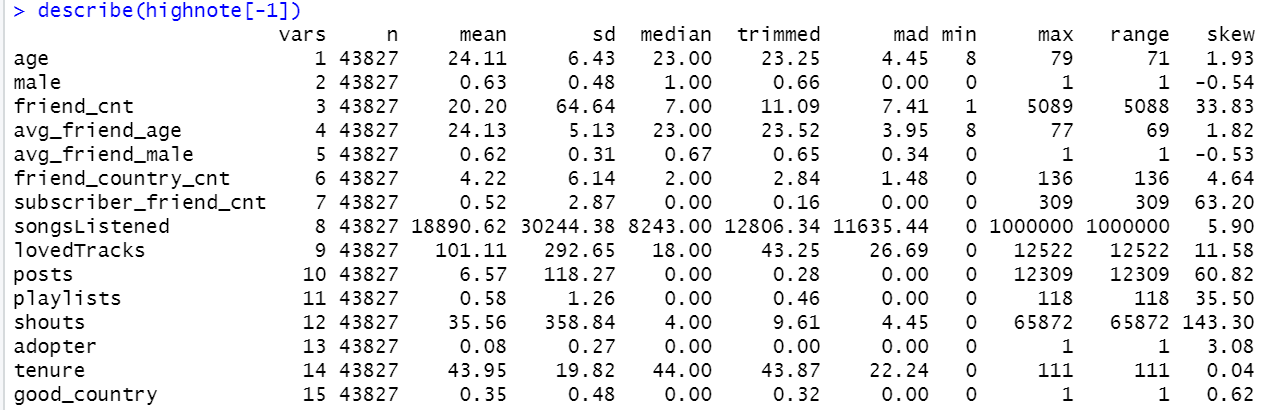
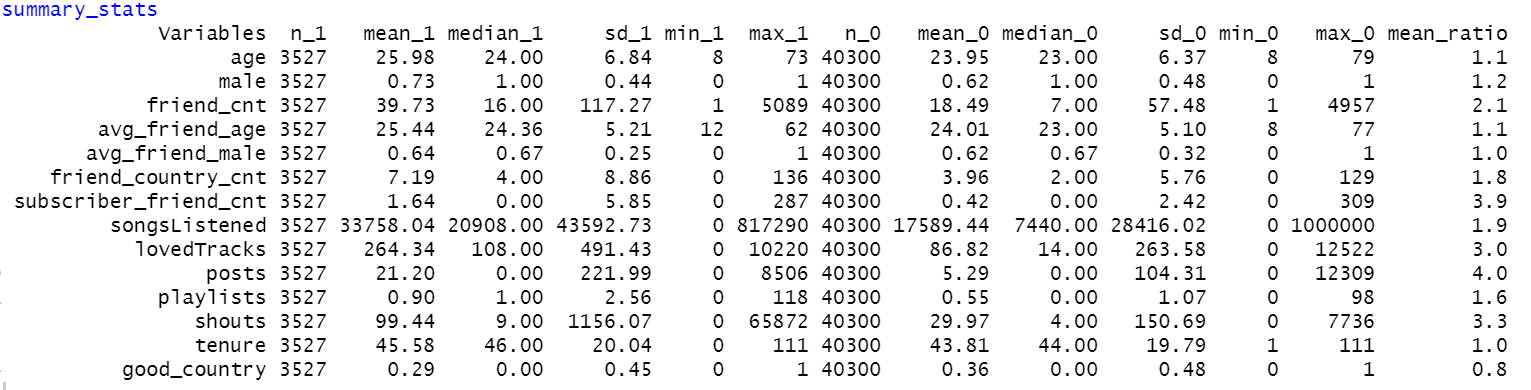
#Summary statistics: Generate descriptive statistics for the key variables in the data set, similar to the table on the last page of the case. (Note that your table will look different because the data set you are analyzing is different from the one used to generate the table in the case.) Analyze the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

The HighNote data has 43827 rows and 16 variables. As shown in the table below, the summary statistics is run on the full data:



For most of the variables there is a big difference in Mean and Median and skewness is >1 or <-1 which implies that the data is skewed

Summary stats on adopter variable is shown below:



1 -> adopter and 0 -> non adopter and the ratio is the ratio of mean\_1 and mean\_0

After comparing the adopter and non-adopter subsamples, it could be concluded that adopter group has following attributes:

Is 2 years older, has 20% more male and not from US, UK or Germany (80% less), spend longer time on the site (2 more months), listened to more songs (90%), love more tracks (3 times), have more posts (4 times), playlist (2 times), and shouts (3 times) compared to non adopters

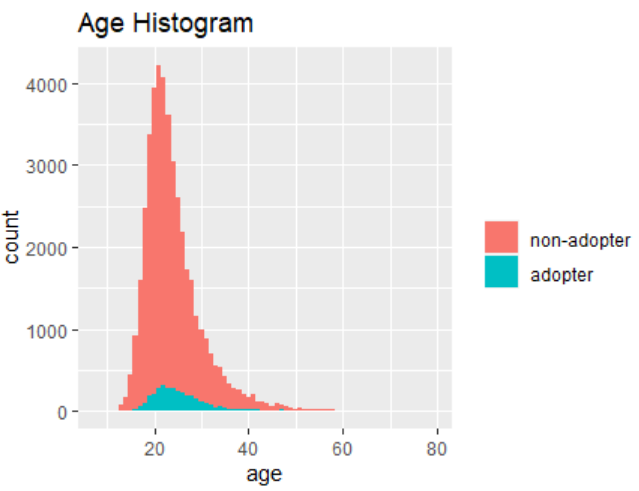
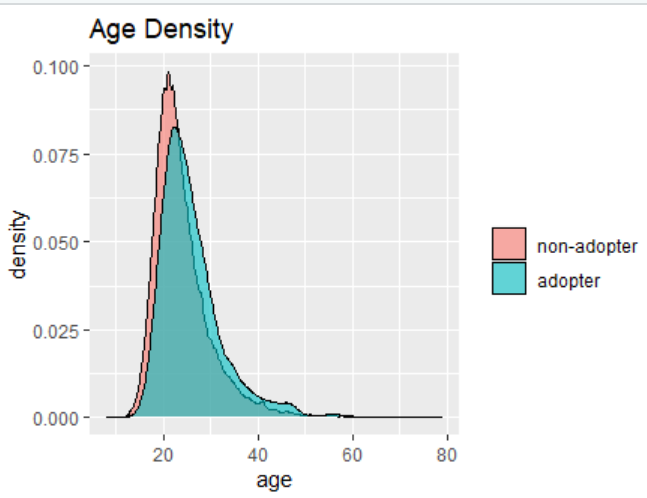
subscribers have on average 111% percent more friends, average friend's age is 1 year older and subscriber’s friends are 290% more likely a subscriber, subscribers consume 90 percent more music, subscribers create 60 percent more playlists, they choose to mark 200 percent more tracks as loved, and they create 230 percent more shouts, subscribers write 300 percent more posts on the site's forums on average than non adopters. Overall, subscribers tend to be more active on the site than the free users.

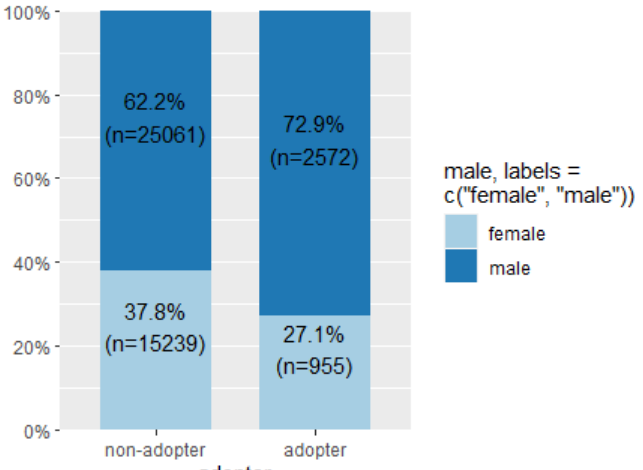
The t tests between adopter and all other variables shows that the difference between the means of adopter and each covariate is statistically significant since the p value is significant at 95% confidence

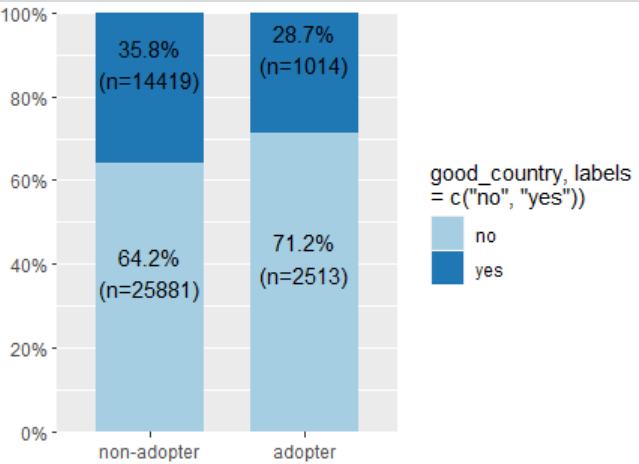
Data Visualization: Generate a set of charts (e.g., scatter plots, box plots, etc) to help visualize how adopters and non-adopters (of the premium subscription service) differ from each other in terms of (i) demographics, (ii) peer influence, and (iii) user engagement. What can you conclude from your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

**Demographics**

The age distribution between non-adopter and adopter is almost the same. The average age of adopter is slightly higher than non-adopter. In the adopter group, the proportion of male is larger in adopter group than non-adopter group. It can be concluded that the typical adopter is older male and does not belong to the good country from the perspective of demographics.

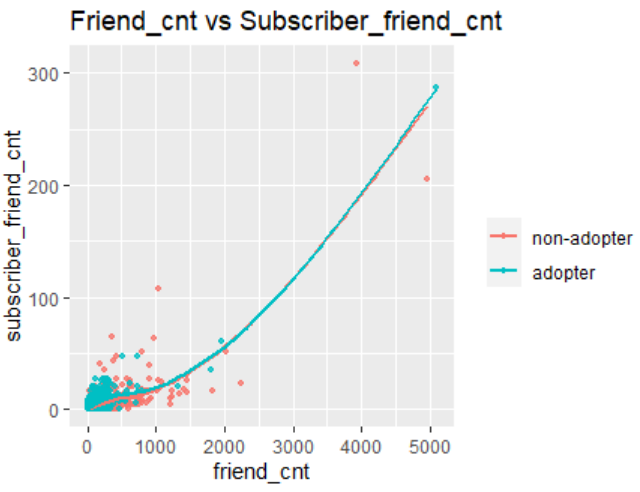
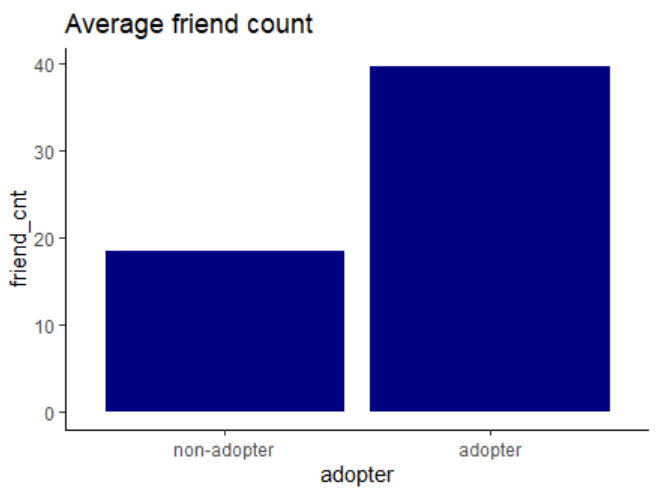


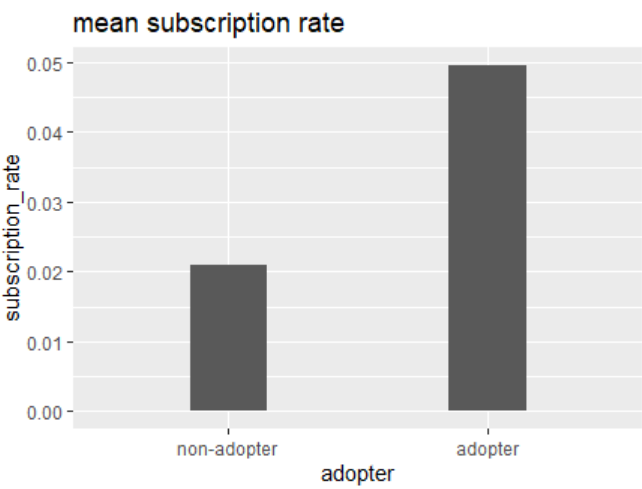
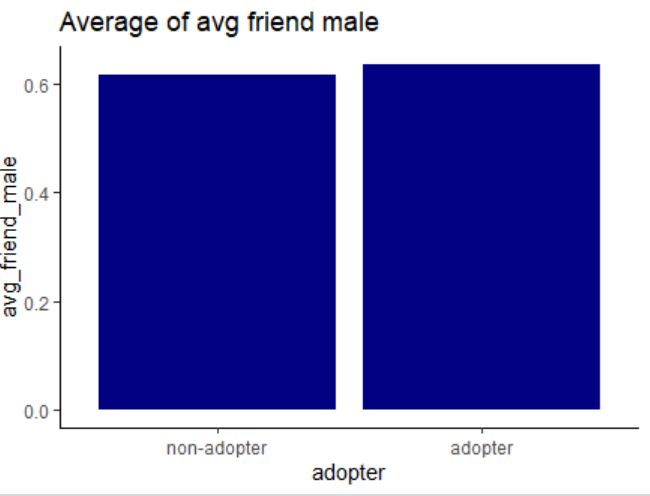
**Gender and Country Distribution:**

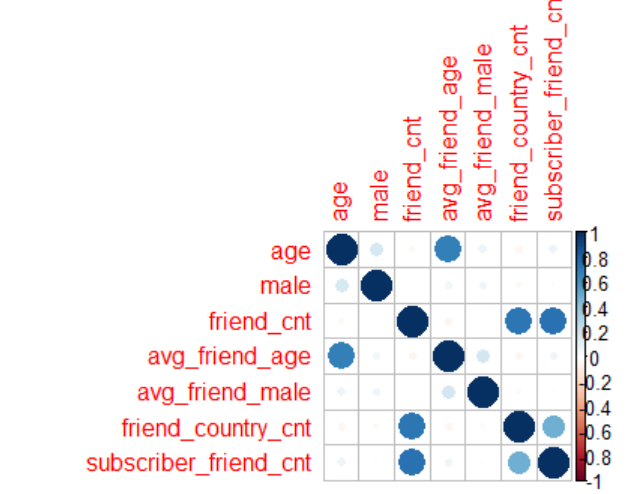


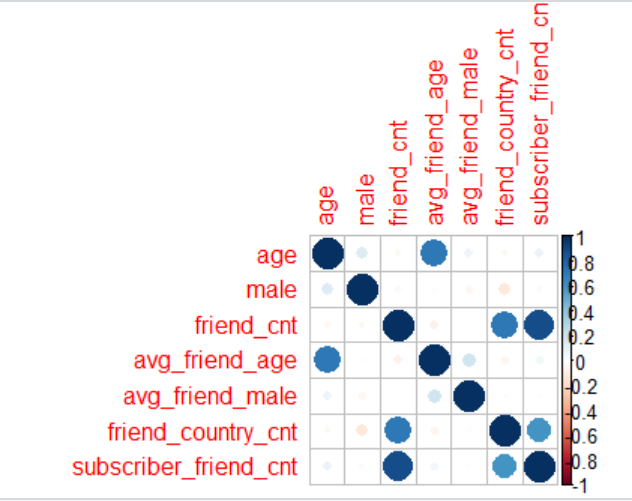
**Peer influence:**

If the subscription rate of friends is defined as subscriber\_friend\_cnt/friend\_cnt,the subscription rate of adopter group is more than 2 times higher than non-adopter group. This means a friend of adopter is more likely to be a subscriber. Also, the relationship between the number of friends and the number of friends who are premium subscribers is close to linear.





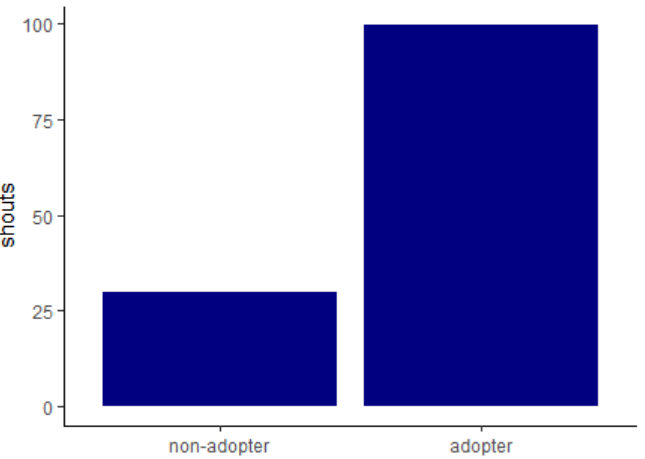
**Adopter and Non Adopter correlation plots**

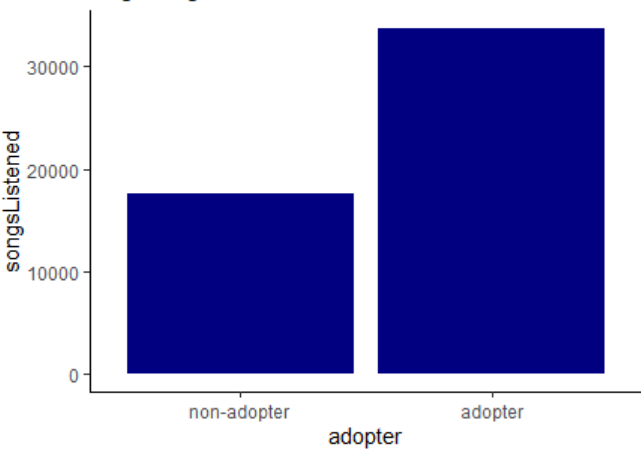
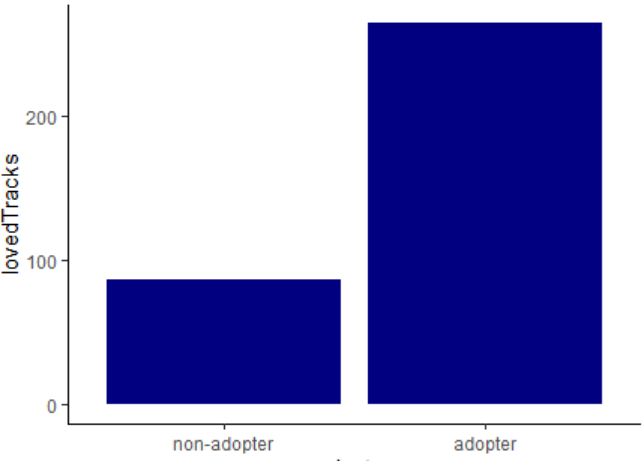


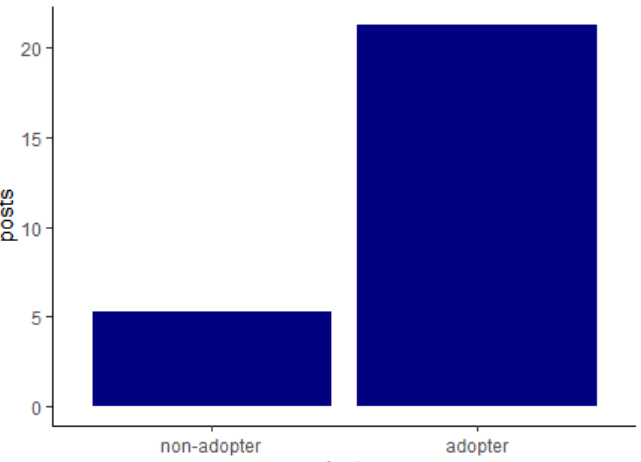
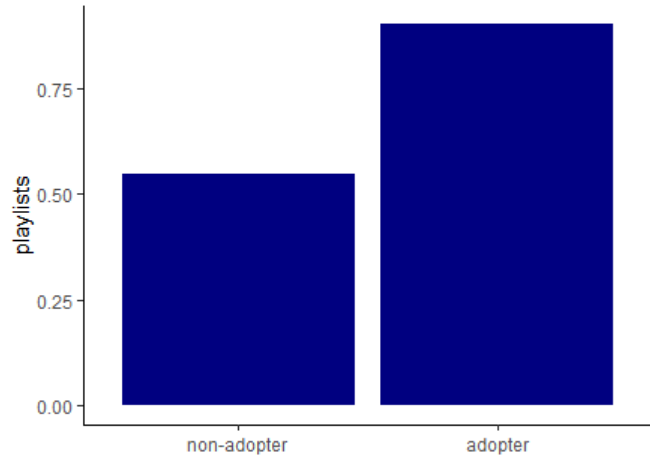
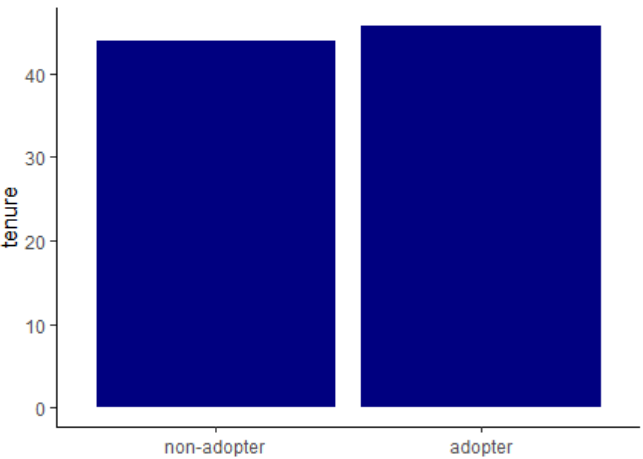
For **adopter** age has 71% correlation with avg\_friend\_age, friend\_cnt has 71% correlation with friend\_country\_cnt and 88% correlation with subscriber\_friend\_cnt. Also,friend\_country\_cnt has 59% correlation with subscriber\_friend\_cnt.

For **non-adopter** age has 68% correlation with avg\_friend\_age, friend\_cnt has 72% correlation with friend\_country\_cnt and 74% correlation with subscriber\_friend\_cnt.Also friend\_country\_cnt has 48% correlation with subscriber\_friend\_cnt. This high correlation implies that there the peer influence exists

**User engagement:**

In each aspects of user engagement, adopter group is higher than non-adopter group **on average**. It could be assumed that the more user engaged on the site, the more likely he/she will become adopter. Also, the adopter tend to be more engaged in every aspect with the site.

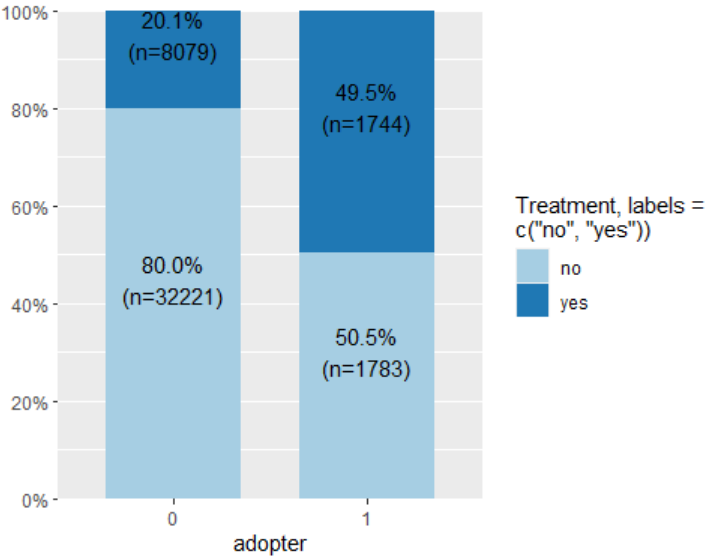




\

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.

After creating the column for the treatment. The below graph shows that adopter column has equal distribution of treatment column but the non adopter column has 80-20 split for the treatment. The t test between the treatment and adopter columns has p-value <2e-16, which implies that there is a significant difference in means

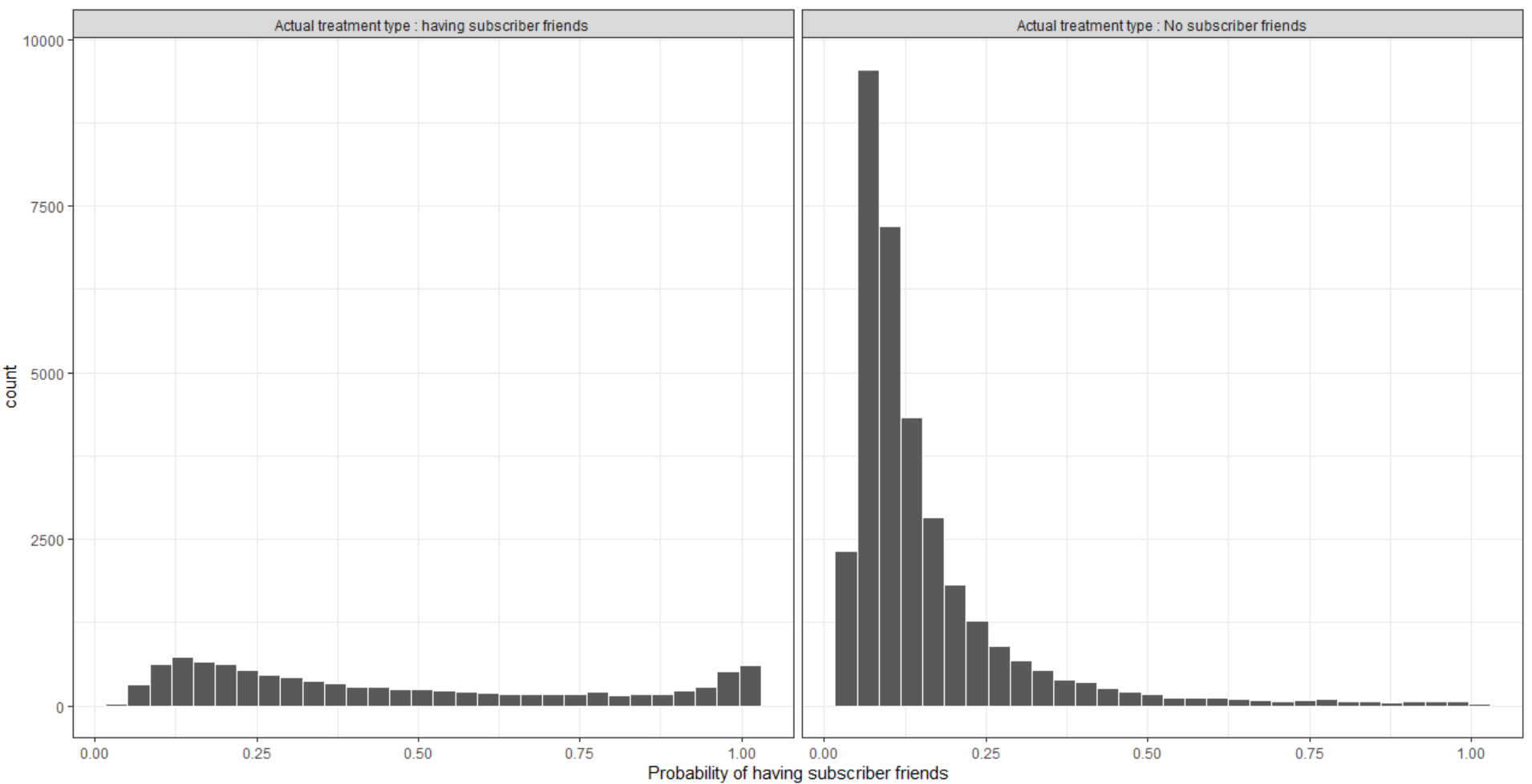


Now, I tested the difference in means on the pre-treatment covariates. Friend\_cnt, friend\_country\_cnt, songlistened, lovedtracks,playlists,posts, shouts and tenure are the variables with significant difference in the means between treatment and control groups in terms of mean values

After performing the t test between the covariates and the treatment column: Considering 0.05 significance level, the diff-in-means of the demographic variable "male" is insignificant with p value of 0.2.

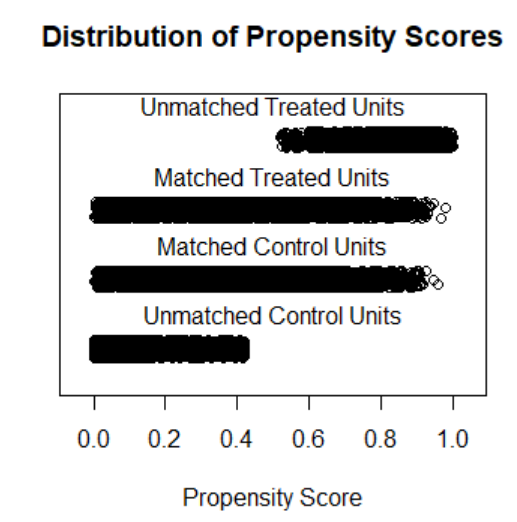
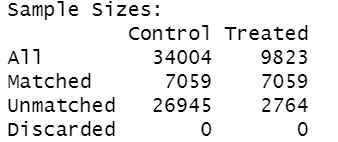
I estimated the propensity score by running a logit model.

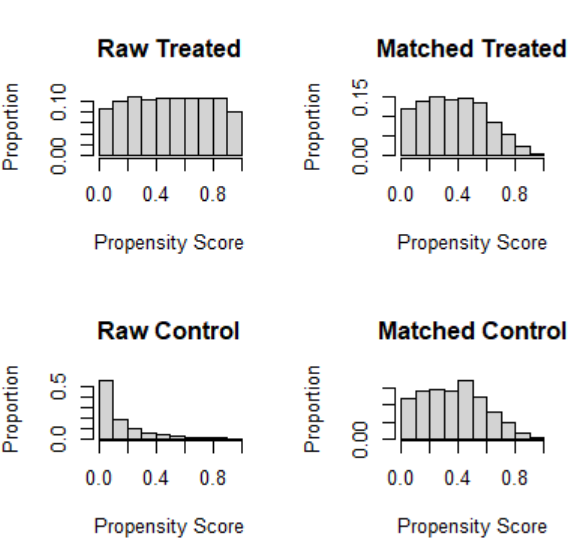
To get a visual idea, I plotted the propensity scores for the treatment and control group. As one can see that there is a difference in frequency distribution between the test and control groups



Next, I used the package MatchIt to run a matching algorithm on the covariates. The matched data is put into a new dataframe called matched\_data. This table has a column called "distance", which gives us the propensity score.

The covariates were log transformed before running the matching algorithm to reduce the skewness. Below is the matched samples numbers. As can be observed in this figure, the section labeled “Unmatched control Units” shows that most of the non-matched individuals were in the lower (0.0 to 0.4) part of the propensity scores.



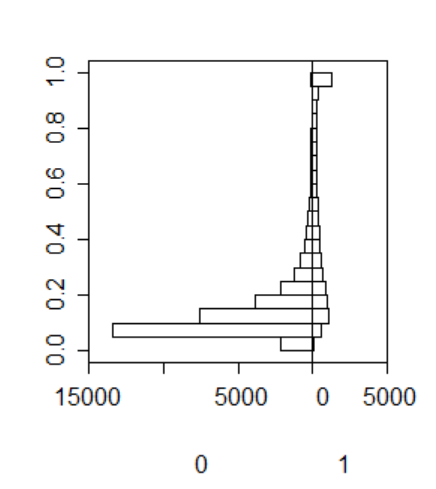
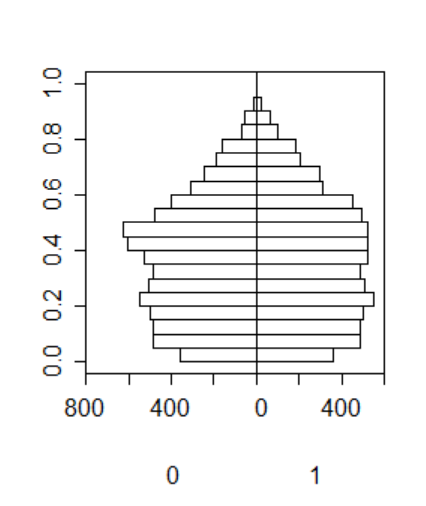


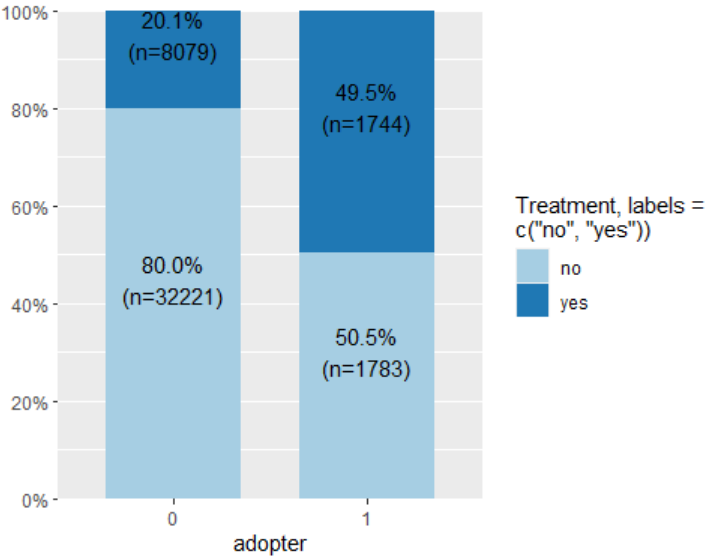
After running the chi square test on the matched\_data, the p value of 0.091 shows that the association between the variables is not significant. The p-value <2e-16 from t test between treatment and adopter columns implies that the means are stastically significant and after controlling all the covariates, the treatment effect is still significant. It is reasonable to claim that the causal relationship exist between having subscriber friends and the likelihood of becoming an adopter. To be specific, having subscriber friends will increase the likelihood of becoming an adopter. To get a numeric idea, I now run a t-test against for all the variables to see if there is still a significant difference in means of the variables after matching. At 95% confidence, I can say that p values are greater than 0.05 for all the variables which implies that the difference in the means is not statistically significant. After comparing the means between adopters and non-adopters I conclude that, the mean of number of non-adopters has increased, while that of number of adopters remains the same. This could be explained in a way that, having no treatment increases the chances of a subscriber to be a non-adopter. However, it does not mean that having treatment increases chances of being an adopter.

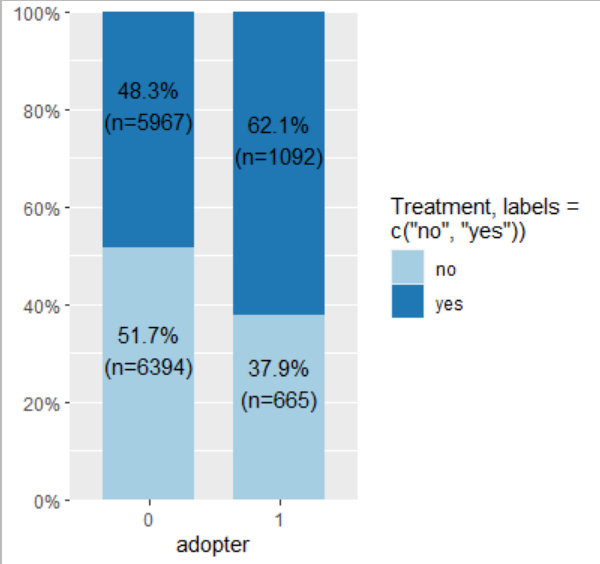
Comparing the distribution of the treatment variable before and after matching:

0 -> has zero subscriber friend and 1 -> has more than 1 subscriber friends

The below graph shows that the treatment and control groups have similar distributed probabilities now.

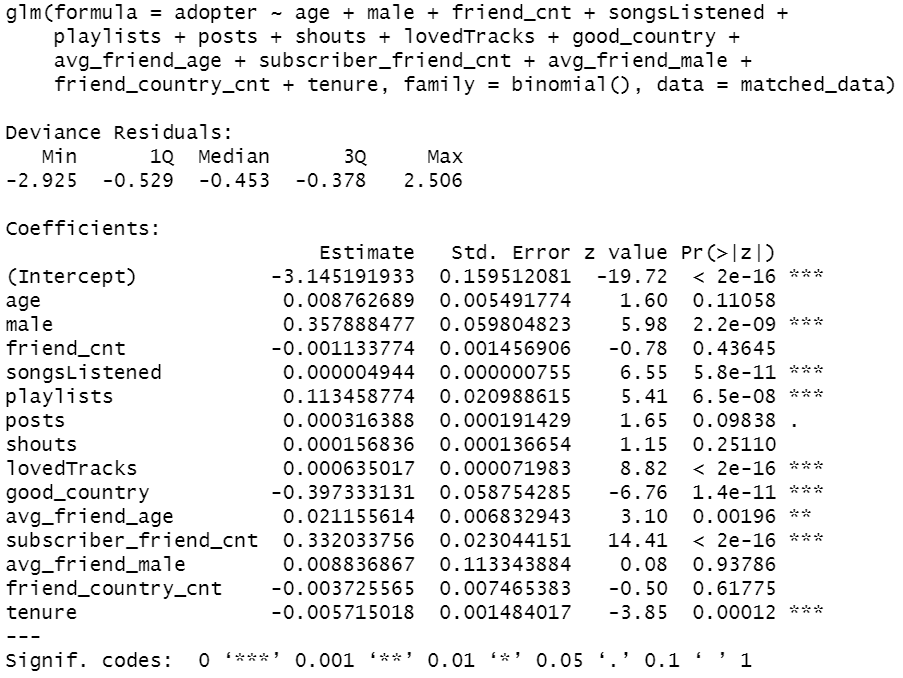
**Before : Y axis is probability scores After: Y axis is propensity scores**

**The distribution of observations between treatment and control groups before and after respectively:**

****

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?

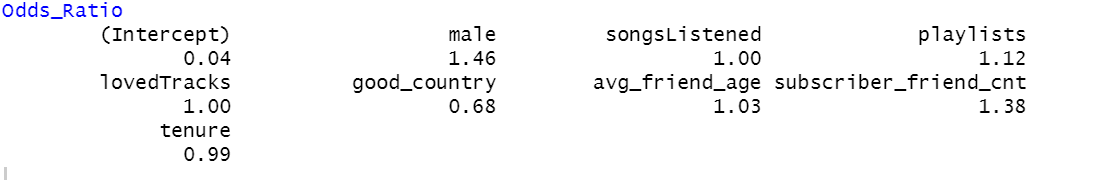
After running the logistic regression on the matched\_data, here are the results:

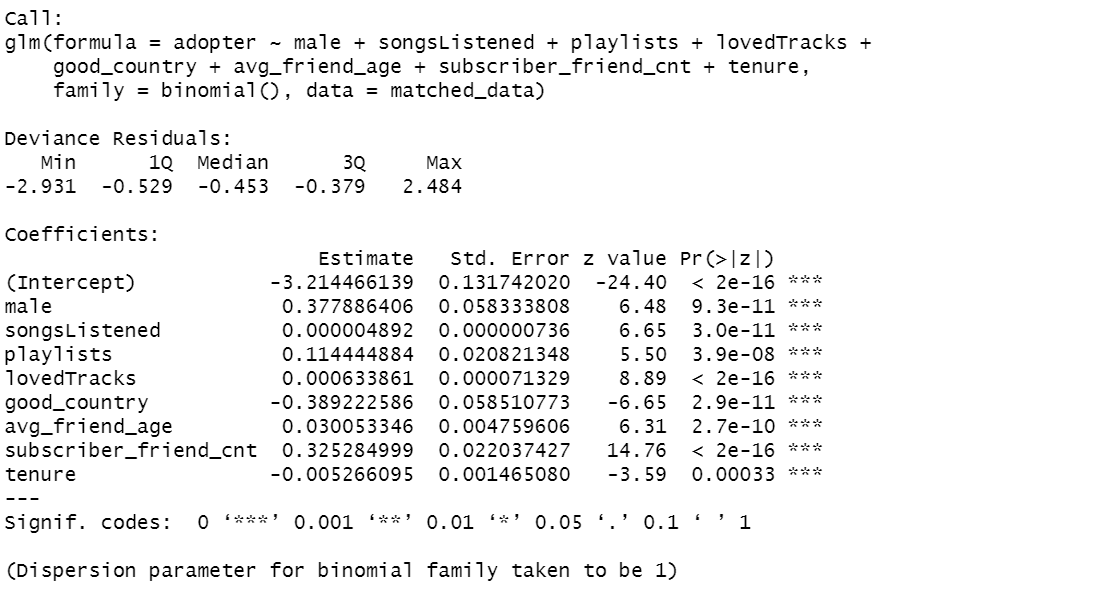


At 95% confidence, after removing the insignificant variables and re running the logistic regression on the matched data, here are the results:

Variables male, playlists, avg\_friend\_age, subscriber\_friend\_cnt will increase the odds of being adopter . Good\_country and tenure reduce the likelihood of being an adopter. Variables lovedtracks and songlistened also have positive impact on increasing the likelihood of being an adopter but the effect is small

Male, Good\_country and subscriber\_friend\_cnt are the variables with high coefficient values which implies that the males with high number of friends who are subscribers and outside the good countries will lead to increase in the probability of being an adopter. And as intercept coefficient is negative that using only adopter will reduce the likelihood of being an adopter





I ran VIF to check the importance of these variables and got the values less than 5 which implies that these variables are significant in predicting the likelihood of becoming an adopter

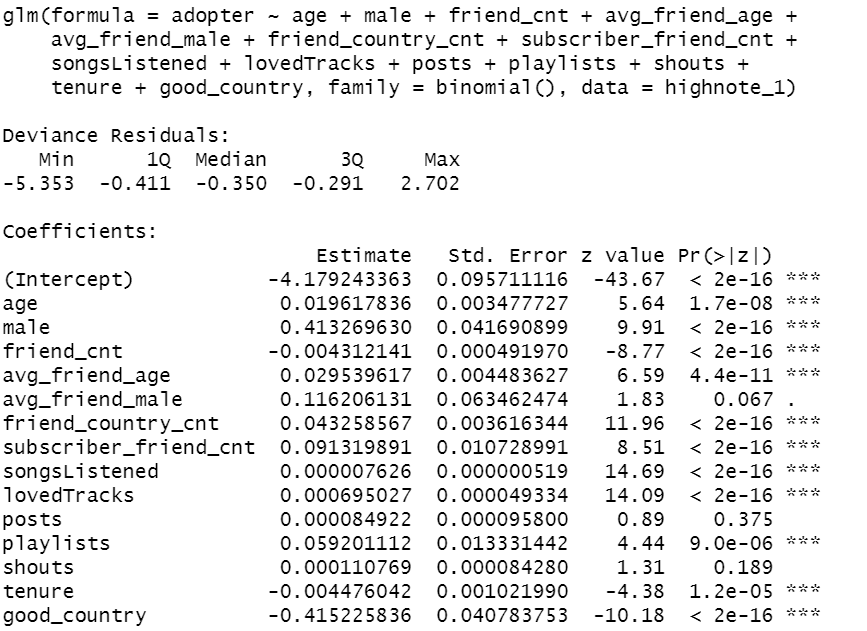


This model has an accuracy of 87%

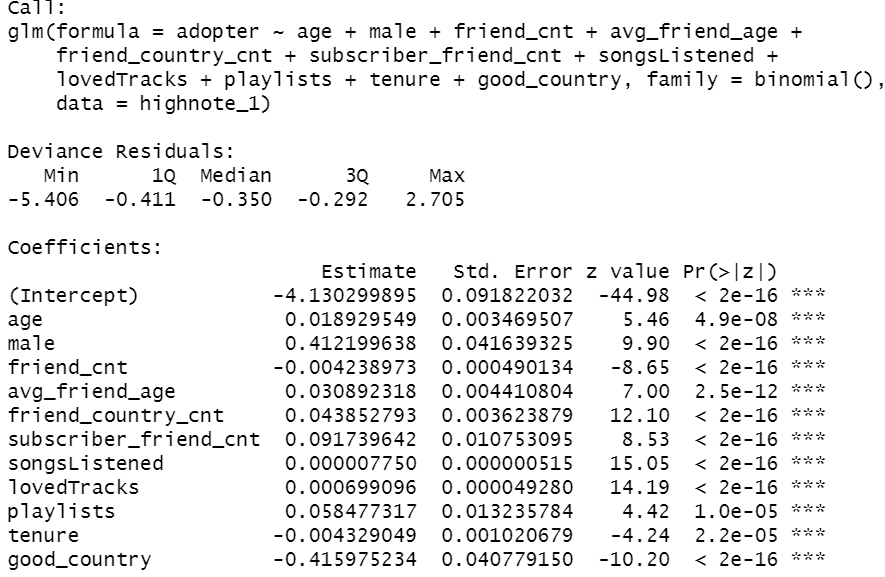
I also ran the model on the original data and here are the results when I include all the variables:

The summary result shows that avg\_friend\_male, posts and shouts are not significant at 95% confidence with p values greater than 0.05. Generally, the following variables have a positive effect on the likelihood of becoming an adopter, including age, male, avg\_friend\_age, friend\_country\_cnt, subscriber\_friend\_cnt, songsListened, lovedTracks, and playlists.The following variables have a negative effect on the likelihood of becoming an adopter, including friend\_cnt, tenure and good\_country.

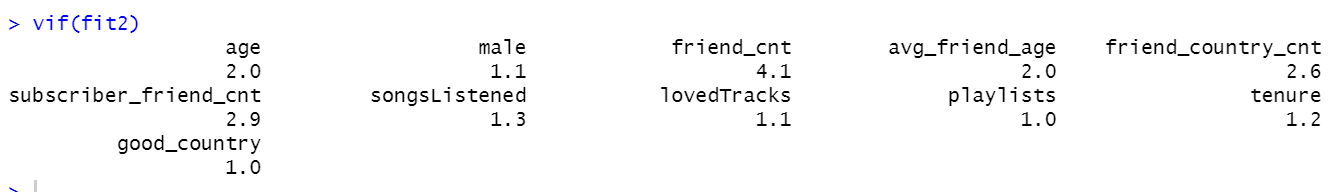
But friend\_cnt, songsListened, lovedTracks, and tenure may not be significant in economical level due to the small coefficient value. Three kinds of variables should be attached more importance to, which are male, subscriber\_friend\_cnt, and good\_country due to the large coefficient value.



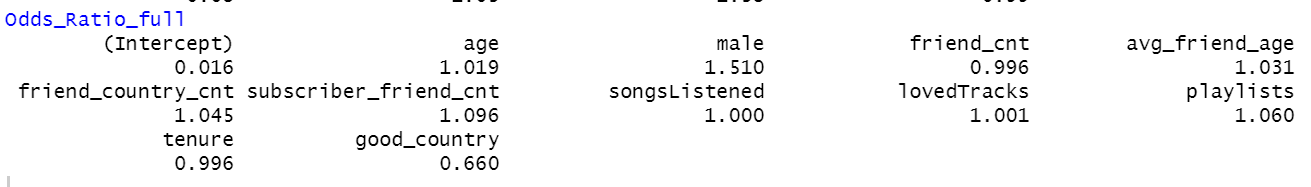
I ran the model after removing the insignificant variables:



I ran VIF to check the importance of these variables and got the values less than 5 which implies that these variables are significant in predicting the likelihood of becoming an adopter



This model has an accuracy of 92%



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

**Target user :** From the descriptive statistics, it is shown that a typical adopter is older male with many friends who are subscribers. In the meantime, the regression analysis shows that older age and older friends will increase the possibility of becoming an adopter. As a result, for High Note, the "free to fee" strategy should focus more on the older male since they are the target users. The company could display more ads to this group of users and develop more function to satisfy their needs to attract them to pay.

**User engagement:** From the data visualization, adopter group is higher than non-adopter group in each aspects of user engagement, meaning that the more user engaged in the site, the more likely he/she will become adopter. Generally, the company should encourage users to interact with the site more frequently, such as recommending more songs to users.

**Globalization:** The company could consider to extend their business globally since users being from good country are not likely to be an adopter. This is because the regression analysis shows that user who is not from US, UK or Germany is more willing to pay.

The analysis concludes that users’ levels of participation are linked to their willingness to pay for premium service. Users who are more active in the website are substantially more likely to pay for premium services. The company should create more content at the global level and target friends of subscribers who are males. Also, to improve the participation of all the non adopters, the company should create content that engage more of such users. Since the participation increases once the customer has converted to premium customer. The company should also take into consideration the tenure since as the tenure increases the likelihood of becoming an adopter decreases. The company should encourage the community building activities, which will encourage the non adopters to engage with the site more.