# Tokyo Olympics 2021

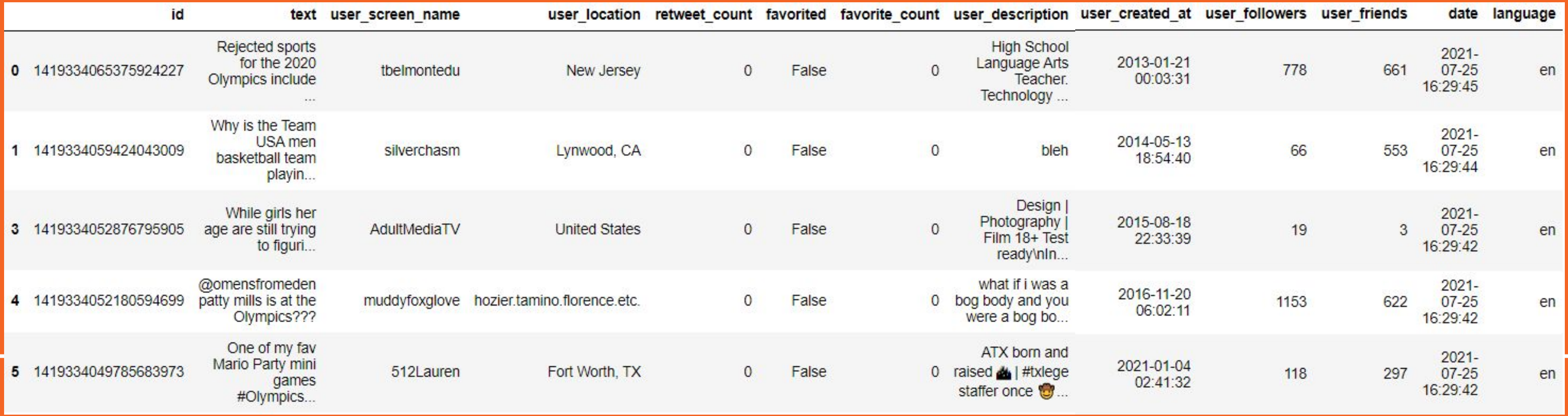
Data Analysis by Barry Murthy and Ryan Yoneshige

# Table of Contents

- EDA and pre-processing
- Most tweeted events
- Tweets by follower count
- Tweets by mention
- Top mentioned countries
- Conclusions

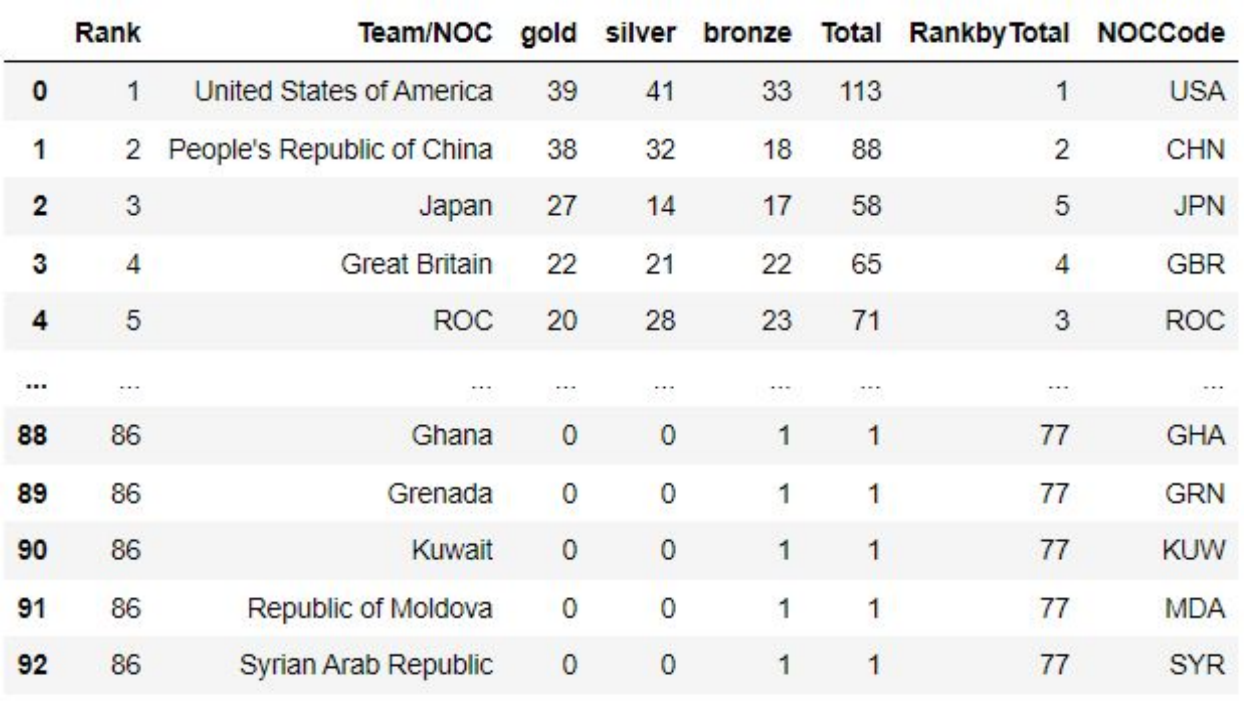
# Exploratory Data Analysis and pre-processing

- ➢ Convert csv file to Dataframe, and reformatted select columns
    - ○ Used datetime library to convert tweet post date and account created date to datetime format
    - ○ Used drop.na to cull missing values and unwanted noise
    - ○ Converted columns from string to int for later processing



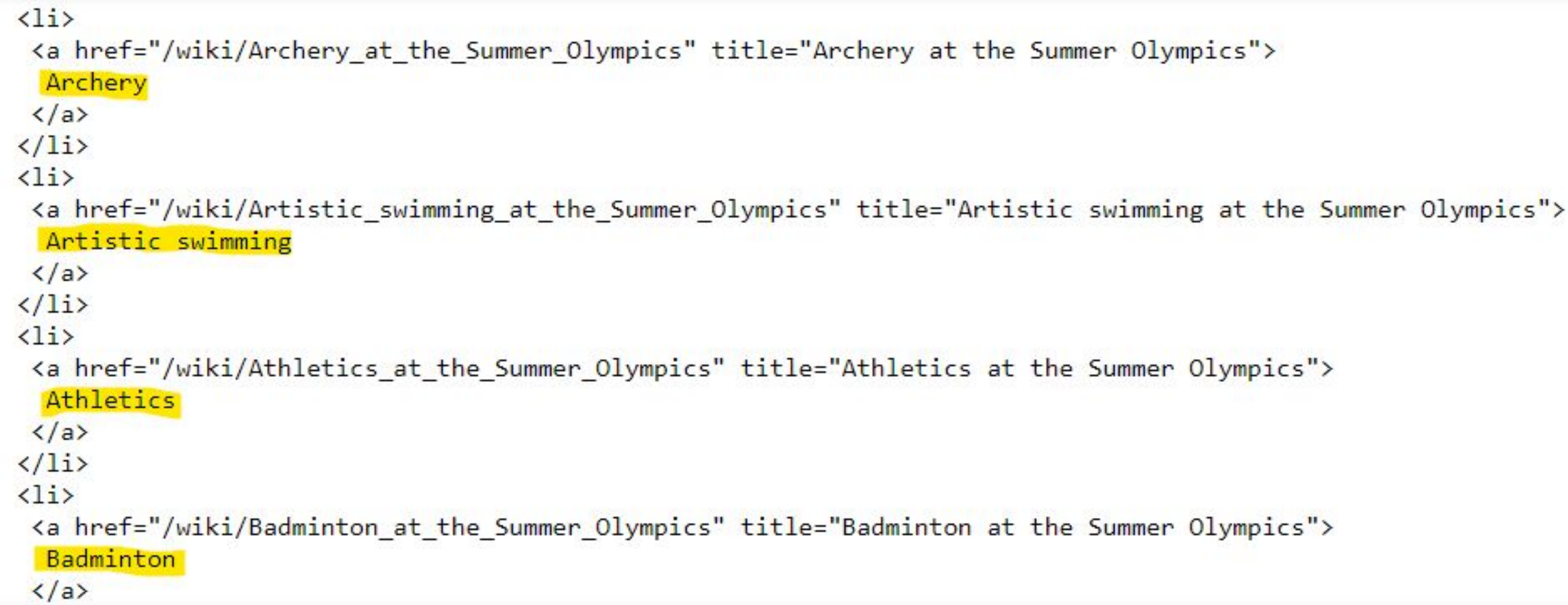
| | id | text | user_screen_name | user_location | retweet_count | favorited | favorite_count | user_description | user_created_at | user_followers | user_friends | date | language |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1419334065375924227 | Rejected sports for the 2020 Olympics include ... | tbelmontedu | New Jersey | 0 | False | 0 | High School Language Arts Teacher. Technology ... | 2013-01-21 00:03:31 | 778 | 661 | 2021-07-25 16:29:45 | en |
| 1 | 1419334059424043009 | Why is the Team USA men basketball team playin... | silverchasm | Lynwood, CA | 0 | False | 0 | bleh | 2014-05-13 18:54:40 | 66 | 553 | 2021-07-25 16:29:44 | en |
| 3 | 1419334052876795905 | While girls her age are still trying to figuri... | AdultMediaTV | United States | 0 | False | 0 | Design \| Photography \| Film 18+ Test ready\n\n... | 2015-08-18 22:33:39 | 19 | 3 | 2021-07-25 16:29:42 | en |
| 4 | 1419334052180594699 | @omensfromeden patty mills is at the Olympics??? | muddyfoxglove | hozier.tamino.florence.etc. | 0 | False | 0 | what if i was a bog body and you were a bog bo... | 2016-11-20 06:02:11 | 1153 | 622 | 2021-07-25 16:29:42 | en |
| 5 | 1419334049785683973 | One of my fav Mario Party mini games #Olympics... | 512Lauren | Fort Worth, TX | 0 | False | 0 | ATX born and raised 🏙 \| #txlege staffer once 🤠... | 2021-01-04 02:41:32 | 118 | 297 | 2021-07-25 16:29:42 | en |

- Created external dataset using read_html to read in the winning medals by Country.



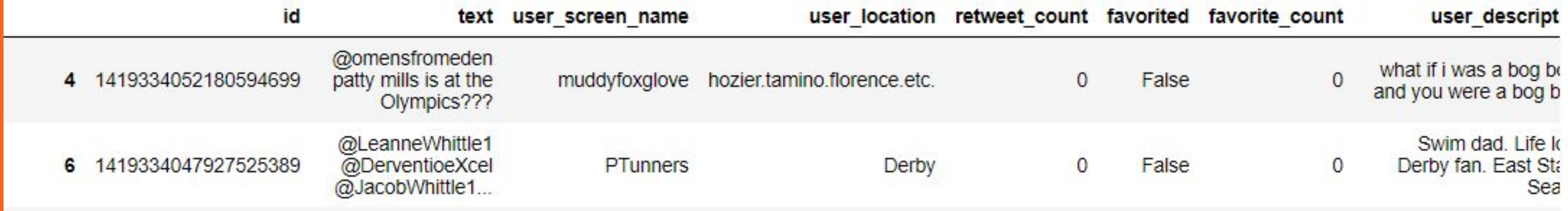
|  | Rank | Team/NOC | gold | silver | bronze | Total | RankbyTotal | NOCCode |
|---|---|---|---|---|---|---|---|---|
| **0** | 1 | United States of America | 39 | 41 | 33 | 113 | 1 | USA |
| **1** | 2 | People's Republic of China | 38 | 32 | 18 | 88 | 2 | CHN |
| **2** | 3 | Japan | 27 | 14 | 17 | 58 | 5 | JPN |
| **3** | 4 | Great Britain | 22 | 21 | 22 | 65 | 4 | GBR |
| **4** | 5 | ROC | 20 | 28 | 23 | 71 | 3 | ROC |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **88** | 86 | Ghana | 0 | 0 | 1 | 1 | 77 | GHA |
| **89** | 86 | Grenada | 0 | 0 | 1 | 1 | 77 | GRN |
| **90** | 86 | Kuwait | 0 | 0 | 1 | 1 | 77 | KUW |
| **91** | 86 | Republic of Moldova | 0 | 0 | 1 | 1 | 77 | MDA |
| **92** | 86 | Syrian Arab Republic | 0 | 0 | 1 | 1 | 77 | SYR |

- Created external dataset using Wikipedia to scrape the event names.

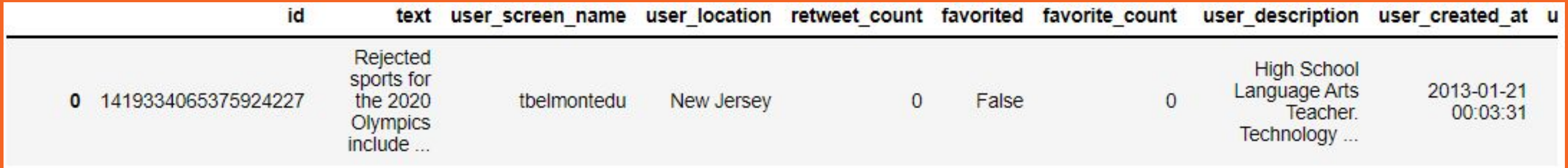


```
<li>
 <a href="/wiki/Archery_at_the_Summer_Olympics" title="Archery at the Summer Olympics">
  Archery
 </a>
</li>
<li>
 <a href="/wiki/Artistic_swimming_at_the_Summer_Olympics" title="Artistic swimming at the Summer Olympics">
  Artistic swimming
 </a>
</li>
<li>
 <a href="/wiki/Athletics_at_the_Summer_Olympics" title="Athletics at the Summer Olympics">
  Athletics
 </a>
</li>
<li>
 <a href="/wiki/Badminton_at_the_Summer_Olympics" title="Badminton at the Summer Olympics">
  Badminton
 </a>
```

➢ Filtered the original tweets DF to be organized by...
mentions (with or without @)

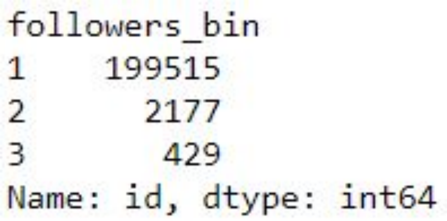


| | id | text | user_screen_name | user_location | retweet_count | favorited | favorite_count | user_descript |
|---|---|---|---|---|---|---|---|---|
| **4** | 1419334052180594699 | @omensfromeden patty mills is at the Olympics??? | muddyfoxglove | hozier.tamino.florence.etc. | 0 | False | 0 | what if i was a bog b<br>and you were a bog b |
| **6** | 1419334047927525389 | @LeanneWhittle1 @DerventioeXcel @JacobWhittle1... | PTunners | Derby | 0 | False | 0 | Swim dad. Life l<br>Derby fan. East Sta<br>Sea |



| | id | text | user_screen_name | user_location | retweet_count | favorited | favorite_count | user_description | user_created_at | u |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1419334065375924227 | Rejected sports for the 2020 Olympics include ... | tbelmontedu | New Jersey | 0 | False | 0 | High School Language Arts Teacher. Technology ... | 2013-01-21 00:03:31 | |

and user follower count.



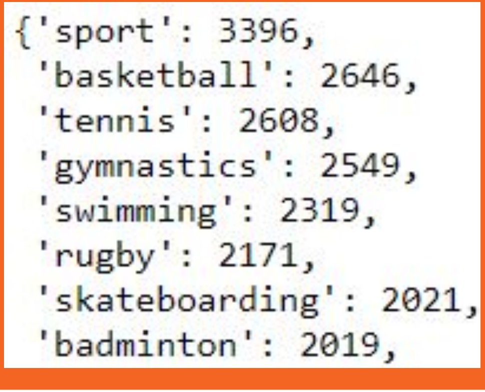
```
followers_bin
1    199515
2      2177
3       429
Name: id, dtype: int64
```

# Question one:
# Which events are the most tweeted?

- Added list of events by using 'https://en.wikipedia.org/wiki/Summer_Olympic_Games' and BeautifulSoup to extract the sports.

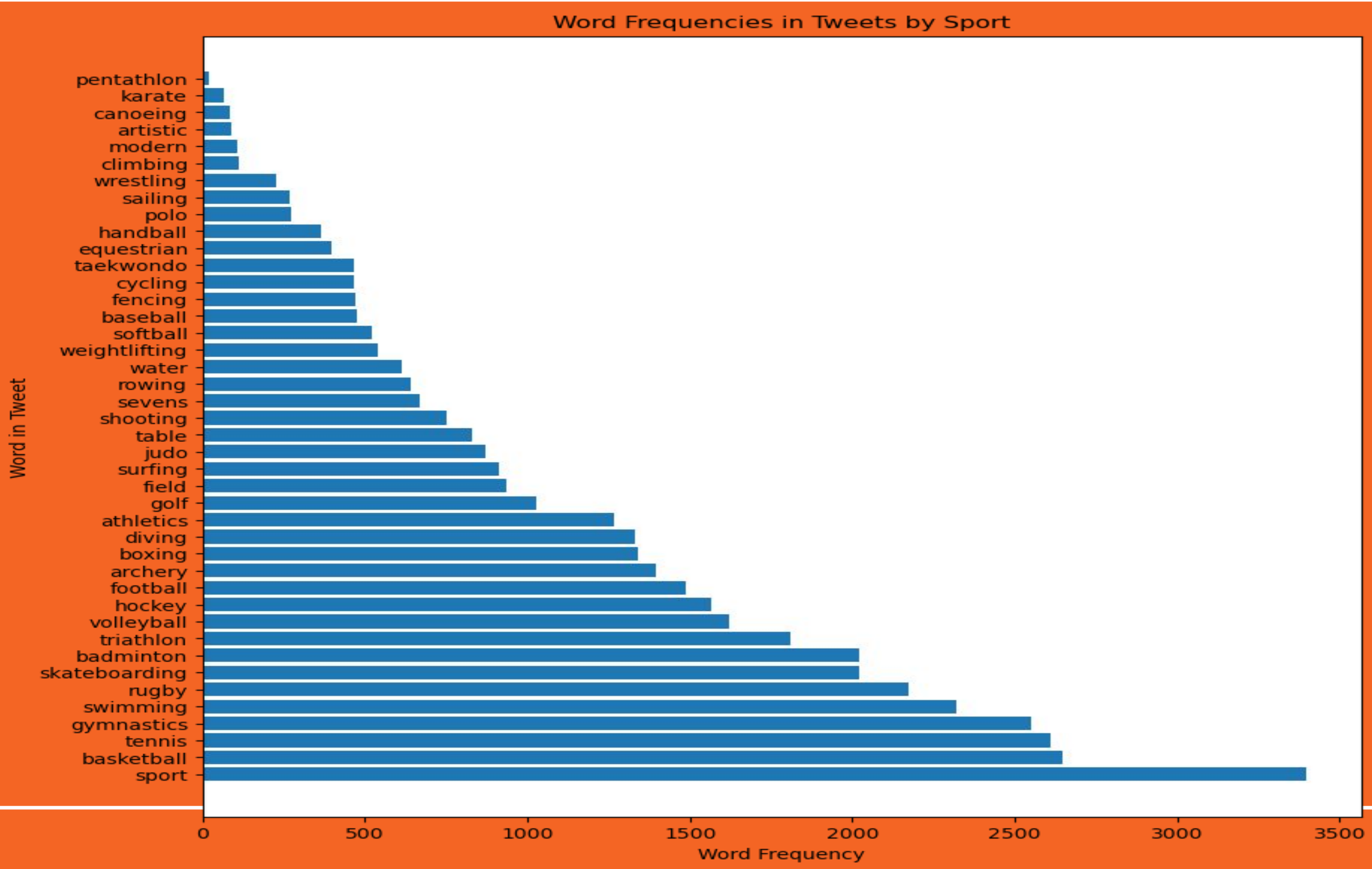
- Used a dictionary of the most popular terms tweeted and cross-referenced it with terms specifically associated with each event.

- Created graph of word frequencies by event.



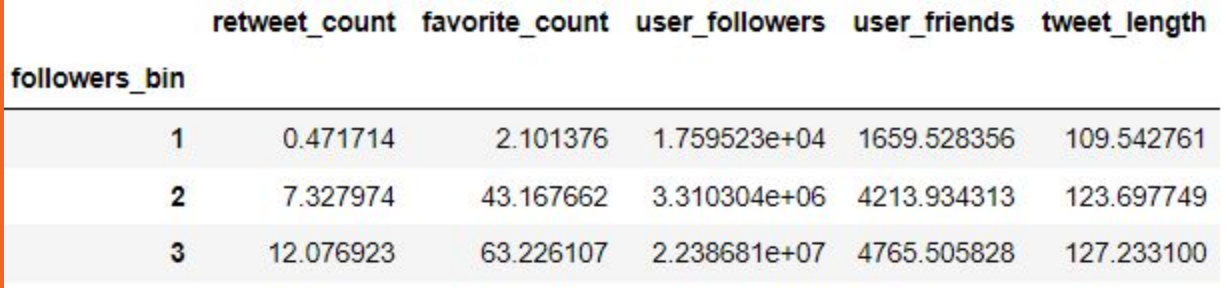
```
{'sport': 3396,
 'basketball': 2646,
 'tennis': 2608,
 'gymnastics': 2549,
 'swimming': 2319,
 'rugby': 2171,
 'skateboarding': 2021,
 'badminton': 2019,
```

# Question one cont: Visualization



# Question two:
# Did users with higher or lower follower counts tweet differently?

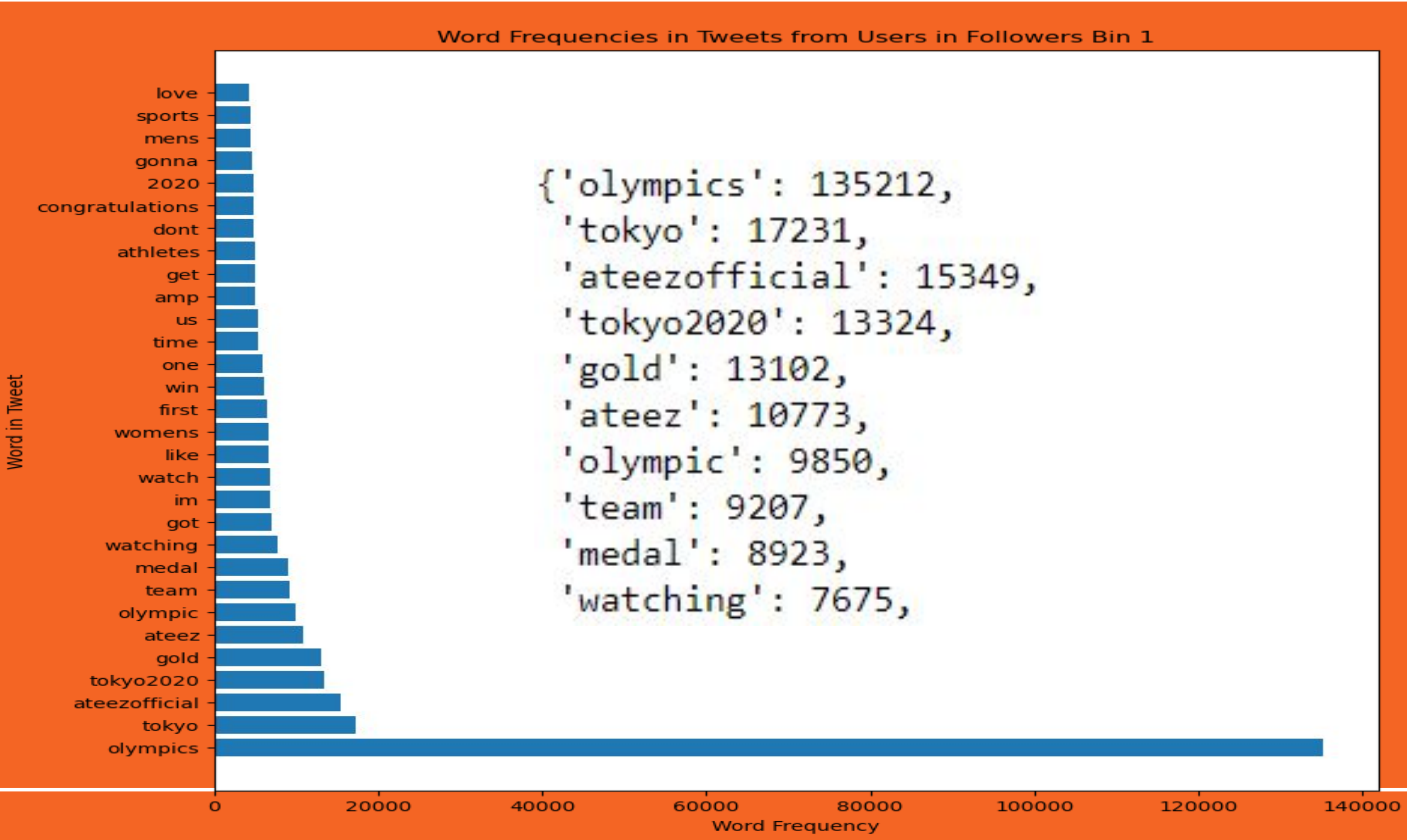
- ➢ Stratified tweets into three bins: less than 1 million, between 1 and 10 million, and 10 million+



| | retweet_count | favorite_count | user_followers | user_friends | tweet_length |
|---|---|---|---|---|---|
| **followers_bin** | | | | | |
| **1** | 0.471714 | 2.101376 | 1.759523e+04 | 1659.528356 | 109.542761 |
| **2** | 7.327974 | 43.167662 | 3.310304e+06 | 4213.934313 | 123.697749 |
| **3** | 12.076923 | 63.226107 | 2.238681e+07 | 4765.505828 | 127.233100 |

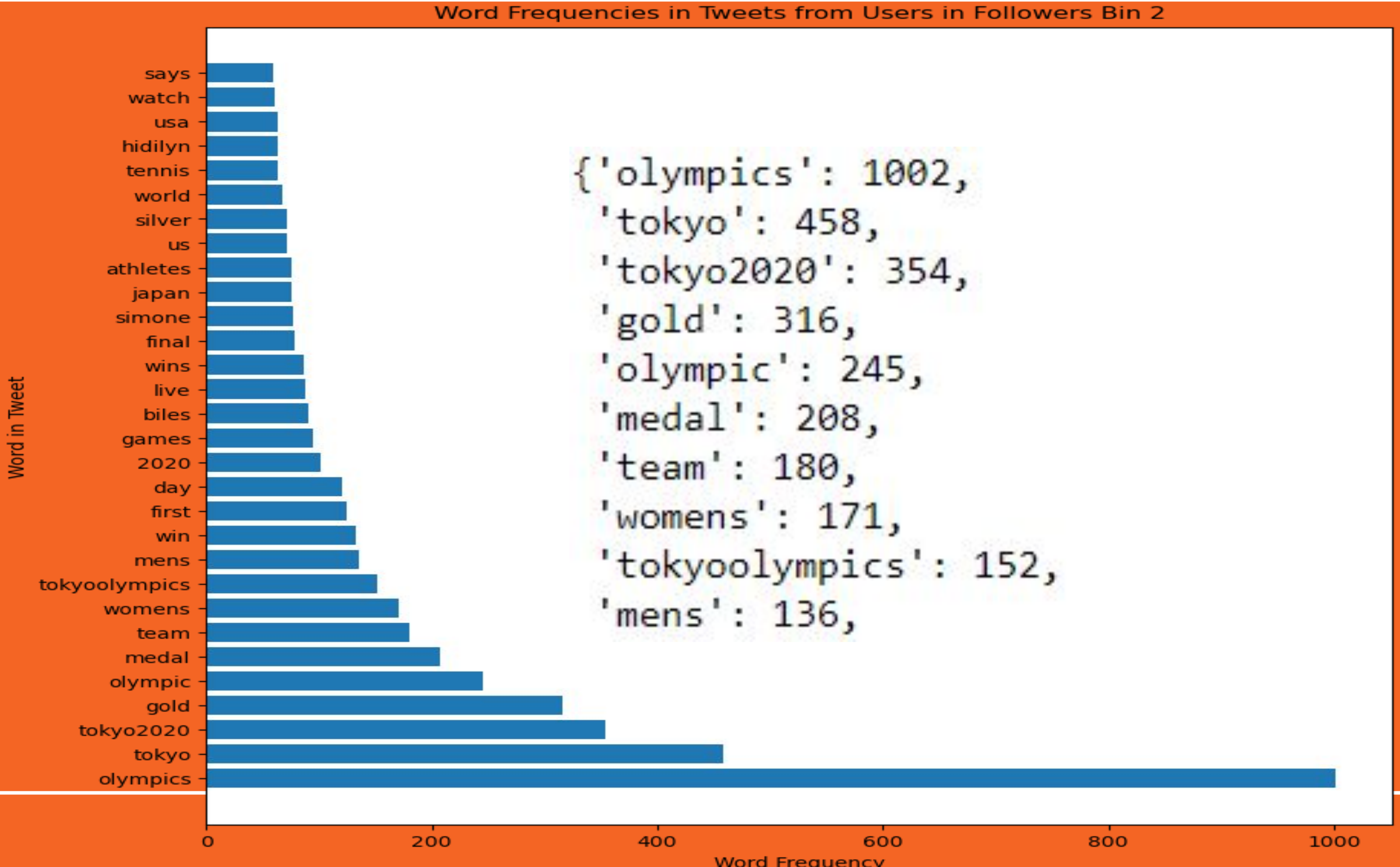
- ➢ Created graph of word frequencies for each of the three groups.

# Question two cont: Visualization Bin 1



```
{'olympics': 135212,
 'tokyo': 17231,
 'ateezofficial': 15349,
 'tokyo2020': 13324,
 'gold': 13102,
 'ateez': 10773,
 'olympic': 9850,
 'team': 9207,
 'medal': 8923,
 'watching': 7675,
```

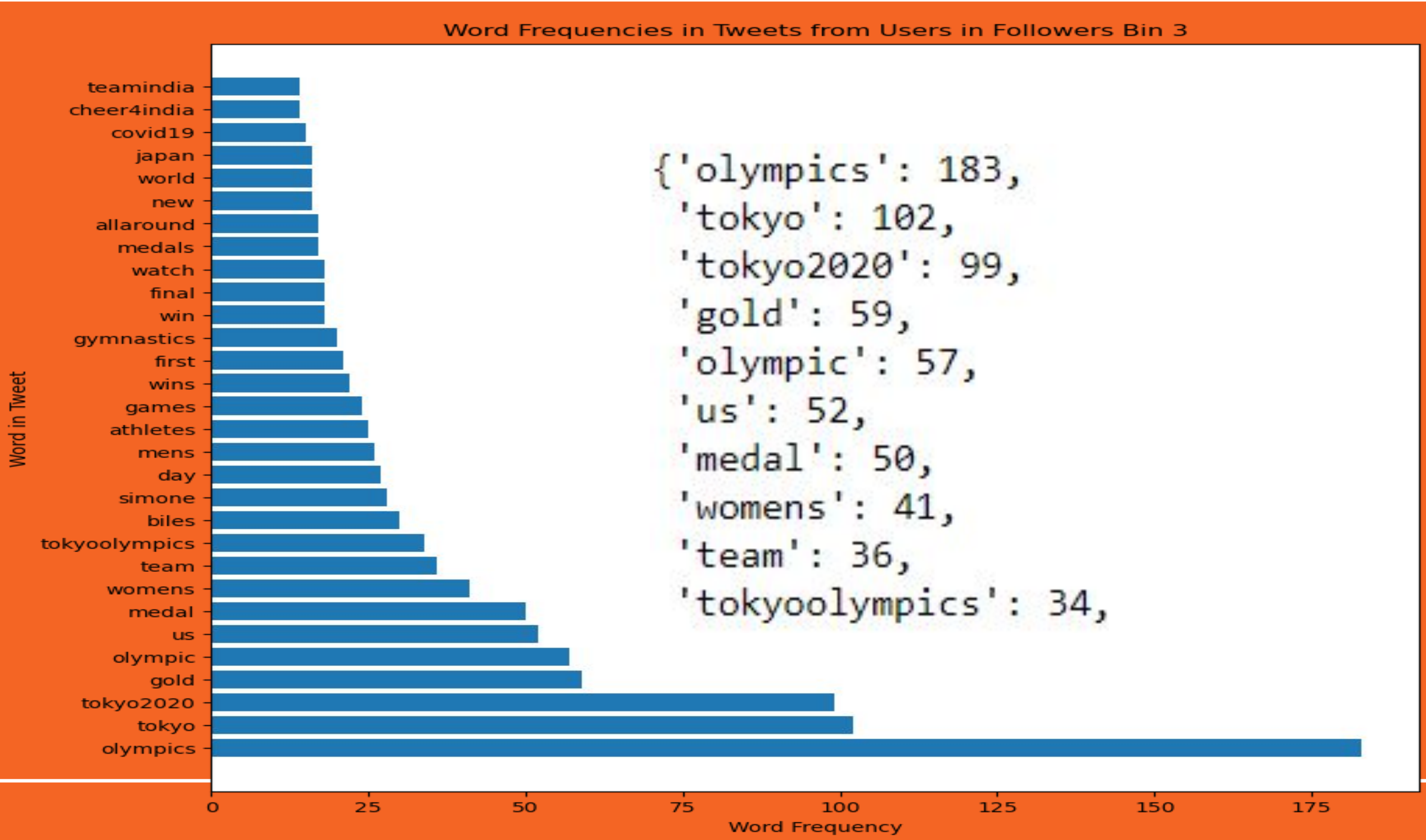
# Question two cont: Visualization Bin 2



Word Frequencies in Tweets from Users in Followers Bin 2

```
{'olympics': 1002,
 'tokyo': 458,
 'tokyo2020': 354,
 'gold': 316,
 'olympic': 245,
 'medal': 208,
 'team': 180,
 'womens': 171,
 'tokyoolympics': 152,
 'mens': 136,
```

# Question two cont: Visualization Bin 3

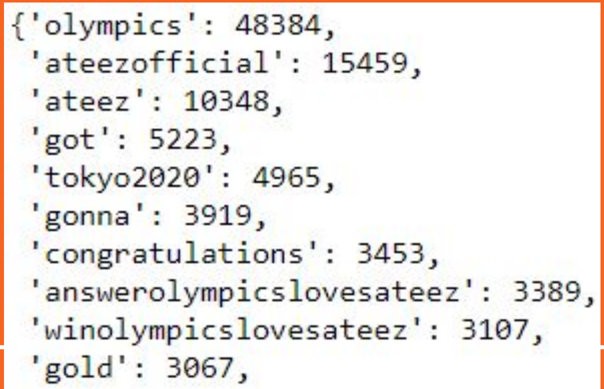


```
{'olympics': 183,
 'tokyo': 102,
 'tokyo2020': 99,
 'gold': 59,
 'olympic': 57,
 'us': 52,
 'medal': 50,
 'womens': 41,
 'team': 36,
 'tokyoolympics': 34,
```

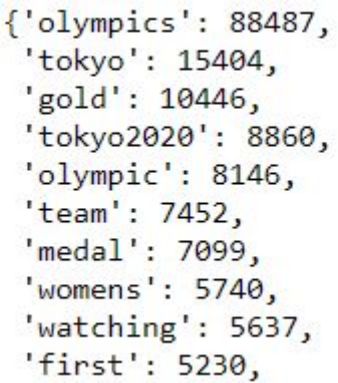
# Question three:
# How were the content in tweets with an "@" different than those without?

- ➢ Split the data frame by tweets that contained an @, and those that did not
- ➢ Created dictionaries of the top 30 words in each group by word frequency

with "@"



```
{'olympics': 48384,
 'ateezofficial': 15459,
 'ateez': 10348,
 'got': 5223,
 'tokyo2020': 4965,
 'gonna': 3919,
 'congratulations': 3453,
 'answerolympicslovesateez': 3389,
 'winolympicslovesateez': 3107,
 'gold': 3067,
```



```
{'olympics': 88487,
 'tokyo': 15404,
 'gold': 10446,
 'tokyo2020': 8860,
 'olympic': 8146,
 'team': 7452,
 'medal': 7099,
 'womens': 5740,
 'watching': 5637,
 'first': 5230,
```

without

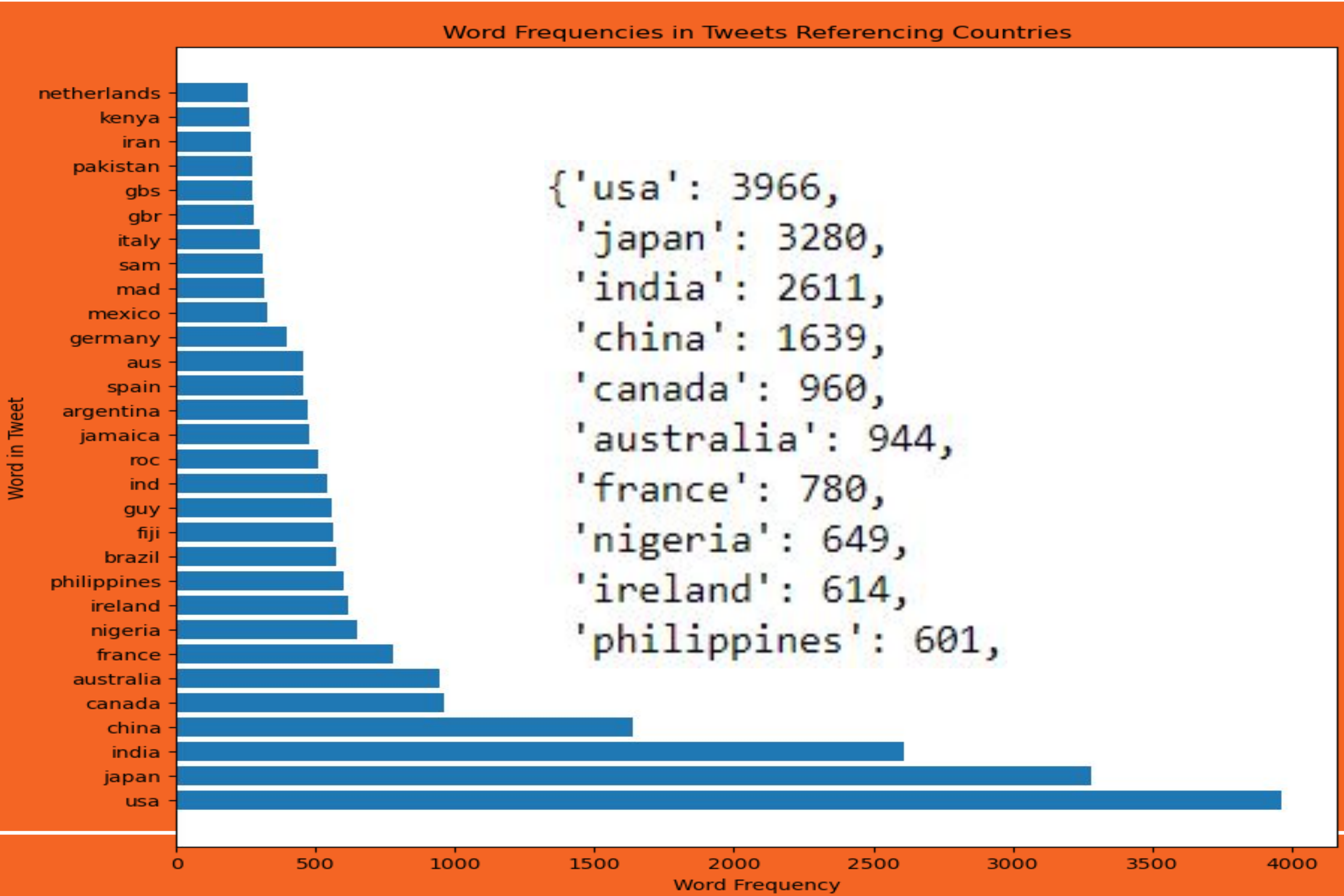
# Question four:
# Which countries were mentioned in tweets the most?

- ➢ Read in external data for a list of country abbreviations (most of our user locations referenced abbreviations.

  https://www.whereig.com/olympics/summer-olympics-participating-countries.html

- ➢ Merged abbreviations with our country field from our medals list. Created word frequency table to track number of times countries were referenced by tweet.

# Question three cont: Visualization



```
{'usa': 3966,
 'japan': 3280,
 'india': 2611,
 'china': 1639,
 'canada': 960,
 'australia': 944,
 'france': 780,
 'nigeria': 649,
 'ireland': 614,
 'philippines': 601,
```

# Conclusions

- Top five most talked about events were Basketball, Tennis, Gymnastics, Rugby, and Skateboarding
- Most common words (other than “olympic(s),” “tokyo,” and “tokyo2020”) across followers were “gold,” “medal,” “team,” and “womens”
- Users in Followers Bin 2 and 3 talked about Simone Biles more than those in Bin 1
- Tokyo 2020 Olympics fans love ATEEZ
- Most talked about countries were USA, Japan, India, China, and Canada.