.

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 the operator’s column and the address column due to missing information which was not relevant to the focus of this study.
* Identified duplicates by doing an inner merge on the incident id column across all three data sets.
* Created queries to address our hypothesis by grouping the data by state and getting the sum of the number of people killed and the number of people injured. We sorted the data in descending order so we could visually see which state had the highest numbers.

**Accidental Death**

A screenshot of a cell phone

Description automatically generated

**Accidental Death Injury**

A screenshot of a cell phone

Description automatically generated

**Mass Shooting**

A screenshot of a cell phone

Description automatically generated

**Load**

After we pulled in the CSV files and loaded them into the data frames, we did an initial connection to the Postgres database using PG admin to store our original clean data sets. We used the quick database website to create the initial table schema that got loaded into the Postgres database that generated the first set of tables. After running the queries and created the new tables with only the relevant information we reconnected to the database and generated additional tables for the data frames.

**Postgres Database:**

**A screenshot of a computer

Description automatically generated**

**Mass Shootings**

A screenshot of a cell phone

Description automatically generated

**Accidental Injuries**

A screenshot of a cell phone

Description automatically generated

**Accidental Deaths**

A screenshot of a cell phone

Description automatically generated

* *See Jupyter notebook for Gun\_Violence-Final.ipynb for more information*

**Summary**

There were some limitations to our findings due to the data available. However, we were able to address our hypothesis question in our initial project proposal below:

H0: There is no difference in deaths of mass shootings as compared to accidental shootings.

* Accidental deaths
* Mass shootings

We took some additional steps by specifically investigating the finding of mass shootings /accidental shootings by state. We did this to help inform our “Campaign Team” to determine our candidates position on gun control issues in the United States. In conclusion, we found that deaths resulting from mass shootings outnumbered deaths resulting from accidental shootings. Thus, we recommend the following:

More gun control in the following states while these states have the highest number killed by mass shootings.

1. California
2. Texas
3. Illinois
4. Pennsylvania
5. Ohio

We recommend more gun education/training in the following states Which involved states with the highest number of deaths from accidental shootings.

1. Texas
2. Georgia
3. Alabama
4. Ohio
5. Mississippi