SWATHI MUDHELLI MIDTERM

Q1) Summary statistics: Generate descriptive statistics for the key variables in the data set, similar to the table on the last page of the case. (Note that your table will look different because the data set you are analyzing is different from the one used to generate the table in the case.) Analyze the differences in the mean values of the variables, comparing the adopter and non-adapter subsamples. What tentative conclusions can you draw from these comparisons?

Adopter (Premium Users)													
Type	Covariate	n	Mean	SD	Median	Trimmed	Mad	Min	Max	Range	Skeness	Kurtosis	se
Social	friend_cnt	3527	39.73	117.27	16	23.69	17.79	1	5089	5088	26.04	1013.79	1.97
	avg_friend_age	3527	25.44	5.21	24.36	24.83	3.91	12	62	50	1.68	5.05	0.09
	avg_friend_male	3527	0.64	0.25	0.67	0.65	0.25	0	1	1	-0.54	-0.05	0
	friend_country_cnt	3527	7.19	8.86	4	5.36	4.45	0	136	136	3.61	24.53	0.15
	subscriber_friend_cnt	3527	1.64	5.85	0	0.84	0	0	287	287	34.05	1609.52	0.1
Engagement	songsListened	3527	33758.04	43592.73	20908	25811.69	23276.82	0	817290	817290	4.71	46.64	734.03
	lovedTracks	3527	264.34	491.43	108	161.68	140.85	0	10220	10220	6.52	80.96	8.27
	posts	3527	21.2	221.99	0	1.44	0	0	8506	8506	26.52	852.38	3.74
	playlists	3527	0.9	2.56	1	0.59	1.48	0	118	118	28.84	1244.31	0.04
	shouts	3527	99.44	1156.07	9	23.89	11.86	0	65872	65872	52.52	2969.09	19.47
Demographics	age	3527	25.98	6.84	24	25.05	4.45	8	73	65	1.68	4.39	0.12
	male	3527	0.73	0.44	1	0.79	0	0	1	1	-1.03	-0.94	0.01
	tenure	3527	45.58	20.04	46	45.6	20.76	0	111	111	0.02	-0.62	0.34
	good_country	3527	0.29	0.45	0	0.23	0	0	1	1	0.94	-1.12	0.01

Non Adopter (Free Users)													
Type	Covariate	n	Mean	SD	Median	Trimmed	Mad	Min	Max	Range	Skeness	Kurtosis	se
Social	friend_cnt	40300	18.49	57.48	7	10.28	7.41	1	4957	4956	32.67	2087.42	0.29
	avg_friend_age	40300	24.01	5.1	23	23.4	3.95	8	77	69	1.84	7.15	0.03
	avg_friend_male	40300	0.62	0.32	0.67	0.65	0.35	0	1	1	-0.52	-0.72	0
	friend_country_cnt	40300	3.96	5.76	2	2.66	1.48	0	129	129	4.74	38.29	0.03
	subscriber_friend_cnt	40300	0.42	2.42	0	0.13	0	0	309	309	72.19	8024.62	0.01
Engagement	songsListened	40300	17589.44	28416.02	7440	11817.64	10576.87	0	1000000	1000000	6.05	105.85	141.55
	lovedTracks	40300	86.82	263.58	14	36.35	20.76	0	12522	12522	13.12	335.93	1.31
	posts	40300	5.29	104.31	0	0.23	0	0	12309	12309	73.92	7005.34	0.52
	playlists	40300	0.55	1.07	0	0.45	0	0	98	98	28.21	1945.28	0.01
	shouts	40300	29.97	150.69	4	8.84	4.45	0	7736	7736	22.53	779.12	0.75
Demographics	age	40300	23.95	6.37	23	23.09	4.45	8	79	71	1.97	6.8	0.03
	male	40300	0.62	0.48	1	0.65	0	0	1	1	-0.5	-1.75	0
	tenure	40300	43.81	19.79	44	43.72	22.24	1	111	110	0.05	-0.7	0.1
	good_country	40300	0.36	0.48	0	0.32	0	0	1	1	0.59	-1.65	0

I categorized the variables into “Social”, “Engagement” and “Demographics” information. From the summary statistics tables, I observed that the means of the two groups (Free users, Premium users) differ significantly.

Among the demographic data, average age of adopters is 25.98, whereas the average age of non-adopters is 23.95. The average tenure of adopters is 45.58 months compared to 43.81 months of non-adapters, indicating that adopters spend more time on the platform before becoming premium subscribers.

Within the engagement metrics, adopters listened to an average of 33000 songs compared to 18000 songs that non-adopters listened to. Adopters loved 264 songs whereas non-adopters loved only 86 songs. Similarly, premium users are highly engaged in creating playlists, posts and shouts when compared to free users. This indicates that premium users are more engaged and consume content differently when compared to free users.

Looking at the social metrics, premium users have higher averages of friends - 39.7 premium users vs 18.9 for free users. Premium users also have 1.64 subscriber friends on an average, when compared to 0.42 of free users. Premium users also have more geographically spread-out friends (based on friend country count) when compared to free users. This indicates that social features have a strong relationship with subscription behavior. We can analyze and validate these relationships further after performing a regression analysis. Regression will help in identifying significant variables and their impact on subscription decision.

I also performed a t-test to understand if the difference in means is statistically significant. Here are the results from the t-test.

$age :Welch Two Sample t-test
data: i by highnote$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

$male :Welch Two Sample t-test
data: i by highnote$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

$friend_cnt : Welch Two Sample t-test
data: i by highnote$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

$avg_friend_male :Welch Two Sample t-test
data: i by highnote$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

$avg_friend_age: Welch Two Sample t-test
data: i by highnote$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

$friend_country_cnt: Welch Two Sample t-test
data: i by highnote$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

$songsListened: Welch Two Sample t-test
data: i by highnote$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

$lovedTracks: Welch Two Sample t-test
data: i by highnote$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

$posts: Welch Two Sample t-test
data: i by highnote$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

$playlists: Welch Two Sample t-test
data: i by highnote$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

$shouts: Welch Two Sample t-test
data: i by highnote$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

$tenure: Welch Two Sample t-test
data: i by highnote$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

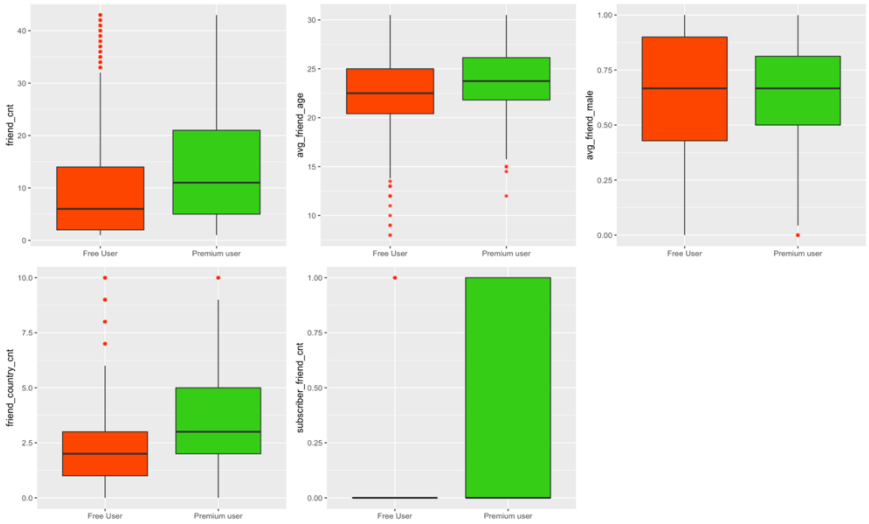
43.80993 45.58322
$good_country: Welch Two Sample t-test
data: i by highnote$adopter
t = 8.8009, df = 4248.5, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.05463587 0.08595434
sample estimates:
mean in group 0 mean in group 1
0.3577916 0.2874965

$subscriber_friend_cnt: Welch Two Sample t-test
data: i by highnote$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

All the t-tests (adopter to each variable) have extremely significant p-values, indicating that the difference in means is significant for free and premium users.

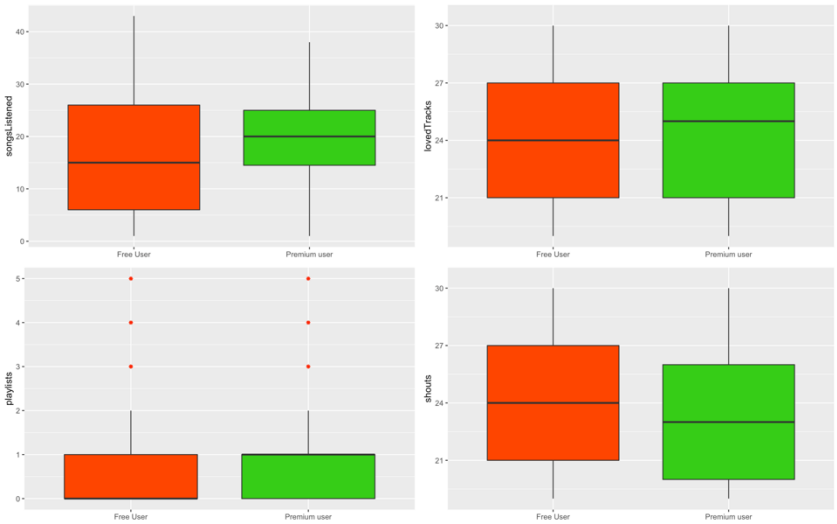
Q2. Data Visualization: Generate a set of charts (e.g., scatter plots, box plots, etc) to help visualize how adopters and non-adopters (of the premium subscription service) differ from each other in terms of (i) demographics, (ii) peer influence, and (iii) user engagement. What can you conclude from your charts?

1) Box plots based on Social / peer influence data



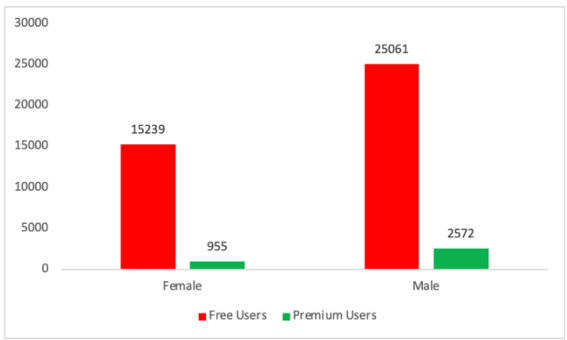
From the charts, premium users have many more subscriber friends when compared to free users. They also have more friends (median > 10) when compared to free users (median<10). Premium users also have friends who are geographically more spread out when compared to free users. This can be a valuable variable if the platform is looking to maximize reach based on sphere of influence.

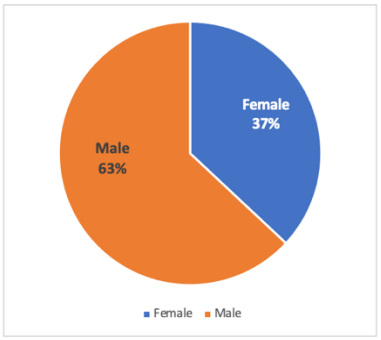
2) Box plots based on engagement



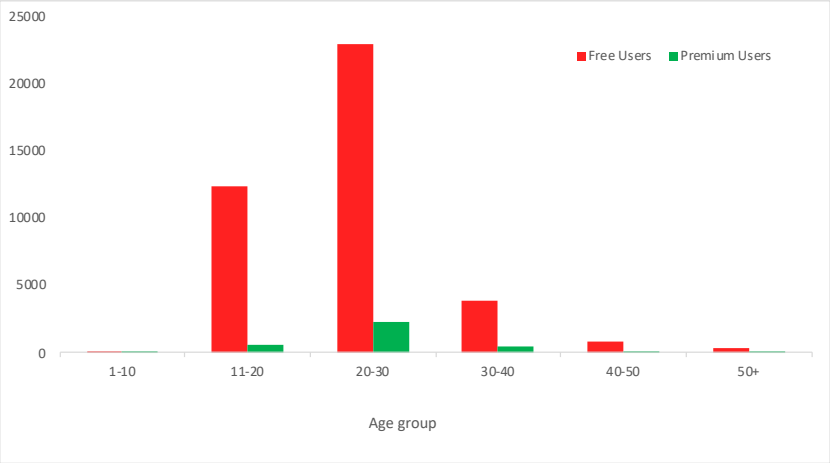
From the charts above, premium users have higher median engagement in terms of songs listened, loved tracks and playlists. However, free users engage more in shouts than premium users.

3) Demographic distribution





From the graph above, most users are male. Of the male users, only ~10% are premium users, whereas less than 10% of female users are premium users.



This graph shows the distribution of free and premium users by age group. Most users are in the “11-20” & “20-30” year age group. Premium users are mostly in the “20-30” age group. This could be due to the fact that “20-30” age group are young workers with not a lot of financial commitments and are open to exploring good products. Highnote can focus on the late teens, as these users can potentially become active users as they age over 20 and start full time jobs (avg tenure for adoption is 3+ years, so targeting male users who 17-20, might be a good place to start)

Q3) Propensity Score Matching (PSM): You will use PSM 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 >= 1), while the "control" group will include users with zero subscriber friends. Use PSM to first create matched treatment and control samples, then test whether there is a significant average treatment effect. Provide an interpretation of your results.

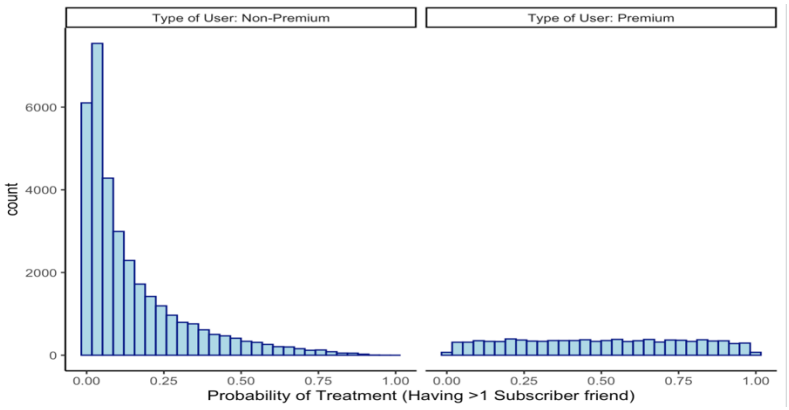
```
m_ps <- glm(treatment ~ age + male + good_country +
              friend_cnt + avg_friend_age + avg_friend_male + friend_country_cnt +
              songsListened + lovedTracks + posts + playlists + shouts + tenure,
            family = binomial(), data = highnote2)
summary(m_ps)
```

Before propensity score matching, I split users into treatment group (>=1 subscriber friend) and control group (0 subscriber friends). Then, I performed a t-test to check if treatment group and control group have significant difference in means. It ended up getting the p-value with statistically significant. So, two groups are different.

Then I performed logistic regression to make sure all the other variables are significant to control group and treatment group. The distributions of all continuous variables are skewed. Thus, I log transformed the variables before logistic regression. The result shows that, based on a significance level (alpha) of 5%, most variables are significant.

```
Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -17.865849   0.302503 -59.060  < 2e-16 ***
log(age + 1)                   0.666546   0.099045   6.730 1.70e-11 ***
male                           0.078832   0.031385   2.512   0.0120 *
good_country                   0.059455   0.030370   1.958   0.0503 .
log(friend_cnt + 1)            1.078821   0.027146  39.741  < 2e-16 ***
log(avg_friend_age + 1)        3.491597   0.125281  27.870  < 2e-16 ***
log(avg_friend_male + 1)       0.380822   0.090651   4.201 2.66e-05 ***
log(friend_country_cnt + 1)    0.559894   0.032010  17.491  < 2e-16 ***
log(songsListened + 1)         0.051806   0.009154   5.659 1.52e-08 ***
log(lovedTracks + 1)           0.084547   0.007936  10.653  < 2e-16 ***
log(posts + 1)                 0.079123   0.014703   5.382 7.39e-08 ***
log(playlists + 1)            -0.152010   0.035584  -4.272 1.94e-05 ***
log(shouts + 1)               -0.029056   0.013678  -2.124   0.0336 *
log(tenure + 1)               -0.366258   0.031053 -11.795  < 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: 46640  on 43826  degrees of freedom
Residual deviance: 31466  on 43813  degrees of freedom
AIC: 31494

Number of Fisher Scoring iterations: 6
```

I then used this model to calculate propensity scores for each user. Plotting the propensity scores of two groups.



The graph indicates that most free (non-premium) users had almost 0 probability of having more than 1 subscriber friend. Its a right skewed distribution, indicating that while

most users had 0 probability of having more than 1 subscriber friend, some users could have 1 or more subscriber friends, because they show higher probability. The histogram of premium users is much more balanced and indicates that most users have a 10-100% probability of having at least 1 subscriber friend.

PROPENSITY SCORE MATCHING:

Then I performed propensity score matching using “matchit” function with “nearest” method. This method matches each user from control group to an identical user in treatment group based on propensity score. The propensity score is in the distance column after matching. According to the propensity score, these subjects are similar. Matching on this distance metric helps ensure the treatment and control groups have similar covariate distributions. I also used a caliper of 0.01 to set a limit or boundary, within which the match has to be made. A 0.01 caliper indicates that the match has to be found within 0.01 standard deviation. The results show that 6885 users have been matched in both groups.

```
Sample Sizes:
          Control Treated
All         34004    9823
Matched      6885    6885
Unmatched   27119    2938
Discarded       0       0
```

Then I tested I for difference in means of variables for these two samples (control and treatment groups). The t-test results show p-values of >5% for all variables. At a significance level of 5%, these p-values are not significant (highly insignificant). Hence, both the groups are alike.

```
Age:    Welch Two Sample t-test
data:  matched_data[, v] by matched_data$treatment
t = 1.6099, df = 13569, p-value = 0.1075
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04010261  0.40873006
sample estimates:
mean in group 0 mean in group 1
      25.06202        24.87771
```

```
Male: Welch Two Sample t-test
data:  matched_data[, v] by matched_data$treatment
```

t = -0.10656, df = 13768, p-value = 0.9151
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01690189 0.01515897
sample estimates:
mean in group 0 mean in group 1
      0.6402324   0.6411038

Good country: Welch Two Sample t-test
data: matched_data[, v] by matched_data$treatment
t = -0.3582, df = 13768, p-value = 0.7202
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01880067 0.01299094
sample estimates:
mean in group 0 mean in group 1
      0.3448076   0.3477124

Friend count: Welch Two Sample t-test
data: matched_data[, v] by matched_data$treatment
t = 1.5928, df = 13730, p-value = 0.1112
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1573843 1.5223807
sample estimates:
mean in group 0 mean in group 1
       25.38054   24.69804

Avg friend age: Welch Two Sample t-test
data: matched_data[, v] by matched_data$treatment
t = 1.8793, df = 13265, p-value = 0.06023
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.007578479 0.359739316
sample estimates:
mean in group 0 mean in group 1
       25.31578   25.13970

Avg friend male: Welch Two Sample t-test
data: matched_data[, v] by matched_data$treatment
t = 0.54738, df = 13766, p-value = 0.5841
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.005820037 0.010330061
sample estimates:

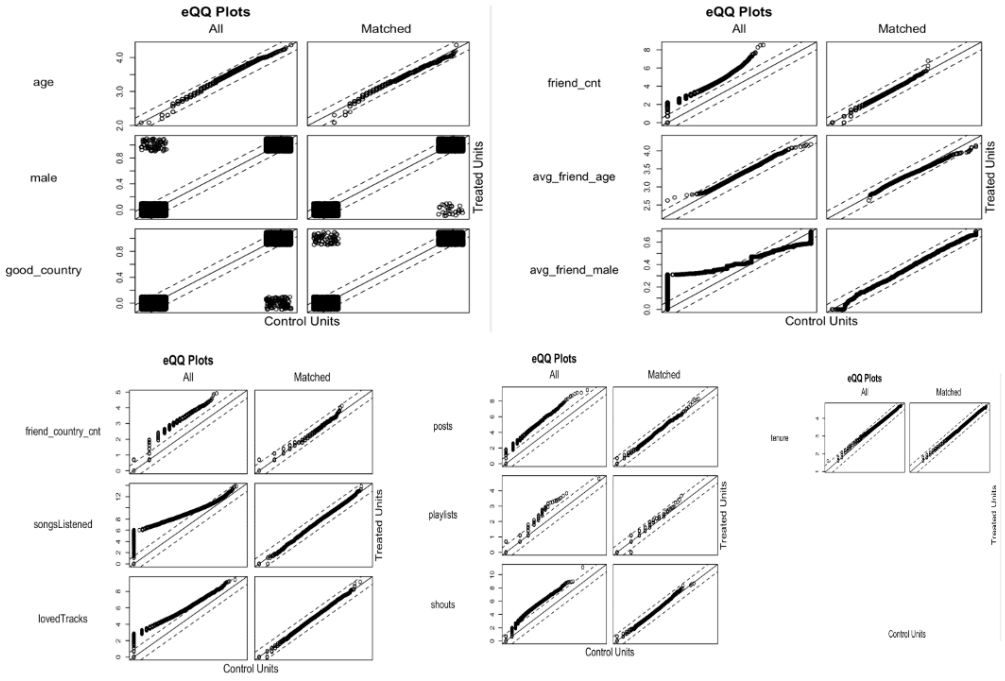
```
mean in group 0 mean in group 1
      0.637071        0.634816

Friend country count: Welch Two Sample t-test
data:  matched_data[, v] by matched_data$treatment
t = 2.1266, df = 13671, p-value = 0.03347
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01375454 0.33773420
sample estimates:
mean in group 0 mean in group 1
      5.722295        5.546550

Songs listened: Welch Two Sample t-test
data:  matched_data[, v] by matched_data$treatment
t = -1.9363, df = 13390, p-value = 0.05285
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2266.37579    13.84275
sample estimates:
mean in group 0 mean in group 1
      26602.70        27728.97

Loved tracks: Welch Two Sample t-test
data:  matched_data[, v] by matched_data$treatment
t = -1.1727, df = 13767, p-value = 0.2409
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.641267   3.931173
sample estimates:
mean in group 0 mean in group 1
       138.260         144.115

posts: Welch Two Sample t-test
data:  matched_data[, v] by matched_data$treatment
t = -1.2548, df = 9660.5, p-value = 0.2096
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.034586  1.324056
sample estimates:
mean in group 0 mean in group 1
      7.089034        9.444299

playlists: Welch Two Sample t-test
data:  matched_data[, v] by matched_data$treatment
```

t = -0.88385, df = 13484, p-value = 0.3768
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05094146 0.01927842
sample estimates:
mean in group 0 mean in group 1
      0.6244009   0.6402324

shouts: Welch Two Sample t-test
data: matched_data[, v] by matched_data$treatment
t = -0.4414, df = 13733, p-value = 0.6589
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.502790 4.112376
sample estimates:
mean in group 0 mean in group 1
       43.05403   44.24924

Tenure: Welch Two Sample t-test
data: matched_data[, v] by matched_data$treatment
t = -0.9618, df = 13768, p-value = 0.3362
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9883963 0.3377064
sample estimates:
mean in group 0 mean in group 1
       45.78853   46.11387

Visually plotting the results of the PSM matching using eQQ plots. For most variables, the observations lie between the dotted lines, on a 45 degree angle, we can consider that the matching was successful and data is balanced.



ESTIMATING TREATMENT EFFECT USING T-TEST: Based on a t-test of adopter variable and treatment variable, the p-value is highly significant. This implies that the difference in means is very significant and that there is a significant treatment effect.

```
        Welch Two Sample t-test

data:  adopter by treatment
t = -10.352, df = 13229, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06892512 -0.04697902
sample estimates:
mean in group 0 mean in group 1
     0.09513435      0.15308642
```

Estimate treatment effect using logistic regression:

- Model 1: Logistic regression of adopter and treatment

```
Call:
lm(formula = adopter ~ treatment, data = matched_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15309 -0.15309 -0.09513 -0.09513  0.90487

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.095134   0.003958   24.03   <2e-16 ***
treatment   0.057952   0.005598   10.35   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3285 on 13768 degrees of freedom
Multiple R-squared:  0.007724,  Adjusted R-squared:  0.007652
F-statistic: 107.2 on 1 and 13768 DF,  p-value: < 2.2e-16
```

From the results, treatment (having >=1 subscriber friend) has a positive and significant impact on the dependent variable. Having 1 or more subscriber friends, increases the odds of becoming adopter by 5.8%.

- Model 2: logistic regression using all variables

Based on this model, most variables are significant predictors of the dependent variable. Most variables have a positive significant impact on probability of becoming an adopter, but some of them have a negative significant effect too. In both models, treatment (having >=1 subscriber friend) increases the probability of becoming a premium user by 5.8%. So, we can conclude that there is significant treatment effect.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.88290 -0.14495 -0.10500 -0.06027  0.99056

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.474e-03  1.323e-02  -0.338 0.735284
treatment      5.730e-02  5.516e-03  10.388  < 2e-16 ***
songsListened  5.325e-07  8.872e-08   6.002 2.00e-09 ***
lovedTracks    1.153e-04  9.685e-06  11.909  < 2e-16 ***
playlists      1.918e-02  2.677e-03   7.165 8.19e-13 ***
posts         -8.831e-06  2.685e-05  -0.329 0.742215
shouts        -9.278e-06  1.967e-05  -0.472 0.637142
friend_cnt     1.303e-04  1.240e-04   1.051 0.293282
age            2.872e-03  4.597e-04   6.249 4.25e-10 ***
male           3.604e-02  5.909e-03   6.099 1.10e-09 ***
tenure        -5.741e-04  1.526e-04  -3.763 0.000169 ***
good_country  -4.089e-02  5.879e-03  -6.956 3.66e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3234 on 13758 degrees of freedom
Multiple R-squared:  0.03858,   Adjusted R-squared:  0.03781
F-statistic: 50.19 on 11 and 13758 DF,  p-value: < 2.2e-16
```

```
> exp(lm_treat2$coefficients)
  (Intercept)     treatment songsListened   lovedTracks     playlists
    0.9955356     1.0589711     1.0000005     1.0001154     1.0193619
        posts        shouts    friend_cnt           age          male
    0.9999912     0.9999907     1.0001303     1.0028765     1.0366926
       tenure  good_country
    0.9994261     0.9599343
```

Q4. Regression Analyses: Now, we will use a logistic regression approach to test which variables (including subscriber friends) are significant for explaining the likelihood of becoming an adopter. Use your judgment and visualization results to decide which variables to include in the regression. Estimate the odds ratios for the key variables. What can you conclude from your results?

I used a logistic regression with adopter as dependent variable and all other variables as independent variables to test the likelihood of becoming an adopter.

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.268e+00  1.619e-01 -20.184  < 2e-16 ***
age                   1.007e-02  5.520e-03   1.824  0.06815 .
male                  4.139e-01  6.090e-02   6.797 1.07e-11 ***
good_country         -4.191e-01  5.952e-02  -7.041 1.91e-12 ***
friend_cnt           -2.832e-04  1.518e-03  -0.187  0.85198
avg_friend_age        2.209e-02  6.792e-03   3.252  0.00114 **
avg_friend_male       5.727e-02  1.147e-01   0.499  0.61757
friend_country_cnt   -5.729e-03  7.730e-03  -0.741  0.45860
songsListened         3.938e-06  7.470e-07   5.271 1.35e-07 ***
subscriber_friend_cnt 3.678e-01  2.471e-02  14.883  < 2e-16 ***
lovedTracks           7.326e-04  7.306e-05  10.027  < 2e-16 ***
posts                -2.084e-04  2.362e-04  -0.882  0.37776
playlists             1.214e-01  2.164e-02   5.610 2.03e-08 ***
shouts               -3.737e-05  1.929e-04  -0.194  0.84639
tenure               -6.011e-03  1.489e-03  -4.038 5.39e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10328  on 13769  degrees of freedom
Residual deviance:  9739  on 13755  degrees of freedom
AIC: 9769

Number of Fisher Scoring iterations: 5

> exp(final_m1$coefficients)
          (Intercept)                   age                  male
           0.03809896             1.01011949             1.51270627
         good_country            friend_cnt        avg_friend_age
           0.65763514             0.99971682             1.02233487
      avg_friend_male    friend_country_cnt         songsListened
           1.05894392             0.99428719             1.00000394
subscriber_friend_cnt          lovedTracks                  posts
           1.44449596             1.00073284             0.99979166
            playlists               shouts                 tenure
           1.12909143             0.99996263             0.99400726
```

Assuming an alpha of 0.05; friend count, avg friend male, friend country count, posts and shouts are not significant predictors of the likelihood of becoming an adopter. So, I ran a regression again with significant variables and estimate odds ratios. I retained age as the significance level is almost close to 0.05.

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -3.327e+00  1.337e-01 -24.881  < 2e-16 ***
age                    1.039e-02  5.467e-03   1.901 0.057346 .
male                   4.232e-01  6.036e-02   7.011 2.37e-12 ***
good_country          -4.130e-01  5.932e-02  -6.962 3.36e-12 ***
subscriber_friend_cnt  3.600e-01  2.391e-02  15.055  < 2e-16 ***
avg_friend_age         2.407e-02  6.547e-03   3.676 0.000237 ***
songsListened          3.770e-06  7.299e-07   5.165 2.40e-07 ***
lovedTracks            7.180e-04  7.226e-05   9.937  < 2e-16 ***
playlists              1.195e-01  2.149e-02   5.560 2.70e-08 ***
tenure                -5.995e-03  1.483e-03  -4.043 5.28e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10328.5  on 13769  degrees of freedom
Residual deviance:  9742.1  on 13760  degrees of freedom
AIC: 9762.1

Number of Fisher Scoring iterations: 5
```

Using the significant variables, the results indicate that subscriber friend count, male, avg friend male are positive significant indicators of the likelihood of becoming an adopter, whereas good_country and tenure are significant negative indicators of the odds of becoming an adopter (premium user).

Odds Ratios:

Variable	Odds Ratio
(Intercept)	0.03649473
Age	1.01044558
male	1.55768478
good_country	0.66438859
subscriber_friend_c	1.43338569
avg_friend_age	1.03302435
songsListened	1.00000373
lovedTracks	1.00071881
playlists	1.13021891
tenure	0.9943869

Interpretation:
Demographics:

- Being a male user (1) , increases odds of becoming an adopter by a factor of 1.58
- Every unit increase in age, increases the odds of becoming an adopter by a factor of 1.01
- For every unit change in tenure, the odds of becoming an adopter decrease by a factor of 0.99
- For a change in country (to US,UK), odds of becoming an adopter decrease by a factor of 0.66

Engagement:

- Songs listened and loved tracks do not impact / increase the likelihood of becoming an adopter significantly.
- Whereas creating playlists, increases the odds of becoming an adopter by a factor of 1.13

Social:

- Every unit increase in subscriber friends increases the odds of becoming an adopter by a factor of 1.43.
- Every unit increase in avg friend age, increases the odds of becoming an adopter by a factor of 1.03.

Q5) Takeaways: Discuss some key takeaways from your analysis. Specifically, how do your results inform a "free-to-fee" strategy for High Note?

- With the matched data, the number of songs listened to is significant, but the impact is almost nil, which is quite intuitive as the mean time spent on the website before becoming a premium member is higher than 3 years.
- The Community Participation data (posts and shouts) are also not significant to predict a customer to become a premium member. As per the "ladder of participation" theory, community involvement plays a significant role in subscription decisions, but for Highnote this is not the case.
- On the contrary, the number of subscriber friends is still associated with a strong positive effect on the user's likelihood to become a premium member, it has a significant association with the subscription decision and the with odd ratio is 1.443. Therefore, I can conclude that **there is a causal relationship between having subscriber friends and becoming a subscriber**. Highnote can focus on "virality" to increase subscriber count.
- Male users have a higher likelihood of becoming premium users when compared to women. This is clear demographic differential, which can be utilized to develop target marketing strategies.
- Also, users from US, UK and Germany seem to have very low likelihood of becoming premium users. This can probably be due to competition in the streaming industry in these countries or subscription price compared to competition.

Of all the variables, being Male and having Subscriber friends have the highest positive impact on the likelihood of becoming an adopter. Highnote can adopt both push and pull strategy. They

should tap into their male premium users and engage them in increased social participation with free users on the platform. At the same time, they can make friend recommendations (of male subscriber friends) to free users and improve the subscriber friend counts. Its clear from the analysis that not “Friend count” but “Subscriber friend count” is what impacts subscription decisions.

Additionally, Highnote should analyze the cause of low probability of subscription in the users from US, UK and Germany. 35% of users are from these countries and its important to understand their behavior. Assuming the same % for the population and the fact that premium users make up a small % of the population, it will be beneficial to include them in Highnote’s “free to fee” strategy.

Its interesting to note that consumption is not a strong influencer of subscription decision. For a freemium platform however, content procurement is a major cost. So focusing on other variables which impact the subscription decision is more benficial than focusing on engagement.

Highnote should take cognizance of the fact that “Social influence” is palying a huge role in subscription decision. They can enhance their platform to encourage more social engagement beyond posts, shouts via blogs, leadership activities, interaction activities etc. Community participation and leadership activities are tools that help users connect and engage with not just others but with the platform itself, which can become a significant driver of subscription behavior. Most of the users are young, and youngsters love to engage via social media. Highnote can either enable social sharing and engagement using popular platforms like Facebook, Snap, Instagram etc or build its own social capabilities that provide users with content, fun and engagement.