Feature Experiment

Introduction

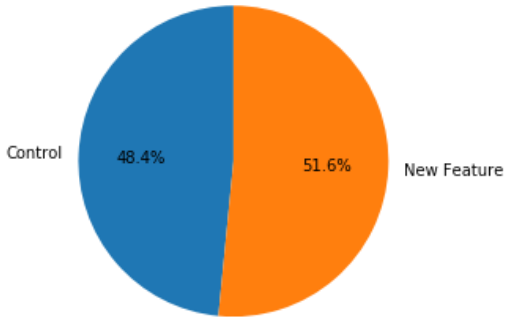
In this project, I analyzed how the enablement of a new limited time feature affected both player engagement and revenue. I used Excel and Python to process the data and conduct my analyses. First, I did a thorough check of the data to ensure that it was valid. I then performed descriptive statistical analysis on all of the features to extract basic insights. Next, I conducted hypothesis tests to see if there was an increase in Average Number of Games Played or Average Revenue while the new feature was enabled. Finally, I attempted a differences-in-differences analysis to determine if there was a causal effect from enabling this feature.

Data Processing

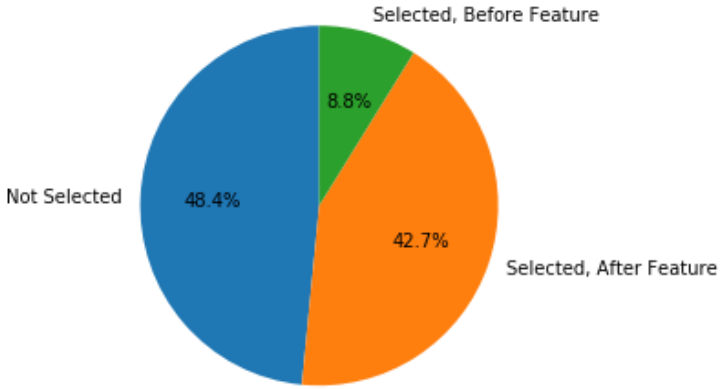
The dataset included 6 variables: User ID, a Feature Enabled dummy, the date a user was selected for the experiment, the date of the observation, number of games played, and revenue. The dataset should have one observation per user.

First, I checked the data using an Excel Pivot Table to ensure that there was only one record per player, per day. I found that User 2199 had 8 duplicate records per day. I deleted the duplicate records, leaving one record per day. Because these were exact duplicate records, I do not believe that any valuable information was lost from removing them.

Second, I checked for any outliers in the dataset. I found that user 4875 had 3 days where they had 999 games played. The next largest number of games played was 104 games. With 999 being significantly larger than 104 games, I felt that this was an error and removed the observations. Leaving this value in may have overinflated the average number of games played and also increased the variance which would later affect the hypothesis test.
Next, I checked how many observations were in the control and experiment groups. The data was split so there were 5837 observations in the control group, and 6211 observations that received the new feature.



However, some of the users selected to receive the new feature did not actually receive the new feature until a few days after being selected for the experiment. Thus, it was best to split the new feature group into observations that had received the new feature and had not received the new feature yet.



This meant the dataset now had 3 groups of observations. The first group is the group that never received the new feature, 5837 observations. The second group is observations of users who were selected to receive the new feature but before the received it, 1066 observations. The final group is observations of users who were selected to receive the new feature, after they received it, 5145 observations.

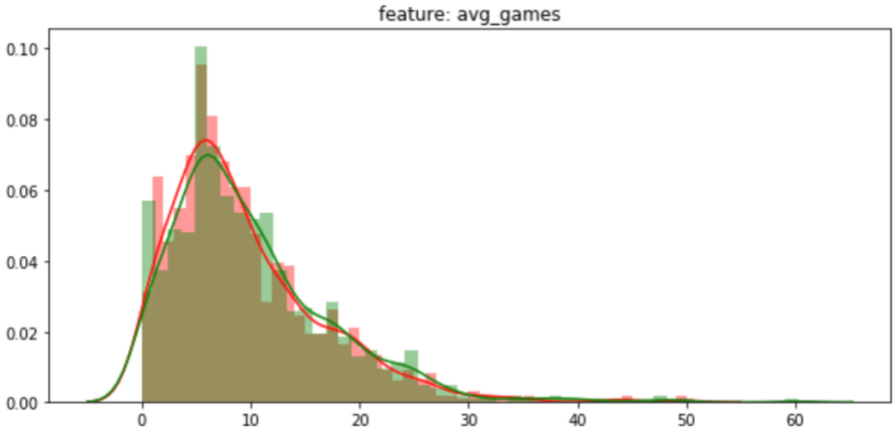
I also added two variables to the dataset for the difference-in-differences analysis. One is a numeric variable for days before or after a user was selected for the experiment. The second is a dummy variable for if an observation is before or after the user was selected.

For further analysis, except for the difference-in-differences analysis, I will be comparing the first group and third group to determine if the feature is a success. Although it may make sense just analyze the users who would receive the feature, before and after they received it, there could be endogenous effects working that could also affect their engagement with the game. This is what I will try to account for in the difference-in-differences analysis.

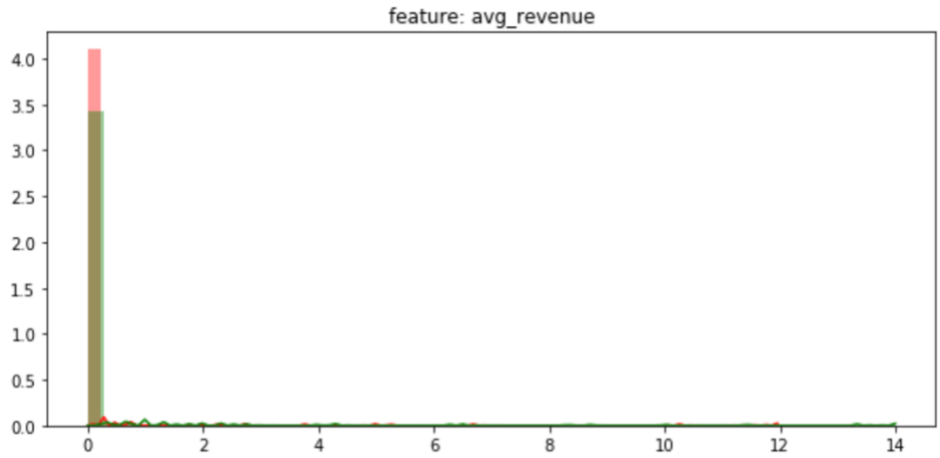
The last step I took to process the data is taking the average number of games played and revenue per user for observations in groups 1 and 3. This means that if a user was selected to receive the feature, I only included the observations after they received the feature to calculate their averages. By doing this, I can assume independence holds between users and conduct my later hypothesis tests. There were 987 users in the control group and 1023 who received the feature.

Descriptive Analysis

I first checked distributions of average number of games played and average revenue per user for each treatment group. In the below, red is for control and green received the feature.

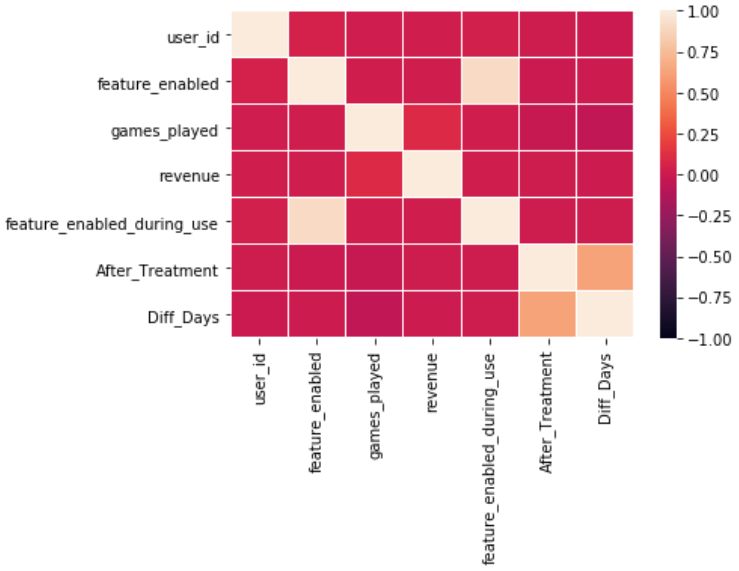


The average number of games per user is right skewed and follows a similar distribution for both treatment and control groups.



The average revenue per user is even more right skewed and while it is hard to tell, also follows a similar distribution.

Lastly, I wanted to check correlations between variables in the original dataset for our difference-in-differences analysis.



There does not seem to be much useful correlation between variables. The only variables with high correlation are all time related with respect to when a user was selected.

Hypothesis Testing

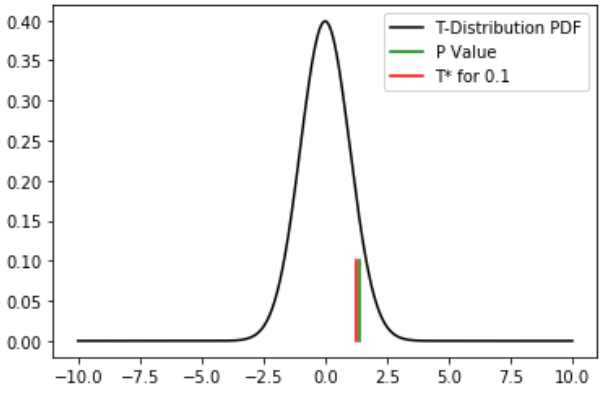
I conducted hypothesis tests to determine whether or not the average number of games played and revenue was significantly different between each group. The null hypotheses for all tests were that the observed averages were equal to each other. The alternative hypotheses were that the received feature group had higher averages than the control group. Below are the summary statistics per user:

Type	Observations	Avg Games Played	Avg Revenue	Games SE	Revenue SE
Control	987	9.473	0.0544	7.17	0.61427
Received Feature	1023	10.05	0.1362	7.597	1.043

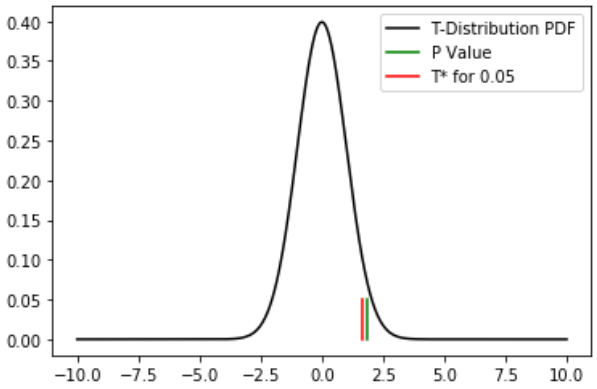
First I conducted two sample t-tests because the distributions of the observations were not normal. Below are the results:

Metric	T Score	P-Value
Number of Games	1.7457434053729304	0.0810084879525783
Revenue	2.1339706618884215	0.03296602717398876

By these scores, Average Number of Games being larger for users who received the feature is significant at a 90% level. Average Revenue being larger for the users who received the feature is significant at a 95% level. Both of these are visualized below.



P-Value for Number of Games Played Test



P-Value for Revenue Test

Due to these distributions not being normal, I also conducted Mann – Whitney U (MWU) hypothesis tests for number of games played. This test also compares the sample averages of both groups but does not require that they follow normal distributions. Below are the results:

Metric	Score	P-Value
Number of Games	481896.0	0.038799978433857615
Revenue	494268.0	0.003741232927946108

By the MWU hypothesis tests, the Average Number of Games for the users who received the feature is larger at a 95% significance level. The Average Revenue for users who received the feature is larger at a 99% significance level.

Difference-In-Differences

Using the original dataset, I conducted a Difference-In-Differences regression analysis. This analysis works by running a regression with Number of Games Played or Revenue as our dependent variable. Our independent variables for Number of Games Played will be Revenue, days before or after the user was selected, a dummy variable for if the user was in the treatment group, a dummy variable for if it is after the experiment started, and an interaction variable

between the previous two dummies (DiD). Our independent variables for Revenue will be the same, except Number of Games Played will now be one of the independent variables instead of Revenue. The coefficients for the interaction terms will show if there is a causal effect on adding the new feature on Number of Games Played or Revenue. Below are the regression results:

OLS Regression Results

Dep. Variable:	Revenue	**R-squared (uncentered):**	0.017
Model:	OLS	**Adj. R-squared (uncentered):**	0.016
Method:	Least Squares	**F-statistic:**	40.71
Date:	Sat, 25 Apr 2020	**Prob (F-statistic):**	1.14e-41
Time:	13:22:30	**Log-Likelihood:**	-20871.
No. Observations:	12045	**AIC:**	4.175e+04
Df Residuals:	12040	**BIC:**	4.179e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
feature_enabled	-0.1117	0.045	-2.478	0.013	-0.200	-0.023
After_Treatment	-0.0752	0.030	-2.478	0.013	-0.135	-0.016
Diff_Days	-0.0017	0.006	-0.289	0.773	-0.013	0.010
games_played	0.0132	0.001	10.871	0.000	0.011	0.016
DiD	0.1970	0.053	3.746	0.000	0.094	0.300

From the first regression it can be seen that the DiD variable has a positive effect on Number of Games Played and low standard error, however the adjusted R squared for the model is only 0.016.

OLS Regression Results

Dep. Variable:	Revenue	**R-squared (uncentered):**	0.543
Model:	OLS	**Adj. R-squared (uncentered):**	0.543
Method:	Least Squares	**F-statistic:**	2864.
Date:	Sat, 25 Apr 2020	**Prob (F-statistic):**	0.00
Time:	13:22:31	**Log-Likelihood:**	-45121.
No. Observations:	12045	**AIC:**	9.025e+04
Df Residuals:	12040	**BIC:**	9.029e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
feature_enabled	11.5644	0.321	36.058	0.000	10.936	12.193
After_Treatment	12.7000	0.196	64.855	0.000	12.316	13.084
Diff_Days	-0.5350	0.043	-12.308	0.000	-0.620	-0.450
revenue	0.7383	0.068	10.871	0.000	0.605	0.871
DiD	-11.0502	0.381	-29.009	0.000	-11.797	-10.304

From the second regression it can be seen that the DiD variable actually has a negative effect on Revenue. This also has a low standard error and while its adjusted R squared is better, it is still only 0.543.

Conclusion and Recommendation

The hypothesis tests all show that the group which received the new feature has a larger Average Number of Games Played and larger Average Revenue. While the results vary in significance 3 of the 4 are above the 95% significance level and the fourth is above the 90% significance level. While the second difference-in-differences regression shows that the feature has a negative effect on revenue, neither of the regressions have large enough adjusted R squared scores for them to be used to determine a causal effect.

Thus, I would determine that this feature is a success. All of the hypothesis tests showed that the Average Number of Games Played and Average Revenue were larger for the group that received the feature, with significant results. While the difference-in-differences analysis provided conflicting results, the results were not statistically significant.

Future Works

To further improve this analysis, I would like access to data about the users' demographics. Having data such as age, location, and other demographics would give a better understanding of what types of users play more games or make more purchases. I would also like data about what else users were doing in the game that led to them either playing another game or make a purchase. Having data such as how long a player plays a game for or how much they've progressed in a session would also give insight into what drives a player to play more games or make purchases. With this data, my goal would be to improve the difference-in-differences model, making it statistically significant by providing it with more data. If the regression results are significant and has a higher adjusted R squared, causal inference of the feature being added can be determined with higher statistical significance.