Problem Chosen	2023	Team Control Number
C	MCM/ICM Summary Sheet	2314475

Predicting Wordle Results

Summary

As the outbreak of the COVID-19 in the wordle puzzle game was spread internationally at the end of December 2019, many users posted their responses on TWitter. This paper mainly studies the characteristics of this game in time series, and describes the word attributes that appear in the game.

For task 1, this paper first establishes a description model of time series based on the life cycle theory. We propose a piecewise linear trend extraction algorithm, which takes the local minimum RMSE as the target to find the appropriate sliding window sequence. Finally, the time series is divided into four segments by clustering, namely, the rising period, the saturation period, the falling period and the stable period, with an average RMSE value of 2563, Finally, the stationary period series of the model is used to predict the 03.01 data as 19368.57. The problem of incomplete use of time series data has not been solved. Our ARIMA (0,1,13) - GARCH (1,1) model complements the prediction part with a goodness of fit of 0.986. We also forecast the 03.01 data to get the 20982 report results. The final result of the correction of the results of the life cycle model is 20175, with a 95% confidence interval of [14689,28354].

In order to study the words in the game, this paper first establishes the continuous index word frequency, word orthogonal number and letter repetition number, as well as the classification index word professionalism, foreign word judgment, word part of speech (noun verb adjective), and uses the Spearman correlation coefficient to calculate the correlation between the percentage of difficult patterns and the word attribute. Our conclusion is that the two are not relevant under the 99% confidence interval, Only the word frequency has a moderate correlation of 0.311.

For task 2, this paper uses the partial least squares regression model to calculate the percentage of different answer times using the previous word attributes. Because there are many discrete data, this paper uses Apriori algorithm to establish the correlation rules between variables and word difficulty, and establishes a mixed index according to the confidence of its rules. The independent variables in the regression model include word frequency, orthogonal number, letter repetition number, and mixed index. The final prediction result is (1=0%, 2=2%, 3=15%, 4=33%, 5=30%, 6=16%, X=4%). The uncertainty factors are social media, contemporary news, global epidemic, etc. The lowest root-mean-square error of the model can reach 0.74.

For task 3, we first use the seven percentages of words to cluster with the system clustering, divide 359 words into three difficulty levels of simple, moderate and difficult, and then use Fisher linear discriminant analysis to judge the classification results with 7:3 training sets and test sets, and draw the conclusion that the clustering accuracy is about 95%, and the model classification ability is strong, Then the seven percentage data of EERIE words predicted in the second question are substituted into the discriminant equation, and the result that EERIE belongs to the first class, namely, simple class, is obtained.

Finally, the sensitivity test is carried out. This paper tests the sensitivity of window threshold, regression equation coefficient, model coefficient, etc., to verify the robustness of the model and the rationality of coefficient selection.

Keywords: Life cycle theory;Piecewise linear trend extraction algorithm; ARFIMA-GARCH Model; correlation coefficient;Apriori algorithm;PLSR;

Contents

1 Introduction

1.1 Background

Wordle, a word guessing game rose to global popularity in the December of 2021[2]. The goal of the game is to guess a five-letter English word within six attempts.Each attempt will provide a prompt to the player through the color change brick, telling the player whether a character is part of the solution and whether it is in the correct position.

In view of the popularity of the game, many people have studied the game, and the medical field has begun to analyze the value of skill development in participating in this daily intelligence activity;At the same time, people are constantly studying how to maximize the game's winning and solve daily challenges.The work by Kandabada [3] suggested manually selected four. words, [SPORT, CHEWY, ADMIX, FLUNK], as the best starting words. Sidh[4] searches forthe best starting word from the perspective of linguistics.Anderson and Meyer[5] use machine learning to find the best strategy for solving puzzles.

1.2 Problem Summary

Wordle is divided into regular mode and hard mode. In the hard mode, players are required to use the correct letter in a word (the tile is yellow or green) in subsequent guesses, which greatly reduces the efficiency of the strategy. We need to use the data set *Problem_C_Data_Wordle.xlsx* provided by MCM, which includes the scores reported by users on Twitter from January 7 to December 31, 2022, Conduct multi-angle analysis on the popularity of the game and the difficulty of daily word titles.

After through in-depth analysis and research on the background of the problem, we can specify that our article should cover the following aspects:

- Establish a model to describe the change of report results, and uuse the model to create a forecast interval for the number of reported 2023 on 1 March.
- This paper explores the influence of the attributes of words on the percentage of players' scores in difficult mode, and discusses the establishment of a regression model to determine the average number of answers of different words, and explores the reasons for the influence of different attributes on players.
- Develop a model to predict the proportion of results players will get on a given solution word at a future date, and explore the prediction of player results for the 2023 word EERIE on March 1, the prediction accuracy of the model is analyzed and evaluated.
- This paper sets up an evaluation model of the word difficulty, classifies the words in the game according to the difficulty, carries on the feature engineering to establish the word classification attribute, and evaluates the word EERIE according to the difficulty, the accuracy of our classification model is also discussed.
- Mining interesting features in a dataset.

- Finally, summarize our results in a one- to two-page letter to the Puzzle Editor of the New York Times.

1.3 Data Sources

Problem_C_Data_Wordle.xlsx contains the result data submitted by players on Twitter every day from January 7 to December 31, 2022,which includes the date, contest number, word of the day, the number of people reporting scores that day, the number of players on hard mode, and the percentage that guessed the word in one try, two tries, three tries, four tries, five tries, six tries, or could not solve the puzzle (indicated by X).

2 Ourwork

3 Assumptions and Symbols

3.1 Assumptions

- **The player result data is stable to a certain extent**: the data will not fluctuate violently, so if the data on a certain day is too different from the surrounding data, it will be considered as error data.
- **The situation that the player fails to pass the customs is considered as seven times**: if the player fails to find the answer six times in the game, the system will give the correct answer, and the player will definitely find the answer the seventh time.
- **The word difficulty in the Wordle game is considered to be evenly distributed**: therefore, we can completely rely on this data set to divide the word difficulty.

3.2 Symbols

Symbols	significance
X_t	Time series
ρ_k	Autocorrelation function
Y_{t-1}	The observations in the t time series
$\hat{Y}_t$	The predicted value under the t time series
u_t	Residual term
σ_t^2	the conditional variance

4 Task1 Report quantity interpretation model (based on life cycle theory)

This paper introduces the life cycle theory to describe the number of changes in the results reported by players in Wordle from January 7, 2022 to December 31, 2022. We use the piecewise linear trend extraction method to extract the shape of the model and then cluster it to identify that the time series model is divided into four stages: growth, maturity, decline and stability.

4.1 Preprocess Data

- **Exception value handling** We observe the existence of four-word abnormal words, because the amount of data is not large, we want to use all the data as possible, so we will replace the correct word;

We also found that one day there was an abnormal mutation in Number of reported results, which is not a normal change in the Number of people playing a game over time, and we corrected it by using spline interpolation.

- **Data Smoothing**

First, smoothing is performed to remove weekly seasonality. The overall trend of a series can be better identified if the impact of low-level fluctuations due to weekly seasonality is removed. Thus, a moving average with a window size of 7 days is applied to each player population series to extract the first approximation of the trend by excluding the impact of weekly seasonality. Here a value at a certain point of the series is approximated by calculating the mean of values within a 7 days window surrounding that data point.Figure 1 is the time series after smoothing.

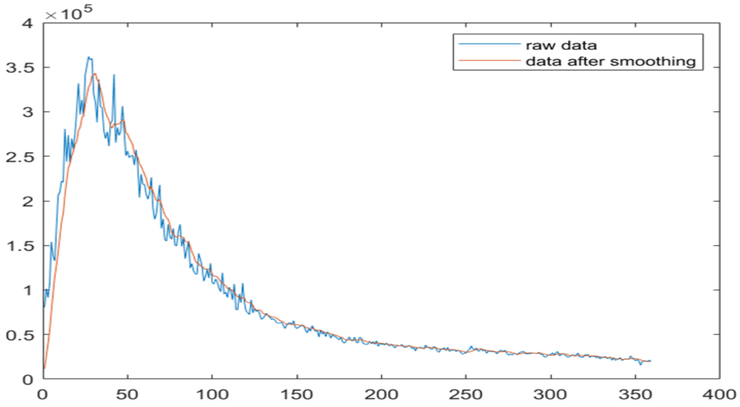


Figure 1: Smooth series to eliminate seasonal trend

- **Data Normalization**

Since the number of players in the game changes greatly, this paper standardizes the data, and the number of submitted results is between 0-1, which provides the basis for the subsequent division of the life cycle of time series. The standardization formula 1 is as follows:

$$\bar{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

4.2 Piecewise Linear Regression for Life Cycle shape extraction

Because we do not know how many segments the time series should be divided into, we use the segmented unknown time series extraction algorithm [1].

The algorithm we designed can extract the time series linearly according to its characteristics by iterative method, take one month (30 days) as the shortest time series, use the sliding window to continuously add a single life cycle segment, take RMSE as the local minimum as the target, and extract the segment. Figure 3 is the schematic diagram of the sliding window.

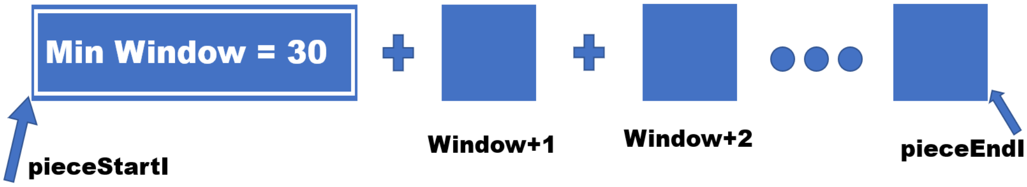


Figure 2: Schematic diagram of sliding window

Find out that the reason why RMSE is the local minimum is that there is a certain contradiction between the error and the length of the time series to a certain extent. Therefore, it is necessary to balance it in the algorithm, and find out that the error of window length+1 (RMSE) is less than or equal to the error of the original window length, so as to find the length of the time series with the local RMSE minimum.

At the same time, in the time series division, we also observed that part of the sequence will increase the error rate with the increase of the window length, which is not consistent with our previous criteria. In view of this situation, we decided to relax the criteria of the local optimal solution and judge it by the artificially given RMSE threshold. The function(2) are as follows:

$$\frac{error_{i+1} - error_i}{error_i} \leq RmseThreshold \tag{2}$$

According to formula(2), when the RmseThreshold value is set to 0, it is the first case mentioned in this paper. With the increasing of the threshold value, it can be assumed that the window of the local optimal solution will also increase. Therefore, to prevent the error from increasing with the increase of the window sequence, we select the RmseThreshold value to 0.01 to avoid excessive increase of the window length and increase the time complexity.

4.3 Results and Discussion

In this section, we use the time series extraction algorithm and clustering algorithm mentioned above to divide the time series of Worlder games into four stages: up, saturation, down and stationary, as shown in Figure 3.

Our article serializes time to X_i,for example, 2022-1-7 is converted to X_1=1, and so on,converting 359 date values to X_i(i=1,2,3...359),and set the number of reports to $f(X_i), i = 1, 2, 3...359$. Figure 4 shows the curve fitting of four time series. In the fitting, we choose the method with the greatest goodness of fit for each curve.

Rising period(X_i = 1 12)**:**f(x) = $1.396 * 10^4 x + 6.202 * 10^4 \quad R^2 = 0.9256$

Saturation period(X_i = 13 56):

f(x) = $2.875 * 10^5 - 2.226 * 10^4 cos(0.137x) - 2.956 * 10^4 sin(0.137x) \quad R^2 : 0.559$

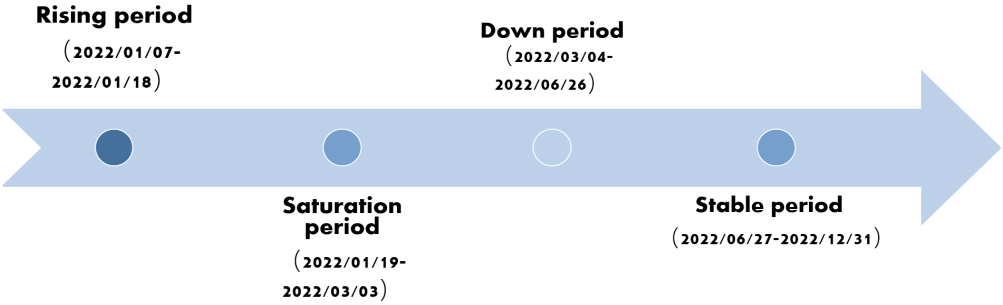
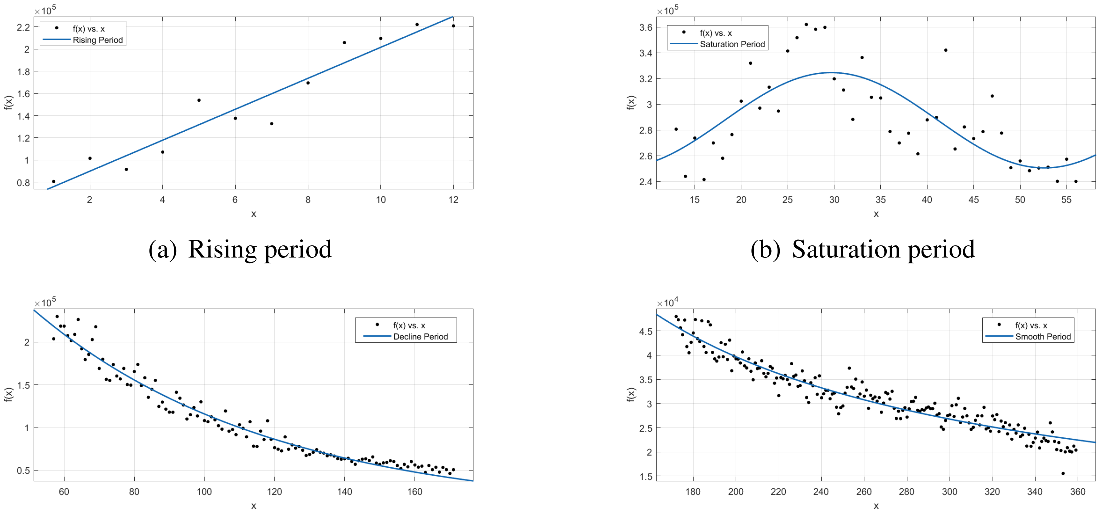


Figure 3: Life cycle time division



(a) Rising period

(b) Saturation period

(c) Down period

(d) Stable period

Figure 4: Life cycle curve fitting

Down period(X_i = 57 171): f(x) = $5.069 * 10^5 exp\,(-0.01477x)$ $\quad R^2 = 0.9648$

Stable period(X_i = 172 359): f(x) = $6.722 * 10^6 x^{-0.96882} = 0.9171$

5 Time series supplementary model

Based on the existing data, the ARIMA model is used to establish the time series regression model of the number of reported results, and to predict the number of reported results in March 1,2023. Results were predicted with 95% confidence intervals for the day.

5.1 Establishment of ARFIMA(p,d,q) model

5.1.1 Stability Test

In the stationary test, the single root test (ADF) is used.The ADF test is to determine whether a sequence has a unit root: if the sequence is smooth, there is no unit root; otherwise, there is a unit root. Therefore, if the significance test statistic is less than three confidence levels (10% , 5% , 1%) , then there should be (90% , 95% , 99%) confidence to reject the original hypothesis.

The table1 is the result of ADF test, including the results of variables, statistics, p value, etc. which used to check whether the sequence is stable.

Table 1: Stability Test

Variable	t	P	Threshold		
			1%	5%	10%
USD	-1.256	0.6490	-3.451	-2.876	-2.570

As shown in Table 1, the p value is significantly greater than 0.05, indicating that the time series data is not stable. At the same time, according to the results of the t-test, the original hypothesis is rejected at the significance level of 1%, 5% and 10%.

5.1.2 First Order difference

Because of the instability of time series data, the first-order difference processing is used to transform the time series formed by adjacent periods, that is, subtracting the previous period from the latter period[6].

$$y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i \tag{3}$$

Using Formula (3), we carry on the difference processing to the data, figure5 is the difference time series chart, can see has tended to be smooth, in the ADF test p value is obviously less than 0.05, at the same time, t-test in 1% , 5% , 10% also accept the original hypothesis, so the time series is stable.

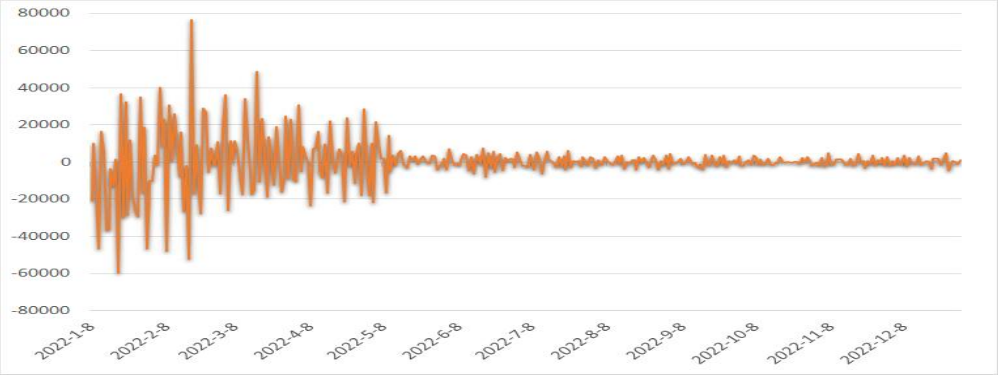


Figure 5: First Order post-difference time series

5.1.3 Build ARFIMA(0,1,13) model

If the time series $X_t, t \in T$ is stable and satisfies the following equation[7]:

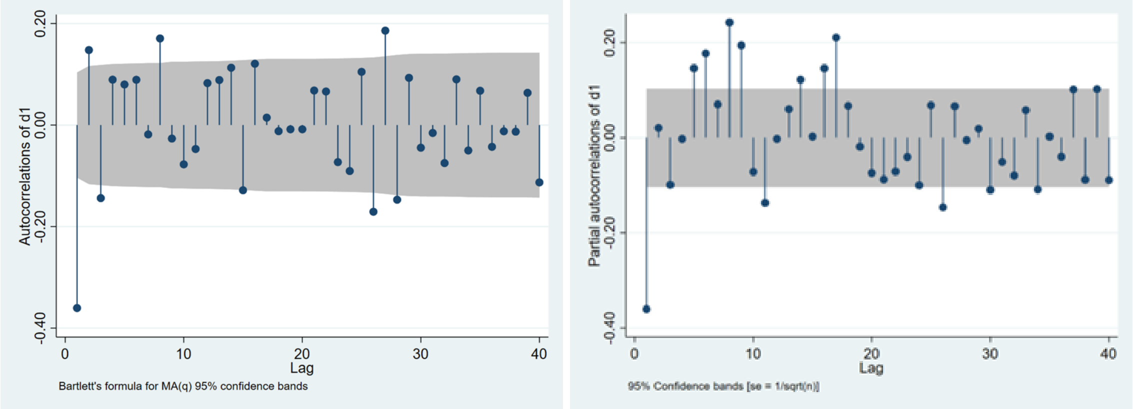
$$\Phi(B)(1-B)^d X_t = \Theta(B)\varepsilon_t, \varepsilon_t \sim N\left(0, \sigma^2\right), t = 1, 2, \cdots, T \tag{4}$$

In addition, for the autocorrelation function, there is a constant $c > 0$. When $k \to \infty$, there is the following relationship:

$$\rho_k \sim ck^{2d-1}, \tag{5}$$

therefore $\{X_t, t \in T\}$ belongs to $ARFIMA(p, d, q)$ process .

For a time series, there may be more than one significantly effective model. At this time, the best values of parameters p and q can be selected by graph of autocorrelation and partial correlation. In general, the upper limit of parameters p and q can be $\sqrt{n}$ or n/10, but in practical applications, the value of p and q generally does not exceed 2.



(a) Autocorrelation

(b) partial correlation

Figure 6: Autocorrelation and partial correlation

From figure 14 we can see that the parameters of the Arima model are p = 0,q= 13 and d = 1 because we use the first order difference. The $ARFIMA(0, 1, 13)$ model was selected ,the least squares estimation result of the model is as follows:

$$y_t = -31.07024 + \varepsilon_t - 0.670\varepsilon_{t-1} + 0.122\varepsilon_{t-2} - 0.245\varepsilon_{t-3} + 0.119\varepsilon_{t-4} + 0.133\varepsilon_{t-5} + 0.076\varepsilon_{t-6} + 0.041\varepsilon_{t-7} + 0.133\varepsilon_{t-8} - 0.109\varepsilon_{t-9} - 0.034\varepsilon_{t-10} + 0.124\varepsilon_{t-11} + 0.272\varepsilon_{t-12} + 0.290\varepsilon_{t-13} \tag{6}$$

where y_{t-1} represents the number of reported results in the previous moment,ε_i represents Residuals at i moments.The goodness of fit of the model is 0.986, which shows that the model has a good fitting effect.

5.2 ARIMA(p,d,q)-GARCH model

In statistics, generally, the assumptions about random perturbation in the time series model are mutually independent and identically distributed, but in the actual modeling process, this assumption is not tenable. The variance of the residual will not only change and have a certain degree of focus. The trend of the fluctuation is often aggregated, and the large fluctuation is followed by the larger fluctuation. This phenomenon is also called heteroscedasticity because the variance of the residual changes. Therefore, this paper uses the GARCH model with autoregressive terms to extract the relevant information left in the residual to predict the square term of the residual[8].

The emergence of GARCH model solves the problem that the traditional autoregressive conditional heteroscedasticity model can not be used to predict the situation with many lagging orders.

5.2.1 A Lagrange multiplier test for the ARCH effect

For many financial time series, especially those related to the stock market, the size of the residual is often related to the nearest residual value, that is, this kind of time series is prone to the phenomenon of conditional heteroscedasticity, that is, the conditional variance σ_t^2 of the residual term u_t depends on the size of u_{t-1}^2 . According to the observation in Figure **??**, it can be noticed that the fluctuation of residual is small in one period of time and very large in another period of time, and there is a phenomenon of "clustering". It can be preliminarily judged that there is a phenomenon of conditional heteroscedasticity.

We use LM test method,if the residual u_t in the mean model satisfies the ARCH (q) process,the following equation (7) can be established:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i u_{ti}^2 \tag{7}$$

If the probability of the coefficients in equation (7) being all zero is large, there is no ARCH effect. In LM test, its hypothesis is:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_q = 0,$$

$$H_1 : \exists \alpha_i \neq 0 (1 \leq i \leq q).$$

The test statistic is,

$$LM = nR^2 \sim \chi^2(q) \tag{8}$$

If $LM > \chi^2_\alpha(q)$, then refuse H_0, thinking exists ARCH effct.

According to the test results, the p value of the residual sequence is significantly equal to 0 at about 20 orders, indicating that the residual sequence of the ARFIMA (0, 1, 13) equation has ARCH effect, so the residual square term of the mean model is applicable to the GARCH model for fitting.

5.2.2 Establish ARFIMA(0,1,13)-GARCH (1,1) model:

The GARCH (1,1) model is established in this paper. The AIC, BIC and HQ values of the model are the smallest as a whole, and its effect is the best. The model expression is as follows:

Mean Value equation:

$$\begin{aligned} y_t = &-136.970 + \varepsilon_t - 0.618\varepsilon_{t-1} - 0.003\varepsilon_{t-2} - 0.033\varepsilon_{t-3} + 0.029\varepsilon_{t-4} + 0.027\varepsilon_{t-5} + 0.026\varepsilon_{t-6} \\ &+ 0.006\varepsilon_{t-7} + 0.007\varepsilon_{t-8} + 0.060\varepsilon_{t-9} - 0.075\varepsilon_{t-10} + 0.025\varepsilon_{t-11} + 0.120\varepsilon_{t-12} + 0.050\varepsilon_{t-13} \end{aligned} \tag{9}$$

Variance equation:

$$\sigma^2 = 240257.6 + 0.250\sigma^2_{t-1} + 0.742\varepsilon^2_{t-1} \tag{10}$$

5.3 Model inspection

Next, we will conduct ARCH-LM inspection.Observe whether the established ARFIMA(0,1,13)-GARCH (1,1) model has ARCH effect.

Table 2: ARCH LM test

Heteroskedasticity Test:ARCH			
F-statistic	0.001065	Prob.F(1,400)	0.974
Obs*R-squared	0.001071	Prob.Chi-Square(1)	0.9739

At this time, the concomitant probability is 0.974, and the original hypothesis is not rejected. It is believed that there is no ARCH effect in the residual sequence, indicating that the GARCH (1,1) model eliminates the conditional heteroscedasticity of the residual sequence.

Test its stability.We use the white noise test, Ljung-Box test[9], namely LB test and randomness test, to test whether the autocorrelation of the sequence within the m-order lag range is significant, or whether the sequence is white noise. The q statistic follows the chi-square distribution with the degree of freedom of m.The white noise test results are shown in the figure7.

According to Figure 7, we can see that the white noise test results of the data after fractional difference are all close to 1, indicating that the autocorrelation of the data is weak, and there is no obvious upward or downward trend, indicating that the data is stable.

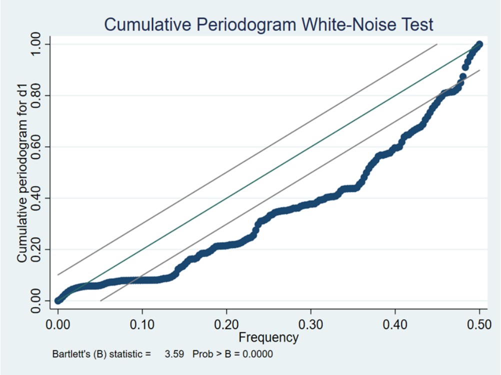


Figure 7: Ljung-Box test

5.4 Model predictions

Next, we use the established a model to predict the data in 2023.03.01. The predicted quantity is 20982, and the 95% forecast interval is [14689,28354].The following is our prediction of 2023 data using the time series prediction model, as shown in Figure 8.

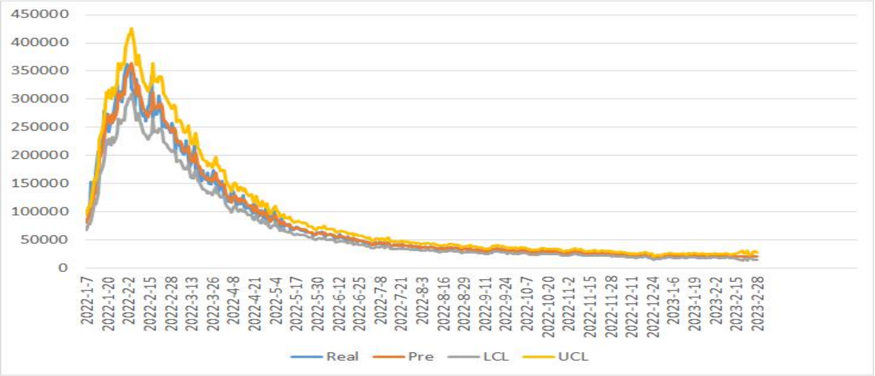


Figure 8: Time series prediction curve

6 Research on the attribute of words

6.1 Word attribute creation

To study attributes of the word affect the percentage of scores reported that were played in Hard Mode.This study used Pearson product-moment correlation coefficient to examine the degree of correlation of six word-related attributes.

- **Word form(Fm):**

The more forms a word has, the more likely it is to be used and the less difficult it is to guess, so the more word forms you set, the greater the property value.

- **Word formProfessionalism(Pf):**

Professional words are difficult to be used in daily life, so they are divided into professional words and non-professional words. our article quotes the concept of $TF - IDF$(term frequency–inverse document frequency)[11],if a word appears many times in an article, but the number of documents containing the word is small, it indicates that the word is professional.

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{11}$$

where $f_{t,d}$ is the raw count of a term in a document, i.e., the number of times that term t occurs in document d.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{12}$$

where N is total number of documents in the corpus$N = |D|$,$|\{d \in D : t \in d\}|$ is number of documents,where the term t appears ($i.e.$tf$(t,d) \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to $1 + |\{d \in D : t \in d\}|$ Then tf-idf is calculated as

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D) \tag{13}$$

- **Frequency(Fq):**

The number of times each word is used per million in the Corpus of Contemporary American English (COCA) ("Corpus of Contemporary American English," n.d.) as its frequency in English.

- **Loanwords(Rp):**

Some words are derived from Latin and are therefore classified into native and foreign words, such as mummy and tiara.

- **orthographic neighbors(Orth):**

This is the number of orthographic neighbors that a string has. An orthographic neighbor is defined as a word of the same length that differs from the original string by only one letter. For example, given the word 'cat', the words 'bat', 'fat', 'mat', 'cab', etc. are considered orthographic neighbors.

6.2 Hypothesis Tests for Correlation Coefficients

Let the Spearman correlation coefficient be r.

In order to verify whether the Spearman correlation coefficient is significantly different from 0, we conduct a hypothesis test on it.

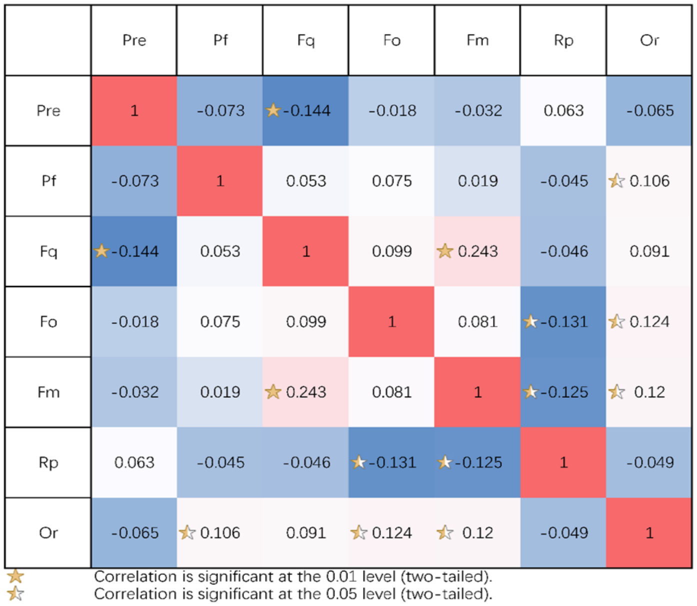
The null and alternative hypotheses are set as:

$$H_0 : r = 0, H_1 : r \neq 0 \tag{14}$$

Under the condition that the null hypothesis is established, a statistic is constructed by using the Spearman correlation coefficient so that it conforms to the standard normal distribution:

$$r_s \sqrt{n-1} \sim N(0,1) \tag{15}$$

We calculate the test value and find the corresponding P value to compare with 0.01, 0.05.



	Pre	Pf	Fq	Fo	Fm	Rp	Or
Pre	1	-0.073	★-0.144	-0.018	-0.032	0.063	-0.065
Pf	-0.073	1	0.053	0.075	0.019	-0.045	☆ 0.106
Fq	★-0.144	0.053	1	0.099	☆ 0.243	-0.046	0.091
Fo	-0.018	0.075	0.099	1	0.081	★-0.131	☆ 0.124
Fm	-0.032	0.019	☆ 0.243	0.081	1	★-0.125	☆ 0.12
Rp	0.063	-0.045	-0.046	★-0.131	★-0.125	1	-0.049
Or	-0.065	☆ 0.106	0.091	☆ 0.124	☆ 0.12	-0.049	1

★ Correlation is significant at the 0.01 level (two-tailed).
☆ Correlation is significant at the 0.05 level (two-tailed).

Figure 9: correlation matrix

For coefficients that are significant at the 0.01 level, we use a complete five-pointed star as a symbol.

For coefficients that are significant at the 0.05 level, we use a half-pointed star as a symbol.

7 Task 2:Report result forecast

7.1 Classification variable processing based on Apriori algorithm

Association rule mining allows us to discover the relationship between items from the dataset. It has many application scenarios in our life, and "shopping basket analysis" is a common scenario. Our paper uses this algorithm to mine the relationship between word classification attributes and word difficulty, and imagine each word as a shopping basket. The attribute of the word is regarded as a commodity. Then the problem is transformed into, when the word has which attributes, it is most likely to be difficult.

Step 1 Data processing:We transform several classification variables of word form (noun, verb, adjective, adverb, word specialization, foreign word) into frequent itemsets between the form search of the fact table and the difficulty of words. In terms of word difficulty processing, we cluster words by k-means according to seven attributes of the number of attempts, and divide them into two categories: difficult and simple. Table 3 is the data we have processed.Among them, 1 represents that the word belongs to this category. For example, the difficulty of main is divided into difficulty, and the word form is adjective, which is not a foreign word.

Step 2 Formulated description:Let I = $i_1, i_2...i_n$ be a set of n binary attributes called items.Here I = Pf,Fo,n,v,adj,adv,pron.Let $D = t_1, t_2...t_m$ be a set of transactions called the database. Here D = manly,molar...cramp transaction in D has a unique transaction ID and contains a subset of the

Table 3: Transaction table

Word	difficulty	Pf	Fo	n	v	adj	adv	pron
manly	1	0	0	0	0	1	0	0
molar	0	1	0	1	0	0	0	0
havoc	1	0	0	0	1	0	0	0
impel	0	0	0	0	1	0	0	0
condo	1	0	1	1	0	0	0	0

items in I. A rule is defined as animplication of the form:

$$X \Rightarrow Y, \text{ where } X, Y \subseteq I \tag{16}$$

Every rule is composed by two different sets of items, also known as itemsets, X and Y, where X is called antecedent or left-hand-side (LHS) andY consequent or right-hand-side (RHS).In order to select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used.Here we use minimum thresholds on support and confidence method, we first need to define the definition of support and confidence. The support of X with respect to T is defined as the proportion of transactionst in the dataset which contains the itemset X.

$$\mathrm{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|} \tag{17}$$

Confidence is an indication of how often the rule has been found to be true.The confidence value of a rule,X) Y , with respect to a set of transactions T, is the proportion of the transactions that containsX which also contains Y.Confidence is defined as:

$$\mathrm{conf}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)} \tag{18}$$

Step 3 Association rule found:The association rules are extracted by the algorithm mentioned above, and the discrete variables are converted into rules and their support is taken as the score of each attribute value, and applied to the regression of the second question. Table4 shows the size of the found association rules and confidence.

According to our association rules4, we use the confidence degree to assign the difficulty to each situation. The difficulty of foreign words in the guessing game is significantly increased. At the same time, the difficulty of adjectives is significantly greater than that of nouns and verbs. At the same time, when the professional words are adjectives, the number of games is also significantly increased. This is consistent with our common sense. Because there are few data in the data set that are filtered as professional words, the confidence degree obtained is deviated from the reality, Therefore, professionalism will be revised in the next calculation.

7.2 Multicollinearity test among attributes

From the first question to mine word attributes, use each attribute as an independent variable to build a model

Table 4: Association rules

rules	confidence	rules	confidence
{Pf} =>{difficulty}	0.272727273	{Pf,n} =>{difficulty}	0.2
{adv} =>{difficulty}	0.318181818	{n,v,adj} =>{difficulty}	0.2
{v} =>{difficulty}	0.416149068	{adj,adv} =>{difficulty}	0.285714
{adj} =>{difficulty}	0.552941176	{Pf,adj} =>{difficulty}	0.333333
{Fo} =>{difficulty}	0.777777778	{Fo,n} =>{difficulty}	0.777778
{n} =>{difficulty}	0.418410042	{Fo,adj} =>{difficulty}	1
{n,v} =>{difficulty}	0.362745098	{n,adj} =>{difficulty}	0.45
{v,adv} =>{difficulty}	0.375	{Fo,n,adj} =>{difficulty}	1
{v,adj} =>{difficulty}	0.416666667	{v,adj,adv} =>{difficulty}	0.4

From the correlation coefficient diagram, it is found that there is a relatively strong correlation between certain attributes, for example, the correlation coefficient between Rp and Fo is -0.16, and the correlation coefficient between Rp and Fm is -0.13, but the correlation is not particularly significant . Then use SPSS to test for multicollinearity:

Then refer to the eigenvalues and conditional indicators in the collinearity diagnosis. When the eigenvalue is approximately equal to 0, the value of the conditional index is greater than 10, and the variance ratio is close to 1 (one of them is enough), it can indicate that there is relatively serious collinearity.

Table 5: Multicollinearity test

Dimension	Eigenvalues	Condition Index	Variance Ratio	Fq	Mi	Rp	Orth
1	2.96	1.00	0.02	0.01	0.02	0.00	0.01
2	0.86	1.86	0.00	0.99	0.00	0.00	0.00
3	0.68	2.09	0.01	0.00	0.03	0.00	0.98
4	0.38	2.78	0.02	0.00	0.22	0.04	0.00
5	0.12	5.00	0.95	0.00	0.73	0.88	0.00

The result of multicollinearity test: 1. The eigenvalue of dimension 5 is approximately equal to 0 2. The variance ratio of Fq dimension 2, the variance ratio of Orth dimension 3, the variance ratio of Mi and Rp dimension 4 are all close to 1 It shows that there is serious collinearity.

7.3 Establishment of Partial Least Squares Regression Analysis Model

Due to the small amount of data in this question and the serious collinearity among the independent variables, the partial minimum double regression model was used to establish the model.

Then we use SPSSPRO to perform partial least squares regression.

We get the prediction result for the word "EERIE"

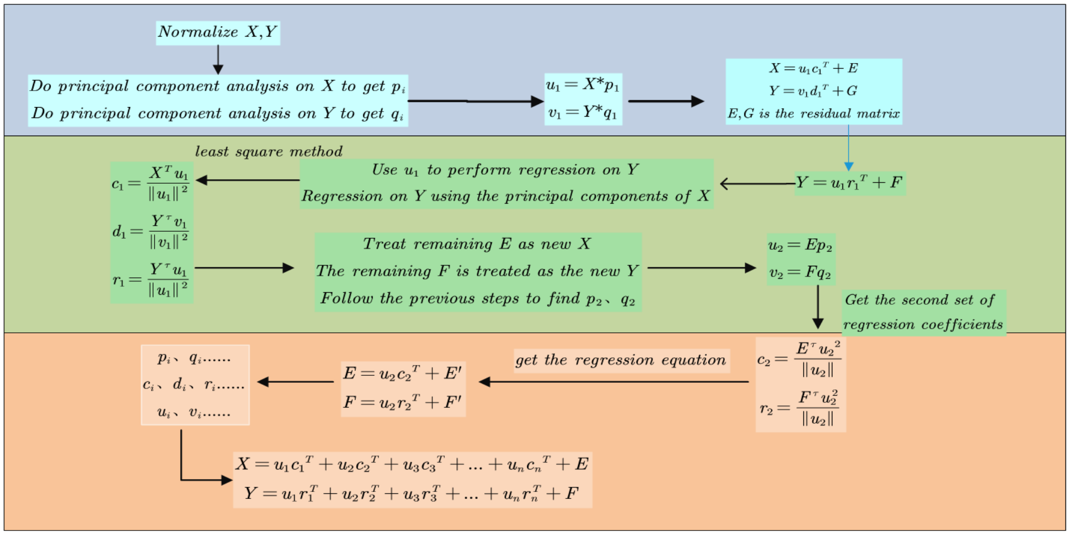

Normalize X,Y
Do principal component analysis on X to get p_i
Do principal component analysis on Y to get q_i
$u_1 = X^* p_1$
$v_1 = Y^* q_1$
$X = u_1 {c_1}^T + E$
$Y = v_1 {d_1}^T + G$
E,G is the residual matrix
least square method
$c_1 = \frac{X^T u_1}{\|u_1\|^2}$
$d_1 = \frac{Y^T v_1}{\|v_1\|^2}$
$r_1 = \frac{Y^T u_1}{\|u_1\|^2}$
Use u_1 to perform regression on Y
Regression on Y using the principal components of X
$Y = u_1 {r_1}^T + F$
Treat remaining E as new X
The remaining F is treated as the new Y
Follow the previous steps to find p_2、q_2
$u_2 = E p_2$
$v_2 = F q_2$
Get the second set of regression coefficients
$c_2 = \frac{E^T {u_2}^2}{\|u_2\|}$
$r_2 = \frac{F^T u_2^2}{\|u_2\|}$
get the regression equation
$E = u_2 {c_2}^T + E'$
$F = u_2 {r_2}^T + F'$
p_i、q_i......
c_i、d_i、r_i......
u_i、v_i......
$X = u_1 {c_1}^T + u_2 {c_2}^T + u_3 {c_3}^T + ... + u_n {c_n}^T + E$
$Y = u_1 r_1^T + u_2 r_2^T + u_3 r_3^T + ... + u_n r_n^T + F$

Figure 10: Flow chart of partial least squares regression analysis

Table 6: EERIE's forecast results

	1 try	**2 tries**	**3 tries**	**4 tries**	**5 tries**	**6 tries**	**7 or more tries (X)**
EERIE	0	2	15	33	30	16	4

7.4 Uncertainties in Models and Forecasts

1. Social media: Social media platforms are an important channel for Wordle games to spread online. If a certain social media platform has many users sharing their game results and experiences, then this may attract more players to try the game.

2. Media coverage: Media coverage and promotion of Wordle will also affect the popularity and popularity of the game. If a major media outlet covered the Wordle game, it might attract more people to try the game.

3. Current news: Certain current news events or hot topics may directly or indirectly affect the popularity of Wordle games. For example, the frequent occurrence of a certain word in a certain news event may drive more people to try that word in the game.

4. Word of mouth and recommendation: Wordle is a game with a very good reputation, and the recommendation among players is also one of the important factors for the popularity of the game. If someone shares their experience playing Wordle in social circles or online forums and receives positive feedback and recommendations from others, then this may attract more people to try the game.

5. Global epidemic: The global epidemic has also had a certain impact on the popularity of Wordle games. Due to people being restricted at home during the epidemic, the demand for games has increased a lot, which may also be one of the important reasons why Wordle games are popular

during the epidemic.

7.5 Evaluate the model prediction results

Table 7: evaluating indicator

	MSE	RMSE	MAPE	MAE
1 try	0.55	0.74	48.45%	0.51
2 tries evaluating indicator	13.75	3.71	66.70%	2.69
3 tries	48.27	6.95	32.44%	5.68
4 tries	23.82	4.88	12.62%	3.72
5 tries	27.83	5.28	20.71%	4.28
6 tries	32.79	5.73	51.97%	4.46
7 or more tries (X)	14.62	3.82	87.99%	2.01

It can be seen from the prediction results that the Mean Absolute Error and Root Mean Square Error are relatively small, indicating that the prediction results of the model are good; and in the evaluation index of Mean Absolute Percentage Error, "7 or more tries (X)" is removed This item, the rest of the MAPE values are within the acceptable range. Therefore, we can consider the model's prediction to be good, and we have confidence in it.

8 Task 3: Word difficulty classification model

In task 3, this paper uses the system clustering method to cluster words into three categories: "difficult", "medium" and "easy", with the contour coefficient of 0.567. In order to test the clustering effect, this paper uses Fisher discriminant analysis to back judge based on training samples, and measures the reliability of clustering analysis by the obtained confusion matrix.

8.1 Hierarchical clustering

We use the seven percentage values of words as the clustering index, and use the system clustering model for clustering.

The flowchart of system clustering is as follows.

We use the idea of cluster recognition to identify the relevant attributes of the classification based on the clustering results, and find out the cluster centers, so that we can use the Euclidean distance to divide the difficulty of words, and divide the difficulty into three categories through the elbow rule of the system clustering, that is: simple , moderate and difficult.

8.2 Feature recognition of categories

We extracted the characteristics of each letter in each word: word frequency (the probability of occurrence of a word per million words based on the COCA corpus), blending indicators (a mixture of word lexicality, whether the word is foreign and whether it has subject specialization), number of letter repetitions (whether the word has duplicate words, the more repetitions of the word, the fewer words meet the requirement), word orthogonality (the more word orthogonality (the more orthogonality of words, the greater the number of similar words).

In clustering, we obtain the cluster center by calculating the average value of all data objects in the class, and then we analyze the attributes of words in each category by comparing the cluster center values of these four attributes in the three categories .

Table 8: Category attribute recognition results

	Frequency	**Blending indicators**	**Repetitions**	**Orthogonality**
Category 1:Easy	2.19	0.16	0.19	1.90
Category 2:Temperate	1.95	0.16	0.64	1.57
Category 3:Difficult	1.73	0.15	0.91	2.30

It can be seen from the cluster center data that the word frequency attribute value of category 1 is relatively large, while the letter repetition value is small. It can be seen that relatively simple words often appear frequently in life, and people can easily think of them; and their letters There are fewer repetitions, and people don't need to worry about how to arrange the repeated letters when playing the game.

For category 2, the indicators of the words in this category are relatively balanced, and there are no extreme values, which shows that the words with moderate difficulty selected by the game are also in line with common sense.

For category 3, its word frequency is the least, indicating that those words that do not appear in life, people often ignore them and it is difficult to think of them; at the same time, because people often subconsciously want to try out more letters when playing games , will cause them to tend to guess different letters when guessing, and it is difficult to find the appearance of the same letter, so there will be more words containing the same letter in the difficult category; the orthogonality value of words in this category is higher, and it may be is because people ignore words that have the same word length but differ by only one letter.

8.3 Fisher linear discriminant analysis

According to the above clustering results, we then perform Fisher linear discriminant analysis based on the percentage data of 359 words to judge the clustering results.

The core idea of Fisher's discrimination is projection, trying to find an optimal projection vector or optimal discriminant function, so that the sample data is projected in this direction, and the discriminant is determined based on the principle that the within-group dispersion is as small as possible and the inter-group dispersion is as large as possible. function, and then determine the sample category according to the discriminant function. Assuming that there are l populations $G_1, G_2, ..., G_l$ and the observation samples are $x_{i1}, x_{i2}, ...x_{iq_i}$ $(i = 1, 2, ..., l)$, then the sum of squares of variance between groups and the sum of squares of variance within a group of sample data x_{ij} are respectively

$$SSA_x = \sum_{i=1}^{l} q_i \left(\overline{x_i} - \overline{x}\right)^2, SSE_x = \sum_{i=1}^{l} \sum_{j=1}^{q_i} \left(\overline{x_{ij}} - \overline{x_i}\right)^2 \tag{19}$$

Suppose the projection vector is p, then the projected data of the observation sample is $y_{ij} = p' x_{ij}$, and the linear relationship is the corresponding discriminant function, then the sum of squared deviations between groups and the sum of squared deviations within a group of projected data are y_{ij} respectively

$$SSA_y = \sum_{i=1}^{l} q_i \left(\overline{y_i} - \overline{y}\right)^2 = p' SSA_x p, SSE_y = \sum_{i=1}^{l} \sum_{j=1}^{q_i} \left(\overline{y_{ij}} - \overline{y_i}\right)^2 = p' SSE_x p \tag{20}$$

When the objective function $f(p) = (p' SSA_x p) / (p' SSE_x p)$ reaches the maximum value, the projection vector p obtained at this time is the best. In order to ensure the uniqueness of the solution, assuming that SSA_x / SSE_x is the unit matrix E, the partial derivative is derived

$$(SSE_x)^{-1} \cdot SSA_x p = \lambda p \tag{21}$$

The maximum eigenvalue of $f(p)$ obtained at this time is the maximum value of the objective function $(SSE_x)^{-1} \cdot SSA_x$, and the corresponding eigenvector is the optimal projection vector, so that the linear discriminant function is obtained as $y_{ij} = p' x_{ij} (i = 1, 2, ..., l; j = 1, 2, ..., q_i)$. According to the discriminant function, the projection matrix of the observation sample and the corresponding Based on the group mean projection matrix, the sample discrimination result can be obtained based on the principle of the minimum distance between the vectors of the distance discrimination method.

By using SPSSPRO, we get the discriminant functions as:

$$\begin{aligned} y1 = & -8739.167 + 152.863 * N_1 + 169.141 * N_2 + 175.642 * N_3 \\ & + 177.733 * N_4 + 170.664 * N_5 + 179.317 * N_6 + 172.743 * N_7 \end{aligned} \tag{22}$$

$$\begin{aligned} y2 = & -8754.238 + 153.61 * N_1 + 169.251 * N_2 + 175.21 * N_3 \\ & + 178.164 * N_4 + 170.975 * N_5 + 179.603 * N_6 + 173.218 * N_7 \end{aligned} \tag{23}$$

$$\begin{aligned} y3 = & -8786.136 + 153.849 * N_1 + 169.284 * N_2 + 175.447 * N_3 \\ & + 178.103 * N_4 + 171.31 * N_5 + 180.62 * N_6 + 173.859 * N_7 \end{aligned} \tag{24}$$

8.4 Classification accuracy

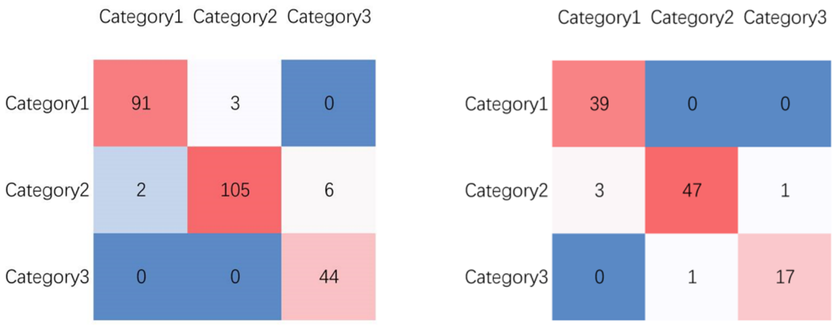
According to the results of system clustering evaluation, the silhouette coefficient is 0.567, which is high in the range of [-1,1], DBI is less than 1, and CH is high, indicating that the clustering effect is better.

Table 9: Evaluation metrics for hierarchical clustering

Contour factor	DBI	CH
0.567	0.919	309.688

Table 10: Evaluation indicators for discriminant analysis

	Accuracy	Recall rate	Positive Accuracy	4 tries
Training set	0.956	0.956	0.958	0.956
Test set	0.944	0.944	0.947	0.945



	Category1	Category2	Category3
Category1	91	3	0
Category2	2	105	6
Category3	0	0	44

(a) Training set confusion matrix

	Category1	Category2	Category3
Category1	39	0	0
Category2	3	47	1
Category3	0	1	17

(b) Testing set confusion matrix

Figure 11: Confusion matrix

In the linear discriminant model, the precision rate, recall rate, positive precision rate, and F1 of the training set and test set are all around 0.95, which means that the clustering reliability is high and the classification accuracy is good.

9 Task 4:Analysis of data set characteristics

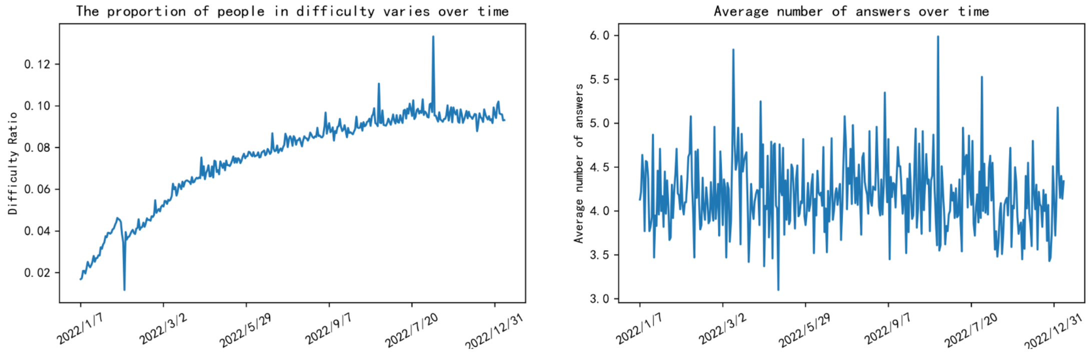
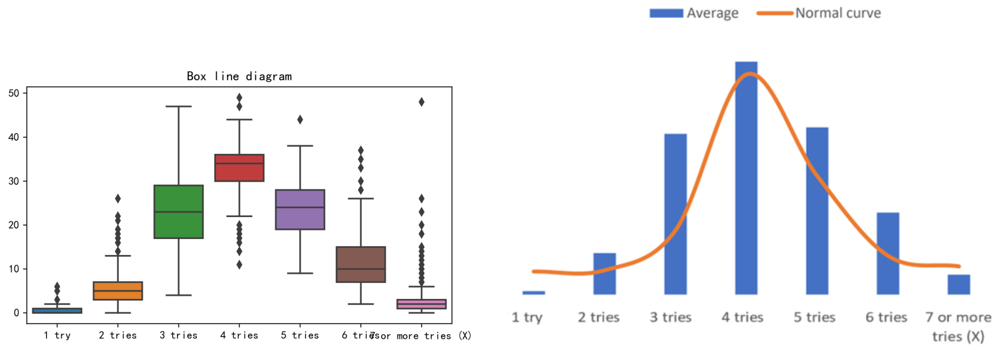


Figure 12: correlation matrix

We will turn on the difficult mode scale by time line chart, you can see the overall trend of the scale is rising, there are two salient points. Due to the outbreak of covid-19 pneumonia at the end of 2020, most people join the game but just experience it. However, the attraction of the game is weak, and people keep losing it with the change of time, the remaining players are more adventurous, at the same time the continuous promotion of the game most people have mastered the rules of the game, then they will continue to be inclined to challenge themselves.

The right graph is The average number of challenges with The change of time series, you can see The overall trend of 4times, only a few times and a number of times The situation, we can think of The difficulty of The problem every day is random, occasionally difficult and simple words appear.



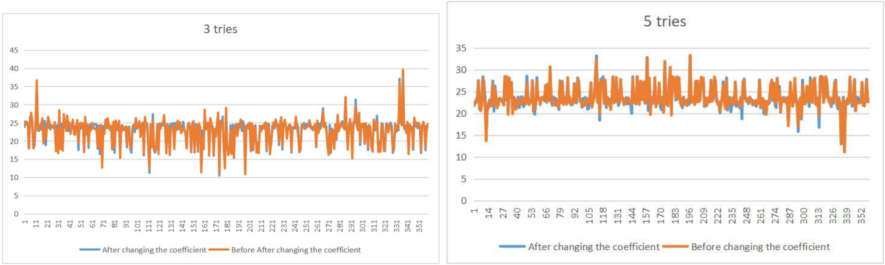
(a) Box diagram

(b) Normal distribution

Figure 13: Analysis of game times

We draw a box chart and a bar chart for the percentage of each number of times. It can be seen that the number of answers conforms to the normal distribution. At the same time, four times are the mode of the number of answers. The number of outliers of number 7 is the largest, and the variance of number 3 is large. Through the analysis of the number of answers, we can see that our previous method of clustering by the number of answers is reasonable.

10 Model sensitivity analysis



(a) Sensitivity test of 3 tries

(b) Sensitivity test of 5 tries

Figure 14: Sensitivity test

In order to determine the sensitivity of our regression model to different time series data, a sensitivity analysis was carried out. When establishing the regression equation prediction, the standard error of the regression coefficient measures the reliability of the estimated value of the regression coefficient. Generally, the smaller the standard error, the higher the sensitivity of the model. We manually change the coefficient of the independent variable of the word occurrence frequency of the regression equation by 10%, and then bring the data in. It can be seen that the percentage of 3 attempts and 5 attempts after the change is relatively small, but the frequency of word occurrence is the biggest factor affecting the regression equation. Therefore, our regression equation has no obvious change to the coefficient, and the model has good sensitivity.

The analysis shows that when the maximum factor changes within a certain range, the change range of the final percentage is 3.68%. Therefore, when other factors with small influence change, the percentage of attempts will not fluctuate significantly, which indicates that the proposed model is quite robust to the word attribute.

11 Strengths and Weaknesses

11.1 Strengths

- The model and strategy based on pairs trading are scientific and reasonable, which can maximize the income. The results obtained have strong confidence.

- The model uses the AP algorithm to assign the classification variables, which is more accurate and accurate than using only the classification variables to build the model. At the same time, the partial least squares regression model is mostly used for multi-output calculation, which is more in line with the actual situation of the topic.

- The model uses Fisher algorithm to test its clustering results in clustering, and the model is more reliable

11.2 Weaknesses

- The word attribute used by the model is only the most commonly used attribute variable in current linguistic research, while there is not much research on uncommon variables, such as vowel consonant combination, irregular spelling and other attributes, which will be cited in our subsequent research.
- The model is directly smoothed in the processing of time series without analyzing factors such as special festivals, which is also reflected in the uncertain factors in Task 3.

References

[1] Wannigamage D. Player Population Patterns in Digital Games: A Data Analytics and Machine Learning Approach[D]. UNSW Sydney, 2021.

[2] De Silva N. Selecting Optimum Seed Words for Wordle using Character Statistics[C]//2022 Moratuwa Engineering Research Conference (MERCon). IEEE, 2022: 1-6.

[3] De Silva N. Selecting Optimum Seed Words for Wordle using Character Statistics[C]//2022 Moratuwa Engineering Research Conference (MERCon). IEEE, 2022: 1-6.

[4] de Silva N. Selecting seed words for wordle using character statistics[J]. arXiv preprint arXiv:2202.03457, 2022.

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

[6] Hamilton J D. Time series analysis[M]. Princeton university press, 2020.

[7] 汪妍. 基于ARFIMA-GARCH-LSTM混合模型的股价预测研究[D].上海师范大学,2022.DOI:10.27312/d.cnki.gshsu.2022.000752.

[8] Bauwens L, Laurent S, Rombouts J V K. Multivariate GARCH models: a survey[J]. Journal of applied econometrics, 2006, 21(1): 79-109.

[9] Hassani H, Yeganegi M R. Selecting optimal lag order in Ljung–Box test[J]. Physica A: Statistical Mechanics and its Applications, 2020, 541: 123700.

[10] the Corpus of Contemporary American English (COCA) (“Corpus of Contemporary American English,” n.d.)http://www.neuro.mcw.edu/mcword/

[11] tf–idf https://en.wikipedia.org/wiki/Tf

12 Strategy Recommendation and Conclusions

Theme:MEMO

Dear Times puzzle editor

I am writing to you as MCM Team 2314475 to summarize our results on forecasting Wordle. After analyzing various data sources and conducting extensive research, we have made some interesting findings that I believe your readers may be interested in

First, we developed a model to explain the change in the number of results reported every day. I found that using * * ARIMA model * * to establish a time series regression model for the number of reported results can best explain this change. ARIMA is a time series analysis method, which is a combination of autoregressive moving average model (ARMA) and differential integration model (I). ARIMA model makes the time series data stable by differential, and on this basis establishes an autoregressive moving average model to predict the future value. Based on this model, we have created a forecast interval for the number of reported results on March 1, 2023. My prediction is that the number of reported results on this day will be in the range of 19368-20982

Secondly, we investigated whether any attribute of a given word would affect the percentage of report scores played in difficult mode. Our analysis found that there was a significant correlation between the frequency of words and the percentage of scores reported in the difficult mode. Specifically, more common words tend to report higher scores in the difficult mode. However, I did not find any significant correlation between other attributes of a given word (such as the irregular spelling of the word) and the percentage of scores reported in the difficult mode

Thirdly, we have developed a model to predict the distribution of report results for a given solution word in the future. Our model uses the relevant attributes that are usually used to describe the clustering results. Based on this model, we made a specific prediction for the word "EERIE" on March 1, 2023. We predict that the reported result distribution of this word will be (1=0%, 2=2%, 3=15%, 4=33%, 5=30%, 6=16%, X=4%)

However, there are also some uncertainties in this prediction, because the distribution of the report results may vary according to the player's choice and the uncertain factors of the day

Fourth, we have developed a model to classify solution words by difficulty. We use the central decision cluster of the system cluster and the attributes of a given word to predict whether it is classified as "difficult", "medium", and "easy". Our analysis found that words with fewer occurrences are often classified as "difficult", while more common words are often classified as "simple". In addition, words with high repetition times are often classified as "simple", while words with high orthogonality are often classified as "difficult". Based on this model, the word "EERIE" is classified as "easy"

Finally, there are other interesting features of this dataset that are worth mentioning. For example, we note that the number of reported results will gradually decrease and stabilize over time, which indicates that the popularity of Wordle games is gradually declining. In addition, the data set contains a wide range of words, ranging from common English words to more obscure terms

In conclusion, we believe that our analysis provides valuable insights into the patterns and trends in the data set provided. We hope that these findings will arouse the interest of your readers, and we look forward to hearing any feedback or comments from you

Sincerely,

Sicerely yours,
Your friends

Appendices

Appendix A Piecewise Linear Trend Extraction

Algorithm : Piecewise Linear Trend Extraction

```
Input : Xdata, Ydata, minPieceSize, threshold
seriesLen = length(Ydata)
pieceStartI = 1
while (pieceStartI < seriesLen) do
    pieceEndI = pieceStartI + minPieceSize - 1
    if pieceEndI > seriesLen then
        pieceEndI = seriesLen
    end if
    bestFit = find the best linear fit for the data subset (pieceStartI : pieceEndI)
    error = calculate the error (RMSE) of the fit
    if error $\leq$ threshold then
        save pieceStartI and pieceEndI
        pieceStartI = pieceEndI + 1
    else pieceEndI = pieceEndI + 1
        while (pieceEndI < seriesLen) do
            newFit = find the best linear fit for the data subset (pieceStartI : pieceEndI)
            newError = calculate the error (RMSE) of the new fit
            if newError > threshold then
                break
            else
                bestFit = newFit
                error = newError
                pieceEndI = pieceEndI + 1
            end if
        end while
    save pieceStartI and pieceEndI
    pieceStartI = pieceEndI
    end if
end while
if remainingLen $\leq$ minPieceSize then
    merge remainingLen to the last piece
end if
```

Appendix B Code for Part 3.1

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Appendix C Code for Part 3.3

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```