By: Jakob Pachucinski, Williams Prevet, and Henry Caso

Table of Contents

1: Letter to the Puzzle Editor of the New York Times

Dear Puzzle Editor,

As you have probably noticed, Wordle's popularity has declined steadily since its peak of 361,908 reported results on February 1st, 2022. In fact, towards the end of 2022, the amount of reported players dipped as low as 15,000. While I doubt these are the numbers you wanted to see, we may have some answers in order to make Wordle's popularity boom once again.

Each Wordle word has its own difficulty associated with how many attempts it takes each player to guess correctly. After doing extensive analysis on letter distributions and letter frequencies, we have two main suggestions.

Certain letters like *t* and *a* are frequently found in easy words. Assuming that most Wordle players like a challenge, it may be better to incorporate uncommon letters like *y*, *f*, and *m*. These letters were often found in hard words. Using these letters could give word puzzle fans more of a rush when they finally discover what the missing letter was.

Not only did we find certain letters increase a word's difficulty, but so do multiple letters. Out of our easy words there were only ***2*** cases where a letter was repeated. However, in the sample of our hard words, ***28*** happened to have either one or two letters that repeated. Using double letters (or even triple letters like *mummy*) could force players to adopt new strategies when playing Wordle.

Not only will using uncommon letters and double letters make Wordle players happier, it could also increase chances of a word going viral. With the relevance of social media in today's society, any tweet that blows up has the potential to bring a substantial amount of new customers to any business. If Wordle challenges start to become harder, the *word* will get around about the recent difficulty and possibly spark more people's interest in the game.

Overall, making the daily Wordle harder with uncommon and repeating letters could provide players with more of a challenge when solving the puzzle. The more difficult the words get, the more time players will have to spend on Wordle trying to figure out what the puzzle is. This could then lead to players asking friends, family, or coworkers for help and posting on social media about how challenging Wordle has become. If you would like to see the details on how we got these numbers and other interesting features we found feel free to read the entirety of our paper. We hope you take this solution into consideration.

Thank you.

2: Introduction

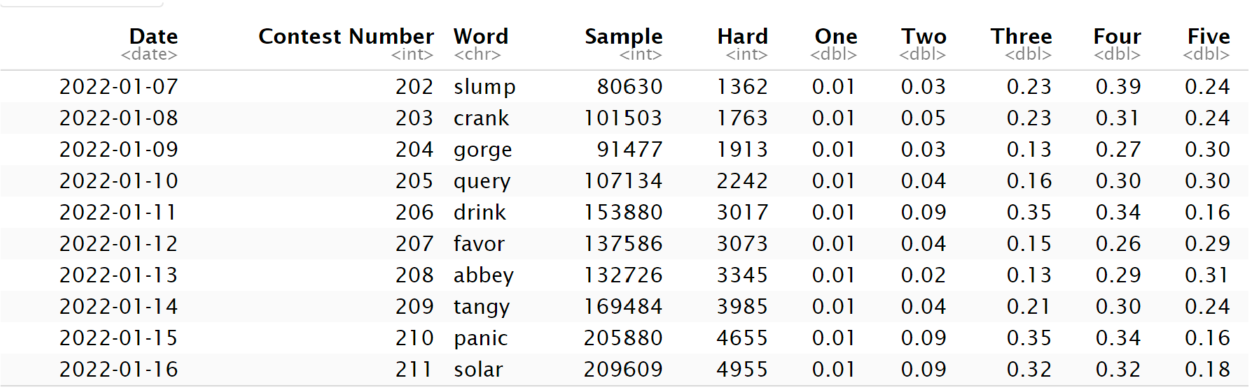
The problem posed to us is fourfold. We must analyze the dataset given to us, containing the solution words for any given day, the number of people reporting their scores on a given day, and the percentage of those people to complete the day's Wordle in 1, 2, 3, 4, 5, 6 guesses or not at all. Using the data provided we must develop 3 models. One model must explain why the number of people reporting their scores to Twitter varies from day to day and be able to predict with a level of accuracy how many people will report their score on a given day, in this case March 1, 2023. We also must determine if and how the aspects of the day's word affect the number of reported scores. The second model must predict the distribution of the reported results, i.e. what percentage of the reported results solve in 1 guess, 2 guesses, 3, etc. We must state our uncertainties and predict the results given the word EERIE for the date March 1, 2023. Our last model must be able to classify a given word by its difficulty for a player to solve it. We must test it again for the word EERIE and describe the accuracy of our models.

3: Analysis of the Data

We used Microsoft Excel, RStudio, and Python to analyze the Wordle data provided. First, we had to clean the data before we manipulated it. Then, we looked at the frequency of letters in the overall words, created a statistic to rank each word's difficulty, and looked at time plots to see any trends.

3.1 Data Manipulation

The main cleaning came when importing the Excel file into RStudio and Python. The first row of the table was imported as what we wanted the column names to be. So, we had to rename each column to fit our needs, and delete the first row of our table. The number of scores for each word was given in percentage, so we changed it to a ratio by dividing each percentage by 100. Below is an example of our cleaned table:



Date	Contest Number	Word	Sample	Hard	One	Two	Three	Four	Five
<date>	<int>	<chr>	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2022-01-07	202	slump	80630	1362	0.01	0.03	0.23	0.39	0.24
2022-01-08	203	crank	101503	1763	0.01	0.05	0.23	0.31	0.24
2022-01-09	204	gorge	91477	1913	0.01	0.03	0.13	0.27	0.30
2022-01-10	205	query	107134	2242	0.01	0.04	0.16	0.30	0.30
2022-01-11	206	drink	153880	3017	0.01	0.09	0.35	0.34	0.16
2022-01-12	207	favor	137586	3073	0.01	0.04	0.15	0.26	0.29
2022-01-13	208	abbey	132726	3345	0.01	0.02	0.13	0.29	0.31
2022-01-14	209	tangy	169484	3985	0.01	0.04	0.21	0.30	0.24
2022-01-15	210	panic	205880	4655	0.01	0.09	0.35	0.34	0.16
2022-01-16	211	solar	209609	4955	0.01	0.09	0.32	0.32	0.18

In the middle of doing our analysis we ran into two different problems:

1. Words Not in the English Language:

    a. rprobe: We took this to be “probe”.
    b. marxh: We took this to be “march”.
    c. clen: We took this to be “clean”.
    d. naïve: We took this to be “naive”.

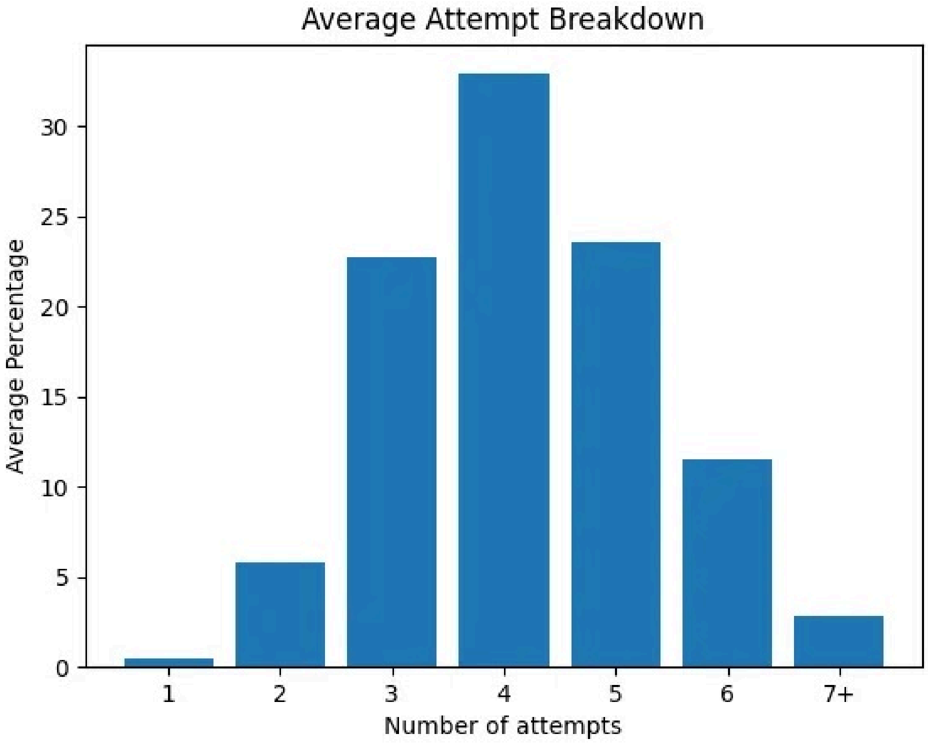
2. Outliers in Number of Reported Results:

    a. On November 30th, 2022 the number of reported results was 2,569. This was about 20,000 off of any other date around it. We took it as 22,569 since the number of reported results around that time were close to 22,000.

We also created new statistics rating the difficulty of each word, and finding the ratio of hard-mode players:

1. **Word Difficulty**:

    For word difficulty we wanted to incorporate the number of tries it took to solve each word. Where if a word took more attempts (on average) to solve, it would be considered more difficult than a word that took less attempts (on average) to solve. Before creating the statistic we looked at the distribution of attempts:



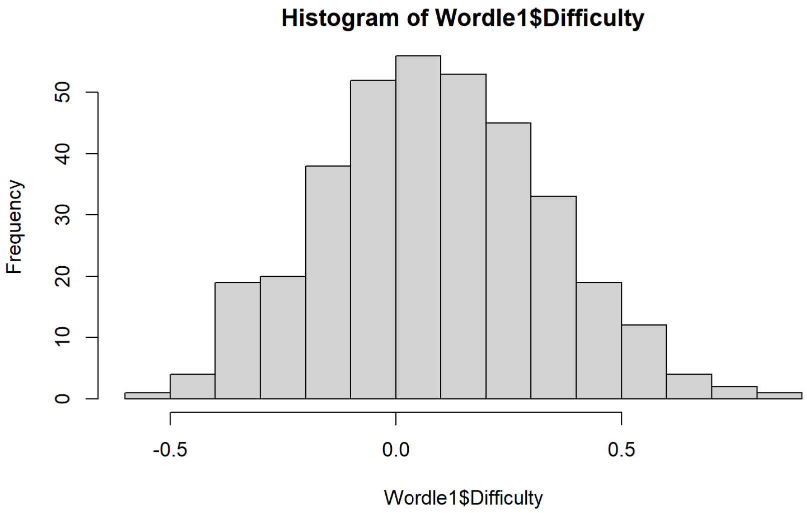
Seeing that this graph was somewhat normal, we will consider difficult being 5 guesses or more and easy being 3 guesses or less. We consider 4 guesses normal and don't apply it in the equation. The word difficulty equation can be seen below:

$$
\begin{aligned}
&Word\ DifficultyScore\ = \\
&[\\
&(proportion\ of\ correct\ guesses\ on\ the\ 5th\ attempt) \\
&+\ (proportion\ of\ correct\ guesses\ on\ the\ 6th\ attempt) \\
&+\ (proportion\ of\ correct\ guesses\ on\ the\ 7th\ attempt\ or\ more) \\
&] \\
&-\ [\\
&(proportion\ of\ correct\ guesses\ on\ the\ 1st\ attempt) \\
&+\ (proportion\ of\ correct\ guesses\ on\ the\ 2nd\ attempt) \\
&+\ (proportion\ of\ correct\ guesses\ on\ the\ 3rd\ attempt) \\
&]
\end{aligned}
$$

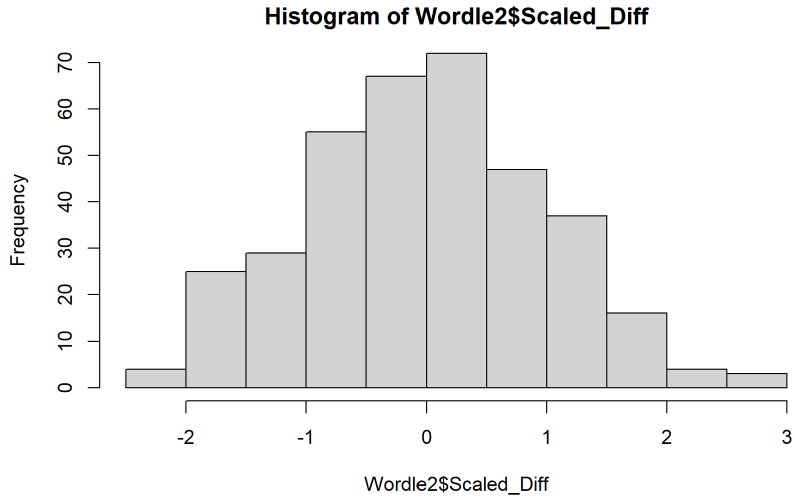
This way if a word's difficulty score is positive, it is considered a hard word. Meaning, on average, players took more than 4 turns to solve the Wordle.

If a word's difficulty score is negative, it is considered an easy word. Meaning, on average, players took less than 4 turns to solve the Wordle.

The spread of word difficulty is shown below:

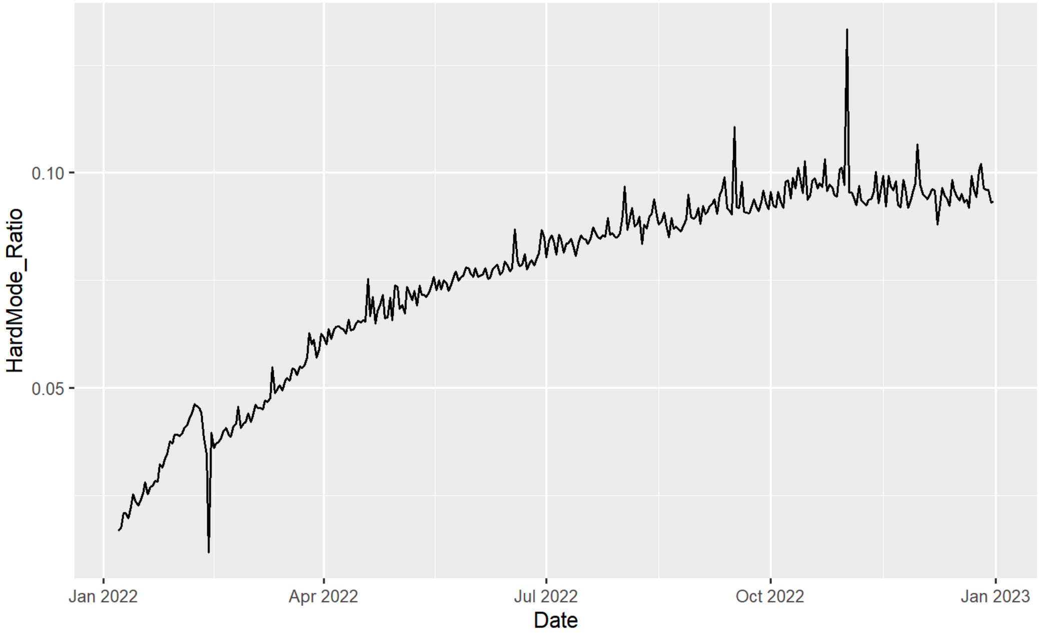


The histogram of word difficulty seemed approximately normal, but was not centered around 0. We decided to normalize it so we can use standard deviations. The normalized distribution of word difficulty is shown below:



2. **Hard Mode Ratio**:

   In order to predict how many people will use Wordle's hard-mode on March 1st, we created a statistic called Hard Mode ratio which divided the number of hard-mode players by the total number of reported scores. The graph of Hard Mode ratio is shown below:



   It appeared to stabilize around October.

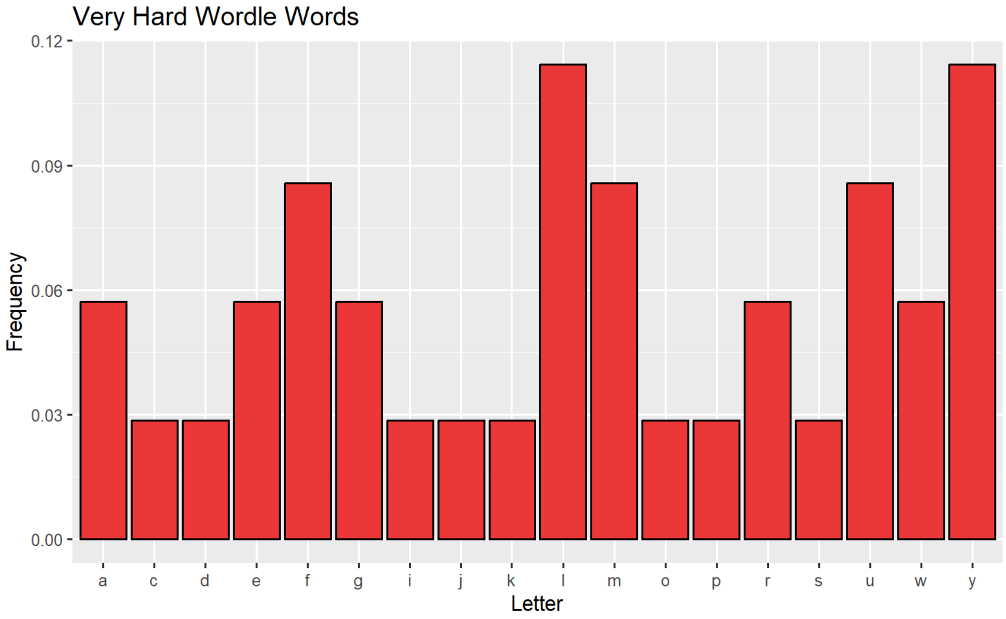
3.2 Word Types

Using our normalized word difficulty score, we decided to classify words as either very easy, easy, normal, hard, or very hard. This classification was determined by where the words fell in the normalized distribution.

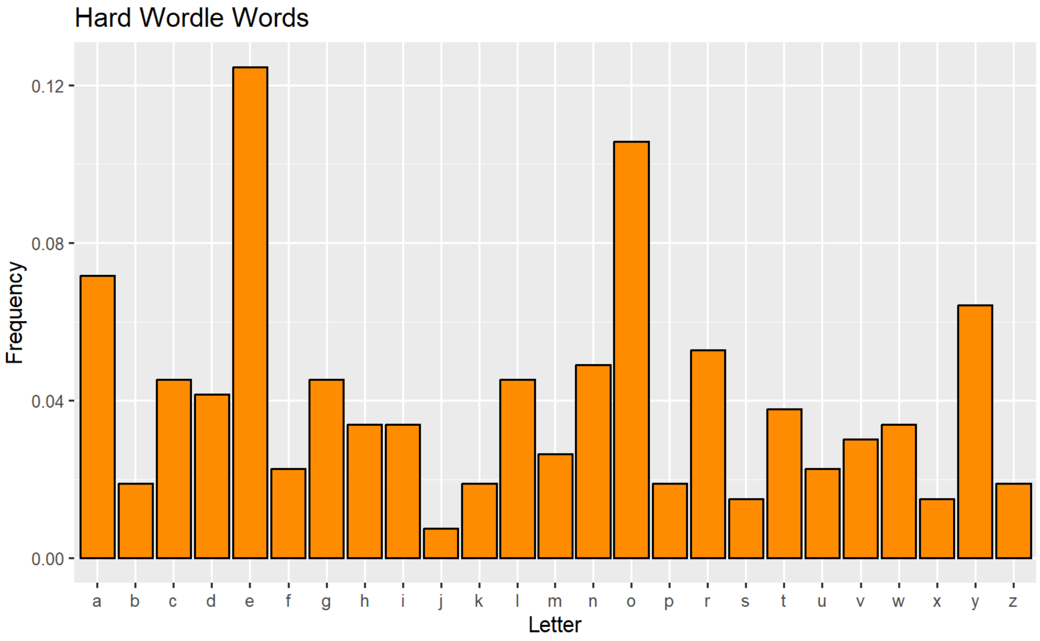
A very easy word is classified as having a normalized difficulty score of less than -2. An easy word is classified as having a normalized difficulty score between -1 and -2. A normal word is classified as having a normalized difficulty score between -1 and 1. A hard word is classified as having a normalized difficulty score between 1 and 2. A very hard word is classified as having a normalized difficulty score greater than 2.

We wanted to use the letter distribution of each category to see what letters typically showed up in very easy words, easy words, and so on. This way we can create a model to predict how hard the Wordle word will be by using the letters in the given word.

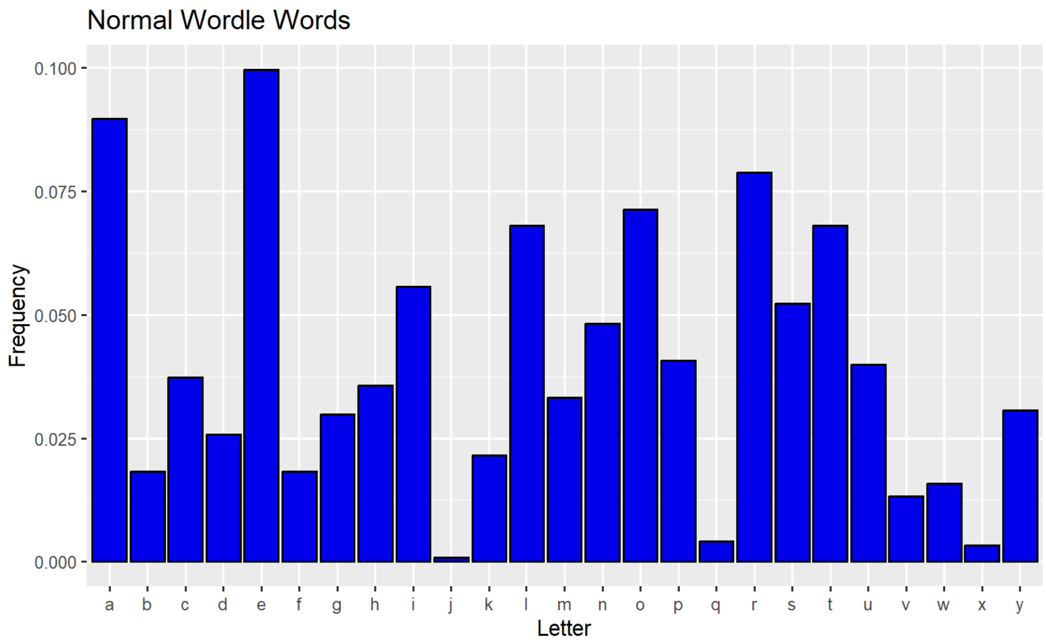
Each categories frequency distribution is shown below:



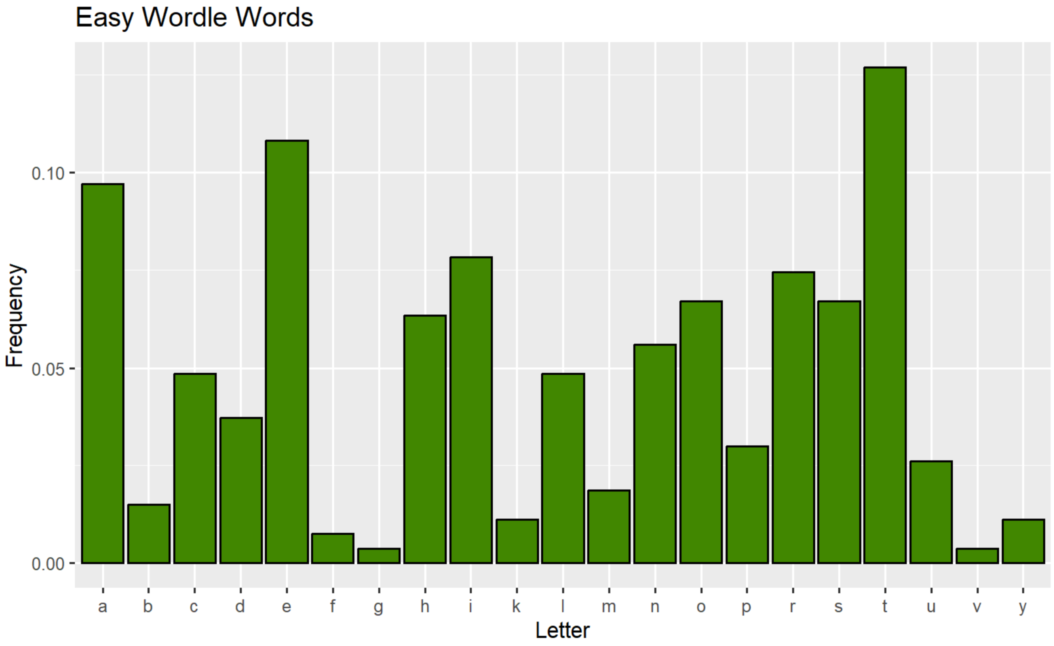
The most frequent letters in very hard words seem to be: “y”, “l”, “m”, “u”, and “f”.



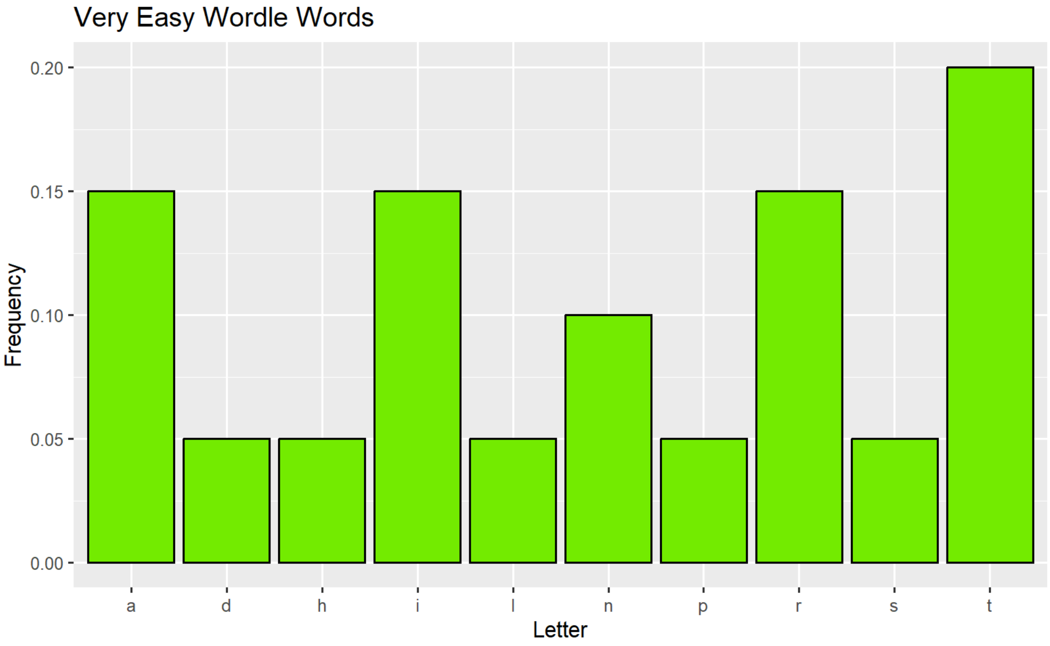
The most frequent letters in hard words seem to be: “e”, and “o”.



The most frequent letters in normal words seem to be: “a”, “e”, and “r”.



The most frequent letters in easy words seem to be: “t”, “e”, “a”, and “i”.



The most frequent letters in very easy words seem to be: “t”, “r”, “i”, “a”, and “n”.

3.3 Interesting Features

Throughout this data analysis we discovered a few interesting features of certain letters and words.

The word *feast* word would be categorized as either a hard or very hard word in our model due to the presence of *f* and *e*. However, it ended up being categorized as normal from the actual results. Looking at the date, *feast* was the word on Thanksgiving which seems to explain why more players guessed it on the first or second try than other words.

Looking at the very hard words table, more than half of the words had two of the same letter. While none of the very easy words had double letters. We thought that was interesting so we looked at double letters in the easy words compared to the hard words.

There were only two instances of easy words having double letters: *tiara* and *photo*. Both happened to be double vowels. Also, their normalized difficulty scores were 0.1 from being considered normal difficulty words.

As for the hard words, there were 28 words that had either one repeating letter, or two. This was extremely interesting because it shows how players tend to use strategies in Wordle. Most likely, once they guess a letter correctly, their guesses will tend to go towards other letters instead of guessing the same letter that they guessed right already.

4: Model Explanation

4.1 Overview

In order to predict the impact that a given word will have on the guess distribution, we considered the difficulty of the letter composition for that word. We also took into account the presence of the same letter multiple times in a single word, and its impact on the word's difficulty.

To determine the total player who will report their scores on a future date, we used a best fit trend line to create our model. From this, we also created a model for the number of Hard Mode players as a consistent fraction of the total players.

4.2 Assumptions

Letter frequency is Related to Word Difficulty: From our letter frequency analysis, we determined that the less frequent a letter is in English words, the more likely it is for words that contain it to be classified as a "difficult" word. We assume that words containing common letters will have a guess distribution skewed towards fewer attempts, and obscure letters will have a guess distribution skewed towards more attempts.

The Letter *q* Brakes This Pattern: While *q* is one of the least common letters (making up only 0.2% of the letters in the data set), it only ever appeared in words classified as a "normal" word. It is assumed that a *q* is almost always followed by a *u*. Thus, correctly guessing *q* also gives *u*–and based on context, guessing *u* sometimes gives *q*–essentially providing twice the value and making the word easier to solve.

Double Letters Make a Word Harder to Solve: Some of the words used for the puzzles contain the same letter more than once. Due to the way Wordle works, guessing a letter correctly provides no information about the count of that letter. This could make it challenging to ascertain if it is worth using a letter multiple times in a single guess. Our analysis showed that double letter words were more likely to be classified as "difficult," and of those, double consonants were more difficult than double vowels. It is assumed that double letters increase the difficulty of a word.

Triple Letters Make a Word Much Harder to Solve: By similar logic to the double letter, a triple letter is even harder to detect in a puzzle. The only triple letter word in the data set, mummy, was classified as "very difficult." It can be assumed that triple letters greatly increase the difficulty of a word.

Words of Similar Difficulty Will Have Similar Attempt Breakdowns: As described in the data analysis section, words were broken into five discrete classifications based on their solve distributions. These classifications can all be assigned a difficulty score range that represents the difficulty level of the words in that category. It is assumed that a word with a difficulty level matching one of the classifications will have an attempt breakdown similar to the average score breakdown of that section

Wordle's Popularity is falling off at a predictable rate: Since its peak reported playership on February 2, 2022, the total number of players posting their scores on Twitter has been decreasing. Plotting the number of players versus the date shows a decaying exponential relationship. It is assumed that Wordle's player base will continue to fall off at this rate.

4.3 Definition of Variables

Notation	Definition
L	Letter difficulty
B	Base word difficulty
D	Difficulty modifier for a double letter
T	Difficulty modifier for a triple letter
W	Final word difficulty
t	Days since June 1, 2022
N	Total number of players
H	Number of players reporting for hard mode
c	Ratio between hard and total players

To assign scores to each of the letters, we first took the weighted average of each letter's frequency across the different word difficulty categories. This ensured that the letter's score reflected the frequency of the letter and what difficulty of words it most often appeared in. We then ordered the letters based on the average and gave them difficulties between -3 and 3, spaced evenly. The values are summarized below.

Letter	Difficulty(L)	Letter	Difficulty(L)
t	-3.000	z	0.000
i	-2.769	v	0.231
a	-2.538	j	0.462
r	-2.308	c	0.692
n	-2.077	o	0.923
h	-1.846	l	1.154
s	-1.615	e	1.385
p	-1.385	w	1.615
d	-1.154	g	1.846
q	-0.923	u	2.077
b	-0.692	m	2.308
x	-0.462	f	2.538
k	-0.231	y	2.769

4.4 Score Distribution

In order to add a formal mathematical interpretation to our difficulty classifications, we assigned numerical scores matching the standard deviations of the scaled distributions. They are as follows:

Classification	Difficulty Range
Very Easy	$-3 \leq W < -2$
Easy	$-2 \leq W < -1$
Normal	$-1 \leq W < 1$
Difficult	$1 \leq W < 2$
Very Difficult	$2 \leq W$

Our model for predicting a difficulty score for a word is:

$$W = B(L) + D + T$$

Where,

- B is the average letter difficulty.
- D is the presence of a double letter. D is equal to 1 if the letter is a vowel and 2 if the letter is a consonant.
- T is the presence of a triple letter. T is equal to 3 for all letters.

The theoretical "easiest" word predicted by this model would be a word containing the letters *t*, *i*, *r*, *a*, and *n* exactly once each, for a score of -2.538. The theoretical "hardest" word possible would be a word containing three of the letter *y* and two of the letter *f*, for a score of 7.677.

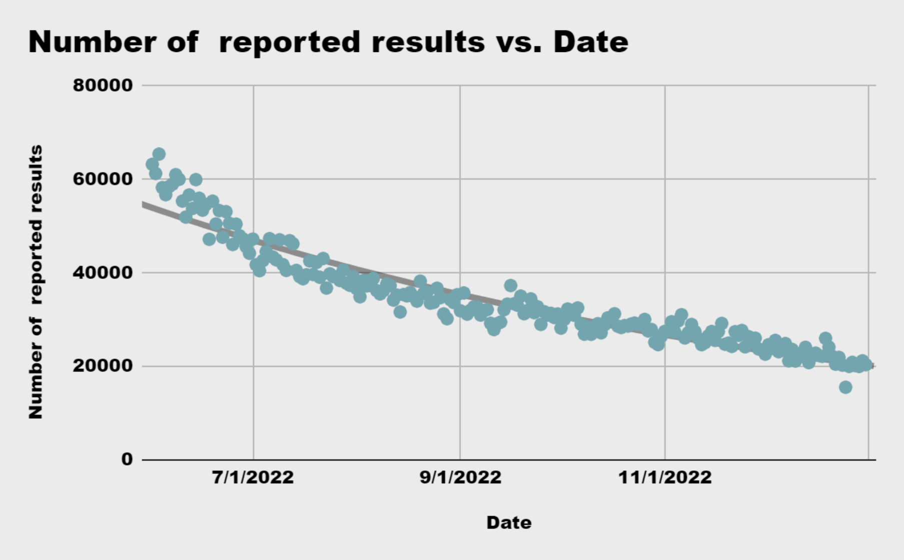
Once a word's difficulty score is calculated, it can be matched to one of the classifications. In order to approximate the attempt breakdown for that word, it is given the average score breakdown for that classification. Those breakdowns are summarized in the table below:

Attempt Distribution for a Given Word Difficulty						
Difficulty Score:		-3 ≤ W < -2	-2 ≤ W < -1	-1 ≤ W < 1	1 ≤ W < 2	2 ≤ W
Classification:		Very Easy	Easy	Normal	Hard	Very Hard
Attempt Distribution (%):	One	2.75	1.11	0.36	0.19	0.14
	Two	19.00	12.06	5.19	2.19	0.71
	Three	38.50	33.72	22.85	12.06	5.29
	Four	27.00	32.31	34.61	28.36	17.71
	Five	10.00	15.19	23.99	30.96	29.14
	Six	3.00	4.87	10.66	20.72	29.71
	Seven+	0.25	0.72	2.20	5.98	17.00

4.5 Player Count

On any given day, the number of people who report their Wordle scores can fluctuate significantly. However, considering the player count over the course of 2022 illustrates a correlation between time and the number of reports. The number of players depending on the day spiked drastically (±10,000 or more) around Wordle's peak popularity, but settled into a more reliable pattern as the year progressed.

In order to make our predictions, we considered the number of reported results from June 1, 2022. We picked this date because all the data onwards have a clear visual correlation. The scatter plot below illustrates the trend in the data.



From the plot, we extracted the best fit line, which as stated before, follows a decaying exponential trend. We took this line for our model of total player count:

$$N(t) = 53980 \cdot e^{\left(-4.6\cdot10^{-3}\right)\cdot t}$$

Where,

- t is the number of days since June 1, 2022

The coefficient of determination for this trend line is $R^2 = 0.911$, which indicates that the trend in the data is very similar to function. While it is true that the larger t becomes, the less accurate this model will be, for dates close to the data points, the model should predict approximately how many remaining players will report their scores.

As described in the data analysis section, since October 2022, the number of people who report their scores in Hard Mode tends to be a consistent fraction of the total reports for any given day. Thus, our model for the number of reported scores in Hard Mode is:

$$H(t) = c \cdot N(t)$$

Where,

- c is the average constant ratio, 0.09.

Since the Hard Mode model is directly proportional to the total reports, its efficacy will fall off at the same rate as the previous model.

5: Model Application

5.1 March 1, 2023 Score Distribution:

In order to test the effectiveness of our model, we were given the Wordle unknown word for March 1, 2023: *eerie*. We were asked to evaluate the word and determine the distribution of solve attempts for that day.

Here is our how our model evaluates *eerie*:

- The word consists of three of the letter e (1.385), and one of the letters r (-2.308) and i (-2.769). This works out to a B value of -0.922.
- The word does not contain a double letter, so D is equal to 0.
- The word does contain a triple letter, so T is equal to 3.

Putting the terms together, we got:

$$W =- 0.922 + 0 + 3 = 2.078$$

A score of 2.078 classifies *eerie* as a "very difficult" word to solve. From this, the expected solve distribution should be approximately:

Attempts	Percentage of Reports
One	0.14
Two	0.71
Three	5.29
Four	17.71
Five	29.14
Six	29.71

Seven+	17.00

5.2 March 1, 2023 Player Count

We were also asked to model how many players would report their scores on March 1.

Here is how our model evaluates March 1, 2023:

- It is 273 days after June 1, 2022, so t is equal to 273.

Plugging t into our function yields:

$$N(273) = 53980 \cdot e^{\left(-4.6\cdot10^{-3}\right)\cdot273} = 15376$$

This indicates that there will be approximately 15,376 people who will report their scores for *eerie*. It is difficult to say how "good" this value is, since the last data point given is two months prior, but if Wordle has remained on a similar decline, the number should be a good ballpark estimate.

Additionally, our model can predict the number of those people who will play the game in hard mode:

$$H(273) = 0.09 \cdot 15376 = 1384$$

Since *eerie* is a "very difficult" word, we can expect that a lot of people will have a high number of attempts, which is especially true for Hard Mode. Since the scores for this word may be bad, we might expect less people to share their scores publicly. This fact is likely even more true for Hard Mode players, as they are the ones who voluntarily make the game more challenging and take pride in being able to overcome tough puzzles. However, the number of responses will be around 1384.

6: Model Analysis

6.1 Strengths

Not only did we take into account different letters' difficulty, we even used the prevalence of double, or triple letters. This will help us predict words that may not have hard letters, but the repeating letters throw off the player's strategy.

Our model is taken straight from the data. This means that previous Wordle results are directly being used to predict the future results. As long as there are no major changes, our use of the previous results should be enough to accurately predict the category of each word.

The model for the amount of players has an extremely good R^2 value. This means that we should be pretty accurate for our prediction of the number of people reporting their Wordle scores on March 1st.

We kept our model simple. This will allow us to more accurately predict within a range of distributions. If we added too many variables there was bound to be a word that broke our scale and created one or multiple outliers.

6.1 Weaknesses

Players who usually report their scores could be sore losers. This would mean that if a player doesn't get the word, or gets the word in more than four tries, they may not report their score or lie and report a better score than they actually got.

If the word is actually *eerie* on March 1st, contestants in this competition may tell their friends to guess the word on the first try. This would skew the results to make *eerie* seem like an easy word. Based on our prediction

Our prediction for the number of players does not include all of the data. We started on June 1st because that seemed to be where exponential decay was most prevalent and allowed for us to have a plausible prediction. Given more time, we would be able to create a better model that could predict the number of players from the beginning of 2022, and use that to see what variables could cause the value to increase or decrease.

We did not use a model for predicting the number of Hard Mode players but instead used the median ratio since June 1st. Given more time, we could fit a logarithmic function to the Hard Mode ratio and find different variables (like relevant articles) that lead to the increase of players using Hard Mode.

6.3 Sensitivity Analysis

Our model is very flexible and can be tweaked to fit any word in the English language. Therefore, if Wordle ever decided to expand the word length to six letters, or decrease to four letters, we could provide predictions on words they could use to spread out the difficulty.