**Predicting Median Housing Prices in New York City**

**Ledia Dobi**

**Conie O’Malley**

**ITEC 621 – Predictive Analytics**

**Final Project Narrative**

Table of Contents

[Introduction: 1](#_Toc192078801)

[Description of the Dataset: 2](#_Toc192078802)

[Data Preprocessing: 2](#_Toc192078803)

[Limitations 3](#_Toc192078804)

[Data Cleaning 3](#_Toc192078805)

[Final Dataset 4](#_Toc192078806)

[Descriptive Analytics, Inspection of Plots, Tests for OLS assumptions: 4](#_Toc192078807)

[Model Specification, Model Methods, and Analysis of Results: 5](#_Toc192078808)

[OLS Model 5](#_Toc192078809)

[WLS Model 5](#_Toc192078810)

[Bootstrap Aggregation 6](#_Toc192078811)

[Cross Validation 6](#_Toc192078812)

[Conclusion: 6](#_Toc192078813)

# Introduction:

In times of economic volatility, we aim to better understand fluctuations in the housing market. By analyzing social and economic predictors, we hope to evaluate which factors are significant in raising or lowering a property value. The real estate market is influenced by economic policy, interest rates, and local and federal social indicators. With this project, we aim to provide insights into market trends, with the eventual goal of arming homebuyers, policymakers, and investors with the proper tools and knowledge to navigate this highly dynamic market.

Our Business Question focuses on using these economic and social factors to predict the price of a single-family home. Our Analytics Question is as follows: In our study, we aim to understand the effect that various economic (federal interest rate, median home sale prices, and others) and social (crime rates, unemployment rates) predicators have on the value of a single-family property. Our analytics question revolves around a quantitative outcome (price) and can be analyzed using quantitative methods, as the data will be numeric.

# Description of the Dataset:

Our dataset analyzes the New York City housing market from January 2000 to December 2024, focusing on monthly trends. We integrated data from multiple sources, including:

* **Zillow** (<https://www.zillow.com/research/data>)**:**
  + Median home sale prices,
  + New constructions sales
  + Mean home values
  + Rental cost index
* **Federal Reserve Economic Data (FRED)** (<https://fred.stlouisfed.org/>)**:**
  + Federal interest rate
  + 15-year and 30-year mortgage rates,
  + NY and national median household income
  + Unemployment rates.
* **NYPD Crime Statistics** (<https://www.nyc.gov/site/nypd/stats/crime-statistics/historical.page>) (aggregated to the city level)**:** 
  + Misdemeanor offenses
  + Major felonies
  + Non-seven major felony offenses

# Data Preprocessing:

* Addressing missing values.
* Standardizing all data to a monthly frequency, using averaging for variables reported at less frequent intervals.
* Ensuring data consistency and accuracy.
* Selecting appropriate variables

Gathering data from a variety of sources allowed us to create a comprehensive analysis of the housing market trends in New York City. Combining additional non-financial data helps us understand the factors that influence housing prices and crime rates. This analysis can help policymakers, investors, and residents make informed decisions about their financial and housing situations.

## Limitations

While we acknowledge potential multicollinearity among these variables, they represent crucial factors influencing homebuyer decisions. This holistic approach allows us to investigate the complex interplay of economic, financial, and socioeconomic variables affecting the New York City housing market. We also acknowledge that we had to exclude certain potentially important variables strictly based on the number of data points that were available (rental index and new construction). We also had to transform some data from yearly and distribute to monthly, which causes artificial smoothing of the data and can affect the overall outcomes. We did our best to mitigate all of the issues throughout our modeling.

## Data Cleaning

Our data required extensive cleaning and formatting, despite the records being very clean.

* **Dates**: We had to manipulate the date columns for all data sets to assure that all date formats and dates we the same to facilitate data merging. Some data sets were set for the last day of the month and others for the first day of the month. We decided to use the first day of the month which required us to use the `lubridate` package to adjust dates within the same months. There were other data that were yearly, median household income and crime data, which we computed to monthly values and distributed those throughout the months of the year. While this does not create a perfect representation of the data since there isn't a way to capture a trend, it helps us in an overall time series analysis as opposed to discluding it based on its periodicity.
* **Dimensions**: Many of the data sets were wide data sets that we had to pivot to long data sets to assign variables to each column and have the “Date” as the joining column, once properly formatted.
* **Missing Values:** We removed any NA values from our data set. This led us to removing variables “New Construction Sales” and “Rental Cost Index” because there weren’t enough data points to thoroughly model. Removing these two variables allowed us to have over 300 observations per variable as opposed to about 70 if we were to retrain them.
* **Economic Crises**: We included dummy variables for the 2008 financial crisis and the COVID-19 pandemic. We inputted 0 for months that were not included in these crises and 1 if they were in these crises. We hope that this will capture some of the outside impacts on housing prices that would not be otherwise captured without the inclusion of dummy variables.
  + 2008 Financial Crisis: 2007-12-01 - 2009-06-01
  + COVID-19 Pandemic: 2020-03-01 - 2023-05-01

## Final Dataset

Once our data manipulation was complete, we joined the data set to give us a final count of 14 variables and 300 observations.

* Date
* mean\_sfr\_value - estimated mean value of homes
* median\_sale\_price - median sale price of homes in a specified month
* Fed\_Interest\_Rate - Federal Reserve Interest Rate
  + calculated by monthly average - $(week\_1 + ... + week\_n) / n$ for each month
* mortgage\_rate\_15\_year - average 15 year mortgage rate in a specified month
* mortgage\_rate\_30\_year - average 30 year mortgage rate in a specified month
* ny\_median\_hh\_income - median household income for residents of New York State
  + calculated as n/12
* national\_median\_hh\_income - median household income for residents of the United States
  + calculated as n/12
* non\_seven\_major\_felonies - non violent felony commissions in NYC
  + calculated as n/12
* major\_felonies - violent felony commissions in NYC
  + calculated as n/12
* misdemeanor\_offenses - misdemeanor commissions in NYC
  + calculated as n/12
* unemployment\_rate - national unemployemnt rate in a specified month
* housing\_crisis - 2008 financial crisis dummy variable
  + 0 = non-crisis, 1 = crisis
* covid\_pandemic - COVID-19 pandemic dummy variable
  + 0 = non-pandemic, 1 = pandemic

# Descriptive Analytics, Inspection of Plots, Tests for OLS assumptions:

Our first objective was to inspect the plots visually. Inspecting a qqplot and a histogram, we found clear signs of non-normality. Although the data generally followed the line in the qqplot, it did deviate at each tail end. The histogram was not extremely skewed, but clearly did skew to the right. Further, when inspecting the residuals vs predicted values and fitted vs residuals plot, we saw clear signs of time series correlation, as evidenced by the cyclical waves and patterns. We conducted a Durbin Watson test, and found heteroskedasticity, ass the statistic was 0.362, showing extreme positive correlation. This is to be expected with this data set, as it is time series data.

Running a ggpairs plot, we gain insights that support what we know from prior business knowledge. The correlation analysis reveals significant multicollinearity among several key economic and social indicators. Most notably, mortgage rates of different terms (15-year and 30-year) demonstrate extremely high correlation, as do the relationships between median home values and median household income. Interest rates show strong connections with multiple housing market indicators, suggesting their fundamental role in shaping housing affordability and accessibility.

# Model Specification, Model Methods, and Analysis of Results:

## OLS Model

Our first model was the full OLS model. After we ran that, we found that several variables were not significant. We excluded the variables: national median household income, non-seven major felonies, date lagged by 1 month, and date lagged by 3 months.

With this final OLS model, we found that all variables except unemployment rate were statistically significant. Our adjusted R2 was 98.7%, indicating that this model did a great job at explaining the variance in our data, but this is abnormally high. In this case, an adjusted R2 of 98.7% is not typical. When analyzing the qqplot and histogram, they both appeared fairly normal. We found that the error variance was not constant, evidenced by the “Heteroskedastic Residuals” plot and the Residuals vs Fitted plot still showed signs of time series issues. Our Durbin Watson test unfortunately confirmed that our model still suffered from severe positive correlation (0.244).

In order to address some of these challenges, we moved on to transforming and lagged date-related variables but found that significant correlation persisted. Our second model specification for this model was a Weighted Least Squares approach. This alternative, we hoped, would manage non-constant variance more effectively. In an effort to maximize our model’s potential and further enhance variable selection, we combined this with a stepwise regression.

## WLS Model

Our WLS model helped moderate the impact of variance differences, however our persistent issue of serial correlation was not resolved. The stepwise regression confirmed that all variables retained from the previous OLS model are statistically significant. Our adjusted R2 is 99.67%, but again, our plots indicated serial correlation. To address this problem of heteroskedasticity, we considered using a logarithmic transformation to adjust variables that may be having an imbalanced effect on the model, with a specific focus on yearly data that we distributed monthly.

## Bootstrap Aggregation

Before we ran our bootstrap model, we first logged yearly variables (crime and income) and ran a full and reduced model. We found our persistent issues of heteroskedasticity and nonnormality, but both are to be expected with time series data. Both models yielded an adjusted R2 of 99.66% and a decrease in standard errors of our coefficients. Again, this adjusted R2 is abnormally high. This gives us a high likelihood of overfitting our model now. To address this, we will perform a bootstrap model to estimate the standard errors and confidence intervals of our coefficients. However, we computed confidence intervals and standard errors and plotted them. None of the individual predictor CIs cross over the 0 line - meaning we have confidence that they are not ambiguous and can be accurately applied to this model going forward. With the bootstrap model, our Durbin Watson statistics were improved (0.35), but still proved we had severe positive correlation. Our adjusted R2 was 99.66%, and we found that 11 out of 13 variables are statistically significant.

## Cross Validation

Finally, we ran a 10-fold Cross Validation (10FCV) to further assess our model’s accuracy. Splitting into 10 folds gave us 252 training points and 18 prediction points per fold iteration. After completion of the 10FCV, we noticed there was no improvement in the overall model performance. remained constant, which is a positive, however the increased from the previous two model runs (Bootstrap and Stepwise WLS). This leads us to believe that some bias has been introduced into the model following the bootstrap model training and tuning. This is to be expected, however, since bootstrap aggregation tends to introduce bias through the repeated use of the same data points.

# Conclusion:

Our exploration into this data underscores how interconnected housing markets, economic conditions, and social indicators are. This 24-year analysis provided valuable insights into the various economic and social factors. Through our various model specifications and modeling methods, we were able to identify key predictors, namely household income, major felonies, federal interest rates, and mortgage rates. The inclusion of dummy variables for the 2008 financial crisis and COVID-19 pandemic allowed us to account for external shocks that significantly impacted the housing market during these periods. Most of our predictors were statistically significant, and these findings emphasize the key areas that policymakers, homebuyers, and investors should narrowly focus their attention. Although our models were able to explain 99.66% of the variance of our data, we still had serial correlation and heteroskedasticity challenges, which were expected due to the time series nature of our data. In order to alleviate these issues, future models may explore additional variables or other techniques to correct for serial correlation. In conclusion, we hope our research provides a comprehensive framework for key stakeholders to navigate and better understand the fluctuation in our housing market and make more informed decisions.