Analysis of Google Play Store Applications

Chad Schaschwary, Aridania Gerardo, Phu Hoang

Stage 1: Analyzing Our Dataset

- Our dataset was published on Kaggle
- Contains data about 1.1 million applications from the Google Play Store
- Strong representation of available apps (around 3.5 million apps altogether)
- 23 columns of descriptive application data
    - App name, category, rating, rating count, # of installs, price, developer, last update, etc.

Preview

```
df.head(5)
```

	App Name	App Id	Category	Rating	Rating Count	Installs	Minimum Installs	Maximum Installs	Free	Price	Currency	Size	Minimum Android	Developer Id	Developer Website
0	HTTrack Website Copier	com.httrack.android	Communication	3.6	2848.0	100,000+	100000.0	351560	True	0.0	USD	2.7M	2.3 and up	Xavier Roche	http://www.httrack.com/
1	World War 2: Offline Strategy	com.skizze.wwii	Strategy	4.3	17297.0	1,000,000+	1000000.0	2161778	True	0.0	USD	86M	5.1 and up	Skizze Games	http://stereo7.com/
2	WPSApp	com.themausoft.wpsapp	Tools	4.2	488639.0	50,000,000+	50000000.0	79304739	True	0.0	USD	5.8M	4.1 and up	TheMauSoft	http://www.themausoft.com
3	OfficeSuite - Office, PDF, Word, Excel, PowerP...	com.mobisystems.office	Business	4.2	1224420.0	100,000,000+	100000000.0	163660067	True	0.0	USD	59M	4.4 and up	MobiSystems	http://www.mobisystems.com
4	Loud Player Free	com.arthelion.loudplayer	Music & Audio	4.2	665.0	50,000+	50000.0	73463	True	0.0	USD	29M	5.0 and up	Arthelion92	http://www.arthelion.com

What we want to know

Goal: We want to understand common characteristics of apps with higher ratings

- What strongly influences the rating of an application?

Machine learning: Can we create a model that predicts the rating of an app based on specific features?

- Regression analysis

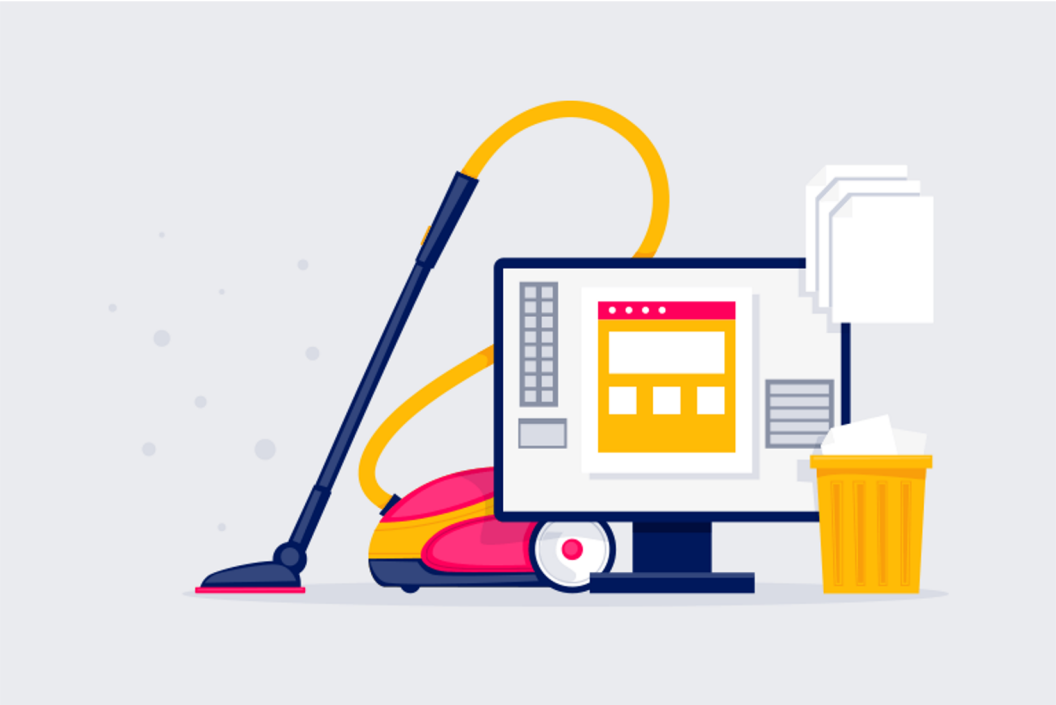
Correlations between two variables

- Do free or paid apps receive higher ratings?
- Are there specific categories of apps that receive better ratings, on average?
- Do applications with more installations get rated more based on whether it being free, ad supported, or an editor's choice?

Benefits of our research

- Understanding consumer preferences can help businesses cater their offerings that meet the needs of end consumers
- Helps application developers determine what features are most important prior to release
- Predict performance of an application in the marketplace (based on customer perception)

STAGE 2: Cleaning and Data Preparation



Cleaning and Data Preparation

How you get your data?

- We acquired our data from a pre - existing dataset from kaggle.com
- Google Play Store Apps | Kaggle

How you clean your data?

- We used pandas data manipulation package to clean up the data. First by removing duplicates. Second by deleting unnecessary categories of data. Lastly by dropping series with values of NaN. From that data, we used a finite range of data to accommodate to unexpected outliers.

Outside Resources to Support Data

No outside datasets were needed or used to support our dataset and our research questions.

STAGE 3: Exploring the Data

Descriptive analysis：What strongly influences the rating of an application?

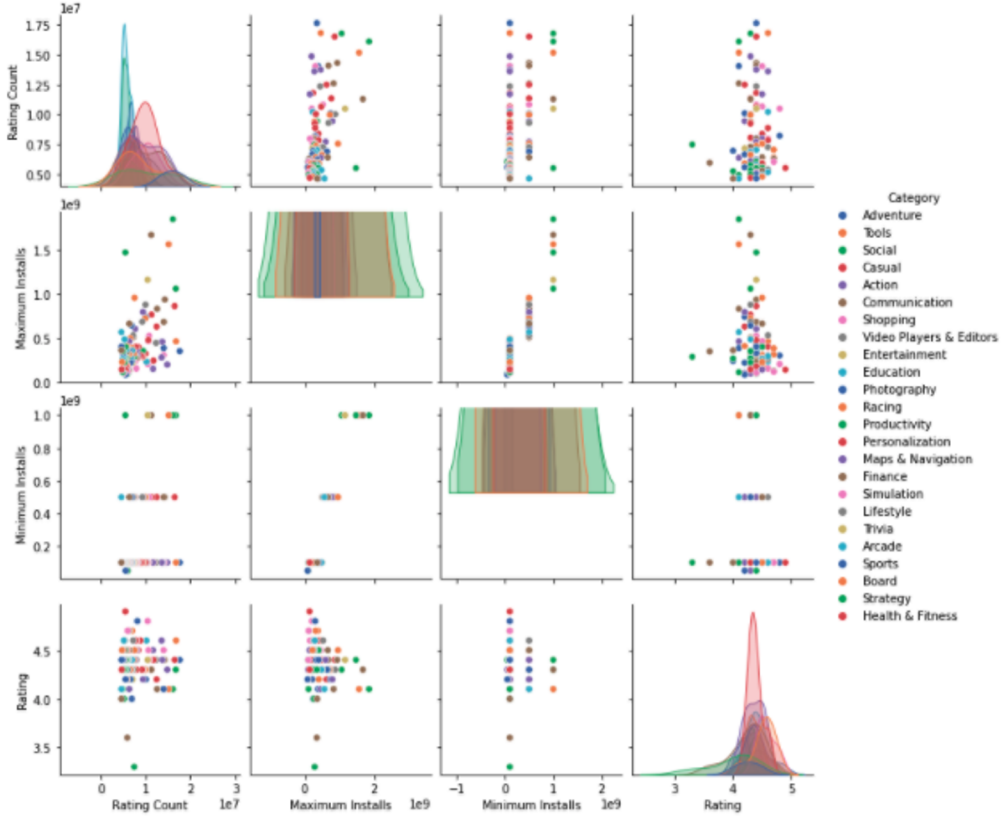


Figure 1. Correlation Matrix of Top 100 Rating Count - Rating Count,Maximum Installs,Minimum Installs, and Rating

Do free or paid apps receive higher ratings?

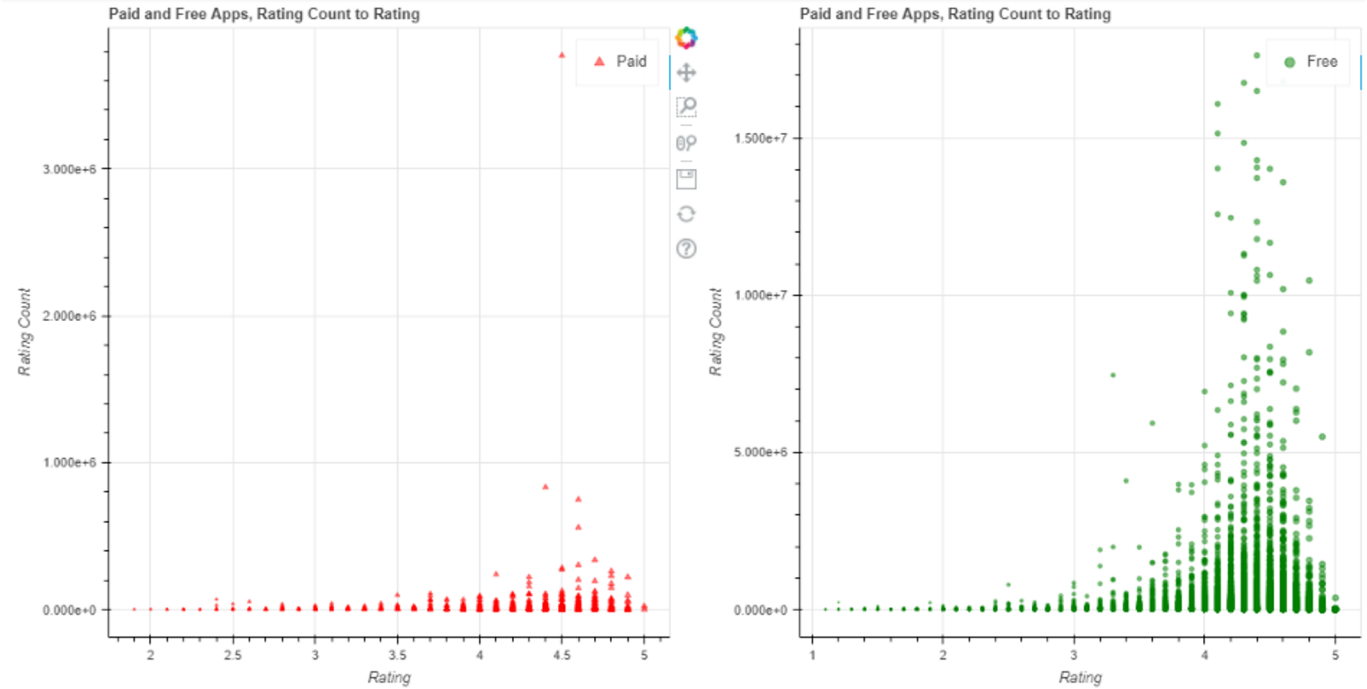
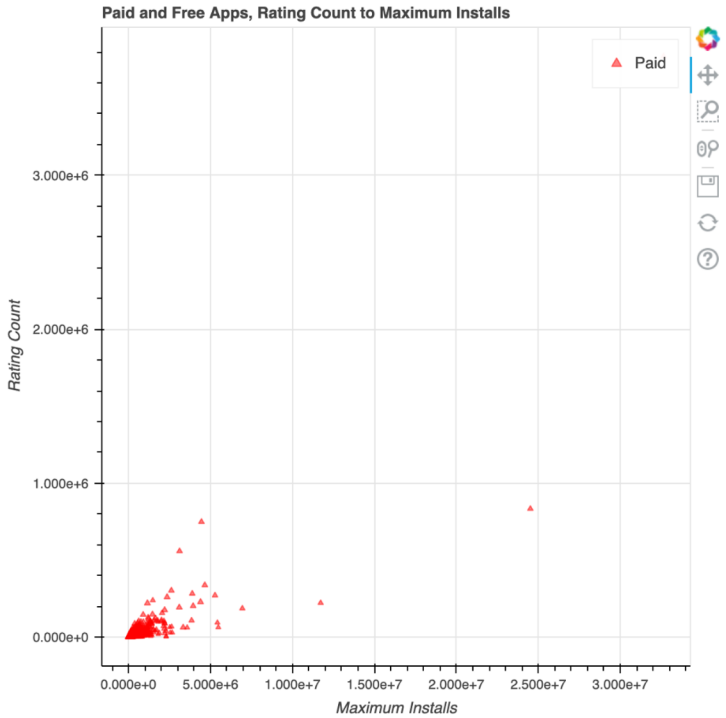


Figure 2. Bokeh Scatter Plot of Paid vs Free Apps, Rating Count to Rating

Do applications with more installations get rated more based on whether it being free, ad supported, or an editor's choice?: Free



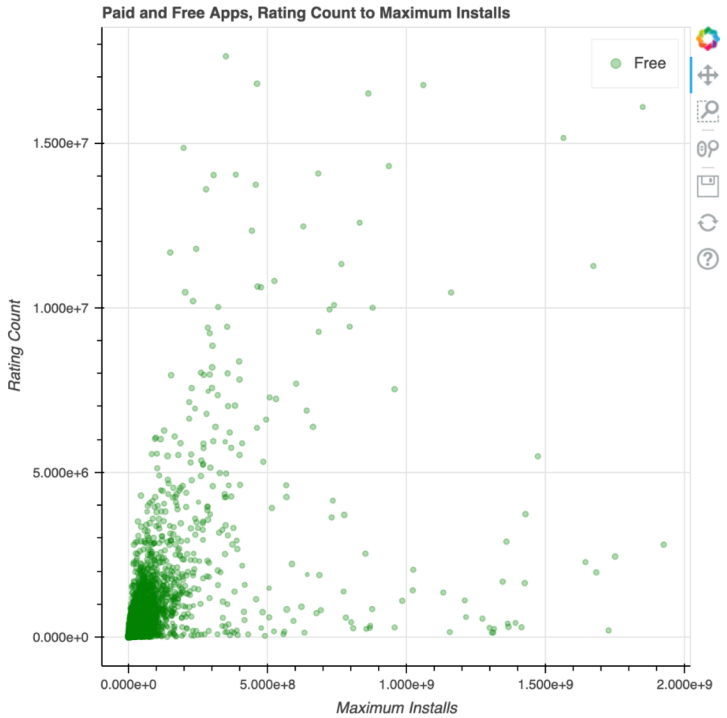


Figure 3. Bokeh Scatter Plots of Paid Apps and Free Apps Rating Count to Maximum Installs

Do applications with more installations get rated more based on whether it being free, ad supported, or an editor's choice?: Ad Supported

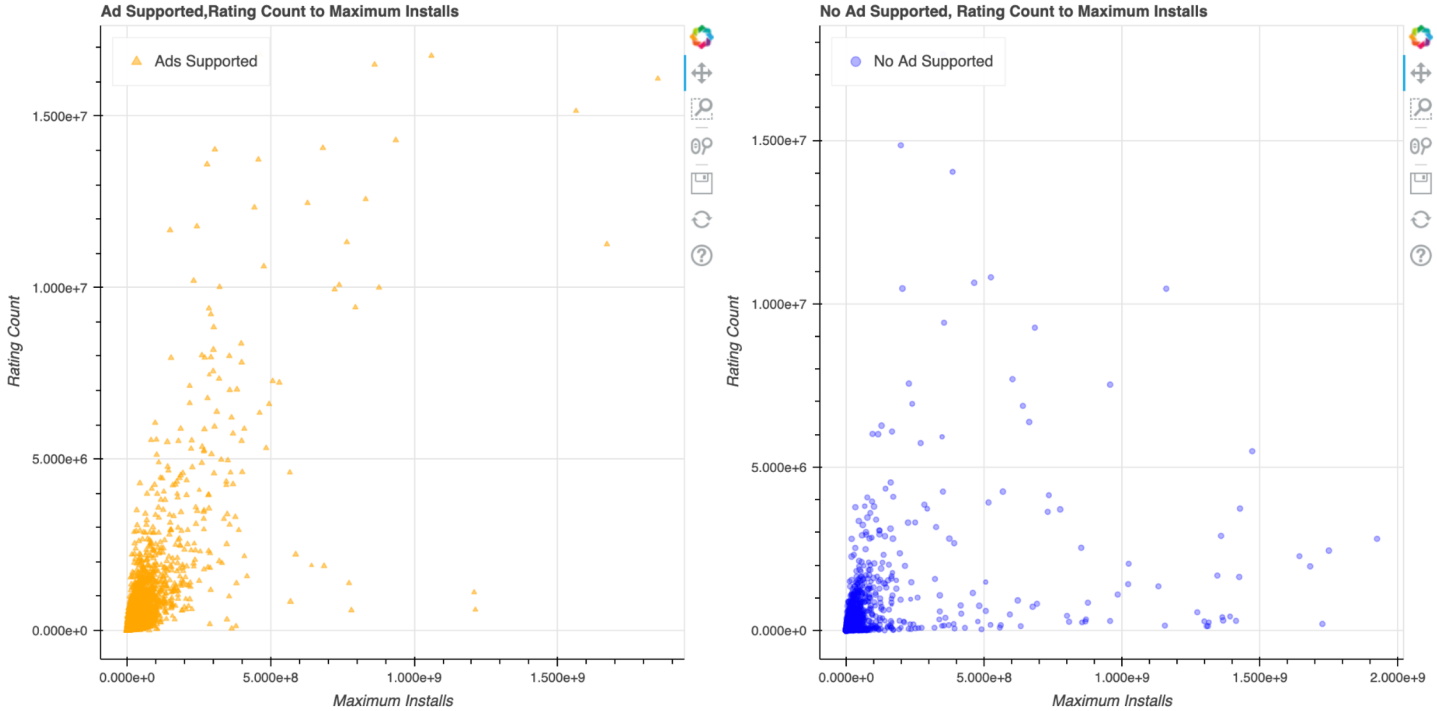


Figure 4. Bokeh Plot of Ads supported Apps Rating Count to Maximum Installs

Do applications with more installations get rated more based on whether it being free, ad supported, or an editor's choice?: Editor's Choice

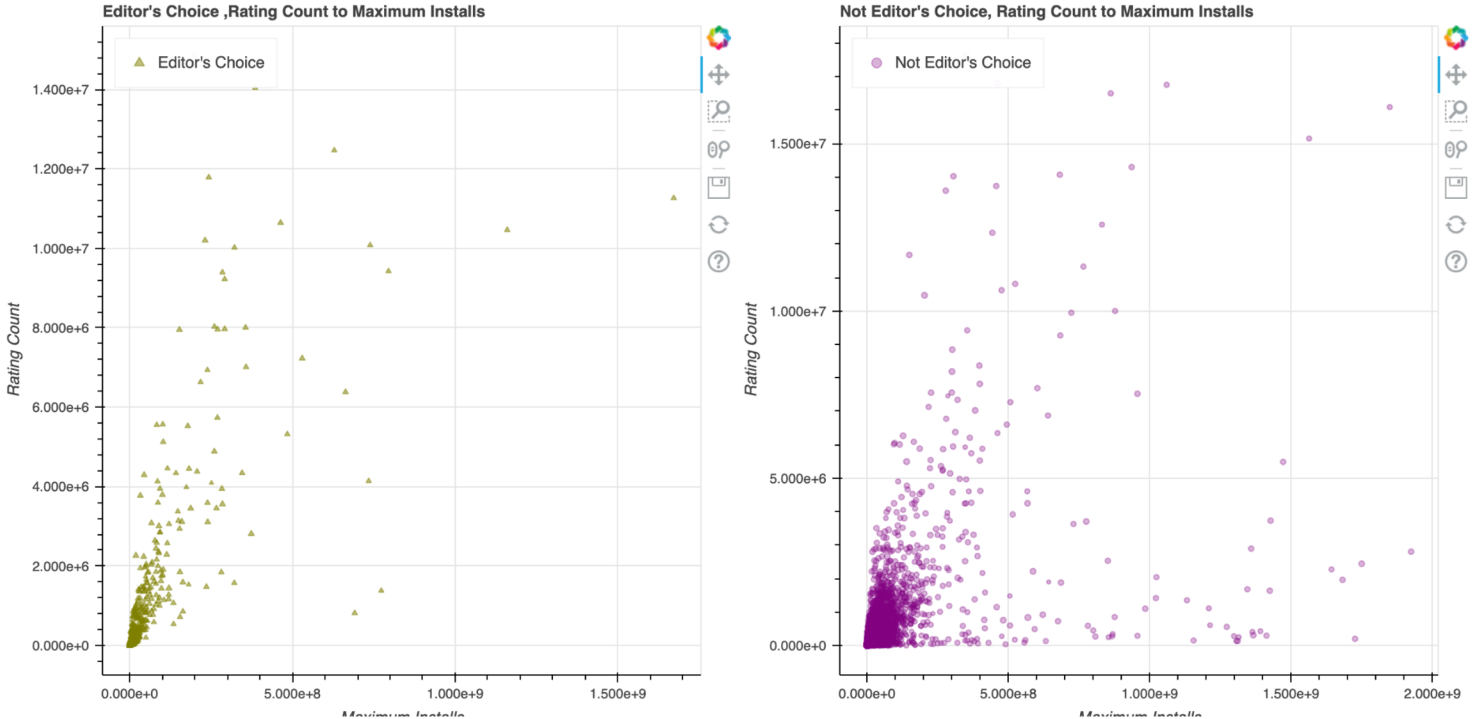


Figure 5. Bokeh Plot of Editor's Choice Apps Rating Count to Maximum Installs

Are there specific categories of apps that receive better ratings, on average?

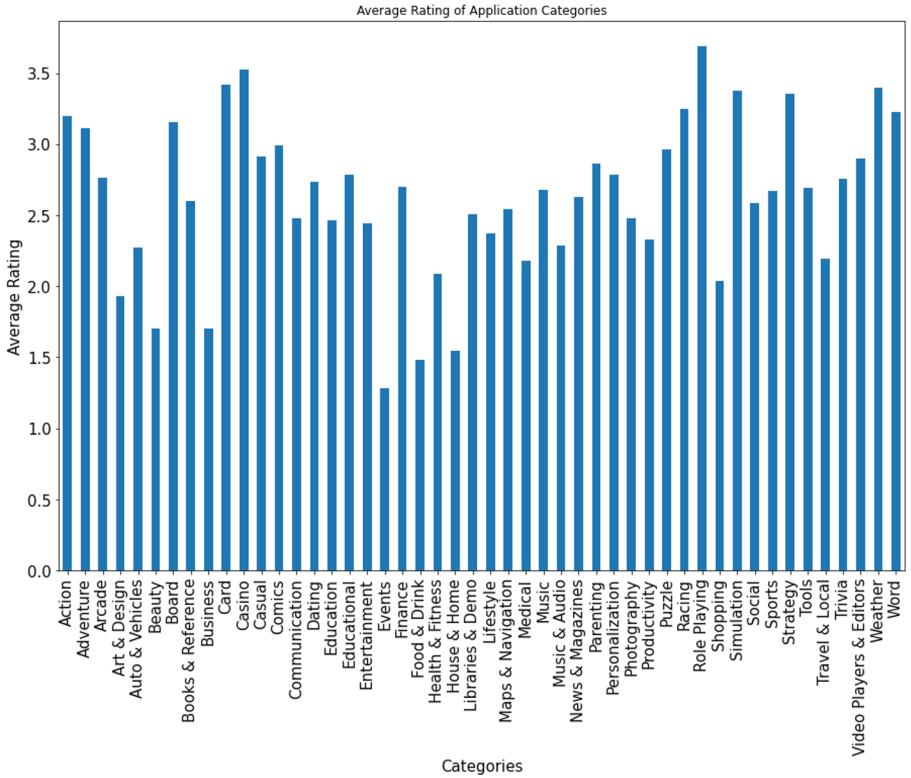


Figure 6. Average Rating of Application Categories

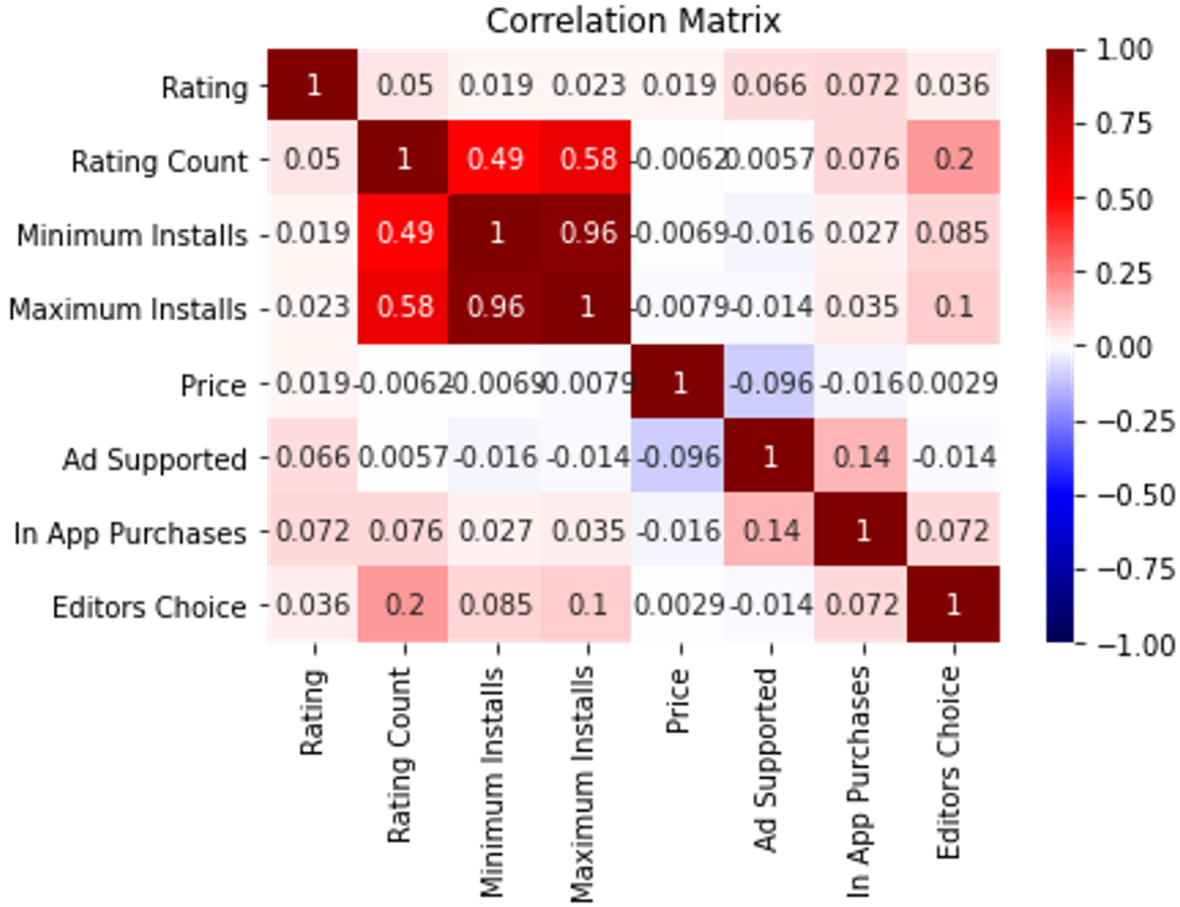


Figure 7. Correlation Matrix

Machine Learning Model

Used K Nearest Neighbor Machine learning to predict if certain data features could identify the type of category it belongs to.

Chose this ML method because the data appears to be non-linear which KNN supports.

We split the training/testing size to a ratio of 80/20 and 70/30.

We also used logistic regression to see if there was a trend of any kind. The LR returned an accuracy score of about 7.5 %

Figure 8: KNN 80/20

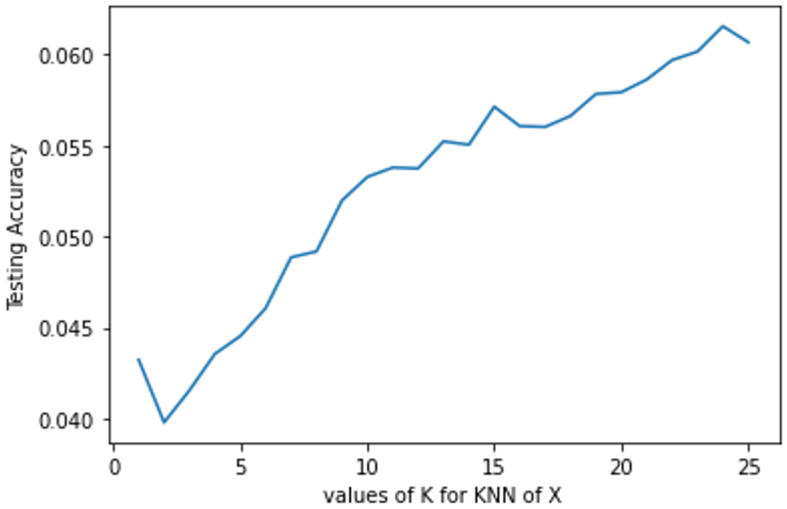
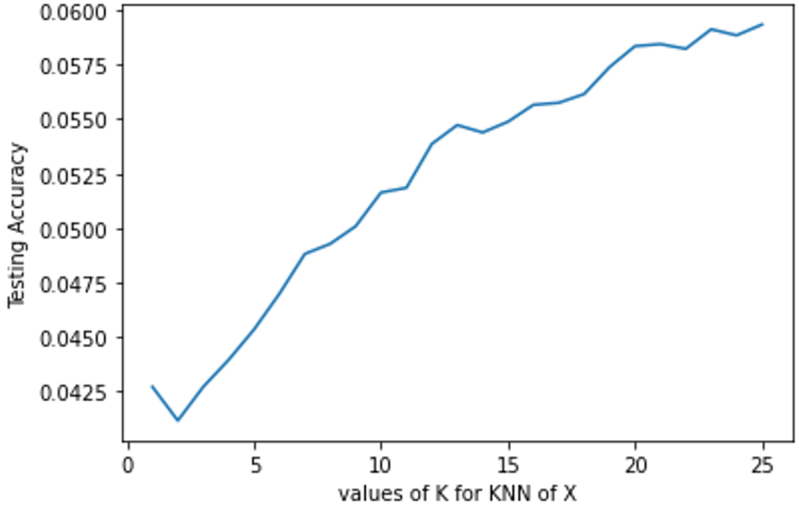
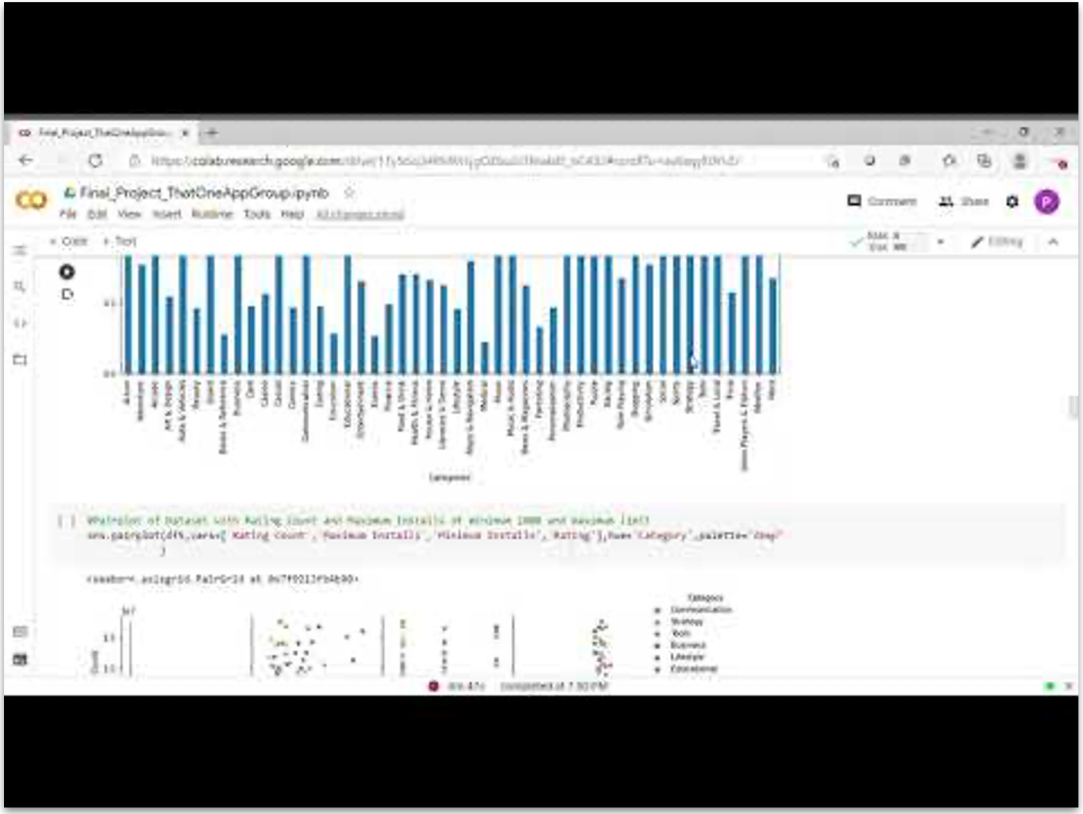


Figure 9: KNN 70/30



How we used Google Collaboratory



Summary of Current Observations

What strongly influences the rating of an application?

- With the current data, nothing suggests a strong influence over the rating of an application

Do applications with more installations get rated more based on whether it being free, ad supported, or an editor's choice?

- It's difficult to conclude since there is an unequal balance of data for free, ad supported, or editor's choice

Do free or paid apps receive higher ratings?

- Free apps have receive higher ratings in majority than paid apps

Are there specific categories of apps that receive better ratings, on average?

- On average, role playing, casino, and card receive better ratings

Can the data predict the type of category?

- After some machine learning, we find that the prediction accuracy of the prediction is really low, which indicates random data.

Issues with Dataset

Problems with Data

- Large database until it was reduced and more manageable to code
- Too many outliers that distort graphs

How to improve dataset for future use

- By grouping genre of categories to have a more straightforward graph instead of the subdivided categories

Are there patterns or anomalies?

- Patterns and Anomalies:
  - No patterns
  - On the Bokeh Scatter Plot of Paid vs Free Apps, it has an anomaly around the 4.5 range on the paid apps chart

Prescriptive analysis

With such a large data that has weak correlations, what can be done is very limited.

This project could be used as reference for future App development.

The project shows some correlation between rating and whether it has In app purchases, Ad Supported, or Editor’s choice. This information may help developers decide whether or not to include these features.

The probable inclusions of features may benefit a developers or an organizations by improving the chances the app would be downloaded and given high ratings.