GOOGLE PLAY-STORE APPS

TEAM MEMBERS:

NIHAL SHETTY - PES1UG19CS298

NISHANTH M S - PES1UG19CS304

NISHANTH NARENDRA - PES1UG19CS305

NUMAN MOHAMMED ABBAS - PES1UG19CS308

SECTION : E

TABLE OF CONTENTS

ABSTRACT

The dataset being studied is Google Playstore Apps. The dataset contains attributes which help us uncover the secret for being the most successful app in the market. The dataset was cleaned and processed using appropriate methods to ensure hassle free analysis. Exploratory data analysis consisted of various graphs and standardisation of the data , which helped us to obtain a better understanding of our dataset and to decipher the correlation between various attributes. We conducted category wise hypotheses testing to shed some light on categories that have more success in the market i.e. an average rating of greater than 4.25 and mean installs of at least 10000000.

INTRODUCTION

Google Play Store houses the applications required for the TWO BILLION Android users across the globe! The ever growing operating system is the most popular in the world and offers a lot of promise to mobile app developers. This dataset is obtained from Kaggle. While many public datasets provide Apple App Store data, there are not many counterpart datasets available for Google Play Store apps anywhere on the web. iTunes App Store page deploys a nicely indexed appendix-like structure to allow for simple and easy web scraping. On the other hand, Google Play Store uses sophisticated modern-day techniques (like dynamic page load) using JQuery making scraping more challenging.

It requires more than a good application to capture this market. It requires smart investments. With the power of statistical analysis , we are set to find out what kind of application offers promise and is worth investing over. The dataset has enormous potential to drive app-making businesses to success. We have tried our best to attain a deeper understanding of the market via this dataset and draw insights for developers to work on and capture the Android market!

DATASET

The Google Play-Store data set has 13 attributes, 9 Categorical , 2 continuous and 2 discrete.
The App attribute is the name of the app.
The Category attribute is a division for apps with similar characteristics.
The Rating attribute gives information of how much people like the app in the scale of 1-5.
The Review attribute specifies the number of feedbacks a user sends to the developer demanding improvement of the app or appreciating their work.
The Size attribute the app size.Its given in MB and KB.
The Installs attribute gives the number of people who have downloaded app around the world.

The Type attribute specifies if the app is a paid app or free for download.
The Price attribute specifies what is the cost of the paid apps.
The ContentRating attribute specifies which age group the app falls under.
The Genre attributes like Category attribute specifies the similarity in purpose/use between other apps .
The Last Updated attribute specifies the date of last update.
The Current ver attribute specifies the version of the app.
The Android ver attribute Specifies the minimum version of the android for the app to function and be available for download.

PRE-PROCESSING : DATA CLEANING

The data set has 13 columns and 10841 rows.
The data set has missing values in the form of 'varies with the device', 'nan', 'unrated' and ' '.

The App attribute has repeated values.The repeated values are dropped.

The category attribute has 16 missing values.These missing values are replaced by the mode of consecutive 20 elements if it falls in between that position.

The ContentRating has 13 missing values.If the app falls in the category of 'Dating' then the missing value is replaced by 'Mature 17' or else 'Everyone'.

Size attribute has 1695 missing values.The missing value is replaced by'0M' in the beginning. The postfix M and K is replaced and multiplied with their value that is, M is replaced and multiplied by 1 and K is replaced and multiplied by 1/1000.The Missing value is replaced with the average size of the apps for that given category.Then it is converted from categorical to numerical.

The 'current ver' attribute has 1460 missing values.Since it's not important for getting insights we drop the entire column.

The 'Android ver' attribute has 1363 missing values and the postfix 'and up'.
The missing values are first replaced by ‘0.0’ and the postfix is removed and the values are converted to integer.The missing values are then replaced with their median.

The Rating of the app has 1474 missing values.Since the rating follows skewed distribution, it's replaced with its median.

The Review attribute has 29 missing values and cannot be interpreted and therefore replaced by 0.

We consider all apps having installs less than 5000 has outliers since it will give misleading insights.
An app with 1 install will be given a 5 star rating and therefore leads to misleading insights.

We don't consider rating values as outliers because they play an important role in giving proper insights.
2 rows with the wrong type of data are dropped.

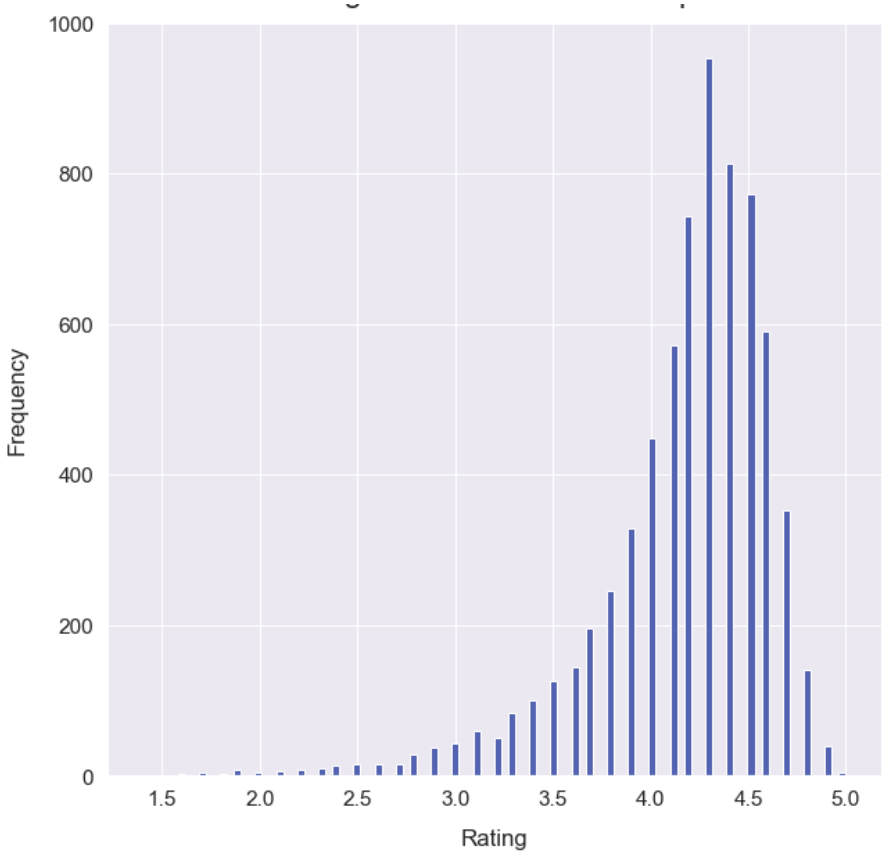
after cleaning the missing values and removing duplicate values and outliers
The data set consists of 12 columns and 6976 rows.

EXPLORATORY DATA ANALYSIS

- **VISUALISATION**

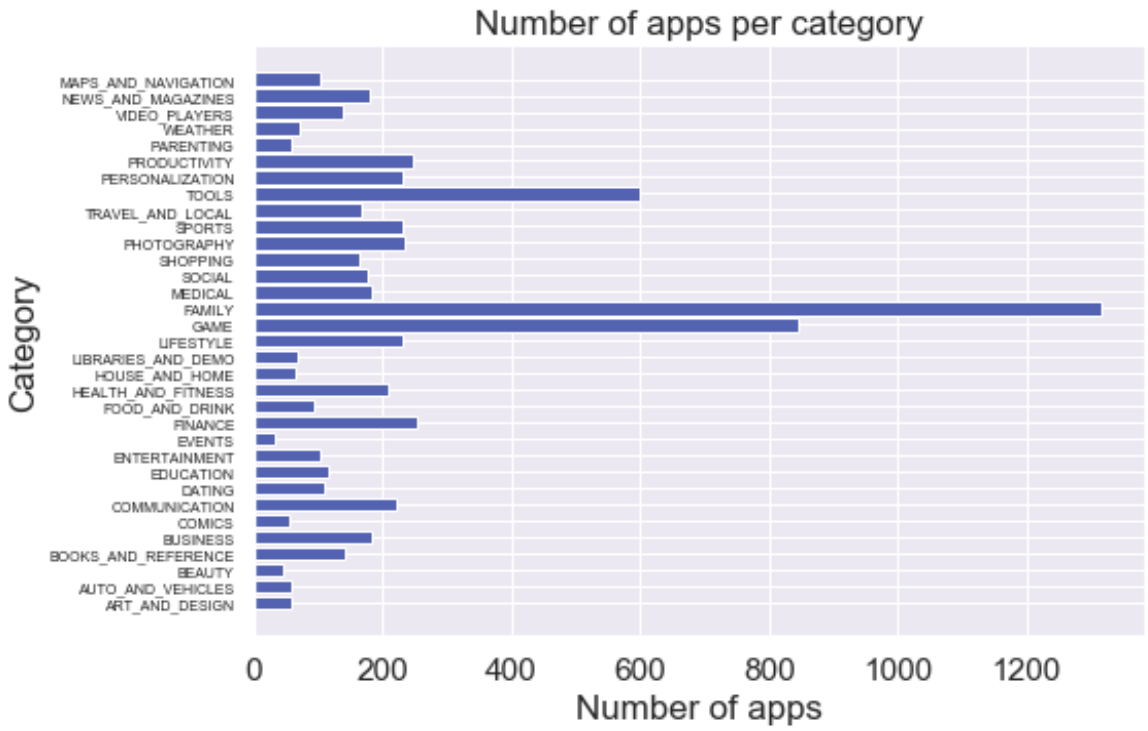
The visualisation step is used to graphically represent the dataset's attributes in various ways in an attempt to find relations between the attributes and make insightful inferences.

1. Rating Distribution - Bar Graph



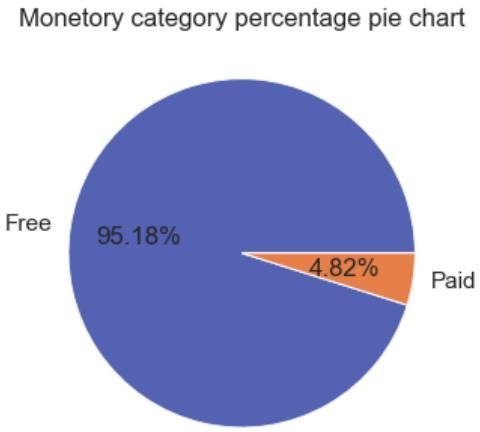
The graph shows that among all the apps in the data set, the rating 4.3 was the highest. It also shows that most of the apps in the data set have a rating above 3.0 .
Making the data set slightly biased towards high rated apps or most apps in general are rated above 3.0 . Either of these conclusions can be true.

2. App Distribution based on category - Horizontal Bar Graph



The graph depicts a relation between the number of apps on google play store differentiated on the basis of the category the app is in.
We can say that the play store has a majority of family, game and tool based apps.
This says that most companies of developers work on family, game or tool apps possibly in hopes for better profits but that might not necessarily be the case.

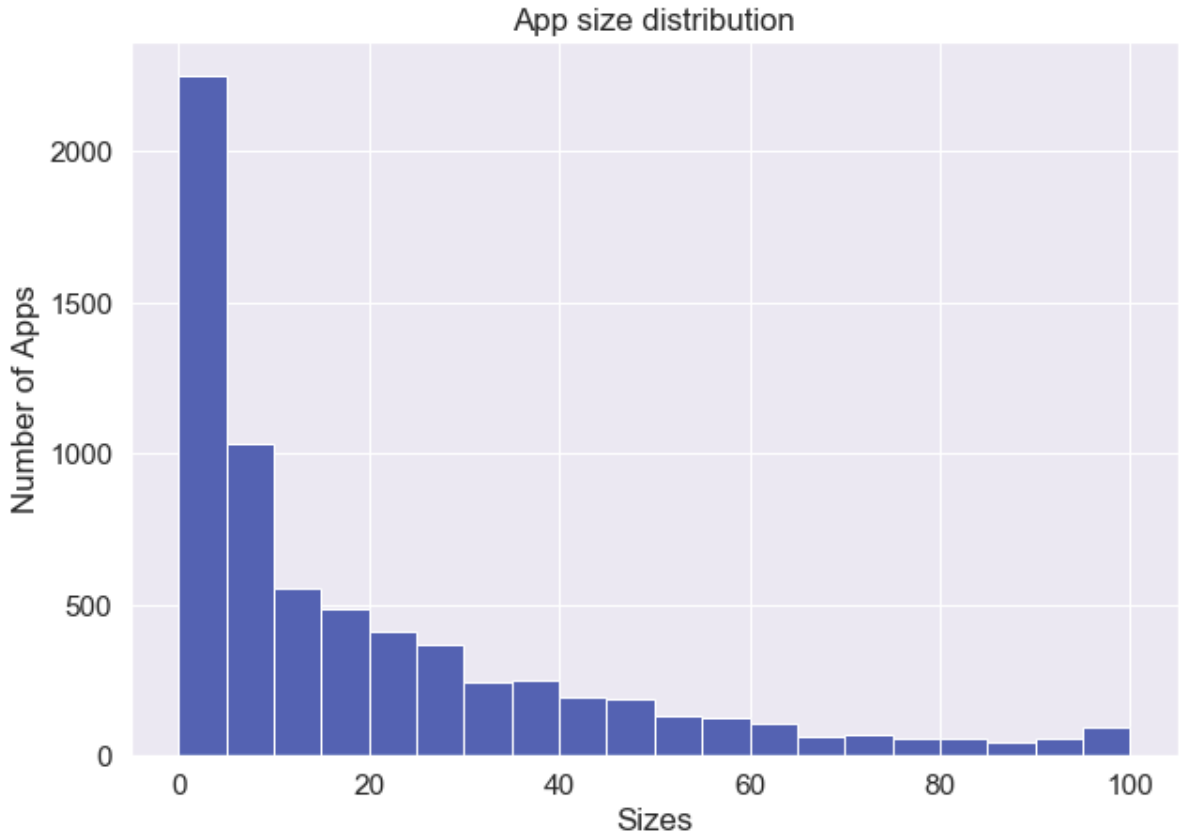
3. <u>App Distribution based on Price - Pie Chart</u>



The pie chart shows the distribution of apps between the free and the paid sector.
This would lead us to believe that a vast majority of the apps on the store are free and yet many companies on play store make their income that way.

This could be because there is nothing to lose for a user to try out a free app since they can try and uninstall the app without any monetary loss unlike paid apps. Paid apps can't be installed just to try since if the user does not like the app, they have already spent money on it.
The likely way that the developers still profit over free apps are the in-app purchases that get them to spend money even though they are using a free to install app.

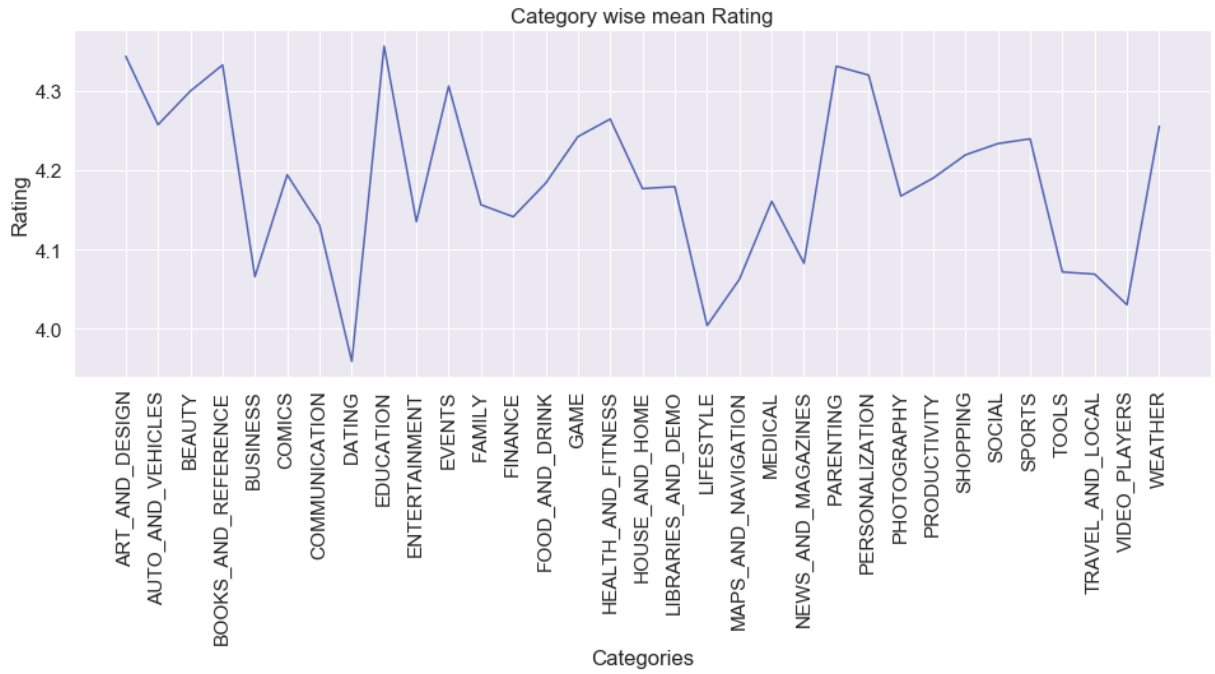
4. App Distribution based on Size(MB) - Histogram



This histogram shows the distribution of apps with respect to the file size of the app.
We can see that a majority of the apps in our data set are below 40.0 MB in size.
The highest accumulation of apps being below 10.0 MB.
This might mean that most apps on the store also accommodate people who do not have space to spare on their phones. Making the decision to try an app or install an app less cumbersome or worrisome for the users as the app takes an insignificant space on their phone.

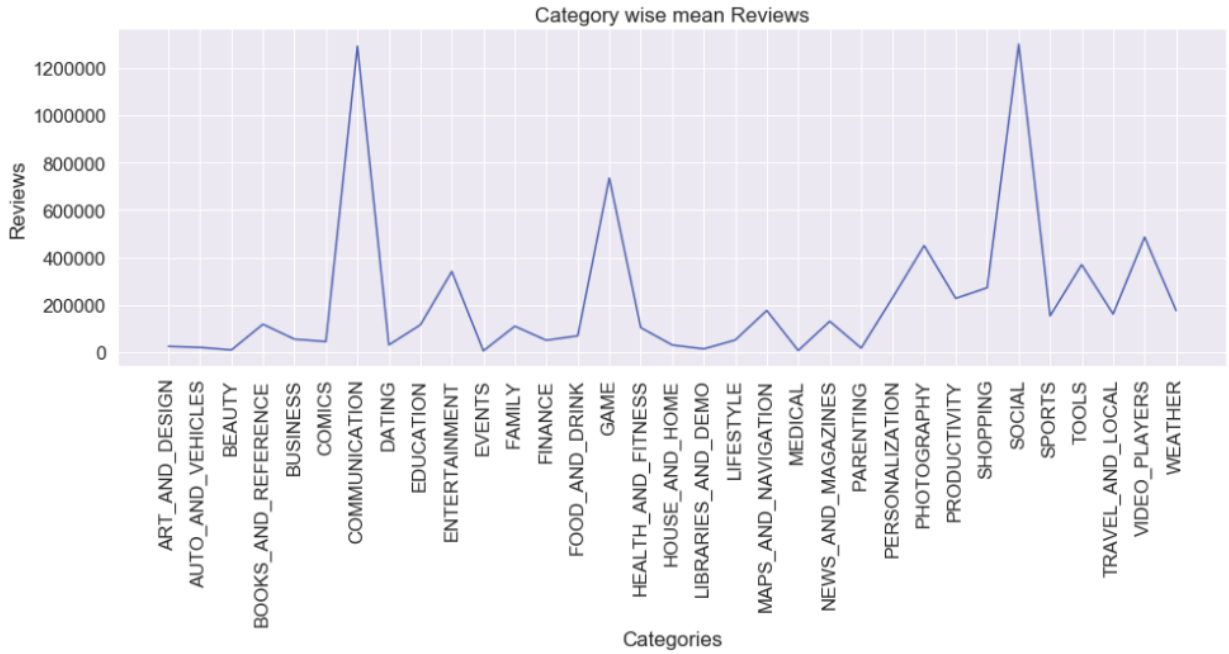
The following graphs are all line graphs that plot out various stats based on category / content rating / monetary type:

a. Mean Rating



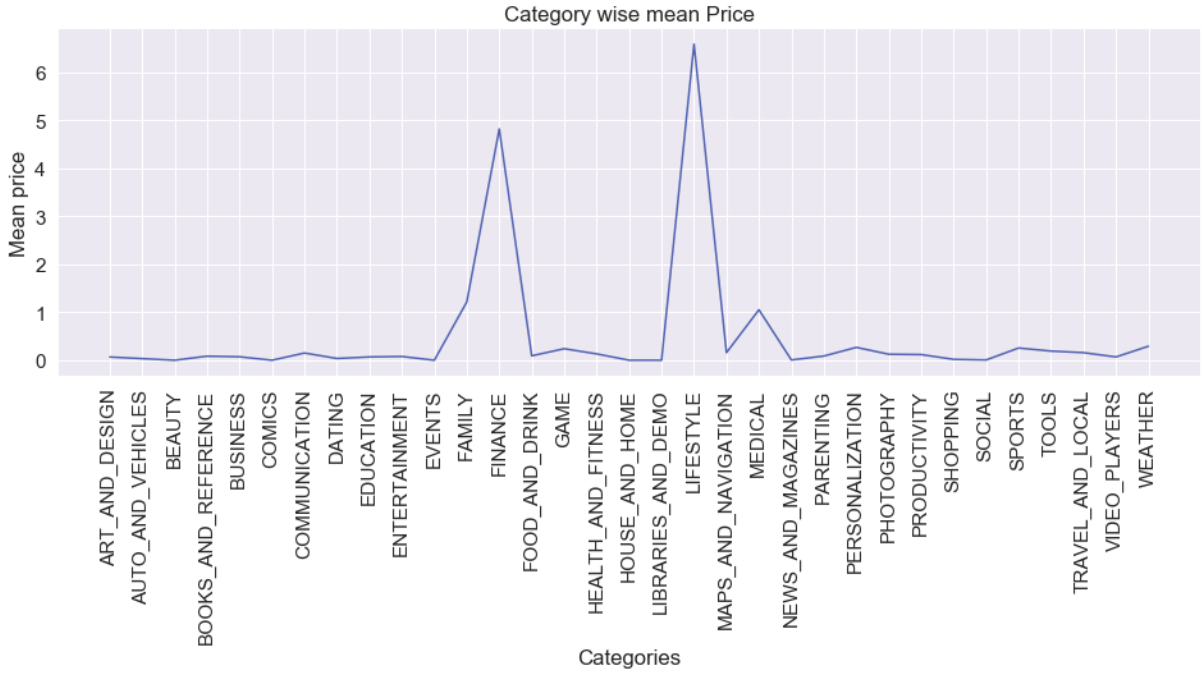
Insights: Categories with the highest average rating- Education, Events, Books and Reference, Art & Design, Parenting and Personalisation.
This shows that categories dealing with learning, academics or cultural aspects such as literature and art are very well rated in the app store.

b. Mean Reviews



Insights: Categories with highest average reviews - Communication, Games, Social
This shows that categories with a social aspect increase the user - developer feedback.

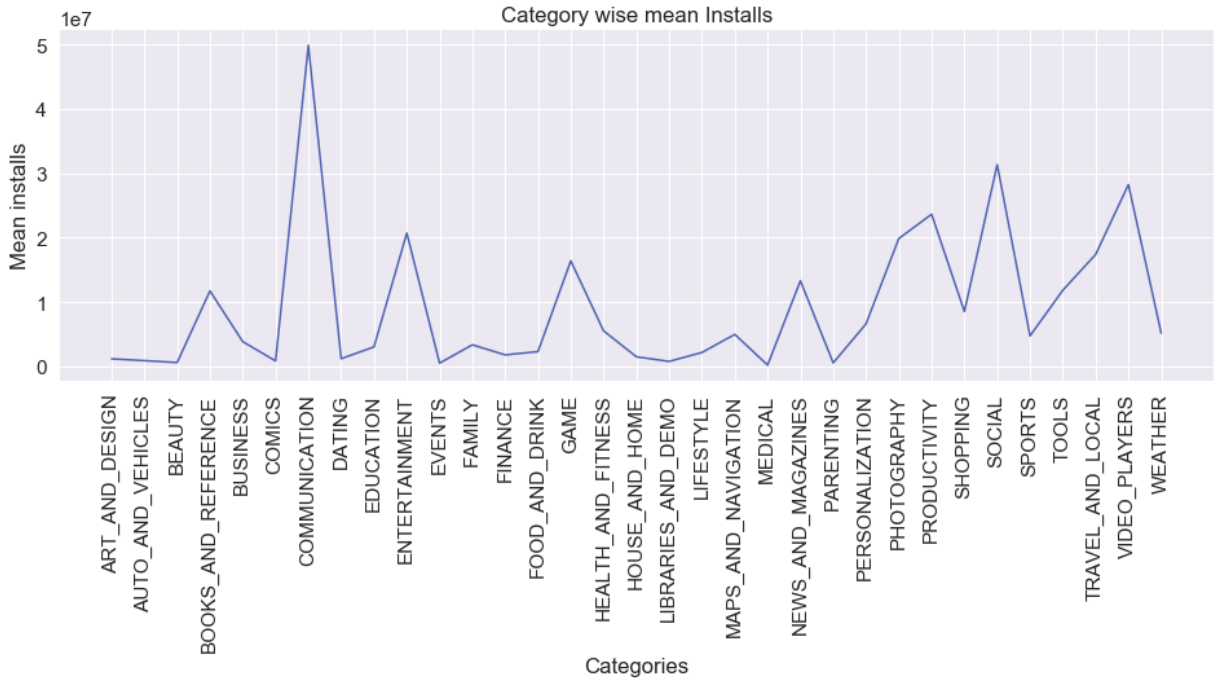
c. <u>Mean Pricing</u>



Insights: Categories with highest mean price - Finance and Lifestyle.
This shows that apps related to monetary transactions are priced higher than others.

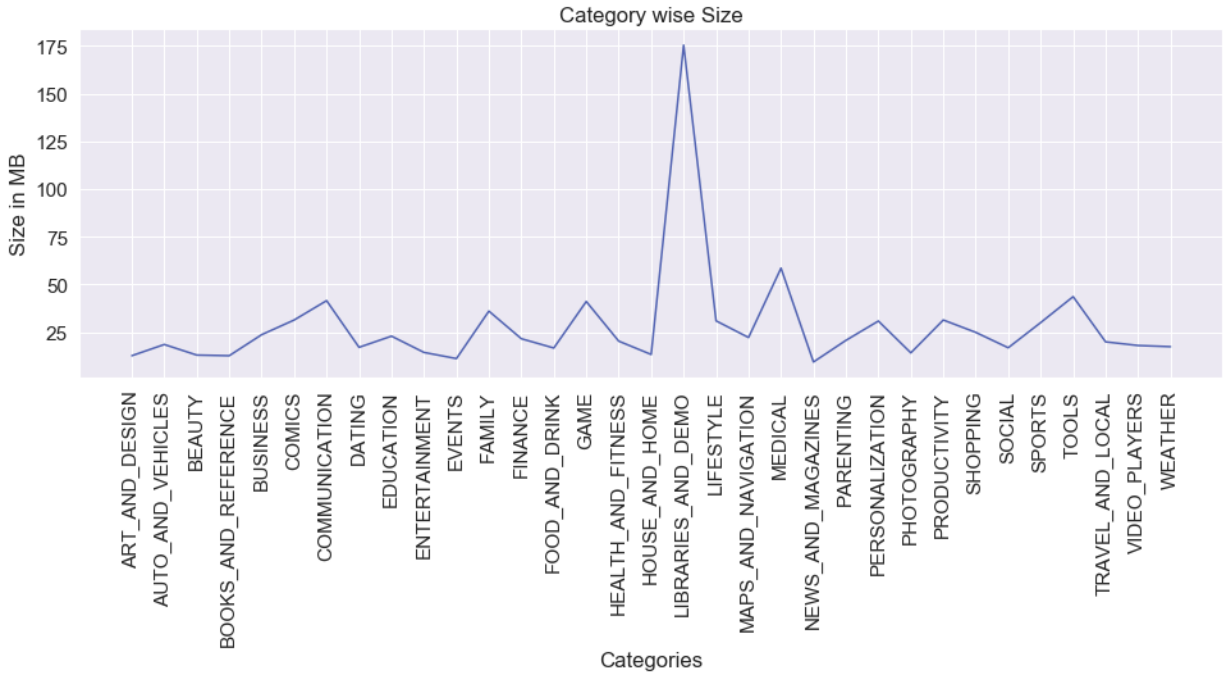
Most probably due to the higher level of security these apps offer.

d. Mean Installs



Insights: Categories with highest average installs - Communication and Social.
This tells us that most installed apps on average in the app store are communication based apps, meaning most people installing apps prioritize communication related apps.

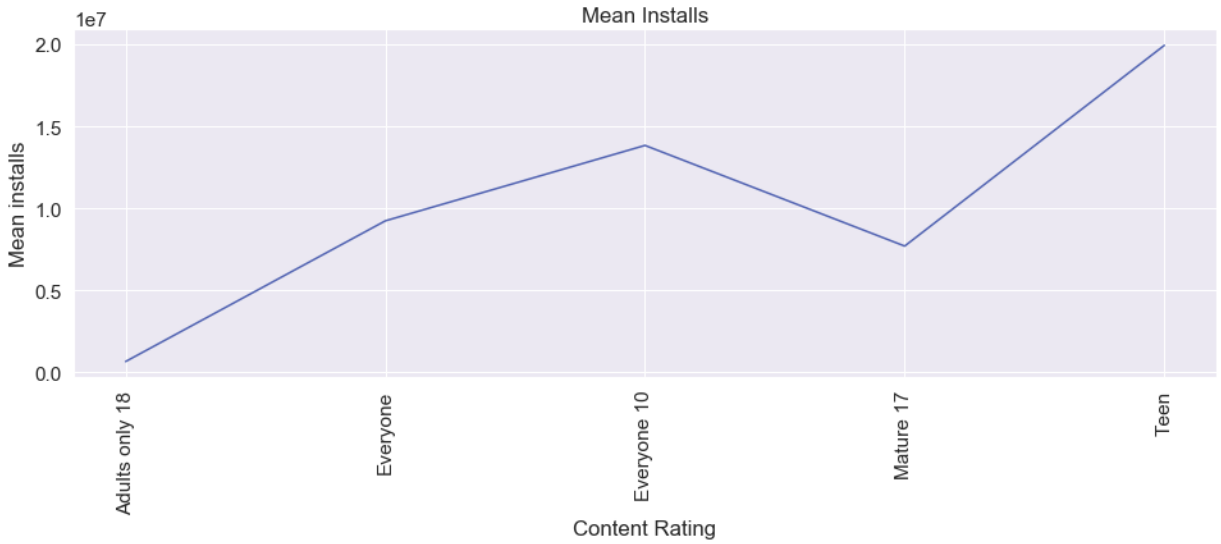
e. Size



Insights: Category with highest size - Libraries.

This is possible because library related apps have a huge amount of data to store and have to keep track of a huge database.

f. <u>Mean Installs</u>

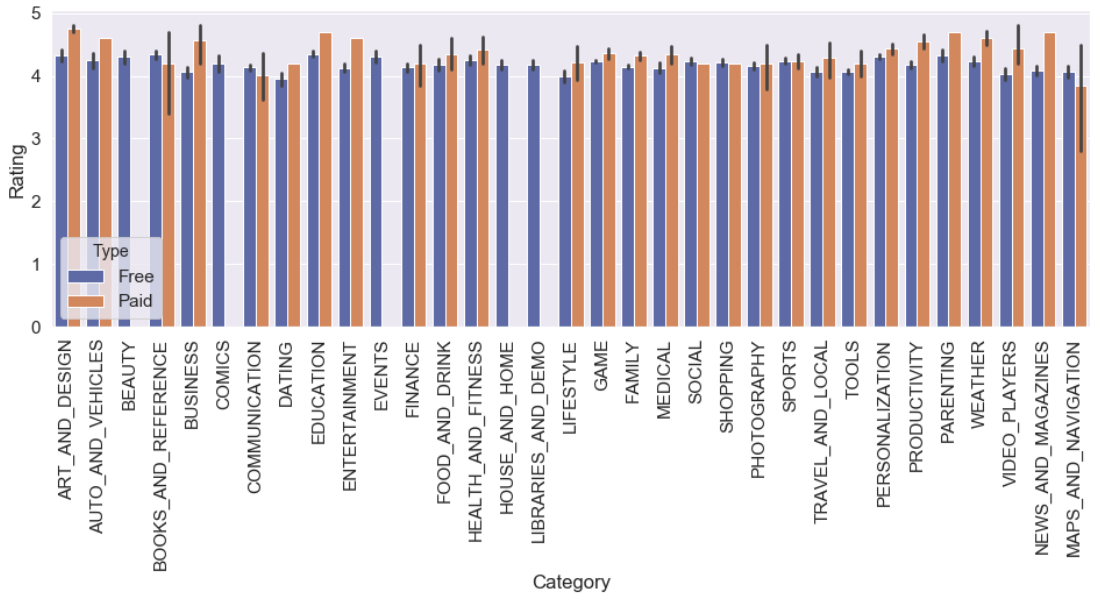


Insights: The teen content rating group has the highest installs.

We can conclude from this that targeting teens with an app will get the app more installs.

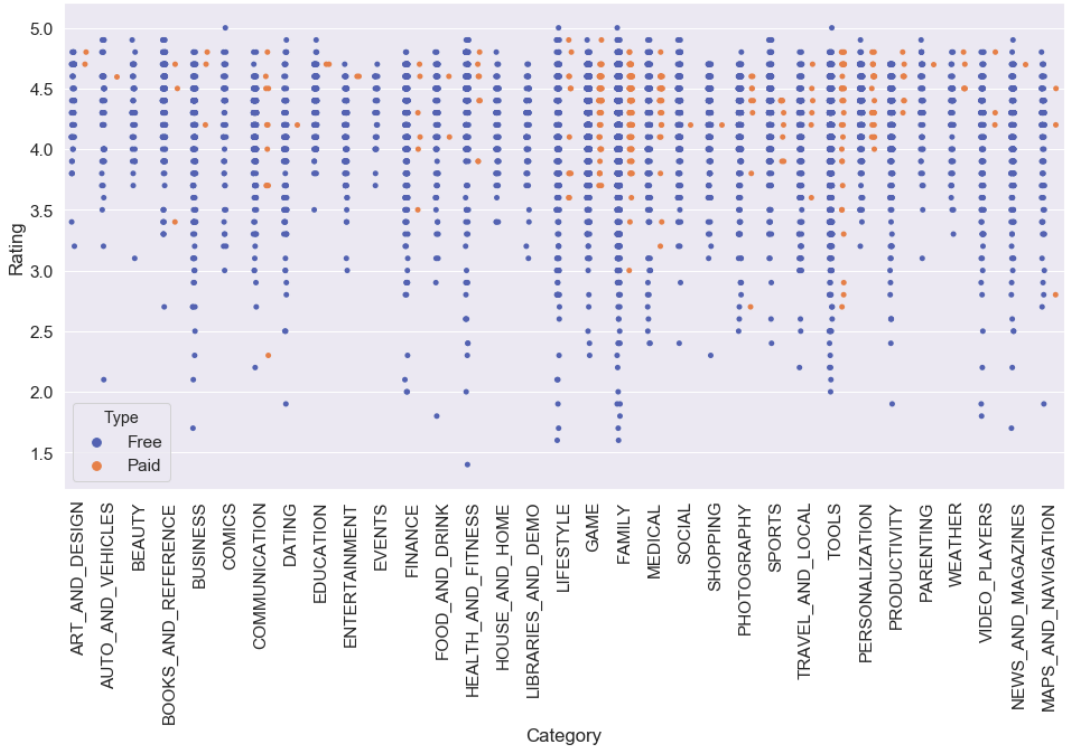
This might also be because there are a larger number of apps with a content rating of teen.

5. Rating - Bar Plot (Based on Monetary Type)



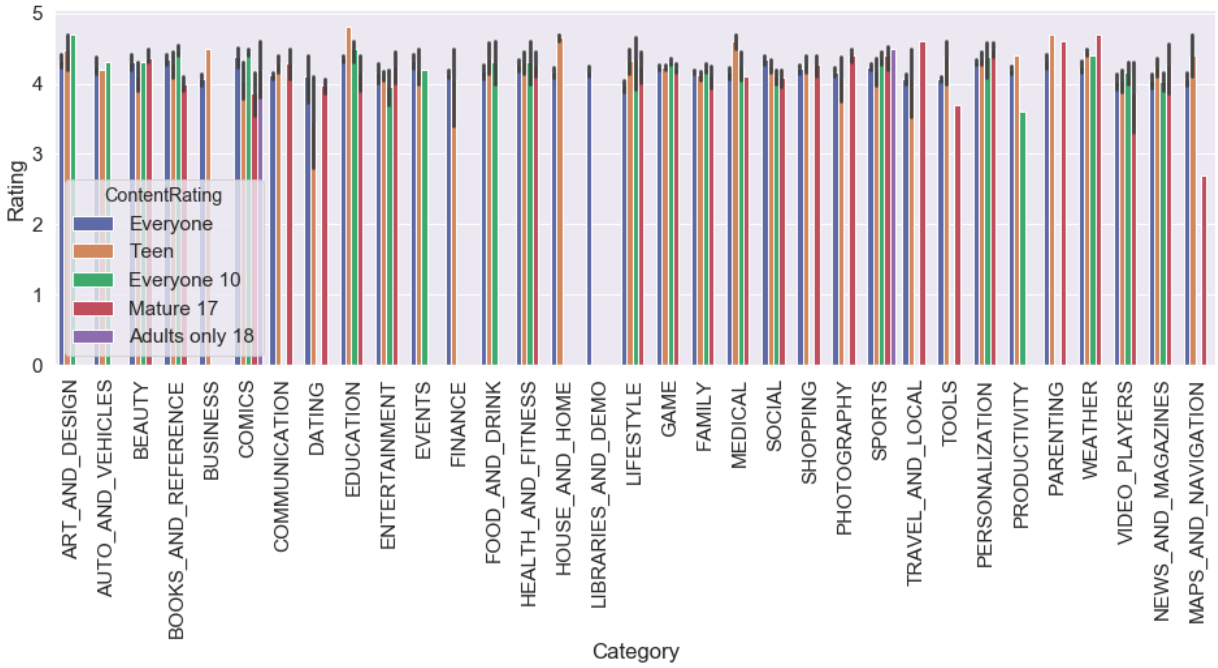
Insights: The graph shows that paid apps are generally rated better than free apps in most categories but not by a huge margin.

6. Rating - Strip Plot



Insights: Taking into consideration the magnitude of data points for free and paid apps, paid apps seem to have a higher average rating compared to free apps that have a larger spectrum.

7. Rating - Bar plot (Based on Content Rating)



Insights: The “adult” rating is sparsely seen incorporated in any categories while the “teen”, “everyone”, and “everyone10” seem to be the most common content rating among all categories.

- **STANDARDISATION AND NORMALISATION**
  Standardization is done in order to put different variables on the same scale in order to make comparisons between them. Standardization rescales data to have a mean of 0 and standard deviation of 1. It is calculated by subtracting the mean of the variable from the raw score and dividing by the standard deviation of the variable. The scores after standardization, referred to as “Z scores” or “Normalized Scores” indicate how many standard deviations the value of the raw score is away from the mean. If the normalized score of the rating of a particular app happened to be 2, it indicates that the rating of the app happened to be 2 standard deviations above the mean rating.

Standardization of Rating variable

	0
0	-0.156077
1	-0.598677
2	1.171722
3	0.729122
4	0.286522
...	...
6971	-0.819977
6972	0.286522
6973	-0.819977
6974	0.729122
6975	0.729122

```
Before Standardization:
[[4.1]
 [3.9]
 [4.7]
 ...
 [3.8]
 [4.5]
 [4.5]]
Mean: 4.170528, StandardDeviation: 0.451876
After Standardization:
[[-0.15607729]
 [-0.59867696]
 [ 1.17172171]
 ...
 [-0.81997679]
 [ 0.72912204]
 [ 0.72912204]]
```

Standardization of Reviews variable

	0
0	-0.139002
1	-0.138625
2	-0.098317
3	-0.038636
4	-0.138625
...	...
6971	-0.138665
6972	-0.139058
6973	-0.138519
6974	-0.139058
6975	0.046442

```
Before Standardization:
[[   159]
 [   967]
 [ 87510]
 ...
 [  1195]
 [    38]
 [398307]]
Mean: 298595.931766, StandardDeviation: 2147003.185489
After Standardization:
[[-0.13900163]
 [-0.13862529]
 [-0.09831654]
 ...
 [-0.13851909]
 [-0.13905798]
 [ 0.04644197]]
```

Similarly, Standardization is done for Installs, Size and Price.

- After Standardization of all variables:

After standardization

	Photo Editor & Candy Camera & Grid & ScrapBook	Coloring book moana	U Launcher Lite – FREE Live Cool Themes Hide Apps	Sketch - Draw & Paint	Pixel Draw - Number Art Coloring Book	Paper flowers instructions	Smoke Effect Photo Maker - Smoke Editor	Infinite Painter	Garden Coloring Book	Kids Paint Free - Drawing Fun	...	Cardio-FR	Frim: get new friends on local chat rooms	Fr Agnel Ambarnath
Rating	-0.156077	-0.598677	1.171722	0.729122	0.286522	0.507822	-0.819977	-0.156077	0.507822	1.171722	...	0.286522	-0.377377	0.065223
Reviews	-0.139002	-0.138625	-0.098317	-0.038636	-0.138625	-0.138998	-0.138993	-0.121929	-0.132652	-0.139019	...	-0.139044	-0.097862	-0.139021
Installs	-0.171617	-0.163698	-0.090980	-0.090980	-0.170162	-0.170970	-0.170970	-0.155619	-0.155619	-0.171617	...	-0.171617	-0.090980	-0.171697
Size	-0.145758	-0.203421	-0.264544	-0.076563	-0.332586	-0.300295	-0.145758	-0.030433	0.015697	-0.329126	...	0.580793	-0.364877	-0.214954
Price	-0.050847	-0.050847	-0.050847	-0.050847	-0.050847	-0.050847	-0.050847	-0.050847	-0.050847	-0.050847	...	-0.050847	-0.050847	-0.050847

5 rows × 6976 columns

Now comparisons can be made about the apps for each variable. The Coloring book moana app has a rating which is 0.59 standard deviations **below** the mean. Meanwhile, the Garden coloring book app has a rating which is 0.5 standard deviations **above** the mean. The Z scores/Normalized Scores are available for every app. Since a large number of the apps were free, those apps were priced 0.05 standard deviations below the mean as can be seen above with all of them having the same Z score.

- Data after Standardization

	App	Category	Rating	Reviews	Size	Installs	Type	Price	ContentRating	Genres	Last Updated
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	-0.156077	-0.139002	-0.145758	-0.171617	Free	-0.050847	Everyone	Art & Design	07-Jan-18
1	Coloring book moana	ART_AND_DESIGN	-0.598677	-0.138625	-0.203421	-0.163698	Free	-0.050847	Everyone	Art & Design;Pretend Play	15-Jan-18
2	U Launcher Lite – FREE Live Cool Themes Hide Apps	ART_AND_DESIGN	1.171722	-0.098317	-0.264544	-0.090980	Free	-0.050847	Everyone	Art & Design	01-Aug-18
3	Sketch - Draw & Paint	ART_AND_DESIGN	0.729122	-0.038636	-0.076563	-0.090980	Free	-0.050847	Teen	Art & Design	08-Jun-18
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	0.286522	-0.138625	-0.332586	-0.170162	Free	-0.050847	Everyone	Art & Design;Creativity	20-Jun-18
...	...	...	...	...	...	...	...	...	...	...	...
10830	News Minecraft.fr	NEWS_AND_MAGAZINES	-0.819977	-0.138665	-0.338352	-0.170162	Free	-0.050847	Everyone	News & Magazines	20-Jan-14
10831	payermonstationnement.fr	MAPS_AND_NAVIGATION	0.286522	-0.139058	-0.251858	-0.171697	Free	-0.050847	Everyone	Maps & Navigation	13-Jun-18
10832	FR Tides	WEATHER	-0.819977	-0.138519	6.347069	-0.170162	Free	-0.050847	Everyone	Weather	16-Feb-14
10836	Sya9a Maroc - FR	FAMILY	0.729122	-0.139058	0.246348	-0.171697	Free	-0.050847	Everyone	Education	25-Jul-17
10840	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	0.729122	0.046442	-0.145758	-0.010182	Free	-0.050847	Everyone	Lifestyle	25-Jul-18

	Rating	Reviews	Size	Installs	Price
count	6.976000e+03	6.976000e+03	6.976000e+03	6.976000e+03	6.976000e+03
mean	-1.955622e-16	-8.148426e-18	-1.629685e-17	-8.148426e-18	2.037106e-18
std	1.000072e+00	1.000072e+00	1.000072e+00	1.000072e+00	1.000072e+00
min	-6.131173e+00	-1.390757e-01	-3.648769e-01	-1.716974e-01	-5.084742e-02
25%	-3.773771e-01	-1.388359e-01	-3.279727e-01	-1.709702e-01	-5.084742e-02
50%	2.865224e-01	-1.357450e-01	-2.264862e-01	-1.636984e-01	-5.084742e-02
75%	7.291220e-01	-1.086978e-01	4.164845e-03	-9.098031e-02	-5.084742e-02
max	1.835621e+00	3.626437e+01	1.100622e+01	1.598780e+01	2.697217e+01

Observations:
It can be clearly seen here that the mean of all the variables is now 0 and the standard deviation of all the variables is 1. All the numerical variables have been standardized. Now we can compare the normalized score (Z score) of each app for every variable with respect to their original mean and standard deviation.

Mean rating: 4.170528, Standard deviation of rating: 0.451876
The app with the **maximum** rating has a rating which is 1.835 standard deviations **above** the mean rating.
The app with the **minimum** rating has a rating which is 6.13 standard deviations **below** the mean rating.

Mean review count: 298595.931766, Standard deviation of review: 2147003.185489
The app with the **most** reviews has a review count which is 36.26 standard deviations **above** the mean review count.
The app with the **least** reviews has a review count which is 0.139 standard deviations **below** the mean review count.

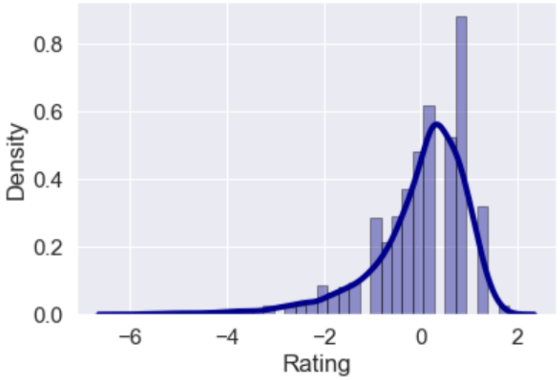
Mean size: 31.638862, Standard deviation of size: 86.711065
The app with **maximum** size has a size which is 11 standard deviations **above** the mean size.
The app with **minimum** size has a size which is 0.365 standard deviations **below** the mean size.

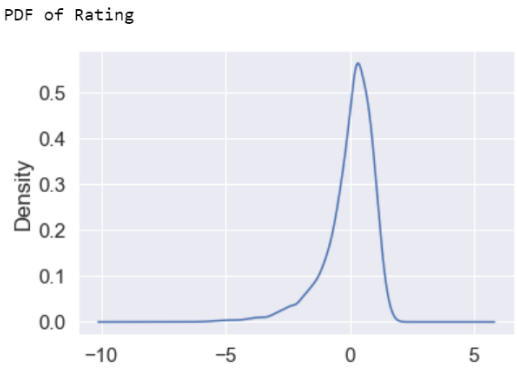
Mean install count: 10630117.545872, Standard deviation of installs: 61882813.703192
The **most** installed app has an installation count which is 15.98 standard deviations **above** the mean install count.
The **least** installed app has an installation count which is 0.17 standard deviations **below** the mean install count.

Mean price: 0.752653, Standard deviation of price: 14.802196
The **most** expensive app has a price which is 26.9 standard deviations **above** the mean price.
The **least** expensive app (free apps) has a price which is 0.05 standard deviations **below** the mean price (exactly what we saw in the beginning).

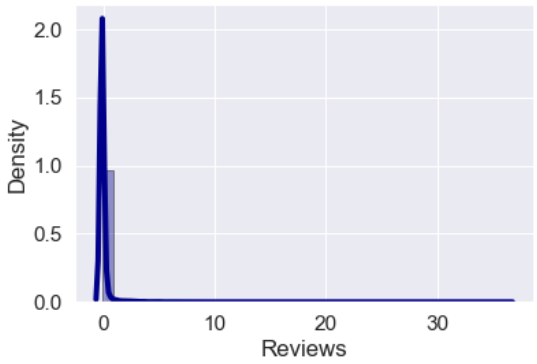
- **Probability Density Functions (PDF)**

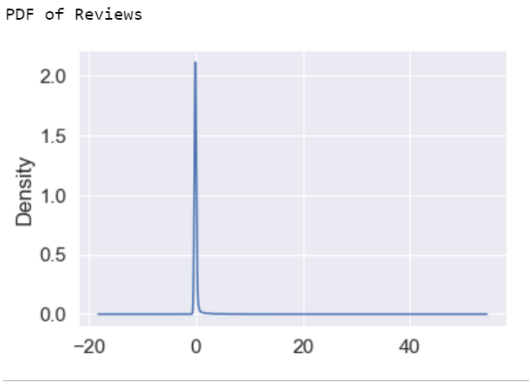
1. PDF of Rating



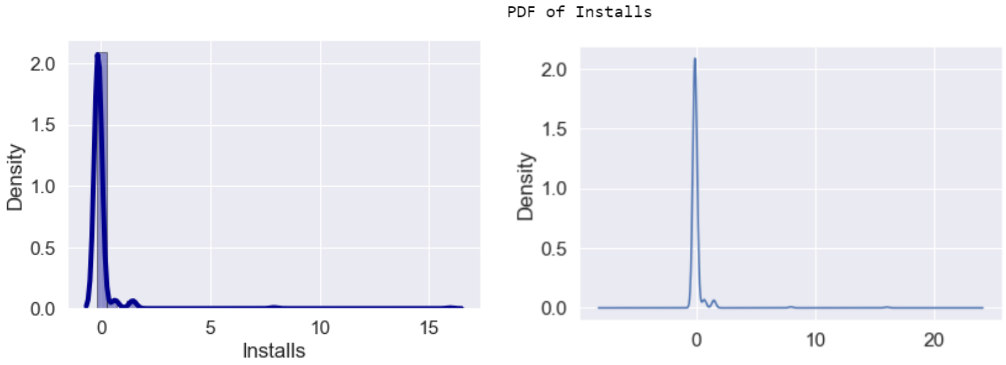


2. PDF of Reviews

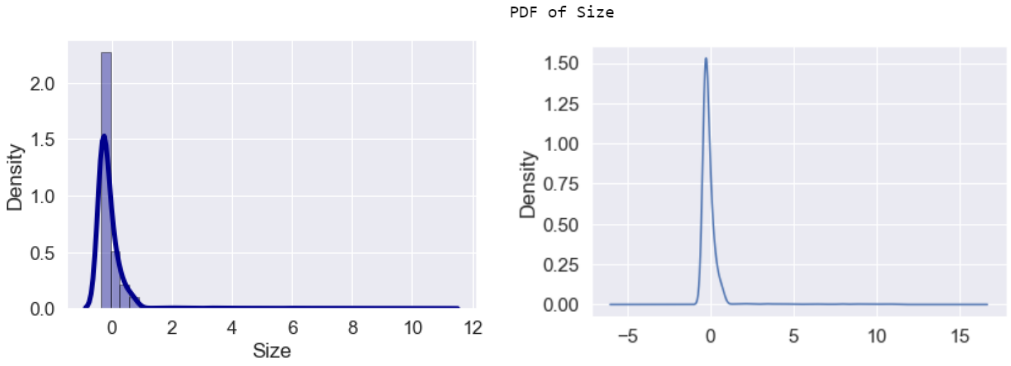




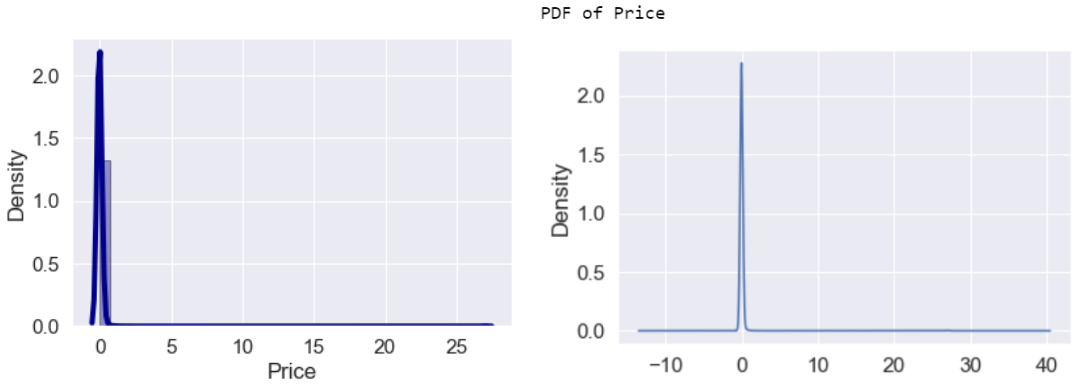
3. PDF of Installs



4. PDF of Size



5. PDF of Price



PDF measures the rates at which probabilities accumulate. Probability Density is probability per unit length. The PDF's of all the variables after standardization are centered at 0 which is what we would expect as the mean of all the variables is now 0. The PDFs replicate a standard normal with mean=0 and unit variance. The area under any region of the curve gives the probability or the likelihood that the variable falls within that particular interval.

- **CORRELATION ANALYSIS**

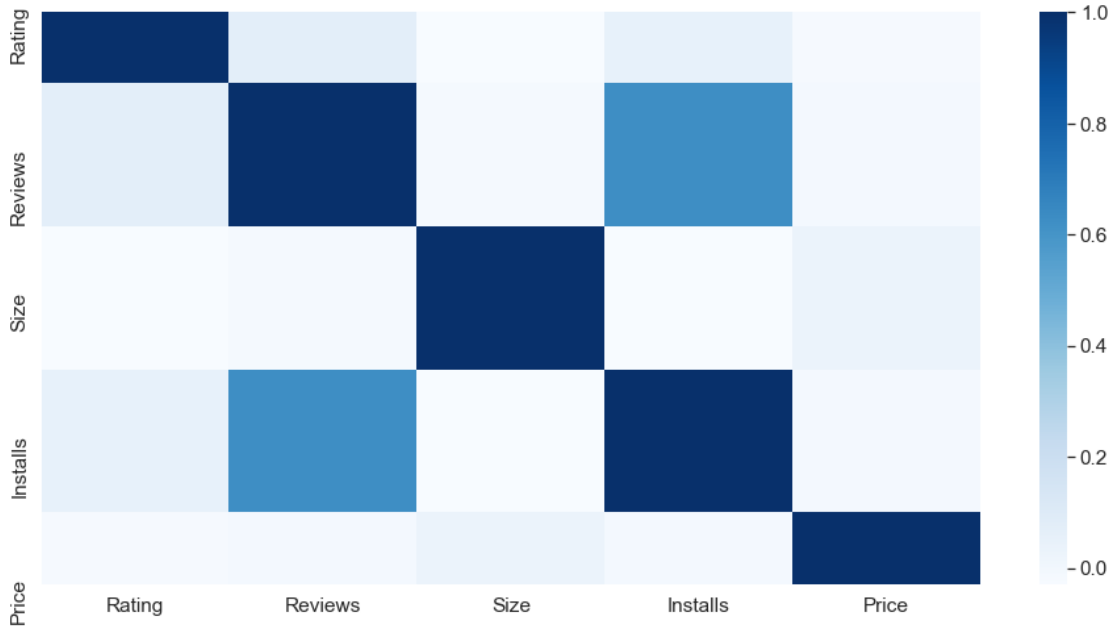
  Standardized values were used for the correlation analysis to get accurate results.

  1. An analysis over all numerical attributes across the dataset

     Correlation table :

	Rating	Reviews	Size	Installs	Price
Rating	1.000000	0.071629	-0.030292	0.052575	-0.018192
Reviews	0.071629	1.000000	-0.010437	0.626313	-0.006914
Size	-0.030292	-0.010437	1.000000	-0.027800	0.033202
Installs	0.052575	0.626313	-0.027800	1.000000	-0.008689
Price	-0.018192	-0.006914	0.033202	-0.008689	1.000000

     Heatmap :



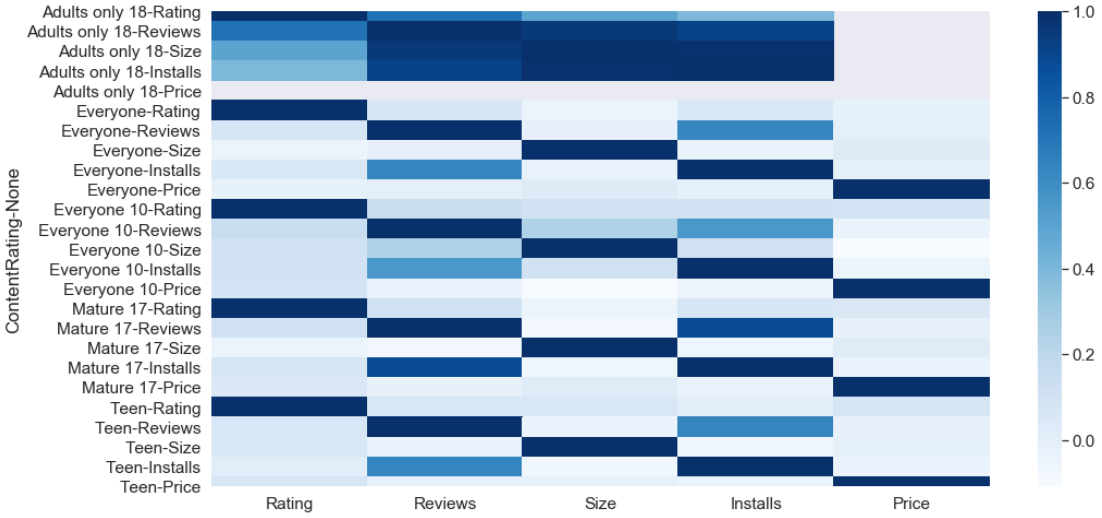
Inference : There is only one strong correlation , between ‘Reviews’ and ‘Installs’ .
This supports the fact and intuitive reasoning that the number of Reviews for an app is higher if the number of Installs is high.

  2. Correlation analysis on data grouped by Content Rating.

Correlation table :

ContentRating		Rating	Reviews	Size	Installs	Price
Adults only 18	Rating	1.000000	0.721995	0.495382	0.397360	NaN
	Reviews	0.721995	1.000000	0.958698	0.921821	NaN
	Size	0.495382	0.958698	1.000000	0.993996	NaN
	Installs	0.397360	0.921821	0.993996	1.000000	NaN
	Price	NaN	NaN	NaN	NaN	NaN
Everyone	Rating	1.000000	0.073088	-0.039607	0.059121	-0.020888
	Reviews	0.073088	1.000000	-0.016712	0.632241	-0.007372
	Size	-0.039607	-0.016712	1.000000	-0.028448	0.034321
	Installs	0.059121	0.632241	-0.028448	1.000000	-0.009137
	Price	-0.020888	-0.007372	0.034321	-0.009137	1.000000
Everyone 10	Rating	1.000000	0.153958	0.107544	0.104072	0.089713
	Reviews	0.153958	1.000000	0.243173	0.555103	-0.033606
	Size	0.107544	0.243173	1.000000	0.112289	-0.108470
	Installs	0.104072	0.555103	0.112289	1.000000	-0.041862
	Price	0.089713	-0.033606	-0.108470	-0.041862	1.000000
Mature 17	Rating	1.000000	0.120361	-0.045101	0.085507	0.053962
	Reviews	0.120361	1.000000	-0.073283	0.887419	-0.025492
	Size	-0.045101	-0.073283	1.000000	-0.053562	0.022802
	Installs	0.085507	0.887419	-0.053562	1.000000	-0.028383
	Price	0.053962	-0.025492	0.022802	-0.028383	1.000000
Teen	Rating	1.000000	0.065825	0.065790	0.016300	0.082943
	Reviews	0.065825	1.000000	-0.033531	0.640126	-0.025130
	Size	0.065790	-0.033531	1.000000	-0.065755	-0.010489
	Installs	0.016300	0.640126	-0.065755	1.000000	-0.031509
	Price	0.082943	-0.025130	-0.010489	-0.031509	1.000000

Heatmap:



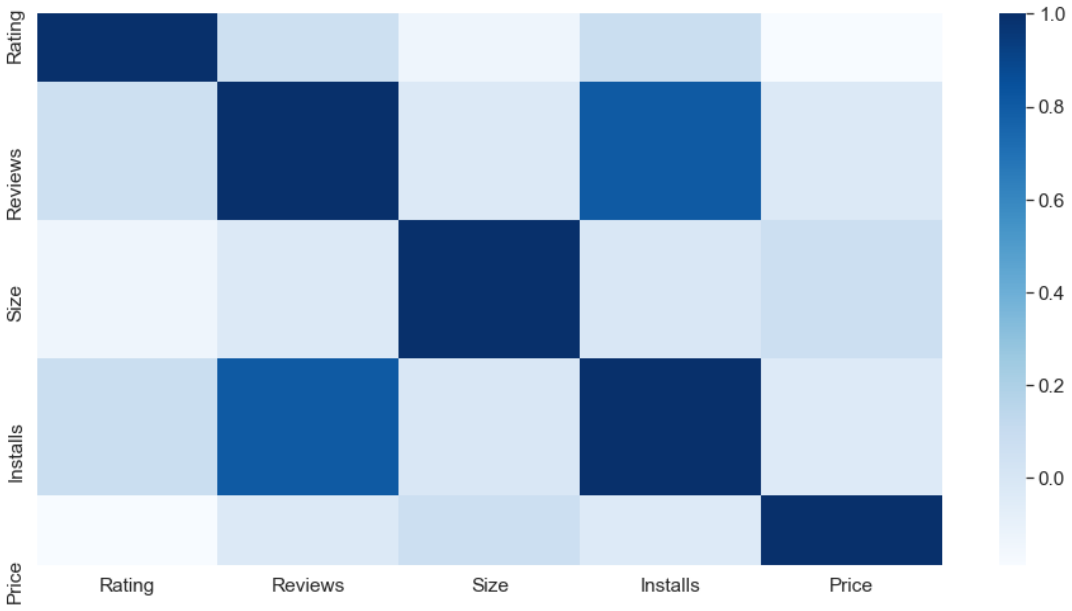
Inference : Though there are multiple correlations under “Adults only 18”, it is misleading because there are only 3 applications with such a field for content rating. Every other ‘Content Rating’ has only one strong correlation as before - Between “Installs” and “Reviews”. The attribute “Price” has NaN as their Pearson correlation coefficient as all the values are the same. This means their standard deviation is 0 , therefore Correlation is not defined.

3. Correlation analysis on data grouped by Type : Paid

   Correlation Table :

	Rating	Reviews	Size	Installs	Price
Rating	1.000000	0.063995	-0.133566	0.082925	-0.188921
Reviews	0.063995	1.000000	-0.024296	0.805224	-0.022883
Size	-0.133566	-0.024296	1.000000	-0.004883	0.075780
Installs	0.082925	0.805224	-0.004883	1.000000	-0.034282
Price	-0.188921	-0.022883	0.075780	-0.034282	1.000000

   Heatmap :



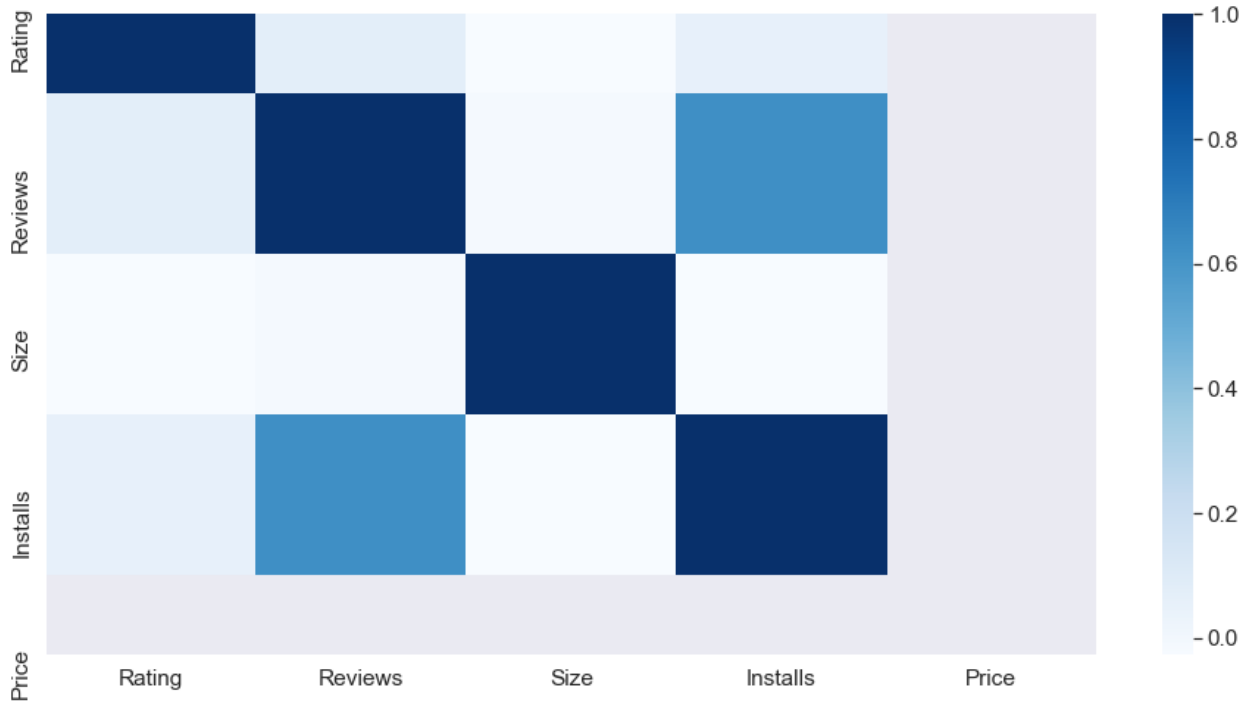
Inference : Amongst the paid apps , there is only one correlation , ie between “Installs” and “Reviews”. It’s imperative to note that “Price” has no strong correlation with “Rating” and “Installs”.

4. Correlation analysis on data grouped by Type : Paid

Correlation table :

	Rating	Reviews	Size	Installs	Price
Rating	1.000000	0.075262	-0.026933	0.056592	NaN
Reviews	0.075262	1.000000	-0.009726	0.625951	NaN
Size	-0.026933	-0.009726	1.000000	-0.027886	NaN
Installs	0.056592	0.625951	-0.027886	1.000000	NaN
Price	NaN	NaN	NaN	NaN	NaN

Heatmap:



Inference : Amongst the free apps , there is only one strong correlation , i.e. between “Installs” and “Reviews”. The attribute “Price” has NaN as their Pearson correlation coefficient as all the values are 0 as they are free apps. This means their standard deviation is 0 , therefore Correlation is not defined.

There is no meaningful correlation between the attributes of the dataset and all important “Rating” and “Reviews” irrespective of how the data is classified.

HYPOTHESIS TESTING

On our quest to find out the successful categories , we performed two hypothesis tests on every category. Based on our observations of the population means of various attributes , we believe that a successful category is the one which has a mean rating of 4.20 or above and has at least 10000000 installs.

Hypothesis test -1

Research Hypothesis Ha : The average rating of the apps belonging to the category is > 4.20

Null Hypothesis Ho : The average rating of the apps belonging to the category is <= 4.20

Hypothesis test-2

Research Hypothesis Ha : The average Installs of the apps belonging to the category is > 10000000

Null Hypothesis Ho : The average Installs of the apps belonging to the category is <= 10000000

Our tests were conducted with a significance level of 5%. Our sample data comprised a minimum of 50% of the population. As these are right tailed tests,we used scipy.stats.norm.sf(), which returns the area to the right, as the P-Value. An example is as follows:

```
Performing hypothesis test with the data belonging to the category :  EDUCATION
Sample size =  58


Hypothesis 1 :
Research Hypothesis Ha : The average rating of the apps belonging to the category :  EDUCATION is >  4.2
Null Hypothesis Ho : The average rating of the apps belonging to the category :  EDUCATION is <=  4.2
Sample mean :  4.389655172413792
Sample std :  0.2559397236963407
Z-Score :  5.643402050283565
P-Value :  8.336119748537628e-09
alpha :  0.5
Null hypothesis Ho is REJECTED



Hypothesis 2 :
Research Hypothesis Ha : The average Installs of the apps belonging to the category :  EDUCATION is >  10000000
Null Hypothesis Ho : The average Installs of the apps belonging to the category :  EDUCATION is <=  10000000
Sample mean :  3958620.6896551726
Sample std :  13224084.679401645
Z-Score :  -3.4792407330628654
P-Value :  0.999748581668441
alpha :  0.5
BOTH Ho and Ha is PLAUSIBLE
In the case of this sample , there is 0.03 % of probablity that the sample obtained is in disagreement with Ho
```

Number of categories : 33
Categories which rejected the first hypothesis : 14
['ART_AND_DESIGN', 'AUTO_AND_VEHICLES', 'BEAUTY',
'BOOKS_AND_REFERENCE', 'COMICS', 'EDUCATION', 'EVENTS',
'HEALTH_AND_FITNESS', 'GAME', 'SOCIAL', 'SHOPPING', 'SPORTS',
'PERSONALIZATION', 'PARENTING']

Categories which rejected the second hypothesis : 10
['BOOKS_AND_REFERENCE', 'COMMUNICATION', 'ENTERTAINMENT', 'GAME',
'SOCIAL', 'PHOTOGRAPHY', 'TRAVEL_AND_LOCAL', 'TOOLS', 'PRODUCTIVITY',
'VIDEO_PLAYERS']

Categories which rejected both hypotheses : 3
['BOOKS_AND_REFERENCE', 'GAME', 'SOCIAL']

It is interesting to observe that though 14 categories rejected hypothesis-1 and 10 rejected hypothesis-2 , only 3 categories are common. This shows that there are many apps which don't

have a lot of installs but offer customer satisfaction. There are other kinds of apps which don't do that well on the ‘Ratings’ front but have a lot of installs nonetheless. This seconds our findings in regard to correlation.
So what can be inferred is that the best bets are the apps belonging to the categories ‘GAME’,‘SOCIAL’ and ‘BOOKS_AND_REFERENCE’.

RESULTS AND DISCUSSION

This project was truly successful in bringing interesting insights into various aspects that makes an application successful.It also proves that one requires more attributes than what was present to accurately predict the success of an app in the playstore.Contrary to the intuitive opinion , as shown by the correlation analysis , both the Rating and the number of Installs of an app is not dependent to its price, size , number of reviews.One might get a slight edge if their targeted audience are teenagers.But based on the trends observed via data visualisation and hypothesis testing, the apps to watch out for generally belong to the following categories :

- ‘GAME’
- ‘SOCIAL’
- ‘BOOKS_AND_REFERENCE’
- 'COMMUNICATION'
- 'EDUCATION'.