Jinghao Yan, Ye Zhang, Brody Vogel, and Zihao Li

ANLY 501

October 5th, 2017

Project Assignment #1: Write-up

**Data Science Problem**

We will investigate the data science question: *Does the U.S. stock market react to professional sports leagues and the performance of their teams, and, if it does, how?*

American professional sports teams both spend and earn incredible amounts of money. As of 2016, the three major professional sports leagues (the NFL, MLB, and NBA) combined for over $27 billion in total revenue.[[1]](#footnote-1) For reference, that’s more than the rest of the world’s 20 largest leagues combined.[[2]](#footnote-2) From a purely monetary standpoint, then, it seems reasonable to think that such a huge industry would affect the U.S. stock market. And, from a less intuitive perspective, professional sports’ ubiquitous popularity leads us to believe the market could be affected by the performance of certain teams, too. For example, if the Cubs are accurately predicted to have over 10 million fans, does the stock market turn up when they perform well or trend down when they don’t?[[3]](#footnote-3) We want to find out.

Especially in light of the NFL’s Rams and Chargers moving to Los Angeles this year with plans of opening a new stadium in 2020, much research has been done on the impact of a city being home to professional sports teams. There is a fair bit of disagreement in the literature. On the one hand, Stanford University, *Business Insider,* and *Marketplace* have all published work that argues professional sports teams have little effect on their cities’ economies.[[4]](#footnote-4) However, the U.S. economy and stock market are far from synonymous, and so, on the other hand, the average value of America’s 50 largest professional teams is up 18% since 2016[[5]](#footnote-5); and, corroborating this, CNN recently ran a feature story highlighting the rapid growth in both popularity and revenue of the NFL and NBA.[[6]](#footnote-6) Our question, it would seem, is at the intersection of these two related discussions: If sports teams have little impact on their local economies, it would follow that they and their performance would not affect the stock market; but, since professional sports and their fan bases are growing at extreme rates, they must be responsible for a huge flow of money, which one would think *must* be reflected in the stock market. Thus, we think our work will produce results applicable to every aspect of the discussion outlined in this section.

**Potential Analyses That Can Be Conducted Using Collected Data**

To get the results we’re after, we collected data from the last 42 years in the U.S. stock market, and then found the outcome of every game played in the NFL, MLB, and NBA since 1975. For stock data, we wanted indexes or large funds that represent the market or one of its sectors, so we used the NASDAQ Composite Index, Dow Jones Industrial Average, S&P 500, Marathon Oil, Chevron Oil, Exxon Mobil, the Franklin Gold and Precious Metals Fund, Sturm and Ruger Guns, American Outdoor, Anheuser-Busch, two sin stocks (VICEX and RIDEX), and index funds for each of: utilities (Vanguard), healthcare (Vanguard), IT (Vanguard), and long-term bonds (Fidelity). For each of these stocks, we collected the daily closing value for every day available on Yahoo! Finance or every day since 1975. Also, if available, we grabbed the opening price, daily high, daily low, and end-of-day volume across that same timeframe. In all, this produced 104 distinct variables. For the sports leagues, we created variables for each team that has played a game since 1975, which turned out 120 variables. Then, indexing by date, we entered the result (‘W’ for win or ‘L’ for loss), score, and opponent for all of each team’s games. Depending on the paths we decide to follow with our analysis, the outcome data could be used to create many more variables.

Our data will help us to answer our question by giving us daily results from the stock market and each of the three major professional sports leagues. In particular, the stock indexes (NASDAQ, Dow Jones, and S&P 500) will give us a picture of the on-goings of the entire market, and the sector-specific funds will let us compare the sports data to the changes of some of the market’s largest subcategories. Since we essentially collected *every* result since 1975, we should have the resources necessary to take our question in many directions.

To give a few examples of the just-mentioned directions our project may take, we’ve discussed each of the following hypotheses:

* The stock market gets a boost when large-market teams perform well, where “performing well” could mean any of: wins one game; is on a winning streak; has an above .500 winning percentage; goes to the playoffs; wins their league’s championship; or any other “good” result.
* The stock market turns down slightly when large-market teams perform poorly, where this means the inverse of everything listed in the last bullet.
* The stock market gets a boost on days with, or the day after days with, a full slate of games or important games. In this case, the market would go up after NFL Sundays, the Super Bowl, the NBA Finals, the World Series, etc.
* The stock market gets a boost while certain sports are in-season, or in their playoffs. In this case, the stock market would be bullish during late October, when each of the three major sports is in an important part of its season.
* The stock market dips when small-market teams perform better relative to large-market teams. So, for instance, when the Tampa Bay Buccaneers played the Oakland Raiders in the 2003 Super Bowl, we’d expect the stock market to show a little dip during that NFL season, at least with respect to the way it typically performs during that time.

These, of course, are just a few vague examples, as we expect to come up with more pointed hypotheses once we start working extensively with the data.

**Data Issues**

For the most part, our data came through quite clean. There are some issues, though. From the stock data, there are a lot of missing entries. Some of these are from instances when the New York Stock Exchange was shut down, while others reflect data (usually the volume) that wasn’t reportable. Finally, the volume of some of the sector-specific funds have extreme day-to-day variances, which will make using these columns as variables almost impossible.

The sports data present some unique challenges, too. At least one name (the St. Louis Cardinals) has concurrently referred to a team in more than one professional league, which causes problems regarding variable names and references. Furthermore, it’s possible for NFL games to end in a tie, and the MLB and NBA have both gone on strike during periods of the dataset. Finally, the biggest challenge with the sports data will be devising a way to handle teams that have changed names, location, or both, often more than once. For instance, the Charlotte Hornets went from the Charlotte Hornets to the New Orleans Hornets, then to the Charlotte Bobcats, and finally back to the Charlotte Hornets, all in the span of a decade. Because of the nature of the data, it would seem that the sports results require a longer cleaning process.

We also have a lot of data, and so some issues won’t be apparent during preliminary scans. To combat this, we’ll run tester cleaners to make sure every value fits the desired format.

**Collecting New Data**

We collected our raw data using two scripts - one for scraping Yahoo! Finance and the other for collecting the sports data. For the finance scraper, we used Selenium’s Webdriver to get the base url for the desired stock and scroll through every date on that specified page. Once the driver gets each page, the script uses Python’s BeautifulSoup module to find the HTML table, rows, and entries in which the data is held, sends those to a growing dictionary, and then finally uses the csv module to write the data to a .csv file. This entire process is encased in a for-loop that iterates through a list of the urls for each stock, producing a new csv file for each stock. It has along runtime (about 45 minutes).

The raw sports data was grabbed using the urllib.request and re modules. For the NFL, MLB, and NBA, there is a portion of the script that loops through a reference site’s data from every season since 1975, collecting the results of each game. Once urlopen gets the page, the script parses through the page’s HTML source code with regular expressions to find the “winner,” “loser,” “date,” and “score,” and then adds them to a pipe-delimited text file. The process produces such a .txt file for each of the leagues. The rest of the script then conglomerates each of the three files into a nicely-formatted dataframe, which is then output to a .csv file. It has a very long runtime (almost three hours).

**Data Cleaning**

A surprisingly large portion of our data cleaning process was re-formatting. This was true, to a lesser extent, for the stock data, and especially for the sports data. To conglomerate the different-sized results from each stock, our script instantiates a giant pandas dataframe, and then passes it a standardized set of columns and dates. The script then fills in this dataframe with the appropriate values from each stock and writes it to a new, final stock csv.

Formatting the sports data was more difficult, because there wasn’t a uniform set of dates for the games. To combat this, our script again creates a large dataframe, and this time passes it every date since 1975 as an index, and every sports team in our study as a separate column. Our script then goes through each of the three .txt files containing data from the professional leagues and uses pandas dataframe methods to fill in cells based on team and date. Again, then, the script writes this dataframe to a final sports csv.

After the re-formatting, the “real” cleaning began. For the stock data, we had to deal with the missing values. To do this, we first found Yahoo!’s default entry for missing values, which is “-”. We then developed our cleaner script in such a way that it runs through every value and, if that value is “-”, changes it to a “nan”. In all, there were 46, 347 such values, most of which came from stocks that never report their daily volume. These stocks without volume, then, were noted, and we decided that if a stock never reports its volume, it’s not useful to have its volume as a variable. So, we dropped all the columns of empty volumes. After this, we checked to make sure every value that should be a number is actually a number, which they were. Thus, from our stocks data, we cleaned 46, 347 missing values and dropped 6 useless variables.

As was alluded to, the sports data was a bit trickier to clean. To start, we decided the data from when the leagues were on strike should be left, since it *is* accurate, and might produce some interesting results once we start analyzing things. For our first act of cleaning, then, we wrote a patch in the scraper to handle the one team name that has concurrently referred to teams in multiple leagues: the St. Louis Cardinals. To do this, we changed all occurrences of the St. Louis Cardinals from the NFL to “St. Louis Cardinals (Football)”. Next, to handle the NFL’s ties, we checked all the game results to see if the winner and loser had the same score. If they did, we changed the outcome from “Winner” and “Loser” to two “Tie”s. After that, we handled the teams that have moved locations and/or changed names. To do this, we created lists of every team that has relocated/changed names and its new name and location. We then used this list to merge the columns of the old teams with the new teams’. This took our dataset from 120 variables down to 95. We made the decision to keep the names of the old teams in the specific cells (eg. ‘W/27/**Baltimore Colts**/17’), though, because we feel that is a more accurate way to store the results. So, for instance, in the column entitled “Arizona Cardinals,” there is data for every game played by the “St. Louis Cardinals”, “Phoenix Cardinals”, and “Arizona Cardinals”. We then ran the new data through the following tests:

* The winner never has fewer points than the loser
* There should never be more than the outcome (W, L, or Tie), opponent, and score in a cell
* Every score should be a number
* Every outcome should be W, L, or Tie
* The opponent should be a string, and should be more than two letters
* There should always be a winner, loser, or two tying teams

As it turns out, none of our entries had any of the bulleted problems.

As a final task, we added four new variables to the sports dataset. These variables count, for every day, how many NBA, MLB, NFL, and total games were played. We’ve commented it out in our cleaning file because it greatly increases the runtime (from a few minutes to eight or nine minutes). This brought us back up to 99 variables, all of which are included in the cleaned dataset.

Below is a summary of all non-formatting cleaning:

|  |  |  |
| --- | --- | --- |
| **Data Problem** | **Cleaning Method** | **Cleaning Result** |
| Stocks Missing Values (‘-’) | Replace with ‘nan’s | Missing Values = nan |
| Empty Stock Volume Columns | Drop the Columns | 6 Fewer Variables |
| Check for Appropriate Stock Values | Various Loops | No Problem Data |
| 2 St. Louis Cardinals Teams | Patch the Scraper | 1 St. Louis Cardinals and  1 St. Louis Cardinals (Football) |
| Teams that Have Moved | Merge Old and New Teams | 25 Fewer Variables |
| Check for Appropriate Sports Results Values | Various Loops | No Problem Data |

Finally, to measure the data cleaning quality before and after cleaning, we simply used the following formula:

*Number of valid (non-missing and correct) values / Number of total values*

Before cleaning, the quality of our sports data was: **.99997**

After cleaning, the quality of our sports data is: **1**

Since we built the dataframe exactly the way we wanted the data formatted, it’s not surprising that our quality scores came out so high. As the above sections, referenced, much of the cleaning came during the data collection phase.

Before cleaning, the quality of our stocks data was: **.9587**

After cleaning, the quality of our stocks data is: **.9972**

Our not having a second perfect quality score is due to the sporadic missing values scraped from Yahoo!.

1. “Which Professional Sports Leagues Make the Most Money?” howmuch.net. https://howmuch.net/articles/sports-leagues-by-revenue (accessed on 10/3/2017). [↑](#footnote-ref-1)
2. ibid. [↑](#footnote-ref-2)
3. Pias, Ashakul. “How Many Fans does Each MLB Team Have?” baseballot.blogspot.com. http://baseballot.blogspot.com/2014/07/how-many-fans-does-each-mlb-team-have.html (accessed on 10/3/2017). [↑](#footnote-ref-3)
4. Reference to:

   Bergman, Ben. “Are Pro Sports Teams Economic Winners for Cities?” marketplace.org. https://www.marketplace.org/2015/03/19/business/are-pro-sports-teams-economic-winners-cities (accessed on 10/4/2017).

   Zimbalist, Andrew. “Sports, Jobs, and Taxes: Are New Stadiums Worth the Cost?” brookings.edu. https://www.brookings.edu/articles/sports-jobs-taxes-are-new-stadiums-worth-the-cost/ (accessed on 10/4/2017).

   Parker, Clifton. “Sports Stadiums Do Not Generate Significant Local Economic Growth.” news.stanford.edu. http://news.stanford.edu/2015/07/30/stadium-economics-noll-073015/ (accessed on 10/4/2017). [↑](#footnote-ref-4)
5. Badenhausen, Kurt. “The World’s 50 Most Valuable Sports Team 2017.” forbes.com. https://www.forbes.com/sites/kurtbadenhausen/2017/07/12/full-list-the-worlds-50-most-valuable-sports-teams-2017/#6f7c9b564a05 (accessed on 10/4/2017). [↑](#footnote-ref-5)
6. Edwards, Pierce. “NFL vs. NBA: The Battle for World Supremacy.” cnn.com. http://www.cnn.com/2016/10/20/sport/nfl-nba-twickenham-london-china/index.html (accessed on 10/4/2017). [↑](#footnote-ref-6)