INVESTIGATING THE RELATIONSHIP BETWEEN NOISE COMPLAINTS AND HOUSING COST IN NEW YORK CITY

MGT 6203 Team 10 Final Report

*Juan David Rodriguez (jrodriguez381), Evgenia Jane Bugai (ebugai3), Muhammad Ahmad (mahmad67),
Nickolas Tyler Reinig (nreinig3), Hisashi Chris Kominami (hkominami3)*

ABSTRACT

As the U.S housing costs continue to rise, the real estate industry is exploring innovative approaches to understand and predict housing prices. One such approach is the use of alternative data sources that go beyond traditional metrics to provide insight into housing price dynamics in urban areas.

In our group project, we investigated the relationship between alternative data sources, such as 311 noise complaints data, and housing prices for both rental properties and home sales in New York City. We started by conducting an exploratory data analysis using the public NYC 311 noise complaints data, IRS income and Zillow's rent/sales price indexes from 2010 to 2023. We then ran multiple linear regression models to evaluate this relationship and predict future prices.

As a result of explanatory analysis, we've discovered a steep change in the noise complaint levels during the COVID-19 pandemic that caused different seasonality, trends, and absolute values in pre-2020 and post-2020 data. Thus, we decided to exclude 2020 records from the analysis and split the data to pre- and post-2020. Altogether, we created 8 linear regression models, consisting of 4 descriptive explanatory and 4 forecast predictive models. We used two types of price indexes, one for rent and one for sales, and applied these models to both pre-2020 and post-2020 time periods.

As a result of regression analysis, sales price models for the period after 2020 showed exceptionally high adjusted R-Squared and low mean absolute error values. On the contrary, rent price models and pre-2020 price models showed higher error and the prediction showed different trend compared to the actual time series. However, it's worth mentioning that our models could be overfit as the bulk of the explanatory power appears to come from the zip code and month variables that were used as control variables for regulations and seasonality. Thus, we concluded that noise complaints may be capturing other exogenous items such as commercial or construction activity and further investigation is needed to understand how noise complaints can contribute to explaining and predicting housing prices, an application with significant business potential.

In conclusion, an interactive Tableau Dashboard was created to effectively display the findings of the analysis and the predictive model, making them applicable for real-world business scenarios.

PROBLEM STATEMENT AND BUSINESS JUSTIFICATION

NYC is famous for its vibrancy, but it can also be a costly and loud place to live. It has become increasingly difficult for people to find quiet places to live that are also affordable. Our objective is to better understand the relationship between noise complaints (a proxy for noise pollution) and housing sales and rent prices in various neighborhoods throughout New York City.

According to a McKinsey article, "Getting ahead of the market: How big data is transforming real estate", using traditional independent variables (i.e., year built, rooms, location) to predict real estate values can be limiting and exclude many non-traditional factors that can significantly impact the price (McKinsey & Co., 2018).

One such factor that can affect the quality of life and possibly account for the housing prices in New York City is the level of noise pollution. Our hypothesis is that the intensity of noise complaints could negatively affect both the value of houses and the cost of renting, resulting in lower prices.

If the hypothesis is true, this information can be used to enhance existing real estate market research and predictive toolkits. By better understanding the factors that impact housing prices and improve their predictability, city planners, developers, and property owners can make better business decisions. They can identify novel areas with potential value increase and avoid investing in areas where the model predicts value erosion. **As a result**, **assuming a 10% addressable market, property owners and developers selling $2.6 billion**

a year in housing and receiving $8.4 billion a year in housing rentals could benefit from our model. If the additional information results in even a 1% improvement in their performance, the delivered value could surpass $100 million per year.

The betas from the model can also be utilized to develop a noise rating system for each geographical area by considering the distribution and types of noise complaints. This noise rating system could be monetized by online rental platforms like Zillow or Redfin, enhancing the value offered to property owners, improving the search experience for their users, and potentially driving more traffic to their websites that could lead to an increase in revenue.

LITERATURE SURVEY:

In addition to research on the New York City real estate market as well as the noise pollution problem in the city, we also conducted a small literature survey of peer-reviewed papers investigating two specific topics related to our project: (1) the analysis of 311 noise complaints in NYC, and (2) the prediction of house prices and rent costs using 'non-traditional' factors. We found two especially relevant papers, *Noise complaint patterns in New York City from January 2010 through February 2021: Socioeconomic disparities and COVID-19 exacerbations* by Ramphal et al., and *Forecasting Residential Real Estate Price Changes from Online Search Activity* by Beracha et al.

From Ramphal's paper, we learned that noise complaints have been steadily increasing in New York City since 2010 and were greatly amplified during the COVID-19 pandemic, particularly affecting those with lower incomes (Ramphal et al., 2022). This finding suggests that noise complaints per capita are higher in lower income groups and supports our initial idea to model income groups separately, as there may be important differences among these groups.

Beracha's article gives further evidence to support the conclusions of McKinsey's study on 'non-traditional' real estate price prediction factors. It shows that there was as much as an 8.5% difference on average in prices between real estate markets that had exceptionally high online search activity (i.e., the intensity of online searches for "real estate" or "rent" in that market), compared to those with exceptionally low search activity, over a 2-year period. However, it's important to note that this result held only over short periods – the study found that after 5 years the difference in prices had disappeared (Beracha et al., 2013).

In summary, these two studies suggest that the topic of our project is relevant and has a solid business justification. Also, the methodologies employed in both gave us a better sense of how to analyze and predict real estate prices and their correlation with noise complaints, which will help to guide us as our project progresses.

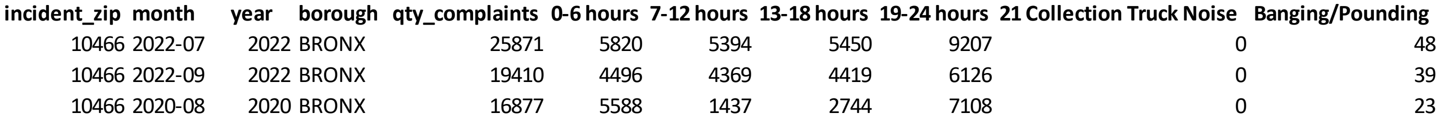
INITIAL HYPOTHESIS:

We hypothesized that housing and rent prices in New York City neighborhoods are negatively correlated with the per capita number of noise complaints in those neighborhoods, and that the relationship is statistically significant. We expected this to be the case especially since the pandemic began, since the increased number of remote workers, increased time spent at home, and numerous pandemic-related challenges has likely made individuals more sensitive to noise.

We also analyzed the types of noise complaints and the time of day when each complaint was made. We hypothesized that types of noise may have a stronger correlation with housing/rent prices in neighborhoods, and other types may be less correlated or even correlated in a different direction. The time of day or night a noise complaint occurs may show a similar phenomenon, with certain times more strongly correlated to housing prices. For example, noise late at night or in the early morning may be more likely to indicate crime activity than noise in the afternoon, with crime rates likely having a relatively strong correlation to housing/rent prices.

EXTRACTION AND PREPARATION OF THE DATASETS:

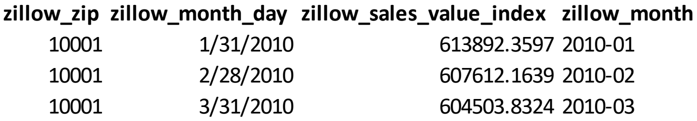
311 NYC calls (New York City Open Data, 2023): 311 New York City calls dataset is a very large dataset comprised of more than 32 million registries and 41 columns will all complaints from 2010 to 2023. We used the available API to access and extract the data by using Python Sodapy library, the code was implemented in AWS Sagemaker due to the processing and memory required by the size of the dataset. The API allowed us to limit the extraction to the complaints containing the keyword "noise" in the description. Due to throughput and performance limitations, we had to perform the extraction for one year at a time. The code included a for loop with automatic retries on timeouts and quality checks to secure the quality of the extraction. The extracted Dataframe included 5'820.662 registries. The hour of each complaint was extracted from the "create_date" field, and then classified in one of four-hour ranges (0-6, 7-12, 13-18, and 19-24 hours). We also created dummy variables for each complaint range and each of the 37 complaint types. Lastly, the variables were grouped by zip code and month to create a panel containing one row per zip code per month (29619 rows by 46 columns).



incident_zip	month	year	borough	qty_complaints	0-6 hours	7-12 hours	13-18 hours	19-24 hours	21 Collection Truck Noise	Banging/Pounding
10466	2022-07	2022	BRONX	25871	5820	5394	5450	9207	0	48
10466	2022-09	2022	BRONX	19410	4496	4369	4419	6126	0	39
10466	2020-08	2020	BRONX	16877	5588	1437	2744	7108	0	23

Figure 1. Resulting 311 Zip-Month Complaint Panel

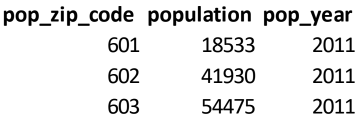
Zillow Housing Research Data - Zillow Home Value Index (ZHVI) and Zillow Observed Rent Index (ZORI) (Zillow, Inc., 2023): We extracted the datasets from the Zillow housing research data. The data schema had one row per Zip code and one column per month. We transformed the data in Python using the melt function from Pandas and filtering the Zip codes available in the 311-complaints dataset.



zillow_zip	zillow_month_day	zillow_sales_value_index	zillow_month
10001	1/31/2010	613892.3597	2010-01
10001	2/28/2010	607612.1639	2010-02
10001	3/31/2010	604503.8324	2010-03

Figure 2. Zillow value index Tibble

US census estimated population series by zip code (United States Census Bureau, 2023): We downloaded the data from the Census Bureau website. The website exports one csv file per year containing the yearly estimated population for each zip code. We used the pandas and a for loop to read all files, select columns, standardize the schema, and concatenate all years in a single file.



pop_zip_code	population	pop_year
601	18533	2011
602	41930	2011
603	54475	2011

Figure 3. Population dataset

Joining all datasets and creating calculated fields: The final step was to join all datasets using the zip code and the month as key. The initial exploratory analysis showed that the distribution of the population per zip code was uneven, and that some zip codes grew their population in the 13-year period. Consequently, we decided to calculate the ratio of complaints by dividing each complaint variable by the estimated population of the year. We also calculated the income per capita, and we created a dummy variable per zip code as explained in the model description. The resulting dataframe has 27,483 rows and 279 columns.

The following table shows the final list of variables to be tested in the model:

Variable	Type	Variable	Type
income_per_capita	numeric	ratio_Noise, Barking Dog (NR5)	numeric
zillow_month	categ	ratio_Noise, Ice Cream Truck (NR4)	numeric
zip_code	categ	ratio_Noise, Other Animals (NR6)	numeric
ratio_qty_complaints	numeric	ratio_Noise: lawn care equipment (NCL)	numeric
ratio_0-6 hours	numeric	ratio_Noise: Air Condition/Ventilation Equip, Commercial (NJ2)	numeric
ratio_7-12 hours	numeric	ratio_Noise: Air Condition/Ventilation Equip, Residential (NJ1)	numeric
ratio_13-18 hours	numeric	ratio_Noise: Alarms (NR3)	numeric
ratio_19-24 hours	numeric	ratio_Noise: Boat(Engine	numeric
ratio_21 Collection Truck Noise	numeric	ratio_Noise: Boat(Engine,Music,Etc) (NR10)	numeric
ratio_Banging/Pounding	numeric	ratio_Noise: Construction Before/After Hours (NM1)	numeric
ratio_Car/Truck Horn	numeric	ratio_Noise: Construction Equipment (NC1)	numeric

ratio_Car/Truck Music	numeric	ratio_Noise: Jack Hammering (NC2)	numeric
ratio_Engine Idling	numeric	ratio_Noise: Loud Music From Siebel System - (NP21)	numeric
ratio_Flying Too Low	numeric	ratio_Noise: Loud Music/Daytime (Mark Date And Time) (NN1)	numeric
ratio_Horn Honking Sign Requested	numeric	ratio_Noise: Loud Music/Nighttime(Mark Date And Time) (NP1)	numeric
ratio_Hovering	numeric	ratio_Noise: Manufacturing Noise (NK1)	numeric
ratio_Loud Music/Party	numeric	ratio_Noise: Other Noise Sources (Use Comments) (NZZ)	numeric
ratio_Loud Talking	numeric	ratio_Noise: Private Carting Noise (NQ1)	numeric
ratio_Loud Television	numeric	ratio_Noise: Vehicle (NR2)	numeric
ratio_NYPD	numeric	ratio_Noise: air condition/ventilation equipment (NV1)	numeric
ratio_News Gathering	numeric	ratio_Other	numeric
ratio_Noise	numeric	ratio_Passing By	numeric
ratio_People Created Noise	numeric		

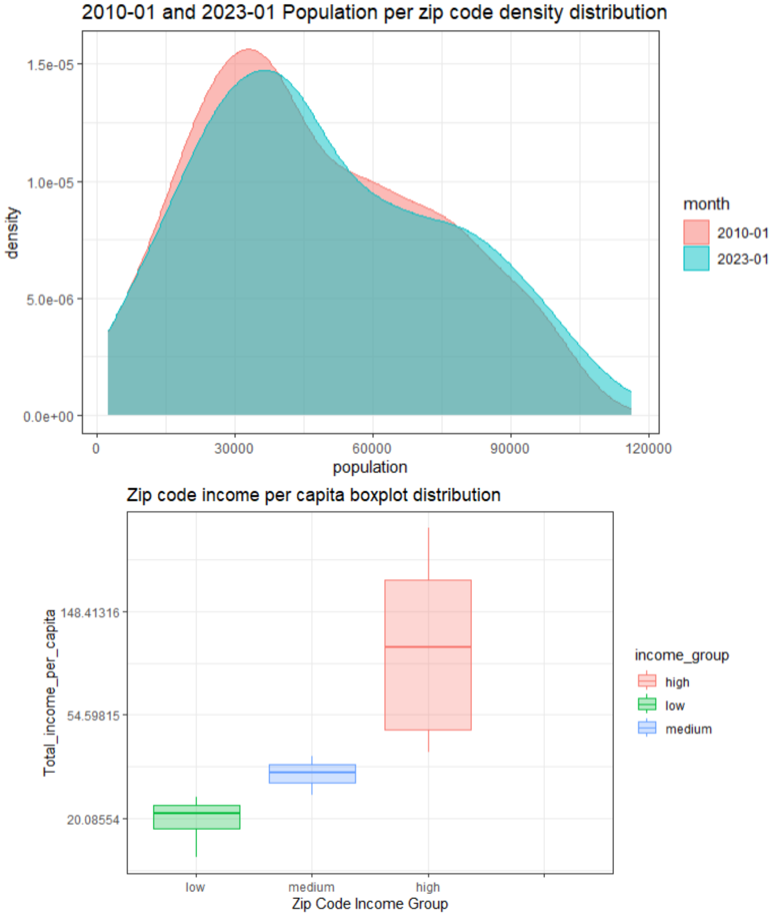


Figure 4. Population and income per capita distribution

zillow_zip	zillow_sales_value_index	zillow_month	borough	qty_complaints	0-6 hours	population	Total_income_per_capita	income_group	comp_ratio_1000_total	comp_ratio_1000_0-6 hours
10001	613892.3597	2010-01	MANHATTAN	98	51	21097	134.1834384	high	4.64521022	2.417405318
10001	607612.1639	2010-02	MANHATTAN	93	35	21097	134.1834384	high	4.4082097	1.65900365
10001	604503.8324	2010-03	MANHATTAN	105	48	21097	134.1834384	high	4.97701095	2.275205005

Figure 5. Population and income per capita distribution

DATA CLEANING: REMOVING NON-RESIDENTIAL ZIP CODES. UNDERSTANTING SEASONALITY, TREND, AND OUTLIERS

The dataset included datapoints for 145 months (about 12 years) between 2010 and 2023. Exploratory data analysis revealed that Manhattan leads the boroughs both in terms of the total number of complaints and the population adjusted complaints between 2015 and 2023 as shown in the figures below.

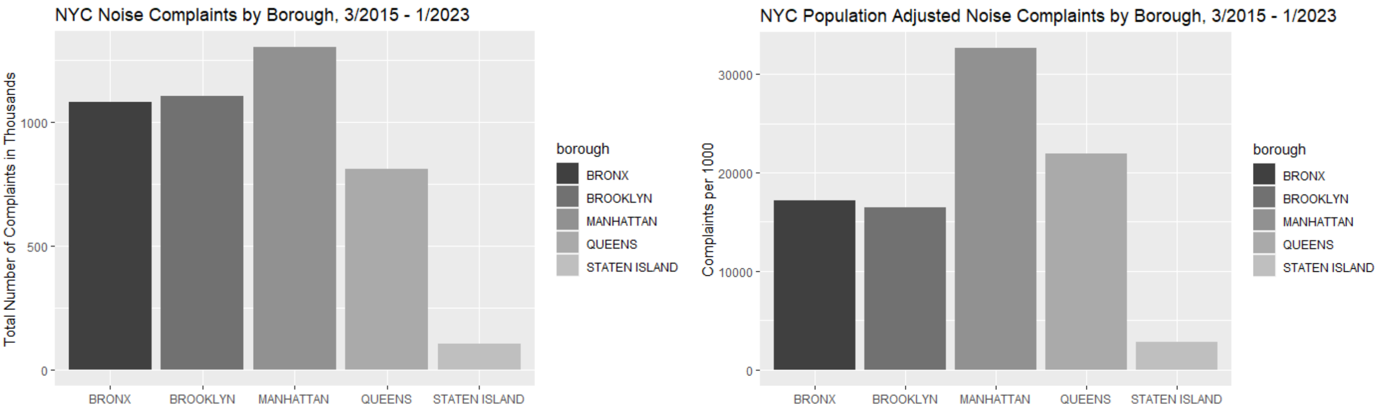


Figure 6. Total Number of Complaints and Population Adjusted Complaints by NYC Boroughs

In addition, our analysis demonstrated that complaints had significant seasonality and trend, while sales price index had seasonality. The analysis also showed that the pandemic in 2020 resulted in a significant change of behavior; 2020 time series shows a spike in the complaint rate, with much higher variability between zip and higher seasonality. Higher complaint rates and seasonality were maintained with some smoothing in 2021, 2022 and 2023. As a result of the analysis, we decided to 1) Remove 2020 measurements from the sample 2) split the data series in two periods: 2010-2019 and 2021-2023 so that separate models could be implemented per period 3) remove trend and seasonality.

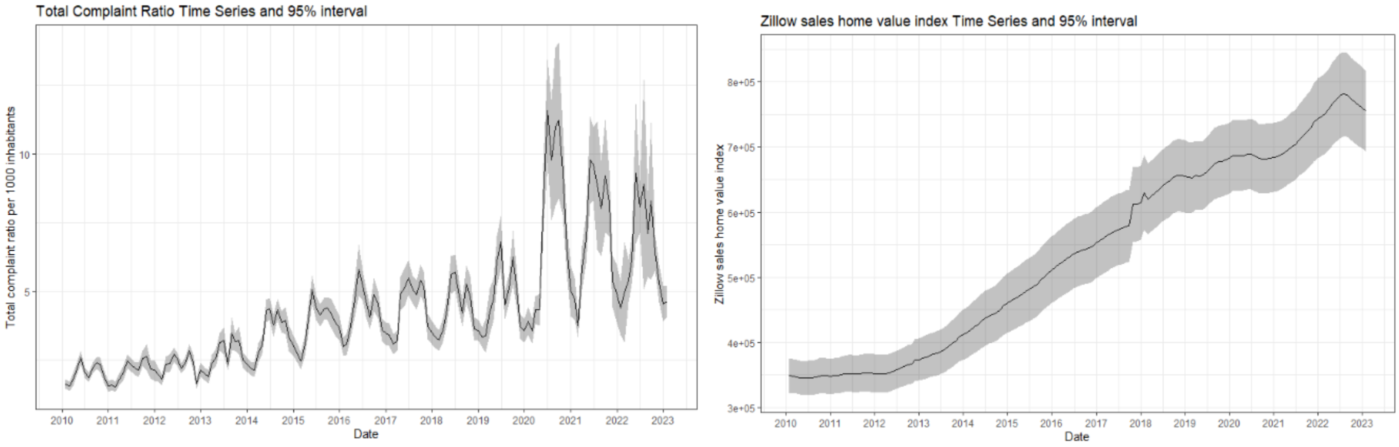


Figure 7. Complaint ratio per 1000 inhabitants and Zillow home sales index time series with 95% interval

Understanding trend and seasonality: We used the "decompose" function from R to understand the marginal contribution of the complaint rate, while excluding the seasonality and trend from both data series. The decomposed analysis confirmed the existence of the seasonality and trend components, and the drastic change in the time series since 2020.

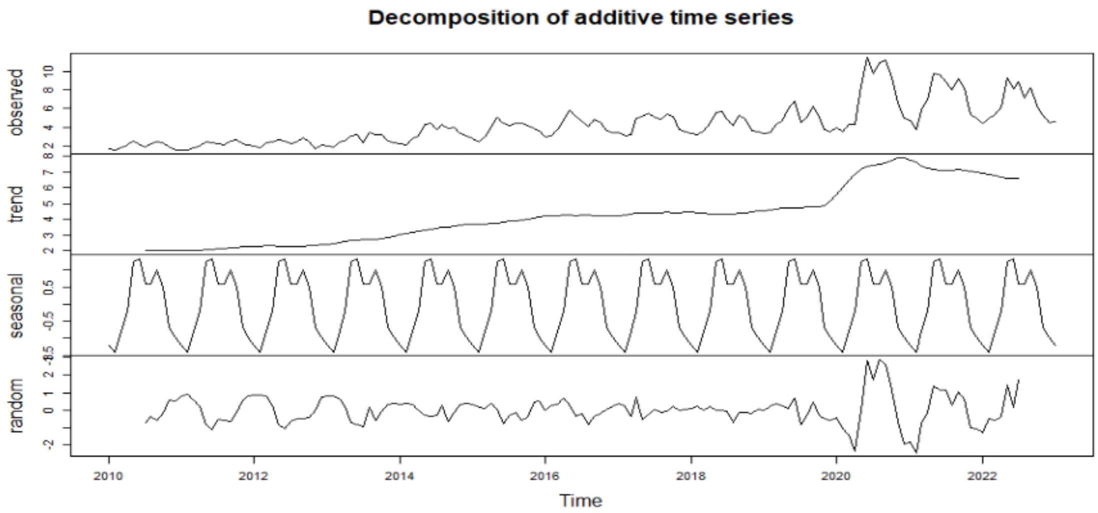


Figure 8. Decomposed time series from the average total complaint rate per month

We tried two methods to remove seasonality and trend:

1) Using seasadj and diff functions from R forecast library
2) Manually implementing a linear regression on the month (categorical) and the n_month (numerical) and then subtracting each component from the time series.

The de-trend de-seasonalized time series was calculated as follows:

$$Complaint\ rate_{det\ rend} = Complaint\ rate(raw) - \Sigma_2^{12}\beta_{i-month} + \alpha_{seasonality} - \beta_{n-month} + \alpha_{trend}$$

Both methods resulted in similar time series, but the second option was selected as it seemed to result in a smaller loss of information, especially for the 2021-2023 time series.

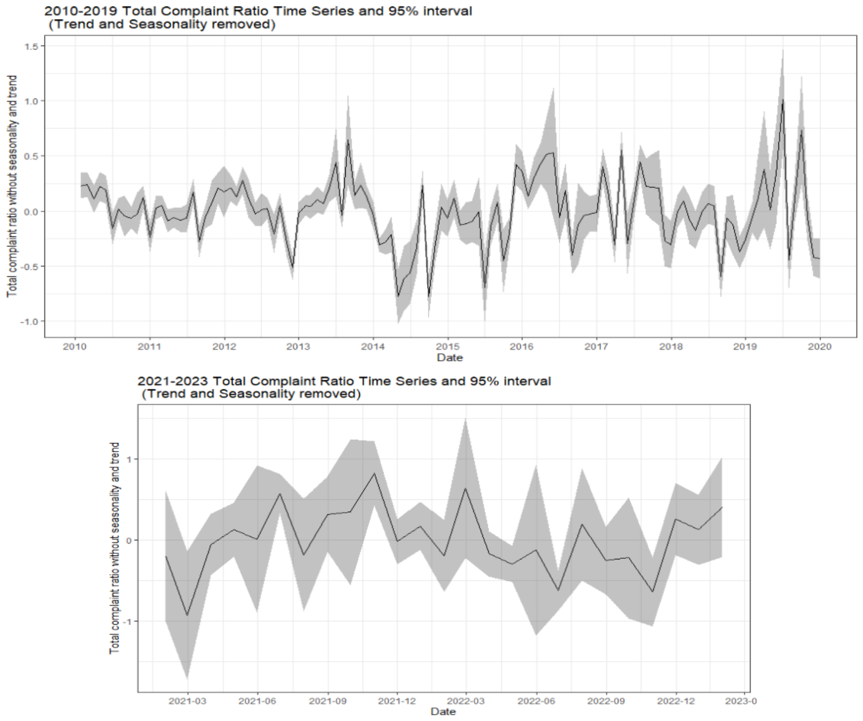


Figure 9. 2010-2019 and 2021-2023 complaint ratio time series after removing seasonality and trend

Removing outliers: The exploratory scatterplot confirmed the presence of low and high outliers in the complaint data. We identified them by calculating 1.5 times the interquartile distance and then filtered them out.

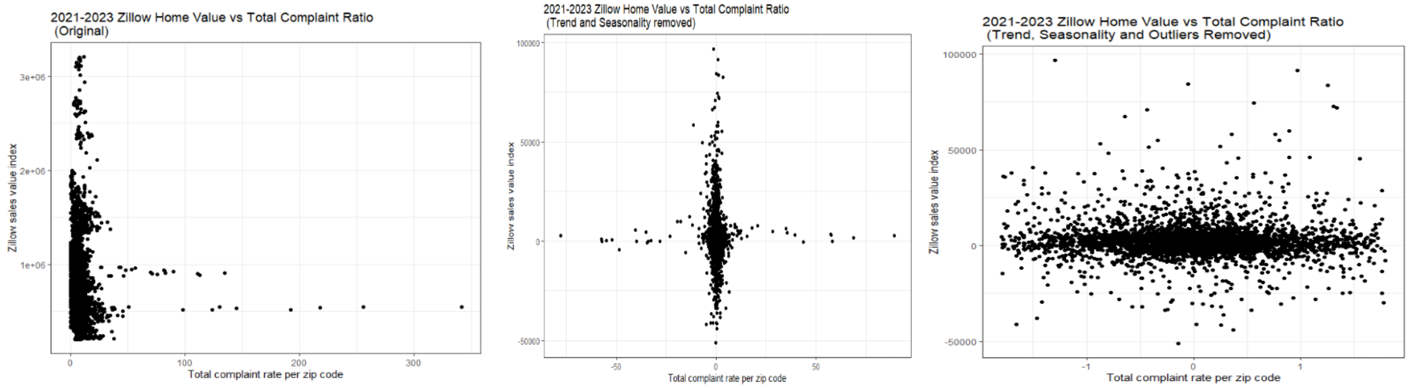


Figure 10. 2010-2019 and 2021-2023 complaint ratio time series after removing seasonality and trend

EXPLORATORY ANALYSIS CONCLUSIONS

The exploratory analysis showed that the average price indexes and the complaint ratio have similar trends, with some lag in the pricing dataset. The analysis also proved that monthly seasonality is strong and has become much larger after 2020. When removing the trend and seasonality components, the scatterplot showed no evident linear relationship between the Zillow sales index and the complaint rate. A linear model with Zillow indexes as dependent variable, and de-trended, de-seasonalized complaint ratio will result in very low R-Squared values and even in non-significant values for complaint variable. However, the complaint time series has an interesting property; once seasonality is removed, it seems to anticipate the change of trend in the prices (excluding 2020).

MODEL DESCRIPTION

In the model, we conducted regression analysis focusing on the impact of noise complaint ratios on the average rental price and residential property prices within a zip code. The noise complaint ratios were calculated by dividing the monthly complaints by the population of the zip code for different types of noise complaints, and these ratios were used as the variable of interest. Our model includes the income per capita for each zip code as a control variable.

We included zip codes as a dummy variable in our regression to account for potential variations among zip codes, such as differing laws and regulations. Following exploratory analysis results, we also included the month of the year as a categorical variable to control for seasonality. The exploratory analysis also showed that the disruption from the 2020 Pandemic resulted in a steep change of behavior in the complaint series and very different seasonality, trend, and absolute values in pre-2020 and post -2022 data. Consequently, we decided to exclude 2020 data from the research and to implement separate models for pre-2020 and post-2020 datasets.

The underlying assumption for our models is that we perceive no reason for endogeneity in our model as the fundamental premise is that fluctuations in noise complaints are exogenous and are not influenced by changes in rental prices. This assumption is reasonable because property owners, in general, do not alter rental rates to manipulate the number of noise complaints, nor do changes in rental rates lead to a variation in the amount of noise generated.

Important Note: Geographical references or zip codes will solely serve as controls, as we intend to employ an index to obfuscate any reference to zip codes in relation to geographical locations. It is important to note that our study does not concern itself with potential differences among zip codes.

Descriptive explanatory models: 4 models were developed to understand the relationship between complaint rates and housing prices (two periods (pre-2020 and post-2020) and two types of price index (rent and sales)):

$$Rnt_{iym} = \beta_0 + \beta_1 TNC_{iym} + \Sigma_2^j \beta_j NC_{ijym} + \rho_i I + \Sigma_2^k \gamma_k \cdot M_{month_m} + \delta_i \cdot Inc.per.capita_i + \epsilon_i$$

$$Pr_{iym} = \beta_0 + \beta_1 TNC_{iym} + \Sigma_2^j \beta_j NC_{ijym} + \rho_i I + \Sigma_2^k \gamma_k \cdot M_{month_m} + \delta_i \cdot Inc.per.capita_i + \epsilon_i$$

Where each income group and each period will have a separate model its corresponding specific coefficients

- β_1: Coefficient that represents the change in rental or sales prices given change in noise complaints.
- β_j: Coefficient that represents the change in rental or sales prices given change in noise complaints of a j type

Variables:

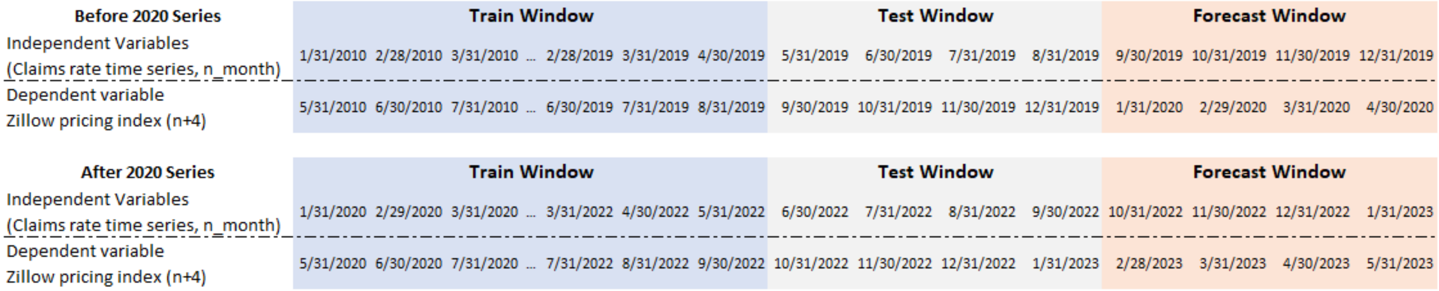
- Rnt_{iym}: Average rent price on ith geographical area, month m and year y
- $\Pr_{iym}$: Average sales price on ith geographical area, month m and year y
- TNC_{iym}: Total number of noise complaints in ith geographical area at month m and year y
- NC_{ijym}: Number of noise complaints in ith geographical area and type j at month m and year y
- I_i: I_i: Index for i^{th} zip-code or geographical area
- M_{month_m}: Month of the year for the ith measurement

We acknowledge that changes in noise complaints are unlikely to have an immediate impact on rental or real-estate prices. In fact, the average search time alone for individuals seeking new rentals in New York can be up to a month, which may vary based on a multitude of variables. Therefore, it can be established that there is a period of time for information to reach both buyers and sellers in the market, ultimately leading to a delay in the effect of noise complaints on pricing. To account for this delay, we decided to test various lag periods to observe how pricing is influenced by the number of noise complaints. This allowed us to determine the most effective lag period to use in our analysis and gain a more comprehensive understanding of the relationship between the noise complaints ratio and prices in the real-estate market. To assess the impact of different lag periods on the significance level of the noise complaints ratio variable, we systematically introduced varying lag periods to our regression model. We found that a lag of four months was overall most effective to the fit.

Forecasting predictive models: 4 models were developed to understand the relationship between complaint rates and future housing prices in the following four months (two periods (pre-2020 and post-2020) and two types of price index (rent and sales)):

$$Rnt_{iy(m+4)} = \beta_0 + \beta_1 TNC_{iym} + \Sigma_2^j \beta_j NC_{ijym} + \rho_i I + \Sigma_2^k \gamma_k \cdot M_{month_m} + \delta_i \cdot Inc\_per\_capita_i + \epsilon_i$$
$$\Pr_{iy(m+4)} = \beta_0 + \beta_1 TNC_{iym} + \Sigma_2^j \beta_j NC_{ijym} + \rho_i I + \Sigma_2^k \gamma_k \cdot M_{month_m} + \delta_i \cdot Inc.per.capita_i + \epsilon_i$$

To avoid a data leak in the forecasting predictive models, and evaluate the real accuracy of the prediction, we separate the dataset into a training window, test window and future forecasting window. The following diagram shows how the data was separated into train, test, and forecasting windows.



Before 2020 Series	Train Window							Test Window				Forecast Window			
Independent Variables (Claims rate time series, n_month)	1/31/2010	2/28/2010	3/31/2010	…	2/28/2019	3/31/2019	4/30/2019	5/31/2019	6/30/2019	7/31/2019	8/31/2019	9/30/2019	10/31/2019	11/30/2019	12/31/2019
Dependent variable Zillow pricing index (n+4)	5/31/2010	6/30/2010	7/31/2010	…	6/30/2019	7/31/2019	8/31/2019	9/30/2019	10/31/2019	11/30/2019	12/31/2019	1/31/2020	2/29/2020	3/31/2020	4/30/2020

After 2020 Series	Train Window							Test Window				Forecast Window			
Independent Variables (Claims rate time series, n_month)	1/31/2020	2/29/2020	3/31/2020	…	3/31/2022	4/30/2022	5/31/2022	6/30/2022	7/31/2022	8/31/2022	9/30/2022	10/31/2022	11/30/2022	12/31/2022	1/31/2023
Dependent variable Zillow pricing index (n+4)	5/31/2020	6/30/2020	7/31/2020	…	7/31/2022	8/31/2022	9/30/2022	10/31/2022	11/30/2022	12/31/2022	1/31/2023	2/28/2023	3/31/2023	4/30/2023	5/31/2023

Figure 12. Train, test and forecast window split

The first stage of feature selection for each model was implemented by selecting only significant variables from the original model.

MODEL RESULTS

Descriptive explanatory models: The following table shows the resulting performance metrics from all exploratory model, top significant complaint variables are ranked by the absolute value of their coefficient, as long as the variable results significant:

	Number of variables	Adjusted R^2	Number of significant complaint variables	Top significant complaint variables and coefficients
Rent prices - Before 2020	208: (16 complaint rates, 1 income, 180 zip dummies, 11 month of year)	0.9835	16	News.Gathering, -289.38 Noise..Manufacturing.Noise..NK1., -143.85 Ice.Cream.Truck..NR4., -92.63 Boat.Engine.Music.Etc...NR10., 62.09 Construction.Equipment..NC1., -55.98
Rent prices - After 2020	207 (16 complaint rates, 1 income, 179 zip dummies, 11 month of year)	0.9233	16	Noise, 2403.23 Noise..Boat.Engine, 1380.39 Noise..Other.Animals..NR6., 552.67 Noise..air.condition.equipment..NV1., -255.07 Noise..Barking.Dog..NR5., 244.06
Sales prices - Before 2020	219 (30 complaint rates, 1 income, 177 zip dummies, 11 month of year)	0.8741	30	Horn.Honking.Sign.Requested..NR9., -1183421.93 Noise...lawn.care.equipment..NCL., 359304.90 Air.Condition.Equip..Residential..NJ1., -319623.76 air.condition.equipment..NV1., 318637.71 NYPD, 299397.70
Sales prices - After 2020	199 (13 complaint rates, 1 income, 174 zip dummies, 11 month of year)	0.9842	13	Collection.Truck.Noise, -630950.39 Noise, 421627.85 News.Gathering, -109170.12 .air.condition.equipment..NV1., -59450.15 Barking.Dog..NR5., 48558.45

Table 1. Descriptive exploratory models' performance

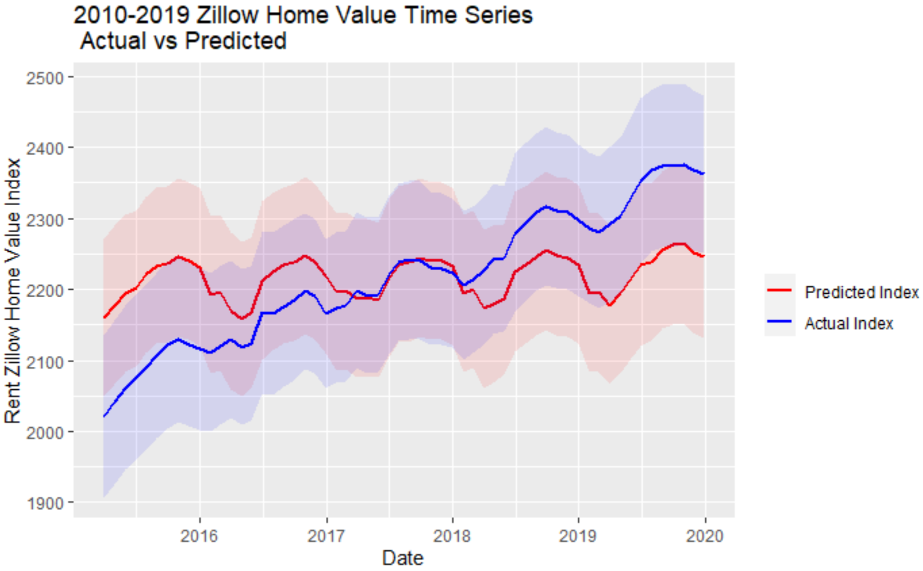


Figure 13. Before 2020 rent prices model plot (predicted price vs. actual price)

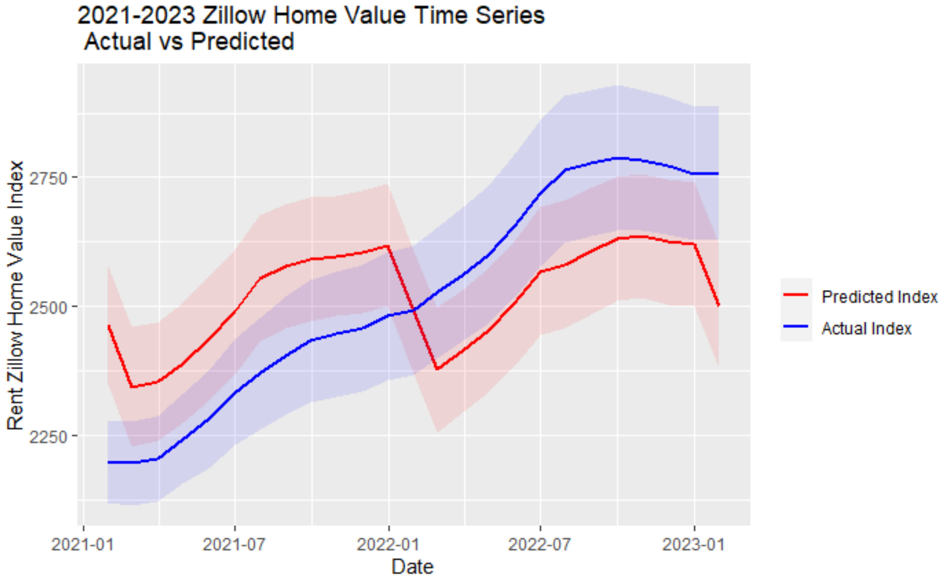


Figure 14. After 2020 rent prices model plot (predicted price vs. actual price)

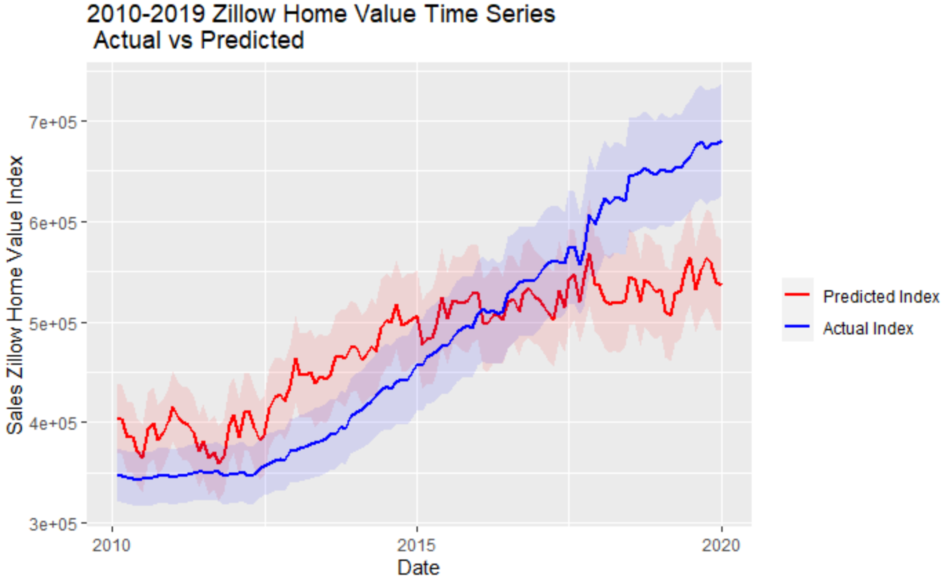


Figure 15. Before 2020 sales prices model plot (predicted price vs. actual price)

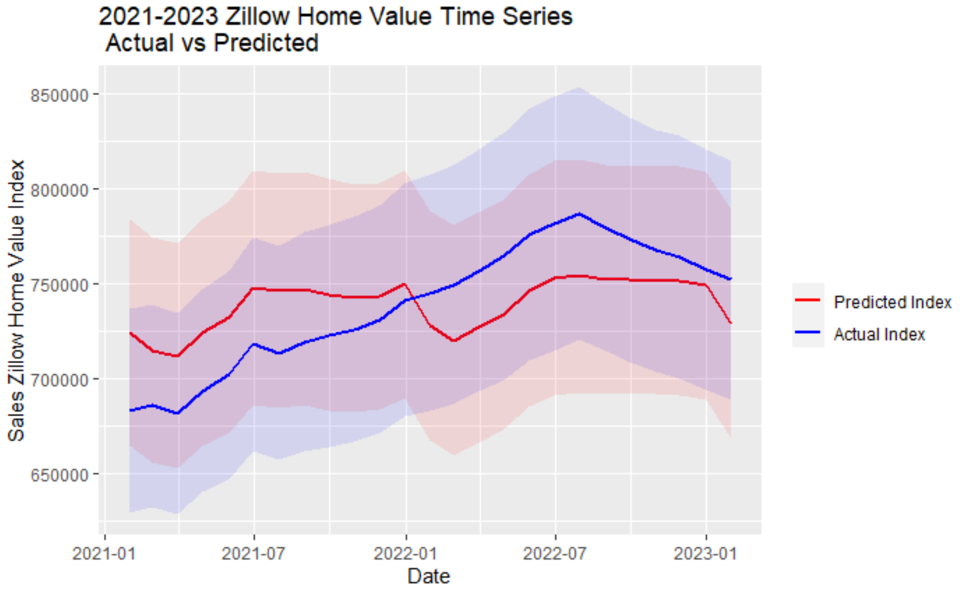


Figure 16. Before 2020 sales prices model plot (predicted price vs. actual price)

Forecast predictive models: The following table shows the resulting performance metrics from all forecast models, top significant complaint variables are ranked by the absolute value of their coefficient, as long as the variable results significant:

	Number of variables	Adjusted in test window R^2	Mean absolute error in test window	Top 5 significant complaint variables and coefficients
Rent prices - Before 2020	195: (3 complaint rates, 1 income, 180 zip dummies, 11 month of year)	0.9937	6.7%	Loud.Music.Nighttime...NP1., 1328.99 Other.Animals..NR6., -186.05 Car.Truck.Music, 62.50
Rent prices - After 2020	203: (12 complaint rates, 1 income, 179 zip dummies, 11 month of year)	0.9668	9.9%	Noise, 6138.83 Collection.Truck.Noise, -3182.52 Barking.Dog..NR5., 258.71 .Ice.Cream.Truck..NR4., 173.27 Jack.Hammering..NC2., 158.76
Sales prices - Before 2020	217: (30 complaint rates, 1 income, 175 zip dummies, 11 month of year)	0.8177	22.6%	Horn.Honking.Sign.Requested..NR9., -1046712.34 NYPD, 350819.80 Noise...lawn.care.equipment..NCL., 325767.26 air.condition.ventilation.equipment..NV1., 325186.36 Air.Condition.Equip..Residential..NJ1., -295804.95
Sales prices - After 2020	195: (9 complaint rates, 1 income, 175 zip dummies, 11 month of year)	0.9886	4.4%	Noise, 807852.11 21.Collection.Truck.Noise, -507155.10 News.Gathering, -258197.58 Noise..Other.Animals..NR6., 155432.46 Barking.Dog..NR5., 48607.29

Table 2. Before 2020 rent prices forecast model performance

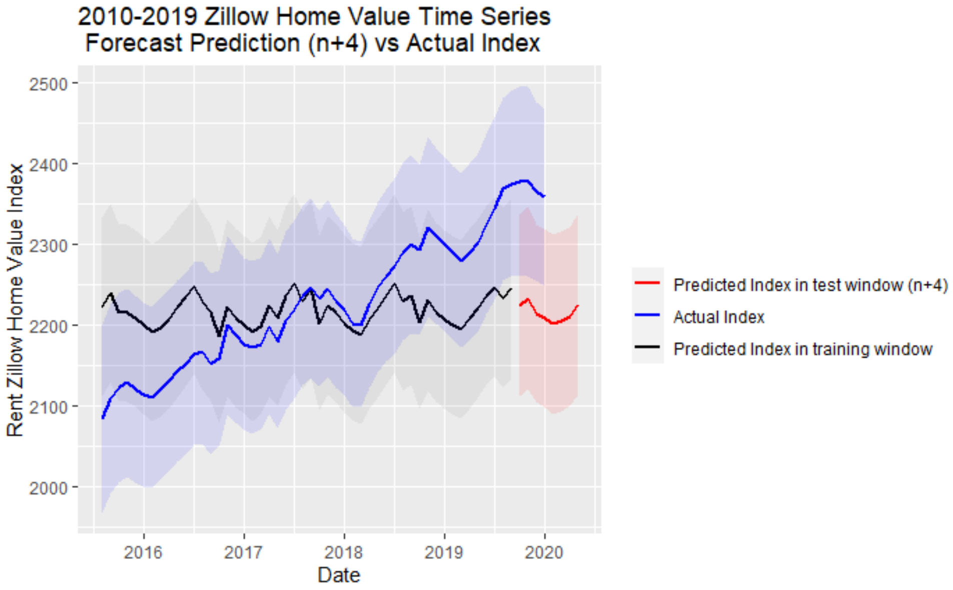


Figure 17. Before 2020 rent prices forecast model plot (predicted price vs. actual price)

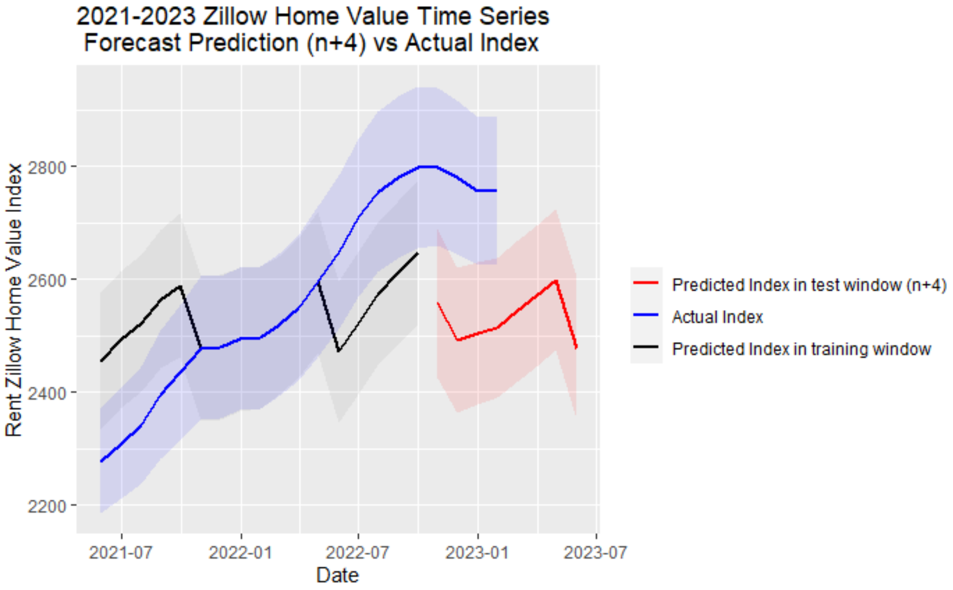


Figure 18. After 2020 rent prices forecast model plot (predicted price vs. actual price)

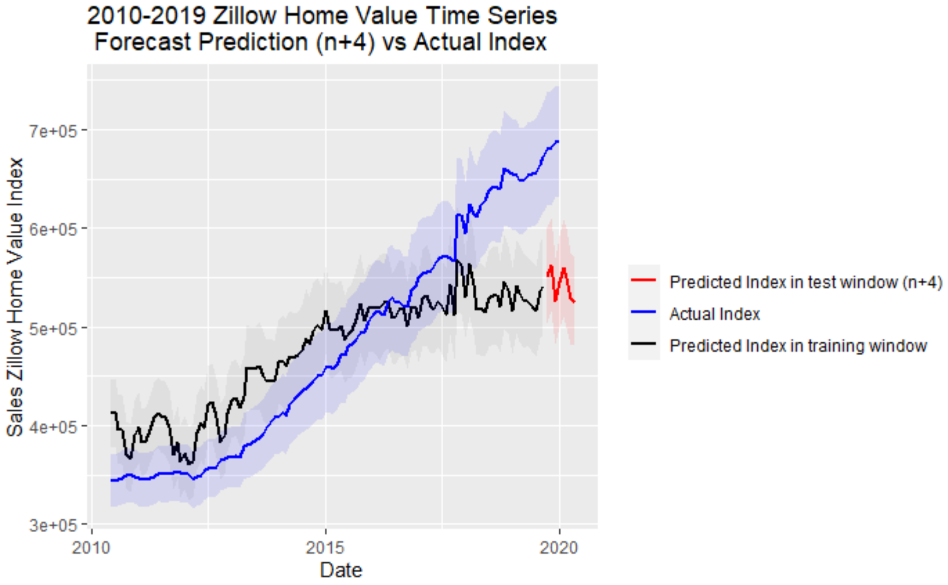


Figure 19. Before 2020 sales prices forecast model plot (predicted price vs. actual price)

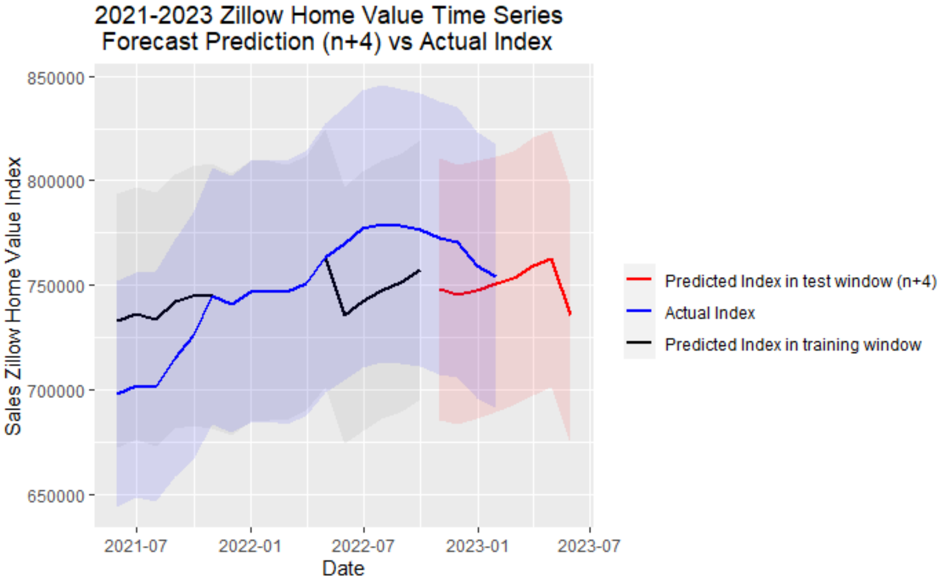


Figure 20. Before 2020 sales prices forecast model plot (predicted price vs. actual price)

ADVANCED FEATURE SELECTION AND OPTIMIZATION

Important aspects of regression modeling include performing goodness of fit testing, evaluating model performance, and exploring the marginal contribution of predicting variables to overall performance. In addition to preliminary feature selection discussed above, we also explored building simpler models using a subset of predicting variables in our original models based on forward selection and backward elimination approaches. Stepwise methods from the ‘olsrr’ package were used because of their ease in implementation and because the high number of dummy variables in our models made it computationally infeasible to explore other approaches like best subset selection. For this effort, we focused on our forecast predictive models to see if simpler models with less predicting variables could perform as well or better than more complex models. Simpler models are not only easier to interpret, but they can also do a better job at balancing tradeoffs between model accuracy and error.

An example graphical summary of forward selection, an example summary output, and a summary table of all the feature selection performed are shown below.

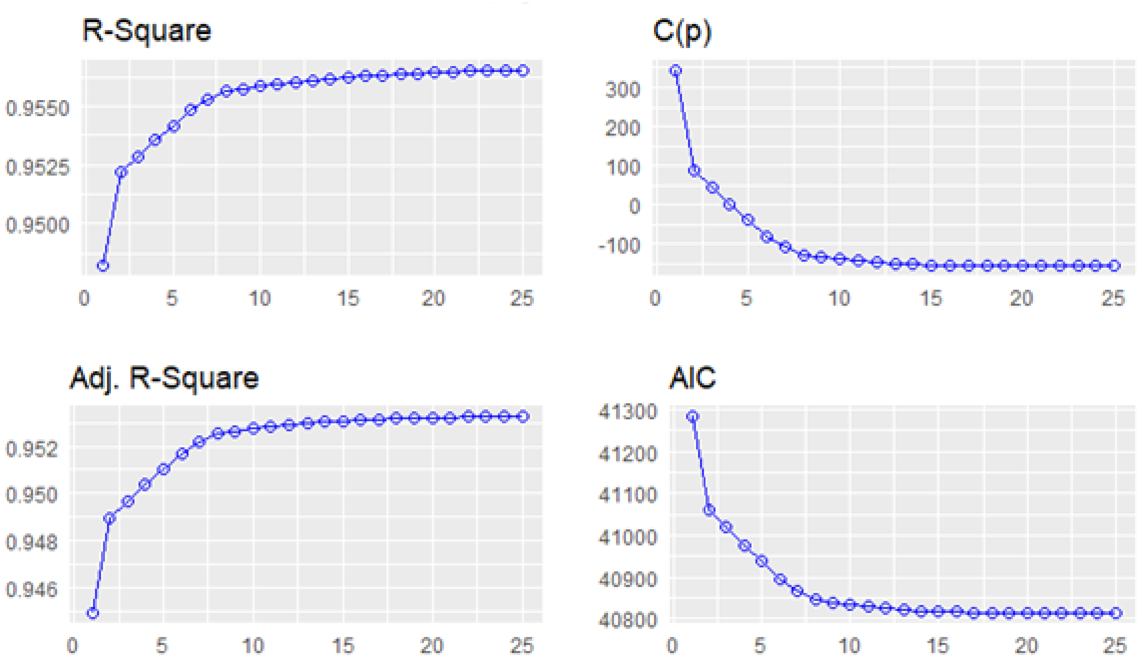


Figure 21. Graphical summary of R-square, adjusted R-square, Mallow’s Cp, and Akaike Information Criterion (AIC) values for forward selection using p-values for the after 2020 rent prices predictive model.

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	var_zip	0.9482	0.9449	345.2300	41283.527	203.8234
2	month	0.9522	0.9489	86.9473	41061.9579	196.2244
3	comp_ratio_1000_21.Collection.Truck.Noise	0.9529	0.9497	42.9866	41018.8533	194.8134
4	comp_ratio_1000_Loud.Music.Party	0.9535	0.9504	0.9803	40977.0328	193.4532
5	comp_ratio_1000_7.12.hours	0.9542	0.9510	-37.8361	40937.8182	192.1847
6	comp_ratio_1000_total	0.9548	0.9517	-79.1545	40895.4596	190.8261
7	comp_ratio_1000_Noise	0.9553	0.9522	-107.3962	40866.1048	189.8813
8	comp_ratio_1000_Noise..Barking.Dog..NR5.	0.9556	0.9525	-127.5324	40844.948	189.1953
9	comp_ratio_1000_13.18.hours	0.9558	0.9526	-133.5932	40838.4716	188.9658
10	comp_ratio_1000_Noise..Alarms..NR3.	0.9559	0.9528	-138.9197	40832.747	188.7598

11	comp_ratio_1000_Noise..air.condition.ventilation.equipment..NV1.	0.9560	0.9528	-143.3472	40827.954	188.5829
12	comp_ratio_1000_Noise..Ice.Cream.Truck..NR4.	0.9561	0.9529	-147.1053	40823.8557	188.4276
13	comp_ratio_1000_Loud.Talking	0.9561	0.9530	-149.2863	40821.42	188.3238
14	comp_ratio_1000_Noise..Jack.Hammering..NC2.	0.9562	0.9530	-151.4879	40818.9559	188.2191
15	comp_ratio_1000_Noise..Boat.Engine.Music.Etc...NR10.	0.9562	0.9531	-152.9507	40817.2719	188.1386
16	comp_ratio_1000_Noise..Construction.Equipment..NC1.	0.9563	0.9531	-154.1541	40815.8599	188.0665
17	comp_ratio_1000_Noise..Manufacturing.Noise..NK1.	0.9563	0.9531	-155.2058	40814.6058	187.9993
18	comp_ratio_1000_Noise..Boat.Engine	0.9564	0.9532	-156.1108	40813.505	187.9369
19	comp_ratio_1000_Car.Truck.Music	0.9564	0.9532	-156.7538	40812.6811	187.883
20	comp_ratio_1000_Car.Truck.Horn	0.9564	0.9532	-156.8196	40812.4721	187.848
21	comp_ratio_1000_Loud.Television	0.9565	0.9532	-156.8528	40812.2965	187.8142
22	comp_ratio_1000_Banging.Pounding	0.9565	0.9532	-156.3756	40812.6659	187.7971
23	comp_ratio_1000_Noise..Other.Animals..NR6.	0.9565	0.9532	-155.8256	40813.1125	187.7824
24	comp_ratio_1000_Noise...lawn.care.equipment..NCL.	0.9565	0.9532	-155.2804	40813.5532	187.7675
25	comp_ratio_1000_Engine.Idling	0.9566	0.9532	-154.4141	40814.3375	187.7632

Table 3. Forward selection output for after 2020 rent prices predictive model.

Predictive Model	**Number of Variables (including intercept)**	**Coefficients Added (+) / Removed (-)**	**Feature Selection Approach**	**Adj. R-square**
Rent Prices - Before 2020	223	31 complaint rates, 180 zip dummies, 11 months of year	-	0.9937
	212	(+): 20 complaint rates, 180 zip dummies, 11 months of year	forward (train)	0.9873
	212	(-): 11 complaint rates	backward (train)	0.9873
Rent Prices - After 2020	224	32 complaint rates, 180 zip dummies, 11 months of year	-	0.9668
	215	(+): 23 complaint rates, 180 zip dummies, 11 months of year	forward (train)	0.9532
	215	(-): 9 complaint rates	backward (train)	0.9532
Sale Prices - Before 2020	224	37 complaint rates, 175 zip dummies, 11 months of year	-	0.8177
	221	(+): 34 complaint rates, 175 zip dummies, 11 months of year	forward (train)	0.8857
	221	(-): 3 complaint rates	backward (train)	0.8857
Sale prices - After 2020	219	32 complaint rates, 175 zip dummies, 11 month of year	-	0.9886
	204	(+): 17 complaint rates, 175 zip dummies, 11 month of year	forward (train)	0.9886
	206	(-): 13 complaint rates	backward (train)	0.9886

Table 4. Summary table of feature selection performed on predictive models.

Our results show that for all but one of the forecast predictive models, the adjusted R-Squared value was higher in the more complex model than in the simpler models. The differences in adjusted R-square between complex models and simpler models was quite small with a maximum difference of ~ 7% observed between the full and reduced models for sale prices before 2020. Both forward selection and backward elimination tended to agree in terms of the number and type of variables that were included in the model as is shown in the feature selection summary table above (Table 4). In all models, the bulk of the explanatory power appears to come from the zip code and month variables, and they tend to be included in the model during the first few steps of forward selection. On the other hand, the complaint ratio variables did not contribute much to increasing the adjusted R-square value of the models and they were the only variables that were removed during backward elimination. These results seem to suggest that housing prices are weakly correlated with noise complaints which is the opposite of what we had initially hypothesized. However, it is also important to consider that our models could be overfit (which is why we're seeing such high R-square and adjusted R-square values) and multicollinearity still appears to be an issue even though we've tried to eliminate unnecessary variables through feature selection. The presence of multicollinearity can be quite problematic with respect to model interpretation as it can lead to unreliable estimated regression coefficients that can change dramatically when comparing different models with different numbers of predictors. Additionally, we may not be able to rely on p-values to detect whether coefficients are statistically significant because that variance of those estimated regression coefficients could be much larger than they normally would be. Future work to improve the models we've built could explore using different predicting variables so that multicollinearity is adequately addressed. Other feature selection approaches such as LASSO and Elastic Net could be explored and shed light on whether noise complaint data has value in explaining and/or predicting housing prices.

An additional consideration when choosing model features is whether the underlying behavior of the feature seems plausible. For example, in one case a variable in the 311 NYC calls dataset appears to have been categorized inconsistently, resulting in a count that declines substantially over time while counts of most other noise complaint variables increase.

Type	**Complaint Count Per Year (for all New York City)**											
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2021	2022
Total noise	192774	195526	220416	253336	333672	379757	413413	437069	430234	473192	761714	733641
Other noise	779	764	221	12	16	23	18	12	19	17	5	10

Table 5. Number of total and 'other' noise complaints by year in New York City, from 2010 to 2022 (excl. 2020)

Table 5 shows the counts of this variable, 'Other Noise Sources (Use Comments) (NZZ)', abbreviated as 'other noise' in the table, compared to the count of all noise complaints combined. While total noise complaints tend to increase over time (with a few minor exceptions), 'other' noise complaints begin at 779 in 2010, decrease substantially to 221 complaints in 2012, then plummet to only 12 complaints in 2013, remaining close to that number through subsequent years. Although there is a chance this phenomenon is real, it seems likely that complaints categorized as 'other' during early years were mapped to different categories over time as the process improved; perhaps new categories were added based on the most common types of complaints appearing in the 'other' category.

Because of this issue, as well as its low sample sizes during most years (especially since it is disaggregated into zip codes for the models), this variable was removed from our models. However, it is important to note that this issue may be present to a lesser extent in other variables, and any further research would benefit from a more thorough investigation of the complaint type variables than time allowed for this project.

CONCLUSIONS

We were able to build explanatory linear models and forecast predictive models with high adjusted R-Squared and low mean absolute error values. However, in most of the cases the fit was because of the zip code dummy variables and the month of the year categorical variable which were intended to control for the specific regulations and conditions of each zip area, and for the seasonality of the complaints' time series. Even though the adjusted R-Squared is high, the practical application of using noise complaints to explain or predict prices does not appear to be useful for all rent pricing models (before and after 2020), and the sales price models before 2020.

Nonetheless, the housing sales price models after 2020 (both the explanatory model and the forecast predictive model), showed a relatively good performance, with several complaints' features having large coefficients. This could be because the trend of the sales prices and the complaint time series are similar. Moreover, in some cases the changes in the rate of noise complaints appeared to precede changes in sales prices, indicating a potential predictive relationship between the two variables. Another interesting result was that the sign of the complaints' coefficients varied depending on the model and the type of complaints. In our best model (sales price after 2020), general noise complaints had a positive coefficient, implying that on average the price increases when the rate of complaints is increased. It could be that noise complaints are acting as a proxy and are capturing other exogenous factors such as economic activity, construction with some of the complaint types.

Further research must be made on the stability and significance of that relationship, but we seemed to have found a useful predictor that can be incorporated to forecast models as an alternative non-traditional source, **but only applicable to housing sales prices after 2020. In general, our model predicts that the average sales price will fall in the next months, with the average going down by 2.2%. 125 out of 175 are predicted to have a lower price and 50 will have a higher price.**

FURTHER RESEARCH

One potential area of focus for further research could be exploring the relationship between noise complaints and other external factors, such as economic activity, to enhance the accuracy of housing price prediction models. While the use of noise complaints as a predictor variable showed some promise in our study, the lack of enough data limited our ability to identify a strong relationship between noise complaints and housing prices. Conducting similar studies across other major cities in North America or other regions with similar characteristics could provide a more comprehensive understanding of the relationship between noise complaints and housing prices, and whether this relationship varies based on specific local factors. Furthermore, incorporating other non-traditional data sources or utilizing difference-in-differences analysis to explore changes before and after COVID-19 could offer a more comprehensive understanding of pricing dynamics in real estate.

Another potential avenue for future research could be to explore the impact of housing supply and demand on housing prices. Understanding the balance between supply and demand in a particular area can provide valuable insights into the drivers of housing prices. In addition, investigating the impact of location factors, such as proximity to schools, transportation, shopping, and amenities, could help to identify the specific factors that drive housing prices in different areas.

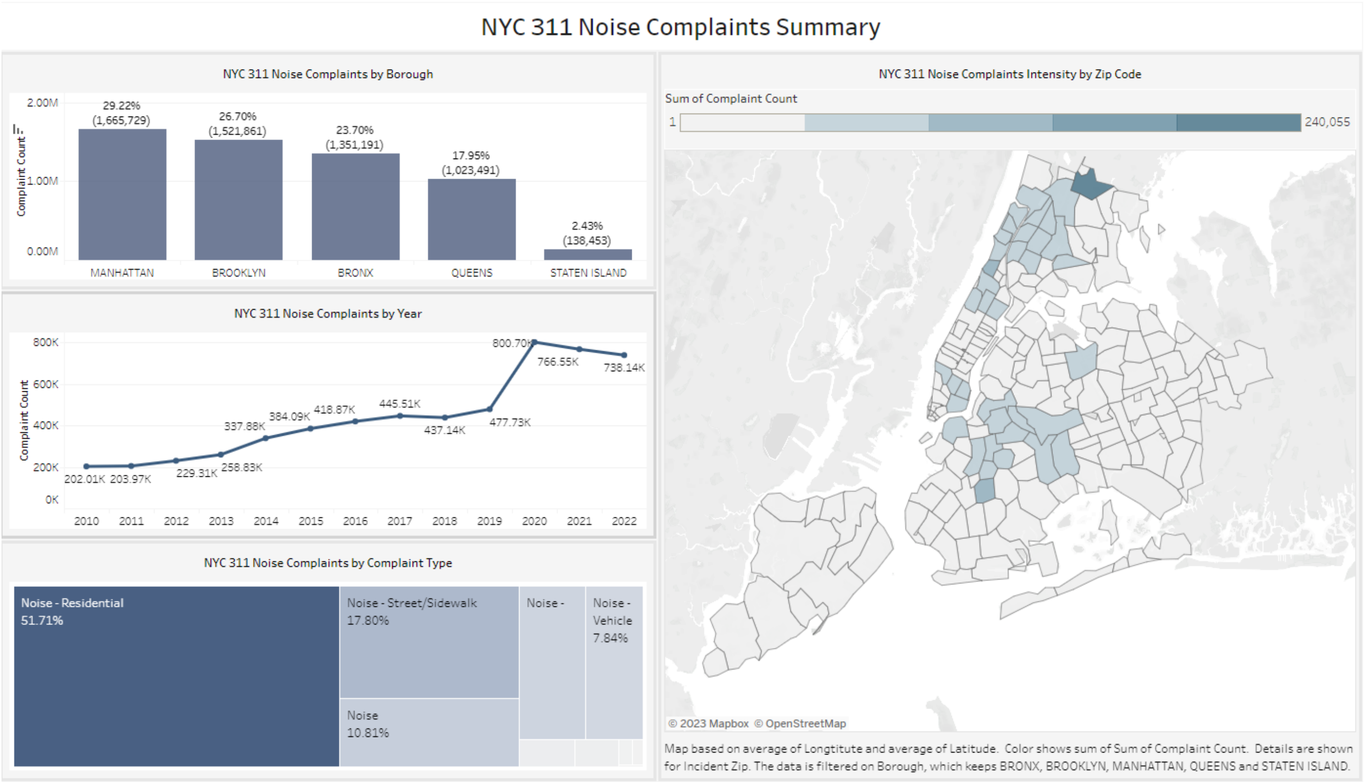
Economic indicators, such as interest rates and inflation, can also be important predictors of housing prices. Incorporating these variables into the models could help to capture the impact of broader economic trends on housing prices. Demographic factors, such as age distribution, can also be important drivers of housing demand and could be incorporated into the models to improve their accuracy. Crime rates and environmental factors, such as air quality, water quality, and natural hazards, could also be explored as potential predictors of housing prices.

Overall, there are several potential areas for future research to enhance the accuracy of housing price prediction models. The selection of variables depends on the research question and the availability and relevance of data.

TABLEAU DASHBOARD WITH RESULTS

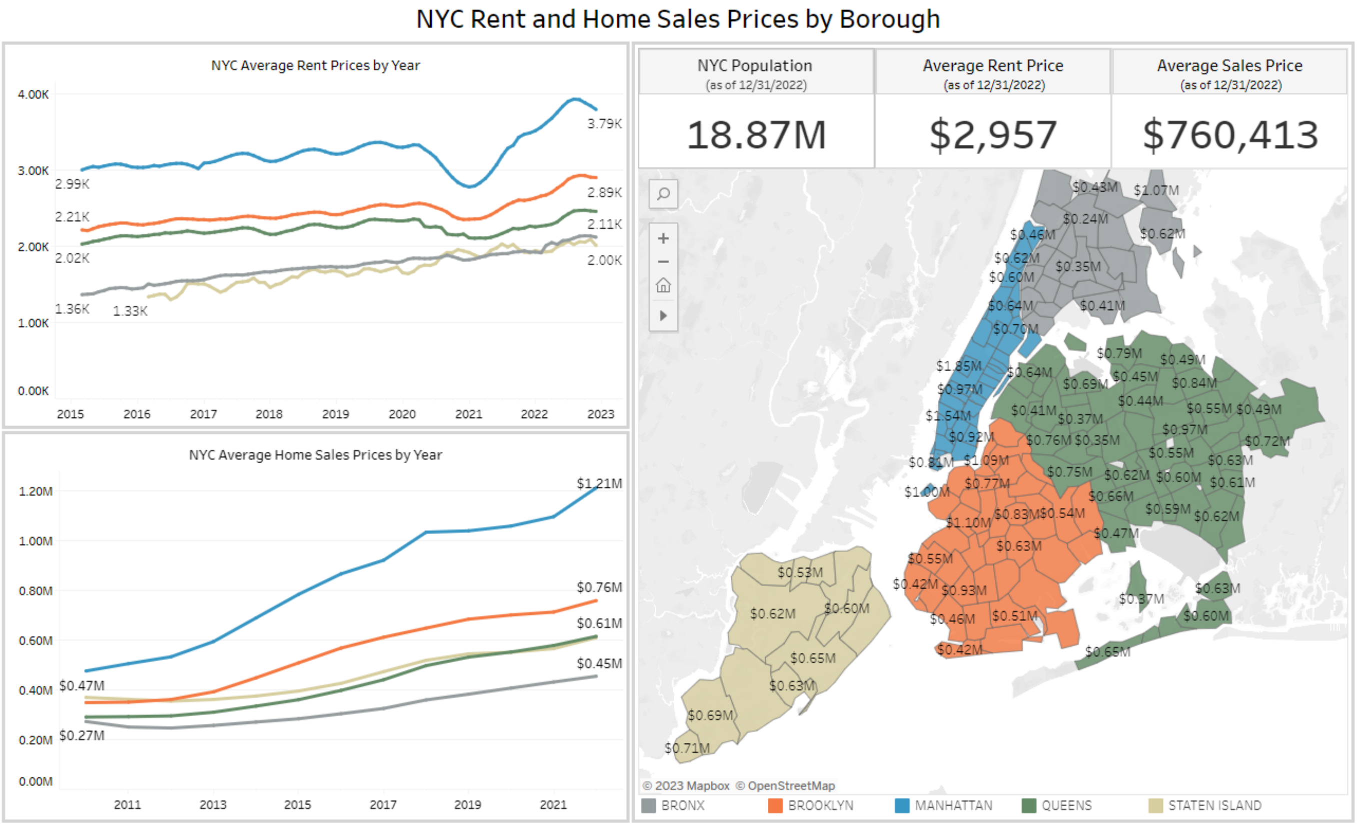
The results of the exploratory data analysis and sales prices predictions using the above linear regression models were summarized in the interactive Tableau dashboard.

Dashboard View 1: NYC 311 Noise Complaints Summary

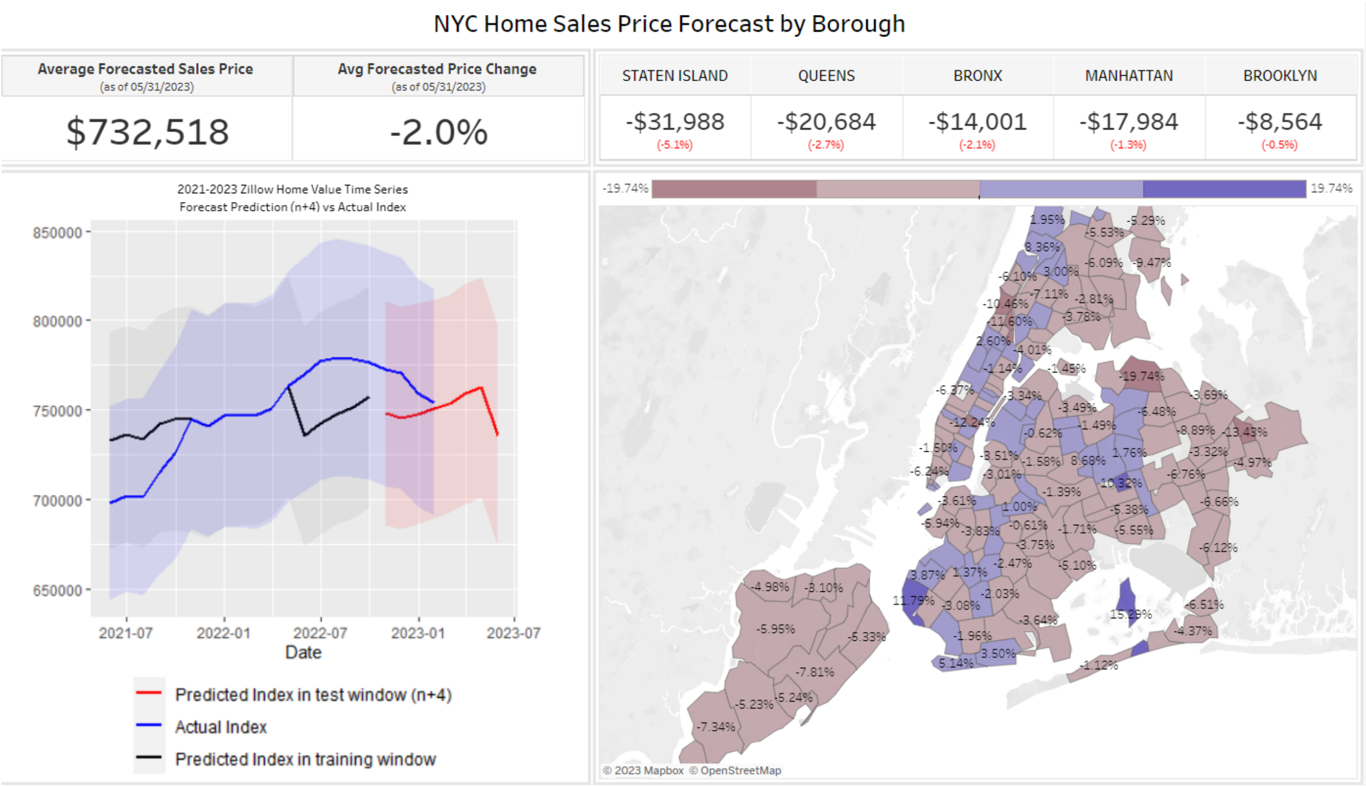


Map based on average of Longtitute and average of Latitude. Color shows sum of Sum of Complaint Count. Details are shown for Incident Zip. The data is filtered on Borough, which keeps BRONX, BROOKLYN, MANHATTAN, QUEENS and STATEN ISLAND.

Dashboard View 2: NYC Rent and Homse Sales Prices by Borough



Dashboard View 3: NYC Rent and Home Sales Predicted Prices by Borough



NYC Home Sales Price Forecast by Borough

Average Forecasted Sales Price (as of 05/31/2023)	Avg Forecasted Price Change (as of 05/31/2023)
$732,518	-2.0%

STATEN ISLAND	QUEENS	BRONX	MANHATTAN	BROOKLYN
-$31,988 (-5.1%)	-$20,684 (-2.7%)	-$14,001 (-2.1%)	-$17,984 (-1.3%)	-$8,564 (-0.5%)

PROJECT TIMELINE/PLANNING

Assignment	Date
Complete draft of project proposal	Mon 3/6
Group meeting to revise proposal draft	Mon 3/6
Meet with TAs to discuss proposal	Tues 3/7
Work on revising/finalizing proposal	Weds 3/8 – Sun 3/12
Turn in project proposal	Sun 3/12
Work on project proposal video	Mon 3/13 – Sun 3/26
Turn in proposal video (4-5 min)	Sun 3/26
Work on project progress report	Mon 3/27 – Sun 4/2
Turn in progress report (4-5 pgs)	Sun 4/2
Work on final report	Sun 4/2 – Sun 4/16
Work on final presentation video	Sun 4/2 – Weds 4/19
Turn in final report (8-10 pgs)	Sun 4/16
Turn in final presentation video (10-12 min)	Weds 4/19
Turn in final slides, code, data, and any other materials	Weds 4/19
Work on outside-of-group peer review	Thurs 4/20 – Weds 4/26
Turn in peer review	Weds 4/26
Complete and turn in within-group performance evaluation	Weds 4/26

REFERENCE LIST

Ariel Property Advisors. (2023, January). *Multifamily Year In Review 10+ Residential Units: New York City | 2022.* Retrieved from https://arielpa.com/report/report-MFYIR-2022

Forbes. (2023, March 5th). *Partying Like It's 2015: Multifamily Housing In New York City.* Retrieved from https://www.forbes.com/sites/shimonshkury/2023/02/09/partying-like-its-2015-multifamily-housing-in-new-york-city/?sh=16bc2efe2736

Hammer, M. S., Swinburn, T. K., & Neitzel, R. L. (2014). Environmental noise pollution in the United States: developing an effective public health response. *Environmental health perspectives*, *122*(2), 115-119.Retrieved from https://ehp.niehs.nih.gov/doi/pdf/10.1289/ehp.1307272

McKinsey & Co. (2018, October 8th). *Getting ahead of the market: How big data is transforming real estate*. Retrieved from https://www.mckinsey.com/industries/real-estate/our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate#

McKinsey & Company. (2022, March). *McKinsey Global Private Markets Review 2022.* Retrieved from https://www.mckinsey.com/~/media/mckinsey/industries/private%20equity%20and%20principal%20investors/our%20insights/mckinseys%20private%20markets%20annual%20review/2022/mckinseys-private-markets-annual-review-private-markets-rally-to-new-heights-vf.pdf

New York City Department of Health and Mental Hygiene (2014). *Ambient Noise Disruption in New York City.* Retrieved from https://www.nyc.gov/assets/doh/downloads/pdf/epi/databrief45.pdf

New York City Department of Health and Mental Hygiene (2013). *Preventing noise-induced hearing loss among young people*. Retrieved from https://www.nyc.gov/assets/doh/downloads/pdf/epi/databrief45.pdf

New York City Department of Housing. (2022, May 16th). *2021 New York City Housing and Vacancy Survey*. Retrieved from https://www.nyc.gov/assets/hpd/downloads/pdfs/services/2021-nychvs-selected-initial-findings.pdf

New York City Open Data. (2023, March 5th). *311 Service Requests from 2010 to Present*. Retrieved from https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

New York City Open Data. (2023, March 5th). *DOF: Summary of Neighborhood Sales by Neighborhood Citywide by Borough*. Retrieved from https://data.cityofnewyork.us/City-Government/DOF-Summary-of-Neighborhood-Sales-by-Neighborhood-/5ebm-myj7

PropertyShark - Yardi Systems, Inc. (2023, March 9th). *Market Trends*. Retrieved from https://www.propertyshark.com/mason/market-trends/residential/nyc-all

United States Census Bureau. (2023, March 5th). *S0101 ACS 5 Year Estimates Subject Tables*. Retrieved from https://data.census.gov/table?g=0100000US$8600000&tid=ACSST5Y2021.S0101

United States Internal Revenue Service - IRS. (2023, March 5th). *SOI Tax Stats - Individual Income Tax Statistics - 2020 ZIP Code Data (SOI)*. Retrieved from https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2020-zip-code-data-soi

World Health Organization. (2011). *Burden of disease from environmental noise: Quantification of healthy life years lost in Europe*. World Health Organization. Regional Office for Europe. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/326424/9789289002295-eng.pdf?sequence=l&isAUowed=y

Zillow, Inc. (2023, March 5th). *Zillow Housing Research Data*. Retrieved from https://www.zillow.com/research/data/

Zillow Inc, Broakerage. (2023, March 9th). *New York, NY Rental Market*. Retrieved from https://www.zillow.com/rental-manager/market-trends/new-york-ny/

Ramphal, B., Dworkin, J.D., Pagliaccio, D., Margolis, A.E. (2022). Noise complaint patterns in New York City from January 2010 through February 2021: Socioeconomic disparities and COVID-19 exacerbations. *Environmental Research,* 206 112254. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0013935121015553

Beracha, E., Wintoki, M.B. (2013). Forecasting Residential Real Estate Price Changes from Online Search Activity. *Journal of Real Estate Research*, 35(3), 283-312. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/10835547.2013.12091364