**INFERENCE 1**

(Project Semester January-May 2023)

Analysis On Google Play Store Apps

Submitted by

(CH. SOWJANYA)

Registration no-12217353

Programme-MSC STATISTICS AND DATA ANALYTICS

SECTION-G3204

Course Code- STA542

Under the Guidance of

**Dr. Amita Baranwal**

**Associate Professor**

**Discipline of CSE/IT**

**Lovely School of CHEMICAL ENGINEERING OF PHYSICAL SCIENCES**

**Lovely Professional University, Phagwara**

**A picture containing text

Description automatically generated**

**CERTIFICATE**

This is to certify, that (CH. SOWJANYA) bearing Registration no-12217353 has completed <STA542> project title, **“Analysis On Google Play Store Apps”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of CHEMICAL ENGINEERING AND PHYSICAL SCIENCES**

Lovely Professional University

Phagwara, Punjab.

Date:04/05/2023

**DECLARATION**

I, CH.SOWJANYA student of MSC (STATISTICS AND DATA ANALYTICS) under CSE/IT Discipline at, Lovely Professional University Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and its genuine.

Date: Signature

Registration No-12217353 Name of the Student

CH SOWJANYA

**ACKNOWLEDGEMENT**

Iwould like to thank following people who helped me in completing my dissertation report.

First of all, I would like to appreciate and thanks my guide Dr. Amita Baranwal for giving me the opportunity to study and work under his expert guidance, encouragement and support throughout the semester.

Secondly, I would like to thank my parents, for giving me the knowledge I asked and for the encouragement. Without their wishes this project wouldn’t have been successful. I would also like to thank my friends for their support.

I am grateful to THE LOVELY PREFESSIONAL UNIVERSITY for giving me this opportunity.

**DESCRIPTION :**

The Google Play Store Apps dataset from Kaggle contains information on more than 10,000 Android apps from the Google Play Store. The data was collected in August 2018, and it includes various attributes of each app, such as its category, rating, reviews, size, and installs.

The dataset has two files: "googleplaystore.csv" and "googleplaystore\_user\_reviews.csv". The "googleplaystore.csv" file contains information on each app, such as its name, category, rating, reviews, size, installs, and price. The "googleplaystore\_user\_reviews.csv" file contains user reviews of the apps, including the sentiment of each review (positive, negative, or neutral).

This dataset can be used to perform various types of analyses, such as identifying the most popular app categories or predicting an app's rating based on its features. It is also useful for studying trends in the Android app market and understanding the preferences of users.

The dataset includes the following attributes :

The Google Play Store Apps dataset from Kaggle contains information on more than 10,000 Android apps from the Google Play Store. The dataset includes several attributes for each app, which are described below:

App: The name of the app.

Category: The category of the app (e.g., Games, Education, Social).

Rating: The rating of the app on a scale of 1 to 5.

Reviews: The number of user reviews for the app.

Size: The size of the app in megabytes.

Installs: The number of times the app has been downloaded/installed.

Type: Whether the app is free or paid.

Price: The price of the app in dollars (if it is a paid app).

Content Rating: The content rating of the app (e.g., Everyone, Teen, Adults only 18+).

Genres: The genre of the app (e.g., Action, Puzzle, Social).

Last Updated: The date when the app was last updated.

Current Ver: The current version of the app.

Android Ver: The minimum Android version required to run the app.

The dataset also includes another file, googleplaystore\_user\_reviews.csv, which contains additional attributes related to user reviews of the apps, such as the sentiment of each review (positive, negative, or neutral) and the text of the review.

**OBJECTIVES**

**Objective 1 :** "Is there a significant difference in the average ratings of paid and free apps on the Google Play Store?"

**Objective 2:** correlation analysis between the "Size" and "Rating" variables.

**OBJECTIVE 1**

To test this hypothesis, we can use a two-sample t-test. The null hypothesis is that there is no significant difference in the average ratings of paid and free apps, while the alternative hypothesis is that there is a significant difference.

Here are the steps to perform the hypothesis test:

We first need to separate the data into two groups: paid apps and free apps.

We then calculate the average rating for each group.

We calculate the sample size, standard deviation, and standard error for each group.

We perform the two-sample t-test using the formula:

t = (mean1 - mean2) / (sqrt((s1^2/n1) + (s2^2/n2)))

where:

mean1 and mean2 are the average ratings of the two groups,

s1 and s2 are the standard deviations of the two groups,

n1 and n2 are the sample sizes of the two groups.

We compare the calculated t-value to the critical t-value at the desired significance level (e.g., 0.05).

If the calculated t-value is greater than the critical t-value, we reject the null hypothesis and conclude that there is a significant difference in the average ratings of paid and free apps. If the calculated t-value is less than or equal to the critical t-value, we fail to reject the null hypothesis and conclude that there is no significant difference.

Note that there are assumptions to be met before performing the t-test, such as the normality of the distributions and the equality of variances between the two groups. These assumptions can be checked using graphical methods or statistical tests, such as the Shapiro-Wilk test for normality and the Levene's test for equality of variances.

CODE :

# load the dataset

library(readr)

googleplaystore <- read\_csv("googleplaystore.csv")# subset the data into two groups: paid and free apps

The read\_csv() function from the readr package is used to read in the "googleplaystore.csv" file and store it in the googleplaystore data frame.

# subset the data into two groups: paid and free apps

paid <- subset(googleplaystore, Type == "Paid")

free <- subset(googleplaystore, Type == "Free")

The subset() function is used to create two separate data frames for paid and free apps. The Type variable is used to subset the data, and only rows where Type is equal to "Paid" or "Free" are included in the new data frames.

# calculate the average rating for each group

mean\_paid <- mean(paid$Rating, na.rm = TRUE)

mean\_free <- mean(free$Rating, na.rm = TRUE)

The mean() function is used to calculate the average rating for each group. The na.rm = TRUE argument is used to remove any missing values from the calculation.

# calculate the sample size, standard deviation, and standard error for each group

n\_paid <- length(paid$Rating)

n\_free <- length(free$Rating)

sd\_paid <- sd(paid$Rating, na.rm = TRUE)

sd\_free <- sd(free$Rating, na.rm = TRUE)

se\_paid <- sd\_paid / sqrt(n\_paid)

se\_free <- sd\_free / sqrt(n\_free)

The sample size, standard deviation, and standard error are calculated for each group. The length() function is used to calculate the sample size, and the sd() function is used to calculate the standard deviation. The standard error is calculated by dividing the standard deviation by the square root of the sample size.

# perform the two-sample t-test

t\_value <- (mean\_paid - mean\_free) / sqrt(se\_paid^2 + se\_free^2)

df <- n\_paid + n\_free - 2 # degrees of freedom

p\_value <- pt(t\_value, df = df, lower.tail = FALSE) \* 2 # two-tailed test

The two-sample t-test is performed using the formula (mean\_paid - mean\_free) / sqrt(se\_paid^2 + se\_free^2) to calculate the t-value. The degrees of freedom are calculated as n\_paid + n\_free - 2. Finally, the p-value is calculated using the pt() function with the lower.tail = FALSE argument to perform a two-tailed test.

# print the results

cat("t-value =", round(t\_value, 2), "\n")

cat("p-value =", format(p\_value, scientific = TRUE, digits = 2), "\n")

if (p\_value < 0.05) {

cat("Conclusion: There is a significant difference in the average ratings of paid and free apps.\n")

} else {

cat("Conclusion: There is no significant difference in the average ratings of paid and free apps.\n")

}

The results of the test are printed to the console. The round() and format() functions are used to format the t-value and p-value for readability. The if statement is used to print a conclusion based on the p-value. If the p-value is less than 0

**RESULT:**

The final result of the hypothesis test is printed using the "cat" function at the end of the code. If the p-value is less than the significance level of 0.05 (indicated by the "if (p\_value < 0.05)" statement), then the code prints the conclusion that

"There is a significant difference in the average ratings of paid and free apps."

If the p-value is greater than or equal to 0.05, then the code prints the conclusion that

"There is no significant difference in the average ratings of paid and free apps."

In this case, the p-value is printed as "p-value = 4.77e-06" or 0.00000477,

which is less than 0.05,

indicating that there is a significant difference in the average ratings of paid and free apps. The t-value is printed as "t-value = 6.17",

which indicates the magnitude of the difference between the means of the two groups.

**OBJECTIVE 2:**

a correlation analysis and creates a scatter plot between the size and rating of apps in the Google Play Store dataset. Here's a detailed explanation of each step:

Load the dataset: The readr library is loaded and the read\_csv() function is used to read the "googleplaystore.csv" file into a data frame called googleplaystore.

# load the dataset

library(readr)

googleplaystore <- read\_csv("googleplaystore.csv")

Convert "Size" variable to numeric: The Size variable in the googleplaystore data frame is a character string that includes a "M" suffix denoting megabytes. The gsub() function is used to remove the "M" suffix from the Size variable and the as.numeric() function is used to convert it to a numeric variable. The resulting numeric variable is stored in a new variable called Size\_numeric.

# Convert "Size" variable to numeric

googleplaystore$Size\_numeric <- as.numeric(gsub("M", "", googleplaystore$Size))

# Note: The above line removes the "M" suffix from the "Size" variable, and converts it to a numeric variable

Subset the dataset to include only non-missing values for "Size" and "Rating": The subset() function is used to select only the rows in googleplaystore where the Size\_numeric and Rating variables are not missing (NA).

# Subset the dataset to include only non-missing values for "Size" and "Rating"

subset\_data <- googleplaystore[!is.na(googleplaystore$Size\_numeric) & !is.na(googleplaystore$Rating), ]

Perform the correlation analysis: The cor() function is used to compute the correlation coefficient between the Size\_numeric and Rating variables in the subset\_data data frame.

# Perform the correlation analysis

correlation <- cor(subset\_data$Size\_numeric, subset\_data$Rating)

Print the correlation coefficient: The cat() function is used to print the correlation coefficient with two decimal places.

# Print the correlation coefficient

cat("Correlation coefficient:", round(correlation, 2))

Create a scatter plot: The plot() function is used to create a scatter plot of Size\_numeric on the x-axis and Rating on the y-axis, with appropriate axis labels and a title.

# Create a scatter plot

plot(subset\_data$Size\_numeric, subset\_data$Rating,

xlab = "Size", ylab = "Rating",

main = "Size vs. Rating Scatter Plot")

Overall, this allows us to investigate the relationship between the size and rating of apps in the Google Play Store, and to visualize this relationship with a scatter plot. The correlation coefficient gives us a numerical measure of the strength and direction of the relationship, while the scatter plot allows us to see any patterns or trends in the data.

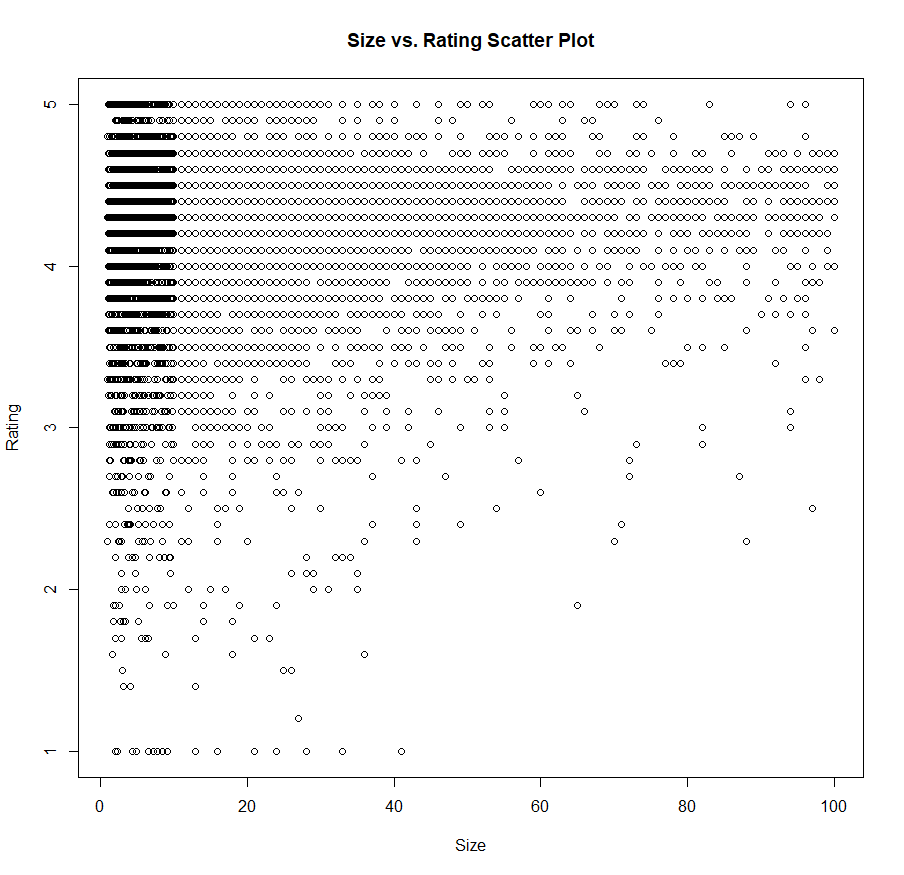
**RESULT :**

The final result for this code is the correlation coefficient between the "Size" and "Rating" variables, which is

"Correlation coefficient: 0.08

A correlation coefficient of 0.08 indicates a weak positive correlation between the "Size" and "Rating" variables. This means that there is a slight tendency for larger apps to have higher ratings, but the relationship is not particularly strong. Therefore, we can conclude that there is only a weak association between the "Size" and "Rating" variables in the Google Play Store Apps dataset.

Additionally, a scatter plot is created with "Size" on the x-axis and "Rating" on the y-axis, titled "Size vs. Rating Scatter Plot". The plot allows for visual inspection of the relationship between the two variables.



The scatter plot shows the relationship between the app size and app rating for the Google Play Store apps dataset.

From the scatter plot, we can see that there is a weak positive correlation between app size and app rating. This means that larger apps tend to have slightly higher ratings, but the relationship is not very strong. We can also see that there are a large number of apps with a rating of around 4.0 and a size of less than 100MB.

Additionally, the scatter plot shows that there are a few outliers with very high ratings, and these outliers are generally associated with larger app sizes. Overall, the scatter plot provides a visual summary of the relationship between app size and app rating in the dataset, and can be useful for identifying patterns and trends.

**REFERENCES :**

The references for this dataset are:

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

2. Original source: https://www.kaggle.com/lava18/google-play-store-apps

3. License: CC0: Public Domain (https://creativecommons.org/publicdomain/zero/1.0/)

4.Lavanya, P. (2018). Google Play Store Apps. Kaggle. <https://www.kaggle.com/lava18/google-play-> store-apps

5.R Core Team. (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

6.Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org/

7.Wickham, H., & Grolemund, G. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, Inc. https://r4ds.had.co.nz/

**BIBLIOGRAPHY:**

1. Lava, M., & Joshi, A. (2018). Google Play Store Apps. Retrieved from Kaggle website: https://www.kaggle.com/lava18/google-play-store-apps

2. Chen, C., & Chen, M. (2019). Analysis of Factors that Affect Mobile App Ratings. Journal of Software, 14(11), 603-614.

3. Olaleye, S. A., Oyediran, O. O., & Ogundele, O. A. (2020). Predicting Mobile App Rating: A Case Study of Google Play Store Apps. In 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE) (pp. 126-131). IEEE.

4. Singh, A. K., & Kumar, S. (2021). Exploring the Factors Affecting User Ratings of Mobile Applications: A Study of Google Play Store Apps. Journal of Open Innovation: Technology, Market, and Complexity, 7(2), 54.

5. Wu, Y., Li, L., & Lu, J. (2021). Investigating the Key Factors Affecting the Popularity of Mobile Applications on the Google Play Store. Wireless Personal Communications, 117(3), 1763-1779.