Analysing Cookie Cats

Data Analytics Project

November 2022

Contents

Cookie Cats

Cookie Cats is a popular mobile puzzle game developed and published by Tactile Entertainment. It's a classic "connect three" style puzzle game where the player must connect tiles of the same color in order to clear the board and win the level. Its main characters are five cats named Rita (pink), Smokey (red), Berry (blue), Ziggy (green), and Belle (yellow).

Figure 1: Cookie cats game.

The game has a huge map of levels. As players progress through the game they encounter gates that force them to wait some time before they can progress or make an in-app purchase. There are also other in-app purchases available to get boosters and powerups to use in levels, refill lives, start levels with an extra booster, and as it has been mentioned, unlock new levels faster.

In this project, we will analyze the behaviour of 90,189 players that were tracked during a week. An A/B test was running during that week, in which the players got randomly version A or B of the game. In particular, the first gate in Cookie Cats was moved from level 30 (original game) into level 40. The designers are wondering whether this has a positive effect on player retention and game rounds. Additionally, the in-app purchases of the users was monitored during that week. Thus, the monetization designers are interested in understanding the spending behaviour of the players.

To perform these analyses, the developers have accessed the data bases and have prepared two datasets.

The first dataset, "cookie_cats_ABtest.csv", contains the following variables:

- userid - a unique number that identifies each player.
- version - whether the player was put in the control group (gate_30 - a gate at level 30) or the test group (gate_40 - a gate at level 40).
- sum_gamerounds - the number of game rounds played by the player during the first week after installation
- retention_1 - did the player come back and play 1 day after installing?
- retention_7 - did the player come back and play 7 days after installing?

When a player installed the game, they were randomly assigned to either gate_30 or gate_40.

The second dataset "cookie_cats_purch.csv" contains the events of in-app purchases of the users. The variables are:

- id - the ID of the player

- purch - the amount of a given in-app purchase in euros

The ID of the user is a unique identifier of every user that was monitored during this week. Thus, given an ID number, it refers to the same player in both data sets. This allows for matching the information of the two data sets, if this is necessary for the analysis. Please, take into consideration that all users in the second data set should appear in the first data set. However, the reverse situation is not necessary, because not all users have made in-app purchases.

The designers of the game need your help in analysing these datasets and answer several research questions:

1. How many users downloaded the game?
2. Are the two groups of users well balanced in terms of number of players that got version A and version B of the game?
3. What is the distribution of game rounds in the players' population? And among the users of version A and version B of the game separately?
4. What is the value of retention at day 1? (percentage of users that are still active the day after installation).
5. What is the value of retention at day 7? (percentage of users that are still active after one week of installation).
6. Are there users that never played the game? How many? In what percentage?
7. What are the values of the most prominent monetisation metrics?
8. Does moving gate to level 40 improve engagement of the users significantly?
9. Can the amount of in-app purchases be related to the number of game rounds of the players?

Let's start!

Ground rules

Team organisation This project needs to be implemented in teams of 4 or 5 people. Every member of the team needs to contribute actively in the project. Dividing the project in several parts can be done but, all the members of the team need to be aware of how the remaining parts have been implemented by the other colleagues. Since it is not easy to work collaboratively in RStudio, I would suggest that you organise your work very well: decide who does every part and how you would combine the different parts. If there are dependencies, decide how you are going to solve them. You can draw a roadmap to organise your work. In the last section of the document, you need to specify how you split your work and what every member was responsible for.

Document to deliver You need to deliver the following documents:

1. Source file (Rmd file)
2. The output of the analysis (html or PDF) with the code, and the result of executing that code.
3. The output of the analysis (html or PDF), but only showing the result of the analysis. This is the same document as before, but disabling the code.
4. The dashboard.

Please, include the names of all the members of the team in all the documents.

Format and contents of the documents You need to format the contents of the document appropriately. Follow the same sections and subsections as the project statement. Keep the same structure and numbering as the problem statement. Add a table of contents, as well. Do not print the full dataset(s) in the document. This would generate hundreds of pages and would make the document unreadable. You can use `head` or `tail` or other functions to show portions of a data set.

Non-cheating manifesto Cheating (sharing partial or complete information) among groups is not allowed. This will conduct to a grade of 0 in the project grade, regardless of the size of the portion copied, and regarless of who copied and who allowed the other ones to copy. Thus, please, do not enter into cheating behaviours.

My suggestion is that you write a non-cheating manifesto in the last section of your document (section 10).

1 Getting to know the datasets

Before we start with the analysis, have a look at the structure of both datasets. First, upload the data sets in R. For a smoother correction on my side, please provide the data frames with the following names:

- For dataset 'cookie_cats_ABtest.csv' use data frame variable named 'DS'.
- For dataset 'cookie_cats_purch.csv", use data frame variable named 'PR'.

For each dataset, show and describe the information that they contain: rows and columns of the dataset, and types of variables.

```
## 'data.frame':    90189 obs. of  5 variables:
##  $ userid        : int  116 337 377 483 488 540 1066 1444 1574 1587 ...
##  $ version       : chr  "gate_30" "gate_30" "gate_40" "gate_40" ...
##  $ sum_gamerounds: int  3 38 165 1 179 187 0 2 108 153 ...
##  $ retention_1   : chr  "FALSE" "TRUE" "TRUE" "FALSE" ...
##  $ retention_7   : logi  FALSE FALSE FALSE FALSE TRUE TRUE ...
```

```
## 'data.frame':    4563 obs. of  2 variables:
##  $ id   : int  5696479 6682624 1698475 4520146 5081239 1177780 4334785 331668 45698 6170568 ...
##  $ purch: chr  "10.99EUR" "2.29EUR" "2.29EUR" "10.99EUR" ...
```

2 Preprocessing

Before we start with the analysis, revise the structure and contents of the data sets to look for inconsistencies, missing values, or errors, and if necessary, perform the necessary transformations. You should also check that all IDs of the second data set (cookie_cats_purch) are valid IDs (i.e., existing IDs in cookie_cats_ABtest). If you apply any transformation, you should include a proper explanation in your report.

2.1 Preprocessing dataset "Cookie Cats AB Testing"

2.2 Preprocessing dataset "Cookie Cats Purchases"

3 Descriptive analytics

Perform descriptive analytics of the datasets visually and numerically. To make this task simpler, we'll do the analysis for each data set separately.

3.1 Dataset "Cookie Cats AB Testing"

Analyse the distribution of every variable separately by using the appropriate plots and some numerical variables of central tendency and variability. After that, show some relationship between the variables, such as game rounds versus the version of the game.

This is an exploratory analysis. Thus, you should try several plots and tables and show the ones that contribute with some insights about the behaviour of the players.

At least, you should follow the next subsections (you can add more if necessary). Please, check questions 1-6 that are listed in the introduction.

3.1.1 Users that downloaded the game

3.1.2 Distribution of users in groups

3.1.3 Game rounds

3.1.4 Retention (day 1 and day 7)

3.1.5 Are there non-playing users?

By looking at the first dataset, are there any users that downloaded the game but never played it? Show the results graphically and/or numerically.

3.2 Dataset "Cookie Cats Purchases"

Perform a descriptive analytis with the dataset `Cookie Cats Purchases` by using the appropriate plots and summary statistics. You can also create new subsections if necessary.

4 Monetization metrics

In this section, we aim at answering Researh Question 7: "What are the values of the most prominent monetisation metrics?" To do that, compute conversion rate, ARPU and ARRPU. Then, build a table where you print the value of these three metrics. Use the data set "Cookie Cats Purchases" to compute these metrics.

Note: to build tables, you can use function `kable` or you can also use html code. Draw plots if you find them useful.

5 A/B testing

In this section, we aim at answering Researh Question 8: "Does moving gate to level 40 improve engagement of the users significantly?"

The designers of the game are interested in knowing whether moving the gate to level 40 has improved the engagement of the users. Thus, we should perform an A/B test to answer the research question, considering engagement as the number of round games of every user. Thus, we would like to know whether moving the gate to level 40 increases engagement with 95% confidence level.

5.1 Hypotheses

Write the null and the alternative hypotheses.

5.2 Visualisation

Draw some plots to investigate whether there seem to be differences on game rounds between the different groups.

5.3 Computation

Apply the A/B testing hypothesis method, and compute and show all the necessary values: the observed statistic, the p value, etc. You need to develop the full computations (you can use `mean` and `sd`, but you can't use functions that already compute the full hypothesis testing).

5.4 Conclusion of A/B testing

Based on the numbers computed before, conclude whether you can reject the null hypothesis or not. And then, answer the research question. Justify your answer.

6 Regression analysis

In this section, we aim at answering Researh Question 9: "Can the amount of in-app purchases be related to the number of game rounds of the players?"

Thus, we would like to investigate whether the amount of in-app purchases depends somehow on the number of game rounds. Use regression analysis to answer this question.

Note: you can build the model with the full population of users, and also with only the paying users.

6.1 Visual representation

Represent visually the potential relationship between the total amount spent by the user and the number of game rounds.

6.2 Building the model

Develop the model with linear regression analysis and visualise the model. Explain how the model relates the total amount of purchases with the game rounds.

6.3 Model fit

Evaluate the quality of the model. Is it an accurate model?

6.4 Predict a case

Predict how much a user would spend if he/she has been playing 50 game rounds.

6.5 Simulation

Run a simulation of how much a player would spend for different game rounds, ranging from 0 until 1000, every 50 rounds. Build a table and draw a plot.

7 Dashboard

Use a visualisation tool (PowerBI/Tableau) to summarise visually the main insights of this analysis. You can generate a PDF page with the dashboard and upload it together with the other documents. Additionally, share the original source file with the dashboard.

8 Conclusion

Write a summary of your analysis, as if you were answering the questions to the designers. Be concise, clear, and use the right terminology.

9 Team contribution

Please, indicate how each member of the team has contributed to the project. Detail how you divided the tasks, and what task every member did. Collective tasks, such as working on the conclusions, sharing the results, etc., needs to be reported as well.

10 Non-cheating manifesto

Write a paragraph where you commit yourselves to a non-cheating manifesto. This is also a way of committing to your other team mates.
