Identifying undervalued condos in Northeastern US cities

Executive Summary

With so many condos to choose from, how can buyers quickly identify fair and bargain value condos? The purpose of this capstone project is to answer the following question: despite being a seller's market (as of June 2022), can we use multiple linear regression to find undervalued condos in the Northeastern cities?

Key Findings

The answer is: Sort of.

I originally wanted to tackle this question solely by creating a model to predict condo prices as accurately as possible. To do this, I first built a multi-city multiple linear regression model. I predicted home prices to be within $363K or 59% of the actual price on average. The accuracy was not great, so I trained a Gradient Boosting model to predict home price, which reduced the error to within $243K and 24% of the actual price on average.

These models found the following features to be predictive of condo price: city, square footage, bathrooms, bedrooms, HOA fee, attached garage, walk score, transit score, commute time, median age, marriage status, median household income greater than 150K, home ownership, and year-built post 1980.

Recognizing that city plays a bigger role in accurately predicting condo prices and my multi-city model is trained on 75% NYC data, I built another multiple linear regression using Philadelphia condos. This model allowed me to predict condo price within $246K and 44% of actual price on average. I also trained a Gradient Boosting model to predict home price, which reduced the error to $135K and 21% of actual price on average. Gradient Boosting outperformed regression in the multi-city and Philadelphia model because it accounts for non-linear relationships. The Philadelphia specific model included a smaller set of predictors to the multi-city model with the addition of neighborhood.

After training these models, I learned there is more to the story. From this process, I realized it's tough to price condos because some factors are hard to label. For example, neighborhood safety affects condo price, but in the United States, crime is an ever-changing situation with recent gun violence activity in previously safe locations. It's also much easier to assess property condition and desirable architectural styles by looking at pictures than to label them.

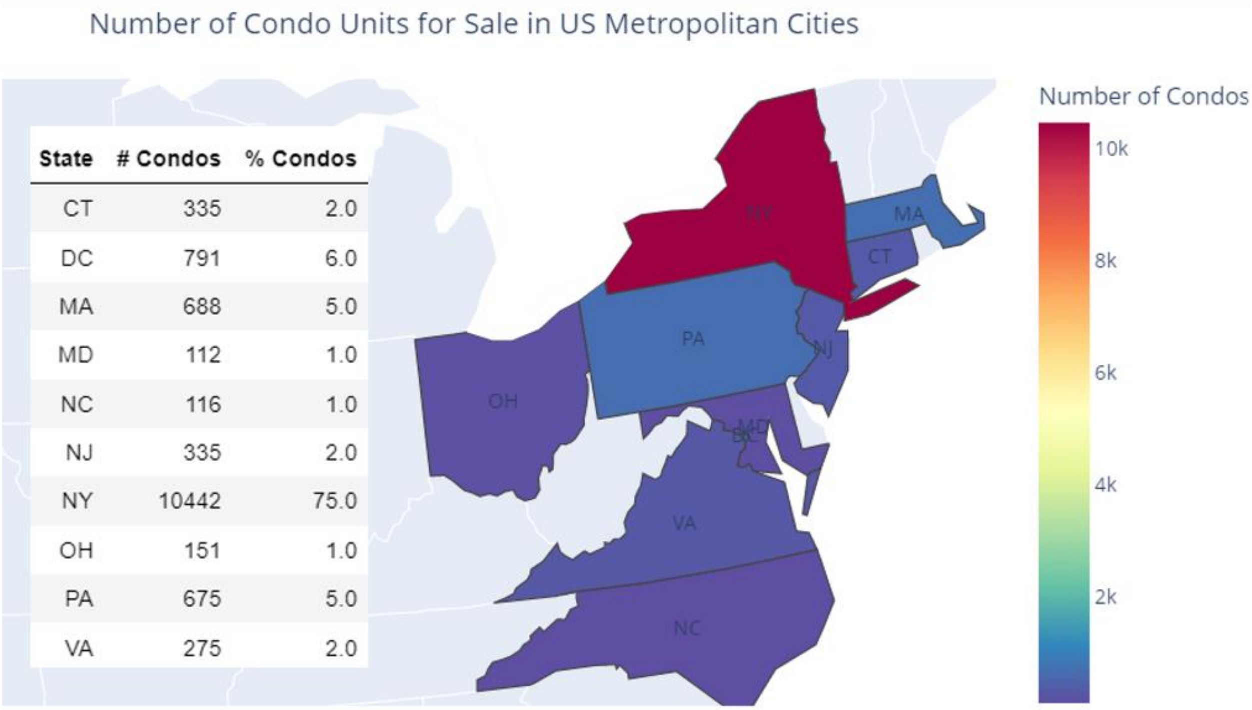
Other issues that make it challenging to accurately model condo prices include data quality (missing or messy data), how to encode eccentric property characteristics (desirable water views, penthouse/condo hotels, owned by famous figures), data availability (there's no simple way to pull crime/neighborhood quality data), and determining the best algorithm to use. Sentiment reviews are important but are complicated to integrate into a model since reviews can come from so many places.

Even Zillow could not profitably automate home buying based on their own model. Zillow's overreliance on their model to automate this process failed: Zillow cut 25% of their workforce and wrote off $500 million this year.

For these reasons, I ultimately created a decision support tool that combines model predictions, a PowerBI dashboard, and in-depth research on promising condos that appear undervalued, which accomplishes the original intent of this project.

Introduction

I acquired property listing data for 13,920 condos in 16 Northeastern cities.

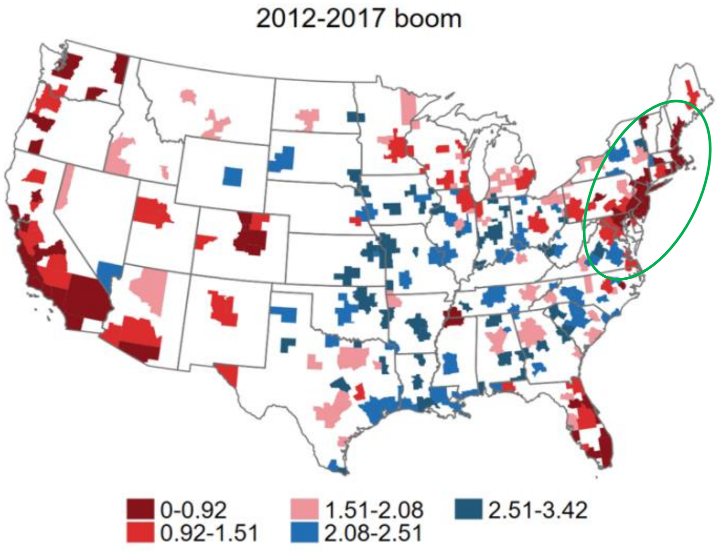


Number of Condo Units for Sale in US Metropolitan Cities

State	# Condos	% Condos
CT	335	2.0
DC	791	6.0
MA	688	5.0
MD	112	1.0
NC	116	1.0
NJ	335	2.0
NY	10442	75.0
OH	151	1.0
PA	675	5.0
VA	275	2.0

Why the Northeast?

1. Elasticity is the change in demand given change in price. The Northeast region exhibits price inelasticity (denoted in red below), meaning that supply cannot keep up with demand and there are a lack of substitutes. Northeastern states are also more densely populated and have bodies of water that limit the land available for construction, so are more comparable than regions with more open space.



Notes: Estimated elasticities for each housing boom. Smaller values refer to lower elasticity areas.

Analytical method

Step 1: Data Collection and Preprocessing

I selected 16 Northeastern US cities to analyze based on where I would like to live, and the fact that the Northeast is characterized by supply constraints and price inelasticity.

Specifically, this dataset is comprised of the following cities.

City	State
Alexandria	Virginia
Arlington	Virginia
Baltimore	Maryland
Boston	Massachusetts
Brooklyn	New York
Charlotte	North Carolina
Columbus	Ohio
Connecticut	Connecticut
Jersey City	New Jersey
Manhattan	New York
Philadelphia	Pennsylvania
Pittsburgh	Pennsylvania
Queens	New York
Staten Island	New York
Washington D.C.	Washington D.C.
Yonkers	New York

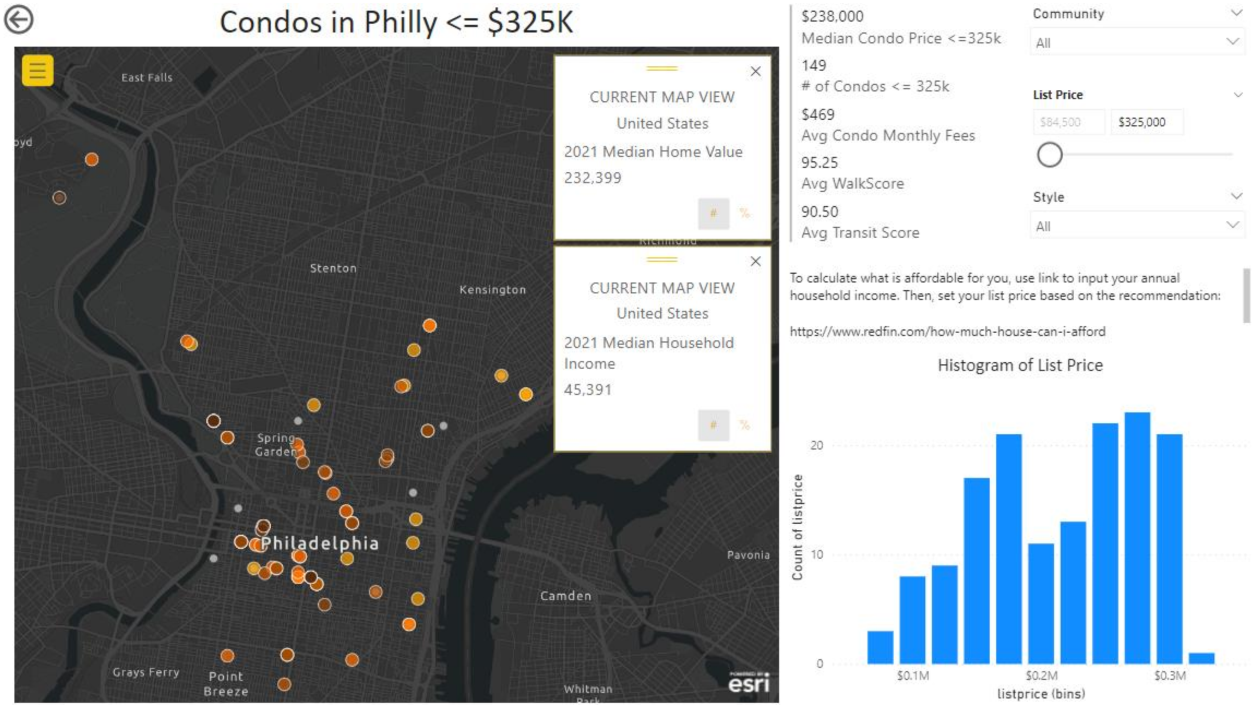
The property dataset contains 13,920 condo records from Rapid API's Zillow.com API, where I pulled from Extended Search, Property Details, and Walk and Transit Score.

Other important features, such as population density, median household income, and commute time, were obtained from data aggregator simplemaps.com.

I cleaned and merged the data sets, imputed missing values, and removed condos with extremely high prices (>$5.995M).

Step 2: Merge Data to Postgres to PowerBI Dashboard

I appended all condo records into one dataset and overlaid local Philly shootings to visualize crimes near condos. This merged dataset was pushed to PostgreSQL, PostgreSQL was connected to PowerBI. Both datasets have lat-long coordinates that allow PowerBI to plot them on a map. This visualization helps gauge the safety of a neighborhood where condo prices are listed on sale, displays geographic patterns in condo pricing, and provides further context to prospective homebuyers.



Condos in Philly <= $325K

CURRENT MAP VIEW
United States
2021 Median Home Value
232,399

CURRENT MAP VIEW
United States
2021 Median Household Income
45,391

$238,000
Median Condo Price <=325k

149
# of Condos <= 325k

$469
Avg Condo Monthly Fees

95.25
Avg WalkScore

90.50
Avg Transit Score

Community: All

List Price: $84,500 – $325,000

Style: All

To calculate what is affordable for you, use link to input your annual household income. Then, set your list price based on the recommendation:

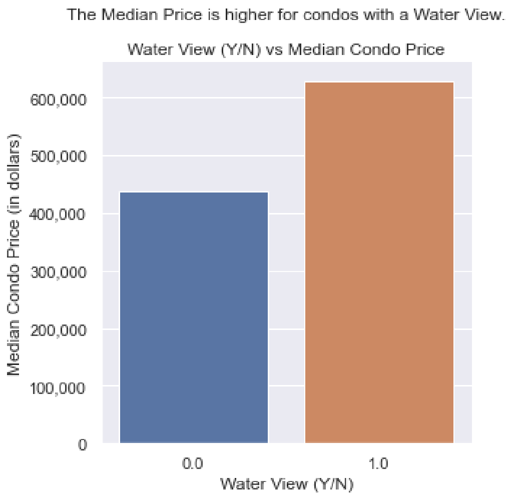
https://www.redfin.com/how-much-house-can-i-afford

Histogram of List Price

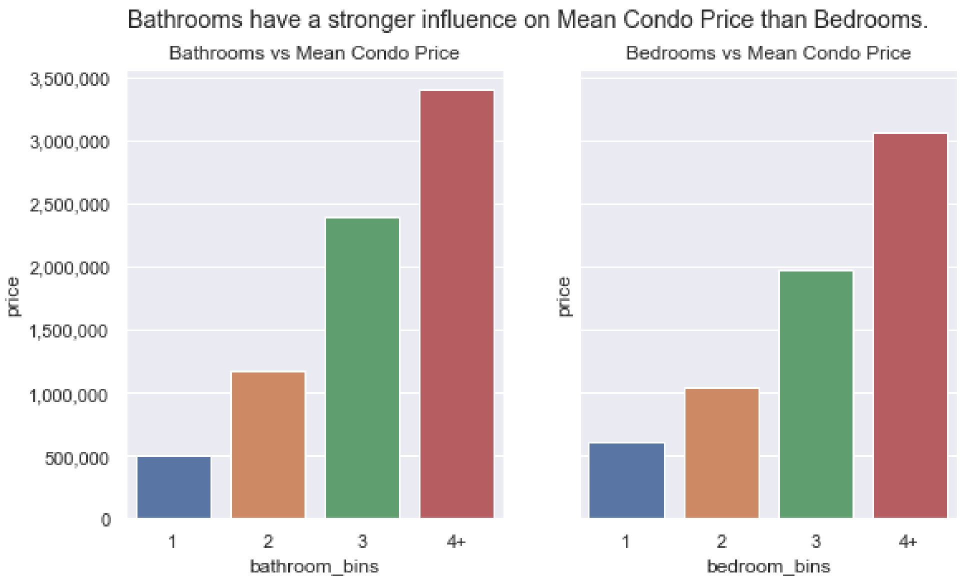
Step 3: Exploratory Data Analysis (EDA)

As part of my quest to identify undervalued condos, I wanted to understand which factors impact condo price, based on a detailed literature review and verifying those trends in my data set.

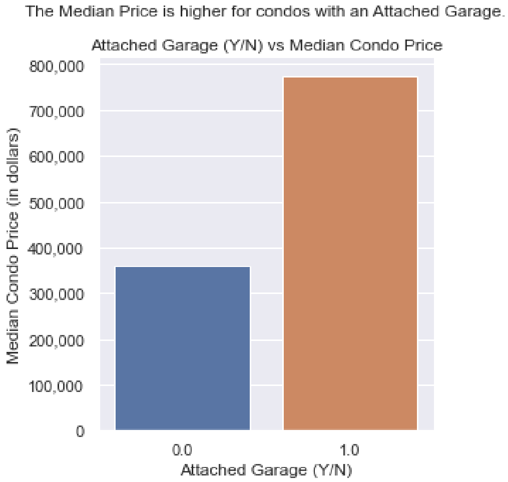
1. Literature review suggests price is higher for condos with a sought-after water view. After inspecting the data, I observed that the median price of a condo with a water view is 50% higher! Not all water views are created equal, however. The locations where my model underestimated the value of a water view – i.e., the most desirable water views – were in Boston Waterfront, Hudson Yards, and Baltimore Harbor.



2. Literature review suggests the more bedrooms and bathrooms, the higher the condo price. After inspecting the data, this is true, and in fact bathrooms have a stronger influence on price than bedrooms. I found that the most common (modal) condo has two bedrooms and one bathroom.

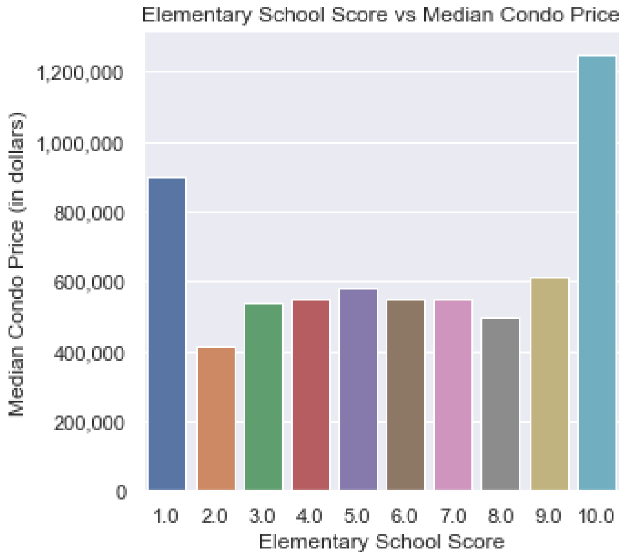


3. Literature review suggests that an attached garage increases condo's price. The data supports this. The median price of a condo with an attached garage is $775k, compared to $360k among condos without an attached garage. In other words, more than double the price!



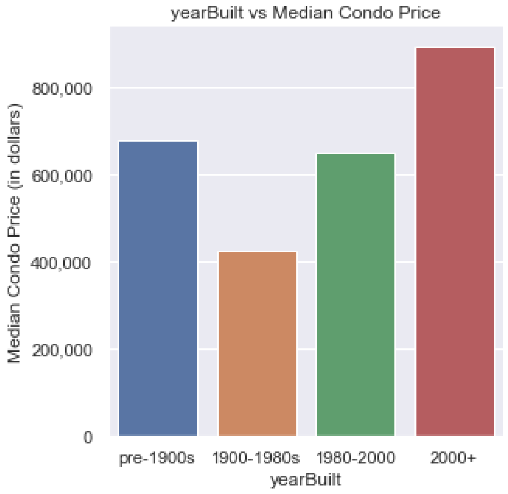
4. Literature review suggests the higher the elementary school score, the higher the condo price. Better elementary schools help students qualify for magnet highs schools that prep student for a better college experience. This environment will help students secure higher paying jobs in the future.

   My data set does not support this theory. Condos in areas with an elementary school score of 10 (the highest rating) have the highest median condo price. However, there is not a strong relationship between elementary school quality and condo price for elementary schools with scores of 1-9. Condos with an elementary school score of 1 have a higher median price, but represent only 1% of records, so may not be representative. In this case, there may be some expensive condos next to poorly rated public schools where students are sent to private schools instead.


The Median Price is highest for condos
where the Elementary School Score has a 10 rating.

I am initially surprised that the correlation coefficient is low (r=.16). One potential explanation is that condo buyer motivations are different. Condo buyers with or plan to have children place value in elementary school score, while those without may not, so that relationship gets obscured.

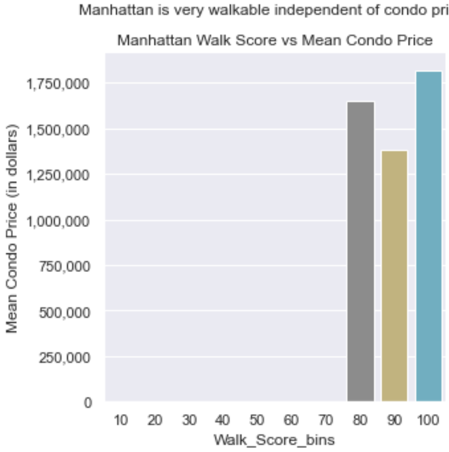
5. Literature review suggests that both historic homes and new builds have higher prices. After inspecting the data, year built has a U-shaped relationship relative to price, which is consistent with the prior research. Newer builds tend to be larger, have modern aesthetics, and do not require as much maintenance or renovation, while historic homes pre-1900s are treasured because of its rich history, architecture, and charm. They are likely in dense, central locations that are attractive for tourism and can offer tax benefits.

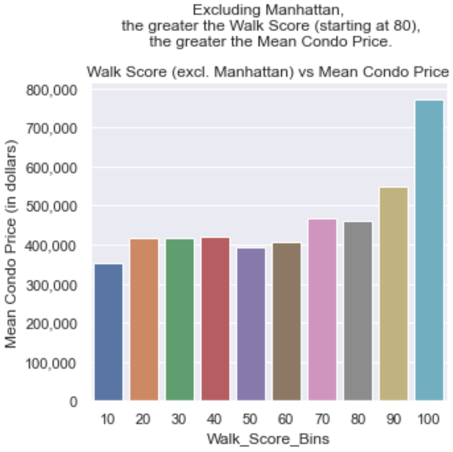

The Median Price is higher for historic and newer condos.

Numerical variable analysis

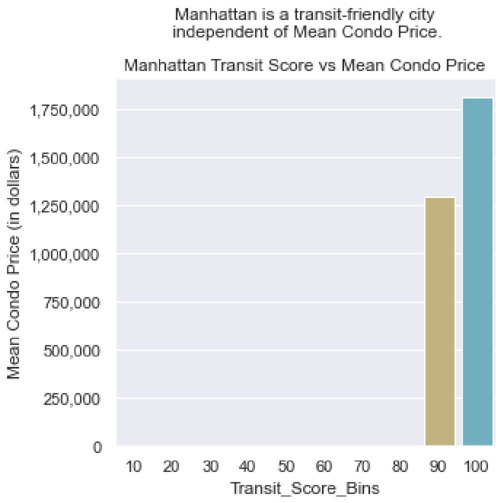
6. Sun score is a composite score that measures your property's solar power potential to install solar panels. While research suggests installing solar panels increases property value, after inspecting the data, sun score appears to have no relationship with price. This could be because many condos – especially in NYC – are in apartment buildings, where individual condo owners cannot choose to install solar panels regardless of the building's solar potential.

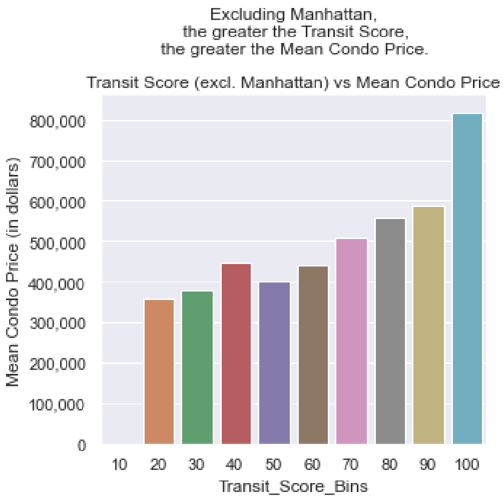
7. Literature review suggests the better the walk score, the higher the condo price. Manhattan is an anomaly since it is a very walkable city independent of condo pricing and very expensive. For that reason, I separated Manhattan from the rest of the data to unearth other relationships between walk score and price. Even after separating Manhattan results, walk score still follows an exponential relationship. Excluding Manhattan, 75% of condo units in my data set are in very walkable areas. Walkability likely increases condo prices because amenities and transit are easily accessible, which is desirable.





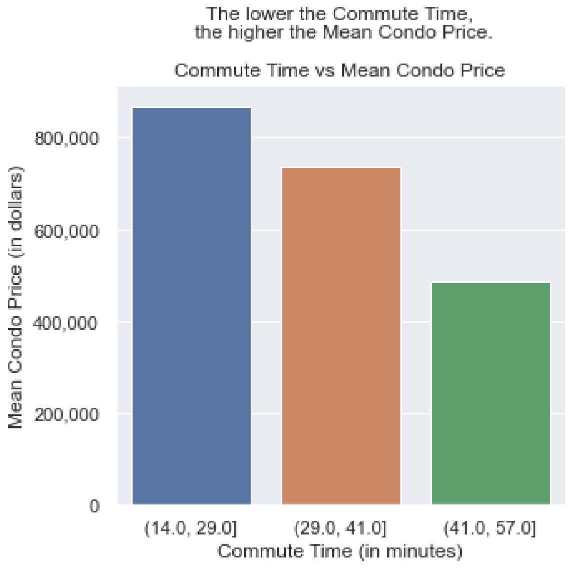
8. Literature review suggests the better the transit score, the higher the property price. After running a correlation matrix, it confirms transit score is highly correlated with walk score (correlation coefficient = 0.75) and have higher prices. Locations with a higher walk score typically have good public transit due to thoughtful urban planning.





9. Based on literature review, the higher the HOA fee, the lower the condo price. After inspecting the data and dropping null HOA fees, it appears monthly HOA fees is weakly positively correlated with price (correlation coefficient = 0.07) and not negatively correlated. Monthly HOA fees could be calculated as a percentage of condo fees and not an offset, which would explain the positive correlation.

10. Literature review suggests the lower commute times, the higher the property value. I binned commute time into 3 quantiles and used the mean price instead. After inspecting the data, this is consistent with literature review, that commute times are negatively correlated with condo price (correlation coefficient=-.37). In general, people are willing to pay for a unit that shortens their commute time, so they have more free time for other activities.



Step 4: Modeling Building

Predictive modeling is the process of applying a statistical model or an algorithm to data to predict future observations.

It is important to use in this analysis because traditional house price prediction is based on cost and sale price comparison and lacks a standard process of evaluation and certification. A predictive model solves for a lack of objective criteria, an information gap, and helps uncover the pricing inefficiencies in real estate.

Multiple linear regression is a machine learning technique used to predict the outcome of a continuous variable based on the value of two or more variables. With multiple linear regression, it is easier to interpret and quantify how much each numerical and categorical feature affects the condo price or predict a future condo price with new data.

Gradient-boosted decision tree is a machine learning technique that accounts for non-linear relationships between factors but is less interpretable. It trains a sequence of decision trees, with each successive tree attempting to minimize the error of the previous tree, which can improve predictive accuracy.

I explored these two algorithms because multiple linear regression resulted in a high error, so I tried a gradient-boosted decision tree to improve the model fit.

Specific error metrics, such as MAE, MAPE, and RMSE, are considered to evaluate the performance of the two algorithms.

- The Mean Absolute Error (MAE) is the average of the absolute error in dollars and is most interpretable. It measures predictive accuracy. From these results, it means, on average, our model predictions are off by ± the MAE value.
- The Mean Absolute Percentage Error (MAPE) measures how far the model's predictions are off on average expressed as a percentage. In other words, the average percentage difference between the predicted value and the actual value is the MAPE.
- The Root mean squared error (RMSE) measures the average magnitude of the error in dollars, and compared to the MAE, penalizes large errors. A high RMSE is not necessarily a bad thing, because large errors can help pinpoint undervalued condos where predicted price is greater than actual price, so RMSE will not be deeply assessed to measure model performance.

Feature Importance

Literature review suggests bathrooms, bedrooms, square footage, expensive cities, attached garages, sought-after water views, good walk/transit scores, year built post-1980s, and positive HOA fees are positively correlated with condo price. A lower commute time also positively impacts condo price. I leveraged Azure Machine ML Studio to quickly understand feature importance of these predictors in the model.

After running feature importance from the gradient boosting model, I noticed all the features above and demographic information like marriage status, income > $150k, and median age, are important in the aggregate model. I'm not at all surprised. 75% of this data reflects characteristics of NYC and their residents/preferences. Excluding demographic attributes reduces R^2 from 0.74 to 0.73 and increases error. Overall, I agree with the features identified as important, but I expected more features to be significant that were not.

Factors I initially thought were important, but the model did not, are elementary school score, population density, sun score, and days on Zillow. These features may be influential depending on the city. Architectural style and property condition are important, but only if their labels are more precise. It's easier to assess these two through pictures or viewing the property in person than in the dataset.

My final linear regression model predicted condo price within $363K, or 59%, on average (MAE and MAPE, respectively). This is a high error given that the average condo price is $1.018 Mil. As a result, I trained a Gradient Boosted Decision Tree model to improve accuracy. Even though the model is not as interpretable, gradient boosting can yield better performance and accuracy because it accounts for non-linear relationships between the features and price.

Step 5a: Validating Errors

An MAPE of 59% means the average prediction's predicted price and actuals price are far off based on the features I fed into the regression model. While the error is large, it is still useful to look at properties where the prediction is higher than the actuals. That signals that the model predicts the condo is undervalued, which may make it more appealing to prospective buyers.

To validate the model, I looked at nine undervalued condos in Philadelphia, which were all the errors where actual price was less than $325K. I ranked the errors from highest to lowest where predicted price is greater than actual price and reviewed each record. Analyzing errors help me determine whether there's a pricing inefficiency I can exploit or there's some qualities of the condo the model just cannot capture.

Multi-city Model - Undervalued Condos

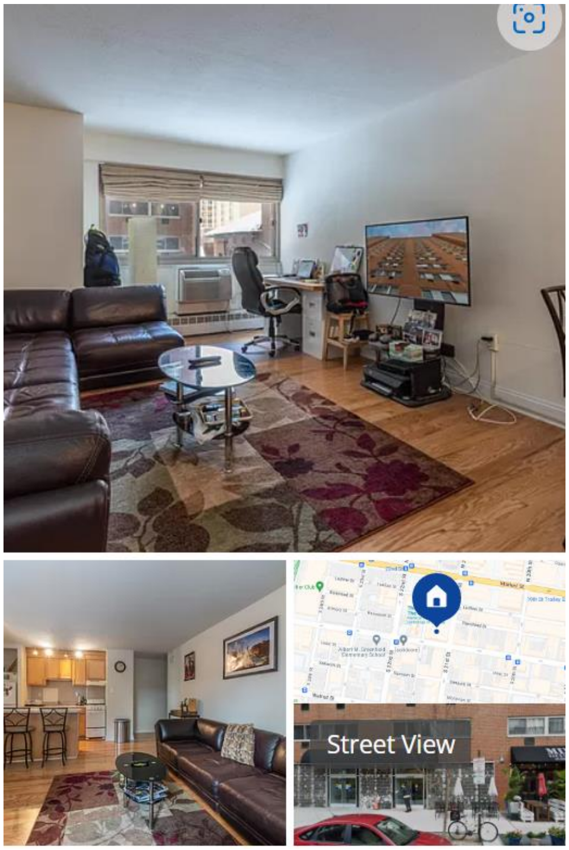
Condos that are undervalued in this price range have the essentials for a great condo! Four of the nine records are Philadelphia condos which have now been sold in the two months since the data was extracted. What the model could not capture is that the remaining five condos that have not sold have some flaw: they are either in a bad neighborhood, have poor management, have poor conduct by workers, are noisy, or a majority of residents of a condo building are senior citizens (that means people might change their behavior to accommodate demographic need). This can only be discovered by reading reviews about the condo building and is not easily encoded in the dataset in a way that can be captured by a predictive model.

1. **Plain but great price.** 118 S 21 St 618. Predicted Price: $170,354, Actual Price: $98,115. This has a pending offer. It is a bargain in Rittenhouse Square, an affluent neighborhood. It offers 24/7 security, a rooftop deck, air conditioning, and is close to food/shops. It could have taken better pictures, but it's a bargain.

2. **Progressive style in a hip neighborhood.** 810 N Hancock St 2. Predicted Price: $338,152, Actual Price: $320,000. This is sold. It's in North Liberties, a young professional neighborhood close to food and entertainment. There's a garage, natural lighting throughout the property, and a rooftop deck. It has a very modern design, which I enjoy.

3. **Average and livable.** 2101 Chestnut 403. Predicted Price: $180,120. Actual Price: $174,500 (Sold Price:$160k). This is sold. Riverwest Condos is in Rittenhouse, close to Trader Joes, restaurants,

and walking distance to universities. The downside is it looks very typical.



Step 5b: Error Findings

NYC represents 75% of the condos in the data set, so the model reflects more of what NYC condo buyers value. This contributes to larger percentage errors for condos in the remaining non-NYC cities in the dataset. The MAE for NYC is $349K, MAPE is 43%, and the MAE for Non-NYC is $281k, MAPE is 61%.

The highest overvalued errors in the original multi-city linear regression model seem to be condos with unique and distinctive qualities such as specific water views, homes where famous figures once resided or utilized, condos located in hotels, and penthouses.

As with the undervalued condos, there are many ways a condo could be unique that dramatically increase its value but are difficult to encode in the data set in a meaningful way, which increases error.

Step 5c: Improving model fit by binning numerical features to capture non-linear relationships

I leveraged Azure ML Studio to efficiently test which algorithm generated the best fit and minimized error. Boosted Decision Tree Regression has much better performance than a multiple linear regression. This could mean that some features have a non-linear relationship with price or have interactions with other features. The multiple linear regression could be improved if it accounted for non-linear relationships such as by binning numeric predictors into groups.

While errors could indicate pricing inefficiency, it is also likely that Multiple Linear Regression is just unable to capture non-linear relationships. It's important to account for this so we can isolate pricing inefficiencies as much as possible. Gradient Boosted Decision Tree produced an R^2 of 0.83. On average, its predictions were off by ± $243k (vs. $346k for Multiple Linear Regression, a $103k improvement). Ultimately, Gradient Boosted Decision Tree yielded a better fit and reduced error.

Step 5d: Multicollinearity

Multicollinearity is detected when independent variables in the regression model are highly correlated with each other. Multicollinearity does not impact model fit, but it does make the coefficients more difficult to interpret. I used Variable Inflation Factor (VIF) to identify variables with a high degree of collinearity, but ended up leaving collinear predictors in the regression to maximize predictive accuracy.

Step 6: Philadelphia Multiple Linear Regression Model

Suspecting that a multi-city model would not be as precise as a local model, I used Azure ML Studio again to build a Boosted Decision Tree Regression on only the Philadelphia condo records. First, feature importance is used to determine which features provided the most information. Then the same features were fed into the multiple linear regression model, but any features with p>.05 were dropped. By focusing on Philadelphia, I was able to add dummy variables for highly desirable Philadelphia neighborhoods, to better tailor the predictions by neighborhood.

In the Philadelphia dataset, the year-built feature is more significant. Bathroom, bedroom, walk score, square footage, and HOA fee are still important. All other features considered important in the multi-city model are no longer important.

None of this is surprising. Philadelphia prides itself on historical preservation (our founding fathers signed the Declaration of Independence in Philadelphia) which is why year built is a key differentiator in this market. A well-maintained home vs one that needs work (beyond aesthetics) can be the determining factor between buying and passing on comparable homes with all the essential features.

Philadelphia Model - Undervalued Condos

I used the Philadelphia Multiple Linear Regression model to identify additional undervalued condos. Capping price at a $325K threshold, I identified 14 total undervalued condos, 3 of which I like:

1. **Clean layout, high ceilings, and natural sunlight.** 444 N 4th St Unit 313. Predicted Price: $368,817, Actual Price: $305K (Sold Price: $299,900).

In Northern Liberties, a young professional neighborhood, the visuals are impressive. The condo uses updated design concepts.

2. **Cozy layout and design.** 1431 N 5th St #1. Predicted Price: $800k, Actual Price: $320K. The predicted price is high because it has 3 bedrooms, 2 bathrooms, and is 1769 sq ft.

While Kensington overall is not known to be safe, Old Kensington is the most gentrified because it's by Northern Liberties, so safety is improving as more young professionals move into this neighborhood.

3. **The living room, kitchen, and bathroom design are appealing.** 3900 Ford Rd APT 3O Predicted Price: $453,690. Actual Price: $315K (Sold Price: $80K).

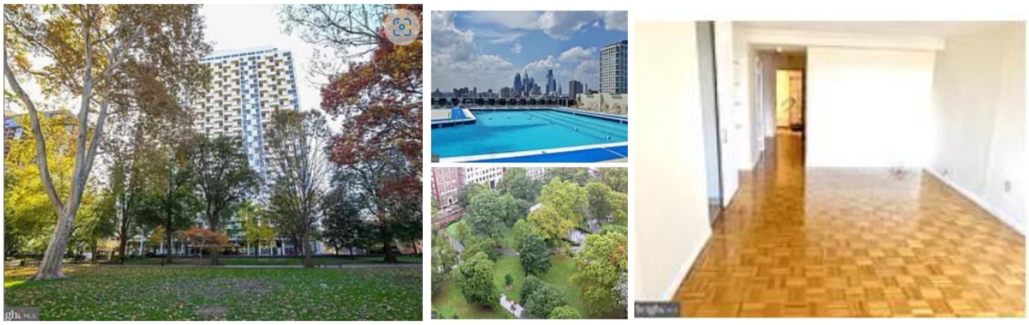
Park Plaza is downtown. It looks like a great place to host friends.

While I found a few hidden gems listed above, most undervalued condos have an average appearance in terms of condition and style. As a prospective buyer, I want to buy something impressive. While my model did not include any features on quality or condition, a more advanced machine learning model may be able to extract a listing's visual attractiveness by analyzing ad scoring the listing's photos.

Philadelphia Model - Overvalued Condos

Next, capping price at a $325K threshold, I looked at overvalued condos to assess the validity of those results. I agree with 12 of the 14 errors. Either they are in buildings that are relatively average or in average neighborhoods. There were 2 overvalued errors that caught my eye.

1. 604 S Washington Sq APT 1806. **Of Sentimental Value.** Predicted Price: $-238K. Actual Price: $217K (Sold Price: $202K). It has a fantastic view but is overvalued by $455k. It has 0 bedrooms and is small at only 600 square feet but is right by a beautiful square. It is one of my favorite spots to talk to friends! Because I included neighborhoods in my model (this neighborhood is cheaper) and no HOA fee was listed, the predicted price came out negative.



2. 1931 Spruce St 1C. **Modern and chic layout.** Predicted Price: $215k. Actual Price: $275K (Sold Price: $256K). It is overvalued by $60k, but the renovations and décor are impressive. The model cannot capture how well maintained or stylistically unique a condo is.

In summary, these two overvalued condos require compromises between less attractive and more attractive features.

By focusing only on Philadelphia, the error metrics improved. MAE for the Philly Gradient Boosted Decision Tree is $135k, vs. $246k MAE for the Philly Multiple Linear Regression. The MAPE for the Philadelphia GBM is 20.6% vs. 44.1% for the MLR.

At both the Northeast regional and local Philadelphia level, Gradient Boosted Decision Trees outperform Multiple Linear Regression in accurately predicting condo price.

The Philly Multiple Linear Regression outperforms the Multi-City Multiple Linear Regression, with an R^2 improvement of 9%. The MAE decreased from $363k to $246k and MAPE went down from 59% to 44%.

Bottom line: it is more useful to model and evaluate undervalued condos at the city-level, because both algorithms yielded more interesting errors to think about compared to the aggregate (multi-city) models. However, even the city-level models cannot distinguish hidden gems (great style, upkeep, great neighborhood) vs condos that have potential but there's no effort to renovate, restore, or have poor reviews. A purchase strategy is if you find a condo you like that has already sold but fits your criteria is to wait until another comparable unit pops up in the same building.

Step 7: Conclusion/Findings

I initially wanted to identify undervalued condos by creating a model to predict condo prices as accurately as possible. At first, I approached this by building a multi-city multiple linear regression model with relevant features. The accuracy was not great, so I trained a gradient boosting model which improved this metric. Sensing that different cities value different condo features, I built a multiple linear regression model using Philadelphia condos, which reduced my errors compared to the original multi-city multiple linear regression model. The Philadelphia specific model included a smaller set of predictors to the multi-city model with the addition of neighborhood. Gradient Boosting outperformed regression in the multi-city and Philadelphia model because it accounts for non-linear relationships.

However, there are still challenges in modeling condo pricing due data quality (missing or messy data), how to encode eccentric property characteristics (desirable water views, penthouse/condo hotels, owned by famous figures), and data availability (there's no simple way to pull crime/neighborhood quality data). Sentiment reviews are important but are complicated to integrate into a model since reviews can come from so many places.

As a result, my approach evolved from strictly using the model to identify undervalued properties, to using it – along with the PowerBI dashboard – as decision support tools alongside in-depth (and manual) research. Oversimplifying the home buying process is too costly and can lead to buyer's remorse. Especially if I am only planning to buy one home in the foreseeable future, I want it to be worth my while.

This process has given me inspiration to explore condo buildings that I have not considered before because I've been downtown-centric in my search. I plan to visit at least one property to make sure what I see is what I expect and read through reviews.

When the zestimate fails:

The zestimate is Zillow's house valuation tool that "allows users to see how much homes are worth. The zestimate is based on information from sources like comparable sales and public data. Zestimates are only as accurate as the data behind them, meaning they may be outdated or incorrect".

I did not use zestimates in this project because Zillow's overreliance of the zestimate led to the company laying off 25% of their workforce and writing off $500M this year.

Zillow leadership’s strategy was to use the zestimate to buy undervalued properties and flip them. The strategy failed since Zillow purchased properties at too high a price. The intent of the zestimate is to be used starting point of conversation about value, not the final decision. Zillow did not get input from agents, brokers, or appraisers (professionals who have been inside the home) to accurately assess the value of the property. Doing so would have reduced the number of overvalued properties they bought.

Another weakness is that Zillow data is not always correct, or there are edge cases like view obstruction/noise that penalize price that humans can easily see that machine learning cannot detect. Finally, Zillow’s zestimate is not designed to adjust to the rapid changes in the volatile real estate market. That is why I use predictive models as a starting point to identify potentially undervalued properties, but ultimately much more information is gained by viewing the property and talking to appraisers.

If there’s anything we’ve learned from Zillow’s failure and the 2008 subprime crisis, it is to verify that your original assumptions are correct before making large transactions.

Final Thoughts

On the topic of the tradeoff between interpretability vs. accuracy with predictive models, most roles require translating technical concepts for non-technical people. That is why it is important to use a model that can be easily explained. However, there are still benefits of using machine learning without entirely understanding how it runs under the hood. If an algorithm continues to work, more insights will follow, so the interpretability lags the technology being deployed. If the algorithm fails, people abandon it, and there’s no longer feedback. It's good to know which algorithm performs the best, just in case this plays a bigger role in the future. For this project, using multiple linear regression was the best choice to keep it simple.

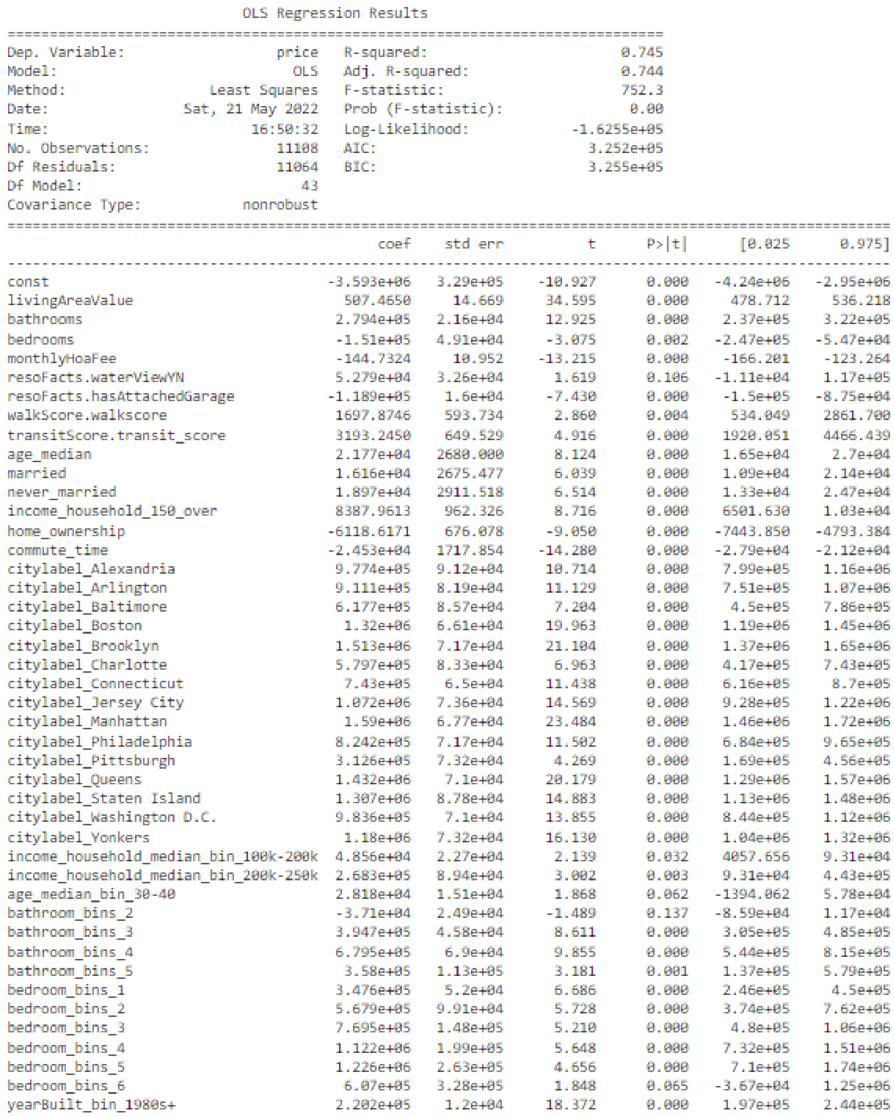
Step 8: Next steps

If there was more time, here are other ways to enhance the project:

- Look at undervalued Philadelphia condo errors above $325k for more inspiration and to understand what else the model can’t account for. Streamline city-level error outputs.
- Analyze condos by city. Each city has a differentiator and faces different risks. A differentiator in Boston is hiking trails and the waterfront view. A risk in Baltimore is environmental flooding. Each housing market is very different, and one may get better results focusing on one city. Note the condo inventory on sale is lower than 1,000 units in every northeastern city except New York City. That means that while one may get better results focusing on one city, there is a tradeoff in terms of data availability to train the model.
- Given that in New York City, Jersey City, and Boston, residents primarily rent because they moved there for work, are more transient, and it’s more expensive to own property, does it make more sense to rent instead of buy in these locations? This project can be replicated and repurposed to identify affordable apartments for rent, using monthly rent as the target variable.

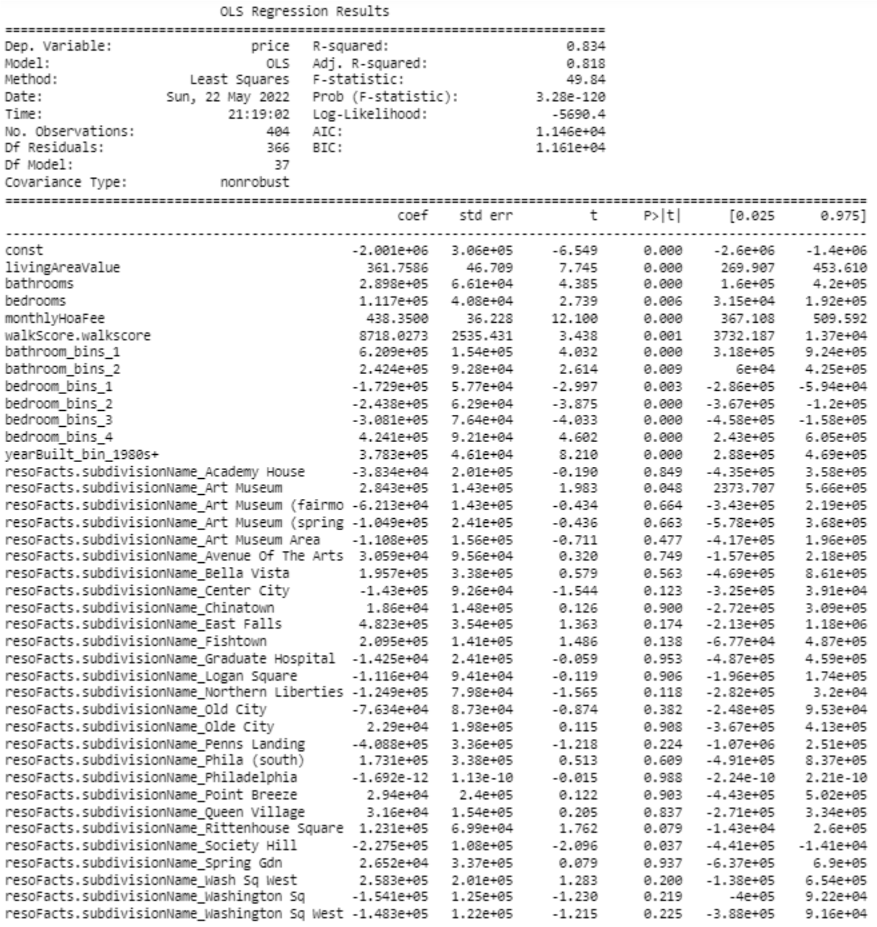
Appendix

Northeastern Cities Multiple Linear Regression Model:



```
                              OLS Regression Results
==============================================================================
Dep. Variable:                 price   R-squared:                       0.745
Model:                           OLS   Adj. R-squared:                  0.744
Method:                Least Squares   F-statistic:                     752.3
Date:               Sat, 21 May 2022   Prob (F-statistic):               0.00
Time:                       16:50:32   Log-Likelihood:            -1.6255e+05
No. Observations:              11108   AIC:                         3.252e+05
Df Residuals:                  11064   BIC:                         3.255e+05
Df Model:                         43
Covariance Type:           nonrobust
=======================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
const                                 -3.593e+06   3.29e+05    -10.927      0.000   -4.24e+06   -2.95e+06
livingAreaValue                         507.4650     14.669     34.595      0.000     478.712     536.218
bathrooms                              2.794e+05   2.16e+04     12.925      0.000    2.37e+05    3.22e+05
bedrooms                               -1.51e+05   4.91e+04     -3.075      0.002   -2.47e+05   -5.47e+04
monthlyHoaFee                          -144.7324     10.952    -13.215      0.000    -166.201    -123.264
resoFacts.waterViewYN                  5.279e+04   3.26e+04      1.619      0.106   -1.11e+04    1.17e+05
resoFacts.hasAttachedGarage           -1.189e+05    1.6e+04     -7.430      0.000    -1.5e+05   -8.75e+04
walkScore.walkscore                    1697.8746    593.734      2.860      0.004     534.049    2861.700
transitScore.transit_score             3193.2450    649.529      4.916      0.000    1920.051    4466.439
age_median                             2.177e+04   2680.000      8.124      0.000    1.65e+04     2.7e+04
married                                1.616e+04   2675.477      6.039      0.000    1.09e+04    2.14e+04
never_married                          1.897e+04   2911.518      6.514      0.000    1.33e+04    2.47e+04
income_household_150_over              8387.9613    962.326      8.716      0.000    6501.630    1.03e+04
home_ownership                        -6118.6171    676.078     -9.050      0.000   -7443.850   -4793.384
commute_time                          -2.453e+04   1717.854    -14.280      0.000   -2.79e+04   -2.12e+04
citylabel_Alexandria                   9.774e+05   9.12e+04     10.714      0.000    7.99e+05    1.16e+06
citylabel_Arlington                    9.111e+05   8.19e+04     11.129      0.000    7.51e+05    1.07e+06
citylabel_Baltimore                    6.177e+05   8.57e+04      7.204      0.000     4.5e+05    7.86e+05
citylabel_Boston                        1.32e+06   6.61e+04     19.963      0.000    1.19e+06    1.45e+06
citylabel_Brooklyn                     1.513e+06   7.17e+04     21.104      0.000    1.37e+06    1.65e+06
citylabel_Charlotte                    5.797e+05   8.33e+04      6.963      0.000    4.17e+05    7.43e+05
citylabel_Connecticut                   7.43e+05    6.5e+04     11.438      0.000    6.16e+05     8.7e+05
citylabel_Jersey City                  1.072e+06   7.36e+04     14.569      0.000    9.28e+05    1.22e+06
citylabel_Manhattan                     1.59e+06   6.77e+04     23.484      0.000    1.46e+06    1.72e+06
citylabel_Philadelphia                 8.242e+05   7.17e+04     11.502      0.000    6.84e+05    9.65e+05
citylabel_Pittsburgh                   3.126e+05   7.32e+04      4.269      0.000    1.69e+05    4.56e+05
citylabel_Queens                       1.432e+06    7.1e+04     20.179      0.000    1.29e+06    1.57e+06
citylabel_Staten Island                1.307e+06   8.78e+04     14.883      0.000    1.13e+06    1.48e+06
citylabel_Washington D.C.              9.836e+05    7.1e+04     13.855      0.000    8.44e+05    1.12e+06
citylabel_Yonkers                       1.18e+06   7.32e+04     16.130      0.000    1.04e+06    1.32e+06
income_household_median_bin_100k-200k  4.856e+04   2.27e+04      2.139      0.032    4057.656    9.31e+04
income_household_median_bin_200k-250k  2.683e+05   8.94e+04      3.002      0.003    9.31e+04    4.43e+05
age_median_bin_30-40                   2.818e+04   1.51e+04      1.868      0.062   -1394.062    5.78e+04
bathroom_bins_2                        -3.71e+04   2.49e+04     -1.489      0.137   -8.59e+04    1.17e+04
bathroom_bins_3                        3.947e+05   4.58e+04      8.611      0.000    3.05e+05    4.85e+05
bathroom_bins_4                        6.795e+05    6.9e+04      9.855      0.000    5.44e+05    8.15e+05
bathroom_bins_5                         3.58e+05   1.13e+05      3.181      0.001    1.37e+05    5.79e+05
bedroom_bins_1                         3.476e+05    5.2e+04      6.686      0.000    2.46e+05     4.5e+05
bedroom_bins_2                         5.679e+05   9.91e+04      5.728      0.000    3.74e+05    7.62e+05
bedroom_bins_3                         7.695e+05   1.48e+05      5.210      0.000     4.8e+05    1.06e+06
bedroom_bins_4                         1.122e+06   1.99e+05      5.648      0.000    7.32e+05    1.51e+06
bedroom_bins_5                         1.226e+06   2.63e+05      4.656      0.000     7.1e+05    1.74e+06
bedroom_bins_6                          6.07e+05   3.28e+05      1.848      0.065   -3.67e+04    1.25e+06
yearBuilt_bin_1980s+                   2.202e+05    1.2e+04     18.372      0.000    1.97e+05    2.44e+05
```

Philadelphia Multiple Linear Regression Model:



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  price   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.818
Method:                 Least Squares   F-statistic:                     49.84
Date:                Sun, 22 May 2022   Prob (F-statistic):          3.28e-120
Time:                        21:19:02   Log-Likelihood:                -5690.4
No. Observations:                 404   AIC:                         1.146e+04
Df Residuals:                     366   BIC:                         1.161e+04
Df Model:                          37                                         
Covariance Type:            nonrobust                                         
================================================================================================================
                                                  coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
const                                       -2.001e+06   3.06e+05     -6.549      0.000    -2.6e+06    -1.4e+06
livingAreaValue                               361.7586     46.709      7.745      0.000     269.907     453.610
bathrooms                                    2.898e+05   6.61e+04      4.385      0.000     1.6e+05     4.2e+05
bedrooms                                     1.117e+05   4.08e+04      2.739      0.006    3.15e+04    1.92e+05
monthlyHoaFee                                 438.3500     36.228     12.100      0.000     367.108     509.592
walkScore.walkscore                          8718.0273   2535.431      3.438      0.001    3732.187    1.37e+04
bathroom_bins_1                              6.209e+05   1.54e+05      4.032      0.000    3.18e+05    9.24e+05
bathroom_bins_2                              2.424e+05   9.28e+04      2.614      0.009       6e+04    4.25e+05
bedroom_bins_1                              -1.729e+05   5.77e+04     -2.997      0.003   -2.86e+05   -5.94e+04
bedroom_bins_2                              -2.438e+05   6.29e+04     -3.875      0.000   -3.67e+05    -1.2e+05
bedroom_bins_3                              -3.081e+05   7.64e+04     -4.033      0.000   -4.58e+05   -1.58e+05
bedroom_bins_4                               4.241e+05   9.21e+04      4.602      0.000    2.43e+05    6.05e+05
yearBuilt_bin_1980s+                         3.783e+05   4.61e+04      8.210      0.000    2.88e+05    4.69e+05
resoFacts.subdivisionName_Academy House     -3.834e+04   2.01e+05     -0.190      0.849   -4.35e+05    3.58e+05
resoFacts.subdivisionName_Art Museum         2.843e+05   1.43e+05      1.983      0.048    2373.707    5.66e+05
resoFacts.subdivisionName_Art Museum (fairmo -6.213e+04   1.43e+05     -0.434      0.664   -3.43e+05    2.19e+05
resoFacts.subdivisionName_Art Museum (spring -1.049e+05   2.41e+05     -0.436      0.663   -5.78e+05    3.68e+05
resoFacts.subdivisionName_Art Museum Area   -1.108e+05   1.56e+05     -0.711      0.477   -4.17e+05    1.96e+05
resoFacts.subdivisionName_Avenue Of The Arts  3.059e+04   9.56e+04      0.320      0.749   -1.57e+05    2.18e+05
resoFacts.subdivisionName_Bella Vista        1.957e+05   3.38e+05      0.579      0.563   -4.69e+05    8.61e+05
resoFacts.subdivisionName_Center City        -1.43e+05   9.26e+04     -1.544      0.123   -3.25e+05    3.91e+04
resoFacts.subdivisionName_Chinatown           1.86e+04   1.48e+05      0.126      0.900   -2.72e+05    3.09e+05
resoFacts.subdivisionName_East Falls         4.823e+05   3.54e+05      1.363      0.174   -2.13e+05    1.18e+06
resoFacts.subdivisionName_Fishtown           2.095e+05   1.41e+05      1.486      0.138   -6.77e+04    4.87e+05
resoFacts.subdivisionName_Graduate Hospital -1.425e+04   2.41e+05     -0.059      0.953   -4.87e+05    4.59e+05
resoFacts.subdivisionName_Logan Square      -1.116e+04   9.41e+04     -0.119      0.906   -1.96e+05    1.74e+05
resoFacts.subdivisionName_Northern Liberties -1.249e+05   7.98e+04     -1.565      0.118   -2.82e+05     3.2e+04
resoFacts.subdivisionName_Old City          -7.634e+04   8.73e+04     -0.874      0.382   -2.48e+05    9.53e+04
resoFacts.subdivisionName_Olde City           2.29e+04   1.98e+05      0.115      0.908   -3.67e+05    4.13e+05
resoFacts.subdivisionName_Penns Landing     -4.088e+05   3.36e+05     -1.218      0.224   -1.07e+06    2.51e+05
resoFacts.subdivisionName_Phila (south)      1.731e+05   3.38e+05      0.513      0.609   -4.91e+05    8.37e+05
resoFacts.subdivisionName_Philadelphia      -1.692e-12   1.13e-10     -0.015      0.988   -2.24e-10    2.21e-10
resoFacts.subdivisionName_Point Breeze        2.94e+04    2.4e+05      0.122      0.903   -4.43e+05    5.02e+05
resoFacts.subdivisionName_Queen Village       3.16e+04   1.54e+05      0.205      0.837   -2.71e+05    3.34e+05
resoFacts.subdivisionName_Rittenhouse Square  1.231e+05   6.99e+04      1.762      0.079   -1.43e+04     2.6e+05
resoFacts.subdivisionName_Society Hill      -2.275e+05   1.08e+05     -2.096      0.037   -4.41e+05   -1.41e+04
resoFacts.subdivisionName_Spring Gdn         2.652e+04   3.37e+05      0.079      0.937   -6.37e+05     6.9e+05
resoFacts.subdivisionName_Wash Sq West       2.583e+05   2.01e+05      1.283      0.200   -1.38e+05    6.54e+05
resoFacts.subdivisionName_Washington Sq     -1.541e+05   1.25e+05     -1.230      0.219      -4e+05    9.22e+04
resoFacts.subdivisionName_Washington Sq West -1.483e+05   1.22e+05     -1.215      0.225   -3.88e+05    9.16e+04
```