GRAND CHALLENGES – REPORT

Introduction and motivation

For my data analysis project for Grand Challenges in Computer Science, I decided to investigate and create data around the New York Times puzzle game, *Wordle*. I found data of the most common letters in the alphabet, in the Wordle answer and guesses lists in order to find the best starting word. I also separated these findings into vowels and consonants and compared the list of common letters to the English dictionary. And finally, I found the best second guess for Wordle and the worst starting word.

However, before starting the investigation, I decided to hypothesis the expected results; "E will be the most common letter in the Wordle list, based on the frequency of letters in the English dictionary, but the guess and answer lists will slightly differ." The answer to the hypothesis was founded and will be discussed later in this report.

I decided on investigating Wordle for my project, as the data and results would provide many benefits for Wordle players. Over 14 million people in the USA play Wordle[1], and starting with a logically better starting word will increase their chances of getting the answer in a less amount of guesses, therefore improving these players statistics and self-confidence for the rest of the day.

Methods and dataset

To achieve the results, I analysed two data sets. The main data set used is a list of all words in the Wordle answer and guesses lists. The data set, created by *Lucas Hohman* in 2022, was created by extracting information from the official Wordle website using JavaScript, meaning the data is accurate and credible.[2] The other data set I used was a list of the frequency of all letters in the alphabet in the English dictionary, created in 2004 by the *Cornell Department of Mathematics*.[3]

To find the results, firstly, a *MATLAB* program was coded, that read from the first data source and calculated the total occurrences of each letter in the alphabet from both the answer and guesses lists. This data was copied from the *MATLAB* program to excel, where the data was neatly organised into tables for the graphs to be generated from. Multiple graphs were created from this set of data, including bar graphs of the most common letters in the answer and guesses lists. The data from the second data set was also copied into excel, and a bar graph comparing the frequency of letters in Wordle to the English dictionary was created.

To find the average amount of vowels and consonants per word in the Wordle answer list, I coded another MATLAB program. The program counted the amount of vowels per word and added this value to an array which kept score of the occurrences for each number of vowels. The average number of consonants is simply the opposite of average amount of vowels. Tables and tree diagrams were created using this data.

Next, to find the 'best' starting word, I used an anagram solver[4] to find all 5-letter words that include the top 5 letters. I coded another Matlab program, which calculated the most common positions of each of these top letters as well. Each of the anagrams were given a score based on the letter's positions, and whichever word had the lowest score, was logically the best word. However, during this process, I found that the top words were both not in the answer list. Therefore, to give a Wordle player a slight chance to get the Wordle in one someday, I added in the sixth most common letter to the anagram solver and repeated the same process.

[1] STATISTIA | Popularity of Wordle by Age Group | January 2022 | J. Clement | Accessed from: https://www.statista.com/statistics/1328012/popularity-of-wordle-by-age-group-usa/
[2] KAGGLE | Wordle Valid Guesses & Answers List | Lucas Hohmann | 2022 | Accessed from: https://www.kaggle.com/datasets/lucashohmann/wordle-valid-guesses-and-answers?resource=download
[3] CORNELL DEPARTMENT of MATHEMATICS | English Letter Frequency | 2004 | Accessed from: https://www.statista.com/statistics/1328012/popularity-of-wordle-by-age-group-usa/
[4] ANAGRAM SOLVER | Word Tips | Accessed from: https://word.tips/anagram-solver/

A similar process was completed again, but with the other letters in the top 10 most common list, to find the best second guess after the first best.

And finally, to find the worst starting word, a list of words with uncommon letters and multiple double letters were compiled. The positions of each letter in each word from the most common letters list were then added up to create a score for each word. The worst word logically has the most uncommon letters and the highest score.

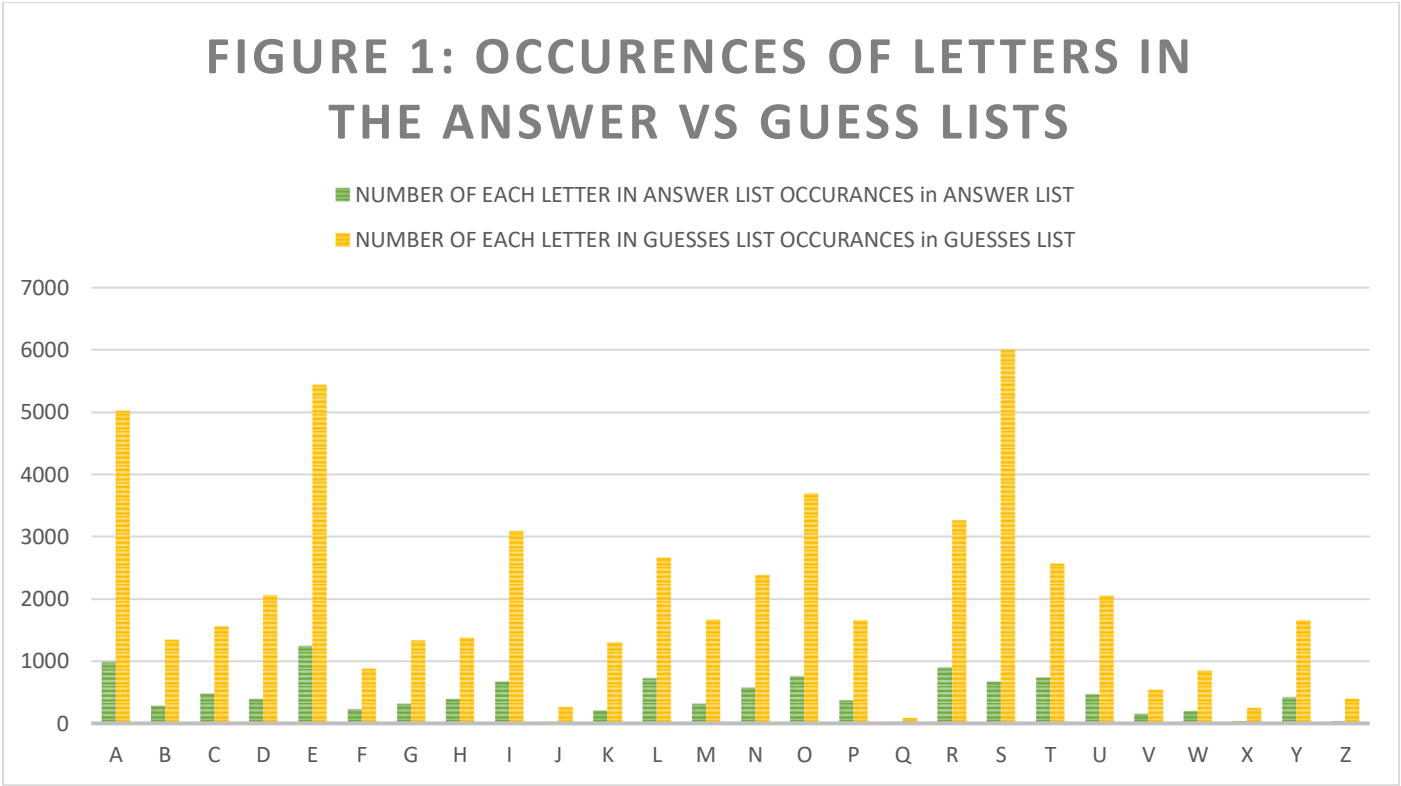
A hypothesis was made before starting the data analysis; "E will be the most common letter in the Wordle list, but the guess and answer lists will slightly differ." This hypothesis was found to be supported and will be discussed further in the results.

Experimental Setup

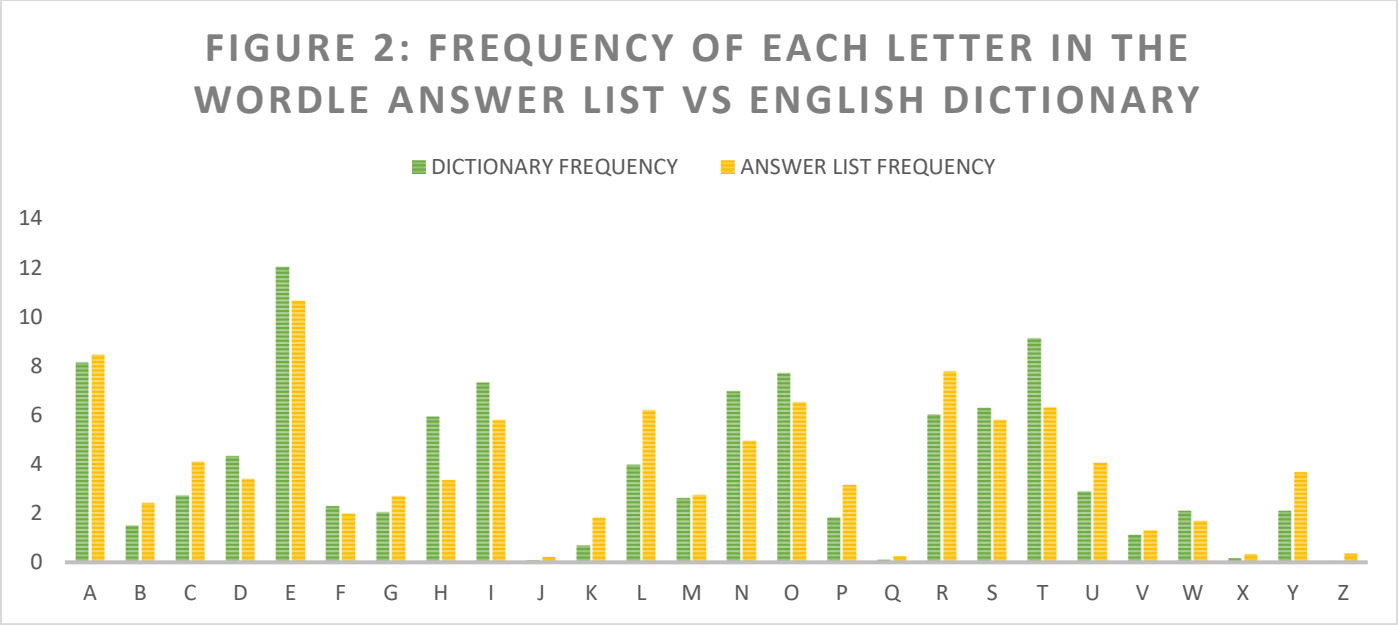
The experimental setup was created using Excel, on a 2023 MacBook Pro. 2309 rows of Data were used to generate more data in Matlab, which was then copied into an Excel file, where tables and graphics were created. The data is organised neatly, and the graphs are colour coded and labelled correctly and effectively. The graphs were created by selecting the specific data and generating a particular type of graph, and the only mathematic formulas used were SUM() and simple addition and division.

Results

Many results and conclusions were formed during my data analysis, the first being the total occurrences of letters in the answer and guess lists. I graphed this information using a bar graph, and as seen in figure 1, there is a clear correlation. However, while E was found to be the most common letter in the answer list, S is the most common letter in the guesses list. These results are mostly due to plural words (e.g. TEARS) being included in the guesses list, but not in the answer list. And as seen, the hypothesis was found to be supported, with E being the most common letter, and a slight difference in common letters between the data sets.



The frequency of letters in the answer list were also compared with the data in my second data set, the English dictionary. Another bar graph was created with this information, as seen in figure 2:



The average amount of vowels and consonants per word in the Wordle answer list was also found, which I displayed using tree maps, seen in figures 3 and 4:

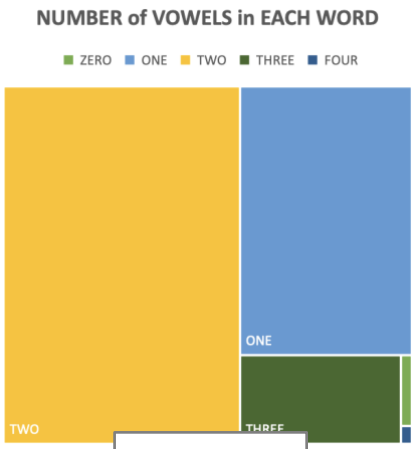


FIGURE 3

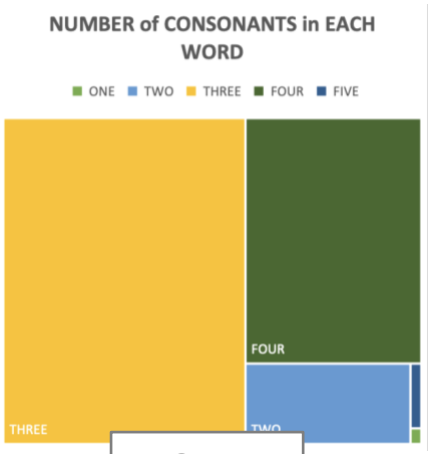


FIGURE 4

The next step in my investigation was to find the most logical starting words, which were found to be ORATE and ROATE. However, as stated in the methodology, these words were found not to be in the answer list. Therefore, I added in the sixth most common letter, leading to ALERT being founded as the most logical starting word, answering my research question of 'What is the best Wordle starting word?'.

I repeated a similar process to find the second-best starting word to use after ALERT, which was found to be either SCION or SONIC, both calculated to have the same score due to their letter's positions.

And lastly, I found the worst starting word; FUZZY. However, this method does not take into account double letters. In the future, I would be to code a program that calculates the average percentage of receiving a green or yellow letter using the guess.

Conclusions and Discussion

During my data investigation, there was one limitation I faced during the analysis process, which was graphing the most common letters. One student during the proposal feedback suggested to create a word map to better visualise the most common letters. However, as there are over 2309 words, and therefore 11,545 words, it was not possible to

create a word map by typing up all letters in the given time frame. I could also not find a website online where you enter the amount of occurrences of each letter/word to create a word map. Because of this limitation, I had to only graph this information using a bar graph.

On the other hand, there are a few changes I would make to my results if I had more time to work on my project. One of them, as I stated before, would be to code a program using MATLAB that reads from the first data source and found an average percentage of receiving green and yellow tiles from a particular starting guess. This would be a more accurate way of finding the best and worst starting guesses.

Finally, there were many conclusions found from my data analysis. It was found that ALERT is the most logical Wordle first guess, while SONIC or SCION are the best second guesses. It was also found that there is a slight difference in common letters from the GUESS and ANSWER lists, and there is also a correlation between common letters in Wordle and the English dictionary. Additionally, FUZZY is one of the worst starting words, and there is an average of TWO vowels and THREE consonants per word in the Wordle answer list.

References

- **STATISTIA** | Popularity of Wordle by Age Group | January 2022 | J. Clement | Accessed from: https://www.statista.com/statistics/1328012/popularity-of-wordle-by-age-group-usa/
- **KAGGLE** | Wordle Valid Guesses & Answers List | Lucas Hohmann | 2022 | Accessed from: https://www.kaggle.com/datasets/lucashohmann/wordle-valid-guesses-and-answers?resource=download
- **STATISTA** | Most Popular Google Searches by Country | Federica Laricchia | Dec 2022 | Accessed from: https://www.statista.com/statistics/1350923/most-popular-google-searches-by-country/
- **CORNEL DEPARTMENT of MATHEMATICS** | English Letter Frequency | 2004 | Accessed from: https://www.statista.com/statistics/1328012/popularity-of-wordle-by-age-group-usa/
- **ANAGRAM SOLVER** | Word Tips | Accessed from: https://word.tips/anagram-solver/

Appendix **(max 2 pages)**

TOPIC	
FEEDBACK	***HOW IT WAS ADDRESSED***
EDWARD: "Project may be short - answer may be found quickly."	2 students suggested for me to make sure my research is in-depth enough. I made sure it was by researching into different sub-questions, including average amount of vowels, worst starting word, comparing the data with the English dictionary, etc.
ISABELLA: "Ensuring that research is in-depth enough"	
JAIDEN: "it's been done."	9 students mentioned that my data had already been found. While this is true, many others found their data using a different method to mine. I also decided to research into other aspects of Wordle as well, which hadn't been researched into before. I also looked into the second best word, as suggested by Bryce's feedback.
JOSH: "There is already a lot of research into this"	
RORY: "Has been done before so could be difficult to make original"	
GEORGIA: "Similar questions may have already been researched into"	
VASILIS: "Question already been researched – best starting word already solved by neural network"	
KHANH: "The question may have already been done by other people."	
JACK: "Has this already been done?"	
KIAN: "similar studies have been done previously"	
BRYCE: "Many other people have already done this, be sure to do something different with your project than others have before, i.e. what about the second? Third? Fourth? (assuming all Grey or something similar)"	

VISUALISATION	
FEEDBACK	***HOW IT WAS ADDRESSED***
PHOENIX: "Maybe use different types of graphs for most common words, e.g word map"	As discussed in the discussion and limitations above, I tried to address this feedback by creating a word map, however, this was found to not be possible in the time frame. Therefore, I decided to use a different type of graph, a tree map instead, for the average amount of vowels and consonants.

DATA SOURCE	
FEEDBACK	***HOW IT WAS ADDRESSED***
AGNES: "Common positions and common letters may not match up"	As stated in the methods, to find the best starting word, I found the anagrams of the top letters, and then gave each a score based on the letter's positions. It was true that the best letters in the anagrams did not have all the best positions, but the word I found was the best out of all anagrams.
ROSELYN: "The app may update and change particular features so the results may not be relevant by a future time"	As of currently, there has only been one removal of words in the answer list, which was only to remove "offensive" words in November 2022. Therefore, it is probably unlikely the list will change any time soon. Even if it does, this is an easy fix and the project can be redone.

ANALYSIS	
FEEDBACK	***HOW IT WAS ADDRESSED***
EDWARD: “Best next moves after first guess?”	Due to these two student’s suggestions, I decided to find a second word after the second guess, which was found to be SCION or SONIC.
RORY: “Best second word based on first guess”	
JAIDEN: “Could this be extended to some of the other -dle games? (future – found second guess which could be useful in _____ for speedruns)”	Three students suggested investigating data of other Wordle variations as well. However, I ran out of time in my data analysis to analyse these variations. But, in the future, this would definitely be interesting to investigate. On the other hand, I did find a second best starting word. As Jaiden suggested, this would be useful for speedruns or in the Wordle variation, Thirtle.
JUSTIN: “Could also investigate other wordle variations.”	
GEORGIA: “Could extend project to look at 6 and 7 letter words and see if the data differs dramatically.”	
JOSH: “What are the worst words?”	Due to these three student’s suggestions, I decided to find the “worst” starting word, which was found to be FUZZY.
VASILIS: “What is the worst starting word(s)?”	
KIAN: “What the worst word also is”	
RYLAN: “Is there any correlation between wordle letter frequency and the frequency of letters spoken in everyday English?”	I decided to investigate the correlation between the Wordle most common list, and the English dictionary’s most common words. And as stated above, there was found to be a correlation!
ANGUS: “May ruin the game of wordle”	Two students suggested that my results might affect the gaming experience of players. However, my investigation’s findings don’t have to be used. Wordle players can have the choice to take my findings or choose their own word to use as a starting guess.
JUSTIN: “may ruin gaming experience for players. “	
WILLIAM: “Does correct position of a letter indicate the best starting word? Deep learning and simulation to solve”	Because of these three students, I took this feedback into account and coded a program to find the common positions for each letter, in order to find the best starting guess. This is a more accurate and logical way of finding the best starting guesses. There were also multiple best starting guesses found using the anagram solver, as Khanh stated. However, by looking into common positions too, I limited this to the words with the best scores.
RYLAN: “Using frequency data only, is it possible to definitively conclude what the best word is? (used common positions too)”	
KHANH: “Are there multiple best starting guesses”	
BRYCE: “More than just the first word? Maybe look at the most common starting words currently used”	During my investigation, I did look into a second starting word. However, I ran out of time to look into most common starting words currently used, but this would be a good idea for future analysis.