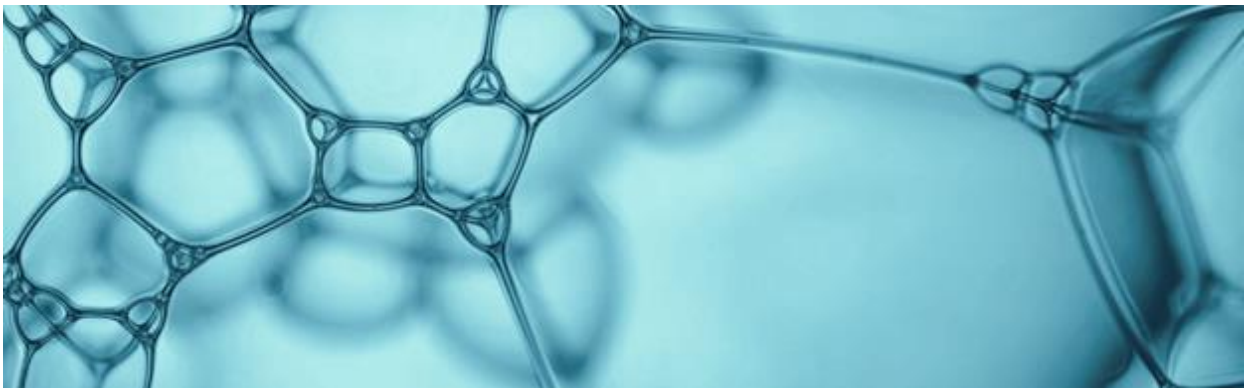


Google-Play-Store Exploratory Analysis

Hi! Welcome to a deep dive into Google Play Store data.

Table of Contents

What should we expect from this presentation? We'll discuss where the data came from and what it contains. After, we'll discuss why this data is of importance. Thirdly, we'll discuss some of the modifications done to the data to make it simpler to analyze. We'll then explore some basic information about the dataset. After that, we could dive deeper into our analysis. We'll conclude this presentation with a summary and key findings.

About The Data

This dataset is available in Kaggle for download. It was updated by a student in India with the username YASH16JR. It is called Google_Play_Store_Cleaned.

It originally contains 9638 rows and 14 columns. Those columns are 'app', 'category', 'rating', 'reviews', 'installs', 'type', 'price', 'content_rating', 'genres', 'current_ver', 'android_ver', 'size(kb)', 'update_month', 'update_year'.

Why Does This Matter?

Businesses and multiple industries have gained valuable insights into their processes, research, sales tactics, and more through data. Our ability to gather this data has increased dramatically since 1983 when the internet was created. Soon after, applications made their debut and changed the way the world communicated.

According to Statista, nearly 90% of Americans, or about 300 million people, use a smartphone. We live in a society where applications are a structural part of life. There is no denying that applications are here to stay, whether for news, communication, entertainment, etc.

This dataset contains information about application features such as ratings, categories, installs, reviews, and prices. Through this information, I hope to gain a deeper understanding of the world of applications.

Data Wrangling

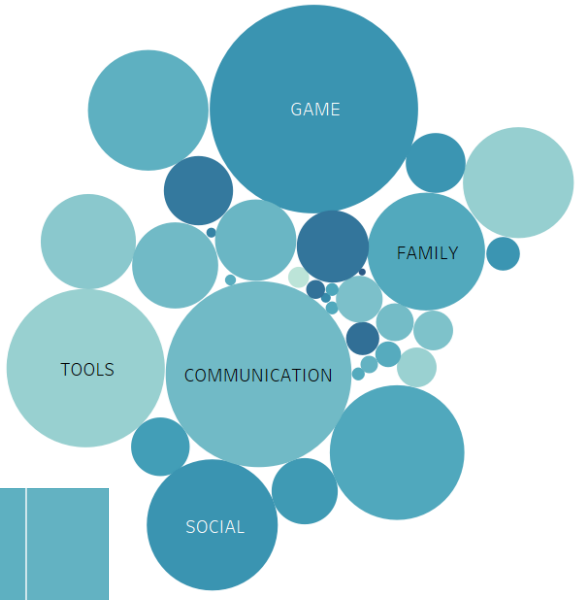
Today, I will not work with all the columns. For this project, I will create a new dataset and only keep the following columns: category, rating, reviews, installs, price, content_rating, and update_year.

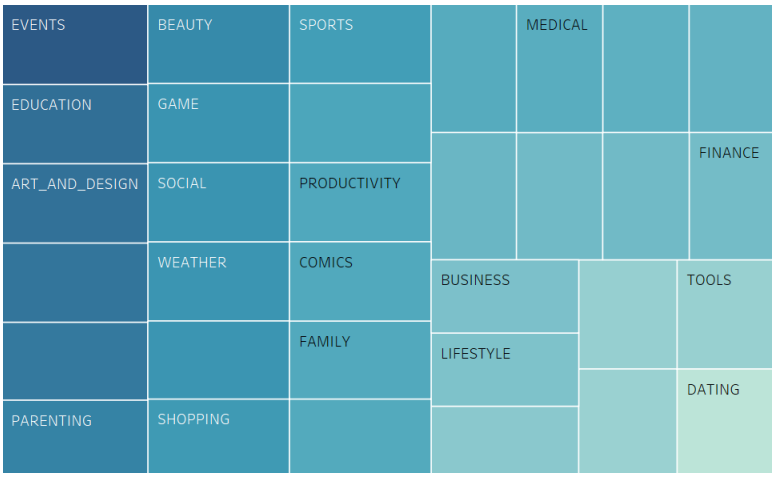
Part of cleaning my data also includes dealing with null values. After some searching, I realized 'rating' was the only column left with null values. Since I'm mostly concentrating on categories, I will replace the null values with the average rating for each category.

I created an additional column with null replacements and will create an updated dataset without the original column. You can find this dataset under ‘Datasets’ folder, under ‘CleanData’.

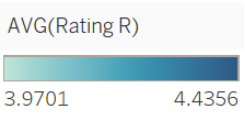
Basic Exploration

This chart shows the top categories with total installs by size, as well as the average rating for each category. It shows the darkest categories having the highest ratings and the lightest having the lowest. While 'Games', 'Communication', and 'Tools' have the most total installs, they lack the highest ratings.





Here is a better view of the highest average rated categories and the lowest. The top rated categories are events, education, and arts and design. The lowest rated categories are tools, maps and navigations, and dating.



Analysis

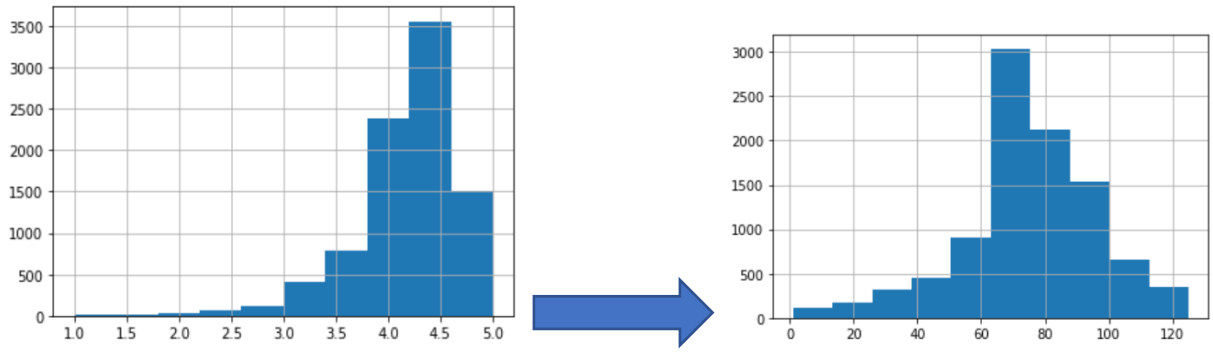
I began to wonder, does an app's rating correlate with other variables? I wanted to create a correlation matrix that helped me see if any variables are correlated with each other. As you can see by this matrix, there is a 63% correlation between installs and reviews. Now, because people tend to review only apps they have installed, this makes sense. No other

correlation seems important.

	reviews	installs	price	update_year	ratingR
reviews	1.0	0.63	-0.0076	0.058	0.055
installs	0.63	1.0	-0.0094	0.069	0.04
price	-0.0076	-0.0094	1.0	-0.0036	-0.02
update_year	0.058	0.069	-0.0036	1.0	0.11
ratingR	0.055	0.04	-0.02	0.11	1.0

In my dive into the dataset, I wanted to see if the difference in rating between free and paid applications was significant. To answer this question, I needed to run an independent t-test. First, I recoded the price column into two groups, Free apps and those with a price tag.

Next, I had to test assumptions. In order to test for normality in my data, I plotted my data and noticed it was negatively skewed. I began the process of transforming my data. Cubing my data made it relatively normal.



Then, I ran the independent t test on my transformed data. Voila! Test showed a significant difference with a p-value of 1.09 E-7. Now, I wanted to see which group had the highest ratings, so I calculated the average rating for each group. The average rating for the free applications was 4.17, while the average rating for the paid applications was 4.25. Paid apps have a slightly higher average rating than free apps. but hmmm... Doesn't seem like there was much of a difference. Despite a significant difference in the test, further examination shows that the difference is not noteworthy.

Even though I did not find a significant difference, I was determined to learn something about this dataset. I discovered that the top-rated categories differed for each group.

Under free applications, 'Books and Reference' was rated third while the first two categories remained unchanged. On the other hand, under paid applications, 'News and Magazines' led the list, while the second and third places remained the same.

The difference in rating between the top and last category is significant in paid applications. The rating is notable when comparing each category individually. Is this why the independent t test showed a significant difference?

In comparing the two groups, I realized that the categories with the most installs differed between free and paid applications. While 'Games' remained the top category for both groups, 'Family' and 'Personalization' took the second and third spots in paid applications. However, it remained the same for free applications. Upon deeper exploration, I also found that many Family apps are also games. There is no doubt that applications have become a source of entertainment for many of us!

Limitations

There were some limitations to this data. There is one in particular within the category 'FAMILY' that stands out. I noticed that in the gender column, there were many subsets of games and other genres that could have been categorized under a different category. This only matters because other categories such as 'tools' and 'personalization' only had one genre and this could have impacted the analysis.

Another limitation was the lack of information about some columns.

Despite being familiar and comfortable with part of this journey, I realized there is much more to analyze within this data, and as I begin my journey as a Data Scientist, I know I have much more to learn.

Summary

Overall, I learned a lot from this dataset.

1. Installs and reviews are strongly correlated.
2. 'Games' are the most installed category of all applications.
3. When comparing each category individually, the difference between free and paid apps is notable even though the average rating does not show much of a difference.
4. Top rated categories differ between free and paid applications.

Thank you!

I'd like to thank you for sitting through this presentation with me, and I hope this gave you some insight into the world of applications.

- Brenda Aguilar