Google Play Store App Analysis

SOR1232 – Hypothesis Testing

Supervised by Dr. David Suda

Chris Frendo (439600L) (Computer Science 1st Year)

Manwel Bugeja (454000L) (Computer Science 1st Year)

Domenico Agius (6500H) (Computer Science 1st Year)

L-Università ta' Malta
Faculty of Science

Department of Statistics & Operations Research

Declaration of Authorship

I, Chris Frendo (1396001), [1] declare that this assignment entitled:

"Google Play Store App Analysis [2]" and the work presented in it are my own.

I confirm that:

1. Where any part of this assignment has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

2. This work is submitted in ~~partial~~ [3] fulfilment of the requirements of the credit SOR1232 [4] offered by the Department of Statistics and Operations Research, Faculty of Science, University of Malta.

3. Where I have consulted the published work of others, this is always clearly attributed.

4. Where I have quoted form the works of others, the source is always given. With the exception of such quotations, this assignment is entirely my own work. 5. I have acknowledged all sources used for the purpose of this work.

6. I have read the guidelines are regulations of the University of Malta regarding plagiarism and understand that the penalties for committing a breach of the regulations include the loss of marks, cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Signature: _____

Date: 12-06-19

[1] Insert name, surname and ID card number
[2] Insert title of Assignment
[3] Remove the word partial where appropriate
[4] Insert code of credit

L-Università ta' Malta
Faculty of Science

Department of Statistics & Operations Research

Declaration of Authorship

I, Manuel Bugeja 454099L [1] declare that this assignment entitled:

"Google Play Store App Analysis [2]" and the work presented in it are my own.

I confirm that:

1. Where any part of this assignment has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

2. This work is submitted in ~~partial~~ [3] fulfilment of the requirements of the credit SOR1232 [4] offered by the Department of Statistics and Operations Research, Faculty of Science, University of Malta.

3. Where I have consulted the published work of others, this is always clearly attributed.

4. Where I have quoted form the works of others, the source is always given. With the exception of such quotations, this assignment is entirely my own work. 5. I have acknowledged all sources used for the purpose of this work.

6. I have read the guidelines are regulations of the University of Malta regarding plagiarism and understand that the penalties for committing a breach of the regulations include the loss of marks, cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Signature: Manuel Bugeja

Date: 12 Jun 2019

[1] Insert name, surname and ID card number
[2] Insert title of Assignment
[3] Remove the word partial where appropriate
[4] Insert code of credit

L-Università ta' Malta
Faculty of Science

Department of Statistics & Operations Research

Declaration of Authorship

I, DOMENICO AGIUS (65004H) [1] declare that this assignment entitled:

"Google Play Store App Analysis [2]" and the work presented in it are my own.

I confirm that:

1. Where any part of this assignment has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

2. This work is submitted in ~~partial~~ [3] fulfilment of the requirements of the credit SOR1232 [4] offered by the Department of Statistics and Operations Research, Faculty of Science, University of Malta.

3. Where I have consulted the published work of others, this is always clearly attributed.

4. Where I have quoted form the works of others, the source is always given. With the exception of such quotations, this assignment is entirely my own work. 5. I have acknowledged all sources used for the purpose of this work.

6. I have read the guidelines are regulations of the University of Malta regarding plagiarism and understand that the penalties for committing a breach of the regulations include the loss of marks, cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Signature: Domenico Agius

Date: 6/6/2019

[1] Insert name, surname and ID card number
[2] Insert title of Assignment
[3] Remove the word partial where appropriate
[4] Insert code of credit

Table of Contents

1 Introduction

The chosen dataset has to do with Google Play Store applications. It can be found at the following link: https://www.kaggle.com/lava18/google-play-store-apps. This dataset was chosen because it contains data that can provide actionable insight on what makes an application successful on this platform. This dataset contains data on around 10,000 Play Store applications which were scraped from the Google Play Store itself. The original dataset contains 13 attributes that describe each application however for the purpose of this assignment only 6 of these were kept. The variables used are listed below:

- Rating (Covariate and Dependent variable)
- Reviews (Covariate)
- Size (Covariate)
- Installs (Factor)
- Type (Factor)
- Content_Rating (Factor)

The variable that is of most interest is *Rating* as it gives the best indication on how successful an app is. The *Reviews* attribute indicates how many reviews (positive or negative ones) an app has. The *Size* variable holds the size in kilobytes for each app. The *Installs* factor is used to indicate how many installs (based on a range) the app has. The *Type* factor indicates if the app is *Free* or *Paid* and the *Content_Rating* factor indicates for which age group the app is targeted.

2 Aims and Objectives

The objective of this assignment was to figure out if there were any correlations between the *Rating* and any of the other variables. This would be useful to identify what makes an application successful on the Google Play Store. Hypothetically it makes sense to assume that an application which is paid should have a higher rating. Moreover, if an application has a large number of installs it also makes sense to expect a higher rating. Also, through the tests the ideal demographical target of an app should be found by finding which factor in the *Content_Rating* variable has the highest rating. Regarding *size* there are two possibilities, either an application with a large size gets a higher rating due to its better quality or else small sized apps get a higher rating because they do not take up as much of the space on their device (which can often be limited).

3 Descriptive Statistics & Illustrations

3.1 Measurements of Location

This section will explain the measurements of locations obtained for the covariate variables and the frequencies obtained for the factors.

3.1.1 Rating

Rating	Mean		3.622	.0145
	95% Confidence Interval for Mean	Lower Bound	3.594	
		Upper Bound	3.651	
	5% Trimmed Mean		3.751	
	Median		4.200	
	Variance		2.293	
	Std. Deviation		1.5142	
	Minimum		.0	
	Maximum		5.0	
	Range		5.0	
	Interquartile Range		.8	
	Skewness		-1.765	.024
	Kurtosis		1.561	.047

Table 1 Descriptives for Rating

Table 1 contains the measurements of location for the *Rating* covariate. The range, minimum and maximum clearly indicate that this rating ranges from 0 to 5. The average rating is 3.622 which shows that more applications in the dataset have a higher rating. In fact, this can be verified by the median which is 4.200 and by the skewness which is -1.765.

This negative skewness shows that the distribution of ratings is skewed to the right: towards the higher values. The kurtosis value (1.561) shows that people prefer to give either a very high or a very low rating instead of a medium rating. The 5% trimmed mean is 3.751 which shows that there is a higher number of lower rated extreme cases since this trimmed mean is greater than the actual mean. The standard deviation is relatively high considering the small range which shows that the ratings are also quite spread.

Figure 1 shows the box plot for the *Rating* covariate. The line inside the box represents the median which also lies around 4.2 and shows that the data is skewed since it is not equidistant from the hinges. It is negatively skewed since it closer to the 75th percentile. The box plot shows that there are many outliers or possible outliers in this dataset.

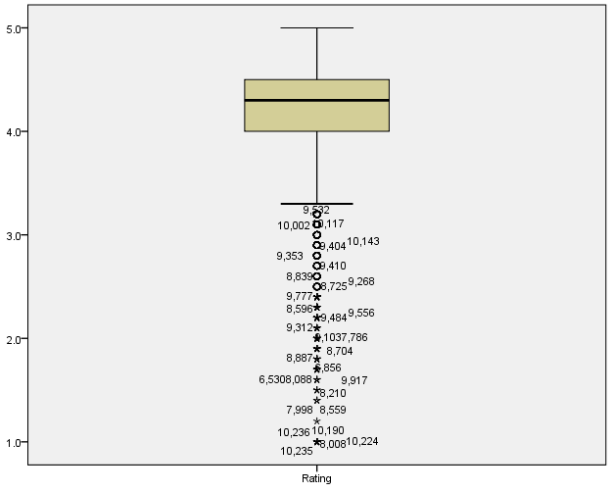


Figure 1 Boxplot for Rating

The histogram for *Rating* can be seen in figure 2. It also shows the negative skewness towards the higher rating. The data in the histogram shows a large kurtosis which suggests that there are more applications in this dataset with a high rating. The histogram seems to have a distribution however it is not a normal one due its skewness and kurtosis.

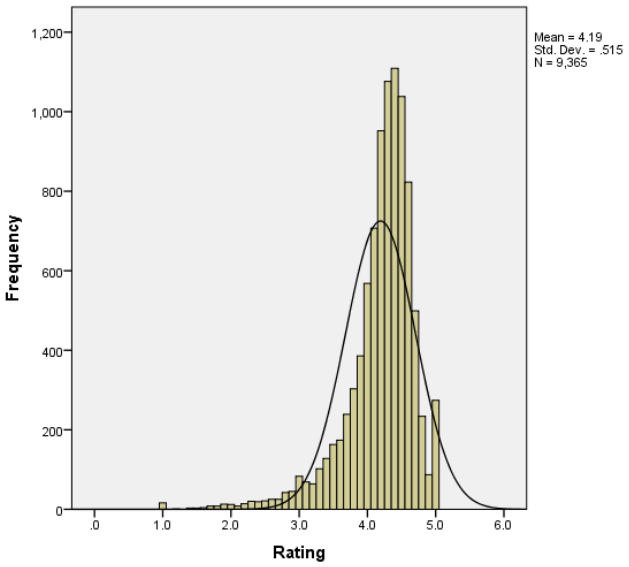


Figure 2 Histogram for Rating

3.1.2 Reviews

Reviews	Mean		444193.87	28122.928
	95% Confidence Interval for Mean	Lower Bound	389067.79	
		Upper Bound	499319.95	
	5% Trimmed Mean		80483.37	
	Median		2094.00	
	Variance		8572554850856.179	
	Std. Deviation		2927892.561	
	Minimum		0	
	Maximum		8E+007	
	Range		78158306	
	Interquartile Range		54760	
	Skewness		16.449	.024
	Kurtosis		341.029	.047

Table 2 Descriptives for Reviews

Immediately it is noticeable that there is a large number of extreme cases within the *Reviews* covariate from the difference between the mean (444193.87) and the 5% trimmed mean (80483.37). The median continues to show the extreme cases because based on the median the average application has 2094 reviews whilst with the 5% trimmed mean the average application has 80483 reviews.

As expected, the range is very large because there are applications that get no reviews and very popular applications that get millions of reviews from people all around the world. However, the skewness indicates that there are more applications that get few reviews than ones that get many reviews since the skewness value (16.449) is quite high: the distribution is shifted to the left. The Kurtosis (341.029) further amplifies the presence of outliers because it is very high which indicates that most of the values are found on the tails of the distribution curve.

Figure 2 shows the boxplot for the *Reviews* covariate. The box itself is not visible due to the scale of the plot however one can easily see the large number of outliers present by the amount of points outside the whiskers.

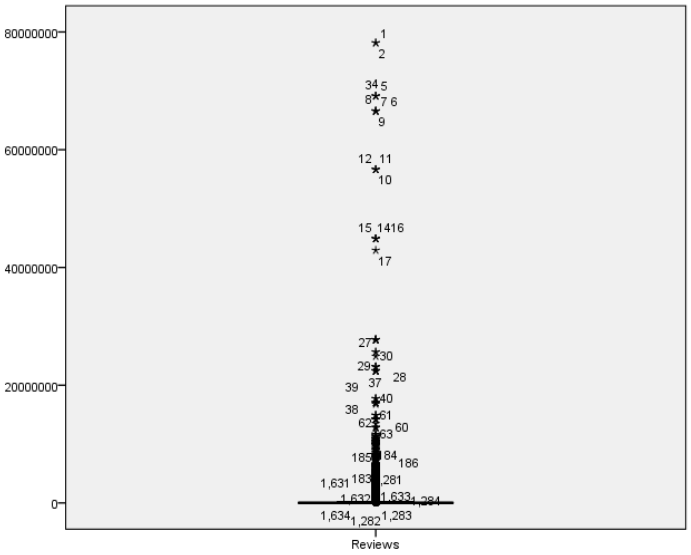


Figure 3 Boxplot for Reviews

Again, due to the large range of the *Reviews* covariate the scale of the histogram makes it very difficult to obtain an ideal curve. From figure 4 it is can be seen that there is positive skewness and hence there are more apps that receive fewer reviews.

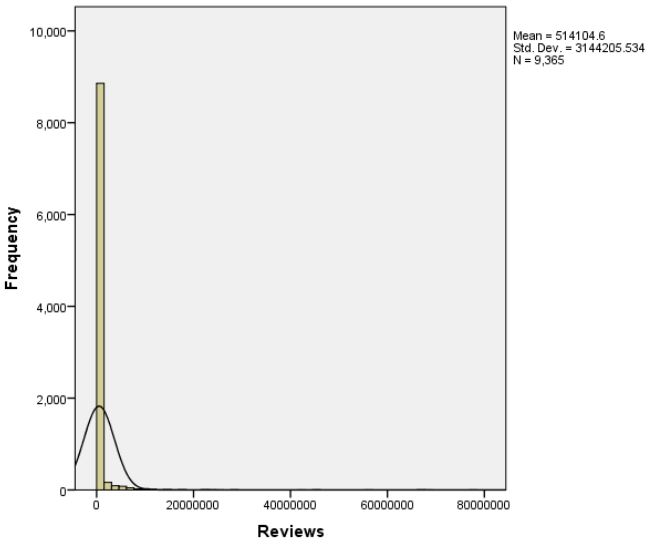


Figure 4 Histogram for Reviews

3.1.3 Size

Size	Mean		18147.60	213.003
	95% Confidence Interval for Mean	Lower Bound	17730.08	
		Upper Bound	18565.13	
	5% Trimmed Mean		15452.56	
	Median		9200.00	
	Variance		491768450.859	
	Std. Deviation		22175.853	
	Minimum		0	
	Maximum		100000	
	Range		100000	
	Interquartile Range		23400	
	Skewness		1.704	.024
	Kurtosis		2.508	.047

Table 3 Descriptives for Size

The mean size for the applications in this dataset is around 18Mb[1]. When the extreme cases are trimmed the average size drops to around 15Mb (5% trimmed mean) which shows that there are more outliers with larger sizes. The median is approximately 9.2Mb which is a better representation of the expected size of an application due to the great number of outliers in this dataset which comes from its relatively large size.

Application sizes vary from less than 1Mb to around 100Mb based on the range. The skewness shows that the distribution of the sizes is shifted to the left since it is positive (1.704) which implies that there are more applications with a small size than large applications. The kurtosis lies at 2.508 which indicates that there is some distribution of sizes along the tails, but it is not too great.

The Boxplot for the size variable shows the positive skewness of this distribution since the median is shifted towards the 25th percentile. It also confirms that there are more outliers with larger sizes.

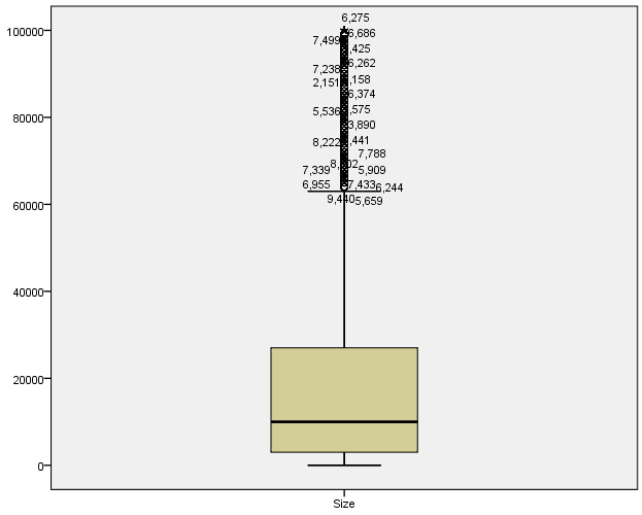


Figure 5 Boxplot for Size

[1] Results are in Kb

Figure 6 shows the histogram for the *Size* covariate. Just like the other histograms before, it does not follow a normal distribution. In this case it has more of an exponential decay because the size of apps rapidly decreases (i.e. there are much fewer large apps). The curve complements the skewness value obtained previously.

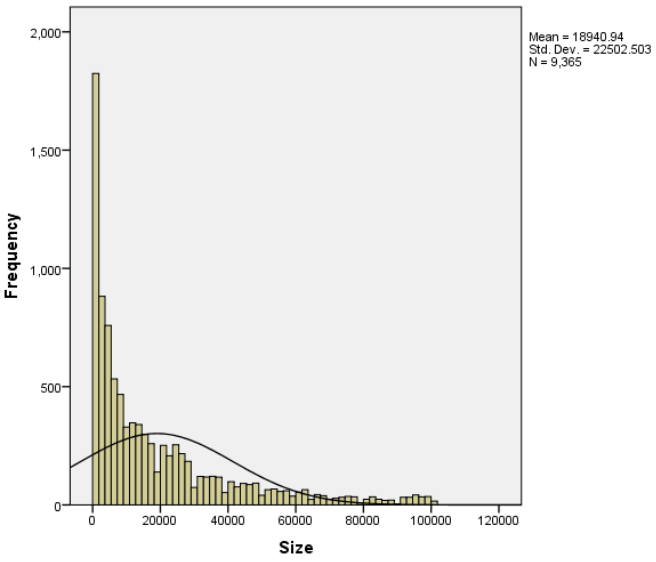


Figure 6 Histogram for Size

3.2 Frequencies

3.2.1 Installs

Installs

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<100K	4720	43.5	43.5	43.5
	100K - 1M	1708	15.8	15.8	59.3
	1M - 10M	2331	21.5	21.5	80.8
	10M - 100M	1541	14.2	14.2	95.0
	100M+	539	5.0	5.0	100.0
	Total	10839	100.0	100.0	

Figure 7 Frequencies for Installs

As can be observed in figure 7, to some extent, the frequency and the number of installs is inversely proportional, i.e. in the data set there are more applications that have a small number of downloads. This is somewhat contradicting, intuitively one would think that if an application has more downloads it would have a higher chance of being selected in this dataset, but the chosen applications do not depend on the installs as they were chosen randomly. This result shows that there are much more applications that have a lower number of installs, while those with a higher number of installs appear more scarcely. These findings can be seen visually in the bar chart below.

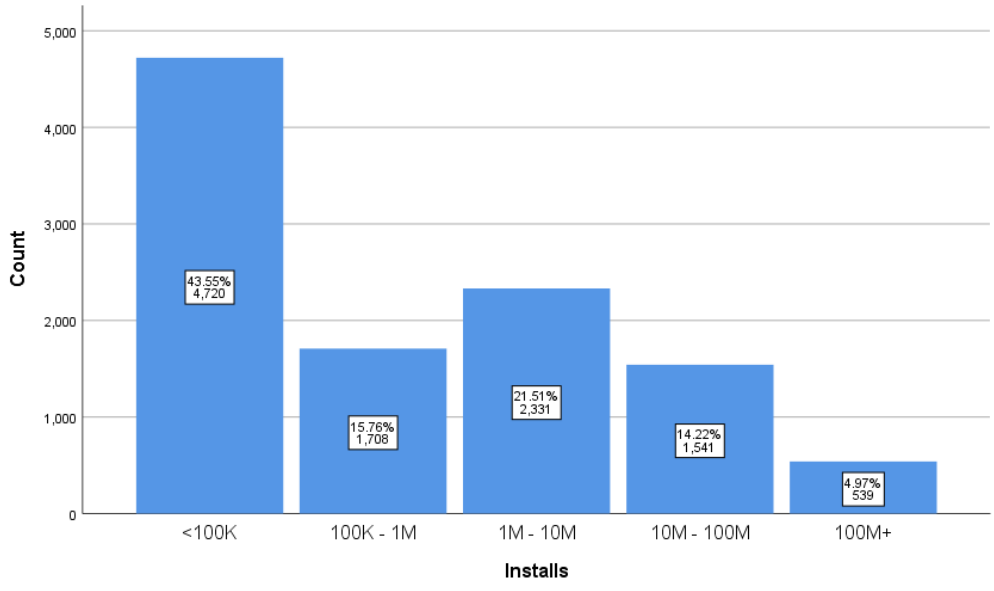


Figure 8 Bar chat for Installs factor

3.2.2 Type

Type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Free	10039	92.6	92.6	92.6
	Paid	800	7.4	7.4	100.0
	Total	10839	100.0	100.0	

Figure 9 Frequencies for Type

In figure 9 the frequency of the number of free and paid applications is observed. As expected, free applications have a much higher frequency than paid ones because developers know that more people will use a free app with advertisements rather than pay for one. They overbalance them with a tremendous 92.6 percent compared with the 7.4 percent for paid applications. The pie chart below gives a visual representation of these frequencies.

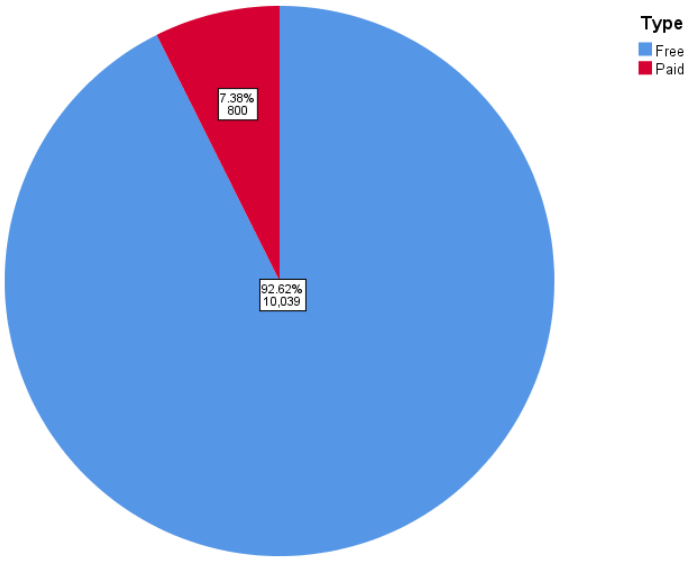


Figure 10 Pie chart for Type factor

3.2.3 Content_Rating

Content_Rating

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unrated	2	.0	.0	.0
	Everyone	8714	80.4	80.4	80.4
	10+	413	3.8	3.8	84.2
	Teen	1208	11.1	11.1	95.4
	17+	502	4.6	4.6	100.0
	Total	10839	100.0	100.0	

Figure 11 Frequencies for Content_Rating

Figure 11 contains the frequencies for the *Content_Rating* factor. 'Everyone' has the highest frequency with 'Teen' being a distant second. This means that applications rated for 'Everyone' are the most common within the Google Play Store based on this dataset. Similarly, applications rated 'Teen' are the second most common with '10+' and '17+' being even more uncommon. However, this is still a lot compared to the mere two applications which are not rated. This data can also be seen in the bar chart below.

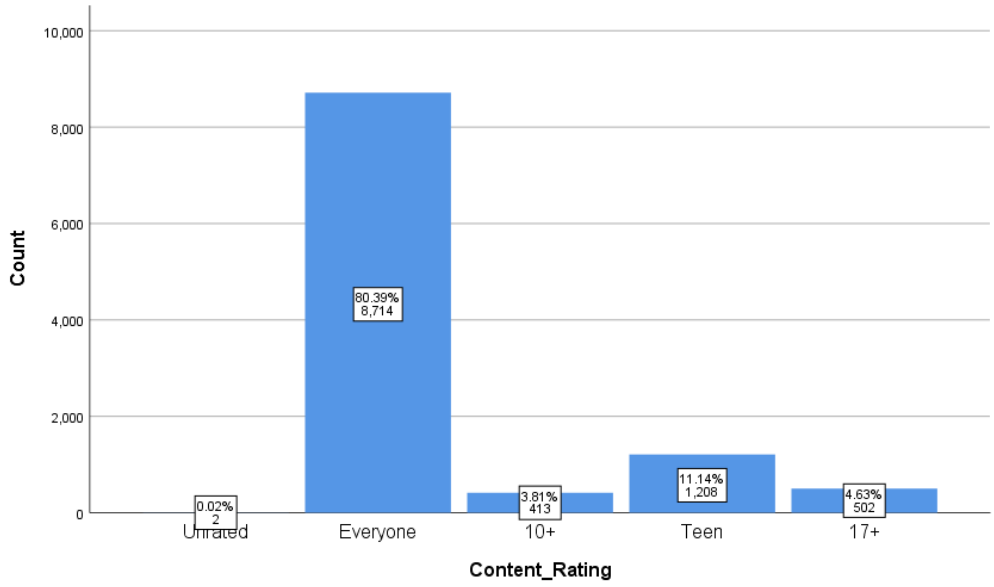


Figure 12 Bar chart for Content_Rating factor

3.3 Boxplots

This section will explain the obtained boxplots. Only boxplots that have to do with the dependent variable (Rating) are discussed because they are the most relevant.

3.3.1 Rating vs Installs

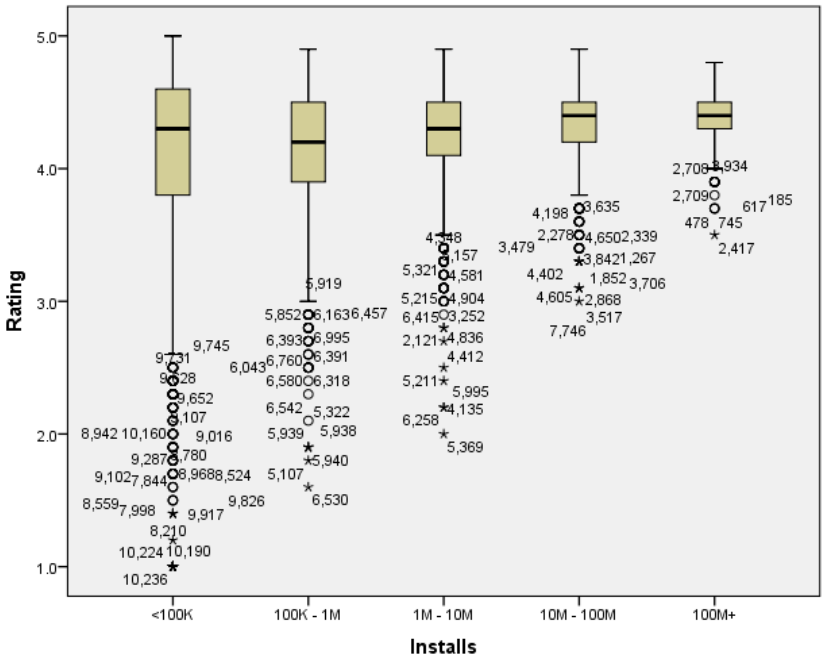


Figure 13 Rating vs Installs Boxplots

Figure 13 shows the boxplot for *Rating* for each category in the *Installs* factor. The median rating always lies in the same general area for all categories however apps that have between 100k and 1 million installs fall slightly behind. There are more outliers for applications that have fewer installs and for the larger number of installs, the range in ratings decreases. This shows that applications with a higher number of installs have a higher average rating. The interquartile range also decreases as the number of installs increases showing that ratings are more condensed and less spread out.

3.3.2 Rating vs Size

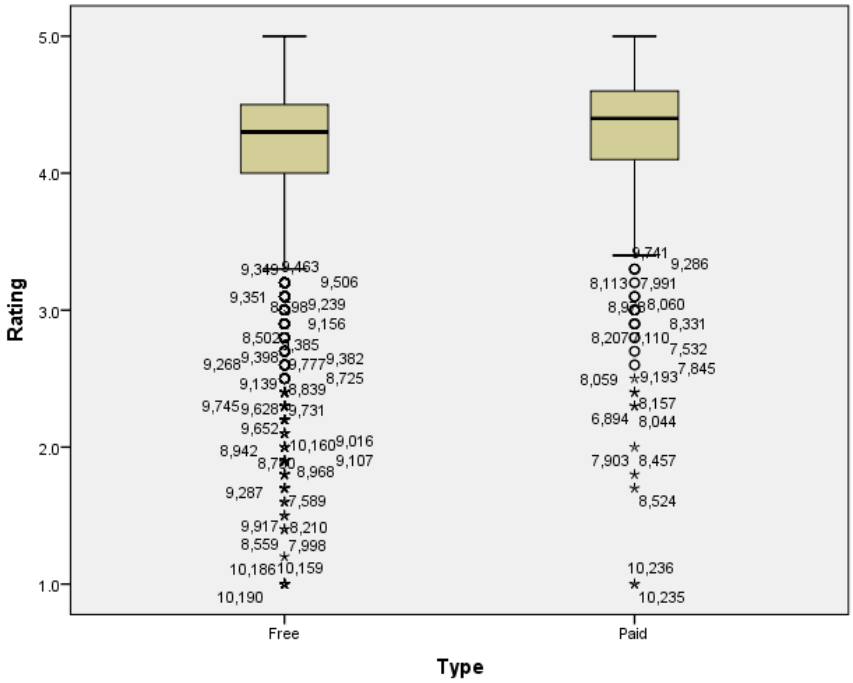


Figure 14 Rating vs Type Boxplots

Figure 14 shows the boxplot for *Rating* for each both categories of the *Type* factor. There is no significant difference between the two categories. The Paid boxplot has a slightly higher median and less outliers that the Free boxplot but not by a large margin. This was not expected because it makes more sense to expect apps that are paid for to have a higher rating than free applications. However, the number of outliers supports this expectation because the Free boxplot has much more outliers on the lower end of the rating scale.

3.3.3 Rating vs Content_Rating

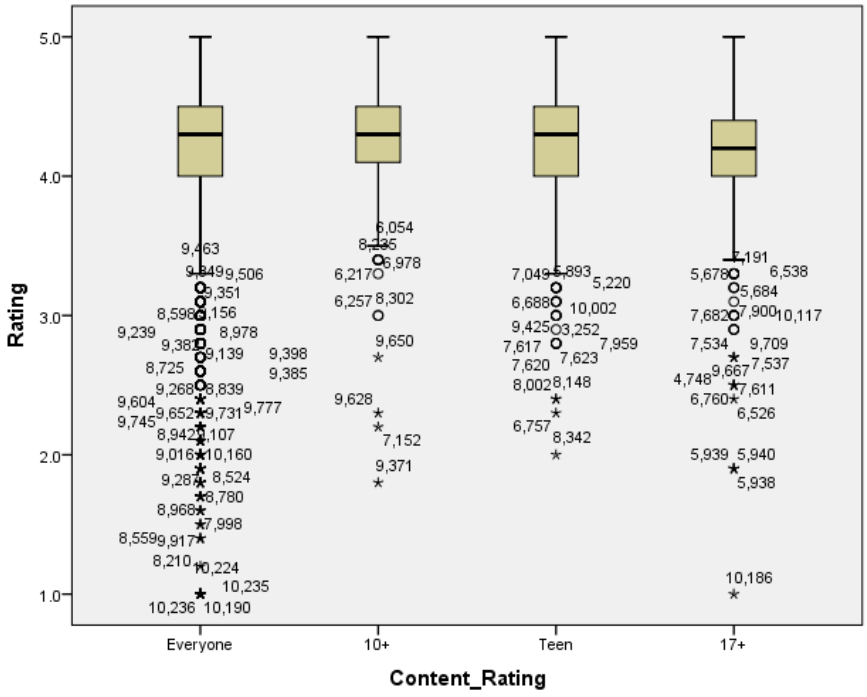


Figure 15 Rating vs Content_Rating Boxplots

Figure 15 shows the boxplot for *Rating* for each both categories of the *Content_Rating* factor. The medians, interquartile ranges and ranges are quite similar for all categories of the *Content_Rating* factor. The Everyone category has many more outliers than the rest, but this is due to the fact that the majority of the applications reside in this category as can be seen in figure 12.

3.4 Scatterplots

The following section will describe how scatterplots were used to visually inspect the data, to see if any relationships between the dependent variable being observed (i.e. *Rating*) and the other covariates (i.e. *Reviews* and *Size*) exist.

3.4.1 Rating vs Reviews

In this case, the *Rating* variable (on the y-axis) is plotted against the *Reviews* variable (on the x-axis), along with a line of best fit and the output is given as follows:

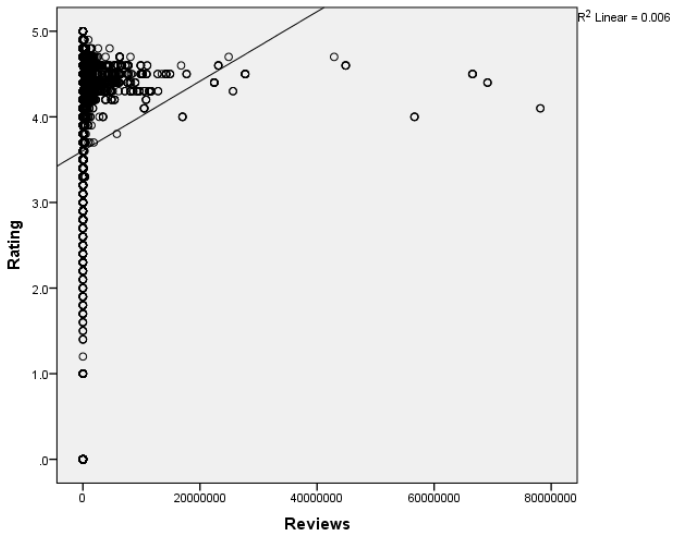


Figure 16: Scatterplot of Ratings against number of reviews.

The output seen in the figure above suggests that a linear regression model might not be a good fit for the data, since many data points seem to deviate from the line of best fit. In fact, the scatterplot suggests that a quadratic model would be more adequate for the data in question. However, this has yet to be determined when performing regression modelling on the data (see Section 5). It is also of note that data points which have a larger number of reviews seem to be quite sparse when compared to those having much less reviews, which may suggest that they are outliers. Moreover, a lot of variability can be observed in the data when the app has no (or few) reviews. This is because when an app has very few reviews, each one has a lot more weight on the final rating of the app. Hence, a single bad or good review can cause the rating of the app to spike or plummet immediately. Nevertheless, as the number of reviews increases, the range of ratings that the app can have can be seen to decrease, usually lying somewhere in the range between 4 and 5.

3.4.2 Rating vs Size

In this case, the *Rating* response variable (on the y-axis) is plotted against the *Size* variable (on the x-axis) along with a line of best fit, to check for any relationships between the two variables. The output is given as follows:

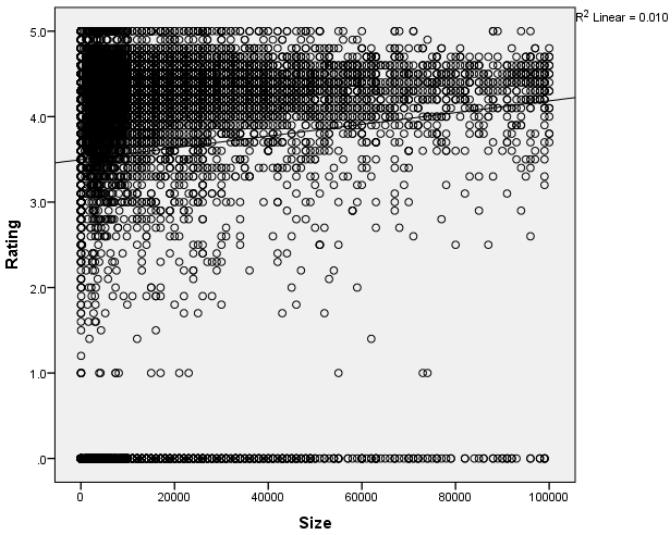


Figure 17 Scatterplot of Rating against Size (in megabytes).

As in the previous case, the above scatterplot also suggests that a linear regression model would not fit the data well, given that more points than before seem to deviate from the line of best fit. Moreover, just as before, this scatterplot also seems to show a quadratic relationship between the variables. However, as one might expect, the correlation between the two variables seems to be far less strong, which is made obvious by the fact that the data points are much more scattered when compared to the data points in the previous scatterplot. Yet, it can still be observed that as the file size of the application increases, the ratings seem to reduce down to a smaller range around the larger ratings, similarly to the previous scatterplot. In addition, it can also be seen that there is a large variability in size for applications with a low rating. Though there does not seem a very clear reason why this would be the case, one possible cause would be lack of correlation between the variables due to reasons such as inflated file sizes, or limited storage capacity on devices making it impossible for users to download the app etc.

3.5 Clustered Bar Charts

This section will explain the information gathered from the clustered bar charts that were created between the *Installs* and the other two factors.

3.5.1 Installs vs Type

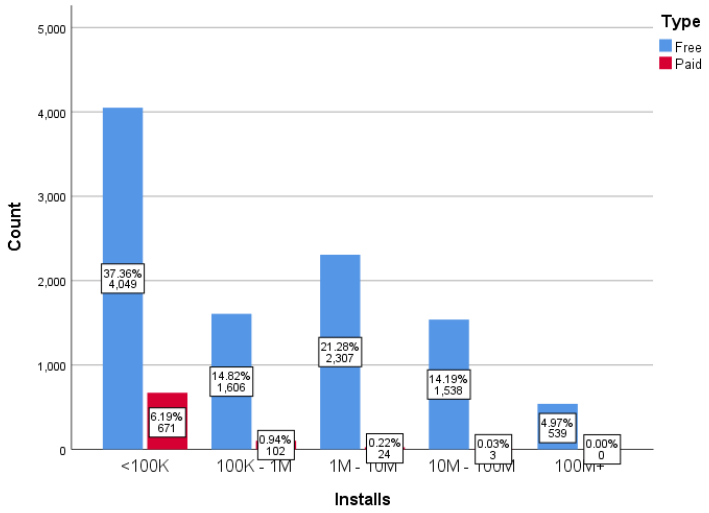


Figure 18 Installs vs Type Clustered Bar Chart

In figure 16 free and paid apps are compared for different amounts of installs. It can be observed that there are much more installs for free applications and the highest count of installed apps which are paid for lies in the <100K range. This shows that not many people but applications on the Google Play Store.

3.5.2 Installs vs Content_Rating

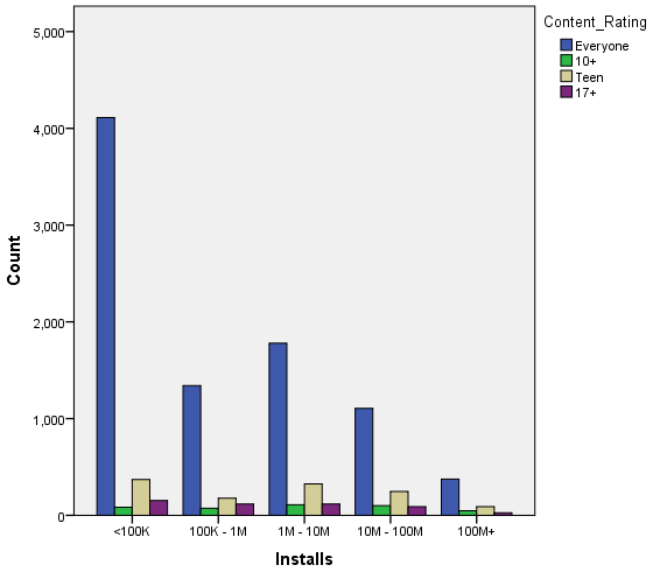


Figure 19 Installs vs Content_Rating Clustered Bar Chart

Figure 17 shows the count of the ratings for different quantities of installs l. The rating for 'Everyone' decreases for larger amounts of installs (except for the 100K-1M). On the contrary the other ratings, decrease at a much smaller rate. From this chart it is evident that applications with a content rating 'Everyone' have a larger share of the Google Play Store.

4 Hypothesis Testing

In this section, parametric/non-parametric tests are used to see if any of the fixed factor variables (i.e. *Installs*, *Type*, and *Content_Rating*) have any significant impact on the mean (or median) of the variable of interest (i.e. the *Rating* variable).

4.1 Rating vs. Installs

For the first test, the effect of the *Installs* variable on the mean (or median) of the *Rating* variable is tested. Since the *Installs* variable has five categories, all of which are independent from each other, only the *One-Way ANOVA* or *Kruskal Wallis* test could be used for this purpose; to determine which one to use, the data is tested to see if it respects the assumptions of the *One-Way ANOVA* test, which include the following:

1. For each group, the response variable must be normally distributed.
2. The variances of all groups must be equal.

The first assumption is tested using the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests, both of which test the following hypotheses:

> **H_0:** *Rating* follows a normal distribution for the given group of the *Installs variable*
>
> **H_1:** *Rating* does not follow a normal distribution for the given group of the *Installs variable*

The outputs of the tests were computed in SPSS and can be seen below:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Installs	Statistic	df	Sig.	Statistic	df	Sig.
Rating	<100K	.231	4718	.000	.764	4718	.000
	100K - 1M	.167	1708	.000	.725	1708	.000
	1M - 10M	.141	2331	.000	.863	2331	.000
	10M - 100M	.128	1541	.000	.938	1541	.000
	100M+	.170	539	.000	.920	539	.000

a. Lilliefors Significance Correction

Figure 20 Outputs of Normality tests for each different 'Installs' category.

For every category, the p-value of both tests is zero, which is far less than the level of significance (which is 0.05). Hence, the null-hypothesis is rejected for all cases, and the response variable does not follow a normal distribution for any category. This can also be

confirmed by looking at the *Q-Q plot charts*, which shows that the data points for each category deviate a lot from the expected normal distribution:

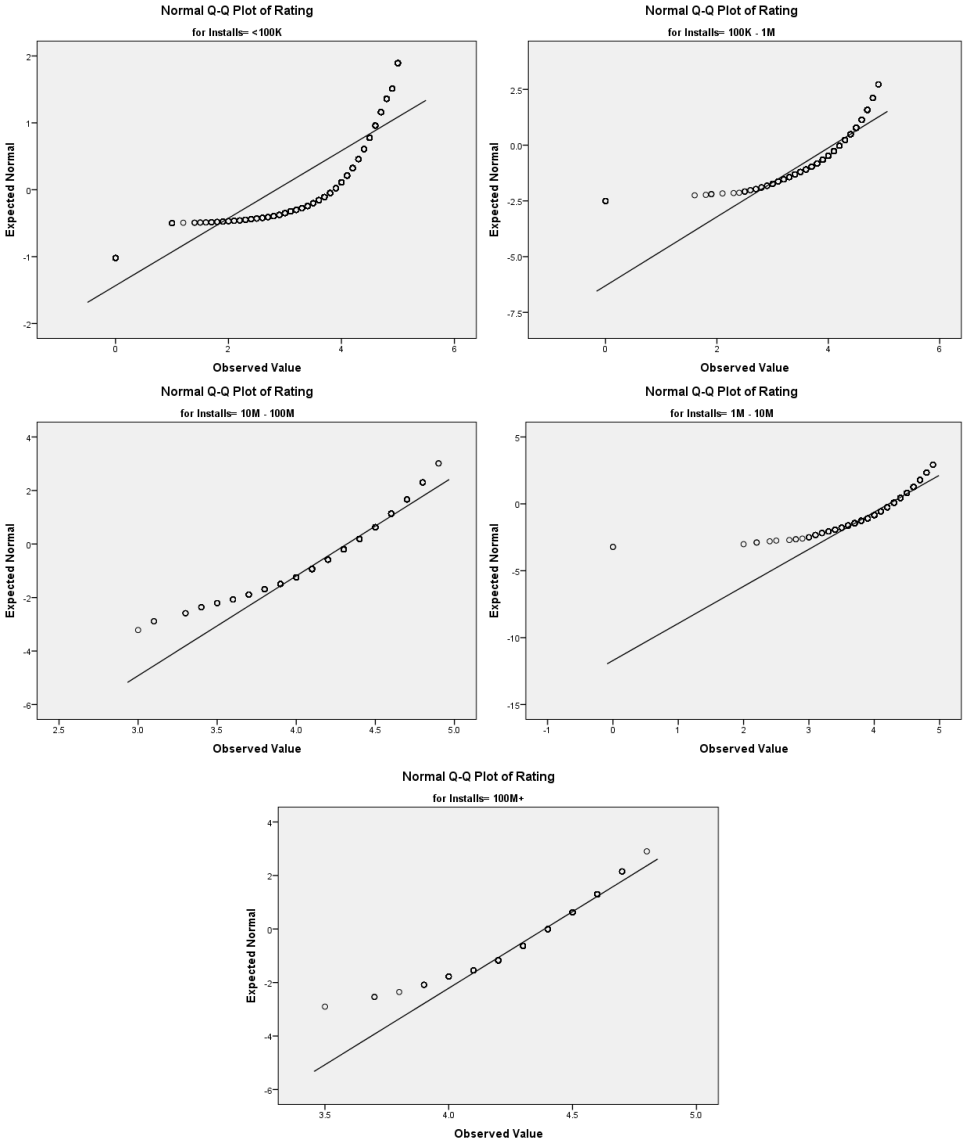


Figure 21 Q-Q plots of all the 'Install' categories

Given that one of the assumptions of the *One-Way ANOVA* test is not upheld, the non-parametric version of the test (i.e. Kruskal Wallis) must be used to check the influence of the *Installs* variable on the *Rating*. In this case, the following hypotheses are tested:

H_0: The median values of *Rating* are the same for all categories of *Installs*

H_1: The median values of *Rating* are different for the categories of *Installs*

The test is performed in SPSS and it gives the following outputs:

Test Statistics[a,b]

	Rating
Chi-Square	1050.744
df	4
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: Installs

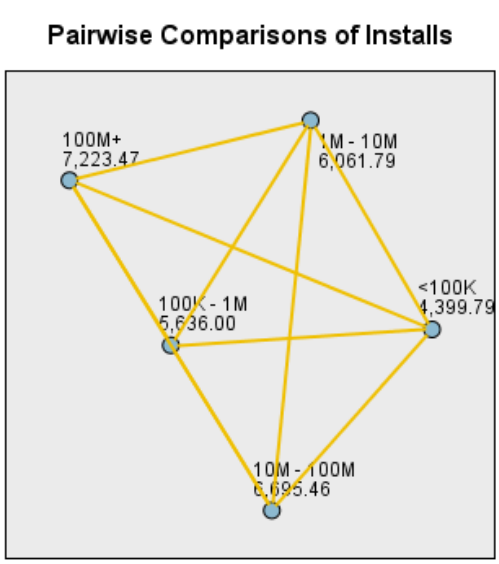
Figure 22 Outputs of the Kruskal Wallis test

As can be seen above, the p-value of the test, which is zero, is much less than the level of significance (0.05). Hence, the null-hypothesis is rejected, and the median *Rating* of the different groups are not the same. Now, since the medians of the groups are different from each other, a Post-Hoc analysis is conducted to verify where these differences lie. Each pairwise comparison consists of a *Mann-Whitney* test which tests the following hypotheses:

H_0: The median ratings of the *Rating* of the two groups are the same.

H_1: The median ratings of the *Rating* of the two groups are different.

The test is conducted in SPSS and the following output is obtained:



Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
<100K-100K - 1M	-1,236.206	88.028	-14.043	.000	.000
<100K-1M - 10M	-1,661.995	78.920	-21.059	.000	.000
<100K-10M - 100M	-2,295.663	91.463	-25.099	.000	.000
<100K-100M+	-2,823.677	141.732	-19.923	.000	.000
100K - 1M-1M - 10M	-425.789	99.287	-4.288	.000	.000
100K - 1M-10M - 100M	-1,059.457	109.522	-9.673	.000	.000
100K - 1M-100M+	-1,587.471	154.005	-10.308	.000	.000
1M - 10M-10M - 100M	-633.668	102.345	-6.191	.000	.000
1M - 10M-100M+	-1,161.681	148.987	-7.797	.000	.000
10M - 100M-100M+	-528.014	155.994	-3.385	.001	.007

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 23 Post-hoc pairwise comparisons of each 'Install' category

Each of the tests conducted above can be seen to have a p-value far below the level of significance (0.05). Hence, for each pairwise comparison, the null-hypothesis is rejected, and each interval of the *Installs* variable has a distinct median value of the *Rating* variable.

Next, to confirm the result found above, the sample mean and median *Rating* of each group were computed:

Report

Rating

Installs	Mean	N	Std. Deviation	Median
<100K	2.844	4718	1.9837	3.900
100K - 1M	4.080	1708	.6473	4.200
1M - 10M	4.224	2331	.3608	4.300
10M - 100M	4.321	1541	.2682	4.400
100M+	4.387	539	.1748	4.400
Total	3.622	10837	1.5140	4.200

Figure 24 Sample means and medians of each category in 'Installs'

The output above confirms the results found using the Post-hoc analysis, suggesting that as the number of installs increases, the rating of the application also increases.

4.2 Rating vs Content Rating

In the second test, the effect of the *Content Rating* variable on the mean (or median) of the *Rating* variable is tested. Similarly to the previous test, since the *Content Rating* variable has four independent categories, either the *One-Way ANOVA* or *Kruskal Wallis* test should be used; the choice between the two is once again made by checking if the data fits the assumptions necessary to use the *One-Way ANOVA* test.

To check if the *Rating* variable is normal for each group in *Content Rating*, the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests are used to verify the following hypothesis:

H_0: *Rating* follows a normal distribution for the given group of the *Installs variable*

H_1: *Rating* does not follow a normal distribution for the given group of the *Installs variable*

The output of the tests given by SPSS can be seen below:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Content Rating	Statistic	df	Sig.	Statistic	df	Sig.
Rating	Everyone	.285	8714	.000			
	10+	.285	413	.000	.510	413	.000
	Teen	.311	1208	.000	.592	1208	.000
	17+	.276	502	.000	.628	502	.000

a. Lilliefors Significance Correction

Figure 25 Output of normality tests for each group of the 'Content Rating' variable

The output above shows that none of the categories have a p-value above the level of significance (0.05). Therefore, the null-hypothesis is rejected, and the *Rating* variable does not follow a normal distribution for any of the *Content Rating* variable's groups. To confirm this result, the *Q-Q plot* charts for the different categories were also created:

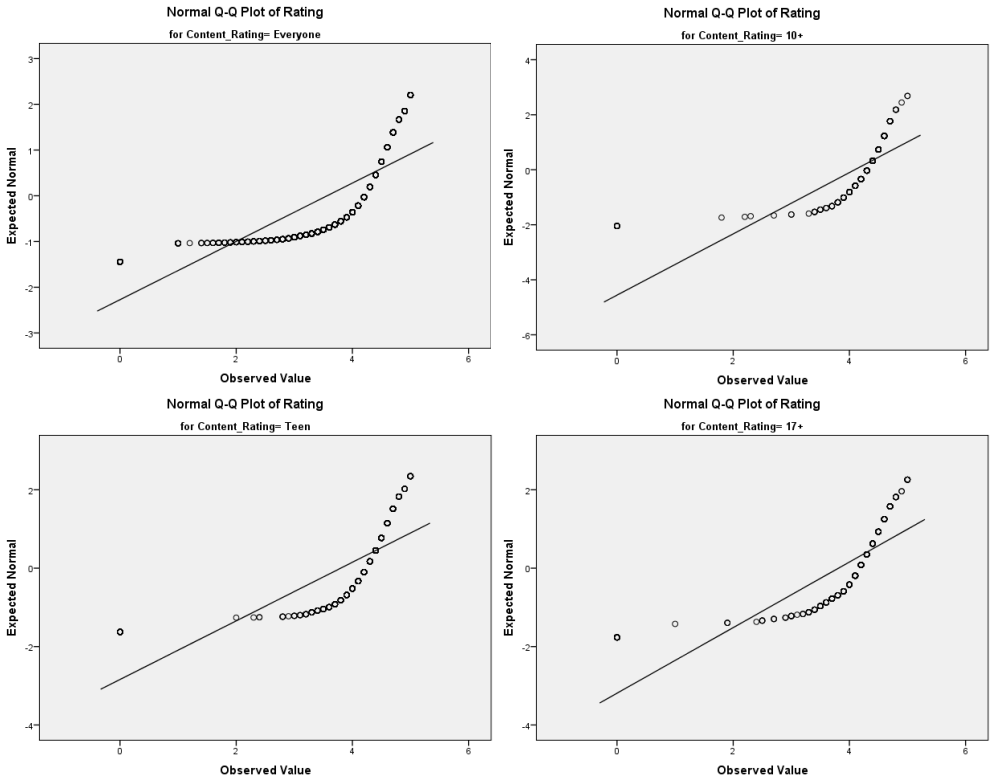


Figure 26 Q-Q plot charts for the different categories of 'Content Rating'

In the images above, the data points for all groups can be seen to deviate from the expected normal distribution, hence the result given by the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests is confirmed.

Hence, since one of the assumptions necessary for the *One-Way ANOVA* test is not upheld, the *Kruskal Wallis test* must be used to test if there is any significant difference between the medians of the different dependent variables. The *Kruskal Wallis test* verifies the following hypotheses:

H_0: The median values of *Rating* are the same for all categories of *Content Rating*

H_1: The median values of *Rating* are different for the categories of *Content Rating*

The output for the test can be seen in the image below:

Test Statistics[a,b]

	Rating
Chi-Square	30.236
df	3
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: Content_Rating

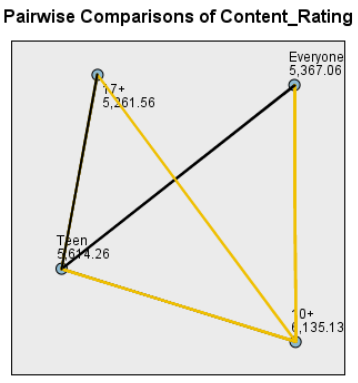
Figure 27 Outputs of Kruskal Wallis test

Since the test gives a p-value of zero, at a level of significance of 0.05 the null-hypothesis is rejected. Thus, the median *Rating* is different for the categories of the *Content Rating* variable. To check where the discrepancies in *Rating* lie specifically, a Post-hoc analysis is performed, and the medians of each group were compared in a pairwise manner using the *Mann-Whitney* test. For each pair of groups, the following hypothesis were tested:

H_0: The median *Rating* of the two groups are the same.

H_1: The median *Rating* of the two groups are different.

When the tests are performed in SPSS, the following output is given:



Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
17+-Everyone	105.500	143.081	.737	.461	1.000
17+-Teen	352.705	165.533	2.131	.033	.199
17+-10+	873.572	207.089	4.218	.000	.000
Everyone-Teen	-247.205	95.704	-2.583	.010	.059
Everyone-10+	-768.072	156.983	-4.893	.000	.000
Teen-10+	520.867	177.687	2.931	.003	.020

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.
Significance values have been adjusted by the Bonferroni correction for

Figure 28 Output of the Post-Hoc analysis for each group in 'Content Rating'

The output above shows that at a 0.05 level of significance, only the tests considering the difference between the '*10+'* category and some other category reject the null-hypothesis. Hence, only the median of the '*10+'* category seems to have any significant difference from the other variables. On the other hand, the categories '*Everyone*', '*Teen'* and '*17+*' all seem to have the same median.

To confirm these findings, the sample means and medians for all the independent groups have been calculated in SPSS, and can be seen in the image below:

Report

Rating

Content_Rating	Mean	N	Std. Deviation	Median
Everyone	3.565	8714	1.5691	4.200
10+	4.092	413	.8979	4.300
Teen	3.799	1208	1.3378	4.200
17+	3.812	502	1.1952	4.200
Total	3.622	10837	1.5140	4.200

Figure 29 Means and medians of all the categories in 'Content Rating'

The output shown in the above table seems to confirm the results of the Post-Hoc analysis; for the means and medians of the '*Everyone*', '*Teen'* and '*17+*' categories all appear to be similar to each other. Moreover, the mean and median of the '10+' category is in fact higher than the other categories, further confirming the results found previously. In addition, this output also suggests that mobile applications targeting to this demographic may achieve a higher rating than if it were to target any other demographic of phone users.

4.3 Rating vs. Type

In the third and final test, the data is verified to see if there is any difference in the mean (or median) *Rating* of paid and free apps, where 'paid' and 'free' are categories in the fixed factor variable called *Type*. Given that *Type* has only two independent categories, this could be done using either the *Independent Samples T-test* or the *Mann-Whitney* test; to determine which one should be used, the data is verified to see if it upheld the assumptions of the *Independent Samples T-test*, which include the following:

1. Both samples must come from normal populations.
2. Both populations must have equal variances.

The first assumption is verified by performing the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests on the *Rating* variable for both 'paid' and 'free' apps. The following hypotheses are tested:

H_0: *Rating* follows a normal distribution for the given type of app (free or paid)

H_1: *Rating* does not follow a normal distribution for the given type of app (free or paid)

The results are computed in SPSS and can be seen in the following table:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Type	Statistic	df	Sig.	Statistic	df	Sig.
Rating	Free	.290	10037	.000			
	Paid	.283	800	.000	.681	800	.000

a. Lilliefors Significance Correction

Figure 30 Outputs of Normality tests for 'paid' and 'free' apps

In this case it can be observed that both sets of p-values are significantly less than the level of significance (0.05). Hence, the null-hypothesis is rejected, and the *Ratings* variable does not follow a normal distribution for 'free' and 'paid' apps. To confirm this result, the *Q-Q plot charts* of the two categories are also created:

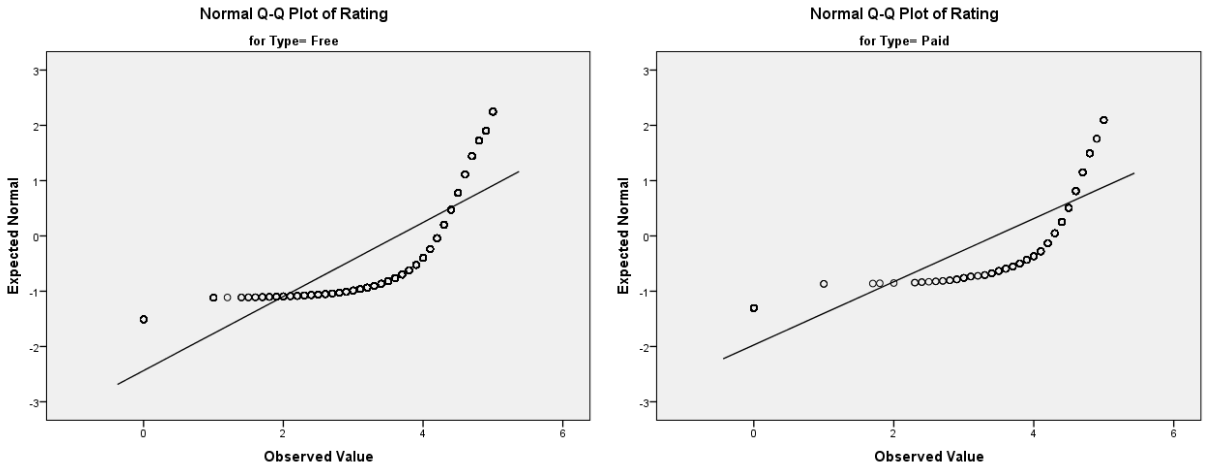


Figure 31 Q-Q plot charts for free and paid apps.

As can be seen in the figure above, the data points deviate significantly from the expected normal distribution, confirming the results found by the *Kolmogorov-Smirnov* and *Shapiro-Wilk* tests.

Therefore, since one of the assumptions of the *Independent Samples T-test* is not upheld, to test if there is any difference between the medians of the two groups, the *Mann-Whitney* test must be used. The *Mann-Whitney* test checks for the following hypotheses:

H_0: The median *Rating* of free and paid apps are the same.

H_1: The median *Rating* of free and paid apps are different.

The results of the test are computed in SPSS, and are given as follows:

Test Statistics[a]

	Rating
Mann-Whitney U	3837602.000
Wilcoxon W	54213305.00
Z	-2.088
Asymp. Sig. (2-tailed)	.037

a. Grouping Variable: Type

Figure 32 Outputs for the Mann-Whitney test

Since the p-value (0.037) of the test is less than the level of significance (0.05), the null-hypothesis is rejected, and the median *Rating* of free and paid applications is not the same.

To verify the results of the hypothesis test, the sample means and medians of free and paid apps were calculated in SPSS:

Report

Rating

Type	Mean	N	Std. Deviation	Median
Free	3.636	10037	1.4928	4.200
Paid	3.451	800	1.7497	4.300
Total	3.622	10837	1.5140	4.200

Figure 33 Table containing sample mean and median of paid and free apps.

Comparing the two medians, it is apparent that they are not the same, and that the median of paid apps is larger, confirming the result found by the *Mann-Whitney* test. However, when comparing the sample means, the mean of free apps is the highest, which contradicts the results of the test as well as the sample medians. As such, the output does not make it clear if a free or paid application is preferable to obtain a higher rating.

4.4 Correlations

This section deals with correlations between the dependent variable (Rating) and the other covariate variables. Pearson's correlation was not used because from the scatter plots in Section 3.4 it is evident that there is not a linear relationship between these variables. Therefore, the Spearman correlation was used. For both of the following cases the following hypothesis tests were used:

H_0: Variables are independent: $\rho = 0$

H_1: A relationship exists between the variables and can be modelled by some monotonic function: $\rho \neq 0$

4.4.1 Rating vs Reviews

Correlations

			Rating	Reviews
Spearman's rho	Rating	Correlation Coefficient	1.000	.156**
		Sig. (2-tailed)	.	.000
		N	9365	9365
	Reviews	Correlation Coefficient	.156**	1.000
		Sig. (2-tailed)	.000	.
		N	9365	9365

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 34 Rating vs Reviews Correlation

Figure 33 shows that the Spearman's correlation coefficient between *Rating* and *Reviews* is 0.156 which indicates that there is a weak positive monotonic relationship between the two variables. This is further confirmed by the p-value which is less than the level of significance. Thus, **H_0** is rejected, and a relationship exists between the two variables.

4.4.2 Rating vs Size

Correlations

			Rating	Size
Spearman's rho	Rating	Correlation Coefficient	1.000	.063**
		Sig. (2-tailed)	.	.000
		N	9365	9365
	Size	Correlation Coefficient	.063**	1.000
		Sig. (2-tailed)	.000	.
		N	9365	9365

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 35 Rating vs Size Correlation

Figure 34 shows that the correlation coefficient between *Rating* and *Size* is 0.063 which indicates that there is a weak positive monotonic relationship between the two variables and is further backed up by the p-value, which is less than the level of significance (0.05). Hence, **H_0** is rejected and the relationship between the variables is significant.

5 Modelling

In this section, the data set is modelled using regression techniques; the *Rating* variable is chosen as the dependent variable, while the rest of the variables (i.e. *Reviews, Size, Installs, Type, and Content Rating*) are used as the predictors. Given that some of the predictors to be used are categorical variables, that the dependent variable is quantitative, and that all observations in the dataset are independent, a *General Linear Model* (*GLM*) is used, specifically an *ANCOVA* model. It is of note that, even though the scatterplots of the covariates in Section 3.4 did not seem show a linear relationship, the data is modelled using a *GLM* to test if such a model would be a good fit for the data set in question. Moreover, even though the cofactors *Reviews* and *Size* have been shown to have little correlation to the *Rating* variable (in Section 4.4), they are initially included to check if they are collinear with any of the other predictors, in which case, they would be removed from the model.

Before the data could be fitted, dummy variables must be created for the categorical variables.

5.1 Collinearity Diagnostics

5.1.1 Spearman Correlation

First, the check for collinearity between the predictors, the correlation coefficients of the variables are found. Pearson could not be used for this since one of the assumptions needed is not fulfilled; as mentioned before, the relationships between some of the covariates do not seem be linear. However, the scatterplot relationships of the variable did seem to show monotonicity, and hence *Spearman* seems to be suitable for this purpose. *Spearman correlation* tests the following hypotheses for each pair of variables:

> **H_0:** Variables are independent: $\rho = 0$
>
> **H_1:** A relationship exists between the variables and can be modelled by some monotonic function: $\rho \neq 0$

When the correlation coefficients are computed in SPSS the following outputs are obtained:

Correlations

			Rating	Reviews	Size	Installs=<100 K	Installs=100K - 1M	Installs=1M - 10M	Installs=10M - 100M	Type=Free	Content_Rating=Everyone	Content_Rating=10+	Content_Rating=Teen
Spearman's rho	Rating	Correlation Coefficient	1.000	.430**	.121**	-.287**	.030**	.108**	.167**	-.020*	-.034**	.046**	.022*
		Sig. (2-tailed)	.	.000	.000	.000	.002	.000	.000	.037	.000	.000	.021
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Reviews	Correlation Coefficient	.430**	1.000	.211**	-.842**	.023*	.364**	.514**	.172**	-.186**	.123**	.125**
		Sig. (2-tailed)	.000	.	.000	.000	.015	.000	.000	.000	.000	.000	.000
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Size	Correlation Coefficient	.121**	.211**	1.000	-.209**	.024*	.113**	.089**	.007	-.011	.012	-.025**
		Sig. (2-tailed)	.000	.000	.	.000	.013	.000	.000	.459	.269	.210	.009
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Installs=<100K	Correlation Coefficient	-.287**	-.842**	-.209**	1.000	-.380**	-.460**	-.358**	-.230**	.149**	-.094**	-.092**
		Sig. (2-tailed)	.000	.000	.000	.	.000	.000	.000	.000	.000	.000	.000
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Installs=100K - 1M	Correlation Coefficient	.030**	.023*	.024*	-.380**	1.000	-.226**	-.176**	.023*	-.021*	.010	-.011
		Sig. (2-tailed)	.002	.015	.013	.000	.	.000	.000	.015	.031	.276	.262
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Installs=1M - 10M	Correlation Coefficient	.108**	.364**	.113**	-.460**	-.226**	1.000	-.213**	.127**	-.053**	.024*	.047**
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.	.000	.000	.000	.014	.000
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Installs=10M - 100M	Correlation Coefficient	.167**	.514**	.089**	-.358**	-.176**	-.213**	1.000	.112**	-.089**	.057**	.062**
		Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.	.000	.000	.000	.000
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Type=Free	Correlation Coefficient	-.020*	.172**	.007	-.230**	.023*	.127**	.112**	1.000	-.046**	-.005	.042**
		Sig. (2-tailed)	.037	.000	.459	.000	.015	.000	.000	.	.000	.630	.000
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Content_Rating=Everyone	Correlation Coefficient	-.034**	-.186**	-.011	.149**	-.021*	-.053**	-.089**	-.046**	1.000	-.403**	-.718**
		Sig. (2-tailed)	.000	.000	.269	.000	.031	.000	.000	.000	.	.000	.000
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Content_Rating=10+	Correlation Coefficient	.046**	.123**	.012	-.094**	.010	.024*	.057**	-.005	-.403**	1.000	-.071**
		Sig. (2-tailed)	.000	.000	.210	.000	.276	.014	.000	.630	.000	.	.000
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837
	Content_Rating=Teen	Correlation Coefficient	.022*	.125**	-.025**	-.092**	-.011	.047**	.062**	.042**	-.718**	-.071**	1.000
		Sig. (2-tailed)	.021	.000	.009	.000	.262	.000	.000	.000	.000	.000	.
		N	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837	10837

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 36 Table of Spearman Corellation coeffiecients between all of the variables

From the table above, it can be seen that the predictors all seem to be weakly related with the response variable *Rating*; the variable with the strongest correlation is *Reviews*, which has a correlation coefficient of 0.43, while the smallest correlation coefficient (-0.020) is that of the *Type* dummy variable. This further implies that some of the independent variables may not be linearly correlated with the dependent variable *Rating*.

Moreover, at a level of significance of 0.05, most of the independent variables seem to be significantly corelated with each other (i.e. $\rho \neq 0$). However most of the correlations are very weak, with the following exceptions:

- The *Reviews* variable seems to have a strong negative correlation with the 'Installs <100K' dummy variable at value of -0.842. To a lesser degree, it is also positively correlated with the 'Installs=1M-10M' and 'Installs=10M-100M' dummy variables, which have correlation coefficients of 0.364 and 0.514 respectively.

- The ‘Installs<100K’ also seems to be somewhat correlated with the other *Installs* dummy variables: ‘Installs=100K-1M’ (with coefficient -0.380), ‘Installs=1M-10M’ (with coefficient -0.460), and ‘Installs=10M-100M’ (with coefficient -0.358).
- The ‘Content Rating=Everyone’ dummy variable also seems to have high negative correlations with the other *Content Rating* dummy variables: with the ‘Content Rating=10+’ the dummy variable has a correlation coefficient of -0.403, while with the ‘Content Rating=Teen’ it has a correlation with a value of -0.718.

However, it is also of note that some of the variables proved to be independent at a level of significance of 0.05:

- The *Size* variable is independent from the ‘Type=Free’, ‘Content Rating=10+’ and ‘Content Rating=Teen’ dummy variables, with p-values of 0.459, 0.269, and 0.210 respectively.
- The ‘Installs=100K-1M’ dummy variable also seems to be independent from the ‘Content Rating=10+’ and ‘Content Rating=Teen’ dummy variables with p-values of 0.276 and 0.262 respectively.

Hence, in summary, while the predictors do not seem to be strongly correlated with the response variable, the dummy variables of the *Installs* variable seem to show a lot of collinearity between themselves as well as with the *Reviews* variable. Moreover, the dummy variables of the *Content Rating* variable also seem to show collinearity between themselves. However, this has yet to be verified using *Condition Indices* and *Variance Proportions*.

5.1.2 Condition Indices and Variance Proportions

In this section, the condition indices and variance proportions of the independent variables are calculated and analysed to check for any serious signs of collinearity between the variables. The output is given as follows:

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Reviews	Size	Installs=<100K	Installs=100K - 1M	Installs=1M - 10M	Installs=10M - 100M	Type=Free	Content_Rating=Everyone	Content_Rating=10+	Content_Rating=Teen
1	1	4.350	1.000	.00	.00	.02	.00	.00	.00	.00	.00	.00	.00	.00
	2	1.204	1.901	.00	.12	.00	.01	.00	.00	.03	.00	.00	.09	.04
	3	1.026	2.059	.00	.09	.00	.01	.00	.06	.02	.00	.00	.05	.04
	4	1.015	2.070	.00	.02	.00	.01	.09	.00	.01	.00	.00	.08	.03
	5	.984	2.103	.00	.01	.00	.00	.05	.03	.01	.00	.00	.12	.06
	6	.936	2.155	.00	.33	.00	.00	.00	.00	.10	.00	.00	.00	.01
	7	.855	2.256	.00	.15	.02	.01	.00	.00	.01	.00	.00	.19	.07
	8	.514	2.909	.00	.00	.93	.00	.00	.01	.01	.00	.00	.00	.00
	9	.063	8.278	.00	.02	.00	.03	.06	.07	.07	.78	.12	.05	.10
	10	.041	10.340	.00	.13	.00	.40	.34	.37	.34	.01	.57	.26	.42
	11	.011	19.831	.99	.14	.02	.53	.45	.44	.41	.21	.29	.15	.22

a. Dependent Variable: Rating

Figure 37 Condition indices and variance proportions for all the independent variables.

In the table above, it can be observed that only the last three eigenvalues have values greater than five, showing that there are four corresponding near dependencies in the columns of the data matrix for the given predictors. Moreover, since the condition number shown by the last eigen value is less than thirty, it can be safely assumed that there is no serious collinearity between the predictors affecting the model, despite the large correlations found (for the *Installs* and *Content Rating* variables) using the *Spearman* collinearity coefficients.

However, when considering the variance proportions for the last eigen vector (19.831), it can be observed that there is a dependency causing damage to the parameter estimates for the constant term (it has a ratio of 0.99 which is larger than 0.5), and the dummy variable 'Installs=<100K' (which has a ratio of 0.53).

Thus, since the *Installs* variable has been shown to have little correlation to the dependent variable (i.e. *Rating*) in Section 5.1.1 and causes a lot of damage, it is removed from the model.

5.2 Fitting the General Linear Model

In this section the GLM is fitted to the data set, and the assumptions necessary are verified. It is of note that to avoid complicating the model, the interaction variables were not included. First of all, the predictors are checked to see if they are significant by looking at the table of *Tests of Between-Subjects Effects* which tests the following hypotheses:

H_0: independent variable is significant

H_1: independent variable is not significant

The following table shows the outputs of the *Tests of Between-Subjects Effects*:

Tests of Between-Subjects Effects

Dependent Variable: Rating

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	565.387[a]	6	94.231	42.046	.000
Intercept	19710.642	1	19710.642	8794.845	.000
Type	15.864	1	15.864	7.079	.008
Content_Rating	148.112	3	49.371	22.029	.000
Reviews	113.779	1	113.779	50.768	.000
Size	247.700	1	247.700	110.523	.000
Error	24271.748	10830	2.241		
Total	167037.530	10837			
Corrected Total	24837.135	10836			

a. R Squared = .023 (Adjusted R Squared = .022)

Figure 38 Table showing Tests of Between-Subjects Effects

At a level of significance of 0.05, all of the variables can be said to be significant, since all of the p-values are less than 0.05. Hence, all of the variables should be kept in the model. The table also shows the *R Squared* value for the model, which is 0.023; since this value is far lower than one, it shows that the model gives very poor predictions. In fact, from the *R Squared value* we can tell that the model is only able to explain 2.3% of the variability in the response variable. Note also, that the value of the *Adjusted R Squared* is 0.22, but since the model is not being compared to any other model, the value is mostly ignored.

Next, the parameter estimates for the General Linear Model are calculated and a hypothesis test tests the following hypotheses:

$$\mathbf{H_0}: \beta_i = 0$$

$$\mathbf{H_1}: \beta_i \neq 0$$

The output from SPSS shown in the following table:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	3.528	.086		41.024	.000		
	Reviews	3.522E-008	.000	.068	7.125	.000	.987	1.013
	Size	6.822E-006	.000	.100	10.513	.000	.999	1.001
	Type=Free	.147	.055	.025	2.661	.008	.996	1.005
	Content_Rating=Everyone	-.234	.069	-.061	-3.410	.001	.278	3.600
	Content_Rating=10+	.245	.100	.031	2.462	.014	.568	1.761
	Content_Rating=Teen	-.015	.080	-.003	-.189	.850	.330	3.031

a. Dependent Variable: Rating

Figure 39 Table of parameter estimates for the GLM

From the output above, one can observe that only the 'Content Rating=Teen' dummy variable accepts the null-hypothesis. However, since all of the other *Content Rating* predictors are significant, it is still kept in the model. Each covariate/dummy variable in the model is represented by the following notation:

Variable Name	Notation
Reviews	R
Size	S
Type=Free	T
Content Rating=Everyone	C
Content Rating=10+	C*
Content Rating=Teen	C**

Using the values given in the table in Figure 39, the following model is obtained:

$$E[Y_i|R_i, S_i, T_i, C_i, C_i *, C_i **] = 3.528 + (3.522 \times 10^{-8})R_i + (6.822 \times 10^{-6})\, S_i + 0.147T_i - 0.234C_i + 0.245C_i * - 0.15C_i ** |$$

According to the model, if application is free, then its rating is increased by 0.147. Likewise, the rating is also increased by 0.245 if the demographic (i.e. *Content Rating*) of the app is '10+'. However, if the demographic is 'Everyone', then the rating is decreased by 0.234.

5.3 Analysis of the created GLM

After fitting the data, the created General Linear Model is tested to check that there are no influential points (outliers). Moreover, the model is also verified to uphold the assumptions necessary for a GLM.

5.3.1 Goodness-of-fit of the model

Apart from the *R Square* value, an *ANOVA* table is also computed to check the goodness of fit of the model. The table tests the following hypotheses:

H_0: Model using only the intercept term is a good fit for the data

H_1: Model fitted fits better than the model with only the intercept term

The output of the table is shown in the following figure:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	565.387	6	94.231	42.046	.000[b]
	Residual	24271.748	10830	2.241		
	Total	24837.135	10836			

a. Dependent Variable: Rating

b. Predictors: (Constant), Content_Rating=Teen, Size, Type=Free, Content_Rating=10+, Reviews, Content_Rating=Everyone

Figure 40 ANOVA table checking for goodness-of-fit

From the table above, a p-value of zero is obtained. Hence, since this is far smaller than the level of significance (0.05), the null-hypothesis is rejected, and our model seems to be adequate. This contradicts the *R Square* value found previously, which suggests that the model is very poor at predicting the dependent variable.

While these two results seem to contradict one another, the descriptive statistics calculated for the predicted values in Figure 41 (in the next section) supports the result suggested by the *R Square* value; it can be observed that the largest predicted value is 6.74, which is far larger than the maximum possible rating of five.

In conclusion, while the *ANOVA* table suggests that the model is adequate, given the other observations made, it is far more likely that the General Linear Model fits the data poorly.

5.3.2 Outlier Diagnostics

Firstly, to check for outliers, the following outlier diagnostics are computed for the data set: *Studentized residuals*, *Cooks Distance*, *Leverage Value*. Then, the descriptives for these values are calculated to check if any of the values in the data set are flagged as outliers. These can be observed in the table below:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Predicted Value for Rating	10837	3.29	6.74	3.6224	.22842
Studentized Residual for Rating	10837	-3.08	1.14	.0000	1.00000
Cook's Distance for Rating	10837	.00	.03	.0001	.00053
Uncentered Leverage Value for Rating	10837	.00	.07	.0006	.00186
Valid N (listwise)	10837				

Figure 41 Table of descriptive statistics for the outlier diagnostics

First of all, from the table it is known that the values of smallest and largest *studentized residuals* are -3.08 and 1.14 respectively. Hence, since this range falls outside of the range of ±2, some points are flagged out by the *Studentized Residuals* as outliers. To check which data points specifically we flagged out, the data is sorted in ascending order by the value of the *Studentized Residuals*, and a large number of points (1472 rows) were found to have a value less than -2.

	Size	Installs	Type	Content_Rating	filter_$	PRE_1	RES_1	SRE_1
1457	9	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1458	8	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1459	8	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1460	7	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1461	6	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1462	5	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1463	5	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1464	3	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1465	3	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1466	3	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1467	2	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1468	2	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1469	2	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1470	1	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1471	0	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1472	0	<100K	Paid	Everyone	Not Selected	3.29	-3.29	-2.20
1473	21000	<100K	Free	17+	Selected	3.82	-2.82	-1.88
1474	55000	<100K	Free	Everyone	Selected	3.82	-2.82	-1.88

Figure 42 Image showing some of the rows flagged out as outliers by the studentized residuals

Next, the data was checked for outliers again using *Cook's Distance*; from the table in Figure 41, the largest value for *Cook's Distance* (0.3) can be obtained. Since this is smaller than one, no influential points were flagged using this method.

Finally, the data was check for outliers one last time using *Leverage Values*; the maximum *Leverage Value* was obtained from the output in Figure 41, and it is equal to 0.07. Considering that the model has seven parameters (p = 8) and a sample size of 10837 (which can be obtained from the table: n = 10837) the largest value observed in the data set is be much higher than the cut-off value of $\frac{2p}{n} = \frac{2(8)}{10837} = 0.0014$. Hence, outliers we also found using this method. Furthermore, when looking at the data many of the points we observed to be above the cut-off value.

	Size	Installs	Type	Content_Rating	filter_$	PRE_1	RES_1	SRE_1	COO_1	LEV_1
1	1600	100M+	Free	Teen	Selected	6.42	-2.32	-1.61	.03	.07
2	48000	100M+	Free	Teen	Selected	6.74	-2.64	-1.82	.03	.07
3	2800	100M+	Free	Everyone	Selected	5.89	-1.49	-1.02	.01	.05
4	9600	100M+	Free	Everyone	Selected	5.94	-1.54	-1.06	.01	.05
5	29000	100M+	Free	Everyone	Selected	6.07	-1.67	-1.15	.01	.05
6	49000	100M+	Free	Teen	Selected	6.34	-1.84	-1.26	.01	.05
7	5000	100M+	Free	Teen	Selected	6.04	-1.54	-1.05	.01	.05
8	9100	100M+	Free	Teen	Selected	6.07	-1.57	-1.07	.01	.05
9	7300	100M+	Free	Teen	Selected	6.05	-1.55	-1.06	.01	.05
10	2500	100M+	Free	Everyone	Selected	5.45	-1.45	-.99	.01	.03
11	11000	100M+	Free	Everyone	Selected	5.51	-1.51	-1.03	.01	.03
12	13000	100M+	Free	Everyone	Selected	5.52	-1.52	-1.04	.01	.03
13	58000	100M+	Free	10+	Selected	5.90	-1.30	-.88	.00	.02
14	3800	100M+	Free	10+	Selected	5.53	-.93	-.63	.00	.02
15	10000	100M+	Free	10+	Selected	5.57	-.97	-.66	.00	.02
16	12000	100M+	Free	10+	Selected	5.58	-.98	-.66	.00	.02
17	67000	100M+	Free	Everyone	Selected	5.41	-.71	-.48	.00	.02

Figure 43 Image showing a subset of the rows flagged as outliers using Leverage Values.

Therefore, given that two of out of the three diagnostics flagged out, it can be concluded that the data set contains multiple outliers which deviate significantly from the expected value of the fitted model. This may partially explain the results found in Section 5.3.1, suggesting that the General Linear Model may not be a good fit for the data.

5.3.3 Tests for ascertaining the assumptions of the GLM

Lastly, the data is tested to see if the assumptions required to fit a General Linear Model are upheld. Some of the assumptions are satisfied due to the nature of the data and variables used:

- The dependent variable (i.e. *Rating*) is a quantitative variable.
- Both between and within groups, all of the observations in the data set are independent, since each observation represents a different application.

However, some of the assumptions have yet to be shown true:

- The residuals must have a normal distribution, and their variance cannot be influenced by the factors' groupings.
- For each factor level, the dependent variable should be linearly related to each covariate.
- The residuals have to be homoscedastic.

The to test the first part of the first assumption, a Kolmogorov-Smirnov test is conducted on the residuals, and it tests the following hypotheses:

H_0: The residuals follow a normal distribution

H_1: The residuals do not follow a normal distribution

The output for the test in SPSS is given as follows:

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
Studentized Residual for Rating	.267	9732	.000

a. Lilliefors Significance Correction

Figure 44 Kolmogorov-Smirnov test for studentized residuals.

The p-value obtained in the table above (0) is less than the level of significance (0.05). Hence, the null-hypothesis is rejected, and the residuals do not follow a normal distribution, and one of the assumptions is not obeyed. Furthermore, a histogram of the residuals also shows that the residuals in fact have a bimodal distribution:

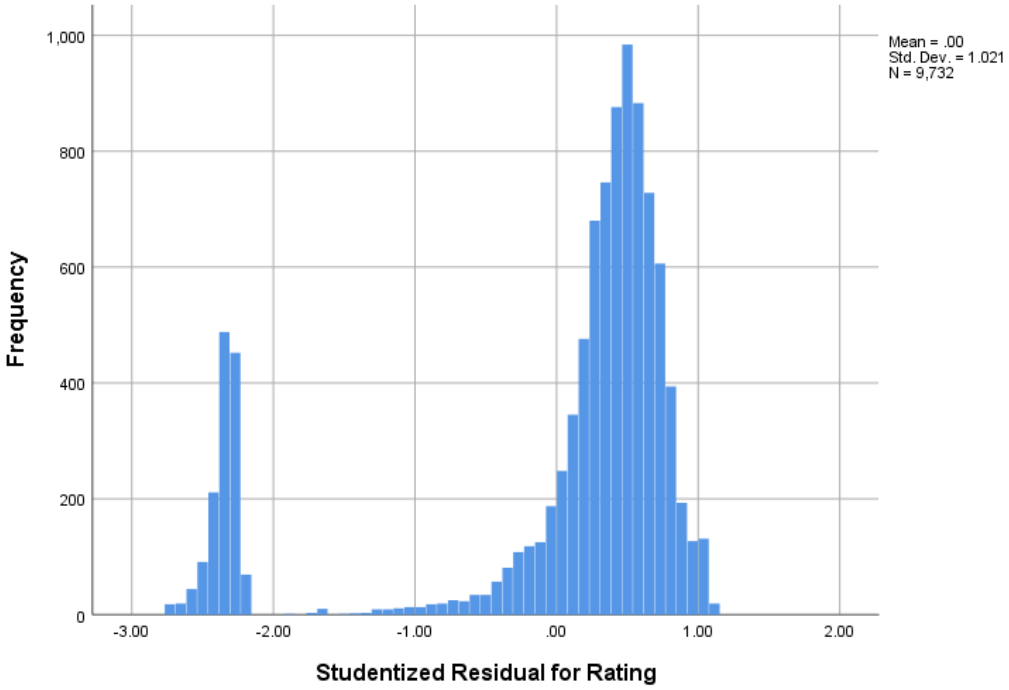


Figure 45 Histogram of standardized residuals.

In addition, the second part of the first assumption is also tested using a *Levene's Test* of Equality of Variances. It tests the following hypotheses:

H_0: The variance of the residuals is equal across all groupings of the factors

H_1: The variance of the residuals is not equal across all groupings of the factors

For the above test the following output was given:

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Rating

F	df1	df2	Sig.
50.338	7	10829	.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + Type + Content_Rating + Reviews + Size

Figure 46 Results for Levene's Test of Equality of Variance of the residuals

From the output in the table above, the p-value can be seen to be zero, which is smaller than the level of significance (0.05). Hence, the null-hypothesis is rejected, and the second part of the assumption is also violated for the data set in question.

Furthermore, the independence of the residuals is also tested. Usually, this would be tested using a Durbin-Watson test, but since this is not available in SPSS, a scatter plot of *predicted values* vs *studentized residuals* is plotted to check if there appear to be any visible patterns:

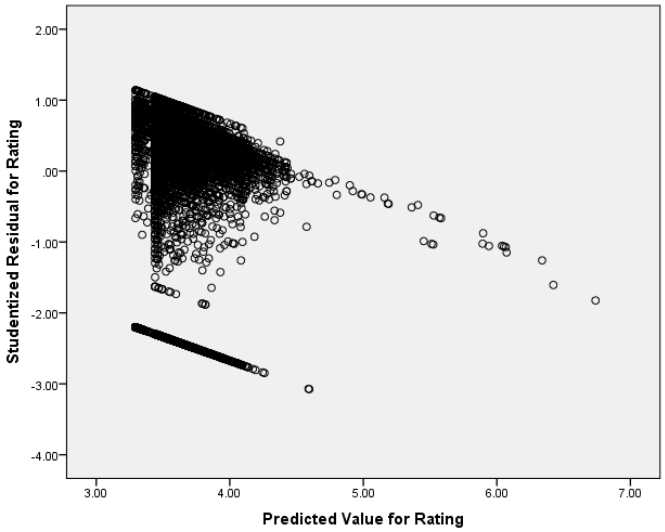


Figure 47 Scatter plot of predicted values vs studentized residuals.

The points in the graph above can be clearly seen to follow some sort of pattern, showing that the model is not well-fitted at all. Hence, it can be assumed that the residuals are not independent

Next, the second assumption (i.e. that the dependent variable is linearly related to the covariates for every factor level) can also be said to be violated, since the scatterplots in Section 3.4 show that the covariates do not have a linear relationship with the *Rating* variable, regardless of the levels of each factor in the model.

Lastly, the homoskedasticity of the residuals is also tested using a *Breusch-Pagan* test, which tests the following hypotheses:

$\mathbf{H_0}$: Residuals are homoscedastic

$\mathbf{H_1}$: Residuals are heteroscedastic

The following output is obtained for the test:

Breusch-Pagan Test for Heteroskedasticity[a,b,c]

Chi-Square	df	Sig.
371.288	1	.000

a. Dependent variable: Rating

b. Tests the null hypothesis that the variance of the errors does not depend on the values of the independent variables.

c. Predicted values from design: Intercept + Type + Content_Rating + Reviews + Size

Figure 48 Outputs of the Breusch-Pagan Test in SPSS

At a level of significance of 0.05, the above test is rejected, and hence the null-hypothesis is rejected, and homoscedasticity is also violated for the data set.

Hence, given that multiple of the assumptions are not upheld by the data set, it cannot be adequately modelled using General Linear Models.

6 Conclusion

From the gathered descriptive statistics and illustrations in section 3 it can be concluded that the most popular applications in the Google Play Store (based on this dataset) are those which are *free* and are targeted to *Everyone*. From the boxplots in subsection 3.3 there are no significant visual discrepancies between the rating for applications in different categories. The scatterplots in subsection 3.4 indicate that there is no linear relationship between the dependent variable (Rating) and the other covariates.

The hypothesis tests on the covariates indicates that an application which has a larger number of installs is more likely to have a higher rating (subsection 4.1). It also suggests that applications which are targeted to the '10+' demographic tend to receive a slightly better rating (subsection 4.2). From subsection 4.3 there were no clear conclusions as to whether a paid app will have a higher rating or vice versa.

In the modelling section (section 5), the Spearman Correlation table indicates that there aren't strong correlations between the predictors and the response variable (Rating). When fitting the general linear model to the dataset it was concluded that if an application was 'Free' then its rating is increased by 0.147. If an application is targeted to the '10+' demographic then its rating increases by 0.245 however if it is targeted to the 'Everyone' demographic, then its rating decreases by 0.234.

From the Outlier Diagnostics in subsection 5.3.2 it was concluded that a large number of values from the dataset were flagged out as outliers, which effected the model in a negative way. Apart from these outliers several assumptions that are needed for the GLM did not hold for this dataset as explained in subsection 5.3.3. This means that this dataset cannot be properly modelled using the GLM.

Another model that could have been used was the Binary Logistic Regression model where the response variable would have been a fixed factor instead of a covariate. The fixed factor version of the *Rating* variable could have two categories: less than 4, greater or equal to 4. The Binary Logistic Regression model would have indicated in which category the rating would fall, but not actually predict it. This was not done due to time limitations. The Beta Regression model could have been used as well, however this was beyond the scope of this study unit.

7 Appendix

7.1 References

M. B. Inguanez, F. Sammut, D. Suda , *Statistical analysis using SPSS and R software*

Dataset: https://www.kaggle.com/lava18/google-play-store-apps