Problem Chosen	2023 MCM Summary Sheet	Team Control Number
C		2320719

A Study on Wordle-based Analysis

Summary

Wordle is a globally popular puzzle game. It is important to analyze how to attract more potential players and forecast the game's future trends. We used cluster analysis to divide the number of wordle players into different period from January 7, 2022 to December 31, 2022, namely, rising period, declining period and stable period.

To forecast the overall player count, we divided the influence factors into game trends and other factors. We obtained a 90% confidence interval and 50% confidence interval for the number of players. For example, we predict that the number of players on March 1, 2023, would be 13,671 to 22,657 in 90% confidence level and 16,737 to 19,591 in 50% confidence level.

We also focused on the distribution of the number of pass, which is influenced by dates and word attributes. To analyze the impact of word attributes on the distribution of Hard Mode, we selected a stable zone with a relatively stable proportion of Hard Mode. Thus we can transformed the problem into the impact of word attributes on the distribution. We found that the following attributes are key influencing factors: whether a word contains repeated letters, whether two identical letters exist,whether the first letter is a vowel, whether the ending is "y/ly" or "e," whether "sh/ch/th/ph/gh" exist, and whether three identical letters exist.

Using fuzzy recognition, we obtained difficulty scores and classifications based on word attributes. On the other hand, we found that the impact of dates on distribution is insignificant. And we can get a combination of the two factors can provide distribution for specific words and dates.

Finally, we provided recommendations for the game's future development, suggesting that the game should introduce more interesting topics that can bring people a joyful mood, such as words with three letters or word "lucky."

Keywords: cluster analysis, fuzzy recognition, confidence interval;

Contents

1 Introduction

1.1 Problem Background

Wordle is a word-guessing game that has recently gained popularity in The New York Times. As it has been widely spread around the world. It has been adapted into various languages and we are studying the classic version of The New York Times. Wordle updates only once a day, and players from all over the world are guessing the same word. Each player has up to six attempts per day to guess a five-letter word using a complete word for each attempt. After each guess, players receive feedback on whether the letters are correct and in the correct position (green), correct but in the wrong position (yellow), or incorrect (gray). In addition, Wordle also has a Hard Mode that requires players to continue using any correct letters they have guessed in subsequent guesses.

As a popular game, many players share their game results on Twitter every day. In addition to studying the solving strategies, it is also very interesting to analyze many features of this game from the perspective of the designer. We have obtained the results data mined by MCM officials on Twitter from January 7, 2022, to December 31, 2022. In the following sections, we will study and analyze these data.

1.2 Restatement of the Probelms

1. (1) Build a model to explain the daily changes in the number of reports.

(2) Use the model to construct a prediction interval for the number of reports on March 1, 2023.

(3) Determine if there is any word attribute that affects the percentage of scores reported that were played in Hard Mode.

2. (1) Construct a model to predict the distribution of results for a given word and date.

(2) Analyze the uncertainty of the model and its prediction.

(3) Use the word "EERIE" on March 1 as an specific example to provide a prediction and analyze its credibility.

3. (1) Develop a model that can classify words by difficulty.

(2) Determine the correlation between each classification and the given word attributes.

(3) Use the model to determine the difficulty of the word "EERIE" and analyze its accuracy.

4. List and describe other interesting features in the data set.

5. Summarize the results of the article in a letter to the Puzzle Editor of The New York Times.

1.3 Our Work

The problem requires establishing an evaluation model for predicting user numbers and assessing game difficulty to make the game more suitable for most people and increase the number of participants. Our work mainly includes:

1. Establishing a product lifecycle model for the game based on past participation data to predict future participation numbers.
2. Listing different word attributes and building a clustering model to determine whether word attributes affect the distribution of number of pass in difficult mode.
3. Establishing word difficulty and date feature models to predict the distribution of number of attempt for a given word and date.
4. Using our model, strategy, and results to write a recommendation letter to the company.

To avoid complex data descriptions and reflect the workflow intuitively, we have established a flowchart.

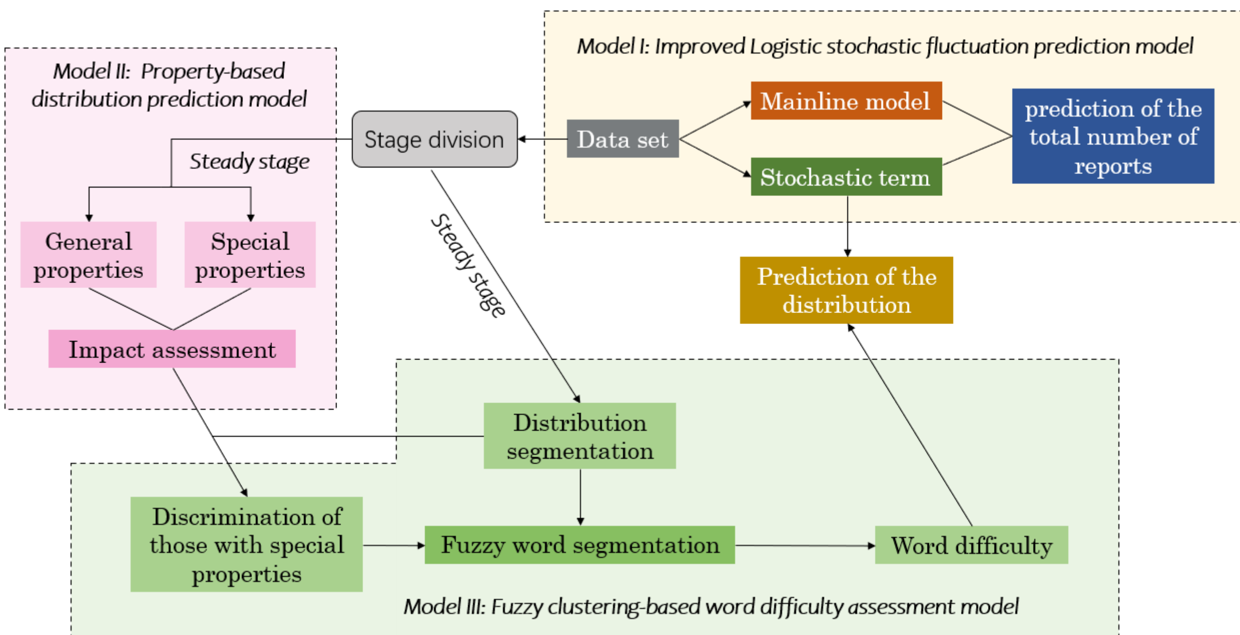


Figure 1: Flow Chart of Our Work

2 Assumptions and Justifications

Assumption 1: The Twitter sharing data can accurately represent the overall situation of number of all game participants.

Justification: In this study, we can only use the given dataset, and we cannot access the backend data of the game itself. In addition, we cannot accurately characterize the number of people who shared this result on Twitter, as well as the distribution ratio of the results, which may differ from the actual situation. Therefore, using this data to represent the overall data is currently the best solution.

Assumption 2: Our belief is that the overall passing frequency from 1 to 6 and the final fail represent a gradual increase in difficulty from low to high.

Justification: In the context of the game, the number of pass attempts represents the level of difficulty in completing the game, which in turn reflects the level of difficulty of a given word under the rules of this game.

3 Notations

The primary notations used in this paper are listed in Table 1.

Table 1: Notations

Symbol	Definition
$N(t)$	the number of results on the day t
$f(t)$	the trend of the game's lifecycle
$\alpha(t)$	interference factors
$S_{increase}$	the number of increaseing results
$S_{decrease}$	the number of decreaseing results
ω_i	modification factors for differnt days of the week
S	the score of the game participants' enjoyment
$\rho(X, Y)$	the degree match between matrix X and matrix Y
M_k	the difficulty level matrix
R	recognition matrix
$G(t)$	the total score of the day
L	one-dimensional distance
R_2	one-dimensional distance

4 Data Preprocessing

Before building the model, we cleaned the overall data. We found three words with certain issues: "tash", "clen", and "rprobe". To address this, we logged into Twitter to check the actual word usage on that day and corrected the words to "trash", "clean", and "probe" respectively. In addition, we noticed that some words had "i" represented as "ï", which caused some interference in our subsequent word filtering and comparison. Therefore, we standardized all the interfered "ï" into a uniform "i".

As the title suggests, the data on the score distribution for each group was rounded after a large sample size was taken from Twitter, so we first normalized these data. Considering that the errors between these data were caused by rounding, the probability of being rounded in the (1, 2, 3, 4, 5, 6, X) scenarios was the same. Hence, we did not adopt the traditional method of directly amplifying to 100 according to the proportion, but evenly distributed the difference between each group's total and one hundred among each type of answering situation using the mean imputation. This made our subsequent research more convenient and accurate.

These data preprocessing steps were the analysis basis of all subsequent models We will apply these changes to all of them.

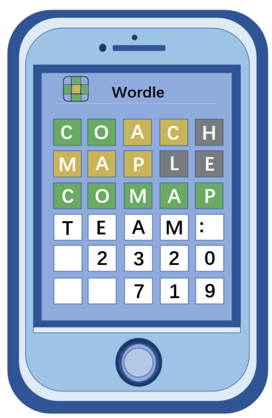


Figure 2: Diagram of Wordle Game

5 Model I: Model for Predicting Total Report Count

In Model 1, we establish a relationship model between date and game player numbers. We define the overall player number model as follows:

$$N(t) = N(f(t), \alpha(t)) \tag{1}$$

where:

$N(t)$ represents the number of participants in the game on day t.

$f(t)$ represents the trend of the game's lifecycle.

$\alpha(t)$ represents the interference factors.

5.1 Game Life Cycle Model

According to the product life cycle curve (see Figure 2),

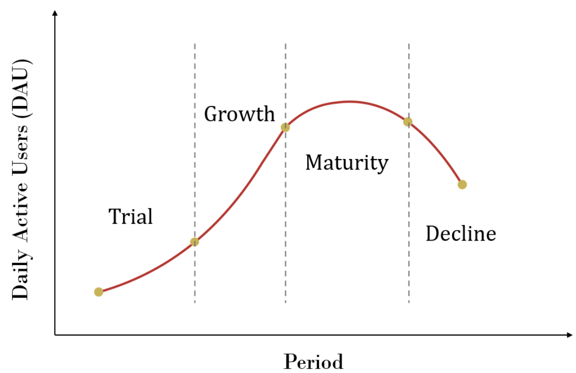


Figure 3: Product Life Cycle Curve

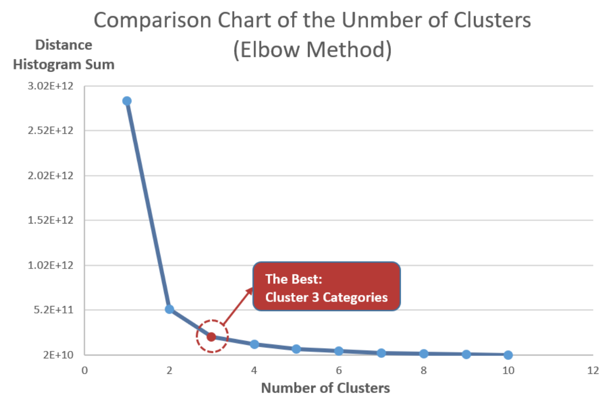


Figure 4: Elbow Method

We establish a model for the changes in time and number of participants in this problem. It is not difficult to find that both have the same characteristics. We consider using factors such as the number of participants, the growth rate of the number of participants, and the proportion of people with difficulty choosing a mode to conduct cluster analysis. By using the elbow method (see Figure 3) to obtain the number of clusters, we choose to cluster them into three categories.

We establish a relationship graph between the number of participants in the game and time, and it is not difficult to find that the clustering is basically consistent.

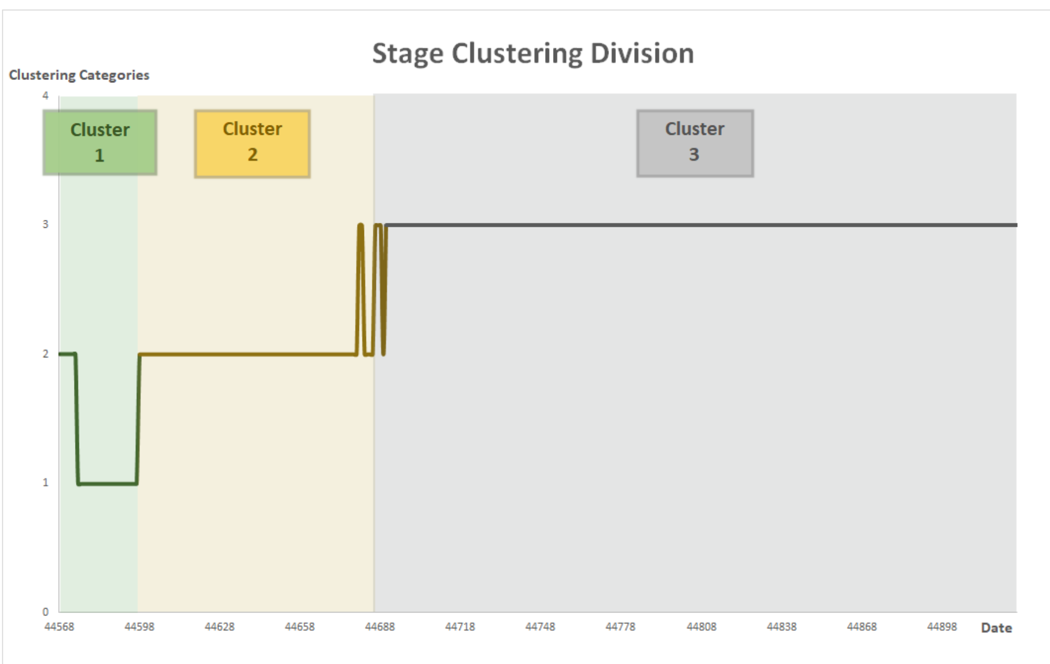


Figure 5: Stage Clustering Division

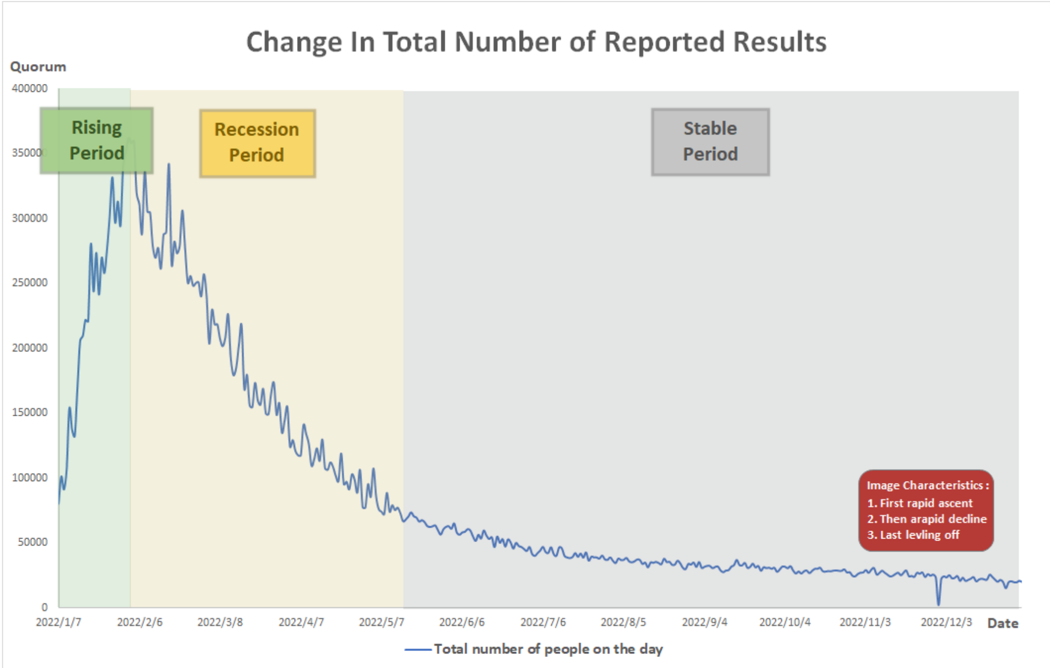


Figure 6: Change In Total Number of Reported Results

Therefore, it can be divided into the rising period, the selection retreat period, and the stable period based on the life cycle stage.

Thus, We can represent the game lifecycle model as follows:

$$f(t) = N(0) + \int S_{increase} + S_{decrease} \tag{2}$$

where

$$S_{increase} = ke^{rt} \cdot \delta t \tag{3}$$

$$S_{decrease} = \frac{1}{(\ln(t) - 1)t} \tag{4}$$

It is clear that the number of players exhibits a significant and stable decline during the course of the game, so we can assume that the growth in the number of players during the latter half of the decline should not be counted. Additionally, during the initial phase, the decrease in the number of players is negative, so it should not be taken into account.

5.2 Correction of the Game Lifecycle Model

Simply considering the lifecycle of the game is not enough. Factors such as the impact of changes in difficulty due to short-term challenges, and differences in user numbers between weekdays and weekends should be taken into account, which is where the significance of the random term lies.

5.2.1 Correction for Periodic Fluctuations

To minimize the effect of the mainline on data processing, we only analyze steady-state data. We calculate the average number of active users from Sunday to Monday on a weekly basis, and obtain the correction factor:

$$\omega_i = \frac{\overline{P_i}}{\overline{P}} \tag{5}$$

Here, $\overline{P_i}$ represents the average daily activity on a certain day of the week, and $\overline{P}$ represents the average daily activity.

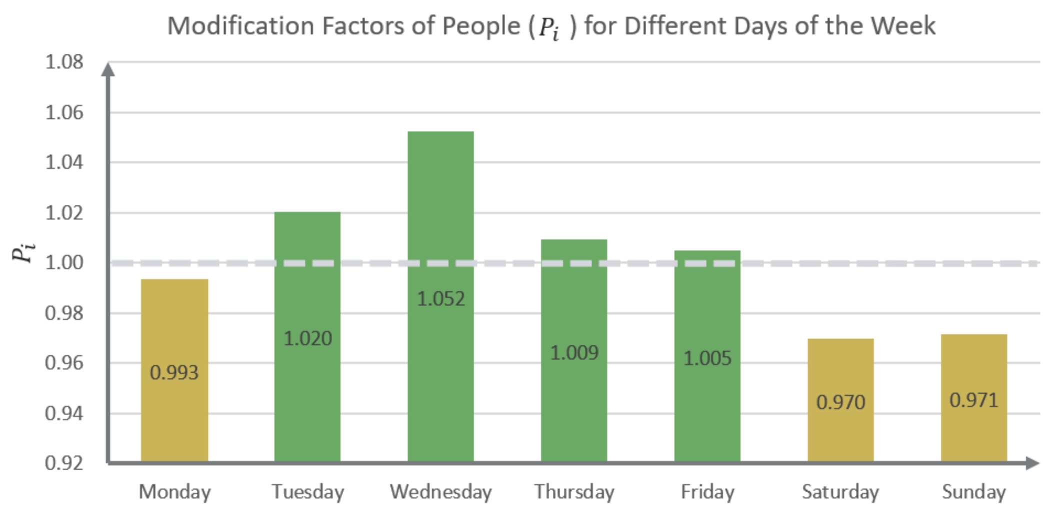


Figure 7: Modification Factors of People for Different Days of the Week

5.2.2 Interval Estimation Based on Immersion Interval Theory

In 1975, Csikszentmihalyi first proposed the theory of immersion, which suggests that anxiety and skill are the main factors that affect user immersion. In other words, if the challenge is too high, the user will lack control over the environment, leading to anxiety or frustration. Conversely, if the challenge is too low, the user will find it boring and lose interest. The state of immersion mainly occurs when the two are balanced.

Today, this theory is widely used in the field of game planning. We use the number of passes to represent the difficulty that the player feels, and use enjoyment as an indicator to describe the pleasure that the player feels when playing the game on a given day. Since the player's passes are distributed from 1 to 7, we calculate the weighted average of different enjoyment levels.

The player's enjoyment is related to the number of times they pass the game and their skills. The first and second passes are more based on luck and lack of logical thinking. The third to fifth passes require more logical thinking, and players can experience the joy and sense of achievement in the game. The sixth pass may lead to frustration if the game is solved through random playing or reaching the maximum number of attempts.

If the player cannot solve the game on the seventh attempt, it may indicate that the player cannot adapt to the game's operations or has not understood the correct way to play, but their enjoyment may not be zero due to some players' competitiveness.

Therefore, we have the following weighted coefficient division:

Table 2: Player Enjoyment

distribution	1	2	3	4	5	6	7
score	3	5	6	8	10	4	2

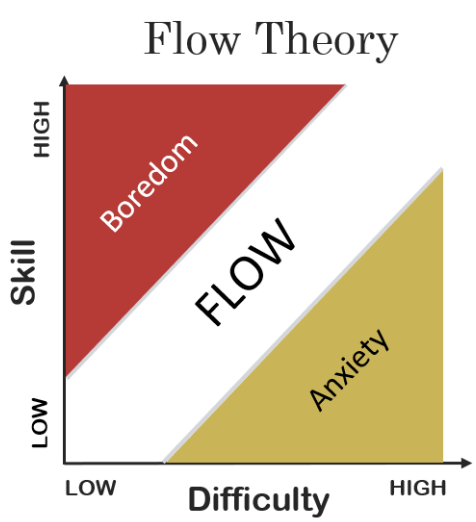


Figure 8: Flow Theory

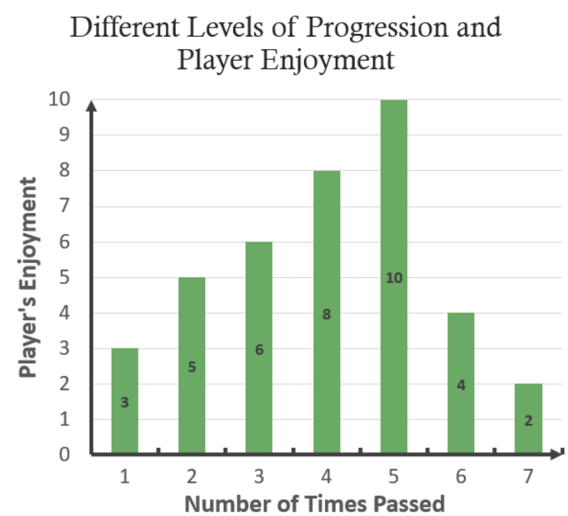


Figure 9: Different Levels of Progression and Player Enjoyment

In addition, the impact on game enjoyment is not only determined by the experience of the day before, but also affected by the Ebbinghaus forgetting curve in the previous period. Although the influence diminishes as the time interval between the experience and the current day increases. The player's memory of the game experience in the previous three days exhibits a decay phenomenon. This led to the determination of the evaluation coefficient s and the normalization coefficient S using the min-max principle.

$$\begin{cases} s = \dfrac{100k}{(\lg d)^c + k} \\ S = \dfrac{s - \min(s)}{\max(s) - \min(s)} \end{cases} \tag{6}$$

In addition, the impact on game enjoyment is not only determined by the experience of the day before, but also affected by the Ebbinghaus forgetting curve in the previous period. Although the influence diminishes as the time interval between the experience and the current day increases, the player's memory of the game experience in the previous three days exhibits a decay phenomenon. This led to the determination of the evaluation coefficient s and the normalization coefficient S using the min-max principle.

Subsequently, we defined the increment for the fourth day, $\delta N = N(t) - N(t-1)$. Obviously, the lower the comprehensive score, the smaller the increment should be (tending towards negative); the higher the comprehensive score, the larger the increment should be (tending towards positive). Since there exist significant random interference factors in the daily changes in the number of players, the plot of the comprehensive score and the data for the fourth day can only provide normal points that satisfy most of the points in the first and third quadrants (about 70%), and white noise points in the second and fourth quadrants (about 30%). The white noise points are caused by random interference factors. The increment result obtained by our calculation is shown below.

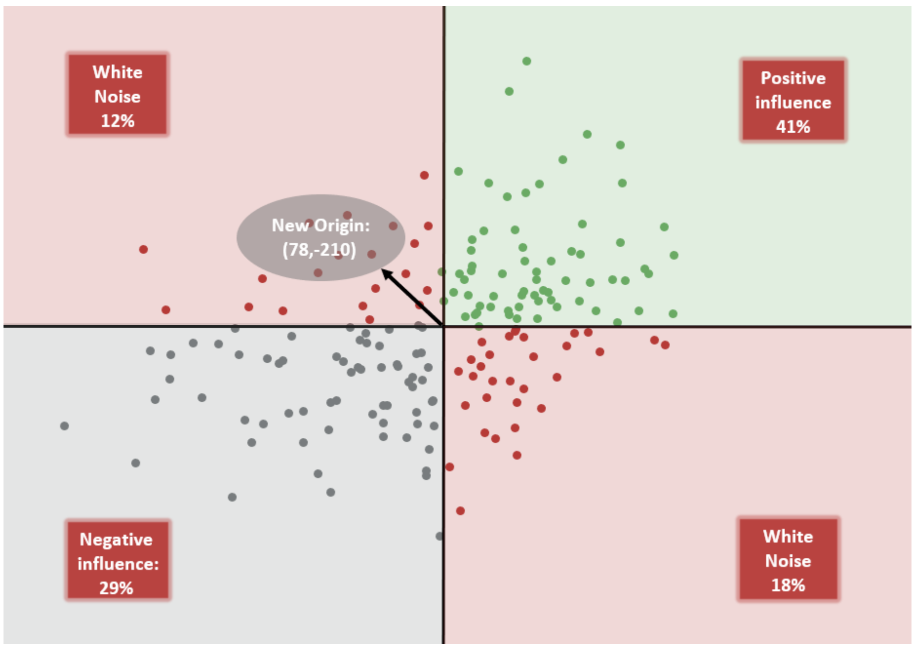


Figure 10: Relationship Between Overall Score and Growth Rate

As we are unable to define the game difficulty coefficient after the steady state, it is the most likely part of our prediction to cause errors. Thus, we set it as the interval prediction part, and take the 90% confidence level interval and 50% confidence level as the predicted boundary $[M_{min}, M_{max}]$.

5.3 Revised Player Number Prediction

In summary, we can express the overall prediction function as follows:

$$N(t) = (f(t) + [M_{min}, M_{max}]) \cdot \omega_i$$

6 Model II: Model of Answering Results Distribution in Hard Mode

Considering the factors that affect the distribution of the number of times questions are answered, the factors that exert influence are the distribution of the number of times questions are answered in hard mode and normal mode, both of which are determined by the words.

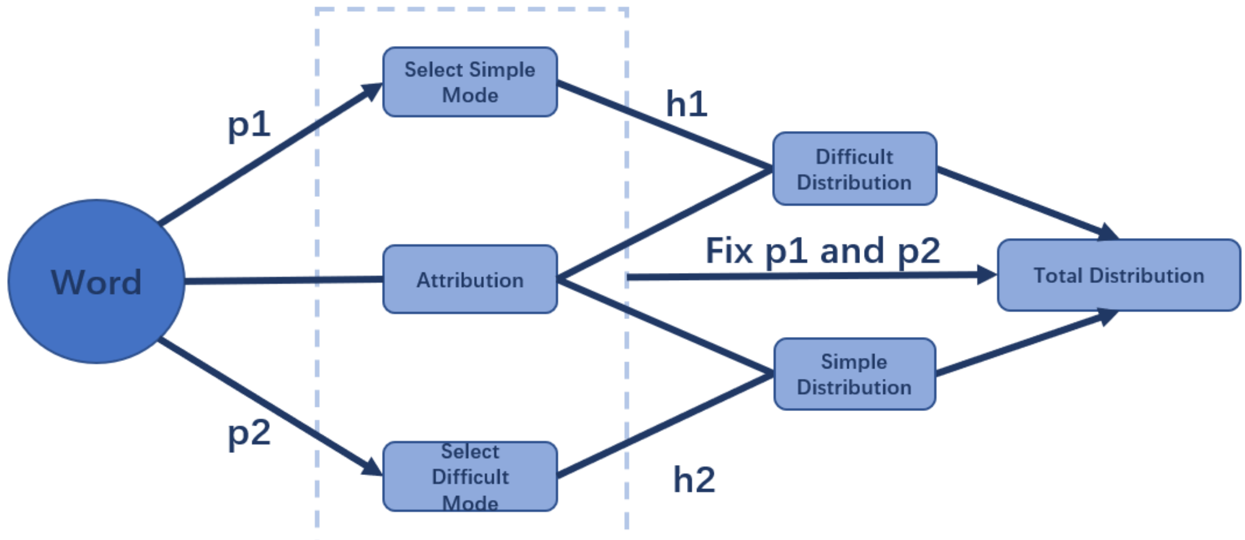


Figure 11: Mind Map for Solving Hard Mode Distribution

Let us assume that:

$$\begin{aligned} g_1 &: \text{ word } \mapsto \text{ distribution}_{\text{hard}} \\ g_2 &: \text{word } \mapsto \text{ distribution}_{\text{normal}} \\ h_1 &: \text{distribution}_{\text{hard}} \mapsto \text{ distribution} \\ h_2 &: \text{ distribution}_{\text{normal}} \mapsto \text{ distribution} \end{aligned} \tag{7}$$

It is not difficult to obtain the following equation:

$$\text{distribution} = \text{selecthard} \cdot h_1(g_1(\text{word})) + \text{selectnormal} \cdot h_2(g_2(\text{word})) \tag{8}$$

It can be seen that, during the stable period, the influence caused by the proportion of selecting the difficult mode can be ignored. Therefore, the problem is transformed into the impact of word attributes on the overall distribution of answer times. Each word is classified into two binary categories according to whether it has the attribute, and the average distribution of each answer times with and without the attribute is calculated. If the deviation between the two is within an acceptable range, it can be preliminarily determined to be irrelevant. However, if the deviation is too large, it indicates that the attribute has an impact on the distribution of answer times, and thus has an impact on the distribution of answer times for selecting the difficult mode.

7 Model III: Word Difficulty Assessment Model

Our developed model can predict the distribution of report results given the future words to be solved and future dates. Therefore, we need to analyze from both the perspectives of words and dates.

7.1 Model Construction and Solution

Building a Word Difficulty Assessment Model is the focus of this chapter. Word difficulty is dependent on its attributes, and Word difficulty external manifestation is the pass rate of answering questions, as shown in the figure below.

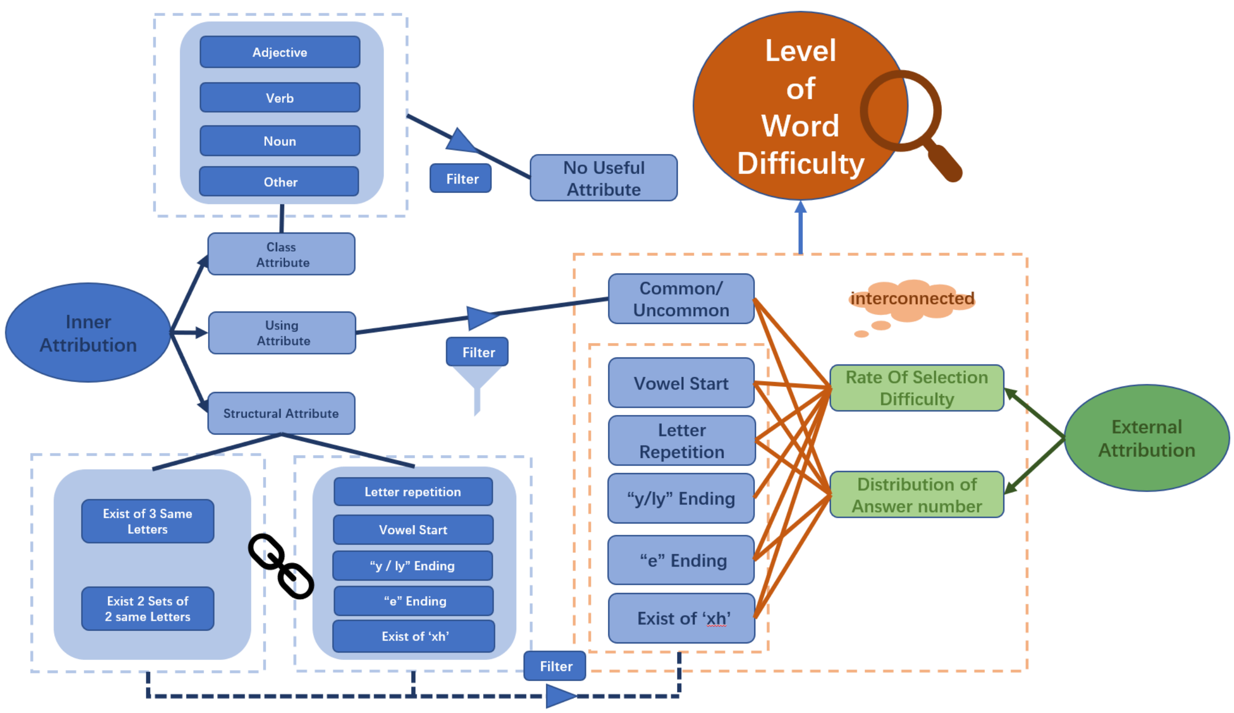


Figure 12: Mind Map for Word Difficulty Assessment Model

Since word difficulty can be directly reflected in the frequency distribution curve, we need to eliminate other factors that affect the frequency distribution before clustering the word difficulty into three difficulty coefficients using seven characteristic values of the frequency distribution function. In this question, not only the difficulty of the word itself affects the frequency distribution, but also the proportion of players choosing to participate in the difficult mode greatly affects the frequency distribution. Therefore, the data we choose to use should avoid the influence of the "difficulty selection ratio" on the frequency distribution as much as possible. By observing the trend chart of the "difficulty selection ratio" over time, we can find that this ratio continues to rise from 2% to 10% in the initial stage and finally fluctuates around 9%. Therefore, we can consider the data where the "difficulty selection ratio" is between 8% and 10% to be unaffected by the "difficulty selection ratio," and finally determine the data that can be used to analyze word difficulty and frequency distribution.

Through the elbow rule, we can find that clustering this data into three categories is the most ideal. Under realistic conditions, these three clusters can represent the difficulty levels of words as "easy," "medium," and "difficult," respectively.

After classification, we should combine all the words in each difficulty level and analyze the attribute characteristics they possess, so that we can use these attribute characteristics to determine the difficulty level of a word. Through analysis, we can use the 12 attribute features selected in the first question to characterize the difficulty level of a word. It was found in the first question analysis that some attribute features did not have a significant effect on the distribution of pass rates. Since the distribution of pass rates can directly reflect the difficulty level of a word in this question, we only need to select the attribute features that have a significant impact on the first question as the basis for dividing word difficulty levels.

We can calculate the probabilities of different feature appearances for words at different difficulty levels (which can also be called membership degrees). Therefore, for each word at a particular difficulty level, different feature attributes have a corresponding membership degree. The larger the membership degree of a certain feature attribute, the greater the probability that the word possesses that feature attribute at the given difficulty level. This allows us to establish the membership degree of different feature attributes for words at different difficulty levels.

Now, if we are given a word, we can determine whether it possesses certain feature attributes. We can represent this information as a multi-dimensional vector W, where a "1" indicates that the feature attribute is present and a "0" indicates that it is not.

Based on research by Shao and Chen on barcode recognition, we established a matching degree model that characterizes the fuzzy matrix Mk and the matrix R to be identified:

$$\rho(M,R) = \bigvee \left(\sum \left(m_{ij} \wedge r_{ij} \right) / \sum \left(m_{ij} \vee r_{ij} \right) \right) \tag{9}$$

In this equation, ρ represents the degree of match between the difficulty level matrix M_k and the recognition matrix R, where $\rho \in [0,1]$. The closer ρ is to 1, the closer the matrices R and M_k are in terms of their characteristics, while a value closer to 0 indicates that the matrices are farther apart. M_k represents the characteristics of words at different difficulty levels, and the matrix M_k has a row for each level with a 7-dimensional vector of information corresponding to that level. The number of rows depends on how many ways there are to characterize the difficulty level, i.e. $M_k = (m_{ij})$. R represents the characteristics of a word to be recognized, and each row has the same 7-dimensional vector of information. The value of $\rho(M_k, R)$ represents the degree of membership of R to M_k and can be interpreted as the maximum value obtained from each value of R for different characterization methods of a certain difficulty level of words that are to be recognized.

7.2 Date Impact Assessment Model

Consider the influence of dates on the distribution, the main factors being overall difficulty trends, weekday and weekend effects.

To analyze the overall difficulty trend, the comprehensive difficulty rating is calculated for each day as follows:

$$G(t) = \sum_{i=1}^{7} i \cdot g_i \tag{10}$$

where $G(t)$ represents the total score for the day, and g_i represents the proportion of people who passed on the ith attempt. The relationship between the comprehensive rating and time is obtained and linear regression is performed to determine the slope and judge the overall difficulty trend.

To analyze the weekday and weekend effects, the mean of the distribution for weekdays and the mean of the distribution for weekends are calculated, and the difference in the mean distribution of the two is tested to determine their effects.

8 Result

8.1 Player count forecast on March 1st

Firstly, using the Kalman filter algorithm to remove interference terms and fitting the model, we obtain the growth function $S_{increase} = 5335e^{0.051x}$ and the decrease function $S_{decrease} = \frac{36182}{(lnt-1)t}$, which can be used to calculate the estimated player count on March 1st by plugging in $t = 420$, giving us $f(420) = 17260$.

Considering the periodic fluctuation correction, we obtained the total number of reports submitted from Monday to Sunday, as shown in the table below.

Table 3: Probability of Different Day

week	1	2	3	4	5
probability	90320.05882	92754.09804	95669.9	91749.33333	91341.69231
week	6	7			
probability	88160.05769	88308.31373			

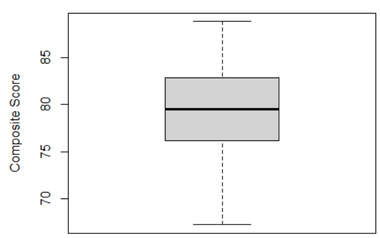
We obtained the correction coefficients, as shown in the table below.

Table 4: Corection Coefficients

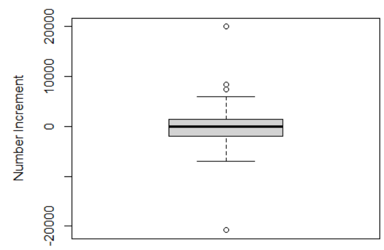
correct factor	1	2	3	4	5
probability	0.993491915	1.020265572	1.052338466	1.00921348	1.004729559
correct factor	6	7			
probability	0.96973259	0.971363359			

Since March 1st falls on a Wednesday, we selected the correction coefficient ω_3.

Based on the interval estimation correction theory, taking into account the impact of weekly cycles, the estimated player count range on March 1st is obtained as follows.



(a) Data's attributues of Abscissa



(b) Data's attributues of Ordinate

Figure 13: Different Boxplots with Confidence Intervals

We used box plot to evaluate the data on March 1st. The outliers of the data were removed and the central tendency of the data was obtained. The interval of the 90% confidence level was [-4493, 4493], and the interval of the 50% confidence level was [-1427, 1427].

After being multiplied by the correlation coefficient w_3, the center of fluctuation of our predicted results on March 1 is 18,164. Within the 90% confidence interval, the predicted value ranges from 13,671 to 22,657, while within the 50% confidence interval, the predicted value ranges from 16,737 to 19,591.

8.2 The Influence of Word Attributes on the Distribution of Difficulty Mode Passes

Based on the general characteristics of words, we divided them into 12 different attributes and calculated the average distribution of passes with and without each attribute. We calculated the distance deviation

$$L = \sum_{i=1}^{7} |a_i - b_i|$$

and the square deviation

$$R_2 = \sum_{i=1}^{7} (a_i - b_i)^2$$

for each attribute and obtained the following results.

Table 5: L and R_2 for Each Attribute

attribution	L	R_2
adjective	2.00	1.20
verb	3.95	2.99
noun	4.34	3.13
other	24.47	142.96
exist 3 same letters	22.68	101.83
Beginning with vowels	5.95	7.07
ending with y/ly	14.80	46.83
ending with e	7.80	13.56
sh/ch/th/gh/ph	16.08	66.09
exist 3 same letters	77.70	1157.30
exist 2 sets of 2 same letters	31.98	269.79
common word	12.40	34.48

It can be seen that verbs, adjectives and nouns, as the three main parts of speech, have nearly the same values and can be used as the standard deviation.

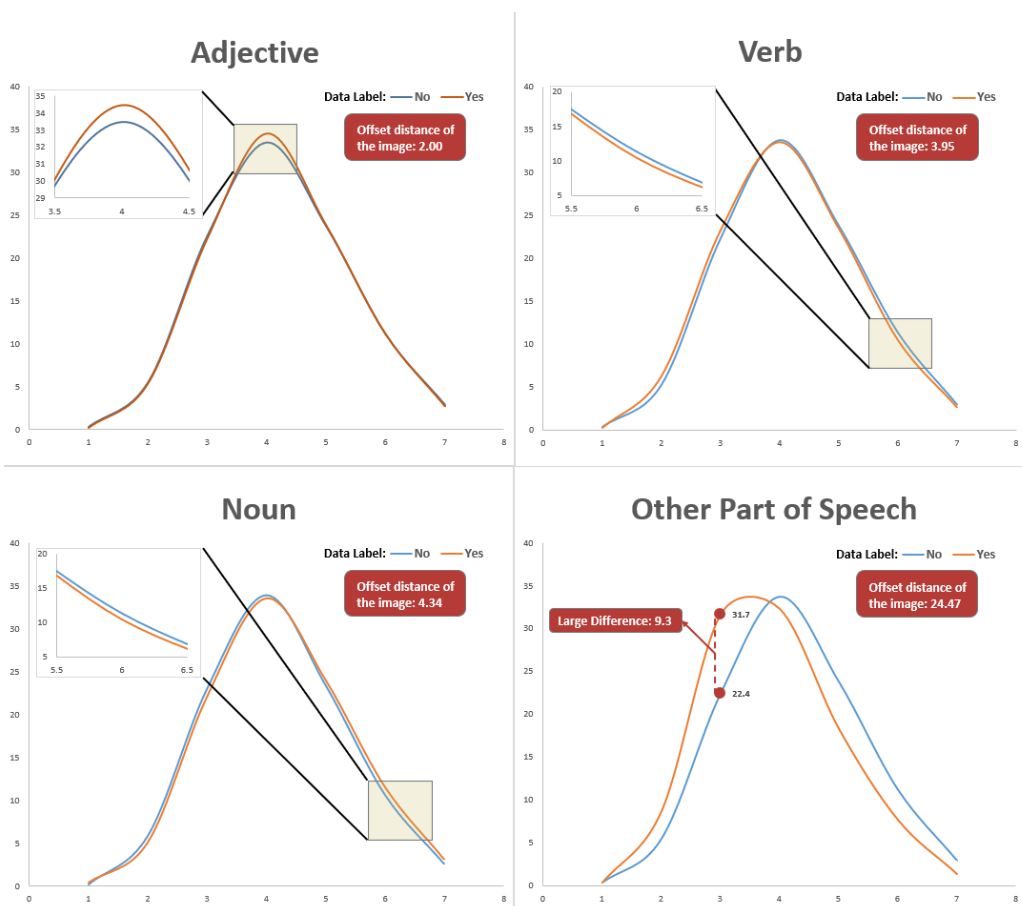
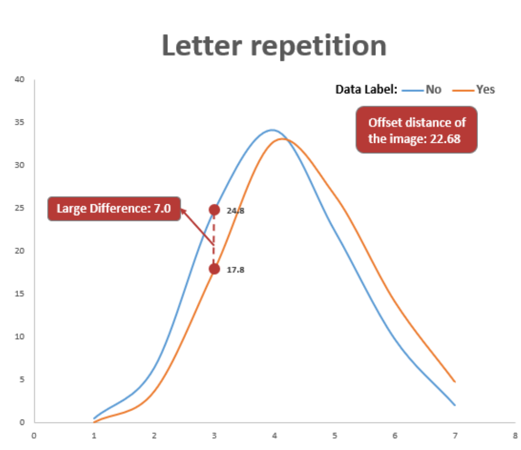
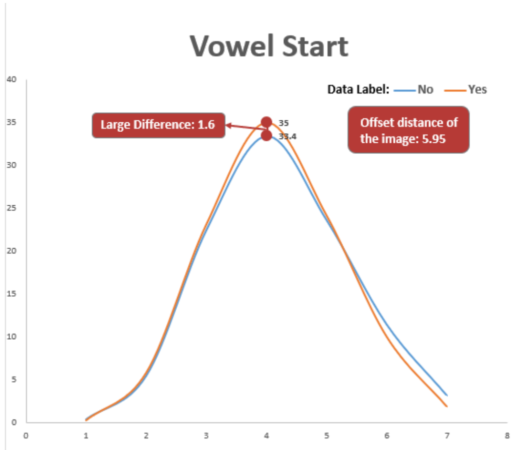


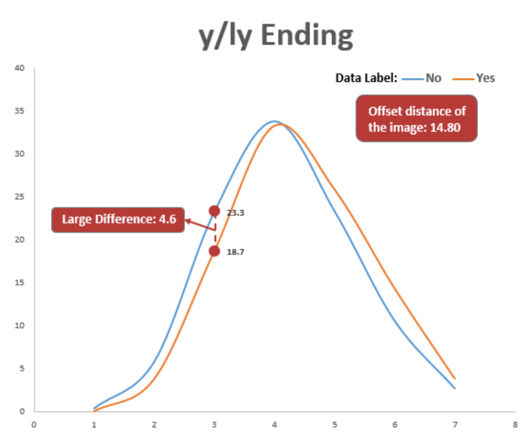
Figure 14: The Distance Deviation of Verbs, Adjectives and Nouns

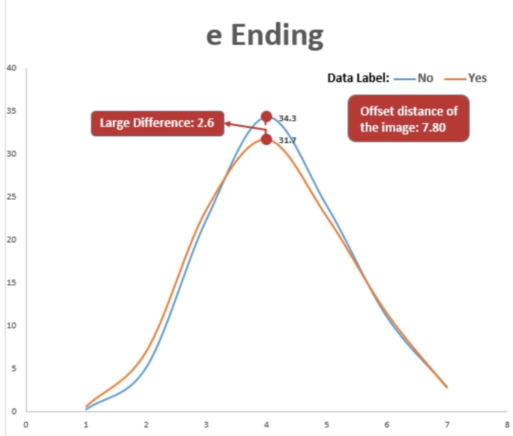
In addition, we get the feature Common Word by comparing the words with the most commonly used words in American English from the book "A Frequency Dictionary of Contemporary American English".

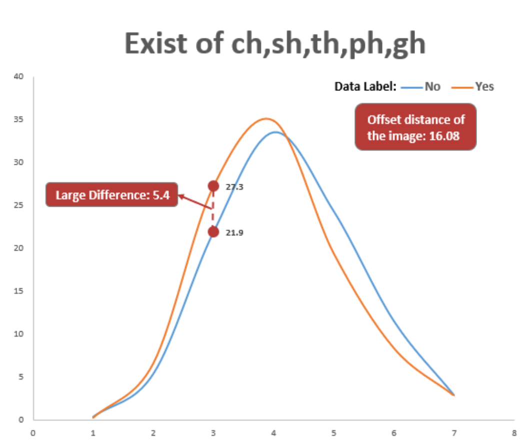
Other attributes have an impact on the distribution.

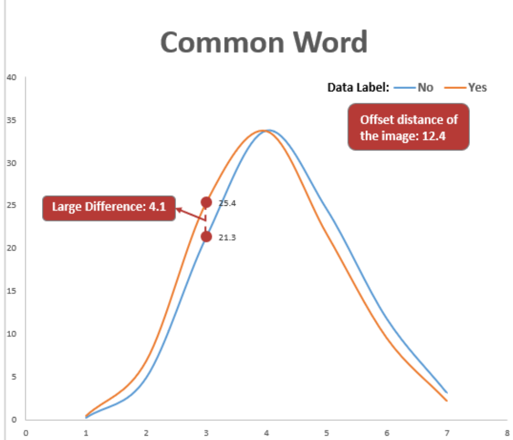












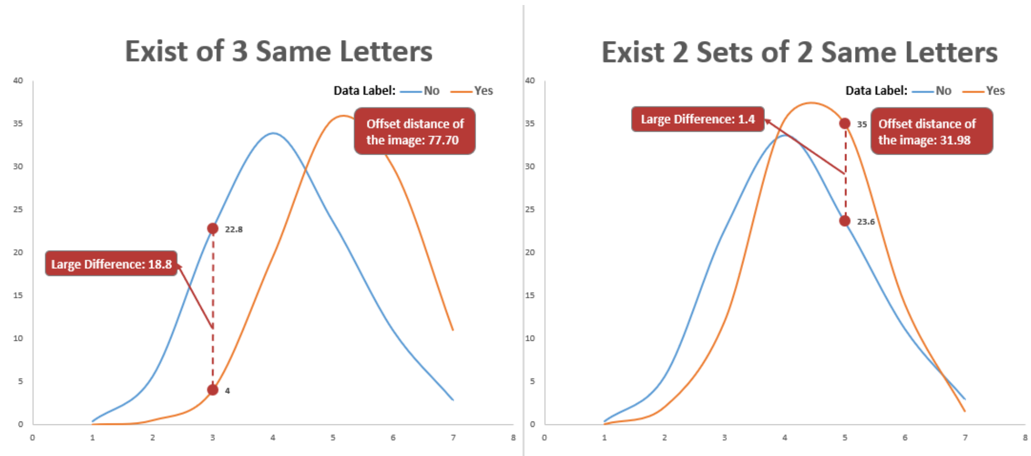


Figure 15: Attributes that Have Impacts on the Distribution

Therefore, we can conclude that whether a word contains repeated letters, whether the first letter is a vowel, whether the ending is "y/ly" or "e," whether "sh/ch/th/ph/gh" exist, and whether three identical letters exist has an impact on the distribution of passes for difficulty selection mode.

8.3 Prediction of Specific Words with Dates

8.3.1 Word Influence

Firstly, we conduct a clustering analysis on words and obtain three categories. We calculate the means of each category's distribution, which can be seen in the figure below.

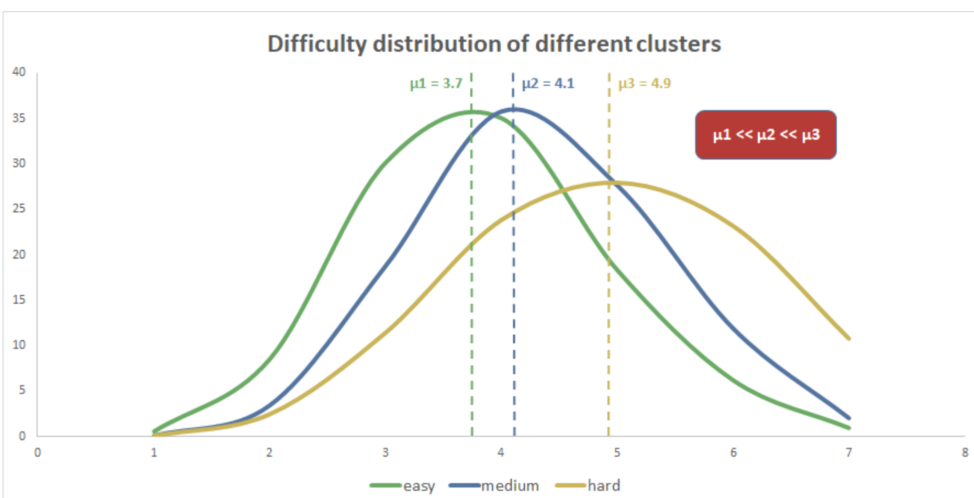


Figure 16: The Distance Deviation of Verbs, Adjectives and Nouns

This means that the clustering effect is good and can reflect the difficulty differences. We name these categories as simple, medium, and difficult, respectively. Through the clustering analysis, we obtain a pattern matrix, namely:

Table 6: Results of Clustering Analysis

	exist 2 same letters	begining with vowels	ending with y/ly	
easy	0.112339355	0.407791211	0.202011133	
midian	0.388333572	0.450139899	0.266022891	
hard	0.499286021	0.142149442	0.532045783	
	ending with e	ch/sh/th/ph/gh	common word	other
easy	0.361326639	0.503085592	0.481386393	0.42
midian	0.244171348	0.198749863	0.259332737	0.39
hard	0.394430639	0.298124795	0.259332737	0.19

Based on the analysis in the previous question, as the distribution of number of pass can directly reflect the difficulty level of the word in this problem. We only need to choose the attribute features that have a significant impact in the first question as the basis for dividing the difficulty level of the word. We can use the following seven attributes to characterize a word:

Table 7: Word Attributes

	attribute
1	Whether there are letter repetitions.
2	Whether it starts with a vowel.
3	Whether it ends with y/ly.
4	Whether it ends with e.
5	Whether there are ch, sh, th, ph, gh.
6	Whether it is a common vocabulary.
7	If none of the above attributes apply, we classify it under "other attributes."

We obtain a pattern matrix for all words.

By calculating the membership degree of words in word difficulty, we obtain the proportions of simple, medium, and difficult categories, which are 36%, 39%, and 25%, respectively, consistent with the clustering factors.

8.3.2 Impact of Dates

Analyzing the overall trend of difficulty changes, we evaluated a comprehensive difficulty score for each day and plotted the score against time. A linear regression was performed, and the slope was found to be -0.03, which can be considered as indicating an unchanged trend in difficulty.

To investigate the influence of weekdays and weekends, we calculated the average difficulty scores for weekdays and weekends, respectively, and examined the difference in the distribution of the two mean values, as shown in the figure. It can be concluded that there is no significant impact of weekdays and weekends on the difficulty.

Overall, the impact of date changes on the pass rate can be considered negligible.

8.4 Word Difficulty Evaluation

By determining the membership degree of each word, we can quantify the degree of difficulty using a weighting system of 2 for "easy", 3 for "medium", and 4 for "difficult" difficulty levels. Therefore, we can obtain the difficulty coefficient of each word.

For the difficulty value confidence level problem, we use the words that have been clustered according to distribution as the test set. Then we use the pattern matrix that represent word attributes to get the word membership of the difficulty's cluster. Compared with the result based on distribution clustering, the following figure is shown

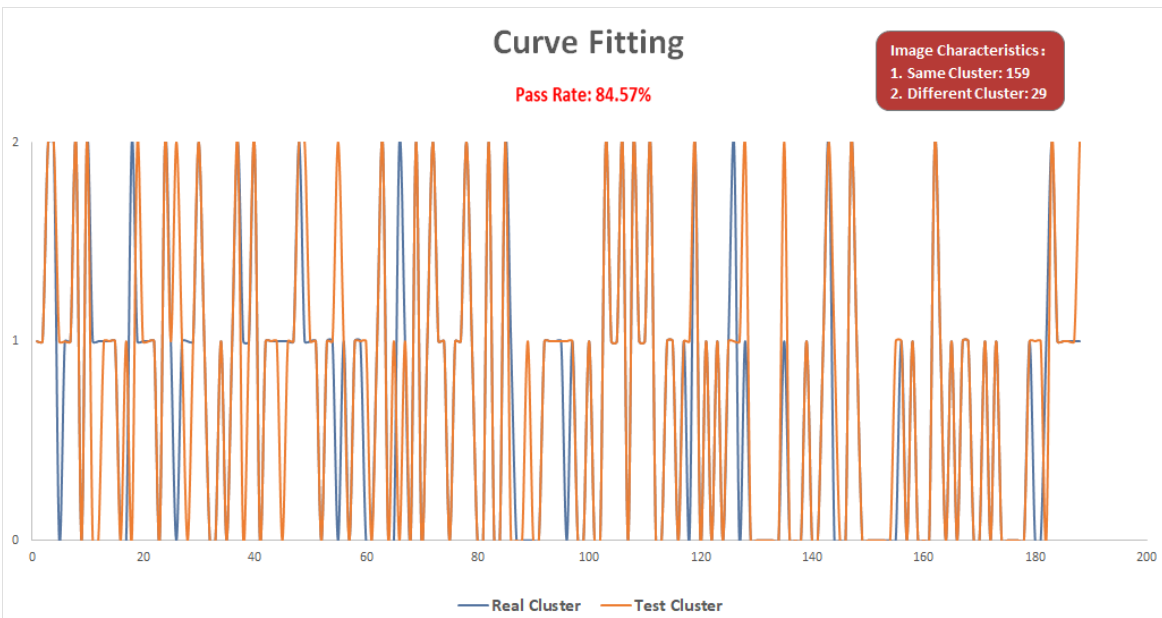


Figure 17: Curve Fitting

It can be seen that the fitting degree reaches 84.57% and the trust value is basically 85%, so it can be trusted.

9 Other Features

9.1 Special Properties of Words

1. Spelling habits of words

We noticed that there are two words in the game, "humor" and "favor," whose spellings are more widely used in the United States, while "humour" and "favour" are more widely used in other regions. Although English speakers from almost all regions can understand the meaning of these words in different spellings, users of "humour" and "favour" may not immediately include them in the range of five-letter guesses when playing the game, which can affect the distribution of results to some extent.

2. Words with three or more repeated letters

When a word has too many repeated letters, such as the word "mummy," it increases the difficulty of guessing the word because the existence of repeated letters makes it harder for guessers to guess the correct letter. Additionally, if a word contains multiple identical letters, guessers may find it difficult to determine their positions in the word, further increasing the difficulty. In addition, too many repeated letters may cause guessers to repeatedly guess the same letter, which can lower the smoothness of guessing or affect established guessing patterns that some players may have.

3. Other five-letter words with similar substitutions

Some words look very similar to commonly used words, such as "libel" and "label," "smelt" and "smell." When guessing, people are more likely to choose more common words like "label" and "smell" instead of proper nouns like "libel" or "smelt." If certain words in the game look very similar to commonly used words, guessers may mistake them for common words instead of specific words.

4. More easily associated with their four-letter form

Some words are more commonly used in their four-letter form, and because of people's usage habits, it may be difficult to associate them with their five-letter form. For example, for the word "needy," people are more likely to use the word "need," making "needy" harder to come up with.

9.2 Special Features of the Total Report Number

1. Special Holidays

For Wordle players around the world, Thanksgiving and Christmas are both important holidays. Interestingly, the number of participants suddenly dropped to 15,000 on Christmas Day, approaching the trough of the fluctuation during that period. Perhaps everyone was spending time with their families that day.

For Thanksgiving, the number of participants on that day was higher than the average for that time period. In such a holiday, many people would associate it with words related to the Thanksgiving dinner. The probability of guessing the correct answer in one try on that day was surprisingly as high as one in twenty.

2. Possible Data Flaws

On November 30, 2022, the total number of reports sharply dropped to 2569. We analyzed the data from the previous few days and investigated whether there were any specific sudden events that day, but we did not find any events that could explain the reason for the appearance of this data point. Therefore, we believe that this data point may have some error and it is likely that a digit is missing from the end of the number. 25690 would be a more reasonable range.

10 Model Evaluation

10.1 Strengths

- The model considers the superimposition of long and short cycles, making the prediction of the number of reportees more accurate and providing confidence levels.
- The word difficulty assessment model is simple and easy to understand, and has good results. This means that any new set of data can be evaluated in a short time and can be quantitatively represented.

10.2 Weaknesses

- Due to the limited number of words on which the model is based, there may be very few examples of a particular word attribute, resulting in high randomness and inaccurate judgments. This problem can be solved by accumulating more data on words in the future.
- Many word attributes are correlated, but in this article, they are considered independent variables. Additionally, attributes such as single syllable and multi-syllable are not taken into account, which may make the model incomplete. To get a better result, we can create new and better attributes or add more words while conducting correlation analysis.
- When analyzing whether word attributes affect the distribution of the number of times the difficult mode is chosen, we only proved whether it has an impact but did not give a specific deviation. This lack of quantitative representation can be addressed by analyzing separate variables during the period when the proportion of the difficult mode selection increases.

Letter

To: Wordle Game Editor of The New York Times
From: Team 2320719
Date: December 21st, 2023
Subject: Recommendations Regarding the Wordle Game

Dear Wordle Game Editor of The New York Times,

I am writing to you as a devoted player of Wordle. I find the game both entertaining and challenging, and I am interested in the future development of Wordle.

Based on recent data from Wordle gameplay, I have come to some conclusions about the current state of the game and would like to share my predictions with you. I would also like to take this opportunity to offer some suggestions for Wordle.

Firstly, it has been observed that the daily participation rate for Wordle is approaching a stable state, with a slight but not significant downward trend. The proportion of players participating in the difficult mode is also approaching stability. This indicates that Wordle already has loyal players, but the number of new players joining is decreasing, making it difficult to increase the number of players.

To address this issue, I have the following suggestions to attract more active players:

Add special words with rare features: As a word-guessing game, Wordle has an inherent logic to its solutions. This means that the more familiar a player is with the vocabulary, the fewer attempts they need to solve the puzzle. However, it is apparent that current players have reached a level of familiarity with the game's vocabulary. To break this stalemate, adding rare words with unique features, such as words with three identical letters, could be an effective solution.

Strengthen the grading system: By adding more words that are easy to guess in the simple mode but challenging in the difficult mode, players of different levels can find their respective tiers.

Evaluate based on the distribution of answer attempts: If the number of attempts for a particular word significantly increases or decreases, it may be due to unreasonable difficulty settings. In this case, using more word difficulty evaluation models could help to solve the issue.

Positive guidance mechanism: As a puzzle game, Wordle may appeal more to children if more positive words such as "smart" are used. It could also include pleasant content such as "lucky".

Provide rewards and competition mechanisms: People often enjoy competition and winning rewards, so Wordle can provide mechanisms such as leaderboards, scores, and rewards to encourage players to play daily and improve their enthusiasm. For example, giving different scores for different successful guesses and adding them up to get the total score for the day, which is then used to rank the players. This can increase both the competitiveness and the level of dependence of players.

Add social interactive functions: Wordle could include social interactive features, such as allowing players to communicate and share their achievements with other players. This can not only increase the fun of the game but also increase player stickiness.

Provide a good user experience: Wordle should provide a good user experience, including a clear and concise interface design, smooth gameplay, and fast loading speed. The player's experience in the game will directly affect their retention rate and word-of-mouth promotion.

In summary, while maintaining the current player base, more positive words can attract more players and generate a good reputation for Wordle.

Sincerely,

Team 2320719

References

[1] Short, M. B. (2022). Winning Wordle Wiselyor How to Ruin a Fun Little Internet Game with Math. The Mathematical Intelligencer, 44(3), 227-237.

[2] Trappey, C. V., & Wu, H. Y. (2008). An evaluation of the time-varying extended logistic, simple logistic, and Gompertz models for forecasting short product lifecycles. Advanced Engineering Informatics, 22(4), 421-430.

[3] Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP conference series: materials science and engineering (Vol. 336, p. 012017). IOP Publishing.

[4] Cowley, B., Charles, D., Black, M., & Hickey, R. (2008). Toward an understanding of flow in video games. Computers in Entertainment (CIE), 6(2), 1-27.

[5] Yang, M. S. (1993). A survey of fuzzy clustering. Mathematical and Computer modelling, 18(11), 1-16.

[6] Ross, S. M. (1995). Stochastic processes. John Wiley & Sons.

[7] Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2014). Mathematical statistics with applications. Cengage Learning.

[8] Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. Journal of the Royal Statistical Society: Series D (The Statistician), 47(1), 183-189.

[9] Murre, J. M., & Dros, J. (2015). Replication and analysis of Ebbinghaus forgetting curve. PloS one, 10(7), e0120644.