ETL Project – Overall Process

1. Project Proposal
   1. Gathering real estate and crime data from multiple sites.
2. Finding Data
   1. We used 2 datasets from Kaggle and 1 dataset from data.world.
3. Project Report
   1. Extract:
      1. The original data was imported from the Kaggle.com website.
      2. We used Pandas to import a file called: population\_by\_zip\_2010 and made it a DataFrame.
      3. We used Pandas to import a file called: RDC\_InventoryCoreMetrics\_Zip\_Hist and made it a DataFrame.
      4. Both datasets were in zipped form as a csv.
   2. Transform:
      1. We used Pandas to Transform the data
      2. For the population file
         1. We dropped columns, geo\_id and renamed zip code columns
         2. We created a DataFrame
         3. Converted the DataFrame into a list of dictionaries
      3. For the Inventory file,
         1. We dropped several columns and only kept the columns we needed
         2. We did a group by to keep consolidate the zip codes
      4. Consolidated files
         1. We merged the data and grouped by zip code
         2. We got a count of the data
   3. Load:
      1. We created a connection to our mongo DB (relational)
      2. Defined the collection
      3. We pushed the records into Mogo DB.
      4. The data is uploaded to GitHub.
4. Created an Entity Relationship Diagram (ERD) to show the relationship of the data.