ETL Project: Zillow Housing Pricing vs Median Income by State (2009-13)

By Kenny Sun, Duong Luu, Hugo Garcia

**Extraction**

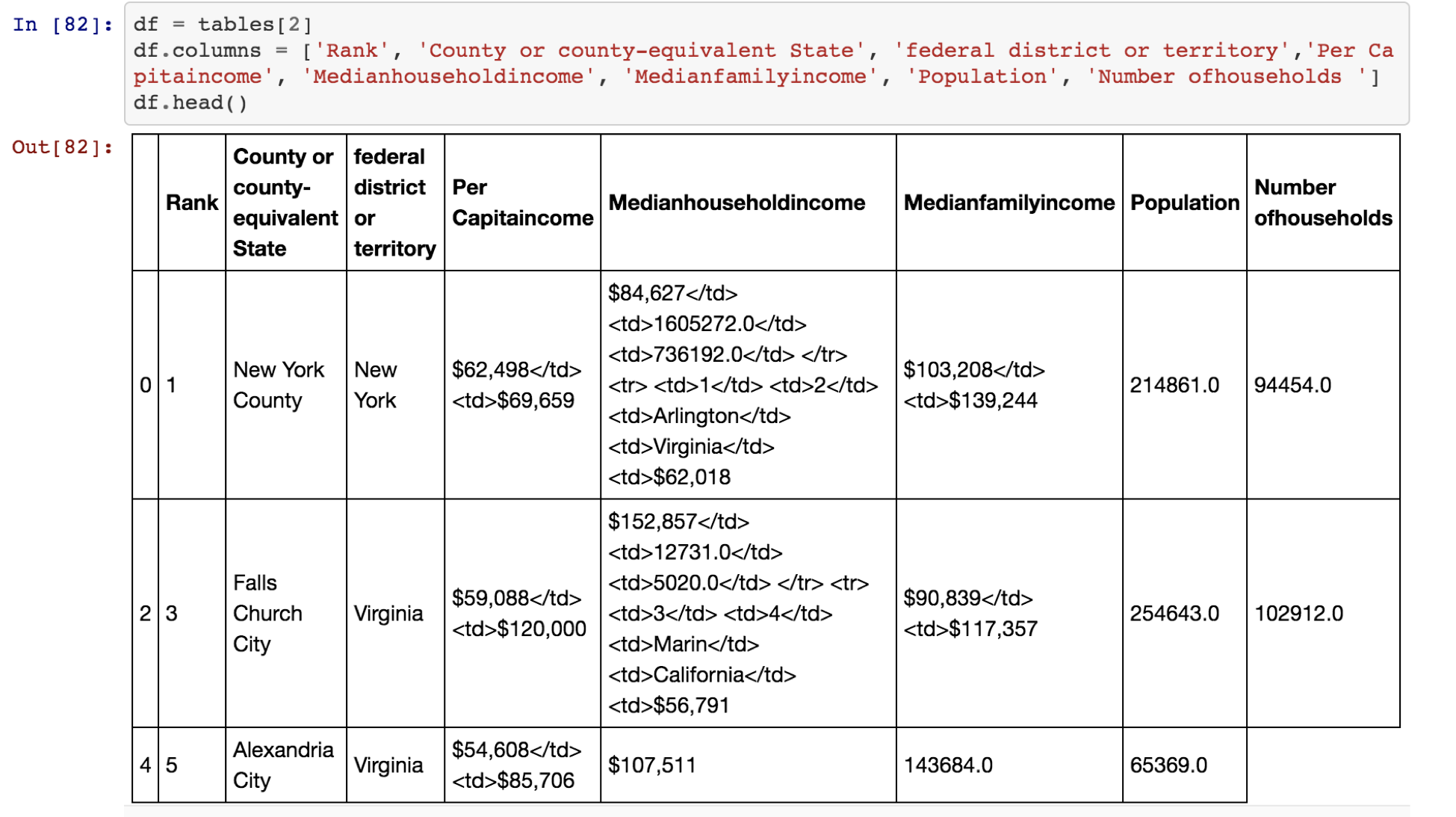
We used 2 datasets from the public platform Kaggle and Wikipedia. All of our data was based on States ranging over various years from 2009 to 2013. These were the most recent ones we could find. The sources for our dataset are as follows

· Zillow House Price Data from Kaggle.

· Median Income by state from Wikipedia.

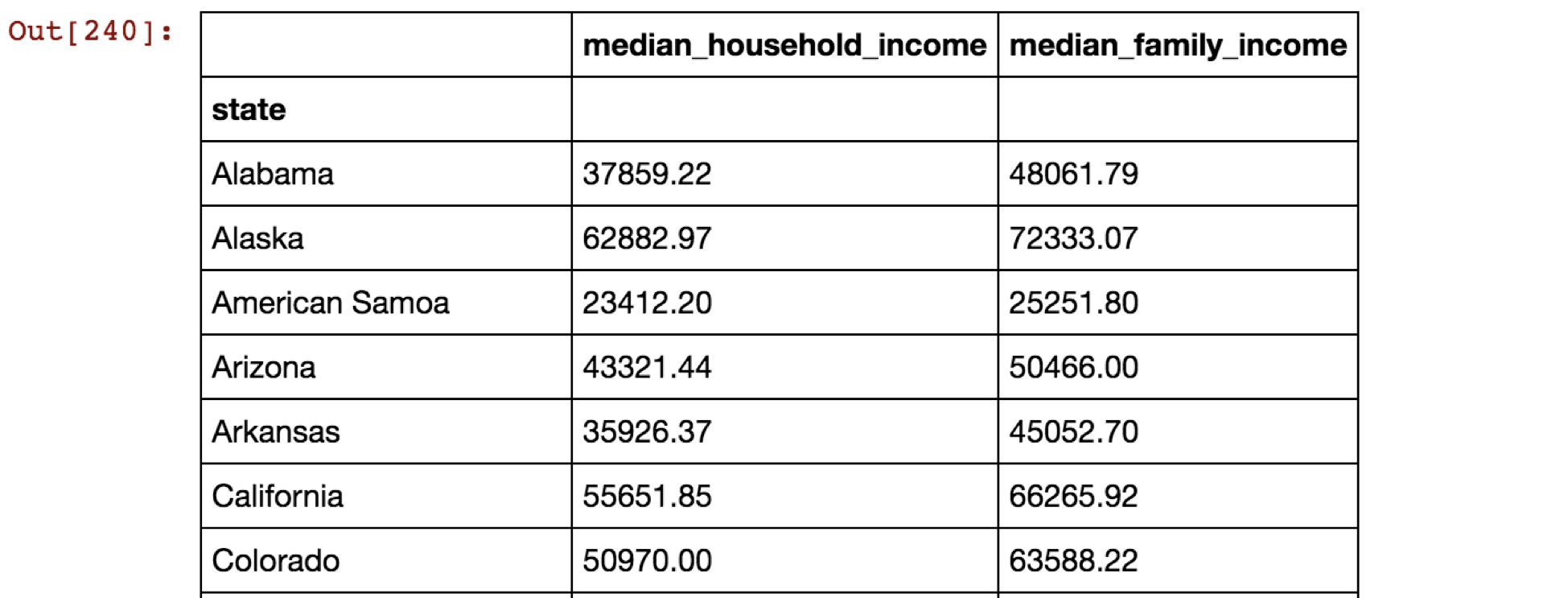
**Transformation**

Our first steps in cleaning up the datasets involved figuring out which variables were not relevant. For the Median Income Dataset (Figure 1), we used read\_html(url) function to extract the tables from Wikipedia, then we only kept the State, Median Household Income and the Median Family Income columns.



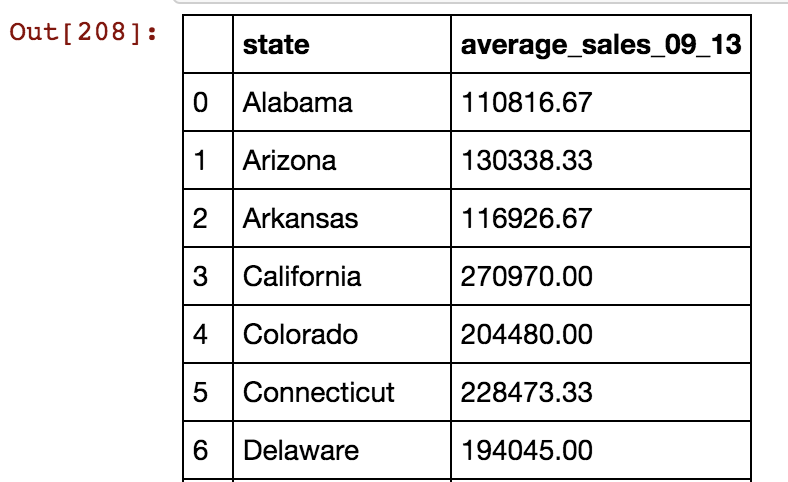
**Figure 1:** Extracted necessary data from Wikipedia

We also changed the “state” columns type from ‘INT to ‘Str’ and set it as index. (Figure 2).



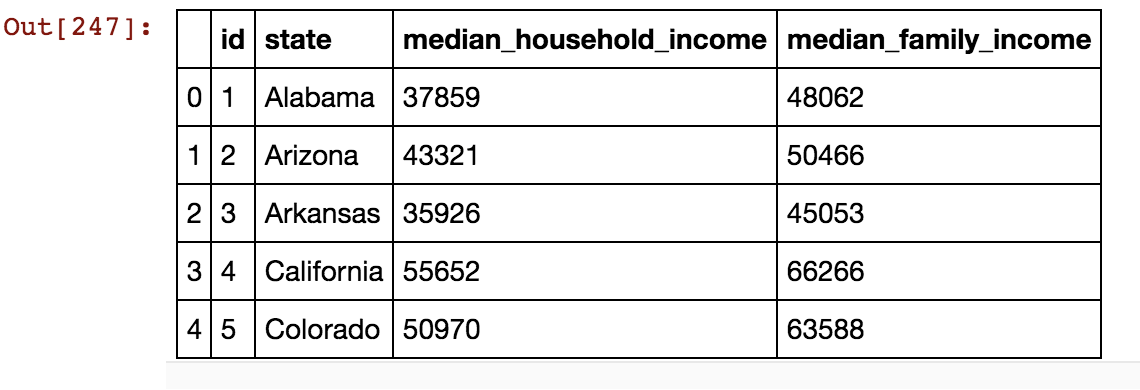
**Figure 2:** Final output of cleaning the data

For the Zillow dataset, we extracted the CSV file and cleaned the data. Since the file contained data from 2008 to 2019, we decided to drop and only keep the years that would be useful. We averaged the sale prices for years 2009 to 2013 to match our Wikipedia dataset.



**Figure 3:** Dropping unnecessary columns and averaged sales price by State for years 2009-2013

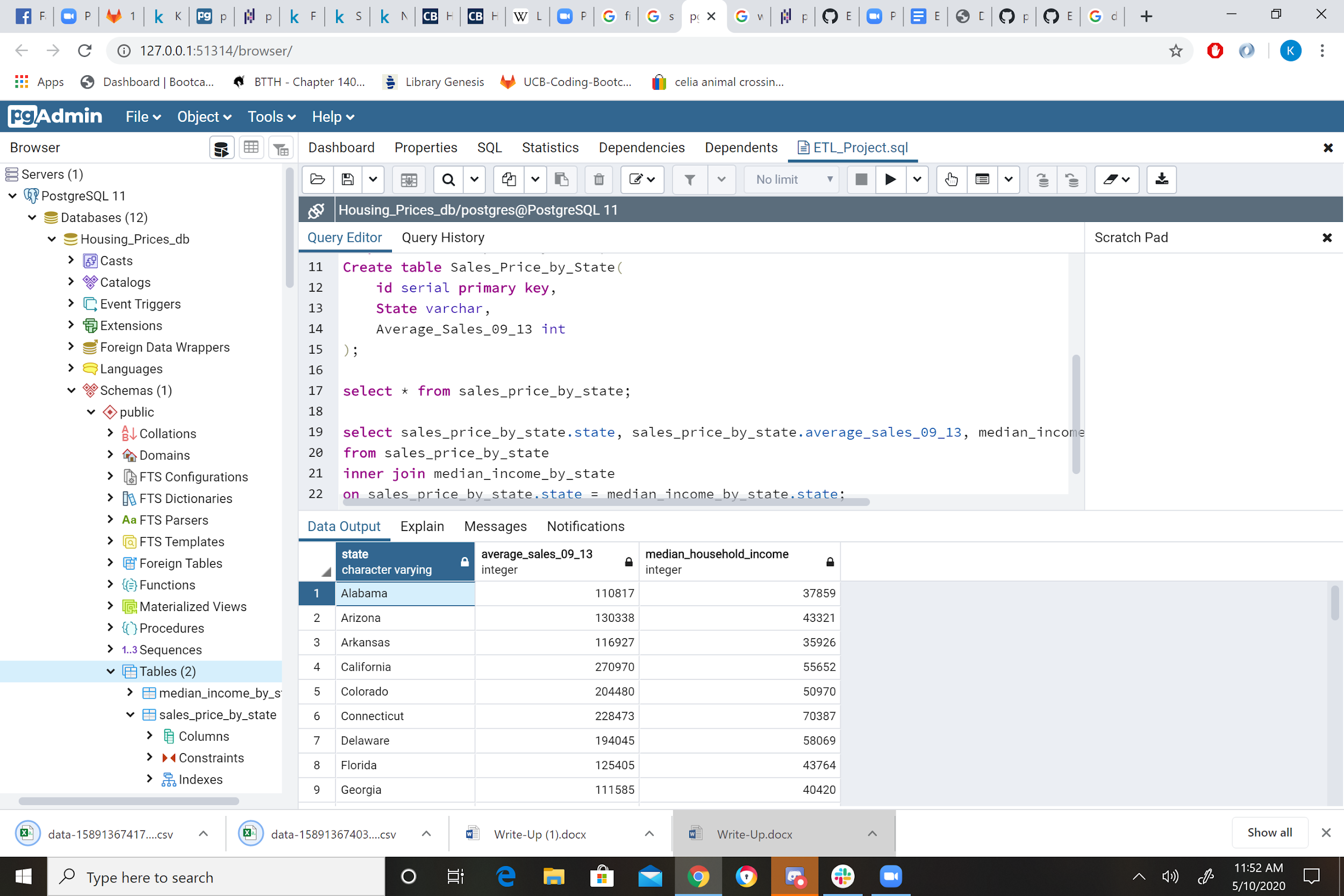
After cleaning the datasets and workable, we created Median\_Income\_By\_State and Sales\_Price\_by\_State tables in the database. In the Median\_Income\_By\_State table, we have the “id”, “State”, “Median\_Household\_Income” and “Median\_Family\_Income” columns and set the “id” column as the primary key. In the “Sales\_Price\_by\_State” table, we have “id”, “State” and “Average\_09\_13” columns and also set the “id” column as the primary key. Finally, we merged these two tables together using an inner join on the “State” column.



**Figure 5:** Final Output after merging all the datasets

**Load**

The last step was to transfer our final output into a DataBase. We created a database and respective tables to match the columns from the final Panda’s Data Frame using PostgreSQL.



**Figure 6:** Sample Query

**Limitations**

Alaska - Not in Sales Dataset/Removed

Vermont - Not in Sales Dataset/Removed

Maine - info missing for 2009-2013

Mississippi - No info for 2009-2013

New Hampshire - No info for 2009-2013

Rhode Island - No info for 2009-2013

**The following four states above that were found missing. Their values will be set to 0 in order to run the aggregate function, average, on the dataset.**

**Summary**

We used these datasets so we could identify the housing sale price and median household income for each state in the United States. The final output will help us to identify which State has the highest house sale price, highest median income and the cheapest cost of living.

· Median Family Income by State.

· Median Household Income by State

· Average House Sale Price by State