**Team:** Ruben Benitez, Mariko Falke, Andy Peters

**Question:** Which regions in California have the highest homeless population rate and what opportunity do they have for shelter based on availability of beds within that region?

**Project Report: Extraction, Transformation, and Load**

**Extraction:** The following data files were extracted from these data sources:

Data from 2018 was used because this was the most complete recent data set.

HUD: <https://www.hud.gov/2019-point-in-time-estimates-of-homelessness-in-US>

* 2007-2019-PIT-Counts-by-CoC (.xlsx file)
  + This data table presents information regarding Point-In-Time (PIT) Counts as a result of conducting counts of the homeless population at various Continuum of Care Programs (COCs) across the nation.
* 2018-Housing-Inventory-Count-Raw-File (.xlsx file)
  + This data table presents information regarding housing (ie. how many beds) available at each COC across the nation.

California Association of Counties: <https://www.counties.org/data-and-research>

* population\_by\_jurisdiction\_and\_by\_county\_-\_1970\_to\_2018\_-\_09-07-2018 (.xlsx file)
  + This data table presents information regarding the total population of each county in California.

**Transform:** The following outlines the process in which the various data files were cleaned and transformed to include only the information necessary to answer our question:

**2007-2019-Point-In-Time(PIT)-Counts-by-CoC**

* Remove columns to narrow scope; Federal data was unnecessarily wide as they subdivided the data across columns. Our project was focused on high level numbers: Total population by county, homeless population by county, and beds available for the homeless
* Remove rows; focus on CA counties.
* Use the most recent data available across all datasets. 2018 data was the most recent year available in all data sets
* Heavy transformation was required on the CoC Name field to extract the county name form a field that is storing mixed data with mixed formatting.
* Reduce columns to focus on County population, homeless population by county, number of available beds.
* Reduce rows
  + There was a lot of homeless data form the 2010 census, but we wanted more recent data. 2018 data was the most recent data available in all the datasets
  + We focused our analysis on the state of California
* Resolve many-to-one between COC number and County Name
* Federal homeless reporting is based on a Continuum of Care number which has one or many counties. We had to resolve a many-to-one relationship to merge datasets form different sources.

**2018-Housing-Inventory-Count-Raw-File**

**population\_by\_jurisdiction\_and\_by\_county\_-\_1970\_to\_2018\_-\_09-07-2018**

* Using Pandas:
  + Two data sets were imported:
    - population\_by\_jurisdiction\_and\_by\_county\_-\_1970\_to\_2018\_-\_09-07-2018
      * To acquire information of how many population within each county
      * Upon importing the 'Population by County' sheet was selected to limit the information.
    - coc\_county\_xref
      * To acquire information of which county belongs to each COC.
  + coc\_county\_xref was updated to include headers for easy querying. Column 0 was renamed to ‘COC Number’ and Column 1 was renamed to ‘County’.
  + population\_by\_jurisdiction\_and\_by\_county\_-\_1970\_to\_2018\_-\_09-07-2018
    - First clean up was to select data from only the 2018 year.
    - Then, the resulting table was merged with the coc\_county\_xref table to bring in the county as they are associated with each COC.
    - Year was removed because that data was unnecessary since all the data is now from 2018.
    - Group.by was used to group by COC and collect population information collected from each associated county.
    - The resulting table was written into a .csv file.

**Load:** the final database, tables/collections, and why this was chosen.

* Both PostGres and Mongo db are viable choices for this project. SQL Relational Database Post Gres was chosen due for the following reasons

1) The size of the dataset is small is does require a big data platform like Mongo db.

2) The data structures are fixed and not changing. In other words, we did not need the data structure flexibility Mongo supports

3) We wanted to create an Entity Relationship diagram to show how the tables from different related to each other.

3) And practically speaking, the team had more experience with SQL and felt they could better support the short deadline by using PostGres.

* The final tables or collections that will be used in the production database.
  + CA homeless population by County
  + CA housing inventory for the homeless population
  + CA population by county