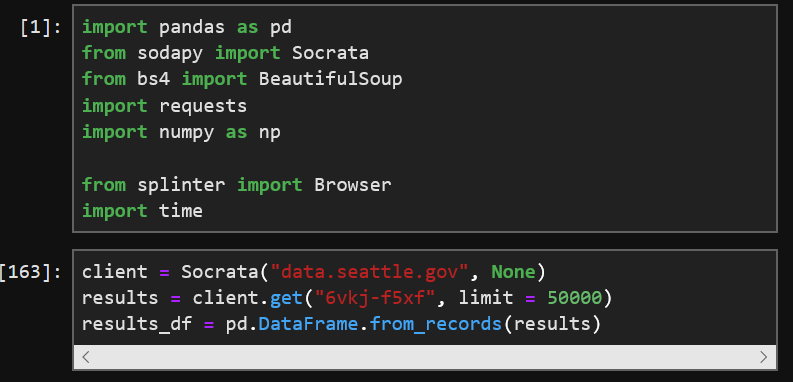
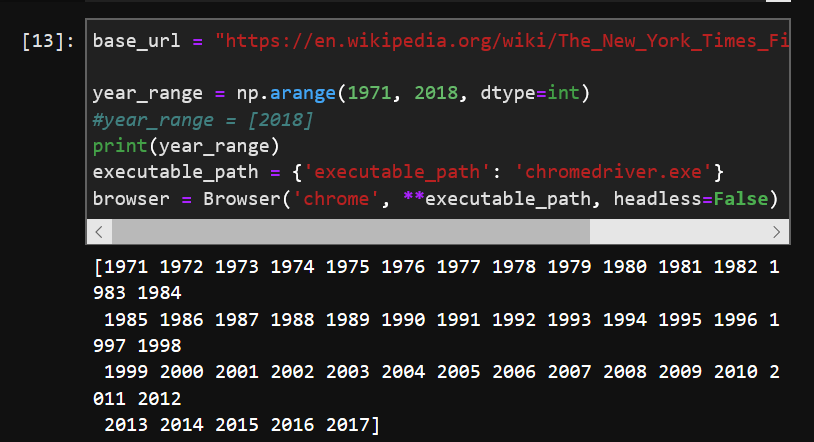
ETL Final Report By Erin Cunningham

And

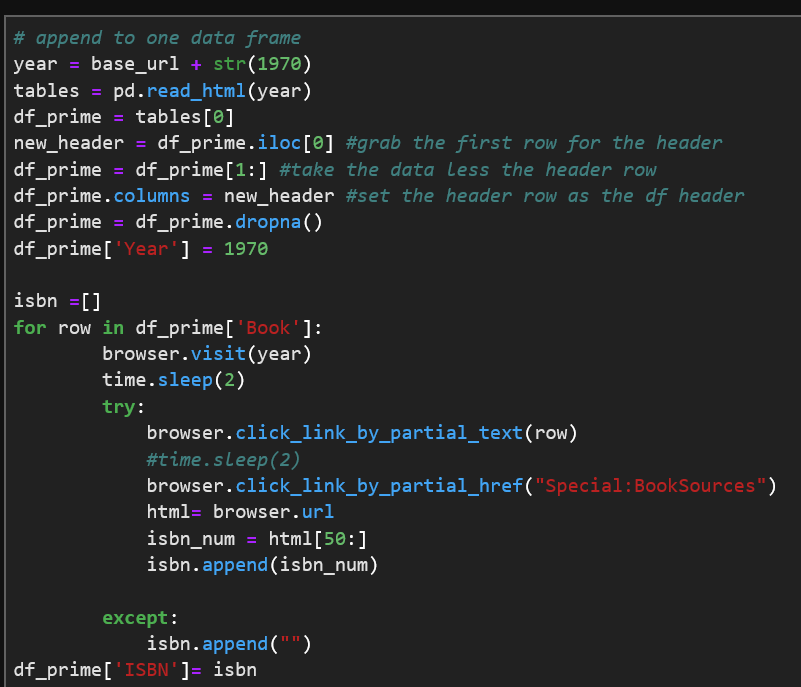
Chidera Nwaubi

# Extraction

For the extraction stage of the ETL process, we chose to use Kaggle as one of our data sources, and Wikipedia as the second source of our data. To extract our Kaggle data, we performed an API call from an open data site/library Socrata. This dataset consisted of the full inventory(50k+) of items contained in Seattle Public Library’s collection. If given more time we could have restricted the data to books listed in the library and not include the DVDs, magazines, etc. For our Wikipedia data source, we performed a web scrape of the New York Time’s best seller’s list data(HTML Tables) on Wikipedia from 1971 to 2017. We achieved this web scrape utilizing Selenium. There were 676 New York Time’s best sellers at that period. We wanted to see how many books in Seattle’s Public Library were New York Time’s best sellers.

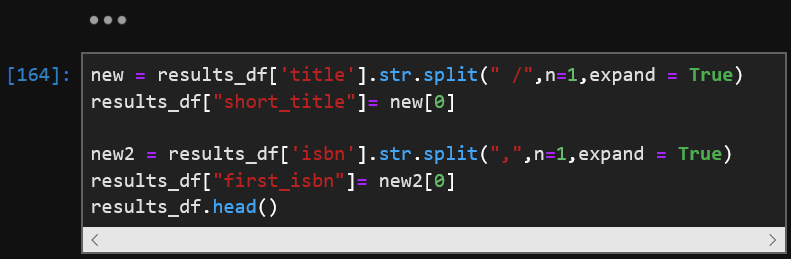
 

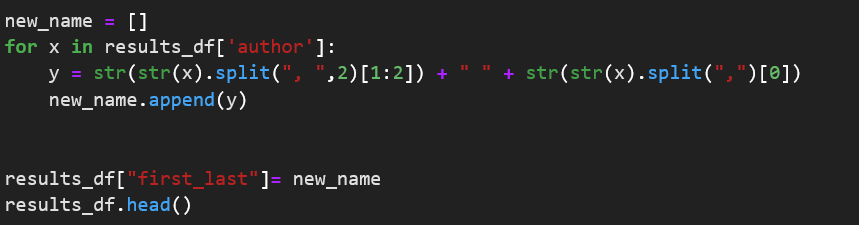
**(Socrata API Data Extraction) (Wikipedia Web scrape)**



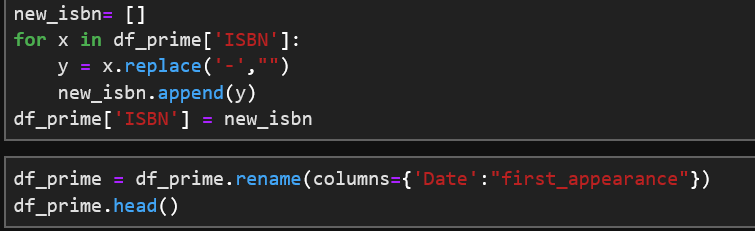
# TRANSFORMATION

We began the cleaning process by manipulating the columns. The Socrata 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. Another issue we encountered was that there were multiple ISBNs in the ISBN column. We cleaned this up by creating a new column that consisted of the first ISBN in the record. Also, the author name column included the author’s name as well as a year. We cleaned this column up by removing the year from the author name column. The Socrata data we loaded unto the data frame results\_df.





For the Wikipedia web scrape, we loaded the extracted output into a data frame df\_prime. We edited the ISBN to remove all dashes from the data frame we also changed the name of the column date to the first appearance to notate in the data frame that only the first year a book was placed on the best seller list was extracted. The edit was performed because some books became NYT best sellers at multiple occasions during the 30 years, but we only wanted to count each book on the list once.



# Loading

For the loading of the data initially, we were going to use CSV as we even wrote the web scrape Wikipedia data onto CSV we decided, however, to write both the results\_df( Socrata data frame) and the df\_prime(Wikipedia NYT best seller’s data frame) to SQL via SQLite. Once that was complete, we performed a join via the ISBN. 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.

