Impacts on Home Pricing

Data Visualization & Analytics: Project-1 Group-8

By: Angela Gosewehr, Naseema Omer, Noah McHale, & Tristan Vazquez

Description

Identify any correlation of certain residential home traits on House Pricing within USA States using a real estate dataset.

Project Questions for Analysis

1. What impact do the number of bedrooms, number of bathrooms, acreage and house size have on House Sale Prices? What has the strongest correlation?
2. How do the above correlations vary by State?
3. What is the average pricing for a 3 bedroom, 2 bathroom home by State?

Data Selection

Selection criteria:

- Public, Free, reliable, with Usability rate of 10
- File Type: CSV
- Size: sufficient workable data

Selected Dataset:

- ‘USA Real Estate Datatest’ available on Kaggle.com
  https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

Dataset review / cleaning requirements

States with sufficient workable data selected for study:

Connecticut , Massachusetts, New Jersey, New York

Key data retained in csv file:

bedroom, bathroom, acreage, house size

Data removed in csv file: Null values, and outliers.

Data Cleansing

Data Frame (df) ‘Clean_states_df’

- Contains the needed Headers and Columns
- Contains data for States: Connecticut, Massachusetts, New Jersey and New York (4 states with most data in the original data set)
- Drop rows with Null Values.
- Calculate IQR to get outliers for each state (function).
- Removed outliers.
- Combined all 4 dfs using *pd.concat(df_list)*
- Write clean df to new csv file: ‘clean_states.csv’

Question 1

Question: For all four states aggregated in our dataset, what impact do the independent variables below have on House Sale Prices? What has the strongest correlation?

Variables:

- # of Bedrooms
- # of Bathrooms
- Lot Size (acres)
- House Size (ft^2)

Question 1 Analysis

Process:

Create scatter plots to display each independent variable in relation to house price. Run linear regression for each plot and calculate the squared r-value for each variable to determine how much of the variation of the home price is explained by our independent variables.

Steps:

- Narrow data frame to only include relevant columns.
- Create function to:
  - populate scatter plots to show correlation
  - calculate R-Squared
  - run linear regression
  - display linear regression equation
- Call each independent variable in function to create desired plot and display R-squared.

Question 1 Results

Result: All independent variables (# of Bedrooms, # of Bathrooms, Lot Size, House Size) had a positive correlation with the Home Price. Home size had the strongest correlation with home price, while lot size had the weakest correlation.

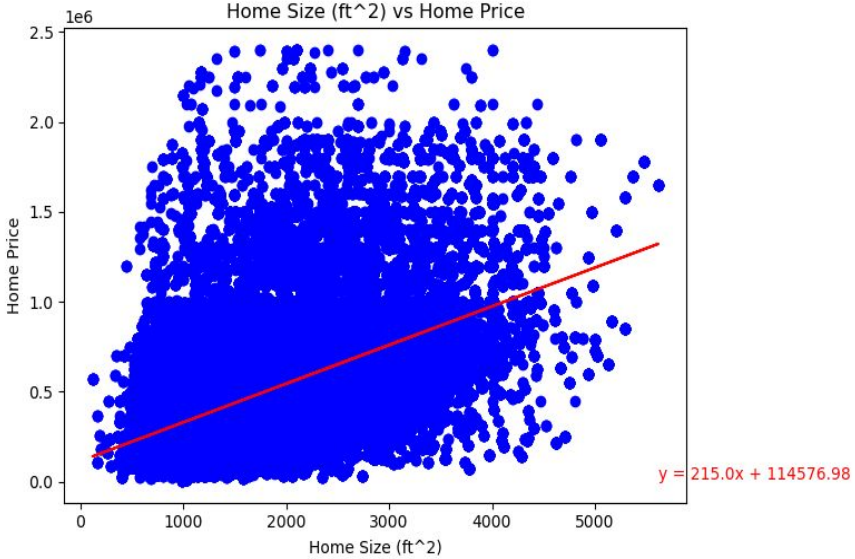
R-squared Values:

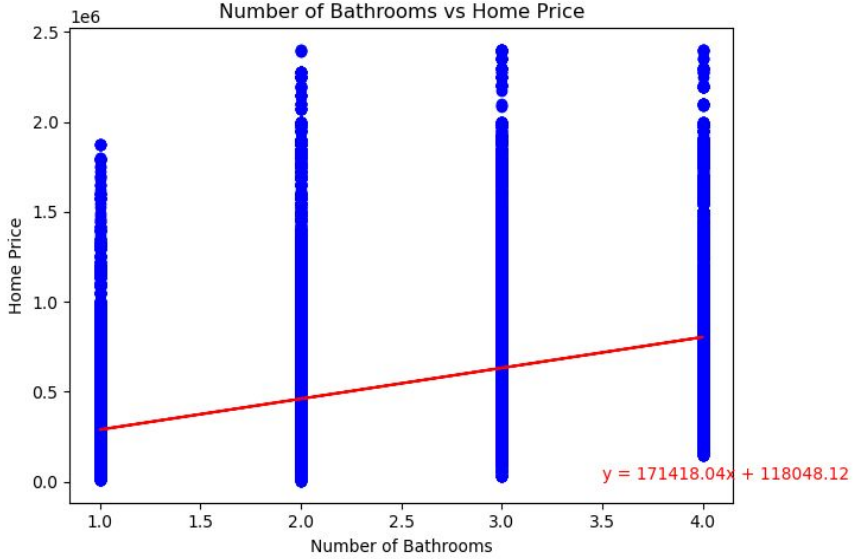
1. House Size: 0.2387
2. # of Bathrooms: 0.2042
3. # of Bedrooms: 0.0559
4. Lot Size: 0.0010

Takeaway:

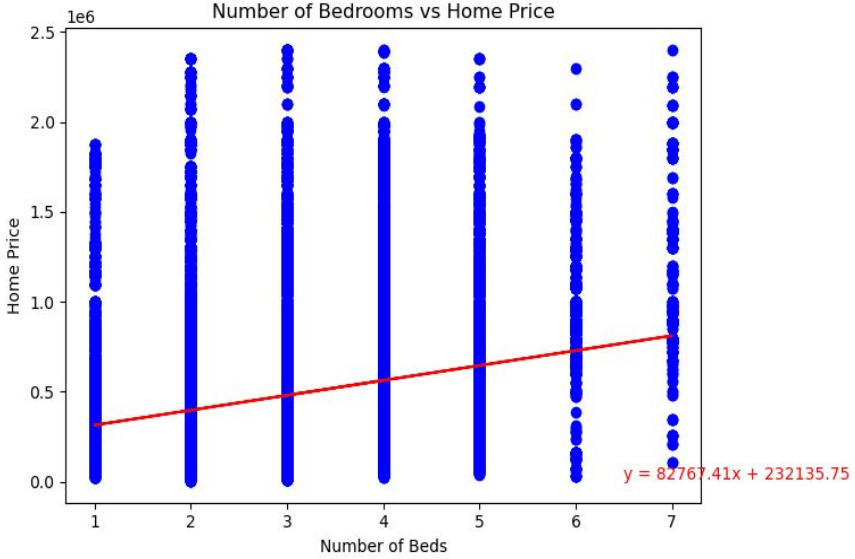
In the aggregate dataset of states we analyzed, ~23.87% the variation in House Sale Price can be explained by the variation in House Size and ~20.48% can be explained by the variation in # of Bathrooms making these variables have the leading impact on price in our regression model. # of Bedrooms and Lot Size had very little impact on the variation in home prices for our dataset at the aggregated level.

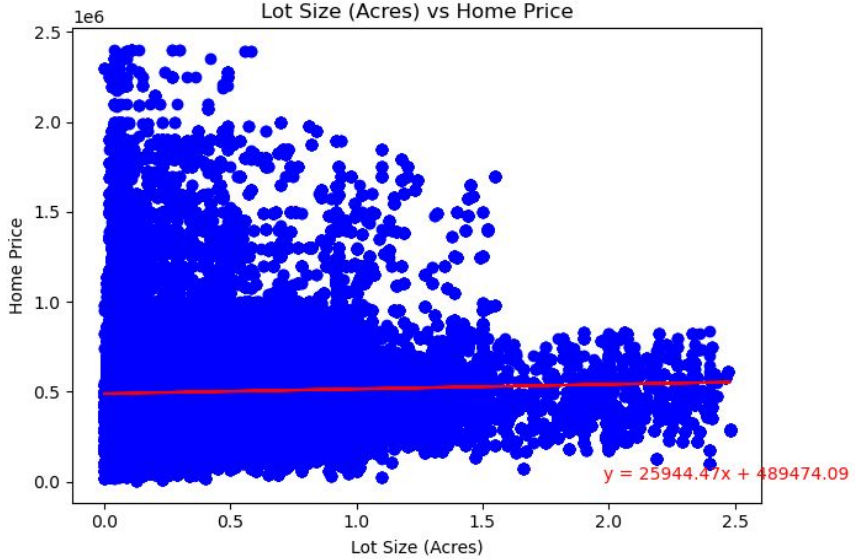
High Impact Variables





Low Impact Variables





Question 2

Question: For each of the four states in our dataset, how do the correlations of House Sale Price with the independent variables below differ state by state.

Variables:

- # of Bedrooms
- # of Bathrooms
- Lot Size (acres)
- House Size (ft^2)

Question 2 Analysis

Process:

Create new data frames for each state by filtering our original data frame. Create scatter plots to display each independent variable in relation to house price. Run linear regression for each plot and calculate the squared r-value for each variable to determine how much of the variation of the home price is explained by our independent variables.

Steps:

- Create new data frames to only include relevant state.
- Use created function to:
  - populate scatter plots to show correlation
  - calculate R-Squared
  - run linear regression
  - display linear regression equation
- Call each independent variable in function to create desired plot and display R-squared for each state.

Question 2 Results: Connecticut

Result: All independent variables (# of Bedrooms, # of Bathrooms, Lot Size, House Size) had a positive correlation with the Home Price. Home size had the strongest correlation with home price, while # of Bedrooms had the weakest correlation.

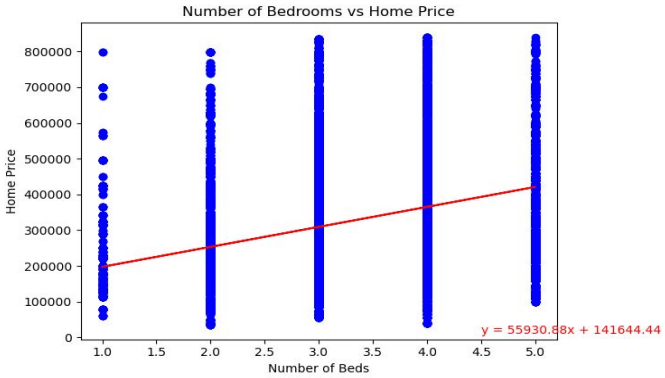
R-squared Values:

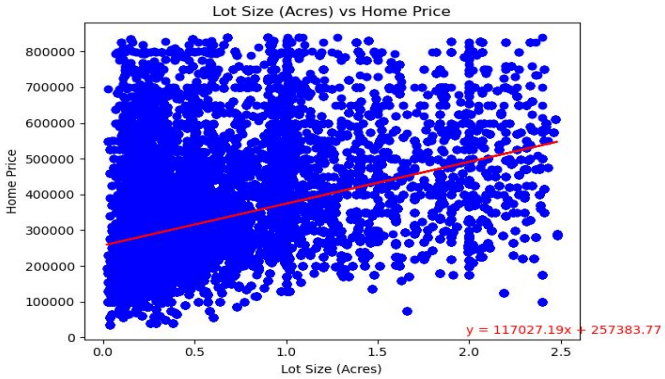
1. House Size: 0.4120
2. # of Bathrooms: 0.3761
3. # of Bedrooms: 0.0879
4. Lot Size: 0.1808

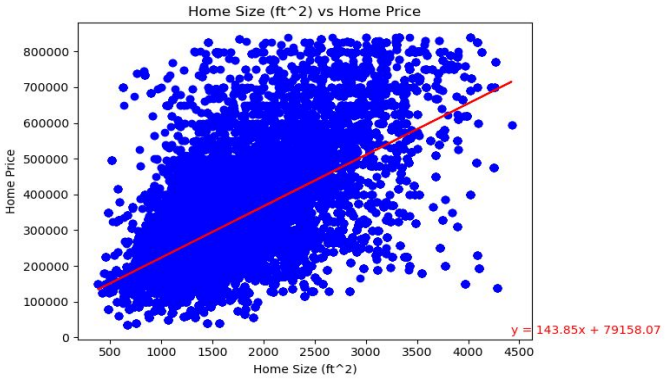
Takeaway:

In the CT specific dataset we analyzed, ~41.20% the variation in House Sale Price can be explained by the variation in House Size, ~37.61% can be explained by the # of Bathrooms, and ~18.08 % can be explained by Lot Size. # of Bedrooms had the smallest impact on the variation in home prices for our dataset specific to Connecticut.

Connecticut Visualizations







Question 2 Results: Massachusetts

Result: All independent variables (# of Bedrooms, # of Bathrooms, Lot Size, House Size) had a positive correlation with the Home Price. # of Bathrooms had the strongest correlation followed closely by House size, while Lot Size had the weakest correlation with House Price.

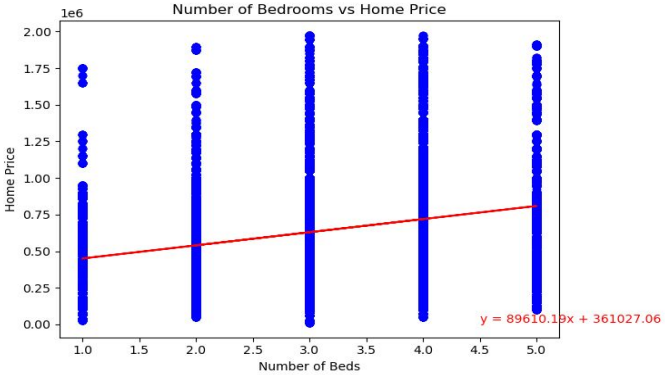
R-squared Values:

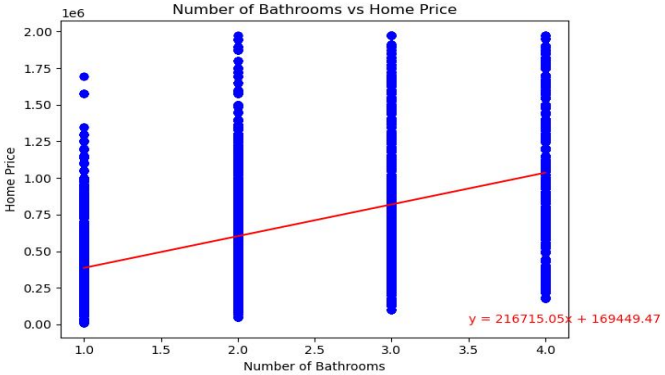
1. House Size: 0.2668
2. # of Bathrooms: 0.2786
3. # of Bedrooms: 0.0550
4. Lot Size: 0.0052

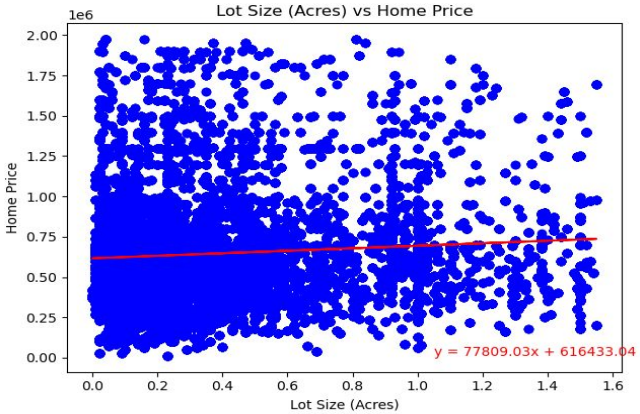
Takeaway:

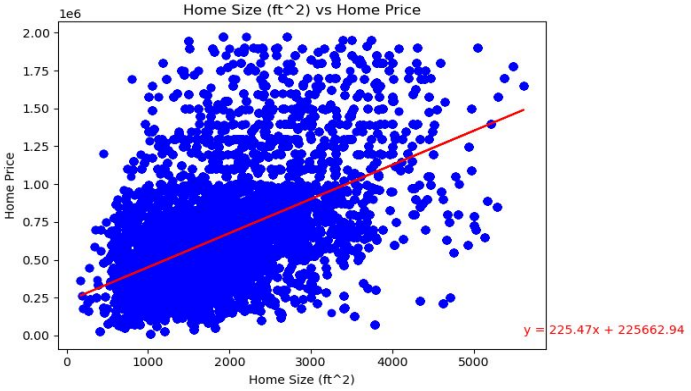
In the MA specific dataset we analyzed, ~26.68% the variation in House Sale Price can be explained by the variation in House Size, and ~27.86% can be explained by the # of Bathrooms. # of Bedrooms and Lot Size had the smallest impact on the variation in home prices for our dataset specific to Massachusetts.

Massachusetts Visualizations









Question 2 Results: New Jersey

Result: All independent variables (# of Bedrooms, # of Bathrooms, Lot Size, House Size) had a positive correlation with the Home Price. House Size had the strongest correlation, while Lot Size had the weakest correlation with House Price.

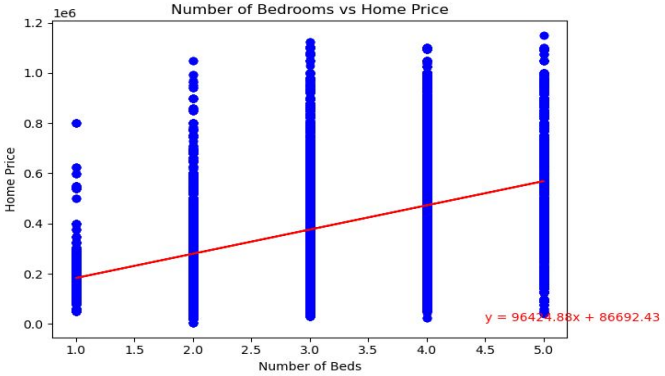
R-squared Values:

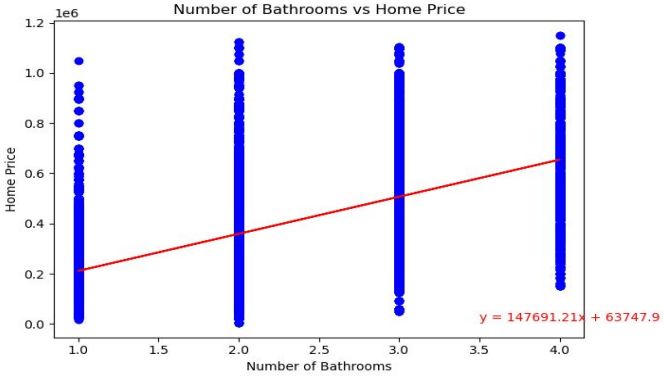
1. House Size: 0.4294
2. # of Bathrooms: 0.3756
3. # of Bedrooms: 0.1815
4. Lot Size: 0.0603

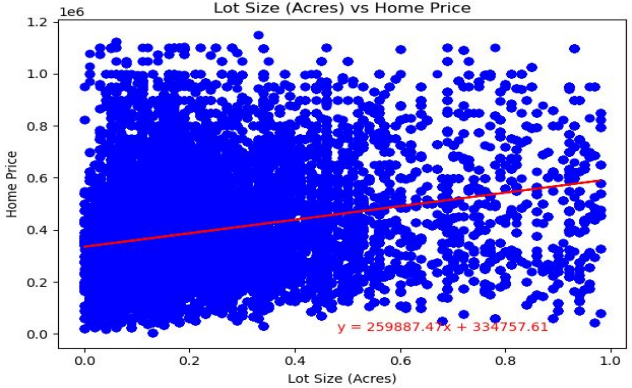
Takeaway:

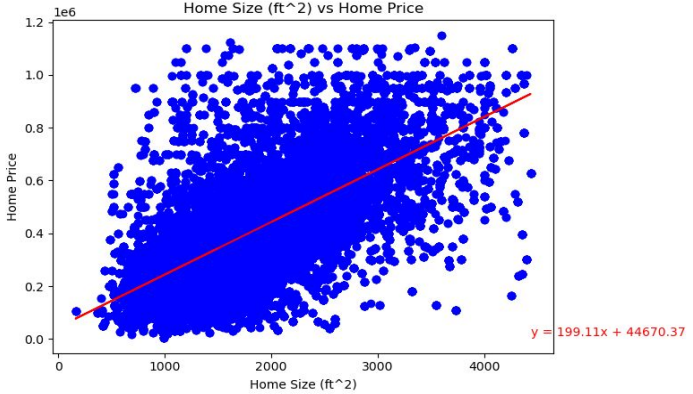
In the NJ specific dataset we analyzed, ~42.94% the variation in House Sale Price can be explained by the variation in House Size, ~37.56% can be explained by the # of Bathrooms, and ~18.15% can be explained by the # of Bedrooms. Lot Size had the smallest impact on the variation in home prices for our dataset specific to New Jersey.

New Jersey Visualizations









Question 2 Results: New York

Result: Three variables (# of Bedrooms, # of Bathrooms, and House Size) had a positive correlation with the Home Price while Lot Size had a negative correlation. House Size had the strongest correlation, while Lot Size had the weakest correlation with House Price.

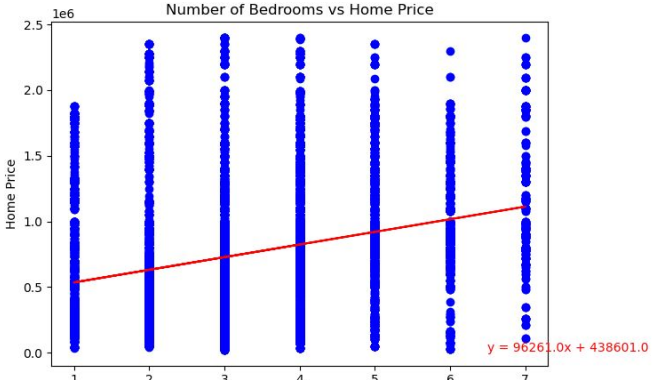
R-squared Values:

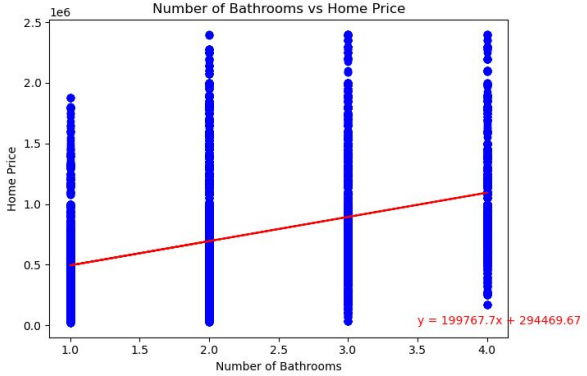
1. House Size: 0.1292
2. # of Bathrooms: 0.1651
3. # of Bedrooms: 0.0863
4. Lot Size: 0.0330

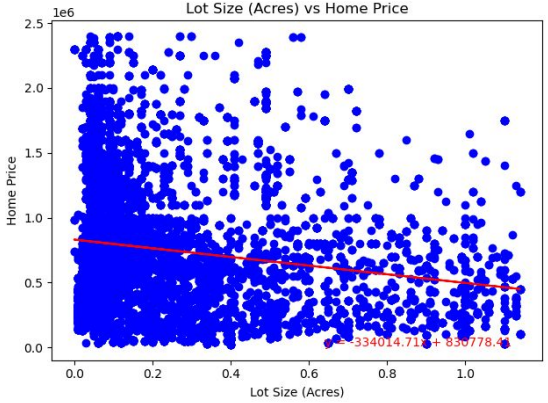
Takeaway:

In the NY specific dataset we analyzed, ~12.92% the variation in House Sale Price can be explained by the variation in House Size, and ~16.51% can be explained by the # of Bathrooms. # of Bedrooms and Lot Size had the smallest impact on the variation in home prices for our dataset specific to New York. Interestingly, New York was the only state with a negative correlation with one of the variables.

New York Visualizations







Q3 Analysis

Question: What is the average pricing for a 3 bedroom/2 bathroom house in the selected states?

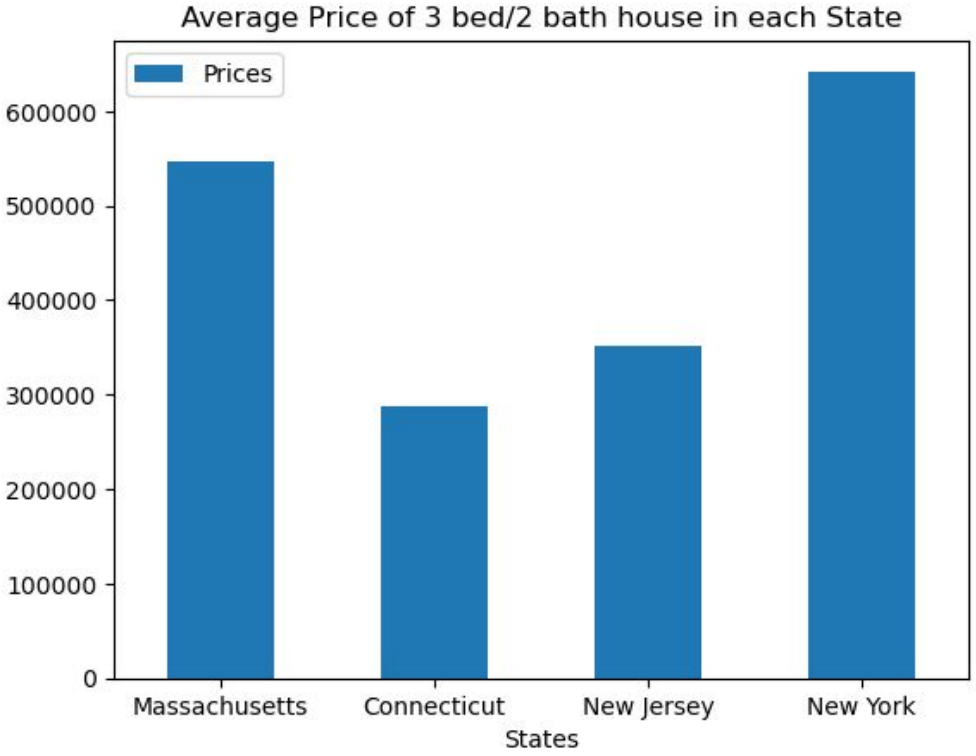
Process:

Create new data frames for each state by filtering our original data frame. Make formulas to find the averages and plot a visual. Also further strengthen the results by creating histograms for each state to see if the higher frequency correlates with the average price that is found.

Steps:

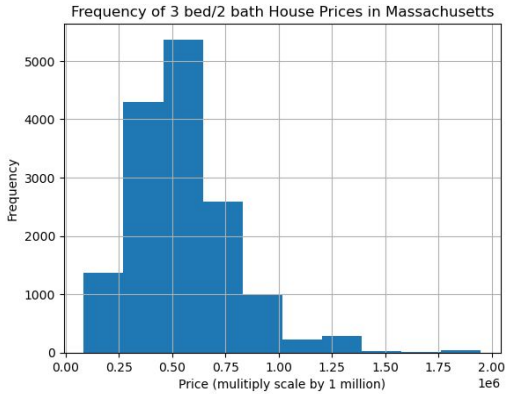
- Create new data frames to only include relevant states and columns.
- Use created dataframes to:
    - Find averages for each state and cumulative
    - Create bar graph of averages
    - Create histograms for each state to show the correlation between frequency of prices and the averages that were found.

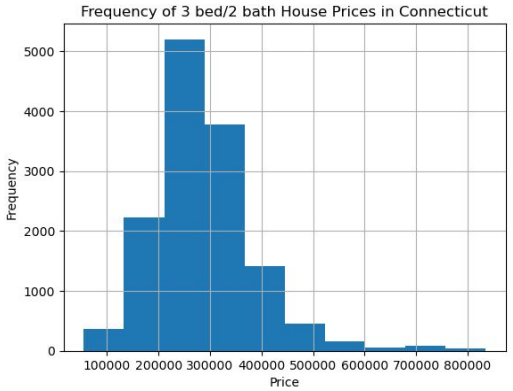
Visualization for Q3 findings

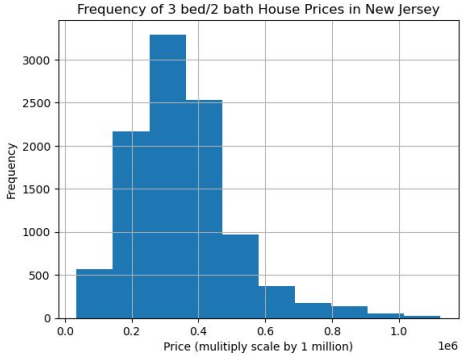


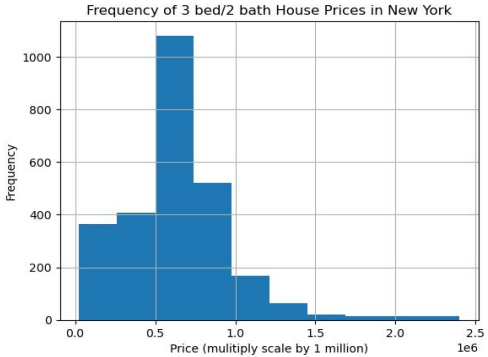
Visualization for Q3 findings (cont.)

In these diagrams, the frequency of the state prices are seen. This shows the average price for each state is consistent with the higher frequency of listings.









Q3 Results

What is the average pricing for a 3 bedroom / 2 bathrooms home by State?

The average New York house price is $642,090.63.

The average New Jersey house price is $350,872.38.

The average Massachusetts house price is $546,770.77.

The average Connecticut house price is $287,699.46.

The average house price for all the states together is $419,604.88.

I found that the average pricing varies between states but is similar between New Jersey and Connecticut and also similar between New York and Massachusetts.

Q4 (Add-On) & Analysis

Q4 a. Identify the Median and Standard Deviation of Price, Bedrooms, Bathrooms, Acreage and Size of homes in all 4 states?

Q4 b. What is the average Price in each of the cities (in the source file) in New York state?

Analysis:

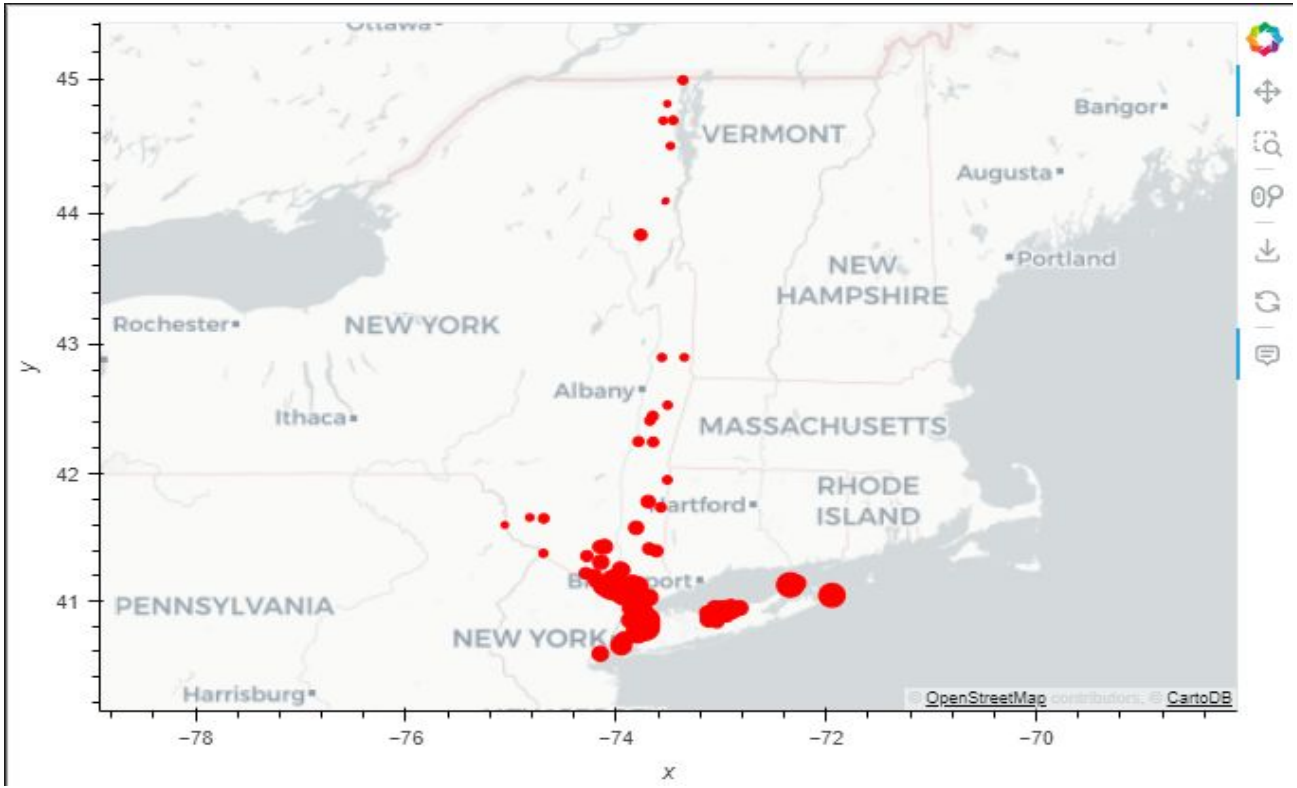
4a. Average Price and Standard Deviation for each state reveals:

1. NY State has the largest Standard Deviation value in comparison with Connecticut, Massachusetts and New Jersey. Reveals it is quite expensive to live in the main cities in and closer to the main cities.
2. Interestingly, the Median Number of Bedroom and Bathroom values is 3 Bedroom and 2 Bathrooms across all 4 States in this study. One may have thought that a 2 Bedroom/ 2 Bathroom may have been the median

4b. Used 'uscities.csv' from kaggle.com to obtain coordinates (latitude and longitude) for each of the 274 cities that were in the clean data source file for NY state.

Only 106 cities of the 274 cities (over less than half) were found in the uscities file. The city map plot for these 106 cities by itself doesn't tell much of a story. Perhaps looking at the average price by city or state over a period of time may be considered in the future utilizing the 'sold date' and the home 'address'.

Q4 Visual NY City Plot



Overall Analysis Report

Key Findings:

- Across all the states that we analyzed the largest impacts on House Price came from the square footage of the house and the number of bathrooms.
- When investigating the correlations on an individual state basis we found that Connecticut and New Jersey behaved differently than Massachusetts and New York.
    - Specifically, Connecticut and New Jersey had higher price correlations with Lot Size than the other two states.
    - Additionally, in New York, there was a negative correlation between price and Lot Size.
- Our Analysis of average price also showed similarities between the two groups of states.
    - Connecticut and New Jersey had average prices much lower than Massachusetts and New York for a typical 3 Bedroom, 2 Bath home.

Takeaway: Our analysis tells us that the differences between the two groups of states must be caused by some other variables aside from the four in our dataset. Our next steps would be to investigate other variables such as cities, population density, income, etc.