**ETL Project Report**

**Team Members:**

* Kara Humanski
* Nick Prieto
* Helena Zhang

**Extract**

* **Data Sources:**

|  |  |  |
| --- | --- | --- |
| File Name | Source | Format |
| Census\_Data\_-\_Selected\_socioeconomic\_indicators\_in\_Chicago\_\_2008\_\_\_2012 | <https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2> | CSV |
| Gun\_Crimes | <https://data.cityofchicago.org/Public-Safety/Gun-Crimes/ta3m-92yk> | CSV |

* **Extracting Process：**

Both files were published data on the above website, we downloaded in CSV format, then we utilized Pandas to import and read.

**Transform**

**Description of the transform process:**

We created three dataframes below from the two CSV files and each contains information. The primary

key of the neighborhoods dataframe will be the community number, this will be the foreign key for both the census dataframe and the gun crimes dateframe.

* Neighborhoods\_df ( from The Census data): explicitly states the name of the neighborhood and each community number, we use it as a reference file later for foreign key.
* Economic\_indicator( from The Census data):
* Gun\_Crime( from gun crime data):

We transformed the data in Pandas (see detail action in the table below):

* Rename Column
* Drop N/A
* Selected Column
* Created dataframes
* Filtered the “Year” column to narrow periods to “2008-2012” to match the census data CSV file.
* Changed the datatype for several column to “Int” for easier SQL recognition.

**Transform Actions**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **File Name** | **Created Dataframes** | **Drop “N/A” Data** | **Rename**  **Columns** | **Col Data Type change** | **Selected Columns** | **Filtered Columns** |
| Census\_Data | Economic \_indicator\_df | Drop N/A data across different columns | Rename several columns to readable text , And add“-“ in the column names | Update the datatype to “int” for several columns | N/A | N/A |
| Census\_Data | Neighborhoods\_df | N/A | N/A |  | Only selected “ Community Area Number” and “Community Area Name” to used as reference file | N/A |
| Gun\_Crimes | Gun \_df | Drop N/A data across different columns | Rename all columns to remove spaces;  Rename the “ Community Area Number “ to match Neighborhoods\_df | Update the datatype to “int” for several columns | Selected relevant columns for potential analysis-“Community Area’, ‘Date', 'Block', 'Description', 'Location Description', 'Arrest', 'Domestic', 'Beat',  'District', 'Ward','Year' | Filter the “Year” to “2008 to 2012” period to match Census report period information |

**Load**

We used sqlalchemy to connect from Jupyter notebook to the local PgAdmin server and inserted the three dataframes as tables into our “gun\_violence” database.

Once the data was inserted, we used sql query tool to add a primary key to each table and make the “community\_area\_number” column in each table as the foreign key which likes them. This will allow for joins to be used for extracting data from the “gun\_violence” database.

We performed the join action between “economic\_indicator” table and “gun\_crime” table to exam the foreign key and each primary key were successfully created and linked.