Final Report

Google Play Store Complainers

A Data Analysis of 2M Google Play Store Reviews

Cameron Carton

UCD School of Computer Science

University College Dublin

May 13, 2022

Table of Contents

Abstract

The Google Play Store is a digital marketplace where people can publish and sell android applications, users that download these apps can then write a review and leave a score rating 1 - 5 stars. With so many apps on the platform (3.48 million as of April 2022), it has become more important than ever to make high quality apps that stand out from the crowd of mediocrity. From Game developers to business savvy Entrepreneurs, there's opportunity for immense lucrative success. Knowing your market is key however, this project researches the reviews left by users on apps, dissecting the sentiment and contextual implications of the review content, and how that influences overall sentiment and score in apps and their genre. Researching if some genre have consistently lower scores and sentiment than others, and why that might be, whether or not high and low scoring reviews have a difference in the topics they discuss, and what factors may influence a users review, and whether or not an apps monetization scheme has any effect on a reviews score or sentiment. This is achieved by collecting more than 2 million reviews from the google play store from across 24 different app genre, calculating their sentiment and analysing their results and implications.

Chapter 1: **Introduction**

This project is apart of my 3rd year Data Science in Computer Science curriculum in University College Dublin. The project starts in the second half of the spring semester, as part of a module called Data Science in Practice directed by Barry Smyth. Students are expected to work in pairs to complete their project, I was unfortunate in that my partner dropped out of the course in the early weeks of the module, however, I was fortunate in that going solo meant I had total control over the project. Students are expected to work together and review their partners code as a team exercise, obviously I didn't have a partner, but myself and another solo project connected on Gitlab and Kreoh so we could review each others code.

Data Science in Practice being a 15 credit module meant Students are expected to work 30-40 hours a week for 10 weeks, providing a short presentation on every Monday and Friday detailing what our plan for the week was and what we achieved that week. We were also expected to meet our demonstrators on a zoom call every Wednesday, where we could ask questions and get feedback on the project.

The first step of the project was the proposal, the original idea for the project topic came from a dataset I found on Kaggle that contained hundreds of thousands of reviews for Dating apps on the google play store. I then decided to widen the scope of the project and include multiple different genres from the Google play store. However, this now meant I needed to collect the data myself, rather than use an existing dataset. The main topic I wanted to focus on in the reviews, was the sentiment present in the review text, and how that was influenced by other factors, like app genre and app quality.

A "boot camp" week was organised in the early weeks of the module, which set out to help students finalise their ideas and grow their project ambition. The week challenged us and taught the basics of tackling a large data science assignment. Examples of data collection and cleaning, as well as fantastic data visualizations gave us a deeper understanding of what was expected from our projects. Of course we were not expected to go as in detail as the sample project made by the university professor with decades of experience, but a high standard was expected, and the project was meant to challenge students in a way previously not explored in other modules.

With the project finalised, it was on to the programming. In the first week all of the data collection and cleaning would take place, while some students still had to find the right dataset they were looking for, I was building a scraper to snatch data on reviews from thousands of apps on the Google Play Store. The next week was allocated to the first research question "Which genre has the most negative marketplace?", the week after that was allocated to the second research question "Which Reviewers are the most responsive?", and the week after that was allocated to the last research question "Does an App's Monetization Scheme influence Review Ratings?". The last 2 weeks of the module are dedicated to making and presenting the project results, as well as writing this report.

1.1 What to expect in this Report

Chapter 2 discusses the basics of the platform, and why it makes a great data science project, outlining the data attached to each review and the information it tells us. Chapter 3 explains the data I have collected for the project, as well as why I collected the data, what it can explain about different apps and their genre, the method in which the data was collected, designing and programming the crawler and scraper used to collect the data, the various forms of filtering and cleaning of the data, and why it needed to be filtered and cleaned. Chapter 4 outlines the 3 research questions this project discusses, "Which genre has the most negative marketplace?" discussed the sentiment in apps and their genre, and how that may be influences by apps and genre score, "Which Reviewers are the most responsive?" discusses the actual text of the review and what factors effect reviewers opinions, and finally "Does an App's Monetization Scheme influence Review Ratings?" discusses the differences in app score and sentiment across different monetization schemes, and how these payment and monetization systems may influence a users opinion on the app, as well as how the app creators may price their apps to produce the most lucrative outcome. Chapter 5 concludes the project and its findings, summarising the results, as well as the limitations of the project, its pitfalls and what could be done better in future analysis.

Chapter 2: What is the Google Play Store?

The Google play store is a digital marketplace where people can publish and sell their android applications, users can then download these applications, and if feel so inclined, write a review and leave a star rating, 1 - 5 stars. The Google play store is the main app marketplace on android devices, the main competitor to Apple's App store. Every android device has immediate connection to the store, acting as an integral part of the android experience.

There are mountains of data available on the Google Play Store. Millions of users and app creators inhabit the platform, and with over 111 billion app downloads in 2021, up from 76 billion in 2018 [1], the market is set to grow even bigger in the coming decade. Hundreds of billions of user reviews available to the public for analysis.

2.1 Star and Score Rating

As mentioned previously, users can leave a star rating on a downloaded app. This rating can range from 1 to 5 stars. Most users leave 5 or 1 star ratings. An apps score is displayed beside the apps name and icon, so users can quickly decide if they want to avoid a low scoring app or target a high scoring app while browsing. An app's score is calculated from its current and relevant reviews, it is very important to have a high app score to attract more users and gain downloads.

This project will analysis these app scores across genre and come to conclusions whether or not an apps genre is important to the apps success. With billions of apps on the market, it is outside the scope of this project to analysis each and every one of them. A small sample size of apps have been taken to analyze.

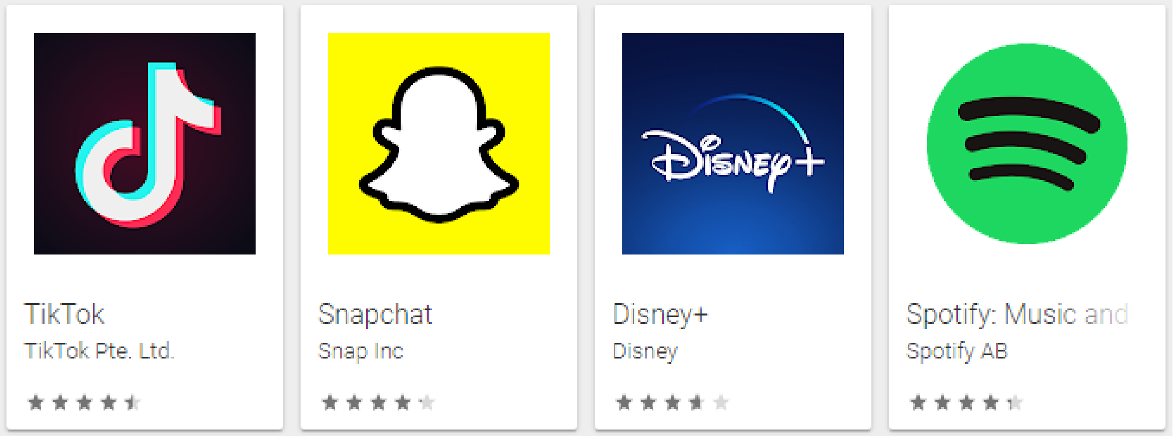


Figure 2.1: A sample of some of the top apps on the store, TikTok, SnapChat, Disney+, Spotify. TikTok alone having over 48 million user reviews.

2.2 Review Likes

Users are able to like and upvote reviews they found helpful. A review's like or thumbs up count indicates how many likes a review has received from other users. This system pushes more informative and mass approved reviews to the top of the relevant reviews list, these reviews are then shown to new potential users checking out the apps page. If these reviews are convincing enough, users may be persuaded to download the app, or if the reviews are bad, scared off entirely [2].

Reviews are the most important part of an app page. They can make or break a successful app. That is why the sentiment in the reviews is interesting to analyze, what makes a review negative, do users support and upvote negative or positive reviews more often, and does this differ genre to genre? This like count statistic can also tell us more about users in a genre, as most people who download apps, do not review them, but can still upvote reviews.

2.3 App Genre

Apps are split up into different genre categories. There are over 30+ genres as well as the Games genre having 17 different sub genres. The genre categories range from silly, wacky tools to complete financial management and banking, having a huge array of diverse genres with apps to follow. Something for everyone, whether you're into crypto and NFTs, or just want to play some chill, casual mobile games, vital apps for Navigating foreign countries, or fun learning apps that teach you new skills. These varied genres give plenty of reviews to analyze from all kinds of people from all backgrounds. Some genre user-bases may have different standards and goals in terms of their apps [3].

2.4 App Monetization

An app can be free or paid, it can contain ads and it can contain In App Products (IAP) which is more content to buy within an app. Most apps on the Google play store are free and contain some kind of monetization like ads or IAP. This is how apps make millionaires, whether the app is a one off purchase or it's flooded with annoying ads and micro-transactions (IAP), most apps have some kind of monetization these days, and it would be interesting to see if app score and sentiment are influenced by these factors.

Chapter 3: Data

The dataset used for analysis in this project was collected using the python library google-play-scraper. The library comes equipped with APIs that make it easy to navigate and access information on the Google Play Store. With more than 30 genre categories, a number of the genres are extremely similar and would share user-bases. For this reason, I have decided to limit the number of genres sampled to 24.

3.1 Data Collection

The first part of Data collection required a large number of app ids. App ids are the digital ids associated with apps and their app pages on the Google Play Store. These app ids are required as input to the review scraper, returning the desired number of reviews for the specified app id. The crawler is designed to take a "seed" app id, and from that id, the algorithm recursively navigates through the similar app ids of the similar app ids, collecting the necessary data from each app. "Similar app ids" is a list containing other app ids that can be accessed from an app id. These similar apps are provided and calculated by the Google Play Store. All of the ids for a "seed" app id, are stored in a file. Setting multiple "seed" app ids, allows a wider range of apps to be collected. After all the app ids have been collected, each of the files are then combined into 1 master app id dataset, removing all the app duplicates.

Once all the app ids were collected, the list was input into the review scraper. The review scraper was designed to take a list of app ids and return the data on the most recent 1000 reviews for an app. The reviews for each app are stored in a csv file. The program takes around 12 - 16 hours to run in jupyter notebook. The review scraper is set up to not retrieve reviews it already has for apps. This is so if the program stops mid-run, it does not need to be restarted from the beginning. Once all the reviews had been collected, each review file for each app was combined into a new master dataset, containing the review's data, and all the data connected to the app that the review was from. Before any additional cleaning or filtering on the dataset, the total number of rows collected was 2479592, consisting of 16 columns of data.

3.2 Data Cleaning

The data didn't require too much cleaning, it was mainly removing columns of data I didn't need, and removing apps and genre with low number of reviews. The Reply Content column usually contains the message content of the reply left on a review, this data was not of interest to me, as most of the replies looked to be left by employees of the app companies, with not much interesting analysis to be had, however, I did re-purpose this data and make the Received Reply column, which returns true if a reply was left on the review, and false if no reply is left. The In App Product Price column contained a text string with the range of In App Products, example, "€7.99 - €59.99 per item", this string was cleaned up and only the euro values were split into two separate columns, "In App Product Price Low" and "In App Product Price High". The Games genre has numerous sub genre, and so a new column called "Sub Genre" was created specifically for the Games genre, which holds the secondary genre for games, other genre just get their original genre repeated in the sub genre column. The review text data was tokenised using the nltk python library. This is used when analyzing the commonality of words in different types of reviews. The tokenisation process splits the words in the review into a big list, then all the unnecessary stopwords are removed from the list (words like "the", "I", "was"), leaving only the more important words.

3.3 App Score

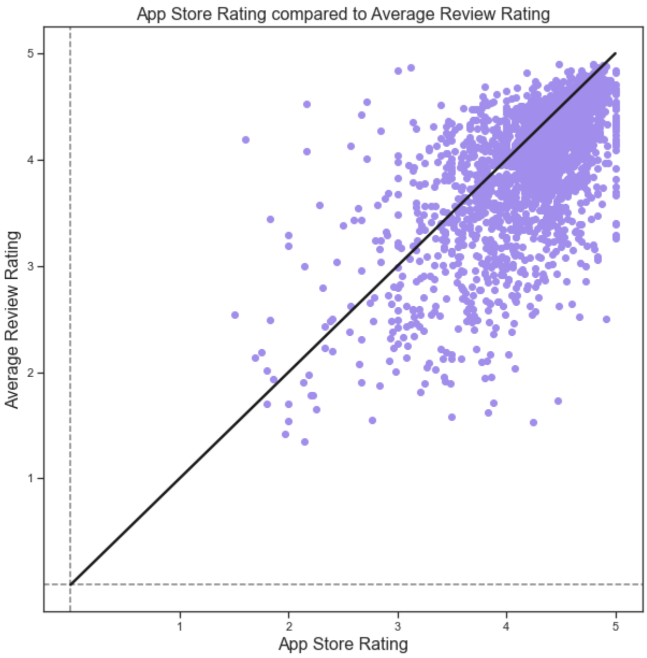
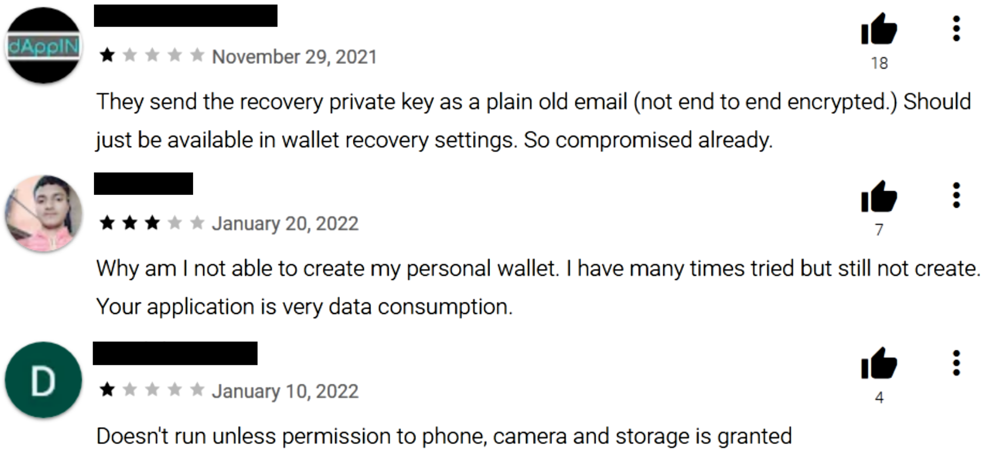


Figure 3.1: Mean Review Score compared with Google Play calculated App score.

The Score rating of an app is calculated by the Google Play Store, based on the most recent review ratings on the app, it is not public information on how recent a review has to be for it to impact an apps score. For this project, I have decided to not use the App Score calculated by Google, and use the mean average score across all the reviews for an app in the dataset. As seen in Figure 3.1, the mean average review score for an app, is not necessarily close to the Google Play Store calculated app score. I have calculated an apps sentiment from its reviews in the dataset, and so that is why I am using the mean average review score instead of the Google Play Score. Whenever this report refers to an apps score, it is referring to the mean review score average, which is what I am using instead of the Google Play calculated app score.

3.4 Sentiment

The main thing I analysed across all research questions was the sentiment of reviews. The sentiment score lying between -1 and positive 1. Reviews with a positive sentiment being greater than 0, and reviews with a negative sentiment being less than 0. The sentiment score was calculated using a python library called textblob. Words like "good" and "great" would return positive sentiment, while words like "bad" and "awful" would return negative sentiment. This sentiment calculator was not perfect however, and although it had high accuracy in most reviews, reviews with "not" weren't accounted for, for example, "this app is not good" would return a positive sentiment because of the word "good", however "not" clearly indicates that the review should be of negative sentiment. This does not significantly impact the overall sentiment of apps and genre, but a more accurate sentiment calculator would be more ideal if I was to do this project again.

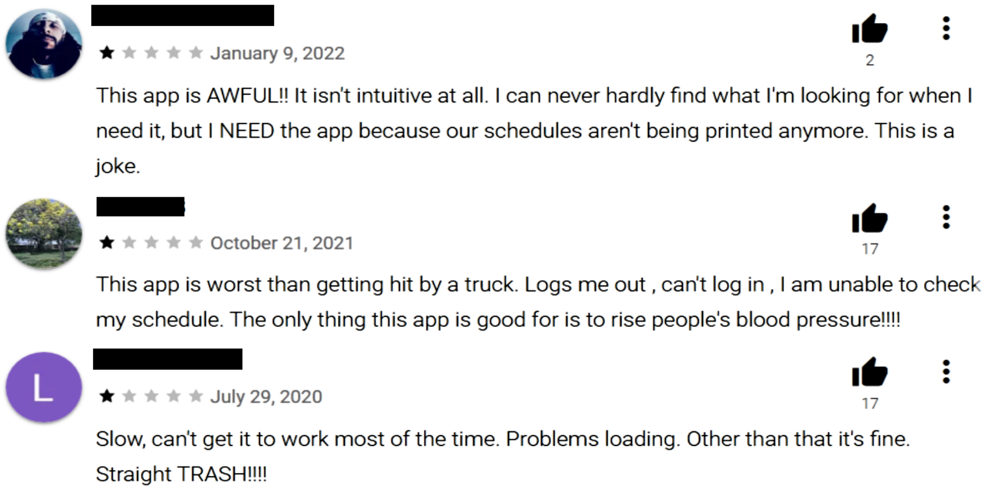


[redacted]
★☆☆☆☆ November 29, 2021
18
They send the recovery private key as a plain old email (not end to end encrypted.) Should just be available in wallet recovery settings. So compromised already.

[redacted]
★★★☆☆ January 20, 2022
7
Why am I not able to create my personal wallet. I have many times tried but still not create. Your application is very data consumption.

[redacted]
★☆☆☆☆ January 10, 2022
4
Doesn't run unless permission to phone, camera and storage is granted

Figure 3.2: Sample of low reviews from the highest sentiment app in the dataset, Talken.



[redacted]
★☆☆☆☆ January 9, 2022
2
This app is AWFUL!! It isn't intuitive at all. I can never hardly find what I'm looking for when I need it, but I NEED the app because our schedules aren't being printed anymore. This is a joke.

[redacted]
★☆☆☆☆ October 21, 2021
17
This app is worst than getting hit by a truck. Logs me out , can't log in , I am unable to check my schedule. The only thing this app is good for is to rise people's blood pressure!!!!

[redacted]
★☆☆☆☆ July 29, 2020
17
Slow, can't get it to work most of the time. Problems loading. Other than that it's fine. Straight TRASH!!!!

Figure 3.3: Sample of low reviews from the lowest sentiment app in the dataset, mywork 1611.

Taking a look at the most positive sentiment app in the dataset, Talken, a multichain crypto and NFT wallet, the low scoring reviews aren't particularly negative (Figure 3.2), but are more constructive, pointing out problems with the app. While the low reviews of the most negative app (mywork 1611) in the dataset use far more dramatic and emotional language (Figure 3.3).

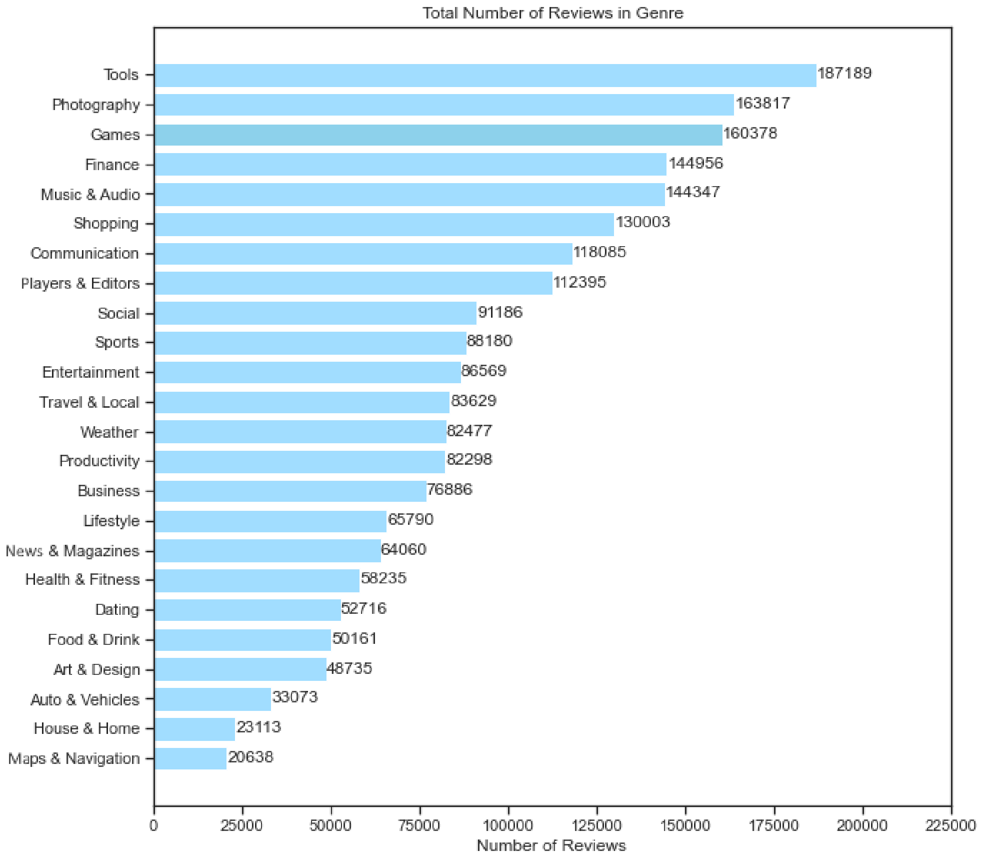


Figure 3.4: Review counts in each genre after cleaning

Chapter 4: Research Questions

In this chapter I present the results and analysis of each research question. In each case I describe the approach taken as well as the findings, providing insight into the results and their implications.

4.1 RQ1: Which Genre has the most Negative Marketplace?

Before diving deep into the sentiment and app score of individual app genres, first the waters needed to be tested, figuring out if app score and sentiment are indeed correlated. The most obvious approach to me at first was to simply average out the sentiment in each review star rating, then begin deeper analysis into all areas of a review.

4.1.1 Does Sentiment correlate with Star Rating given?

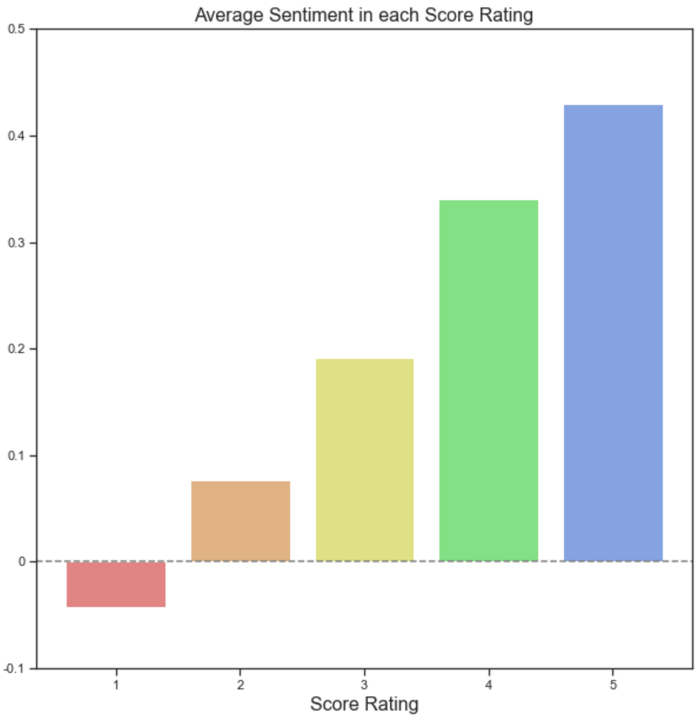


Figure 4.1: Mean Review Sentiment for each Star Rating.

With the progressively higher mean sentiment the higher the star rating, it is very apparent that sentiment and score are directly correlated. Interestingly, the 1 star rating does not have nearly the same degree of negativity as the 5 star rating has positivity. People tend to stay more positive on average when reviewing an app in general. This is why finding the negative outliers could give greater insight into problematic genres.

Now that we know review sentiment and star rating are correlated, it is important to discuss the sentiment and score correlation in the apps rather than just their individual reviews. The sentiment and score of an app is the mean values of all their review sentiment and star ratings. This can provide insight into whether or not apps with more or less reviews have greater correlation with sentiment.

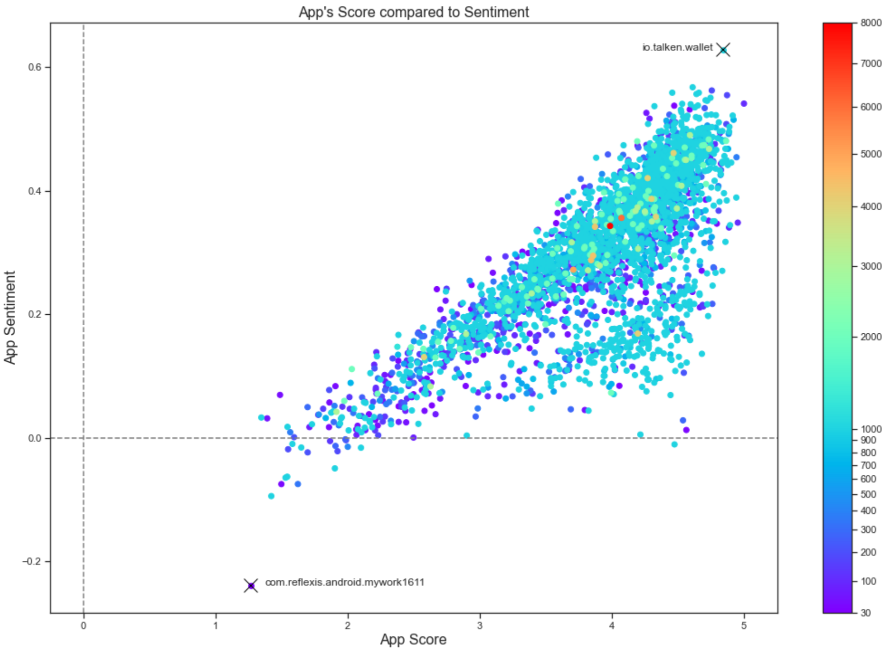


Figure 4.2: Individual Apps with their Sentiment and App score, the color bar relating to how many reviews an app has in the dataset.

The apps with low review counts (<500 reviews marked in blue and purple) have a relatively wide distribution in Figure 4.2, apps with higher review counts (1000 reviews marked in turquoise) have a tighter correlation with their sentiment and app score, and stick closer together, apart from a large group of outliers that have relatively low sentiment to their app score. This may mean apps and genres with lower review counts might not have the most accurate results, and with more reviews in these apps, their sentiment and app score would have a closer correlation. That is something to keep in mind with further analysis in the future. Figure 4.3, Highlighting the group of outliers and we can see it's actually the Games genre. The entire genre has significantly lower sentiment relative to their app score, lower than any other genre. This may be due to the genre typically having a reviewer-base with significantly more young people and children who may use more dramatic and emotional language when reviewing. A better view of these results can be seen in Figure 4.4. Games in particular having a huge drop in the sentiment rankings relative to its app score rankings. Which is different than a genre like Dating which has both low app score and low sentiment. On the other end of it, we have Maps and Navigation which has a much higher sentiment ranking relative to its app score ranking. Games is clearly a huge outlier in this dataset having such a large drop in sentiment.

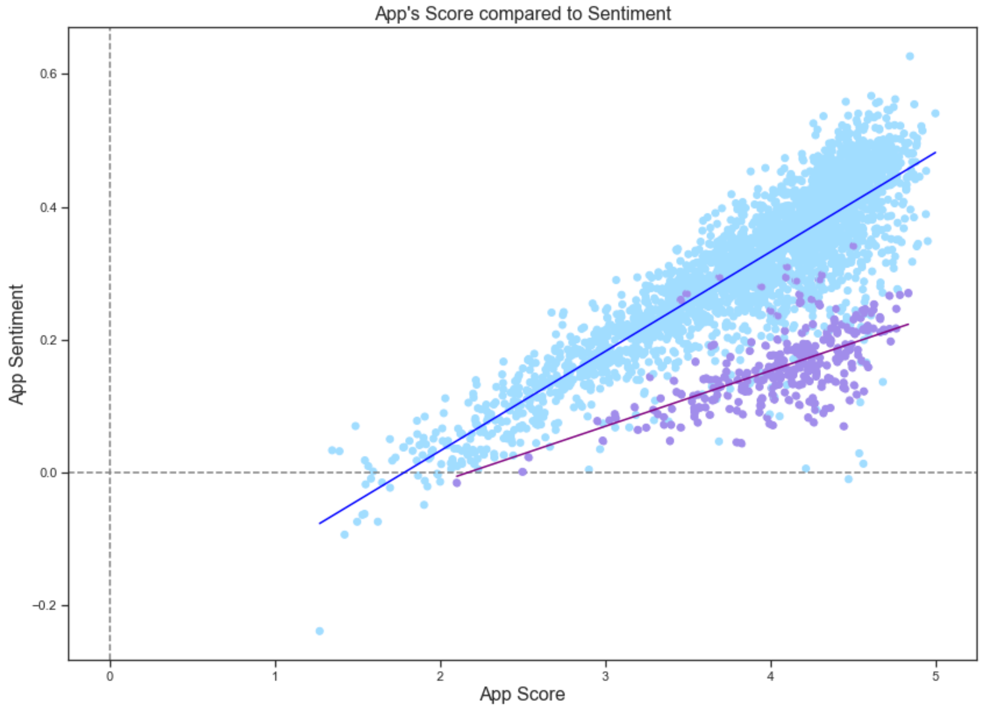


Figure 4.3: Games Genre highlighted in purple. Lines of best fit added to Games and other genres.

4.1.2 What type of Reviews get more support?

A review is marked "good" if it has a star rating of 4 or 5, a review is marked "bad" if it has a star rating of 1 or 2, 3 star reviews are marked "Mid", they are not of interest to us. Most people tend to leave 5 or 1 stars when rating an app. Figure 4.5 lets us see the percentages of both good and bad reviews in each genre. Games actually has the second least percentage of bad reviews, whilst unsurprisingly Dating has the highest percentage of bad reviews. Maps and Navigation having a high percentage of bad reviews is an example where having more reviews in the dataset could change this result, Maps and Navigation has the lowest amount of reviews for a genre.

An important factor that distinguishes positive and negative genre, is the support users show to the reviews in the genre. The positive and negative reviews in this section of analysis are actually calculated genre specific, so a positive review for a genre, is a review with sentiment greater than the median sentiment for the genre, and a negative review would be reviews with sentiment lower than the genres median. This is to essentially normalize what a positive and negative review is in each genre. As a secondary measure to ensure that reviews are at the extreme ends of the spectrum (very positive or very negative), the negative reviews only include 1 and 2 star reviews (Bad Reviews), while the positive reviews only include the 4 and 5 star reviews (Good Reviews). This is to eliminate the outliers in review sentiment that may be caused by miscalculations from the sentiment calculator python library TextBlob, example, "This app is not bad" would be calculated as negative sentiment when in reality it should be positive sentiment. Negative reviews tend to get significantly more likes/support. This may give a further glimpse into the sentiment of a genre, as users who don't necessarily leave reviews on apps, may still like and upvote the reviews they find most helpful. However, the like values on the reviews would be influenced by other factors like, how much traffic there is going through a genre, genres with a bigger user base would naturally have bigger like counts on both their positive and negative reviews.

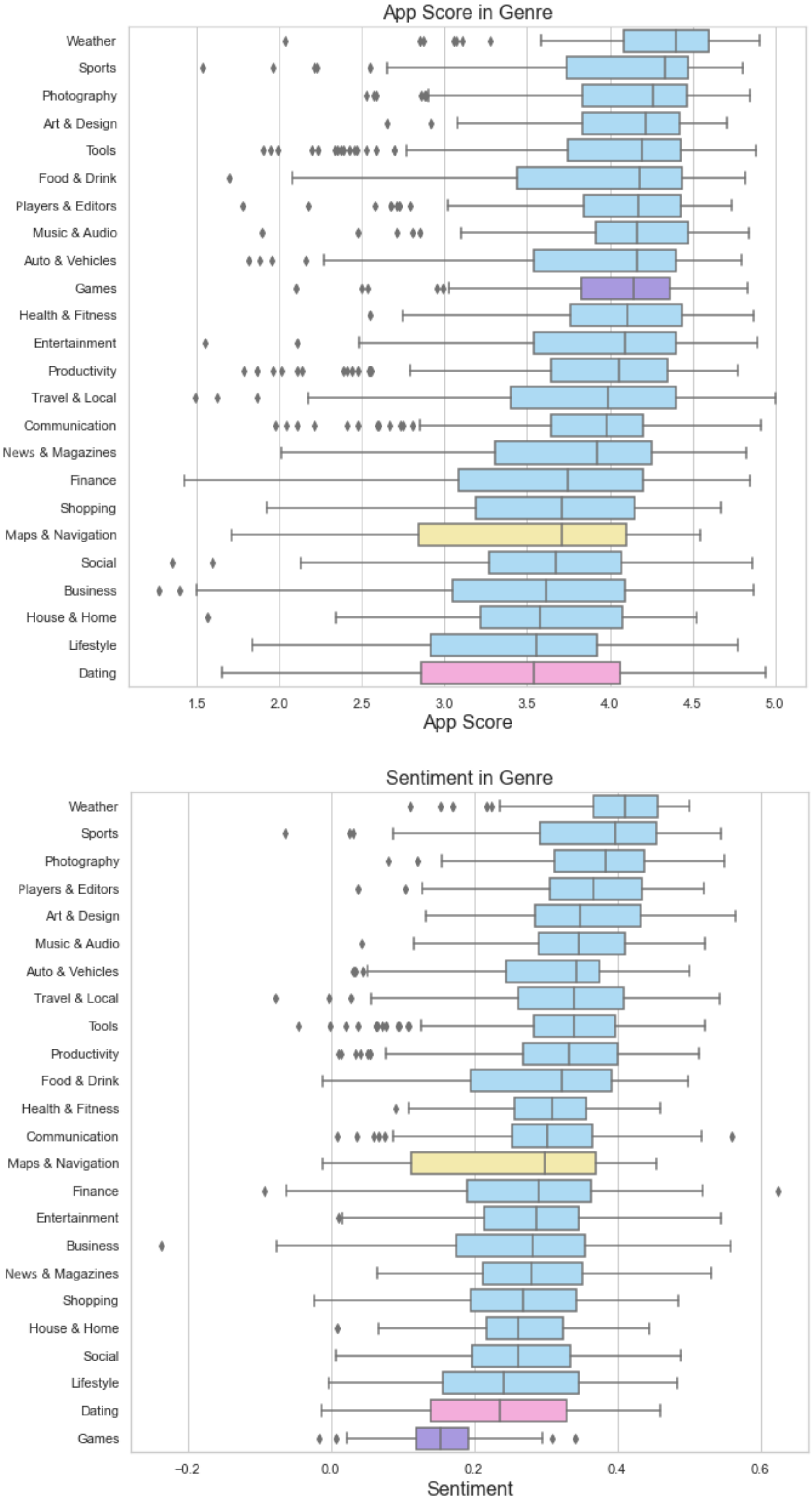


Figure 4.4: Genres ranked in order of their median Score in the first boxplot graph, and sentiment score in the second boxplot graph, the Games genre highlighted in purple, the Dating genre highlighted in pink, the Maps and Navigation genre highlighted in yellow. The Games genre takes a large dip in sentiment relative to its app score.

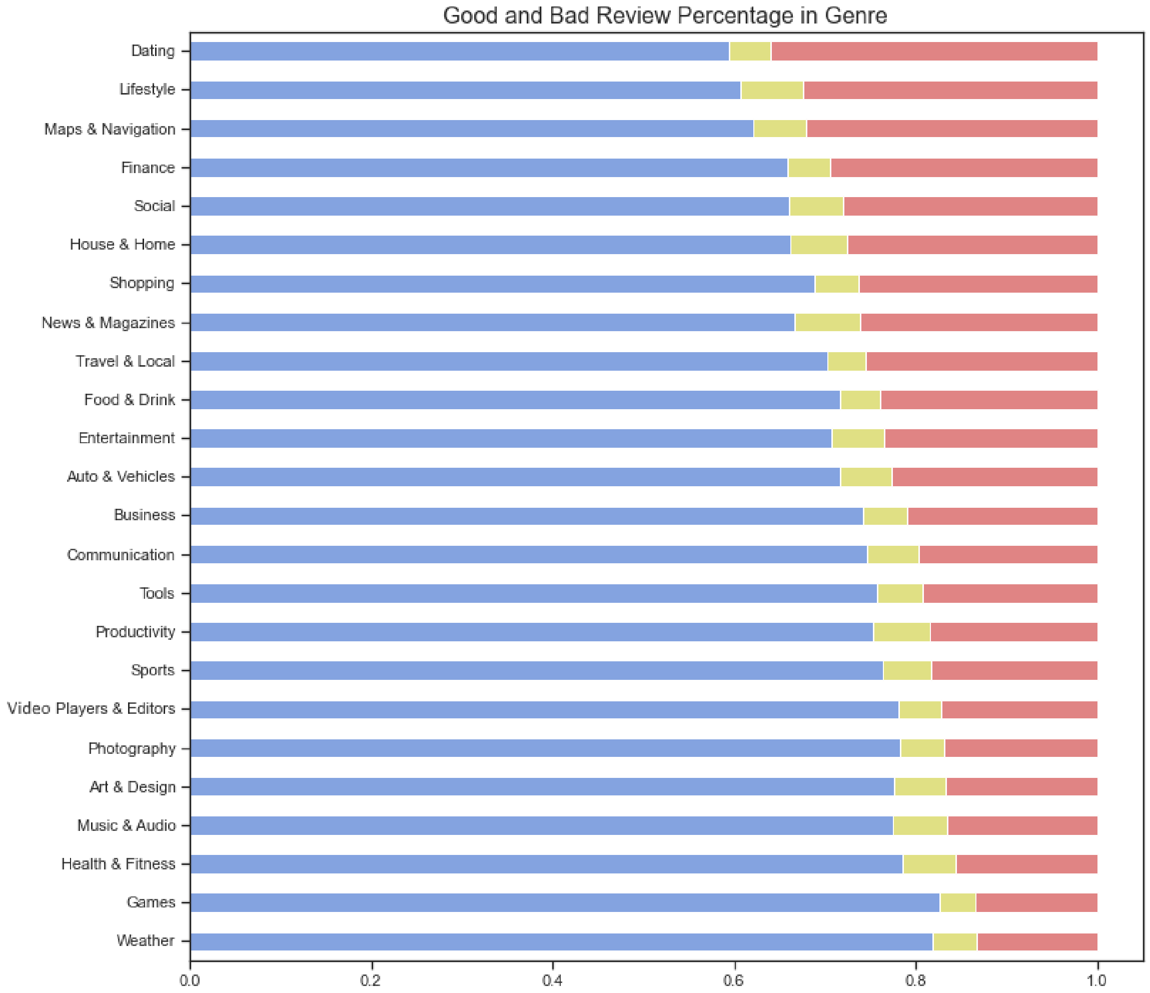


Figure 4.5: Good and Bad percentages in the total reviews of each genre, Good Reviews in blue, Bad Reviews in red, Mid Reviews in yellow.

The positive to negative review like ratio seen in Figure 4.6, gives a much clearer indication of the support shown on negative reviews in the genre. The Games genre falls to the lower ranks of the list, with just over 2 likes on its negative reviews for every 1 like on its positive reviews, significantly more positive than the Dating genre, which has just over 4 times the amount of likes on its negative reviews relative to its positive reviews. Maps and Navigation leading the way in positivity with less than 2 likes on its negative reviews. Surprisingly, the Weather genre has the highest amount of likes on its negative reviews, even though it is top rank in both app score and sentiment. This may be due to the apps having a very low threshold in regards to the quality, most weather apps meeting the standard expected, but outlier apps having very obvious and intolerable problems, leading to negative reviews on those apps with mass approval from the general user-base. The Dating genre gaining significant support on its negative reviews may be a similar case to the Weather genre, except rather than having high score and sentiment overall, the genre has low score and sentiment overall, meaning most apps have not met the standards of the user-base, with their negative reviews still gaining significant approval from other users. The Games genre on the other hand having high score but low sentiment, gets relatively low support on its negative reviews. This may be an indication that although dramatic language may be used by reviewers to convey their disapproval of the app, their reviews are not widely supported by other users.

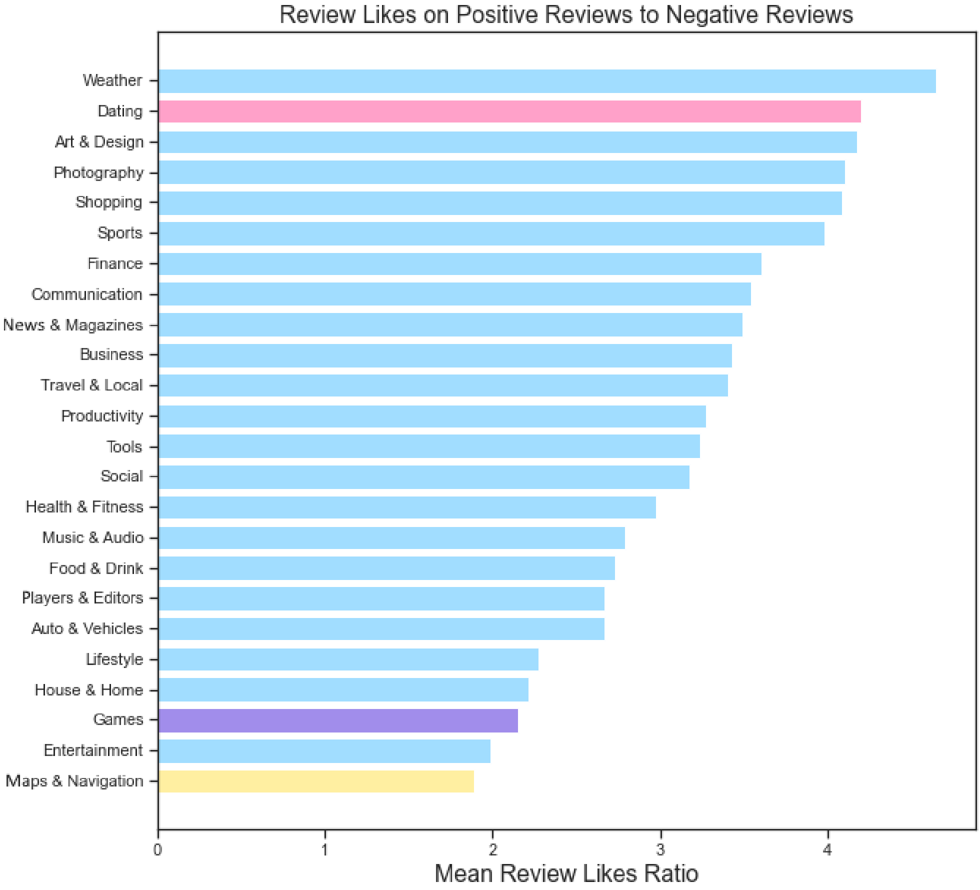


Figure 4.6: Genre Positive:Negative like ratio, for every 1 like on a genres positive reviews, there are this many likes on their negative reviews.

4.1.3 Summary

This research question was meant to identify the negativity and positivity of the genres and apps on the Google Play Store. Discussing what made the genres and apps Negative or Positive. First, we looked at the correlation between sentiment and the star rating given to an app, discovering that higher star rating indicates higher sentiment in a review, and that apps with higher review counts tended to show greater correlation with their sentiment and app score. Outlier genres like Games having significantly lower sentiment relative to their app score, or Maps and Navigation having higher sentiment relative to their app score. While genres like Dating and Weather have similar sentiment rankings relative to their app score rankings, both low or both high. The most interesting results coming from discussing the likes on the positive and negative reviews in each genre. Indicating that users like and agree more with low sentiment, negative reviews. With users in more negative genres like Dating, agreeing far more with its negative reviews than most other genre, and genres like Games, with users not showing significant support to its negative reviewer base.

4.2 RQ2: Which Reviewers are the most responsive?

Now that we are aware of the overall sentiment of reviews and genre, it is important to discuss the actual text of reviews, the content in the review and factors that may influence a users review opinion.

4.2.1 The Length of User Reviews

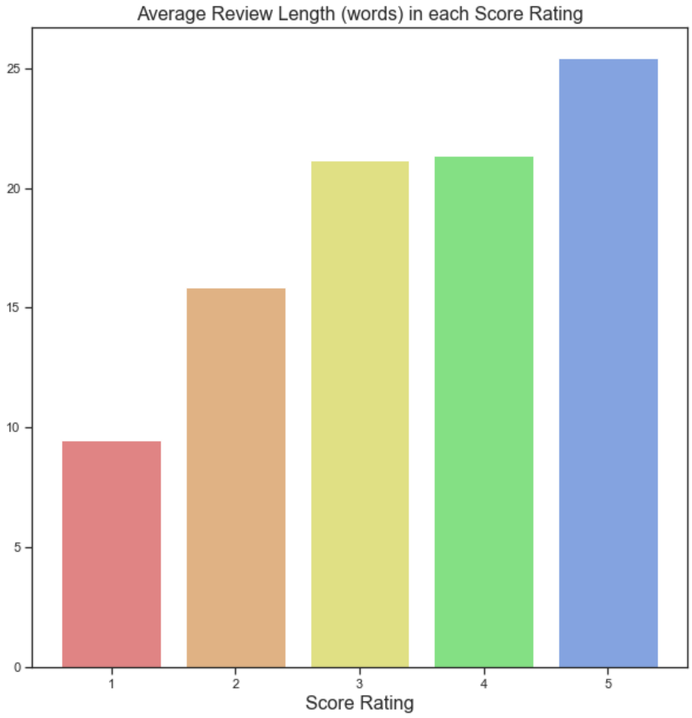


Figure 4.7: Each Star ratings mean average review length in words.

As seen in Figure 4.7, the length of a review tends to get longer the higher the star rating given. Averaging about 25 words in a 5 star review, and just under 10 words in a 1 star review. I would have thought this would be the other way round, with high score reviews having low review lengths and low score reviews having high review lengths. Initially thinking that people would leave short reviews with high appraisal more often, "This app is excellent!", "this app is great", but in fact it turns out to be the opposite. High scoring reviews usually encourage other users to download the app, outlining all the many reasons why it is worth getting, leading to a lengthy review, while low scoring reviews tend to point out 1 or 2 things that were problematic with the app, leading to shorter reviews overall.

Review sentiment does not appear to have any kind of bias toward review length, most reviews don't break the 100 word mark, which is near enough the 500 character limit for a review.

From Figure 4.8 we can see the massive difference in review length across the genre. Health and Fitness having the highest average review length at around 23 words, and Video Players and Editors having the lowest average review length at around 9 words. We can see now that genre review length does not necessarily correlate with the genres app score, there must be other factors influencing review length in genre, perhaps some genre have vastly more feedback and praise to give their apps.

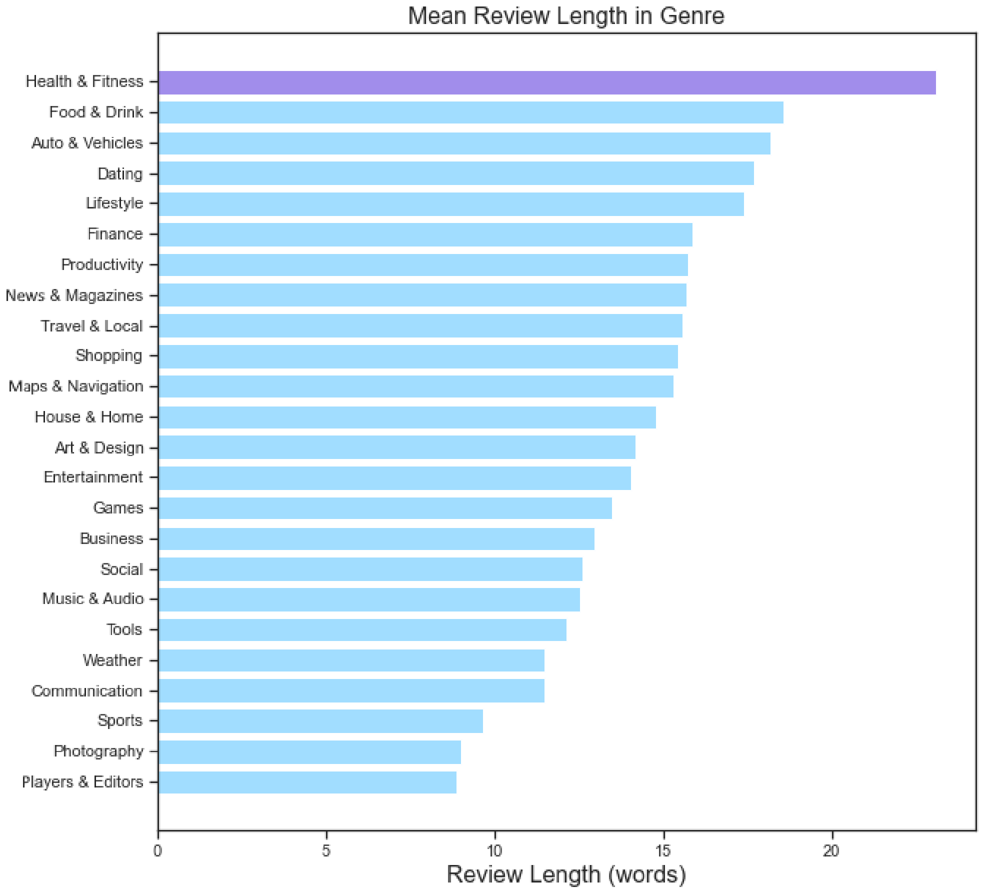


Figure 4.8: Each Genres mean average review length in words.

Reviews that received replies were longer on average, this may be due to longer reviews involving questions that prompt responses, rather than shorter reviews that just state their approval, "I love this App".

The average likes on reviews relative to their length can be seen in Figure 4.9. Reviews under 100 words appear to have a very direct correlation with the number of likes they received, with longer reviews gaining more likes than shorter reviews. However, after 100 words, things take a steep dive in like counts and then their correlations seem to evaporate, likes and review length not correlated at all. I think this may be due to most of the reviews with length >100 being outliers in the overall dataset, perhaps if more data was present with reviews >100 in length, there would be a clearer correlation.

I also looked at the score given on reviews relative to the time of day, however, the effect on score was too minimal and the results may have been misleading, more data is required to come to a conclusion on whether or not the time of day has an effect on the score given in a review.

4.2.2 Common words in Reviews

Looking at the Table 4.1, We can see the most common words in good and bad reviews. I did filter out most of the words relating to sentiment, for example, good, bad, awful, great. This allowed more interesting words to float to the top of the list, that might give a better insight into what users really have to say about apps. Looking at the good review words we can see words like "easy", "useful". This provides insight into why users may be rating these apps 4 and 5 stars, an app being "easy" to use, or "useful". The bad review common words tell a different story. Multiple of the top words relate to an apps monetization scheme, "money", "pay", "ad", as well as more general things like "useless" and "update". It appears that a large amount of bad reviews critique the apps payment or ad scheme, with most of these reviews being negative and low scoring.

4.2.3 Summary

This research question set out to discuss the text content in user reviews on the Google Play Store. The length of a review varying widely in genre, with Health and Fitness having the most opinionated reviews. Higher scoring reviews generally being longer than low scoring reviews. Reviews that were longer were also more likely to receive a reply from the app creator, and had a greater chance to be liked by other users (as long as the review was under 100 words in length). Good reviews tended to focus on the "easy" usability of the app, while bad reviews tended to criticize the payment or ad scheme of an app, complaining about "ads" and "money", as well as the apps being "useless" and needing an "update".

Rank	Good_words	Bad_words
1	easy	work
2	fun	update
3	work	money
4	cool	please
5	useful	working
6	thank	fix
7	simple	pay
8	perfect	open
9	new	service
10	please	new
11	add	ads
12	free	try
13	keep	useless
14	find	tried
15	everything	free
16	recommend	keeps
17	always	people
18	fast	waste
19	ok	nothing
20	lot	download

Table 4.1: The top 20 most common words in Good reviews and Bad reviews

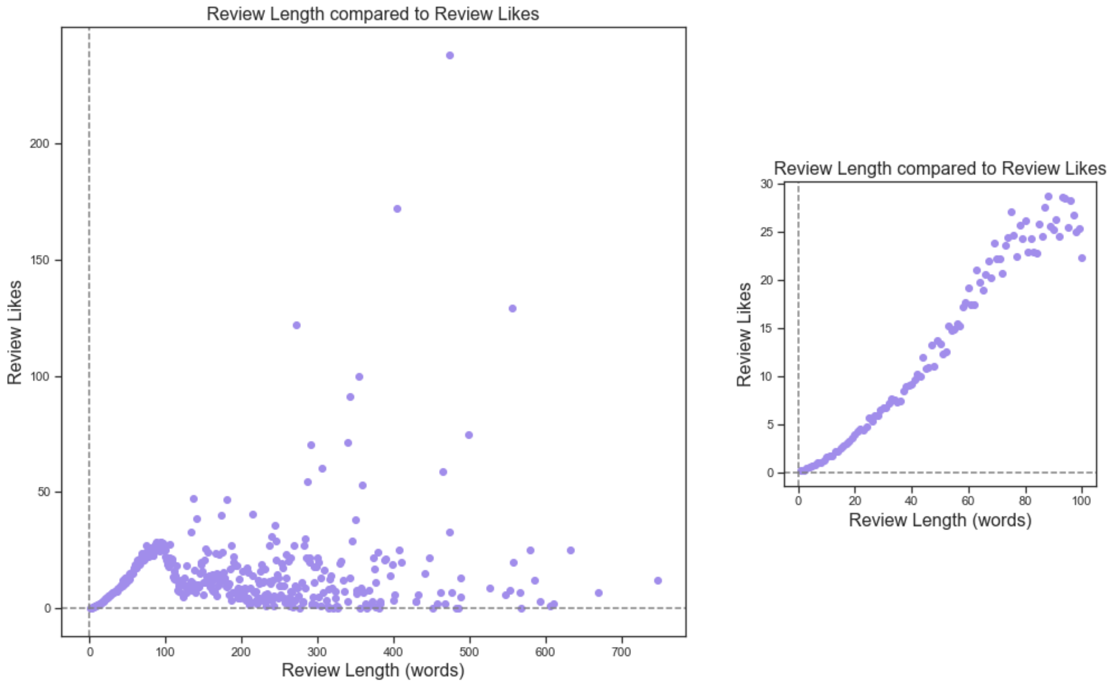


Figure 4.9: Mean Average likes on reviews length.

4.3 RQ3: Does an App's Monetization Scheme influence Review Ratings?

From the results in the second research question, we can see users often complained about an apps payment system when rating an app low stars. It is important to know if an apps monetization scheme has a significant effect on the overall score of an app, and how the sentiment of an app may be influenced by these payment and monetization options. An app can be free or paid, it can contain ads and it can contain In App Products (IAP) which is essentially more things to buy within an app. Most apps on the Google play store are free and contain some kind of monetization like ads or IAP. Most of the data collected is on free apps.

Monetization	app_count
Free	976
Free + Ads + IAP	947
Free + IAP	506
Free + Ads	419
Paid	155
Paid + IAP	46
Paid + Ads	3
Paid + Ads + IAP	2

Table 4.2: App Monetization categories and their app count in each

The apps in the bottom 2 categories in the Table 4.2, "Paid + Ads" and "Paid + Ads + IAP", are excluded from further analysis. There are not enough apps in this dataset in those categories for me to accurately analyse trends in those monetization schemes. It's clearly not a monetization scheme that's used very often, most users would avoid these types of apps.

4.3.1 Monetization Sentiment and Score

Paid apps tend to have a higher app score than free apps, but are lower in sentiment. The high app score may be due to paid apps being higher in quality relative to most free apps, but users having a higher expectation tied to this premium price, leading to lower sentiment overall. Free apps usually set a lower bar when it comes to user expectations. If the free app isn't up to your standard, you can delete it and move on, but if a paid app isn't up to your standard, you just lost money and gained nothing. People don't complain as much about things they get for free.

Apps that support Ads or IAP have both higher app score and higher sentiment than apps that don't support these monetization systems. This may be due to these apps being of higher quality and being able to take advantage of these forms of monetization, without it effecting their ratings. On the other hand, maybe most users aren't bothered by ads or IAP if the app is free. As mentioned previously, most of the data is on free apps, so I would imagine most of the results in the graph above are heavily skewed towards free rather than paid apps.

Looking at Figure 4.11, we can see the ranking of each monetization scheme from their median values. Free with ads being top in both score and sentiment, whilst paid apps with IAP being lowest in both score and sentiment. Paid and Free with ads almost neck and neck in the top ranks of score, but Free with Ads just ahead. I would imagine most people don't want ads in their apps, but when the app is free and useful, they tolerate the ads, giving it a higher app score on average, and being a free app, it carries low expectations, so naturally, sentiment is also very high in these apps. Paid apps being of high quality but users having high expectations, leading to low sentiment relative to its app score, all other free app categories overtaking paid app categories in sentiment. Paid with IAP being lowest in score and sentiment may be due to users disagreeing with their greedy monetization tactics. When buying an app, users don't generally want to see that more of the app is available through another paywall. Free apps with no monetization do poor in both score and sentiment, perhaps these apps are of low quality and that's why their creators don't bother monetizing them.

4.3.2 In App Product Range

The In App Product (IAP) Range of an app is the lowest to the highest price of an IAP within the app. These products are usually extra content in the app or more commonly in Games, these products are in the form of an in-game currency that can be purchased with real money. IAP are far more common in free apps then they are in paid apps.

Figure 4.12, Free apps with only IAP have the highest and also the widest range of product prices, with an average high of 84.13 and low of 5.07 euro. This is a significant difference from the high and low of Paid apps with IAP, with an average high of 14.38 and low of 1.80 euro. Free with Ads and IAP acting as almost a direct in between, with an average high of 52.99 and low of 2.55 euro. It seems as if the more monetization factors in the app the lower the range in In App Products (ads being an additional monetization factor, and paid being an even heavier monetization factor). App companies may be designing these price ranges to squeeze the most out of the customers, without arousing too much suspicion and alerting users to their predatory extortionate pricing models. Unfortunately, the majority of these In App Products are offered to fix a problem the app deliberately implemented. Often coming in "cool downs" for Games (periods in which users cannot play the game and have to wait a set amount of time between play sessions, this of course can be bypassed for a price paid with in-game currency that can only be bought with real money), or in the case of Dating apps like Tinder, users only get a limited amount of likes they can send to potential matches each day, this limit can be extended with paying real money, this may be a reason for Dating apps generally scoring low on app score and sentiment, having such a basic feature nerfed and locked behind a paywall.

4.3.3 Summary

This research question discussed if an apps monetization and payment scheme had a significant effect on the apps rating. First, looking at the categories of Free vs Paid apps and whether or not the apps supporting Ads and In App Products (IAP) had significant difference in app score and sentiment. Breaking these apps into further categories, having each combination of monetization, for example, Free, Free with Ads, Free with IAP, and Free with both Ads and IAP, and discussing whether or not certain payment schemes had better or worse sentiment and score. Free with Ads dominating in both App Score and sentiment, while Paid with IAP lay lowest in both app score and sentiment. The range of IAP being significantly lower in paid apps than in Free apps and Free apps with Ads.

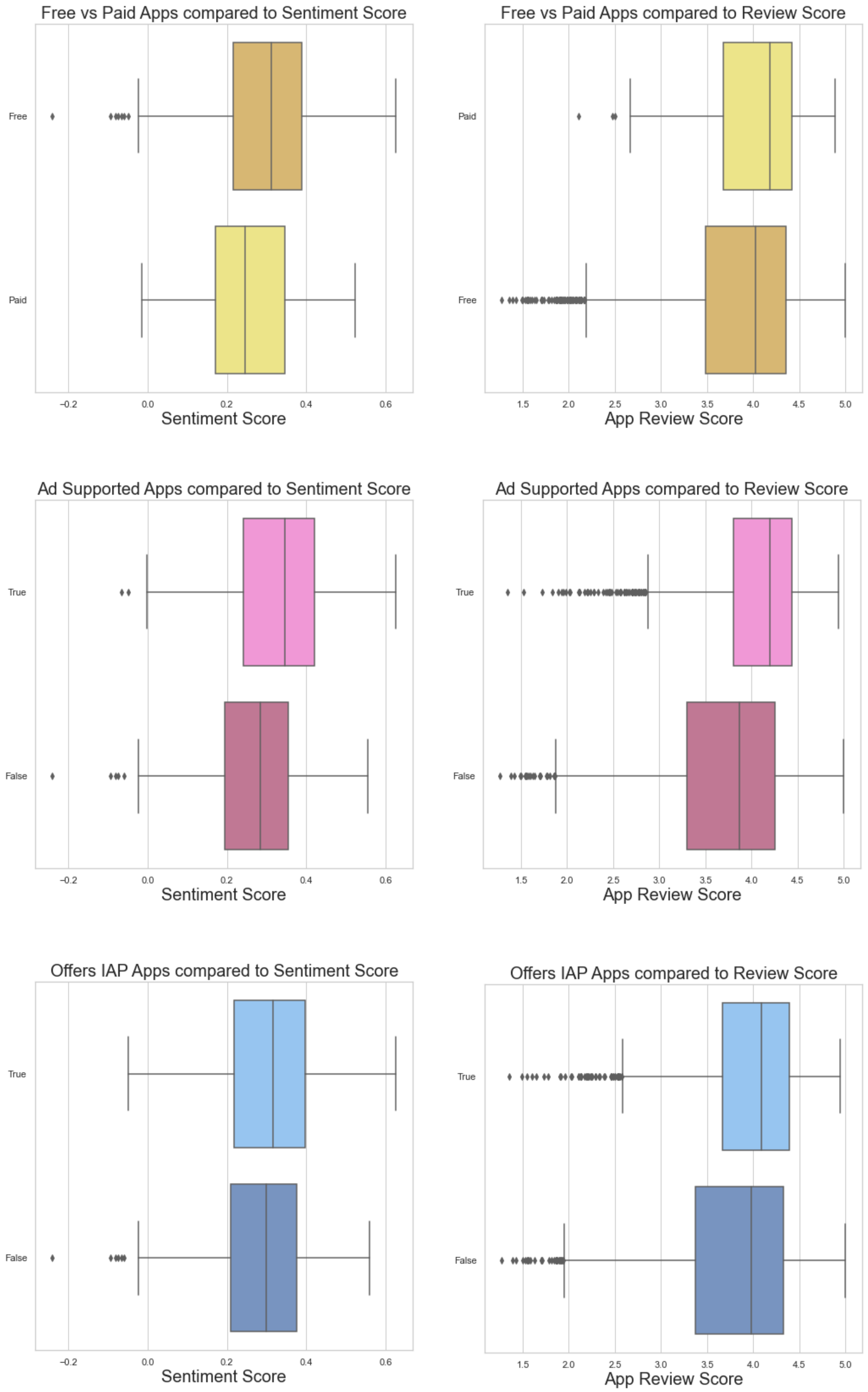


Figure 4.10: The median score and sentiment in Free apps vs Paid apps in the first 2 graphs, then Ad supported apps in the middle 2 graphs, then the apps with IAPs in the bottom two graphs.

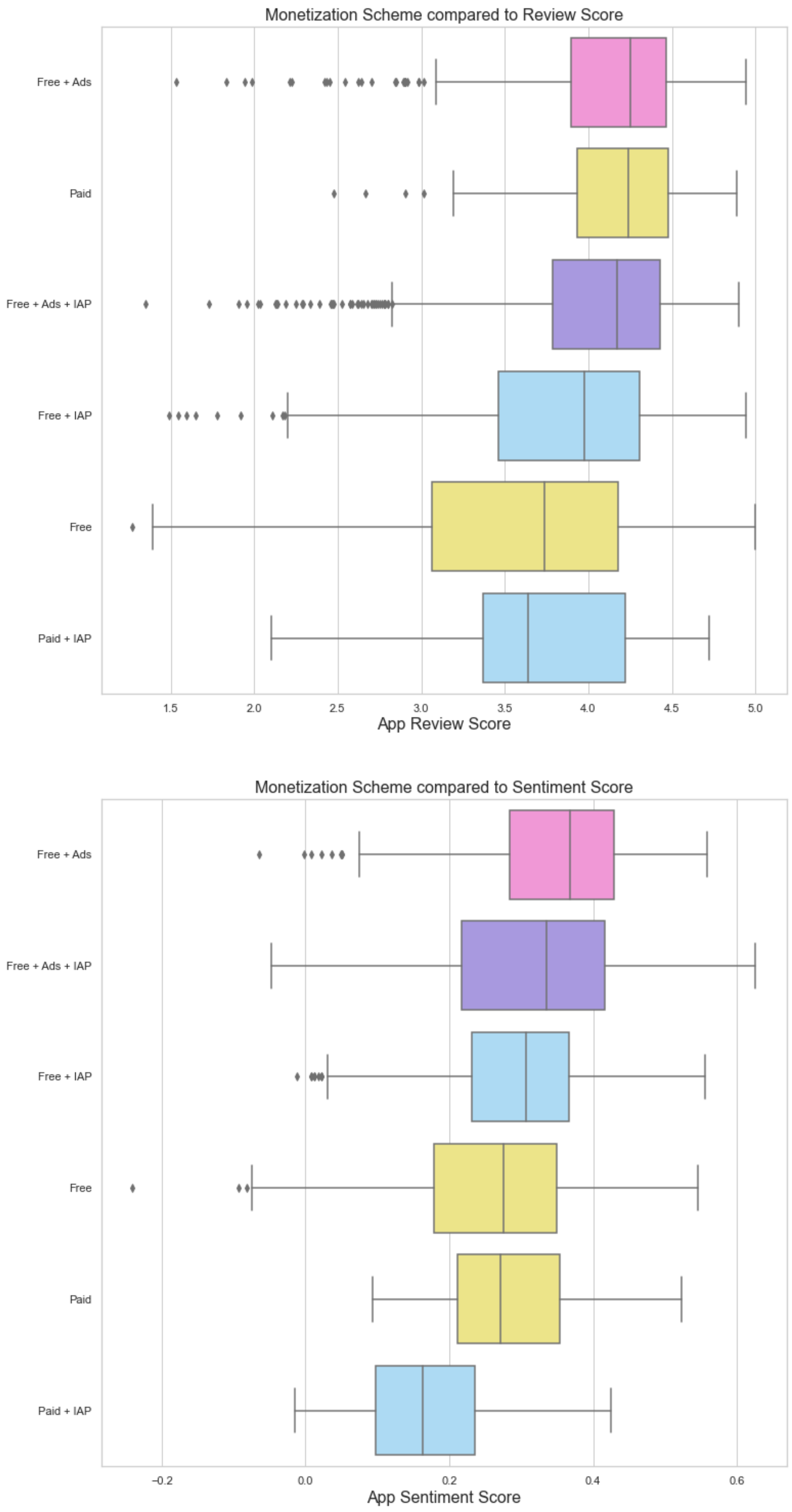


Figure 4.11: Median values of Score and Sentiment across all monetization categories.

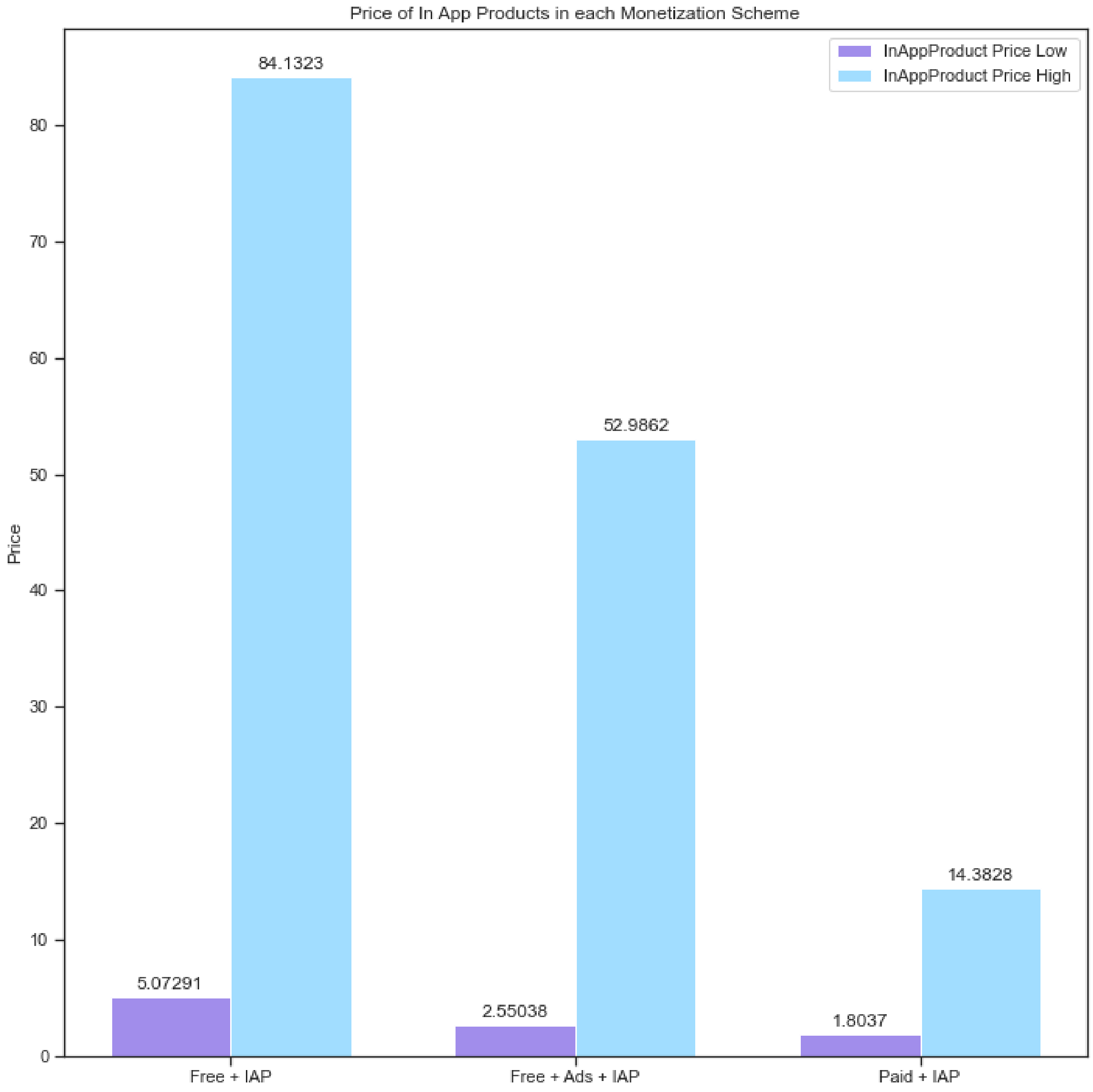


Figure 4.12: Mean Average High and Low In App Product Prices in monetization category.

Chapter 5: **Conclusions**

The Google Play Store is a digital marketplace where people publish and sell their android applications. With a growing user and creator-base, the platform is a vase ecosystem and digital terrain for businesses and hobbyists. Users can rate apps 1 to 5 stars, and upvote reviews they found helpful.

This report analyses the correlation and influence the sentiment and score of a review have with themselves and other factors. First describing how the review scraper I built harvested more than 2 million reviews spread across 24 different app genre. Discussing which data points from these reviews were of significance, and what parts of the data needed cleaning up. The research questions returned some interesting results:

1. Genres do have different scores of negative and positive sentiment. Higher star rating indicates higher sentiment in a review, but not in genre, and that apps with higher review counts tended to show greater correlation with their sentiment and app score. Outlier genres like Games having significantly lower sentiment relative to their app score, or Maps and Navigation having higher sentiment relative to their app score, as well as genres like Dating with both low score and sentiment. Users like and agree more with low sentiment, negative reviews, with user-bases in more negative genres like Dating, agreeing far more with its negative reviews (4 times more likes on its negative reviews to its positive reviews).
2. The length of a review varying widely in genre, with Health and Fitness having the most opinionated reviews. Reviews that were longer were also more likely to receive a reply from the app creator, and had a greater chance to be liked by other users (as long as the review was under 100 words in length). Bad reviews tended to complain about the apps monetization scheme.
3. Apps scores and sentiment are influenced by the type of monetization on the app. Free with Ads dominating in both App Score and sentiment, while Paid with IAP lowest in both app score and sentiment. The range of IAP being significantly lower in paid apps than in Free apps and Free apps with Ads.

The main part of this project that I would do differently (do better), would be change how the sentiment for the reviews are calculated. The python library I used for this part of the project, was not the most accurate for some of the reviews. I would look into finding or making a better sentiment calculator. It would also be more ideal if the dataset was bigger. Sampling only 3000 apps may not by enough to get a clear image of the overall sentiment and score on the Google Play Store, some genre had less than 100 apps in the dataset. All of these downfalls were taken into account when analysing the results of this project.

Acknowledgements

Thank you Barry Smyth for providing a wonderful sample project and presentation for myself and other students, as well as coaching us through the process of finishing a large Data science assignment.

Thank you Eoghan Cunningham for the great feedback every week, helping me grow the ambition of the project and answering all of my project queries.

Chapter 6: Code

The data collection is in this file: GooglePlayStoreAppReviews_dataCollection.ipynb

The first research question is in this file: GooglePlayStoreAppReviews_RQ1.ipynb

The second research question is in this file: GooglePlayStoreAppReviews_RQ2.ipynb

The third research question is in this file: GooglePlayStoreAppReviews_RQ3.ipynb

Bibliography

1. Annual number of app downloads from the Google Play Store worldwide from 2016 to 2021. *Statista.* https://www.statista.com/statistics/734332/google-play-app-installs-per-year/ (2022).
2. Colgan, M. HOW IMPORTANT ARE MOBILE APP RATINGS REVIEWS? *Tapadoo.* https://tapadoo.com/mobile-app-ratings-reviews/ (2019).
3. App Store Play Store Category Ranking: Does It Still Matter? *Gummicube.* https://www.gummicube.com/blog/app-store-play-store-category-ranking-does-it-still-matter (2018).