ETL PROCESS DESCRIPTION

We began our ETL project with two distinct databases, or collections, downloaded from Kaggle. The first is the Zillow median home prices data set CSV, which includes geographic information broken down to the town level, and price information spanning from April 1996 to August of 2016. This is interesting, as it provides a good twenty year time span that could potentially be useful for analysis, as it is time series data and panel data, for respective regions.

The second dataset CSV we downloaded was one on entrepreneurship and start up companies from Crunchbase. It includes geographic information as well, so we can match it to the Zillow dataset, as well as target market of the company, funding total, category, and comprehensive funding status information iterating through various rounds.

We began the ETL process by downloading these datasets, and then reading them into dataframes via jupyter notebooks, using pandas. Once this was completed, we selected the relevant columns that we wanted to retain from the very comprehensive datasets. This comprised the cleaning of the data, and included things like dropping the international entries for the startup data, and non relevant years from the Zillow data, as well as editing observation entries to have the cleanest and most intuitive form, for instance removing vertical lines from the category column of the startup dataset. For the investment table, we have the columns name, category list, market, funding total in USD, status, country code, state code, region, city, and number of funding rounds. For the Zillow database we included region name, state, metro (city), size rank, and the median house value for the year and month August of 2016.

We then connected these datasets to postgres SQL by utilizing SQLalchemy within the jupyter notebook, having predefined the database columns for the columns we want to keep, in SQL. We then double checked for the tables to make sure they were there, and then used pandas to upload the CSV modified into dataframes, into SQL databases. We then confirmed that this was successful by querying the SQL databases and returning the first few rows. This marked the completion of the Loading aspect of the ETL project.

ANALYSIS

Correlation between median home value and start up funding level? Is it the case that wealthier areas are more amenable to higher funding levels? Or is it the case that there is room for good deals, where home prices are reasonable and more resources can be devoted to startups?

TABLES

### investment\_data

Name

category list

Market

funding total in USD

Status

country code

state code

Region

City

number of funding rounds

### home\_price\_data

region name

State

metro (city)

size rank

median house value for the year and month August of 2016.