Extract, Transform, Load Project

TEAM MEMBERS: Abdurrahman Darvesh, Alex Burch, Roderick Reynolds, Rodney Kirkendoll, Serena Leung

PROJECT DESCRIPTION: Our client is a real estate developer interested in figuring out which market to invest in that will have the highest growth potential. We will look at population growth over the course of five years, income, and poverty levels.

We pulled these particular variables because linking the Census data together with the median per square foot price will help use determine which zip code area has had the most growth in price historically and which area had the highest population growth and income growth.

Extraction:

Our data came from the United States Census Bureau and Zillow. We downloaded each year of census data through the Census API into a csv file. The Zillow data was downloaded from the Zillow website’s data tab. We filtered the data we wanted by using the search criteria of median value per square foot of all homes by zip code across 2012 to 2017. This was exported into a csv file.

We extracted columns from the Census Data that will help us understand what are the trends of population change, income change, and poverty in the last 5 years. The seven columns we extracted include population, median age, household income, per capita income, poverty count, and poverty rate. We extracted data for 6 years, each year has 33120 rows. The total rows extracted from census data was 198,720.

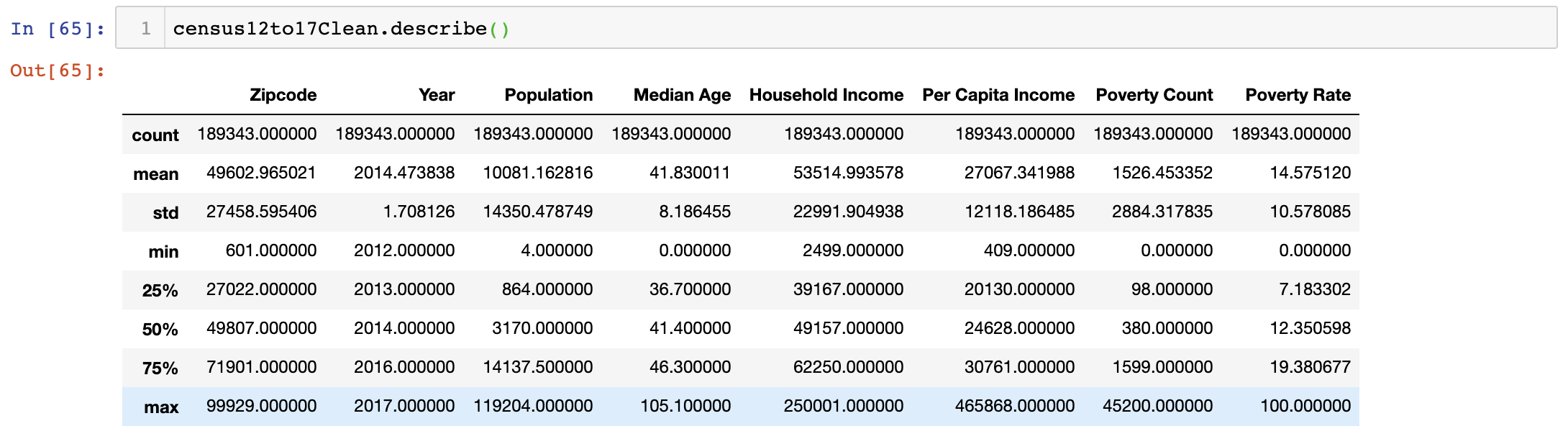
We extracted columns from the Zillow data that will help us look at prices over time by zip code. From the Zillow Data, we extracted 287 columns, 15,002 rows. The data was sorted by rank of population in a zip code. Out of the top 15 cities, six zip codes belonged to New York City, and three belonged to Chicago.

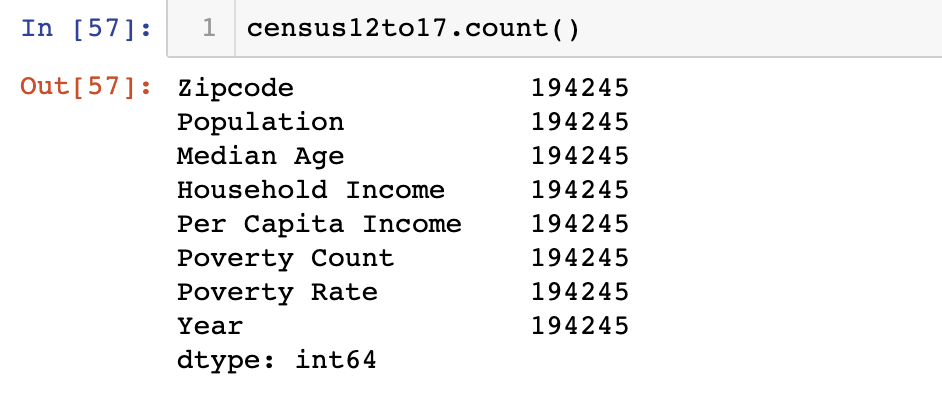
Transformation: what data cleaning or transformation was required.

**United States Census Bureau** (<https://api.census.gov/data.html>, <https://github.com/CommerceDataService/census-wrapper>)

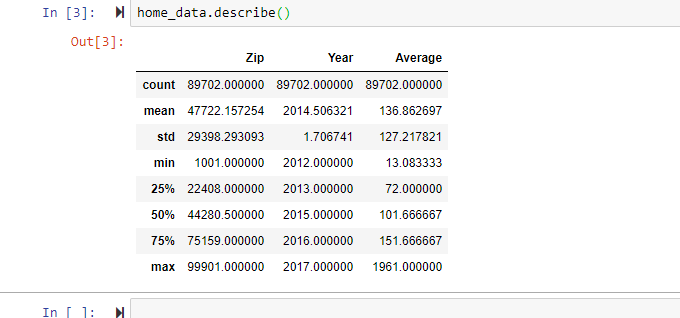
We added a year field for each corresponding year and merged the six years of Census data vertically using the merge function from the Python Pandas library. We reviewed descriptive statistics and counts to look for anomalies in the data. We found that there were nulls and negative values for Age, income, and per capita income. We removed any rows with nulls using the dropna() function. We also removed the negative values Median Age, Household Income, and Per Capita Income, which were removed. There were 4905 rows that had -666666666 value, which was removed from the dataset.

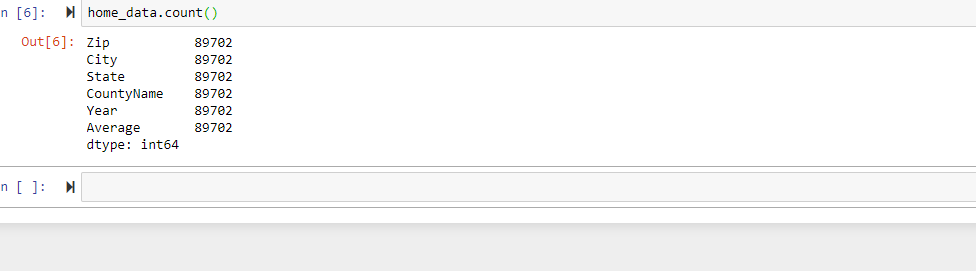
Here are the descriptive statistics of our cleaned Census Data:





Here is the descriptive statistics of our cleaned Zillow Data:

****



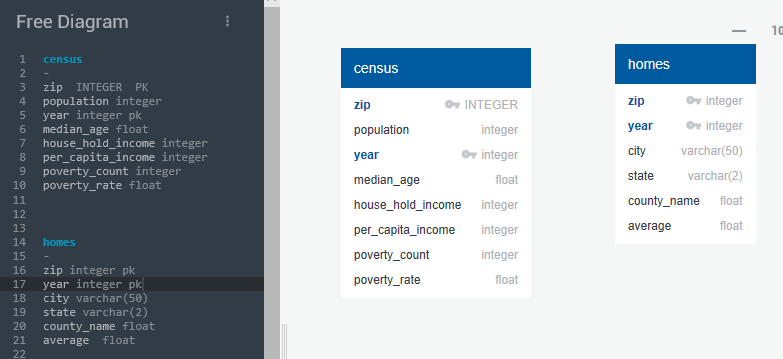
**Zillow Data** (<https://www.zillow.com/research/data/>):

The Zillow data is by month so we have to aggregate this by year and zip code. We calculated averages for each zip code by year. We reformatted the table to have columns that have stacked year and average.

The Census and Zillow can be merged together using the zip code and year columns.

We created the schema using the quick database diagrams website (<https://app.quickdatabasediagrams.com/#/d/AeajAV>).

The schemas we set up have two primary keys: zip code and year.



In the Census Data, we looked population, median age, household income, per capita income, poverty count, and poverty rate.

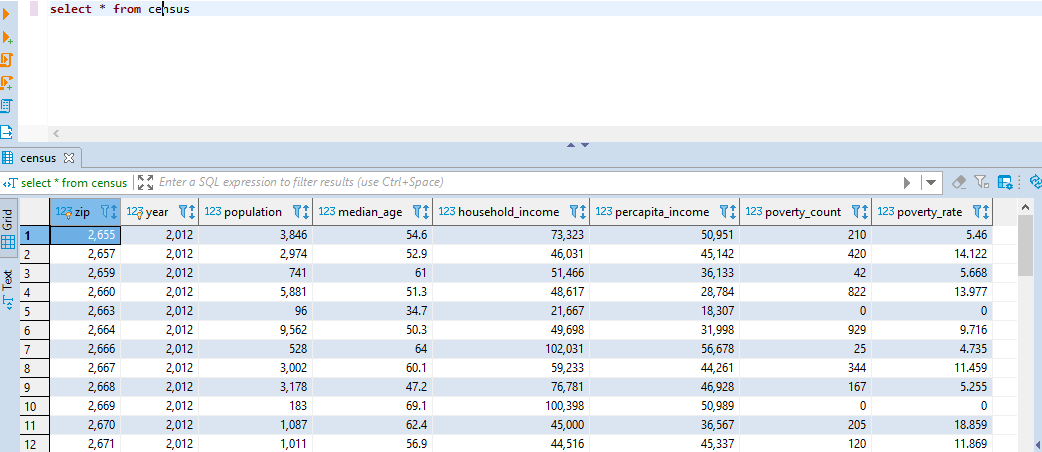
In the Zillow data, we looked at average of the median price per square foot by county, city, state, by year and zip.

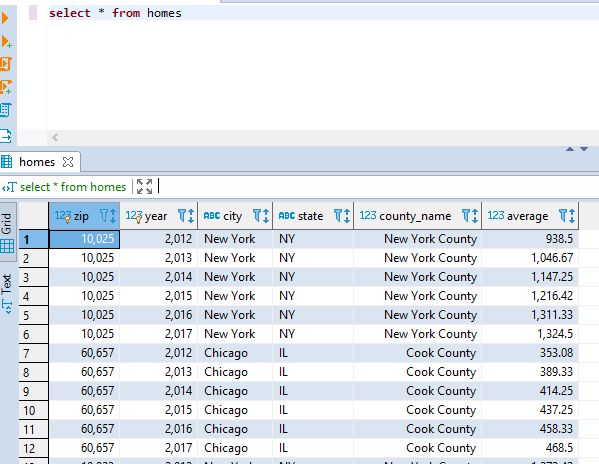
Load: the final database, tables/collections, and why this was chosen.

We imported the datasets into DBeaver from csv files. We then ran some queries to verify and test data. We chose SQLite as out type of database. We chose to use DBeaver because it is easy to use and view the data. We are able to query and link the data using simple SQL queries.

We chose this database structure because the two tables have two keys that allows us to merge the data and perform the analysis.

Screenshots of loaded tables below:





The database can be accessed via Github: <https://github.com/reynoldsroderick/ETL-Project>

.

To perform the analysis, our client can pull the data using the query below:

select census.zip

, census.year

, census.household\_income

, census.population

, homes.city

, homes.state

, homes.county\_name

, homes.average

from census

JOIN  homes

on (homes.zip = census.zip)

Using this data, our client can plot a line chart of population and income change over the last 5 years. Our client can also look at percentage change in income on a year over year basis to see which zip code had the highest percentage growth.