Google Play Store Data Analysis

Shankar Mohanathas

Introduction

This project focuses on data cleaning and analysis of apps from the Google Play Store, to understand their relationships.

Data Source

The source of raw data was found in (https://www.kaggle.com/datasets/lava18/google-play-store-apps), which has crucial information in categories such as, category of data, price, reviews, size, version, last updated, and content rating. The original CSV file contained over 10 000 data entries.

Data Cleaning

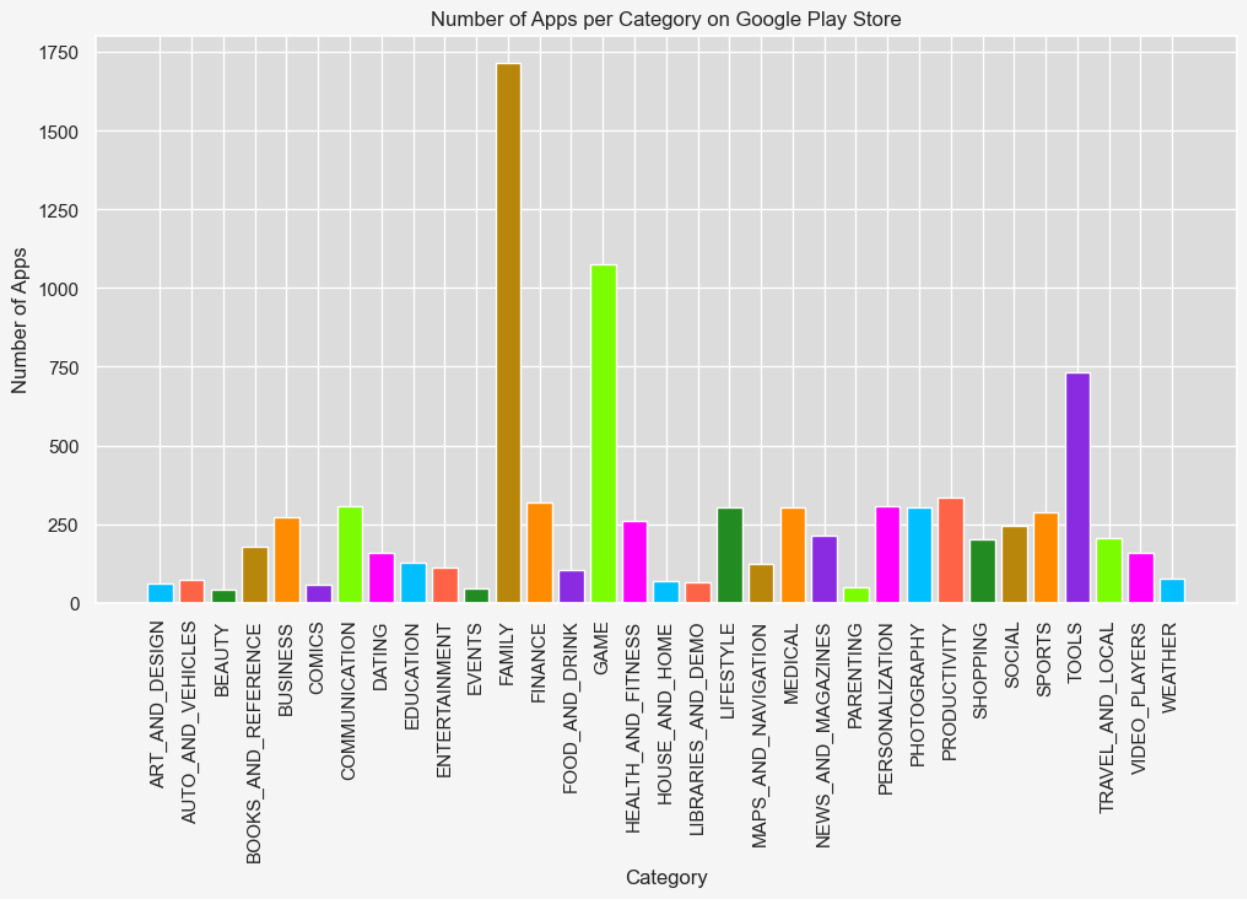
The methodology for tackling a data set of this size was to first import all the information and grasping what information that we had access to. The dataset was filtered through, and missing values were identified and rows containing them were removed. Next step was to look for duplicate values in the data and removing them. Furthermore, non-numeric characters were removed from various columns, and the were converted to appropriate data types. Post these changes, a new cleaned dataset was created and exported to be used for further analysis, through our questions.

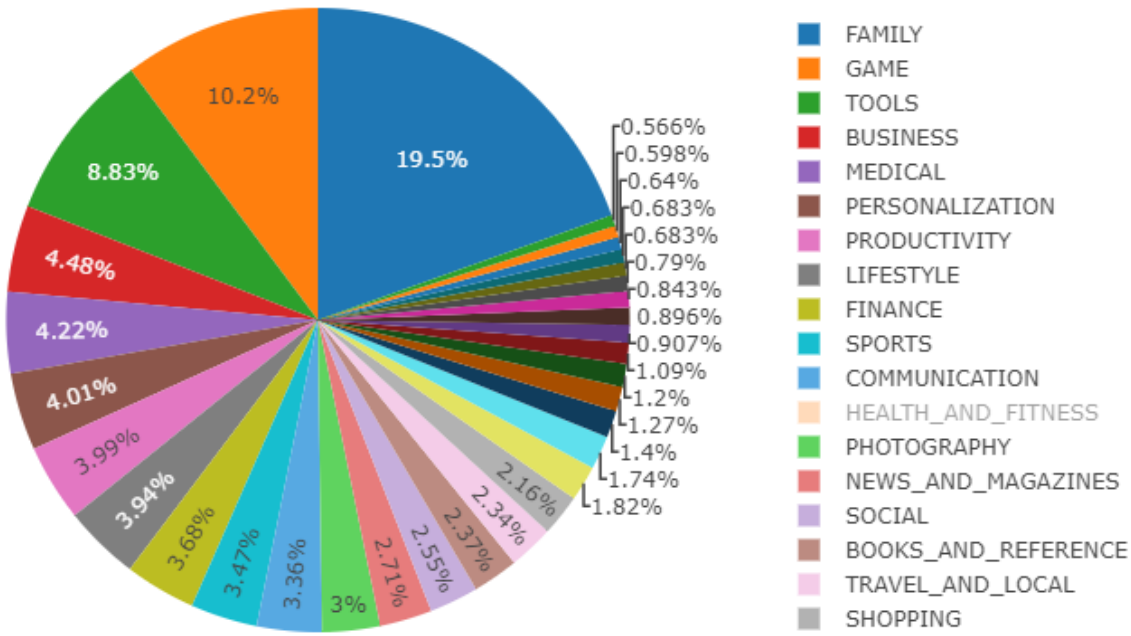
Limitations

One of the more challenging aspects of working with this dataset is depicting on a graph when comparing a small number with respect to a very big number. For instance, the rating system goes only up to 5 whereas the reviews exceeds 60 million in some categories, so when illustrating a graph to show this relationship, it made it a challenge.

Question 1: What is the distribution of app categories in the Google Play Store?

A bar plot was used to visualize the distribution of app categories. We found that the top 3 categories with the most apps in the Google Play Store are Family, Game, and Tools. The least represented categories are Beauty, Comics, and Parenting.

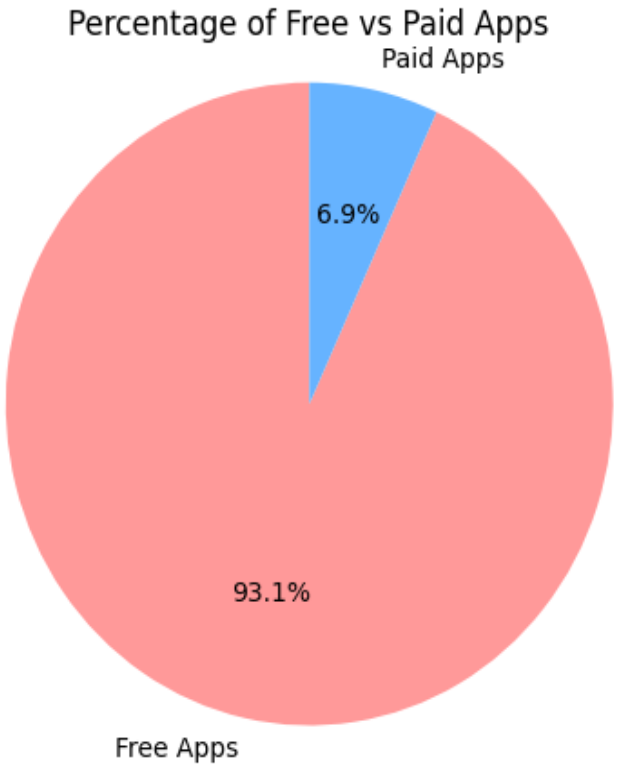


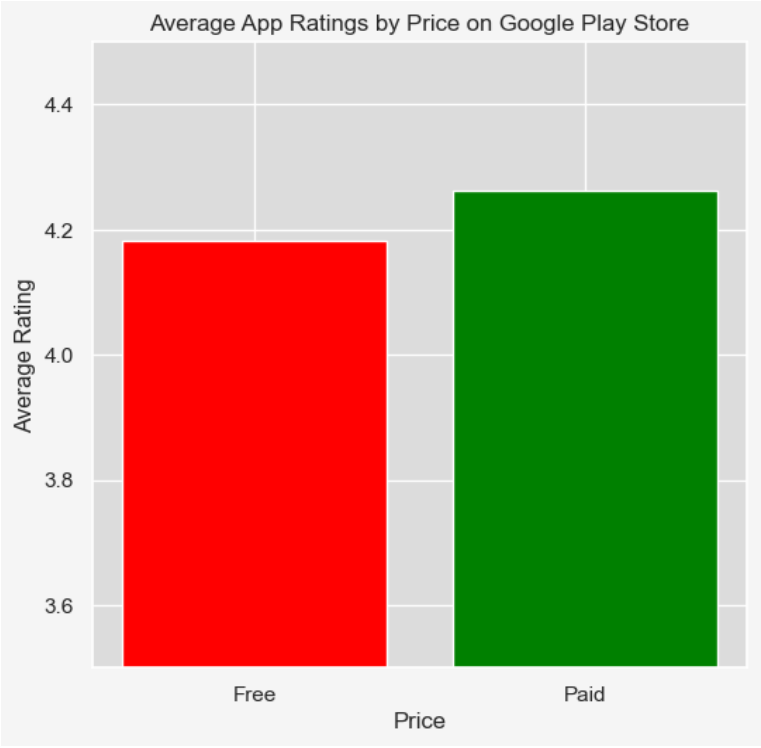


A Pie Chart was created to show the exact same information as above to depict the breakdown of categories.

Question 2: What is the percentage of free vs paid apps in the Google Play Store?

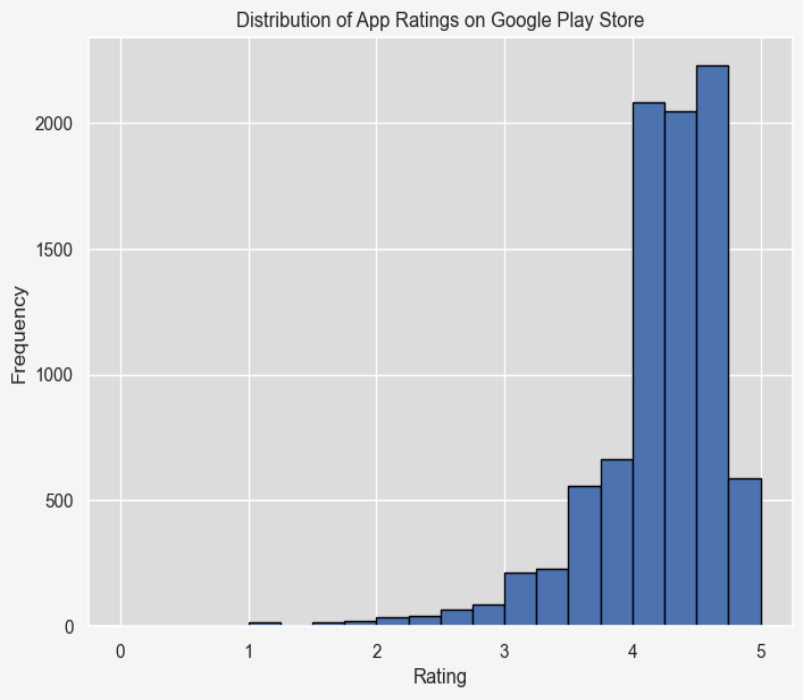
The pie graph on the left shows us that the majority of apps in the Google Play Store are free (93.1%) while only a small fraction are paid (6.9%). A bar chart was used to visualize this distribution to make a comparison between price and average rating. It can be seen that the average rating of free app is slightly less than 4.2 (max rating is 5), where as for paid apps is hovers slightly higher than 4.2.

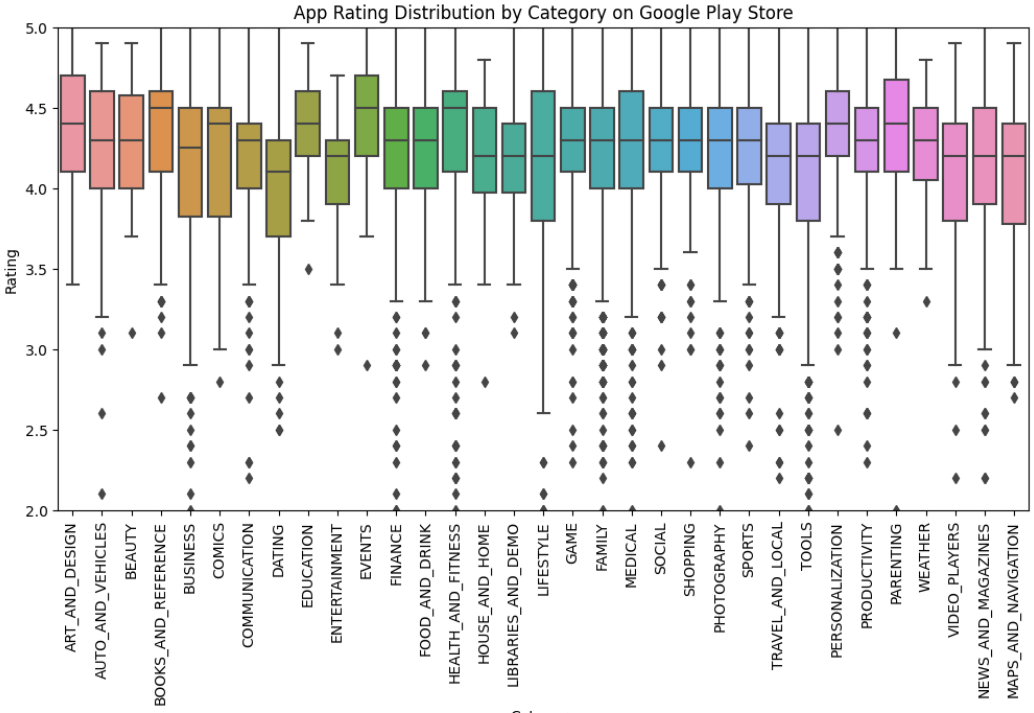




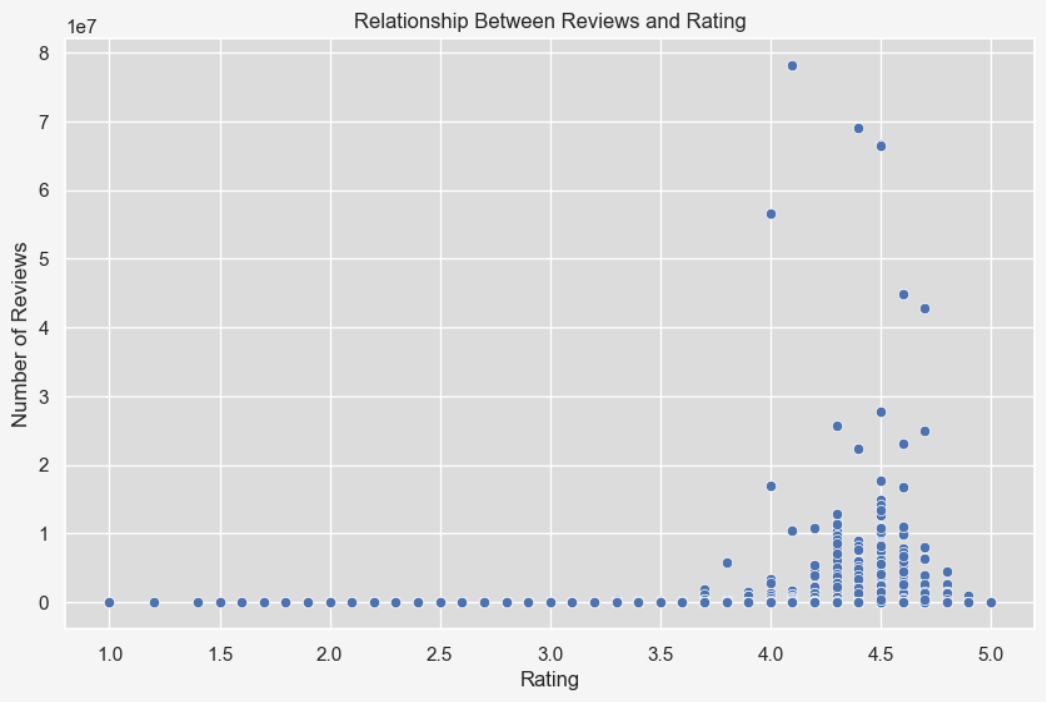
Question 3: What is the distribution of app ratings in the Google Play Store?

We found the majority of apps in the Google Play Store have a rating between 4.0 and 4.5. The distribution of ratings is skewed to the right, indicating that there are more apps with higher ratings. A histogram was used to visualize the distribution of app ratings. To further illustrate this statement a box plot was created comparing each Category against Ratings.



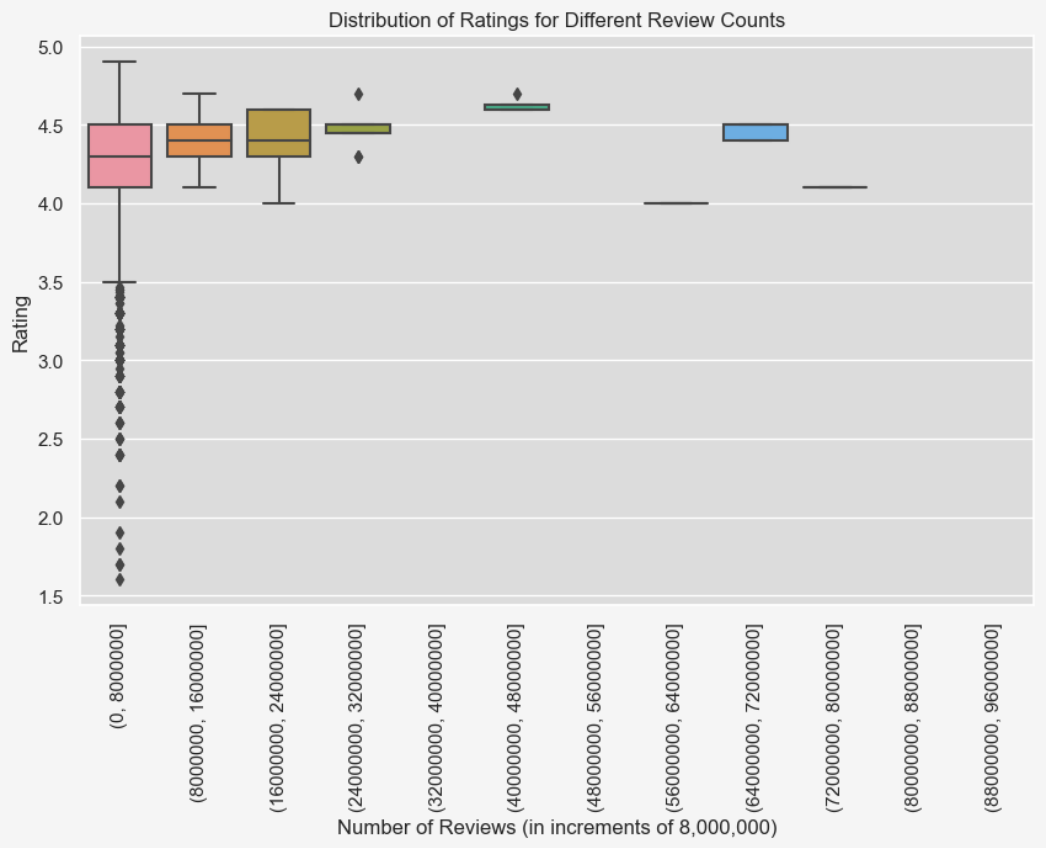


Question 4: How do the number of reviews correlate with app ratings?



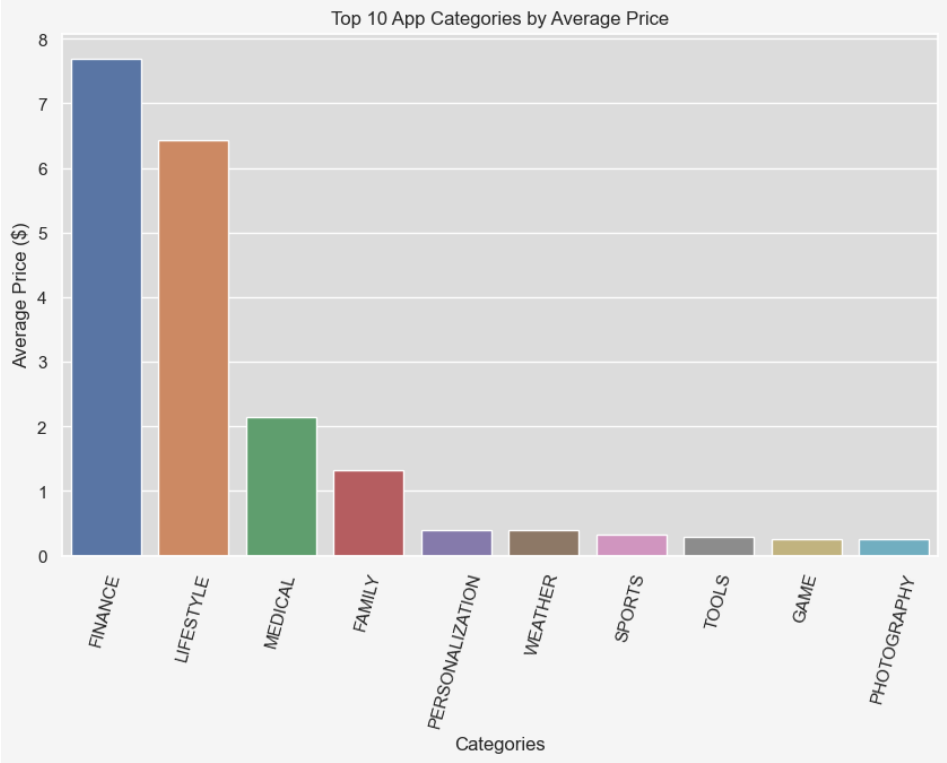
There is a weak positive correlation between the number of reviews and app ratings.

This means that apps with more reviews tend to have slightly higher ratings, but the correlation is not very strong. A scatter plot was used to visualize this correlation.

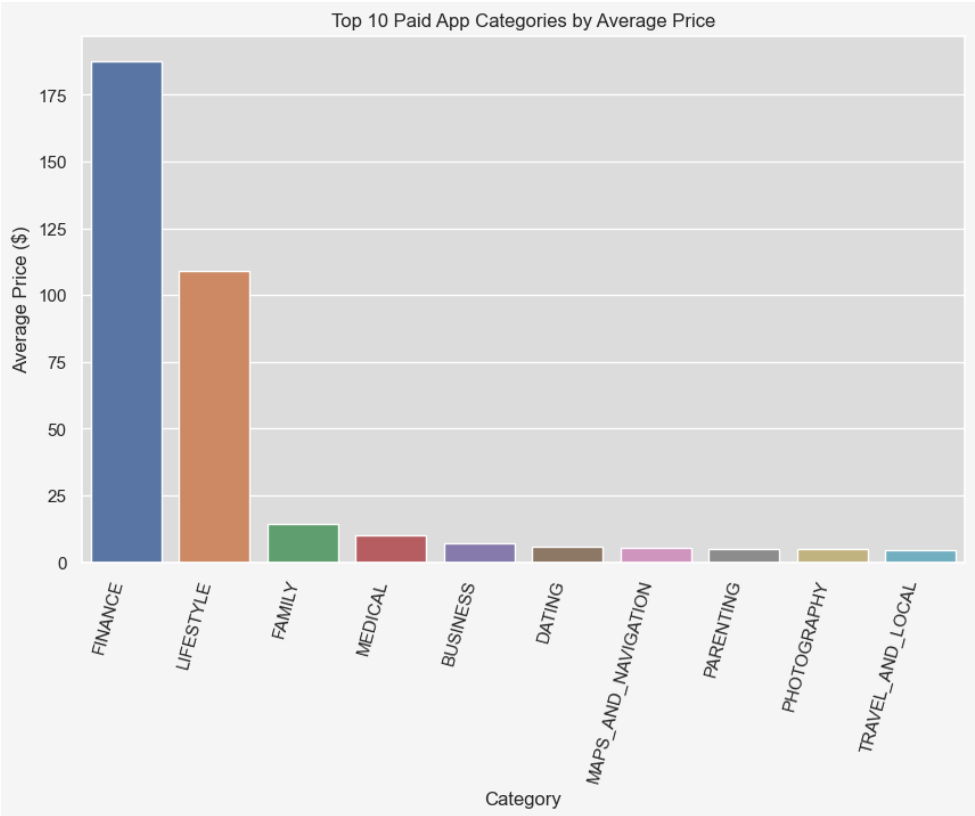


This boxplot shows the distribution of ratings for different review counts. The boxplots are divided into 10 categories based on the number of reviews (in increments of 8 million). The plot shows that apps with higher numbers of reviews tend to have slightly higher ratings, with fewer outliers.

Question 5: What are the most expensive app categories on the Google Play Store?



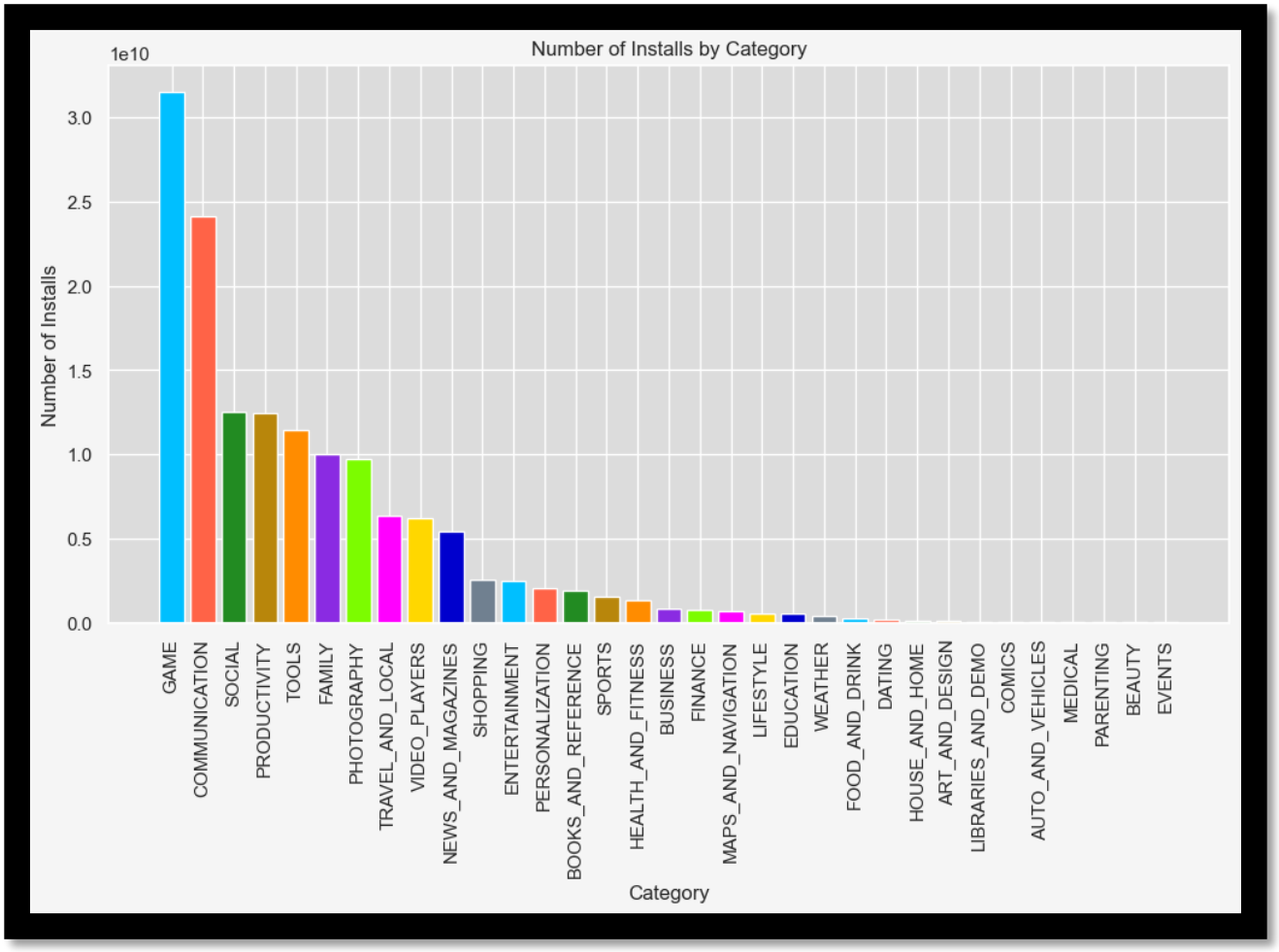
To tackle this question, two scenarios were looked at. First the average price of apps vs categories were compared(top graph) resulting in Finance, Lifestyle and Medical being the top 3 categories. With the highest average price just above $7. However we do have to keep in mind that over 93% of the apps were free.



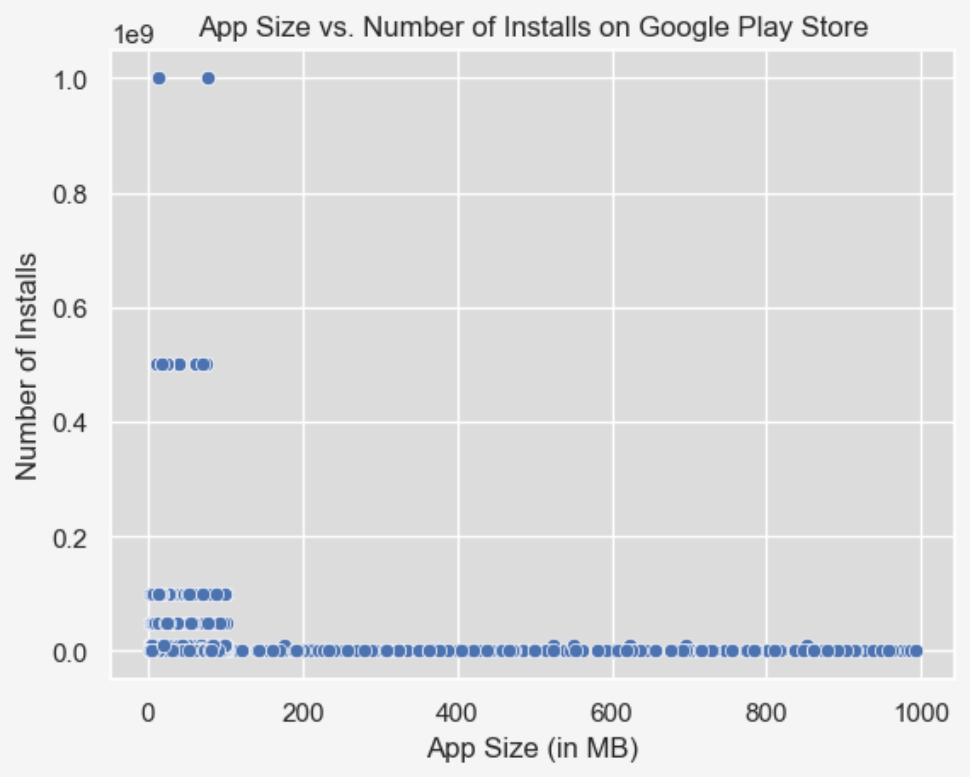
Secondly the comparison was between average price of all paid apps, vs the categories and found that Finance, Lifestyle and Family were the top three. There are some minor changes in the smaller categories, but for the majority it remains the same.

Question 6: Which app categories have the highest number of installs?

We found that the app categories with the highest number of installs are Games, Communication, Social, and Video Players & Editors. These categories have the most popular apps in terms of number of installs. A bar plot was used to visualize the total number of installs per category.

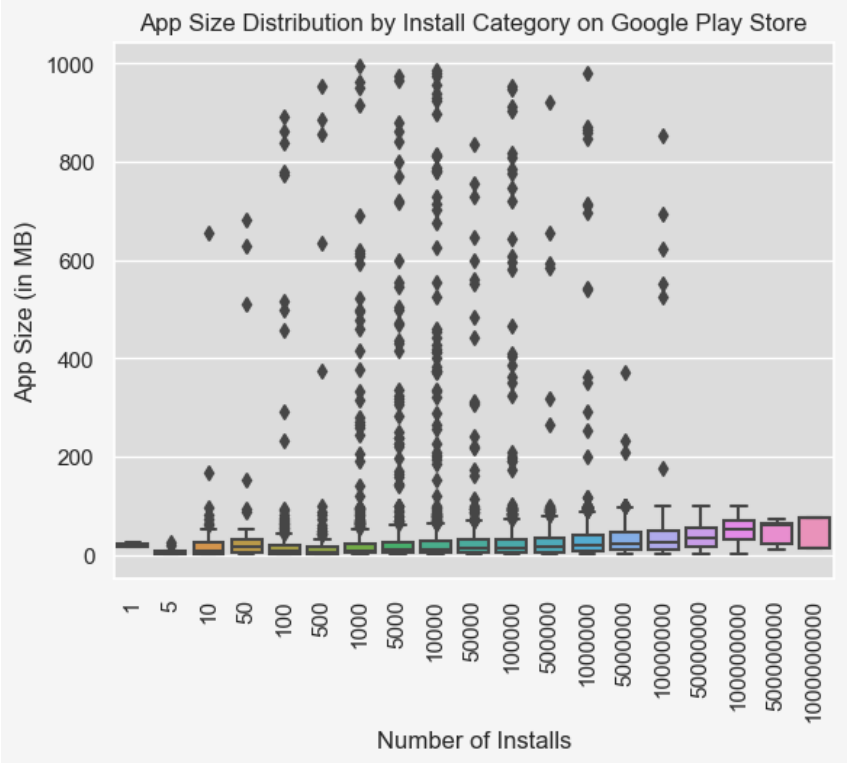


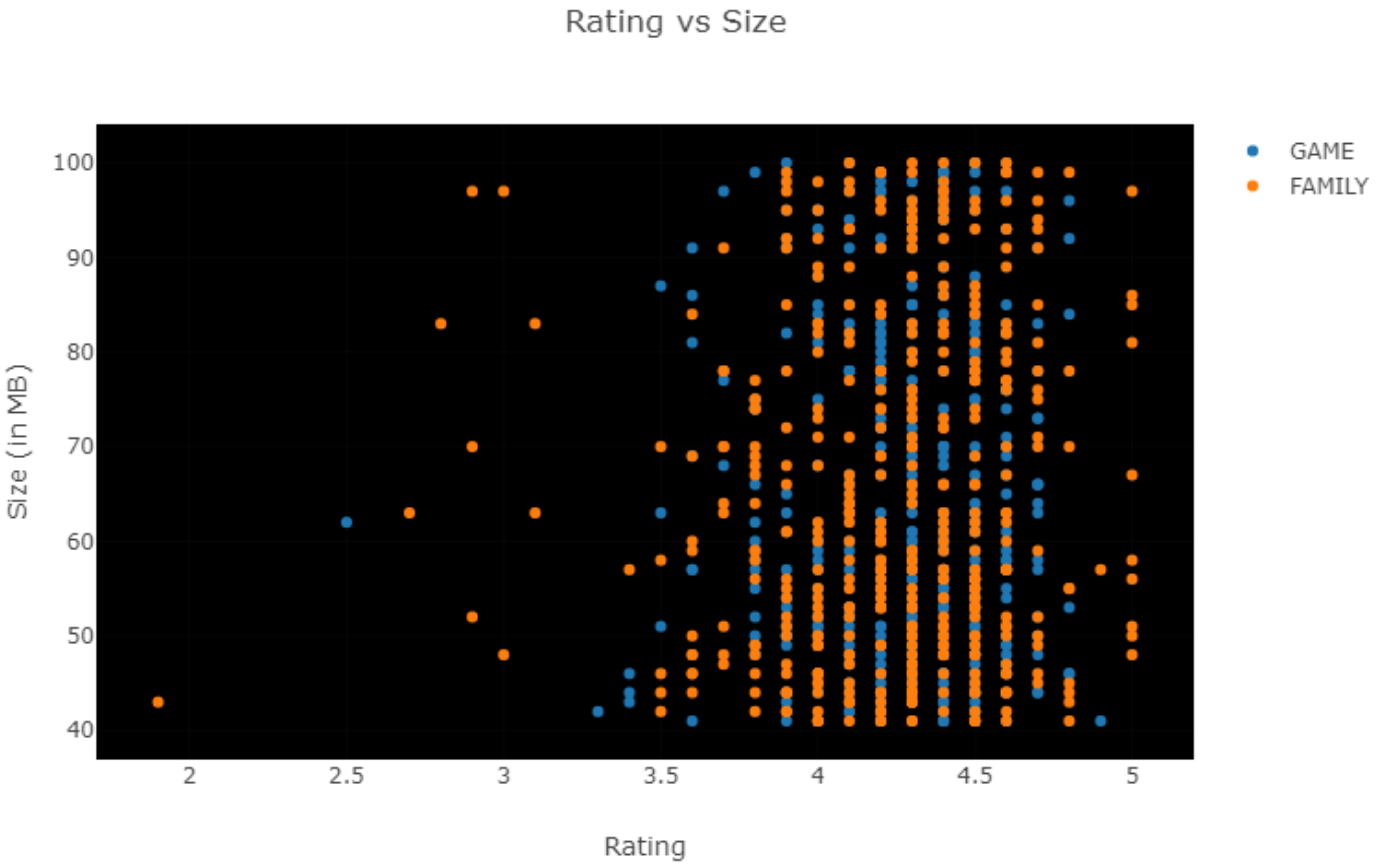
Question 7: How does app size affect the number of installs?



There is a weak positive correlation between app size and number of installs. This means that apps with larger sizes tend to have slightly higher number of installs, but the correlation is not very strong. A scatter plot was used to visualize this correlation.

This graph goes to show that similarly like above that apps with higher size have a greater number of installs (correlation is weak).





This is a sample of a graph to better illustrate a comparison between the Game and Family category. It can be seen that the bulkier apps (greater in MB) tend to have a higher rating.

Statistical Analysis

Sample t-tests were taken. The first t-test was a two sample test on the average ratings where it was concluded that average rating of paid apps were greatly different than the average rating of the free apps.

The second is a one-sample t-test on the "Finance" category prices with the null hypothesis being that the average price of the "Finance" category is equal to the overall average price of all categories. If the p-value is less than the significance level of 0.05, we reject the null hypothesis and conclude that the average price of the "Finance" category is significantly different from the overall average price of all categories.

```python
from scipy.stats import ttest_ind

# Split the data into paid apps and free apps
paid_apps = df[df['Type'] == 'Paid']
free_apps = df[df['Type'] == 'Free']

# Perform a two-sample t-test on the average ratings
from scipy.stats import ttest_ind

t, p = ttest_ind(paid_apps['Rating'], free_apps['Rating'], equal_var=False)

print('T-statistic: {:.2f}'.format(t))
print('P-value: {:.2f}'.format(p))

if p < 0.05:
    print('Reject null hypothesis: the average ratings of paid apps are' +
          'significantly different from the average ratings of free apps')
else:
    print('Fail to reject null hypothesis: there is not enough evidence to suggest that' +
          'the average ratings of paid apps are significantly different from the average ratings of free apps')
```

✓ 0.1s

```
T-statistic: 3.40
P-value: 0.00
Reject null hypothesis: the average ratings of paid apps aresignificantly different from the average ratings of free apps
```

```python
from scipy.stats import ttest_ind
# Obtain the sample data for the "Finance" category
finance_prices = df[df['Category'] == 'FINANCE']['Price']

# Compute the mean price for all categories
overall_mean_price = df['Price'].mean()

# Perform the one-sample t-test
t_statistic, p_value = stats.ttest_1samp(finance_prices, overall_mean_price)

# Print the results
print("T-statistic: {:.2f}".format(t_statistic))
print("P-value: {:.2f}".format(p_value))
if p_value < 0.05:
    print("Reject null hypothesis")
else:
    print("Fail to reject null hypothesis")
```

```
T-statistic: 2.20
P-value: 0.03
Reject null hypothesis
```

Conclusion

This was an interesting data set, in which insight was gained. Relationships between various factors of applications were analyzed and compared. From this data set, I would make some key takeaways which are but not limited to :

- Bulkier apps tend to get better ratings
- Free applications tend to get a harsher critic in terms of rating vs paid apps
- Individuals primary downloaded application is for gaming purposes

With more accurate information and more categories, such as revenue per game, in app purchasing, locations data we can provide further insight to this data and gain more understanding.