Joe Tucci

September 13, 2019

IS590

1. My interest area for this project centered around two topics; video games and/or baseball. Since I could not land a more specific topic area within the realm of video games, I knew that my topic for baseball would work better for this project. I am mainly interested in looking at effects on a baseball team’s outcome that are not directly game-related. So, looking at things like the weather, stadium, attendance records, and whether the stadium’s amenities and food safety are a deterrent or attractor for fans.
2. This will be very similar to my first answer above, but within the realm of baseball I am interested in looking at the external factors that can have affects on baseball games and teams.

* Weather- One of those includes weather since baseball is played from late-March and early April until October. My studies may or may not look into the playoffs, which can then drag baseball deep into October and sometimes even November. This could have different effects on the weather data, but the number of teams that make it to the playoffs are a small number, especially teams who make it to the World Series at the end of a season.
* Health inspection reports - My main topic of interest was in looking at the food and bathrooms at stadiums for each baseball team. There are some not so official studies and surveys of the best popcorn, hot dogs, etc. that are out there, but I think potentially focusing on more official reports like health reports for each restaurant within the stadium may be more beneficial.
* Baseball statistics - I then want to compare this data with attendance records to compare those sets of data along with the baseball teams’ record in a given year, team batting average, and win/loss percentage.

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. 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.
  5. 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.
  6. Baseball stadium food safety violations
  7. Corey Nachman. Business Insider.
  8. <https://www.businessinsider.com/mlb-stadium-food-safety-violations-2011-3>
  9. 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.
  10. 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.
  11. Baseball-reference.com
  12. N/A
  13. [www.baseball-reference.com](http://www.baseball-reference.com)
  14. 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.
  15. 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.
  16. 2017 Fan Reviews collection
  17. Unknown? Review Trackers.
  18. <https://www.reviewtrackers.com/reports/best-baseball-stadiums/>
  19. 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.
  20. 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.
  21. Baseball Stadium statistics
  22. Andrew Clem?
  23. <http://www.andrewclem.com/Baseball/Stadium_statistics.html>
  24. 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.
  25. 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.

1. Stuff
   1. NOAA US major city weather data
   2. NOAA
   3. <https://www.ncdc.noaa.gov/cag/city>
   4. 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.
   5. 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.
2. Stuff
   1. National Weather Services Climate Services Web Page
   2. National Weather Service
   3. <https://w2.weather.gov/climate/>
   4. 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.
   5. 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.

Narrative

As I’ve talked about a little bit in this report already, developing this research question came about by looking at my interests and hobbies first and foremost. I initially wanted to look within the space of video games, but my lack of a true focus on that topic made it difficult to focus on any specific data. I knew that baseball has and still is a very data and statistics-rich entity that has stats that go back well over 100 years. Not that I want to use data from the early 1900’s, but I knew that being able to find any kind of data from the past 10-15 years would not be terribly difficult.

There is often talk about the many factors that can affect a baseball team at any given time, and I’ve always been fascinated by food at stadiums and how well they are liked by fans. I was mainly just curious to know how much data there was out there about baseball statistics that weren’t about the teams or players. Turns out there certainly is that kind of data out there. I remembered a video we had to watch from 452 last year of a Python demonstration that looked at and transformed a lot of food and health report data from the Chicago area. It went over health code violations and with the data you could determine the best and worst places to eat and the regional areas that were notoriously good or bad.

My idea right now is to hopefully look at all 30 Major League Baseball teams and their stadiums and how their food ranks and does with health code violations from the past 10-15 years. It may turn out that I only look at about a 2-3 year window, but that all depends on the kind of data I can find that is consistent throughout each topic and field.

Finding data about baseball teams and players is the easiest to find, and there are tons of websites that can provide that information for me. I figured there would be just as much information about the stadiums and things like attendance, seating capacity, etc. I was able to find suitable information for that inquiry pretty well, and I think those resources will work well no matter what angle I decide to take with this project.

The food data was a bit more difficult to find. A part of this difficulty comes from needing to look more into each city’s food health department to find the reports through there. It would make that aspect of the project a lot more tedious, but thankfully I have found some already gathered data that sort of summarizes a lot of things I would want to gather and put into this project anyway. Things like basic health code violations and how each baseball stadiums does on average per health inspection. Though, this data isn’t as specific as I would like it to be and does not always look at the frequency of health inspections. Some of the sources I have found may be able to help me fill in that gap, but I also think that I may not even need to get that minute anyway.

Finding the weather data was surprisingly not as easy as I had hoped it would be, but it also wasn’t too difficult either. There are just a lot of variables to consider when researching weather data that made me really think about what I want out of this project. In the end I did find two good sources that will be able to provide me with good data about average temperatures for major cities I would want. It just may require more effort if I want monthly or even daily weather data.