Poonam Ahuja Student id no. : 31167068

BANA 277 Midterm: High Note Musical company Case

Question 1:

Summary statistics: Generate descriptive statistics for the key variables in the data set. 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?

Answer:

Below are the screenshots of the summary statistics of the High Note Users. Group 0, indicates the free users and Group 1 indicates the Paid/premium users on High Note Platform

Descriptive statistics by group
group: 0

	vars	n	mean	sd	median	trimmed	mad	min	max	range
age	1	40300	23.95	6.37	23.00	23.09	4.45	8	79	71
male	2	40300	0.62	0.48	1.00	0.65	0.00	0	1	1
friend_cnt	3	40300	18.49	57.48	7.00	10.28	7.41	1	4957	4956
avg_friend_age	4	40300	24.01	5.10	23.00	23.40	3.95	8	77	69
avg_friend_male	5	40300	0.62	0.32	0.67	0.65	0.35	0	1	1
friend_country_cnt	6	40300	3.96	5.76	2.00	2.66	1.48	0	129	129
subscriber_friend_cnt	7	40300	0.42	2.42	0.00	0.13	0.00	0	309	309
songsListened	8	40300	17589.44	28416.02	7440.00	11817.64	10576.87	0	1000000	1000000
lovedTracks	9	40300	86.82	263.58	14.00	36.35	20.76	0	12522	12522
posts	10	40300	5.29	104.31	0.00	0.23	0.00	0	12309	12309
playlists	11	40300	0.55	1.07	0.00	0.45	0.00	0	98	98
shouts	12	40300	29.97	150.69	4.00	8.84	4.45	0	7736	7736
adopter	13	40300	0.00	0.00	0.00	0.00	0.00	0	0	0
tenure	14	40300	43.81	19.79	44.00	43.72	22.24	1	111	110
good_country	15	40300	0.36	0.48	0.00	0.32	0.00	0	1	1

	skew	kurtosis	se
age	1.97	6.80	0.03
male	-0.50	-1.75	0.00
friend_cnt	32.67	2087.42	0.29
avg_friend_age	1.84	7.15	0.03
avg_friend_male	-0.52	-0.72	0.00
friend_country_cnt	4.74	38.29	0.03
subscriber_friend_cnt	72.19	8024.62	0.01
songsListened	6.05	105.85	141.55
lovedTracks	13.12	335.93	1.31
posts	73.92	7005.34	0.52
playlists	28.21	1945.28	0.01
shouts	22.53	779.12	0.75
adopter	NaN	NaN	0.00
tenure	0.05	-0.70	0.10
good_country	0.59	-1.65	0.00

---

group: 1

	vars	n	mean	sd	median	trimmed	mad	min	max	range
age	1	3527	25.98	6.84	24.00	25.05	4.45	8	73	65
male	2	3527	0.73	0.44	1.00	0.79	0.00	0	1	1
friend_cnt	3	3527	39.73	117.27	16.00	23.69	17.79	1	5089	5088
avg_friend_age	4	3527	25.44	5.21	24.36	24.83	3.91	12	62	50
avg_friend_male	5	3527	0.64	0.25	0.67	0.65	0.25	0	1	1
friend_country_cnt	6	3527	7.19	8.86	4.00	5.36	4.45	0	136	136
subscriber_friend_cnt	7	3527	1.64	5.85	0.00	0.84	0.00	0	287	287
songsListened	8	3527	33758.04	43592.73	20908.00	25811.69	23276.82	0	817290	817290
lovedTracks	9	3527	264.34	491.43	108.00	161.68	140.85	0	10220	10220
posts	10	3527	21.20	221.99	0.00	1.44	0.00	0	8506	8506
playlists	11	3527	0.90	2.56	1.00	0.59	1.48	0	118	118
shouts	12	3527	99.44	1156.07	9.00	23.89	11.86	0	65872	65872
adopter	13	3527	1.00	0.00	1.00	1.00	0.00	1	1	0
tenure	14	3527	45.58	20.04	46.00	45.60	20.76	0	111	111
good_country	15	3527	0.29	0.45	0.00	0.23	0.00	0	1	1

	skew	kurtosis	se
age	1.68	4.39	0.12
male	-1.03	-0.94	0.01
friend_cnt	26.04	1013.79	1.97
avg_friend_age	1.68	5.05	0.09
avg_friend_male	-0.54	-0.05	0.00
friend_country_cnt	3.61	24.53	0.15
subscriber_friend_cnt	34.05	1609.52	0.10
songsListened	4.71	46.64	734.03
lovedTracks	6.52	80.96	8.27
posts	26.52	852.38	3.74
playlists	28.84	1244.31	0.04
shouts	52.52	2969.09	19.47
adopter	NaN	NaN	0.00
tenure	0.02	-0.62	0.34
good_country	0.94	-1.12	0.01

T-Test screenshot for all the key variables:


```
        Welch Two Sample t-test

data:  i by data$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 data$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 data$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 data$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
```

Poonam Ahuja Student id no. : 31167068


```
$avg_friend_age

        Welch Two Sample t-test

data:  i by data$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 data$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 data$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 data$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        264.24080
```


```
$posts

        Welch Two Sample t-test

data:  i by data$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 data$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 data$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 data$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
```

```
$tenure

        Welch Two Sample t-test

data:  i by data$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 data$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 data$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
```

Poonam Ahuja Student id no. : 31167068

Summary of the above descriptive statistics:

1.There are 40,300 free users (group 0/adopter = 0) and 3527 paid/premium users (group 1/adopter = 1). Only 8% of the users on High Note Platform are in the paid category, majority of them 92% are in the free category.

2. For premium users, almost all variables have higher mean values than that of free users. Details below:

a. The average age of premium users is 25.98 years old as compared to 23.95 years for the free users. This could imply that younger users prefer free service on High Note platform due to their income or could be any other reason.

b. Premium users have more than double the average friend count (39.73) than free users(18.49).

c. Both premium and free users have similar proportion of average male friends count ie. 0.62

d. Premium users have friends from a more spread of countries as compared to the free users. This is indicated by the scale of 3.96 for the free users and 7.19 for the premium users. Infact it is more than double that of free users.

e. Subscriber friend count for the premium users(1.62) more than double than the free users(0.42)

g. In terms of the user engagement parameters too, the variables have a higher number for the premium users as compared to the free users. For eg – mean of the cumulative number of songs heard by premium users(33,758) is almost double than that of the free users(17,589). Similar is the case, for the number of posts, number of playlists, number of shouts received.

h. The tenure for which the paid users have been on High Note platform on an average is 45.58 months as compared to the free users which is 43.81 months, ie paid users on an average, have been on the platform for 2 months more than the paid users.

i. Since the country has binary details, we can't assess anything from the mean score . Nevertheless, 35% of the users on HighNote platform are from US, UK & Germany. Almost 65% of the users are from the rest of the world.

To summarize, the overall performance of the key variables for premium users are much better than free users. Since almost all demographics, peer influence and user engagement variables of paid users is better, it could be safe to assume that premium users are more likely to be engaged and loyal customers on the High Note music platform as compared to the free users

Also, post conducting the t-test on all variables for adopters (premium users) and non-adopters (free users), p-values of all variables are statistically significant, indicating there is a difference between the two groups, in terms of all the variables.

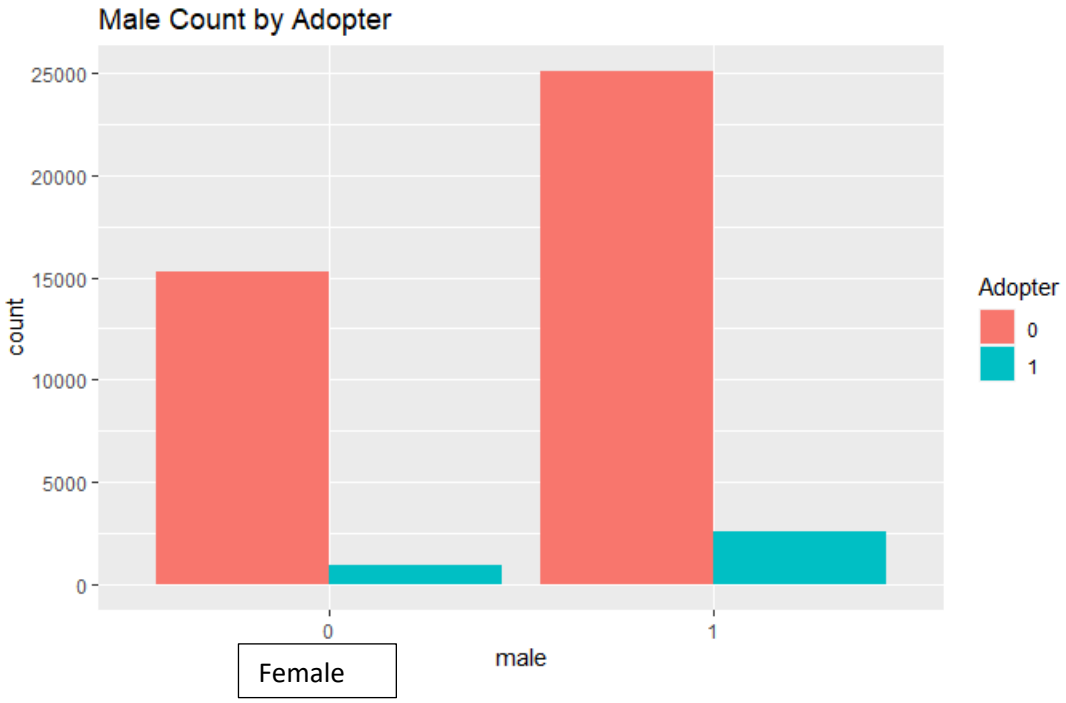
Question 2:

Data Visualization: Generate a set of charts to help visualize how adopters and non-adopters differ from each other in terms of (i) demographics, (ii) peer influence, and (iii) user engagement. What can you conclude from your charts?

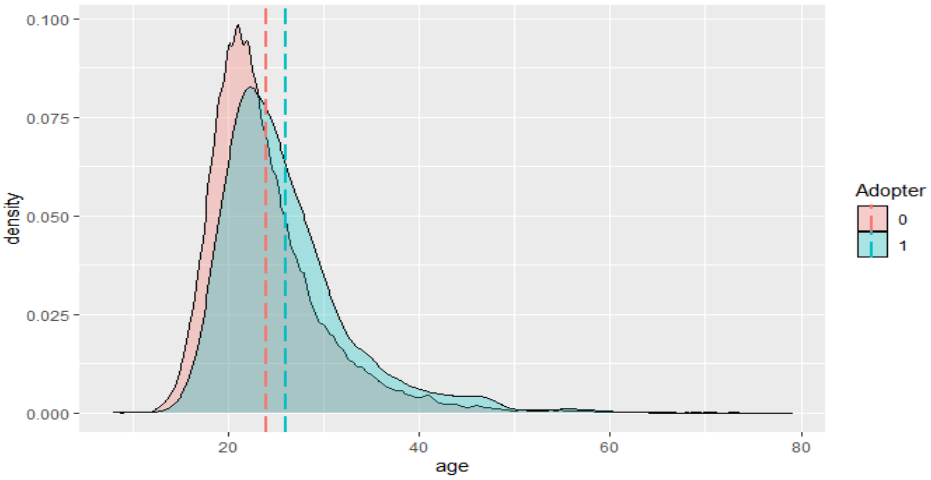
Answer 2a. Graphical Representation by Demographics – Age, country, gender

Adopter 0- free user, Adopter 1= paid user for all the graphs below.

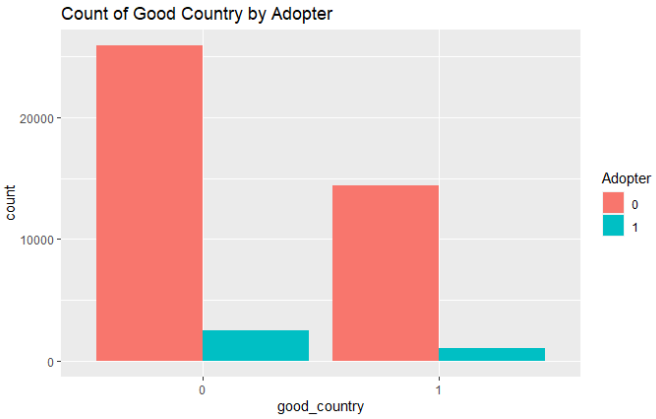
From the graphical representation of the demographic datapoints (age, gender, country), we can see–



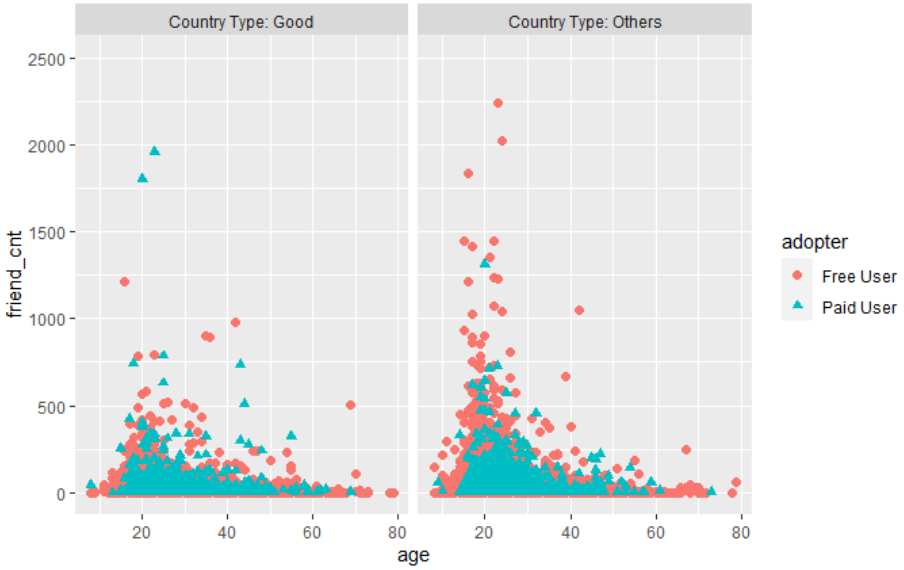
✓ Overall High Note platform has more male users than females, in both free and paid category.



✓ The average age of the free user is 23.95 years and that of the paid user is 25.96, indicating that free users would be relatively younger in age than the paid users. This could be based on their spending power earnings etc, lesser in age, no income and therefore availing more of free services. Overall, we can see the distribution too from the density versus age graph.



✓ Overall, users both free and paid seem to be higher in the rest of the world than in good countries (US, UK and Germany), potentially rest of the countries could be tapped more from business perspective (Free to fee model)



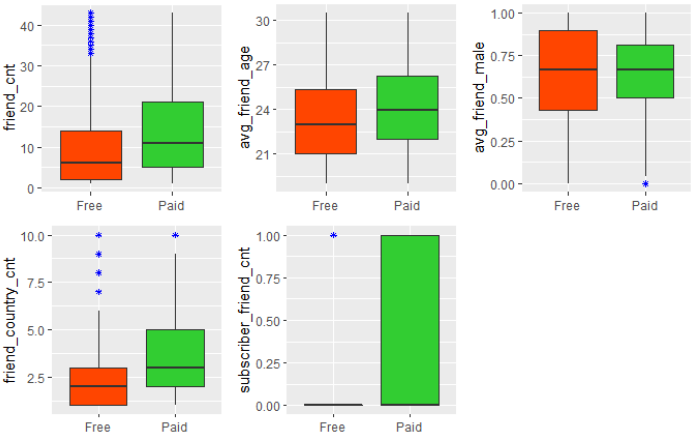
✓ This is an interesting plot of age versus friend count in the country indicating that the number of friends is relatively higher in rest of the world countries as compared to US, UK and Germany. Paid users are normally distributed along age specifically between 20 and 40 years.

Box Plots for Demographics to study the distribution across adopters and non adopters as well. Major points are already covered above.

2b. Descriptive Graphs for Peer Influence

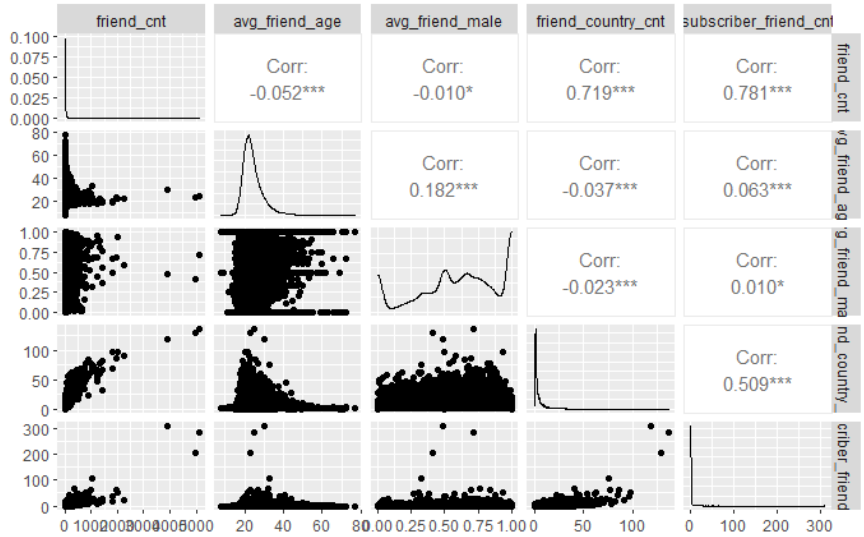
In terms of variables related to peer influence, it includes friend_cnt, avg_friend_age, avg_friend_male, friend_country_cnt, subscriber_friend_cnt etc.

Box graphs to understand the spread of the 'peer influence variables' across adopters and non-adopters. Mean of these variables are plotted.



Based on the above box plots, it seems that, the premium users have higher mean values of peer influence than free users have. This implies that premium users interact more with their friends and more likely to have a peer influence

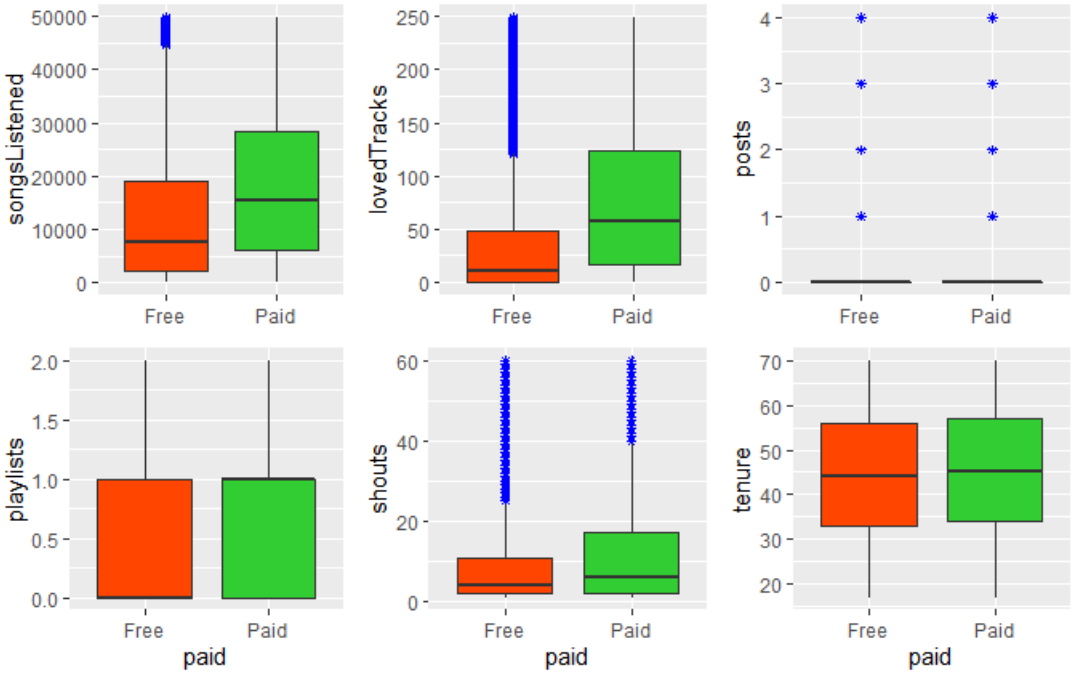
Plotted the graph below of peer influence variables



2c. Descriptive Graphs for User Engagement

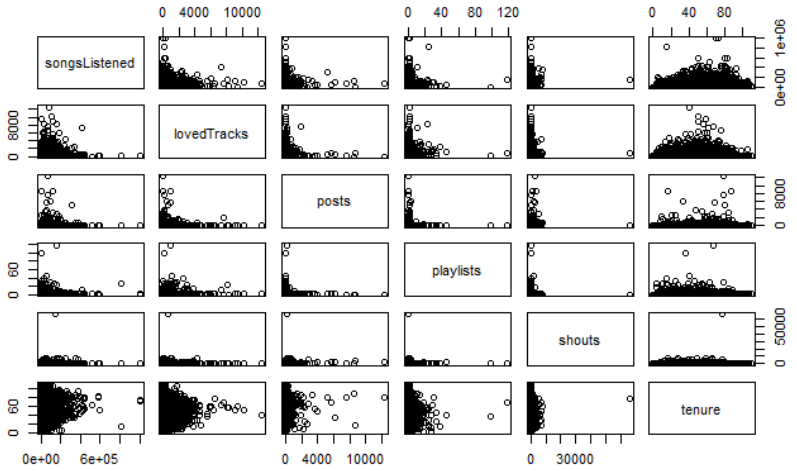
In terms of the user engagement variables, it includes songListened, lovedTracks, posts, playlists, shouts, and tenure. Below is a box plot of these user engagement variables.

- ✓ Paid users seem to have higher user engagement than free users, in terms of the key mean values of these variables. Looks like, the premium users are more active on the High Note music platform as compared to free.



Poonam Ahuja Student id no. : 31167068

Plotted the graph below of user engagement variables



Question 3: Propensity Score Matching (PSM): 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.

Answer: Propensity score matching (PSM) is done by running a logit model, where the outcome variable is a binary variable indicating treatment status. It helps understand covariates (Independent Variables), we should include for the matching to give a causal estimate. In the end, we need to include any covariate that is related to both the treatment assignment and potential outcomes.

Here, in our case, we define treatment as users who have one or more than 1 subscriber friends and study the variation of treatment across users from the two adopter classes (free and paid)

Before doing PSM, as suggested above, I split treatment group (1 or more subscriber friends) and control group (zero subscriber friends). Variation of treatment across users from the 2 adopter classes (free/paid users) is studied. Then, I did T-test to check if treatment group and control group have statistically significant difference on the behavior of being a free and premium user (adopter). Here, p-value is statistically significant.

So, treatment and control groups are different with regards to the adopter variable. The mean value of adopter for treatment group is 0.178 and that of control group is 0.052 or so. So, by doing PSM we will make two groups more balance.

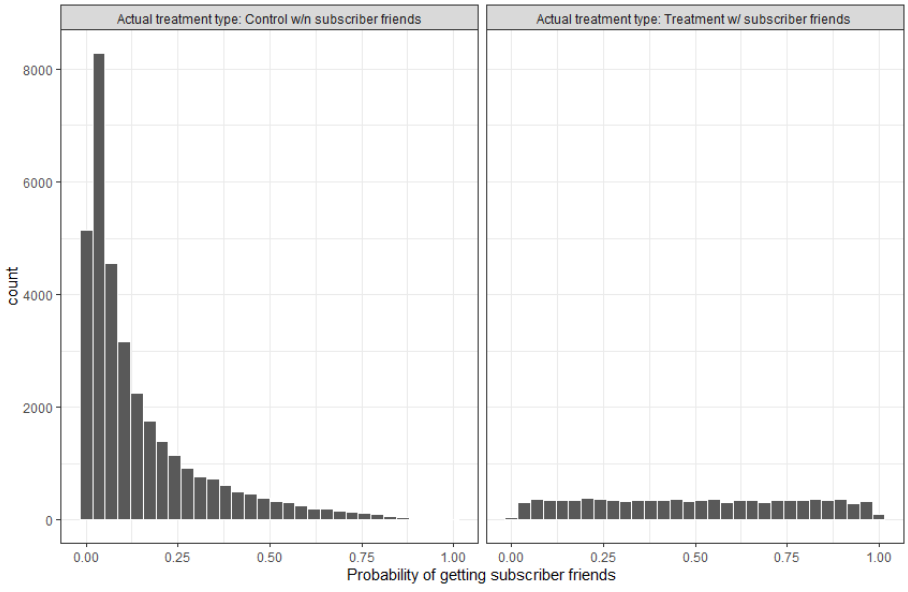

```
e. paid customer).
> #split treatment and control group
> data_desc <- read.csv("HighNote Data Midterm.csv")
> data_desc<- mutate(data_desc, treatment = ifelse(subscriber_friend_cnt>=1, 1, 0))
> data_desc %>%
+   group_by(adopter) %>% summarise(mean_treatment = mean(treatment),users=n())
# A tibble: 2 x 3
  adopter mean_treatment users
*   <int>          <dbl> <int>
1       0          0.200 40300
2       1          0.494  3527
>
> #run t-test before PSM
> with(data_desc, t.test(adopter ~ treatment)) #c-t: 0.052 - 0.178

        Welch Two Sample t-test

data:  adopter by treatment
t = -30.961, df = 11815, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1330281 -0.1171869
sample estimates:
mean in group 0 mean in group 1
     0.05243501      0.17754250
```

T test was done for each covariate by treatment status.

Logistic regression for PSM is done. To make sure all the other variables are significant to control group and treatment group, I did logistic regression for all variables, excluding variables related to treatment and adopter. We know that distribution of all continuous variables are skewed hence have taken log transformation before doing logistic regression. The result shows that, if the alpha of model is 0.05, then all variables are significantly correlated to the dependent variable - the treatment. AIC is 31493.

After logistic regression, we use the model to predict propensity scores of treatment and control group.

Propensity score by two groups was plotted - treatment and control group, which refers to the below distribution plot. It shows that the treatment group has consistent number of users among different levels of propensity scores. Yet the control group has right skewed distribution. It has more people with lower propensity score, who are less likely to turn into adopters ie. Paid users.



According to the pre-matching table(Refer R script) , some variables between treatment group and control group are significantly different, in terms of standardized mean difference (SMD) (*When SMD of a variable is larger than 0.1, we can say that this variable between two groups are different.). This is the case here as seen.

	Stratified by treatment 0	1	SMD
n	34004	9823	
age (mean (SD))	23.75 (6.22)	25.37 (6.97)	0.246
male (mean (SD))	0.63 (0.48)	0.64 (0.48)	0.015
friend_cnt (mean (SD))	10.43 (15.28)	54.02 (127.91)	0.479
avg_friend_age (mean (SD))	23.76 (5.06)	25.39 (5.17)	0.319
avg_friend_male (mean (SD))	0.61 (0.33)	0.64 (0.23)	0.079
friend_country_cnt (mean (SD))	2.73 (3.10)	9.39 (10.01)	0.899
songsListened (mean (SD))	14602.22 (23214.29)	33735.64 (43952.34)	0.544
lovedTracks (mean (SD))	65.21 (181.48)	225.36 (498.23)	0.427
posts (mean (SD))	2.54 (33.79)	20.52 (241.27)	0.104
playlists (mean (SD))	0.53 (0.97)	0.74 (1.96)	0.139
shouts (mean (SD))	16.42 (79.74)	101.82 (739.51)	0.162
tenure (mean (SD))	43.20 (19.72)	46.55 (19.92)	0.169
good_country (mean (SD))	0.35 (0.48)	0.34 (0.47)	0.024

```
>
```

Propensity Score Matching – It is simply the user's predicted probability of being Treated, given the estimates from the logit model. **Method used was nearest neighbor matching**. After finding differences in the treatment status, we use MatchIt package to find pairs of instances that have very similar propensity scores. The nearest method is to pair a treated subject to a control who is close in terms of its distance in covariate space.

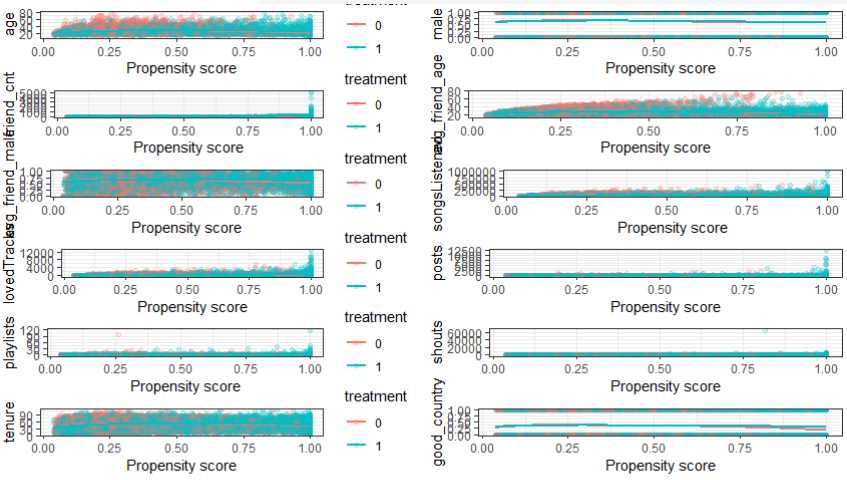
Post matching we see, that 9823 samples got matched. For all data without matching, the distances (= propensity score) between treatment and control are 0.4982 and 0.1450. However, the distances of matched data between treatment and control are 0.4982 and 0.3447. The mean difference of distance of the matched data is 0.1535, lower than the distance of all data, 0.3532. There are 56.54% balance improvement of mean different of distance.

We do a plot of matchit data. We can observe eQQ Plots. We see here, how post propensity score matching, the QQ plots are normalised for the covariants, whereas, all on the left side denoted before Matching *(Refer R script for the graphs)*

Create a dataframe that has the propensity score as well as the users actual treatment status. Then we do Mean difference test and the T test. I still got significant difference of all variables between control group and treatment group after nearest method matching. t-test between adopter and treatment also validates the same. So, I chose another matching method, the subclass method, to see if it can balance observed covariates of two groups(treatment & control) better.

We can observe that there are covariates whose values are matched between treatment and control, so we do a visual inspection of it.

Visual inspection to examine covariate balance in the matched sample. If the scatter points are located near the trend line, it means the matching is good. Yet the below plots show the nearest method still have a room for improvement of well matching the treatment and control samples in all covariates.



Propensity Score Matching – Method 2 used was subclassification (subclass). In a simplified manner, the subclass method is to put "similar'' observations (both treatment and control) into the same "subclass". With a lot of permutation and combinations of sub class, based on the low mean differences and mean distances of treatment and control, we get the final sub class.

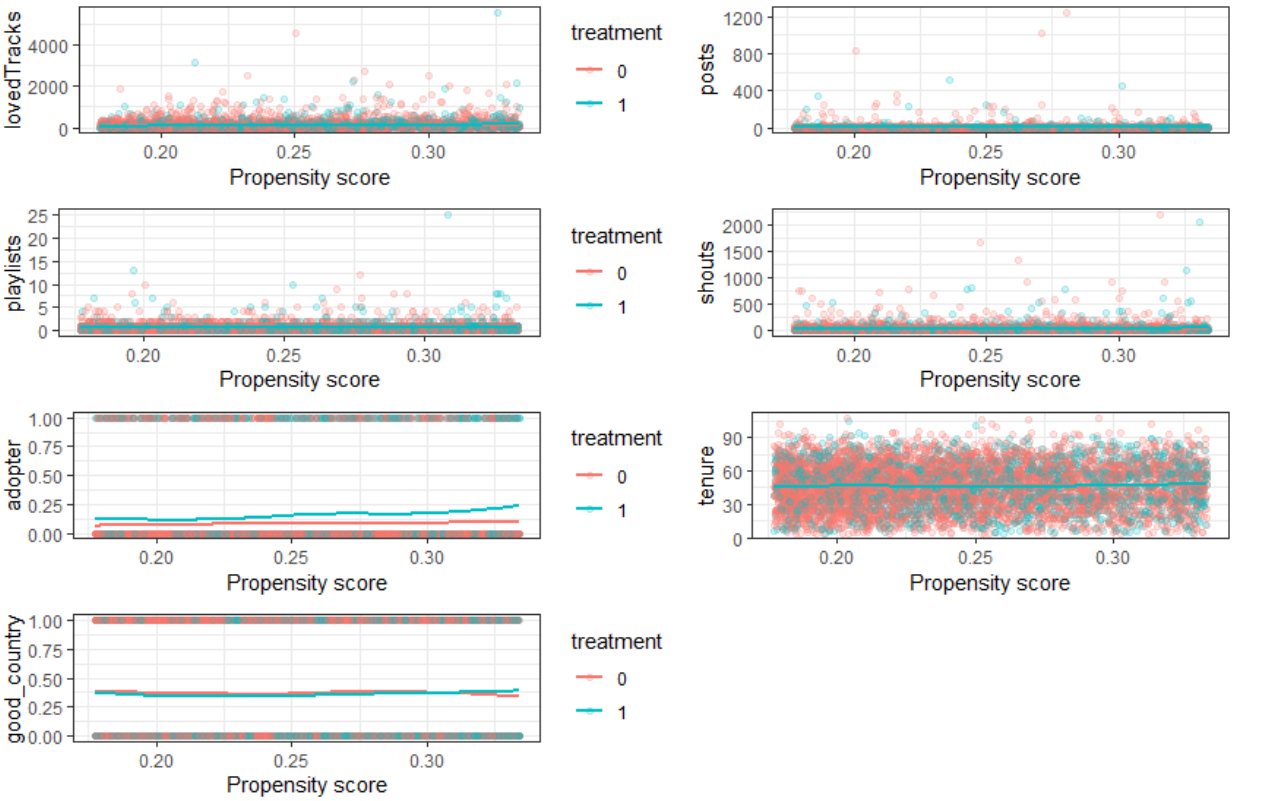
For subclass 2, their treatment and control samples are 5026 and 1637. Looking to the overall performance of all subclasses, we got the mean distances of treatment and control 0.4982 and 0.4792, respectively. The mean difference is 0.0092. It improves balance of two groups by 94.6247%, in terms of their distance (propensity score).

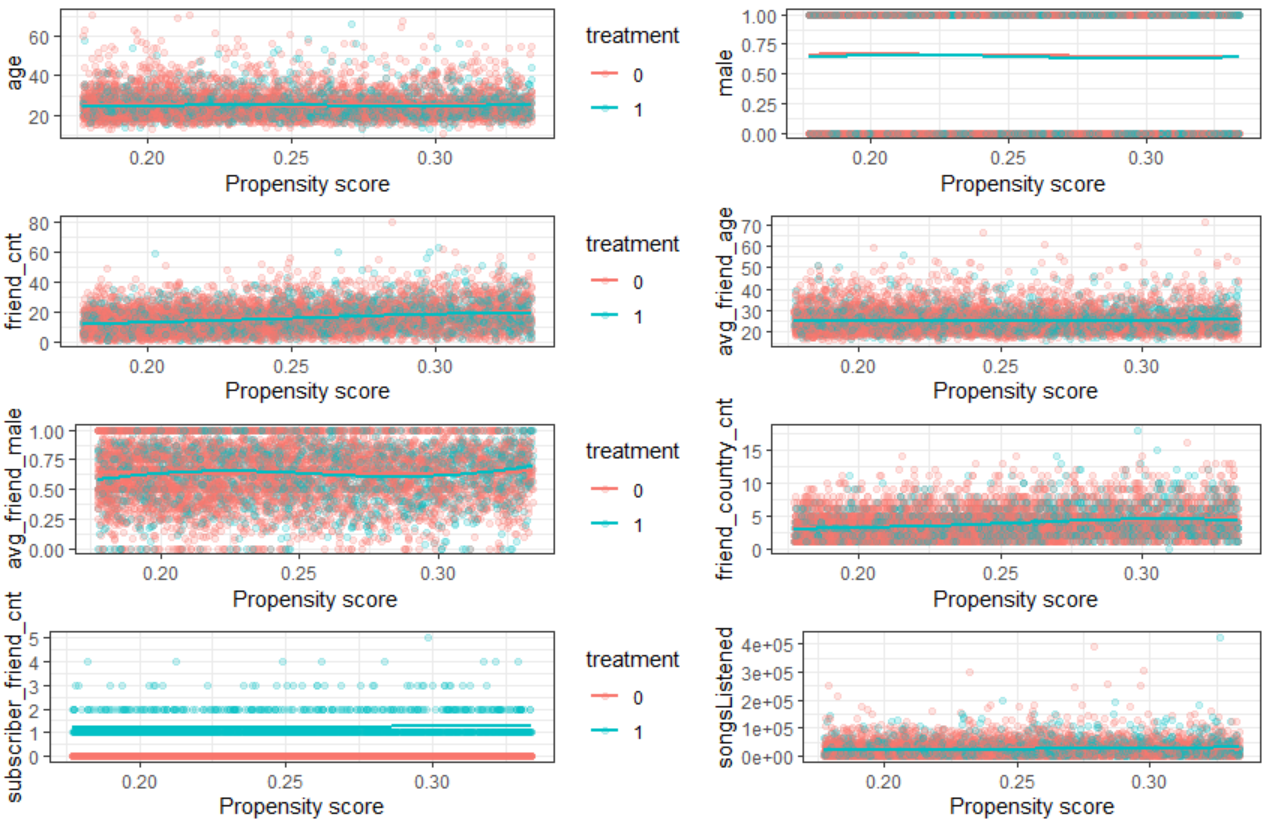
Mean difference for all variables was checked too. All the standardized mean differences, SMD, are lower than 0.1. It means that there is no huge difference between treated and control for all variables.

Looking at the t-test, all p-value in our matched subclass are not statistically significant (>0.05). So, there is no different between treatment and control group, which means that the observed covariates of two groups are balance.

On visual inspection we can see that most of points are located near the trend line. Graphs in the next page

Poonam Ahuja **Student id no. : 31167068**





Question 4:

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?

Answer: We build 4 models:

1. **Model 1:** With only the treatment variable to check if it is a significant predictor of the adopter variable.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5682  -0.4025  -0.4025  -0.4025   2.2599

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.47268    0.05266 -46.954   <2e-16 ***
treatment    0.73064    0.08712   8.387   <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: 4192.9  on 6662  degrees of freedom
Residual deviance: 4126.3  on 6661  degrees of freedom
AIC: 4130.3

Number of Fisher Scoring iterations: 5

> exp(coef(model_test1))
(Intercept)   treatment
 0.08435814  2.07640321
>
```

The treatment i.e. having subscriber friends is a significant predictor in assessing that if the user will be a free or a paid user. The odds ratio of the treatment is 2.07 which is greater than 1 and hence supports the point. Also as seen from the 3 stars against treatment, indicating that our treatment is statistically significant.

2. **Model 2:** Here we use all variables in dataset(continuous variables is taken log transformation) into logistic regression. Adopter (Free/paid) is the dependent variable.

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -0.877026   0.248840  -3.524 0.000427 ***
log(age)                     0.076590   0.026307   2.911 0.003611 **
male                         0.022135   0.007869   2.813 0.004922 **
log(friend_cnt + 1)          0.038629   0.016762   2.304 0.021226 *
log(avg_friend_age + 1)      0.159405   0.058842   2.709 0.006766 **
log(avg_friend_male + 1)     0.032685   0.024377   1.341 0.180029
log(friend_country_cnt + 1)  0.015553   0.010934   1.423 0.154927
log(subscriber_friend_cnt + 1)  0.091302   0.010156   8.990  < 2e-16 ***
log(songsListened + 1)       0.014650   0.002539   5.769 8.31e-09 ***
log(lovedTracks + 1)         0.025319   0.002260  11.205  < 2e-16 ***
log(posts + 1)               0.011735   0.004238   2.769 0.005640 **
log(playlists + 1)           0.031406   0.009318   3.370 0.000755 ***
log(shouts + 1)             -0.017596   0.003401  -5.174 2.36e-07 ***
log(tenure + 1)             -0.039041   0.009375  -4.164 3.16e-05 ***
good_country                -0.036847   0.007424  -4.963 7.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2854 on 6648 degrees of freedom
Multiple R-squared:  0.05756,   Adjusted R-squared:  0.05557
F-statistic:    29 on 14 and 6648 DF,  p-value: < 2.2e-16
```

The summary result shows that friend_cnt, avg_friend_male_log and friend_country_cnt_log are not significant, based on the alpha = 0.05. So, I removed these two variables in the next logistic regression model.

Model 3:

Here in this model, we have included all variables which are statistically significant. This is enhanced version of model 2. Looking into the significance we have calculated the odds ratio as well.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3804  -0.4839  -0.3602  -0.2528   3.0144

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -8.35850    1.05981  -7.887 3.10e-15 ***
log(age)                    0.54983    0.29232   1.881 0.059988 .
male                        0.21807    0.10072   2.165 0.030384 *
log(avg_friend_age + 1)     0.77195    0.39231   1.968 0.049101 *
log(subscriber_friend_cnt + 1)  0.95732    0.10775   8.885  < 2e-16 ***
log(songsListened + 1)      0.21621    0.03871   5.585 2.34e-08 ***
log(lovedTracks + 1)        0.30735    0.02752  11.168  < 2e-16 ***
log(posts + 1)              0.10442    0.04962   2.104 0.035352 *
log(playlists + 1)          0.35726    0.10243   3.488 0.000487 ***
log(shouts + 1)            -0.19948    0.04026  -4.955 7.24e-07 ***
log(tenure + 1)            -0.35799    0.09418  -3.801 0.000144 ***
good_country               -0.51378    0.09576  -5.365 8.09e-08 ***
---

> exp(coef(model_test3))
                 (Intercept)                     log(age)                         male
              0.0002343951                1.7329544235                1.2436714857
      log(avg_friend_age + 1) log(subscriber_friend_cnt + 1)        log(songsListened + 1)
              2.1639782822                2.6047103007                1.2413691946
         log(lovedTracks + 1)                log(posts + 1)           log(playlists + 1)
              1.3598109624                1.1100645864                1.4294101712
           log(shouts + 1)               log(tenure + 1)                good_country
              0.8191596959                0.6990774327                0.5982306131
>
```

To decode this in a simplified manner, for every 1 unit increase in log(age), adopter(free/paid) will increase by the factor of exp()

Ie. For every 1 unit increase in age, adopter will increase by a factor of 1.732. For every 1 unit increase in male, adopter will increase by a factor of 1.243. For every 1 unit increase in average friend's age (avg_friend_age), adopter will increase by a factor of 2.163. For every 1 unit increase in subscriber friend count (subscriber_friend_cnt), adopter will increase by a factor of 2.604. For every 1 unit increase in songListened, adopter will increase by a factor of 1.241. For every 1 unit increase in loveTracked, adopter will increase by a factor of 1.359. For every 1 unit increase in posts, adopter will increase by a factor of 1.110. For every 1 unit increase in playlists, adopter will increase by a factor of 1.429 For every 1 unit increase in shouts, adopter will decrease by a factor of 0.819. For every 1 unit increase in tenure, adopter will decrease by a factor of 0.699. For every 1 unit increase in good country, adopter will decrease by a factor of 0.598**.**

In a nutshell, the age, male, friend average age, subscriber friend count, song listened, love tracked, posts, and playlists are the variables which has positive correlation with adopter. The variables of shouts, tenure, good country are negatively correlated with adopter, which means that with these three variables increasing overall, it is less likely to become paid users (adopter = 1).

Model 4 - Model 3 with original data

We build a model with the same significant predictors from model 3 but with the original dataset which has not been matched by propensity score, to try finding if the odds ratio are similar.

We find that the odds ratio is different for few parameters in model 4(original data set) as compared to the model 3, where we have balanced with the propensity score matched data . This indicates that it is important to do a PSM else we might get wrong interpretations and thereby wrong business interventions, thereby impacting business effectiveness overall. Key Observations on differences are mentioned below.

```
Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -9.33103    0.36540 -25.536  < 2e-16 ***
log(age)                    0.80080    0.11754   6.813 9.55e-12 ***
male                        0.33243    0.04332   7.674 1.67e-14 ***
log(avg_friend_age + 1)     0.70790    0.15414   4.593 4.38e-06 ***
log(subscriber_friend_cnt + 1)  0.63427    0.03409  18.607  < 2e-16 ***
log(songsListened + 1)      0.22801    0.01401  16.271  < 2e-16 ***
log(lovedTracks + 1)        0.29762    0.01127  26.419  < 2e-16 ***
log(posts + 1)              0.12409    0.01754   7.075 1.49e-12 ***
log(playlists + 1)          0.15065    0.04308   3.497 0.000471 ***
log(shouts + 1)            -0.12971    0.01568  -8.270  < 2e-16 ***
log(tenure + 1)            -0.32347    0.03960  -8.168 3.15e-16 ***
good_country               -0.48338    0.04165 -11.606  < 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: 24537  on 43826  degrees of freedom
Residual deviance: 20710  on 43815  degrees of freedom
AIC: 20734

Number of Fisher Scoring iterations: 6
```

Below is the odds ratio of model 4

```
> exp(model_test4$coefficients)
                 (Intercept)                      log(age)                         male
                8.863118e-05                  2.227333e+00                  1.394351e+00
      log(avg_friend_age + 1) log(subscriber_friend_cnt + 1)        log(songsListened + 1)
                2.029731e+00                  1.885650e+00                  1.256097e+00
         log(lovedTracks + 1)                log(posts + 1)            log(playlists + 1)
                1.346656e+00                  1.132119e+00                  1.162586e+00
              log(shouts + 1)               log(tenure + 1)                  good_country
                8.783480e-01                  7.236365e-01                  6.166976e-01
>
```

Odds ratio of model 3

```
> exp(coef(model_test3))
                  (Intercept)                       log(age)                          male
                1.0002343951                1.7329544235                1.2436714857
      log(avg_friend_age + 1) log(subscriber_friend_cnt + 1)       log(songsListened + 1)
                2.1639782822                2.6047103007                1.2413691946
         log(lovedTracks + 1)                log(posts + 1)           log(playlists + 1)
                1.3598109624                1.1100645864                1.4294101712
              log(shouts + 1)               log(tenure + 1)                good_country
                0.8191596959                0.6990774327                0.5982306131
> |
```

Key Points between odds ratio of model 3 and 4:

1. Variables like male, average friend's age, songs listed, lovedTracks, posts, playlists have almost similar odds ratio in both model 3 and 4
2. Variables like shout, tenure, good country, which were showing a negative correlation with the adopter in model 3, are now showing a positive impact in model 4. This could be misleading and will impact the overall business interventions, hampering business growth overall, if the dataset is taken as it is.

So overall, we think model 3 – with the matched dataset including all the significant variables. is better and serves our purpose as well to design the right marketing interventions for bringing about business growth for High Note Platform.

Question 5:

Key Takeaways for a "free-to-fee" strategy for High Note?

Answer: Looking at the analysis done by Descriptive statistics, graphical representation based on demographics, peer influence and user engagement and Regression it is seen that –

Demographic variable impact:

- ✓ Male and age are positively corelated to the adopter ie males in higher age range are more likely to become a paid user on High Note. So, it is a good strategy for High Note, to do STP (segmentation, targeting and positioning) based on this. With focused targeted marketing with this segment, will lead to more paid users.
- ✓ Looking at the user trend in US, UK and Germany (Good country) as compared to the rest of the countries and also from data which shows that if the user is from US, UK and Germany, the odds ratio of the user converting to paid is lesser, it is recommended that High Note should expand its international markets, beyond US, UK and Germany. It is also seen that paid users and overall free users are also are higher from other countries beyond US, UK, Germany. This could be due to a high competitive intensity in these countries, Therefore, it is suggested that High Note should explore these other international frontiers beyond US, UK and Germany. With a good peer influence and user engagement drive as mentioned below with a customized and targeted marketing with value added packages/offers, High Note can improve " Free to Fee" conversions.

Peer influence variable impact:

- ✓ Overall, paid users have a higher level of peer influence than free users, indicating that premium users interact more with their friends and more likely to be affected by their friends. Subscriber friends overall, has a positive effect in increasing the odds of becoming a paid customer. Also, Friend Country count has a positive effect on the odds ratio of becoming an adopter, which means that these users have more friends from other

countries (globally diverse person) and with their engagement, High Note can further increase its paid users globally as well.

- ✓ In a nutshell, if a user actively interacts with lots of subscriber friends who are in the "targeted age range" of High Note, the user is more likely to continuously engage and become a paid user.
- ✓ So, High Note can create more marketing and engagement activities among users – viral marketing, create coupons or group packages, referral schemes etc. to create more buzz amongst the users and further boost peer influence and thereby convert free users to premium users.

User engagement variable impact:

- ✓ User engagement activities viz. songs listened, loved tracks, playlists etc. all have a positive odds ratio on adopter binary variable (i.e. it increases the paid users) Looking at this and the graphical representations which show that paid users have higher user engagement than free users, High Note could create more interactions (number of songs offered, explore customization via regional songs etc) between paid and free users which will influence the free users to become paid. Engagement drives could be via various PR campaigns on various aspects on sharing/re-sharing listened songs, loved tracks or playlists. This will keep them hooked onto the platform and they will get used to using not only the basic features of the platform but also try using the paid features. Also, a great move would be to have "paid user" sharing experiences which will further have a "ripple effect" not only amongst the paid user category but also in the free category. This is similar to an advocacy which can be done by this loyal and highlight engaged customer. They will be motivated to try the paid features and then with continuous engagement will become "loyal paid users".
- ✓ Also, it was very evident from data that as the user tenure increases, the likelihood of converting to a paid customer reduces, so High Note should make focused efforts and target the user in the initial period itself with more advertisements, user engagement drive etc. to get them hooked on to High Note platform, in the initial phase itself.
- ✓ Recommendation to High Note should be to continuously revamp not only the functions or services on their platform but also build good engaging content to create the "marketing pull" needed and for users to get glued on.

In summary, age, male, friend average age, subscriber friend count, song listened, love tracked, posts, and playlists are the variables which show a positive correlation with adopter ie. More these variable values can be enhanced via marketing and engagement activities, better success of High Note for the "Free to Fee" conversion model. The variables of shouts, tenure, good country (users from US, UK, Germany) are negatively correlated with adopter, which indicates that the higher these three variables are, it is less likely to become paid users. So, High Note should intervene and target the right markets and engage with users in the initial period itself for better success to get paid users.