Web Sudoku Analysis Report

DSO 599 Game Analytics Project 1

Alex Furrow, Chengqin Kui, Juno Wen, Yangzi Zhang

November 13, 2019

USC Marshall

School of Business

1. Executive summary

Web Sudoku is a free-to-play webpage puzzle game whose revenues primarily come from display advertising. Player performance statistics are tracked and summarized at different difficulty levels. Optional demographic information is collected if provided by signed-up users. Current players can be segmented into five groups, each demonstrating distinctive playing behaviors and differs in demographic composition. Based on these patterns, specific features or feature changes to improve user retention are suggested, in addition to recommendations on target marketing.

2. Recommendation

The players of *Web Sudoku* can be separated into 5 categories based on the way they play the game (see Table 1). The first four categories each focus on a specific level of puzzles with the first category having the largest population, the highest churn rate, and the least time engagement. The fifth category houses the most hardcore fans who spend an outstanding amount of time on all levels of puzzles.

Table 1: Player Category Summary Statistics

	cluster	solved	quit_aft5%	tot_time	avg_time	accuracy%	easy%	medium%	hard%	evil%	age	male%	size	pct%
1	Dip the Toe	50	54.3	6:45:56	14:27	66.9	93.8	4.0	1.2	1.0	43	53.7	69454	56.9
2	Play and Chill	117	38.7	21:31:39	18:29	68.8	13.4	79.9	5.4	1.3	51	62.3	22362	18.3
3	Challenge Me	167	34.1	35:11:49	20:48	70.1	5.2	6.0	83.4	5.4	52	72.6	16279	13.3
4	More Evil Than Devil	212	30.6	51:51:48	22:53	66.1	3.8	1.7	5.0	89.5	51	76.6	13518	11.1
5	King of Puzzles	8173	0.0	1233:50:14	11:09	81.9	17.5	18.8	24.5	39.1	64	76.9	555	0.5

The key takeaways regarding player behaviors and demographics are:

- Most players tend to focus on one certain level of puzzles
- Players who love harder games generally have lower churn rate, play more rounds and spend more time in total
- Players at early 50's or older are more loyal to the game
- Harder games attract more male players than female players

These insights, along with industry and competing goods research, lead to the following five recommendations on user retention and marketing. Considering its feasibility and potential return, each recommendation is rated as A, B or C, with A being the best.

2.1. User Retention

The first four recommendations are aimed at reducing churn rate and maximizing user lifetime by easing the learning curve, improving user experience and introducing social mechanisms. All suggested feature additions or changes should be A/B tested to ensure impacts are as intended.

2.1.1. Visual Aids (A)

The current design of the puzzle table is very plain and has little to none visual aids to help players' logical thinking. Especially since the major users of the game are elder people who might be slower thinkers and have a hard time reading on a screen, the following user interface designs are recommended to enhance game accessibility.

- The game web page can be made interactive so that when a user clicks on a number, all same numbers will be highlighted (see Appendix A). It allows players to quickly spot visual clues in columns and rows to help them deduct the next number to fill.
- The current color of user filled numbers is light purple, which is not very distinct from black. This color can be changed to a more noticeable color or the website can allow users to choose their preferred color.
- In addition to the "How am I doing?" button that only checks if there is a mistake, a "Need a Hint?" button can be introduced to provide clues for a potential next step.
- Larger fonts should be used on all pages.

2.1.2. Tutorials for Beginners (A)

The game currently doesn't have any design to differentiate expert gamers and new gamers who never played sudoku. A simple survey including questions like "Have you ever played sudoku before?", or "How good are you in sudoku? (from 1 to 5, 5 means expert, 1 means beginners)" is highly recommended. After the survey, the system can redirect beginners to a tutorial page where basic mechanisms are explained and useful techniques are introduced. A link to this tutorial page should always be visible on the front page in case users need a refreshment. For people who have fair experience with sudoku, the system can suggest Medium or Hard level games for them to start with so that the users won't feel bored because the game is not challenging enough.

2.1.3. Level Suggestions (B)

Given that users who play harder puzzles usually spend more time on the game in general and most users who decide to stay after the first few games tend to develop a habit of playing puzzles at the same level over and over again, migrating users from easier levels to harder levels at early stage can help maximize user lifetime and time engagement. For example, if the system detects that a user solved five Easy level puzzles in a row with accuracy rate and speed much higher than average, it can launch a pop-up window asking "You are nailing it! Do you want to try a Medium level puzzle?". If the player realizes that the harder level is a great fit, he or she may very likely stay at this level and start spending more time on the game without even noticing.

2.1.4. Competition and Interaction Mechanisms (C)

One way to increase users' stickiness is to encourage users to compete and interact. Currently, ranking among all players within a level is available only when a player solves a puzzle but not on the "My Statistics" page. It is highly recommended to make this ranking information available at all times because competitive users may want to check this information frequently and play more puzzles just to hold their rank.

Another feature to be considered is a friend network where users can add their real-life friends and check friends' performance statistics. An even advanced feature can be a PK option where one user can invite another user, a friend or a randomly matched stranger, to compete on the same puzzle and PK winning counts will be displayed and ranked on the "My Statistics" page.

2.2. Marketing (B)

Men in their 50s and 60s are most likely to become loyal players of *Web Sudoku*, thus target marketing at this population can maximize per dollar return. Since older adults are relatively harder to target compared to younger adults or teenagers, a combination of both traditional (i.e. TV and radio commercials, catalog) and digital (i.e. direct email, social media, search engine optimization, pay per click ads) marketing approaches should be tested to see what works best for the game's specific customers.

3. Data Overview and Cleaning

Two datasets are used for analysis. The *dbsudoku-counts* dataset provides gameplay summaries and contains 311,919 rows of entries. Each entry describes the number of puzzles solved by a player, the number of puzzles solved with or without errors, the best solving time and the total time spent on a certain difficulty level. Altogether, 196,480 players appeared in this database. In addition to that, the *dbsudoku-users* dataset provides demographic information including birth year, country, US state and gender for 264,490 players. The two datasets are joined based on unique user IDs to perform analysis.

Web Sudoku has four difficulty levels, Easy, Medium, Hard and Evil. Although naturally, harder level puzzles take a longer time to solve, players are more addicted to harder level puzzles in the sense that they play more puzzles and for a longer accumulative time at harder levels. Among all players, 16.5% quit after just one game and 47.2% quit after 5 games or less. The average best time to solve a puzzle increases when puzzles get harder, but at a slower speed compared to the average time to solve a puzzle. The Evil level has the lowest average accuracy rate of 58.1% and the Easy level has the highest average accuracy rate of 69.2%.

The player population is averaged at 47.1 years old, with the two biggest age groups being young adults, aged between 25 and 35, and retired or close-to-retired elder people, aged between 55 and 70. Sixty percent of players are male. The top country of origin is the United States with nearly 38% of players coming from it, of which 12.5% comes from California.
A total of 113,535 (36.4%) entries are removed from the original dataset because they are missing either gameplay statistics, birth year or gender information. The detailed data cleaning process and exploratory analysis can be referred to in Appendix B.

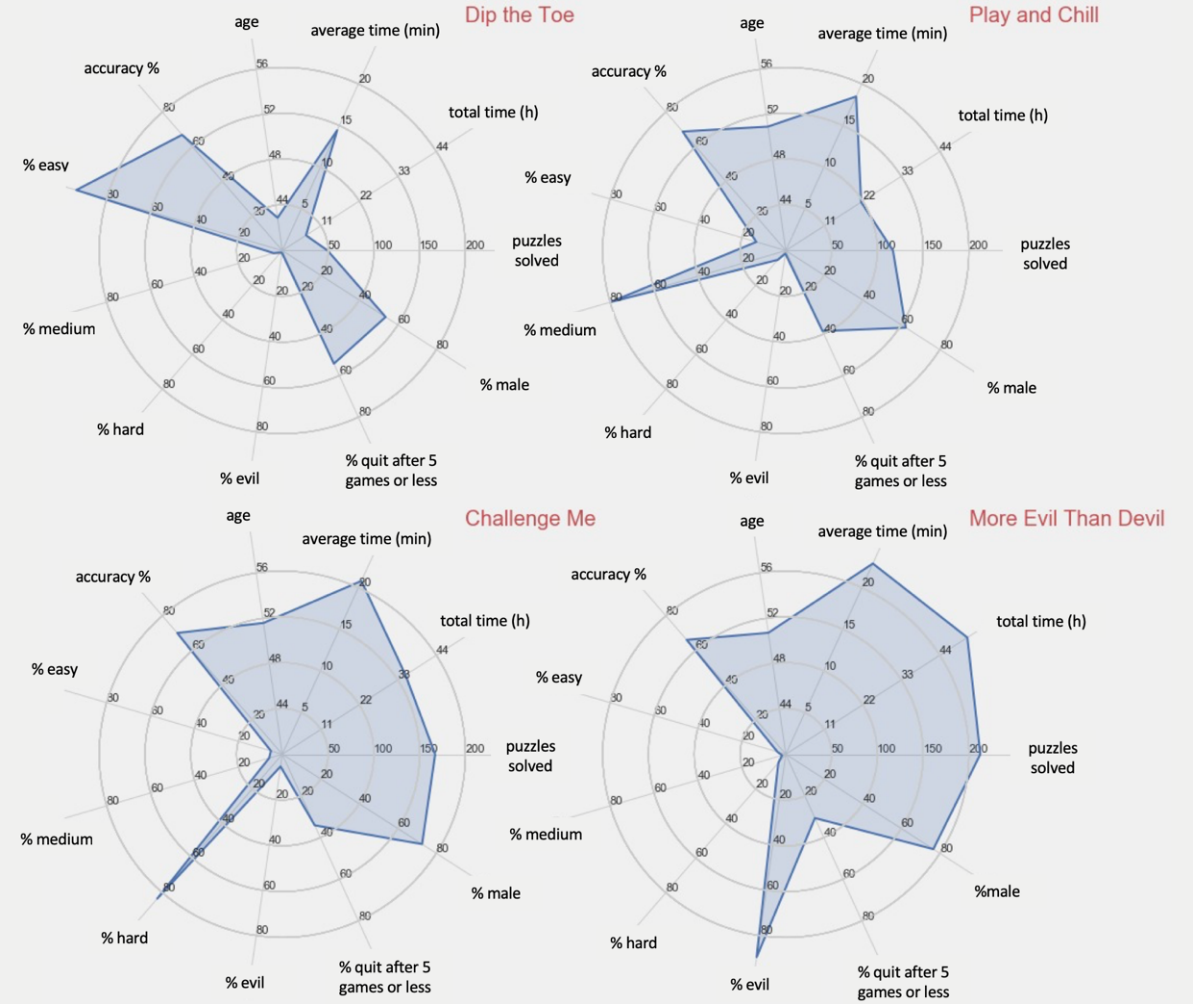
4. Player Segmentation

The combined and cleaned dataset is aggregated into one row for each player. Ten features including total puzzles solved, total time spent, the average time to solve a puzzle, percentage of puzzles solved at each difficulty level, average accuracy rate, age, and gender are calculated and normalized for model fitting (see Appendix C for details). K-means clustering algorithm is used

to segment players into five distinctive categories. The characteristics of each category are then examined and studied.

Current Web Sudoku players can be segmented into five distinctive types: *Dip the Toe*, *Play and Chill*, *Challenge Me*, *More Evil Than Devil* and *King of Puzzles*. The characteristics of each type of players are summarized in Table 1 (see Page 1) and visualized in Figure 1[1].

Figure 1: Player Characteristics by Type



Dip the Toe is the largest player group, consisting of 56.9% of total players. These players are casual players who almost exclusively work on Easy level puzzles and usually spend less than 15 minutes on one puzzle. Although on average they play 50 games and spend nearly 7 hours in their game lifetime, more than half of them quit after 5 games or less. On the contrary, an extreme 1% of them play more than 900 puzzles, and they play very quick (around 6 minutes per game) and at a very high accuracy rate (80.8%). The biggest commonality of players in the group

[1] Only *Dip the Toe*, *Play and Chill*, *Challenge Me* and *More Evil Than Devil* types are visualized because the *King of Puzzles* type has disproportionately large values in most fields compared to other types and thus is not suitable to be visualized along with the rest on the same scale.

is their love for Easy level puzzles. They are also the youngest among all groups, averaged at 43 years old, and they have the highest percentage of females among them (46.3%).

Play and Chill players consist of 18.3% of total players and four out of five times when they start a new game, the game is at Medium level. A typical player in this group is around 51 years old, plays 117 puzzles in their game lifetime and spends over 21 hours in total. Churn rate (percent of players who quit after 5 games or less) is at 38.7%. Three out of five players are male.

Challenge Me players consist of 13.3% of total players and their specialty is nailing Hard level puzzles. A typical player in this group plays 167 puzzles in their game lifetime and spend over 35 hours on the game. Churn rate is at 34.1% and 72.6% of players are male.

More Evil Than Devil players consist of 11.1% of total players and nine out of ten puzzles they play are at Evil level. A typical player in this group plays 212 puzzles in their game lifetime and spend nearly 52 hours on the game. Churn rate is 30.6% at and 76.6% players are male.

The last group of players are the true hardcore fans of *Web Sudoku*. *King of Puzzles* players only consist of 0.5% of total players. They on average play an extraordinary 8,173 puzzles and spend over 1,200 hours on the game. They play puzzles at all levels, with a special focus on the Hard and Evil levels. They solve puzzles extremely fast at only 11 minutes per round and with a high accuracy rate of 81.9%. These players tend to be older than other groups, averaged at 64 years old, and 77% of them are male.

5. Conclusions

Web Sudoku's most loyal users are typically males over 50. Most players tend to focus on one certain difficulty level and players who love harder games usually have lower churn rates, play more rounds and play longer. Based on these findings, features such as visual aids, tutorials for beginners, level suggestions, and competition and interaction mechanisms are highly recommended as methods to improve user retention. In the meantime, marketing strategies should be tailored to the target player population, with a specific emphasis on prospective users similar to current loyal players.

All recommendations should be A/B tested before full deployment to validate and maximize impact. More detailed game telemetry data such as information on puzzles unsolved or help button usage are urged to further study player behaviors and refine game designs.

Appendix A: Visual Aids Example

Source: Sudoku iOS app from Mind The Frog, Inc.

Appendix B: Data Exploration and Data Cleaning

1. Level

Web Sudoku has four difficulty levels, Easy, Medium, Hard and Evil. Based on the way the dataset is set up, a user should not have a record for a level that he or she has never solved a puzzle at. However, there are 776 records with no number of puzzles solved data or any gameplay data. Therefore, these records should be dropped before further analysis. The dataset is now down to 311,143 rows.

Most players tried and stayed at the Easy level. Data shows that 73.1% of players have tried the Easy Level, near 60% of which stayed at that level and never tried any other level until they quit the game (see Table 2 in Appendix D). On the contrary, 71-78% of users who played the Medium, Hard or Evil level would also try at least one other level.

2. Puzzles Solved, Average Time, Best Time and Total Time

Among all players, 16.5% quit after just one game and 47.2% quit after 5 games or less. For those who quit after one game, 67.3% quit at the Easy level.

Players solved more puzzles and spent more time on harder levels (see Table 3 in Appendix D). Although naturally, harder level puzzles take a longer time to solve, players are more addicted to harder level puzzles. In the ballpark, an Evil level puzzle takes around 24 minutes to solve, almost doubling the time needed to solve an Easy level puzzle. Yet each Evil level player normally solved 147 Evil level puzzles, nearly tripling the number of puzzles Easy level player solved at Easy level. As a result, hardcore players typically spent more than 35 hours on Evil level puzzles while casual players would only spend less than 6 hours on Easy level puzzles. The pattern remains true for all four difficulty levels. This inspired the hypothesis that the player population can be separated into different categories by looking at what level of games they like to play. This hypothesis is tested in Section 4.

There are 193 entries with the best time to solve a puzzle recorded as 0. These missing values are filled with the average best time among all players at the given level. The average best time to solve a puzzle increases when puzzles get harder, but at a slower speed compared to the average time to solve a puzzle.

3. Accuracy Rate

Accuracy rates are calculated as the number of puzzles solved without errors divided by the total number of puzzles solved. Unfortunately, whether a puzzle is solved without error is only recorded when a player is logged in to the website and there are 65,804 (21%) entries with no such data at all. Therefore, the accuracy rate of these entries will be filled with the average accuracy rate at the given level.

The more difficult a level is, the lower the accuracy rate gets, although the accuracy rates of the easiest three levels don't differ much (66-69%). Evil level puzzles have the lowest accuracy rate of around 58%.

4. Age

Thirty-three percent of the players are missing birth year information. Compared to those who registered birth year information, players who did not register generally played fewer games for less overall game time, thus are less likely to be the preferred user who can bring large revenues (see Table 4 in Appendix D). Therefore 94,395 (30.3%) records that don't have user birth year information are dropped form the dataset, with 216,748 rows remained.

Web Sudoku players averaged at 47.1 years old, with a bimodal age distribution (see Figure 2 in Appendix E). The two biggest age groups are young adults, aged between 25 and 35, and retired or close-to-retired elder people, aged between 55 and 70.

5. Country and U.S. State

Among the players who registered birth year information, 11.8% are missing country information. There are 230 different countries recorded in the dataset, much more than there are in the world, indicating many of these entries are not valid. Also, more than 37.9% of the players are from the United States, with over 60% of the players coming from the top five countries (United States, India, United Kingdom, Canada, and Australia) (see Figure 3 in Appendix E). On the other hand, California has the largest player population in the United States, followed by Texas, Florida, New York, and Illinois (see Figure 4 in Appendix E). Thirty-seven percent of American players are from one of these states.

Preliminary analyses didn't show evidence of county or state of origin being correlated to player behaviors. Considering that the goal of this analysis is to create player segments and clustering algorithms, in general, don't work well with categorical variables of such large cardinality, country and state information are included in the segmentation model.

6. Gender

After deleting users without birth year information, 7.5% of the remaining users did not provide gender information. These users will be dropped, removing 18,364 entries and resulting in a final dataset of 198,384 records from 122,168 players. Sixty percent of the players are male.

Appendix C: Feature Descriptions

The following 10 features are created for each player:

solved:	number of total puzzles solved across all levels
tot_time:	total time spent on all puzzles solved
avg_time:	average time spent on each puzzle solved, calculated as *tot_time* divided by *solved*
level_1:	ratio of Easy level puzzles solved to all puzzles solved
level_2:	ratio of Medium level puzzles solved to all puzzles solved
level_3:	ratio of Hard level puzzles solved to all puzzles solved
level_4:	ratio of Evil level puzzles solved to all puzzles solved
accuracy:	weighted average accuracy rate across all levels, calculated as $accuracy_{level\_1} \times level\_1 + accuracy_{level_2} \times level\_2 + accuracy_{level\_3} \times level\_3 + accuracy_{level_4} \times level\_4$
age:	age of the player as of 2019
male:	1 if the player is male, 0 if the player is female

Appendix D: Tables

Table 2: Player Composition by Level

	Easy	Medium	Hard	Evil
Player Count	143343	73154	53374	41272
Player %	73.1	37.3	27.2	21
Only This Level Count	84022	15858	11920	11791
Only This Level %	58.6	21.7	22.3	28.6

Table 3: Summary Player Statistics by Level

	Easy	Medium	Hard	Evil
Puzzles Solved	49	71	101	147
Total Time	5:42:11	12:05:27	20:54:02	35:18:41
Average Time	0:13:33	0:17:58	0:21:26	0:24:04
Best Time	0:09:57	0:13:03	0:15:13	0:16:42
Accuracy %	69.2	68	66.1	58.1

Table 4: Summary Statistics for Players with and without Birth Year Data

	Puzzles Solved	Total Time
Players Without Birth Year Data	56	10:32:18
Players With Birth Year Data	85	15:08:01

Appendix E: Figuress

Figure 2: Player Age Distribution

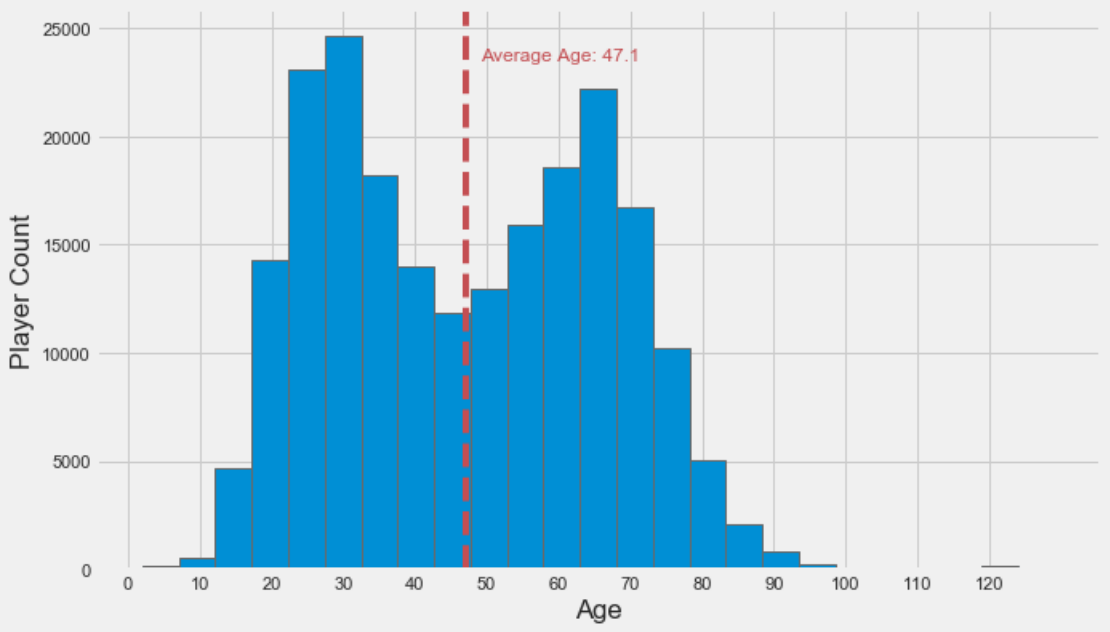


Figure 3: Player Composition by Country (Top 20)

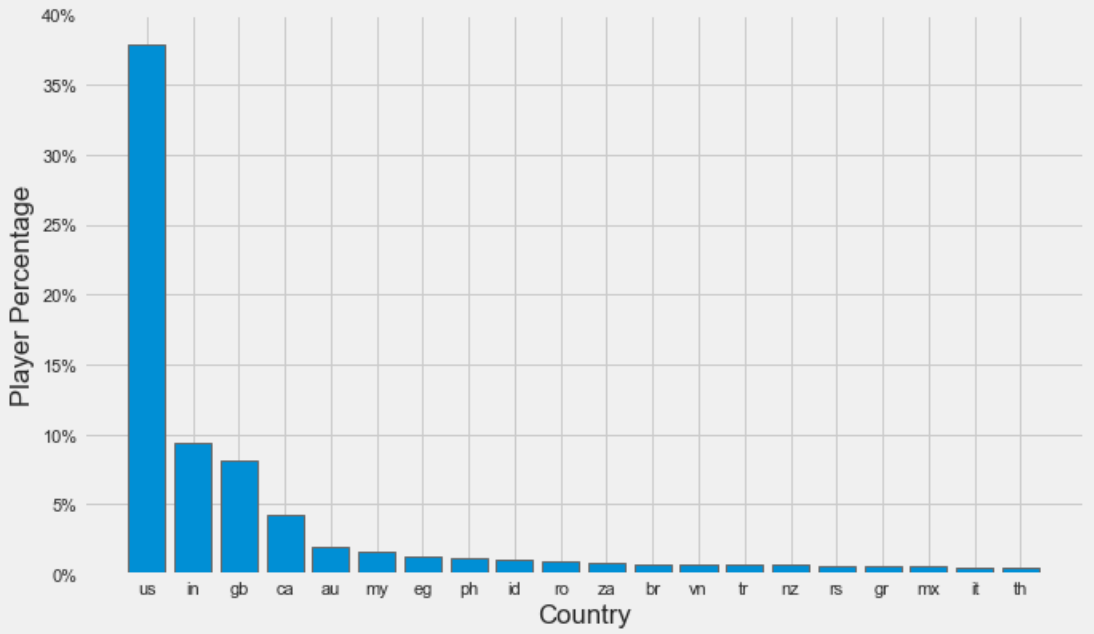


Figure 4: US Player Composition by State (Top 20)