HighNote-Free to Fee Strategy

After loading the data into R markdown, the data is separated into two groups based on whether the users are premium users or not. (adopter=1 premium users; adopter=0 free users). The summary statistics by groups is shown below:

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	0.00	0.00	0.00	0.00	0.00	0	0	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	1.00	0.00	1.00	1.00	0.00	1	1	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

Compared the mean difference for the two groups:

dt2$adopter <int>	age <dbl>	male <dbl>	friend_cnt <dbl>	avg_friend_age <dbl>	avg_friend_male <dbl>	friend_country_cnt <dbl>	subscriber_friend_cnt <dbl>
0	23.94844	0.6218610	18.49166	24.01142	0.6165888	3.957891	0.417469
1	25.97987	0.7292316	39.73377	25.44131	0.6365983	7.188829	1.636802

subscriber_friend_cnt <dbl>	songsListened <dbl>	lovedTracks <dbl>	posts <dbl>	playlists <dbl>	shouts <dbl>	adopter <dbl>	tenure <dbl>	good_country <dbl>
0.417469	17589.44	86.82263	5.293002	0.5492804	29.97266	0	43.80993	0.3577916
1.636802	33758.04	264.34080	21.200454	0.9007655	99.43975	1	45.58322	0.2874965

From the mean difference, one can see that:

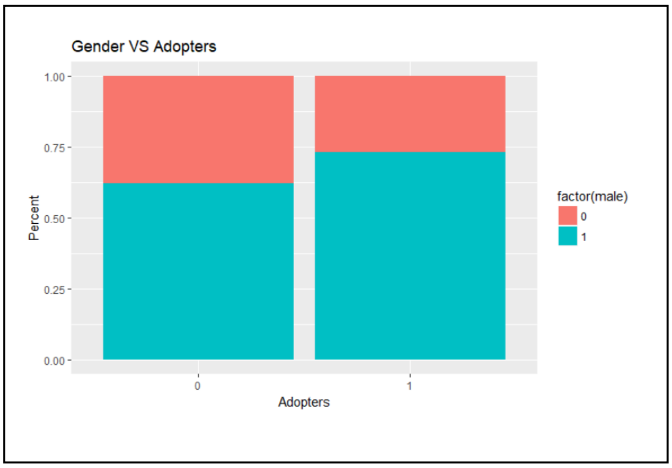
- The adopter group subscribers on average are(switched from Free to Fee) older, have more male, more friends with average friends age is larger than non-adopter group. Also, the adopter subscribers have higher proportion of male friends and more diverse friend and higher percentage of friends who are also adopters.

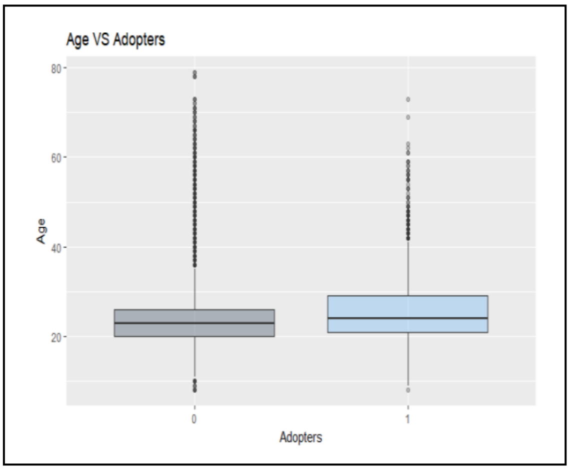
Therefore, peer influence and user engagement may affect users' decisions to pay for a premium subscription.

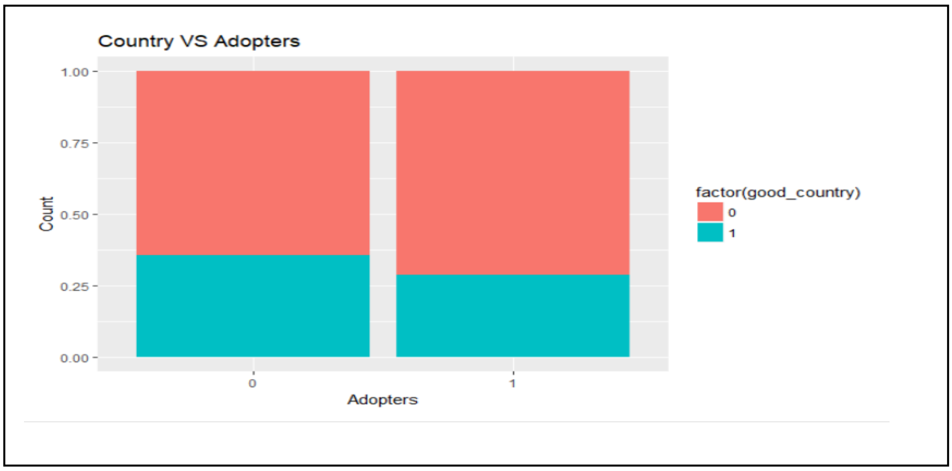
- Also, adopters are more engaged with larger number of songs listened and more posts, playlists and shouts.
- what is more, HighNote has fewer users in US,UK and Germany than other countries.

Then a set of charts are drawn to visualize how adopters (adopter=1) and non-adopters (adopter=0) differ from each other in terms of (1) demographics, (2) peer influence, (3) user engagement.

- (1) demographic include characteristics such as age, gender and country:

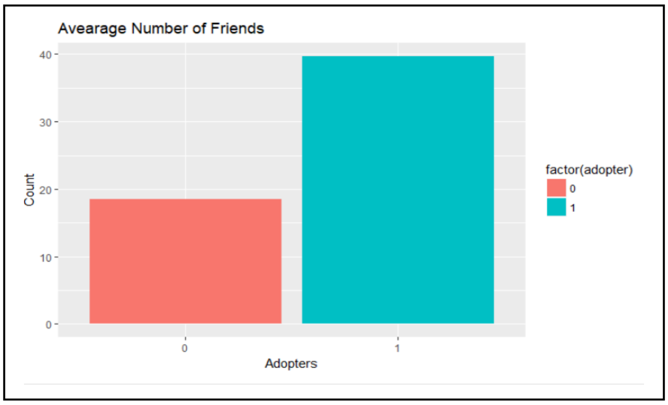


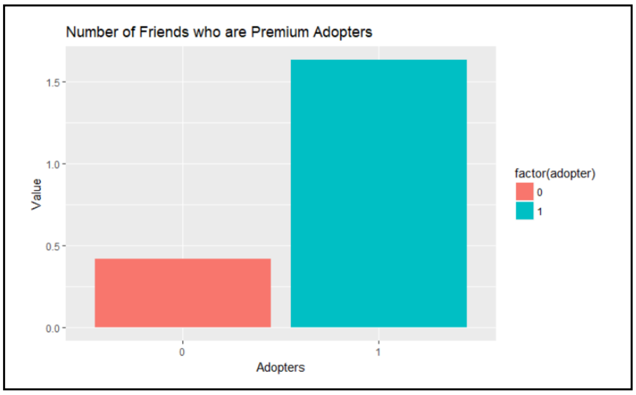




✓ As it shown on the graph above, adopters are more males, are older than non-adopters and more come from countries other than US, UK and Germany. The results are consistent with the previous observation.

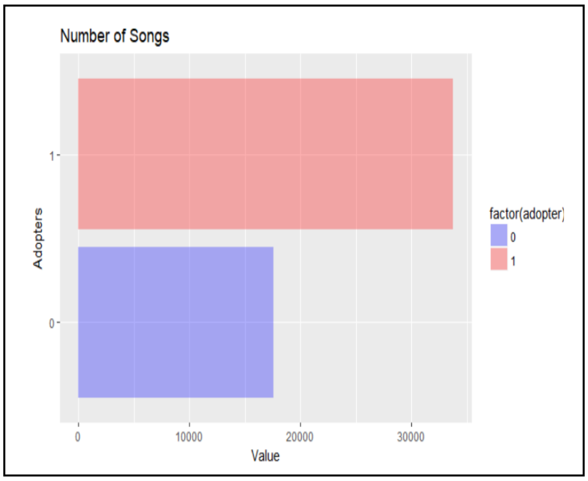
- (2) peer influence includes characteristics such as number of friends and number of friends who are adopters.

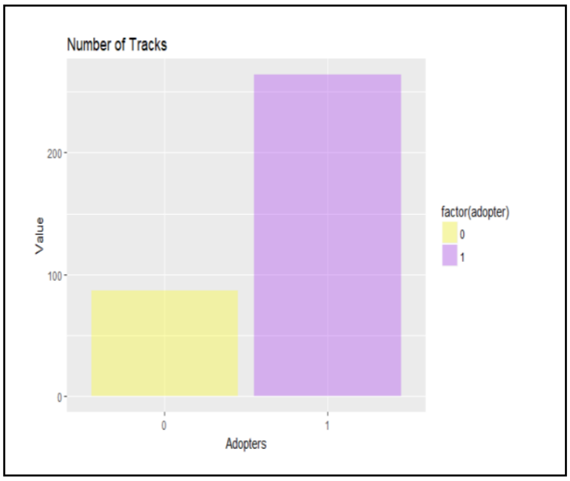


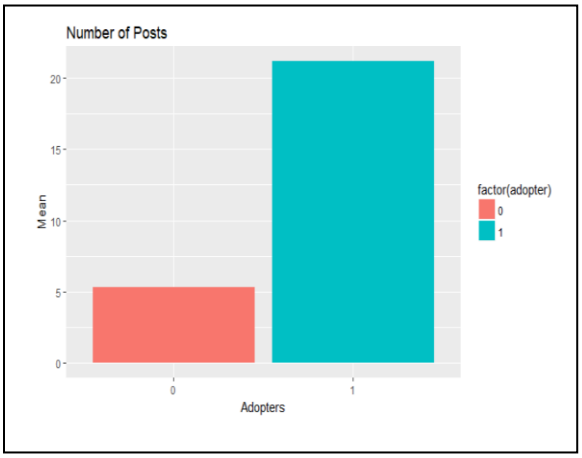


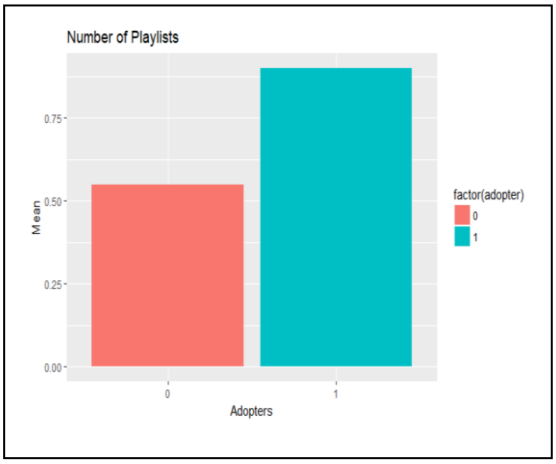
- ✓ As it shown above, adopters have more friends and they have more friends who are premium subscribers. Therefore, there might have peer influence exists.

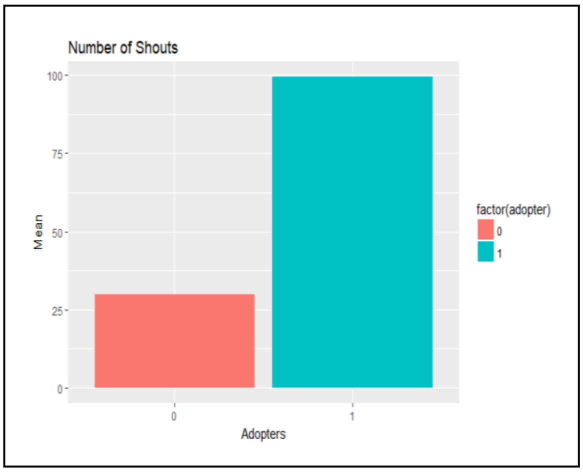
- (3) Engagement level data on activities performed when using the service, which include the number of songs the user has listened to, playlists created, "shouts" sent to friends, etc.

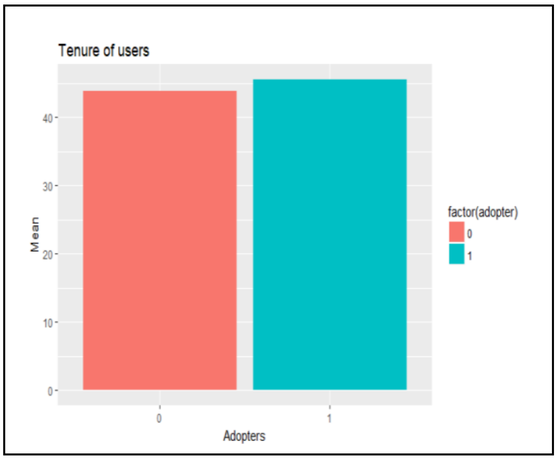












✓ As the graphs showed, premium users on average have more songs listened, more tracks, posts, playlists, shouts than free users. Therefore, premium users might be more engaged or more engaged users are more likely to be premium users. However, the tenure difference is not huge between the two group of users.

After visualization, PSM is used to test whether having subscriber friends affects the likelihood of becoming an adopter (i.e., fee customer). For this purpose, the "treatment" group will be users that have one or more subscriber friends (subscriber_friend_cnt), while the "control" group will include users with zero subscriber friends. (codes can be seen in the attached R file)

- Before PSM, the t-tests are carried out to evaluate whether means of all variables are statistically distinguishable:
  Here as it showed, mean of "male" are not statistically distinguishable (p>0.05), so it will be left out from the PSM. (details can be found in R file)

```
[[2]]

        Welch Two Sample t-test

data:  dt2[, v] by dt2[, "test"]
t = -1.3459, df = 15986, p-value = 0.1784
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.018236129  0.003388028
sample estimates:
mean in group 0 mean in group 1 
      0.6288378       0.6362618 
```

The following model is used to calculate the PS for each user:

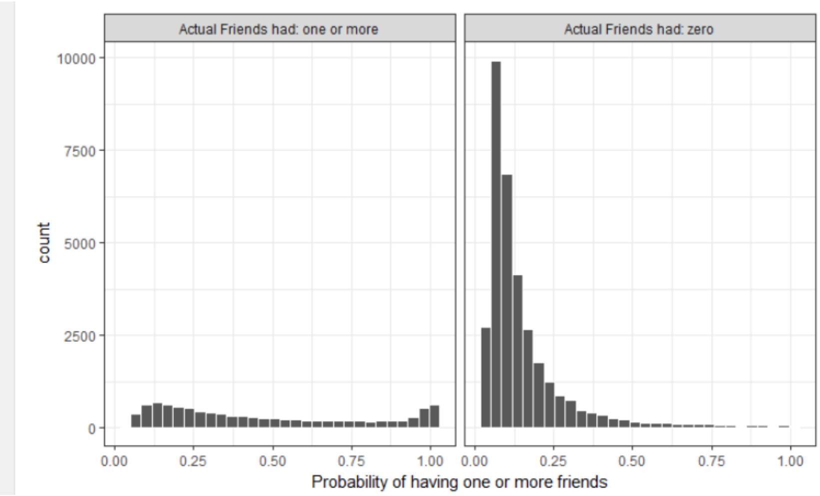
```
Call:
glm(formula = test ~ age + friend_cnt + avg_friend_age + avg_friend_male + 
    friend_country_cnt + songsListened + lovedTracks + posts + 
    playlists + shouts + adopter + tenure + good_country, family = binomial, 
    data = dt2)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.3520  -0.5621  -0.4143  -0.2960   2.5603  

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -5.111e+00  7.619e-02 -67.086  < 2e-16 ***
age                 1.838e-02  2.786e-03   6.596 4.23e-11 ***
friend_cnt          3.084e-02  1.038e-03  29.713  < 2e-16 ***
avg_friend_age      7.785e-02  3.479e-03  22.376  < 2e-16 ***
avg_friend_male     2.484e-01  5.058e-02   4.910 9.09e-07 ***
friend_country_cnt  1.104e-01  4.767e-03  23.154  < 2e-16 ***
songsListened       6.417e-06  5.136e-07  12.494  < 2e-16 ***
lovedTracks         5.557e-04  5.612e-05   9.901  < 2e-16 ***
posts               5.248e-04  2.613e-04   2.008  0.04465 *  
playlists          -5.993e-03  1.271e-02  -0.472  0.63717    
shouts             -5.729e-05  3.790e-05  -1.512  0.13058    
adopter             8.039e-01  4.423e-02  18.176  < 2e-16 ***
tenure             -2.094e-03  7.799e-04  -2.684  0.00727 ** 
good_country        5.934e-02  2.939e-02   2.019  0.04351 *  
---
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: 33856  on 43813  degrees of freedom
AIC: 33884

Number of Fisher Scoring iterations: 6
```

The PS calculated for each users are shown below:

	pr_score <dbl>	treatment <dbl>
1	0.08417235	0
2	0.13814668	0
3	0.07621456	0
4	0.23069166	1
5	0.67574975	0
6	0.20448624	0

6 rows

After estimating the PS, a histogram is plotted to examine the region of common support:



- Therefore, from the graph, one can see that for treatment group, there are many samples who should have more friends had zero friend; while there are many samples who should have zero friends had many friends, making our matching feasible.

Then a match algorithm was executed by using the package MatchIt based on the method of choice ("nearest"). After matching, the final dataset is smaller than the original: it contains 19646 observations, meaning that 9823 pairs of treated and control observations were matched.

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

```
[1] 19646    18
```

Next, difference-in-mean is tested to assess covariate balance in the matched sample:

test <dbl>	age <dbl>	friend_cnt <dbl>	avg_friend_age <dbl>	avg_friend_male <dbl>	friend_country_cnt <dbl>	songsListened <dbl>	lovedTracks <dbl>
0	26.28016	21.21358	26.49360	0.6532951	5.053039	26952.95	134.9268
1	25.37321	54.02097	25.39043	0.6358077	9.385626	33735.64	225.3647

songsListened <dbl>	lovedTracks <dbl>	posts <dbl>	playlists <dbl>	shouts <dbl>	adopter <dbl>	tenure <dbl>	good_country <dbl>
26952.95	134.9268	6.02260	0.6605925	37.19037	0.1467983	47.58984	0.3673012
33735.64	225.3647	20.52296	0.7440700	101.81951	0.1775425	46.54871	0.3432760

Also, the treatment effects are estimated using a t-test:

```
        Welch Two Sample t-test

data:  adopter by test
t = -5.8501, df = 19529, p-value = 4.992e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04104508 -0.02044326
sample estimates:
mean in group 0 mean in group 1
      0.1467983       0.1775425
```

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0899  -0.6091  -0.5569  -0.4832   2.2109

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.884e+00  1.093e-01 -17.235  < 2e-16 ***
test                1.626e-01  4.090e-02   3.976 7.00e-05 ***
age                 2.151e-02  3.614e-03   5.951 2.67e-09 ***
friend_cnt          4.245e-04  2.547e-04   1.666  0.09566 .
avg_friend_age     -1.305e-02  4.961e-03  -2.630  0.00853 **
avg_friend_male     1.580e-02  8.216e-02   0.192  0.84746
friend_country_cnt -1.103e-02  3.563e-03  -3.095  0.00197 **
songsListened       3.370e-06  4.944e-07   6.817 9.31e-12 ***
lovedTracks         4.918e-04  4.490e-05  10.953  < 2e-16 ***
posts               1.510e-04  8.829e-05   1.710  0.08731 .
playlists           7.433e-02  1.338e-02   5.557 2.75e-08 ***
shouts              1.089e-04  7.159e-05   1.521  0.12816
tenure             -3.526e-03  1.105e-03  -3.191  0.00142 **
good_country       -3.638e-01  4.330e-02  -8.402  < 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: 17416  on 19645  degrees of freedom
Residual deviance: 16996  on 19632  degrees of freedom
AIC: 17024

Number of Fisher Scoring iterations: 4
```

```{r}
print(exp(glm_treat2$coefficients[1:14]))
```

```
       (Intercept)               test                age         friend_cnt     avg_friend_age    avg_friend_male
         0.1520344          1.1766083          1.0217399          1.0004246          0.9870352          1.0159300
friend_country_cnt      songsListened        lovedTracks              posts          playlists             shouts
         0.9890336          1.0000034          1.0004919          1.0001510          1.0771601          1.0001089
            tenure       good_country
         0.9964807          0.6950245
```

- After doing propensity score matching, as the result shows, keeping other covariates constant, compared with users who have zero subscriber friend, users have one or more subscriber friends have 17% higher chance to be an adopter. The P-value <0.05, which means there's significant treatment effect.

- Keeping other covariates constant, 1 unit increases in age would result in 2.1% increase in the chance of being an adopter.
- Keeping other covariates constant, 1 unit increases in average_friend_age would result 2% decrease in the chance of being an adopter.
- Keeping other covariates constant, 1 unit increases in friend_country_cnt would result 2% decrease in the chance of being an adopter.
- Keeping other covariates constant, 1 unit increases in (songslistened/lovedtrack/posts/playlist) would result 0 % increase in the chance of being an adopter.
- Keeping other covariates constant, 1 unit increases in tenure would result 1 % decrease in the chance of being an adopter.
- Keeping other covariates constant, 1 unit increases in good_country would result 30 % decrease in the chance of being an adopter.

Next, a logistic regression is done using the original dataset without matching:

- Based on the visualization in part 2, variables to be included in the model could be:

reg_val<-c("male","subscriber_friend_cnt","good_country","friend_cnt","avg_friend_age","friend_country_cnt","songsListened","lovedTracks","posts","playlists","shouts","adopter")

```
Call:
glm(formula = adopter ~ male + good_country + subscriber_friend_cnt +
    friend_cnt + avg_friend_age + friend_country_cnt + songsListened +
    lovedTracks + posts + playlists + shouts, family = binomial,
    data = dt2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.4957  -0.4126  -0.3509  -0.2932   2.7104

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.158e+00  9.014e-02 -46.133  < 2e-16 ***
male                   4.448e-01  4.102e-02  10.843  < 2e-16 ***
good_country          -4.142e-01  4.054e-02 -10.217  < 2e-16 ***
subscriber_friend_cnt  9.725e-02  1.077e-02   9.031  < 2e-16 ***
friend_cnt            -4.515e-03  5.000e-04  -9.029  < 2e-16 ***
avg_friend_age         4.273e-02  3.253e-03  13.134  < 2e-16 ***
friend_country_cnt     4.401e-02  3.658e-03  12.032  < 2e-16 ***
songsListened          7.016e-06  4.914e-07  14.278  < 2e-16 ***
lovedTracks            7.075e-04  4.932e-05  14.344  < 2e-16 ***
posts                  7.162e-05  9.507e-05   0.753    0.451
playlists              6.223e-02  1.349e-02   4.613 3.96e-06 ***
shouts                 9.798e-05  8.227e-05   1.191    0.234
---
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: 22659  on 43815  degrees of freedom
AIC: 22683

Number of Fisher Scoring iterations: 5
```

```
#odds ratio
```{r}
print(exp(fit.glm$coefficients[1:12]))

```
```

```
        (Intercept)                male        good_country subscriber_friend_cnt          friend_cnt
         0.01563431          1.56015560          0.66084227            1.10214027          0.99549538
     avg_friend_age  friend_country_cnt       songsListened           lovedTracks               posts
         1.04365596          1.04499132          1.00000702            1.00070773          1.00007162
          playlists              shouts
         1.06420469          1.00009799
```

Interpretation of the results:

1 unit increase in "male" (which is male) increases the odd of switching to Fee user by a factor of 1.56

1 unit increase in "good_country" (which is users from US, UK and Germany) decreases the odd of switching to Fee user by a factor of 0.66

1 unit increase in "subscriber_friend_cnt" (which is users have one or more subscriber friends) increases the odd of switching to Fee user by a factor of 1.10

1 unit increase in "friend_cnt" decreases the odd of switching to Fee user by a factor of 1.10
1 unit increase in "average_friend_age" increases the odd of switching to Fee user by a factor of 1.04

1 unit increase in "friend_country_cnt" increases the odd of switching to Fee user by a factor of 1.04
1 unit increase in "songsListened" increases the odd of switching to Fee user by a factor of 1
1 unit increase in "lovedTracks" increases the odd of switching to Fee user by a factor of 1
1 unit increase in "playlists" increases the odd of switching to Fee user by a factor of 1.06

Takeaways:
From the analysis, we can see that male users that are from countries other than US, UK and Germany are more likely to become fee users. As a result, Highnote could target this group of customers. Also, number of friends a user has is negatively correlated with the variable "adopter" while number of subscribers friends is positively correlated. Therefore, it provides insights for the company that the "quality" of the friends for users outweighed the "quantity" of friends for users. If Highnote would like to send promotions to users, they should target users who have more subscriber friends. In addition, users who are more engaged are more likely to switch to fee users as they have more posts, more loved tracks etc. therefore, Highnote could find strategies in improving user engagement. For example, Highnote can boost user activity with frequency updates.