Problem Chosen	2023	Team Control Number
C	MCM/ICM Summary Sheet	2305317

Prediction of Wordle game

Summary

At the beginning of this research on Wordle, we need to define the attribute characteristics of words. Therefore, we have consulted the Collins Dictionary, Oxford University Dictionary and other relevant materials and information related to word research. After combining our personal experience in learning English words, we divided the properties of words, including: 1) the frequency of use of words 2) the number of vowels of words 3) the number of repeated letters of words.

We have established three sets of models. For question 1, in the process of preliminary observation of data, we believe that the number of reported results is stable and relatively stable, so we choose ARIMA time series model among many prediction models. According to the ARIMA model established, after the ADF test, the significance P value is less than 0.05, and the number of reported results is considered to be a stable time series. At the same time, the ARIMA model requires that the residual of the model should not be correlated. After checking the P value of the established model, it is considered that the ARIMA model does not have a white noise. After checking R ^{2} That is, after the fitting degree of time series, it is found that R ^{2}= 0.982 is close to 1, and the model performs well. According to the ARIMA model, the number of reported results in 2023/03/01 is 10456 person-times. According to the standard deviation obtained from the model and the false set reliability is 95%, the range of the number of reported results in 2023/03/01 is [9537,11367].

For problem 2, we need to predict the percentage distribution of players with different attempts for a certain word on a certain day, and for this we chose to use a BP neural network as our prediction model, and for the word uncertainty factor, we chose to use the number of repeated letters of the word and the word frequency as sample input features, and when we considered transforming the time uncertainty factor into the total reported score as one of the feature inputs, we found that the model fit was worse instead. The output feature is then the percentage of players with different number of attempts. The model shows a better performance after we trained it with 356 samples.However, due to the small training sample, the credibility of the model is not that high. By inputting the features of the word EERIE, we obtained its corresponding distribution roughly as 0, 1, 4, 19, 35, 30, and 11.

For problem 3, we want to develop a model that classifies words by their difficulty. Therefore, we first consider features that can be used to characterize the difficulty of words. The number of repeated letters of words and the word frequency of words used in the previous problem do not directly reflect the difficulty of words, so we choose to use the percentage of players with different attempts for each word as a characterization of word difficulty. We used the K-MEANS clustering algorithm to classify the words. In choosing the k value, we used the elbow rule and determined that clustering was best when k = 3. The profile coefficient used to evaluate the K-MEANS model was 0.361 and the CH value was 330.84, indicating that the model aggregation was better. By analyzing the clustering results, we found that the words in category 1 had the most players who attempted 4 attempts, the words in category 2 had the most players who attempted 5 or 6 attempts, indicating the most difficult, the words in category 3 had the highest percentage of players who attempted 3 attempts, indicating the easiest, and the word EERIE belonged to category 2, which is the most difficult class.

Keywords: WORDLE, ARIMA, BP Neural Network, K-MEANS cluster algorithm

Contents

1 Introduction 1
1.1 Problem Background . . . 1
1.2 Restatement of the Problem . . . 1
1.3 Our Approach . . . 1

2 Preparation of the Models 2
2.1 Assumptions . . . 2
2.2 Notations . . . 2

3 Data Preprocessing 3
3.1 Default Value Processing . . . 3
3.2 Abnormal Value Processing . . . 3

4 Model Construction 4
4.1 Using Autoregressive Integrated Moving Average model[ARMIA] . . . 4
4.2 Predicting Reported Results Percentages with BP Neural Network . . . 9
4.2.1 Back Propagation Neural Network . . . 9
4.2.2 Model in Details . . . 10
4.2.3 Model Results . . . 11
4.2.4 Model Analysis . . . 12
4.3 Word Classification Model . . . 12
4.3.1 K-MEANS Algorithm . . . 12
4.3.2 Model Details . . . 13
4.3.3 Model Results . . . 14

5 Strengths and Weaknesses 15
5.1 Strengths . . . 15
5.2 Weaknesses . . . 16

6 conclusion and other interesting findings 16

A letter to the Puzzle Editor of the New York Times 17

References 19

Appendices 20

1 Introduction

1.1 Problem Background

Wordle is a new popular puzzle provided daily by The New York Times. We need to guess a word.The way it works is that every time you submit a word that actually exists, the color changes, and yellow means the letter is in the answer, but in the wrong place. Green means the letter exists and is in the correct position, and gray means the letter does not exist in the answer word at all. Players can submit words in six chances and guess the answer. Wordle has grown in popularity, with many users now Posting their answers on Twitter. To do so, MCM collected background data on wordle answers from 2022 to 2023. Our team is going to complete the related tasks released by MCM.

1.2 Restatement of the Problem

1. The number of reported results vary daily. Develop a model to explain this variation and use your model to create a prediction interval for the number of reported results on March 1, 2023. Do any attributes of the word affect the percentage of scores reported that were played in Hard Mode? If so, how? If not, why not?

2. For a given future solution word on a future date, develop a model that allows you to predict the distribution of the reported results. In other words, to predict the associated percentages of (1, 2, 3, 4, 5, 6, X) for a future date. What uncertainties are associated with your model and predictions? Give a specific example of your prediction for the word EERIE on March 1, 2023. How confident are you in your model's prediction?

3. Develop and summarize a model to classify solution words by difficulty. Identify the attributes of a given word that are associated with each classification. Using your model, how difficult is the word EERIE? Discuss the accuracy of your classification model.

4. List and describe some other interesting features of this data set.

1.3 Our Approach

The question asks us to analyze a year's worth of data on wordle's answers in the background. And develop a model that can predict the answers on a given day in the future, including but not limited to the total number of answers, the success rate of 6 answers and so on. Therefore, the main work of our team consists of the following parts:

1. Build a model that can predict future answers based on the historical data of background answers in the past year.

2. Analyze whether the attributes of words will affect the answer scores through data and find out the exact reasons.

3. Build a model that classifies words as difficult.

4. Build a predictive percentage of future answers based on the difficulty of a word. And find out the uncertainty of the model prediction. Then predictive analysis is carried out through some random specific examples.

5. List other interesting features that describe the data set.

2 Preparation of the Models

2.1 Assumptions

To simplify the problem, we make the following basic assumptions, each of which is justified.

1. Assumption1.

There is a correlation between the total reported score for a word and the date.

2. Assumption2.

Words in the WORDLE game have the same word frequencies as real life words.

3. Assumption3.

No significant interaction between date and word.

2.2 Notations

Important notations used in this paper are listed in Table 1

Table 1: Notations

Symbol	Definition
t_i	Date and time
r_{t_i}	The number of reported results
h	The Heat Transfer Coefficient
c	The Heat Capacity of Water (=4200 J/(kg·°C))
ρ	The Density of Water (=10^3 kg/m^3)
Y_i	The i-th expected value
$f(x_i)$	The i-th predicted value of the output layer
L	The loss function
X_j	The coordinates of the jth data sample
C_{ij}	The i-th clustering center relative to the j-th data sample
n_k	the number of observations in cluster k
C_k	the centroid of cluster k
C	the centroid of the dataset
K	the number of clusters
X_{i_k}	the i-th observation of cluster k
N	number of observations
p	stratum
q	order
r	Difference order
Q_{BP}	The square of the residual
$D(r_t)$	variance
R^2	Goodness of fit

3 Data Preprocessing

3.1 Default Value Processing

We added some data to the calculation. The difficulty mode ratio is derived from the number of difficulty outcomes/number of reported outcomes. Add the number of repeated words of the word, the number of vowels of the word. We also obtained the data of word frequency according to the research of word frequency in Linse Dictionary, Oxford University Dictionary and so on. In order to facilitate the analysis and comparison of data, we conducted Zero standardization processing on some data.

3.2 Abnormal Value Processing

While processing the data, we also encountered some abnormal data. For example, the word rprobe does not exist and should be changed to probe. There are two more words of length 4, which will affect our data structure, so we have decided to remove them. Secondly, when we analyzed the time series, we found that the data on November 20, 2022 was abnormal compared to the surrounding data!

Such data, we are selective to delete retention. More detailed processing of the data will be discussed in detail later in the model building.

4 Model Construction

4.1 Using Autoregressive Integrated Moving Average model[ARMIA]

After processing all the data, our team found that the number of reported results of wordle was from low at the beginning to sudden increase to slow decrease, which can reflect two possibilities. The first one is that the number of reported results of wordle gradually increased, and then its popularity began to decline and slowly decreased. The second is that wordle's daily answer word difficulty affects the number of reported results. Although the first is obviously more likely, the second case will be explained later.

The graph below reporting the number of results versus time shows that there is a strong correlation between the two.

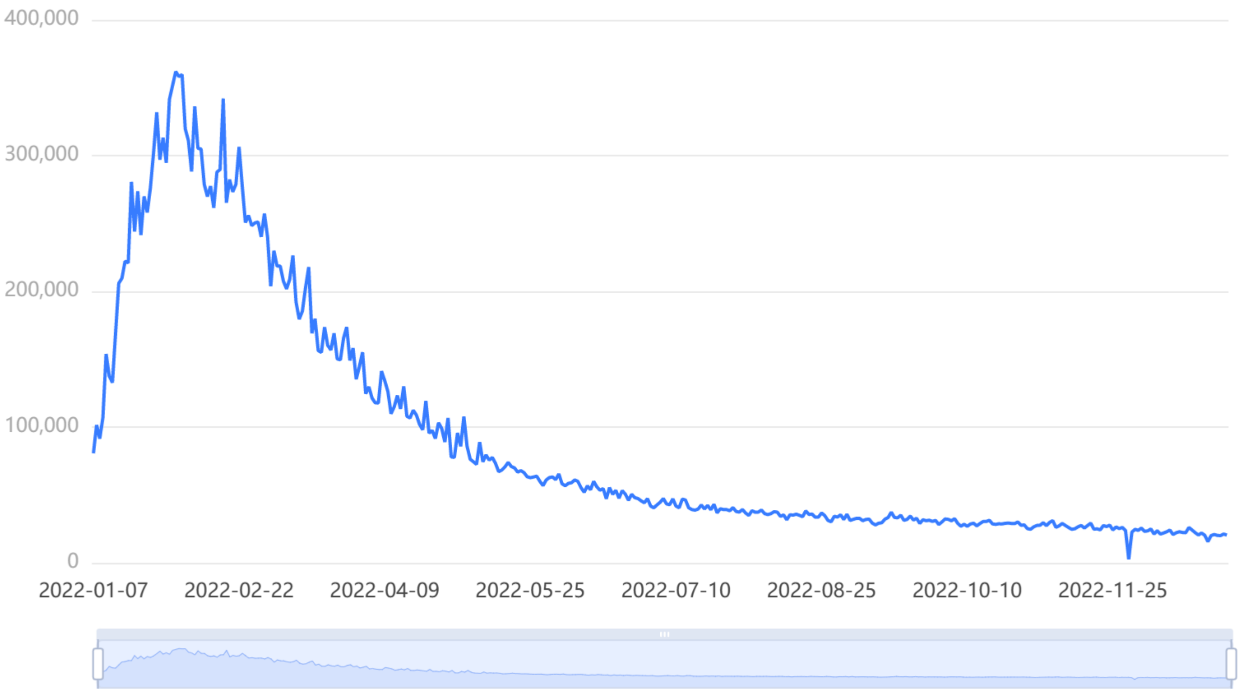


Figure 1: Number of reported results as a function of time

Since ARMIA uses a set of variables to observe at a series of moments, the time series of a series of discrete numbers fits the situation well. We will use the time series model to analyze and deal with this problem.

We need two kinds of data from excel, namely time, date and number of report results, which are represented by r_{t_i} and t_i:

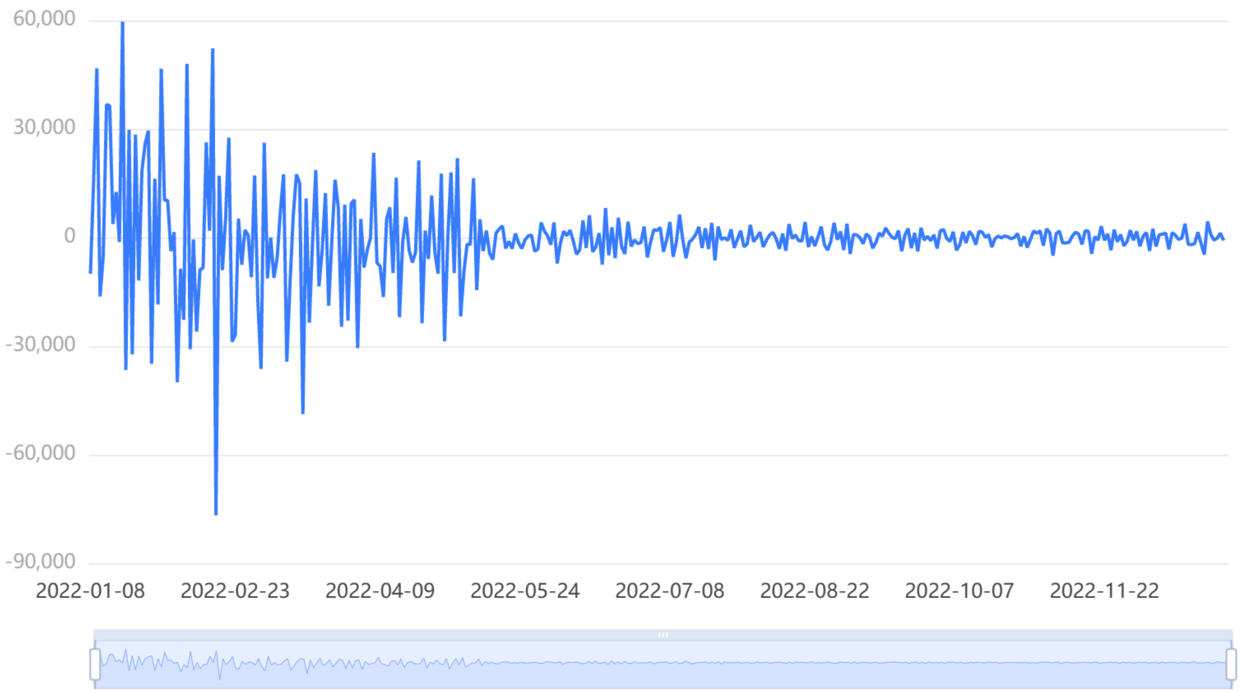
$$(t_1, t_2 \ldots\ldots t_k)$$
$$(r_{t_1}, r_{t_2}, \ldots\ldots r_{t_k})$$

Our team will first transform the original sequence model into a stationary model, and carry out n-order difference for the sequence:

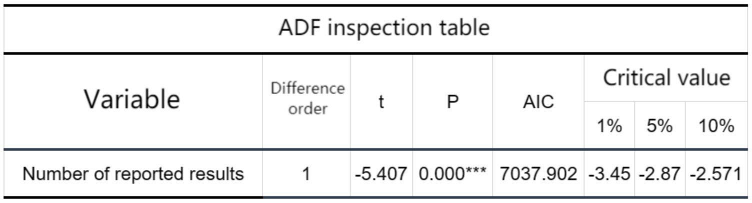
$$y_t = \phi_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + a_t + \sum_{i=1}^{q} \theta_i a_{t-i}$$

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

The first order difference processing was carried out on the original data, and the timing diagram was obtained as follows:

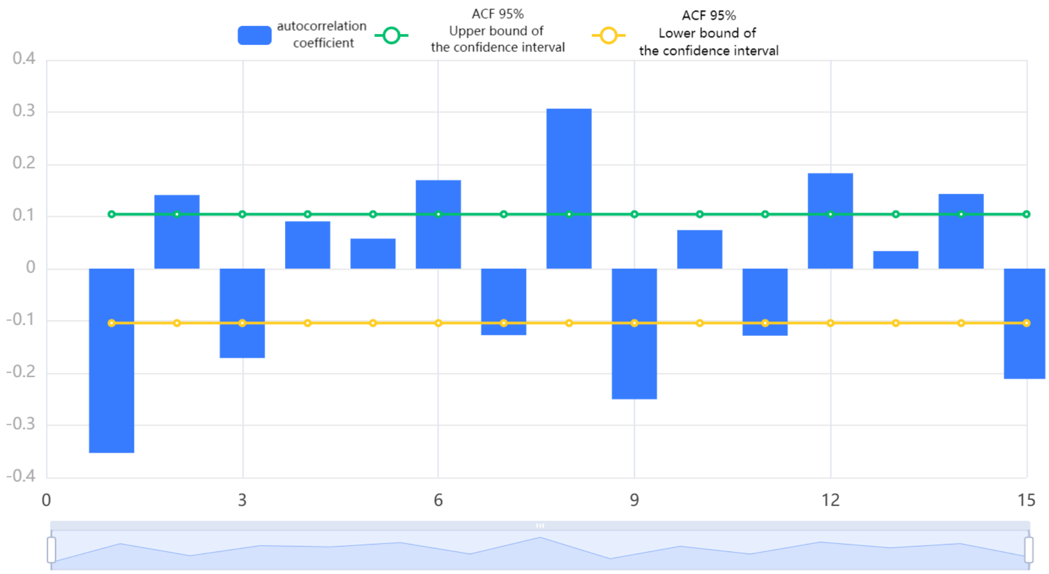


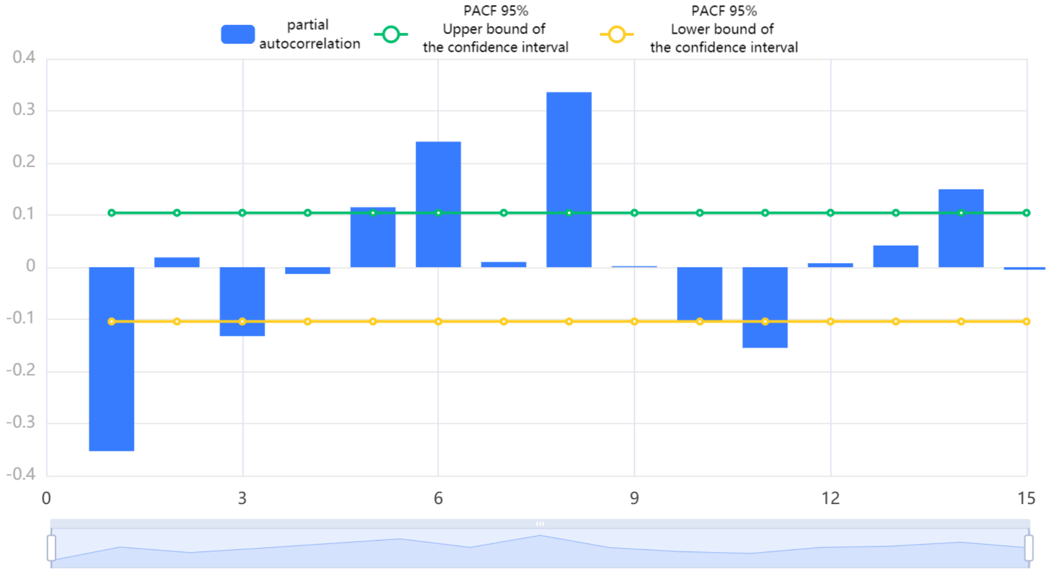
It is not difficult to see that the data after the first difference is visually stable. After passing the stationariness test (ADT), the P value is much smaller and 0.05, so the model presents significance. Then the series is a stationary time series, and the difference order is set as 1. (See the ADT test diagram)


ADF inspection table

Variable	Difference order	t	P	AIC	Critical value		
					1%	5%	10%
Number of reported results	1	-5.407	0.000***	7037.902	-3.45	-2.87	-2.571

Next, determine the level p and the order q. The autocorrelation and deviation correlation diagrams can be derived from the previous stationary sequence diagrams as follows.





It can be observed that the autocorrelation coefficient graph shows that there are 4 orders that exceed the confidence boundary, and the partial autocorrelation coefficient graph shows that the coefficient exceeds the confidence boundary when the second order to the eighth order, and then Narrows to 0. Using Ljung-Box Test method, calculate the square Q_{BP} of residual error:

$$Q_{\mathrm{BP}} = n \sum_{k=1}^{h} \hat{\rho}_k^2$$

Then, the goodness of fit R² is calculated as:

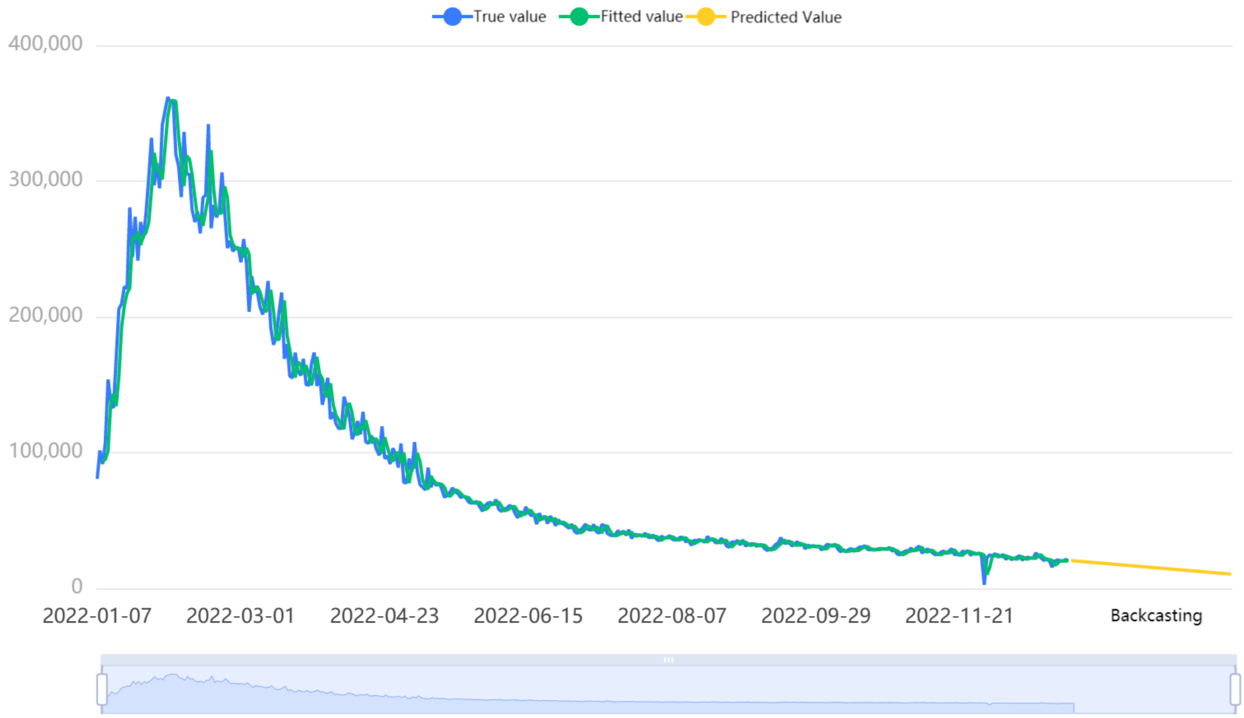
$$R^2 = 1 - \frac{Q_{BP}}{D(r_t)}$$

Where D (rt) is the variance of rt. After calculation, we get the goodness of fit **R²=0.982**, the model performance is excellent, basically meet the requirements. Therefore, the model parameter will be set to ARIMA (1,1,0).

After that, **ARIMA (1,1,0)** model in matlab was used for calculation, and the formula of the model was calculated as follows:

$$\mathrm{y(t)} = -168.296 - 0.362^{*}\mathrm{y(t-1)}$$

Plug in the parameters and simulate the number of reported results after 60 days as shown below:



Then calculate the upper and lower limits of prediction error. According to the table, coefficient k of 95% confidence is: 1.96, and the upper and lower limits are: 915.32 Finally, we predict that the number of results reported on the next 60 days, March 1, 2023, will be **[9540.68,11371.32]**.

Next, in order to solve the problem of whether the attributes of words will affect the ratio of difficult patterns, we will classify the attributes of words, which can be roughly divided into the number of repeated letters, the number of vowels, word frequency and word length. In most cases, the word length is 5, so it is not considered.

We classify words according to the rules of the game. Wordle's gameplay reveals something that looks like a test of luck, but is actually a test of analytics of words. Therefore, some words may be easy and some may be difficult, which affects the rate at which users choose the difficult mode. For example, if we guess a word with a repeating letter, it may be harder to tell if it contains that letter, and the placement of that letter may be more difficult, leading to a lower rate of difficulty patterns. For another example, if the number of vowels is more, the positioning of words will be more accurate and faster, and the ratio of difficult patterns may increase. The same is true for word frequency, and some words may be unfamiliar and affect users' answers. It is all possible! Of course, the above is only a rough analysis, as for whether there is any impact, still need to analyze the data to know.

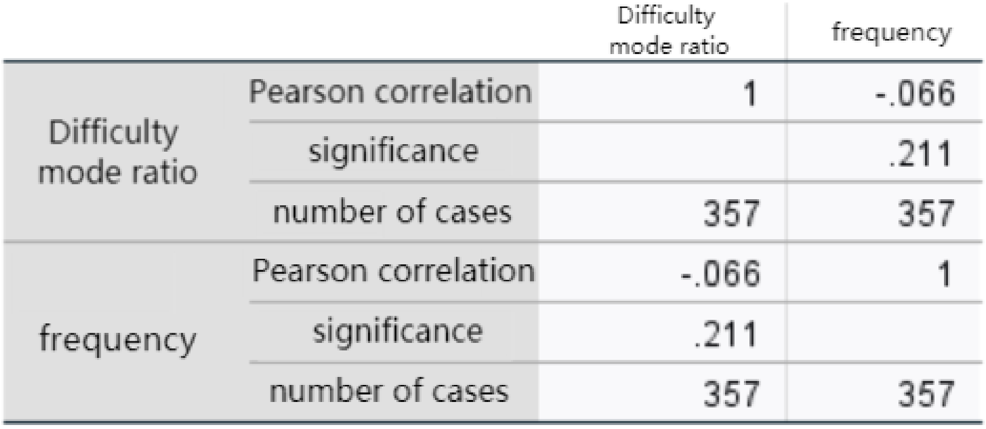
For example, the word photo has 1 repeated letters, 2 vowels, and 5 stars in word frequency. Therefore, under the influence of multiple factors, the difficulty ratio may be affected to some extent.

Next, we will analyze and calculate the two groups of data to confirm whether there is a correlation between them. A good correlation model is Pearson correlation coefficient. The relevant formula is:

$$Pearson = \frac{\sum_{i=1}^{p}(x_t - \bar{x_t})(y_t - \bar{y_t})}{\sqrt{\sum_{i=1}^{p}(x_t - \bar{x_t})^2}\sqrt{\sum_{i=1}^{p}(y_t - \bar{y_t})^2}}$$

Firstly, the correlation between difficulty pattern ratio and word frequency was analyzed.

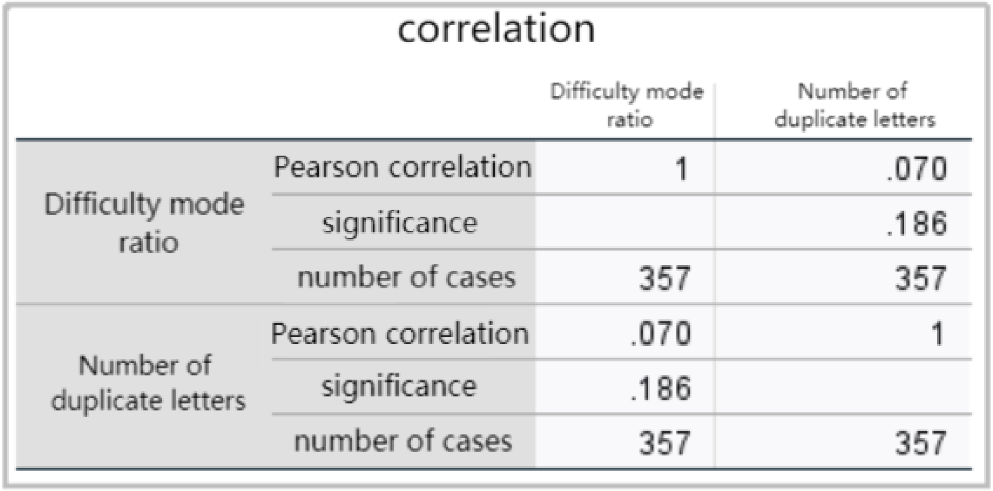
After the bivariate calculation of correlation in spssIBM, the following figure is obtained:



		Difficulty mode ratio	frequency
Difficulty mode ratio	Pearson correlation	1	-.066
	significance		.211
	number of cases	357	357
frequency	Pearson correlation	-.066	1
	significance	.211	
	number of cases	357	357

Among them, Pearson's correlation **P=0.066** , which is smaller than 0.1, indicates that the correlation between the difficult mode ratio and word frequency is very poor, and there is no obvious correlation between the two. Therefore, the word attribute does not affect the difficult mode ratio.

The other two attributes (the number of vowels and the number of repeated letters) showed similar results:



correlation

		Difficulty mode ratio	Number of duplicate letters
Difficulty mode ratio	Pearson correlation	1	.070
	significance		.186
	number of cases	357	357
Number of duplicate letters	Pearson correlation	.070	1
	significance	.186	
	number of cases	357	357

The Pearson correlation coefficient is very small, which again indicates that the difficulty pattern ratio is not affected by these two word attributes.

In summary, we conclude that any attribute of the word has no effect on the score percentage in difficult mode.

We take into consideration that the popularity of wordle game began to rise in 2022, the number of successful answers rose from thousands to tens of thousands, and then began to decline gradually, while the percentage of hard mode scores has been slowly increasing, which may be related to the number of players. There is a good argument that wordle's popularity slowly wanes, is a time for screening players, and the better players will probably stay and play the game, increasing the hard mode score by a slow increase. Of course, that's just a guess, and it's out of our scope of study.

4.2 Predicting Reported Results Percentages with BP Neural Network

4.2.1 Back Propagation Neural Network

For the second problem, in order to obtain the distribution of players with different attempts for a word at a future date, we thought that a supervised machine learning approach would be a good strategy to solve the problem. Our team used a more typical 3-layer BP neural network for prediction. The 3-layer neural network is an input layer, an implicit layer, and an output layer, and each layer contains multiple neurons that can be mapped between adjacent network layers.

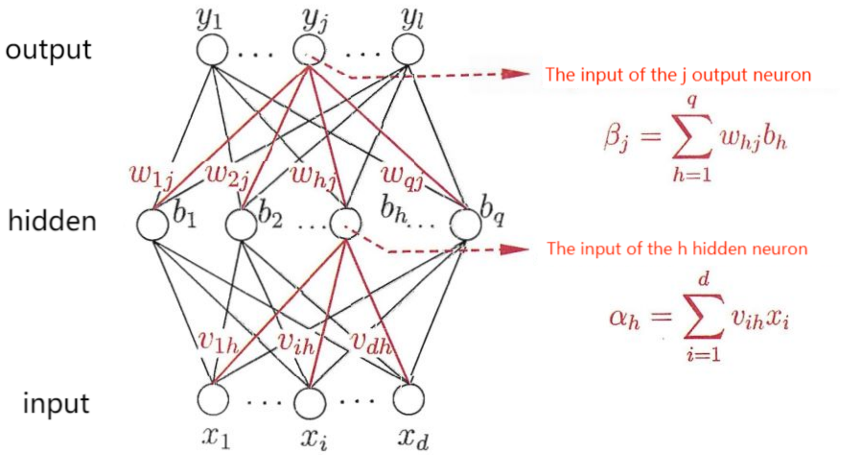

output
hidden
input
y_1
y_j
y_l
The input of the j output neuron
$\beta_j = \sum_{h=1}^{q} w_{hj} b_h$
w_{1j}
w_{2j}
w_{hj}
w_{qj}
b_1
b_2
b_h
b_q
The input of the h hidden neuron
$\alpha_h = \sum_{i=1}^{d} v_{ih} x_i$
v_{1h}
v_{ih}
v_{dh}
x_1
x_i
x_d

Figure 2: Neural network structure

Figure 2 illustrates a three-layer neural network in which the input layer receives three relevant data (features) from one sample during forward propagation, and the neurons in the hidden layer are a weighted combination of the neurons in the input layer, showing that the network places different emphasis on different features. The output layer is still a weighted sum of the values of the individual neurons in the hidden layer.

$$L(Y \mid f(x)) = \frac{1}{n} \sum_{i=1}^{N} \left(Y_i - f\left(x_i\right)\right)^2 \tag{1}$$

In the backpropagation process, Equation 1 shows the method for calculating the loss between the predicted and true values. The loss is calculated for the optimization of the parameters in the model, and the common method is the gradient descent method.The optimization algorithm updates the current parameters by multiplying the derivative value by the learning rate by the loss function to derive the derivatives for the corresponding parameters. A model with better parameters is obtained by continuously repeating the training.

4.2.2 Model in Details

The percentage of players with different attempts is highly correlated with words. Based on the first problem, we selected two features, the word frequency of words and the number of repeated letters in words, as the input of a sample, while we used the percentage of (1, 2, 3, 4, 5, 6, X) of the sample as the expected value of the output in order to let the neural network learn by itself.In total, we used 356 processed data for model training.That is, the input is a 356 X 2 matrix and the output is a 365 X 7 matrix.The dataset is divided randomly, and the training, validation, and test sets are **0.7, 0.2, and 0.1**, respectively. The neural network is trained using the Levenberg-Marquardt algorithm, which is mainly used for solving nonlinear least squares problems, and converges faster using gradient descent when the current solution is far from the optimal solution, and uses Gauss-Newton to slow down the convergence when it is close to the optimal solution. Figure 3 shows the structure of our neural network.

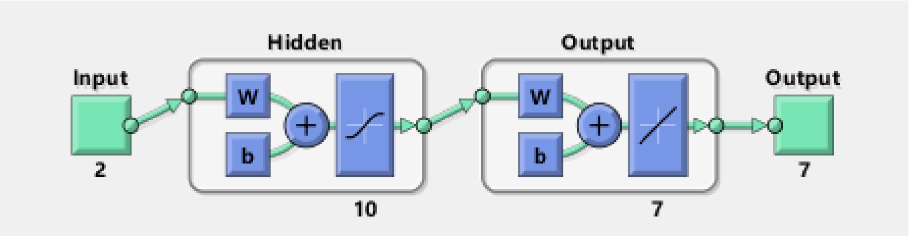


Figure 3: Model Structure

We also tried to consider the effect of the factor of time on our prediction results, and we entered the total reported score into the network as a feature of the sample as well, however its training effect was worse in most cases. Therefore we finally settled on the two features mentioned above.

4.2.3 Model Results

The model is preset for 1000 iterations and converges to the target value after 17 actual iterations, verifying that the model performs best in the 11th round. Figure 4 shows the error of model fitting on each part of the dataset and the total error. the closer the R is to 1, the better the model fits. the R value of our model is 0.9234, which indicates that the model fits better and the results are more credible. The figure shows the training state of our model. Figure 5 shows the training state of our model.For the word EERIE, we input its word repetition number and word frequency characteristics, and finally get the results: **0%, 1%, 4%, 19%, 35%, 30%, and 11%.**

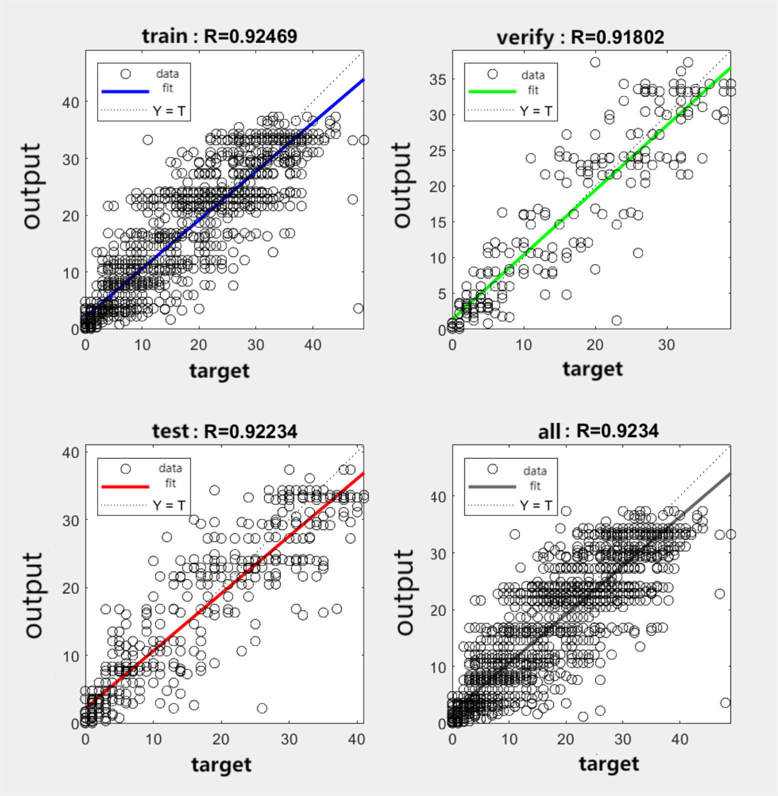


Figure 4: Errors of BP Neural Network Model

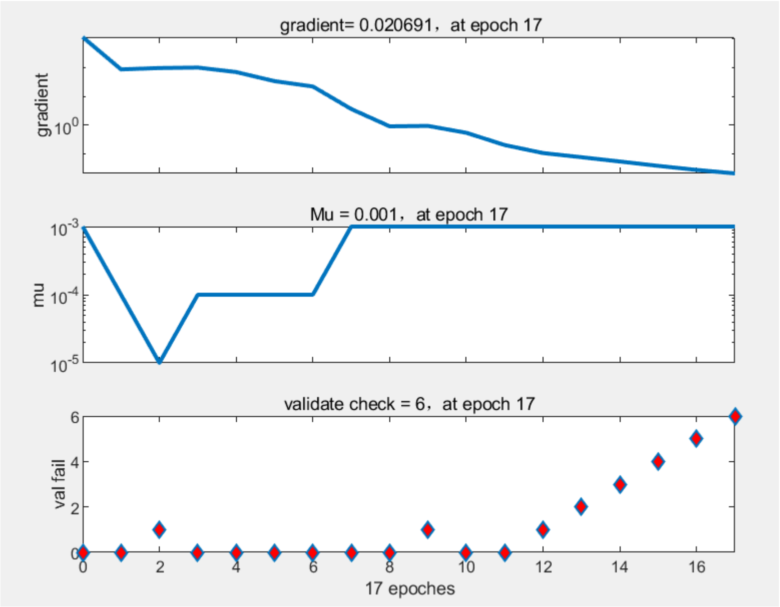


Figure 5: Train State

4.2.4 Model Analysis

There are still many uncertainties in the prediction, first of all about whether the date will affect the prediction results we can not yet make a clear interpretation, if the effect of the date is not considered, then the distribution of each indicator of a particular word can be considered the same on any day, however this will make the distribution of the word EERIE on March 1, 2023 required by the question less meaningful or redundant, because the distribution of the word EERIE on March 1, 2023. This is because the reported distribution of results is independent of time. However, the training model R-values are in most cases smaller and the model fits worse when we include the total score of the reported results.

Also, since the total amount of data is just over 300, which is still too little for training a neural network, the model may be difficult to fit well, and as we show the data, the R-value of the model is only 0.9234, which still has room for improvement. And we often get negative results during our simulations, which also convince us that our BP neural network is not yet good enough to make predictions. In 10 simulations, only about 1 or 2 results are normal, and the confidence of the model is about 0.1-0.2.

4.3 Word Classification Model

4.3.1 K-MEANS Algorithm

For word classification we use the clustering algorithm, which is an unsupervised machine learning algorithm that requires only inputting the feature values of the sample and the model can learn itself.

$$d\left(X, C_i\right) = \sqrt{\sum_{j=1}^{m}\left(X_j - C_{ij}\right)^2} \tag{2}$$

The most commonly used clustering algorithm is the k-means algorithm, the main step is to randomly select k objects from the data as the initial clustering center Ci, according to the mean value of each clustering object (the central object), calculate the Euclidean distance between each object and the central object, equation 2 shows the formula for calculating the Euclidean distance, according to the minimum distance to the object division, and then recalculate the object in each cluster of the The iterative process continuously reduces the Sum of Squared Error (SSE) of the class clusters, and the clustering is completed when the SSE no longer changes or reaches the maximum number of iterations.

4.3.2 Model Details

In selecting a reasonable value of k, we used the elbow rule. the K-means algorithm uses the square root of the error between the sample and the centroid as the objective function, and we can describe the degree of distortion of a cluster in terms of SSE; the lower the distortion, the tighter the cluster members are, and the looser they are anyway. We are able to obtain the SSE values in different cases by setting different k values to repeat the training model, and when the SSE at a certain point is significantly reduced, we can select that k value as the number of clusters.

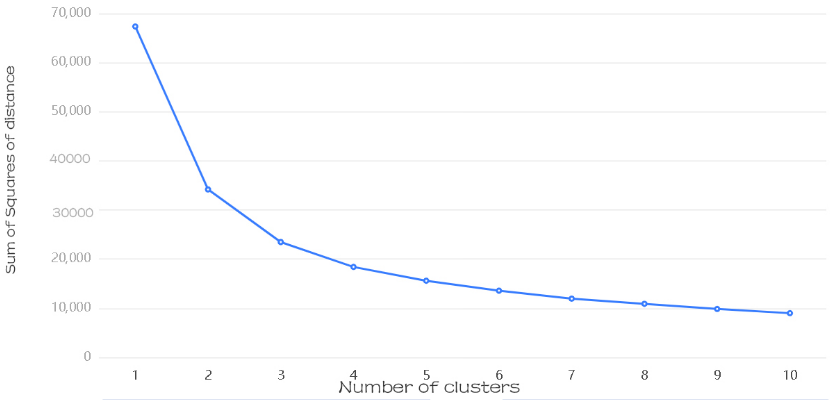


Figure 6: Comparison of the Number of Clusters

From the curves shown in figure 6, we can see that the SSE drops to a reasonable level at a k value of 3. The decline tends to be significantly slower, so we can be confident in choosing 3 as the number of clusters for our model.

In order to be able to classify words according to their difficulty, we still tried to choose some indicators to describe the difficulty of words. First of all, from the point of view of word characteristics, in the previous problem we chose the number of letter repetitions of words and the word frequency as word characteristics for prediction, however, word frequency obviously does not affect word difficulty, and we cannot determine for the time being whether the number of letter repetitions of words is related to word difficulty.

Based on this, we chose the percentage of players with different attempts of the words in the given data as a feature of each sample as a way to characterize the difficulty of the sample words.We let the model learn by inputting 7 features of a sample, a total of 356 samples, and used this to complete the clustering.

In evaluating the clustering effect we selected two metrics: contour coefficient, and CH score. The contour coefficient combines two factors: cohesion and separation. a_i indicates the average distance from sample i to other samples in the same cluster, the smaller a_i means that the sample is more closely related to the cluster, indicating the intra-cluster dissimilarity of the sample. b_ij indicates the average distance from sample i to all samples in other cluster C_j, the larger b_ij means that the less sample i belongs to the cluster, indicating the inter-cluster dissimilarity of the sample, from which we can define the contour coefficient as equation 3 .

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & a(i) > b(i) \end{cases} \tag{3}$$

Another metric is the Calinski-Harbasz Score , which calculates the sum of squares of the distances between points within a cluster and the class center to measure the intra-class tightness, and the sum of squares of the distances between cluster centroids and the dataset centroids to measure the separation of the dataset.The intra-cluster tightness we use BGSS to express, calculated as in equation 4 . The separation between clusters we use WGSS to express, calculated as in equation 5.

$$BGSS = \sum_{k=1}^{K} n_k \times \|C_k - C\|^2 \tag{4}$$

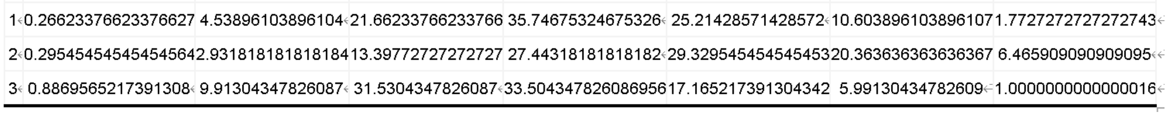
$$WGSS_k = \sum_{i=1}^{n_k} \|X_{ik} - C_k\|^2 \tag{5}$$

The CH value is obtained from the ratio of separation to tightness, and equation 6 shows how CH is calculated. a larger CH indicates better clustering.

$$CH = \frac{\frac{BGSS}{K-1}}{\frac{WGSS}{N-K}} = \frac{BGSS}{WGSS} \times \frac{N-K}{K-1} \tag{6}$$

4.3.3 Model Results

The data in the three classes of the final clustering accounted for 43.14%, 24.65%,32.21% of the total, respectively.The coordinates of the three clustering centers are shown in figure 7.



1	0.26623376623376627	4.53896103896104	21.66233766233766	35.74675324675326	25.21428571428572	10.603896103896107	1.7727272727272743
2	0.29545454545454564	2.9318181818181817	13.397727272727272	27.44318181818182	29.329545454545453	20.363636363636367	6.465909090909095
3	0.8869565217391308	9.91304347826087	31.5304347826087	33.504347826086956	17.165217391304342	5.99130434782609	1.0000000000000016

Figure 7: Cluster Centers

When we visualize the scatter distribution after clustering, we can use a graph 8 to show the final clustering results.

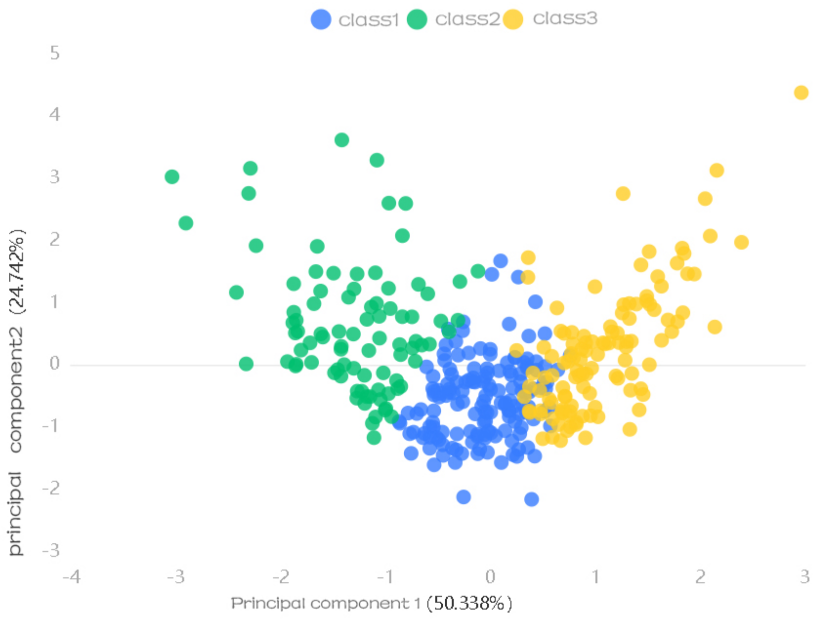


Figure 8: Clustered Scatter Plots

By calculation, the contour coefficient value of **our model is 0.361 and CH reaches 330.84,** and the clustering effect of the model performs well.

We looked at the categorical index of each sample, and through comparative analysis, we found that the samples in category 1 were mainly characterized by the highest percentage of players who tried 4 times, which means that the words in category 1 were moderately difficult, the samples in category 2 were mainly characterized by the highest percentage of players who tried 5 or 6 times, which means that the words in category 2 were the most difficult, and the samples in category 3 were mostly characterized by the highest percentage of players who tried 3 times. Clearly, the words in category 3 are relatively the easiest. Combining the distribution of the percentage of players with different attempts for the word EERIE obtained from the second problem, we can classify EERIE into **category2**, which means that EERIE belongs to the most difficult level of the game.

5 Strengths and Weaknesses

5.1 Strengths

- Strength1.

  Our ARIMA model has low white noise, stable data and good accuracy.

- Strength2.

  The prediction model for the total score of word reporting has good robustness

- Strength3.

  High accuracy and reliability of word classification model results

5.2 Weaknesses

- Weakness1.

  The number of reported results has a relatively large growth rate in January 2022. In essence, the ARIMA time series prediction model can only capture the linear relationship, but not the nonlinear relationship.

- Weakness2.

  BP neural network prediction results may be wrong, and may be accurate once in many predictions

- Weakness3.

  The model may not consider certain features in the prediction process, resulting in poor model prediction accuracy.

6 conclusion and other interesting findings

Conclusion I *We found out that if the Wordle column does not change in the future, the number of reported results will continue to decline, and it is predicted to decline to 10456 person-times in 2022/03/01.*

Conclusion II *We get the conclusion that the length of words, the number of repeated letters of words and the number of vowels of words had no effect on the proportion of people who challenged the difficult mode*

Conclusion III *According to the constructed BP neural network, we believe that the proportion of people who try from 0-7 (or more times) is 0%, 1%, 4%, 19%, 35%, 30%, and 11% respectively on the day of 2023/03/01.*

Conclusion IV *Through the clustering algorithm, we find that the difficulty of words with 4 attempts is medium, and the difficulty of words with 5-6 attempts is difficult, The difficulty of words with less than 3 attempts is simple. Similarly, we assume that the word on the day of 2023/03/01 is EERIE. According to the model, we believe that the difficulty of the word is high.*

Interesting finding I *The number of reported results showed an exponential growth in January 2022, and then began to decline month by month, indicating that Wordle game once attracted many players because of its novelty in the promotion process, and then many players lost because of its simplicity.*

Interesting finding II *The proportion of people challenging the difficult mode shows an increasing trend, indicating that players are more willing to challenge the difficult mode in the process of mastering the Wordle game, indicating that people are full of challenge spirit.*

A letter to the Puzzle Editor of the New York Times

Dear the Puzzle Editor of the New York Times:

We are college students who have a strong interest in the column wordle. After analyzing the previous problems of Wordle, I have come to the conclusion that I hope to share with you.

Firstly, from the published data, we can clearly find that the number of reported results showed a rapid growth trend in January 2022, reaching the peak of 361908 at one time, and then decreasing month by month. According to the existing data, the number of reported results fluctuates around 22000 on a single day, we decided to use Autoregressive Integrated Moving Average model to analyze. We found out that if the Wordle column does not change in the future, the number of reported results will continue to decline, and it is predicted to decline to 10456 person-times in 2022/03/01. According to the data, the number of people challenging the difficult model is similar to the number of reported results. Both of them showed a rapid growth trend in January 2022, and then decreased month by month. Through time series prediction and analysis, we predict that the number of people challenging the difficult mode on the day of 2022/03/01 will be 2008. It can be observed from the data that the proportion of people who challenge the difficult model is increasing month by month, which indicates that participants are more willing to challenge and full of challenge spirit after completing many times of wordle.

Next, we try to explore the impact of attributes of words on the proportion of people who challenge the difficult mode. We assume a variety of attributes of words, including the length of words, the number of repeated letters of words, the number of vowels in words and the frequency of use of words. We counted the relevant characteristics of words in 2022/01/07-2022/12/31, and then used Pearson correlation analysis and Spearman correlation analysis to get the conclusion that the length of words, the number of repeated letters of words and the number of vowels of words had no effect on the proportion of people who challenged the difficult mode. Then we find Harper Collins's word use frequency table and classify words according to the word use frequency. Divide the words appearing in the past into six levels, and the frequency of using words increases from 0-5. We use the typical correlation analysis to find that the frequency of using words has no effect on the proportion of people who challenge the difficult mode. To sum up, we believe that the proportion of people challenging difficult patterns has nothing to do with the characteristics of words.

Then we try to use the supervised three-layer BP neural network to analyze the distribution of players who try different ways on a certain day. In the BP neural network, we take the word frequency characteristics and word letter repetition times of 2022/01/07-2022/12/30 as input, and obtain an accurate model after 1000 iterations of the neural network. We assume that the word entered on the day of 2023/03/01 is EERIE. According to the constructed BP neural network, we believe that the proportion of people who try from 0-7 (or more times) is 0%, 1%, 4%, 19%, 35%, 30%, and 11% respectively.

In addition, we also try to classify the difficulty of each word by using word characteristics. We use the percentage of players with different attempts as a factor to describe the difficulty of a word. Using K-MEANS clustering analysis algorithm, we cluster players with different attempts and find that players can be clearly divided into three categories. Through the clustering algorithm, we find that the difficulty of words with 4 attempts is medium, and the difficulty of words with 5-6 attempts is difficult, The difficulty of words with less than 3 attempts is simple. According to the characteristics of clustering, we can help you adjust the difficulty of the word. Similarly, we assume that the word on the day of 2023/03/01 is EERIE. According to the model, we believe that the difficulty of the word is high.

Based on the data analysis of wordle, we hope to make some suggestions. 1) We hope to modify the rules of the Wordle game, add more interesting playing methods, and improve the fun of the Wordle game. 2) Accumulate points for players who have completed the Wordle game. Rewards can be provided when the points reach a certain amount. 3) Wordle games can be divided into difficulty gradients, providing the difficulty of 4 alphabetic words (easy), 5 alphabetic words (medium), and 6 alphabetic words (difficult).

Finally, I hope the Wordle game will be better and better, because it will help us improve our love of English words.

Students who love the Wordle game 2023/02/20

References

[1] Bailyn M., A Survey of Thermodynamics. *American Institute of Physics*, 1997:23.

[2] Louis C.B., Convective Heat Transfer (2nd ed.). *Wiley-Interscience*, 2003:107.

[3] Frank I., Theodore L.B., David D., Adrienne S.L., Fundamentals of Heat and Mass Transfer (6th ed.). *John Wiley & Sons.*, 2007:260-261.

[4] Marek R., Straub J., Analysis of the Evaporation Coeffcient and the Condensation Coeffcient of Water. *International Journal of Heat and Mass Transfer*, 2001(44):39-53.

[5] Tao W.Q., Heat Transfer (4th ed.) (in Chinese). *Higher Educational Press*, 2006:2-5.

[6] Lu J.A., Shang A., Xie J., Gu P., MATLAB-Solution of Partial Differential Equations (in Chinese). *Wuhan University Press*, 2001.

[7] Zhu Xianhui,Yu Yue,Shi Nan,Xu Liang,Jian Youwei.A study on hierarchical optimization of BP neural network and its application in wind power prediction[J]. High Voltage Electronics,2022,58(02):158-163+170.DOI:10.13296/j.1001-1609.hva.2022.02.021.

[8] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software].

[9] Wang Yan. Application of time series analysis [M]. Beijing: China Renmin University Press 2005.

Appendices

Appendix1

```
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf',C=0.8, random_state = 0,class_weight={0:20,1:24})
#classifier = SVC(kernel = 'rbf',C=1.2, random_state = 0)
classifier.fit(X_train, y_train)
# Predicting the Test set results
y_pred = classifier.predict(X_test)
```

```
    import numpy
    import pandas
    from spsspro.algorithm import statistical_model_analysis
    data = pandas.DataFrame({
    "A": numpy.random.random(size=20)
    })
    result = statistical_model_analysis.arima_analysis(data=data, p=1, d=1, q=0,
        forecast_num=10)
    print(result)
```