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] - 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.

[1] www.websudoku.com

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

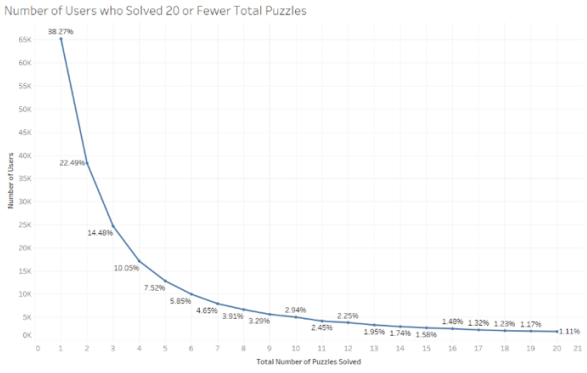
Prior to diving deep into the data, we first conducted surface level exploratory data analysis to identify any notable trends. The dataset we received consists of two files indexed by User IDs. For our analysis, we assume that each User ID is associated with a unique player. Since providing demographic 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%

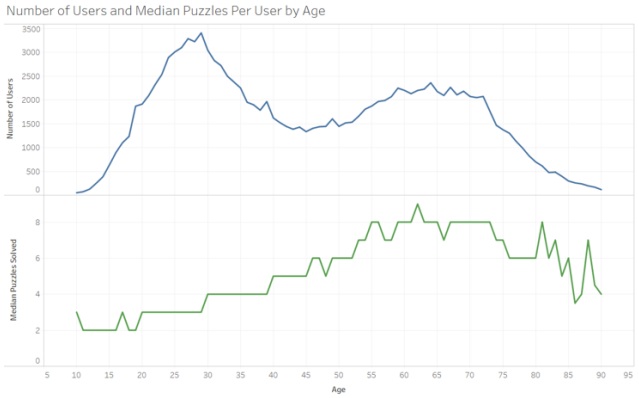
When comparing the original data to our cleaned data, there was a noticeable impact to the average number of puzzles completed per user. In the original data, one-third of all users only ever completed one puzzle, and over 80% of players completed five or fewer. However, this proportion drops to approximately 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.

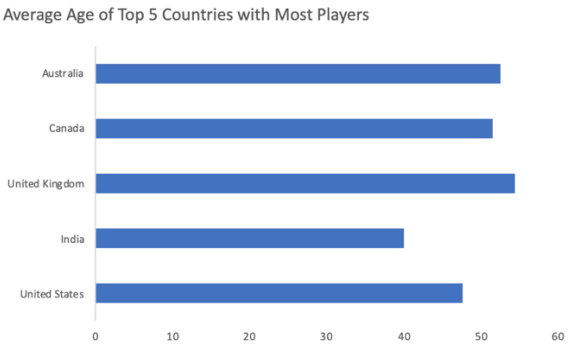
Figure 2 - Puzzle completion rate



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





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

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.6001	2.707	-1.330	0.184	-8.907	1.706
Age	2.9462	0.076	39.014	0.000	2.798	3.094
f	-11.4161	1.978	-5.771	0.000	-15.293	-7.539
m	7.8160	2.025	3.860	0.000	3.847	11.785
Average_time	-0.0525	0.002	-29.556	0.000	-0.056	-0.049

Omnibus:	434056.571	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	95376235237.784
Skew:	41.914	**Prob(JB):**	0.00
Kurtosis:	3953.999	**Cond. No.**	6.01e+18

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

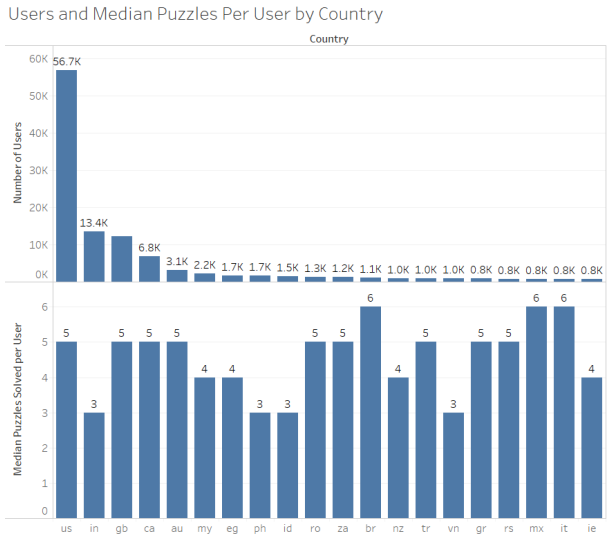
Level	% of Players who did not complete any puzzle
Level 1 (Easy)	0.32%
Level 2 (Medium)	0.21%
Level 3 (Hard)	0.18%
Level 4 (Evil)	0.16%
Overall	**0.19%**

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:



Level 1:

	Country	Num_User	Average_Age	Average_Time	Average_Puzzle_Solved
0	United States	38117	45.568119	805.331949	58.883543
1	India	8976	36.938837	812.017469	19.912656
2	United Kingdom	7511	51.305552	752.732962	57.203834
3	Canada	3987	49.731628	777.253922	56.398294
4	Australia	2002	50.343157	800.000598	75.063936

Level 2:

	Country	Num_User	Average_Age	Average_Time	Average_Puzzle_Solved
0	United States	18288	49.248250	1047.311604	77.674486
1	United Kingdom	4195	56.329678	1046.702995	82.688915
2	India	4087	40.101541	1093.006190	34.787130
3	Canada	2316	53.130397	1055.801433	119.581174
4	Australia	921	54.024973	1067.483457	120.720955

Level 3:

	Country	Num_User	Average_Age	Average_Time	Average_Puzzle_Solved
0	United States	11766	51.002635	1248.259823	120.913310
1	India	3591	43.231412	1367.568223	62.602339
2	United Kingdom	2785	58.185278	1248.520810	111.584201
3	Canada	1608	53.710199	1202.987657	128.102612
4	Australia	705	54.885106	1284.912043	106.520567

Level 4:

	Country	Num_User	Average_Age	Average_Time	Average_Puzzle_Solved
0	United States	9279	48.578834	1387.535303	155.971118
1	India	2523	46.324217	1555.824674	169.327784
2	United Kingdom	1951	56.219887	1415.151887	150.276269
3	Canada	1142	51.685639	1370.606031	185.900175
4	Australia	518	54.324324	1404.137696	161.538610

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

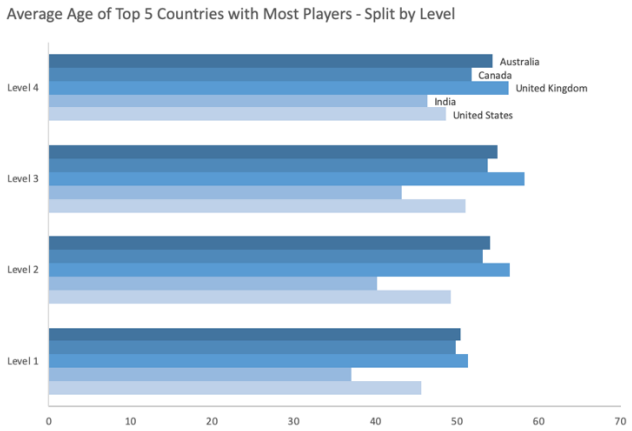


Figure 10 – Regression results for each difficulty level

Level1

OLS Regression Results

Dep. Variable:	Puzzles Solved	**R-squared:**	0.014
Model:	OLS	**Adj. R-squared:**	0.014
Method:	Least Squares	**F-statistic:**	322.0
Date:	Sat, 09 Nov 2019	**Prob (F-statistic):**	1.33e-207
Time:	16:30:18	**Log-Likelihood:**	-5.0449e+05
No. Observations:	67817	**AIC:**	1.009e+06
Df Residuals:	67813	**BIC:**	1.009e+06
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.9721	4.429	1.349	0.177	-2.708	14.652
Age	1.8627	0.083	22.388	0.000	1.700	2.026
Average_time	-0.0094	0.005	-1.775	0.076	-0.020	0.001
Best Time (s)	-0.0551	0.006	-9.093	0.000	-0.067	-0.043

Omnibus:	205741.923	**Durbin-Watson:**	1.981
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	30410177904.524
Skew:	45.137	**Prob(JB):**	0.00
Kurtosis:	3282.301	**Cond. No.**	3.52e+03

Level 2

OLS Regression Results

Dep. Variable:	Puzzles Solved	**R-squared:**	0.017
Model:	OLS	**Adj. R-squared:**	0.017
Method:	Least Squares	**F-statistic:**	191.2
Date:	Sat, 09 Nov 2019	**Prob (F-statistic):**	6.07e-123
Time:	16:31:03	**Log-Likelihood:**	-2.6277e+05
No. Observations:	33832	**AIC:**	5.256e+05
Df Residuals:	33828	**BIC:**	5.256e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	10.3999	9.287	1.120	0.263	-7.804	28.604
Age	2.6714	0.166	16.118	0.000	2.347	2.996
Average_time	-0.0036	0.009	-0.401	0.688	-0.021	0.014
Best Time (s)	-0.0778	0.010	-7.784	0.000	-0.097	-0.058

Omnibus:	103559.522	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16044274590.301
Skew:	46.203	**Prob(JB):**	0.00
Kurtosis:	3375.398	**Cond. No.**	4.82e+03

Level 3

OLS Regression Results

Dep. Variable:	Puzzles Solved	**R-squared:**	0.019
Model:	OLS	**Adj. R-squared:**	0.019
Method:	Least Squares	**F-statistic:**	161.0
Date:	Sat, 09 Nov 2019	**Prob (F-statistic):**	2.12e-103
Time:	16:31:50	**Log-Likelihood:**	-2.0006e+05
No. Observations:	25044	**AIC:**	4.001e+05
Df Residuals:	25040	**BIC:**	4.002e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	29.7424	14.397	2.066	0.039	1.524	57.961
Age	3.4169	0.253	13.525	0.000	2.922	3.912
Average_time	-0.0128	0.011	-1.146	0.252	-0.035	0.009
Best Time (s)	-0.0811	0.012	-6.720	0.000	-0.105	-0.057

Omnibus:	83731.140	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	36400477178.288
Skew:	59.083	**Prob(JB):**	0.00
Kurtosis:	5908.006	**Cond. No.**	6.16e+03

Level 4

OLS Regression Results

Dep. Variable:	Puzzles Solved	**R-squared:**	0.036
Model:	OLS	**Adj. R-squared:**	0.036
Method:	Least Squares	**F-statistic:**	247.3
Date:	Sat, 09 Nov 2019	**Prob (F-statistic):**	1.46e-157
Time:	16:32:29	**Log-Likelihood:**	-1.5940e+05
No. Observations:	19875	**AIC:**	3.188e+05
Df Residuals:	19871	**BIC:**	3.188e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	51.9706	16.326	3.183	0.001	19.970	83.971
Age	4.6541	0.293	15.898	0.000	4.080	5.228
Average_time	0.0064	0.011	0.585	0.559	-0.015	0.028
Best Time (s)	-0.1262	0.012	-10.666	0.000	-0.149	-0.103

Omnibus:	36193.549	**Durbin-Watson:**	2.008
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	67207913.295
Skew:	13.398	**Prob(JB):**	0.00
Kurtosis:	286.617	**Cond. No.**	6.77e+03