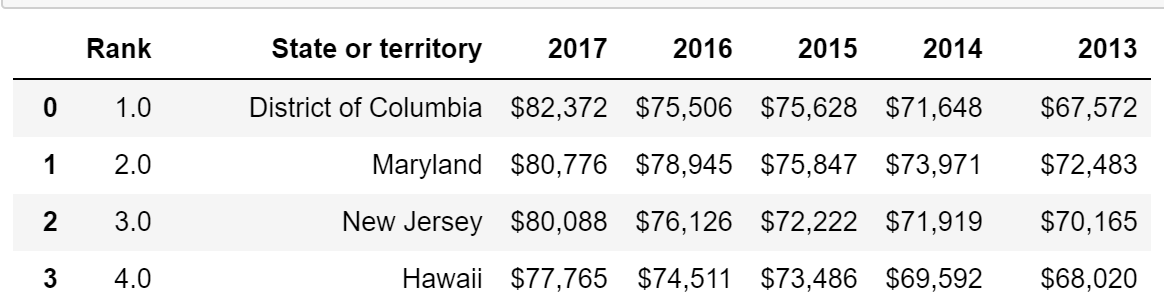
ETL Project – Health and Income Outcomes by State

Team Members: Tanvir Bhullar, Jess Alcalde

For our ETL project, we chose to create a relational database in PostGres that would give a user access to income data per state and causes of death per state, and do further correlation analyses.

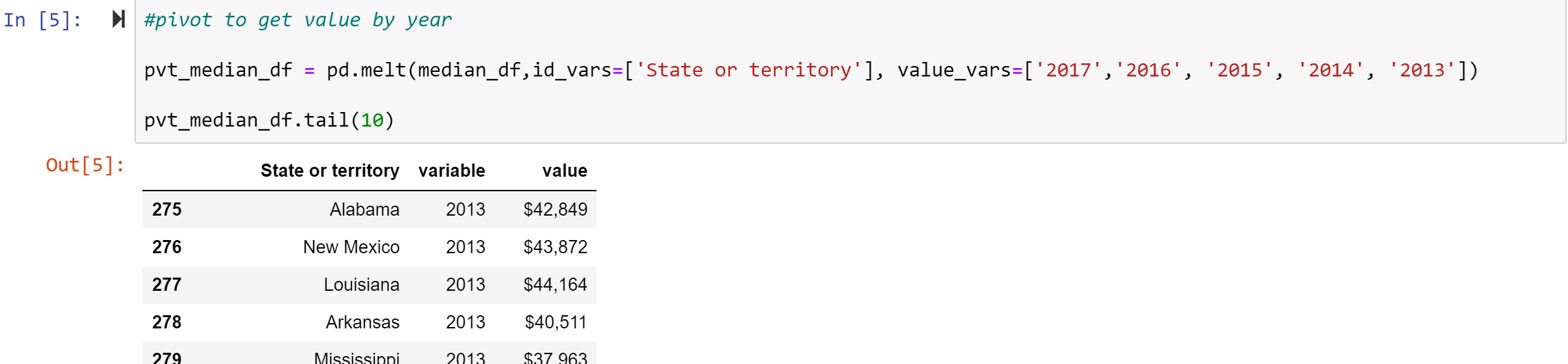
Extraction and Transformation:

**Income by State:**

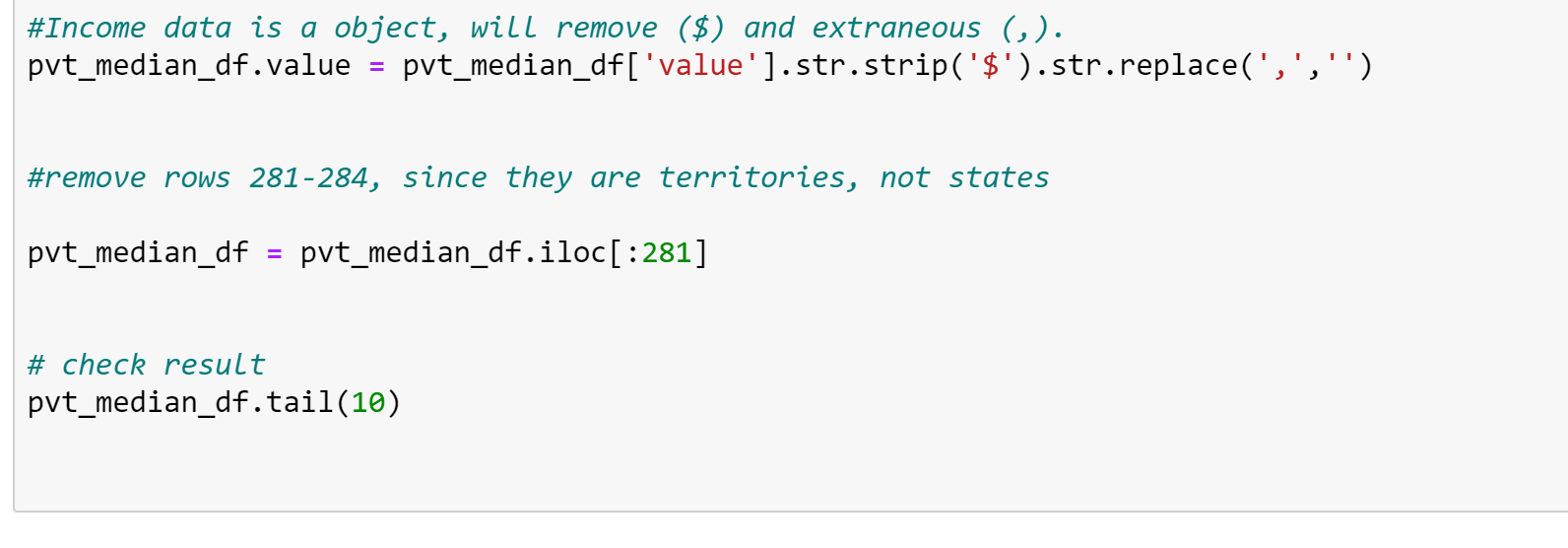
Our income data per state was scraped from a Wikipedia page(<https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income> ). The data included income data by year and state. 

We used pandas (read\_html) to read the contents of the website into a dataframe (since it was in tabular format), then used pandas again to clean up extraneous symbols from the columns ($) and pass the contents into a csv.

* Transformations:
  + Index dataframe to only data from income table
  + Rename 2013 column to remove [note 2] text
  + Utilize pd.melt to unpivot data along the Year variable.



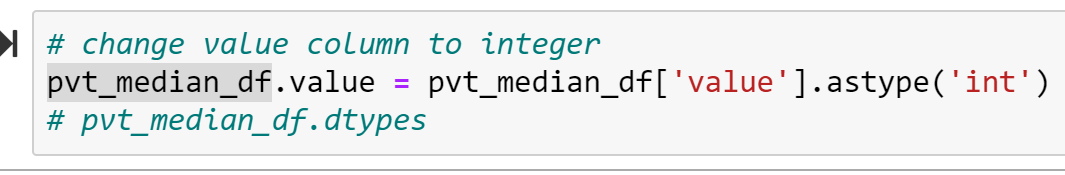
* + Because the value column contained textual data such as “$” and “,”, we remove those.



* + We had several rows that contained US territories and not states, so we removed them



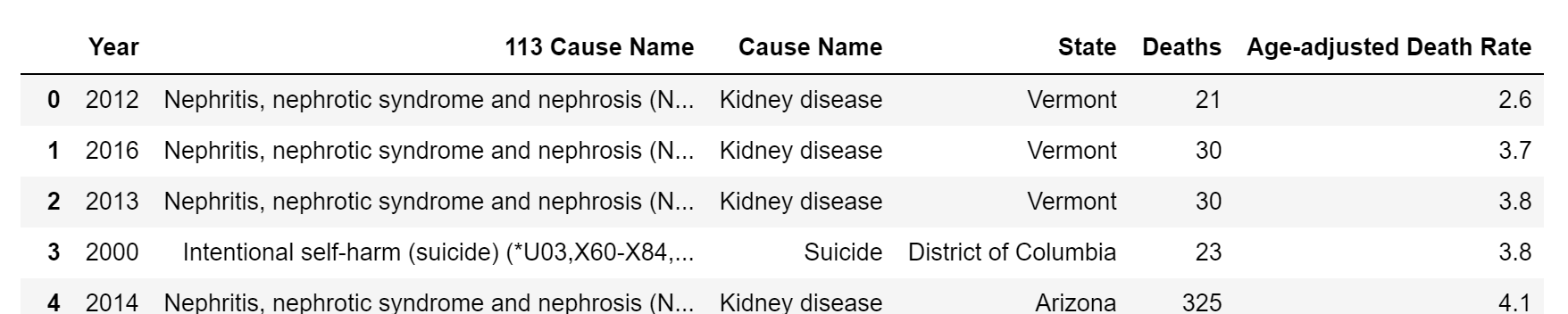
* + Convert value column to integers to make numeric operations possible



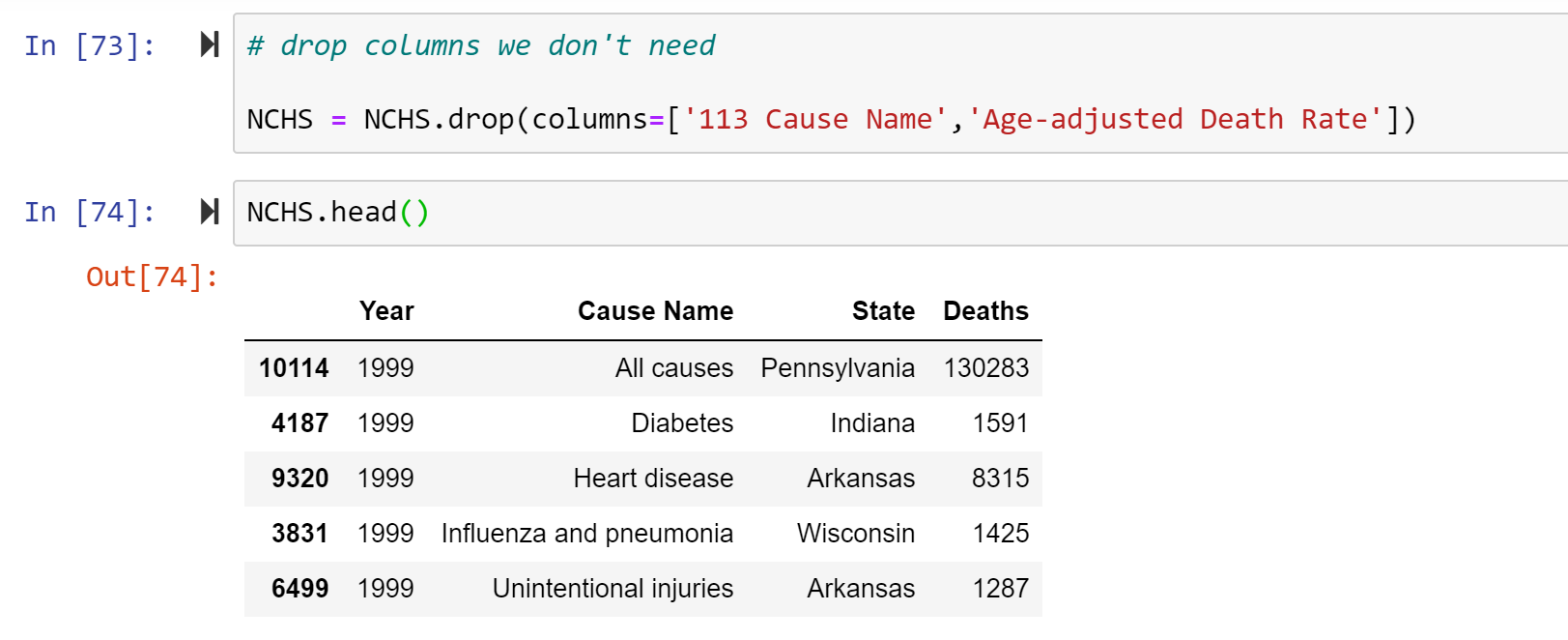
* + Use to\_csv to output into a csv file.

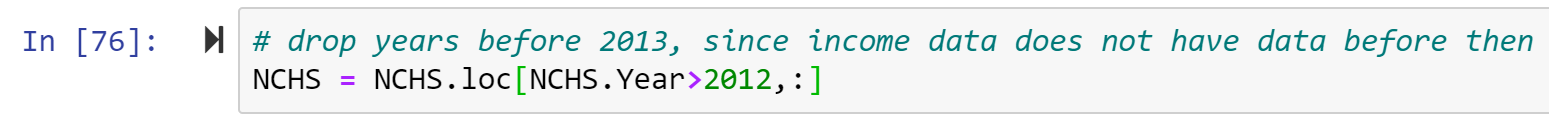
**Deaths by State:**

Our death data was sourced from the NCHS (NCHS.gov). The data was originally sourced from a csv directly from the website. The NCHS dataset included year, Cause Name (more descriptive), Cause Name, State, Deaths, and Age-adjusted Death Rate.



* Transformations:
  + Sort Values by state and year
  + Drop columns we wouldn’t use for the analysis (113 Cause Name, Age-adjusted Death Rate) – These columns seemed too difficult to interpret.



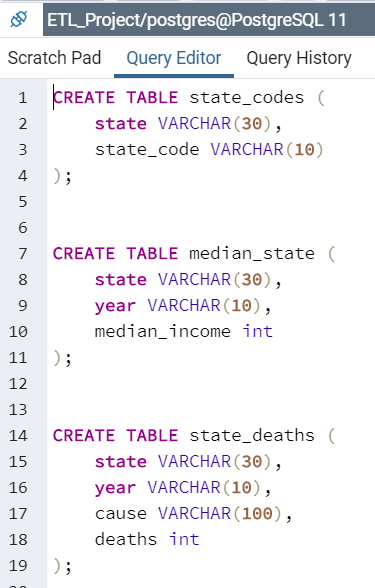
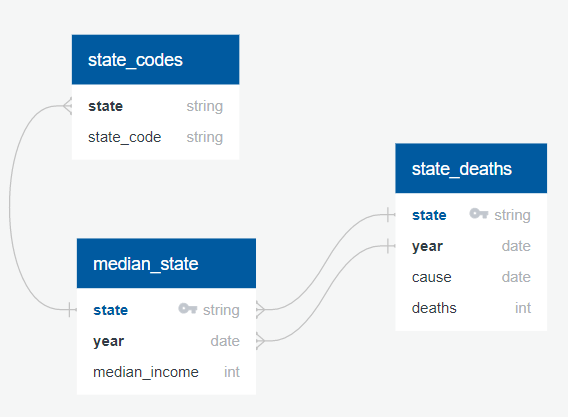
* + Drop years before 2013, since the income data we had did not have data before then: 
  + Change the ‘Year’ column to a string:



* + Use to\_csv to output into a csv file.

Loading:

For our database, we used PostGres to input our tables into. The below code was used to create the schema for our tables:

We used the native import function to import the csvs into our database.

Sample Queries:

