Joe Tucci

590 OM/OMO

Final Project Proposal

9/24/19

There’s so much data that comes from baseball that transforming that data to determine, for example, the worth of a center fielder can get even more specific with advanced stats like sabermetrics. Teams are often looking at these advanced stats when considering salaries and trades. But what if teams put as much consideration into the quality of their hot dogs at games as they do statistics? The question I want to try to answer revolves around the external factors that contribute to the game of baseball. Those external factors include the food at stadiums, average fan attendance in a given year, weather data from each city the team plays in, and the success of a team’s season and player’s success. All of these different datasets will inform one another to give some more insight into the question of what kinds of external factors might have an effect on the game itself.

When it comes to food at a baseball stadium, I want to know if the quality of the food has any effect on the fans’ overall experience at the stadium. And the first aspect of that comes in their attendance at games. Is there any kind of correlation between the quality of food at a stadium and the attendance each year? The quality of food will be determined from a few different factors. First, I want to look at health code reports from the stadium. This metric is more of a departmental grade of how well a stadium does when it comes to food prep and cleanliness when serving food. Then I have found user data from reviews of Yelp and Google reviews of stadiums, and this data compiled mentions certain kinds of food and facilities.

I’m hoping that the bulk of my data will be centered on this food health code reports and reviews data. Ultimately it would be cool to have some kind of metric that could say “X team’s hot dogs are good, so therefore this explains their moderate to good record.” But to ground my project a little bit more so that it isn’t totally goofy to compare hot dogs to batting averages, I will also incorporate weather data. This will contribute to a few external factors to baseball, as well as show some more direct correlations as well. While the weather will provide insight into the success of a team, it will also be compared to food health reports and quality of food. If the weather is particularly cold or rainy that day, does that have an effect on food prep?

The result of combining all this data will involve basic baseball team statistics like their win-loss record, team batting average, and team ERA, combined with the external factors mentioned previously like weather and food. This data will not provide or create any kind of aggregate score or rating of teams or teams’ stadiums but rather provide insight into different kinds of data for each team that will have a different conclusion for each team and stadium.

Data Sources

* 1. ESPN Health Inspection Data
  2. Paula Lavigne. ESPN.
  3. <http://www.espn.com/espn/feature/story/_/id/25316231/health-inspection-reports-find-critical-violations-nfl-nhl-nba-mlb-stadiums-2018-espn-lines#!>
  4. License information: Since ESPN is now owned by the Disney corporation, the terms of use for ESPN-related content come directly through Disney. I was able to find a [terms of use page](https://disneytermsofuse.com/english/#section6-claims) that says that their services, which include content from ESPN, “are for your personal, noncommercial use and are intended for informational and entertainment purposes only.”
  5. This is a crucial part of my dataset in that it will provide substantial data about each team stadium and their health code reports. Depending on whether I want to stick with just baseball, or focus on a specific city’s stadiums, this source provides data for the four major American sports that will help with finding health code information. You can do a search by city or filter by sport. At its most basic level it provides a rate at which each stadium commits a health code violation with a percentage number.   
       
     This data can provide me with a wide variety of data. I know that if I stick to just baseball, I have options when it comes to either just taking the overall “score” percentage of violation rates, or more data from each entry for each stadium.
  6. The biggest weakness that this source has for me is that the webpage is very design-heavy in that you can customize what data you want to see. And scraping some of that data may be difficult. This data also synthesizes and gives a sort of average or composite “score” ranking that is already mashing together some data. So, if I were able to find those original datasets that would be very helpful.
  7. This source, unfortunately, is at about a moderate to hard level right now. While it does provide very basic information like the percentage of violations committed within each stadium, it’s obvious that there is more data that makes up this entire article. Where it’s at now it would involve web scraping the website for information. I’m hoping that the actual dataset can be found, and that may even be elsewhere that’s not on this ESPN article.  
       
     If I cannot find the actual dataset somewhere on the backend or through some additional research, which is the most probable outcome, it will require some tricky web scraping. I’m uncertain right now as to the structure of the webpage, and combing through that and trying to understand how all of the information is structured will be the biggest hurdle. The issue then isn’t necessarily the web scraping, but more so how I will even get to the part where I will want/need to web scrape.  
       
     Another issue that relates to web scraping is working with the ESPN API, which according to [this web page](http://www.espn.com/apis/devcenter/blog/read/publicretirement.html), says that they do not offer an API at all.
  8. Baseball stadium food safety violations
  9. Corey Nachman. Business Insider.
  10. <https://www.businessinsider.com/mlb-stadium-food-safety-violations-2011-3>
  11. License information: The terms of service page for Business Insider states that their content may be used for personal and informational use, and I cannot reproduce anything from their site. That can be found [here](https://www.insider-inc.com/terms/).
  12. This is a good source that points me in the right direction for finding the specific city’s department of health website. The actual data from this website just simply states the number of violations per stadium, but not every stadium has a number attached to that subset. Others merely just write out what some code violations were. This source is more of a good reference point for looking at health department websites for if/when I need to narrow down my project.
  13. There is not a lot of consistent data from the actual web page that will make it easy to gather effectively. Some of the actual data that is on this webpage appears to be cherry-picked almost at random, which would make it even more difficult to gather.
  14. Right now this is at a hard to impossible source, primarily because of the lack of data contained on the actual webpage. Some web scraping could be used to gather the listed number of violations for each stadium, but that does not provide a whole lot in the way of anything else for more data. The impossible aspect comes from the fact that each stadium includes a link to the health department website for the corresponding city. This will require a lot of gathering.  
        
      This source is one that is only useful if I want to really go into each city’s health code website to dig even deeper to find health code reports for the stadiums. This will require a lot of hunting and gathering on my part. The upside of that, though, is that typically those websites won’t require a lot of web scraping and will just provide some sort of CSV or excel file. It would just add up to a lot of datasets and require a lot of organization on my part.
  15. Baseball-reference.com
  16. N/A
  17. [www.baseball-reference.com](http://www.baseball-reference.com)
  18. License information: The site is very friendly to information gathering and states that it is for informational use only. It even includes a section about allowing for lots of gathering and printing for teachers. Here’s their [terms of service](https://www.sports-reference.com/termsofuse.html) page.
  19. This source is basically a one-stop shop for anything baseball statistics. I can easily look up by a specific team or by year. The website offers lots of options in the way of being able to filter out the specific data that I need for any given query. If I want to find each team’s win/loss record and team batting average for a specific year, I can get a list of every team and those statistics. Or if I want to go more specific and get more in depth with an individual team I can look at statistics over a set of years (or their entire existence) as well as players on a given team. With this source I can also find basic attendance data for each time in any given year, but it does not appear at this time that I can find attendance for each individual game.
  20. The only real weakness with any of these datasets is that there might be too much information for whatever kind of dataset I’m able to download. And with that overabundance of information it can be difficult to pair down in a granular manner, like finding attendance for specific days as opposed to each year.
  21. This source is primarily at an easy source, because most of the time it allows for the option to download a CSV file of the data that you have looked up. No matter what I throw together in a search query for any given purpose, within that table there are options to download the dataset. The only real issue will be getting the proper dataset or datasets that I want and that will fit my project well.  
        
      This source, while easy in that I can almost always download a CSV file, will be quite time consuming. It will be a lot of planning and organization to keep all of these separate files in place. Then when it comes time to start using these data sets it will require even more work and time to combine them in a thoughtful way that fits into the scope of my project.
  22. 2017 Fan Reviews collection
  23. Unknown? Review Trackers.
  24. <https://www.reviewtrackers.com/reports/best-baseball-stadiums/>
  25. License information: Much of the content this website deals with comes from other review websites, and they address this in their [terms of use](https://www.reviewtrackers.com/terms-service/) page.
  26. This dataset is a collection of internet reviews of baseball stadiums based on a few factors; facility, family appeal, fan experience, food and drink, and an overall score. The basic dataset that it provides at the beginning of the article is an aggregate of a number of different questions. It then asks even more questions like best hot dogs, best food and beer, favorite kinds of food, etc.   
      This is a very long list of data about all kinds of stadium things for baseball stadiums that are all based on reviews. My particular interest in this dataset was its focus on keywords like “hot dogs” and “seats” that can potentially allow for a closer focus within my project.
  27. Some of the weaknesses from this dataset may come from its scope. It only looks at very specific keywords within reviews, so that may pose an issue for me and the scope of my project later on in the process. Also the ways that it quantifies things like family appeal or sentiment are unclear and will be data I don’t want.
  28. The difficulty of this source is somewhere between a moderate to impossible. The moderate parts come from the literal dataset that seems very web scrapable because of its easy layout. There’s a strong chance that I do wind up just using that data and nothing else. But if I want to use the other data from the graphs, that’s where the challenge comes in. As it appears now, they look like only static images, so grabbing the data from those specific images could prove very difficult for the purposes of this project.   
        
      While this may be a source that I don’t get the main chunk of data I really want from it, there is still enough on the webpage to work with for me. There is some data that I can web scrape pretty easily that would still make this source worthwhile. The ways it quantifies that data, though, may require some extra thought and analyses when I mash it together with my other datasets.
  29. Baseball Stadium statistics
  30. Andrew Clem?
  31. <http://www.andrewclem.com/Baseball/Stadium_statistics.html>
  32. License information: The [terms of use](http://www.andrewclem.com/Terms.html) for this web page is a lot more informal, stating that the site is “based on the guiding principles of fair play, truthfulness, good taste, public spiritedness, common sense, reasonableness, and a passion for learning and the dissemination of knowledge.”
  33. This source includes a lot of specific data about baseball stadiums that includes things like the dimensions of the field of play and foul territory, the seating capacity, how long the stadium has been operating, etc. This source will be one that has at least one good thing to grab from it potentially if I decide to go in a different route with what I want to look at pertaining to baseball stadiums. There’s also the possibility to look more in depth at each baseball stadium individually if I want to get all of that data or if I decide to go another route with looking at certain cities.
  34. There aren’t too many weaknesses with this data that aren’t directly related to what Baseball-reference.com offers, which is that it can potentially be too much information. There isn’t a way to download a CSV file from this, so parsing through too much information will just be it a bit more tedious.
  35. This source will be a moderate difficulty because all information from any page you click on with this website is displayed pretty neatly in a table format. So, with that table format from the website it will require some web scraping to gather that data, but given how well structured that data is it should not be too difficult to web scrape.   
        
      Some of the data from this source is overkill in that I can get it from other sources I’ve already mentioned. But this will give me additional data that can be good supplementary data if I decide I need more for my project and ultimate end product.
  36. NOAA US major city weather data
  37. NOAA
  38. <https://www.ncdc.noaa.gov/cag/city>
  39. License information: I surprisingly could not find much legal information from the NCDC website, but I have to imagine with it being a .gov website it wouldn’t be terribly difficult to work with.
  40. This source provides some versatility with how I want to get weather data as it pertains to major US cities. I can specify the data within a certain date range or yearly range and even look at certain elements like precipitation or weather. Much of the data is provided with mean averages whether that’s yearly or sometimes even monthly, so my potential take on daily weather data may not work for this source. That project pipeline in general may prove to be way more minute than I was even anticipating in the first place, so this source is great for most if not all things weather related in the United States.
  41. The weaknesses with the data all depend on how each city has structured their data, and it will depend on what kind of weather and precipitation data I want. Gathering it all will be the most difficult aspect of it.
  42. This source will most likely be an easy difficult, because once I give the correct parameters for what I’m looking for with the data it gives me a variety of download options. You can download the .xml file in a nice an organized format, and I remember working with xml in 452 so I feel comfortable that I can work with that. There are sometimes even options to just download excel or CSV files from the website as well, but it does not appear that may always be the case. Regardless, this source should not require much web scraping if at all.  
        
      Finding all of the weather data I want will most certainly be the most time consuming and difficult datasets to gather for this project. It will require a lot of individual searching for cities and then gathering the right dataset. Then, much like with team data from baseball-reference, I will need to organize the data effectively so that I can utilize it during the data mashup part of my project when all of the data will start to come together. Another worry is being able to find consistent data points between each dataset so that there is consistency throughout for each city.
  43. National Weather Services Climate Services Web Page
  44. National Weather Service
  45. <https://w2.weather.gov/climate/>
  46. License information: Their [disclaimer page](https://www.nws.noaa.gov/disclaimer.php) says that all of their information is in the public domain and free for me to use.
  47. This is more of a backup or supplementary source for weather that I found when doing research for weather data. It was often a little more difficult to find weather data about major US cities for specific years or months that I wanted, and it often took a few clicks or search parameters to really get there. This is one of those sources that requires a few more clicks than I would like, but from what I have seen and searched through it looks like I will be able to get some kind of data that I would be looking for as it pertains to temperature averages per year for major US cities where baseball stadiums are located. This site is not as easy to use, and can often be confusing on when I can even get the data as opposed to when they just provide a nice fancy graphic to go along with the data that is not provided.
  48. The weaknesses for this source won’t really be made clear to me until I’ve gathered all of them together, but that itself becomes a part of the weakness. Much like with my potential other source for weather data, being able to find all of these datasets and find consistency among them will be a challenge.
  49. This source will probably be moderate difficulty for gathering data. Most of the time the data can be displayed in some format within the actual webpage itself, so it would just be a matter of gathering it correctly and finding the right parameters and then web scraping the web page. The more difficult aspect would be in web scraping multiple pages, like when they list each city alphabetically and there are about 10 pages worth of US cities A-Z. I did see some options for xml files as well, but once again it may not always be the case for the kind of data I’m looking for.  
        
      Since this is so similar to my previous weather source, much of the same principles apply here as well. It will require a lot of individual searching for cities and then gathering the right dataset, and I will need to organize the data effectively so that I can utilize it effectively. Another worry is being able to find consistent data points between each dataset so that there is consistency throughout for each city.

Technologies:

I will mainly code in Python using PyCharm as that’s still what I’m most familiar with. I would consider myself a novice at best, and since PyCharm is primarily used as a learning tool in this regard, I will stick with it. So long as I stay organized, it should function well for my project.

I am going to try to utilize Jupyter notebooks to the best of my ability. This will be supplementary to, if not completely replace PyCharm. My hope is that I can break away from PyCharm with something like this that will allow me to transfer work from my laptop to my desktop much easier.

Web scraping is something I have already done for my 452 final project. I felt pretty comfortable web scraping by downloading an xml file and then parsing through that locally, instead of hitting the web page every time I ran new code. I had tried BeautifulSoup in the past but could not quite get it to work the way I wanted to.

I’ll use GitHub again much like I did with my final project, only this time I hope to do it throughout my project and not at the very end. This will develop my project management skills and make me document my process better along the way.

I will probably also need to use SQL for combining and mashing data once I’ve got it all. This was an activity I remember doing in 452 and not having too much trouble with, but this project should develop those skills further and make me more comfortable with it.

What I like about my project:

I like this project that I chose both because I love baseball and because I’ve already done some basic web scraping in the past. This allows me to dig deeper into web scraping and I hope to gain more confidence to do it with more personal projects in the future. The subject matter is what will keep me most engaged and interested in my project throughout the semester.

What will my biggest challenges be?

One of my challenges will be in collecting certain datasets from my sources. The scope of my project right now is a little big for what I want, and I worry that I won’t be able to do it and will have to scale it back to look at data from one year or from a more centralized location rather than an entire sport. I will probably want to seek help from either the TA or Elizabeth directly to properly assess what I’m capable of at that time.

Another challenge is just in basic Python and web scraping. I’ve done it before and know that I can adapt my approach from last time to this project, but there are always difficulties and with a lot more data, I’m anticipating that. I often turn to online forums or YouTube videos for specifics on how to do certain kinds of code, but I also can get access to the 452 moodle page from last semester and review techniques and class recordings from last spring.

One challenge I’m especially worried about is finding proper weather data and health code data. I have my sources that can point me towards that data for both topics, but for 30 stadiums and 30 different cities it will require a lot of time and searching. Mainly though, searching through those sources can be confusing and difficult to understand if I’m in the right place or not. I hope that it will just get easier as I continue on in the process.

I think that my biggest challenge throughout this semester will mainly be time management and planning. A large part of this class centers on just that sort of management and planning, but I know myself too well to know that it won’t be incredibly easy for me. The main solution for me is to just keep myself on task and set strict goals and deadlines for myself that I need to meet throughout the semester.