Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2318036

Exploring Wordle: Insights into Puzzle Solving and Tweet Shares Pattern

Summary

Wordle, a word puzzle that has attracted millions of people, is now owned by The New York Times. For the company's game editor, how the game is solved and shared on social media is critical information, as it can be used to guide future puzzle design and ultimately maximise the total number of players. **This paper aims to build a quantitative model based on word attributes and result reports on Twitter to predict the future pattern of players.**

After examining and cleaning the raw data, we first define 12 attribute indicators measuring its **familiarity** (how often used), **degree of association**, **degree of confusion** and **word composition features**. They are computed in advance because the following models will frequently use these indicators.

For Problem 1, we build a dynamic system called **Target-two-Players-Lost (T2PL)** based on the **SIR Model** to explain the daily fluctuation of Wordle reports. Players are additionally divided into two categories: general players and loyal players, each with a different attrition rate. This allows the model to simulate unequal decline rates over different time periods better. The relationship between word attributes and the number of hard mode players is also explored, and it is found that certain attributes affect the percentage of Hard Mode reports.

For Problem 2, we develop a **P&S Model**, which is a model that uses **simulation algorithms** and **gradient descent** to mimic the behavior of players in guessing words and sharing the game results. The simulator works by eliminating all unsatisfactory words using observable information, then randomly sampling words from the remaining word list using word frequency as the weight. However, we found that the simulation result could not perfectly match the true distribution. Therefore, we rescaled the distribution with 7 variables representing how players are likely to share their score when given different scores. They are optimised by gradient descent, and better distribution predictions could be generated. Using the P&S Model, we predict **the distribution of the word EERIE on March 1, 2023 is (0, 0, 9%, 29%, 45%, 14%, 3%).**

For Problem 3, we are required to classify puzzles by difficulty. We perform a cluster analysis on all reported trial distributions using **3 clusters K-means**, with each cluster labelled easy, medium and hard. We fit a **Random Forest Model** to divide the words into these three categories using the attribute indicators defined at the beginning. The correlation coefficient between each indicator and the difficulty is calculated, showing the direction in which these indicators affect the difficulty of the puzzles. The sensitivity of the clustering is discussed as well. Based on our model, **the difficulty of EERIE is hard**.

For Problem 4, we further explore the effects of word difficulty. Using **Linear Regression**, we found that word difficulty has an obvious effect on the number of results reported: harder puzzles lead to fewer reports. Difficulty also correlates with the percentage of people choosing Hard Mode, as we mentioned earlier. Through this part of the study, we find that the correlation is formed by word difficulty affecting the number of Normal Mode players.

With all the uncovered interactions between word attributes, puzzle difficulty, and game report patterns, Wordle operators could gain a deeper understanding of their players. Several sensible suggestions could also be made based on this discovery.

Keywords: Wordle; Dynamic system; Simulation; K-means; Random Forest

Contents

1 Introduction

1.1 Background and Literature Review

The word puzzle Wordle invented by Josh Wardle has attract millions of people due to its simplicity and myriad variation. Beyond the game, perhaps the major factor that cause Wordle went viral is its integrated sharing format consists of emoji squares, which spread widely through Twitter. In January 2022, Wordle was purchased by The New York Times Company and operated by them ever since. Only one piece of puzzle is released every day at the game's official website and this scarcity is also believed to contribute to Wordle's success.

In Wordle, player aim to crack a five-letter word within six guesses. Feedback is given after each guess is submitted: Letters highlighted in green indicates that the answer has the same letter at the same location. Yellow indicates this letter appears in the answer, but at another place. Grey indicates the letter is absent in the answer. Generally, it requires three to five tries for an average player, but it could vary significantly among different words. Addition to the normal version, there is also a Hard Mode Wordle, stipulating each discovered correct word (in Yellow or Green) must be maintained in the following guess[11].

Much research has focused on finding optimal strategy on solving the puzzle[1][4]. However, it seems that player's pattern is worthy to explore as well. As a major product under The New York Times Games, its operator would like to trace and predict the number of shared games on Twitter. Besides, released word should be well-considered, since easy problems could not challenge experienced player, while rare word like "rebus" or "tapir" make most fans frustrated[10]. Therefore, a quantitative model to predict distribution of attempts according to a given word is also expected.

1.2 Restatement of the Problem

Considering the background, in this paper we are required to solve the following problems:

- **Task 1:** Combine the game mechanics of Wordle to build a set of indicators that reflect the attributes of words, and apply them to the subsequent model.
- **Task 2:** Develop a model that explains the trends in the number of reported results and the percentage of scores reported that were played in Hard Mode, and use it to predict the number of reported results on March 1, 2023. Further, analyze the effect of word attributes on the percentage of scores reported that were played in Hard Mode.
- **Task 3:** Develop a model that can predict the distribution of the reported results based on words and use it to predict the distribution for the word EERIE on March 1, 2023. In addition, illustrate the uncertainty and accuracy of the model.
- **Task 4:** Classify words according to their difficulty and explain the relationship between the attributes of the words and the difficulty of the words.
- **Task 5:** Perform a comprehensive analysis of the dataset, and give some interesting conclusions.

1.3 Our Work

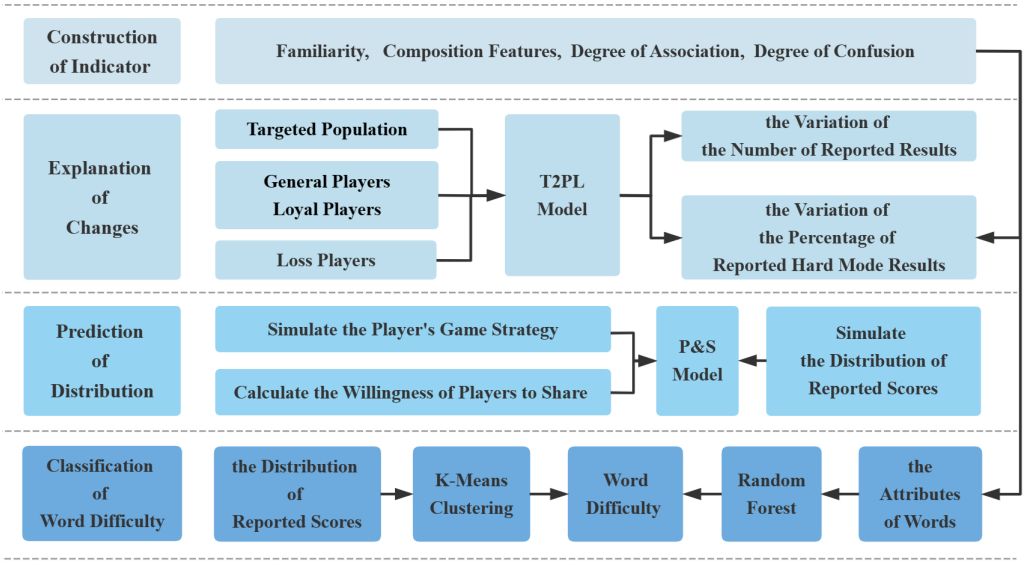

Figure 1: Flow chart of our work

Firstly, we constructed four types of indicators that can measure the familiarity, composition features, degree of association and degree of confusion of words, and used these indicators to reflect the attributes of words.

Secondly, we developed the T2PL Model based on the SIR model, a dynamic model that can well explain the overall trends in the number of reported results and the percentage of reported Hard Mode results. Based on this, we explored the effect of word attributes on the percentage of reported Hard Mode results.

Thirdly, we used the algorithm to simulate the strategies of wordle players when guessing words, so as to simulate the initial distribution of results. Considering the psychological characteristics of players, we added parameters indicating players' willingness to share their scores, and simulated the final distribution of reported results.

Fourthly, we clustered the words according to the distribution of scores and classified the words into 3 classes based on difficulty. The clustering results were used as labels to construct a Random Forest Model for classifying words' difficulty based on their attributes.

Finally, based on the results of the above model, we conducted further exploration and found some interesting conclusions.

2 Assumptions and Notations

2.1 Model Assumptions

Considering the conditions required for modeling, we make following assumptions:

- **Assumption 1:** There will not be a shift in the general trend of Wordle's daily user number.
  Justification: This is required to predict future trend based on observed daily usage.

- **Assumption 2:** Most players use rational strategies.
  Justification: To establish a mathematic model for potential player, it is necessary to assume that they are actually using a strategy and will not take unessential moves. Otherwise, it would become meaningless to simulate result based on potential strategies.

- **Assumption 3:** There is no significant change in players' skill along time.
  Justification: As Wordle is played for a period, players are expected to improve their strategies which might affect attempt times distributions at different date. However, experienced players are giving up Wordle while rookies are joining simultaneously, producing an opposite effect. It would be too complicated to consider these possibilities.

- **Assumption 4:** In task 2 (T2PL Model), player and those who share their result are not distinguished.
  Justification: For convenience, players are modelled in Task 2, although Twitter report numbers are actually used. This is because there is not enough information to distinguish between the two categories in this step, and it makes sense to switch from modelling players to modelling players who share their results.

2.2 Notations

Symbol	Definition	Symbol	Definition
N	Number of reported results	P	Number of all players
H	Number of reports in Hard Mode	P_{loy}	Number of loyal players
P_H	Percentage of scores reported that were played in Hard Mode. $P_H = H/N \times 100\%$	P_{gen}	Number of general players
D	All words of length 5 in dictionary data	W_{sj}	Probability of sharing when finishing with j tries
T	Number of Targeted Population	p_{ij}	Probability of solving Wordle #i with j tries

Table 1: Symbol table.
Word attribute indicators is not included, because they are explaned in detail below.

3 Data Preprocessing

The dataset `Problem_C_Data_Wordle.xlsx` contains 359 days of Wordle report information. Each row consists of date, the word of the day, the number of reported results, the hard mode results, and the distribution of each number of attempts. There are no missing values in the table, but on closer inspection several words are misspelled. We manually correct each of these by searching for the correct Wordle answer using the question number for that day.

Date	Contest number	original word	correct word
2022/12/16	545	rprobe	probe
2022/12/11	540	naïve	naive
2022/11/26	525	clen	clean
2022/10/5	473	marxh	marsh
2022/4/29	314	tash	trash

Table 2: All corrected words

We also checked the sum of the percentages of each distribution and found that the sum of question 281 ”nymph” on 2022/3/27 was 126, which is likely to be an outlier. We do not know which of the percentages is wrong, so we cannot simply scale it back. Therefore, this row of data is not used in the prediction percentage problem. Question number 529 ”study” on 2022/11/30 seems like an outlier as well. The Number of reported results in this row is 2569 and the Number in hard mode is 2405, which is obviously very different from the data in other rows, so it is corrected to 25690.

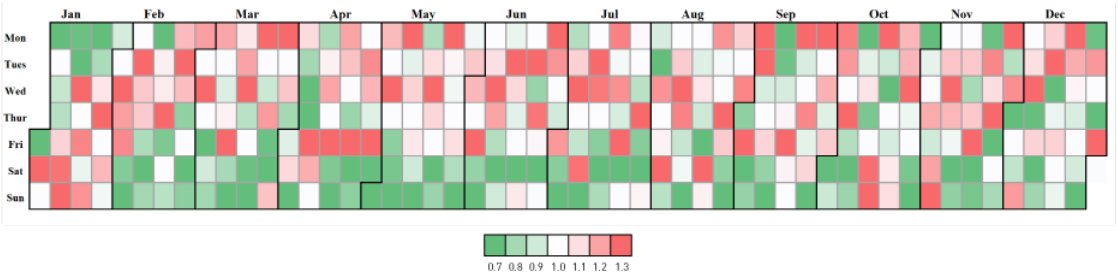

Figure 2: calendar map of submissions

Considering a possible weekly periodic pattern in the number of submissions, a calendar graph (Figure 2) is created to visualise the weekly variation. Each day is referenced to the average value of the week in which it is located. The more it exceeds the average value, the redder it is, and the more it falls below the average value, the greener it is. It can be seen that although there are some differences between each day of the week, there is no general pattern. Therefore, we assume that the fluctuation is only based on the general trend of the reports and the difficulty of the word.

4 Task 1: Word Attribute Indicators

Considering the game characteristics of Wordle, the indicators we constructed should reflect the spelling structure of the word and the ease with which people associate it as much as possible. To achieve this, we constructed the following framework of indicators.

- **Familiarity:** how familiar people are with the word and how commonly the word is used.
- **Composition features:** structural features of words, such as whether they contain the same letters.

- **Degree of association:** the likelihood that the Wordle will give more valid information (i.e., green or yellow tiles) when the player guesses the word.
- **Degree of confusion:** the degree of similarity between the word and other words. When the degree of similarity is too great, the player may need to spend more times to verify which of the many similar words is the correct answer.

Based on the above framework, we constructed 12 indicators:

Measured Feature	Symbol	Meanings
Familiarity	F	Frequency of the word
Composition feature	N_c	Number of the most repeated letter in the word
	N_v	Number of vowels in the word
Degree of association	LF_i	Frequency that the ith letter of the word appears in the ith position of all words. ($i = 1, 2, 3, 4, 5$)
	LF_{sum}	Letter Frequency Sum
	N_{FI_2}	Number of frequent 2-itmes
Degree of confusion	N_{D_1}	Number of all the words in D whose Levenshtein distance is 1 from the word
	N_{D_2}	Number of all the words in D whose Levenshtein distance is 2 from the word

Table 3: Symbol table of all selected parameters

A detailed description of certain symbols are present below:

- **Number of the most repeated letters (N_c):** For example, N_c("apple") = 2, N_c("mummy") = 3. Because players don't usually try two or three of the same letters at the same time, we predict the greater N_c, the less likely the word is to be guessed.
- **Letter Frequency of the ith Letter (LF_i):** For example, the first letter of apple is a, then LF_1("apple") is the proportion of words with the first letter a in D. The greater LF_i is , the greater the probability of a green tile appears at the ith position.
- **Letter Frequency Sum (LF_{sum}):** Indicator LF_{sum}, representing the sum of letter frequency in the word. First, letter frequency(LF) is calculated by:

$$\text{(initialize)} \quad \underset{c\in C}{\forall}\Big[LF(c) = 0\Big] \tag{1}$$

$$\text{(repeat)} \quad \underset{W\in D}{\forall}\Big[\underset{c_1\in C}{\forall}\Big[LF(c_1) = LF(c_1) + \sum_{c_2\in W}\begin{cases}1 & if\ c_1 = c_2\\ 0 & otherwise\end{cases}\Big]\Big] \tag{2}$$

$$\text{(finally)} \quad \underset{c_1\in C}{\forall}\Big[LF(c_1) = \frac{LF(c_1)}{\sum\limits_{W\in D} len(W)}\Big] \tag{3}$$

Where C is the alphabet set, D is the word set, $len(w)$ represent the length of the word.

After calculating LF, LF_{sum} could be produced by summing all characters in a word:

$$LF_{sum}(\text{"}c_1c_2c_3c_4c_5\text{"}) = \sum_{i=1}^{5} LF(c_i) \tag{4}$$

The greater the LF_sum, the greater the probability of yellow tiles appearing when playing Wordle.

- **Number of frequent 2-items (N_{FI_2}):** This indicator count the number of frequent items contained in a word in the frequent 2-itenset(FI_2). FI_2 is calculated by FP-Growth algorithm, an algorithm for frequent pattern mining, with supporting degree $\frac{|D|}{10}$. Greater N_{FI_2} increase the probability of information gain by close guesses, so it is assumed to be easier when N_{FI_2} is high.
- **Number of 1/2 Levenshtein Distanced Words (N_{D_1}, N_{D_2}):** Levenshtein distance is the minimum number of editing operations required to change from one to the other between two strings, so all words with Levenshtein distance equals to 1 or 2 are all similar words to the answer. When N_{D_1} and N_{D_2} are high, player may need to spend more times to verify which of the many similar words is the correct answer, and more likely fail to guess a word.

5 Task 2: Predicting Daily Reports & Hard Mode Percentage

5.1 Problem Analysis

It is obvious that Wordle has become viral last year. People send these iconic Wordle result tweets, attracting more people to try the game and share it. At the beginning of February, shared tweets rise in an exponential rate, then it drops gradually after March. Thus, this tendency could well be captured by SIR(Suspected-Infected-Recovered) model, which is originally intended to explain infectious diseases[9]. Establishing similar dynamic model to explain game user amount fluctuations is also observed in many researches, proved to be successful[5].

5.2 Establishment of the Model

Like SIR model, we can assume three groups of people: targeted population, current players, and lost players. Targeted population has the potential to become player, and this rate of transformation is proportional to total current players, owing to the advertising effect of their tweets. Therefore, decrease of target population could be written as the multiple of targeted population, current players and a constant α. For current players, beside new players joining in, there are also players who became tired and give up the game. These group of people is related only to the number of current players and constant β. With these conditions, a dynamic system could be formulated as follow:

$$\begin{cases} \dfrac{\partial T}{\partial t} = -\alpha \cdot T \cdot P \\ \dfrac{\partial P}{\partial t} = \alpha \cdot T \cdot P - \beta \cdot P \\ \dfrac{\partial L}{\partial t} = \beta \cdot P \end{cases} \tag{5}$$

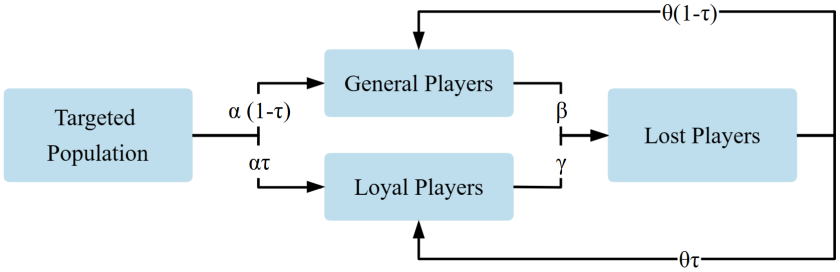

Figure 3: two-type player dynamic system

However, this model failed to satisfy both the sharp decrease after peaking in February and mild decrease in the second half of the year. After all, while many people might gradually abandon the game, others might maintain their fondness for Wordle. Modelling all players having the same player leaving rate β might not work well. Therefore, we improved the model by dividing all players into loyal players who has lower tendency to get tired with Wordle and keep sending Wordle tweets, and general players who follow the trend and leave with the trend, Which is named **T2PL(Target-2-players-Lost) model**. Each two category has a unique decay rate β and γ. Both types of player's tweets would have similar effects, so the sum of two categories is used when calculating decrease in potential population. What percentage do players fall into each category also need to be determined. We set variable τ as the rate of new players becoming a loyal player and variable p as the initial split of two types. In case there were players who wants to try Wordle again after leaving for a while, an additional variable θ is used to describe the lost-return rate, and this means the initial value of lost player L_0 also has to be modeled in advance. Now the final formulation would be:

$$\begin{cases} \dfrac{\partial T}{\partial t} = -\alpha T \cdot (P_{loy} + P_{gen}) \\ \dfrac{\partial P_{loy}}{\partial t} = \tau \cdot \left[\alpha T \cdot (P_{loy} + P_{gen}) + \theta L\right] - \beta \cdot P_{loy} \\ \dfrac{\partial P_{gen}}{\partial t} = (1-\tau) \cdot \left[\alpha T \cdot (P_{loy} + P_{gen}) + \theta L\right] - \gamma \cdot P_{gen} \\ \dfrac{\partial L}{\partial t} = \beta \cdot P_{loy} + \gamma \cdot P_{gen} - \theta L \end{cases} \tag{6}$$

With initial values: $T_0,\ P_{loy_0} = P_0 \cdot p,\ P_{gen_0} = P_0 \cdot (1-p),\ L_0$

5.3 Solving the Model

By estimating a set of initial values ($\alpha,\ \beta,\ \gamma,\ T_0,\ L_0,\ \tau,\ \theta,\ p$) and observed initial value P_0, the dynamic system can be numerically solved using forward Euler method. Our goal is to fit the dynamic system towards the result curve. The objective function is the mean square error between 1 and the ratio between estimated value and actual value at each sample point. It is re-scaled because the range between maximum and minimum value at the ground truth curve is too large.

$$\min_{\alpha,\beta,\gamma,T_0,\tau,p} \sum_{t=1}^{N} \left(1 - \frac{\hat{P}_i(\alpha,\ \beta,\ \gamma,\ T_0,\ \tau,\ p)}{P_i}\right)^2 \tag{7}$$

Without explicitly solving the function, our team use auto-gradient package (Pytorch) to optimize the parameters. Each parameter is updated by gradient descent. As long as the initial value are roughly the correct scale, it will find the local minimum of the parameters that produce a well-fit curve.

5.4 Solution and Result

After 5000 epoch of training with learning rate 0.01, we find the optimal value of each variable.

	α	β	γ	θ	T_0	L_0	τ	p
initial value	5×10^{-7}	0.02	2×10^{-3}	2×10^{-4}	310000	30000	0.2	0.2
optimal value	3.03×10^{-7}	0.0228	3.7×10^{-3}	2.0×10^{-4}	431193	29918	0.0992	0.1683

Table 4: initial value and optimal value of all variable

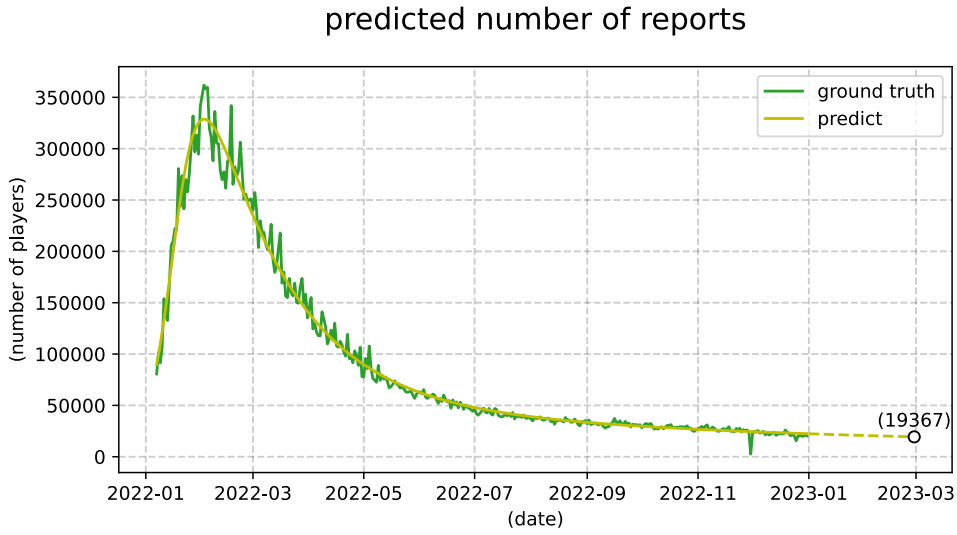

Figure 4: Predicted number of reports. Predicted days are drawn with dotted line.

The plotted curve is shown in Figure 4. The model is proved to excellently capture the tendency of reported Wordle result in 2022. In 2023.3.1, it is predicted to have 19367 users sending Wordle

results. Because the model cannot provide a range of possible prediction by itself, we estimate by considering the MAPE (Mean Absolute Percentage Error), which is often used as a measurement of prediction accuracy. It is defined as:

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{\hat{P}_t - P_t}{P_t} \right| \tag{8}$$

Given MAPE = 0.083 on observable date, we can assume the result will fall within ±MAPE. Therefore, the estimated region is given as:

$$\text{Confidence Interval}(t) = \left[(1 - \text{MAPE}) \cdot \hat{P}_t, (1 + \text{MAPE}) \cdot \hat{P}_t \right] \tag{9}$$

Therefore, the prediction interval for the number of reports is $[17760, 20976]$.

In addition, this model could provide additional insight into the structure of Wordle players. We have plotted the change in each category in 2022. According to Figure5, general players increase rapidly in February and then decrease to a very low level at the end. Loyal players, on the other hand, don't change much throughout the process. The model predicts that there will be only 19% general players on 31 December 2022 and 21% on 1 March 2023, showing that the majority of players will be loyal and probably more skilled players in the near future.

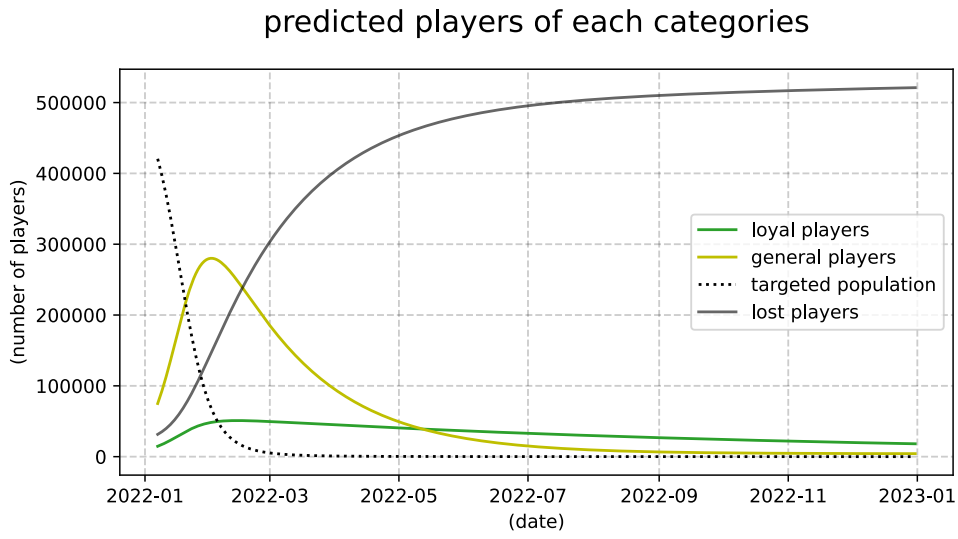

Figure 5: two-type player dynamic system

5.5 Hard Mode Percentage Estimation

Unlike the trend in the number of reported results, the percentage of scores reported that were played in Hard Mode posting slowly increased from 2% to 10% and then remained stable. This overall trend can perhaps be explained by the change in player structure mentioned above, where loyal players

are more likely to choose Hard Mode because of their higher level of play and love for the game. As shown in the figure below, the trend of the percentage of loyal players in the overall player base is basically the same as the trend of the percentage of scores reported that were played in Hard Mode.

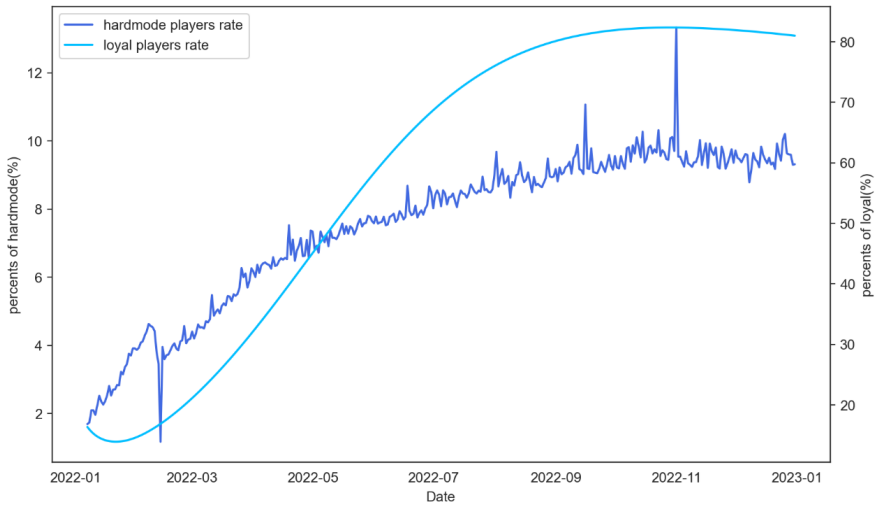

Figure 6: Curve of hard mode reports' percentage and loyal players' percentage, re-scaled.

In addition, there were always small fluctuations in the percentage of reported scores played in Hard Mode around the overall trend, and these fluctuations may be related to the attributes of the word of the day. To explore this relationship, we established a multiple linear regression model with the percentage of Hard Mode reports as the dependent variable, and all indicators reflecting word attributes as independent variables. In order to avoid the influence of the overall trend on the exploration of fluctuations, the percentage of scores reported in the previous day is selected as the control variable.

The regression results were as follows (only the results of variables with significant regression coefficients are given here):

P_H on the previous day	F	N_c	LF_4	LF_{sum}	N_{FI_2}
0.9608^{***}	$-6.342 \times 10^{-12*}$	$2.776 \times 10^{-3***}$	0.0291^{*}	-0.0423^{***}	$6.777 \times 10^{-4*}$
(0.0111)	(2.814×10^{-12})	$(5.959 \times 10{-}4)$	(0.0113)	(0.0110)	(2.761×10^{-4})

Table 5: Regression result, showing coefficient and corresponding standard error.
Note: *, **, *** represent $p < 0.05$, $p < 0.01$, $p < 0.001$.

The results show that F and LF_{sum} have a significant negative effect on the percentage of scores reported that were played in Hard Mode, while N_c, LF_4 and N_{FI_2} have a significant positive effect on it. And it can be found through the later analysis that these indicators are the important factors that determine the difficulty of words.

Select the two attributes with the strongest salience and plot the categorical box plot of the percentage of scores reported that were played in Hard Mode about them. The results show that the percentage of scores reported that were played in Hard Mode is higher when the words contain the same letters, and lower when the words have a higher frequency of letters.

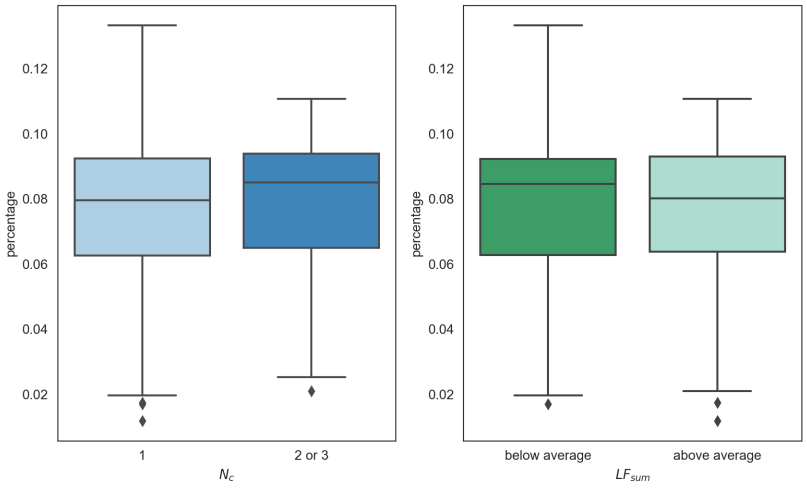

Figure 7: Box plot result

6 Task 3: Predicting Report Distribution

6.1 Problem Analysis

Wordle is an interactive game. When a player fills in a word, Wordle gives feedback and the player chooses a new word based on Wordle's feedback. Each step of the player's strategy is related to the word filled in the previous step, which provides so many possibilities for the game to proceed that it is almost impossible to calculate the probability of the player guessing the word within a specified number of times. In order to predict the associated percentages of (1, 2, 3, 4, 5, 6, X) for a future date, we designed a simulation algorithm to simulate the player's strategy of playing Wordle, and obtained the simulated distribution by performing 5000 simulations for each word.

6.2 Establishment of the Model

The main idea of the simulation algorithm is: Players randomly select a word from the dictionary to guess, and get the game state of Wordle. For the state of each letter:

1. If it is green, only the word with the letter in the corresponding position will be selected for subsequent guesses;
2. If it is yellow, only words with that letter in other positions will be selected for subsequent guesses;
3. If it is gray and not among the yellow letters, the subsequent guesses will not select words containing this letter.

Above process is repeated until the player has succeeded or guessed more than 6 times.

Note: Here we assume that players know all the words in the dictionary D.

Algorithm 1 Simulate(W_S)

(**Note:** Evaluate-Guess(W_S,W_G) return a game state of wordle. For example: W_S =”tools”, W_G=”brook”, it return ”NNYMN”. ”Y” means the letter is in the answer and in the correct position. ”M” means the letter is in the answer but in a different position. ”N” means the letter is not in the answer.)

Input: Solution word W_S

Output: The list of game states L_G

$D_W \leftarrow D$;
$L_G \leftarrow \phi$;
for i from 1 to 6 **do**
    random choose a guess word W_G from D_W;
    $S_i \leftarrow$Evaluate-Guess(W_S, W_G);
    $L_G \leftarrow L_G \bigcup \{S_i\}$;
    if S_i =”YYYYY” **then** break;
    end if
    for j from 1 to 5 **do**
        if S_i^j =”Y” **then**
            remove from D_W the words W satisfying $W^j \neq W_G^j$;
        else if S_i^j =”M” **then**
            filter out the word W satisfying $W^k = W_G^j, k \neq j$ from D_w;
        else if S_i^j =”N” & $W_G^j \notin \{W_G^k | S_i^k =$”M”, $k = 0, 1, 2, 3, 4\}$ **then**
            remove from D_W the words W satisfying $W^j = W_G^j$;
        end if
    end for
end for
return L_G;

We subtracted the probabilities on the corresponding points of the simulated distribution and the real distribution, and took the absolute value as error. The value of error is 5.36%, which was not satisfactory. By comparing the mean value of the simulated distribution with the mean value of the real distribution, we found that the mean value of 79.6% of the simulated distribution was larger than that of the real distribution, i.e., the simulation results showed that players needed more attempts to guess. We believe this is due to the fact that players who guessed correctly with fewer attempts were more likely to share their results on Twitter, while players who guessed correctly after many attempts were more likely not to share on Twitter.

According to above considerations, we propose **P&S(Played and Shared)** model. Probability of a player sharing after playing Wordle(either successfully or failed) is defined as W_{si}, indicating the willingness to share after making i attempts to success($W_{si} \in [0, 1]$). We corrected the simulated distribution according to W_{si} and used gradient descent method to minimize error. The corrected formula is as follows:

$$\hat{p_{ij}} = \frac{p_{ij} W_{sj}}{\sum_{k=1}^{5} p_{ik} W_{sk}} \times 100\% \tag{10}$$

p_{ij} is the percentage of players who succeeded after j attempts for the ith word, $\hat{p_{ij}}$ is the percentage of players who succeeded after j attempts for the ith word after correction.

The optimal parameters obtained by the gradient descent method are shown below:

W_{s1}	W_{s2}	W_{s3}	W_{s4}	W_{s5}	W_{s6}	W_{s7}
1.000	0.993	0.940	0.755	0.698	0.696	0.301

Table 6: Scale parameters' optimal value

We corrected the simulated distribution with the above parameters and compared it with the real distribution again, and found that the corrected distribution could fit the distribution of the data set better. The error is reduced from 5.36% to 1.50%, and only 45.6% of the simulated distributions have a mean value greater than the real distribution, which is a more desirable result. We named this model P&S Model, where P stands for playing the game and S stands for sharing the results.

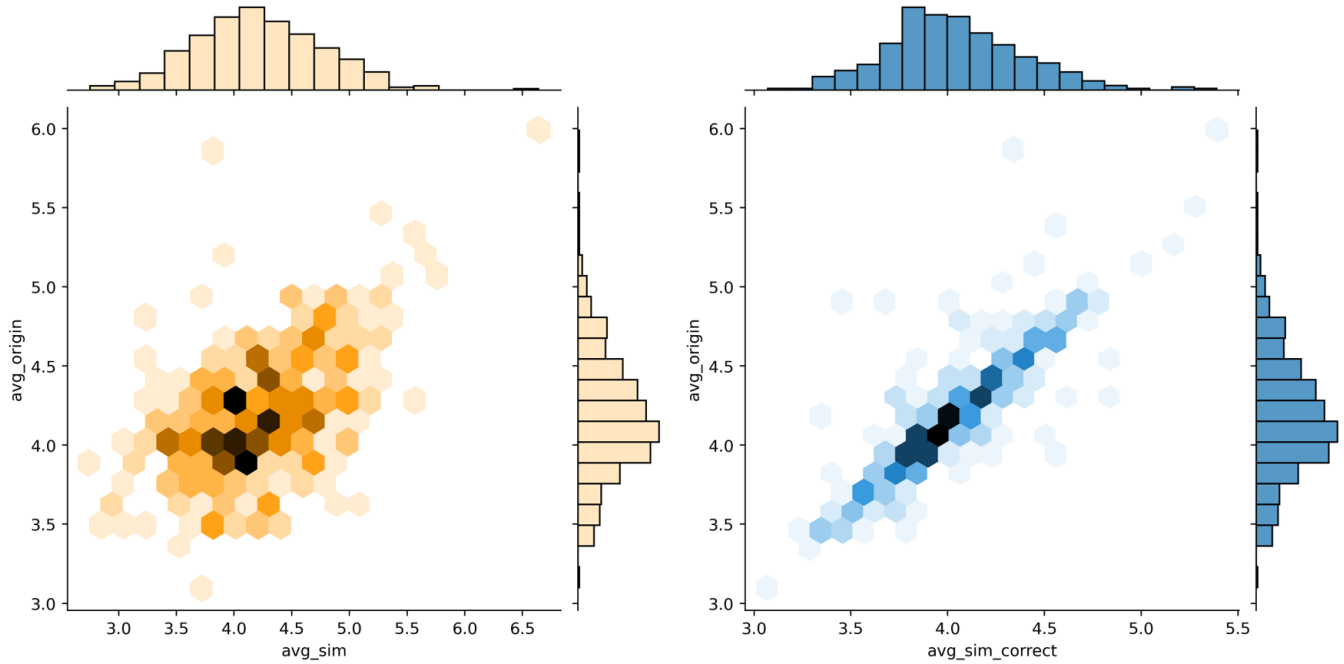

Figure 8: Visualize the simulated distribution versus the true distribution, before and after correction.

6.3 Predict Confidence and Uncertainties

Based on our model, the distribution of the word EERIE on March 1, 2023 is (0, 0, 9%, 29%, 45%, 14%, 3%). Because the value of error is 1.50%, we speculated that the probability of each point of the real distribution will probably fall within plus or minus 1.50% of the simulated distribution.

Due to the stochastic nature of the simulation, the prediction results of the P&S Model will not always be the same, which will lead to some uncertainty in the prediction results.

7 Task 4: Word Difficulty Classification

In the previous section we built a model to analyse the possible distribution for each word. However, puzzle designers seeking insight from this model would find it difficult to use this information. It is expected that a comprehensive label describing the difficulty of the puzzle will be created and that the classification into these categories will be produced by explainable attributes of each 5-letter word. Therefore, we perform a cluster analysis on the score distributions in the dataset, hoping to identify 3 types of difficulty: Easy, Medium and Hard. We then enumerate all possible attributes in each word and run a classification algorithm into three difficulty levels using these attributes. In this way, the most important attributes for classification could be discovered, and then we could gain a thoughtful understanding of how each category relates to the word feature.

7.1 Cluster Analysis

For clustering algorithms, we compare two commonly used methods: k-means[3] and Gaussian Mixture Model (GMM)[7].

K-means tries to divide the data into k categories, generating k category centroids C_k. For each point, its nearest centroid is expected to be the centroid of the category to which it belongs. By initiating several random points as the centroid, it is iteratively updated using this formula:

$$x_n \in C_k^{(t+1)}, \text{ if } k = \arg\min_k d(x_n, c_k^{(t)}) \tag{11}$$

$$c_k^{(t+1)} = \frac{1}{|C_k^{(t+1)}|} \sum_{x_n \in C_k^{(t+1)}} x_n \tag{12}$$

GMM considers all observed data sampled from k normal distributions $N_k(\mu_k, \Sigma_k)$ and tries to find the maximum likelihood pairs of parameters μ_k, Σ_k. It can be solved using the expectation maximisation (EM) algorithm. In the E-step, the responsibility (posterior distribution of clusters) of each sample point is calculated.

$$p^{(t)}(k|x_n) = \frac{p(x_n; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{i=1}^{K} p(x_n; \mu_i^{(t)}, \Sigma_i^{(t)})} \tag{13}$$

Then in M-step, the responsibility is fixed and the model parameters are updated as:

$$\mu_k^{(t+1)} = \frac{\sum_{n=1}^{N} p^{(t)}(k|x_n) x_n}{\sum_{n=1}^{N} p^{(t)}(k|x_n)} \qquad \Sigma_k^{(t+1)} = \frac{\sum_{n=1}^{N} p^{(t)}(k|x_n)(x_n - \mu_k^{(t+1)})(x_n - \mu_k^{(t+1)})^T}{\sum_{n=1}^{N} p^{(t)}(k|x_n)} \tag{14}$$

After convergence, each sample point belongs to the distribution with maximum responsibility.

$$x_n \in C_k, \text{ if } k = \arg\max_k p(k|x_n) \tag{15}$$

In order to decide the better algorithm and choose the appropriate number of clusters, we compare the results using the Silhouette Coefficient[8] and the Calinski-Harabasz Index[6]. The silhouette

coefficient is obtained by averaging the silhouette values of all sample points, which describe how close the sample point is to its cluster compared to others. The greater silhouette values is, more distinguishable the cluster would be. Its mathematical formulation is:

$$SC = \frac{1}{N}\sum_{i=1}^{N} s(x_n)\ , \qquad s(x_n) = \frac{b(x_n) - a(x_n)}{\max(a(x_n), b(x_n)} \tag{16}$$

$$\text{Where}: a(x_n) = \frac{1}{C_i - 1}\sum_{x_m \in C_i, m \neq n} d(x_m, x_n) \qquad \text{(same cluster closeness)}$$

$$b(x_n) = \min_{j \neq i} \frac{1}{C_j}\sum_{x_m \in C_j} d(x_m, x_n) \qquad \text{(different cluster closeness)}$$

The Calinski-Harabasz index is a ratio between the variance of the cluster centres and the variance of each cluster. If the centres of each cluster are spread apart and the sample points are clustered around each cluster centroids, then the Calinski-Harabasz Index is expected to be high.

$$\text{C-H Index} = \left[\frac{\sum_{k=1}^{K} |C_k| \cdot ||c_k - \mu||^2}{K-1}\right] \Big/ \left[\frac{\sum_{k=1}^{K} \sum_{x \in C_k} ||x - c_k||^2}{N-K}\right] \qquad (c_k \text{ is the centriod of } C_k) \tag{17}$$

Score of both indicators in shown in the table:

	k-means (3)	GNN(3)	k-means(2)	GNN(2)
silhouette coefficient	0.369	0.306	0.425	0.366
Calinski-Harabasz index	311.9	203.4	329.6	122.3

Table 7: scores for different kinds of cluster algorithm

Finally, we decide to create 3 clusters using k-means. Although 2 class k-means seems to have a higher evaluation score, detailed classification is better for drawing more specific conclusions, and the performance of 3 class k-means is not too far from it. To observe the clustering effect more intuitively, we use the t-SNE algorithm to reduce the data to 2 dimensions. Figure 9 visualize all three clusters, labeled "easy", "medium" and "hard" according to expected attempts of each class centroids. Therefore, we successfully find a method to allocate each word into three difficulties based on how they are reported to be solved.

7.2 Difficulty Classification

Now that all the word features have been quantized into 12 numerical values and assigned a difficulty level, a supervised classification model could be built to predict the difficulty level based on word attributes. We use Random Forest classification[2] in this task. Random Forest has been shown to provide stable classification results on a wide range of data. The motivation of the algorithm is to reduce variance by bagging on the training set and feature space. B decision tree is trained on a sampled

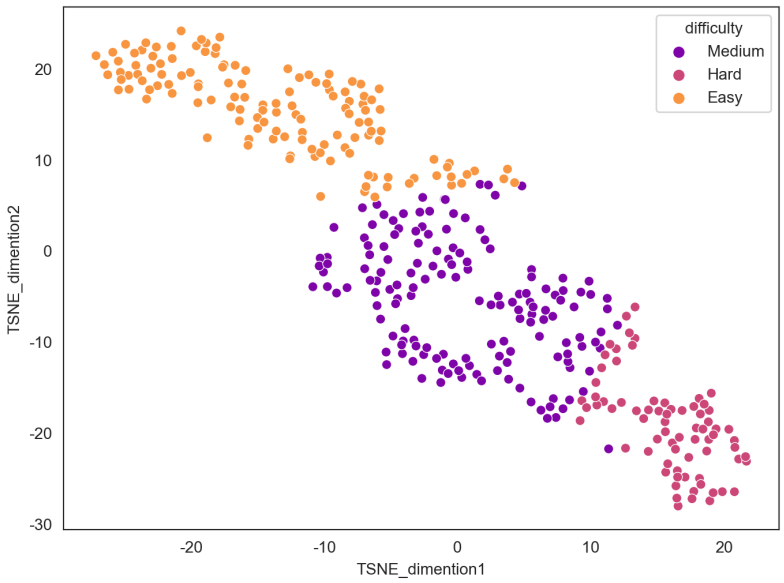

Figure 9: Visualized 3 clusters projected on 2 dimensions

set of all data with certain feature is masked, and classification result is produced by summarising all B votes from each decision tree.

$$\text{(train)} \qquad f_b = \arg\min_{f_b} L(f_b(X_b), Y_b) \qquad (b = 1, ..., B) \tag{18}$$

$$\text{(predict)} \qquad \hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(X) \tag{19}$$

All record is split into 80% training set and 20% testing set. After training the model, Classification metric on test data is shown in table 8. It seems that classification score is generally even among each categories and accuracy(0.68) is high, meaning the model could successfully grasp the relation between word attributes and its difficulty.

	precision	recall	f1-score	support number
easy	0.63	0.70	0.67	27
medium	0.77	0.68	0.72	34
hard	0.58	0.64	0.61	11
accuracy			0.68	72
weighted average	0.69	0.69	0.68	72

Table 8: Classification metric.
Beside total accuracy, precision, recall and f1-score is calculated on each categories and averaged by weight.

Our model also provides insight into how each feature affects puzzle difficulty. Figure 10 shows the correlation of each feature with word difficulty. This factor is obtained by assigning values to each difficulty level (easy=0, medium=1, hard=2) and calculating the covariance matrix with each feature.

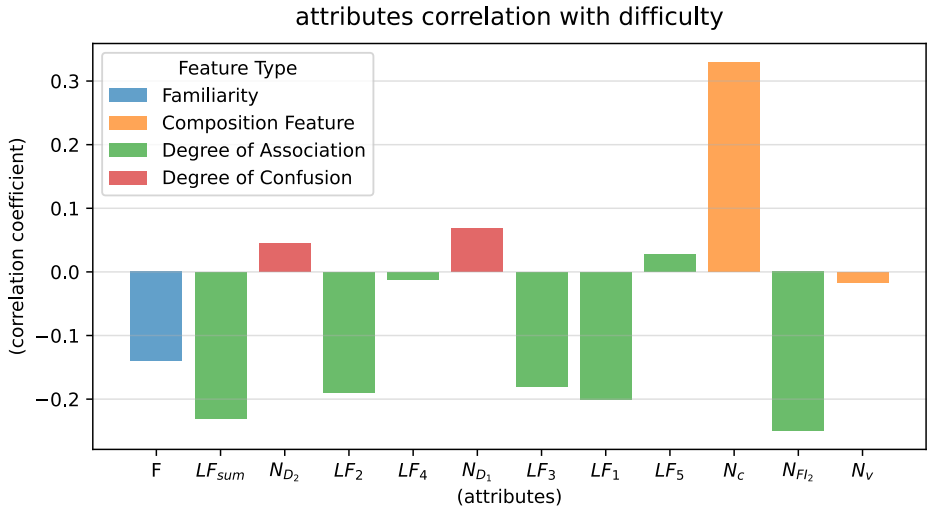

Figure 10: Each feature's correlation with difficulty, ranked by its importance in model

The pattern is very obvious: if a word is used more often (familiarity) or can be easily associated with another word, it would have a negative correlation with difficulty, i.e. it is easier to guess. Furthermore, if it is often confused with another word, it becomes more difficult. The effect of compositional features can also be explained: It is relatively less instinctive to try words with multiple identical letters, and human players prefer to solve Wordle by cracking the vowel first, which makes the word with multiple vowels easier. Therefore, the number of same letters is positively correlated with difficulty, and the number of vowels is negatively correlated with difficulty.

Based on our model, **"EERIE"** is classified as a hard answer. Precision of this category is 0.58 on testing set.

7.3 Sensitivity Analysis

Sensitivity analysis determines how the variations of an independent variable affect a particular dependent result under a given set of assumptions, in which way we can test the robustness of the results. To test whether our models' results are sensitive to the change of the input parameters, we need to carry out sensitivity analysis. In the actual data calculation, the percentages in data may not sum to 100% due to rounding. In order to analyze the stability of our clustering model when the original data is inaccurate, we use sensitivity analysis to evaluate the model. To simulate the loss of precision caused by rounding, we add a stochastic disturbance term ε_i to the data. We assume the stochastic disturbance fluctuates randomly in the range of 5% of the original data. We define the data after adding the stochastic disturbance as:

$$p'_{ij} = p_{ij} \times \varepsilon_i \qquad \varepsilon_i \sim U(0.95, 1.05) \tag{20}$$

Then, we use the new data for clustering. By comparing the clustering results with and without stochastic disturbance, we find that 3.34% of the clustering results changed, which indicates that the clustering results were not seriously affected by data errors and our model is relatively stable.

	Easy	Medium	Hard
without stochastic disturbance	133	153	73
with stochastic disturbance	132	159	68

Table 9: Clustering result with or without disturbance

8 Task 5: Other Features

8.1 Fluctuations in the Number of Reported Results

Similar to the change in the percentage of scores reported that were played in Hard Mode, the number of reported results always fluctuates up and down around the overall trend. Since people are always more willing to share out their better scores, we speculate that this fluctuation may be related to the difficulty of the word. To explore the relationship between the two, we subtracted the results fitted by the T2PL Model in part 4 from the real number of reported results and used it as the dependent variable, and used the difficulty level derived from clustering in part 6 as the independent variable to establish a linear regression model.

The results are as follows:

Easy	Medium	Hard	R^2
6589.5^{***}	1087.5	-2649.3^{***}	0.09
(1143.7)	(1371.6)	(781.3)	

Table 10: The effect of word difficulty on the number of reported results
Note: Because difficulty level is a categorical variable, dummy variables are set for it in the regression.

The regression results show that the difficulty of the word can explain the change of the number of reported results to some extent. Specifically, the number of reported results will be 6589.5 higher than the fitted results when the word is easy, 1087.5 higher than the fitted results when the word is moderately difficult, and 2649.3 less than the fitted results when the word is hard. We corrected the interval predicted by the T2PL Model. Because the word EERIE is hard, so we subtracted 2649 from the endpoint value of the original interval at the same time to obtain the corrected prediction interval [15111, 18327].

8.2 Effect of Word Difficulty on Hard Mode Reports Percentage

Combining the findings from Part 5 and Part 7, we can find that the word attributes that affect the percentage of scores reported that were played in Hard Mode happen to be important attributes that affect the difficulty of the words, so it is reasonable to believe that there is some relationship between the percentage of scores reported that were played in Hard Mode and the difficulty of the words. Similar to 7.1, we explored the relationship by creating a linear regression equation (adding P_H on the previous day as a control variable, and the same for each of the following regressions).

The results are as follows:

Easy	Hard
-0.0021938^{***}	0.0021471^{***}
(0.0005296)	(0.0006389)

Table 11: The effect of word difficulty on the percentage of reported Hard Mode results.
Note: This section replaces dummy variables of moderate difficulty with constant terms, so the coefficients of median are no longer reported.

Surprisingly, the percentage of scores reported that were played in Hard Mode increases when the word is difficult, while decreasing when the word is easy. To investigate the reason for this phenomenon, we ran another two regressions with the number of scores reported that were played in Normal Mode (N-H) and the number of scores reported that were played in Hard Mode as dependent variables.

The results are as follows:

	Easy	Hard
$N-H$	6374^{***}	-5125^{***}
	(1430)	(1718)
H	166.8	229.0
	(90.27)	(158.5)

Table 12: The effect of word difficulty on $N-H$ and H

It can be found that the number of scores reported that were played in Normal Mode decreases significantly when the word is difficult, and on the contrary, increases significantly when the word is easy. While the word difficulty has no significant effect on the number of scores reported that were played in Hard Mode. Therefore, we can analyze the reason for this strange phenomenon: players who choose Hard Mode may have a higher level of word guessing and enjoy the game more, so they are able to maintain a more stable level of guessing the correct word even when the word difficulty increases. Furthermore, they are eager to share their game results even if they cannot guess the word. But players in Normal Mode may not share their game results when they cannot guess the word. As a result, when the word is difficult, the percentage of scores reported that were played in Hard Mode increases.

9 Strengths and Weaknesses

9.1 Strengths

1. **The selection of evaluation indicators is scientific and easy to understand.** We have taken into account the game characteristics of Wordle and constructed four categories of indicators that can comprehensively reflect the spelling structure of words and the ease with which people associate it.

2. **The model is constructed with full consideration of the realistic situation and has good interpretability.** Considering the faithfulness of players to Wordle, the T2PL Model is improved

on the basis of SIR Model by dividing players into two categories: general players and loyal players. The results prove that this improvement enables the model to simulate and explain the changing trend of the number of reported results very well. The P&S Model, based on the simulation of players' strategies of guessing words, also sets the sharing probability for players with different numbers of tries, which well simulates the psychological characteristics of people.

3. **The consideration is very comprehensive.** When exploring the trends in the number of reported results and the percentage of scores reported that were played in Hard Mode, we analyzed not only their overall trends, but also their fluctuations around the overall trends. Furthermore, we explained the reasons for the formation of the fluctuations. Based on the results of this analysis, we revised the prediction interval of the first question.

4. **The models are interlinked and corroborate each other.** The indicators we constructed to reflect word attributes were applied to both the analysis in Part 5 and Part 7. The distribution of EERIE predicted by the P&S Model was clustered for difficulty, and the result was consistent with that of the classifier in Part 7.

5. **Overfitting problem is avoided.** Considering that the amount of data is not very large, we did not use a model containing a large number of parameters to avoid the overfitting problem.

9.2 Weaknesses

1. **There is still a gap between the model and the reality.** Different people will use different strategies to guess word, and different people have different vocabularies, but we did not sufficiently consider in the P&S Model.

2. **The accuracy of the classification model is not high enough.** The accuracy of the classifier in Part 7 is only 68%, which is not particularly good, probably because we omitted some characteristics of the words when constructing the indicators.

9.3 Further Discussion

9.3.1 Model Improvement

Considering that different players have different strategies and vocabularies, we can use different algorithms to simulate different players' strategies, use different lexicons to simulate different players' vocabularies, and finally use the optimization algorithm to calculate the proportion of each type of player in the overall players.

9.3.2 Model Extension

The T2PL Model we constructed can well simulate the popularity of Wordle and explains the changes in the structure of the players. By modifying some parameters, this model can be applied to other games.

References

[1] Benton J Anderson and Jesse G Meyer. Finding the optimal human strategy for wordle using maximum correct letter probabilities and reinforcement learning. *arXiv preprint arXiv:2202.00557*, 2022.

[2] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[3] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[4] Peter G Jensen, Kim G Larsen, and Marius Mikučionis. Playing wordle with uppaal stratego. In *A Journey from Process Algebra via Timed Automata to Model Learning: Essays Dedicated to Frits Vaandrager on the Occasion of His 60th Birthday*, pages 283–305. Springer, 2022.

[5] Xian Jiang, Jing Hua, and Yimin Li. Dynamic research on game communication. *Mathematic in Practice and Theory*, (277-283), 2016.

[6] Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, 24(12):1650–1654, 2002.

[7] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[8] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[9] David Smith, Lang Moore, et al. The sir model for spread of disease-the differential equation model. *Convergence*, 2004.

[10] Daniel Victor. Wordle is a love story. *The New York Times*, 2022.

[11] Wikipedia contributors. Wordle Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Wordle&oldid=1139735175`, 2023.

LETTER

Dear Sir/Madam:

Hello, we are a group of Wordle fanatics. We wake up every day to new Wordle puzzles on your app. The game has certainly brought a lot of joy to our daily lives, and we want to do something to improve the game in return. So we are building a mathematical model of all Twitter Wordle reports in 2022. We hope our analysis can help your team improve Wordle.

Because the rules of Wordle are fixed, perhaps the most (and only) challenging part of running the game is choosing an appropriate word. But how do we know if a word is appropriate enough? Is it better to be common or rare? And more fundamentally, how hard should the puzzle be? To answer these questions, we need to consider the complex relationship between four important factors: word features, distribution of guessing times, puzzle difficulty, and number of players. We could build models to explore them and the relation between.

First, we extract 12 indicators from each word, measuring its familiarity (how often it is used), degree of association, degree of confusion, and word composition features, including all dimensions of how different words influence guessing.

Next, we modelled the general trend of total reports with a dynamic system called **Target-two-Players-Lost (T2PL)**, dividing players into general and loyal two categories. This model can predict future report numbers (for example, reports on 1 March are expected to fall between [17760, 20976]). It can also explain player structues, shows that most Wordle players now are loyal players.

Then we try to predict the solution distribution using simulation algorithms. To improve the result, the distribution is rescaled by considering how likely people would share their result if they received a different score, hence it is called **P&S (Played and Shared) Model**. For any word (e.g. "eerie") the distribution is expected to be (0, 0, 9%, 29%, 45%, 14%, 3%) according to P&S Model.

The difficulty of each puzzle needs to be defined. We run **k-means** on the report distributions and categorize them into 3 types, labelled easy, medium and hard. Then we classify word attributes into three difficulty levels. We know the difficulty of each word (the example "eerie" used above is of course labelled hard) and we know how indicators affect word difficulty from correlation analysis.

The last piece of the puzzle is the relationship between word difficulty and number of reports. Using linear regression, we found that easier Wordle puzzles encourage more shared tweets. We also discovered that word attribute could influence the percentage of Hard Mode players. Although this is a little counterintuitive, we ultimately explain this by saying that Normal Mode players are less likely to share when the word is difficult, while Hard Mode players are not.

For your team, we wonder if our model can be used to predict player response before each puzzle is released and estimate the number of future players. Also, a difficult word might not be an appropriate choice as it might discourage intermediate players from sharing their results. We recommend words with only one or two indicators that make it difficult, so it doesn't fall into the hard category. If the chosen indicators that make words difficult could vary from day to day (e.g. one day with less word frequency, another day with less vowels), it might even give daily Wordle players a sense of freshness.

We would be happy to discuss the game further with your team. We look forward to playing great Wordle puzzles in the future.

Yours sincerely,
Team #2318036