Vaajoor (Persian Wordle) data analysis

Bahar Oveis Gharan, Mohamadreza Bahrami

[a]*Sharif University of Technology*

1. Introduction

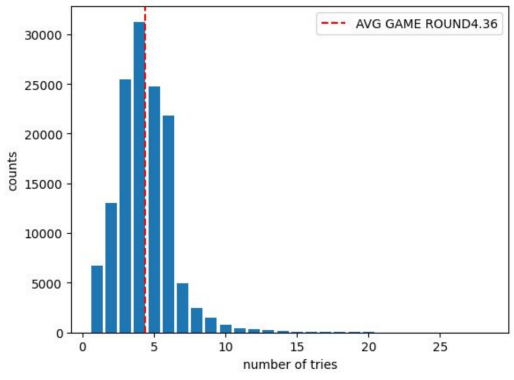
First of all, I have been utilizing available data in Big Query for a month (2022-03). This dataset includes information about guesses and the timestamps of game-play sessions. Unfortunately, it lacks the *userId field* required for analyzing individuals' gameplay attitudes. This field was recently added to the data. My primary purpose is to explore this data to gain understanding of people's behaviors, focusing more on behavioral analysis rather than technical analysis.

2. Overall Analysis

The combination of *ga-session-number, ga-session-id*, and **id** indicates a unique game that a player plays, while the field indexes show the order of guesses that the user has made. The field "id" indicates which word game the user is playing.

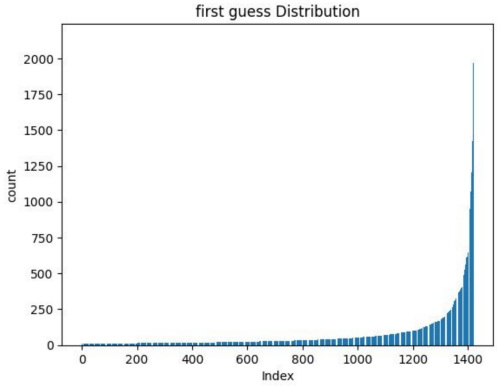
First, I cleaned the data and removed games with incorrect logs.

Below, you will find a table displaying the counts of guessing attempts, and on average, people make 4 attempts to find the answer.



This number is consistent with the Wordle game report, where people make four attempts on average. Approximately 30 percent of people are able to win within four attempts.

I see the distribution of first guesses. As you can see, there are certain words that function as starters and are commonly used as the initial guess. Furthermore, the distribution of first words is not uniform.

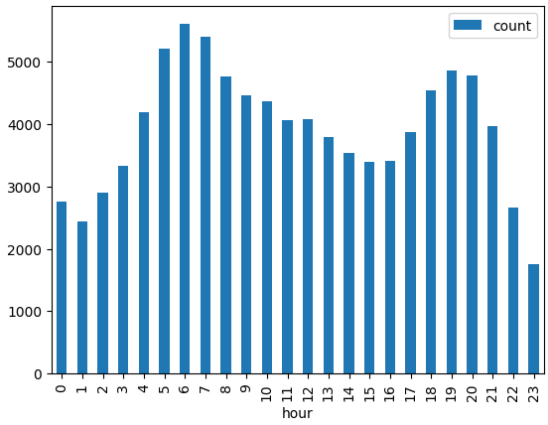


The table displays the most frequently used words as the first guess. By examining the most commonly occurring letters in five-letter words, it becomes apparent that these letters are present in those words. Therefore, it can be inferred that people employ a certain strategy for their initial guess, rather than making a random choice.

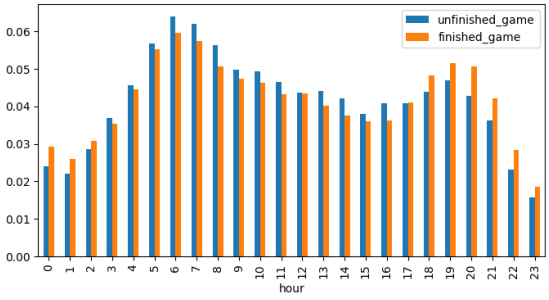
guess	count
درمان	2123
سلامت	1967
دیوار	1613
سماور	1407
نوروز	1398
سیمان	1383
رامین	1289
دیوان	1089
تهران	1072
ویران	1041
سنتور	977
کرمان	950
ایوان	943
مرداب	929

```
[('10544 ,ب'),
 ('11138 ,ل'),
 ('13298 ,ت'),
 ('14813 ,و'),
 ('15188 ,ن'),
 ('15604 ,ه'),
 ('17264 ,م'),
 ('18925 ,ر'),
 ('19282 ,ی'),
 ('31126 ,ا')]
```

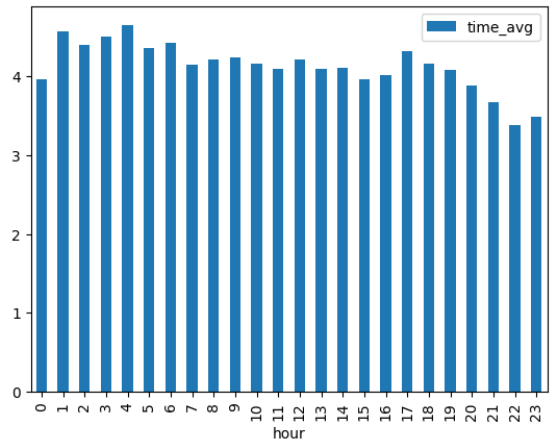
I examine the timing of games and track the number of games played in each hour of the day. Considering that people play this game in Iran, I use the Iran timezone to convert timestamps. As seen in the plot, people play the game between 6 to 7 AM more frequently than other time of the day, which is quite surprising. Additionally, the total number of games increases around 8 pm again. I believe these two times are the best opportunities to update game word and attract more users.



I have categorized games into "finished games," where users find the answers, and "unfinished games," where users drop the game. I am interested in determining whether the time of day has an impact on players' motivation to find the answers. To investigate this, I plotted a normalized graph of game counts for each hour of the day. The results indicate that in the morning, the percentage of people who drop the game is higher than at night. However, overall, the figures remain the same.



During the observation period, I recorded the time it took for users to find the answer in each game. I categorized these time durations based on the hour of the day in order to determine whether the hour had an effect on the duration of finding the answer or not. As we can see, there is no significant difference



in the average time it takes to find an answer depending on the hour. Therefore, it can be concluded that the duration of the game is independent of the time of the day.It take about four minutes that users finds answer in average.

3. Game Strategy

I want to explore people's game strategies. There are some algorithms based on entropy, but I don't think they are a good way to analyze people's gameplay. This is because the list of words that exist in individuals' minds varies and depends on factors such as age, situation, and many other parameters.

I try to analyze people's games in a different manner. I begin with the simplest approach: defining a scoring system for guesses. In this system, each yellow-colored letter is assigned 1 point, while green letters are assigned 2 points. Therefore, the score for a guess is determined by summing up the points for its letters. Additionally, the answer receives 10 points.

It is evident that individuals can achieve a score equal to or greater than their previous guess score, suggesting that their guesses must be monotonically non-decreasing values. However, there are some people who prefer seeking new information rather than relying on their prior knowledge of the discovered letters and the answer. To initiate this approach, we can define strategic games wherein users aim to gather more information instead of solely guessing the final answer. This is indicated by their guesses' scores not being non-decreasing, as they strive to uncover most of the word's letters in their initial attempts before making the final guess. In this simplest definition of a strategic game, I observe that only **30** percent of people choose to play strategically.

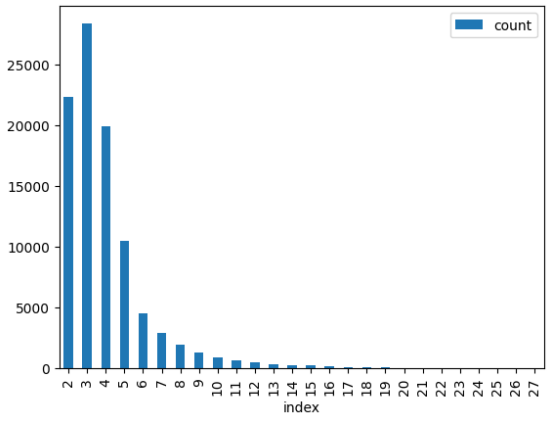
4. Guess group

I split the guesses into three groups:
The first group is considered wise guesses, which means a per-

son uses all previous information and tries to guess the final answer, so green letters and yellow letters from previous guesses are used in the right place. I set tag "2" to these guesses in Dataframe.
The second group is the guesses that try to find more information. So people don't use green letters and also don't use yellow letters in the place that they used before. These guesses tag is "1". First guesses are always in this group.
The last group is bad guesses, that people used the previous red letter or yellow letter in the place that they knew was wrong. This group's tag is "0".
About 40% of the guesses are considered wise, while 23% of them are regarded as bad guesses.36% of the guesses are neither wise nor bad.

But we know that the first guess always falls into the second group. If we don't consider it, only 3% of the guesses attempt to find new information instead of guessing the final answer(wise guesses group). This shows that people are more willing to try guessing the final answer rather than seeking new information at the beginning of the game.

The average time for each guessing group is as follows: wise guess - 75 seconds, bad guess - 76 seconds, and not wise-not bad - 67 seconds. It is interesting to note that wise guesses take less time compared to bad guesses. This can be attributed to the difficulty people face when attempting to use all the information to make a single-word guess. Wise guesses have a smaller group of words to be considered, unlike bad guesses that have a larger group of words to be choosen from.
Bad guesses can be the result of people becoming tired of guessing, causing them to use random words near their options. Based on my observations, I have noticed that people tend to make the most bad guesses, as you can see below.



5. Feel-Lucky game

Feel-lucky" games are games in which people improve their guesses with each attempt in order to find the final answer. In this group, participants tend to employ more strategic and informed guesses after their initial attempts until the conclusion of the game.

I split the game into "Feel Lucky" game and other games. The "Feel Lucky" game finds the answer within an average of 3.9 attempts and 240 seconds. However, the other games receive the answer in an average of 5.2 attempts and 307 seconds. These amazing results show that if people use previous information, it is easier for them to find the final word in their mind rather than knowing letters that are not part of the final answer.

6. Future analysis ideas

- Explore individuals' behaviors and their playing time.
- Explore individuals' strategies for starter words.
- Improve the definition of a strategy game and its score
- Observe the starter words when checking with the dictionary's word mode was unable.