Rating of Google Play Store (Android market)

Nazanin Komeilizadeh

SpringBoard Capstone Project

- Dataset from Kaggle https://www.kaggle.com/lava18/google-play-store-apps

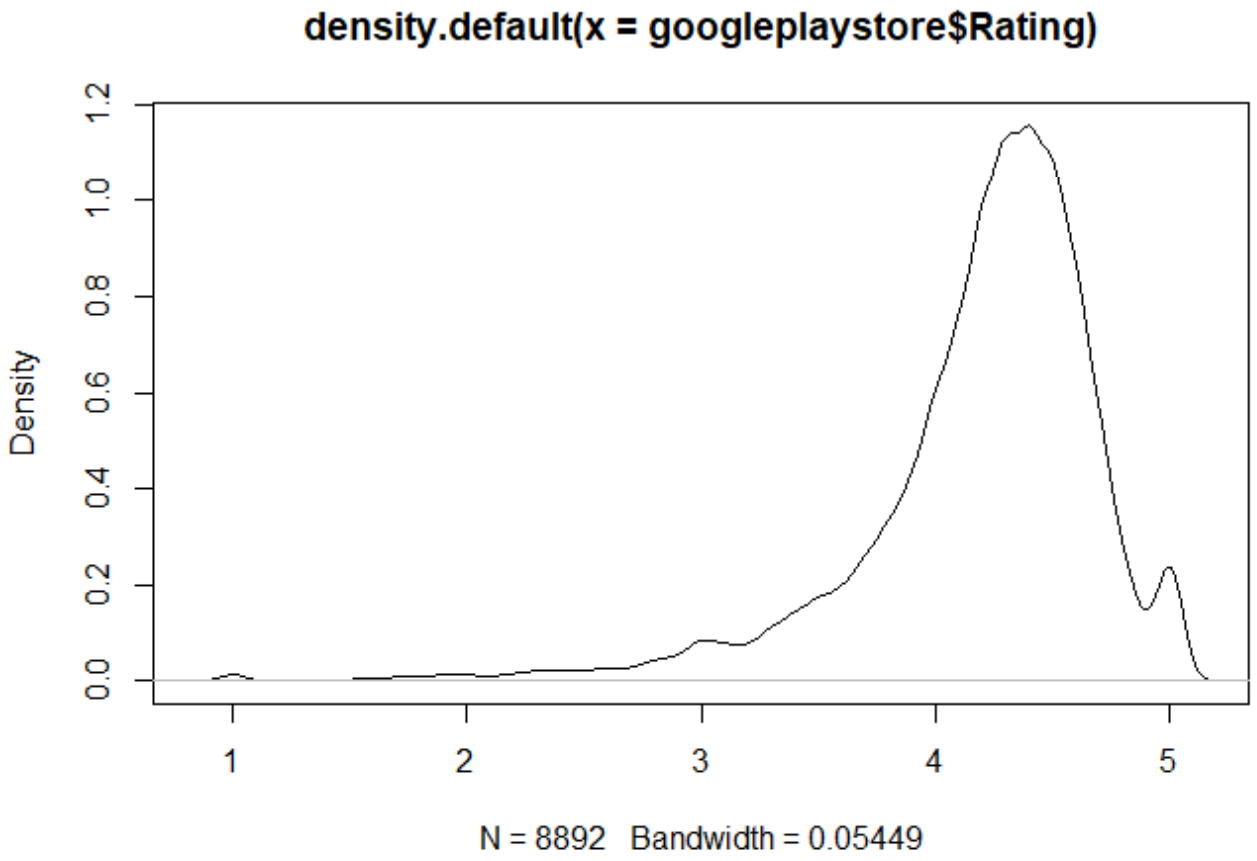
- Through various views of this dataset, we reveal patterns to assist companies/individuals in Android app development to focus on the most profitable strategy for app development

- We attempt to predict the rating of app based on random forest machine learning

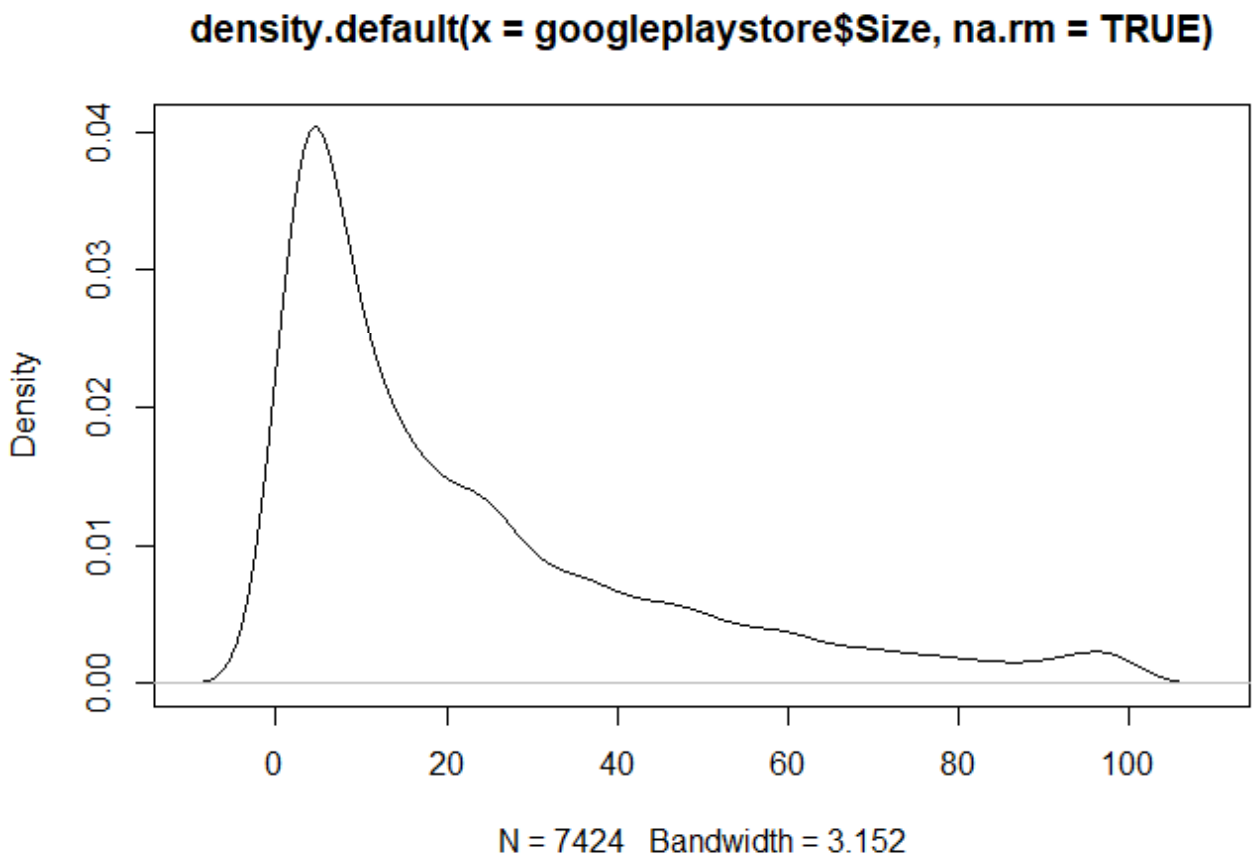
Data Wrangling

Density plot of `Rating`

Big portion of the `Rating` is between 4 and 5

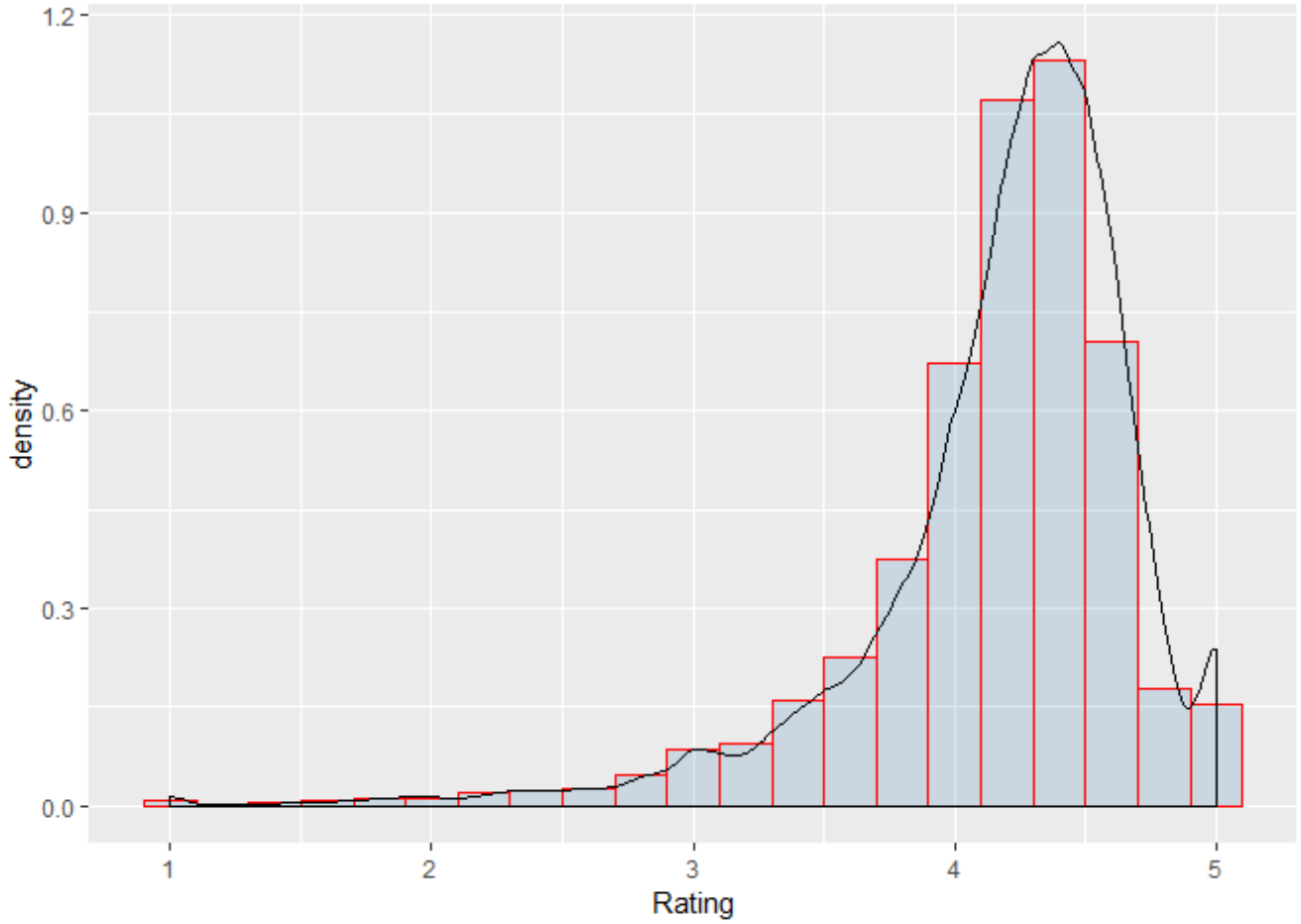


Density of the `Size` variable

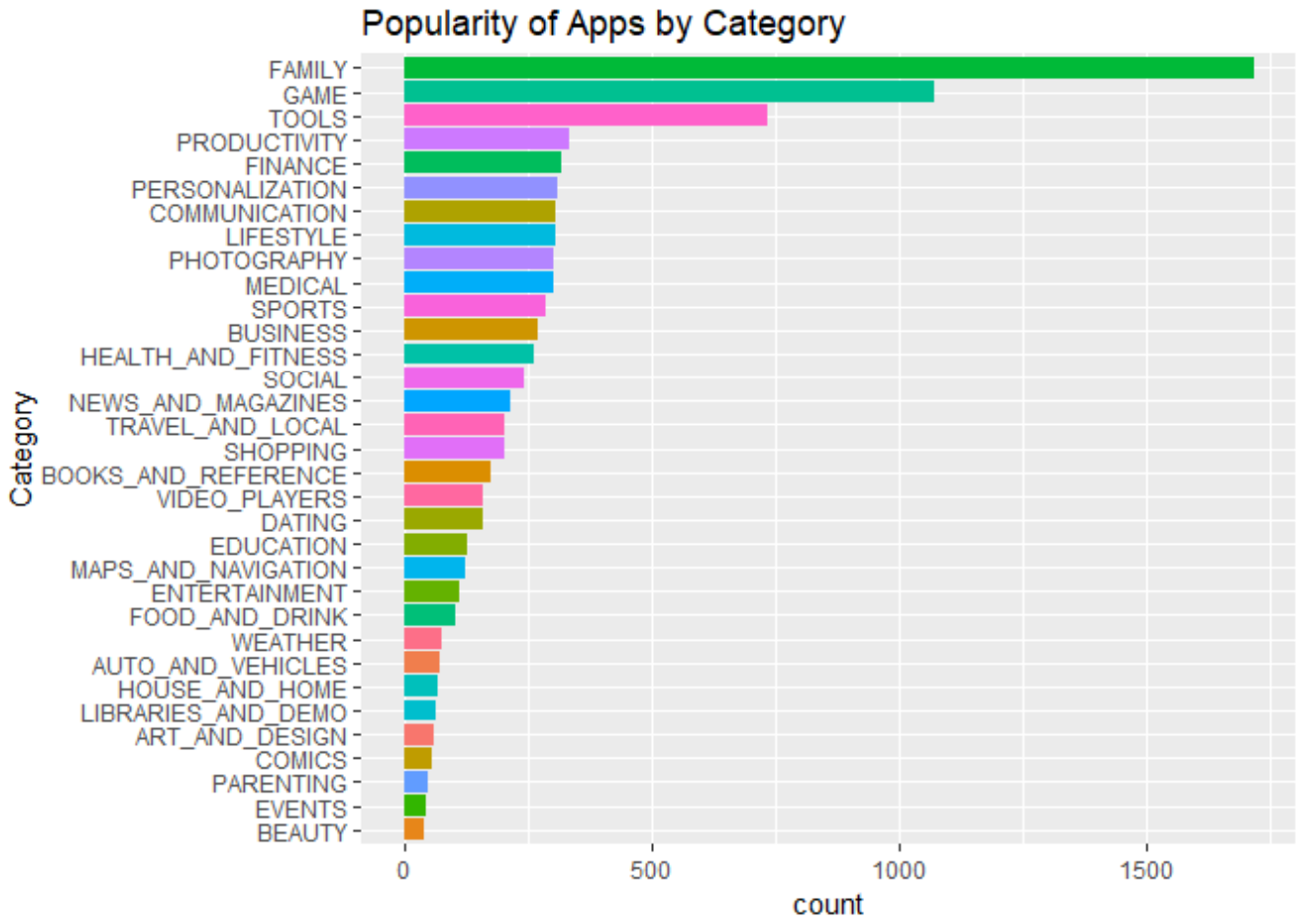


Exploratory Data Analysis

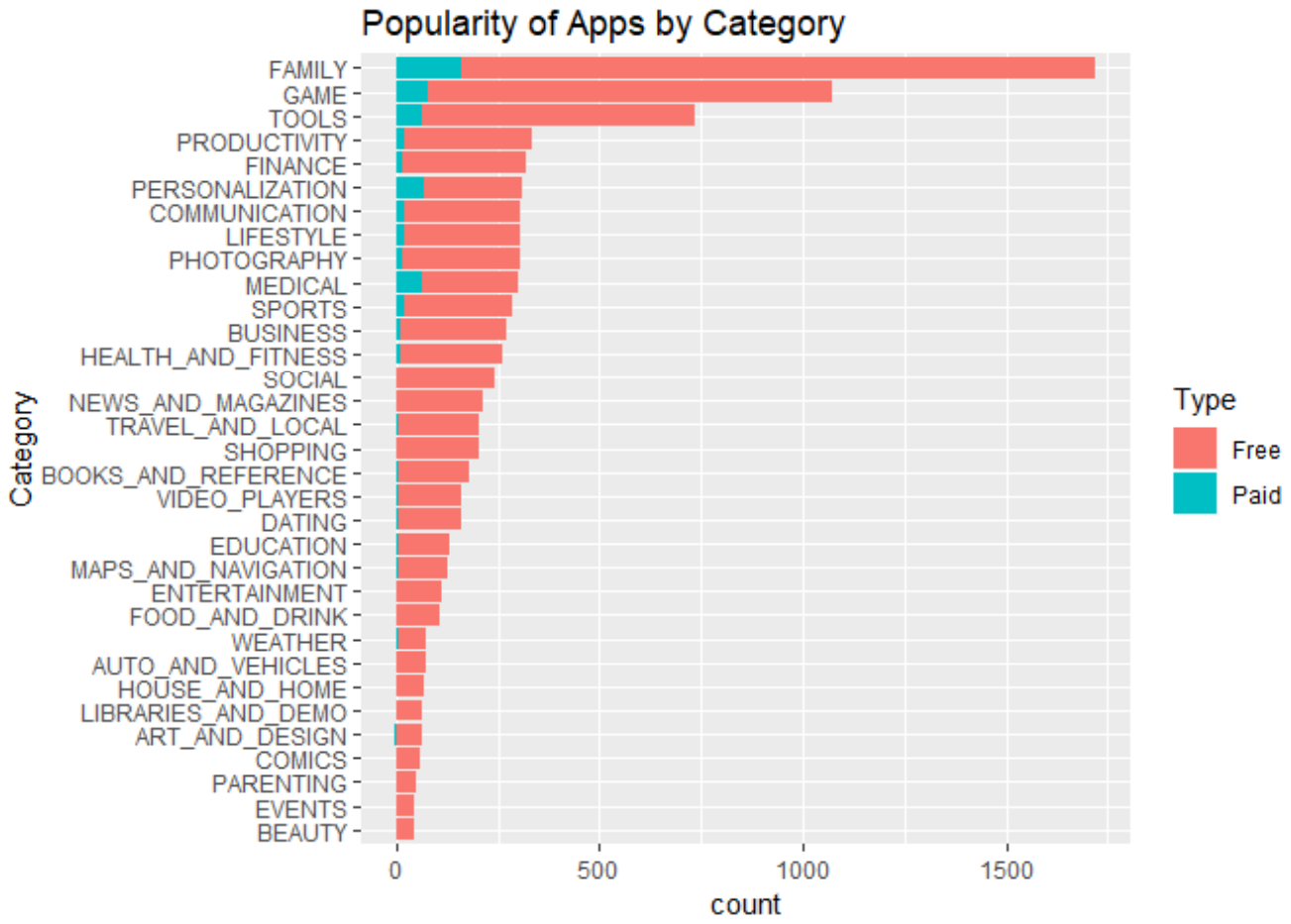
Histogram of Rating



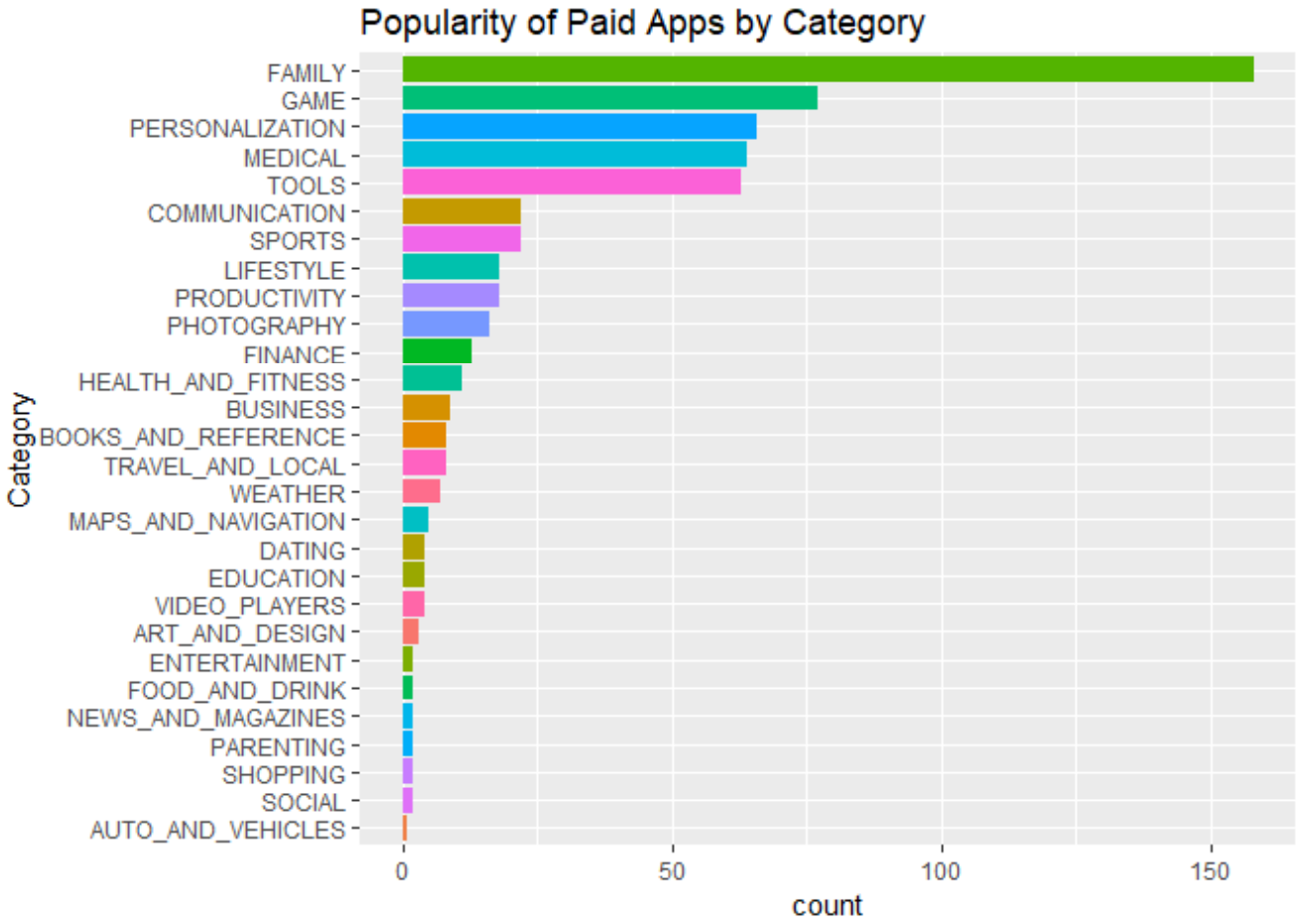
Popularity of Apps by Category



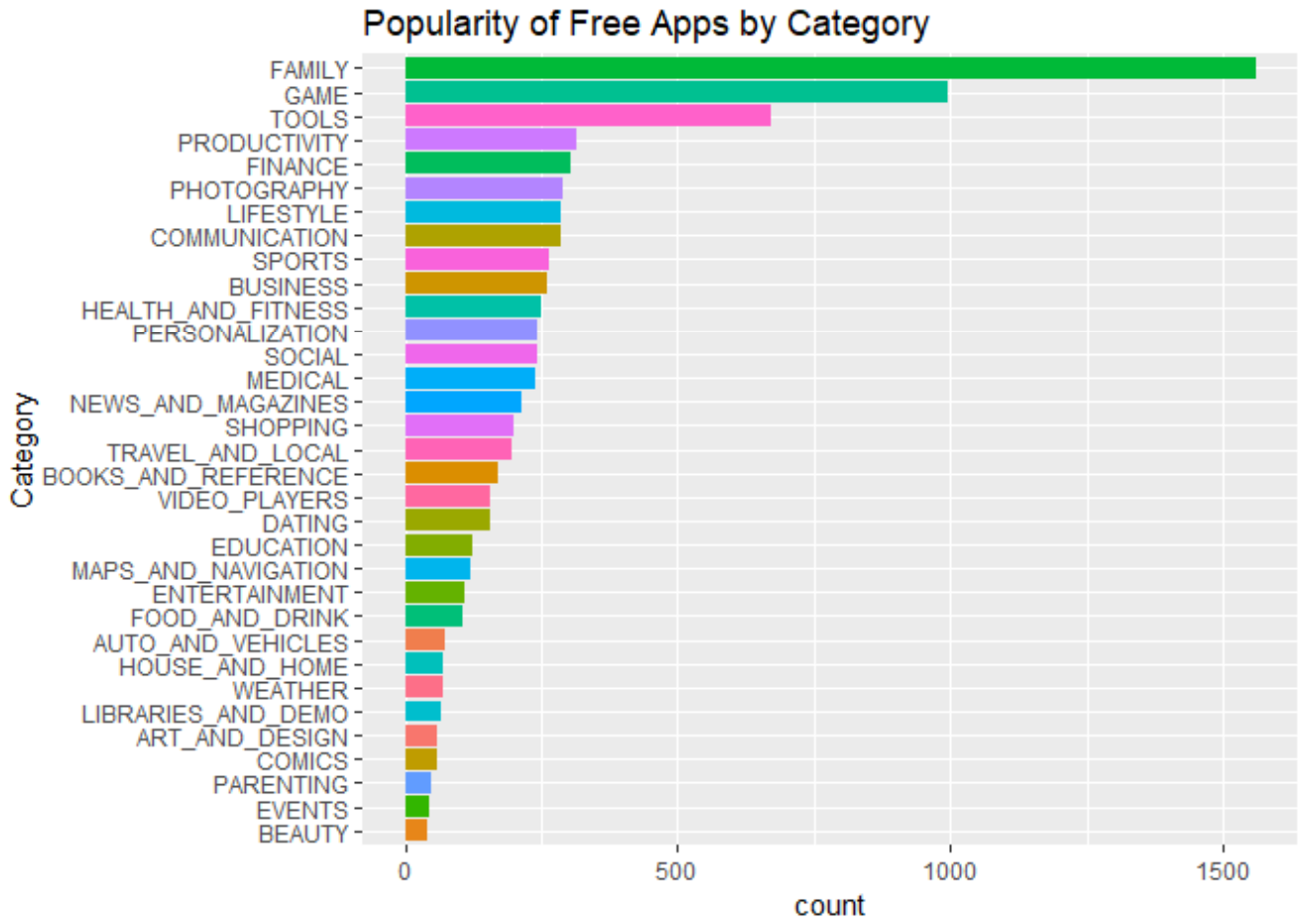
Popularity of Apps by Category



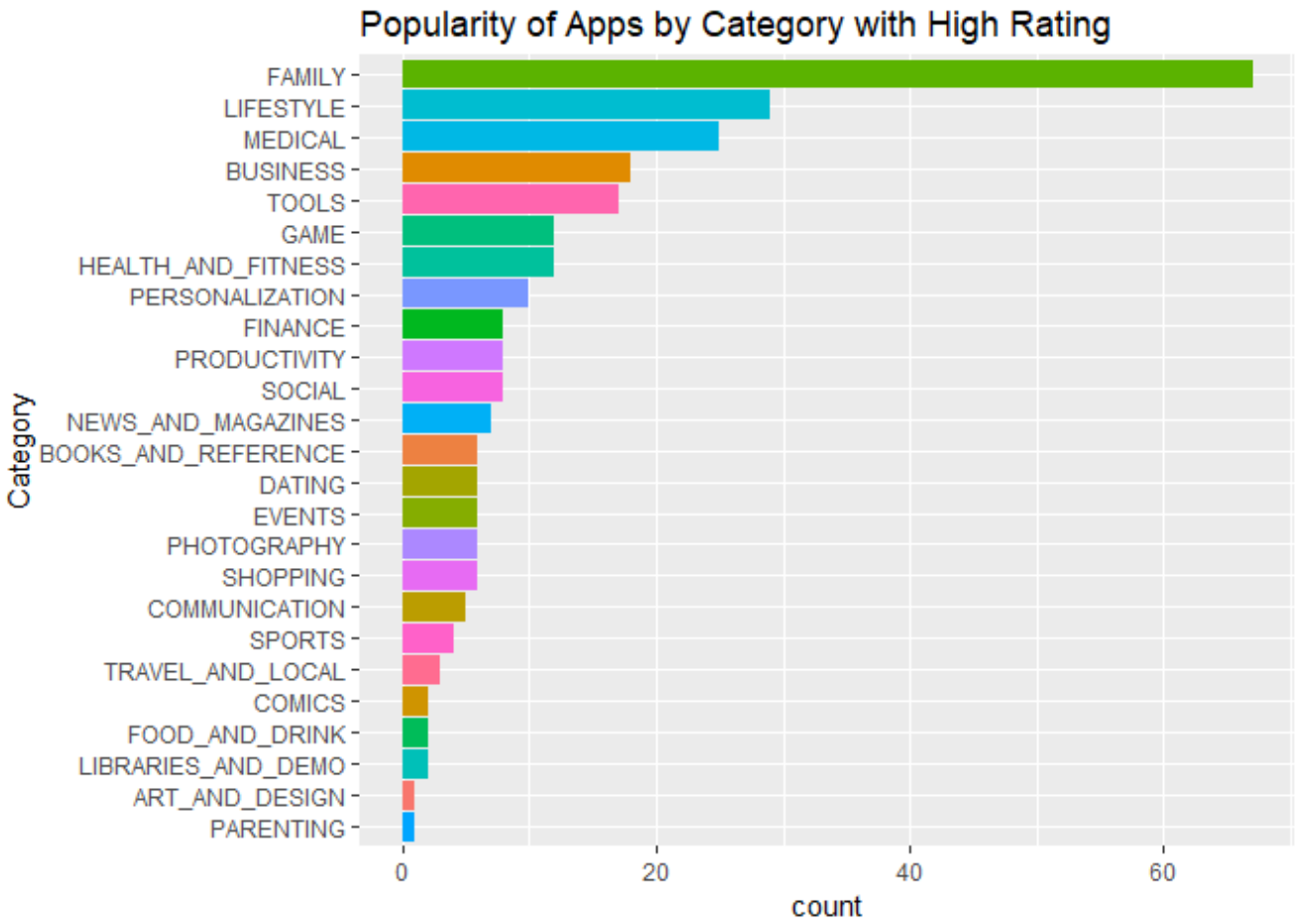
Popularity of Paid Apps by Category



Popularity of Free Apps by Category



Popularity of Apps by Category with High Rating (>4.9)

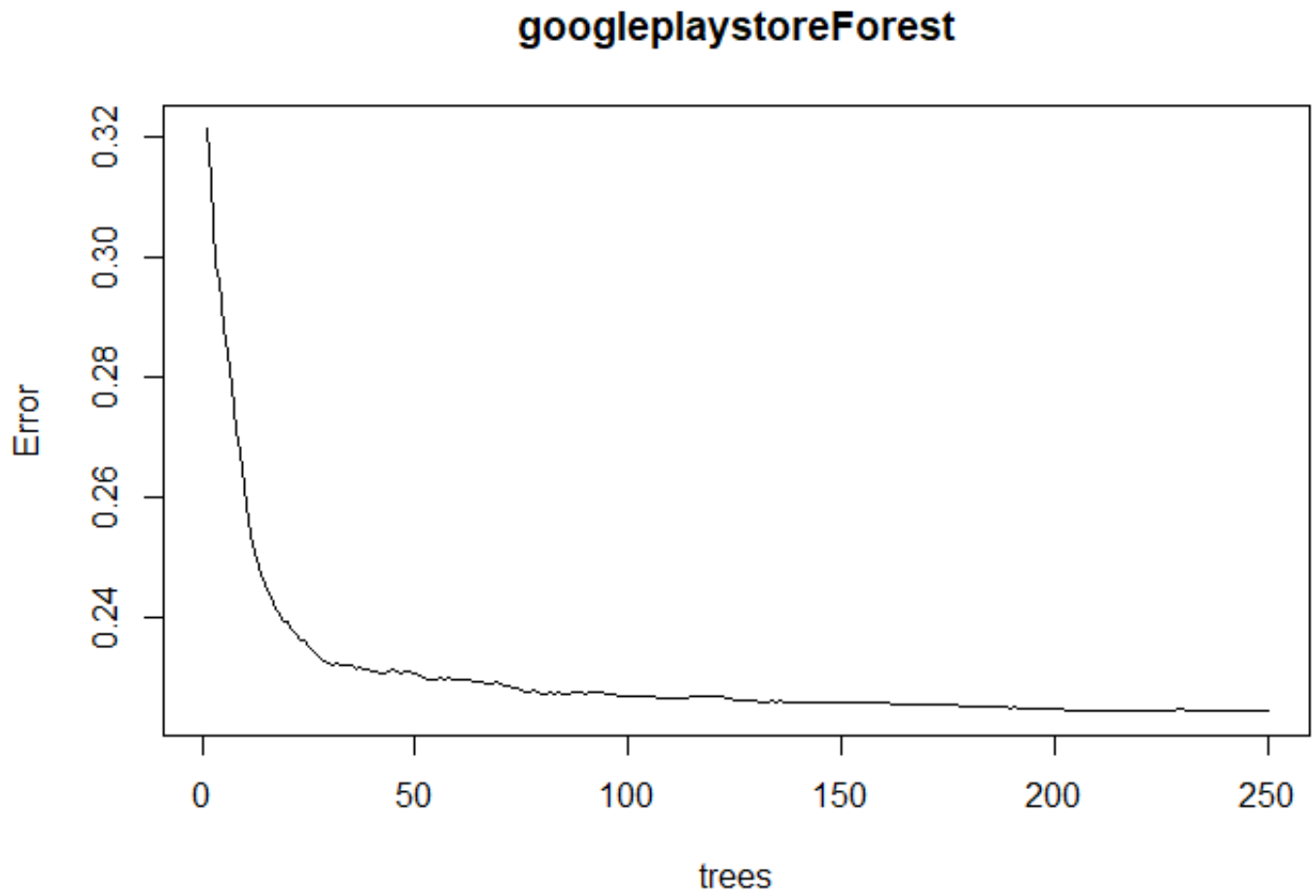


Machine Learning

- Dataset is a Supervised regression type
- We use `randomForest()` to predict the outcome variable `Rating` on other variables

```
     App                Category           Rating          Reviews                Size
 Length:8892        Length:8892        Min.   :1.000   Min.   :       1   Min.   :  0.0083
 Class :character   Class :character   1st Qu.:4.000   1st Qu.:     164   1st Qu.:  5.1000
 Mode  :character   Mode  :character   Median :4.300   Median :    4714   Median : 14.0000
                                       Mean   :4.188   Mean   :  472776   Mean   : 22.7473
                                       3rd Qu.:4.500   3rd Qu.:   71267   3rd Qu.: 33.0000
                                       Max.   :5.000   Max.   :78158306   Max.   :100.0000
                                                                          NA's   :1468
    Installs             Type              Price          Content Rating
 Min.   :1.000e+00   Length:8892        Min.   :  0.0000   Length:8892
 1st Qu.:1.000e+04   Class :character   1st Qu.:  0.0000   Class :character
 Median :5.000e+05   Mode  :character   Median :  0.0000   Mode  :character
 Mean   :1.649e+07                      Mean   :  0.9632
 3rd Qu.:5.000e+06                      3rd Qu.:  0.0000
 Max.   :1.000e+09                      Max.   :400.0000

    Genres           Last Updated         Current Ver         Android Ver
 Length:8892        Min.   :2010-05-21   Length:8892        Length:8892
 Class :character   1st Qu.:2017-09-21   Class :character   Class :character
 Mode  :character   Median :2018-05-28   Mode  :character   Mode  :character
                    Mean   :2017-11-21
                    3rd Qu.:2018-07-23
```

Random Forest



Classification Trees

- Classify `Rating` variable to different classes of "Bad“ (<= 3), (3<)"Moderate“(<=4),

  (4<) "Good" (<= 4.5) and "Excellent" (>4.5)

Bad	Excellent	Good	Moderate
362	1838	4570	2122

- We predict the outcome variable `RatingClass` based on other variable in the dataset

SMOTE Computation `SmoteClassif()`

- `RatingClass` in the dataset shows the four classes are imbalanced observations

Bad	Excellent	Good	Moderate
362	1838	4570	2122

- we now use the SMOTE (Synthetic Minority Oversampling Technique) for the imbalanced datasets to oversample the rare event

Summary of New Dataset

```
            Category        Reviews                Size                Installs          Type
FAMILY           :1591   Min.    :       1   Min.    :  0.0083   Min.    :1.000e+00   Free:6877
GAME             : 959   1st Qu.:      99   1st Qu.:  5.1000   1st Qu.:1.000e+04   Paid: 547
TOOLS            : 634   Median :    2067   Median : 14.0000   Median :1.000e+05
PERSONALIZATION  : 279   Mean    :  278774   Mean    : 22.7473   Mean    :7.824e+06
MEDICAL          : 277   3rd Qu.:   36895   3rd Qu.: 33.0000   3rd Qu.:1.000e+06
LIFESTYLE        : 273   Max.    :44893888   Max.    :100.0000   Max.    :1.000e+09
(Other)          :3411
     Price                    ContentRating                Genres                Version
Min.    :  0.000   Adults only 18+:   2   Tools           : 634   4.1 and up  :1864
1st Qu.:  0.000   Everyone       :5958   Education       : 471   4.0.3 and up:1153
Median :  0.000   Everyone 10+   : 299   Entertainment   : 455   4.0 and up  :1073
Mean    :  1.117   Mature 17+     : 332   Action          : 332   4.4 and up  : 731
3rd Qu.:  0.000   Teen           : 832   Personalization: 279   2.3 and up  : 558
Max.    :400.000   Unrated        :   1   Medical         : 277   5.0 and up  : 446
                                          (Other)         :4976   (Other)     :1599
   RatingClass
Bad      : 344
Excellent:1590
Good     :3617
Moderate :1873
```

Proportion of `Rating Class` observations before and after SMOTE

```
       Bad  Excellent       Good   Moderate
0.04633621 0.21417026 0.48720366 0.25228987

      Bad Excellent      Good  Moderate
0.2499326 0.2499326 0.2500674 0.2500674
```

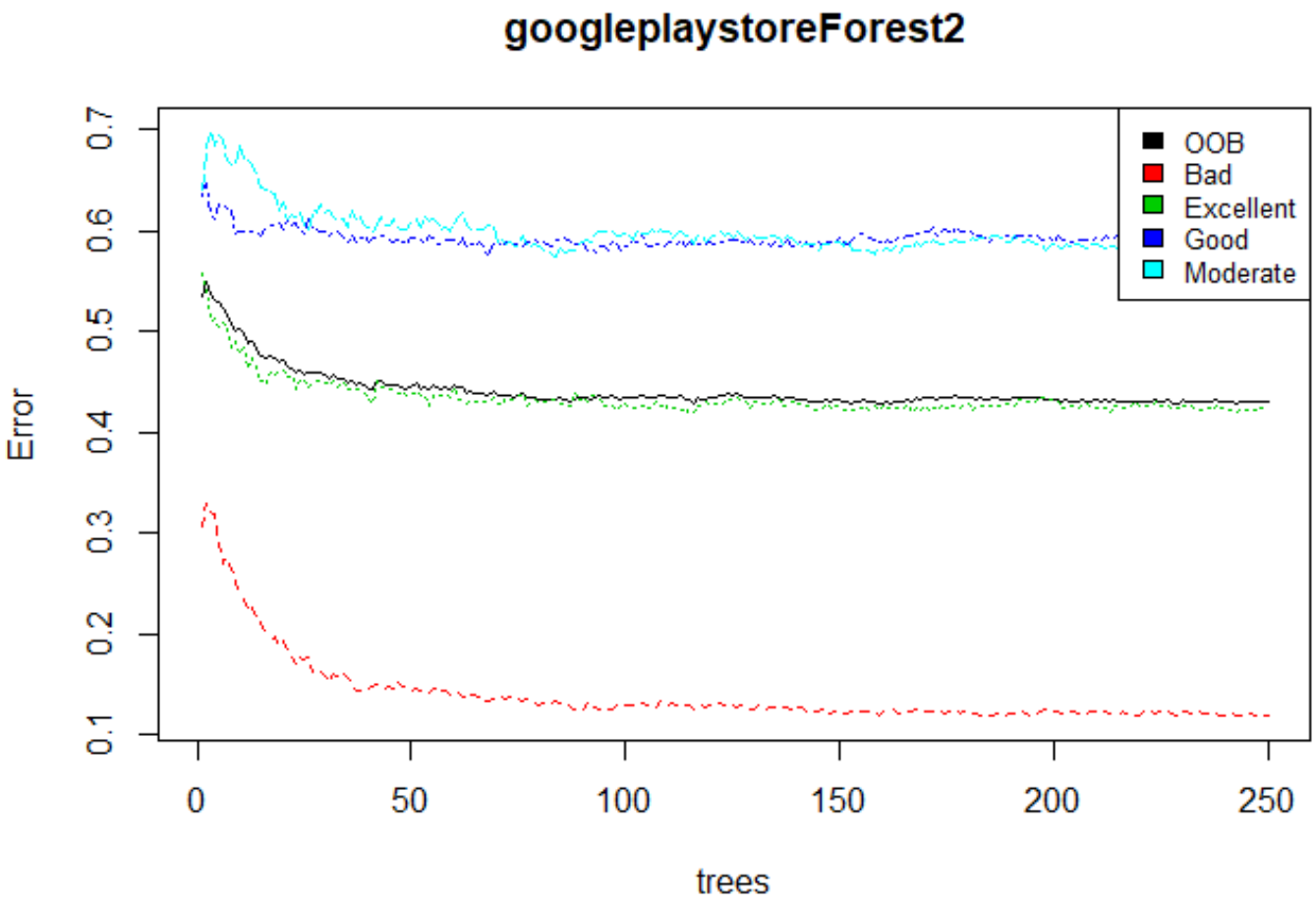
Random Forest on SMOTE(d) Dataset

```
Confusion Matrix and Statistics

            Reference
Prediction   Bad Excellent Good Moderate
  Bad        498        54   52      122
  Excellent   32       327  138       71
  Good         6        95  201      110
  Moderate    21        81  166      254

Overall Statistics

                Accuracy : 0.5745
                  95% CI : (0.5537, 0.5952)
    No Information Rate : 0.25
    P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.4327
 Mcnemar's Test P-Value : < 2.2e-16
```

Statistics by Class:

	Class: Bad	Class: Excellent	Class: Good	Class: Moderate
Sensitivity	0.8941	0.5871	0.36086	0.4560
Specificity	0.8636	0.8558	0.87373	0.8396
Pos Pred Value	0.6860	0.5757	0.48786	0.4866
Neg Pred Value	0.9607	0.8614	0.80396	0.8224
Prevalence	0.2500	0.2500	0.25000	0.2500
Detection Rate	0.2235	0.1468	0.09022	0.1140
Detection Prevalence	0.3259	0.2549	0.18492	0.2343
Balanced Accuracy	0.8788	0.7214	0.61730	0.6478

Random Forest on SMOTE(d) Dataset



Conclusion and Outlook

- The aim of this project is to carry out extensive data exploration and machine learning on GooglePlayStore dataset to reveal insights for the Android App development sphere

- EDA of dataset revealed that Popularity of Apps by Category is lead by Family and Game followed by Tools and Productivity

- Types of Free and Paid apps show that Medical and Personalization apps are the two categories with substantial number of paid apps, despite the fact that Family, Games and Tools are still the top three of the Paid and Free apps

- Considering apps with high Rating (Rating of greater than 4.9 out of 5) Family, Lifestyle and Medical are the top three of the apps

Conclusion and Outlook

- Machine Learning of supervised regression dataset through randomForest() predicts the outcome of “Rating” based on the other variables.
- Random Forest shows, `ntree = 250` to be a good tuning parameter for the number of trees in the Random Forest model. Moreover, the regression RF shows a smaller error rate compared to the CART model once the error rate was plotted
- Random Forest on Classification trees and by classifying `Rating` variable into "Bad", "Moderate", "Good" and "Excellent" rating classes; using SMOTE computation, the results show the best prediction for predicting the "Bad" class followed by "Excellent", "Moderate" and "Good" class with an overall accuracy of 0.5929.
- Further investigations were performed by reducing the number of independent variables to `Category`, `Reviews`, `Size`, `Installs`, `ContentRating` and `Genres`, however, the overall accuracy proved to reduce to 0.5624 so no further improvements were shown.