

Lab 3: Data Visualization

Describe a data analysis task for the chosen dataset

Objectives

I choose the `google-play-store-apps` as the database to proceed the Data Visualization task.

In this dataset I want to know the relation between **the Rating and Installs** to choose a App with both popularity and quality, the **Size of the App and Total Number of the Apps** in this range of size, the **Price of the App and Total Number** of the Apps which the price in below this price, the **Content Rating Distribution**, the **Rating of the App and Total Number** of the Apps.

And I also want to know the information from **different category and different charge Type of Apps**. For example, I want to know the characteristics of all the Apps which are belong to ART_AND_DESIGN category. And if I classify all the Apps into different category, I can compare those data from other category, then I can hold a global view of the information in the Google Play Store.

So I am supposed to focus on these columns of the .csv file: `App`, `Category`, `Rating`, `Size`, `Installs`, `Price`, `Content Rating`

Characteristics

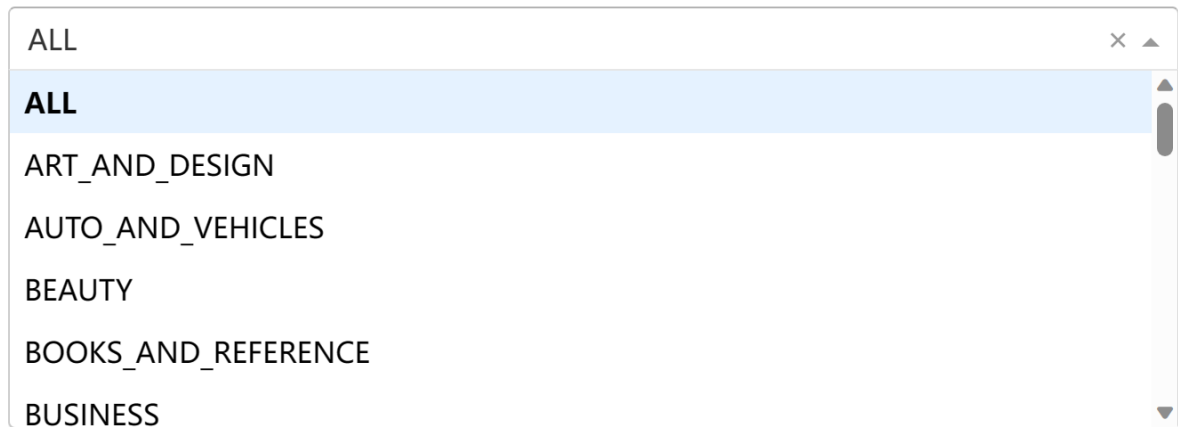
I will illustrate the characteristics from different aspects.

Category

- It has 33 different categories.

And I add 'ALL' to identify the global view. Then I use the Category information to build a `dcc.Dropdown` :

MYADD:'ALL'				
'ART_AND_DESIGN'	'AUTO_AND_VEHICLES'	'BEAUTY'	'BOOKS_AND_REFERENCE'	'BUSINESS'
'COMICS'	'COMMUNICATION'	'DATING'	'EDUCATION'	'ENTERTAINMENT'
'EVENTS'	'FINANCE'	'FOOD_AND_DRINK'	'HEALTH_AND_FITNESS'	'HOUSE_AND_HOME'
'LIBRARIES_AND_DEMO'	'LIFESTYLE'	'GAME'	'FAMILY'	'MEDICAL'
'SOCIAL'	'SHOPPING'	'PHOTOGRAPHY'	'SPORTS'	'TRAVEL_AND_LOCAL'
'TOOLS'	'PERSONALIZATION'	'PRODUCTIVITY'	'PARENTING'	'WEATHER'
'VIDEO_PLAYERS'	'NEWS_AND_MAGAZINES'	'MAPS_AND_NAVIGATION'		



Charge Type

- There are far more Free Apps than Paid Apps.

To observe Apps of Free or Paid types separately, I am supposed to use Paid and Free to build a `dcc.RadioItems` to change the range of Apps.

☐ Free ☐ Paid ☒ All

Size

- The size of the Apps has different representations. Varies with device, means that the size of some Apps vary from different devices, and the units of size are K or M.
- The precise size are between 8.5k~1020.0k, 1.0M~100.0M.

To observe the Size distribution of Apps, I should set the range classification by myself. Instead of balanced classification, I classified lightweight Apps in units of 300K under 1M and larger Apps in units of 25M above 1M.

```
size_category = [  
    '0~300K', '300K~600K', '600K~900K', '900K~25M', '25M~50M',  
    '50M~75M', '75M~100M', 'Varies with device'  
]
```

Price

- The price of Paid Apps are between \$0.99~\$400.0.

In order to reflect the more specific price (\$0 or very high) and the minimum unit is 1 cent, I choose direct numerical statistics.

Content Rating

- It has 6 different content ratings.



I use a pie graph to present the relation of content rating.

Rating Count

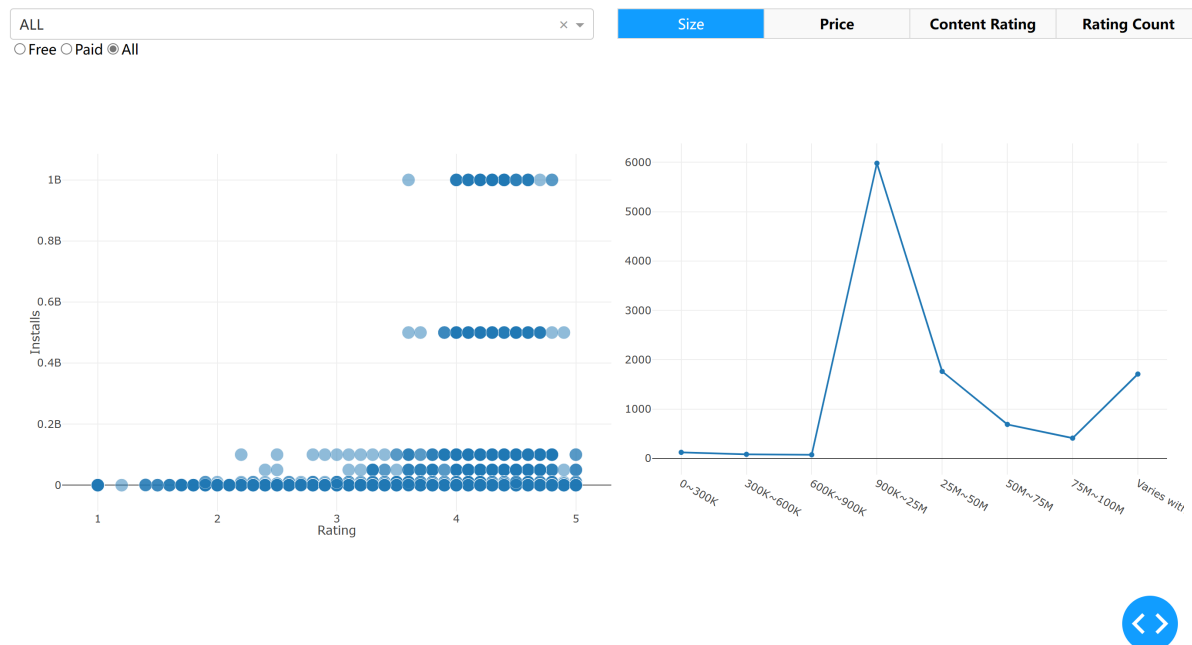
- The rating of Apps are between 1.0~5.0, with NaN value which means that this App has not been rated in the Google Play Store.
- The rating distribution of the Apps in almost every category approximates the normal distribution.

Summary

- It has 33 different categories.
- There are far more Free Apps than Paid Apps.
- The size of the Apps has different representations. Varies with device, means that the size of some Apps vary from different devices, and the units of size are K or M.
- The precise size are between 8.5k~1020.0k, 1.0M~100.0M.
- The price of Paid Apps are between \$0.99~\$400.0.
- It has 6 different content ratings.
- The rating of Apps are between 1.0~5.0, with NaN value which means that this App has not been rated in the Google Play Store.
- The rating distribution of the Apps in almost every category approximates the normal distribution.

Describe the layout of designed dashboard and briefly describe the patterns revealed in the figures.

Gloal View



In the upper-right corner, select the category (34 types includes 'ALL') and type of charges (Free, Paid, All).

Rating and Installs Scatter Graph

The scatter chart of rating and installs is always shown on the left, which is convenient to find certain Apps with particularly outstanding quality or popularity by category.

the patterns revealed

- Only the Apps with rating more than 3.6 can get more than 500,000,000(500M).
- Most of the installs of Apps are below 20M, but a few of them hold the installs number more than 1,000,000,000(1B).
- The relation of rating and installs are not entirely linear, which means that it not what we think "higher rating means more installs number", and we can conclude some of the good quality Apps are more professional or have very little publicity.
- Choose the 'Paid' Radio Item, we can know if the Apps are charged, the installs number will be at a really low level. Most of the users don't want to pay for the software they use.

Size and AppCount Line(Scatter) graph

Click the first "Size" of the view toggle button on the right, and Size and AppCount Line(Scatter) graph is shown.

the patterns revealed

- The majority size of Apps are between 900K~25M. It holds in almost every categories.
- Only a few number of Apps are between 600K~900K, and also only a few number of Apps are more than 100M from the data in the database.
- **The relation of the attribute —The distribution of size approximates normal distribution.**
- And there are also some Apps which hold the size vary from different devices.

Price and AppCount Bar graph

Click the second "Price" of the view toggle button on the right, and Size and Price and AppCount Bar graph is shown.

the patterns revealed

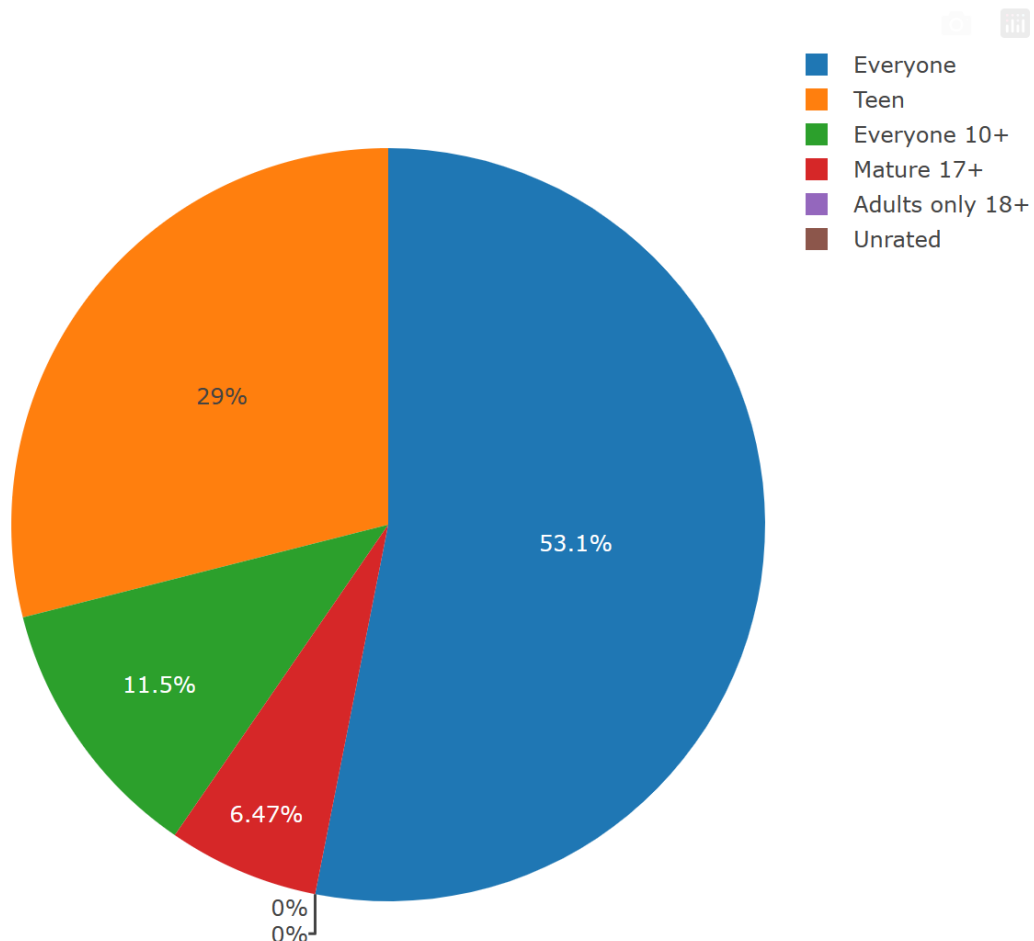
- We can have a obviously view that most of the Apps are free. It holds in every category.

Content Rating Pie graph

Click the third "Content Rating " of the view toggle button on the right, and Content Rating Pie graph is shown.

the patterns revealed

- The majority of Apps are open for everyone to install.
- There are some restriction for Teen in some of the category, which we can guess that there kind of Apps are not suitable for teens.
- There are over 29% Apps for teen which they are belongs to GAME category.



- Almost few Apps are Unrated, which we can learn that the Google holds a strictly restriction for publishing software in comparatively speaking.

Rating and AppCount Bar graph

Click the fourth "Rating Count " of the view toggle button on the right,and Rating and AppCount Bar graph is shown.

the patterns revealed

- The ratings of Apps are normally distributed,and 4.4 is the center value.
- Apps with a rating of 3.5 to 5 make up the vast majority.

