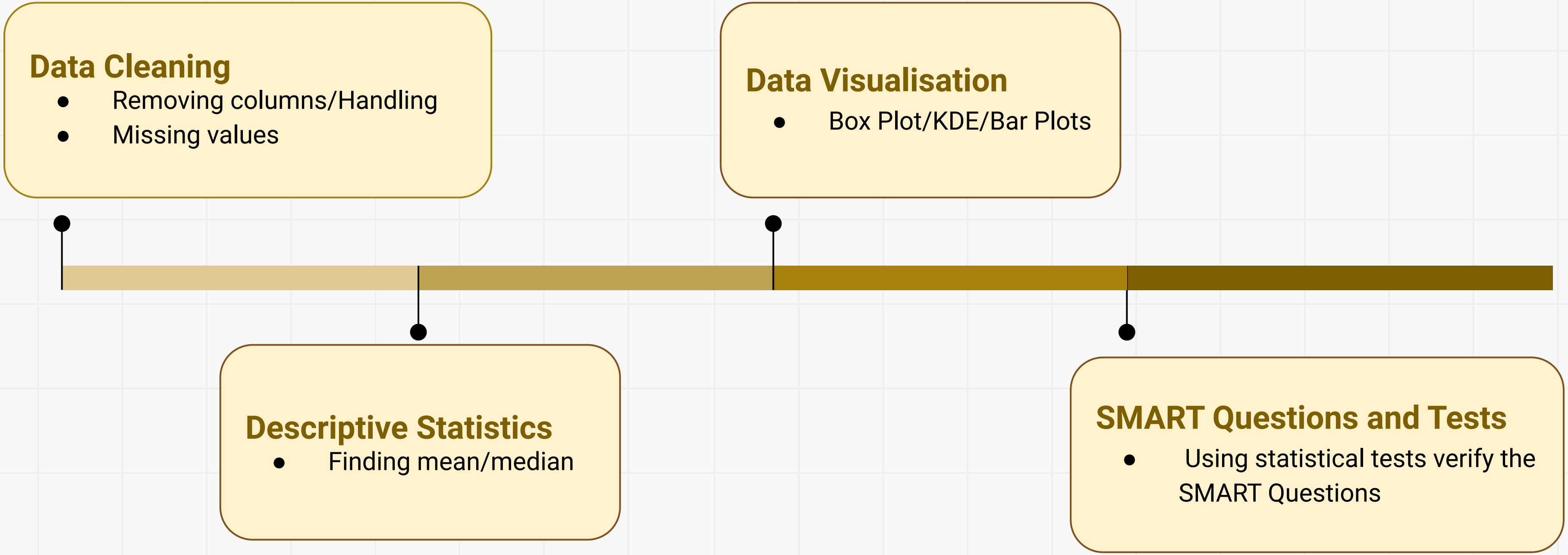
DATS 6101

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

# Exploratory Data Analysis:
## Google Play Store Apps

Snehitha Tadapaneni
Sai Rachana Kandikattu
Amrutha Jayachandradhara
Pramod Krishnachari
Wilona Nguyen

# Methodology



### Data Cleaning

- Removing columns/Handling
- Missing values

### Descriptive Statistics

- Finding mean/median

### Data Visualisation

- Box Plot/KDE/Bar Plots

### SMART Questions and Tests

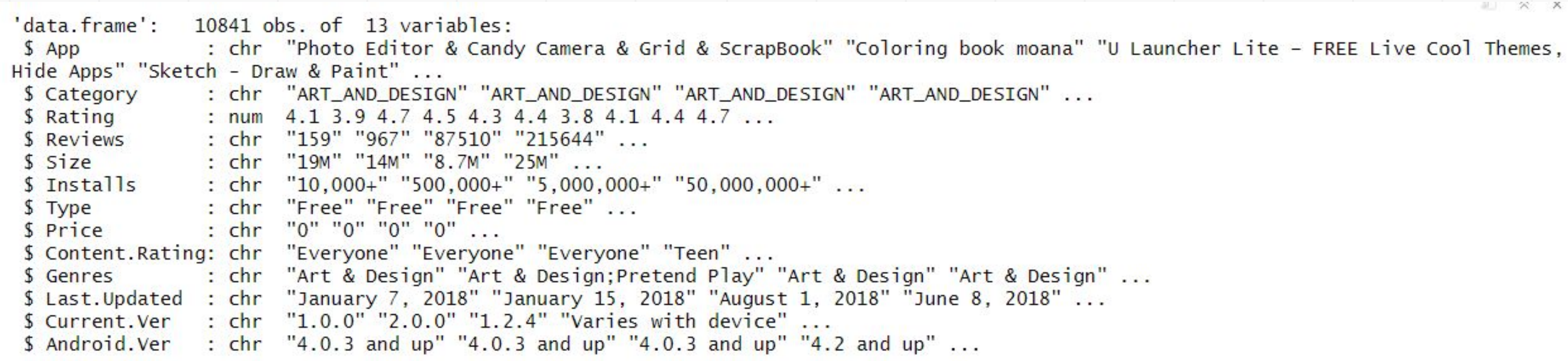
- Using statistical tests verify the SMART Questions

# Dataset Overview

**10, 841**
Observations

**13**
Variables

**Source**
Kaggle



```
'data.frame':	10841 obs. of  13 variables:
 $ App           : chr  "Photo Editor & Candy Camera & Grid & ScrapBook" "Coloring book moana" "U Launcher Lite – FREE Live Cool Themes,
Hide Apps" "Sketch - Draw & Paint" ...
 $ Category      : chr  "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" "ART_AND_DESIGN" ...
 $ Rating        : num  4.1 3.9 4.7 4.5 4.3 4.4 3.8 4.1 4.4 4.7 ...
 $ Reviews       : chr  "159" "967" "87510" "215644" ...
 $ Size          : chr  "19M" "14M" "8.7M" "25M" ...
 $ Installs      : chr  "10,000+" "500,000+" "5,000,000+" "50,000,000+" ...
 $ Type          : chr  "Free" "Free" "Free" "Free" ...
 $ Price         : chr  "0" "0" "0" "0" ...
 $ Content.Rating: chr  "Everyone" "Everyone" "Everyone" "Teen" ...
 $ Genres        : chr  "Art & Design" "Art & Design;Pretend Play" "Art & Design" "Art & Design" ...
 $ Last.Updated  : chr  "January 7, 2018" "January 15, 2018" "August 1, 2018" "June 8, 2018" ...
 $ Current.Ver   : chr  "1.0.0" "2.0.0" "1.2.4" "Varies with device" ...
 $ Android.Ver   : chr  "4.0.3 and up" "4.0.3 and up" "4.0.3 and up" "4.2 and up" ...
```

# Features

### App

Application name

### Category

Category the app belongs to

### Rating

Overall user rating of the app

### Reviews

Number of user reviews for the app

### Size

Size of the app

### Installs

Number of user installs for the app

### Type

Paid or Free

# Features

### Price

Price of the app

### Content Rating

Age group that app is targeted at

### Genres

Genre of the app within its category

### Last Updated

Date when the app was last updated

### Current Ver

Current version of the app

### Android Ver

Minimum required Android version

# Smart Question

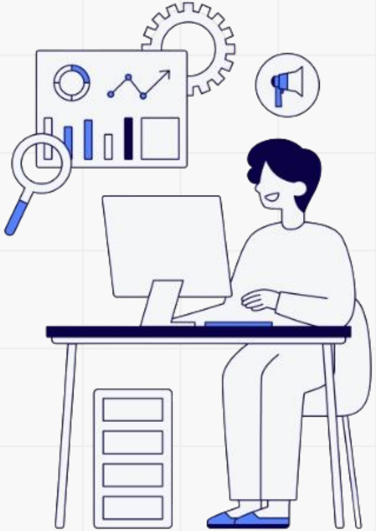
What is the impact of content rating, required App version, category, size, last updated and pricing on predicting app success in terms of positive rating, high user reviews, as well as the number of installs, using data from Google Play Store apps from 2010 to 2018?

# Data Cleaning

Removed Duplicated Apps

Replaced missing value with mean

Dropped Rows with missing price values

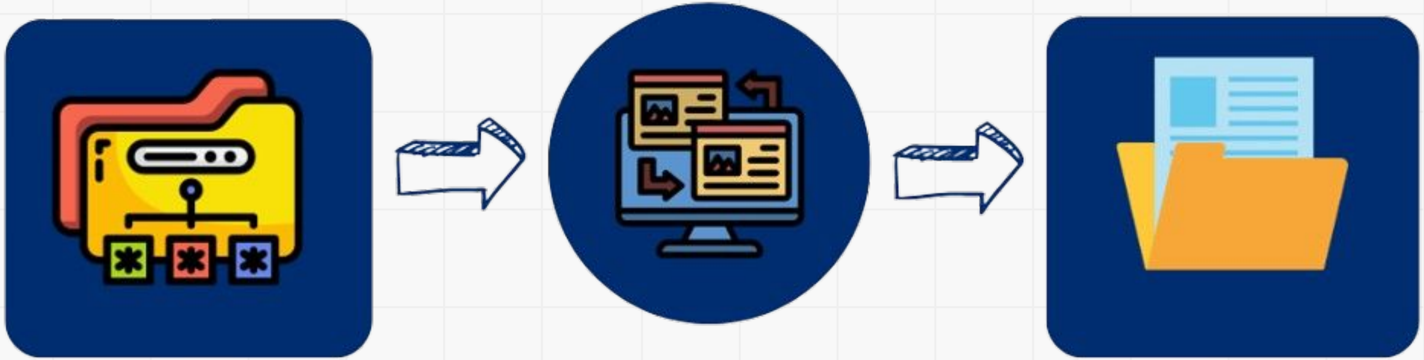


Dropped Irrelevant Columns

Data Format Conversion

All Good!

# After dropping duplicated Apps



Initial Dataset

10,841

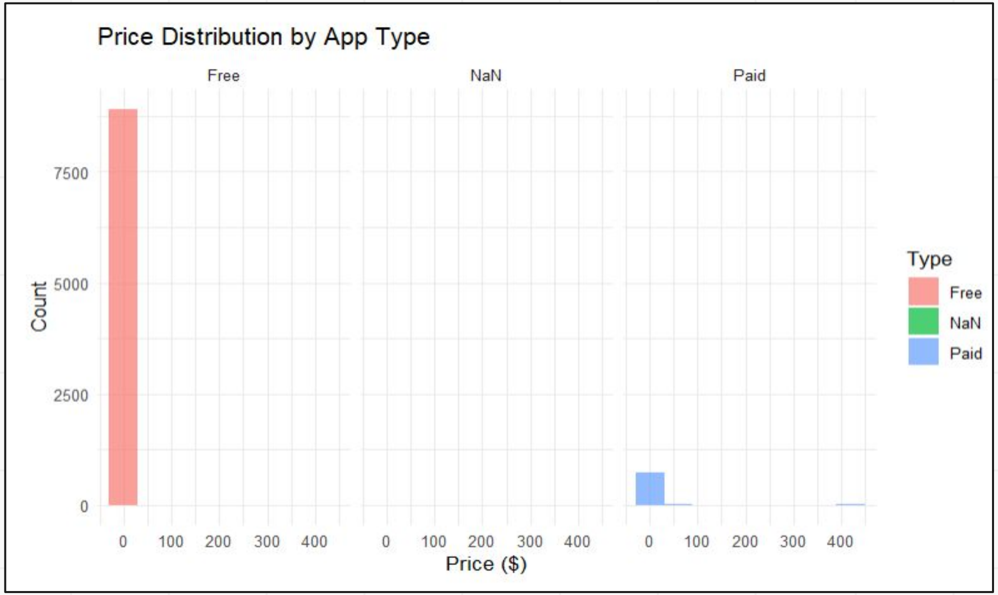
404

Duplicate Apps

After Dropping

9,659

# Dropped Type Column



After dropping

**9,659**

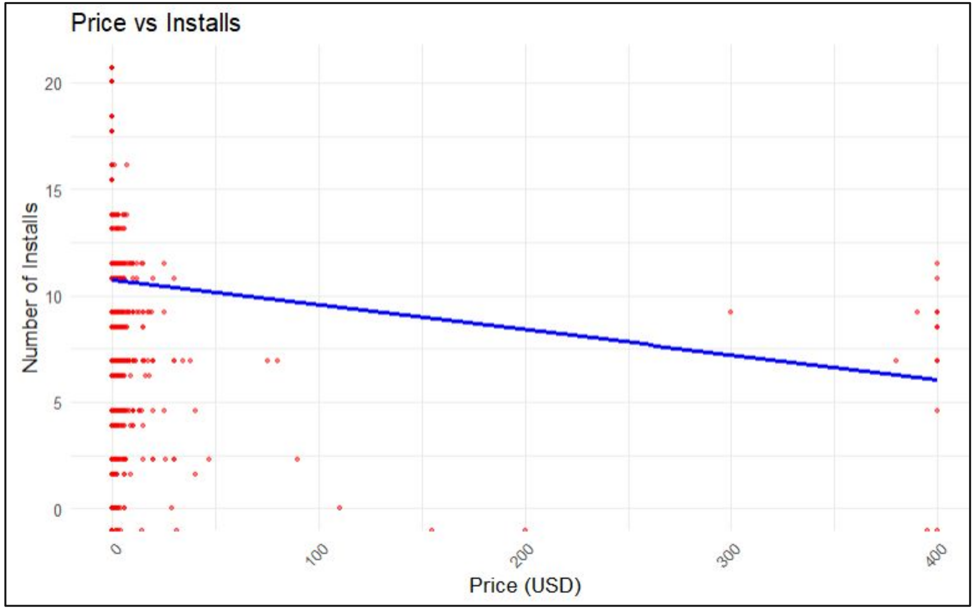
Observations

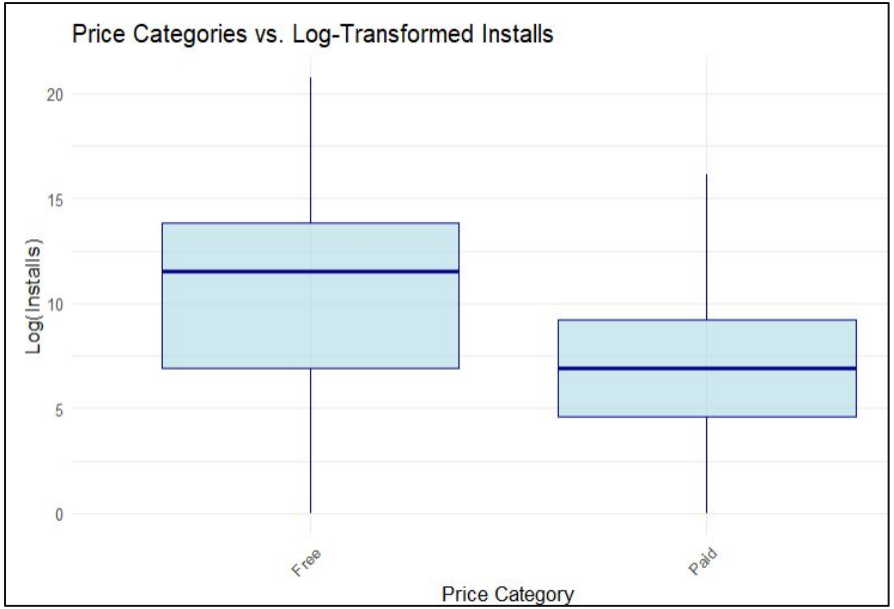
**12**

Variables

# Smart Q1 : Installs vs price

Does price significantly impact the popularity of an app in terms of installs?





# Statistical Test: t-test

### T-test Result:

- → **Test Statistic (t)**: 29.042
- → **Degrees of Freedom (df)**: 977.19
- → **P-value**: < 2.2e-16

| Price Category | Mean Log(Installs) | App Count |
|---|---|---|
| Free | 11.002709 | 8898 |
| Paid | 7.284829 | 746 |

Since the p-value is extremely small, we reject the null hypothesis concluding that the difference in mean log installs between free and paid apps is statistically significant.

# Genre, Current Version

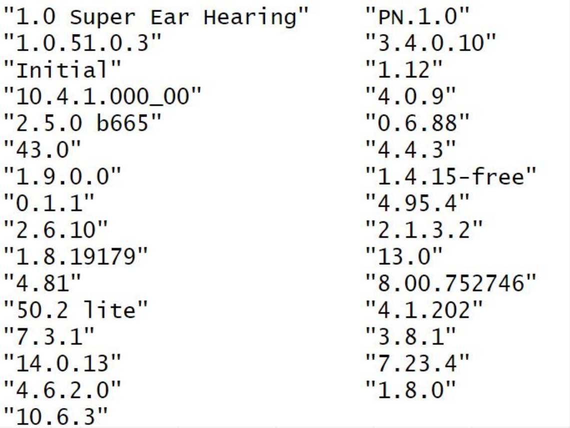
**33**

Categories

**118**

Genres

Drop Genres, proceed with Category



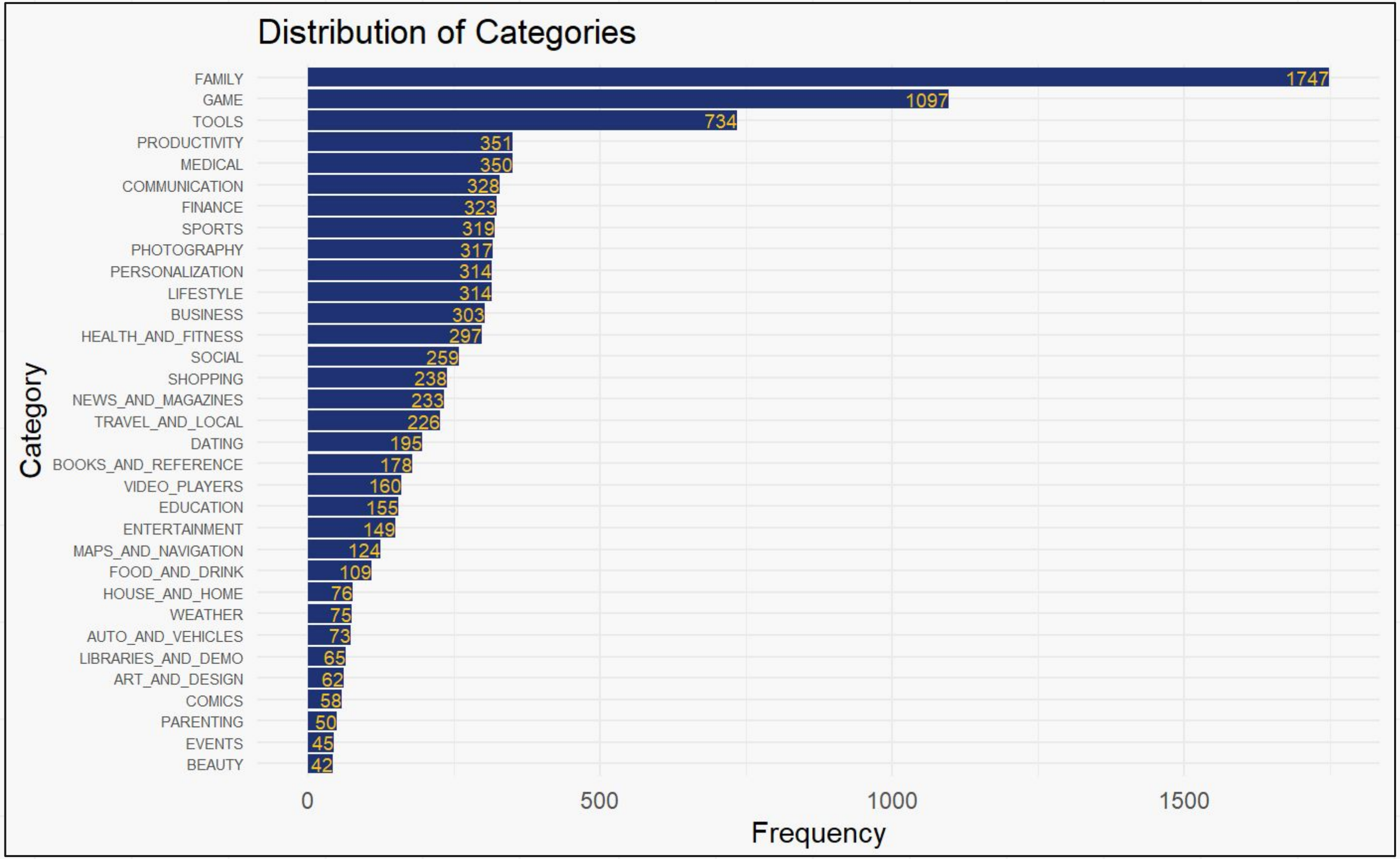
```
"1.0 Super Ear Hearing"     "PN.1.0"
"1.0.51.0.3"                "3.4.0.10"
"Initial"                   "1.12"
"10.4.1.000_00"             "4.0.9"
"2.5.0 b665"                "0.6.88"
"43.0"                      "4.4.3"
"1.9.0.0"                   "1.4.15-free"
"0.1.1"                     "4.95.4"
"2.6.10"                    "2.1.3.2"
"1.8.19179"                 "13.0"
"4.81"                      "8.00.752746"
"50.2 lite"                 "4.1.202"
"7.3.1"                     "3.8.1"
"14.0.13"                   "7.23.4"
"4.6.2.0"                   "1.8.0"
"10.6.3"
```

**Inconsistent Formatting**

Current.Ver

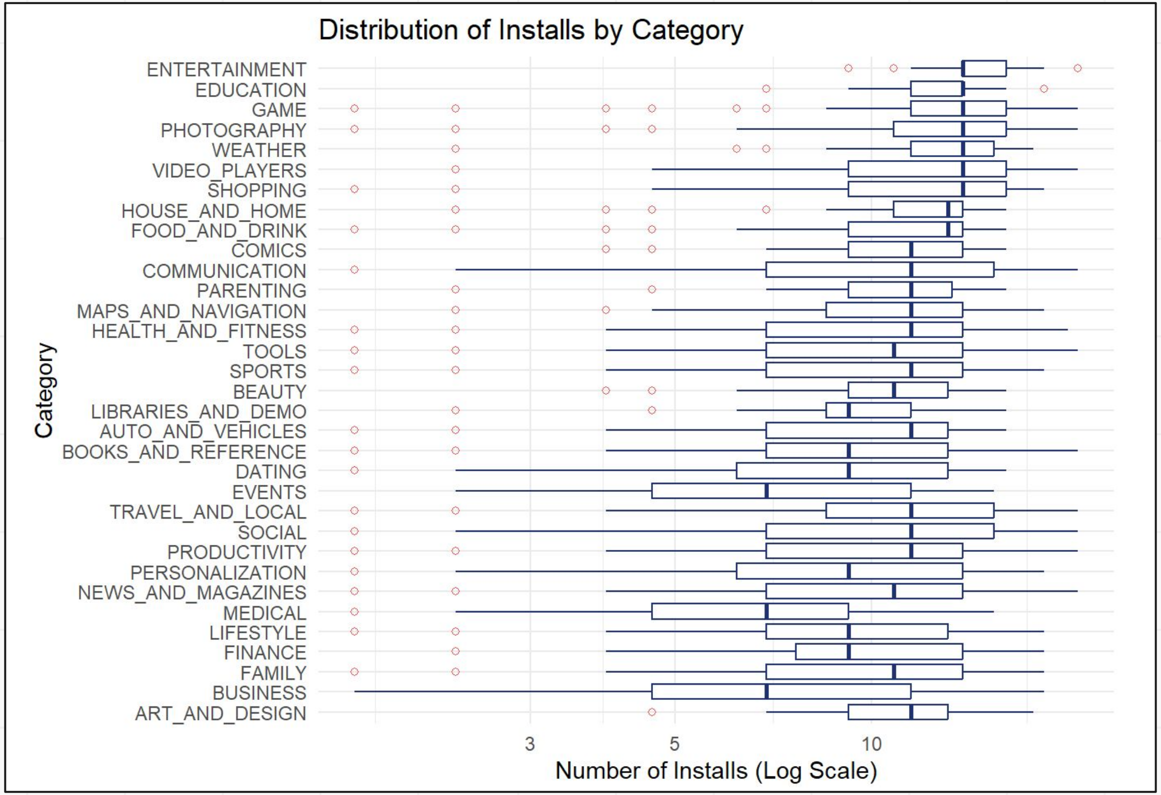
Drop Current.Ver, excluded from the analysis

# Category

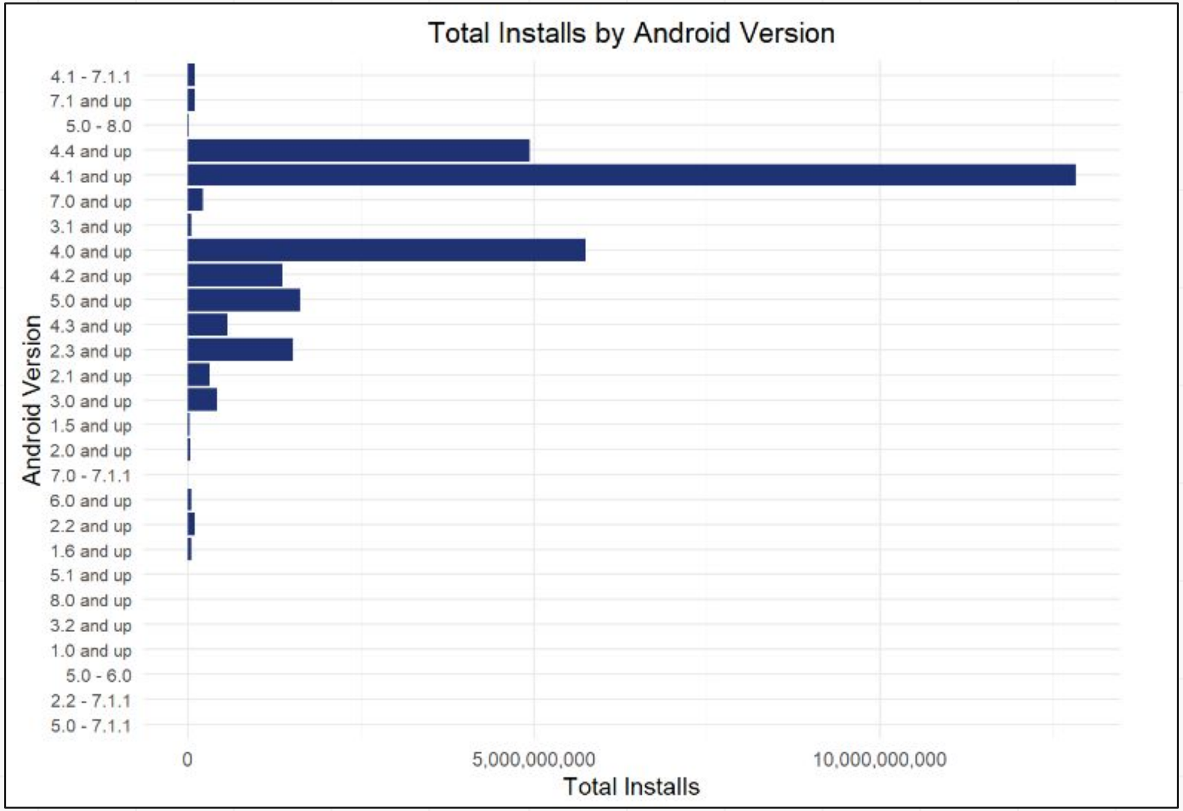


Distribution of Categories
FAMILY 1747
GAME 1097
TOOLS 734
PRODUCTIVITY 351
MEDICAL 350
COMMUNICATION 328
FINANCE 323
SPORTS 319
PHOTOGRAPHY 317
PERSONALIZATION 314
LIFESTYLE 314
BUSINESS 303
HEALTH_AND_FITNESS 297
SOCIAL 259
SHOPPING 238
NEWS_AND_MAGAZINES 233
TRAVEL_AND_LOCAL 226
DATING 195
BOOKS_AND_REFERENCE 178
VIDEO_PLAYERS 160
EDUCATION 155
ENTERTAINMENT 149
MAPS_AND_NAVIGATION 124
FOOD_AND_DRINK 109
HOUSE_AND_HOME 76
WEATHER 75
AUTO_AND_VEHICLES 73
LIBRARIES_AND_DEMO 65
ART_AND_DESIGN 62
COMICS 58
PARENTING 50
EVENTS 45
BEAUTY 42
Category
0
500
1000
1500
Frequency

# Category vs. Installs



# Android Ver vs. Installs



# Rating

STEP 1:

**Checking missing values**

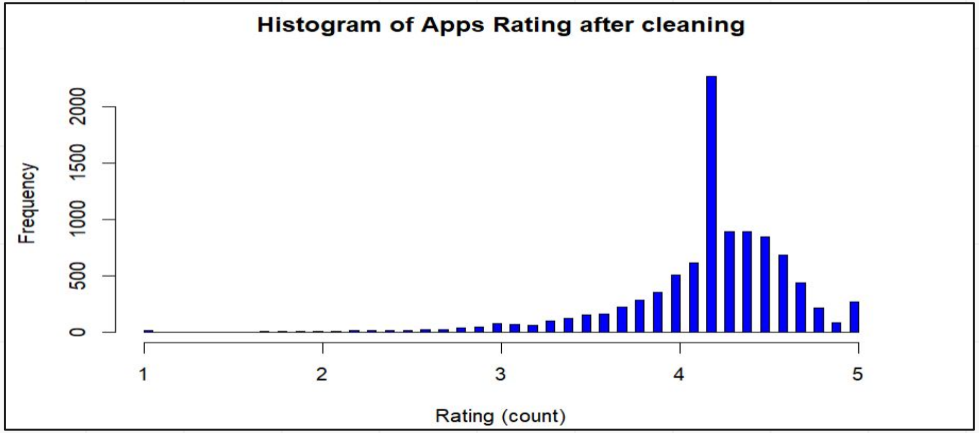
# 1463 NA values

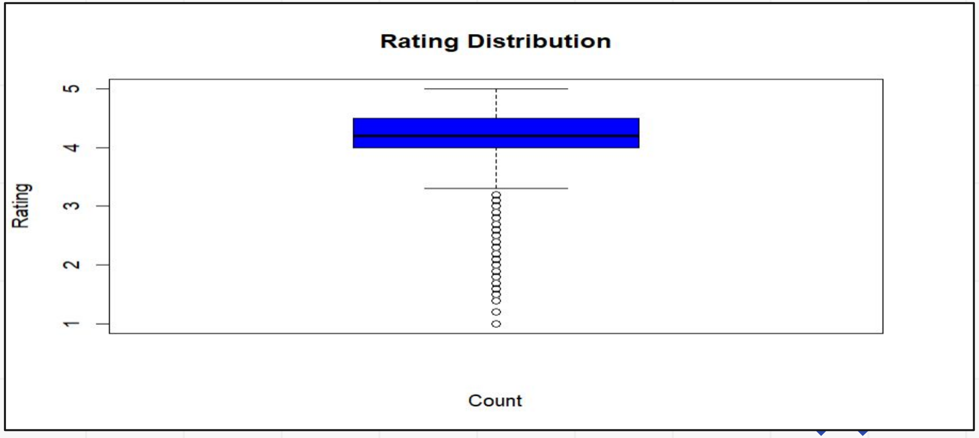
STEP 2:

**Replaced NA values**

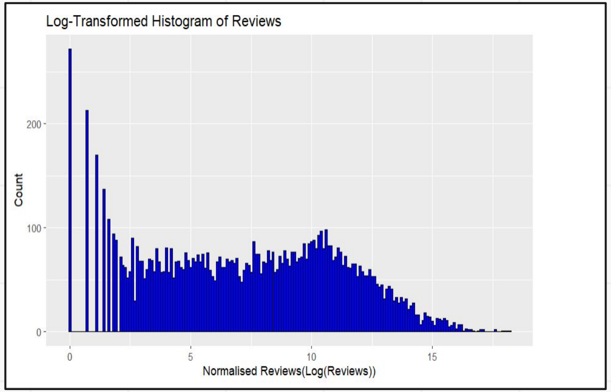
Replaced NA values with mean value.

**Majority Rating values ~[ 3-5]**
**Outliers ~ [1-3)**



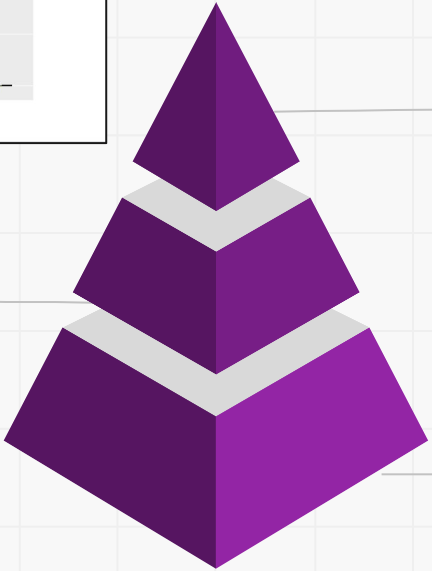


# Reviews



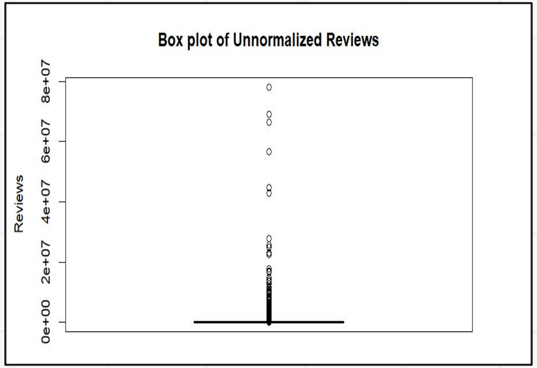
## Review distribution

2 Checked the review distribution, used various visualisations and **normalised reviews** for better visualisation.



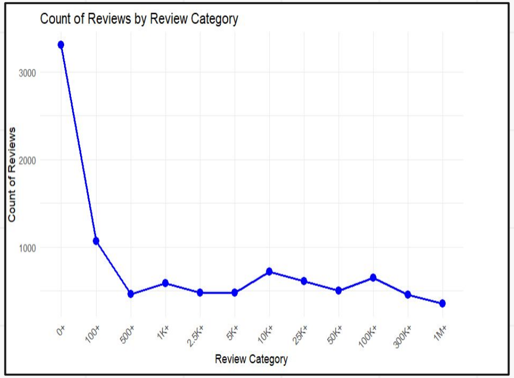
## Data Type Check

1 Converted Review variable data type from **CHAR -> INT** (numeric value)

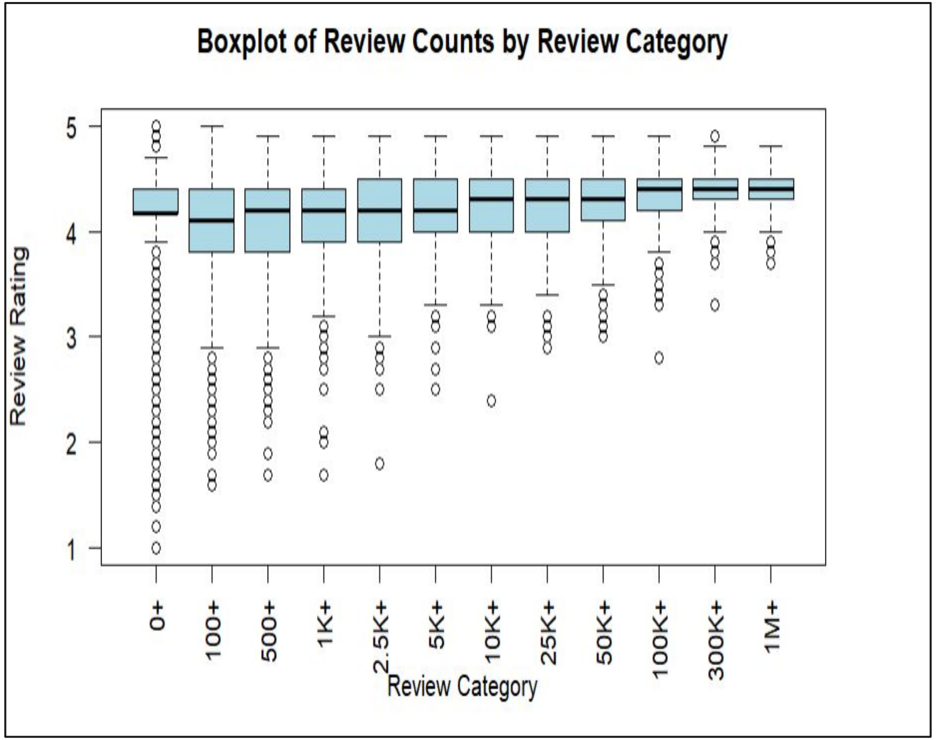


## Review_category variable

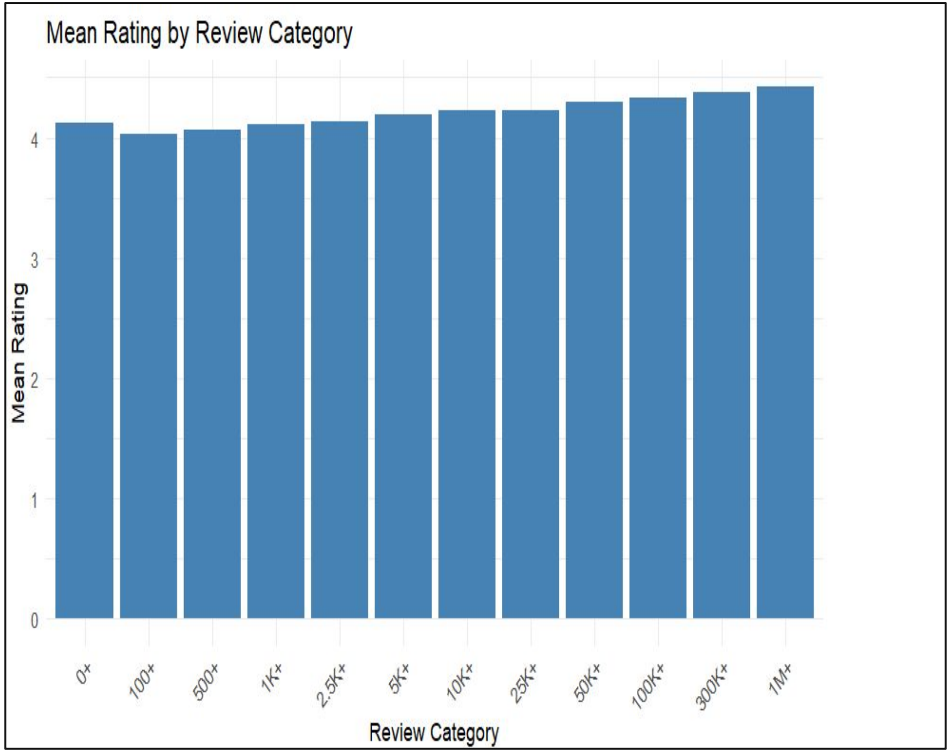
3 Divided Reviews into equal number of reviews sections **(Binning)** for easy representation and further analysis into **review_category.**



# Rating vs Reviews



Boxplot of Review Counts by Review Category



Mean Rating by Review Category

# Statistical Test: ANOVA-test

## ANOVA Result:

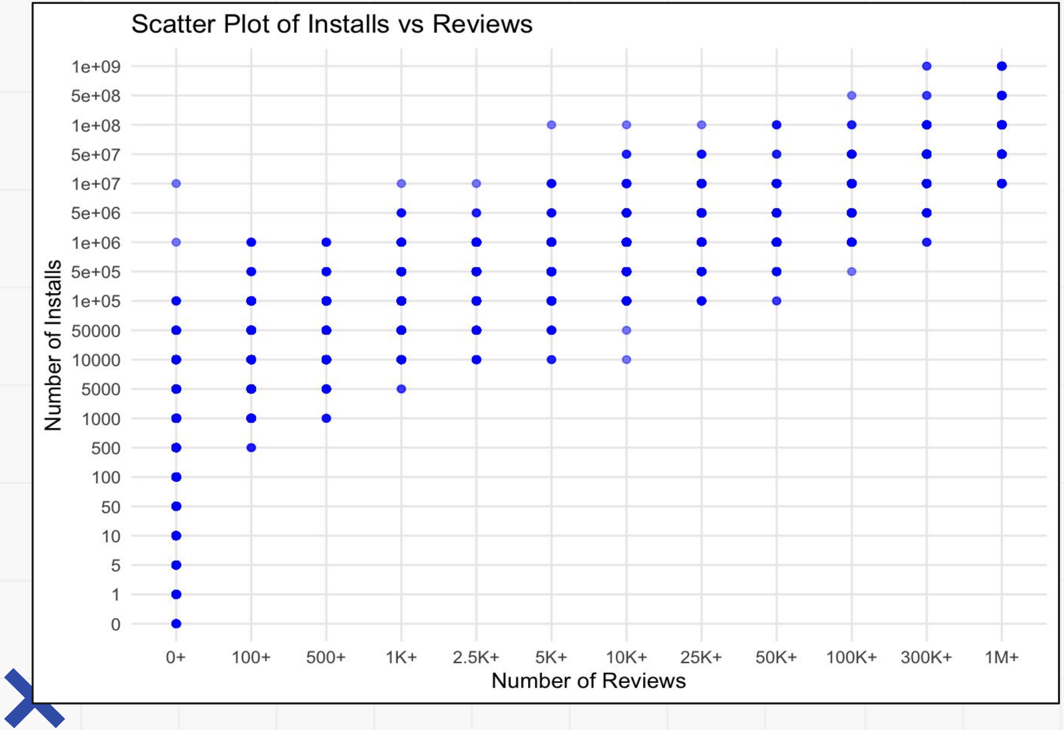
- ➔ **F value (f)**: 41.3
- ➔ **Degrees of Freedom (df)**: 11
- ➔ **P-value**: < 2e-16

| Review_Category | Mean_Rating |
|---|---|
| 0+ | 4.13 |
| 100+ | 4.03 |
| 500+ | 4.06 |
| 1K+ | 4.11 |
| 2.5K+ | 4.13 |
| 5K+ | 4.19 |
| 10K+ | 4.22 |
| 25K+ | 4.23 |
| 50K+ | 4.29 |
| 100K+ | 4.33 |
| 300K+ | 4.38 |
| 1M+ | 4.43 |

Since the p-value is extremely small, we reject the null hypothesis concluding that the difference in mean rating for different categories not the same.

# Smart Q2: Installs vs Reviews

Does Reviews  and Installs have significantly impact the popularity of an app in terms of installs?



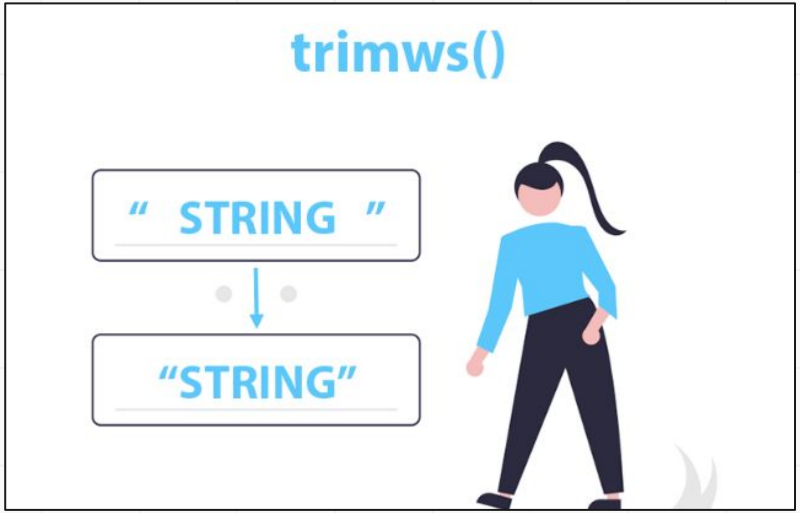
| | Df | Sum Sq | Mean Sq | F Value | Pr(>F) |
|---|---|---|---|---|---|
| Review_Category | 11 | 6.94e+18 | 6.31e+17 | 290 | <2e-16 |
| Residuals | 9647 | 2.10e+19 | 2.17e+15 | | |

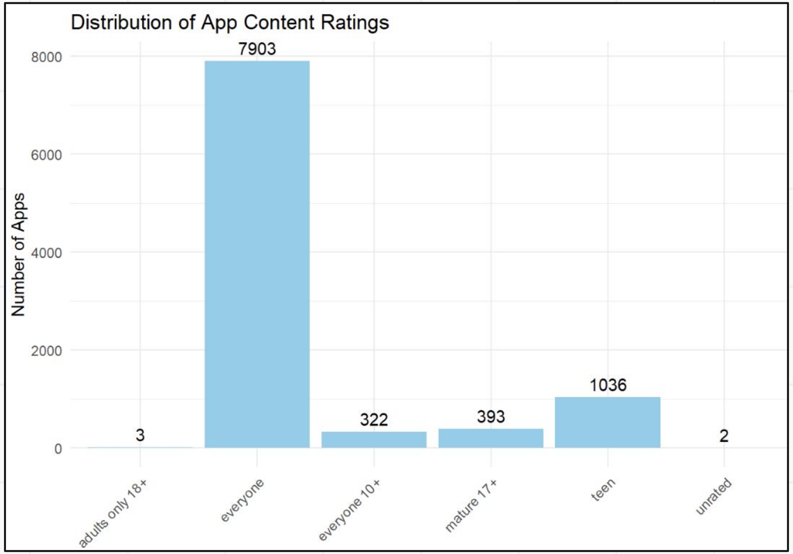
Since the p-value is extremely small, we reject the null hypothesis concluding that the difference in mean Installs for different Review categories not the same.

# Content Rating

Data Cleaning:

- The `trimws()` function in R : To Remove Leading and Trailing spaces.
- Converted to lowerCase



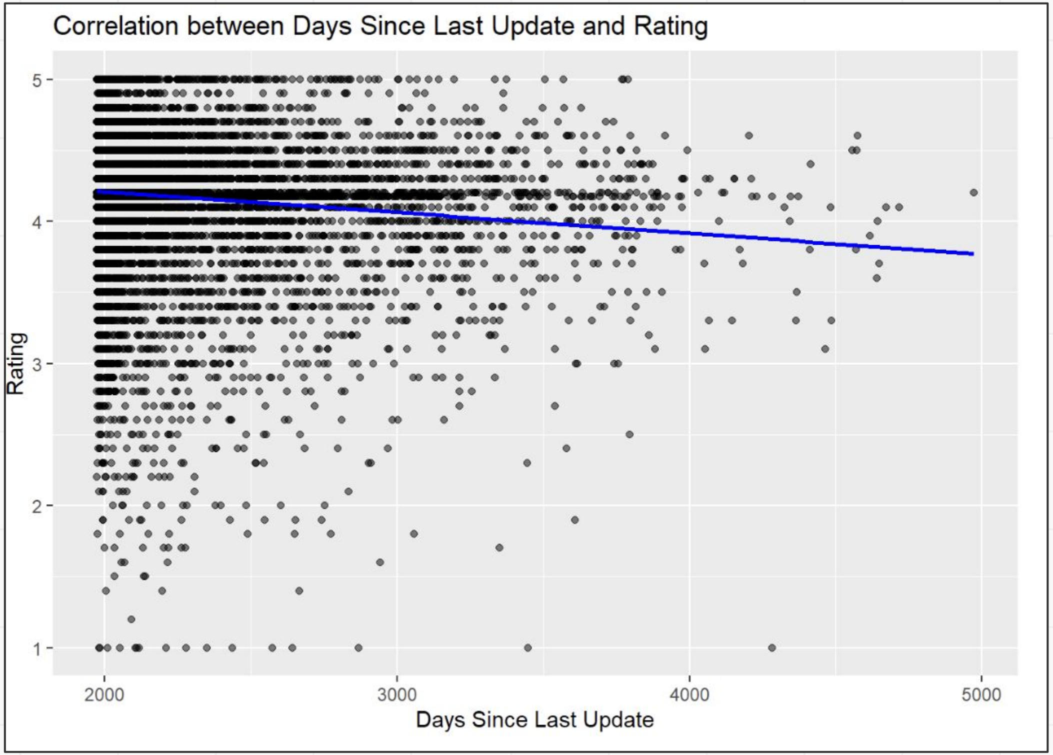


Unique Values:

- Everyone: 7903
- Teen: 1036
- Mature 17+: 393
- Everyone 10+ : 322
- Adults only 18+: 3
- Unrated: 2

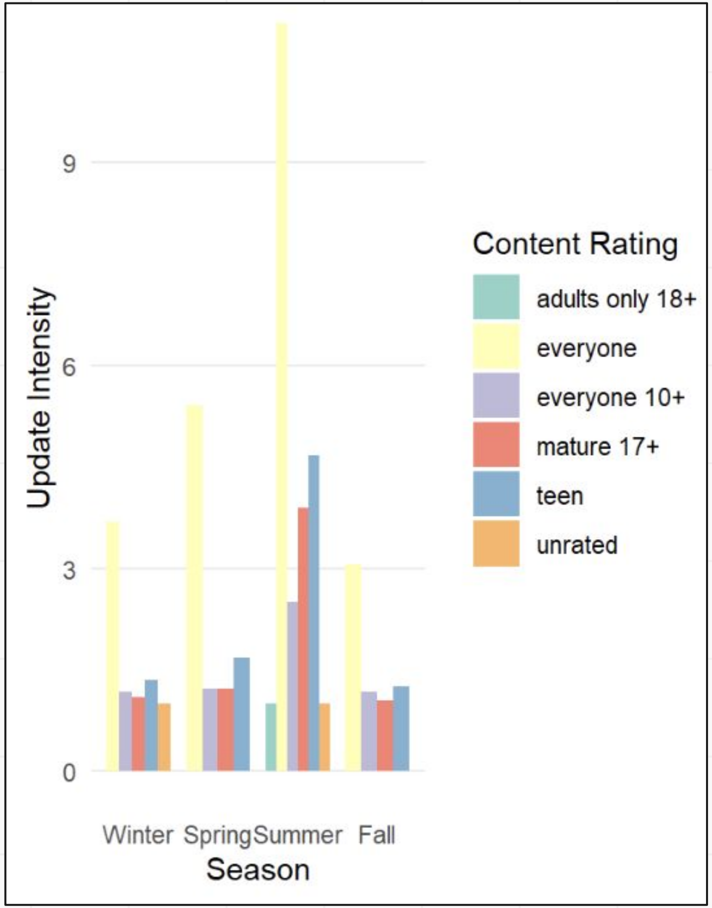
# Smart Q3 : Rating vs Last update

Does Update have significant impact on Rating?



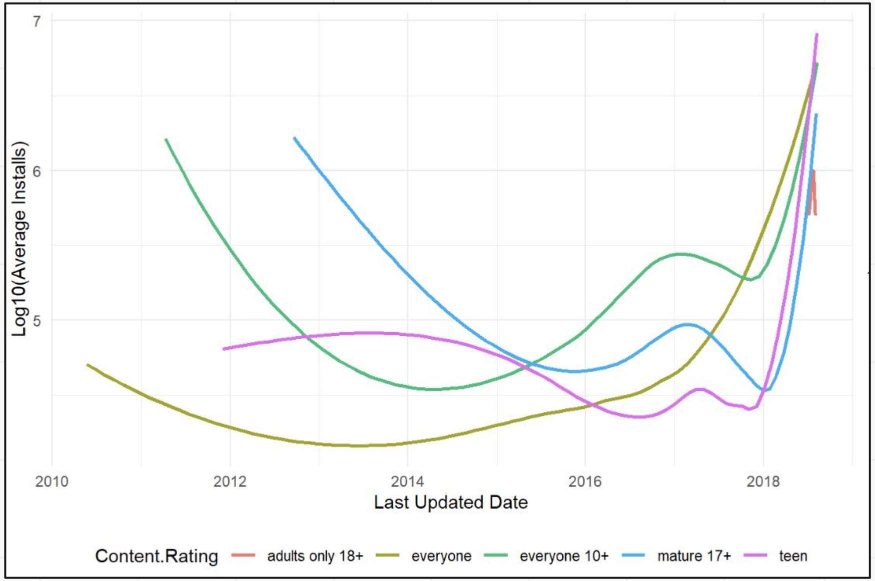
# Update Intensity by Content Rating

- The "Everyone" category peaks in update intensity during summer, while "Teen" and "Everyone 10+" show consistent updates year-round, slightly increasing in summer. "Adults only 18+" and "Unrated" apps have the lowest and most infrequent updates all year.

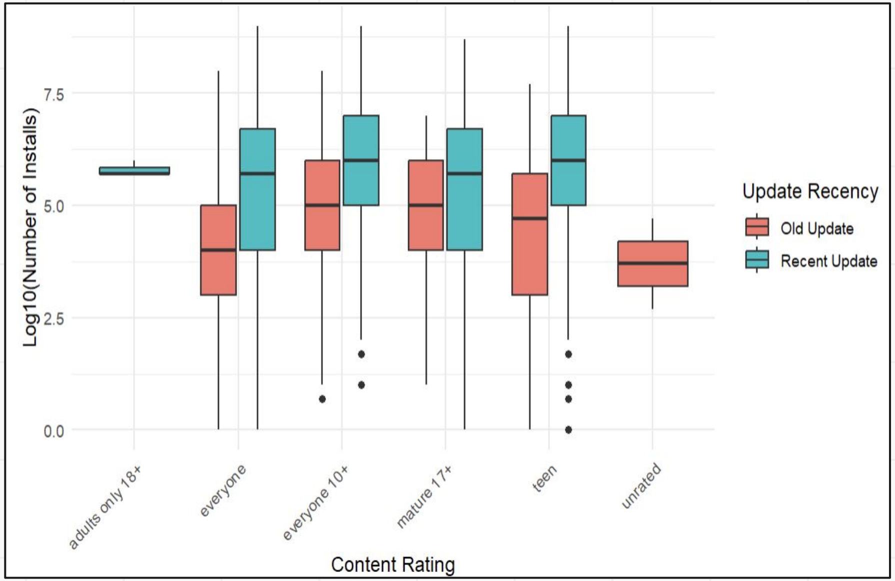


# Smart Q4 : Installs vs Content Rating  vs Last Updated

Does Content Rating and Last Updated  significantly impact the popularity of an app in terms of installs?



trend of average app installs over time for different content ratings



distribution of app installs across different content ratings, segmented by update recency (old vs. recent)

# Statistical Test: ANOVA - test (Last Updated)

**Days Since Last Update by Reviews:**

- **F-value**: 41.95
- **p-value**: 9.82e-11

**Days Since Last Update by Installs:**

- **F-value**: 58.92
- **p-value**: 1.8e-14

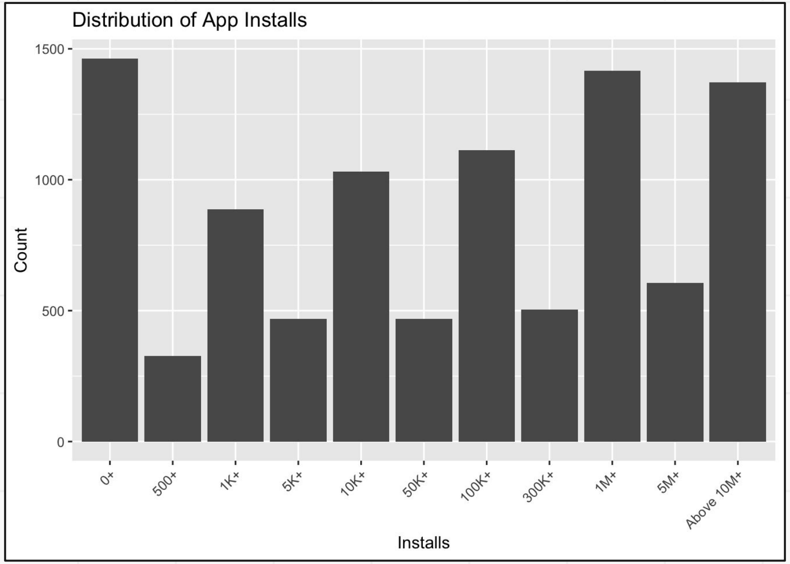
**Days Since Last Update by Ratings:**

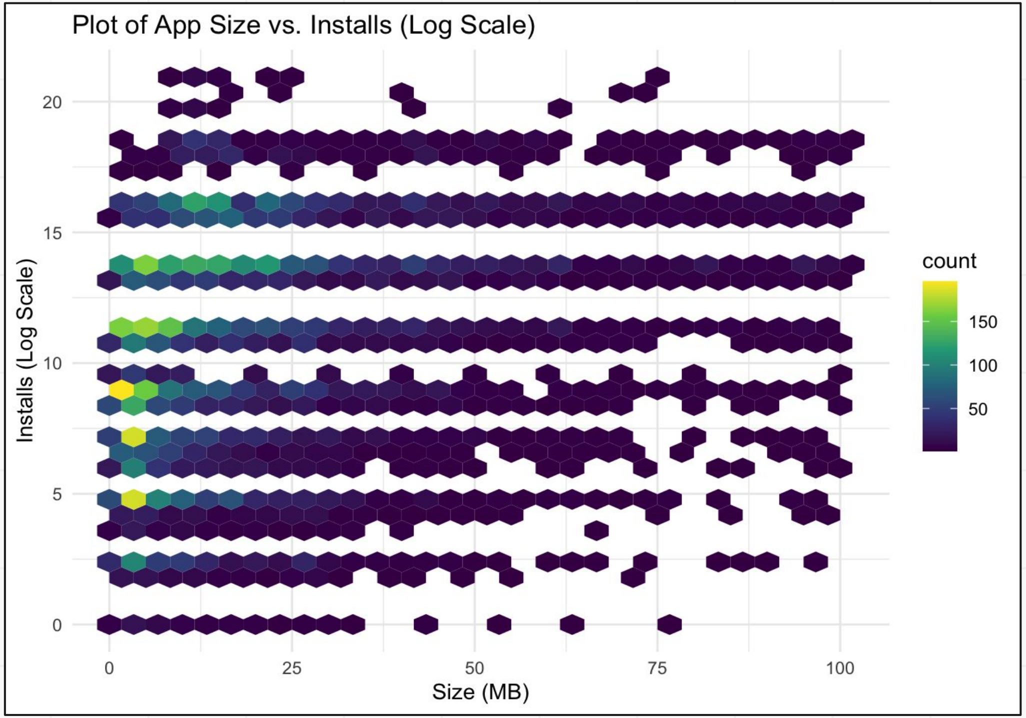
- **F-value**: 143.8
- **p-value**: <2e-16

All three factors—Reviews, Installs, and Ratings- all have very small p - value so they are strongly correlated to Last Updated Factor

# Smart Q5: Size vs Installs

Does app size have significant impact on the number of Installs ?

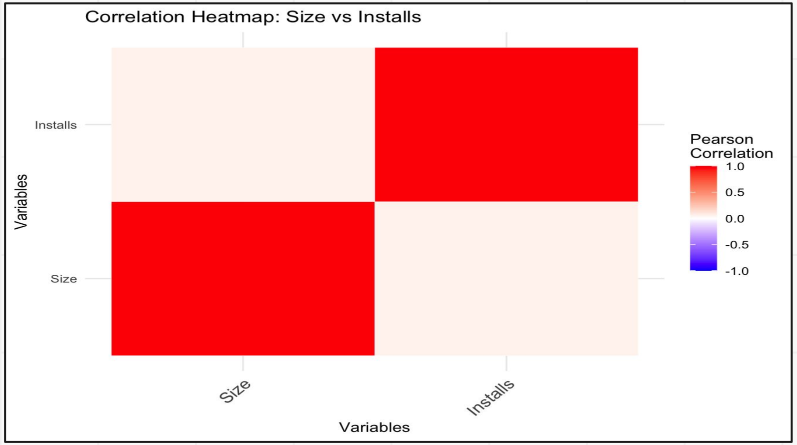




# Statistical Test: correlation coefficient and p-value Install VS Size

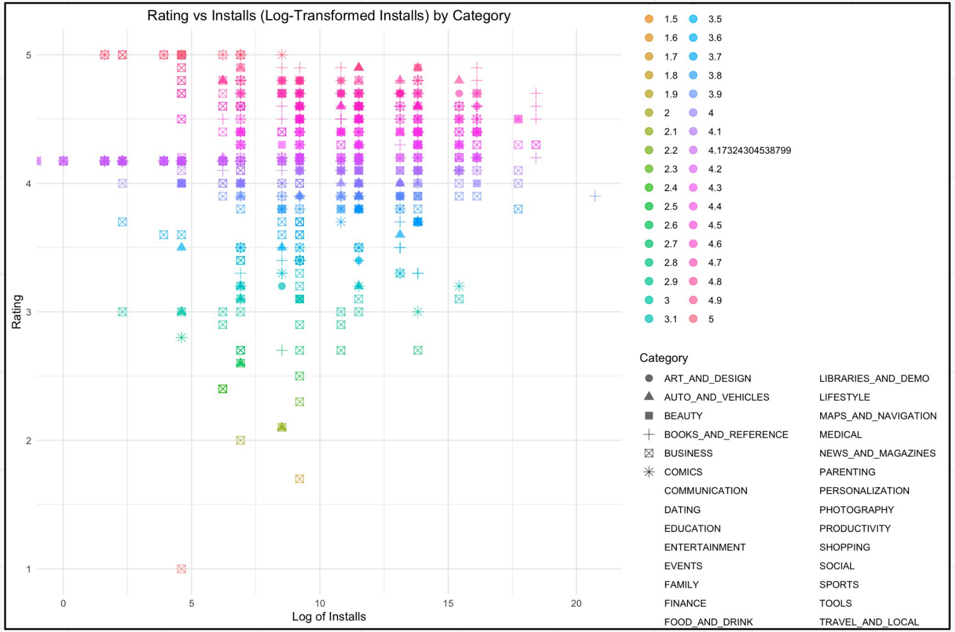
## T-test Result:

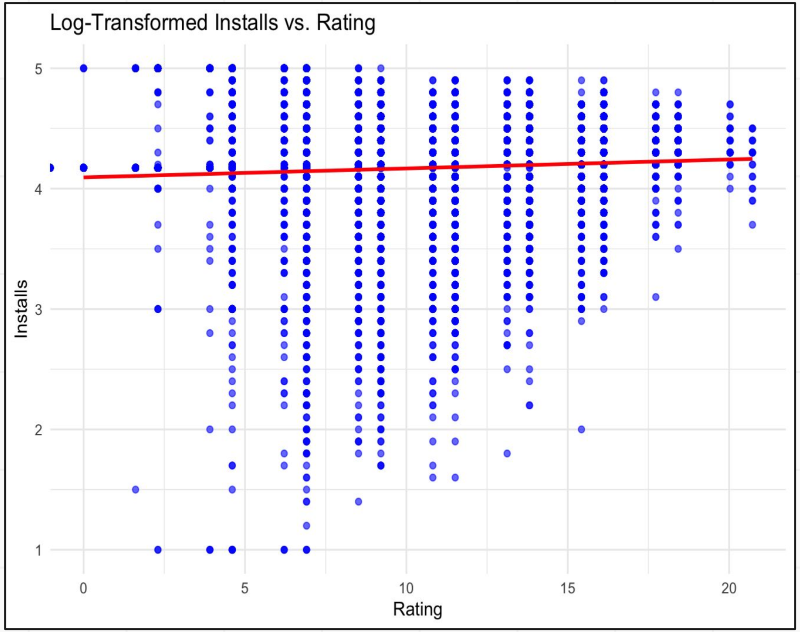
- **Test Statistic (t)**: 4.0069
- **Degrees of Freedom (df)**: 9657
- **P-value**: 6.198e-05
- **Correlation Coefficient**:0.0407
- **95% Confidence Interval**: 0.0208 to 0.0606.



Since the p-value is extremely small, we reject the null hypothesis concluding that the difference in mean log installs for different sizes .

# How do app installs vary by category and rating trends?





# Conclusion

- **Review & Ratings VS Installs**:High install counts are associated with a positive correlation with ratings and reviews.Hence, could conclude that:

  **High Installs -> Positive Rating -> High Reviews**

  **HIGH POPULARITY -> HIGH INSTALLS**

- **High Installs is seen in**

  **Category :** Top 3 categories (Entertainment, Education and Game)

  **Content Rating :** Everyone, Everyone(10+) have highest number of installs

  **Last Updated(+ve Correlation) :** Latest the update, higher the Installs

  **Price(-ve Installs) :** Lesser the price, higher the installs

  **Size :** This is could not be seen as higher size better, as the higher size might also be correlated to categorical apps such as Gaming etc

Share your app idea with us, and we'll estimate its potential installs, reviews and rating.

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

# Thank You!