Brooklyn Houses Pricing Prediction
CSP 571 - Data Preparation and Analysis
Course Project Report

Dr. Jawahar Panchal
Department of Computer Science
Illinois Institute of Technology, Chicago

Pavani Reddy Podduturi [A20545675]

Table of Contents

Abstract

We will use linear regression to explain the housing prices in the borough of Brooklyn, New York, from the years of 2016-2020. We will prepare and analyze the data provided by the City of New York for these five years to understand how the prices varied overall as well as trying to understand how the prices varied for a very specific period. Also, since the borough of Brooklyn is densely populated, we will only be considering single-unit apartments, condos, and single-family residences for this project.

Introduction

The housing market in Brooklyn, New York, has long been a subject of interest and scrutiny due to its dynamic nature and the borough's significance within the larger real estate landscape of New York City. Over the years, factors such as demographic shifts, economic trends, and urban development projects have all played crucial roles in influencing housing prices. This study delves into an in-depth analysis of Brooklyn housing prices from 2016 to 2020. We will prepare and analyze the data provided by the City of New York for these five years to understand how the prices varied overall, as well as to understand how the prices changed for a very specific period. By examining data provided by the City of New York, we aim to uncover the underlying patterns and determinants that have shaped the housing market within this specific time frame.

To ensure a focused analysis, we narrow our attention to single-unit apartments, condos, and single-family residences. These property types constitute a significant portion of Brooklyn's housing stock and are known to exhibit distinct pricing dynamics compared to multi-unit or commercial properties. By concentrating on these residential units, we aim to provide a comprehensive understanding of the trends and forces driving the housing market for individuals and families within the borough.

The importance of this study lies in its potential to offer valuable insights to various stakeholders, including potential homebuyers, real estate investors, policymakers, and urban planners. Understanding the nuances of housing price trends in Brooklyn can inform decision-making processes related to property investments, urban development strategies, and housing policies. Moreover, this research contributes to the broader discourse on urban economics and real estate

dynamics, shedding light on how local factors interact with broader economic trends to influence housing markets within dynamic urban centers like Brooklyn. Through rigorous data analysis and interpretation, we seek to unravel the complex web of variables that have shaped housing prices in this vibrant borough.

Goal

The ultimate goal of this research is to get a better understanding of how a major borough's residential market ties in with the larger metropolitan statistical area's story. Being able to apply a model on an existing set of real-time data and understanding how each variable affects the fluctuations in the market will help us delve deeper into getting a better understanding of how to predict future prices and ultimately understand why Brooklyn is the way it is and how that affects New York City's housing market as a whole. Now, as the project continues; and based on further findings, the goal may deviate from the initial path in order to accommodate any supplemental information.

Proposed Methodology

As mentioned in the project proposal document earlier, we stated that we will make use of linear regression on our data for five years (2016-2020. Pre-Covid), to understand which variables had the highest effect on the housing market and how they changed over time. In order to achieve this, it is important to lay down a few rules. First, we divide each year into quarters (Labeled Q1 through Q4). This will help us get a better understanding of what happened in each quarter and whether that change is dependent on any other socio-economic developments in the region during that time period. Second, we will divide our data by performing a split. A major portion of it will be utilized for training the linear regression model while the remaining will be used for validation purposes.

Next, it is important to distinguish between each appropriate variable that will have an effect on our project. Variables include quantities such as square footage, number of floors, land value, the tax class of the individual, zip code, etc. We will use the model for data extraction, initial data analysis, data cleaning, and feature engineering methods. With the removal of white spaces, commas, various marks, and dashes, as well as verifying that the formats of various indications

were correct, some columns in the data might be deemed unnecessary and hence will be excluded from the final output. Finally, we will make use of graphs, boxplots, histograms, etc. to visualize how this will look and then deliver a detailed project report with our findings and thoughts.

Data Analysis

Conducting a thorough examination of the brooklyn housing dataset for the years 2016-2020 demanded a meticulous approach to data preprocessing and analysis, ensuring the dataset's coherence and reliability. Initial steps involved the removal of redundant columns, making sure the dataset's alignment with this project’s objectives. To enhance interpretability, column names were revised, giving a far more insightful dataset.

The pursuit of data quality commenced with the elimination of null values and the optimization of data types for efficient analysis. String values, when present, underwent a conversion to numeric data types, facilitating numerical operations and enabling better statistical computations. We also directed our attention towards numeric columns containing commas, necessitating their removal to transform them into suitable formats for subsequent analyses.

Each housing dataset file was then imported into individual data frames based upon the already specified format, setting the course for cleaning procedures. Leveraging the capabilities of the date class emerged as a key refinement in our project, where we meticulously restructured the dates pertaining to house acquisitions. Subsequently, our focus shifted to enriching the temporal aspect of the dataset. We accomplished this by segmenting dates into quarters while preserving the granularity of the original month information. This nuanced strategy brings a heightened level of sophistication to our temporal analysis, facilitating a more detailed exploration of trends. It positions us to conduct insightful analyses, honing in on subtleties like seasonality patterns and year-over-year variations, crucial aspects for our project's comprehensive understanding.

Initiating these systematic procedures, we integrated the individual data frames into a consolidated dataframe, shaping a completely new dataset crucial for our project. This formed the bedrock of our analysis, presenting a comprehensive view of housing trends over the

designated five-year period. The successful outcome provides us with invaluable insights into the dynamic landscape of property acquisition and the nuanced fluctuations in property values. This contribution significantly augments our comprehension of the dynamics within the housing sector for the borough of Brooklyn.

Data Processing

The labeling method that we employed to standardize our data consisting of string and numeric types is given in the table below. It helps a lot in the data-cleaning process:

neighborhood	The neighborhood in Brooklyn
bldclasscat	Building Class Category
taxclasscurr	Tax Class at Present
block	Block
lot	Lot
easement	Easement
bldgclasscurr	Building Class at Present
address	Address
aptnumber	Apartment Unit Number
zip	Zip code
resunits	Number of Residential Units
comunits	Number of Commercial Units
totunits	Number of Total Units
landsqft	Total land square footage

grosssqft	Total gross built-up square footage
yrbuilt	Year Built
taxclasssale	Tax Class at the time of sale
bldgclasssale	Building class at the time of sale
price	Sale Price
date	Sale Date

Once we consolidate all the files into one and apply this labeling method to standardize the date, it makes it easier to find information without any confusion and then operate on it. Also, in the cleaning step, we will eliminate any $0 purchases as that value is far below the market average and it generally means that it was a deal between close family members.

From the dataset we obtained for the housing data from the borough of brooklyn in New York City for the years between 2003-2020, we decided to focus on the years 2016-2020 for this study, primarily because this was the time before the coronavirus pandemic, but it also includes two years of the pandemic (2019 & 2020). This set of years gives us a proper insight into how the housing market changes due to unforeseen circumstances such as a global pandemic, but also the yearly or quarterly changes based on the city's and the borough's progress.

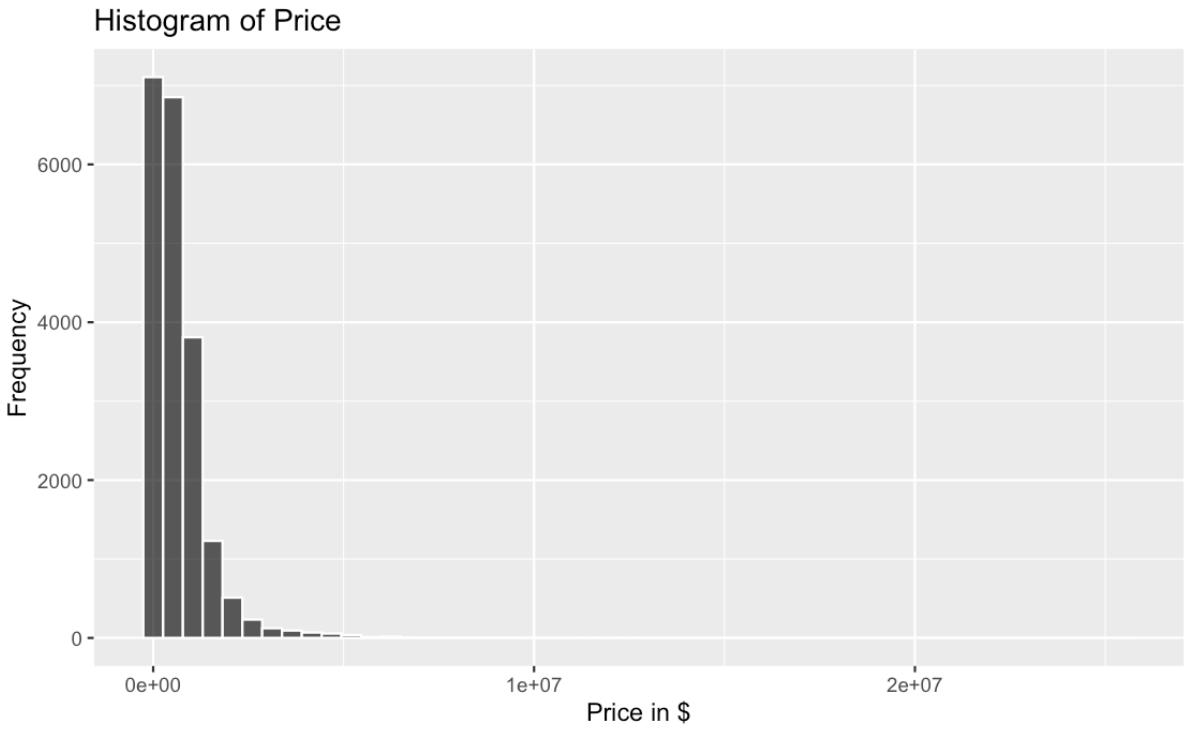
Now, with regard to the data we collected, we realized after looking through the respective files that it would be best if we generalized each column and consolidated its appropriate values. So, the first thing we did in the data preprocessing phase was to thoroughly go through each file, and find out which columns can be consolidated, which can be omitted, etc. A thorough analysis of the data helped us minimize most downstream dependency related errors. Once the thorough analysis was conducted, we gave each relevant column a name, so it would be easier for us to refer to it and then make any necessary changes, should we choose to do so.

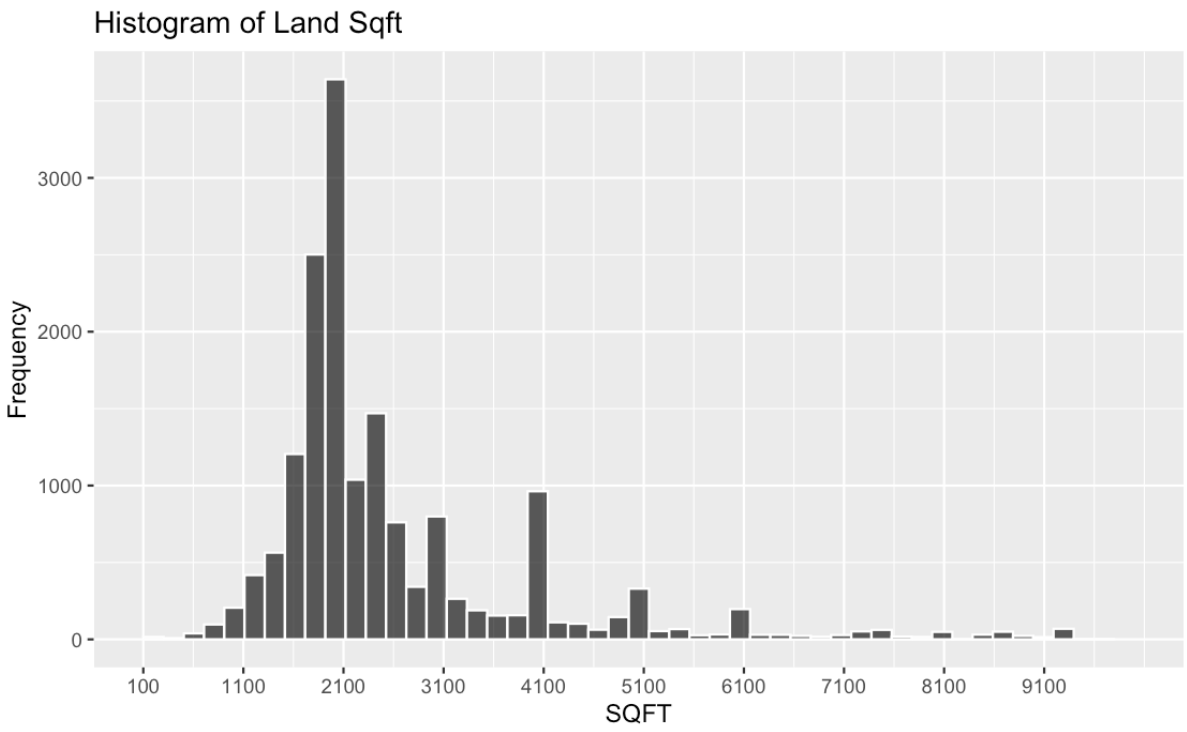
The next step was to import each individual file into the environment. Importing it into a dataframe tailored to our naming conventions. After importing the data for years 2016-2020, we created five separate dataframes to store each individual year's data before binding all those data frames together to create one final data frame. This data frame had 119,374 observations; a big set of data to work with and operate on. After generally binding all five dataframes, we did a thorough analysis of the resulting data frame to get a better understanding of how to minimize any issues that might have future impact. We concluded that the "price" variable is our dependent variable, so we chose that and decided to make adjustments to the dataframe based on that.

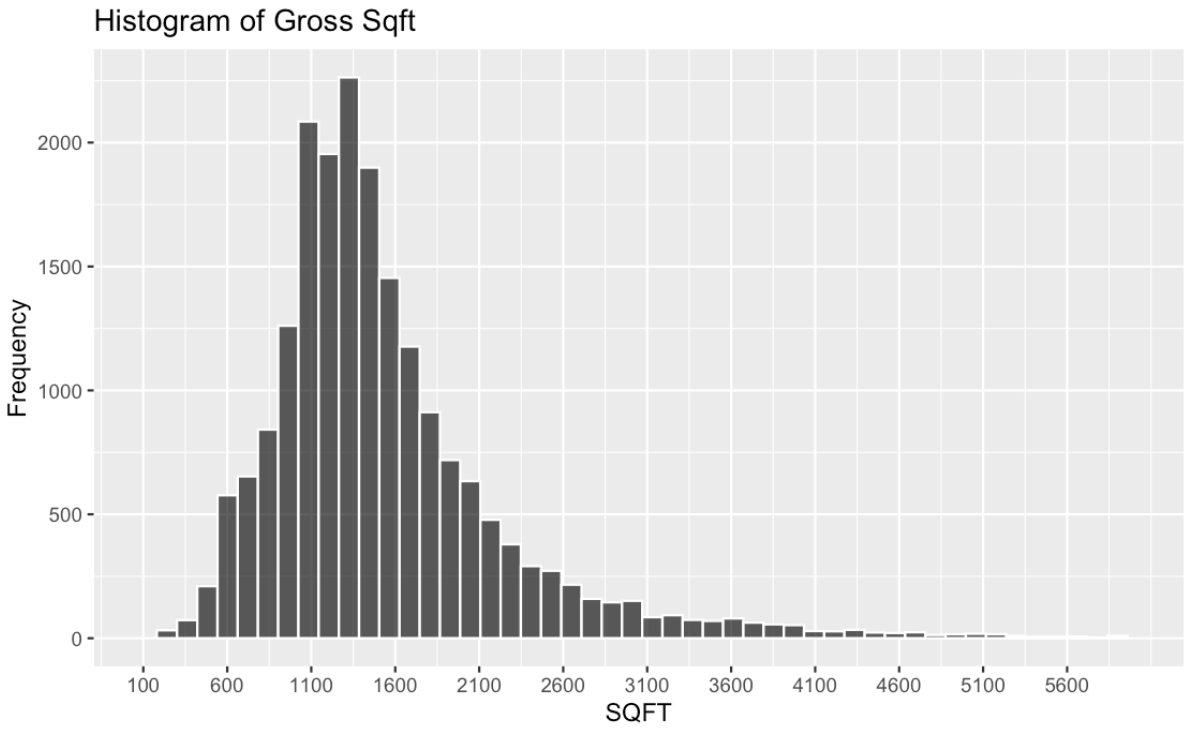
One issue we found out is that a lot of columns needed to be changed to numeric values before we could operate on them. Furthermore, we realized that some numeric values had commas in them. So, though they look numeric, they really are not as we need to remove those commas. Furthermore, we made note of a few columns that would be of no use for us, so we went ahead and dropped them (Eg: the column "easement"). As mentioned in the project proposal and earlier, we are working on single unit homes, so the next task was to find out how many single unit homes there are among the data and limit those, so all the operations would be performed upon them.

Another important issue here was to convert each year into quarters based on the months in a year. This would help us to make more accurate predictions, quarter to quarter, rather than from year to year. So, that is what we focused on next. We took each year and based on the months, we divided each year into four quarters (Q1, Q2, Q3, Q4). We also went ahead and did a few other things more so to get a better understanding of what we were dealing with, regardless of its impact on the project. These included finding out in which year, how many single unit residences were built, applying restrictions based on our constraints (Which drastically reduced the number of records in the final data set), playing around with the values pertaining to gross footage, etc. This little part of the project gave us a better understanding of what we had and which variables to use, which to completely ignore, etc.

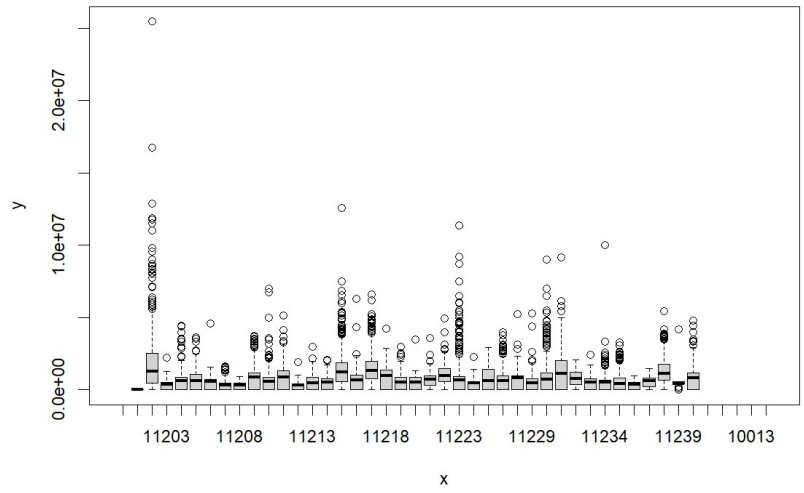
Before we fit and train the models, we decided to visualize some of the data. Primarily, the focus was on price, the square footage of the land and finally the gross square footage. We just wanted to know the frequency of units with respect to these three parameters. The results are given below:


Histogram of Price
Frequency
6000
4000
2000
0
0e+00
1e+07
2e+07
Price in $

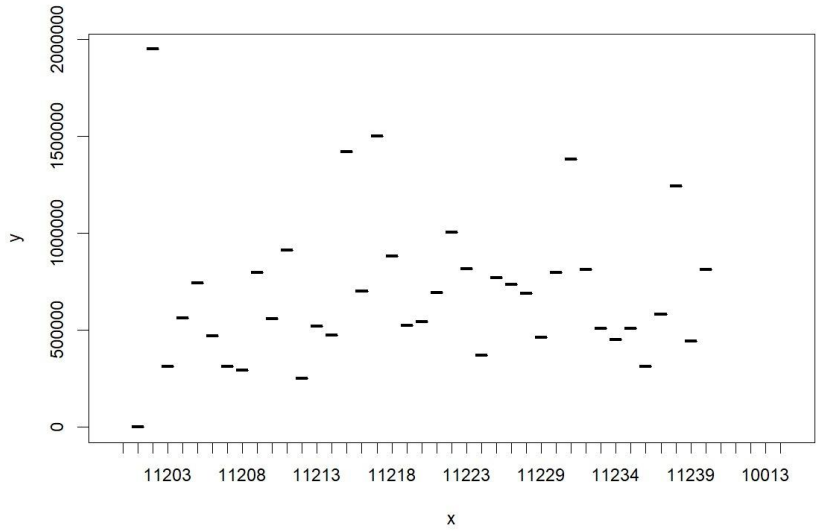

Histogram of Land Sqft
Frequency
3000
2000
1000
0
100
1100
2100
3100
4100
5100
6100
7100
8100
9100
SQFT


Histogram of Gross Sqft
Frequency
2000
1500
1000
500
0
100
600
1100
1600
2100
2600
3100
3600
4100
4600
5100
5600
SQFT

Now, after this step, we decided to group the zip codes into unique buckets and rank them based on these parameters. Final steps in the EDA included the meticulous bucketing of these zip codes and then visualizing the results using libraries such as corrplot and ggplot. We utilized both the mean and the median while making these plots. The results for the same are below:



With Median



With Median

Now, the correlation plot and the features of correlation are given below:

	taxclasscurr	landsqft	price	resunits	totunits	yrbuilt	taxclasssale
taxclasscurr	1.00000000	0.13593808	0.0452871801	-0.479745999	0.0566112622	0.37686996	0.73691132
landsqft	0.13593808	1.00000000	0.1067544908	-0.050763909	0.0436298232	0.15552551	0.15849301
price	0.04528718	0.10675449	1.0000000000	0.110065485	-0.0005511958	0.02380221	0.04234858
resunits	-0.47974600	-0.05076391	0.1100654850	1.000000000	0.0066524887	-0.14601830	-0.28157367
totunits	0.05661126	0.04362982	-0.0005511958	0.006652489	1.0000000000	0.09256420	0.05778891
yrbuilt	0.37686996	0.15552551	0.0238022128	-0.146018300	0.0925641964	1.00000000	0.30233786
taxclasssale	0.73691132	0.15849301	0.0423485765	-0.281573673	0.0577889064	0.30233786	1.00000000

Correlation features

Correlation plot

Next, we changed zipcodes to zero based on addresses and then found the specific median and mean price before we created a zip rank column. This was just a way to get a feel for all the variables before doing any necessary cleaning processes. The results are given below:

Description: df [39 × 2]

zip <fctr>	n <int>
0	1
11201	535
11203	958
11204	751
11205	235
11206	216
11207	557
11208	486
11209	692
11210	1188

1–10 of 39 rows

We repeated another process for the price variable as it is our dependent variable. That included getting rid of unrealistically high priced homes (with respect to this study. For example, homes priced above $7 Million). The subsequent steps included dealing with the land square feet column. We noticed many empty spaces in that column, so with the new information based on the zip code bracket and the averages found earlier, we cleaned up that data and made sure each nonexistent value was filled with the average from that area. Repeated the same process for the year built column as well. The results for that are showcased below:

```
 $ neighborhood: Factor w/ 124 levels "","BATH BEACH              ",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ bldclasscat : chr  "01ONEFAMILYDWELLINGS" "01ONEFAMILYDWELLINGS" "01ONEFAMILYDWELLINGS"
"01ONEFAMILYDWELLINGS" ...
 $ taxclasscurr: num  3 3 3 3 3 3 3 3 3 3 ...
 $ block       : Factor w/ 6692 levels "","1","1000",..: 3710 3712 3720 3735 3743 3743 3744 3747
3752 3757 ...
 $ lot         : Factor w/ 2097 levels "","1","10","100",..: 113 1301 1318 1219 702 777 1061 1116
1116 506 ...
 $ bldclasscurr: Factor w/ 168 levels "","  ","A0","A1",..: 8 8 8 8 8 8 4 10 4 8 ...
 $ address     : Factor w/ 98432 levels "","1 12TH   ST EXTENSION                    ",..: 17256
16865 18990 4402 19101 19097 19070 19918 19432 8478 ...
 $ aptnum      : Factor w/ 4556 levels "","            ",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ zip         : Factor w/ 45 levels "","0","11201",..: 28 15 15 28 15 15 15 15 15 28 ...
 $ resunits    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ comunits    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ totunits    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ landsqft    : num  2900 1950 2223 2469 2417 ...
 $ grosssqft   : num  1660 972 2520 1836 1462 ...
 $ yrbuilt     : num  57 77 57 67 52 52 51 42 37 62 ...
 $ taxclasssale: num  2 2 2 2 2 2 2 2 2 2 ...
 $ bldclasssale: chr  "A5" "A5" "A5" "A5" ...
 $ price       : num  829000 790000 788000 920000 839000 854000 750000 699000 950000 699000 ...
 $ date        : Date, format: "2016-04-05" "2016-06-21" "2016-03-31" "2016-02-04" ...
 $ quarter     : chr  "2016/02" "2016/02" "2016/01" "2016/01" ...
 $ zip_rk      : chr  "zip_3" "zip_5" "zip_5" "zip_3" ...
```

Model Training

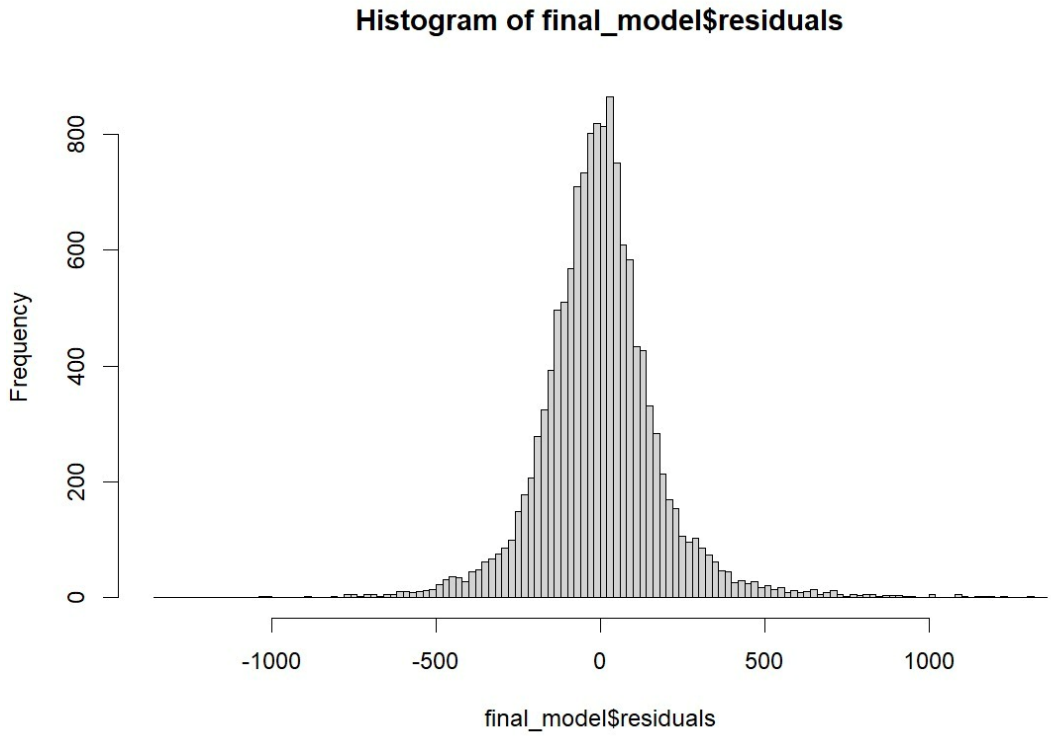
Now, after the data analysis and data preprocessing was done, all we needed to do was utilize linear regression and train models before finding out which one's the best for validation, and which model gave us the best insights. We employed four models and compared between those four to figure out which one best suited our progress in this project. We looked at the R-Squared, RMSE values and the degrees of freedom before deciding on a model. The results of this phase are depicted below:

```
#Starting the models
model1 <- lm(price ~ bldclasscat + bldclasssale + grosssqft + yrbuilt + quarter + zip_rk, data =
fullDf)
summary(model1)
############### r^2 = 0.5518 , Degrees of Freedom = 43 ######################

#RMSE
sqrt(mean(model1$residuals^2))
########## RMSE = 497361.6 ##################

#model2
model2 <- lm(price ~ bldclasscat + log(landsqft) + sqrt(grosssqft )*zip_rk + yrbuilt + quarter , data =
fullDf)
summary(model2)
##############  r^2 = 0.6146 , Degrees of Freedom = 39 ####################

#RMSE model2
sqrt(mean(model2$residuals^2))
################ RMSE = 460563.2 ############

#model3 square root of price
model3 <- lm(sqrt(price) ~ bldclasscat + yrbuilt + landsqft + grosssqft + quarter + zip_rk, data =
fullDf)
summary(model3)
############## r^2 = 0.5774 , Degrees of Freedom = 35 #########

#RMSE model 3
sqrt(mean((fullDf$price - model3$fitted.values^2)^2))
################ RMSE = 480768.5 ############

#final model, add interaction between zip_rk and grosssqft
final_model <- lm(sqrt(price) ~ bldclasscat + yrbuilt + quarter + sqrt(grosssqft)*(zip_rk)
+log(landsqft), data = fullDf)
summary(final_model)
############## r^2 = 0.6232 , Degrees of Freedom = 39 #############

#RMSE final model
sqrt(mean((fullDf$price - final_model$fitted.values^2)^2))
############# RMSE = 460493 ###############
```

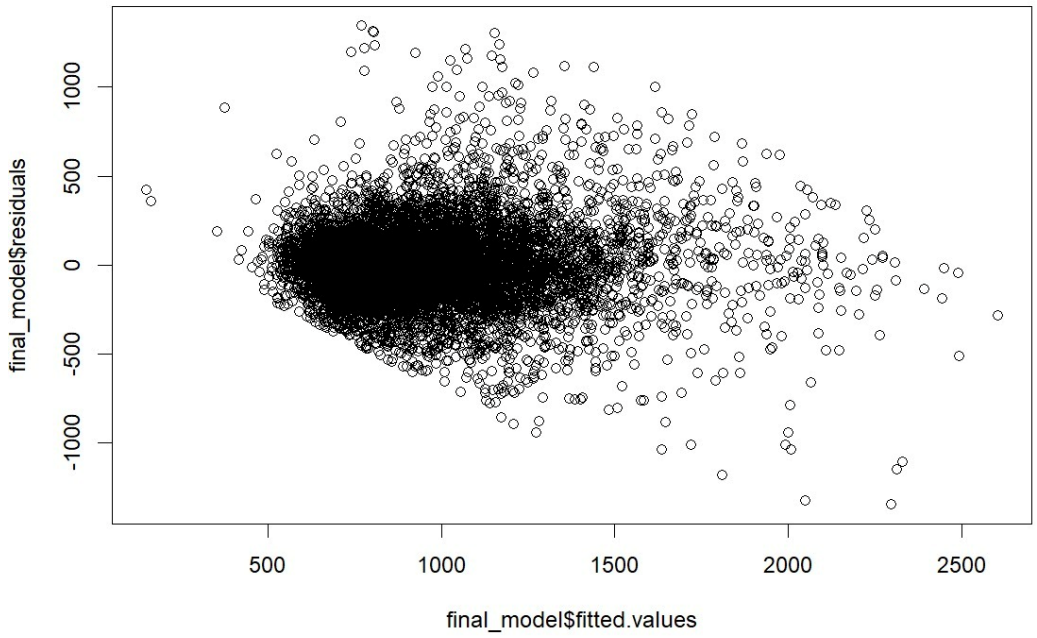
Model Validation

So, after going through all of the models above, we decided that the final model (model 4) was the best in aiding our progress in this project. While the second model was also a good fit, we felt that model 4 would better aid us based on the subtle differences in R-Squared, RMSE values and the degrees of freedom.

After choosing this model and fitting it to the data, we obtained results that we visualized. We mainly wanted to visualize the frequency to residuals in one graph and then between residuals and fitted values. The visuals for both are below:



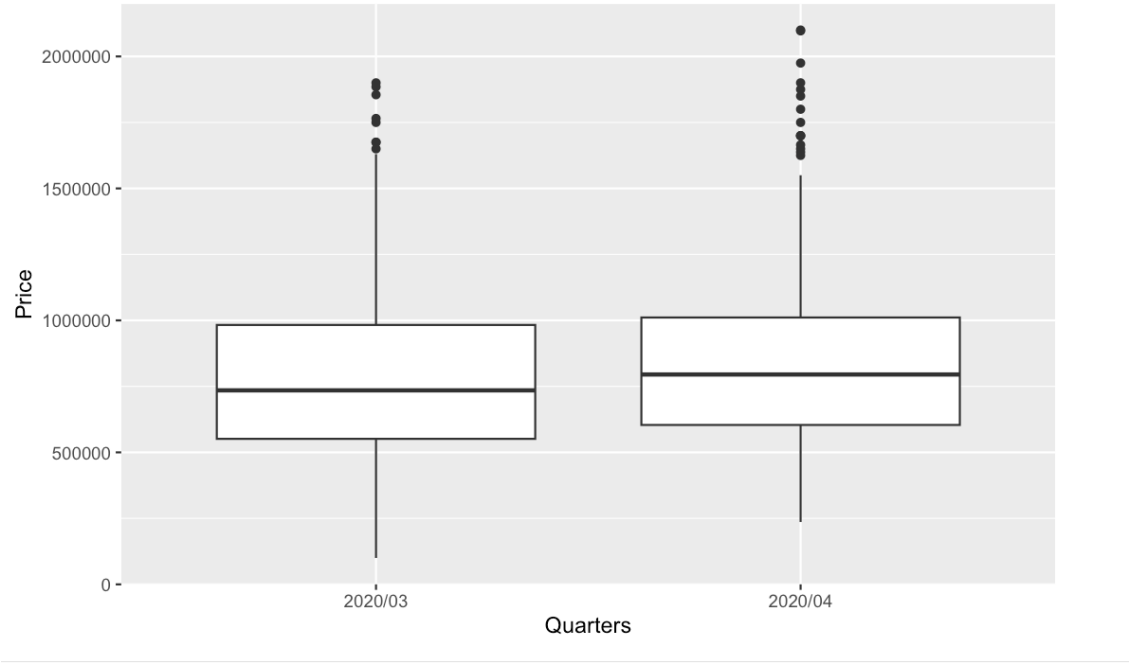
Frequency vs. Residuals

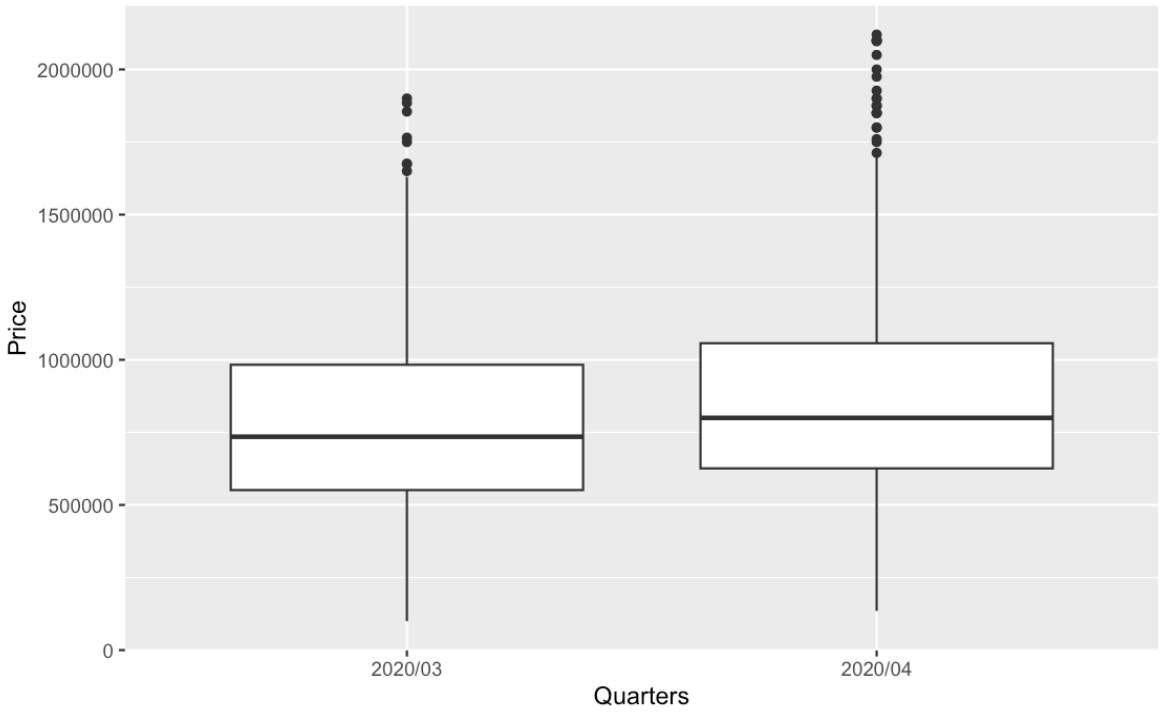


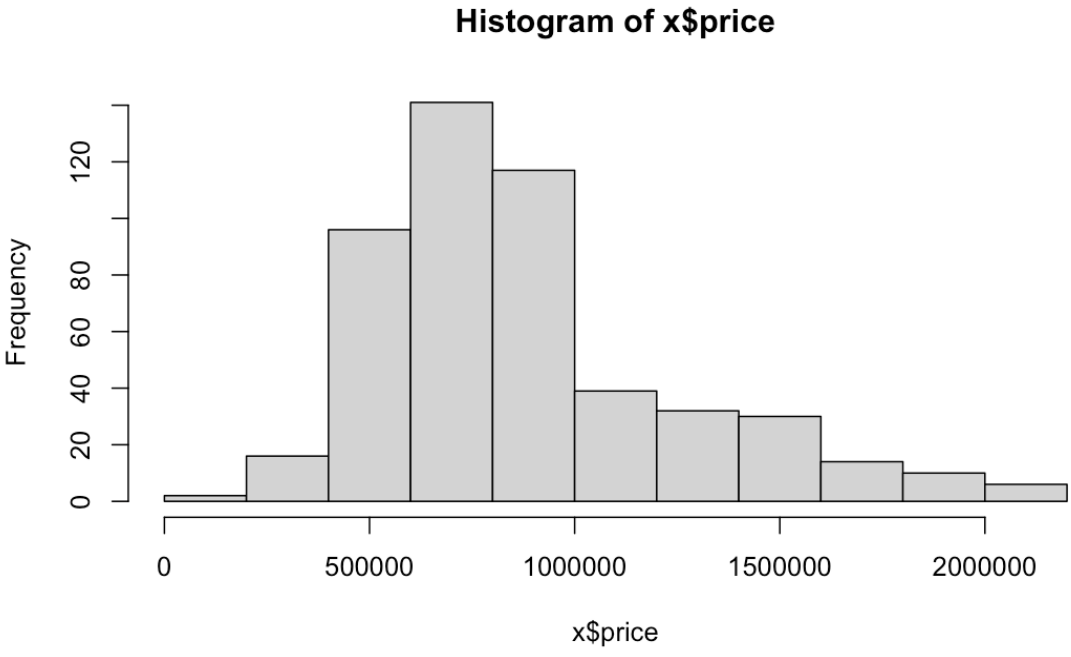
Residuals vs. Fitted Values

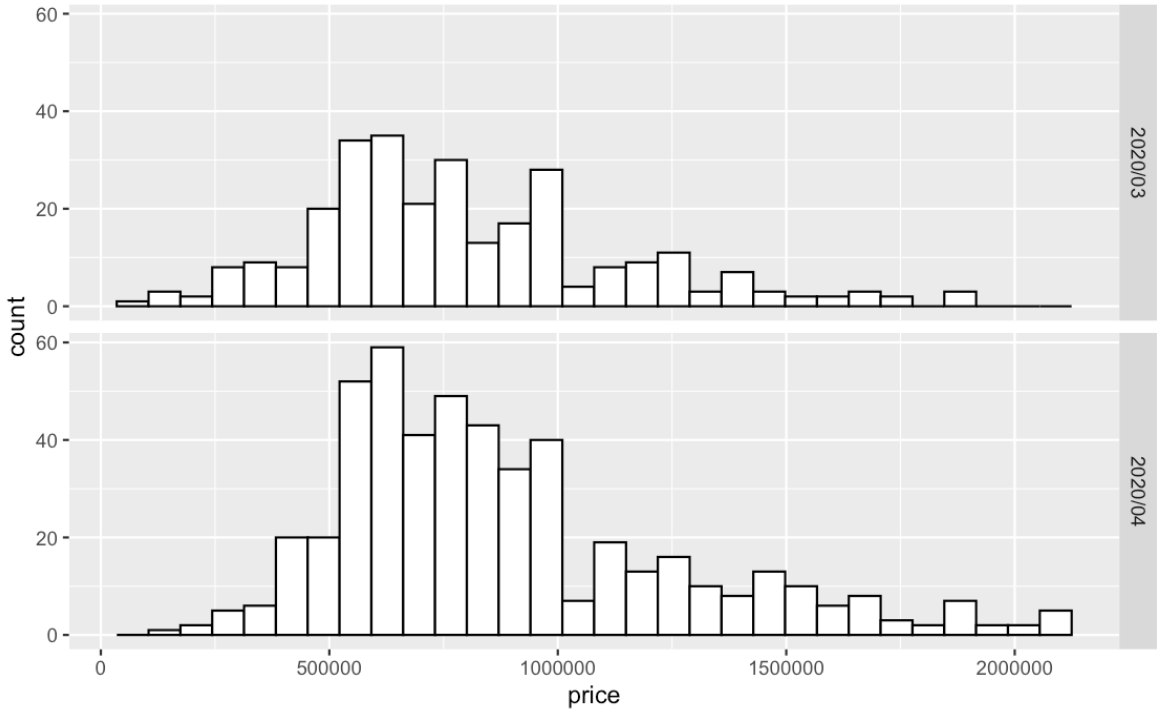
Now after this, we wanted to know the price change from quarter to quarter. So, after a lot of tinkering, we decided to showcase the results for Q3 to Q4 in the year 2016. We noticed that

between these quarters, there was a slight increase in the price, perhaps owing to economic factors affecting the borough of brooklyn. Generally speaking though, unless its a huge crisis, we noticed the trend might slightly increase or decrease; in other words, the change is not too significant. The plots for the same are given below:




2000000
1500000
1000000
500000
0
Price
2020/03
2020/04
Quarters


Histogram of x$price
Frequency
0
20
40
60
80
120
0
500000
1000000
1500000
2000000
x$price



Conclusion

To conclude, this study of Brooklyn, New York, housing prices from 2016 to 2020 using linear regression analysis has been a major undertaking with important ramifications. This research was intended to provide useful information to a range of stakeholders in addition to deciphering the complex patterns of home price fluctuations over a certain period of time.

The results of this study have important implications for legislators, real estate investors, prospective homeowners, and urban planners. Our objective was to provide stakeholders with well-informed decision-making tools by exploring the intricate patterns of housing pricing. The information gathered from this research can direct real estate investments, mold urban development plans, and have an impact on housing regulations, all of which will support Brooklyn's sustainable development.

This research contributes a valuable layer to the wider conversation on urban economics and real estate dynamics, in addition to its direct practical applications. Beyond a retrospective examination, this research's ultimate goal was to look forward. Our objective was to apply a

strong model to real-time data and interpret the intricate network of factors that influences market movements. By offering insights into future price trends, this predictive component of our study seeked to improve our comprehension of Brooklyn's place within the greater metropolitan statistical area. We added to the continuing story of urban development and offered a basis for well-informed decision-making in the ever-changing real estate market by dissecting Brooklyn's distinct history and its influence on the housing market in New York City.

Data Sources

Libraries/Packages used:

tidyverse (and everything included in this library)
lmtest
plyr
dplyr
corrplot
readr
stringr
ggplot2
lubridate
MASS

Software:

R, RStudio

Dataset:

Kaggle

Associated Tools:

MS Office, Adobe Acrobat PDF Reader, GitHub

Bibliography

Furman Center. "Brooklyn Neighborhoods." Furman Center. November, 2023. https://furmancenter.org/neighborhoods/view/brooklyn.

NYC Department of Finance. "Property Rolling Sales Data." NYC Department of Finance. November, 2023. https://www.nyc.gov/site/finance/taxes/property-rolling-sales-data.page.

Tianhao Wu. "Brooklyn Homes 2003 to 2017." Kaggle. November, 2023. https://www.kaggle.com/datasets/tianhwu/brooklynhomes2003to2017/code.

SPMukherjee. "Predictions on Brooklyn Housing Prices." Kaggle. November, 2023. https://www.kaggle.com/code/spmukherjee/predictions-on-brooklyn-housing-prices.

"Urban Housing Trends." Journal of Urban Studies, vol. 25, no. 3, 2005, pp. 123-145. November, 2023. https://www.sciencedirect.com/science/article/abs/pii/S1051137705000495.

"New Perspectives on Urban Planning." Journal of Urban Planning, vol. 42, no. 2, 2021, pp. 78-95. November, 2023. https://journals.sagepub.com/doi/full/10.1177/00027642211003149.

NYC Department of City Planning. "PLUTO Data Dictionary." NYC Department of City Planning. October, 2023. https://www.nyc.gov/assets/planning/download/pdf/data-maps/open-data/pluto_datadictionary.pdf?v=18v1.