Music App Experiment

Danny Li

I. Executive Summary

Based on the results of my analysis, I would recommend launching this change; with the caveat of further research if possible. Initial analysis of the experiment shows that the average time spent listening during the experiment is smaller in the treatment group than the control group. When comparing average minutes listened by gender, region, and age, time spent listening dropped nearly across the board. Statistical analyses in R shows that the smaller average in the treatment group is statistically significant. Therefore, the experiment did have a significant impact on minutes listened, but mostly caused users to listen less.

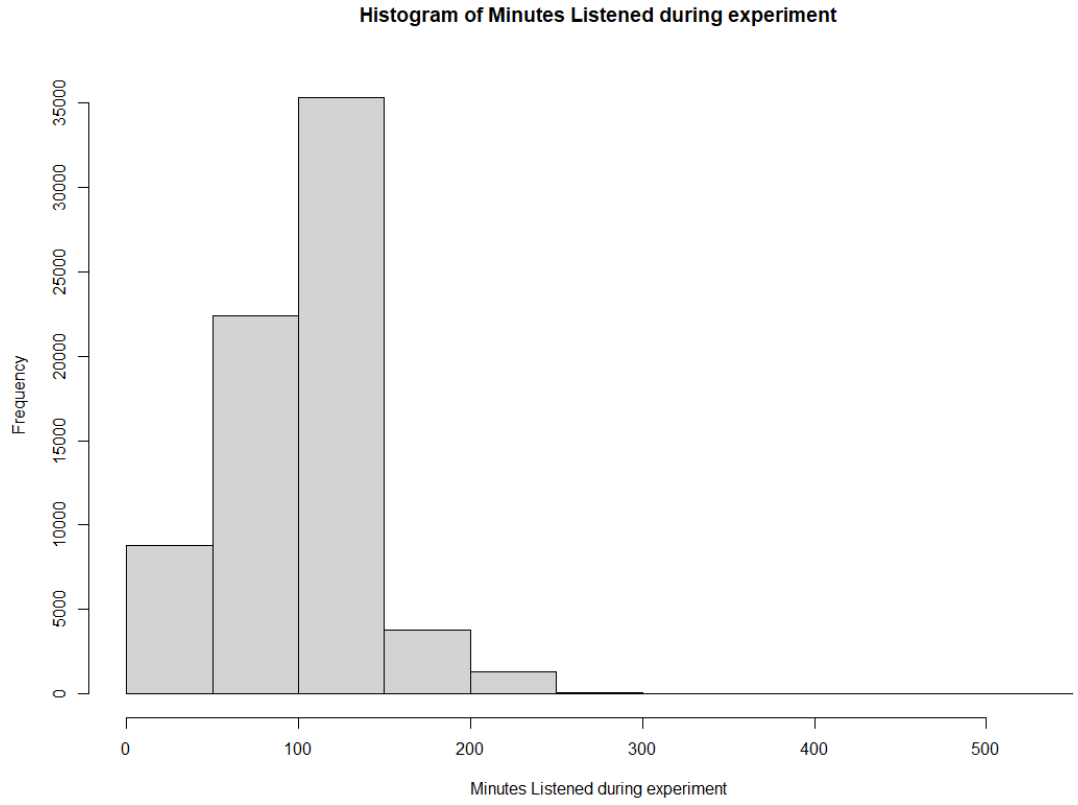
Despite the grim initial outlook, further statistical analyses indicates that device type has a statically significant impact on minutes listened with 95% confidence. With all other factors held constant, the experiment resulted in an increase in average minutes listened across the board. This seemingly contradictory result resulted from a skewed sample. Average minutes listened by desktop users is by far the largest, with averages of 125 and 160 minutes in the control and treatment groups, respectively. The average minutes listened by control and treatment groups is skewed when observing other parameters because the proportion of desktop users in the control group is very high at 67%, but only 32% in the treatment group. This information, combined with the fact that the treatment group is only around 20% the size of the control group, is what leads to our seemingly contradictory results.

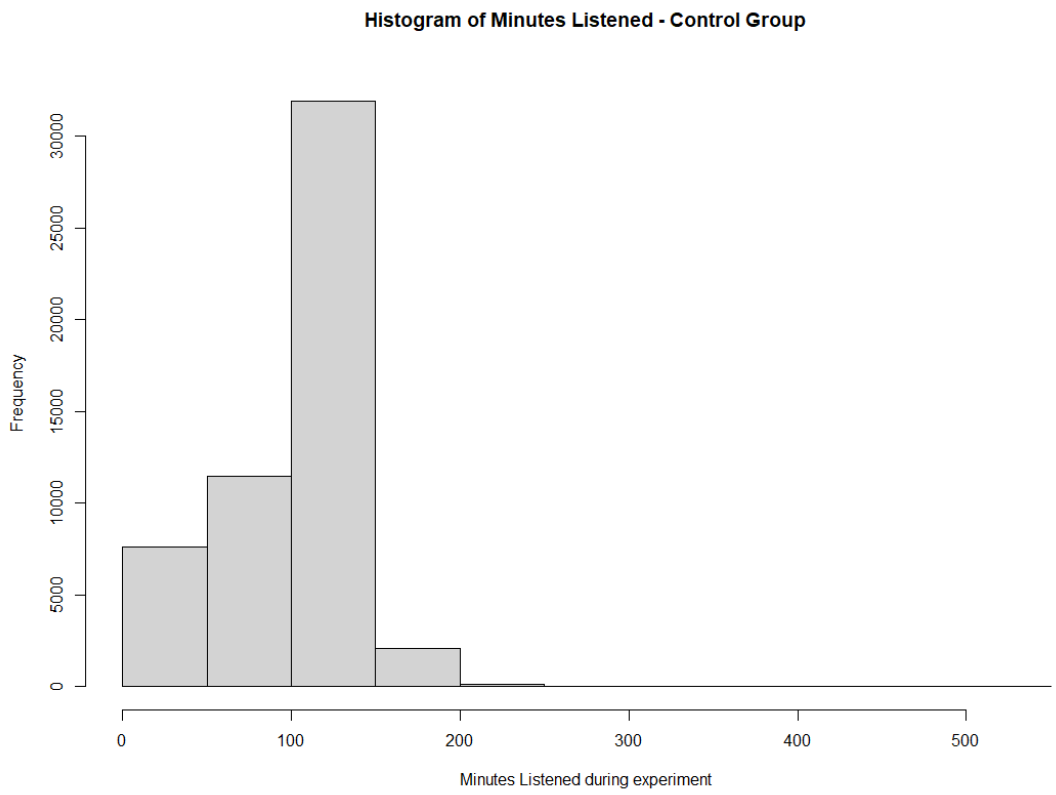
II. Analysis and Results

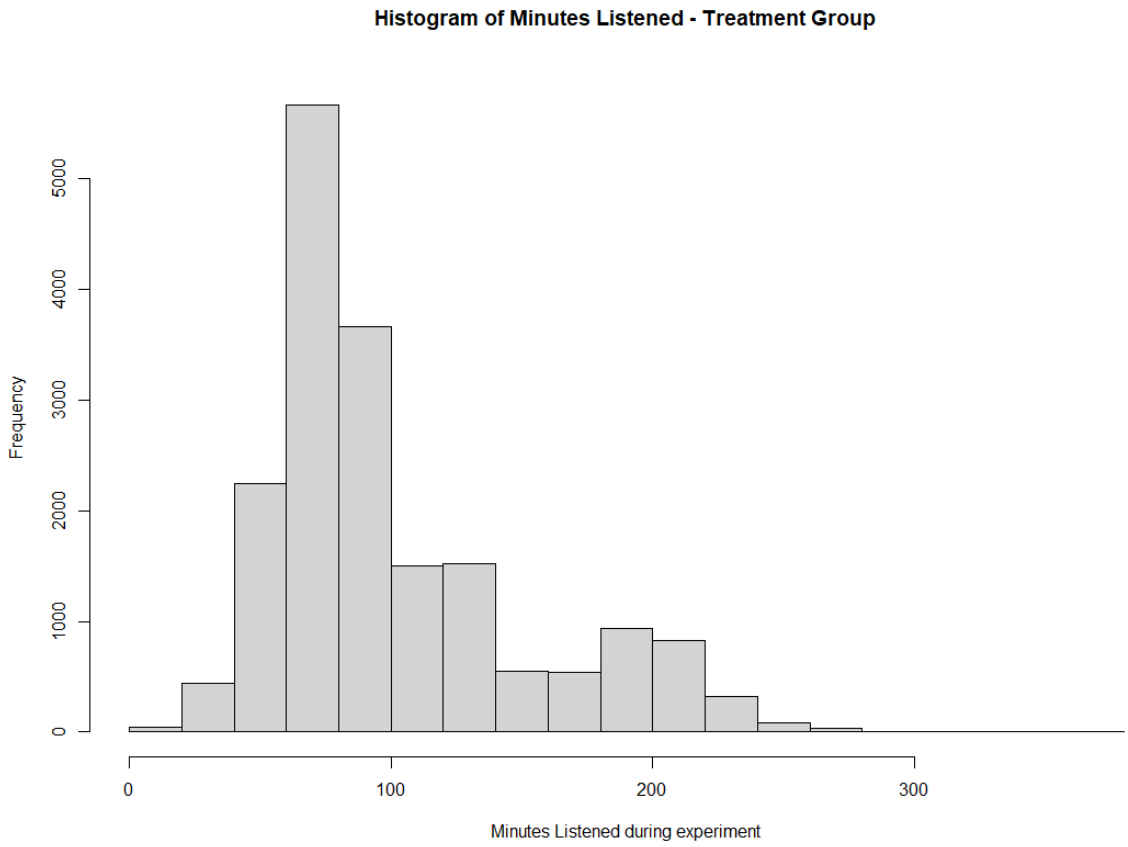
During my analyses, I made several assumptions for the sake of time and simplicity, listed below:

1. Since sample size is very large, I applied the Central Limit Theorem. Therefore, I am assuming sample mean is approximately normally distributed.
2. Independence of observations – assumed independent variables are not dependent on one another.
3. Homoscedasticity – the variation around the averages for each group are similar among all groups
4. Outliers observed are valid results of the experiment

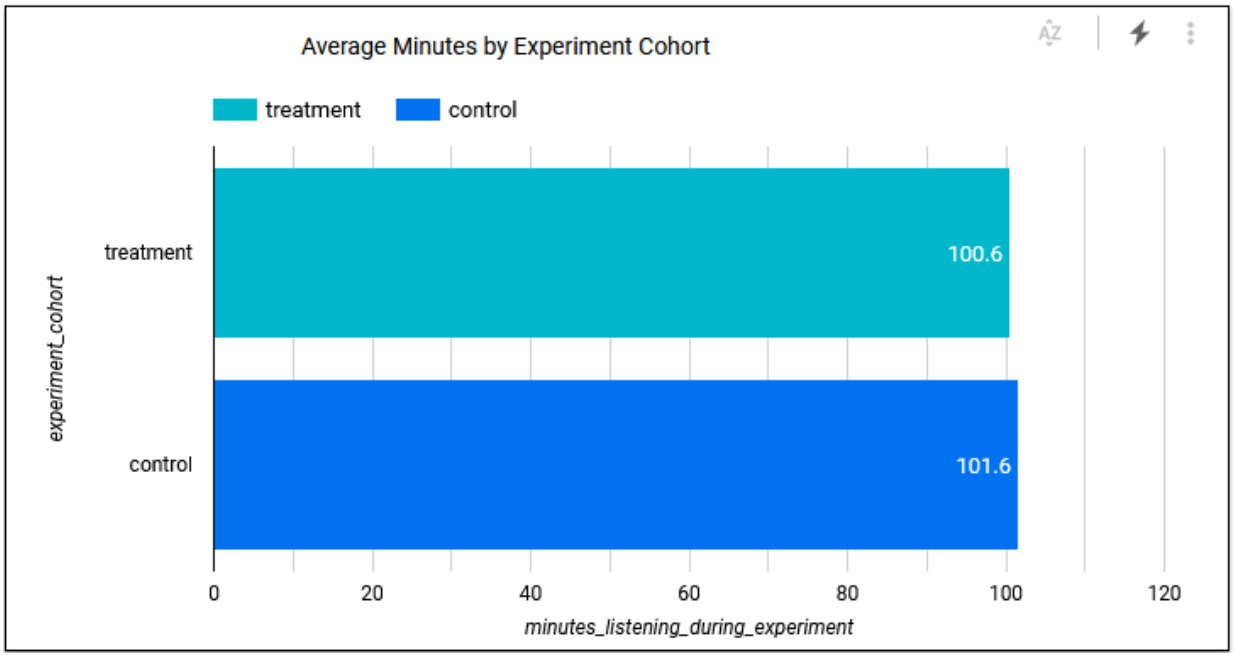
First, I examined the results of the experiment in R, which I also used for statistical analyses. I observed that most participants spent between 50 and 150 minutes listening during the experiment. I also observed some outliers, members of the control group who listened to music for over 8 hours a day. I chose to keep them (refer to assumption #4).

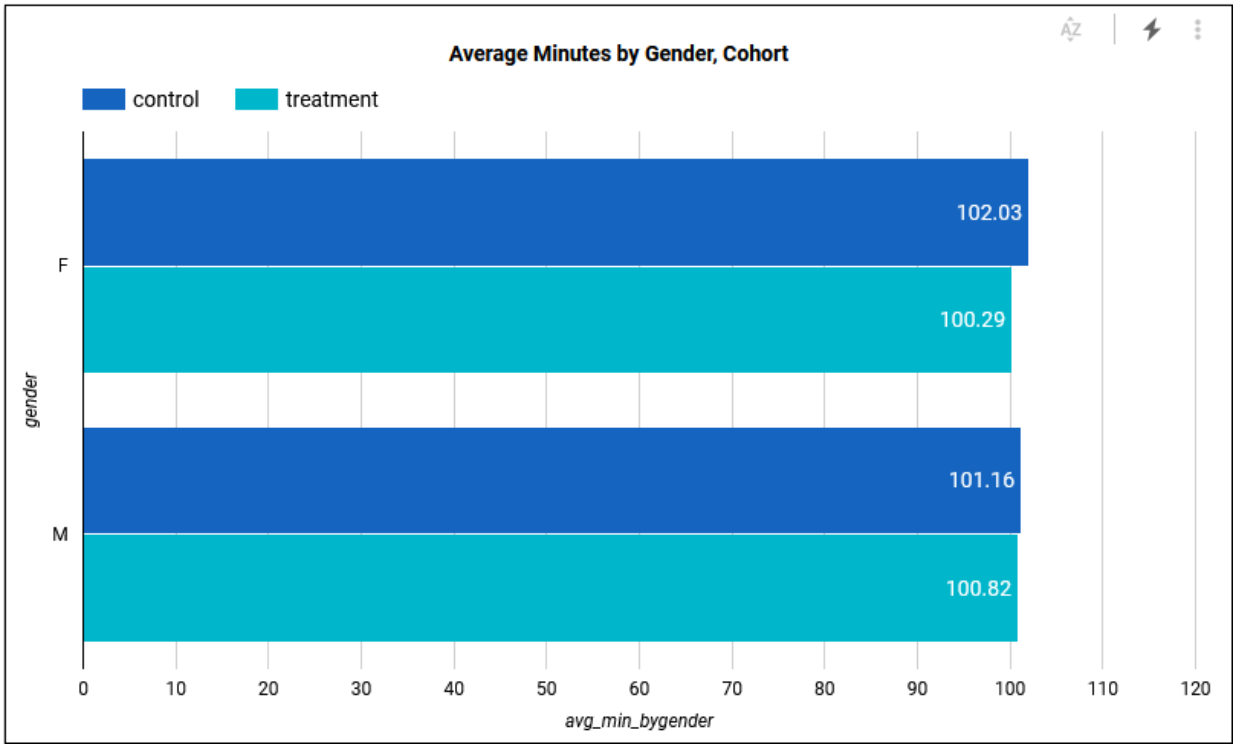


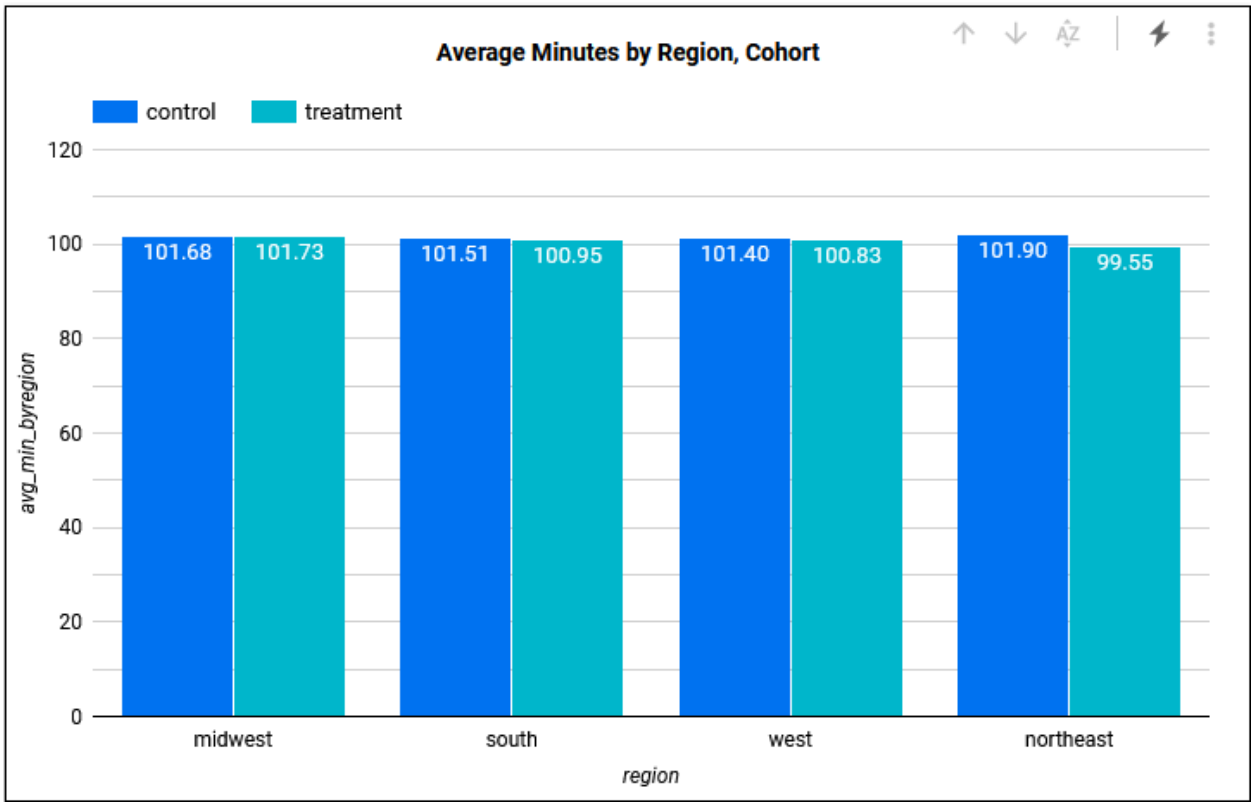

Histogram of Minutes Listened - Control Group
Frequency
0
5000
10000
15000
20000
25000
30000
0
100
200
300
400
500
Minutes Listened during experiment


Histogram of Minutes Listened - Treatment Group
Frequency
0
1000
2000
3000
4000
5000
0
100
200
300
Minutes Listened during experiment

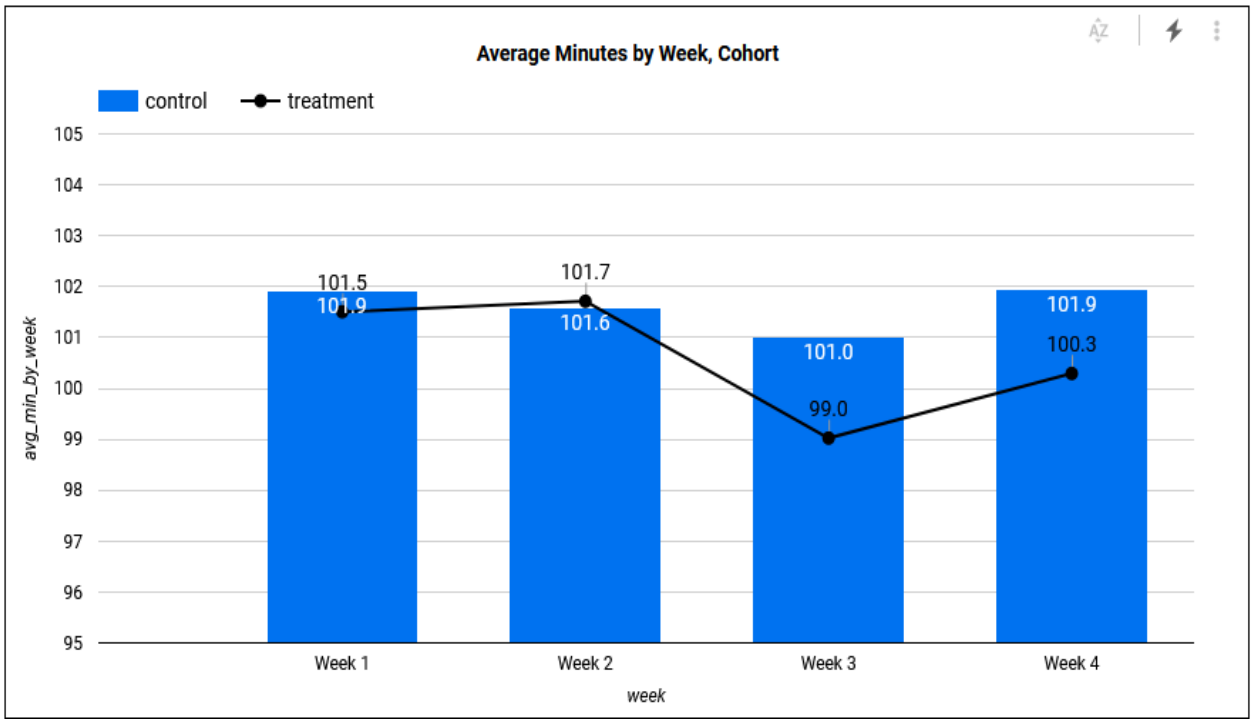
Then, I segmented the data to study the results of the experiment by category. The following charts were created in Google Data Studio for ease of access. We can see that when the results are categorized by different parameters, the outlook isn't good; the average minutes listened from the treatment group decrease when segmenting by gender, region, and week of the experiment.



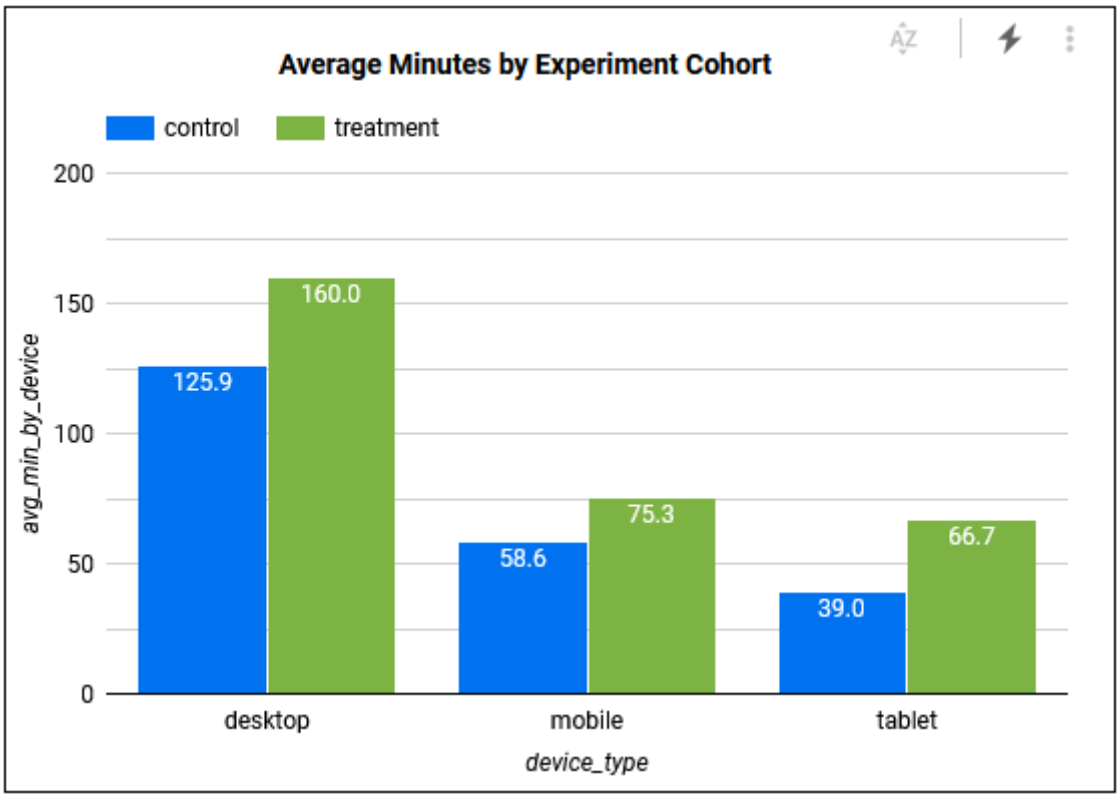




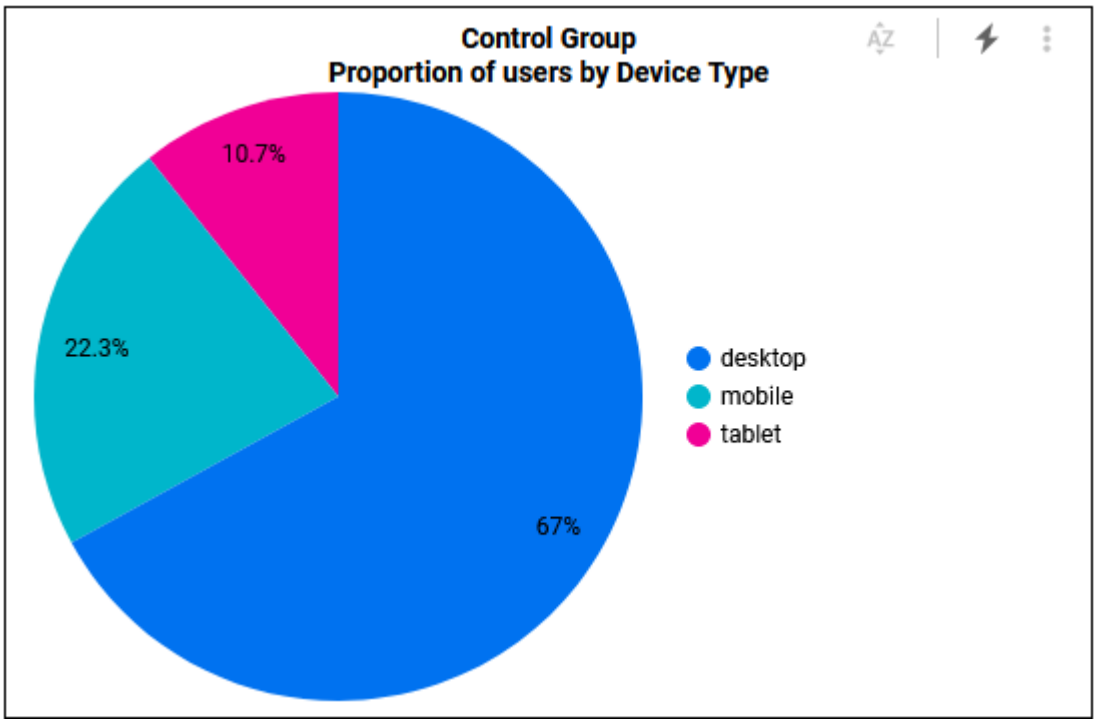
The Midwest region showed a slight increase. Whether or not the increase is statistically significant is examined later.

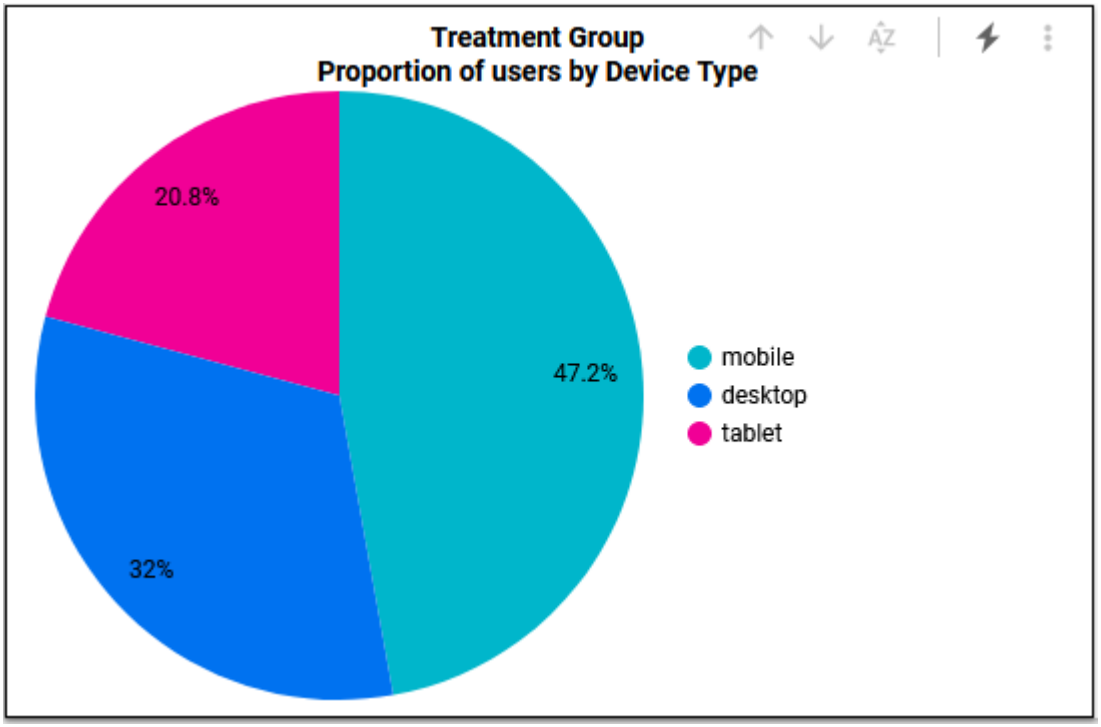


However, when segmenting by device type, it appears that the experiment has been successful! I observed significant increase in the average minutes spent listening in the control group. This will also be tested later in R.



Further observations into the data revealed the reason that we saw different results when segmenting by device type. As seen in the chart above, participants on desktop devices had significantly higher average minutes than those using other devices; on average, desktop users spend at least twice as much time listening. However, the ratio of desktop users to other devices is skewed; there were many more desktop users selected as part of the control group than the treatment group. This brought up the average minutes listened in the control group when compared to the treatment group, which had (as a percentage), less than half the proportion of desktop users selected.





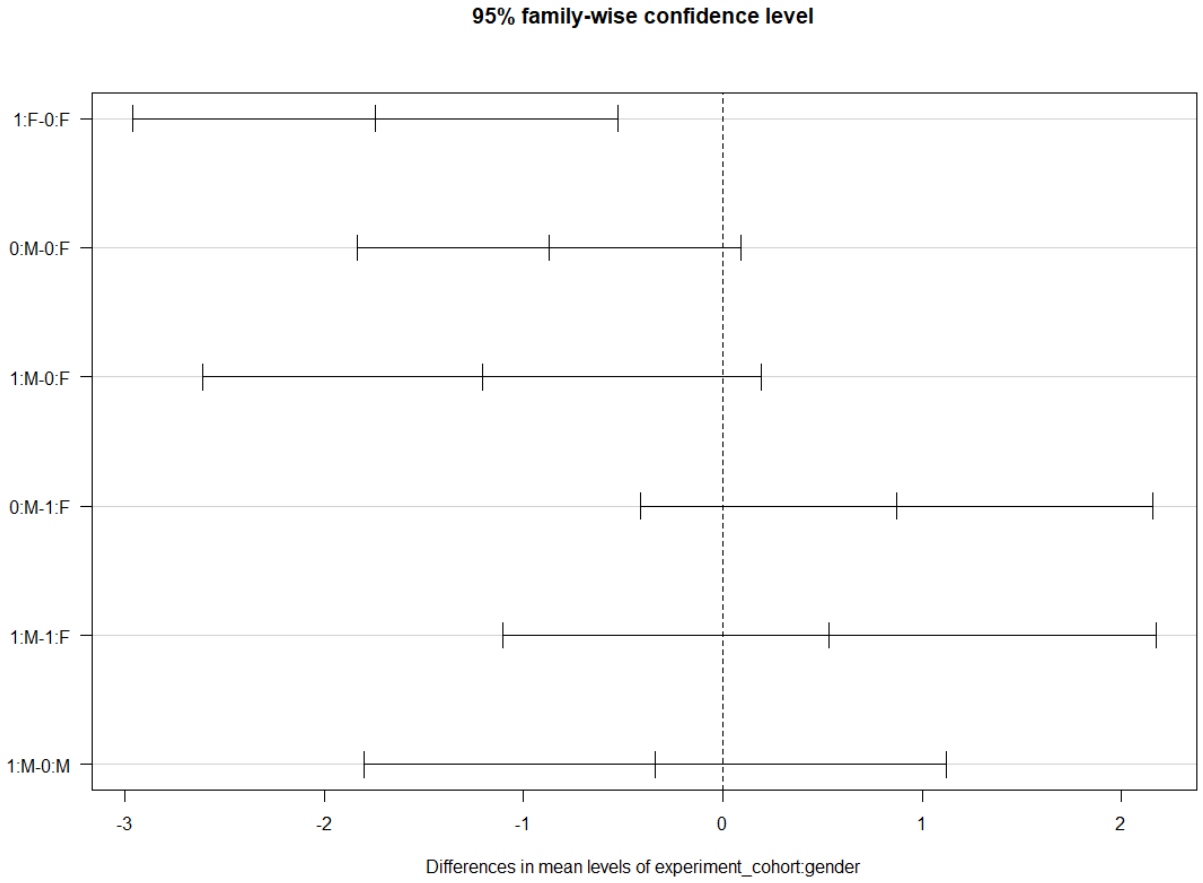
Next, I used R to perform some statistical analyses. First, I performed a two-sample t-test to check whether the difference in average minutes listened between the control and treatment group is statistically significant. The low p-value of 0.0128 indicates that it is 95% likely that the new user experience influenced the average minutes listened during the experiment, and that our results weren't due to chance alone. However, the t-value of -2.488 indicates that the experiment resulted in a decrease of average minutes listened.

```
data:  t_data$minutes_listening_during_experiment and c_data$minutes_listening_during_experiment
t = -2.4882, df = 26628, p-value = 0.01284
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.7805548 -0.2114182
sample estimates:
mean of x mean of y
 100.6232  101.6192
```

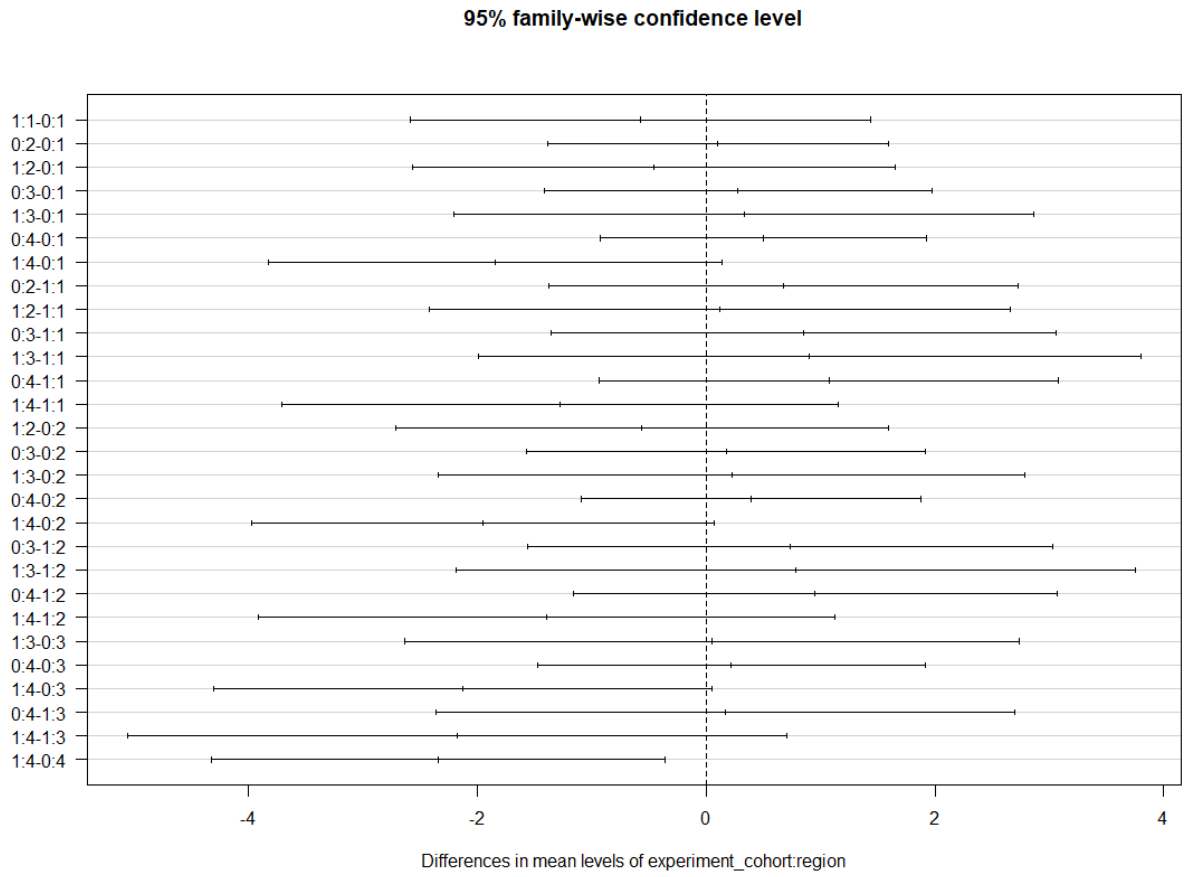
Next, I tried removing the 2 participants who spent over 8 hours a day listening from the control group to see if that would change anything. This proved uneventful, as both the t-value and p-value are largely unchanged.

```
data:  t_data$minutes_listening_during_experiment and c2$minutes_listening_during_experiment
t = -2.4517, df = 26594, p-value = 0.01422
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.7653472 -0.1967478
sample estimates:
mean of x mean of y
 100.6232  101.6043
```

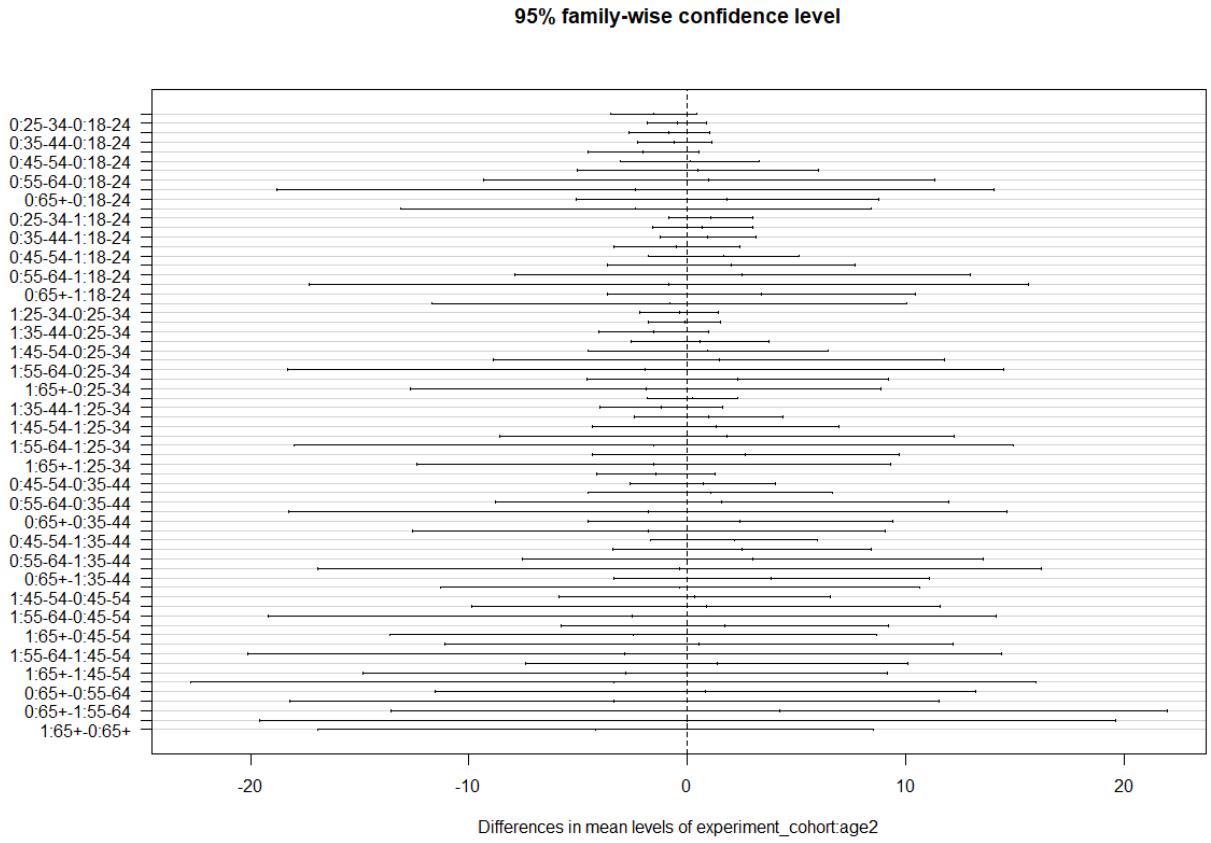
I then applied two-way ANOVA (analysis of variance) to test whether and how the independent variables affected minutes listened. When observing the data, I noticed that some users were identified with an 'X', as opposed to 'M' for male, and 'F' for female. I chose to exclude these data points as there is no reasonable way to me to extrapolate the gender. I created the charts below to find out which group means are statistically different from each other. The y-axis in the charts below follow a format of 'A:B-C:D', where 'A' and 'C' represent the group being tested (in our case, control group was assigned a value of zero, and the treatment group was assigned a value of 1), and 'B' and 'D' represent the subset of the parameter being tested. For example, in the first chart we analyze the difference in means for gender; the top-most value in the y-axis of '1:F-0:F' means we are comparing the means for female users in the treatment group with means for female users in the control group.



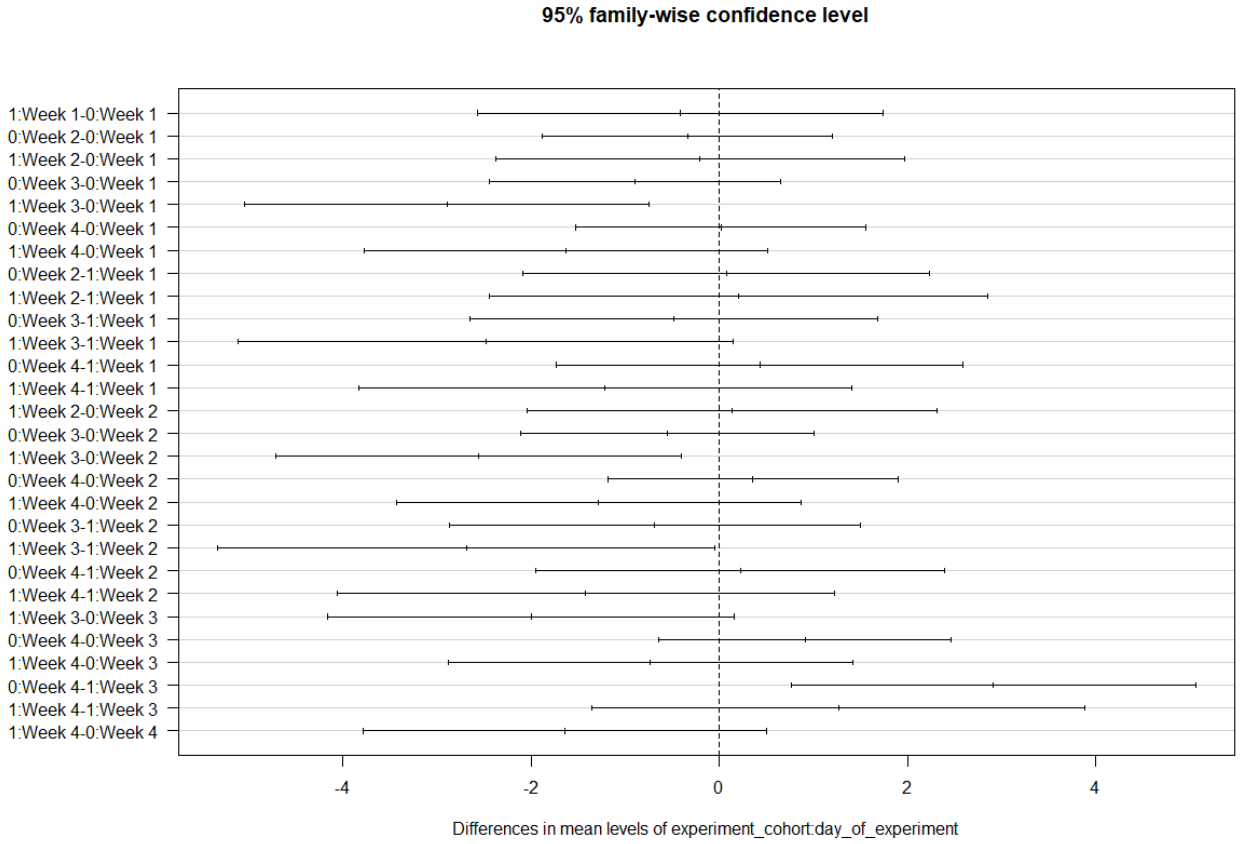
The horizontal lines in the chart above represent the confidence interval. Because we chose a 95% confidence interval, the lines represent the range of values where the true difference in means lies with 95% confidence. If the line crosses the value of zero in the x-axis, it means that the difference in means is not statistically significant for our confidence level of 95%. It appears that the results for gender are not very significant, with the only significant differences being between females in the control and treatment group, which also shows a decrease in average minutes listened.



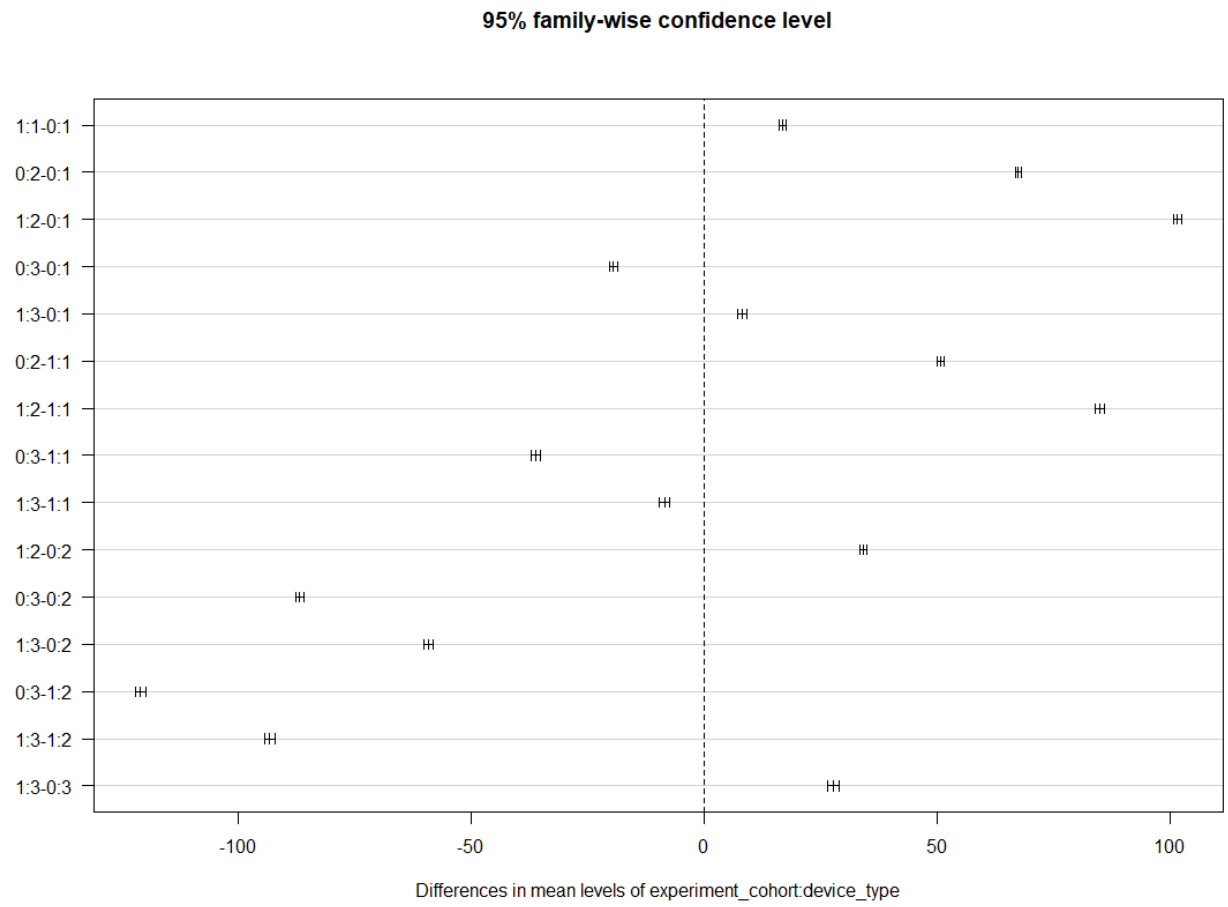
Regions were mapped numerically, with ‘1’ representing the west, ‘2’ representing the south, ‘3‘ representing the Midwest, and ‘4’ representing the northeast. The chart for region shows similar results; most combinations of experiment cohort and region did not yield significant differences, except for the northeast region, which again shows a decrease in average minutes listening.



This chart shows the difference in means between different combinations of the experiment cohort, and age. Ages were split and categorized into the groups: ’18 to 24’, ’25 to 34’, ’35 to 44’, ’45 to 54’, ’55 to 64’ and ‘65+’. I did not include ages 17 and under as the data only had individuals aged 18 or older. I observed no significant differences when considering age as a factor.



Next up is the time of the experiment. I grouped the days of the experiment into 4 categories: days 1 though 7 are represented as ‘Week 1’, days 8 through 14 as ‘Week 2’, and so on. The chart shows significant negative differences when comparing week 3 of the treatment group with weeks 1, 2, and 3 of the control group. However, it is difficult to ascertain whether these differences were caused by the novelty effect, when users are resistant to changes, and thus respond negatively.



Above is a chart for analyzing the effect of device type. Device types were mapped to the following values: '1' for mobile, '2' for desktop, and '3' for tablet. As we can see, all combinations of experiment cohort and device type had statistically significant differences. This chart suggests that device type has a significant impact on the amount of time spent listening in both treatment, and control groups.

Lastly, I ran a linear regression in R with minutes listened during the experiment as a function of all other variables to check my ANOVA results.

```
Call:
lm(formula = minutes_listening_during_experiment ~ ., data = d2)

Residuals:
    Min      1Q  Median      3Q     Max
-117.75  -11.58   -1.24    9.25  461.38

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 5.428e+01  5.449e-01  99.609   <2e-16 ***
experiment_cohort1          2.585e+01  1.817e-01 142.284   <2e-16 ***
day_of_experimentWeek 2     2.728e-01  2.133e-01   1.279   0.2009
day_of_experimentWeek 3    -2.496e-01  2.134e-01  -1.170   0.2421
day_of_experimentWeek 4    -1.932e-04  2.122e-01  -0.001   0.9993
device_type2                7.234e+01  1.795e-01 403.013   <2e-16 ***
device_type3               -1.496e+01  2.504e-01 -59.769   <2e-16 ***
genderM                     3.410e-02  1.569e-01   0.217   0.8280
genderX                     2.110e-01  3.726e-01   0.566   0.5713
region2                     2.213e-01  2.052e-01   1.079   0.2808
region3                     1.536e-01  2.336e-01   0.658   0.5108
region4                     3.173e-01  1.970e-01   1.611   0.1072
age                         1.713e-03  1.955e-02   0.088   0.9302
total_artists_last_100_days -9.850e-04  1.785e-03  -0.552   0.5811
user_id                     8.407e-10  4.100e-10   2.050   0.0403 *
age25-34                    7.376e-02  2.348e-01   0.314   0.7534
age35-44                   -3.098e-01  4.046e-01  -0.766   0.4438
age45-54                   -9.724e-02  6.685e-01  -0.145   0.8843
age55-64                   -3.012e-01  1.483e+00  -0.203   0.8390
age65+                     -6.496e-01  9.840e-01  -0.660   0.5091
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.19 on 71692 degrees of freedom
Multiple R-squared:  0.7654,    Adjusted R-squared:  0.7653
F-statistic: 1.231e+04 on 19 and 71692 DF,  p-value: < 2.2e-16
```

As expected, the experiment cohort and device type are the only statistically significant factors when studying effects on minutes listened. User ID was also marked with a ‘*’ to signify statistical significance, but we’re going to ignore that as User ID was only used for identification in our experiment.

III. Discussion

We found statistically significant differences in average minutes listened between the treatment and control group, meaning that our experiment influenced minutes listened. We also found that most factors that we considered except for device type did not have a significant impact on minutes listened. However, this could be because of the method used to select our sample. The sampling method of selecting every 5^{th} user from a randomized list could have been the reason for our seemingly contradictory results of decreased average minutes when segmenting by all categories except for device type.