Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2316429

Wordle Report: An ARIMA and BPNN-Based Evaluation and Prediction Model

Abstract

The game of "Wordle" has recently gained popularity as a succinct word-based game. It is now necessary to analyze and predict the reported and shared counts of this game on social media, as well as to conduct an analysis of the game's own patterns, in order to facilitate further research by news professionals and game developers. Therefore, our team has established a model based on ARIMA time series analysis, BP neural network construction, and K-means clustering methods to evaluate words and predict reporting details.

First, we evaluated the correlation between the total number of shared result reports and various potential influencing factors. To achieve this, we first established an ARMA model inherited from the classic forecasting model. This method is a typical way to study the rational spectrum of stationary random processes. Subsequently, after performing certain data cleaning and GARCH model construction, we realized that if the fluctuation of the distribution of the total number of reports over time was extracted, the characteristic of this random process did not vary with time, which means there was a stationary time series. Thus, we employed an improved ARIMA model, which was better able to fit the actual regularity of the total number of reports over time. The model's various parameters performed well after evaluation. Through the construction of this model, we were able to predict the range of the total number of reports on March 1st. Furthermore, by analyzing the data on the proportion of Hard Mode reports to the total number of reports and conducting a correlation test, we judged the correlation between word attributes and this proportion.

Secondly, to analyze and predict the proportion distribution of reports of various attempt numbers among all shared game outcomes in Wordle, we used a model constructed based on the BP neural network algorithm and NLP methods to address this issue. Since natural language has a certain level of abstraction, we digitally processed the dimensional features of the word itself and established a connection between the neural network training and the proportion distribution of attempt numbers. Based on human language habits and related data, we extracted two types of feature attributes for each word: word isolation level and word priority. We then determined the attribute details of each word in the dataset and established a neural network between the various factors.

Finally, to evaluate the word difficulty in Wordle, we used the K-means clustering algorithm to preliminarily classify each word based on its attributes. Then, we used the proportion distribution of attempt numbers as a calibration for the classification dimension, and established a Wordle word library classification standard with three dimensions: "easy," "moderate," and "difficult." We flexibly combined parameter search and classification models to develop a model specifically designed for the Wordle word evaluation system. To evaluate this model, we compared it with a simple model that classifies based on a single feature such as letter repetition frequency or word frequency, and found that our model has better generalization ability and applicability. Its classification ability remains relatively stable even when the conditions change.

Keywords: BP Neural Network Model ARIMA Time Series Forecasting Model NLP K-means Cluster Analysis Wordle

Contents

1 Introduction

1.1 Problem Background

Wordle is a daily guessing game where a five-letter word is selected from a database each day, and players have six chances to guess the word. After each guess, the game provides feedback: grey blocks indicate that the word does not contain that letter, yellow blocks indicate that the word contains the letter but in the wrong position, and green blocks indicate that the word contains the letter in the correct position.

In addition, the game has two modes: regular mode and hard mode. In hard mode, players are required to use the green or yellow letters they have already guessed in subsequent guesses until the word is correctly guessed.

We now have a table that records the information from 1/7/2022-12/31/2022, including the daily word, number of reported results, number in hard mode, and the proportion of 1-6 and X tries. We will do the following to determine the sales strategy and functional design:

- Predict the possible range of the number of participants on March 1, 2023, and determine if word properties will affect the number of people choosing the hard mode.
- Develop a model will predict the percentage of future attempts for each feedback category (1, 2, 3, 4, 5, 6, X). Explain uncertainties are associated with your model and predictions and an attempt will be made to predict the word EERIE.
- Develop and summarize a model to classify solution words by difficulty, and evaluate the difficulty of the word EERIE.
- List and describe some other interesting features of this data set.
- Summary and write a letter to the editor of the New York Times.

1.2 Problem Analysis and Model Preview

The first question needs to build a model to predict the reported results number. We used the time-series model ARIMA to analyze and forecast the time and the number of people participating in the report. The results show that the two have a strong correlation, which proves the rationality of our model, and ARIMA can be used to predict the number of reported people in the future days.

The second question requires building a model that, given a word, predicts the percentage of attempts. However, since the information that can be extracted from each word is abstract, we need to digitize the information of the word, and then establish a relationship with the number of tries. We extract two features that may affect them, word isolation level and word priority. Determine the two attributes of each word through related data sets, and then establish a BP neural network between the two attributes and the number of tries to build a connection.

The third question needs to build a model to evaluate the difficulty of words in Wordle games. The evaluation of difficulty requires some related indicators and evaluation criteria. These characteristics conform to the k-means clustering model, so we use this model to evaluate the difficulty. After that, we hope to use the existing data to classify words. The data that can serve as the classification standard is

the number of tries. The proportion distribution of its components can reflect the difficulty of the word to a certain extent.

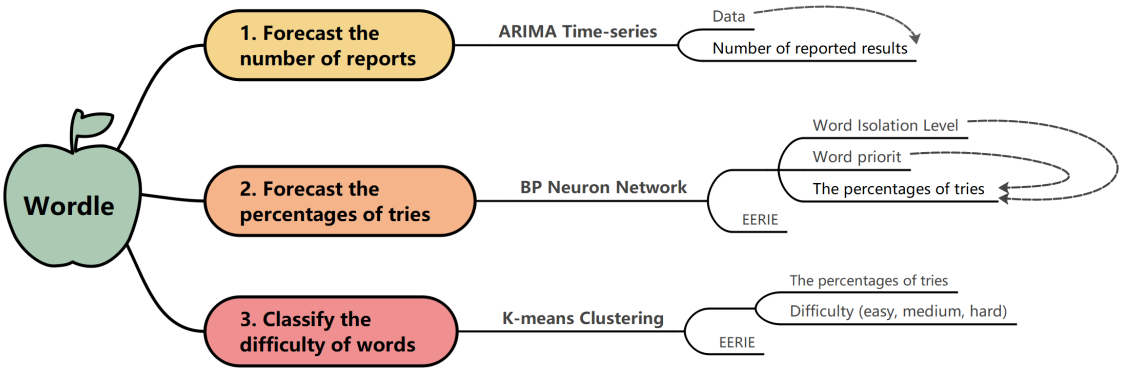

Figure 1: Flow Chart of Our Work

2 Assumptions and Symbols

2.1 Model Assumptions

- Assuming that no extreme situations happened, which rapidly influences total number of wordle player and players' distribution for English level.
- Assuming that the percentage of guessing the correct answer in one attempt has nothing to do with the word's feature, and only depends on word frequency.
- Assuming that the wordle's player is making an earnest effort to guess the word, and is not aware of the answer.
- Assuming that the wordle's player is trying to guess the word in the fewest possible attempts after careful consideration.
- Assuming that all players have a basic vocabulary when playing wordle.
- Assuming that all players use the same strategy: when there are three or four characters determined, players would try words contains these characters. When there are less than three characters, players would try their best to reducing the number of possible words.

These assumptions ensure that our analysis focuses on how the player solves the puzzle, independent of any external factors, and builds a reasonable model.

2.2 Symbols and Definitions

Table 1: Notations

Symbols	Description
$lev(a, b)$	Levenshtein distance of string a and b
$C(\omega, n)$	nth isolation level of word ω
$p(\omega)$	word priority of word ω
$E(\omega)$	Elimination Value of word $omega$
W	set of all 5 letter words in English
$\|W\|$	cardinality of set W

Definitions of Levenshtein distance, isolation level, word priority and Elimination Value are given in chapter 4.

3 Time-Series Analysis and Forecast Model

3.1 Data Cleaning

We first checked the data for outliers, missing values, duplicate values, and erroneous data to ensure the data quality. Statistical methods and visualizations were used to identify and address any issues.

We processed and cleaned the data based on the real situation. Firstly, after the check, we found that the words "tash" and "clen", whose contest number are "314" and "525", both have only four letters in the word, which mean they are the erroneous data. Then we search the all allowed-words in the Wordle game, guess that the two words are "taish" and "clean". Secondly, the "473" word "marxh" does not exist in the dictionary, the correct word ought to be "march". Thirdly, the number of reported results of word "study" may be wrong, because it is almost the same as the number in hard mode. Every step of the cleaning process was recorded for future reference. Every step of the cleaning process was recorded for future reference.

Contest number	Word
314	tash
473	marxh
525	clen
529	study

Figure 2: Data Cleaning

After cleaning the data, we performed another round of checks to ensure the data was cleaned correctly and that no new issues were introduced.

3.2 Introduction to ARIMA

ARIMA, short for Auto-Regressive Integrated Moving Average, is a statistical model that analyzes time-series data to forecast future trends. It is a famous time-series forecasting method proposed by George Edward Pelham Box and Gwilym Meirion Jenkins in the early 1970s, so it is also known as the box-Jenkins model. The main concept behind ARIMA is to use past and present values of the time series to predict future values. This model is used aiming to predict the future number of reported results based on the contest number of the certain word.

3.2.1 Overview of ARIMA model

The ARIMA model is generally composed of two models, the auto-regression (AR) model and the moving average (MA) model. The AR model is a statistical method for dealing with time series:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{1}$$

Then, the MA model uses a linear combination of past residual terms to examine future residuals:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2}$$

The combination of the AR model and the MA model yields the ARMA model:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{3}$$

For above formula, p is the order and q is the white noise. The above equation is also written as ARMA(p,q). Next, we define the Lag operator:

$$LX_t = X_{t-1} \, or \, X_t = LX_{t+1} \tag{4}$$

So we get the I model:

$$(1 - L)^d X_t \tag{5}$$

The ARIMA model is actually an ARMA model after transforming the time series into a smooth series using the I model. Thus, the ARIMA prediction model can be expressed as the formula:

$$\hat{p}^{\{t\}} = p_0 + \sum_{j=1}^{p} \gamma_j p^{\{t-j\}} + \sum_{j=1}^{q} \theta_j \varepsilon^{\{t-j\}} \tag{6}$$

where p is the order of Auto-regressive Model (AR), q is the order of Moving Average Model (AM), δ $\{t\}$ is the Error term between time t and $t-1$, γ_j and θ_j are the fitting coefficients, p_0 is constant term.

3.2.2 ADF and Ljung-Box test

When using ARIMA, it is required that the time series is smooth. the full name of the ADF test is Augmented Dickey-Fuller test, which can be used when there is a lagged correlation of higher order in the series: assuming that there is a unit root in the series, if the significance test statistic obtained is less than three confidence levels (10%, 5%, 1%), it corresponds to having (90%, 95%, 99%) of confidence to reject the original hypothesis.

Differential order	t	P	AIC	Threshold		
				1%	5%	10%
0	-3.867	0.002***	7203.313	-3.45	-2.87	-2.571
1	-4.242	0.001***	7195.208	-3.45	-2.87	-2.571
2	-10.663	0.000***	7176.608	-3.45	-2.87	-2.571
Note: ***, **, * represent 1%, 5%, 10% significance levels, respectively						

Figure 3: ADF Inspection Form

As the table displays, the value of P shows significance (P<0.05), indicating that the non-smooth hypothesis is rejected and the series is a smooth time series; the comparison of statistical values and ADF Test result for different degrees of rejection of the original hypothesis at the critical values of 1%, 5%, and 10%, and the ADF Test result is less than the statistical values of 1%, 5%, and 10% at the same time, indicating that the hypothesis is very well rejected.

The AIC value is a measure of the goodness of fit of a statistical model. As all the values shown, the model shows good stability at the difference orders of 0, 1, and 2, respectively.

Use Ljung-Box Q test to guarantee residuals are independent. The Statistic of it is:

$$Q(q) = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{7}$$

where n is the sample size, $\hat{\rho}_k$ is the sample auto-correlation at lag k, and h is the number of lags being tested. And the statistic Q asymptotically follows a x^2 distribution. [3]

3.2.3 Selection of parameters p, q

The trailing and truncating shapes of the ACF (auto-correlation function) and PACF (partial auto-correlation function) images determine how the parameters p, q should be chosen. Respectively, they are both functions to evaluate the Linearity of the value in the model.

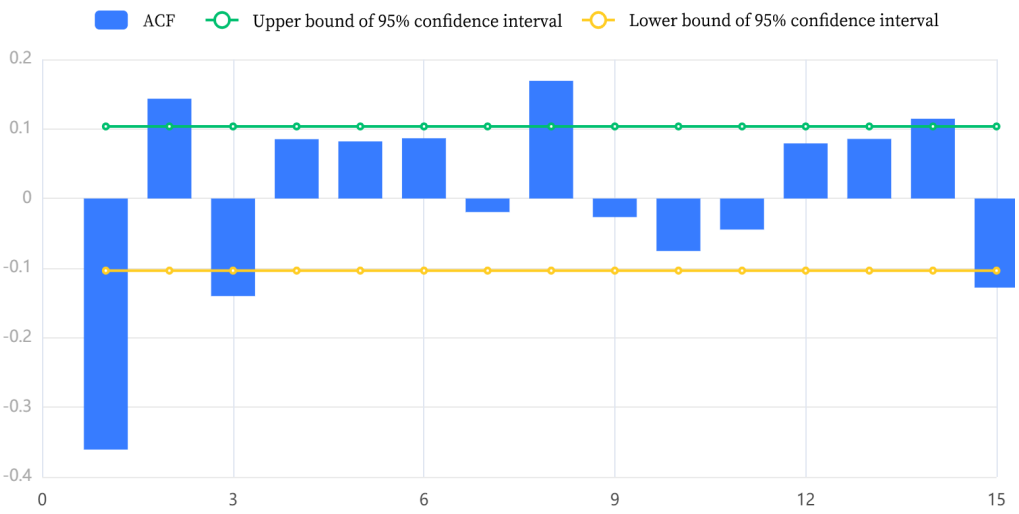

Figure 4: Final Differential Data ACF Graph

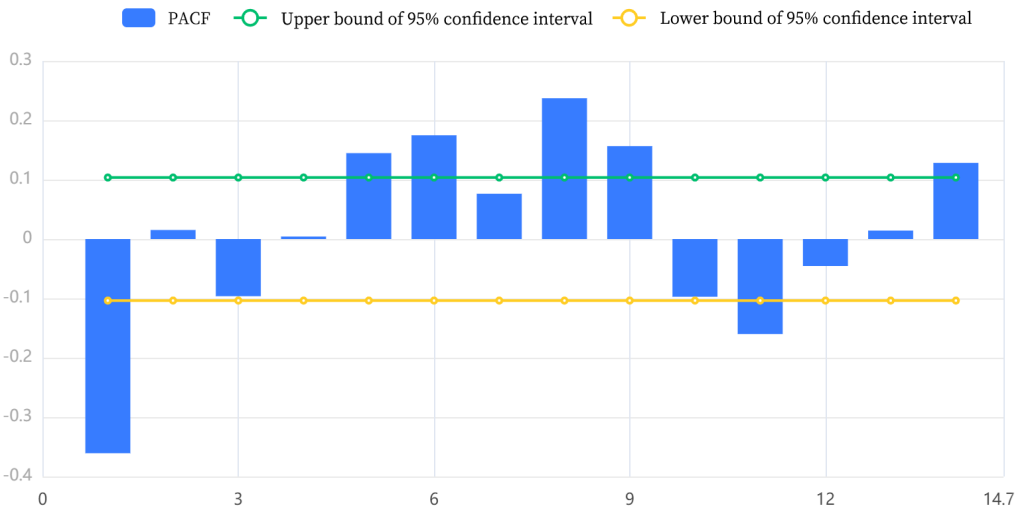

Figure 5: Final Differential Data PACF Graph

Both ACF and PACF graphs drag the tails, and the most significant order in PACF and ACF graphs can be used to decide the p and q values.

3.3 Results Analysis

3.3.1 Results of ARIMA prediction model

The system automatically finds the optimal parameters based on the AIC information criterion, and the model results are the ARIMA model (1,1,0) test table; from the analysis of the Q statistic result that the value of Q6 is small and the value of its p is more than 0.1, the hypothesis that the residual of the

ARIMA Model (1,1,0) Inspection Table					
Q-statistic	Q6	Q12	Q18	Q24	Q30
	0.019	24.347	54.224	82.246	98.285
R^2	0.982				

Figure 6: Model Test Results

model is a white noise sequence cannot be rejected. That is, there is no auto-correlation in the residuals of the model, the model residuals are white noise; at the same time, the goodness-of-fit R^2 of the model is 0.982, the model performance is excellent, and the model basically meets the requirements.

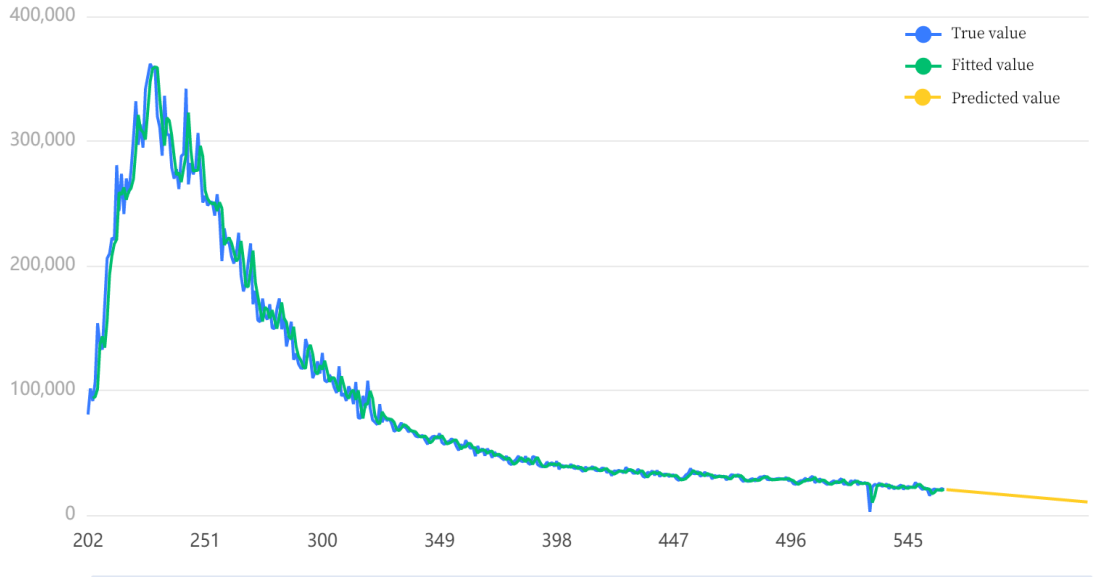

Figure 7: Results of ARIMA Model Prediction

According to the above theories, we apply the ARIMA model to forecast the number of reported results on March 1, 2023. We use all the data to train the model with the parameter vector (p, q, d) = (1, 1, 0) and predict the on March 1 the number will be in the interval (10288, 10624). Our result of prediction is shown as the following Figure.

3.3.2 Relationship between Hard Mode reports and word attributes

To answer the question of whether any attributes of the word affect the percentage of scores reported that were played in Hard Mode, we first need to perform some data analysis.

We first calculated the proportion of report numbers in Hard Mode to the total number of reports. We then performed descriptive statistical analysis and a normality test on this data. The results showed that the standard deviation of the proportion data was 0.009, the coefficient of variation (CV) was 0.104, and the data was relatively normally distributed. These findings suggest that there were not many hidden factors that could have influenced the data.

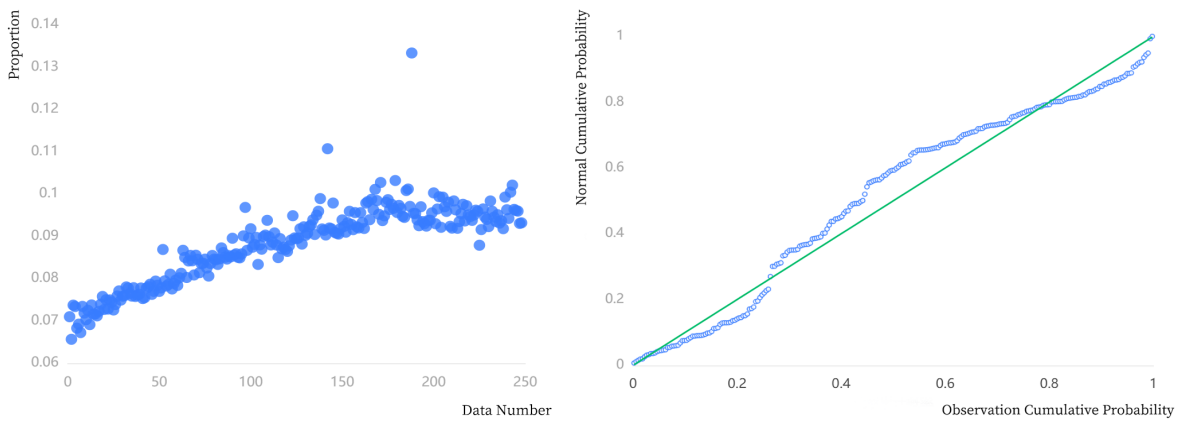

Figure 8: Statistical Analysis and Normality Test

Furthermore, we evaluated the approximate difficulty of each word by converting the proportion of attempts for each difficulty level. We then performed a correlation analysis between the difficulty level and the proportion of scores reported in Hard Mode, and found a weak correlation between the two variables.

Overall, our data analysis suggests that there is no strong correlation between the attributes of a word and the percentage of scores reported that were played in Hard Mode. These findings may be attributed to the fact that the game mechanics and individual player preferences have a greater influence on the choice of difficulty level.

3.3.3 Analysis of error sources and sample feature

The amount of data in this sample is small, and the model may be overfitting, resulting in a decrease in the accuracy of the model. The phenomenon that the number of reports decreases over time indicates that the influence of a word game with a relatively simple model will gradually and slowly decay. At the same time, considering the variability of online user behavior, it is possible that new games of the same type will appear on the Internet, resulting in a sudden drop in the number of Wordle users.

4 Attempts Percentage Prediction Model

The relationship between words and the percentage of future attempts is quite unclear. By intuition, on the one hand, time doesn't influence the percentage , because the distribution of English level among players remains the same when number of reports varies. On the other hand, word is definitely important, as how common the word is, how the word is build and so on determines the difficulties to guess.

We extracted four features based on the word itself and the word's frequency in English, tested the features' correlation by Pearson correlation coefficient. The features are non-linear according the test. Therefore, we use BP neural network on those features to predict the percentage of tries except one try.

Percentage of one try is set to be 0.0063% in our model, as the probability of it is fixed to $\frac{1}{15920}$. The number 15920 means there are 15920 5-letter words in English.

4.1 Word Features in Model

4.1.1 Word Isolation Level

Measuring the difference between two words is useful in estimate difficulty of word, which influences the percentage. If a word is less difference with more words, there's a greater possibility to gather more information from previous guesses. Meanwhile, words with only one letter difference can also lead to a situation: multiple guesses are needed to determine answer even with four green tiles. For example, words "hands" and "pands".

To measure the difference, we introduced the Levenshtein distance. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. It's a string metric widely used in natural language processing, for measuring the difference between two sequences.

The Levenshtein distance between two strings a, b is given by $lev(a, b)$, where

$$lev(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ lev(tail(a), tail(b)) & \text{if } a[0] = b[0], \\ 1 + min \begin{cases} lev(tail(a), b) \\ lev(a, tail(b)) \\ lev(tail(a), tail(b)) \end{cases} & \text{otherwise} \end{cases} \tag{8}$$

Where,

$tail(a)$ represents a string of all but the first character of a,

$a[0]$ represents the first character of a,

$|a|$ represents the length of a.

As described in the beginning of section, the difference influence guesses in multiple ways. To cover these ways well, we use isolation level to estimate how unique a word is, among all 5 letter words. The nth isolation level of word ω is the number of words, where the Levenshtein distance between ω and these words are all n. It's giving formally by:

$$C(\omega, n) = |\{v \in W \mid lev(w, v) = n\}| \tag{9}$$

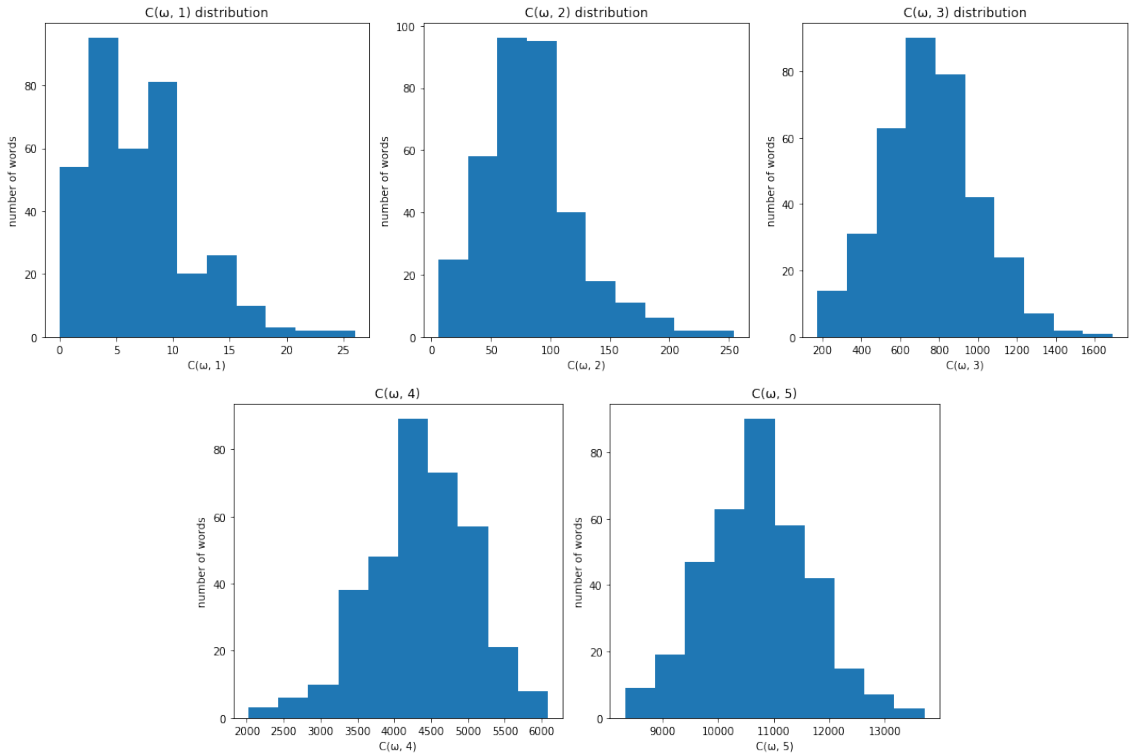

Figure 9: The Distribution of $C(\omega, n), \omega \in W$

According to assumption 4, we take $C(\omega, 1)$ and $C(\omega, 2)$ as two features to train BP neuron network.

4.1.2 Word priority

Word frequency is a important feature in prediction. For a brief view, the mean frequency of all five letter words in English is 6.55e-06, while then mean of words occurred in dataset is 5.17e-05. Also, players' vocabulary depends on word frequency.

In experience, we assume all players have a similar possibility to choose most frequent N_{common} words. Then, the possibility decreases with word frequency.

According the graph below, The raw word frequency data distribution doesn't match our assumption. Therefore, we introduced word priority $p(\omega)$ for preprocessing.

The priority of ith most frequent word p_i is given by evenly space all 5 letter words(sorted by frequency) to $[-10 + \frac{N_{common}}{|W|}, 10 + \frac{N_{common}}{|W|}]$, then apply sigmoid function to fit the result to our assumption.

More formally:

$$sigmoid(x) = \frac{1}{1 + \epsilon^{-x}} \tag{10}$$

$$p_i = sigmoid(10(\frac{n_{com}}{|W|} - 1 + \frac{i}{|W|})), i \in \{0, 1, \dots |W|\} \tag{11}$$

And $p(\omega)$ is determined by first find the index j of ω in all 5 letter words which sorted by frequency, then:

$$p(\omega) = p_j \tag{12}$$

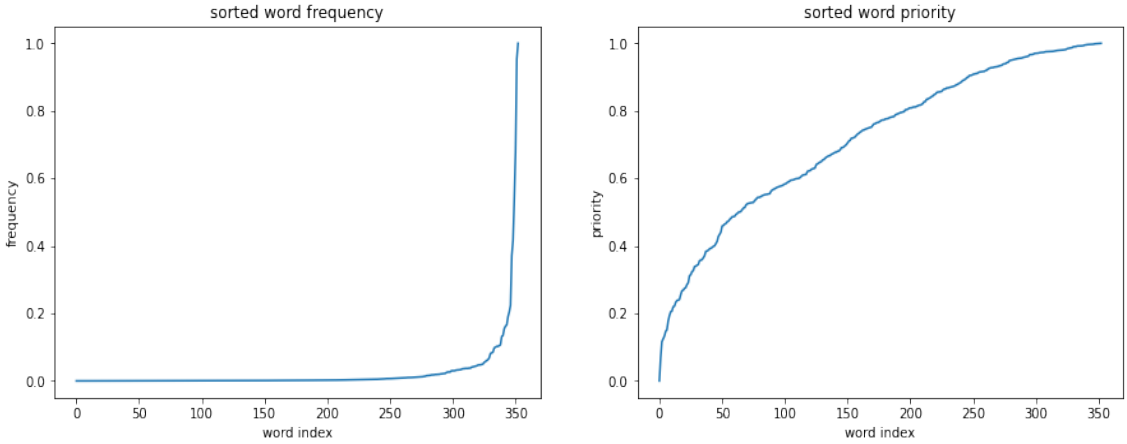

Figure 10: The Values of Sorted Word Frequency and Priority

Figure 3 compares the distribution of sorted words' frequency and priority, in the given dataset. The priority is less steep than frequency, makes words with medium frequency more likely to be chosen. We use $p(\omega)$ as a feature in BP neural network.

4.1.3 Elimination Value

Most people use common opening word to start a wordle game. The opening word is the word you first guess in a game. It is important as it eliminate the most possible words during the whole game.

The best opening word is proved to be "salte" by information theory. It gives the most information gain over all opening words. We define Elimination Value $E(\omega)$ by setting answer as ω, then use "salte" to play one step, then count how many words are still available after the guess. $E(\omega)$ equals the word number counted. For example, if the result of "a" is green tile and other tile is grey, then "aroma" is available while "apple" is not.

It's obvious that less remaining word means less steps needed after first step. So, $E(\omega)$ is counted as a feature.

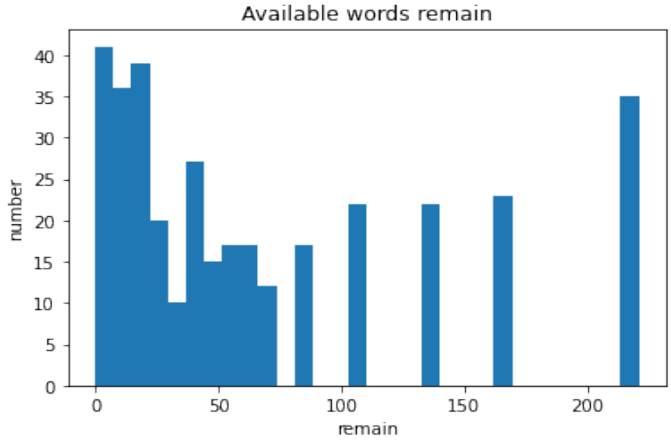

Figure 11: Distribution of Elimination Value

4.2 Correlation Test

The correlation of features and time to percentages is tested by Pearson correlation coefficient.

As a result, all features is considered not moderately linearly correlated to percentage(no $|r| > 0.5$ case). The total result is huge, so we only put the coefficient calculated by features and average percentages for an example:

Table 2: Pearson correlation coefficients

Contest Number	$E(\omega)$	$C(\omega, 1)$	$C(\omega, 1)$	$p(\omega)$
-0.071	0.3	0.09	0.018	-0.304

As the correlation is non-linear, we decided to use BP neuron network for regression.

4.3 BP Neuron Network

For the non-linear model, we tried some machine learning algorithms, and BP neuron network gives the best result.

BP Neural Network is a nonlinear, multi-layer system, which can build nonlinear model with multiple inputs and outputs. BPNN contains three layers: input layer, hidden layer and output layer, showed in Figure 5.

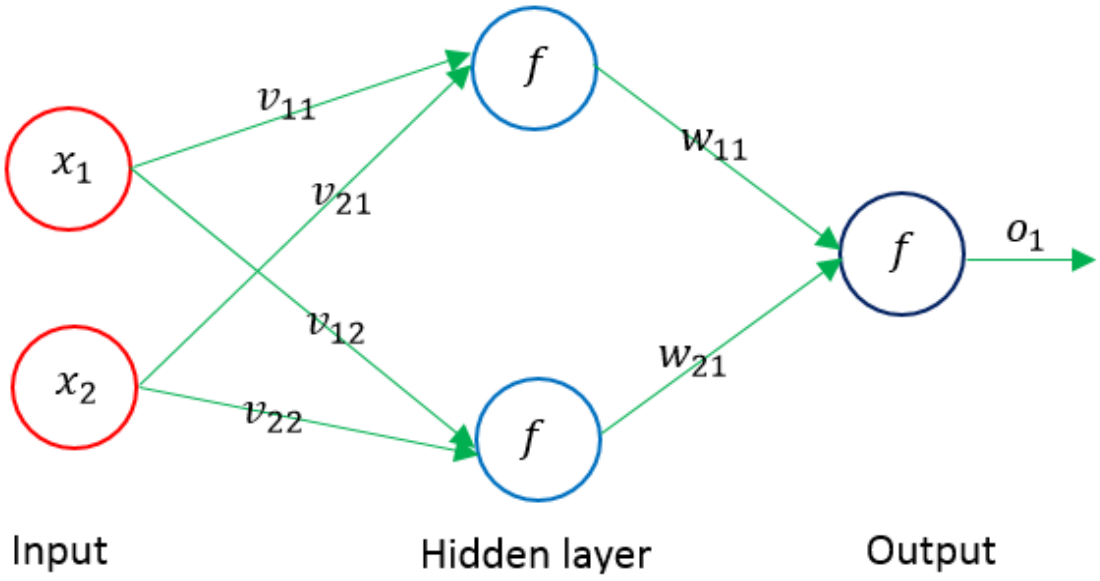

Figure 12: BP Neuron Network

The BP Neural Network and training is set as:

- input: $C(\omega, 1)$, $C(\omega, 2)$, $p(\omega)$, $E(\omega)$
- output: percentage of attempts

- 70% training set, 30% testing set
- Activation Function: ReLU
- Learning rate: 0.1
- one hidden layer with 50 neurons

The evaluation of the network is listed in Figure 9 and table below:

Table 3: BP neuron network evaluation result

	MSE	RMSE	MAE	MAPE	R^2
training set	12.172	3.489	2.567	51.565	0.21
test set	12.807	3.579	2.659	45.281	0.258

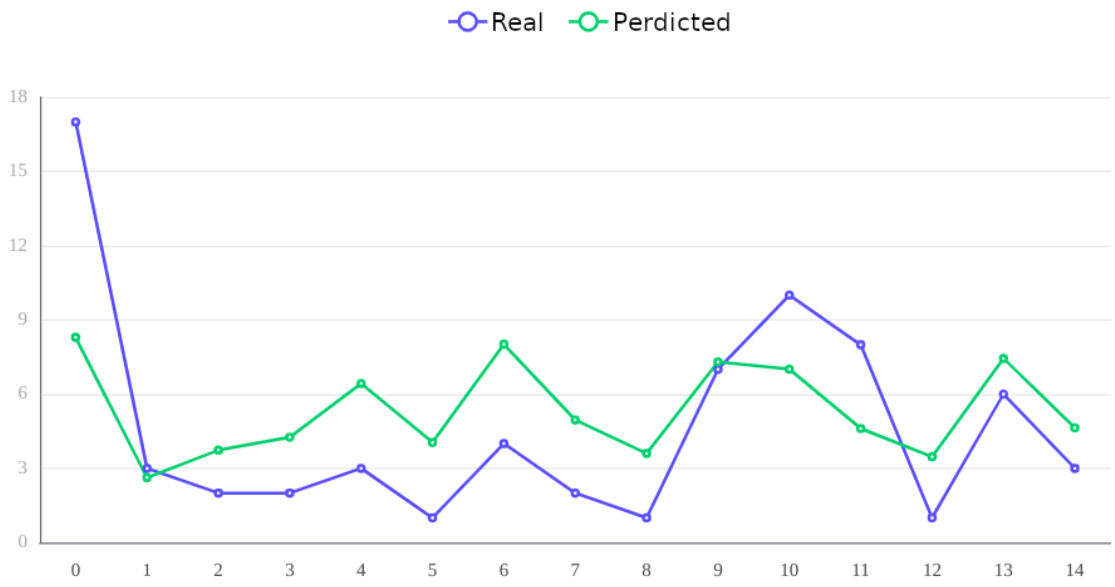

Figure 13: BPNN Over Test Set

The result is the best over some machine learning methods, for example, GBDT ends with overfitting(training set $R^2 = 0.99$, test set $R^2 = -0.18$), SVR has more error(negative R^2 for bot testing and training set). The model used here can be significantly improved with more dataset.

4.4 Results

4.4.1 Uncertainties

Our model has certain uncertainties:

- Current dataset is quite small (about 359 lines) for all machine learning algorithms. The result is uncertain in such small set.

- The correlation between features and output is non-linear. There's no strong proof for the correlation.
- Real wordle game use a special subset of all 5 letter words. Analysis on all 5 letter words may have error.

4.4.2 Predictions

The percentage of attempts for word "EERIE" predicted by our model is:

Table 4: EERIE number of tries prediction

	1	2	3	4	5	6	7 or more
raw	0	5.80537	22.78431	33.137455	23.80060	11.63913	2.83313
truncated	0	6	23	33	24	11	3

5 Words Difficulty Classification Model

To establish a model to evaluate the difficulty of words in the Wordle game. The evaluation of difficulty is currently based on subjective factors and the indicators and standards for evaluation are difficult to determine in a qualitative way, making it an unsupervised problem.

To use existing data to classify words as an alternative to subjective evaluation. The data that can serve as classification criteria is tries number, and the proportion of each component of the distribution can reflect the difficulty of the word to some extent. We also had planned to use the participation rate in the difficult mode as a classification criterion, but based on the results of the first question, we think that the word itself does not affect the participation in difficult mode, so we will not use the number of participants in difficult mode as a classification criterion.

Therefore, our approach is to use the proportion of tries in the existing data to classify words into three levels of difficulty.

5.1 Introduction to K-means clustering

We need a model that can solve unsupervised problems, is good at dealing with small data sets, and can determine the number of classes.

5.1.1 Brief introduction

The K-means algorithm is a clustering algorithm that divides data points into K clusters by computing the distances between them. The algorithm starts by randomly selecting K cluster centers and assigning data points to the cluster with the nearest center. Then, for each cluster, its center is updated until a certain termination condition is met. The result is K clusters, where each cluster contains data points that are closest to each other and furthest from the points in other clusters.

Given a set of observations $(x_1, x_2, ..., x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq$ n) sets S = $S_1, S_2, ..., S_k$ so as to minimize the within-cluster sum of squares (WCSS). Formally, the objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{13}$$

where μ_i is the mean (also called centroid) of points in S_i

$$\boldsymbol{\mu_i} = \frac{1}{|S_i|} \sum_{\mathbf{x} \in S_i} \mathbf{x} \tag{14}$$

5.1.2 Apply to the data

The number of tries from 2 to x times was taken as the dimension of clustering. The words were clustered from these six dimensions and the k-value was located at 3.

5.2 The Model Result

5.2.1 Cluster center coordinate

The three clusters obtained by the clustering algorithm are in the center of the six dimensions, and the average number of attempts of the three classes is calculated, ranking them in the simple, medium and difficult categories. The specific data results are shown in the table below:

Table 5: Cluster center coordinates (When calculating the average tries, let x=7)

	cluster1	cluster2	cluster3
try2	3.98649	9.32331	2.86111
try3	20.27027	30.64662	12.77778
try4	35.73649	33.66917	25.98611
try5	26.39189	17.90226	28.86111
try6	11.41892	6.52632	21.33333
tryX	1.93243	1.10526	7.83333
average tries	4.25730	3.81669	4.75139
Difficulty	medium	Easy	hard

5.2.2 Scatter distribution of words

After determining the difficulty of the three clusters, we plotted the distribution of the words involved in the data table in the three clusters, as shown in the scatter plot:

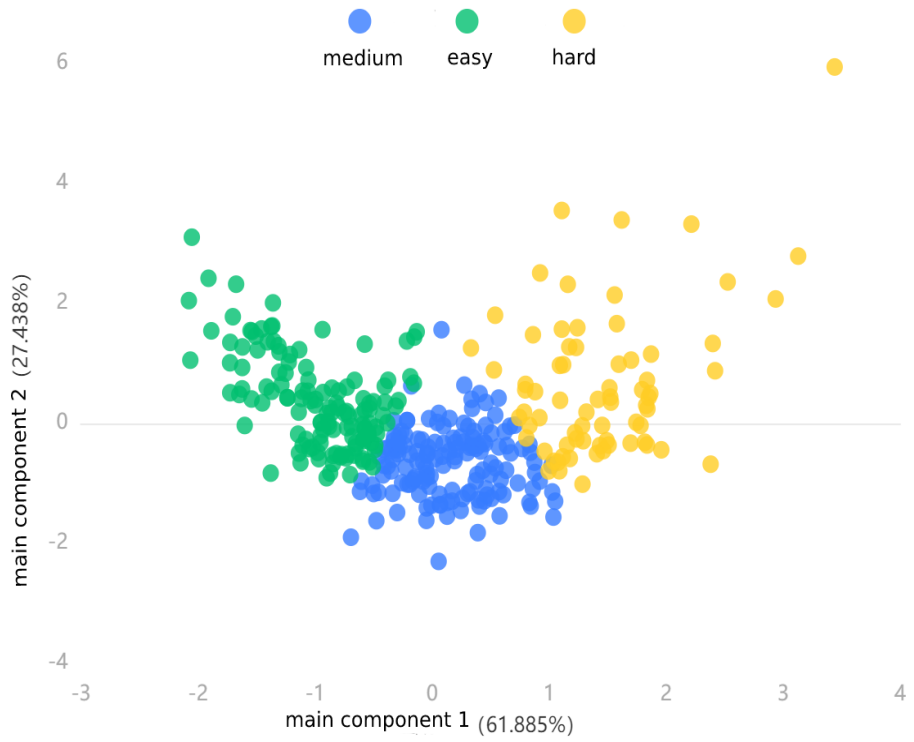

Figure 14: Cluster scatter plot

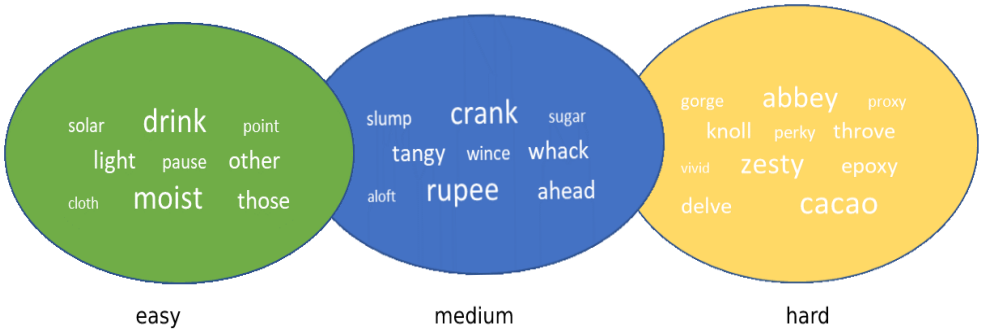

Figure 15: Words scatter plot

5.2.3 Cluster generalization

Based on the center point of the three clusters, it is possible to evaluate which cluster any word belongs to after estimating the distribution of the proportion of attempts for any word:

$$min \sum_{n=2}^{X} (x_n - \bar{x_{ni}})^2 \tag{15}$$

x_n represents the proportion of NTH attempts, and x_{ni} represents the central value of the NTH attempt for a cluster,We need to substitute it into three groups of clusters to calculate, which cluster gets the smallest value, indicating which cluster the word belongs to.

5.2.4 EERIE difficulty classification

From the second question we predict that the frequency of the possible distribution of the number of EERIE word tries is:

Table 6: EERIE number of tries prediction

	try1	try2	try3	try4	try5	try6	tryX
EERIE	0	5.80537	22.78431	33.137455	23.80060	11.63913	2.83313

The above data were substituted into the model for calculation: the distance square of EERIE for simple clusters was 138.57027, for medium clusters was 24.75103, and for difficult clusters was 302.27994.The medium cluster has the smallest value, so EERIE should be of medium difficulty.

5.2.5 Accuracy of model

- Accuracy depends to some extent on the accuracy of the second question
- Clustering only gets the better solution from the distance, which is not necessarily the best classification criterion to distinguish the difficulty of words

6 Strengths and Weaknesses

6.1 Strengths

- Model for predicting number of participants by ARIMA is adaptive to the dataset and has ability to produce accurate forecasts. The result is reasonable, proved by various tests.
- Different features are extracted from words to build the percentage prediction model. These features are properly based on NLP and information theory, make the model more reasonable.
- Fine tuned BP neuron network is used to solve the non-linear relationship between word features and percentages. BPNN is proved to be adaptive and suitable for such complex problem.
- Difficulty classification is based on K-means clustering on tries number. This avoid subjective evaluations on difficulty.

6.2 Weaknesses

- ARIMA model is easy to overfit. The model will fit the training data too close, make poor prediction on new data. Even the test shows that our model performs well, the model may not handle some situations.
- Our features on percentage prediction model is non-linear, makes it hard to verify how features contribute to the model. The dataset is also too small for any machine learning algorithm. That leads to a poor performance of the model.

- The relationship between difficulty and tries percentage is unknown. The result of K-means is not based on theories and may have some error.

6.3 Model Promotion

All models can be significantly improved by adding more complete data.
Beside, For each specific model:

- Model for predicting number of participants can be improved by using other advanced model like GARCH model.
- BPNN can be further tuned by change hyperparameters and perprocessing data. More proves and features also can be included to train BPNN.
- Using k-means to classify difficulty is too intuitive and maybe inaccuracy. A objective difficulty model combined with characteristic of word and tries percentage is more suitable for the problem.

7 Conclusion and Letter

Our team successfully developed reasonable models to solve the problems. The detailed conclusion, analysis and data are shown below in the letter.

7.1 Letter to Editor

To: Editor of the New York Times
From: MCM Team 2316429
Subject: Prediction and summary for Wordle game data
Date: February 20, 2023

Dear the Puzzle Editor of the New York Times:

For the problems that need to be solved, we establish models to solve the corresponding problems, based on the time series model ARIMA to solve the prediction of the reported results number, based on the BP neural network model to establish the proportional association between words and the number of tries, and based on the k-means clustering model to classify the difficulty of words.

Here are the results:
Forecast the contest number

- In the prediction reported results number problem, we analyze it based on the time series, and forecast backwards from December 31, 2022, predicting the reported results number on March 1, 2023: the number of people on the 60th day backward, our model prediction is a fixed value, and we take the value of the previous day and the day after March 1st to simulate its range. The range of the reports number on March 1, 2023 will be 10,288 – 10,624.

- When assessing whether the proportion of selected hard mode is affected by words, we first removed the anomalous data, including words with only four letters: tash, clen, and non-existent words: marxh, and study where the proportion of hard mode and normal mode was extremely anomalous. After that, we analyzed the remaining proportions and found that the data in this set of data was a certain linear distribution and relatively stable, and it was believed that the words do not affect the percentage of scores reported that were played in Hard Mode.

Forecast the associated percentages of tries number

- In predicting the associated percentages of (1, 2, 3, 4, 5, 6, X) for a future date with a given word. First, we rounded the data with the number of tries of 1 and did not count it into subsequent analysis. The result was uniformly considered to be 0. After that, we built the BP neural network model to make predictions. The isolation level and priority of the word were selected to build the neural network. However, the results predicted by the model will be different from the real situation, and the specific uncertainty is caused by the fact that the number of data sets given is too small to prove that the correlation between features and outputs is linear. Finally, the associated percentages of tries numbers of EERIE were predicted: 0, 6, 23, 33, 24, 11, 3.

Classify word difficulty

- When classifying words by difficulty, the data with the number of tries of 1 is also rounded out and is not counted in the analysis. After that, After that, we used the k-means clustering algorithm. The remaining six tries numbers were clustered as six dimensions. At the same time, let K=3, respectively, easy, medium, difficult. The center points of the three clusters we get are detailed in Table 2. Finally, the associated proportion of tries number to predict EERIE is substituted into the model, and the squared distances from the three clusters are 138.57, 24.75, and 302.28. So EERIE should be medium difficulty.

Yours sincerely,
Team 2316429

References

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

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

[3] G. M. Ljung; G. E. P. Box (1978). "On a Measure of a Lack of Fit in Time Series Models". Biometrika. 65 (2): 297–303. doi:10.1093/biomet/65.2.297

[4] Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". Knowledge and Information Systems. 52 (2): 341–378. doi:10.1007/s10115-016-1004-2. ISSN 0219-1377. S2CID 40772241

[5] 3Blue1Brown. Solving Wordle using information theory [Video]. YouTube. https://www.youtube.com/watch?v=v68zYyaEmEA&ab_channel=3Blue1Brown

[6] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.