Google Play Store App Analysis

Exploratory Data Analysis

Najja Osiomwan

November 18, 2020

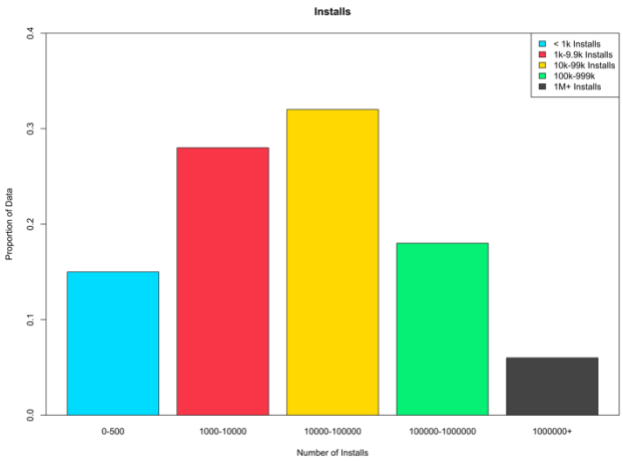
Overview

To recap, the purpose of this project is to analyze the Google Play Store data, explore relationships between variables and determine which ones would be good predictors to determine app Installs by using logistic regression. In this document I will discuss the findings of the univariate analysis of each variable, explain why I chose to manipulate the data and what functions I used to do so.

Installs

This will be my dependent variable for the project, so it is important that I have a good understanding of the variable and ensure that it is normally distributed.

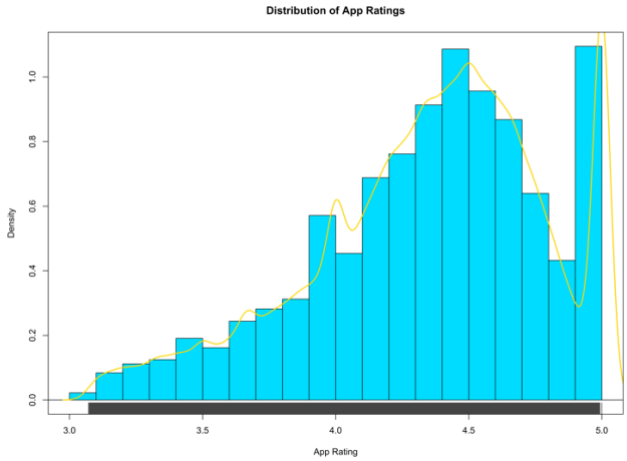
- After reviewing Installs, a numeric variable, I saw that the data was extremely left skewed as there were many apps with installs under 100,000 and just a few apps with installs greater than 1,000,000.
- I decided it would be best to convert this variable to a factor and group installs into 5 buckets cut() so that apps with installs greater than 1,000,000 would not have to be excluded from the data as outliers.
- This data transformation from a numeric to categorical data type is the reason I needed to use logistic regression glm() instead of linear regression lm() in later sections.



Install Group	Proportion of Data
< 1k Installs	15%
1k-9.9k Installs	28%
10k-99k Installs	32%
100k-999k	18%
1M+ Installs	6%

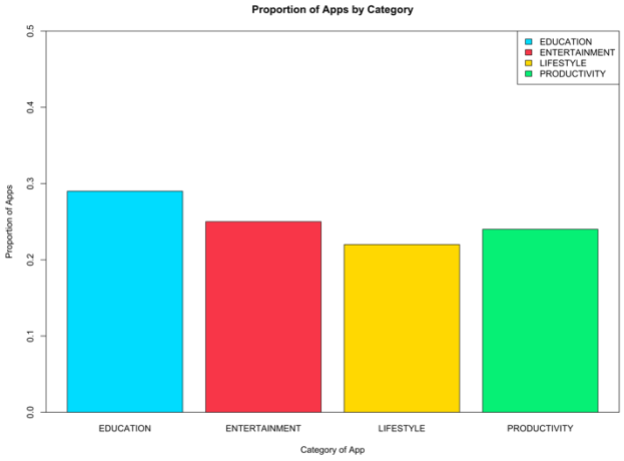
Rating

- I started by converting Rating to a numeric variable as.numeric() and creating a histogram hist() to view the distribution of ratings. The graph shows that ratings are heavily left skewed.
- Further analysis shows that almost 97% of the apps are rated 3 or higher which is the reason for the skewed data. I attempted the log log() and square root sqrt() data transformations and saw no change so I decided to remove outliers from the dataset. Outliers were determined to be any app with a rating lower than 3.07.
- Although still left skewed, the data has moved closer to a normal distribution by removing the outliers. The log and square root transformations do not improve the shape of the bell curve so I will leave the data as is.



Category

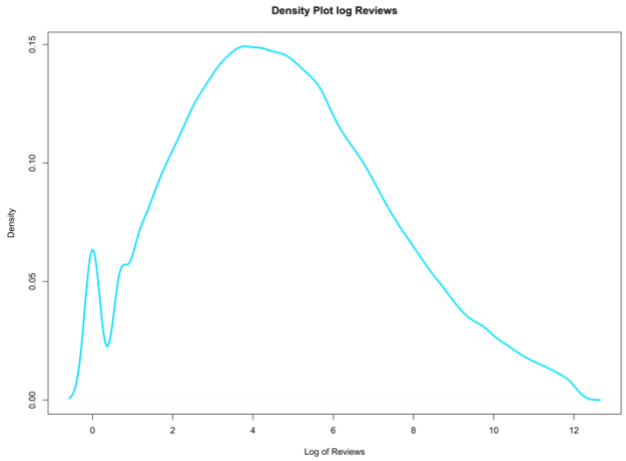
- Creating a table table() for Category showed that there were 68 categories in total. A few categories with a lot of apps and some categories with little to no apps. In order to achieve significant test results in the statistical analysis portion of this project, I decided to combine these categories into 4 groups: Education, Lifestyle, Productivity and Entertainment.
- Producing a bar plot barplot() shows that the groups are proportionate.



App Category	Proportion of Data
EDUCATION	29%
ENTERTAINMENT	25%
LIFESTYLE	22%
PRODUCTIVITY	24%

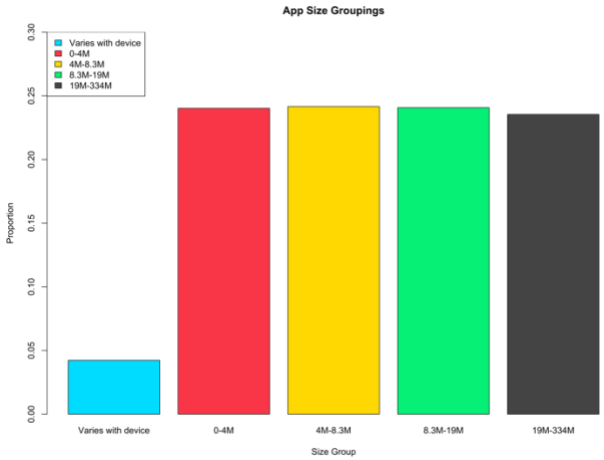
Reviews

- The data here shows up that there are a few apps that have received an extremely high number of reviews causing the variable to be right skewed.
- Performing a log transformation log() of the data does move the data closer to a normal distribution however there are still some upper outliers.
- By using the log transformation and removing the upper outliers we are able to achieve a somewhat normal distribution.



Size

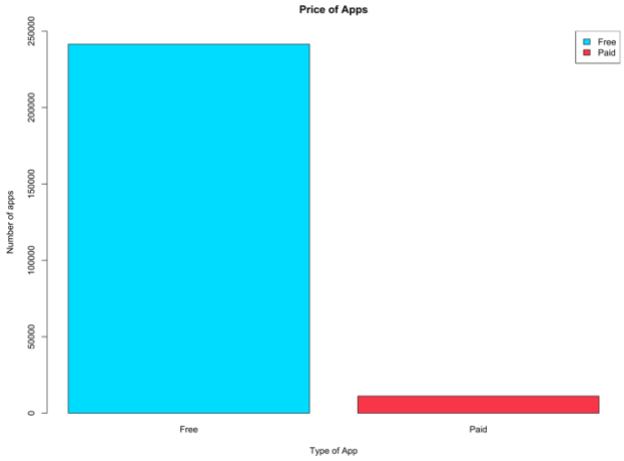
- Size was another variable that I thought would be numeric. But after examining the data I see that the size of many apps is “Varies by device.”
- Rather than exclude those apps from the data, I chose to make Size a factor and group by buckets.
- I first checked and saw that the minimum value for Size was 3.1 fivenum(). Knowing this, I could convert all “Varies by device” values to a “1” and I transform the variable to numeric as.numeric().
- I created a new column, Size.Groups, and used the cut() function to create groups.
- The bar plot shows that groups are proportional with the exception of “Varies by device”



Size Group	Proportion of Data
Varies with device	4%
0-4M	24%
4M-8.3M	24%
8.3M-19M	24%
19M-334M	24%

Price

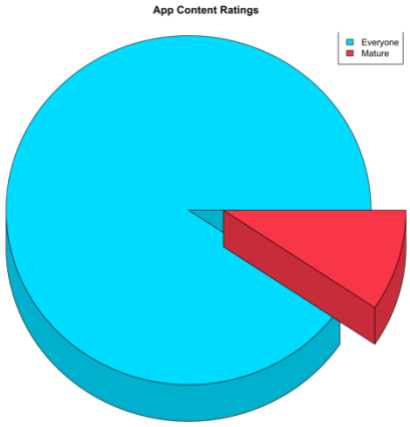
- Skewed data was an issue with price as well. I needed to convert to a factor since analysis showed that over 95% of the apps in this dataset were free.
- Price was split into two groups, Paid and Free.



Price Group	Proportion of Data
Free	96%
Paid	4%

Content Rating

- Although there were 5 levels for Content Rating, apps rated Everyone comprised of over 90% of the data leaving all other groups with less than 10% of data.
- To combat this, I grouped Everyone 10+, Teen, Mature 17+ and Unrated into one bucket called “Mature.”



Content Rating Group	Proportion of Data
Everyone	91%
Mature	9%