**Joshua Marcus**

**Rutgers Data Science Bootcamp**

**Excel Homework**

**July 25, 2019**

**ETL Project**

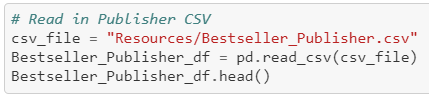
**Extraction**

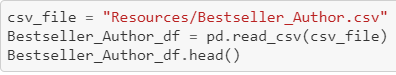
The dataset used was from the public platform Data World and consisted of New York Times Bestsellers from 2011 to 2018. The dataset was split into two csv files via excel.

New York Times Bestsellers from 2011 to 2018

<https://data.world/typhon/new-york-times-bestsellers-from-2011-to-2018>

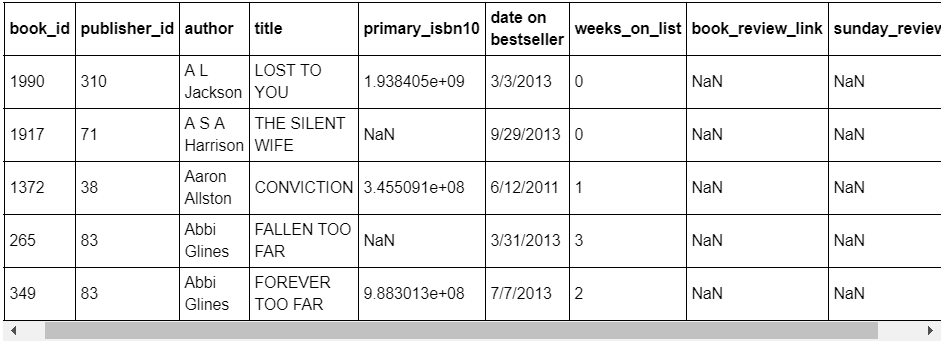






**Transformation**

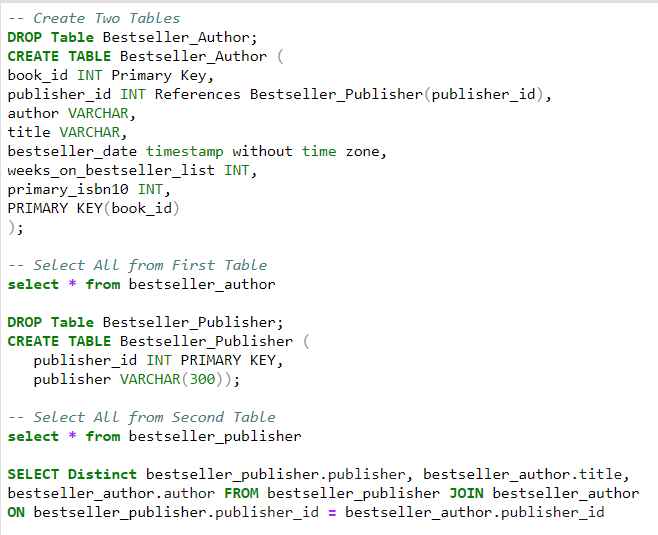
The first steps in cleaning up the datasets involved figuring out which columns were not useful. The amounts of employees for all of the states were removed. In addition, unique keys were created for book\_id and publisher\_id.



Which brings us to our new data with our seven selected columns in table 1 and 2 columns in table 2.

**Load**

The last step was to load dataframes into tables in PostgresSQL. We created a database and two tables in PostgresSQL to match the columns from pandas dataframes.



**Summary**

We used these datasets to identify the publisher used by each author and to identify the unique book that won best seller. The datasets are from 2008 – 2018, however, they are general enough to be added to as a running list.