ETL Project Proposal

Kabrina Ramnath

# Data Sets

* Building permits by state (and metro areas) by year
* New privately owned housing units completed by purpose and design (by region)
* Average sales price of single-family houses sold in 2005 compared with that of houses actually sold
* Zillow API

**Potential Useful Investigations**

* Review building trends by volume
* Review completion trends
* Review the relationship between permitting/ownership/sales price
* Use Zillow to look at changes in pricing- may select one major state/metropolitan area for the scope for this project. New construction and older homes information available.

**Type of Database**

* Relational database
* This type of database would be ideal as I am trying to define the impact that each of these variables have on the other.

ETL Final Project

Kabrina Ramnath

* Data Sources
  + https://www.census.gov/construction/bps/uspermits.html
  + <https://www.census.gov/construction/bps/xls/table1c_cust.xls>
  + <https://www.census.gov/construction/nrc/xls/permits_cust.xls>
  + <https://www.nahb.org/-/media/Sites/NAHB/research/housing-economics/construction-statistics/national/starts-and-permits/single-family-starts-121918.ashx?la=en&hash=6B5EA4A6B701A6A9330850BC590BBBA4EADF4B03>
  + <https://www.census.gov/construction/nrc/xls/quarterly_starts_completions_cust.xls>
  + https://www.zillow.com/tx/home-values/
  + <https://www.zillow.com/market-report/02-19/54/tx.xls?rt=14>
  + Zillow Webscraping with Python script
* Detailing the process of the extraction, transformation, and loading steps
  + Converted excel files from websites into CSVs
  + Used Pandas/Python to create dataframes and merge tables to see applicable data side by side
  + Per the proposal, I intended to webscrape Zillow.com (API does not have listings) in order to create a database of prices based on zip code to pull averages. Initially trying to use beautiful soup, I ran into the issue where I needed to get through a catchpha. The code that I used does not pull pricing for only certain zip codes but needs a specific address. With that, I could not move further to create a general database but the code exists for when a database needs to be created for a more specific area.
* What data sources you chose, and why
  + Used the Census Bureau as a main source. For this type of information, public websites typically require payment.
  + Used Zillow as there are no limitations. All listings are posted regardless of selling firm, location etc.
* Explication why you have performed the types of transformations you did
  + I combined a lot of the data from the census bureaus to have them side by side in order to better read trends. For example, I merged the tables showing the percentage change in property value and the percentage change in building permits acquired for Texas metropolitan areas so easily see if the supply of building permits would have an effect on property value (simple supply and demand).
* Why you chose the type of the final database
  + I went with the SQL database as I taking multiple types of statistics related to this subject to try to establish a trend.
* Hypothetical use cases for your database
  + Track changes in pricing for zip codes
  + Homebuilders to assess where new business is viable
  + Private businesses to track new residential areas to evaluate potential new business
    - E.g. building a grocery near a newly established residential development.