Foundations of Data Science
(UCS548)

Dashboard Submission

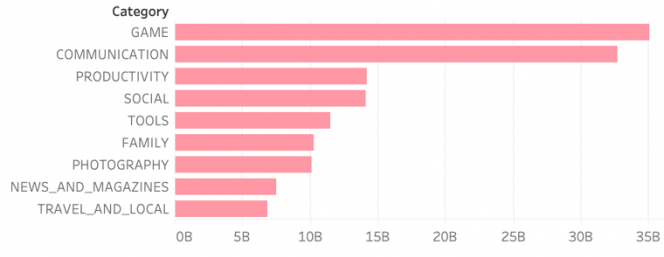
Inferences from Google Play Store Apps Dataset

Submitted By:
Anshh Chaturvedi
102003665
3COE26

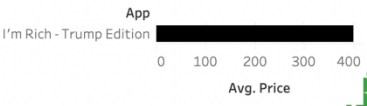
Submitted To:
Dr. Sharad Saxena

Dashboard

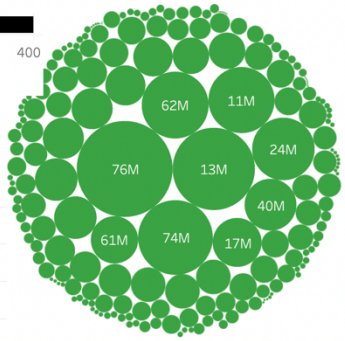
Category vs Number of Installs



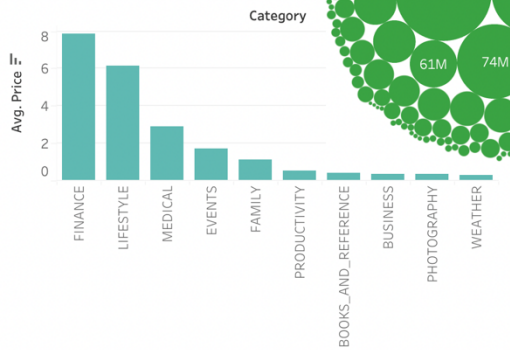
Most Expensive Apps



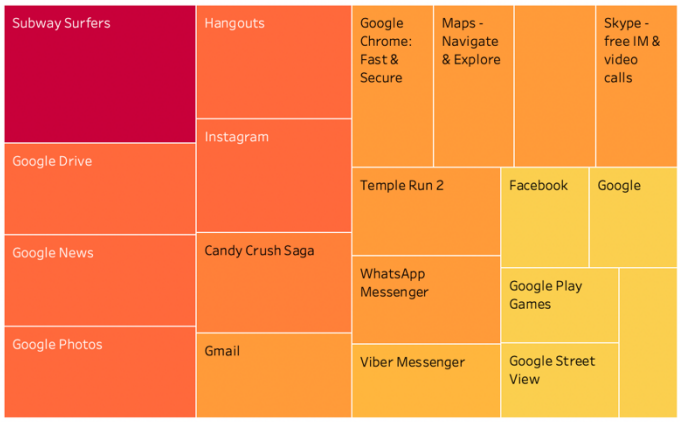
Size vs No. of Installs



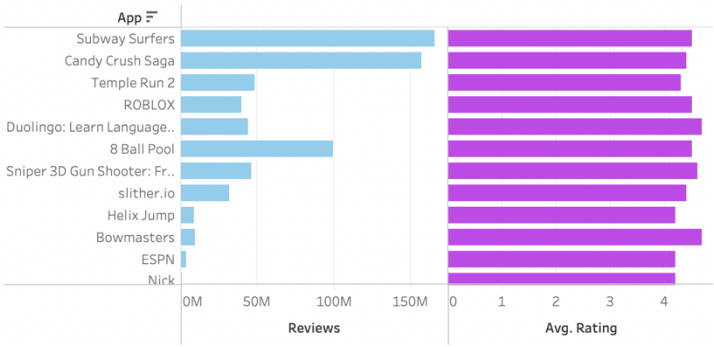
Category vs Avg. Price



Most installed apps



Best Rated Apps



The Datasets

Dataset 1 :

ID	App	Category	Rating	Reviews	Size
1	Photo Editor	NA	4.1	NaN	19M
2	Coloring boo	ART_AND_D	3.9	967	NA
3	U Launcher L	ART_AND_D	4.7	87510	8.7M
4	Sketch - Dra	ART_AND_D	4.5	215644	25M
5	Pixel Draw -	NA	4.3	NaN	2.8M
6	Paper flower	ART_AND_D	4.4	167	5.6M
7	Smoke Effec	ART_AND_D	3.8	178	19M
8	Infinite Paint	ART_AND_D	4.1	36815	NA
9	Garden Colo	NA	4.4	NaN	33M
10	Kids Paint Fr	ART_AND_D	4.7	121	3.1M
11	Text on Phot	ART_AND_D	4.4	13880	28M
12	Name Art Ph	ART_AND_D	4.4	8788	12M
13	Tattoo Name	NA	4.2	44829	20M
14	Mandala Col	ART_AND_D	4.6	4326	NA
15	3D Color Pixe	ART_AND_D	4.4	NaN	37M
16	Learn To Dra	ART_AND_D	3.2	55	2.7M
17	Photo Desigr	NA	4.7	3632	5.5M
18	350 Diy Roor	ART_AND_D	4.5	27	17M
19	FlipaClip - Ca	ART_AND_D	4.3	194216	NA
20	ibis Paint X	ART_AND_D	4.6	NaN	31M
21	Logo Maker -	ART_AND_D	4	450	14M

Dataset 2 :

ID	Installs	Type	Price	Content Rating
1	NA	Free	0	NA
2	5,00,000	NA	0	Everyone
3	50,00,000	Free	NA	Everyone
4	5,00,00,000	Free	0	Teen
5	NA	Free	0	Everyone
6	50,000	Free	0	Everyone
7	50,000	Free	0	Everyone
8	10,00,000	Free	0	Everyone
9	10,00,000	Free	0	Everyone
10	NA	Free	0	NA
11	10,00,000	Free	0	Everyone
12	10,00,000	NA	0	Everyone
13	1,00,00,000	Free	0	Teen
14	1,00,000	Free	NA	Everyone
15	1,00,000	Free	0	NA
16	NA	Free	0	Everyone
17	5,00,000	Free	0	Everyone
18	10,000	Free	0	Everyone
19	50,00,000	NA	0	Everyone

Dataset 3:

ID	Genres	Last Updated	Current Ver	Android Ver
1	NA	January 7, 2018	NA	NA
2	Art & Design;Pretend Play	NA	2.0.0	4.0.3 and up
3	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
4	Art & Design	June 8, 2018	NA	4.2 and up
5	Art & Design;Creativity	June 20, 2018	1.1	NA
6	NA	March 26, 2017	1	2.3 and up
7	Art & Design	April 26, 2018	1.1	4.0.3 and up
8	Art & Design	June 14, 2018	NA	4.2 and up
9	Art & Design	September 20, 2017	2.9.2	NA
10	Art & Design;Creativity	July 3, 2018	2.8	4.0.3 and up
11	Art & Design	October 27, 2017	1.0.4	4.1 and up
12	NA	July 31, 2018	1.0.15	4.0 and up
13	Art & Design	April 2, 2018	NA	4.1 and up
14	Art & Design	June 26, 2018	1.0.4	NA
15	Art & Design	August 3, 2018	1.2.3	2.3 and up
16	Art & Design	June 6, 2018	NaN	4.2 and up
17	Art & Design	July 31, 2018	NA	4.1 and up
18	NA	November 7, 2017	1	2.3 and up
19	Art & Design	August 3, 2018	2.2.5	NA

Now, we have to combine these datasets using R, using library openxlsx and readxl.

```
df1<-read_excel('/Users/anshhchaturvedi/Desktop/Assessment.xlsx')
df1
df2<-read_excel('/Users/anshhchaturvedi/Desktop/Assessment2.xlsx')
df2
df3<-read_excel('/Users/anshhchaturvedi/Desktop/Assessment3.xlsx')
df3
```

Now we merge these into a single excel file, that we've named Assessment_Combined.

```
#Merging divided datasets
df4<-merge(df1,df2,by="ID")
df5<-merge(df4,df3,by="ID")
write.xlsx(df5,'/Users/anshhchaturvedi/Desktop/Assessment_Combined.xlsx')
```

The combined dataset :

ID	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Ratin	Genres	Last Updated	Current Ver	Android Ver
1	Photo Editor (	NA	4.099999999	NaN	19M	NA	Free	0	NA	NA	January 7, 20	NA	NA
2	Coloring book	ART_AND_DES	3.9	967	NA	500000	NA	0	Everyone	Art & Design;F	NA	2.0.0	4.0.3 and up
3	U Launcher Li	ART_AND_DES	4.7	87510	8.7M	5000000	Free	NA	Everyone	Art & Design	August 1, 201	1.2.4	4.0.3 and up
4	Sketch - Draw	ART_AND_DES	4.5	215644	25M	50000000	Free	0	Teen	Art & Design	June 8, 2018	NA	4.2 and up
5	Pixel Draw - N	NA	4.3	NaN	2.8M	NA	Free	0	Everyone	Art & Design;C	June 20, 2018	1.100000000	NA
6	Paper flowers	ART_AND_DES	4.400000000	167	5.6M	50000	Free	0	Everyone	NA	March 26, 20	1	2.3 and up
7	Smoke Effect F	ART_AND_DES	3.8	178	19M	50000	Free	0	Everyone	Art & Design	April 26, 201	1.100000000	4.0.3 and up
8	Infinite Painte	ART_AND_DES	4.099999999	36815	NA	1000000	Free	0	Everyone	Art & Design	June 14, 2018	NA	4.2 and up
9	Garden Colori	NA	4.400000000	NaN	33M	1000000	Free	0	Everyone	Art & Design	September 20	2.9.2	NA
10	Kids Paint Fre	ART_AND_DES	4.7	121	3.1M	NA	Free	0	NA	Art & Design;C	July 3, 2018	2.8	4.0.3 and up
11	Text on Photo	ART_AND_DES	4.400000000	13880	28M	1000000	Free	0	Everyone	Art & Design	October 27, 2	1.0.4	4.1 and up
12	Name Art Pho	ART_AND_DES	4.400000000	8788	12M	1000000	NA	0	Everyone	NA	July 31, 2018	1.0.15	4.0 and up
13	Tattoo Name (	NA	4.2	44829	20M	10000000	Free	0	Teen	Art & Design	April 2, 2018	NA	4.1 and up
14	Mandala Colo	ART_AND_DES	4.599999999	4326	NA	100000	Free	NA	Everyone	Art & Design	June 26, 2018	1.0.4	NA
15	3D Color Pixe	ART_AND_DES	4.400000000	NaN	37M	100000	Free	0	NA	Art & Design	August 3, 201	1.2.3	2.3 and up
16	Learn To Draw	ART_AND_DES	3.2	55	2.7M	NA	Free	0	Everyone	Art & Design	June 6, 2018	NaN	4.2 and up
17	Photo Design	NA	4.7	3632	5.5M	500000	Free	0	Everyone	Art & Design	July 31, 2018	NA	4.1 and up
18	350 Diy Room	ART_AND_DES	4.5	27	17M	10000	Free	0	Everyone	NA	November 7, 1	1	2.3 and up
19	FlipaClip - Car	ART_AND_DES	4.3	194216	NA	5000000	NA	0	Everyone	Art & Design	August 3, 201	2.2.5	NA
20	ibis Paint X	ART_AND_DES	4.599999999	NaN	31M	10000000	Free	0	Everyone	Art & Design	July 30, 2018	5.5.4	4.1 and up
21	Logo Maker - S	ART_AND_DES	4	450	14M	NA	Free	0	NA	Art & Design	April 20, 201	NA	4.1 and up
22	Boys Photo Ec	ART_AND_DES	4.099999999	654	12M	100000	Free	0	Everyone	Art & Design	NA	1.100000000	4.0.3 and up
23	Superheroes \	NA	4.7	7699	4.2M	500000	Free	0	Everyone 10+	Art & Design	July 12, 2018	2.2.6.2	NA
24	Mcqueen Col	ART_AND_DES	NaN	61	7.0M	100000	Free	0	Everyone	Art & Design;A	March 7, 201	1.0.0	4.1 and up
25	HD Mickey Mi	ART_AND_DES	4.7	NaN	23M	50000	Free	NA	Everyone	Art & Design	July 7, 2018	1.1.3	4.1 and up
26	Harley Quinn	ART_AND_DES	4.8	192	6.0M	10000	NA	0	Everyone	Art & Design	April 25, 201	1.5	3.0 and up
27	Colorfit - Drav	ART_AND_DES	4.7	20260	NA	500000	Free	0	Everyone	Art & Design;C	October 11, 2	1.0.8	4.0.3 and up
28	Animated Phc	ART_AND_DES	4.099999999	203	6.1M	NA	Free	0	Everyone	NA	March 21, 20	1.03	4.0.3 and up
29	Pencil Sketch	NA	3.9	136	4.6M	10000	Free	0	Everyone	Art & Design	NA	6	NA

About the combined dataset

This dataset provides details about the performance of various applications that are present on the Google Play Store App.

The dataset contains 10840 tuples and 13 attributes.

The attributes are:

1) App-> Name of the application
2) Category-> Broadly categorises the app on the basis of functionality, for example Art and Design.
3) Rating-> The average rating of the app
4) Reviews-> Number of reviews the app has been given.
5) Size-> Storage required by the app
6) Installs-> Number of times the app has been installed.
7) Type-> Indicates whether the app is free or not.
8) Price-> Indicates the price of the app.
9) Content.Rating-> Age based rating of the app.
10) Genres-> Describes the group of the app.
11) Last.Updated-> Date on which the app was last updated.
12) Current.Ver-> Current version of the app.
13) Android.Ver-> Latest version of the android on which the app is available.

Rating, Reviews, Installs and Price are all integer values.
The rest are string values.

The dataset contains NA and NaN values.

Data cleaning

We have to replace the NA and NaN values.
First, I've changed all the NA and NaN values to NA in R, since excel stores these in string format.

```
df[df=="NA"]<-NA
df[df=="NaN"]<-NA
```

I've replaced the NaN values with mean and the NA values with mode

Function for mode:

```
mymode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

Column names for reference:

```
> colnames(df)
 [1] "ID"           "App"          "Category"     "Rating"       "Reviews"
 [6] "Size"         "Installs"     "Type"         "Price"        "Content.Rating"
[11] "Genres"       "Last.Updated" "Current.Ver"  "Android.Ver"
```

Removing NA values from Category, Rating and Reviews columns:

```
df$Category[is.na(df$Category)]<-mymode(df$Category)
df$Category
df$Rating[is.na(df$Rating)]<-mean(!is.na(df$Rating))
df$Rating
df$Reviews[is.na(df$Reviews)]<-mean(!is.na(df$Reviews))
df$Reviews
```

Some of the values in the Size column were measured in kBs, while most of them were in MB.

```
input <- df$Size
output <- sapply(input, function(x) {
  ifelse(grepl("k$", x), paste0(0.001*as.numeric(sub("(\\d+(?:\\.\\d+)?)k", "\\1", x)), "M"), x)
})
output
```

```
df$Size<-output
df$Size[is.na(df$Size)]<-mymode(df$Size)
df$Size
```

I've replaced the NA size values with the mode, since in this column the mode is "Varies with Device"

Installs column in the dataset was in string format, I've converted them to integer format.

```
#Changing data type of Installs from string to numeric
df$Installs<-as.integer(df$Installs)
class(df$Installs)
df$Installs[is.na(df$Installs)]<-mean(!is.na(df$Installs))
df$Installs
```

Same for Price values

```
#Changing price values to integer from string
df$Price<-as.integer(df$Price)
class(df$Price)
df$Price[is.na(df$Price)]<-mymode(df$Price)
df$Price
```

For Type column :

```
df$Type[is.na(df$Type)]<-mymode(df$Type)
df$Type
```

For Content.Rating, Genres and Last.Updated columns

```
df$Content.Rating[is.na(df$Content.Rating)]<-mymode(df$Content.Rating)
df$Content.Rating
df$Genres[is.na(df$Genres)]<-mymode(df$Genres)
df$Genres
df$Last.Updated[is.na(df$Last.Updated)]<-mymode(df$Last.Updated)
df$Last.Updated
```

For Android.Ver

```
df$Android.Ver[is.na(df$Android.Ver)]<-mymode(df$Android.Ver)
df$Android.Ver
write.xlsx(df,'/Users/anshhchaturvedi/Desktop/Assessment_Final.xlsx')
```

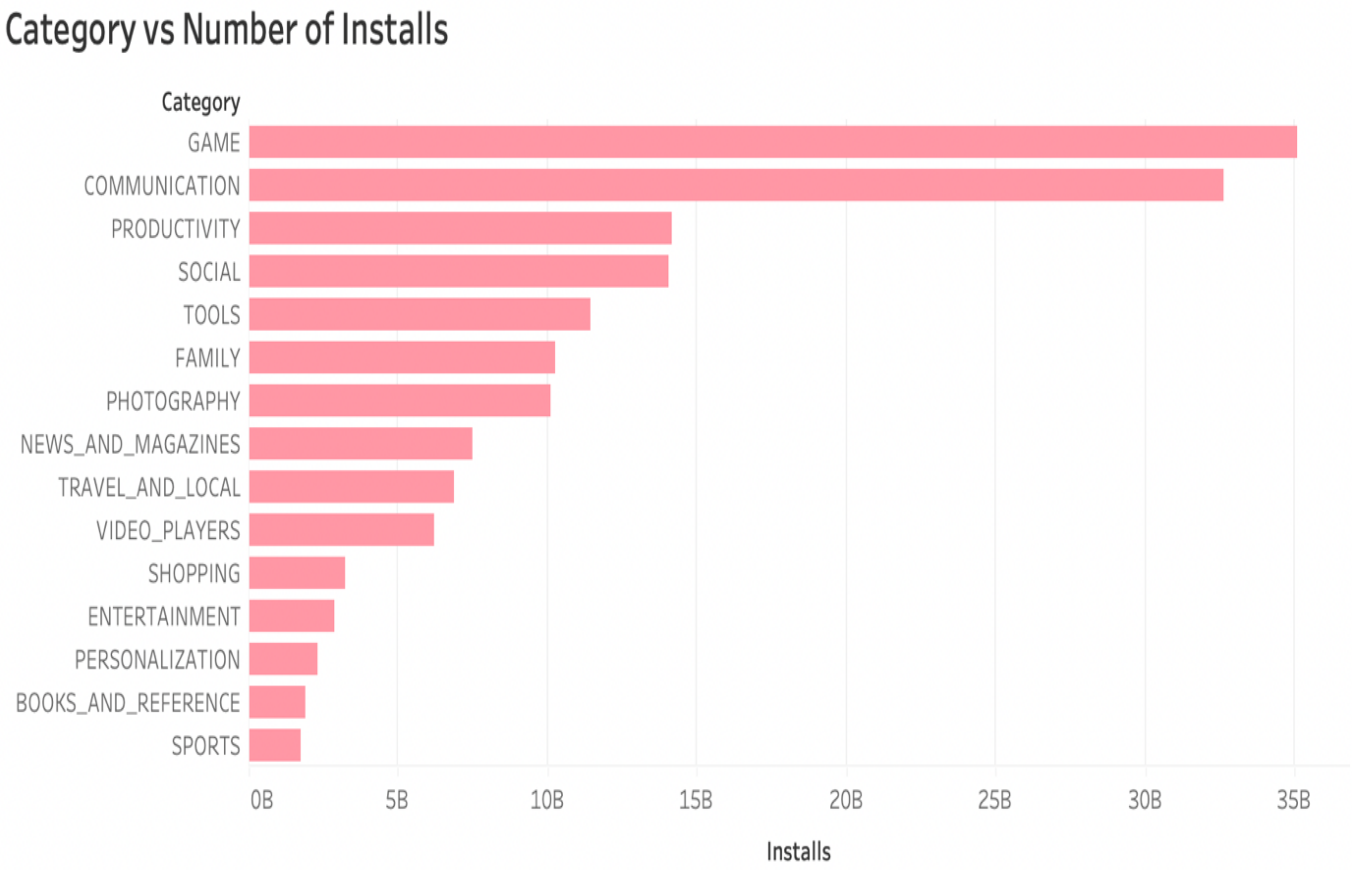
After replacing the NAs, I've written the data to a new excel file called Assessment_Final.

Inferences using Tableau

1) Which are the most popular categories on the Google Play Store?

->The GAME category is the most popular, with 35 Billion+ installations

The top 15 are :



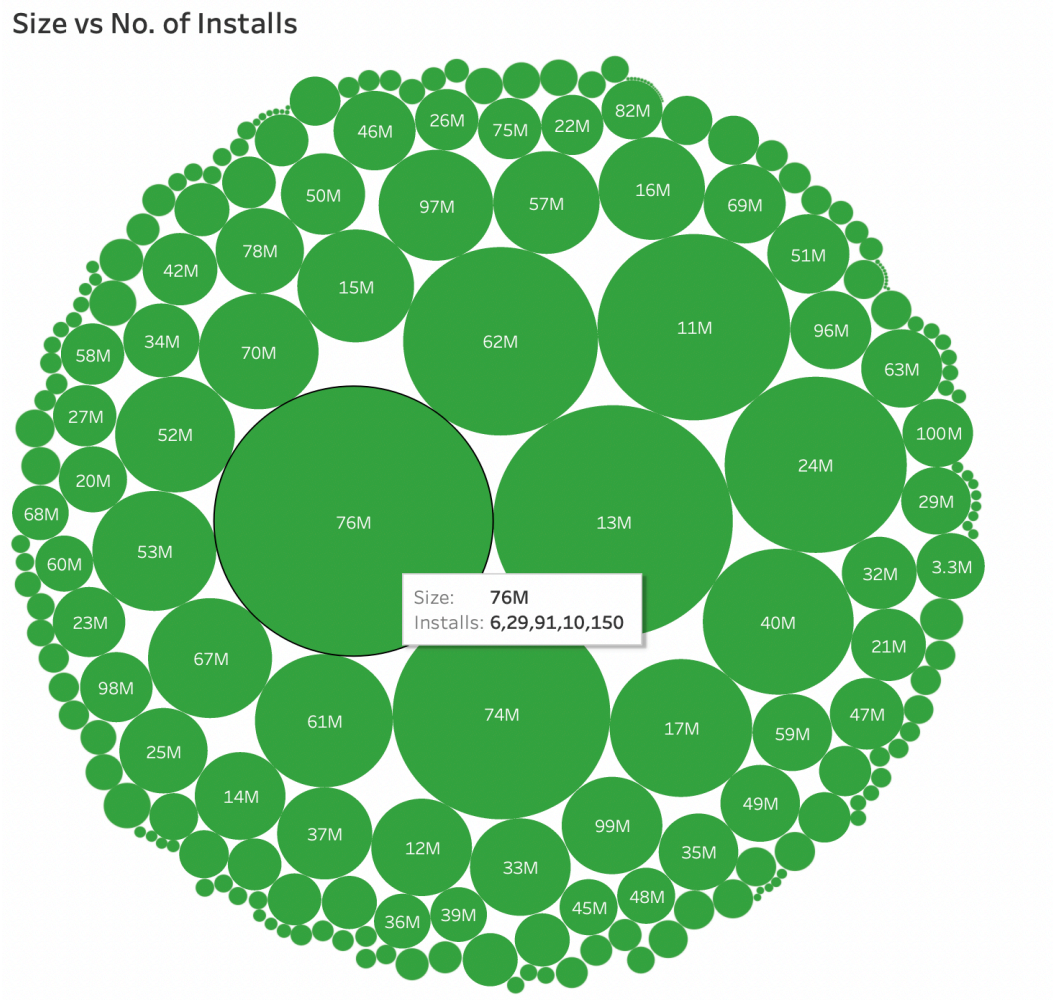
2) Which are the most costliest categories on the Play Store?

->Finance is the costliest category, with an average price of $7.88

The top 10 are:

Q.3) What is the most appropriate app size?

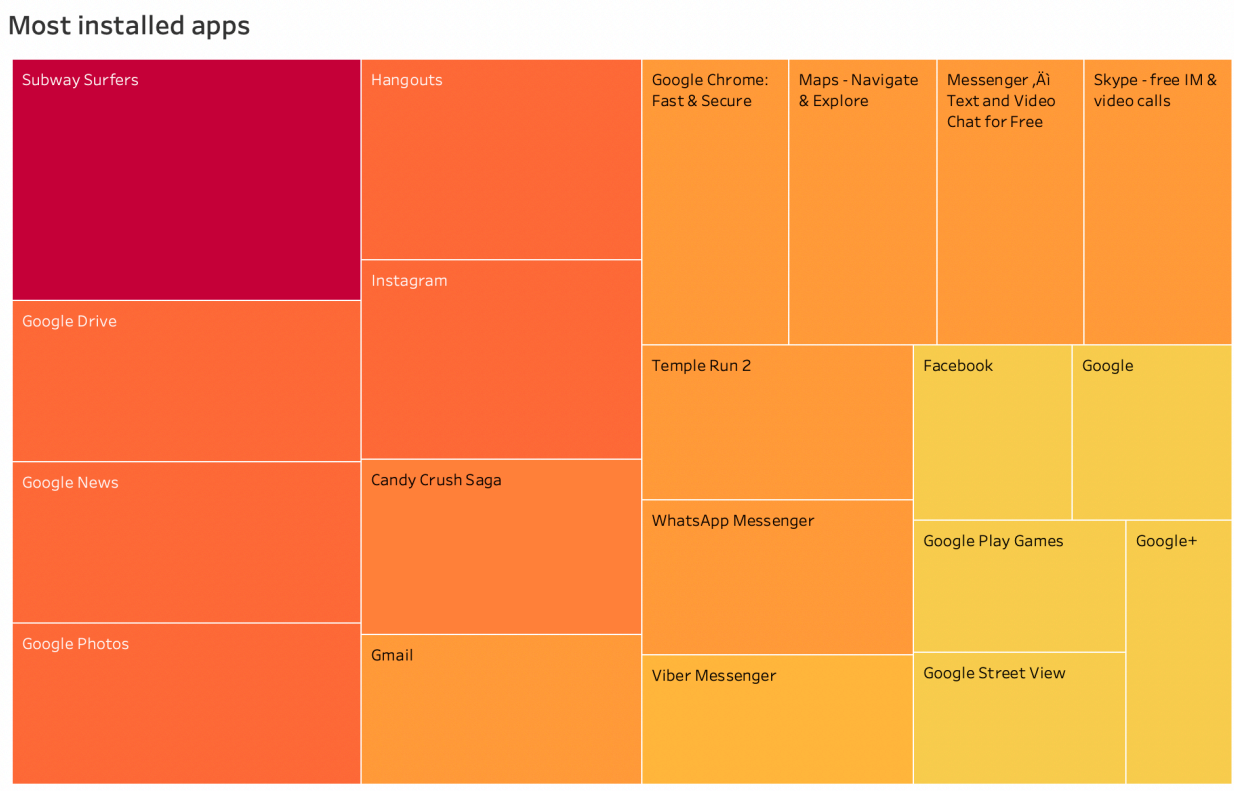
->The most appropriate app size lies between 10-80 MBs



Q.4) Which are the most famous apps in terms of number of installations?

->The most famous app is Subway Surfers

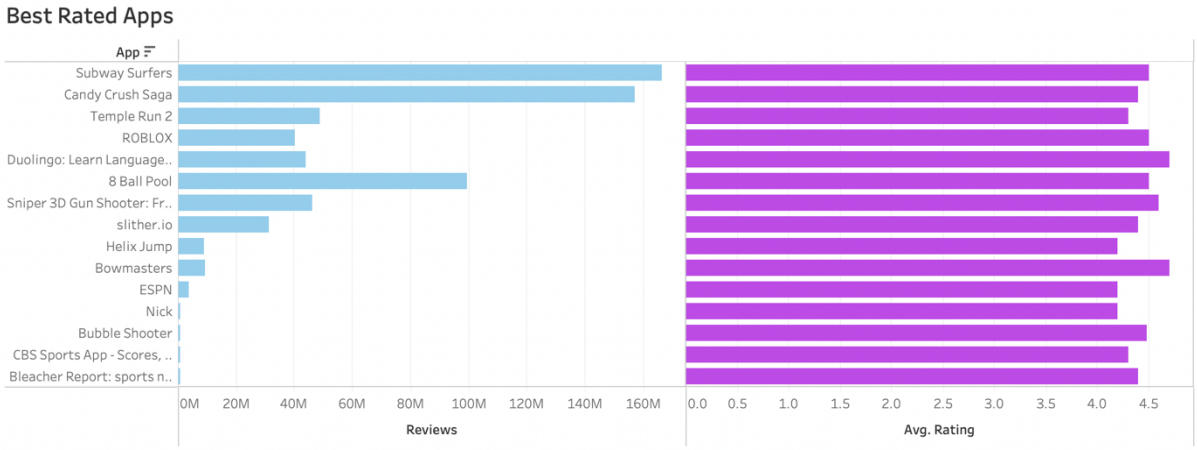
The top 20 are:



Q.5) Which are the best rated apps, combined with number of reviews?

-> The answer is Subway Surfers again!

The top 15 are :



Q.6) Which is the most costly app present on the Play Store? The most expensive app on the Play Store is “I’m Rich- Trump Edition” and it costs $400