**ETL Project – 5/26/2020**

**Fareedah Sabree Pete**

**Incarceration data in the U.S.**

This project takes a look at crime and incarceration statistics by U.S. state. The data takes a look at the number incarcerated from 2001-2016. The data has a breakdown by crime commited, age (17 and below), and non-Americans by state and year.

**Extract**

I used 3 datasets from two sources:

1. [www.kaggle.com](https://www.kaggle.com/christophercorrea/prisoners-and-crime-in-united-states)
   1. Crime and Incarceration in the United States (.csv file)
2. [Bureau of Justice Statistics](https://bjs.gov/)
   1. Number of non-U.S. citizens held in custody in federal or state prisons, December 31, 1998-2016 (.xlsx file)
   2. Number of inmates age 17 or younger held in custody in federal or state prisons, December 31, 2000-2016 (.xlsx file)

I leveraged the Panda functions, **read\_csv** and **read\_excel** to read the data from the csv and excel files. Once completed, I was able to transform the data.

**Transform**

**File 1 -** **Crime and Incarceration in the United States**

**Steps to Clean the data**

* Updated the dataframe to only include columns that were needed.
* Renamed column, ‘jurisdiction’ to “Jurisdiction”, to match the spelling in the other 2 tables in order to join the tables if needed.
* Set the Jurisdiction as the index on the table.

**File 2 – Less than 18 data set**

**Steps to Clean the data**

This data set required a little more cleaning as it was data directly from the BJS site with minimal changes.

* This table was a wide table with ‘Year’ as the columns. Update the table using the **melt** function to make it long with fewer columns. I thought that this would be a better way to display the data where users would know what was being captured. The table went from 18 columns to three column: Jurisdiciton, year, and lessthan\_18\_count.
* Converted all the Jurisdiction values to Upper Case to match the Crime and Incarceration in the United States file.
* Set Jurisdiction as the index on the table.
* Renamed the index values for the Jurisdictions. Some values had extra characters that needed removal in order to join the tables. Example ALASKA/B
* Removed extraneous columns like the ‘US Total’ and ‘State’ as these could be derived by doing a count of the records in the table.
* Removed special characters in the lessthan\_18\_count column. I replaced these with null values as the special characters represented that data was not provided. I wanted to distinguish this from a zero count.

**File 3 – Non-citizen data set**

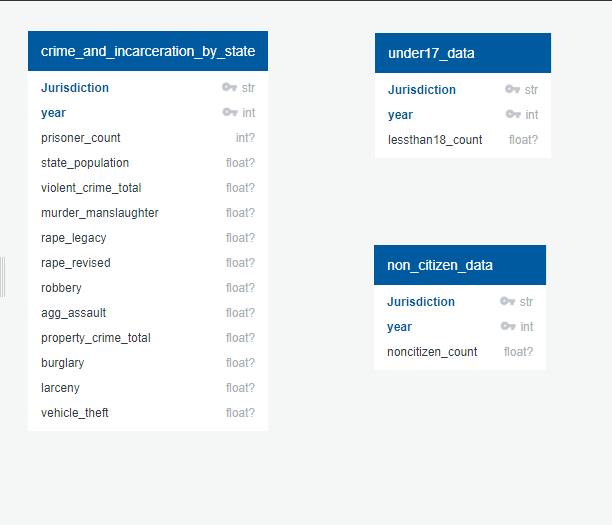
**Steps to Clean the data**

This data set was similar to the **Less than 18 data set** and required the same type of cleaning as this data was also directly from the BJS site with minimal changes

* This table was a wide table with ‘Year’ as the columns. Updated the table using the **melt** function to make it long with fewer columns. Three columns were created Jurisdiciton, year, and noncitizen\_count.
* I renamed one column, 2016/a, to remove the special characters.
* Converted all the Jurisdiction values to Upper Case to match the Crime and Incarceration in the United States file.
* Set Jurisdiction as the index on the table.
* Renamed the index values for the Jurisdictions. Some values had extra characters that needed to be removed.
* Removed extraneous columns like the ‘US Total’ and ‘State’ as these could be derived by doing a count of the records in the table.
* Removed special characters in the noncitizen\_count column. I replaced these with null values as the special characters represented that data was not provided. I wanted to distinguish this from a zero count.

**Load**

1. Created relational database in Postgresql called us\_incarceration\_db.
2. Leveraged the <https://www.quickdatabasediagrams.com/> site to create data models and exported the scripts to “CREATE TABLE” in Postgresql



1. Three tables were created using [these](https://github.com/fsabree/ETL-project/blob/master/Incarceration_db_scripts-export.sql) scripts:
   1. crime\_and\_incarceration\_by\_state
   2. under17\_data
   3. non\_citizen\_data
2. Once tables were created, created a database connection to the us\_incarceration\_db in postgres by using sqlalchemy **create\_engine** function.
3. Loaded the tables from the following dataframes created in Pandas
   1. crime\_and\_incarceration\_by\_state (incarceration\_transformed dataframe)
   2. under17\_data (lessthan18\_df\_transformed dataframe)
   3. non\_citizen\_data (noncitizen\_df\_transformed dataframe)
4. Completed a data check and pulled counts of all 3 tables to make sure that all data was loaded:

select count(\*)

from crime\_and\_incarceration\_by\_state;

--816 rows

select count(\*)

from under17\_data;

--832 rows

select count(\*)

from non\_citizen\_data;

--832 rows