[Title]

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

Table of Contents

[1 Introduction 2](#_Toc10495814)

[2 Aims and Objectives 3](#_Toc10495815)

[3 Descriptive Statistics & Illustrations 4](#_Toc10495816)

[3.1 Scatterplots 4](#_Toc10495817)

[3.1.1 Rating vs Reviews 4](#_Toc10495818)

[3.1.2 Rating vs Size 5](#_Toc10495819)

[4 Modelling 6](#_Toc10495820)

[5 Appendix 7](#_Toc10495821)

[5.1 References 7](#_Toc10495822)

# 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 that were 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.

# 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.

# Descriptive Statistics & Illustrations

## 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.

### 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 1: 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 2). 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 little) 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.

### 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 2 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.

# 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).

## 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:

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

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

**H0:** *Rating* follows a normal distribution for the given group of the *Installs variable*

**H1:** *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:



Figure 3 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:

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:

Figure 4 Q-Q plots of all the 'Install' categories



**H0:** The median values of *Rating* are the same for all categories of *Installs*

**H1:** The median values of *Rating* are different for the categories of *Installs*

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



Figure 5 Outputs of the Kruskal Wallis test

As 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:

**H0:** The median ratings of the *Rating* of the two groups are the same.

**H1:** The median ratings of the *Rating* of the two groups are different.

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



Figure 6 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:



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.

## 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:

**H0:** *Rating* follows a normal distribution for the given group of the *Installs variable*

**H1:** *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:



Figure 7 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:

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.

Figure 8 Q-Q plot charts for the different categories of 'Content Rating'



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:

**H0:** The median values of *Rating* are the same for all categories of *Content Rating*

**H1:** 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:



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:

**H0:** The median ratings of the *Rating* of the two groups are the same.

**H1:** The median ratings of the *Rating* of the two groups are different.

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

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.

Figure 9 Output of the Post-Hoc analysis for each group in 'Content Rating'



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:



Figure 10 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 marketed to this demographic may achieve a higher rating than apps in other demographics.

# Modelling

# Appendix

## References

M. B. Inguanez, F. Sammut, D. Suda , *Statistical analysis using SPSS and R software*, pages 108-111