To: Professor Amini

From: Wanyi (Evelyn) Dong, Matt Andersen, Jenny Shang

Subject: Sudoku Report

Date: November 13, 2019

## **Introduction & Executive Summary**

This report serves as a summary of our analysis of data provided by Web Sudoku[[1]](#footnote-1) - a free-to-play online sudoku website. Our objective is to analyze the data to address the following key questions in order to optimize business performance:

1. Who are the most valuable sudoku players we want to solicit and retain?
2. How can we attract and engage more of these high-value players?

After a series of analyses, we propose the following recommendations which we believe best answer the questions above, and are most useful in propelling the business forward:

1. Focus on repeat engagement of existing users by introducing email campaigns and new features to motivate players to attempt additional puzzles across multiple levels (A-level recommendation).
2. Target international users, and users in the 35 to 55-year-old age range (A-level recommendation).
3. Increase incentivization for new users to register accounts, by introducing features that are only accessible to registered users and by clearly explaining the value of creating an account on Web Sudoku’s home page (B-level recommendation).

In the following sections, we outline the approaches we took in conducting our analyses, elaborate on the insights that we derived, and further explain each of our recommendations.

## **Background & Assumptions**

Web Sudoku operates an online, free-to-play sudoku game with four levels of difficulty. The website tracks player data on puzzles completed, and upon solving each puzzle the player is advised on his or her ranking relative to all players who played the same level of difficulty. The website is available to play without an account, but offers additional functionality for those who register, such as personal statistics tracking and the ability to challenge friends. There is also a premium version available for a one-time payment of $14.95 that offers additional features.

This website is a typical example of an ad-supported free-to-play game because the player is served Google banner ads rather than being asked to spend any money directly on the game, other than the option to purchase the deluxe version. Due to the data that was made available to our group, and to limit the scope of our analyses, we decided to focus solely on optimizing ad revenue rather than on increasing revenue from purchases of the deluxe edition. Our group has assumed that ad revenue is facilitated through Google, and it is primarily based on page visits.

Based on these assumptions, we have chosen to define the “value” of a sudoku player based on the number of puzzles that the player completes.  This is because Web Sudoku displays an ad on each individual puzzle, which means that the more puzzles a single user attempts, the more ads they will be served, and therefore the user will bring in more revenue than a user who only attempts one puzzle.

## **Exploratory Data Analysis**

Prior to diving deep into the data, we first conducted surface level exploratory data analysis to identify any notable trends. The dataset we received consists of two files indexed by User IDs. For our analysis, we assume that each User ID is associated with a unique player. Since providing demographic data is optional when a user registers for an account, the dataset contains many missing demographic values, such as Birth Year, Gender, and Country. Approximately one-third of all players did not populate at least one field and approximately 27% did not populate any of the three fields. To ensure our analysis is more reliable, we cleaned the data and created a separate dataset where we removed entries that contained missing values.

*Figure 1 - Missing values*

|  |  |  |  |
| --- | --- | --- | --- |
| Field Missing Value | # of Missing Values | Total # of Players | % of Missing Values |
| Birth Year | 90,231 | 264,690 | 34.09% |
| Gender | 93,569 | 264,690 | 35.35% |
| Country | 91,351 | 264,690 | 34.51% |
| All 3 | 70,360 | 264,690 | 26.58% |

When comparing the original data to our cleaned data, there was a noticeable impact to the average number of puzzles completed per user. In the original data, one-third of all users only ever completed one puzzle, and over 80% of players completed five or fewer. However, this proportion drops to approximately 46% within the cleaned dataset. We also noticed that the percentage of players solving five or fewer puzzles decreases as the level of difficulty increases (see Figure 5 in Appendix), as does the percentage of players who were not able to complete any puzzles at all (see Figure 6 in Appendix). Therefore, we can conclude that players who attempt more difficult levels are more likely to be highly engaged and are more likely to attempt multiple puzzles. These players are highly valuable and should be a priority to retain.

*Figure 2 - Puzzle completion rate*

A close up of a map

Description automatically generated

We next looked at the players’ demographic distribution. When we grouped players based on gender, we found that the population skews slightly male, with the proportion of female players decreasing as the level of difficulty increases (see Figure 7 in Appendix). Furthermore, we noticed that the majority of users are in English-speaking countries such as the U.S. (see Figure 8 in Appendix). In terms of player age, the range with the most users is currently between 25 and 30 years old, but when we narrow the data down to the top five countries with the most users, we noticed an increase in the average age (see Figure 9 in Appendix). This implies that there is a potentially untapped market of players in non-English speaking countries, and that they are younger on average than current Sudoku players in the U.S. and Canada.

*Figure 3 -  Player demographics*

A map of a person

Description automatically generated A screenshot of a cell phone

Description automatically generated

## **Regression Analysis**

To derive additional insights from the dataset, we decided to conduct a regression analysis with the goal of identifying variables that are useful in predicting the number of puzzles a user will solve. This is important because as mentioned previously, players who complete more puzzles are more valuable to the website. The regression results are summarized in Figure 4*.* Although the demographic variables cannot very effectively explain the variability in the number of puzzles solved, they are still statistically significant in making predictions. Overall, we find the number of puzzles solved for a player to be positively related with age, and negatively related to the average time it takes to complete a puzzle. In fact, it is interesting to note that this relationship strengthens as the level of difficulty increases. This is reflected in the regression results for each level of difficulty (see Figure 10 in Appendix).

*Figure 4 - Regression result*

A screenshot of a cell phone

Description automatically generated

## **Recommendations**

After analyzing the data and conducting the regression analysis, we generated useful insights that guided us to the following recommendations for Web Sudoku:

***Recommendation 1 (A-level):***

Since Web Sudoku receives ad revenue per puzzle, the business should incentivize players to complete as many puzzles as possible and seek to retain these active, high-value users.  The cost of retaining an existing player is much less than the cost of soliciting a new player, and therefore Web Sudoku’s resources are more effectively spent on driving repeat engagement.  It is very important for Web Sudoku to quickly identify potential high-value players upon registration and maintain connections to incentivize them to regularly return to the site. As mentioned earlier in the report, one third of users only complete one puzzle and 80% completed five or fewer, which means that there is ample room for growth in this area.

One way that Web Sudoku can drive repeat engagement is to introduce email campaigns targeting users who have registered and completed one or two puzzles but have not revisited the site since then. These emails could serve as a gentle reminder to try a new puzzle, or as a challenge to attempt a more difficult puzzle. Web Sudoku can also introduce new features to motivate players to attempt additional puzzles, such as the ability to “level up” your account by completing a certain number of puzzles, or a global leaderboard ranking players by the number of puzzles completed, average time, or fastest time. These would provide competitive motivation to try additional puzzles in order to move up the ranks or achieve a higher account level than your friends. It could also be beneficial to motivate players to attempt puzzles across multiple difficulty levels, since players who reach higher levels are more likely to complete multiple puzzles.

***Recommendation 2 (A-level):***

Our second recommendation is to prioritize user acquisition efforts by targeting international users, as well as users in the 35 to 55-year-old age range.  Based on our demographic analysis, the vast majority of existing players are in the U.S..  Additionally, based on Figure 3, Web Sudoku has an established presence in the 20 to 30 age range and the 55 to 75 age range. Although the age range with the most users is between 20 and 30, Figure 2 shows that users above 30 complete more puzzles per person on average, which makes them a more valuable age group. Our cleaned data summary statistics based on countries suggest that the average age of users from English-speaking countries is greater than the average age of all users. This could indicate that there is a large population of international players outside these five countries. Therefore, we propose adding additional languages to the website to attract international users.

***Recommendation 3 (B-level):***

Our final recommendation is to increase the incentivization for users to register accounts rather than play the game without one, so that Web Sudoku is able to compile more player data.  As shown in our data analysis, about one third of all players did not provide the complete set of information. However, users who did provide information tended to be more engaged, with over 50% of them completing more than five puzzles.  In addition, players who provide a birth date complete twice as many puzzles per person on average.

Demographic data is very useful when creating a model to predict player value. Although the variables we used in the regression model were statistically significant, the R-squared value was relatively low, which means that it does not explain a large variance in the data.  Collecting additional demographic data and filling in the many null values in the existing dataset could allow us to create a more accurate model, which would allow Web Sudoku to prioritize high-value users more effectively, save costs, and increase ad revenue in the long term.

To convince new users to register accounts, we recommend that Web Sudoku more clearly explain the value of registration on their home page. Creating an account currently unlocks a lot of functionality for new users, such as the ability to track personal statistics and challenge a friend to beat your puzzle times, however these features are not communicated until after the account creation process. By adding clear and visible text advertising all of the advantages of registration, we believe that the website will be able to drive new accounts and collect a large amount of new demographic data. Incentivization can be increased even further by making any of the new features mentioned in Recommendation 1 exclusive to registered players as well.

## **Conclusion**

After analyzing the data from Web Sudoku, we generated three recommendations to help improve business performance through identification of high-value players. We believe that Web Sudoku should implement efforts to incentivize players in completing more puzzles, increase exposure to international users and adults in the 35 to 55-year-old age range, and introduce features to segment registered users.

# Appendix

*Figure 5 - percentage of players who solved 5 or fewer puzzles, broken out by difficulty level*

|  |  |
| --- | --- |
| Level | % of Players who solve <=5 Puzzles |
| Level 1 (Easy) | 58.86% |
| Level 2 (Medium) | 53.07% |
| Level 3 (Hard) | 47.84% |
| Level 4 (Evil) | 48.43% |
| **Overall** | **46.17%** |

*Figure 6 - percentage of players who did not solve any puzzles, broken out by difficulty level*

|  |  |
| --- | --- |
| Level | % of Players who did not complete any puzzle |
| Level 1 (Easy) | 0.32% |
| Level 2 (Medium) | 0.21% |
| Level 3 (Hard) | 0.18% |
| Level 4 (Evil) | 0.16% |
| **Overall** | **0.19%** |

*Figure 7 - Proportion of Male users broken out by difficulty level*

|  |  |
| --- | --- |
| Level | Proportion of Male Players |
| Level 1 (Easy) | 62.87% |
| Level 2 (Medium) | 64.99% |
| Level 3(Hard) | 69.40% |
| Level 4 (Evil) | 74.15% |

*Figure 8 – Summary statistic of players based on geographic location (Overall and broken out by difficulty level)*

*Overall:*

A screenshot of a cell phone

Description automatically generated

*Level 1:*

A screenshot of a cell phone

Description automatically generated

*Level 2:*

A screenshot of a cell phone

Description automatically generated

*Level 3:*

A screenshot of a cell phone

Description automatically generated

*Level 4:*

A screenshot of a cell phone

Description automatically generated

*Figure 9 - Average age of top 5 countries with most players broken out by difficulty level*

A screenshot of a cell phone

Description automatically generated

*Figure 10 – Regression results for each difficulty level*

Level1 Level 2

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

Level 3 Level 4

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

1. www.websudoku.com [↑](#footnote-ref-1)