I Code Summary

The attached code comprises of three pieces of separate sections:

1_load_data loads all the source data and performs simple analysis on whether gender plays a role in determining the user behavior.

2_Session_Analysis analyzes the users' listening patterns and groups user listening to sessions, which can be thought of as a continuous sequence of individual songs played. The code also discovers how different demographic factors can alter users' listening patterns.

3_Regression conducts a naïve regression analysis on the link between the users' listening pattern and their demographic information and their willingness to pay for the premium product.

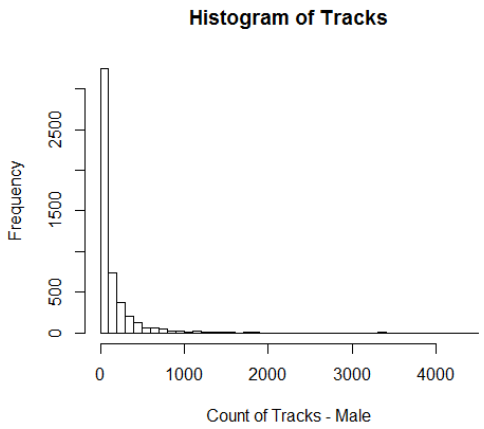
II Analysis 1 – Are Male and Female Listeners Significantly in Their Overall Listening?

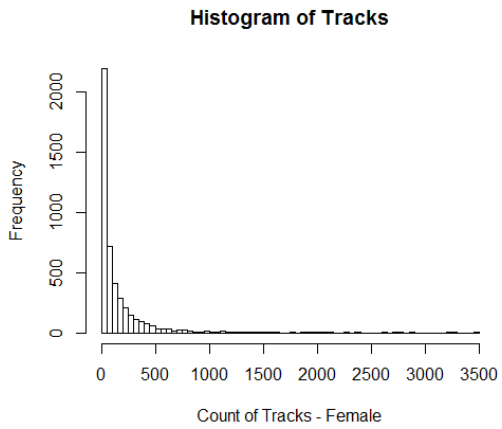
The answer is No.

The distribution of the total number of tracks/mins a listener spends is similar to a chi squared distribution with the counts floored at zero and long right tail.

The average tracks listened by a male and female listeners are:

```
    gender        V1
1:  female 142.36360
2:    male 138.87909
3: unknown  85.92308
```

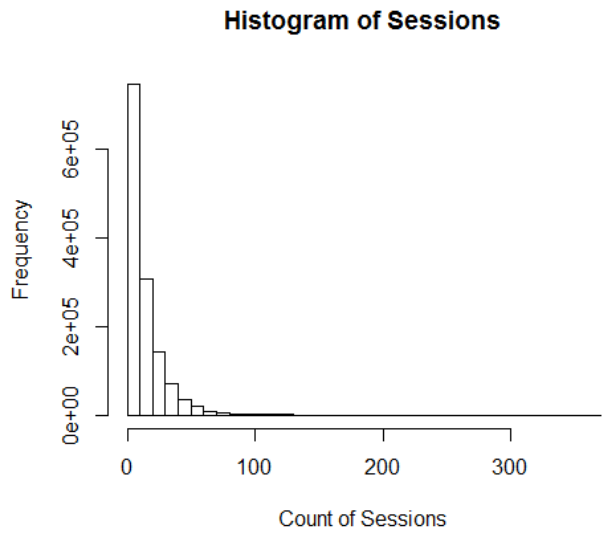




II Listener Session Analysis

Approach to breaking the listening to sessions

1. Find the median length of gaps between two end timestamps.
    a. The median gap is 210 seconds
2. Group different raw timestamps to sessions.
    a. If the gap between two consecutive timestamps > N * median gap, we think these two timestamps are in different sessions.
    b. In this exercise I use N = 5. This is an arbitrary assumption. Higher N results in fewer sessions which is more conservative in counting the total number of sessions.



Correlation between Users' Demographic Information and Sessions

a. Gender

Gender doesn't seem to matter which is consistent with our earlier finding.

```
    gender       V1
1:    male 12.56471
2:  female 12.37187
3: unknown 11.82223
```

b. Age

People between 25 and 54 on average spend more time than the rest of the population.

```
   age_range       V1
1:           13.53544
2:   25 - 29 13.15713
3:   45 - 54 13.14233
4:   30 - 34 12.76832
5:    0 - 17 12.39308
6:   35 - 44 12.36747
```

```
7:       55+ 12.35958
8:   18 - 24 11.93202
```

c. Country

I group the countries to four regions, North America, Europe, Latin America, and Other. Users in North America and Europe spend most time on the app.

```
          region       V1
1: North America 12.88451
2:        Europe 12.70678
3:         Other 12.03623
4: Latin America 11.38453
```

III How to Make a User Pay for the Product?

I conduct a logistic regression to try to understand what makes a user willing to pay for the premium account. Factors considered include demographic information (age, region, gender) and user behavior information.

The predictive power of the logistic regression is determined using AUC. In this case the AUC is 0.659 which suggests the model has some predictive power.

Model Result

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.3276213  1.2213752   0.268   0.7885
gendermale           0.0953533  0.0552601   1.726   0.0844 .
genderunknown       -1.7319737  1.0685131  -1.621   0.1050
age_range0 - 17     -3.1057825  1.2243018  -2.537   0.0112 *
age_range18 - 24    -2.7712064  1.2204168  -2.271   0.0232 *
age_range25 - 29    -2.3545968  1.2207596  -1.929   0.0538 .
age_range30 - 34    -2.2260210  1.2212993  -1.823   0.0684 .
age_range35 - 44    -2.0938065  1.2208168  -1.715   0.0863 .
age_range45 - 54    -2.2272184  1.2222908  -1.822   0.0684 .
age_range55+        -2.1318215  1.2233278  -1.743   0.0814 .
acct_age_weeks       0.0036330  0.0003433  10.581   <2e-16 ***
num_session          0.0144607  0.0015350   9.421   <2e-16 ***
regionLatin America -0.0668325  0.0953000  -0.701   0.4831
regionNorth America  0.1523360  0.0659772   2.309   0.0209 *
regionOther          0.2083708  0.0853579   2.441   0.0146 *
```

Interpretation of the model

The vintage of the account (number of weeks since the account was open) and the frequency of using the app are the strongest factors. Intuitively, it makes sense because the longer and more often a user has used the product, the more likely he or she is willing to pay for it.

- Young people under age 24 tend to use the free version possibly because of their limited financial resources.
- People in North America are more willing to pay for the product than people in Europe and Latin America are.
- There is a slight more chance a male user will pay for the product than a female but this needs further validation.

Suggestion to the business

- Cash Cow. Monetize on users over age 25 in North America as they are the most likely paid users.
- Forward Looking. Attract young users even they are not paying for the product now. It is worth waiting for them to become more economically established.
- Growth Opportunities. Try to understand what holds users in Europe and Latin Americ back from paying for the product. Should we provide more features? Or should we make the free version a little bit worse (price discrimination)?