Joe Tucci

590 Open Data Mashups

October 20, 2019

Data Management and Project Management Plans

**What file formats will you be gathering or creating for this project?**

* .py PyCharm
* .JSON
* .CSV
* .html
* .txt
* .docx
* .xml

**How are you storing your data?**

Right now I have my main file directory saved on my laptop within my school folder for other classes. It’s called 590 ODM, and within that is a folder called “Data” with individual folders for each dataset. I have the folder linked to GitHub Desktop, but admittedly I cannot get it to work as intended. It basically takes everything I have in my 590 ODM folder except for my “Data” folder. So, right now it’s not entirely useful, but I’m still trying to get that up and running.

I have the basic Windows recovery options set up for my laptop. But mostly, I back up most of my data manually by doing two things. First, whenever I’m done working for the day I will copy and upload my entire “590 ODM” folder to my UIUC OneDrive account. Instead of writing over the current copy, I make another copy of the folder with a current version name to go with it, in case I need to look back at older versions of my folder and data for whatever reason. I also do the same method with a USB drive as well.

**How are you preserving and protecting all data values from changes, either accidental or purposeful?**

In addition to keeping previous copies of all of my data folders by having multiple copies and versions of everything, I make sure that my original, unedited datasets are kept in a specific folder that goes untouched.

**What will your final data file look like?**

My final data file will most likely be some kind of CSV file, or potentially a JSON file to match. The format should be tied to each stadium, so that each entry is identified by either the stadium name or a stadium ID (which is basically the stadium name.) Each row of data will include elements from each of my datasets like weather, attendance, and health code reports. Because there are only 30 baseball teams in the league, there will only be 30 total rows with many columns for each row. My final data file should read like some of my initial datasets like the Baseball-reference.com dataset that includes statistics for each team, but much more developed and expanded upon.

The file size won’t be terrible large since I’m only working with 30 rows of data, so I fully expect it to come in well under 1GB, if even 500MB. The range of data will most likely stick to the year 2017, because one of my datasets was limited to that year and it made for a better dataset as a result. The weather is also limited to the months when baseball is played, which I estimated to be between April and October.

**How do you plan to disseminate the data once the project is complete?**

I’m still not quite sure where I would want to publish this dataset, because some of the data came from ESPN, and their copyright can be a little tricky. Despite their data coming from public sources like health code reports, the way that I gathered the data was from ESPN and not from another source, so that may complicate things. I think that making my project public on GitHub should be fine.

**Describe hand edits, curation, cleaning, or other alterations that are needed for the data.**

1. ESPN-health-data

Much of this data is hand-picked by me as I web scrape and gather from the .json file that was nested within the .html file. This .json included stadium data from all four major US sports, so I had to first get only the MLB stadiums and their data. Some of the data like the critical score needs some sort of explanation or thought, in that in the article it was listed as a percentage in terms of the stadium’s rate of health code violations. I will also have to contextualize the high-level violations per inspection in both venue and area, and how to best quantify that number per stadium.

My version control with this dataset is based on two different sets. The first being the original .json file because it will have all of the unedited data, and then I also have my initial .CSV where I will have gathered my curated data. I will keep a log of my curated data, why I chose it, and how I changed it if I do within the readme file in the dataset’s folder.

2. stadium-food-safety

I have abandoned this source because I was able to get the health code data that I wanted from the ESPN article. This source also linked to specific city website that housed health code data, and the more I dug into it the more I saw broken links.

3. baseball-reference-team-data

This data was partly hand picked in that the website allowed for the curation before downloading the .CSV file provided. Since this dataset did not require any scraping whatsoever, version control is easy to maintain since it’s a simple .CSV from the website that can easily be gathered again. I may decide to transform some of the attendance data in this dataset on my own. Each team plays half of their games at home, so 81 games of the 162 game season. So, I would multiply the seating capacity that each stadium holds (taken from source number 5 later) by 81 games, and then create a new number by subtracting the total attendance by what would be a projected attendance. This is all relative to how useful this stat would be in my dataset. This dataset also includes data that I may not even decide to use, like payroll or runs scored as well.

Maintaining the original for this is easy with the data staying put on the website, and I also ensure that my original downloaded .CSV file stays untouched as I progress with using this data moving forward.

4. Voice-of-Fan-Report-2017

This data was largely copied into an excel file and saved as a .CSV. The data as it is may need some transformation. It’s based on user reviews and aggregated on a plus and minus 1.0 scale. Some of this data may have some overlap with the actual health code data, and may be transformed and compared to side-by-side with the fan reviews. I’m not sure if this will actually work in any meaningful way, but the option is still there if I can figure out how to best do it.

Version control is the same as every other data source I’m using, in that I will keep an unedited version that won’t ever be touched, and a copy which will be the one I mess around with. This dataset may have a part where I potentially use some of the original and transformed data, but I need to figure out how and why that might happen.

5. Baseball-stadium-stats

This data was very easily web scraped from a table in the .html file, so it was all on me to curate and pick out the pieces of data that I wanted. There was once again overlap with the team name, stadium, and city, but that data is needed as identifiers for the seating capacity and dimensions of the ballparks. Where the seating capacity will be used for a potential new stat elsewhere, I will need to use a list of current stadiums from other sources to further gather only the active 30 baseball stadiums. This list is a compilation of most, if not all, baseball stadiums in history. I have not yet done that, but simply gathering a list of the current baseball stadiums in one single list and then running that through code here will help me get the data I want from this source.

Keeping an original of this dataset will be important, and I will have two separate “originals” in that I will have my initial web scrape of all the stadiums and pieces of data I wanted from each stadium, as well as another original dataset of the active 30 baseball stadiums only.

6. NOAA-weather-service

This data was similar to the baseball-reference.com data in that I was able to give it parameters for the data I wanted, and then from there I was able to just download a .CSV according to the parameters I set. I had to curate the weather dates according to a typical baseball season, from about April to October in a 6-month span, which limits the data. These 6-month datasets act as my originals and unedited copies. This comes from an online database so getting the data back from the internet again won’t be too difficult if I need to. This data also works on a range from the 6-month span, so it largely defines how the rest of my data will be transformed according to this 6-month look (which ties directly to the attendance data as well as that is basically a 6-month span as well.)

7. National-weather-service

I am most likely going to abandon this source. It was mainly a backup in case I could not figure out source 6, but I will still occasionally take a look at this source if I figure out a way to get the kind of data I want.

**Describe the stages of the project (that you determined from your workflows exercise from class) and the associated due dates you anticipate for them.**

Most of the due dates that I’m about to set for myself are very tentative and will most certainly be flexed for a number of reasons. I will either flex my due dates based on other classes or personal matters, and based on each class and my general level of comfort and confidence after each week.

Stages of my project:

Data gathering: All of the data that I wanted for my project is now pretty much collected. I may decided to gather some further datasets, but I have ones essential to my project.

Data organizing and version control: Due date 10/31

I should be close to and have a general idea of how each of my datasets needs to be cleaned and organized by this point. This will require me to thoroughly go back through each dataset and really understand each dataset’s structure and syntax. I’ll also determine where and how I’ll want to transform certain datasets.

Along with this, a week or so from this point I want to have a much better grasp on how my project will function with GitHub desktop and Jupyter notebooks for publishing purposes. I will be at the stage where my time will be split between working locally and working on either of the previously mentioned tools to ensure version control.

Data cleaning: Due date 11/21

This is the week where Jupyter notebook drafts are due, and I’d like to have a firm grasp on my data being cleaned and ready to go.

Data mashing(?): Due date early December

Since this is basically the final portion of the project that I’m working towards, the general range of when I want to be done with this will be somewhere after Thanksgiving to early December, at least I think.