Problem Chosen	2023 MCM/ICM Summary Sheet	Team Control Number
C		2322823

Summary

Wordle is a popular puzzle currently offered daily by the ***New York Times***. Players try to solve the puzzle by guessing a five-letter word in six tries or less, receiving feedback with every guess. In this work, we focus on predicting number of reported results and distribution of them, finding out attributes that affect the variation, and constructing a suitable standard of classifications.

First, we analyze the variance of daily number of reported results. According to the study of Time series, we establish the **ARIMA Model**. Based on this model, it is estimated that on March 1, for the word "**EERIE**", the number of reported is 16726. What's more, an analysis of the prediction, as well as attributes that affect percentage of attempts are also presented.

In addition to predict the number of reports, we continue to focus on key features of a word that may affect the distribution, and hence to make prediction of percentages of tries. Comparing all the words given, we conclude features like vowel and consonant, repeated letters, and letter frequency. By applying these into the **Random Forest Regression**, we develop and train a predicting model to compute distributions for word "**EERIE**". The percentages are: 0, 4, 20, 36, 28, 11, 2, respectively for every attempt.

Furthermore, to classify all words by difficulty, we apply the **Cluster K-Means method** to features collected in previous analysis and then make a detailed introduction. Finally, we discuss some interesting facts when observing the given data set. Although there is uncertainty in our model, the model established in this paper can also be applied to solve problems in studying the Wordle game and make prediction.

Keywords: Wordle; five-letter word; ARIMA Model; Random Forest Regression; Cluster K-Means method

Table of Contents

1. Introduction

1.1. problem restatement

Wordle, a worldwide popular puzzle provided by the ***New York Times***, has attracted eyeballs around the world. Players are expected to find a certain five-letter word in at most six attempts, which they are given feedbacks after every try. While thousands of people getting immersed in this fantastic game and reporting their results via ***Twitter***, it is interesting to analyze how different words affects results. Now, based on previous daily data, assuming players well follow the rule of Wordle, we need to establish mathematical models to solve the following problems:

1. Explain the variation of number of daily reported results and make a prediction of number on March 1, 2023. Investigate features of a word which potentially influence the distribution of results in ***Hard Mode***.
2. Predict relevant percentages of numbers of attempts (1, 2, 3, 4, 5, 6, X) for future, and apply it for "EERIE" in March 1,2023. Analyze uncertainty and confidence level for model.
3. Categorize solution words by difficulty using model and identify factors that attributes to difficulty, apply it for "EERIE". Discuss the accuracy.

1.2. problem analysis

Since the observed data are reported randomly and separately, we ought to make a series of assumptions and justifications to guarantee the statistical meaning of results (will be further explained in section 2), excluding insignificant situations which cannot reflect the fact. For example, we assume that a player cannot play many times in order to get a "confusing" higher score. Specifically, we apply the relevant knowledge about mathematics, statistics, communication, linguistics, and social study to optimize our models.

Firstly, when explaining the variating relation between date and total reported number, it is reasonable to associate it with the time series model. After checking the stationarity and white noise. We establish the ARIMA model for predicting future numbers. In order to ensure the reality, the trend of change should between a minimum and maximum value, and the number, should be non-negative.

The problem becomes more complicated when we try to investigate potential features that affect percentages of number of tries and predict future distributions of attempts. We assume the characteristics of different words are remarkable so that every attempt is bound to offer players different comentropy(information). After finding these features, different regression model, especially the Random Forest Regressor model can effectively identify characteristic values and then compute a reasonable distribution of attempts.

Finally, in order to classify word by their difficulty, we introduce the Cluster K-Means method. Synthesizing previous results, we can group all words by a series of process, and frequency of words' appearance seems to be a significant factor while testing our model, as well as the distribution of tries.

2. Assumptions and Justifications

By adequate analysis of the problem, to optimize our models to solve the four questions above, we make the following well-justified assumptions:

1. Players obey the rule of Wordle (only one try every day), such that every reported result truly reflects the answer situation for a player.
2. There is almost no difference for between weekend and workdays, since there is no remarkable difference of number of players between them.
3. Since all prediction and analysis work is strictly according to the historical data there exist no factors that may affect these recorded data.
4. There is no change of Wordle's rule, such that all data is on a relatively steady process.
5. All solution words are randomly chosen from the list which is not published, it is a fair game for everyone.

3. Question One

3.1. Model Preparation

To predict the number of reported results on March 1, 2023, we develop a model based on the ARIMA (Autoregressive Integrated Moving Average) model, since the results are published varies days, which fits the time series.

3.2. Introduction of the ARIMA model

According to Hyndman, R.J. and Athanasopoulos, G (2014), exponential smoothing and ARIMA models are widely used for time series forecasting, providing complementary approaches. This approach uses techniques of stationarity and differencing to indicate the autocorrelations among data given. Specifically, it is combined by two separated models: AR and MA.

ARIMA is the combination of the AR model and MA model together with a difference d. The full model can be written as

$$y_t' = c + \phi_1 y_{t-1}' + \cdots + \phi_p y_{t-p}' + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

which is denoted as ARIMA (p, d, q), where d is the difference between consecutive observations, p is the number of previous values (lags) that affect the current value, q is the number of previous error terms (lags of the forecast errors) that affect the current value.

3.3. Data set

The data we used are daily reported results for January 7, 2022 through December 31, 2022, 359 data sets totally.

Date	Contest number	Word	Number of reported results	Number in hard mode
2022/3/18	272	saute	179830	9304
2022/9/19	457	trice	35050	3430
2022/2/4	230	pleat	359679	14813
2022/9/29	467	scald	30477	2829
2022/8/27	434	ruder	31241	2784
2022/3/1	255	rupee	240137	10577
2022/11/15	514	snarl	27475	2650
2022/6/29	375	gawky	45645	3957
2022/11/18	517	glyph	29208	2899
2022/10/26	494	flout	30063	2904
2022/7/2	378	egret	41765	3515
2022/7/25	401	elope	39228	3339
2022/12/2	531	chafe	24646	2343
2022/10/4	472	bough	32014	3060
2022/3/22	276	slosh	160161	8807
2022/2/19	245	swill	282327	11241
2022/4/28	313	zesty	88974	6315
2022/8/2	409	coyly	34909	3380
2022/8/16	423	gruel	35105	3087
2022/7/21	397	aphid	39086	3367

Figure 3-1: Examples of data sets

3.4. Prediction

3.4.1. Determination of parameters

First, we need to check the steadiness of the data. The data is called steady if it has no obvious trend. However, we can see clearly that the data set has a peak and a trend of decline later. Thus, we need to make the data steady by process of difference. The following plot shows the data after two times differences, which has no obvious trend and is generally symmetric about the 0.

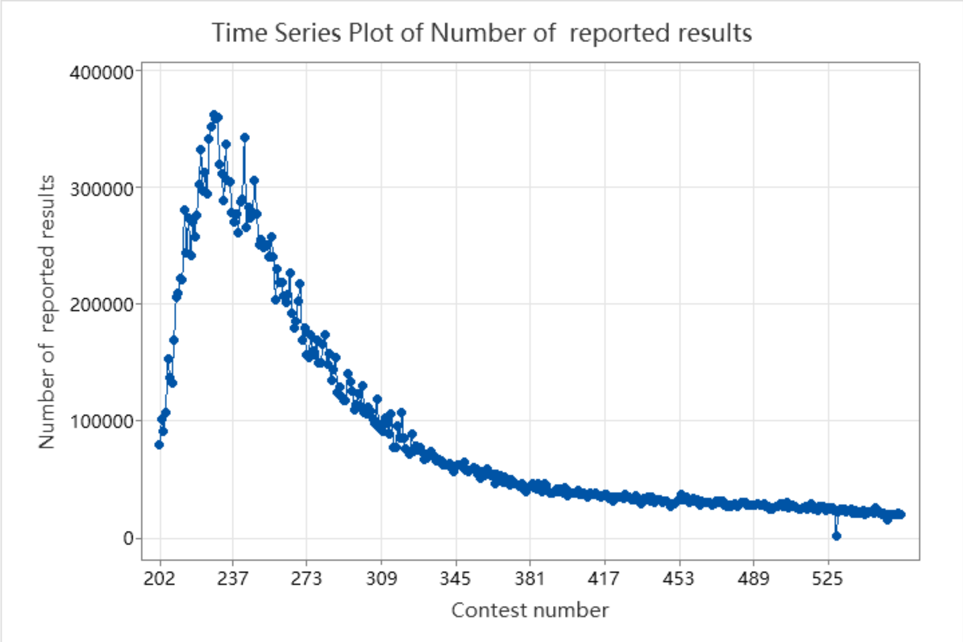

Figure 3-2-1: Time series plot of original data

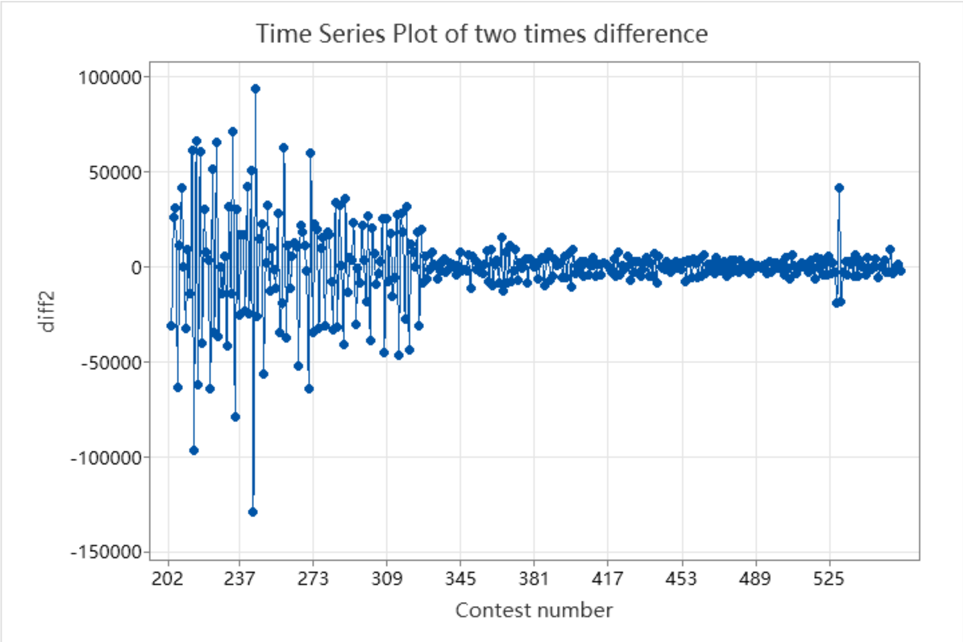

Figure 3-2-2: Time series plot of data after two times of differences

By observing the ACF and PACF plots, additionally with the auto-choice of program, we finally determine (p, d, q) as ***(5, 2, 3)***, which has a perfectly acceptable p-value, i.e., less or equal to 0.05.

Meanwhile, there is few evidence to show that there exists correlation between residuals and the observations, since most of the autocorrelations lie in the 95% confident interval although a few of them go out. However, this can happen at higher order lags that are not seasonal lags.

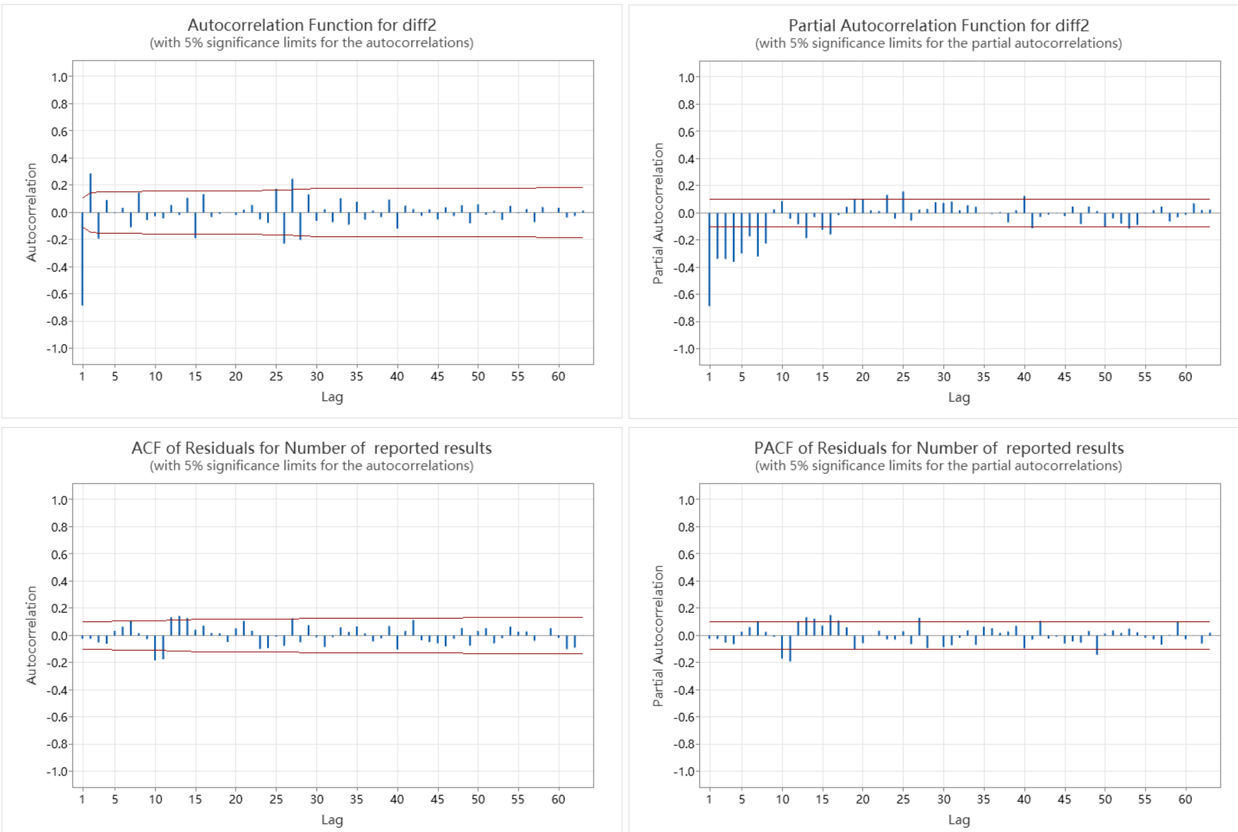

Figure 3-3: ACF/PACF plots of two times difference and residuals

3.4.2. Result of prediction

After 60 steps of predictions, we finally reach March 1, 2023, with the number of reported results of ***16726***, which lie in the interval determined as (0, 433648).

Forecasts from period 359

Period	Forecast	95% Limits Lower	Upper
360	21309. 1	−959	43577
361	19281. 0	−6748	45310
362	20845. 2	−9530	51220
363	20446. 1	−12282	53174
364	19801. 5	−17019	56622
365	20598. 2	−20958	62154
366	19985. 4	−26254	66224
367	19690. 6	−30432	69813
368	20566. 1	−34529	75661
369	19502. 2	−39991	78996
370	19864. 6	−44578	84307
371	20153. 8	−49452	89759
372	19263. 8	−55131	93659
373	19924. 2	−59766	99614
374	19729. 7	−65371	104830
375	19146. 1	−71128	109421
376	19878. 9	−76144	115902
377	19286. 3	−82291	120864
378	19159. 9	−88033	126353
379	19676. 3	−93557	132910
380	18935. 4	−100087	137958
381	19196. 1	−105861	144253
382	19360. 0	−111966	150686
383	18704. 5	−118680	156089
384	19189. 2	−124614	162992
385	18984. 7	−131276	169245
386	18590. 7	−138047	175228
387	19086. 7	−144286	182459
388	18622. 5	−151397	188642
389	18549. 8	−158185	195284
390	18879. 1	−164847	202605
391	18325. 9	−172258	208910
392	18523. 8	−179108	216156
393	18590. 5	−186243	223424
394	18119. 0	−193816	230054
395	18459. 3	−200829	237747
396	18267. 0	−208410	244944
397	17992. 9	−216056	252042
398	18324. 1	−223342	259990
399	17957. 8	−231285	267201
400	17914. 8	−238981	274811
401	18114. 0	−246625	282853
402	17699. 4	−254818	290217
403	17842. 4	−262604	298289
404	17849. 1	−270637	306335
405	17505. 9	−278981	313993
406	17738. 4	−286934	322411
407	17563. 4	−295331	330458
408	17368. 5	−303765	338502
409	17582. 1	−311966	347131
410	17291. 9	−320662	355246
411	17262. 2	−329173	363698
412	17372. 8	−337682	372428
413	17059. 0	−346594	380712
414	17156. 3	−355217	389529
415	17126. 7	−364051	398305
416	16872. 9	−373108	406853
417	17025. 2	−381899	415950
418	16868. 7	−391038	424775
419	16725. 5	−400197	433648

Table 3-1: Predictions until March 1, 2023

3.4.3. Analysis of the result

According to the model we constructed, the number of reported results varies mainly along the time, which corresponds with the characteristic of a time series model. In the given data, one of the most possible reasons for the variation of the number can be the popularity of this game. Thus, we can see a peak in the middle of February, since there always exists a lag when the game is spreading out on the internet. After that, the trend of number appears as a continuous decrease. Our prediction shows a decline with waves, which can be considered as an acceptable result.

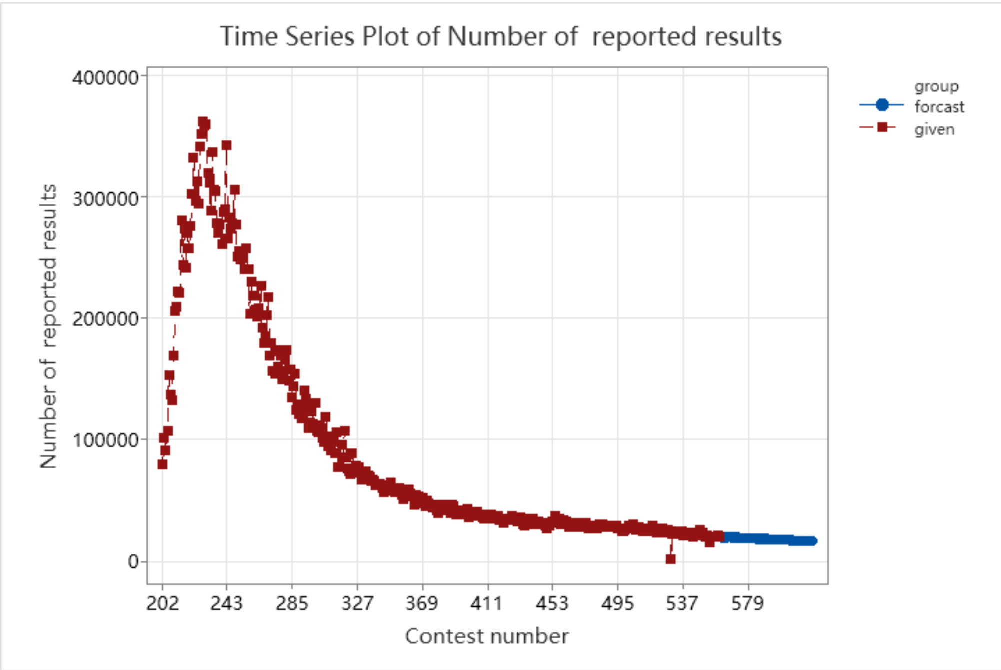

Figure 3-4: Time series plot with predictions

3.5. Attributes that affect percentage

Since we only have the data of the total reported results without the specific percent of hard mode. However, the data of hard mode can be considered to have a high correlation with the number of reported results. Therefore, we can use the percent of total result to study the percent of hard mode.

Repeated letters

We divided all the words into two clusters by whether containing repeated letters, and denote 'yes' as 0, 'no' as 1. For example, the word 'taunt' is denoted as 0 since it has a repeated letter 't'. By observing the plot of the two clusters against each percentage of tries, we find that it is generally easier when the word has repeated letters, since its percentage before 4 tries is higher while after 4 tries is lower.

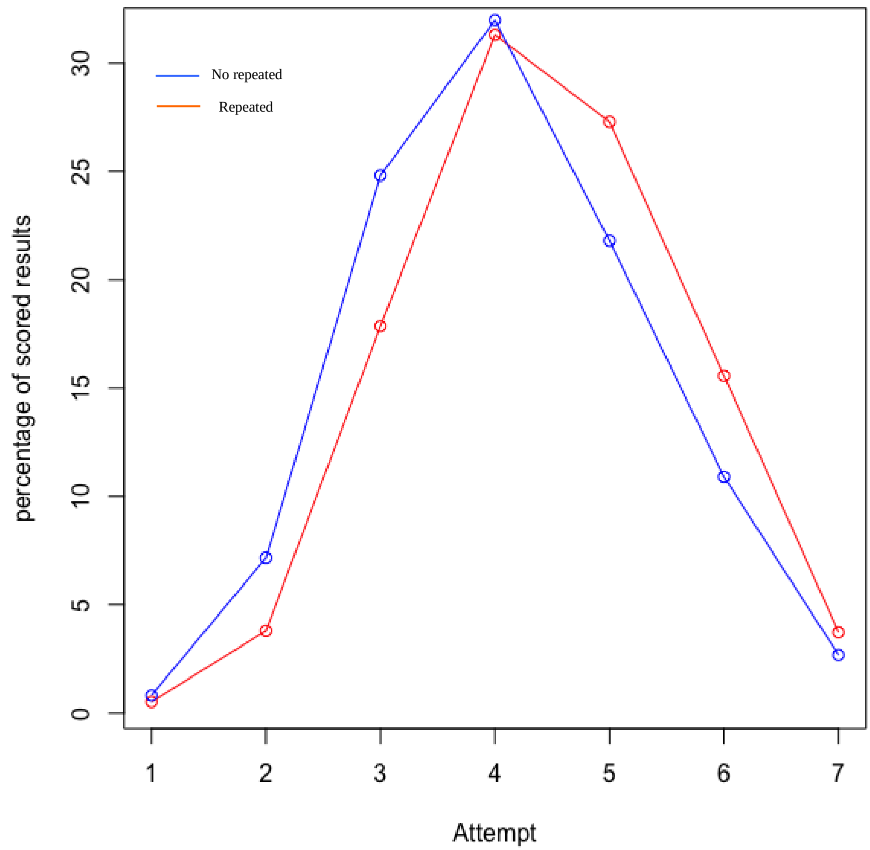

Figure 3-5: Distribution of percent of words that have/don't have repeated letters

Probability of existing letters

Through observing the given data set, the probability of every letter appearing in the data set varies greatly. Therefore, we categorized the words in each letter from a to z. Taking ‘a’ as an example, we found that there are 148 words containing ‘a’; and for ‘z’, we only found 5 words containing it. Then, we divided 26 letters into four groups by comparing the quantity of different letters and constructed the figure of the probability of existing letters against each percent. The figure below demonstrated that the higher probability of existing letters means people can guess the correct words in fewer attempts.

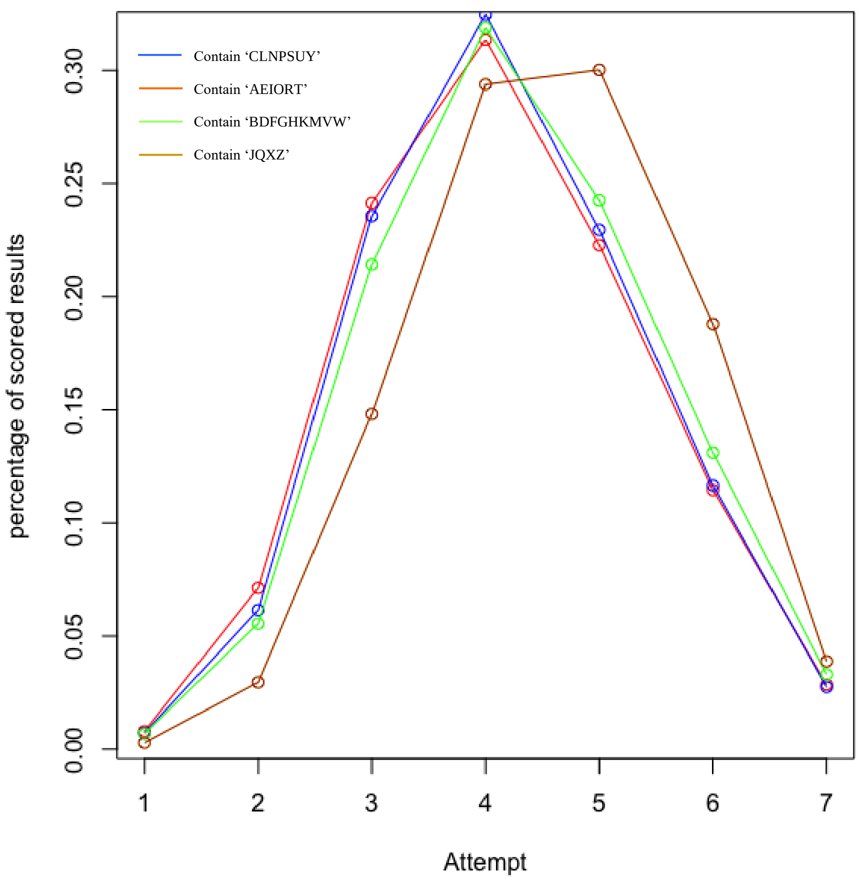

Figure 3-6: Distribution of percent of words that contains different letters

Frequency

We collected data of frequency of the words according to the Corpus of Contemporary American English (COCA) and constructed the figure of the frequency against each percent, finding that higher frequency often means the difficulty is lower.

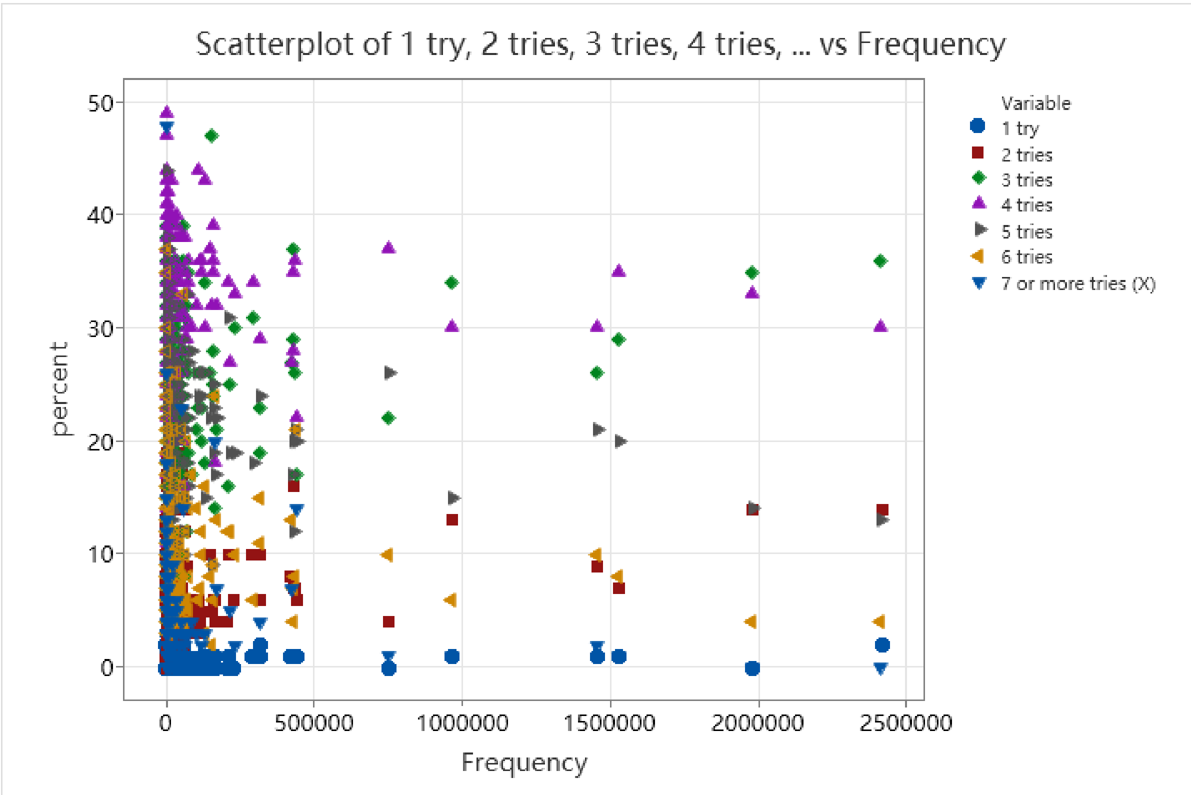

Figure 3-7: Scatterplot of number of tries versus frequency of words

4. Question Two

4.1. Model preparation

In this section, our model is established based on the Random Forest Regression model. Random Forest Model is applied to predict results by selecting strong features that are strong predictors for data sets. Hence, it guarantees high accuracy while fitting the model and is suitable for our prediction, since we can derive a series of characteristics from a word. The model is developed on decision trees, a popular statistical method. Containing a series of decisions, decision trees are always used to complete category and regression tasks. Tree learning "comes closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie et al (2008). " However, each decision tree has a high variance, but when we combine all decision trees in parallel, the variance of the integration result is low and the output does not depend on one decision tree, but on multiple decision trees, since each decision tree is perfectly trained on a particular sample data. The lower variance effectively reduces errors for each decision, in particularly ensures a higher accuracy when having various predictors.

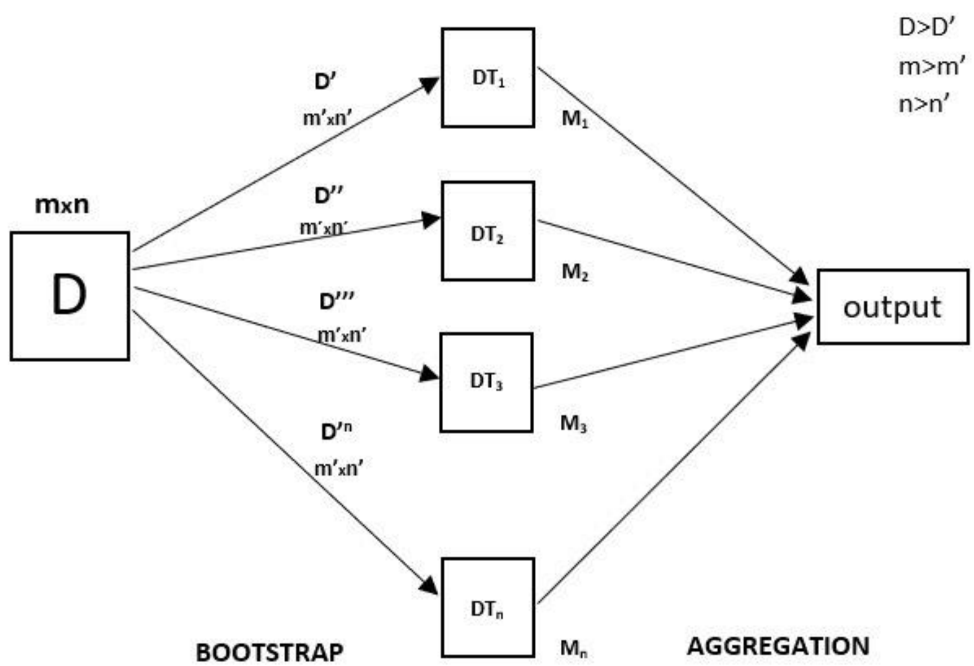

Figure 4-1: Process of Random Forest

The procedure Figure 4-1 above visualizes specific steps in the whole model.

In the case of the regression problem, the final output is the average of all the outputs. This part is called bagging.

Specifically, given a training set $X_b = x_{b_1}, \ldots x_{b_n}$ with responses $Y_b = y_{b_1}, \ldots, y_{b_n}$, bagging repeatedly (B times) selects a random sample with features of the training set and fits trees to these samples, for every index $b = 1, \ldots, B$. After bagging, the final model is the average of all individual regression trees:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

The uncertainty of the prediction can be presented as the standard deviation:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}$$

The training and test error tend to level off after some numbers of trees have been fit.

4.2. Data set

More complicated situations appear when predicting percentage distributions of a solution word. We ought to make a brief but clear analysis of the process of Wordle: As mentioned before, we assume the characteristics of different words are remarkable so that every attempt is bound to offer players different comentropy(information). Therefore, after gathering more information, players in later tries will always have higher probability to get the answer. Hence, we conclude that difficulty of a word affects distributions of attempts, because difficult words require more information to be guessed, and difficulty is always represented by a series of features of a word, as mentioned before.

4.2.1. "Deconstruction" of words

When analyzing solutions words, there should be some noteworthy features. The simplest but most detailed characteristics are letters themselves. Different letters have different frequency to appear in a randomly selected five-letter word. Also, focusing on a letter, different positions may affect difficulty of a word. Besides, there are other remarkable features, like vowel and consonant, repeated letters. By deconstructing words, we derive a series of characteristics which can be applied as the training data set in Random Forest, shown in Table 4-1:

Date	Contest number	Word	Number of reported results	Number in hard mode	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	...	w4	w5	Vowel_fre	Consonant_fre
2022-01-07	202.0	slump	80630.0	1362.0	1.0	3.0	23.0	39.0	24.0	9.0	...	13	16.0	1	4
2022-01-08	203.0	crank	101503.0	1763.0	1.0	5.0	23.0	31.0	24.0	14.0	...	14	11.0	1	4
2022-01-09	204.0	gorge	91477.0	1913.0	1.0	3.0	13.0	27.0	30.0	22.0	...	7	5.0	2	3
2022-01-10	205.0	query	107134.0	2242.0	1.0	4.0	16.0	30.0	30.0	17.0	...	18	25.0	2	3
2022-01-11	206.0	drink	153880.0	3017.0	1.0	9.0	35.0	34.0	16.0	5.0	...	14	11.0	1	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2022-12-28	557.0	impel	20160.0	1937.0	0.0	3.0	21.0	40.0	25.0	9.0	...	5	12.0	2	3
2022-12-29	558.0	havoc	20001.0	1919.0	0.0	2.0	16.0	38.0	30.0	12.0	...	15	3.0	2	3
2022-12-30	559.0	molar	21204.0	1973.0	0.0	4.0	21.0	38.0	26.0	9.0	...	1	18.0	2	3
2022-12-31	560.0	manly	20380.0	1899.0	0.0	2.0	17.0	37.0	29.0	12.0	...	12	25.0	1	4
2023-03-01	NaN	eerie	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	9	5.0	4	1

Table 4-1: Data set of percentages and features for training in our model

(We number each letter by their order in alphabet)

4.3. Analysis of result

After identifying the feature and percentage data set, we apply the model to the word "eerie" in March 1, 2023.

In order to eliminate the occasionality, we compute 100 times by our model and average the sum of result, which is demonstrated below, in Table 4-2:

Number of tries	**1** try	**2** tries	**3** tries	**4** tries	**5** tries	**6** tries	**X**
Predicted percentages	0.04	3.73	20.21	35.69	27.76	11.23	2.1

The rounding results are: ***0, 4, 20, 36, 28, 11, 2***

Table 4-2: Predictions

4.4. Model evaluation

Comparing with other regression models, it has several advantages. By applying the decision trees, the model corresponds word features highly to the prediction and this promotes the accuracy of prediction. In addition, the Random Forest Model effectively eliminates the variance and simplifies the complex circumstance brought by different characteristics of each word. However, we still should consider the uncertainty of our model due to the small data size. Since we only have data set of 300-400, this is no so sufficient that it remains uncertainty. In conclusion, it is a convincing model.

5. Question Three

5.1. Model Preparation

Our model is basically developed from the Cluster K-Means Analysis, which enable us to classify words into a given number of clusters. Cluster K-Means is used to group observations into clusters that share common characteristics. This method is appropriate when you have sufficient information to make good starting cluster designations for the clusters. Cluster K-means uses a non-hierarchical procedure to group observations. Therefore, in the clustering process, two observations might be split into separate clusters after they are joined together.

The following figure shows the process of K-means clustering.

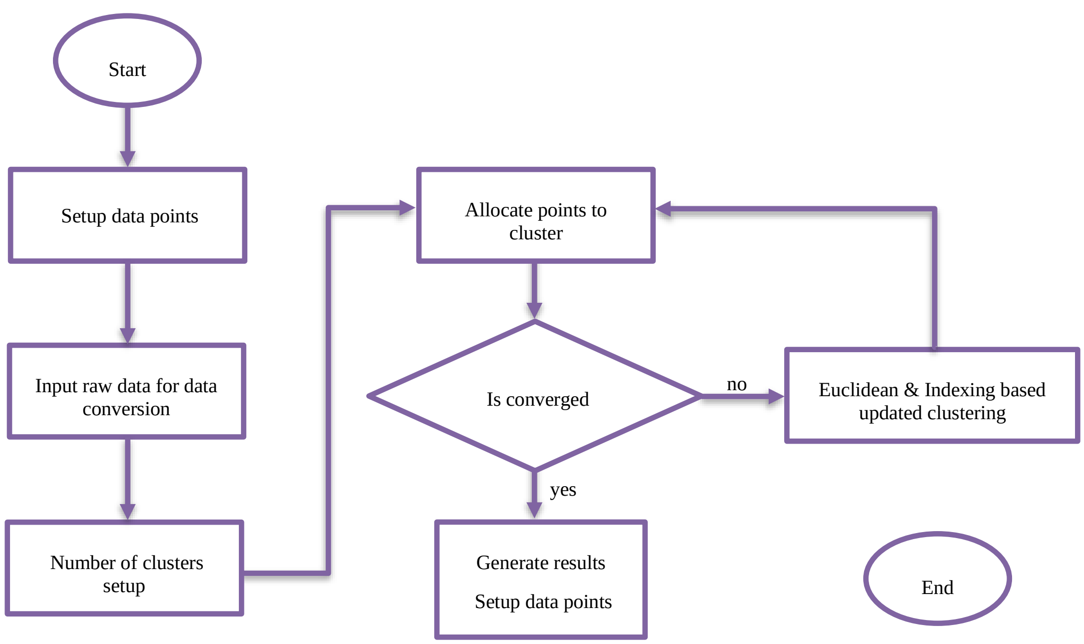

Figure 5-1: low chart of Modified K-means clustering
(Hyndman, R.J. & Athanasopoulos, G, 2014)

5.1.1. Data set

The model is fed by the data of percentage of each number of tries together with frequency and existence of repeated letters.

↓	C1-D	C2	C3-T	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
	Date	Contest number	Word	Number of reported results	Number in hard mode	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries (X)	Frequency	repeated
1	2022/1/7	202	slump	80630	1362	1	3	23	39	24	9	1	3317	1
2	2022/1/8	203	crank	101503	1763	1	5	23	31	24	14	2	3678	1
3	2022/1/9	204	gorge	91477	1913	1	3	13	27	30	22	4	2318	0
4	2022/1/10	205	query	107134	2242	1	4	16	30	30	17	2	3913	1
5	2022/1/11	206	drink	153880	3017	1	9	35	34	16	5	1	73839	1
6	2022/1/12	207	favor	137586	3073	1	4	15	26	29	21	4	53423	1
7	2022/1/13	208	abbey	132726	3345	1	2	13	29	31	20	3	3835	0
8	2022/1/14	209	tangy	169484	3985	1	4	21	30	24	15	5	994	1
9	2022/1/15	210	panic	205880	4655	1	9	35	34	16	5	1	18372	1
10	2022/1/16	211	solar	209609	4955	1	9	32	32	18	7	1	37826	1
11	2022/1/17	212	shire	222197	5640	1	8	32	32	18	8	2	474	1
12	2022/1/18	213	proxy	220950	6206	1	2	11	24	31	26	6	5239	1
13	2022/1/19	214	point	280622	7094	1	16	37	28	12	4	1	430363	1
14	2022/1/20	215	robot	243964	6589	1	8	29	34	20	8	1	12234	0
15	2022/1/21	216	prick	273727	7409	1	8	30	33	19	7	1	3494	1
16	2022/1/22	217	wince	241489	6850	1	3	17	33	29	15	3	1022	1

Figure 5-2: Examples of data set

5.1.2. Determination of classifications

After a series of attempts and comparisons of average distance from centroid, number of classifications are finally determined as 3: easy (denoted as level 1), normal (denoted as level 2), hard (denoted as level 3). Specifically, the initial cluster memberships are chosen as follows:

Word	Contest number	Classification
their	275	Easy (1)
cheek	301	Normal (2)
coyly	409	Hard (3)

Table 5-1: Initial cluster memberships

They are determined by the integrated ranks of percentage of tries and frequency. Generally speaking, the higher percentage of higher tries always shows a more difficulty, a less difficulty usually appears with a higher word frequency, and 'normal' words always have its percentage mostly in 3 or 4 tries.

5.2. Analysis of result

All the words are finally divided as follows. The accuracy of the classifications is mainly reflected by the 'average distance from centroid'. The biggest distance 1.953 appears at Cluster3 (Hard), which is acceptable. Additionally, most of word are classified as 'Normal', while Cluster1 (easy) and Cluster3 (hard) distribute symmetrically on the two side. Hence, this kind of classifications is generally good.

Final Partition

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1 (Easy)	81	335.699	1.687	7.315
Cluster2 (Normal)	190	347.718	1.266	2.775
Cluster3 (Hard)	89	489.124	1.953	10.772

Table 5-2: Final partition of the results

Applying the mode to the word 'eerie', it is classified as '***Normal***', according to its predicted variation of percentage as well as its frequency and repeated letters. The process of the classification is finished by Minitab.

This result is generally accepted by our observations that words having adjacent letter duplicate are easier and its highest percentage appears at 3 and 4 tries, which means that this word can be classified as 'Normal'.

6. Question Four (Other Interesting Features)

6.1. Difference in Repeated letter

As we all know, repeated-letter words can be different in modality. For example, for 'gloom', 'bluff', 'patty', the letters they repeated are adjacent to each other. However, as for 'vivid', 'fewer', 'label', the letters they repeated are separated from each other. Therefore, we divided the repeated letters into two groups by observing these two different features and constructed the figure of this against the percentage of scored results. The figure below indicates that participants tend to guess correctly in fewer attempts when the right answer is in the latter modality.

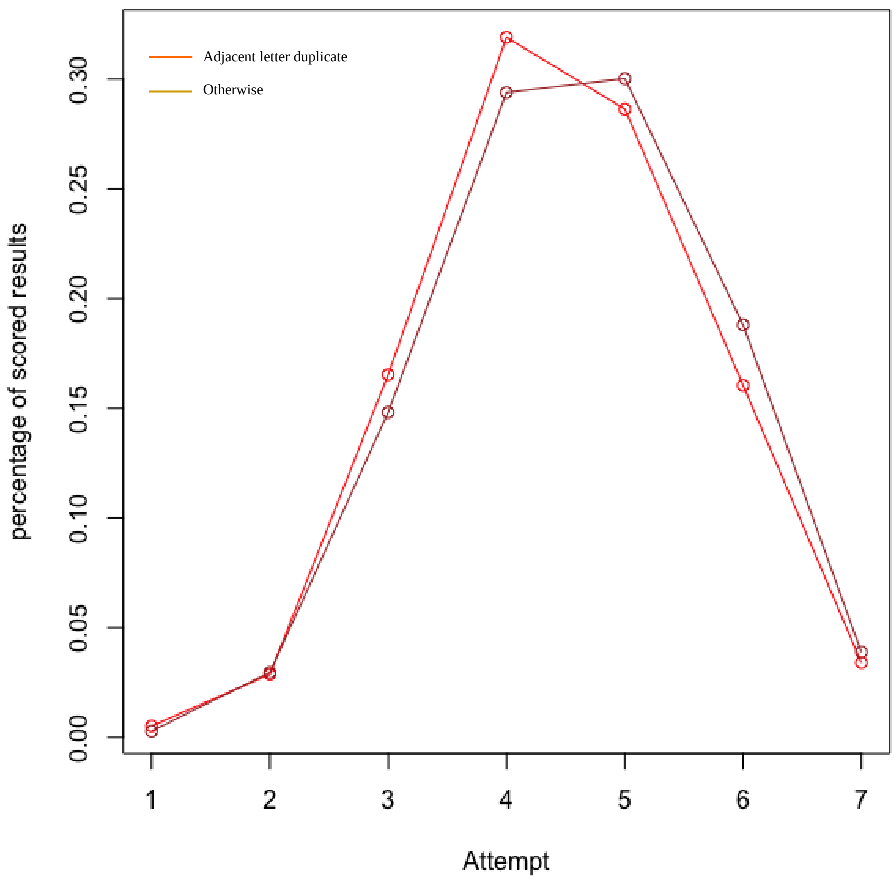

Figure 6-1: Plot of words having different types of repeated letters

6.2. Difference in the ratio of vowel

Another interesting feature that we found is that the difference in the ratio of vowels will affect the accuracy of participants. For example, ‘tryst’ has no vowel, ‘their’ has 2 vowels in total. Therefore, according to the ratio of vowels, we divided these words into four groups. The words in first group have no vowel like ‘tryst’ as mentioned before, the words in second group have only 1 vowel such as ‘glass’, ‘happy’, and the words in the third group have 2 vowels like ‘their’, ‘focus’, and the words in the last group have 3 vowels in total such as ‘axiom’, ‘alike’. By establishing the graph, we found that participants tend to guess correctly in fewer attempts when the right answer has no vowel. However, there is no big difference in the other three situations.

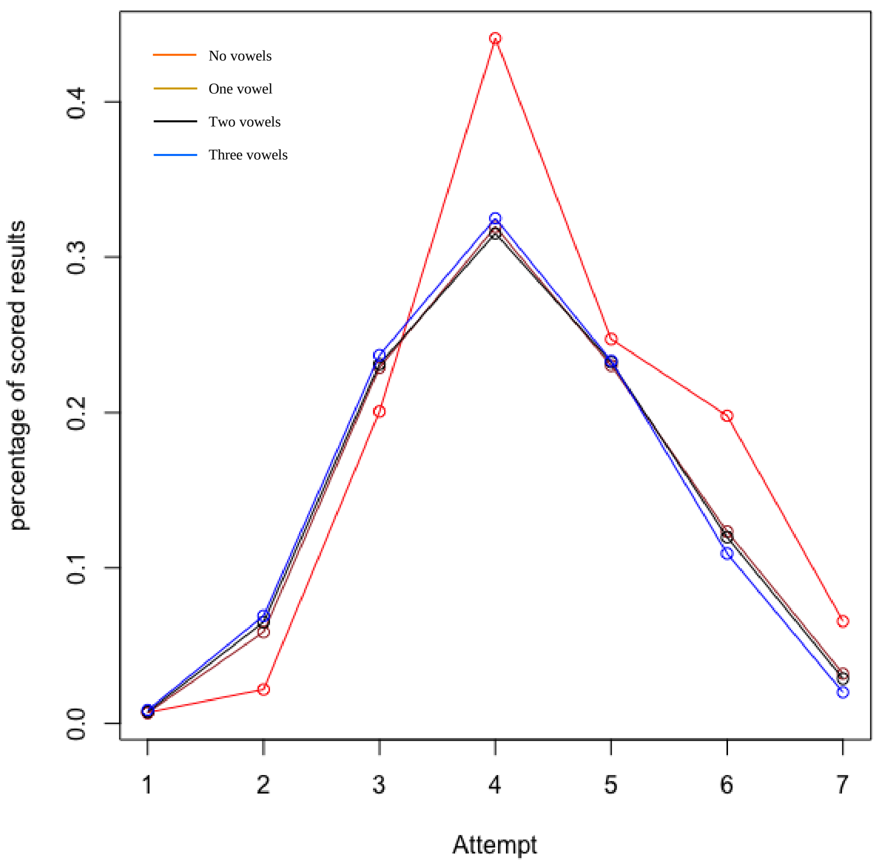

Figure 6-2: Plot of words having different ratio of vowels

7. References

[1] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.

[2] Hyndman, R.J. & Athanasopoulos, G. Forecasting: principles and practice (2nd ed), S.l.]: OTexts, 2014.

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