# Final Project

Stephenson Gokingco, Akash Thakkar, Caroline Hao, James Cornejo

## Abstract

Our team investigated data of the Scripps Spelling Bee spanning the years of 2011 to 2018. We posed two different research questions. First, are some words harder than others? Are there patterns in characteristics for words that are misspelled? While the first question focuses on the word, the second question instead revolves around the competitors. Can demographic characteristics predict a participant's success? Through extensive descriptive data visualization and attempting to employ a classification tree model to predict elimination using word characteristics, we uncovered null results; word characteristics were ineffective in predicting elimination. The random forest classification tree model was, however, more useful in answering the second question. Random forest was able to run a regression on the medium income and city count of a region to show that extrema of the income, i.e. poor or rich, seemed to predict poorly whether a contestant would do well, but around the mean average income, there was greater predictability. Similarly for city size, we found a trend that larger cities could predict max round better than smaller cities. Thus, there were some characteristics that could predict how well a contestant would do.

# Introduction

The Scripps National Spelling Bee (formerly the Scripps Howard National Spelling Bee and commonly called the National Spelling Bee) is an annual spelling bee held in the United States. The bee is run on a not-for-profit basis by The E. W. Scripps Company and is held at a hotel or convention center in Washington, D.C. during the week following Memorial Day weekend. Interestingly, on May 30, 2019, the Spelling Bee ran out of words that might challenge the contestants. They ended up having 8 winners instead of the traditional 1 or 2. The ultimate mission of the Spelling Bee is to help students improve their spelling, increase their vocabulary, and develop correct English usage that will help them all their lives.

Although most of its participants are from the U.S., students from countries such as The Bahamas, Canada, the People's Republic of China, India, Ghana, Japan, Jamaica, Mexico, and New Zealand have also competed in recent years. Historically, the competition has been open to, and remains open to, the winners of sponsored regional spelling bees in the U.S. (including territories such as Guam, American Samoa, Puerto Rico, the Navajo Nation, and the U.S. Virgin Islands, along with overseas military bases in Germany and South Korea). Participants from countries other than the U.S. must be regional spelling-bee winners as well.

Because the Scripps National Spelling Bee is broadcasted live on ESPN and because Scripps posts the round-by-round results, we were able to amass data from 2011 onward. One might consider the Spelling Bee to be an intellectual and competitive sport of memorization and performance, taking into account the difficulty of the words and what it takes to qualify for the national stage. Although there is a vocabulary aptitude component to qualifying for the Finals, we decided to focus solely on data collected during the live, broadcasted spelling rounds. This is the event that people traditionally picture in their minds when they hear

“spelling bee.” We focused our analysis around two main questions: 1) are there discernible patterns or ways to measure word difficulty in a particular year or across multiple years? and 2) how might we be able to predict a contestant’s performance given certain factors?

# Methods

Scripps has released datasets detailing the national level competitions it has put on in the past several decades. These datasets provide information over the words given (e.g. origin, definition, round) and the spellers competing (location, speller order). We organized the available data into two datasets to evaluate word characteristics and speller characteristics respectively.

In regards to word characteristics, we were interested if there were specific features of words given that may determine how likely a contestant is to get eliminated (e.g. if words of Japanese origin or contain a high degree of vowels may be more predictive of speller elimination). In this dataset, we used elimination (a variable specifying whether a speller was eliminated for misspelling that word or not). Some characteristics we hypothesized may affect the difficulty of a word to spell include the origin, the length, the number of vowels, the ratio of the number of vowels to total number of letters, the scrabble score of the word, and the ratio of the scrabble score over the word length. It is important to note that a handful of words had more than one correct spelling. For these words, we thus calculated the word length, scrabble point value, and vowel count as the average of the spelling variations. We incorporated word origin by joining a competition round record dataset to a dataset containing words given and their associated origin. The rest of the characteristics were derived through calculated fields.

To determine whether each of these variables usefully predict speller elimination, we first created visualizations to determine whether there appeared to be a relationship between elimination and the chosen variable. It was important to conduct this initial exploratory phase because we did not expect all features identified to have a substantial predictive power on elimination. After all, participants at the National Spelling Bee are all high-level spellers. After identifying variables that appeared to have a relationship with elimination, we used a classification tree model to evaluate this relationship. Classification trees are especially useful in this scenario where there are not many feature variables, the relationship between the predictive and feature variables is unlikely to be linear, and there may be interactions between the variables.

In evaluating speller characteristics, we were interested to determine if there were identifiable characteristics of spellers that would make them more likely to perform well in the spelling bee. We chose the highest round a speller had advanced to in the bee as a measure of success. The variables we hypothesized may impact speller success included their socio-economic circumstance and whether they come from a spelling bee “hub”. To proxy for socio-economic status, we joined the competitive results dataset to a census median income dataset on the city where the speller qualified from. This estimate may be skewed toward higher socio-economic status because spellers from smaller districts may choose to travel to larger districts in order to compete to qualify for the national competition. To evaluate how competitive a speller’s district may be, we counted the number of spellers who had qualified from each city.

To evaluate the relationship between speller performance and income and city competitiveness, we used a similar classification tree approach as in the word characteristic part of the problem for similar reasons: there are not many feature variables, the relationship between the predictive and feature variables is unlikely to be linear, and there may be interactions between the variables.

# Results

As discussed earlier, this paper has two main research objectives. First, it seeks to uncover whether or not word characteristics can predict elimination. And second, it aims to understand if socioeconomic demographic features can help predict success at the individual participant level.

To answer the question over word characteristics, the first step in our visual data analysis was creating boxplots for word length, scrabble value, vowel count, and the two ratio variables by elimination. The left boxes plot the distribution of the feature variables for words that were spelled correctly and the right boxes plot the distribution of the feature variables for words that were spelled incorrectly, and thus recorded as eliminations (Appendix Figures 1-5). All five word characteristics are distributed almost identically across correctly and incorrectly spelled words. And thus, these word features fail to predict elimination.

The last word characteristic to be explored is word origin. The bar plot in Figure 1 of elimination frequency by origin group shows that word origin may provide some variation in explaining

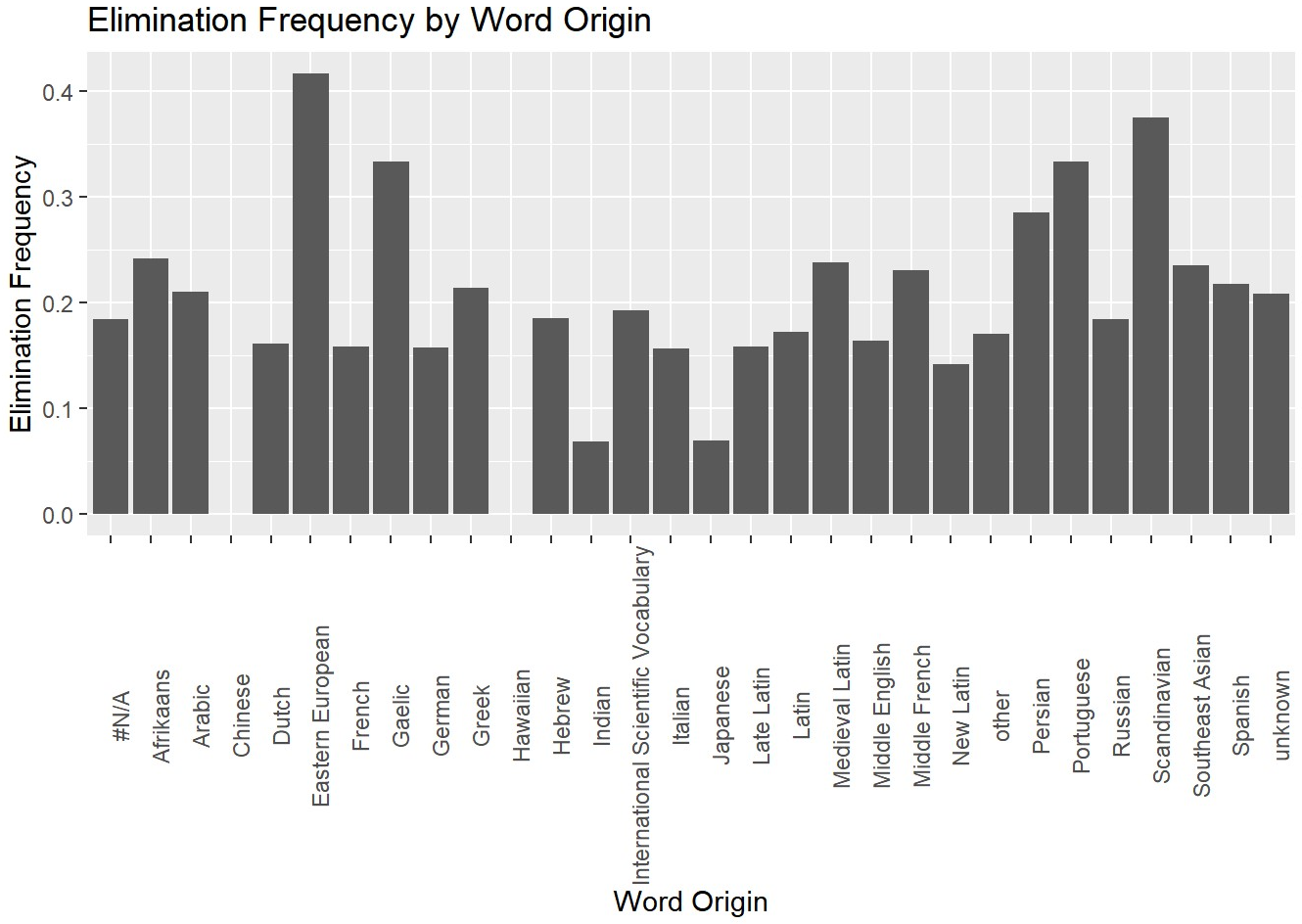



Figure 1 Elimination frequency by origin group. Elimination appears to vary across different origin groups.

elimination. As such, we continued our visual analysis by plotting elimination frequencies for the five prior word characteristics interacted with word origin. The plots, which are included in the appendix, were largely similar across different word origins (Appendix Figures 6-9). However, although unlikely, there was still some potential for variation based on unobserved interactions.

Thus, our team proceeded to grow a classification tree model to predict elimination. As discussed earlier, a critical benefit of using a tree model is that it naturally tests potential interaction terms. All six feature variables were included in the model and the unpruned tree, depicted below, grew to a size of 209 terminal nodes (Figure 2). However, once the tree was pruned, the model failed to produce any branches. As the size of the tree grew, the estimated error also increased. Thus, the error was minimized at a tree that was just the root. The plot of error is shown below and is distinctly marked with an increasing slope (Figure 3). This is a clear indication that the tree model fails in this application.

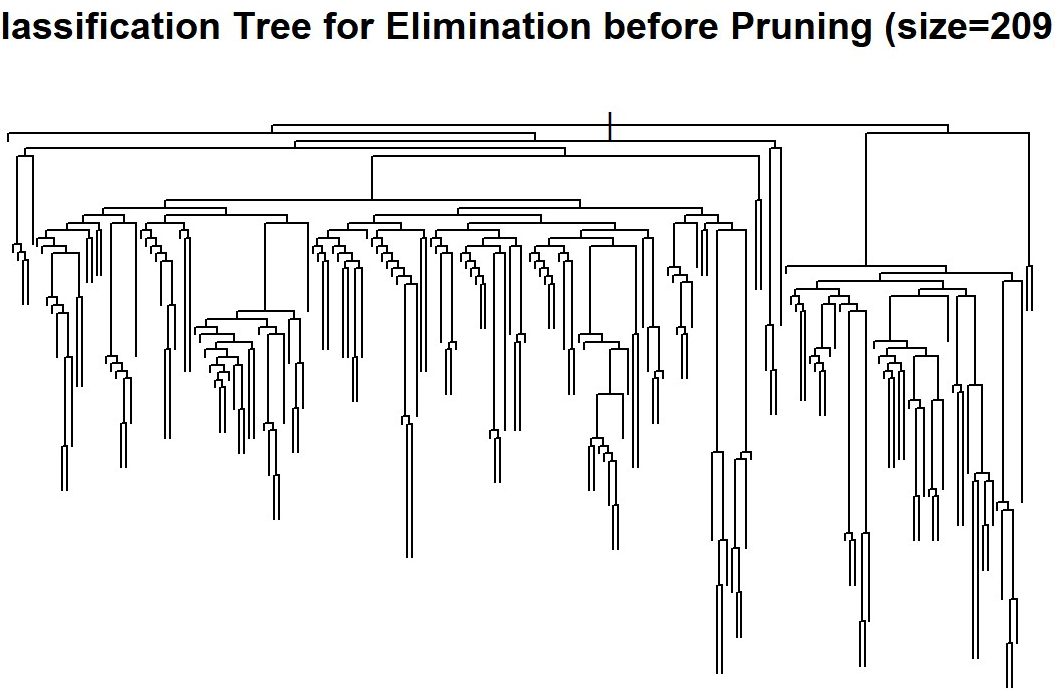



Figure 2 Classification tree predicting elimination based on word characteristics. Contains 209 terminal nodes prior to pruning.

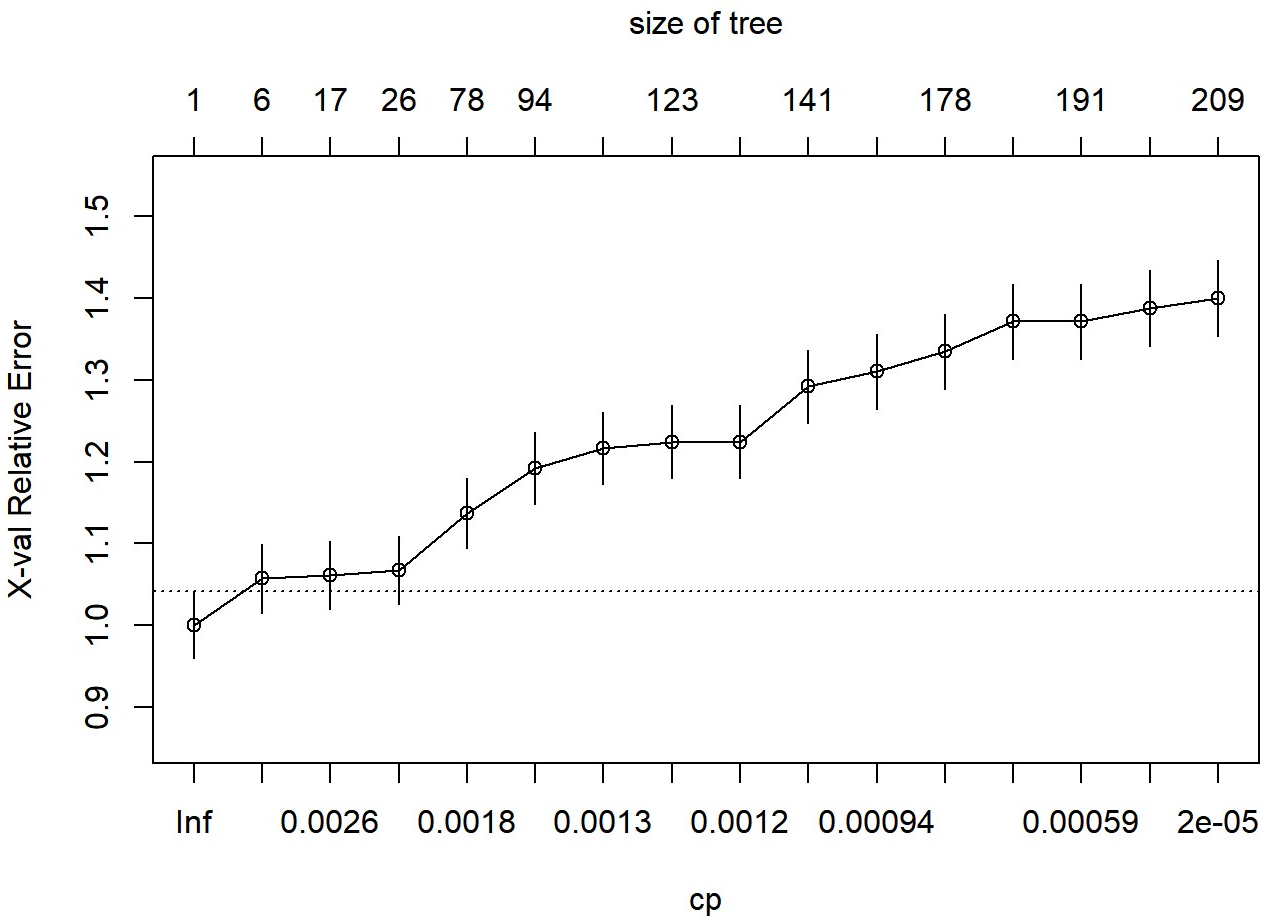



Figure 3 Error plot of classification tree. RMSE increases as the tree grows, indicating that none of the variables included improves the model's performance in predicting elimination outcome.

There's a relatively positive linear relationship between highest round attained and median income until we hit around $150,000. Perhaps spellers who come from significantly high income backgrounds are not as pressured to succeed, and spellers with significantly low income backgrounds either 1) don't have as much resources as their peers in higher brackets or 2) have other commitments and pressures in their lives besides winning a spelling bee.

For our second question over speller characteristics predicting success, we ran a random forest algorithm. The features that were considered in the random forests algorithm were {'med_income', 'pass_3', 'city_count', 'big_city'} in predicting 'max_round'. The max_round feature means the round a certain contestant got to before being eliminated. This feature was predicted by looking at the medium income of the city where the contestant was from and whether or not it was a big city. This used the features 'med_income' and 'big_city', where 'big_city' was defined as if more than four contestants came from a certain city, this is classified as a 'big_city'. With this regression, we wanted to answer the question whether or not socioeconomic status had an effect on performance, i.e. does a higher medium income and/or a big_city affect the maximum rounds of a contestant. For this question, we split our data into a training and testing set for both of the random forests. The first random forest was trained and then tested against an out-of-sample data set. These training and testing splits defined the predictive features by numerical value of each feature. For example, within the 'med_income' column, an income of, lets say, 40,000 would be compared against an income of 60,000 to predict the max round. Another random forest was fitted using the

train/test split also. This random forest predicted ‘max_round’ by ‘med_income’ and ‘city_count’, but the entire column list was included and not separated like in the first random forest. In addition to these forest and variable importance plots, we took the partial derivatives of the second forest with respect to ‘med_income’ and ‘city_count’ respectively. This showed how changing each of these variables would affect the predicted value of maximum number of rounds.

After creating the train and test split, these factors were then converted to numeric values to run a regression instead of classification. From randomForest 1, after predicting, we got an RMSE value of 2.25 for our out-of-sample test. In addition, the plot (Appendix Figure. 11) showed the convergence of the solution as the forest grew. At around 150 trees, it seemed that the error leveled out and became a steady-state solution. Since there were no big fluctuations occurring, this meant to us that around this amount of trees, the predictive ability of randomForest 1 converged to an error value of 6.6 - 6.7. After this, we plotted the variable importance plot of this randomForest. Seen here (Figure. 4), this plot shows the importance of each variable in predicting max_round. Some notable features that stand out is that med_income with values 47,694 and 65,852 seemed to have a stronger importance in predicting versus values with 102,185 or city_count. Like above, this could mean that spellers from significantly higher income backgrounds don’t care as much. It seemed to show that spellers below the mean average income of ~75,000 were predicted to be able to achieve a higher max round.

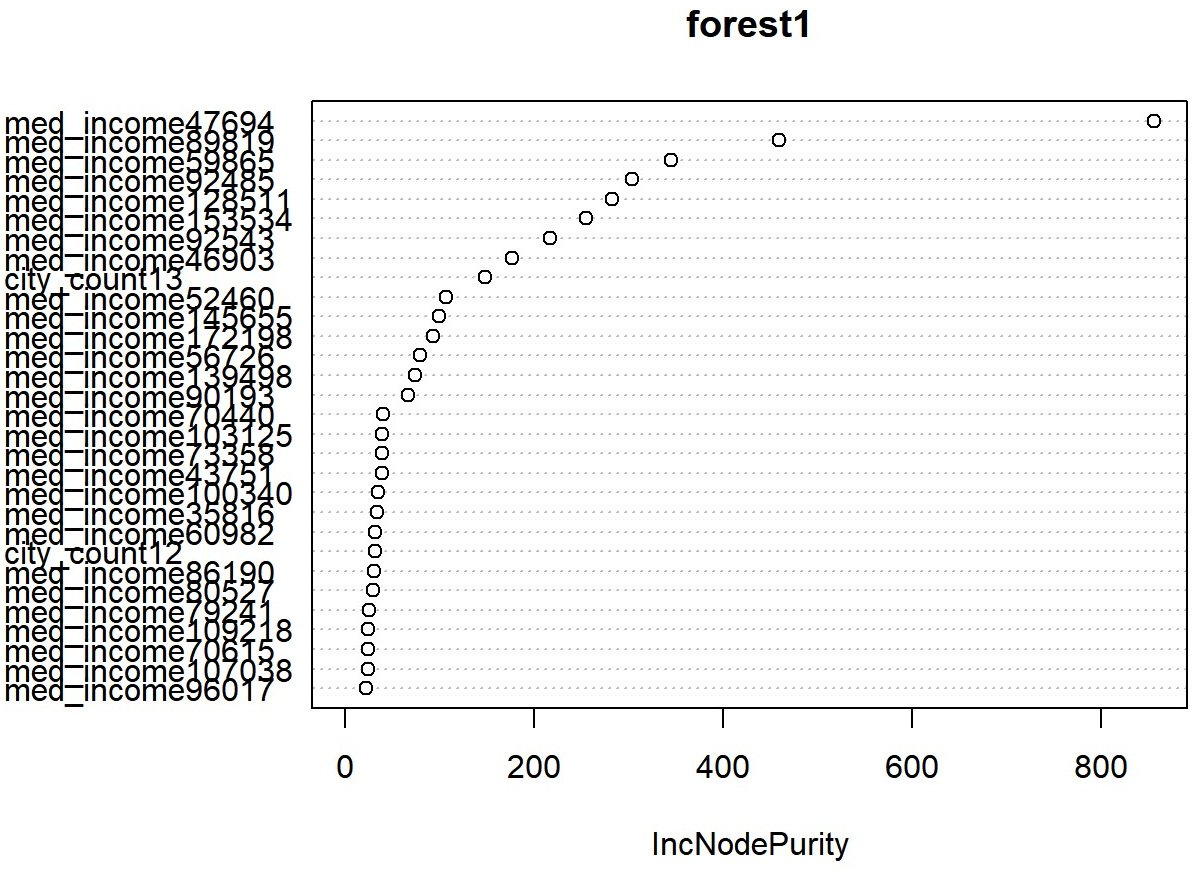



Figure 4: Importance of variables in forest1 predicting max_round. High IncNodePurity of a variable indicates better prediction

Continuing with the second random forest, we similarly set the parameters as “max_round’ becoming predicted by ‘med_income’ and ‘city_count’ from the training set, but didn’t use a matrix for the predictors. This allowed for prediction of the entire variable ‘med_income’ or ‘city_count’ compared to each other, instead of the previous forest, which compared various values within each predictor variable. From randomForest 2, this gave us an RMSE value of 2.35 between the values predicted by this forest and the out-of-sample test test data, ‘scripps_test’. This is very comparable to the RMSE value from randomForest 1. With plotting this forest, we don’t see the huge spike like in randomForest 1, but sort of a plot that resembles a negative exponential. Similarly, the error seems to converge to a value under 6.00 after 200 trees (Appendix Figure. 12). Since this forest didn’t compare values within each predictor variable, the variable importance plot just shows that ‘med_income’ has a stronger predictive value of max_round than ‘city_count’ (Appendix Figure. 10). This doesn’t show how different med_incomes compare to each other, which was shown in randomForest 1. Although, we were able to take the partial derivative and plot each of these variables against randomForest 2. From this plot (Figure. 5), we see that towards the end of the med_income, 50,000 and 250,000, there is not as strong of a correlation as around the mean average of the med_income. From values 60,000 to 150,000, there is a greater predictability of max_round based on contestants from this socioeconomic demographic.

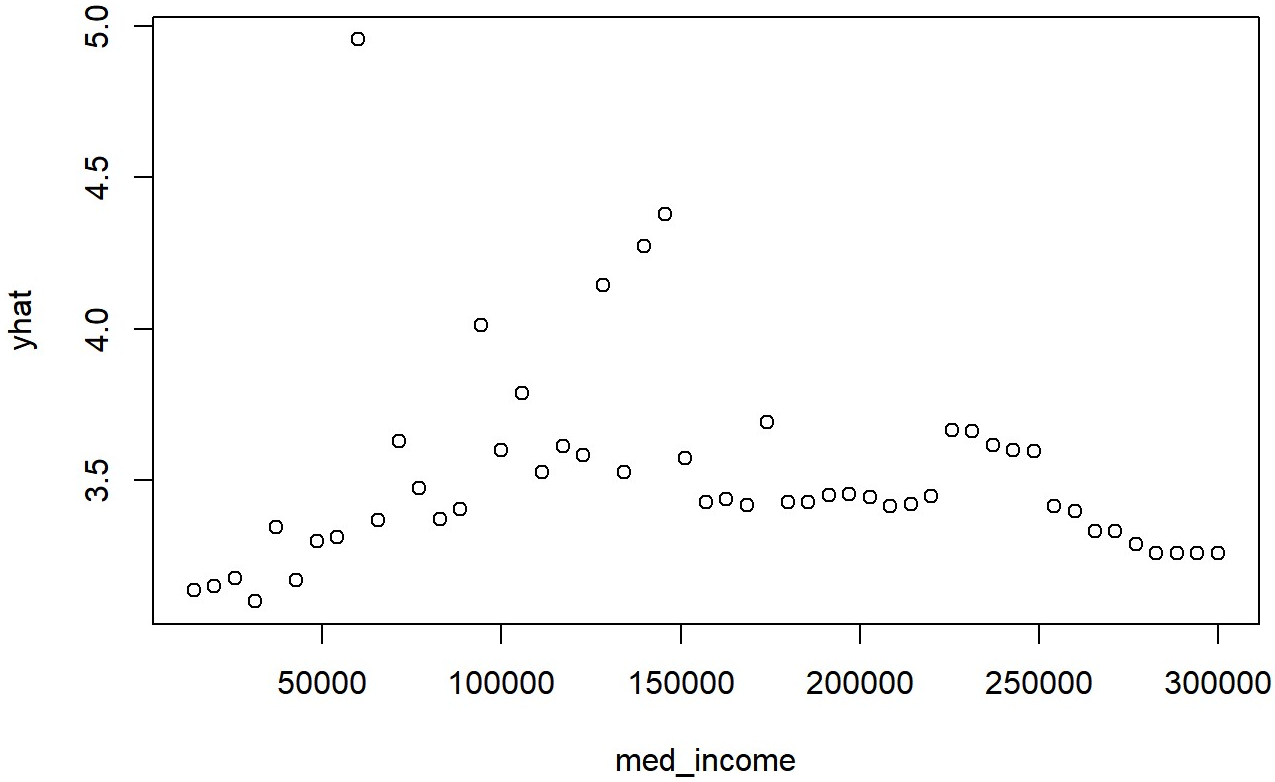



Figure 5: Affect of various med_income values of max_round, there is a slight uptrend toward the center, mean of the plot

With city_count, we see a slight uptrend in larger cities (Figure 6). This makes sense because we expected larger cities to perform better due to greater academic resources to the spellers and sort of “richness” of the area. With smaller cities, we expected less of a predictive ability. In fact, there is an outlier at around a value of 12-13 city_count. Parsing through the original data, we found that these cities were New York, New York, Athens, Ohio, Houston, Texas, and Miami, Florida, respectively. These cities are relatively large and can pull from a larger pool of students to send to the spelling bee, which could answer this large predictive outlier.

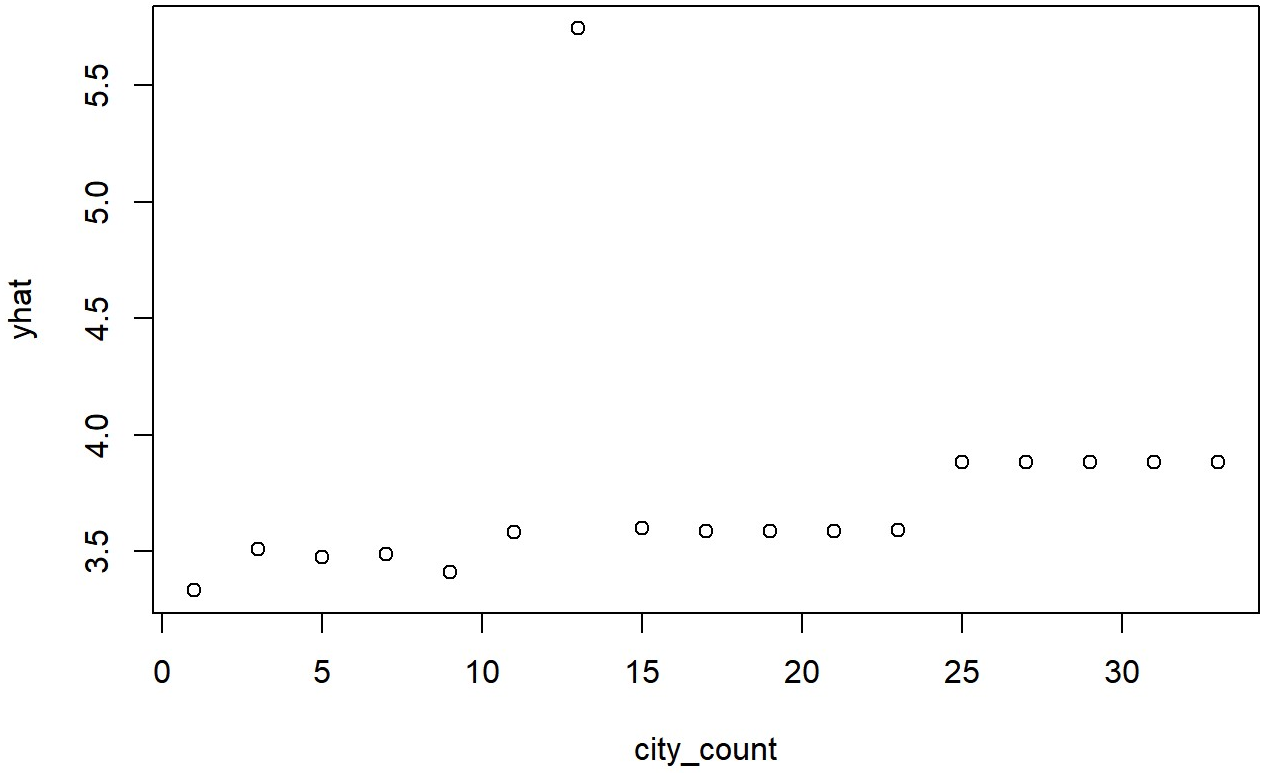


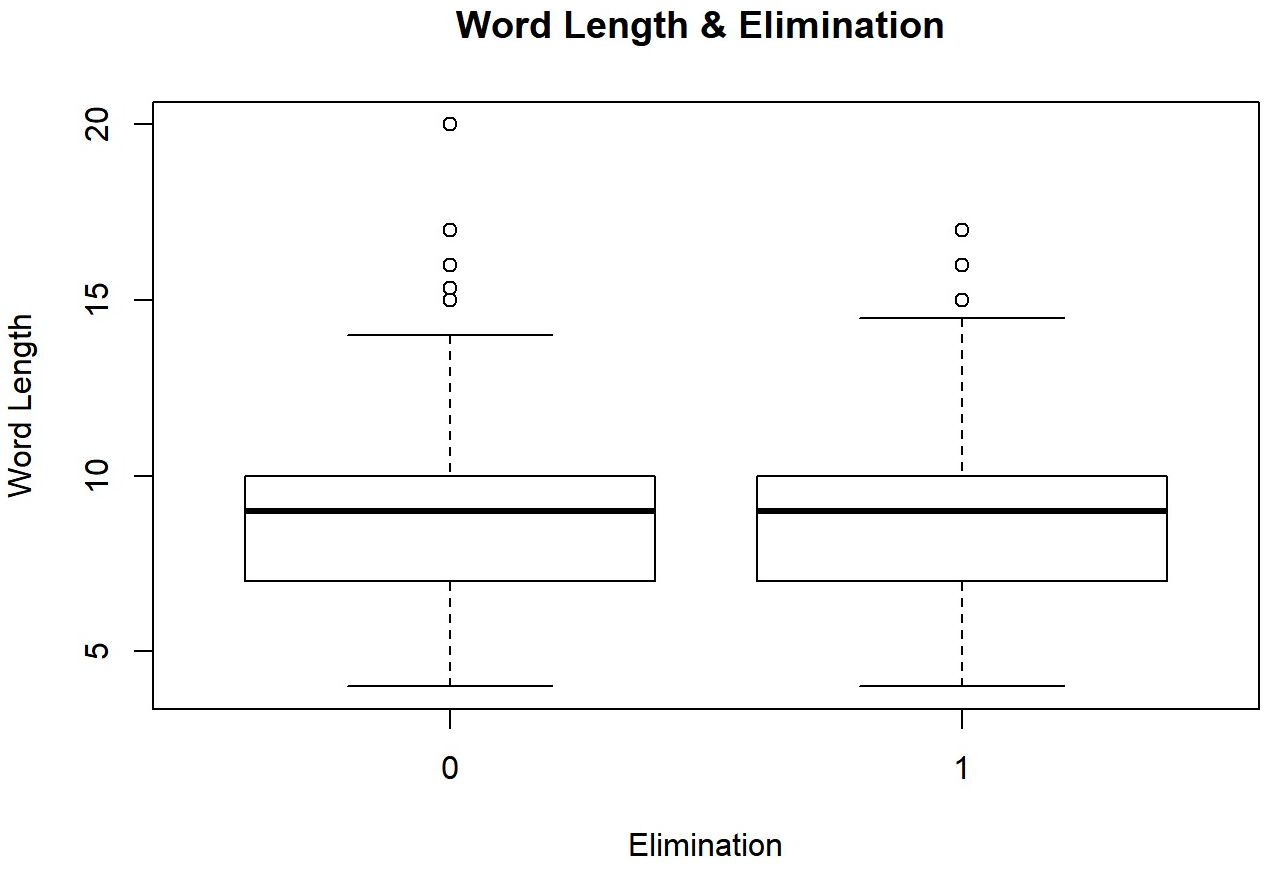

Figure 6: Trend of city_count affect on max_round, there is a positive trend as city_count increases, meaning better predictability of max_round

# Conclusion

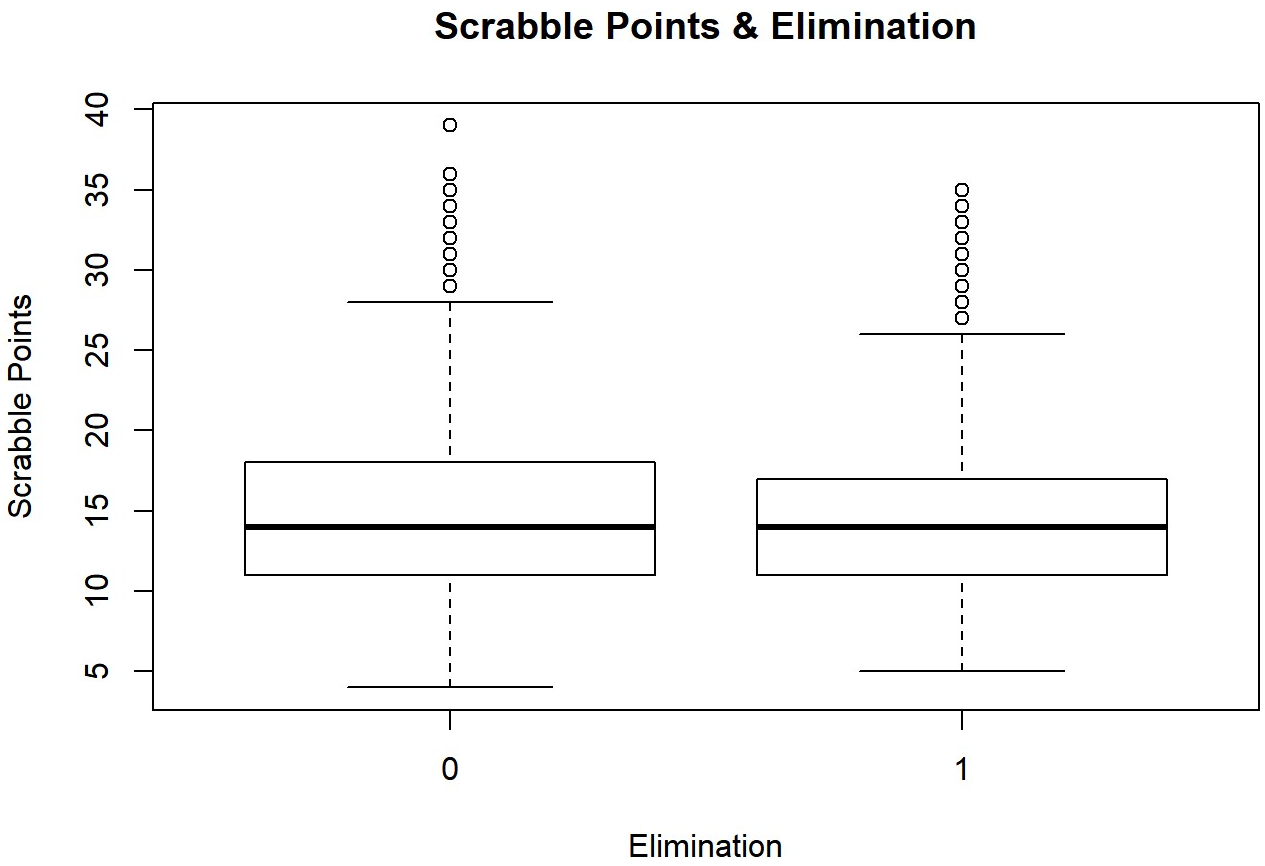
There are two main takeaways from our analysis: 1) our classification methods have demonstrated that there are no significant differences in difficulty between words given certain features like language of origin or our calculated scrabble score to describe word complexity; 2) spellers that come from more fortunate socioeconomic backgrounds in terms of median household income by state or spellers that come from “feeder” states perform significantly better as evidenced by the data and methods. Perhaps the reasoning behind our first finding stems from the fact that the finalists are all accomplished spellers or that there isn’t much variation among words in the first place in terms of the features we showcased. Our second finding makes intuitive sense because a tougher competition in regionals conditions a speller better than a speller who makes it through a pool that’s not as competitive. Additionally, a speller with more financial resources up

to a certain point is better equipped to be successful in competition. If we wanted to take our analysis further, we would like to explore whether or not turn order has an impact on speller performance. This is one way in which we can analyze factors besides the words themselves or where the spellers come from to predict speller performance.

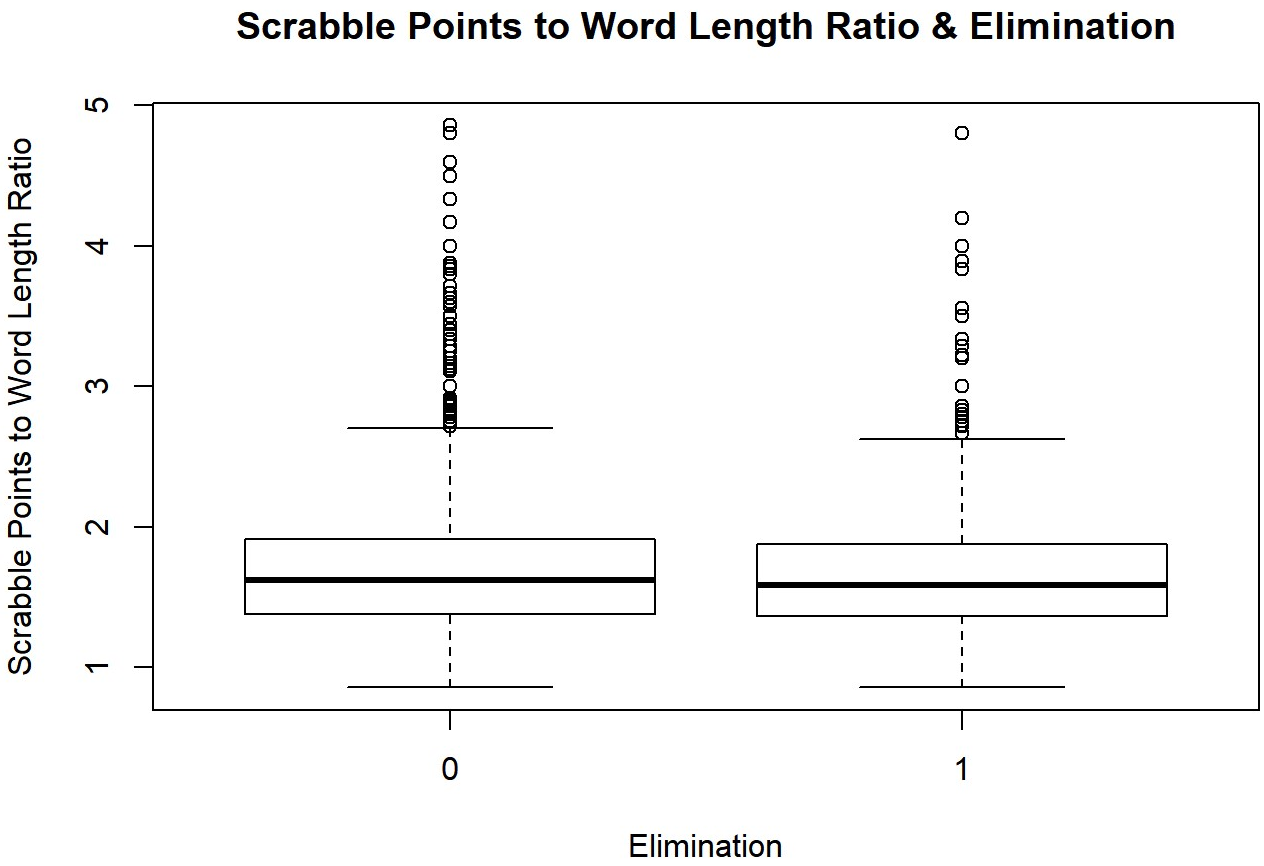
# Appendix



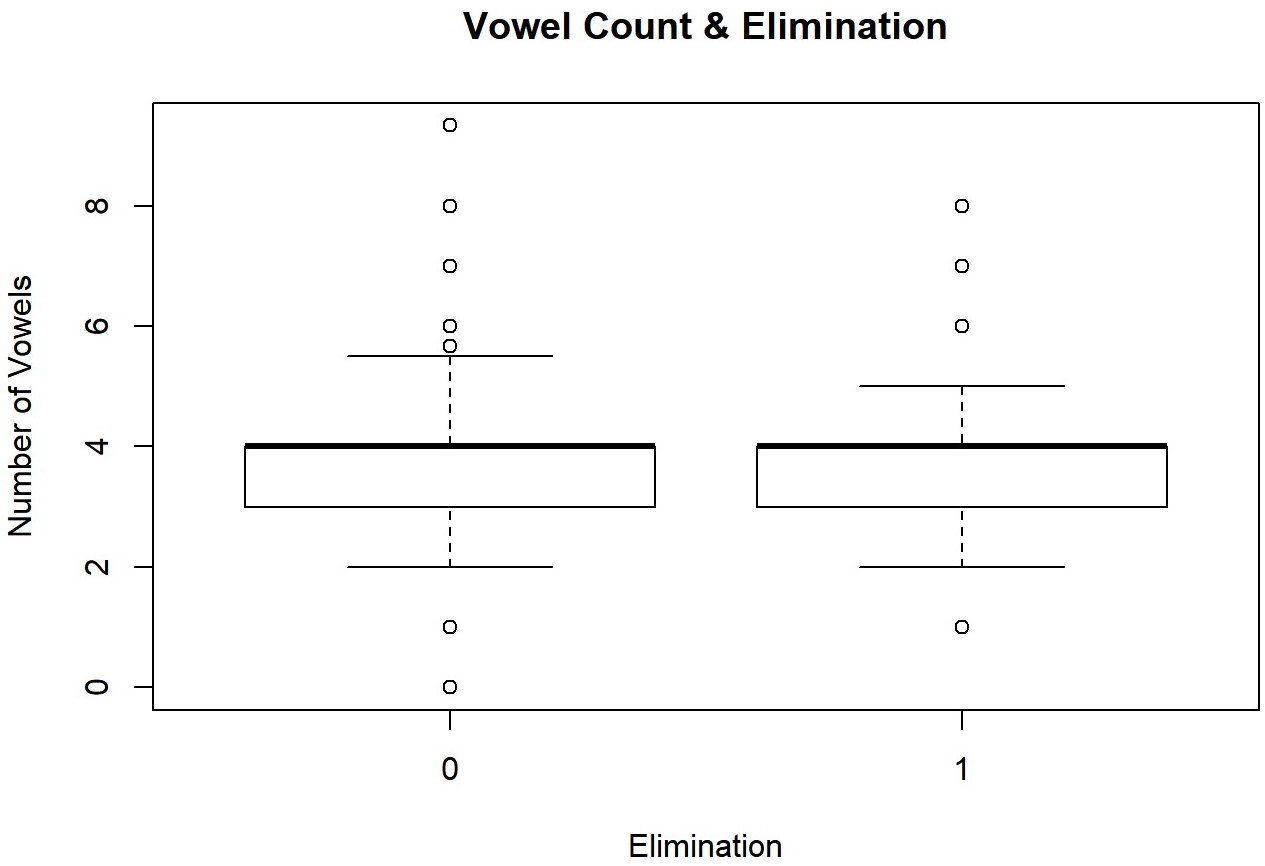

Appendix Figure 1 Elimination likelihood as predicted by word length. There is no discernable difference in average word length between the words contestants were eliminated on versus not.



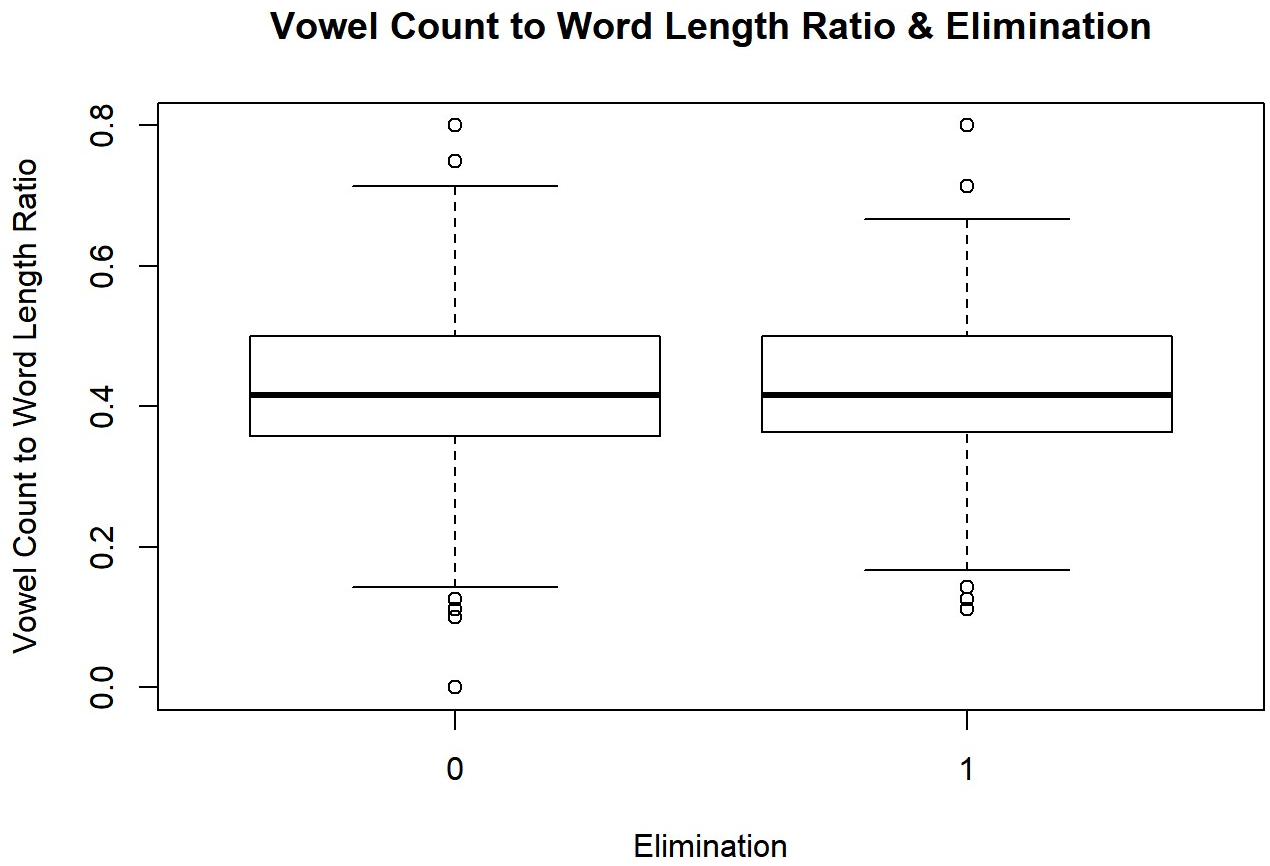

Appendix Figure 2 Elimination likelihood as predicted by scrabble points. There is no discernable difference in average scrabble points between the words contestants were eliminated on versus not.



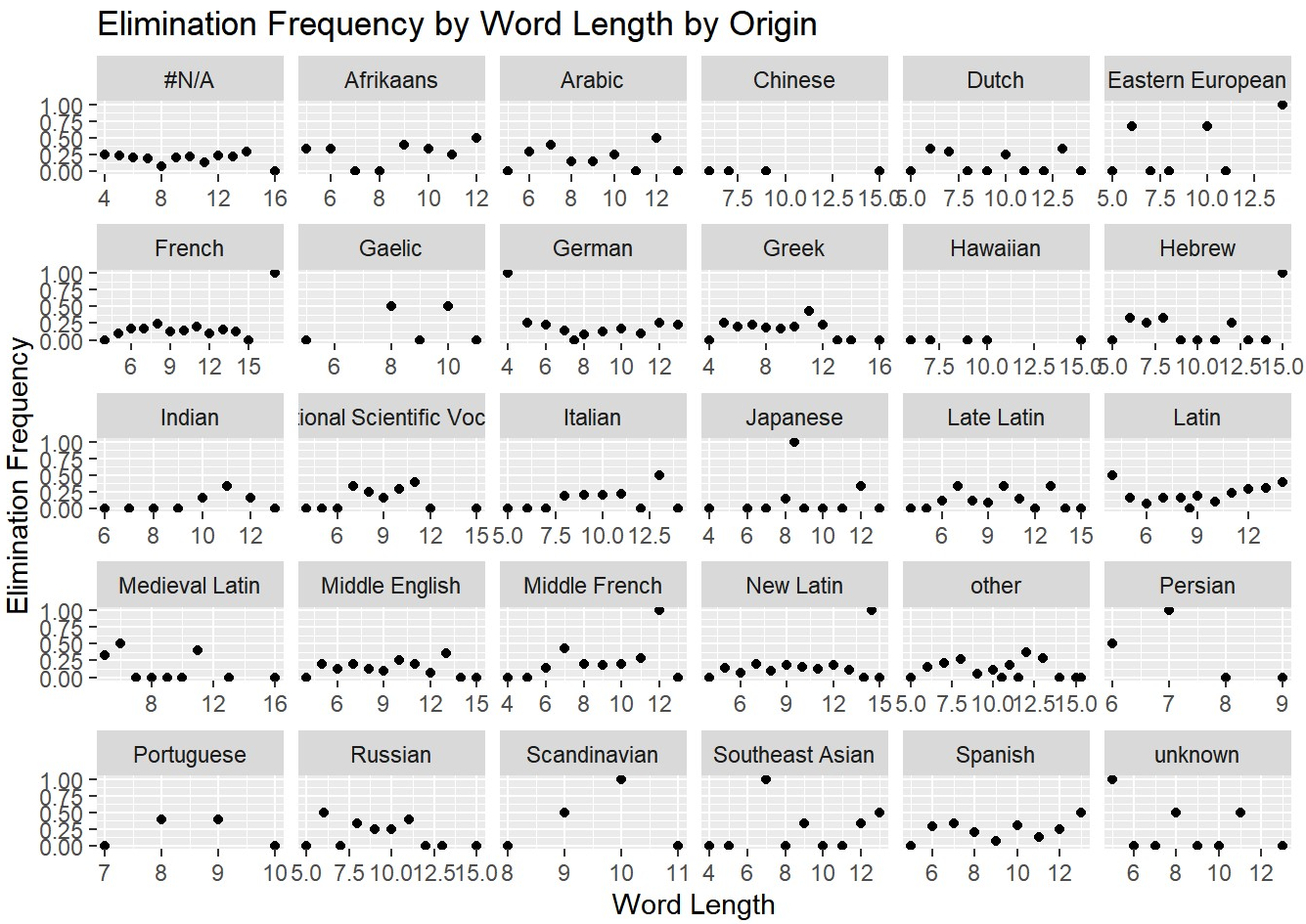

Appendix Figure 3 Elimination likelihood as predicted by scrabble points to word length ratio. There is no discernable difference in average scrabble points to word length ratio between the words contestants were eliminated on versus not.



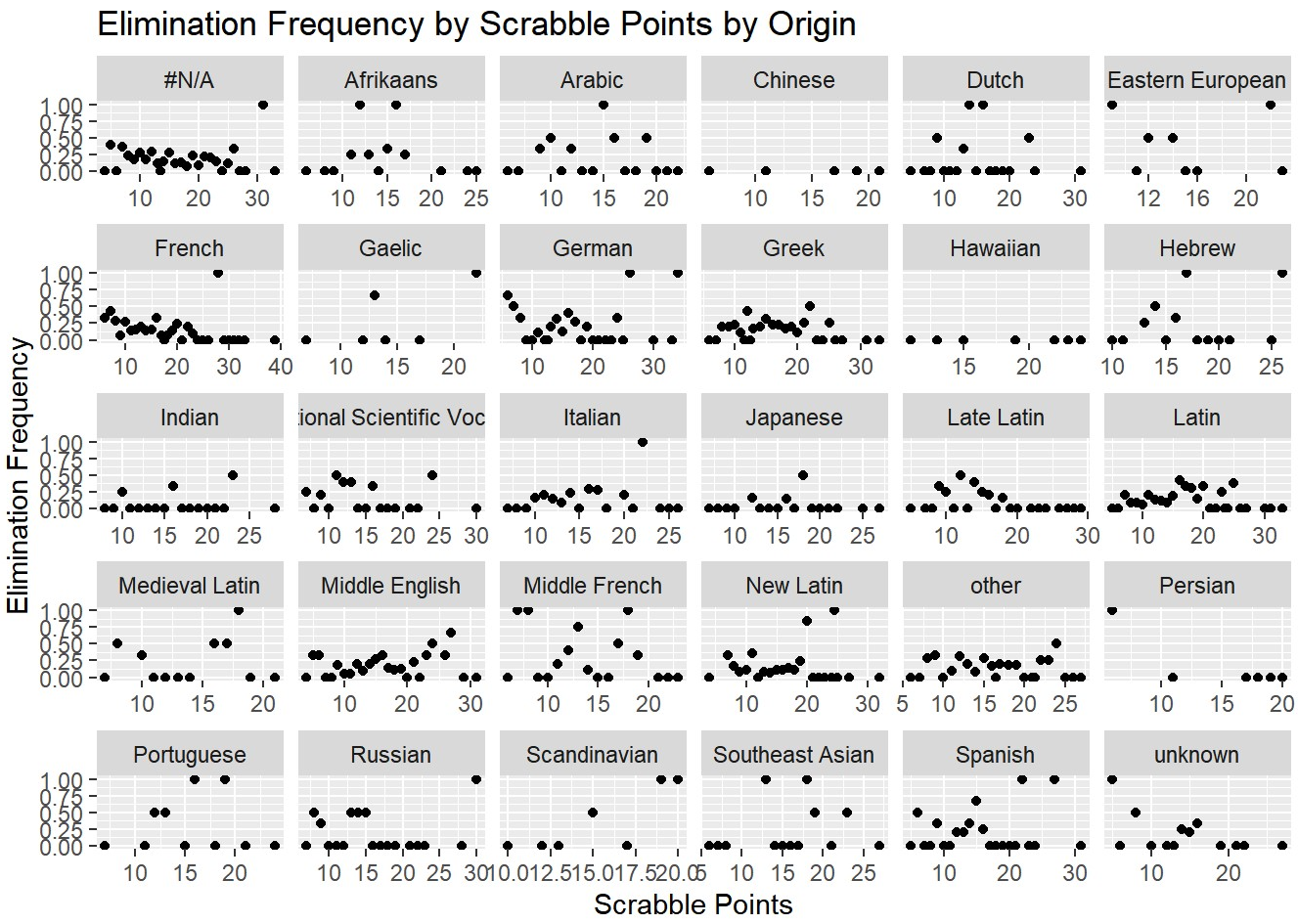

Appendix Figure 4 Elimination likelihood as predicted by vowel count. There is no discernable difference in average vowel count between the words contestants were eliminated on versus not.



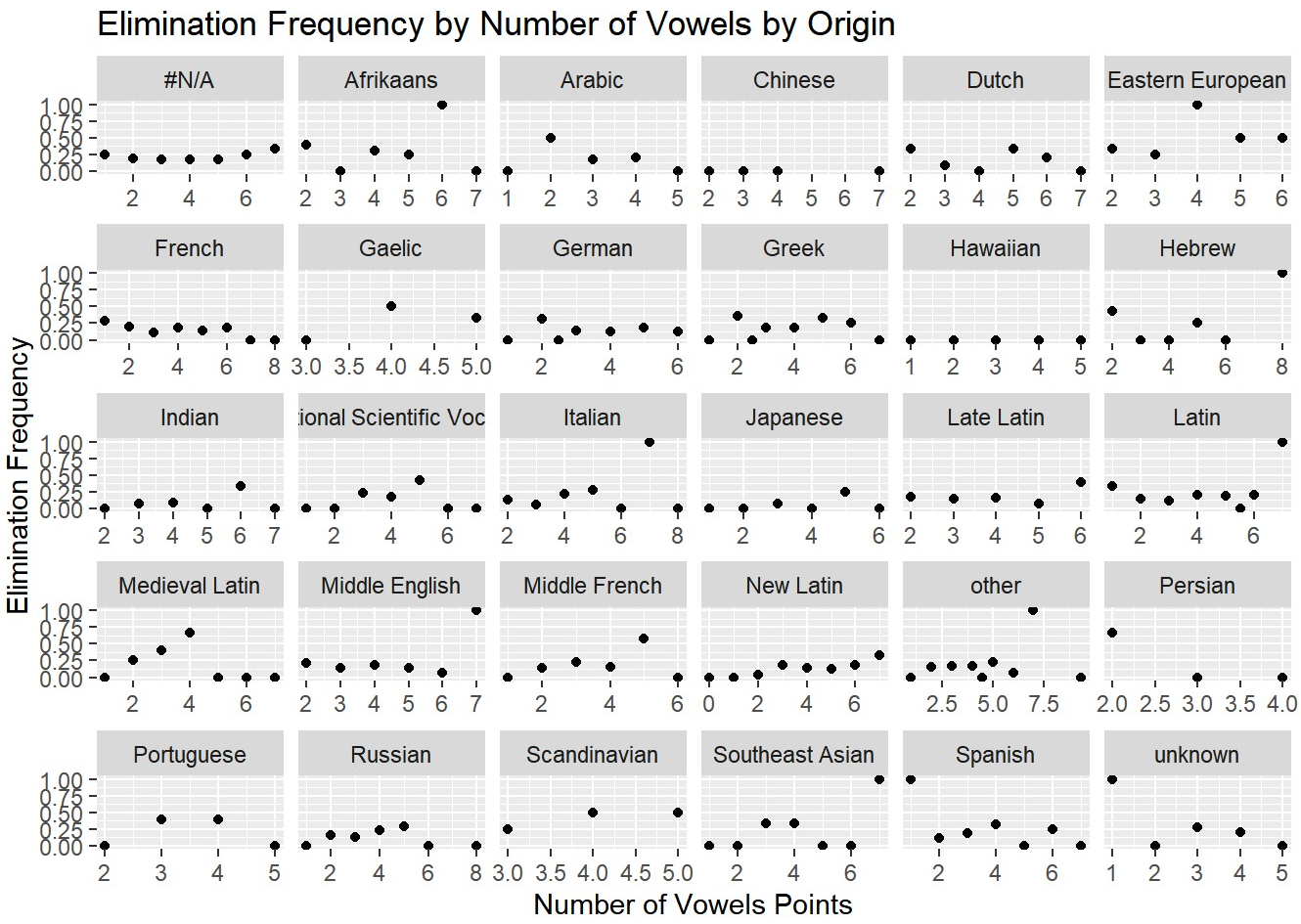

Appendix Figure 5 Elimination likelihood as predicted by vowel count to word length ratio. There is no discernable difference in average vowel count to word length ratio between the words contestants were eliminated on versus not.



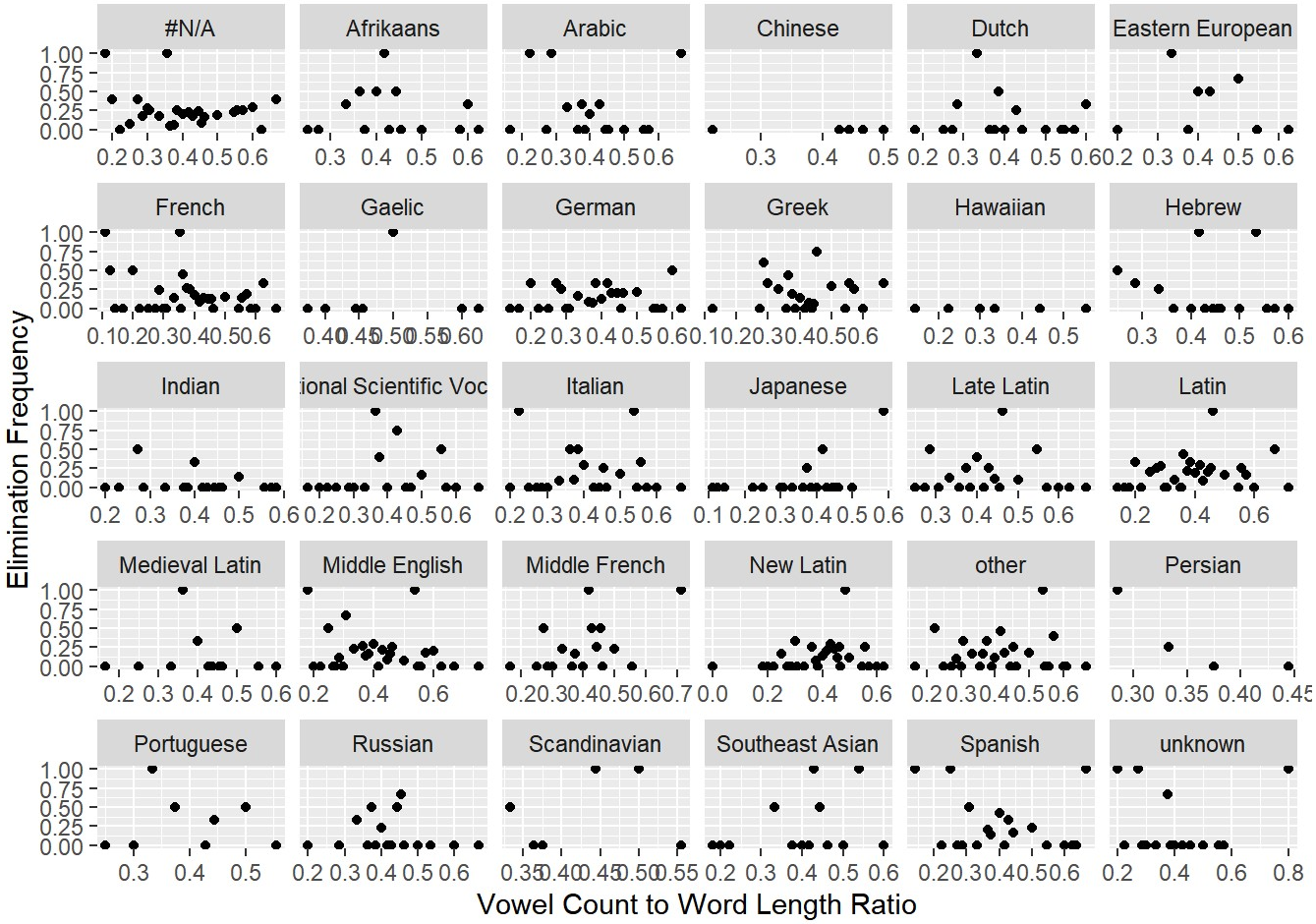

Appendix Figure 6 Elimination as varies with word length moderated by origin. There appears to be a slight moderating influence of origin in the relationship.



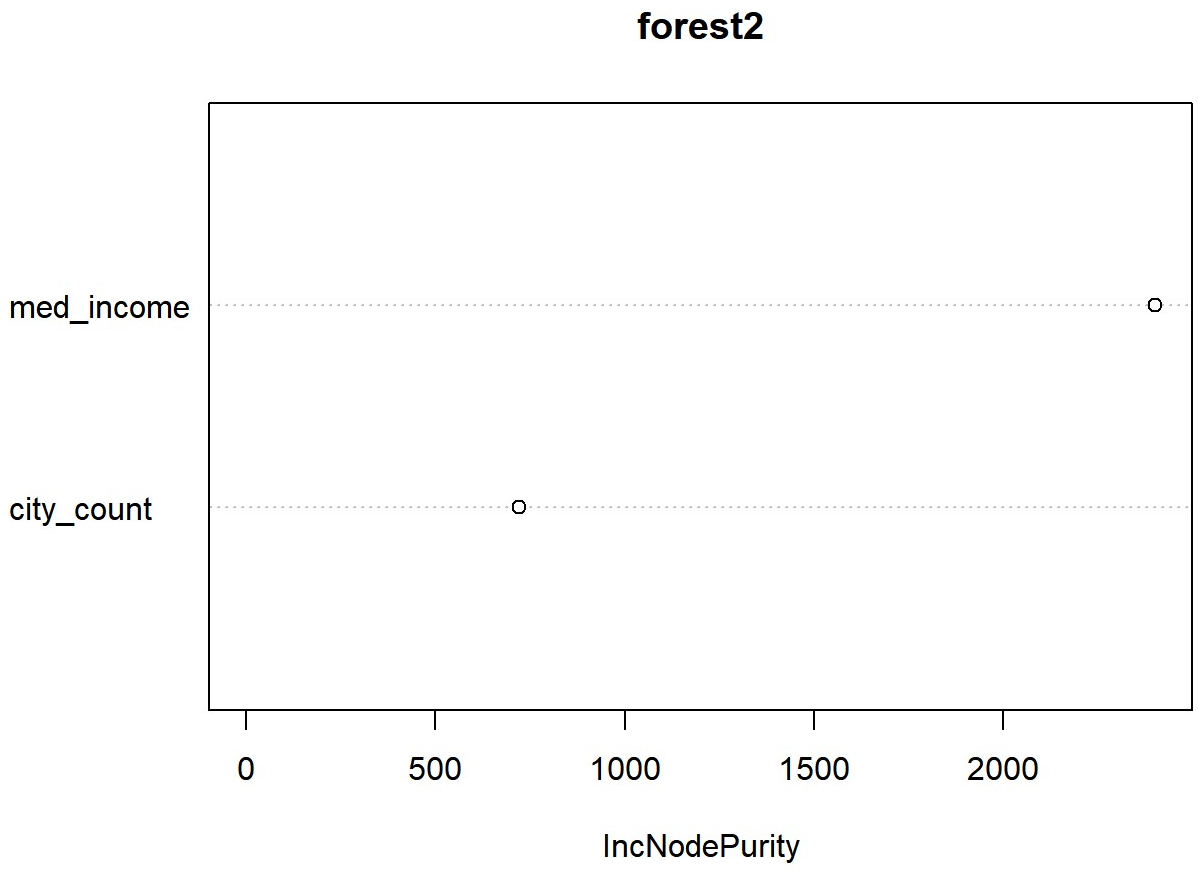

Appendix Figure 7 Elimination as varies with scrabble points moderated by origin. There appears to be a slight moderating influence of origin in the relationship.



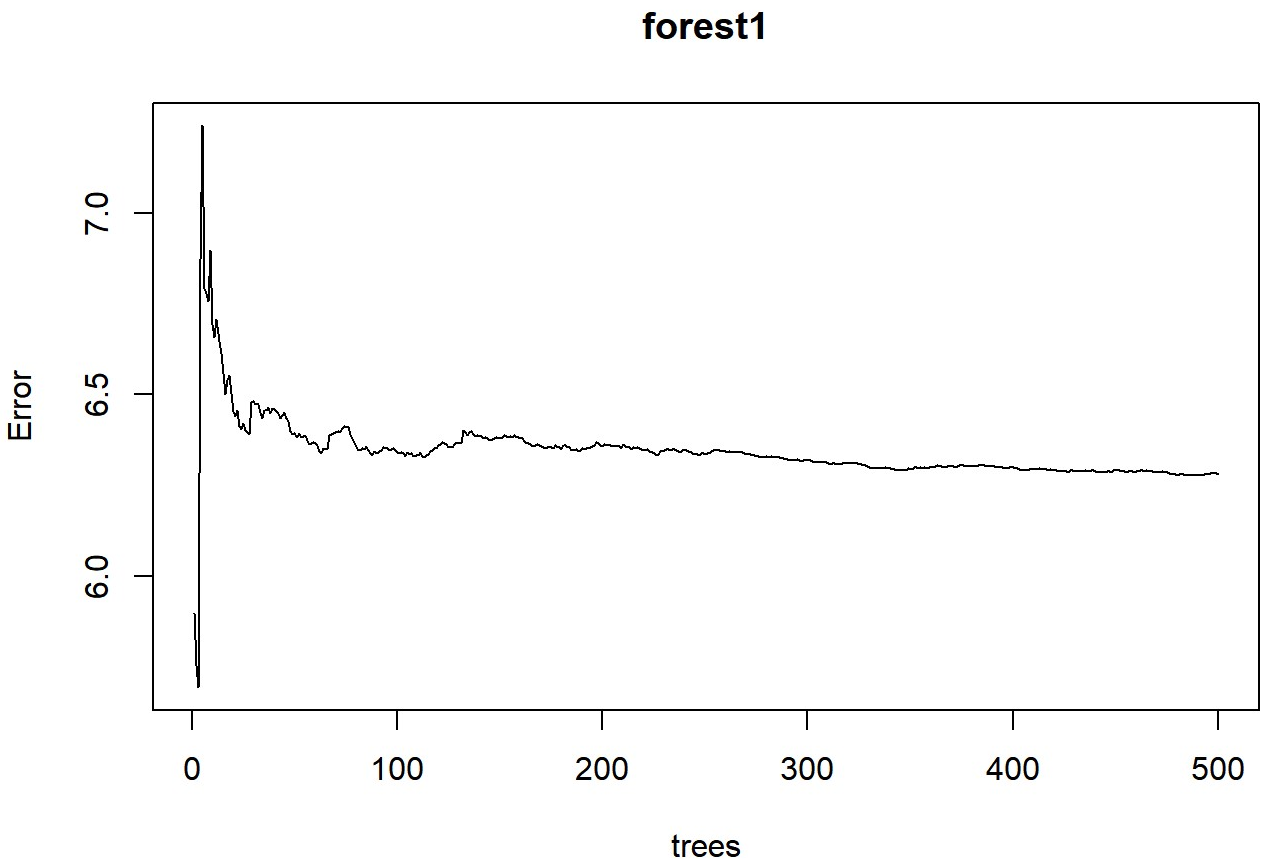

Appendix Figure 8 Elimination as varies with vowel points moderated by origin. There appears to be a slight moderating influence of origin in the relationship.



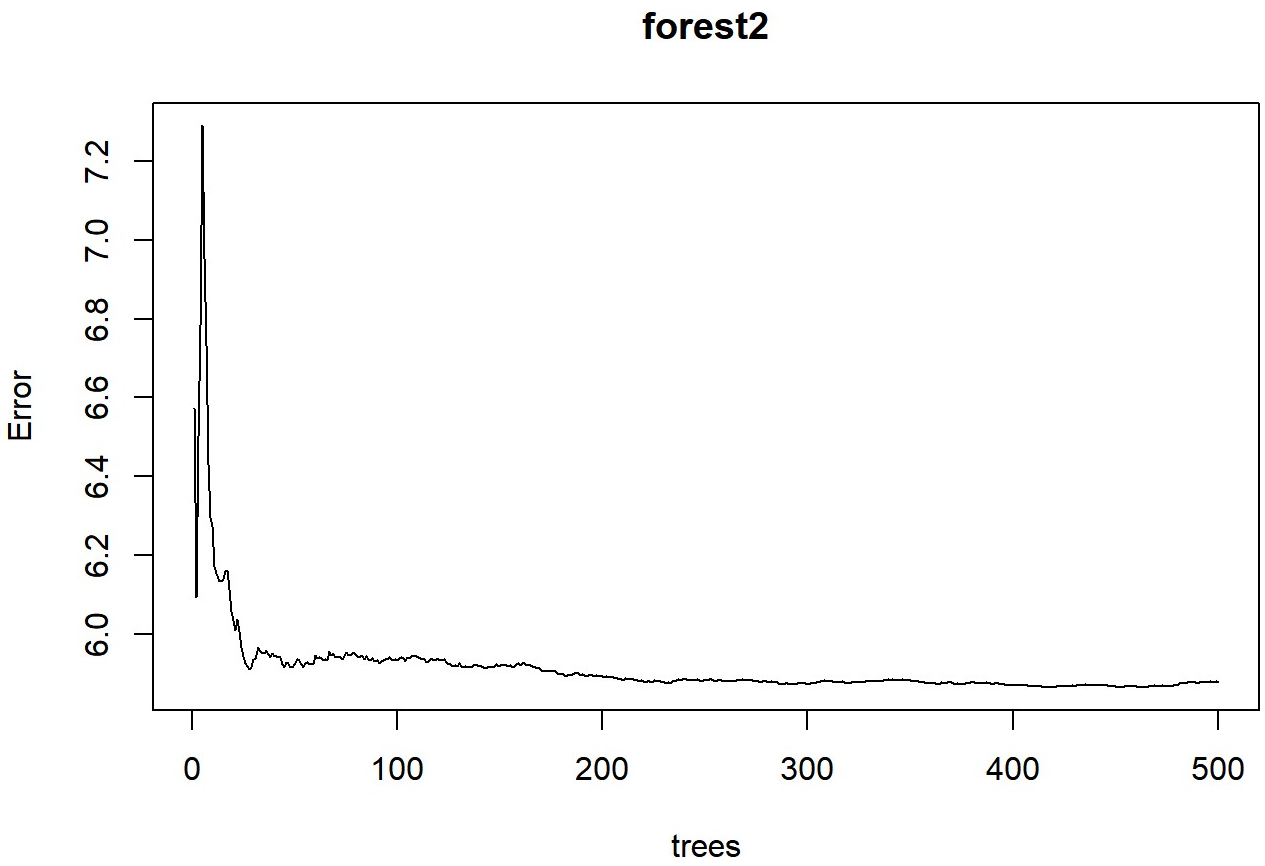

Appendix Figure 9 Elimination as varies with vowel count to word length ratio moderated by origin. There appears to be a slight moderating influence of origin in the relationship.




Appendix Figure 10 IncNodePurity for forest 2, this shows that med_income is a better predictor of max_rounds than city_count.




Appendix Figure 11 For forest1, as the forest grew, the error converged to around a value of 6.25.




Appendix Figure 12: For forest2, as the forest grew the error converged to value under 6.0. This is better than forest1 and has the shape of a negative exponential.