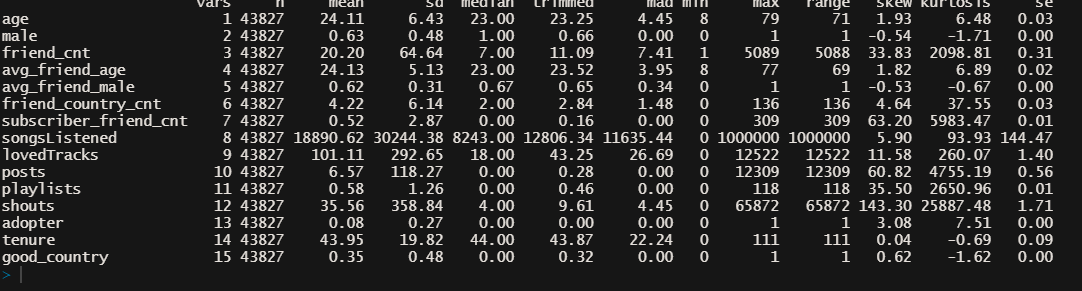
John Chaffey

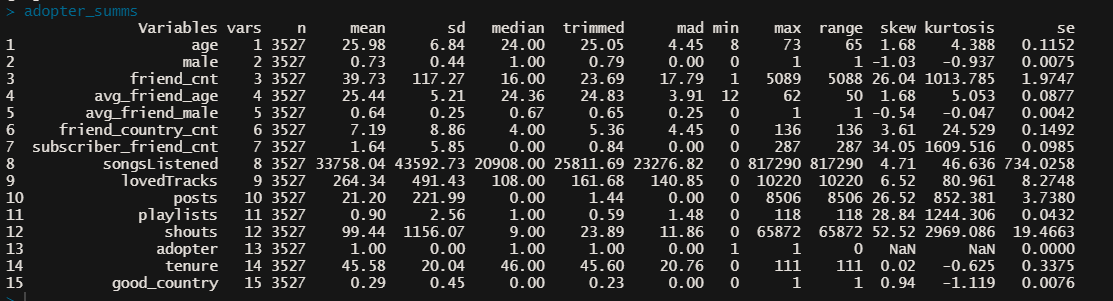
Customer and Social Analytics

Final

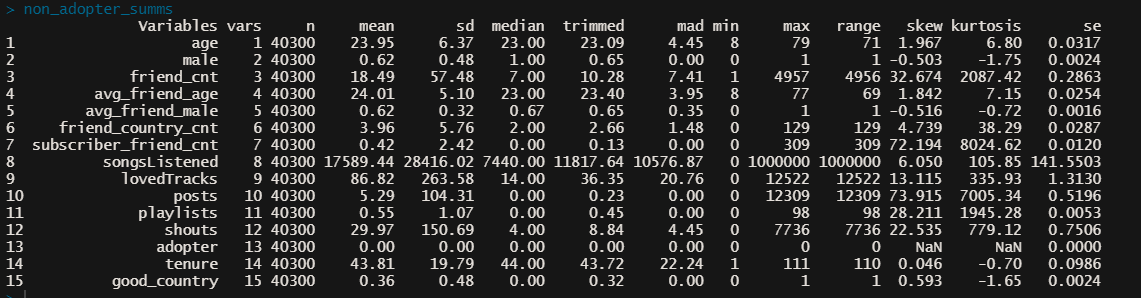
1. **Summary statistics: Generate descriptive statistics for the key variables in the data set, similar to the table on the last page of the case. (Note that your table will look different because the data set you are analyzing is different from the one used to generate the table in the case.) Analyze 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?**



*Here is a summary of the full dataset without subsetting by adopter/non-adopter generated with describe()*

**

*This picture shows the summary statistics of only the adopters in the dataset*



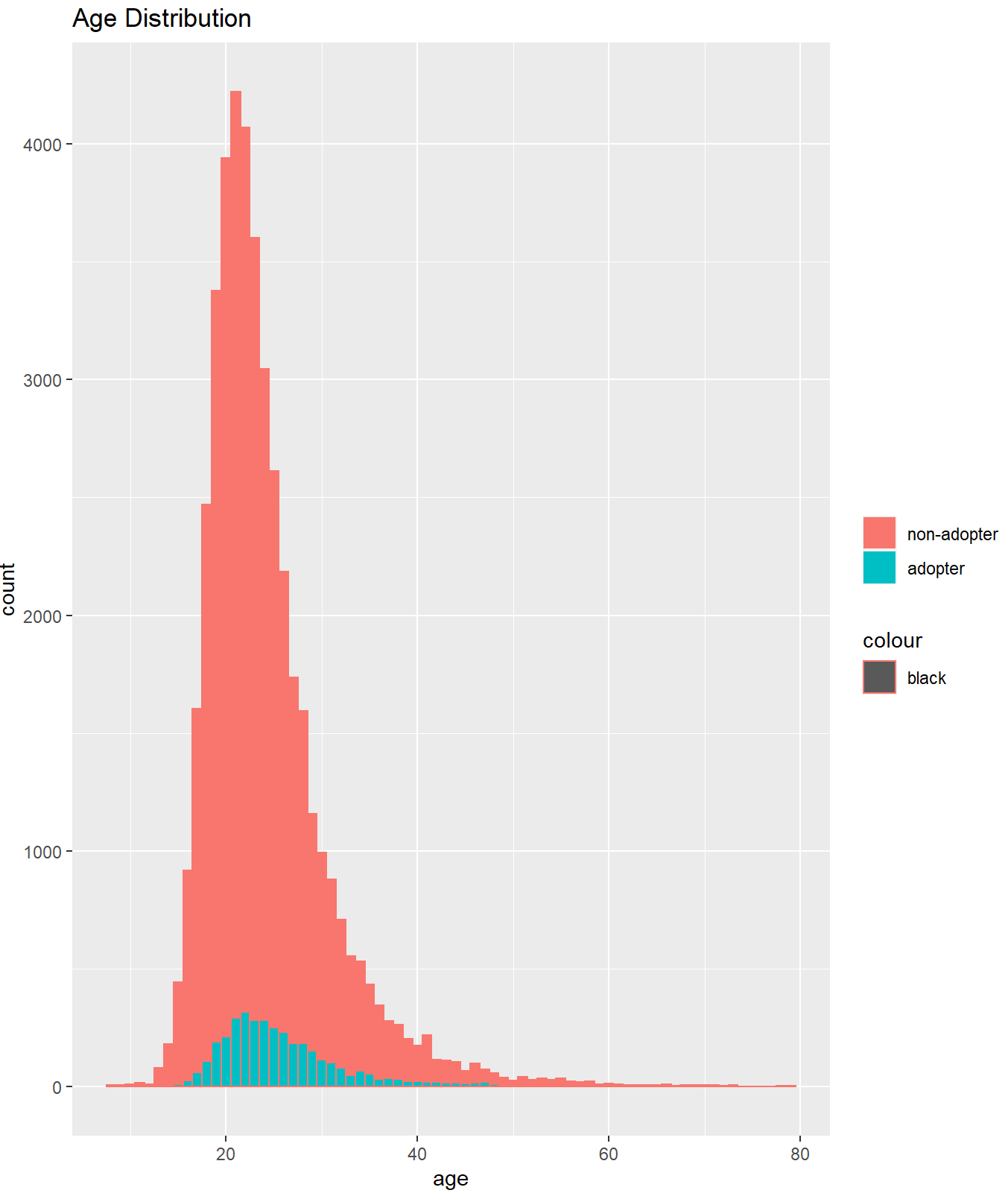
*This picture shows the summary statistics of only the non-adopters in the dataset*

In regards to differences between the adopter and non-adopter subgroups we can see an approximate two year difference in age; with adopters having an approximate mean age of 26 and non-adopters having an approximate mean age of 24. There is also a difference in gender ratios between the adopters/non-adopters with adopters having a higher percentage of males (at ~73%) while males in the non-adopter group make up ~62% of the total population. Adopters also spend more time on the site (43.81 for non adopters vs 45.58 for adopters). Finally we can see that adopters have more playlists, listen to more songs, have more friends, love more tracks, and have higher numbers of shouts when compared to non-adopters.

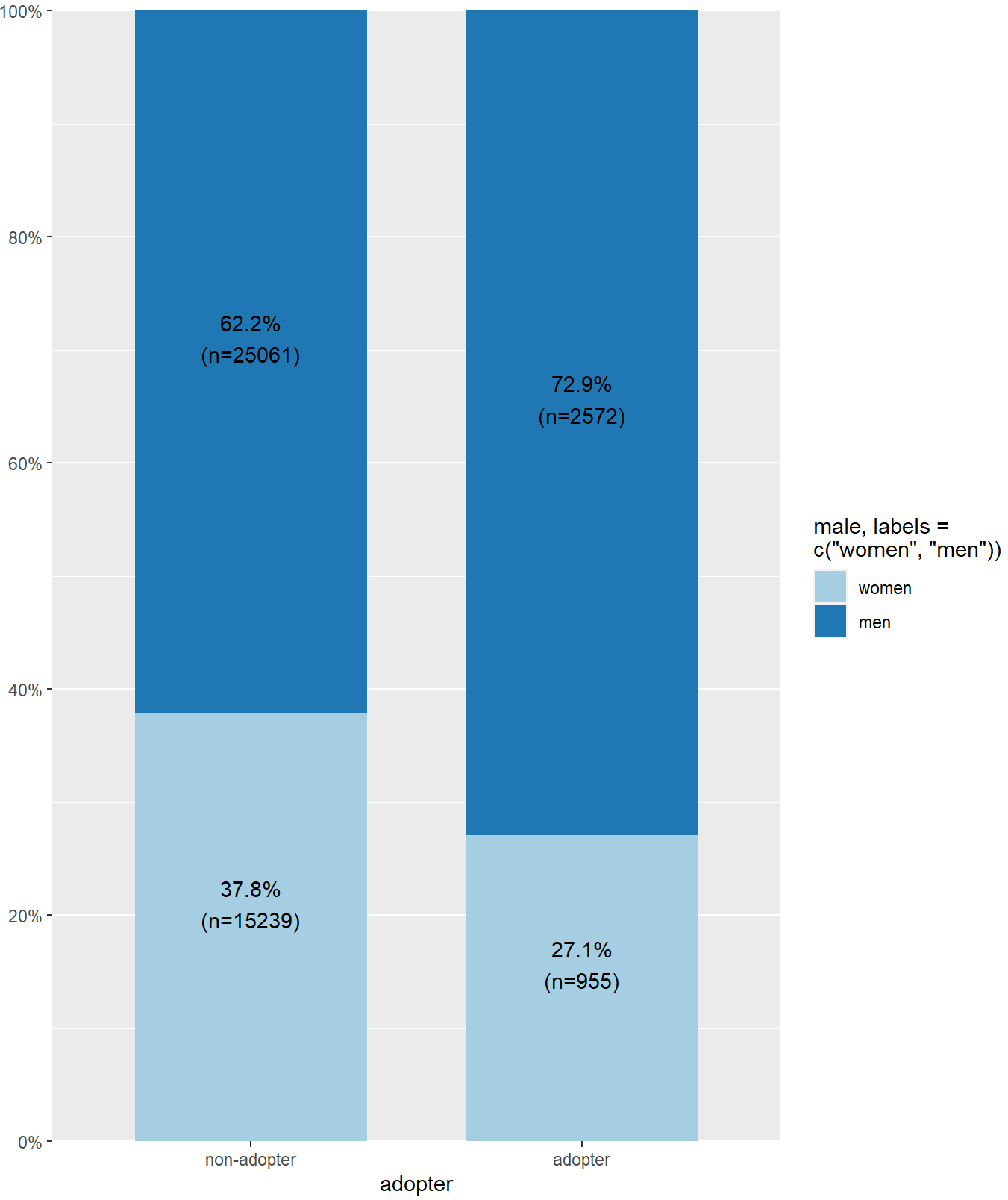
Overall, adopters tend to be far more active on the site than non-adopters.

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?**

i.Demographics

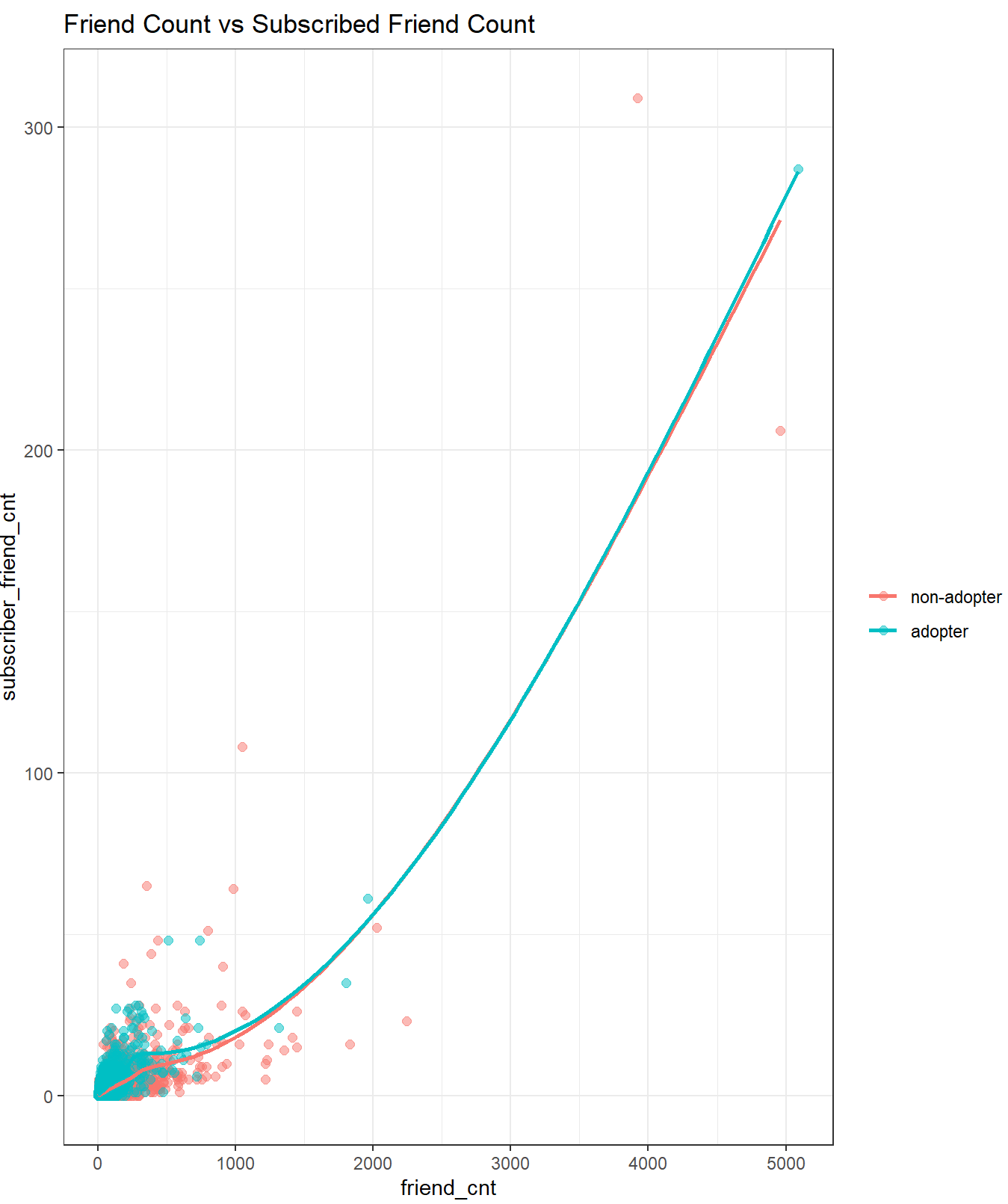
.

*(this graph shows the distribution of adopter/non-adopter varying by age, adopter peaks at early twenties)*

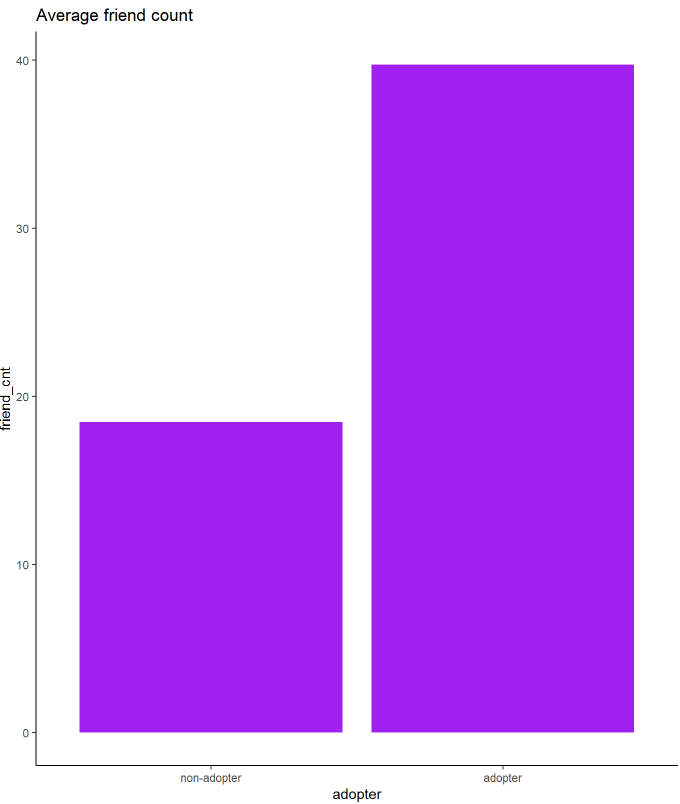


*(gender distribution of adopters vs non-adopters, adopters are heavily skewed towards being male)*

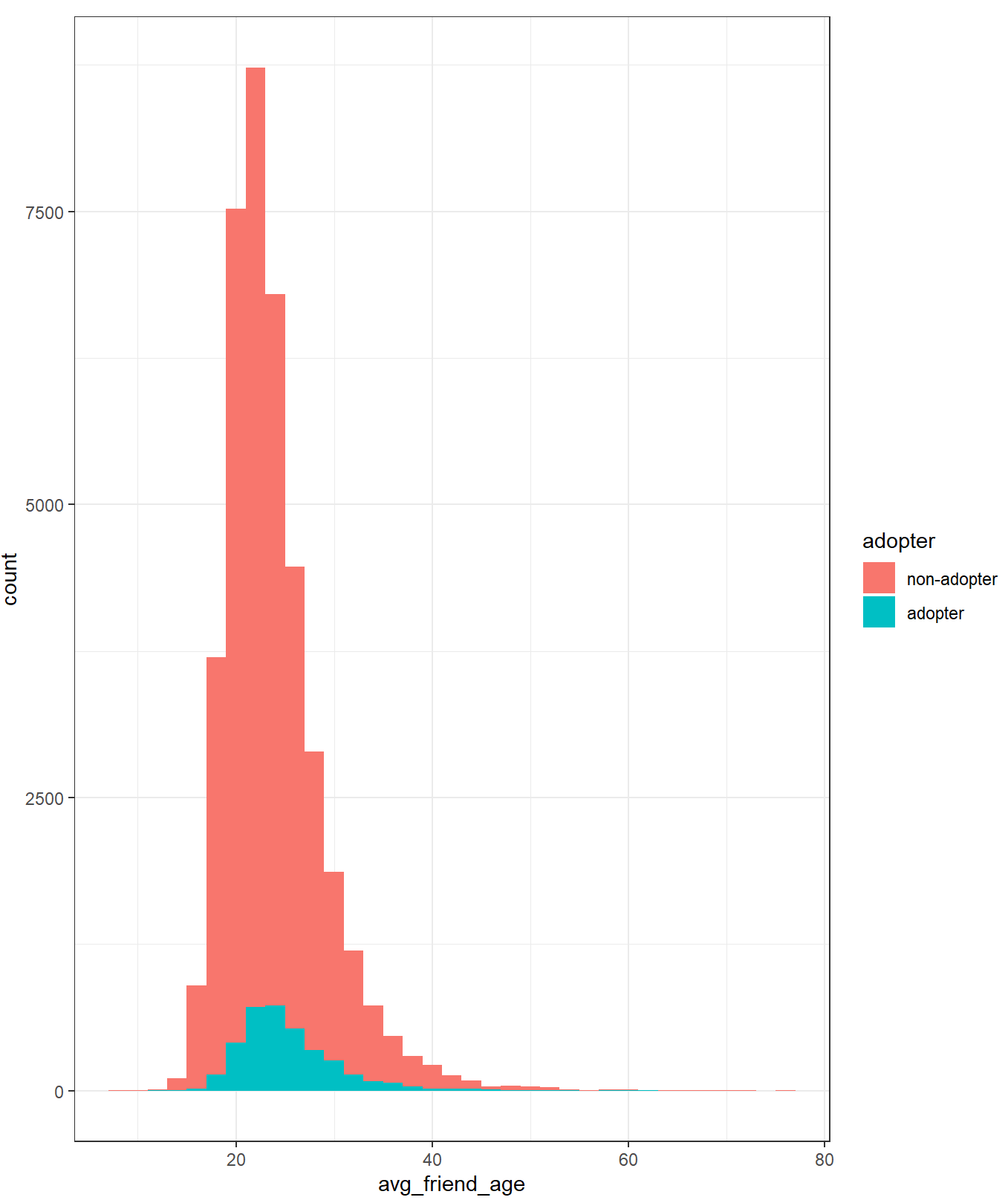
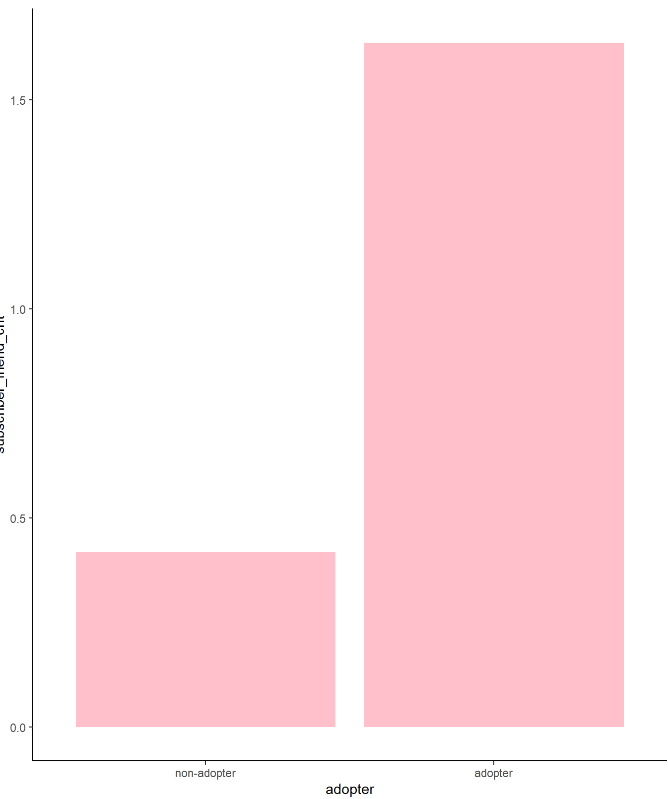
We can see from these charts that there is a heavy skewing regarding both age and gender in regards to the populations for both adopters and non-adopters. From the first chart we can see that the majority of adopters are in their early twenties (the number of adopters by age declines sharply after approximately twenty four). In the second chart we can see that the gender ratio for both adopters and non-adopters is skewed in favor of men with approximately 72.9% of adopters being male and only 27.1% of adopters being female.

ii. Peer Influence

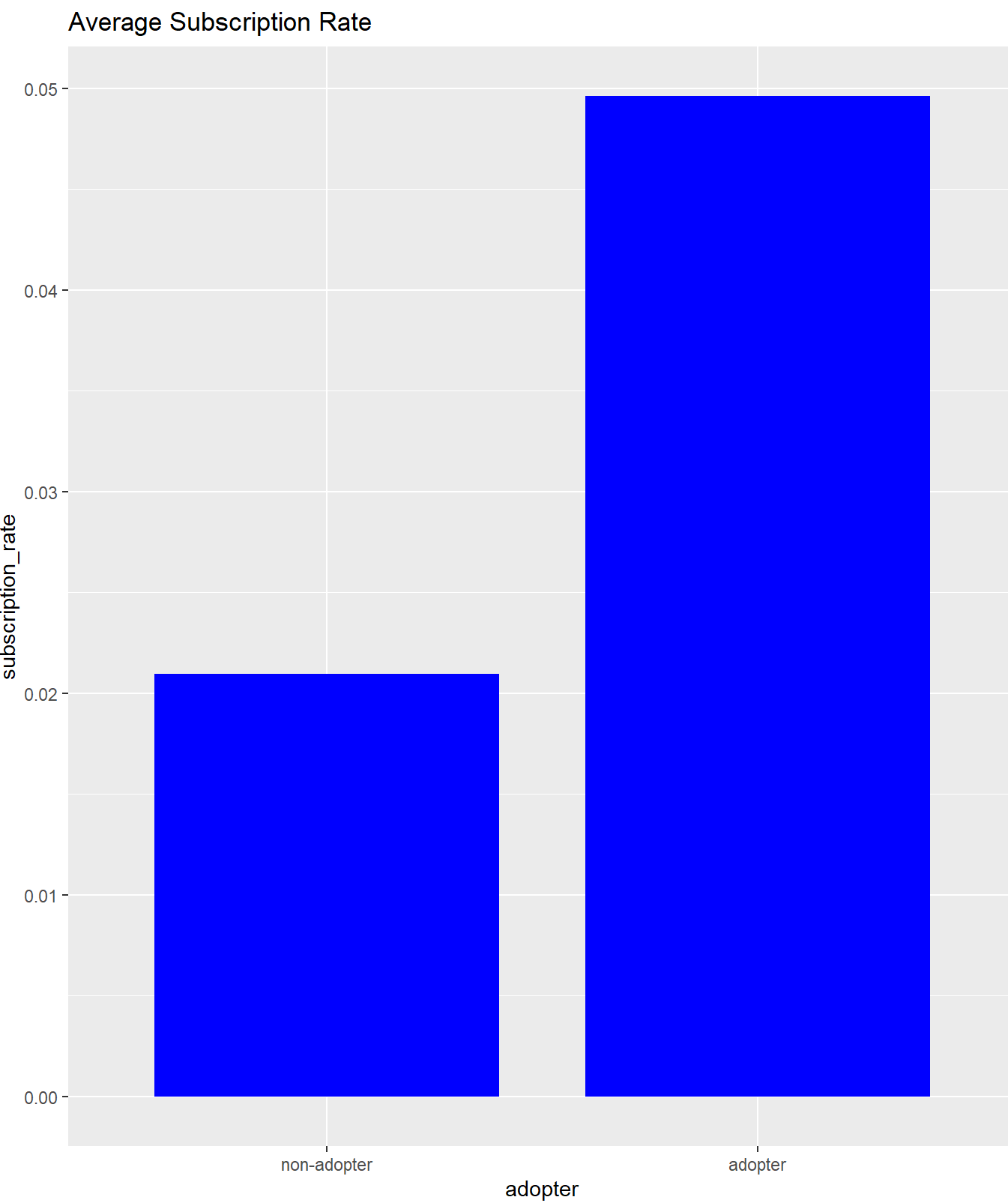
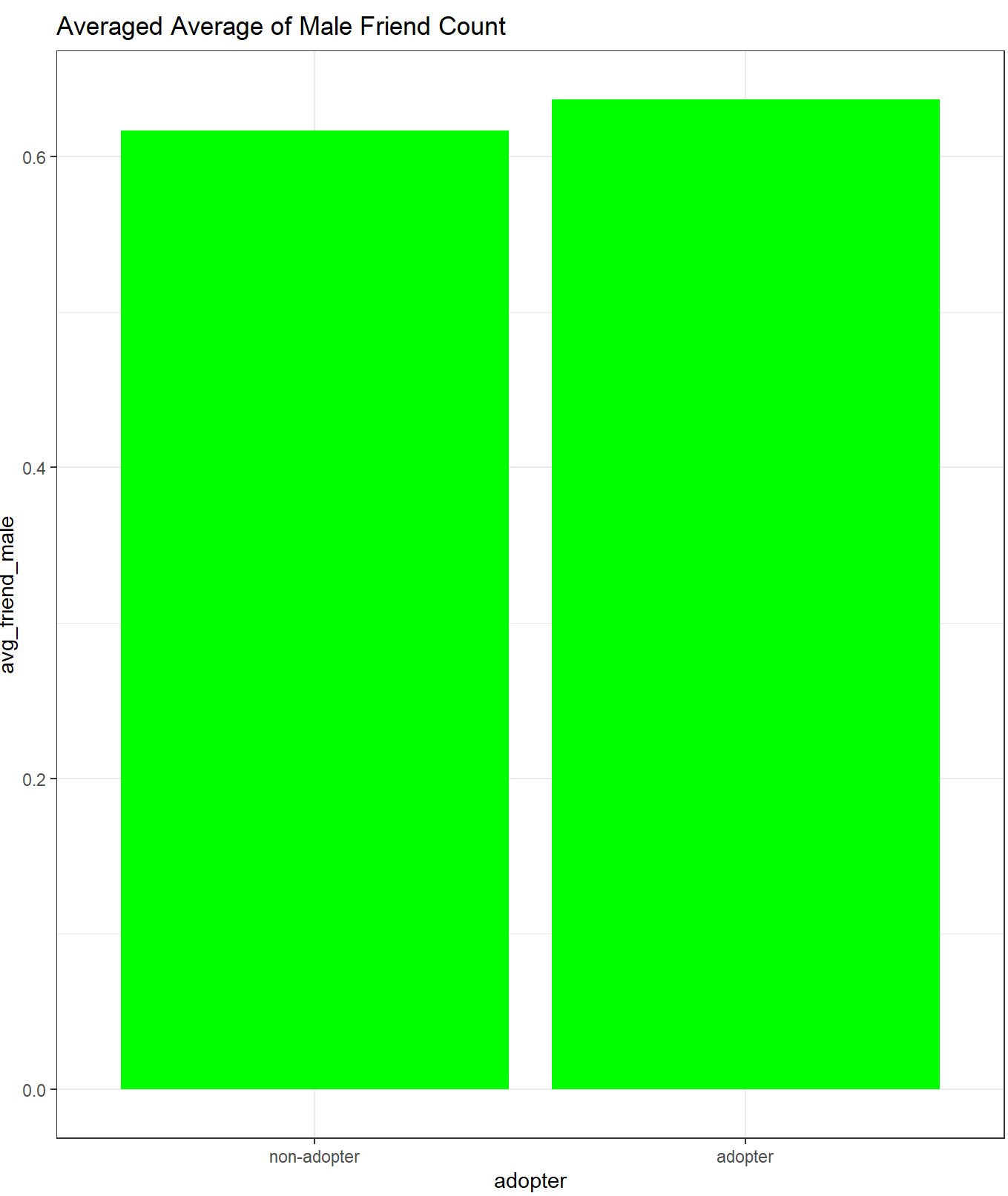
*can be seen in the above chart, subscribed friend counts naturally increase as general friend count increases)*

**

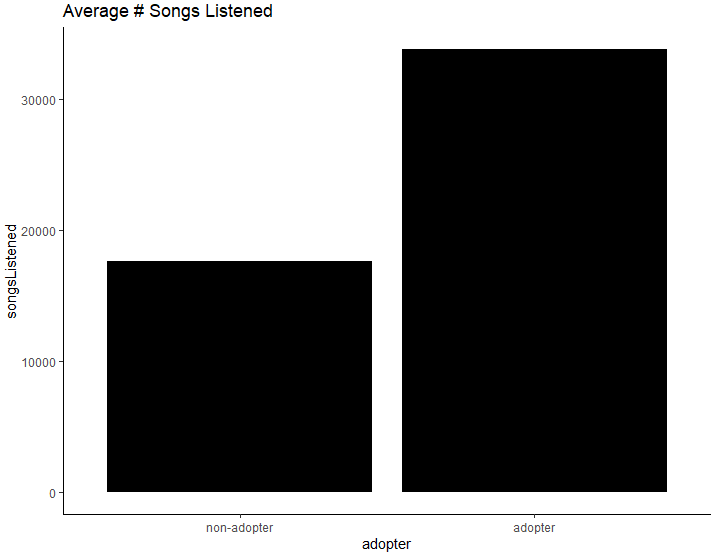
*(this chart showcases the friend counts for adopter vs non-adopter. As can be seen here adopters have on average a significantly higher friend count)*



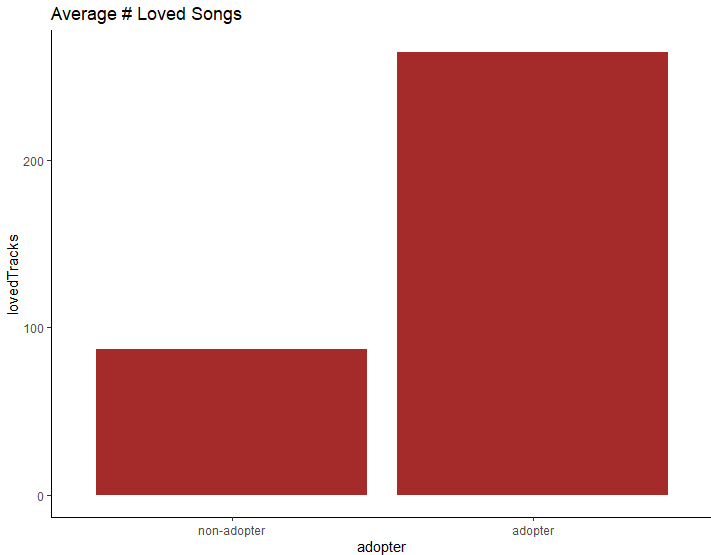
*(this graph shows the relationship between adopters and whether their friends are subscribed or not; as can be seen, adopters typically have a much higher number of friends who are subscribed as well)*

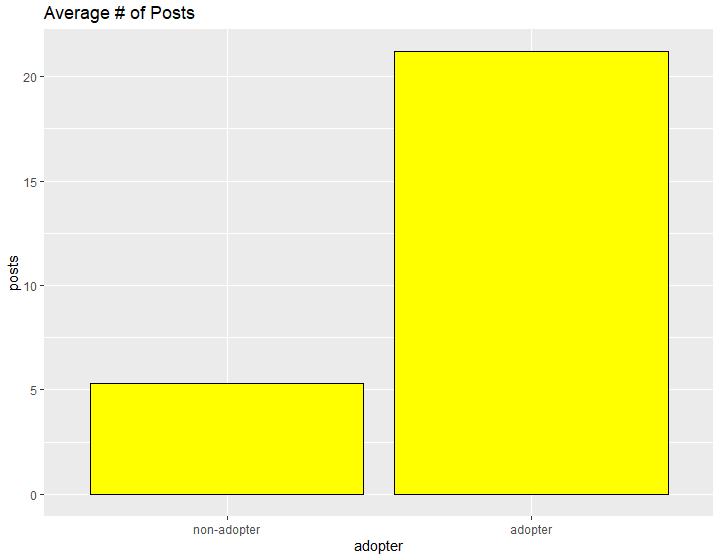
*(adopters have a much higher subscription rate as can be seen in the graph above)*

Assuming that subscription rate of friends can be defined with the equation: “subscriber friend count” divided by “total friend count”, the subscription rate of adopter group is more than twice as high when compared to the non-adopter group. To further clarify, a friend of an adopter is more likely to be (or become) a subscriber than a friend of a non-adopter. Additionally, the relationship between the number of friends and the number of friends who can be classified as premium subscribers is almost linear.

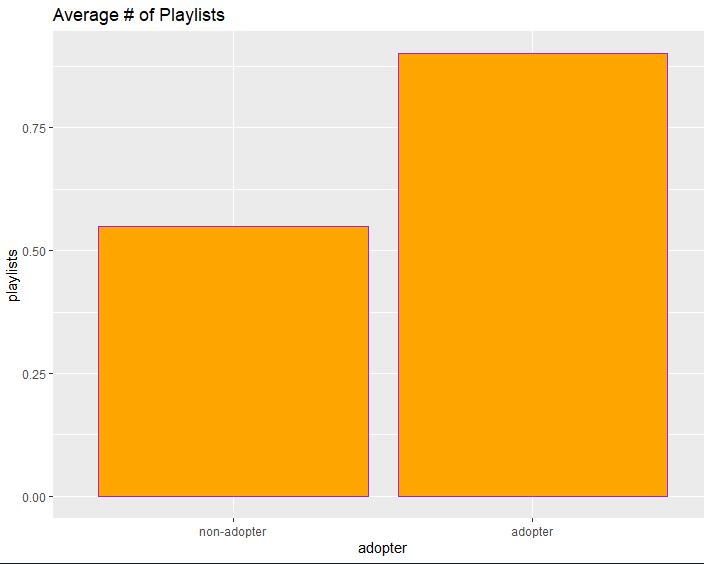
iii.User Engagement

*(this graph shows the listening behavior between adopters and non-adopters; as can be seen adopters on average listen to more songs)*

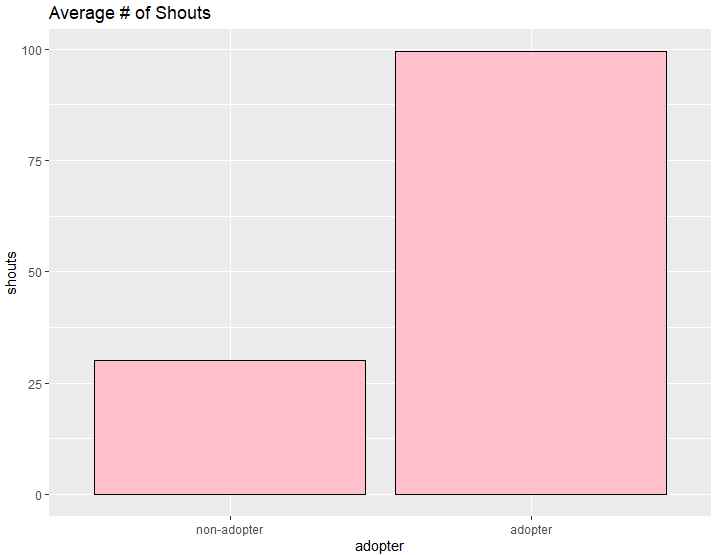
(*average number of loved tracks grouped by adopter vs non-adopter)*



*(average number of posts by adopter/non-adopter)*



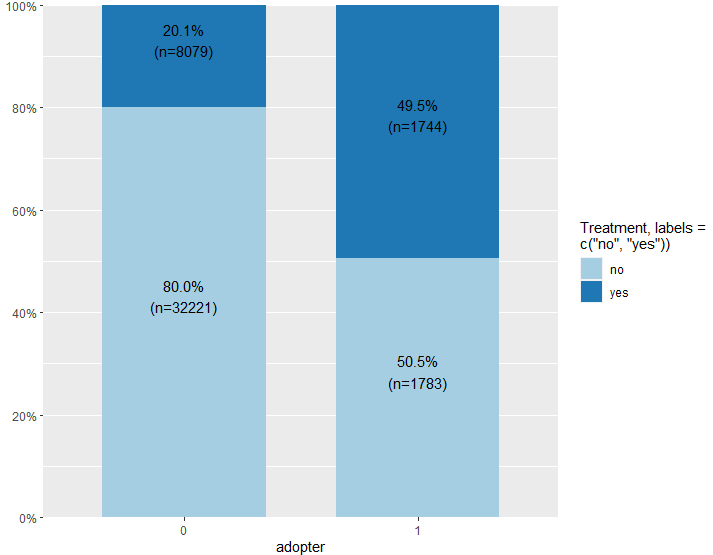
*(Number of playlists, broken down by adopter/non-adopter)*



*(Average number of shouts broken down by adopter vs non adopter)*

When looking at all of the graphs in question 2iii, every category/metric which could be used to track user engagement (ie number of shouts, playlists, posts, “loved” songs, as well as average number of songs listened to) has a clear skewing towards the adopter group. While this could possibly be attributed to the fact that people who are high traffic users on the site are more likely to become adopters, I think in this dataset that we can correlate adopters to being individuals who use the site more often rather than the other way around. The graphs clearly show that when looking at the average users, adopters have a far greater engagement with the site than non-adopters.

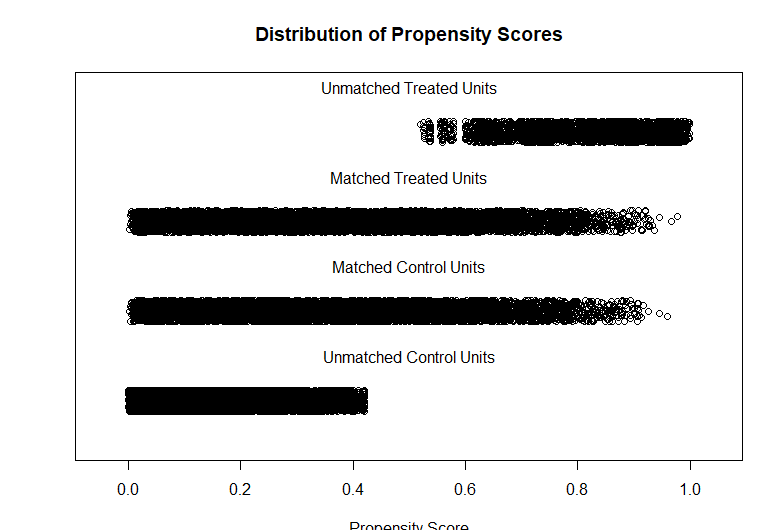
1. **Propensity Score Matching (PSM): You will use PSM to test whether having subscriber friends affects the likelihood of becoming an adopter (i.e., fee customer). For this purpose, the "treatment" group will be users that have one or more subscriber friends (subscriber\_friend\_cnt >= 1), while the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

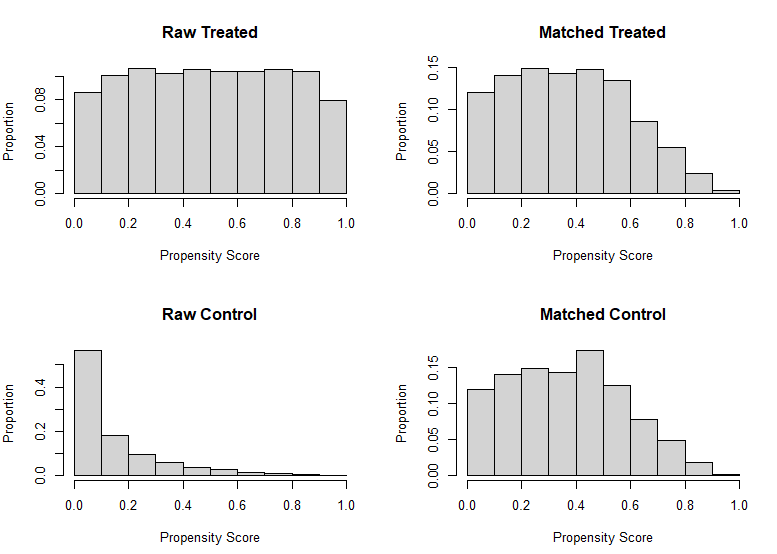
****

*(0 = non-adopter, 1 = adopter; no = no treatment, yes = yes treatment)*

We can see from this stacked bar chart that the adopter column has an even distribution between individuals from the treatment and non-treatment groups while the non-adopters are skewed, with the majority of the non-adopters belonging to the non-treatment group as well.

Some of the variables I chose to test include: friend count (friend\_cnt), friend country count (friend\_country\_cnt), songslistened, lovedtracks, posts, tenure and shouts. I chose these variables (alongside a few more) due to them being significant at the 0.05 mark on a linear model I ran.

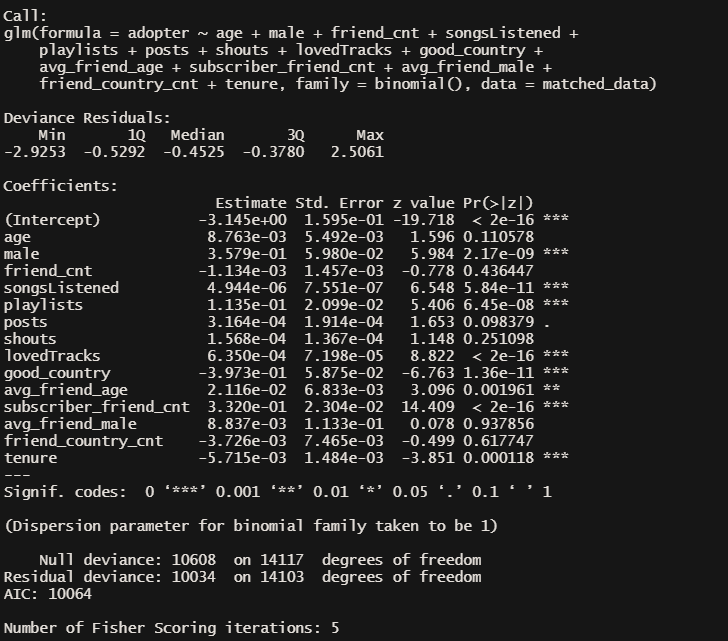
****

****

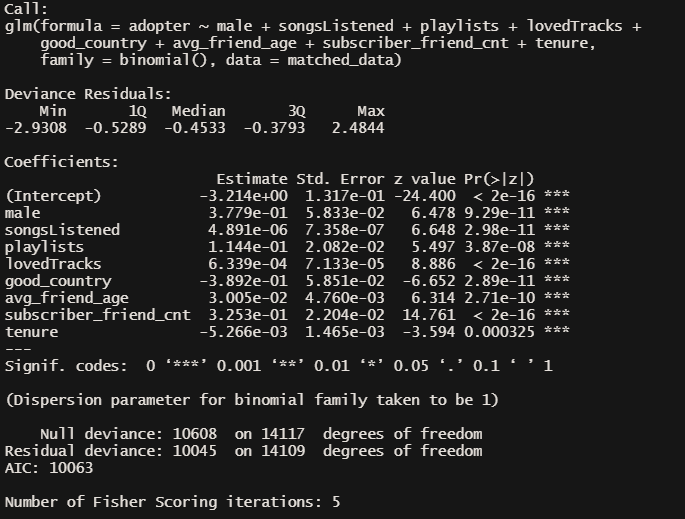
Using the matchit package (after logging variables to get rid of some bias) to pair “similar” individuals in both groups against each other we can see from the graphs above that the majority of the unmatched units trend towards holding a lower propensity score while the majority of the matched units trend towards holding a higher propensity score. From these tests (and a t-test result of <2e-16 from testing between treatment and adopter indicating that the treatment is significant) I believe that there is a causal relationship between having friends who are subscribers and the chances of becoming an adopter; or in other words the more subscriber friends a person in the dataset has the higher their chances of becoming/being an adopter. To get a numeric idea, I ran t-tests for all the variables to see if there is still a significant difference in means of the variables post being matched. Working under a 95% confidence threshold, I found that the p-values are greater than 5% for all the variables which implies that the difference in means is not statistically significant. In conclusion, after comparing the means between adopters and non-adopters I believe that the mean of non-adopters increased, while the number of adopters remained the same. I interpret this to mean that having no treatment increases the chances of a subscriber to be a non-adopter; however, presence of treatment does not necessarily “guarantee” an increased chanced of being an adopter.

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

My first model I ran with all of the variables (output below), however noticed that there were a few insignificant variables (ie like age, friend\_cnt, friend\_country\_cnt, and avg\_friend\_male) which I decided to remove to try and improve model accuracy/reduce penalty associated with higher number of variables.

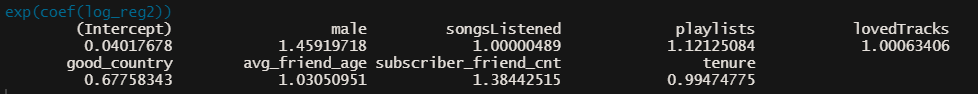


After removal of insignificant variables:



Examining the variables we can see that variables: male (ie whether the subject is male), songsListened, playlists, lovedTracks, avg\_friend\_age, and subscriber\_friend\_cnt all have positive slopes, meaning that an increase in that variable also increases the odds of an individual being or becoming an adopter while good\_country, and tenure have negative slopes which mean that for every increase in those variables there is a negative effect on the odds of an individual becoming or being an adopter.

It is also worth noting that in regards to the variables songsListened, lovedTracks, and tenure that the increase/decrease in odds of the dependent variable is quite small; thus if increasing these metrics (in the case of positively sloped variables like songsListened and lovedTracks) or decreasing the negatively sloped variables (in this case tenure) in an individual customer is expensive it may not be worth the investment due to low cost vs reward ratio. Below is a visual of the exponentiated intercepts from the previous model to estimate the odds ratio for each.



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

Ignoring the possibility for differing costs for reducing/increasing (as desirable) each variable used to predict whether an individual will become an adopter or not, we can see from the descriptive stats that the average adopter is in their twenties, male, and has many friends who are also subscribers. Thus the “free-to-fee” conversion strategy should focus on this demographic in the non-adopter category. The company could try a number of things ranging from increased ad showing to this group, to strategies which incentivize/reward adopters who bring their friends onto the platform themselves.

A different strategy which could be used to inform marketing campaigns could revolve around narrowing ad focus to free users/non-adopters who use the site a lot and convert them from being free to fee based on traffic to the website. When comparing adopters vs non-adopters in every metric used to track “engagement” of an individual with the website, adopters outperformed non-adopters (ie adopters have a greater number of shouts, playlists, listens etc) so another way the company could try and market could be to target those individuals in the free category with messaging intended to cajole them into paying for a service they use so often.

Finally, we can also use customer location to inform marketing strategy. The variable good\_country signifies whether an individual is from the US, UK or Germany and also has a negative slope when examined in the light of it being a predictor variable for adoption of the platform. This means that individuals in the above mentioned countries were less likely to become an adopter, so another effective strategy may be to market to users outside of these locations to try and get increased conversion from non-adopter to adopter.