Problem Chosen	2023	Team Control Number
C	MCM/ICM Summary Sheet	2313161

Percentage Prediction of Tries Based on Partial Least Square Method and ARIMA Model

Wordle used to be a particularly popular puzzle that has taken Twitter and other social platforms by storm. The puzzle requires players to guess a five-letter word within six tries. Over the past few months, Wordle's player traffic has gradually declined due to the solidified game mode and gameplay. This does not bode well for Wordle's developers, and finding the reasons for the loss of users is the key to saving the game's drop in player traffic.

In view of the problem of player traffic, our study conducts ADF test on the time series of **number of reported results (NORR)** . On the basis of judging the stability of the series, the **ARIMA model** is constructed to perform time series analysis and predict the change of the **NORR** in the future. To explore whether there was a correlation between the words and the ratio of **number in hard mode (NIHM)** to **NORR**. In our study, the word attributes are divided into **the frequency of various types of letters, whether there is repetition and whether the first letter is common or uncommon initial**. And then the correlation between these attributes and the above ratio is analyzed. The results show that the spearman correlation coefficients between the three attributes and the above ratios are 0.01, 0.09, and -0.045, respectively. This degree of correlation is considered statistically insignificant.

In view of the relationship between user data, word attributes and the number of tries, our study uses **partial least squares method to construct regression model (PLS)**, and according to the Variable Importance in Projection of independent variables on the latent factors which relatively important, it is concluded that **frequency and repetition will play a more important role in constructing the regression model**. On the premise of model construction, our study based on the regression equation calculates the predicted values of the original data and obtains **mean of absolute residual** of these data, which is used as a test of credibility. Combined with the ARIMA model, the percentage of tries of the word "EERIE" under the user proportion on March 1,2023 is predicted and our study obtains the corresponding confidence interval.

In view of the difficulty level of word, our study uses the K-means algorithm to cluster the currently used words. Three word clusters with different difficulty levels are obtained: **Simple**, **Medium**, and **Difficult.** According to the above clustering model, it is concluded that **"EERIE" belongs to Difficult word**. In order to judge the accuracy of the clustering model, our study calculated the silhouette coefficient, DBI and CH of the model, and the results were 0.687, 0.53 and 508.931, respectively, which represented the high accuracy of the model.

In view of the clustering results described above, the following interesting conclusions are drawn: **word difficulty is negatively and significantly correlated with the percentage of the first three tries and positively and significantly correlated with the percentage of more than five tries**. Moreover, in the prediction results of **NORR** and **NIHM** by ARIMA, it can be seen that **the ratio of the two will continue to increase in the future**.

Finally, our study discusses and analyzes the established model, and makes an objective comprehensive evaluation of its advantages and disadvantages. On this basis, our study compares the fitting situation of ARIMA model and Long Short-Term Memory artificial neural Network (LSTM) under the current data situation, and provides advice for the prediction of large amounts of data in the future: **if the amount of data is still small, the ARIMA model should be maintained. However, it’s necessary to consider switching to LSTM model if the amount of data continues to increase in the future.**

Keywords: ARIMA; Correlation Analysis; Partial Least Squares; K-Means

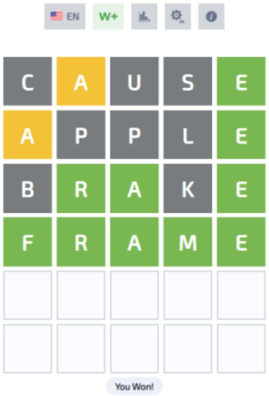
Contents

1 Introduction

1.1 Background

Wordle is a particularly popular puzzle currently in the ***New York Times***. Players try to guess, without prompting, a five-letter word that must be an officially recognized word. After players submit their word, the color of the tiles will change. If the letter in that tile is in the correct word and in the correct location, the color of the tile will be green. If the letter in that tile is in the correct word but in the wrong location, the color of the tile will be yellow. Besides, a gray tile shows that the letter in that tile isn't in the correct word. In the case of the example shown in **Figure 1**, the correct word is "FRAME". The first guess is "CAUSE". The correct word includes the letters "A" and "E". Because the "E" in the guessed word is in the correct location, it is shown in green; while the "A" is in the wrong location, it is shown in yellow; the rest of the letters are shown in gray because they are not included in the correct word. Players have six tries at most. The fewer guesses a player makes, the higher the score. **Figure 2** shows the pie chart of the percentage of attempts by different players to guess the word "slate" in the data file attached to the problems. (The percentage is the result of data preprocessing)

In this regard, the attributes of different words will be bound to affect the difficulty of the puzzle, such as the frequency of letters in the word, whether there are repeated letters and so on, thus affecting the frequency of tries. In order to explore the number of players playing Wordle in the future, the percentage of tries of guessing the word correctly and the difficulty of the word, our study carries out relevant analysis through the data file attached to the topic (Problem C Data Wordle.xlsx), and establishes a model that can achieve the above goals.



C	A	U	S	E
A	P	P	L	E
B	R	A	K	E
F	R	A	M	E

Figure 1: Wordle schematic diagram

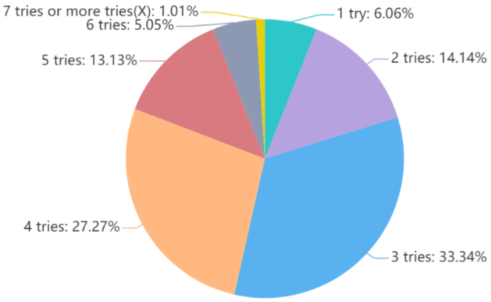


Figure 2: Pie chart of percentage of tries to guess the word "slate"

1.2 Statement of the problem

Problem I: Every day, the people playing Wordle are different and the number of people playing Wordle is also different, which means that **Number of reported results** also changes every day, that is, the number of people playing Wordle on the day of 2023.3.1 needs to be predicted by the number of people who have played the game in the past year given by the topic. In addition, different words have different attributes, so it is necessary to explore whether the attributes of different words will affect the percentage of scores reported that were played in Hard Mode, and explain it.

Problem II: The difficulty of guessing different words is not the same, and the percentage of tries to guess a word must also be different. This problem want to develop a model to predict the distribution of the percentage of guess tries for a particular word, with values 1,2,3,4,5,6,X, and try to predict the reported result of the word "EERIE." In addition, it is necessary to explain the uncertainty of this model and evaluate the prediction effect of this model.

Problem III: As can be seen from Problem I, the attributes of a word will affect the difficulty of guessing. This problem needs to develop a model that can classify the difficulty of different words and obtain the difficulty of the word "EERIE". Finally, the accuracy of the model needs to be further verified and discussed.

Problem IV: There are some other interesting features of the data set attached to the topic. This problem requires to find them and describe them.

1.3 Our Work

In view of the four problems given by the title, our work mainly includes the following:

- Since the **Number of reported results** also has a regular change with time, our study uses the ARIMA to build a time series prediction model to complete the prediction of **Number of reported results** on 2023.3.1.
- In order to explore the influence of word attributes on the percentage of scores reported that were played in Hard Mode, we summarize the word attributes that will affect the difficulty of the puzzle, such as the frequency of letters in the word, whether there are repeated letters and so on, and analyze the correlation between these attributes and the percentage of scores reported that were played in Hard Mode. The results are used to determine whether word attributes affect the percentage of scores reported that were played in Hard Mode.
- In order to predict the percentage of tries for a word corresponding to a date, our study adopted a regression model based on partial least squares method for the feature that the number of independent variables was less than the number of dependent variables. Our study predicted the percentage of tries according to the correlation equations about the independent variable and dependent variable obtained by the model, and evaluated the credibility of the model according to the mean of absolute residual.
- The difficulty of a word is affected by some attributes, so we can use K-means clustering algorithm to build a model according to these attributes, and cluster these words into several different levels of difficulty. Then we can analyze the difficulty level of the word "EERIE" through this model, and further evaluate the model by some evaluation indicators.
- Through data analysis, we found some interesting conclusions: there was a correlation between the difficulty of words and the percentage of the number of tries; the percentage of scores reported that were played in Hard Mode was decreasing, and the popularity of Wordle first increased sharply and then decreased, and the decline rate was slowing down.

Figure 3 below is a brief diagram of our entire workflow:

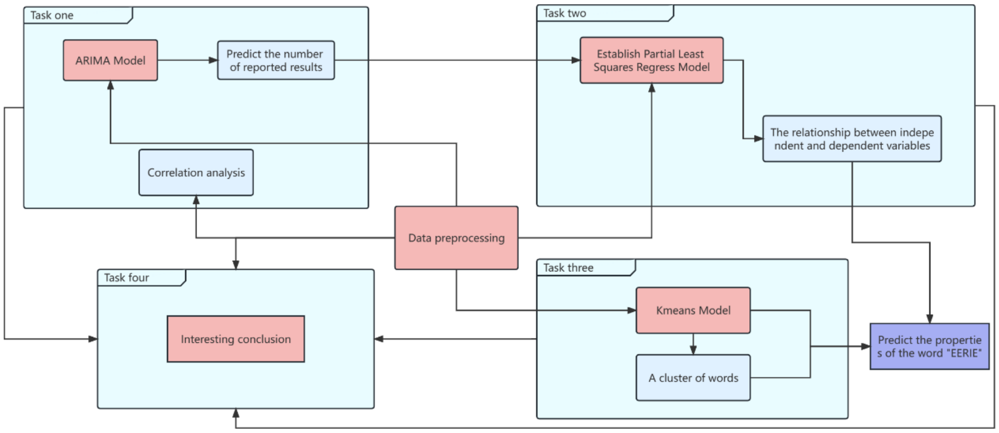

Task one
ARIMA Model
Predict the number of reported results
Correlation analysis
Task two
Establish Partial Least Squares Regress Model
The relationship between indepe ndent and dependent variables
Data preprocessing
Task four
Interesting conclusion
Task three
Kmeans Model
A cluster of words
Predict the propertie s of the word "EERIE"

Figure 3: Diagram of the workflow

2 Assumptions

Assumption 1: We assume that players fill in the data according to their physical truth of guesses.

Assumption 2: We assume that players don't use other tools to complete the game.

Assumption 3: When using the ARIMA algorithm for regression forecasting, our study assume that the other disturbance terms follow an independent normal distribution.

Assumption 4: We assume that players' IQ follows a normal distribution.

3 Notations

Symbol	Description
α_i	The percentage of guessing words correctly with i times
β_i	The percentage of guessing words correctly with i times after data processing
β_i^*	The percentage of guessing words correctly with i times by predicting
T_a	The frequency of the letter A （The frequency of the letter B is denoted by T_b, and so on）
η	the percentage of scores reported that were played in Hard Mode
f_i	frequency indicator before normalization
f_i^*	final normalized frequency indicator
y_t	The value of the function on the t-th day
$\boldsymbol{r}_\mathbf{s}$	spearman correlation coefficient
p	The number of variables initially involved in the analysis
h	The final number of iterations
w_{jk}	the weight of the j-th variable to be mapped at the k-th iteration
$\boldsymbol{X}$	the independent variable matrix
$\boldsymbol{Y}$	the dependent variable matrix
$\boldsymbol{\gamma}$	the constant vector
$\boldsymbol{R}$	the coefficient matrix of PLS model

4 Data Preprocessing

4.1 Abnormal Value Processing

Since the Wordle is a puzzle of guessing five-letter words, the "Word" in the data file should be all five-letter words. In the process of data processing, our study deleted several groups of data in which the “Word” did not meet the requirements of five letters, like "rprobe" whose corresponding contest number is 545. Besides, there are some words that aren’t an officially recognized words, like " naïve " whose corresponding contest number is 540. Our study also deleted the data corresponding to the word.

In addition, our study deleted the abnormal value of the "the number of reported results" to make sure predictions are more accurate.

4.2 The Percentage Of Different Tries Processing

In the data file, for some groups of data, the percentages of different tries might not sum to 100% due to rounding. To normalize this, our study assumes that the percentage that guesses the word in i tries is α_{i} (i = 1,2,...,6), and the percentage that could not solve the puzzle is α_7. For words whose percentages sum to less or more than 100% of all tries, our study assumes that the corrected percentage that guesses the word in i tries is β_{i} (i = 1,2,...,6), and the corrected percentage that could not solve the puzzle is β_7, the formula is as follows:

$$\beta_{\mathrm{i}} = \frac{\alpha_{\mathrm{i}}}{\sum_{i=1}^{7} \alpha_{\mathrm{i}}} \tag{1}$$

Thus the percentages of different tries of the word sum to 100%.

5 Model Construction

5.1 Modeling and solving of Task 1

5.1.1 Method Overview of Task 1

This problem first needs to prediction the **number of reported results** on 2023.3.1. Because the influence of external conditions is small, our study uses the ARIMA model for time series prediction to obtain the results. In addition, the problem also asks us to explore the influence of word attributes on the the percentage of scores reported that were played in Hard Mode, that is, this problem needs to first calculate the the percentage of scores reported that were played in Hard Mode in each word. Because there must be a certain relationship between the percentage and the difficulty of the word, our study analyzes three attributes of the word. That is, the frequency of letters, whether there are repeated letters, and the frequency of words using different initial letters. These three attributes are related to the difficulty of words to a certain extent. Therefore, the correlation analysis between these three attributes and the difficulty of words can be done to draw the conclusion, and the flow chart of ideas is shown in **Figure 4** below.

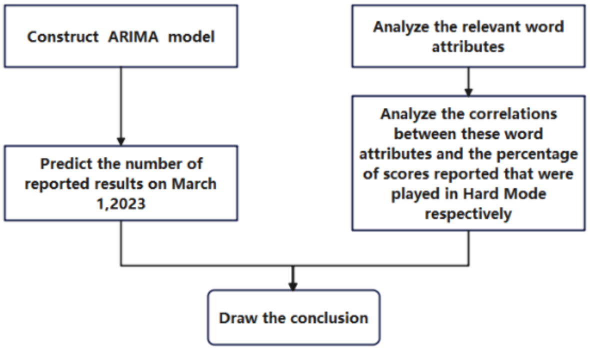

Construct ARIMA model
Predict the number of reported results on March 1,2023
Analyze the relevant word attributes
Analyze the correlations between these word attributes and the percentage of scores reported that were played in Hard Mode respectively
Draw the conclusion

Figure 4: The Flow Chart of Task I

5.1.2 Autoregressive Integrated Moving Average Model

ARIMA (Autoregressive Integrated Moving Average Model) is one of the time series prediction methods. ARIMA(p, d, q) which actually refers to the combination of the autoregressive model and the moving average model can effectively eliminate the random fluctuations in the prediction and complete the prediction according to its own data, the formula is defined as follows:

$$y_t = \mu + \sum_{i=1}^{p} y_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \Theta_t \varepsilon_{t-i} \tag{2}$$

Where μ is a const, ε_t is the error value, p is the auto-regressiveitem, q is the moving average item, d refers to the number of differencing done when the time series becomes stationary. The difference between this model and ARMA is that it transforms non-stationary time series into stationary time series, and the way of transformation needs d order difference operation. ∇ is the difference operator, and y_t represents the function value of t-th day, which should satisfy the following formula:

$$\nabla y_t = y_t - y_{t-1} \tag{3}$$

$$\nabla^2 y_t = \nabla(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2} \tag{4}$$

Only the first-order and second-order differences operations are shown. For higher-order differences, we can do the similar operations. After the transformation is completed, the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) of the stationary time series can be calculated.

The formula of the Autocorrelation Function (ACF) is as follows:

$$ACF(k) = \frac{Cov(y_t,\ y_t - k)}{Var(y_t)} \tag{5}$$

The range is [-1, 1], but the function does not simply refer to the relationship between y_t and y_{t-k}. It is also influenced by the k-1 intermediate variables y_{t-1}, y_{t-2}, …, y_{t-k+1}, while the Partial Autocorrelation Function (PACF) eliminates the correlation degree of the influence of the above variables, and obtains the strict correlation between the two variables y_t and y_{t-k}.

Finally, p and q can be determined through the above two functions. And then the model establishment is completed.

5.1.3 Calculate the percentage of scores reported that were played in Hard Mode

Assume the **number in hard mode** of one day as m_1 and the **number of reported results** of the same day as m_2. Then the percentage of scores reported that were played in Hard Mode on that day is calculated as follows:

$$\eta = \frac{m_1}{m_2} \times 100\% \tag{6}$$

5.1.4 Attributes of words

5.1.4.1 frequency

To guess the correct word faster, the word with high frequency letters will have a higher probability of guessing. Therefore, the attribute of letter frequency is an important basis for correlation analysis, because it has an important impact on the difficulty of guessing. In our study, according to all the words given by the data, the frequency of the occurrence of all letters is counted, and the bar chart obtained is shown in **Figure 5** below.

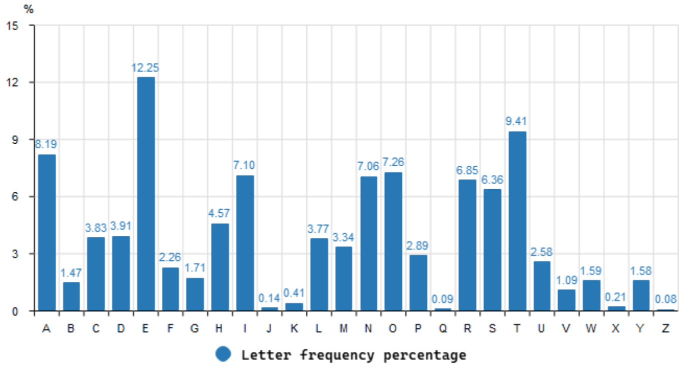


Figure 5: Bar Graph of Letter Frequency

Assume the frequency of different letters be T_a, T_b, T_c…The frequency indicator of a word can be obtained by adding the frequencies of each letter in which it appears. Assume frequency indicator as f_i. For example, the value of the f_i of "APPLE" is $T_a+2T_p+T_l+T_e$.

In order to reduce the error caused by the large difference of the values, it is necessary to standardize the frequency indicator f_i. Assume the mean and standard deviation of the frequency indicator for all words in the given data as μ and σ, respectively. Firstly, it is necessary to process the given data into standard normal distribution, so we need to pre-normalize it at first. Assume the frequency indicator after pre-standardization be f_i', and the conversion formula is as follows:

$$f_i' = \frac{f_i - \mu}{\sigma} \tag{7}$$

The higher the value of f_i', the less difficult the word. In order to achieve the purpose that the higher the frequency indicator, the more difficult the word, we need to do the following transformation, the frequency indicator after transformation is f_i'':

$$f_i'' = \frac{\mu}{\sigma} - f_i' \tag{8}$$

In order to map the data to [0, 1] to reduce the error caused by large difference of the values, we need to discretize the data. In our study, it is believed that f_i '' will not exceed $\frac{2\mu}{\sigma}$ (the given data is indeed less than this value). Therefore, assume the final normalized frequency indicator is f_i^*, the discretization process can be performed as follows:

$$f_i^* = \frac{f_i''\sigma}{2\mu} \tag{9}$$

The final standardized indicator can be obtained, and the correlation analysis between f_i^* and the percentage of scores reported that were played in Hard Mode can be directly performed after processing. The following **Table 1** shows the values of f_i and f_i^* of some words in the data.

Table 1: The Value of Frequency before and after normalization for some words

Word	f_i	f_i^*
slump	18.94	0.6824
crank	26.34	0.5583
gorge	29.78	0.5007
query	23.35	0.6085

drink	25.33	0.5753
abbey	24.96	0.5815

5.1.4.2 repetition

Word will be more difficult to guess when it has repeated letters, because most of the time it is assumed that the same letter only appears once, so our study thinks that it also has some influence on the percentage of scores reported that were played in Hard Mode. To verify the correlation between them, our study assume an indicator named “repetition”. Repetition indicator of words with repetitive letters are represented by a “1”, but otherwise by a “0”, as shown in **Table 2** below. **Table 3** below shows the repetition indicator for some words in the data.

Table 2: The Meaning of Repetition Indicator

repetition	meaning
0	words without repeated letters
1	words with repeated letters

Table 3: The Value of Repetition Indicator of Some Words

Word	slump	crank	gorge	query	drink	abbey
repetition	0	0	1	0	0	1

5.1.4.3 initial

In addition to the above two attributes that may have an effect on the percentage of scores reported that were played in Hard Mode, we also found that the frequency of use of words with different initial letters also had an effect. According to the conclusion on the occurrence frequency of the first letter in literature [2], our study lists the first letter t, a, w, i and s as key words, and the first letter j, k, q, u, v, x and z as non-key words. And our study assume an indicator named “initial”. Besides, "2" is used to represent key words, "0" is used to represent non-key words. Words that are neither key words nor non-key words are represented by "1", as shown in **Table 4** below.

Table 4: The Meaning of Initial Indicator

initial	meaning
0	non-key words
1	neither key words nor non-key words
2	key words

Table 5: The Value of Initial Indicator of Some Words

word	slump	crank	gorge	query	drink	abbey
initial	2	1	1	0	1	2

To more fully represent the size of the various attributes of words, **Figure 6** below is a radar diagram of the three attributes of the six words as examples above.

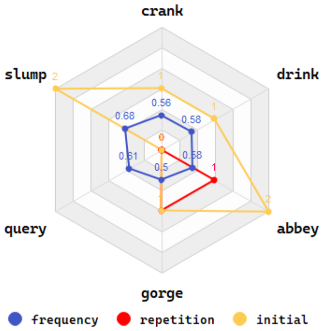


Figure 6: Radar map of different attributes of words

5.1.5 Model Solution

5.1.5.1 ADF test and ARIMA parameter determination

In order to determine whether this data is capable of using an ARIMA model, an ADF test is also required. The purpose of this test is to analyze whether the corresponding variable can significantly reject the original hypothesis that the series is unstable. If it shows significance ($P<0.05$), the original hypothesis is rejected and the series is a stationary time series. In order to facilitate the solution of subsequent problems, not only the **Number of reported results** but also the **Number in hard mode** will be predicted in this problem. The following **Table 6** shows the ADF test table of the variable **Number of reported results** and **Number in hard mode**. The AIC indicator is used as a standard to measure the goodness of fit of statistical models, and the smaller the value is, the better the fit is. Critical values are fixed values corresponding to a given significance level.

Table 6: ADF test table of the variable Number of reported results and Number in hard mode

ADF Test Table							
Variable	Order of difference	t	P	AIC	Critical values		
					1%	5%	10%
Number of reported results	0	-3.773	0.003***	7085.808	-3.45	-2.87	-2.571
	1	-4.235	0.001***	7076.079	-3.45	-2.87	-2.571
	2	-10.359	0.000***	7057.039	-3.45	-2.87	-2.571
Number in hard mode	0	-1.334	0.613	5375.777	-3.449	-2.87	-2.571
	1	-7.963	0.000***	5359.083	-3.449	-2.87	-2.571
	2	-7.842	0.000***	5372.468	-3.45	-2.87	-2.571

Note: *, **, * represent significance levels of 1%, 5%, and 10% respectively**

As can be seen from the above table, the **P** of the Number of reported results on the difference order of 0, 1 and 2 are all far less than 0.05, showing significance on the level, rejecting the original hypothesis, so the series is a stationary time series. The **P** of Number in hard mode on the difference order of 1 and 2 is far less than 0.05, showing significance on the level, rejecting the original hypothesis, and the series is a stationary time series. Therefore, the parameter d is determined to be 1 by synthesizing the two variables.

The parameters q and p can be determined according to the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF) of the first-order difference time series of the above two variables, as shown in **Figure 7** and **Figure 8** below.

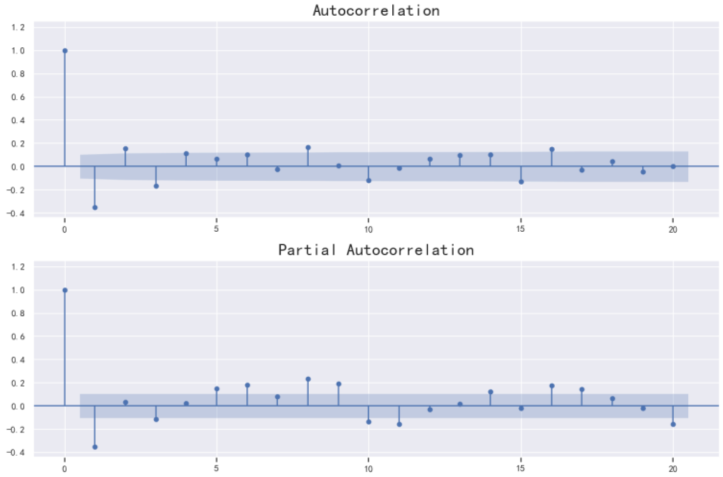


Figure 7: ACF and PACF of the first-order difference time series of the Number of reported results

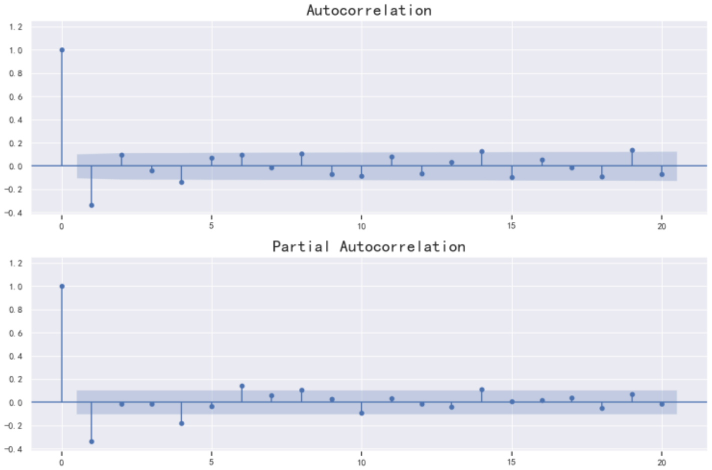


Figure 8: ACF and PACF of the first-order difference time series of the Number in hard mode

As can be seen from the above two figures, ACF and PACF of Number of reported results and Number in hard mode are both censored at order 1 (that is, ACF and PACF fluctuate randomly around 0 after the order is greater than 1), so the parameters p and q are determined as 1 and 0 respectively.

Tables of parameters for ARIMA(1,1,0) model testing **Number of reported results** and **Number in hard mode** are shown in **Table 7** and **Table 8** below.

Table 7: The parameter table of predicting Number of reported results

ARIMA model（1,1,0）Test Table		
Item	Symbol	Value
Sample size	Df Residuals	351
	N	354
Q statistical magnitude	Q6(the value of P)	0.061(0.805)
	Q12(the value of P)	26.624(0.000***)
	Q18(the value of P)	58.252(0.000***)
	Q24(the value of P)	86.144(0.000***)
	Q30(the value of P)	99.504(0.000***)
Information criterion	AIC	7636.932
	BIC	7648.532
Goodness of fit	R^2	0.982

Note: *, **, * represent significance levels of 1%, 5%, and 10% respectively**

Table 8: The parameter table of predicting Number in hard mode

ARIMA model（1,1,0）Test Table		
Item	Item	Item
Sample size	Df Residuals	351
	N	354
Q statistical magnitude	Q6(the value of P)	0.012(0.911)
	Q12(the value of P)	20.614(0.002***)
	Q18(the value of P)	33.221(0.001***)

	Q24(the value of P)	44.854(0.000***)
	Q30(the value of P)	57.562(0.000***)
Information criterion	AIC	5658.408
	BIC	5670.007
Goodness of fit	R²	0.947

Note: *, **, * represent significance levels of 1%, 5%, and 10% respectively**

It can be seen from the above two tables that for the variables **Number in hard mode** and **Number of reported results**, it can be obtained from the analysis of Q statistics magnitude results that Q6 does not show significant in the level, and the hypothesis that the residual of the model is a white noise sequence cannot be rejected. The goodness of fit R² of the former model is 0.947, and the goodness of fit R² of the latter model is 0.982. It can be seen that the model has excellent performance and basically meets the requirements.

5.1.5.2 Number in hard mode and Number of reported results predictions on March 1, 2023

Through the above determined ARIMA(1,1,0) model, our study have predicted the Number in hard mode and the Number of reported results in the next three months, and the prediction graphs are shown in **Figure 9** below.

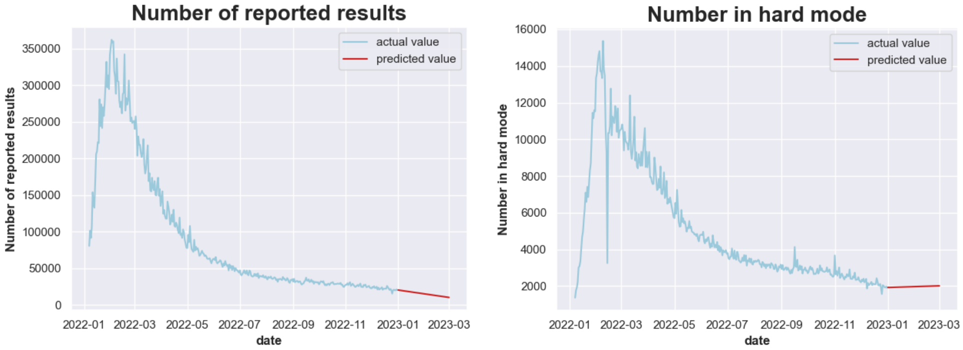


Figure 9: The Prediction Graphs of The Number in hard mode and The Number of reported results

The results of predictions for Number in hard mode and Number of reported **results** on March 1, 2023 (which have been rounded) are shown in **Table 9** below.

Table 9: The Results of Predictions for Number in hard mode and Number of reported results on March 1, 2023

Time	Number in hard mode	Number of reported results
2023.3.1	10310	2009

5.1.5.3 Spearman Correlation Coefficient

In order to facilitate the correlation analysis, Spearman Correlation Coefficient is used in our study. Firstly, it is necessary to clarify the concept of rank. The rank of a number refers to the position of the number in the data which is ordered from smallest to largest. **Table 10** below shows a simple example of this concept, where A and B denote two different groups of variable data.

Table 10: An example of the concept of rank

A	The rank of A	B	The rank of B	Rank Difference
3	2	5	1	1
8	5	10	4.5	0.5
4	3	8	3	0
7	4	10	4.5	-0.5

2	1	6	2	-1

The rank difference is the difference between the rank of two variables. Assume the two variables of the i-th group of data be A_i and B_i, d_i denotes the rank difference between the two, and n denotes the total amount of data, then the Spearman correlation coefficient is calculated as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{10}$$

It is easy to know that r_s is range to [-1, 1]. The closer it is to -1, the stronger the negative correlation between two variables is. The closer it is to 1, the stronger the positive correlation between two variables is. And the closer it is to 0, and the weaker the correlation is.

5.1.5.4 Correlation Analysis and Results

For the three word attributes summarized in 5.1.4, Spearman Correlation Coefficient was performed between these attributes and the percentage of scores reported that were played in Hard Mode respectively. Firstly, whether there is a statistically significant relationship between the two variables (P<0.05) was tested. And then analyze the positive and negative direction of the correlation coefficient and the degree of correlation. The Spearman Correlation Coefficient obtained is shown in **Figure 10** below.

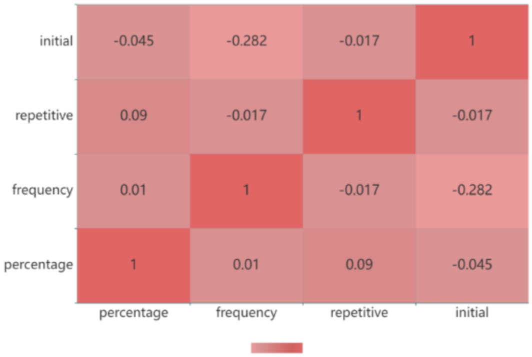


Figure 10: Heatmaps of Spearman Correlation Coefficient between word attributes and the percentage of scores reported that were played in Hard Mode

The Spearman Correlation Coefficient is shown as a heat map, with darker colors showing a stronger positive correlation and lighter colors showing a stronger negative correlation. The following **Table 11** shows the relationship between Spearman Correlation Coefficient r_s and correlation.

Table 11: The Relationship between Spearman Correlation Coefficient r_s and correlation

The range of $\|r_s\|$	0-0.19	0.20-0.39	0.40-0.69	0.70-0.89	0.90-1.00
correlation	extremely low	low	moderate	strong	extremely low

The comparison shows that the three attributes (**frequency**, **repetition**, and **initial**) have very low correlation with percentage, **so it can be considered that the attributes of the word are not directly related to the percentage of scores reported that were played in Hard Mode.**

5.2 Modeling and Solving of Task 2

5.2.1 Method Overview of Task 2

In order to predict the percentage of tries for a word by all users on a future date, our study explores the independent variables (user attributes and word attributes) that have a high degree of explanation for these dependent variables, and constructs the partial least squares regression model between these independent variables and the dependent variable, and obtains the correlation equations between the independent variables and the dependent variables. Given the date, the word to be predicted, and the correlation equations, our study can get the percentage of tries for "EERIE" by all users on 2023.3.1. Finally, the credibility of the model was tested by calculating the mean absolute value of the residual between the actual value and the predicted value, and the confidence interval of the percentage of tries of the word to be predicted was given. The flow chart of the idea is shown in **Figure 11** below.

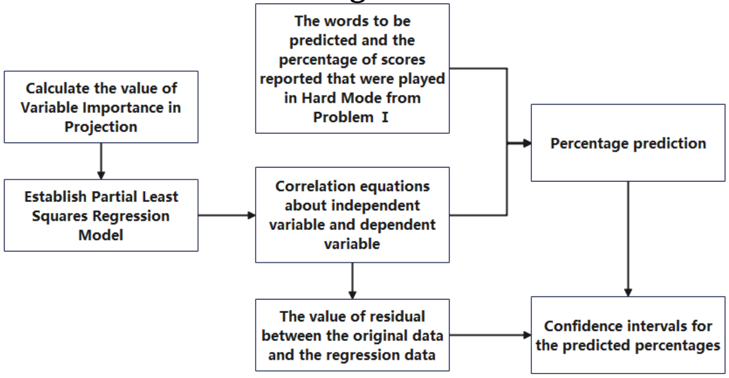

Calculate the value of Variable Importance in Projection
The words to be predicted and the percentage of scores reported that were played in Hard Mode from Problem Ⅰ
Percentage prediction
Establish Partial Least Squares Regression Model
Correlation equations about independent variable and dependent variable
The value of residual between the original data and the regression data
Confidence intervals for the predicted percentages

Figure 11: The Flow Chart of Task II

5.2.2 Partial Least Squares Model

Partial Least Squares (PLS) model is mostly used to find the basic relationship between two matrices (independent variable matrix X and dependent variable matrix Y), that is, a hidden variable method to model the covariance structure in these two vector spaces. In particular, PLS performs better than other models when the independent variables has variables whose number is less than the number of variables to be predicted. VIP refers to Variable Importance in Projection, which is mainly used to filter variables in PLS model. Its calculation formula is as follows:

$$VIP_j = \sqrt{\frac{p \sum_{k=1}^{h} \left(\hat{c}_k^2 t'_k t_k\right) \left(w_{jk}\right)^2}{\sum_{k=1}^{h} \hat{c}_k^2 t'_k t_k}} \tag{11}$$

where p is the number of variables initially included in the analysis, h is the final number of iterations, w_{jk} represents the weight of the j-th variable to be mapped at the k-th iteration; $\hat{c}_k^2 t_k' t_k$ is the degree of explanation of dependent and independent variables for the k-th iteration.

Since the sum of squares of the values of VIP of all variables is equal to 1, 1 is often used as the threshold value for VIP determination. **In our study, variables whose value of VIP greater than 1 explain the model to a high degree, while variables whose value of VIP less than 1 explain the model to a low degree.**

The algorithm process of this model is as follows:

1. Standardize the independent variable matrix X and the dependent variable matrix Y.
2. Assume the first principal component of X as p_1, and assume the first principal component of Y as q_1. Both of which have been unitized.
3. Assume $u_1 = Xp_1$， $v_1 = Yq_1$.

4. Var (u_1)→max， Var (v_1)→max, that is, the variance of the projection onto the principal components is maximized.

5. Corr (u_1, v_1)→max, that is, the correlation coefficient is maximized.

6. Integrating 4 and 5, the optimization objective is obtained: $\mathrm{Cov}(u_1,v_1) = \sqrt{\mathrm{Var}(u_1)\mathrm{Var}(v_1)}\mathrm{Corr}(u_1,v_1) \to$ max, that is, the covariance is maximized.

5.2.3 Model Solution

Here is a table of factor variance explanations:

Table 12: The Table of Factor Variance Explanations

latent-factor	variance of X	cumulative variance of X	variance of Y	cumulative variance of Y(R^2)	adjusted R^2
1	0.323	0.323	0.176	0.176	0.174
2	0.271	0.594	0.018	0.194	0.19
3	0.23	0.824	0.006	0.2	0.193
4	0.176	1	0.003	0.203	0.194

As shown in **Table 12** above, the first three latent factors can explain 80% of the information of the independent variable. However, all the latent factors cannot explain 80% of the information of the dependent variable.

Here is the **Variable Importance in Projection** Graph of independent variable:

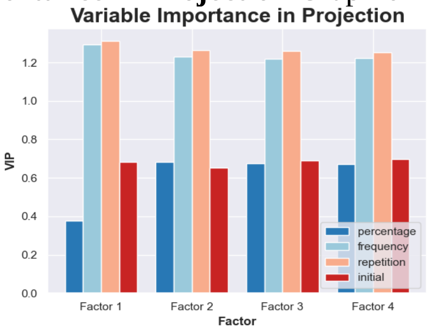


Figure 12: The Variable Importance in Projection of Independent Variable

Figure 12 shows the situation of VIP (Cumulative projected Importance) of each variable. The results intuitively show that frequency and repetition occupy a high proportion in the process of constructing the latent factor, and these two attributes will play a more critical role in the subsequent use of partial least squares regression.

The following **Table 13** is a table of the component matrices obtained after dimension reduction by PCA on the independent and dependent variables.

Table 13: The Component Matrices Table

variable	Factor 1	Factor 2	Factor 3	Factor 4
percentage	0.189	0.937	0.191	-0.101
frequency	0.646	0.039	-0.172	0.735
repetition	0.656	-0.343	0.689	-0.234
initial	-0.342	0.062	0.692	0.65
1 try	4.991	1.186	-22.068	-16.853
2 tries	-3.738	-1.854	14.028	9.046
3 tries	-0.287	0.445	-1.006	-0.036
4 tries	1.578	2.108	-7.618	-2.416
5 tries	5.136	2.518	-17.704	-16.363

6 tries	-9.241	-6.989	35.917	33.202
7 or more tries (X)	7.604	4.831	-25.897	-28.519

Table 14: Table of PLS Model Coefficient Results

	1 try	**2 tries**	**3 tries**	**4 tries**	**5 tries**	**6 tries**	**7 or more tries (X)**
const	2.183	14.449	35.292	26.873	14.051	6.191	0.962
percentage	-11.042	-13.25	-11.323	40.894	7.664	-21.37	8.427
frequency	-1.592	-15.207	-22.883	5.995	17.841	13.682	2.164
repetitive	-0.361	-3.16	-6.911	-0.864	5.108	4.503	1.687
initial	0.007	0.628	1.228	0.208	-0.976	-0.84	-0.256

The above table mainly includes the coefficients of the model, that is, the coefficient table used to construct the constant vector $\boldsymbol{\gamma}$ and the coefficient matrix $\boldsymbol{R}$. These coefficients can be used to analyze the influence relationship between the independent variable $\boldsymbol{X}$ and the dependent variable $\boldsymbol{Y}$.

The constant vector $\boldsymbol{\gamma}$ is as follows:

$$\boldsymbol{\gamma} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_7 \end{pmatrix} \tag{12}$$

where c_i represents the value of the const indicator corresponding to i tries in **Table 14.**

$\boldsymbol{R}$ is the coefficient matrix of PLS model:

$$\boldsymbol{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{17} \\ r_{21} & r_{22} & \cdots & r_{27} \\ \vdots & \vdots & \ddots & \vdots \\ r_{41} & r_{42} & \cdots & r_{47} \end{pmatrix} \tag{13}$$

where r_{ij} denotes the coefficient of the i-th independent variable with respect to the j-th dependent variable.According to the above two matrices (vectors), the equations of independent variable and dependent variable are obtained:

$$\boldsymbol{Y} = \boldsymbol{R}^T \boldsymbol{X} + \boldsymbol{\gamma} \tag{14}$$

In this formula:

$$\boldsymbol{Y} = \begin{pmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_7^* \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} d_1 \\ \vdots \\ d_4 \end{pmatrix}$$

where $\beta_{\mathrm{i}}{}^{*}$ denotes the value of predicting the percentage of guessing the word correctly with i tries, and d_{i} denotes the value of the independent variable. After the above processing, the sum of the predictions needs to be updated to 100%, that is, $\beta_{\mathrm{i}}{}^{*}$ will be updated to $\frac{\beta_{\mathrm{i}}^{*}}{\sum_{\mathrm{j=1}}^{\mathrm{j<7}}\beta_{\mathrm{j}}^{*}}$.

In addition, in the prediction process, the predicted value may be negative, for which our study uses amortization to deal with the overflow of negative values.

$$\beta_i^* = \begin{cases} 0, & \beta_i^* < 0 \\ \beta_i^* + \Phi(i) \sum_{j=0}^{j<7 \ and \ \beta_j^* < 0} \beta_j^*, & \beta_i^* \geqslant 0 \end{cases} \tag{15}$$

Here, $\Phi(x)$ is the percentage of the x-th prediction over the valid prediction (valid prediction is the fraction of the predicted value greater than 0):

$$\Phi(x) = \begin{cases} 0, & \beta_i^* < 0 \\ \frac{\beta_x^*}{\sum_{i=0}^{i<7 \ and \ \beta_i^* \geqslant 0} \beta_i^*}, & \beta_i^* \geqslant 0 \end{cases} \tag{16}$$

Table 15 shows the confidence of the model:

Table 15: The Confidence of The Model

attempts	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
Mean of absolute residual value(%)	0.396	2.288	4.931	3.378	3.677	4.158	2.117

Mean of absolute residual value of all the attempts: 2.992. In our study, it is considered that such a degree of error is within the acceptable range.

5.2.4 The results of prediction

According to the above equations and the mean of absolute residual value, the predicted value of the percentage of guess tries and its confidence interval for the word "EERIE" can be obtained, as shown in **Table 16** below and **Figure 13** below. In our study, it is believed that the predicted value is credible in the confidence interval.

Table 16: The Predicted Value of The Percentage of Guess Tries and its Confidence Interval for the word "EERIE"

attempts	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)
Predicted Value	0	7.335	23.547	34.324	22.063	7.994	4.738
Confidence Interval	[0, 0.396]	[5.047,9.623]	[18.616,28.478]	[30.946,37.702]	[18.386,25.740]	[3.836,12.152]	[2.621,6.855]

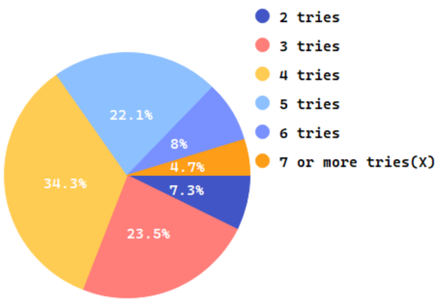


Figure 13: The Predicted Value of The Percentage of Guess Tries for the word "EERIE"

5.3 Modeling and solving of Task 3

5.3.1 Method Overview of Task 3

It can be known from Problem I that the frequency, repetition and initial are three types of attributes that may affect the difficulty of a word. Therefore, the K-means clustering algorithm can be used to establish a model through these three types of attributes to divide the words into three levels of difficulty: easy, medium and difficult. Then the difficulty of the word "EERIE" can be determined by analyzing the three types of attributes. Finally, our study will use some evaluation indicators, such as silhouette coefficient, DBI(Davies-bouldin), CH(Calinski-Harbasz Score) and so on, to further verify and discuss the accuracy of the model. The flowchart of the idea is shown in **Figure 14**.

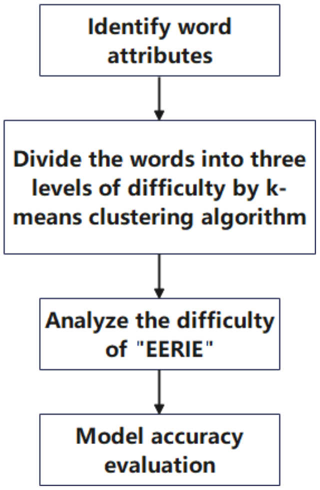

Identify word attributes
Divide the words into three levels of difficulty by k-means clustering algorithm
Analyze the difficulty of "EERIE"
Model accuracy evaluation

Figure 14: The Flow Chart of Task III

5.3.2 K-means clustering algorithm

The core of clustering analysis is that according to the valuable information of data samples, the samples with high similarity are divided into the same cluster according to the similarity, so as to obtain multiple clusters with different similarity. K-means clustering is an iterative solution clustering algorithm. At first, it needs to be determined to divide the data into K groups. Then define K cluster centers, and then calculate the "Euclidean Distance" between each other object and each cluster center, and assign each other object to its nearest cluster center. Once the assignment is complete, each cluster center and the objects assigned to it are called a cluster, and the cluster center for each cluster will be recomputed, iteratively repeating the previous step until the cluster centers no longer change.

The specific implementation steps are as follows. Let the given sample set {$\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …, $\boldsymbol{x}_m$}, and then choose K initial cluster centers randomly from the data set to be M ={C_1, C_2, …, C_k}. The "Euclidean Distance" calculation formula between data sample and the cluster center in the space is:

$$d(\boldsymbol{x}_i, C_i) = \sqrt{\sum_{j=1}^{m} (x_{ij} - C_{ij})^2} \tag{17}$$

where x_{ij} is the j-th attribute value of $\boldsymbol{x}_i$, C_i is the i-th cluster center, C_{ij} is the j-th attribute value of C_i. And the sum of the Squared Errors of the entire data set is defined as E whose calculation formula is:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |d(\boldsymbol{x}_i, C_i)|^2 \tag{18}$$

The smaller the value of E, the higher the similarity of the samples within the cluster, so after one assignment, the average value of all samples is calculated as the new cluster center, and then the above steps are repeated iteratively to make the value of E lower and improve the accuracy of clustering, until the clustering result remains unchanged. Then the result can be obtained according to the current cluster division.

The pseudocode of K-means clustering algorithm is as follows:

```
Algorithm : K-means clustering algorithm
    Input: Sample Set $D = \{x_1, x_2, \dots, x_m\}$, The Number of Clusters K
    Output: Cluster Partition $O = \{O_1, O_2, \dots, O_k\}$
    choose K initial cluster centers randomly from $D$ to be $M = \{C_1, C_2, \dots, C_k\}$
    repeat
        $O_i = Ø \ (1 \le i \le K)$
        for j = 1,2, …,m do
            calculate the "Euclidean Distance" between $x_j$ and each cluster centers
            assign $x_j$ to its nearest cluster center
        end for
        a cluster center $C_i$ and the objects assigned to it to be a cluster $O_i \ (1 \le i \le K)$
        recompute new cluster centers to obtain new data set $M$
    until $M$ no longer change
```

5.3.3 Difficulty Clustering Solution

In order to clearly show the difficulty level of words, our study divides the difficulty levels of words into three categories: easy, medium and difficult. Therefore, in the K-means algorithm, different words are clustered into three clusters according to their similarity, which is based on three attributes of words: **frequency**, **repetition** and **initial**. Then, according to the characteristics of frequency, repetition and initial of these three cluster centers, the higher the value of frequency and repetition indicator are, the higher the difficulty of the word is, and the higher the value of initial indicator is, the lower the difficulty of the word is, so they are divided into three categories: easy, medium and difficult respectively, as shown in **Figure 15** below. For the distribution of the clustered words in the three difficulty levels of easy, medium and hard, the results are shown in **Figure 16** below.

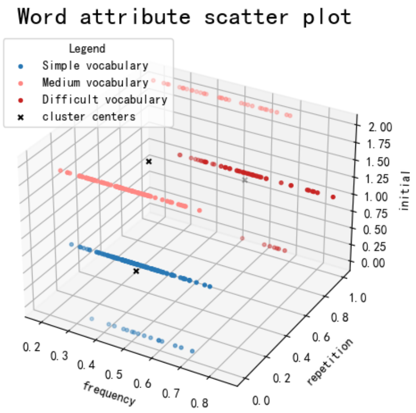


Figure 15: Clustering Results of Each Sample

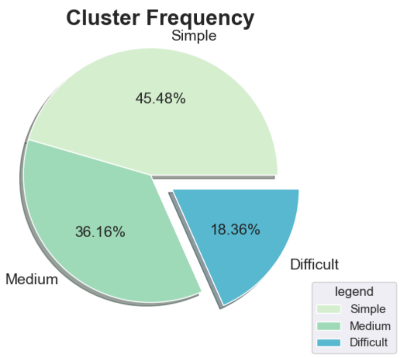


Figure 16: The proportion of the words in the three difficulty levels of easy, medium and hard

5.3.4 Difficulty Classification of "EERIE"

For the word “EERIE”, the values of the word attributes (**frequency**, **repetition** and **initial**) can be obtained as described above. Then, the distances between the sample and the three cluster centers are calculated respectively and compared with each other. And it is classified into the cluster with the smallest distance. The respective distance calculation results between the sample and the three cluster centers are shown in **Table 17** below:

Table 17: Distance of sample "EERIE" from three cluster centers

The Level of Difficulty	simple	medium	difficult
Distance	1.2742	1.0710	0.3954

Therefore, it can be concluded that the word "EERIE" should be classified as a word of the difficulty level.

5.3.5 Model Accuracy Evaluation

In order to determine the accuracy of the model, our study evaluates model with the silhouette coefficient, DBI(Davies-bouldin), and CH(Calinski-Harbasz Score). According to literature [3], our study obtains the following explanation of contour coefficient, DBI and CH.

silhouette coefficient: For a particular sample $\boldsymbol{x}_i$, assume the average of the distances from this sample to all other points in the cluster that contains this sample as $a(\boldsymbol{x}_i)$ and assume the minimum value of the average distance between this sample and all points in a cluster that does not contain this sample as $b(\boldsymbol{x}_i)$, then the silhouette coefficient of this sample can be shown below:

$$S(\boldsymbol{x}_i) = \frac{b(\boldsymbol{x}_i) - a(\boldsymbol{x}_i)}{\max\{a(\boldsymbol{x}_i),\ b(\boldsymbol{x}_i)\}} \tag{19}$$

It follows from its definition that $S(\boldsymbol{x}_i) \in [-1,1]$. For a sample set, its silhouette coefficient is the average of the silhouette coefficient of all samples. If the value of the silhouette coefficient is higher, the clustering effect is better.

DBI(Davies-bouldin): The meaning of it is the ratio of the sum of the intra-cluster distances to the inter-cluster distances about any two clusters. Assume i-th cluster’s intra-cluster distances as S_i which can be used to measure the dispersion degree of sample points, and it can be calculated as follows:

$$S_i = \left\{\frac{1}{T_i}\sum_{j=1}^{T_i} |d(\boldsymbol{x}_j,\ C_i)|^q\right\}^{\frac{1}{q}} \tag{20}$$

where T_i is the number of the i-th cluster sample, $d(\boldsymbol{x}_j, C_i)$ is the distance between the j-th point in the i-th cluster and the corresponding cluster center point. When q is equal to 1, S_i represents the average from each point to the cluster center, and when q is equal to 2, S_i represents the standard deviation from each point to the cluster center, both of which can be used to measure the intra-cluster distance.

And then inter-cluster distance can be calculated as follows:

$$G_{ij} = \left\{\sum_{k=1}^{N} |C_{ik} - C_{jk}|^p\right\}^{\frac{1}{p}} \tag{21}$$

where C_{ik} is the k-th attribute of the i-th cluster’s cluster center, N is the dimension of C_i, p has the same meaning as q above. Then DBI can be calculated as follows:

$$D = \frac{S_i + S_j}{G_{ij}} \tag{22}$$

It can be seen that a smaller value of DBI indicates a better clustering effect.

CH(Calinski-Harbasz Score): CH can be obtained by the ratio of separation and closeness. Separation refers to the sum of squared distances between each sample and the corresponding cluster center, and closeness refers to the sum of squared distances between the center point of the cluster and the center point of the sample set. If the value of CH indicator is larger, the clustering effect is better.

Through the above methods, our study obtains the silhouette coefficient, DBI and CH of this model, as shown in Table 18 below.

Table 18: The Value of the silhouette coefficient, DBI and CH of this model

silhouette coefficient	DBI	CH
0.687	0.53	508.931

It can be found that the contour coefficient and CH of the clustering model are high, and DH is relatively low, so **it can be concluded that the clustering model has high accuracy in the difficulty classification of words**.

5.4 Description of interesting data characteristics about Task 4

According to the data given by the title, we found several interesting characteristics, as shown in **Figure 17** below, which will be described in detail below.

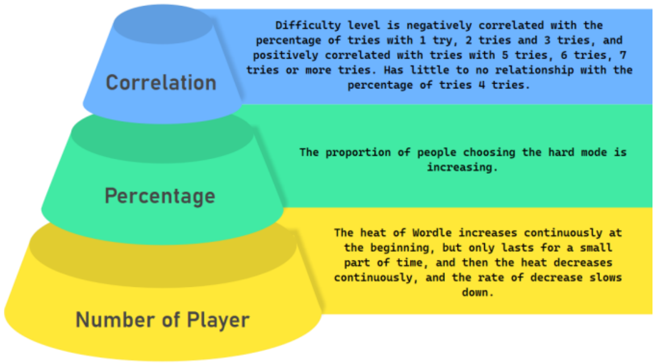

Correlation
Difficulty level is negatively correlated with the percentage of tries with 1 try, 2 tries and 3 tries, and positively correlated with tries with 5 tries, 6 tries, 7 tries or more tries. Has little to no relationship with the percentage of tries 4 tries.
Percentage
The proportion of people choosing the hard mode is increasing.
Number of Player
The heat of Wordle increases continuously at the beginning, but only lasts for a small part of time, and then the heat decreases continuously, and the rate of decrease slows down.

Figure 17: Interesting feature diagram

5.4.1 Correlation Characteristics Between Cluster Indicator And The Percentage Of Different Tries

In 5.3, we have grouped words into three class clusters according to their attributes, which are simple, medium and difficult respectively. We set an indicator name as "**cluster**", which is used to distinguish the class cluster of the corresponding word. The value of **cluster** corresponds to the difficulty of the word as shown in Table 19 below. Table 20 below shows the cluster indicator for some words in the data.

Table 19: Mapping table of cluster values

level of difficulty	**cluster**
simple	0
medium	1
difficult	2

Table 20: The Value of Cluster Indicator of Some Words

Word	slump	crank	gorge	query	drink
cluster	1	0	2	0	0

Spearman correlation analysis was conducted on the cluster value and the percentage of tries 1,2,3,4,5,6 and X respectively, and the results were obtained as shown in the following heat **Figure 18**.

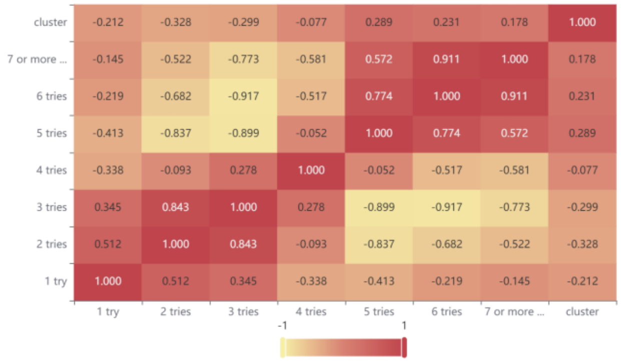


Figure 18: Spearman correlation coefficient heat map of cluster value and percentage of tries

According to Table 11 above, **cluster** is negatively correlated with the percentage of tries with 1 try, 2 tries and 3 tries, and positively correlated with tries with 5 tries, 6 tries, 7 tries or more tries. Has little to no relationship with the percentage of tries 4 tries. In other words, the difficulty of the word will affect the number of tries, that is, the easier the word, the fewer the tries. Interestingly, though, the difficulty of a word makes little difference in trying 4 tries, since their correlation is close to 0.

5.4.2 The Feature Of The Percentage Of People Choosing The Hard Mode

We have calculated the percentage of people who choose the hard mode in question 1. **Figure 19** shows the line chart of the proportion of people who choose the hard mode from January 7, 2022 to December 31, 2022.

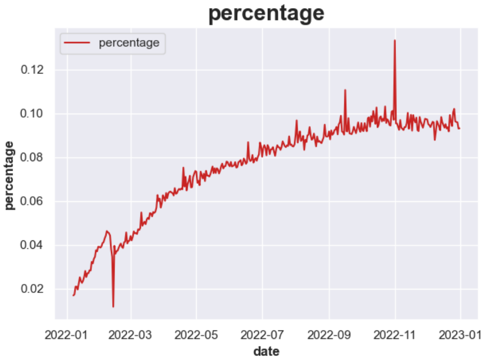


Figure 19: Line chart of percentage of people choosing hard mode

It can be found that the proportion of the total number of people playing wordle who choose hard mode every day is increasing. That is to say, with the change of time, there is a growing tendency among Wordle players to play hard mode, and this proportion may increase in the future.

5.4.3 Change In The Number Of People Playing Wordle

In the observation data, we found that the number of people playing Wordle changed greatly, as shown in **Figure 20** below. The darker the color, the larger the Number of people playing Wordle. From left to right, it changes over time, with January 7, 2022 at the far left end and December 31, 2022 at the far right end. **Figure 21** shows the change rate of the number of people who play Wordle every day, i.e. the change of the difference between the day before and the day after playing Wordle.

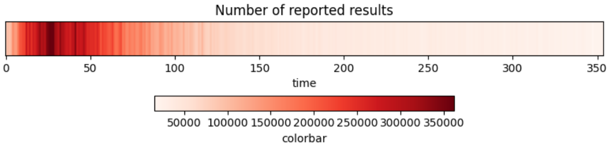


Figure 20: Change in the number of Wordle players

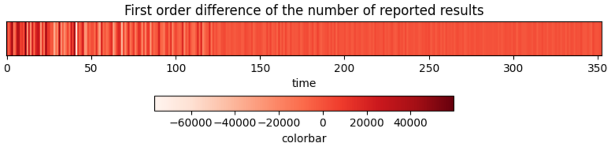


Figure 21: The change rate of the number of Wordle players

From this, it can be seen that there was a sharp increase in the number of people playing Wordle for a period of time, which peaked on a certain day, and then decreased over time. Therefore, an interesting phenomenon can be inferred, that is, the heat of Wordle increases continuously at the beginning, but only lasts for a small part of time, and then the heat decreases continuously, and the rate of decrease slows down.

6 Strengths and weaknesses

6.1 Strengths of Model

- In the process of predicting the **Number of reported results** and the **Number in hard mode**, the method used in our study is the ARIMA model, which only needs to use the endogenous variables in the given data and does not need to use other exogenous variables, so the model is relatively simple.
- According to the characteristic that the number of independent variables is less than the number of dependent variables, the **partial least squares regression** model is determined, and the method is accurate.
- The user data and the word difficulty attribute partition data obtained by time series prediction are used to are used as the independent variables of the **partial least squares regression** model, which has strong coupling and is innovative.
- In our study, K-means algorithm is used to realize the fast clustering of discrete data, and the algorithm has strong interpretability.
- The results obtained by fusing multiple models for prediction and classification are more comprehensive.

6.2 Weaknesses of Model

- The use of ARIMA model is based on **Assumption 3**. However, this assumption has limitations. When the disturbance variable does not obey the normal distribution in practice (for example, the server is down for updates), it may lead to the transformation of the stationary time series into unstable time series, which makes the ARIMA model fail in part of the time series.
- Although the division of word attributes includes most of the variables related to difficulty, it is still subjective.

7 Conclusions and extension

According to the characteristics of small amount of data, our study established an ARIMA model for the prediction of time series data, but when the amount of data is large, the stationarity of time series is more difficult to be guaranteed. In this case, we can choose the Long Short-Term Memory (LSTM) model instead.

The advantages of LSTM model are as follows:

- The problem of long-term dependence in RNN is improved. LSTM model generally performs better than temporal recurrent neural networks and Hidden Markov Models (HMM). As a nonlinear model, LSTM can be used as a complex nonlinear unit to construct larger deep neural networks.
- LSTM uses various gate functions to retain important features, which can effectively slow down the gradient disappearance or explosion that may occur in long sequence problems. Although this phenomenon cannot be eliminated, it performs better than traditional RNN on longer sequence problems

However, the drawbacks of this model are obvious:

- When the data size is small, the performance on the test set is not better than other models, and the fit may not be high, that is, the overfitting may be high.
- The internal structure of this model is relatively complex, resulting in more computing time.

For the current data, our study also tried to use the LSTM model to fit the **Number of reported results**, but the results of the fitting were not satisfactory, as shown in **Figure 22** below. (The blue line represents the raw data, the yellow line represents the fitted values)

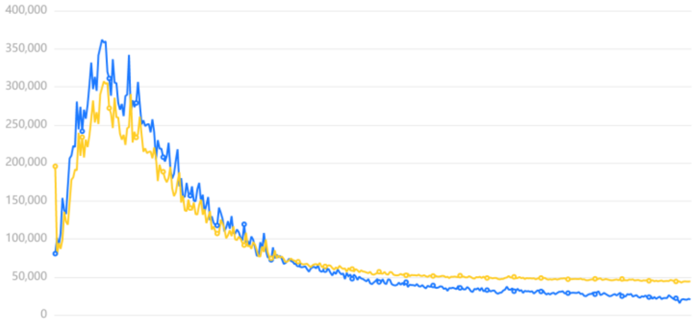


Figure 22: LSTM Fitting Results

For this fitting result, our study believes that it is caused by the drawbacks I of LSTM model. Therefore, our study did not use the LSTM model instead of the ARIMA model. But if we have more data, we can still consider using LSTM model to fit the time series.

References

[1] XU Weichao. Review of correlation coefficient research. Journal of Guangdong University of Technology, 2012, 29(3):12-17

[2] ZHAO Hang, MA Ji, ZHANG Fuchun, LIU Xiaoning. Analysis of variance was used to study the frequency of words with different initial letters in CET-4 and CET-6. English Square(academic research), 2012, (12):87-88

[3] Saroj,Kavita.Review:study on simple k mean and modified K mean clustering technique.International Journal of Computer Science Engineering and Technology,2016,6(7)：279-281.

Our letter

From：#Team2313161
To：New York Times
Date：February 20，2023

Dear Editor,

Thank you for hiring our team as your advisor! We have received your specified requirements and have fully evaluated the feasibility of our task. Here, we give you a detailed description of the problems we found and the corresponding solutions.

In the process of data analysis, we found that the number of reported results of Wordle reached a peak between 2022.1 and 2022.2. After the peak, the number of reported results gradually decreases, but there are still a group of loyal players who are still willing to play Wordle, which may be the reason why the change rate gradually decreases. In the process of analyzing number in hard mode, we found that in the later period of release, the change rate of this item of data decline gradually decreased, and even had an upward trend. It may be that a subset of challenging users gradually move from normal mode to hard mode.

Subsequently, in the process of analyzing the words posted each day, we found that the main attributes that determine the difficulty of a word are whether there are repeated letters in the word and whether the first letter of the word is common. According to such a rule, we find that the frequency of difficult words is small in the immediate aftermath of the puzzle's release, and the frequency of difficult words is increasing in the later stage of the release, which may be the reason for the gradual loss of users playing the normal mode.

In view of the above problems, we put forward the following suggestions for you:

a) According to our model, before releasing a new word each day, it’s necessary to predict the number of players in both categories in advance, and use this to predict the percentage of times that the word is tried, and select the words for which this percentage is more consistent with normality, which is not easy to make the percentage distribution appear in the situation of September 17, 2020 (most players fail to complete the game), resulting in user churn.

b) The difficulty of the words should be appropriately reduced, and the key words starting with the letters t, a, w, i and s should be increased, and the non-key words starting with the letters j, k, q, u, v, x and z should be reduced, so that the players playing the normal mode have a higher completion rate and attract more players.

c) Adding variations to the game (e.g., having separate words for normal and hard, increasing the number of letters in hard words, decreasing the number of hard questions, etal.) might help attract more players from normal to hard.

We sincerely hope that Wordle can regain its former popularity. If you want more details, please refer to our paper. We'd be happy to discuss the details of the solution with you.

Yours Sincerely,
#Team2313161