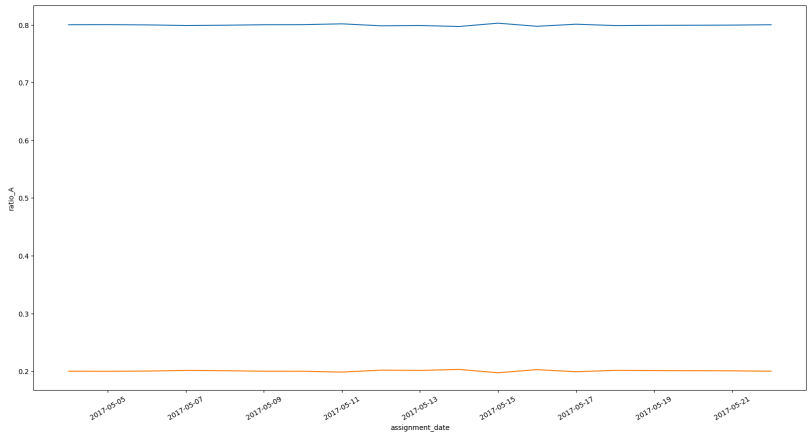
King A/B Test Use Case

Fateme Ahmadi

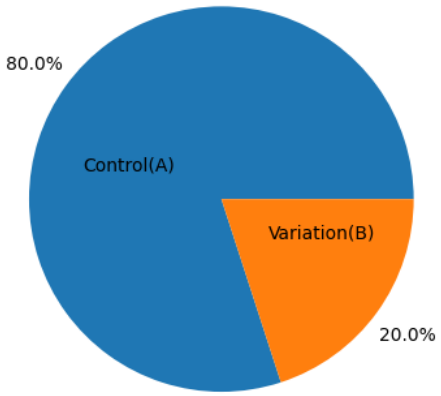
Experiment Analysis

Experiment Statistics

The experiment ran for two weeks, from 4th May 2017 to 22nd May 2017, referred to as the treatment window. Throughout this period, players were randomly assigned to either the control group (A) or the variation group (B). As depicted in the plot, the ratio of group assignment remained almost consistent at 0.2 for group A and 0.8 for group B every day.



In total, the experiment included a sample of 10,331,056 players, with 8,265,610 assigned to the control group (A) and 2,065,446 assigned to the variation group (B).



To study the effect of the treatment, we will also examine the group activity during the 2 weeks prior to the start of the treatment. This period will be referred to as the *pre-treatment window*.

Player Features

Each player in this game has a set of features that can be extracted from the available data:

1. **Age in App**: The number of days since the player's installation at the time of assignment to the experiment.
2. **Age of Convert**: The number of days since installation when the player first converts.

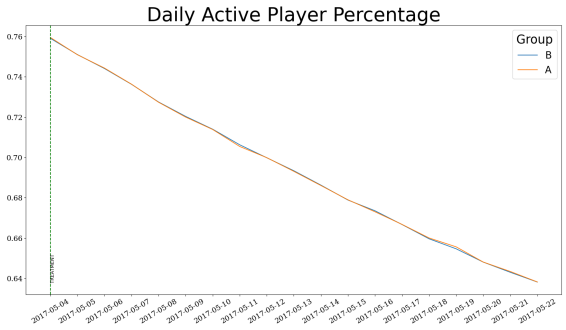
Based on their activity during the pre-treatment window, we can also extract several behavioral features:

1. **Active Days Ratio**: The ratio of days the player has been active since the start of the pre-treatment window.
2. **Purchase Value**: The average purchase value per activity.
3. **Conversion Days**: The number of days the player has made at least one purchase.
4. **Game Rounds Ratio**: The number of game rounds per active day.

Performance metrics

Active Daily Players

In order to better compare both groups and make the comparison independent of group sizes, let's look at the percentage of daily active players in each treatment group.



It is illustrated that both groups have almost the same number of active players for the treatment windows.

To delve deeper, we calculate the number of active days for each player during their participation in the experiment. Given that the test duration varies and users enter the experiment on different days, but all conclude by 22nd May 2017, we also compute the active days per duration, known as the active days ratio.

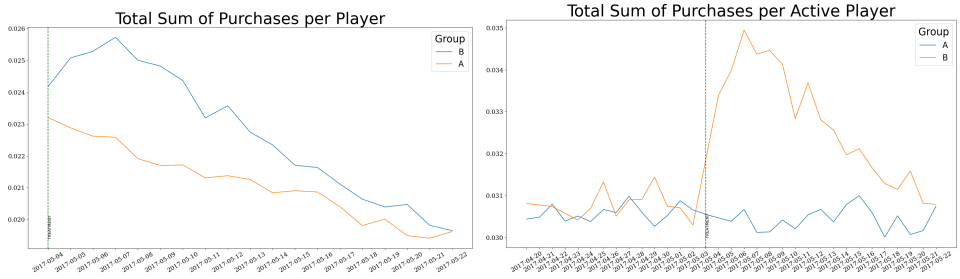
	Number of Active Days	**Active Days Ratio**
Group A	12.028104	0.686115
Group B	12.023382	0.686051
p-value	0.21	0.71

After running a t-test, we obtained a p-value of 0.21 for the Number of Active Days and a p-value of 0.71 for the Active Days Ratio. Therefore, we cannot reject the null hypothesis, indicating that the difference between the two groups is not statistically significant. **This variation has not affected the number of active days for users**.

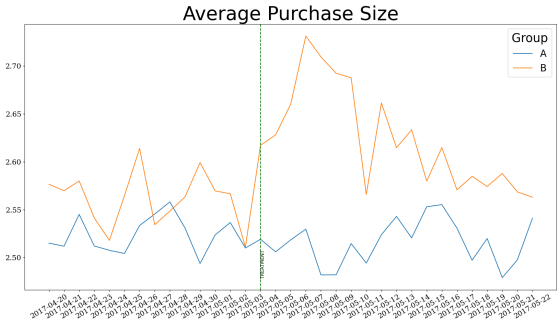
Revenue

If we multiply the total number of purchases by the value of each purchase, we get the revenue. Assuming the value of each purchase is fixed, the number of purchases can serve as a good proxy for total revenue. Since the treatment group sizes differ, we cannot directly compare the total number of purchases between groups. Therefore, we define three related metrics that are independent of group sizes:

The *total Sum of purchases per player*, which serves as a good proxy for the total revenue gained from each group.



The average purchase size, which gives us an idea of how profitable each group's players are.

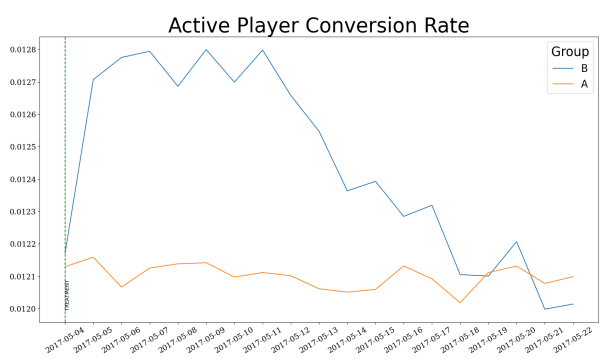


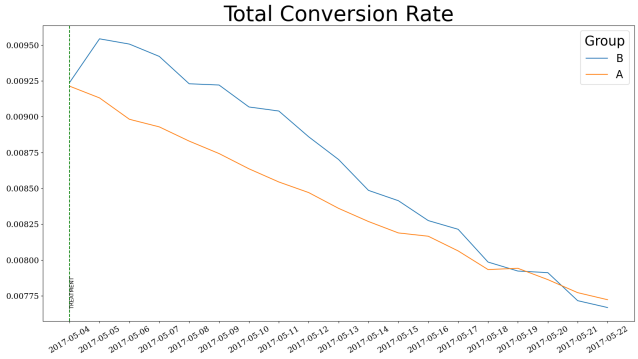
	Player Total Purchase	**Players Purchase per Active Days**
Group A	0.366211	0.030647
Group B	0.39257	0.032558
p-value	0.00046	0.0010570

After conducting a t-test and obtaining p-values all less than 0.001, we reject the null hypothesis at any significance level higher than 0.001. This confirms that **the observed difference in group purchase values is statistically significant.** However, it's important to note that the difference between the two groups may not be substantial enough to have a high impact.

Conversion Rate

If a player makes a purchase, they are considered “converted”. We calculate the conversion rate in two ways:
The *total conversion rate* is based on the total group size. And the *active player conversion rate* is based on daily active players.





	Player Converted Days	**Players Converted per Active Days**
Group A	0.145542	0.012087
Group B	0.149601	0.012457
p-value	1.76e-06	5.25e-08

The above results show that **the players' conversion metrics between the two groups are statistically significant.**
A *high conversion rate* indicates effective monetization strategies and a players’ willingness to engage in these revenue-generating activities.

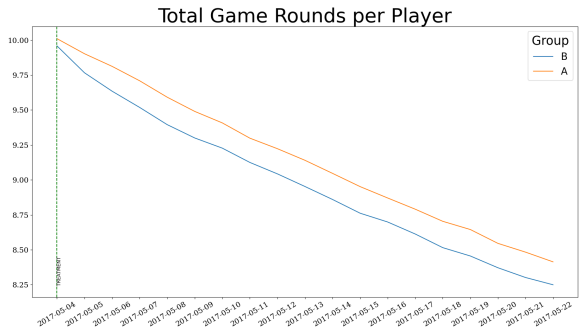
Both conversion rate and revenue metrics illustrate an increase compared to the control group. However, the plots reveal two potential concerns:

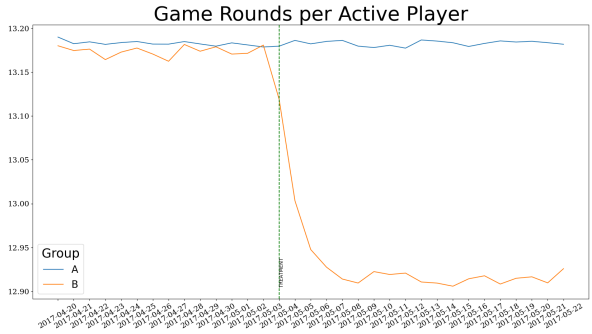
- First, the metrics in the variation group (B) decrease over time, and by the last days of the experiment, they either fall below the control group's rates (conversion rate metrics) or reach the control group's value (revenue metrics).
- Second, the pre-treatment performance of the groups differs, which may indicate a lack of randomness in treatment selection.

Both concerns require further investigation, which we will address later.

Game Rounds

To study our users' engagement with the app, examining the number of game rounds can serve as a good proxy for the number or duration of daily sessions. This metric indicates how consistently users engage with the app.



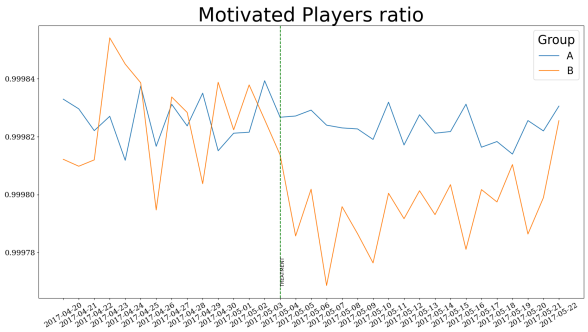


	Player Total Game Rounds	**Players Game Rounds per Active Days**
Group A	158.563439	13.182673
Group B	155.490497	12.948998
p-value	1.47e-181	1.84e-217

The above-calculated average game rounds between the two groups and the p-values resulting from the t-test demonstrate that **the number of game rounds, which indicates user engagement with the app, decreased in this variation.** The differences are statistically significant.

Here I suppose game rounds refer to playing a whole one round of a game respectful of moving to the next level or remaining on the current level. So it is a proxy for session time, or daily duration time. If it is not the case and game rounds are the same as levels, it can not be a proxy for user engagement.

Motivated players are users who have finished at least one game round. This means they have not just opened the app, but have also engaged with it by completing one or more rounds of the game.



Motivated days for each player are defined as the number of days they have played at least one round of the game. This metric can serve as a proxy for app engagement and customer retention.

	Player Total Motivated Days	Players Motivated Days Ratio
Group A	12.025982	0.999824
Group B	12.020927	0.999798
p-value	0.175	6.409e-09

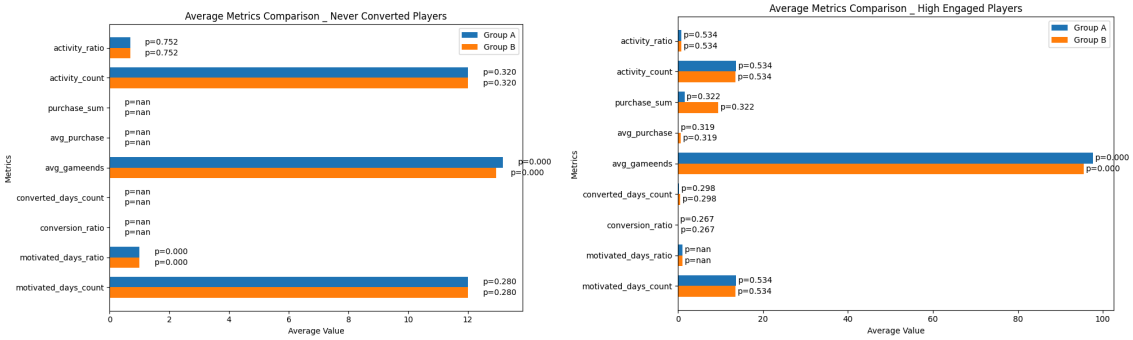
The above-calculated average motivated days between the two groups and the p-values resulting from the t-test demonstrate that, in terms of total motivated days, we cannot reject the null hypothesis, indicating that there are no significant differences. However, **in terms of the motivated days ratio, the difference is statistically significant.** Despite the small p-value indicating statistical significance, the actual difference between the two means is small, suggesting that the practical impact may be limited.

Segmentation Analysis

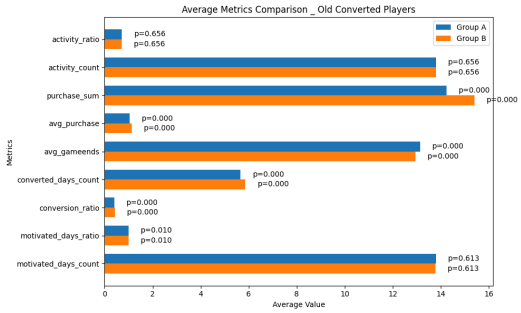
Segments Defined:

- **Old Converted Players:** Players whose app age was more than 10 days at the start of the treatment and who had already converted before the treatment began.
- **New Converted Players:** Players whose app age was less than 2 days at the start of the treatment and who converted after the treatment period.
- **New Players:** Players whose app age was less than 2 days at the start of the treatment.
- **None Converted Players:** Players who have never made a purchase.
- **Active Players:** Players who had at least 10 active days before the experiment assignment.
- **Profitable Players:** Players who had at least one purchase before the experiment assignment.
- **High Engaged Players**: Players who had more than average (100) game rounds before the start of the experiment.
- **After Treatment Converted Players**: Players who were converted during the treatment period.

Upon examining various metrics across these segments, it is evident that not all segments show statistically significant differences in aggregated metrics between groups.
For example, in the non-converted and highly engaged segments, nearly all revenue and conversion-related metrics show no statistically significant differences. This suggests that the variation does not prompt players who never make purchases or those without access to payment methods to make a purchase. The same applies to highly engaged players, as measured by the number of game rounds per day.



Conversely, in the old converted segment, we observe statistically significant differences in revenue and conversion-related metrics between groups.



Experiment Validation Test

Common players

There are no players common to both the control and variation groups.

Player Features

We can check for player features to see if there are any statistically significant differences between them.

Age in App

The age of the customer in the app is defined as the number of days between the player installation date and the assignment date. The calculated averages are as follows:

Group A: 142.943370
Group B: 142.912342
p-value: 0.78
With the calculated p-value, we cannot reject the null hypothesis of the t-test, indicating that there are no significant differences between these values.

Never Converted Players

We checked the ratio of never-converted players, whose converted date is null, between the two groups:

Group A: 0.972
Group B: 0.972
 P-value: 0.3014

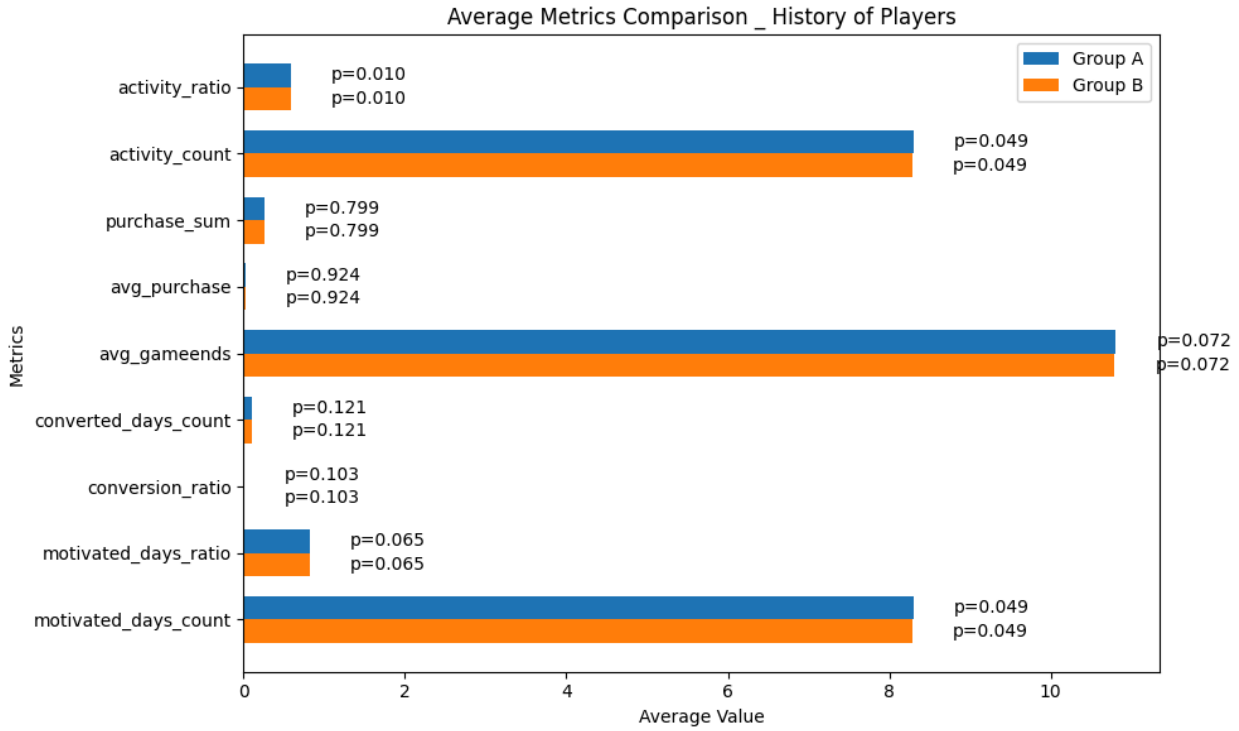
To determine if the difference in the ratio of non-converted players between the groups is statistically significant, we ran a chi-square test for independence. This test is used to see if there is a significant association between the group and the presence of non-converted players.
The p-value of 0.3014 indicates that there is no statistically significant difference in the ratio of non-converted players between the two groups.

Converted Players

We checked for players who converted before or after the assignment and examined if this ratio differs between the groups. According to the p-value of the chi-square test, there is no statistically significant difference between the groups regarding both pre-treatment conversion and post-treatment conversion.

Historical Behaviors

Analyzing customer pre-treatment activities, we observe that metrics such as activity ratio, motivated day count, and average game ends differ significantly between groups at significance levels of 0.01 or 0.07. However, metrics like conversion ratio and revenue-related metrics do not show statistically significant differences.



Conclusion

In this experiment analysis, we examine various metrics crucial to customer behavior and business KPIs. The variation has impacted several metrics: **it increases group revenue and conversion metrics**, **decreases player engagement with the app**, but **does not affect the number of activity days**. However, **the size of activity in terms of game rounds has decreased**.

To validate the test, we conducted sanity checks to ensure that players' behavioral or demographic features were not different at the time of assignment. We also verified for any errors in A/B test setup, such as having common players between groups.

This variation may be linked to a change to the difficulty of game levels, potentially reducing player motivation to continue playing and prompting them to make payments to progress. If we had data on the player's number of levels completed, which serves as a proxy for player skill, we could identify and analyze the segment of skilled players to assess how this variation affects them. However, highly engaged players, measured by the number of game rounds per day, can also serve as a proxy for skilled players. In this experiment, we observe that this variation does not influence their purchasing behavior.

In the segmentation analysis, we demonstrate that the old converted segment exhibits statistically significant differences in revenue and conversion-related metrics between groups. Although the decrease in engagement is statistically significant, the actual value difference is not substantial. This suggests that this segment aligns more closely with the new game version strategy.