



Google Play

# Google Play Store App Preference Strategy

Xinlian Huang  
Jiahui Bi  
Hashem Orabee  
Ping-Lun Yeh

BIA-672 Marketing Analytics

1. Introduction
2. Related Research
3. Data Overview
4. Exploratory Data Analysis
5. Methodology Approach
6. Conclusion



# Content

# Introduction

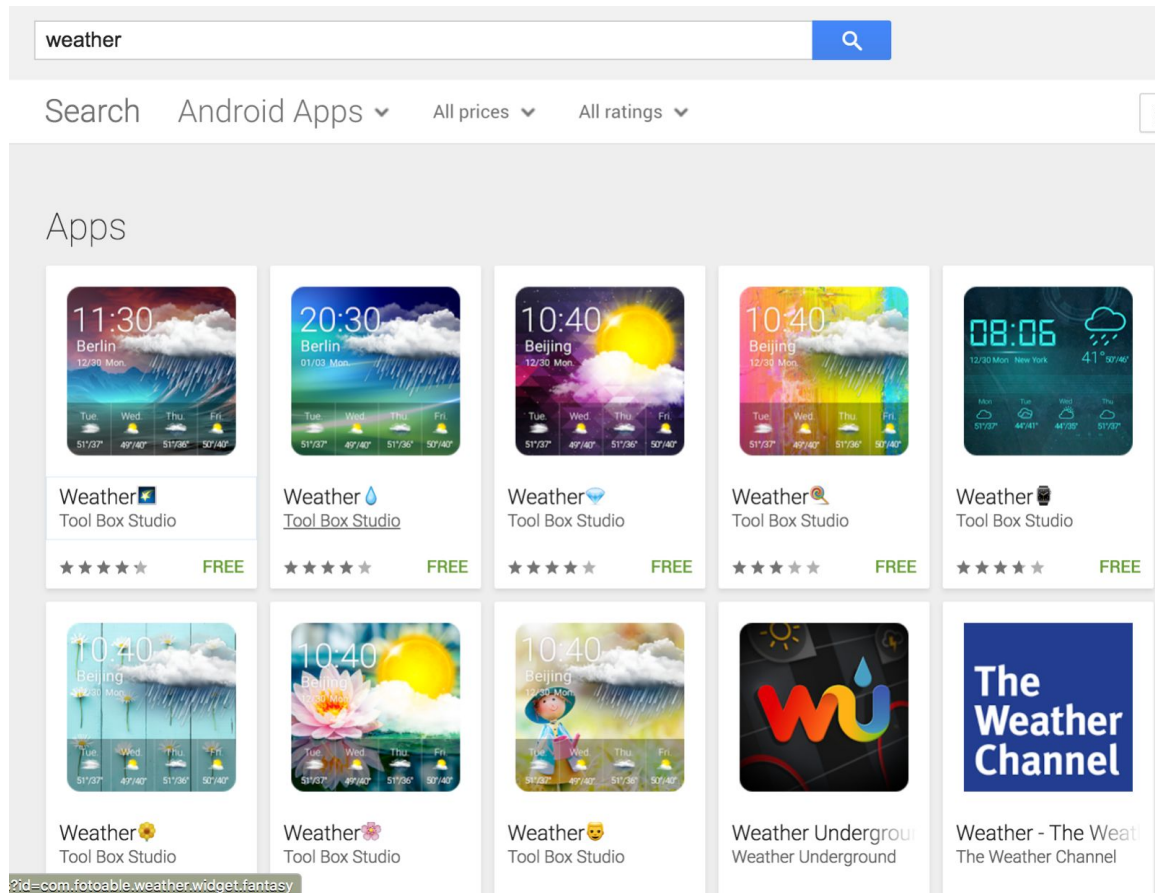
- Google Play Store is a digital distribution service operated and developed by Google, and it serves as a digital media store, offering music, books, movies, and television programs.
- Google Play Store is becoming bigger and competitive. The number of applications available in the Google Play Store in Jan 2017 were 2.2 Ml by end of 2017 were approximately 3.5 ML
- Ability to publish rapidly to over 2 billion active Android devices, Google Play helps users grow a global audience for apps and games and earn revenue.
- Applications are available through Google Play either free of charge or at a cost. They can be downloaded directly on an Android device through the Play Store mobile app or by deploying the application to a device from the Google Play website.



# Dashboard

## Common concerns for users:

- Which one to download?
- How many installations?
- How is the review?
- What is the price?
- What is the size?



# Related Research














## Optimize Google Play store App

- Keyword & Market research
- Page optimization + A/B testing
- Tracking / monitoring

---

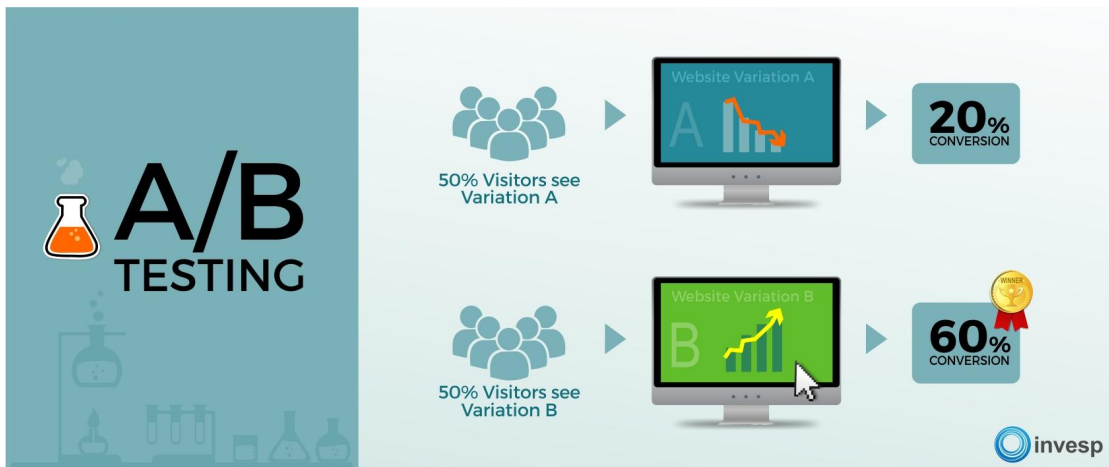
# Keyword & Market research

- How to choose keywords and keyword combinations for your app?
  - Brainstorm the keywords – it could be anything that comes to your mind that is relevant to your app and to the tasks it performs, or its main features.

	KEYWORD ↕	DIFFICULTY ↕	TRAFFIC ↕	APPS ↕	RANK ↕	CHANGE ↕	
 +	 clans of clash	High	79 	247	1	-	
 +	 clash clans	High	78 	249	1	-	
 +	 clans	High	84 	249	1	-	
 +	 coc	High	83 	246	1	-	
 +	 clash of clans	High	86 	249	1	-	
 +	 clash	Very High	46 	250	2	-	
 +	 games	Medium	99 	92	7	 44	

# Page optimization + A/B testing

- What is A/B testing?
  - To perform an A/B test, you will need to create 2 different versions of your listing element, and compare them against each other. During the experiment, half of your traffic will go to version A, and another half to version B, and then both options will be compared to determine the best result.



# Tracking / monitoring

- The App Store Optimization process never stops, the market changes as well as the Google Play Store itself, so it should be always up-to-date with the current market situation.
- **User reviews** is a part that is not totally under control, but if you are constantly monitoring it, you will get valuable information about your product, keywords, and the weak points of your product and / or your communication strategy with users.

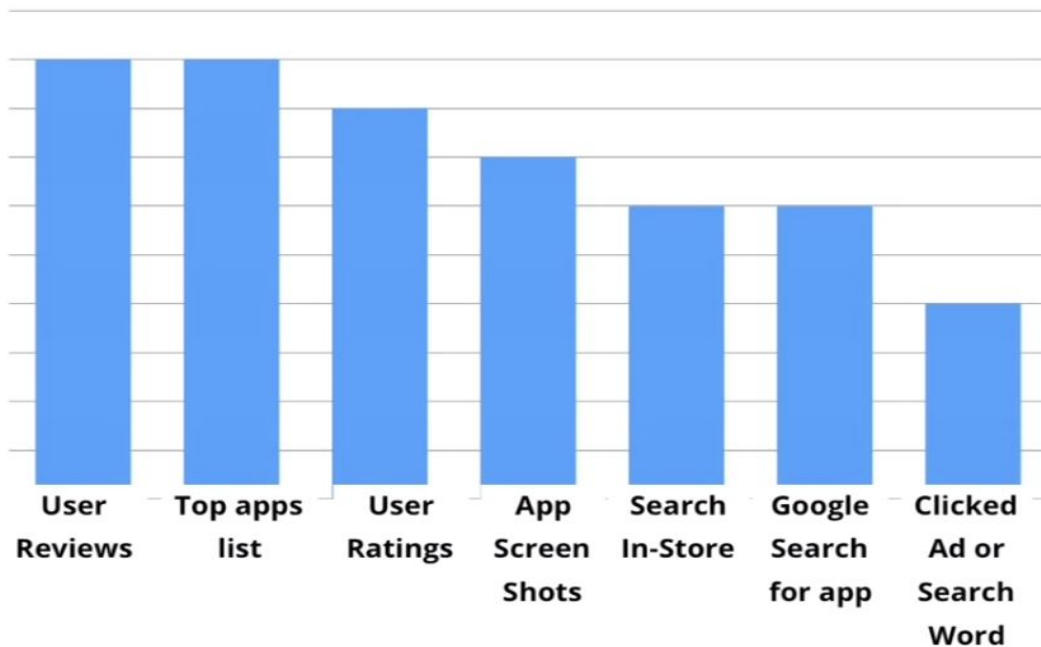


# Facts- App Install Influences

---

## In terms of installs, we know what works

What influences your decision to install an app?



# Objectives

- Analyze variables that drive the Google Play Store install/download
- Understand the importance of each variable
- Determine the most critical variable/s for downloading the Google Play Store

# Data Overview

- Google Play Store Dataset
- User Review Dataset

---

# Google Play Store Dataset

- 13 features with regard to Google Play Store
- 10K records

App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19M	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
Coloring book moana	ART_AND_DESIGN	3.9	967	14M	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8.7M	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25M	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2.8M	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

# User Review Dataset

- Users' reviews on Google Play Store
- 64K records

App	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
10 Best Foods for You	I like eat delicious food. That's I'm cooking ...	Positive	1.00	0.533333
10 Best Foods for You	This help eating healthy exercise regular basis	Positive	0.25	0.288462
10 Best Foods for You	NaN	NaN	NaN	NaN
10 Best Foods for You	Works great especially going grocery store	Positive	0.40	0.875000
10 Best Foods for You	Best idea us	Positive	1.00	0.300000

# Exploratory Data Analysis

Tools:

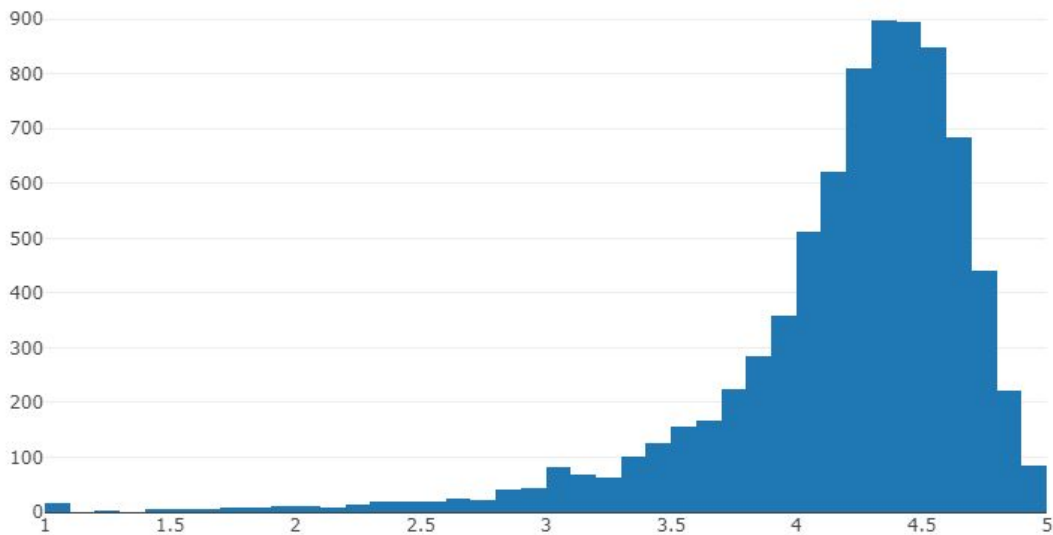
- Tableau
- Python Jupyter Notebook

---

# Rating Counts

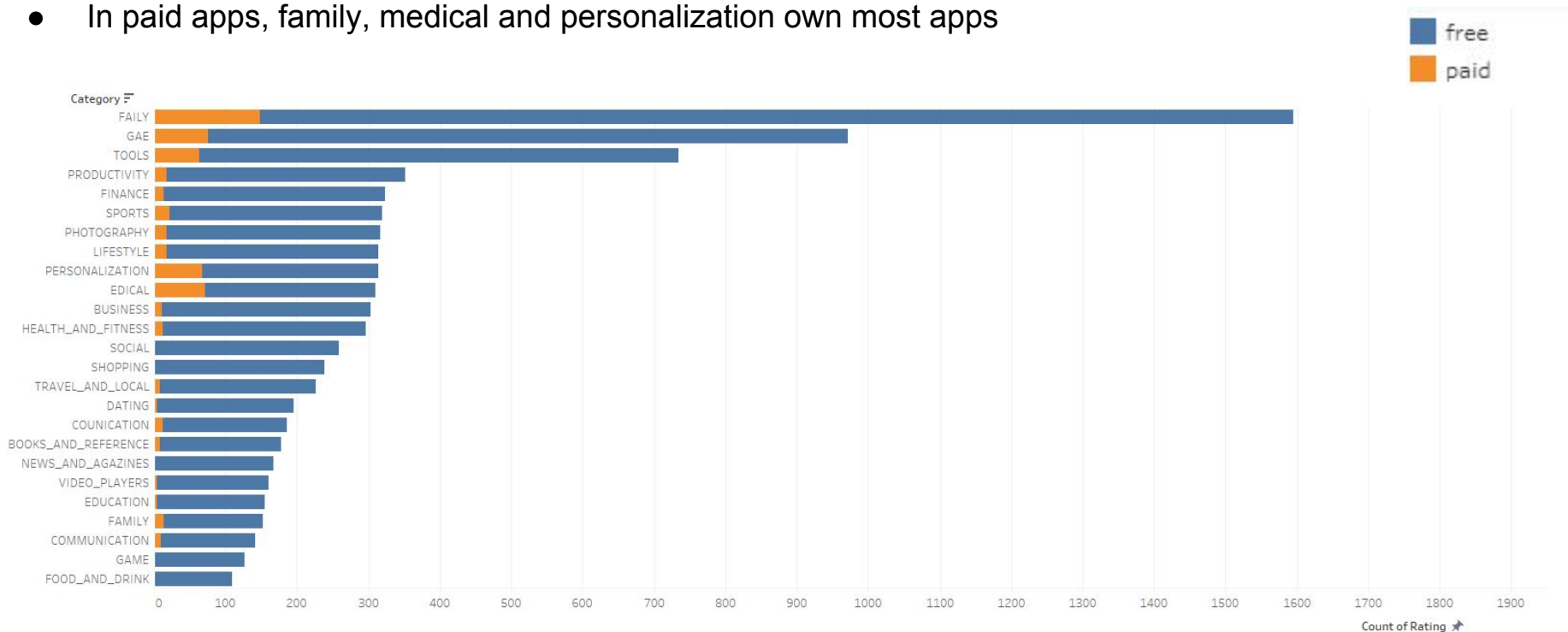
Average app rating = 4.173243045387998

Generally, most apps do well with an average rating of 4.17.



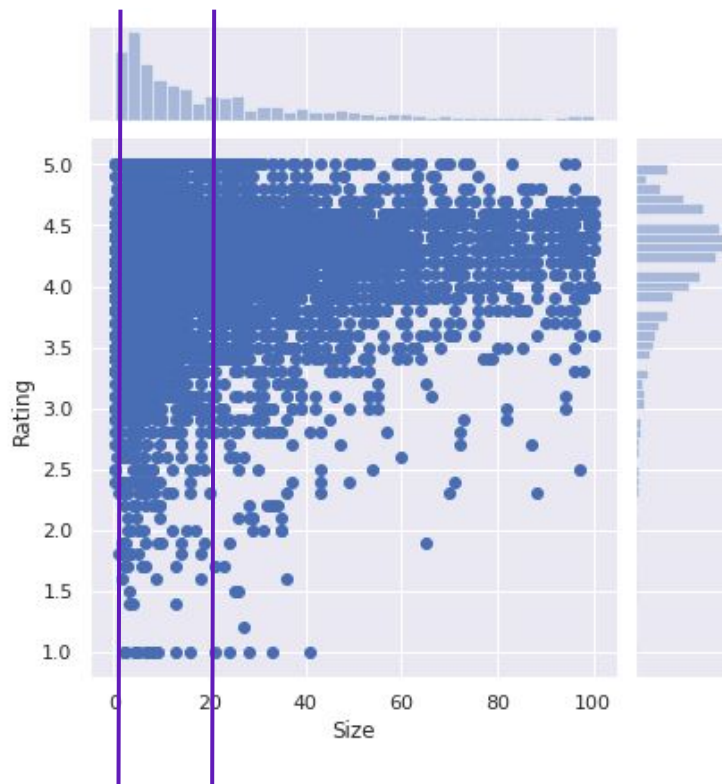
# App Counts Comparison

- Overall family, game and tools ranked the top 3
- In paid apps, family, medical and personalization own most apps



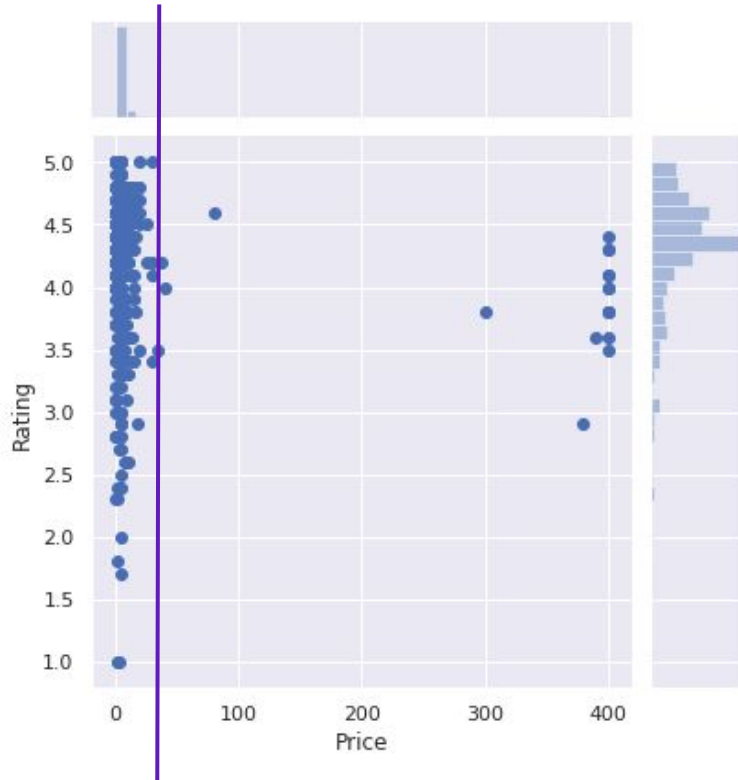


# Size Strategy



Most top rated apps are optimally sized between  
**0MB to ~20MB**

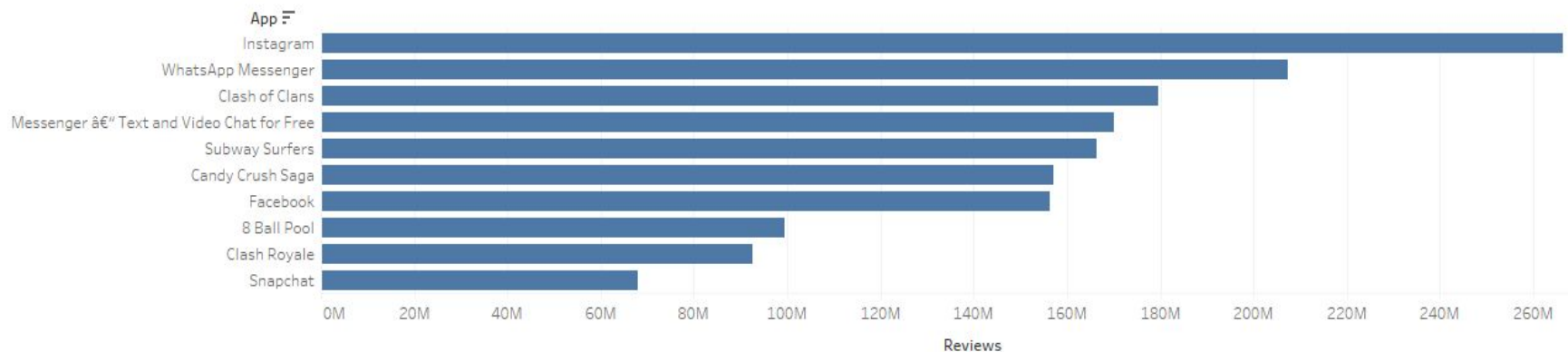
# Pricing Strategy



Most top rated apps are optimally priced between **~1\$ to ~20\$**. There are only a very few apps priced above 30\$.

# Apps with Most Reviews

- Apps with most reviews are chatting; social entertainments and game
- Instagram, What'sApp and Clash of clans ranked the top



# EDA Takeaways

- Average rating of apps on Google Play Store is **4.17**.
- Most of the top rated apps are **optimally sized between ~2MB to ~20MB**
- Most of the top rated paid apps are **optimally priced between ~1\$ to ~30\$**
- Users tend to download a given app more if it has been reviewed by a large number of people.
- Apps with most reviews are chatting; social entertainments and game

# Reviews

- Category Review Rating
- Sentiment Analysis
- WordCloud

# Pipeline

*“I love app, quick & easy use, right size read. I’m able meditate anywhere w/o laptop. I don’t access email meditation experience, I handy app. Thanks much.”*

## Data Cleaning

- Delete duplicates
- Drop missing values

## Data Preprocessing

- Merge rating and review dataset

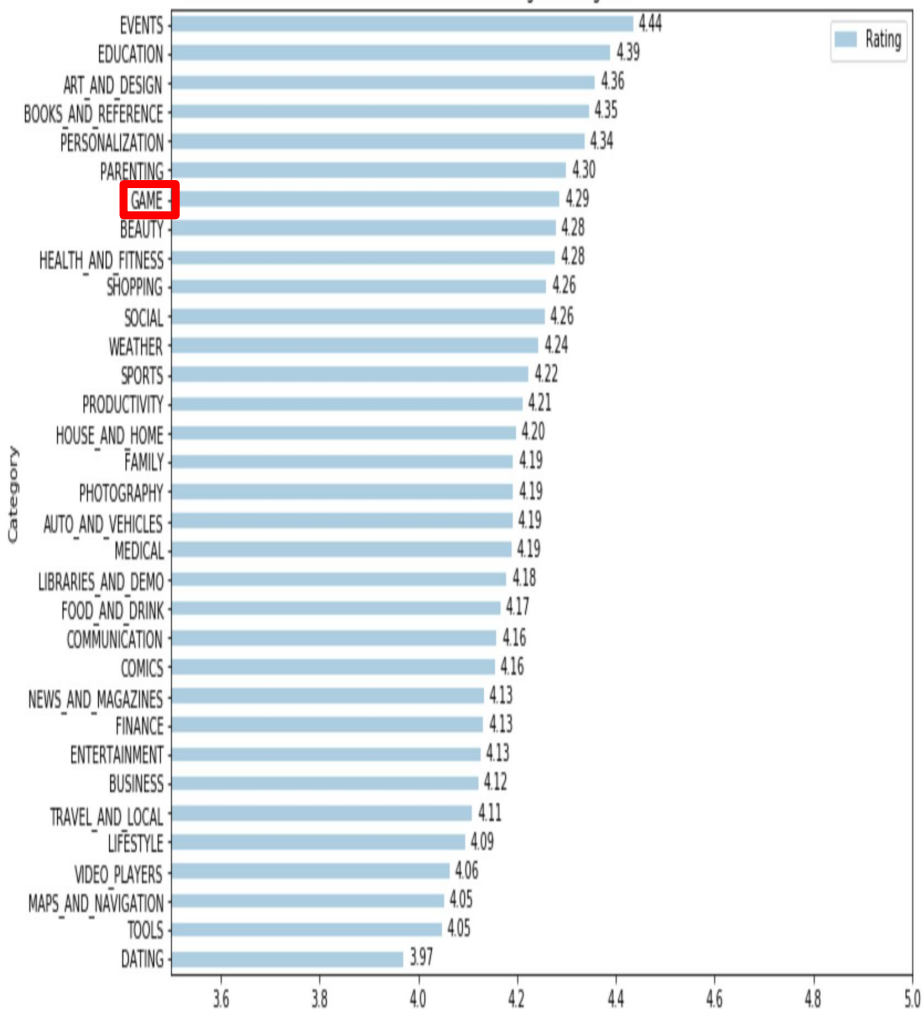
## Sentiment Analysis

- Sentiment Analysis
- Compare rating and sentiment score

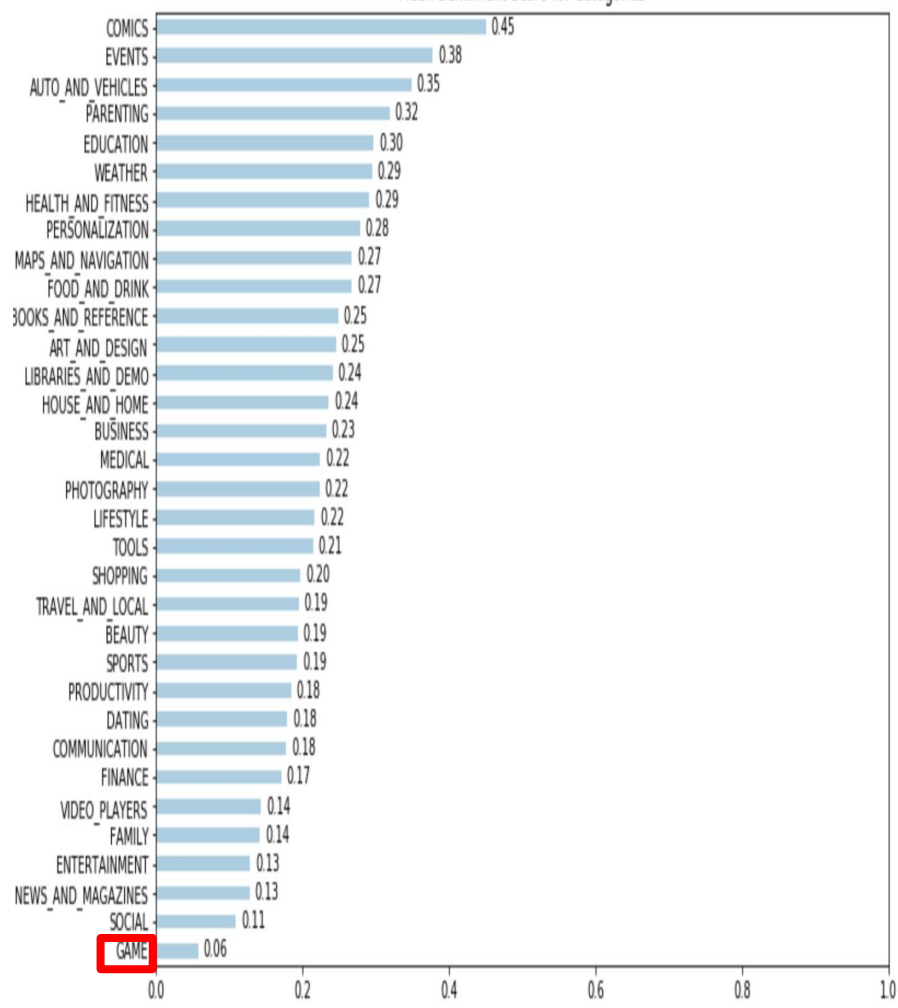
## WordCloud

- Remove Stopwords
- Text Fragment for negative reviews

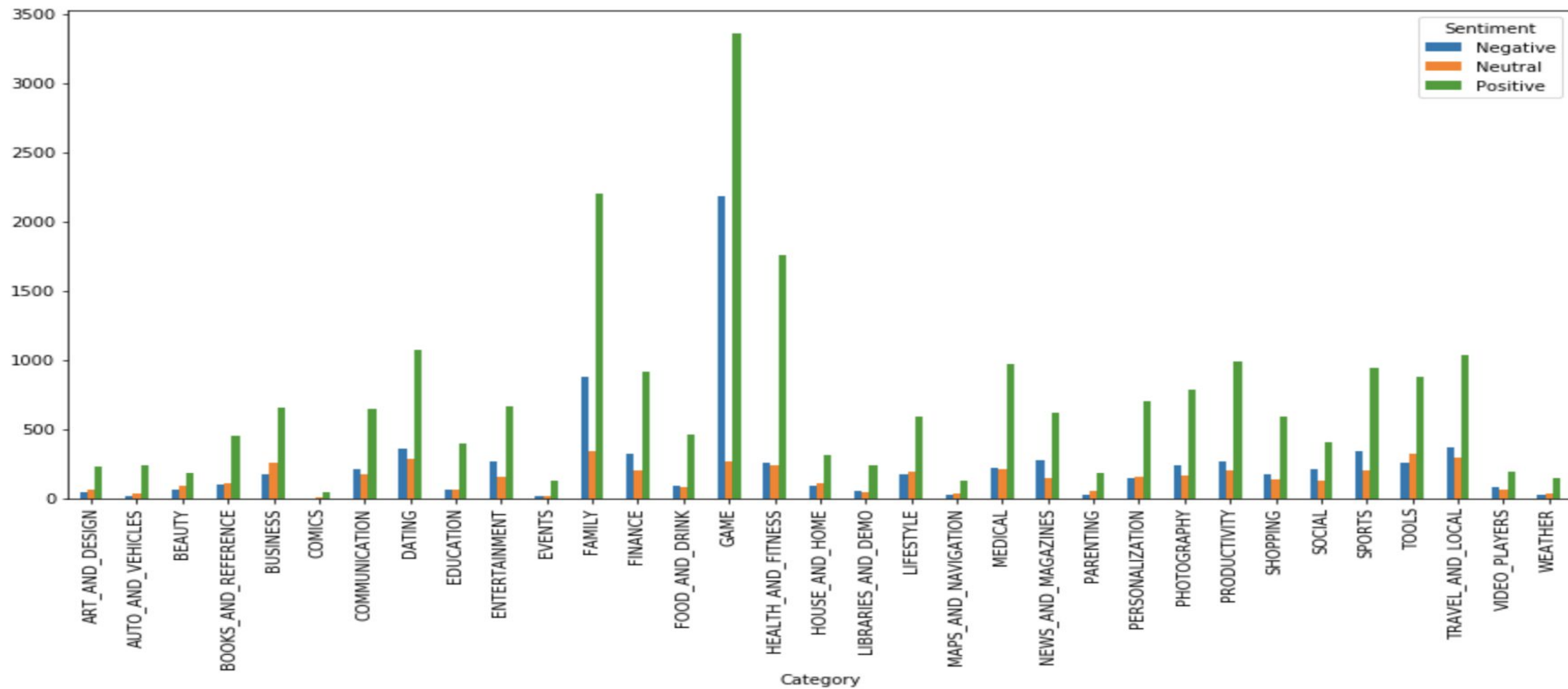
Mean Rating for Categories



Mean Sentiment Score for Categories



# Sentiment Class Distribution





# Negative Reviews Fragment

## Game

Negative Reviews: 2181/5802

Rating: 4.286

Sentiment Score: 0.057



## Health & Fitness

Negative Reviews: 257/2249

Rating: 4.277

Sentiment Score: 0.29



## Travel & Local

Negative Reviews: 1034/1692

Rating: 4.109

Sentiment Score: 0.195



# Methodology Approach

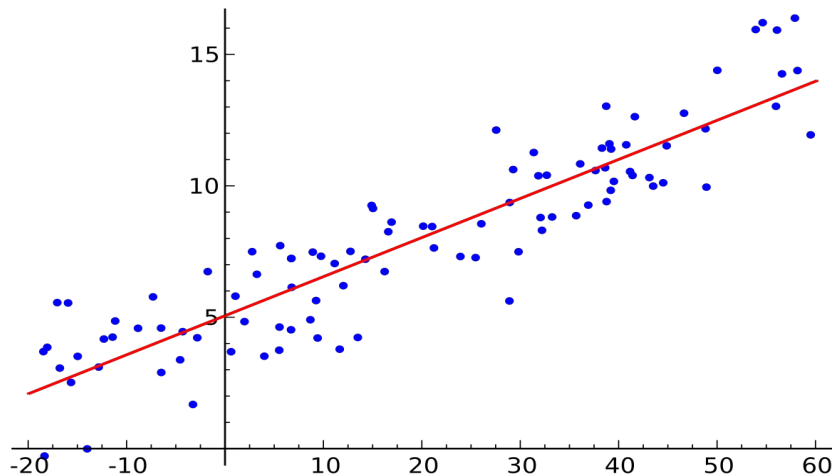
Methodology:

- Linear Regression
- Random Forest Classification

---

# Linear Regression Approach

- Converted “**categorical**” variables into model-understandable “**numerical**” data.
- Dropped “Rating” variable and applied the rest variables into the Linear Regression model.
- Split the dataset into training (80%) and testing (20%) subsets.
- A linear regression line has an equation of the form  $Y = a + bX$ , where  $X$  is the explanatory variable and  $Y$  is the dependent variable.
  - $X$ : Category, Reviews, Size, Price ...
  - $Y$ : Rating



# Linear Regression Model

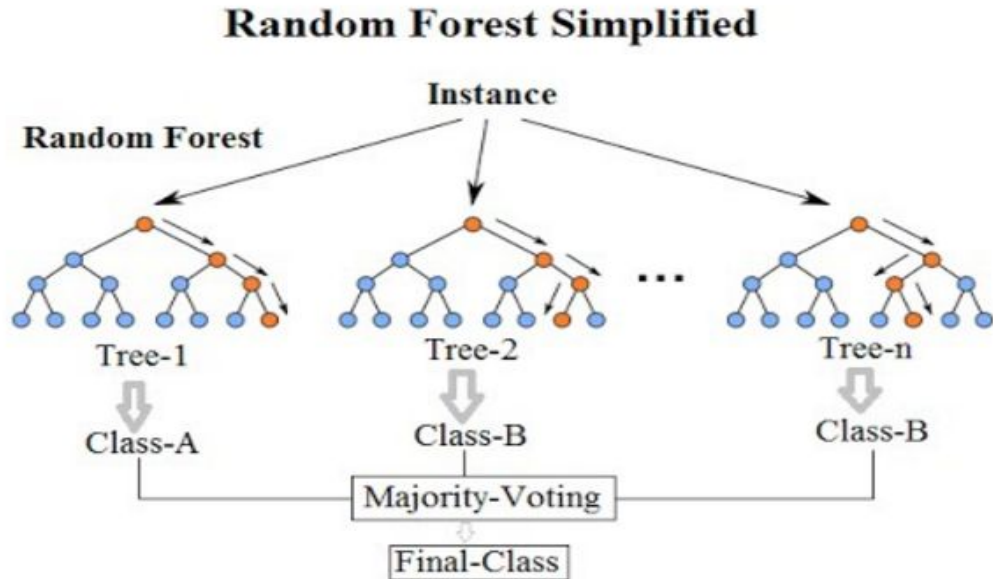
- “Category”, “Type” and “Content Rating” have positive relations with “Rating” variable.
- “Reviews” and “Price” are not that significant comparing to other variables.
- R-Squared: 0.852

<b>Dep. Variable:</b>	Rating	<b>R-squared:</b>	0.852
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.852
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	6730.
<b>Date:</b>	Tue, 11 Dec 2018	<b>Prob (F-statistic):</b>	0.00
<b>Time:</b>	01:56:04	<b>Log-Likelihood:</b>	-17824.
<b>No. Observations:</b>	9360	<b>AIC:</b>	3.566e+04
<b>Df Residuals:</b>	9352	<b>BIC:</b>	3.572e+04
<b>Df Model:</b>	8		
<b>Covariance Type:</b>	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
<b>Category</b>	0.2035	0.003	68.664	0.000	0.198	0.209
<b>Reviews</b>	-1.769e-08	7.03e-09	-2.518	0.012	-3.15e-08	-3.92e-09
<b>Size</b>	3.178e-08	6.69e-10	47.511	0.000	3.05e-08	3.31e-08
<b>Installs</b>	1.015e-09	2.4e-10	4.224	0.000	5.44e-10	1.49e-09
<b>Type</b>	0.7884	0.068	11.588	0.000	0.655	0.922
<b>Price</b>	7.881e-06	1.09e-05	0.723	0.470	-1.35e-05	2.93e-05
<b>Content Rating</b>	0.3928	0.021	18.332	0.000	0.351	0.435
<b>Genres</b>	-0.0147	0.001	-16.669	0.000	-0.016	-0.013
<b>Omnibus:</b>	21.069	<b>Durbin-Watson:</b>	1.088			
<b>Prob(Omnibus):</b>	0.000	<b>Jarque-Bera (JB):</b>	19.868			
<b>Skew:</b>	-0.085	<b>Prob(JB):</b>	4.85e-05			
<b>Kurtosis:</b>	2.852	<b>Cond. No.</b>	3.78e+08			

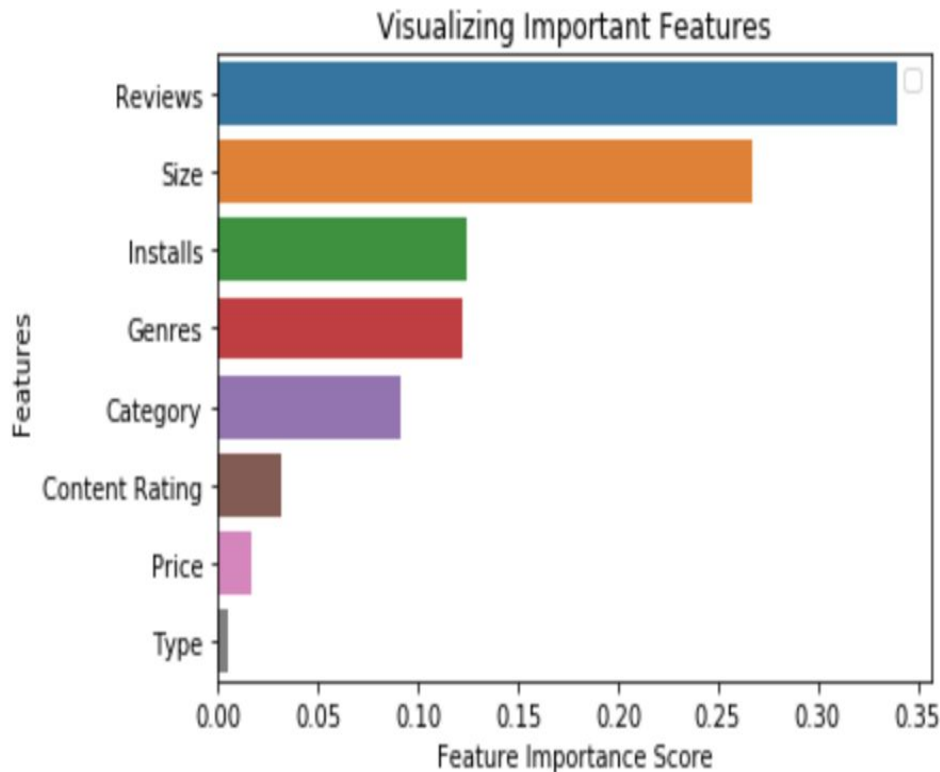
# Random Forest Classification

- A decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogenous).
- Since the average of Rating is around 4.1, we decide to set rating  $\geq 4.5$  to be “Excellent”. And the rest to be “Normal”.
- As a result, the bottom variables of each tree should be homogenous to each other.



# Random Forest Classification

- Model Accuracy: 0.7628
- Feature Importance by scores:
  - From the right figure, we can tell that “Reviews” and “Size” are the top two significant variables corresponding to the “Rating” variable.
  - On the other hand, “Price” and “Type” are the least important two variables.



# Conclusion

—

# Model Results

- Positive Relation Variables: (Linear Regression)
  - Content Rating
  - Category
  - Type
- Positive Relation Variables: (Random Forest Classification)
  - Reviews
  - Size

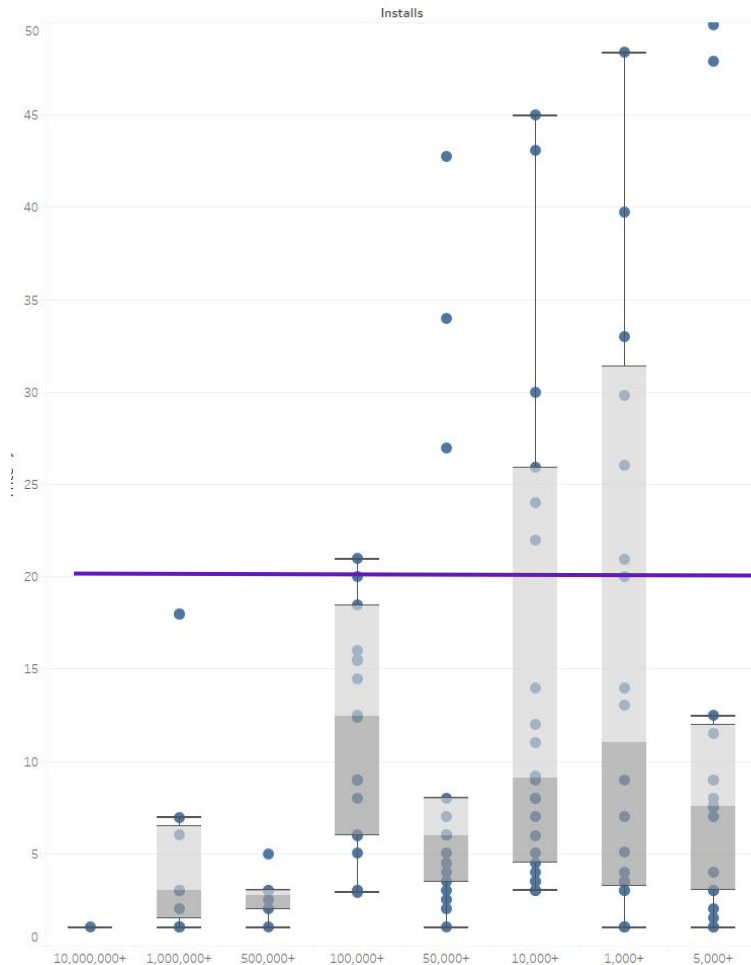


# Appendix

—

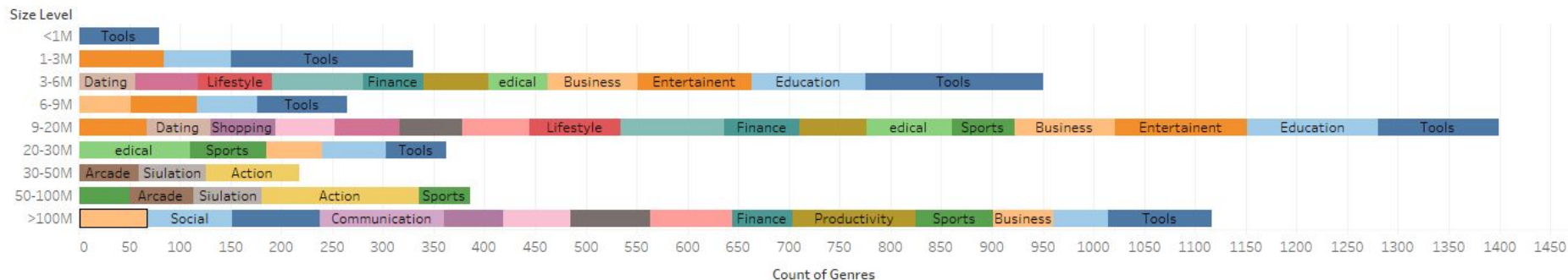
# Price Spread Out

- The more installs, the cheaper the app is
- In paid apps, majority apps priced under \$20
- App with more installs usually less spread out than those with less installation



# Size Varies by Genre

- Most apps have size at 9-20MB
- Tools, education, business and entertainment apps spread across different sizes

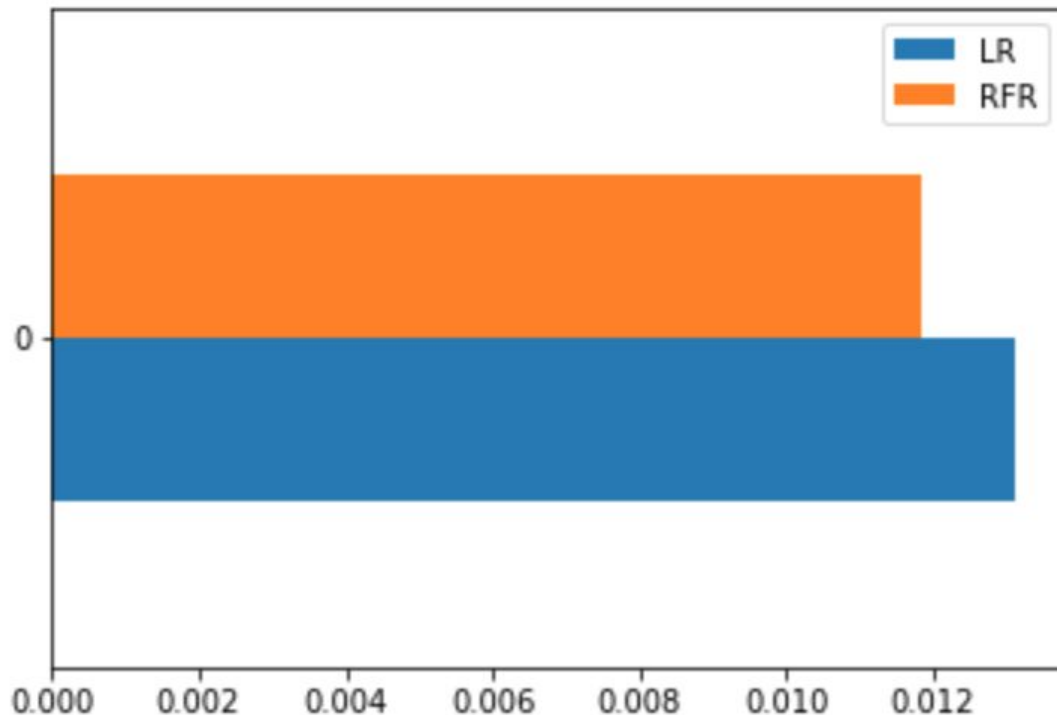


# Mean Squared Log Error (MSE)

	LR	RFR
0	0.013118	0.011838

## Methodology Conclusion:

1. MSE:  
LR > RFR
2. Performance:  
RFR > LR



# Market Breakdown

- **Family** and **Game** apps have the highest market prevalence.
- Interestingly, **Tools**, **Business** and **Medical** apps are also catching up.

