Music platform membership(adopter) analysis

Eric Yang

Summary statistics

```
library(pastecs)
s1<-stat.desc(data_adopter)
s2<-stat.desc(data_non_adopter)
setwd('/Users/yanghaoying/Desktop/')
write.csv(s1,'s1.csv')
write.csv(s2,'s2.csv')
```

Statistics table

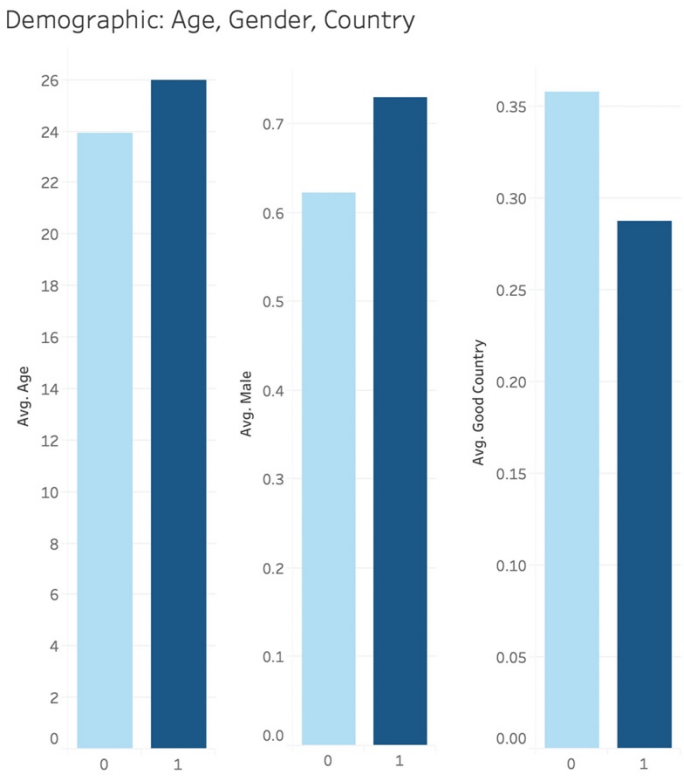
adopter	N Obs	Lable	mean	missing	min	max	median	std.dev
1	3527	age	25.98	0.00	8.00	73.00	24.00	6.84
		male	0.73	955.00	0.00	1.00	1.00	0.44
		friend_cnt	39.73	0.00	1.00	5089.00	16.00	117.27
		avg_friend_age	25.44	0.00	12.00	62.00	24.36	5.21
		avg_friend_male	0.64	130.00	0.00	1.00	0.67	0.25
		friend_country_cnt	7.19	7.00	0.00	136.00	4.00	8.86
		subscriber_friend_cnt	1.64	1783.00	0.00	287.00	0.00	5.85
		songsListened	33758.04	1.00	0.00	817290.00	20908.00	43592.73
		lovedTracks	264.34	197.00	0.00	10220.00	108.00	491.43
		posts	21.20	2158.00	0.00	8506.00	0.00	221.99
		playlists	0.90	1598.00	0.00	118.00	1.00	2.56
		shouts	99.44	241.00	0.00	65872.00	9.00	1156.07
		adopter	1.00	0.00	1.00	1.00	1.00	0.00
		tenure	45.58	1.00	0.00	111.00	46.00	20.04
		good_country	0.29	2513.00	0.00	1.00	0.00	0.45
0	40300	age	23.95	0.00	8.00	79.00	23.00	6.37
		male	0.62	15239.00	0.00	1.00	1.00	0.48
		friend_cnt	18.49	0.00	1.00	4957.00	7.00	57.48
		avg_friend_age	24.01	0.00	8.00	77.00	23.00	5.10
		avg_friend_male	0.62	4398.00	0.00	1.00	0.67	0.32
		friend_country_cnt	3.96	262.00	0.00	129.00	2.00	5.76
		subscriber_friend_cnt	0.42	32221.00	0.00	309.00	0.00	2.42
		songsListened	17589.44	1446.00	0.00	1000000.00	7440.00	28416.02
		lovedTracks	86.82	9607.00	0.00	12522.00	14.00	263.58
		posts	5.29	31464.00	0.00	12309.00	0.00	104.31
		playlists	0.55	21880.00	0.00	98.00	0.00	1.07
		shouts	29.97	3311.00	0.00	7736.00	4.00	150.69
		adopter	0.00	40300.00	0.00	0.00	0.00	0.00
		tenure	43.81	0.00	1.00	111.00	44.00	19.79
		good_country	0.36	25881.00	0.00	1.00	0.00	0.48

From summary statistics, we can draw following tentative conclusions from mean values:

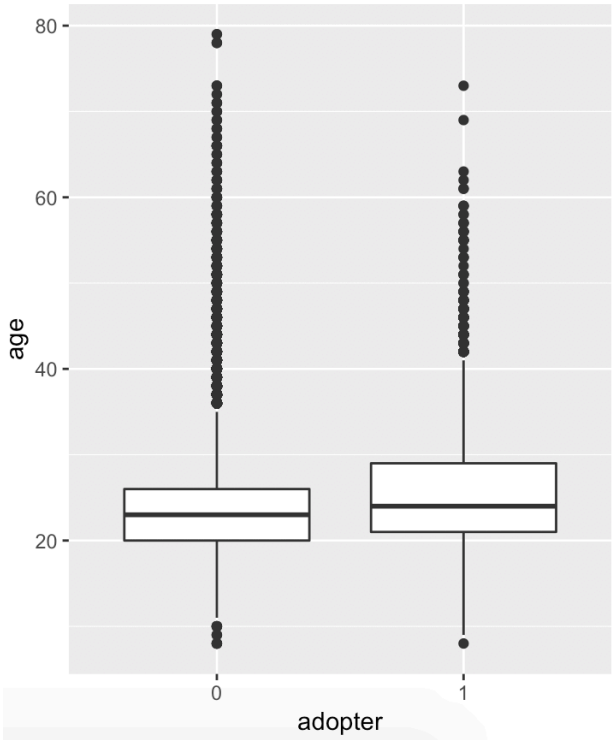
1. Average age of adopters and their friends are older than non-adopters and their friends.
2. Adopters have more friends, friend_country_cnt, songsListened, lovedTracks, posts, playlists and shouts than non-adopters, which means adopters are more active than non-adopter.
3. Adopters have been on the site longer than non-adopters, they are elder users.

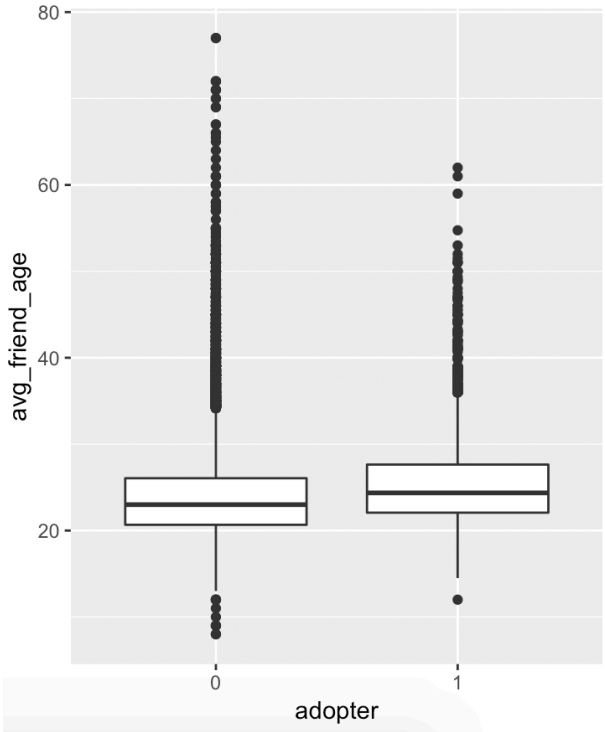
Data Visualization

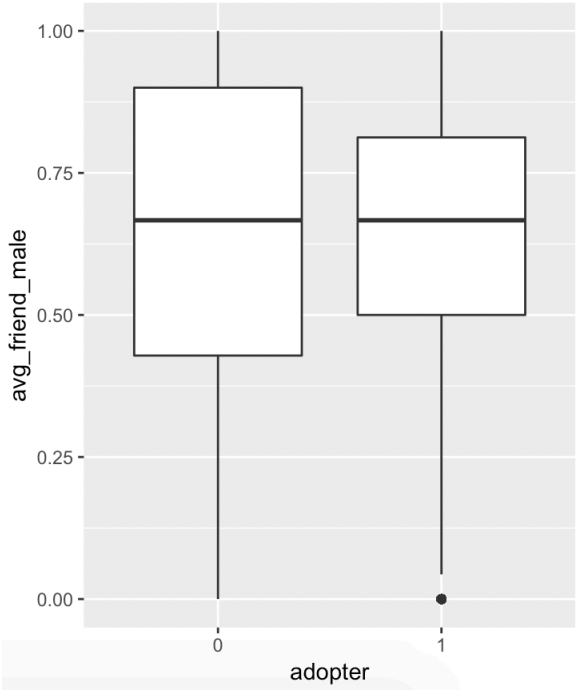
1. Demographics variables: age, avg_friend_age, male, avg_friend_male, country



Boxplot:

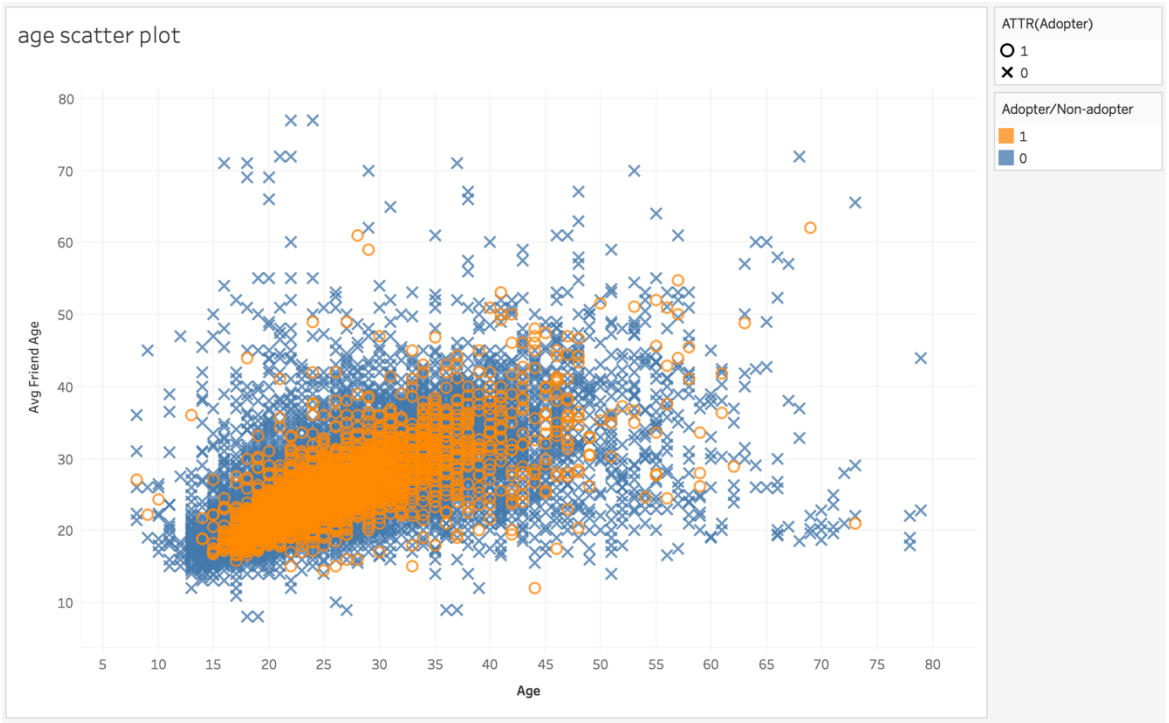






From boxplot, we can see the age, male distribution of adopters and non-adopters are similar, and both have many outliers.

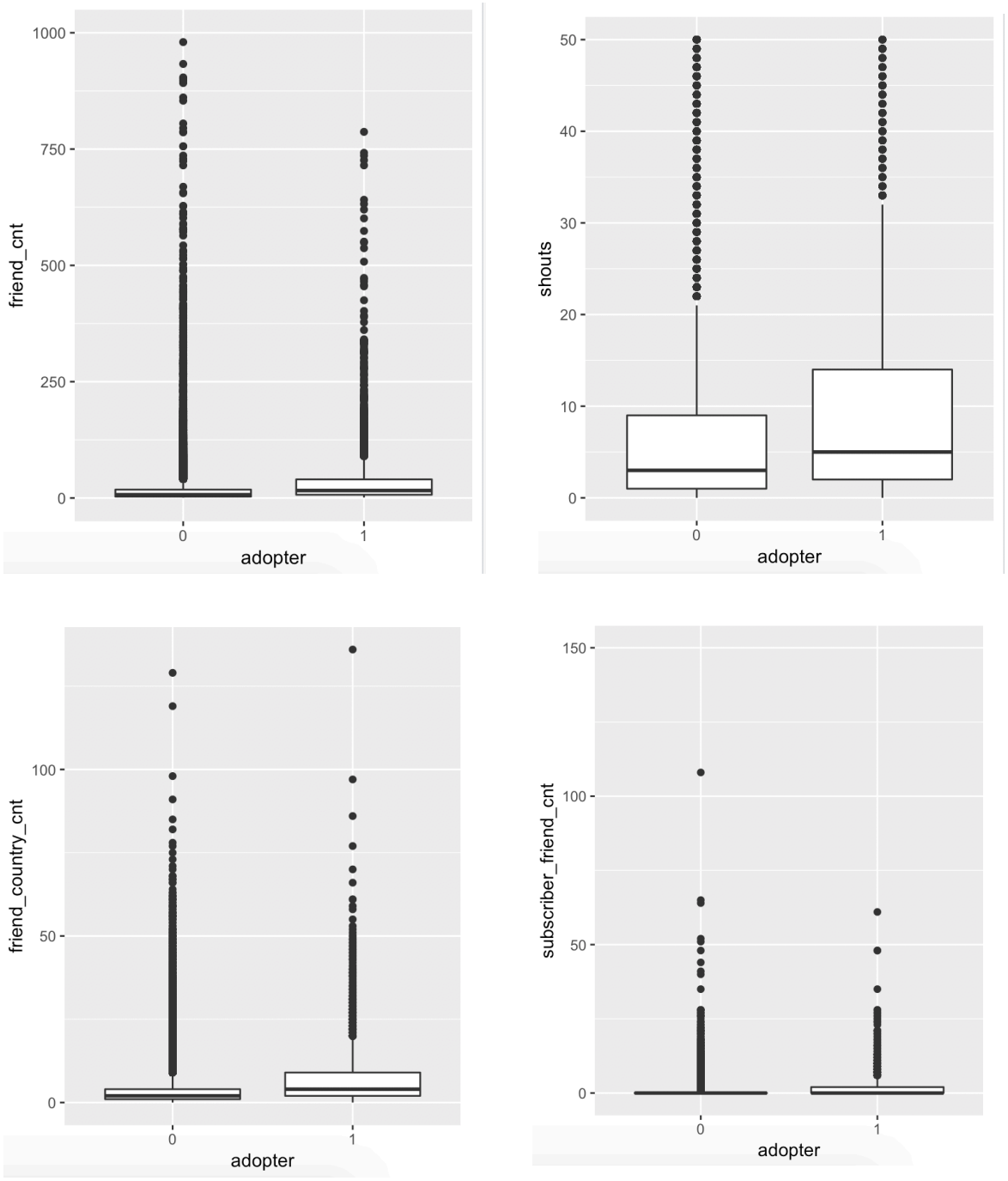
Scatter plot of age scale of adopter and non-adopter:



From the scatter plot we can see the distribution of users' age and their friends' age does not have significant difference between adopter and non-adopter.

2. Peer influence: difference between adopter and non-adopter on Friend_Cnt, Shouts, Friend_Country_Cnt and Subscriber_Friend_Cnt.

Boxplot:



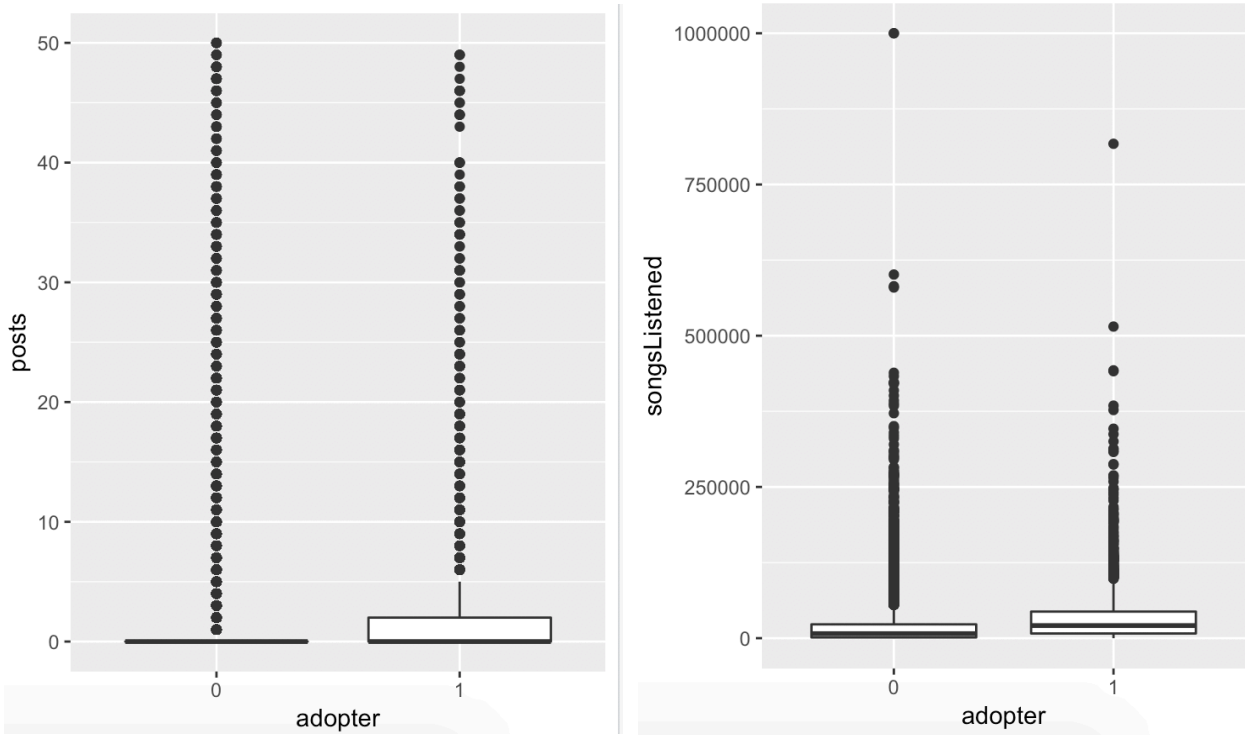
From boxplot, we can see these variables have many outliers.

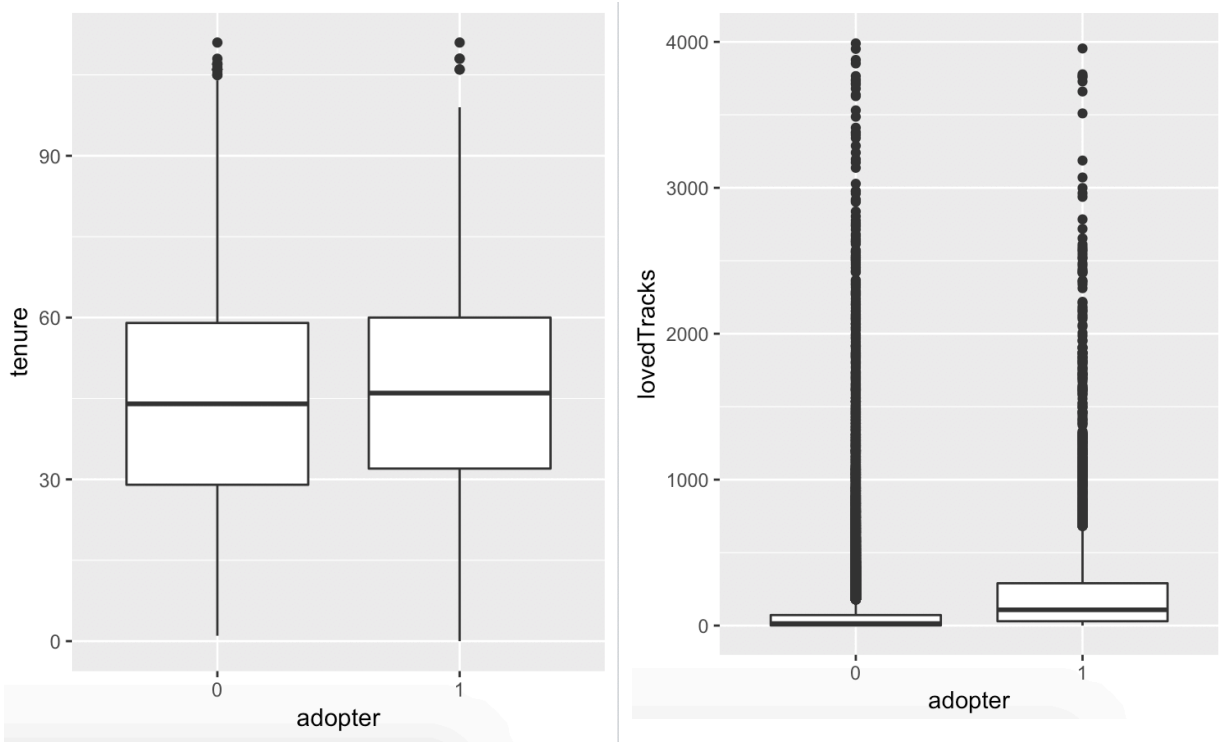
Mean value plot:

From peer influence plot, it tells us adopter's peer influence are obviously higher than non-adopter.

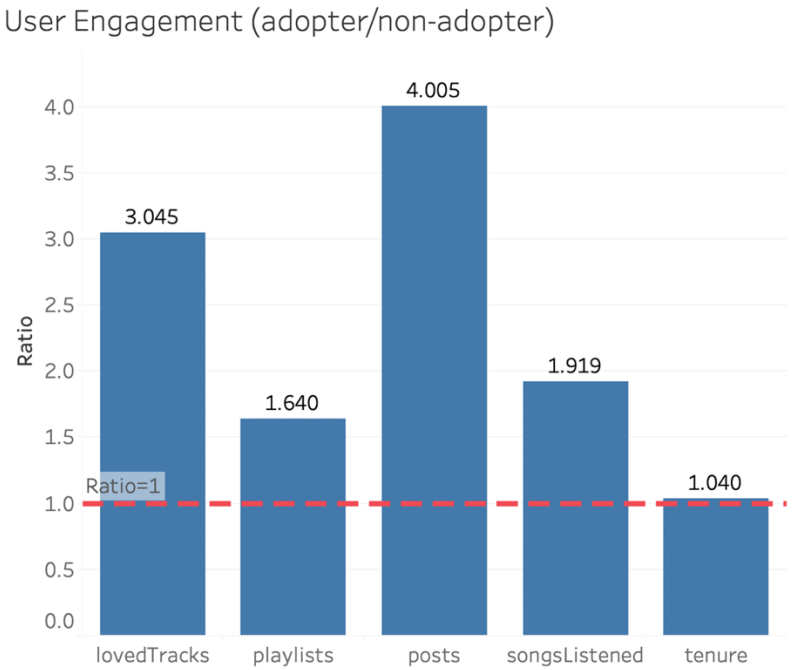
3. User engagement: Ratios (adopter/ non-adopter) on lovedTracks, Playlists, posts, songsListened and tenure.

Boxplot:





Ratio plot:



The ratios are all bigger than 1, which means adopter has more active engagement than non-adopter.

Propensity Score Matching

Use R to create regression model and use the model to calculate propensity score, then do ps matching to create an after-matching dataset and use the dataset to estimate treatment effect.

```
#create new variable called subscriber_friend
data<-mutate(data, subscriber_friend=ifelse(subscriber_friend_cnt==0, 0, 1))
# Pre-analysis using non-matched data
t.test(data$subscriber_friend,data$adopter)

cov <- c('age','male', 'friend_cnt',  'avg_friend_age', 'avg_friend_male',
         'friend_country_cnt', 'songsListened', 'lovedTracks',
         'posts', 'playlists',  'shouts', 'tenure', 'good_country')

lapply(cov, function(v) {
  t.test(data[, v] ~ data$subscriber_friend)
})

r<-glm(subscriber_friend~age+male+friend_cnt+avg_friend_age+avg_friend_male+
         friend_country_cnt+songsListened+lovedTracks+posts+playlists+
         shouts+tenure+good_country,
       family = binomial(), data = data)
summary(r)
#for variables 'male', 'playlists', 'shouts', 'good_country'are insignificant
r1<-glm(subscriber_friend~age+friend_cnt+avg_friend_age+avg_friend_male+
         friend_country_cnt+songsListened+lovedTracks+tenure+posts,
       family = binomial(), data = data)
summary(r1)

# Using this model, we can now calculate the propensity score for each student.
prs_df <- data.frame(pr_score = predict(r1, type = "response"),
                     subscriber_friend=r$model$subscriber_friend)
head(prs_df)
head(r$model)

 # eliminate missing values
 data_nomiss <- data %>%
   select(adopter,subscriber_friend_cnt, subscriber_friend,one_of(cov)) %>%
   na.omit()
 #matchit
 mod_match <- matchit(subscriber_friend ~ age+male+friend_cnt+avg_friend_age+avg_friend_male+
                        friend_country_cnt+songsListened+lovedTracks+posts+playlists+
                        shouts+tenure+good_country,
                      method = "nearest", data = data_nomiss)
 summary(mod_match)
 plot(mod_match)
 # To create a dataframe containing only the matched observations.
 dta_m <- match.data(mod_match)
 dim(dta_m)
 # Difference of means
 dta_m %>%
   group_by(subscriber_friend) %>%
   select(one_of(cov)) %>%
   summarise_all(funs(mean))

 lapply(cov, function(v) {
   t.test(dta_m[, v] ~ dta_m$subscriber_friend)
 })
 # Estimating treatment effects
 t.test(dta_m$subscriber_friend ~ dta_m$adopter)
 r_sf<-lm(adopter~subscriber_friend, data = dta_m)
 summary(r_sf)
```

After running the t-test and regression on estimating treatment effects, the results are as follow:

t-test result:

```
data:  dta_m$subscriber_friend by dta_m$adopter
t = -19.808, df = 3550.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2172408 -0.1781085
sample estimates:
mean in group 0 mean in group 1
      0.4738694       0.6715441
```

regression result:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.086837   0.003387   25.64   <2e-16 ***
subscriber_friend 0.090705   0.004790   18.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3357 on 19644 degrees of freedom
Multiple R-squared:  0.01793,   Adjusted R-squared:  0.01788
F-statistic: 358.7 on 1 and 19644 DF,  p-value: < 2.2e-16
```

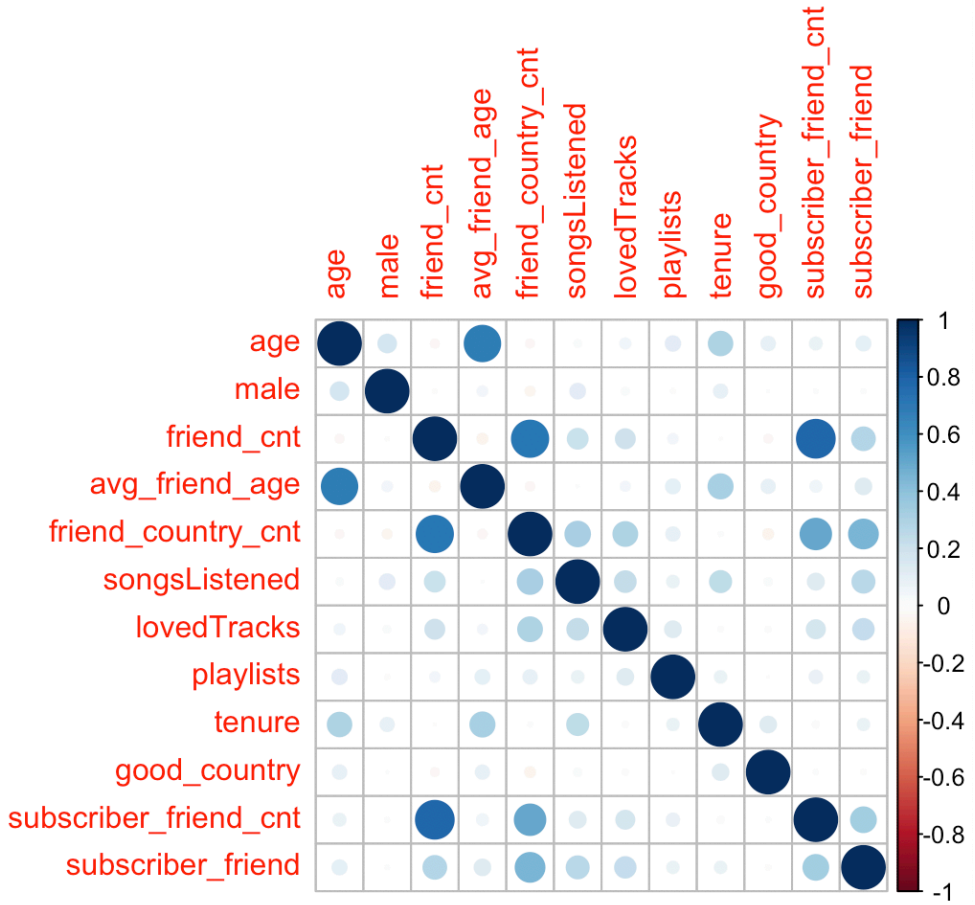
As results shows, p-value is smaller than 0.05, we can conclude there is a significant average treatment effect.

Regression Analysis:

After regression, I found posts,shouts and avg_friend_male are insignificant variables, so we take off these three variables from regression model. Then I did a correlation among significant variables.

```
#regression
summary(regression<- glm(adopter~age+male+friend_cnt+avg_friend_age+avg_friend_male+
                           friend_country_cnt+songsListened+lovedTracks+posts+playlists+
                           shouts+tenure+good_country+subscriber_friend_cnt+subscriber_friend,
                 family = binomial(), data = data))
# posts,shouts and avg_friend_male are insignificant.
summary(regression<- glm(adopter~age+male+friend_cnt+avg_friend_age+friend_country_cnt+
                         songsListened+lovedTracks+playlists+
                         tenure+good_country+subscriber_friend_cnt+subscriber_friend,
                         family = binomial(), data = data))

#Report the correlations among the all variables.
s_data<-subset(data,select=c(age,male,friend_cnt,avg_friend_age,
                             friend_country_cnt,songsListened,lovedTracks,playlists,
                             tenure,good_country,subscriber_friend_cnt,subscriber_friend))
M<-cor(s_data)
library('corrplot')
corrplot(M, method = "circle")
```



The result shows that subscriber_friend_cnt, friend_cnt and friend_country_cnt are highly correlated. I choose to only keep subscriber_friend_cnt. Age and avg_friend_age are highly correlated, so I choose to only keep age.

After simplifying variables, I did regression again to generate final regression model.

```
#subscriber_friend_cnt, friend_cnt and friend_country_cnt are highly correlated,
#so I only keep subscriber_friend_cnt.
#age and avg_friend_age are highly correlated, so I only keep age.
#after simplify variables, we do regression again.
regression<- glm(adopter~age+male+songsListened+lovedTracks+playlists+
                           tenure+good_country+subscriber_friend_cnt+subscriber_friend,
                         family = binomial(), data = data)
summary(regression)

odds_ratio<-data.frame(exp(regression$coefficients))
odds_ratio
```

From exp(coefficient), we can conclude odds ratio of significant variables.

significant variables	odds ratio
age	1.032024
male	1.458309
songsListened	1.000006
lovedTracks	1.000649
playlists	1.070197
tenure	0.995936
good_country	0.671268
subscriber_friend_cnt	1.016316
subscriber_friend	2.784875

Odds ratio of being an adopter for age is 1.032
The odds of being an adopter for males is 1.458 times that of females.
Odds ratio of being an adopter for songsListened is 1.000
Odds ratio of being an adopter for lovedTracks is 1.000
Odds ratio of being an adopter for playlists is 1.070
Odds ratio of being an adopter for tenure is 0.996
Odds ratio of being an adopter for good_country is 0.671
Odds ratio of being an adopter for subscriber_friend_cnt is 1.016
Odds ratio of being an adopter for subscriber_friend_cnt is 1.032

Takeaways

From results of statistics and regression analysis, we can conclude that these variables, age, gender, number of listened songs, loved tracks, playlists and subscriber friends, tenure, good_country and whether have subscriber friends are significant variables for being an adopter.

Among these significant variables, Age, songsListened, lovedTracks, playlists, male, subscriber_friend_cnt and subscriber_friend have positive effect on adopter. Good_country, tenure has negative effect on adopter.

Especially, for male, subscriber_friend and good_country, they have more obvious effect on being an adopter, so in order to effectively change free users into fee users, we should pay more attention for these three variables.

For user's gender, it shows male are more likely to be an adopter, which implies that it will be more effectively if HN make more strategies focused on male. For example, they can create some recommend music lists targeted men, attracting them being adopters so that they can enjoy these music in a better experience.

For subscriber_friend, it shows that if a person has a subscriber friend, he is more likely to be an adopter, which also means subscriber friends can influence his friends to be adopters. So, I suggest HN that they can give subscriber some bonus if they successfully invited his friends to be adopters. Also, we want to inform non-adopters that which of their friends have been adopters, and what are the benefits for joining them.

For good_country, it shows people from US, UK or Germany are less likely to be adopters, maybe for those big developed countries, they have too many similar options in the market, HN has more competitors in those countries. So I suggest HN could expand their marketing scale to other countries to gain more subscribers.