Group 9: Crime Statistics vs Property Value in NYC Boroughs
Team: Max Wilde, Monica Chen, Szeyeung Luk
Topic: Are higher crime rates indicative of lower property values in each NYC borough?

Introduction

The focus of this project is to organize a collection of data to study the relationship between property values and crime statistics in New York City. The first dataset records every building unit sold in New York City from September 2015 to September 2016, including the address, borough and neighborhood, sale price, and sale date. The second dataset records all reported crimes in New York City from 2014 to 2015. This dataset includes the borough where the crime was committed, a description of the offense, time and date of occurrence, and the precinct where the crime was committed. Using these datasets, we will merge the two sources to create a new dataframe that will help determine if the number of crimes committed is related to the average property value in each borough.

Data Sources

The two sources of data used in this project are both CSV files extracted from Kaggle. The link to both sources are included below:

1. NYC Property Sales: https://www.kaggle.com/new-york-city/nyc-property-sales
2. New York City Crimes: https://www.kaggle.com/adamschroeder/crimes-new-york-city

Step 1: Extract

The datasets listed above were both pulled into Jupyter Notebook as CSV's via the above kaggle links. Because both files are in CSV format, our team decided to use a structured/relational database to organize the data because the data source is already organized in tables containing rows and columns.

Step 2: Transform

After the data was then pulled into Jupyter Notebook using Pandas, we cleaned the datasets by editing down the tables to minimize the number of columns present. The housing data was consolidated from 22 columns to 7, while the crime data was taken from 24 columns to 8. This cleaning allowed the datasets to become significantly more manageable and remove data that was not pertinent to our question.

Next, we removed all rows where the sale price in the NYC Property Dataset was omitted from the data. This preprocessing step allows the user to directly analyze the data in SQL without having to perform the extra step of removing unnecessary rows in PostGres.

Finally, after reviewing the data in the sale price column in the NYC Property Dataset, we noticed that several sale prices were recorded as unrealistic values. For example, several

rows showed the sale price as under $10. We decided to set a filter to only include sale prices over $1000. Although $1000 is still a low estimate of a realistic property value in New York City, this step will eliminate any rows where the low sale price was entered erroneously as unrealistic values.

Finally, we merged the two datasets based on the borough as a common element. This allows the user to study trends in property values and crime statistics in relation to each New York City borough.

Step 3: Load

The merged dataset and the transformed versions of the NYC Property Sales and New York City Crimes Datasets were loaded into pgAdmin via a connection string that allowed the final dataset to be seen as a SQL table. The final merged table we loaded into SQL is a relational dataset that includes all the necessary data to address the topic of our project, including location and sale prices of all building units sold in New York City in a year and the crime location and description of the crime rates in New York City for the period of a year.