**Regression Modelling Methodology**

**Data Preparation**

The first part of our analysis consisted of importing the dataset that included data on property sales in New York City. After we uploaded the dataset, we conducted an exploratory data analysis to understand its organization, factors, and any problems that required attention.

During the exploratory data analysis, we faced one of the first hurdles with the discovery of missing values in multiple columns. To tackle this problem, we methodically pinpointed the variables with incomplete data and assessed the percentage of missing values in each instance. After careful deliberation, we made the choice to eliminate data points containing incomplete information, as they comprised less than 5% of the entire data set. This method enabled us to keep a large part of the data while reducing the effect of missing values on future analyses.

After addressing missing data, we focused on the distribution of numerical variables in the dataset. We noticed that many variables showed noticeable skewness, suggesting possible departures from normal distribution. To tackle this problem, we utilized Tukey's method for identifying and eliminating outliers. Our goal was to enhance the robustness of future analyses by improving the distributional properties of numeric variables through the identification and exclusion of outliers from the dataset.

Additionally, we examined the connections between variables using correlation analysis. This included computing correlation coefficients for pairs of numerical variables in order to evaluate the magnitude and orientation of their relationships. During this examination, we found multiple variables that showed strong positive relationships with each other, along with variables that had weaker or negative relationships. These results offered valuable information on potential factors that could predict real estate prices and guided the choice of variables for inclusion in regression analysis.

In addition, we analysed changes over time by graphing the time-based pattern of property values in New York City. This examination showed a rising pattern in mean sale prices over time, with variations in certain time frames. Through the visualization of time-based trends, we achieved a better comprehension of the fluctuations within the real estate market of New York City, pinpointing potential influences on property price fluctuations.

**Regression Modelling**

After completing thorough data preparation and exploratory analysis, we focused on the main objective of our research: using regression modelling to forecast real estate prices in New York City. The method we used involved carefully building linear regression models, using different predictor variables to understand the complex factors influencing real estate prices.

Leading our regression modelling was the incorporation of important predictor variables that were considered to have a significant impact on real estate prices. These variables included a wide range of factors, each providing valuable perspectives on the intricate fabric of the New York City real estate market. Included in these predictors were the quantity of housing units in a property, offering an insight into its size and ability to house residents. The categorization of taxes for properties at various points highlighted their financial status and legal ramifications, revealing insights into the larger economic and legal environments in which these properties function.

Furthermore, the year in which the building was constructed was identified as a crucial factor in predicting real estate values, giving a historical perspective on how they change over time. We aimed to capture temporal trends and identify any seasonal or cyclical patterns that could affect pricing dynamics by including the sale date of properties in our models. Moreover, factors like total area and land area were essential in evaluating the physical size and spatial characteristics of properties, enhancing our comprehension of their inherent value.

By conducting thorough regression analysis, we discovered significant statistical connections between these predictor variables and property prices, revealing the complex interaction of factors that influence pricing decisions in the real estate market of New York City. Nevertheless, even though our models were strong, the adjusted R-squared values suggested that there might be additional variability that was not accounted for, indicating the presence of hidden factors that were not included in our analysis.