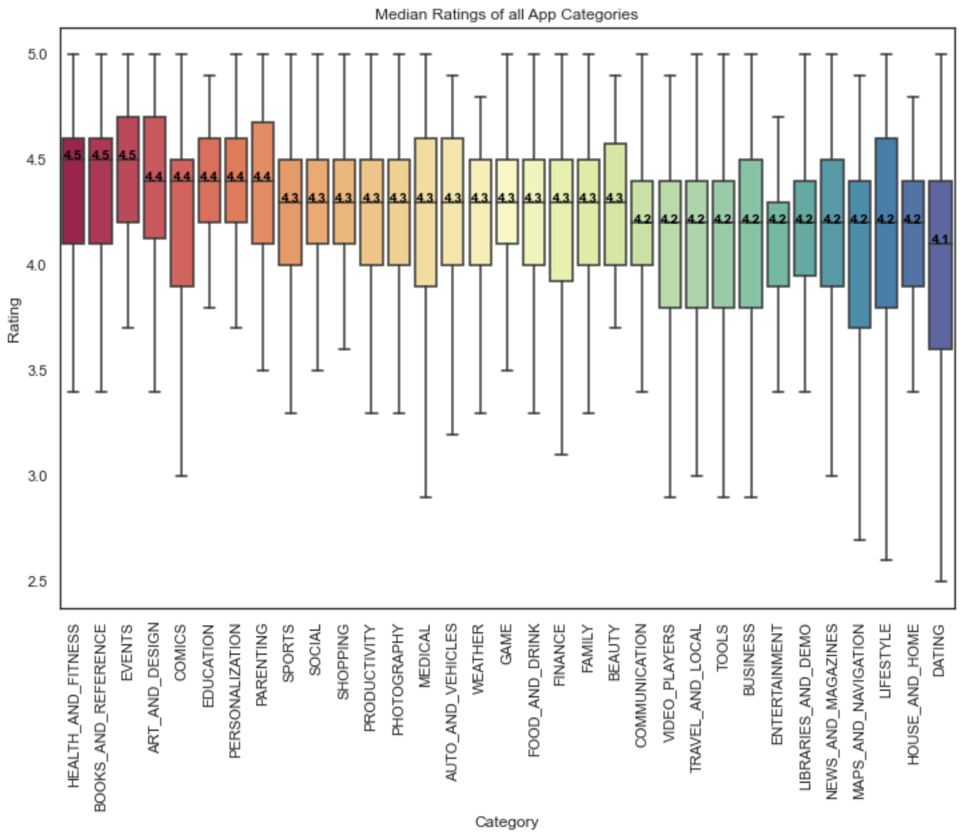
## 5) Ratings vs. Number of Reviews



*Observations:*

- Most app ratings are in the range **3.5 and 5** as can be seen from the above graph.
- The apps which have **most number of reviews** are the apps rated in the approximate range of **4.25-4.75**.
- For most apps, the number of reviews are **less than 0.5 million**.
- There are a few apps which have more than 4 million reviews and even close to 80 million reviews which we have treated as outliers in this chart.
- Not many apps have low ratings and less reviews.

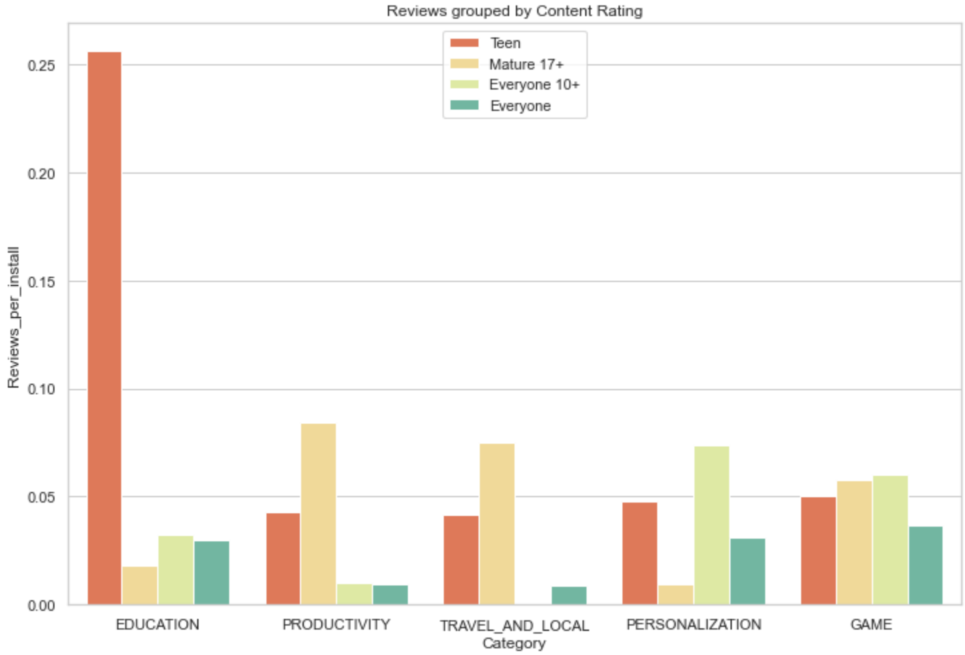
# 6) Median Ratings of all App Categories with Lower and Upper Quartiles



*Observations:*

- The categories with the **highest median rating of 4.5** are **Health & Fitness, Books and Events**.
- The **lowest median rating** of a category in the dataset is **4.1 for Dating**.
- Barring outliers which have been removed from this plot, the **lowest rated app** also belongs to the **Dating** category.
- Some categories like **Education, Weather, Beauty and Maps** do not have a **single app rated 5 stars**.
- The **Entertainment** category has the **lowest upper quartile and the lowest maximum rating**.

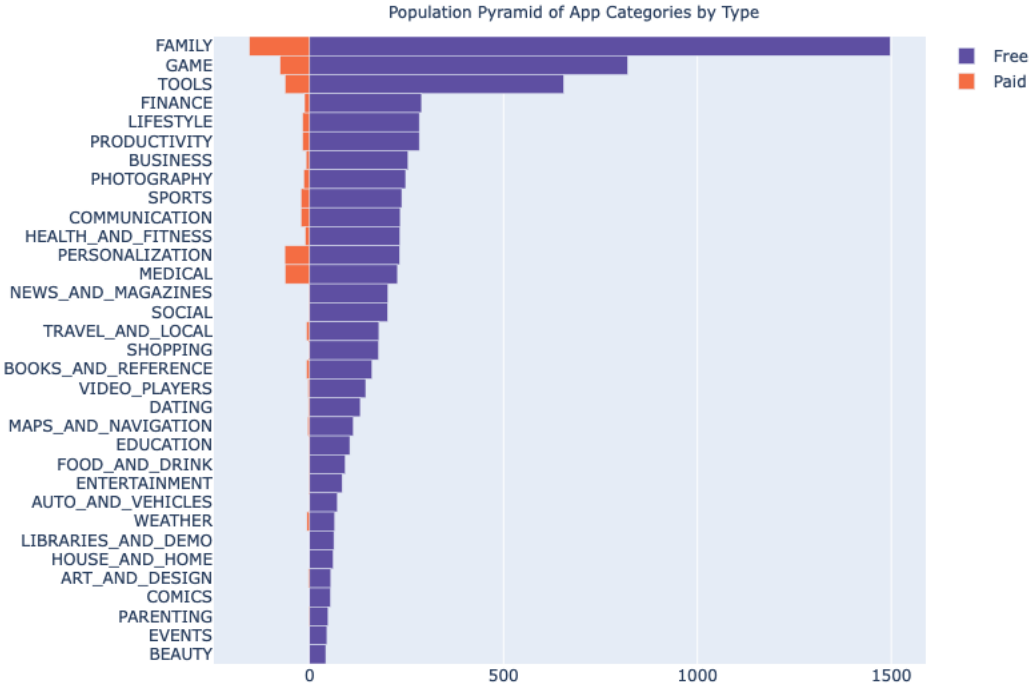
## 7) Top 5 Most Reviewed App Categories based on Content Rating



*Observations:*

- Taking reviews per install into consideration for this chart as well, **Education** seems to be **most reviewed by Teens** which seems accurate.
- **Games** category is almost equally reviewed by **all age groups**.
- **Productivity and Travel** are reviewed mostly by the **Mature 17+ category** followed by **Teens**.

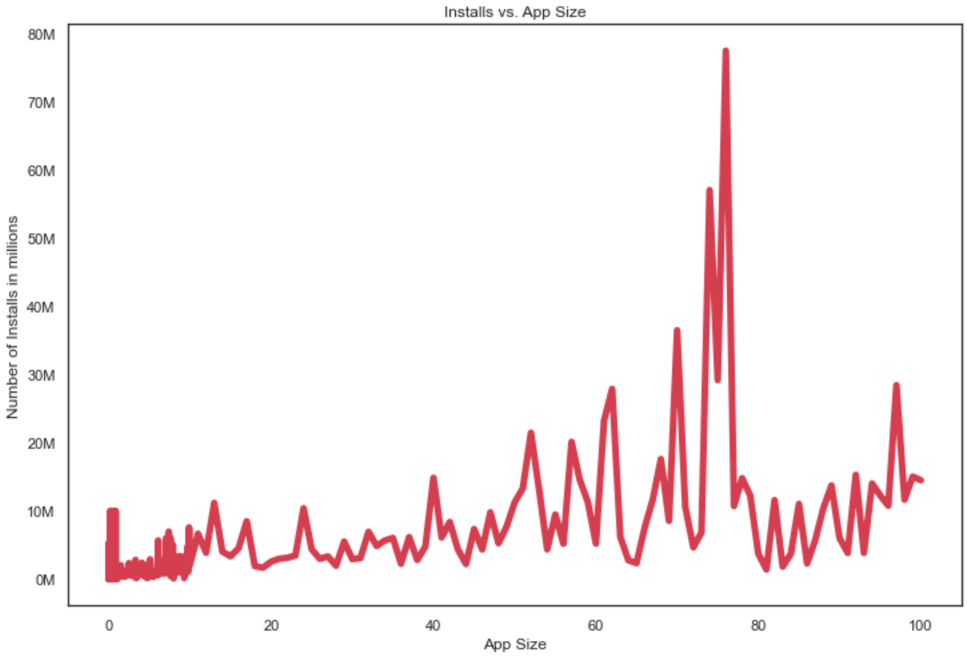
# 8) Population pyramid of App Categories by Type



*Observations:*

- The population pyramid of app count clearly tells us that the **majority of apps** on the Playstore are **free**.
- Looking at the ratio of Free to Paid apps in all categories on Playstore, most number of apps as well as the **most number of paid apps** belong to the **Family** category.
- There are aboout **76 paid apps** in the **Game** category.
- **Tools, Personalization and Medical** also have about 60 paid apps.
- **Comics, Libraries, Home, Vehicles, Events and Beauty** categories only have **free apps**.

## 9) App Sizes vs. Installs



*Observations:*

- Apps of around **70-80 MB** size have the **highest** number of installs **(~80M)**.
- Apps of **close to 100 MB** size have around **30M installs**.
- Apps of size **<10MB** are mostly downloaded lesser **(~<10M installs)**.