[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](#_Toc10025714)

[2 Descriptive Statistics & Illustrations 3](#_Toc10025715)

[2.1 Scatterplots 3](#_Toc10025716)

[2.1.1 Rating vs Reviews 3](#_Toc10025717)

[2.1.2 Rating vs Size 4](#_Toc10025718)

[3 Modelling 5](#_Toc10025719)

[4 Appendix 6](#_Toc10025720)

[4.1 References 6](#_Toc10025721)

# 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)
* 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.

# 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) was plotted against the *Reviews* variable (on the x-axis), along with a line of best fit and the output was 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) was 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 was 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.

# Modelling

# Appendix

## References

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