DAO2702
PROGRAMMING FOR BUSINESS ANALYTICS

Google Play Store Applications Market

Tutorial Group: G10 **Date of Submission:** 23 Nov 2018

Group Members	Matric Number
Ng Heng Rui	A0168628N
Darren Darius Tan Jia Wei	A0173269W
Tan Sumin	A0171133U

Table of Contents

1. Introduction to Google Play Store Market

1.1 Mobile Application Industry

The mobile application industry is proliferating at an exponential rate. According to Statista, in 2015, global mobile app sales amounted to 69.7 billion U.S. dollars. It is projected that by 2020, mobile apps will be able to generate 188.9 billion U.S. dollars in revenues via app stores and in-app advertising.

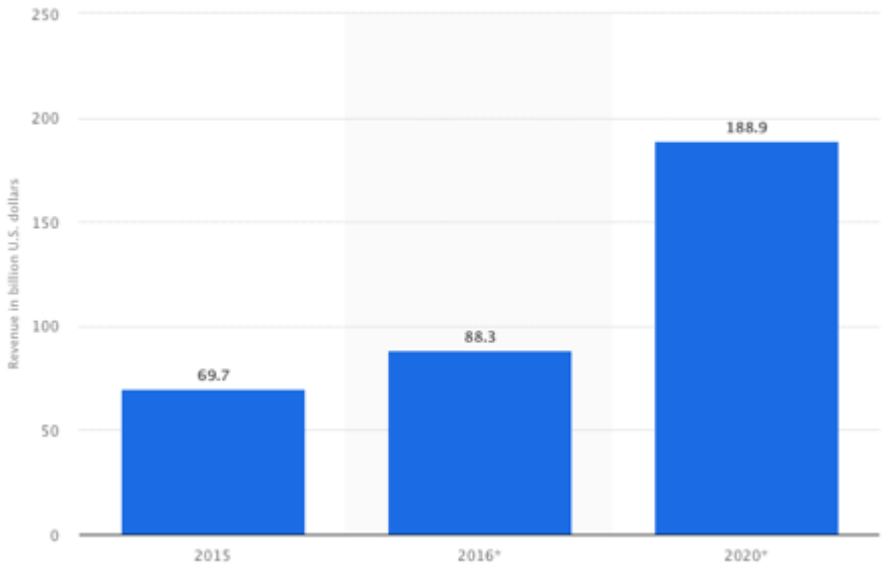
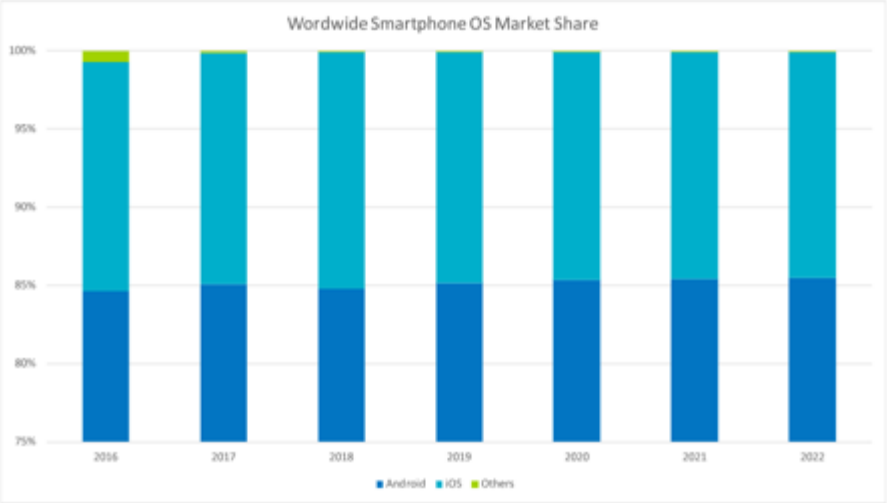


Figure 1: Projection of mobile application industry

From: ***https://www.statista.com/statistics/269025/worldwide-mobile-app-revenue-forecast/***

The mobile application market is dominated by two shareholder who is Google Play and Apple App Store. According to IDC, Google controls 84.8% of the market while Apple holds 15.1% of the market. The Android platform is in the lead with an average salary of $97.6K amongst developers. Android is a well-regarded platform because of its continued market penetration. Apple's iOS platform mid-level iOS developers make about $96.6K a year. Both Android and iOS generate a large revenue to their top developers.



Year	2016	2017	2018	2019	2020	2021	2022
Android	84,6%	85,1%	84,8%	85,2%	85,3%	85,4%	85,5%
iOS	14,7%	14,7%	15,1%	14,8%	14,6%	14,6%	14,5%
Others	0,7%	0,2%	0,1%	0,1%	0,1%	0,1%	0,1%
TOTAL	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Figure 2:

From: ***https://www.idc.com/promo/smartphone-market-share/os***

1.2 Background

Our team, Positivevibes, is a group of new application developers who decided to enter the Google Play Market. This is because the Android market is cheaper and significantly larger, owning more than 80% of the market share. To distribute through the official marketplaces, developers must pay an up-front membership fee. The fee for Google Play is $25 and the fee for Apple is $99. Furthermore, we would like to test our application before bringing it to Apple market. Hence, we would start our application on the Android market first.

1.3 Business Problem

1.3.1 Google Play Dataset

Our Team has obtained a data set of Google Play with over 10,000 application providing data such as the application name, category, overall user rating, number of user reviews, size of the app, number of user downloads, paid or free, price and genre. Data Obtained from: https://www.kaggle.com/lava18/google-play-store-apps

We decided to analyze this data set to find out what type of application would fit our company's goals and vision.

1.3.2 Goals and Vision

Being a new start-up and having limited resource and manpower, we aim to create a great application to kickstart our company. Such application will should have a high rating to build a strong customer base. According to studies by Google, it shows that good rating is the key to raise awareness in the store and ultimately drive user downloads. Of the top 100 free iPhone apps, 67% earned ratings of four stars or higher, and 92% of them have ratings higher than three stars. These ratings have a huge impact on whether users will download a foreign app. And with high turnover rates in the app space, users are almost always looking for better substitutes.

1.3.3 Data Source

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000000.0	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700000.0	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000000.0	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800000.0	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up

Hence, our business problem would be to **determine the mix of variables that help to maximise Rating on Google Play** in our given data file, before we look towards creating an Application with attributes of those identified variables. We will also be **looking at other forms of research data** to assist in our decision on this particular mix of variables.

1.4 Model of Analysis

1.4.1 Modelling Methodologies

Positive Vibes plans to use both **Simple Linear Regression** and **Multiple Linear Regression** to identify any trends between any of the given variables in the data and Rating. As we plan to maximise Rating, it becomes a y-variable that is also a dependent variable on other variables. For all other given variables in the data, they are x-variables which are independent variables affecting Rating. All the variables from our csv file we are including in our regression analyses are summarized in Table 1 below, with Table 2 explaining the variables we are intentionally excluding thereafter below..

Variable	Relationship	Variable Type	Conversion (Later Stages)
Y: Rating	Independent Variable	Numerical	N.A.
X1: Category	Dependent Variable	Categorical	N.A.
X2: Reviews		Numerical	N.A.
X3: Size		Numerical	N.A.
X4: Installs		Categorical	To Numerical
X5: Price		Numerical	To Binary (Categorical)

Table 1: Variables for Regression Analysis

For Simple Linear Regression, we maximise Rating by determining how much it increases for every unit increase in the x-variables. This is based on any positive relationships identified from any positive coefficients between x-variables and Rating. Subsequently, to condense the various relationships between each of the x-variables and Rating, Multiple Linear Regression is then conducted to form only 1 equation involving all the x-variables and Rating. Furthermore,

multiple linear regression is conducted to predict a single x-variable from one or more independent x-variables.

After finding out the relationship between the respective y and x variables in simple and multiple regression, we aim to target our operations towards increasing the results of the x-variable that has the strongest relationship with the y-variable, Rating. This is so as to boost the Ratings with the greatest certainty which is our objective.

1.4.2 Modelling Assumptions

The following assumptions would have to be made before we could model our regression line, summarized with Table 2 below. Our assumptions would then be evaluated in Section 6 below due to the existence of some limitations to our model.

No.	Assumptions	Explanation
1	All x-variables are **independent** of one another and are only potentially correlated Rating.	Linear Regression assumes that there is **little or no multicollinearity** between our x-variables, so that the change in value of one variable would not affect others. Independence between each x-variable would then estimate the relationship accurately between each x-variable and Rating because the x-variables will not change in unison with one another for every change in Rating.
2	Positive Vibes are able to **directly control** the values of x-variables in order to increase Rating	This is to ensure not only the absence of correlation between x-variables, but also **absence of causation** of one x-variable over another. If this happens, there would be dependency and no control over the latter x-variable on the former.
3	Based on independence of x-variables, the **ability to remove outliers** in each x-variable's data distribution so that there is greatest central tendency around the mean value.	Linear Regression also assumes **multivariate normality** of each independent x-variable, which ideally translates into a symmetric normal distribution of x-values over each value of Rating. Thus in here we assume remove outliers is possible to ensure each x-variable has the least distributional error for each value of Rating.
4	**Homoscedasticity** of each independent x-variable.	We assume that the variance around the regression line is the same for all values of the predictor x-variable. This is

		important because with **unequal variances of error** for each x-value observed, the cases with larger variance disturbances have larger distortion on the value of Rating.

1.4.3 Conclusion For Data Analysis

Assuming that all x-variables are independent, the data shall be first cleaned, manipulated and simplified so that it can be useful for Simple Linear regression. To obtain the general descriptive statistics of each variable such as the **central tendency and spread**, as well as the **distribution of each variable**, data visualisation is conducted on each x-variable to obtain its distribution. Based on each x-variable's data distribution, if it is not normally distributed such that there is **asymmetric error or variance around the mean**, the range of outliers in each distribution is then identified. Subsequently, by using the **sample analogue method** to **estimate our parameters around the mean value** of each x-variable which covers about 95% of our data, the extreme end(s) of each x-variable's distribution is removed such that it **models a normal distribution** as much as possible.

Next with 95% of coverage probability in the data for each x-variable, we then proceed to conduct hypothesis testing of each x-variable against Rating to test if there is **statistical significance** of the data of each x-variable against Rating. If proven statistically significant, we then conduct **Simple Linear Regression** to obtain the relationship between any two x-y variables. The **relationship will also be aided with data visualisation** for 2-variables to check if the relationship is indeed linear.

If the x-y relationships are non-linear, **transformations** would be conducted into a linear form. Subsequently, we then proceed to consider if the x-variables could possibly have any **interaction** effect on other variables, through slicing each x-variables data into different ranges to check for the presence of **confounders**.

Penultimately, we then conduct **Multiple Linear Regression** taking into consideration interaction and transformation effects.

Lastly, we evaluate the validity of our regression analysis in retrospect to the assumptions made in Table 2. Based on the syllabus, more focus will be placed on parameter estimation and

determining the **confidence intervals** of each x-variables by **error-minimization method**, even if other limitations are found. We will then conclude with the interpretations of our results of the regression model based on the syllabus concepts, and bring in other forms of data to support our Business Problem.

2. Data Manipulation and Cleaning

2.1 Drop Missing Data

Firstly, we drop all the missing data (i.e. NaN) in any of the variables, which changes our population into a sample. In addition, our group have decided to drop all the unwanted columns below to only obtain data with the variables above. Here are the following reasons for each of the variables dropped:

Unused Variables	Reasons:
Type	'Type' classifies 'Price' as a binary variable, instead of its absolute numerical price, on whether an Application is Paid or Free. Due to overlapping variables, 'Type' is removed.
Content Rating	Positive Vibes aims to establish a general application for all to use, and would not want to look at particular age groups.
Genres	'Genre' is a subset of the variable 'Category' where even though it gives better elaboration of information in each Category, there are already 33 categories. Thus, it is better to simplify the Applications based on Categories rather than genres for data analysis, such that the data frame does not become too complicated.
Last Updated Ver	An existing application would first have to be in the market, and thus the these three sections would not be relevant as Positive Vibes has not yet launched the application.
Current Ver	
Android Ver	

Table 2: Variables excluded from regression analysis

2.2 Conversion to Numerical Variable

Secondly, upon using the value count and describe functions for all variables above to see the general statistics of each variable, independent of other variables. In the process, we realise that some numerical variables data types are objects. We then convert 'Installs' and 'Price' which are numerical data to floats, while keeping all other numerical data as floats or integers and categorical data as objects.

Thirdly, we used the Split-apply-combine model to obtain the general statistics of each numerical variable, after taking into account the effects of each categorical variable, "Installs" and "Category". At this juncture, due to the presence of too many categories and sub-categories, we have decided to treat "Installs" as a discrete numerical variable rather than as a Category, so as to simplify the regression analysis. Below is a table of the descriptive statistics of each variable:

Variable	Mean	S.D.	Category Counts	Distribution
Rating	4.17	0.544	N.A.	Requires Data Visualisation
Category	N.A.	N.A.	33	
Reviews	2.95 x 10^5	1.86 x 10^6	N.A.	
Size	2.29 x 10^7	2.34 x 10^7	N.A.	
Installs	8.41 x 10^6	5.01 x 10^7	19	
Price	1.12	17.4	N.A.	

3. Data Visualisation

3.1 Distribution of Variables

We aim to check for outliers in our given data for each x-variable against Rating through data visualisation of the relationship between the 2 variables. To do so, we first conduct data visualization for 1 variable (i.e. each x-variable or y-variable) by first determining the range of each variable's data with low-frequency counts. We do so by data visualisation for the distribution pattern of each respective variable, combined with estimation o ref the parameters of each variable by the sample analogue method.

For visualising the distribution of categorical variables, we used a histogram. Meanwhile for visualising the numerical variables, as we were unsure of the distribution, we used the Kernel Density Estimation plot. As expected, the data mostly did not reflect a normal distribution, but rather an F-distribution or chi-square distribution which is not in our syllabus.

3.2 Descriptive Measures

Next, we use the sample analogue method to obtain a range of data (Table 3). With the help of data visualisation of each variable, we estimate the percentiles at which we obtain about 95% of our data. The remaining 5% of our data, which may potentially be outliers due to the low-frequency counts, lie on the extreme lower or upper end or both of each variable which we then aim to get rid of. By reducing the range of each variable to concentrate at where there is higher frequency counts, it helps in improving our sample.

Variable	Central Tendency		Standard Deviation	Distribution Type	Location of Parameters (Percentiles)		Parameters Obtained	
	Mean	Median			Lower	Upper	Lower	Upper
Rating	4.16	4.30	0.553	Chi-square (F-dist.)	0	95	1.0	4.8
Reviews	5.18e+03	1.82e+04	1.36e+05	Exponential	0	95	1.0	295510.29999999993

Size	2.16e+07	1.30e+07	2.24e+07	Chi-square (F-dist.)	0	95	8500.0	72000000.0
Installs	9.740460	10.000000	3.038665	Binomial	2.5	97.5	4.0	14.0
Price	1.179341	0.000000	17.8	Exponential (Bimodal)	0	95	0.0	2.49

Table 3: Descriptive measures of variables

3.3 Relationship Analysis

Subsequently, we then visualise the relationship between each of the given x-variable and Ratings to determine the range of data in each of the x-variables which contains the outliers. As Rating is a Numerical Variable, we used a scatter plot with histograms for plotting against numerical variables while we used side by side boxplots for plotting against categorical variables. After the visualisation between bivariate numerical data, we have also obtained correlation coefficients between each x-variable and Rating (Section 4).

3.3.1 Reviews and Price

Based on our findings, “Reviews” and “Price” had the greatest range of outliers which we would want to drop. However, we could not drop the upper range of “Reviews” and upper range of “Price” simultaneously, because many of the Applications that had high reviews were mostly of low prices and that Applications with high prices were mostly of low reviews. This is because based on research, people would rather install applications that are of low prices in the first place before they could give a review, leading to a higher frequency of reviews. The converse is true as well, where higher prices of applications would lead to fewer installations, leading to a lower frequency of reviews. Thus if we were to get rid of data from both the upper ends of “Reviews” and “Prices”, there would be very few data rows left for analysis.

Based on the correlation between “Reviews” and “Price” x-variables, our group has then decided to drop the upper range of “Reviews” while splitting the “Price” data into 2 categories, where one is for cheaper applications below $4.99 and another for expensive applications above $4.99.

Thus, our data frame has been reduced to exclude Reviews that are more than 1 million to remove the majority of the outliers, bearing in mind our company's objective to introduce new applications which could not possi itbly reach a high number of reviews over a short period of time. Furthermore, given the causation of "Installs" on "Reviews", we have identified the existence of some correlation between the x-variables which will be evaluated in our report later in section 7.

Hence, due to the limitations of removing outliers, we have only removed data sets with 1 million or higher "Reviews" so that there is a smaller spread of data. However given the correlation of the already removed "Reviews" data with other x-variables (as they are not independent), it is not possible to determine a 95% Confidence Interval for each x-variable which by the sample analogue method might have removed some of the data within the Confidence Interval of that particular x-variable.

4. Simple Linear Regression Modelling

After treatment of some outliers based on "Reviews", we then obtain a resultant set of data for each x and y variable. At this stage, we perform hypothesis testing between each x-variable on Rating, to see if each x-variable is of statistical significance to correlate with Rating.

H0 : $\beta 1 = 0$ (there is no relationship between X and Y)
H1: $\beta 1 \neq 0$

To determine whether we should reject H0 or not, we then perform simple linear regression and calculate the p-value to see if each x-variable (as shown in Table 1) is statistically significant in order to determine whether there is linear relationship between each x-variable and Rating. The lower the p-value, the higher the statistical significance, the stronger the evidence against the null hypothesis, to prove there is relationship between X and Y.

X variables	
X_1	Category
X_2	Reviews
X_3	Size
X_4	Installs
X_5	Price Initial stage: Prices above and below $4.99 [numerical variables] Current stage: Free (i.e. Price = 0) or Paid (i.e. Price > 0) [categorical variables]

Table 4: List of X variables for simple linear regression

4.1 Category, Reviews, Size and Installs

For X_1 variable [Category], based on the regression results, Table 5 below shows that 20 out of 33 categories are statistically significant and have an impact Ratings, even though the relationship may be very weak. We might need to consider some factors to evaluate the reasons in Section 6.3 below.

1. Auto and Vehicles
2. Business
3. Comics
4. Communication
5. Dating
6. Entertainment
7. Family
8. Finance
9. Food and Drinks
10. House and Home
11. Lifestyle
12. Maps and Navigation
13. Medical
14. News and Magazines
15. Photography
16. Productivity
17. Sports
18. Tools
19. Travel and Local
20. Video Players

Categories:	R2 Value	Coefficient	p-value
Auto and Vehicles	0.030	-0.2134	0.000
Business		-0.2426	0.002

Comics		-0.2304	0.029
Communication		-0.2876	0.000
Dating		-0.4032	0.000
Entertainment		-0.2110	0.022
Family		-0.1780	0.014
Finance		-0.2523	0.001
Food and Drinks		-0.2634	0.005
House and Home		-0.1985	0.050
Lifestyle		-0.2671	0.000
Maps and Navigation		-0.3514	0.000
Medical		-0.1768	0.020
News and Magazines		-0.2172	0.000
Photography		-0.2341	0.000
Productivity		-0.2353	0.000
Sports		-0.1627	0.040
Tools		-0.3623	0.000
Travel and Local		-0.3179	0.000
Video Players		-0.3530	0.000

Table 5: List of Categories Significant against Rating

In addition, the p-values for Reviews (X_2), Size (X_3) and Installs (X_4) are around 0, which is less than 0.05, showing that the variables are statistically significant too. This is reflected in Table 6 below.

X-variable against Rating	R2 value	Coefficients	p-value
Reviews	0.018	5.473e-07	0.000
Size	0.004	1.494e-09	0.000
Installs	0.003	0.0100	0.000
Price	0.000	-0.0006	0.085

Table 6: List of Other Variables Significant Against Rating

4.2 Price

However, the p-value for Price (X_5) is 0.085, which is more than 0.05, showing that X_5 variable is not statistically significant. Hence, based on the above classification where Prices were split into above and below $4.99, we decided to change "Price" from a numerical variable into a binary variable, where it is either Free (i.e. Price = 0) or Paid (i.e. Price > 0). After the modification, the results show that there is statistical significance between Price as a binary variable against Rating. Before conducting simple linear regression for each x-variable and Rating, the variables have been finalised into Table 1 as above.

5. Multiple Linear Regression Modelling

5.1 Initial Multiple Regression Model

We combine all the x-variables and y-variable into a Multiple Linear Regression Line initially without any modifications done to the regression equation. Below are the summarized results:

OLS Regression Results

Dep. Variable:	Rating	R-squared:	0.048
Model:	OLS	Adj. R-squared:	0.043
Method:	Least Squares	F-statistic:	10.22
Date:	Wed, 21 Nov 2018	Prob (F-statistic):	4.56e-55
Time:	09:20:24	Log-Likelihood:	-5929.4
No. Observations:	7390	AIC:	1.193e+04
Df Residuals:	7353	BIC:	1.219e+04
Df Model:	36		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.3856	0.074	59.061	0.000	4.240	4.531
Category[T.AUTO_AND_VEHICLES]	-0.2170	0.098	-2.212	0.027	-0.409	-0.025
Category[T.BEAUTY]	-0.0645	0.113	-0.569	0.570	-0.287	0.158
Category[T.BOOKS_AND_REFERENCE]	-0.0491	0.084	-0.587	0.557	-0.213	0.115
Category[T.BUSINESS]	-0.2479	0.079	-3.154	0.002	-0.402	-0.094
Category[T.COMICS]	-0.2281	0.105	-2.181	0.029	-0.433	-0.023
Category[T.COMMUNICATION]	-0.3075	0.081	-3.816	0.000	-0.465	-0.150
Category[T.DATING]	-0.4090	0.082	-5.011	0.000	-0.569	-0.249
Category[T.EDUCATION]	0.0050	0.087	0.058	0.954	-0.166	0.176
Category[T.ENTERTAINMENT]	-0.2644	0.091	-2.896	0.004	-0.443	-0.085
Category[T.EVENTS]	0.1195	0.113	1.061	0.289	-0.101	0.340
Category[T.FAMILY]	-0.2009	0.072	-2.791	0.005	-0.342	-0.060
Category[T.FINANCE]	-0.2588	0.078	-3.316	0.001	-0.412	-0.106
Category[T.FOOD_AND_DRINK]	-0.2818	0.092	-3.063	0.002	-0.462	-0.101
Category[T.GAME]	-0.1926	0.074	-2.615	0.009	-0.337	-0.048
Category[T.HEALTH_AND_FITNESS]	-0.1551	0.079	-1.956	0.050	-0.311	0.000
Category[T.HOUSE_AND_HOME]	-0.2015	0.101	-1.995	0.046	-0.399	-0.004
Category[T.LIBRARIES_AND_DEMO]	-0.1580	0.098	-1.606	0.108	-0.351	0.035
Category[T.LIFESTYLE]	-0.2693	0.078	-3.470	0.001	-0.421	-0.117
Category[T.MAPS_AND_NAVIGATION]	-0.3571	0.090	-3.972	0.000	-0.533	-0.181
Category[T.MEDICAL]	-0.1786	0.077	-2.325	0.020	-0.329	-0.028
Category[T.NEWS_AND_MAGAZINES]	-0.2402	0.082	-2.935	0.003	-0.401	-0.080
Category[T.PARENTING]	-0.0173	0.108	-0.160	0.873	-0.229	0.194
Category[T.PERSONALIZATION]	-0.0507	0.078	-0.653	0.514	-0.203	0.102
Category[T.PHOTOGRAPHY]	-0.2775	0.079	-3.491	0.000	-0.433	-0.122
Category[T.PRODUCTIVITY]	-0.2470	0.079	-3.116	0.002	-0.402	-0.092
Category[T.SHOPPING]	-0.1783	0.082	-2.185	0.029	-0.338	-0.018
Category[T.SOCIAL]	-0.1469	0.082	-1.789	0.074	-0.308	0.014
Category[T.SPORTS]	-0.1920	0.079	-2.436	0.015	-0.346	-0.037
Category[T.TOOLS]	-0.3727	0.074	-5.053	0.000	-0.517	-0.228
Category[T.TRAVEL_AND_LOCAL]	-0.3376	0.083	-4.091	0.000	-0.499	-0.176
Category[T.VIDEO_PLAYERS]	-0.3843	0.087	-4.412	0.000	-0.555	-0.214
Category[T.WEATHER]	-0.1246	0.104	-1.198	0.231	-0.329	0.079
Reviews	5.745e-07	5.45e-08	10.539	0.000	4.68e-07	6.81e-07
Size	3.884e-10	3.24e-10	1.201	0.230	-2.46e-10	1.02e-09
Installs	-0.0042	0.003	-1.647	0.100	-0.009	0.001
Price	-0.0005	0.000	-1.404	0.160	-0.001	0.000

Omnibus:	2492.857	Durbin-Watson:	1.846
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10344.197
Skew:	-1.621	Prob(JB):	0.00
Kurtosis:	7.804	Cond. No.	2.02e+09

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.02e+09. This might indicate that there are
strong multicollinearity or other numerical problems.

5.2 Transformations

Firstly, based on data visualisation, we see that Reviews better suit a model of square root function with Ratings when compared in Section 4 in Simple Linear Modelling with the summarized results in Table 6. We apply Reviews as a square root function of Rating and find that there is a stronger relationship than between Reviews as a function of Rating.

OLS Regression Results

Dep. Variable:	Rating	R-squared:	0.030
Model:	OLS	Adj. R-squared:	0.030
Method:	Least Squares	F-statistic:	227.4
Date:	Wed, 21 Nov 2018	Prob (F-statistic):	1.22e-50
Time:	16:00:46	Log-Likelihood:	-5997.8
No. Observations:	7390	AIC:	1.200e+04
Df Residuals:	7388	BIC:	1.201e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.0968	0.008	533.875	0.000	4.082	4.112
np.power(Reviews, 0.5)	0.0005	3.37e-05	15.081	0.000	0.000	0.001

Omnibus:	2340.890	Durbin-Watson:	1.829
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9198.754
Skew:	-1.533	Prob(JB):	0.00
Kurtosis:	7.524	Cond. No.	276.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Secondly, again based on data visualisation, we see that the relationship between the logarithmic function of Size against Rating better suits the relation between Size as a function of Rating as seen in Table 6 above.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Rating   R-squared:                       0.005
Model:                            OLS   Adj. R-squared:                  0.005
Method:                 Least Squares   F-statistic:                     38.30
Date:                Wed, 21 Nov 2018   Prob (F-statistic):           6.40e-10
Time:                        16:00:48   Log-Likelihood:                -6090.8
No. Observations:                7390   AIC:                         1.219e+04
Df Residuals:                    7388   BIC:                         1.220e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.6639      0.081     45.372      0.000       3.506       3.822
np.log(Size)   0.0306      0.005      6.188      0.000       0.021       0.040
==============================================================================
Omnibus:                     2550.545   Durbin-Watson:                   1.811
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10423.316
Skew:                          -1.669   Prob(JB):                         0.00
Kurtosis:                       7.765   Cond. No.                         206.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Lastly, based on data visualisation, we see that the relationship between the square of Installs as a function of Rating is better than that Installs as a function of Ratings, as seen in Table 6 above.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Rating   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     54.20
Date:                Wed, 21 Nov 2018   Prob (F-statistic):           2.01e-13
Time:                        09:20:21   Log-Likelihood:                -6082.8
No. Observations:                7390   AIC:                         1.217e+04
Df Residuals:                    7388   BIC:                         1.218e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.0756      0.013    304.689      0.000       4.049       4.102
np.power(Installs, 2)     0.0008      0.000      7.362      0.000       0.001       0.001
==============================================================================
Omnibus:                     2336.072   Durbin-Watson:                   1.822
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8868.122
Skew:                          -1.544   Prob(JB):                         0.00
Kurtosis:                       7.390   Cond. No.                         247.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

5.2 Interactions

Based on research, there is an association between Reviews and Categories because review styles vary according to software categories and thus may be biased.

OLS Regression Results

Dep. Variable:	Rating	R-squared:	0.056
Model:	OLS	Adj. R-squared:	0.048
Method:	Least Squares	F-statistic:	6.446
Date:	Wed, 21 Nov 2018	Prob (F-statistic):	1.48e-53
Time:	09:20:27	Log-Likelihood:	-5895.0
No. Observations:	7390	AIC:	1.193e+04
Df Residuals:	7321	BIC:	1.240e+04
Df Model:	68		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.3955	0.078	56.274	0.000	4.242	4.549
Category[T.AUTO_AND_VEHICLES]	-0.2359	0.103	-2.283	0.022	-0.439	-0.033
Category[T.BEAUTY]	-0.1047	0.122	-0.857	0.392	-0.344	0.135
Category[T.BOOKS_AND_REFERENCE]	-0.0397	0.089	-0.448	0.654	-0.213	0.134
Category[T.BUSINESS]	-0.2529	0.083	-3.055	0.002	-0.415	-0.091
Category[T.COMICS]	-0.2154	0.110	-1.949	0.051	-0.432	0.001
Category[T.COMMUNICATION]	-0.3205	0.086	-3.724	0.000	-0.489	-0.152
Category[T.DATING]	-0.4339	0.087	-5.003	0.000	-0.604	-0.264
Category[T.EDUCATION]	-0.0095	0.099	-0.097	0.923	-0.203	0.184
Category[T.ENTERTAINMENT]	-0.2360	0.102	-2.312	0.021	-0.436	-0.036
Category[T.EVENTS]	0.1744	0.121	1.436	0.151	-0.064	0.413
Category[T.FAMILY]	-0.1840	0.076	-2.415	0.016	-0.333	-0.035
Category[T.FINANCE]	-0.2553	0.082	-3.101	0.002	-0.417	-0.094
Category[T.FOOD_AND_DRINK]	-0.2887	0.099	-2.918	0.004	-0.483	-0.095
Category[T.GAME]	-0.1732	0.078	-2.210	0.027	-0.327	-0.020
Category[T.HEALTH_AND_FITNESS]	-0.1973	0.085	-2.328	0.020	-0.363	-0.031
Category[T.HOUSE_AND_HOME]	-0.2067	0.107	-1.930	0.054	-0.417	0.003
Category[T.LIBRARIES_AND_DEMO]	-0.1321	0.104	-1.273	0.203	-0.336	0.071
Category[T.LIFESTYLE]	-0.2670	0.082	-3.266	0.001	-0.427	-0.107
Category[T.MAPS_AND_NAVIGATION]	-0.3939	0.096	-4.088	0.000	-0.583	-0.205
Category[T.MEDICAL]	-0.2034	0.081	-2.510	0.012	-0.362	-0.045
Category[T.NEWS_AND_MAGAZINES]	-0.2047	0.087	-2.357	0.018	-0.375	-0.034
Category[T.PARENTING]	-0.0081	0.112	-0.073	0.942	-0.227	0.211
Category[T.PERSONALIZATION]	-0.0345	0.082	-0.420	0.675	-0.196	0.127
Category[T.PHOTOGRAPHY]	-0.2761	0.085	-3.238	0.001	-0.443	-0.109
Category[T.PRODUCTIVITY]	-0.2506	0.084	-2.986	0.003	-0.415	-0.086
Category[T.SHOPPING]	-0.1446	0.087	-1.655	0.098	-0.316	0.027
Category[T.SOCIAL]	-0.0974	0.088	-1.113	0.266	-0.269	0.074
Category[T.SPORTS]	-0.1646	0.084	-1.962	0.050	-0.329	-0.000
Category[T.TOOLS]	-0.3779	0.078	-4.843	0.000	-0.531	-0.225

Category[T.TRAVEL_AND_LOCAL]	-0.3666	0.088	-4.175	0.000	-0.539	-0.194
Category[T.VIDEO_PLAYERS]	-0.3901	0.093	-4.195	0.000	-0.572	-0.208
Category[T.WEATHER]	-0.1241	0.115	-1.083	0.279	-0.349	0.100
Reviews	1.147e-06	1.36e-06	0.841	0.400	-1.53e-06	3.82e-06
Category[T.AUTO_AND_VEHICLES]:Reviews	1.199e-06	1.94e-06	0.617	0.537	-2.61e-06	5.01e-06
Category[T.BEAUTY]:Reviews	9.272e-06	7.81e-06	1.187	0.235	-6.04e-06	2.46e-05
Category[T.BOOKS_AND_REFERENCE]:Reviews	-5.724e-07	1.48e-06	-0.387	0.698	-3.47e-06	2.32e-06
Category[T.BUSINESS]:Reviews	1.659e-07	1.46e-06	0.114	0.910	-2.7e-06	3.03e-06
Category[T.COMICS]:Reviews	-7.858e-07	2.5e-06	-0.314	0.754	-5.69e-06	4.12e-06
Category[T.COMMUNICATION]:Reviews	-1.354e-07	1.4e-06	-0.097	0.923	-2.88e-06	2.61e-06
Category[T.DATING]:Reviews	9.566e-07	1.54e-06	0.621	0.535	-2.06e-06	3.98e-06
Category[T.EDUCATION]:Reviews	-1.227e-07	1.49e-06	-0.083	0.934	-3.04e-06	2.79e-06
Category[T.ENTERTAINMENT]:Reviews	-6.679e-07	1.4e-06	-0.478	0.633	-3.41e-06	2.07e-06
Category[T.EVENTS]:Reviews	-1.499e-05	1.18e-05	-1.275	0.202	-3.8e-05	8.05e-06
Category[T.FAMILY]:Reviews	-7.387e-07	1.37e-06	-0.541	0.589	-3.42e-06	1.94e-06
Category[T.FINANCE]:Reviews	-3.951e-07	1.4e-06	-0.282	0.778	-3.14e-06	2.35e-06
Category[T.FOOD_AND_DRINK]:Reviews	-1.98e-07	1.46e-06	-0.135	0.892	-3.06e-06	2.67e-06
Category[T.GAME]:Reviews	-6.375e-07	1.37e-06	-0.467	0.641	-3.32e-06	2.04e-06
Category[T.HEALTH_AND_FITNESS]:Reviews	5.963e-07	1.41e-06	0.422	0.673	-2.18e-06	3.37e-06
Category[T.HOUSE_AND_HOME]:Reviews	4.575e-08	1.63e-06	0.028	0.978	-3.14e-06	3.23e-06
Category[T.LIBRARIES_AND_DEMO]:Reviews	-1.559e-06	1.93e-06	-0.809	0.419	-5.34e-06	2.22e-06
Category[T.LIFESTYLE]:Reviews	-2.732e-07	1.41e-06	-0.194	0.846	-3.04e-06	2.49e-06
Category[T.MAPS_AND_NAVIGATION]:Reviews	1.381e-06	1.71e-06	0.807	0.420	-1.97e-06	4.74e-06
Category[T.MEDICAL]:Reviews	6.276e-06	2.25e-06	2.789	0.005	1.87e-06	1.07e-05
Category[T.NEWS_AND_MAGAZINES]:Reviews	-1.001e-06	1.39e-06	-0.720	0.471	-3.72e-06	1.72e-06
Category[T.PARENTING]:Reviews	-5.253e-07	1.59e-06	-0.330	0.742	-3.65e-06	2.6e-06
Category[T.PERSONALIZATION]:Reviews	-7.935e-07	1.41e-06	-0.562	0.574	-3.56e-06	1.98e-06
Category[T.PHOTOGRAPHY]:Reviews	-4.545e-07	1.38e-06	-0.330	0.741	-3.15e-06	2.24e-06
Category[T.PRODUCTIVITY]:Reviews	-2.122e-07	1.4e-06	-0.151	0.880	-2.96e-06	2.54e-06
Category[T.SHOPPING]:Reviews	-7.913e-07	1.38e-06	-0.574	0.566	-3.5e-06	1.91e-06
Category[T.SOCIAL]:Reviews	-1.055e-06	1.38e-06	-0.764	0.445	-3.76e-06	1.65e-06
Category[T.SPORTS]:Reviews	-8.275e-07	1.38e-06	-0.599	0.549	-3.54e-06	1.88e-06
Category[T.TOOLS]:Reviews	-1.642e-07	1.37e-06	-0.119	0.905	-2.86e-06	2.53e-06
Category[T.TRAVEL_AND_LOCAL]:Reviews	2.529e-07	1.41e-06	0.179	0.858	-2.51e-06	3.02e-06
Category[T.VIDEO_PLAYERS]:Reviews	-3.376e-07	1.39e-06	-0.243	0.808	-3.07e-06	2.39e-06
Category[T.WEATHER]:Reviews	-1.524e-07	1.92e-06	-0.079	0.937	-3.92e-06	3.62e-06
Size	4.648e-10	3.25e-10	1.429	0.153	-1.73e-10	1.1e-09
Installs	-0.0064	0.003	-2.488	0.013	-0.011	-0.001
Price	-0.0005	0.000	-1.397	0.162	-0.001	0.000

Omnibus:	2512.698	Durbin-Watson:	1.852
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10607.032
Skew:	-1.629	Prob(JB):	0.00
Kurtosis:	7.882	Cond. No.	2.14e+09

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.14e+09. This might indicate that there are
strong multicollinearity or other numerical problems.

Also as mentioned above, price (X_5) may affect the number of installs (X_4) because in the first place if consumers need to pay, they would not want to install the application.

OLS Regression Results

Dep. Variable:	Rating	R-squared:	0.048
Model:	OLS	Adj. R-squared:	0.043
Method:	Least Squares	F-statistic:	9.947
Date:	Wed, 21 Nov 2018	Prob (F-statistic):	1.47e-54
Time:	09:20:29	Log-Likelihood:	-5929.4
No. Observations:	7390	AIC:	1.193e+04
Df Residuals:	7352	BIC:	1.220e+04
Df Model:	37		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.3856	0.074	59.050	0.000	4.240	4.531
Category[T.AUTO_AND_VEHICLES]	-0.2170	0.098	-2.212	0.027	-0.409	-0.025
Category[T.BEAUTY]	-0.0645	0.113	-0.569	0.570	-0.287	0.158
Category[T.BOOKS_AND_REFERENCE]	-0.0491	0.084	-0.587	0.557	-0.213	0.115
Category[T.BUSINESS]	-0.2479	0.079	-3.154	0.002	-0.402	-0.094
Category[T.COMICS]	-0.2281	0.105	-2.181	0.029	-0.433	-0.023
Category[T.COMMUNICATION]	-0.3075	0.081	-3.816	0.000	-0.465	-0.150
Category[T.DATING]	-0.4090	0.082	-5.011	0.000	-0.569	-0.249
Category[T.EDUCATION]	0.0050	0.087	0.058	0.954	-0.166	0.176
Category[T.ENTERTAINMENT]	-0.2644	0.091	-2.896	0.004	-0.443	-0.085
Category[T.EVENTS]	0.1195	0.113	1.061	0.289	-0.101	0.340
Category[T.FAMILY]	-0.2009	0.072	-2.791	0.005	-0.342	-0.060
Category[T.FINANCE]	-0.2587	0.078	-3.315	0.001	-0.412	-0.106
Category[T.FOOD_AND_DRINK]	-0.2818	0.092	-3.063	0.002	-0.462	-0.101
Category[T.GAME]	-0.1926	0.074	-2.615	0.009	-0.337	-0.048
Category[T.HEALTH_AND_FITNESS]	-0.1551	0.079	-1.956	0.050	-0.311	0.000
Category[T.HOUSE_AND_HOME]	-0.2015	0.101	-1.995	0.046	-0.400	-0.003
Category[T.LIBRARIES_AND_DEMO]	-0.1580	0.098	-1.606	0.108	-0.351	0.035
Category[T.LIFESTYLE]	-0.2694	0.078	-3.470	0.001	-0.422	-0.117
Category[T.MAPS_AND_NAVIGATION]	-0.3571	0.090	-3.972	0.000	-0.533	-0.181
Category[T.MEDICAL]	-0.1786	0.077	-2.324	0.020	-0.329	-0.028
Category[T.NEWS_AND_MAGAZINES]	-0.2402	0.082	-2.935	0.003	-0.401	-0.080
Category[T.PARENTING]	-0.0173	0.108	-0.160	0.873	-0.229	0.194
Category[T.PERSONALIZATION]	-0.0507	0.078	-0.653	0.514	-0.203	0.102
Category[T.PHOTOGRAPHY]	-0.2775	0.079	-3.491	0.000	-0.433	-0.122
Category[T.PRODUCTIVITY]	-0.2470	0.079	-3.116	0.002	-0.402	-0.092
Category[T.SHOPPING]	-0.1783	0.082	-2.184	0.029	-0.338	-0.018
Category[T.SOCIAL]	-0.1469	0.082	-1.789	0.074	-0.308	0.014
Category[T.SPORTS]	-0.1920	0.079	-2.435	0.015	-0.347	-0.037
Category[T.TOOLS]	-0.3727	0.074	-5.052	0.000	-0.517	-0.228
Category[T.TRAVEL_AND_LOCAL]	-0.3376	0.083	-4.091	0.000	-0.499	-0.176
Category[T.VIDEO_PLAYERS]	-0.3843	0.087	-4.412	0.000	-0.555	-0.214
Category[T.WEATHER]	-0.1246	0.104	-1.198	0.231	-0.329	0.079
Reviews	5.746e-07	5.45e-08	10.536	0.000	4.68e-07	6.81e-07
Size	3.886e-10	3.24e-10	1.201	0.230	-2.46e-10	1.02e-09
Installs	-0.0042	0.003	-1.647	0.100	-0.009	0.001
Price	-0.0006	0.002	-0.328	0.743	-0.004	0.003
Installs:Price	1.167e-05	0.000	0.048	0.962	-0.000	0.000

Omnibus:	2492.880	Durbin-Watson:	1.846
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10344.486
Skew:	-1.621	Prob(JB):	0.00
Kurtosis:	7.804	Cond. No.	2.02e+09

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.02e+09. This might indicate that there are
strong multicollinearity or other numerical problems.

5.3 Fitting of Final Multiple Regression Model

We plot the relationship between Rating and all the x-variables simultaneously, using multiple linear regression. Over here, we take into account the interactions and transformation of the regression equation.

OLS Regression Results

Dep. Variable:	Rating	R-squared:	0.082
Model:	OLS	Adj. R-squared:	0.073
Method:	Least Squares	F-statistic:	9.486
Date:	Wed, 21 Nov 2018	Prob (F-statistic):	1.93e-91
Time:	09:20:34	Log-Likelihood:	-5793.4
No. Observations:	7390	AIC:	1.173e+04
Df Residuals:	7320	BIC:	1.221e+04
Df Model:	69		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.2959	0.119	35.991	0.000	4.062	4.530
Category[T.AUTO_AND_VEHICLES]	-0.2709	0.112	-2.408	0.016	-0.491	-0.050
Category[T.BUSINESS]	-0.2945	0.090	-3.269	0.001	-0.471	-0.118
Category[T.COMMUNICATION]	-0.3635	0.095	-3.843	0.000	-0.549	-0.178
Category[T.DATING]	-0.4648	0.095	-4.880	0.000	-0.651	-0.278
Category[T.FAMILY]	-0.1902	0.083	-2.298	0.022	-0.352	-0.028
Category[T.FINANCE]	-0.2981	0.090	-3.330	0.001	-0.474	-0.123
Category[T.FOOD_AND_DRINK]	-0.2954	0.111	-2.659	0.008	-0.513	-0.078
Category[T.GAME]	-0.1998	0.086	-2.336	0.020	-0.368	-0.032
Category[T.HEALTH_AND_FITNESS]	-0.2949	0.094	-3.151	0.002	-0.478	-0.111
Category[T.LIFESTYLE]	-0.2869	0.089	-3.223	0.001	-0.461	-0.112
Category[T.MAPS_AND_NAVIGATION]	-0.4756	0.107	-4.462	0.000	-0.684	-0.267
Category[T.MEDICAL]	-0.2653	0.088	-3.013	0.003	-0.438	-0.093
Category[T.NEWS_AND_MAGAZINES]	-0.2208	0.095	-2.327	0.020	-0.407	-0.035
Category[T.PHOTOGRAPHY]	-0.3293	0.094	-3.486	0.000	-0.514	-0.144
Category[T.PRODUCTIVITY]	-0.2983	0.092	-3.244	0.001	-0.479	-0.118
Category[T.TOOLS]	-0.4144	0.085	-4.879	0.000	-0.581	-0.248
Category[T.TRAVEL_AND_LOCAL]	-0.4100	0.096	-4.253	0.000	-0.599	-0.221
Category[T.VIDEO_PLAYERS]	-0.4203	0.102	-4.119	0.000	-0.620	-0.220
Category[T.MEDICAL]:np.power(Reviews, 0.5)	0.0019	0.001	2.483	0.013	0.000	0.003
np.power(Installs, 2)	-0.0014	0.000	-7.392	0.000	-0.002	-0.001
Priceb[T.Paid]:np.power(Installs, 2)	0.0021	0.001	3.114	0.002	0.001	0.003

5.4 Model Assessment

Next, we conduct a model assessment to see the combined effect of all variables, transformations and interactions when put in a multiple linear regression line, checking for the goodness of fit and model parsimony of our resultant model.

Based on our results, there is a very weak relationship between the x-variables and Rating indicated by the R2 value, showing that none of the 5 variables affects Rating, but possibly some other factors. However given our Adjusted R2 value which is lower than the R2 value, it shows that our model could have a better fit and it should ideally match R2 at 0.082. The AIC and BIC value are 11730 and 12210 respectively which is relatively high, suggesting that our multiple linear regression model might consist of too many x-variables as predictors and that possibly, we should take into consideration results from Simple Linear Regression as it only considers 1 x-variable at a time as a predictor of Rating.

6. Evaluation: Parameter Estimation and Confidence Intervals

Lastly, as we suspect that our estimation of parameters by coercing a 95% significant data range through sample analogue method and through observation from data visualisation is ineffective, we would evaluate our Confidence Intervals (CI) of our x-variable through Error Minimisation method.

6.1 Limitations to Modelling Assumptions

Since we have removed those data points that have Reviews above 1 million (Section 3.3.1), the associated values of other x-variables that has more than 1 million reviews might be in the original confidence intervals of the other x-variables. This results in the old confidence intervals having many missing data points. Hence, we will be calculating the coverage probabilities (Section 6.2) to ensure that our sample data is significant at the 95% confidence level.

6.2 Error-Minimisation and Confidence Intervals (CI)

Error-minimization model: $Y_i = \mu + \epsilon_i$

where i is the application index in the sample, and $i = 1, 2, \ldots 7390$

μ is the population parameter to represent averages of various variables (i.e. reviews, size, installs, price)

ϵ is the application i's error, the deviation from the average

Through Error Minimisation method, we obtain the best error-minimising mean values (i.e. x_bar and y_bar) (Table 6). Through assuming a normal distribution under Central Limit Theorem, it helps us to determine how much the set of data covers the best mean value of each variable, also known as the Confidence Intervals. The calculated parameters of each variable's CIs then determine the coverage probability of each variable set of data amongst all the outliers. This step is conducted as after taking into account all the transformation of the variables, we do not know which data points are considered outliers. This is because we do not know the true population variance, as well as not knowing the real data distribution after the removal of outliers.

Variable	Best New Mean	Coverage Probability of Data

		(given best mean value)
Reviews	51,759.9484	0.9467
Size	21,633,404.1819	0.9497
Installs	9.7393	0.9561
Price	1.1910	0.9515
Rating	4.1620	Not needed

Table 7: Coverage probability of data for each variable

Hence, after constructing the CIs, we can determine if the remaining data in each x-variable is still statistically significant to conclude an association with y-variable, Rating. It is done by referring to the coverage probabilities of all the numerical x-variables which reveals that are still very close to 0.95. Therefore as the coverage probability is remained unchanged, all the x-variables still pose statistical significant with Rating. Thus, we can conclude that despite modifications in the data set, the relationship of each x-variable with Rating still remains valid when conducting Simple and Multiple Linear Regression.

6.3 Evaluation of Modelling Methodology

However, as seen in the R2 value of the Simple linear and Multiple linear regression models above in Table 5 and 6, there is a very weak relationship between each individually transformed x-variable and Rating, and between all variables and Rating in the final regression model. Given the very weak relationship between the x-variables and Rating in Simple and Multiple Linear Regression models, we then question our assumptions made in our models to see if they are true in the first place.

Our first assumption were that all the x-variables are independent on each other, however after calculating the product-moment-correlation-coefficients between each x-variable in Table 7 below, we found that there is a non-zero correlation between any 2 x-variables. This effectively translates into all x-variables having significant interdependency on one another, instead of all being independent of one another. Therefore, this goes against our assumption of the absence of multicollinearity between each x-variable.

Variable 1	Variable 2	Correlation Coefficient
Reviews	Size	0.2349556
Reviews	Installs	0.50302441
Reviews	Price	-0.02429811
Size	Installs	0.27068609
Size	Price	-0.0242148
Price	Installs	-0.05762671

Table 7: Correlation coefficients

In addition, our second assumption is deemed invalid as we have proven the causation of "Price" variable on "Installs" which will then further cause an effect on "Reviews" in Section 3.3.1. This means that not only there is correlation between all the x-variables which are supposed to be independent variables, but there are also causation of some x-variables over one another.

The invalidity of the first two assumptions would then mean that if we try to remove outliers outside of the confidence interval of the first x-variable so as to reduce the spread of error over a

mean value of x (homoscedasticity), these outliers which are dependent on the other x-variables might be correlated to data points that are within the confidence intervals of the other x-variables. Hence if we were to remove outliers in the first x-variable, relevant data points within confidence intervals of other x-variables will be removed as well. Therefore, the assumption of multivariate normality in the latter x-variables will be challenged while ensuring a normal distribution of errors in any one x-variable. This is because of reduced homoscedasticity in the latter x-variables if data points were removed asymmetrically, rendering the third assumption invalid too.

Lastly, the removal of data points within the confidence intervals of other-x-variables might further reduce the coverage probability of its data over its calculated best mean value, which may even result in the x-variable being statistically insignificant to affect the y-variable Rating. Thus, we have conducted Parameter Estimation using Error-Minimization method to ensure the coverage probability of our x-variables over their respective mean values.

In conclusion, with the above assumptions rendered invalid, there are significant limitations to our model (which we will not resolve as it is not in our DAO2702 syllabus). This explains the very weak relationships between our supposed independent variables and dependent variables; and our interpretations of our data results would have to be supported with other research findings, both elaborated in Section 7 below.

.

7. Conclusion: Interpretation of Findings

All the x variables and y variables are correlated to one another, with some variables not having enough data to provide a good coverage probability of over 95% over the best mean value. These makes it difficult to have a significant trend to identify. There are only two variable mix which has p-value lower than 0.05 and has a positive coefficient. They are price [paid] and category: medical with reviews. Hence, Positive Vibes conclude that we will be creating an paid, medical application with high quantity of review. This would ensure our company goal being fulfilled bringing high satisfaction to our customer.

Variable Mix	p-value	Top Ranked Coefficients
Price [Paid]	0.002	0.0021
Category: Medical with Reviews	0.013	0.0019

References:

Guppda L. (2018 GooglePlay DataSet. Retrieved 23 November 2018 from ***https://www.kaggle.com/lava18/google-play-store-apps***

IDC (2018) Smartphone OS market share. Retrieved 23 November 2018 from ***https://www.idc.com/promo/smartphone-market-share/os***

Mackenzie T (2012). App store fees, percentages, and payouts: What developers need to know Retrieved 23 November 2018 from ***https://www.techrepublic.com/blog/software-engineer/app-store-fees-percentages-and-payouts-what-developers-need-to-know/***

Tiongson J (2015) Mobile app marketing insights: How consumers really find and use your apps Retrieved 23 November 2018 from ***https://www.thinkwithgoogle.com/consumer-insights/mobile-app-marketing-insights/***

Statista (2018) Worldwide mobile app revenues in 2015, 2016 and 2020 (in billion U.S. dollars) . Retrieved 23 November 2018 from ***https://www.statista.com/statistics/269025/worldwide-mobile-app-revenue-forecast/***

Wano M. and Lio J. (2014) Relationship between Reviews at App Store and the Categories for Software. Retrieved 23 November 2018 from ***https://ieeexplore.ieee.org/document/7024016/authors#authors***