[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. 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 hypotheses, 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 space.

# Descriptive Statistics & Illustrations

## Measurements of Location

### 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 |

The above table 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.

### 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 |

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.

The range, as expected, 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.

### 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 |

The mean size for the applications in this dataset is around 18Mb[[1]](#footnote-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.

## 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 : 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 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

1. Results are in Kb [↑](#footnote-ref-1)