Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2322040

Regressing Wordle

Abstract

Analyzing social media posts for the puzzle game Wordle is a challenging problem. We must take data from the past year of Worldle-related twitter posts, and use it to predict future results. From this data, we can also extract information of what makes a Wordle word difficult and use that to predict future score distributions.

To predict the number of reported results on a given day, we regress the number of reports on time by the Fréchet distribution, and extrapolated that function to predict results. We also use sigmoid function to model the percentage of hard mode players over time. We conclude that both models fit well into the observed data.

Meanwhile, we use multivariate linear regression to model word difficulty and predict future score distributions. We find that five word attributes have significant impact on the reported score: starting letter frequency, ending letter frequency, common letter combination, duplicate letters, and frequency of the word. While having duplicate letters correlates with increased difficulty, the other four attributes all make the word easier. Thus, we use the attributes' coefficient on the average score to create an index for word difficulty.

We also include time and percentage of hard mode players in our linear model. We find that time has a negative relationship with the average number of attempts while percentage of hard mode players has a positive relationship.

Our models predict that the player count will continue its current slow decline, with the proportion of hard mode players leveling off at 9.675%. Our models also predict that, on March 1, with word "EERIE," the number of reported results will be 11,883 with 95% confidence interval [7,314, 16,668], the percentage of hard mode players will be 9.646%, and the score distribution will be (1 try: 0.4%, 2 tries: 6.2%, 3 tries: 25.8%, 4 tries: 34.7%, 5 tries: 22.2%, 6 tries: 8.9%, failure: 1.8%).

Contents

1 Introduction

1.1 The Problem

For this problem, we must create models that can do the following tasks:

- Predict the number of reported results on a given date in the future,
- Classify words based on difficulty,
- Predict the distribution of reported results, given a word and a future date.

2 Assumptions

- We ignore outside effects from sources such as twitter management, world events, etc.
- We assume Wordle will not undergo drastic change soon
- We ignore the effects of cheaters
- We assume that word difficulty and time are independent variables
- We assume that the majority of players after the peak are not new, (i.e. they are regulars)
- The variance in regression error is homogeneous and normally distributed.

3 Models

3.1 Predicting Future Number of Results

3.1.1 Note on Daily Variation

When looking for potential causes of day to day variation in the data, we examined the relationship between the number of posted results and the following factors:

- Probability of the word appearing based on how often it appears in the English language [2]
- Location of vowels
- Number of repeated letters
- Whether there are consecutive repeated letters
- Location of repeated letters
- Average number of guesses to get the word right
- Whether there are high-frequency, similar words (i.e. words differing only by one letter, as suggested by Waldron [10])

and we found **no statistically significant relationship**.

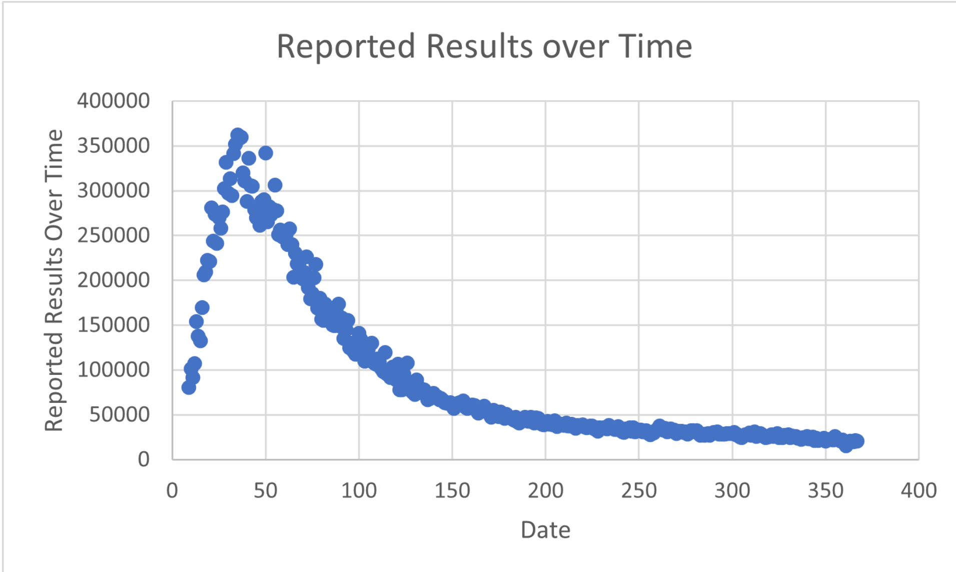

Figure 1: Results posted over time compared to date, where Jan 7th is day 1.

3.1.2 Fréchet Distribution Model

When viewing the reported results over time (figure 1), we can see similarity to the Fréchet distribution, a special case of the generalized extreme value function [9], with the probability distribution function

$$f(x) = \frac{\alpha}{s}\left(\frac{x-m}{s}\right)^{-\alpha-1} \exp\left\{-\left(\frac{x-m}{s}\right)^{-\alpha}\right\} \tag{1}$$

Where $m \leq x$ is the location parameter, $s > 0$ is the shape parameter and $\alpha > 0$ is the scale parameter.

Historically, the Fréchet distribution has been applied to several real-world problems, from predicting the rainfall from a flood [3], to predicting rates of oil extraction from a well [5]. More recently, the Fréchet distribution has been used to model internet fads [1]. Wordle is an internet fad, so we apply this distribution to understand the fluctuating number of reported results from Wordle. The Fréchet distribution is particularly well-suited to this problem because of its long tail, which mimics the mild popularity that Wordle experienced in the second half of 2022. We created a model that uses the Fréchet distribution as a basis, and regressed it to our data (figure 2).

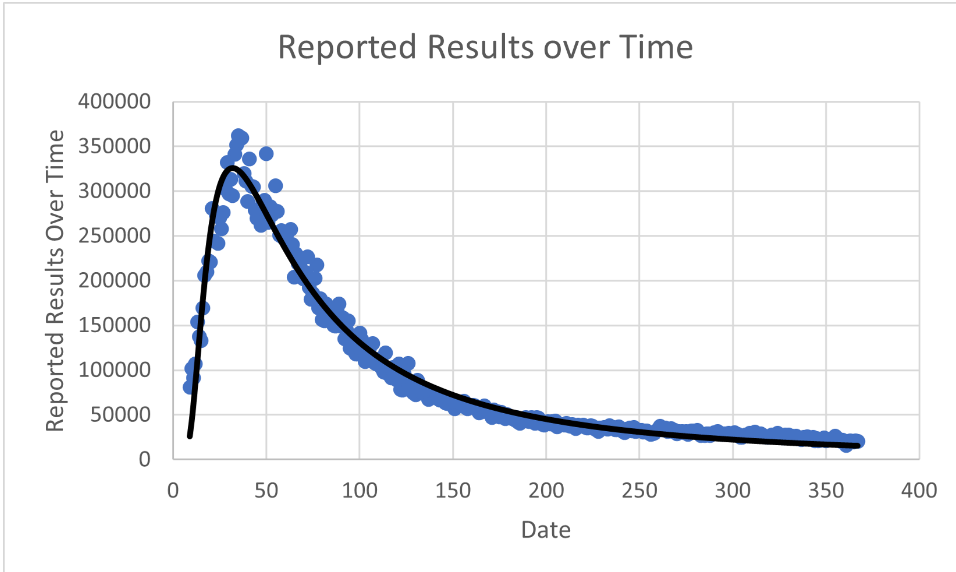

Figure 2: Results posted over time, with the fitted Fréchet distribution in black.

3.1.3 Predicting Hard Mode Results

We examined attributes of the words to find correlation with the percentage of players in hard mode, and found that all of the following factors had an R^2 of less than 0.03:

- Location of vowels
- Probability of each letter appearing using a letter count corpus [6]
- Number of repeated letters
- Number of Vowels
- Identity of the letters in each location
- Consecutive repeated letters
- Location of repeated letters
- Average number of guesses to get the word right
- Letter frequency [6]
- Whether there are high-frequency, similar words (i.e. differing only by one letter as suggested by Waldron [10])

The only factor with a statistically significant correlation was time. When viewing the percentage of players in hard mode, as shown in figure 3, we see that the data closely resembles a sigmoid function. The standard sigmoid function is of the form

$$f(x) = \frac{1}{1+e^{-x}} \tag{2}$$

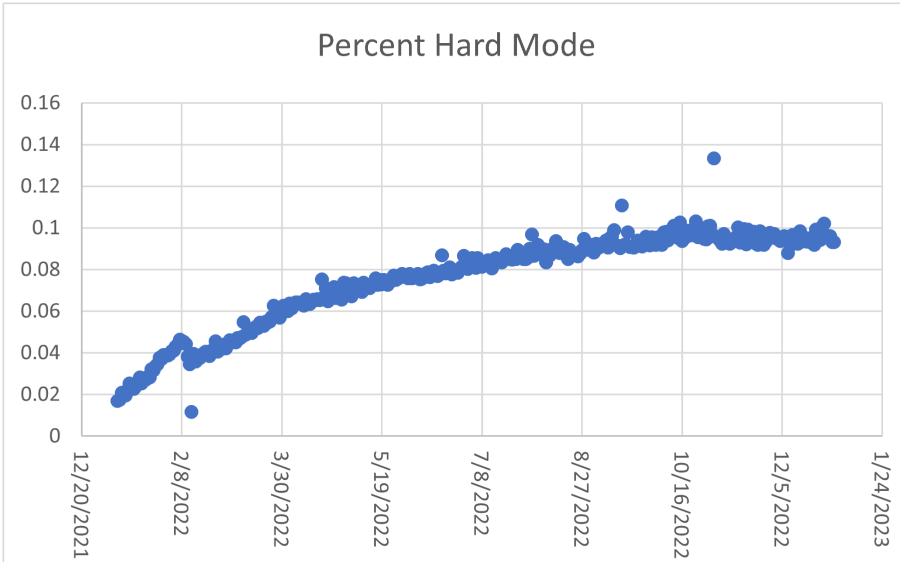

Figure 3: The percentage of results posted in hard mode, with the X axis being date.

We chose a the sigmoid as the based for our model of this data. We used the solver function in Excel to minimize the mean absolute deviation between the observed percent of hard mode players and the sigmoid. This resulted in the following function:

$$f(x) = \frac{0.09675}{1 + e^{-0.0162(x-262)}} \tag{3}$$

Our results, displayed in figure 4, closely match the dataset, with an R^2 value of 0.997. **On March 1st, we estimate that 9.65% of results will be from hard mode players.**

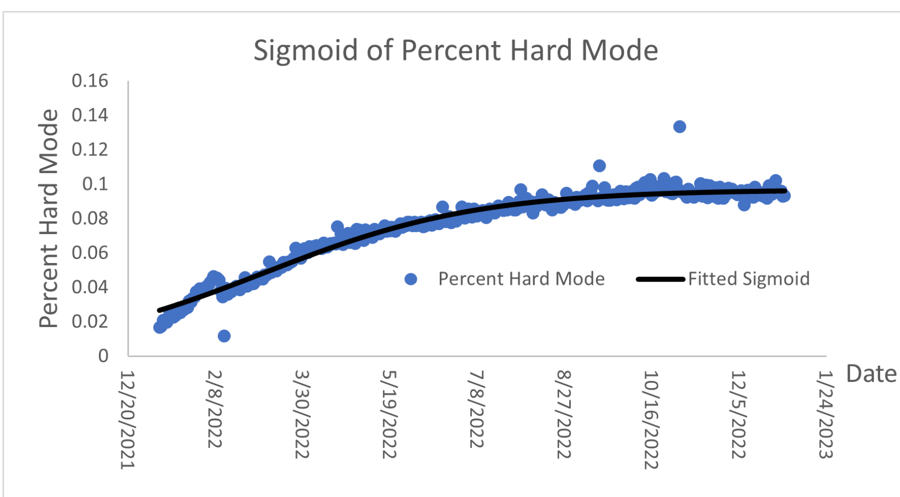

Figure 4: This graph shows a the sigmoid function from equation 3 overlayed on the data described in figure 3

Symbol	Meaning
$\mathbf{X}$	Vector containing relevant variables (described below) for data point i
$s_i(\mathbf{X})$	Ratio of reported results that take i tries. $i \in \{1,2,3,4,5,6\}$, because ratio of failures (i=7) is determined by one minus the rest of the ratios
β_j	Scalar modifying x_j
x_1	Boolean variable for if the word **starts** with one of the 5 most common starting letters
x_2	Boolean variable for if the word **ends** with one of the 5 most common ending letters
x_3	Boolean variable for if the word contains one of the most frequent two or three letter combinations in English
x_4	Boolean varaible for if the word has duplicate letters
x_5	The logarithm of the word's frequency
t	Time in days, with Jan 7 2022 being day 0
r	Ratio of hard mode players
C	Constant term
ε_i	Error term

Table 1: Variables for equation 4. Letter frequencies from [6]. Word frequencies from [2].

3.2 Word Difficulty and Score Distribution Prediction

3.2.1 Previous work

In February of 2022, data analyst David Waldron used data available at the time from twitter user @WordleStats to analyze what makes a Wordle word difficult [10]. His criteron consisted of a mix of the following factors:

- Duplicate Letters
- Similarity to other common words
- Scrabble score
- Word obscurity

Our approach to this problem was inspired by Waldron's work. We first tried to reproduce his results using our newer dataset [8], but were unable to. We hypothesize that Waldron's results were highly influenced by the spiking popularity of Wordle and the small dataset size Waldron used. Waldron's dataset was roughly 25% the size of ours and came from before New York Times acquisition of Wordle.

3.2.2 Multivariate Linear Model

We utilize a multivariate linear model, based on the following equation:

$$S_i(\mathbf{X}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + C + \varepsilon \tag{4}$$

With the variables being as described in table 1. This model depends on independence between variables. This requirement is satisfied for our variables, as we tested the variables and found no statistically significant correlation.

3.2.3 Multivariate Linear Model with Date

One flaw with the previous model is that it cannot account for the change in unobserved factors over time (e.g., composition of players who post their results), we can add a term t to the regression, which is as listed in 1. Then, we have the following regression:

$$S_i(\mathbf{X}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 t + C + \varepsilon \tag{5}$$

Again, the assumption above is justified, as time does not correlate to the words picked in common sense. However, regressing Score on date alone does not show significant results.

3.2.4 Multivariate Linear Model With Percent of Hard Mode Players

Outside the four word difficulty predictors and the time predictor, we also investigate the impact of ratio of hard mode players (r) on the score distribution:

$$S_i(\mathbf{X}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 t + \beta_7 r + C + \varepsilon \tag{6}$$

Note: r and t are have significant correlation, which violates the assumption that all variables are independent. On the other hand, this additional variable significantly improved the R^2 value.

Similar to date, the ratio of hard mode players does not exhibit a significant relationship to the score when regressed alone.

4 Data and Computation

4.1 Data

We used the New York Times dataset from January 7, 2022, to December 31, 2022, for date, word, number and distribution of reported results, and the percentage of hard mode players. The original data has a few input errors, so we checked with the past answers archived on yourdictionary.com to correct the wrong words [7]. We also removed the observation on November 30, 2022, for having an unusually small number of reports.

We retrieved the frequency of letters from Cornell University's Letter Frequency Table [4]. We retrieved the frequency of English words from Google Web Trillion Word Corpus, distributed by University of Pennsylvania [2].

Furthermore, we obtained a list of the most common letters that start or end an English word and a list of the most common two- or three-letter combinations in English from Emory University [6].

4.2 Computational Methods

To compute the average score, we assign that all reported failures a score of 7, and compute the weighted average using each score's percentage (each score will be represented as a percent, e.g. 300 means 300%).

We define the dummy variable "x_1" as if a word starts with one of the five most common starting letters ("t," "a," "i," "s," "o") and "x_2" as if the word ends with one of the five most common ending letters (“e,” “s,” “d,” “t,” “n”). The attribute “x_3” reflects whether a word contains any of the common two- or three-letter combinations on the list.

We employed Excel to clean the data and derive the desired variables. We also regressed the Fréchet and Sigmoid model by using the Solver function on Excel, which implements GRG nonlinear solving. For the multivariate linear model and the word difficulty classification model, we employed Stata to run the regressions and collect the results.

5 Results and Analysis

5.1 Future Result Model

By using the Solver function in Excel, which implements GRG nonlinear solving, we found the following coefficients for equation 1 to be:

- $\alpha = 0.99$
- $s = 64.7$
- $m = 193$

resulting in the function

$$f(x) = \frac{0.99}{64.7}\left(\frac{x-193}{64.7}\right)^{-1.99} \exp\left\{-\left(\frac{x-193}{64.7}\right)^{-0.99}\right\} \tag{7}$$

By evaluating this equation, **We predict that there will be 11,883 Wordle players on March 1st, 2023**.

Our 95% confidence interval for this range is [7,314, 16,668]. Our R^2 value is 0.978

We obtained this confidence interval by recognizing that the ratio of reported results to expected results is within 0.39 of 1 for 95% of the data points in our sample. We used the ratio of reported to expected instead of the difference between these two because we noticed that the residuals were more varied for higher values of expected number of players.

$$11,883 * 0.61 = 7,314$$

$$11,883 * 1.39 = 16,668$$

This is a large confidence interval. Even though the R^2 value for our model is very close to 1, there is still significant variation around the values predicted by the Frechet model.

5.2 Score Distribution Model

Outside the predictors used in this model. We have also linearly regressed scores on overall letter frequency, location of vowels, number of vowels, scrabble scores, but none of those have shown significant results.

Our regression results are shown in detail in the appendix and more simply in table 2, and they are mostly consistent across models but vary substantially across the scores. In general, Model 3 has a

Table 2: calculated coefficients described in equation 6

Tries	x_1	x_2	x_3	x_4	x_5	t	r	Constant
1	0.11502	0.19477	0.02027	-0.25119	0.14996	0.0025	-20.838	0.55476
2	1.7090	2.0793	1.5006	-2.1484	0.9879	0.00984	-47.730	-0.55895
3	3.3841	2.3562	2.2501	-5.3698	1.585	0.04634	-199.21	17.072
4	0.20635	-1.3374	-1.2678	-1.1637	-0.07339	0.02062	-52.201	35.305
5	-2.401	-1.6973	-2.1761	3.9281	-1.1782	-0.02673	112.09	29.444
6	-2.154	-1.1207	-0.88388	3.5691	-1.0507	-0.0337	115.24	16.484

significantly higher adjusted R^2 value than the baseline model, indicating that time and the percentage of hard mode players are meaningful predictors.

When predicting the percentage of 1 try, x_3 ("having common letter combinations") has a insignificant coefficient (p = 756) while all others are significant. This makes sense because while a word with higher frequency or common starting/ending letters is more likely to be guessed when a player has no information, the common letter combinations need the player to know at least one letter to be an useful information.

When predicting the percentage of 2,3, and 5 tries, all coefficients are statistically significant. Both the R^2 values and adjusted R^2 values are above 0.3, which fits into the idea that more attempts allow players to use their skills more, thus making the results more predictive. When predicting the percentages of 4 and 6 tries, all but x_3 and x^5 have significant coefficients. Both regressions also have lower R^2 values. For 6 tries, the high variance in coefficients may be due to the relatively small sample size; the insignificance for 4 tries is hard to explain. However, given the predictors' strong performance in other scores, we still conclude that all predictors are significant.

The word on March 1, 2023, "**EERIE**," has a common ending letter, common combination "ER," and duplicate letters, but not a common starting letter according to our criteria. It also has a logarithm of word frequency of 5.89. Therefore, calculating using equation 6, **we predict the distribution of reported results on March 1st with word "EERIE" to be (1 try: 0.4%, 2 tries: 6.2%, 3 tries: 25.8%, 4 tries: 34.7%, 5 tries: 22.2%, 6 tries: 8.9%, failure: 1.8%).** The confidence interval of this distribution is difficult to compute, and will be a focus in the future work

6 Strengths and Weaknesses

Future Result Model

Strengths:

- High R^2 value.
- Draws from previous literature. [1]

Weaknesses:

- There is high error at the peak.
- The curve is an imperfect fit (it can be seen to go slightly above the data in the middle section, and slightly below towards the end).

- This model has a very large confidence interval.

Score Distribution and Difficulty Classification Models

Strengths:

- For score distribution prediction and difficulty classification, the linear models have clear indications on the marginal impact of each factor.
- The R^2 of 0.2 - 0.4 indicates the predictive power of the models is fairly strong.

Weaknesses:

- Treating the dataset as cross-sectional and using time as a linear predictor may not perform as well as time series models.
- This model does not provide valid forecasts in the long run, as the predicted values will eventually exceed the bounds of percentage between 0 and 100.
- Lacking confidence intervals.

7 Future Steps

Future Result Model

With more time, we could analyze the data using other functions as our basis for regression, and see if any fit better than what we currently have. Specifically, the extreme value distribution, which is a more generalized form of the Fréchet distribution, is worth investigating.

We could analyze the word characteristics on previous days. If a word is challenging, it may make players turn hard mode off for the next day if that player was unable to guess the word the previous day.

Score Distribution and Difficulty Classification Models

We could benefit from trying models that fit into the range of probability, such as the logit or sigmoid model. The correlation between scores and the predictors towards more recent months also warrants more investigation, as the first few months were turbulent in terms of result count. This may disturb the correlation between variables and result in weaker observed relationships.

Another approach to the problem could be treating the dataset as a time series, or potentially looking into more positional frequencies and find a way to convert the boolean variables from 1 to continuous quantities.

We must also make a final note that a method for calculating confidence intervals is needed. This is a difficult problem to solve, but it would greatly improve our model.

8 Letter to the New York Times

To: Puzzle Editor of the New York Times

First of all, we would like to thank you for making Wordle available and for offering this invaluable data modeling exercise. We have enjoyed playing this game for a while, and it feels great for us to be able to dive deep into the data behind this game. Now, we would like to present our discoveries to you.

The number of reported results on Twitter has a high day-to-day variance; however, we find the overall pattern of it over time a very good fit to the Fréchet distribution, which has previously been observed in the popularity trend of many internet fads, such as memes [1]. The number of reports started at around 80,000 in January 2020, quickly peaked at about 360,000 in early February, and then gradually declined over time. Many distributions have patterns similar to the data provided to us, but we find Fréchet distribution to be the best fit; moreover, the Fréchet distribution is special as it can be explained as a result of "interplay processes of growing and declining attention," according to a paper written by Christian Bauckhage et. al. from University of Bonn [1]. In other words, the number of reported results of Wordle attracts attention depends on the game's perceived novelty and loses attention based on the amount of interest it has received so far.

As for the daily variance of the number of reported results, we have tested it using attributes of words such as letter frequency, word frequency, number of vowels, starting letter, ending letter, and repetitive letters; however, we do not find any significant correlation between those factors. We did the same test on the percentage of reports in hard mode, and they show no significant results either. Thus, we cannot conclude that attributes of the words play a role in those variances.

On the other hand, we find the pattern of the percentage of hard mode players fits well with a modified sigmoid function, whose value always increases over time but converges to 9.675% in the long run. This could indicate that while Wordle is gradually losing its players, it has a relatively small number of loyal players who enjoy the challenge of hard mode and are less likely to quit.

We used the word attributes above to predict the distribution of results, and find no significant impact by letter frequency, position of vowels, or number of vowels; however, having a common starting letter, ending letter, or letter combination statistically decreases the average number of attempts by 0.16, 0.13, and 0.08, respectively. Having a word be 10-times more frequent in the English language also decreases the average number by 0.09. Having duplicate letters increases the average number of attempts by 0.25.

Those attributes also help we classify the difficulty of a word. "EERIE," for example, has a common ending letter, a common combination "ER," and duplicate letters, but not a common starting letter. It also has a frequency of 772,484 according to a dictionary data from University of Pennsylvania [2]. Given these data, our model indicates that "EERIE" is one of the top 25% most difficult among appeared Wordle words.

We also add time and percentage of hard mode results in our prediction model. As expected, time has a negative impact on the average number of attempts, while the percentage of hard mode results have a positive impact. On top of that, if we remove the data from the first 70 days (when the number was drastically increasing or decreasing), the estimated impact of time during the slow declination period almost doubles compared to the estimation with all data. This supports our hypothesis above that Wordle has a steady player base after its peak, as the impact of time can be explained as the players improve their skills over time, reducing the average number of attempts. According to our theory, this effect is weakened when we use all the data because a lot of players came and left each day during the first few months, making the improved skill of the steady players a smaller factor on the results.

With all the information above, we are able to make the prediction you requested. Our models predict that there will be 11,883 reported results on March 1, 2023, and 9.646% of them will be in hard mode. Although it is extremely unlikely to be the exact number we will see, we are 95% confident that the number of reported results will fall between 7,314 and 16,668. With the word "EERIE," we also predict the score distribution to be (1 try: 0.4%, 2 tries: 6.2%, 3 tries: 25.8%, 4 tries: 34.7%, 5 tries: 22.2%, 6 tries: 8.9%, failure: 1.8%).

Regardless of how our prediction turns out, this has been a fun and invaluable experience for us. Again, we want to thank you for this opportunity to apply our knowledge to this problem.

Sincerely,
Team 2322040

References

[1] Christian Bauckhage, Kristian Kersting, and Fabian Hadiji. "Mathematical models of fads explain the temporal dynamics of internet memes". In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1 (2021), pp. 22–30. DOI: 10.1609/icwsm.v7i1.14392.

[2] Brants, Thorsten, and Alex Franz. *Web 1T 5-gram Version 1 LDC2006T13*. Sept. 2006. DOI: https://doi.org/10.35111/cqpa-a498. URL: https://catalog.ldc.upenn.edu/LDC2006T13.

[3] Stuart Coles. *An introduction to statistical modeling of extreme values*. Springer, 2011.

[4] *English Letter Frequency (based on a sample of 40,000 words)*. URL: https://pi.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html.

[5] Se Yoon Lee and Bani K. Mallick. "Bayesian hierarchical modeling: Application towards production results in the Eagle Ford Shale of south Texas". In: *Sankhya B* 84.1 (2021), pp. 1–43. DOI: 10.1007/s13571-020-00245-8.

[6] *Letter Frequencies in English*. URL: https://mathcenter.oxford.emory.edu/site/math125/englishLetterFreqs/.

[7] *List of wordle answers*. URL: https://wordfinder.yourdictionary.com/wordle/answers.

[8] Mathematical Contest in Modeling. *Problem C Data Wordle*. The Consortium for Mathematics and Its Applications. Feb. 2023. URL: https://www.contest.comap.com/undergraduate/contests/mcm/contests/2023/problems/.

[9] Pedro L. Ramos et al. "The Fréchet Distribution: Estimation and Application - An Overview". In: *Journal of Statistics and Management Systems* 23.3 (2019), pp. 549–578. DOI: 10.1080/09720510.2019.1645400.

[10] David Waldron. *What makes a Wordle word hard?* Feb. 2022. URL: https://www.waldrn.com/what-makes-a-wordle-word-hard/.

Appendix

1 Try

Standard Multivariate Linear Model (3.2.2)

Source	SS	df	MS
Model	27.3700725	5	5.47401449
Residual	191.850598	352	.545030108
Total	219.22067	357	.614063502

Number of obs	=	358
F(5, 352)	=	10.04
Prob > F	=	0.0000
R-squared	=	0.1249
Adj R-squared	=	0.1124
Root MSE	=	.73826

try	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	.137743	.0837116	1.65	0.101	-.0268947	.3023807
endcom	.2179328	.0802781	2.71	0.007	.0600477	.3758179
hascom	.025602	.0823196	0.31	0.756	-.1362981	.187502
hasdup	-.3013528	.0874956	-3.44	0.001	-.4734328	-.1292729
logfreq	.1946639	.0440677	4.42	0.000	.1079948	.2813331
_cons	-.8620626	.2919821	-2.95	0.003	-1.436311	-.2878137

Multivariate Linear Model With Date (3.2.3)

Source	SS	df	MS
Model	37.0481104	6	6.17468507
Residual	182.17256	351	.519010142
Total	219.22067	357	.614063502

Number of obs	=	358
F(6, 351)	=	11.90
Prob > F	=	0.0000
R-squared	=	0.1690
Adj R-squared	=	0.1548
Root MSE	=	.72042

try	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	.1355937	.0816904	1.66	0.098	-.0250706	.296258
endcom	.2199857	.0783399	2.81	0.005	.0659111	.3740604
hascom	.0293931	.0803354	0.37	0.715	-.1286061	.1873923
hasdup	-.28532	.0854622	-3.34	0.001	-.4534025	-.1172375
logfreq	.1821571	.0431004	4.23	0.000	.0973896	.2669245
time	-.0015942	.0003692	-4.32	0.000	-.0023202	-.0008681
_cons	-.5035103	.2967792	-1.70	0.091	-1.087199	.0801788

Multivariate Linear Model With Percent of Hard Mode Players (3.2.4)

Source	SS	df	MS
Model	48.0592022	7	6.86560031
Residual	171.161468	350	.489032766
Total	219.22067	357	.614063502

Number of obs	=	358
F(7, 350)	=	14.04
Prob > F	=	0.0000
R-squared	=	0.2192
Adj R-squared	=	0.2036
Root MSE	=	.69931

try	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	.1150203	.0794146	1.45	0.148	-.0411696	.2712101
endcom	.1947741	.0762292	2.56	0.011	.0448491	.3446991
hascom	.0202749	.0780045	0.26	0.795	-.1331416	.1736914
hasdup	-.2511959	.0832686	-3.02	0.003	-.4149656	-.0874262
logfreq	.1499631	.0423837	3.54	0.000	.0666043	.2333219
time	.002506	.0009354	2.68	0.008	.0006662	.0043458
percenthardmode	-20.83811	4.391493	-4.75	0.000	-29.47515	-12.20108
_cons	.5547661	.3643221	1.52	0.129	-.1617699	1.271302

2 Tries

Standard Multivariate Linear Model (3.2.2)

Source	SS	df	MS
Model	1940.76306	5	388.152611
Residual	4008.47717	352	11.3877192
Total	5949.24022	357	16.6645384

Number of obs	=	358
F(5, 352)	=	34.09
Prob > F	=	0.0000
R-squared	=	0.3262
Adj R-squared	=	0.3166
Root MSE	=	3.3746

tries	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	1.75559	.3826428	4.59	0.000	1.003036	2.508144
endcom	2.137694	.3669488	5.83	0.000	1.416006	2.859382
hascom	1.522584	.3762801	4.05	0.000	.7825436	2.262623
hasdup	-2.222076	.3999396	-5.56	0.000	-3.008648	-1.435504
logfreq	1.058117	.2014321	5.25	0.000	.6619551	1.454279
_cons	-2.88116	1.334641	-2.16	0.032	-5.506033	-.2562873

Multivariate Linear Model With Date (3.2.3)

Source	SS	df	MS
Model	1941.54354	6	323.59059
Residual	4007.69668	351	11.4179393
Total	5949.24022	357	16.6645384

Number of obs	=	358
F(6, 351)	=	28.34
Prob > F	=	0.0000
R-squared	=	0.3264
Adj R-squared	=	0.3148
Root MSE	=	3.379

tries	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	1.7562	.3831573	4.58	0.000	1.002627	2.509773
endcom	2.137111	.3674421	5.82	0.000	1.414446	2.859776
hascom	1.521507	.3768015	4.04	0.000	.7804342	2.26258
hasdup	-2.226629	.4008484	-5.55	0.000	-3.014996	-1.438262
logfreq	1.061669	.2021562	5.25	0.000	.6640789	1.459258
time	.0004527	.0017315	0.26	0.794	-.0029528	.0038582
_cons	-2.982982	1.392	-2.14	0.033	-5.720693	-.2452717

Multivariate Linear Model With Percent of Hard Mode Players (3.2.4)

Source	SS	df	MS
Model	1999.31424	7	285.616319
Residual	3949.92599	350	11.2855028
Total	5949.24022	357	16.6645384

Number of obs	=	358
F(7, 350)	=	25.31
Prob > F	=	0.0000
R-squared	=	0.3361
Adj R-squared	=	0.3228
Root MSE	=	3.3594

tries	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	1.709076	.3814977	4.48	0.000	.9587596	2.459392
endcom	2.079363	.3661955	5.68	0.000	1.359142	2.799583
hascom	1.500621	.3747236	4.00	0.000	.7636279	2.237615
hasdup	-2.148466	.4000115	-5.37	0.000	-2.935195	-1.361738
logfreq	.987927	.2036059	4.85	0.000	.587482	1.388372
time	.0098443	.0044937	2.19	0.029	.0010062	.0186824
percenthardmode	-47.73061	21.09617	-2.26	0.024	-89.22183	-6.239399
_cons	-.5589539	1.750157	-0.32	0.750	-4.001101	2.883194

3 Tries

Standard Multivariate Linear Model (3.2.2)

Source	SS	df	MS			
				Number of obs	=	358
				F(5, 352)	=	30.99
Model	6608.83673	5	1321.76735	Prob > F	=	0.0000
Residual	15013.3672	352	42.6516113	R-squared	=	0.3057
				Adj R-squared	=	0.2958
Total	21622.2039	357	60.5663975	Root MSE	=	6.5308

v13	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	3.571217	.7405304	4.82	0.000	2.114797	5.027638
endcom	2.606519	.7101577	3.67	0.000	1.209833	4.003205
hascom	2.354291	.7282166	3.23	0.001	.9220886	3.786494
hasdup	-5.624222	.7740051	-7.27	0.000	-7.146478	-4.101966
logfreq	1.836791	.3898325	4.71	0.000	1.070097	2.603485
_cons	8.561768	2.582937	3.31	0.001	3.481839	13.6417

Multivariate Linear Model With Date (3.2.3)

Source	SS	df	MS			
				Number of obs	=	358
				F(6, 351)	=	26.86
Model	6803.08533	6	1133.84756	Prob > F	=	0.0000
Residual	14819.1186	351	42.219711	R-squared	=	0.3146
				Adj R-squared	=	0.3029
Total	21622.2039	357	60.5663975	Root MSE	=	6.4977

v13	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	3.580846	.7367852	4.86	0.000	2.131777	5.029915
endcom	2.597322	.7065659	3.68	0.000	1.207686	3.986957
hascom	2.337307	.7245635	3.23	0.001	.9122747	3.762339
hasdup	-5.69605	.770804	-7.39	0.000	-7.212026	-4.180075
logfreq	1.892822	.3887324	4.87	0.000	1.128285	2.65736
time	.0071419	.0033296	2.14	0.033	.0005934	.0136905
_cons	6.955427	2.676721	2.60	0.010	1.690997	12.21986

Multivariate Linear Model With Percent of Hard Mode Players (3.2.4)

Source	SS	df	MS			
				Number of obs	=	358
				F(7, 350)	=	28.27
Model	7809.48373	7	1115.64053	Prob > F	=	0.0000
Residual	13812.7202	350	39.4649148	R-squared	=	0.3612
				Adj R-squared	=	0.3484
Total	21622.2039	357	60.5663975	Root MSE	=	6.2821

v13	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	3.384159	.7134065	4.74	0.000	1.981056	4.787262
endcom	2.356291	.6847912	3.44	0.001	1.009468	3.703115
hascom	2.250134	.7007389	3.21	0.001	.8719454	3.628323
hasdup	-5.369815	.7480276	-7.18	0.000	-6.841009	-3.89862
logfreq	1.58504	.3807462	4.16	0.000	.8362013	2.333878
time	.0463404	.0084033	5.51	0.000	.029813	.0628678
percenthardmode	-199.2178	39.45016	-5.05	0.000	-276.807	-121.6286
_cons	17.07283	3.27282	5.22	0.000	10.63596	23.50969

4 Tries

Standard Multivariate Linear Model (3.2.2)

Source	SS	df	MS
Model	314.009359	5	62.8018718
Residual	9948.10237	352	28.2616545
Total	10262.1117	357	28.745411

Number of obs	=	358
F(5, 352)	=	2.22
Prob > F	=	0.0517
R-squared	=	0.0306
Adj R-squared	=	0.0168
Root MSE	=	5.3162

v14	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	.2439382	.602801	0.40	0.686	-.9416063	1.429483
endcom	-1.260953	.5780772	-2.18	0.030	-2.397872	-.1240331
hascom	-1.220372	.5927774	-2.06	0.040	-2.386203	-.0545409
hasdup	-1.145146	.6300498	-1.82	0.070	-2.384281	.0939895
logfreq	-.0739706	.3173286	-0.23	0.816	-.698069	.5501278
_cons	34.98338	2.102543	16.64	0.000	30.84826	39.11851

Multivariate Linear Model With Date (3.2.3)

Source	SS	df	MS
Model	722.184723	6	120.364121
Residual	9539.92701	351	27.1792792
Total	10262.1117	357	28.745411

Number of obs	=	358
F(6, 351)	=	4.43
Prob > F	=	0.0002
R-squared	=	0.0704
Adj R-squared	=	0.0545
Root MSE	=	5.2134

v14	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	.2578962	.5911561	0.44	0.663	-.9047576	1.42055
endcom	-1.274285	.5669099	-2.25	0.025	-2.389253	-.1593176
hascom	-1.244993	.5813501	-2.14	0.033	-2.38836	-.1016248
hasdup	-1.249267	.618451	-2.02	0.044	-2.465603	-.0329315
logfreq	.0072522	.3118977	0.02	0.981	-.6061712	.6206755
time	.0103529	.0026715	3.88	0.000	.0050987	.015607
_cons	32.65485	2.147655	15.20	0.000	28.43096	36.87874

Multivariate Linear Model With Percent of Hard Mode Players (3.2.4)

Source	SS	df	MS
Model	791.283947	7	113.040564
Residual	9470.82779	350	27.059508
Total	10262.1117	357	28.745411

Number of obs	=	358
F(7, 350)	=	4.18
Prob > F	=	0.0002
R-squared	=	0.0771
Adj R-squared	=	0.0586
Root MSE	=	5.2019

v14	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	.2063581	.5907332	0.35	0.727	-.9554754	1.368192
endcom	-1.337442	.5670384	-2.36	0.019	-2.452674	-.2222109
hascom	-1.267834	.5802439	-2.19	0.030	-2.409038	-.1266311
hasdup	-1.163784	.6194011	-1.88	0.061	-2.382	.0544329
logfreq	-.0733963	.3152753	-0.23	0.816	-.6934687	.5466762
time	.0206241	.0069584	2.96	0.003	.0069386	.0343095
percenthardmode	-52.20112	32.66654	-1.60	0.111	-116.4485	12.04629
_cons	35.30592	2.710045	13.03	0.000	29.9759	40.63594

5 Tries

Standard Multivariate Linear Model (3.2.2)

Source	SS	df	MS
Model	3634.57407	5	726.914815
Residual	9004.78068	352	25.5817633
Total	12639.3547	357	35.404355

Number of obs	=	358
F(5, 352)	=	28.42
Prob > F	=	0.0000
R-squared	=	0.2876
Adj R-squared	=	0.2774
Root MSE	=	5.0578

v15	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	-2.506233	.5735092	-4.37	0.000	-3.634169	-1.378297
endcom	-1.839	.5499868	-3.34	0.001	-2.920674	-.7573268
hascom	-2.236269	.5639727	-3.97	0.000	-3.345449	-1.127089
hasdup	4.064731	.5994339	6.78	0.000	2.885809	5.243654
logfreq	-1.314778	.3019087	-4.35	0.000	-1.90855	-.7210067
_cons	34.08517	2.000374	17.04	0.000	30.15098	38.01936

Multivariate Linear Model With Date (3.2.3)

Source	SS	df	MS
Model	3717.81931	6	619.636552
Residual	8921.53544	351	25.4174799
Total	12639.3547	357	35.404355

Number of obs	=	358
F(6, 351)	=	24.38
Prob > F	=	0.0000
R-squared	=	0.2941
Adj R-squared	=	0.2821
Root MSE	=	5.0416

v15	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	-2.512536	.5716754	-4.40	0.000	-3.636876	-1.388196
endcom	-1.832979	.5482281	-3.34	0.001	-2.911205	-.7547542
hascom	-2.22515	.5621925	-3.96	0.000	-3.33084	-1.11946
hasdup	4.111753	.5980707	6.88	0.000	2.9355	5.288006
logfreq	-1.351459	.3016195	-4.48	0.000	-1.944668	-.75825
time	-.0046754	.0025835	-1.81	0.071	-.0097564	.0004056
_cons	35.13674	2.076882	16.92	0.000	31.05205	39.22144

Multivariate Linear Model With Percent of Hard Mode Players (3.2.4)

Source	SS	df	MS
Model	4036.4231	7	576.631871
Residual	8602.93165	350	24.5798047
Total	12639.3547	357	35.404355

Number of obs	=	358
F(7, 350)	=	23.46
Prob > F	=	0.0000
R-squared	=	0.3194
Adj R-squared	=	0.3057
Root MSE	=	4.9578

v15	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	-2.40187	.5630159	-4.27	0.000	-3.50919	-1.29455
endcom	-1.697363	.5404329	-3.14	0.002	-2.760268	-.6344586
hascom	-2.176102	.5530187	-3.93	0.000	-3.26376	-1.088444
hasdup	3.928195	.5903387	6.65	0.000	2.767138	5.089253
logfreq	-1.178284	.3004825	-3.92	0.000	-1.769263	-.5873055
time	-.0267305	.0066319	-4.03	0.000	-.0397738	-.0136872
percenthardmode	112.0905	31.13382	3.60	0.000	50.85756	173.3234
_cons	29.44416	2.582889	11.40	0.000	24.36422	34.5241

6 Tries

Standard Multivariate Linear Model (3.2.2)

Source	SS	df	MS
Model	2348.15745	5	469.631491
Residual	11443.8649	352	32.5109798
Total	13792.0223	357	38.6331158

Number of obs	=	358
F(5, 352)	=	14.45
Prob > F	=	0.0000
R-squared	=	0.1703
Adj R-squared	=	0.1585
Root MSE	=	5.7018

v16	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	-2.253021	.6465323	-3.48	0.001	-3.524572	-.9814686
endcom	-1.274401	.6200148	-2.06	0.041	-2.4938	-.0550015
hascom	-.9605455	.6357815	-1.51	0.132	-2.210954	.2898627
hasdup	3.646892	.6757579	5.40	0.000	2.317861	4.975923
logfreq	-1.142286	.3403497	-3.36	0.001	-1.81166	-.4729108
_cons	19.85662	2.255075	8.81	0.000	15.4215	24.29173

Multivariate Linear Model With Date (3.2.3)

Source	SS	df	MS
Model	2811.47985	6	468.579974
Residual	10980.5425	351	31.2835969
Total	13792.0223	357	38.6331158

Number of obs	=	358
F(6, 351)	=	14.98
Prob > F	=	0.0000
R-squared	=	0.2038
Adj R-squared	=	0.1902
Root MSE	=	5.5932

v16	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	-2.267892	.6342224	-3.58	0.000	-3.515246	-1.020537
endcom	-1.260196	.6082098	-2.07	0.039	-2.45639	-.0640026
hascom	-.9343142	.623702	-1.50	0.135	-2.160977	.2923489
hasdup	3.757824	.6635057	5.66	0.000	2.452878	5.062771
logfreq	-1.228821	.3346197	-3.67	0.000	-1.886933	-.5707096
time	-.0110301	.0028661	-3.85	0.000	-.016667	-.0053931
_cons	22.33747	2.304113	9.69	0.000	17.80586	26.86907

Multivariate Linear Model With Percent of Hard Mode Players (3.2.4)

Source	SS	df	MS
Model	3148.27012	7	449.752875
Residual	10643.7522	350	30.4107206
Total	13792.0223	357	38.6331158

Number of obs	=	358
F(7, 350)	=	14.79
Prob > F	=	0.0000
R-squared	=	0.2283
Adj R-squared	=	0.2128
Root MSE	=	5.5146

v16	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
startcom	-2.15411	.6262458	-3.44	0.001	-3.385788	-.9224318
endcom	-1.120763	.6011266	-1.86	0.063	-2.303038	.0615114
hascom	-.8838858	.6151259	-1.44	0.152	-2.093694	.3259222
hasdup	3.569101	.6566371	5.44	0.000	2.27765	4.860552
logfreq	-1.050773	.3342284	-3.14	0.002	-1.708121	-.3934239
time	-.033706	.0073767	-4.57	0.000	-.0482141	-.0191978
percenthardmode	115.2452	34.63033	3.33	0.001	47.13553	183.355
_cons	16.48467	2.872962	5.74	0.000	10.83423	22.13511