IMD033 - Probabilidade

Aula 04 - Projeto #1

Ivanovitch Silva
Fevereiro 2019

Atualizar o repositório

```
git clone https://github.com/ivanovitchm/imd0033_2019_1
```

Ou ....

```
git pull
```

Lists of Lists

Files

Dictionary

Function

All in One

Hello World

Lists

For Loops

Conditional Statements

Real Python

Mobile Apps

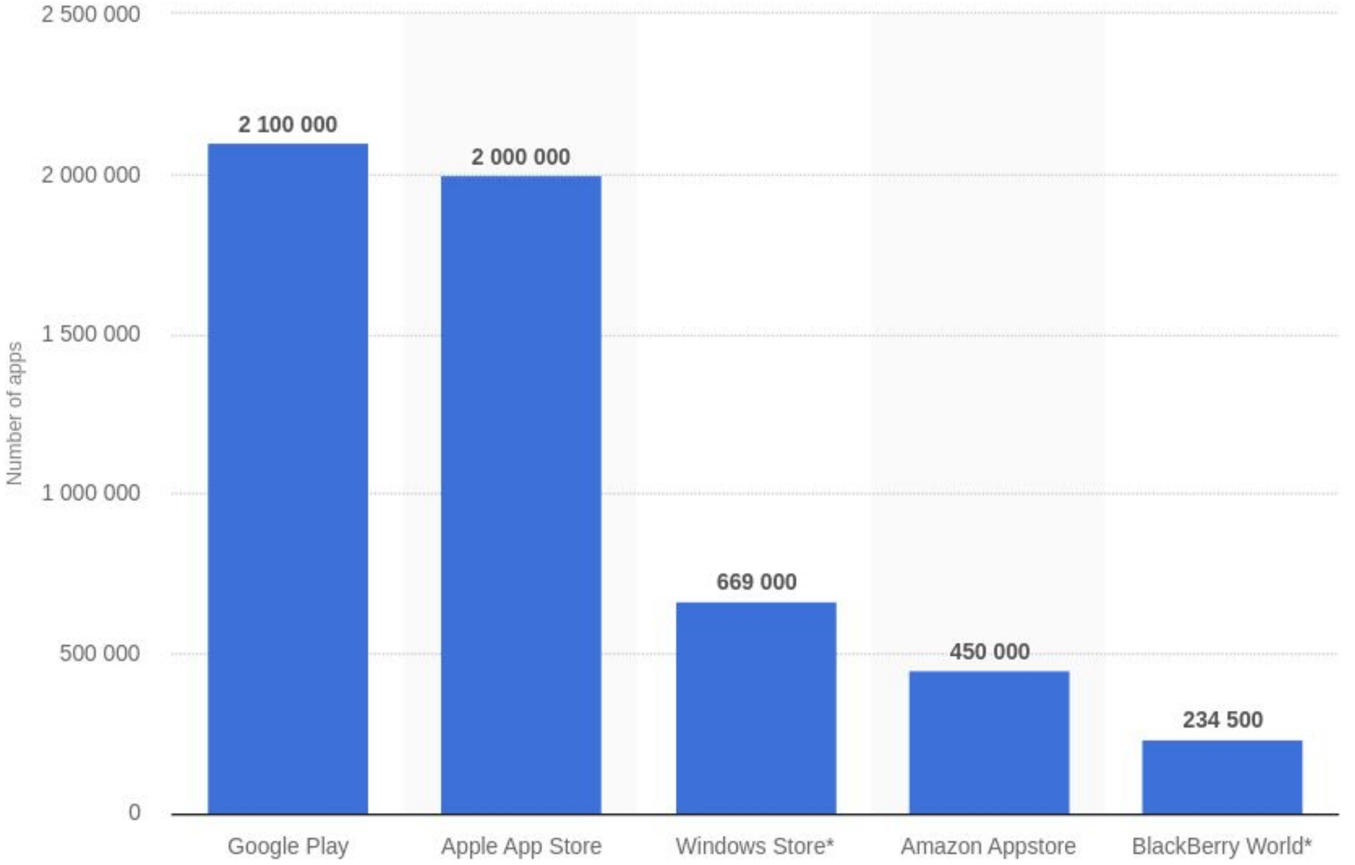
BigBasket
Shop By Category
Fruits and Vegetables
Groceries and Staples
Bread, Dairy and Eggs
DISCOUNTS
5%
25%
Starter Kit
Gaming & Toys
Sports
Colure
Home
Search
Deal
Profile
Friday
Network Hospitals
Health & Wellness
My Coverage
My Family
Policy Guide
My Claims
Other Contacts
Claims on the
Ads
FREE APP!

The revenue for any given app is mostly influenced by the number of users who use our app

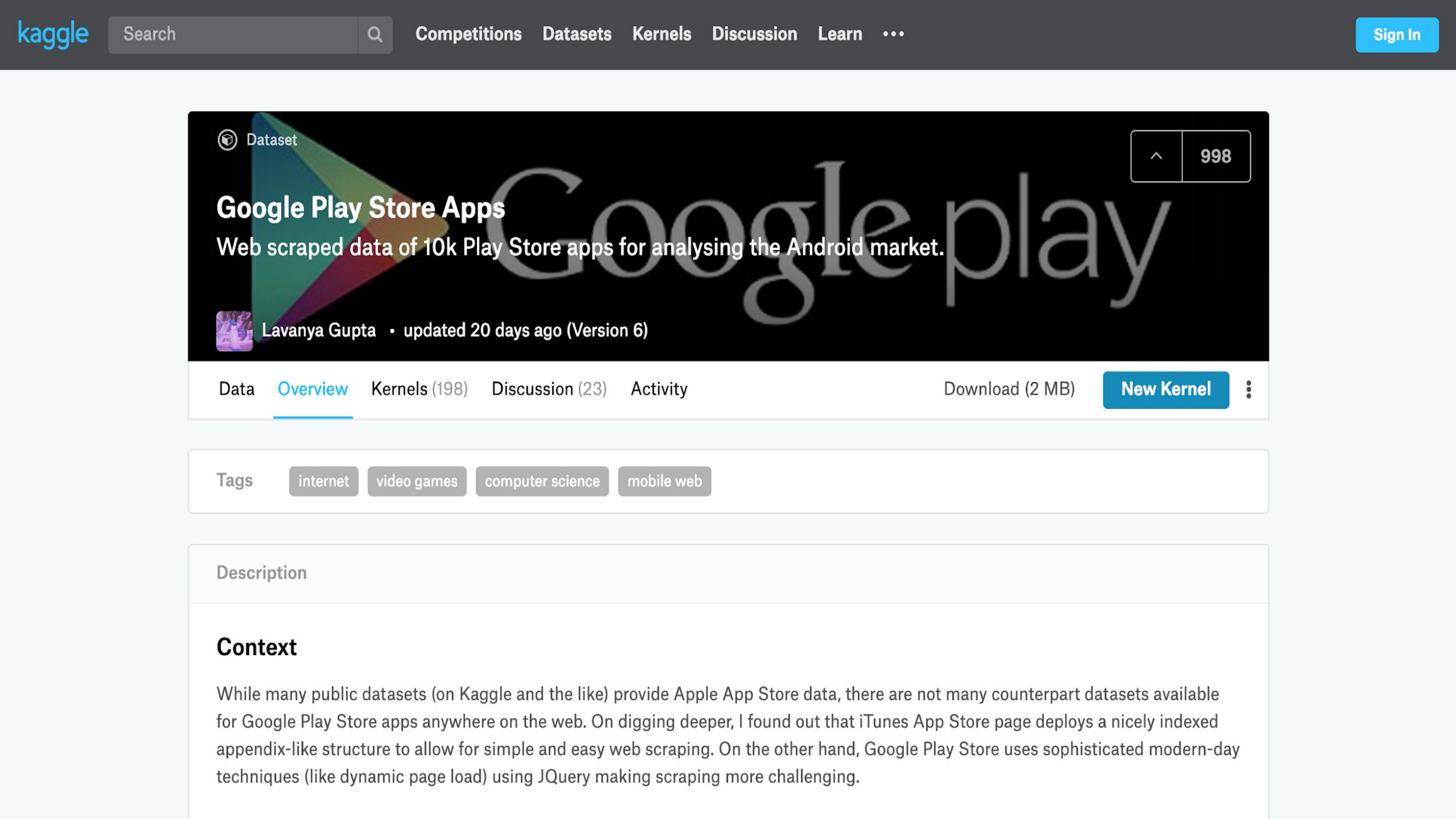
Our goal for this project is to analyze data to help our developers understand what **KINDS OF APPS** are likely to attract **MORE USERS.**

1

Your portfolio should **exemplify your awesome development (and design) skills.**



© Statista 2018



Google Play Store Apps

Web scraped data of 10k Play Store apps for analysing the Android market.

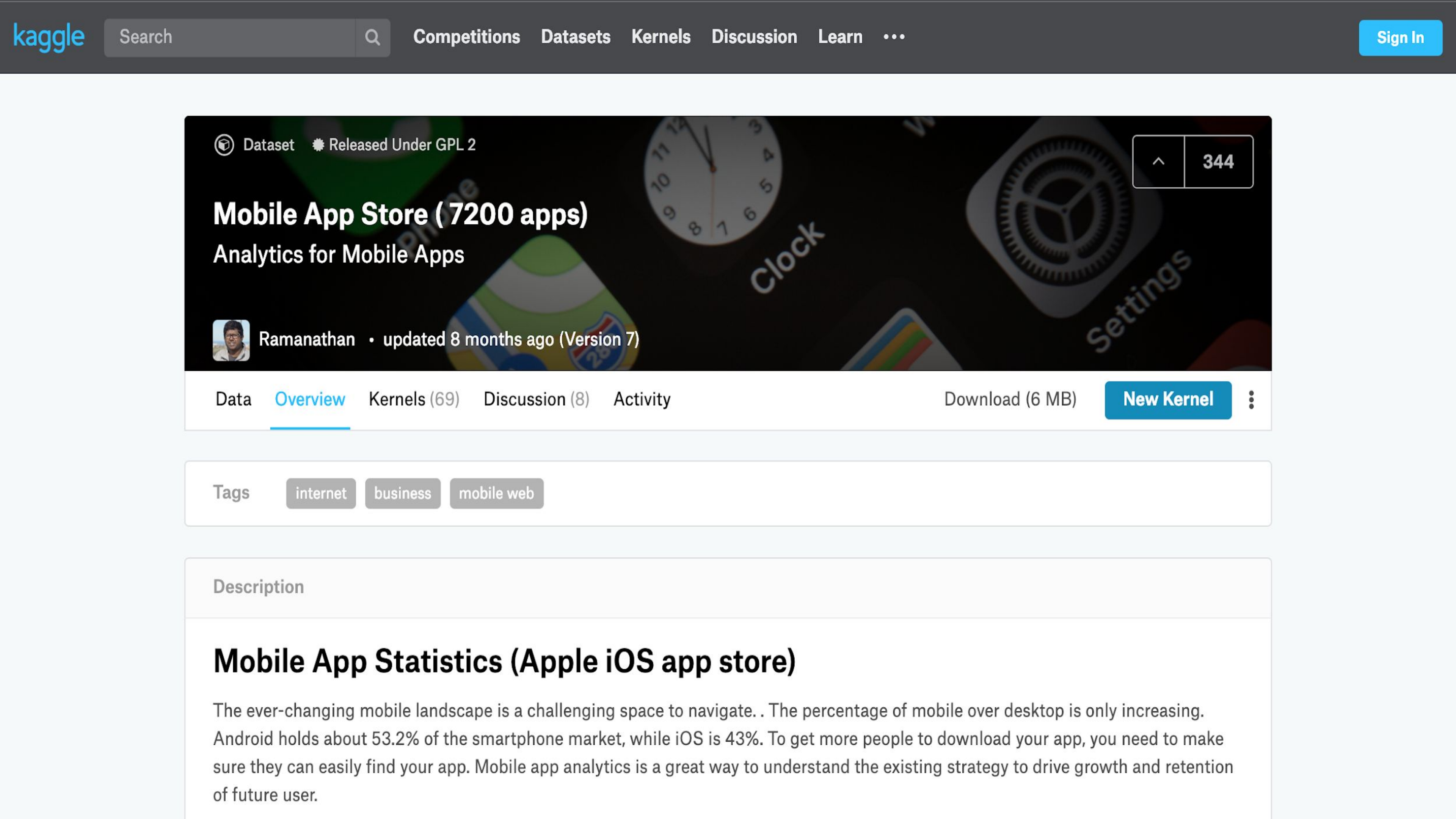
Lavanya Gupta • updated 20 days ago (Version 6)

Tags: internet, video games, computer science, mobile web

Description

Context

While many public datasets (on Kaggle and the like) provide Apple App Store data, there are not many counterpart datasets available for Google Play Store apps anywhere on the web. On digging deeper, I found out that 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.



Mobile App Store (7200 apps)

Analytics for Mobile Apps

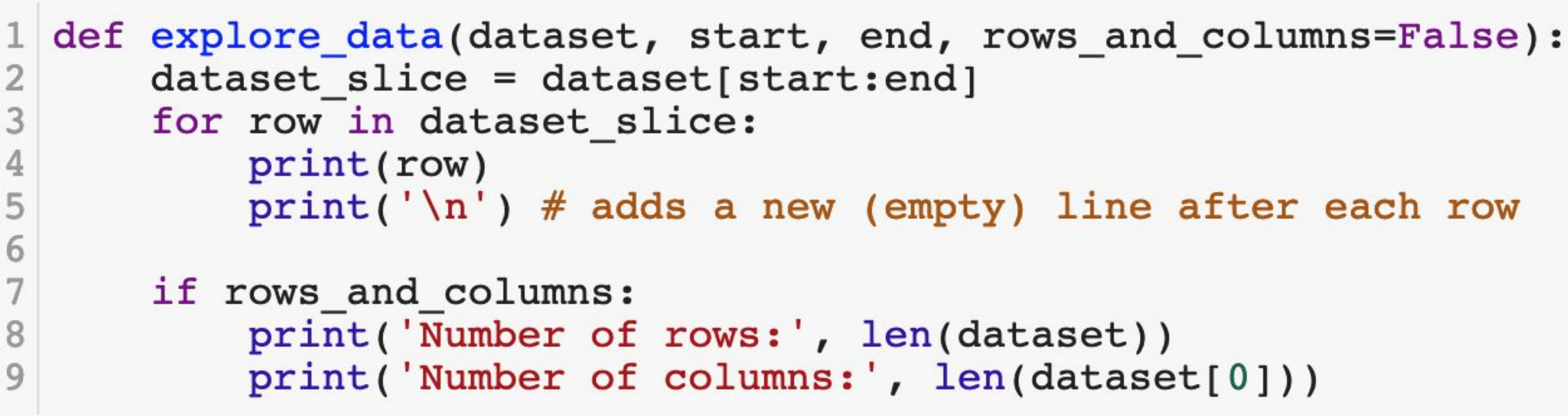
Ramanathan • updated 8 months ago (Version 7)

Mobile App Statistics (Apple iOS app store)

The ever-changing mobile landscape is a challenging space to navigate. . The percentage of mobile over desktop is only increasing. Android holds about 53.2% of the smartphone market, while iOS is 43%. To get more people to download your app, you need to make sure they can easily find your app. Mobile app analytics is a great way to understand the existing strategy to drive growth and retention of future user.

2

Opening and explore the data



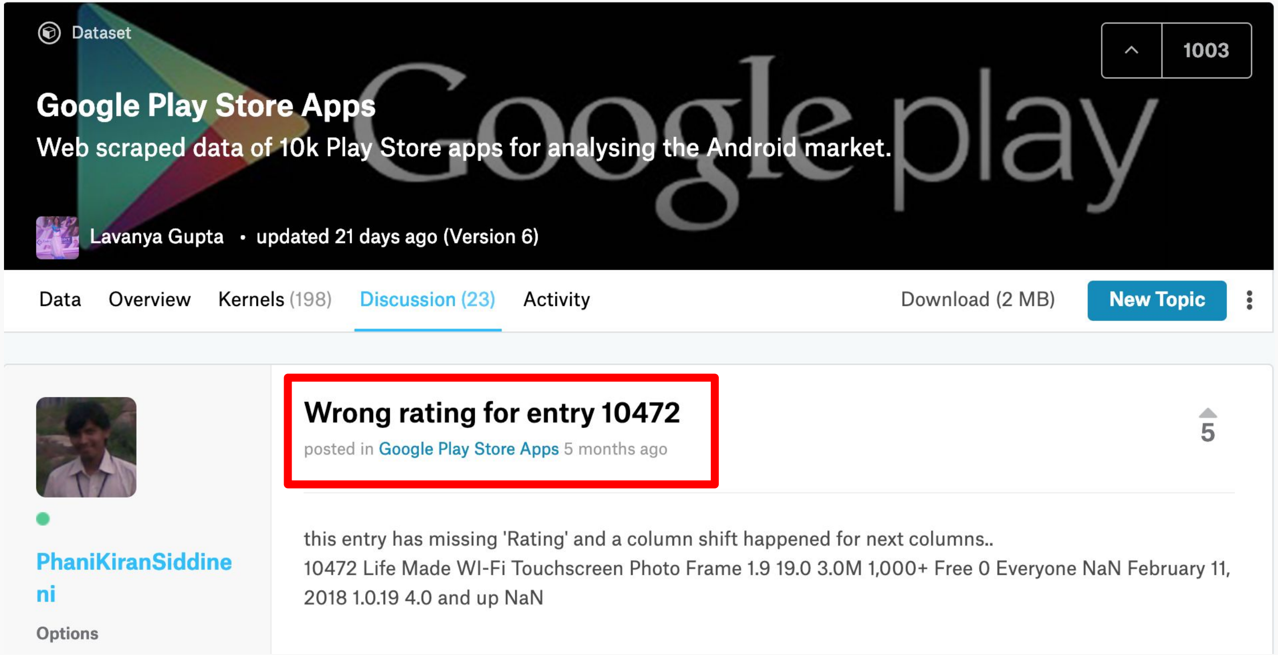
```python
def explore_data(dataset, start, end, rows_and_columns=False):
    dataset_slice = dataset[start:end]
    for row in dataset_slice:
        print(row)
        print('\n') # adds a new (empty) line after each row

    if rows_and_columns:
        print('Number of rows:', len(dataset))
        print('Number of columns:', len(dataset[0]))
```

80:20

- Detect inaccurate data and correct (or remove) it
- Detect duplicate data and remove the duplicates
- Remove non-English apps
- Remove non-free apps

3

Deleting wrong data



Google Play Store Apps

Web scraped data of 10k Play Store apps for analysing the Android market.

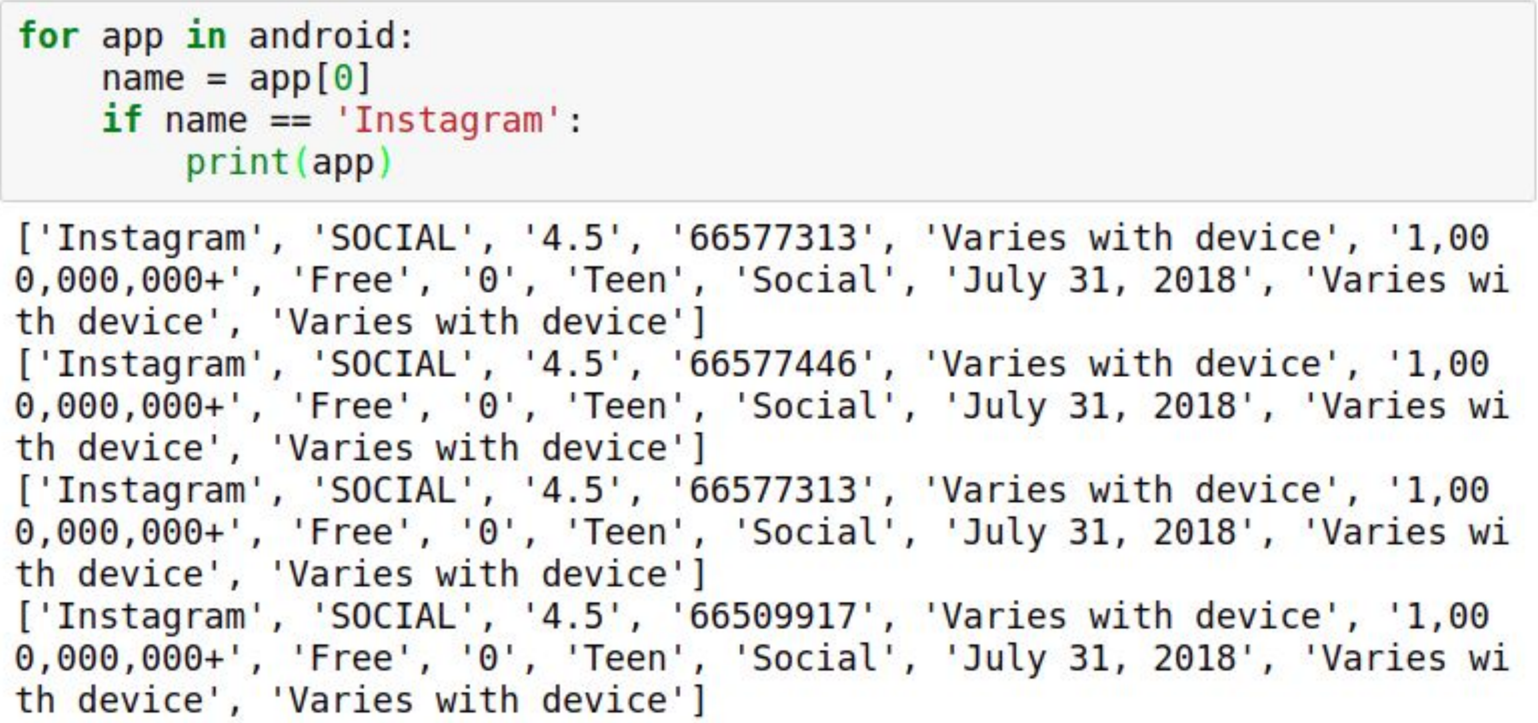
Lavanya Gupta • updated 21 days ago (Version 6)

Wrong rating for entry 10472

posted in Google Play Store Apps 5 months ago

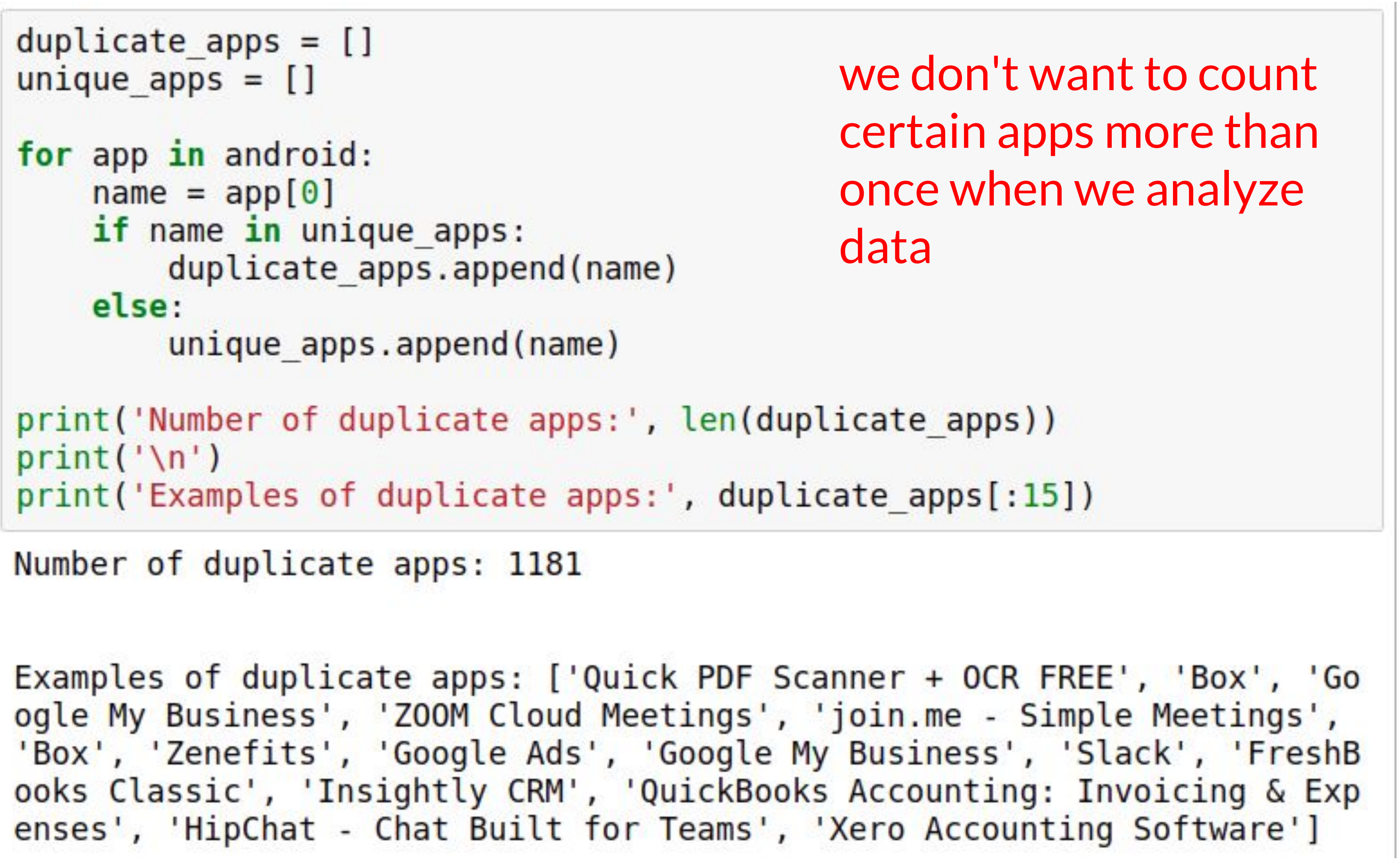
this entry has missing 'Rating' and a column shift happened for next columns..
10472 Life Made WI-Fi Touchscreen Photo Frame 1.9 19.0 3.0M 1,000+ Free 0 Everyone NaN February 11, 2018 1.0.19 4.0 and up NaN

You'll notice some apps have duplicate entries



```python
for app in android:
    name = app[0]
    if name == 'Instagram':
        print(app)
```

```
['Instagram', 'SOCIAL', '4.5', '66577313', 'Varies with device', '1,00
0,000,000+', 'Free', '0', 'Teen', 'Social', 'July 31, 2018', 'Varies wi
th device', 'Varies with device']
['Instagram', 'SOCIAL', '4.5', '66577446', 'Varies with device', '1,00
0,000,000+', 'Free', '0', 'Teen', 'Social', 'July 31, 2018', 'Varies wi
th device', 'Varies with device']
['Instagram', 'SOCIAL', '4.5', '66577313', 'Varies with device', '1,00
0,000,000+', 'Free', '0', 'Teen', 'Social', 'July 31, 2018', 'Varies wi
th device', 'Varies with device']
['Instagram', 'SOCIAL', '4.5', '66509917', 'Varies with device', '1,00
0,000,000+', 'Free', '0', 'Teen', 'Social', 'July 31, 2018', 'Varies wi
th device', 'Varies with device']
```

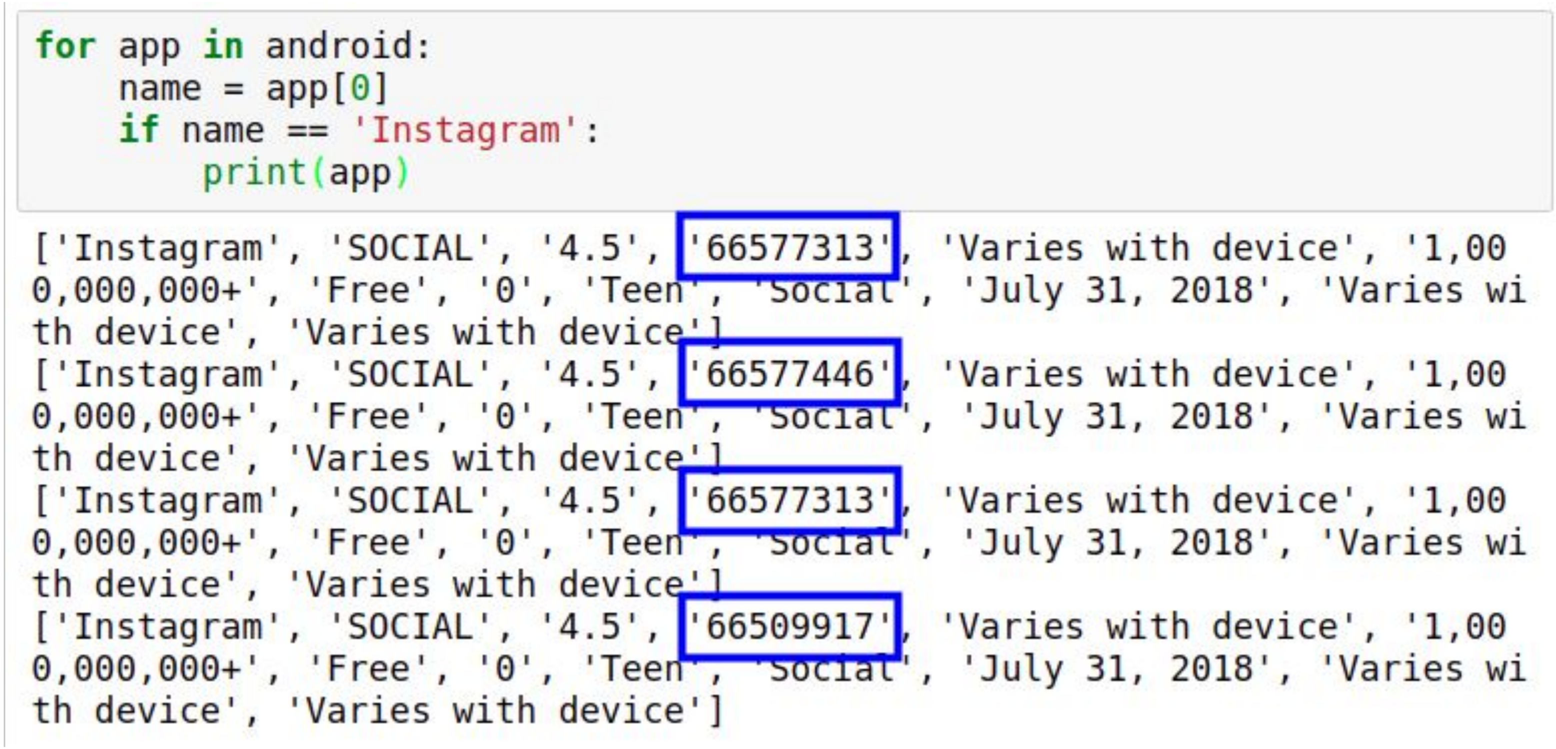


```python
duplicate_apps = []
unique_apps = []

for app in android:
    name = app[0]
    if name in unique_apps:
        duplicate_apps.append(name)
    else:
        unique_apps.append(name)

print('Number of duplicate apps:', len(duplicate_apps))
print('\n')
print('Examples of duplicate apps:', duplicate_apps[:15])
```

```
Number of duplicate apps: 1181


Examples of duplicate apps: ['Quick PDF Scanner + OCR FREE', 'Box', 'Go
ogle My Business', 'ZOOM Cloud Meetings', 'join.me - Simple Meetings',
'Box', 'Zenefits', 'Google Ads', 'Google My Business', 'Slack', 'FreshB
ooks Classic', 'Insightly CRM', 'QuickBooks Accounting: Invoicing & Exp
enses', 'HipChat - Chat Built for Teams', 'Xero Accounting Software']
```

we don't want to count certain apps more than once when we analyze data

We could remove the duplicate rows randomly, but we could probably find a better way ...



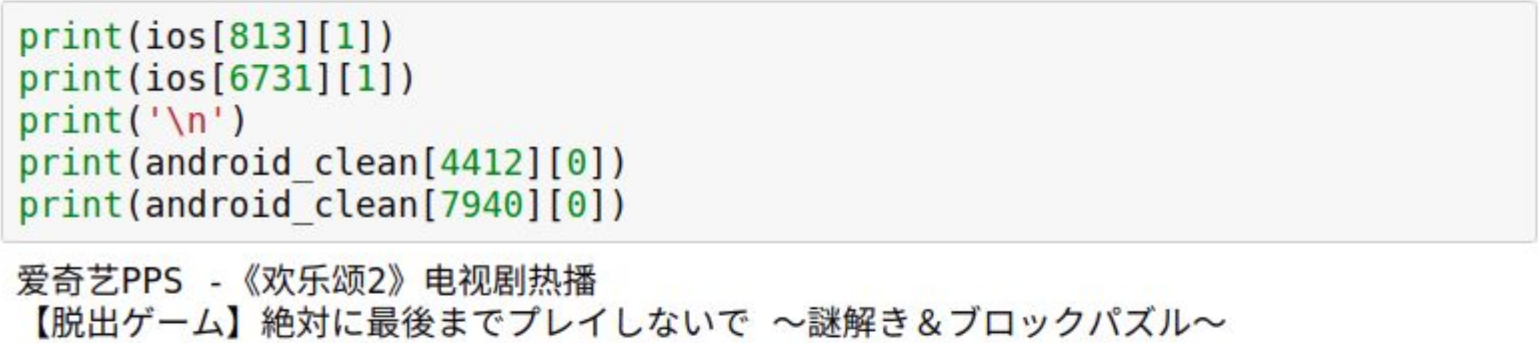
```python
for app in android:
    name = app[0]
    if name == 'Instagram':
        print(app)
```

```
['Instagram', 'SOCIAL', '4.5', '66577313', 'Varies with device', '1,00
0,000,000+', 'Free', '0', 'Teen', 'Social', 'July 31, 2018', 'Varies wi
th device', 'Varies with device']
['Instagram', 'SOCIAL', '4.5', '66577446', 'Varies with device', '1,00
0,000,000+', 'Free', '0', 'Teen', 'Social', 'July 31, 2018', 'Varies wi
th device', 'Varies with device']
['Instagram', 'SOCIAL', '4.5', '66577313', 'Varies with device', '1,00
0,000,000+', 'Free', '0', 'Teen', 'Social', 'July 31, 2018', 'Varies wi
th device', 'Varies with device']
['Instagram', 'SOCIAL', '4.5', '66509917', 'Varies with device', '1,00
0,000,000+', 'Free', '0', 'Teen', 'Social', 'July 31, 2018', 'Varies wi
th device', 'Varies with device']
```

4 5

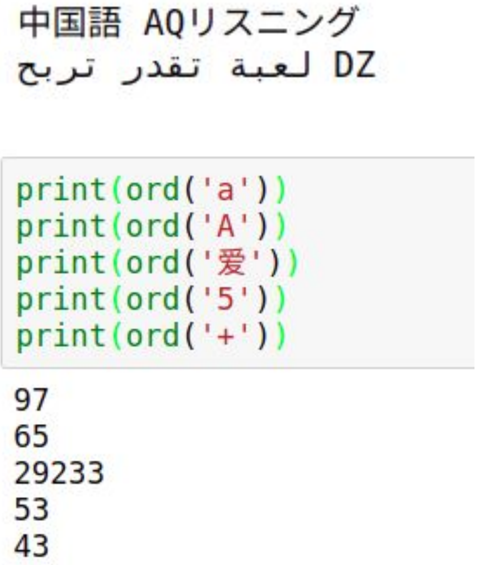
Remove duplicate entries

- Create a dictionary where each key is a unique app name and the corresponding dictionary value is the highest number of reviews of that app
- Use the dictionary you created above to remove the duplicate rows



```
print(ios[813][1])
print(ios[6731][1])
print('\n')
print(android_clean[4412][0])
print(android_clean[7940][0])
```

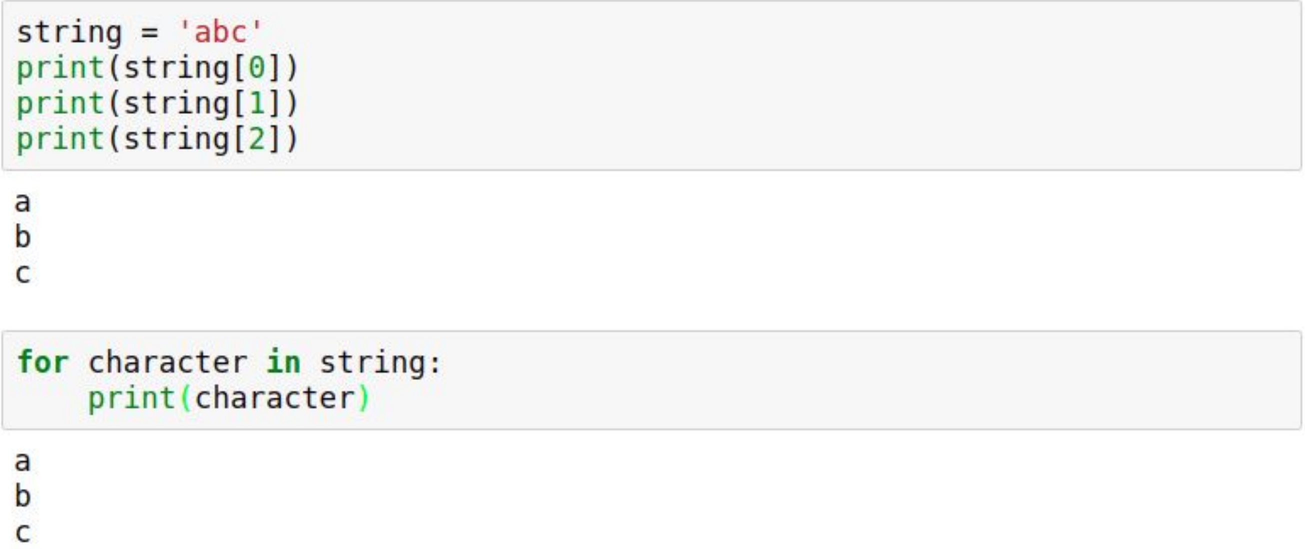
```
爱奇艺PPS -《欢乐颂2》电视剧热播
【脱出ゲーム】絶対に最後までプレイしないで 〜謎解き＆ブロックパズル〜
```



```
中国語 AQリスニング
لعبة تقدر تربح DZ
```

```
print(ord('a'))
print(ord('A'))
print(ord('爱'))
print(ord('5'))
print(ord('+'))
```

```
97
65
29233
53
43
```



```
string = 'abc'
print(string[0])
print(string[1])
print(string[2])
```

```
a
b
c
```

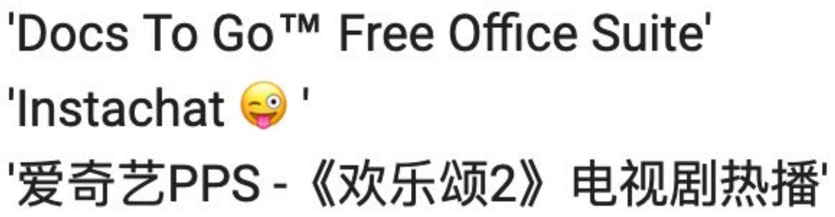
```
for character in string:
    print(character)
```

```
a
b
c
```

6 7

Remove non-english apps

Write a function that takes in a string and returns False if there's any character in the string that doesn't belong to the set of common English characters, otherwise it returns True



'Docs To Go™ Free Office Suite'

'Instachat 😜'

'爱奇艺PPS -《欢乐颂2》电视剧热播'

8

Isolating free apps

Loop through each data set to isolate the free apps in separate lists

After you isolate the free apps, check the length of each data set to see how many apps you have remaining (Android - 8864, iOS - 3222)

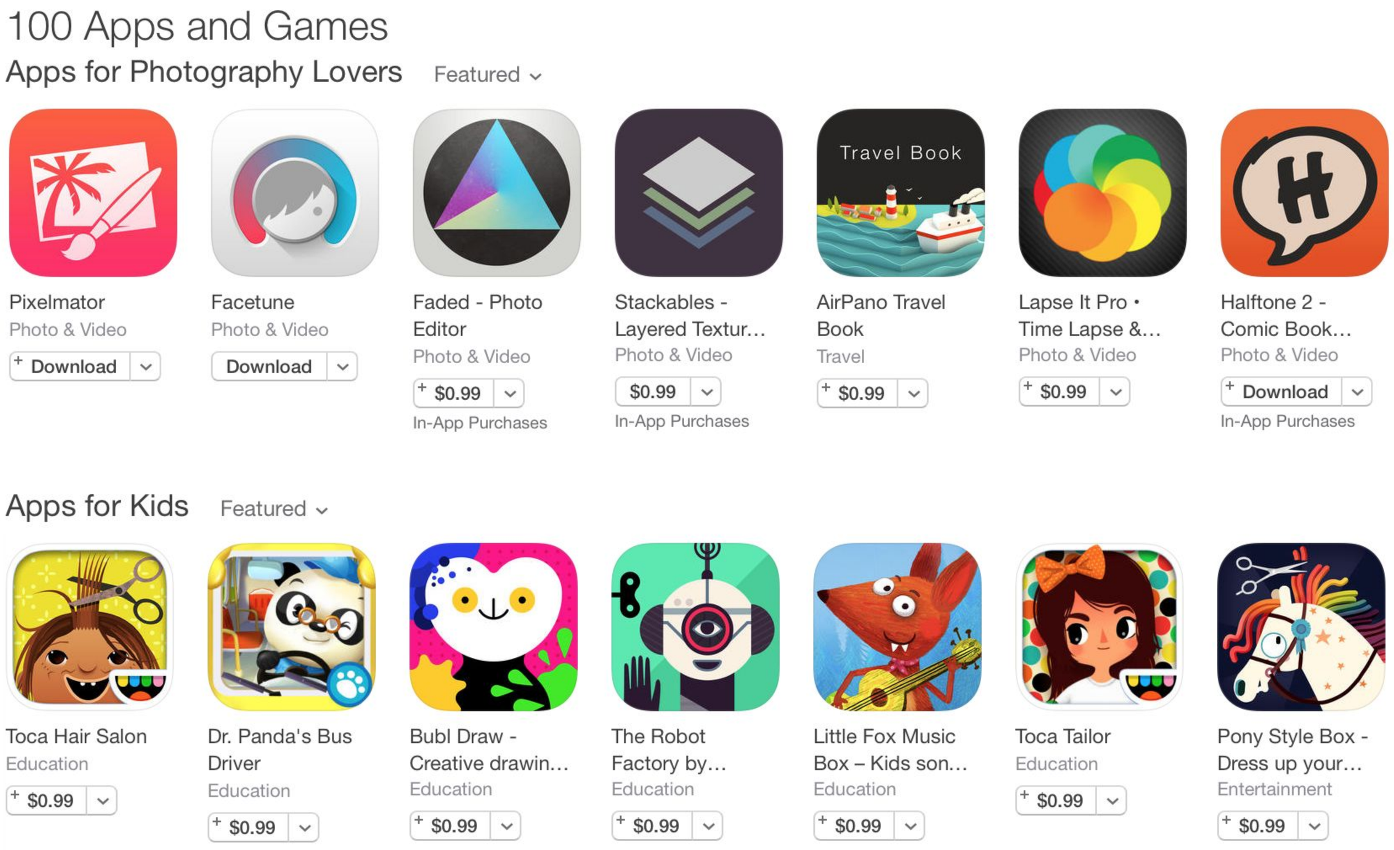
Analysis

Our aim is to determine the kinds of apps that are likely to attract more users because the revenue is highly influenced by the number of people using our apps

IDEA VALIDATION

1. Build a minimal Android version of the app, and add it to Google Play.
2. If the app has a good response from users, we then develop it further.
3. If the app is profitable after six months, we also build an iOS version of the app and add it to the App Store.

What are the most common genres for each market?



100 Apps and Games

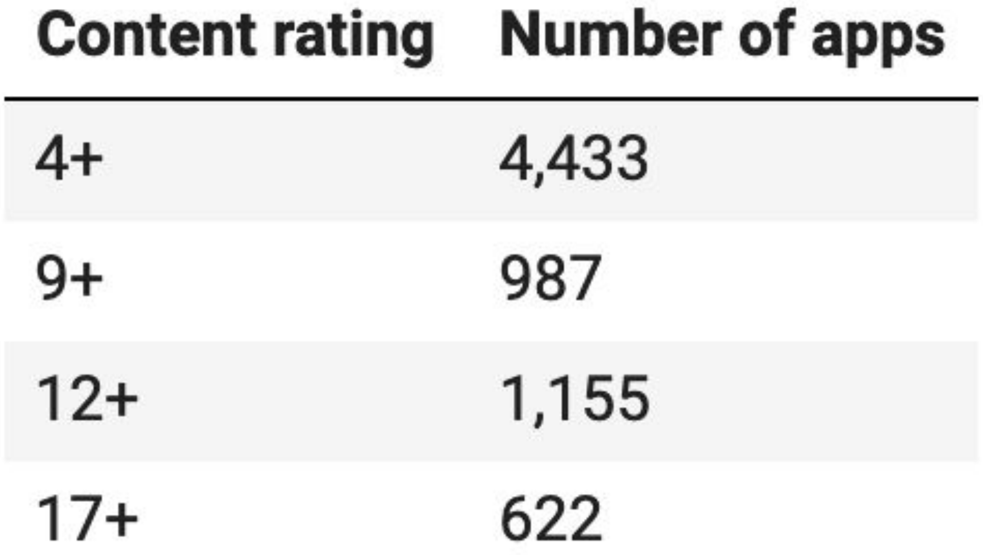
Apps for Photography Lovers Featured ⌄

- Pixelmator
  Photo & Video
  + Download ⌄
- Facetune
  Photo & Video
  Download ⌄
- Faded - Photo Editor
  Photo & Video
  + $0.99 ⌄
  In-App Purchases
- Stackables - Layered Textur…
  Photo & Video
  $0.99 ⌄
  In-App Purchases
- AirPano Travel Book
  Travel
  + $0.99 ⌄
- Lapse It Pro • Time Lapse &…
  Photo & Video
  + $0.99 ⌄
- Halftone 2 - Comic Book…
  Photo & Video
  + Download ⌄
  In-App Purchases

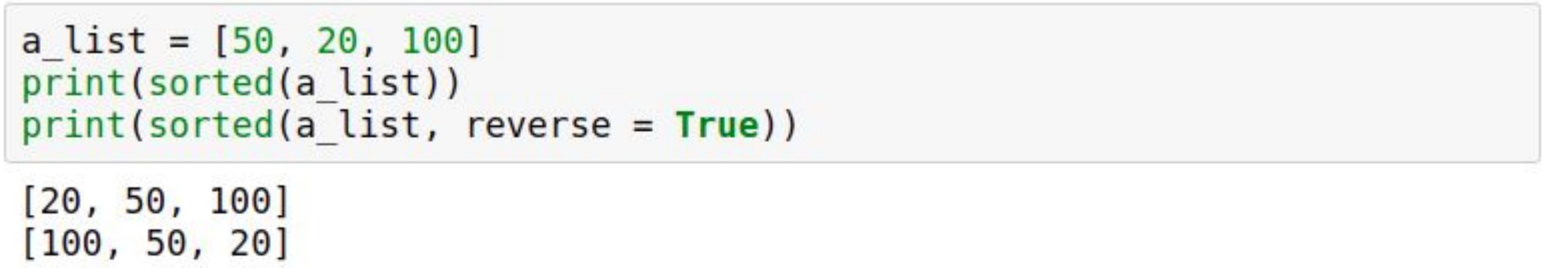
Apps for Kids Featured ⌄

- Toca Hair Salon
  Education
  + $0.99 ⌄
- Dr. Panda's Bus Driver
  Education
  + $0.99 ⌄
- Bubl Draw - Creative drawin…
  Education
  + $0.99 ⌄
- The Robot Factory by…
  Education
  + $0.99 ⌄
- Little Fox Music Box – Kids son…
  Education
  + $0.99 ⌄
- Toca Tailor
  Education
  + $0.99 ⌄
- Pony Style Box - Dress up your…
  Entertainment
  + $0.99 ⌄

Frequency Table (lesson #3)

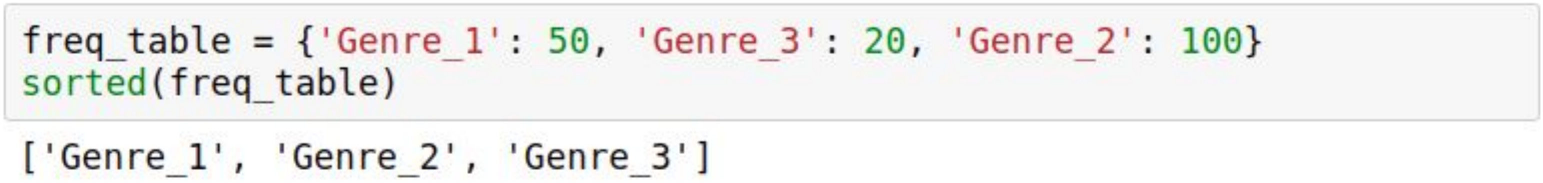


Content rating	Number of apps
4+	4,433
9+	987
12+	1,155
17+	622



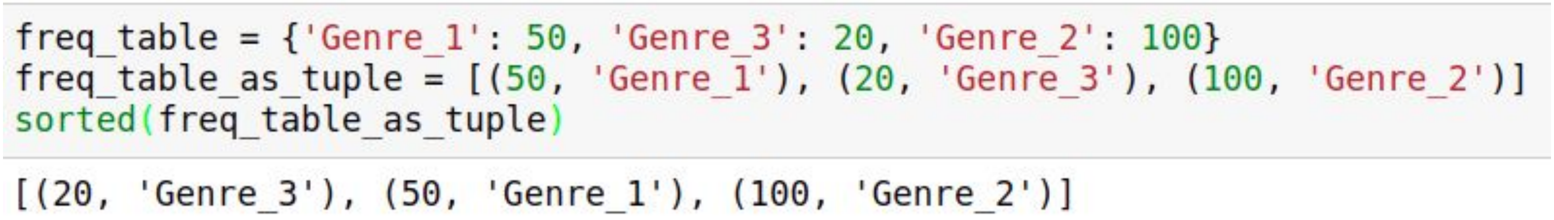
```
a_list = [50, 20, 100]
print(sorted(a_list))
print(sorted(a_list, reverse = True))
```

```
[20, 50, 100]
[100, 50, 20]
```



```
freq_table = {'Genre_1': 50, 'Genre_3': 20, 'Genre_2': 100}
sorted(freq_table)
```

```
['Genre_1', 'Genre_2', 'Genre_3']
```



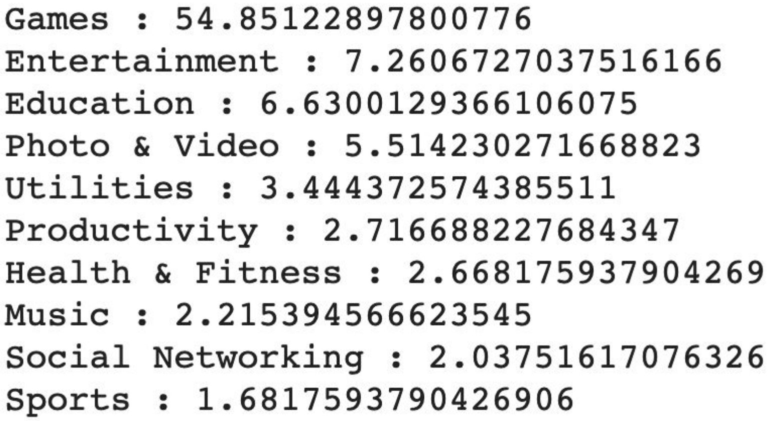
```
freq_table = {'Genre_1': 50, 'Genre_3': 20, 'Genre_2': 100}
freq_table_as_tuple = [(50, 'Genre_1'), (20, 'Genre_3'), (100, 'Genre_2')]
sorted(freq_table_as_tuple)
```

```
[(20, 'Genre_3'), (50, 'Genre_1'), (100, 'Genre_2')]
```

9 10 11

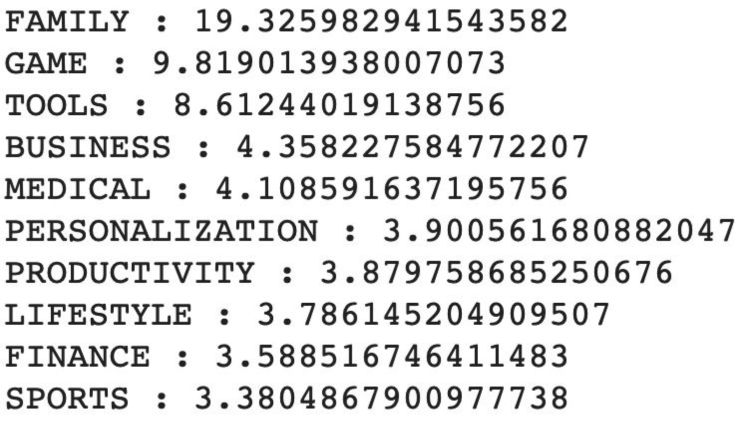
Most common apps by genre

iOS



```
Games : 54.85122897800776
Entertainment : 7.260672703751616
Education : 6.630012936610607
Photo & Video : 5.514230271668823
Utilities : 3.444372574385511
Productivity : 2.716688227684347
Health & Fitness : 2.668175937904269
Music : 2.215394566623545
Social Networking : 2.037516170763260
Sports : 1.681759379042690
```

Android



```
FAMILY : 19.325982941543582
GAME : 9.819013938007073
TOOLS : 8.612440191387560
BUSINESS : 4.358227584772207
MEDICAL : 4.108591637195756
PERSONALIZATION : 3.900561680882047
PRODUCTIVITY : 3.879758685250676
LIFESTYLE : 3.786145204909507
FINANCE : 3.588516746411483
SPORTS : 3.380486790097738
```

App Store is dominated by apps designed for fun, while Google Play shows a more balanced landscape of both practical and for-fun apps

12 13

Most popular apps by genre

Calculate the average number of installs for each app genre

- For the Google Play
  - Installs column
- For App Store
  - rating_count_tot column

Project#01.ipynb