ETL Project

For this project, our team extracted data from Lending Club, US Census, and Zillow. The Lending Club data was a SQLite file from [Kaggle.com](https://www.kaggle.com/wendykan/lending-club-loan-data#database.sqlite). The US Census and Zillow data were csv files obtained from [Kaggle.com](https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations). and [Zillow.com](https://www.zillow.com/research/data/) respectively.

We cleaned the data from each source using Jupyter Notebook and Pandas, before putting the data into a relational SQL database, called ZipCodes, via pgAdmin4. The tables in the DB are home\_values, income\_by\_zip, loan\_data, and loan\_information.

**Data Sources**

Lending Club Loan Data SQLite File

<https://www.kaggle.com/wendykan/lending-club-loan-data#database.sqlite>

US Household Income Statistics CSV File

<https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations>

Zillow Home Values Index (ZHVI) All Homes (SFR, Condo/Co-op) Time Series ($) Data Type with the Zip Code as the Geography (Zip\_Zhvi\_AllHomes.csv)

<https://www.zillow.com/research/data/>

**Transformation**

Zillow Real Estate Value Data: Source data set had the median real estate sale price by Zip Code for every month in the last 20 years along with a few other columns mostly indicating location. All entries before 2011 were dropped. An Abr\_Zip column was created - this simply extracted the first three numbers of the Zip\_Code column. The yearly averages were then computed for each year after and including 2011. The rest of the monthly median sale prices were then dropped. The resulting table, home\_values, that was loaded into the ZipCodes database has the following columns: Abr\_Zip, Zip\_Code, City, State, CountyName, Rank, Metro, 2011\_Current\_Average, 2012\_Current\_Average, 2013\_Current\_Average, 2014\_Current\_Average, 2015\_Current\_Average, 2016\_Current\_Average, 2017\_Current\_Average, 2018\_Current\_Average, 2019\_Current\_Average.

Kaggle Income Data: From the source data set, a dataframe was created that grouped the Median Incomes based on the Zip\_Code. A third column, Abr\_Zip, was created that extracted the first three digits from the Zip Code column. The income\_by\_zip table, with three columns - Abr\_Zip, Zip\_Code, Median\_Income - was loaded into the ZipCodes database.

Lending Club Data: The Source data had only had the first 3 numbers of the zip code. A table, loan\_data, grouped the numeric variables based on the abbreviated zip code. (The abbreviated zip code of this table, zip\_codes, serves as the primary key and a foreign key in the other tables for this database). The table has 9 columns: zip\_code (pK), loan\_amount, funded\_amnt, funded\_amnt\_inv, term, int\_rate, installment, annual\_inc, dti.

Another table kept the non numeric columns along with the abbreviated zip code. This table, loan\_information, has 11 columns: zip\_code, funded\_amnt\_inv, sub\_grade, home\_ownership, emp\_title, emp\_length, issue\_d, loan\_status, pymnt\_plan, purpose, addr\_state.

More information about the database can be found in Entity Relationship Diagram.png and Database Documentation.pdf