WHAT IS THE BEST WORDLE STARTING WORD?

Grand Challenges PRESENTATION - Georgia Mason

INTRODUCTION

In this presentation, I will be sharing the data found by analyzing Wordle, including:

- Most common letters
- Average amount of vowels
- Comparison with the English dictionary
- The best 2 starting words
- The worst starting word

Relevance:

- Using a better starting word, players have an increased chance to get the answer in a less amount of guesses.

METHODOLOGY

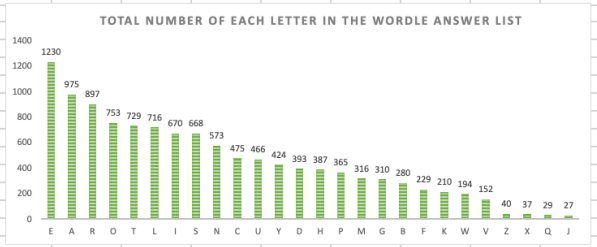
Coded a MATLAB program that read from the data source & found the number of occurrences of each letter in the alphabet:

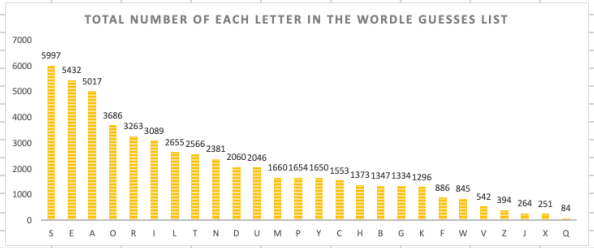
```
clear;
clc;

alphabet = zeros(1, 26);

file = fopen('valid_guesses.txt', 'rb'); %reads from file
fseek(file, 0, 'eof');
fileSize = ftell(file);
frewind(file);
data = fread(file, fileSize, 'uint8');
numLines = sum(data == 10);
fclose(file);

file = fopen('valid_guesses.txt');

for i = 1:numLines
    line = fgetl(file);  % read one line excluding newline character
    word = char(line);
    for i = 1:5
        if word(i) == 'a'
            alphabet(1) = alphabet(1) + 1;
        elseif word(i) == 'b'
            alphabet(2) = alphabet(2) + 1;
        elseif word(i) == 'c'
            alphabet(3) = alphabet(3) + 1;
        elseif word(i) == 'd'
            alphabet(4) = alphabet(4) + 1;
        elseif word(i) == 'e'
            alphabet(5) = alphabet(5) + 1;
        elseif word(i) == 'f'
            alphabet(6) = alphabet(6) + 1;
        elseif word(i) == 'g'
            alphabet(7) = alphabet(7) + 1;
elseif word(i) == 'h'
    alphabet(8) = alphabet(8) + 1;
elseif word(i) == 'i'
    alphabet(9) = alphabet(9) + 1;
elseif word(i) == 'j'
    alphabet(10) = alphabet(10) + 1;
elseif word(i) == 'k'
    alphabet(11) = alphabet(11) + 1;
elseif word(i) == 'l'
    alphabet(12) = alphabet(12) + 1;
elseif word(i) == 'm'
    alphabet(13) = alphabet(13) + 1;
elseif word(i) == 'n'
    alphabet(14) = alphabet(14) + 1;
elseif word(i) == 'o'
    alphabet(15) = alphabet(15) + 1;
elseif word(i) == 'p'
    alphabet(16) = alphabet(16) + 1;
elseif word(i) == 'q'
    alphabet(17) = alphabet(17) + 1;
elseif word(i) == 'r'
    alphabet(18) = alphabet(18) + 1;
elseif word(i) == 's'
    alphabet(19) = alphabet(19) + 1;
elseif word(i) == 't'
    alphabet(20) = alphabet(20) + 1;
elseif word(i) == 'u'
            alphabet(20) = alphabet(20) + 1;
        elseif word(i) == 'u'
            alphabet(21) = alphabet(21) + 1;
        elseif word(i) == 'v'
            alphabet(22) = alphabet(22) + 1;
        elseif word(i) == 'w'
            alphabet(23) = alphabet(23) + 1;
        elseif word(i) == 'x'
            alphabet(24) = alphabet(24) + 1;
        elseif word(i) == 'y'
            alphabet(25) = alphabet(25) + 1;
        elseif word(i) == 'z'
            alphabet(26) = alphabet(26) + 1;
        end
    end
end

for i = 1:26
    fprintf("%d\n", alphabet(i));
end
```

OVERVIEW of the EXCEL SPREADSHEET:

NUMBER OF EACH LETTER IN ANSWER LIST	
LETTER	OCCURANCES in ANSWER LIST
E	1230
A	975
R	897
O	753
T	729
L	716
I	670
S	668
N	573
C	475
U	466
Y	424
D	393
H	387
P	365
M	316
G	310
B	280
F	229
K	210
W	194
V	152
Z	40
X	37
Q	29
J	27

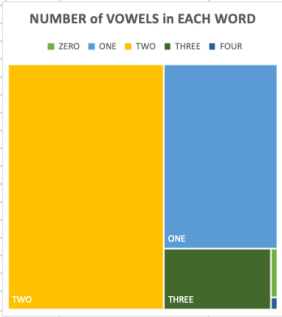
NUMBER OF EACH LETTER IN GUESSES LIST	
LETTER	OCCURANCES in GUESSES LIST
S	5997
E	5432
A	5017
O	3686
R	3263
I	3089
L	2655
T	2566
N	2381
D	2060
U	2046
M	1660
P	1654
Y	1650
C	1553
H	1373
B	1347
G	1334
K	1296
F	886
W	845
V	542
Z	394
J	264
X	251
Q	84

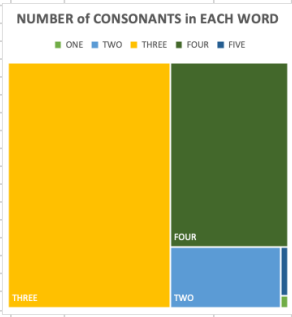


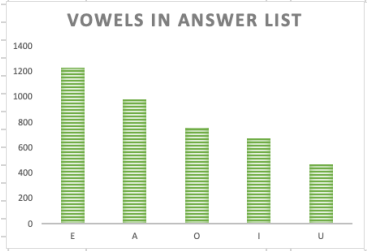


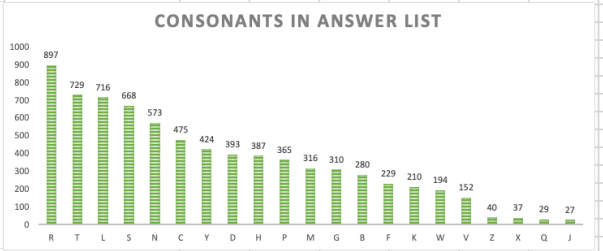
NUMBER of VOWELS in EACH WORD		
VOWELS	NUMBER of OCCURANCES	PERCENTAGE
ZERO	12	0.5%
ONE	732	31.7%
TWO	1336	57.9%
THREE	226	9.8%
FOUR	3	0.1%
	2309	

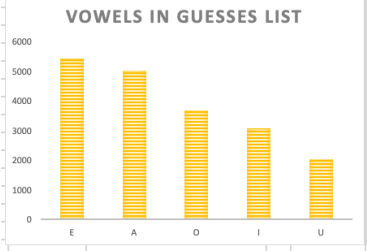
NUMBER of CONSONANTS in EACH WORD		
VOWELS	NUMBER of OCCURANCES	PERCENTAGE
ONE	3	0.1%
TWO	226	9.8%
THREE	1336	57.9%
FOUR	732	31.7%
FIVE	12	0.5%
	2309	

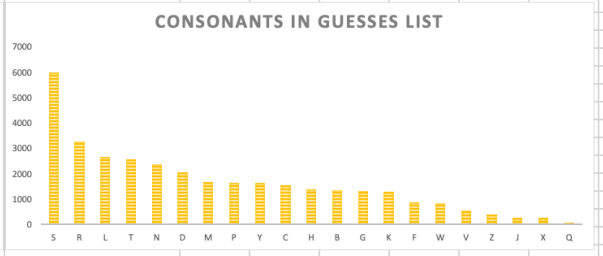


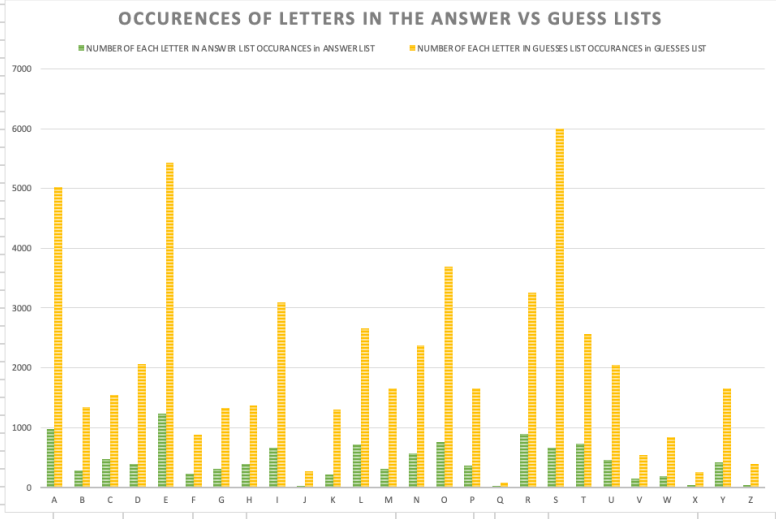




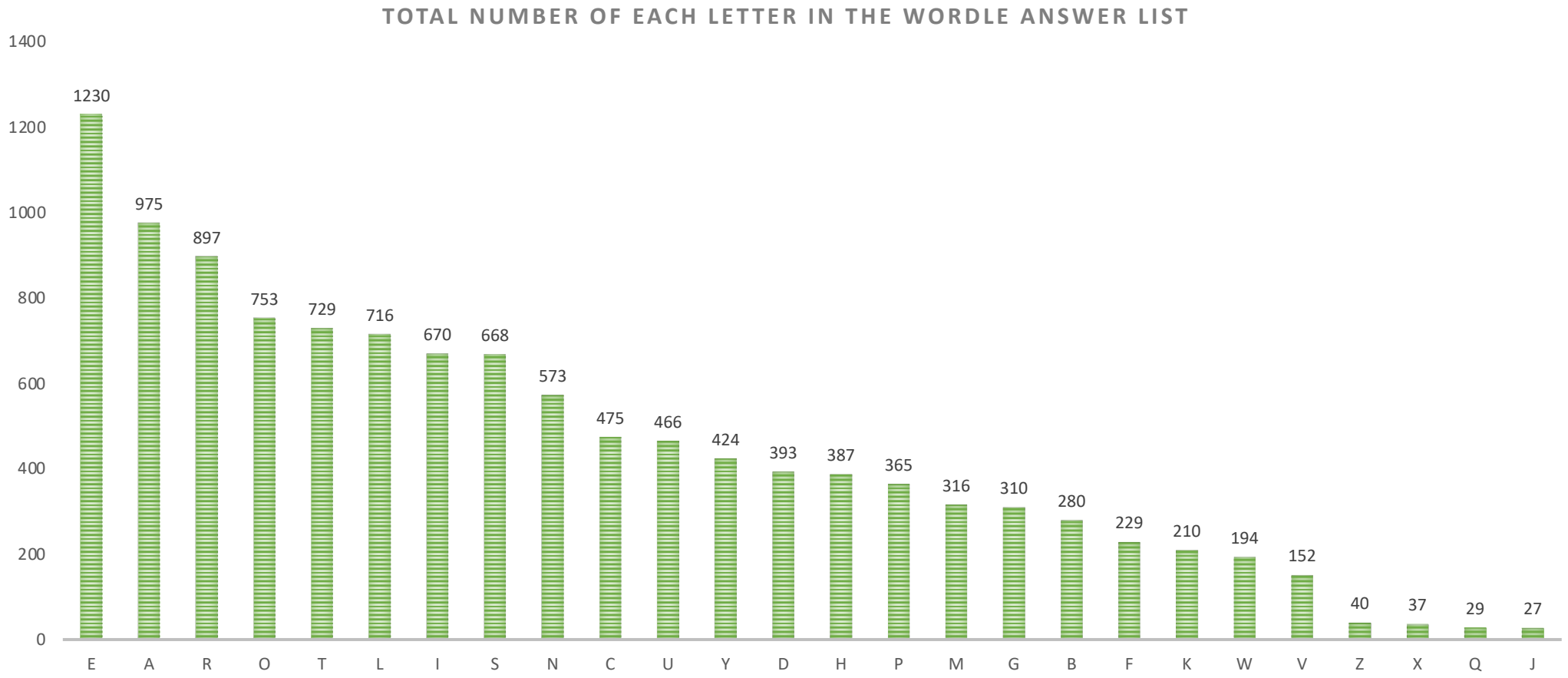




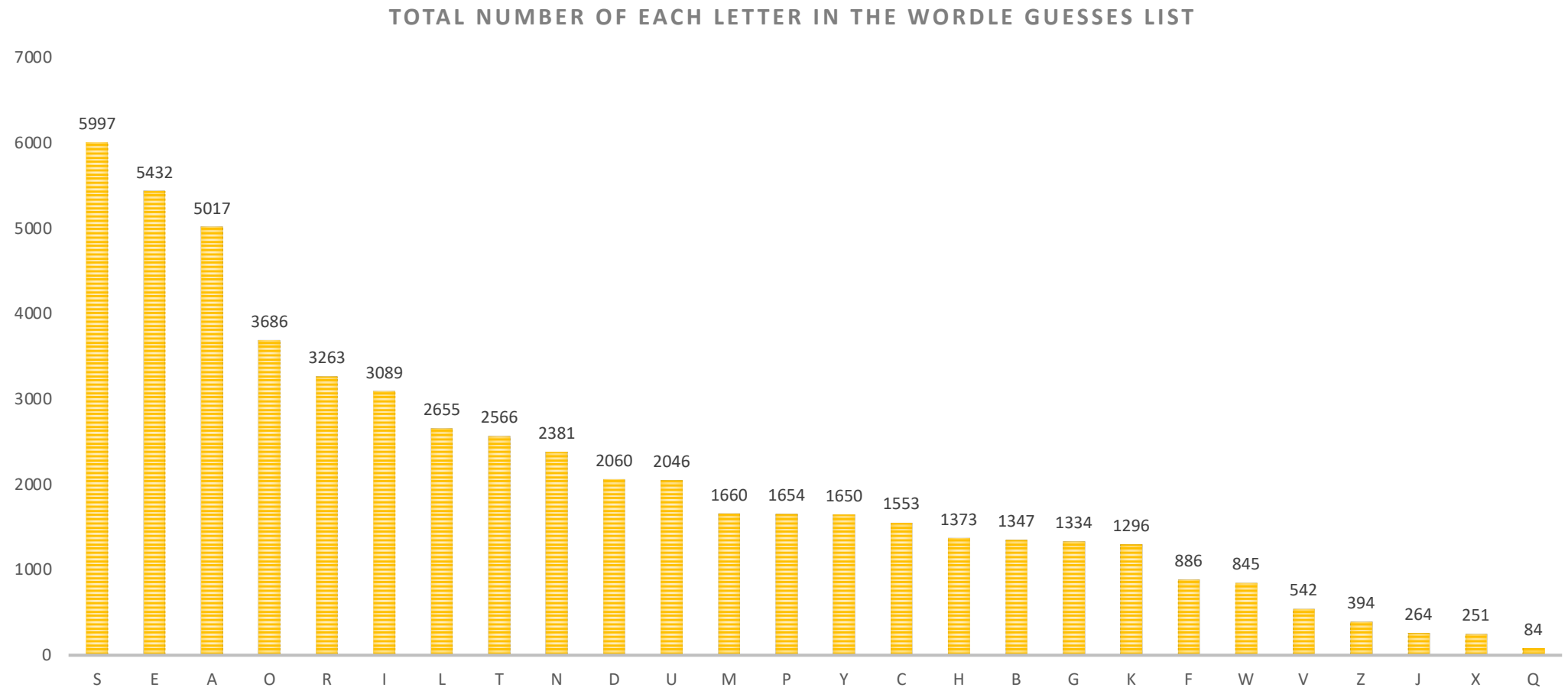




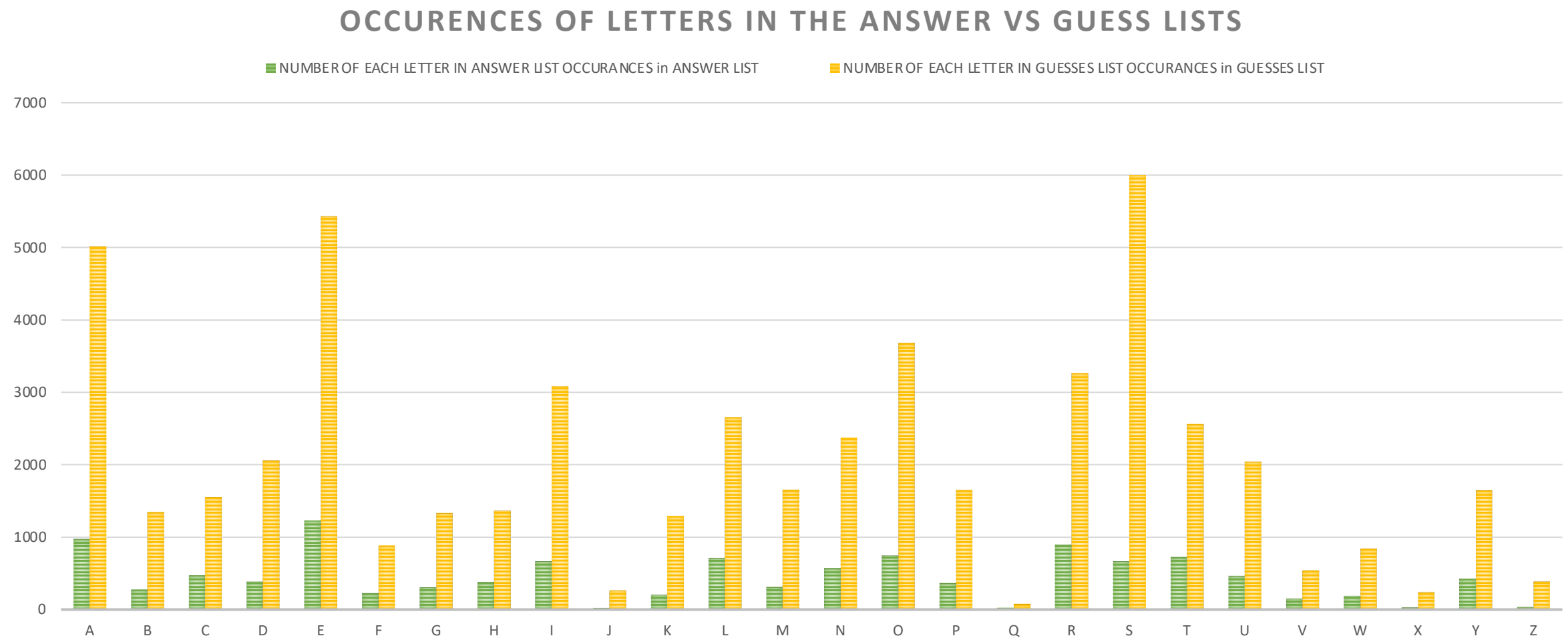
GRAPHICS


TOTAL NUMBER OF EACH LETTER IN THE WORDLE ANSWER LIST
1400
1200
1000
800
600
400
200
0
1230
975
897
753
729
716
670
668
573
475
466
424
393
387
365
316
310
280
229
210
194
152
40
37
29
27
E
A
R
O
T
L
I
S
N
C
U
Y
D
H
P
M
G
B
F
K
W
V
Z
X
Q
J

GRAPHICS

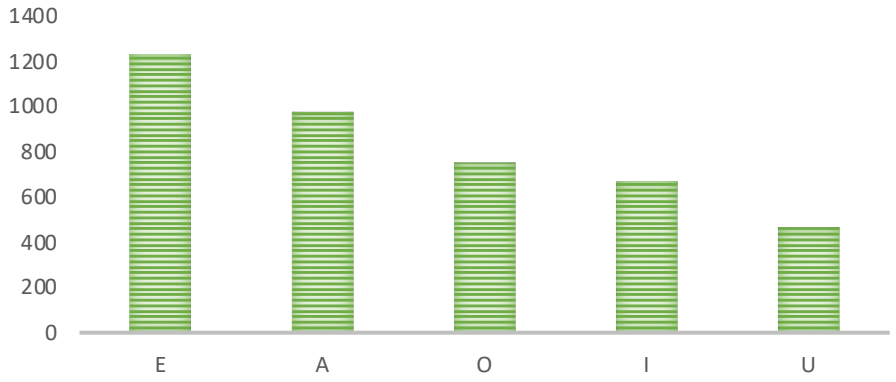

TOTAL NUMBER OF EACH LETTER IN THE WORDLE GUESSES LIST
7000
6000
5000
4000
3000
2000
1000
0
5997
5432
5017
3686
3263
3089
2655
2566
2381
2060
2046
1660
1654
1650
1553
1373
1347
1334
1296
886
845
542
394
264
251
84
S
E
A
O
R
I
L
T
N
D
U
M
P
Y
C
H
B
G
K
F
W
V
Z
J
X
Q

GRAPHICS

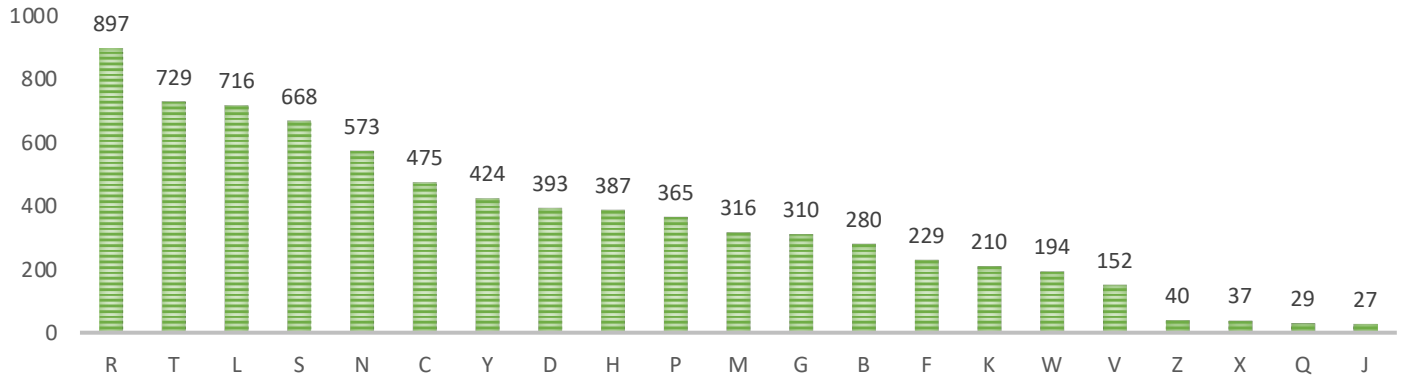

OCCURENCES OF LETTERS IN THE ANSWER VS GUESS LISTS
NUMBER OF EACH LETTER IN ANSWER LIST OCCURANCES in ANSWER LIST
NUMBER OF EACH LETTER IN GUESSES LIST OCCURANCES in GUESSES LIST
7000
6000
5000
4000
3000
2000
1000
0
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

GRAPHICS

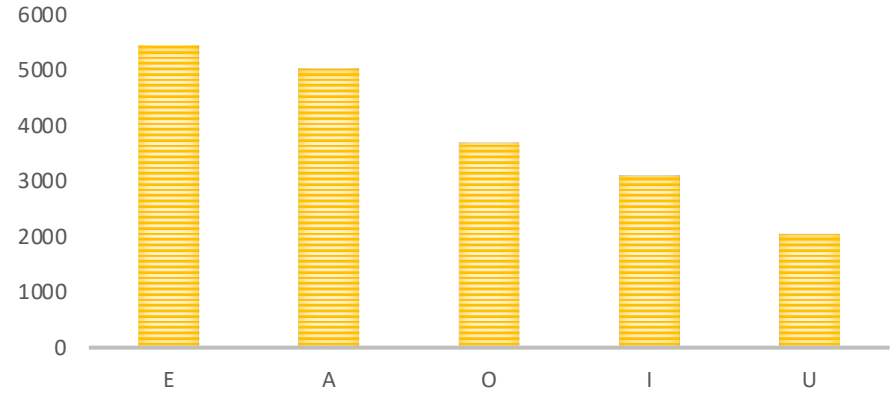
VOWELS IN ANSWER LIST



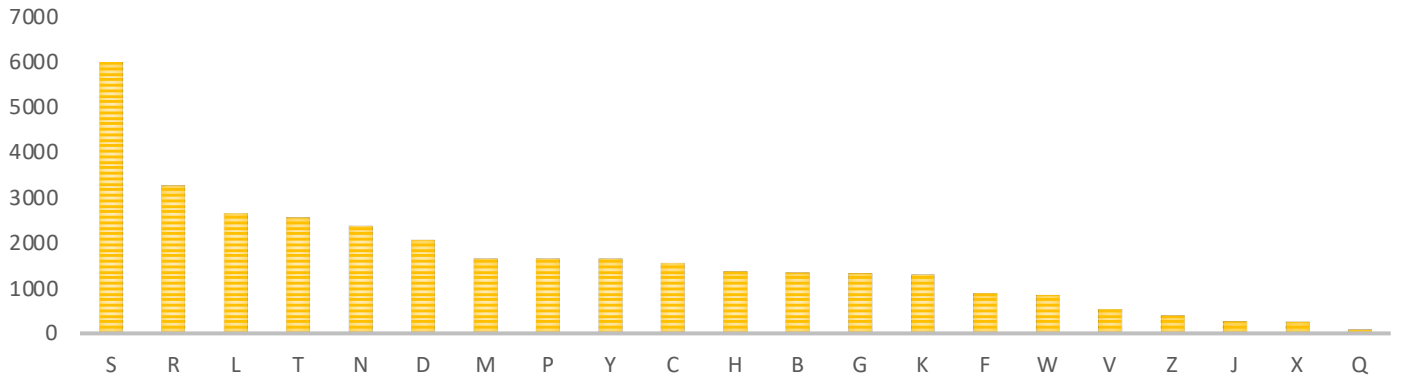
CONSONANTS IN ANSWER LIST



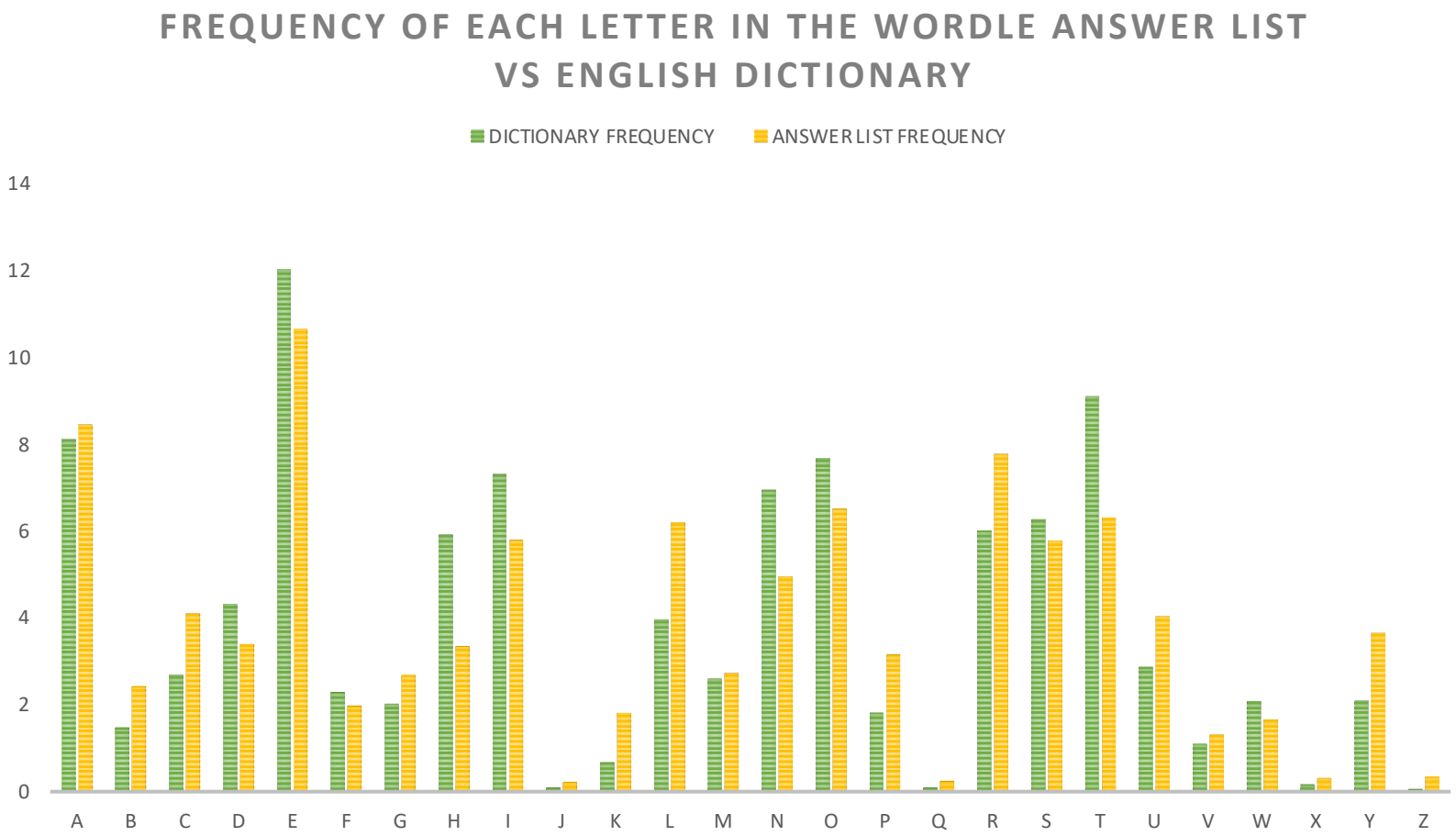
VOWELS IN GUESSES LIST



CONSONANTS IN GUESSES LIST

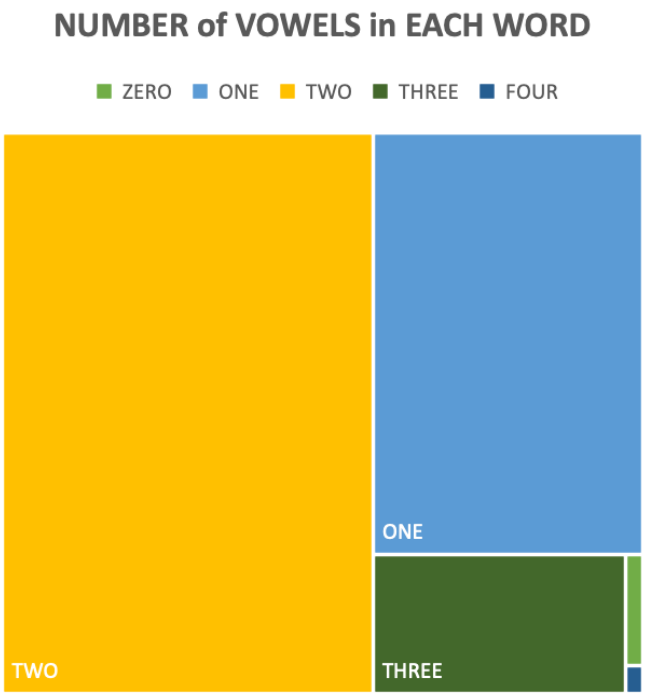


DATA SET USED: *Cornell Department of Mathematics – 2004 – English Letter Frequency*



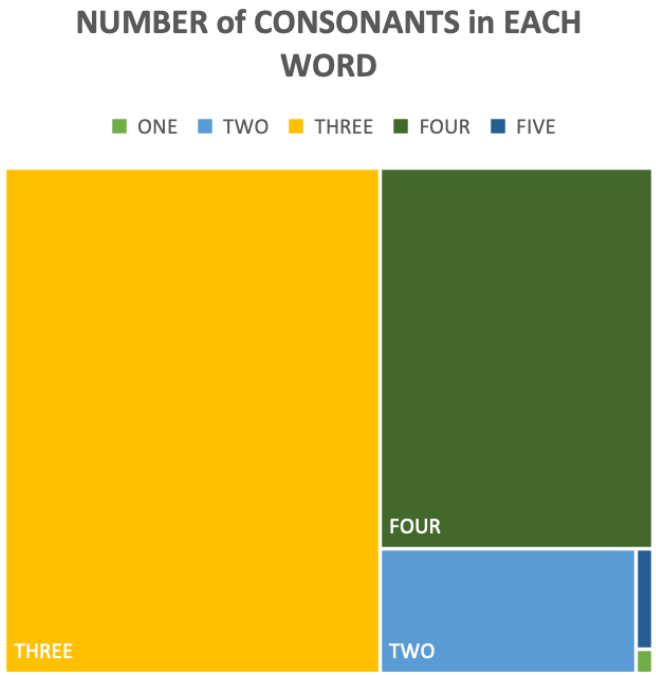
DICTIONARY		ANSWER LIST	
LETTER	FREQUENCY	LETTER	FREQUENCY
A	8.12	A	8.45
B	1.49	B	2.43
C	2.71	C	4.11
D	4.32	D	3.4
E	12.02	E	10.65
F	2.3	F	1.98
G	2.03	G	2.69
H	5.92	H	3.35
I	7.31	I	5.8
J	0.1	J	0.23
K	0.69	K	1.82
L	3.98	L	6.2
M	2.61	M	2.74
N	6.95	N	4.96
O	7.68	O	6.52
P	1.82	P	3.16
Q	0.11	Q	0.25
R	6.02	R	7.77
S	6.28	S	5.79
T	9.1	T	6.31
U	2.88	U	4.04
V	1.11	V	1.32
W	2.09	W	1.68
X	0.17	X	0.32
Y	2.11	Y	3.67
Z	0.07	Z	0.35

AVERAGE number of VOWELS + CONSONANTS per word in the answer list:



NUMBER of CONSONANTS in EACH WORD		
VOWELS	NUMBER of OCCURANCES	PERCENTAGE
ONE	3	0.1%
TWO	226	9.8%
THREE	1336	57.9%
FOUR	732	31.7%
FIVE	12	0.5%
	2309	

NUMBER of VOWELS in EACH WORD		
VOWELS	NUMBER of OCCURANCES	PERCENTAGE
ZERO	12	0.5%
ONE	732	31.7%
TWO	1336	57.9%
THREE	226	9.8%
FOUR	3	0.1%
	2309	



FINDING THE BEST STARTING WORD

1. Coded a program to find most common positions of common letters:

A COMMON POSITIONS	
POSITION	OCCURANCES
3	306
2	304
4	162
1	140
5	63

E COMMON POSITIONS	
POSITION	OCCURANCES
5	422
4	318
2	241
3	177
1	72

R COMMON POSITIONS	
POSITION	OCCURANCES
2	267
5	212
3	163
4	150
1	105

O COMMON POSITIONS	
POSITION	OCCURANCES
2	279
3	243
4	132
5	58
1	41

T COMMON POSITIONS	
POSITION	OCCURANCES
5	253
1	149
4	139
3	111
2	77

2. Found the anagrams of common letters and found which word has the letters in the most common positions:

OATER:	
O	5
A	2
T	4
E	2
R	2
TOTAL	15

ORATE:	
O	5
R	1
A	1
T	3
E	1
TOTAL	11

ROATE:	
R	5
O	1
A	1
T	3
E	1
TOTAL	11

```
clear;
clc;

position = zeros(1, 5);

file = fopen('valid_answers.txt', 'rb'); %reads from file
fseek(file, 0, 'eof');
fileSize = ftell(file);
frewind(file);
data = fread(file, fileSize, 'uint8');
numLines = sum(data == 10);
fclose(file);

file = fopen('valid_answers.txt');

for i = 1:numLines
    line = fgetl(file);  % read one line excluding newline character
    word = char(line);
    num_of_vowels = 0;
    for i = 1:5
        if word(i) == 'c'
            if i == 1;
                position(1) = position(1) + 1;
            elseif i == 2
                position(2) = position(2) + 1;
            elseif i == 3
                position(3) = position(3) + 1;
            elseif i == 4
                position(4) = position(4) + 1;
            elseif i == 5
                position(5) = position(5) + 1;
            end
        end
    end
end

for i = 1:5
    fprintf("%d\n", position(i));
end
```

I added in the sixth most common letter to the anagram solver, and condensed this list to words that ARE in the answer list.

I repeated the same process, and found that ALERT is the best starting word. This word has also not been the answer yet, so it is possible to get this word in one some day.

L COMMON POSITIONS	
POSITION	OCCURANCES
2	200
4	162
5	155
3	112
1	87

ALERT:	
A	4
L	1
E	4
R	4
T	1
TOTAL	14

ALTER:	
A	4
L	1
T	4
E	2
R	2
TOTAL	15

LATER:	
L	5
A	2
T	4
E	2
R	2
TOTAL	15

I then found another word to use after ALERT.

I repeated the same process as finding the best word, but using the letters, S, I, N, C, and O, as this letter was not used in the best guess.

The two anagrams in the answer list found both achieved the same score, meaning either could be used in a second guess.

S COMMON POSITIONS	
POSITION	OCCURANCES
1	365
4	171
3	80
5	36
2	16

I COMMON POSITIONS	
POSITION	OCCURANCES
3	266
2	201
4	158
1	34
5	11

N COMMON POSITIONS	
POSITION	OCCURANCES
4	182
3	137
5	130
2	87
1	37

C COMMON POSITIONS	
POSITION	OCCURANCES
1	198
4	150
3	56
2	40
5	31

SCION	
S	1
C	4
I	1
O	3
N	3
TOTAL	12

SONIC	
S	1
O	1
N	2
I	3
C	5
TOTAL	12

FINDING THE WORST WORD

I compiled a list of the words with very uncommon letters and multiple double letters, then added up the score of the position in the most common letters list.

The highest score, with the letters that are the least common was FUZZY.

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.

JAZZY	
J	26
A	2
Z	23
Z	23
Y	12
TOTAL	86

XYLYL	
X	24
Y	12
L	6
Y	12
L	6
TOTAL	60

FUFFY	
F	22
U	11
F	22
F	22
Y	12
TOTAL	89

FUZZY	
F	22
U	11
Z	23
Z	23
Y	12
TOTAL	91

JEEZE	
J	26
E	1
E	0
Z	23
E	0
TOTAL	50

PHPHT	
P	15
H	14
P	25
H	14
T	5
TOTAL	73

QAJAQ	
Q	25
A	2
J	26
A	2
Q	25
TOTAL	80

ZIZIT	
Z	23
I	7
Z	23
I	7
T	8
TOTAL	68

MAMMA	
M	16
A	2
M	16
M	16
A	2
TOTAL	52

CONCLUSIONS

- ALERT is the best first guess
- SONIC or SCION is the best second guess
- There is a slight difference in common letters from the GUESS and ANSWER lists
- There is a correlation between common letters in Wordle and the English dictionary
- FUZZY is one of the worst starting words
- There is an average of TWO vowels and THREE consonants per word

FOR LISTENING ☺

BIBLIOGRAPHY

CORNEL DEPARTMENT of MATHEMATICS – English Letter Frequency - 2004

- https://www.statista.com/statistics/1328012/popularity-of-wordle-by-age-group-usa/

KAGGLE - Wordle Valid Guesses & Answers List - Lucas Hohmann – 2022

- https://www.kaggle.com/datasets/lucashohmann/wordle-valid-guesses-and-answers?resource=download

ANAGRAM SOLVER – Word Tips

- https://word.tips/anagram-solver/