**Predicted\_Gentrification\_in\_California**

**Extract, Transform, and Load Overview**

The Extract, Transform, and Load process is when data is gathered from a source which is then altered and sent out to be used in the machine learning code. This process was very important when creating our dataset for our project. Since we did not have a proper dataset and the project demanded multiple dependencies. It was crucial to create a functional etl code so that the data would seemlessly be integrated in the AWS S3 bucket and the machine learning model.

**ETL Process**

In our code, we had three different ETL processes to convert the data we needed to create proper analysis and visualiztion.

The first file was the etl\_datasets.ipynb. In this file we gathered and created a dataframe from the USA.com datasets. We created a function that would read the html of the dataset and then transform it into a pandas dataframe. The datasets metrics we gathered were avg\_edu, house\_median\_value, median\_income, median\_rent, public\_transportation, total\_pop, and white\_pop. We gathered each datasets from the year 2000 and 2014 which came out to a total of 14 dataframses. From the dataframe the unneccessary columns would then be edited out. The zipcode/total population column was then split and the total population data was deleted. The dataframes were first uploaded as a csv file and everything was transferred locally but then our group integrated the AWS S3 Buckets so from here we changed the uploads to transfer the csv files on the cloud.

The second file then gathered each and every csv file from the cloud and transformed the dataset. We grabbed a total of 14 different datasets that were then converted and formatted to fit the machine learning model. We transformed the datasets by removing the special characters from the data values so that each value could be manipulated as a float rather than a string. The related dataframes were then merged to create new dataframes for each metric that would be measured: Average Education Index, Median House Price, Median Income, Median Rent, Public Transportation %, Total Population, and White Population %. For each new dataframe a new % change metric column was created by finding the difference. After the different datasets were then aggregated into two dataframes called economics and demoggraphics. The economics dataframe had the metrics pertaining to income and prices while the demogrpahics dataframe had metrics pertaining to details about the population. The resulting dataframes are then converted into a csv and transferred into the AWS S3 bucket.

The final ETL process is transferring the final ML predicted model from the AWS S3 bucket onto a local csv file which would then be used in the Tableau Notebook to create visualization model.