COMP8430 Data wrangling project

u5976992

Word count: 1339 words

1. Data and problem description and overall strategy

1.1 Problem description

Nowadays, Android and iOS are the most popular mobile systems. I'm very interested in the difference of user preference for apps between the two different systems. I choose Google play store data set and Apple app store data set. The hypothetical end-use of them is to find the linkage of rating in Google play store and rating in Apple app store for the same app.

1.2 Overall strategy

Firstly, we need to analysis each data set. In this step, I will decide features that are used in the future work and find some useful information such as missing values, incorrect values, features type. Using the information to get data quality. Secondly, we need to do data preparation and cleaning. In this step, I will remove duplicate records and deal with records that contain missing value or incorrect value. I will change features type to correct one. Thirdly, we need to do record linkage. I will simplify those app name to find more record linage. Merge them as a new table.

2. Data description and data exploration

2.1 Google play store data set description

Google play store data set comes from https://www.kaggle.com/lava18/google-play-store-apps. It's collected in 2018 by Lavanya Gupta, a software developer at HSBC Software Development. The data set contains 10841 records. There are 13 features as follow,

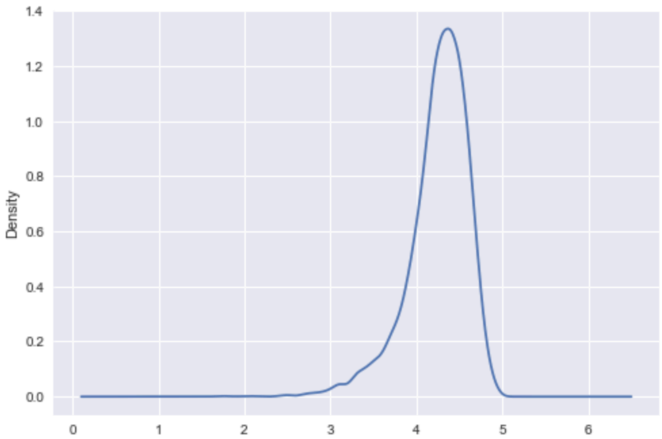
No	Feature Name	Type	Description
1	App	string	application name
2	Category	string	category the app belongs to
3	Rating	float	overall user rating of the app
4	Reviews	string	the number of user reviews for the app
5	Size	string	size of the app
6	Installs	string	the number of user downloads/installs for the app
7	Type	string	paid/free
8	Price	string	price of the app
9	Content Rating	string	age group the app is targeted at children/mature21+/adult
10	Genres	string	genres the app belongs to
11	Last Updated	string	date when the app was last updated on Play Store

12	Current Ver	string	current version
13	Android Ver	string	minimax required Android version

Since we need to find the linkage of rating between Google play store and Apple app store for the same app, I want to focus on such features as follow,

1. App: as the key to merge tables
2. Rating: use it the compute the correlation
3. Reviews: we need to use this to remove those apps that too few people rate since those ratings may be very bias.

According to my experiment, there is only one record rating greater than 5, which is an unexpected value. After removing that, we can find the density of rating as below, which imply that most apps rating is the range of 3 to 5.



the density of rating of Google play store data set

2.2 Apple app store data set description

Apple app store data set comes from https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps. The data was extracted from the iTunes Search API in July 2017 by Ramanathan, a research engineer at KIT. The data set contains 7198 records. There are 17 features as follow,

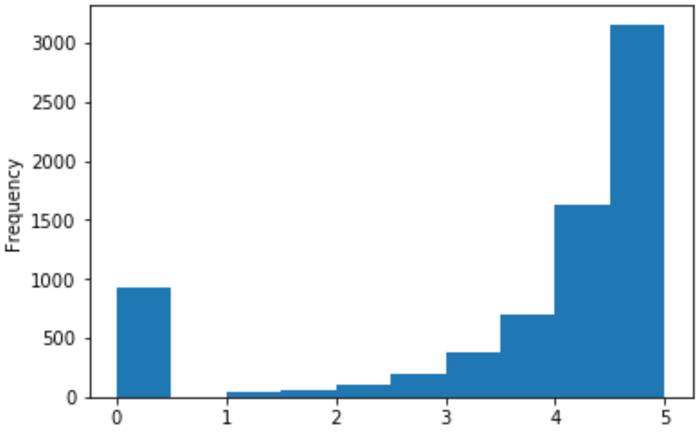
No	**Feature Name**	**Type**	**Description**
1	Unnamed: 0	int	unknown
2	id	int	App ID
3	track_name	string	application name
4	size_bytes	int	size of the app
5	currency	string	currency type
6	price	float	price of the app
7	rating_count_tot	int	user rating counts for all version
8	rating_count_ver	int	user rating counts for current version

9	user_rating	float	user rating for all version
10	user_rating_ver	float	user rating for current version
11	ver	string	latest version
12	cont_rating	string	content rating
13	prime_genre	string	genres the app belongs to
14	sup_devices.num	int	the number of supporting devices
15	ipadSc_urls.num	int	the number of screenshots showed for display
16	lang.num	int	the number of supported languages
17	vpp_lic	int	vpp device based licensing enabled

For this data set, we need features as follow,

1. track_name: as the key to merge tables
2. user_rating: use it to compute correlation
3. rating_count_tot: we need to use this to remove those apps that too few people rate since those ratings may be very bias.

Notice that the rating in the Google play store data set is discrete. However, in the Apple app store data set, the rating is divided exactly by 0.5. Therefore, we can find the histogram of rating as below. There are too many 0 and 5 value records, maybe since too few people rate those apps.



the histogram of rating of Apple app store data set

3. Data quality assessment

3.1 Google play store data set

- App name: the value should be a string. There are 9660 unique values. We need to deal with those duplicate records later.
- Rating: the value should be a number and in the range of 0 to 5. There are 1474 records missing this value. There is one record which Rating value is 19.0.

- Reviews: the value should be a number and greater than or equal to 0. There is one record which Review value is ‘3.0M’. We need to convert it to 3000000.
- Installs: the value should be one of ‘0+’, ‘1+’, ‘5+’, ‘10+’, ‘50+’, ‘100+’, ‘500+’, ..., ‘1,000,000,000+’. There is one record which Installs is ‘Free’ and one record which Install is ‘0’.
- Price: the value should be greater than or equal to 0. There is one record which Price value is ‘Everyone’.

	App name	**Rating**	**Reviews**
Completeness	100%	86.4%	100%
Uniqueness	89.1%	No need	No need
Validity	100%	99.9%	99.9%

3.2 Apple app store data set

track_name: the value should be a string. There are 7195 unique values.

user_rating: the value should be divided exactly by 0.5. All records values are valid.

rating_count_tot: the value should be a number and greater than or equal to 0. All records values are valid.

	track_name	**user_rating**	**rating_count_tot**
Completeness	100%	100%	100%
Uniqueness	99.9%	No need	No need
Validity	100%	100%	100%

4. Data preparation and cleaning

4.1 Google play store data set

Firstly, we need to remove duplicate record for the same app. I find the records that contains the same app name and keep the latest version one.

Secondly, we remove the records that rating is greater than 5 or smaller than 0.

Thirdly, I convert one record that Review is ‘3.0M’ to 3000000. Then change Review type from Sting to int.

Finally, I remove those records that Reviews is smaller than 1000. I think those records’ rating maybe be too bias.

In the end, we have 4802 records in this data set.

4.2 Apple app store data set

basically we do the same thing. Except, we don’t need to deal with the rating_count_tot since it’s already good enough and I choose 100 as threshold for Apple app store since there the number of users in Apple app store is largely smaller than the number of users in Google play store.

In the end, we have 4489 records in this data set.

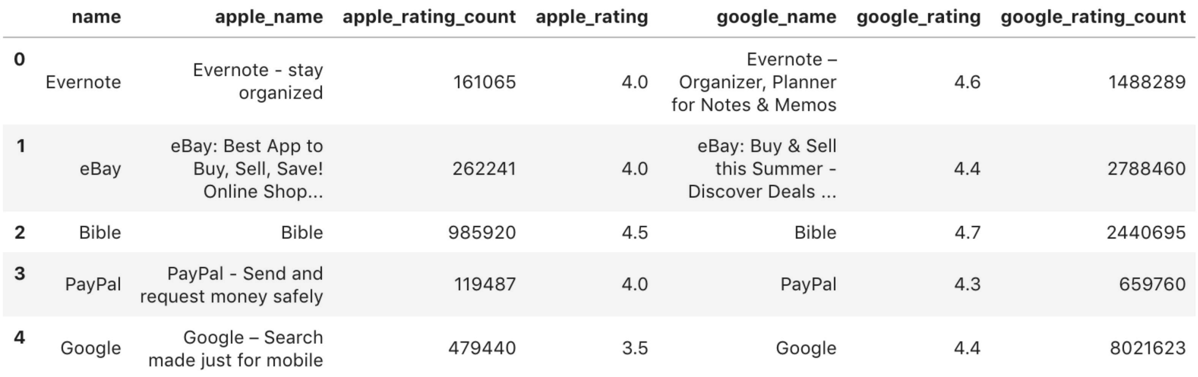
5. Data integration or record linkage

The most difficult part is to find the same app in those two different data sets since the app name is much different in those two data sets and we don't have other features to deal with it. According to my experiment, many apps name follows the format like

- real name: some description
- real name - some description
- real name – some description

Therefore, I extract the real name from the original app name and make it as a new feature. Then, I change the feature name to make the result much more readable.

After merge two tables. We have 579 records. The result table looks like as below,



	name	apple_name	apple_rating_count	apple_rating	google_name	google_rating	google_rating_count
0	Evernote	Evernote - stay organized	161065	4.0	Evernote – Organizer, Planner for Notes & Memos	4.6	1488289
1	eBay	eBay: Best App to Buy, Sell, Save! Online Shop...	262241	4.0	eBay: Buy & Sell this Summer - Discover Deals ...	4.4	2788460
2	Bible	Bible	985920	4.5	Bible	4.7	2440695
3	PayPal	PayPal - Send and request money safely	119487	4.0	PayPal	4.3	659760
4	Google	Google – Search made just for mobile	479440	3.5	Google	4.4	8021623

sample of the final result

We will know the simplified name of the app and their names in different app stores. Also, we have the rating of the app and the number of rating count in different app stores. We can use those data to compute the correlation. The correlation between the rating in Google play store and the rating in Apple app store is 0.63.