Lab 3: Data Visualization

Make use of dash to visualize one of the three dataset.

❗❗❗ PLEASE READ README.md FOR DETAILED INTRODUCTION

Data Analysis Task

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 Reviews and Installs**, 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.

And I also want to know the information from different category 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`, `Reviews`, `Size`, `Installs`, `Price`, `Content Rating`

Characteristics

The first thing I have to do is that I should know the characteristic of each attributes and the basic relationship between different attributes before plotting my dash graphs.

I write the `fetch_attribute.py` to fetch the attributes I need. (However, after I get the result, I use this .py file to write another logic code😭, so I have not put this .py file in the `src` folder😫)

The characteristics of all the attributes I used are list here:

Category

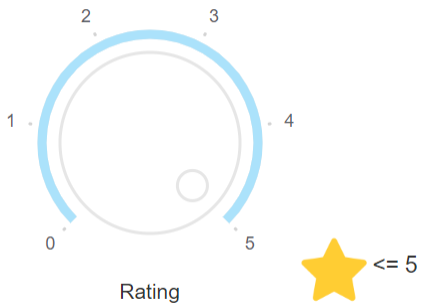
- It have 33 different values, and I use the Category information to build a `dcc.Dropdown`:

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

ART_AND_DESIGN
WEATHER
VIDEO_PLAYERS
NEWS_AND_MAGAZINES
MAPS_AND_NAVIGATION

Rating

- NaN value which means that this App have not been rated in the Google Play Store
- A float number between 1.0~5.0, and the step I choose is 0.1
- I use a `daq.Knob` to present it, and allow the user to interact with the graph by dial the knob.



Size

- Varies with device, which means that the size of some Apps vary from different devices, so I have to manage this kind of Apps specially.
- There are 278 items which the scale of the Apps is "KBytes", and the precise size are between 8.5k~1020.0k. And there is an interacting thing I observed, there is only one item which the last number after the point is 5, others are all integer. So, I can use 1 as a step to build my graph for attribute —— size.
- And there are 182 items which the scale is "MBytes", and the precise size are between 1.0M~100.0M. Use 0.1 for here as a step will be better.
- Finally, I classify the attribute —— size as follow:

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

Installs

- It have 20 different values in attribute —— installs, and they are as follow:

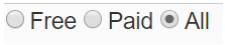
0+
1+
5+
10+
50+
100+
500+
1,000+
5,000+
10,000+
50,000+
100,000+
500,000+
1,000,000+
5,000,000+
10,000,000+
50,000,000+
100,000,000+
500,000,000+
1,000,000,000+

Price

- 0, which means that the App is free for download. And I manage this kind of value specially.
- $0.99~$400.0, which means that the App is charged. And I choose 0.01 as step to copy with the price.
- Finally, I classify the attribute —— price as follow:

```
price_category = [
        '0','$1~$50','$50~$100','$100~$150','$150~$200',
        '$200~$250','$250~$300','$300~$350','$350~$400'
    ]
```

- I use a `dcc.RadioItems` which hold 3 choices: `Free`, `Paid`, `All`

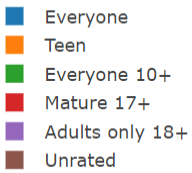


Content Rating

- It have 6different values in attribute —— content rating, and they are as follow:

Everyone
Teen
Everyone 10+
Mature 17+
Adults only 18+
Unrated

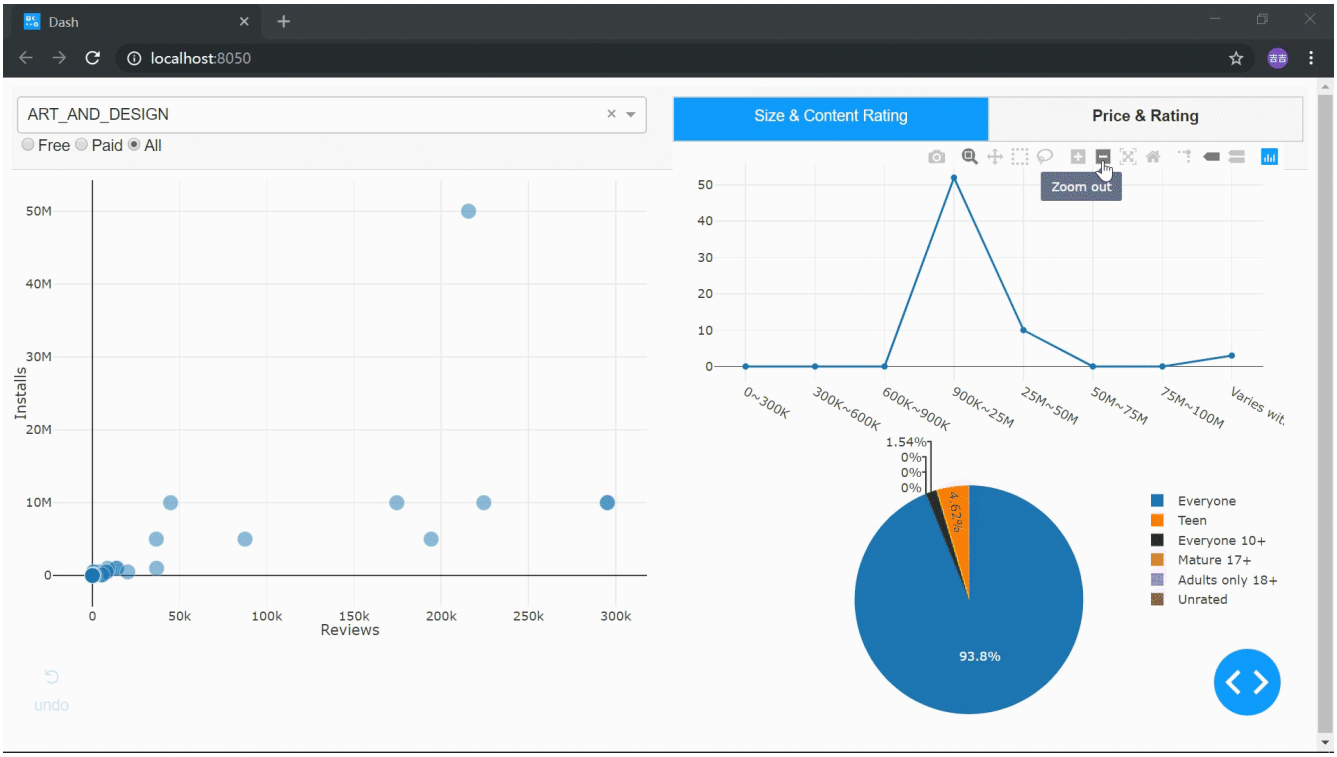
- And I use a pie graph to present the relation of content rating and other attributes and information in this database



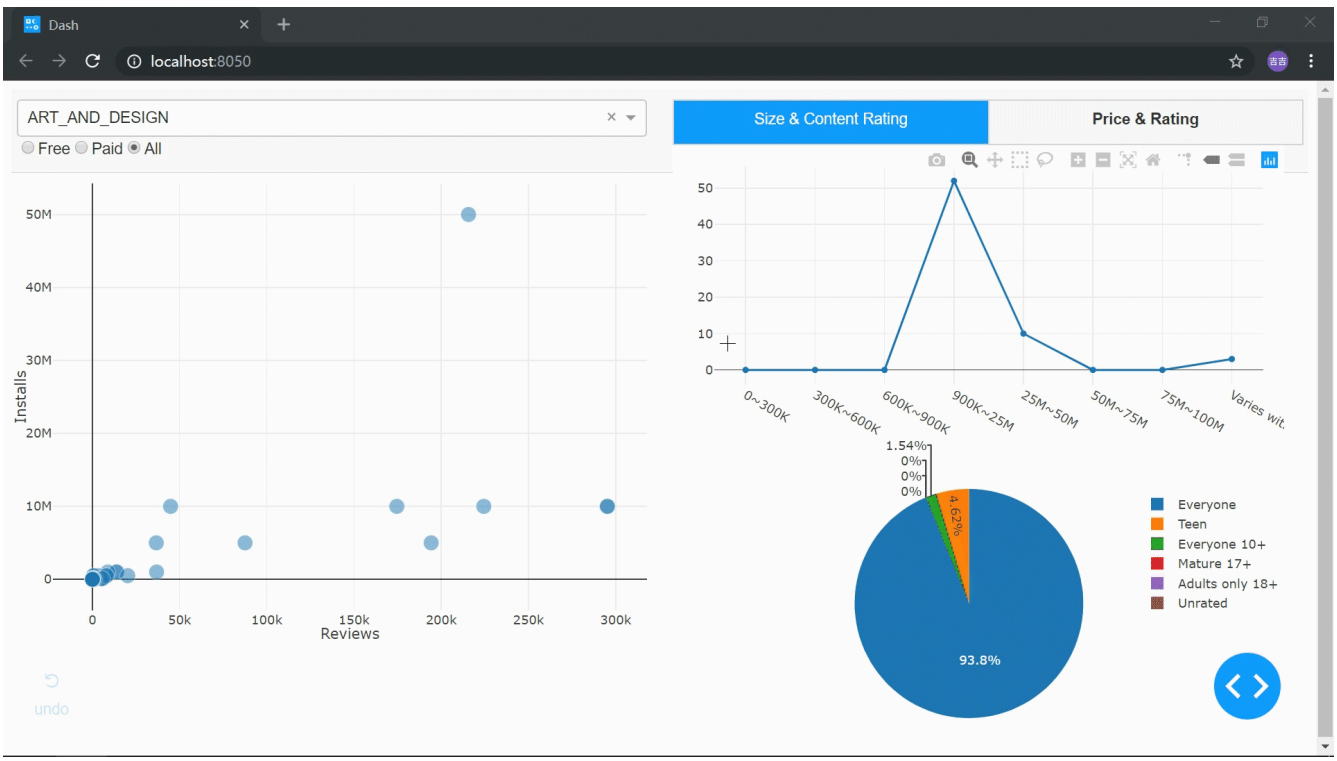
Interact Between User & Dashboard

❗Please read the .md file to see the gif for interaction.

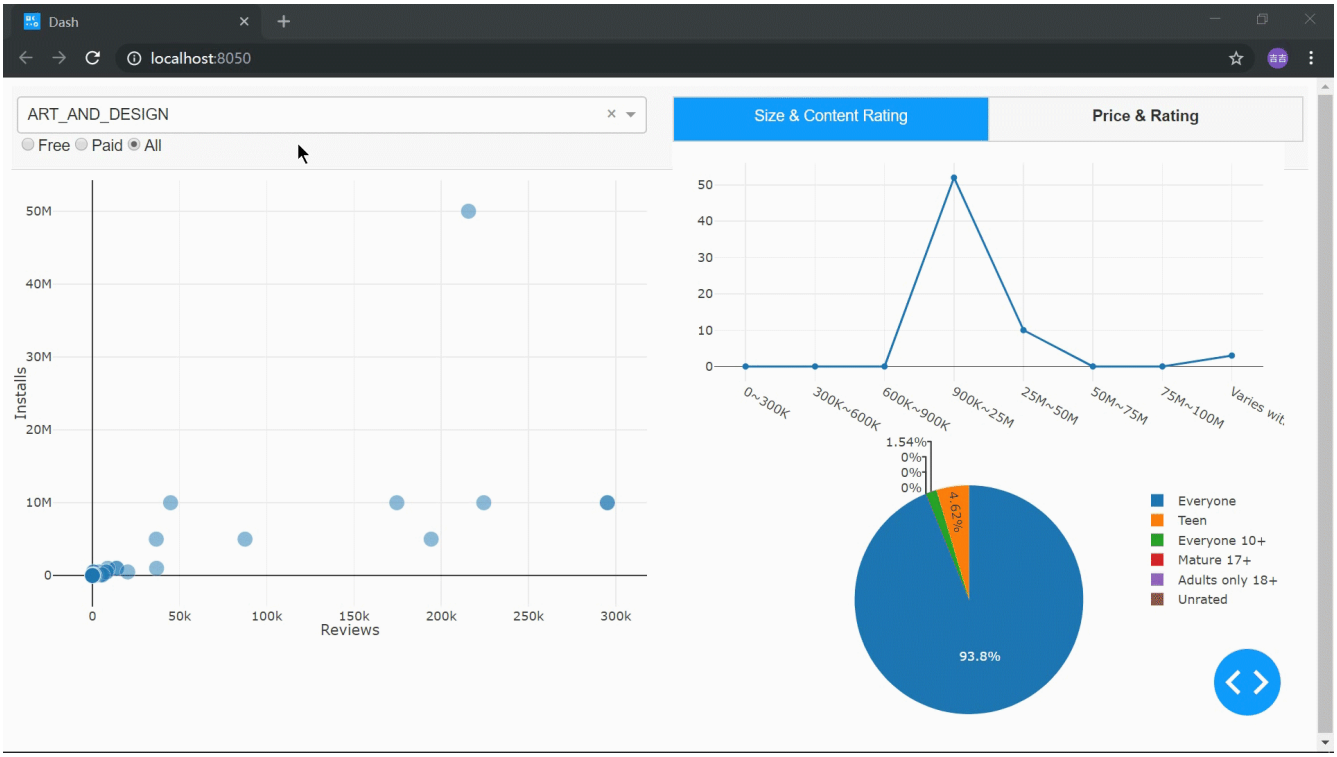
Switch from different Rating



Switch from Free or Paid

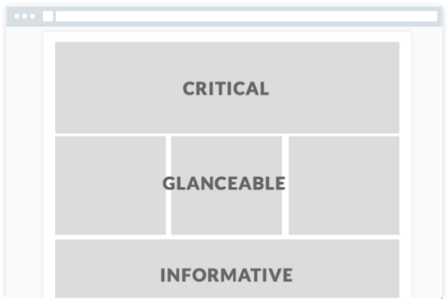


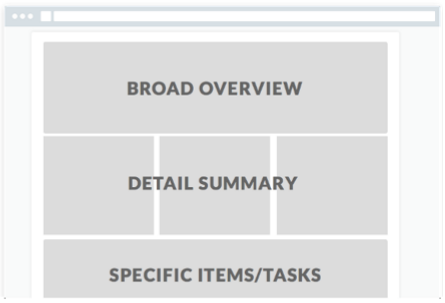
Switch from different item in the category



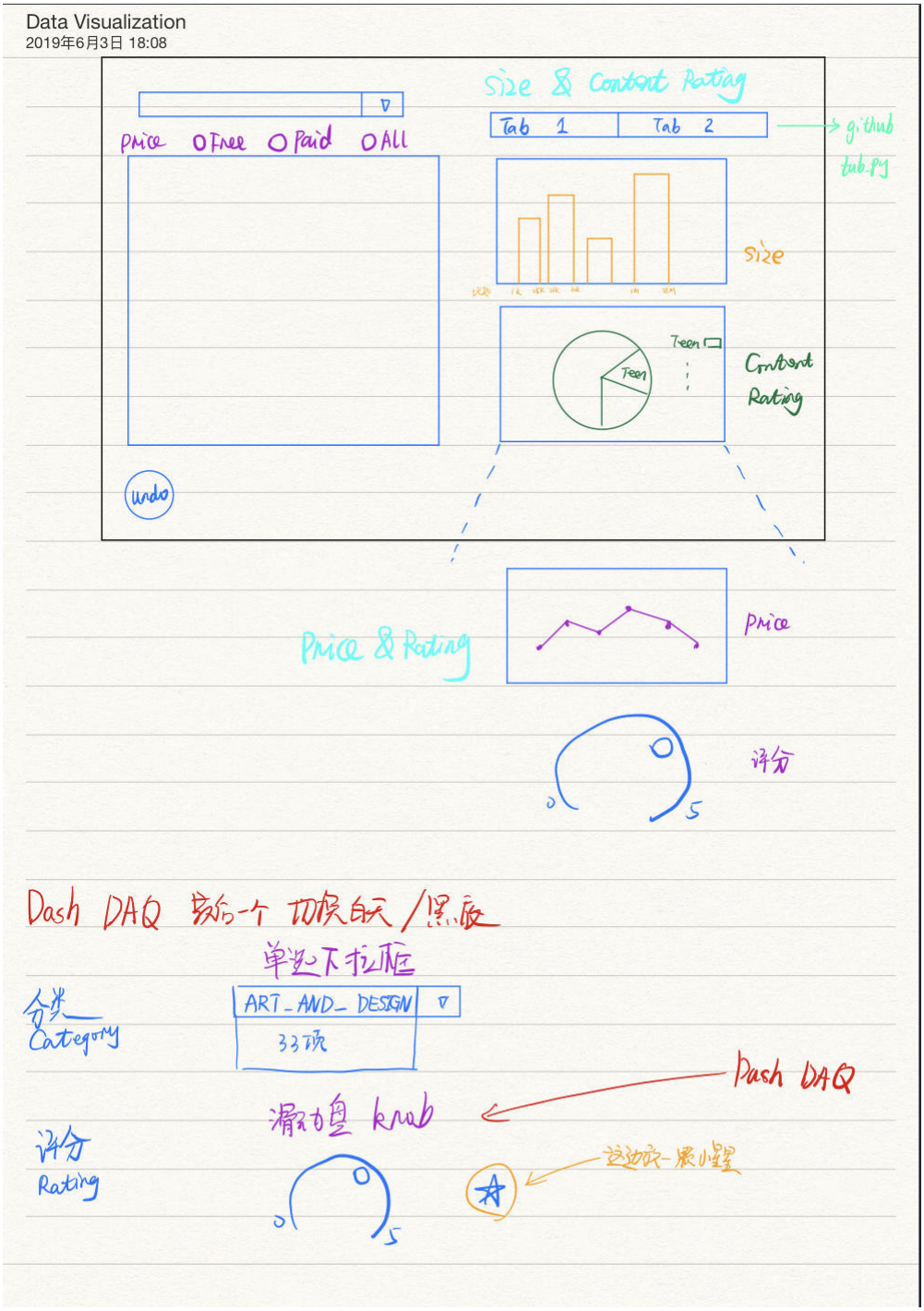
Layout of Designed Dashboard

- After I finish the **Data Analysis Task**, I have an ideal overview of this database, and then I concentrate more on the design of the **Data Visualization**.
- I take the recommend layout shown in teacher's slide into consideration seriously. But as a consequence of the particularity of this database, a majority of attributes are related with the category and the price, so I transfer my mind and create another layout.

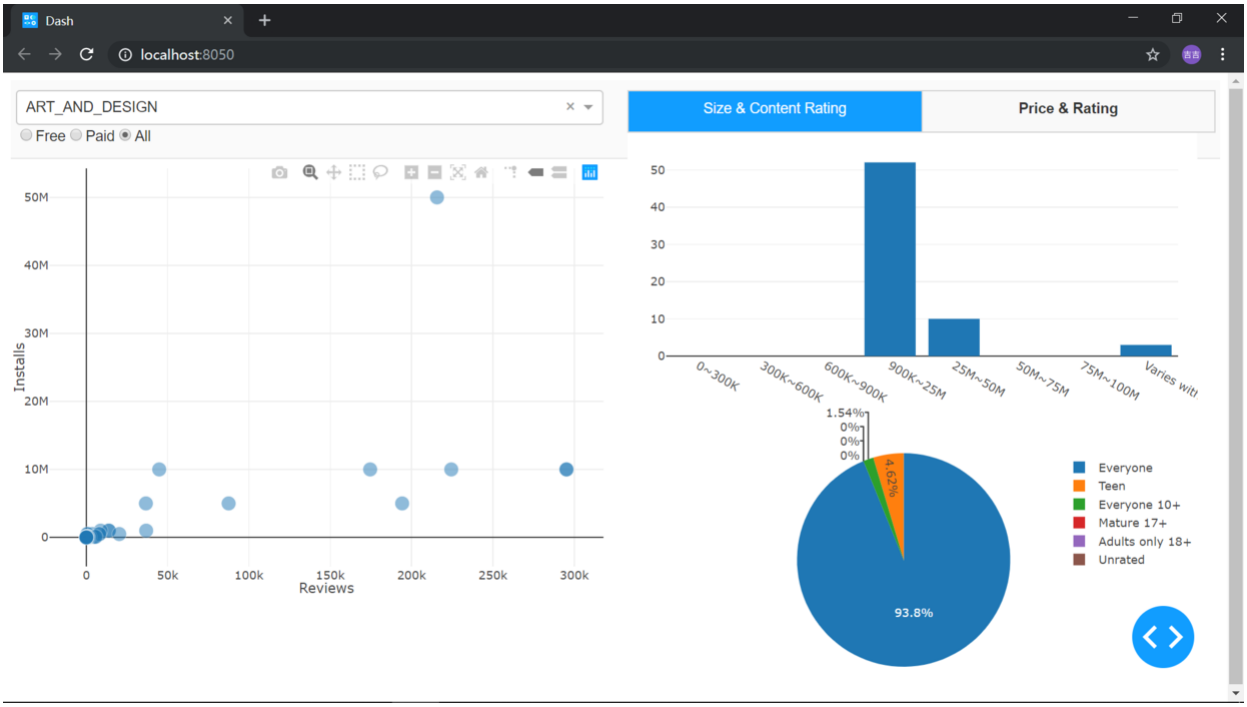


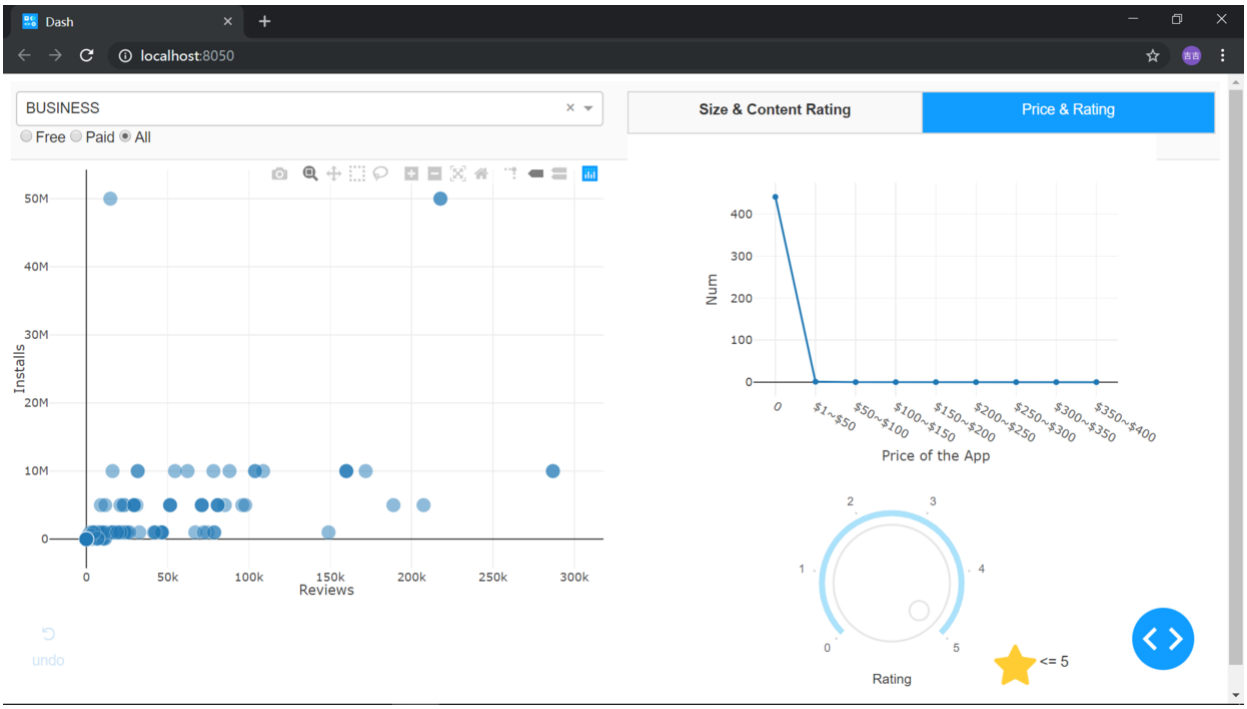


- Firstly, I draw a prototype on my iPad(please ignore my poor paint standard😐)



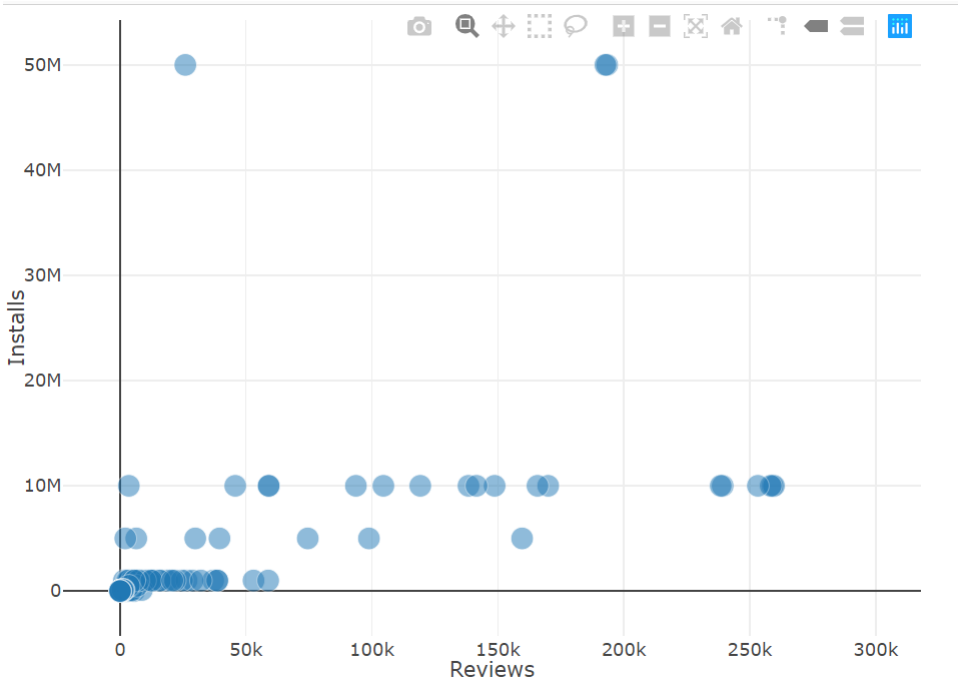
- Then, I use some `Dash Core Components(dcc.Dropdown, dcc.RadioItems, dcc.Tab, dcc.Graph)`, `Dash DAQ Components(daq.Knob)` to manufacture my dashboard. Some interfaces are as shown below:





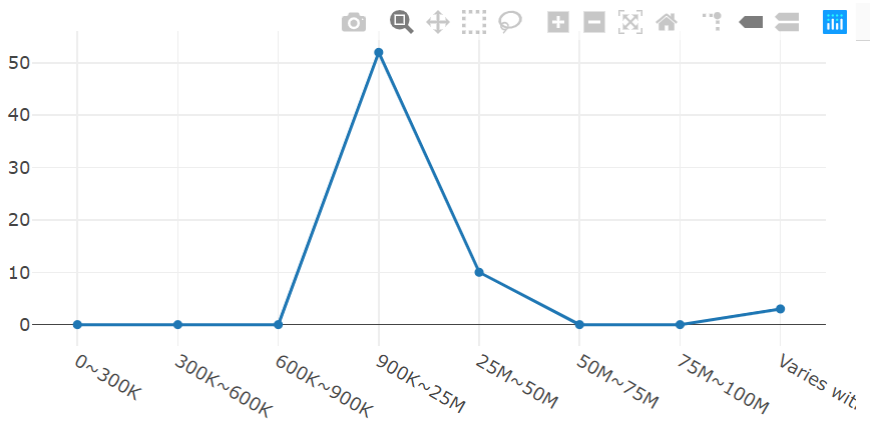
Patterns Revealed in the Figures

- I draw 4 graphs for this database:
    1. Main-Scatter graph, the model is markers



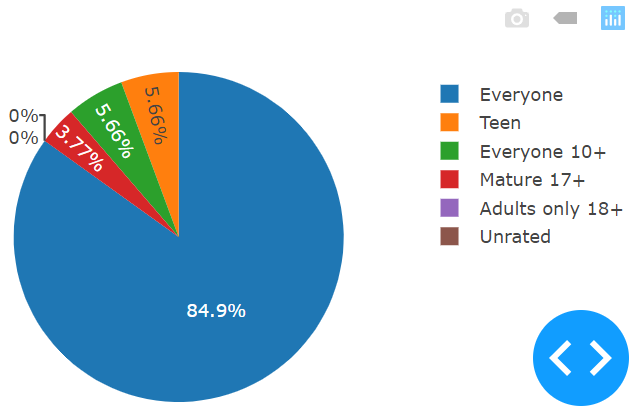

```
'data': [
    go.Scatter(
        x=Reviews,       # x轴为评论数
        y=Installs,      # y轴为安装数
        text=App,        # 点信息为App的名字
        mode='markers',
        marker={
            'size': 15,
            'opacity': 0.5,
            'line': {
                'width': 0.5,
                'color': 'white'
                }
})],
```

2. Size-Bar(Scatter) graph



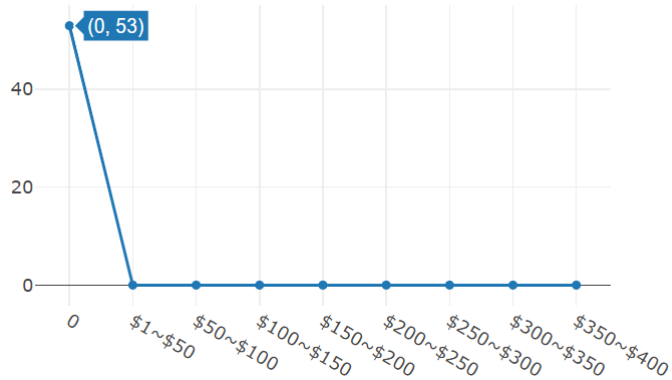
```
'data': [{
    'x': size_catalogue,      # x轴为软件大小分类
    'y': size_list,           # y轴为软件大小列表
    'type': 'Scatter',
}],
```

3. Content-Rating-Pie graph




```
 trace = go.Pie(
    labels=content_rating_catalogue,      # x轴为应用分级分类
    values=content_rating_list,           # y轴为应用分级列表
)
'data': [trace],
```

4. Price-Scatter graph




```
'data': [{
    "x": price_category,          # x轴为软件价格分类
    "y": price_list,              # y轴为软件价格列表
    "mode": "lines+markers",
    'type': 'Scatter',
    'name': 'Line'
}],
```

- The *Reviews and Installs* **Sctter Graph**:

    - From the graph we can find most of the installs of Apps are below 20M, but a few of them hold the installs number more than 50M🔥.
    - More Apps hold reviews number below 50k, but a few of them can reach more than 300k 🔥.
    - The relation of reviews and installs are not entirely linear, which means that it not what we think "more reviews means more installs number", and we can conclude there maybe someone who don't install the App but give his or her reviews, or he or she unload the App.
- The *Size* **Bar Graph**:
    - The majority size of Apps are between 600K~900K and 25M~50M.
    - Only a few number of Apps are less than 300K, and also only a few number of Apps are more than 100M from the data in the database.
    - **The relation of the attribute ——Size obeys normal distribution**.
    - And there are also some Apps which hold the size vary from different devices.
- The *Content Rating* **Pie Graph**:
    - The majority of Apps are open for everyone to download.
    - 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😨. I have some concern to the physical and psychological health for teen.
    - Only a few Apps are Unrated, which we can learn that the Google hold a strictly restriction for publishing software in comparatively speaking.
- The *Price* **Scatter Graph**:
    - We can have a obviously view that most of the Apps are free.
    - And if the Apps are charged, the installs number will be at a really low level.
    - We can know that most of the users don't want to pay for the software they use😢.
- The *Rating* **Knob**:
    - Along with we dial the rating of the knob we can know that most of the Apps hold a rate over 3.0 point.
    - Different category Apps hold different rating.
    - `TRAVEL_AND_LOCAL`, `SPORTS`, `SOCIAL`, `LIFESTYLE` and .ect hold the higher rating than other category from the rough observation.
    -

Some Questions

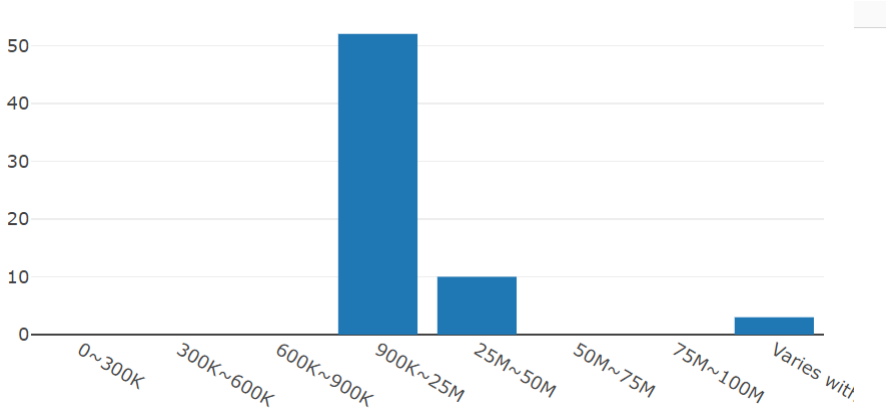
After the shallow learn of the `dash`, I have some question of this framework:

The Blank of the Bar Graph

Firstly, I write the code of the Size-Graph like this:

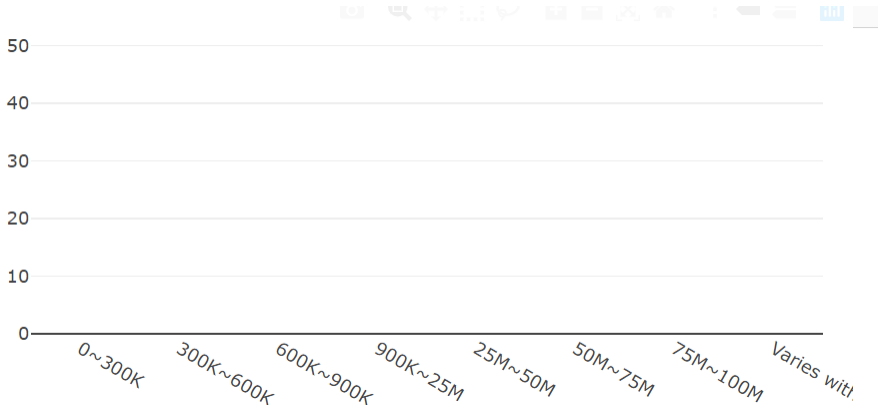
```
'data': [{
   'x': size_category,      # x轴为软件大小分类
   'y': size_list,          # y轴为软件大小列表
   'type': 'Bar',
}],
```

And the result in the dashboard is like this:



It is perfect, and it can describe the relationship between the Size and the Number clearly.

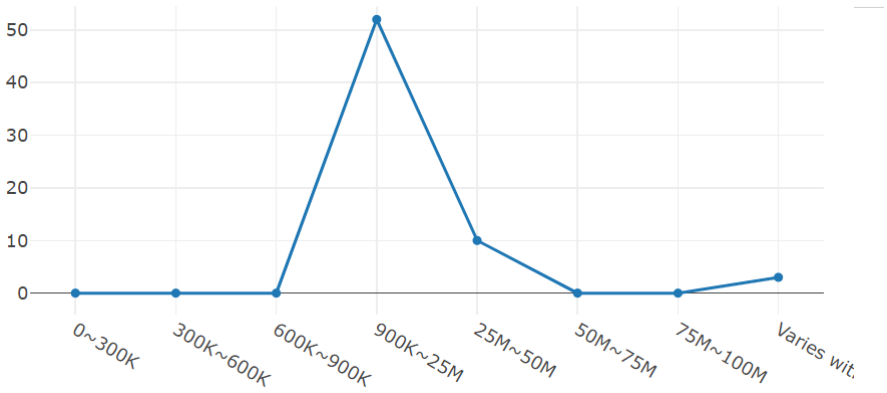
However, if I change the category from the `dcc.Dropdown`, there will be blank

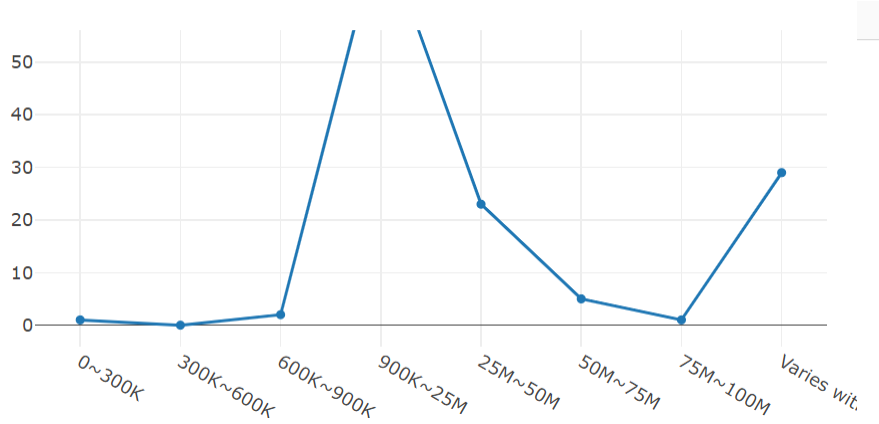


I print the content of the x-label and the y-label, which the data is correct all the time, but it wouldn't render it in the html pages

```
['0~300K', '300K~600K', '600K~900K', '900K~25M', '25M~50M', '50M~75M', '75M~100M', 'Varies with device'] [0, 0, 0, 52, 10, 0, 0, 3]
['0~300K', '300K~600K', '600K~900K', '900K~25M', '25M~50M', '50M~75M', '75M~100M', 'Varies with device'] [1, 0, 1, 50, 16, 5, 2, 10]
```

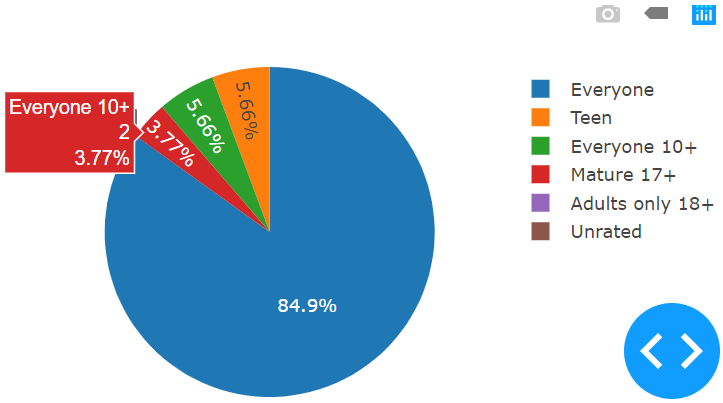
After I change the Bar Graph to Scatter Graph, it will be correctly.





The Deranged of the Pie Graph

Sometimes, the label and the value do not match each other, but I print the detailed information of labels and the values, they are correct, as the same as the Bar Grapy, when they are render to the html pages, it will be deranged.




```
['Everyone', 'Teen', 'Everyone 10+', 'Mature 17+', 'Adults only 18+', 'Unrated'] [61, 3, 1, 0, 0, 0]
['Everyone', 'Teen', 'Everyone 10+', 'Mature 17+', 'Adults only 18+', 'Unrated'] [83, 1, 1, 0, 0, 0]
['Everyone', 'Teen', 'Everyone 10+', 'Mature 17+', 'Adults only 18+', 'Unrated'] [445, 13, 1, 1, 0, 0]
['Everyone', 'Teen', 'Everyone 10+', 'Mature 17+', 'Adults only 18+', 'Unrated'] [102, 125, 2, 67, 0, 0]
```

I hope the teacher Shen can solve my inexplicable errors and give me a reply. Thanks for the hard teacher🙏.

About the author

ID 1754060

name Zhe Zhang

adviser Ying Shen

contact email: doubleZ0108@gmail.com