



Analysis of Top Android Game Categories

Jasmine Diep

Taron Sarksyian

Professor Shilpa Balan

CIS 5250

December 14, 2021



Table of Contents

A. Introduction.....	1
B. Data Set URL	3
C. Data Description.....	3
D. Data Cleaning.....	5
Scoping in Data:	6
1) Dropping out of scope data	6
2) Remove the letter "M" from the values in the “installs” column.	8
3) Change the field type of the “installs” column to integers.	9
4) Remove rows with “NA” values.	10
5) Rename the “installs” column to “installs in millions.”	11
6) Clean column names for all columns.	12
7) Remove unused columns: rank, title, average_rating, growth_(30 days), growth_(60 days), and paid.	13
E. Analysis & Visualizations	14
1) What is the total number of installs in millions for each game category?.....	14
Analysis:	16
2) What are the total star ratings for each numbered star rating by category?.....	16
Analysis:	18
3) What are the total prices for each game category?.....	20
Analysis:	22
F. Statistical Summary and Functions	22
Overall View of Statistical Summary	22
Statistical Summary of Total Ratings	23
Analysis:	23
Statistical Summary of Installs in Millions.....	24
Analysis:	25
RStudio - R Script Feature.....	25
R Script:.....	26
R Script Code:	26
RStudio - User Defined Function	27
G. Conclusion	28
References:	29

A. Introduction

Video games have been around for decades beginning with arcade systems, to home consoles, to handheld consoles, and now to smartphones. Today, playing video games on mobile devices may arguably be the biggest common hobby for millions of young adults around the world, especially since a majority of the world's population own a smartphone or have access to one. In fact, approximately two-thirds of American households consist of household members that regularly play video games and contribute to the \$100 billion global video gaming industry (History, 2019). According to an article from Statista, smartphone games have generated approximately 50 percent of revenue globally in the video gaming industry in 2020, which is projected to increase beyond 100 billion dollars by 2023 (Clement, 2021). The accelerated growth of mobile gaming has also impacted the video gaming landscape of the United States. Being one of the largest video game markets in the world, the United States recorded revenue of 10.73 billion dollars in 2020 due to new technology, innovation, and as a result of the lockdowns during the coronavirus (COVID-19) pandemic (Clement, 2021). This is 19.52 percent more as compared to the revenue recorded in 2019, which indicates a significant increase in users transitioning to mobile gaming on their smartphones (Statista, 2021).

Due to the unprecedented circumstances from COVID-19, there was a limitation on activities people can do outdoors as everyone had to abide by the stay-at-home orders, which led to an increase for people to search for indoor activities to do at home such as playing mobile games. This motivated us to select the data set on Android games for our project analysis. Android games are easily downloaded from the Google Playstore application through one's smartphone. What also motivated us to work on this data set was a genuine interest in mobile gaming, especially to one of us. Currently, one of us owns a Razer Kishi (image in cover page),

which is a controller that attaches to your smartphone. This new technology enhances the experience of playing games on smartphones with the added features of a controller as opposed to the touch screen buttons of the smartphone.

In exploring and visualizing the data set on Android games, we wanted to identify what categories of games attracted consumers to play the most. Their preferences drive many factors that companies use to track and understand consumer interest to produce more games, such as the number of installation downloads for each game, total ratings, star ratings, and price. Most games available in the Google Play Store are free for consumers to download and play instantly, however we wanted to also analyze which game categories consumers were willing to purchase. In this project, we will analyze and visualize the Android Games data to address the following questions:

- What is the total number of installs in millions for each game category?
- What are the total star ratings for each numbered star rating by category?
- What are the total prices for each game category?

B. Data Set URL

Data Source: <https://www.kaggle.com/dhruvildave/top-play-store-games>

Data Size: 1,731 rows and 15 columns

Data Set Screenshot:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
2	1	Garena Free Fire- World Series	86273129	500.0 M	4	2.1	6.9	0	GAME ACTION	63546786	4949507	3158756	2122183	12495915	FALSE
3	2	PUBG MOBILE - Traverse	37276732	500.0 M	4	1.8	3.6	0	GAME ACTION	28339753	2164478	1253185	809821	4709492	FALSE
4	3	Mobile Legends: Bang Bang	26663595	100.0 M	4	1.5	3.2	0	GAME ACTION	18777988	1812094	1050600	713912	4308998	FALSE
5	4	Brawl Stars	17971552	100.0 M	4	1.4	4.4	0	GAME ACTION	13018610	1552950	774012	406184	2219794	FALSE
6	5	Sniper 3D: Fun Free Online FPS Shooting Game	14464235	500.0 M	4	0.8	1.5	0	GAME ACTION	9827328	2124154	1047741	380670	1084340	FALSE
7	6	Call of Duty®: Mobile - Season 4: Spurned & Burned	13572148	100.0 M	4	2	4	0	GAME ACTION	10501443	1274162	517273	268489	1010776	FALSE
8	7	Among Us	11936964	100.0 M	3	1.8	5.6	0	GAME ACTION	5954262	1041297	853099	719378	3368926	FALSE
9	8	Temple Run 2	9633929	500.0 M	4	0.3	0.8	0	GAME ACTION	5793669	991341	614643	349003	1099571	FALSE
10	9	PUBG MOBILE LITE	7578630	100.0 M	4	1	2.5	0	GAME ACTION	5302545	500696	351523	238986	1104879	FALSE
11	10	Gangstar Vegas: World of Crime	6268377	100.0 M	4	0.4	1	0	GAME ACTION	4509647	605510	319332	167792	666094	FALSE
12	11	Pixel Gun 3D: FPS Shooter & Battle Royale	5681934	100.0 M	4	0.2	0.5	0	GAME ACTION	3866308	527159	326294	186593	775578	FALSE
13	12	Bowmasters	5471344	100.0 M	4	2	4.1	0	GAME ACTION	4319351	514574	229990	98901	308525	FALSE
14	13	Standoff 2	4801658	50.0 M	4	3	7.6	0	GAME ACTION	3884644	365380	135150	65422	351059	FALSE
15	14	Talking Tom Gold Run	4710639	100.0 M	4	1.3	4012.8	0	GAME ACTION	3581634	397794	200800	116460	413948	FALSE
16	15	Garena LÃn QuÃn Mobile	4564398	50.0 M	4	1	2.7	0	GAME ACTION	3142312	386513	204843	111640	719082	FALSE
17	16	Banana Kong	4496723	100.0 M	4	0.5	1	0	GAME ACTION	3497817	518453	223009	72115	185328	FALSE
18	17	Crossy Road	4464668	100.0 M	4	0.1	0.2	0	GAME ACTION	3379550	576816	234463	68644	205194	FALSE
19	18	War Robots. 6v6 Tactical Multiplayer Battles	4221389	50.0 M	3	692.7	0.6	0	GAME ACTION	2634036	450688	267931	164087	704645	FALSE
20	19	MORTAL KOMBAT: The Ultimate Fighting Game!	4215808	50.0 M	4	0.5	1.2	0	GAME ACTION	2874353	457806	254154	136153	493340	FALSE
21	20	Last Day on Earth: Survival	4159837	50.0 M	4	0.7	1.6	0	GAME ACTION	3073534	463932	205166	103198	314004	FALSE
22	21	Agar.io	4112075	100.0 M	3	0.2	0.3	0	GAME ACTION	2493810	428315	289428	156522	743997	FALSE
23	22	DEER HUNTER CLASSIC	3987062	100.0 M	4	0	0	0	GAME ACTION	2776350	614755	242898	81199	272058	FALSE
24	23	Modern Combat 5: eSports FPS	3516024	100.0 M	4	0.5	1	0	GAME ACTION	2419540	421218	214049	100008	361206	FALSE
25	24	Kick the Buddy	3213257	100.0 M	4	1	2.3	0	GAME ACTION	2174466	265019	189630	118612	465528	FALSE
26	25	Special Forces Group 2	3195503	100.0 M	4	1.4	7584.9	0	GAME ACTION	2486283	232254	118927	70889	287147	FALSE
27	26	DEAD TRIGGER 2 - Zombie Game FPS shooter	3140510	50.0 M	4	0.2	0.5	0	GAME ACTION	2332406	409344	156460	63835	178462	FALSE
28	27	Mini Militia - Doodle Army 2	3133361	100.0 M	4	0.9	2.1	0	GAME ACTION	2107121	271167	174340	107203	473528	FALSE
29	28	Call of Duty®: Mobile - Garena	2695923	10.0 M	4	3.2	5.5	0	GAME ACTION	1922909	223625	133776	79762	335849	FALSE
30	29	CATS: Crash Arena Turbo Stars	2670589	50.0 M	4	1082.3	1	0	GAME ACTION	1944649	272948	134237	69945	248907	FALSE
31	30	Injustice: Gods Among Us	2540439	10.0 M	4	0.1	0.2	0	GAME ACTION	1879504	251048	115871	60732	233281	FALSE
32	31	Swamp Attack	2472561	100.0 M	4	0.3	0.7	0	GAME ACTION	1841035	284687	138333	56785	151718	FALSE
33	32	GUNSHIP BATTLE: Helicopter 3D	2380639	50.0 M	4	0.2	0.3	0	GAME ACTION	1582549	267812	149090	73581	307605	FALSE
34	33	Real Steel World Robot Boxing	2365838	50.0 M	4	0.1	0.2	0	GAME ACTION	1849560	211891	101452	42652	160281	FALSE
35	34	Critical Ops: Online Multiplayer FPS Shooting Game	2271977	50.0 M	4	0.4	0.9	0	GAME ACTION	1561779	253442	130615	65118	261021	FALSE
36	35	LINE Rangers - a tower defense RPG w/Brown & Cony!	2131549	10.0 M	4	0.1	0.2	0	GAME ACTION	1508264	249413	138706	48233	186930	FALSE
37	36	Worms Zone .Io - Voracious Snake	2100984	100.0 M	4	3.1	6.7	0	GAME ACTION	1509125	187398	116123	66573	221763	FALSE
38	37	Six-Guns: Gang Showdown	2083413	10.0 M	4	0.1	0.1	0	GAME ACTION	1394048	187327	118080	65989	317967	FALSE

C. Data Description

The data set used is selected from the Kaggle site as a CSV file and the total file size is 179 KB. The topic of the data set is on the recent top 100 games by category from the Google Play Store as of June 2021. The below table includes the column fields and its description as it is displayed in the CSV file. The table contains usable columns such as the game's rank, title, total ratings, installs, price, and category.

	Field Name	Field Description	Field Used
1	Rank	This field indicates the rank for each game from 1 to 100 by a particular category.	N
2	Title	Identifies the name of the game.	N
3	Total Ratings	Records the total number of ratings for each game.	Y
4	Installs	Records the total number of download installs for each game.	Y
5	Average Rating	Records the total average ratings out of 5 for each game.	N
6	Growth (30 days)	Indicates the percent growth in 30 days per game.	N
7	Growth (60 days)	Indicates the percent growth in 60 days per game.	N
8	Price	This field contains the price in USD consumers paid for the game.	Y
9	Category	This field contains the game genre.	Y
10	5 Star Ratings	Stores the total number of 5-star ratings based on the consumer's preference for the game.	Y
11	4 Star Ratings	Stores the total number of 4-star ratings based on the consumer's preference for the game.	Y
12	3 Star Ratings	Stores the total number of 3-star ratings based on the consumer's preference for the game.	Y
13	2 Star Ratings	Stores the total number of 2-star ratings based on the consumer's preference for the game.	Y
14	1 Star Ratings	Stores the total number of 1-star ratings based on the consumer's preference for the game.	Y
15	Paid	This field identifies the purchase status of the game as "TRUE" or "FALSE."	N

Below is the screenshot of importing the data set (CSV file) into RStudio and viewing the data set.

```
> setwd("C:/Users/taron/Desktop/Fall '21/CIS 5250/R Directory")
> library(readr)
> android_games <- read_csv("android-games.csv")
Rows: 1730 Columns: 15

-- Column specification -----
delimiter: ","
chr (3): title, installs, category
dbl (11): rank, total ratings, average rating, growth (30 days), growth (60 days), price, 5 star ratings, 4 star ratings, 3 star r...
lgl (1): paid

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
> view(android_games)
> |
```

```
> setwd("C:/Users/taron/Desktop/Fall '21/CIS 5250/R Directory")
```

```
> library(readr)
```

```
> android_games <- read_csv("android-games.csv")
```

```
> View(android_games)
```

	rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	1	Garena Free Fire- World Series	86273129	500.0 M	4	2.1	6.9	0	GAME ACTION	63546766	4949507	3158756	2122183	12495915	FALSE
2	2	PUBG MOBILE - Traverse	37276732	500.0 M	4	1.8	3.6	0	GAME ACTION	28339753	2164478	1253185	809821	4709492	FALSE
3	3	Mobile Legends: Bang Bang	26663595	100.0 M	4	1.5	3.2	0	GAME ACTION	18777988	1812094	1050600	713912	4308998	FALSE
4	4	Brawl Stars	17971552	100.0 M	4	1.4	4.4	0	GAME ACTION	13018610	1552950	774012	406184	2219794	FALSE
5	5	Sniper 3D: Fun Free Online FPS Shooting Game	14464235	500.0 M	4	0.8	1.5	0	GAME ACTION	9827328	2124154	1047741	380670	1084340	FALSE
6	6	Call of Duty®: Mobile - Season 4: Spurred & Burned	13572148	100.0 M	4	2.0	4.0	0	GAME ACTION	10501443	1274162	517273	268489	1010778	FALSE
7	7	Among Us	11936964	100.0 M	3	1.8	5.6	0	GAME ACTION	5954262	1041297	853099	719378	3368926	FALSE
8	8	Temple Run 2	9633929	500.0 M	4	0.3	0.8	0	GAME ACTION	6579389	991341	614643	349003	1099571	FALSE
9	9	PUBG MOBILE LITE	7578630	100.0 M	4	1.0	2.5	0	GAME ACTION	5382545	500696	351523	238986	1104879	FALSE
10	10	Gangstar Vegas: World of Crime	6268377	100.0 M	4	0.4	1.0	0	GAME ACTION	4509647	605510	319332	167792	666094	FALSE

Showing 1 to 11 of 1,730 entries, 15 total columns

D. Data Cleaning

In analyzing the raw data, we noticed there were duplicate game titles among different categories. We cleaned the data using the following steps in RStudio to scope in our data by excluding categories with duplicate titles and including categories with a price.

Scoping in Data:

1) Dropping out of scope data.

We extracted rows with duplicate titles to determine which categories to scope out of our data set.

```
> duplicate_titles<-android_games[duplicated(android_games$title),]
> View(duplicate_titles)
```

```
> duplicate_titles<-android_games[duplicated(android_games$title),]
```

```
> View(duplicate_titles)
```

rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	3 Solitaire	1590733	100.0 M	4	0.4	0.9	0	GAME CARD	431463	90350	40065	15998	70189	FALSE
2	3 Solitaire	1590733	100.0 M	4	0.4	0.9	0	GAME CARD	418115	47121	16871	6574	18567	FALSE
3	3 Solitaire	1590733	100.0 M	4	0.4	0.9	0	GAME CARD	299715	68177	20617	4959	12653	FALSE
4	10 Solitaire	648068	50.0 M	4	0.0	0.0	0	GAME CARD	1117298	308402	82191	25750	57090	FALSE
5	10 Solitaire	648068	50.0 M	4	0.0	0.0	0	GAME CARD	431463	90350	40065	15998	70189	FALSE
6	10 Solitaire	648068	50.0 M	4	0.0	0.0	0	GAME CARD	418115	47121	16871	6574	18567	FALSE
7	10 Solitaire	648068	50.0 M	4	0.0	0.0	0	GAME CARD	299715	68177	20617	4959	12653	FALSE
8	13 Solitaire	507250	50.0 M	4	0.9	1.9	0	GAME CARD	1117298	308402	82191	25750	57090	FALSE
9	13 Solitaire	507250	50.0 M	4	0.9	1.9	0	GAME CARD	431463	90350	40065	15998	70189	FALSE
10	13 Solitaire	507250	50.0 M	4	0.9	1.9	0	GAME CARD	418115	47121	16871	6574	18567	FALSE

Showing 1 to 11 of 55 entries, 15 total columns

Then, we extracted rows to identify which categories have a price.

```
> library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':
```

```
filter, lag
```

```
The following objects are masked from 'package:base':
```

```
intersect, setdiff, setequal, union
```

```
> price_zero<-filter(android_games, price != "0")
> View(price_zero)
```

```
> library(dplyr)
```

```
> price_not_zero<-filter(android_games, price != "0")
```

```
> View(price_not_zero)
```


rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	Minecraft	4100452	10.0 M	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE
2	Geometry Dash	777941	1.0 M	4	0.7	1.3	1.99	GAME ARCADE	636912	73308	24749	9026	33945	TRUE
3	Reigns	146718	500.0 k	4	0.1	0.1	2.99	GAME CARD	108356	25417	5887	2868	4187	TRUE
4	Cytus II	111679	1.0 M	4	0.8	1.7	1.99	GAME MUSIC	88842	11945	4620	1947	4322	TRUE
5	Muse Dash	34140	100.0 k	4	768.5	6.7	1.49	GAME MUSIC	29210	2774	718	429	1007	TRUE
6	Evertale	311126	1.0 M	4	725.3	5.1	0.99	GAME ROLE PLAYING	222761	32884	15877	9298	30304	TRUE
7	Trivia Crack (No Ads)	75712	500.0 k	4	0.7	1.4	1.99	GAME TRIVIA	56454	12628	3284	1088	2256	TRUE

Showing 1 to 7 of 7 entries, 15 total columns

Once we identified which categories had duplicate titles, we deleted all the rows associated with those categories with duplicate titles. We also deleted all the rows for categories that did not contain a price.

```
> android_games<-filter(android_games, !(category %in% c("GAME ACTION", "GAME ADVENTURE", "GAME BOARD", "GAME CARD", "GAME CASINO", "GAME CASUAL", "GAME EDUCATIONAL", "GAME PUZZLE", "GAME RACING", "GAME SIMULATION", "GAME SPORTS", "GAME STRATEGY", "GAME WORD")))
> view(android_games)
```

```
> android_games<-filter(android_games, !(category %in% c("GAME ACTION", "GAME ADVENTURE", "GAME BOARD", "GAME CARD", "GAME CASINO", "GAME CASUAL", "GAME EDUCATIONAL", "GAME PUZZLE", "GAME RACING", "GAME SIMULATION", "GAME SPORTS", "GAME STRATEGY", "GAME WORD")))
```

Before Cleaning:

rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	Garena Free Fire- World Series	86273129	500.0 M	4	2.1	6.9	0	GAME ACTION	63546766	4949507	3158756	2122183	12495915	FALSE
2	PUBG MOBILE - Traverse	37276732	500.0 M	4	1.8	3.6	0	GAME ACTION	28339753	2164478	1253185	809821	4709492	FALSE
3	Mobile Legends: Bang Bang	26663595	100.0 M	4	1.5	3.2	0	GAME ACTION	18777988	1812094	1050600	713912	4308998	FALSE
4	Brawl Stars	17971552	100.0 M	4	1.4	4.4	0	GAME ACTION	13018610	1552950	774012	406184	2219794	FALSE
5	Sniper 3D: Fun Free Online FPS Shooting Game	14464235	500.0 M	4	0.8	1.5	0	GAME ACTION	9827328	2124154	1047741	380670	1084340	FALSE
6	Call of Duty®: Mobile - Season 4: Spurred & Burned	13572148	100.0 M	4	2.0	4.0	0	GAME ACTION	10501443	1274162	517273	268489	1010778	FALSE
7	Among Us	11936964	100.0 M	3	1.8	5.6	0	GAME ACTION	5954262	1041297	853099	719378	3368926	FALSE
8	Temple Run 2	9633929	500.0 M	4	0.3	0.8	0	GAME ACTION	6579369	991341	614643	349003	1099571	FALSE
9	PUBG MOBILE LITE	7578630	100.0 M	4	1.0	2.5	0	GAME ACTION	5382545	500696	351523	238986	1104879	FALSE
10	Gangstar Vegas: World of Crime	6268377	100.0 M	4	0.4	1.0	0	GAME ACTION	4509647	605510	319332	167792	666094	FALSE
11	Pixel Gun 3D: FPS Shooter & Battle Royale	5681934	100.0 M	4	0.2	0.5	0	GAME ACTION	3866308	527159	326294	186593	775578	FALSE
12	Bowmasters	5471344	100.0 M	4	2.0	4.1	0	GAME ACTION	4319351	514574	229990	98901	308525	FALSE
13	Standoff 2	4801658	50.0 M	4	3.0	7.6	0	GAME ACTION	3884644	365380	135150	65422	351099	FALSE
14	Talking Tom Gold Run	4710639	100.0 M	4	1.3	4012.8	0	GAME ACTION	3581634	397794	200800	116460	413948	FALSE
15	Garena Liên Quân Mobile	4564398	50.0 M	4	1.0	2.7	0	GAME ACTION	3142312	386513	204843	111645	719082	FALSE
16	Banana Kong	4486723	100.0 M	4	0.5	1.0	0	GAME ACTION	3497817	518453	223009	72115	185328	FALSE

Showing 1 to 17 of 1,730 entries, 15 total columns

After Cleaning:

rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	Subway Surfers	35665901	1000.0 M	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	Hungry Shark Evolution - Offline survival game	7202013	100.0 M	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	Geometry Dash Lite	6960814	100.0 M	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	Zombie Tsunami	5903477	100.0 M	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	Sonic Dash - Endless Running & Racing Game	5402149	100.0 M	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	88100	268825	FALSE
6	Fruit Ninja®	5367599	100.0 M	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	Jetpack Joyride	4919742	100.0 M	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	Temple Run	4816448	500.0 M	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	Cooking Fever – Restaurant Game	4565220	100.0 M	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	Minecraft	4100452	10.0 M	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE
11	Vector	3398125	100.0 M	4	0.2	0.5	0.00	GAME ARCADE	2395375	354636	204730	99007	344374	FALSE
12	Granny	3290970	100.0 M	4	1.4	2.9	0.00	GAME ARCADE	2393224	283587	152239	95032	366885	FALSE
13	DragonFlight	2858076	10.0 M	4	0.0	0.2	0.00	GAME ARCADE	1795281	475749	286985	88464	211595	FALSE
14	Red Ball 4	2342083	100.0 M	4	1.0	2.1	0.00	GAME ARCADE	1816379	198405	90386	47422	189489	FALSE
15	Minecraft Trial	2222817	100.0 M	4	3.5	7.7	0.00	GAME ARCADE	1461686	192188	120639	79343	368958	FALSE
16	Block Craft 3D: Building Simulator Games For Free	2156973	100.0 M	4	1.8	3.9	0.00	GAME ARCADE	1510574	191501	113211	71395	270289	FALSE

Showing 1 to 17 of 400 entries, 15 total columns

2) Remove the letter "M" from the values in the “installs” column.

This will allow the values to be summed. We noted there are 14 rows where the values were in thousands “k” and this cleaning technique will make these values NA. The NA values will be cleaned in the cleaning step #4.

```
> android_games$installs=gsub("\\M", "", android_games$installs)
```

```
> android_games$installs=gsub("\\M", "", android_games$installs)
```

Before Cleaning:

rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	Subway Surfers	35665901	1000.0 M	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	Hungry Shark Evolution - Offline survival game	7202013	100.0 M	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	Geometry Dash Lite	6960814	100.0 M	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	Zombie Tsunami	5903477	100.0 M	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	Sonic Dash - Endless Running & Racing Game	5402149	100.0 M	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	88100	268825	FALSE
6	Fruit Ninja®	5367599	100.0 M	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	Jetpack Joyride	4919742	100.0 M	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	Temple Run	4816448	500.0 M	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	Cooking Fever – Restaurant Game	4565220	100.0 M	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	Minecraft	4100452	10.0 M	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE
11	Vector	3398125	100.0 M	4	0.2	0.5	0.00	GAME ARCADE	2395375	354636	204730	99007	344374	FALSE
12	Granny	3290970	100.0 M	4	1.4	2.9	0.00	GAME ARCADE	2393224	283587	152239	95032	366885	FALSE
13	DragonFlight	2858076	10.0 M	4	0.0	0.2	0.00	GAME ARCADE	1795281	475749	286985	88464	211595	FALSE
14	Red Ball 4	2342083	100.0 M	4	1.0	2.1	0.00	GAME ARCADE	1816379	198405	90386	47422	189489	FALSE
15	Minecraft Trial	2222817	100.0 M	4	3.5	7.7	0.00	GAME ARCADE	1461686	192188	120639	79343	368958	FALSE
16	Block Craft 3D: Building Simulator Games For Free	2156973	100.0 M	4	1.8	3.9	0.00	GAME ARCADE	1510574	191501	113211	71395	270289	FALSE

Showing 1 to 17 of 400 entries, 15 total columns

After Cleaning:

rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	Subway Surfers	35665901	1000.0	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	Hungry Shark Evolution - Offline survival game	7202013	100.0	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	Geometry Dash Lite	6960814	100.0	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	Zombie Tsunami	5903477	100.0	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	Sonic Dash - Endless Running & Racing Game	5402149	100.0	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	88100	268825	FALSE
6	Fruit Ninja®	5367599	100.0	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	Jetpack Joyride	4919742	100.0	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	Temple Run	4816448	500.0	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	Cooking Fever – Restaurant Game	4565220	100.0	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	Minecraft	4100452	10.0	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE

Showing 1 to 11 of 400 entries, 15 total columns

3) Change the field type of the “installs” column to integers.

```
> android_games$installs<-as.integer(android_games$installs)
warning message:
NAS introduced by coercion
```

```
> android_games$installs<-as.integer(android_games$installs)
```

Before Cleaning:

rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	Subway Surfers	35665901	1000.0	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	Hungry Shark Evolution - Offline survival game	7202013	100.0	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	Geometry Dash Lite	6960814	100.0	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	Zombie Tsunami	5903477	100.0	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	Sonic Dash - Endless Running & Racing Game	5402149	100.0	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	88100	268825	FALSE
6	Fruit Ninja®	5367599	100.0	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	Jetpack Joyride	4919742	100.0	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	Temple Run	4816448	500.0	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	Cooking Fever – Restaurant Game	4565220	100.0	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	Minecraft	4100452	10.0	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE

Showing 1 to 11 of 400 entries, 15 total columns

After Cleaning:

rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	Subway Surfers	35665901	1000	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	Hungry Shark Evolution - Offline survival game	7202013	100	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	Geometry Dash Lite	6960814	100	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	Zombie Tsunami	5903477	100	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	Sonic Dash - Endless Running & Racing Game	5402149	100	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	88100	268825	FALSE
6	Fruit Ninja®	5367599	100	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	Jetpack Joyride	4919742	100	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	Temple Run	4816448	500	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	Cooking Fever – Restaurant Game	4565220	100	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	Minecraft	4100452	10	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE

Showing 1 to 11 of 400 entries, 15 total columns

4) Remove rows with “NA” values.

We will only scope in values of the data where the number of installs is in millions.

```
> android_games<-na.omit(android_games)
```

```
> android_games<-na.omit(android_games)
```

Before Cleaning:

rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings
160	60 Beat Bop: Pop Star Clicker	69242	NA	4	0.1	0.1	0.00	GAME MUSIC	51096	9117	5438	
161	61 Don't Forget the Lyrics	68321	5	4	0.7	1.5	0.00	GAME MUSIC	39709	11266	6684	
162	62 6ix9ine Runner	67388	5	4	2.1	4.2	0.00	GAME MUSIC	54348	4150	2189	
163	63 Beat Roller - Music Ball Race	65897	10	4	0.7	426.0	0.00	GAME MUSIC	39278	12259	3699	
164	64 アイドルマスター シンデレラガールズ スターライトステ...	65210	1	4	0.2	0.4	0.00	GAME MUSIC	44969	8419	4498	
165	65 Marshmello Music Dance	64028	10	4	3.0	6.0	0.00	GAME MUSIC	52167	3567	2048	
166	66 Drum kit (Drums) free	60920	10	4	0.1	0.2	0.00	GAME MUSIC	38319	6683	5388	
167	67 Tebak Lagu Indonesia	56963	1	4	0.0	0.0	0.00	GAME MUSIC	31363	9224	9663	
168	68 Lost in Harmony	56107	1	4	0.1	0.3	0.00	GAME MUSIC	45083	5276	2158	
169	69 Real Piano Teacher 2	55578	1	4	0.8	1.9	0.00	GAME MUSIC	37028	7639	4729	
170	70 SHOW BY ROCK!!![爽快音ゲーショウバイロック]	55357	NA	4	0.0	0.0	0.00	GAME MUSIC	30635	15762	5493	
171	71 Beat Racer	55015	1	4	0.2	0.5	0.00	GAME MUSIC	35154	7096	4617	

Showing 160 to 171 of 400 entries, 15 total columns

After Cleaning:

rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings
159	59 Love Live! School Idol festival - 音楽節奏遊戯	71008	1	4	0.0	0.0	0.00	GAME MUSIC	51391	7698	4139	
160	61 Don't Forget the Lyrics	68321	5	4	0.7	1.5	0.00	GAME MUSIC	39709	11266	6684	
161	62 6ix9ine Runner	67388	5	4	2.1	4.2	0.00	GAME MUSIC	54348	4150	2189	
162	63 Beat Roller - Music Ball Race	65897	10	4	0.7	426.0	0.00	GAME MUSIC	39278	12259	3699	
163	64 アイドルマスター シンデレラガールズ スターライトステ...	65210	1	4	0.2	0.4	0.00	GAME MUSIC	44969	8419	4498	
164	65 Marshmello Music Dance	64028	10	4	3.0	6.0	0.00	GAME MUSIC	52167	3567	2048	
165	66 Drum kit (Drums) free	60920	10	4	0.1	0.2	0.00	GAME MUSIC	38319	6683	5388	
166	67 Tebak Lagu Indonesia	56963	1	4	0.0	0.0	0.00	GAME MUSIC	31363	9224	9663	
167	68 Lost in Harmony	56107	1	4	0.1	0.3	0.00	GAME MUSIC	45083	5276	2158	
168	69 Real Piano Teacher 2	55578	1	4	0.8	1.9	0.00	GAME MUSIC	37028	7639	4729	
169	71 Beat Racer	55015	1	4	0.2	0.5	0.00	GAME MUSIC	35154	7096	4617	
170	72 AvuDance Mobile	54079	1	3	0.3	0.9	0.00	GAME MUSIC	29528	3838	4178	

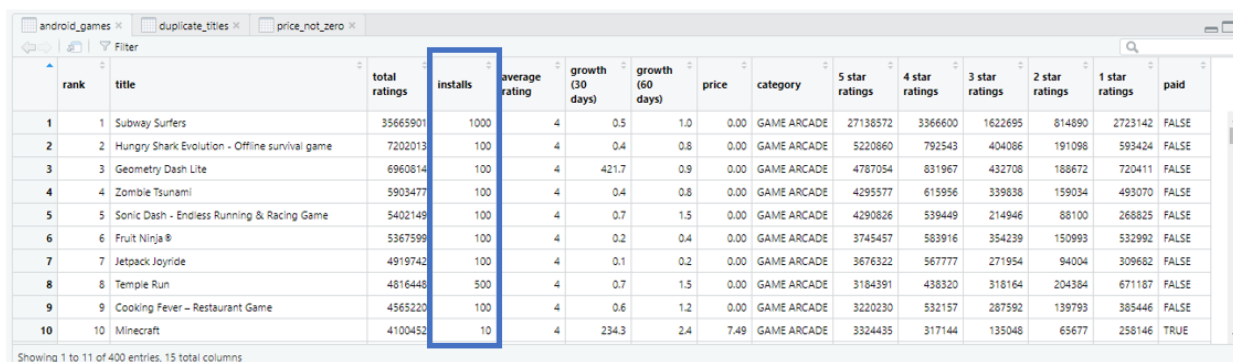
Showing 158 to 170 of 386 entries, 15 total columns

5) Rename the “installs” column to “installs in millions.”

```
> names(android_games)[names(android_games) == "installs"]<-"installs in millions"
```

```
> names(android_games)[names(android_games) == "installs"]<-"installs in millions"
```

Before Cleaning:



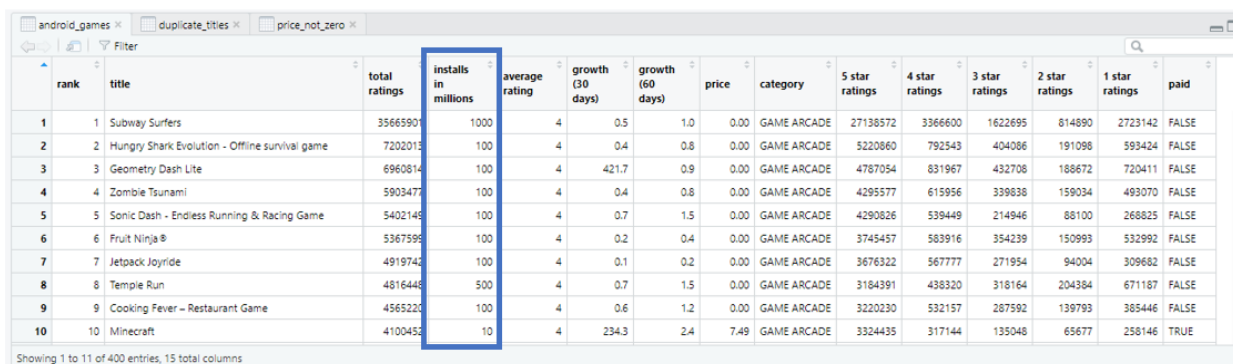
android_games × duplicate_titles × price_not_zero ×

Filter

	rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	1	Subway Surfers	35665901	1000	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	2	Hungry Shark Evolution - Offline survival game	7202013	100	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	3	Geometry Dash Lite	6960814	100	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	4	Zombie Tsunami	5903477	100	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	5	Sonic Dash - Endless Running & Racing Game	5402149	100	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	88100	268825	FALSE
6	6	Fruit Ninja ®	5367599	100	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	7	Jetpack Joyride	4919742	100	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	8	Temple Run	4816448	500	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	9	Cooking Fever - Restaurant Game	4565220	100	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	10	Minecraft	4100452	10	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE

Showing 1 to 11 of 400 entries, 15 total columns

After Cleaning:



android_games × duplicate_titles × price_not_zero ×

Filter

	rank	title	total ratings	installs in millions	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	1	Subway Surfers	35665901	1000	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	2	Hungry Shark Evolution - Offline survival game	7202013	100	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	3	Geometry Dash Lite	6960814	100	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	4	Zombie Tsunami	5903477	100	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	5	Sonic Dash - Endless Running & Racing Game	5402149	100	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	88100	268825	FALSE
6	6	Fruit Ninja ®	5367599	100	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	7	Jetpack Joyride	4919742	100	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	8	Temple Run	4816448	500	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	9	Cooking Fever - Restaurant Game	4565220	100	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	10	Minecraft	4100452	10	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE

Showing 1 to 11 of 400 entries, 15 total columns

6) Clean column names for all columns.

We added an underscore to remove spaces between the text and added a variable “x” in front of the column names for star ratings to avoid using a number as the first character of a column name.

```
> library(janitor)
Attaching package: 'janitor'
The following objects are masked from 'package:stats':
  chisq.test, fisher.test
warning message:
package 'janitor' was built under R version 4.1.2
> android_games <- android_games %>%
+ clean_names
> view(android_games)
```

```
> library(janitor)
```

```
> android_games <- android_games %>%
```

```
+ clean_names
```

Before Cleaning:

rank	title	total ratings	installs in millions	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
1	Subway Surfers	33663901	1000	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	Hungry Shark Evolution - Offline survival game	7202013	100	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	Geometry Dash Lite	6960814	100	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	Zombie Tsunami	5903477	100	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	Sonic Dash - Endless Running & Racing Game	5402149	100	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	88100	268825	FALSE
6	Fruit Ninja®	5367599	100	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	Jetpack Joyride	4919742	100	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	Temple Run	4816448	500	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	Cooking Fever - Restaurant Game	4565220	100	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	Minecraft	4100452	10	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE

Showing 1 to 11 of 400 entries, 15 total columns

After Cleaning:

rank	title	total_ratings	installs_in_millions	average_rating	growth_30_days	growth_60_days	price	category	x5_star_ratings	x4_star_ratings	x3_star_ratings	x2_star_ratings	x1_star_ratings	paid
1	Subway Surfers	33663901	1000	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	Hungry Shark Evolution - Offline survival game	7202013	100	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	Geometry Dash Lite	6960814	100	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	Zombie Tsunami	5903477	100	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	Sonic Dash - Endless Running & Racing Game	5402149	100	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	88100	268825	FALSE
6	Fruit Ninja®	5367599	100	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	Jetpack Joyride	4919742	100	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	Temple Run	4816448	500	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	Cooking Fever - Restaurant Game	4565220	100	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	Minecraft	4100452	10	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE
11	Vector	3398125	100	4	0.2	0.5	0.00	GAME ARCADE	2395375	354636	204730	99007	344374	FALSE

Showing 1 to 12 of 400 entries, 15 total columns

7) Remove unused columns: rank, title, average_rating, growth_(30 days), growth_(60 days), and paid.

```
> android_games$rank <- android_games$title <- android_games$average_rating <-  
android_games$growth_30_days <- android_games$growth_60_days <- android_games$paid  
<- NULL
```

```
> android_games$rank <- android_games$title <- android_games$average_rating <- android_games$growth_30_days <- android_games$growth_60_days  
<- android_games$paid <- NULL
```

Before Cleaning:

rank	title	total_ratings	installs_in_millions	average_rating	growth_30_days	growth_60_days	price	category	x5_star_ratings	x4_star_ratings	x3_star_ratings	x2_star_ratings	x1_star_ratings	paid
1	Subway Surfers	35665901	1000	4	0.5	1.0	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142	FALSE
2	Hungry Shark Evolution - Offline survival game	7202013	100	4	0.4	0.8	0.00	GAME ARCADE	5220860	792543	404086	191098	593424	FALSE
3	Geometry Dash Lite	6960814	100	4	421.7	0.9	0.00	GAME ARCADE	4787054	831967	432708	188672	720411	FALSE
4	Zombie Tsunami	5903477	100	4	0.4	0.8	0.00	GAME ARCADE	4295577	615956	339838	159034	493070	FALSE
5	Sonic Dash - Endless Running & Racing Game	5402149	100	4	0.7	1.5	0.00	GAME ARCADE	4290826	539449	214946	86100	268825	FALSE
6	Fruit Ninja®	5367599	100	4	0.2	0.4	0.00	GAME ARCADE	3745457	583916	354239	150993	532992	FALSE
7	Jetpack Joyride	4919742	100	4	0.1	0.2	0.00	GAME ARCADE	3676322	567777	271954	94004	309682	FALSE
8	Temple Run	4816448	500	4	0.7	1.5	0.00	GAME ARCADE	3184391	438320	318164	204384	671187	FALSE
9	Cooking Fever - Restaurant Game	4565220	100	4	0.6	1.2	0.00	GAME ARCADE	3220230	532157	287592	139793	385446	FALSE
10	Minecraft	4100452	10	4	234.3	2.4	7.49	GAME ARCADE	3324435	317144	135048	65677	258146	TRUE
11	Vector	3398125	100	4	0.2	0.5	0.00	GAME ARCADE	2395375	354636	204730	99007	344374	FALSE

Showing 1 to 12 of 400 entries, 15 total columns

After Cleaning:

	total_ratings	installs_in_millions	price	category	x5_star_ratings	x4_star_ratings	x3_star_ratings	x2_star_ratings	x1_star_ratings
1	35665901	1000	0.00	GAME ARCADE	27138572	3366600	1622695	814890	2723142
2	7202013	100	0.00	GAME ARCADE	5220860	792543	404086	191098	593424
3	6960814	100	0.00	GAME ARCADE	4787054	831967	432708	188672	720411
4	5903477	100	0.00	GAME ARCADE	4295577	615956	339838	159034	493070
5	5402149	100	0.00	GAME ARCADE	4290826	539449	214946	86100	268825
6	5367599	100	0.00	GAME ARCADE	3745457	583916	354239	150993	532992
7	4919742	100	0.00	GAME ARCADE	3676322	567777	271954	94004	309682
8	4816448	500	0.00	GAME ARCADE	3184391	438320	318164	204384	671187
9	4565220	100	0.00	GAME ARCADE	3220230	532157	287592	139793	385446
10	4100452	10	7.49	GAME ARCADE	3324435	317144	135048	65677	258146
11	3398125	100	0.00	GAME ARCADE	2395375	354636	204730	99007	344374

Showing 1 to 12 of 400 entries, 9 total columns

E. Analysis & Visualizations

After cleaning the data set and identifying our scope, we created the below visualizations for each question of analysis to further analyze the data.

1) What is the total number of installs in millions for each game category?

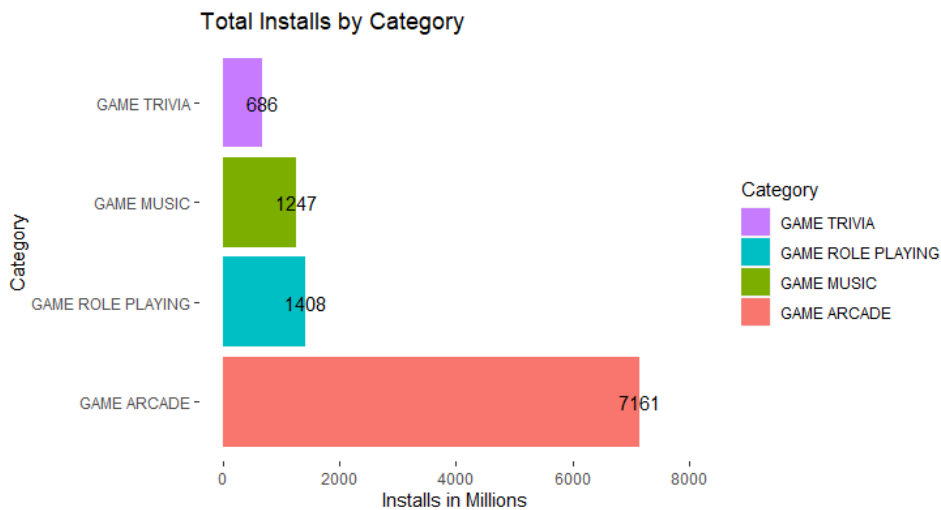


Figure 1 - Horizontal Bar Chart for Total Installs in Millions by Category

R Features:

- Chart Type: Horizontal Bar Chart
- Functions: aesthetics (aes), geom_bar, labs, theme, scale_x_continuous
- Libraries: ggplot2

The below data frame is created to view only the game category and installations in millions.

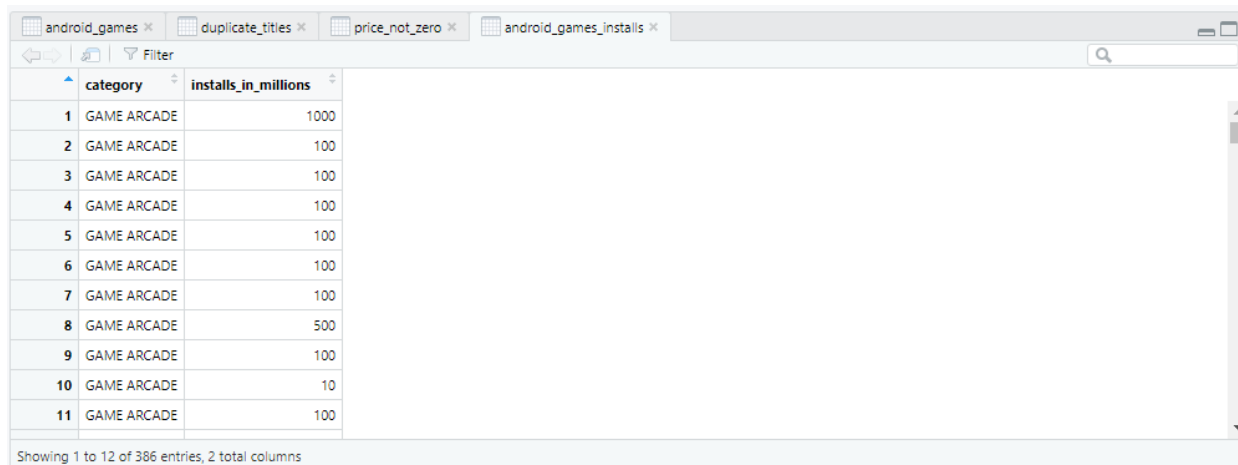
```
> android_games_installs <- android_games %>%
+   clean_names %>%
+   select(category, installs_in_millions)
> view(android_games_installs)
```

```
> android_games_installs <- android_games %>%
```

```
> clean_names %>%
```

```
> select(category, installs_in_millions)
```

```
> View(android_games_installs)
```

	category	installs_in_millions
1	GAME ARCADE	1000
2	GAME ARCADE	100
3	GAME ARCADE	100
4	GAME ARCADE	100
5	GAME ARCADE	100
6	GAME ARCADE	100
7	GAME ARCADE	100
8	GAME ARCADE	500
9	GAME ARCADE	100
10	GAME ARCADE	10
11	GAME ARCADE	100

Showing 1 to 12 of 386 entries, 2 total columns

The below data frame is created to group the installations in millions as totals for each game category.

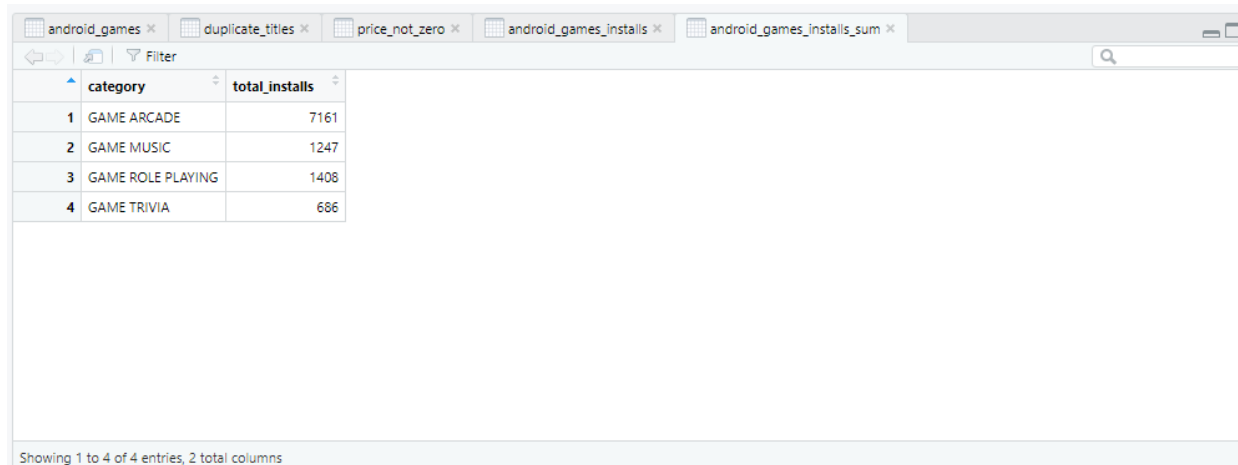
```
> android_games_installs_sum <- android_games_installs %>%
+   group_by(category) %>%
+   summarize(total_installs = sum(installs_in_millions))
> view(android_games_installs_sum)
```

```
> android_games_installs_sum <- android_games_installs %>%
```

```
+ group_by(category) %>%
```

```
+ summarize(total_installs = sum(installs_in_millions))
```

```
> View(android_games_installs_sum)
```



	category	total_installs
1	GAME ARCADE	7161
2	GAME MUSIC	1247
3	GAME ROLE PLAYING	1408
4	GAME TRIVIA	686

Showing 1 to 4 of 4 entries, 2 total columns

```
> ggplot(android_games_installs_sum, aes(x= total_installs, y=reorder(category, -total_installs), fill=category)) + geom_bar(stat
= "identity") + geom_text(aes(label = total_installs)) + scale_fill_discrete(name = "Category", guide=guide_legend(reverse=T)) +
labs(title= "Total Installs by Category", x = "Installs in Millions", y = "Category") + theme(panel.background=element_blank())
+ scale_x_continuous(name="Installs in Millions", limits=c(0,8000))
```

```
> ggplot(android_games_installs_sum, aes(x= total_installs, y=reorder(category, -total_installs),
fill=category)) + geom_bar(stat = "identity") + geom_text(aes(label = total_installs)) +
scale_fill_discrete(name = "Category", guide=guide_legend(reverse=T)) + labs(title= "Total
Installs by Category", x = "Installs in Millions", y = "Category") +
theme(panel.background=element_blank()) + scale_x_continuous(name="Installs in Millions",
limits=c(0,8000))
```

Analysis:

We wanted to see which categories had the highest installations as this is a good indicator of consumer interest and preference for games within each category. As we can see, the category of Game Trivia had 686 million installations, which is the lowest number of installations among the four categories. Whereas Game Music and Game Role Playing had 1,247 and 1,408 million installations, respectively, which is close to twice as many installations as Game Trivia. Finally, Game Arcade had 7,161 million installs, which is the highest number of installations compared to the other categories. Therefore, consumers prefer to play Game Arcade games more than Game Trivia, Game Music, and Game Role Playing.

2) What are the total star ratings for each numbered star rating by category?

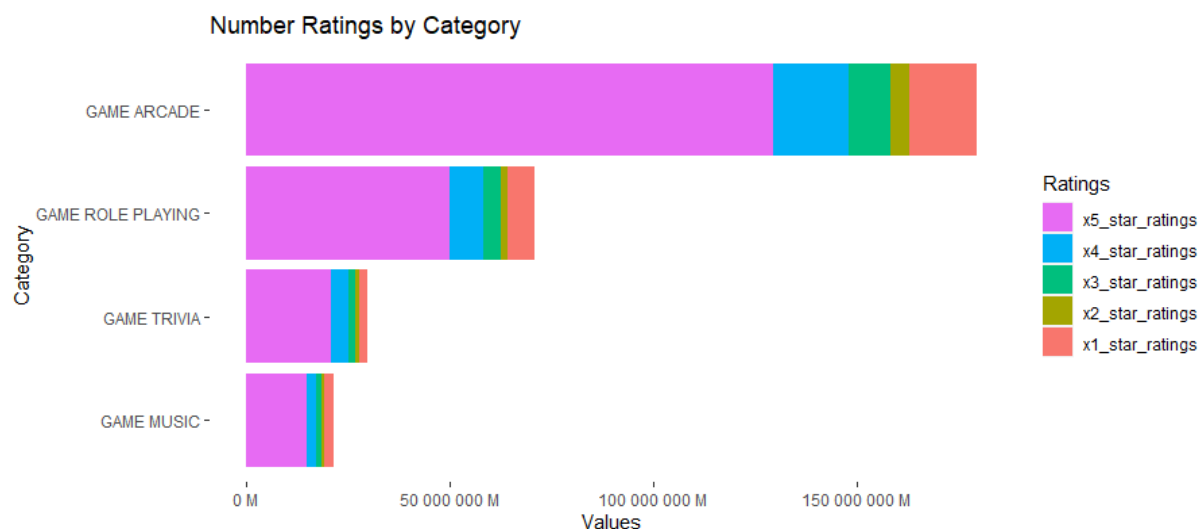


Figure 2 - Horizontal Stacked Bar Chart for Star Ratings by Category

R Features:

- Chart Type: Horizontal Stacked Bar Chart
- Functions: aesthetics (aes), geom_bar, geom_text, scale_fill_discrete, labs, theme, scale_x_continuous
- Libraries: ggplot2

The below data frame is created to view the different number of star ratings by category.

```
> android_games_ratings <- android_games %>%
+   clean_names %>%
+   select(category:x1_star_ratings)
> view(android_games_ratings)
```

```
> android_games_ratings <- android_games %>%
```

```
+ clean_names %>%
```

```
+ select(category:x1_star_ratings)
```

```
> View(android_games_ratings)
```

category	x5_star_ratings	x4_star_ratings	x3_star_ratings	x2_star_ratings	x1_star_ratings
1 GAME ARCADE	27138572	3366600	1622695	814890	2723142
2 GAME ARCADE	5220860	792543	404086	191098	593424
3 GAME ARCADE	4787054	831967	432708	188672	720411
4 GAME ARCADE	4295577	615956	339838	159034	493070
5 GAME ARCADE	4290826	539449	214946	88100	268825
6 GAME ARCADE	3745457	583916	354239	150993	532992
7 GAME ARCADE	3676322	567777	271954	94004	309682
8 GAME ARCADE	3184391	438320	318164	204384	671187
9 GAME ARCADE	3220230	532157	287592	139793	385446
10 GAME ARCADE	3324435	317144	135048	65677	258146
11 GAME ARCADE	2395375	354636	204730	99007	344374

Showing 1 to 12 of 400 entries, 6 total columns

The below data frame is created to gather and consolidate all the star ratings (1 thru 5) from multiple columns into one column called “Ratings” with its corresponding values in another column called “Values”.

```

> library(tidyverse)
-- Attaching packages ----- tidyverse 1.3.1 --
v ggplot2 3.3.5      v purrr  0.3.4
v tibble  3.1.5      v stringr 1.4.0
v tidyr   1.1.4      v forcats 0.5.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
Warning message:
package 'tidyverse' was built under R version 4.1.2
> android_games_ratings_format <- android_games_ratings %>%
+ gather(key = "Ratings", value = "Values", 2:6)
> view(android_games_ratings_format)

```

```
> library(tidyverse)
```

```
> android_games_ratings_format <- android_games_ratings %>%
```

```
+ gather(key = "Ratings", value = "Values", 2:6)
```

```
> View(android_games_ratings_format)
```

	category	Ratings	Values
1	GAME ARCADE	x5_star_ratings	27138572
2	GAME ARCADE	x5_star_ratings	5220860
3	GAME ARCADE	x5_star_ratings	4787054
4	GAME ARCADE	x5_star_ratings	4295577
5	GAME ARCADE	x5_star_ratings	4290826
6	GAME ARCADE	x5_star_ratings	3745457
7	GAME ARCADE	x5_star_ratings	3676322
8	GAME ARCADE	x5_star_ratings	3184391
9	GAME ARCADE	x5_star_ratings	3220230
10	GAME ARCADE	x5_star_ratings	3324435
11	GAME ARCADE	x5_star_ratings	2395375

```

> ggplot(android_games_ratings_format, aes(x= reorder(category, Values), y=Values, fill =
Ratings)) + geom_bar(stat="identity", position="stack") + geom_text(aes(label = Values),
position = position_stack(vjust=0.5), size=3) +
scale_fill_discrete(guide=guide_legend(reverse=T)) + labs(x="Category", y="Values", title =
"Number Ratings by Category") + theme(panel.background=element_blank()) +
scale_y_continuous(labels=unit_format(unit = "M", scale = 1e+00)) + coord_flip()

```

Analysis:

We did not add the labels on the horizontal stacked bar chart as it would clutter the visual. Instead, we created the below data frame to view the sum of individual star ratings (1 to 5) by category to validate the results of the horizontal stacked bar chart.

```

> android_games_ratings_sum <- android_games_ratings %>%
+   group_by(category) %>%
+   summarize(x5_star_ratings = sum(x5_star_ratings), x4_star_ratings = sum(x4_star_ratings), x3_star_ratings = sum(x3_star_ratings), x2_star_ratings
= sum(x2_star_ratings), x1_star_ratings = sum(x1_star_ratings))
> view(android_games_ratings_sum)

```

category	x5_star_ratings	x4_star_ratings	x3_star_ratings	x2_star_ratings	x1_star_ratings
1 GAME ARCADE	129602978	18547895	10020924	4846981	16358970
2 GAME MUSIC	15057603	2180468	1286267	677903	2427712
3 GAME ROLE PLAYING	49902323	8671082	3956948	1812902	6532985
4 GAME TRIVIA	20989759	4288082	1724477	699355	2120237

Showing 1 to 4 of 4 entries, 6 total columns

By creating a horizontal stacked bar chart, we wanted to identify which category had the highest total star ratings to see what consumers enjoyed playing the most. Game Arcade has approximately 129.6 million 5-star ratings, 18.5 million 4-star ratings, 10 million 3-star ratings, 4.8 million 2-star ratings, and 16.3 million 1-star ratings, which is the highest compared to the other categories. Whereas, Game Music has an estimate of 15 million 5-star ratings, 2.1 million 4-star ratings, 1.2 million 3-star ratings, 677 thousand 2-star ratings, and 2.4 million 1-star ratings, which is the least compared to the other categories. On the other hand, Game Role Playing has 49.9 5-star ratings, 8.6 million 4-star ratings, 3.9 million 3-star ratings, 1.8 million 2-star ratings, and 6.5 million 1-star ratings. Finally, Game Trivia has 20.98 5-star ratings, 4.2 million 4-star ratings, 1.7 million 3-star ratings, 699 thousand 2-star ratings, and 2.1 million 1-star ratings. When considering all the values of star ratings for each category, Game Arcade has the highest star ratings which shows that Game Arcade is still the most noteworthy category in Google Play compared to the other three categories, while Game Music has the least number of star ratings. This is consistent with the horizontal stacked bar chart as the bar for Game Arcade is the longest among the categories.

3) What are the total prices for each game category?

Price by Category

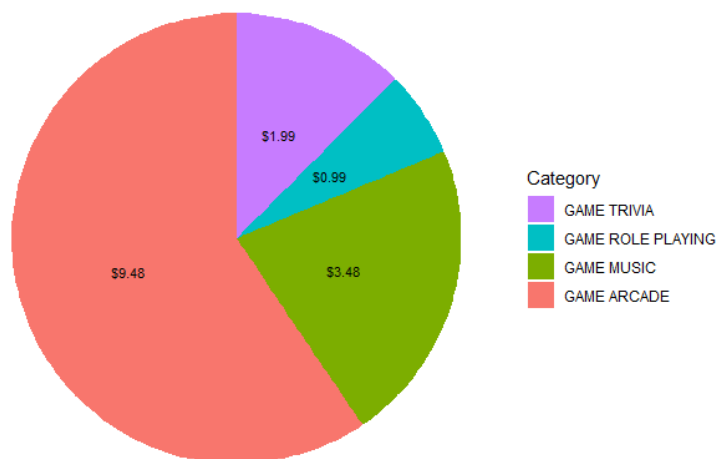


Figure 3 - Pie Chart for Total Prices by Category

R Features:

- Chart Type: Pie Chart
- Functions: aesthetics (aes), geom_bar, coord_polar, geom_text, scale_fill_discrete, labs, theme, theme_void()
- Libraries: ggplot2

The below data frame is to view the price by game category.

```
> android_games_cleaned_price <- android_games %>%
+ clean_names %>%
+ select(price, category)
> view(android_games_cleaned_price)
```

```
> android_games_price <- android_games %>%
```

```
+ clean_names %>%
```

```
+ select(price, category)
```

```
> View(android_games_price)
```

	price	category
1	0.00	GAME ARCADE
2	0.00	GAME ARCADE
3	0.00	GAME ARCADE
4	0.00	GAME ARCADE
5	0.00	GAME ARCADE
6	0.00	GAME ARCADE
7	0.00	GAME ARCADE
8	0.00	GAME ARCADE
9	0.00	GAME ARCADE
10	7.49	GAME ARCADE
11	0.00	GAME ARCADE

Showing 1 to 12 of 400 entries, 2 total columns

The below data frame is created to sum the values of price for each game category.

```
> android_games_price_sum <- android_games_cleaned_price %>%
+ group_by(category) %>%
+ summarize(total_price = sum(price))
> View(android_games_price_sum)
```

```
> android_games_price_sum <- android_games_cleaned_price %>%
```

```
+ group_by(category) %>%
```

```
+ summarize(total_price = sum(price))
```

```
> View(android_games_price_sum)
```

	category	total_price
1	GAME ARCADE	9.48
2	GAME MUSIC	3.48
3	GAME ROLE PLAYING	0.99
4	GAME TRIVIA	1.99

Showing 1 to 4 of 4 entries, 2 total columns

```
> ggplot(android_games_price_sum, aes(x="", y=total_price, fill=category)) + geom_bar(stat="identity", width=1) + coord_polar("y", start=0) + geom_text(aes(label = dollar(total_price)), position = position_stack(vjust=0.5), size=3) + scale_fill_discrete(name = "Category") + labs(x = NULL, y=NULL, title = "Price by Category") + theme(axis.line = element_blank(), axis.ticks = element_blank()) + theme_void()
```

```
> ggplot(android_games_price_sum, aes(x="", y=total_price, fill=category)) +
geom_bar(stat="identity", width=1) + coord_polar("y", start=0) + geom_text(aes(label =
dollar(total_price)), position = position_stack(vjust=0.5), size=3) + scale_fill_discrete(name =
"Category", guide=guide_legend(reverse=T)) + labs(x = NULL, y=NULL, title = "Price by
Category") + theme(axis.line = element_blank(), axis.ticks = element_blank()) + theme_void()
```

Analysis:

We wanted to determine which game category had the highest total price as this helps to identify how much consumers value the games within each category. Based on the pie chart, the total price for games within the Game Trivia category is \$1.99, whereas the total price for games within the Game Role Playing category is \$0.99. The games under the Game Music category is \$3.48, which is higher than Game Trivia and Game Role Playing. Finally, the total price for games under Game Arcade is \$9.48, which is significantly more than the other categories. Therefore, consumers highly value Game Arcade games as they are willing to pay a higher price for games in this category.

F. Statistical Summary and Functions

Below is the overall view statistical summary of the android_games dataset. When running the code below, this displays the results of the minimum, 1st quartile, median, mean, 3rd quartile, and maximum values of every column in the data set.

Overall View of Statistical Summary

```
> summary(android_games)
total_ratings   installs_in_millions   price   category   x5_star_ratings   x4_star_ratings   x3_star_ratings   x2_star_ratings
Min.   : 32993   Min.   : 1.00   Min.   :0.00000   Length:386   Min.   : 13975   Min.   : 2451   Min.   : 1265   Min.   : 528
1st Qu.: 112175 1st Qu.: 5.00   1st Qu.:0.00000   Class :character   1st Qu.: 73578   1st Qu.: 12729   1st Qu.: 6899   1st Qu.: 3379
Median : 381810 Median : 10.00   Median :0.00000   Mode  :character   Median : 265534 Median : 43155   Median : 21371 Median : 9256
Mean   : 779508 Mean   : 27.21   Mean   :0.03228   Mean   : 556785 Mean   : 87028   Mean   : 43914 Mean   : 20788
3rd Qu.: 728267 3rd Qu.: 10.00   3rd Qu.:0.00000   3rd Qu.: 526702 3rd Qu.: 87271   3rd Qu.: 44332   3rd Qu.: 21212
Max.   :35665901 Max.   :1000.00   Max.   :7.49000   Max.   :27138572 Max.   :3366600   Max.   :1622695 Max.   :814890

x1_star_ratings
Min.   : 1116
1st Qu.: 10282
Median : 31298
Mean   : 70991
3rd Qu.: 73807
Max.   :2723142
```

```
> summary(android_games)
```


Statistical Summary of Total Ratings

```
> summary(android_games$total_ratings)
   Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
 32993   112175   381810   779508   728267 35665901
> min(android_games$total_ratings)
[1] 32993
> max(android_games$total_ratings)
[1] 35665901
> mean(android_games$total_ratings)
[1] 779508
> median(android_games$total_ratings)
[1] 381810.5
> sd(android_games$total_ratings)
[1] 2066869
```

```
> summary(android_games$total_ratings)
```

```
> min(android_games$total_ratings)
```

```
> max(android_games$total_ratings)
```

```
> mean(android_games$total_ratings)
```

```
> median(android_games$total_ratings)
```

```
> sd(android_games$total_ratings)
```

Analysis:

One specific column that was valuable for the statistical summary was "total_ratings."

The screenshot below shows the minimum, 1st quartile, median, mean, 3rd quartile, and maximum, and standard deviation values of Total Ratings. The difference between the code for the overall view of the statistical summary versus the individual statistical codes for the total ratings is the "\$" sign, which specifies which column to focus on. The minimum number of total ratings in the dataset is 32,993 within the game categories and the maximum number of total ratings in the dataset is 35,665,901 within the game categories. The mean is the average or the most common value in a collection of numbers. The mean of the total ratings is 779,508, which shows that there are many games in the Google Play Store that do not have too many reviews. It is possible that the most popular games in the four categories that this project emphasizes have around 700,000 ratings. The median is the middle number in a list. The middle amount for total

ratings is 381,810.50. The first quartile is considered the lowest half of the data, whereas the third quartile is the median of the top half of the data (ThoughtCo, 2019). The first quartile of total ratings is 112175 and the third quartile of total ratings is 728,267. Finally, the standard deviation calculates the distribution of the data in relation to the mean and includes the variance (Investopedia, 2021). The standard deviation for the total ratings is 2,066,869. This large number for the standard deviation signifies that there is very little variance in the observed data and it is clustered tightly around the mean.

Statistical Summary of Installs in Millions

```
> summary(android_games$installs_in_millions)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.00   5.00   10.00   27.21   10.00 1000.00
> min(android_games$installs_in_millions)
[1] 1
> max(android_games$installs_in_millions)
[1] 1000
> mean(android_games$installs_in_millions)
[1] 27.20725
> median(android_games$installs_in_millions)
[1] 10
> sd(android_games$installs_in_millions)
[1] 64.34266
```

```
> summary(android_games$installs_in_millions)
```

```
> min(android_games$installs_in_millions)
```

```
> max(android_games$installs_in_millions)
```

```
> mean(android_games$installs_in_millions)
```

```
> median(android_games$installs_in_millions)
```

```
> sd(android_games$installs_in_millions)
```

Analysis:

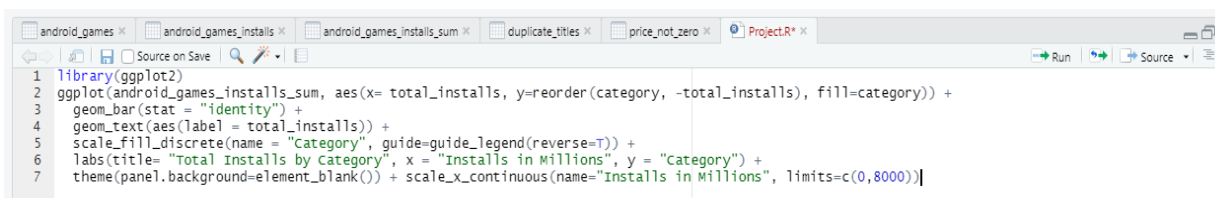
The second specific column that was valuable for our statistical summary was "install_in_millions." The screenshot below shows the minimum, 1st quartile, median, mean, 3rd quartile, maximum, and standard deviation values of Total Installs in Millions. The minimum number of total installs in this dataset is 1 million for a specific title. The maximum number of total ratings in this dataset is 1 billion for a specific title. In this column, the mean is an estimate of 27.2 million. This shows that the amount of people installing games on their phones is still a low amount despite that there are over five billion smartphone users worldwide. One theory is that more people download apps instead of games in which the result of the mean would be much higher for apps than games. There is a disadvantage for the competition when only a few apps have reached the 1 billion mark of total installs in our dataset. The middle amount for total installs is 10 million. The 1st quartile of total ratings is 5 million and the 3rd quartile is 10 million. The standard deviation for total installs is in an estimate of 63.34 million. This low number for the standard deviation signifies that there is very little variance in the observed data, and it is clustered tightly around the mean.

RStudio - R Script Feature

RStudio contains a feature called "script" which differentiates with the way the code is run in the "console" feature. R script is a plain text file that saves the R code in. Thus, script is more of an advantage than just running the code in the console. The console feature cannot save the codes so the codes need to be copied and pasted to Notepad++ or Microsoft Word. Also, the advantage of using script is to place the code in a list format so it is easier to understand the code line by line. There is the option to highlight all the codes so they run simultaneously or run one

line at a time. The screenshot below is what R script looks like in RStudio. The code below is the ggplot bar chart that was shown in the Analysis & Visualizations section. In order to make sure that the code is functional, ggplot2 must be loaded using the "library" code. To create the bar chart, the main code begins with ggplot following the dataset. Next, the "aes" code is required to assign the x and y axis. To add color to the bars, a fill code is required, and it must be assigned to the correct column. The rest of the codes are the aesthetics of the chart.

R Script:

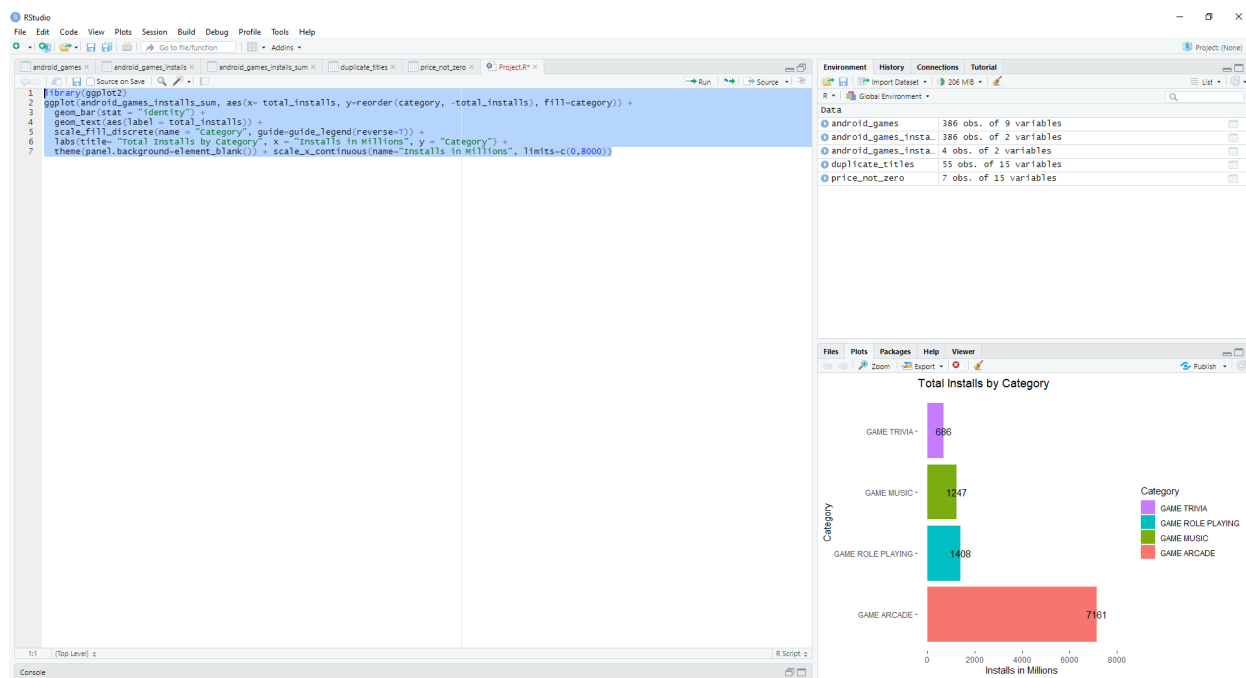


R Script Code:

```
library(ggplot2)
ggplot(android_games_installs_sum, aes(x= total_installs, y=reorder(category, -total_installs),
fill=category)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = total_installs)) +
  scale_fill_discrete(name = "Category", guide=guide_legend(reverse=T)) +
  labs(title= "Total Installs by Category", x = "Installs in Millions", y = "Category") +
  theme(panel.background=element_blank()) + scale_x_continuous(name="Installs in
Millions", limits=c(0,8000))
```

Below is what RStudio looks like and after highlighting the entire code in the script section, the code is ready to be executed by clicking the "run" button. The result is a bar chart shown on the bottom right.

Running the Code in RStudio Script:



RStudio - User Defined Function

A user defined function is highly resourceful when a specific line of code is used multiple times. One code that helped us make the charts look more appealing was running "labs" for our ggplot charts. Below is an example of what a labs function looks like.

- `labs(x="Category", y="Values", title = "Number Ratings by Category")`

By running this code, the title labels of the x and y axis need to be renamed. Otherwise, the chart titles are displayed just like the titles in the column headers of the dataset. The title can be added

to the top of the chart which is required when presenting the data. This way, the viewer can understand the meaning of the chart.

G. Conclusion

In summary, the goal of this R project analysis was to understand consumer preferences for different game categories considering various factors such as the number of installations in Google Play, star ratings, and price. After analyzing and visualizing the data set using RStudio, we discovered that Game Arcade is the most preferable game category across all the factors of our analysis. Game Trivia had the lowest number of installs, whereas Game Music had the lowest star ratings. Game Role Playing had the lowest price, which indicated consumers did not enjoy playing the games within this category compared to the other categories. In summary, we were able to better understand the types of games people enjoyed during the pandemic.

References:

- Clement, J. (2021, November 12). *Mobile gaming market in the United States - statistics & facts*. Data Novia. (2021, December 2021). *DATA MANIPULATION IN R*. Retrieved from Data Novia: <https://www.datanovia.com/en/lessons/rename-data-frame-columns-in-r/>
- Hargrave, M., Somer, A., & Bellucco-Chatham, A. (2021, December 8). *Standard Deviation*. Retrieved from Investopedia: <https://www.investopedia.com/terms/s/standarddeviation.asp>
- History. (2019, June 10). *Video Game History*. Retrieved from History: <https://www.history.com/topics/inventions/history-of-video-games>
- O'Reilly. (2021, December 14). *1. The Very Basics*. Retrieved from O'Reilly: https://learning.oreilly.com/library/view/hands-on-programming-with/9781449359089/ch01.html#_the_r_user_interface
- Schork, J. (2021, December 14). *Remove Rows with NA in R Data Frame (6 Examples) | Some or All Missing*. Retrieved from Statistics Globe: <https://statisticsglobe.com/r-remove-data-frame-rows-with-some-or-all-na>
- Stack Overflow. (2013, June 14). *Find duplicate values in R [duplicate]*. Retrieved from Stack Overflow: <https://stackoverflow.com/questions/16905425/find-duplicate-values-in-r>
- Stack Overflow. (2017, October 14). *R: Deleting rows based on a value in a column from a large data set in R [duplicate]*. Retrieved from Stack Overflow: <https://stackoverflow.com/questions/46574230/r-deleting-rows-based-on-a-value-in-a-column-from-a-large-data-set-in-r>
- Stack Overflow. (2019, March 14). *Stacked bar plot using R and ggplot*. Retrieved from Stack Overflow: <https://stackoverflow.com/questions/55002355/stacked-bar-plot-using-r-and-ggplot?rq=1>
- Statista. (2021, September 9). *Digital Market Outlook: video games revenue change U.S. 2018–2025, by category*. <https://www.statista.com/statistics/460115/digital-games-revenue-growth-category-digital-market-outlook-usa/>
- Statista. (2021, September 9). *Digital video games revenue growth in the United States from 2018 to 2025, by category*. Retrieved from Statista: <https://www.statista.com/statistics/460115/digital-games-revenue-growth-category-digital-market-outlook-usa/>
- ThoughtCo. (2019, April 27). *What Are the First and Third Quartiles?* Retrieved from ThoughtCo: <https://www.thoughtco.com/what-are-first-and-third-quartiles-3126235>