Kathleen Sebastian
BANA 270: Customer & Social Analytics
Midterm Exam

HighNote Case Study: Propensity Score Matching

1. **<u>Summary statistics</u>: 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?**

The descriptive statistics in Figure 1 illustrate the variation in means for each variable in the High Note dataset between group = 0 or non-adopters and group = 1 or adopters. Some notable statistical points to emphasize is that the average age for users in the non-adopter subsample is ~24 years old, while the average age for users in the adopter subsample is ~25 years old. As expected, and on average, it is also evident that the adopter subsample listened to significantly more songs than those in the non-adopter subsample, despite containing a larger number of sample datapoints.

Figure 1: Descriptive Statistics by Adopter and Non-Adopter Subsample

```
> describe.by(df, group=df$adopter)

 Descriptive statistics by group
group: 0
                     vars     n     mean       sd   median trimmed      mad min     max   range  skew kurtosis     se
ID                      1 40300 20150.50 11633.75 20150.50 20150.50 14937.19   1   40300   40299  0.00    -1.20  57.95
age                     2 40300    23.95     6.37    23.00    23.09     4.45   8      79      71  1.97     6.80   0.03
male                    3 40300     0.62     0.48     1.00     0.65     0.00   0       1       1 -0.50    -1.75   0.00
friend_cnt              4 40300    18.49    57.48     7.00    10.28     7.41   1    4957    4956 32.67  2087.42   0.29
avg_friend_age          5 40300    24.01     5.10    23.00    23.40     3.95   8      77      69  1.84     7.15   0.03
avg_friend_male         6 40300     0.62     0.32     0.67     0.65     0.35   0       1       1 -0.52    -0.72   0.00
friend_country_cnt      7 40300     3.96     5.76     2.00     2.66     1.48   0     129     129  4.74    38.29   0.03
subscriber_friend_cnt   8 40300     0.20     0.40     0.00     0.13     0.00   0       1       1  1.50     0.24   0.00
songsListened           9 40300 17589.44 28416.02  7440.00 11817.64 10576.87   0 1000000 1000000  6.05   105.85 141.55
lovedTracks            10 40300    86.82   263.58    14.00    36.35    20.76   0   12522   12522 13.12   335.93   1.31
posts                  11 40300     5.29   104.31     0.00     0.23     0.00   0   12309   12309 73.92  7005.34   0.52
playlists              12 40300     0.55     1.07     0.00     0.45     0.00   0      98      98 28.21  1945.28   0.01
shouts                 13 40300    29.97   150.69     4.00     8.84     4.45   0    7736    7736 22.53   779.12   0.75
adopter                14 40300     0.00     0.00     0.00     0.00     0.00   0       0       0   NaN      NaN   0.00
tenure                 15 40300    43.81    19.79    44.00    43.72    22.24   1     111     110  0.05    -0.70   0.10
good_country           16 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
ID                      1 3527 42064.00  1018.30 42064.00 42064.00  1307.65 40301  43827   3526  0.00    -1.20  17.15
age                     2 3527    25.98     6.84    24.00    25.05     4.45     8     73     65  1.68     4.39   0.12
male                    3 3527     0.73     0.44     1.00     0.79     0.00     0      1      1 -1.03    -0.94   0.01
friend_cnt              4 3527    39.73   117.27    16.00    23.69    17.79     1   5089   5088 26.04  1013.79   1.97
avg_friend_age          5 3527    25.44     5.21    24.36    24.83     3.91    12     62     50  1.68     5.05   0.09
avg_friend_male         6 3527     0.64     0.25     0.67     0.65     0.25     0      1      1 -0.54    -0.05   0.00
friend_country_cnt      7 3527     7.19     8.86     4.00     5.36     4.45     0    136    136  3.61    24.53   0.15
subscriber_friend_cnt   8 3527     0.49     0.50     0.00     0.49     0.00     0      1      1  0.02    -2.00   0.01
songsListened           9 3527 33758.04 43592.73 20908.00 25811.69 23276.82     0 817290 817290  4.71    46.64 734.03
lovedTracks            10 3527   264.34   491.43   108.00   161.68   140.85     0  10220  10220  6.52    80.96   8.27
posts                  11 3527    21.20   221.99     0.00     1.44     0.00     0   8506   8506 26.52   852.38   3.74
playlists              12 3527     0.90     2.56     1.00     0.59     1.48     0    118    118 28.84  1244.31   0.04
shouts                 13 3527    99.44  1156.07     9.00    23.89    11.86     0  65872  65872 52.52  2969.09  19.47
adopter                14 3527     1.00     0.00     1.00     1.00     0.00     1      1      0   NaN      NaN   0.00
tenure                 15 3527    45.58    20.04    46.00    45.60    20.76     0    111    111  0.02    -0.62   0.34
good_country           16 3527     0.29     0.45     0.00     0.23     0.00     0      1      1  0.94    -1.12   0.01
```

To further analyze differences in the means between the variables in the data, Figure 2 manifests the output of t-tests conducted for each variable based on the adopter and non-adopter subsample. Given that the p-values from each variable tested is less than the assumed alpha at 0.05, we have sufficient evident to reject the null hypothesis and indicate that there is a significant difference in the means between the adopter and non-adopter subsamples for each potential predictor or variable in this particular dataset.

Figure 2: T-Test to Test for Differences in Means Comparing Adopter and Non-Adopter Subsample

```
> lapply(df[,c('age','male' , 'friend_cnt' , 'avg_friend_male' ,'avg_friend_age',
+                   'friend_country_cnt'  , 'songsListened' , 'lovedTracks' ,
+                   'posts' , 'playlists' ,'shouts' , 'tenure' ,'good_country', 'subscriber_friend_cnt')], function(a) t.test(a ~ df$adopter))
$age

        Welch Two Sample t-test

data:  a by df$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:  a by df$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:  a by df$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:  a by df$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

$lovedTracks

        Welch Two Sample t-test

data:  a by df$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:  a by df$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:  a by df$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

$avg_friend_age

        Welch Two Sample t-test

data:  a by df$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:  a by df$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:  a by df$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

$shouts

        Welch Two Sample t-test

data:  a by df$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:  a by df$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:  a by df$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
```

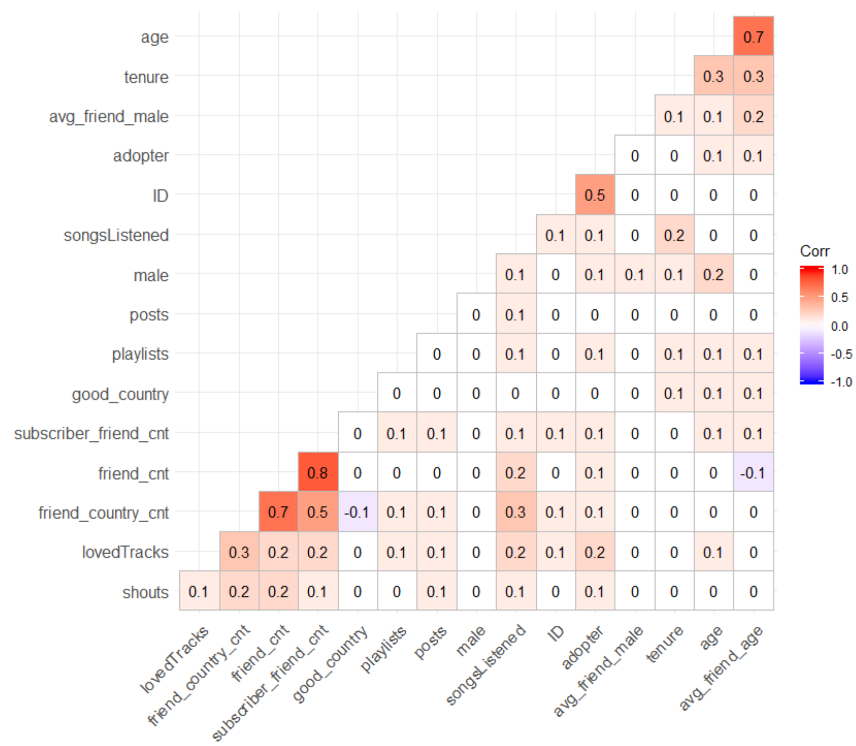
```
$subscriber_friend_cnt

        Welch Two Sample t-test

data:  a by df$adopter
t = -33.978, df = 3931.7, p-value < 2.2e-16
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
```

2. **<u>Data Visualization</u>: 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?**

Figure 3 depicts a basic correlation matrix generated to visualize the correlation coefficients for each variable. This visualization indicates a strong positive correlation between the following pairs of variables:

- Friend_cnt and subscriber_friend_cnt
- Friend_country_cnt and friend_cnt
- Age and average_friend_age

Figure 3 – Correlation Matrix for HighNote Dataset Variables



Based on the visualizations generated in Figure 4, we can conclude that there is a wider range in ages for adopters in comparison to those who are non-adopters. When observing the bar chart for comparing adopters and non-adopters by country and gender, it is clear that there is a significantly larger amount of users who originate from countries other than UK, US or Germany and more males for users in the data for both adopter and non-adopters subsamples.

Figure 4 – Visualizations for Demographic Related Variables

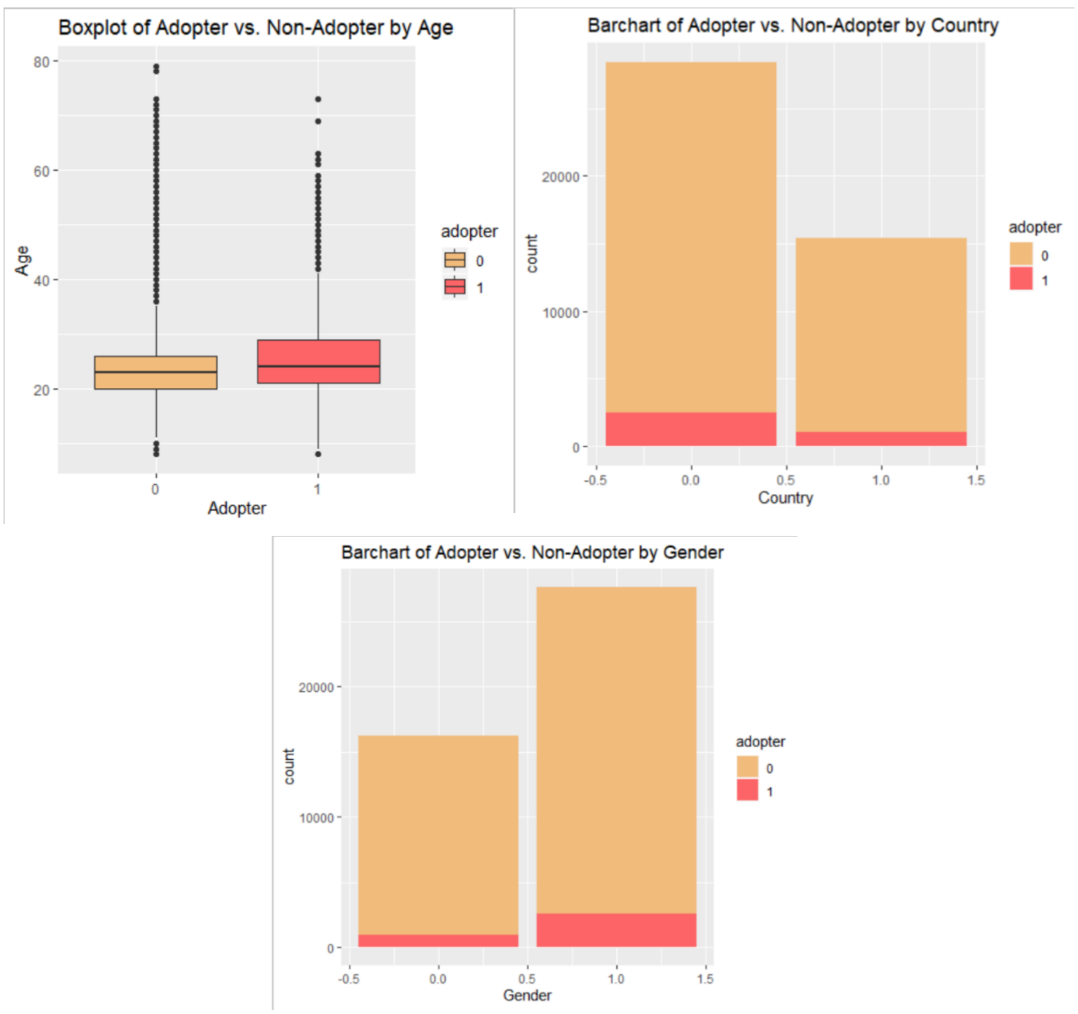


Figure 5 illustrates visualizations pertaining to the peer influence related variables in the data. When observing the scatter chart between the age and friend_count variable, we notice a relatively similar distribution of datapoints for most adopters and non-adopters. However, it is important to note that there are several outliers for non-adopters with high friend counts; this may be valuable for High Note to consider, especially when trying to determine marketing strategies on the types of non-paying users to target (i.e. non-paying subscribers with high number of friends may be influential).

Next, the histogram in Figure 5 illustrates that there are generally younger in age have a higher friend count. This is especially for those users in the non-adopter group. The histogram depicts that the average friend age for non-adopters is clustered within those who are in their early to late twenties. The subsequent bar chart illustrating the friend country count tells us that a large proportion of non-adopters belong to a country other than UK, US or Germany. The last bar chart indicates that the average subscriber friend count for adopters is exponentially greater than that of non-adopters.

Figure 5 – Visualizations for Peer Influence Related Variables

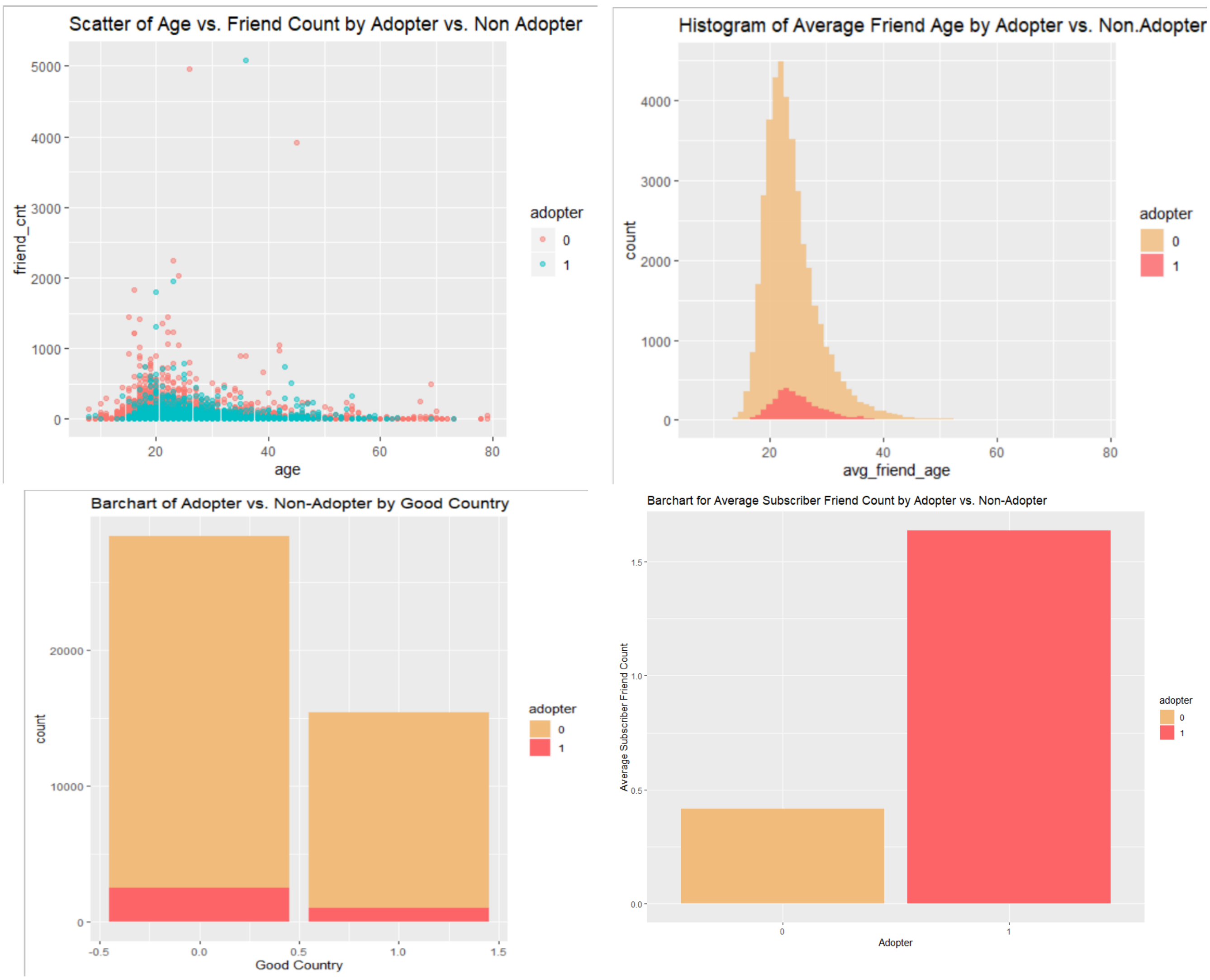
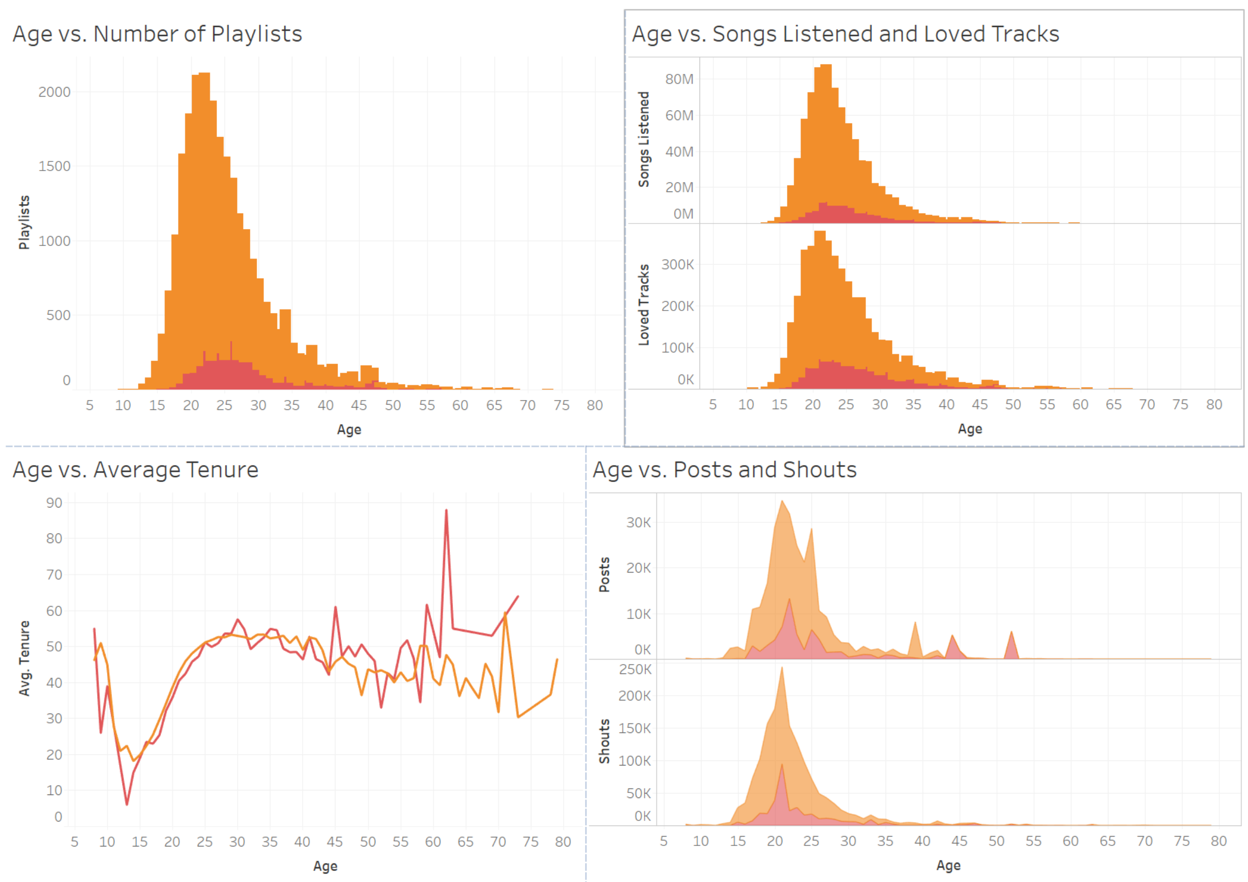


Figure 6 captures some of the user related variables in the High Note dataset. To generate more insightful visualizations, these variables were plotted against the age demographic variable. In the first graph, we notice that most non-adopters who are within the younger age range possess many playlists, with the maximum value being a group of 22-year old's and owning over 1800 playlists in total. We notice a similar pattern of high non-adopter usability and interaction with the High Note platform when looking at the visualization pertaining to Age vs. the Songs Listened and Loved Tracks, where a group of 22 years old's listened to the highest number of songs and loved the greatest number of tracks. When comparing age vs. the user's average tenure, as expected, we notice a greater average tenure for adopters who are older in age. Surprisingly, for both adopters and non-adopters, there is an extremely low average tenure for those who are between ages 10-15. Another interesting observation from Figure 6 is that non-adopters in the data are interacting a lot with the platform, especially in the form of posts and shouts. We also see a similar trend for adopters and see a few users in the older age range (between 45-55) who contribute to a large number of posts on the platform.

Figure 6 – Visualizations for User-Related Variables



3. <u>Propensity Score Matching (PSM)</u>: **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.**

To conduct propensity score matching to test whether having a subscriber friends affects the likelihood of a user becoming an adopter or a paying customer, we pre-analyzed the data using the non-matched data. First, dummy variables were created for the "subscriber_friend_cnt" variable to classify between the control group and treatment group. 1 is assigned to data rows where the customer has subscriber friend count of equal to or greater to 1 (treatment group). On the contrary, the 0 is assigned to data rows where customer's the subscriber friend count is 0 (control group). Figure 6 illustrates the results from a t-test conducted to determine whether the difference in means between the two groups are statistically significant at a 95% confidence level. Given a p value that is less than the assumed alpha at 0.05, we have enough evidence to reject the null hypothesis, illustrating a statistical significance in the differences between the two means of the control group and treatment group.

Figure 6 – Two Sample T-Test Comparing Subscriber Friend Count Groups

```
> df$subscriber_friend_cnt <- ifelse(df$subscriber_friend_cnt >=1,1,0)
> with(df, t.test(subscriber_friend_cnt ~ adopter))

        Welch Two Sample t-test

data:  subscriber_friend_cnt by adopter
t = -33.978, df = 3931.7, p-value < 2.2e-16
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
```

Figure 7 illustrates the following step in the PSM test to calculate the mean and test for significance for each covariate based on the treatment status. Given that Figure 2 illustrates statistical significance for all the variables in the data, all variables were considered as covariates. In addition, when further t-tests were conducted by the treatment and control group, it is evident that all the variables are statistically significant (see R script for t-test output results).

Figure 7 – Difference in Means for Pre-Treatment Covariates

```r
# Difference in means for pre-treatment covarities
df_cov <- c('age','male' , 'friend_cnt' , 'avg_friend_male' ,'avg_friend_age',
            'friend_country_cnt'  , 'songsListened' , 'lovedTracks' ,
            'posts' , 'playlists' ,'shouts' , 'tenure' ,'good_country', 'subscriber_friend_cnt')
diff_in_means_covariteis <- df %>%
  group_by(adopter) %>%
  select(one_of(df_cov)) %>%
  summarise_all(funs(mean(., na.rm = T)))
```

	adopter	age	male	friend_cnt	avg_friend_male	avg_friend_age	friend_country_cnt	songsListened	lovedTracks	posts	playlists	shouts	tenure	good_country
1	0	23.94844	0.6218610	18.49166	0.6165888	24.01142	3.957891	17589.44	86.82263	5.293002	0.5492804	29.97266	43.80993	0.3577916
2	1	25.97987	0.7292316	39.73377	0.6365983	25.44131	7.188829	33758.04	264.34080	21.200454	0.9007655	99.43975	45.58322	0.2874965

Figure 8 illustrates the propensity score matching results, after running a logit model and using all the covariates depicted in Figure 7. Then, Figure 9 illustrates the first several rows of predicted propensity score output after using the propensity score matching model in Figure 8 to calculate the propensity score for each High Note subscriber. In other words, the propensity score is the probability of a High Note subscriber being assigned to the treatment group or specifically, those users with one or more subscriber friends, given the set of covariates utilized to generate the model.

Figure 8 – Propensity Score Matching Results

```
> # Propensity Score Matching; given all means significant for all, use all features related to the outcome variable
> hn_ps <- glm(subscriber_friend_cnt ~ age + male + good_country + avg_friend_age + avg_friend_male + friend_country_cnt + songsListened +
+               lovedTracks + posts + playlists + shouts + tenure + friend_cnt,
+             family = binomial(), data = df)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(hn_ps)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4206  -0.5671  -0.4220  -0.3001   2.5619

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.144e+00  7.703e-02 -66.782  < 2e-16 ***
age                 1.970e-02  2.808e-03   7.015 2.30e-12 ***
male                4.311e-02  2.998e-02   1.438 0.150419
good_country        3.201e-02  2.922e-02   1.096 0.273235
avg_friend_age      7.955e-02  3.481e-03  22.850  < 2e-16 ***
avg_friend_male     2.514e-01  5.029e-02   4.999 5.75e-07 ***
friend_country_cnt  1.110e-01  4.765e-03  23.302  < 2e-16 ***
songsListened       6.906e-06  5.156e-07  13.396  < 2e-16 ***
lovedTracks         6.671e-04  5.645e-05  11.817  < 2e-16 ***
posts               5.699e-04  2.682e-04   2.125 0.033613 *
playlists           5.639e-03  1.190e-02   0.474 0.635530
shouts             -4.909e-05  3.707e-05  -1.324 0.185434
tenure             -2.571e-03  7.769e-04  -3.309 0.000935 ***
friend_cnt          3.132e-02  1.034e-03  30.301  < 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: 34170  on 43813  degrees of freedom
AIC: 34198

Number of Fisher Scoring iterations: 7
```

Figure 9 – Propensity Score Output

```
> # Use model to calculate propensity score and generate new dataframe
> ps_df <- data.frame(pr_score = predict(hn_ps, type = "response"),
+                     subscriber_friend_cnt = hn_ps$model$subscriber_friend_cnt)
> head(ps_df)
    pr_score subscriber_friend_cnt
1 0.08597334                     0
2 0.14417767                     0
3 0.08217010                     0
4 0.23894067                     1
5 0.69552208                     0
6 0.22306633                     0
```

Figure 10 plots the propensity scores into a histogram based on the same estimated propensity scores by treatment status values generated in Figure 9. Here, we notice that the histogram for subscribers in the control group are right skewed, with many subscribers possessing lower propensity scores (~0.00-0.25). On the contrary, we observe a different trend for the histogram representing subscribers in the treatment group. Here, we see a relatively equal distribution of subscribers with lower and higher propensity scores or in other words, an equal distribution of subscribers being assigned to the treatment group (subscriber friend count equal to or greater than 1).

Figure 10 – Histogram Plot of Propensity Scores by Treatment Status

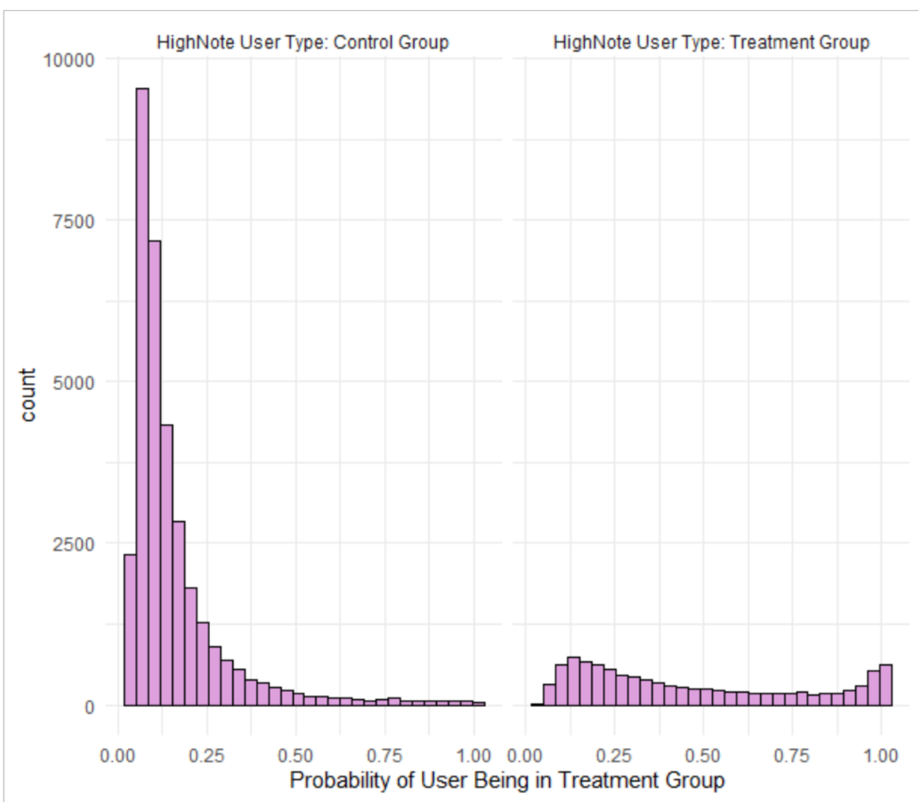


Figure 11 depicts the output when the match it package is initiated to identify pairs of observations with similar propensity scores but are distinct in treatment status and more specifically, how successful the matching process was. In this new matched dataset, there are 9823 data records that were treated and subsequently matches it 9823 records from the Control group based on the nearest propensity score estimate.

Figure 11 – Tabular Output for Matching Algorithm for Estimating the Treatment Effect

```
> summary(mod_match)

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

Summary of balance for all data:
                   Means Treated Means Control SD Control Mean Diff    eQQ Med
distance                  0.4635        0.1550     0.1436    0.3086     0.2506
age                      25.3732       23.7476     6.2245    1.6256     1.0000
male                      0.6363        0.6288     0.4831    0.0074     0.0000
good_country              0.3433        0.3547     0.4784   -0.0114     0.0000
avg_friend_age           25.3904       23.7614     5.0577    1.6291     1.5909
avg_friend_male           0.6358        0.6131     0.3343    0.0227     0.0738
friend_country_cnt        9.3856        2.7251     3.1024    6.6606     5.0000
songsListened         33735.6404    14602.2205 23214.2898 19133.4199 15471.0000
lovedTracks             225.3647       65.2137   181.4812  160.1510    65.0000
posts                    20.5230        2.5434    33.7947   17.9796     0.0000
playlists                 0.7441        0.5295     0.9673    0.2146     0.0000
shouts                  101.8195       16.4230    79.7381   85.3965    15.0000
tenure                   46.5487       43.2027    19.7212    3.3460     3.0000
friend_cnt               54.0210       10.4313    15.2769   43.5896    22.0000
                     eQQ Mean     eQQ Max
distance               0.3086      0.6840
age                    1.6296      5.0000
male                   0.0074      1.0000
good_country           0.0114      1.0000
avg_friend_age         1.6369     11.5000
avg_friend_male        0.0958      0.3636
friend_country_cnt     6.6598     95.0000
songsListened      19126.1623 653702.0000
lovedTracks          159.9562   6343.0000
posts                 17.8829   9535.0000
playlists              0.2092     26.0000
shouts                85.1764  59168.0000
tenure                 3.3473     10.0000
friend_cnt            43.5838   4794.0000

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

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
good_country         -30.1771   0.0000  -30.3571   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
friend_cnt            25.3162  45.4545   25.3062   0.0000

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

>
```

Thus, the QQ plots generated in Figure 12, illustrates a comparison between the probability distributions of the treated vs. control groups provided the covariates used. The “All” plots the probabilities for the covariates before matching and the “Matched” plots illustrated the probabilities after propensity score matching was conducted. Ideally, we want to see an improvement in the positioning of the points in the “Matched” plots, which should fall closer within the dotted lined in the comparison to the “All” plots. And we certainly observe this pattern from the plots in Figure 12.

Figure 12 – QQ Plots for Matching Algorithm

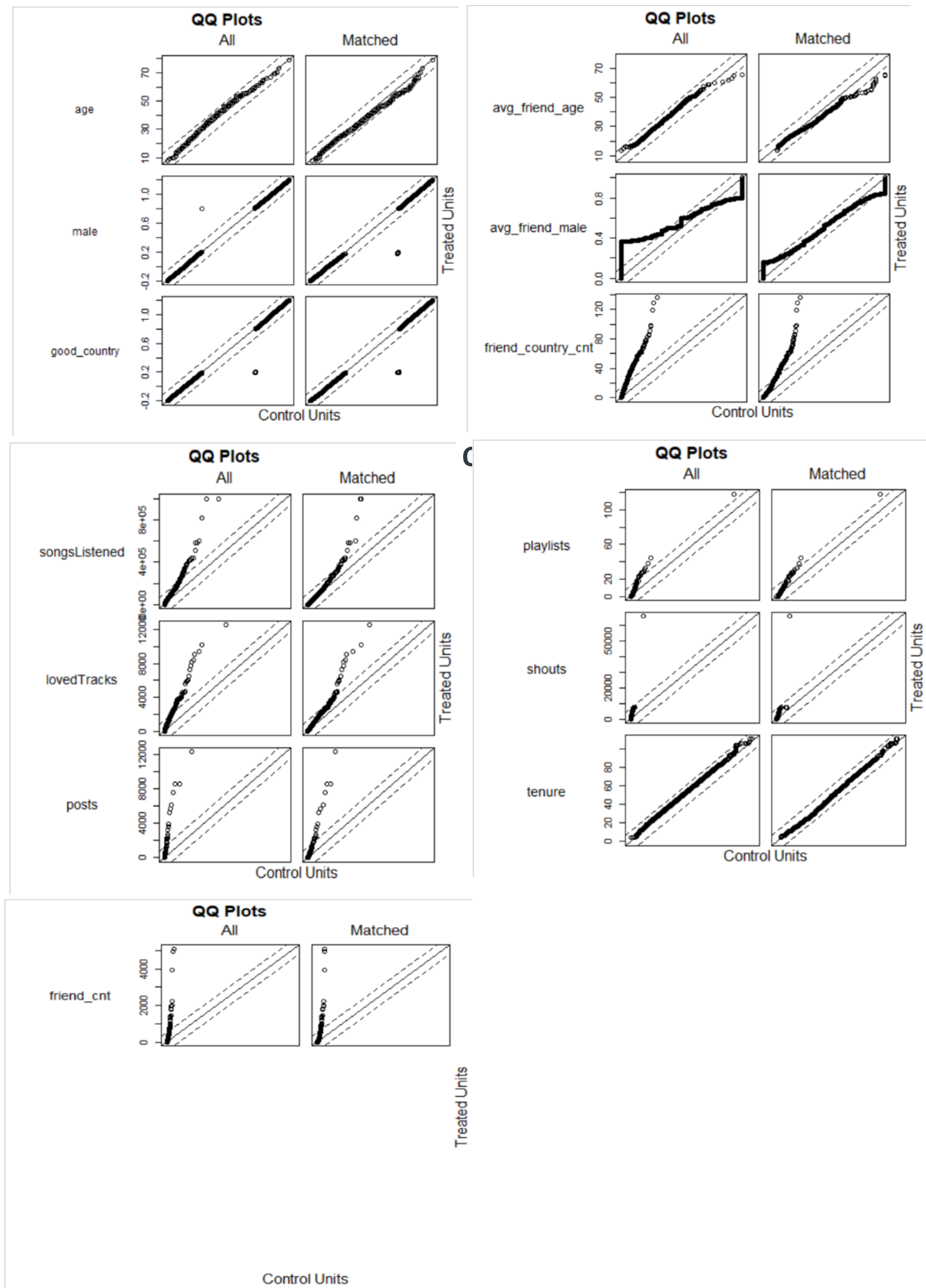
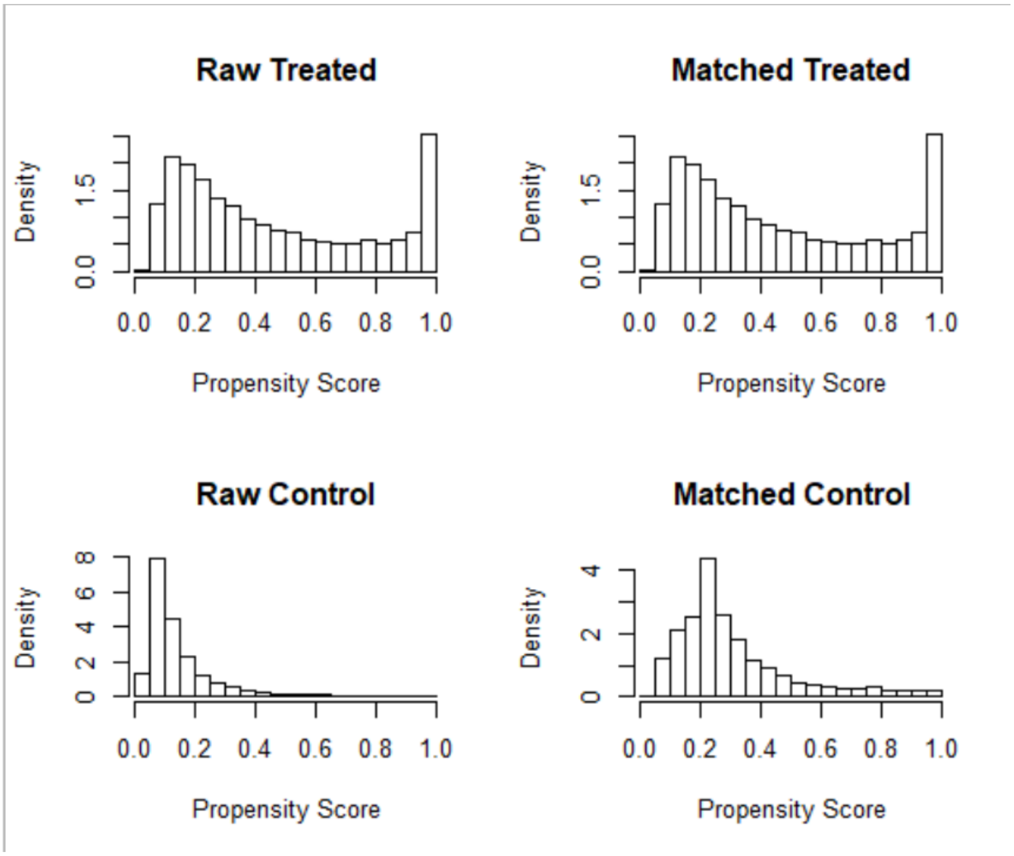


Figure 13 – Histogram of Matching Algorithm Output



Two types of method used to evaluate the covariate balance in matched sample can be done by conducting a a difference in means tests comparing the means across treatment groups for the covariates and estimating treatment effect using OLS with and without covariates. Figure 15 illustrates the model output generated when estimating treatment effect using OLS with and without covariates. When we add the covariates into the model, we notice that they are all statistically significant, except for the “friend_cnt” and “avg_friend_male” variables.

Figure 15 – Estimating Treatment Using OLS With and Without Covariates

```
> summary(lm_treat1)

Call:
lm(formula = adopter ~ subscriber_friend_cnt, data = matched_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.17754 -0.17754 -0.08684 -0.08684  0.91316 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           0.086837   0.003387   25.64   <2e-16 ***
subscriber_friend_cnt 0.090705   0.004790   18.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3357 on 19644 degrees of freedom
Multiple R-squared:  0.01793,	Adjusted R-squared:  0.01788 
F-statistic: 358.7 on 1 and 19644 DF,  p-value: < 2.2e-16
```

```
> lm_treat2 <- lm(adopter ~ subscriber_friend_cnt +  age + male + good_country + friend_cnt + avg_friend_age
+                 + avg_friend_male + friend_country_cnt + songsListened + lovedTracks + posts + playlists + s
houts + tenure, data = matched_df)
> summary(lm_treat2)

Call:
lm(formula = adopter ~ subscriber_friend_cnt + age + male + good_country + 
    friend_cnt + avg_friend_age + avg_friend_male + friend_country_cnt + 
    songsListened + lovedTracks + posts + playlists + shouts + 
    tenure, data = matched_df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.27876 -0.15553 -0.10616 -0.05705  1.00012 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -3.618e-02  1.338e-02  -2.705 0.006844 ** 
subscriber_friend_cnt  7.663e-02  4.901e-03  15.635  < 2e-16 ***
age                    1.764e-03  4.596e-04   3.838 0.000125 ***
male                   3.070e-02  5.125e-03   5.991 2.12e-09 ***
good_country          -3.889e-02  5.014e-03  -7.756 9.22e-15 ***
friend_cnt            -1.620e-05  3.595e-05  -0.451 0.652317    
avg_friend_age         1.556e-03  5.799e-04   2.684 0.007282 ** 
avg_friend_male        7.275e-03  9.820e-03   0.741 0.458791    
friend_country_cnt     1.055e-03  4.454e-04   2.368 0.017884 *  
songsListened          6.209e-07  6.610e-08   9.394  < 2e-16 ***
lovedTracks            8.509e-05  5.971e-06  14.250  < 2e-16 ***
posts                  2.819e-05  1.355e-05   2.081 0.037461 *  
playlists              7.252e-03  1.407e-03   5.155 2.56e-07 ***
shouts                 1.295e-05  4.562e-06   2.838 0.004538 ** 
tenure                -3.098e-04  1.322e-04  -2.344 0.019095 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3304 on 19631 degrees of freedom
Multiple R-squared:  0.04936,	Adjusted R-squared:  0.04868 
F-statistic: 72.81 on 14 and 19631 DF,  p-value: < 2.2e-16
```

******NOTE: SEE ATTACHED R-SCRIPT FOR THE SAME PSM ANALYSIS USING LOGGED VARIABLES.***

4. **<u>Regression Analyses</u>: 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?**

Figure 16 manifests the initial logistic model generated, using all the covariates in the model. As evident from the output, predictor variables such as friend count, shouts, posts, and having an average male friend are not significant when determining the likelihood of a user being an adopter or a paying subscriber.

Figure 16 – Model #1: Logistic Regression Output With All Covariates

```
> summary(hn_lr)

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

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.6288  -0.3990  -0.3240  -0.2678   2.7604  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -4.213e+00  9.562e-02 -44.062  < 2e-16 ***
male                   4.139e-01  4.175e-02   9.914  < 2e-16 ***
age                    2.103e-02  3.517e-03   5.979 2.24e-09 ***
subscriber_friend_cnt  9.719e-01  4.211e-02  23.080  < 2e-16 ***
friend_cnt            -4.584e-04  2.972e-04  -1.543 0.122942    
avg_friend_age         2.369e-02  4.637e-03   5.108 3.25e-07 ***
friend_country_cnt     1.401e-02  3.646e-03   3.843 0.000122 ***
songsListened          6.152e-06  5.212e-07  11.805  < 2e-16 ***
lovedTracks            6.148e-04  4.828e-05  12.734  < 2e-16 ***
good_country          -3.939e-01  4.077e-02  -9.661  < 2e-16 ***
playlists              6.467e-02  1.310e-02   4.938 7.89e-07 ***
tenure                -4.929e-03  1.024e-03  -4.812 1.49e-06 ***
shouts                 7.416e-05  6.476e-05   1.145 0.252113    
posts                  1.074e-04  9.027e-05   1.189 0.234260    
avg_friend_male        1.047e-01  6.555e-02   1.597 0.110222    
---
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: 22198  on 43812  degrees of freedom
AIC: 22228

Number of Fisher Scoring iterations: 5
```

Figure 17 illustrates an optimized logistic regression model, using only the significant covariates from the first initial model, which also includes a calculation of the odds ratios of each respective predictor variables in the model. This final model concludes that the odds ratio will increase for a non-paying user to convert to a paying user at High Note if the user has friends who are also subscribers to the High Note platform. Here, for everyone unit change in the subscriber friend count, the odds of a user becoming an adopter increases by a factor of approximately 2.66. Additional variables that also increase the likelihood of converting a non-paying user to a paying user at High Note is older age and males. More specifically, for every unit change in the user’s age, the likelihood of a user becoming an adopter increases by a factor of approximately 1.02 and given that a user is a male, the likelihood of them becoming an adopter increases by a factor of approximately, 1.51. Thus, it is also important to note that all the predictors in the model generated positive odd ratio coefficients.

Figure 17 – Model #2 (FINAL MODEL): Logistic Regression Output with Only Significant Covariates and Odds Ratio

```
> # Optimize model to include only significant variables
> hn_lr_opt <- hn_lr <- glm(adopter ~ male + age + subscriber_friend_cnt + avg_friend_age + friend_
country_cnt + songsListened + lovedTracks + good_country + playlists + tenure,
+                           family = binomial(), data = df)
> # Optimize model to include only significant variables
> hn_lr_opt <- hn_lr <- glm(adopter ~ male + age + subscriber_friend_cnt + avg_friend_age + friend_
country_cnt + songsListened + lovedTracks + good_country + playlists + tenure,
+                           family = binomial(), data = df)
> summary(hn_lr_opt)

Call:
glm(formula = adopter ~ male + age + subscriber_friend_cnt +
    avg_friend_age + friend_country_cnt + songsListened + lovedTracks +
    good_country + playlists + tenure, family = binomial(), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6465  -0.3981  -0.3235  -0.2683   2.7675

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.163e+00  9.122e-02 -45.638  < 2e-16 ***
male                   4.091e-01  4.163e-02   9.828  < 2e-16 ***
age                    2.041e-02  3.506e-03   5.822 5.82e-09 ***
subscriber_friend_cnt  9.794e-01  4.183e-02  23.414  < 2e-16 ***
avg_friend_age         2.502e-02  4.555e-03   5.492 3.98e-08 ***
friend_country_cnt     1.062e-02  2.499e-03   4.250 2.14e-05 ***
songsListened          6.307e-06  5.167e-07  12.205  < 2e-16 ***
lovedTracks            6.215e-04  4.818e-05  12.899  < 2e-16 ***
good_country          -3.963e-01  4.076e-02  -9.722  < 2e-16 ***
playlists              6.465e-02  1.304e-02   4.957 7.14e-07 ***
tenure                -4.798e-03  1.023e-03  -4.691 2.71e-06 ***
---
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: 22208  on 43816  degrees of freedom
AIC: 22230

Number of Fisher Scoring iterations: 5
```

```
> exp(hn_lr_opt$coefficients)
          (Intercept)                  male                   age subscriber_friend_cnt
           0.01555748            1.50543746            1.02062333            2.66293682
       avg_friend_age    friend_country_cnt         songsListened           lovedTracks
           1.02533301            1.01067806            1.00000631            1.00062174
         good_country             playlists                tenure
           0.67282465            1.06678846            0.99521304
```

******NOTE: SEE ATTACHED R-SCRIPT FOR THE SAME LOGISTIC REGRESSION MODELS USING LOGGED VARIABLES.***

5. **<u>Takeaways</u>: Discuss some key takeaways from your analysis. Specifically, how do your results inform a "free-to-fee" strategy for High Note?**

- The three types of predictor variables such as demographics, user engagement, and user engagement is significant and are effective in informing the free to fee strategy for High Note. Some potential strategies that can help High Note implement this strategy include:
    - Allocating focus on customers in countries that are not in the UK, US or Germany.
        - Users who originate from the UK, US, or Germany (a good country) reduces the likelihood of someone converting to a paying customer.
    - Focusing on non-paying customers with high count of subscriber friends, male, and are older in age will increase the likelihood of them becoming a paying customer of High Note.

    - They may also consider creating and implementing strategies for customers who are currently in their early to late twenties (a large number of their non-paying customers are clustered around this specific age range)
- Given that the odds radio for tenure is slightly below 1.0, High Note may consider shifting their focus and.
- Focus on retaining customers for longer periods of time (make them sticky to the platform), given that tenure has some impact converting non-paying users.
    - Strategy may include interacting with customers, offering promotions, creating a personalized experience, and, consistently enhancing the platform's user experience to ensure that customers remain on the platform for a prolonged time period.
- The music content on High Note is also crucial. Therefore, they may consider increasing their music libraries and expanding their content to ensure that both their user types have variation in the music content that is accessible to them on the platform.
    - High Note can also create special platform features for paying users → encouraging those who are non-paying users to sign up.
- Given the high user interaction for customers with ages 21-22, High Note may consider offering a special pricing model and marketing strategy catered for college students.
    - For example, mirroring Hulu's partnership with Spotify, offering students with a bundle of Spotify Premium and Hulu.