Assessment 01 - Increasing User Engagement

Data Analysis by Alexander Raouf

Problem Statement

As our music listening app grows in popularity, one of our main high-level objectives is increasing user engagement, as measured by the number of minutes spent listening to music in the app. In conjunction with the data science team, our product team has developed an in app feature that they hypothesize will increase user engagement. After developing an experiment on active and eligible users, we have measured the effect of this feature on user engagement.

Proposed Solution

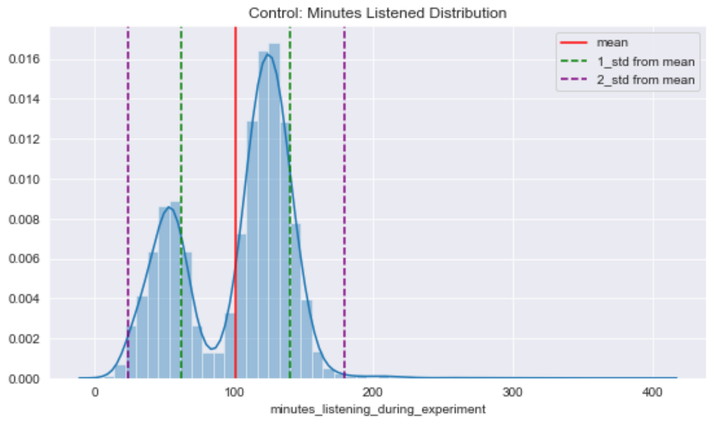
After performing analysis on the experiment data, I have concluded that the change under consideration should **not** be implemented. The data shows that the control group participants (users who did not use the new feature) listened to music on average a longer time in minutes than the treatment group.

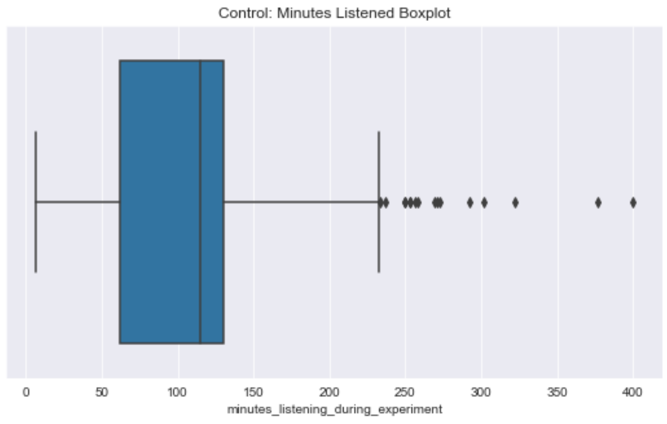
Detailed Report

The experiment that we developed was essentially an A/B test. Half of all participants received an in app experience with the new feature (the treatment group), while the other half of the participants had no change to their in app experience (control group).

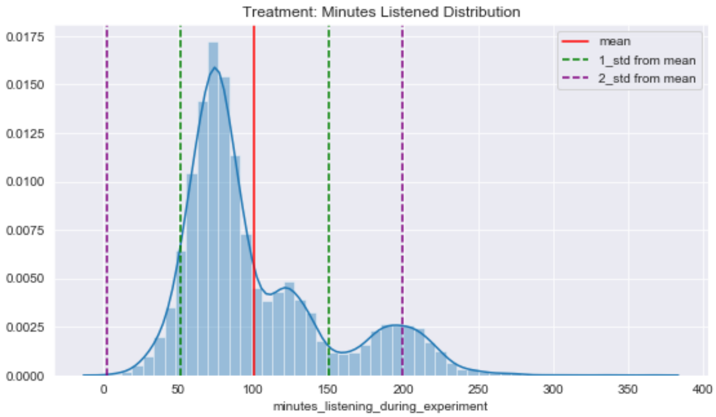
In order to determine which app experience had higher user engagement, we needed to use a metric that we could use to compare the two groups to each other. In this case I decided to use the mean, or average time in minutes each group spent listening to music on the app. On average the treatment group listened to music ~**100.63** minutes in a 24 hour window. The control group listened to on average ~**101.61** minutes in a 24 hour window. We can see the listening minutes distributions and box plots for both control and treatment groups below.

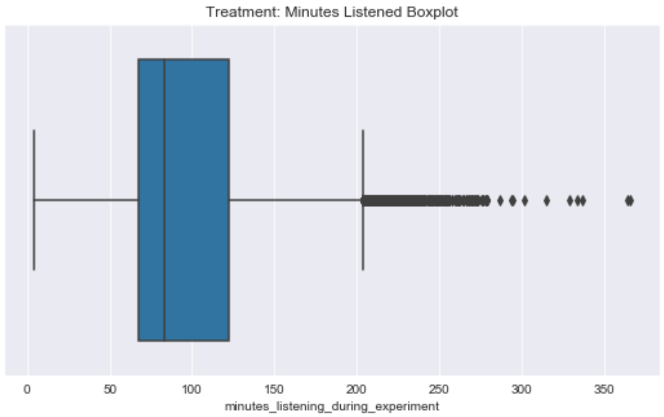
Control





Treatment





Although there is a small difference between the means of both groups, the control group is listening to music about 0.98 minutes on average. This would dispute our hypothesis that the treatment group would listen to music more minutes on average. In order to make sure the control group is *actually* listening to music more on average we have to test whether this difference is statistically significant, or if it's just by chance.

We can test for statistic significance by using a 2 sample t-test. In this instance since we are testing whether the treatment's average listening time is greater than the control groups (this was our hypothesis), we are using a 1-sided t-test. After performing the t-test, we received a resulting p-value of 0.0064. What this means is that there is a less than 1% chance that our results — that the average listening time for the control group is greater than that of the treatment group —is due to random chance, and thus confirms what is indicated in the data.

Since there seems to a lot of outliers in both groups, I also decided to perform a t-test on the data with the outliers removed. The result was an even smaller p-value than my previous test which further proves that the control group is in fact listening to music longer than the treatment group. This led me to the conclusion that the new feature is actually decreasing user engagement.

BONUS QUESTION

How many users does the business have in total on the last day of the experiment?
On the last day of experiment the business has **363,376 total users**. I arrived at that number by multiplying the total number of participants in our experiment by (100 / 19.73). Since about 19.73% of our eligible users were in our experiment, we can multiply the number of users in our experiment by (100 divided by 19.73).

Appendix

Sample Size + Potential Sampling Error

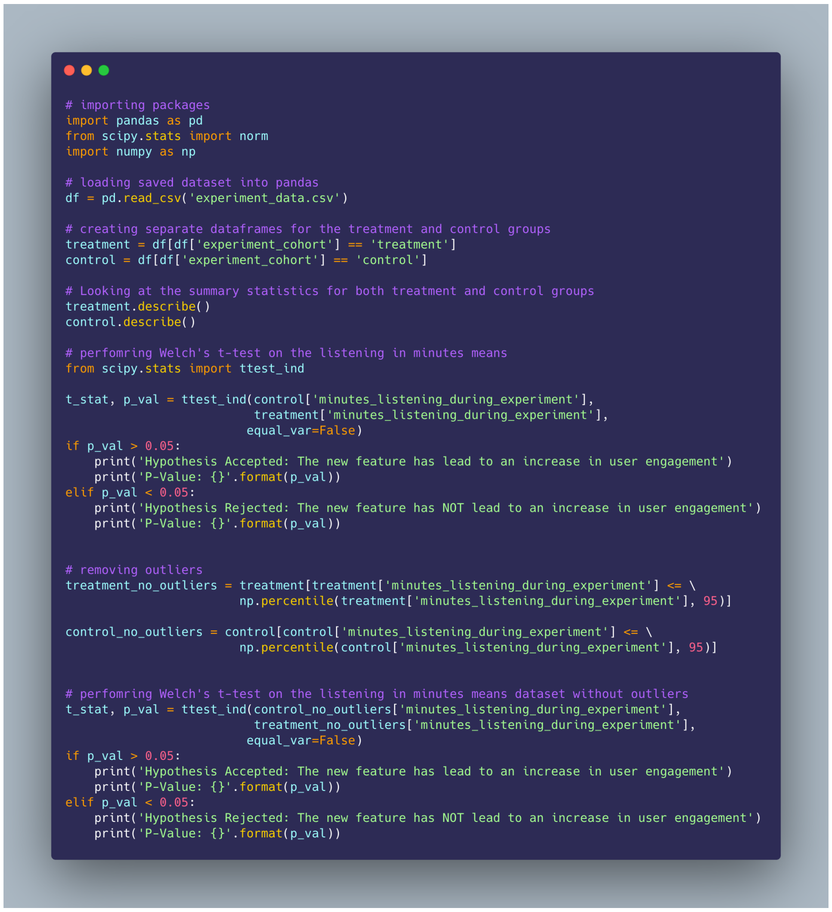
While Analyzing the data I noticed that the sample sizes were way off between the control and treatment groups. There were roughly ~53,000 control participants and roughly ~18,000 treatment participants. When performing a standard t-test you ideally want the sample sizes to be the same. If they are off it can lead to less reliable results. Since our experiment was designed to have participants be placed into either the control or treatment group based on 50/50 odds, we expect the group sample sizes to be relatively equal. However, there is a large disparity between the groups, and this leads me to believe there was an error in the sampling design of our experiment.

T-Test Information

The t-test used in my experiment was Welch's t-test. As described above in *Sample Size + Potential Sampling Error,* the sample sizes of our control and treatment groups were way off. In addition the listening time in minutes variance's between the two groups were not equal, and there was about a 20% difference between the control listening time in minutes variance and the treatment listening time in minutes variance. Luckily Welch's t-test accounts for non-equal sample sizes and non-equal variances, and allowed us to account for these disparities while performing a test on the means. I performed the t-test with a confidence interval of 95%, and therefore rejected the null hypothesis if the p-value was less than 0.05.

Python Code

Here is the python code I wrote to explore the data in addition to SQL, and perform testing.



```python
# importing packages
import pandas as pd
from scipy.stats import norm
import numpy as np

# loading saved dataset into pandas
df = pd.read_csv('experiment_data.csv')

# creating separate dataframes for the treatment and control groups
treatment = df[df['experiment_cohort'] == 'treatment']
control = df[df['experiment_cohort'] == 'control']

# Looking at the summary statistics for both treatment and control groups
treatment.describe()
control.describe()

# perfomring Welch's t-test on the listening in minutes means
from scipy.stats import ttest_ind

t_stat, p_val = ttest_ind(control['minutes_listening_during_experiment'],
                          treatment['minutes_listening_during_experiment'],
                         equal_var=False)
if p_val > 0.05:
    print('Hypothesis Accepted: The new feature has lead to an increase in user engagement')
    print('P-Value: {}'.format(p_val))
elif p_val < 0.05:
    print('Hypothesis Rejected: The new feature has NOT lead to an increase in user engagement')
    print('P-Value: {}'.format(p_val))


# removing outliers
treatment_no_outliers = treatment[treatment['minutes_listening_during_experiment'] <= \
                        np.percentile(treatment['minutes_listening_during_experiment'], 95)]

control_no_outliers = control[control['minutes_listening_during_experiment'] <= \
                        np.percentile(control['minutes_listening_during_experiment'], 95)]


# perfomring Welch's t-test on the listening in minutes means dataset without outliers
t_stat, p_val = ttest_ind(control_no_outliers['minutes_listening_during_experiment'],
                          treatment_no_outliers['minutes_listening_during_experiment'],
                         equal_var=False)
if p_val > 0.05:
    print('Hypothesis Accepted: The new feature has lead to an increase in user engagement')
    print('P-Value: {}'.format(p_val))
elif p_val < 0.05:
    print('Hypothesis Rejected: The new feature has NOT lead to an increase in user engagement')
    print('P-Value: {}'.format(p_val))
```