Google Play Store Statistical Analysis

OMIS 324 Fall 2018

Report and Visuals created By:

John Acton (Johnacton68@Gmail.com)

Marcus Diaz (marcus.diaz1994@gmail.com)

Parth Shah (pshah5404@gmail.com)

Brandon Delgado (bgado25@gmail.com)

Jeremy Watson (oapwatson78@gmail.com)

Jake Werber (jake.werber@gmail.com)

Background

The Google Play Store is available on devices all around the world by being the primary operating platform of Android and Chrome users. The global digital distribution center was launched on October 28, 2008 and is ran by Google LLC. The main purpose of the Play Store is to allow users to browse, download, and review applications on a large platform accessible by users anywhere around the world. Being operated and run by Google is what allows the service to be available worldwide and in hundreds of languages. Additionally, Google is a highly profitable tech company that allows the store to run solely online and smoothly. The Google Play Store has thousands of apps ranging from free to $400 in price, and has hundreds of different categories ranging from games, entertainment, music, business, and much more. The store has something for everyone to utilize. The store also allows users who have downloaded the app to rate the app on a scale of zero to five with five being the best. This is a key feature because it allows other users to read rating and reviews from other users who have downloaded the app, and eventually make the decision on whether to buy the app and use it themselves. The Google play store as of 2017 has over 2 billion users worldwide, making it one of the largest e-commerce platforms in the world.

Main Goal

The main goal of our analysis on the Google Play Store is to determine if it is worth purchasing a paid application. We chose to focus solely on paid applications which brought our total dataset from over 1,000 values down to 627. After further examination of the data we noticed duplicates in the data set and removed them to bring our data down to 562. Of those

562, we chose a random sample of 100 applications under the price of $100 to analyze. We did this by collecting statistical data between paid apps, content ratings, installs, and the number of reviews they received. Our regression model was created by using a dependent variable of ratings, and four independent variables of price, reviews, installs, and content rating. In order to use content rating in our regression analysis we needed to create dummy variables for each type of content rating. We will also find which genre produces the highest average rating, and determine which genre produces the most applications to help us produce a better background on our Google Play analysis. The analysis and visuals we construct will help us determine our tested hypothesis and whether we reject or not.

Statistical Measures of Variables

To conduct our analysis we will use several variables. The first data type we use is nominal data, these are variables that are used to classify the data using words, letters or alphanumeric symbols. The nominal variables we use are Genre, and Content Rating. Genre is used to classify which type of app it is (Eg. Game, Business etc…). Content Rating classifies which age group is appropriate for the given app (Eg. Everyone, Teen. etc…). Next we use ordinal data that indicates an ordering of the data. We use one ordinal data type in our analysis and that's ratings. Ratings help determine whether users like the app on scale of one to five and if it is worth purchasing. Finally the last type of data we use is ratio. Ratio data has measurable intervals (Ex. An app with 20 installs or reviews has twice as many as one with 10). We use three ratio data types that are installs which count the number of installs the app has. Next we use reviews that tell us how many reviews a specific app has received from users. The last variable is

price which simply just tells us how much a specific app costs to download and help determine an application riskiness. By using these variables we hope we hope to find the relationships between the variables and what effect they have on one another.

Hypothesis

Null Hypothesis: Price, Number of Reviews, Installs and Content Rating do not have a significant effect on the app's rating.

Alternate hypothesis: Price, Number of Reviews, Installs, and Content Rating do have a significant effect on the app's rating.

Statistical Analysis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.68157113
R Square	0.46453921
Adjusted R Square	0.44199349
Standard Error	0.08964923
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	0.66238645	0.16559661	20.6043212	3.00282E-12
Residual	95	0.76351355	0.00803698		
Total	99	1.4259			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	4.73328273	0.02702692	175.132174	4.687E-121	4.679627516	4.78693794	4.67962752	4.78693794
Price	-0.0054703	0.00523852	-1.0442483	0.29902047	-0.015870091	0.00492946	-0.0158701	0.00492946
Reviews	-1.886E-06	1.3442E-06	-1.4030623	0.16385819	-4.55443E-06	7.8255E-07	-4.554E-06	7.8255E-07
Installs	-6.402E-08	1.7643E-07	-0.3628864	0.71749593	-4.14286E-07	2.8624E-07	-4.143E-07	2.8624E-07
Content Rating	0.04663935	0.00652113	7.15203316	1.7606E-10	0.033693268	0.05958544	0.03369327	0.05958544

For our first statistical analysis, we decided we would run regression analysis. In our

regression analysis, we decided we would look at the R Square, the Significance F, and the P-values. We ended up getting 0.46 for our R Square. This value shows that there is correlation between our variables. The correlation isn't strong, but it is within the acceptable range of correlation. We know the perfect R Square fit is 1.0 and that this is not around 1.0, but this is still an acceptable R Square. The Significance F is also lower than 0.05 which shows that the independent variables are significant within the dataset. What the Significance F is testing is whether the independent variables are significant in affecting the dependant variable. After that analysis we looked at the P Values. What the P Values are measuring is whether or not we should reject or accept the null hypothesis. We reject the null hypothesis if the P Values are below 0.05 because this shows that there is strong evidence against the null hypothesis. We chose 0.05 because our confidence interval is at 95%. We have three variables that have high P Values, but we have one variable that has a lower P Value than 0.05, and because of this we will reject the null hypothesis and we will go with the alternative hypothesis.

	Rating	*Price*	*Reviews*	*Installs*	*Content Rating*
Rating	1				
Price	-0.104916	1			
Reviews	0.147922	0.1950173	1		
Installs	0.1430544	0.1901139	0.8150994	1	
Content Ratir	0.0736806	-0.292446	-0.132979	-0.22904	1

For the second analysis we wanted to show was the correlation between the variables. In this chart it demonstrates how each variable correlates with each other. For the first column (Rating) our visual depicts that the relationship between rating and price is an extremely strong negative correlation due to the negative value of -0.104916. This value shows that the direction

of the line would slope downwards for its negative linear relationship. In the next column over, the relationship that is shared between price and reviews is a value of 0.1950173. What we can determine based off of this amount is that the direction of this line would trend upwards, however due to the value being relatively near 0, that informs us that the relationship between price and reviews is a weaker linear relationship. For reviews in comparison to installs, we calculated a value of 0.8150994. With this value, we assume that multicollinearity exists since this value exceeds the threshold of +/- 0.7. When this is the case, we have to be weary of the fact that our p-value may be inflated due to our independent variables sharing a strong correlation. From this information, we can conclude that all of the relationships with the exception of reviews by installs are reliable to use because their linear relationships do not result in multicollinearity existing. Since multicollinearity exists for reviews by installs, any small change in our x-value will yield a large change in the estimate afterwards.

For the third piece of our analysis, we wanted to decipher which genres produced the highest amount of reviews. In order to efficiently discover this information, we wanted to first create a pivot table in Microsoft Excel that would allow us to filter which fields we wanted to utilize and set the parameters that we were looking for accordingly. The chart on the left with the help of conditional formatting we can easily see which applications have the highest number of reviews. The number of reviews tells us the number of users that have interacted with an application in the genre and posted a review of that app. This also helps see some of the

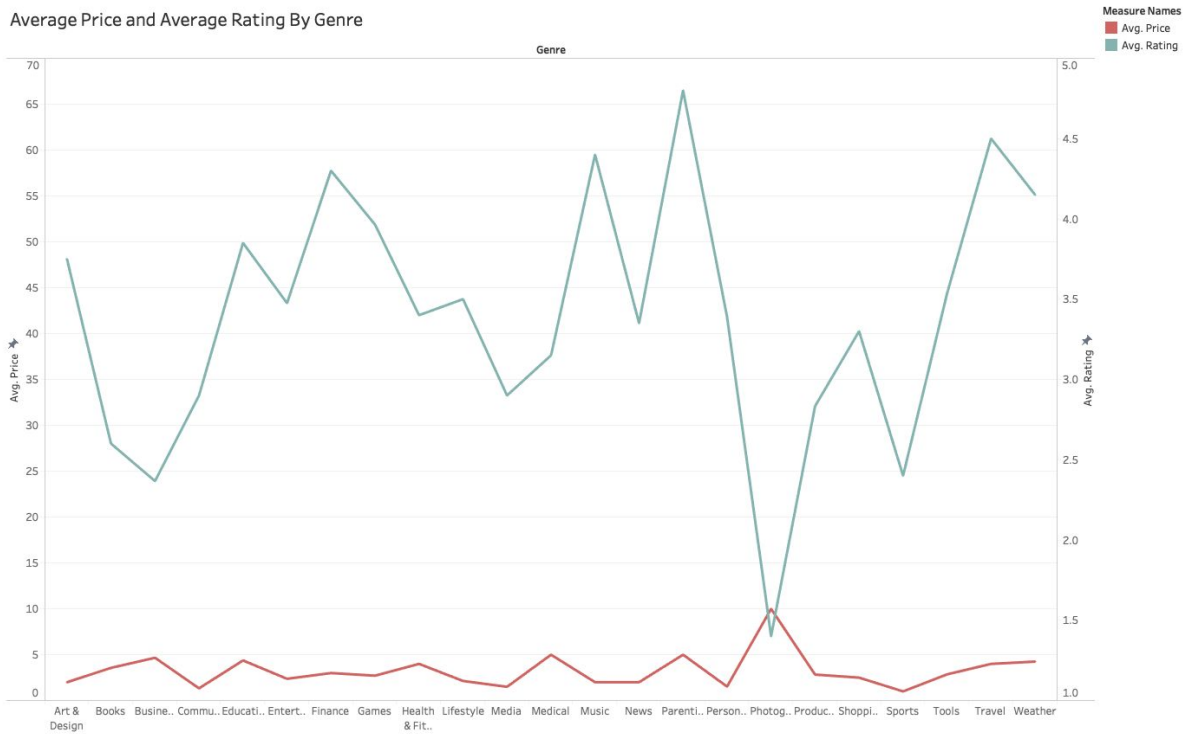
skewness of the data but this can be due to some genres not having as many apps as others. The chart on the right also with conditional formatting shows us the average rating of each application in the given genre of our sample and the number of installs. Conditional formatting really helps in this case because we are able to see that the genre has a applications in it that are rated very high in green but don't have a lot of installs which could mean the prices of app in that genre are high compared to other genres.

Row Labels	Count of Reviews
Personalization	18
Tools	14
Games	14
Education	12
Lifestyle	4
Books	4
Entertainment	4
Medical	4
Health & Fitness	3
Productivity	3
Communication	3
Business	3
News	2
Art & Design	2
Weather	2
Shopping	1
Travel	1
Sports	1
Photography	1
Music	1
Media	1
Parenting	1
Finance	1
Grand Total	**100**

Row Labels	Average of Rating	Sum of Installs
Parenting	4.8	50000
Travel	4.5	5000
Music	4.4	100000
Finance	4.3	100000
Weather	4.2	200000
Games	4.0	613170
Education	3.9	1201330
Art & Design	3.8	15000
Tools	3.5	450620
Lifestyle	3.5	15110
Entertainment	3.5	800
Health & Fitness	3.4	111000
Personalization	3.4	179620
News	3.4	5500
Shopping	3.3	100
Medical	3.2	10700
Communication	2.9	5200
Media	2.9	10000
Productivity	2.8	600010
Books	2.6	20150
Sports	2.4	1000
Business	2.4	201000
Photography	1.4	1000
Grand Total	**3.481**	**3896310**

Finally, our last visual shows the relationship between Price and Rating by Genre. By

creating this visual, we produced an observable correlation between the highest price and the average ratings in the given genre to determine which is worth exploring more. For example, photography has the highest average price at $9.99, while also having the lowest average rating at 1.4 out of 5. This shows us that we do not want to explore the photography genre because they have a negative relationship with an average rating being far lower than average price. By determining theses correlations, we are able to give users who are willing to pay for an application a good reference to base their search on. With this visual it related to the pivot table charts above but this gives us and the user the ability to see it in a more visual friendly way that is more appealing.



Conclusion and Recommendation

In conclusion, the main goal of our analysis was to figure out if it is worth purchasing a paid application in the Google Play store. We did this by running a regression analysis and creating multiple visuals to gain insightful knowledge about our dataset. The results of our regression showed that we would reject the null hypothesis since the p-value of content rating is less than .05. By rejecting our null hypothesis this proves our belief that Price, Number of Reviews, Number of Installs, and Content Rating do have a significant effect on the app's rating.

The Google Play Store has hundreds of thousands of applications, and users normally do not want to waste their time on an app that will not satisfy their needs. Our analysis led to the conclusion that apps with high ratings are influenced by a number of variables as stated in our analysis. This provided great information from the potential buyer of the application because a number of these variables that we tested come from first hand users of the application. This is especially important for the user because once an application is purchased there are no refunds on the application, so it provides the buyer with the satisfaction they are getting a good product. Therefore, we recommend that once the buyer finds an application they are interested in buying,

after they look at the price and the content rating is appropriate for them, they should then look at the number of installs it has and the rating. If the installs are high along with a high rating we believe the buyer will be happy with the application and will not have buyers remorse.

Work Cited

kaggle datasets download -d lava18/google-play-store-apps