MGT 6203 Team 10 Progress Report

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

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's article, "Getting ahead of the market: How big data is transforming real estate", using traditional independent variables to predict property market values (i.e., year built, rooms, location) 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, leading to lower prices.

If the hypothesis is true, this information can be used to enhance existing market research and predictive toolkits. By better understanding the factors that impact price and improving its predictability, city planners, developers, and property owners can make better business decisions. They can identify novel areas with potential value and avoid investing in areas where the model predicts value erosion. The model can also improve pricing by considering factors that are relevant to tenants and buyers: **Assuming a 10% addressable market, property owners and developers selling $2.6 billion a year in houses 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 model will also enable the identification of types of noise complaints that impact prices and the magnitude of their effects. The betas from the model can 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 listing their properties, as well as improving the search experience for their users, and potentially driving more traffic to their websites, resulting in increased revenue.

PROGRESS OVERVIEW:

The initial phase of the project has allowed us to better understand the problem, get additional sources as detailed in the survey and advanced in the extraction, cleaning and preliminary modeling of the datasets. The extraction of the datasets was completed successfully, we were able to create a join dataset panel including the Zillow sales and rent index, population, income, and complaint rate per type of complaint for 177 zip codes in New York City. Exploratory data analysis and visualizations have helped us to better understand the pattern of noise complaints in each of the boroughs. An analysis of the time series showed that the complaints dataset has strong seasonality and trend, while the price dataset has trend, it also evidenced the presence of outliers. As part of the data cleaning process, the datasets were transformed to remove seasonality, trend, outliers, and exclude zip codes that represent office buildings without residential units.

LITERATURE SURVEY:

In addition to researching the New York City real estate market and the extent of the noise pollution problem in the city (as discussed above), we also conducted a small literature survey of peer-reviewed papers investigating two specific topics related to our project: (1) the analysis of noise complaints in NYC (using data from the NYC 311 dataset), 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 that they increased dramatically during the first year of the COVID-19 pandemic, with that increase highest among people at lower income levels (Ramphal et al., 2022). This suggests that the noise problem in the city is likely worsening, which makes our project especially relevant. And the fact that increases in noise complaints per capita are higher in lower income groups seems to support our 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 article on 'non-traditional' real estate price prediction factors mentioned previously. 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). Overall, these two studies helped convince us that our project is worthwhile, and that its business justification is strong. Also, the methodologies employed in both gave us a better sense of how to analyze and predict real estate prices and their correlation to noise complaints, which will help to guide us as our project progresses.

ANTICIPATED CONCLUSIONS/HYPOTHESIS:

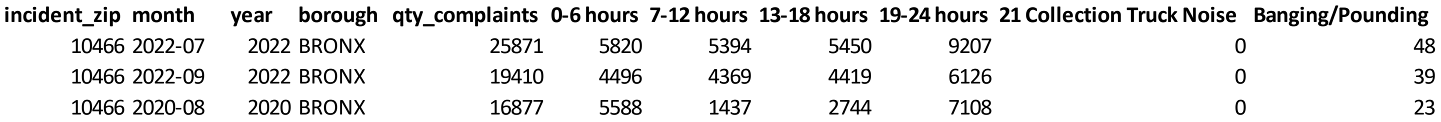
We hypothesize that housing and rent prices in New York City neighborhoods are negatively correlated with the per capita number of noise complaints in those neighborhoods, and that the relationship is statistically significant. We expect this to be the case especially since the pandemic began, since the increased number of remote workers and increased time spent at home has likely made individuals more sensitive to noise. We also predict that there could be a different relationship between noise complaints and housing/rent prices among different income levels; thus, we will try using separate models for different income levels to determine whether that approach yields better-fitting models.

If our hypothesized correlation is found to exist, we plan to further explore the relative timing of significant changes in noise complaints and changes in housing prices. If, for example, noise complaint increases precede housing/rent price decline (instead of occurring after those declines), then it may be possible to use noise complaints as a predictor or early warning sign of potential price deterioration in that region.

We will also analyze types of noise complaints and the time of day when each complaint was made. Certain 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 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 categorized the income level in three tiers using equal ntiles (high income, medium income, low income). We created a dummy variable per zip code as explained in the model description. The resulting dataframe has 27.483 rows and 279 columns.

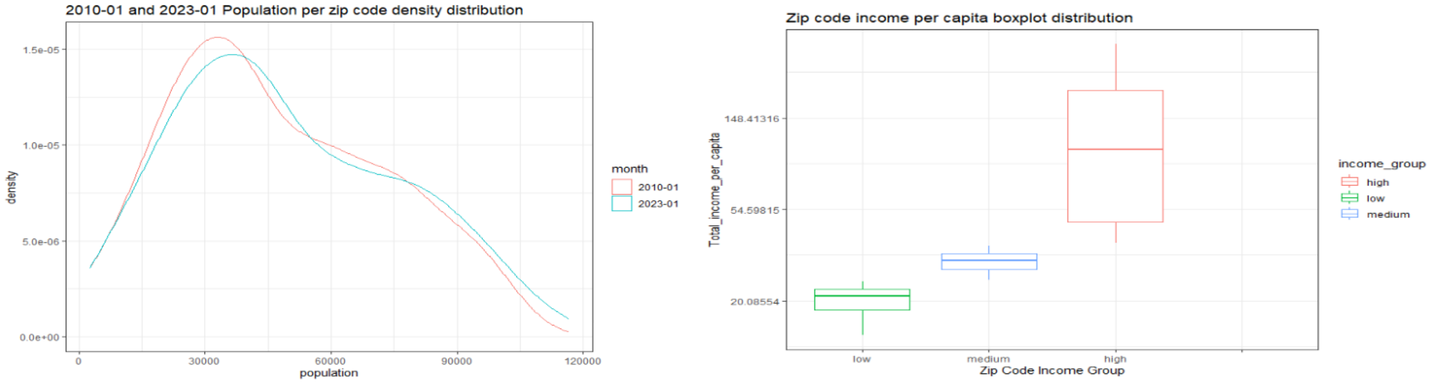


Figure 4. Population and income per capita distribution

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, SEASONALITY, TREND, AND OUTLIERS

The dataset included datapoints from 145 months 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.

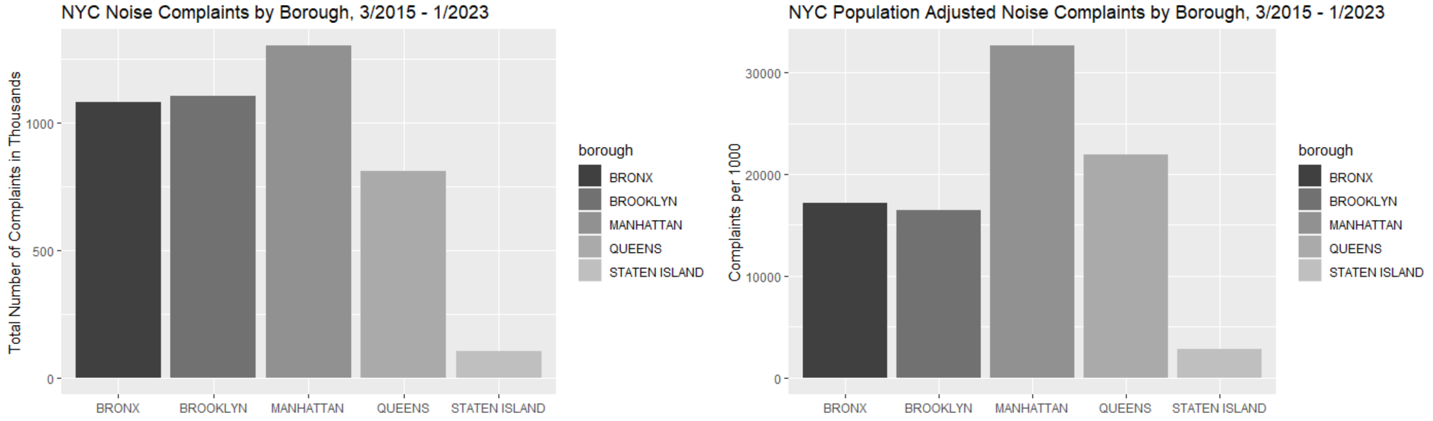


Figure 6. Total Number of Complaints and Population Adjusted Complaints by NYC Boroughs

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.

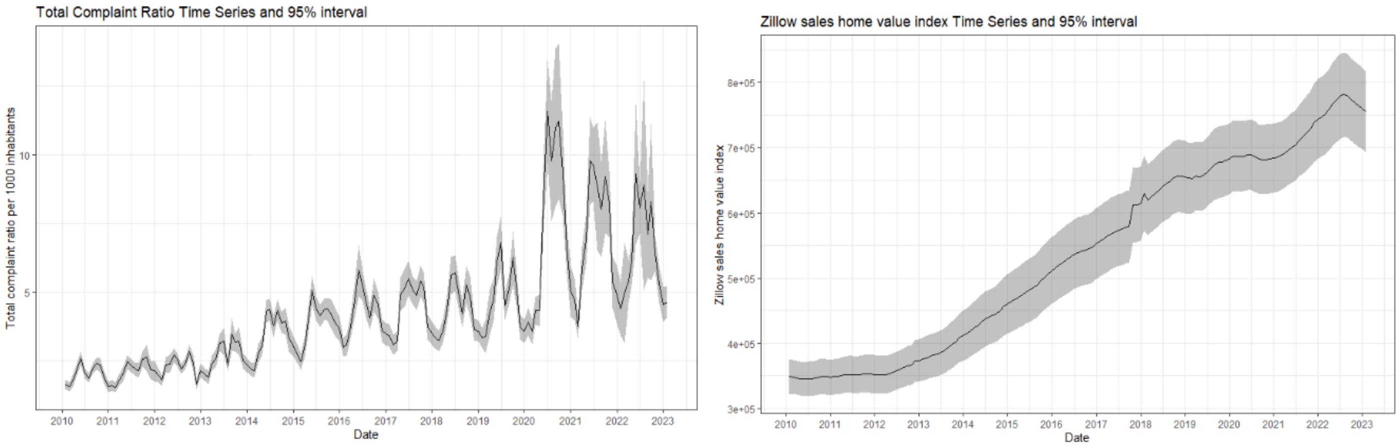


Figure 7. Complaint ratio per 1000 inhabitants and Zillow home sales index time series with 95% interval

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.

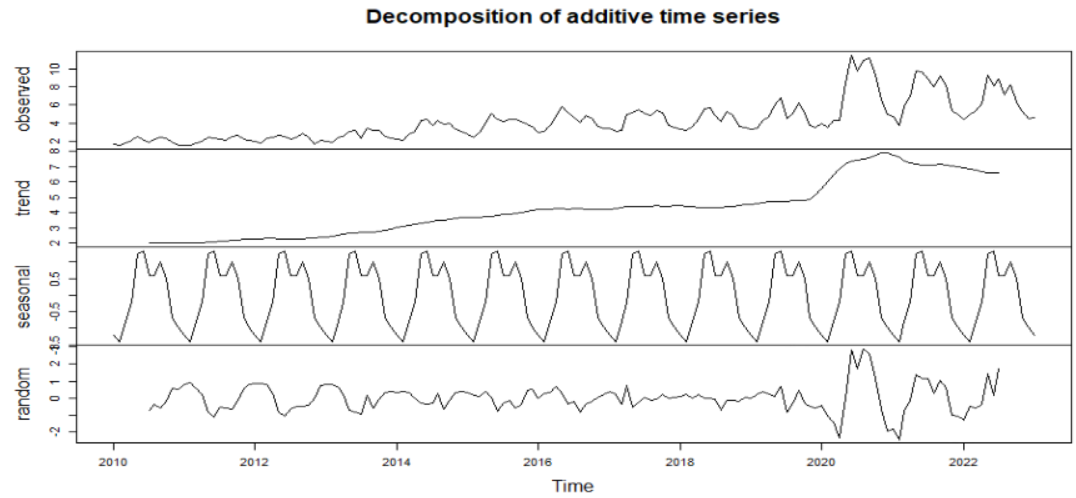


Figure 8. Decomposed time series from the average total complaint rate per month

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:

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

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.

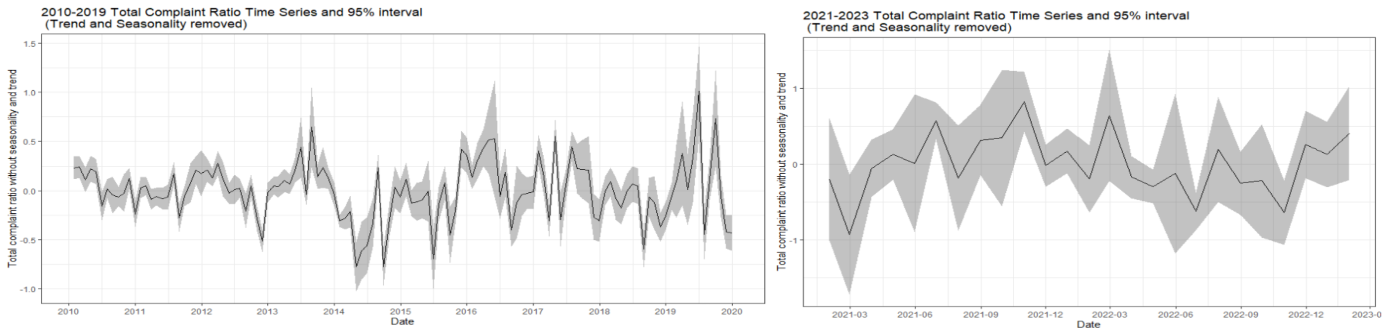


Figure 9. 2010-2019 and 2021-2023 complaint ratio time series after removing seasonality and trend

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.

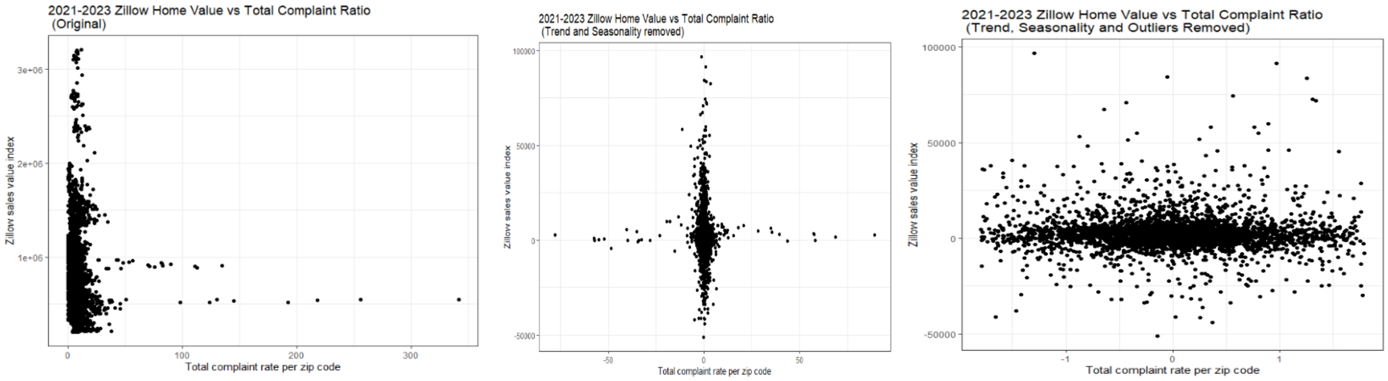


Figure 10. 2010-2019 and 2021-2023 complaint ratio time series after removing seasonality and trend

INITIAL CONCLUSIONS AND CHALLENGES

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 raw complaint ratio, together with the month of the year, and the zip code as independent variables will most likely result in in very high R2 values, however, the marginal contribution of the complaint ratio itself will most likely be low, hence resulting in low estimator values or even lack of statistical significance. The complaint time series also has a very interesting property; once seasonality is removed, it seems to anticipate the change of trend in the prices (excluding 2020). If this is corroborated by models, the gradient model has a potential to become a valuable business tool to predict pricing trends as suggested in the article by McKinsey.

MODEL DESCRIPTION:

In our model, we conduct a regression analysis of the average rental price and residential property prices in a given year within a Zip Code while accounting for the impact of population growth. We consider population growth as a suitable proxy to account for natural changes in rental income that are not related to noise complaints. Additionally, we will explore detrending the average rental income and using it as an independent variable. However, this approach may result in a loss of information that could potentially weaken the impact of noise.

To avoid the over-representation of income in our analysis, we will employ separate regression models for each income level, recognizing that the relationship between noise and rental income may differ across income levels. Income levels will be categorized based on the number of standard deviations from the average income level in New York City (i.e., threshold of three standard deviations below the average income level will define the low-income group). We will also consider incorporating higher dimensions of income as an explanatory variable to account for non-linear relationships at the extreme ends of the income spectrum.

We will include Zip Code as a dummy variable in our regression to account for potential variations among Zip Codes, such as differing laws and regulations. However, we acknowledge that this approach may adversely affect the performance of our model. From an economic research standpoint, controlling for these variables is still recommended, but our primary focus is on prediction performance. Therefore, we will also seek access to data on rent control and rent stabilization policies to control for the variation due to different rental policies.

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.

We are aware that depending on the income level, the relationship may be affected. Therefore, we plan to run separate models for different income levels.

Regression model with Zip Code as a proxy for each income level:

$$Rnt_{iym} \ = \beta_0 + \beta_1 TNC_{iym} + \Sigma\,_2^j \beta_j NC_{ijym} + \rho_i I + \Sigma\,_2^k \gamma_k \cdot M_{month_m} + \epsilon_i$$

$$Pr_{iym} \ = \beta_0 + \beta_1 TNC_{iym} + \Sigma\,_2^j \beta_j NC_{ijym} + \rho_i I + \Sigma\,_2^k \gamma_k \cdot M_{month_m} + \epsilon_i$$

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

Gradient (change) predictive model to predict rent and sales prices n months later:

$$\Delta Rnt_{iy(m+n)} \ = \beta_{0(income\ of\ i)} + \beta_{1(inc\ of\ i)} \Delta TNC_{iym} + \Sigma\,_2^j \beta_{j(inc\ of\ i)} \Delta NC_{ijym} + \rho_i I + \Sigma\,_2^k \gamma_{k(inc\ of\ i)} \cdot M_{month_m} + \epsilon_i$$

$$\Delta \mathrm{Pr}_{iy(m+n)} = \beta_{0(income\ of\ i)} + \beta_{1(inc\ of\ i)} \Delta TNC_{iym} + \Sigma\,_2^j \beta_{j(inc\ of\ i)} \Delta NC_{ijym} + \rho_i I + \Sigma\,_2^k \gamma_{k(inc\ of\ i)} \cdot M_{month_m} + \epsilon_i$$

- 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

The overarching concept remains consistent; however, our team is currently exploring alternative methods to better understand the relationship between noise complaints and rental or real-estate prices. One approach we are implementing involves performing data cleaning techniques to remove any underlying trends or seasonal variations within the data. The cleaned data will then be utilized in a regression model to better analyze the relationship between noise complaints and prices.

Another significant modification we have made is acknowledging that changes in noise complaints are unlikely to have an immediate impact on rental or real-estate prices. In fact, the average search time for individuals seeking new rentals in New York can be up to 32 days, 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 have decided to test various lag periods to observe how pricing is influenced by the number of noise complaints. This will allow us to determine the most effective lag period to use in our analysis and gain a more comprehensive understanding of the relationship between noise complaints and prices in the real-estate market. In order to assess the impact of different lag periods on the significance level of the noise complaints variable, we will systematically add varying lag periods to our regression model. Additionally, we will examine the R-squared value to evaluate the overall effectiveness of the model's fit. Given that the time delay between noise complaints and changes in pricing may vary across different income groups, we anticipate that different lag periods may be necessary for each group. Therefore, we plan to conduct separate analyses for each income level and explore the optimal lag period for each group. By conducting these analyses, we hope to gain a more nuanced understanding of how noise complaints impact pricing across different income brackets.

Furthermore, we have already conducted preliminary analyses comparing the effectiveness of our models with and without lag periods. Our initial findings indicate that incorporating a lag period yields more insightful results. However, in order to ensure the robustness and validity of our findings, we have yet to finalize our analyses and are not yet publishing our results.

In addition to assessing the optimal lag period for our regression models, we also plan to determine the elasticity of price on noise complaints across different income configurations. By examining these dynamics, we hope to gain a deeper understanding of the impact of noise complaints on real-estate prices across various income levels. Ultimately, these analyses will contribute to a more comprehensive understanding of the complex relationship between noise complaints and real-estate prices in the market.

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