Author: MJ

HighNote User Information data analysis:
Paid subscription conversion strategy

1. Data Overview & Exploratory Analysis & Summary statistics:

   Below I generate descriptive statistics for the key variables in the data set. I mainly observed the differences in the mean values of the variables, comparing the adopter and non-adapter subsamples.

```
 Descriptive statistics by group
group: 0
                    vars     n     mean       sd  median  trimmed      mad min     max   range  skew kurtosis     se
age                    1 40300    23.95     6.37   23.00    23.09     4.45   8      79      71  1.97     6.80   0.03
male                   2 40300     0.62     0.48    1.00     0.65     0.00   0       1       1 -0.50    -1.75   0.00
friend_cnt             3 40300    18.49    57.48    7.00    10.28     7.41   1    4957    4956 32.67  2087.42   0.29
avg_friend_age         4 40300    24.01     5.10   23.00    23.40     3.95   8      77      69  1.84     7.15   0.03
avg_friend_male        5 40300     0.62     0.32    0.67     0.65     0.35   0       1       1 -0.52    -0.72   0.00
friend_country_cnt     6 40300     3.96     5.76    2.00     2.66     1.48   0     129     129  4.74    38.29   0.03
subscriber_friend_cnt  7 40300     0.42     2.42    0.00     0.13     0.00   0     309     309 72.19  8024.62   0.01
songsListened          8 40300 17589.44 28416.02 7440.00 11817.64 10576.87   0 1000000 1000000  6.05   105.85 141.55
lovedTracks            9 40300    86.82   263.58   14.00    36.35    20.76   0   12522   12522 13.12   335.93   1.31
posts                 10 40300     5.29   104.31    0.00     0.23     0.00   0   12309   12309 73.92  7005.34   0.52
playlists             11 40300     0.55     1.07    0.00     0.45     0.00   0      98      98 28.21  1945.28   0.01
shouts                12 40300    29.97   150.69    4.00     8.84     4.45   0    7736    7736 22.53   779.12   0.75
adopter*              13 40300     1.00     0.00    1.00     1.00     0.00   1       1       0   NaN      NaN   0.00
tenure                14 40300    43.81    19.79   44.00    43.72    22.24   1     111     110  0.05    -0.70   0.10
good_country          15 40300     0.36     0.48    0.00     0.32     0.00   0       1       1  0.59    -1.65   0.00
------------------------------------------------------------------------------------------------------------------
group: 1
                    vars    n     mean       sd   median  trimmed      mad min    max  range  skew kurtosis     se
age                    1 3527    25.98     6.84    24.00    25.05     4.45   8     73     65  1.68     4.39   0.12
male                   2 3527     0.73     0.44     1.00     0.79     0.00   0      1      1 -1.03    -0.94   0.01
friend_cnt             3 3527    39.73   117.27    16.00    23.69    17.79   1   5089   5088 26.04  1013.79   1.97
avg_friend_age         4 3527    25.44     5.21    24.36    24.83     3.91  12     62     50  1.68     5.05   0.09
avg_friend_male        5 3527     0.64     0.25     0.67     0.65     0.25   0      1      1 -0.54    -0.05   0.00
friend_country_cnt     6 3527     7.19     8.86     4.00     5.36     4.45   0    136    136  3.61    24.53   0.15
subscriber_friend_cnt  7 3527     1.64     5.85     0.00     0.84     0.00   0    287    287 34.05  1609.52   0.10
songsListened          8 3527 33758.04 43592.73 20908.00 25811.69 23276.82   0 817290 817290  4.71    46.64 734.03
lovedTracks            9 3527   264.34   491.43   108.00   161.68   140.85   0  10220  10220  6.52    80.96   8.27
posts                 10 3527    21.20   221.99     0.00     1.44     0.00   0   8506   8506 26.52   852.38   3.74
playlists             11 3527     0.90     2.56     1.00     0.59     1.48   0    118    118 28.84  1244.31   0.04
shouts                12 3527    99.44  1156.07     9.00    23.89    11.86   0  65872  65872 52.52  2969.09  19.47
adopter*              13 3527     2.00     0.00     2.00     2.00     0.00   2      2      0   NaN      NaN   0.00
tenure                14 3527    45.58    20.04    46.00    45.60    20.76   0    111    111  0.02    -0.62   0.34
good_country          15 3527     0.29     0.45     0.00     0.23     0.00   0      1      1  0.94    -1.12   0.01
```

```
> gap
                              mean            sd   median min      max    range
age                     2.03143285    0.47176558     1.00   0       -6       -6
male                    0.10737060   -0.04050889     0.00   0        0        0
friend_cnt             21.24210554   59.79372662     9.00   0      132      132
avg_friend_age          1.42989144    0.10491273     1.36   4      -15      -19
avg_friend_male         0.02000953   -0.06869740     0.00   0        0        0
friend_country_cnt      3.23093821    3.09565085     2.00   0        7        7
subscriber_friend_cnt   1.21933283    3.43183010     0.00   0      -22      -22
songsListened       16168.59903072 15176.70461315 13468.00   0 -182710 -182710
lovedTracks           177.51816927   227.84631551    94.00   0   -2302    -2302
posts                  15.90745116   117.68401476     0.00   0   -3803    -3803
playlists               0.35148513    1.49143627     1.00   0      20       20
adopter                 1.00000000    0.00000000     1.00   1       1        0
tenure                  1.77328964    0.25489922     2.00  -1       0        1
good_country           -0.07029511   -0.02669711     0.00   0       0        0
```

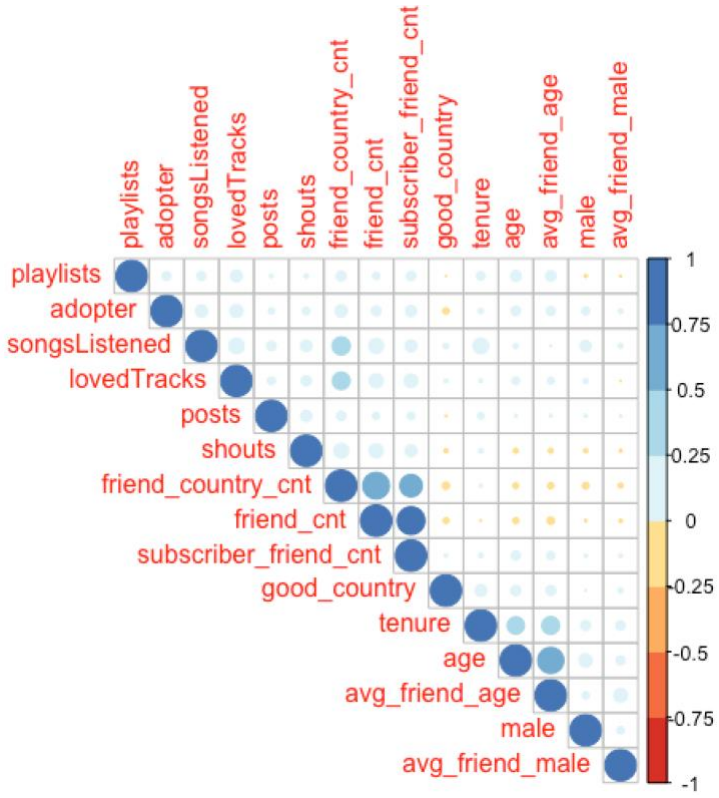
Many differences are observed in Adopter (1) and Non-subscriber (0) dataset.
On average, **subscribers are observed to be older** about 2 years more, have more **male proportion**, 21+ more friends, and their friends are about 1.4 years old. And they have **more diverse friend** base from 3.2 more countries, have 1.2 more subscriber friends**, more active on the platform** with 16,000+ more songs listened, liked 177 more tracks, made 15 more posts and slightly less share of subscribers from US, UK, Germany.

Variable	Group 0	Group 1	Difference	P-Value
Age	23.94844	25.97987	-2.03	< 0.00000000000000022
Male	0.621861	0.7292316	-0.11	< 0.00000000000000022
friend_cnt	18.49166	39.73377	-21.24	< 0.00000000000000022
avg_friend_age by adopter	24.01142	25.44131	-1.43	< 0.00000000000000022
avg_friend_male by adopter	0.6165888	0.6365983	-0.02	< 0.000009097
friend_country_cnt by adopter	3.957891	7.188829	-3.23	< 0.00000000000000022
subscriber_friend_cnt by adopter	0.417469	1.636802	-1.22	< 0.00000000000000022
songsListened by adopter	17589.44	33758.04	-16168.60	< 0.00000000000000022
lovedTracks by adopter	86.82263	264.3408	-177.52	< 0.00000000000000022
playlists by adopter	0.5492804	0.9007655	-0.35	= 0.0000000000000008619
shouts by adopter	29.97266	99.43975	-69.47	= 0.0003674
tenure	43.80993	45.58322	-1.77	= 0.0000004768
good_country	0.3577916	0.2874965	0.07	< 0.00000000000000022

2. Data Visualization:

Below sets of visualizations are generated to help understand how adopters and non-adopters of the premium subscription service differ from each other in terms of demographics, peer influence, and user engagement.

1) Correlation heatmap for all users (subscriber and non-sub combined) and all variables



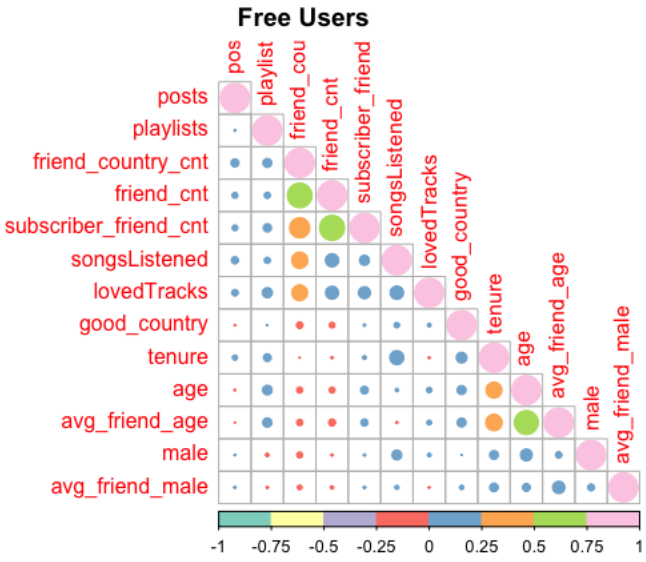
Strong Positive Correlated Relationship:

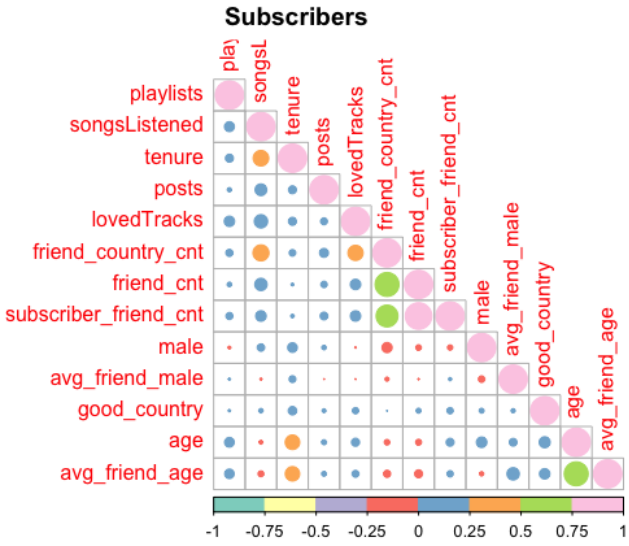
1. Number of countries that friends are from and number of friends
2. Number of friends and number of subscriber friends
3. Number of subscriber friends and number of countries that friends are from
4. Number of friends' country and songs listened
5. Number of loved tracks and number of friends
6. User's age and average of friends' age
7. Tenure and age
8. Tenure and average friends' age

Weak Negative Correlated Relationship:

1. Number of friends' country vs. tenure, age, friends' age, male, male friends
2. Shouts vs. tenure, age, friends' age, male, male friends
3. Number of friends vs. tenure, age, friends' age

2) Correlation heatmap for Subscribers vs Free Users

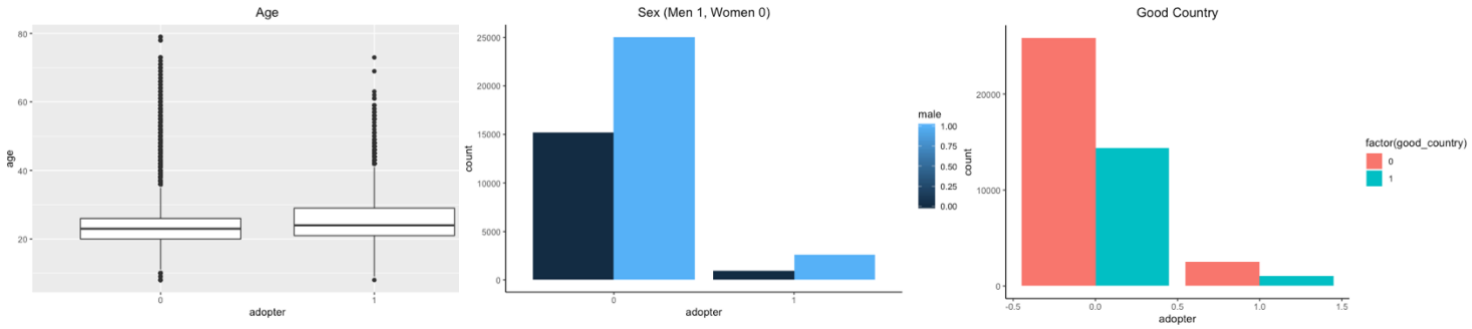




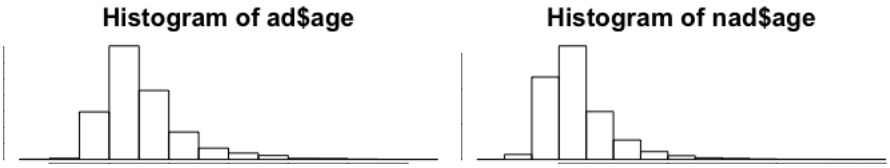
Subscriber dataset shows a very unique strong relationship between friend count and subscriber friend count. This can be understood that **there is Strong Peer Influence within subscriber g**roup to form a relationship with another subscriber or convert other users as a subscriber, on the other hand, Free users showed lower subscriber friend count

showed weaker tie with other variables. For both groups, strong relationships were found in <Age, friend age> and <friend count, friend country count>.

3) Demographic

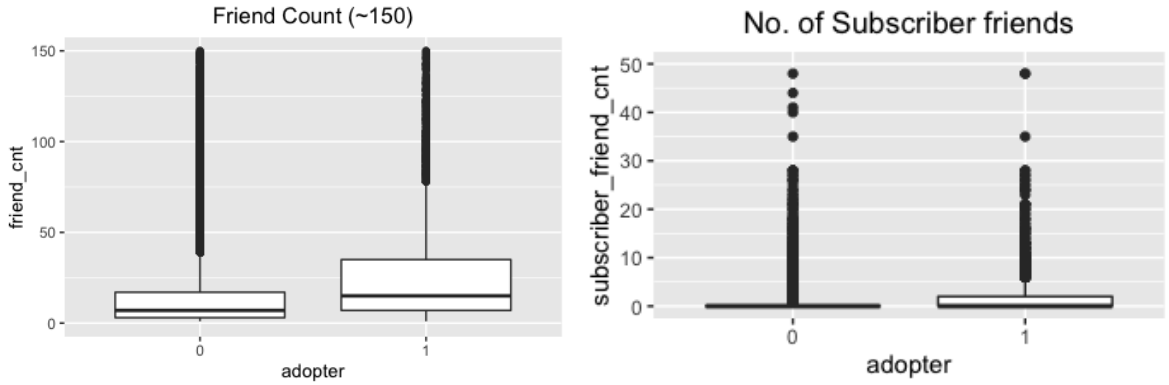


On average, subscribers are older than free users by about two years. 75% of both paid and free users are in their 20s. Examining histogram of each group, we can discover that free users (nad) are heavier on youth (early 20) and has a long tail compared to subscribers as shown below.

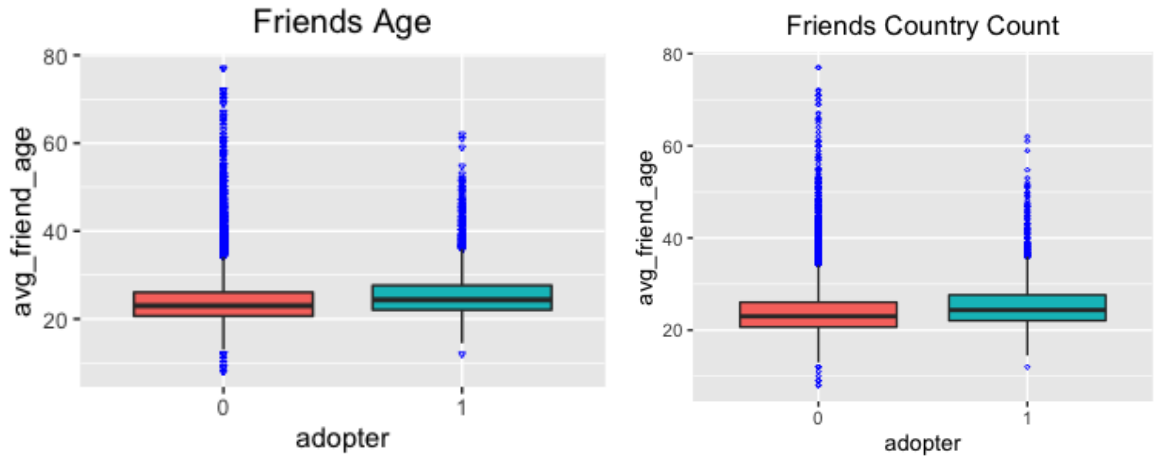


There are more male users than female users on HighNote, Also, the ratio of male user in paid group is larger than the ratio of male in free user group. This could mean that **male user's willingness of pay is higher than female users**.

4) Peer Influence

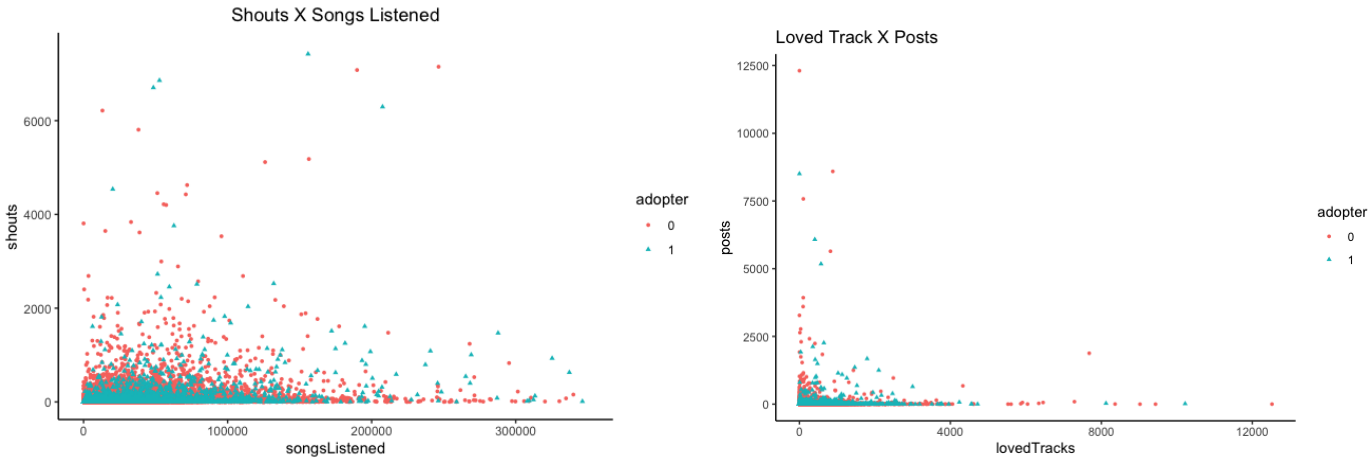


Outlier observations with a record of Friend count > 150, Subscriber friends > 50, are removed to better examine the dataset for Friend Count plot. It is not surprising to observe that paid users (adopter, 1) are more socially engaged than non-subscribers (adopter, 0) in terms of number of friends (friend_cnt) and number of paid user friend (subscriber_friend_cnt). What is worthy of noticing is that 75%+ of non paid users have 0 paid user friend as shown on the right hand side chart; On No. of subscriber friends charts, non subscribers(adopter 0) has a boxplot ranging from Q1 to Q3 on the 0 of y-axis meaning median 50% of the observed have 0 subscriber friend. Since number of friends cannot go negative, this indicates that at least 75% of the nonsubscribed users have 0 paid user friend.



Overall, stronger influence on subscription is observed among subscribers. Paid users are generally better connected as shown in their number of friends and number of countries where their friends are from. Also, we notice that both subscription and non-subscription user's friends's age plot is almost identical to Age of the user plot.

5) User Engagement



Surprisingly, shouts and Number of songs played does not show positive correlation for both users. For both users, 'Shout' isn't as actively used as they listen to songs. Paid users are more engaged to do both listen and shouts at the same time. Given that paid users have more friends as seen above, we can guess that it partially because paid users have more friends that they can recommend a song for.

According to the Loved Track X Post graph, both free and paid users do either liking albums or posting about them but not both. Surprisingly, it's rare that a user makes a post of tracks that they love.

3. Propensity Score Matching (PSM):

Here PSM is used to test whether having subscriber friends affects the likelihood of becoming an adopter (i.e., subscriber). For this purpose, 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. I will first use PSM to first create matched treatment and control samples, then test whether there is a significant average treatment effect.

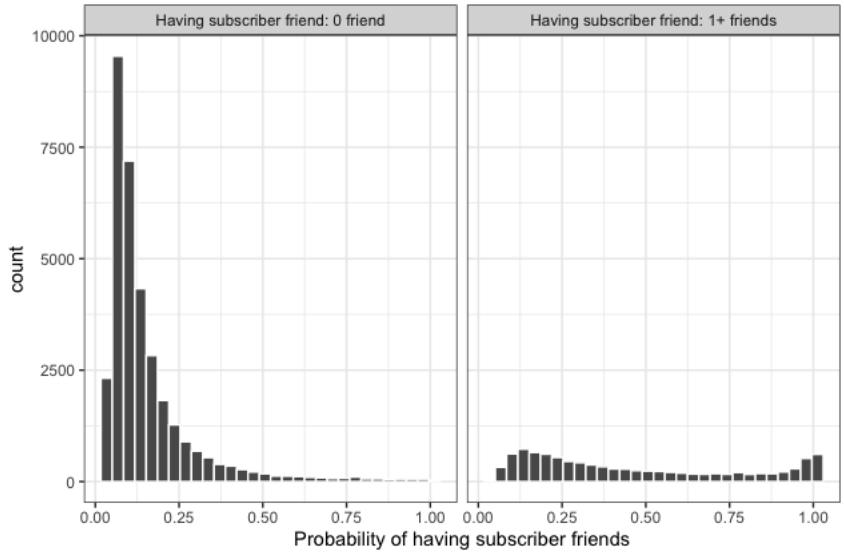
```
        Welch Two Sample t-test

data:  treatment by adopter
t = -33.978, df = 3931.7, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3109641 -0.2770354
sample estimates:
mean in group 0 mean in group 1
      0.2004715       0.4944712
```

With p-value being under 0.0001%, there's statistically significant difference in the treatment and control group.

Propensity score (distance) is simply the user's predicted probability of being treated (in this case, having more than 1 friends) given the set of observed covariates from the logit model. Propensity scores are used to reduce selection bias by equating groups based on these covariates. For the matching to give a causal estimate in the end, I included any covariate that is related to both the treatment assignment and potential outcomes.

Histograms of the estimated propensity scores by treatment status



Above historgram plots the respective propensity score (x-axis) by treatment status.
Left side histogram represents the group who don't have a subscriber friend (subscriber_friend_count = 0) and their probability of having subscriber friend. The histogram is highly left-skewed toward 0 indicating that a lot of the members are not likely to have a subscriber friend. On the other hand right-hand side histogram, users with 1+ subscriber friends, are evenly spread. LHS histogram is bigger as there are more members in the 0 subscriber friend group.

Then we can use MatchIt to estimate the propensity score in the background and then matches observations based on the method of choice ("nearest" in this case).

```
Call:
matchit(formula = treatment ~ age + male + friend_cnt + avg_friend_age + avg_friend_male +
friend_country_cnt + songsListened + lovedTracks + posts + playlists + shouts + tenure + good_country, data
= hn2, method = "nearest")
```

Summary of balance for matched data:

	Means Treated	Means Control	SD Control	Mean Diff	eQQ Med	eQQ Mean	eQQ Max
distance	0.4635	0.3040	0.1913	0.1596	0.1077	0.1596	0.4517
age	25.3732	26.3324	7.9056	-0.9592	1.0000	0.9592	7.0000
male	0.6363	0.6576	0.4745	-0.0214	0.0000	0.0214	1.0000
friend_cnt	54.0210	21.4666	23.5251	32.5544	12.0000	32.5544	4794.0000
avg_friend_age	25.3904	26.5572	6.7320	-1.1668	0.4376	1.2763	14.0000
avg_friend_male	0.6358	0.6551	0.2643	-0.0193	0.0158	0.0326	0.1602
friend_country_cnt	9.3856	5.0914	4.6473	4.2942	2.0000	4.2942	95.0000
songsListened	33735.6404	27360.8630	33892.7804	6374.7775	4680.0000	6374.7775	566867.0000
lovedTracks	225.3647	134.5440	299.1995	90.8206	38.0000	90.8206	6180.0000
posts	20.5230	6.2773	60.2598	14.2456	0.0000	14.2456	9535.0000
playlists	0.7441	0.6723	1.4015	0.0718	0.0000	0.1035	22.0000
shouts	101.8195	37.2362	138.8781	64.5833	10.0000	64.5833	59168.0000
tenure	46.5487	47.7039	19.0357	-1.1551	1.0000	1.2995	4.0000
good_country	0.3433	0.3581	0.4795	-0.0149	0.0000	0.0149	1.0000

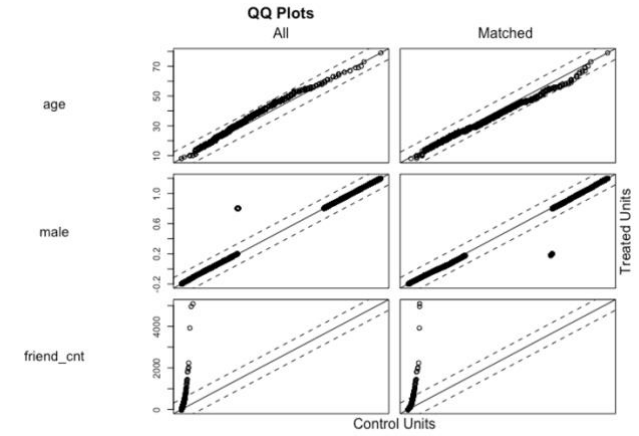
Percent Balance Improvement:

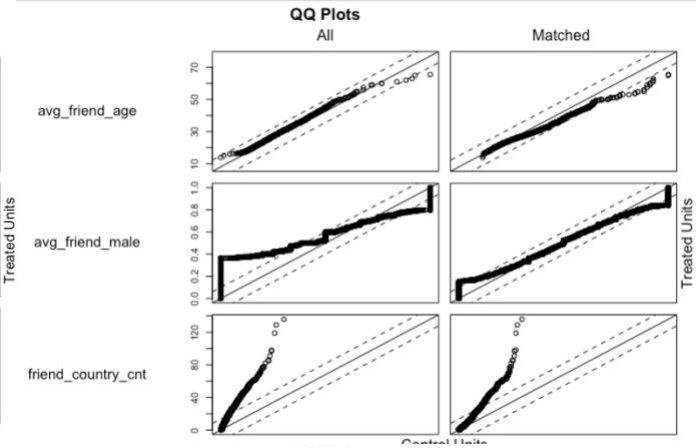
	Mean Diff.	eQQ Med	eQQ Mean	eQQ Max
distance	48.2930	57.0083	48.2908	33.9658
age	40.9972	0.0000	41.1419	-40.0000
male	-187.9614	0.0000	-187.6712	0.0000
friend_cnt	25.3162	45.4545	25.3062	0.0000
avg_friend_age	28.3760	72.4916	22.0309	-21.7391
avg_friend_male	14.7957	78.6165	65.9532	55.9466
friend_country_cnt	35.5279	60.0000	35.5203	0.0000
songsListened	66.6825	69.7499	66.6699	13.2836
lovedTracks	43.2906	41.5385	43.2216	2.5698
posts	20.7676	0.0000	20.3394	0.0000
playlists	66.5567	0.0000	50.5109	15.3846
shouts	24.3724	33.3333	24.1770	0.0000
tenure	65.4771	66.6667	61.1782	60.0000
good_country	-30.1771	0.0000	-30.3571	0.0000

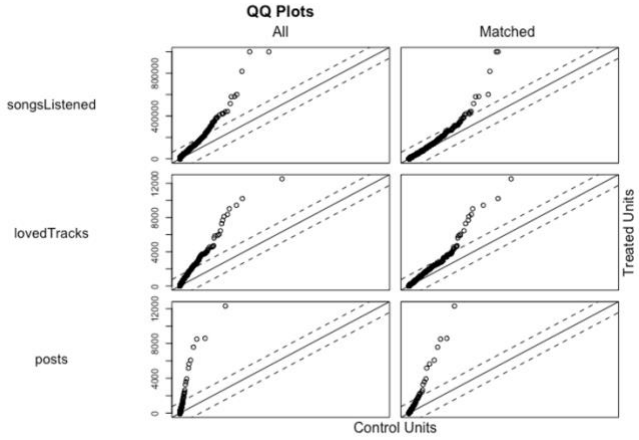
Sample sizes:

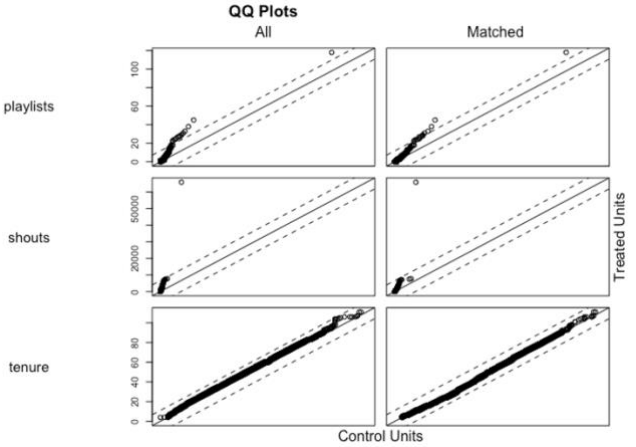
	Control	Treated
All	34004	9823
Matched	9823	9823
Unmatched	24181	0
Discarded	0	0

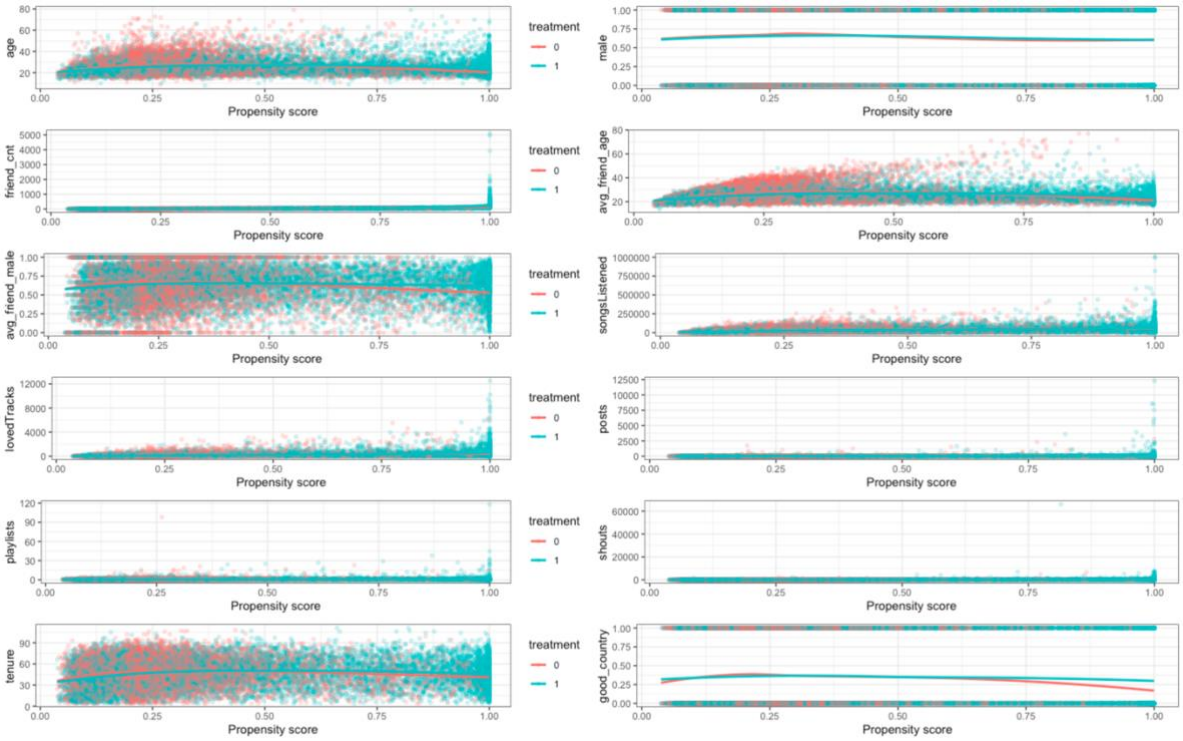
There are 9,823 pairs of treated and controlled records (total of 19,646) that have been matched based on Propensity score (variable called distance). Quantile-quantile (QQ) plot compares the probability distributions of the treated and control groups on a given covariate by plotting their quantiles against each other. As shown in the below QQ plot chart, the results show that although the points are not located on the y=x line exactly after matching, it is slightly improved as compared to original data.












```
> propensity
    pr_score treatment
1  0.08597334         0
2  0.14417767         0
3  0.08217010         0
4  0.23894067         1
5  0.69552208         0
6  0.22306633         0
7  0.12644080         0
8  0.79381453         1
9  0.06475981         0
10 0.27014010         0
11 0.27304943         1
12 0.07225773         0
13 0.04481833         0
14 0.16221702         0
15 0.28976083         0
16 0.08571043         0
17 0.19854136         1
18 0.04583879         0
19 0.06664671         0
20 0.12940208         1
21 0.15150629         1
22 0.15875058         1
23 0.07700667         0
```

4. Regression Analyses:

Now I will use a logistic regression approach to test which variables (including subscriber friends) are significant for explaining the likelihood of becoming an adopter. I used my best judgment and visualization results to decide which variables to include in the regression. After that, I estimate the odds ratios for the key variables.

1) Logistic Regression Model with all the variables including treatment to test which variables are significant for explaining the likelihood of becoming an adopter

```
Call:
glm(formula = adopter ~ treatment + age + male + friend_cnt +
    avg_friend_age + avg_friend_male + friend_country_cnt + songsListened +
    lovedTracks + posts + playlists + shouts + tenure + good_country,
    family = "binomial", data = dta_m)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2240  -0.5668  -0.4562  -0.3697   2.5257

Coefficients:
                     Estimate    Std. Error z value             Pr(>|z|)
(Intercept)       -3.3709684810  0.1261268190 -26.727 < 0.0000000000000002 ***
treatment          0.7292753804  0.0468153008  15.578 < 0.0000000000000002 ***
age                0.0141945749  0.0040691111   3.488              0.000486 ***
male               0.3039659743  0.0489909251   6.205 0.000000000548584009 ***
friend_cnt        -0.0002001575  0.0002792679  -0.717              0.473545
avg_friend_age     0.0130425316  0.0053490465   2.438              0.014757 *
avg_friend_male    0.0600688475  0.0925246717   0.649              0.516196
friend_country_cnt 0.0072774704  0.0036486080   1.995              0.046088 *
songsListened      0.0000042553  0.0000005301   8.027 0.000000000000000997 ***
lovedTracks        0.0005210828  0.0000469224  11.105 < 0.0000000000000002 ***
posts              0.0001186059  0.0000889116   1.334              0.182212
playlists          0.0446485958  0.0119444028   3.738              0.000185 ***
shouts             0.0001119478  0.0000745439   1.502              0.133156
tenure            -0.0024335424  0.0012170876  -1.999              0.045556 *
good_country      -0.3694917638  0.0480180878  -7.695 0.000000000000014167 ***
```

Given the P-Value, variables such as friend count, average friend being a male, posts and shouts are NOT statistically significant.

On the other hand, Variables like age, sex, subscriber friends, songs listened, loved tracks, number of friends' country and playlists, etc have significant impact on the dependent variable, adopter.

For model improvement, variables with statistical significance will be used only.

2) Optimized model only with statistically significant variables.

Regression: glm(formula = adopter ~ treatment + age + male + avg_friend_age + friend_country_cnt + songsListened + lovedTracks + playlists + tenure + good_country, data = dta_m)

Variables used based on Significance (P-value): treatment, age, male, avg_friend_age, friend_country_cnt, songsListened, lovedTracks, playlists, tenure, good_country

```
Call:
glm(formula = adopter ~ treatment + age + male + avg_friend_age +
    friend_country_cnt + songsListened + lovedTracks + playlists +
    tenure + good_country, data = dta_m)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.28859  -0.15589  -0.10624  -0.05655   0.99818

Coefficients:
                     Estimate     Std. Error t value             Pr(>|t|)
(Intercept)        0.96622744020  0.01265349611  76.361 < 0.0000000000000002 ***
treatment          0.07693526599  0.00490039844  15.700 < 0.0000000000000002 ***
age                0.00171597141  0.00045852615   3.742              0.000183 ***
male               0.03029474232  0.00511504251   5.923  0.00000003220222844 ***
avg_friend_age     0.00165086368  0.00056940879   2.899              0.003745 **
friend_country_cnt 0.00108730614  0.00032514409   3.344              0.000827 ***
songsListened      0.00000063991  0.00000006592   9.708 < 0.0000000000000002 ***
lovedTracks        0.00008598994  0.00000596592  14.414 < 0.0000000000000002 ***
playlists          0.00719924354  0.00140642590   5.119  0.00000031035110903 ***
tenure            -0.00028898123  0.00013195338  -2.190              0.028534 *
good_country      -0.03911872213  0.00501421899  -7.802  0.00000000000000642 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1091944)

    Null deviance: 2253.7  on 19645  degrees of freedom
Residual deviance: 2144.0  on 19635  degrees of freedom
AIC: 12257

Number of Fisher Scoring iterations: 2
```

3) Odds ratio: Used exp function to enhance the interpretability of the coefficient

```
> #Odds ratio of the variables
> exp(glm_imp$coefficients)
       (Intercept)          treatment                age               male      avg_friend_age
         0.9667914          1.0799722          1.0017174          1.0307583          1.0016522
friend_country_cnt      songsListened        lovedTracks          playlists             tenure
         1.0010879          1.0000006          1.0000860          1.0072252          0.9997111
      good_country
         0.9616365
```

- Treatment odds ratio 1.079 means: Having 1 or more subscriber friends (treatment) increases the chance of being an adopter by 7.9%.
- Male odds ratio 1.0307 : Male users have 3% higher chance of being an adopter
- Playlists odds ratio 1.00722 : Adopter increases 0.7% at every time when one playlist added.
- Age, average friends' age, number of friends' country, number of songs listened, number of loved tracks have positive impact in converting to an adopter, however, the impact is not significant (appx. 0.1).
- Odds ratio for Tenure and Country (UK, US, Germany) are less than 1, meaning they are negatively impacting the chances of being adopter. The longer the users use HighNote for free, the less likely HighNote convert them as a paid user.

Input variables in the model are not log-transformed as neither logistic nor OLS regression requires normally distributed independent variables. GLM is used as predicted output is binary. For logistic regressions, even the residuals don't need to be normally distributed. [1]

5. Takeaways & strategy suggestions to covert free users to paid user subscription (Free-to-fee)

Strategy Focus: Social network relationship

[1] Assumptions of Logistic Regression: https://www.statisticssolutions.com/assumptions-of-logistic-regression/

The analysis result after propensity score matching suggests that being friends with a subscribed user has the most influence on free users to convert them to a paid user. 1 subscribed user increase in a non-subscribe user's network will increase chance by 7.9% for the non-paid user to convert premium.

Current situation: HighNote has 43,827 users on its platform. Currently only 8.75% of the users are paid users and the rest (91.25%) is unpaid users. This means that HighNote has large potential customers to capture by leveraging existing customer base.

Strategy suggestions: New user engagement through referral from paid users, promoting active interaction and networking for paid users with free users, Encouraging in-platform activities for paid users.

- Family or friends plan (as phone plans): a group of 4~6 free users invited by paid users get discount under the family plan
- Promoting music feed of the premium users: premium users will get publicity on their likes, playlists, new friends, posts so that free users can be exposed to premium users more often for future connection and understand premium perks (publicity and social influence) better. Overall user engagement will be increased.
- Incentives for paid users for referral recruiting: Offer 1 month free coupon at every new user engagement milestone achieved. This will keep the super users longer within the platform and help recruit new paid users through current users.
- Targeting male users more: the study shows that male users are more likely to convert, therefore, HighNote can put an AD targeting male population in terms of contents or channel. For instance, HighNote can consider inserting an YouTube AD before a sports game review video.