Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2316597

"The art of simplicity is a puzzle of complexity."
— Douglas Horton

The World of Wordle: An Analysis of Wordle Game Based on Machine Learning

As Douglas Horton said, cramming complexity into simplicity is the art of intelligence. Wordle, as a brilliant combination of mathematics and computer science, is the best illustration. In this paper, we will mainly address three basic and highly correlated problems: the prediction of daily report number, the prediction of the distribution of number of trials, and the classification of word difficulty.

In the first part which requires us to predict the number of reported results and observe the percentage of those played in Hard Mode, we use an ARIMA model for time series analysis, and an NLR model for curve fitting. The ARIMA model achieved a mean squared error of 2.91×10^{-3}, and the NLR model achieved approximately 2.00×10^{-3}, which are both within the acceptable range. Furthermore, they led to very similar prediction result on the number of reported results for the target date March 1st, 2023. More specifically, one predicted 16756 and the other predicted 16774. We also discovered the relationship between the occurrences and repeatedness of letters with the proportion of results reported in Hard Mode.

In the second part we proposed several new attributes of the words and preform the Pearson Correlation Analysis to select the ones that can affect the distribution of the number of trials used for guessing a word. The finally selected attributes are the Variety of Letters (VOL), the Expectation of Yellow Hit (EYH), and the Expectation of Green Hit (EGH). Using these three attributes and the position information of the letters, we train a Decision Tree Regression (DTR) model and a Random Forest Regression (RFR) model to predict the distribution of the number of trials. The two mutually corroborating models both achieved mean absolute errors of at most 3%, and produced almost identical prediction results for the distribution of number of trials for the word "EERIE". It is a Gaussian-like distribution, with the most players using four trials to succeed.

In the third part we developed two classification models of word difficulty. The Expectation of Number of Trials (EOT) model simply treated the number of trials as a random variable and computed its expectation. Selecting the quartiles as boundaries, EOT can provide a quite intuitive yet powerful classification of word difficulty. The other model is based on k-Means, where we add predictions in the previous part to the attributes of the words and perform an unsupervised clustering. The two models almost agreed on the difficulty level of the word "EERIE", which is a little bit over the third quartile if we sort the words in an ascending order of difficulty and overall considered as hard.

In the fourth part we analyze some additional patterns of word difficulty. We found some combinations of letters that tend to lead to medium or easy words. We also analyzed the aforementioned attributes VOL, EYH, and EGH. Large VOL values generally lead to high difficulty levels, and low EYH and EGH values often correspond to difficult words.

Keywords ARIMA • Nonlinear Regression • Decision Tree • Random Forest • k-Means

Contents

1 Introduction

1.1 Background

Wordle is a web-based word game created and developed by Welsh software engineer Josh Wardle and owned and published by the New York Times Company since 2022. It prevailed since its first release and many people still spend some time every day to play this game as a relax. Wordle releases a five-letter solution word for guessing every day, without any prompts, and it entitles six attempts to each player. To make difference from blind guessing, Wordle further provides a subsequent hint to the player after each guess. Any letter that is placed correctly in the position in each guess will be marked as green. Those letters that exist in the solution word but are misplaced will be marked as yellow. Letters that do not appear in the solution word will be marked as grey. A major restriction of this game is that each guess must be a word exist in its word bank, containing 2315 relatively common 5-letter English words, which imposes a challenge on the players' vocabulary. But with proper strategies, players will be able to significantly narrow down possible solutions, and eventually come up with a correct answer.

Wordle also provides a Hard Mode, in which players must use the previous hint in the succeeding guess. This rule prohibits players from guessing completely different words in purpose of narrowing down the range of possible solutions more significantly. An example is shown as in Figure 1. For instance, after the second guess, I may try "CLERK" in the normal mode to exclude more possibilities, but this is not allowed in Hard Mode since I would have to use "A", "S", and "T". In fact, after guessing "CLERK", the only possibilities left are "STEAM" and "SWEAT", so it would be an $1/2$ probability to win with 4 trials and otherwise with 5 trials.

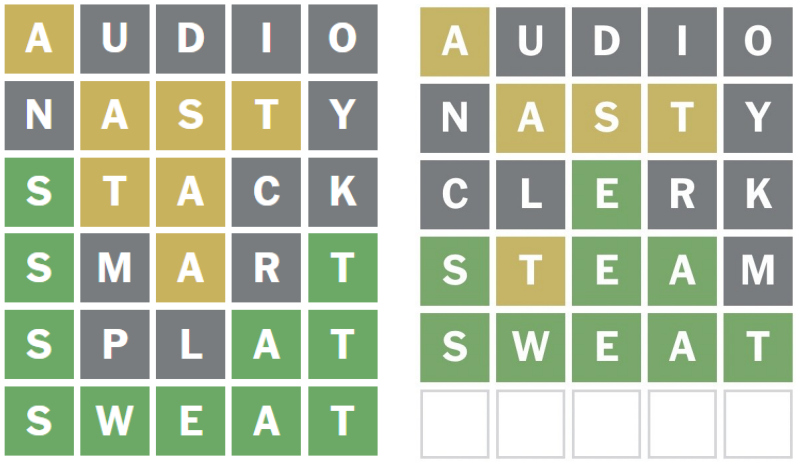


Figure 1: An example of Wordle played in Hard Mode (left) and in normal mode (right).

1.2 Problem Restatement

Some players may report their results, which is then collected by New York Times. New York Times releases daily statistics including the solution word for the day, the number of reported results, the

number of reported results that were played in Hard Mode, and the distribution of the number of trials for guessing the correct answer. Based on these statistics, we are asked to analyze and predict the number of reports (and the number of reports in Hard Mode) for the future. Moreover, we also need to develop a prediction model that takes as input a five-letter solution world and provides an approximation on the distribution of the number of trials used by the players to obtain the correct answer. Finally, we need to investigate into the features of the solution words and provide a classification based on difficulty. Other interesting observations on the given data set are also preferable.

1.3 Literature Review

The Wordle game covers a variety of fields around linguistics, big data, machine learning, and psychology. Considering current research relevant to our problem, we discovered that there have already existed mature game-play strategies depending on machine learning, which nearly operates a perfect and optimal trial strategy for robots to solve the Wordle game. Dimitris and Alex have developed a deep learning model to make the guess as quick and accurate as possible [3]. Additionally, there are plenty of research that aimed to make a ranking system for the difficulty of English words based on their commonality. Hence, we construct our model based on machine learning strategies but within the constraint of human natural behaviors, as well as of word difficulty under the limitation of Wordle's word bank and its intimate relevance to the unique guessing system of Wordle.

1.4 Our Work and Organization

In this part, we will briefly introduce the structure of the rest of this technical report. In Section 2, we will introduce some important preliminary assumptions and the notations that will be applied throughout this paper. In Section 3, we will introduce our model for predicting the number of reported results on a future date, as well as our observations on the relation between the words and the percentage of reported results played in Hard Mode among all reported results. We will also use it for predicting the number of reported result on March 1st, 2023. In Section 4, we will propose some attributes for the words, with which we develop a model to predict the distribution of number of trials for a given word. In Section 5, we will introduce our models for classifying the words into different difficulty levels. In both Section 4 and Section 5, we will apply our models to make predictions and classifications on the word "EERIE". Finally in Section 6, we will further discuss the relations between the occurrences of letters in the words and the difficulty level, as well as between the previously proposed attributes and the difficulty level. Section 7 will be the conclusion of our work and some possible future improvements.

2 Overview

2.1 Preliminary Assumptions

We make the following preliminary assumptions that apply to all the models mentioned in this report:

- *We consider the attributes of solution words is the main contributor to the difficulty of the Wordle game, and specifically, we do not take the recognition difficulty of the solution words into account.* Based on our research, the majority of the player in certain language version consists of native speakers, hence we consider the commonality of the solution words to have little impact on defining the difficulty of the word that can be dismissed in our problem.
- *We assume that players all behave naturally and empirically, in other words, we do not consider players' unexpected behaviors. For instance, we do not allow a vast amount of players to use cheating machines or search for plot news instead of behaving on their knowledge and motivations.* Referring to the literature review, there has already existed a mature system of algorithms and strategies purely for machines to churn out optimized solutions in the Wordle game. There is no need nor room for us to make people robots in this problem.

2.2 Notations

Throughout the rest this paper, we will stick to the following notations:

Reported Results	
#RRT	*the total number of reported results*
#RRHM	*the number of reported results played in Hard Mode*
HM/T (%)	*the percentage of reported results played in Hard Mode*
Distribution of Trials	
#TRIALS	*the number of trials used for guessing the word*
VOL	*Variety of Letters* (attr.)
POV	*Percentage of Vowels* (attr.)
$\mathbb{E}$[YH]	*Expectation of Yellow Hit* (attr.)
$\mathbb{E}$[GH]	*Expectation of Green Hit* (attr.)

3 #RRT and #RRHM: Time Series Analysis

Recall that #RRT stands for the total number of reported results and #RRHM stands for the number of reported results played in Hard Mode. In this section, we will introduce our model for predicting #RRT on a future date, and our observation on the relationship between HM/T and the structure of the given word. At the end of Section 3.1, we will provide our prediction of #RRT on March 1st, 2023.

3.1 Predicting #RRT

First let us have an overall view on the change of #RRT with the passage of time, as is shown in Figure 2. Since its first publication by New York Times, the number of reported results has surged to

a peak of over 350,000 in the first two months. After the first wave of fanaticism has passed, #RRT has started dropping gradually until it reached a relatively stable and slowly descending state in around July.

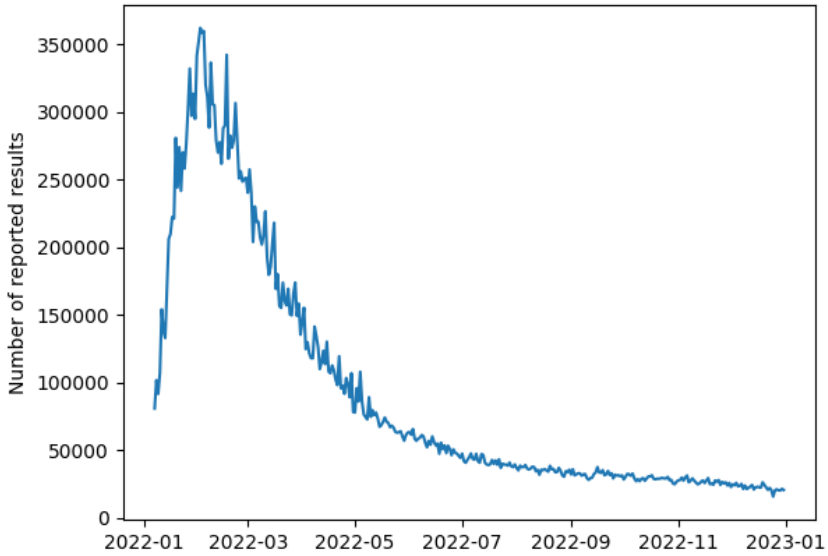


Figure 2: The relation between #RRT and the date.

From the observations above, we can tell that another peak is very unlikely to occur in the future unless some new versions of new features were released. Assuming this, we can wipe out the first 80 dates of data and only consider the descending part. Moreover, since we are going to propose prediction models, in order to compute meaningful prediction errors for model validation, we are will normalize the data set so that all #RRT values lie in the range $[0, 1]$. The parsed and normalized time series is shown as in Figure 3.

In the rest of this subsection, we will use the parsed and normalized time series in Figure 3, and propose two types of models. One is the Autoregressive Integrated Moving Average (ARIMA) [5] model, and the other one is a Nonlinear Regression (NLR) model.

3.1.1 The ARIMA Model

The ARIMA model uses the method of differencing to transform a non-stationary time series with lots of small perturbations into the stationary series, thus providing a prediction for the future. In order to determine if the ARIMA model is suitable for the given data, we run the augmented Dickey-Fuller unit root test, which tests the null hypothesis that there exists a unit root for this time series. The resulting p-value is 1.48×10^{-8}, small enough to imply the statistical significance, suggesting that the null hypothesis should be rejected. The absence of a unit root further indicates the stationarity of the data, so we can safely apply the ARIMA model. Now it suffices to determine the orders of differencing, regression, and moving average respectively. To do this, we refer to the autocorrelations and partial autocorrelations, which are shown as in Figure 4.

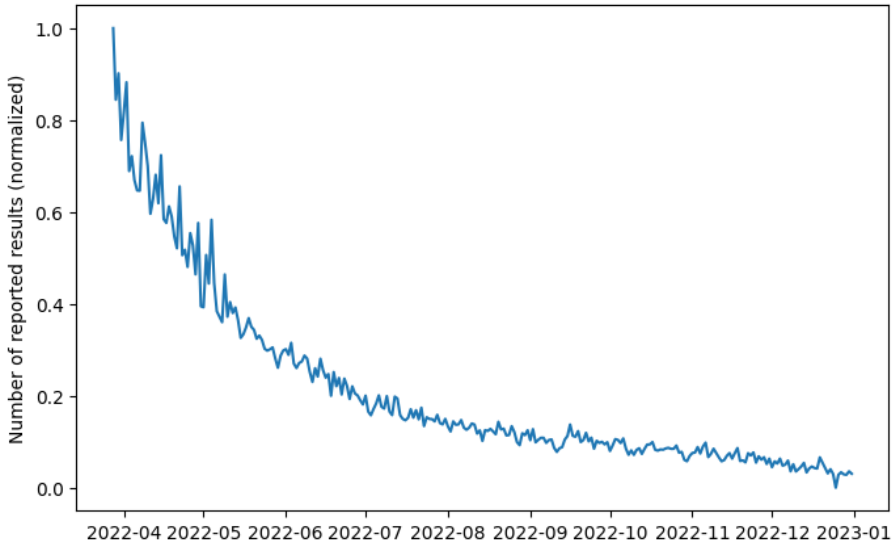


Figure 3: The relation between normalized # RRT and the date, excluding the peak at the beginning.

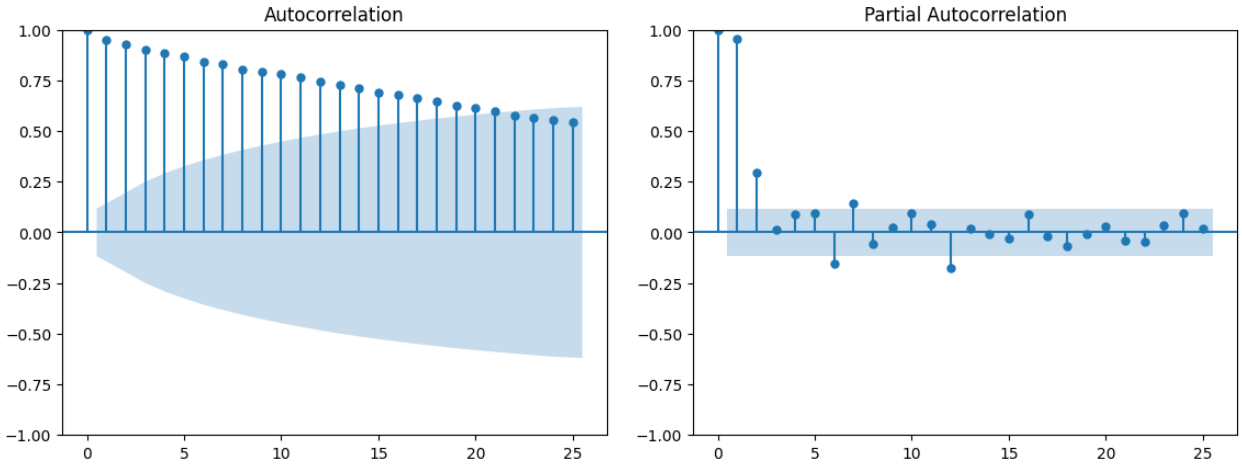


Figure 4: The autocorrelations (left) and partial autocorrelations (right) of the time series.

From Figure 4, the autocorrelations are large for all values of lags, but it could be only due to the propagation effect of small values of lags. Indeed, if we look at the plot of partial autocorrelations, there are significant spikes only with lag of 1 and 2. Based on the observations above, we determine our model to be `ARIMA(2,2,5)`, which means that the orders of regression and differencing are both 2, while the order of moving average is 5. The Akaike Information Critera (AIC) for `ARIMA(2,2,5)` is evaluated to be -1161.979, a pretty satisfying value. Now we split the time series into a training set and a testing set, with the testing set involving data of the last month (*i.e.*, December, 2022) and the training set involving the rest. The training and prediction results using the `ARIMA(2,2,5)` model is shown as in Figure 5.

To validate our model, we checked the mean-squared error between the prediction and the true # RRT values on the testing set (the orange part in Figure 5). The error is reported to be approximately

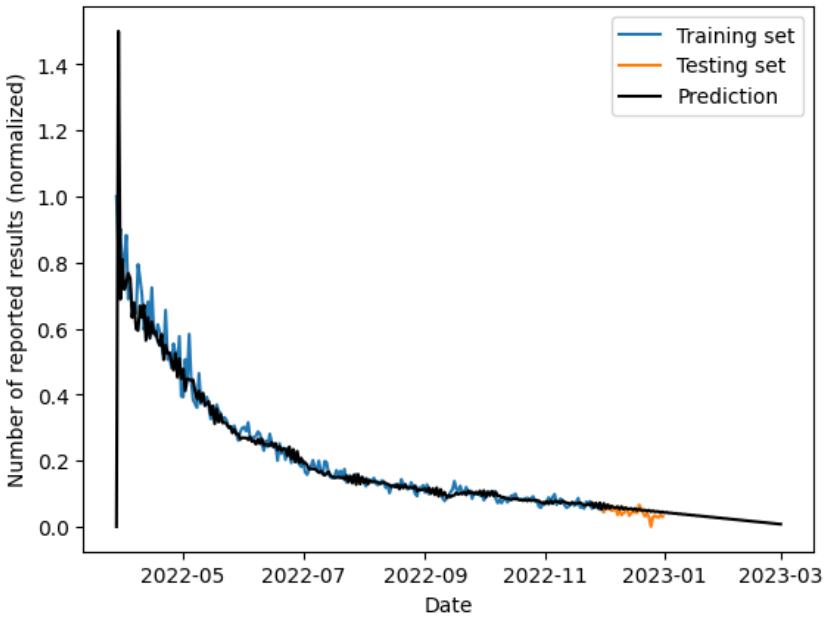


Figure 5: The training and prediction results of the normalized # RRT using the `ARIMA(2,2,5)` model.

2.91×10^{-3}, meaning that the prediction is a gives a decent stationary fitting of the testing set data. If we extend the prediction to March 1st, 2023, the target date, and transform back to the original scale, we can obtain that

$$\#\,\mathrm{RR_T}(2023/03/01) \approx 16756. \tag{1}$$

3.1.2 The NLR Model

We are aware that the ARIMA model can only provide stationary prediction, and are commonly used only for short-term prediction. With this in mind, we will use an NLR model to support the ARIMA prediction (1). We will use nonlinear least squares to fit a specific class of curves onto the time series, so it suffices to determine the class of curves to use. Judging from the shape of the data points, we should use some curves that have similar shapes to an inversely proportional curve. The class of curves we select are in the form of

$$f(x) = a(x+b)^{-2/3} + c, \tag{2}$$

where a, b, c are arbitrary parameters for fitting. We pick this class of curves over other classes such as negative exponential functions after massive experiments, and the power $-2/3$ is chosen deliberately so that the error on the test set can be minimized after training. Now we split the time series into the same training set and testing set as we did in the ARIMA model, train our NLR model on the training set, and validate it using the mean squared error on the testing set. The visualized result is shown as in Figure 6. Note that the dates are treated as consecutive integers starting from 0 in this model for the sake of simplicity.

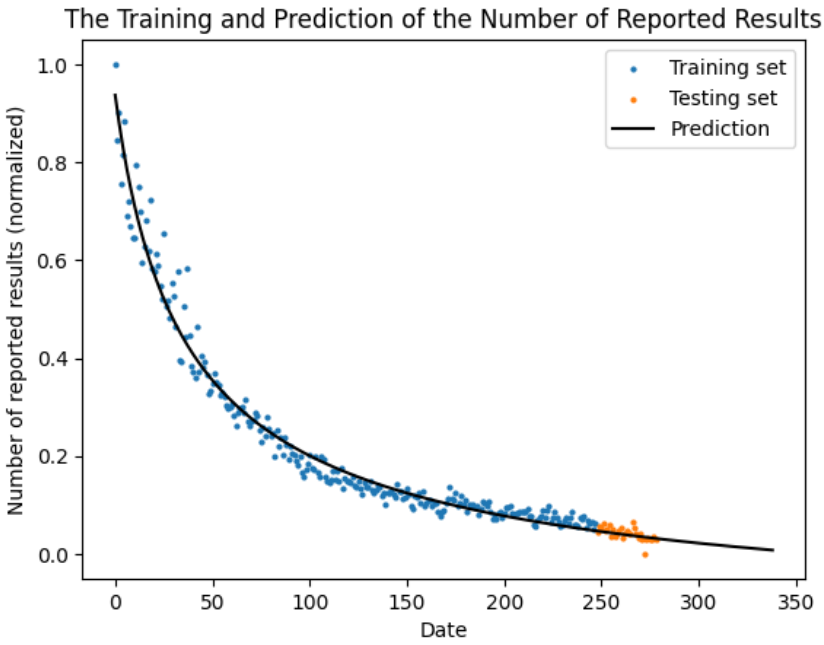


Figure 6: The training and prediction results of the normalized #RRT using the NLR model, fitting with curves in the form of $f(x) = a(x+b)^{-2/3} + c$.

The trained curve, as in shown in Figure 6, is approximately

$$f(x) = \frac{236}{25}\left(x + \frac{616}{25}\right)^{-2/3} - \frac{89}{500}. \tag{3}$$

The mean-squared error between the prediction and the true #RRT values on the testing set (the orange part in Figure 6) is reported to be approximately 2.00×10^{-3}, which is even smaller than that of the ARIMA model. If we extend the prediction to March 1st, 2023, the target date, and transform back to the original scale, we can obtain that

$$\#\mathrm{RR_T}(2023/03/01) \approx 16774. \tag{4}$$

This value is very close to the prediction result (1) using the ARIMA model, adding to the confidence of our models. Given that predicting #RRT three months in the future using only fewer than a year of data, we allow a ±10% error to our prediction, leading to the final prediction interval of approximately [15000, 18000].

3.2 Influencing Factors of HM/T

Recall that

$$\mathrm{HM/T}\ (\%) = \frac{\#\mathrm{RR_{HM}}}{\#\mathrm{RR_T}} \times 100\%, \tag{5}$$

representing the percentage of reported results played in Hard Mode among all reported results. In this part, we will discuss some features of the words that may affect HM/T.

Letter occurrence. The occurrence of some specific letters may be related to the HM/T value. To see this, we will break each word down into five letters and calculate for each letter its corresponding average HM/T value as the arithmetic mean over the HM/T values of all the words that include this letter. The plot is shown as in Figure 7.

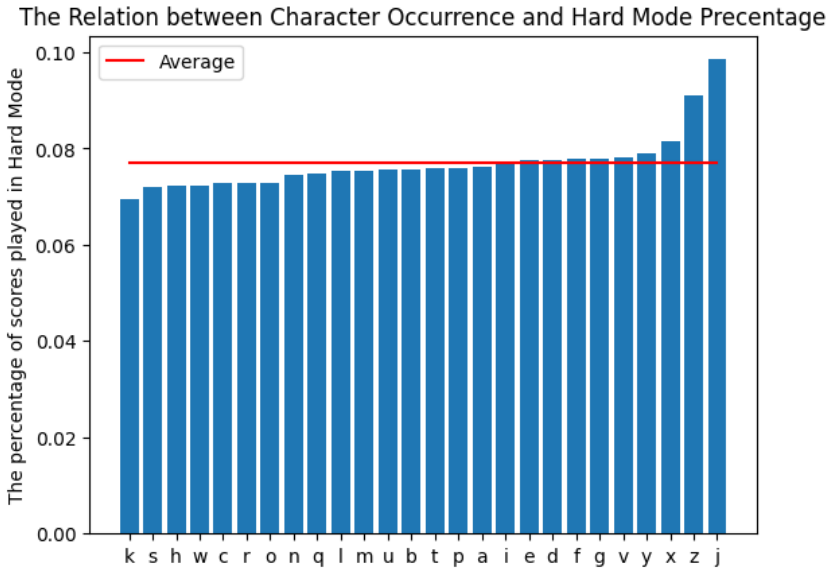


Figure 7: The average HM/T value of each letter among all occurrences, sorted in ascending order.

From Figure 7, we can see that the average HM/T values of the letters "J", "Z", and "X" are significantly higher than that of the other letters. The average HM/T values of "Y", "V", "G", "F", "D", and "E" are also above the average.

Letter frequency. The frequency of the five letters in a word, *i.e.*, how often those letters occur among all words chosen by Wordle, may be another influencing factor on the HM/T value of the word. To see this, we compute the average HM/T value of each letter in the same way as above, and count its frequency of occurrence. The trend can be vague since the size of the alphabet is small, so we use a nonlinear curve to fit the data points so as to reveal the trend. The plot is shown as in Figure 8.

From Figure 8, we cannot tell the relation of most between the frequencies and the average HM/T values of most of the letters. However, for the letters that occur rarely, we can see that the average HM/T are significantly higher. This is indeed reasonable: once a player discover that a letter of critically low occurring frequency exists in the word, he/she is likely to stick to this clue since this single clue can narrow down the possible word candidates a lot due to its rareness.

4 The Distribution of # TRIALS: Regression Analysis

Recall that # TRIALS stands for the number of trials used for guessing the correct word of the day. In this section, we will introduce and determine some attributes of words that influence the distribution

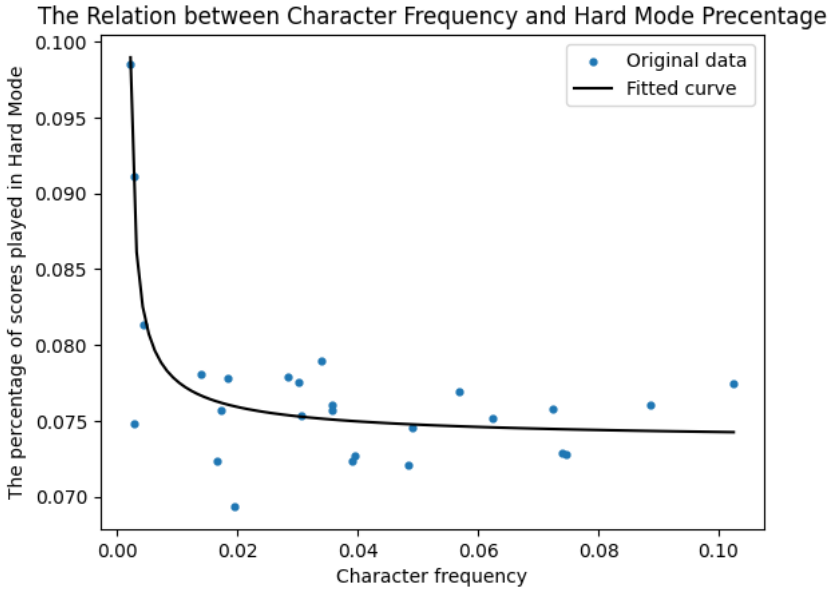


Figure 8: The average HM/T value of each letter among all occurrences, sorted in ascending order.

of #Trials, and further develop corresponding models to predict that distribution. At the end of this section, we will provide our prediction of the distribution of #Trials for the solution word "EERIE".

4.1 Word Attributes Model

In this section, we first propose a set of new attributes that may affect the distribution of #Trials. By performing the Pearson correlation analysis, we will determine which one(s) of these attributes may serve as an influencing factor on the distribution of #Trials. Selected attributes will be used for further analysis.

4.1.1 Definitions of Word Attributes

Variety of Letters (`VOL`). To aid in our analysis of #Trials, we introduce a novel attribute called the "Variety of Letters" (`VOL`). Under such consideration, we suppose many players tend not to repeat the same letter after it is marked green or yellow. For instance, "MUMMY" is a rather simple word, but few people got the answer within four trials, a larger proportion of people tend to use more trials despite the commonality of this word. The probable reason is because the effect of its low `VOL`. In our model, this attribute is calculated by viewing a five-letter word as a probability distribution of the 26 letters in the English alphabet, where the probability of each letter is assigned based on its frequency in the word. Then, we calculate the information entropy $H(X)$ of the probability distribution to obtain `VOL`. For each word w, the `VOL` attribute can be formulated as

$$\texttt{VOL}(w) = -\sum_{\text{letter } l \in w} p_x(l) \cdot \log p_x(l), \tag{6}$$

where for each letter l, we denote that

$$p_x(l) = \frac{\text{\# of letter } l \text{ in } w}{\text{the length of } w}. \tag{7}$$

Percentage of Vowels (POV). We introduce another attribute called the "Percentage of Vowels" (POV), which is calculated as the percentage of vowels in a given word. Most of the players tend to use words with a high POV such as "AUDIO" for their first trial, and if the real word also has a high POV, players will be able to gain a lot of information from the very beginning. For each word w, the POV attribute can be formulated as

$$\texttt{POV}(w) = \frac{\text{\# of vowels in } w}{\text{the length of } w}. \tag{8}$$

Expectation of Yellow & Green Hit ($\mathbb{E}$[YH]& $\mathbb{E}$[GH]). To conduct our analysis, we use a word bank [9] sourced from the New York Times, which includes all the words that are allowed in the Wordle game. We assume that players will randomly select words from this word bank during the game. Since yellow hits and green hits provides much benefit for finding a solution, we assume $\mathbb{E}$[YH] and $\mathbb{E}$[GH] have a negative correlation with game's difficulty. That is, when the solution have higher $\mathbb{E}$[YH] and $\mathbb{E}$[GH] people tend to solve this game in less trials. For each word w, the $\mathbb{E}$[YH] and $\mathbb{E}$[GH] attributes can be formulated as

$$\mathbb{E}[\texttt{YH}](w) = \sum_{\text{letter } l \in w} \frac{\text{\# of words in the word bank that have letter } l}{\text{the size of the word bank}}, \tag{9}$$

$$\mathbb{E}[\texttt{GH}](w) = \sum_{\text{letter } l \in w} \frac{\text{\# of words in the word bank with letter } l \text{ in the same position as in } w}{\text{the size of the word bank}}. \tag{10}$$

Seen from these formula, we know that $\mathbb{E}$[YH] and $\mathbb{E}$[GH] can also describe how close a word is to the word bank. Thus, words with higher $\mathbb{E}$[YH] and $\mathbb{E}$[GH] are usually words more familiar to players. Some examples of these proposed attributes are shown as in Table 1. We can see that VOL is the highest when all the letters in a word are distinct. POV is high for words containing many vowels. $\mathbb{E}$[YH] and $\mathbb{E}$[GH] are more interesting and lack an intuitive interpretation without revealing the word bank.

Words	VOL	POV	$\mathbb{E}$[YH]	$\mathbb{E}$[GH]
MUMMY	0.9503	0.2	0.5067	0.3400
MIMIC	1.0549	0.4	0.6020	0.2417
APPLE	1.3322	0.4	1.2772	0.3647
AUDIO	1.6094	0.8	1.3209	0.2668

Table 1: Some examples of the attributes VOL, POV, $\mathbb{E}$[YH], and $\mathbb{E}$[GH].

4.1.2 Pearson Correlation Analysis

In this part, we aim at selecting valid influencing factors on the distribution of #Trials from the attributes proposed above. To do this, we perform a one-to-one parallel correlation analysis between

the aforementioned attributes and the distribution of # Trials using the Pearson correlation coefficient, in order to check their credibility and validity respectively. In the Pearson correlation analysis, the Pearson coefficient p is defined as

$$p = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \cdot \sum(y_i - \overline{y})^2}}. \tag{11}$$

The results of pairwise Pearson correlation analyses are shown in Figure 9. We apply the interpretation in [8], such that $p > 0$ implies a positive correlation, while $p < 0$ implies a negative correlation. Moreover, $|p| \geq 0.5$ typically implies a strong correlation, $0.3 \leq |p| < 0.5$ a moderate correlation, and $|p| < 0.3$ a weak correlation.

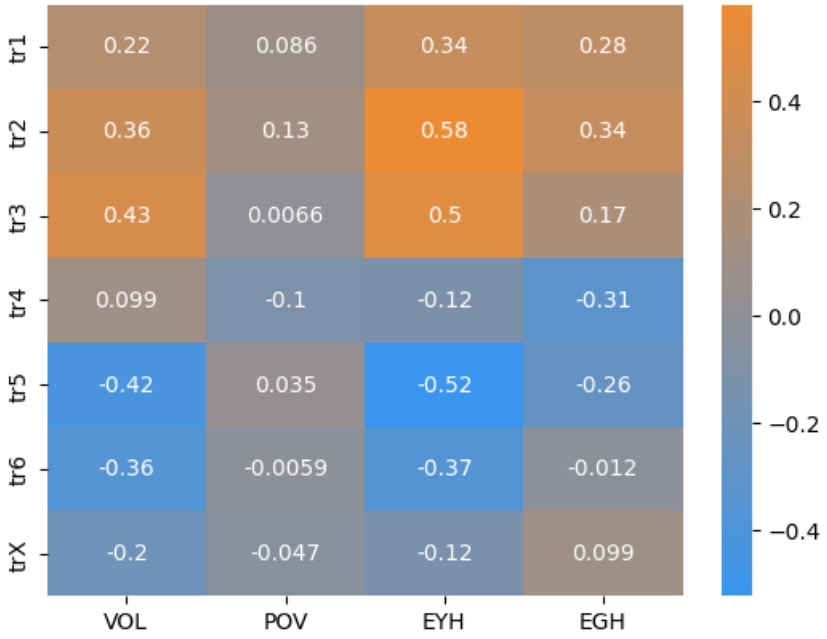


Figure 9: The Pearson correlation coefficients between the attributes VOL, POV, $\mathbb{E}$[YH], and $\mathbb{E}$[GH]and the distribution of # Trials.

From Figure 9, we can see that orange colors indicate positive correlations and blue colors the inverse. Moreover, the brighter the color is, the stronger the correlation exist between attributes and hm/t (%). On one hand, due to the weak correlation between POV and each # Trials, we do not consider it as a proper influencing factor of the distribution of # Trials. On the other hand, VOL, $\mathbb{E}$[YH], and $\mathbb{E}$[GH] are all at least moderately correlated with some of the # Trials. Such correlations are especially strong if we consider the attributes VOL and $\mathbb{E}$[YH], and when # Trials $= 2, 3, 5, 6$. It is also reasonable that # Trials $= 1, 4, \geq 7$ are not as strongly correlated with these attributes as other values of # Trials. Indeed, players that can blindly guess the correct word within the first trial are always rare, and most of the players will be able to obtain the correct answer with four guesses given that # Trialsfollows a Gaussian-like distribution.

To conclude, the attribute POV will be dropped and the attributes VOL, $\mathbb{E}$[YH], and $\mathbb{E}$[GH] will be applied in the following sections.

4.2 Predicting the Distribution of # Trials

Aside from the attributes VOL, $\mathbb{E}$[YH], and $\mathbb{E}$[GH] which we previously determined as influencing factors of the distribution of # Trials, we also break each word down into five letters and assign them corresponding numerical values such that $(\mathrm{a}, \cdots, \mathrm{z}) \mapsto (0, \cdots, 25)$. The specific letters in the specific positions can hardly be replaced by any statistical evaluation, thus playing an essential role in predicting the distribution of # Trials.

To formulate this problem, we define the input space $\mathcal{X}$ and the outcome space $\mathcal{Y}$, such that

$$\mathcal{X} = \{\text{letter } l_i\}_{i=1}^{5} \cup \{\texttt{VOL}, \mathbb{E}[\texttt{YH}], \mathbb{E}[\texttt{GH}]\}, \tag{12}$$

$$\mathcal{Y} = \{\#\,\textsc{Trials}_{=j}\}_{j=1}^{6} \cup \{\#\,\textsc{Trials}_{\geq 7}\}. \tag{13}$$

Then, our goal is to find a prediction function $f : \mathcal{X} \to \mathcal{Y}$. To achieve this, we will use the Decision Tree Regression (DTR) [7] model and the Random Forest Regression (RFR) [4] model.

4.2.1 The DTR Model

We split the data set into 75% for the training set and 25% for the testing set and perform the DTR. Note, however, that the accuracy of the DTR model is largely dependent on the maximum depth assigned to the decision tree. With massive experiments, we determine that decision trees of maximum depth 4 are the best choice for this data set, in the sense that they can provide the least overall mean absolute error on the testing set. The resulting errors of our trained model on the training set and the testing set are shown as below, which are indeed within an acceptable range.

```
Mean absolute error (trial_1): train = 0.2361, test = 0.2000
Mean absolute error (trial_2): train = 1.0632, test = 1.5722
Mean absolute error (trial_3): train = 2.2435, test = 3.0556
Mean absolute error (trial_4): train = 1.8569, test = 1.9389
Mean absolute error (trial_5): train = 1.7602, test = 2.4278
Mean absolute error (trial_6): train = 1.9294, test = 2.2611
Mean absolute error (trial_x): train = 0.8829, test = 1.0611
```

After validating our model, we train it again with the same attributes, but using the whole data set instead of just the training set. We also created "EERIE", the target word for prediction, in the input space $\mathcal{X}$ such that

$$\text{"EERIE"} = \begin{array}{|c|c|c|c|c|c|c|c|} \hline l_1 & l_2 & l_3 & l_4 & l_5 & \texttt{VOL} & \mathbb{E}[\texttt{YH}] & \mathbb{E}[\texttt{GH}] \\ \hline 4 & 4 & 17 & 8 & 4 & 0.9503 & 1.0974 & 0.4573 \\ \hline \end{array}. \tag{14}$$

The prediction result shows that none of the players is likely to guess the word "EERIE" with only one guess. 4% of the players can get the answer within two trials and 18% within three trials. A further 32% and 27% of players can obtain the correct word with four and five guesses, which is consistent with most of the previous data, where the most of the players always fall in the four-trials or five-trials categories. Finally, 15% of the players need to use up their opportunities to obtain the correct answer, while 4% will fail. The predicted distribution of # Trials is shown as in Figure 10.

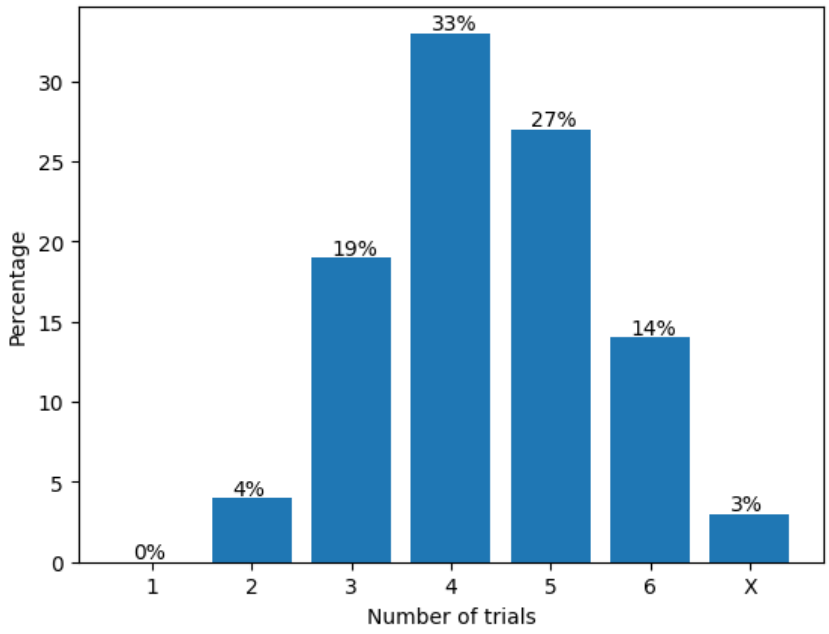


Figure 10: The predicted distribution of # Trials for the word "EERIE" using the DTR model.

4.2.2 The RFR Model

The RFR model, in contrast to the DTR model, randomly samples several decision trees to form a random forest instead of using all data for a single decision tree. We apply exactly the same strategy as that for the DTR model. The overall mean absolute error improves slighly, but the mean absolute error on each trial does not show a significant improvement. The prediction result is also almost identical to that using the DTR model, except for the 1% difference for # Trials = 3 and # Trials = 4, as can be seen in Figure 11. Meanwhile, this again validates the correctness of the RFR model and adds to the confidence of the prediction result.

5 Word Difficulty: Classification

In this section, we will propose two different models for difficulty classification of the solution words. One of them is based on the expectation of # Trials (EOT), and the other one is developed on the basis of the k-Means clustering (KMeans) [6] model. At the end of this section, we will provide the difficulty level of the word "EERIE".

5.1 The EOT Model

The number of trials # Trials is the most direct way to measure the difficulty of guessing a word. Words that are considered hard usually take more # Trials whereas words that are considered easy

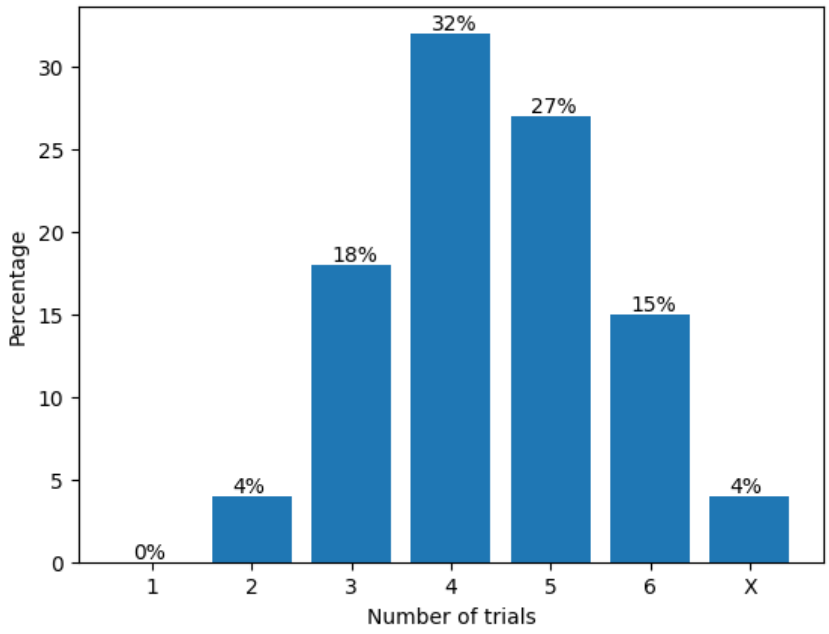


Figure 11: The predicted distribution of # Trials for the word "EERIE" using the RFR model.

take less # Trials. We treat it as a random variable and calculate its expectation as

$$\mathbb{E}[\#\,\textsc{Trials}](w) = \sum \#\,\textsc{Trials} \cdot \text{the corresponding percentage}. \tag{15}$$

Note that sometimes we will abbreviate $\mathbb{E}[\#\,\textsc{Trials}]$ as EOT. We plot the distribution of EOT as in Figure 12 across all words in the data set.

As a convention (like box plots), we choose the first and third quartiles (*i.e.*, the 25th and the 75th percentiles) as the boundaries of classification. Words in the top 25% of EOT are considered hard, those in the bottom 25% are considered easy, and the remaining words are considered of medium difficulty. This would be reasonable because most of words (in this way of classification, 50%) are considered normal, while few words (in this way, 25%) are considered as deviation from normal (either too easy or too hard). By computation, we have that $\mathbb{E}[\#\,\textsc{Trials}] = 3.94$ is the marks the transition from the easy to the medium difficulty, and $\mathbb{E}[\#\,\textsc{Trials}] = 4.42$ marks that from the medium to the hard difficulty. In Figure 12 we also plot these (transitional) boundaries in blue dotted lines.

Initializing the word "EERIE" via its predicted distribution of # Trials as in Figure 11, we have

$$\mathbb{E}[\#\,\textsc{Trials}](\text{"EERIE"}) = 4.43 > 4.42, \tag{16}$$

and thus we consider the word "EERIE" among the top 25% hard ones, but close to the boundary between the medium and the hard.

5.2 The KMeans Model

Different from the EOT model which directly applies the prediction of the distribution of # Trials to a fixed difficulty scale, we also propose the a k-Means-based model that again takes into account all

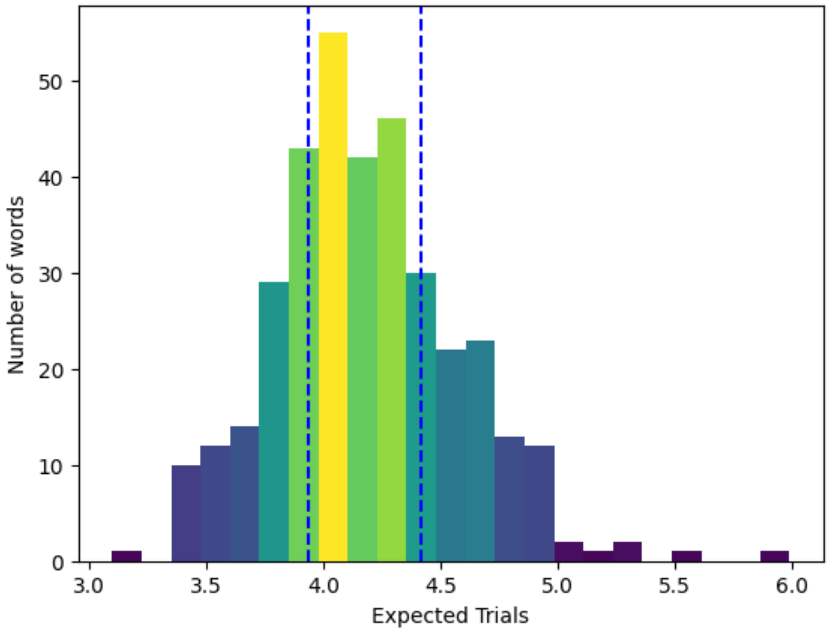


Figure 12: The distribution of $\mathbb{E}[\#\textsc{Trials}]$. Blue dotted lines represent the first and the third quartiles.

the attributes used for the prediction.

In order to apply the k-Means clustering method, we first standardize all the attributes by removing the mean and scaling to unit variance in the data set. As for determining the number of clusters k, we experiment for $k \in [2, 10]$ and compute the Silhouette scores respectively. By the Elbow method, we select $k = 7$, which corresponds to the minimal Silhouette score of approximately 9.84×10^{-2}, and thus implying the most clustering accuracy.

Furthermore, in order to obtain a meaningful visualization of high-dimensional data, we use the Principal Component Analysis (PCA) model [1] for dimension reduction. Through massive experiments, we first reduce the dimension to $\mathbb{R}^7$, then take the first two principal components determined by PCA. The clustering result is visualized as in Figure 13.

Now it suffices to determine the level of difficulty for each clustering class. Same as in the EOT model, we will classify them into easy, medium, and hard. This classification will be determined by analyzing the mean #Trials for each class of words, and we will assign a difficulty level for each two of the clustering classes, with the ascending expectation of #Trials corresponding to higher levels of difficulty. The relation between the distribution of #Trials and different clustering classes is shown as in Figure 14, and the assignment of difficulty levels is as follows:

Class #	6	4	1	0	2	3	5
Difficulty	very easy	easy	medium-easy	medium	medium-hard	hard	very hard

Initializing the word “EERIE” based on its attributes (14) and its predicted distribution of #Trials

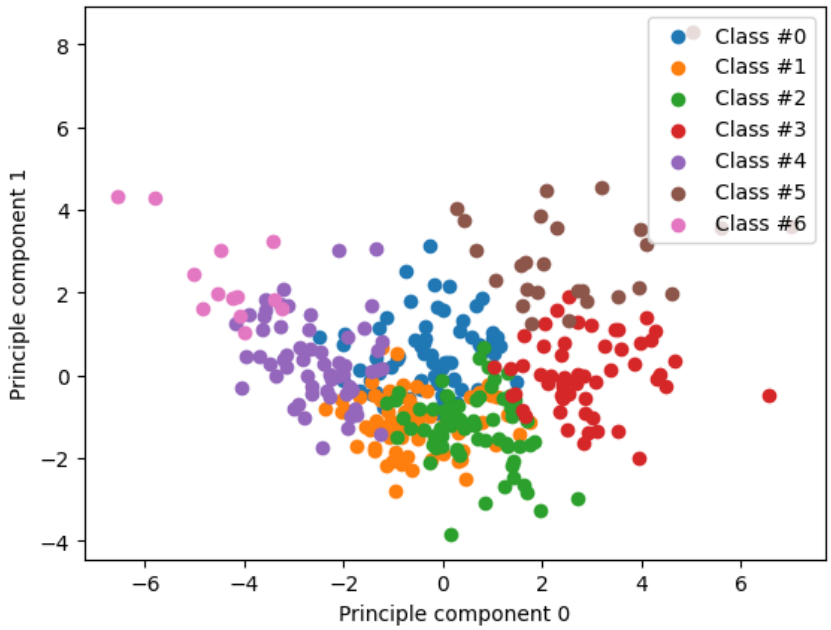


Figure 13: The visualization of the k-Means clustering result using $k = 7$.

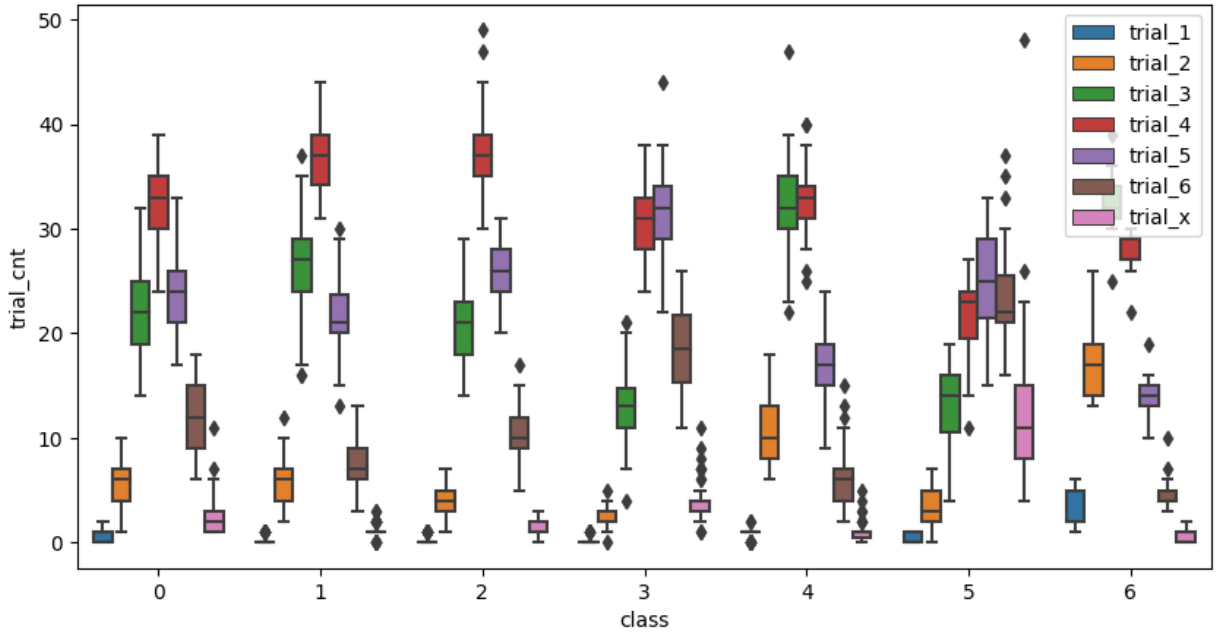


Figure 14: The relation between the distribution of # TRIALS and different clustering classes.

as in Figure 11, we have that

"EERIE" =

# TRIALS = i							ith letter l_i					attributes		
1	2	3	4	5	6	≥ 7	l_1	l_2	l_3	l_4	l_5	VOL	$\mathbb{E}$[YH]	$\mathbb{E}$[GH]
0	4	18	32	27	15	4	4	4	17	8	4	0.9503	1.0974	0.4573

. (17)

Applying the same transformation as for standardizing the original data, our KMeans model categorizes the word "EERIE" as in Class #3, which is of hard difficulty. This is the same conclusion as that of the EOT model. By examining the sizes of the clusters, its predicted expectation of #Trials is approximately at the 78th percentile, again agreeing with the EOT model result that the expectation of #Trials for "EERIE" is a little bit over the third quartile.

6 Other Features: Pattern Discovery

In this final analysis section, we will discover some patterns in a word that may implicitly relate to its difficulty level. Specifically, we are interested in certain combinations of letters in the words. To do this, we first characterize each word as a zero-one vector, marking its occurrences of letters and its difficulty level. For instance, the word "EERIE" is of hard difficulty, and contains the letters "E", "R", and "I", so it will be characterized as

"EERIE" =	difficulty			letters										
	⋯	hard	⋯	"A"	⋯	"D"	"E"	"F"	⋯	"I"	⋯	"R"	⋯	.
	⋯	1	⋯	0	⋯	0	1	0	⋯	1	⋯	1	⋯	

(18)

Then, we apply the Apriori algorithm [2] to find the best association rules among the data set, and select only the rules that involve some combination of letters and a certain level of difficulty. With at least 4% of support, we discover the following:

- The combinations ("A", "E"), ("A", "R"), and ("E", "R") often lead to words of medium level.
- The combination ("E", "T") often leads to words of medium-easy level.
- The combinations ("A", "E"), ("E", "S"), and ("E", "T") often lead to words of easy level.

If we extend the pattern to length 3 and lower the least required support to 2%, we can see that the combination ("A", "E", "R") often leads to words of medium level difficulty. However, there are very few words in the data set that has such a combination of letters, so this conclusion remains for further discussion.

Except from the explicit letters that make up the words, recall the attributes `VOL`, $\mathbb{E}$[`YH`], and $\mathbb{E}$[`GH`], which are the abbreviations (notations) for the Variety of Letters, the Expectation of Yellow Hit, and the Expectation of Green Hit, respectively. We are also interested in their relation with the difficulty of a word, which we will discuss as follows.

`VOL` versus difficulty. The relation between the attribute `VOL` and the word difficulty is shown as in Figure 15. Since the solution words of Wordle consist of only five letters, the possible values of `VOL` are very limited. There is no obvious relation when we consider high `VOL` values, but we can clearly see that low `VOL` values only lead to high difficulty levels (*i.e.*, hard and very hard). This is reasonable since players do not tend to guess duplicate letters, resulting in more trials to find out the correct answer.

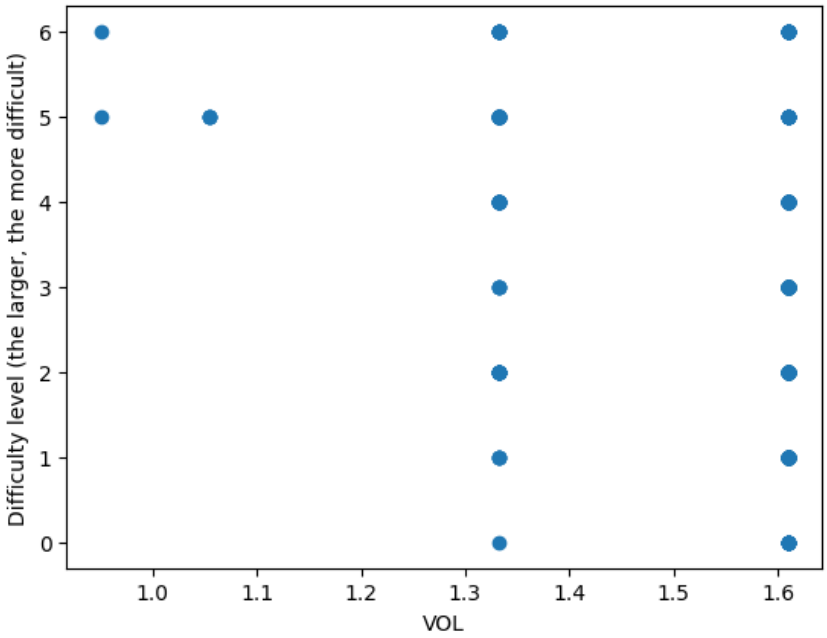


Figure 15: The relation between VOL and word difficulty.

$\mathbb{E}$[YH] and $\mathbb{E}$[GH] versus difficulty. The relation between the attributes $\mathbb{E}$[YH] and $\mathbb{E}$[GH] and the word difficulty is shown as in Figure 16. From the plot on the left-hand side of Figure 16 We can see that high $\mathbb{E}$[YH] values generally do not lead to high difficulty levels, and low $\mathbb{E}$[YH] values tend to imply that the word is at least of the medium-hard level of difficulty. The pattern of $\mathbb{E}$[GH] is similar, except that there are few samples with low $\mathbb{E}$[GH] values that can lead to a meaningful conclusion.

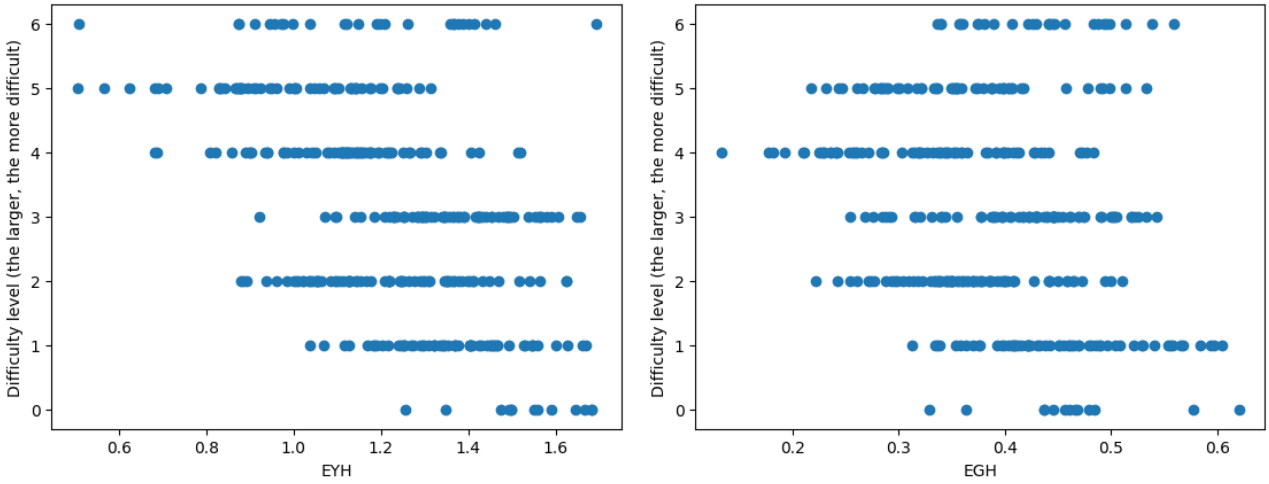


Figure 16: The relation between $\mathbb{E}$[YH] (left), respectively $\mathbb{E}$[GH] (right), and word difficulty.

7 Conclusion

We have proposed and trained various models for forecasting, prediction, and classification on the given data set of Wordle reports. We have also forecast # RRT on March 1st, 2023, predicted the distribution of # TRIALS for the solution word "EERIE", and classified its difficulty level, as required. For the rest of this section, we will discuss the advantages and disadvantages of our models, and propose some possible future improvements.

Advantages. Generally, our models have the following advantages:

- In each part, we have two mutually corroborating models which can cross validate the result of each other, thus adding to the credibility of the models in addition to checking the training errors.
- We do not only consider the superficial attributes such as the positions and occurrences of different letters, but also propose new attributes that may affect the distribution of # TRIALS and the classification of difficulty according to our own experience playing Wordle. The correlation analysis also demonstrates the feasibility of some of these attributes.

Disadvantages and future improvements. However, there are several disadvantages in our models:

- In Section 3, is it really valid to ignore the peak in the first few months after the publication of Wordle? Is the ARIMA model the best in this case, given it only provides stationary forecasts and mostly applies only to the near future? Our models do not take into account these questions. We may try to consider more time series models such as the Prophet model, more types of curves such as the negative exponential curves, and take into account the peak in the data set during our analysis.
- In Section 4, there must be more hidden attributes of the words yet to be discovered. Furthermore, the Fully Connected Neural Network model may give a better prediction, but we are not familiar with that.

At last, we would like to thank you for your meticulous reading.

February 21, 2023

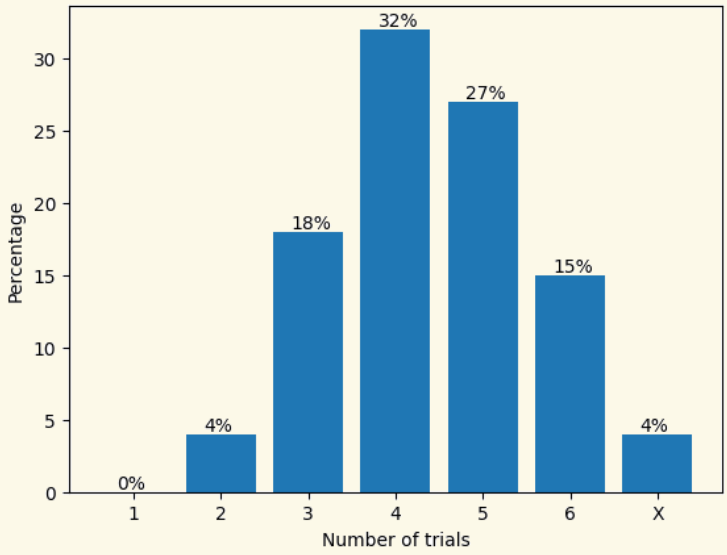
Dear Editor Bennett,

Greetings! We are three Math enthusiasts invited by the Mathematical Contest in Modeling (MCM) to analyze a file you provided to us, containing the data collected from a worldwide puzzle game you edit, Wordle. We have built up vigorous mathematical models and managed to elaborate on a proper answer to your questions. We have confidence in saying that what we have done is not only refined and logical but also intriguing and inspiring.

The first main question you assigned is to develop a model expounding the daily variation of the total number of reported results, and further predict an interval for a target date, March 1, 2023. Therefore, we conduct a time series analysis and a nonlinear regression to make a reasonable prediction. **Our prediction interval of the total number of reports on March 31, 2023 is [15000, 18000].** For your information, we provide a braver guess of the number based on our vigorous computation, which is **around 16760**. This is nothing but a bold bet.

In addition to our interval prediction, we evaluated the correlations of two attributes of words with the proportion of results reported in Hard Mode, which are the letter occurrences and the letter frequency. We find that the words involving the letters "J", "Z", or "X" have significant positive correlations with the proportion of Hard Mode players. Moreover, if a word contains letters that occur rarely, that proportion would be significantly higher than normal.

The second main question you asked is to deal with a prediction of the distribution of the reported results. To solve this problem, we proposed several novel word attributes and performed a correlation analysis to select some potential influencing factors of that distribution. Training a Decision Tree and a Random Forest, we obtained a robust prediction model. We then input "EERIE" and **the distribution prediction is shown below.**



However, nothing is perfectly still. The uncertainties include whether there will be a future trend to use word searchers in playing the Wordle game, whether there exist hardly estimated psychological factors people may have, and whether there are more word attributes that could largely influence the outcome of our prediction model (we consider the last one as the most uncertain factor), still exist due to the constraint of time and data limits. Nevertheless, we have confidence in our prediction since we use two backup models to form a dual control of the prediction, and the outcome of these two models are very close and similar to each other, we are convinced that this method keeps the accuracy of our prediction in a relatively lower level.

The third main question you inquire about is to classify Wordle's solution words by difficulty. To tackle this problem, we set up two mutually corroborating models based on the previously predicted distribution of the number of trials. One of them is based on the expectation of the number of trials and the other is based on clustering. Both models suggested the same category for "EERIE", which is the hard realm near the hard-medium boundary. More specifically, **"EERIE" is on the boundary of the top 25% most difficult words to guess.** The two approaches cross validate each other and thereby add to the robustness of our difficulty evaluation system.

Last but not least, here we list some other interesting features we find in the data set you provided. The combinations ("A", "E"), ("A", "R"), and ("E", "R") often lead to words of medium level, ("E", "T") often leads to words of medium-easy level, while ("A", "E"), ("E", "S"), and ("E", "T") often lead to words of easy level. Moreover, VOL is positively correlated to the difficulty level, while $\mathbb{E}$[YH] and $\mathbb{E}$[GH] are negatively correlated to it.

Finally, we would appreciate it very much if you could consider our results. Hope our analyses could help you better understand what is going on under the surface of the simplicity of Wordle.

Sincerely,

Team 2316597

References

[1] Hervé Abdi and Lynne J. Williams. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, July 2010.

[2] Mohammed Al-Maolegi and Bassam Arkok. An Improved Apriori Algorithm for Association Rules. 2014.

[3] Dimitris Bertsimas and Alex Paskov. An Exact and Interpretable Solution to Wordle. 2022.

[4] Gérard Biau and Erwan Scornet. A Random Forest Guided Tour. *TEST*, 25(2):197–227, June 2016.

[5] S.L. Ho and M. Xie. The Use of ARIMA Models for Reliability Forecasting and Analysis. *Computers & Industrial Engineering*, 35(1-2):213–216, October 1998.

[6] Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. The Global K-Means Clustering Algorithm. *Pattern Recognition*, 36(2):451–461, February 2003.

[7] Anthony J. Myles, Robert N. Feudale, Yang Liu, Nathaniel A. Woody, and Steven D. Brown. An Introduction to Decision Tree Modeling. *Journal of Chemometrics*, 18(6):275–285, June 2004.

[8] Patrick Schober, Christa Boer, and Lothar A. Schwarte. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768, May 2018.

[9] Owen Yin. Here Lies Wordle: 2021–2027, February 2023.