ETL Final Report By Erin Cunningham

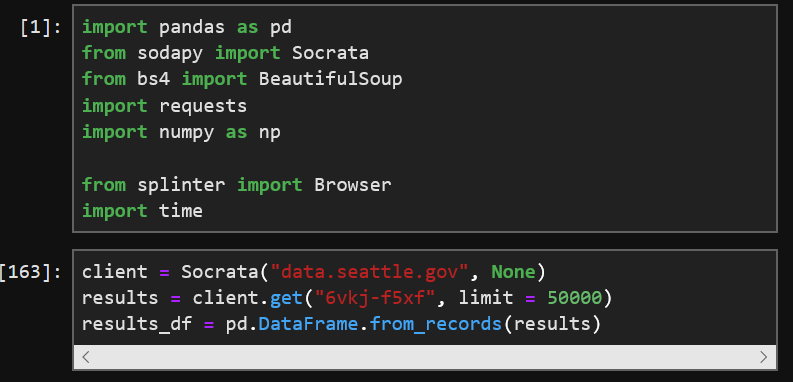
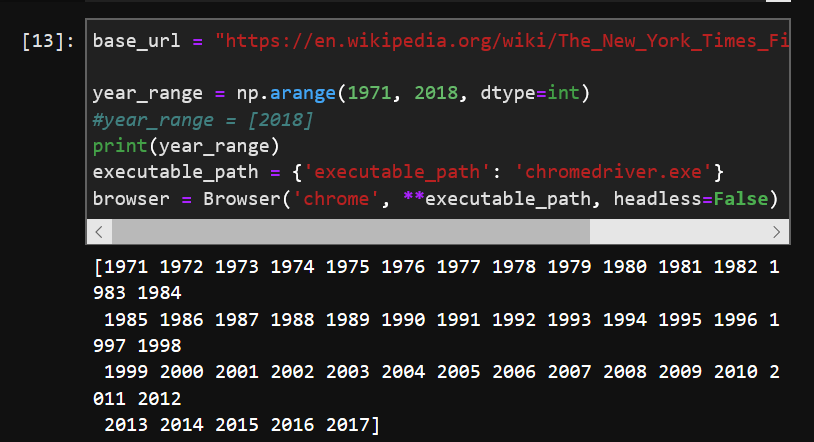
And

Chidera Nwaubi

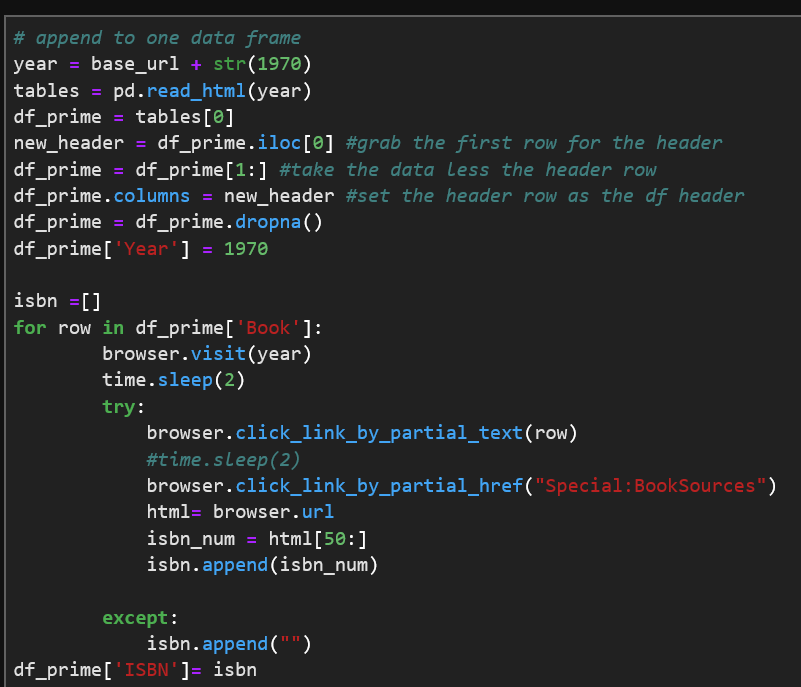
# Extraction

Our first data set came from Socrata via Kaggle – [The Library Collection Inventory for the City of Seattle.](https://dev.socrata.com/foundry/data.seattle.gov/6vkj-f5xf) Socrata offers an API for the open data sets on its site and there is a Python library, sodapy, that streamlines that process. We pulled just a fraction of the total collection, 50,000 items. We did not limit this to any particular type of item, so the data includes books and other assorted entries.

Our second data set came from a web-scraping Wikipedia to gather a list of the [New York Times Bestsellers from 1970 to 2017](https://en.wikipedia.org/wiki/Lists_of_The_New_York_Times_Fiction_Best_Sellers). The webpage for each year contained a table with the week, book title, and author. We were able to use Pandas to scrape those tables and drop them directly into a Dataframe. We then created a Selenium scrip that clicked into each of the book titles, where available, and returned the book’s ISBN number. We chose this method because it was surprisingly hard to locate a database of ISBN numbers without paying for them.

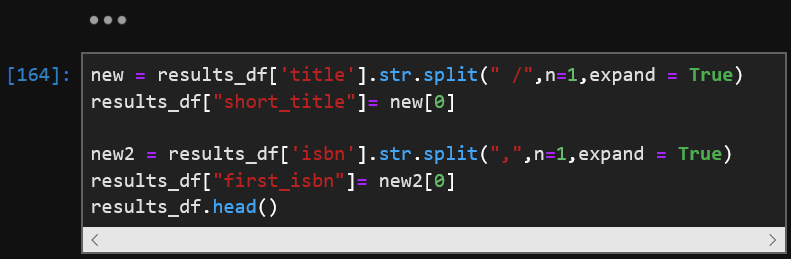
 

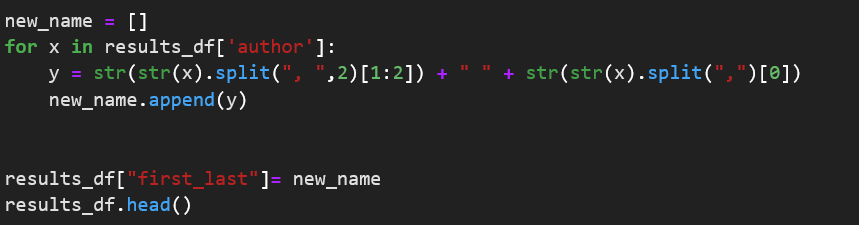
**(Socrata API Data Extraction) (Wikipedia Web scrape)**



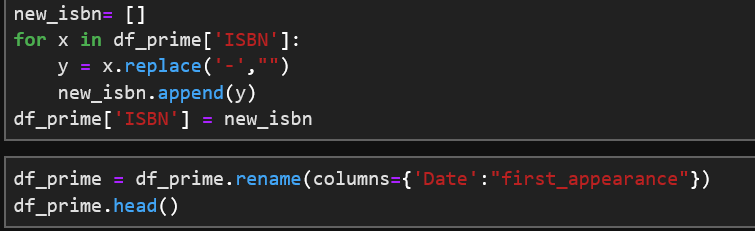
# TRANSFORMATION

We dropped both data sets into Pandas for easy viewing and began the cleaning process by manipulating the columns. One of the main issues we ran into in the Socrata data was that some of the fields contained lists or strings that were hard to manipulate in a large data set. For example, the ISBN field contained all the ISBN numbers for a title, representing multiple editions of the book, making it hard to perform a straight join with our NYT’s data. We were able to manipulate the title data column consisted of the title, type, and author’s name, so we extracted just the actual title and created a new column called short\_title.





When we pulled the NYT data into Pandas, we had a few issues. Some titles appeared on the list for multiple weeks at a time. Pandas would show these weeks as having null values. We decided to drop the null values and the duplicate values and just show the first appearance of a title on the bestseller list. We also added a year column to the Dataframe during the web-scraping process. We then removed the dashes from the ISBN so that the format would better match the ISBN from out Socrata data. We attempted to split the author data into first and last name columns, but ran into issues where the author title did not meet the expected format; for example if there were multiple authors or the author’s first name contained spaces. Thanks J K Rowling.



# Loading

Because we had both of our datasets in Pandas, loading into SQLite was relatively easy. Once that was complete, we attempted to performed a join via the ISBN and by book title. If given more time we would have likely found a better source of ISBNs besides Wikipedia because one thing we noticed was that a book could have multiple ISBNs. We decided that the title was a better join than ISBN. While there were only 15 results with the ISBN join, we ended up having 61 matches with the join on book title name.

