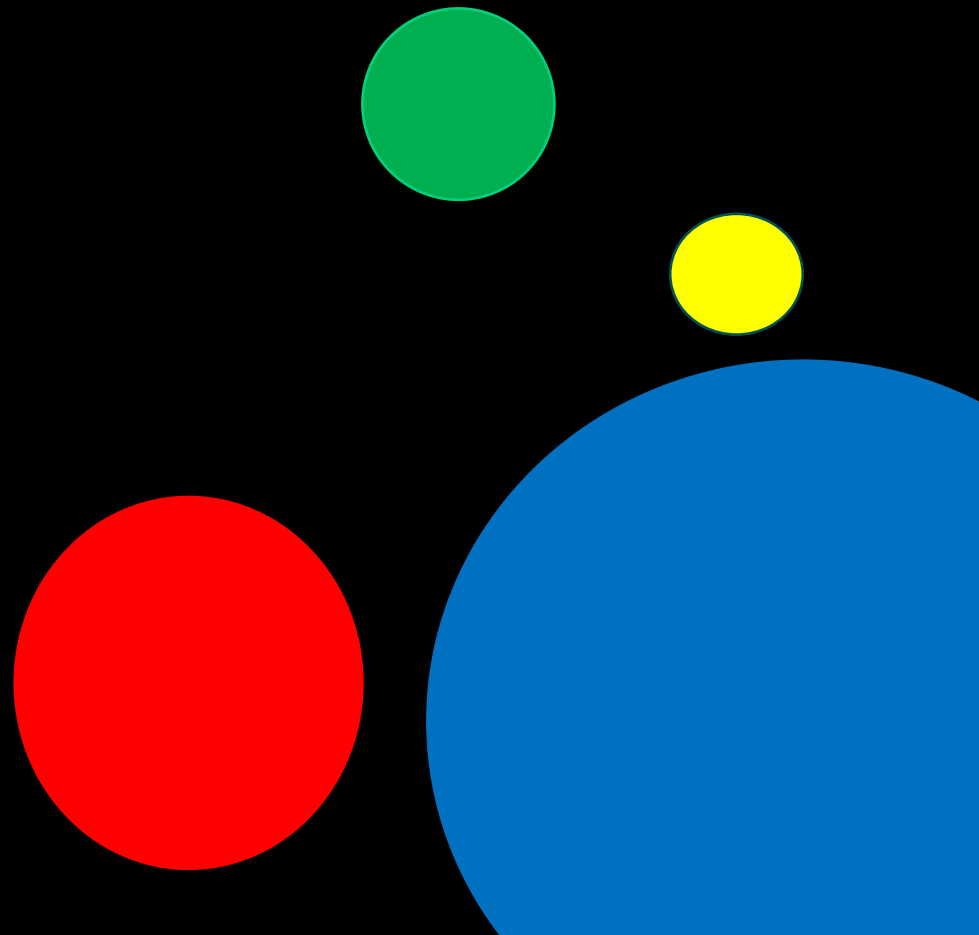
Google Play Store Downloads Forecast

Team members: Aruneema, Leah, Spruha, Nancy

Instructor: Christopher Dunham

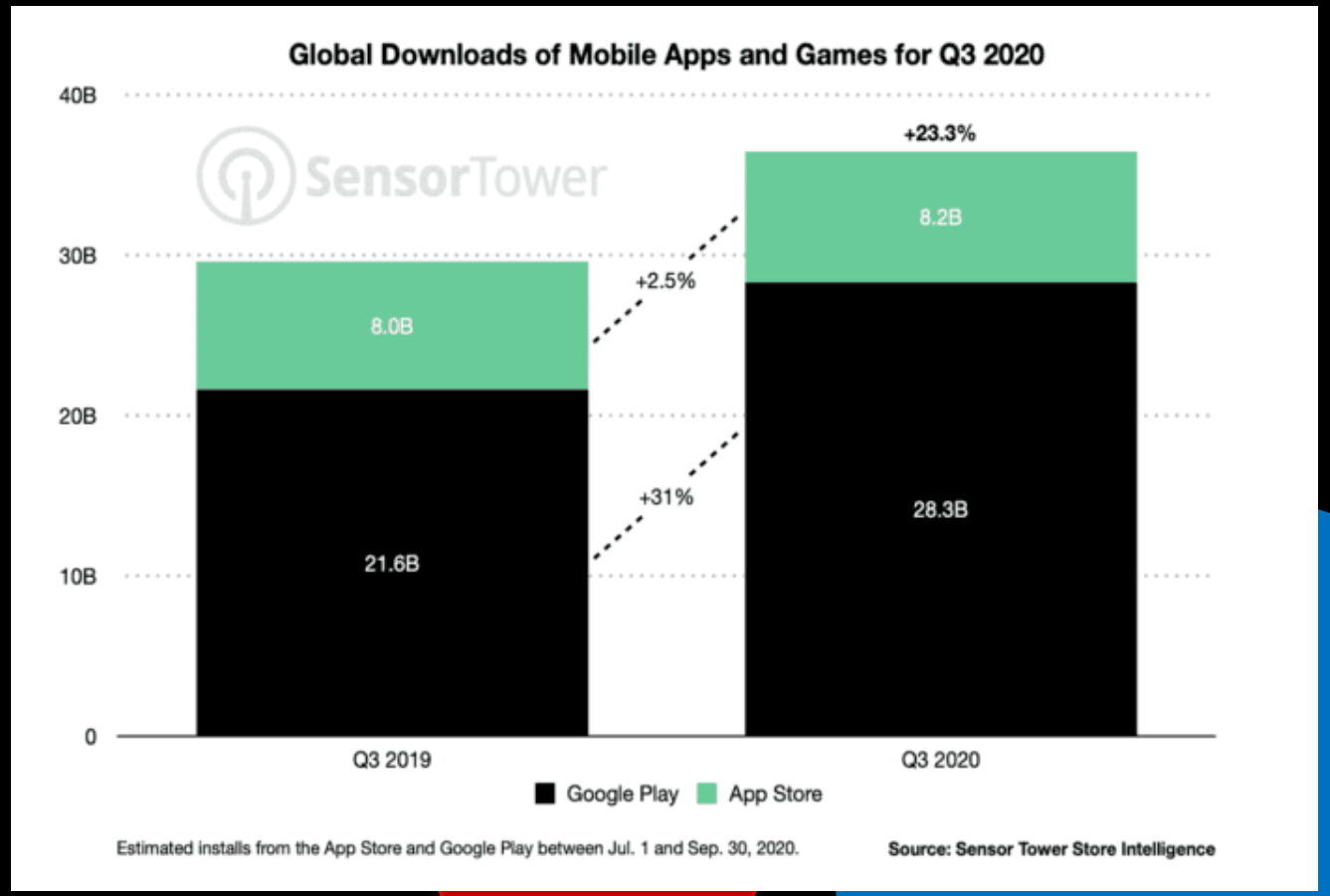
Agenda

- Objective
- The Data
- EDA
- Overview of the Supervised models
- Gradient descent
- Association rules
- Insights



Why Google Play Store

- Google Play makes it easy for more than 2.5 billion monthly users across 190+ countries worldwide to discover millions of high-quality apps and delightful content.
- Allows 97%of developers to distribute their free apps and take advantage of all Google Play has to offer at no charge.



Objective

Primary objective:

To develop a predictive model that **estimates the number of downloads** for apps on the Google Play Store.

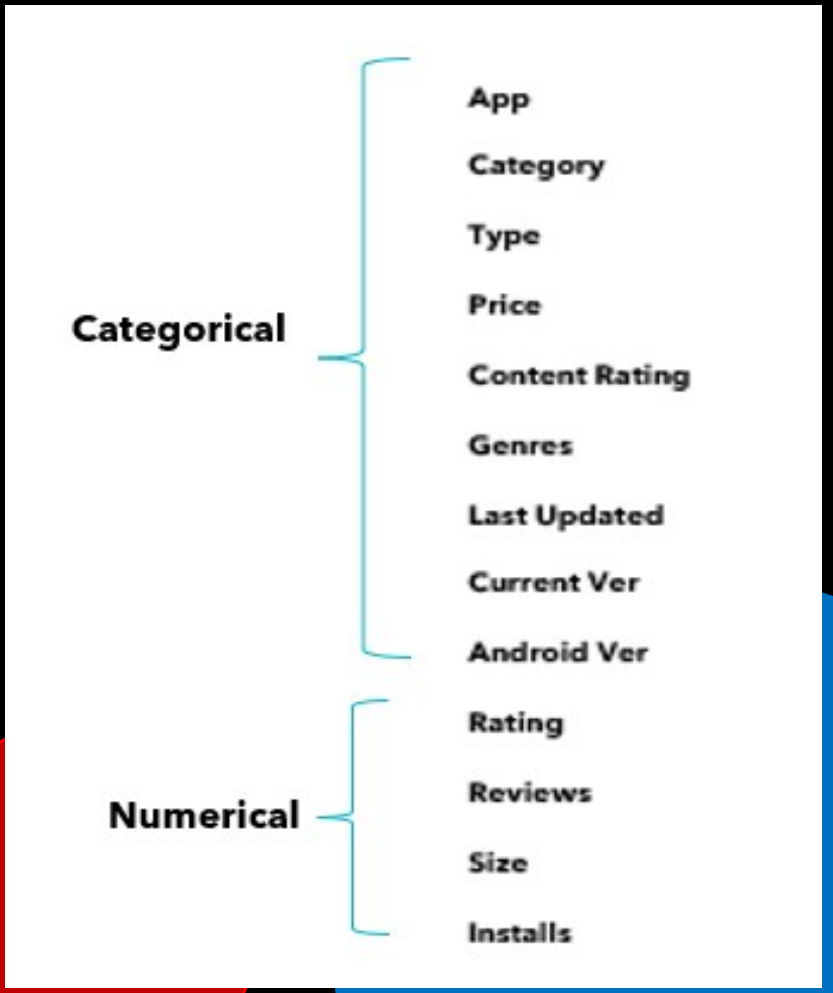
Why:

Build a valuable tool for app developers and publishers to gain insights into the numbers of downloads from their potential user base before they're published.

Enhance **decision-making processes** by offering data-driven insights into the market and helping allocate resources effectively.

The Data

- Rows: 10,841
- Columns: 14
- 8 categorical variables , 6 numerical variables
- Unique values:
  - 'App’ - 7026
  - 'Category’- 33
  - 'Type’- 2
  - 'Content Rating’- 5
  - 'Genres’- 112
  - 'Current Ver’- 2476
  - 'Android Ver’- 32



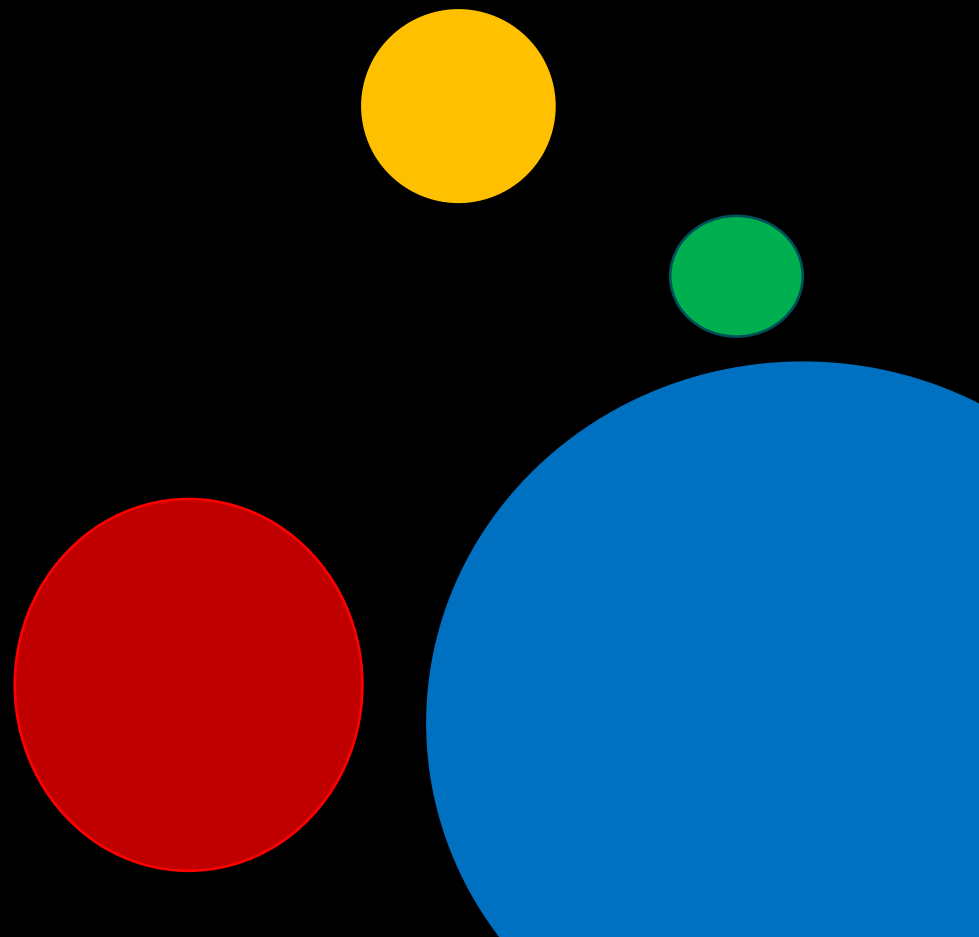
Data Cleaning

Data Type issues, removing alphanumeric and special characters, removing nulls

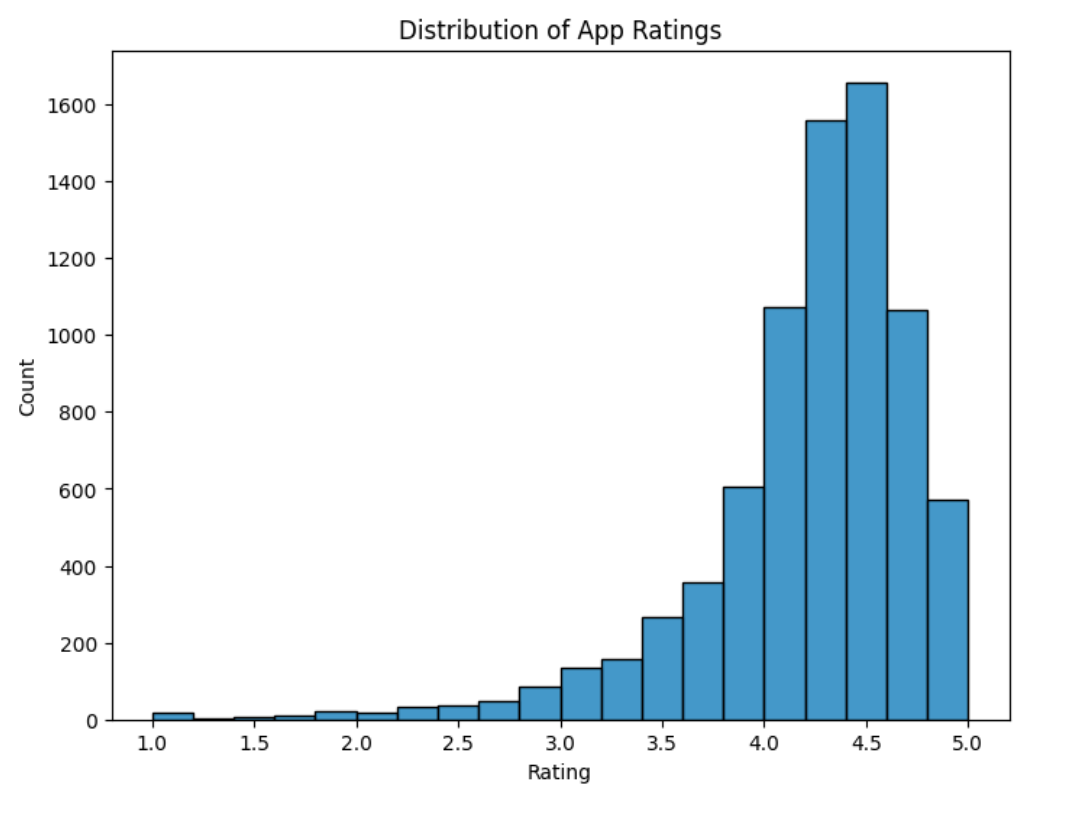
- Size => Kb → Mb
- RegEx to remove ‘$, +’

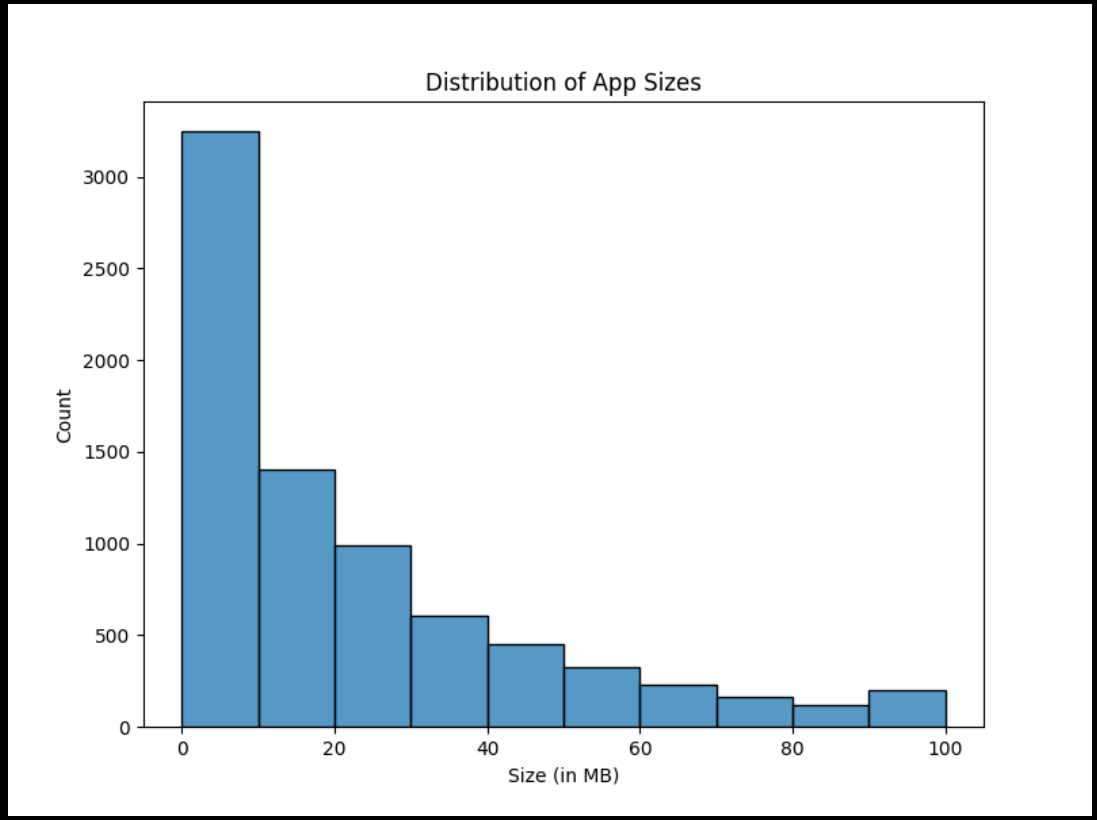
Post data cleaning :

- No of rows 7726
- No of columns 13



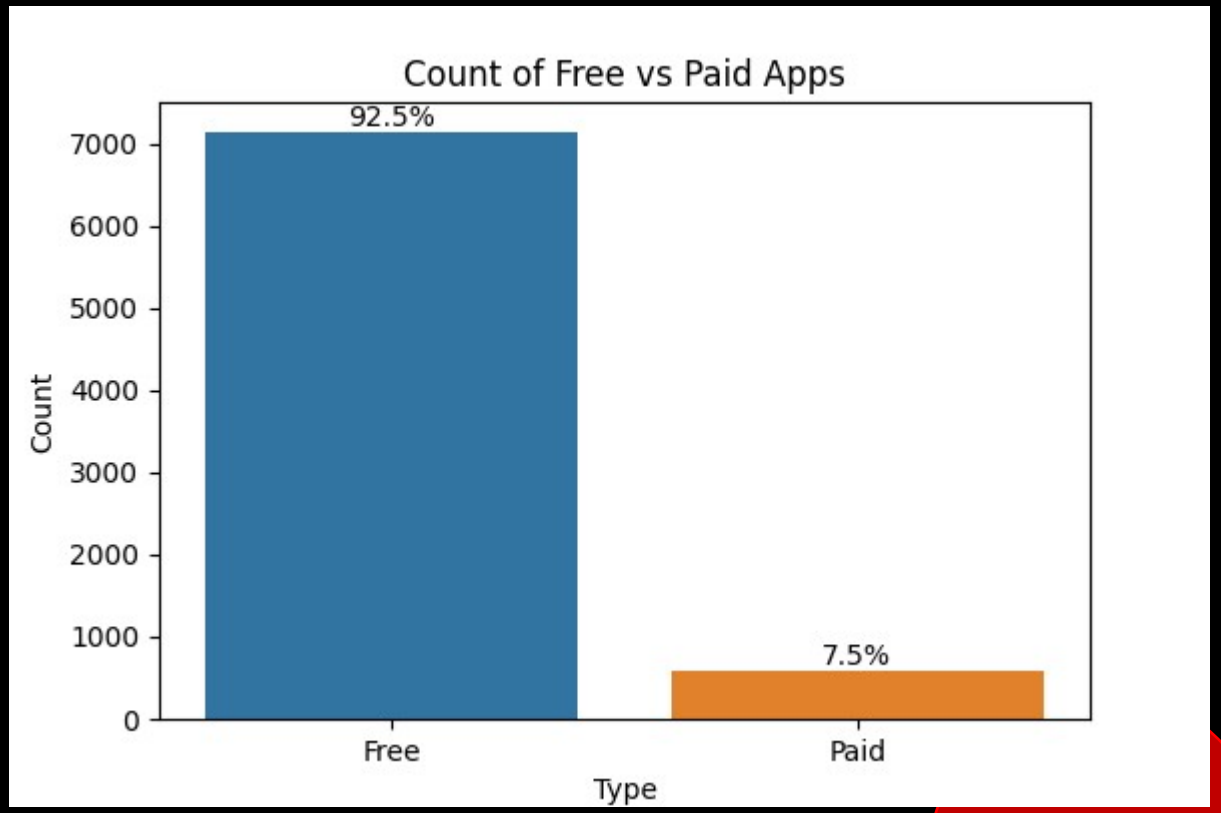
Exploratory Data Analysis: Ratings & Size





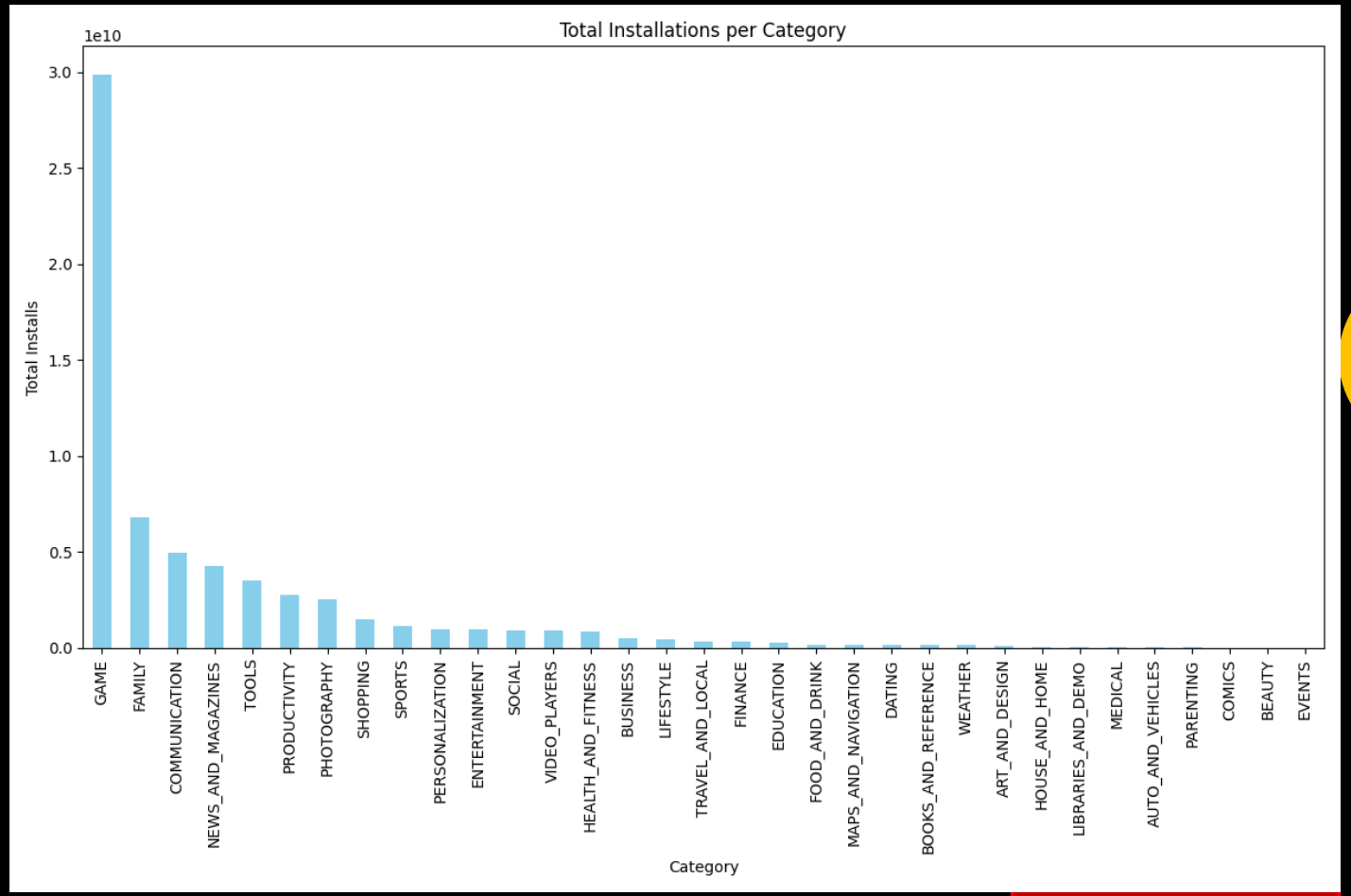
- Majority of apps have ratings between 4.1 to 4.4 . Among these , the largest portion has ratings 4.4 . App size < 10 MB has the highest distribution

Type

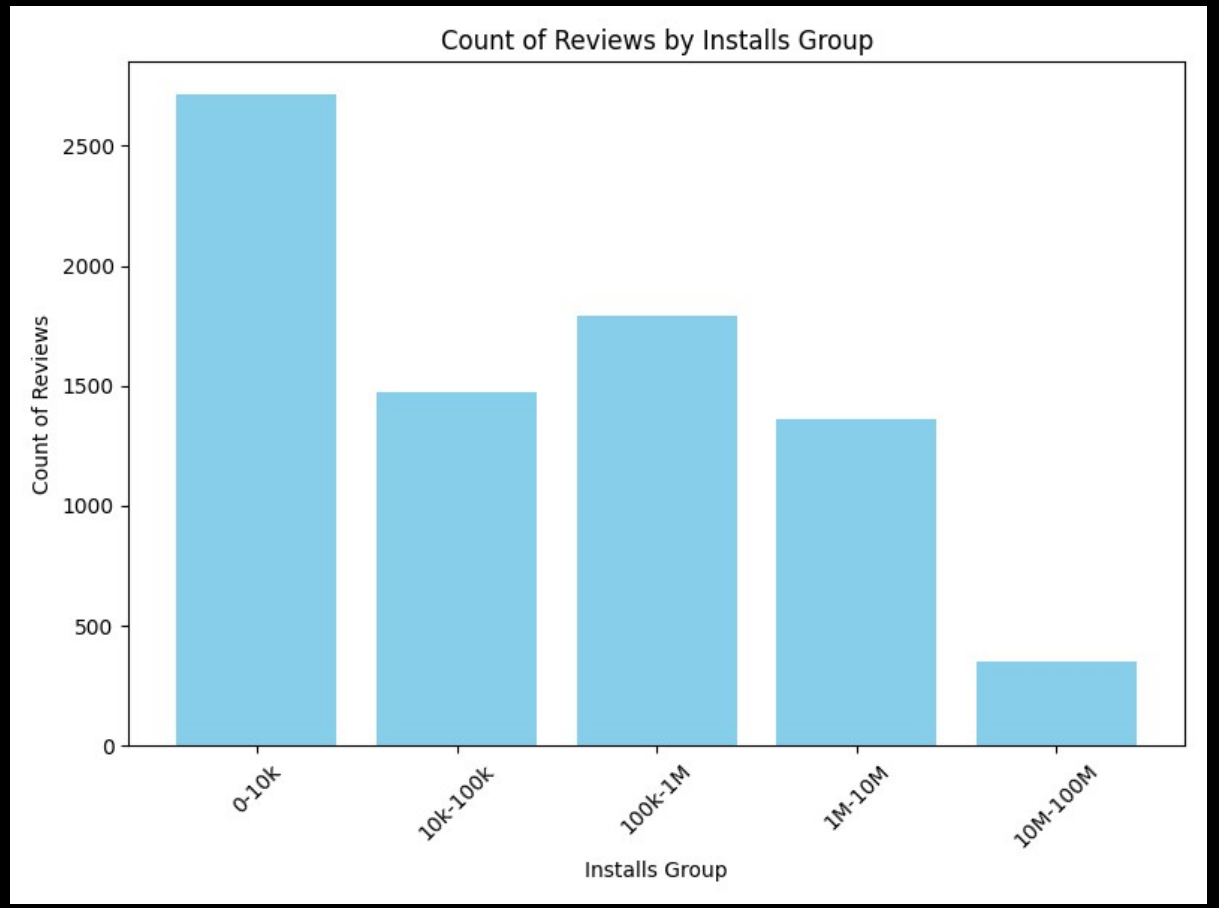


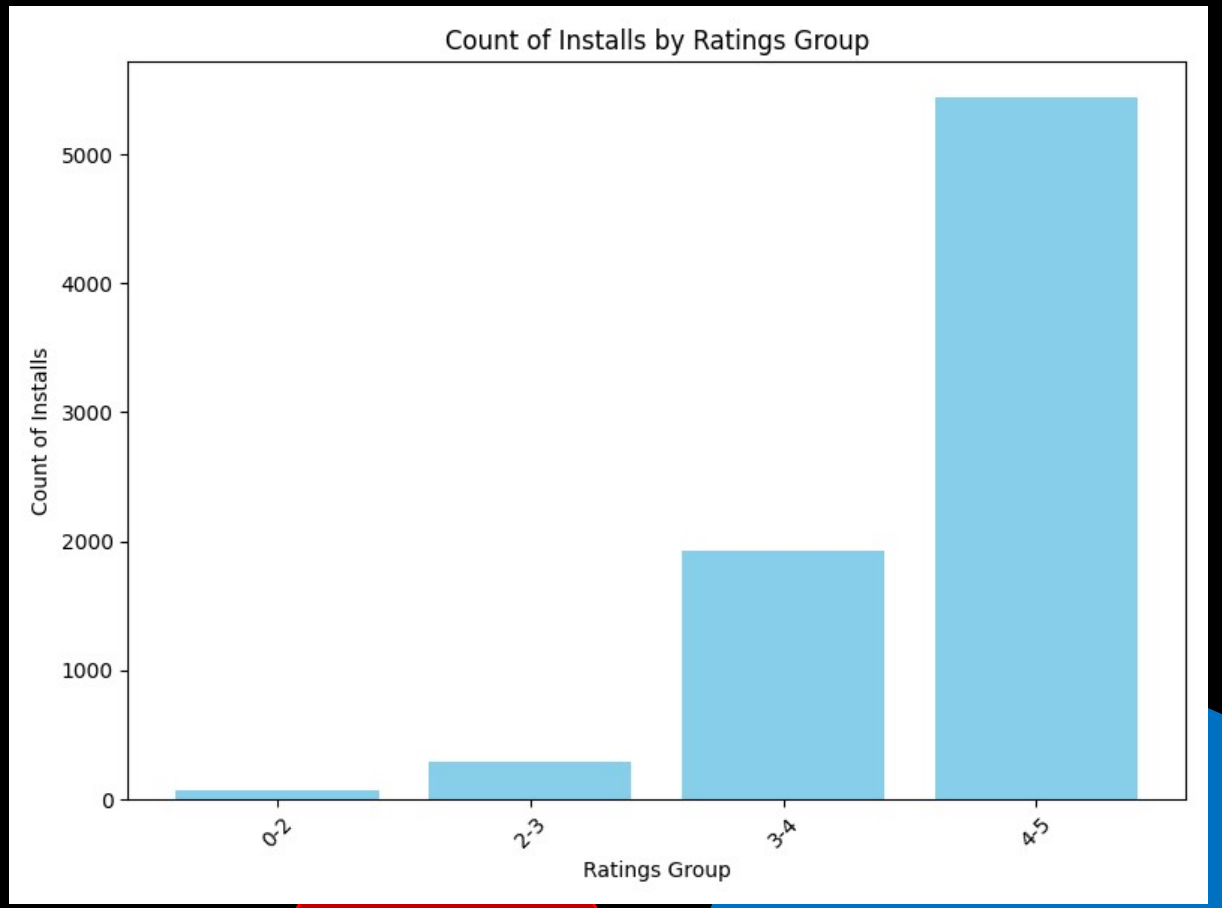
92.5 % apps are free and only 7% apps are paid apps

Installations

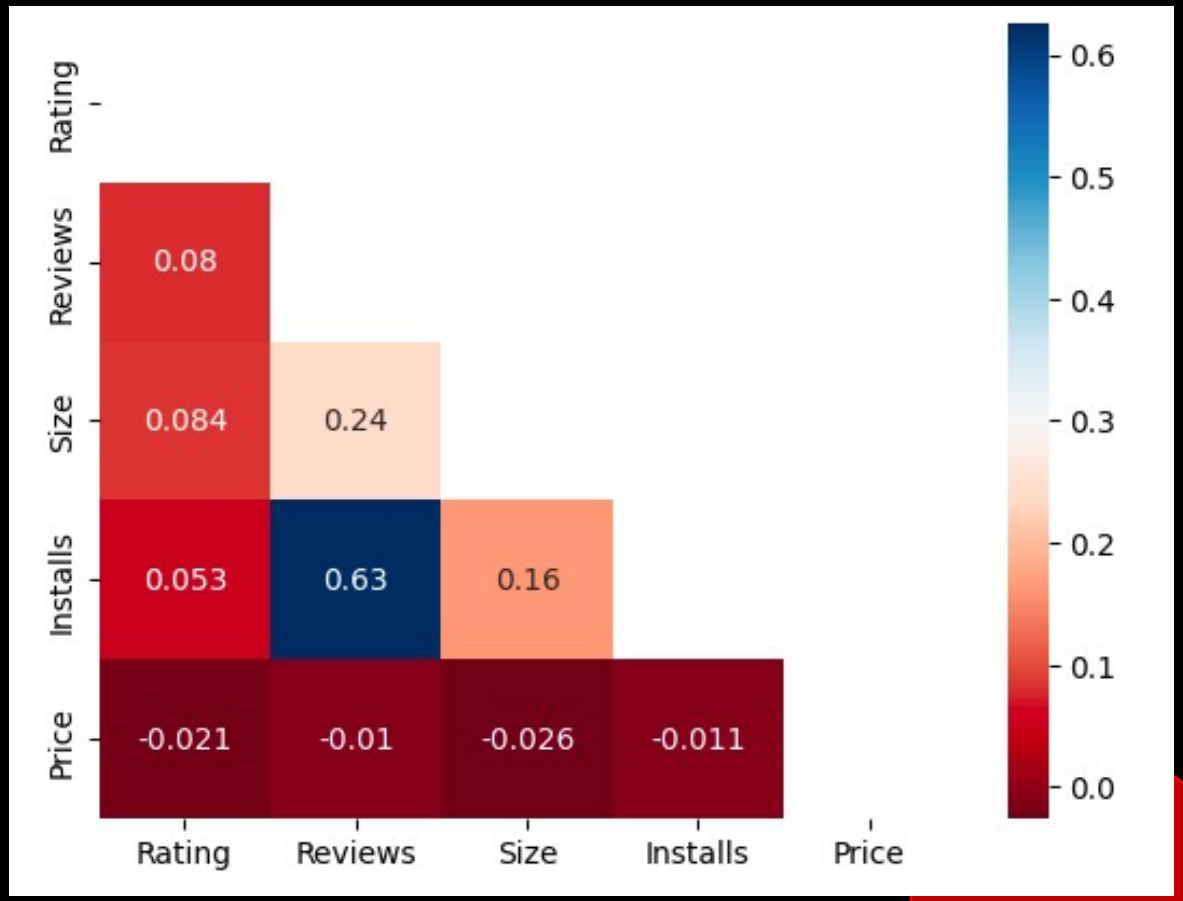


There are 34 unique categories of which "Family" category has the highest app distribution. Apps under the category of games has the highest no of installs


Count of Reviews by Installs Group
Count of Reviews
2500
2000
1500
1000
500
0
0-10K
10k-100k
100k-1M
1M-10M
10M-100M
Installs Group


Count of Installs by Ratings Group
Count of Installs
5000
4000
3000
2000
1000
0
0-2
2-3
3-4
4-5
Ratings Group

Correlation Matrix



There’s a strong correlation between ‘Installs’ and ‘Reviews’

Overview of Supervised Models

Model	Inputs	RMSE
Multivariate Regression Model	Rating, Reviews, Size, Price	44,019,633
Decision Tree with Numeric Variables	Rating, Reviews, Size, Price	37,400,492
Decision Tree with Categorical variables	Rating, Reviews, Size, Price, Category, Type, Content Rating, Genres	33,449,809
Random Forest Model with All Variables	Rating, Reviews, Size, Price, Category, Type, Content Rating, Genres	31,573,385
Gradient Boost	Rating, Reviews, Size, Price	23,112,198

Gradient Boost

Hyperparameters:

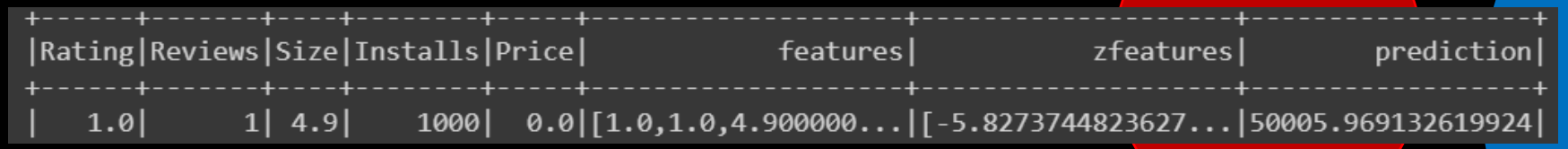
maxDepth=4, maxIter=100, stepSize=0.1

Feature: Size, Importance: 0.507

Feature: Reviews, Importance: 0.285

Feature: Rating, Importance: 0.207

Feature: Price, Importance: 4.132e-05



```
+------+-------+----+--------+-----+--------------------+--------------------+------------------+
|Rating|Reviews|Size|Installs|Price|            features|           zfeatures|        prediction|
+------+-------+----+--------+-----+--------------------+--------------------+------------------+
|   1.0|      1| 4.9|    1000|  0.0|[1.0,1.0,4.900000...|[-5.8273744823627...|50005.969132619924|
```

Association Rules- Tagging

We subset the data for apps with high installations (>500K)

Then we tried to use associations rules to predict the right genre tagging for a particular app so that it can get high installations

Apps	Category_Indices	prediction
THE KING OF FIGHT...	[13]	[14, 0]
Zombie Death Shooter	[16]	[14, 0]
Life market	[54]	[]
Where is my Train...	[7, 60]	[]

13 = Video
54 = Role Playing
7 = Home
60 = Tool

0= App
14 = Action
15 = Action
16 = Adventure
53 = Racing

Insights

1. From the linear Regression: Apps with high ratings do not necessarily have high installs, whereas higher price indicates lower downloads.

Installations = 5780403.39 -652434.40 * Ratings + 20.93 * Reviews -3860.36 * Size -10493.48 * Price

2. The Gradient Boost has the best performance. The feature importance from Gradient Boost suggests that ‘size’ is the most important numeric feature, followed by reviews, ratings and price respectively.

3. From the rules that we built, we can suggest the Google App Store to build up a better auto tagging system to help content creator improving the installing.