**ETL Project**

Team Members: Heesung, Caitlin, Kathryn, William

**E**xtract:

* How many data sources?

4 datasets – 1x p/p

* What format is each data source in?

atleast 2 types e.g. JSON, CSV

* Where are the data sources from?

Kaggle

* What are our data validation rules?

Main rule – all connected via lon / lat

All datasets are Bay Area specific

* What data validation did we conduct / did the data pull from the sources correctly.

Transform:

* What type of transformation needed for each data source? Examples as follows –
  + Select certain columns to loads.
  + Translate code values (1 = Male | 2 = Female, convert to M = Male, F = Female)
  + Sorting
  + Joining
  + Aggregating
  + Generating surrogate keys
* Why did we decide to use these transformations?
* How did we perform each transformation e.g. challenges faced?

Load: