**ETL Project Report**

*Group 6:* *Mahesh, Shuo, Alfredo, and Dan*

1. **Overview**

For our ETL project we elected to compile a database capable of providing insight into mortgages and home prices in the U.S. To do so, we focused on U.S. single family home prices and Federal mortgage data for the years 2016 and 2018. We extracted, transformed, and loaded the data into a production-ready relational database.

1. **Data Sources**

* [**Zillow Home Value Index (ZHVI)**](https://www.zillow.com/research/data/): A smoothed, seasonally adjusted measure of the typical home value and market changes across a given region and housing type. Specifically, we chose Zillow’s ZHVI single-family residences data ($, typical value for all single-family homes in a given region); aggregating prices by zip code and calculating averages for the (2) years analyzed.
* [**Public Use Database - Federal Home Loan Bank System**](https://www.fhfa.gov/DataTools/Downloads/Pages/Public-Use-Databases.aspx): The FHLB makes available to the public, in a form that is useful to the public (including forms accessible electronically), and to the extent practicable, census tract level data relating to mortgages purchased by each Federal Home Loan Bank. Such data is widely considered to be a benchmark for nationwide mortgage origination activity, commonly used on the Street. Similar to the Zillow data and for consistency in our eventual database, we elected to focus solely on Single-Family mortgages.
* **US Zip codes CSV:** Downloaded publicly available US zip csv from the internet. This csv contains ZIP code, against city, and county FIPS. The FHLB data used county FIPS (rather than ZIP code) and one county FIP can have multiple ZIP codes. As such, the ZIP code csv attached within our repository helped us to obtain ZIP code by way of FIPS + City.

1. **Data Extraction**

As noted above, our data was extracted from the web and in all cases (Zillow, FHLB and US Zip code csv) via publicly available CSV files.

1. **Data Transformation**
2. Downloaded CSV files and used Python (Panda’s package) to read CSVs into Pandas DataFrames
3. Used pandas to clean and transform our data:
   1. **Zillow data**:
      * Dropped unnecessary columns (i.e. our data set when back to ~1996; some of this was performed in the CSV)
      * Once we had the years we wanted, we calculated average home prices for each year (and added AvgHomeVal column). The data was previously broken into column months, but since the data itself was already seasonally adjusted by Zillow, we decided to calculate a simple average of the 12-monthly ZHVI figures.
      * Re-format ZIP codes. Since ZIP codes were automatically recognized by pandas as integers, the leading zeros were removed. So, understanding ZIP codes would be the basis of our merging down the road, we used a function to add the leading zeros back to the ZIP code column of our data 🡪 *.apply(lambda x: x.zfill(5))*
      * Grouped data by ZIP code and dropped unnecessary columns
      * Export DataFrame to clean CSV file for loading to pgAdmin
   2. **FHLB data**:
      * Selected only columns useful in comparing with Zillow data
      * Using pandas, converted “FIPSCountyCode” and “FIPSStateCode” into string values and combined them to make “stcountyfp” and converted the value back to integer
      * Changed all column headings from upper case to lower case to utilize in queries once loaded to Postgresql
   3. **US ZIP codes**:
      * Selected only columns which we need to get zip and city by using “stcountyfp” field from FHLB data
4. **Data Loading**

We used Pandas and pgAdmin to establish, build and load our data into the database.

* 1. **Database creation**: Using pgAdmin, we created our database.
  2. **Table creation**: We loaded our clean CSV files into Pandas DataFrames and used the SQLalchemy engine to load the DataFrames into pgAdmin tables.
  3. **Queries and Table joining**: Using SQL in pgAdmin we used joins to merge the Zillow tables (2016 and 2018) and then an inner join to merge the Zillow result with the FHLB mortgage data. In order to join the Zillow data with FHLB data we elected to use an inner join on Zip code as we understood that neither data set was exactly the same (i.e. Zillow did not have data on certain ZIP codes and vice-versa).

Final Tables: ***single\_fam\_data\_master*** *(*pgAdmin*),* ***single\_fam\_data\_master2016-18.csv***(located in CleanData directory)

* This table exists within our production (relational) database, *housing\_db* in pgAdmin. We also exported the database table as a CSV for easy loading and analysis in Python.