**Team 2 - Project 1**

**Motivation:**

Nowadays people can’t live without cell phones, and people can’t live without phone apps. Tech industry has been lucrative and booming, developing an app that fits the market taste would be a smart choice for business.

**Introduction:**

We are working with one csv that lists close to 270,000 android apps. We attempted to use other sources such as iOS and Goggle Play APIs, but those cost a lot of money. This csv was pulled from a Kaggle project, which was originally pulled from a Google Play API from April 2019. We wanted to use this dataset because it involves an everyday device which makes it easy to relate to. This makes our analysis easier, and hopefully our presentation more accessible and engaging.

Our main objective with this dataset is to answer: What app should we make?

Before we can answer this question, we would also need to ask:

* What are the relationships between each variable of this dataset?
* What kinds of apps are the most profitable/popular?
* What factors should we consider when making this app?

Note that we’re mostly answering these questions based on what is available to us from this dataset.

**Data Cleaning:**

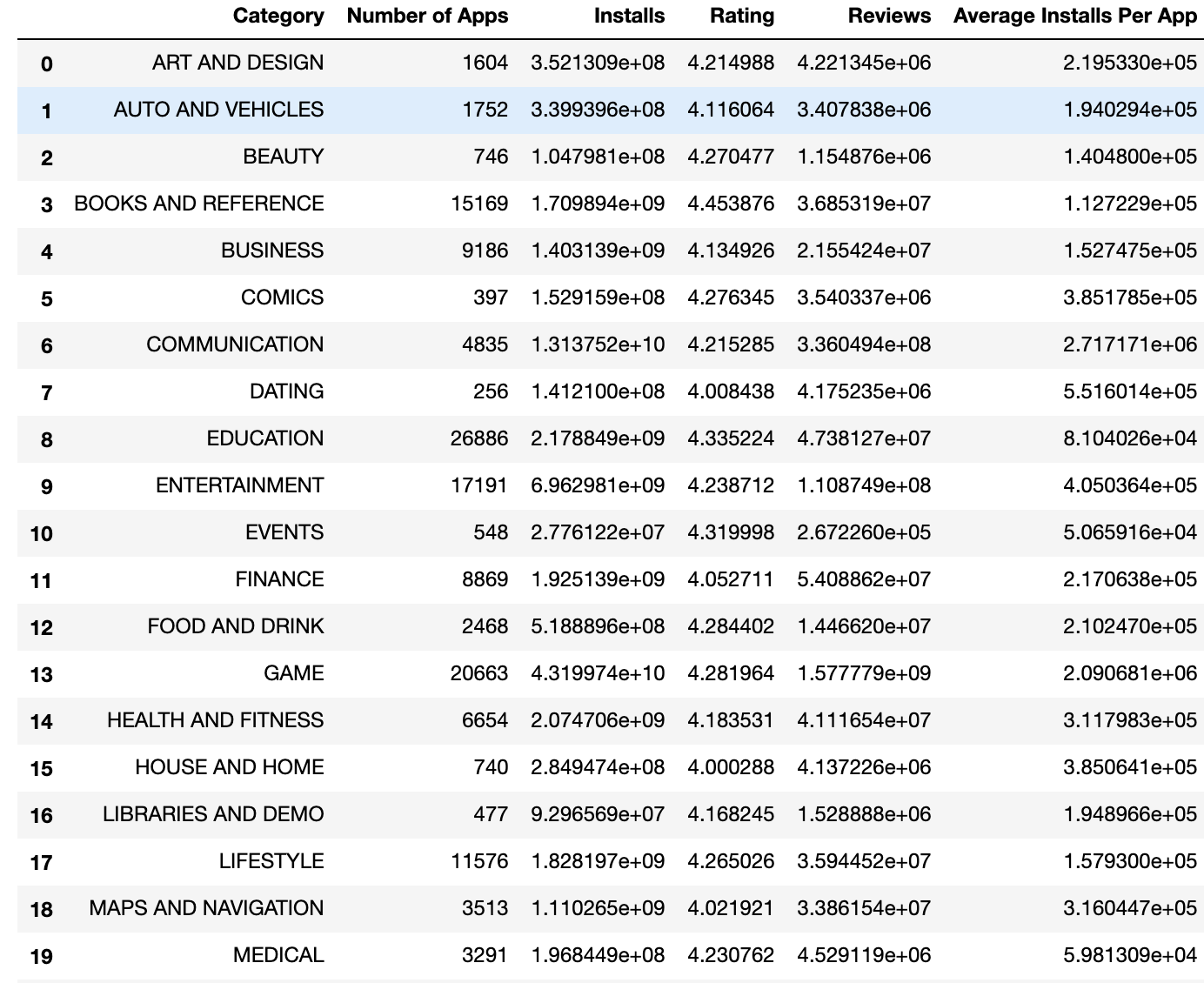
During the data cleaning process, we needed to convert certain variables into numbers in order to analyze them. We had to remove characters that made these variables into Strings. We then converted these strings into floats. 

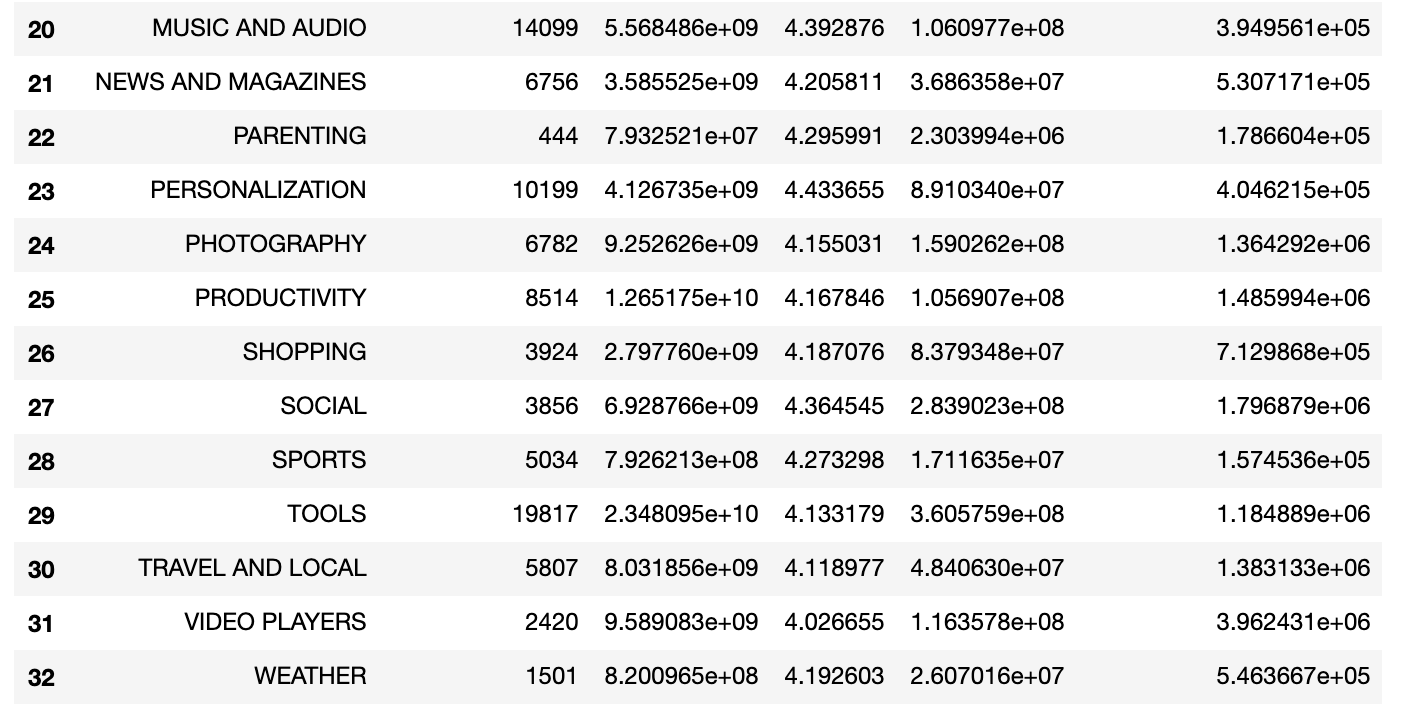




After converting, we grouped by categories in order to determine what the most popular category is.

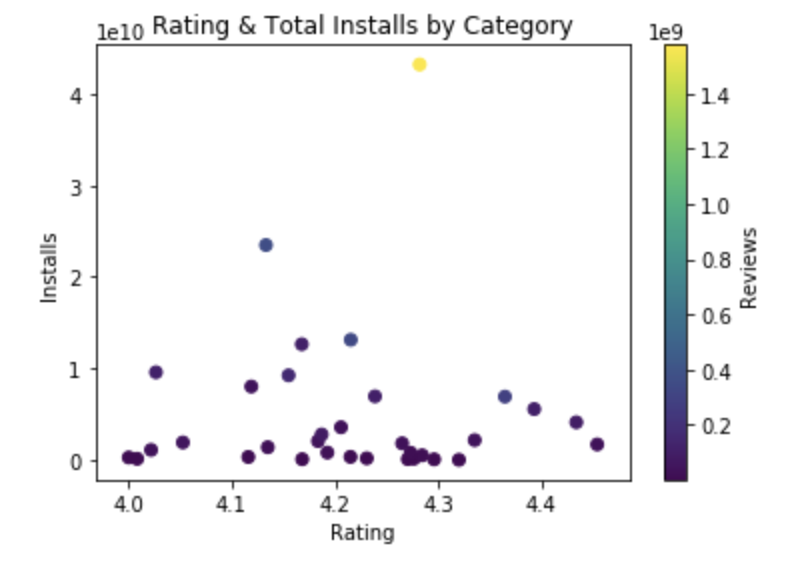
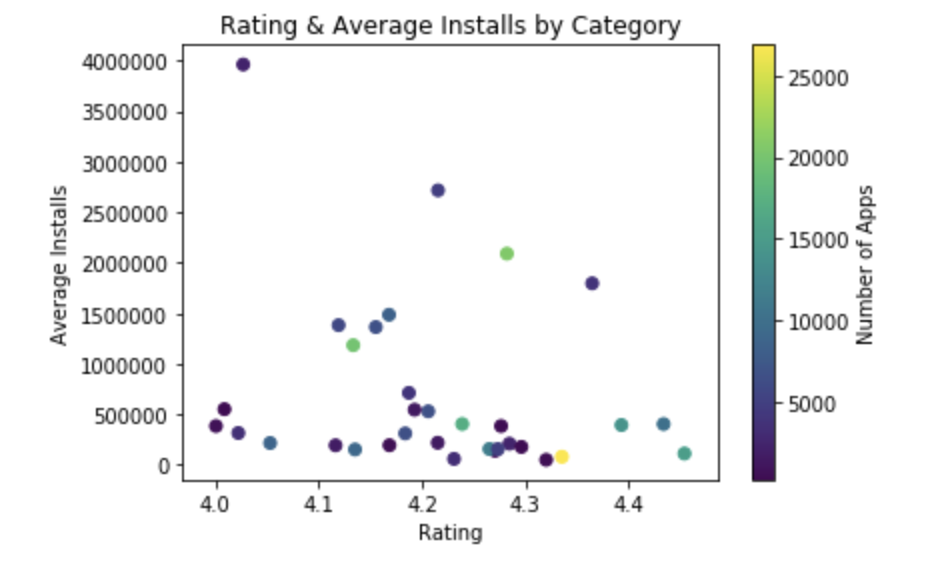
From there, we merged the required columns and made a new data frame of our cleaned data.

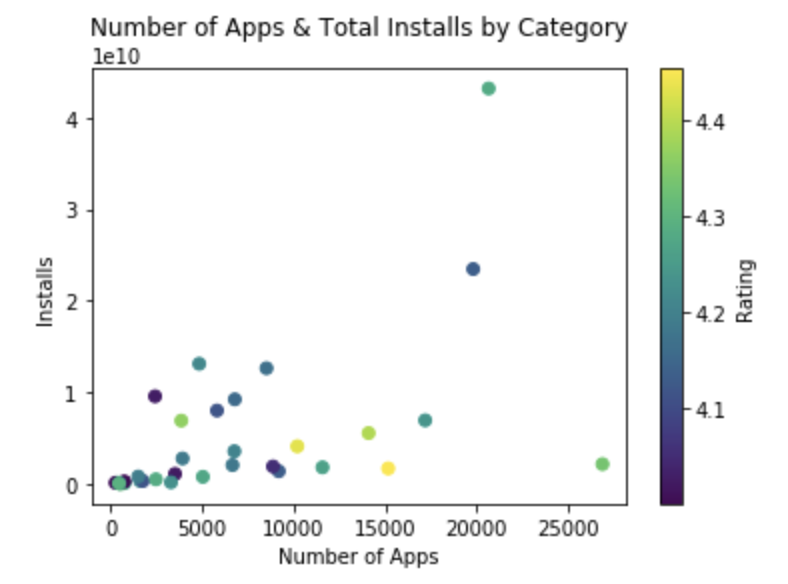
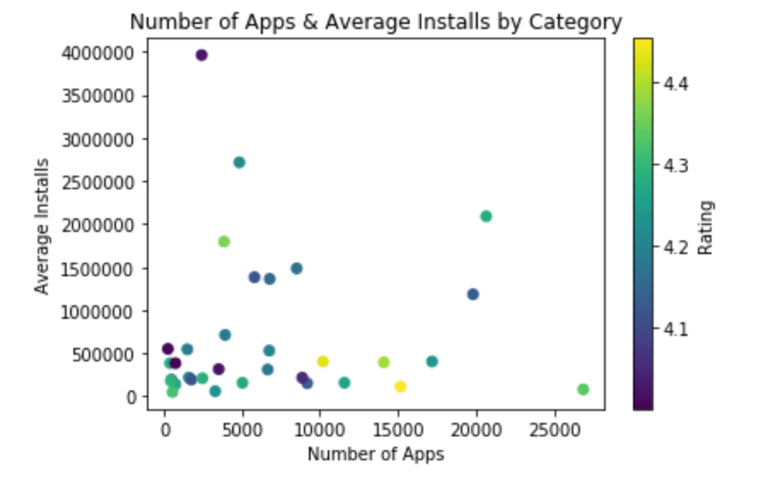


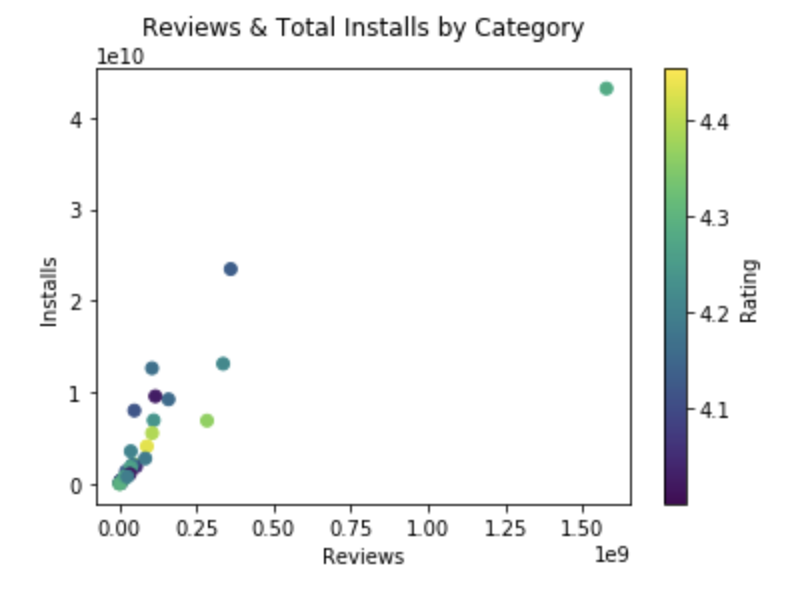


**Scatterplots:**

Exploring the relationships of this dataset with several scatterplots, we’ve found only one correlation between two variables: Total reviews and total installs. It might sound obvious at first; if more people are using your app, then you would have more reviews.



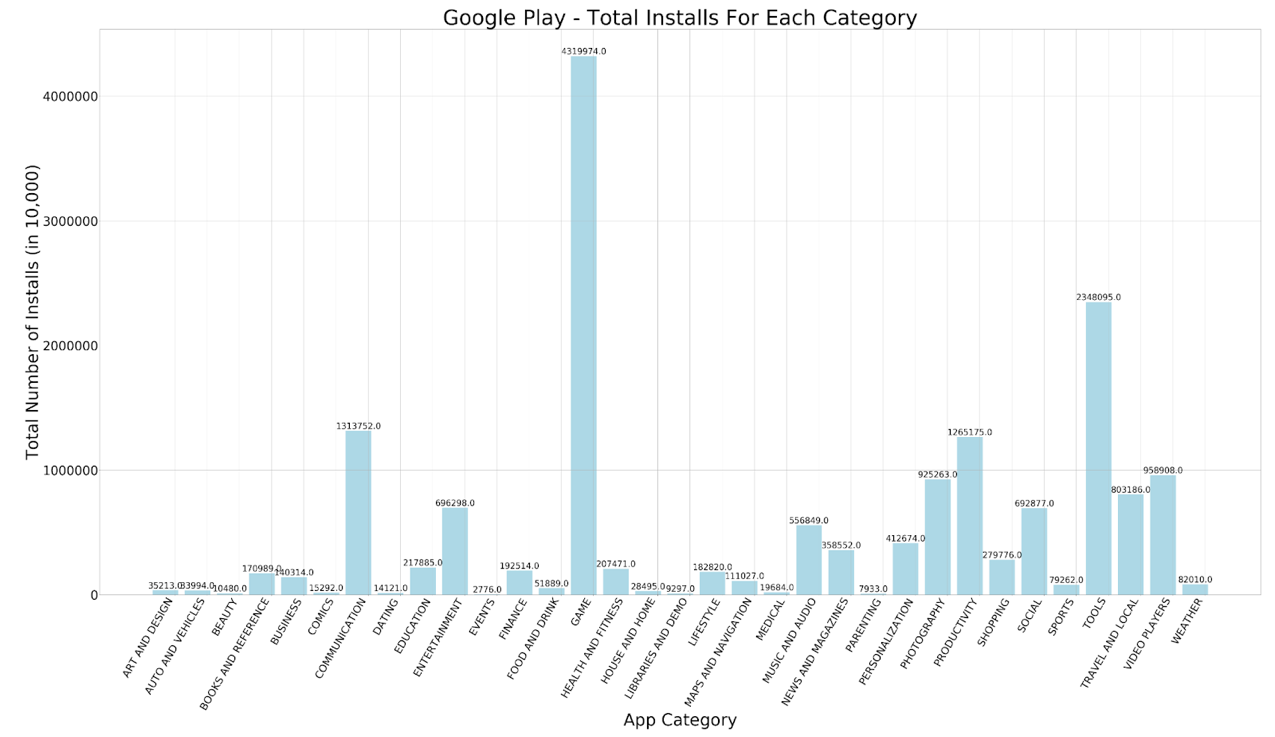


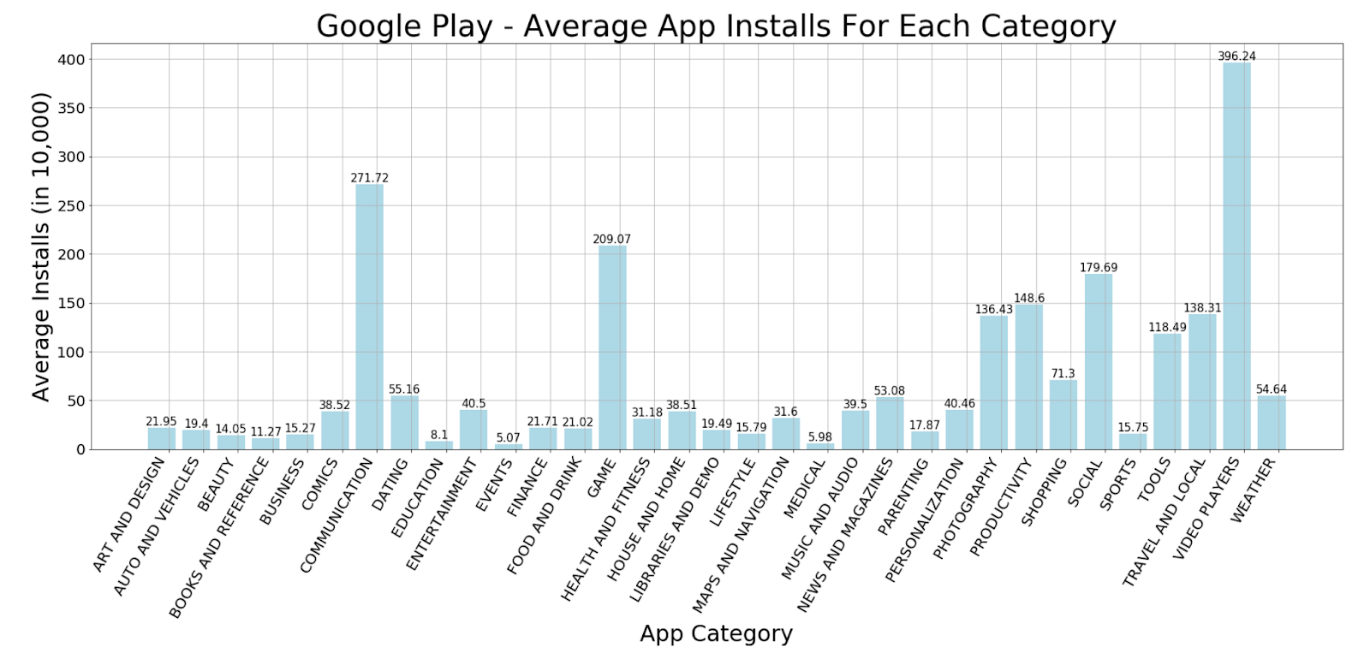


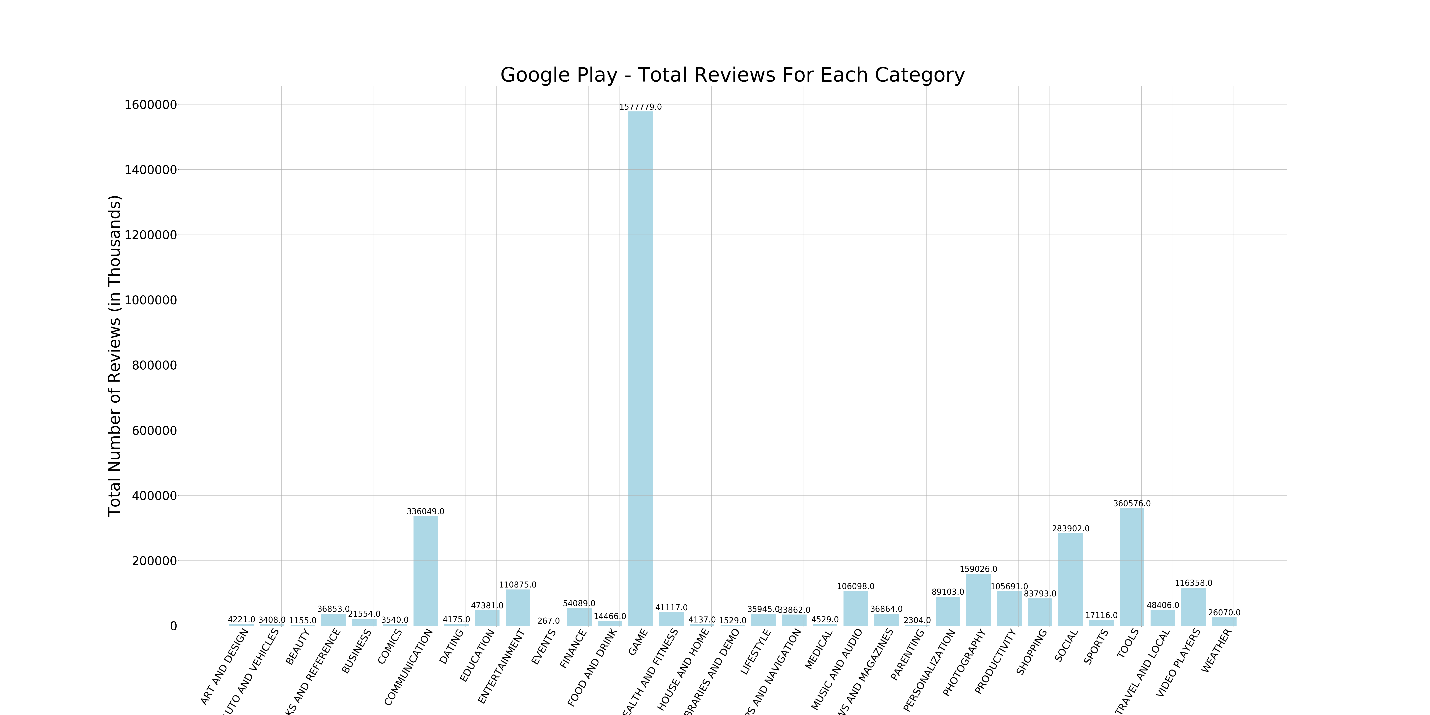
But if you didn’t know, Google Play Store’s search algorithm factors in many variables, such as number of downloads, overall rating, and number of reviews. A better way to look at this correlation is that by having more reviews for your app, it increases your app’s visibility, therefore increasing your number of installs. This relationship should be considered when making an app.

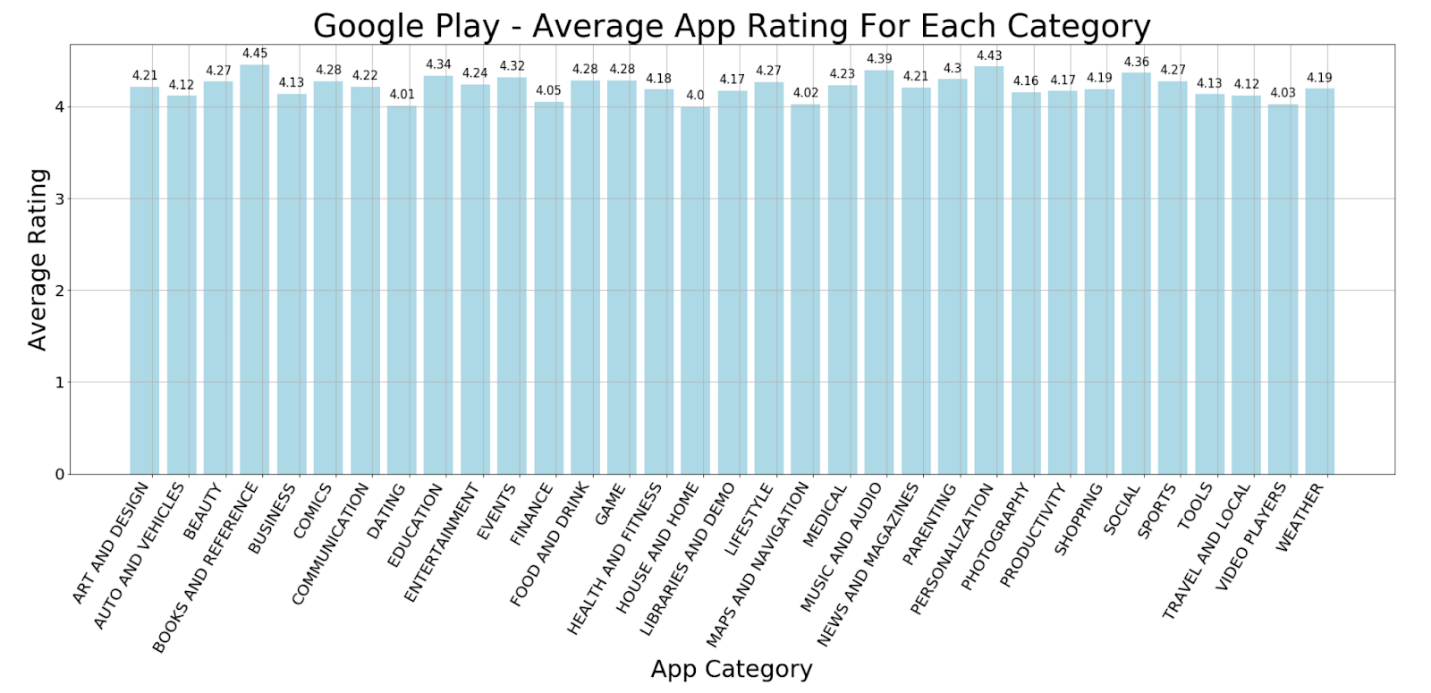
**Bar graphs:**

We have several bar graphs that features the totals and averages of each variable. We’ve found there are several outliers for these, particularly for the Communication, Game and Video Player category. Ratings, however, seem to all float around at the same value with all categories averaging above 4. This suggests that users tend to review an app only if they like it.

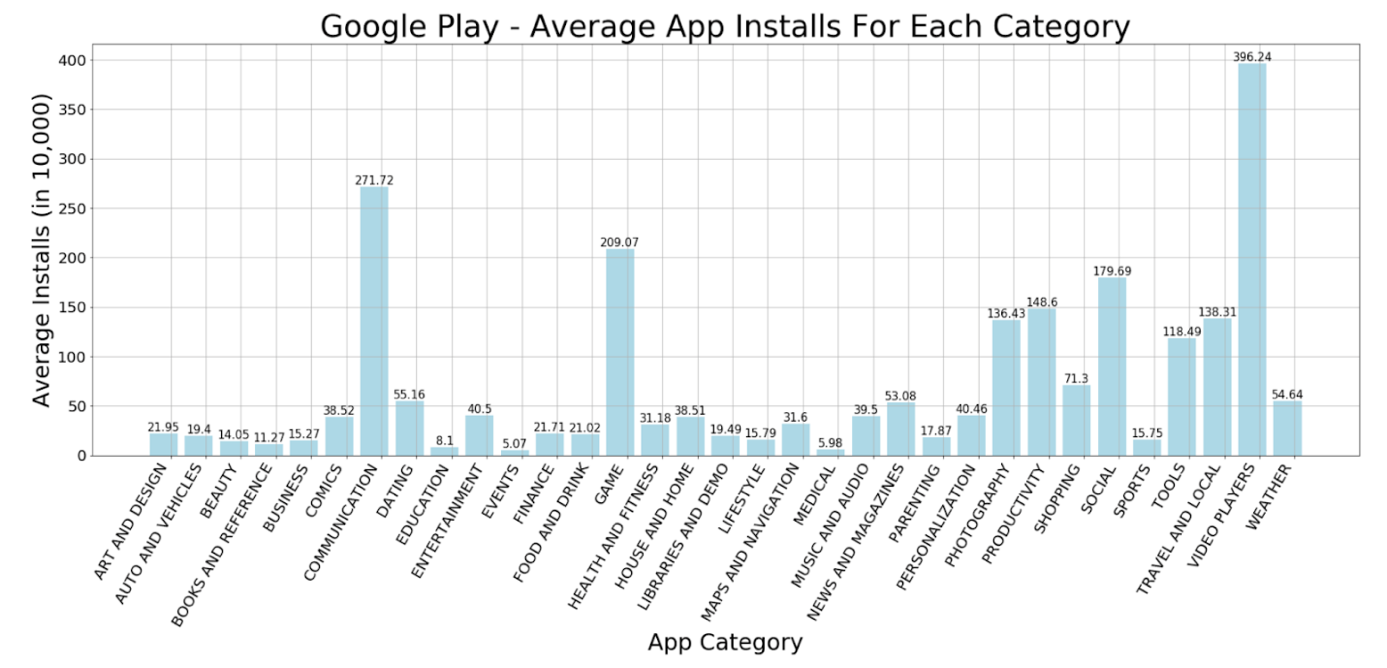




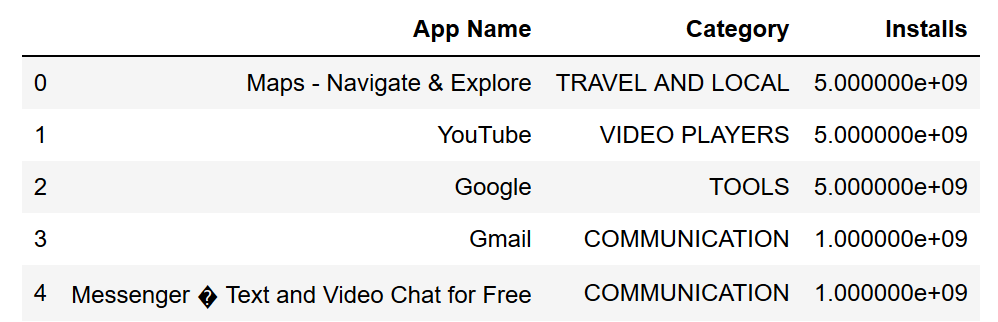




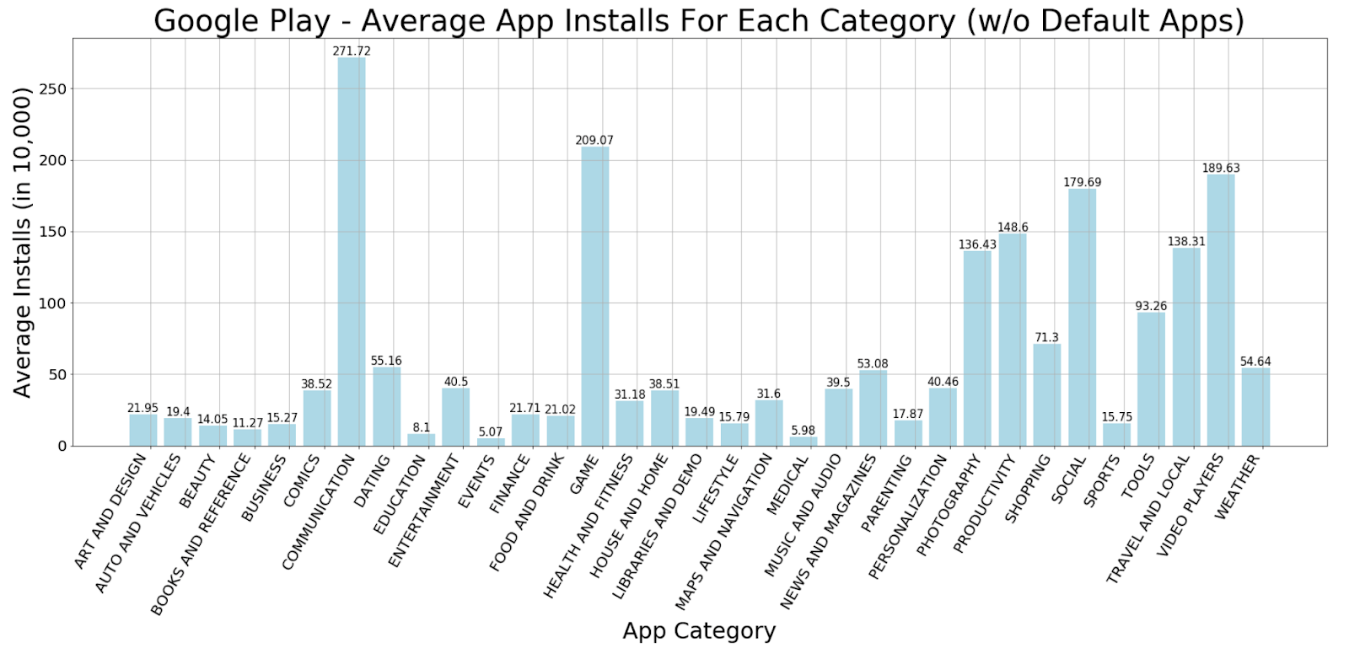
Moving on, our main variable when evaluating popularity is installs. Game apps have the most installs, while Video Players have the most average installs per app. This might be a little deceptive, as the number of installs is greatly skewed for certain categories.



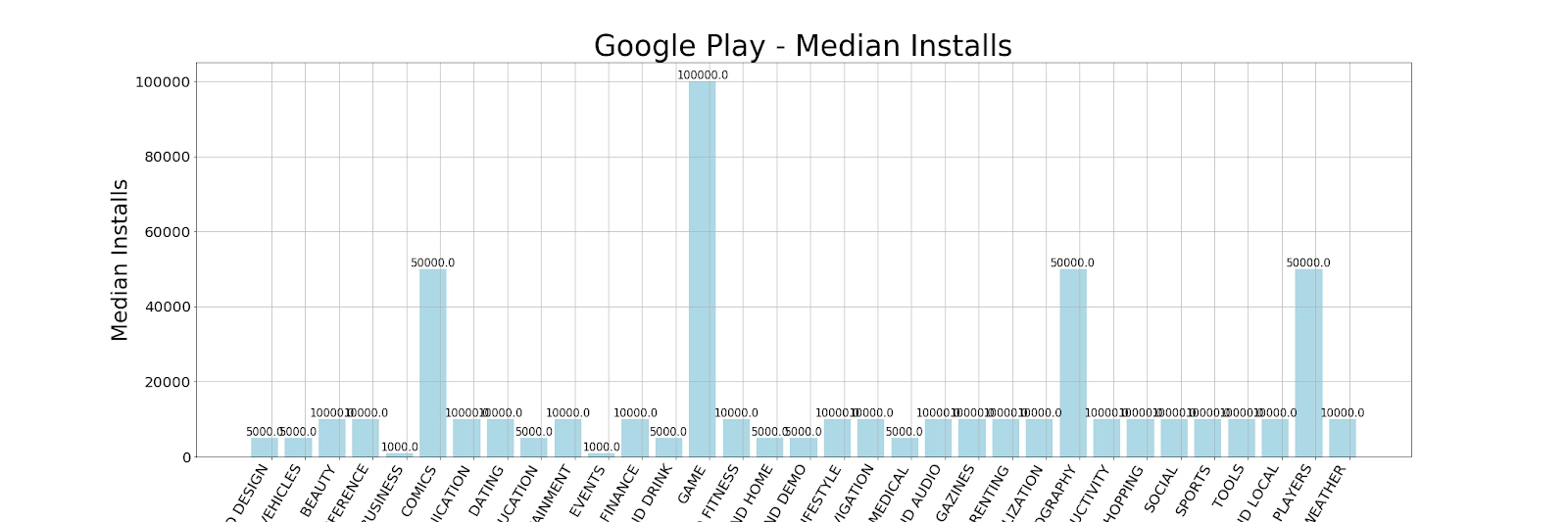
If you look at the top five installed apps, you can see that Google and YouTube are by far the highest, with a huge drop off from second to third. That’s because both apps come pre-installed on all android devices, whether it’s a Samsung, Google phone, LTG, etc.



If we were to remove Google and YouTube in a new data frame, our average installations per app would look a little different, resulting in Communications and Games to be the top categories.

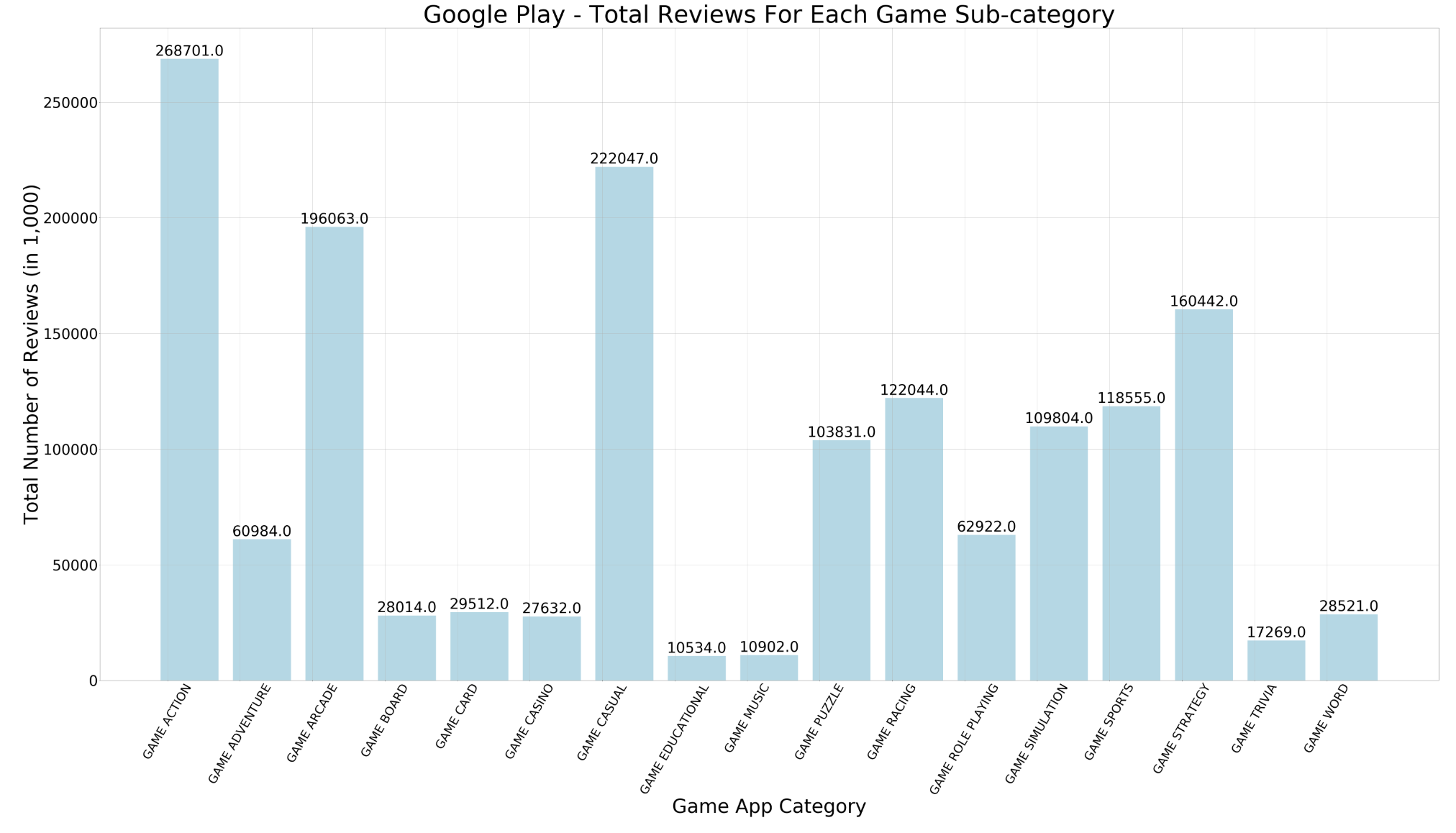


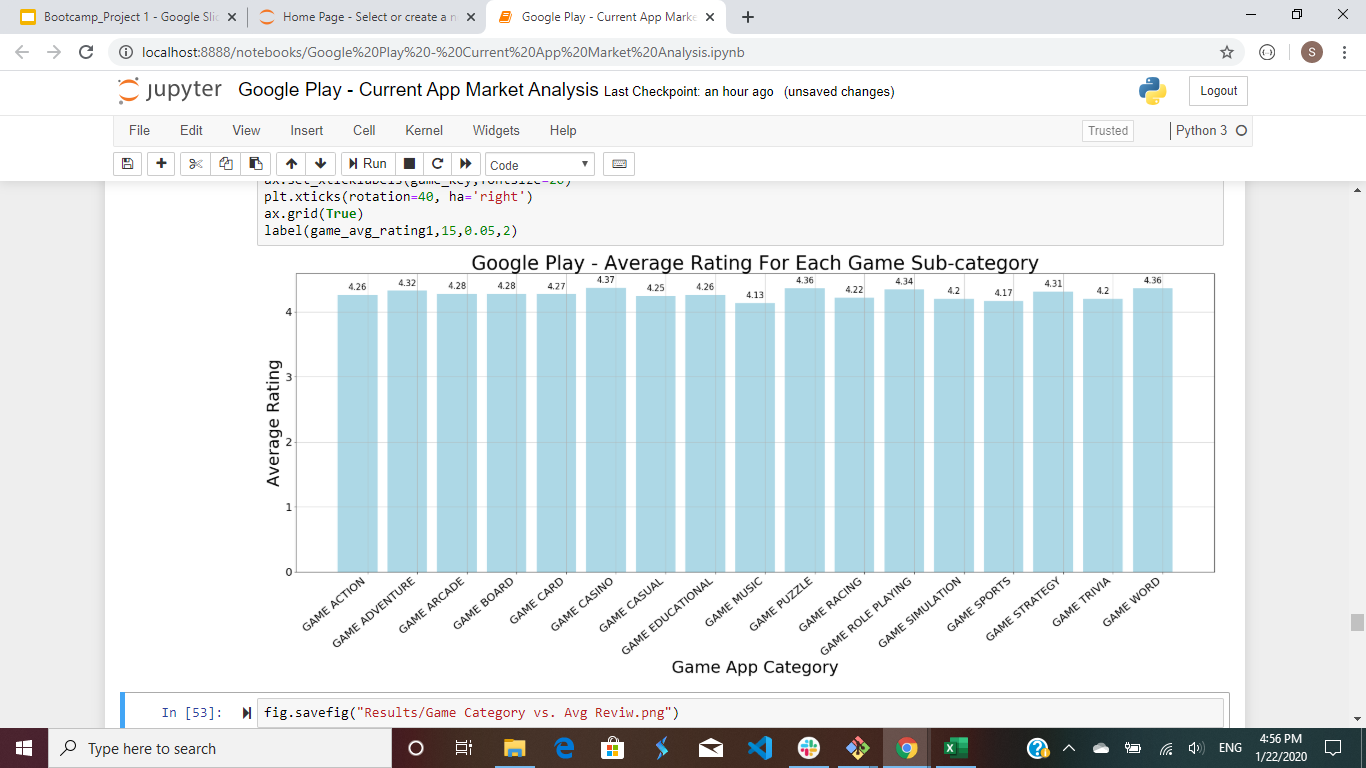
We should also consider the median of installations. Looking at the median, we can determine which markets are top-heavy. While Communications, which are generally messaging and texting apps, has a very high install-average, it has a low median, meaning their high install-average is heavily influenced by their top apps. Meanwhile, the Game category remains high.

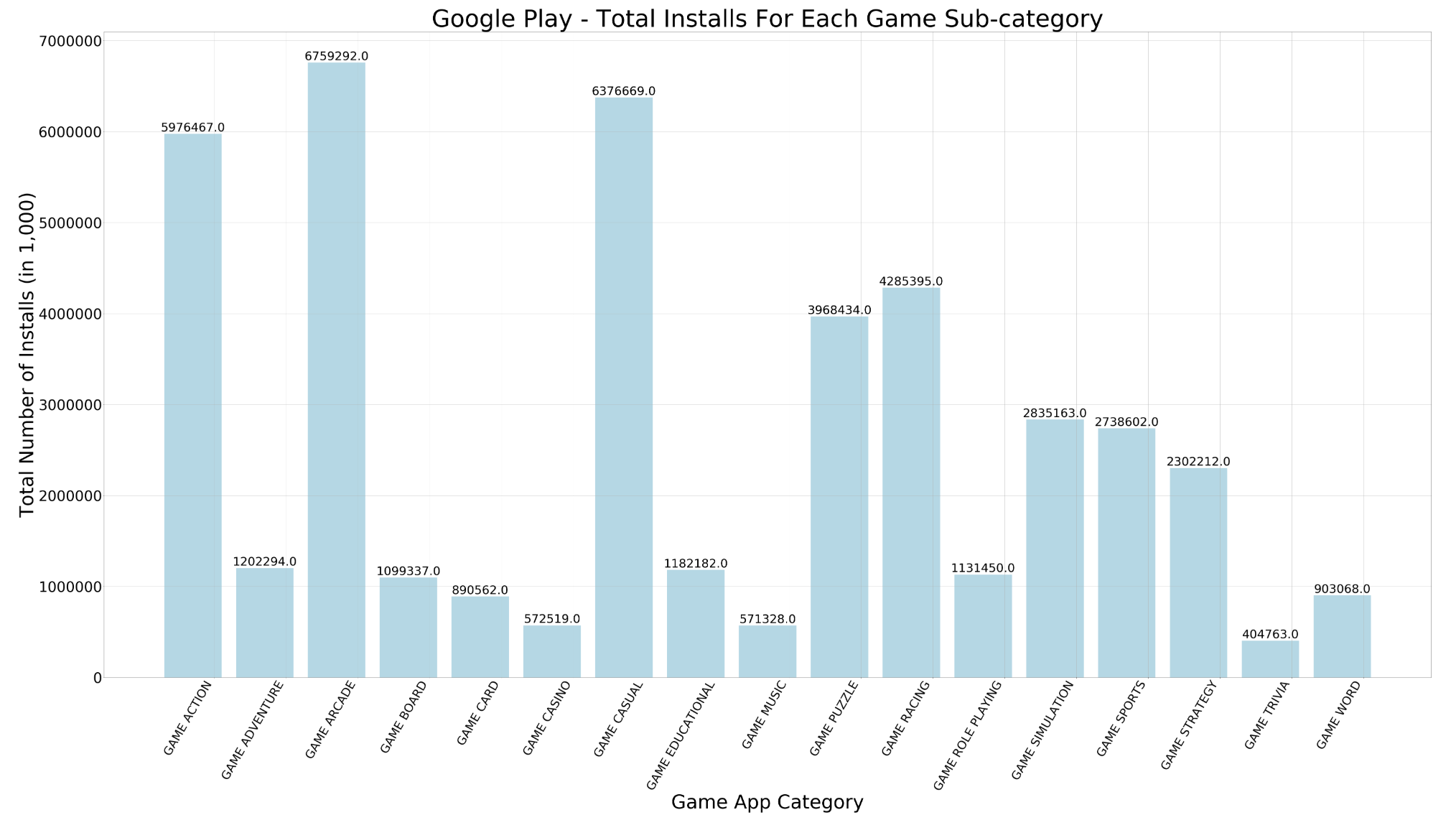


Bringing all these charts together, we’ve learned that the game category has the highest total installs, 2nd highest average installs, highest median installs, and highest total reviews. Therefore, we conclude that we should make a game app.

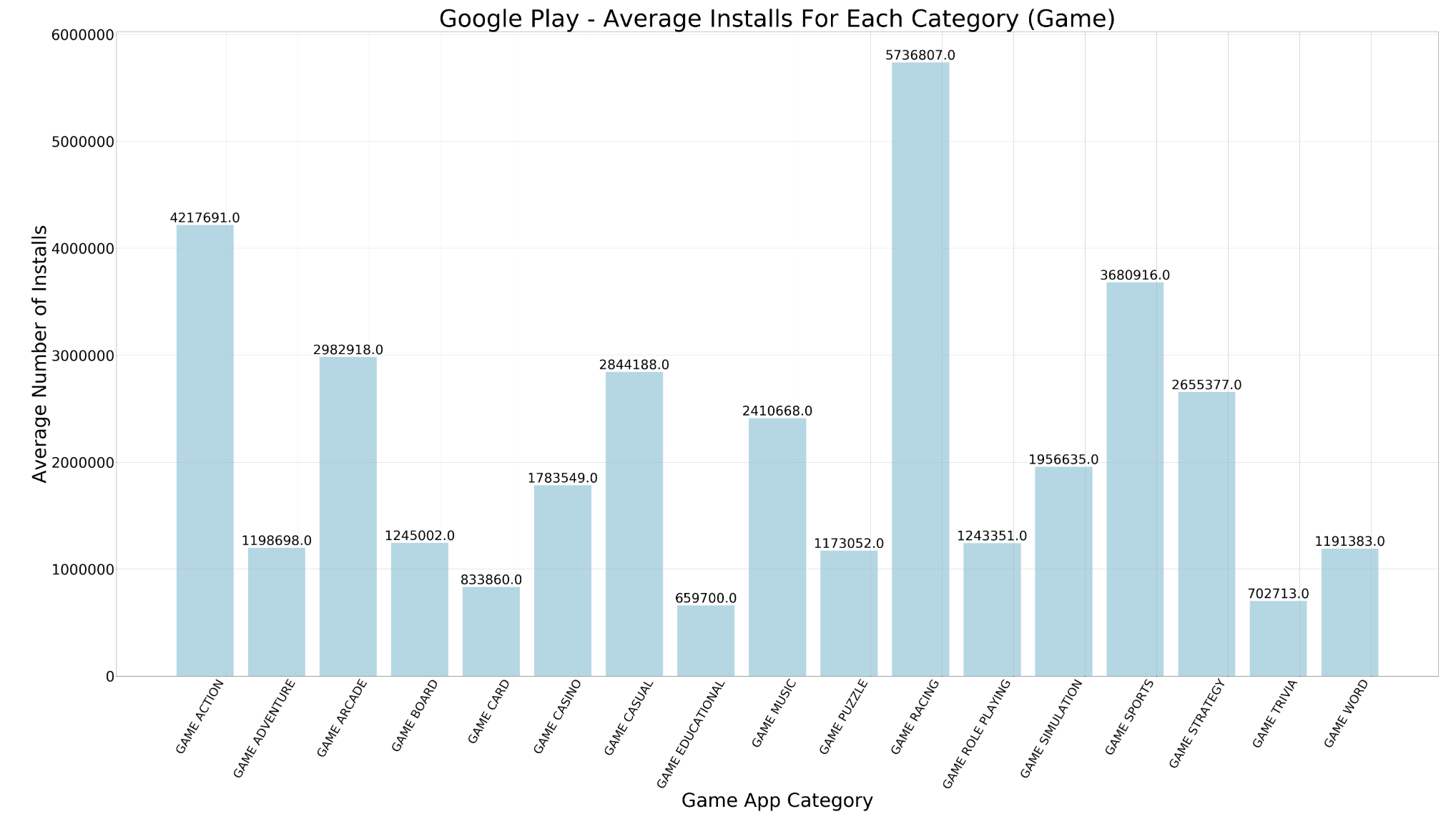
**Gaming Apps:**

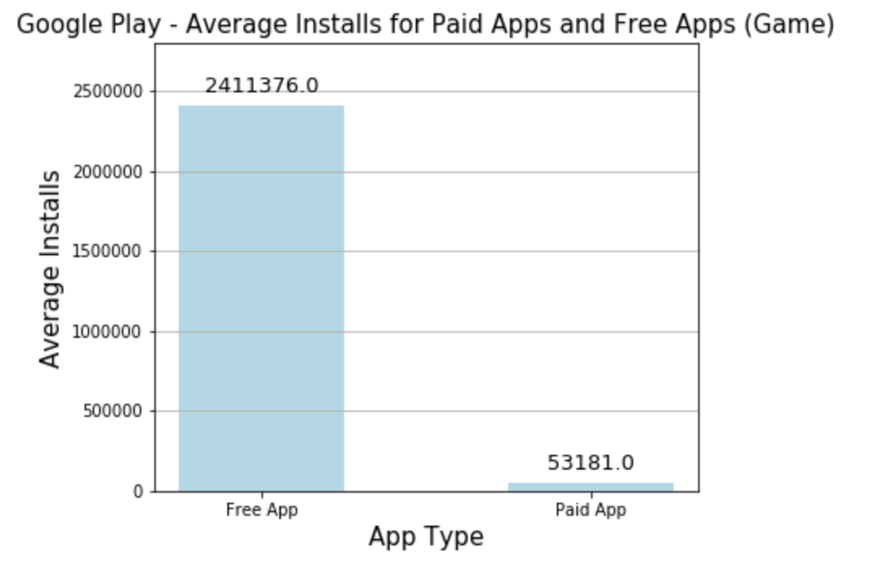


****Action games, casual games, and arcade games are the most popular game subcategories based on the number of total reviews. And if you look at the average rating for different subcategories, the ratings are fairly close and are all above 4, meaning that user will only rate an app if it is actually good.



The total installs, exhibits a similar pattern as the total review graph above.



But if you look at the average installs for each subcategory, the most popular subcategory became racing games, which is what we are trying to find: the subcategory that is popular (because it has a lot of installs), but there are not enough apps in the market/ are in high demand.

In terms of paid and unpaid apps, it is clear that free apps dominate the market. In fact, during our analysis, the only category that has an average cost per app over 1 USD is health apps. The result is not a surprise to the team: it makes sense that nobody is willing to pay for entertainment apps, if there are so many free apps available in the market. People only pay if it is for functionalities.

As a conclusion, and also to answer the questions we brought up at the beginning, the data shows that the “right” app to make should have following characters:

* Free (but can still have in app purchases
* Game app
* Racing game
* Should incentivize users to rate us on app store

However, the data/ analysis we had still have a lot of limitations:

* Information is not free. We were not able to get the Apple store app information and therefore cannot compare the two biggest phone app market data.
* Communication is the second popular apps in the market by installs/ average installs. If we have more time, we could have analyzed communication more with similar codes that we used for gaming apps.
* If we can have access to data for the past few years, we can run more analysis on based on time: Which type of apps are trending up/down? Which category has growing apps (more competition) and which category has a high demand (high installs but not a growth in number of apps over time)?
* Analyze the app data based on location: geographical information determines the users’ preference of the apps as well. Anyone outside the U.S. will probably not install a “Shake LA” earthquake app. Find the right target group f is important.
* Analyze the app data based on user info to set the target group: gender/ age group. However, a lot of the user information are tied to each individual app. To gather user info for the entire app market would be very challenging.