|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Adopter** | **N Obs** | **Label** | **Mean** | **Std Dev** | **Missing** | **Median** | **Minimum** | **Maximum** | **Range** |
| 0 | 40300 | age | 23.95 | 6.37 | 0 | 23 | 8 | 79 | 71 |
|  |  | male | 0.62 | 0.49 | 15239 | 1 | 0 | 1 | 1 |
|  |  | friend\_cnt | 18.49 | 57.48 | 0 | 7 | 1 | 4957 | 4956 |
|  |  | avg\_friend\_age | 24.01 | 5.1 | 0 | 23 | 8 | 77 | 69 |
|  |  | avg\_friend\_male | 0.62 | 0.32 | 4398 | 0.66 | 0 | 1 | 1 |
|  |  | friend\_country\_cnt | 3.96 | 5.76 | 262 | 2 | 0 | 129 | 129 |
|  |  | subscriber\_friend\_cnt | 0.42 | 2.42 | 32221 | 0 | 0 | 309 | 309 |
|  |  | songsListened | 17589 | 28416 | 1446 | 7440 | 0 | 100000 | 100000 |
|  |  | lovedTracks | 86.82 | 263.58 | 9607 | 14 | 0 | 12522 | 12522 |
|  |  | posts | 5.29 | 104.31 | 31464 | 0 | 0 | 12309 | 12309 |
|  |  | playlists | 0.55 | 1.07 | 21880 | 0 | 0 | 98 | 98 |
|  |  | shouts | 29.97 | 150.69 | 3311 | 4 | 0 | 7736 | 7736 |
|  |  | adopter | 0 | 0 | 40300 | 0 | 0 | 0 | 0 |
|  |  | tenure | 43.81 | 19.78 | 0 | 44 | 1 | 111 | 110 |
|  |  | good\_country | 0.36 | 0.48 | 25881 | 0 | 0 | 1 | 1 |
| 1 | 3527 | age | 25.98 | 6.84 | 0 | 24 | 8 | 73 | 65 |
|  |  | male | 0.73 | 0.44 | 955 | 1 | 0 | 1 | 1 |
|  |  | friend\_cnt | 39.73 | 117.27 | 0 | 16 | 1 | 5089 | 5088 |
|  |  | avg\_friend\_age | 25.44 | 5.21 | 0 | 24.36 | 12 | 62 | 50 |
|  |  | avg\_friend\_male | 0.64 | 0.25 | 130 | 0.66 | 0 | 1 | 1 |
|  |  | friend\_country\_cnt | 7.19 | 8.86 | 7 | 4 | 0 | 136 | 136 |
|  |  | subscriber\_friend\_cnt | 1.64 | 5.85 | 1783 | 0 | 0 | 287 | 287 |
|  |  | songsListened | 33758 | 43593 | 1 | 20908 | 0 | 817290 | 817290 |
|  |  | lovedTracks | 264.34 | 491.43 | 197 | 108 | 0 | 10220 | 10220 |
|  |  | posts | 21.2 | 221.99 | 2158 | 0 | 0 | 8506 | 8506 |
|  |  | playlists | 0.901 | 2.56 | 1598 | 1 | 0 | 118 | 118 |
|  |  | shouts | 99.44 | 1156.07 | 241 | 9 | 0 | 65872 | 65872 |
|  |  | adopter | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
|  |  | tenure | 45.58 | 20.04 | 1 | 46 | 0 | 111 | 111 |
|  |  | good\_country | 0.29 | 0.45 | 2513 | 0 | 0 | 1 | 1 |

1. **Summary statistics: Generate descriptive statistics for the key variables in the data set, similar to the table on the last page of 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?**

There are differences in mean values between adopter and non-adopter subsamples. The average age is higher in adopter subsample. On average, this subsample also consists of higher proportion of male, higher number of friends, friends’ age, male friends, friends’ country, subscriber friends, songs listened, loved tracks, posts, playlists, shouts and even tenure. The only variable in which adopter is lower in average compared to non-adopter is good country. This means that users in non-adopter subsample tend to come from a country such as US, UK, or Germany.

1. **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**

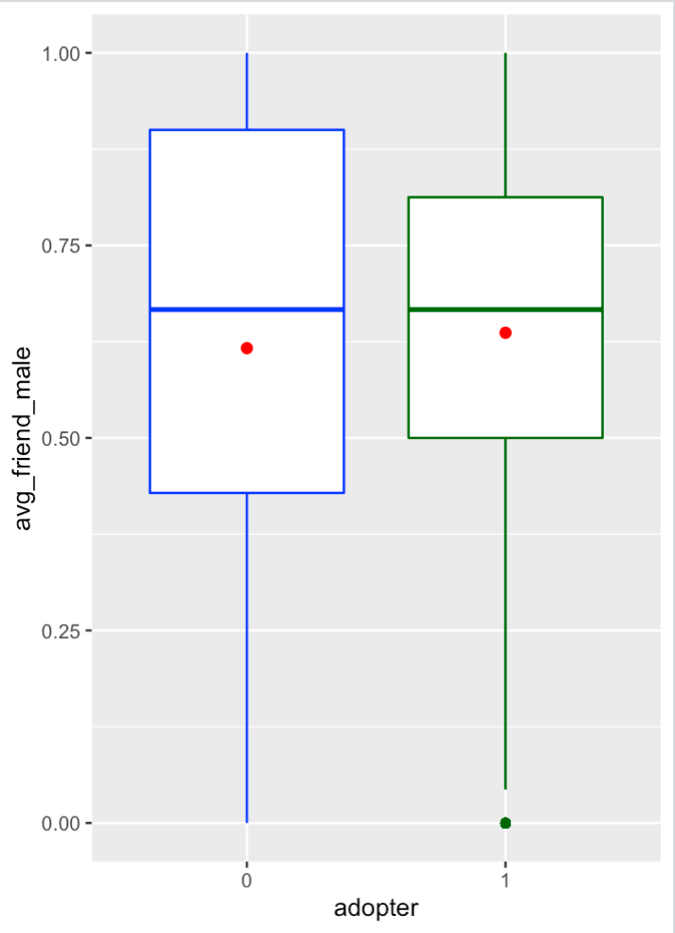
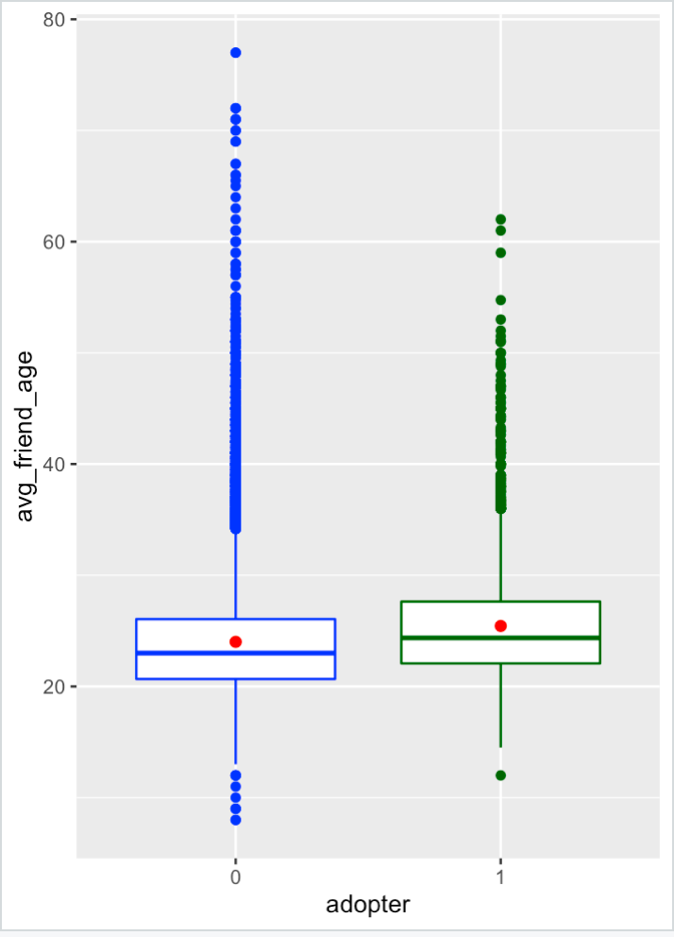
The demographic factors can be analyzed by knowing the average number of age, gender, and good country between adopters (adopter with variable 1) and non-adopters (adopter with variable 0).

Based on the bar graphs below, High Note adopters are usually around the age of 26, high average proportion of males, and come from the rest part of the world besides US, UK, or Germany.

Meanwhile, non-adopters tend to have lesser average proportion of males, around the age of 24 and come from a good country like US, UK, or Germany.



Other factors that can be determined in terms of demographics are High Note users’ friend’s age and gender. It is shown that the average age of users’ friends, indicated by the red dot, is higher in adopter. Similarly, there is a higher average in male proportion for the adopter group.

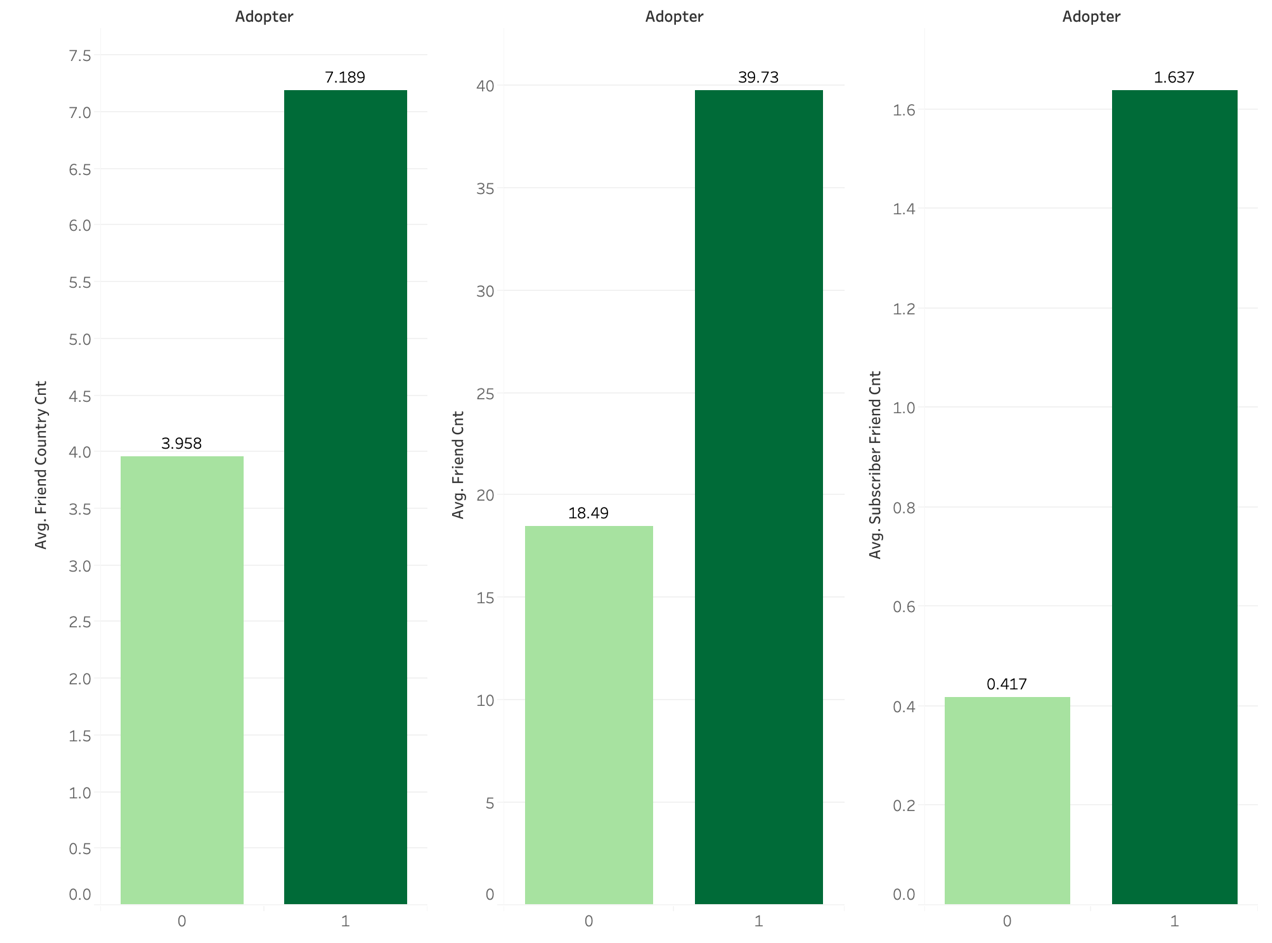


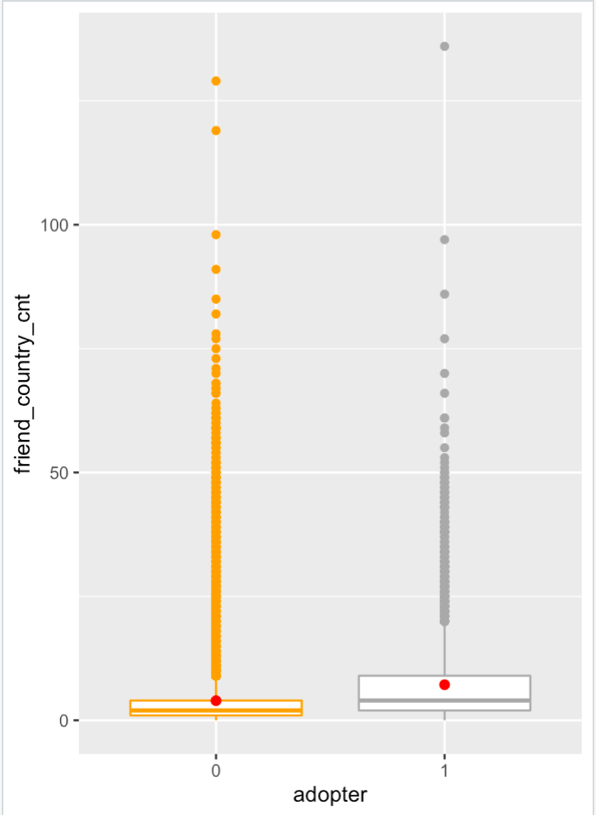
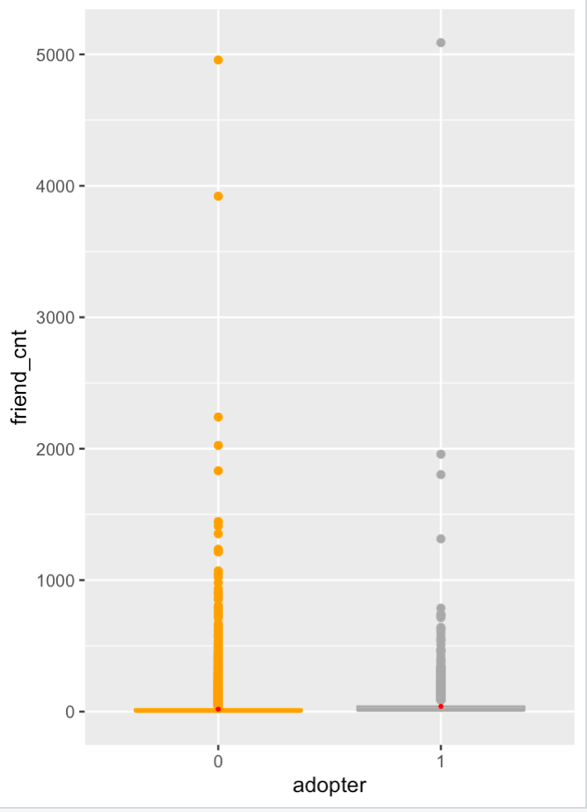
**Peer Influence**

The peer influence-related factors are shown by the mean differences in friend counts, friend country counts, as well as subscriber friend counts.

According to the bar graphs, the average number of different countries from High Note users’ friends is almost twice higher in the adopter group. The adopter group also shows a twice greater average in number of friends than the non-adopter group. Additionally, the adopters tend to be premium subscribers of High Note.

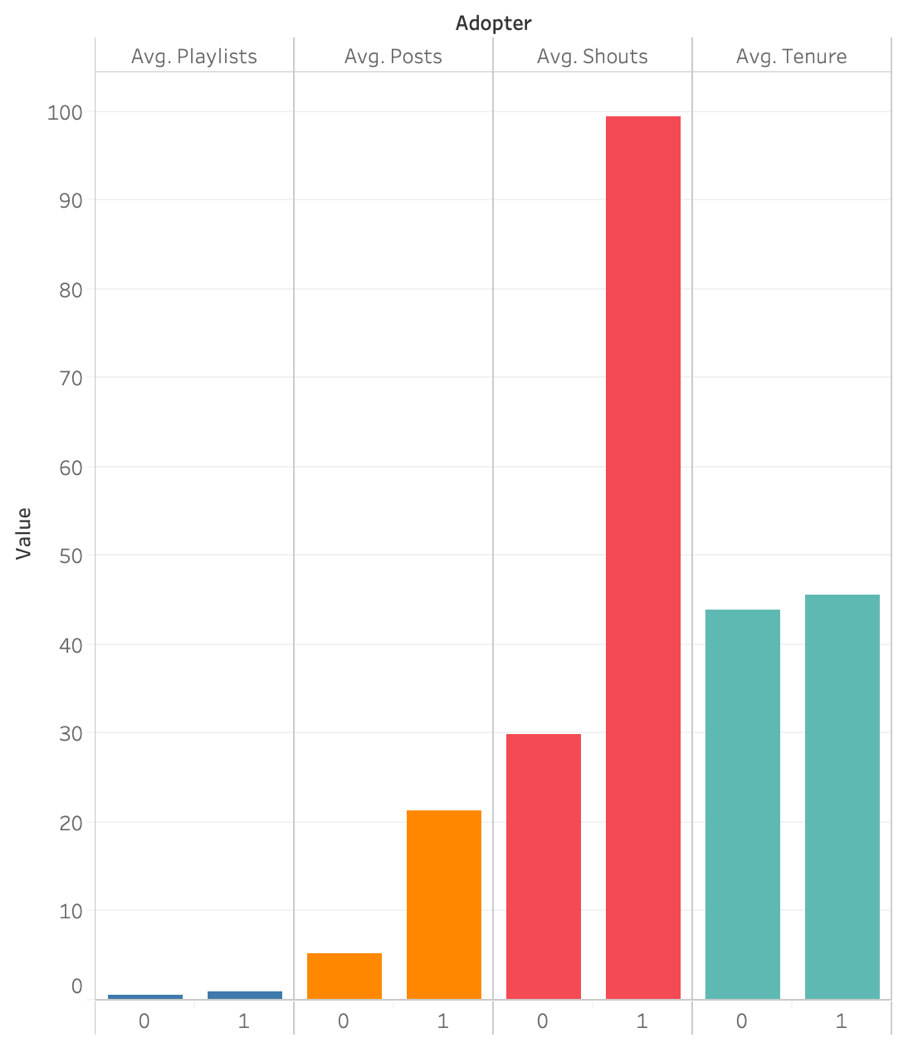
The boxplot for friend counts indicates small interquartile ranges for both adopter and non-adopter group. As mentioned above, the adopter group has a higher mean for number of friends. The boxplot for friend country counts consist of user friends who are adopters and come from various and different countries. This is because the plot in the adopter group has a wider range than the non-adopter’s.

****

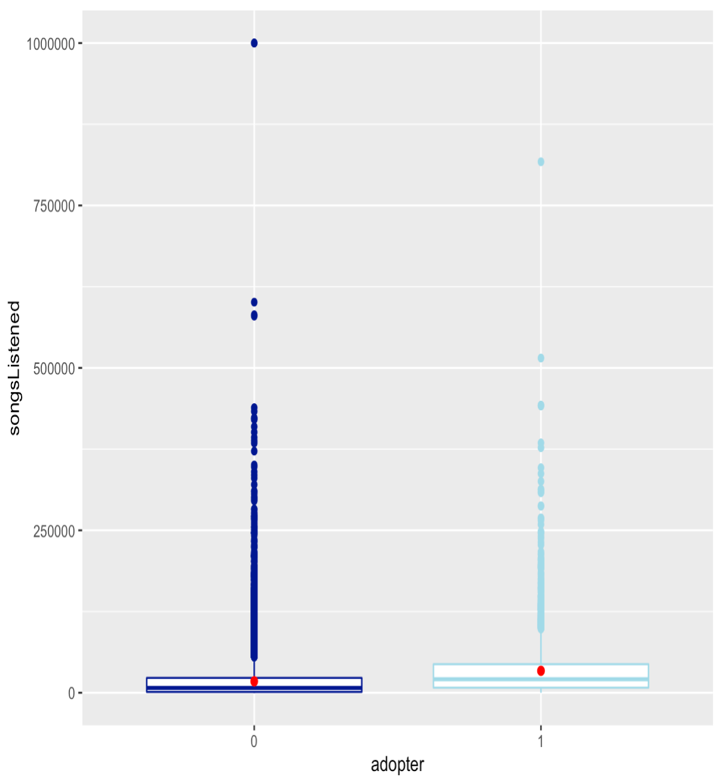
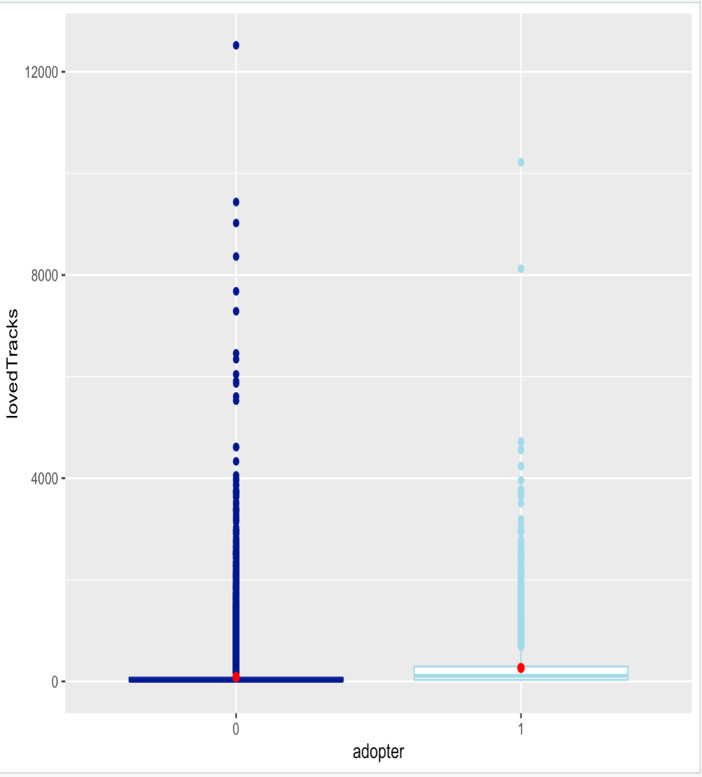
****

**User Engagement**

Average number of playlists, posts, shouts and tenure of the users are used to determine the user engagement. The bar graphs below show that the adopter group has higher average number of engagement in all aspects. This shows that premium subscribers are true music lovers who are active and responsive when it comes to music playing.



The adopter groups also show a higher mean in number of loved tracks and songs listened, as showed by the higher position of red dots compared to the non-adopter group. These may indicate that their premium accounts allow browsing and listening to music better.

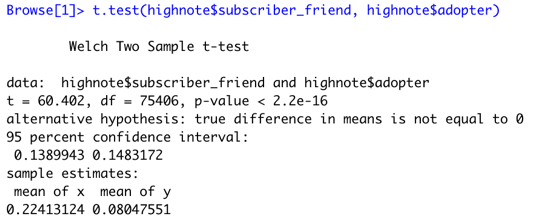


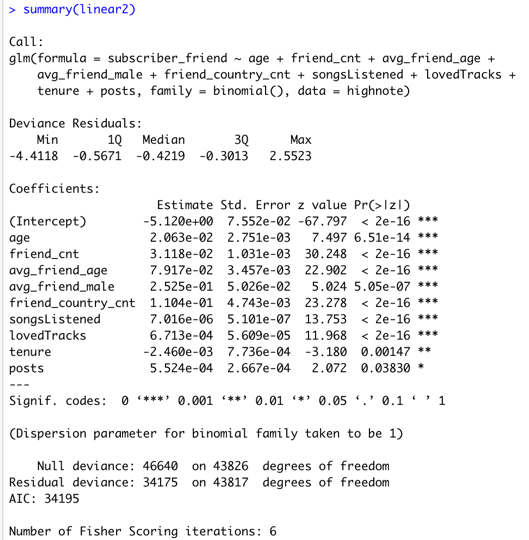
1. **Propensity Score Matching (PSM): You will use PSM to test whether having subscriber friends affects the likelihood of becoming an adopter (i.e., fee customer). For this purpose, the "treatment" group will be users that have one or more subscriber friends (subscriber\_friend\_cnt >= 1), while the "control" group will include users with zero subscriber friends.**

**Code:**



**Result:**



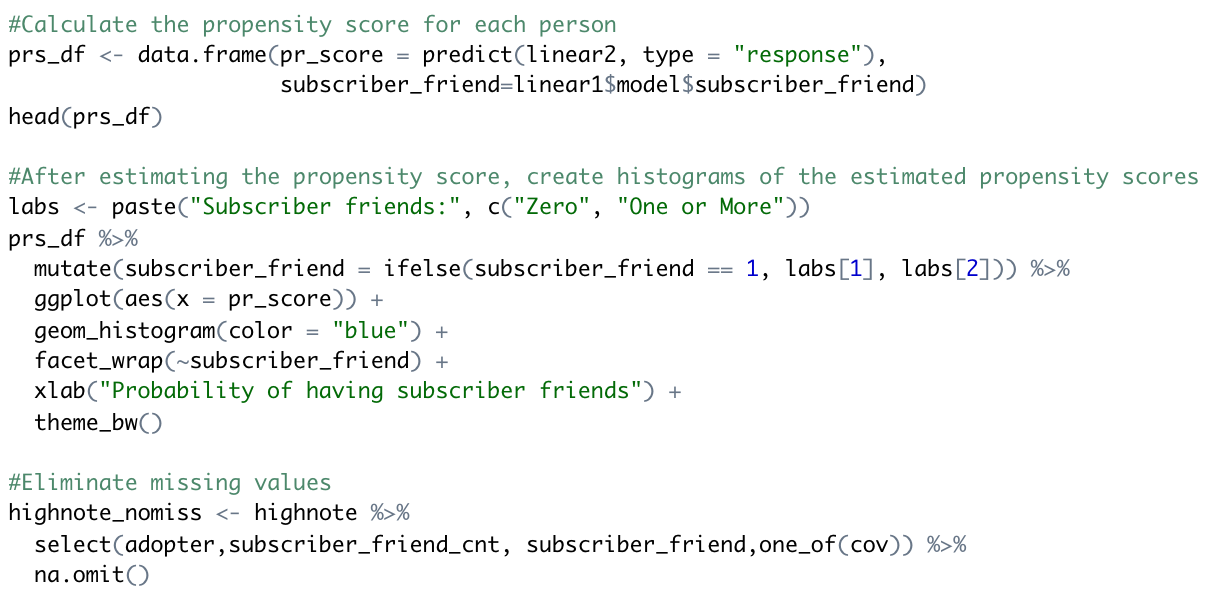


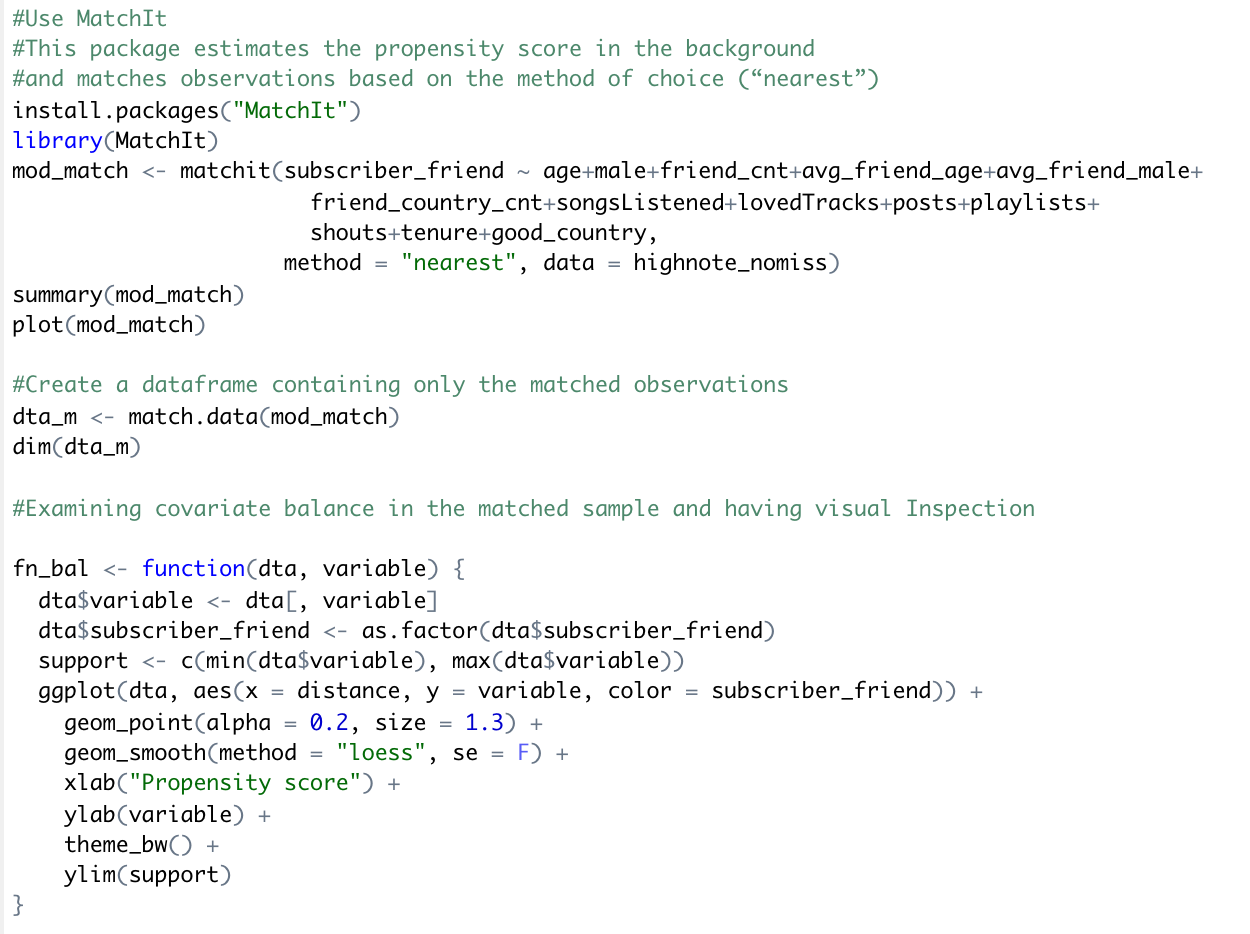
The t-test result shows that the variables “subscriber\_friend” and “adopter” are statistically significant to each other in the 95% confidence interval. This is indicated by the red elliptical mark at the screenshot above.

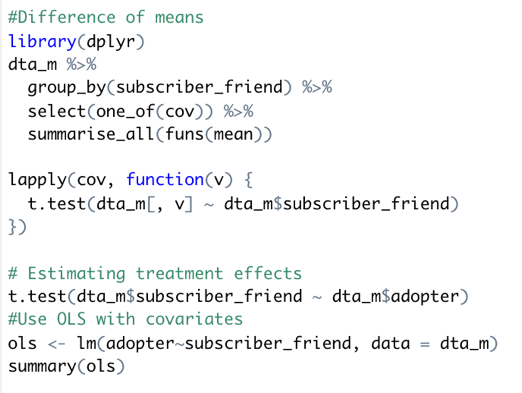
Moreover, the result of generalized linear model (after running the 2nd time) shows all variables that have p-values lower than 0.05. It is shown by the red rectangle mark above. This result is achieved due to removal of variables “male”, “playlists”, “shouts” and “good country” that have p-values greater than 0.05.

**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.**

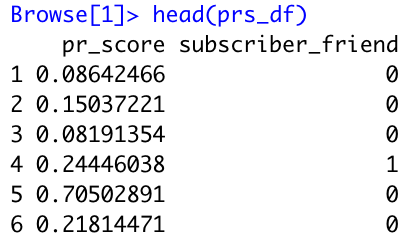
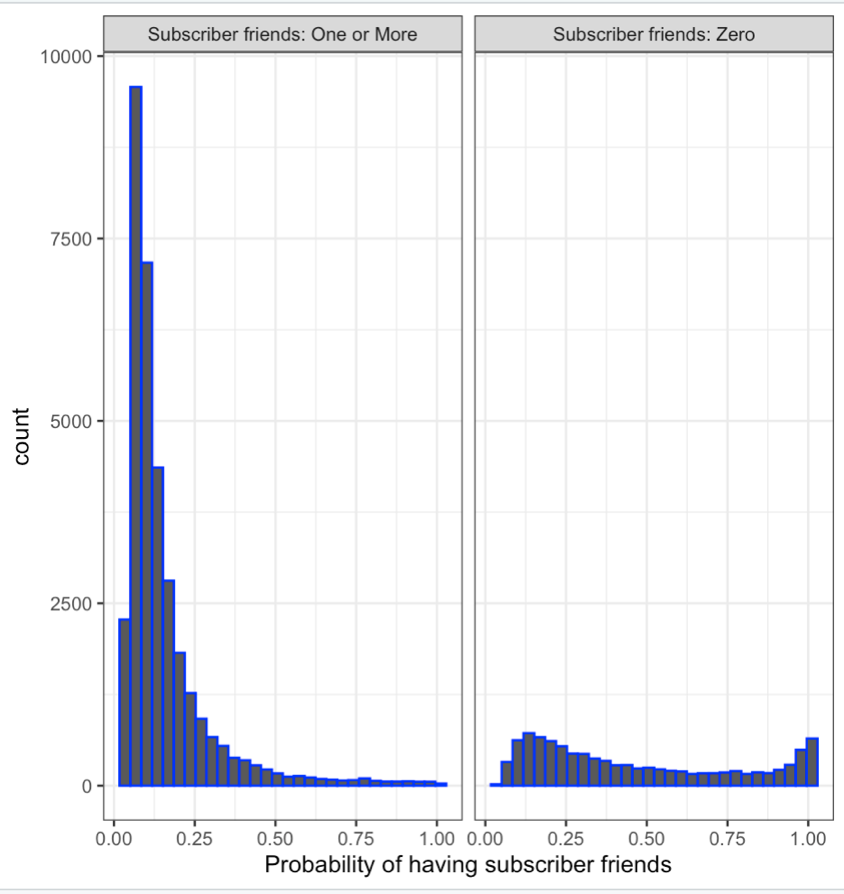
**Code:**



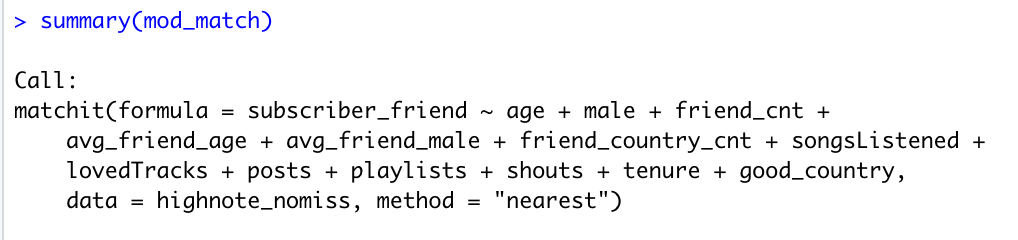




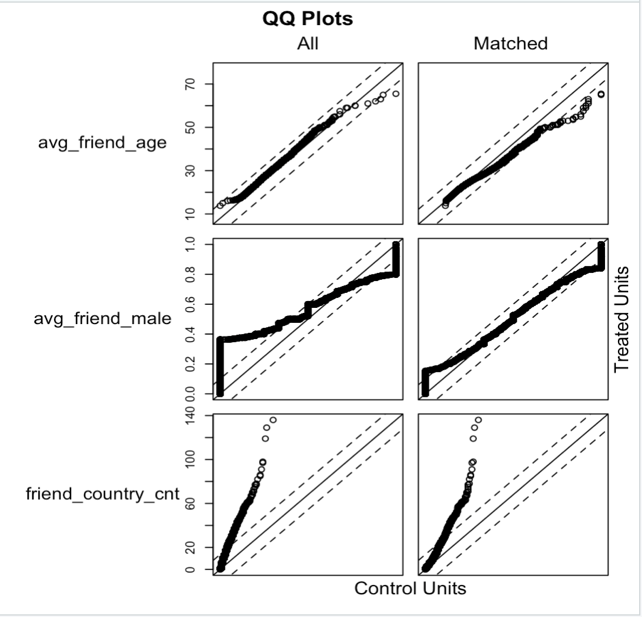
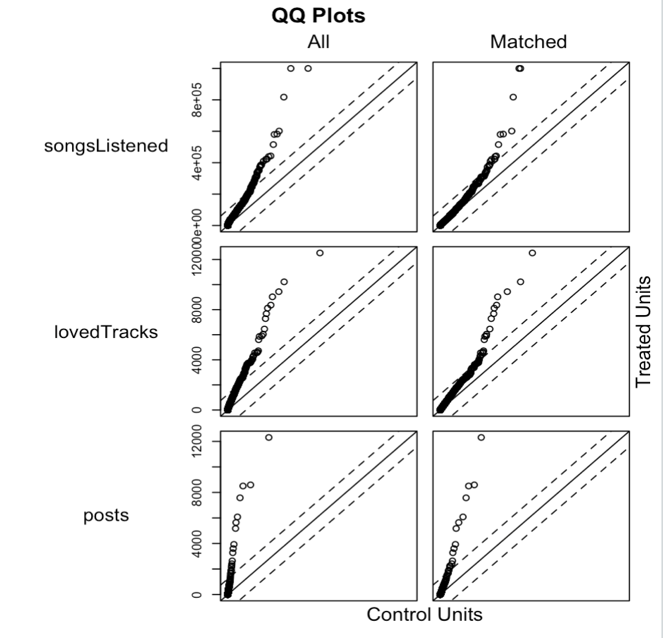
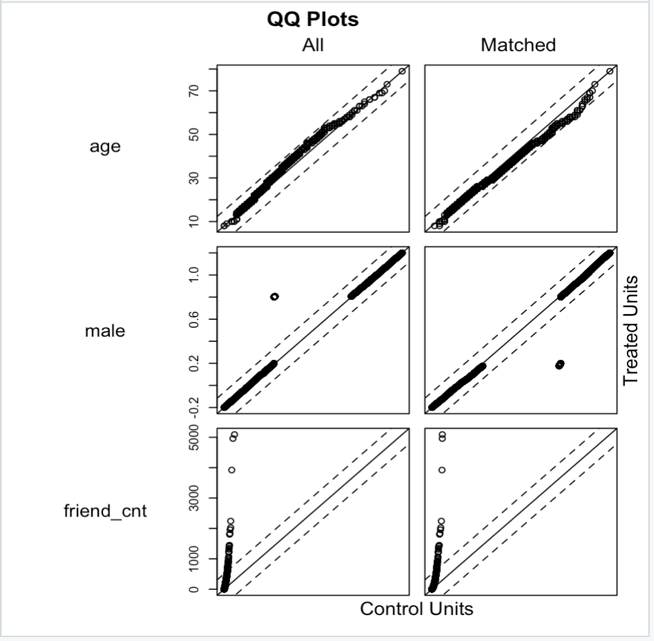
**Result:**

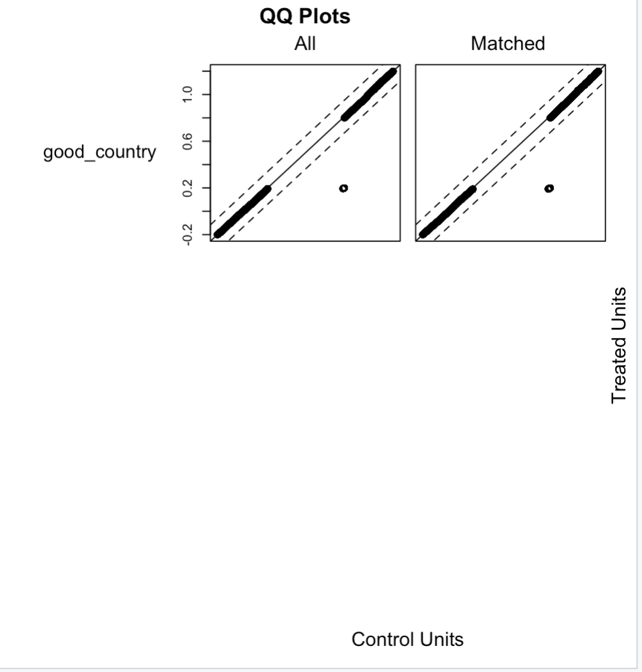
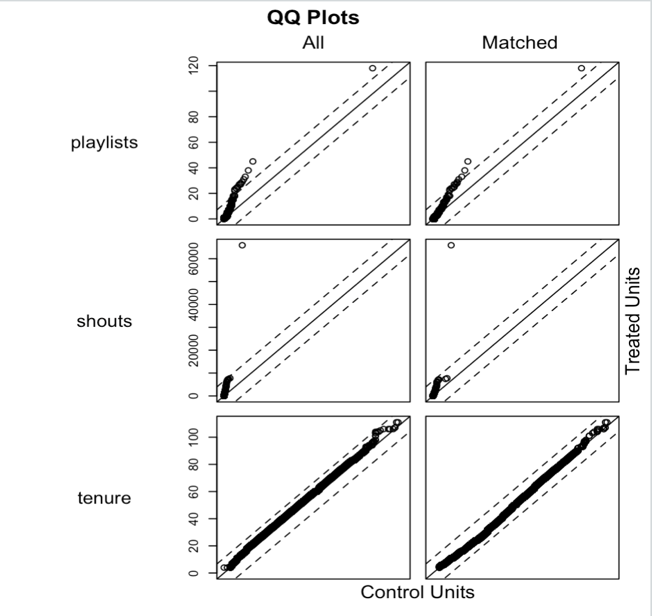
****

The histograms indicate the distributions estimated propensity scores when subscriber\_friend equals to 1 (left-side picture) and 0 (right-side picture). These estimation graphs show the predicted probability of treatment and control group derived from the fitted linear regression. These distributions, however, still consist of unmatched observations between the two groups.

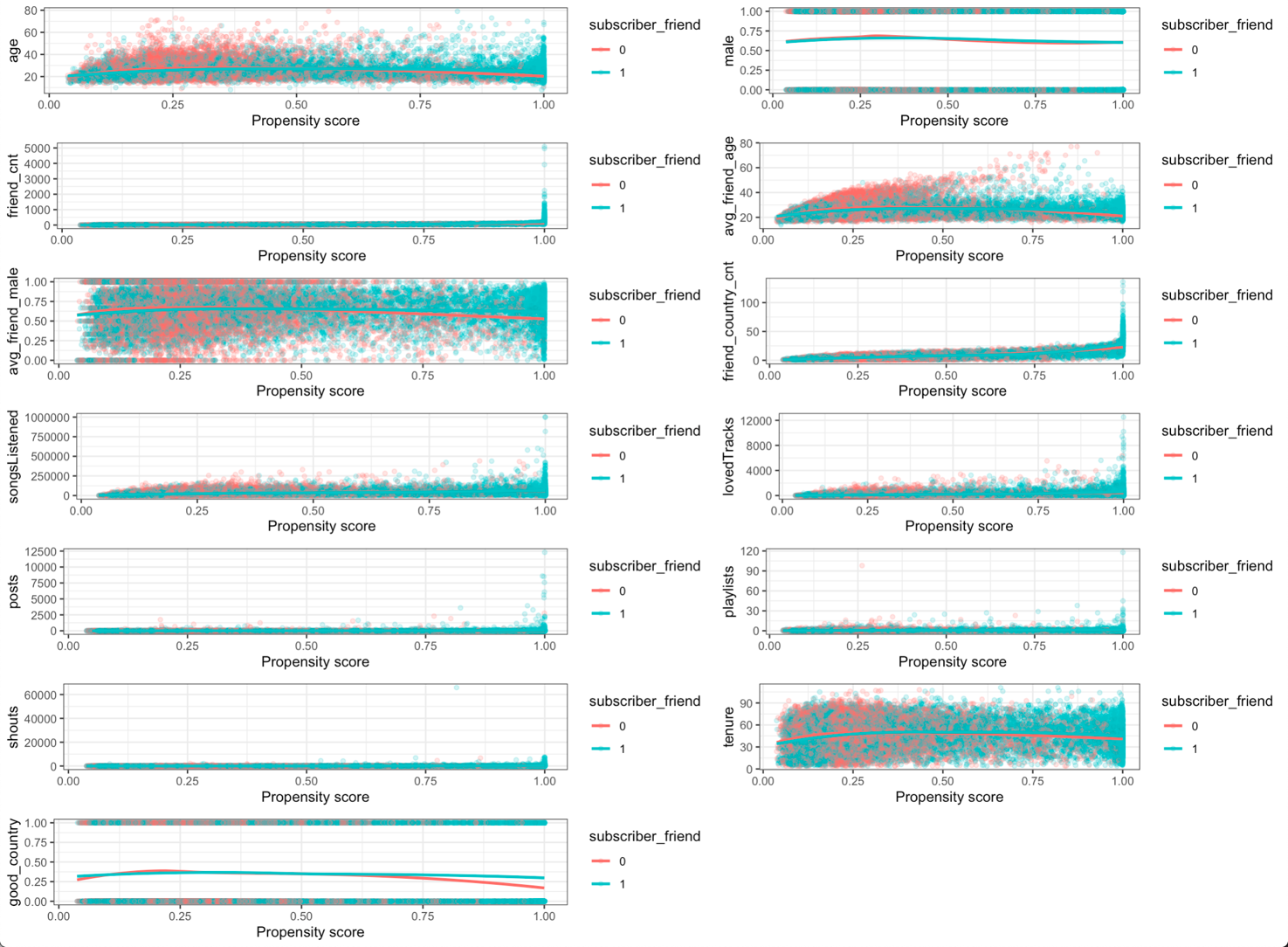
****

The MatchIt package helps to produce QQ plots which plotting two sets of quantiles between one variable against the other. If both sets of quantiles come from the same distribution, the points forming a line would be roughly straight. In this case, the only plots that show straight line and little skew are age and tenure.

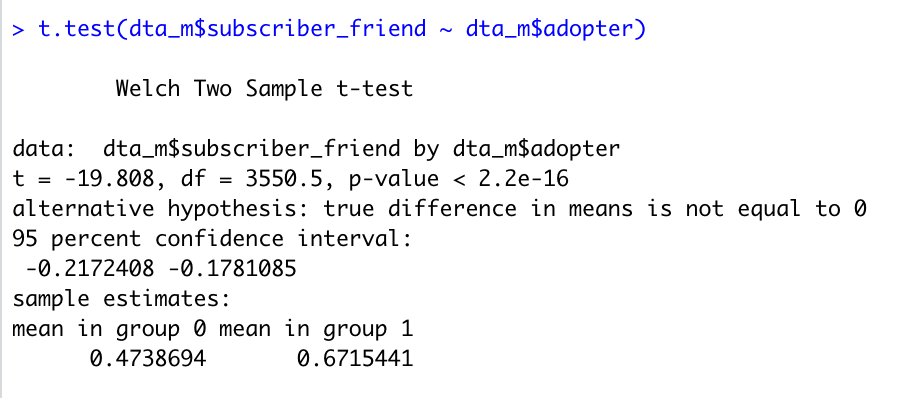
****

****

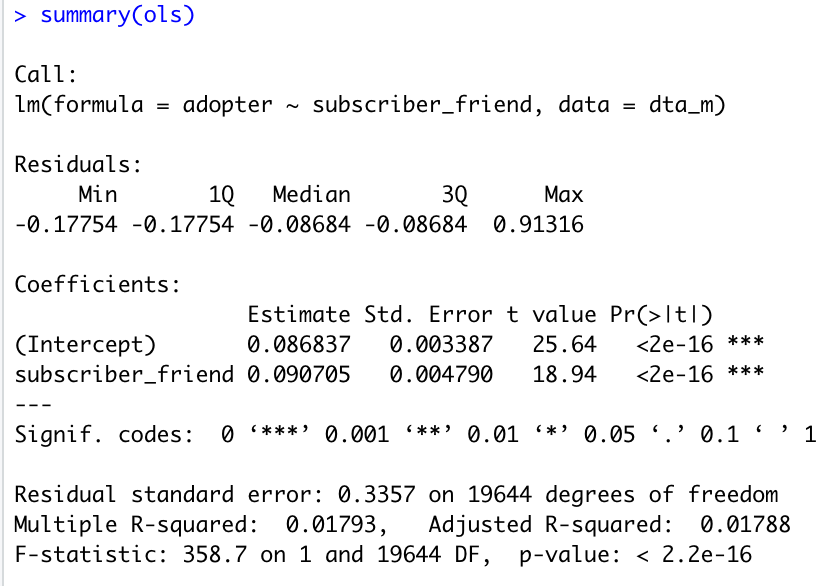
These PSM graphs indicate observations that has been matched. It can be seen that each variable has line graphs of the treatment group (pink color) and control group (blue color), which lie upon each other due to the matched attempts.

****

The t-test helps to identify if there is a significant average treatment affect after doing PSM. PSM is able to control the bias by making the groups having zero and non-zero subscriber friends comparable with respect to the control group. Since the p-value is less than 0.05, shown by the elliptical red mark, the dataset is comparable on all observed covariates.

****

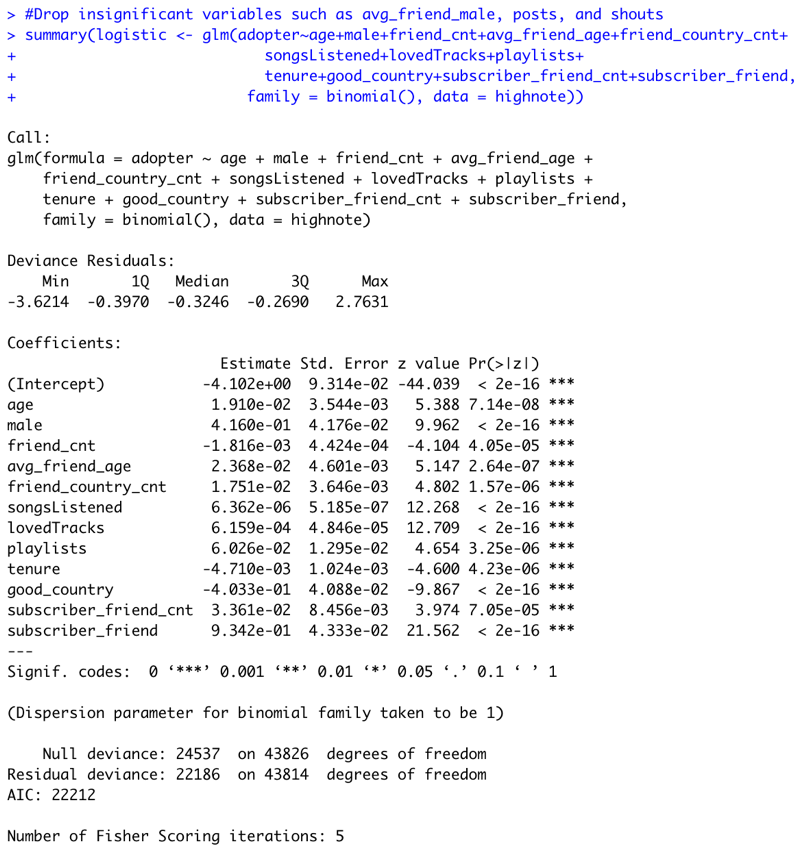
The OLS (Ordinary Least-Squares) regression is used to model a relationship between subscriber\_friend and adopter in the matched observations. It can be seen below that the subscriber\_friend is statistically significant in the 95% confidence interval, shown by the elliptical red mark.

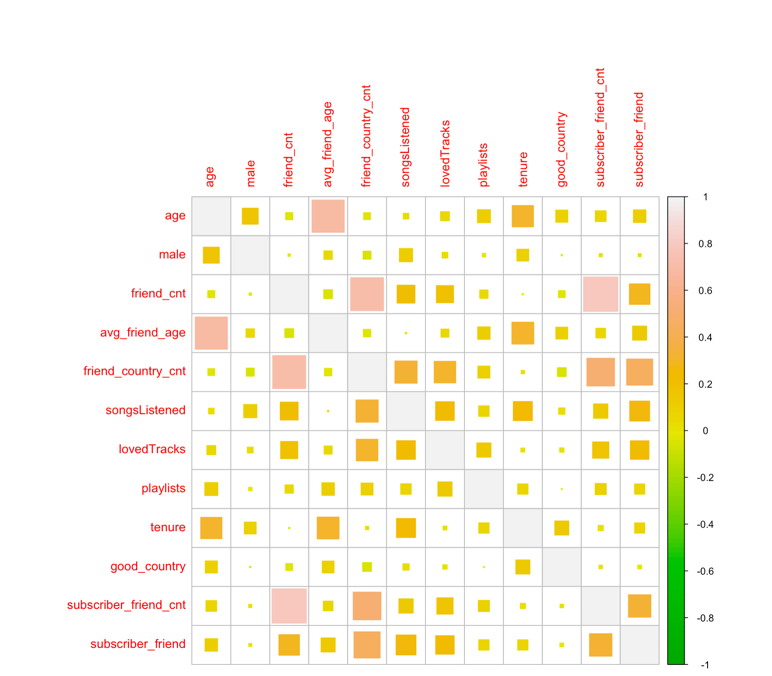
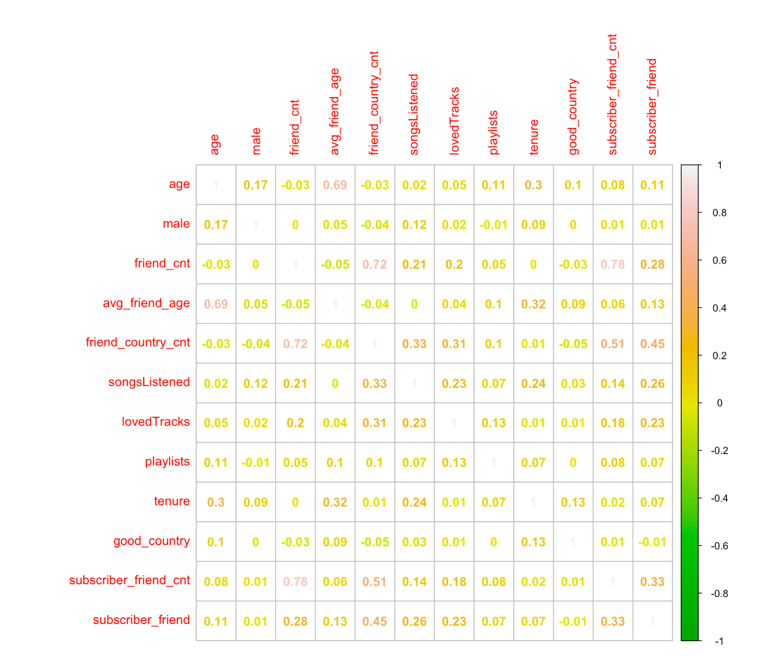
****

1. **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.**



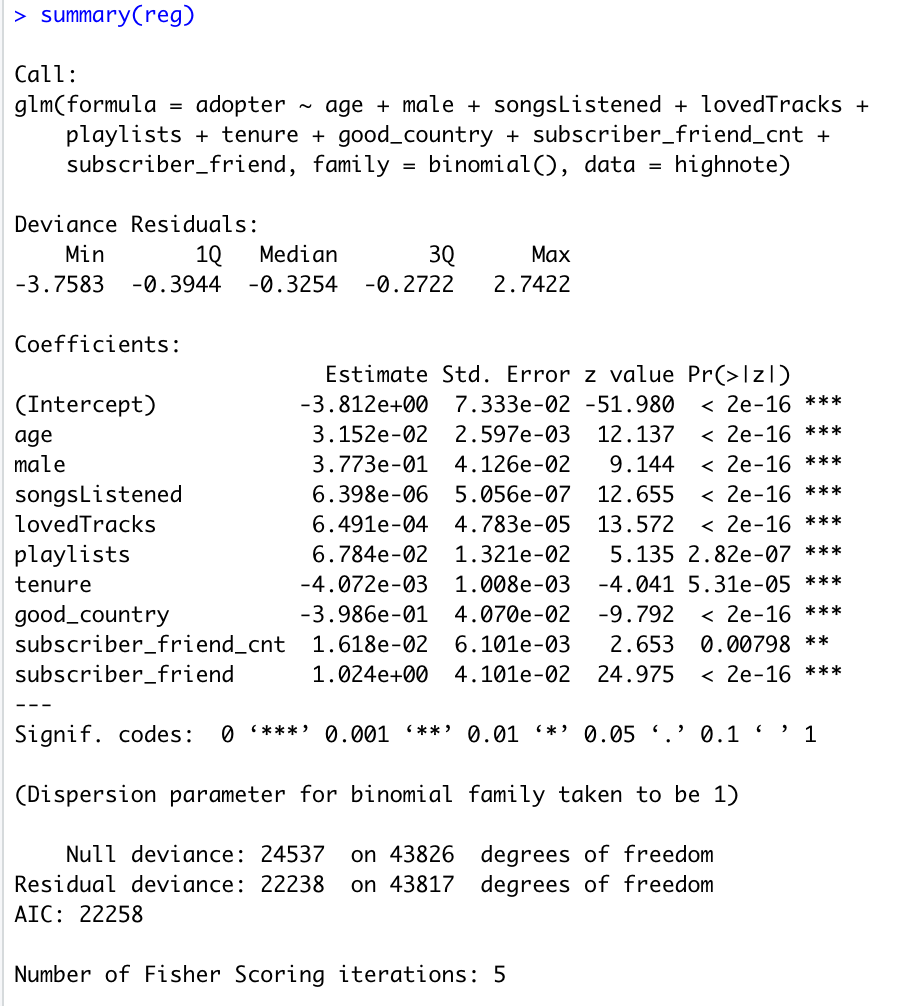
**Use your judgment and visualization results to decide which variables to include the regression.**

After running the first regression model, variable avg\_friend\_male, posts, and shouts show p-values greater than 0.05. This means that they are statistically insignificant in the 95% confidence interval. The removal of those variables creates the second regression model, indicated by the picture on the left. It is showing that all variables are significant for explaining the likelihood of becoming an adopter, since their p-values are below 0.05.

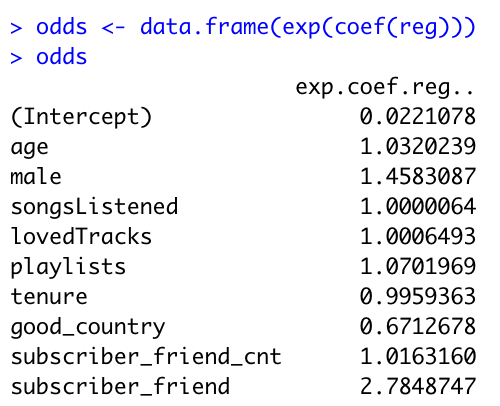


These plots clearly explain that subscriber\_friend\_cnt is highly correlated to both friend\_country\_cnt and friend\_cnt. It is because their correlation values are above 0.50, meaning that they have a positive and strong correlation. Furthermore, age is also highly correlated with avg\_friend\_age, as their correlation value is 0.69.

**Estimate the odds ratios for the key variables. What can you conclude from your results?**



A regression model can be made by simplifying variables that are highly correlated to each other. Variables like avg\_friend\_age, friend\_cnt, and friend\_country\_cnt are dropped, and the rest of variables turn out to be statistically significant to adopter at 95% confidence interval. It is shown by the red rectangular mark at the picture above.



The odds ratio compares the odds of exposure to becoming adopter among treatment groups to odds of exposure to becoming adopter among control group. It is highlighted using the red rectangular mark at the picture above.

If the ratio equals to 1, assuming variable like “tenure”, it means that the odds of exposure between two groups is the same. If the ratio is greater than 1, the odds of exposure among treatment group is then greater than the odds of exposure among control group. For instance, variables like age, male, songsListened, lovedTracks, playlists, subscriber\_friend\_cnt, and subscriber\_friend play a bigger role in explaining the likelihood of becoming an adopter.

When the odd ratio is less than 1, as shown in variable like good\_country, the odds of exposure among treatment group is lower than the odds of exposure among controls. Hence, this variable may hinder the likelihood for users to become adopters.

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

Based on the regression model, age, male, songsListened, lovedTracks, playlists, subscriber\_friend\_cnt, subscriber\_friend, tenure and good\_country are statistically significant to the adopter group.

For instance, the gender “male” makes up a huge proportion for being an adopter. This information can be used to inform a “free-to-fee” strategy for High Note. The company should target male users even more by understanding their tastes about particular song genres that they keep listening to. By giving out different recommended playlists (as playlists are one of the contributing factors as well) which contain ad-free songs for the first few months, those users will experience the enjoyable side of listening to music. Hence, males can be more interested in becoming adopters.

Moreover, variable like “subscriber\_friend” indicates that users tend to be adopters when they are friends with adopters. This is a huge opportunity for High Note to strategize in the marketing point of view. For example, the premium subscribers will get a discount in a two-month payment from High Note if they refer their non-adopter users to become a premium member. The subscribers just need to send a referral code to the non-adopters. Although terms and conditions should be applied during the period, having the referral from subscriber friends may influence users to join the adopter group.

In addition, not all variables that are significant also have high odds of turning users into adopters. A variable like “good\_country” play a smaller role to explain likelihood of becoming an adopter, as it has a low odds ratio. This means that users from good countries like US, UK, or Germany are less likely to become adopters. These developed countries might have more competitors in terms of music and entertainment, so it is better for High Note to expand their services into less developed countries. By navigating other countries besides US, UK, or Germany, High Note would be able to gain attention of users to be premium subscribers and thrive in business as online music platform is not readily and widely available to them.