ETL Analysis

With this project giving us the liberty to choose almost any topic we wanted, we came to the decision to analyze the housing market in relation to crime occurring within the city. With us potentially having such a broad scope in teams if the data to be collected, we decided to limit our scope to only cities in Orange County, California.

The first step was looking into where we were going to begin pulling the data from. We had first looked into Kaggle, but soon realized that it may not be the best source for us seeing as how the housing csv's available were from many years prior and we wanted more current data sets. We decided that the best course of action would be to extract our own data differing websites. In order to pull the housing market data, we implemented what we had learned in class prior about web scraping, and used it to pull the house listing info from Compass.com. We had tried to pull information from other popular realtor websites but had difficulty in doing so due to the fact that they had set up pop up blockers to prevent people from trying to scrape their websites. But, Compass offered us the listings that we needed and had made it less difficult to pull the data from their site. For the crime statistics, we found that the FBI had a public API that was available for use. After requesting an API key we had to perform two separate API calls in order to extract all the data that we needed. The first API call was to be able to get the police department information, and the second was to be able to get a list of the different types of occurring crime with their frequency.

After extracting the data we then proceeded to transform it into a more usable formatting. For the housing data, seeing as it was taken for a website, we needed to transform it from an Html file into a usable string. Then we proceeded to remove any extra unwanted characters such as "$" and ", " to make the data easier to manipulate. After which we used a split function to separate the values of "bed & baths" into two separate cells, to better differentiate between the two. These manipulations made it easier to create a Pandas Dataframe and then export the information as a csv file. When it came to the crime data, we had two perform two separate API calls to pull the information. We dropped in unnecessary information such as the county, seeing as we were only looking in Orange County, and made sure to show the “Ori” which was a unique identifier per city. The data printed out was in a JSON formatting that then had to be turned into a string to be able to turn into Pandas DataFrames then csv flies.

After having extracted and transforming the data we then had to decide what we wanted to do with it. Since we had generated a few csv files we wanted to use SQL, seeing it’s a relational based data management tool. With all the information gathered a commonality among them all was the city. We were then able to create an ERD image to help us further visualize how our SQL tables were to look. Seeing as how we were primarily through python, we implemented the use of Sqlalchemy to help facilitate the transition between our databases to sql statements.