Courtney Walker

Final Project for Data Sciences

https://github.com/cmw72d/FinalProject_dataset

Due Friday, May 13 at Midnight

I chose the world series/baseball dataset which is job_id 2575, query q=#worldseries, and description baseball one it is finally baseball season and my team won the world series last year (GO ROYALS) and two it is not in Ukraine. I also did Mizzou dataset for the first few questions job_id 4257, query q=#MIZZOU, and description Mizzou. I did this dataset purely out of interest to see what it was about. I would also like to point out that I do not have a twitter so this topic was not the best for me to know what everything was or what to write about.

Descriptive Statistics

1. How many tweets are in the collection?

   There are 2,525,687 tweets in the collection for baseball.
   There are 748,811 tweets in the collection for Mizzou.

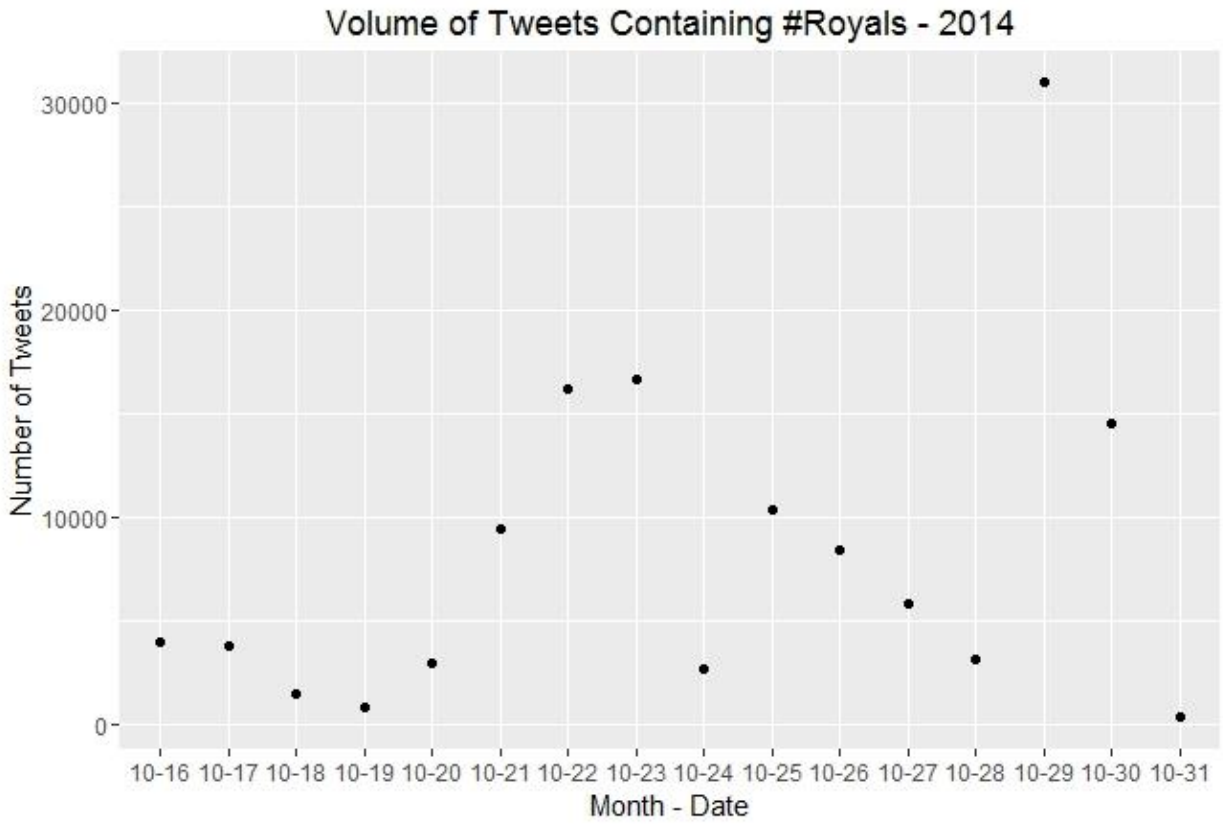
2. When do they start?

   Baseball
   The first tweet started on October 16th, 2014 at 00:40:05.
   Text RT @BaseballTitans: Congratulations to former Titan Christian Colon and the @Royals on advancing to the #WorldSeries! #TusksUp.
   Mizzou
   The first tweet started on November 11th, 2015 at 06:04:50.
   Text @YikYakApp how are you handling the terrorist threats on your app at #Mizzou #PrayForMizzou

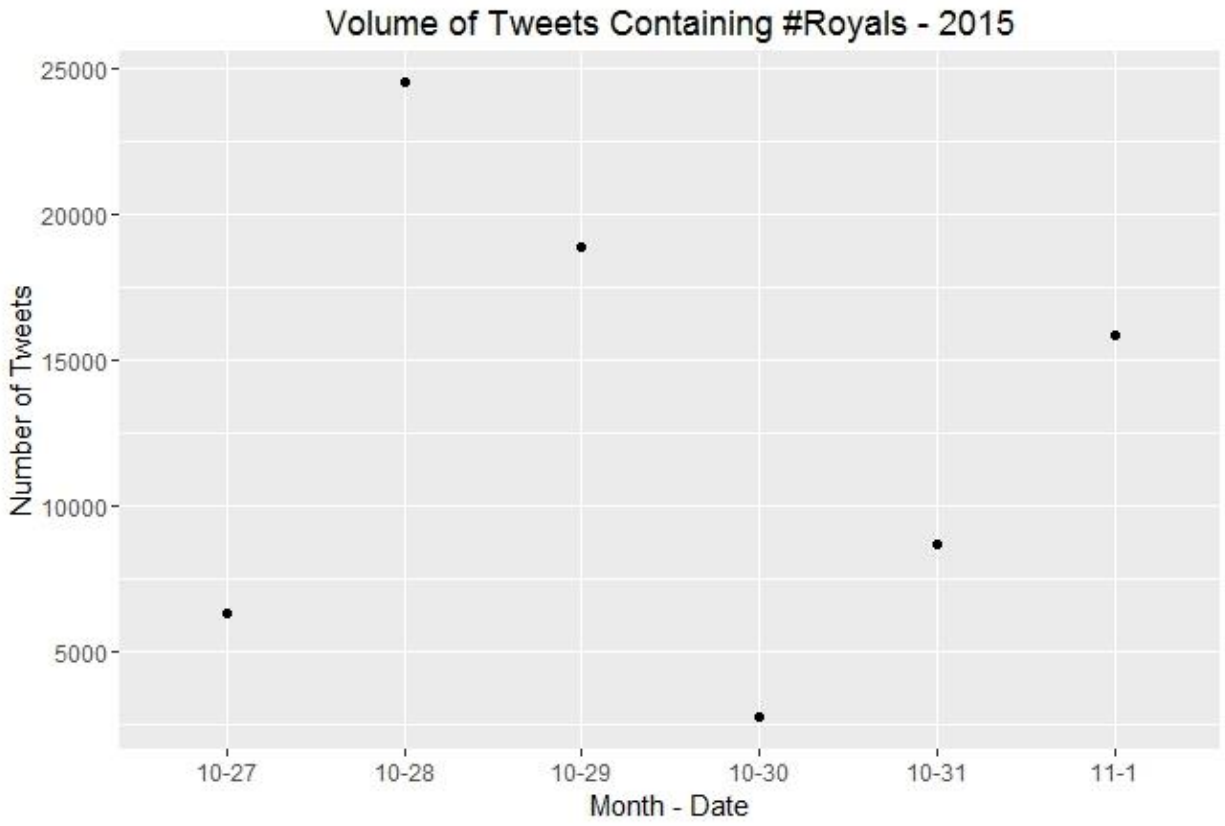
3. When do they end?

   Baseball
   The last tweet was on April 21st, 2016 at 20:40:53.
   Mizzou
   The last tweet was on April 21st, 2016 at 21:20:16.

4. What is the trend for tweet volume?



I chose to look at tweets strictly from the begging of the series Wild Card game was on September 30. So in my R file I started the search from 9-30-14 and ending on game seven of the World Series 10-31-14. 2014 we did not win so notice the drop in text for game seven. However, the peak in game six was a huge win for the Royals. From October 25th to October 28th we had a consistent downfall in tweets.



For some reason my graph did not turn out very well for this year even though I used the same code. This graph seemed to have plotted each peak and once again we loyal fans decided for after game 3 on October 30th to not say much to support our team. Even after our big win on November 1st tweets still did not reach their top peak on October 28th during game two win. It would be interesting to see what it was everyone was tweeting about during game two because we smoked them.

5. If you look at the most common words over the lifetime of the search, do you notice any particular trends associated with those words?

1. #WorldSeries. 86855
2. #Game7 54927
3. #WorldSeries! 44430
4. #MLB 40487
5. #OctoberTogether 38323
6. #Giants 35269
7. #WORLDSERIES 28768
8. #WorldSeries: 25291
9. #WorldSeries, 20245

10. #royals 19446
11. #WorldSeries? 19145
12. #YaGottaBelieve 18954
13. #WickedCity 15460
14. #KCRoyals 15307
15. #AgentsofSHIELD 15086
16. #postseason 14947
17. #earnfromhome 14799
18. #MadBum 14075
19. #KC 13491
20. #WORLDSERIES!!! 12930
21. #NYMvsKC 12784
22. #MetsWIN! 12727
23. #ForeverRoyal 12135
24. #mets 11730
25. #mlb 11385
26. #baseball 11380
27. #KansasCity 10315
28. #NationalAnthem 10046
29. #sfgaints 9103
30. #SFvsKC 8975

This is the first 30 common words over the lifetime.

6. What external events might correspond with the differences in the trends of most common words?

The 30 most common words came as no surprise comparing. Except seeing how many different ways people could hashtag world series. Like I said before I do not have a tweeter and no nothing about it so I do not even understand what a hashtag is for. To me it would make more sense to do an @ symbol so it can go to their tweeter account which will allow them to see it. Anyways, the top word which is World Series is a perfect example of how it corresponds with the event. The World Series is one of the datasets you gave us so it would only make sense that the top word in the dataset would be WORLD SERIES. Two other common words that tie in with the event would be gaints and royals because those are the teams that played each other in 2014. Mets are also in there for 2105. We have not yet played 2016 and the Royals are not doing so great so I was not interested in the 2016 dataset.

7. What hashtags show up as most prominent in each month of the lifecycle?

This is a great question #TakeTheCrown is my favorite hashtag that shows up most prominent during each month along with #Royals and #WorldSeries. #TakeTheCrown is awesome and has so much meaning to our boys in blue so I was glad to see the fans going crazy in October tweeting it for the series. Actually, after our win in 2015 the crown has been newly updated!

8. Which twitter users are the most mentioned?

The top five twitter users that most mentioned with the @ symbol that I said needed to be used earlier are: @MLB with 251,671, @Royals with 176,882, @MLBFanCave with 153,653, @Mets with 74,471, and @SFGiants with 71,931. Once again this came as no surprise with the years we are looking at for the dataset. However, I would be interested to know why @MLBFanCave beat out @Mets and @SFGiants. I could take a guess saying that mets and giants do not have as good of a fan base as royals.

9. How frequently is each user mentioned during each month of the lifecycle?

This question really isn't any different than number 8. If you take each of the five twitter users and divided the number of times they were mentioned in a tweet in the four-month period two months for 2014 and two for 2015, I looked at @MLB was tweeted at 62,917. @Royals was tweeted at 44,220. @MLBFanCave was tweeted at 38,413. @Mets was tweeted at 18,617. Last but not least @SFGiants was tweeted at 17,982.

10. What is the relationship between the volume of tweets you selected and the volume of tweets for other collections in the data set?

I only looked at the volume of tweets for 2014 and 2015. I chose not to look at 2016 because, for starters royals are doing horrible and they are the only team really follow besides cardinals. Second, it is just the beginning of the season and looking at the tweet stats would not be very interesting. Third, the tweets for the World Series which this dataset is based around so the seeing the volume of tweets peak around those couple months is

awesome! I did not look too far into any other dataset's however, I am going to go with a different one for the research question. But I liked this dataset because even though baseball season is seven months long World Series is only a month and a half.

Identifying Research Questions

1. Select one research question to pursue.'
   For my research question I am going to pursue the weather dataset.

2. For that research question, identify candidate machine learning approaches for making sense of this data. This may include the four main approaches discussed already, or it may include new approaches.

   I did my descriptive statistics questions on the baseball/world series dataset but I would like to switch it up for my research question. My research question is looking at the user_timezone for each dataset which time zone has the most natural disaster's occur.

Answering Research Questions

1. Prepare the data set for analysis

2. Create a GitHub Repository for your final project

   https://github.com/cmw72d/FinalProject_dataset

3. Describe the specific data set you are using for the final project (which job_id's).

   The job_id I am looking at is 1160, the query is q=#tornado, and the description is weather. Along with job_id 2335, query = #earthquake, and the description is blank but I am going to tie it in with weather. They could both be described as natural disasters.

4. Build one directory to answer the research question. Include all code and data set references (SQL) there.

5. REPORT: For the research question:
    a. Describe the data carpentry work and software carpentry work you did to obtain both descriptive statistics and answer the questions

        First I used JupyterHub to get inside of the database. Once I was in JupyterHub I would insert and pull the dataset into a data frame in Python. I extracted the data from the job_id = 1160 for tornado's and job_id = 2335 for earthquake's into sequel pro. I used sequel pro to run my queries using the dataset you provided so I could see what was inside each dataset. Finally, I created for/if loops to find the top hashtags, top words, and top mentions in the datasets I looked up that users tweeted.

    b. Provide one, short 3-5 paragraph explanation of your results for your question

```
In [*]: import csv
        import pandas as pd
        from collections import Counter

        try:
            con = mdb.connect('128.206.116.195', 'tg4_ro', '?3stEt7!3hUbRa-R', 'tw4_db');

            #All tweets pertaining to #weather
            df = pd.read_sql_query("SELECT t.text, t.created_at, t.from_user_timezone, MONTH(created_at) AS theMonth, WEEK(created_at) AS theWeek, DAY(created_at) AS the
            df = pd.read_sql_query("SELECT t.text, t.created_at, t.from_user_timezone, MONTH(created_at) AS theMonth, WEEK(created_at) AS theWeek, DAY(created_at) AS the

            df1 = pd.read_sql_query("SELECT MONTH(created_at) AS theMonth, DAY(created_at) AS theDay, YEAR(created_at) AS theYear, TIME (from_user_timezone) AS theTime,
            df2 = pd.read_sql_query("SELECT MONTH(created_at) As theMonth, DAY(created_at) AS theDay, YEAR(created_at) AS theYear, TIME (from_user_timezone) AS theTime,

        except mdb.Error as e:

            print ("Error %d: %s" % (e.args[0],e.args[1]))
            sys.exit(1)


        finally:

            if con:
                con.close()

/usr/lib64/python3.4/site-packages/pandas/io/sql.py:1560: Warning: Truncated incorrect time value: 'Caracas'
  cur.execute(*args)
/usr/lib64/python3.4/site-packages/pandas/io/sql.py:1560: Warning: Truncated incorrect time value: 'Eastern Time (US & Canada)'
  cur.execute(*args)
/usr/lib64/python3.4/site-packages/pandas/io/sql.py:1560: Warning: Truncated incorrect time value: 'Pacific Time (US & Canada)'
  cur.execute(*args)
/usr/lib64/python3.4/site-packages/pandas/io/sql.py:1560: Warning: Truncated incorrect time value: 'Athens'
  cur.execute(*args)
/usr/lib64/python3.4/site-packages/pandas/io/sql.py:1560: Warning: Truncated incorrect time value: 'La Paz'
  cur.execute(*args)
/usr/lib64/python3.4/site-packages/pandas/io/sql.py:1560: Warning: Truncated incorrect time value: 'Madrid'
  cur.execute(*args)
/usr/lib64/python3.4/site-packages/pandas/io/sql.py:1560: Warning: Truncated incorrect time value: 'Paris'
```

I thought this was a great thing to happen during a research question. I was having a huge problem coding from_user_timezone I could not figure out what value it was using. I tried timezone, location, city, and finally time. Time worked I just received warning because my version of JupyterHub considers them incorrect time values. However, I found online while googling the error that more recent

versions do not consider it an error. It was awesome to be able to see all the different time zones such as: Caracas, Eastern Time (US & Canada), Athens, La Paz, etc. To me these all looked like cities until I got into Eastern and Pacific time which I am used to seeing. This is what made it so hard to determine what to use for my query. The trial and error of research is what makes it so much fun.

Here are a few of the top hashtags in the two datasets combined:

1. #earthquake 440371
2. #Earthquake 155447
3. #saigai 110824
4. #jishin 108697
5. #earthquake. 56073
6. #?? 54593
7. #quake 37752
8. #iPhone 30209
9. #tsunami 18975
10. #EarthQuake 17404
11. #alert 14661
12. #Sismo 13795
13. #jish 13533
14. #Earthquake: 13513
15. # 13286
16. #Japan 11978
17. #??? 11789
18. #UnitedStates 11581
19. #Breaking?#earthquake?M 11404
20. #prayfromjapan 11267
21. #???#earthquake 10858
22. #eart 9287
23. #Ecuador 8296
24. #Quake 8237
25. #???? 8159

It is not until really far down the line that we start seeing tweets about tornados. It has only been tweeted 99 times in the dataset from April, 2014 through March, 2016. Even though both are horrible natural disaster's is seems as though earthquakes happen more frequently and cause a bigger disaster when they strike hard. Looking at the time zones I notice that Tokyo is in one of the first few most mentioned and comparing it to the top hashtags we can see some similarities. Japan was in the top most used hashtags as well.

This was a very time consuming project which, is why I am turning it in last minute. Thankfully you narrowed it down on top of letting us chose what dataset we wanted to use so it did not make it so horrible. I would be interested to continue to watch the World Series tweets with the royals doing so well the last couple of years. However, weather has always fascinated me even though I am

terrified of tornados and never seen one in person. Earthquakes I remember studying what causes them in school when I was younger. In that two datasets I noticed that tsunami was in the top hashtag tweets which, could be because some earthquakes can cause them. This might be why earthquakes are tweeted a significant amount more. Time zone was an interesting research question so I can know what parts of the World to stay away from such as: London, Paris, Hawaii, Athens, Tokyo, and Brasilia.