Earthquakes and Tweets: Turkey

Group 1: Leslie Medina, Felix Wu, Merve Dumlu, Stacy Deng, Leo Ng, Mianbo Hu

Introduction

On February 6th, 2023, a devastating earthquake of magnitude 7.8 struck southern and central Turkiye (Turkey) and northern and western Syria. Many aftershocks happened not long after, leaving thousands of buildings collapsed and tens of thousands buried under the rubble. The widespread damage affected millions of people, and damaged roads impeded relief efforts. Many people called out for help via Twitter, a popular social media service, but unfortunately, most went unseen.

The goal of our project is to analyze tweets - messages in the form of text, photos, or videos from Twitter - and to identify which tweets have the most reach in order to help with earthquake relief efforts. Our hypothesis is that of the tweets from Turkey, the more likes and retweets (reshares on Twitter) - together referred to as visibility - you receive on a tweet, the more likely it is that frequent keywords are used in said tweet.

Data Collection and Cleaning

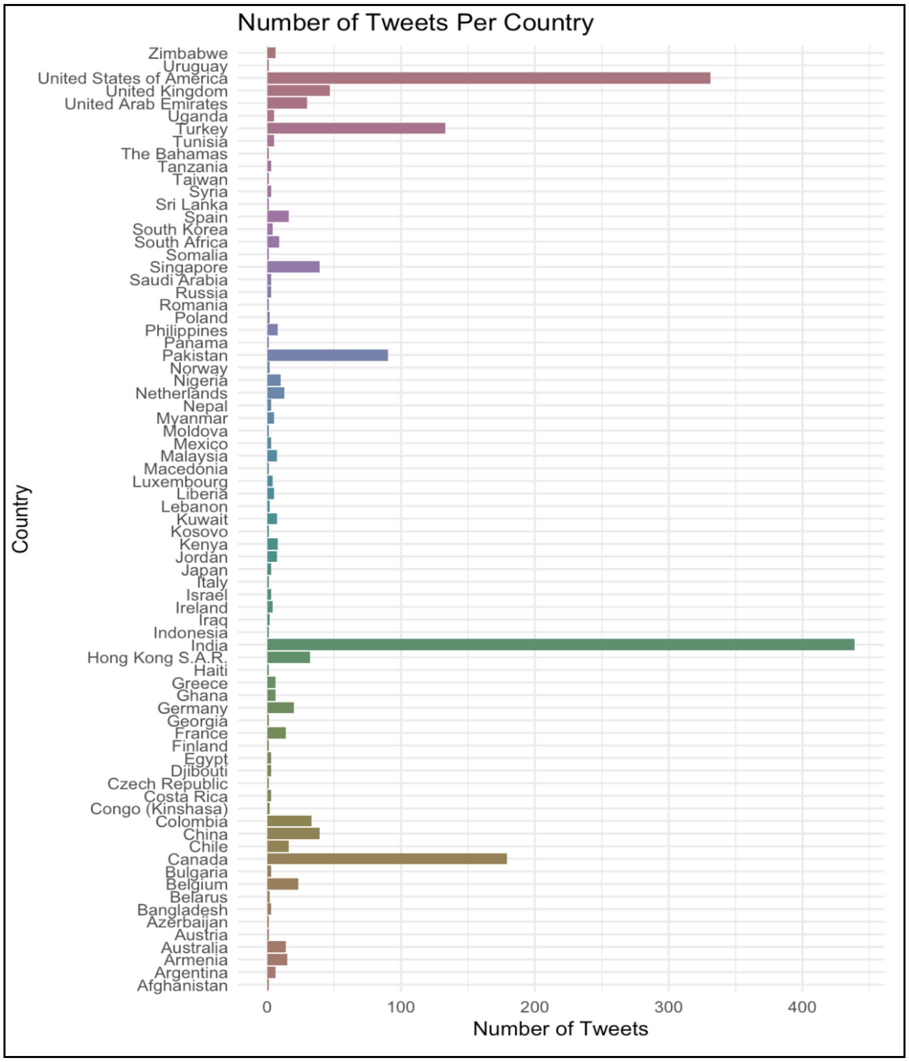
The dataset we used is the Turkey Earthquake Tweets dataset by Gabriel Preda, found on Kaggle. The dataset contains 28,844 observations (rows) and 16 variables (columns), with each column providing information about the Twitter user who made the tweet (user_name, user_location, user_description, user_created, user_followers, user_friends, user_favorites, user_verified) or the tweet itself (date, text, hashtags, source, retweets, favorites, is_retweet), along with one identifying column (id).

For our project, we eliminated most of the user related data, keeping only “user_name” so that we could identify who tweets belonged to, and “user_location”, which would help us determine where the user is located. Of the tweet related data, we removed the “is_retweet” column that determined whether a tweet was a retweet. We did this because the majority of the column had “NA” values, while two rows were labeled “No”, hence it was determined to be invaluable. The remaining columns for tweets were all kept. We also kept the “id” column which identified each tweet, narrowing our data down to 9 columns, shown in the table below.

Table 1: Variables and Descriptions

Variable	Description
id	The tweet's ID number
user_name	The username of the Twitter user
user_location	The location of the Twitter user
date	The date and time that a tweet was made
text	The text of the tweet
hashtags	Hastags that were used in a tweet
source	What device a tweet was made from
retweets	The number of times a tweet was retweeted by others
favorites	The number of times a tweet was liked/favoried by others

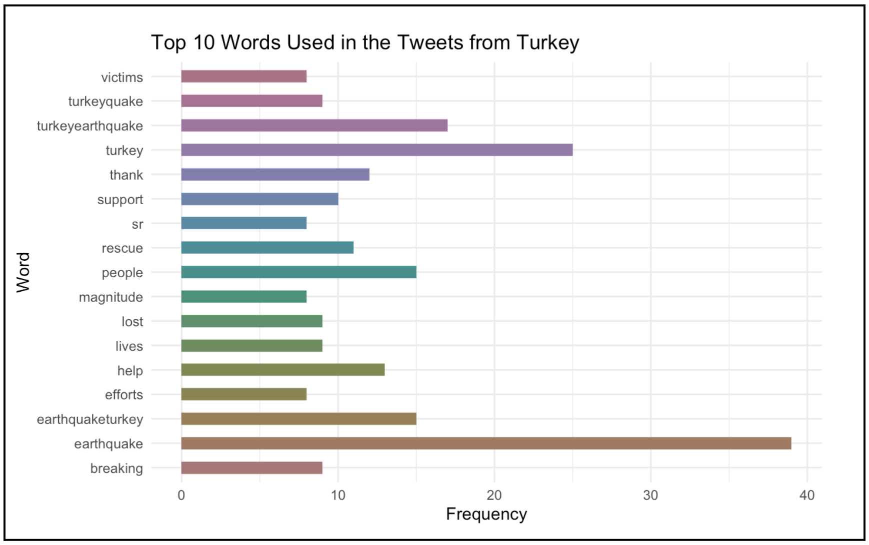
To test our hypothesis, we wanted to use the variable "user_location" to identify which tweets actually originated from Turkey. However, Twitter allows users to input their own location, which lead to values in the "user_location" column such as "Resident of Planet Earth", "Universe", or "BTS 💜 ", along with many empty values as a location is not required to use Twitter. We uploaded a dataset–from the GitHub user girijesh18–that depicted each country and cities associated. Using this dataset and the package "countrycode", we were able to extract all of the tweets made from legitimate locations, narrowing our dataset of 29,665 observations down to only 1705 observations. Of these 1705 observations, 133 of these tweets were actually from Turkey.

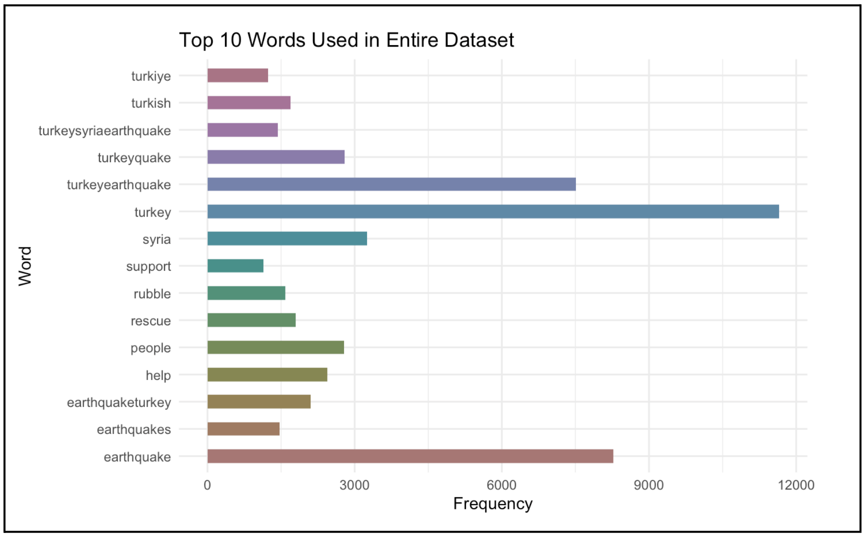


Finally, we split our dataset into three different datasets: the dataset with only the observations with country name "Turkey", the dataset with all of the countries, and the main dataset that includes all locations regardless of country value being NA.

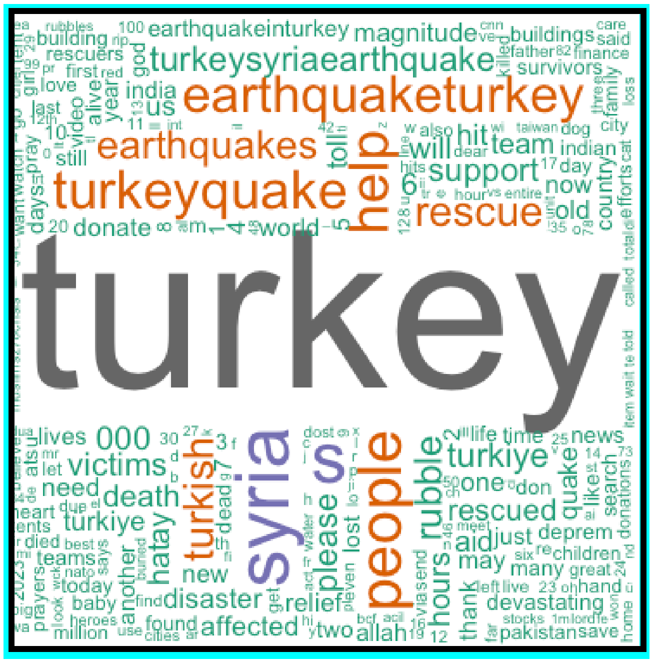
Methodology and Observations

We wanted to find out which words were the most commonly used in tweets about the Turkey earthquakes. To do so, we used the 'qdap' package, a library targeted towards transcript analysis, which allowed us to find the most frequent words used across our dataset. Initially, the most commonly used words displayed were words like articles (a, an, the), pronouns (you, we, our, etc.), and other function words (is, was, etc.). We removed these words from the text using the 'removeWords' function from the package "tm", which is a library used for text mining. The links from the "text" column, which were typically links to images, were also removed since we are only interested in the text of the tweets for this project. The most commonly used words from tweets that come from Turkey and tweets from all countries combined were as follows:

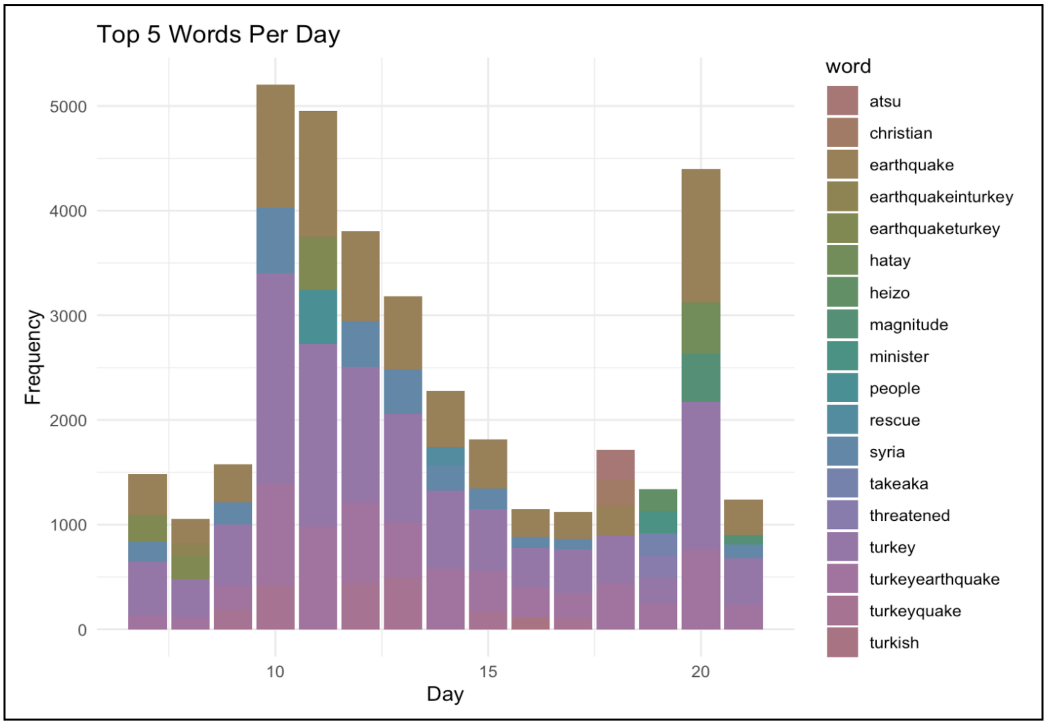




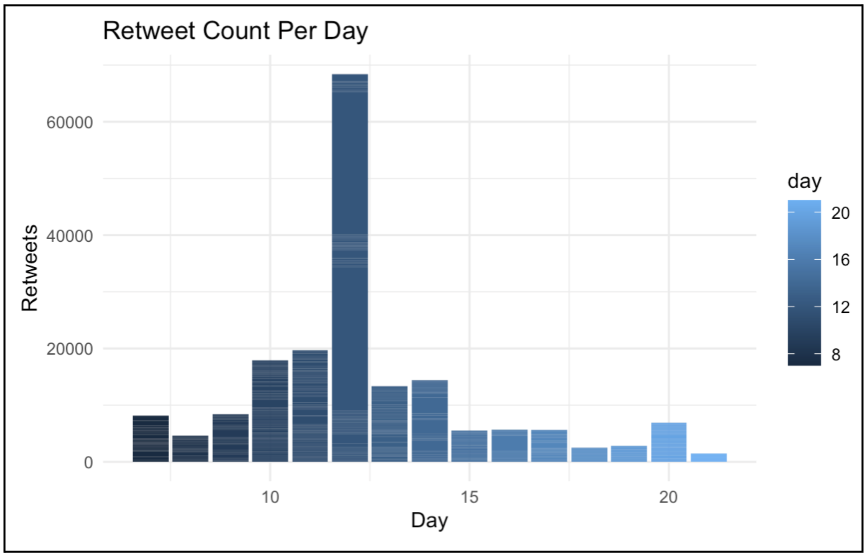
Both groups had similar keywords, with the words "Turkey" and "earthquake" being the most common words. Terms like "earthquaketurkey" and "turkeyquake" were also common terms found, and that's likely because these are the hashtags used on Twitter in regards to the Turkey earthquakes. The following word cloud shows that many of the keywords used in tweets are words like "help", "rescue", "donate", or "support", which we expected, as many would likely turn to social media for help in hopes of being seen.



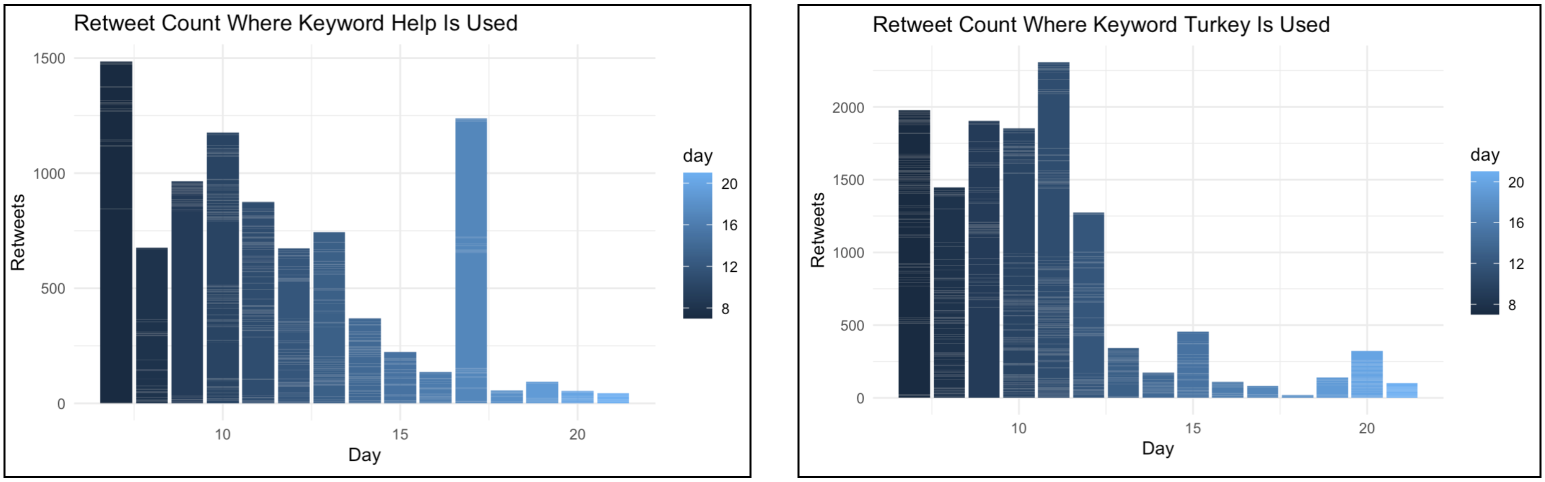
We also plotted a graph showing which keywords were the most tweeted across the month, and as expected, words present in our word cloud were the most prevalent, with the word “earthquake” being the most tweeted keyword consistently throughout the month of February.



Knowing the keywords being used in tweets, we also wanted to see the relevance of tweets related to the Turkey earthquakes over the month, as we suspected that keyword usage would follow a similar pattern. We plotted the following graph for retweets over the month of February for our dataset of all countries:



The keywords "Turkey" and "help" were chosen arbitrarily among the 15 most common words, with the purpose to compare their usage to the retweets.



As we can see from the graphs, the retweet counts of tweets that include the words "Turkey" and "help" follow a similar pattern, where there is a high retweet count towards the beginning of the month when the earthquakes first occurred, before tapering off towards the end of the month, where a significant amount of time has already passed. Although the pattern does not exactly match that of the graph of the retweet count per day near the beginning of the month, we can see that their general pattern is similar; the longer the time has passed since the earthquakes occurred, the less retweets and likely visibility it receives.

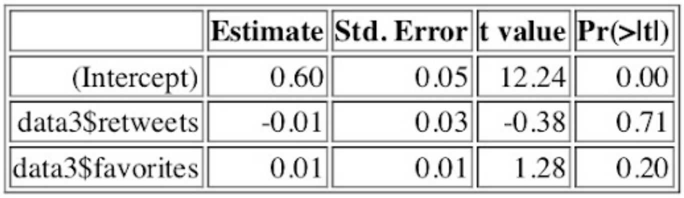
Hypothesis Testing

Our goal is to investigate the connection between tweet visibility (as measured by the number of retweets and likes) and the use of commonly occurring keywords. To this end, we have formulated a null hypothesis: there is no association between the frequency of retweets and likes and the use of frequently occurring words. Our alternative hypothesis, on the other hand, posits that a greater number of retweets and likes is correlated with the use of the most commonly occurring keywords. Our study will primarily examine tweets originating from Turkey.

To test our hypothesis on the relationship between tweet visibility and the likelihood of using frequent keywords, we employed logistic regression analysis. This method was selected as it allows for the examination of the association between the response variable (is_top_15_turkey) and the explanatory variables (number of retweets and number of likes).

To conduct the logistic regression analysis, we created a dummy variable called "is_top_15_turkey" to indicate if a tweet used any of the top 15 commonly occurring keywords. We then employed a logistic regression model with "is_top_15_turkey" as the response variable and "number of retweets" and "number of likes" as the explanatory variables.

Upon analyzing the results of the logistic regression model, we found that the p-values of both explanatory variables were not statistically significant (larger than 0.05). This suggests that, based on the data used in our analysis, we do not have enough evidence to reject the null hypothesis that there is no relationship between the number of retweets and likes and the use of the most frequent words. Here is the result:

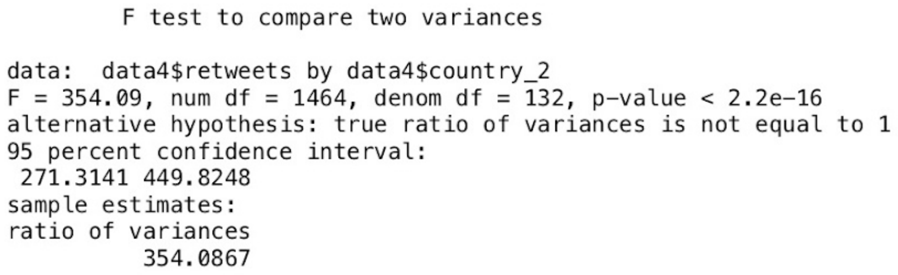


	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.60	0.05	12.24	0.00
data3$retweets	-0.01	0.03	-0.38	0.71
data3$favorites	0.01	0.01	1.28	0.20

Other Hypothesis Testing

In addition to the hypothesis test described earlier, we conducted another hypothesis test to investigate whether tweets originating from Turkey have a higher likelihood of exposure (measured by the number of retweets) compared to tweets from other countries. To this end, we conducted a two-sample t-test to compare the number of retweets between Turkey and other countries.

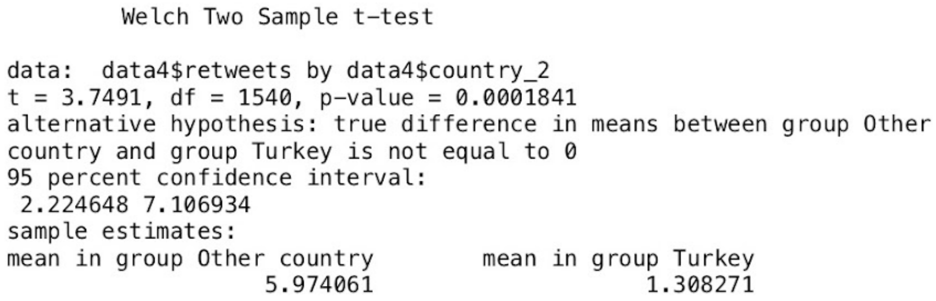
Before conducting the t-test, we needed to determine whether the two samples had equal variances. We conducted a variance test in R and found that the variances of the two samples were significantly different. Therefore, we could not assume equal variances for the two samples. The result is shown below.



```
        F test to compare two variances

data:  data4$retweets by data4$country_2
F = 354.09, num df = 1464, denom df = 132, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 271.3141 449.8248
sample estimates:
ratio of variances
          354.0867
```

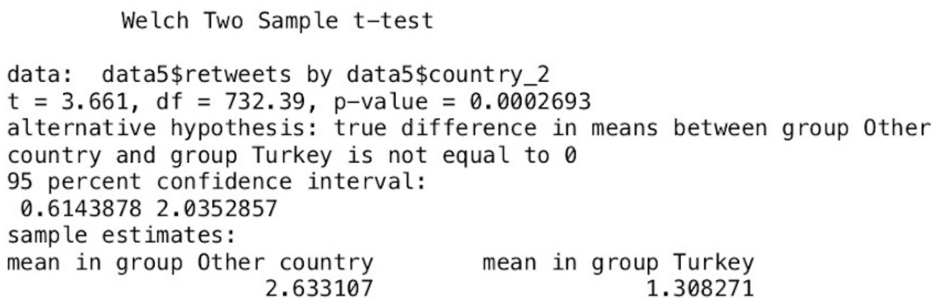
We proceeded to conduct a t-test while assuming unequal variances. Our results indicated that the means of the number of retweets between Turkey and other countries were significantly different. Furthermore, tweets originating from other countries had a higher average number of retweets compared to those from Turkey. The result is shown below.



```
        Welch Two Sample t-test

data:  data4$retweets by data4$country_2
t = 3.7491, df = 1540, p-value = 0.0001841
alternative hypothesis: true difference in means between group Other
country and group Turkey is not equal to 0
95 percent confidence interval:
 2.224648 7.106934
sample estimates:
mean in group Other country       mean in group Turkey
                   5.974061                   1.308271
```

Furthermore, our analysis revealed that tweets originating from India exhibited an abnormally higher frequency of retweets compared to other countries. This observation led us to suspect that certain tweets from India could have employed automated accounts or bots to manipulate the number of retweets. Consequently, we excluded India from the sample and re-conducted a two-sample t-test. However, even after this exclusion, the outcome remained unchanged, reinforcing the finding that tweets from Turkey have a comparatively lower number of retweets relative to other nations. The result is shown below.



```
        Welch Two Sample t-test

data:  data5$retweets by data5$country_2
t = 3.661, df = 732.39, p-value = 0.0002693
alternative hypothesis: true difference in means between group Other
country and group Turkey is not equal to 0
95 percent confidence interval:
 0.6143878 2.0352857
sample estimates:
mean in group Other country        mean in group Turkey
                   2.633107                    1.308271
```

Results

Upon conducting an analysis of tweets related to the earthquake in Turkey, we determined that frequently used keywords pertained to human lives, such as "victim," "rescue," and "help," which were observed in tweets from both Turkey and other nations. We then attempted to establish a link between the usage of these common keywords and tweet visibility in Turkey, as measured by retweets and likes. However, hypothesis testing did not yield sufficient evidence to validate this association. Furthermore, we investigated whether tweets from Turkey were more likely to be viewed, but our statistical analyses revealed that tweets from Turkey had a lower average visibility in comparison to other countries. Additionally, our examination of retweet behavior uncovered that tweets from India garnered substantially more likes and retweets compared to all other countries, raising suspicions of potential bot activity.

Conclusion

Our analysis of tweets related to the earthquake in Turkey found that commonly used keywords pertained to human lives, such as "victim," "rescue," and "help," which were prevalent in tweets from both Turkey and other nations. However, we were unable to establish a clear link between the usage of these keywords and tweet visibility in Turkey, as hypothesis testing did not yield sufficient evidence to validate this association. Moreover, our statistical analyses revealed that tweets from Turkey had a lower average visibility in comparison to other countries. We also found suspicious bot activity in retweets from India, which received substantially more likes and retweets than all other countries.

Given the small sample size of only 133 observations after dataset cleaning, the hypothesis testing conducted to establish a potential relationship between keyword usage and tweet visibility in Turkey may lack sufficient statistical evidence to conclusively reject the null hypothesis of no association. Therefore, further research with a larger sample size is necessary to obtain more reliable results. Moreover, our analysis highlights the importance of investigating various factors that influence tweet visibility and dissemination related to natural disasters across different countries to gain a comprehensive understanding of social media behavior in disaster relief scenarios.