Product Life Cycle Group Project

Google PlayStore Android App Data

Ateeksha Chaudhary, Jai Agrawal, Jennie Ryu, Prashasti Sharma, Patrick Yip, Tanisha Arora

BUSINESS GOAL

1. Determine the factors that are significant in predicting a high rating on the Google PlayStore

2. Determine the best product decision through data analysis and modeling

DATA

- Data Source : Kaggle
- Data URL : https://www.kaggle.com/datasets/gauthamp10/google-playstore-apps?resource=download
- Number of Observations
  - Columns : 24
  - Rows : 2,312,944

DEPENDENT & INDEPENDENT VARIABLES

- Dependent Variable: “Maximum Installs”
- Independent Variable: 17 variables(14 Original + 3 Calculated)
- **Calculated Variables**
  - Days since App release = (Scraped Date - Release Date)
  - Days since App update = (Scraped Date - Last Updated Date)
  - Rating Density = (Rating Count/Maximum Installs)

DATA DICTIONARY-1

#	Variable	Type	Description	
1	AppName	String	Name of the app	
2	Rating	Int	Average rating	
3	Rating Count	Int	Number of rating	
4	Maximum Installs	Int	Approximate maximum app install count	***Dependent variable***
5	Free	Boolean	Whether app is Free or Paid	
6	Price	Int	App price	
7	Size	Int	Size of application package	
8	Editors Choice	Boolean	Whether rated as Editor Choice	
9	In-App purchase	Boolean	In-App purchases in app	
10	Content Rating	Object	Maturity level of app	

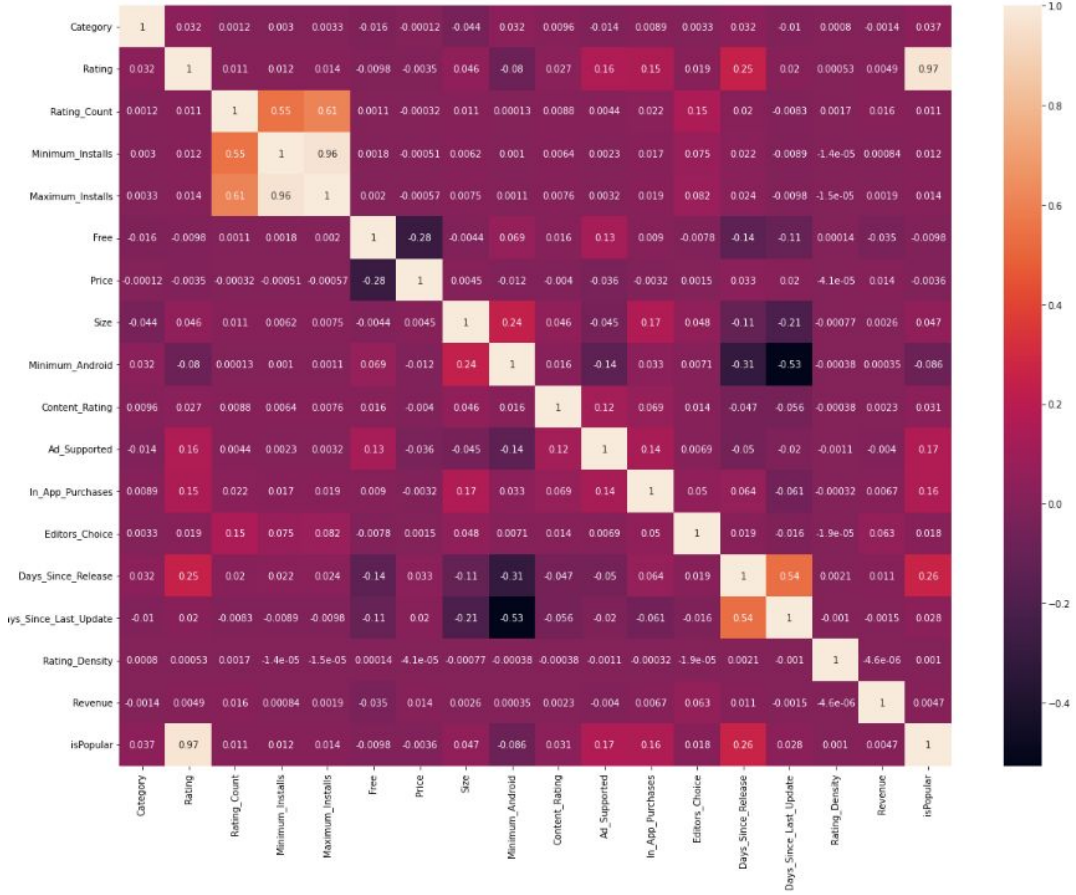
DATA DICTIONARY-2

#	Variable	Type	Description
11	Category	Object	App category
12	Released	Date	App launch date on Google Playstore
13	Ad Supported	Boolean	Ad support in app
14	Last Updated	Date	Last app update date
15	Scraped time	DateTime	Scraped date-time in GMT
16	Days since App Release	Int	Scraped Date - Release Date
17	Days since Last Update	Int	Scraped Date - Last Updated Date
18	Rating Density	Int	Rating Count/Maximum Installs
19	isPopular	Boolean	Whether the app is popular or not; Rating < 2.5 App is not popular

Calculated variable

Interactive System

Correlation Matrix



Models - Linear Regression (Implemented on SAS)

Model: MODEL1
Dependent Variable: Maximum_Installs

Number of Observations Read	1013035
Number of Observations Used	1008064
Number of Observations with Missing Values	4971

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	4.118448E20	2.57403E19	984764	<.0001
Error	1.01E6	2.634887E19	2.613853E13		
Corrected Total	1.01E6	4.381936E20			

Root MSE	5112586	R-Square	0.9399
Dependent Mean	307412	Adj R-Sq	0.9399
Coeff Var	1663.10618		

Models - Logistic Regression (Implemented on Python)

```
#Accuracy
y_pred = logreg.predict(x_test_logistic)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(x_test_logistic, y_test_logistic)))
```

```
Accuracy of logistic regression classifier on test set: 0.73
```

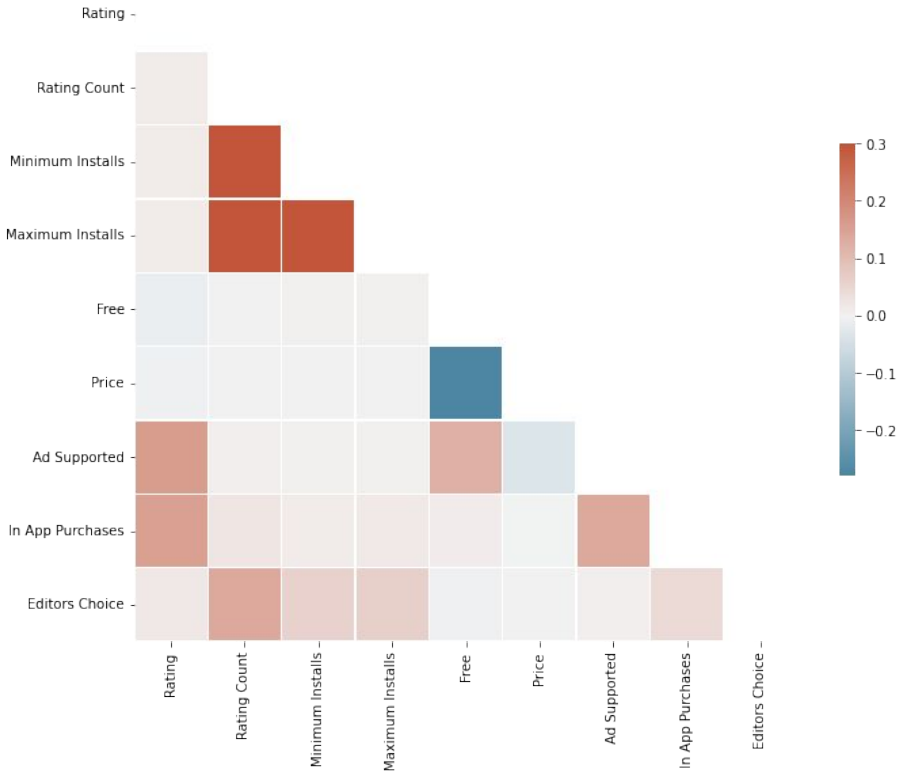
```
# Confusion Matrix
confusion_matrix = confusion_matrix(y_test_logistic, y_pred)
print(confusion_matrix)
```

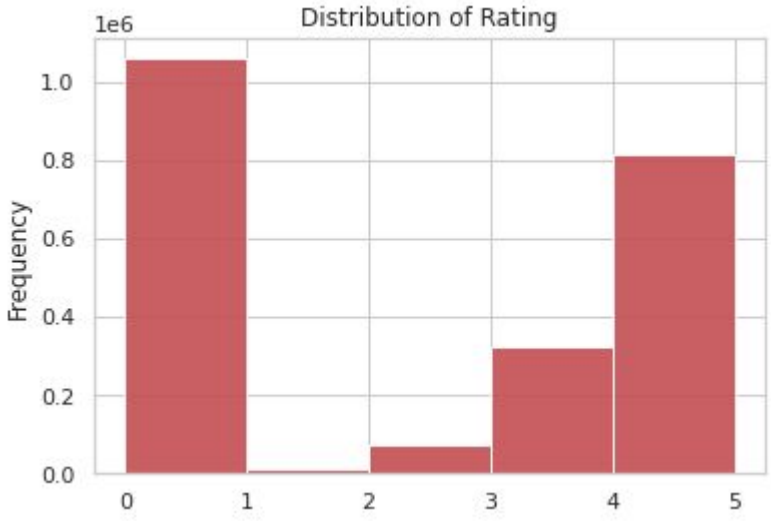
```
[[58786 13974]
 [26318 52878]]
```

```
# Classification
print(classification_report(y_test_logistic, y_pred))
```

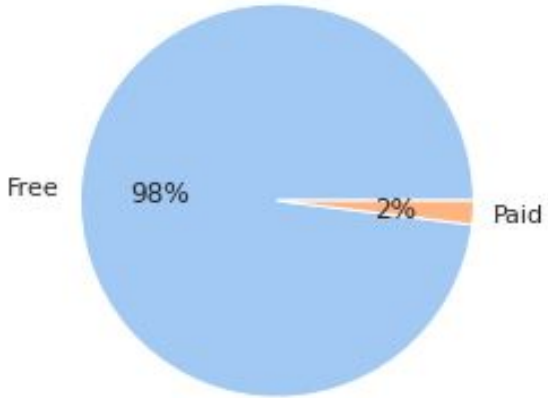
```
              precision    recall  f1-score   support

           0       0.69      0.81      0.74     72760
           1       0.79      0.67      0.72     79196

    accuracy                           0.73    151956
   macro avg       0.74      0.74      0.73    151956
weighted avg       0.74      0.73      0.73    151956
```

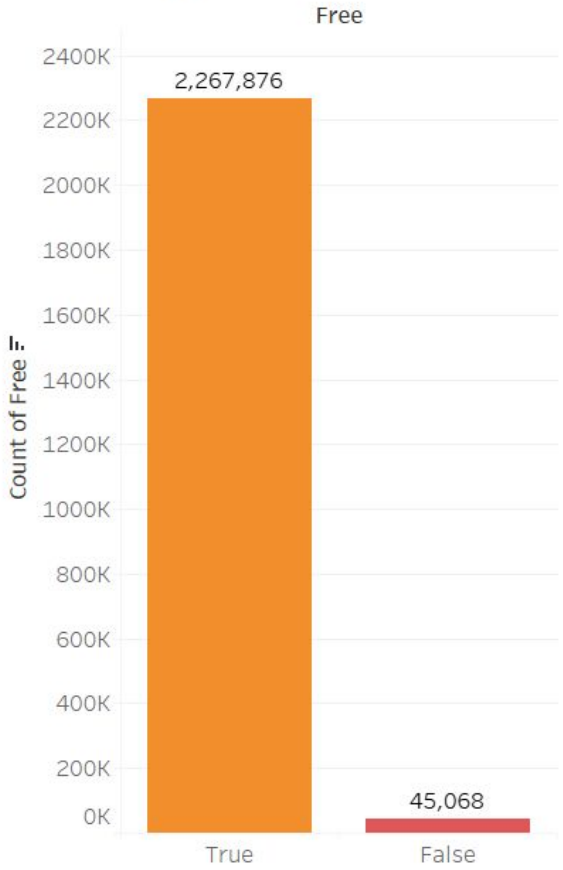
CORRELATION





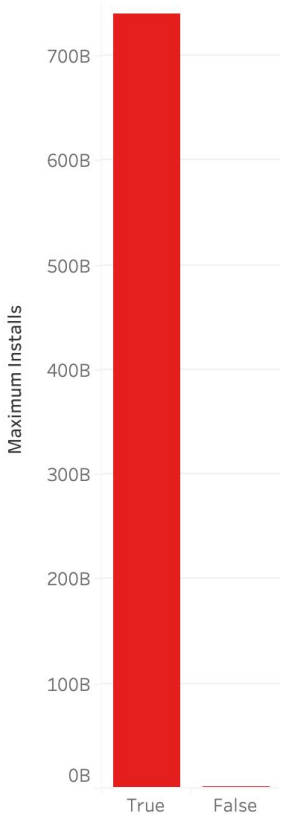
FREE vs PAID APP DISTRIBUTION



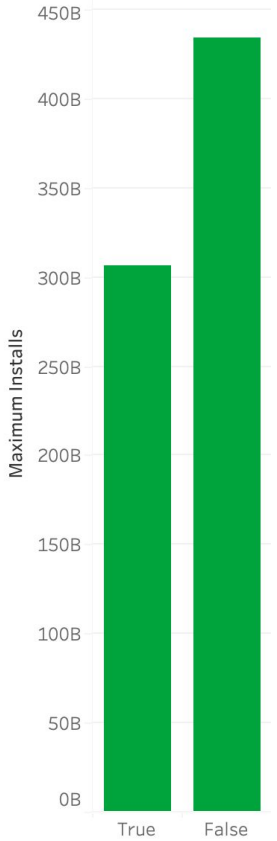


INSTALL DISTRIBUTION

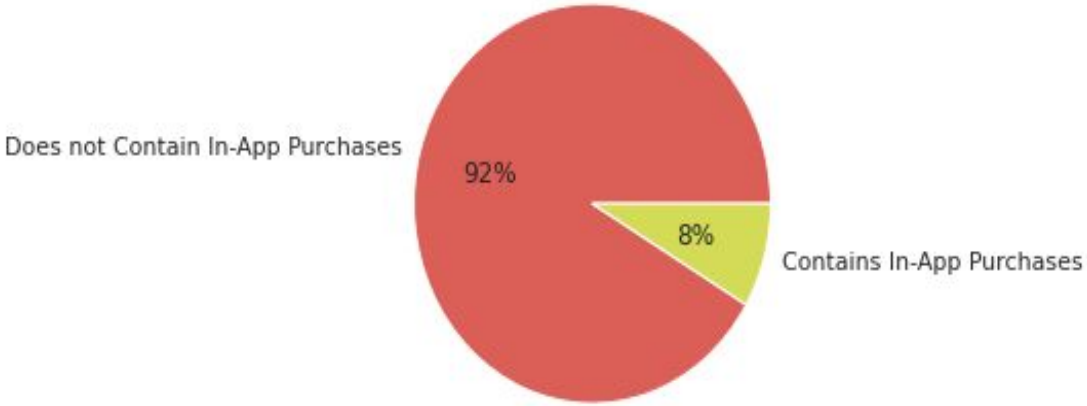
Free App & Installs

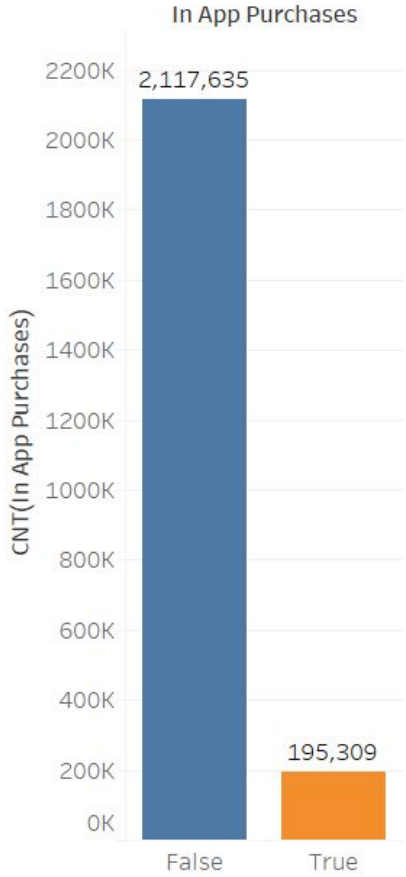


In App Purchase & Installs

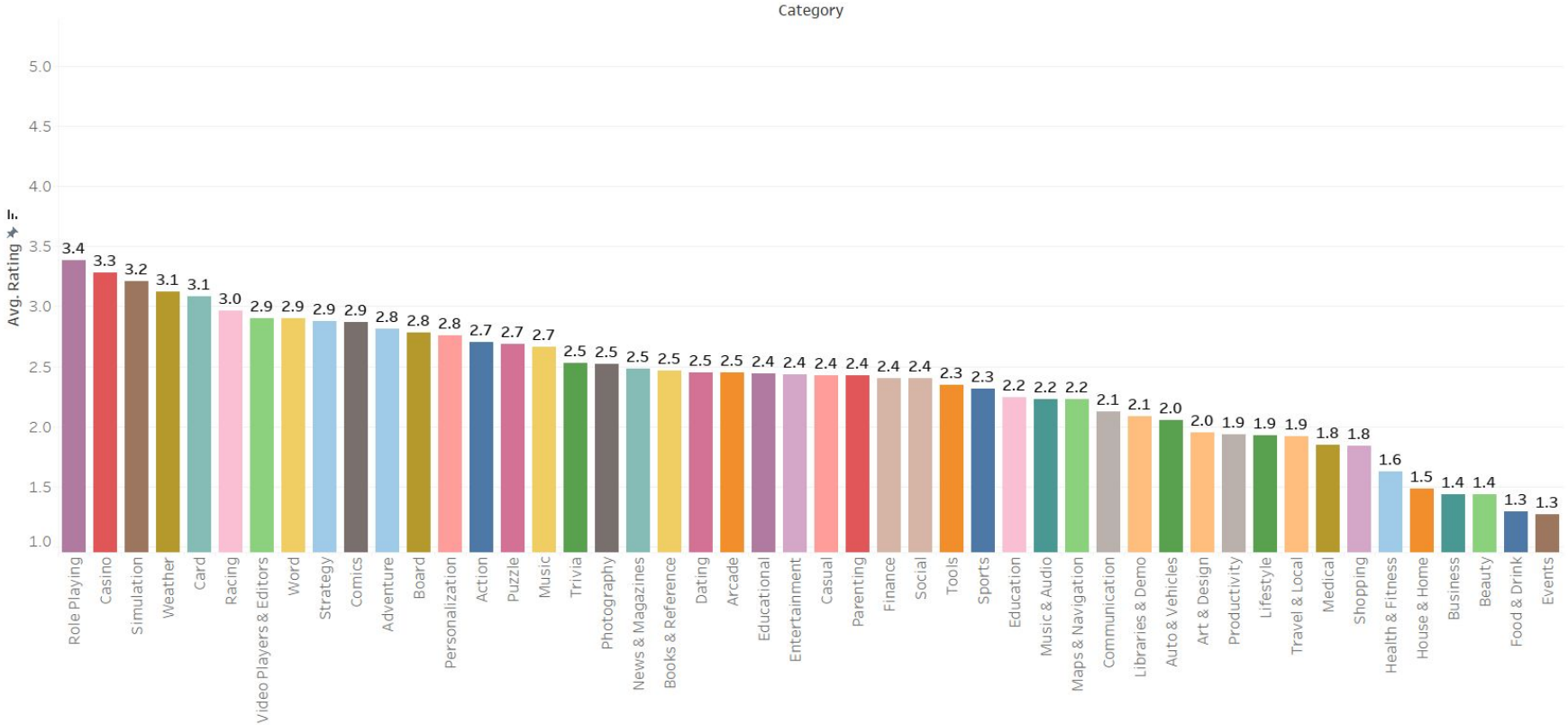


IN-APP PURCHASE AVAILABILITY BY % AND COUNT

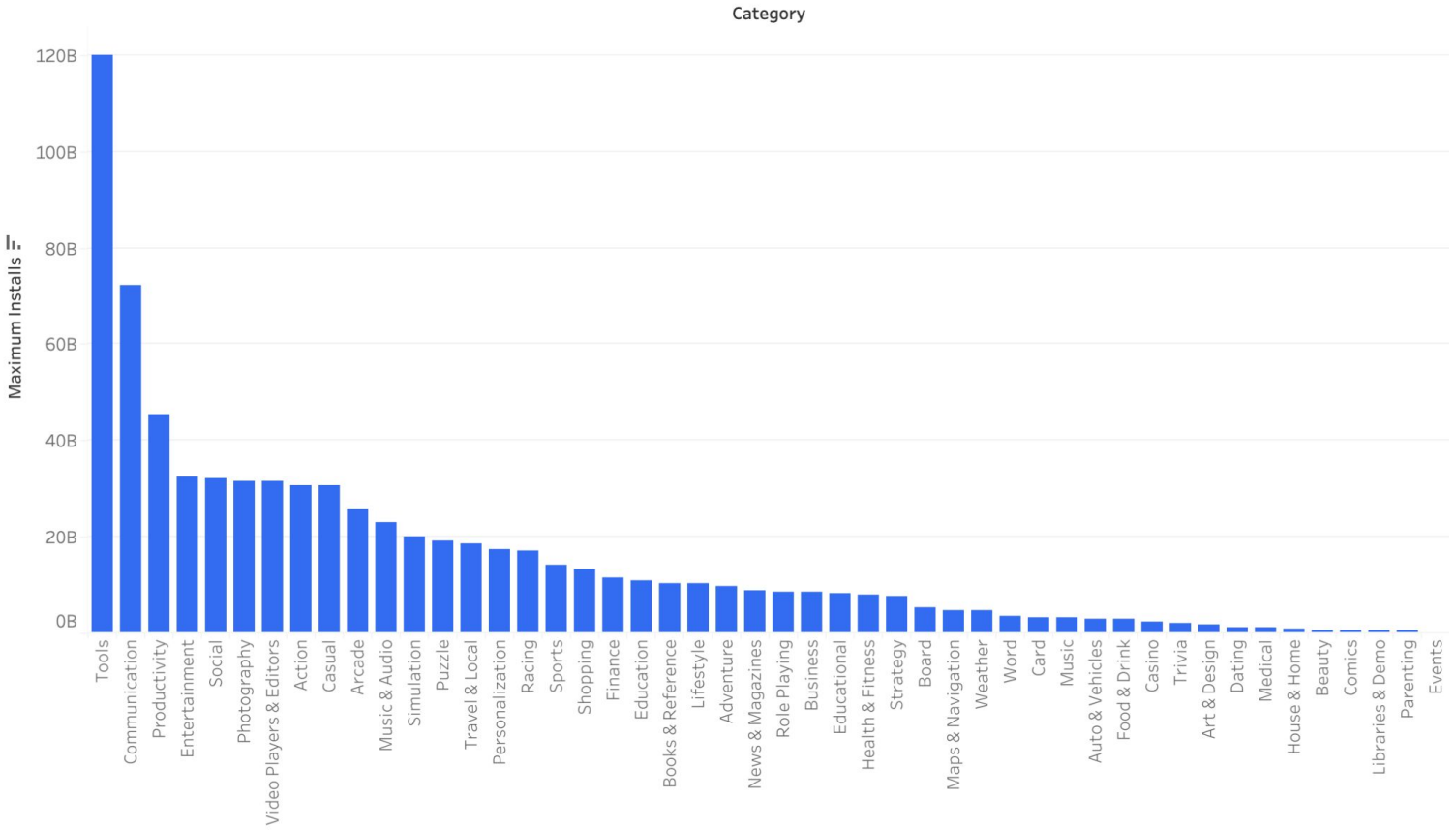




AVG RATING BY CATEGORY



MAX INSTALLS BY CATEGORY



DESCRIPTIVE STATISTICS

	Rating	Rating Count	Minimum Installs	Maximum Installs	Price
count	2290061.000	2290061.000	2312837.000	2312944.000	2312944.000
mean	2.203	2864.839	183445.214	320201.713	0.103
std	2.106	212162.571	15131439.060	23554954.887	2.633
min	0.000	0.000	0.000	0.000	0.000
25%	0.000	0.000	50.000	84.000	0.000
50%	2.900	6.000	500.000	695.000	0.000
75%	4.300	42.000	5000.000	7354.000	0.000
max	5.000	138557570.000	10000000000.000	12057627016.000	400.000