Axel Broome, Jack Cook, Joseph Hnat, Bryan Kariuki

ETL Technical Report

Vanderbilt DBC

February 14th, 2020

*Social Media Influence and the NBA: Revaluating the Worst and the Best from the 2016-2017 Season*

1. *Essential Questions for Project (in Theory)*

Our group was set on answering the following questions regarding, arguably the most exciting and best, NBA season:

* Who were the least favorite players (aka Cast-aways) according to Twitter statistics from the 2016-2017 season?
  + Potentially is there a correlation between the players social media presence (Twitter) and crowd attendance at games?
* Would this team of “Cast-aways” be able to exist under the salary cap of the 2016-2017 season?
  + If not, how much over the salary cap would that team be?
* In theory, how competitive would the ‘cast-aways’ be versus the Golden State Warriors (NBA Champs 2016-2017)?

1. *Essential Questions for Project (in Practice)*

After reviewing the available data sets from our source, we found it best to answer the following questions about the same NBA season:

* Is there a correlation between who were the most popular players in the league and how well the performed for their respective teams?
* Is there a correlation between how much teams were paying players and how effective those players were on the court?

Both sets of essential questions gave us plenty to work with. Below is a more detailed report of the process we used to **E**xtract, **T**ransform and **L**oad our data to get answers to the questions we were eager to find. Enjoy!

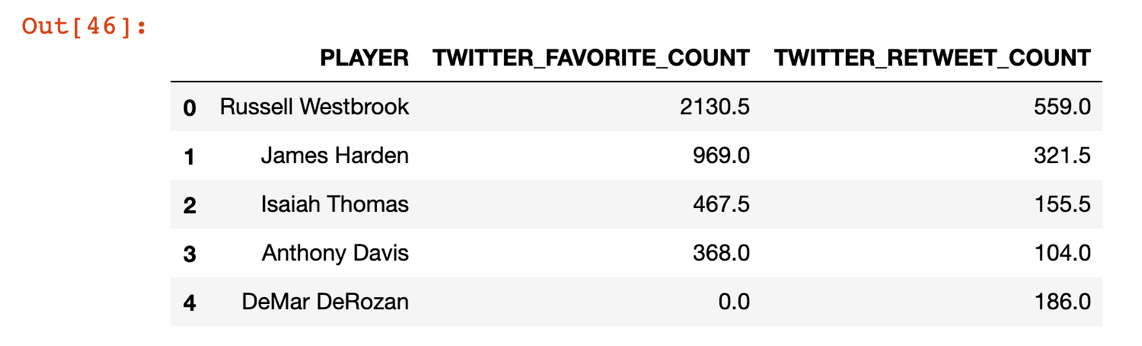
1. *Extraction*

We used 2 data sets from the public platform Kaggle. More specifically our data set came from the Social Power NBA database. All our data focused on the 2016-2017 NBA season. The two specific data sets that we used for our analysis were as followed:

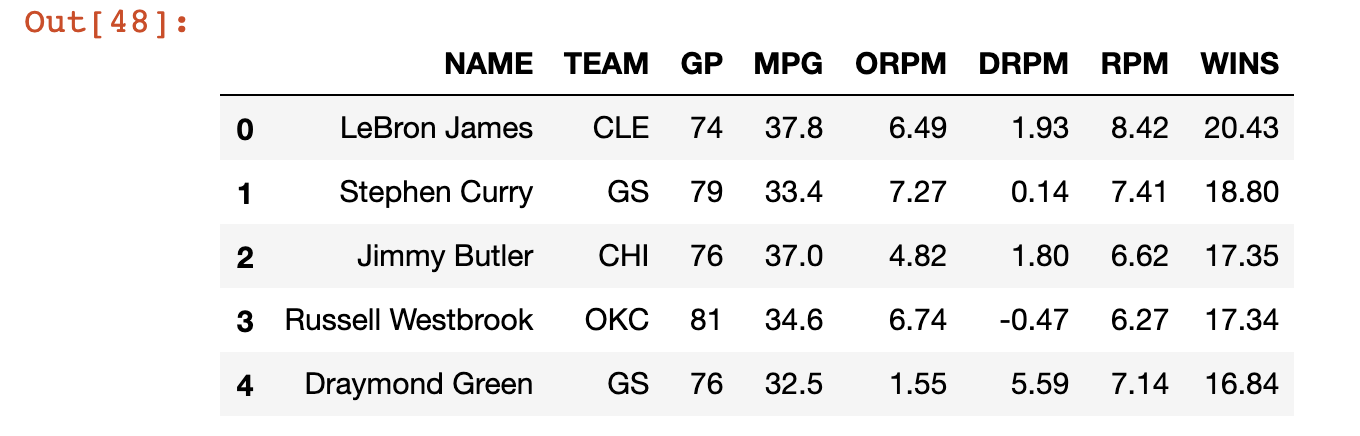
* Twitter Player Stats
  + This was a sample of 100 NBA players during that season which included information on both the number of times a specific player’s tweet was either favorited or retweeted.
* Player Real Plus Minus (RPM)
  + As mentioned in our script, RPM is a stat that describes a player's average impact in terms of net point differential per 100 offensive and defensive possessions.

1. Transformation

Our first steps in cleaning up the data sets involved loading our respective csv files using pandas and jupyter notebook and then determining which column value we would need to merge our data sets on, and if there were any changes that needed to be made prior to the merge. Below are views of our two primary data sets that we worked with prior to any merging.

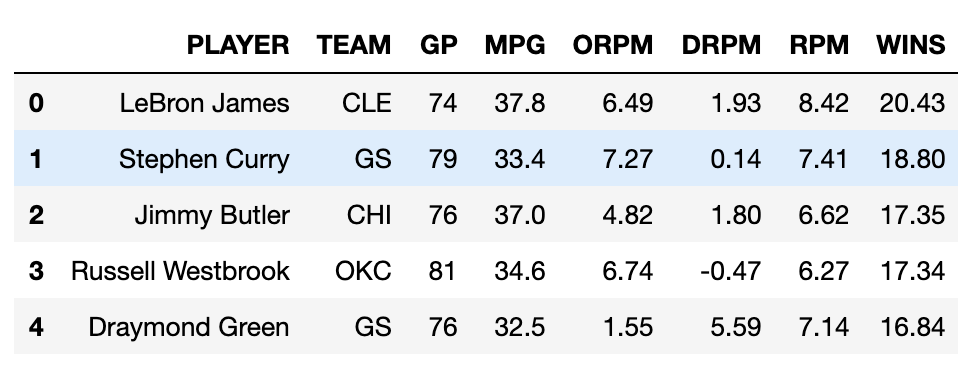


**Figure 1**: Twitter Stats data frame



**Figure 2**: Player Real Plus Minus Data

Our next step was to consider how we were going to merge these two data sets. Prior to merging these two data sets, we knew that the column ID for the desired merged-“on” data set, would have to be the same. This is why we performed a .rename(columns={}) to make sure that our second data set matched on the column PLAYER. The updated RPM data frame can be seen below:



**Figure 3**: Updated RPM data frame

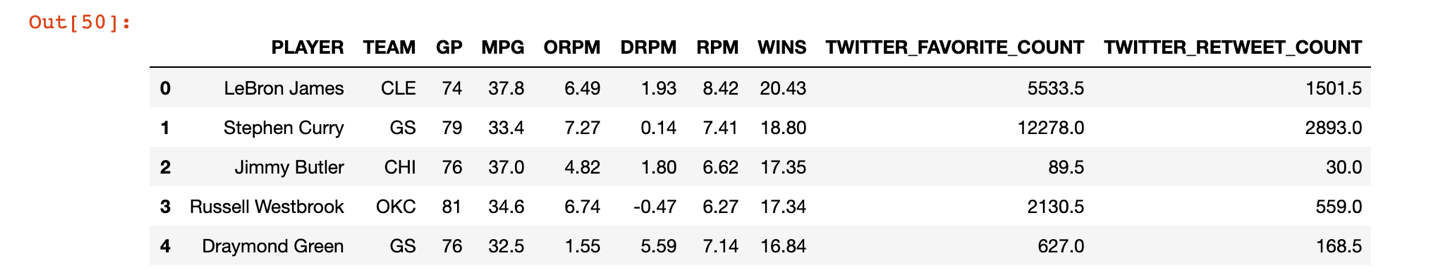
Something that is not displayed above, but was a critical part of our project, was the original data frame from the RPM stats. In that original file, the player column also included the player positions. Below we will talk about, why cleaning up the original data set in excel to remove the player position was critical to the loading process for this project.

1. *Load*

Our last phase of the phase for this project was to load/merge our data sets into on data file or frame. This is the part of project where we ran into our first limitations. To be frank our first attempt to perform a pd.merge on the figures above did not work. We continued to get the same type of gaps in our merged file. When we performed a right merge on the data set, it would return only the player and twitter statistics. When we performed a left merge on the data set, it would only return the real plus minus data and not the twitter stats.

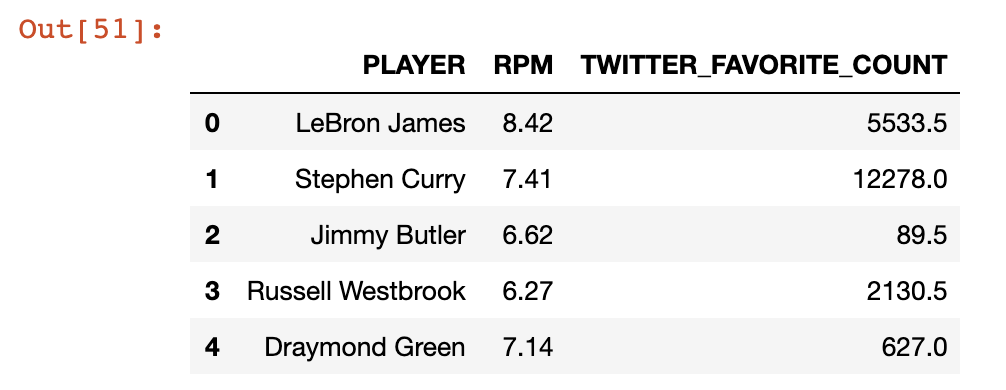
We researched how to perform merges on at least 3 files because we thought we could join our merged df with the data frame that was not showing up. That also didn’t work. After brainstorming what was the root issue with our code, we found out that there was an error on our merged-"on" data set.

In the original CSV file for our RPM file, the player names also included the players basketball position. We found out that you are unable to perform a successful merge unless both sets of data on the selected merge-"on" criteria are the same. In other words, since our twitter stats only had the player name, we needed to make sure that our RPM data set only had the player name as well. After we hard edited the RPM csv to not include the position of the player, we then reloaded that csv into the file path for our script. And wahlah! All the data in our pd.merge finally showed up! That figure can be seen below:



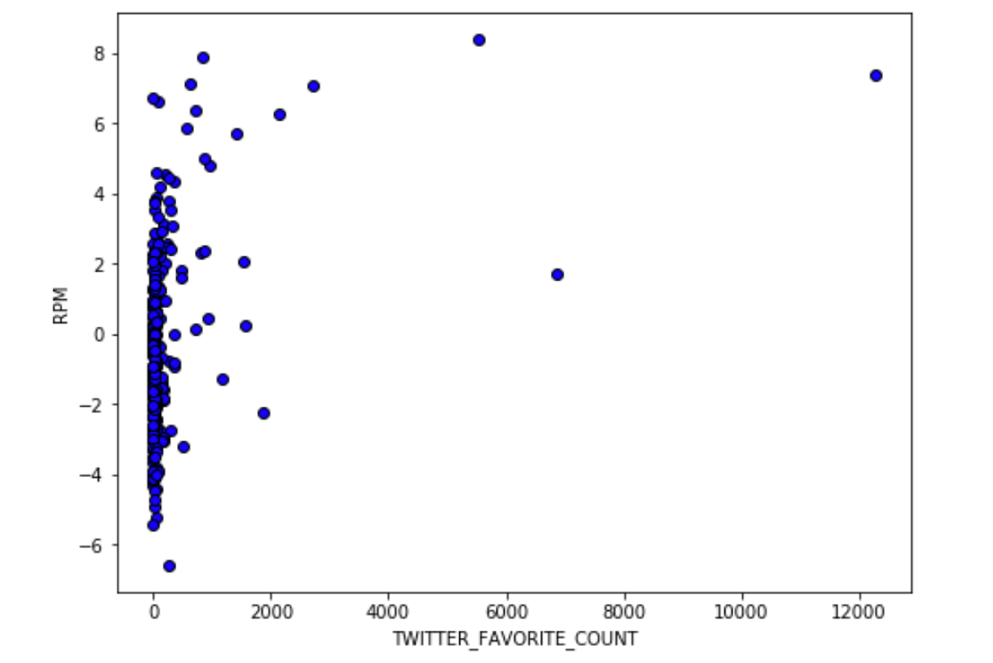
**Figure 4**: Merged Data Frame

After getting our merged data set using the two different csv files, we then cleaned up the df, by eliminating the column values we did not need for our first analysis. That figure can be found below:



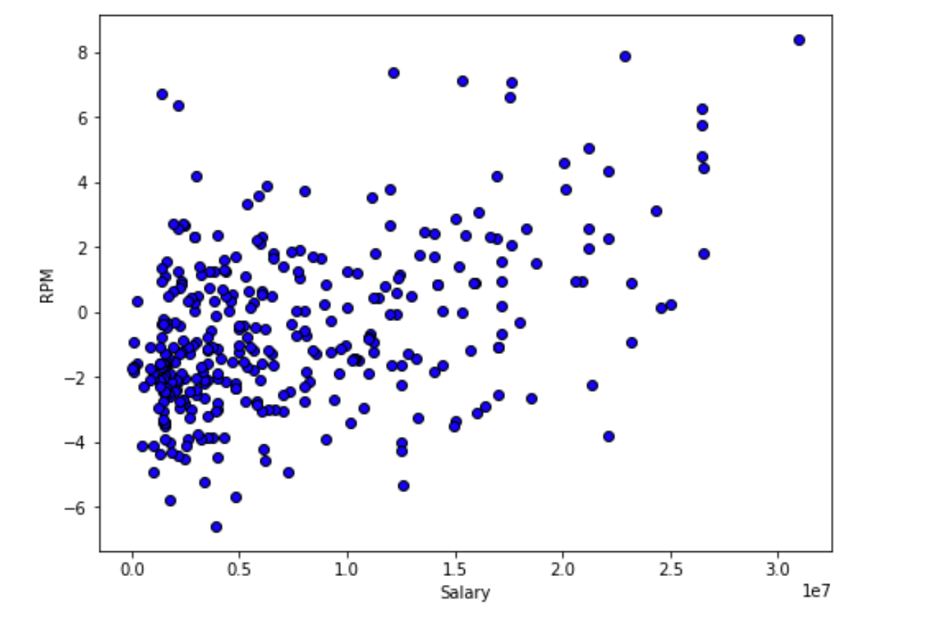
**Figure 6**: Reduced Merged Data Frame

While this was not a requirement of the project, we wanted to get a visual to help us answer our essential question to find out whether or not a correlation existed between a player's popularity and how well they performed for their respective team. We used matplotlib scatter plots to help us visualize the data. What we found was that there was no correlation between how many times a player's tweet was favorited and how they performed on the court. The figure for this analysis can be found below:



**Figure 7**: Visualization 1

For our second research question, we followed the same logical progression as the first half of this script. We followed the exact same logical progression to get our visualization. As seen below, we were able to land on there actually being a positive correlation between how much players were paid that season and their overall production on the basketball court.



**Figure 8**: Visualization 2

1. *Conclusion*

Overall this project pushed our group to have a deeper understanding of the entire ETL process for when you want to use csv files and python. Additionally, it pushed our group to think about fixes to concepts we thought we had sure fire solutions for. With more time, we would do a more in-depth analysis to determine correlations between RPM and attendance and RPM for whole team comparisons with selected players from the league.