Problem Chosen	2023	Team Control Number
C	MCM/ICM Summary Sheet	2322596

Predicting Wordle Results

Wordle has become an extremely popular game on social media, resulting in a wealth of user-generated data that can be utilized to develop predictive models and gain insights into various factors that affect the game's score distribution, reported results, and difficulty of word solutions.

In the first part of our study, we aimed to expand the dataset and investigate the relationship between time, popularity, and the number of reported results in Wordle. We constructed four regression models to explore the factors influencing the number of reported results in Wordle over time. Our analysis revealed that the exponential regression model with time as a predictor variable was the most accurate. We expressed the revised exponential decay equation as $y(t) = 292004 \cdot e^{-0.01724t} + 18000$, where t represents the number of days since the initial observation, and C accounts for persistent participants. By substituting $t = 322$ into the equation, we predicted the number of reports on March 1st, 2023, to be 19139.

In the second part of our study, we introduced a predictive model that simulates the Wordle game in hard mode. We developed a simulator that mimics human-like Wordle players, utilizing the Monte-Carlo method to ensure accuracy and consistency. We achieved the best results by combining a random word generated from the player's vocabulary bank with a list of optimal words, resulting in an RMSE of 0.321 and Pearson coefficient of 0.54824. We categorized solution words by difficulty, determining the attributes associated with each classification. After running the model through 5000 simulations, we concluded that the word "EERIE" is relatively hard, and we predicted a score distribution of (0.0, 0.2, 6.1, 31.2, 43.8, 16.0, 7.0) for March 1st, 2023. Our study also revealed that challenging words typically contain duplicate and rare letters, and a word's difficulty level as its atrribute does not influence the number of submissions in hard mode.

Key Words: Regression; Machine Learning; Monte Carlo simulation

Contents

1 Introduction 1

2 Data Analysis 2

2.1 Dataset Exploration . . . 2

2.2 Data Engineering . . . 3

3 Prediction of Wordle Results 4

3.1 Dataset Expansion . . . 4

3.2 Assumption . . . 4

3.3 Models . . . 5

3.4 Results . . . 6

3.5 Strengths and Weaknesses . . . 8

3.5.1 Strengths . . . 8

3.5.2 Weaknesses . . . 9

3.6 Consideration and Future Work . . . 9

4 Regression Analysis for Reported Results Prediction 10

4.1 Dataset Expansion . . . 10

4.2 Assumption . . . 10

4.3 Models . . . 11

4.3.1 Polynomial Regression . . . 11

4.3.2 Exponential Regression . . . 13

4.3.3 Metrics . . . 14

4.4 Results . . . 14

4.5 Strengths and Weaknesses . . . 16

4.5.1 Strengths . . . 16

4.5.2 Weaknesses . . . 16

4.6 Consideration and Future Work . . . 16

5 Conclusion 17

1 Introduction

The COVID-19 pandemic has resulted in an unprecedented shift towards online entertainment, including the development of new online games. One such game is **Wordle**, an online word-guessing game that has garnered immense popularity since its inception in 2021. The game's meteoric rise in popularity can be attributed to its user-friendly interface, uncomplicated rules, and a sharing feature that allows players to quickly and effortlessly share their scores with friends and family. Within weeks of its release, Wordle became a global sensation, with millions of players worldwide logging in daily to attempt to guess the day's five-letter word.

Wordle challenges players to solve a new puzzle each day within six attempts. As the player inputs their guesses, the letters are color-coded to indicate their accuracy: gray indicates that the letter is not present in the word, yellow means the letter is present but in the wrong position, and green signifies that the letter is both present and correctly positioned. The goal of the game is to guess all five green letters and identify the correct word. With its addictive nature and increasing popularity, Wordle has gained widespread attention on social media platforms such as Twitter, where users enthusiastically share their scores and engage in discussions about the game. Players can play in regular mode or "Hard Mode." Wordle's Hard Mode makes the game more difficult by requiring that once a player has found a correct letter in a word (the tile is yellow or green), those letters must be used in subsequent guesses.

This study aims to analyze a dataset consisting of the results of the Wordle game from January 7, 2022, to December 31, 2022. Through the analysis of this dataset, our study aims to develop models that can achieve the following research goals:

- explain the variation in the number of reported results on Wordle each day, and create a prediction interval for the number of reported results on March 1, 2023
- investigate whether any attributes of the Wordle solution word affect the percentage of scores reported that were played in Hard Mode, and explain the findings in our model
- predict the distribution of the reported results, expressed as associated percentages of (1, 2, 3, 4, 5, 6, X) for a given future solution word on a future date, and identify uncertainties associated with the model and its predictions
- classify solution words by difficulty, and identify the attributes of a given word that are associated with each classification
- assess the difficulty of the word, and evaluate the accuracy of our classification model.

2 Data Analysis

2.1 Dataset Exploration

Our study utilized a dataset provided by the MCM, which was sourced from Twitter and comprised daily results spanning from January 7, 2022, to December 31, 2023 [1]. This dataset comprised various parameters, including the date, contest number, the word of the day, the count of individuals reporting scores on a given day, the number of players participating in hard mode, and the percentage of individuals who successfully guessed the word within one to six attempts, as well as those who were unable to solve the puzzle (indicated by the letter ’X’). Through an analysis of the correlation between the date and the number of reports, we observed a marked increase in the number of reported results (both in normal mode and hard mode) from January to March 2022, followed by a steady decline, which can be visualized through Figure 1 and Figure 2.

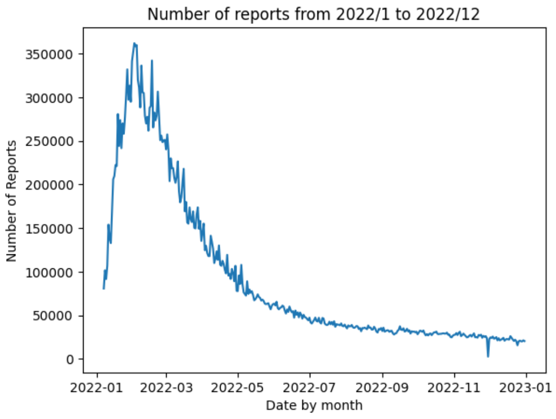

Figure 1: Reports Variation vs Time

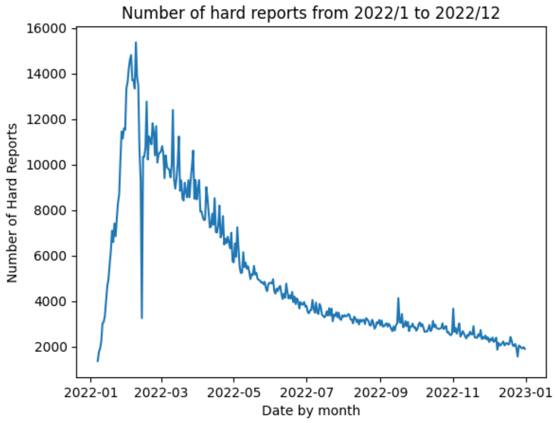

Figure 2: Hard Reports variation vs Time

We hypothesized that the dramatic surge in reports during this period was due to the fact that the game was first launched in October 2021 and later acquired by The New York Times Company in January 2022 [2]. The promotional efforts undertaken by The New York Times by including the game in their app, in conjunction with the ease of sharing score on Twitter which leads to increase visibility and popularity of the game, could plausibly account for the observed increase in the number of reports.

Moreover, we discovered from our analysis that the majority of players take around 3 or 4 attempts to complete a Wordle game, which is illustrated in Figure 3.

Upon further analysis of the dataset, we have identified several errors. Specifically, we found that the dataset contains two four-letter words, "clen" and "tash," which are not recognized in the English dictionary. Additionally, we also found one word that contains a non-English alphabet character, "naïve."

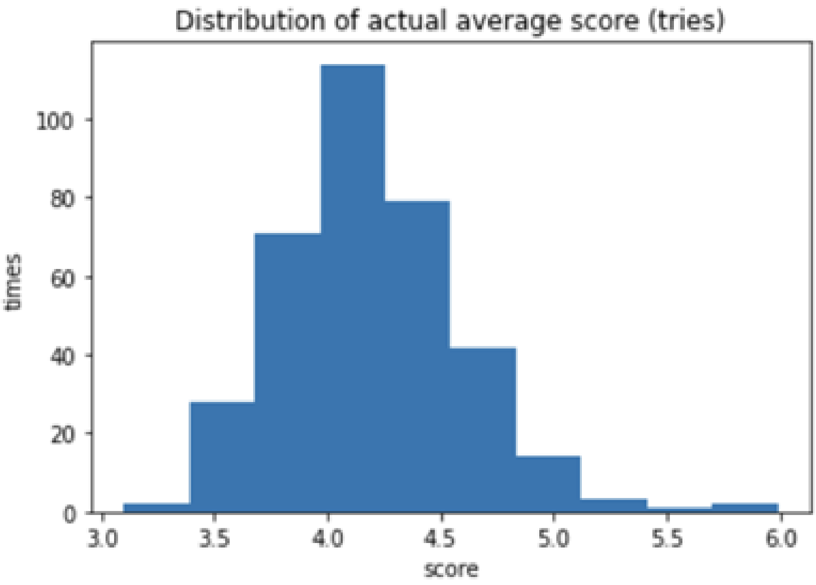

Figure 3: Score Distribution

2.2 Data Engineering

As alluded to in the preceding section, our examination of the dataset revealed two issues related to incorrect formatting. Specifically, we found instances of words that deviated from the required 5-letter length, as well as instances where non-English letters, such as "ï", were present in the words. To address the issue of words that were not 5 letters in length, since we could not ascertain the actual word, we opted to remove the relevant rows from the dataset. With respect to the problem of non-English characters, we surmised that it was likely an artifact of text detection, and therefore replaced the offending characters with their English equivalents, such as replacing "ï" with "i".

Our analysis revealed a marked increase in the number of reports in the Wordle game during the period from January to March 2022. We corroborated this observation by referring to data from Google Trend, which indicated that the popularity index had reached 100 by February 13, 2022 [3]. However, our initial attempts at developing regression models proved challenging due to the presence of peaks in the data. The regression model tended to overfit the data, learning to predict based on these rare occurrences rather than the underlying patterns. To mitigate this issue, we chose to preprocess the data by removing the anomalous patterns, specifically, by discarding the first 35 rows of the dataset (corresponding to the period from January 7 to February 13). Following this cutoff, the updated data exhibited a clearer and more consistent pattern (Figure 4) that was better suited to fitting the regression model. Moving forward, we analyzed our dataset with our first observation as the one on February 14, 2022.

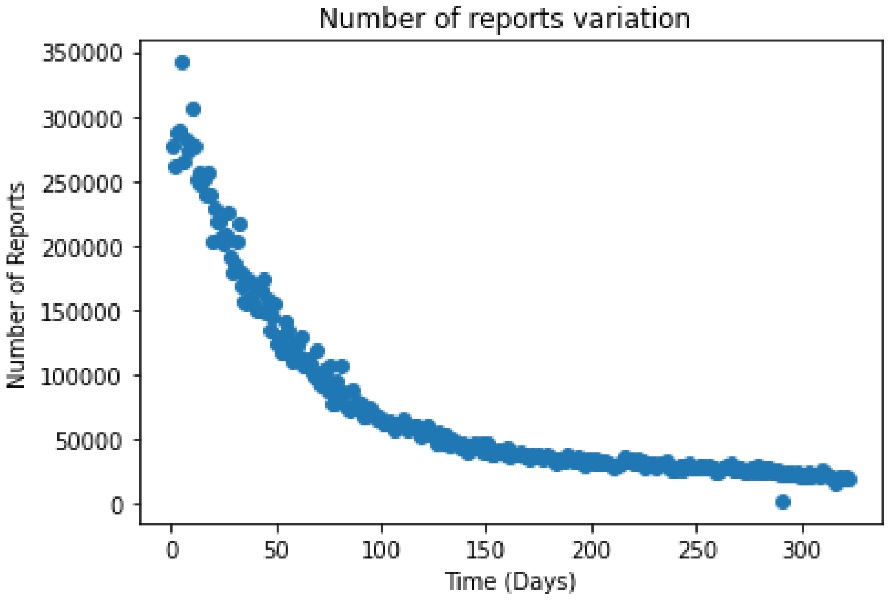

Figure 4: Reports variation after preprocessing

3 Prediction of Wordle Results

In the first part of our study, we focus on our goal to develop a model that predicts the distribution of reported results for a given future word and date, to classify solution words by difficulty, identify associated attributes, and assess the difficulty of specific words.

3.1 Dataset Expansion

To build our predictive model, we decided to enrich our dataset by acquiring a dictionary of the English language with associated frequencies from Kaggle [4]. The dataset contains the frequency count of the 333,333 most frequently used single words on the English language web. To focus on five-letter words and their respective frequencies, we filtered the original dataset accordingly. Our final dictionary consists of 8092 number of five-letter words with their corresponding frequencies.

3.2 Assumption

We present a model that simulates the Wordle game in hard mode, assuming the player is a competent native English speaker. We start with the assumption that the player has a limited vocabulary word bank, comprising words that are expected to be familiar to them, with around 90% of the words being known to them based on our English dictionary, while the remaining 10% are unknown. Our model also incorporates the idea that the more frequently a word is used in the English language, the more likely the player is to be familiar with it.

We further assume that during gameplay, the player will predominantly choose words that they are familiar with, particularly those that are commonly used and have been shown to have a high success rate. In addition, we assume that after each guess, the player will use the color-coded hints provided to make an informed next guess. In the event that the player exhausts their vocabulary bank, we assume that they will use all the given constraints and resort to randomly filling in the blanks.

3.3 Models

We introduced a simulator that emulates human-like Wordle players. To ensure the reliability and consistency of our simulation results, we adopted the Monte-Carlo method. This approach is particularly useful for simulations that involve randomness and statistical factors, as it enables us to run the model multiple times, generating new random inputs for each iteration, and thus providing a more robust analysis.

The first step in developing the simulator is to generate a vocabulary bank for the player, based on our assumption that the player is more likely to be familiar with more frequently used words. To achieve this, we obtain the frequency percentage of each word, and then use this information to determine the number of words that the player is expected to know. Specifically, we use a variable 'v' to represent the percentage of possible words that are likely to be in the player's vocabulary, and then randomly generate the vocabulary bank with a weighting system based on the frequency percentage of all possible words.

`Define variable`:

- **vocab**: The original vocabulary bank of the player.
- **percentage**: An array containing the percentage associated with each word remaining in the vocabulary bank.
- **filtered_vocab**: The vocabulary bank left after filtering through new constraints.
- **constraints**: Contains all the constraints up to the current point in the game.
- **filtered_not_in_vocab**: A list of words not present in the player's vocabulary bank, which have been filtered by the current set of constraints.

Our Wordle simulator follows the steps below to simulate a game of Wordle:

1. At the start of the game, the simulator randomly selects a five-letter target word from the set of possible words that the player has knowledge of.
2. In order to increase the chances of winning, experienced Wordle players tend to choose specific words for their first guess. We identified a list of such words through a recent research study, which includes “salet”, “reast”, “trace”, “crate”, “slate” [5]. Therefore, our Wordle simulator selects the first guess by randomly choosing from this list of "good" words as well as a word from the player's vocabulary bank. The selection of the latter word is weighted by the frequency percentage of each word left in the filtered vocabulary bank, if applicable. This approach

ensures that more common words are selected more often, adding to the challenge of the game.

```
choice <- vocab.sample(n=1, weights=percentage)
guess <- random.choice([’salet’,’reast’,’trace’,’crate’,’slate’, choice])
```

3. The simulator updates the color-coded constraints (hints) based on the player’s guess.

```
constraints <- update_constraints(constraints)
```

4. The simulator filters the player’s vocabulary bank to fit the updated constraints.

```
fltered_vocab <- filter(vocab, constraints)
```

5. If the player still has words left in their vocabulary bank, the simulator calculates the frequency percentage of each word left in the filtered vocabulary bank, based on the data you obtained on the frequency of all possible target words.

6. The player can guess at random, but with the selection weighted by the frequency percentage of each word left in the filtered vocabulary bank (if applicable). This means that more common words will be selected more often, making the game more challenging.

```
Guess <- filtered_vocab.sample(n=1, weights=percentage)
```

7. If the player runs out of words in their limited vocabulary bank before guessing the target word correctly, the player will randomly select a word from the set of possible words that he doesn’t already know, but which fits the current constraints (hints). (like putting random letter in the spaces)

```
guess <- filtered_not_in_vocab.sample(n=1)
```

This would run until the player guesses the target word correctly and return the number of turns it takes the player to guess the target.

3.4 Results

In the examination of the initial guess for the player, multiple options were considered. These encompassed the implementation of a stochastic term drawn from the player’s vocabulary bank with weighted proportions, the selection of words from a list of lexemes deemed optimal by academics, and a combination of the two techniques.

Upon careful scrutiny of the outcomes, it was discovered that the combination of both approaches yielded the highest level of correlation between the generated scores and the actual scores reported on Twitter.

To evaluate the performance of our best model, we performed the simulation 5000 times and compared the mean score of the simulation with the mean score of the available empirical data. Our model achieved a Root Mean Square Error (RMSE) value of

0.321 and a Pearson correlation coefficient of 0.54824, signifying its dependability in forecasting forthcoming outcomes. Since our model is not time-sensitive and still manages to produce a positive correlation, we can deduce that time does not affect the results. Hence, we can predict the percentage distribution of a future instance based on the input word without the need to incorporate time.

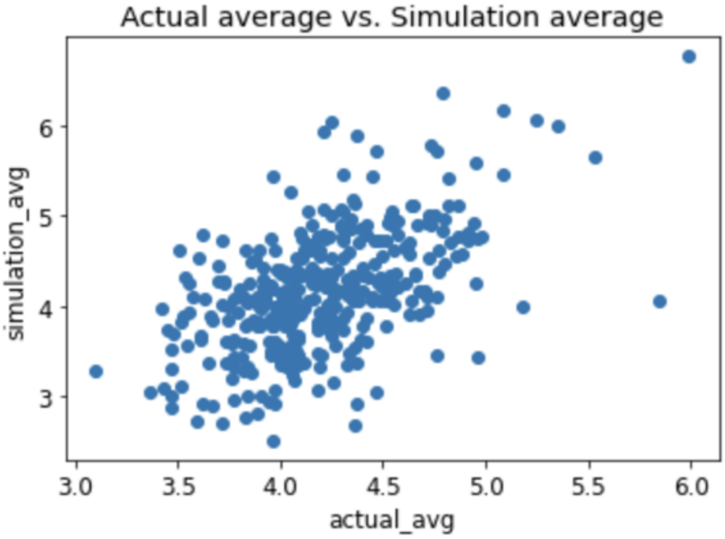

Figure 5: Actual Average vs Simulation Average

In order to predict the future distribution, we first need to classify words by difficulty. We obtained the average score from the actual data for each day by calculating the expected value of the distribution of score

$$E = \sum_{i=1}^{7} dist[x] \cdot x$$

To classify solution words by difficulty, we used the average score as the criterion and identified the attributes of a given word that are associated with each classification. We presented our classification of the difficulty of a word solution in Table 1. Using our model, we then ran our model through 5000 simulations to determine the distribution of the reported results of the word "EERIE" and difficulty of the word "EERIE". we got a distribution of the results in figure, so we predict that the distribution of the score for "ERRIE" will be (0.0, 0.2, 6.1, 31.2, 43.8, 16.0, 7.0) on March 1st, 2023. In addition, we calculated an average score of 4.7. We can say that this word is relatively hard.

Table 1: Word Difficulty Classification

Score	**Category**
≤ 3.5	Easy (accounted for 6.3%)
$3.5 \leq avg \leq 4.5$	Medium (accounted for 78.7%)
$4.5 \leq avg$	Hard (accounted for 15%)

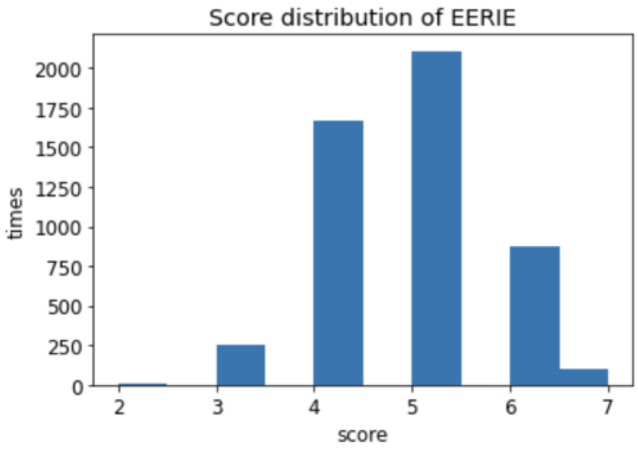

Figure 6: Score Distribution for EERIE

Given the results from the simulators and actual data, we see that most difficult/ extreme words contain duplicate letters and abnormal letters.We can observe that the difficulty (quantified by the average score) does not affect the number of submissions in the hard mode.

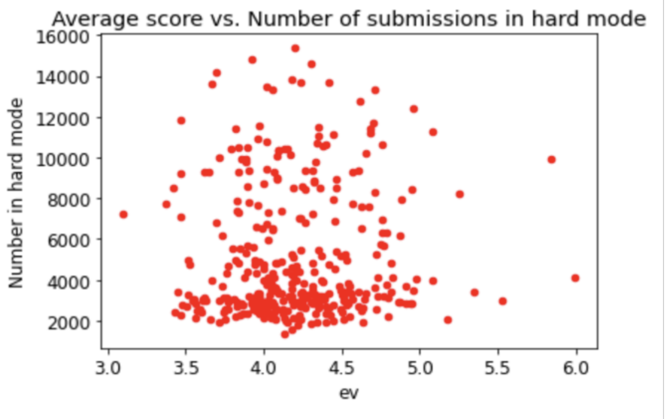

Figure 7: Average Score vs Number of Submissions in Hard Mode

3.5 Strengths and Weaknesses

3.5.1 Strengths

The simulator developed to predict the average score of players is notable for its simplicity, transparency, and ease of implementation. Based on a logical and straightforward set of rules, the model is less prone to overfitting to noisy or biased data, providing a clear understanding of the underlying factors influencing the score distribution.

The positive correlation observed in the model allows for the accurate prediction of solution word difficulty from real players' perspectives, as well as the distribution of scores. The model also provides valuable insights into how regular players engage with the Wordle game, which may not be discernible using more complex prediction models. As such, the simulator has the potential to inform future game design and development decisions, as well as enhance our understanding of player behavior in other word games.

3.5.2 Weaknesses

The simulation conducted in this study exhibits room for improvement, specifically in the area of running time, which currently poses a hindrance to conducting a large number of simulations and obtaining a more comprehensive understanding of player behavior.

Additionally, it is important to acknowledge that the simulation framework employed may not fully encompass the complexity and diversity of player behavior. Given that players have unique strategies and approaches for solving the game, any modeling approach has inherent limitations. Our model, in particular, may struggle to account for the subtleties of player behavior that are not represented by our current set of features or the simulation framework itself.

3.6 Consideration and Future Work

There are several considerations and potential avenues for future work related to the analysis of the Wordle game dataset. One key consideration is the need for careful evaluation and interpretation of the results to ensure their validity and reliability. This may involve using additional techniques, such as cross-validation or bootstrapping, to assess the robustness and generalizability of the models.

For future work, it may be beneficial to address the identified weaknesses to improve the accuracy and generalizability of the predictions. First, the current running time of the simulation may limit the number of iterations and insights into player behavior that can be gained. Optimizing the simulator program and utilizing more efficient computational techniques can enable more comprehensive and faster simulations. Second, to address limitations in the model's ability to capture all player behaviors, further research may involve incorporating a wider range of datasets, additional features, or modeling techniques. This could include capturing more nuanced and complex aspects of player behavior, such as individual differences in problem-solving strategies or personal preferences for certain word types or game mechanics. Additionally, applying the simulation framework to other similar word games to compare and contrast player behavior across different game types and mechanics may provide insights into the factors that influence player engagement and performance in word games more broadly, as well as inform the development of more effective predictive models for these types of games.

4 Regression Analysis for Reported Results Prediction

4.1 Dataset Expansion

To develop predictive models for future reported results in the Wordle game and to gain insights into the factors that influence the number of reported results over time, we sought to expand our dataset. We posited that the decrease in popularity through time was the reason for the decrease in number of reports. Accordingly, we introduced a new column into the original dataset to include time as a predictor variable, assigning a timestamp to each observation based on the number of days elapsed since the first recorded observation on February 14, 2022. For instance, an observation on February 15 , 2022, would be assigned a timestamp of 1 (day), while an observation on February 16 would have a timestamp of 2 (days), and so on.

The popularity of the Wordle game is a crucial factor that can impact the number of reported results. To quantify popularity, we leveraged Google Trends to explore the popularity index of the search term "Wordle" during the period covered in the dataset in Figure 8. The popularity index serves as a proxy for public interest in the game, and we obtained the data in CSV format and added it as a new column in the dataset, serving as the second predictor variable.

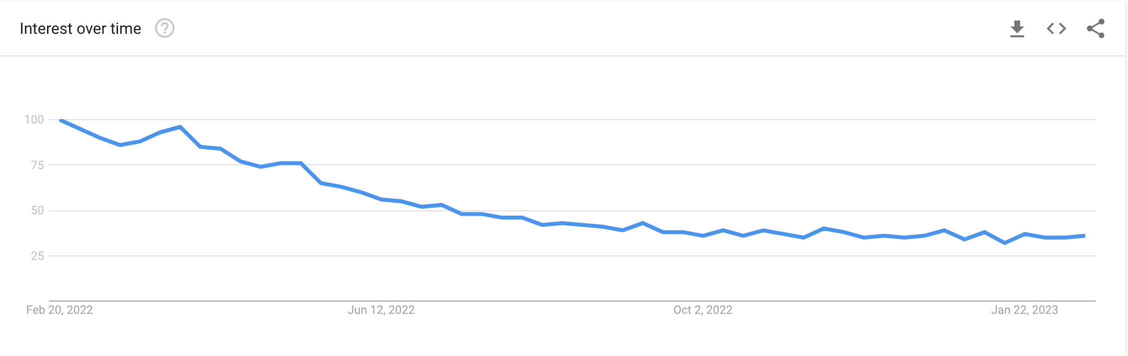

Figure 8: Google Trends Popularity Index

4.2 Assumption

In our analysis, we make several assumptions regarding the data and its relationship to the popularity of the Wordle game. Firstly, we assume that Google Trends provides a reliable index of the game's popularity. Secondly, we assume that both relationship between the popularity of the game and the number of reported results and relationship between time and the reported results are non-linear. Finally, we assume that the data will follow a certain trend, which we will analyze using statistical methods to gain insights into the popularity and usage patterns of the game over time. These assumptions will guide our analysis and interpretation of the data, and will help us to draw meaningful conclusions about the Wordle game and its impact.

4.3 Models

The first primary objective of this study is to investigate the factors that affect the number of reported results in Wordle over time and to develop predictive models for future reported results. We developed four models to accomplish this goal, with time and popularity identified as the primary predictors. Due to the high correlation we observed between time and popularity in our dataset, we opted to use only one feature at a time for our model. Although high correlation does not necessarily degrade the model's prediction performance, it can result in collinearity and affect the estimated coefficients, leading to an unreliable interpretation of feature significance. Our heatmap analysis revealed strong correlations between time and popularity as observed in Figure 9 . Therefore, we believe that choosing only one feature would be a suitable approach to address the concerns of collinearity and obtain more reliable interpretations of feature significance.

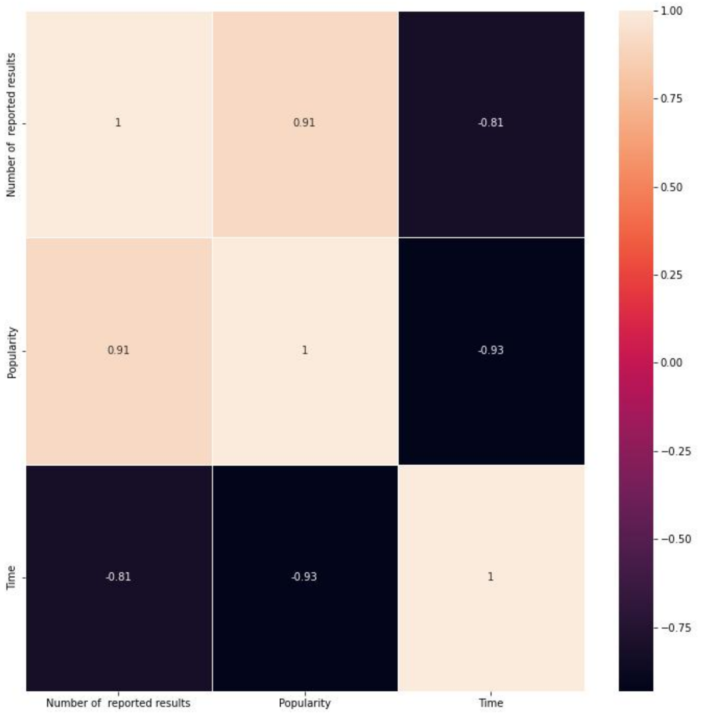

Figure 9: Heatmap of Poppularity and Time

4.3.1 Polynomial Regression

We selected polynomial regression as our modeling approach because it is well-suited for capturing non-linear relationships between input and output variables. After analyzing Figure 4 and Figure 10, it became evident that the relationships between popularity vs the number of expected outputs and time vs the number of expected outputs were unlikely to be linear.

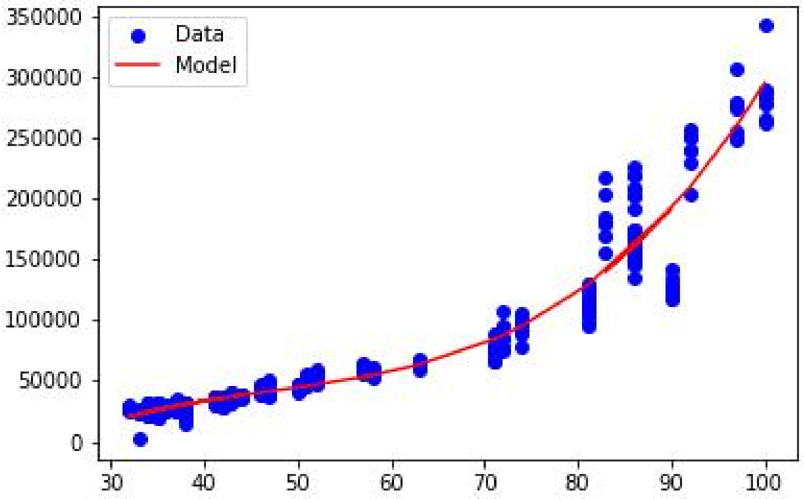

Figure 10: Reports Variation vs Popularity

We constructed two separate models: one using popularity as a feature and the other using time as a feature. We developed two hypothesis functions using polynomial regression: one using the popularity feature and the other using the time feature. Our popularity hypothesis function can be expressed as:

$$f(p) = c_0 + c_1 p + c_2 p^2 + \cdots + c_n p^n$$

where c represents the weights and p represents popularity. Our time hypothesis function can be expressed as:

$$f(p) = c_0 + c_1 t + c_2 t^2 + \cdots + c_n t^n$$

where t represents time.

To estimate the optimal values of the coefficients that minimize the sum of the squared differences between the predicted values and the actual values, we trained our models using the least squares method. The coefficients of the polynomial regression model are denoted by $c_0, c_1, \ldots, c_n$.

To obtain the optimal values of the coefficients and to prevent overfitting, we used Scikit-Learn regression algorithms that allow for parameter tuning. We divided our data into a training set and a testing set in an 80:20 ratio. The training set was used to model the relationship between the target and predictor variables, and the testing set was used to evaluate the performance of the model on new, unseen data. This prevents the machine learning algorithm from fitting the noise in the training data and helps to ensure that the model generalizes well to new data. The least squares method was used during the training phase to estimate the optimal values of the coefficients that minimize the sum of the squared differences between the predicted values and the actual values.

4.3.2 Exponential Regression

We considered an exponential regression as a viable alternative because it is frequently utilized to represent cases in which a variable increases and decreases over time, as is the scenario with our dataset. With popularity, we see exponential growth. With time, we see an exponential decay. Therefore, we think modeling both exponential growth for popularity and exponential decay for time are two good models. This model is especially effective when we expect the rate of decrease to slow down over time.

The exponential decay model describes a process in which the rate of decrease of a quantity is proportional to the current value of the quantity. In our case, we express it as follows:

$$y = a \cdot e^{-kt} + C$$

where:

y represents the number of reported results

a represents the initial value

k is the decay constant, which determines the rate at which the quantity decreases

t represents time

C represents the constant, which denotes players who consistently report irrespective of external factors

The exponential growth model describes a process in which a variable increases exponentially with time. In our case, we express it as follows:

$$y = a \cdot e^{kp} + C$$

where:

y represents the number of reported results

a represents the initial value

k is the decay constant, which determines the rate at which the quantity increases

p represents popularity

C represents the constant

We used the `scipy.optimize.curve_fit()` function to fit the exponential curve to our dataset. This function estimates the optimal values for the model parameters (in this case, a and k) that best describe the data. The optimization process used by `curve_fit()` is based on the least squares method. It finds the optimal values of a and k by minimizing the sum of the squared differences between the model and the data at each time point. This is done iteratively using a nonlinear optimization algorithm.

4.3.3 Metrics

After performing the experiments and hyperparameter tuning on the regression models discussed in the previous section, we compared their performance both on the red wine and white wine datasets. We evaluate the performance of our models using 3 metrics, which are R^2, MSE, and MAPE because these metrics provide a comprehensive evaluation of the model's accuracy and precision in capturing the relationship between the target and predictor variables.

- R^2:
  R^2 measures the goodness of fit of the model and provides insight into how well the model explains the variability in the target variable.
- Mean Squared Error (MSE):
  MSE is used to measure the average squared error between the predicted values and the actual values, which helps to quantify the overall accuracy of the model.
- Mean Absolute Percentage Error (MAPE):
  MAPE measures the average absolute percentage difference between the predicted values and the actual values, providing a more interpretable measure of the model's performance, particularly in cases where there are large differences in scale between the target and predictor variables.

By using these three metrics, we had a well-rounded evaluation of the model's performance, which helped us to fine-tune and improve our models.

4.4 Results

Table 2: Model Performance Metrics

Model	R^2	MSE	MAPE
Time Polynomial	0.98	90329214.611	0.156
Popularity Polynomial	0.938	297825619.611	0.153
Time Exponential Decay	0.983	78220780.244	0.144
Popularity Exponential Growth	0.939	324229576.909	0.1725

Our models demonstrate a strong correlation between time and the number of reported results, as well as popularity and the number of reported results. However, there remains some variability in the results, which may be attributed to the gradual decline in the popularity of the Wordle game over time. As such, it is essential to consider the context of the data when interpreting the models' results and conclusions. It may be valuable to conduct further analyses to investigate the underlying reasons for the game's decline in popularity and explore potential solutions to revitalize its popularity.

Following the completion of our experiments and hyperparameter tuning on the regression models outlined in the previous section, we proceeded to compare their respective performances on our dataset using our results in Table 2. We determined the time exponential decay model to be the optimal choice for analyzing the correlation between the number of reports and the number of days following the first observation date because it has significantly lower MSE than the other three models and provides a good R^2 score and MAPE. The conventional equation of exponential decay, $y = a \cdot e^{-kt}$, was initially employed, but the predictions derived from it exhibited considerable deviations from the actual values. A comprehensive analysis revealed that certain individuals regularly participated in the Wordle game and reported their results on Twitter, regardless of external influences, leading to inaccurate predictions. To address this issue, a constant term denoted as C was incorporated into the model, with its value determined from the number of reports generated during the final week of the dataset, resulting in an estimated value of 18000.

Subsequently, we conducted rigorous tuning of the model by incorporating the constant term and obtained the coefficients $a = 292004.6476$ and $k = 0.01724$. The revised exponential decay equation may be expressed as follows:

$$y(t) = 292004 \cdot e^{-0.01724t} + 18000$$

This equation demonstrates the relationship between the number of reports and the number of days following the first observation date, with a and k representing the initial value and the decay constant, respectively, while C accounts for the persistent participants. We provide a visualization in Figure 11. This figure can help to illustrate how the number of reports changes over time, taking into account the various factors represented by the parameters in the equation.

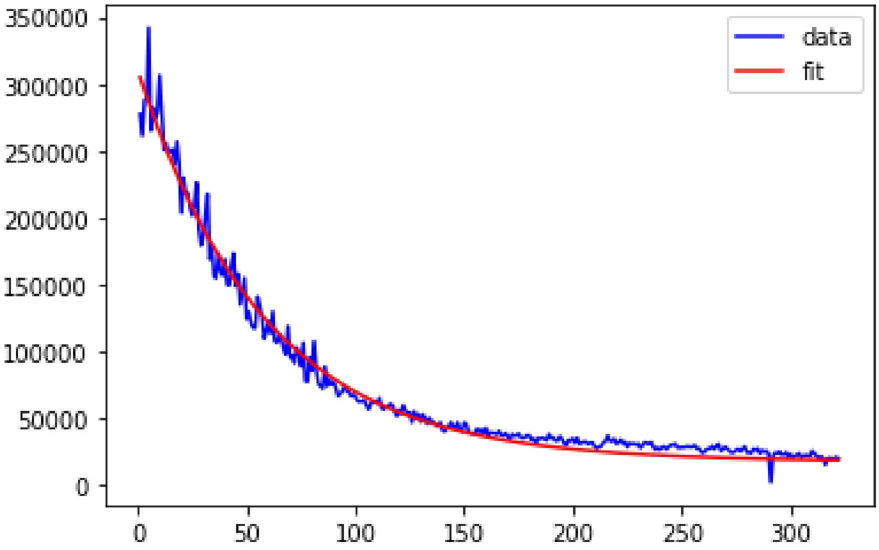

Figure 11: Reports Variation by Exponential Decay

Upon substitution of t = 322, representing the number of days elapsed since the initial observation, into the revised exponential decay equation, the predicted number of reports on March 1st 2023 was calculated to be 19139. This estimate provides valuable insights into the expected number of reports and may be used to develop informed decisions concerning the future progress of the Wordle game.

4.5 Strengths and Weaknesses

4.5.1 Strengths

Our regression model demonstrates a high level of accuracy in predicting the number of reported results in the Wordle game based on days following the initial observation date and game popularity. The model's accuracy is evidenced by its high R-squared value and low mean absolute percentage error (MAPE), despite the large range of expected results. Our rigorous hyperparameter tuning and model selection approach contributed to the model's success. Thus, we anticipate that our model can provide valuable insights into the game's performance and popularity trends over time.

4.5.2 Weaknesses

To enhance the performance of our machine learning models, we made the decision to remove a spike in the dataset by eliminating some of the data because we believed that the peaks in the dataset were not representative of the true underlying distribution, and if we left them in, the model might learn to be biased towards the peaks and perform poorly on data that does not have similar peaks. This can result in poor performance on the test data. However, removing data can negatively impact the representativeness and generalizability of the dataset, which is a potential limitation of our approach.

Furthermore, there are other factors beyond the current features that may impact the distribution of results in the Wordle game, which we have not taken into account. For instance, the difficulty of the words used in the game or the number of similar games available to users may also play a role. These factors could influence user behavior and preferences, which in turn could affect the distribution of results. This is another potential limitation of our approach. Additionally, while our model produced high R-squared and low MAPE values, the high mean squared error (MSE) values we observed are a cause for concern. The substantial range of expected results (100,000-300,000) may contribute to a higher degree of prediction error, leading to inflated MSE values. Additionally, the limited number of observations and features in our model, as well as their high correlation, could result in overfitting, which could also contribute to high MSE values. Therefore, we need to interpret the MSE values with caution and avoid overly optimistic conclusions about the overall accuracy of our model.

4.6 Consideration and Future Work

There are several considerations and potential avenues for future work related to the analysis of the Wordle game dataset. One key consideration is the need for careful evaluation and interpretation of the results to ensure their validity and reliability. This may involve employing additional techniques, such as cross-validation or bootstrapping, to assess the robustness and generalizability of the models.

In terms of future work, several strategies can be employed to improve the performance of machine learning models in analyzing the Wordle game dataset. One such approach is to expand the dataset by obtaining additional data to increase the size and diversity of the training set. This can help mitigate the issue of overfitting, which can result in poor performance on new, unseen data. In addition, incorporating additional features into the model could help capture the influence of factors beyond those currently considered. For example, natural language processing techniques could be used to analyze the difficulty of the words used in the Wordle game, or data on other similar games available to users could be collected to better understand the impact of competition on user behavior.

Furthermore, exploring alternative machine learning algorithms, such as support vector machines, neural networks, decision trees, or ensemble methods, can help identify the most suitable model for the dataset. Testing a range of models can provide a better understanding of the underlying relationships between the features and the outcome variable, and may help overcome some of the limitations of the current models, such as high mean squared error values or overfitting. By doing so, it may be possible to improve the performance and interpretability of the models, leading to more reliable and accurate predictions. As a result, these findings can contribute to a better understanding of user behavior in online gaming environments, with potential implications for the development of more effective predictive models in the future.

5 Conclusion

In conclusion, our study highlights the potential of user-generated data to develop predictive models and gain insights into the factors that impact the Wordle game. By exploring the relationship between time, popularity, and the number of reported results, we found that the exponential regression model with time as a predictor variable provided the most accurate results. We also introduced a predictive model that simulates the Wordle game in hard mode, and our analysis revealed that challenging words often contain duplicate and rare letters, but their difficulty level does not affect the number of submissions. Our study's findings provide a foundation for further research on the Wordle game and demonstrate the potential of data-driven approaches to study popular online games.

References

[1] Twitter. (2022). *Wordle Stats*. Retrieved from https://twitter.com/WordleStats

[2] Photutorial. (2022, October 4). *Wordle Statistics, Facts, & Strategies*. Retrieved from https://photutorial.com/wordle-statistics/#:~:text=Wordle%20was%20created%20by%20Josh,but%20the%20number%20is%20declining

[3] Google. (2022). *Wordle*. Google Trends. Retrieved from https://trends.google.com/trends/explore?date=2022-01-07%202022-12-31&geo=US&q=wordle

[4] Tatman, R. (2017). English Word Frequency. [Dataset]. Kaggle. https://www.kaggle.com/rtatman/english-word-frequency.

[5] Bertsimas, D. & Paskov, A. (2022). An Exact and Interpretable Solution to Wordle.