WORDLE

DATA ANALYTICS PROJECT

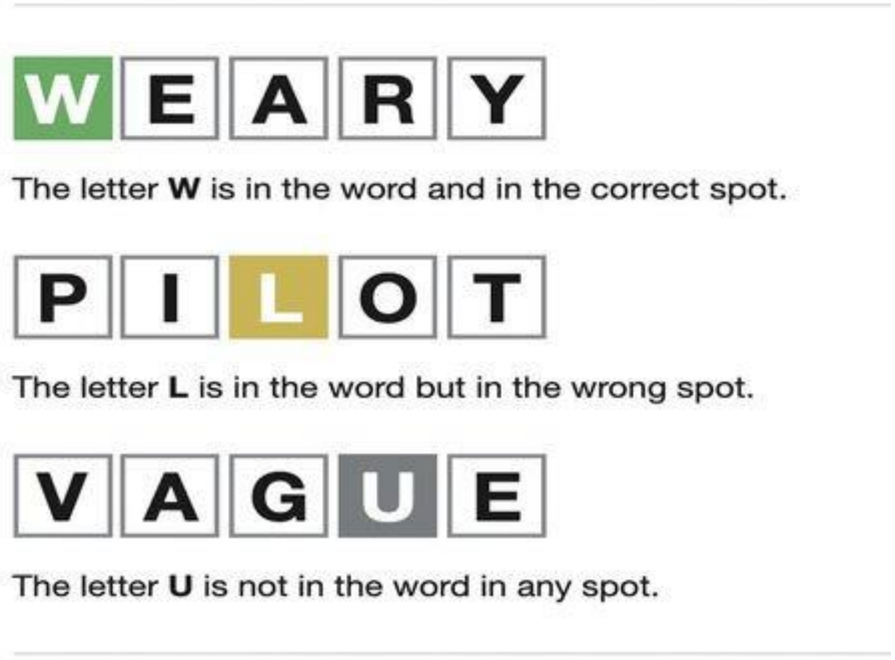
Perla, Berkelee, Tingna, Alice, Daniella

What is Wordle?

Wordle was created by Josh Wardle, a software engineer from New York. He created a prototype in 2013 and dusted it off during the pandemic for his partner, who likes playing word games.

He sold it to NYT in 2022 for over $1mil

It follows simple rules and the word resets daily at 12am.



W E A R Y

The letter **W** is in the word and in the correct spot.

P I L O T

The letter **L** is in the word but in the wrong spot.

V A G U E

The letter **U** is not in the word in any spot.

Data Sourcing

Dictionary Database:
https://www.kaggle.com/datasets/dfydata/the-online-plain-text-english-dictionary-opted

Wordle Allowed Guesses:
https://gist.github.com/cfreshman/cdcdf777450c5b5301e439061d29694c

Wordle Answers:
https://gist.github.com/cfreshman/a03ef2cba789d8cf00c08f767e0fad7b

Data Review

Dictionary Database:

- 4 Columns:
    - Word
    - Count (characters in word)
    - POS (Part of Speech)
    - Definition

Wordle Allowed Guesses:

- Txt file. Single column of words separated by ‘\n’

Wordle Answers:

- Txt file. Single column of words separated by ‘\n’

Data Cleansing

Creating a “5-letter dictionary”

- Used Excel to filter words with max 5 characters
- Concatenated POS into one row

-We used this dictionary to help us reduce run time so when we wanted to find more info on words (like POS) we could more easily loop through the words in the dictionary.

Packages Used:
Pandas, numpy, beautifulSoup, seaborn, requests

General EDA

What are the most popular/unpopular letters?

What is the frequency of multiple occurrences of letters in Wordle words?

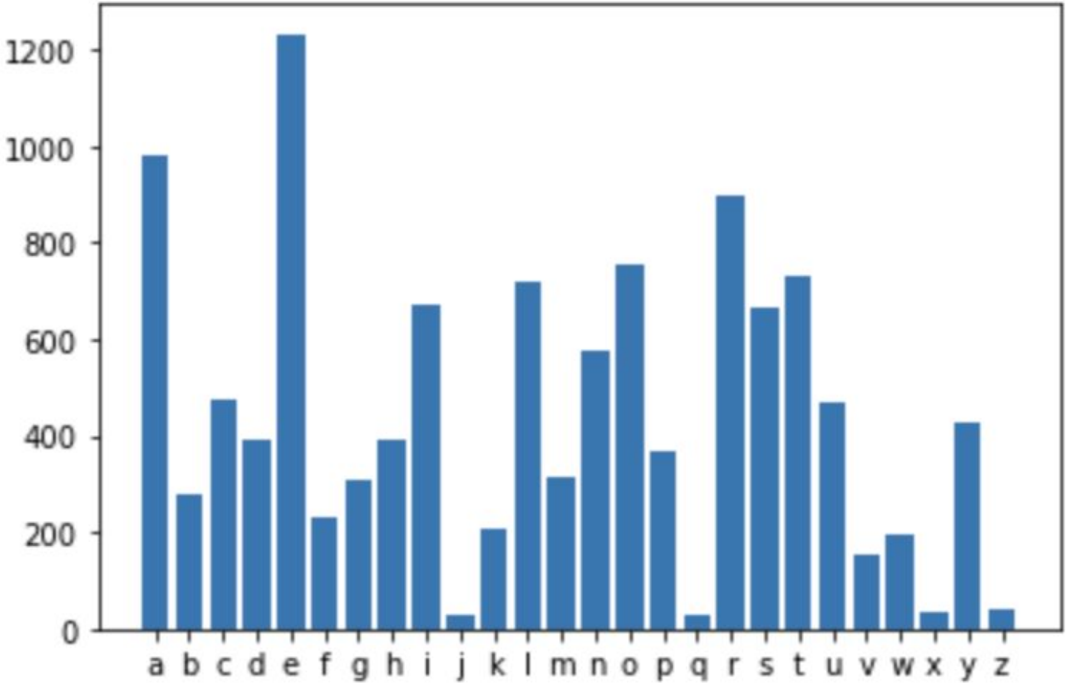
What is the most common type (adj, verb, noun, etc) of word?

What is the “**BEST**” Wordle starting word?

Letter Popularity

How we found the most popular letters:

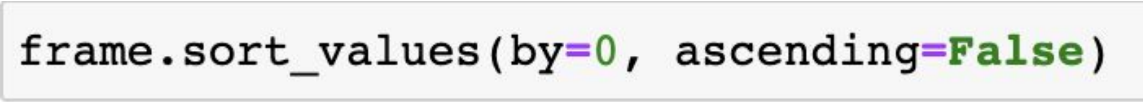
- Divided words by frequency of letter
- Created sums of the columns to get the total number of occurrences for each letter



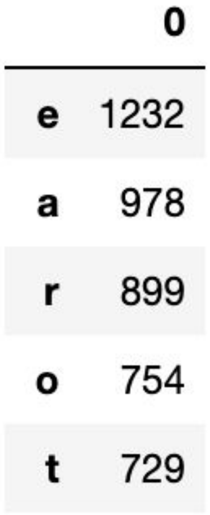
Top Five Letter Popularity

How we found the top five most popular letters:

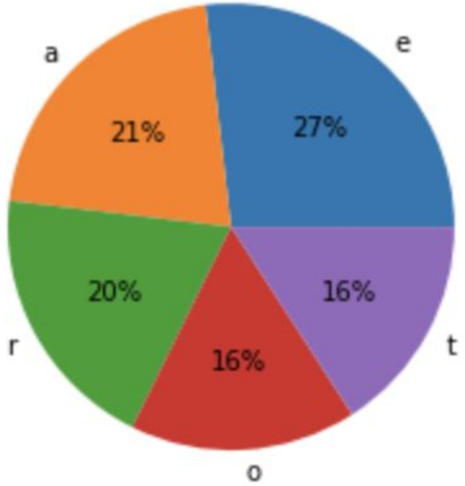
- Ordered the frequency values of letters from most to least (desc)
- Selected the first five rows



```
frame.sort_values(by=0, ascending=False)
```



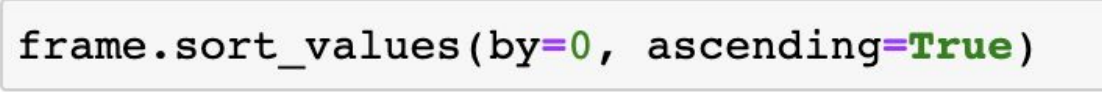
	0
e	1232
a	978
r	899
o	754
t	729



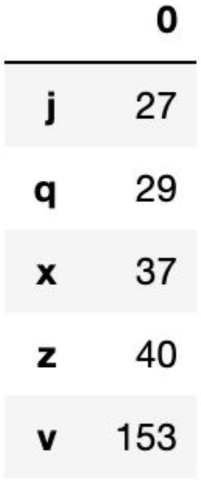
Bottom Five Letter Popularity

How we found the bottom five letter popularity:

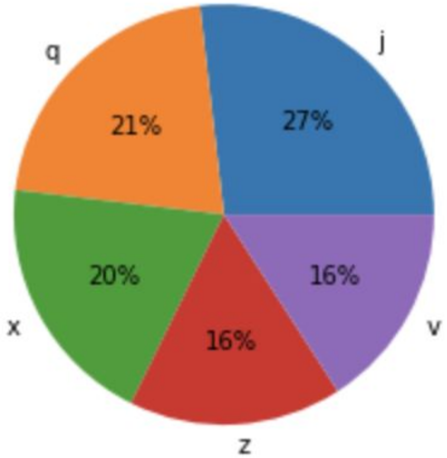
- Ordered the frequency values of letters from least to most (asc)
- Selected the first five rows



```
frame.sort_values(by=0, ascending=True)
```



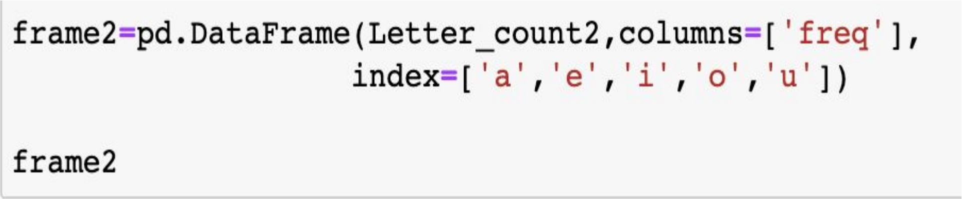
	0
j	27
q	29
x	37
z	40
v	153



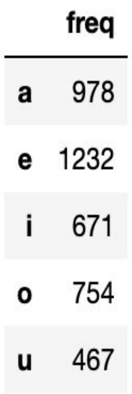
Vowel exploration

How we found the vowel frequency:

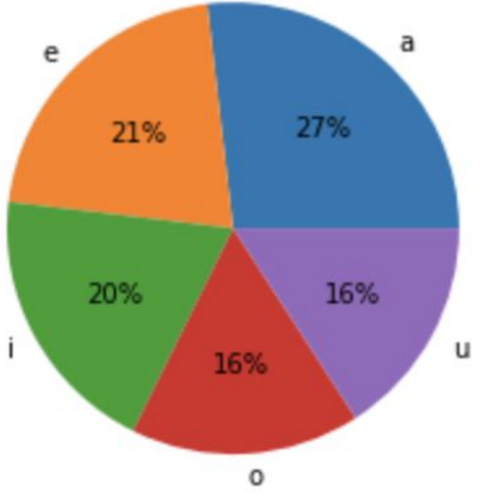
- Made the index= to vowels “a,” “e,” “i,” “o,” “u” > output= frequency of vowels



```
frame2=pd.DataFrame(Letter_count2,columns=['freq'],
                        index=['a','e','i','o','u'])

frame2
```



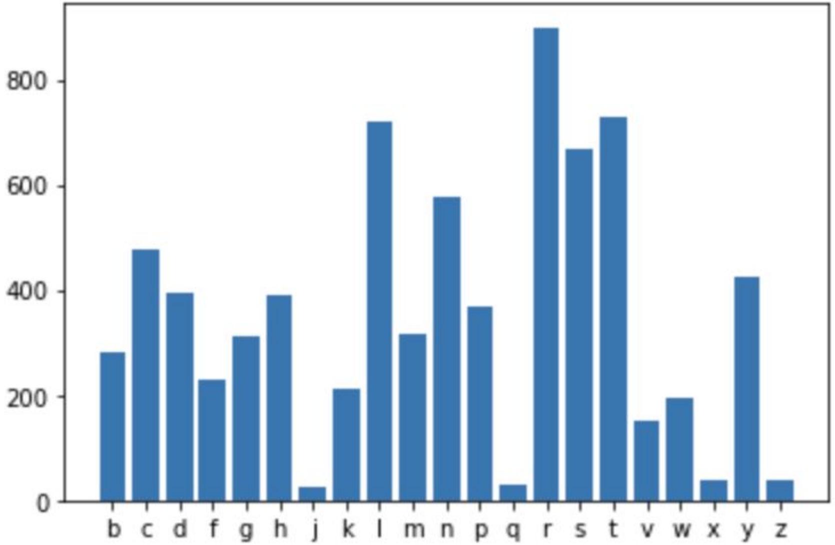
	freq
a	978
e	1232
i	671
o	754
u	467

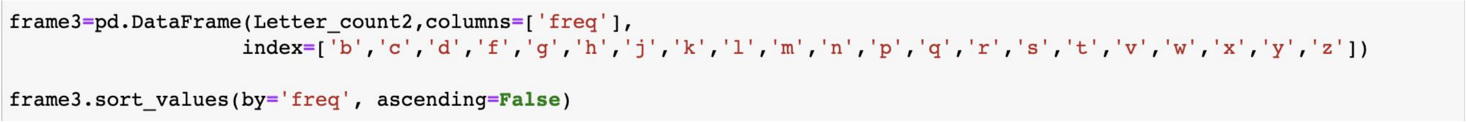


Consonants exploration

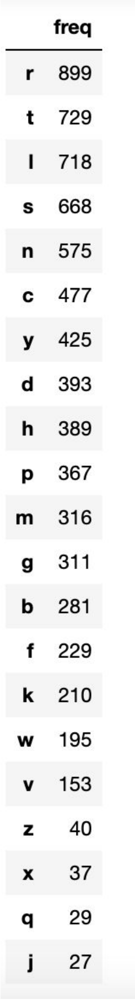
How we found the consonants frequency:

- Excluded vowels for index > output= frequency of consonants





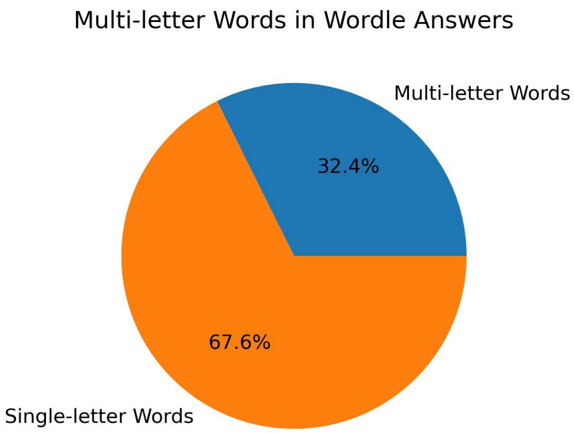
```
frame3=pd.DataFrame(Letter_count2,columns=['freq'],
                    index=['b','c','d','f','g','h','j','k','l','m','n','p','q','r','s','t','v','w','x','y','z'])

frame3.sort_values(by='freq', ascending=False)
```

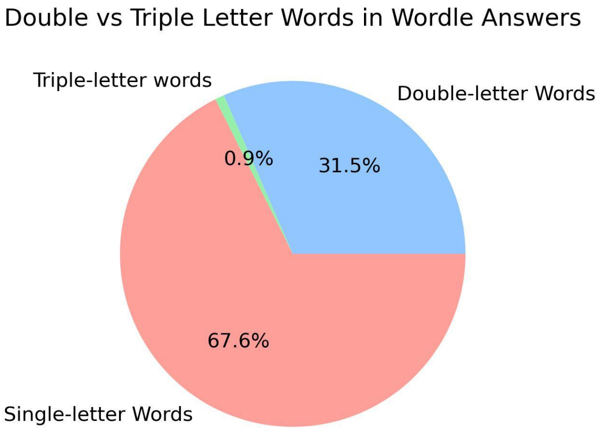


	freq
r	899
t	729
l	718
s	668
n	575
c	477
y	425
d	393
h	389
p	367
m	316
g	311
b	281
f	229
k	210
w	195
v	153
z	40
x	37
q	29
j	27

Spread of Multi or Double Lettering

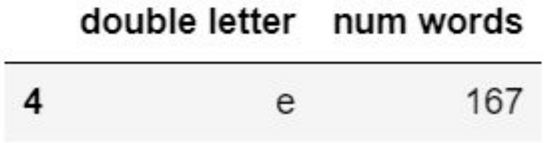
Out of the 2,315 wordle answers from the timing of downloading this dataset, only 749 words had more than 1 of the same letter, 730 words with only double letters, and 20 words with only triple letters.



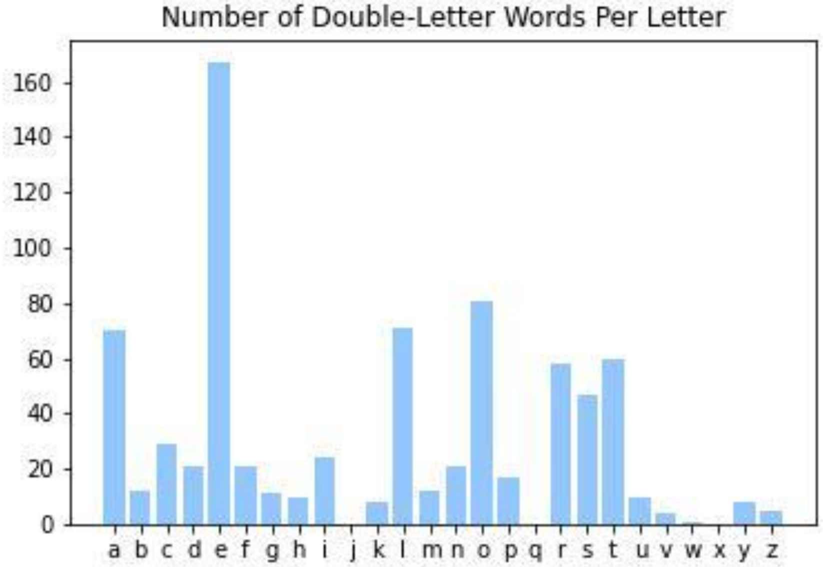


Most Common Double Letter

Out of the 730 words with double letters only, the most common double letter was double “e” with 167 words. (i.e. “seedy”, “beefy”, “cheek”, etc)

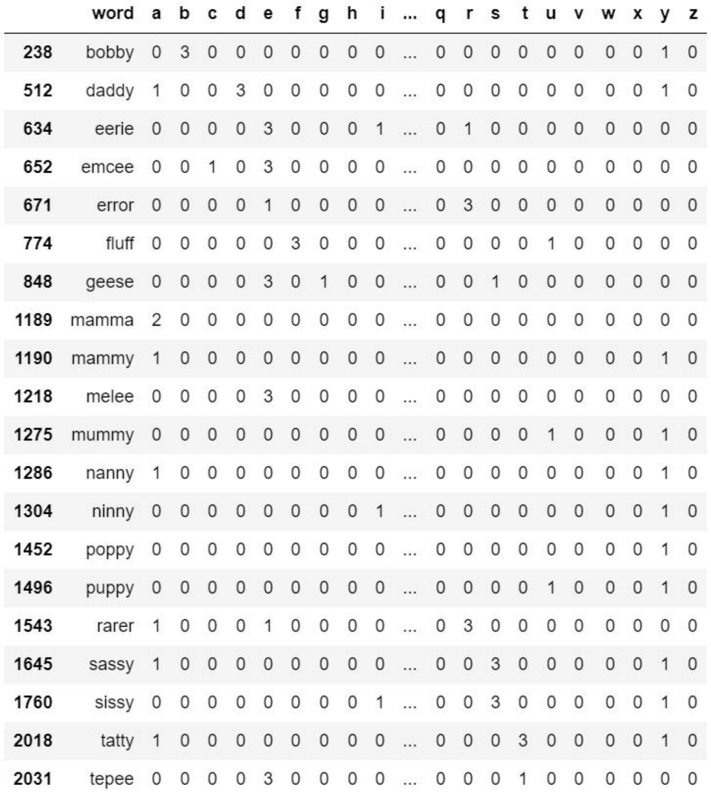


	double letter	num words
4	e	167



Triple Letter, Triple Threat

Out of the 2,315 possible wordle answers, only 20 of them have triple letters. This suggests that the chances of the word of the day having triple letters is VERY rare, but that also makes it VERY lethal to players' streaks.



	word	a	b	c	d	e	f	g	h	i	...	q	r	s	t	u	v	w	x	y	z
238	bobby	0	3	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
512	daddy	1	0	0	3	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
634	eerie	0	0	0	0	3	0	0	0	1	...	0	1	0	0	0	0	0	0	0	0
652	emcee	0	0	1	0	3	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
671	error	0	0	0	0	1	0	0	0	0	...	0	3	0	0	0	0	0	0	0	0
774	fluff	0	0	0	0	0	3	0	0	0	...	0	0	0	0	1	0	0	0	0	0
848	geese	0	0	0	0	3	0	1	0	0	...	0	0	1	0	0	0	0	0	0	0
1189	mamma	2	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1190	mammy	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
1218	melee	0	0	0	0	3	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1275	mummy	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	1	0
1286	nanny	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
1304	ninny	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	1	0
1452	poppy	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
1496	puppy	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	1	0
1543	rarer	1	0	0	0	1	0	0	0	0	...	0	3	0	0	0	0	0	0	0	0
1645	sassy	1	0	0	0	0	0	0	0	0	...	0	0	3	0	0	0	0	0	1	0
1760	sissy	0	0	0	0	0	0	0	0	1	...	0	0	3	0	0	0	0	0	1	0
2018	tatty	1	0	0	0	0	0	0	0	0	...	0	0	0	3	0	0	0	0	1	0
2031	tepee	0	0	0	0	3	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0

Part of Speech Exploration:

- We decided to separate the wordle answer data to only keep the 5 letter words.
- After removing missing values and outliers we were left with 16,900 5-letter words
- Top 5 categories:

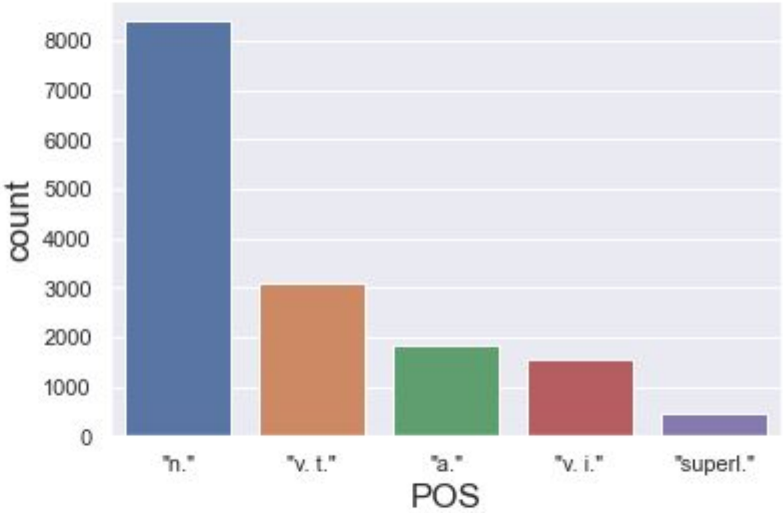
-Noun: 8329

-Verb Tense:3096

-Adverb: 1830

-Intransitive Verb: 1533

- Superlative Adverb: 458



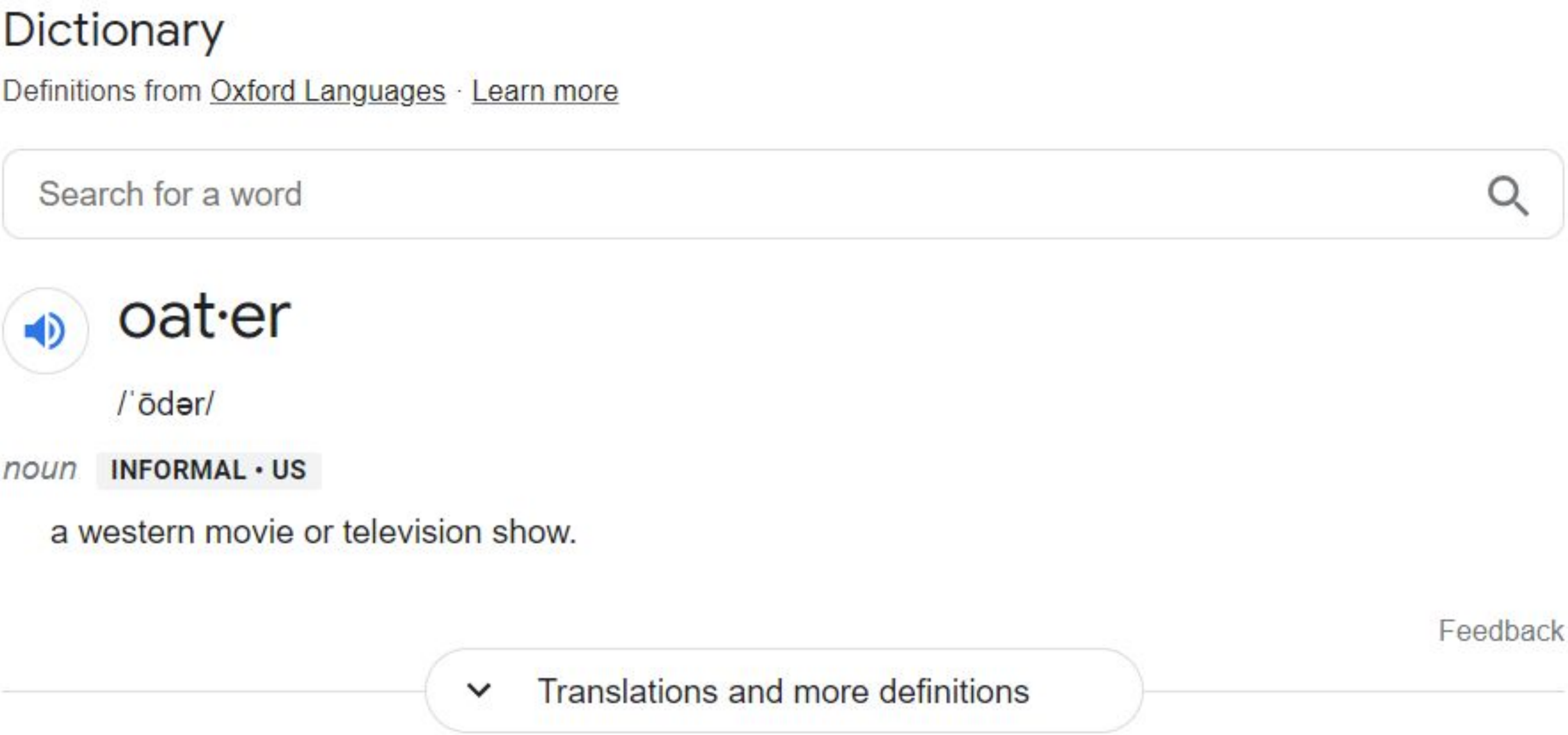
Finding the “Best” starting word

[Alice show them the notebooks…]

Conclusions

- The top five most common letters in the list of possible wordle answers are **E, A, R, O, T**
- There's a **~32% chance that a word has a double letter** (so be careful!)
- The most common POS for correct wordle answers was **noun**

The best Wordle starting word based on our data is **"Oater"** which has all of the top five words in it, it is a noun, and it has no double letters.



Dictionary

Definitions from Oxford Languages · Learn more

oat·er

/ˈōdər/

noun **INFORMAL • US**

a western movie or television show.

Error Analysis + Ideas for the future:

Data Discrepancies:

- Word in the Wordle allowed words and possible Wordle answers DID NOT overlap (an oversight on our part)

Possible enhancements:

- Exploration of letter placement
- Good letter couplets
- Average number of vowels and consonants per word
- Better weights for the testing tool

THANK

YOU!