Data Analyst Test Task - Game Analysis Report

Puneeth Nikin Krishnan

10/07/2020

Objective

The objective of this analysis is to identify the optimum time to display the shop. It is imperative to establish the relationship between the available parameters and ARPU (Average Revenue Per User) to be able to gauge the effect modifying the controllable factors will have on the overall gameplay experience and revenue generation.

Exploratory Data Analysis

I have performed exploratory data analysis to analyse and understand these relationships so as to deliver a strategy that will optimise ARPU.

About the DataSet

The zip file provided at google drive contains three csv files namely :
1. data_daily_activity.csv - This file contains daily activity of users.

userId	date	countryCode	platform	abTestGroup
j83udscs5b5bfmb	2020-05-03	FR	ios	test_group_a
6vmia2xkmuo7ubm	2020-05-10	BR	ios	test_group_a
kov82tsj6he7hvp	2020-05-12	BR	android	control_group
85el8huhiwealui	2020-05-11	GB	android	test_group_b

2. data_daily_matches.csv - This file contains number of matches played by each user for a given day. I added a column matches_till_date(number of matches played till date) by computing cumulative sum of matches.

userId	date	matches	matches_till_date
000m8owu19gmejx	2020-05-12	7	7
000m8owu19gmejx	2020-05-13	23	30
000m8owu19gmejx	2020-05-14	14	44
000m8owu19gmejx	2020-05-15	17	61

3. data_in_app_purchases.csv - This file contains details of purchases made by users.

userId	date	product	cost
t054hrczly04vrf	2020-05-04	promotiondeal1	0.99
nmkzr6rjop5igcz	2020-05-05	promotiondeal1	0.99
rpa7dpseh6bddw2	2020-05-10	cashinjection	0.99
sxlb2r42xb2sss0	2020-05-04	cashinjection	0.99

Wrangling

I have combined all three datasets/dataframes into one single dataframe user_data to simplify analysis.I have added two columns acquisition_date(date user was added) and days_since_acquisition(number of days passed since acquisiotion) to user_data.

userId	date	countryCode	platform	abTestGroup	matches	matches_till_date	product
j83udscs5b5bfmb	2020-05-03	FR	ios	test_group_a	2	7	NA
6vmia2xkmuo7ubm	2020-05-10	BR	ios	test_group_a	1	42	NA
kov82tsj6he7hvp	2020-05-12	BR	android	control_group	7	62	NA
85el8huhiwealui	2020-05-11	GB	android	test_group_b	NA	NA	NA

cost	acquisition_date	days_since_acquisition
NA	2020-05-01	2
NA	2020-05-02	8
NA	2020-05-03	9
NA	2020-05-03	8

Using user_data I have derived a new dataframe called N_Day_Analysis to perform date wise analysis and compute Retention Rate, Cumulative ARPU and Cumulative Revenue.

abTestGroup	acquisition_date	days_since_acquisition	n	revenue	conversion	retention_rate
control_group	2020-05-01	0	3178	656.81	169	1.0000000
control_group	2020-05-01	1	900	240.79	71	0.2831970
control_group	2020-05-01	2	639	395.86	64	0.2010699
control_group	2020-05-01	3	496	180.92	58	0.1560730

Cumulative_ARPU	Cumulative_conversion
0.2066740	0.0531781
0.2824418	0.0755192
0.4070044	0.0956576
0.4639333	0.1139081

Analyse Realationship between different parameters

Correlation Analysis (Number of users and Revenue)

While it is intuitive that more number of active users will lead to higher revenue it is critical to analyse the relationship.

```
##                       n   revenue conversion retention_rate
## n              1.0000000 0.7849676  0.8022070      0.9757916
## revenue        0.7849676 1.0000000  0.8838052      0.7651753
## conversion     0.8022070 0.8838052  1.0000000      0.7738279
## retention_rate 0.9757916 0.7651753  0.7738279      1.0000000
```

The correlation matrix shows that all the parameters are positively correlated. A higher retention rate will lead to higher conversion and consequently higher revenues.

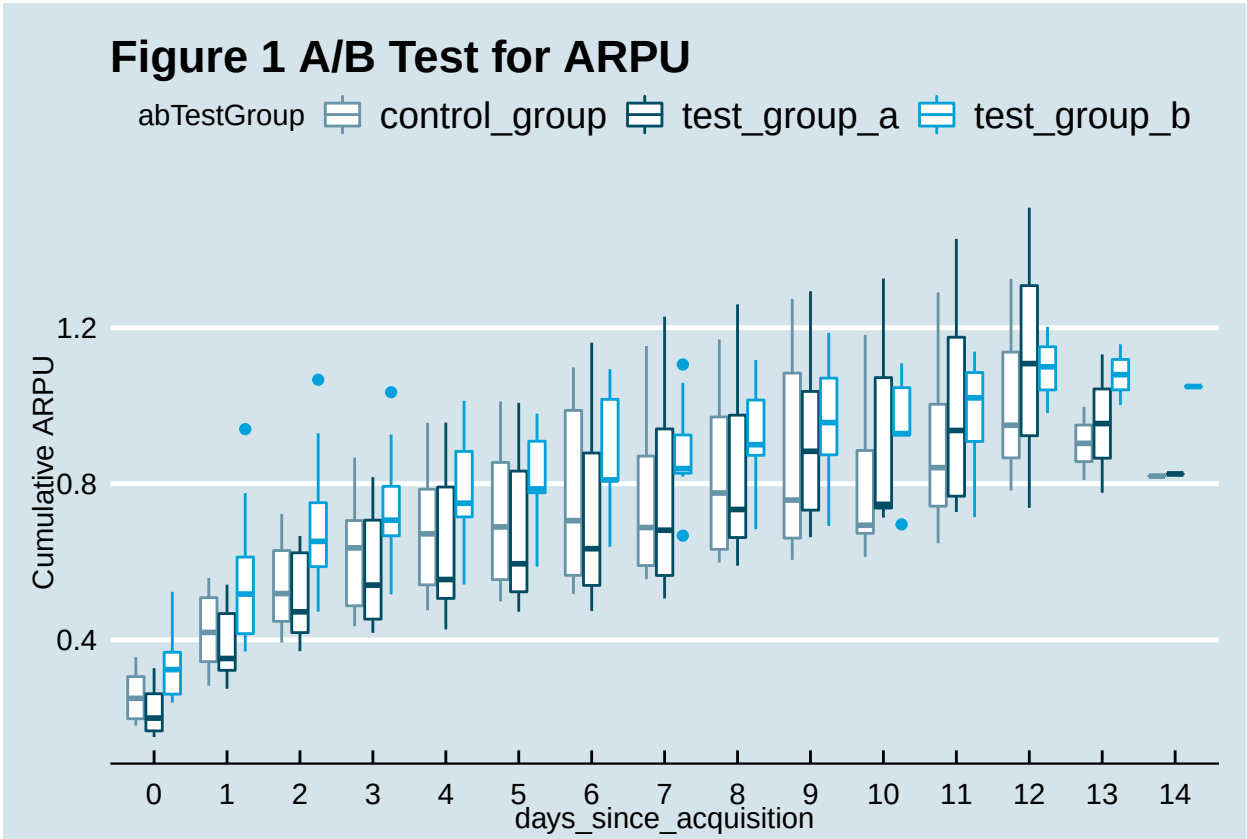


Figure 1 A/B Test for ARPU
abTestGroup
control_group
test_group_a
test_group_b
Cumulative ARPU
1.2
0.8
0.4
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
days_since_acquisition

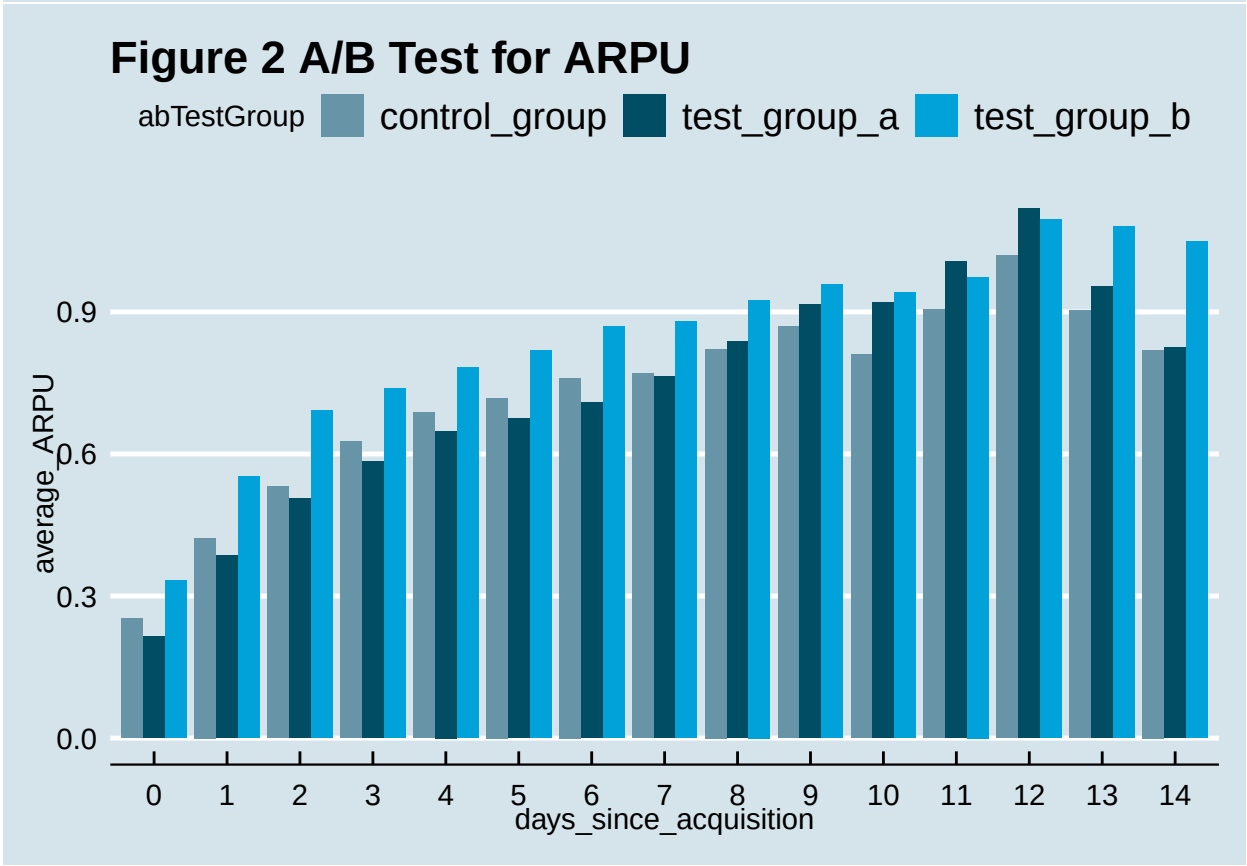


Figure 2 A/B Test for ARPU
abTestGroup
control_group
test_group_a
test_group_b
average_ARPU
0.9
0.6
0.3
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
days_since_acquisition

Figure 1 is a boxplot of Cumulative ARPU over different days since acquisition for each group. Figure 2 shows the average cumulative ARPU over diffent acquisition days. While ARPU increases everyday it can be seen that the test_group_a takes a strong lead in the initial days. This can be evidence of the fact that there is significant purchasing activity in the early stages of the game. Figure 2 also shows that rate of increase of cumulative ARPU for test_group_a and control_group is faster than test_group_b and hence these two groups tend to catch up with test_group_b over the period of time. This can be evidence of the fact that retention rate increases when display of shop is delayed.

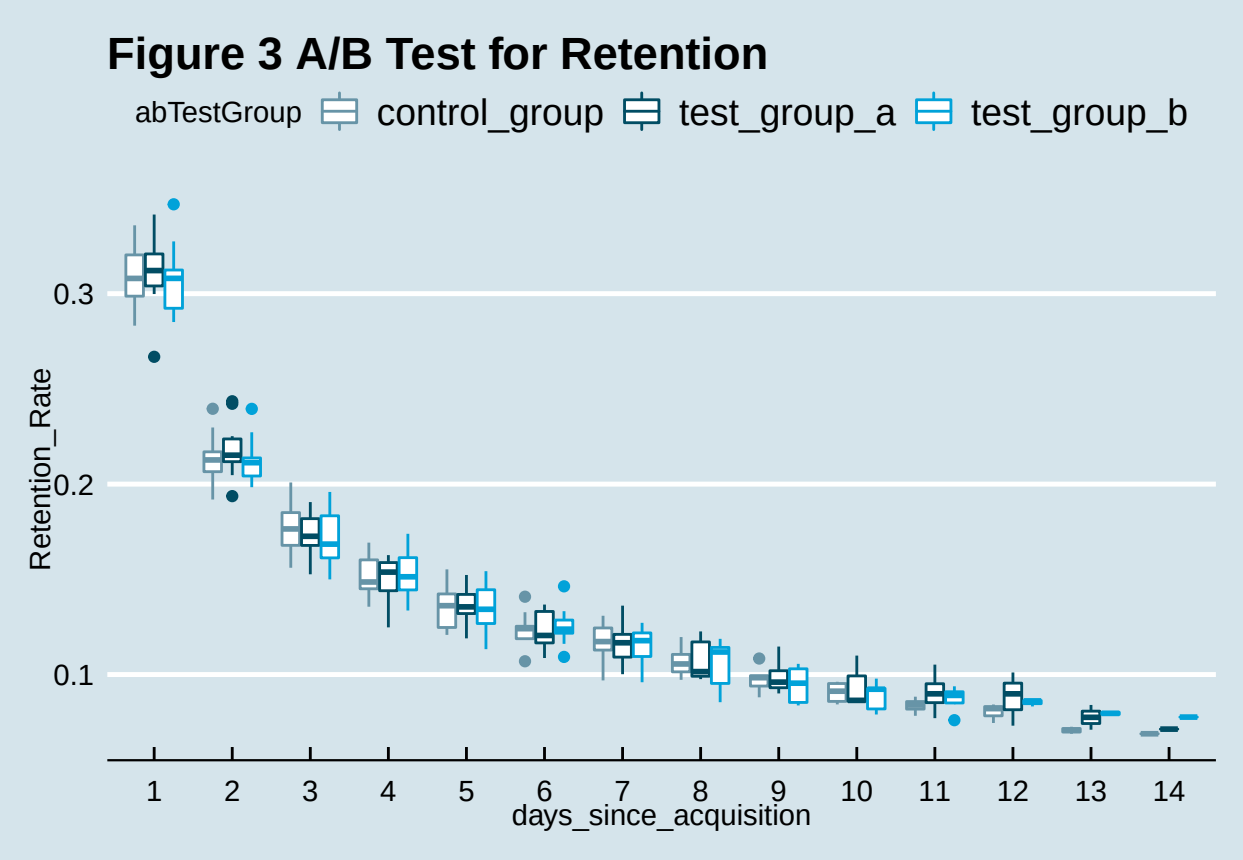


Figure 3 shows that retention rate drops rapidly during the period of analysis for all groups in general. This denotes that the period of analysis that is 15 days can be assumed to be the overall lifecycle for users with minimal revenue generation post that for a specific acquisition day.The graph also shows that retention rate is higher for control_group and test_group_a in the early stages. This could be evidence of higher retention rates for groups in which the display of shop was delayed.

Optimum Number of Matches Prior to display of Shop

It is clear from this analysis that higher retention rate can lead to higher conversion rates and consequently higher Cumulative ARPU. Also test_group_a performs the best in terms of cumulative ARPU over a period of time. While it can be seen that delaying the display of shop has a potential increase in retention_rates, their is also a loss in revenue when the shop is not displayed. A strategy needs to be devised that balances the need for revenue while maintaining retention rates.

To achieve this goal, analysis for number of matches prior to purchase of a product is performed.

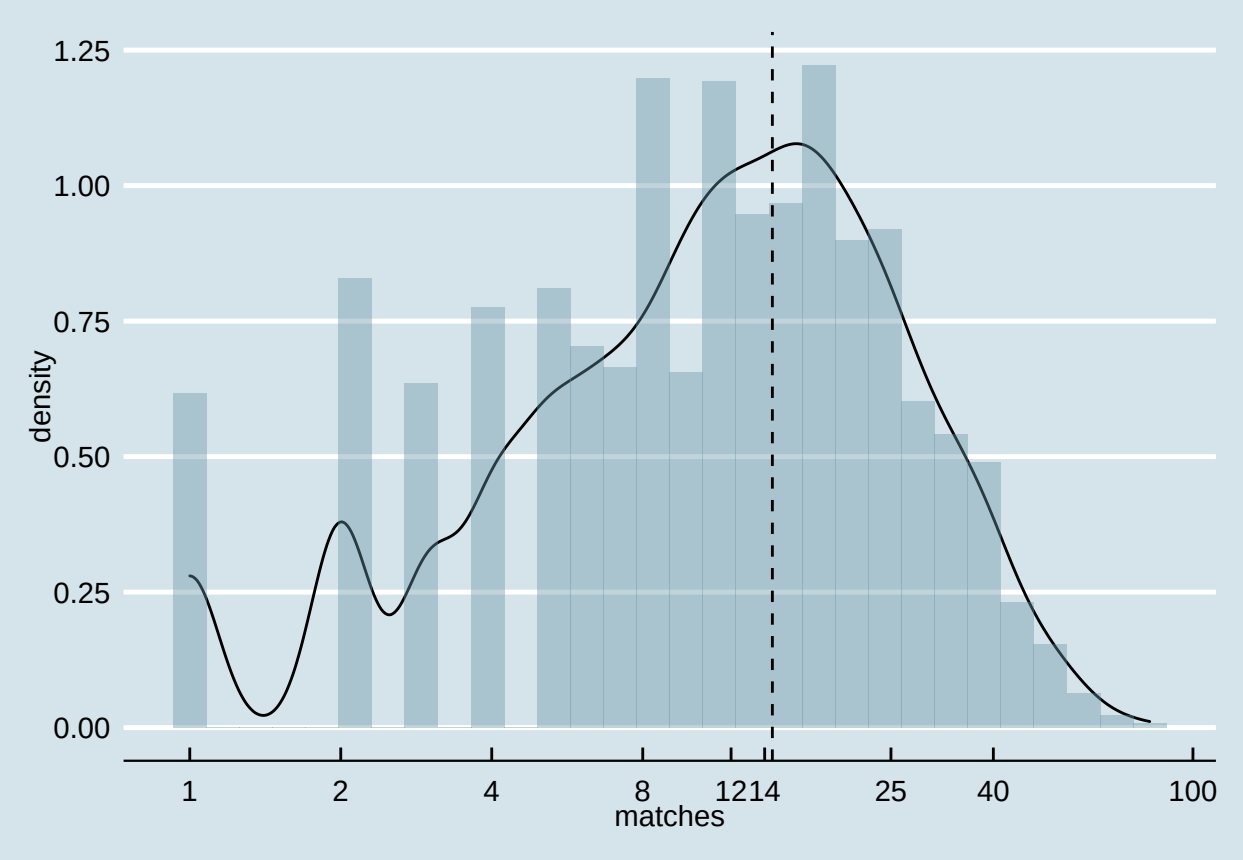
Assumption

The dataframe user_data contains total number of matches played for a given day and if a product has been purchased on that day. I have assumed that the product has been purchased at the end of the day. If a player

plays 10 games on a given day and purchases product "cashinjection" on that day, I have assumed that the product was purchased after playing those 10 games.

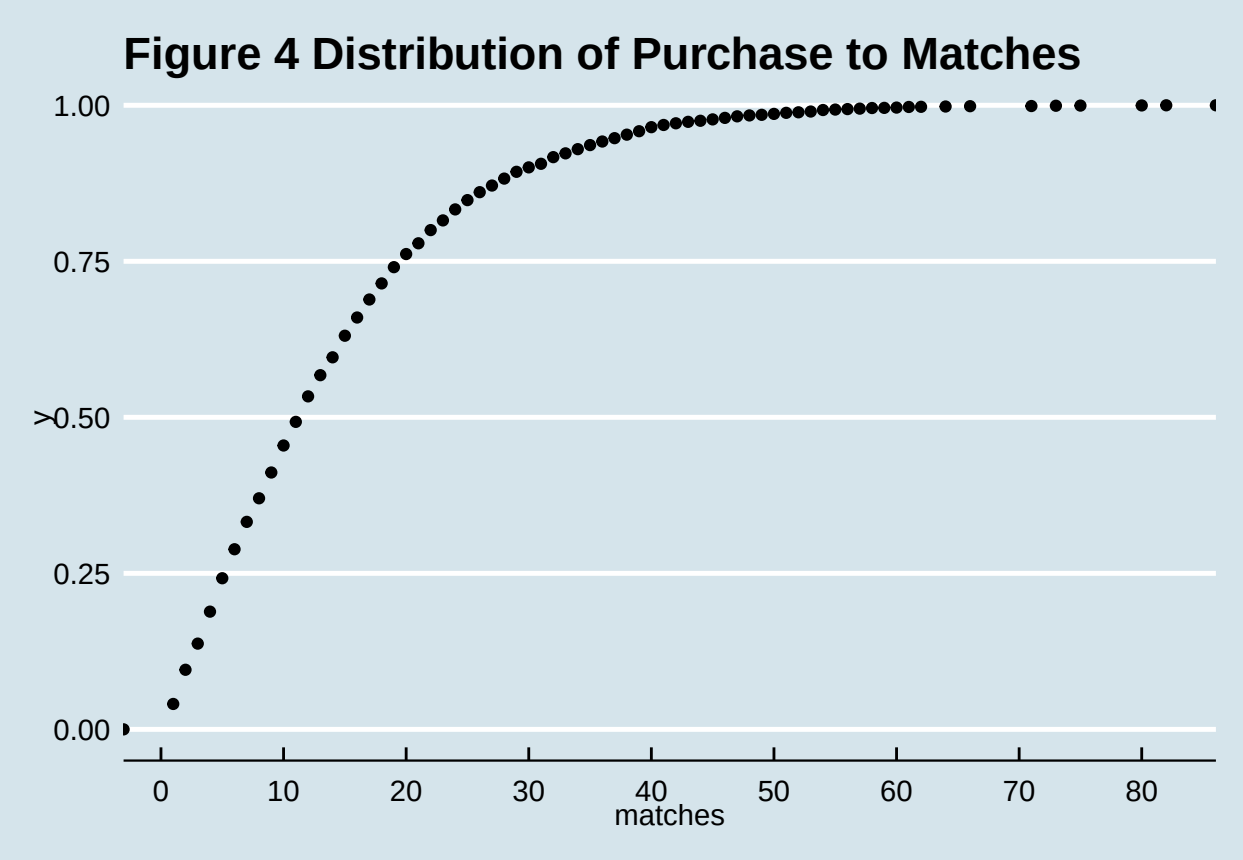
Distribution Analysis

Users tend to purchase products at different stages of their engagement with the game and it is necessary to study the distribution of matches prior for users first purchase of any product.



The distribution shows that that most users tend to purchase a product between the 5th game and 25th game. The average number of games played prior to first purchase is approx. 14. There is a spike in number of users purchasing a product after 1st, 2nd and 4th game. This could be potentially due to display of shop for test groups at these junctures. Purchase of a product peaks post the 8th game.

To study this further emperical Cumulative Distribution Function (eCDF) is generated.




```
d_fun<-ecdf(purchased$matches)
# probability of purchase before 3rd game
d_fun(3)*100
```

```
## [1] 13.73915
```

```
# probability of purchase before 5th game
d_fun(5)*100
```

```
## [1] 24.21464
```

```
# probability of purchase before 5th game
(d_fun(25)-d_fun(5))*100
```

```
## [1] 60.59161
```

This study shows that 13.7 % of users purchase a product before the 3rd game, 24.21% before the 5th, 60.59 % between 5th to 25th game and the rest later. This data confirms the fact that the purchasing is strongly skewed towards the early stages of the game.