**Deliverable 2: Preliminary Data Analysis Report**

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, and rental cost index.
* **Federal Reserve Economic Data (FRED)** (https://fred.stlouisfed.org/)**:** Federal interest rate, 15-year, and 30-year mortgage rates, NY median household income, national median household income, and unemployment rates.
* **NYPD Crime Statistics** (<https://www.nyc.gov/site/nypd/stats/crime-statistics/historical.page>)**:** Misdemeanor offenses, major felonies, and non-seven major felony offenses, aggregated to the city level.

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.

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.

Our data required extensive cleaning and formatting, despite the records being very clean. Data preprocessing involved addressing missing values, standardizing all data to a monthly frequency, using averaging for variables reported at less frequent intervals, and creating two dummy variables to account for the 2008 financial crisis and the COVID-19 pandemic. We were left with 14 variables and 300 observations.

Initial Set of Predictors:

Our initial set of predicators is closely related to our dataset, since we built a veritable Frankenstein’s Monster of a dataset. By pulling various variables from different sources, we were able to define an initial set of economic predictors that we suspect could influence median single-family home prices. Right now, we are running the regression with all the predictors:

* Mean Single-family home value
* Federal Interest Rate
* 15-year Mortgage Rate
* 30-year Mortgage Rate
* NY Median Household Income
* National Median Household Income
* Major Felonies
* Non-Severn Major Felonies
* Misdemeanor Offenses
* Unemployment Rate

We will later use a stepwise formula to determine which predictors are significant enough to keep in our final model.

OLS Regression: see appendix 1

Since we have concluded that our data suffers from serial correlation, we must transform our variables and lag the data. Therefore, we will use logistic regression to build our model.

We found that most of our predicators are significant. Those that were not significant include national median household income and non-seven major felonies.

Descriptive Analytics, Inspection of Plots, Tests for OLS assumptions: see appendix 2

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.

Conclusion:

Our initial exploration into this data underscores how interconnected housing markets, economic conditions, and social indicators are. As we continue developing this model, we hope that it will lead to greater understanding that will shape comprehensive policy interventions that account for these complex relationships rather than addressing each factor in isolation.

Our OLS regression analysis indicated that most predictors in our model were significant, with the notable exceptions of national median household income and non-seven major felonies. We also found clear evidence of serial correlation and heteroskedasticity, which is expected given the time-series nature of our data. The Durbin-Watson statistics of 0.362 confirmed extreme positive correlation, which will inform us of our next steps as we transform and lag variables.

**Appendix**

Code

# import cleaned data set

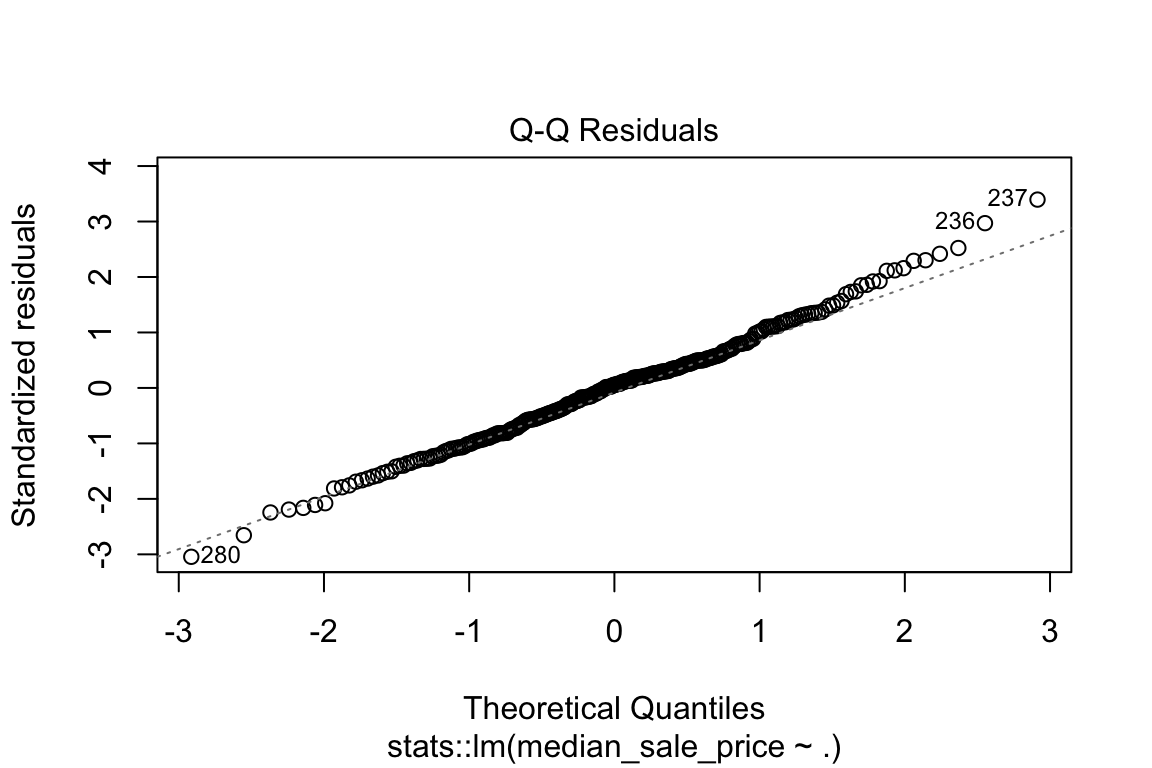
ny\_housing\_data\_clean <- readr::read\_csv("data/ny\_housing\_data\_clean.csv")

Code

# plot a reduced lm model

lm\_model <- stats::lm(median\_sale\_price ~ ., data = ny\_housing\_data\_clean)

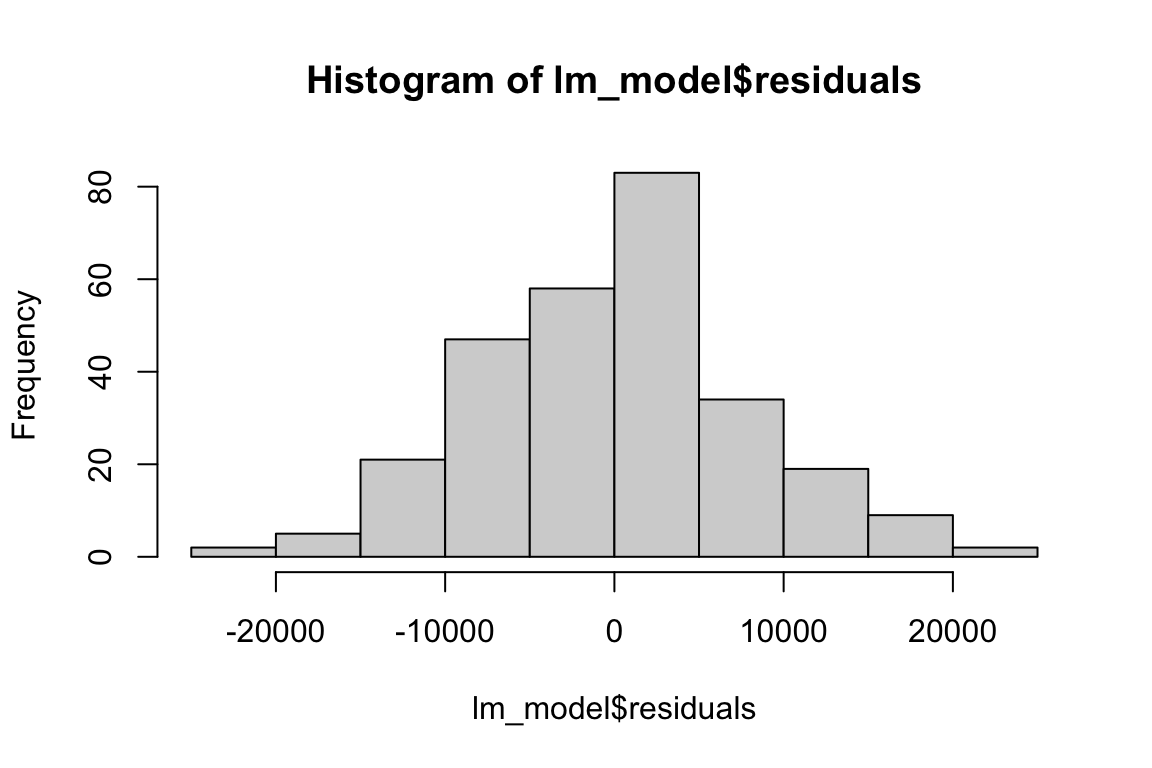
graphics::plot(lm\_model, which = 2)



Appendix 1 – Q-Q Residuals Plot

Code

graphics::hist(lm\_model$residuals)



Appendix 2 – Residuals Histogram

Code

summary(lm\_model)

Call:

stats::lm(formula = median\_sale\_price ~ ., data = ny\_housing\_data\_clean)

Residuals:

Min 1Q Median 3Q Max

-23306.1 -5769.8 435.6 4436.0 24177.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -8.365e+05 8.063e+06 -0.104 0.917

...1 -2.340e+03 2.197e+04 -0.107 0.915

Date 9.751e+01 7.216e+02 0.135 0.893

mean\_sfr\_value 6.852e-01 1.711e-02 40.053 < 2e-16 \*\*\*

Fed\_Interest\_Rate 4.915e+03 8.410e+02 5.844 1.49e-08 \*\*\*

mortgage\_rate\_15\_year 4.445e+04 6.366e+03 6.983 2.33e-11 \*\*\*

mortgage\_rate\_30\_year -4.859e+04 6.032e+03 -8.056 2.70e-14 \*\*\*

ny\_median\_hh\_income 1.786e+01 3.950e+00 4.522 9.24e-06 \*\*\*

national\_median\_hh\_income -5.965e-01 6.315e+00 -0.094 0.925

non\_seven\_major\_felonies -1.515e+00 3.251e+00 -0.466 0.642

major\_felonies -4.458e+00 1.121e+00 -3.977 9.00e-05 \*\*\*

misdemeanor\_offenses -3.946e+00 5.504e-01 -7.170 7.50e-12 \*\*\*

unemployment\_rate -2.781e+03 4.929e+02 -5.642 4.32e-08 \*\*\*

housing\_crisis 1.520e+04 2.812e+03 5.405 1.45e-07 \*\*\*

covid\_pandemic 1.707e+04 3.067e+03 5.565 6.43e-08 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8180 on 265 degrees of freedom

Multiple R-squared: 0.9968, Adjusted R-squared: 0.9966

F-statistic: 5886 on 14 and 265 DF, p-value: < 2.2e-16

Appendix 3 – Summary Statistics of OLS Model

Code

#heteroskedasticity check

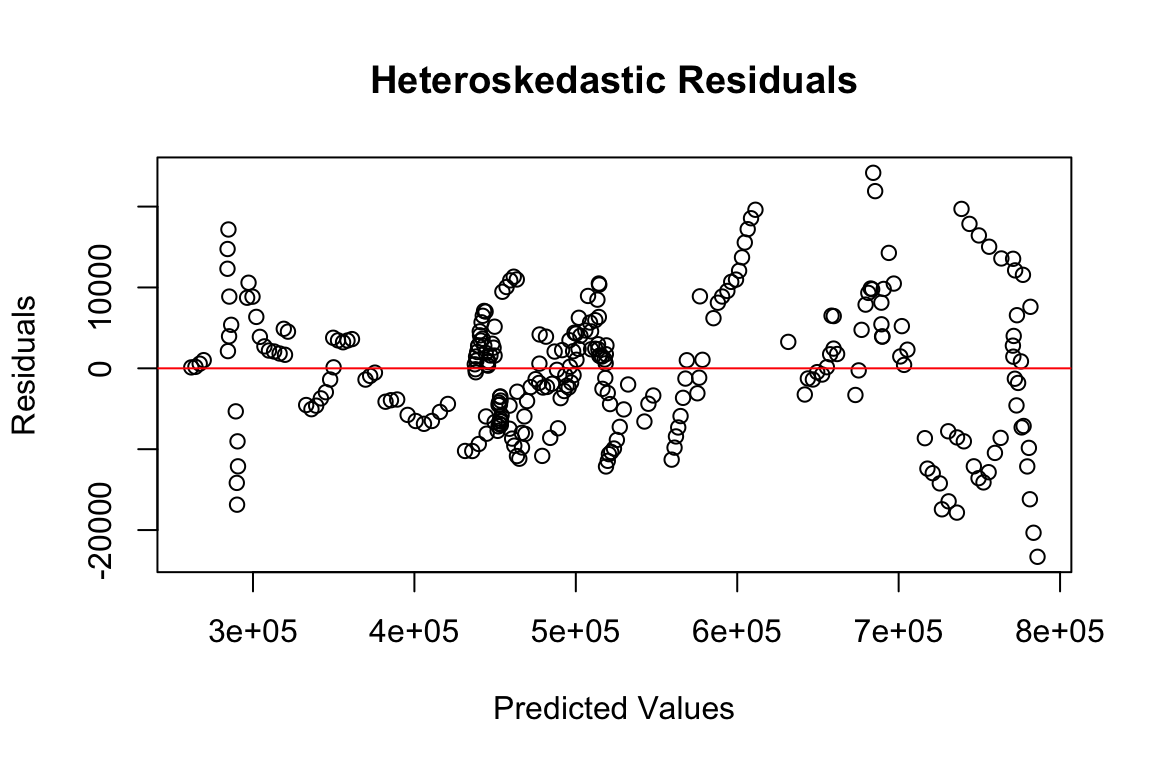
graphics::plot(lm\_model$residuals ~ lm\_model$fitted.values,

main = "Heteroskedastic Residuals",

xlab = "Predicted Values",

ylab = "Residuals")

graphics::abline(h=0, col="red")

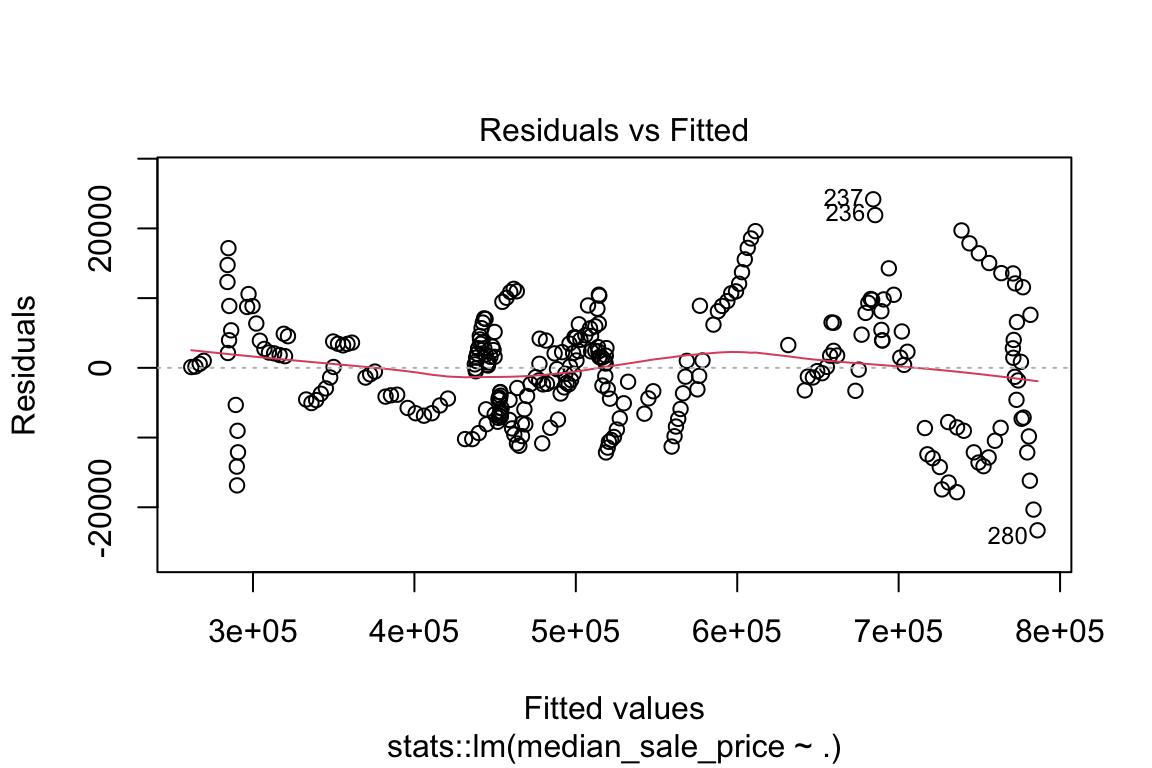


Appendix 4 – Heteroskedasticity Testing (Residual Plot)

Code

#residuals vs fitted plot

graphics::plot(lm\_model, which = 1)



Appendix 5 – Residual vs. Fitted Plot

Code

#checking for serial correlation

lmtest::dwtest(lm\_model)

Durbin-Watson test

data: lm\_model

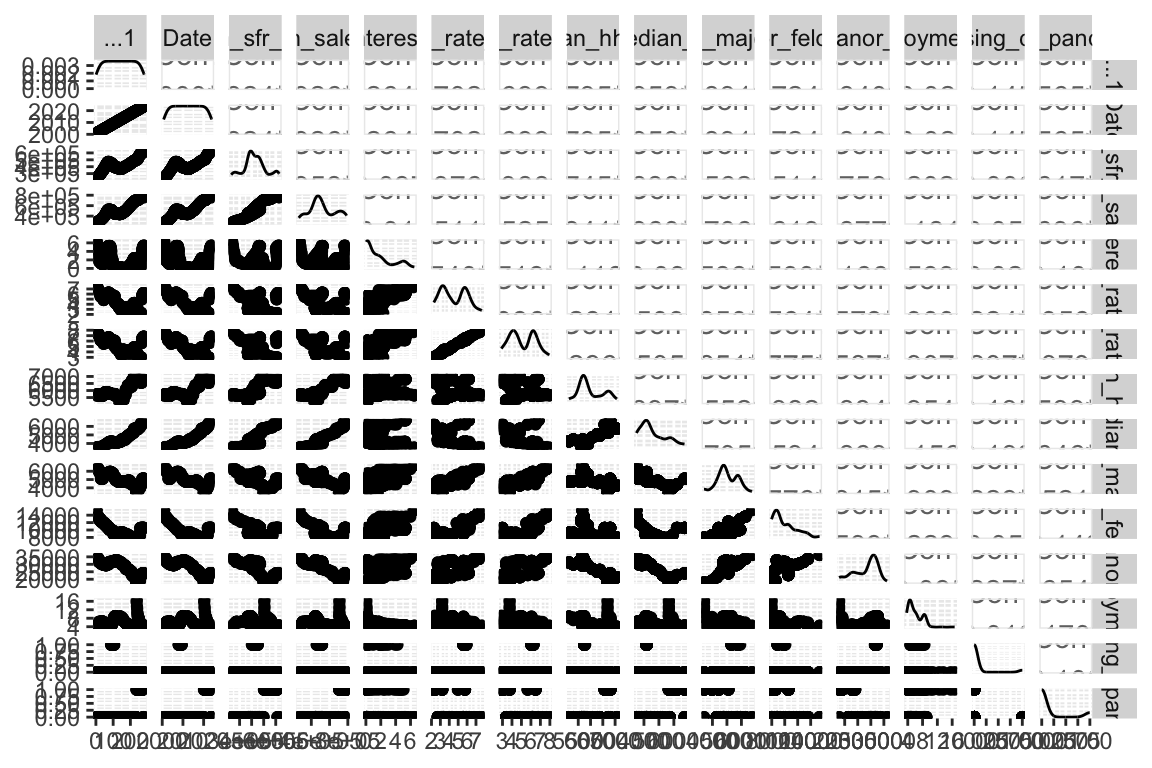
DW = 0.36236, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is greater than 0

Code

#run ggpairs

GGally::ggpairs(ny\_housing\_data\_clean)



Appendix 6 – Pairwise Comparison Plot