  

SEBAHAWKWOOD

Members:

- Ghassan **Seba** gseba7@gmail.com gseba
- Matthew **Hawk**ins hawkins.mattd@gmail.com matthew-david-hawkins
- Mike **Wood**all mike@savvysolutions.com savvysol

EXTRACT

The extract portion of this project was completed using five separate data sources, California Cities and Population, California Cities Crime Rate, Unemployment Rate California Cities, Median Home Price California Cities, and California Median Age by City. The data for California Median Age by City, California Cities and Population, and California Cities Crime Rate were extracted using a Pandas Web Scrape. Initial edits utilized a Jupyter Notebook and various Pandas functions, i.e. "`str.split`", "`df.drop`", "`df.set_index`", "`df.rename`", `etc.`. The Unemployment Rate California Cities and Median Home Price California Cities were CSV downloads that needed minimal editing to maintain data integrity.

TRANSFORM

Inputs to the transformation code were a set of five Pandas dataframes, one for each data source. The dataframes were "inner" merged on the city name to create a set of records with consistent data. Cleaning of this data required text manipulation to strip extra characters off the city name fields of each source data frame and casting numpy datatypes to python datatypes for easy handling with pymongo. The output of the transformation code was a list of dictionaries. Organizing the data into a list of dictionaries required an iteration through each row of the merged dataframe.

LOAD

On the heels of our homework, our team chose to use a MongoDB to store our information. We liked Mongo because it is quick to set up and inserting information as a dictionary is simple.

The major drawback to this format was querying the data was not as straightforward as a SQL select statement. But overall we were pleased with our decision.