.

ETL Project

By: Kayla K., Ryan B., Alex G., and Paul P.

**Pre-Processing**

With the worlds focus on clean renewable energy, our group wanted to collect energy usage/production data in the United States, compared to population data. The data analyst will use our database to find the most current data available to observe states utilizing cleaner energy sources and which energy source each state uses. This database can be used to rank the best and worst states based on their energy consumption.

|  |  |  |  |
| --- | --- | --- | --- |
| Step | Action | Notes | Observation |
| 1 | **Topic** | Energy Consumption by Population |  |
| 2 | **Data Sets** | US Energy Information Admin   * State Total Energy Rankings, 2018:   Link: <https://www.eia.gov/state/?sid=US>   * Primary Energy, Electricity, and Total Energy Price Estimates, 2018   Link: <https://www.eia.gov/state/seds/data.php?incfile=/state/seds/sep_sum/html/sum_pr_tot.html&sid=US>  Census Bureau   * Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2019 (NST-EST2019-01)   Link: <https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html>  American Council for an Energy-Efficient Economy   * The 2018 State Energy Efficiency Scorecard: <https://www.aceee.org/sites/default/files/publications/researchreports/u1808.pdf> | Main data set begins only contains data from 2018. We will extract the data from all other data sets and only used the year 2018. |
| 3 | **Verify Data** | Ensured there was a primary key to link all the tables (States) and that each data set had usable data. |  |

**Extraction**

We used 4 different datasets from the US Energy Information Admin, Census Bureau and American Council for an Energy-Efficient Economy . The data in the three files included the following information:

1. Energy Types by State
2. Energy Consumed by State
3. Production/Consumption/Expenses Ranks
4. Population by state

\*Included data from 2018 years.

The following sources for our datasets used:

* Primary Energy, Electricity, and Total Energy Price Estimates, 2018 (Dollars per Million Btu) <https://www.eia.gov/state/seds/data.php?incfile=/state/seds/sep_sum/html/sum_pr_tot.html&sid=US>
* State Total Energy Rankings, 2018 **:** <https://www.eia.gov/state/?sid=US>
* Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2019 (NST-EST2019-01:<https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html>
* The 2018 State Energy Efficiency Scorecard: <https://www.aceee.org/sites/default/files/publications/researchreports/u1808.pdf>

**Transformation**

In order to transform the public data and use it in our study we performed the following:

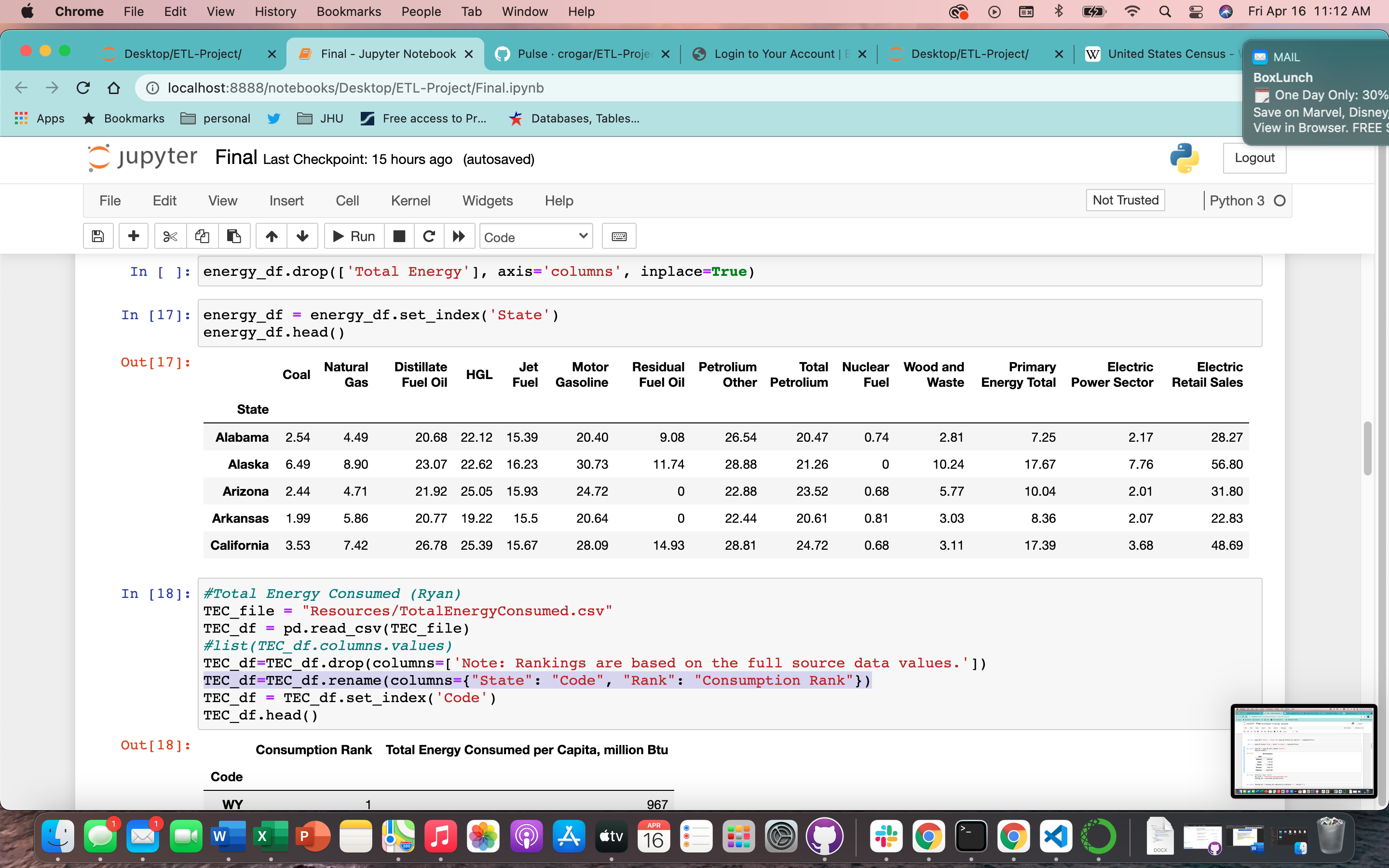
* Used Pandas functions in Jupyter Notebook to load all three CSV files.
* Reviewed the files and transformed into data frames
* Removed columns that was not relevant to the focus of this study.
* Replaced empty data fields with necessary data inputs, (i.e “---” to “0.00”).
* Merged datasets on “State” and “State: Code”
* Exported data frames to csv.
* Created table schemas in our queries to individually access cleaned data
* Created a data based by grouping all data by state .

Jupyter Notebook DF Screenshots:

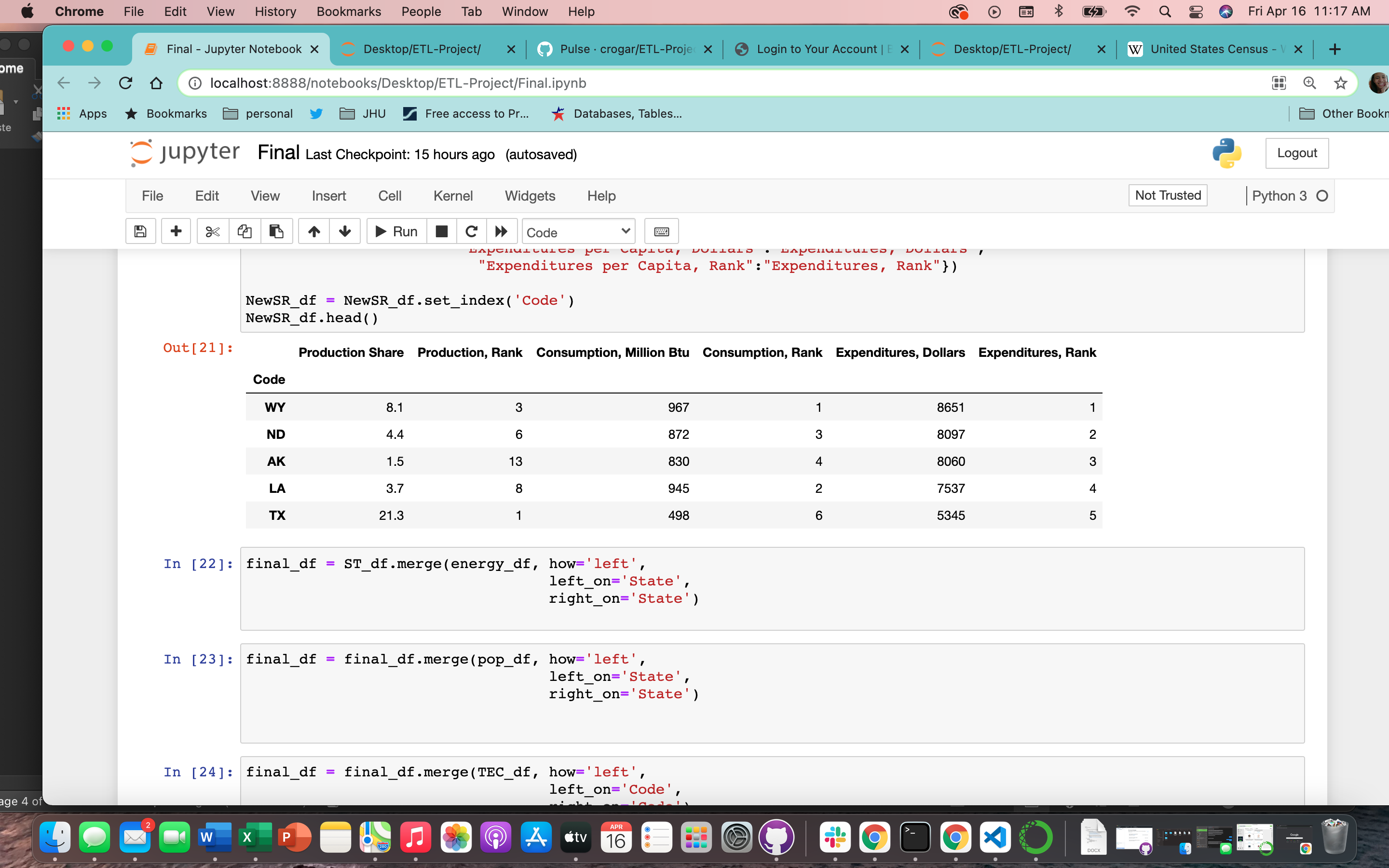
* **State Population DF, Sourced from Census**



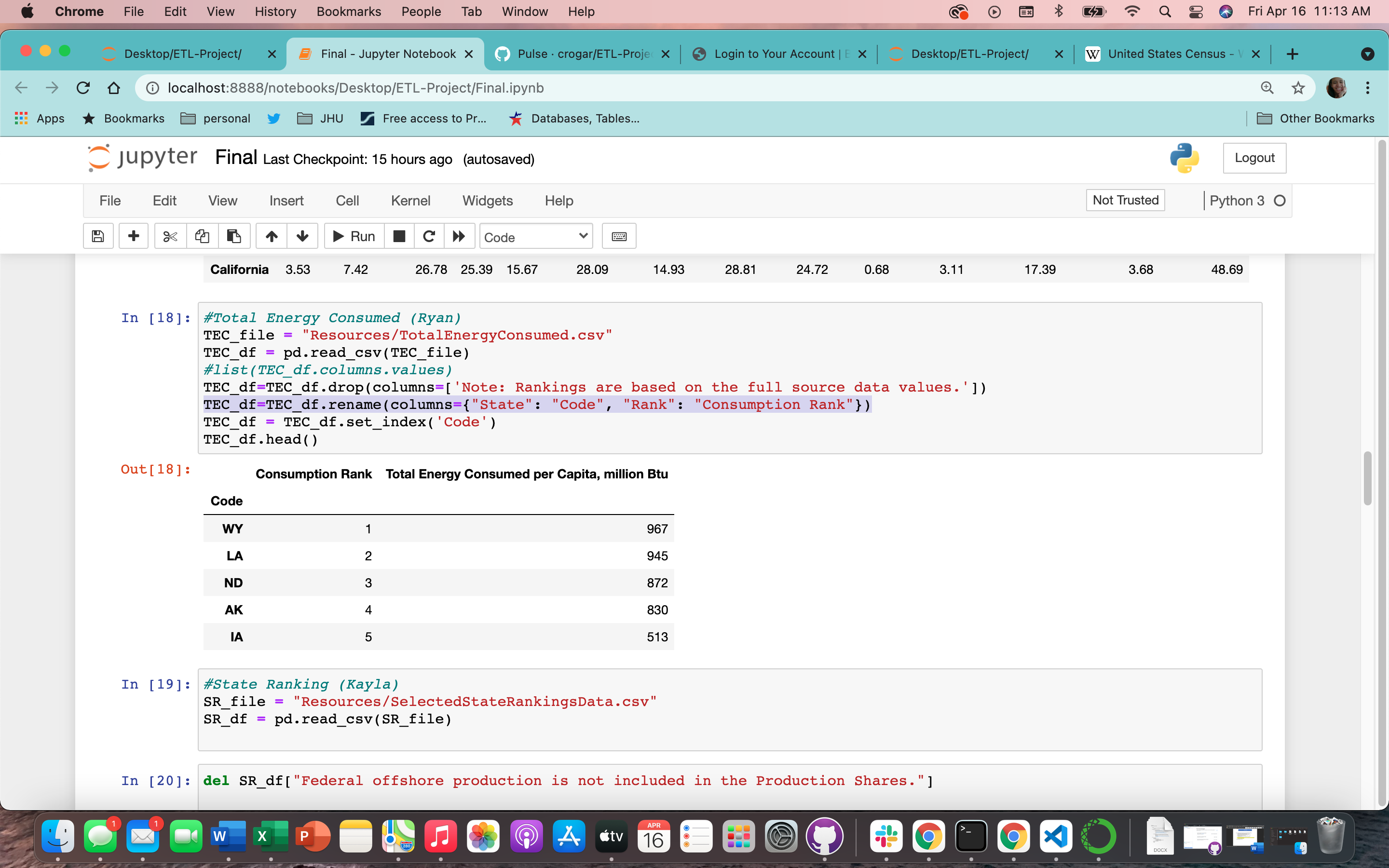
* **Energy Type DF, Sourced from EIA**



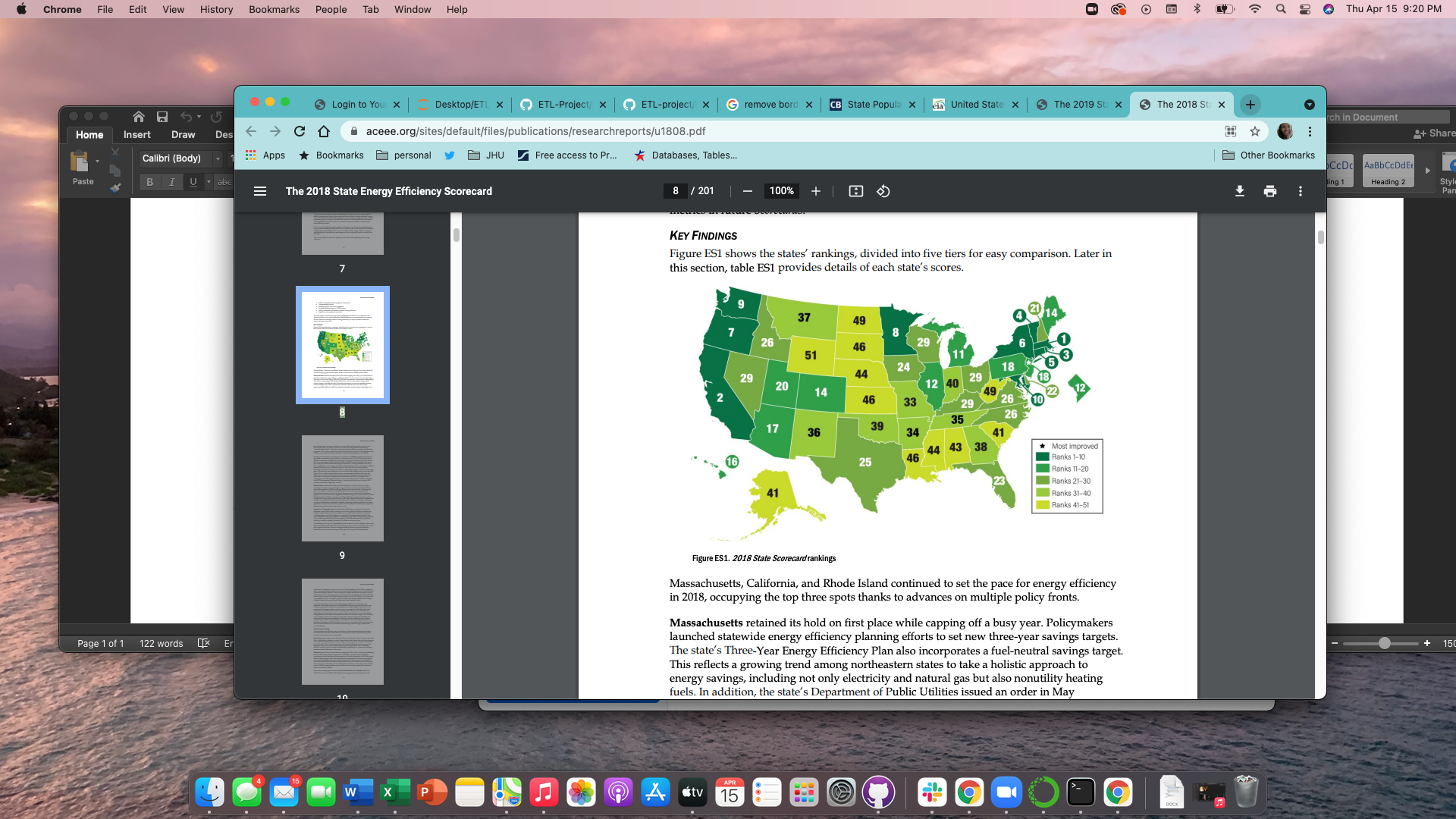
* **State Energy Ranking DF, Sourced from EIA**



* **Total Energy Consumption DF, Sources ACEEE**



**🡪 To view the DF process please see FINAL.ipynb file in folder.**



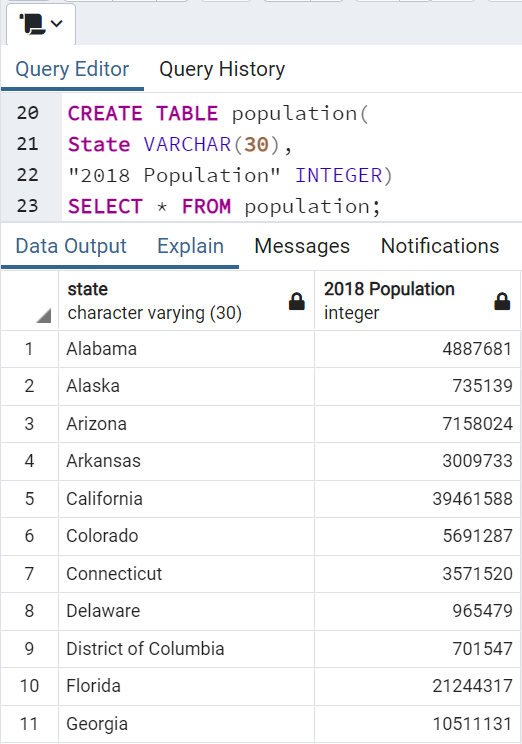
* Photo: <https://www.aceee.org/sites/default/files/publications/researchreports/u1808.pdf> Pg. 8

**Load**

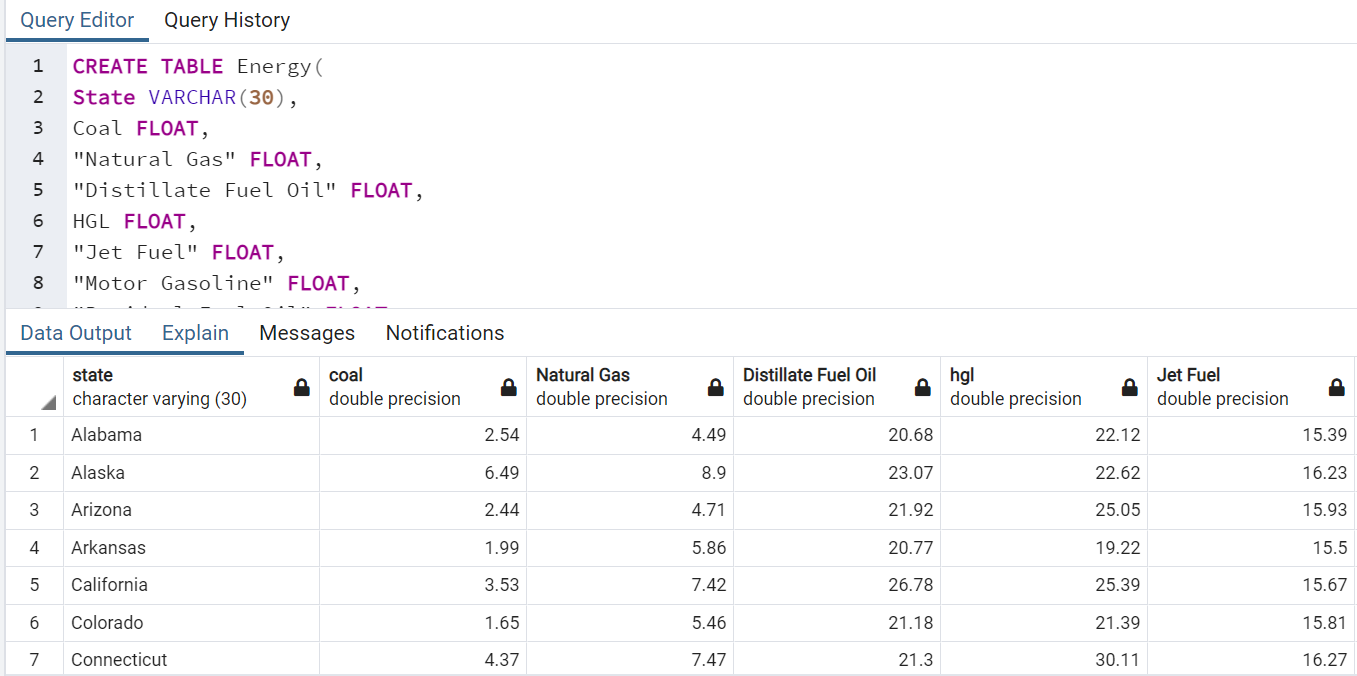
After we pulling the CSV files, we connected to the Postgres database in PG admin to store our original clean data sets. We used the query editor to create the initial table schema that got loaded into the Postgres database that generated the all of the tables. After running the queries, Postgres created the new tables with only the relevant information and proper formatting we reconnected to the database and generated additional tables for the data frames.

Postgres Database Screenshot:

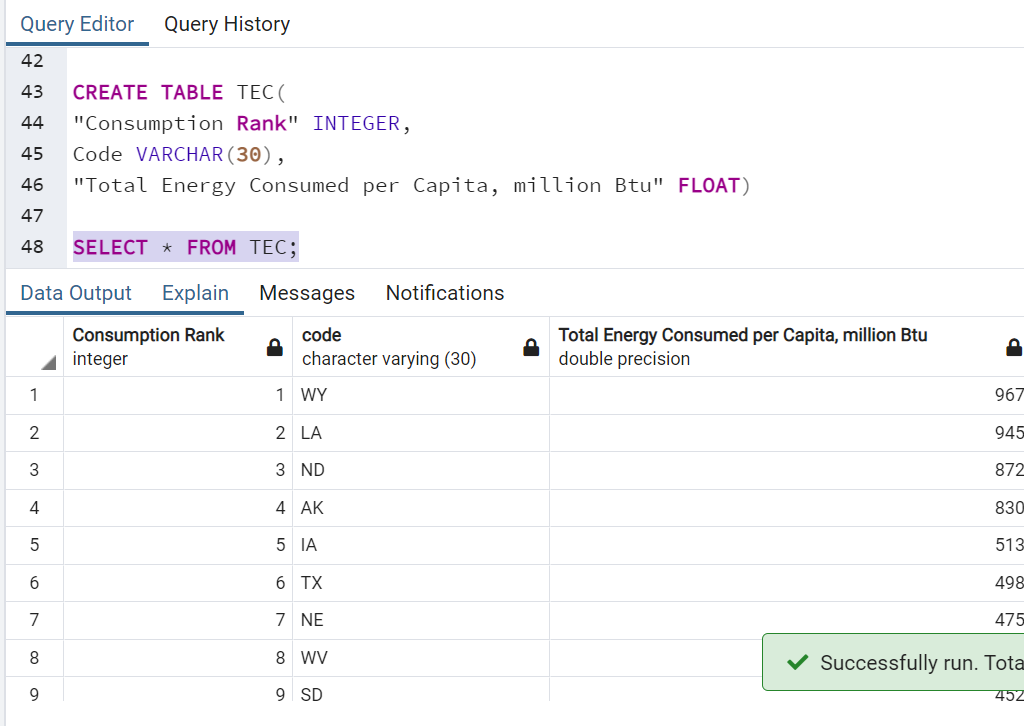
* **State Population Database**

****

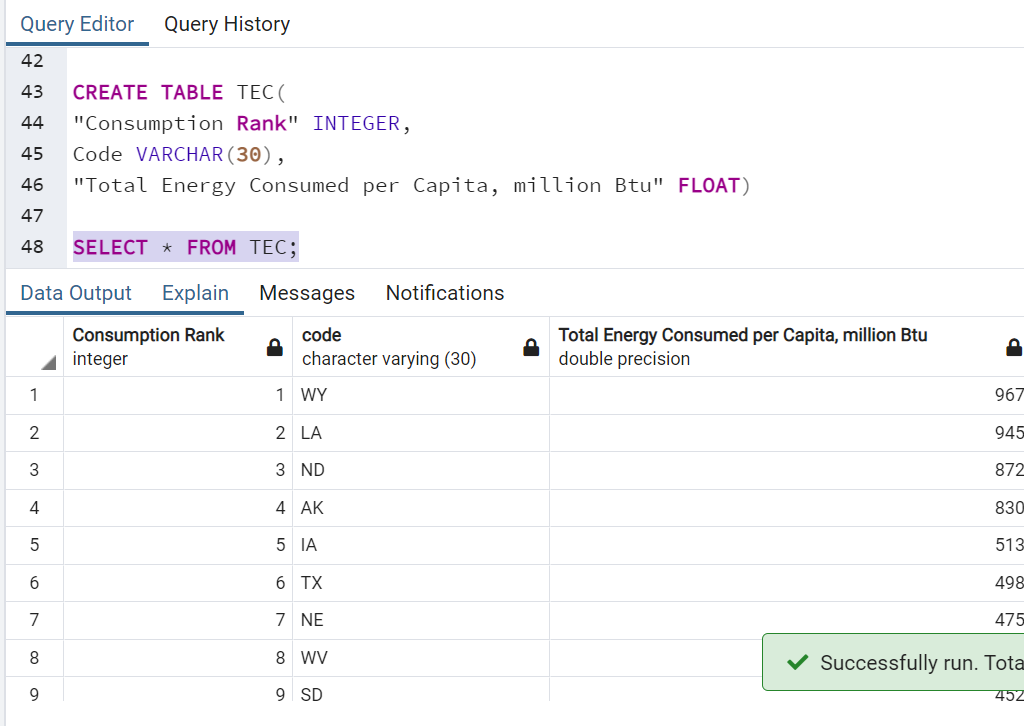
* **Energy Type Database**



* **State Energy Ranking Database**



* **Total Energy Consumption Database**

****

Summary/User Guide

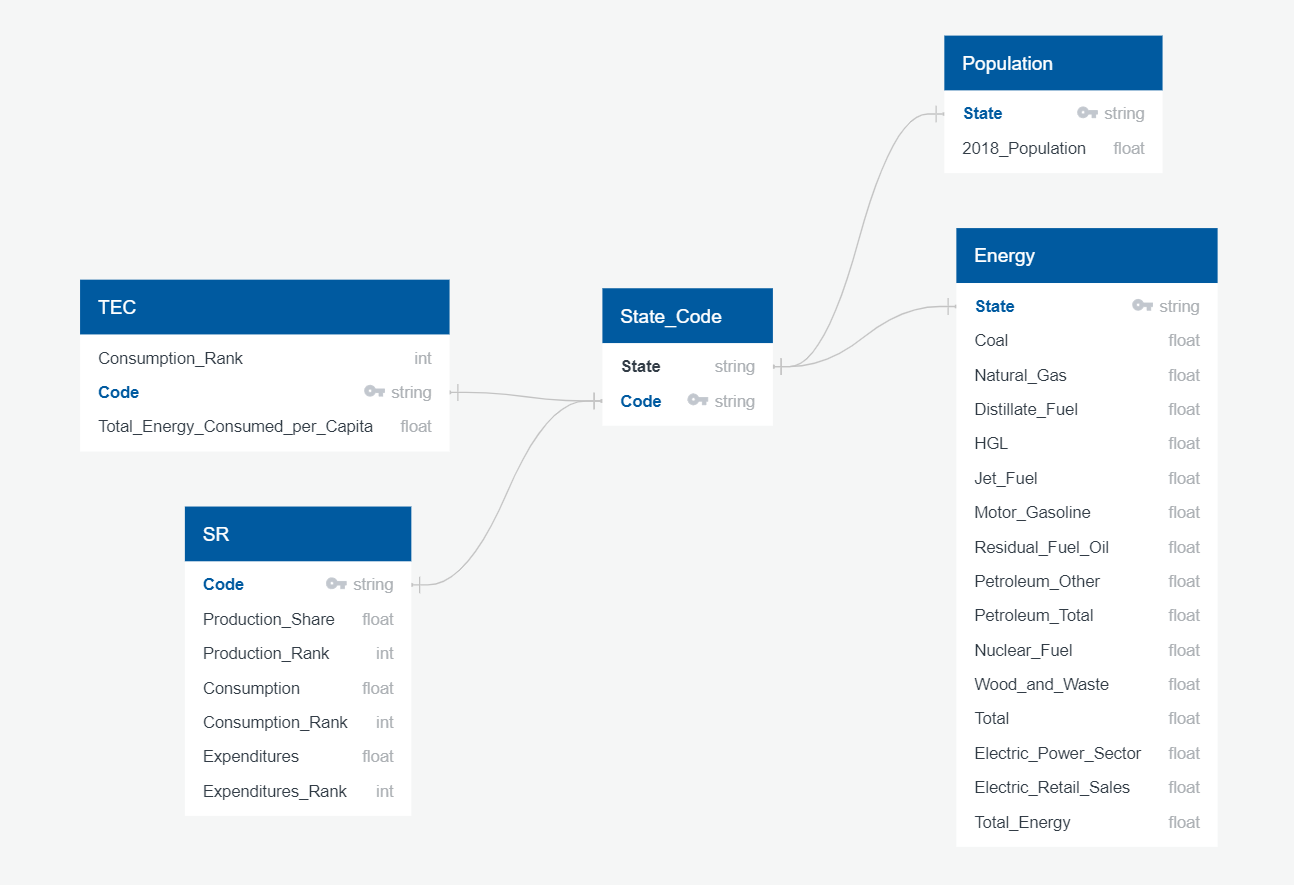
There were some limitations to our findings due to the data available. One limitation was our main energy was from 2018. However, we were able to provide information for our purpose.

\*Please note that all energy values are reported as Million BTU (British Thermal Unit). \*

Each table can be viewed in the Postgres database individually based on topic and data source such as Total Energy Consumption (TEC) or Energy by State (Energy). We have also created an all-inclusive summary table for the analyst where they can view all the data by state. This is additional table will provide all of the data provided by each database in one complete table. (Please see Finding Example section to view the best and worst states based on their energy production rank)

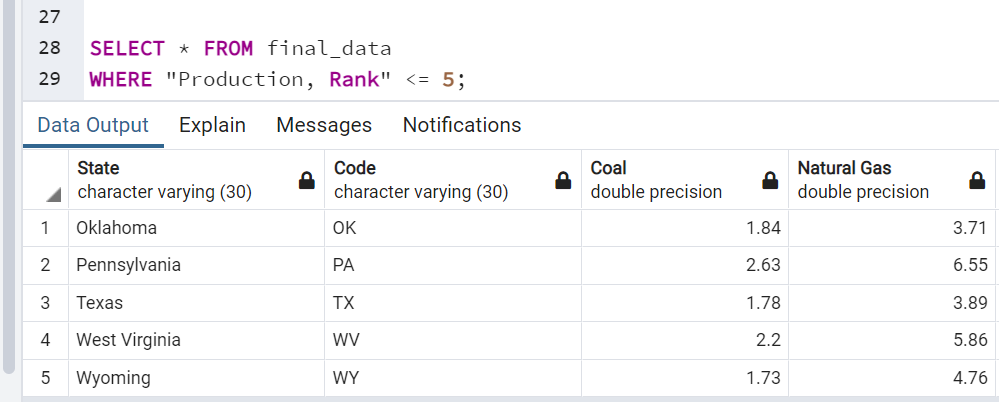
**ERD:**

**ERD is included in this report for the analyst for the preservation of the sourcing of the data.**

****

**Finding Examples:**

**Top 5: This image presents a sample query of the aggregate table and presents the top five states. (uses the database that can parse ALL data by State)**



**Worst 5: This image presents a sample query of the aggregate table and presents the worst five states. (uses the database that can parse ALL data by State)**

