CS4001: Data Science Final Project - #WorldSeries

Rachel Hankins

May 13th, 2016

https://github.com/rlh5z6/data_science_final

DATA SET

I am using the #WorldSeries data set which is a collection of tweets containing the created date, the tweet, the user's information, the location, and various other information. The collection is about 2.5 million tweets and spans the length of just over 16 months.

Note: The 2014 World Series was the Kansas City Royals vs. the San Francisco Giants and the 2015 World Series was the Kansas City Royals vs. the New York Mets.

DESCRIPTIVE STATISTICS

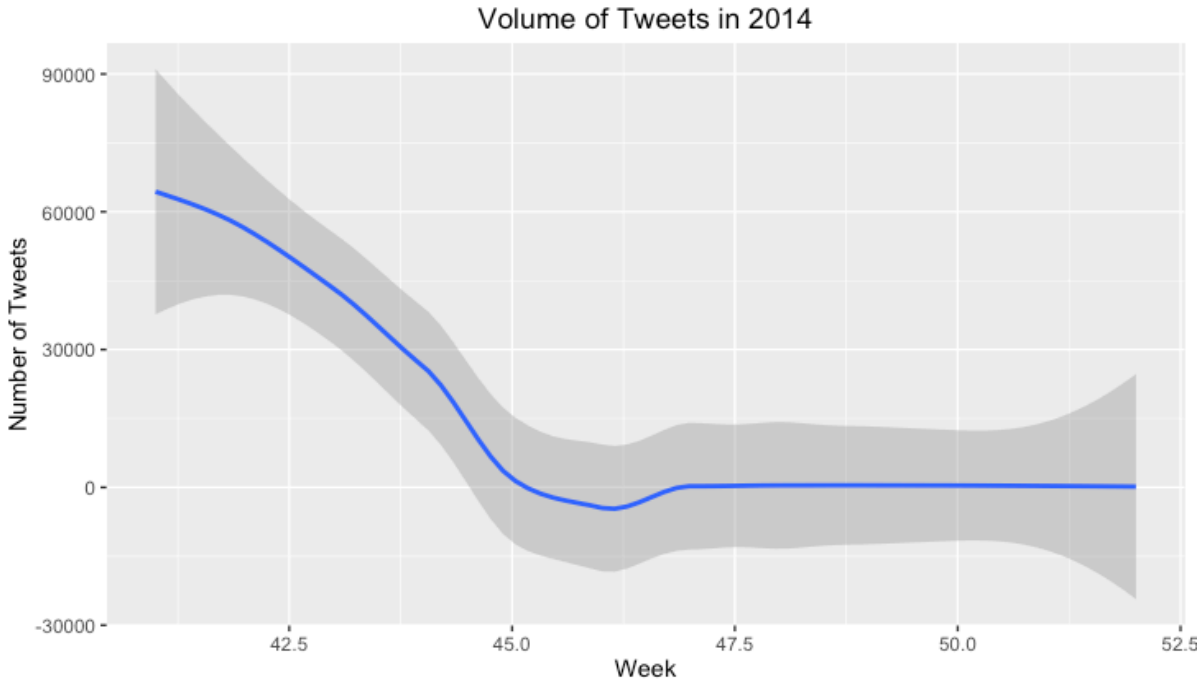
1. **How many tweets are in the collection?**
   2,525,687 tweets

2. **When do they start?**
   2014-10-16 00:40:05

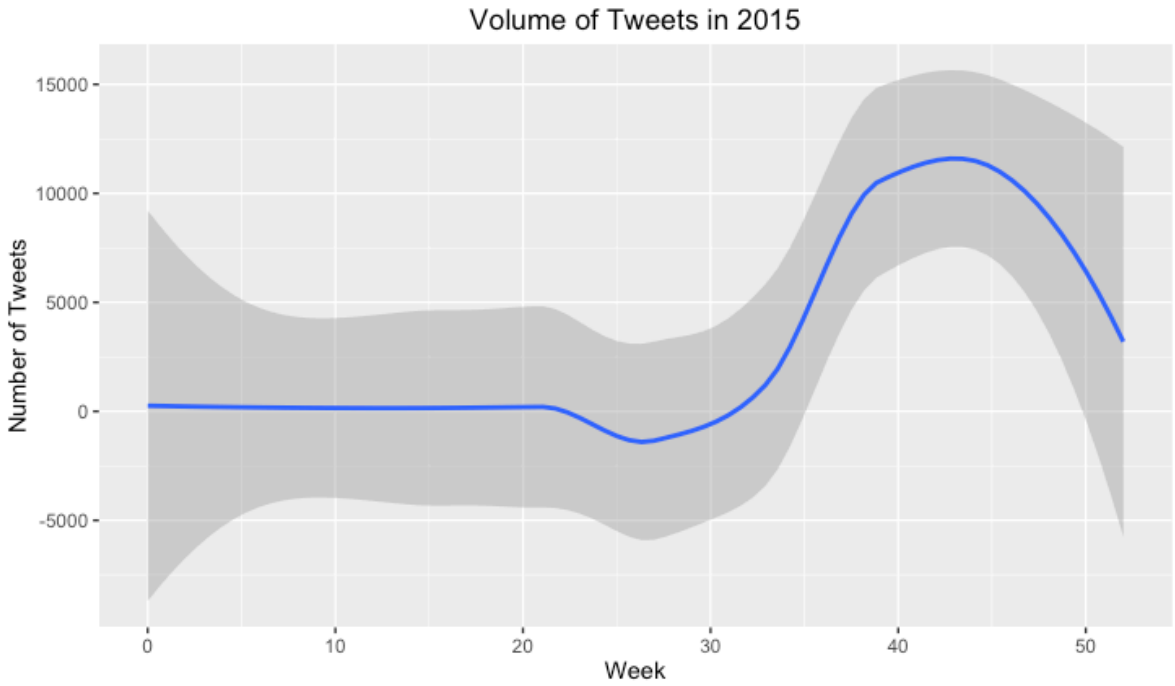
3. **When do they end?**
   2016-04-21 20:40:53

4. **What is the trend for tweet volume?**

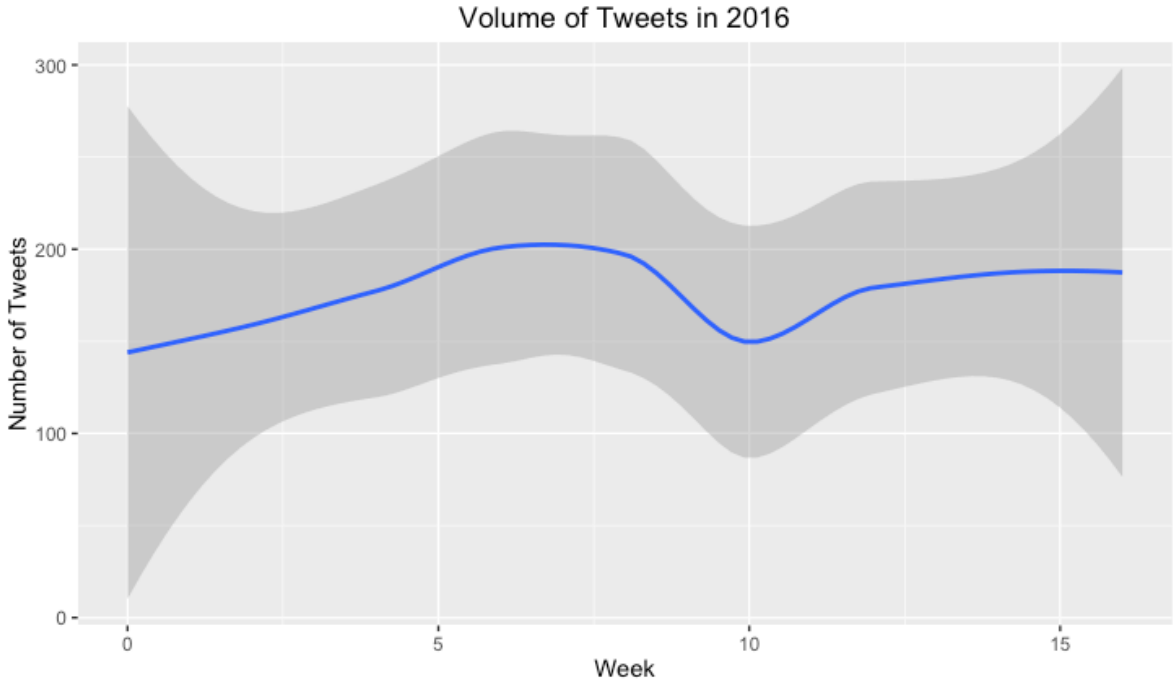
2014 – The data collection starts on October 16, 2014, right in the peak of the playoffs. This caused a high volume of tweets during October and into early November. After the playoffs were over, the tweet volume decreased and eventually staggered out.



2015 – Around Week 27 (beginning of July), the #WorldSeries tweets starting picking up. The volume gradually increased over the course of the next ten or so weeks, peaking around week 42, mid-October, right when the World Series is played.



2016 – Being that the data is only collected until late April, there is not much to show yet for 2016. The data is basically staggered going through small increases and decreases over the first few months of the year.



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) Game, 249678
    2) win, 207606
    3) game, 110724
    4) 1, 86444
    5) Royals, 83139
    6) 2, 69539
    7) Mets, 64829
    8) first, 60567
    9) Kansas, 56162
    10) World, 54419
    11) 3, 52226
    12) 4, 45430
    13) Series, 43870
    14) 7, 41596
    15) lead, 39505
    16) chance, 38240
    17) baseball, 36669
    18) home, 35505
    19) Madison, 32927
    20) Bumgarner, 31504
    21) Joe, 25570
    22) Congrats, 24090

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

    The most popular event that would trigger trends of the most common words is the World Series. The World Series is the final series played in the Major League Baseball playoffs. The entire playoffs season lasts from about late September to late October/early November. The most popular words are variations of "World" and "Series." These words are most commonly found when the World Series winner is crowned. Other spikes in common words are in July and August, when trade deadlines and playoff season is upon. Because baseball is played April through October/November (depending on year), words, hashtags, and mentions can be found throughout that time.

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

    The following hashtags are the most prominent in the month of October, playoffs season. Since the hashtags all spike in October, I decided to look into it a little further. I subset the data into the 2015 year and then further into primarily

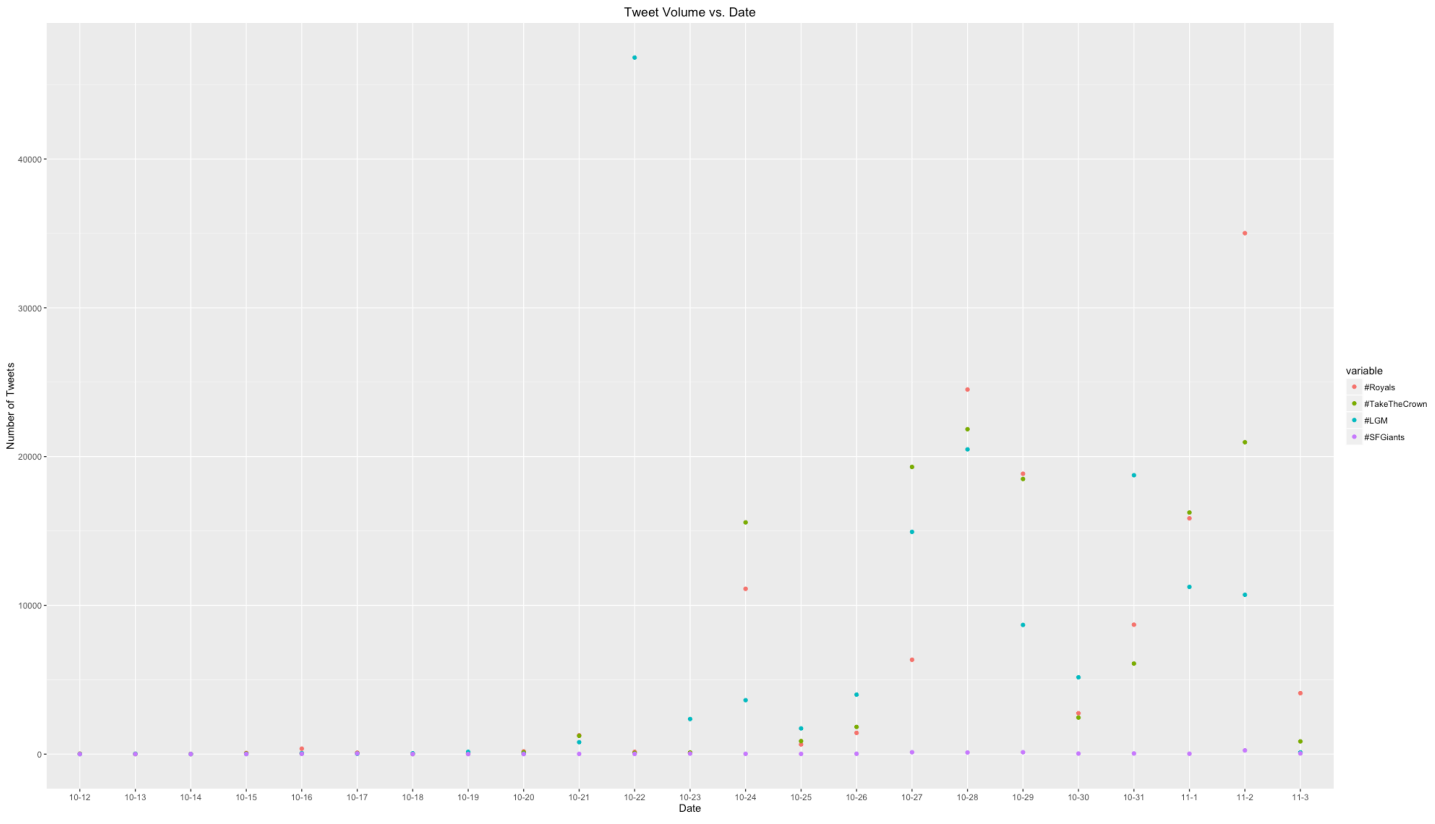
the playoffs season. I broke it down day by day for the top 4 hashtags. The graphs can be found on GitHub under the hashtags folder. Since every tweet contained #WorldSeries to be found in the data set, I discounted it to get a better measurement on the other four hashtags.

#TakeTheCrown - #TakeTheCrown spiked on October 24th and October 27th-November 2nd. This can be further explained by the fact that the Royals clinched the World Series (won the ALCS) on October 24th and the World Series was played through the other week. Once the Royals won the World Series, #TakeTheCrown changed to #TookTheCrown.

#Royals - #Royals gradually increased throughout the second half of October, but spiked on November 2nd, when the Royals won the World Series.

#LGM - #LGM stands for Let's Go Mets. It spiked on October 22nd and kind of died off through the World Series. On October 22nd the Mets clinched the World Series (won the NLCS). Since the Royals dominated the Mets in the series, the social media fan base for the Mets slowed down.

#SFGiants - The San Francisco Giants were the 2014 World Series winners. Because I only took October 12th through November 3rd into account, the SF Giants social media presence is very lack thereof.
*Note: This plot can be found in my GitHub repository in the **hashtags_plots** folder.*



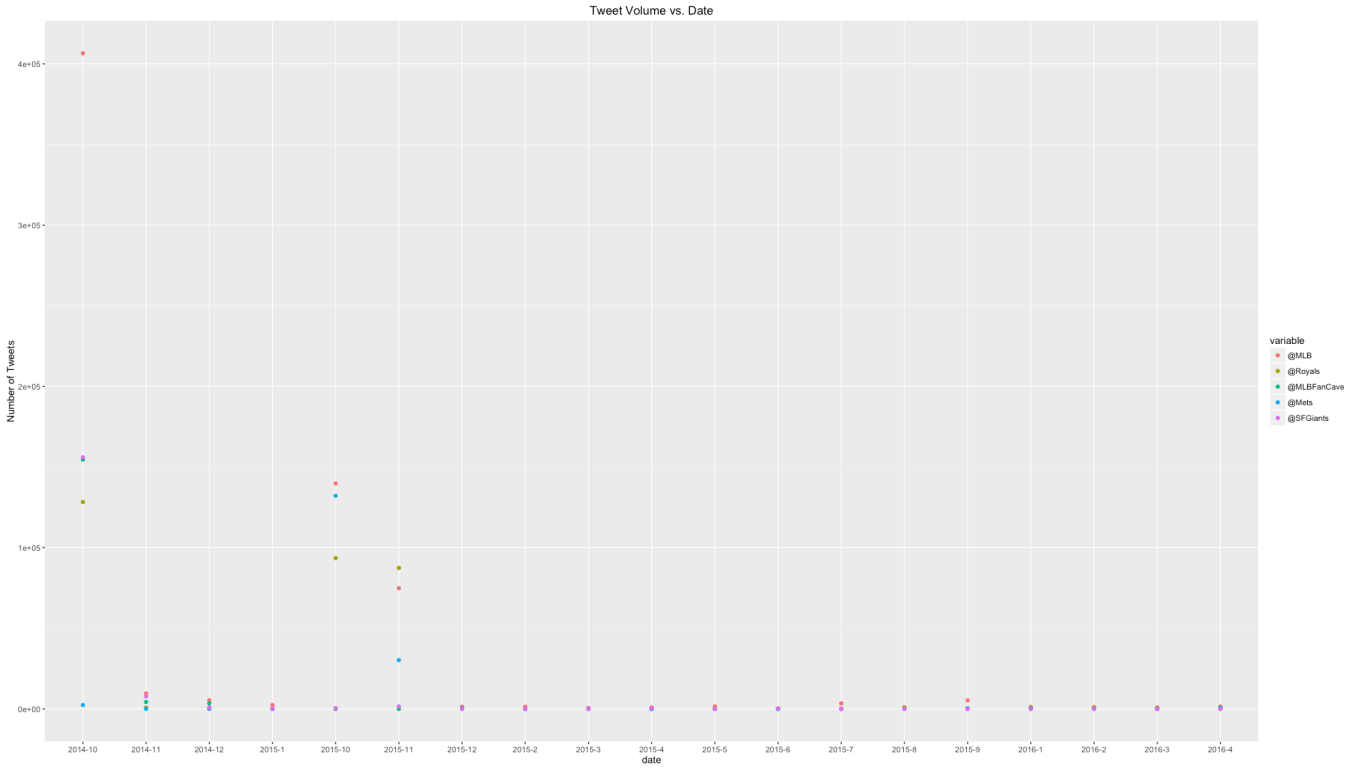
1) #WorldSeries, 2035883
2) #Royals, 233375
3) #TakeTheCrown, 163508
4) #LGM, 144025
5) #SFGiants, 135765
6) #Mets, 101511
7) #Game7, 54927
8) #MLB, 40487
9) #OctoberTogether, 38323
10) #Giants, 35269
11) #YaGottaBelieve, 18954
12) #WickedCity, 15460
13) #postseason, 14947
14) #NYMvsKC, 12784
15) #ForeverRoyal, 12135

8. Which twitter users are the most mentioned?

1) @MLB, 251671
2) @Royals, 176882
3) @MLBFanCave, 153653
4) @Mets, 74471
5) @SFGiants, 71931
6) @MLBONFOX, 70106
7) @MLBMeme, 27511
8) @MLBNetwork, 25389
9) @MLBGIFs, 22864
10) @sluggernation, 20298
11) @HeymanHustle, 17545
12) @BBTN, 15394

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

The following users are some of the top mentions found in the data set. Once again these users are all spiked around October since #WorldSeries isn't used as much outside of that month. The top mentions are either Major League Baseball accounts or the three teams who played in the 2014 or 2015 World Series. Both @MLB and @MLBFanCave greatly plateaued off. After further investigation @MLBFanCave account does not exist anymore. *Note: This plot can be found in my GitHub repository in the **mentions_plots** folder.*



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

The volume of tweets in the #WorldSeries data set definitely spikes around the central focus of the World Series. It peaks throughout playoff season (October). Unlike other data sets, which are event specific, tweets can be found throughout the year as baseball is a seven, month sport. Some of the other data sets are event triggered and only last for a small window of time outside of the event. Another interesting point on this data set is there is a build up until the event. Unlike unexpected events such as the Paris terrorist attacks or Florida tornados, this data set has a gradual increase up until the World Series.

RESEARCH REPORT

Research Question
What external events are causing the trend in the top 5 players tweeted about? How might these events affect their social media presence? What makes a player talked about on social media?

Machine Learning
First, I will start with finding the top 5 players mentioned in tweets. This can be by their first name, last name, or nickname. I will further break the data set down by the hour. After I go through and find the top players, I will then pull the tweets where they are mentioned and add them to a data frame in Python. I will use my word count function to find the top words associated with each player. Last, I will do further research to better understand why these top words are associated with them. I would use linear regression to see a correlation between the top players and the top words associated with them.

Data Carpentry
Twitter Data

1. Extract the data for the job_id = 2577 in MySQL. This pulls only the data related to the #WorldSeries.
2. Using Jupyter, upload the data set that was pulled and insert it into a data frame in Python.
3. Create a function that uses the Counter() function to pull the most common words in the tweets from the data frame.
4. Convert data frame to a CSV.
5. Using, R, plot the words, hashtags, or mentions usage over a certain period of time.

Results
The following five players are the most tweeted about players during the 2014 and/or 2015 World Series. Each player is tweeted about for a varierty of different reasons, but my goal was to form a further grasp on why these players are being talked about. Is it because they're phenomenal athletes? Did they make a game winning play or maybe a game costing error? Or are they just highly talked about as people instead of players? I dug further into this question by finding five of the top words also in tweets that mentioned them. Some players had more of a variety in their words while others had a more centralized event that caused them to be talked about more.

Top 5 Tweeted About Players

1. **Madison Bumgarner – San Francisco Giants (Pitcher)**
    1) **Shutout** – Madison Bumgarner pitched a shutout (no runs scored by opposing team) in Game 5 of the 2014 World Series.

2) **ERA** – Madison Bumgarner had an ERA (earned run average) of .25 in the World Series. This is the lowest ERA ever in the World Series.
3) **History** – Bumgarner pitched an incredible series and broke several records. His series pitching is known to go down in history.
4) **#AceStatus** – An ace in baseball is the best starting pitcher on a team and nearly always the first pitcher in the starting rotation. After Bumgarner's incredible starts, the hashtag, #AcePitcher trended.
5) **MVP** – Bumgarner was named the World Series MVP after the San Francisco Giants won the World Series in five games. He was selected for his 2-0 record, 1 save after 3 appearances, his shutout, and his 17 strikeouts.

2. Daniel Murphy – New York Mets (2nd Base)

1) **2B** – Daniel Murphy is the starting second baseman for the New York Mets.
2) **Goat** – This trend was decently interesting. There are two explanations behind the word "goat" being a trending word. First, the name of the cursed goat of the Chicago Cubs is Murphy. Second, when Daniel Murphy made the critical, game-costing error in Game 4, many people took to social media to call him the cursed goat of the New York Mets.
3) **Human** – The New York Mets twitter account published a tweet that said, "DANIEL MURPHY YOU ARE NOT HUMAN." followed by a GIF of three of Daniel Murphy's playoff homeruns. Murphy led the team in all aspects of hitting throughout the series.

4) **4** – Daniel Murphy committed an error in the top of the 8th in Game 4 that cost them the lead and ultimately the game. This error caused him to go from New York hero to most hated man in New York in a matter of minutes.
5) **Duda** – Like Murphy, he committed an error in Game 5 that cost Mets the lead, ultimately the game

3. Matt Harvey – New York Mets (Pitcher)

1) **Mound** – In Game 5 of the World Series, Matt Harvey asked to be left on the mound in the ninth inning to complete a full nine innings. After a slew of hits and errors, Harvey was pulled. Fans took to Twitter to cause an uproar over if Harvey should have been left on the mound. Note: Harvey also sprinted to the mound in the ninth inning and fans found it quite funny.
2) **Ninth** – This word also falls under the category of Harvey staying in to pitch in the ninth inning.
3) **Collins** – This word also falls under the category of manager Terry Collings choosing to leave Harvey in to pitch in the ninth inning.
4) **#HarveyDay** – November 1st (Game 5) of the World Series was deemed Harvey Day by the New York Mets. Harvey was on fire going into Game 5 and would be the starting pitcher.
5) **Hero** – "You either come out of the game soon enough to be the hero, or you stay in long enough to become the villain." Harvey started as the zero in the game, but ended as the villain. Harvey is also the name is also the

name of a character associated with Batman which can contribute to the reasoning behind the word “hero” being used so frequently.

4. **Hunter Pence – San Francisco Giants (Right Field)**
    1) **RBI** – Hunter Pence had a large number of RBIs (runs batted in) throughout the playoffs as well as scored of several RBIs.
    2) **Sandoval** – Hunter Pence and Pablo Sandoval became the third set of teammates in World Series history to each have 12 hits.
    3) **Home Alone** – Fans took to Twitter to talk about how Hunter Pence looks like the robber from Home Alone.
    4) **Tie** – In Game 4 of the 2014 World Series, Hunter Pence tagged up on a fly ball hit to center to tie the game. Fans also took to twitter to point on that his shoes were untied at a few points throughout the playoffs.
    5) **Signs** – Fans took to social media to post some of the hilarious, yet ridiculous signs made about Hunter Pence.

5. **Edinson Volquez Kansas City Royals (Pitcher)**
    1) **Cueto** – After the team learned of Volquez's father's death, Cueto dedicated Game 2 to Volquez's father. Cueto pitched an incredible game.
    2) **Father** – Edinson Volquez's father died before Game 1 (October 27th) of the World Series. Volquez was set to start this game. The management and his family made the decision not to tell Volquez prior to the game's start. He pitched a phenomenal six innings before being pulled and given the news. Fans took to Twitter to give their condolences and discuss Volquez's performance.
    3) **Dad** – This word also falls under the category of Volquez's father's death.
    4) **Heart** – This word also falls under the category of Volquez's father's death. He died of a heart condition.
    5) **Knowing** – There was large controversy whether Volquez knew prior to Game 1 of his father's passing. It was revealed later that he did not.

After further investigation using the Internet, Twitter, and plotting the top players and words in R, I have come to the conclusion that there are a few different reasons a player may be talked about. You can find plots of my findings in the **top_players_plots** folder in my GitHub repository.

Madison Bumgarner – Bumgarner was the World Series MVP in 2014 and pitched phenomenally throughout the entire playoffs. He was tweeted about over a longer period of time, but spiking when he did pitch or won the World Series MVP title.

Daniel Murphy – Murphy was the hero for the New York Mets, especially with his offensive skills. After Murphy completed the crucial error in Game 4, discussion on him sailed. He went from being the hero to the most hated man in New York. His numbers half spiked when he did well but peaked when he made the historic error. It just shows that people remember the bad more then they remember the good. Note: The Chicago Cubs cursed goat being named Murphy also may have skewed his numbers.

Matt Harvey – Matt Harvey is a starting pitcher for the New York Mets. He was a hero to them all season and was hoping to carry them to a World Series win in Game 5. Harvey pitched an incredible first eight innings and begged manager, Terry Collins to keep him in. The Royals offense opened up and Matt Harvey found himself in trouble. Harvey was remembered as both a hero and for making a selfish decision to stay in.

Hunter Pence – Although Pence did well throughout the World Series that is not what fans took to twitter to talk about him for. They talked about everything from his Home Alone doppelganger to the ridiculous fan created signs. Pence is known to be an out there player, so fans can find it easy to talk about him. His social media presence is almost more based on his personality then on his skills.

Edinson Volquez – Volquez's father died hours before he was supposed to start Game 1 of the World Series. Fans actually knew of his father's death before he did and took to Twitter to contribute to the conversation to pay their condolences. Although Volquez did well over the course of the series, the loss of his father is the primary reasoning behind why he was discussed on Twitter.

Ultimately, I have found that players are highly discussed on Twitter for four different reasons:

1. High caliber athlete, potential MVP candidate, makes more than one phenomenal play
2. Makes a single critical error
3. Easily talked about on anything besides athletic abilities
4. Tragic event

Potential Problems with the Data Set

This is a list of issues that I encountered throughout my project that may have skewed the data.

- Retweets alter the amount of times words, hashtags, and users are counted
- Analyzing data over multiple World Series and the Royals being in it both years
- Capitalization and punctuation
- Varying time zones of tweets