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

**Project Question:** What is the homeless population in CA by county and what opportunity do they have for shelter?

**Project Report**

At the end of the week, your team will submit a Final Report that describes the Extract, Transform, Load (ETL) process:

* **E**xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).
  + Data.World
  + Hud
  + https://www.counties.org/data-and-research
* **T**ransform: what data cleaning or transformation was required
  + 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.
* **L**oad: 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