Analysis of the housing Market in Two US Cities from 2020-2022

Nikhil Muthuvenkatesh
Drexel University
nm3297@drexel.edu

Rohan Ukkalam
Drexel University
rsu24@drexel.edu

Luqing Qi
Drexel University
lq53@drexel.edu

Seyi Oyesiku
Drexel University
so536@drexel..edu

***Abstract* - The aim of this study is to construct, using a multitude of datasets on property values over the years, a model that can predict possible future prices for various real estate. There are two datasets used for this study. The first dataset is a realtor dataset that contains information about the built status, price, bed, bath, acre lot size, full address with city and state, zip code, and the size of the house. The other dataset contains more demographic information about buyers and sellers of property across the United States; however, these datasets were later filtered to Pennsylvania and New York. This report explores the data using visualizations to ascertain what preprocessing needs to be done and which machine learning models could be used to meet the goals.**

Keywords: Real Estate, Pennsylvania, New York, Pricing Prediction

I. Introduction

Since the COVID-19 pandemic in 2020, real estate pricing has been quite volatile. The initial lockdowns led to a large drop in property values across the country, but shortly thereafter there was a huge spike that has continued up until now. This project will implement a model that is trained with real estate data over the last two years to predict the future prices for housing. The goal of this capstone project is to correctly predict when someone who is interested in purchasing real estate should purchase a home in a city on the East Coast, such as Philadelphia or New York.

The motivation behind this project comes from the team members. We are all graduate students who are interested in purchasing real estate in the future, and we would like to know what to expect from this market in the future. It could also potentially be useful for banks that give out home loans, mortgage companies, and escrow companies. In addition, anyone that is interested in purchasing a home in one of the cities that we use for analysis in the next couple of years would find this information useful. A model that can predict house pricing in these two cities within the next couple of months

- The training set will include real estate data from both cities from the period of 2020-2022
- The end product will include visualizations for laymen to better understand the trends for the past few years
- The end product will include several areas/locations of interest

II. Dataset Description

Several datasets were chosen for this project. The first dataset originally came from www.realtor.com - A real estate listing website operated by the News Corp subsidiary Move, Inc. and based in Santa Clara, California. The dataset we are using is realtor data from Kaggle. However, this dataset contains properties from across the United States. As we are for the context of this project only focusing on the states of Pennsylvania and New York, a filter was created on the dataset to only get properties in those two states. The price will be the column that we are trying to predict for these cities, and the other data features are possible predictors of that price.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 57778 entries, 465089 to 921687
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   status        57778 non-null  object
 1   price         57778 non-null  float64
 2   bed           50791 non-null  float64
 3   bath          56258 non-null  float64
 4   acre_lot      9508 non-null   float64
 5   full_address  57778 non-null  object
 6   street        57758 non-null  object
 7   city          57778 non-null  object
 8   state         57778 non-null  object
 9   zip_code      57776 non-null  float64
 10  house_size    37209 non-null  float64
 11  sold_date     33279 non-null  object
dtypes: float64(6), object(6)
memory usage: 5.7+ MB
```

Fig. 1. Dataset Variables

The second database used is a public-use database by Fannie Mae and Freddie Mac. Loan-level Public-Use Databases (PUDBs) are released annually to meet FHFA's requirement under 12 U.S.C 4543 and 4546(d) to publicly disclose data about the Enterprises' single-family and multifamily mortgage acquisitions. The datasets supply mortgage lenders, planners, researchers, policymakers, and housing advocates with information concerning the flow of mortgage credit in America's neighborhoods. The PUDB single-family datasets include loan-level records that include data elements on the income, race, and gender of each borrower as well as the census track location of the property, loan-to-value(LTV) ratio, age of mortgage note, and affordability of the mortgage.

The third database used is the Zillow Home Value Index. The Zillow Home Value Index measures monthly changes in median price at property-level in different cities in the United States. We use this dataset to study the housing price change in Philadelphia and New York from the years 2020 to 2022.

The fourth database used is the Rolling Sales Data at NYC. The Department of Finance's Rolling Sales file lists properties sold in the last twelve-month period in New York City for tax classes 1,2, and 4. These fields include the neighborhood, the building type, the square footage of the real estate, and the sale price.

III. Exploring Data

1. Realtor Data

A. Imbalanced Data

Imbalanced data refers to classification problems where we have unequal instances for different classes of the target variable. The only data in this dataset that seemed imbalanced were the zip codes, but it makes sense because they are assigned to each property, as opposed to being an actual feature of the property. The zip code is based on how the city organizes different parts of itself. Therefore, we had expected there to be some imbalance in this feature. The rest of the features we are using for our analysis are balanced already.

B. Null Values

The values in the dataset that can be considered null are the NaN values for the sets we are working with. These represent values that are missing from the columns in our datasets. Figure 1 shows all the columns in our realtor dataset that have null values. Figure 2 below shows the distribution of the null values in our dataset.

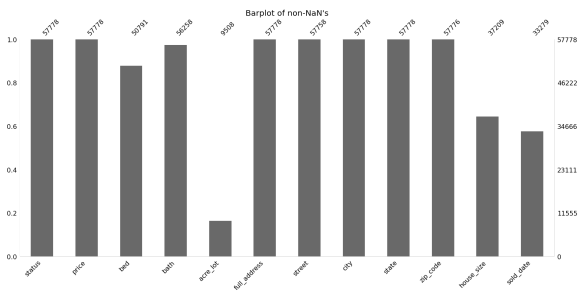

Fig. 2. NaN Values Visualization

C. Categorical Features

Categorical features are normally non-numeric, represented by strings, variables that represent a limited number of values that assign qualitative property. They usually represent categories. The categorical features within this data are status, full_address, street, city, and state. Due to this nature, for future machine learning models, we will have to preprocess these categorical features using various encoding techniques such as label encoding, one hot encoding, or hash encoding. The following graph is a heat map of various Philadelphia zip codes and the average price of property within said zip code.

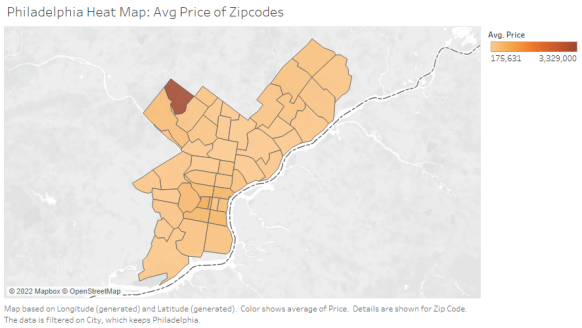

Fig. 3. Heat Map of Average Price

D. Continuous Features

Continuous features are normally numerical features that are quantified, but they have no limit in terms of the range of values that can be encompassed in them. Our continuous features in this study are price, acre lot size, and house size.

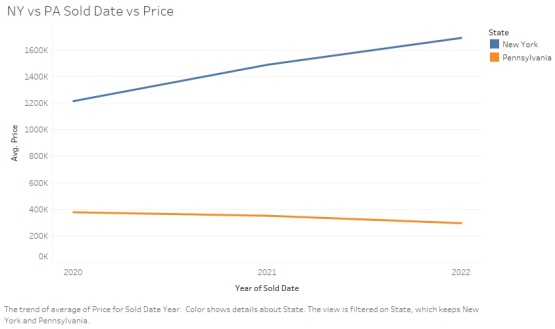

Fig. 4. Average Price over Time

E. Discrete Features

Within the data are several discrete data features, such as bed and bath, that are of interest to us. The number of bedrooms and bathrooms for a given property would indicate its potential price. The following figure shows the average amount of beds a property has per its location. .

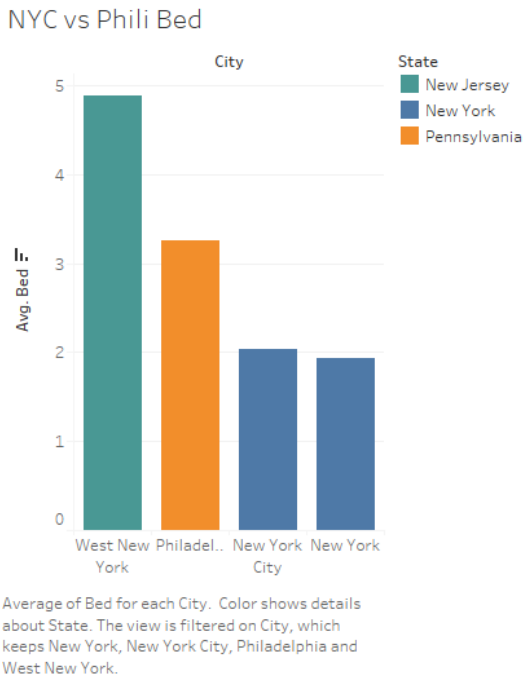

Fig. 5. Average beds in a property

2. Loan-level Public-Use Databases (PUDBs)

In the dataset, we noticed that the bank allowed us to have two different borrowers apply the loan for each mortgage loan. We didn't find any NA value in our dataset and did some basic demographics analysis about borrowers. The Average Total Monthly Income Amount is equal to $10872.35. Most Mortgages browser has 360 months to pay back their loans. The Average Mortgage interest range is equivalent to 2.84. The average Mortgage amount is equal to $227,637.9.The average age for Borrower one is equal to 44, and the average age for Borrower two is equal to 68. We got 34 categorical features and 22 continuous features.

A. Data Features

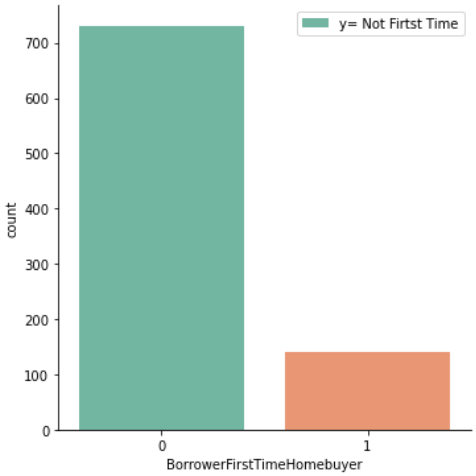

Fig. 6. Countplot of First-Time House Owner

In this graph, the number zero represents someone who is not buying a house for the first time. Number one represents someone who is a first-time home buyer. From this graph, we could tell, the majority of the borrowers are not buying a house for the first time.

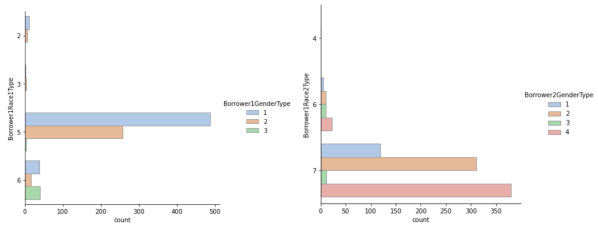

Fig. 7

From the upper graph, For the hue-axis, we can tell that the Numeric code indicates the sex of the first or primary Borrower. Number one represents males. Number two represents females. Number three represents information not provided by the Borrower. Number four represents No Co-Borrower, for the primary borrower is male. For co-borrower, we could found most loans don't have co-borrower. For the y-axis, we can tell the race of the Borrower. Numeric code indicates the Borrower's race. 2 = Asian, 3= Black or African American, 4 = Native Hawaiian or other Pacific Islander, 5 = White, 6= Information not provided by Borrower, 7= Not Applicable. From the left graph, we can tell that most of the borrowers are white males. From the right chart, we can find that a borrower is an institution, a corporation, or a partnership.

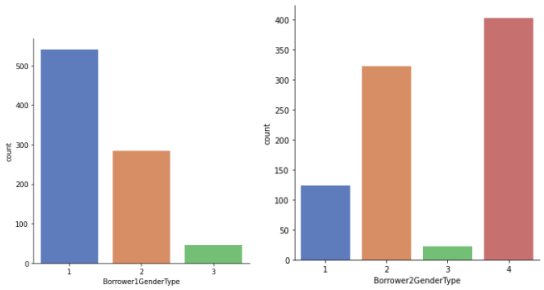

Fig. 8. Gender Distribution of Buyers

This graph shows the gender distribution for borrower one and borrower two. Numeric code indicating the sex of the co-borrower. 1= Male, 2 = Female, 3 = Information not provide by borrower, 4= No-borrower. Our primary borrowers have twice as many male borrowers as female borrowers. For our co-borrower, we could see more females than males. We could also find that a certain amount of lone just got one borrower.

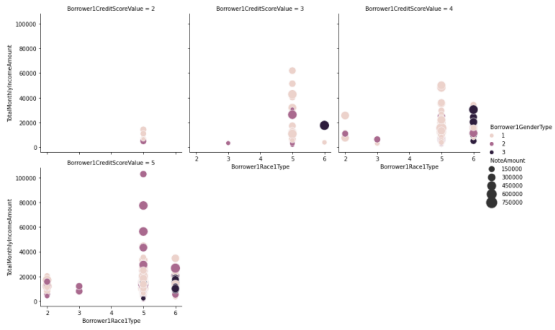

Fig. 9. Loan Distribution of Borrowers

These graphs show the number of loans under different credit ratings for primary borrowers of different races and monthly incomes. Numeric code indicating the race of the Borrower. 1= American Indian or Alaska Native, 2= Asian, 3= Black or African American, 4= Native Hawaiian

or other Pacific Islander, 5= White, 6= Information not provided by Borrower. Numeric code indicating the sex of the first or pr primary borrower. 1= Male, 2=Female, 3= Information not provided by the borrower. From this graph, we could see that whites can borrow more than other groups with the same credit score. We also found that white women tended to borrow more with a credit rating of five.

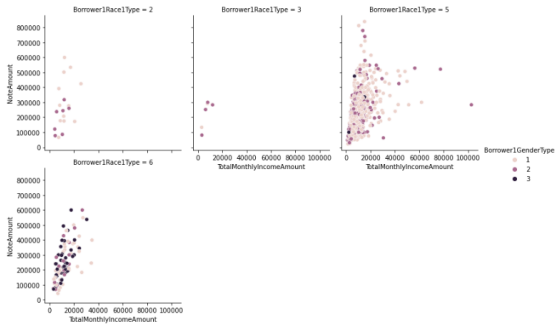

Fig. 10.

The numbers in this graph and the graph above represent the same race group and gender. We found a positive correlation between monthly income and the number of loans. At the same time, we found that whites have more income and loans than other ethnic groups.

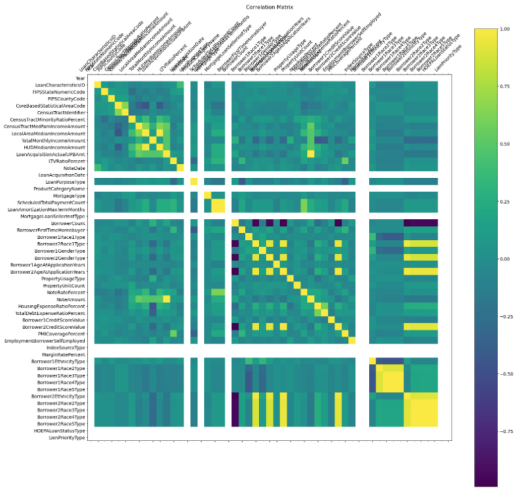

Fig. 11. Correlation Matrix

From this correlation matrix, we found that Loan Acquisition Actual UPB Amount, Census Tract Median Family Income amount, HUD Median Income Amount, Local Area Median Income Amount, and Total Monthly Income Amount. From this graph, we can conclude that three important factors can affect the number of loans you can take:

1. The number of loans you borrow from the bank.
2. The median income for the house purchased.
3. The average monthly payment of borrower.

3. Zillow Home Value Index

We got 893 records and 278 dimensions in our dataset. We filtered data into Philadelphia and New York to study the housing marketing price change during the pandemic. Most of the variables are numeric variables. We picked up variables from 1/01/2020 to 9/30/2022. We did data wrangling through Python and made data visualization through Excel.

A. Data Features

Fig. 12. Housing Price over time

Before we started our work, we tried to understand the impact of the pandemic on housing prices in different regions. From this chart, we can see that the average home price in New York is twice that of Philadelphia, and property prices in both places have steadily increased during the pandemic.

4. Rolling Sales Data at NYC

We pick our dataset from 10/1/2021 to 9/30/2022. We got some missing value problems in the dataset. We dropped the variables that we don't need, such as EASEMENT and APARTMENTNUMBER. We use model imputation for our category variables. We used mean imputation for our numeric variables. We got 6 category variables and 13 numeric variables.

A. Data Features

NYC Real Estate Build Year Distribution

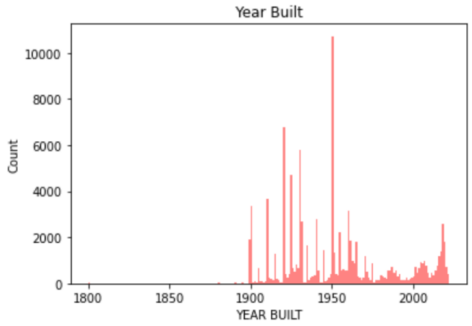

Fig. 13. NYC Build Year Distribution

From this chart, we can see that from 1900 to 1950, there was an upward trend in new Real Estate construction in New York City, reaching its peak in 1950.

Average Sq Ft Across NYC

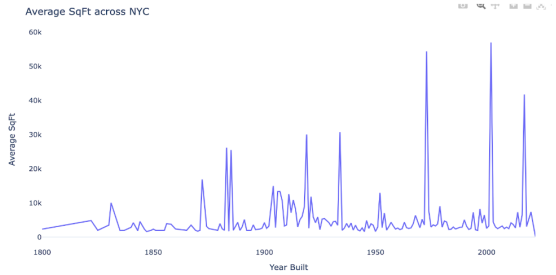

Fig. 14. NYC Average Square Footage

We can find that in the period between 1900 and 1950, there were two peaks. The first peak was in 1920, and the 20s were also the era of the great boom in the real estate market, and it became fashionable to build skyscrapers between 1920 to 1930. New York City built more skyscrapers in these ten years than at any other time in history. The Chrysler Building was the symbol of that era. The second peak may involve Roosevelt's New Deal, which dramatically increased investment in infrastructure. The Lincoln Tunnel and the Empire State Building have become icons of this period.

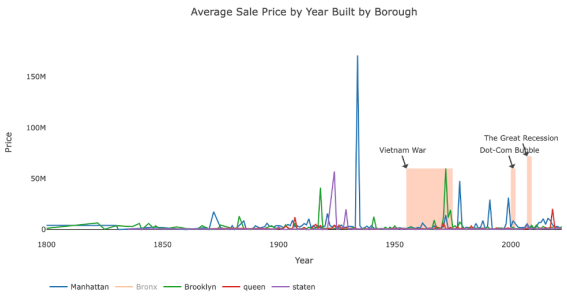

Fig. 15. Avg Price of NYC borough over time

We can see the same trend in this graph of average sales price change in different New York City areas. Manhattan had the highest average home price between 1900 to 1950. In the later stages of the Vietnam War, the average home price in the Brooklyn area were higher than in Manhattan. We found that the New York area strongly correlated with the U.S. economy. For example, the dot-com bubble and the 08 subprime mortgage crisis significantly impacted the real estate market.

Average Sale Price in Different Boroughs in NYC

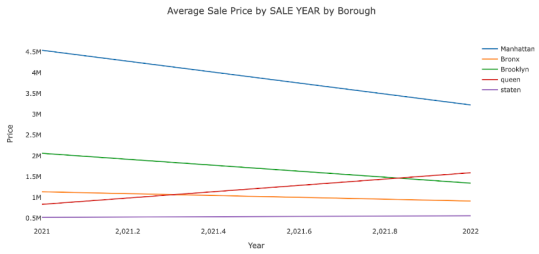

Fig. 16. NYC Sale Price over Time

We also explored average sale prices in different boroughs of New York City. From this graph, we can tell that the average sale price in Manhattan dropped from Oct 2021 to Set 2022. In contrast, we can see a clear trend improvement in Queens' average sale price.

IV. Summary

Real Estate pricing since the start of the pandemic has been extremely volatile. With the initial steep decline, there was also an extreme spike in the pricing over the last couple of years

in housing pricing. This study focuses on the two cities of Philadelphia and New York. The datasets used here were filtered to these two cities for analysis. The goal was to use continuous features along with the categorical features in our datasets to try to predict the future pricing of housing in these two major cities.

The first step was acquiring the datasets that were needed for this analysis. The realtor data was originally from www.realtor.com, but we accessed it through Kaggle. It details the pricing of real estate across the United States, but we wanted to focus solely on Pennsylvania and New York. Therefore, a mask had to be created for the data to be filtered to just those two states.

The second dataset comes from a Loan-Level Public-Use Database. It shows information about all mortgage acquisitions of single-family and multi-family homes over the last few years we are using for analysis. It also provides details of loans and demographic information about the loan holders.

The third dataset is a Zillow Home Value Index that provides us with the median price of homes in the United States.

The final dataset we used was rolling sales data from properties in NYC. The Department of Finance is where this data was found and it includes property information as well.

The next step was exploring the data that we had. We were able to identify the missing values and separate the categorical and continuous features in our datasets. We also had to identify imbalanced features.

We will be continuing our work by choosing models to use for predicting pricing with our data next term. Since we do have a target variable pricing, we will need to use only supervised learning methods. Since the number of features we might use for analysis would be fewer than 20, we could potentially use K-Nearest Neighbors. In addition, other options for us would be a Random Forest or Logistic Regression. For our machine learning model, we want to use Loan-level Public-Use Databases (PUDBs) data to predict the most important factor influencing the interest ratio and interest amount. The model we could use include: Decision tree, Random Forest, Neural Network, and Linear Regression.

We will keep exploring the feature in the public data for the real estate market in New York and Philadelphia. The new dataset we plan to use is Condominium Comparable Rental Income in NYC.

V. References

Santarelli, M. (2022, December 7). *Is it a good time to buy a house or should I wait until 2023-2024*. Norada Real Estate Investments. Retrieved December 8, 2022, from https://www.noradarealestate.com/blog/is-it-a-good-time-to-buy-a-house/

Santarelli, M. (2022, November 26).Housing market trends 2022: Will house prices fall ?. Norada Real Estate Investments. Retrieved December 8, 2022, from https://www.noradarealestate.com/blog/housing-market-trends/