Text

Description automatically generated

**Course:** 2023-0405 IST 652 Scripting for Data Analysis

**Name:** Jacob N. Conard

**Assignment:** Final Project

**Date due:** 09 June 2023

**Date submitted:** 04 June 2023

**Number of Pages:** 16

**Final Project**

**About the Data:**

The datasets that I am working with for the Final Project is two College Basketball Datasets. The first dataset is from Division 1 College Basketball seasons from 2013 – 2019. The dataset has 2,455 rows and 24 columns. The dataset is in a csv that can be downloaded from Kaggle (link is provided below). There are many variables that are included in this dataset that are computed stats related to College Basketball team performance. Those variables and their definitions are listed below:

* TEAM: The Division I college basketball school
* CONF: The Athletic Conference in which the school participates in (A10 = Atlantic 10, ACC = Atlantic Coast Conference, AE = America East, Amer = American, ASun = ASUN, B10 = Big Ten, B12 = Big 12, BE = Big East, BSky = Big Sky, BSth = Big South, BW = Big West, CAA = Colonial Athletic Association, CUSA = Conference USA, Horz = Horizon League, Ivy = Ivy League, MAAC = Metro Atlantic Athletic Conference, MAC = Mid-American Conference, MEAC = Mid-Eastern Athletic Conference, MVC = Missouri Valley Conference, MWC = Mountain West, NEC = Northeast Conference, OVC = Ohio Valley Conference, P12 = Pac-12, Pat = Patriot League, SB = Sun Belt, SC = Southern Conference, SEC = South Eastern Conference, Slnd = Southland Conference, Sum = Summit League, SWAC = Southwestern Athletic Conference, WAC = Western Athletic Conference, WCC = West Coast Conference)
* G: Number of games played
* W: Number of games won
* ADJOE: Adjusted Offensive Efficiency (An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division I defense)
* ADJDE: Adjusted Defensive Efficiency (An estimate of the defensive efficiency (points allowed per 100 possessions) a team would have against the average Division I offense)
* BARTHAG: Power Rating (Chance of beating an average Division I team)
* EFG\_O: Effective Field Goal Percentage Shot
* EFG\_D: Effective Field Goal Percentage Allowed
* TOR: Turnover Percentage Allowed (Turnover Rate)
* TORD: Turnover Percentage Committed (Steal Rate)
* ORB: Offensive Rebound Rate
* DRB: Offensive Rebound Rate Allowed
* FTR : Free Throw Rate (How often the given team shoots Free Throws)
* FTRD: Free Throw Rate Allowed
* 2P\_O: Two-Point Shooting Percentage
* 2P\_D: Two-Point Shooting Percentage Allowed
* 3P\_O: Three-Point Shooting Percentage
* 3P\_D: Three-Point Shooting Percentage Allowed
* ADJ\_T: Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo)
* WAB: Wins Above Bubble (The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it)
* POSTSEASON: Round where the given team was eliminated or where their season ended (R68 = First Four, R64 = Round of 64, R32 = Round of 32, S16 = Sweet Sixteen, E8 = Elite Eight, F4 = Final Four, 2ND = Runner-up, Champion = Winner of the NCAA March Madness Tournament for that given year)
* SEED: Seed in the NCAA March Madness Tournament
* YEAR: Season

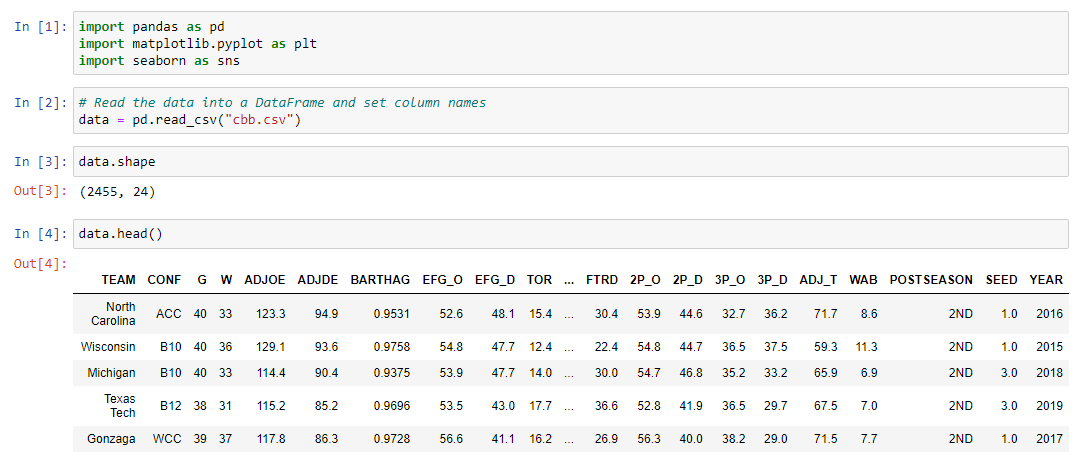
The first dataset can be found on Kaggle at the following link: <https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset>

The second dataset that I am working with for the Final Project is a College Basketball Twitter Dataset. The data includes College Basketball Tweets from March 27th 2017. The file is a text file of JSON data that includes the following:

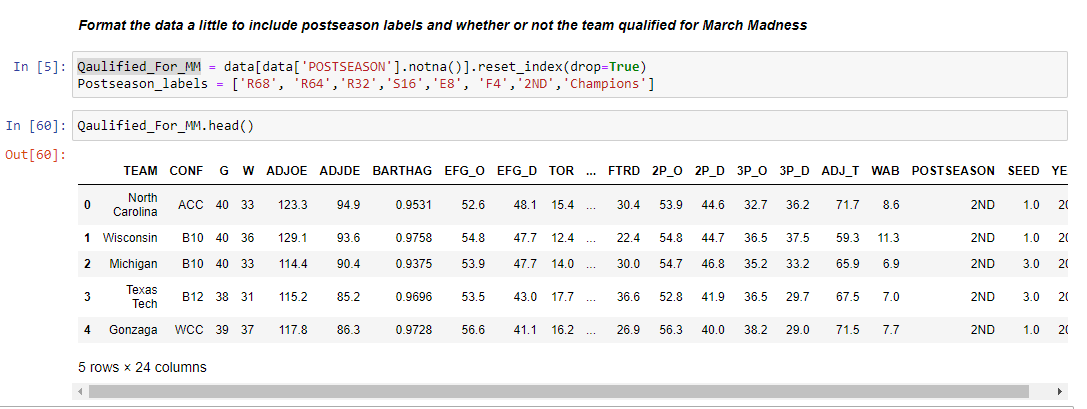
* User
* Count of Friends
* Count of Followers
* Screen name
* Profile image URL
* Tweet text
* Tweet Creation time
* Name
* Twitter profile description
* Favorites
* Likes

**Data Preparation:**

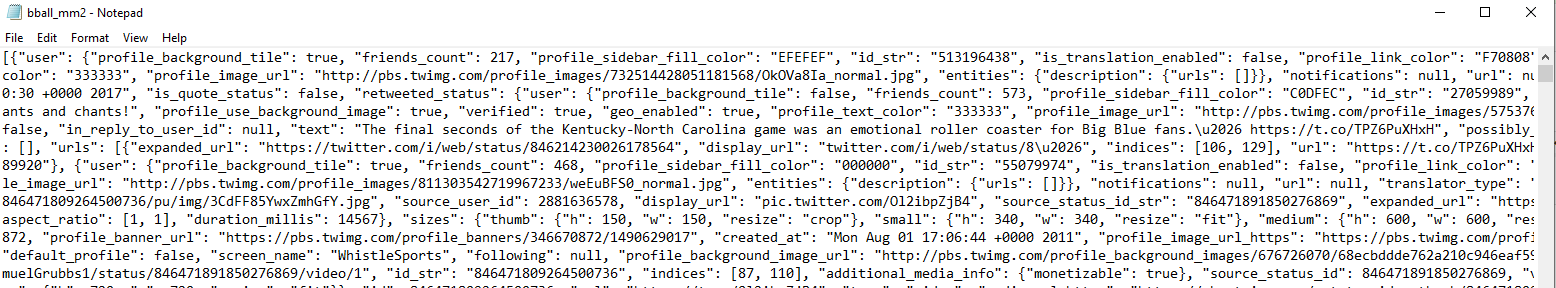
The first dataset was provided from Kaggle in a csv and did not need much data cleaning; however, some data cleaning and manipulation was performed. First, the data was read in to a pandas dataframe:



Once the data was read-in to a pandas dataframe, I needed to only include teams that made the postseason tournament, often called ‘March Madness’. The purpose behind only including teams that participated in the post season tournament was because all of my analysis was focused on the variables that led to teams participating and winning the March Madness tournament. To ensure that I was only dealing with teams that made the tournament I needed to clean the data a little to only include teams that made the tournament in a different dataframe:

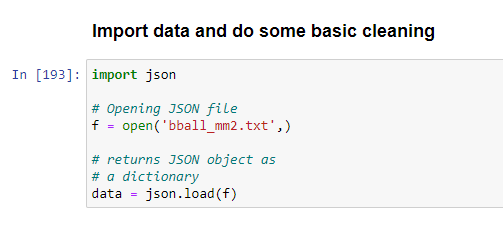


The second dataset was provided from the IST 652 asynchronous material, but aligned perfectly to my College Basketball Final Project. The data cleaning of this data was unique and somewhat complicated due to the nested nature of JSON data. Depending on what I was trying to accomplish I had to clean the data and prepare the data in different ways. To show an example of what the data looks like see the below screenshots to see the data in the native JSON format and a screen capture of the JSON data from CyberChef (<https://gchq.github.io/CyberChef/>):





I utilized the python JSON library to read in the data and created pandas dataframes as needed depending on the specific question that I was trying to answer.



**Questions:**

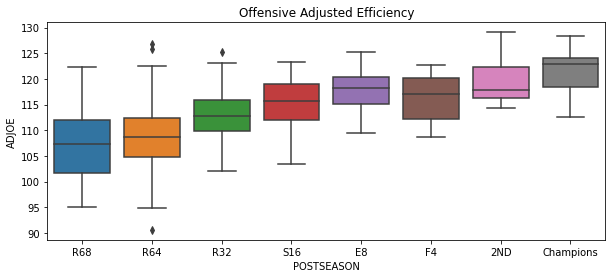
There are ten key questions that I would like to answer based on this dataset.

1. What Stat had the highest correlation with championship caliber teams? (dataset 1)
2. What teams won the most Championships from 2013 - 2019? (dataset 1)
3. How did 2-point percentage and 3-point percentages effect Championship bids? (dataset 1)
4. What was the correlation between offensive and defensive stats? (dataset 1)
5. What were the most common ‘words’/’tokens’ that were tweeted? (dataset 2)
6. What was the time that most tweets were sent? (dataset 2)
7. What username has the most friends and which username has the most followers? (dataset 2)
8. What are some of the most active usernames? (dataset 2)
9. What are some of the most active Screen names? (dataset 2)
10. What are the most popular operating systems for the users in this dataset using twitter? (dataset 2)

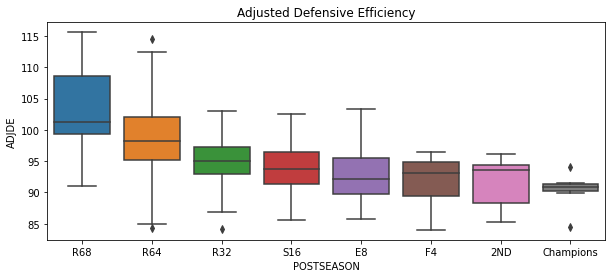
**What Stat had the highest correlation with championship caliber teams?**

There are a large amount of stats and calculated stats in this dataset and when analyzing this dataset, I wanted to know which stat at the highest correlation with Championship caliber teams. Was it Offensive Adjusted Efficiency? Was it Defensive Efficiency? Was it Field Goal Percentage? Was it the tempo of the team? Was it the Power Rating? To seek to find this answer I utilized all of the aforementioned stats and plotted them in various boxplots to see the variability of the stats and how those stats may effect a Championship run in the march Madness tournament.

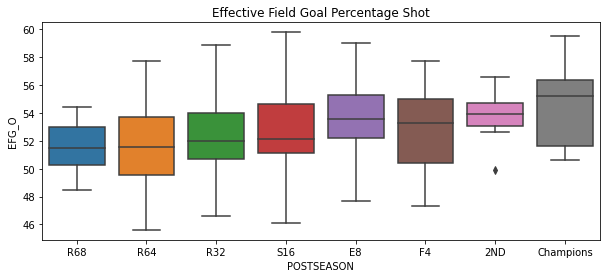
**Offensive Adjusted Efficiency**

****

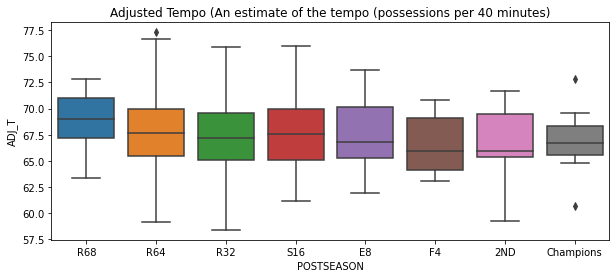
**Defensive Adjusted Efficiency**

****

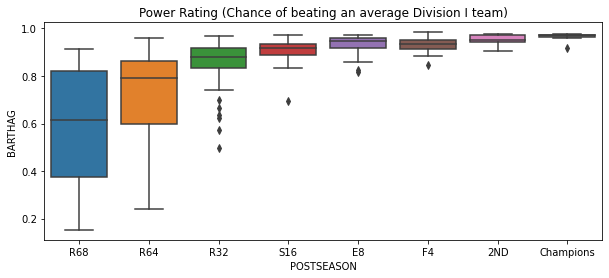
**Effective Field Goal Percentage**

****

**Adjusted Tempo**

****

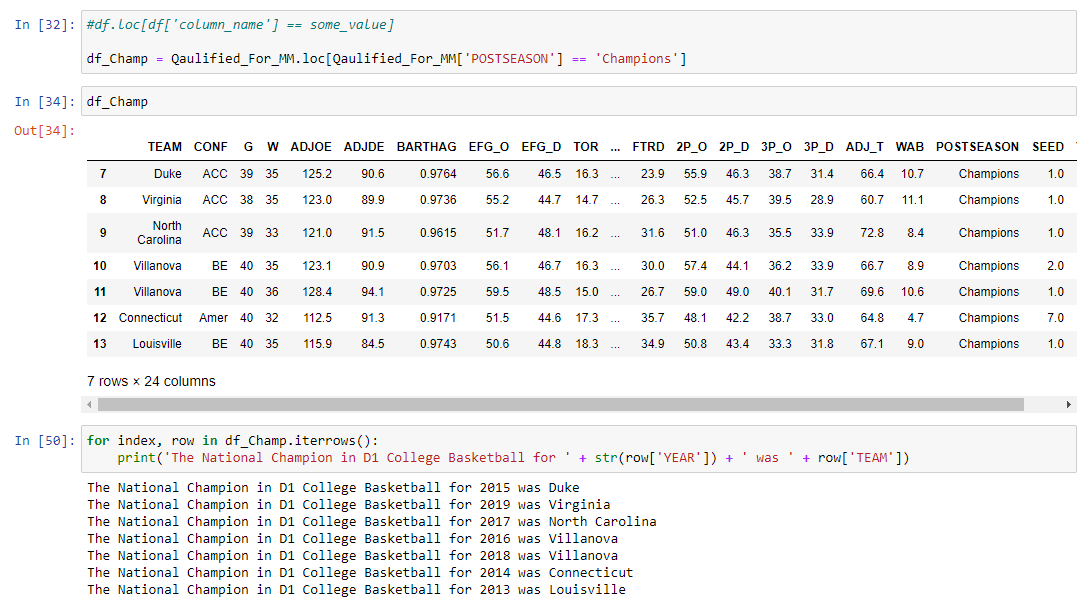
**Power Rating**

****

It looks like there are two stats that are the best determining factors on if a team is going to make a Championship run in the march Madness tournament. Those stats are the Power Rating (Chance of beating an average Division 1 Team) and Adjusted Offensive Efficiency (An estimate of the offensive efficiency (points scored per 100 possessions) a team would have against the average Division I defense).

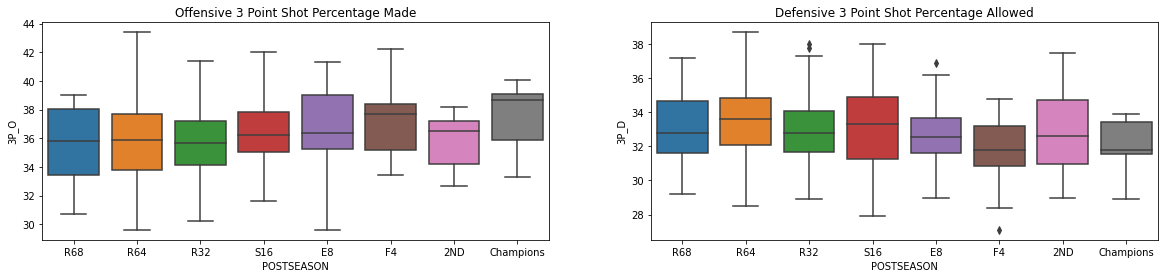
**What teams won the most Championships from 2013 - 2019?**

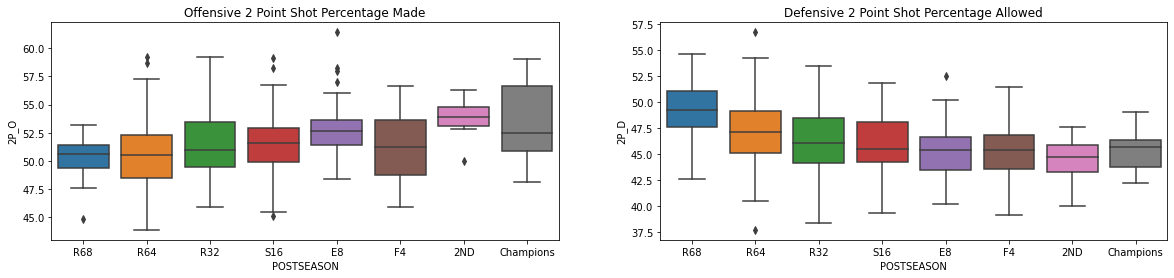
I also wanted to know which team won the most Championships during the collection of this data from 2013 – 2019. To determine this we first needed to narrow down the dataframe to only include teams that had a POSTSEASON value of ‘Championship’, and then just print the teams that won and the year that they won.



The team that won the most National Championships from 2013 – 2019 was Villanova with 2. They won the National Championship in 2016 and again in 2018.

**How did 2-point percentage and 3-point percentages effect Championship bids?**

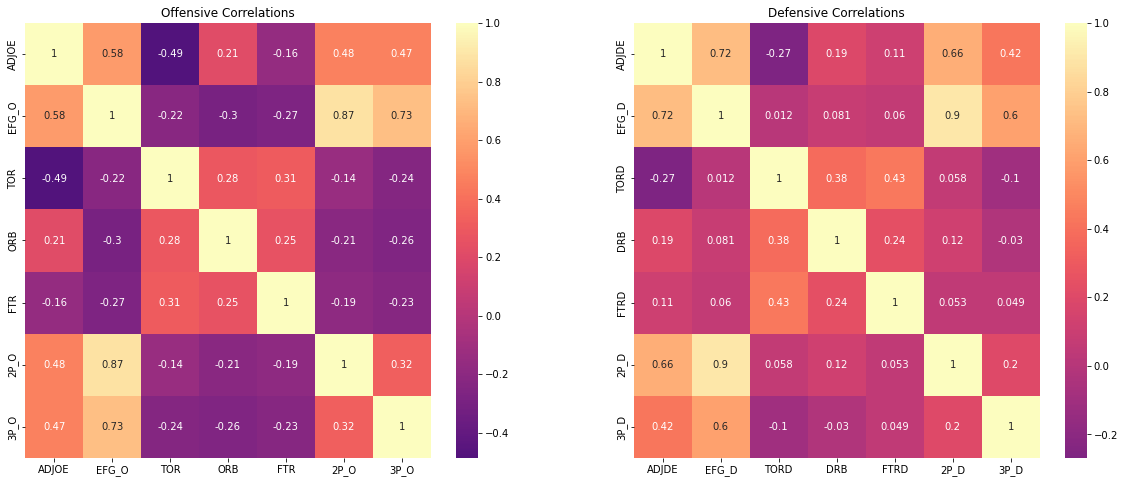
I also wanted to figure out how 2-point and 3-point percentages effected the Championship bids. Meaning did having a higher 3-point or 2-point shooting percentage effect how far a team went in the March Madness tournament. 



The conclusion for if 2-point and 3-point percentages had an effect on the Championship bid was that over the last 6 years the team with a higher 3-point percentage (around 39%) went further in the tournament and was more statistically significant than the importance of 2-point field goal percentage.

**What was the correlation between offensive and defensive stats?**

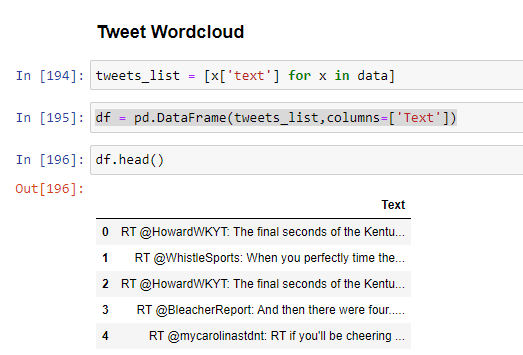
Finally, I wanted to create a visualization showing the correlations between the offensive stats and the defensive stats to determine if any of those stats were closely related using a heatmap in seaborne.



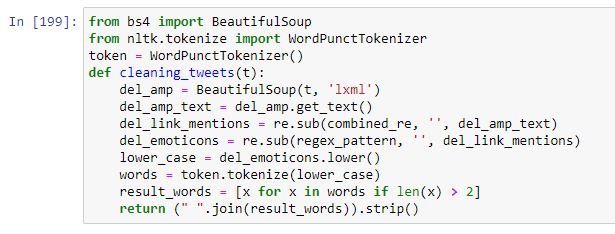
The highest correlated offensive stats were 2-point shooting percentage and effective field goal percentage, and 3-point shooting percentage and overall effective field goal percentage. The highest correlated defensive stats were 2-point shooting percentage allowed and effective field goal percentage allowed, and Adjusted Defensive Efficiency and effective field goal percentage allowed. This visualization helped understand how these different stat categories either are or are not related.

**What were the most common ‘words’/’tokens’ that were tweeted?**

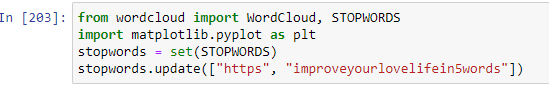
To determine the most common ‘words’ that were tweeted some unique data cleaning needed to occur. I first read the twitter text data into a list and placed that list into a pandas dataframe for processing.



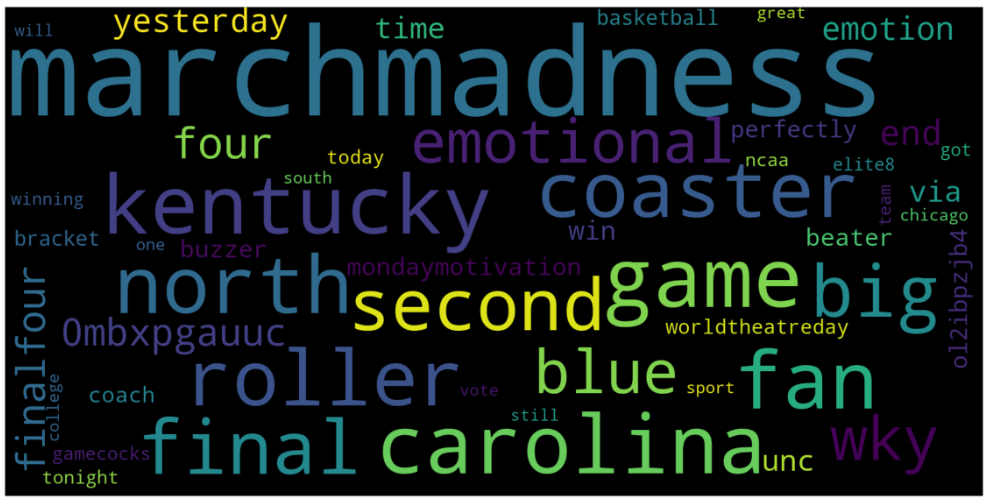
I then needed to clean some of the text data to remove hyperlinks, emojis, and mentions (@). To accomplish this data cleaning, I created a function that would clean up the data and used a package inside the nltk library called WordPunctTokenizer to “tokenize” the words into unique tokens.



I then wanted to display the data in a wordcloud and used the WordCloud package in python as well as removed some stopwords.



Finally, the wordcloud was displayed showing the most common ‘tokens’ in all of the tweets related to March Madness to help determine the most common ‘tokens’ throughout all of the tweets.



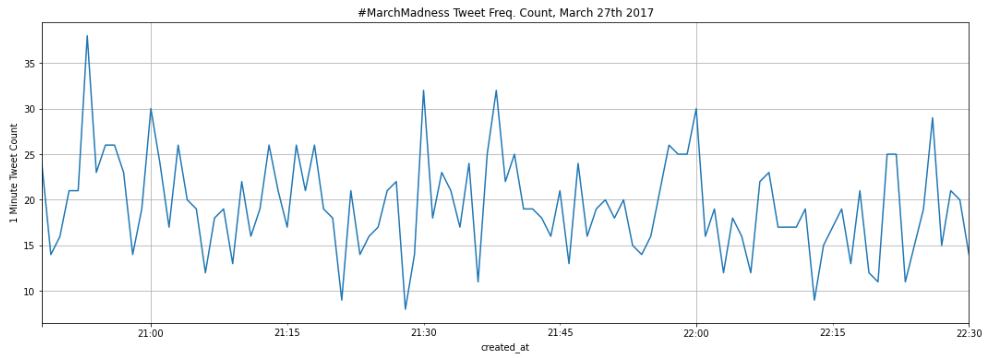
As expected, the most common token tweeted in this dataset was the token ‘marchmadness’, but some of the other most common tokens were Kentucky and Carolina which were too teams that were a part of the tournament in 2017.

**What was the time that most tweets were sent?**

The next question that I wanted to answer from this dataset was trying to determine the time that most tweets were sent from this dataset. The data was read into a pandas dataframe using the read\_json() function, and some basic grouping was completed to bin the times into different buckets.



The following graph was created to help answer the question about what time the most tweets were sent in this college basketball dataset:



All of the tweets were sent on the 27th of March of 2017 with the most tweets sent around 8:45 PM.

**What username has the most friends and which username has the most followers?**

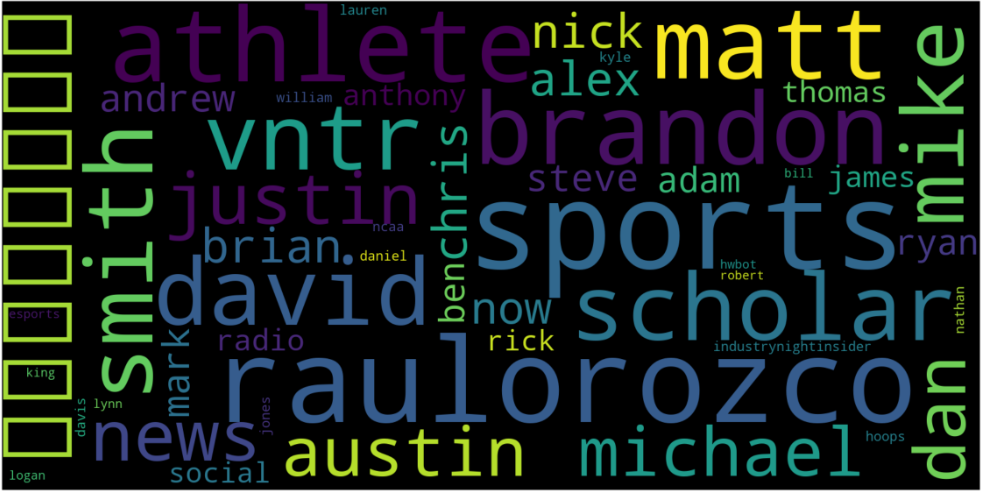
The next question that I wanted to answer was which username had the most friends and which username had the most followers on their twitter account. To answer this question, I first needed to iterate through the JSON file and create separate lists and then read those lists into a dataframe for processing. I then sorted the list by the friends and the followers count to answer the question.



The username with the most followers and the most friends was “Producer 9-0” with 831k friends and 1 million followers.

**What are some of the most active usernames?**

To answer the question about which usernames were seen the most in this dataset, I opted to use a wordcloud again to answer this question. The same data cleaning to create this wordcloud was used and explained in the creation of the previous wordcloud.



The username with the most tweets from this dataset was ‘raulorozco’.

**What are some of the most active Screen names?**

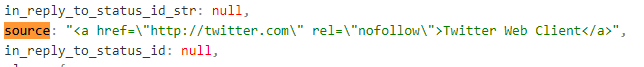
To answer the question about which Screen Names were seen the most in this dataset, I opted to use a wordcloud again to answer this question. The same data cleaning to create this wordcloud was used and explained in the creation of the previous wordcloud.



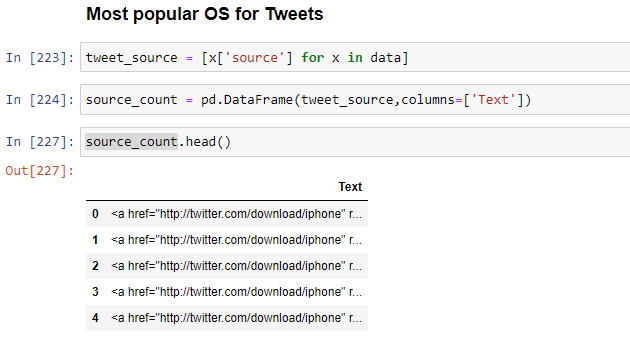
The most popular Screen Name in this dataset was also ‘raulorozco’ shortly followed by ‘scholarathlete’.

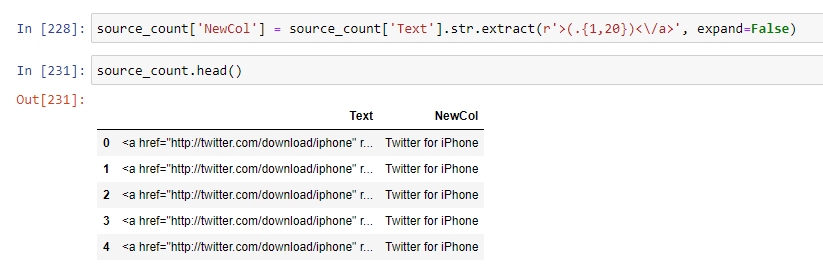
**What are the most popular operating systems for the users in this dataset using twitter?**

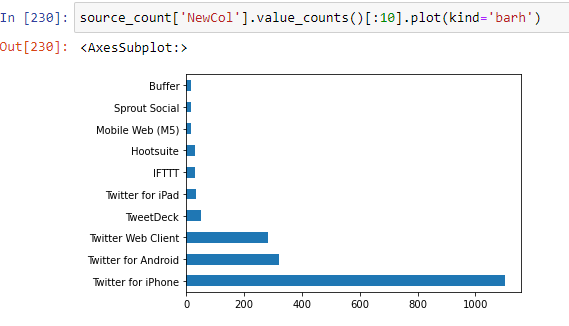
During the analysis of the data, I noticed that in the ‘source’ key you could determine which operating system the user was utilizing at the time the tweet was created.



Based on this data point, I thought it would be interesting to try to determine and graph which operating system was the most utilized. As you can see from the example above, the operating system is between ‘>’ and ‘</a>’ so with some simple regex extraction and capture groups we can create a new column in the pandas dataframe with the operating system extracted. Then we can create a simple graph of the most common operating systems using the .value\_counts() function in python.







The most popular operating system of users in this College Basketball dataset is the iPhone by far with Android being in second place.

**Conclusion**

The biggest take away from analyzing the Division 1 College Basketball Kaggle dataset is that Adjusted Offensive efficiency is the most important stat in determining which teams have the best likelihood of winning the March Madness tournament. Additionally, the team that has won the most National Championships from 2013 – 2019 is Villanova with two. Finally, 3-point shooting percentage is slightly more important than 2-point shooting percentage when it comes to predicting a team that will win the march Madness tournament. Additionally, the biggest take away from analyzing the Division 1 College Basketball Twitter dataset is that the most common token tweeted regarding College Basketball from this dataset is ‘marchmadness’. The most active time for tweets during this dataset was around 8:45 PM. The username with the most friends and followers was ‘Producer 9-0’ with 831k Friends and over a million followers. The most active username and screen name was ‘raulorozco’. Finally, the most popular operating system utilized by users in this dataset was iPhone.