Google Play Store App Data Analysis

Statistical Analysis

Najja Osiomwan

November 18, 2020

Overview

Again, the purpose of this project is to analyze the Google Play Store data, explore relationships between variables and determine if Rating is a good predictor to determine app Installs by using logistic regression. In this section I look at the relationship between variables and answer questions about how certain variables may affect how consumers either rate, review or install these apps. In the previous section, I ensured that each variable was as close to a normal distribution as possible. Given the large dataset I am working with, over 267k samples, we can assume the central limit theorem applies.

Bivariate Relationships

Ratings & Reviews (figure 2.1)

Because these are the only numeric variables, I would like to see if there is a correlation between these variables and how strong it is. I first check to make sure there are no missing values in the data is.na() and there are not. A correlation test cor() from the corrplot package shows us that I can reject the null hypothesis that there is no correlation between the two variables. There is a slight negative correlation between Ratings and Reviews. The very low p-value tells us that this test is significant. The low correlation between the two variables also allows me to add both to the regression model since multicollinearity ill not be a problem.

Correlation	p-value
-0.1726915	< 0.00000000000000022

Figure 2.1

Installs & Size

In the previous section I transformed both the Size and Installs to categorical variables to make the data more manageable. A chi-squared test was necessary to see if there was a significant relationship between these Size and Installs. Because the p-value of the chi-squared test was 1, I conclude that size of app has no effect on the number of installs an app receives. This may actually be true if we assume that many consumers do not look at app size before downloading an app. However, the small amount of data we have on "varies with device" could be the reason the p-value is a 1. More data would be needed to conduct a more successful test.

Installs & Rating (figure 2.2)

Here we can see that apps with fewer installs have higher ratings on average. This makes sense because apps with a more installs would have a large number of ratings. Each rating would carry less weight and at the same time a large number of people rating an app would likely introduce a wider range of ratings, driving the average rating down over time.

The p-value for the ANOVA test was low enough to reject the null hypothesis.
I also used diagnostics plots to verify that the ANOVA assumptions were correct. The data does have some outliers but in general the plots look good. More specifically, the Residuals vs. Fitted plot shows that there is homogeneity among variances as there is no relationship between the residuals of each group. The Normal Q-Q plot tells us that residuals are normally distributed.

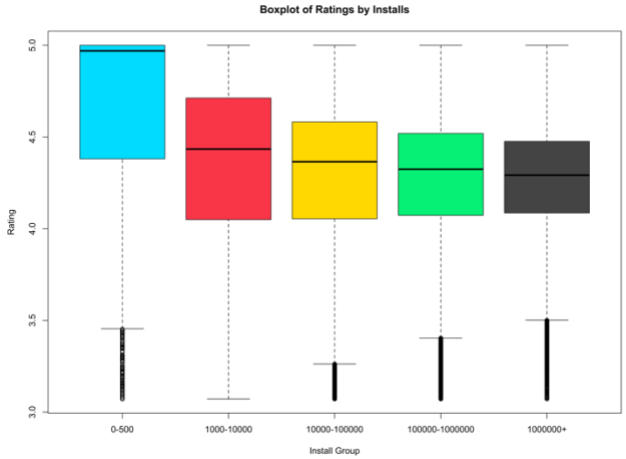

Figure 2.2

Installs & Reviews (figure 2.3)

As expected, the apps with over 1 million installs have the highest number of reviews on average and apps with fewer than 1,000 installs have the lowest number of reviews. The p-value for the ANOVA test is low enough to make the assumption that these results are correct. Installs and Reviews have a positive relationship with each other. Diagnostic plots look good as well.

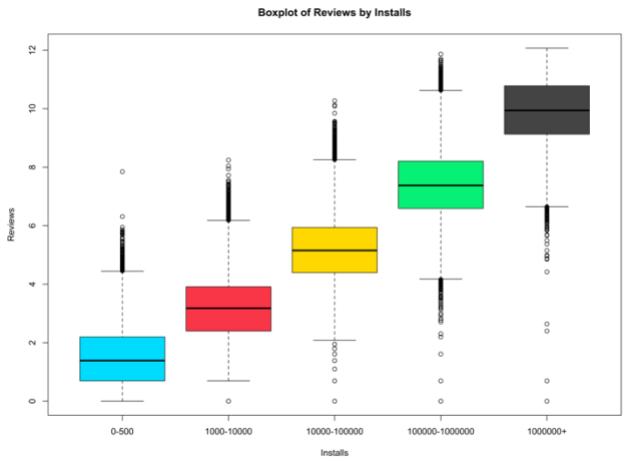

Figure 2.3

Installs & Price

A chi-squared test was performed on Installs and Price however the p-value came out to 0.99 causing me to accept the null hypothesis. This may be due to the disproportionate data available on Price with 95% of the apps in the dataset being free. I will not make any conclusions about Installs compared to the Price of an app.

Installs & Content Rating

The test results for Installs and Content Rating determined that relationship between these two variables were not significant. The p-value of 0.99 could be due to the disproportionate data available on Content Rating with 90% of the apps were rated "Everyone."

Installs & Category

In the data I have it looks as if the Entertainment category receives the highest proportion of installs over 1 million. But with a chi-squared p-value of 1, these results are not statistically significant.

Hypothesis Testing

I have made some assumptions about the Google Play store data and I attempt to verify those assumptions by running a series of statistical tests.

Rating & Category (figure 2.4)

Because education apps help people improve their skills, I figured that consumers who have a positive experience with an app and are actually learning from the app are more likely to take the time to leave a 5-star rating. This notion has helped me form my first hypothesis to analyze how users rate apps on the Google Play apps store.

> Hypothesis: On average, education apps will have higher ratings that apps in other categories.
>
> Results: Because Rating is a numeric variable and Category is a categorical variable I am using an ANOVA test aov() to test for differences in means between categories. By using

a box plot boxplot() you can see a slight difference in means. The test results tell us that these differences in means are significant, and we can reject the null hypothesis. In fact, Education apps are rated higher than all other categories of apps.

I used diagnostics plots plot() to verify the results of the ANOVA test. While the data does have some outliers, we can see that there is no correlation between residuals and fitted values and the Normal Q-Q plot approximately follows a 45-degree angle. The Scale-Location and the Residuals vs Leverage plots look good as well.

Category	diff	lwr	upr	p adj
ENTERTAINMENT-EDUCATION	-0.0220154	-0.0281104	-0.0159205	0
LIFESTYLE-EDUCATION	-0.0547612	-0.0610499	-0.0484725	0
PRODUCTIVITY-EDUCATION	-0.114749	-0.1209444	-0.1085536	0
LIFESTYLE-ENTERTAINMENT	-0.0327457	-0.0392447	-0.0262468	0
PRODUCTIVITY-ENTERTAINMENT	-0.0927336	-0.0991423	-0.0863248	0
PRODUCTIVITY-LIFESTYLE	-0.0599878	-0.066581	-0.0533946	0

Figure 2.4

Reviews & Category (figure 2.5)

Following the same mindset, I figured that Education apps would have more reviews than Lifestyle apps on average. Because I am analyzing just two categories this time, I am going to use a one-sided t-test to test for significance here.

Hypothesis: On average, Education apps will have a higher number of reviews than Lifestyle apps.

Results: With a p-value of 1 I will accept the null hypothesis. The difference in means between Education and Lifestyle apps is not statistically significant.

mean (LIFESTYLE)	mean (EDUCATION)	p-value
4.535357	4.372955	1

Figure 2.5

Ratings & Price (figure 2.6)

It seems that consumers who pay to download an app would be more satisfied with the app and rate the app higher. To prove this, I used an ANOVA test on the Rating and Price variables.

Hypothesis: On average, Paid apps will be rated higher than free apps.

Results: The test results are significant with a very low p-value. The diagnostics plots for this test also validate the ANOVA test result can be trusted.

Price	diff	lwr	upr	p adj
Paid-Free	0.03953342	0.03119399	0.04787285	0

Figure 2.6

Reviews & Content Rating (figure 2.7)
I used an ANOVA test to explore if there were any differences in the number of reviews left on apps with different content ratings. This could be used as a sort or proxy to determine if adults or children are more likely to leave reviews.

Hypothesis: Apps Rated Mature will have more reviews than apps for Everyone

Results: I reject the null hypothesis and conclude that apps rated mature have a higher number of reviews on average.

Content Rating	diff	lwr	upr	p adj
Mature-Everyone	0.779582	0.7452323	0.8139318	0

Figure 2.7

Regression

Simple Regression (figure 2.8)
In a previous section, I created a new variable named "InstallCat." This is a binary variable classifying each app as "High" installs for apps with 100k installs or more and "Low" installs for apps with less than 100k installs. I used InstallCat as the target variable for a simple logistic regression model to determine how well *Rating* couple predict the number of installs an app would receive.

Results
Call: formula = InstallCat ~ Rating

The coefficient for rating shows a positive relationship between these two variables. As the rating of an app increases, the log likelihood that the number of installs will increase also increases. These results are statistically significant with a p-value less than 0.05.

By using McFadden's R squared calculation for logistic regression, I see that the simple logistic regression model only explains about 1% of the variance in Installs. I will need a multiple logistic regression model to improve accuracy.

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-1.5498513	0.0444911	-34.83509	7.16E-266
Rating	0.6151776	0.01030832	59.67777	0.00E+00

Figure 2.8

Multiple Regression (figure 2.9)

To find a regression model with a high R squared value I will use multiple logistic regression, keeping InstallCat as the target variable. I will use forward selection, backward elimination and stepwise regression to help determine which variables should be used in the model.

Results

Call: formula = Installs ~ Reviews + Price + Rating + Category +
Size.Groups + Content.Rating

Each stepwise regression method returned the same model with the same AIC value of 89460. The final model below uses all relevant variables in the dataset and all variables are statistically significant at a 95% confidence level however, McFadden's R squared calculation shows that it only explains about 58% of the variance in the *InstallCat* variable. This tells me that the model could benefit from more data to improve prediction. If I were to calculate the accuracy of the model, perhaps setting the probability threshold to a value greater than 0.5 would improve the McFadden's R squared value. However, using test data to calculate the accuracy of the model is beyond the scope of this project.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.577643	0.107844	23.902	< 2e-16
Reviews	1.590614	0.008831	180.12	< 2e-16
PricePaid	-2.980053	0.036327	-82.034	< 2e-16
Rating	-1.097728	0.018374	-59.745	< 2e-16
CategoryENT	-0.217857	0.023541	-9.254	< 2e-16
CategoryLIFE	-0.173653	0.023423	-7.414	1.23E-13
CategoryPRO	-0.243129	0.024475	-9.934	< 2e-16
Size.Groups0	-0.251971	0.066287	-3.801	0.000144
Size.Groups4	-0.342221	0.066499	-5.146	2.66E-07
Size.Groups8	-0.307805	0.066705	-4.614	3.94E-06
Size.Groups1	-0.292086	0.066933	-4.364	1.28E-05
Content.Ratı	0.069374	0.032315	2.147	0.031811

Figure 2.9