Project II

Trevor Martin

3/24/2022

Introduction

Wordle is a web-based game created and developed by Welsh software engineer Josh Wardle, and owned and published by The New York Times Company since 2022. In Wordle, the player is given 6 attempts to guess the hidden 5 letter word. As you guess words, letters in those guesses may appear either yellow or green based on their presence and location in the hidden word. The game can only be played once a day and every player gets the same word each day. This game has become a viral sensation, and hundreds of thousands of players will tweet their results daily. Included in these results is the number of guesses it took the player to correctly identify the hidden word, as well as the pattern of their guesses.

Is it possible to use these tweets to determine which Wordle games were more difficult than others? Which Wordle this week was the hardest? The easiest?

Data Collection

In order to collect this data, each Wordle game needed to be distinguished from one another. Luckily the creator of the game included an identifier to tell the games apart. Each game is numbered, starting with game 0 which took place on June 19th, 2021. Each day since then, the game number has increased by 1. Using this information, a data set was created with all Wordle game numbers and their corresponding dates from the beginning to the end of Wordle (October 20th, 2027).

```
wordle_dates <- data.frame("Date" = as.Date(seq(as.Date("2021-06-19"),
                                                as.Date("2027-10-20"), by = 1)),
                          "Wordle" = seq(0,2314, by = 1))
```

Using the current date, a list of the most recent Wordle games is created.

```
recent_dates <- data.frame("Date" = seq(Sys.Date()-6, Sys.Date(), by = 1))

recent_wordles <- merge(recent_dates, wordle_dates, on = "Date")

recent_wordles
```

```
##         Date Wordle
## 1 2022-03-18    272
## 2 2022-03-19    273
## 3 2022-03-20    274
## 4 2022-03-21    275
## 5 2022-03-22    276
## 6 2022-03-23    277
## 7 2022-03-24    278
```

The only thing missing at this point was the Wordle results from Twitter users. This was the most complicated portion of data collection, and involved a few steps. The first step was to create a matrix to store the Twitter data in.

```
wordle_tweets <- as.data.frame(matrix(0, ncol = 1, nrow = 2500))
```

This matrix was then filled with 2500 rows of data for each of the 7 most recent Wordle games. This number was chosen because it resulted in 17500 tweets collected, and the rate limit on the Twitter API is 18000 tweets per 15 minute interval. The only portion of the Twitter data that was relevant to this project was the "text" information, aka the content of the tweet. Every other piece of Twitter data was removed.

```
for (i in seq_along(recent_wordles$Date)){
  each_wordle_tweets <- search_tweets(q = paste("Wordle", recent_wordles[i,"Wordle"]),
                                      n = 2500, include_rts = FALSE)
  wordle_tweets[,as.character(recent_wordles[i,"Wordle"])] <- each_wordle_tweets[,"text"]
}
```

Once the data frame was full of Wordle and Twitter data, it had to be properly formatted to create the visualization. The first column was removed because that was the dummy column created when creating the placeholder matrix. Then, all of the columns for each Wordle game were removed and instead added on as additional rows, creating an identifier for which game they belonged to.

```
wordle_tweet_df <- wordle_tweets[,-1]

wordle <- as.data.frame(matrix(recent_wordles[1,2], nrow = 2500, ncol = 1))
wordle <- cbind(wordle, wordle_tweet_df[,1])
names(wordle) <- c("Wordle", "Tweets")

for (j in 2:7){
  wordle_temp <- as.data.frame(matrix(recent_wordles[j,2], nrow = 2500, ncol = 1))
  wordle_temp <- cbind(wordle_temp, wordle_tweet_df[,j])
  names(wordle_temp) <- c("Wordle", "Tweets")
  wordle <- rbind(wordle, wordle_temp)
}

head(data.frame("Wordle" = wordle$Wordle, "Tweets" = substr(wordle$Tweets,1,15)))
```

```
##   Wordle           Tweets
## 1    272  Wordle 272 5/6*
## 2    272 Wordle 272 4/6\n
## 3    272  nerve-wracking
## 4    272 Wordle 273 5/6\n
## 5    272 Wordle 272 2/6\n
## 6    272 Wordle 272 5/6\n
```

With just 2 columns to work with, this last step became much easier. The score had to be extracted from each of the tweets. In order to accomplish this, the pattern of the tweets was identified. In the default tweet generated from the Wordle game's share button, the game number is followed directly by the score. For example, a tweet will typically read "Wordle 276 4/6. . . " which made it simple to locate the final score as most players do not modify the default tweet from the share feature.

```
for (row_counter in 1:17500){
  wordle$solved[row_counter] <-
    as.character(substr(sub(paste(".*", wordle$Wordle[row_counter], ""),
                            "", wordle$Tweets[row_counter]),1,1))
}

head(data.frame("Wordle" = wordle$Wordle,
                "Tweets" = substr(wordle$Tweets,1,15),
                "Guesses" = wordle$solved))
```

```
##   Wordle          Tweets Guesses
## 1    272  Wordle 272 5/6*       5
## 2    272 Wordle 272 4/6\n       4
## 3    272  nerve-wracking       5
## 4    272 Wordle 273 5/6\n       6
## 5    272 Wordle 272 2/6\n       2
## 6    272 Wordle 272 5/6\n       5
```

The possible results for Wordle are guessing in either 1, 2, 3, 4, 5, or 6 guesses, or not guessing the word at all, represented by an "X". In order to filter out tweets that did not contain results, the character after the game number was checked against this list of possible results. If this character was not a possible game result, that tweet was removed.

```
possible_results <- c("1","2","3","4","5","6","X")

wordle_clean <- wordle[wordle$solved %in% possible_results,]

head(data.frame("Wordle" = wordle_clean$Wordle,
                "Tweets" = substr(wordle_clean$Tweets,1,15),
                "Guesses" = wordle_clean$solved))
```

```
##   Wordle          Tweets Guesses
## 1    272  Wordle 272 5/6*       5
## 2    272 Wordle 272 4/6\n       4
## 3    272  nerve-wracking       5
## 4    272 Wordle 273 5/6\n       6
## 5    272 Wordle 272 2/6\n       2
## 6    272 Wordle 272 5/6\n       5
```

```
wordle_clean$Wordle <- as.factor(wordle_clean$Wordle)

remaining_data <- as.data.frame(table(wordle_clean$Wordle))
names(remaining_data) <- c("Wordle", "Tweets Remaining")
remaining_data
```

```
##   Wordle Tweets Remaining
## 1    272             2432
## 2    273             2429
## 3    274             2453
## 4    275             2452
## 5    276             2467
## 6    277             2462
## 7    278             2494
```

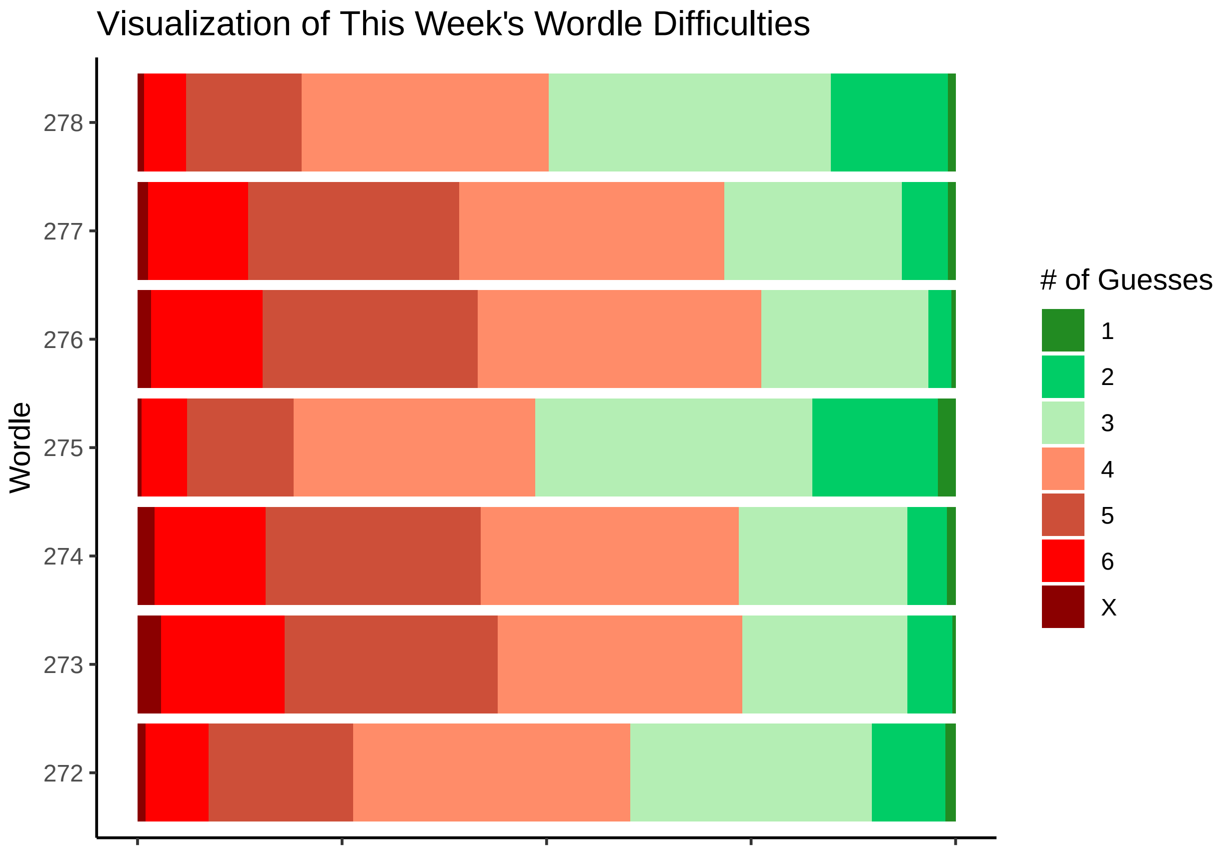
Luckily, very few tweets were not about player's results. All games started with 2500 tweets. The remaining data is more than large enough to draw meaningful conclusions from.

Analysis

The main barometer for the difficulty of a Wordle game is the number of guesses it takes for the player to guess the hidden word. Games where it takes five or six tries are generally considered more difficult than games that take three or four tries. Below is a visualization of the number of guesses taken by Twitter players with the worst scores being in red and the best scores being in green.

```r
wordle_clean$Percent <- 1

ggplot(wordle_clean, aes(x = Percent, y = Wordle, fill = solved)) +
  geom_bar(stat = "identity", position= "fill", orientation = "y") +
  theme_classic() +
  labs(x = "%", y = "Wordle", title = "Visualization of This Week's Wordle Difficulties",
       fill = "# of Guesses") +
  scale_fill_manual(values = c("forestgreen", "springgreen3", "darkseagreen2", "salmon1",
                               "tomato3", "red1", "red4")) +
  theme(axis.title.x=element_blank(), axis.text.x=element_blank())
```



Conclusion

From this visualization, there are clear differences in the difficulty of the Wordle games. It can be seen that Wordle 275 & Wordle 278 were the easiest games this week, while both Wordle 276 and 273 appeared to give players some trouble. Wordle 273 had the highest number of players not guess the word at all, while Wordle 276 had very few players guess the word in one or two tries.

Discussion

These results show that many players have similar play-styles and are guessing similar letters to other players. There are a few ‘optimal’ starting words in the Wordle community such as Adieu, Trace, and Irate. These words reveal a lot of vowels or place the most frequently used letters in their most frequently used positions which helps the player narrow down their guesses. With this type of community, it makes sense that there can be some words that are more quickly guessed by a large number of players, since many players are using similar starting words. On days where these words work, you will see a lot of green on this visualization, similar to Wordle 275 or Wordle 278. The hidden words for these days were ‘their’ and ‘chest’. Many popular starting words contain many of these letters, so the community as a whole benefits from an ‘easier’ hidden word.

In order to improve this study, more historical data could be used. The Twitter API used for this project only allows the collection of tweets over the past 6-9 days, so that was not possible for this project. With access to the full historical data as well as the list of Wordle answers, some interesting analysis could be conducted on how the hidden words letters correlates with the average number of guesses.

There is also most likely the presence of survivorship bias, as people are less likely to tweet out their worse scores compared to their better scores. Using these results to try to predict the results of the average player might not be ideal, as these results most likely contain an unrealistically low number of bad scores.