**Data Clean Up**

Data Sets:

* The initial idea was to the Landmarks data set to make a map and view what parts of Chicago are denser in landmarks. Then, the Murals data set was added to the mix.
* Looking at both CSV’s, both had very similar data that could be merged into like columns with a few header name changes.
* An additional column was manually added to both CSV, via Excel, listing the type as a Landmark or Mural to be able to map either or in the future.

Data manipulation for both data frames

* Rename like columns to match up and merge data sets in those columns
  + landmark\_db:
    - landmark name > name
    - architect > artist
    - date built > year installed
  + murals\_db
    - artist credit > artist
    - zip > zip code
    - street address > address
    - artwork title > name
* Select only desired columns and drop the others that will not be used towards future maps.
  + Removed due to no functionality:
    - ID
    - Year Restored
    - Media
  + Removed due to information lacking in the other data set:
    - Affiliated Organization
    - Year Restored
  + Removed due to use of zip code or longitude and latitude:
    - Ward
    - Community Area Number
    - Location
  + Drop NaN values from certain columns, where the info was crucial:
    - Longitude
    - Latitude
    - Zip Code
* To have a clean set of dates to use for the maps and analysis, a new column was made with the extracted year. Then the original column was removed
  + Landmarks:
    - designation\_date need to have just the year extracted from mm/dd/yyyy format using the Datetime index
    - install\_date had a range of dates, and to get the first year listed, the first four digits from the column were extracted from the string
  + Murals:
  + year\_installed had the first 4 digits extract to remove the decimal place. It could not be converted to INT with the NaN values.
* Columns rearranged for functionality

Merging the two data frames:

* Both data frame column header names changed to lower case formatting
* Dtypes converted to merge certain columns
  + murals\_df[‘longitude’] converted STR to numeric
  + murals\_df[‘zip\_code’] converted STR to INT
* The final chicago\_df was done as an outer merge to include all remaining columns
* chicago\_df was merged on multiple columns:
  + name
  + address
  + artist
  + latitude
  + longitude
  + zip code
  + type
  + install date
  + designation date
  + description of artwork

Changes to the final merged data frame:

* All column headers converted to “snake\_case” for a smoother transition to SQL and future inquiries
* NaN values replaced with the word “unknown” or “none” in certain columns
  + name
  + address
  + artist
  + designation date
  + zip code
  + description of artwork
* Zip code converted to string per request for future analysis
* Column header names changed to lower case and “snake\_case” formatting for smoother transition to SQL

Postgres and SQL

* Table created in PgAdmin to receive the chicago\_df date frame
* The table was exported as a CSV
* SQL would not except any strings as VARCHAR
  + they were labeled at TEXT
* When exported as a CSV, the columns in the excel file were not lining up.
  + There were a few random symbols, commas, and “/n” being taken as delimiters
  + Symbols were removed, commas became semi-colons, and the “\n” were replaced with proper space.