“Free-to-Fee” Strategy for High Note

Summary Statistics

The differences in the mean values of all covariates are all less than 0.05 which means statistically significant, and I come up with the following tentative conclusions from these comparisons. Users who are older and male tend to purchase premium service, so do their similar-age friends. Only good_country has a slightly lower mean that is negligible (0.07). Users who have more friends, and those friends are from different countries tend to become premium members. Social network influence is a powerful force in getting users from free to premium levels. When friends have premium membership, you tend to be fee-users too. More active users who listened to more songs, loved tracks, made posts, received shouts, had been registered for a long time, are more likely to access premium subscriptions.

Adopter Summary

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
age	3,527	25.98	6.84	8	21	24	29	73
male	3,527	0.73	0.44	0	0	1	1	1
friend_cnt	3,527	39.73	117.27	1	7	16	40	5,089
avg_friend_age	3,527	25.44	5.21	12.00	22.07	24.36	27.64	62.00
avg_friend_male	3,527	0.64	0.25	0.00	0.50	0.67	0.81	1.00
friend_country_cnt	3,527	7.19	8.86	0	2	4	9	136
subscriber_friend_cnt	3,527	1.64	5.85	0	0	0	2	287
songsListened	3,527	33,758.04	43,592.73	0	7,804.5	20,908	43,989.5	817,290
lovedTracks	3,527	264.34	491.43	0	30	108	292	10,220
posts	3,527	21.20	221.99	0	0	0	2	8,506
playlists	3,527	0.90	2.56	0	0	1	1	118
shouts	3,527	99.44	1,156.07	0	2	9	41	65,872
adopter	3,527	1.00	0.00	1	1	1	1	1
tenure	3,527	45.58	20.04	0	32	46	60	111
good_country	3,527	0.29	0.45	0	0	0	1	1

```
Non-Adopter Summary
====================================================================================
Statistic               N       Mean    St. Dev.  Min  Pctl(25) Median Pctl(75)  Max
------------------------------------------------------------------------------------
age                   40,300   23.95     6.37      8     20      23      26       79
male                  40,300   0.62      0.48      0     0       1       1        1
friend_cnt            40,300   18.49     57.48     1     3       7       18     4,957
avg_friend_age        40,300   24.01     5.10    8.00  20.67   23.00   26.06    77.00
avg_friend_male       40,300   0.62      0.32    0.00   0.43    0.67    0.90     1.00
friend_country_cnt    40,300   3.96      5.76      0     1       2       4       129
subscriber_friend_cnt 40,300   0.42      2.42      0     0       0       0       309
songsListened         40,300 17,589.44 28,416.02   0   1,252   7,440  22,892.8 1,000,000
lovedTracks           40,300   86.82    263.58     0     1       14      72     12,522
posts                 40,300   5.29     104.31     0     0       0       0      12,309
playlists             40,300   0.55      1.07      0     0       0       1        98
shouts                40,300   29.97    150.69     0     1       4       15     7,736
adopter               40,300   0.00      0.00      0     0       0       0        0
tenure                40,300   43.81     19.79     1     29      44      59      111
good_country          40,300   0.36      0.48      0     0       0       1        1
------------------------------------------------------------------------------------
```

```
    age  male friend_cnt avg_friend_age avg_friend_male friend_country_cnt subscriber_friend_cnt songsListened lovedTracks posts playlists shouts tenure good_country
  <dbl> <dbl>      <dbl>          <dbl>           <dbl>              <dbl>                 <dbl>         <dbl>       <dbl> <dbl>     <dbl>  <dbl>  <dbl>        <dbl>
1  23.9 0.622       18.5           24.0           0.617               3.96                 0.417        17589.        86.8  5.29     0.549   30.0   43.8        0.358
2  26.0 0.729       39.7           25.4           0.637               7.19                 1.64         33758.       264.  21.2      0.901   99.4   45.6        0.287
```

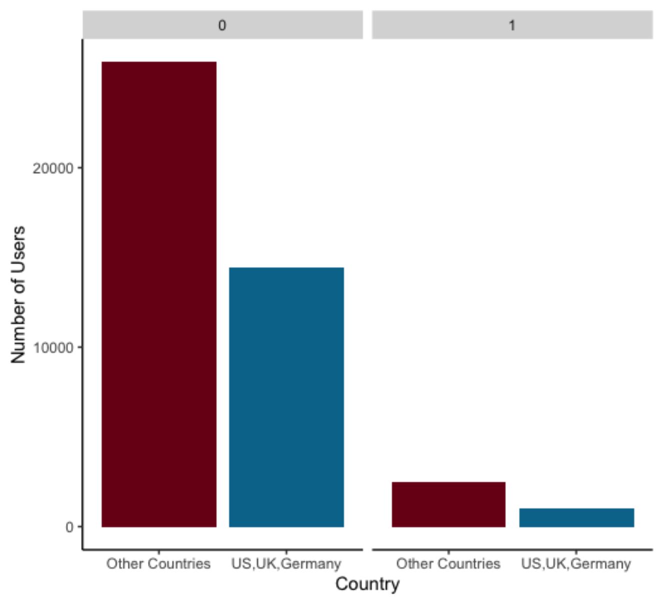
R Code

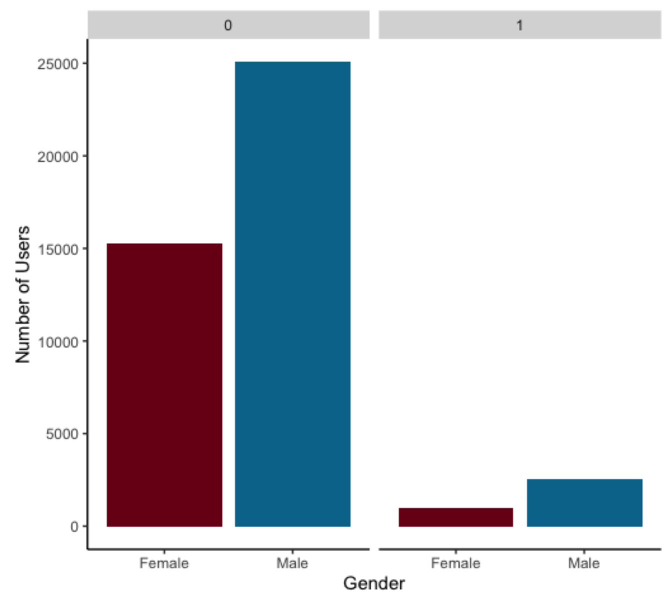
```
hn <- read.csv("~/Documents/UCI/CSA/HighNote Data.csv")
View(hn)
library(ggplot2)
library(dplyr)
library(stargazer)
library(tidyverse)
sum(is.na(hn))

# Summary Statistics
premium <- subset(hn,adopter==1)
free <- subset(hn,adopter==0)
stargazer(premium[,2:16],type="text",title="Adopter Summary",median=TRUE,iqr=TRUE,digits=2)
stargazer(free[,2:16],type="text",title="Non-Adopter
Summary",median=TRUE,iqr=TRUE,digits=2)

hn_cov <- c('age', 'male', 'friend_cnt', 'avg_friend_age',
'avg_friend_male','friend_country_cnt',
         'subscriber_friend_cnt', 'songsListened','lovedTracks', 'posts',
'playlists','shouts',
         'tenure','good_country')
hn %>% group_by(adopter) %>% summarise_all(funs(mean)) %>% select(one_of(hn_cov))
lapply(hn_cov, function(v) {t.test(hn[, v] ~ hn$adopter)})
```

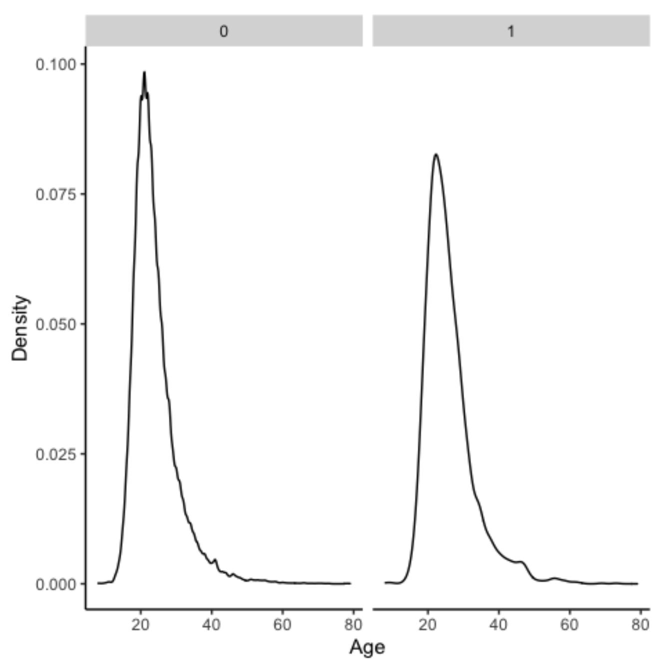
Data Visualization

From the following visualizations, I have the same conclusion as I mentioned earlier. Users who are male, older, have more friends from different countries, are more likely to be fee-users. There is a potential for social influence going from the users who are subscribed to the premium service. More premium member friends users have, they also tend to purchase premium service. Active users who listened to more songs, received more shouts, loved more tracks, made more posts, and created more playlists are more likely to be adopters.

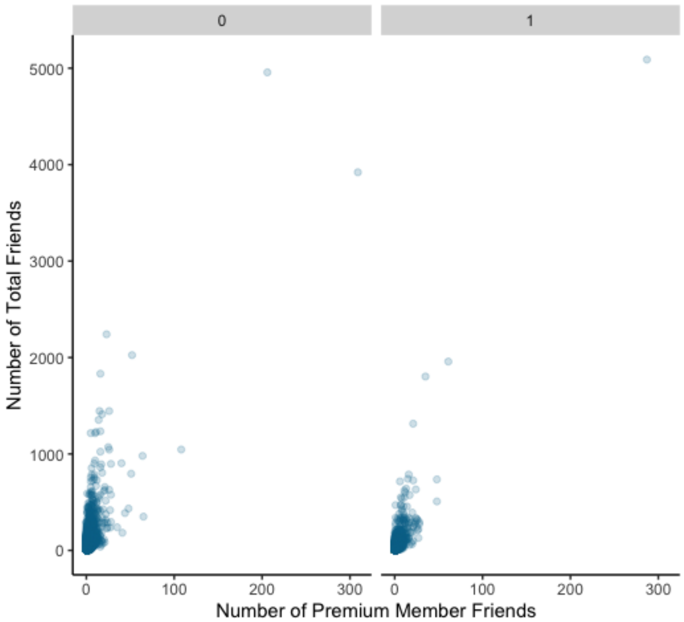
i) Demographics

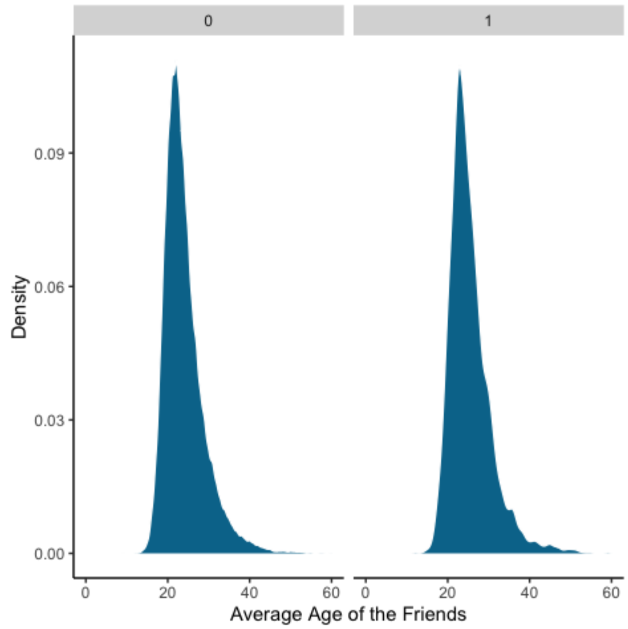


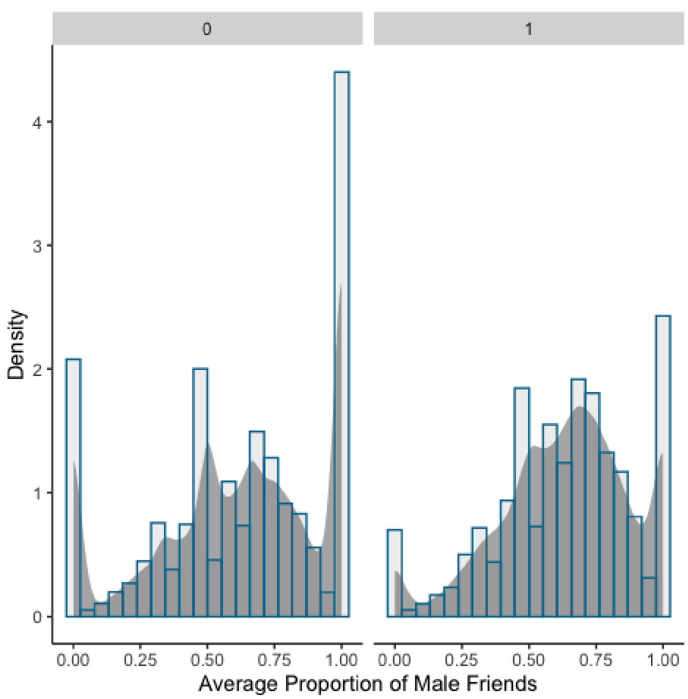


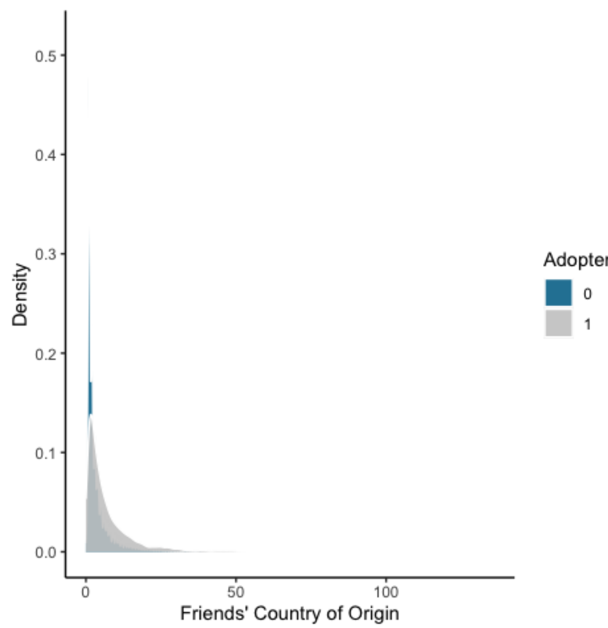


ii) Peer Influence

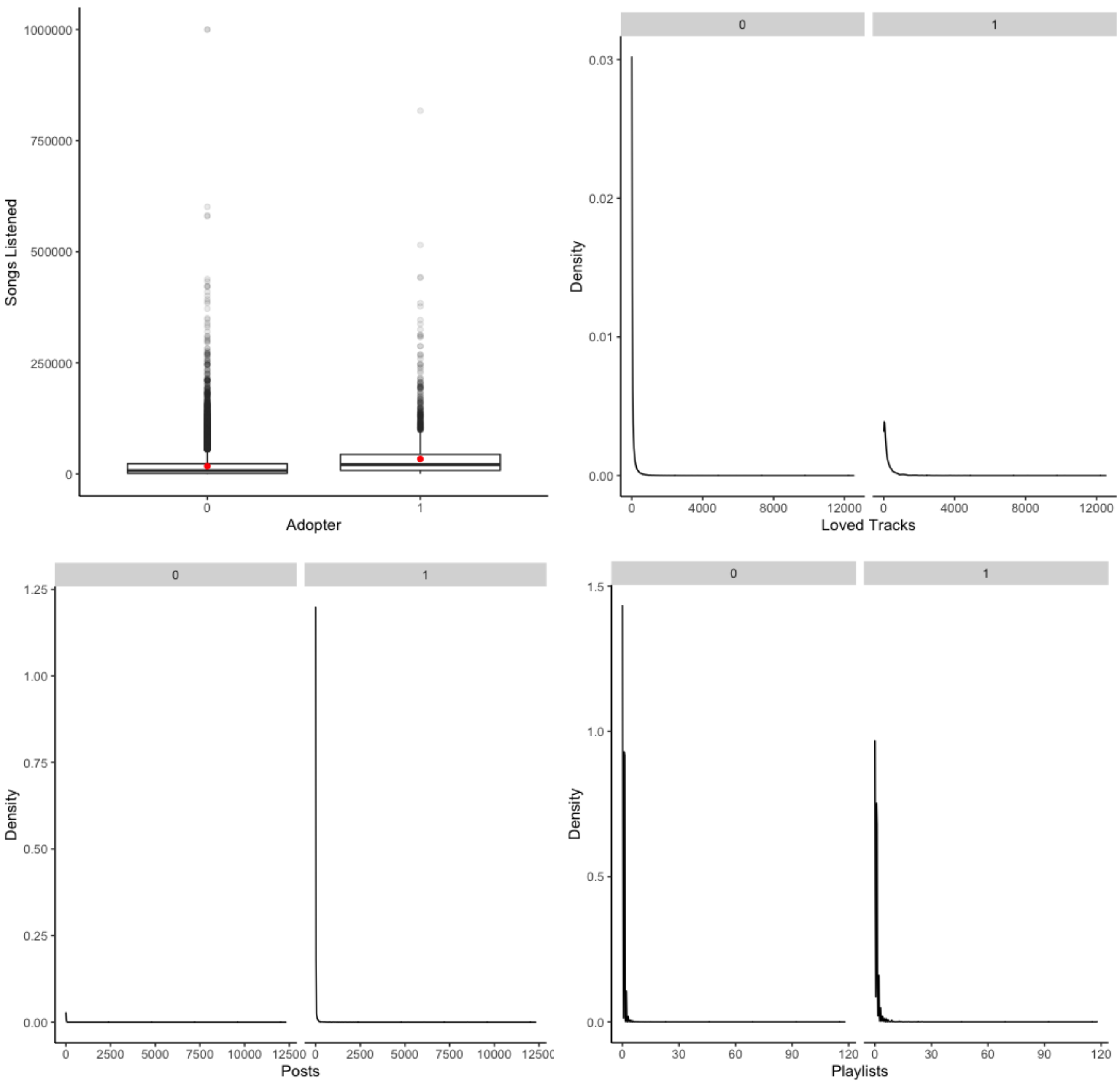


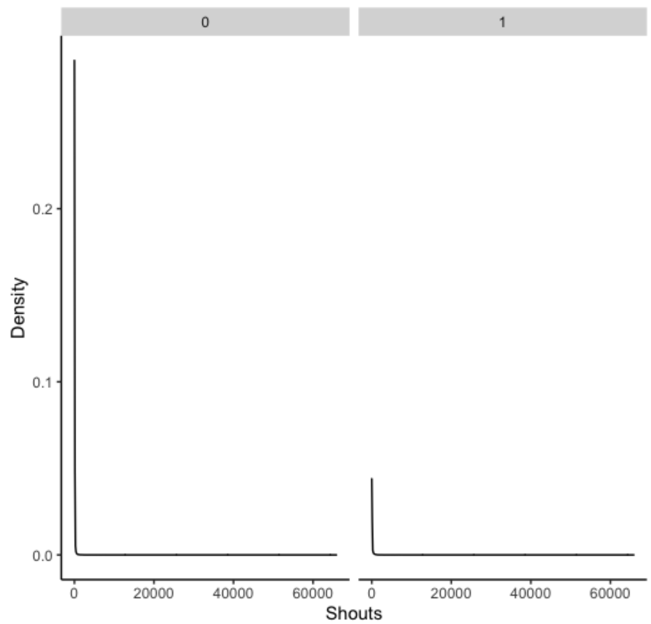


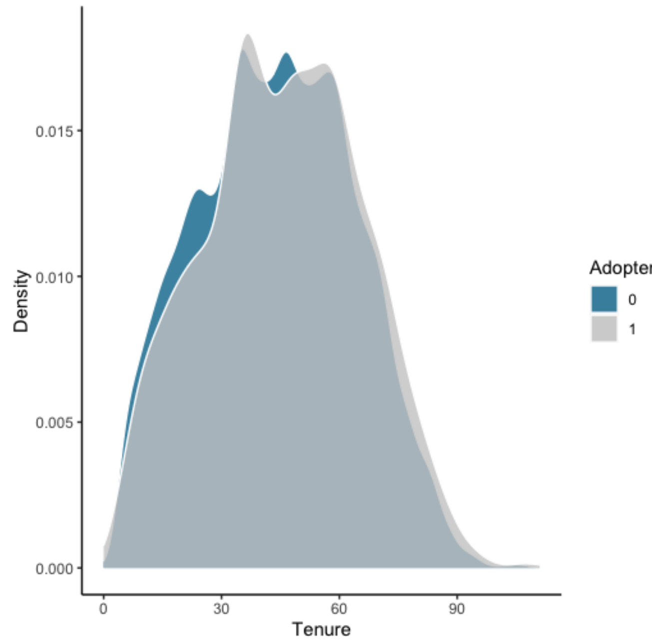




iii) User Engagement


Songs Listened
1000000
750000
500000
250000
0
Adopter
0
1
Density
0.03
0.02
0.01
0.00
0
4000
8000
12000
Loved Tracks
1.25
1.00
0.75
0.50
0.25
0.00
2500
5000
7500
10000
12500
Posts
1.5
1.0
0.5
0.0
30
60
90
120
Playlists





R Code

```
# i) Demographics
hn %>% ggplot(aes(x=factor(good_country),fill=factor(good_country)))+geom_bar()+
  scale_x_discrete(labels = c("Other Countries","US,UK,Germany"))+
  scale_fill_manual(values = c("#7a0019", "#00759a"))+ labs(x="Country",y="Number of
Users")+
  theme(legend.position="none",axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())+facet_wrap(~ adopter)

hn %>% ggplot(aes(x=factor(male),fill=factor(male)))+geom_bar()+
  scale_x_discrete(labels = c("Female","Male"))+
  scale_fill_manual(values = c("#7a0019", "#00759a"))+ labs(x="Gender",y="Number of
Users")+
  theme(legend.position="none",axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())+facet_wrap(~ adopter)

hn %>% ggplot(aes(x=age))+geom_density()+labs(x="Age",y="Density")+facet_wrap(~ adopter)+
  theme(axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())

# ii) Peer Influence
```

```
hn %>% ggplot(aes(subscriber_friend_cnt,friend_cnt))+geom_point(color="#00759a",alpha =
0.2)+
  labs(x="Number of Premium Member Friends",y="Number of Total Friends")+
  theme(axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())+facet_wrap(~ adopter)

hn %>% ggplot(aes(x = avg_friend_age))+geom_density(fill = "#00759a", color = "#ffffff")+
  labs(x = "Average Age of the Friends", y = "Density") +xlim(0,60)+
  theme(axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())+facet_wrap(~ adopter)

hn %>% ggplot(aes(x = avg_friend_male)) + geom_density(fill = "grey70", color = "#ffffff")
+
  geom_histogram(aes(y=..density..),bins = 20, color = "#00759a", alpha = 0.1) +
  labs(x = "Average Proportion of Male Friends", y = "Density")+facet_wrap(~ adopter)+
  theme(axis.line = element_line(color="black"),
      panel.background=element_blank(),
      panel.grid.minor=element_blank(),
      panel.grid.major.y=element_blank(),
      panel.grid.major.x=element_blank())

hn %>% ggplot(aes(x=friend_country_cnt,fill=factor(adopter)))+
  geom_density(color = "white", alpha = 0.9)+scale_fill_manual(values=c("#00759a",
"grey80"))+
  labs(x="Friends' Country of Origin",y="Density",fill="Adopter")+
  theme(axis.line = element_line(color="black"),panel.background=element_blank(),
        panel.grid.minor=element_blank(),panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())

# iii) User Engagement
hn %>% ggplot(aes(x =factor(adopter), y=songsListened))+geom_boxplot(alpha=0.1)+
  labs(x="Adopter",y = "Songs Listened") + stat_summary(fun.y=mean, geom="point",
color="red", fill="red")+
  theme(axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())

hn %>% ggplot(aes(x=lovedTracks))+geom_density()+labs(x="Loved
Tracks",y="Density")+facet_wrap(~ adopter)+
  theme(axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())

hn %>% ggplot(aes(x=posts))+geom_density()+labs(x="Posts",y="Density")+facet_wrap(~
adopter)+
```

```
  theme(axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())

hn %>% ggplot(aes(x=playlists))+geom_density()+labs(x="Playlists",y="Density")+facet_wrap(~
adopter)+
  theme(axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())

hn %>% ggplot(aes(x=shouts))+geom_density()+labs(x="Shouts",y="Density")+facet_wrap(~
adopter)+
  theme(axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())

hn %>% ggplot(aes(x=tenure,fill=factor(adopter)))+
  geom_density(color = "white", alpha = 0.8)+scale_fill_manual(values=c("#00759a",
"grey80"))+
  labs(x="Tenure",y="Density",fill="Adopter")+
  theme(axis.line = element_line(color="black"),panel.background=element_blank(),
        panel.grid.minor=element_blank(),panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())
```

Propensity Score Matching

The treatment group will be users that have one or more subscriber friends (subscriber_friend_cnt >= 1), while the control group will include users with zero subscriber friends.

1. Pre-analysis Using Non-Matched Data

We can see that the difference-in-means is statistically significant as adopter is the outcome variable (group 0: non-adopter, group 1: adopter).

```
	Welch Two Sample t-test

data:  hn$adopter by hn$ynsf
t = -30.961, df = 11815, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.1330281 -0.1171869
sample estimates:
mean in group 0 mean in group 1
     0.05243501      0.17754250
```

Next, I calculated the mean for each covariates and used t-test to evaluate those means that are statistically different. The output shows that only “male” is not statistically distinguishable (p-value = 0.1784).

```
# A tibble: 2 × 13
    age  male friend_cnt avg_friend_age avg_friend_male friend_country_cnt songsListened lovedTracks posts playlists shouts tenure good_country
  <dbl> <dbl>      <dbl>          <dbl>           <dbl>              <dbl>         <dbl>       <dbl> <dbl>     <dbl>  <dbl>  <dbl>        <dbl>
1  23.7 0.629       10.4           23.8           0.613               2.73        14602.        65.2  2.54     0.529   16.4   43.2        0.355
2  25.4 0.636       54.0           25.4           0.636               9.39        33736.       225.   20.5     0.744  102.    46.5        0.343
```

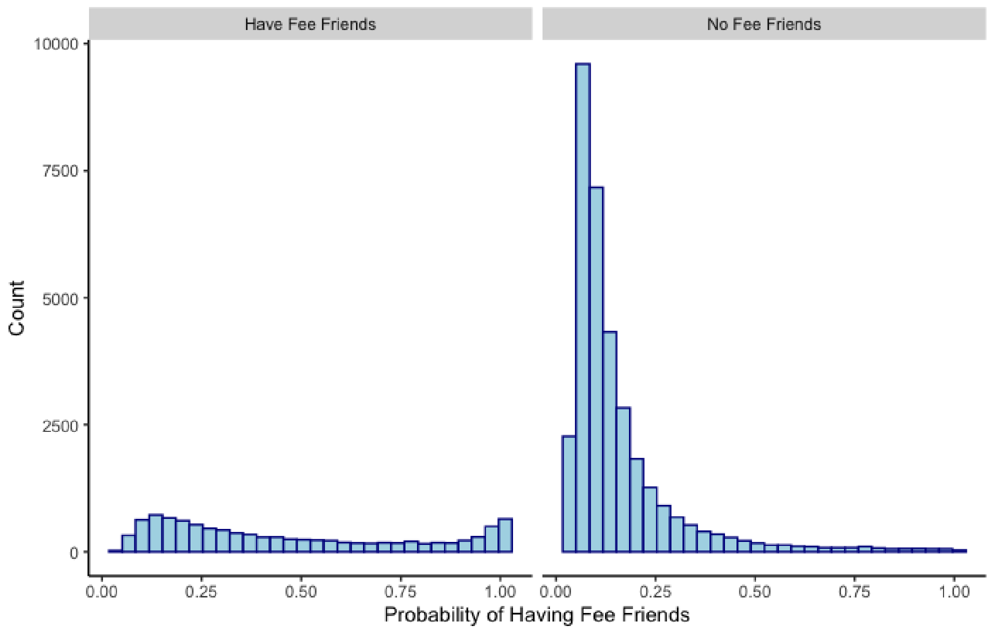
2. Propensity Score Estimation

I excluded “male” when estimating the propensity score by running a logit model where the outcome variable is a binary variable indicating treatment status.

```
Call:
glm(formula = ynsf ~ age + friend_cnt + avg_friend_age + avg_friend_male +
    friend_country_cnt + songsListened_1k + lovedTracks + posts +
    playlists + shouts + tenure + good_country, family = binomial(),
    data = hn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4154  -0.5668  -0.4221  -0.3009   2.5520

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.124e+00  7.566e-02 -67.720  < 2e-16 ***
age                 2.043e-02  2.757e-03   7.409 1.27e-13 ***
friend_cnt          3.131e-02  1.033e-03  30.295  < 2e-16 ***
avg_friend_age      7.904e-02  3.460e-03  22.843  < 2e-16 ***
avg_friend_male     2.528e-01  5.027e-02   5.030 4.92e-07 ***
friend_country_cnt  1.105e-01  4.751e-03  23.266  < 2e-16 ***
songsListened_1k    7.012e-03  5.107e-04  13.731  < 2e-16 ***
lovedTracks         6.685e-04  5.644e-05  11.845  < 2e-16 ***
posts               5.753e-04  2.686e-04   2.142   0.0322 *
playlists           5.249e-03  1.191e-02   0.441   0.6593
shouts             -5.027e-05  3.678e-05  -1.367   0.1717
tenure             -2.534e-03  7.766e-04  -3.262   0.0011 **
good_country        3.088e-02  2.921e-02   1.057   0.2903
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46640  on 43826  degrees of freedom
Residual deviance: 34173  on 43814  degrees of freedom
AIC: 34199

Number of Fisher Scoring iterations: 8
```

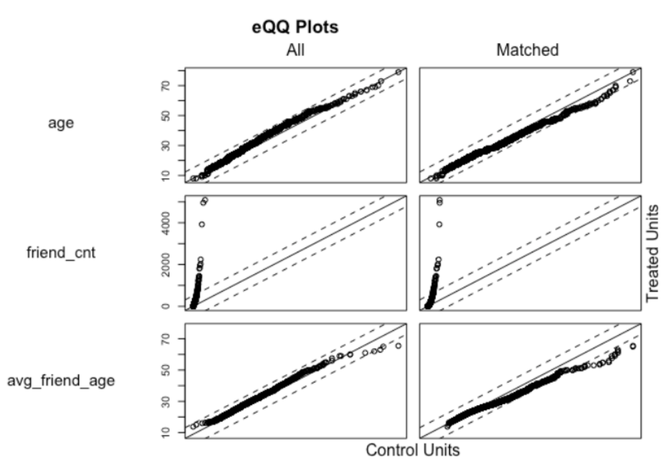
Using this model, I calculated the propensity for each user. It is the user’s predicted probability of being treated, given the estimates from the logit model.

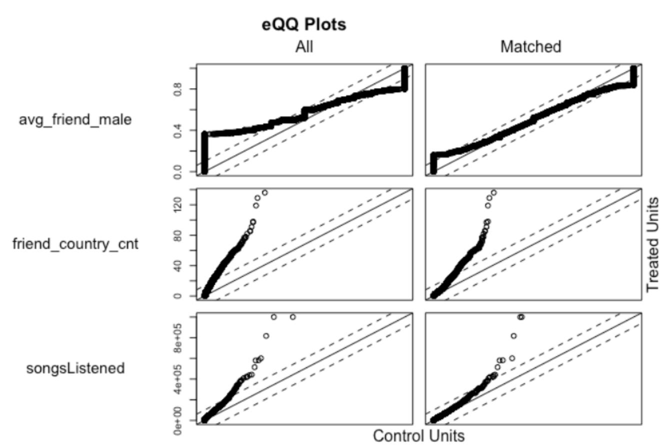
```
    pr_score ynsf
1 0.08810050    0
2 0.14832644    0
3 0.08121395    0
4 0.24291404    1
5 0.70270131    0
6 0.22199154    0
```

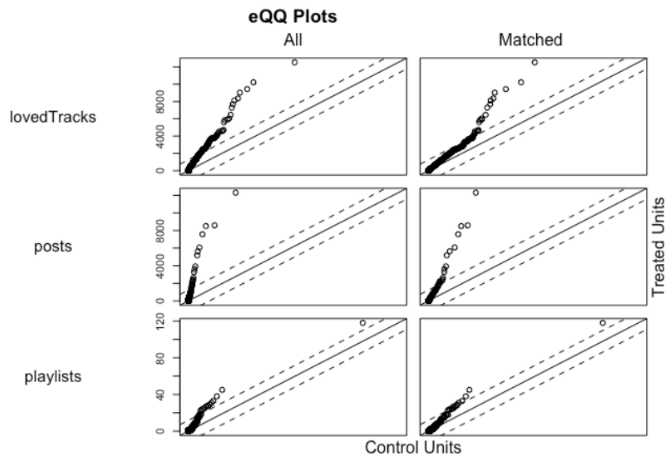


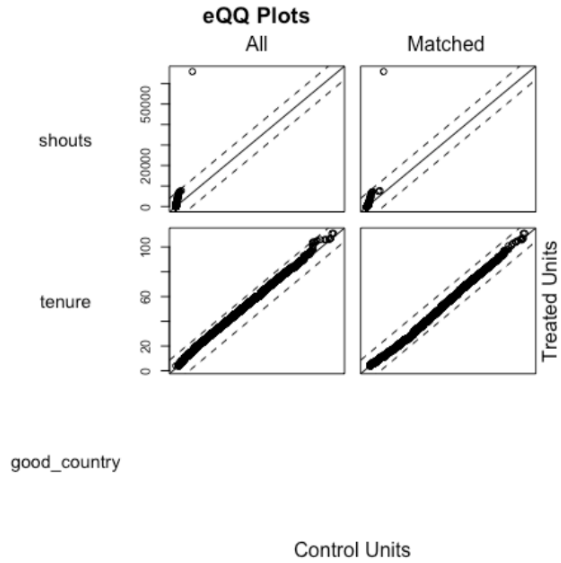
3. Matched Sampling

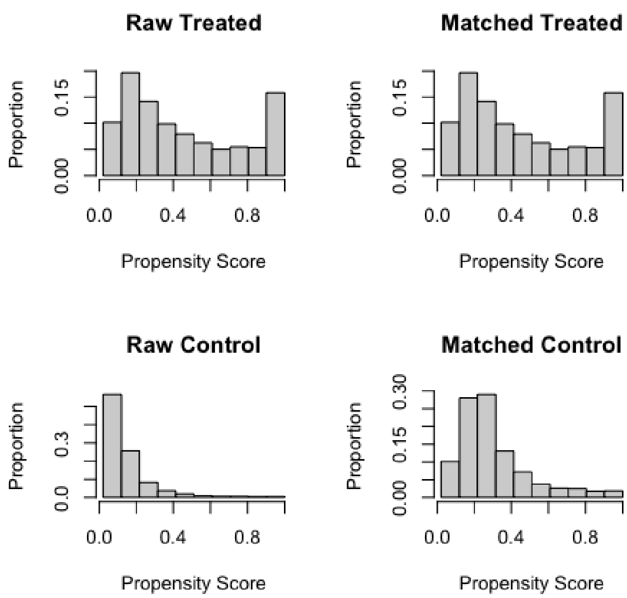
I found a pair of observations that have very similar propensity scores, but that differ in their treatment status. Below are some visualizations of how successful the matching works.











Then, I created a dataframe containing only the matched observations. The final dataset has 19646 observations, and 9823 pairs of control and treated observations are matched. It also contains a variable called distance, which is the propensity score.

```
Sample Sizes:
          Control Treated
All         34004    9823
Matched      9823    9823
Unmatched   24181       0
Discarded       0       0
```

4. Covariate Balance in the Matched Sample

I calculated the mean for each covariates and used t-test to estimate the treatment effect with the matched sample. T-value was changed from -30.961 before matching to -18.938. Having subscriber friends has a higher probability of being adopter than those who don't have subscriber friends.

```
	Welch Two Sample t-test

data:  dta_m$adopter by dta_m$ynsf
t = -18.938, df = 18060, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -0.10009352 -0.08131745
sample estimates:
mean in group 0 mean in group 1
     0.08683702      0.17754250
```

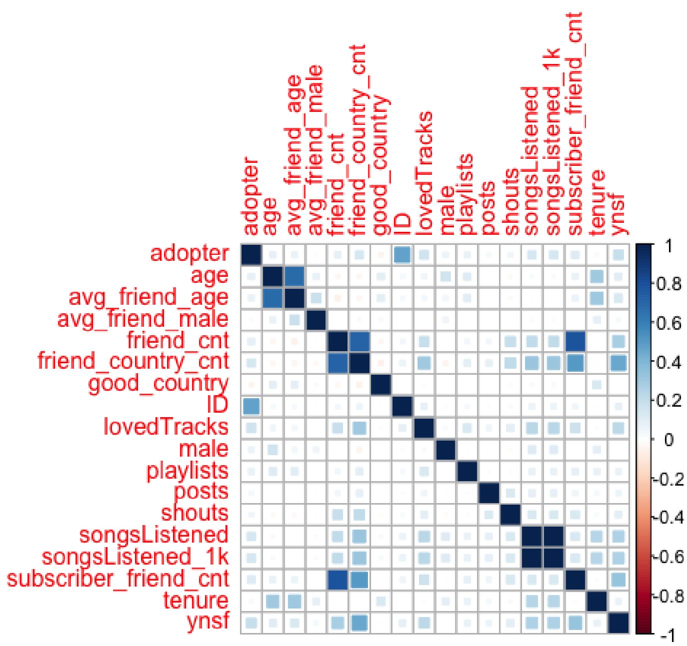
R Code

```
hn$ynsf = ifelse(hn$subscriber_friend_cnt >= 1, 1, 0)
t.test(hn$adopter~hn$ynsf)
hn_cov2 <-c('age', 'male', 'friend_cnt', 'avg_friend_age', 'avg_friend_male',
'friend_country_cnt',
             'songsListened', 'lovedTracks', 'posts', 'playlists','shouts', 'tenure',
'good_country')
hn %>%group_by(ynsf)%>%summarise_all(funs(mean))%>%select(one_of(hn_cov2))
lapply(hn_cov2, function(v) {t.test(hn[,v] ~ hn[,'ynsf'])})

hn <- hn %>% mutate(songsListened_1k = songsListened/1000)
m_ps <- glm(ynsf ~ age + friend_cnt + avg_friend_age + avg_friend_male +
friend_country_cnt+
              songsListened_1k + lovedTracks + posts + playlists+ shouts + tenure +
              good_country, family = binomial(), data = hn)
summary(m_ps)

prs_df <-data.frame(pr_score = predict(m_ps, type = "response"),ynsf = m_ps$model$ynsf)
head(prs_df)
head(m_ps$model)

prs_df %>%mutate(ynsf = recode(ynsf,"0" = "No Fee Friends","1" = "Have Fee Friends"))%>%
  ggplot(aes(x = pr_score)) +geom_histogram(color="darkblue",fill = "lightblue") +
  facet_wrap(~ynsf) +labs(x="Probability of Having Fee Friends",y="Count")+
  theme(axis.line = element_line(color="black"),
        panel.background=element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major.y=element_blank(),
        panel.grid.major.x=element_blank())

library(MatchIt)
hn_nomiss <- hn %>%select(adopter, ynsf, one_of(hn_cov2)) %>%na.omit()
mod_match <- matchit(ynsf ~ age + friend_cnt + avg_friend_age + avg_friend_male+
                       friend_country_cnt+ songsListened + lovedTracks + posts +
                       playlists+ shouts + tenure + good_country,method="nearest",
                     data = hn_nomiss)
summary(mod_match)
```

```
plot(mod_match)
plot(mod_match,type="hist")

dta_m <- match.data(mod_match)
dim(dta_m)

dta_m %>%group_by(ynsf) %>%select(one_of(hn_cov2)) %>%summarise_all(funs(mean))
lapply(hn_cov2, function(v) {t.test(dta_m[, v] ~ dta_m$ynsf)})

t.test(dta_m$adopter~dta_m$ynsf)

treat1 <- glm(adopter ~ ynsf, family = binomial(), data = dta_m)
summary(treat1)
treat2 <- glm(adopter ~ ynsf + age + friend_cnt + avg_friend_age +
                    avg_friend_male +friend_country_cnt+ lovedTracks +
                    posts + playlists+ shouts + tenure + good_country+
                    I(songsListened / 1000), family = binomial(), data = dta_m)
summary(treat2)
```

Regression Analysis

I used a logistic regression approach to test which variables (including subscriber friends) are significant for explaining the likelihood of becoming an adopter. Here is the graph of the correlation matrix and the summary table of putting all the variables into the model. I discovered that some independent variables are relatively highly correlated, therefore, I would exclude those in the next regression model, which are age & avg_friend_age, friend_cnt & friend_country_cnt, friend_cnt & subscriber_friend_cnt, subscriber_friend_cnt & friend_country_cnt. In addition, avg_friend_male, posts, and shouts are not statistically significant, which also needs to be excluded.



```
Call:
glm(formula = adopter ~ age + male + friend_cnt + avg_friend_age +
    avg_friend_male + friend_country_cnt + subscriber_friend_cnt +
    lovedTracks + posts + playlists + songsListened_1k + shouts +
    tenure + good_country, family = binomial(), data = hn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3526  -0.4114  -0.3500  -0.2913   2.7018

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.179e+00  9.571e-02 -43.665  < 2e-16 ***
age                    1.962e-02  3.478e-03   5.641 1.69e-08 ***
male                   4.133e-01  4.169e-02   9.913  < 2e-16 ***
friend_cnt            -4.312e-03  4.920e-04  -8.765  < 2e-16 ***
avg_friend_age         2.954e-02  4.484e-03   6.588 4.45e-11 ***
avg_friend_male        1.162e-01  6.346e-02   1.831   0.0671 .
friend_country_cnt     4.326e-02  3.616e-03  11.962  < 2e-16 ***
subscriber_friend_cnt  9.132e-02  1.073e-02   8.512  < 2e-16 ***
lovedTracks            6.950e-04  4.933e-05  14.088  < 2e-16 ***
posts                  8.492e-05  9.580e-05   0.886   0.3754
playlists              5.920e-02  1.333e-02   4.441 8.97e-06 ***
songsListened_1k       7.626e-03  5.192e-04  14.687  < 2e-16 ***
shouts                 1.108e-04  8.428e-05   1.314   0.1887
tenure                -4.476e-03  1.022e-03  -4.380 1.19e-05 ***
good_country          -4.152e-01  4.078e-02 -10.181  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24537  on 43826  degrees of freedom
Residual deviance: 22613  on 43812  degrees of freedom
AIC: 22643

Number of Fisher Scoring iterations: 5
```

Looking at correlations only among pairs of predictors, however, is limiting. A variance inflation factor (VIF) quantifies how much the variance is inflated. If some of the predictors are correlated with the predictor, then the variance is inflated. The general rule of thumb is that VIFs exceeding 4 warrant further investigation. The result shows that friend_cnt has a multicollinearity problem with the predictor variables, therefore I would exclude it in the next optimized model.

$$VIF_k = \frac{1}{1 - R_k^2}$$

```
                 age                 male           friend_cnt        avg_friend_age
            2.028083             1.061966             4.295009              2.061113
      avg_friend_male   friend_country_cnt subscriber_friend_cnt          lovedTracks
            1.042020             2.621221             3.007514              1.150339
               posts            playlists      songsListened_1k               shouts
            1.088116             1.044297             1.280630              1.337860
              tenure         good_country
            1.213634             1.029508
```

```
                  age                  male            friend_cnt        avg_friend_age       avg_friend_male
                FALSE                 FALSE                  TRUE                 FALSE                 FALSE
   friend_country_cnt subscriber_friend_cnt           lovedTracks                 posts             playlists
                FALSE                 FALSE                 FALSE                 FALSE                 FALSE
     songsListened_1k                shouts                tenure          good_country
                FALSE                 FALSE                 FALSE                 FALSE
```

```
       rstudent unadjusted p-value Bonferroni p
32663 -5.837848         5.2879e-09   0.00023175
```

After excluding some variables as mentioned above, the new model with the key variables no longer has a multicollinearity problem, but some outliers exit.

```
Call:
glm(formula = adopter ~ age + male + subscriber_friend_cnt + 
    lovedTracks + songsListened_1k + playlists + tenure + good_country, 
    family = binomial(), data = hn)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-7.5127  -0.4112  -0.3562  -0.3032   2.6629  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -3.668e+00  7.209e-02 -50.880  < 2e-16 ***
age                    3.489e-02  2.555e-03  13.656  < 2e-16 ***
male                   3.459e-01  4.110e-02   8.417  < 2e-16 ***
subscriber_friend_cnt  9.817e-02  8.364e-03  11.737  < 2e-16 ***
lovedTracks            7.710e-04  4.932e-05  15.633  < 2e-16 ***
songsListened_1k       8.293e-03  5.017e-04  16.532  < 2e-16 ***
playlists              7.003e-02  1.367e-02   5.123    3e-07 ***
tenure                -3.452e-03  1.003e-03  -3.443 0.000576 ***
good_country          -4.229e-01  4.060e-02 -10.417  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24537  on 43826  degrees of freedom
Residual deviance: 22796  on 43818  degrees of freedom
AIC: 22814

Number of Fisher Scoring iterations: 7
```

```
                  age                  male subscriber_friend_cnt           lovedTracks      songsListened_1k
             1.138048              1.040311              1.130367              1.123615              1.204860
            playlists                tenure          good_country
             1.041063              1.180435              1.024255
```

```
          age                  male subscriber_friend_cnt           lovedTracks      songsListened_1k 
        FALSE                 FALSE                 FALSE                 FALSE                 FALSE 
    playlists                tenure          good_country 
        FALSE                 FALSE                 FALSE 
```

```
      rstudent unadjusted p-value Bonferroni p
32663 -7.942491         1.9816e-15   8.6848e-11
21293 -6.257061         3.9230e-10   1.7193e-05
10623 -4.999455         5.7492e-07   2.5197e-02
```

I deleted three outliers, and all variables are statistically significant. The AIC value decreased from 22814 to 22670. It compares the fit of models and the model with the lowest AIC value is best.

```
Call:
glm(formula = adopter ~ age + male + subscriber_friend_cnt + 
    lovedTracks + songsListened_1k + playlists + tenure + good_country, 
    family = binomial(), data = new_hn)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.9047  -0.4090  -0.3543  -0.3005   2.6713  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -3.671e+00  7.246e-02 -50.665  < 2e-16 ***
age                    3.377e-02  2.572e-03  13.131  < 2e-16 ***
male                   3.622e-01  4.130e-02   8.771  < 2e-16 ***
subscriber_friend_cnt  1.435e-01  8.814e-03  16.285  < 2e-16 ***
lovedTracks            7.228e-04  4.948e-05  14.608  < 2e-16 ***
songsListened_1k       7.789e-03  5.042e-04  15.449  < 2e-16 ***
playlists              6.625e-02  1.365e-02   4.855  1.2e-06 ***
tenure                -3.307e-03  1.005e-03  -3.289    0.001 ** 
good_country          -4.228e-01  4.076e-02 -10.372  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24536  on 43823  degrees of freedom
Residual deviance: 22652  on 43815  degrees of freedom
AIC: 22670

Number of Fisher Scoring iterations: 5
```

Here is the regression coefficient:

```
          (Intercept)                   age                  male subscriber_friend_cnt           lovedTracks 
         -3.671304231           0.033769983           0.362194493           0.143534431           0.000722798 
     songsListened_1k             playlists                tenure          good_country 
          0.007788749           0.066251411          -0.003306948          -0.422794625 
```

In a logistic regression the response being modeled is the log(odds) that Y = 1. The regression coefficient gives the change in log(odds) in the response for a unit change in the predictor variable, because log(odds) are difficult to interpret, I can exponentiate them. 1 is the cutoff point. For example, the age's odds ratio is 1.0343. If you are older than one year, the odds ratio increases. For male, the odds of being a fee-user increases by 43%, and so on. Tenure and good_country are the only two variables that decrease the odds of conversion rate. Changing good_country by 1, the odds ratio goes down by 35%.

```
      (Intercept)                 age                male subscriber_friend_cnt           lovedTracks
       0.02544326          1.03434666          1.43647830            1.15434655            1.00072306
  songsListened_1k           playlists              tenure          good_country
       1.00781916          1.06849532          0.99669851            0.65521318
```

R Code

```
cor <- cor(hn)
library(corrplot)
corrplot(cor,method="square",order="alphabet")
library(car)
fit1 <- glm(adopter ~ age + male + friend_cnt + avg_friend_age + avg_friend_male +
friend_country_cnt
                + subscriber_friend_cnt + lovedTracks + posts + playlists +
songsListened_1k
                + shouts + tenure + good_country, family = binomial(), data = hn)
summary(fit1)
car::vif(fit1)
vif(fit1)> 4
outlierTest(fit1)

fit2 <- glm(adopter ~ age + male+subscriber_friend_cnt+lovedTracks+
              songsListened_1k+playlists+tenure+good_country, family = binomial(), data =
hn)
summary(fit2)
vif(fit2)
vif(fit2)>4
outlierTest(fit2)

new_hn <- hn[c(-32663,-21293,-10623),]
fit3 <- glm(adopter ~ age + male+subscriber_friend_cnt+lovedTracks+
              songsListened_1k+playlists+tenure+good_country, family = binomial(), data =
new_hn)
summary(fit3)
coef(fit3)
exp(coef(fit3))
```

Takeaways

From the above data analyst, here are some potential insights to inform a "free-to-fee" strategy for High Note:

- Target users' age: 20's

- High user engagement: more accurately estimating the likelihood that users will listen based on interactions, preferences on services to keep high engagement (note: posts and shouts are not important)
- Peer influence: making HIgh Note more interactive and community-focused, and provide recommendations based on their social experience, listening history, and trends (note: subscriber friend does matter)
- Location: reaching out to more users from different countries (note: other than the US, UK, Germany)