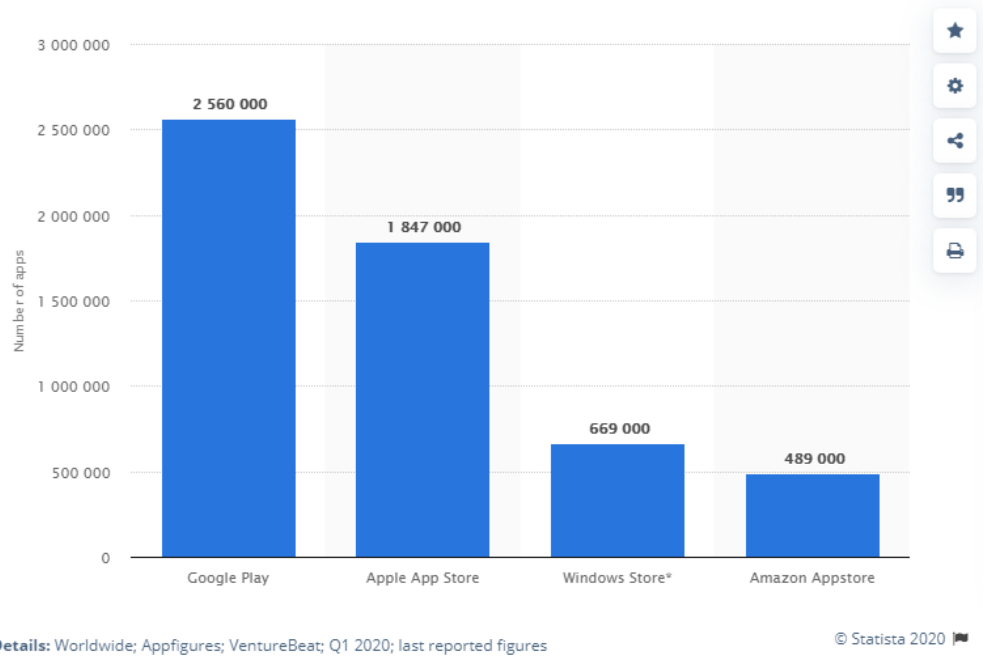
Subject: Mobile App Analytics

Background:

Creating an app is only half of the battle. Once the app is finished, you need to understand your users. What are they tapping, swiping, watching and buying? How often do they use your app and how long do they stay in the app? The best way of knowing that is using a mobile app analytics platform. In the two of the leading app stores as of 1st quarter 2020, there are more than 4 million apps and it is becoming increasingly important to track the habits of the users and their behavior. Mobile app analytics collect and present the data with the insight in all the platforms. They help us achieve our goals.

Number of apps available in leading app stores as of 1st quarter 2020



Details: Worldwide; Appfigures; VentureBeat; Q1 2020; last reported figures

Today, an average user in some countries has over 100 apps on their phone. There are more attributed installs than ever. Attributed installs are growing at a rate of 39%, while non-attributed installs are growing at a slower pace. That means more marketers are tracking the user journey to their app or through their app.

Inspiration:

According to a Harvard Business Review study, researchers found that "businesses that integrate multiple sources of customer and marketing data significantly outperform other companies..". They also had dramatically higher total shareholder returns. When it comes to creating new mobile app, it's an overwhelming experience to decide on category, price, size and to predict its success as there are so many apps on the market and more are being added every day!

Problem Statement:

Derive insights on how users discover your app or how users search the App Store.

How to track app store impressions, user engagement, as well as segmentation of users.

How sales and trends section allow you to understand which of your apps or in-app subscriptions are the most popular.

Few Interesting Questions to address through visualization…

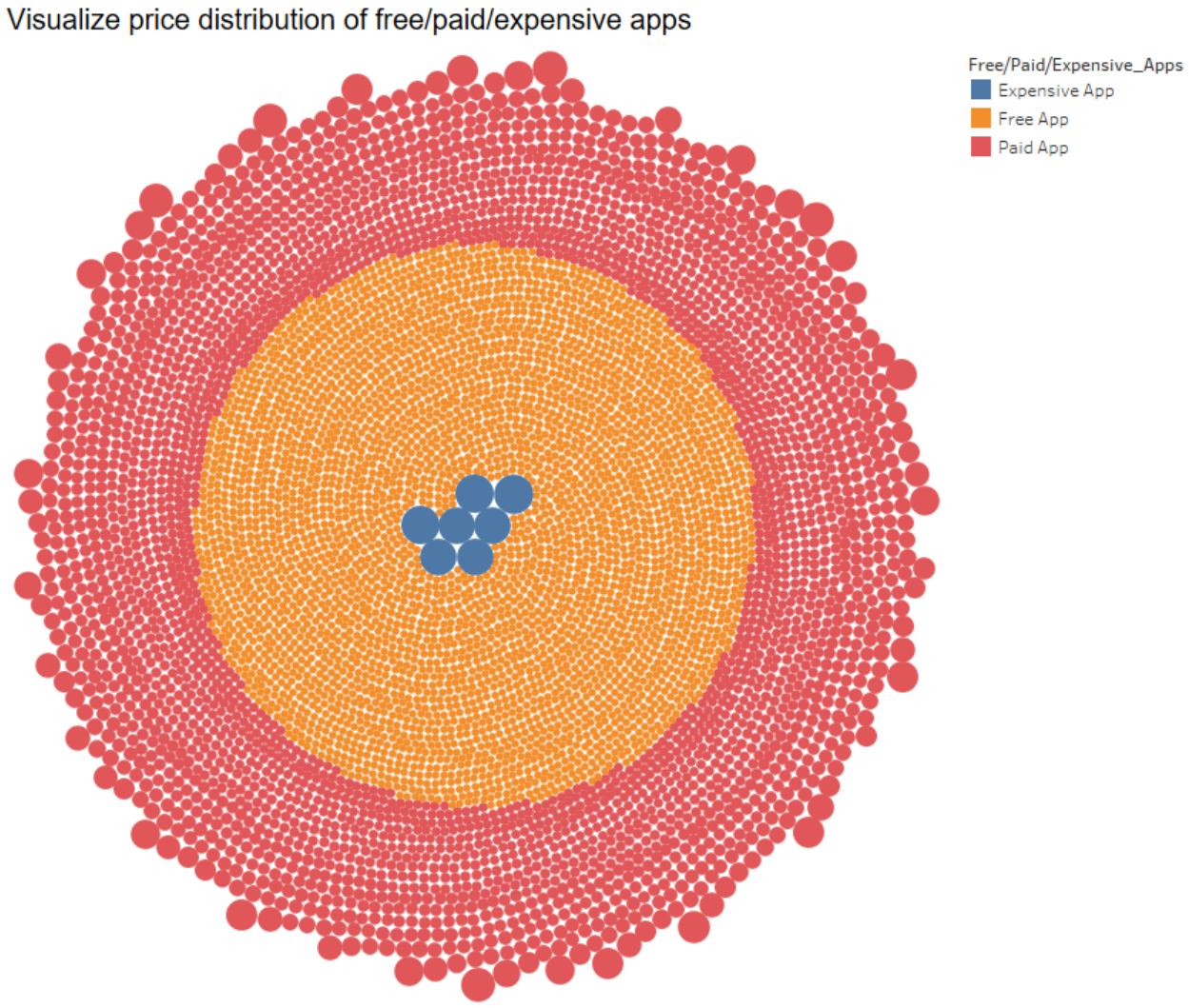
1. How do you visualize price distribution of paid apps ?
2. How does the price distribution get affected by category ?
3. What about paid apps Vs Free apps ?
4. Are paid apps good enough ?
5. As the size of the app increases do they get pricier ?
6. How are the apps distributed category wise ? can we split by paid category ?
7. How to predict success of App?

Dataset: The data set comprises of information on 7200 apps on App store with following imp details.
Dimension of the data set;
7197 rows and 16 columns

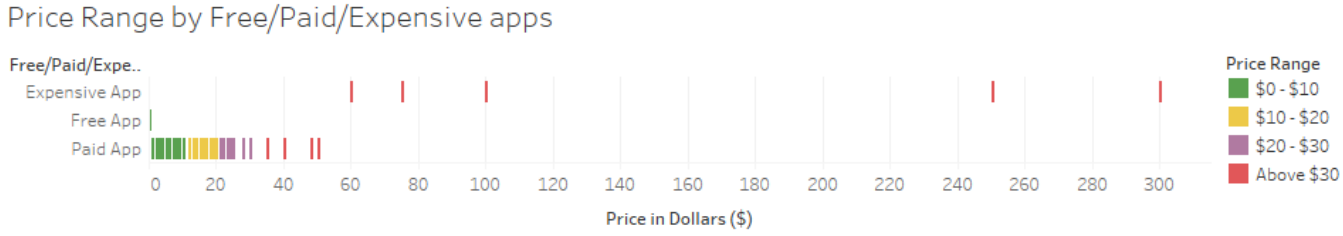
appleStore.csv
"id" : App ID
"track_name": App Name
"size_bytes": Size (in Bytes)
"currency": Currency Type
"price": Price amount
"ratingcounttot": User Rating counts (for all version)
"ratingcountver": User Rating counts (for current version)
"user_rating" : Average User Rating value (for all version)
"userratingver": Average User Rating value (for current version)
"ver" : Latest version code
"cont_rating": Content Rating
"prime_genre": Primary Genre
"sup_devices.num": Number of supporting devices
"ipadSc_urls.num": Number of screenshots showed for display
"lang.num": Number of supported languages
"vpp_lic": Vpp Device Based Licensing Enabled

appleStore_description.csv
id : App ID
track_name: Application name
size_bytes: Memory size (in Bytes)
app_desc: Application description

I. How do you visualize price distribution of paid apps ?

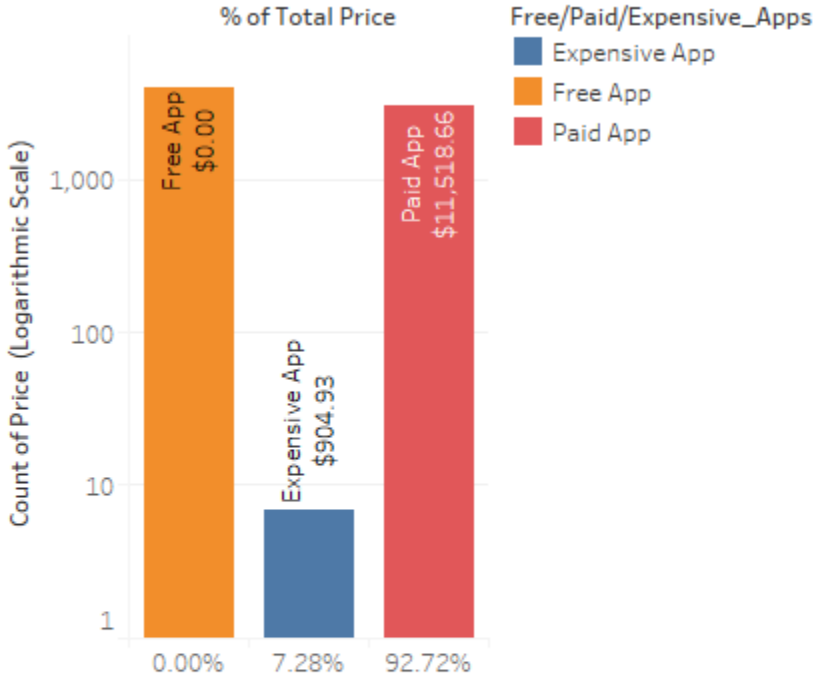


Free/Paid/Expensive_Apps and track name (appleStore!description). Color shows details about Free/Paid/Expensive_Apps . Size shows details about sum of Price. The marks are labeled by Free/Paid/Expensive_Apps and track name (appleStore!description). The view is filtered on Free/Paid/Expensive_Apps , which keeps Expensive App, Free App and Paid App. Percents are based on the whole table.



Price for each Free/Paid/Expensive_Apps . Color shows details about Price Range.

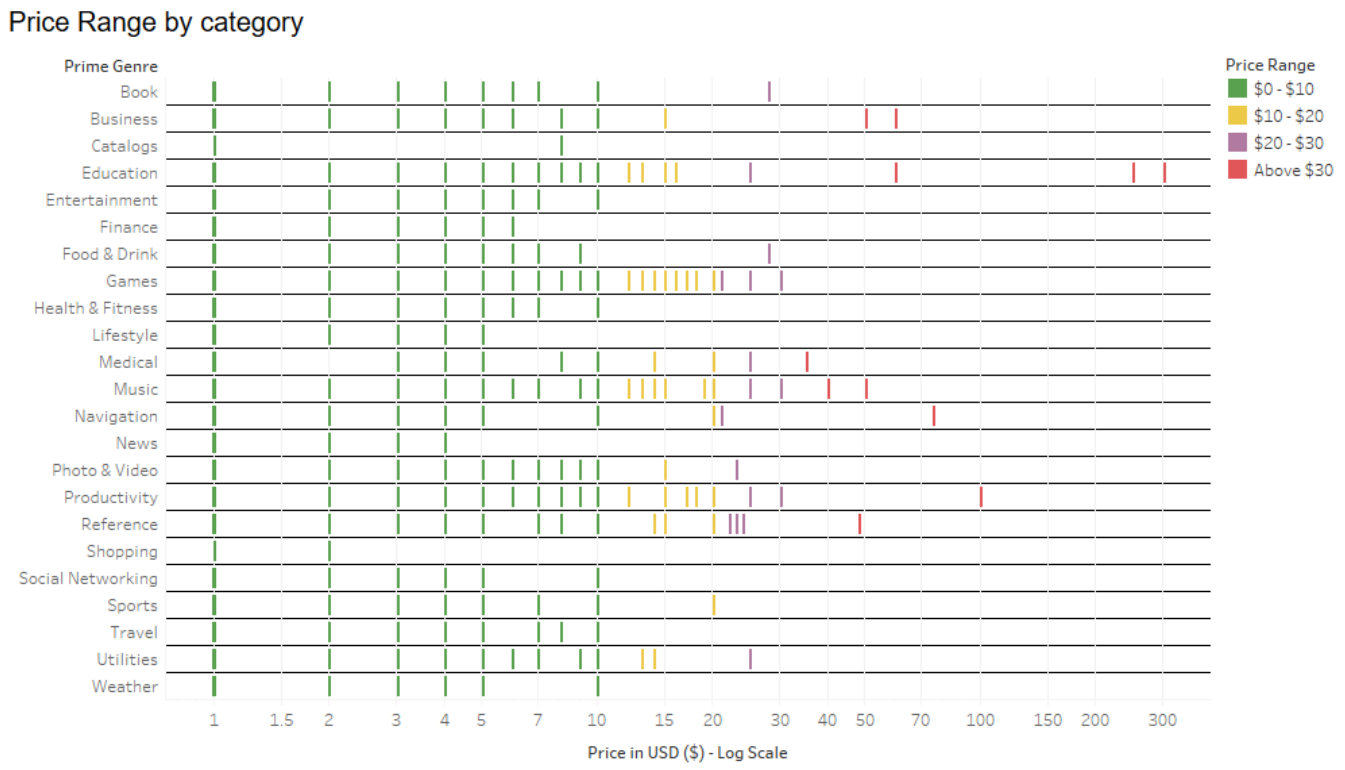
Price distribution of apps (Histogram)

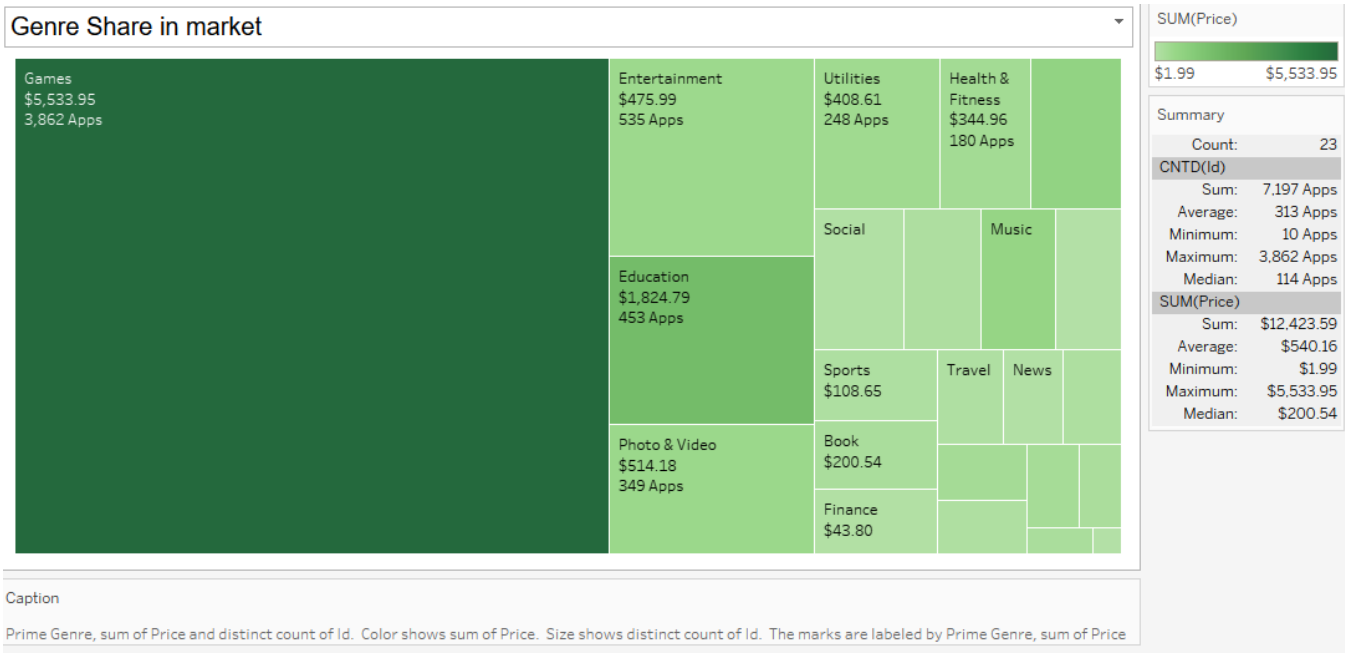


Count of Price for each % of Total Price. Color shows details about Free/Paid/Expensive_Apps . The marks are labeled by Free/Paid/Expensive_Apps and sum of Price. Details are shown for Free/Paid/Expensive_Apps . The view is filtered on Free/Paid/Expensive_Apps , which keeps Expensive App, Free App and Paid App.

- Count of paid apps is exponentially decreasing as the price increases.
- Count of Free apps ($0) is nearly same as Paid apps ($0-$30).
- Paid Apps in the range of $0-30$ price holds 93% of market share with $11.5k
- Expensive Apps in the range of $30-$300 range holds just 7% of market share with $900.
- There are not many apps in the app store that are above $30, so it's important to keep the price of the app under $30.

2. How does the price distribution get affected by category?

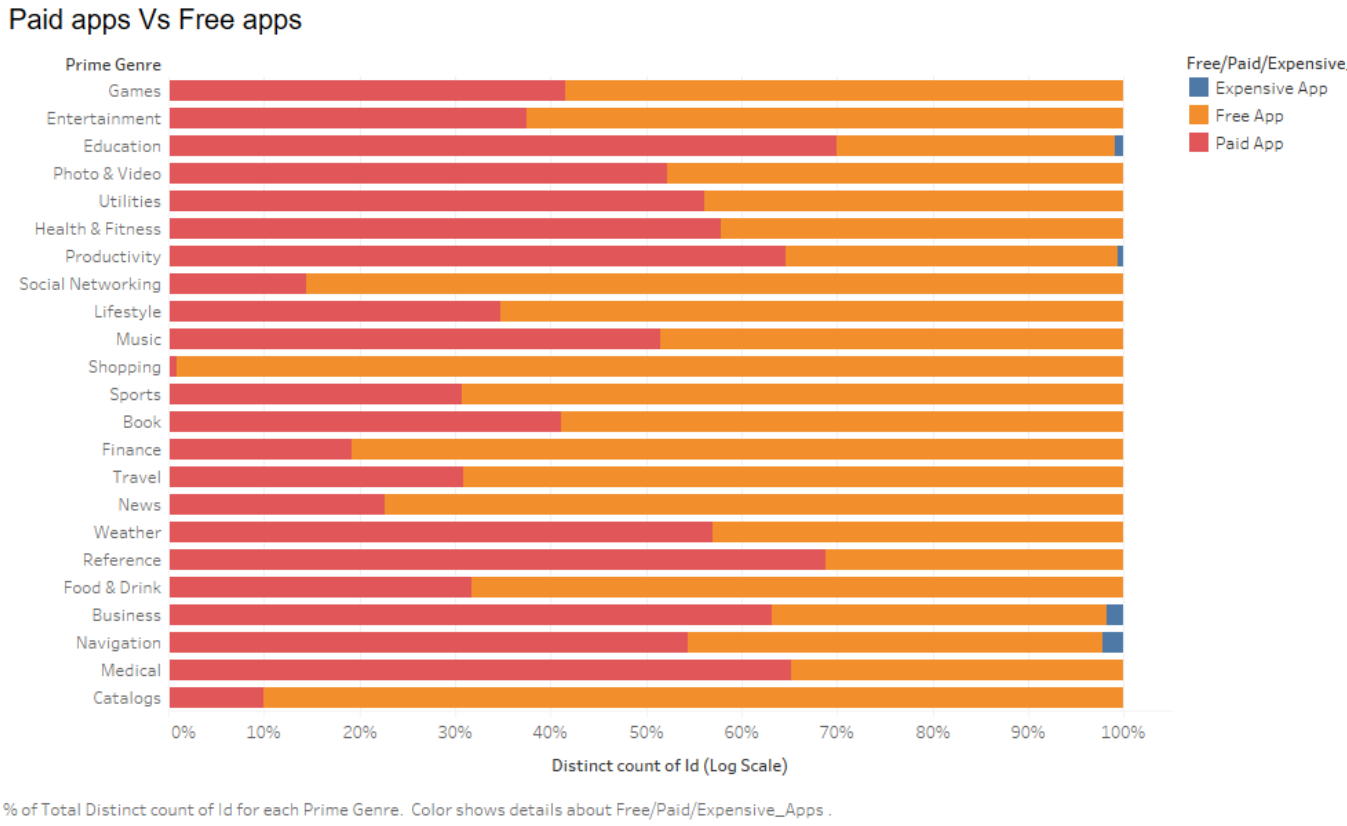




- Games Genre holds highest share in Market in terms of Count of Apps and Value of Apps

- There are no apps in **Catalogues, Finance, Lifestyle, Entertainment, New, Shopping and Travel** Genre that are beyond $0 - $10 range. So, it's important to keep the the price in this range if you are interested to develop app in these Categories.
- There are no apps in **Books, Education, Food & Drink, Games, Medical, Music, Navigation, Photo & Video, Utilities** Genre that are beyond $0-$30 price range.
- **Business, Education, Productivity and Navigation are** Genres that have scope for Expensive apps ($30-Above)

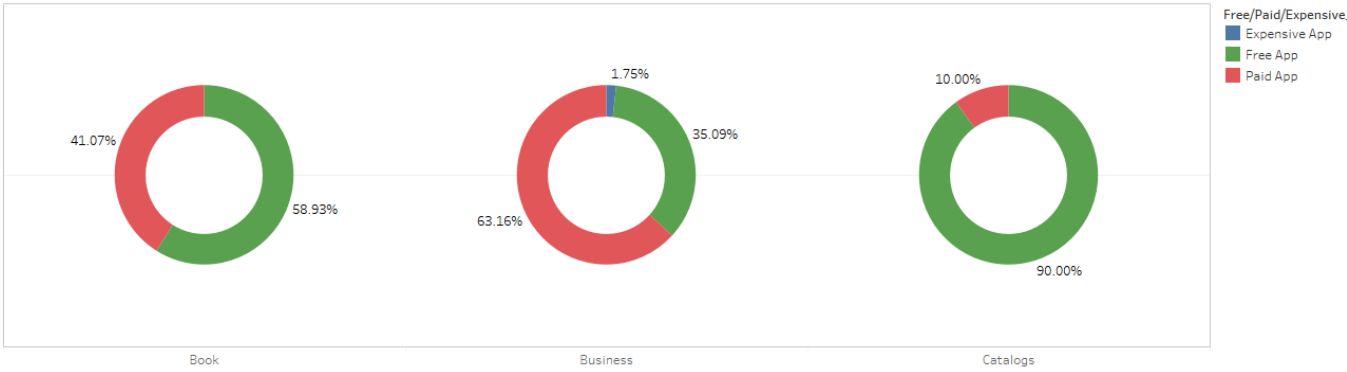
3. What about paid apps Vs Free apps ?



- **Education, Medical, Reference, Productivity, Business, Health & Fitness Categories hold** relatively highest Paid app market share.
- **Shopping, Catalogues, Social Networking and Finance** Genres hold the lowest Paid app market share.

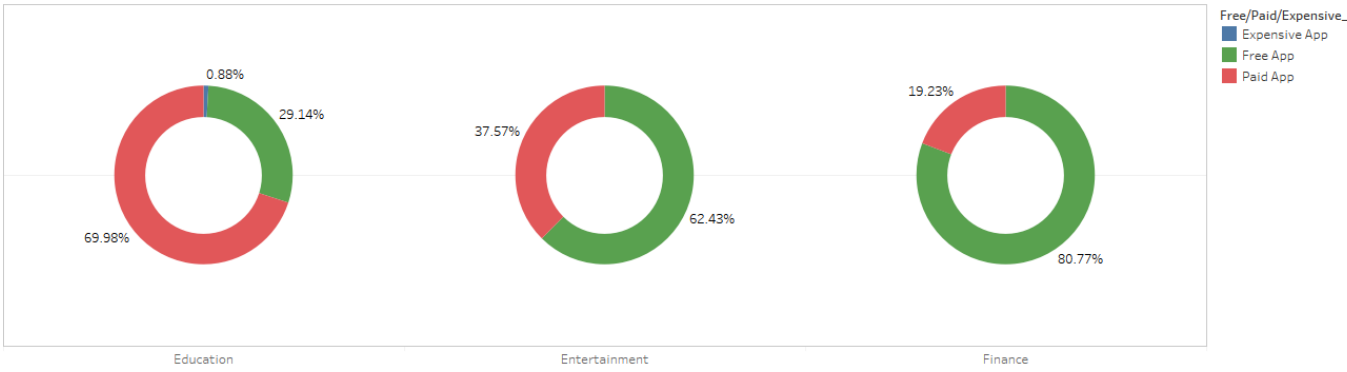
We can see the Individual Genre market share of Paid and Free Apps as below.

Paid apps Vs Free apps by Genre



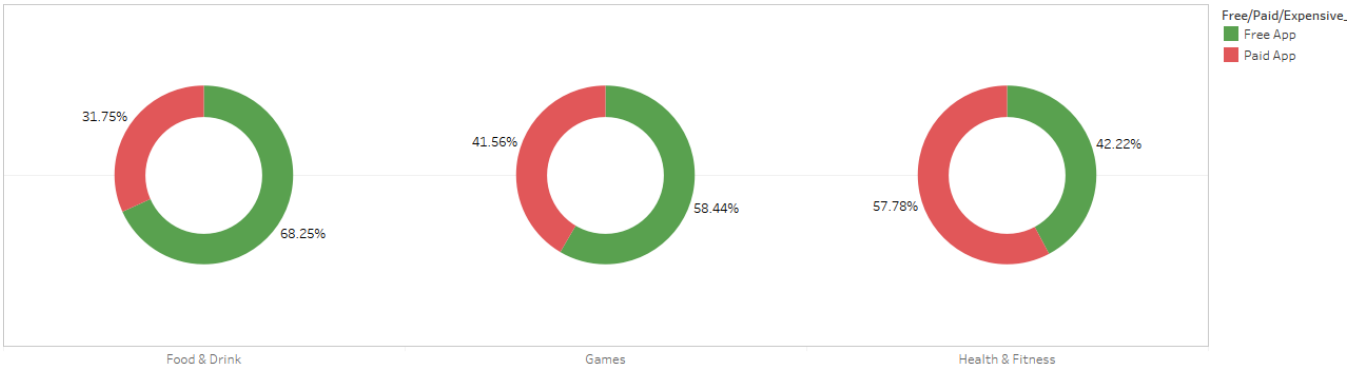
Minimum of 1 and minimum of 1 for each Prime Genre. For pane Minimum of 1: Color shows details about Free/Paid/Expensive_Apps . The marks are labeled by % of Total Distinct count of Id. The view is filtered on Prime Genre, which keeps Book, Business and Catalogs.

Paid apps Vs Free apps by Genre



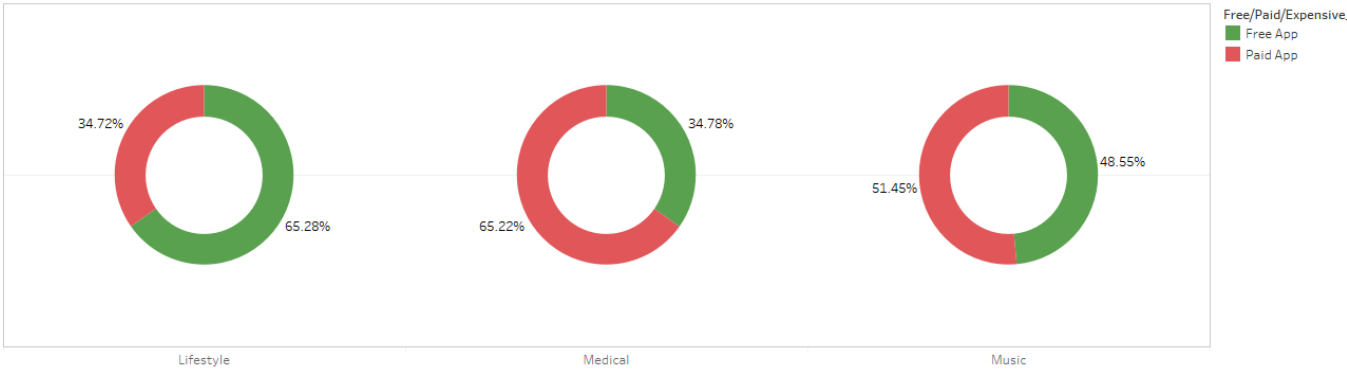
Minimum of 1 and minimum of 1 for each Prime Genre. For pane Minimum of 1: Color shows details about Free/Paid/Expensive_Apps . The marks are labeled by % of Total Distinct count of Id. The view is filtered on Prime Genre, which keeps Education, Entertainment and Finance.

Paid apps Vs Free apps by Genre



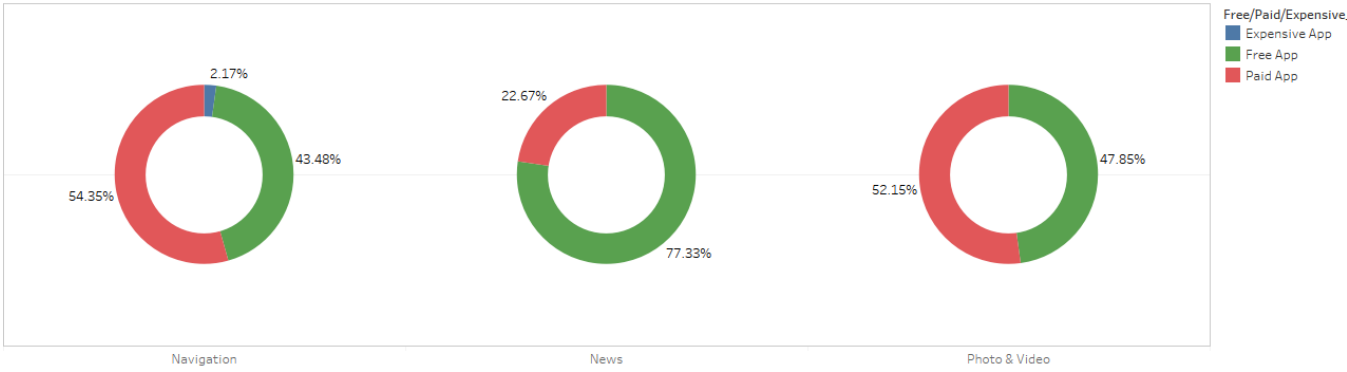
Minimum of 1 and minimum of 1 for each Prime Genre. For pane Minimum of 1: Color shows details about Free/Paid/Expensive_Apps . The marks are labeled by % of Total Distinct count of Id. The view is filtered on Prime Genre, which keeps Food & Drink, Games and Health & Fitness.

Paid apps Vs Free apps by Genre



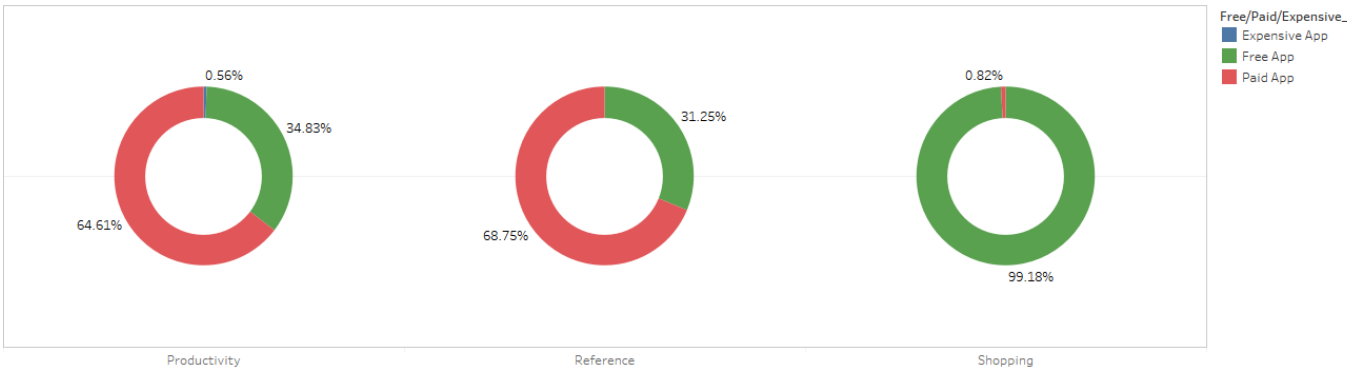
Minimum of 1 and minimum of 1 for each Prime Genre. For pane Minimum of 1: Color shows details about Free/Paid/Expensive_Apps . The marks are labeled by % of Total Distinct count of Id. The view is filtered on Prime Genre, which keeps Lifestyle, Medical and Music.

Paid apps Vs Free apps by Genre



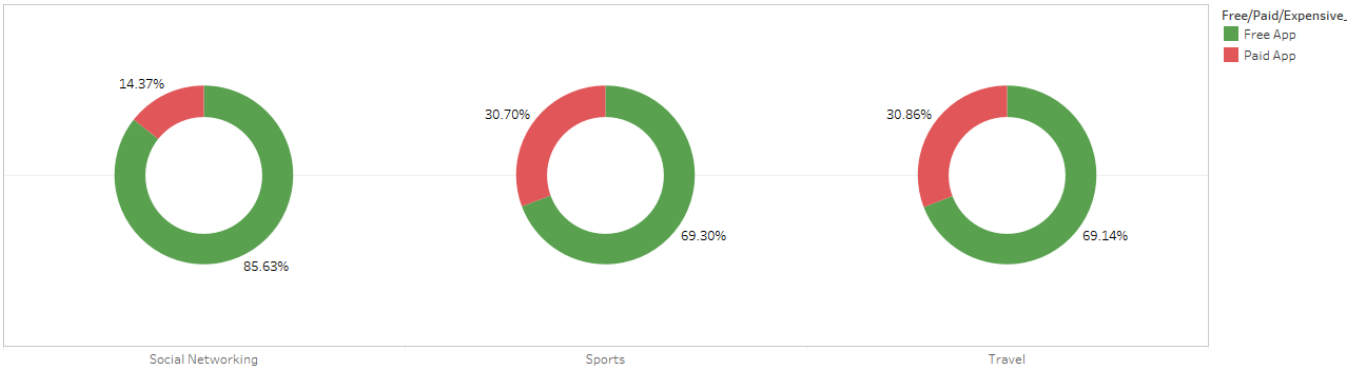
Minimum of 1 and minimum of 1 for each Prime Genre. For pane Minimum of 1: Color shows details about Free/Paid/Expensive_Apps . The marks are labeled by % of Total Distinct count of Id. The view is filtered on Prime Genre, which keeps Navigation, News and Photo & Video.

Paid apps Vs Free apps by Genre



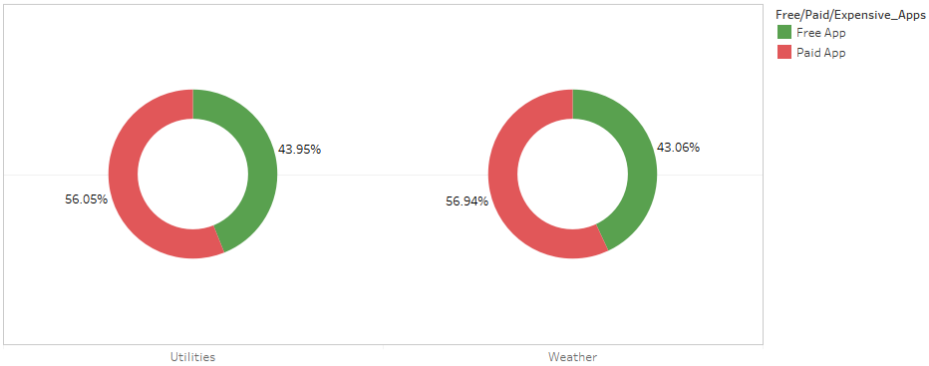
Minimum of 1 and minimum of 1 for each Prime Genre. For pane Minimum of 1: Color shows details about Free/Paid/Expensive_Apps . The marks are labeled by % of Total Distinct count of Id. The view is filtered on Prime Genre, which keeps Productivity, Reference and Shopping.

Paid apps Vs Free apps by Genre



Minimum of 1 and minimum of 1 for each Prime Genre. For pane Minimum of 1: Color shows details about Free/Paid/Expensive_Apps . The marks are labeled by % of Total Distinct count of Id. The view is filtered on Prime Genre, which keeps Social Networking, Sports and Travel.

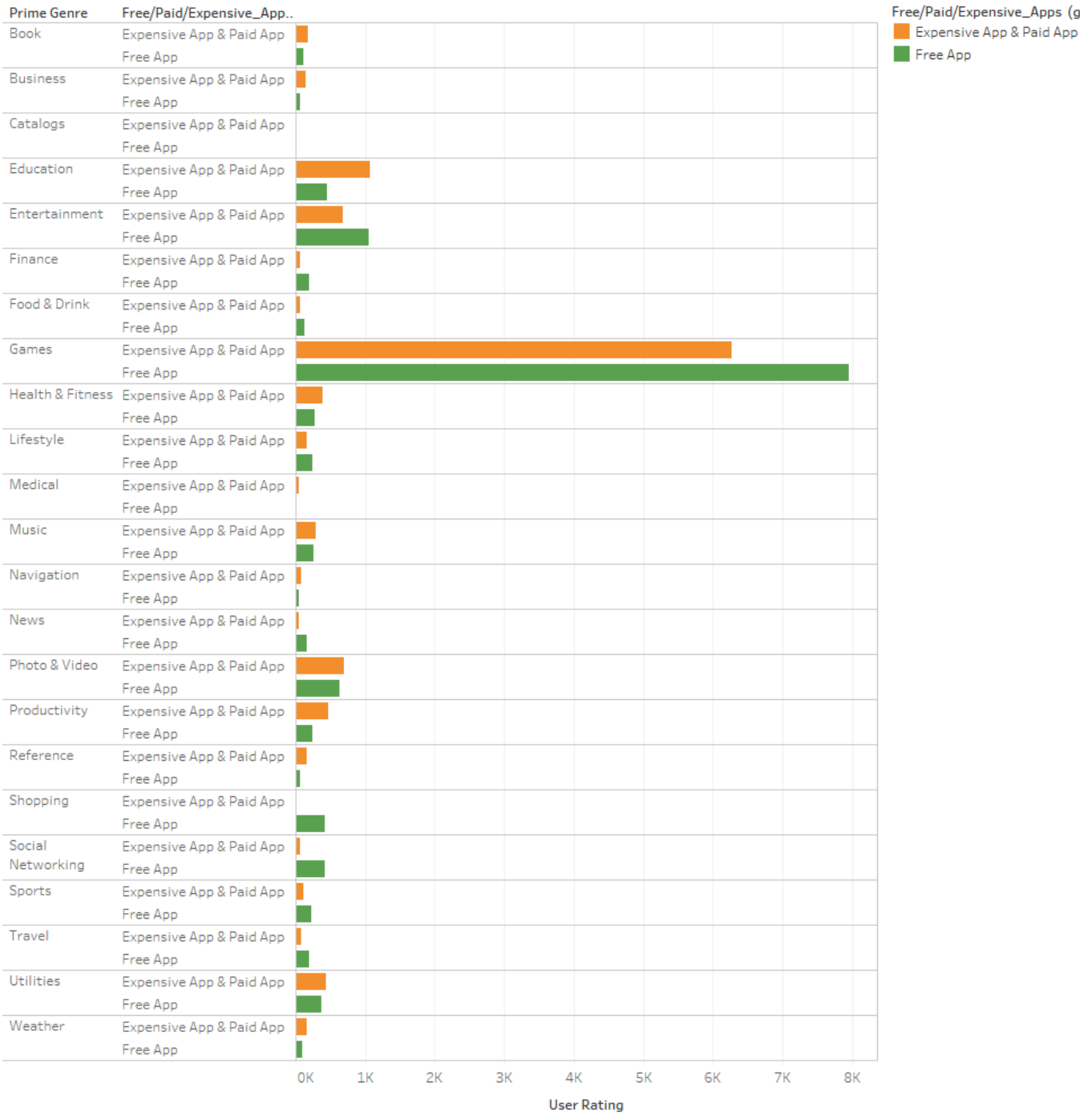
Paid apps Vs Free apps by Genre



Minimum of 1 and minimum of 1 for each Prime Genre. For pane Minimum of 1: Color shows details about Free/Paid/Expensive_Apps . The marks are labeled by % of Total Distinct count of Id. The view is filtered on Prime Genre, which keeps Utilities and Weather.

3. Are paid apps good enough?

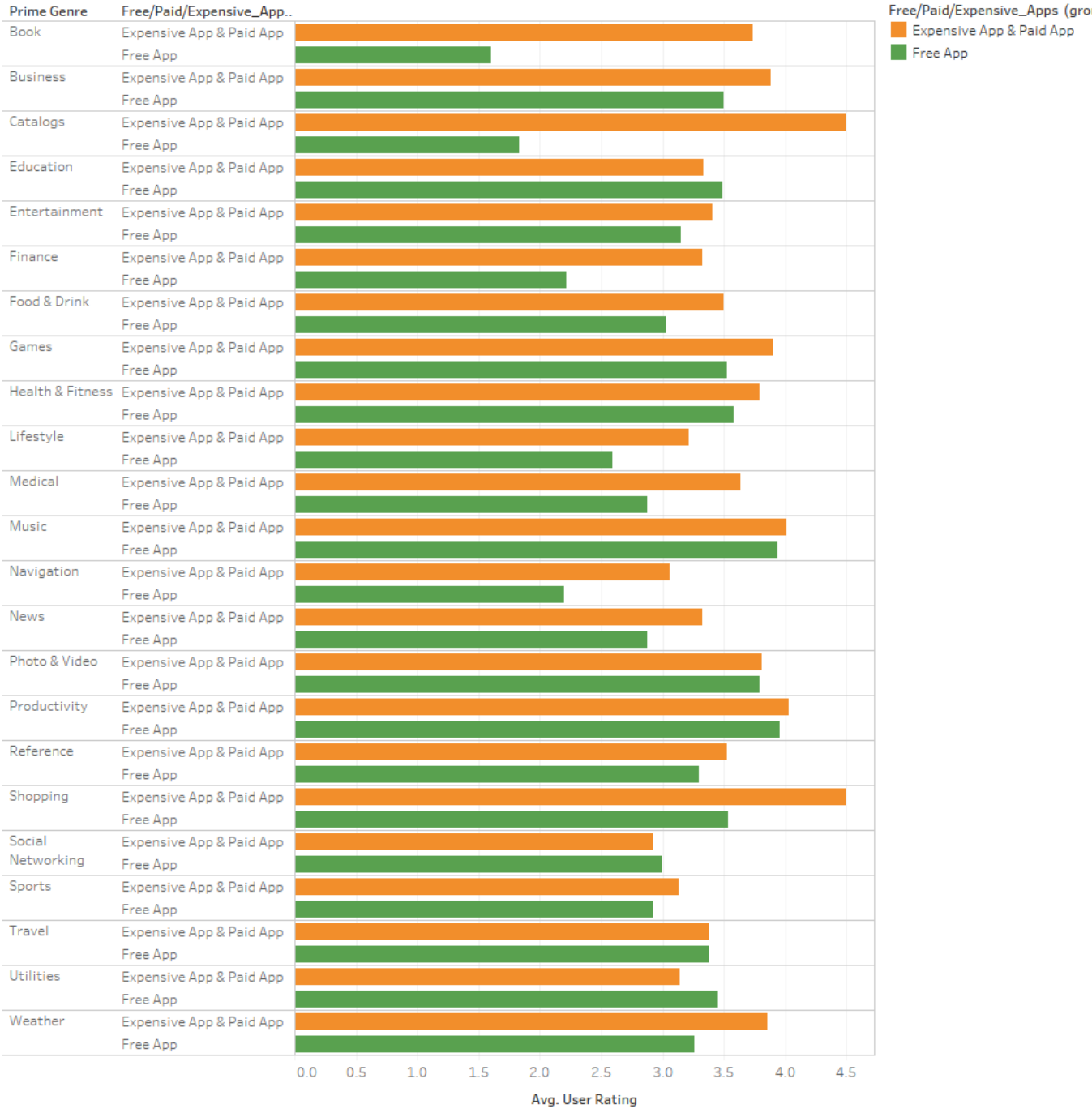
Q-4



Sum of User Rating for each Free/Paid/Expensive_Apps (group) broken down by Prime Genre. Color shows details about Free/Paid/Expensive_Apps (group).

- **Games** Genre has got the highest number Ratings amongst all Genres following **Entertainment, Education.**

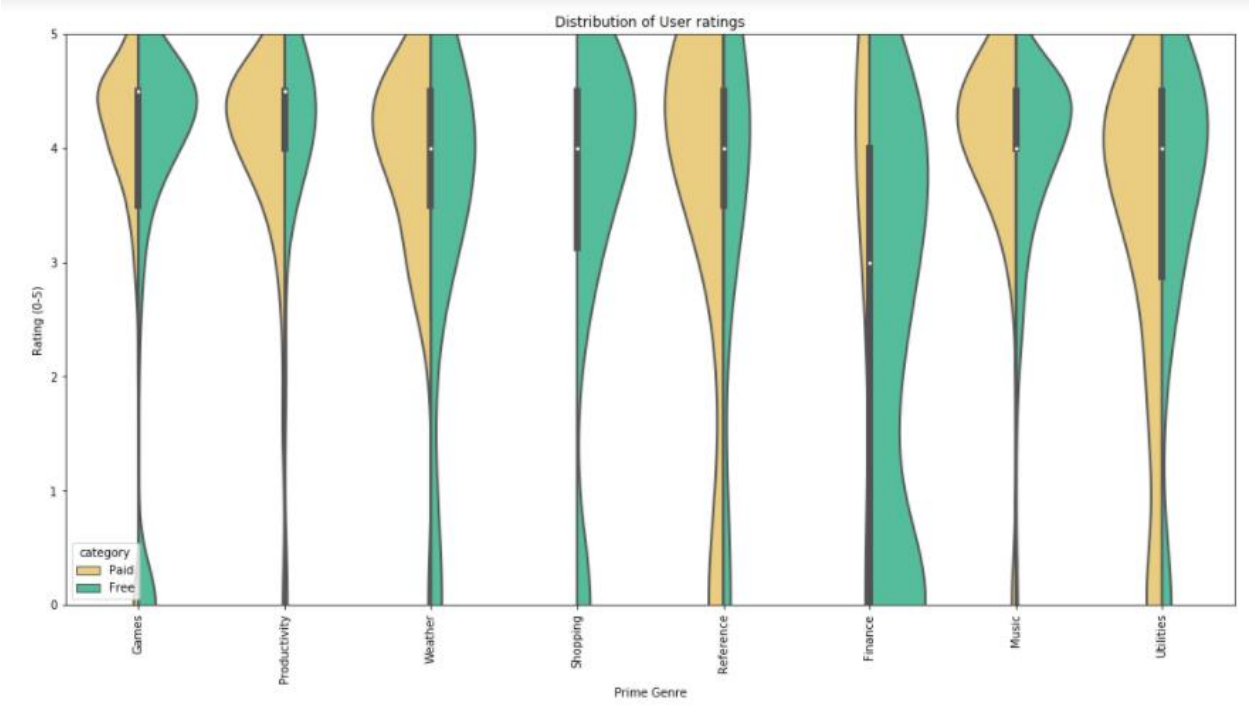
Average User rating per Genre

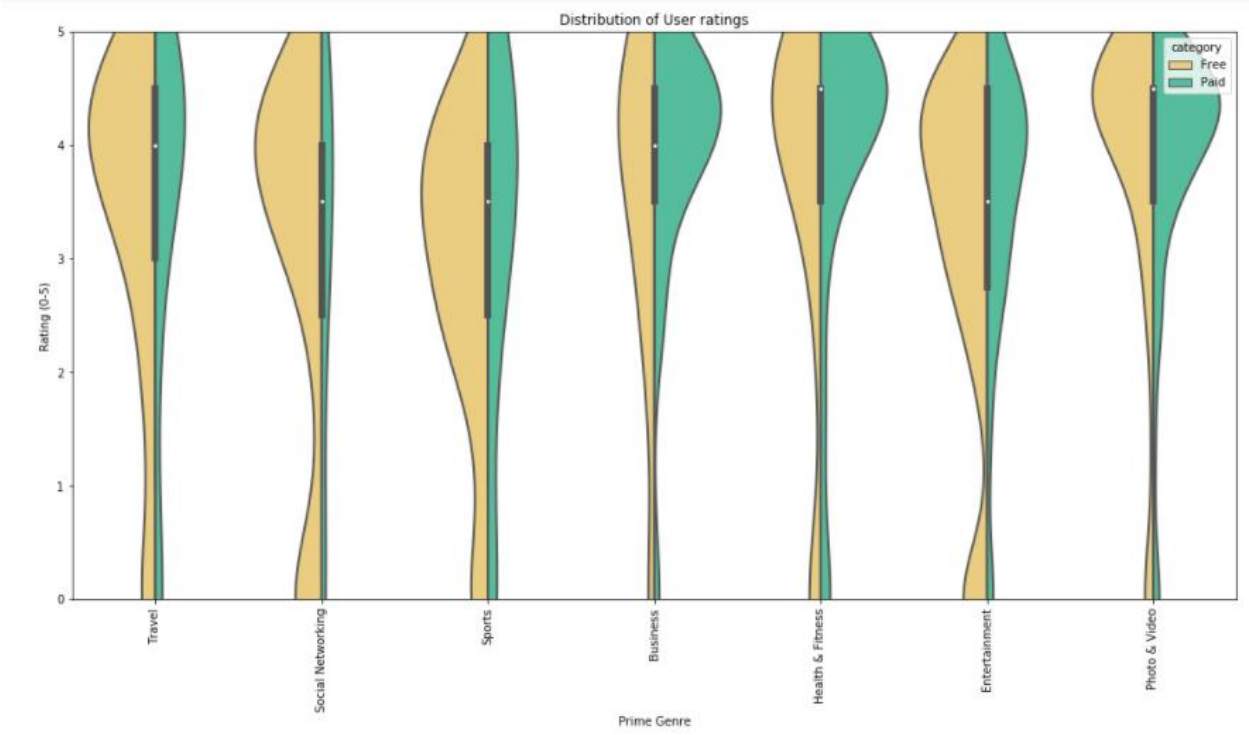


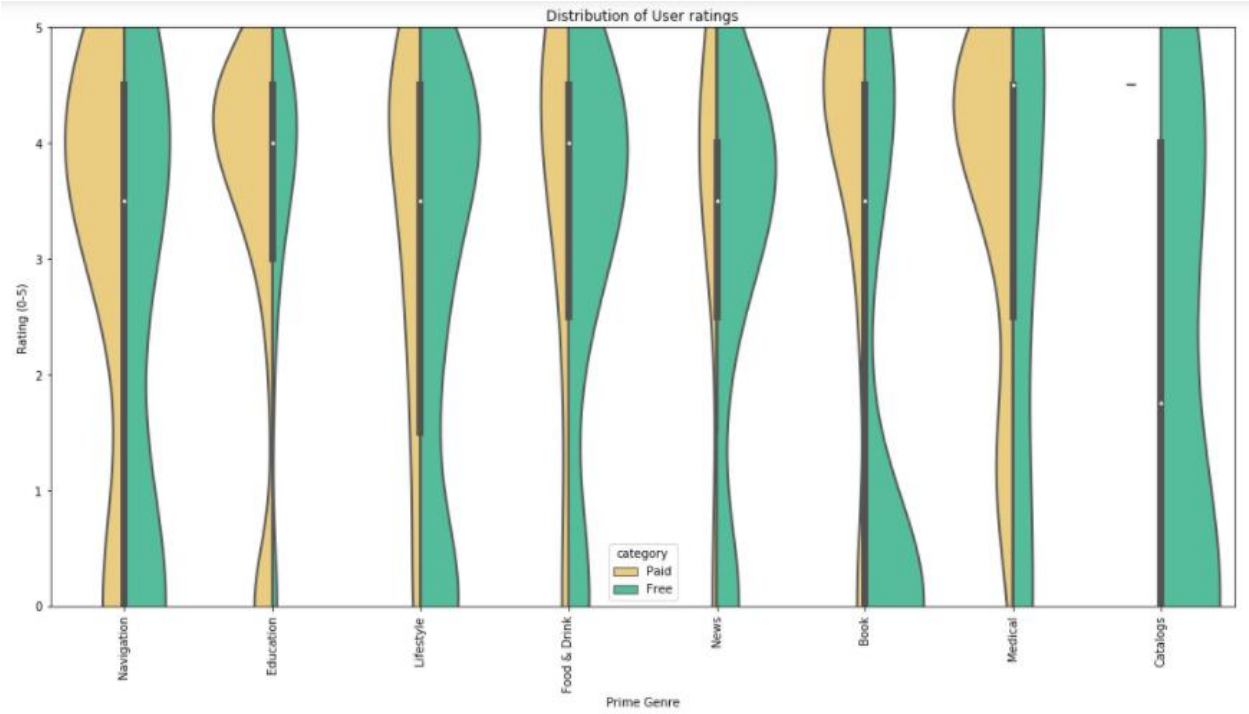
Average of User Rating for each Free/Paid/Expensive_Apps (group) broken down by Prime Genre. Color shows details about Free/Paid/Expensive_Apps (group).

- Catalogues, Books, Shopping are the three Genres where Paid Apps have got relatively higher ratings than Free Apps.
- Rest of the Genre's Free Apps have got equally Good Rating as Paid Apps.

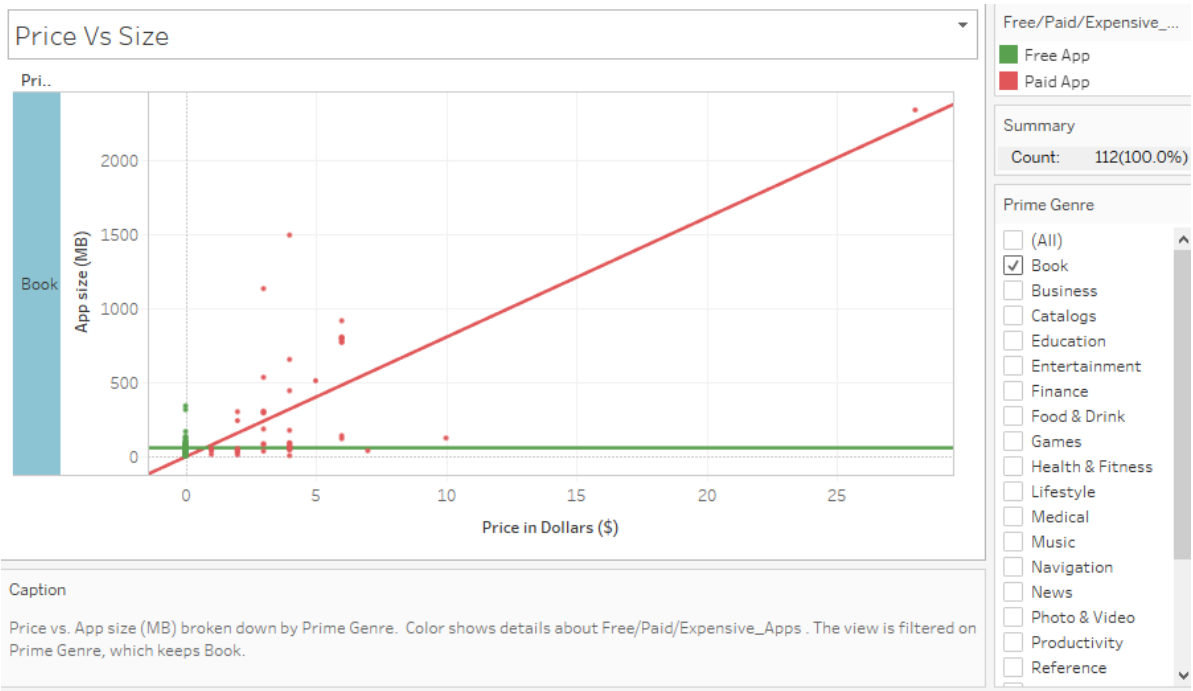
Let's see the distribution of User Ratings per Genre.



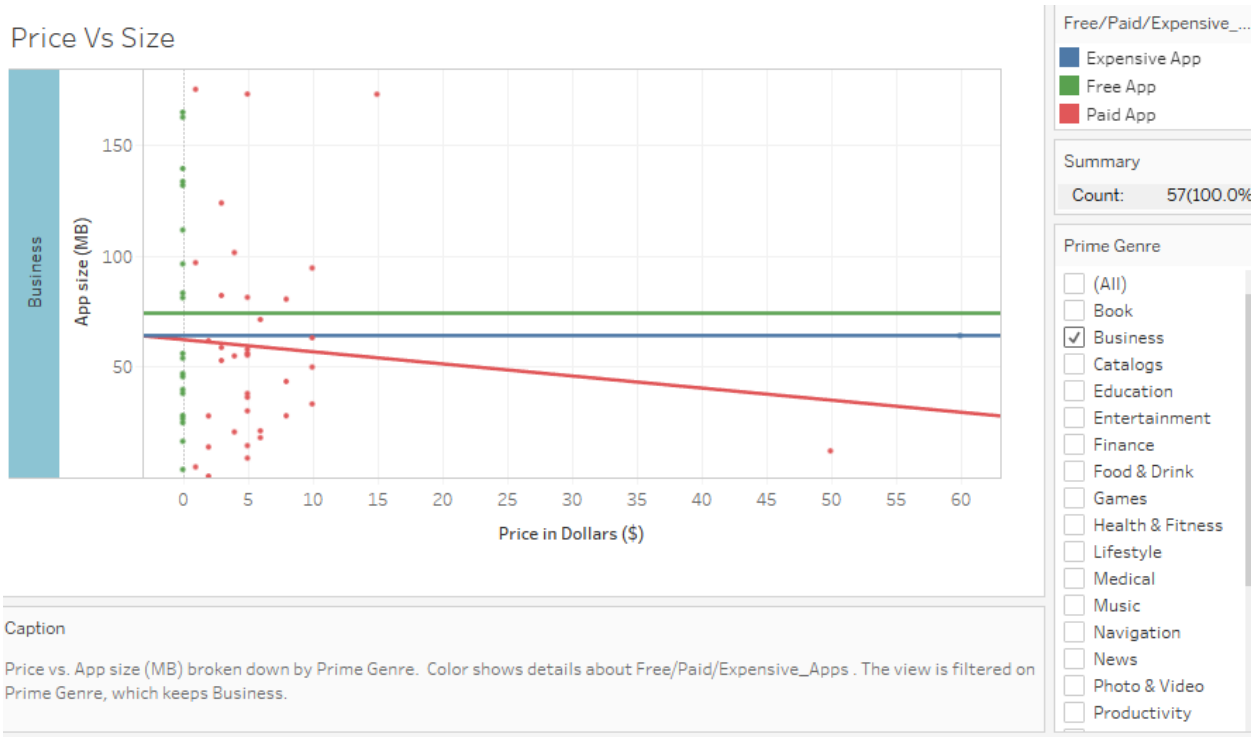




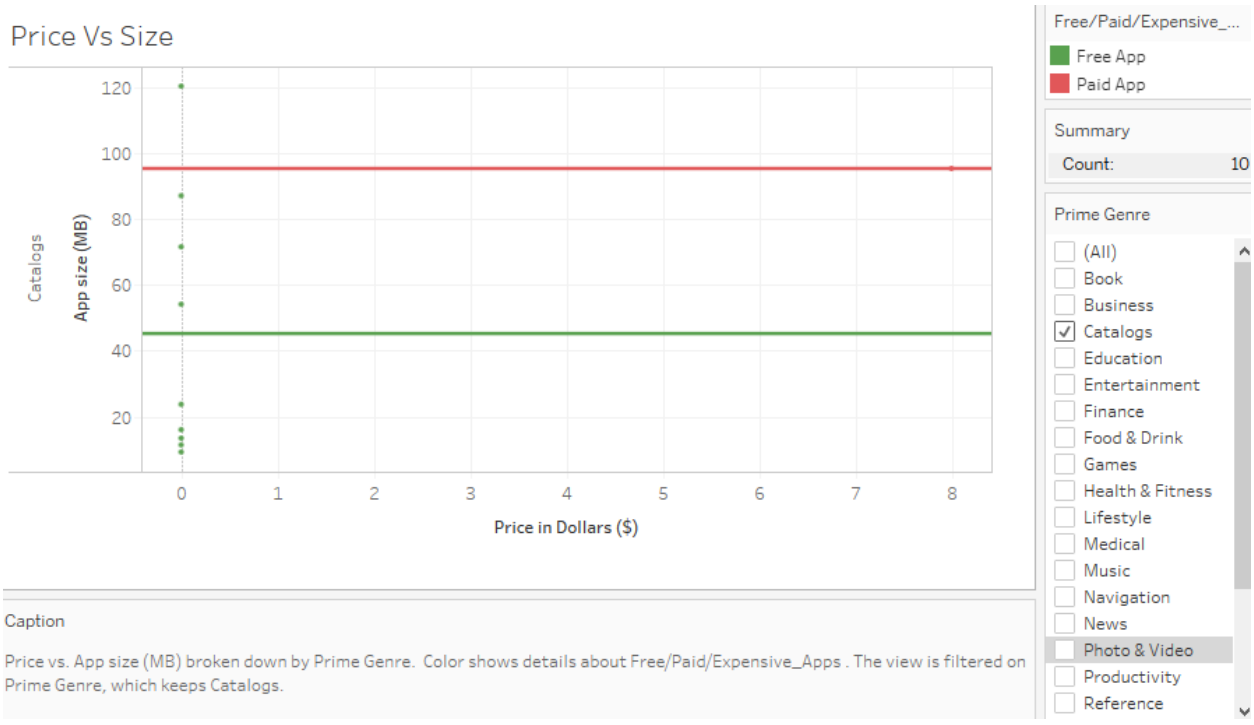
4. As the size of the app increases do they get pricier?



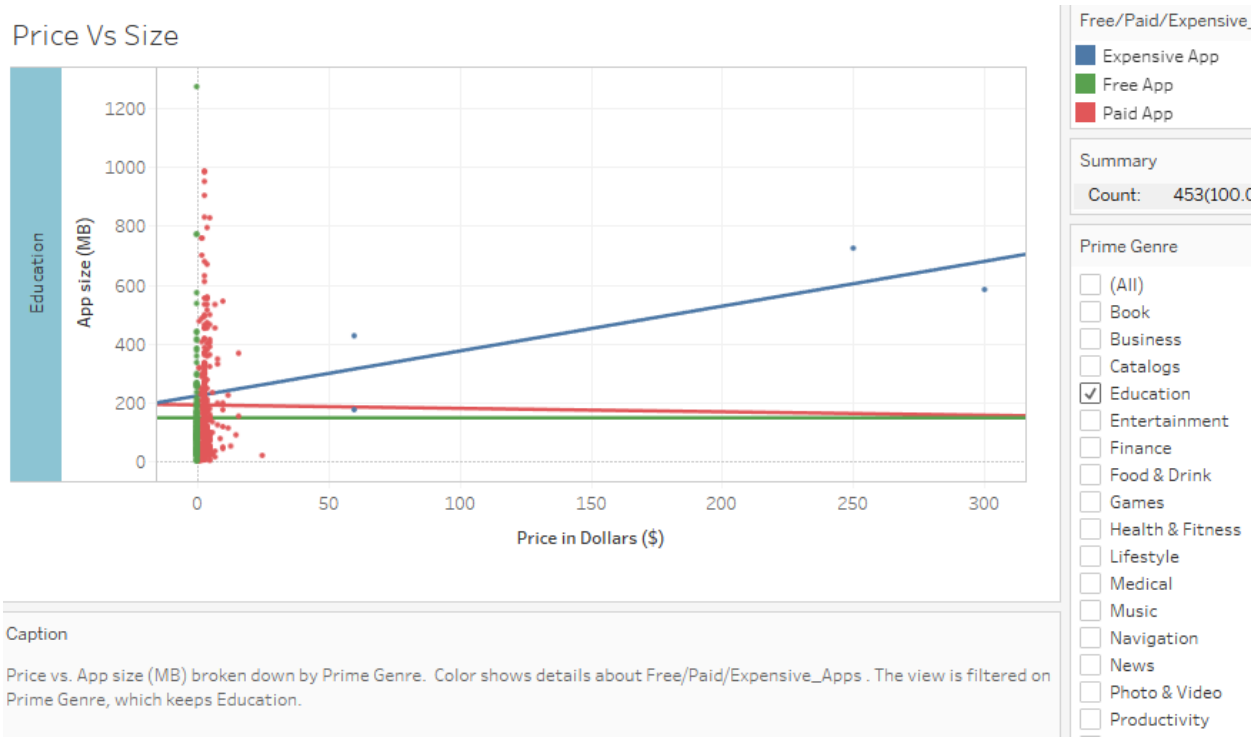
- Yes in Book Genre



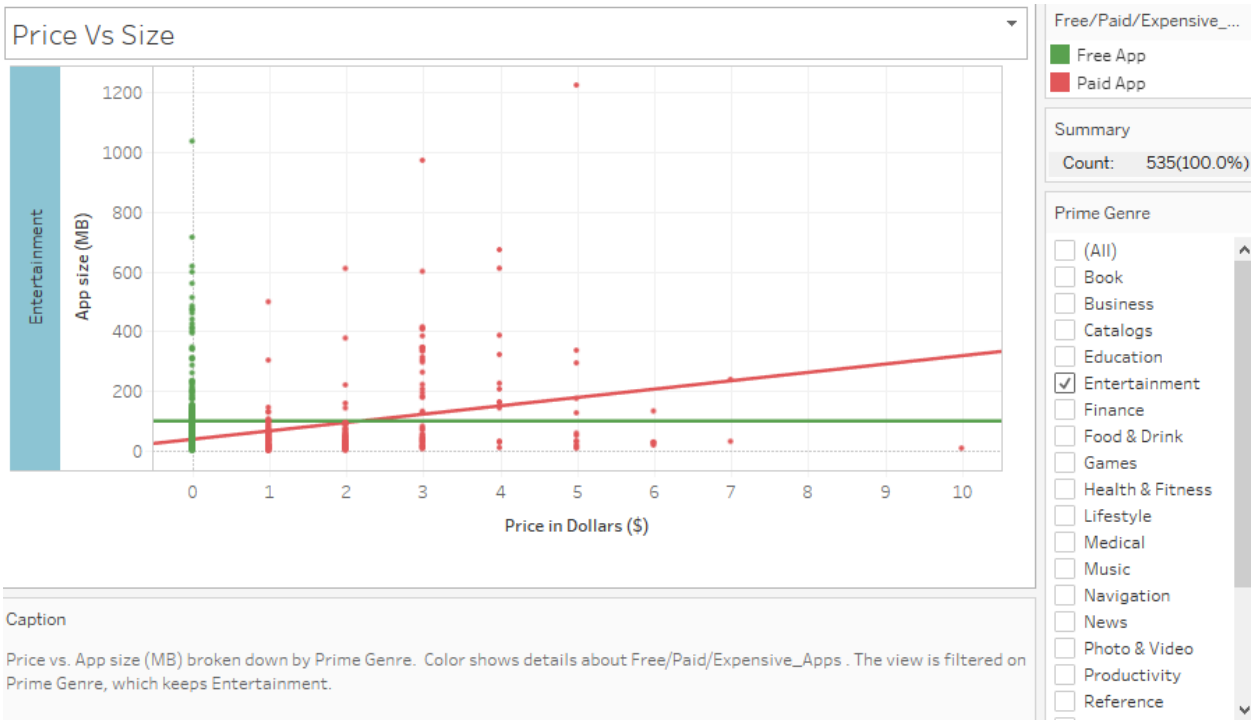
- No in Business Genre



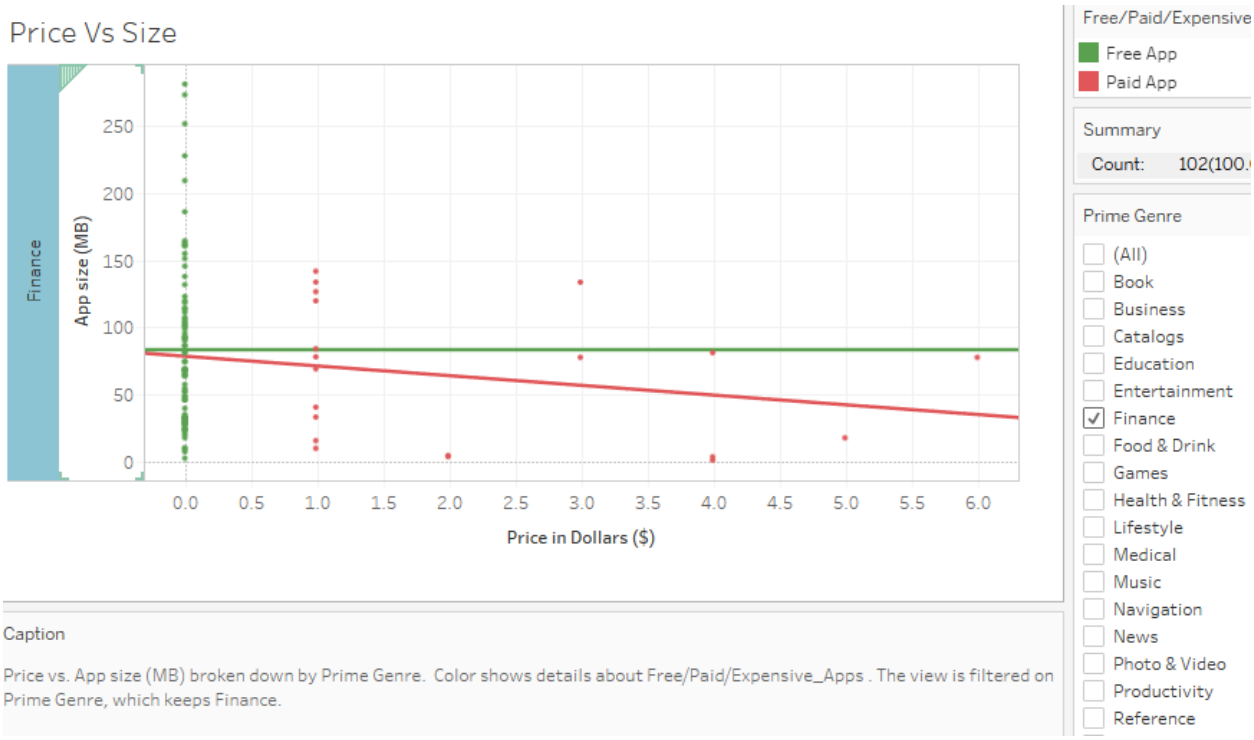
- No in Catalogs Genre



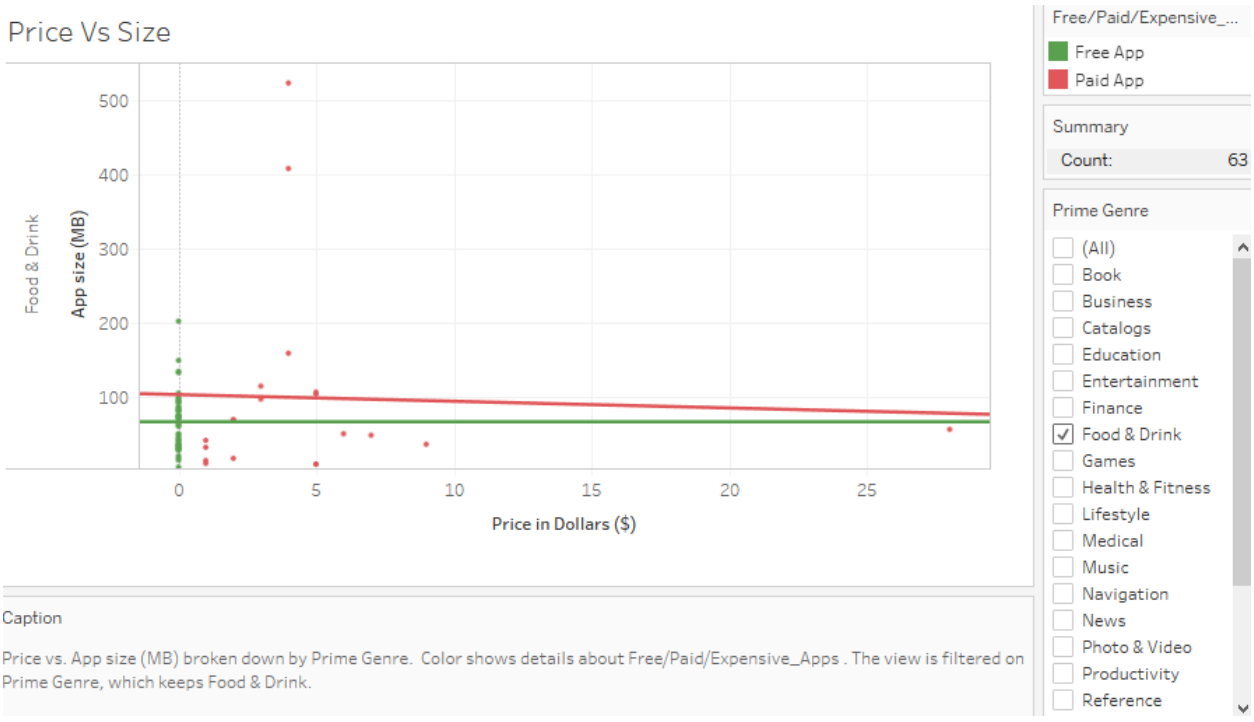
- No in Education Genre



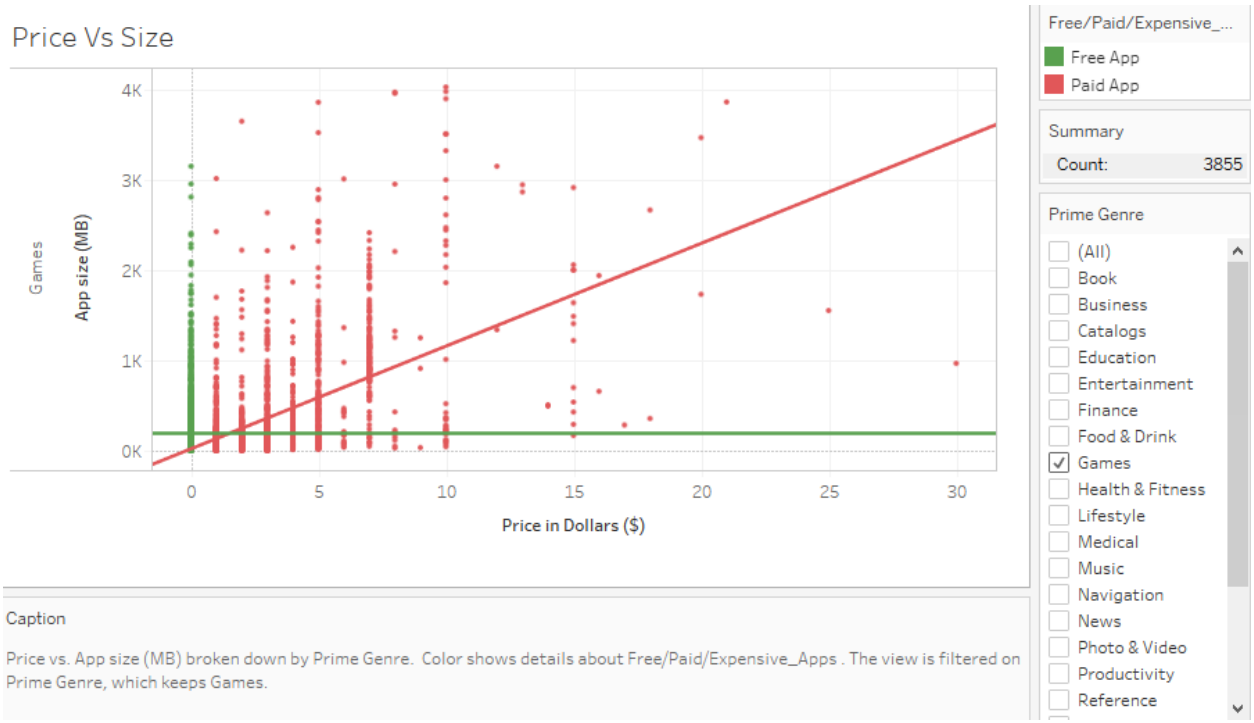
- Yes in Entertainment Genre



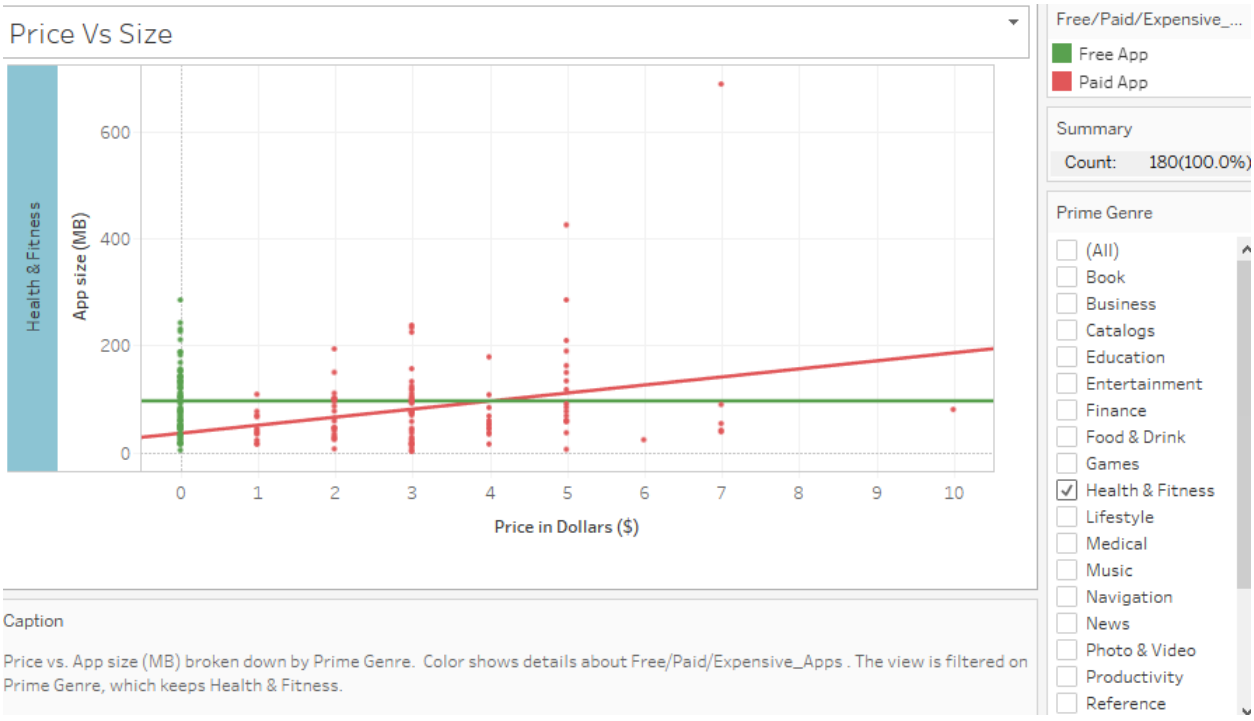
- No in Finance Genre



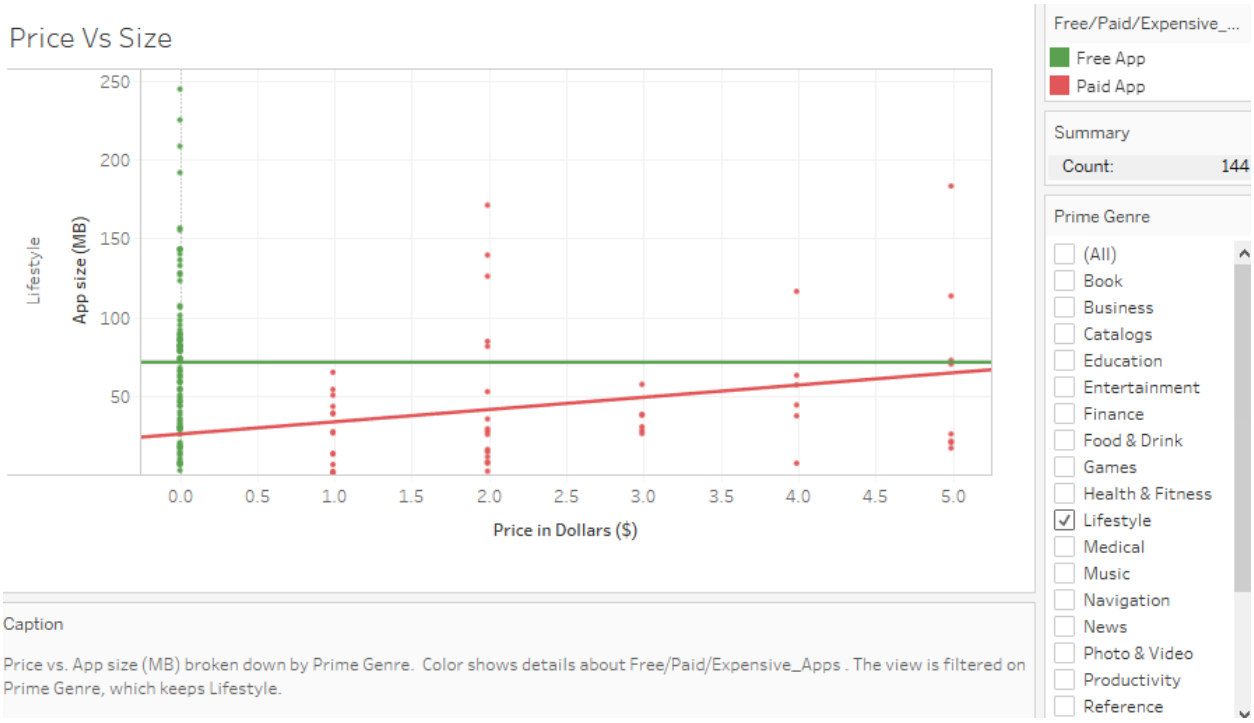
- No in Food & Drink Genre



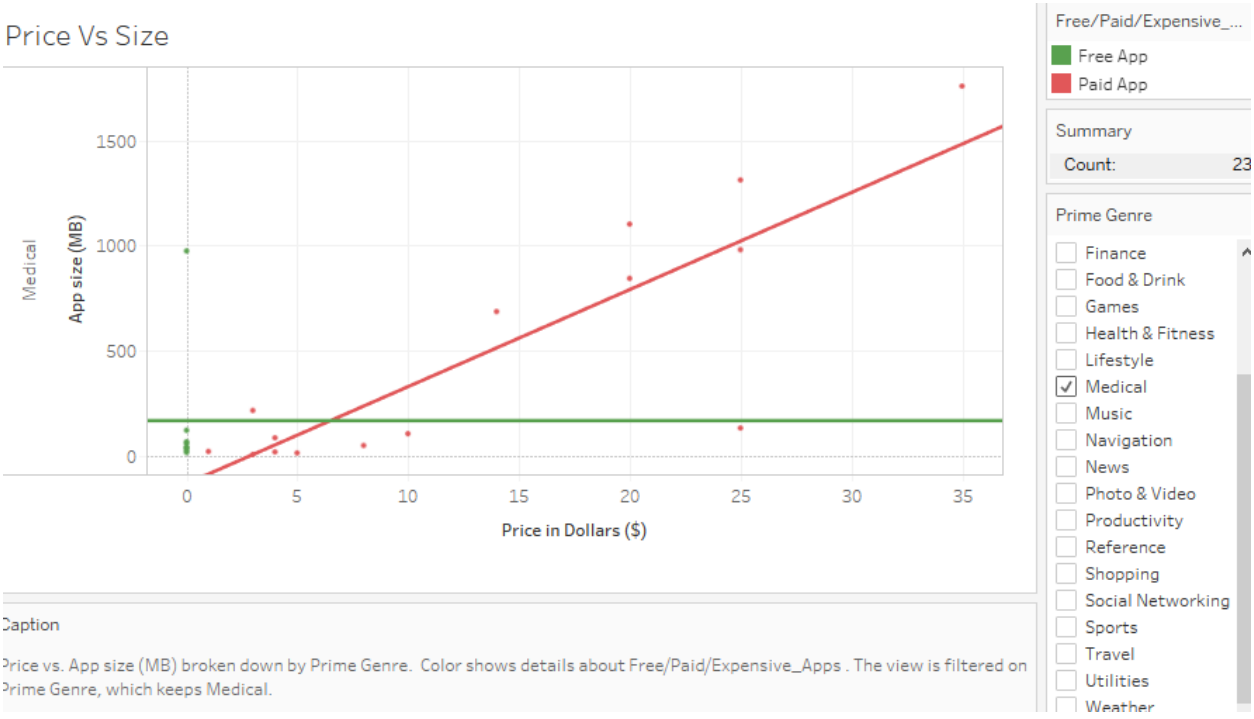
- No in Games Genre



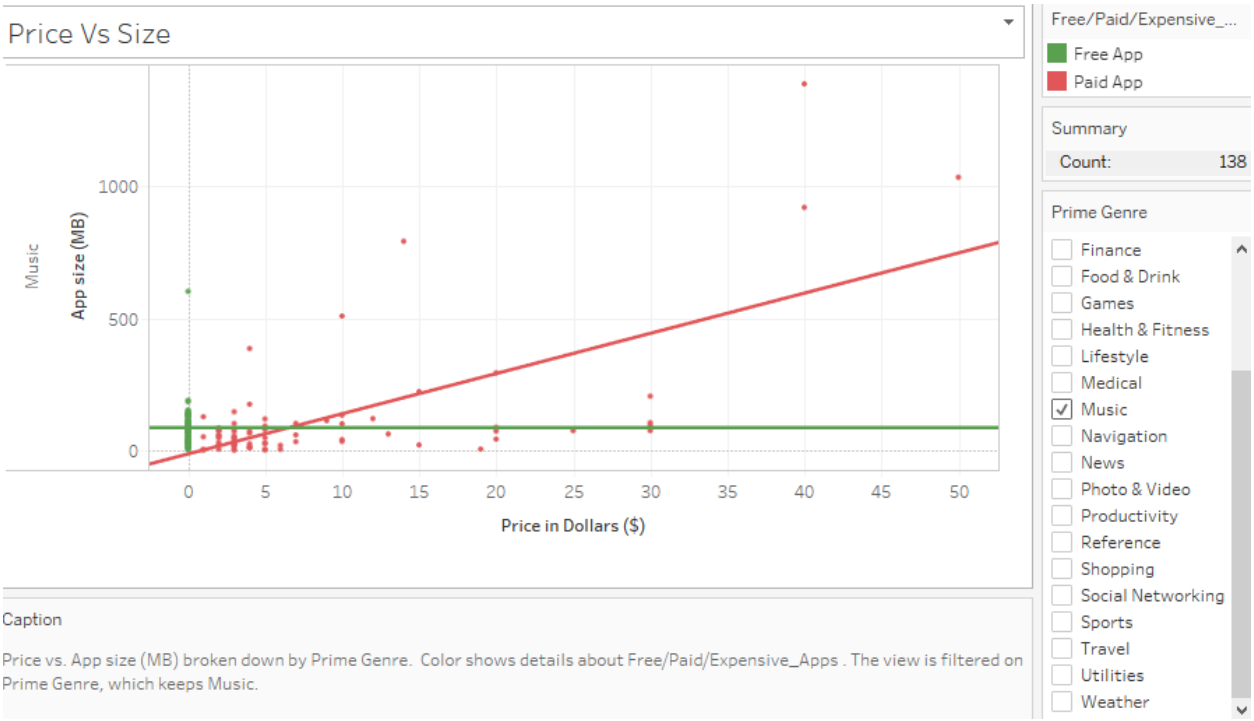
- No in Health & Fitness Genre



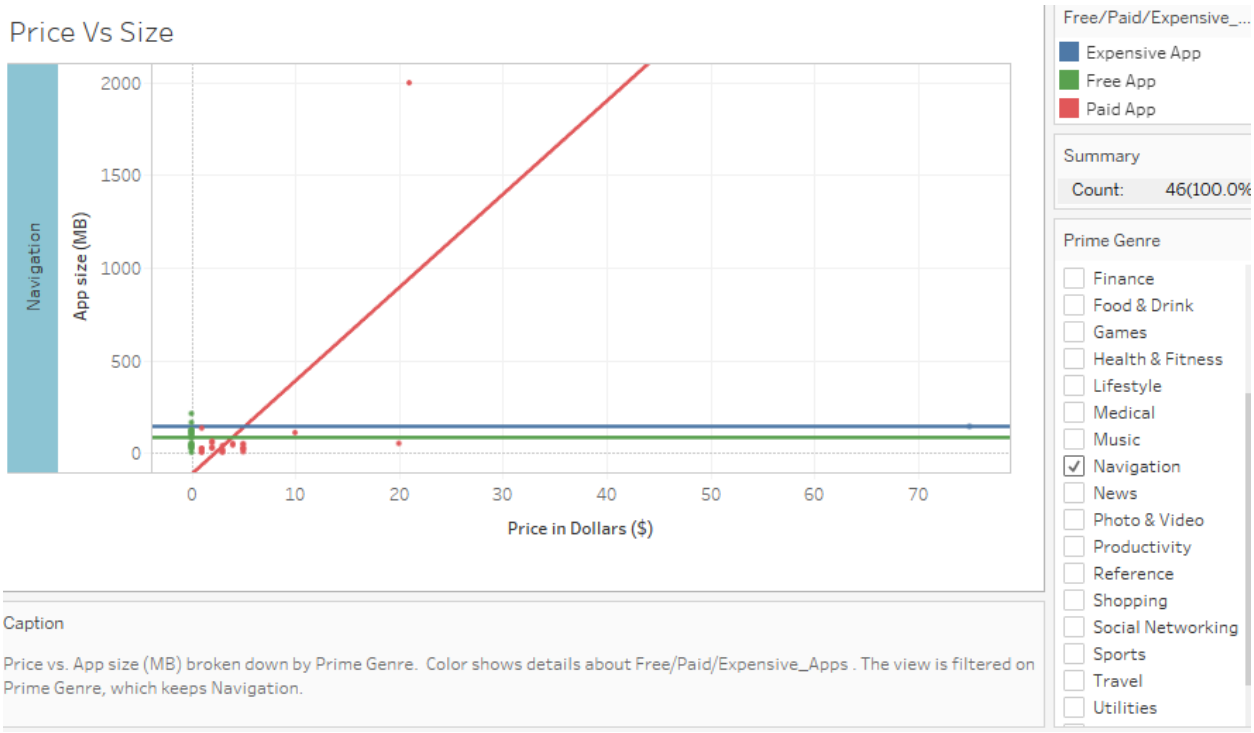
- No in Lifestyle Genre



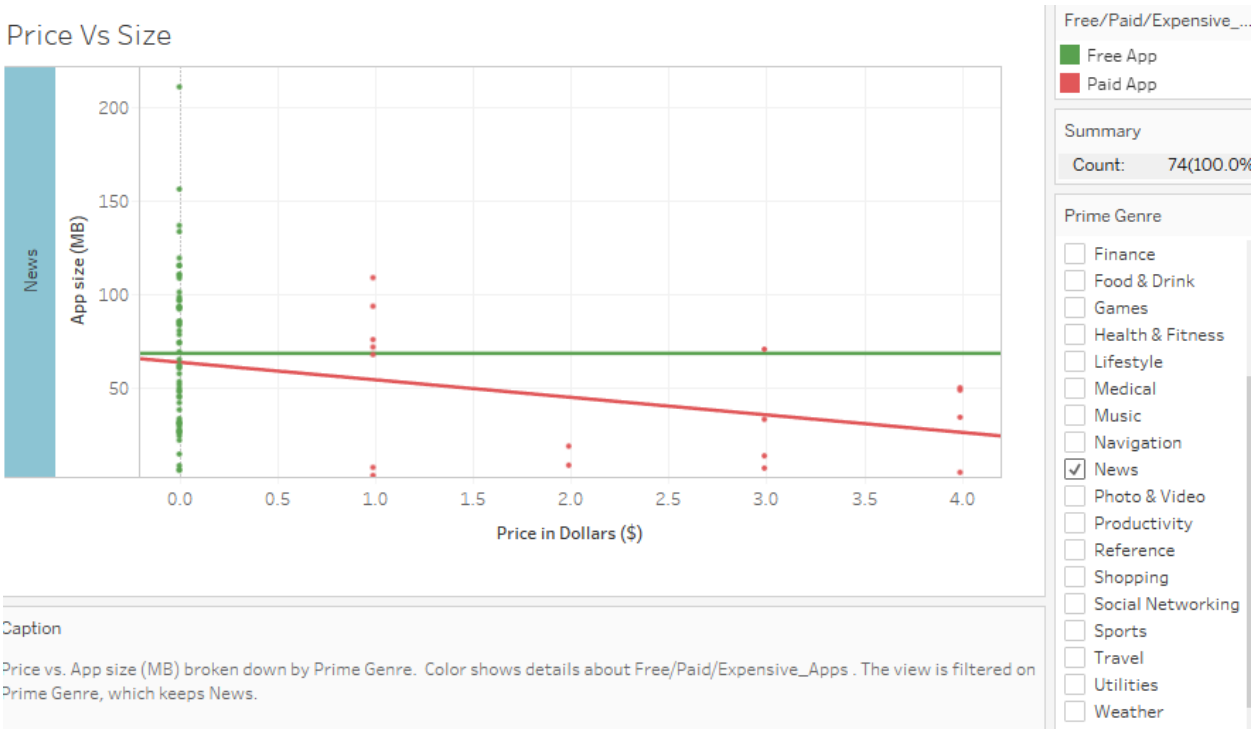
- Yes in Medical Genre



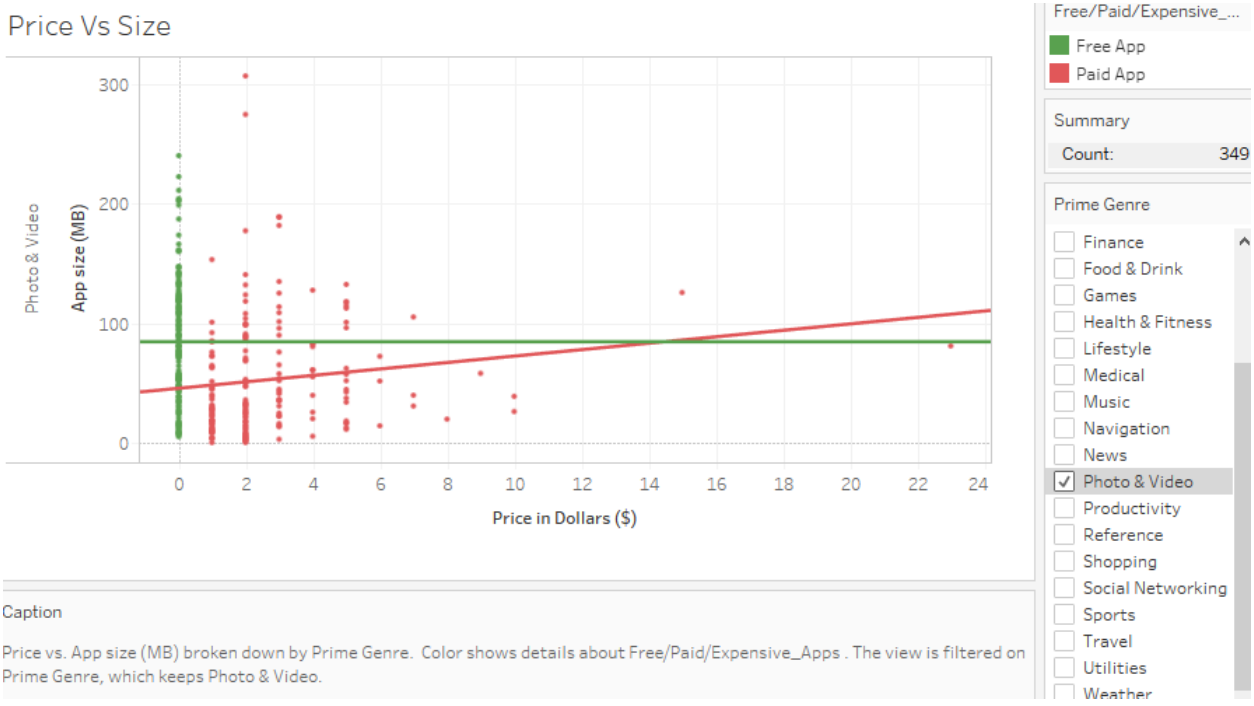
- Yes in Music Genre



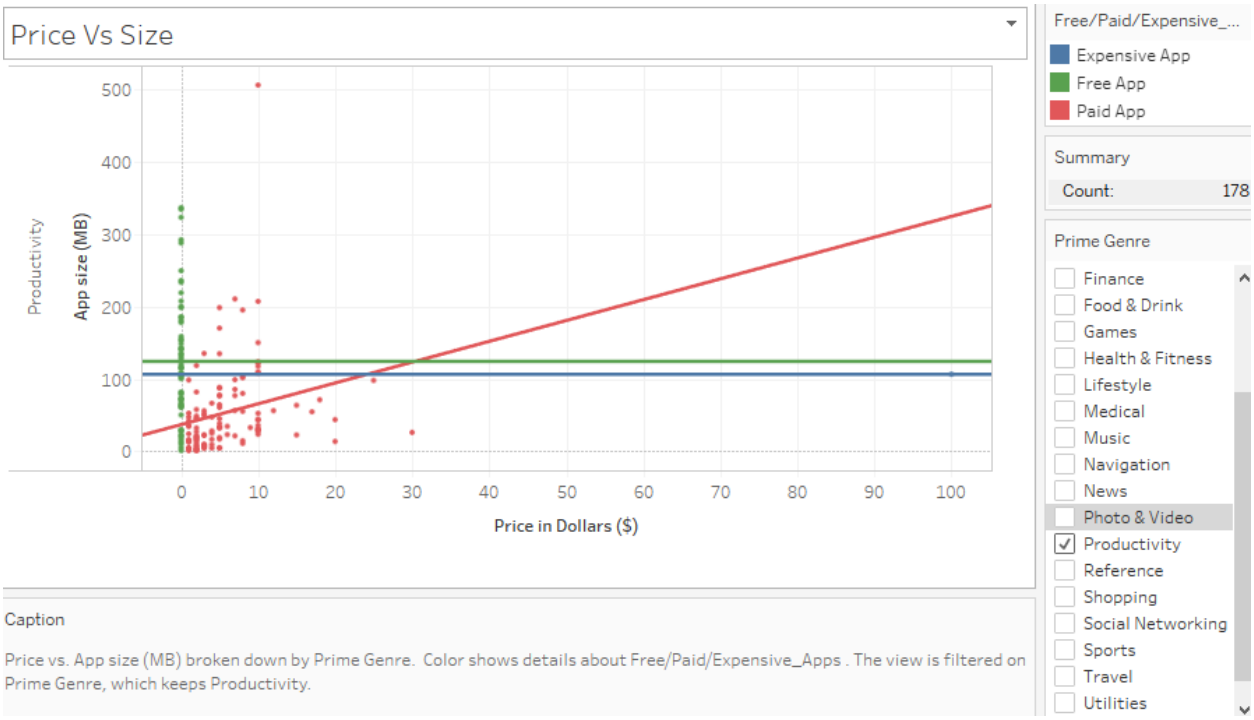
- No in Navigation Genre after excluding outliners



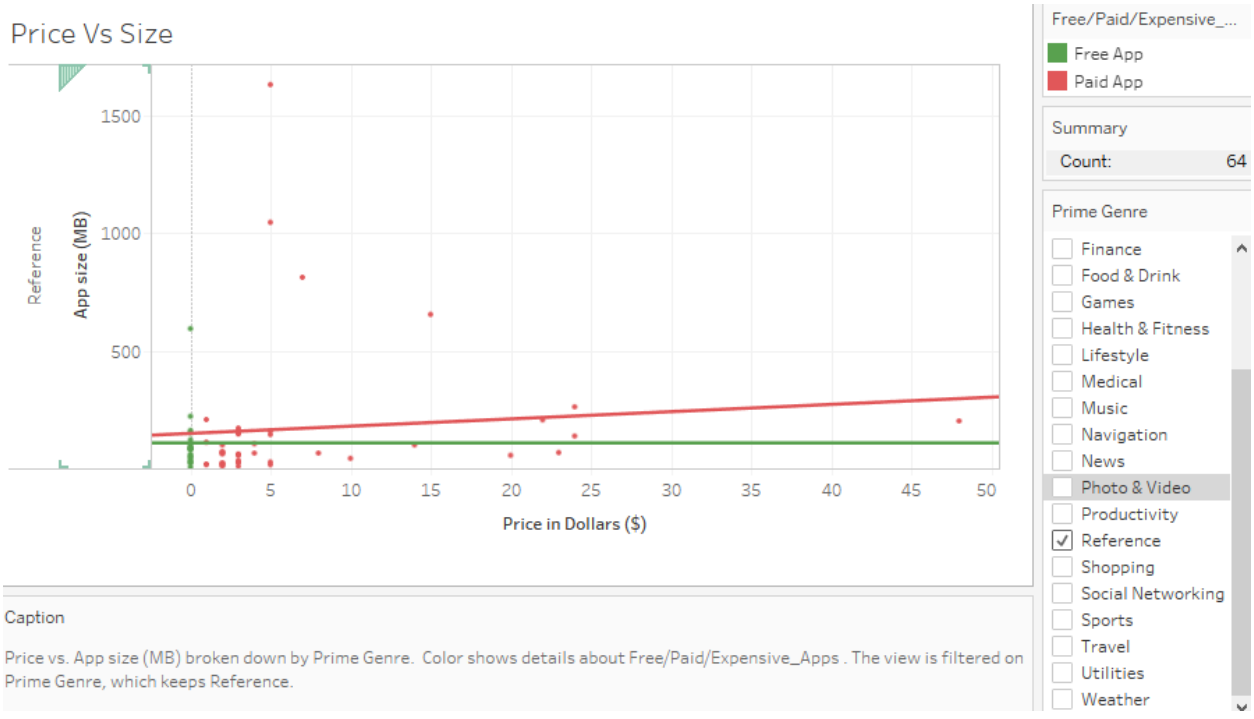
- No in News Genre



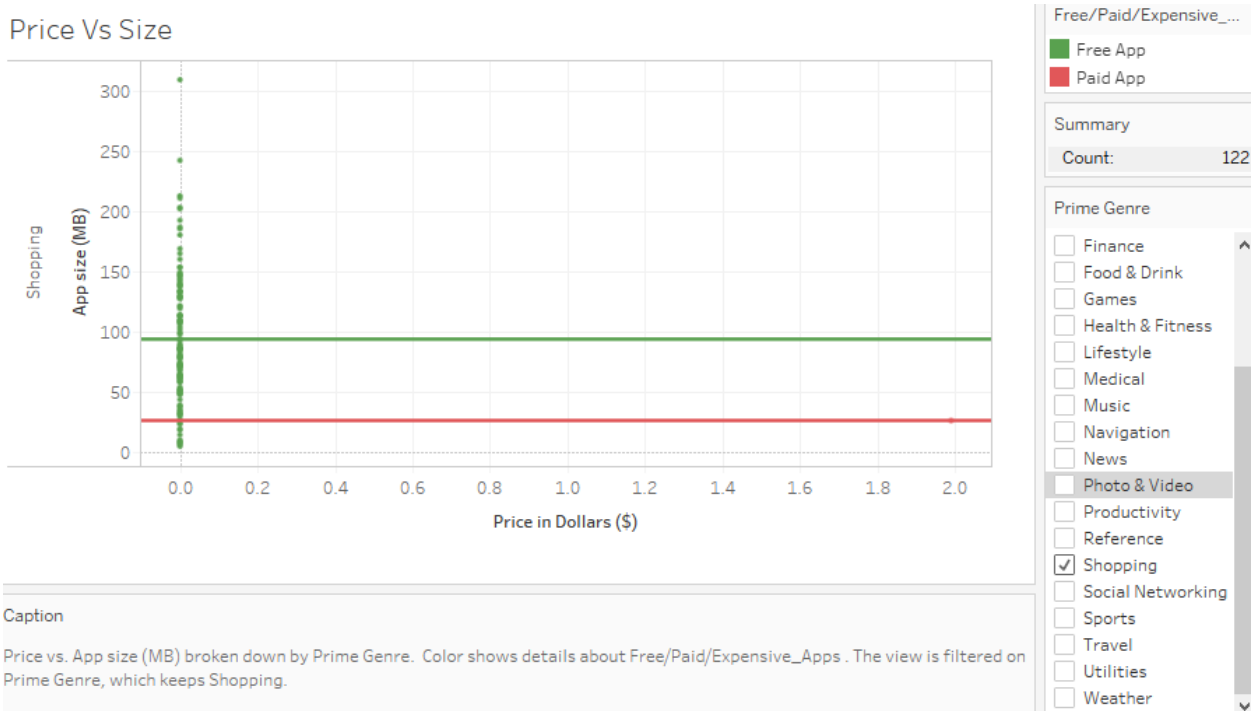
- No in Photo & Video Genre



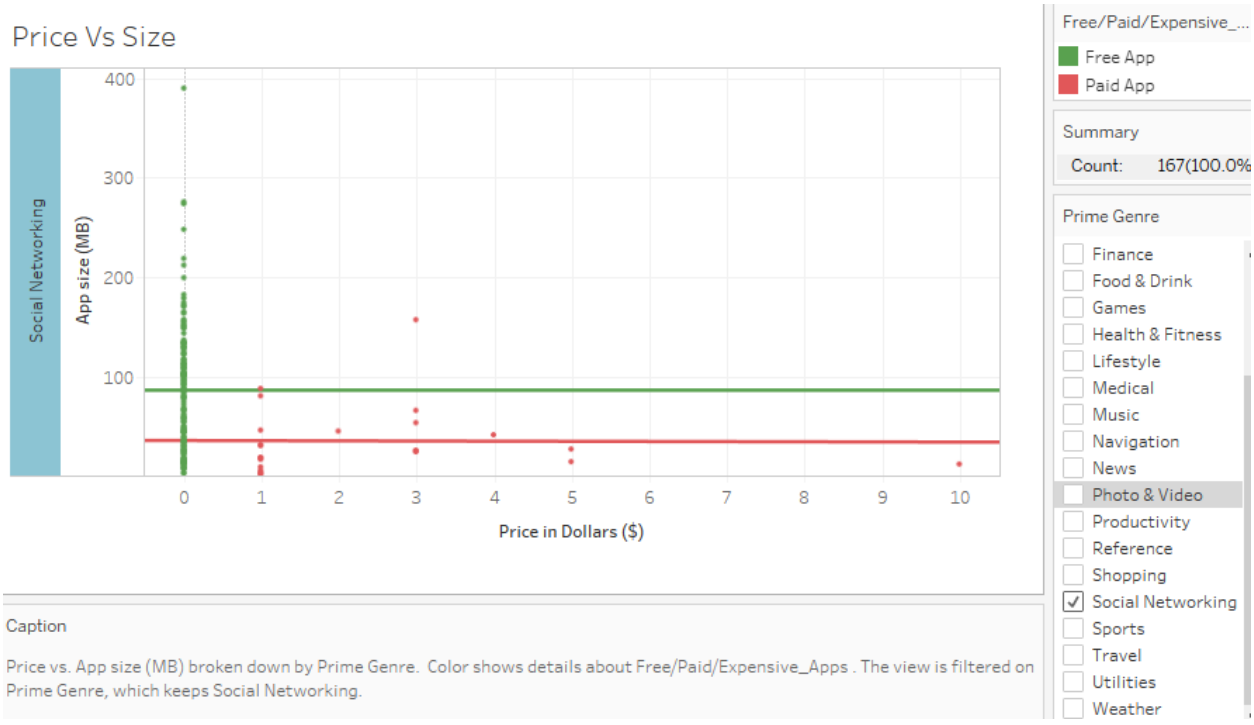
- No in Productivity Genre



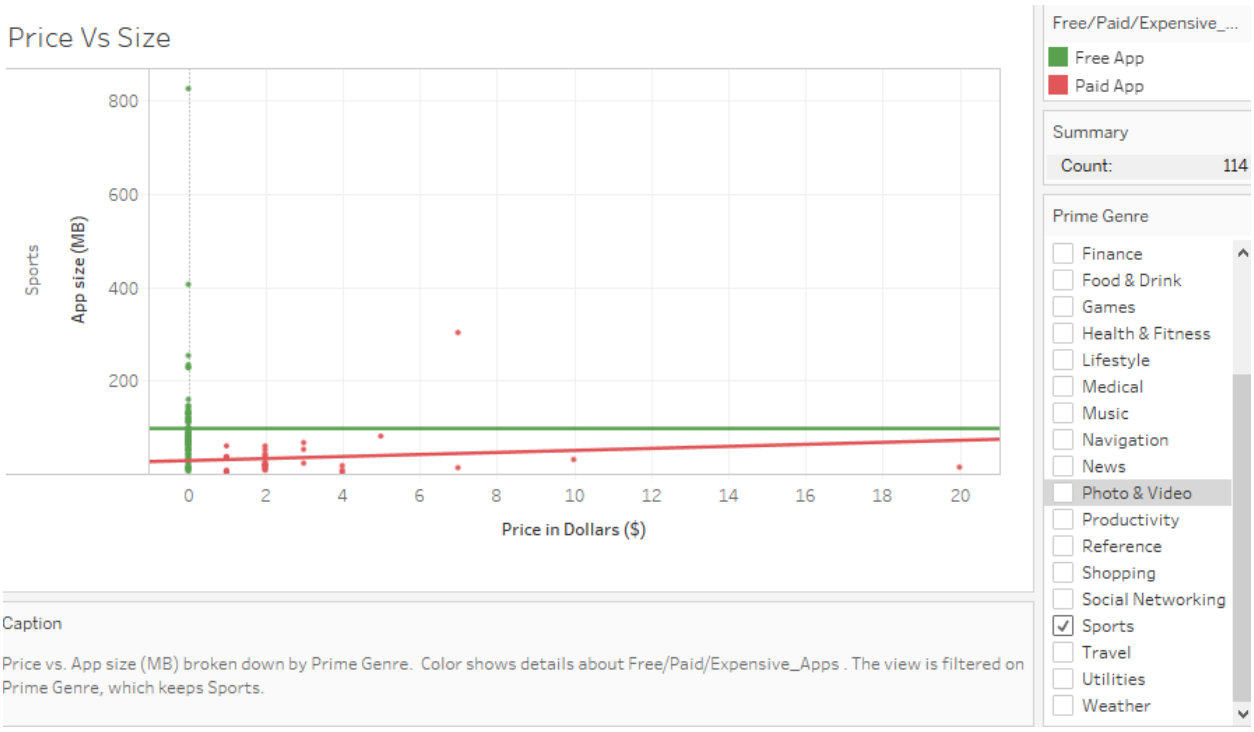
- No in Reference Genre



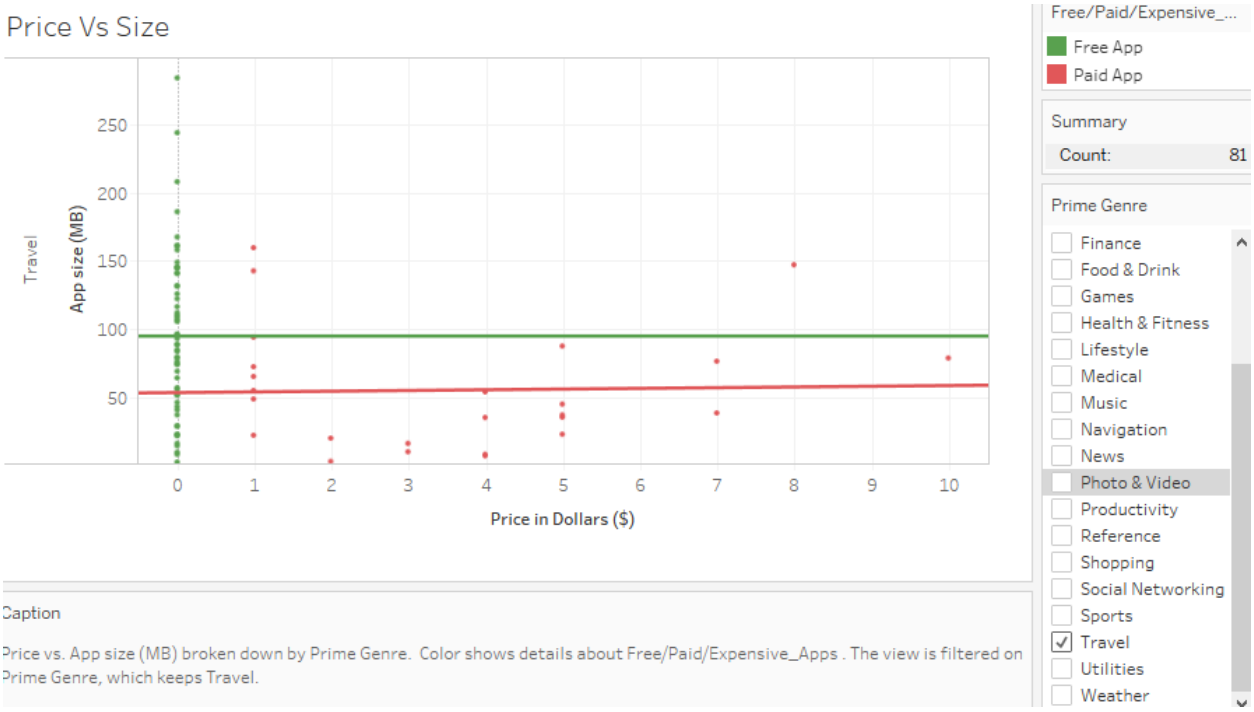
- No in Shopping Genre



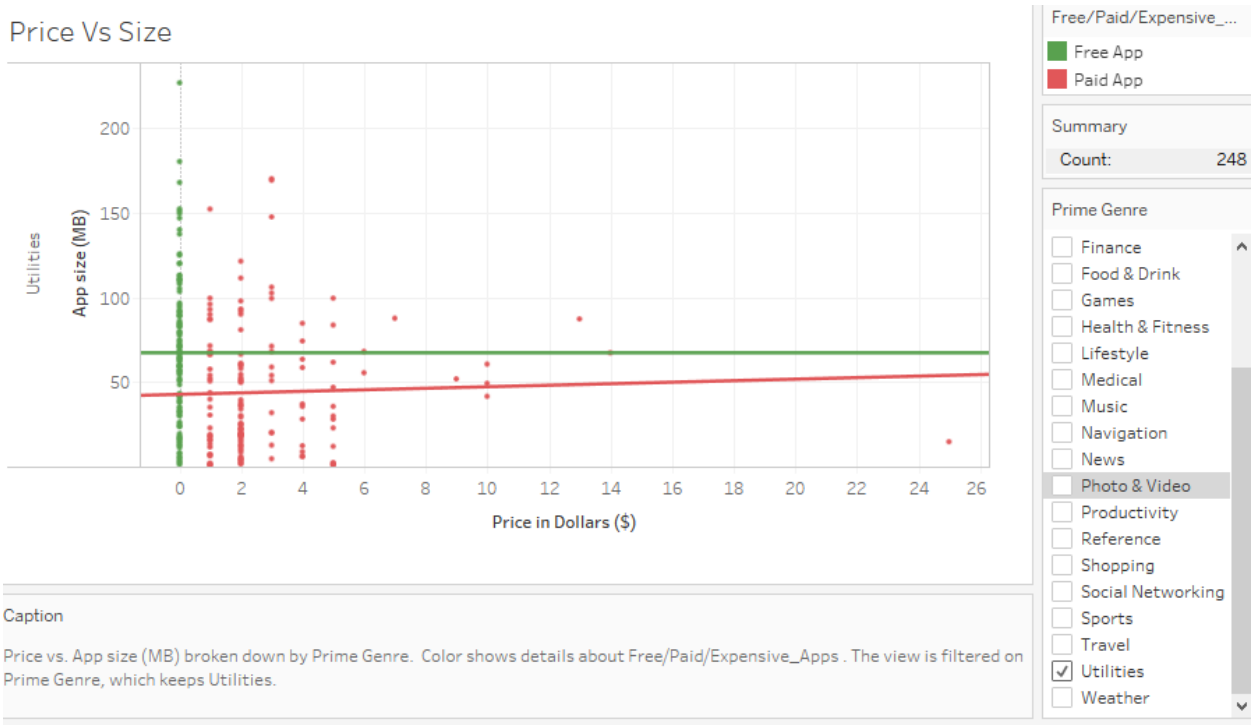
- No in Social Networking Genre



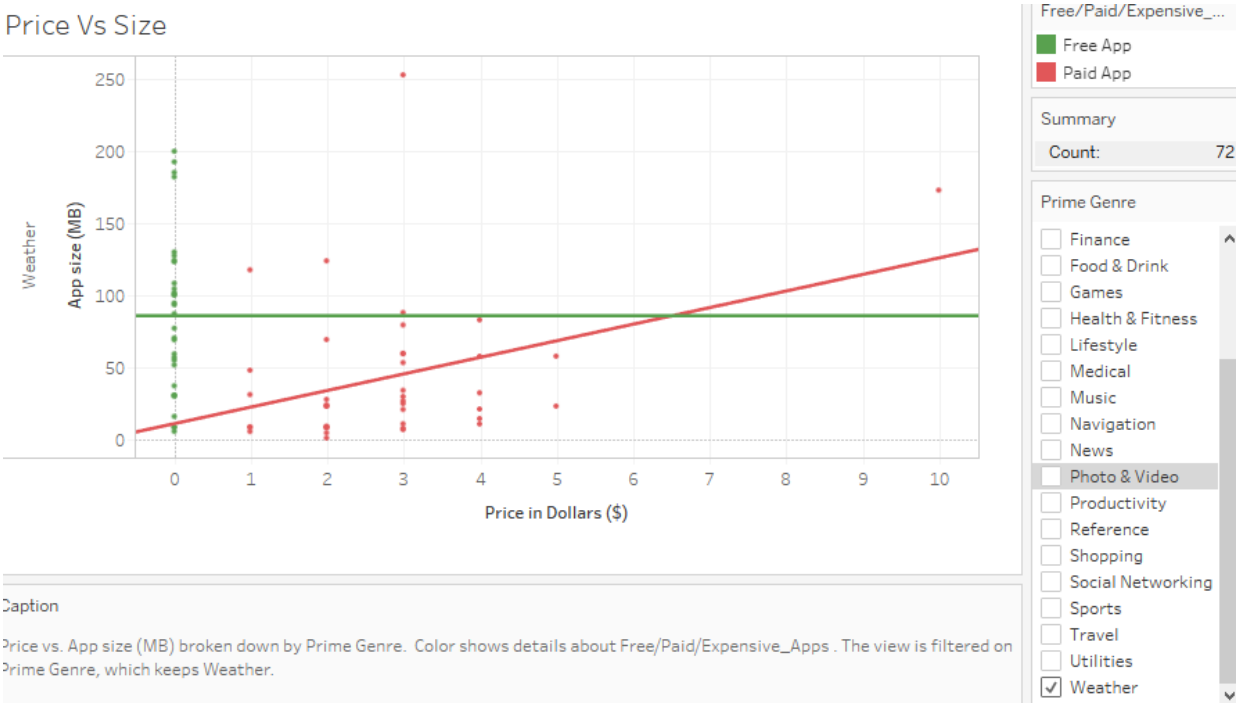
- No in Sports Genre



- No in Travel Genre



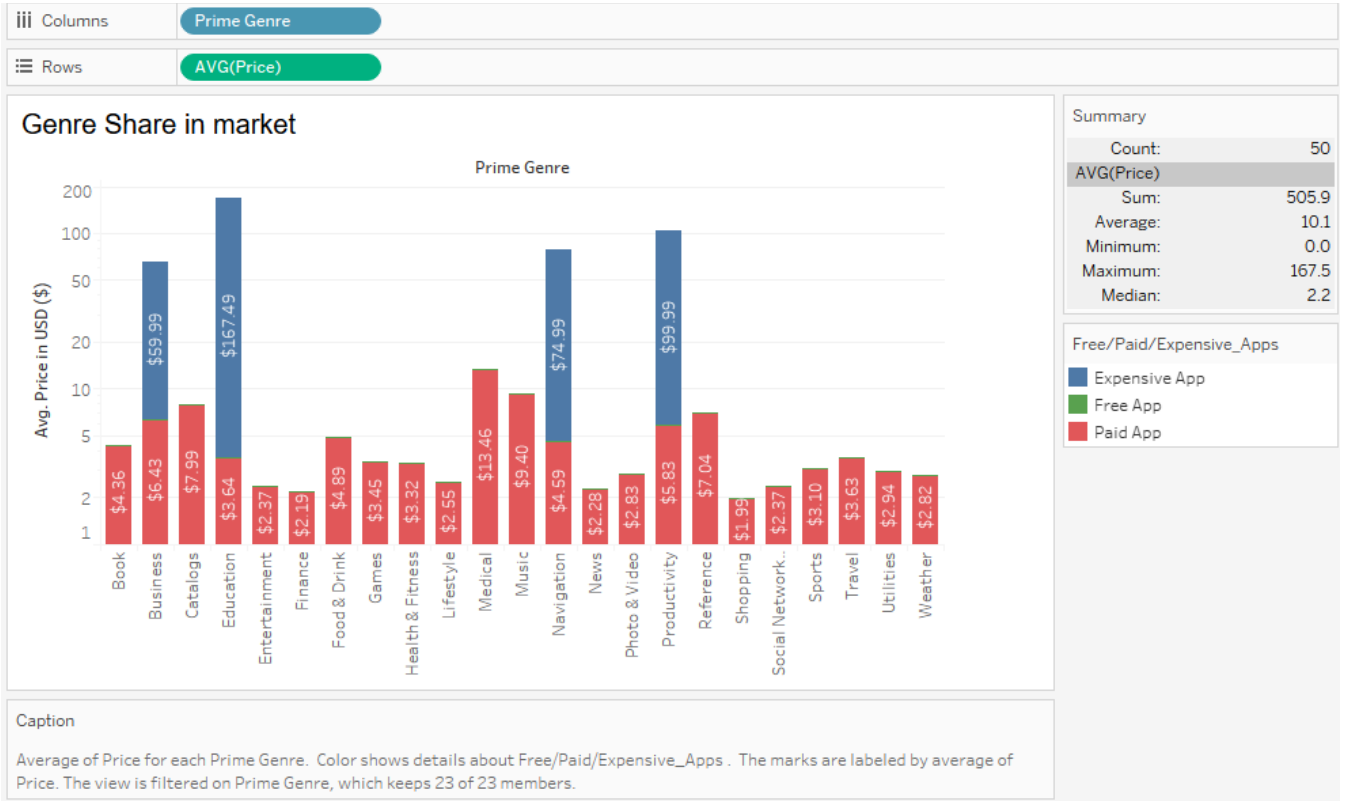
- No in Utilities Genre



- No in Weather Genre after excluding Outliners

- As the size of the app increases, it gets pricier in **Books, Games, Medical, Entertainment, Productivity** are the Genres while it doesn't get any pricier in rest of the Genres.

6. How are the apps distributed category wise? can we split by paid category?



Shopping, Finance, Lifestyle Genres have the lowest average App price, while **Medical, Music, Catalogs** Genres have highest average App price.

7. How to predict success of App?

```
from sklearn.metrics import confusion_matrix
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split
import xgboost as xgb
```

```
#Cleanup the data for modeling
data['rating_count_before'] = data['rating_count_tot'] - data['rating_count_ver']
data['paid'] = data['price'].apply(lambda x: 1 if x > 0 else 0)
data['size_MB'] = data['size_bytes'] / (1024 * 1024.0)
data.head(2)
```

]:

	id	track_name	size_bytes	currency	price	rating_count_tot	rating_count_ver	user_rating	user_rating_ver	ver	cont_rating	prime_genre	su
0	281656475	PAC-MAN Premium	100788224	USD	3.99	21292	26	4.0	4.5	6.3.5	4+	Games	
1	281796108	Evernote - stay organized	158578688	USD	0.00	161065	26	4.0	3.5	8.2.2	4+	Productivity	

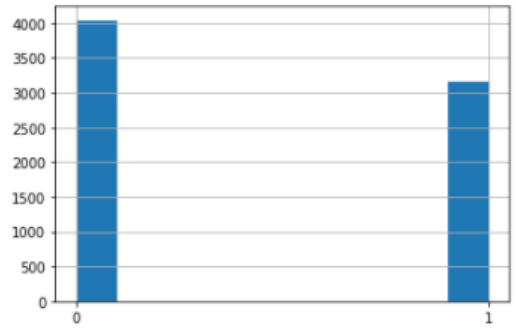
```
df_train = data[['size_MB', 'paid', 'price', 'rating_count_before', 'sup_devices.num', 'ipadSc_urls.num', 'lang.num',
                 'vpp_lic', 'prime_genre']]
target = data['user_rating']

df_train = pd.get_dummies(df_train)

def categorize_rating(x):
    if x <= 4:
        return 0
    else:
        return 1

target = target.apply(categorize_rating)

target.astype(str).hist();
```




```
X_train, X_test, y_train, y_test = train_test_split(df_train.values, target, test_size=0.2, random_state=1989, stratify=targe

print('X_train shape:', X_train.shape)
print('X_test shape:', X_test.shape)
```

```
X_train shape: (5757, 31)
X_test shape: (1440, 31)
```

Success prediction

```
from sklearn.metrics import accuracy_score
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score, cross_validate
from sklearn.ensemble import RandomForestClassifier
from lightgbm import LGBMClassifier
from xgboost import XGBClassifier
```

```
models = [RandomForestClassifier(), LGBMClassifier(), XGBClassifier()]

kfold = KFold(n_splits=5, random_state=1989)

clf_comparison = pd.DataFrame(columns=['Classfier_name', 'fit_time', 'train_score', 'test_score'])

for i, model in enumerate(models):
    clf = model
    cv_result = cross_validate(model, X_train, y_train, cv=kfold, scoring='accuracy')
    cv_score = cross_val_score(model, X_train, y_train, cv=kfold)
    clf_comparison.loc[i, 'Classfier_name'] = model.__class__.__name__
    clf_comparison.loc[i, 'fit_time'] = cv_score[2].mean()
    clf_comparison.loc[i, 'train_score'] = cv_score[1].mean()
    clf_comparison.loc[i, 'test_score'] = cv_score[0].mean()

clf_comparison
```

Out[77]:

	Classfier_name	fit_time	train_score	test_score
0	RandomForestClassifier	0.67159	0.638889	0.644965
1	LGBMClassifier	0.682016	0.702257	0.677951
2	XGBClassifier	0.681147	0.684028	0.677951

At the beginning, I set a multi class classfication problem to predict user_rating using this dataset.
I categorized user ratings into 5 groups, (1) 0 ~ 1 (2) 1 ~ 2 (3) 2 ~ 3 (4) 3 ~ 4 (5) 4 ~ 5.
But, the output was bad.

For now, I set a binary classfication problem to predict success of apps. I supposed that app which has more than 4 user_rating is succefull.
For this approach, the prediction accuracy is about 70%. It is not bad.

References:

https://hbr.org/resources/pdfs/comm/google/19688GoogleFeb2016.pdf

https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps

https://medium.com/swlh/top-11-mobile-app-analytics-platforms-pricing-included-cdc553578fd