Taylor Imhof

Bellevue University | DSC 540

Final Project: Write-up

03/04/2022

I must say that this project has been a wild ride. I was able to find a topic that I was interested in exploring pretty early on. I had done some basic exploring of IMDb data before starting this course. I have always been curious about what factors contribute most to whether a movie will be successful. Of course, blockbusters like the Avengers or movies from popular directors will usually draw big box-office numbers. However, I was more interested in seeing if I could identify any underlying trends for other metrics.

I was able to find a rather large .tsv file that housed movie data from IMDb. The people who curated the IMDb data files wound up splitting them into a few different files. For this reason, I did have to combine two of the files together, but this was easy enough using the merge() function built-in with the Pandas package. I did wind up filtering the data to only include information about movies; it came with data for TV shows, which I was not as interested in exploring. During this step, I started to get comfortable with the common data cleaning steps.

I think I was a little too confident after working with the flat-file data because the web scraping milestone gave me much trouble. Working through the BeautifulSoup assignment from the book did not prepare me for the work required for this step. I wound up having to scour quite a few guides and YouTube videos. However, after getting my cleaned dataset at the end of milestone three, I must say that the power enabled by bs4 is incredible.

I learned so much during this step, not just handling Pandas data but also general steps to working with data across the Internet (e.g., handling HTTP errors and adding headers to URL requests). I am so thankful for the immense number of helpful coders both in this program and online. Without their help and troubleshooting efforts, I do not believe I would have gotten this part of the project accomplished.

I initially wanted to work with the official IMDb API endpoints for the API milestone. However, after requesting access a few times (they had me re-submit my applications twice), I was denied access to their free API endpoint. Fortunately, during one of the weeks where we learned about making API requests, the textbook referenced OMDb’s API, which was a godsend for this project section. I wound up using some of the logic from the chapter, but I had to tweak it quite a bit for my project’s needs.

The final step involving databases went rather smoothly. I have used SQL quite a bit, and the Python syntax for sqlite3 was very easy to pick up. It was also easy for me to join the data from my three data sources, as two of them housed the same IMDb identification number, and I joined the last dataframe on matching movie titles. Also, it was nice to expand further my understanding of the Seaborn library for the visualization requirements. That library always impresses me with how simple the code creates pleasing visualizations.

Overall, I enjoyed working on this project. It might have had me banging my head against the wall a few nights, but I feel that I gained so much useful experience working through each milestone. I am looking forward to further increasing my skills in future projects!