SPEARS
SCHOOL OF BUSINESS

Database Marketing Project

Team 5

Ahmed Sodeinde (A20187045)
Ahmadreza Homayouni (A20160011)
Subhash Daggubati (A20090700)
Suhao Chen (A20164586)

Introduction of Google Play Dataset

Google Play is a digital distribution service operated and developed by Google Inc. It serves as the official app store for the Android operating system, allowing users to browse and download applications developed with the Android software development kit (SDK) and published through Google. Google Play also serves as a digital media store, offering music, books, movies, and television programs. *(Wikipedia)*

The store has over 2.6 million apps with a wide array of different applications ranging from kid's games to language translation apps. These apps are placed in different categories based on minimum age required, intended usage (*health and fitness, educational, shopping etc.*), cost *(free or paid) etc.*

Data analysis for these applications will attempt to answer four pertinent questions from the sample dataset with 10,839 Google play apps. The dataset obtained from *Kaggle* contains the app names, categories, rating scores, reviews, number of installations, genres, size, type, etc.

Overview of our research

Target audience: Android application developers

Research objectives: Explore the relationship of application rating scores with numbers of reviews, numbers of installation, type(free/paid) etc.

Research approach: Use descriptive, inferential statistics & analytical techniques including distributions, frequencies, visualizations, hypothesis tests, etc.

Data dictionary

Name	Label	Type	Values
App	App name	Character	application names
Category	App category	Character	Art and design, education, sports etc.
Rating	App rating score	Numeric	1.0-5.0
Reviews	Number of app reviews	Numeric	Numbers
Size	Size of an app	Character	Number of m-bytes or varies with device
Installs	Installation category of an app	Character	0+, 100+, 100M+, etc.
Type	Type of an app	Character	free/paid
Price	Price of an app	Numeric	$0.00 etc.
Content Rating	Content rating of app	Character	Everyone/teen/mature 17+ etc.
Genres	Genres of the app	Character	Art design, auto vehicles, etc.
Last Updated	Date the app last updated	Numeric	20JUL2018 etc.
Current Version	Current version of the app	Character	2.2.6.2 etc.
Android Version	Android version need to run the app	Character	4.1 and up etc.
Installs_Numeric	Numeric of Installs	Numeric	1000, 500,000 etc.
Installs_Categories_N	Re-categorized Installs	Character	Below 1k, 1k-10k, etc.

Data cleaning

Challenges:

- There is one row with completely irrelevant data.
- There is one row with Type = NaN
- There are more than thousand records in the dataset with Ratings having no value (missing value).

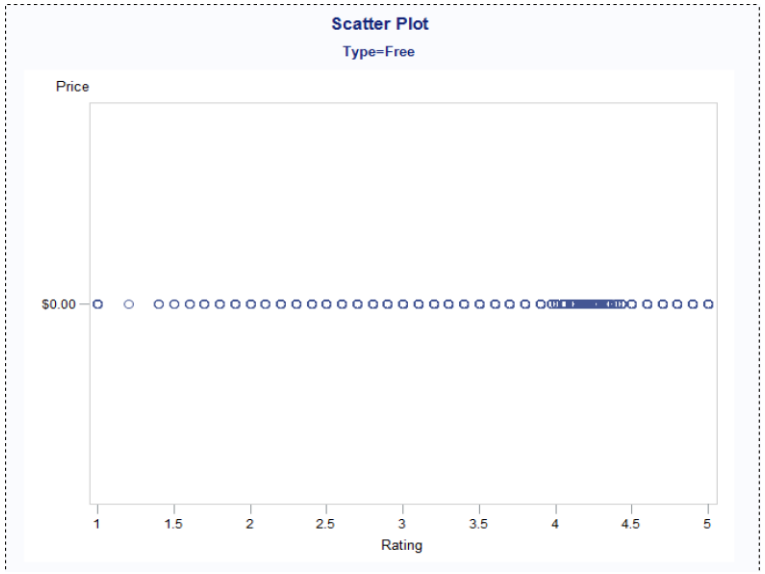
Steps:

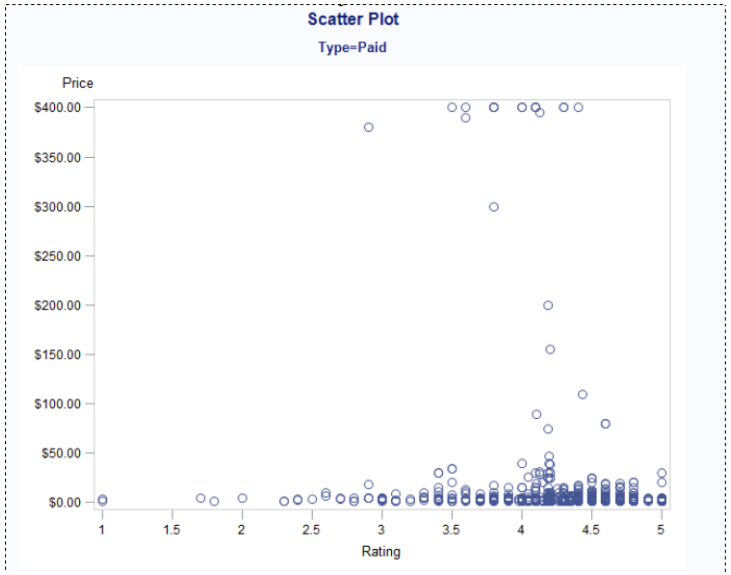
Step 1: Deleted the row with irrelevant data.

Step 2: Deleted the row with Type = NaN as deleting single record will not impact us much.

Step 3: As we have more than thousand records with no value for Rating, these rows are imputed with the mean rating of the corresponding app category rather than deleting the rows.

Question 1: Do paid apps have higher rating scores compared to free ones?





In the free apps scatter plot, there appears to be a cluster along the 3.8 to 4.5 rating score indicating a higher rating for $*0.00* apps (free apps), which is also evident in the paid apps scatter plot.

One-Way Analysis of Variance

Results

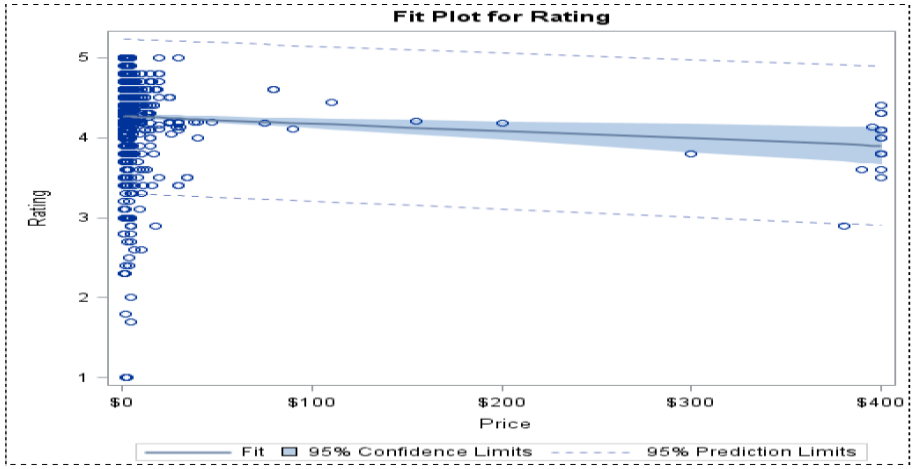
The ANOVA Procedure

Dependent Variable: Rating

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.719737	1.859869	8.08	0.0003
Error	10837	2494.670726	0.230199		
Corrected Total	10839	2498.390464			

R-Square	Coeff Var	Root MSE	Rating Mean
0.001489	11.45122	0.479791	4.189868

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Type	2	3.71973739	1.85986870	8.08	0.0003



Hypothesis Test:

$$\begin{cases} H_0: Average\ rating\ of\ the\ paid\ applications\ is\ equal\ to\ the\ average\ rating\ of\ free\ applications. \\ H_a: Average\ rating\ of\ paid\ applications\ and\ free\ applications\ are\ not\ equal. \end{cases}$$

P-value for the ANOVA test is 0.0003 which is less than the α value 0.05. Hence, we reject Null Hypothesis (H0)
This indicates there is a significant difference between average rating for paid apps and average rating for free apps.

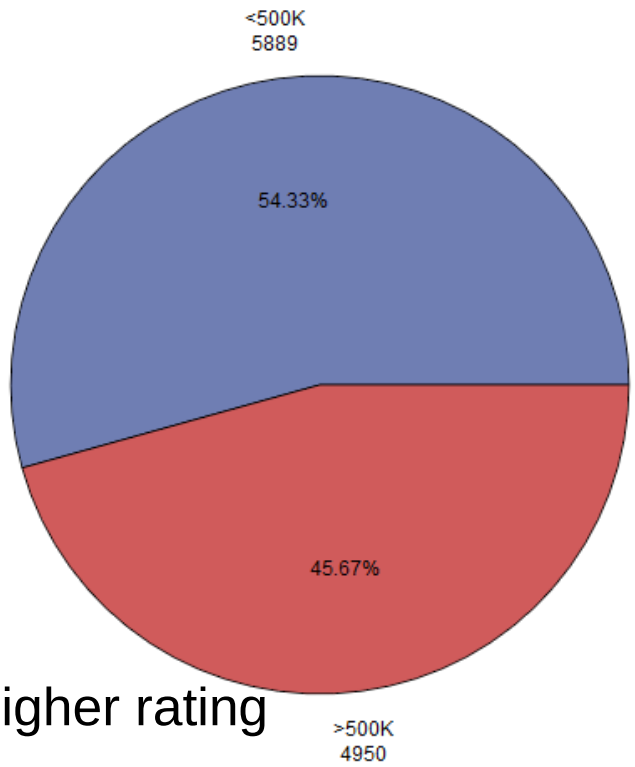
Note: We don't have equal number of data rows for free apps and paid apps so the results of this ANOVA might not be effective.

Conclusion:

The paid apps don't have higher rating scores compared to free ones as we can see from the fit plot slope of the line is negative between the rating and price.

Question 2: Do more than 500K installed applications have higher rating scores compared to less than 500K installed ones?

Installs_Categories_N	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Below 1K	1803	16.63	1803	16.63
1K-10K	1384	12.77	3187	29.4
10K-50K	1054	9.72	4241	39.12
50K-100K	479	4.42	4720	43.54
100K-500K	1169	10.79	5889	54.33
500K-1M	539	4.97	6428	59.3
1M-5M	1637	15.1	8065	74.4
5M-10M	752	6.94	8817	81.34
10M-50M	1661	15.32	10478	96.66
Above 50M	361	3.33	10839	99.99



Original question: Do more than 1M installed applications have higher rating scores compared to less than 1M installed ones?

By examining the frequencies, we find that 500K is a good checkpoint to divide the dataset into two categories with near equal number of records.

One-Way Analysis of Variance

Results

The ANOVA Procedure

Dependent Variable: Rating

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	54.978363	54.978363	243.84	<.0001
Error	10837	2443.412098	0.225469		
Corrected Total	10838	2498.390461			

R-Square	Coeff Var	Root MSE	Rating Mean
0.022006	11.33296	0.474836	4.189868

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Installs_Binary_N	1	54.97836273	54.97836273	243.84	<.0001

One-Way Analysis of Variance

Results

The ANOVA Procedure

Levene's Test for Homogeneity of Rating Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Installs_Binary_N	1	130.8	130.8	306.69	<.0001
Error	10837	4622.3	0.4265		

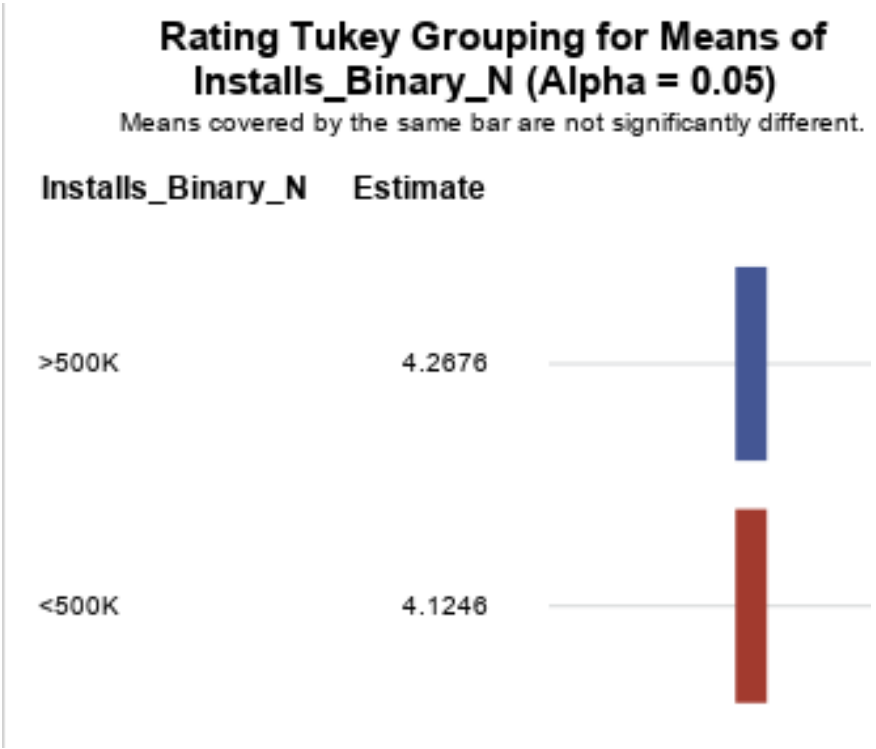
Welch's ANOVA for Rating			
Source	DF	F Value	Pr > F
Installs_Binary_N	1.0000	266.42	<.0001
Error	9602.7		

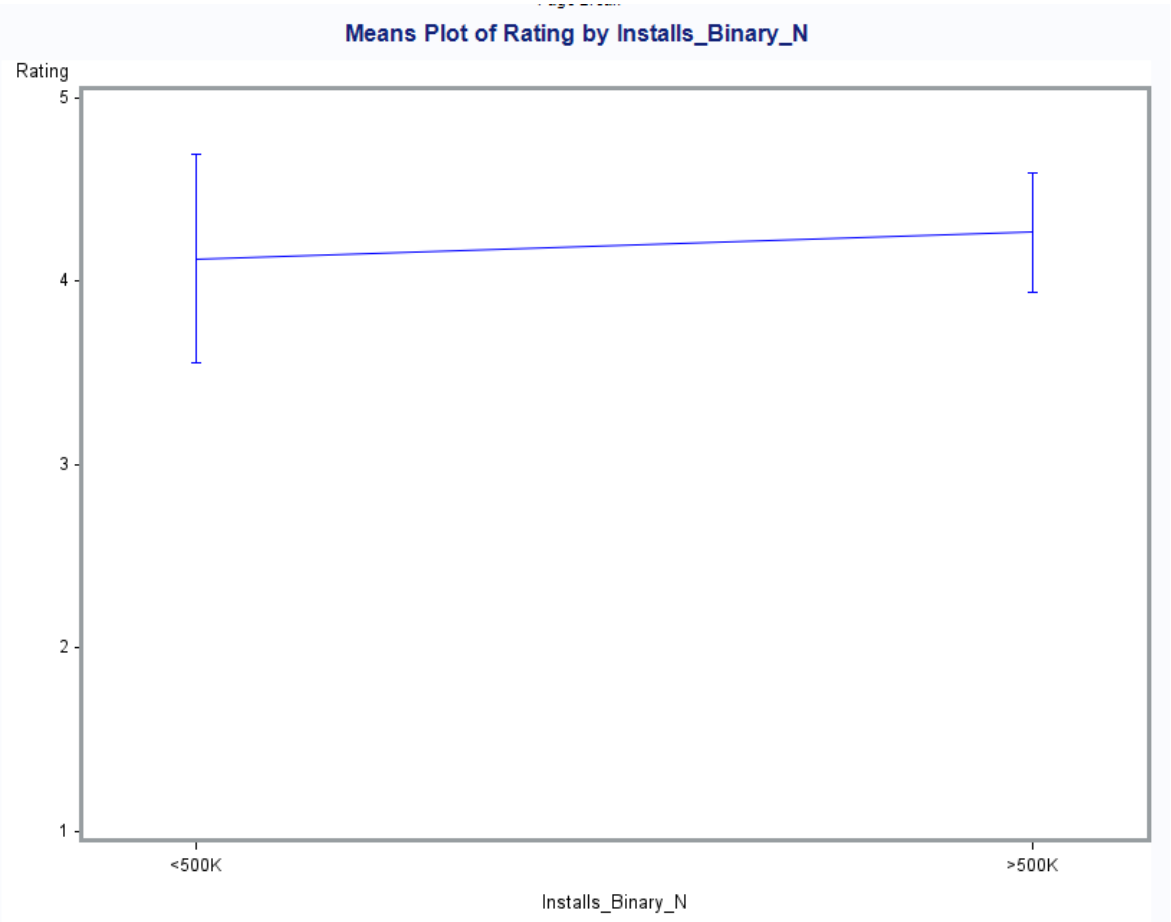
Hypothesis Test:

$$\begin{cases} H_0: Average\ rating\ of\ applications\ installed\ more\ than\ 500K\ is\ equal\ to\ the\ average\ rating\ of\ those\ installed\ less\ than\ 500K. \\ H_a: Otherwise \end{cases}$$

P-value for the ANOVA test is less than the α value 0.05.

This indicates there is a significant difference between average rating for apps installed more than 500K and those installed less than 500k.





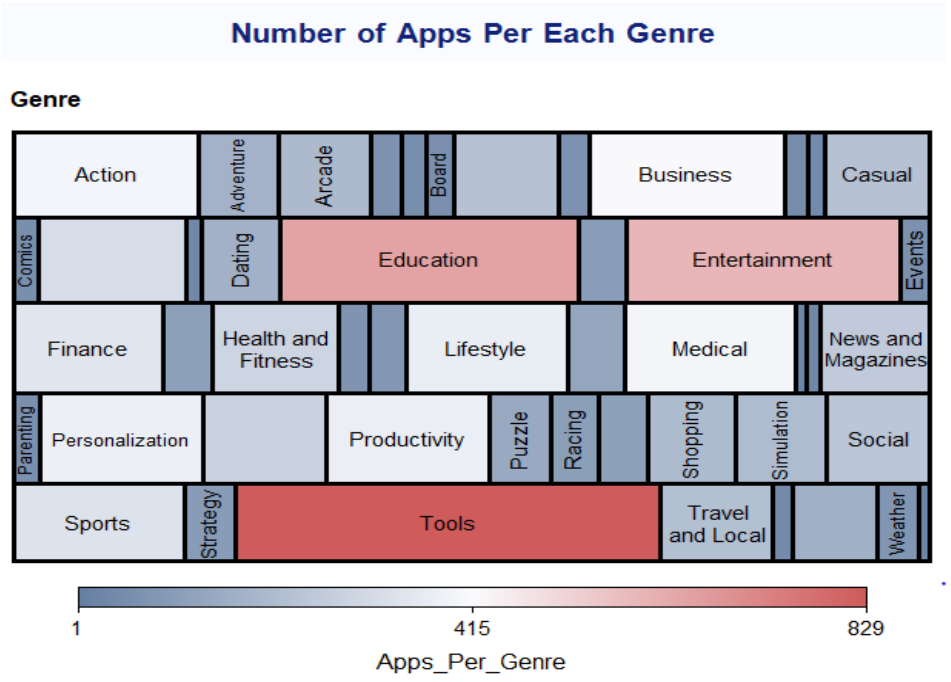
Conclusion: From the two above plots, we can see that applications installed more than 500K have a higher average rating score.

Question 3: In each genre, which three applications are the most popular ones?

There are 50 individual genres. Each application has been assigned with one or more genres.

Here we are considering each combination of the genres as one category and top three popular applications are being filtered.

Popularity: We are defining popular app as the app with higher installations. In case of equal number of installations, we are considering the app with highest rating as the more popular application.



Steps:

Step 1: Run Rank task to rank Installs_Numeric Column by Genres column.

Ranking Method:

- Rank largest to lowest
- If set of Ranking column values are equal then assign the smallest rank.

Step 2: Run Rank task to rank Ratings column by Genres and rank_Installs_Numeric columns.

Ranking Method:

- Same as in step 1.

Step 3: Apply Filter Task to filter the apps with both rank_Installs_Numeric and rank_Ratings columns in (1, 2, 3)

Step 4: Order the list by Genres, Rank_Installations_Numeric and Rank_Ratings.

Step 5: Use PROC SQL statement to pick top 3 rows of each genre and find union of them.

	Genres	App	Category	Rating	Installs_Categories_N	rank_Installs_Numeric	rank_Rating
1	Action	Bowmasters	GAME	4.7	Above 50M	1	1
2	Action	CATS: Crash A...	GAME	4.7	Above 50M	1	1
3	Action	War Robots	GAME	4.6	Above 50M	1	3
4	Action;Action &...	BEYBLADE BU...	GAME	4.5	10M-50M	1	1
5	Action;Action &...	Strawberry Sho...	FAMILY	4.3	10M-50M	1	2
6	Action;Action &...	Strawberry Sho...	GAME	4.3	10M-50M	1	2
7	Adventure	â-º MultiCraft â...	GAME	4.3	Above 50M	1	1
8	Adventure	Eyes - The Sca...	GAME	4.4	10M-50M	2	1
9	Adventure	Harry Potter: H...	GAME	4.4	10M-50M	2	1
10	Adventure	Jungle Monkey...	GAME	4.4	10M-50M	2	1
11	Adventure;Acti...	ROBLOX	FAMILY	4.5	10M-50M	1	1
12	Adventure;Acti...	ROBLOX	GAME	4.5	10M-50M	1	1
13	Adventure;Brai...	The Hunt for th...	FAMILY	4.6	100K-500K	1	1
14	Adventure;Edu...	Masha and the...	FAMILY	4.1	10M-50M	1	1
15	Arcade	Geometry Das...	GAME	4.6	Above 50M	1	1
16	Arcade	Granny	GAME	4.5	Above 50M	1	2
17	Arcade	Red Ball 4	GAME	4.4	Above 50M	1	3
18	Arcade;Action...	Disney Crossy...	FAMILY	4.5	10M-50M	1	1
19	Arcade;Action...	Minecraft	FAMILY	4.5	10M-50M	1	1

Adjacent figure shows some of the results.

Among the popular applications of each genres combination we can see that 298 are Free applications and 15 are Paid applications.

	Type	Num_Applications
1	Free	298
2	Paid	15

Conclusion:

Free applications are more in number among the popular application of each of the genres.

Question 4: Is there any relation between the number of installations and reviews?

One-Way Analysis of Variance

Results

The ANOVA Procedure

Dependent Variable: Reviews

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.5348143E15	2.5348143E15	303.95	<.0001
Error	10837	9.0374535E16	8.3394422E12		
Corrected Total	10838	9.2909349E16			

R-Square	Coeff Var	Root MSE	Reviews Mean
0.027283	650.1236	2887809	444193.9

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Installs_Binary_N	1	2.5348143E15	2.5348143E15	303.95	<.0001

One-Way Analysis of Variance

Results

The ANOVA Procedure

Levene's Test for Homogeneity of Reviews Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Installs_Binary_N	1	8.965E29	8.965E29	36.95	<.0001
Error	10837	2.629E32	2.426E28		

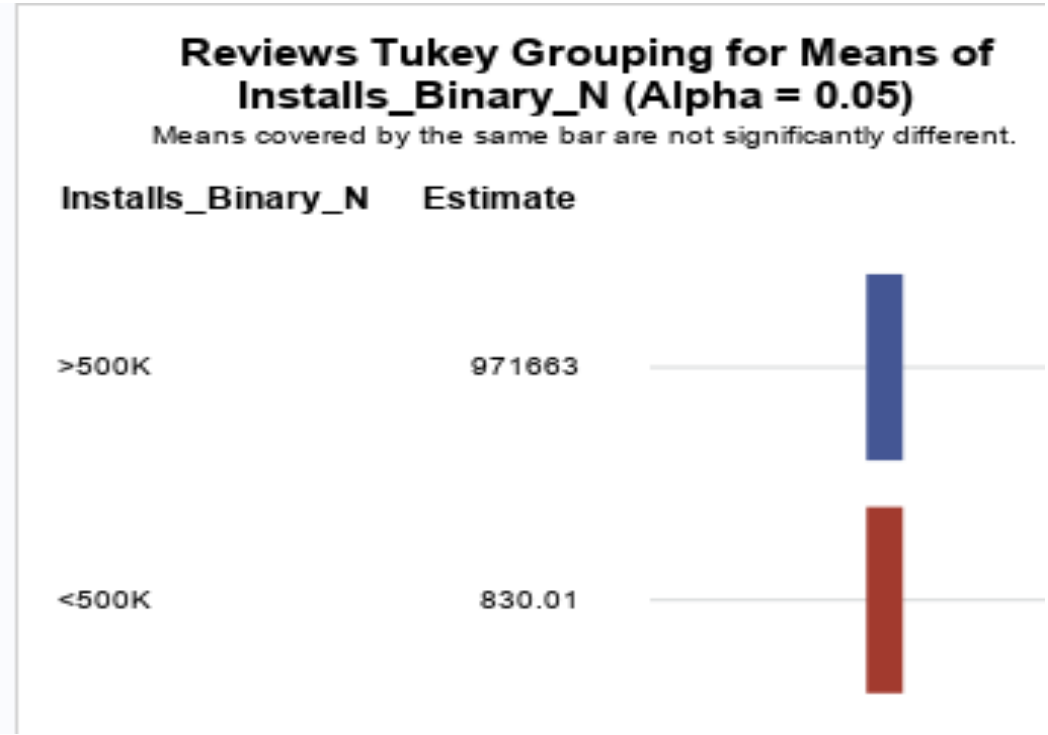
Welch's ANOVA for Reviews

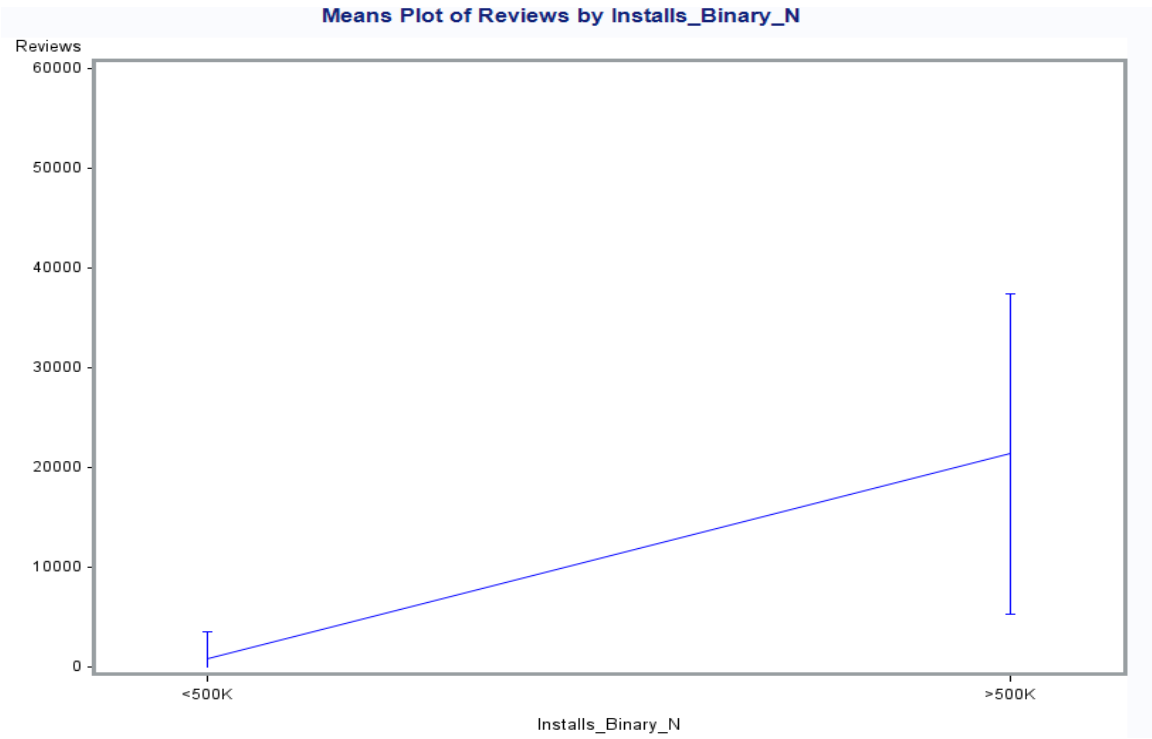
Source	DF	F Value	Pr > F
Installs_Binary_N	1.0000	255.48	<.0001
Error	4949.0		

Hypothesis Test:

$$\begin{cases} H_0: Mean\ of\ number\ of\ reviews\ of\ apps\ installed\ more\ than\ 500K\ is\ equal\ to\ mean\ of\ number\ of\ reviews\ of\ apps\ installed\ less\ than\ 500K. \\ H_a: Otherwise \end{cases}$$

P-value for the ANOVA test is less than the α value 0.05.





This indicates there is a significant difference between mean of number of reviews of apps installed more than $500K$ and mean of number of reviews of apps installed less than $500K$.

<u>Conclusion:</u> More installed applications have much more average reviews.