MIDTERM – BANA 277

DARSHANA DAGA

DATE – 02/25/2021

Contents

Question 1 – DESCRIPTIVE STATISTICS

Understanding High notes customer segments to better analyze the behavior of free and premium app users. And to analyze and quantify social engagement's effect on revenue. It's to be kept in mind that each additional premium users provides 24 times more revenue than a free customer.

We start by analyzing the basic descriptive statistics on the given data grouped by the adopter. Where Adopter = 0 are free customers and Adopters = 1 are premium customer

ADOPTERS

Statistics

Variable	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
age	25.98	6.84	8.00	21.00	24.00	29.00	73.00
male	0.73	0.44	0.00	0.00	1.00	1.00	1.00
friend_cnt	39.73	117.27	1.00	7.00	16.00	40.00	5089.00
avg_friend_age	25.44	5.21	12.00	22.07	24.36	27.64	62.00
avg_friend_male	0.64	0.25	0.00	0.50	0.67	0.81	1.00
friend_country_cnt	7.19	8.86	0.00	2.00	4.00	9.00	136.00
subscriber_friend_cnt	1.64	5.85	0.00	0.00	0.00	2.00	287.00
songsListened	33758.04	43592.73	0.00	7803.00	20908.00	44040.00	817290.00
lovedTracks	264.34	491.43	0.00	30.00	108.00	292.00	10220.00
posts	21.20	221.99	0.00	0.00	0.00	2.00	8506.00
playlists	0.90	2.56	0.00	0.00	1.00	1.00	118.00
shouts	99.44	1156.07	0.00	2.00	9.00	41.00	65872.00
adopter	1.00	0.00	1.00	1.00	1.00	1.00	1.00
tenure	45.58	20.04	0.00	32.00	46.00	60.00	111.00
good_country	0.29	0.45	0.00	0.00	0.00	1.00	1.00

NON-ADOPTERS

Statistics

Variable	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
age	23.95	6.37	8.00	20.00	23.00	26.00	79.00
male	0.62	0.48	0.00	0.00	1.00	1.00	1.00
friend_cnt	18.49	57.48	1.00	3.00	7.00	18.00	4957.00
avg_friend_age	24.01	5.10	8.00	20.67	23.00	26.06	77.00
avg_friend_male	0.62	0.32	0.00	0.43	0.67	0.90	1.00
friend_country_cnt	3.96	5.76	0.00	1.00	2.00	4.00	129.00
subscriber_friend_cnt	0.42	2.42	0.00	0.00	0.00	0.00	309.00
songsListened	17589.44	28416.02	0.00	1252.00	7440.00	22894.25	1000000.00
lovedTracks	86.82	263.58	0.00	1.00	14.00	72.00	12522.00
posts	5.29	104.31	0.00	0.00	0.00	0.00	12309.00
playlists	0.55	1.07	0.00	0.00	0.00	1.00	98.00
shouts	29.97	150.69	0.00	1.00	4.00	15.00	7736.00
adopter	0.00	0.00	0.00	0.00	0.00	0.00	0.00
tenure	43.81	19.79	1.00	29.00	44.00	59.00	111.00
good_country	0.36	0.48	0.00	0.00	0.00	1.00	1.00

OBSERVATION

- There is no missing data in the provided data set and the number of rows is 43827 and number of columns are 16.
- Also there are 40300 Free Users and 3572 Premium Users.

DEMOGRAPHICS

- The mean age of adopters is higher than that of non-adopters. The data is skewed, and it will be important to look at the median, which also is 24 higher than 23 for non-adopters.
- Both adopters and non-adopters are majorly male. Adopters are slightly more male dominated.

PEER INFLUENCE

- For all the peer influence factors Adopters seems to have more engagement with peers, they have more friends, and the friends are older compared to non-adopters.
- They also seem to have friends who are also subscribers more compared to non-adopters.

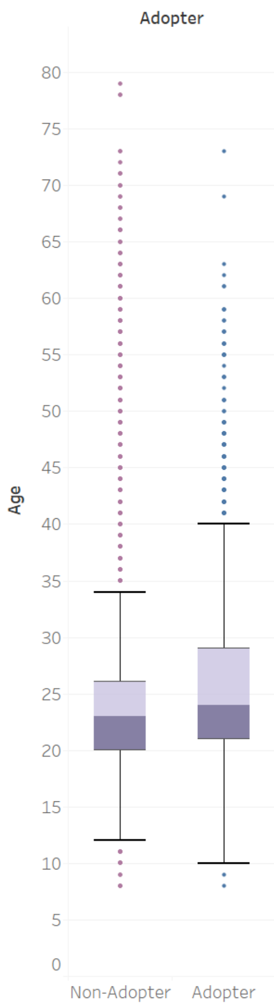
USER ENGAGEMENT

- Adopters have loved songs almost 3 times compared to non- adopters.
- They post 4 times more compared to non-premium customers.
- Adopters have 231% more shouts than non-adopters.
- Overall Adopters are much more engaged than non-adopters.

Question 2 - VIZUALIZATION

CUSTOMER DEMOGRAPHICS

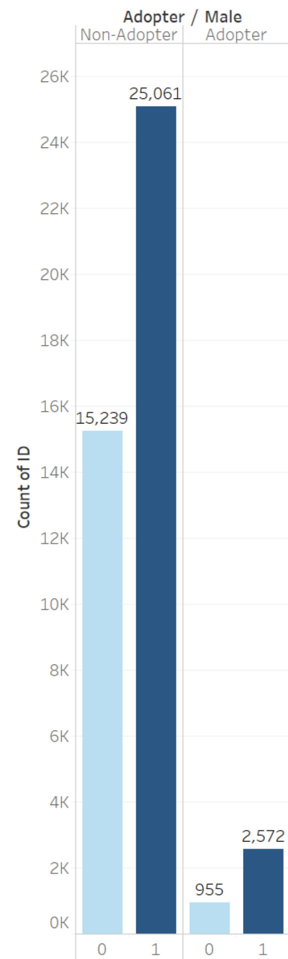
Age



Age for each Adopter. Color shows details about Adopter.

Adopter
- Non-Adopter
- Adopter

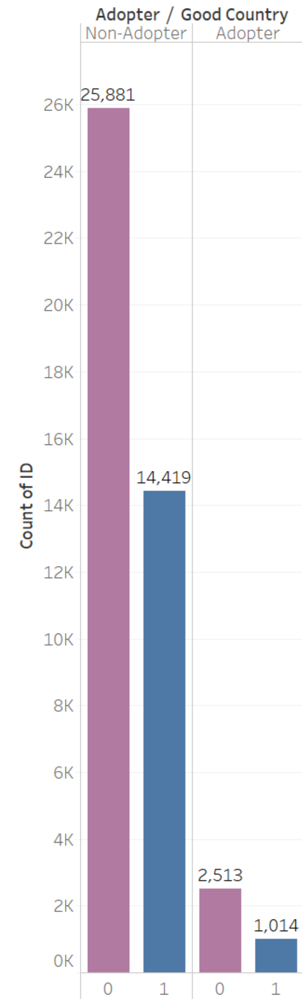
Male



Count of ID for each Male as an attribute broken down by Adopter. Color shows details about Male. The marks are labeled by count of ID.

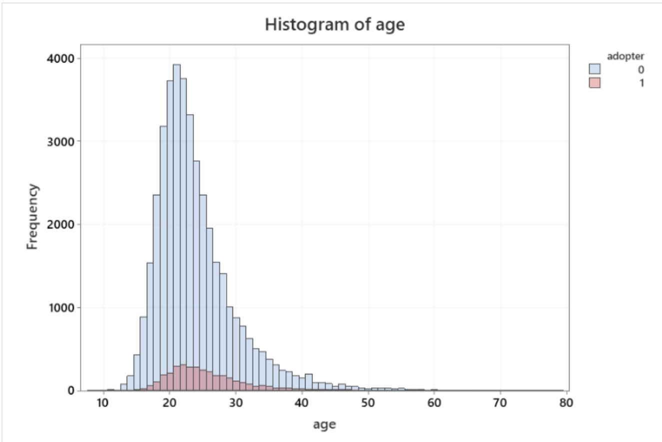
Male
0 1

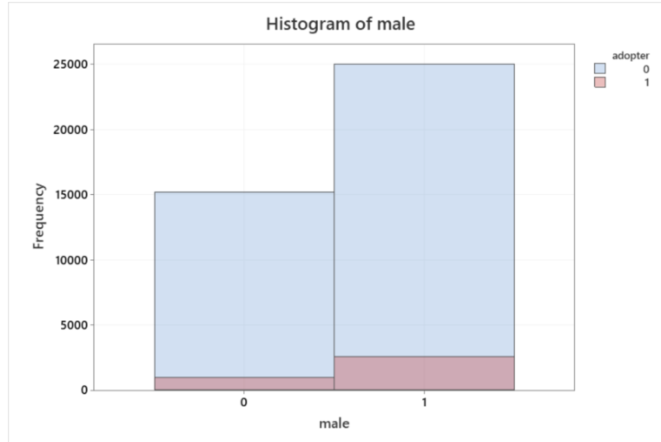
Good Country

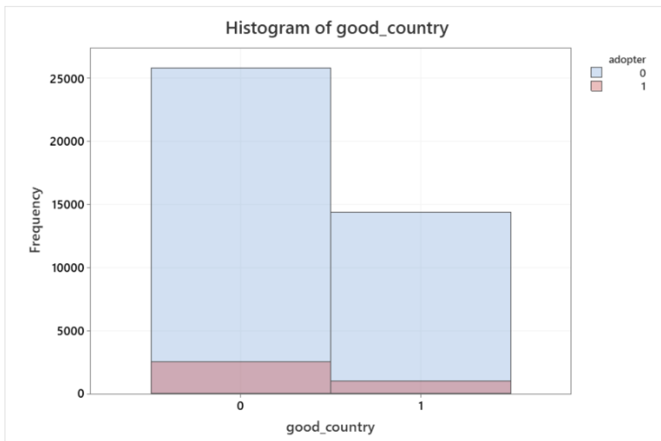


Count of ID for each Good Country broken down by Adopter. Color shows details about Good Country. The marks are labeled by count of ID.

Good Country
- 0
- 1





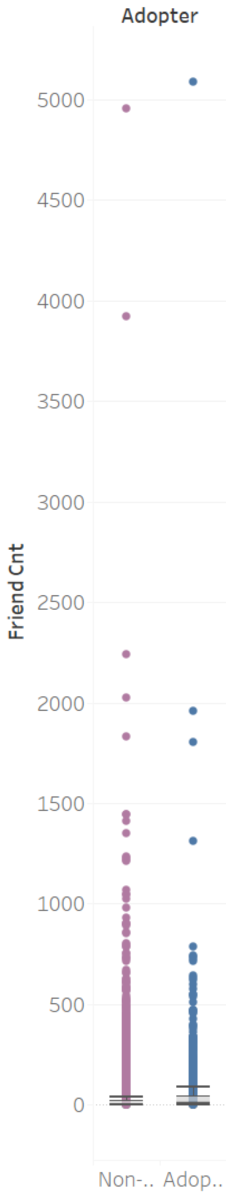


OBSERVATIONS

- Age is highly skewed and has a significant difference between adopters and non-adopters. Adopters are older compared to non-adopters.
- In both the cases the customers are more males, females are significantly lower in numbers.
- US, UK and Germany are smaller markets compared to rest of the world for both adopters and non-adopters.

PEER INFLUENCE

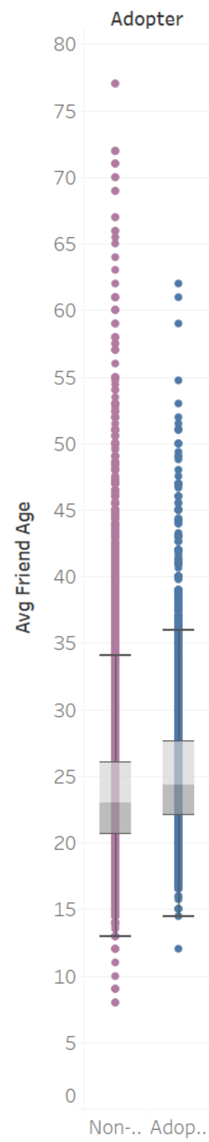
Friend Count



Friend Cnt for each Adopter. Color shows details about Adopter.

Adopter
- Non-Adopter
- Adopter

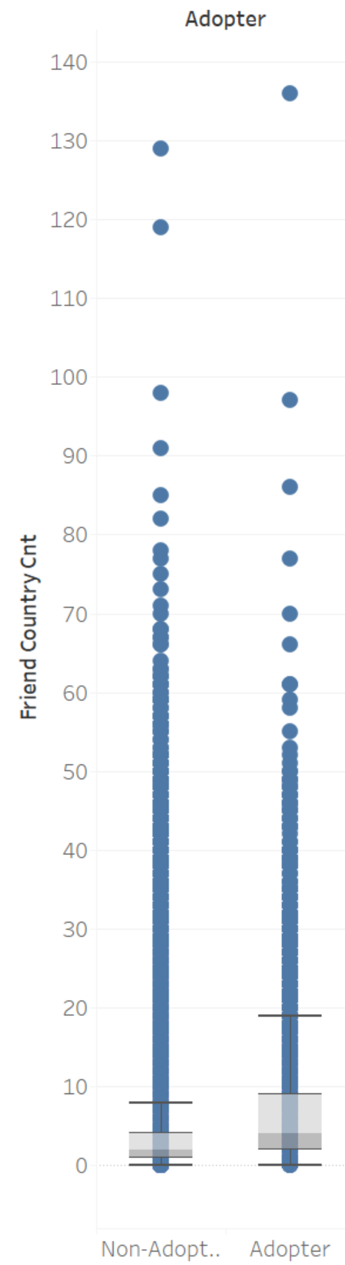
Avg_Friend Age



Avg Friend Age for each Adopter. Color shows details about Adopter.

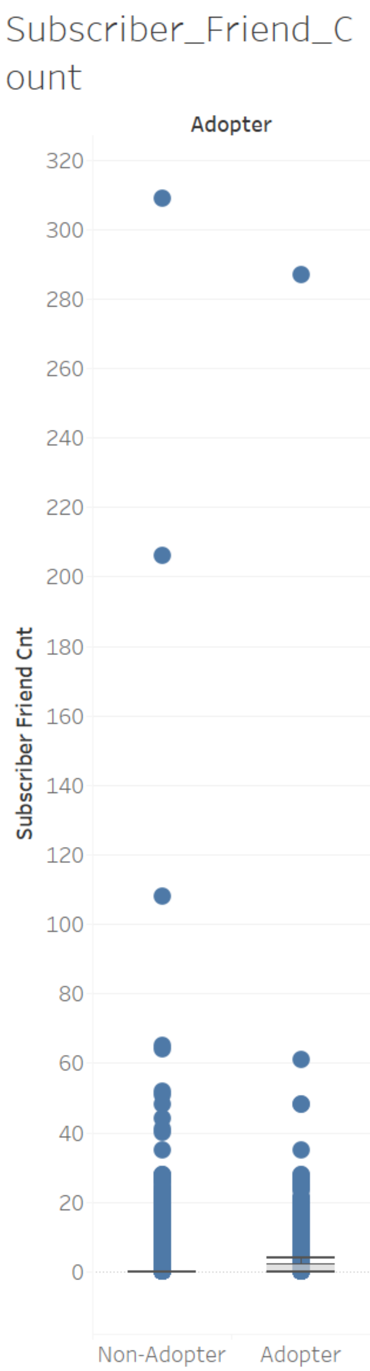
Adopter
- Non-Adopter
- Adopter

Friend_Coun-try_Count



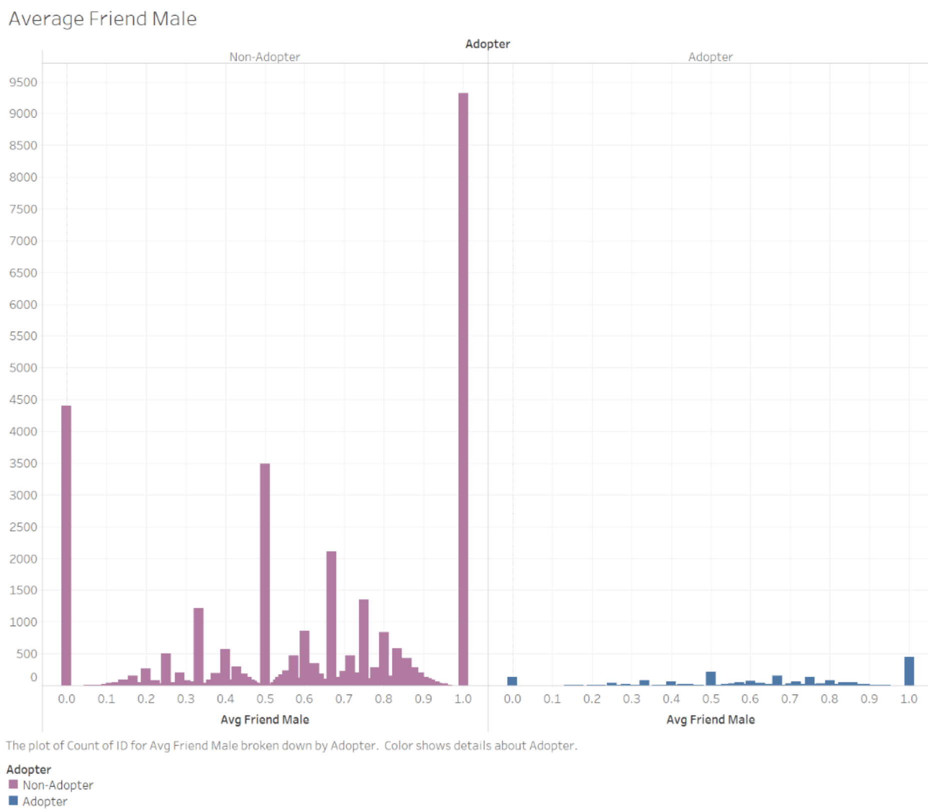
Friend Country Cnt for each Adopter.

PEER INFLUENCE



Subscriber Friend Cnt for each Adopter.

PEER INFLUENCE

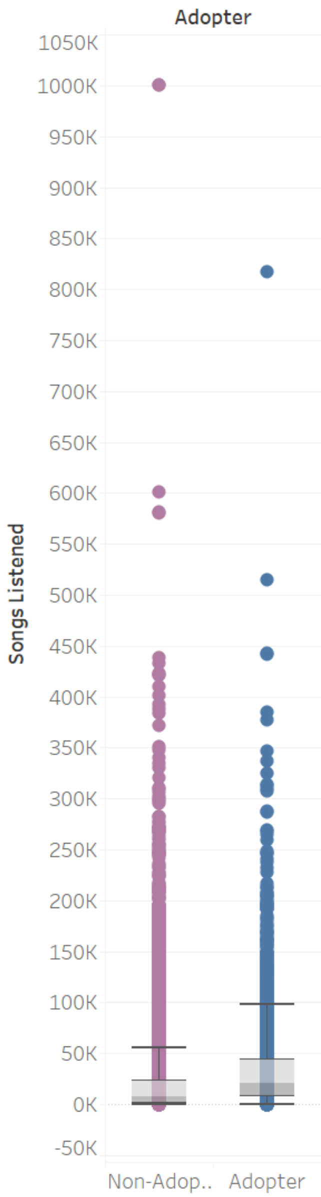


OBSERVATIONS

- Most of the peer influence covariates are higher for premium users or the adopters.
- The average age of adopters is lower compared to non-adopters.
- Adopters friends are more well-travelled almost 2 times of non-adopters.
- Non-adopters have lower friend counts when compared to adopters but have a lot more outliers.
- In both adopters and non-adopters, the males are more dominant as friends.
- Overall, there seems to be a lot of outliers and data seems to be skewed.

USER ENGAGEMENT

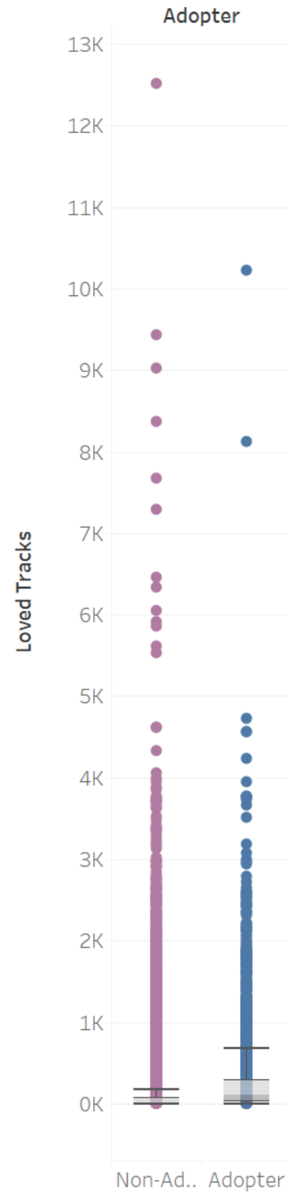
Song Listened



Songs Listened for each Adopter. Color shows details about Adopter.

Adopter
- Non-Adopter
- Adopter

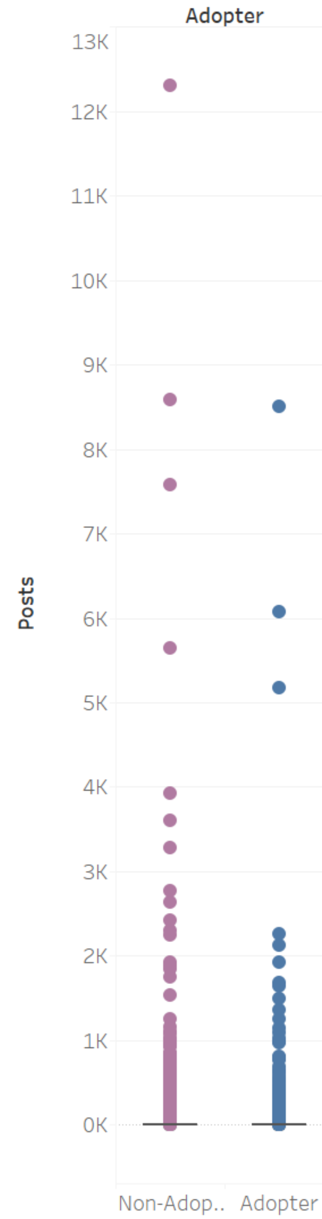
Loved Tracks



Loved Tracks for each Adopter. Color shows details about Adopter.

Adopter
- Non-Adopter
- Adopter

Posts

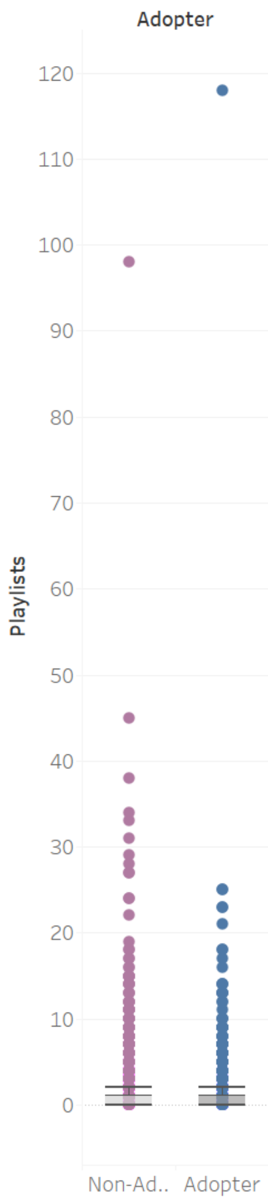


Posts for each Adopter. Color shows details about Adopter.

Adopter
- Non-Adopter
- Adopter

USER ENGAGEMENT

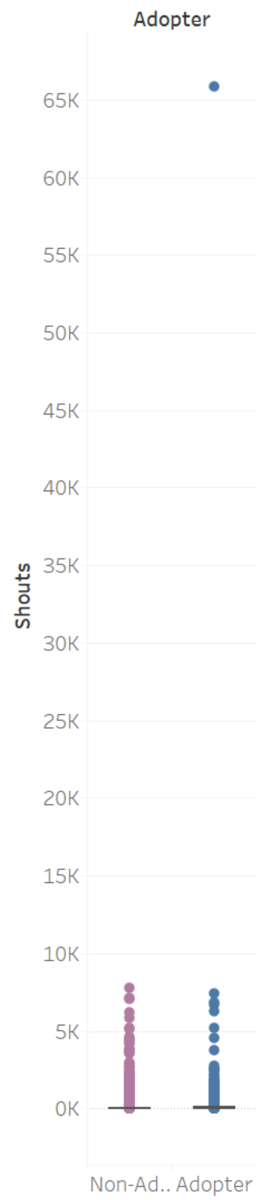
Playlist



Playlists for each Adopter. Color shows details about Adopter.

Adopter
- Non-Adopter
- Adopter

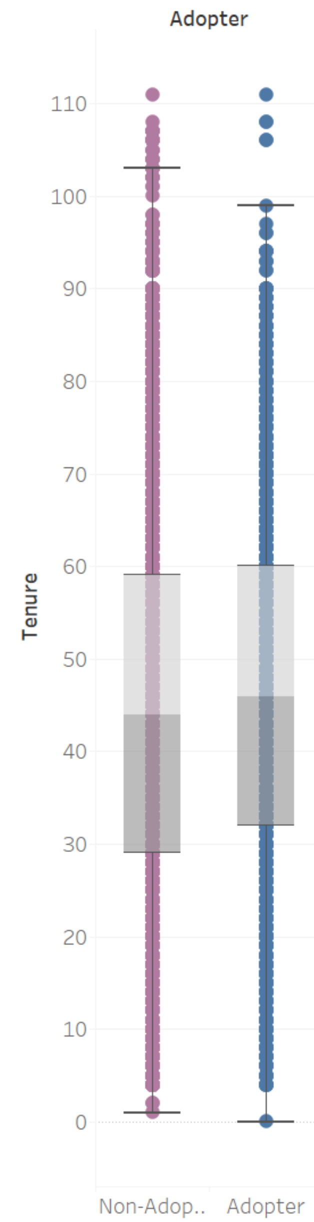
Shouts



Shouts for each Adopter. Color shows details about Adopter.

Adopter
- Non-Adopter
- Adopter

Tenure



Tenure for each Adopter. Color shows details about Adopter.

Adopter
- Non-Adopter
- Adopter

OBSERVATIONS

- Again, is visible that most of the data has outliers and is highly skewed.
- Overall Adopters seems to be more engaged and interacting with the platform features like creating playlists or number of songs listened.
- The song listened is much higher numbers compared to other variables.

T TEST

We start by running a t test to see if the covariates are significant or not in relation to adopters/premium users. We use all the variable to run t test against adopters, as we can see below all the variables are significant and hence are important in determining the subscription for premium services.

```
$age

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -16.996, df = 4079.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.265768 -1.797097
sample estimates:
mean in group 0 mean in group 1 
       23.94844        25.97987 


$male

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -13.654, df = 4295, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.12278707 -0.09195413
sample estimates:
mean in group 0 mean in group 1 
      0.6218610       0.7292316 


$friend_cnt

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -10.646, df = 3675.7, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.15422 -17.32999
sample estimates:
mean in group 0 mean in group 1 
       18.49166        39.73377 
```

```
$friend_cnt

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -10.646, df = 3675.7, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.15422 -17.32999
sample estimates:
mean in group 0 mean in group 1
       18.49166        39.73377


$avg_friend_male

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -4.4426, df = 4591.6, p-value = 9.097e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02883955 -0.01117951
sample estimates:
mean in group 0 mean in group 1
      0.6165888       0.6365983


$avg_friend_age

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -15.658, df = 4140.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.608931 -1.250852
sample estimates:
mean in group 0 mean in group 1
       24.01142        25.44131
```

```
$friend_country_cnt

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -21.267, df = 3791.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.528795 -2.933081
sample estimates:
mean in group 0 mean in group 1
       3.957891        7.188829


$songsListened

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -21.629, df = 3792.7, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17634.24 -14702.96
sample estimates:
mean in group 0 mean in group 1
       17589.44        33758.04


$lovedTracks

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -21.188, df = 3705.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -193.9447 -161.0917
sample estimates:
mean in group 0 mean in group 1
       86.82263       264.34080
```

```
$posts

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -4.2151, df = 3663.5, p-value = 2.557e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -23.30665  -8.50825
sample estimates:
mean in group 0 mean in group 1
       5.293002       21.200454


$playlists

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -8.0816, df = 3634.7, p-value = 8.619e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4367565 -0.2662138
sample estimates:
mean in group 0 mean in group 1
      0.5492804       0.9007655


$shouts

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -3.5659, df = 3536.5, p-value = 0.0003674
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -107.66170  -31.27249
sample estimates:
mean in group 0 mean in group 1
       29.97266        99.43975



$tenure

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -5.0434, df = 4150.6, p-value = 4.768e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.462620 -1.083959
sample estimates:
mean in group 0 mean in group 1
       43.80993        45.58322


$good_country

        Welch Two Sample t-test

data:  x by highnote$adopter
t = 8.8009, df = 4248.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.05463587 0.08595434
sample estimates:
mean in group 0 mean in group 1
      0.3577916       0.2874965


$subscriber_friend_cnt

        Welch Two Sample t-test

data:  x by highnote$adopter
t = -12.287, df = 3632.2, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.413899 -1.024766
sample estimates:
mean in group 0 mean in group 1
       0.417469        1.636802
```

Question 3

“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.”

Based on the above description we divide the subscriber_friend_count into two category 0 & 1. Using this binary variable, we will calculate the propensity score for customers to understand how treatment effect similar group of people and use this information to increase the premium subscribers.

As the data is skewed in many of the variables, I decided to take logs for better estimations and ease of analysis.

PROPENSITY SCORE MODEL

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5576  -0.5682  -0.2997  -0.1170   3.3903

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -17.758481   0.301395 -58.921  < 2e-16 ***
log(age)                      0.647953   0.095402   6.792 1.11e-11 ***
male                          0.078393   0.031386   2.498   0.0125 *
log(friend_cnt + 1)           1.078788   0.027146  39.740  < 2e-16 ***
log(avg_friend_age + 1)       3.486321   0.125283  27.828  < 2e-16 ***
log(avg_friend_male + 1)      0.381212   0.090651   4.205 2.61e-05 ***
log(friend_country_cnt + 1)   0.560058   0.032010  17.496  < 2e-16 ***
log(songsListened + 1)        0.051766   0.009154   5.655 1.56e-08 ***
log(lovedTracks + 1)          0.084539   0.007936  10.652  < 2e-16 ***
log(posts + 1)                0.079169   0.014703   5.385 7.26e-08 ***
log(playlists + 1)           -0.152139   0.035585  -4.275 1.91e-05 ***
log(shouts + 1)              -0.028970   0.013678  -2.118   0.0342 *
log(tenure + 1)              -0.367109   0.031064 -11.818  < 2e-16 ***
good_country                  0.059487   0.030370   1.959   0.0501 .
---
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: 31465  on 43813  degrees of freedom
AIC: 31493

Number of Fisher Scoring iterations: 6
```

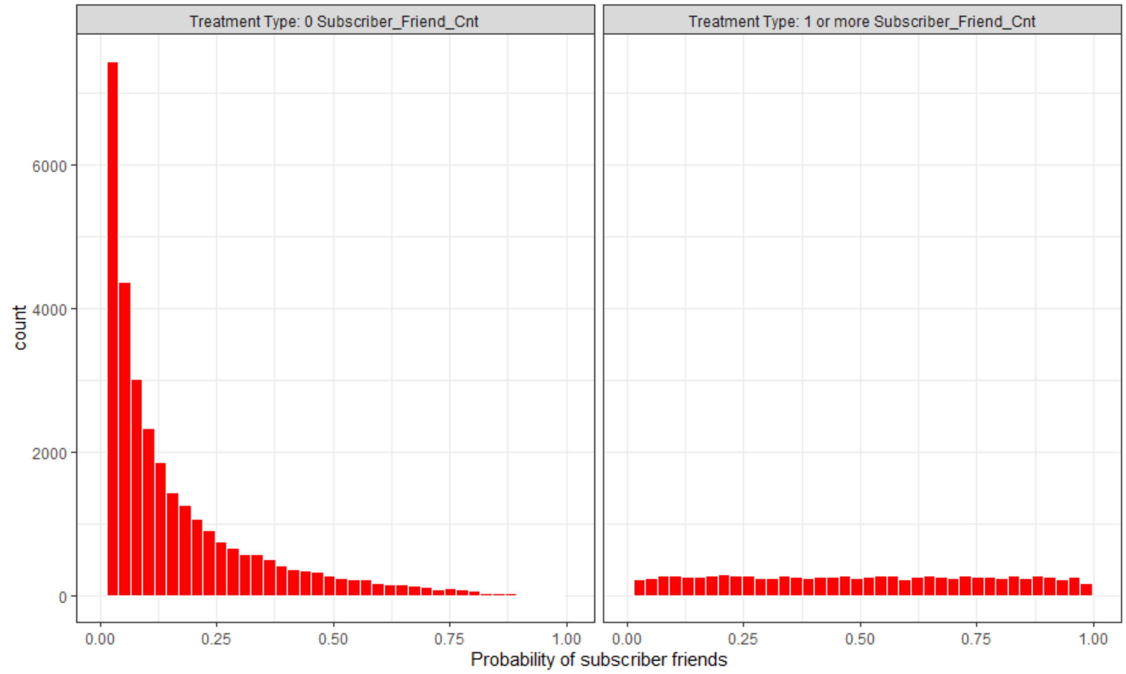
All the variables are highly significant for formulating propensity scores as the P values are < 0.05.

PREDICT MODEL

To find the probability of a customer being treated, we use these propensity scores, calculated for each customers and run a predict function and create a data frame to show the propensity scores as well as the customer’s actual treatment status.

```
    pr_score subscriber_friend_cnt
1 0.07337272                     0
2 0.04897846                     0
3 0.02140195                     0
4 0.42334070                     1
5 0.64329312                     0
6 0.15366186                     0
```

After estimating the propensity score, it is useful to plot histograms of the estimated propensity scores by treatment status:



In the above graph we can see there are customers in the control group of having 0 subscriber friends who are similar to customers in the treatment group. And if we are able to understand these customers, they can be used in treatment too.

MATCHIT

We Used MatchIt function We also used caliper at 0.5. Calipers ensure paired units are close to each other on the calipered covariates, which can ensure good balance in the matched sample.

Below is the number of matched data in treatment groups:

```
Sample Sizes:
          Control Treated
All         34004    9823
Matched      6979    6979
Unmatched   27025    2844
Discarded       0       0
```

After running the MatchIt function we found 6979 customers from the control group matched with other 6979 customers in the treatment group.

The new matched data set has 13958 rows and 18 columns.

```
> dta_m<-match.data(mod_match)
> dim(dta_m)
[1] 13958    18
```

DIFFRENCE IN MEAN

	subscriber_friend_cnt	age	male	friend_cnt	avg_friend_age	avg_friend_male	friend_country_cnt	songsListened	lovedTracks	posts	playlists	shouts	tenure	good_country
1	0	25.07838	0.6400630	25.37828	25.33174	0.6369536	5.730334	26636.51	139.2917	7.051870	0.6251612	43.19602	45.85858	0.3440321
2	1	24.92291	0.6436452	25.20089	25.17927	0.6328983	5.614845	27911.12	147.5908	8.551942	0.6447915	45.22998	46.10890	0.3437455

We are trying to understand how the variable are means different among these two matched groups. As we ca see the matched groups are pretty similar in most of their attributes. A major differentiator though is the song listened.

MATCHED T TEST

Let's know analyze if the mean difference are significant or not. We essentially want similar group of people are reduce the difference in mean as much as possible.

```
[[1]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = 1.361, df = 13785, p-value = 0.1735
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06843961  0.37937240
sample estimates:
mean in group 0 mean in group 1
       25.07838        24.92291


[[2]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = -0.44132, df = 13956, p-value = 0.659
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01949255  0.01232820
sample estimates:
mean in group 0 mean in group 1
      0.6400630       0.6436452


[[3]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = 0.41217, df = 13951, p-value = 0.6802
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6662162  1.0209948
sample estimates:
mean in group 0 mean in group 1
       25.37828        25.20089


[[4]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = 1.6292, df = 13516, p-value = 0.1033
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03097078  0.33590408
sample estimates:
mean in group 0 mean in group 1
       25.33174        25.17927
```

```
[[5]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = 0.99228, df = 13955, p-value = 0.3211
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.003955457  0.012066020
sample estimates:
mean in group 0 mean in group 1
      0.6369536       0.6328983


[[6]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = 1.3985, df = 13900, p-value = 0.162
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04637843  0.27735708
sample estimates:
mean in group 0 mean in group 1
       5.730334        5.614845


[[7]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = -2.2818, df = 13776, p-value = 0.02252
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2369.5484  -179.6716
sample estimates:
mean in group 0 mean in group 1
       26636.51        27911.12


[[8]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = -1.58, df = 13833, p-value = 0.1141
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.594897   1.996817
sample estimates:
mean in group 0 mean in group 1
       139.2917        147.5908
```

```
[[9]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = -0.93434, df = 10832, p-value = 0.3502
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.647126  1.646983
sample estimates:
mean in group 0 mean in group 1
       7.051870        8.551942


[[10]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = -1.0985, df = 13632, p-value = 0.272
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05465835  0.01539771
sample estimates:
mean in group 0 mean in group 1
      0.6251612       0.6447915


[[11]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = -0.77749, df = 13954, p-value = 0.4369
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.161777  3.093859
sample estimates:
mean in group 0 mean in group 1
       43.19602        45.22998


[[12]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = -0.74397, df = 13956, p-value = 0.4569
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9098423  0.4091976
sample estimates:
mean in group 0 mean in group 1
       45.85858        46.10890


[[13]]

        Welch Two Sample t-test

data:  dta_m[, v] by dta_m$subscriber_friend_cnt
t = 0.035636, df = 13956, p-value = 0.9716
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01547621  0.01604936
sample estimates:
mean in group 0 mean in group 1
      0.3440321       0.3437455
```

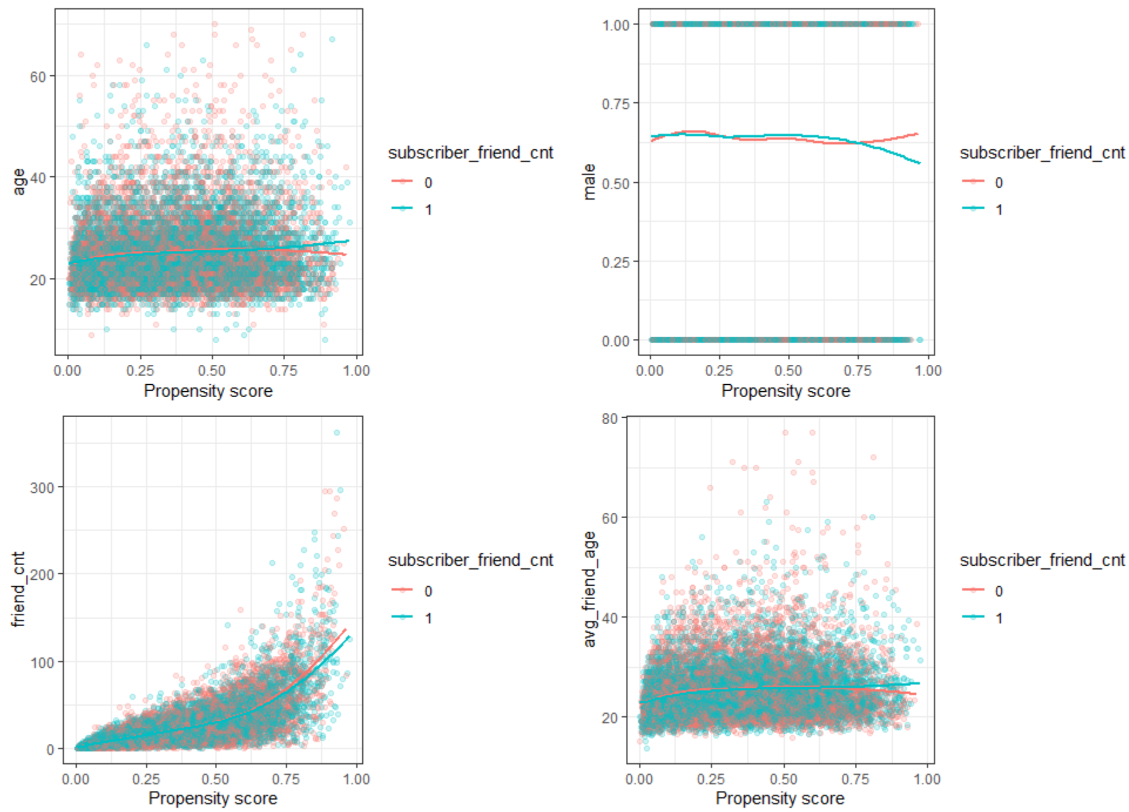
All the variables have no significant difference in their means except the song listened. This makes us agree that we cannot reject the null hypothesis. And that the matched data in treatment and control groups are similar.

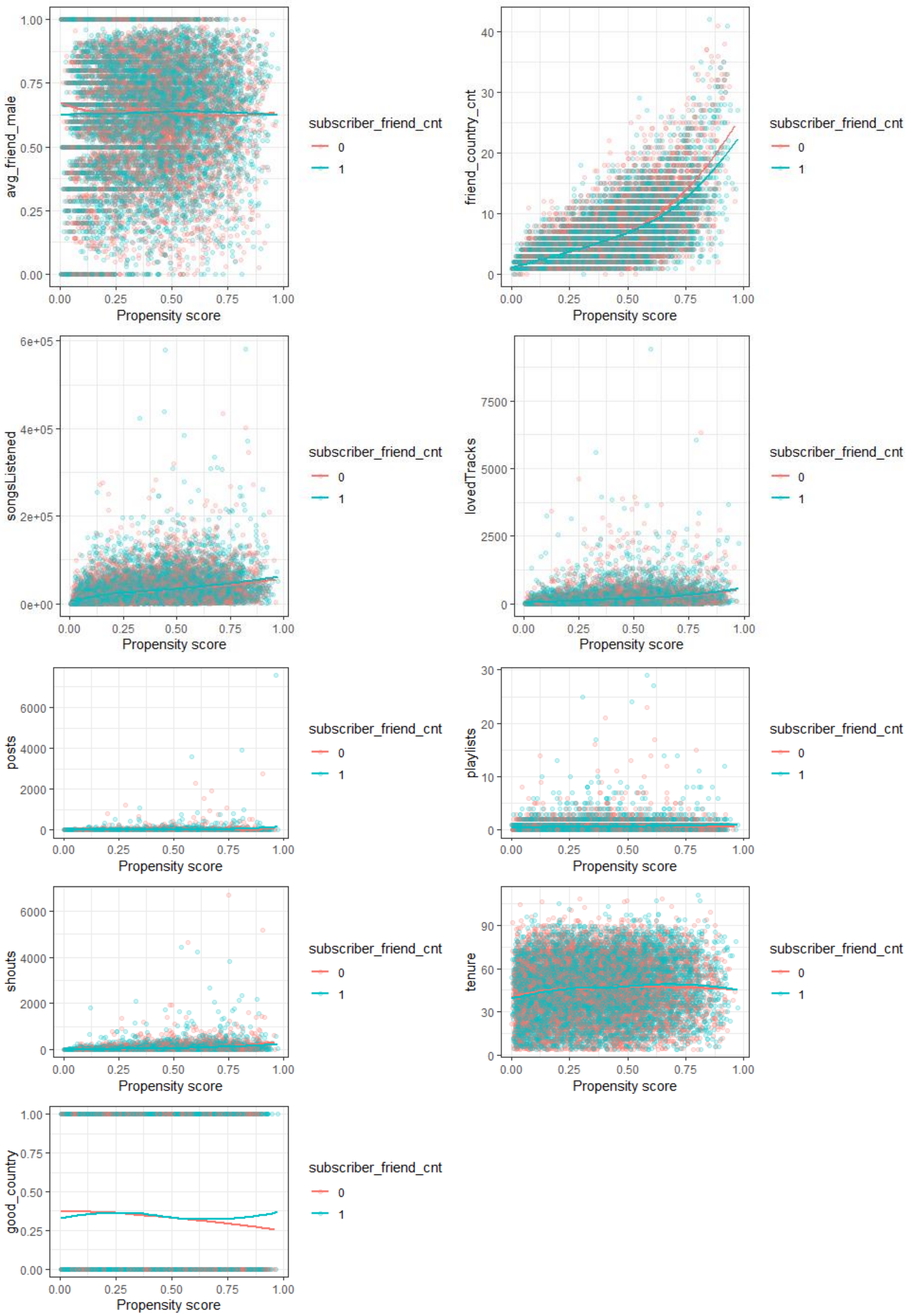
VIZUALIZATION MATCHED DATA

VISUAL INSPECTION

["It is useful to plot the mean of each covariate against the estimated propensity score, separately by treatment status. If matching is done well, the treatment and control groups will have (near) identical means of each covariate at each value of the propensity score."]

Below is an example using the all the covariates in our model.




avg_friend_male
Propensity score
subscriber_friend_cnt
0
1
friend_country_cnt
Propensity score
subscriber_friend_cnt
0
1
songsListened
Propensity score
subscriber_friend_cnt
0
1
lovedTracks
Propensity score
subscriber_friend_cnt
0
1
posts
Propensity score
subscriber_friend_cnt
0
1
playlists
Propensity score
subscriber_friend_cnt
0
1
shouts
Propensity score
subscriber_friend_cnt
0
1
tenure
Propensity score
subscriber_friend_cnt
0
1
good_country
Propensity score
subscriber_friend_cnt
0
1

As we can see the propensity scores of treatment groups are almost matching for all variables, except few like friend_country_cnt . Overall, I am happy with the model and believe this subgroup of matched customers can be a good set to understand the effect of treatment.

Question 4

The next step is to run regression on these matched customers to see which variables are significant in explaining if a customer can become a premium customer.

MODEL1

```
Call:
glm(formula = adopter ~ subscriber_friend_cnt, family = binomial(),
    data = Q4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5705  -0.5705  -0.4493  -0.4493   2.1649

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.24245    0.04063 -55.195   <2e-16 ***
subscriber_friend_cnt  0.50915    0.05266   9.668   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> cbind(exp(coef(Model1)))
                           [,1]
(Intercept)           0.1061975
subscriber_friend_cnt 1.6638688
```

I ran a GLM model with just treatment of having subscriber friend or not. It is very clear that having friends who are also subscribers relate to premium users significantly. With a one percent increase in the subscriber_friend_cnt, the odds of increase in adopters is 0.50915 or approximately 0.51%.

We build another model by including all the covariates in the model and see how the effect the prediction of a free or premium user.

MODEL 2 – ALL VARIABLES AND MATCHED DATA

```
Call:
glm(formula = adopter ~ log(age) + male + log(friend_cnt + 1) +
    log(avg_friend_age + 1) + log(avg_friend_male + 1) + log(friend_country_cnt +
    1) + log(songsListened + 1) + log(lovedTracks + 1) + log(posts +
    1) + log(playlists + 1) + log(shouts + 1) + log(tenure +
    1) + good_country + subscriber_friend_cnt, family = binomial(),
    data = Q4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3894  -0.5578  -0.4207  -0.2880   2.9670

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -9.19019    0.62127 -14.793  < 2e-16 ***
log(age)                     0.57341    0.17737   3.233  0.00123 **
male                         0.33556    0.06197   5.415 6.13e-08 ***
log(friend_cnt + 1)          0.09797    0.05242   1.869  0.06163 .
log(avg_friend_age + 1)      1.02219    0.23974   4.264 2.01e-05 ***
log(avg_friend_male + 1)     0.09275    0.18484   0.502  0.61581
log(friend_country_cnt + 1) -0.02112    0.05909  -0.357  0.72080
log(songsListened + 1)       0.21267    0.02338   9.096  < 2e-16 ***
log(lovedTracks + 1)         0.25163    0.01629  15.447  < 2e-16 ***
log(posts + 1)               0.13789    0.02435   5.662 1.49e-08 ***
log(playlists + 1)           0.20234    0.06214   3.256  0.00113 **
log(shouts + 1)             -0.13771    0.02443  -5.637 1.73e-08 ***
log(tenure + 1)             -0.39152    0.05801  -6.749 1.49e-11 ***
good_country                -0.50021    0.06030  -8.296  < 2e-16 ***
subscriber_friend_cnt        0.53459    0.05427   9.850  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> cbind(exp(coef(Model2)))
                                    [,1]
(Intercept)                 0.0001020358
log(age)                    1.7743131450
male                        1.3987185294
log(friend_cnt + 1)         1.1029302380
log(avg_friend_age + 1)     2.7792616882
log(avg_friend_male + 1)    1.0971913491
log(friend_country_cnt + 1) 0.9791024600
log(songsListened + 1)      1.2369810703
log(lovedTracks + 1)        1.2861145844
log(posts + 1)              1.1478500137
log(playlists + 1)          1.2242593024
log(shouts + 1)             0.8713527305
log(tenure + 1)             0.6760272450
good_country                0.6064010791
subscriber_friend_cnt       1.7067551287
```

Average friend who is male and how many different counties friends are from are not significant in finding if a user is going to be a free or premium customer. Friend counts is marginally significant, but for a business term 90% confidence level is a good significance. Engagement metrics seems to have significant effect on adopters and non-adopters decisions. Like playlist can have odd of 0.20% increase in odds of a customer becoming premium customer.

MODEL 3 – ALL VARIABLES AND INITIAL DATA

I used all the variables to train the model on the original data set. The reason I have used all the variables is to understand if the effects change on the original dataset.

```
Call:
glm(formula = adopter ~ log(age) + male + log(friend_cnt + 1) +
    log(avg_friend_age + 1) + log(avg_friend_male + 1) + log(friend_country_cnt +
    1) + log(songsListened + 1) + log(lovedTracks + 1) + log(posts +
    1) + log(playlists + 1) + log(shouts + 1) + log(tenure +
    1) + good_country + subscriber_friend_cnt, family = binomial(),
    data = highnote)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4122  -0.4357  -0.2894  -0.1774   3.3041

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -11.056173   0.366351 -30.179  < 2e-16 ***
log(age)                  0.951769   0.117803   8.079 6.51e-16 ***
male                      0.330428   0.043097   7.667 1.76e-14 ***
log(friend_cnt + 1)       0.184890   0.034273   5.395 6.87e-08 ***
log(avg_friend_age + 1)   1.041240   0.157305   6.619 3.61e-11 ***
log(avg_friend_male + 1)  0.185774   0.110910   1.675 0.093934 .
log(friend_country_cnt + 1) 0.052172 0.043470  1.200 0.230074
log(songsListened + 1)    0.213483   0.014522  14.701  < 2e-16 ***
log(lovedTracks + 1)      0.302947   0.011421  26.527  < 2e-16 ***
log(posts + 1)            0.136825   0.017245   7.934 2.12e-15 ***
log(playlists + 1)        0.152221   0.042715   3.564 0.000366 ***
log(shouts + 1)          -0.126367   0.017482  -7.228 4.89e-13 ***
log(tenure + 1)          -0.371831   0.039484  -9.417  < 2e-16 ***
good_country             -0.458880   0.041554 -11.043  < 2e-16 ***
subscriber_friend_cnt     0.023857   0.007711   3.094 0.001974 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                                    [,1]
(Intercept)                 1.578937e-05
log(age)                    2.590287e+00
male                        1.391563e+00
log(friend_cnt + 1)         1.203086e+00
log(avg_friend_age + 1)     2.832729e+00
log(avg_friend_male + 1)    1.204150e+00
log(friend_country_cnt + 1) 1.053556e+00
log(songsListened + 1)      1.237983e+00
log(lovedTracks + 1)        1.353843e+00
log(posts + 1)              1.146628e+00
log(playlists + 1)          1.164417e+00
log(shouts + 1)             8.812911e-01
log(tenure + 1)             6.894709e-01
good_country                6.319910e-01
subscriber_friend_cnt       1.024144e+00
```

CONCLUSION

Both the models have near similar results. I would like to conclude by saying the below;

- Age seems to have a real good effect on customer converting to paid user. With 1 year increase in average age of the customers, we can have odds of 0.95% to become a paid customer, increasing our revenues by 24 times.
- Song listened and love track also have positive effect on pushing customers from "free to fee" customer stage.

- Overall Demographic, Peer Influence and User Engagement all have positive effect on the likelihood of a non-adopter becoming an adopter of Highnote's premium services.
- Shouts, tenure and good_country don't really seem to be good indicator to pursue customers on, as they have negative effect on the likelihood of the conversion.

WHAT CAN HIGHNOTE DO TO INCREASE "FREE TO FEE" CUSTOMER CONVERSIONS?

- The good country has negative effect on the likelihood of adoption, US, UK and Germany have many other players. And its hard to get a share of the pie. A good ide will be to have a stronger global presence and find markets where High note can become unique provider.
- Age seems to have the most effect. Income levels also are highly correlated to the age and one way to look at it may have corporate tie-ups, people who are working and have disposal income are more likely to adopt and the age factor can be utilized this way.
- Develop community type engagement even further. Engagement metrics have a positive effect on the adoption of paid services. May be develop a weekly leader board to increase engagement.
- Incentivize people to have more friends on the map. Make tier account based on how many people a customer has a friend, introduce badges that customers can brag about (Similar to elite status in Yelp).

BIBLIOGRAPHY AND RESOURCES

https://cran.r-project.org/web/packages/MatchIt/vignettes/matching-methods.html

https://www.rdocumentation.org/packages/optmatch/versions/0.9-13/topics/caliper

https://sejdemyr.github.io/r-tutorials/statistics/tutorial8.html