BIG DATA (X-LARGE)

WSDM – KKBox's Music Recommendation Challenge

LANGUAGES

LANGUAGES

R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing.

The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

[reference] https://www.kaggle.com/pallabs/pallab-sarkar-kkbox-music-recommendation-eda-i

LANGUAGES

Python is powerful and fast; plays well with others; runs everywhere; is friendly & easy to learn; is Open.

Python is developed under an OSI-approved open source license, making it freely usable and distributable, even for commercial use. Python's license is administered by the Python Software Foundation.

Python can be used in multi-object (productivity), uses other languages' functions (versatility), possesses rich ecosystem of community(combination) , and has openness.

[reference] https://www.Kaggle.com/bellar/basic-eda/notebook

R

Load Libraries

```
# general visualisation
library('ggplot2') # visualisation
library('scales') # visualisation
library('grid') # visualisation
library('ggthemes') # visualisation
library('gridExtra') # visualisation
library('RColorBrewer') # visualisation
library('corrplot') # visualisation
library('ggpubr') # visualisation

# general data manipulation
library('dplyr') # data manipulation
library('readr') # input/output
library('data.table') # data manipulation
library('tibble') # data wrangling
library('tidyr') # data wrangling
```

Load Data

```
> train <- as.tibble(fread('C:/Users/wnlwn/Desktop/input/train.csv',nrows = 1e6))
> summary(train)
     msno              song_id          source_system_tab  source_screen_name
 Length:1000000     Length:1000000     Length:1000000     Length:1000000
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



 source_type           target
 Length:1000000     Min.   :0.0000
 Class :character   1st Qu.:0.0000
 Mode  :character   Median :1.0000
                    Mean   :0.6918
                    3rd Qu.:1.0000
                    Max.   :1.0000
```

Problem : How we are going to deal with large data. We need GPU!

Data Summary

Summary(train)

```
##      msno              song_id         source_system_tab
##  Length:7377418     Length:7377418     Length:7377418
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  source_screen_name source_type          target
##  Length:7377418     Length:7377418     Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :1.0000
##                                        Mean   :0.5035
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
```

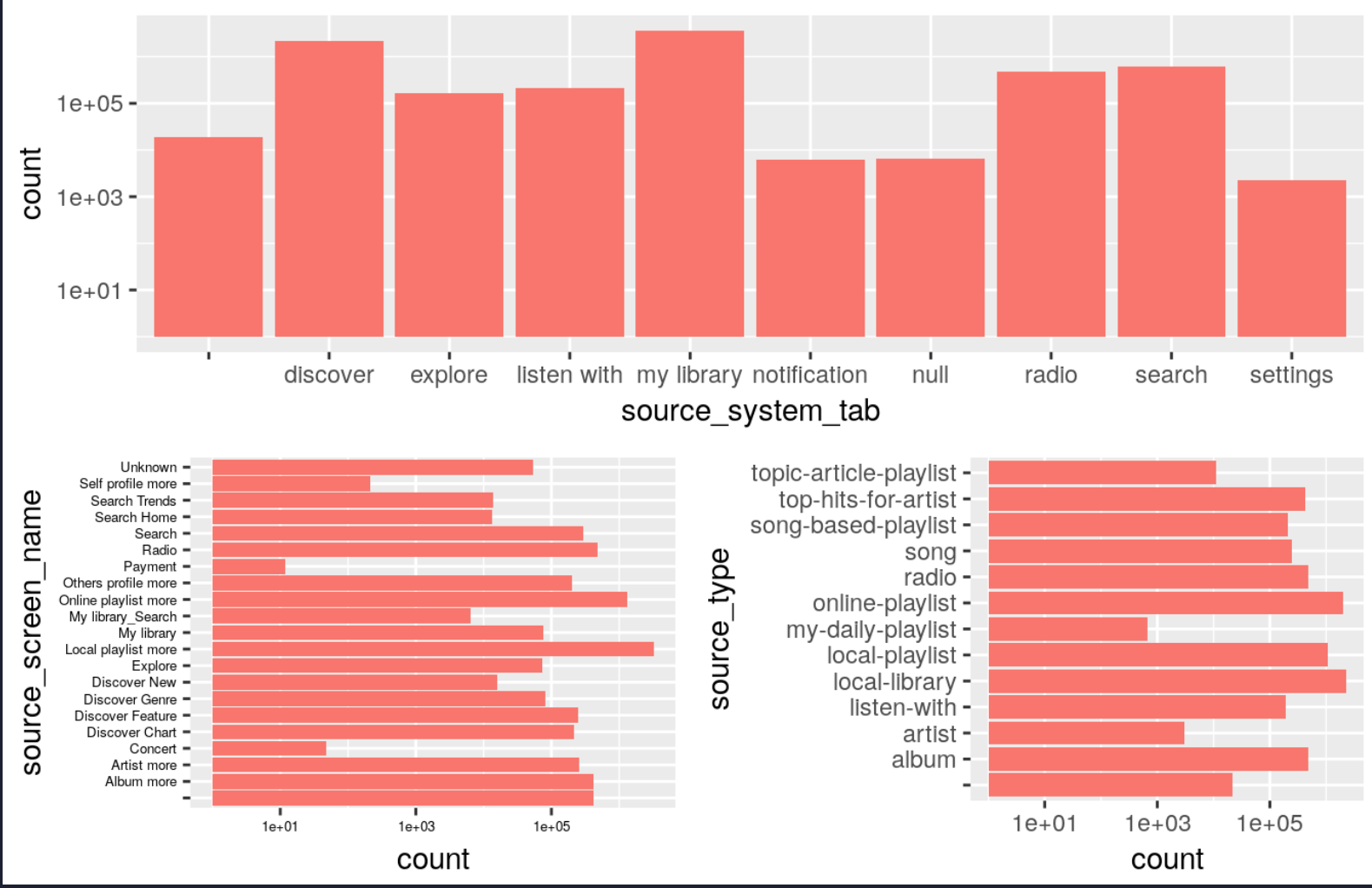
Unique(songs$composer)

```
   [2] "TEDDY| FUTURE BOUNCE| Bekuh BOOM"


   [3] ""


   [4] "疫\xaf弱\xeb쁜"


   [5] "Traditional"


   [6] "Joe Hisaishi"


   [7] "Jonathan Lee"


   [8] "\xe5쮜\xe8\xad"


   [9] "JJ Lin"


  [10] "Stephen Garrigan| Mark Prendergast| Vincent May| Jacknife Lee| Jas
on Boland"
```

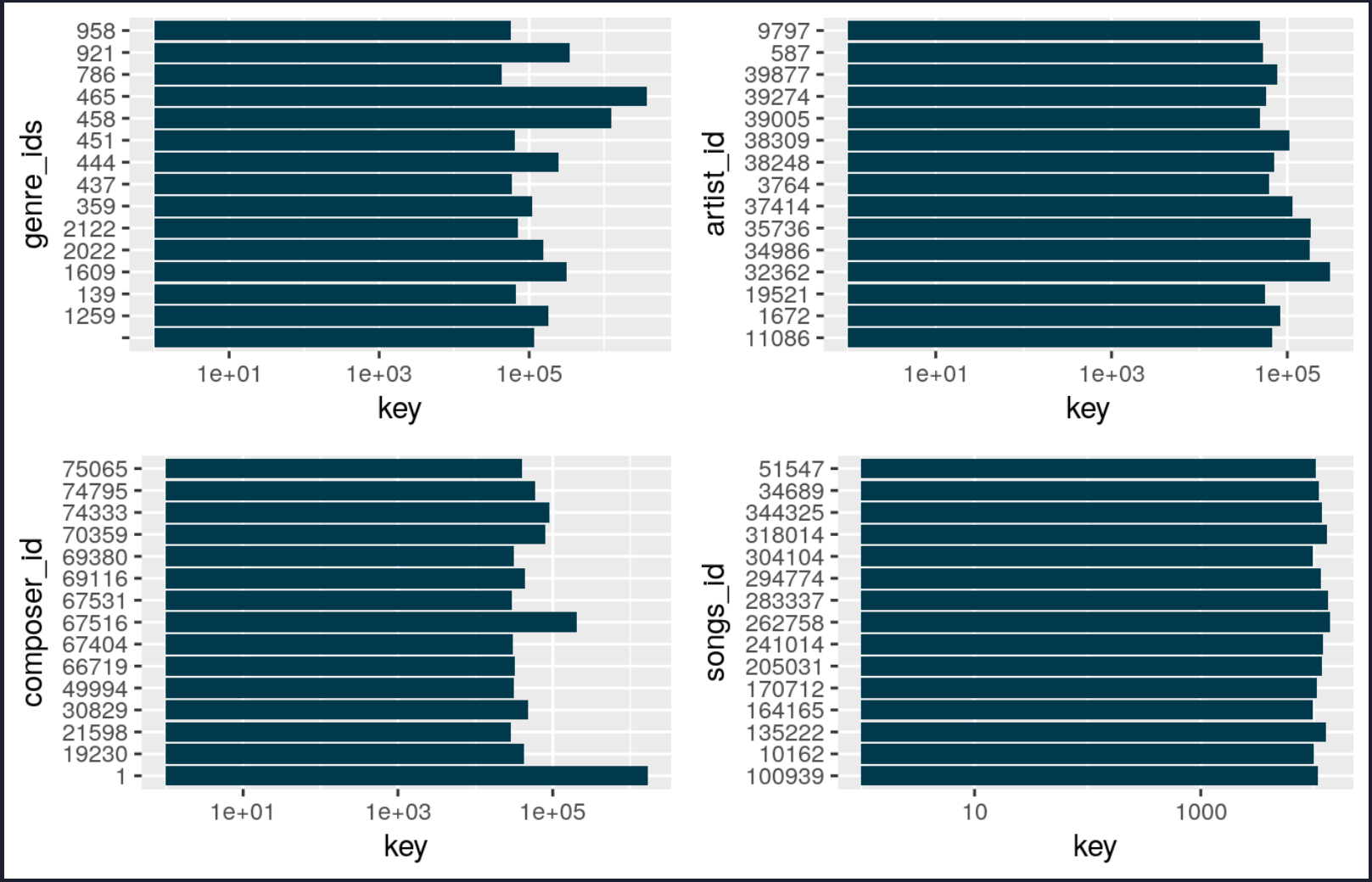
Data Manipulation

```
train$key<-paste(train$msno,train$song_id,sep="_")
test$key<-paste(test$msno,test$song_id,sep="_")

train$id<-row.names(train)

test$target<-''

train$type<-'train'
test$type<-'test'

train1<-train[,c('key','type','id','msno','song_id','source_system_tab','source_screen_name','source_
type','target')]
test1<-test[,c('key','type','id','msno','song_id','source_system_tab','source_screen_name','source_ty
pe','target')]

rm(train)
rm(test)

master_df<-rbind(train1,test1)
```

Data Visualization



Data Visualization



PYTHON

LIBRARYIES

numpy

A library that plays a key role when used in computational science

Provides high-performance multidimensional array objects and tools to handle them

pandas

Provides detailed information processing functions and functions related to matplotilb graphic output

Widely used for data munging and preparation for structured data manipulation and manipulation

Matplotlib.pyplot

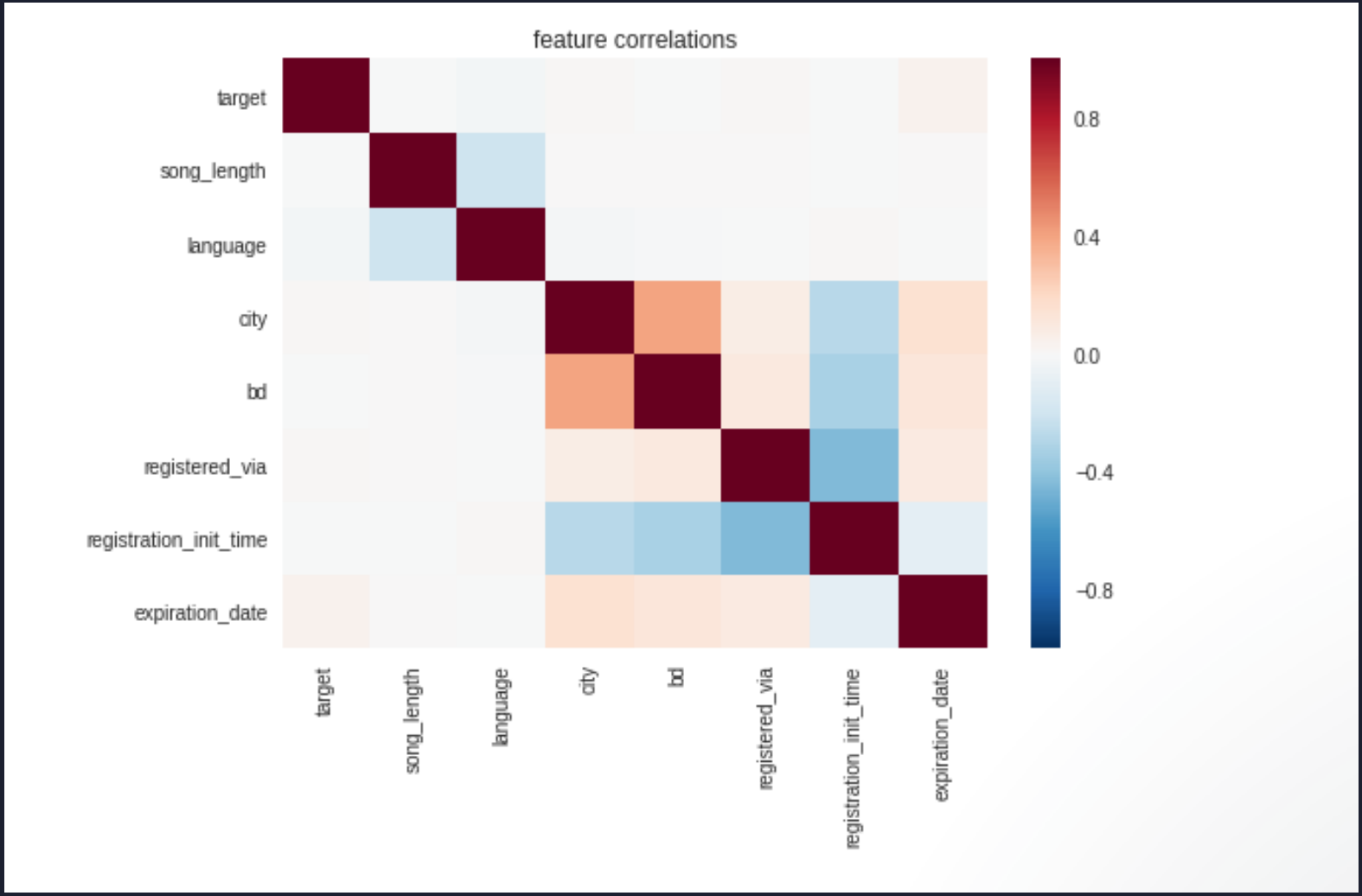
Modules that provide functionality similar to the MATLAB plotting system

seaborn

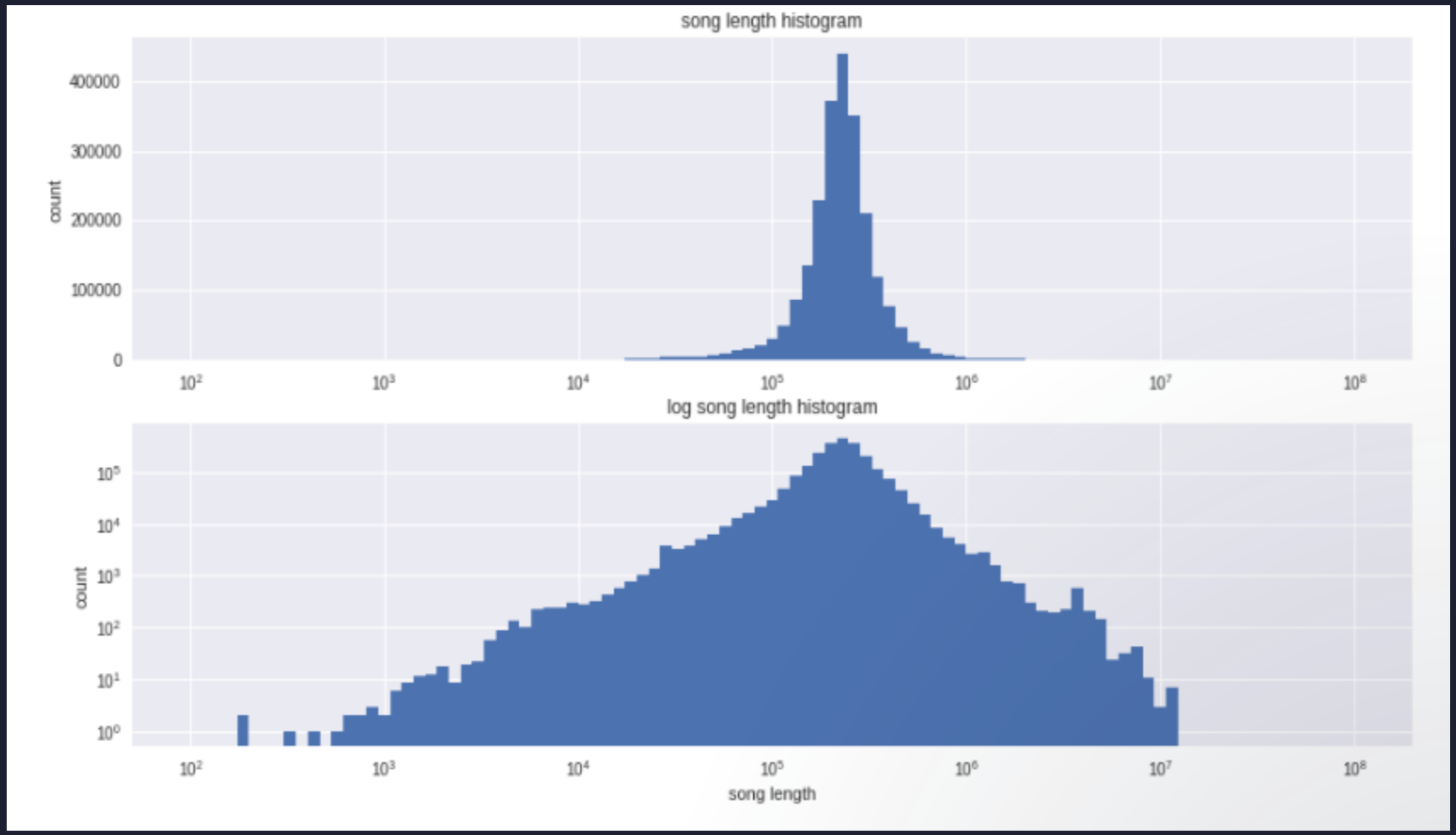
Python visualization library based on Matplotlib

Show attractive statistical graph

Data Visualization



Data Visualization



FEATURES of DATA

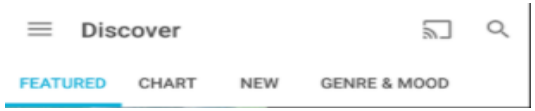
train.csv

- msno: user id
- song_id: song id
- source_system_tab: the name of the tab where the event was triggered. System tabs are used to categorize KKBOX mobile apps functions. For example, tab `my library` contains functions to manipulate the local storage, and tab `search` contains functions relating to search.
- source_screen_name: name of the layout a user sees.
- source_type: an entry point a user first plays music on mobile apps. An entry point could be `album`, `online-playlist`, `song` .. etc.
- target: this is the target variable. `target=1` means there are recurring listening event(s) triggered within a month after the user's very first observable listening event, `target=0` otherwise .

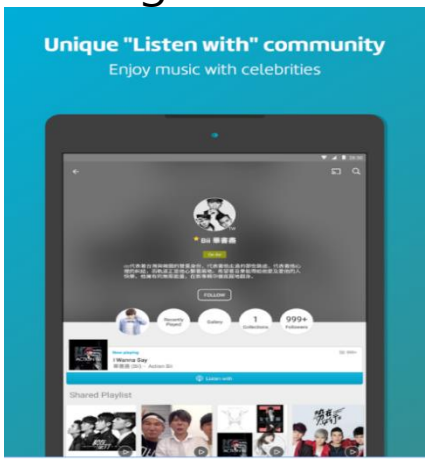
train.csv

Source_system_tab:

- Discover: show the musics by chart, featured(today's recommendation), new, genre&mood



- Explore: suggest more songs based on listening history (by Genre, Home)
- Listen with: listen songs that celebrities or friends recommand
- My library: Menu that my playlist is gathered
- Notification: notification when some artist's album is coming out
- Radio: listen songs like radio
- Search: search songs
- Setting: sort songs by filters
- Null



train.csv

Source_screen_name:

- Unknown
- Self-profile-more
- Others profile more
- Local-playlist-more
- My-library-search
- My-library
- Search-home
- Search-trends
- Search
- Radio
- Payment
- Explore
- Discover new
- Discover Genres
- Discover Feature
- Discover Chart
- Concert
- Artist more
- Album more

train.csv

Source_type:

- Topic-article-playlist
- Song-based-playlist
- Song
- Radio
- Online-playlist
- My-daily-playlist
- Local-playlist
- Local-library
- Listen-with
- Artist
- Album

songs.csv

The songs. Note that data is in unicode.

- song_id
- song_length: in ms
- genre_ids: genre category. Some songs have multiple genres and they are separated by |
- artist_name
- composer
- lyricist
- language

* What about Release date?

members.csv

user information.

- msno
- city
- bd: age. Note: this column has outlier values, please use your judgement.
- gender
- registered_via: registration method
- registration_init_time: format `%Y%m%d`
- expiration_date: format `%Y%m%d`

Domain Analysis
on Customer Behaviors

Why we need domain analysis

Labels in Classification Data

While computers treat a label as just one of many labels in the models, domain analysis lets the human analyzers aware of the underlying structures and target behaviors from the label name.

Model Enhancement

Analyzers can enhance, separate and combine models based on the domain knowledge. Analyzers can find which prototypes of models are suitable from the domain knowledge.

Last week

We made hypotheses of the domain, especially for people behaviors

Recall the problem Statement :

From the records of song listening,
determine whether the listeners
listen to the songs again on next 30 days.

Hypotheses on Customer Behaviors

1. Searchers Hypothesis

   Searching songs can represent the active interest of the customer for the music. Therefore, searchers are likely to listen the songs again.

2. My Library Hypothesis

   Songs saved in My Library are intended to listen later. Therefore, people are tend to listen those songs again.

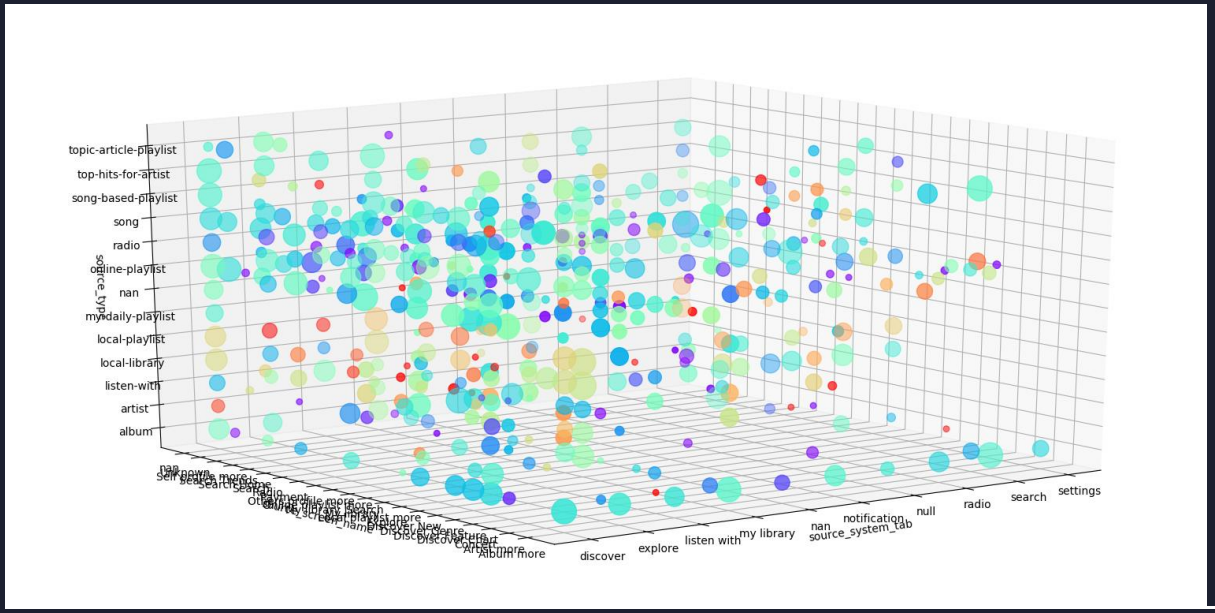
3. Radio Listeners Hypothesis

   Songs from radio are volatile. As radio is the most passive way to listen songs, people would not listen again unless the songs are popular or they actively look for the songs.

Verifying Hypotheses

Hypotheses can be verified easily by analyzing train data only.

Train data contains: target value, 3 in-app menus of songs



Red: High Ratio
Green: About 50%
Blue: Low Ratio

Log-scaled Dot Size on data count

Significant: >10%p deviation from the average (50.35%)

Naïve Ratio-by-menus

Units: Ratio(%), Deviation(%p)

source_system_tab	Count	Ratio	Deviation
discover	2179252	0.41577	-8.77474
explore	167949	0.422146	-8.13711
listen with	212266	0.326581	-17.6936
my library	3684730	0.619659	11.61419
nan	18371	0.537423	3.390603
notification	6185	0.378011	-12.5506
null	6478	0.433621	-6.98956
radio	476701	0.222662	-28.0855
search	623286	0.421362	-8.21551
settings	2200	0.590909	8.739201

source_type	Count	Ratio	Deviation
album	477344	0.393421	-11.0096
artist	3038	0.572745	6.922814
listen-with	192842	0.31965	-18.3867
local-library	2261399	0.632098	12.8581
local-playlist	1079503	0.657719	15.42023
my-daily-playlist	663	0.375566	-12.7951
nan	21539	0.525233	2.171621
online-playlist	1967924	0.424985	-7.85317
radio	483109	0.219735	-28.3782
song	244722	0.43746	-6.60574
song-based-playlist	210527	0.38046	-12.3058
top-hits-for-artist	423614	0.418537	-8.49804
topic-article-playlist	11194	0.494283	-0.92344

source_screen_name	Count	Ratio	Deviation
Album more	420156	0.390553	-11.2965
Artist more	252429	0.416858	-8.66593
Concert	47	0.510638	0.712121
Discover Chart	213658	0.517032	1.351481
Discover Feature	244246	0.364104	-13.9413
Discover Genre	82202	0.347887	-15.563
Discover New	15955	0.455531	-4.79859
Explore	72342	0.448149	-5.5368
Local playlist more	3228202	0.636983	13.34663
My library	75980	0.657173	15.36559
My library_Search	6451	0.611223	10.7706
Online playlist more	1294689	0.414939	-8.85781
Others profile more	201795	0.312629	-19.0888
Payment	12	0.666667	16.31496
Radio	474467	0.217256	-28.6261
Search	298487	0.471749	-3.17679
Search Home	13482	0.353583	-14.9935
Search Trends	13632	0.376981	-12.6536
Self profile more	212	0.424528	-7.89888
Unknown	54170	0.339155	-16.4363
nan	414804	0.469788	-3.37289

High Ratio: my library, local library/playlists

Low Ratio: radio(less than -28%p), listen-with, Others profile

Acceptance of Hypotheses

1. Searchers Hypothesis (Not accepted)

   Searching have no significant difference from the other sources of songs, on whether listening the songs again.
2. My Library Hypothesis (Accepted)
3. Radio Listeners Hypothesis (Accepted, strong)

4. New Phenomenon Found: Listen-with others profile

   Though less significant than radio listeners, listeners from listen-with page tend to have low re-listening rate.

THANKYOU

WSDM – KKBox's Music Recommendation Challenge