North Carolina School District Grades Project

Description

For this project, I mainly performed various data wrangling and munging tasks to merge data from various sources(csv files, web scraping, JSON, SQL, etc.) into one central data file for analysis. I did not have to do any analysis or modeling for this project. It was strictly about gathering data from various sources using Python code.

In order to accomplish this project, I had to perform different types of data wrangling and transformations on all three data sets. The most challenging data set I had to work with is the **School District Location Data.** This data is not in an easy to read tabular format. This data is written into chunks on a website page with separate headings for each district location. I needed to use a web page scraper to extract this information into a single table identifying the geographical region for each school district. I then had to create a join with this table and the **School District Location Data** and create a calculated field to look up each school district and provide the district region for each school. This portion was the most difficult because I did not know how to do either the web scraping or the database lookup/join.

Finally, I had to join the **School Grades Data** with the **School Database** on the *school code number* from the **School Grades Data** and the *facility id* from the **School Database.** These are the unique keys for each school that I can link between the two datasets. This part was not as difficult as there is a one to one relationship for each school in each database. The School Grades Data and the School Database are relatively clean data sources. However, I had to do some analysis to review for duplicates, errors in the data, and identifying any additional inconsistencies in the data. Also, I had to ensure every school received a grade. If the school did not receive a grade, I removed those schools. I investigated the reason for a school not having a grade further to ensure the removal will not throw off the analysis, and the schools did not receive a grade for a specific reason.

Navigation

- Data Wrangling - School Locations (JSON).ipynb: This file is where I clean data from a JSON file into a pandas dataframe.
- Data Wrangling - School Region Dataset (Web Scraping).ipynb: This file is where I scrape data from a web page and clean it into a pandas dataframe
- Data Wrangling - School Code Dataset (Excel).ipynb: This file is where I clean data from an Excel file into a pandas dataframe.
- Data Wrangling - Merge School Data (SQL).ipynb: This file is where I use SQL and the sqlite library in python to merge all three clean data sources into one dataset.
- Additional Code Resources: This is where source data is located for running the Python code above

Tools

For this project I used Python to import/clean the data, perform the exploratory data analysis. I used various Python libraries including Pandas, BeautifulSoup, Requests (reading JSON data), and Sqlite3,to combine data from various sources into one single dataset that could be fed into a data model pipeline for production use.

Data

1. **School District Location(Website)**:
   - This site contains all of the school districts in North Carolina, as well as the education district region they reside within.
   - Link to data:
   - https://stateboard.ncpublicschools.gov/about-sbe/education-districts
2. **School Grades(CSV/Excel)**:
   - This file contains the school grades for every school in North Carolina for the 2018-2019 school year.
   - Link to data:
   - https://files.nc.gov/dpi/documents/accountability/reporting/spg-report2019_october.xlsx
3. **School Database (JSON)**
   - This API contains location information about each school within North Carolina
   - Linkt to data:
   - https://www.nconemap.gov/datasets/dea6ff0e8b4743a0ba361e13a85a4c70_3/data?orderBy=OBJECTID&orderByAsc=false

Methods/Evaluation/Techniques

I first had to extract data from a variety of different sources including a simple csv/excel file to web scraping and APIs. Data wrangling from the csv file was the easiest wrangling task of this project. The file was in a mostly clean format and just needed minor tweaks. Where I really had to dig in and learn new concepts was the web scraping and API ingesting tools. These techniques took quite a bit of data wrangling in order to obtain and clean the data. The data from the API was in JSON format and I had to learn JSON and how to unpack it with Python. For web scraping, I had to decipher the HTML of the web page that contained the data I wanted to capture. Once I identified where the data was stored in the HTML, I had to unpack it and create a data frame out of the column heading and body of the data.

After I had all three of these data sources in a clean, easy to understand format, I had to write these data frames to database tables using SQLite3 in Python. This was an excellent tool that made importing pandas data frames into a SQL database easy. Once the three data sources were in a SQL database, I then had to identify the common keys to merge the data sets on. This was the school ID for the School Grades and School County databases. For the School Region database, this was the school county name. This took some trial and error to understand the different join types and how I had to join for each table. Once all three tables were merged, I was able to properly perform data analysis on my combined data set.

Data visualizations were the third major thing I had to learn and complete for my term project. I used Matplotlib as well as Geopandas to create my visualizations for this project. Matplotlib is an excellent module for data visualizations in Python allowing for countless amounts of customization. Geopandas was an interesting package I came across on the internet that allowed me to create a visualization on the state of North Carolina.

Conclusion

Overall, this was a project that showcased not only my ability to extract data from various sources across the internet, but also my ability to solve problems with critical thinking. There is not a simple solution to extracting data from various sources across the internet and each data source presents it’s own challenges.