Phase 1 Case Study
Team Exploracity

08/28/2020

The Team

Seth Chart

Math loving data nerd.

Leah Pope

Likes software and data (and books and cats)

Jaklyn Soler

Likes books. Humor darker than black market kidney transplants.

The Challenge

Data

Google Play Store App Data collected in 2018

Source: https://www.kaggle.com/lava18/google-play-store-apps

Context

Our client suspects there is valuable information waiting to be uncovered in this dataset, enabling them to better position themselves in the app provider market.

Problem statement

Answer the client's top 3 burning questions.

Surprise and delight the client by answering a bonus question - one they didn't even know to ask.

Client Question 1

The Best App Category is?

Which app category, in your opinion, has the best ratings?

Art & Design, Health & Fitness, and Education.

How are you measuring best ratings?

We defined ‘best rated’ by the highest median rating.

Q1. We Aggregated Ratings By Category

- The median rating tells us how an app that is in the “middle of the pack” performs. This measure is not sensitive to unusually high ratings or unusually low ratings.

```
median_rating_by_category = google_play_df[['Category', 'Rating']].groupby('Ca
tegory').median().sort_values(by='Rating', ascending=False)
median_rating_by_category
```

	Rating
Category	
ART_AND_DESIGN	4.40
HEALTH_AND_FITNESS	4.40
EDUCATION	4.40
COMICS	4.35
PARENTING	4.35
BOOKS_AND_REFERENCE	4.30
SHOPPING	4.30

Q1. We Defined Best Rated by Maximum Median

- We defined top_median as the maximum in the medians of the ratings.
- We even made our results extra pretty by replacing the underscores with spaces.

```
top_median = median_rating_by_category[median_rating_by_category['Rating'] ==
median_rating_by_category['Rating'].max()]
for category in top_median.index:
    print(category.replace('_', ' ').title()+'\n')
```

```
Art And Design

Health And Fitness

Education
```

Did we have a good time in the terminal today?

Client Question 2

Are Ratings and Size Related?

Is there a relationship between ratings and size?

At best, a weak relationship.

How did you measure the relationship?

By correlation calculation.

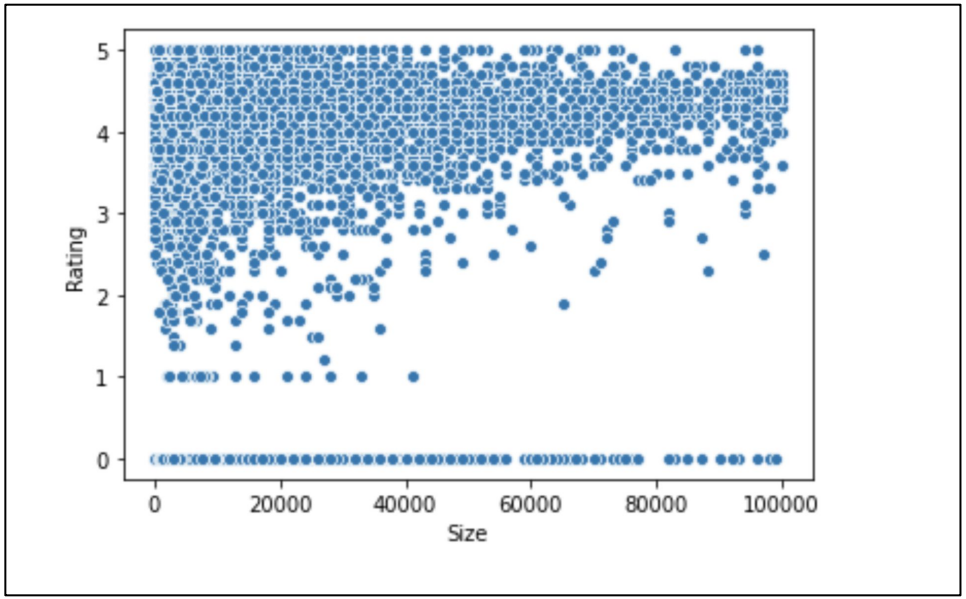
Why did you choose this measurement?

To determine the relationship between the two variables.

Q2. Rating & Size Scatter Plot

- We calculated a correlation coefficient of 0.165324, which indicates a weak positive relationship.

```
sns.scatterplot(x='Size', y='Rating', data=google_play_df)
```



Client Question 3

What Genres are in the ‘Game’ Category ?

How many genres are represented in the ‘Game’ Category?

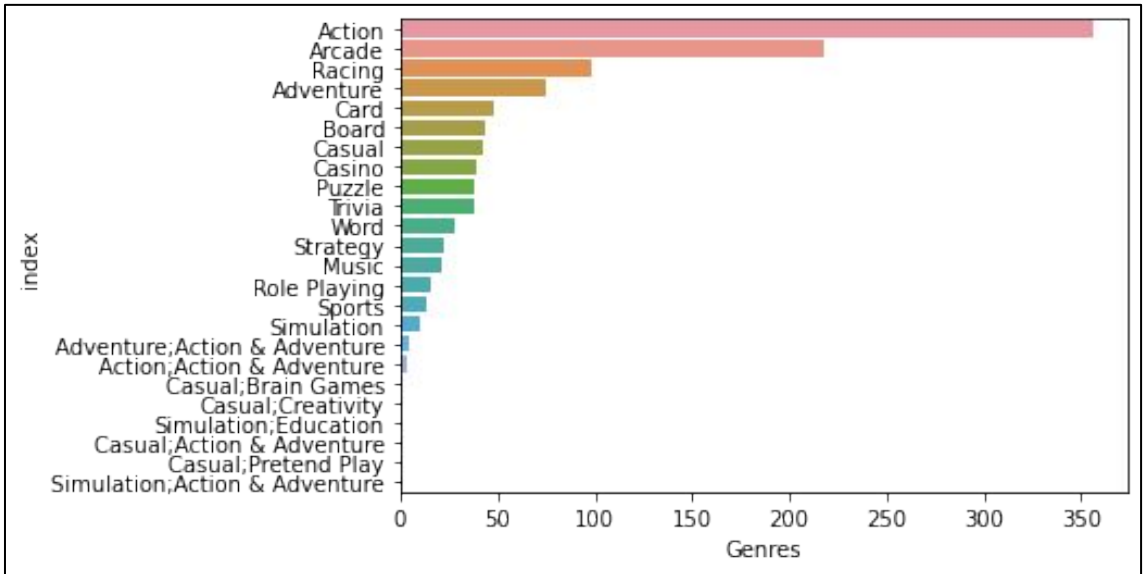
There are 24 genres in ‘Game’.

What are their counts?

Counts are provided in following barchart.

Q3: Counts by Genre in the ‘Games’ Category

- We used the following bar chart to show counts by Genre for the ‘Games’ Category.
- Highest count is ‘Action’ with 356
- Lowest count(s) are Casual;Brain Games, Casual;Creativity, Simulation;Education, Casual;Action & Adventure, Casual;Pretend Play, and Simulation;Action & Adventure. All tied at 1




```
In [27]: game_df['Genres'].value_counts()

Out[27]: Action                          356
         Arcade                          218
         Racing                           98
         Adventure                        75
         Card                             48
         Board                            44
         Casual                           43
         Casino                           39
         Puzzle                           38
         Trivia                           38
         Word                             28
         Strategy                         22
         Music                            21
         Role Playing                     16
         Sports                           13
         Simulation                       10
         Adventure;Action & Adventure      5
         Action;Action & Adventure         3
         Casual;Brain Games                1
         Casual;Creativity                 1
         Simulation;Education              1
         Casual;Action & Adventure         1
         Casual;Pretend Play               1
         Simulation;Action & Adventure     1
         Name: Genres, dtype: int64
```

Team Question 4

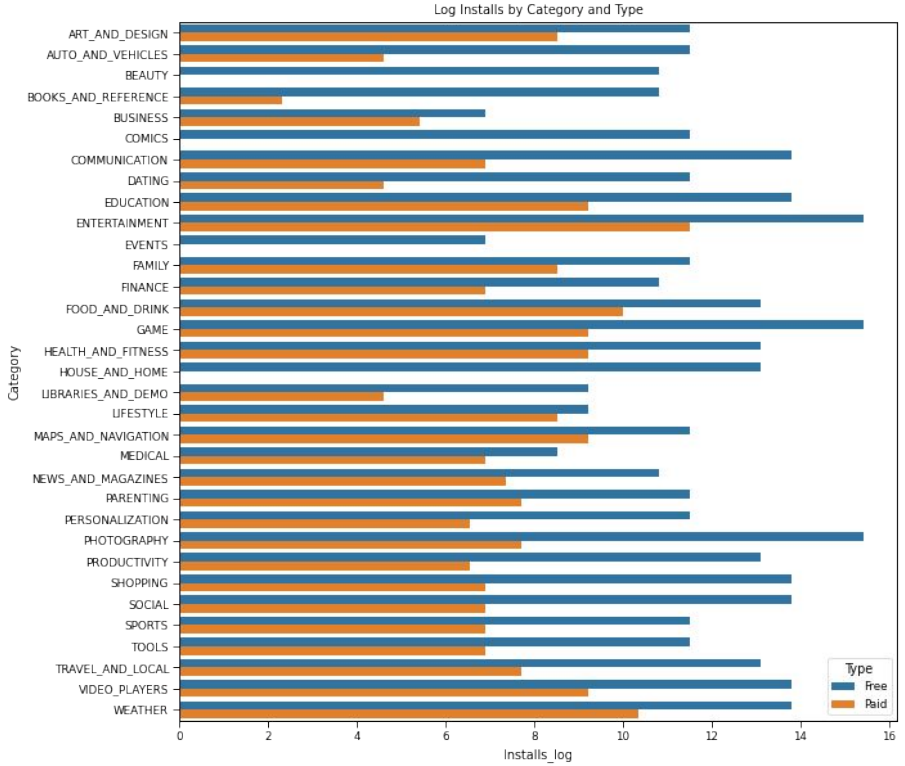
What did our client not know to ask about?

In each Category, how does number of Installs depend on Type (Free vs Paid)?

It’s apparent that Free apps have a greater number of Installs, but if our client wants to release a Paid app, we have some insights to guide them.

Q4: Where are customers willing to pay for apps?

- Used the log of installs so that the Free counts wouldn't overwhelm the chart.
- We see that the Install count for Free apps are higher across all app Categories.
- However, there are a few Categories where the gap between Paid and Free installs is smaller, suggesting that customers are more open to Paid apps:
  - Lifestyle
  - Business
- Also there are a few Categories where customers seem very unwilling to install Paid apps:
  - Beauty
  - Comics
  - House and Home



Future Work

Idea One

Monitor changes relative to app releases. How do these values change as new app versions are released?

- Rating
- Installs
- Android Version

Idea Two

Improve data cleaning/ data normalization.

- Some Categories appeared to have Subgenres (ex: Casual;Creativity).
- It might be better to roll subgenres up into the main Genre

Too soon..? :’-)


Time spent analyzing data
Time spent cleaning data
@iopsychmemes

Data Cleaning Methods for Questions 1-4

- We removed 483 duplicate rows from the Play Store Listings data.
- We filled 1465 missing Ratings with zero. There were no zero ratings in the original data.
- We found that, even after removing duplicate rows, there were still multiple listings for some Apps.
- We removed the ‘1.9’ category since it did not appear to be a legitimate category. Only one row in our data had this category.
- The Size variable was stored as a string with units of Megabytes or Kilobytes. There were also Apps with the string value ‘Varies with Device’. We converted strings to a float representing the number of kilobytes. For devices with ‘Varies by Device’ we filled with NaN.
- The Installs variable refers to the number of installs of the app through the Play Store. This variable was stored as a string with extraneous punctuation. We removed punctuation and converted to a float.

Did you enjoy our presentation?

Thanks to...

- our dataset provider, Lavanya Gupta
- our instructor, Rafael Carrasco
- our cohort members
- our support networks @ Flatiron and @ Home

Reach out to us with your questions and remember Team Exploracity's motto:

"Have the audacity to explore!"