Joe Tucci

Final Dataset Design Plan

590 ODM

November 10, 2019

Structure of final file:

The general structure of my final data file will be centered on each baseball stadium. So, in total it will be about 31 rows since there are 30 baseball stadiums. This means that each row will directly correlate with each stadium and baseball team. The columns will generally be as follows:

* Stadium Name
* Team Name
* Location (City, State)
* Win/Loss record
* Home Record
* Attendance (for 2017)
* Attendance overall difference (Attendance for 2017 - (Seating capacity x 81))
* Venue Score
* Area Score
* Total Inspections
* Critical Inspections
* Critical Score
* Average Score (Voice of Fan report)
* Average Precipitation
* Anomaly Precipitation
* Average Weather
* Anomaly Weather

Some of the concerns I have with creating this file is that all of the things that I want to include, which are listed above, make for a very large file. I think my biggest concern is exactly how best to format the data I’ve collected and how to really go about this process. I understand all of the elements from each of my 4-5 datasets well enough, and I know how they will all be able to relate to one another in meaningful-enough ways for the final dataset. But I assume that one giant file like this maybe isn’t necessarily the way to go. I could certainly clump together things like food scores and weather, or attendance and weather/food and facilities; weather and team’s records; health inspection data and fan reviews; etc. So, my biggest concern is how to best manage this, organize this data, and present it in a neat and legible way.

To do list:

The main things I have left at this point are data cleaning with my five datasets. I have cleaned about two datasets to my liking at this point, with a few more coming along. The few that still need cleaning require some more cleaning in the way of getting rid of a lot of unnecessary data that came with the initial CSV file. The weather datasets all contain the major US cities within the US, so getting rid of about 200 or so rows has required some thought to get through.

At this point, I also need to do way more with Jupyter Notebooks and getting all of the necessary files and code that I need on there. I have not really started that process at all just yet, partly because I’m unsure of how/why I need to utilize it for my final project. I need to spend some time researching and looking through other Jupyter Notebooks to see how I could fit my own project into it.

The last thing I really need to do is tidying up my folders, folder structure, and unnecessary files I don’t need anymore. I had a lot of test files created throughout my web scraping processes along the way. With some of those I even created multiple .PY files to test certain things. I also moved some CSV’s to other folders that didn’t match that dataset for filtering purposes, and while I still want to include that in the final product for documentation purposes, I probably need to rename something like that to better differentiate it.