# Abstract

The data set we have chosen is regarding Google Play Store Apps with various aggregation types including app rating, app function category (e.g. game, business, communication, etc.), number of user reviews, size of app, etc.
https://www.kaggle.com/lava18/google-play-store-apps

**Variables**

- App: application name
- Rating: overall user rating of the app (when scraped)
- Reviews: number of user reviews for the app (when scraped)
- Installs: number of user downloads for the app (when scraped)
- Type: Paid or Free-To-Install (App)
- Genre: genre that the application belongs to

**Summary Statistics**

```
> summary(gp[,c(1, 3, 4, 6, 7, 10)])
                                          App           Rating          Reviews         Installs      Type           Genres
 ROBLOX                                   :    9   Min.   : 1.000   0      :  596   1000000 :1579   0   :    1   Tools        : 842
 CBS Sports App - Scores, News, Stats & Watch Live:    8   1st Qu.: 4.000   1      :  272   10000000:1252   Free:10039   Entertainment: 623
 8 Ball Pool                              :    7   Median : 4.300   2      :  214   100000  :1169   Paid:  800   Education    : 549
 Candy Crush Saga                         :    7   Mean   : 4.193   3      :  175   10000   :1054                Medical      : 463
 Duolingo: Learn Languages Free           :    7   3rd Qu.: 4.500   4      :  137   1000    : 907                Business     : 460
 ESPN                                     :    7   Max.   :19.000   5      :  108   5000000 : 752                Productivity : 424
 (Other)                                  :10796   NA's   :1474     (Other):9339   (Other) :4128                (Other)      :7480
```

The questions we have come up after initial investigation of the data are as follows:

1. Do free-to-install apps get significantly more installs than paid ones on the Google Play Store?
2. Are app ratings correlated with the number of installs on the Google Play Store?
3. Is number of reviews correlated with the number of installs on the Google Play Store?
4. Are the average numbers of installs significantly different across app genres on the Google Play Store?

We believe these questions are relevant for app developers in analyzing their products to increase popularity, visibility, and profitability of their products.

Through one-sided two-sided t-test for question 1, two-sided test for correlation for question 2 and 3, and ANOVA F-test for question 4, we found the following: Free apps are installed more

than paid apps, app rating to installs correlation is weak across apps of all popularities, number of reviews positively correlated to the number of installs per app, and not all app genres are created equally and app developers should consult the most popular genres carefully and ensure that they are not modeling their genres after outliers.

End of abstract.

# Introduction

It is without a doubt that software development promises great potential for success, so it is no wonder that the Google Play Store and the Apple Store are saturated with thousands of different apps that have become an essential part of our lives today. Perhaps the market is oversaturated; for example, there are hundreds of calculator apps to choose from, many of which are no different from the others. This creates one of the greatest challenges modern-day software developers have to face, and it is in the best interest of these creators to figure out strategies that would optimize their success in this highly competitive market.

Although performance and quality are important in making the application stand out amongst others, we must investigate other factors that could potentially draw the users into downloading the product. For example, before publishing a new app, the developer must make several decisions such as whether to make it downloadable for free or at a certain price. Within the program, some developers even implement reminders that encourage users to rate or write reviews in hopes to promote their apps even more. Maybe developers should focus on a specific app genre in order to maximize the number of installs. This then raises several questions that we can investigate:

1. Do free-to-install apps get significantly more installs than paid ones on the Google Play Store?

2. Are app rating scores correlated with the number of installs on the Google Play Store?

3. Is number of reviews correlated with the number of installs on the Google Play Store?

4. Are the average numbers of installs significantly different across app genres on the Google Play Store?

All of the statistical operations and tests were conducted through the R software, analyzing a Kaggle data frame that details a variety of information about all the apps available on the Google Play Store in 2018.

# Background Information and Methods

Significant tests with this 2018 data will be used for predictive analysis for near future market trends. As this data contains apps on the Play Store since its conception, we can effectively conclude that, for all intents and purposes, this data of 10k apps is an accurate representation of all apps on the Play Store, and is sufficient for the questions we will be answering in this report.

The team first modified the original dataset via MS Excel to properly format the elements of the "Installs" column, as the original formatting included commas and "+"s at the end of each element. The team erased all instances of the characters mentioned above.

1. *Do free-to-install apps get significantly more installs than paid ones on the Google Play Store?*

In order to answer this question, we will use the one-sided two-sample t test for the difference in means between two different groups of data: Free and Paid, which are options for the Type column of the dataframe. The test choice is appropriate as we want to evaluate if the population mean for installation counts in free apps is greater than that of paid apps, and the one-sided two-sample t test for the difference in means does just that.

The hypotheses are as follows:

$H_0 : \mu_f = \mu_p$

$H_a : \mu_f > \mu_p$

$\mu_f$ = population mean installs for free apps on the Google Play Store in 2019, $\mu_p$ = population mean installs for paid apps on the Google Play Store in 2019

The following are assumptions for question 1 and their rationales:

1. The samples follow a normal distribution.
    a. Sample sizes are large enough, most definitely greater than 30, so by the central limit theorem, the samples are approximately normal.
2. The samples are independent between and within.
    a. For independence between, we make the assumption that the number of installs does not impact that of the other when comparing free and paid apps, disregarding special cases such as when one type of app fills up the rest of the device storage and prevents further downloads of the other type. We claim independence within by assuming that the sum of Google Play Store apps in the cumulative catalog across time is at least ten times greater than each paid and free app sample.
3. The samples were selected without sampling bias.
    a. Usually whether or not simple random sampling was used or not is the condition that is checked, but refer to the first 2 sentences of this section, Background Information and Methods.

2. *Are app rating scores correlated with the number of installs on the Google Play Store?*

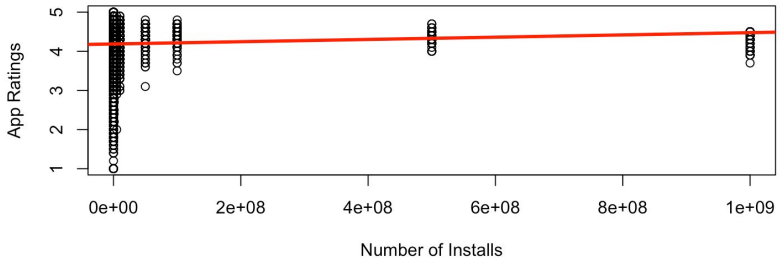
For this question, we will conduct a two-sided test for correlation (Pearson Correlation) in order to visualize the correlation between app ratings and number of installs on the Google Play Store. App Ratings is considered as a continuous variable and Number of Installs is considered as a discrete variable, but since we converted Number of Installs to a numerical, we are able to compare the two and see how they correlate. A two-sided test for correlation is a good way to test this as it tests if app ratings and number of installs are correlated or not (correlation coefficient = 0).

The hypotheses:

$H_0: r = 0$

$H_a: r \neq 0$

*r = the correlation coefficient that measures the strength of a linear relationship between population app ratings and population number of installs.*




```
> x=Installs[ !is.nan(Rating) & Rating < 15]
> y=Rating[ !is.nan(Rating) & Rating < 15]
> plot(x, y, xlab = 'Number of Installs', ylab = 'App Ratings')
> mod=lm(y~x, data=ds)
> abline(mod, col="red", lwd=3)
```

The following are the assumptions of question 2:

1. Level of measurements
    a. Both variables are continuous, so a Pearson correlation works here (number of Installs was modified to fit this requirement, as discussed above).
2. Linearity
    a. Data seems to have a “straight-line” relationship, so the condition is met.
3. Related pairs
    a. Each app in the data has an “install number” and rating (1 sample that does not meet this requirement was removed, discussed in Results).
4. Absence of outliers

a. No visible outliers after filtering the data, so the condition is met.

3. *Is number of reviews correlated with the number of installs on the Google Play Store?*

This question is very similar to question 2 with a different variable. We will also conduct a two-sided test for correlation between the number of Installs and the number of Reviews. Number of Reviews is considered as a continuous variable and number of Installs is considered as a discrete variable, but since we converted Number of Installs to a numerical, we are able to compare the two and see how they correlate. A two-sided test for correlation is a good way to test this as it tests if the number of reviews of apps and number of installs of apps are correlated (correlation coefficient = 0).

The hypotheses:
$H_0: r = 0$
$H_a: r \neq 0$
*r = the correlation coefficient that measures the strength of a linear relationship between population number of reviews and population number of installs.*

The following are the assumptions of question 3:

1. Level of measurements
    a. Both variables are continuous, so a Pearson correlation works here (number of Installs was modified to fit this requirement, as discussed above).
2. Linearity
    a. Data seems to have a "straight-line" relationship, so the condition is met.
3. Related pairs
    a. Each app in the data has an "install number" and number of reviews (1 sample that does not meet this requirement was removed, discussed in Results).
4. Absence of outliers
    b. No visible outliers after filtering the data, so the condition is met.

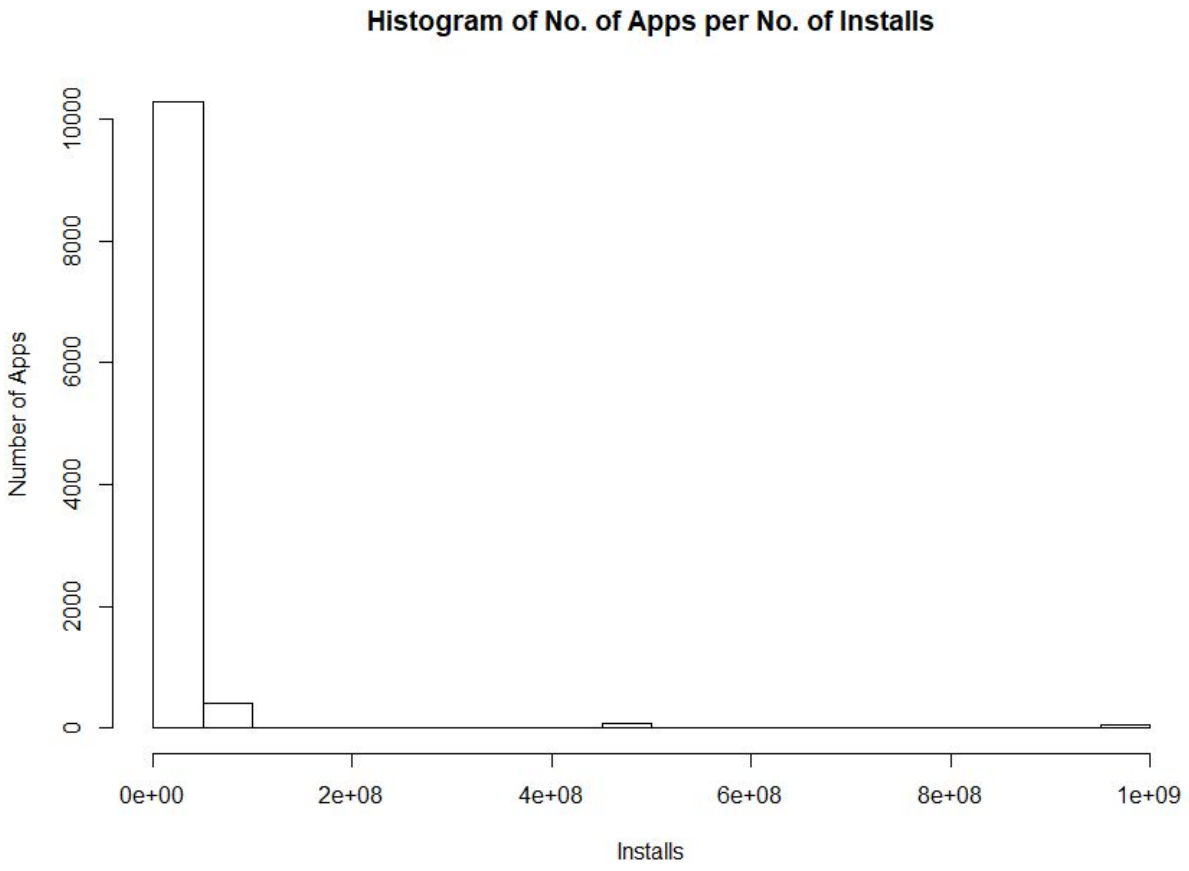
**Note:** The team was aware that, because the popularity of apps (i.e. number of installs for all apps) followed the Pareto principle, most apps received 0 to insignificant number of installs while a select few number of apps occupied most of the market (refer to figure below). This means that the sample size for the number of reviews of a rarely-installed app would potentially be too small and may potentially disrupt our tests for questions 2 and 3, since every app regardless of popularity has equal weight in our model. Taking this issue into consideration, the team decided to apply questions 2 and 3 into three samples:

1. The original dataset with all numbers of installs (n = 10840)
2. Dataset that only contained elements with 1,000+ installs (n = 9037)
3. Dataset that only contained elements with 1,000,000+ installs (n = 2081)

The team determined these cutoff points for the three respective samples for the following reasons:

1. To test the original data without any manipulation as control
2. To exclude all data that the team considered too unpopular (too few installs) to be significant
3. To test only highly popular apps.

---



4. *Are the average numbers of installs significantly different across app genres?*

In order to answer this question, we will use ANOVA comparing the average number of installations across different app genres present in the data set. Like in previous questions, Installs was converted to a numeric type beforehand. An ANOVA F-test was deemed appropriate by the team as we wanted to see if there were any statistically significant differences in the average number of app installs between the individual, independent app genres. This test would allow us to conclude that app genre does matter when it comes to gaining the highest chance of success in the market.

The hypotheses are as follows:

**Null Hypothesis:** $\mu_1 = \mu_2 = \cdots = \mu_i$

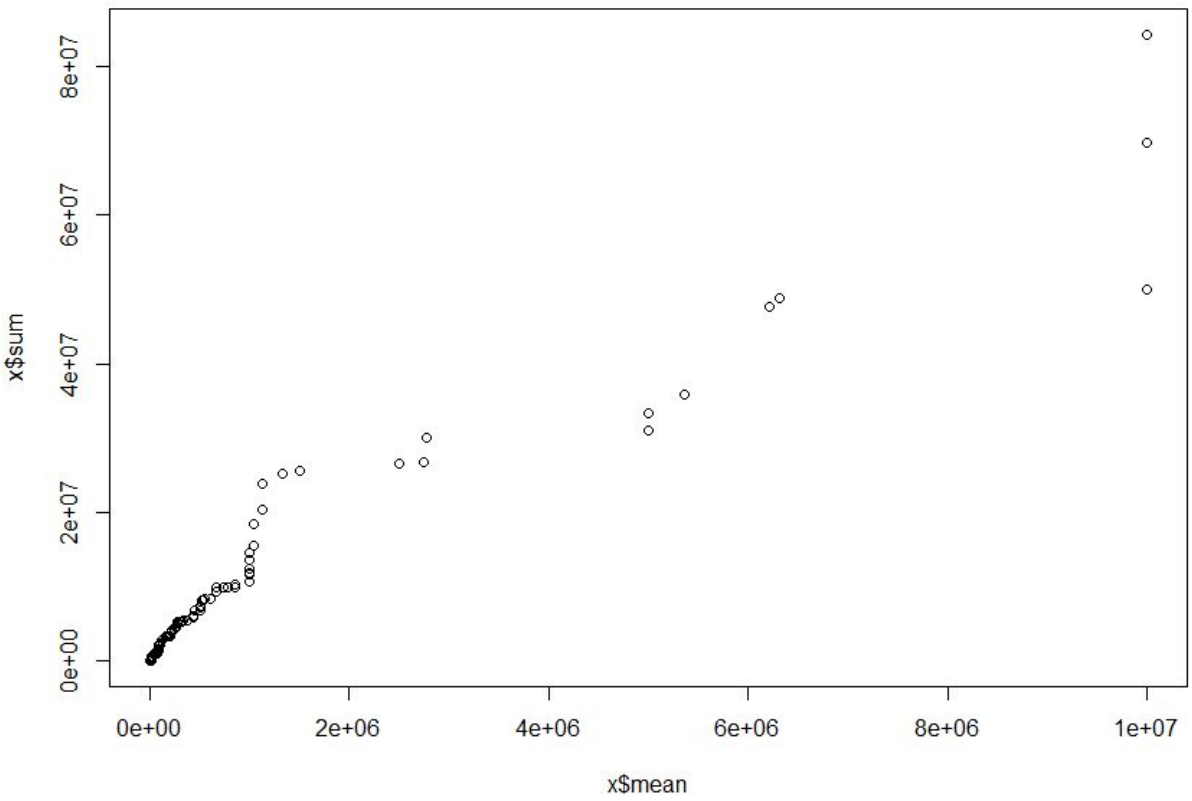
**Alternative Hypothesis:** At least one $\mu_i$ differs

$\mu_i$ = population mean number of installations of all unique app genres.

The following are assumptions for question 4 and their rationales:

1. Randomness and independence of samples
    a. Usually whether or not simple random sampling was used or not is the condition that is checked, but refer to the first 2 sentences of this section, Background Information and Methods.
2. Equal variances
    a. Condition is met since all groups have roughly equal variance.
3. Normality
    a. We tested the normality of the average number of installations per genre by dividing the total number of installs per genre by the total count of apps per genre. The resulting values are plugged into a qq plot:

```
> meanInstalls = aggregate(mydata10841[,6], by = list(mydata10841$Genres), FUN = mean)
> x = as.data.frame(table(mydata10841$Genres))
> y = meanInstalls[order(meanInstalls$Group.1, decreasing = FALSE),]
> x[,"mean"] = y[,2] / x[,2]
> x[,"sum"] = y[,2]
> qqplot(x$mean, x$sum)
```




The team is aware the qq plot does not look like a traditional linear line due to the number of installs per genre following the Pareto distribution and thus having fewer samples on the upper right quadrant. Thus, the abundance in outliers may provide inaccuracies. However, the team considers this plot still roughly linear and thus passed the normality test, keeping in mind the likelihood of inaccuracies in results.

# Results

The R code for each research question is shown below.

```
> ds[10473,]
                                                 App Category Rating
10473 Life Made WI-Fi Touchscreen Photo Frame      1.9     19
      Reviews   Size Installs Type    Price Content.Rating
10473    3.0M 1,000+     Free    0 Everyone
                Genres Last.Updated Current.Ver Android.Ver
10473 February 11, 2018       1.0.19  4.0 and up
> ds = ds[-10473,]
```

The team ran this code before performing the tests. This deletes an element with bad data; values in the Installs column of the dataframe should be numeric data, but instead is “Free” and makes no sense. Values in the Reviews column is numeric, but instead is “3.0M” which does not follow the format of others, while converting the entry to 3000000 leaves it at a disparate outlier and thus raises suspicion to its authenticity. Such inconsistencies in the data convinced the team to just remove this element.

Note, also, that the team used “ds”, “gp”, and “mydata” interchangeably as variables all referring to the original dataset.

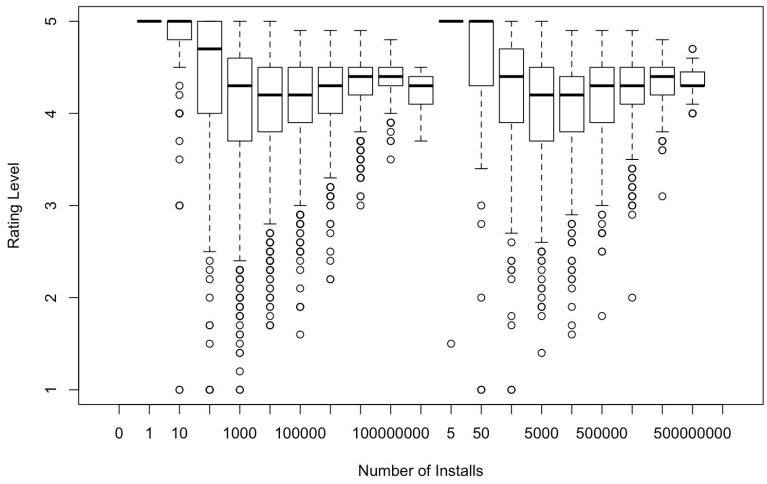
1. Do free-to-install apps get significantly more installs than paid ones on the Google Play Store?

```
> #Do free apps get significantly more downloads than paid apps (two sample t-test diff in means)
> gp = read.csv(file='~/Desktop/googleplaystore-1.csv')
>
> freedownload = as.numeric(as.character(gp[gp$Type == 'Free',]$Installs))
> paiddownload = as.numeric(as.character(gp[gp$Type == 'Paid',]$Installs))
>
> t.test(freedownload, paiddownload, alternative = "greater", var.equal = FALSE)

        Welch Two Sample t-test

data:  freedownload and paiddownload
t = 18.842, df = 10051, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 15150537      Inf
sample estimates:
 mean of x  mean of y
16690953.0    91195.1
```

- **Null Hypothesis:** There is no significant difference in number of installs between paid and unpaid apps.
- **Alternative Hypothesis:** Number of installs of unpaid apps is significantly greater than that of paid apps.
- **Conclusion:** p-value = 2e-16 < alpha = 0.05 → reject the null hypothesis. We have sufficient evidence to conclude that free-to-install apps do get more installs than paid ones on the Google Play Store.

2. Are app rating scores correlated with downloads? **(two-sided test for correlation between paired samples)**

- **Null Hypothesis:** There is no correlation between the Ratings and number of Installs.
- **Alternative Hypothesis:** There is a correlation between the Ratings and the number of Installs.



```
> boxplot(ds$Rating~ds$Installs, xlab='Number of Installs', ylab='Rating Level')
> ds$Installs = as.numeric(as.character(ds$Installs))
> cor.test(ds$Rating, ds$Installs, alternative='two.sided', conf.level = 0.90)

        Pearson's product-moment correlation

data:  ds$Rating and ds$Installs
t = 4.976, df = 9364, p-value = 6.606e-07
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.03438736 0.06829218
sample estimates:
       cor
0.05135457
```

**Conclusion:** p-value = 2e-16 < alpha = 0.05 → reject the null hypothesis. Since the correlation coefficient is 0.05, there is a weak correlation between number of Installs and Ratings when considering all apps on the Google Play store. We think this is due to the large number of apps with few or without any installs.

(Apps with 1000 or more Installs)

```
> newInstalls=Installs[!is.nan(Rating) & as.numeric(Installs)>=1000]
> newRatings=Rating[!is.nan(Rating) & as.numeric(Installs)>=1000]
> cor.test(newRatings, newInstalls, alternative = 'two.sided', conf.level = 0.90)

        Pearson's product-moment correlation

data:  newRatings and newInstalls
t = 5.8299, df = 8717, p-value = 5.743e-09
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.04475343 0.07984996
sample estimates:
       cor
0.06232096
```

**Conclusion:** p-value = 5.74e-09 < alpha = 0.05 → reject the null hypothesis.

(Apps with 1,000,000 or more Installs)

```
> newInstalls=Installs[!is.nan(Rating) & as.numeric(Installs)>=1000000]
> newRatings=Rating[!is.nan(Rating) & as.numeric(Installs)>=1000000]
> cor.test(newRatings, newInstalls, alternative = 'two.sided', conf.level = 0.90)

        Pearson's product-moment correlation

data:  newRatings and newInstalls
t = 3.0057, df = 4407, p-value = 0.002665
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.02047805 0.06992703
sample estimates:
       cor
0.04523025
```

**Conclusion:** p-value = 0.002665 < alpha = 0.05 → reject the null hypothesis.
**Final Conclusion:** The correlation between App Ratings and Installs is about 0.05 regardless of app popularity.

3. Are numbers of reviews correlated with downloads? **(two-sided test for correlation between paired samples)**
   **Null Hypothesis:** There is no evidence for significant correlation between number of installs and ratings.
   **Alternative Hypothesis:** There is evidence for significant correlation between number of installs and ratings.

The team performed this test for the dataset with all elements:

```
> Installs = as.numeric(as.character(mydata$Installs))
> reviews = as.numeric(as.character(mydata$Reviews))
```

```
> cor.test(Installs,reviews, alternative = "two.sided")

        Pearson's product-moment correlation

data:  Installs and reviews
t = 87.433, df = 10838, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6319477 0.6540291
sample estimates:
      cor
0.6431221
```

**Conclusion:** p < 0.05, reject H0. Since the value is 0.64, there is a moderate correlation between number of installs and the number of reviews.

The team performed this test for the dataset with elements with 1,000+ installs:

```
> cor.test(mydata1000$Installs, mydata1000$Reviews, alternative = "two.sided")

        Pearson's product-moment correlation

data:  mydata1000$Installs and mydata1000$Reviews
t = 79.42, df = 9034, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6288997 0.6531874
sample estimates:
      cor
0.6412041
```

**Conclusion:** p < 0.05, reject H0. Since the value is 0.64, there is a moderate correlation between number of installs and the number of reviews.

The team performed this test for the dataset with 1,000,000+ installs:

```
> cor.test(mydata1M$Installs, mydata1M$Reviews, alternative = "two.sided")
```

```
        Pearson's product-moment correlation

data:  mydata1M$Installs and mydata1M$Reviews
t = 34.319, df = 2078, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5733045 0.6282039
sample estimates:
      cor
0.6014638
```

**Conclusion:** $p < 0.05$, reject H0. Since the value is 0.60, there is a moderate correlation between number of installs and the number of reviews.

**Final conclusion:** The correlation value of number of installs and number of reviews regardless of app popularity is about 0.6.

4. Are the average numbers of installs significantly different across app genres?

   **Null Hypothesis:** $\mu 1 = \mu 2 = \cdots = \mu i$ ($\mu i$ = mean installs for different app genres)
   **Alternative Hypothesis:** At least one $\mu i$ differs
   $\mu i$ = population mean number of installations of app genre i.

```
> ds$Installs = as.numeric(as.character(ds$Installs))
> summary(aov(lm(ds$Installs~as.factor(ds$Genres))))
                       Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(ds$Genres)   118 3.992e+18 3.383e+16   4.877 <2e-16 ***
Residuals            10720 7.437e+19 6.937e+15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

- **Conclusion:** $p < 0.05$, reject the null hypothesis. We have statistically significant evidence that the average number of installs significantly differ across app genres.

Taking into account the large number of outliers in this experiment, this conclusion should be accepted with trepidation.

If we wish to see the genres with the highest average installs, a significant portion of them are outliers:

```
> x[order(x$mean, decreasing = TRUE),]
                                    Var1 Freq          mean         sum
71                  Lifestyle;Pretend Play    1 1.000000e+07 10000000.0
95                Role Playing;Brain Games    1 1.000000e+07 10000000.0
111                         Tools;Education    1 1.000000e+07 10000000.0
49          Educational;Action & Adventure    4 6.315625e+06 25262500.0
86               Puzzle;Action & Adventure    5 6.204000e+06 31020000.0
5              Adventure;Action & Adventure   13 5.361834e+06 69703846.2
7                       Adventure;Education    2 5.000000e+06 10000000.0
34                    Casual;Music & Video    2 5.000000e+06 10000000.0
118 Video Players & Editors;Music & Video    3 2.777778e+06  8333333.3
26                 Card;Action & Adventure    2 2.750000e+06  5500000.0
117     Video Players & Editors;Creativity    2 2.500000e+06  5000000.0
60               Entertainment;Pretend Play    2 1.500000e+06  3000000.0
57                 Entertainment;Creativity    3 1.333333e+06  4000000.0
30                Casual;Action & Adventure   21 1.133789e+06 23809571.4
76                      Music;Music & Video    3 1.122333e+06  3367000.0
94          Role Playing;Action & Adventure    7 1.042857e+06  7300000.0
43                    Education;Brain Games    5 1.040040e+06  5200200.0
10                      Arcade;Pretend Play    1 1.000000e+06  1000000.0
58                  Entertainment;Education    1 1.000000e+06  1000000.0
65     Health & Fitness;Action & Adventure    1 1.000000e+06  1000000.0
79                    Parenting;Brain Games    1 1.000000e+06  1000000.0
92                      Racing;Pretend Play    1 1.000000e+06  1000000.0
108                     Strategy;Creativity    1 1.000000e+06  1000000.0
32                        Casual;Creativity    7 8.571429e+05  6000000.0
97                Role Playing;Pretend Play    5 8.480000e+05  4240000.0
55        Entertainment;Action & Adventure    3 7.777778e+05  2333333.3
50                  Educational;Brain Games    6 7.388889e+05  4433333.3
105               Sports;Action & Adventure    4 6.687500e+05  2675000.0
18                 Board;Action & Adventure    3 6.672222e+05  2001666.7
74                                    Music   22 6.614060e+05 14550931.8
100          Simulation;Action & Adventure   11 6.115702e+05  6727272.7
53                Educational;Pretend Play   19 5.461637e+05 10377110.5
91                Racing;Action & Adventure   20 5.306500e+05 10613000.0
```

Only in the lower sections do we begin seeing more comfortable numbers of apps in each genre.

| | | | | | |
|---|---|---|---|---|---|
| 3 | Action;Action & Adventure | 17 | 3.045692e+05 | 5177676.5 |
| 51 | Educational;Creativity | 5 | 2.802200e+05 | 1401100.0 |
| 107 | Strategy;Action & Adventure | 2 | 2.750000e+05 | 550000.0 |
| 29 | Casual | 193 | 2.594118e+05 | 50066480.5 |
| 9 | Arcade;Action & Adventure | 16 | 2.547266e+05 | 4075625.1 |
| 14 | Art & Design;Pretend Play | 2 | 2.500000e+05 | 500000.0 |
| 33 | Casual;Education | 3 | 2.233333e+05 | 670000.0 |
| 8 | Arcade | 220 | 2.216349e+05 | 48759678.0 |
| 35 | Casual;Pretend Play | 31 | 2.199386e+05 | 6818096.8 |
| 38 | Communication | 387 | 2.179842e+05 | 84359887.0 |
| 90 | Racing | 98 | 2.082723e+05 | 20410686.9 |
| 116 | Video Players & Editors | 173 | 2.073909e+05 | 35878628.4 |
| 81 | Parenting;Music & Video | 6 | 1.863889e+05 | 1118333.3 |
| 106 | Strategy | 107 | 1.724332e+05 | 18450351.4 |
| 103 | Social | 295 | 1.616762e+05 | 47694467.5 |
| 31 | Casual;Brain Games | 13 | 1.604201e+05 | 2085461.5 |
| 102 | Simulation;Pretend Play | 4 | 1.318750e+05 | 527500.0 |
| 42 | Education;Action & Adventure | 6 | 1.116667e+05 | 670000.0 |
| 112 | Travel & Local | 257 | 1.039953e+05 | 26726798.2 |
| 6 | Adventure;Brain Games | 1 | 1.000000e+05 | 100000.0 |
| 27 | Card;Brain Games | 1 | 1.000000e+05 | 100000.0 |
| 66 | Health & Fitness;Education | 1 | 1.000000e+05 | 100000.0 |
| 70 | Lifestyle;Education | 1 | 1.000000e+05 | 100000.0 |
| 89 | Puzzle;Education | 1 | 1.000000e+05 | 100000.0 |
| 113 | Travel & Local;Action & Adventure | 1 | 1.000000e+05 | 100000.0 |
| 77 | News & Magazines | 283 | 9.359984e+04 | 26488755.3 |
| 83 | Photography | 335 | 8.989305e+04 | 30114172.1 |
| 114 | Trivia | 38 | 8.906258e+04 | 3384378.2 |
| 17 | Board | 44 | 8.632604e+04 | 3798345.7 |
| 28 | Casino | 39 | 8.564149e+04 | 3340017.9 |
| 85 | Puzzle | 140 | 8.512637e+04 | 11917691.5 |
| 84 | Productivity | 424 | 7.885419e+04 | 33434177.8 |
| 4 | Adventure | 75 | 7.230139e+04 | 5422604.3 |
| 25 | Card | 48 | 7.067105e+04 | 3392210.4 |
| 2 | Action | 365 | 7.012227e+04 | 25594627.9 |
| 101 | Simulation;Education | 3 | 6.672222e+04 | 200166.7 |
| 80 | Parenting;Education | 7 | 6.469388e+04 | 452857.1 |
| 119 | Weather | 82 | 6.337010e+04 | 5196347.8 |
| 12 | Art & Design;Action & Adventure | 2 | 5.000000e+04 | 100000.0 |
| 37 | Comics;Creativity | 1 | 5.000000e+04 | 50000.0 |
| 98 | Shopping | 260 | 4.804510e+04 | 12491726.1 |
| 13 | Art & Design;Creativity | 7 | 4.510204e+04 | 315714.3 |
| 72 | Maps & Navigation | 137 | 3.858926e+04 | 5286729.1 |
| 52 | Educational;Education | 41 | 3.742478e+04 | 1534416.1 |
| 11 | Art & Design | 58 | 3.642334e+04 | 2112553.4 |
| 21 | Books & Reference | 231 | 3.600887e+04 | 8318050.1 |
| 19 | Board;Brain Games | 15 | 3.503156e+04 | 525473.3 |
| 47 | Education;Pretend Play | 23 | 3.249149e+04 | 747304.3 |
| 93 | Role Playing | 109 | 2.964739e+04 | 3231565.8 |
| 67 | House & Home | 88 | 2.178622e+04 | 1917187.1 |
| 99 | Simulation | 200 | 2.036495e+04 | 4072989.1 |
| 104 | Sports | 398 | 1.810039e+04 | 7203956.0 |

# Overall Conclusion

From the factors that we analyzed and the results accrued from the Google Play Store dataset, we can conclude that free-to-install apps generate more installs than apps that require a paywall to install. However, this does not necessarily imply that free-to-install apps generate more revenue, as there can be microtransactions in each app.

From the findings in the dataset, app rating to app installs correlation is weak across apps of all popularities. We may presume that quality and user-experience for an app does not seem to be related to app popularity, or that app ratings may not be accurate reflections of user satisfaction.

Number of reviews positively correlate to the number of installs per app, regardless of the popularity of the app. Note that these are not evidence of causation. The team cannot advise whether marketing for higher number of reviews would result in more installs or if high number of ratings of any level is just an effect of more installs, or if both statements are true.

Genres are not all created equally, as they follow a Pareto distribution in average installs of apps in each genre. An app developer should consult the most popular genres carefully and ensure that they are not modeling their genres after outliers.

Thus, from the data the team has gathered, we advise that app developers push for free-to-install apps, encourage user reviews regardless of rating scores, and keep in mind certain app genres are more successful in terms of getting installations.

# References

Gupta, Lavanya. (2018; December). Google Play Store Apps, Version 6. Retrieved November 18, 2019 from https://www.kaggle.com/lava18/google-play-store-apps