Modeling Increasing Housing Costs in Brooklyn

Alexander Cole and Callum Roberts

April 24, 2024

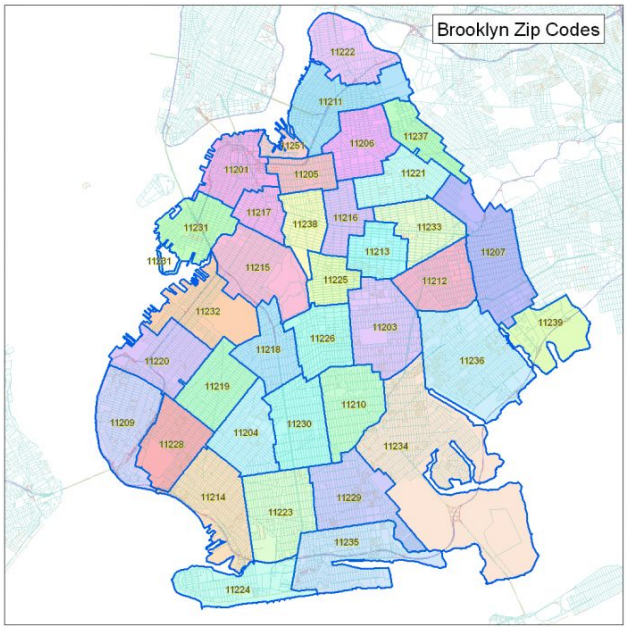


Figure 1: Brooklyn Separated by Zip Codes

1 Introduction

1.1 Project Motivation and Outline

For this project, we sought to select a problem within our current society's climate that interested us and attempted to model it as a means to more deeply understand it and potentially predict future outcomes. The phenomenon in the real world that we found particularly intriguing was housing prices for their constant, wild, fluctuation. More specifically, we wanted to choose a location known for its rapidly developing market. With all of this in consideration, we decided to model the change in housing prices in the borough of Brooklyn, New York. Brooklyn is commonly thought of as one of the most gentrified parts of New York City, and by extension, the country. Often in the news for its seemingly momentous growth in living costs, long-time Brooklyn residents are finding it to be increasingly difficult to afford

to remain in their communities; especially as more and more, wealthier, newcomers chose to move to the largest, most iconic city in the United States.

We will begin this paper by discussing each of the variables that were chosen for this model, as well as the type of model we used and why. Next, we will discuss the data collection techniques, and display the model and its output. We will end the paper by discussing the results and how these findings can be applied in the future, as well as discussing our methods for modeling and displaying our data.

1.2 Variables Chosen

In order to construct this model, it was important to decide which variables could be used to measure growing costs, and how we could measure what is behind said costs. Of course in a real-world scenario, there are countless factors that come into the housing prices of a neighborhood. Even though we would have loved to incorporate every element, it would be impractical and impossible to construct a mathematical model with all of these variables. Because of this, we decided to isolate our model to the following variables:

- **Population** - The first attribute we thought would be important for predicting the living costs of a zip code is its population. Naturally, if a given zip code has a large or increasing population, one would assume that there is a higher demand to live in that location, thus driving up housing prices. Currently, this is a topic that doesn't have a formal answer, but there are many studies such as [1] that are in support of this claim.

- **Median Income** - The next attribute we looked at was median income. If the average income in a zip code is higher, we would expect individuals to be able to pay much more for their homes. Furthermore, one would also assume that a neighbourhood with a growing median income would attract new, higher income residents. This has been studied thoroughly, with there being much research in support of an increase in living costs with an increase in the income of a population. The two variables have a positive correlation but are not as indicative as one might think [2].

- **Number of Arts and Entertainment Venues** - This attribute was chosen as indicator of how attractive an area may be to new residents. More venues for art and entertainment will lead to an increase in tourists, alongside an increase in an area's attractions for newcomers. A study was recently done on tourism's effect on housing affordability in a few cities in Croatia that experience high levels of tourism. It was found that with an increasing amount of tourism, housing becomes more and more unaffordable as prices rise [3]. We expect that the number of arts and entertainment venues will be able to model tourism and and predict increasing costs within an area.

- **Percent of Young Adults Enrolled in School** - For this last attribute we assume that a higher rate of young adults graduating from college will increase income and subsequently living costs. Studies have found that more individuals with college degrees stimulate the local economy. It is claimed that the average college degree holder will contribute $278,000 towards their local economy over their lifetime compared to a high

school graduate counterpart [4]. From this, we assume that a more stimulated economy will boost housing costs, much as arts and entertainment venues, and median incomes will.

1.3 Data Collection

Once we selected our project and the variables that we wanted to use, we first had to select the zip codes within Brooklyn that we wanted to target. After searching through all the zip codes we decided to target five of the most populated areas. Of these five, we wanted to select a few that had some of the highest home prices and some that had the lowest. For example, 11206 is the zip code for Williamsburg, one of the more famously gentrified areas of Brooklyn, known for its large increase in housing costs in recent years. In contrast, 11212 is the zip code for Brownsville, one of the poorest areas of the borough. We wanted to see how these zip codes had gained their price increases and if the variables we chose above had the relationships that we initially assumed them to have. The full list of zip codes we decided on are **11206, 11212, 11219, 11230, and 11235**.

Next we had to find the data, and for all of which we were able to use the U.S. Government's Census Bureau to find them[5]. We extracted our specific variables from public data tables for each zip code from 2012 to 2022 (fig. 2).

zip 11206	population	median income	average monthly housing cost	arts/entertainment	percent of 20-24 year old enrolled in school
2012	80086	28584	875	14	30.4
2013	81525	28559	908	12	31
2014	83180	30686	962	18	31
2015	84806	30779	976	22	29.9
2016	86486	31549	1002	26	32
2017	87767	34122	1039	31	32
2018	88349	36404	1117	36	33.4
2019	88422	39753	1187	42	33.8
2020	87599	43065	1257	43	32.7
2021	90903	49013	1353	37	31.4
2022	89949	51507	1497		29.1

(a) 11206 data

zip 11212	population	median income	average monthly housing cost	arts/entertainment	percent of 20-24 year old enrolled in school
2012	81267	27901	948	6	30.8
2013	84520	28348	990	6	35.2
2014	87751	28146	991	8	34.7
2015	88668	28207	992	7	33.6
2016	86469	28495	1023	5	33.3
2017	82831	25677	1020	5	39.2
2018	76527	26239	1029	6	40.3
2019	75605	26521	1032	5	39
2020	74037	29385	1046	6	39
2021	78296	30733	1077	8	41
2022	84006	35840	1196		35

(b) 11212 data

zip 11230	population	median income	average monthly housing cost	arts/entertainment	percent of 20-24 year old enrolled in school
2012	84707	42568	1182	6	48.5
2013	84219	42170	1204	8	48.5
2014	87064	41820	1257	7	49
2015	88589	41068	1276	7	47.8
2016	88933	43344	1311	7	49.1
2017	90257	46013	1348	10	48.4
2018	89075	49541	1407	16	49.7
2019	86139	53070	1446	18	53.4
2020	87188	57770	1484	19	53
2021	91789	61017	1588	15	55
2022	90245	65051	1697		58.1

(c) 11230 data

zip 11219	population	median income	average monthly housing cost	arts/entertainment	percent of 20-24 year old enrolled in school
2012	95069	34590	1195	9	43.3
2013	96971	34316	1223	9	42.4
2014	98719	35083	1256	6	38.4
2015	97670	35974	1280	7	38.8
2016	96287	36573	1315	10	37.7
2017	93979	37665	1357	12	38.9
2018	90036	39295	1420	15	40.1
2019	89371	40683	1464	16	41.8
2020	87812	41907	1475	19	41.6
2021	93119	44450	1498	19	45.2
2022	92283	51194	1662		42.3

(d) 11219 data

zip 11235	population	median income	average monthly housing cost	arts/entertainment	percent of 20-24 year old enrolled in school
2012	72447	42298	1118	20	57.1
2013	74630	41639	1151	19	56.8
2014	75622	42257	1161	22	56.6
2015	76668	42818	1162	26	58.1
2016	78237	45578	1195	26	54.9
2017	80222	49653	1241	26	53.7
2018	78128	52538	1297	26	48.5
2019	78775	54646	1338	29	47.4
2020	76921	56308	1375	27	40.5
2021	84859	58669	1417	25	38.6
2022	83069	61320	1561		39.4

(e) 11235 data

Figure 2: Data for Each Zip Code

2 Modeling

2.1 Type of Model

The model we decided was appropriate for our project was the linear regression model. We felt that this model was appropriate because the price of housing tends to rise somewhat linearly, as does inflation and many of the variables discussed above. Moreover, we expect these variables to have a linear impact on housing prices. Before we use our data points to predict the cost of housing, we normalized all of the variables to have values between 0 and 1. This is due to the fact that without this normalization, the variables with large values (population and median income) would appear to have very small coefficients, although they provide significance to the model.

2.2 Target Equation and Variable Syntax

Our linear regression target model is described below

$$H(t) = a * P(t) + b * I(t) + c * E(t) + d * S(t) \tag{1}$$

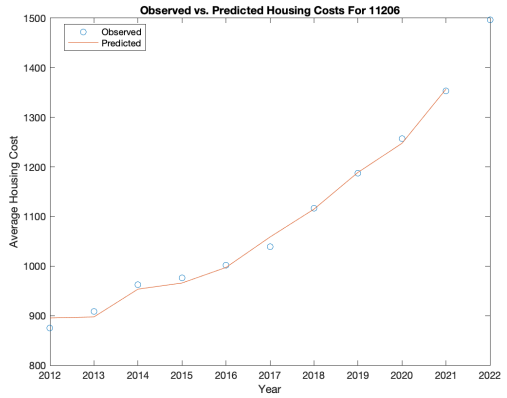
This equation describes the housing cost H(t), using variables P(t), I(t), E(t), and S(t) which represent the normalized population, median income, number of arts and entertainment venues, and percent of young adults in school for the zip code respectively. Moreover, the constants a, b, c, and d represent the different weights for each variable.

3 Results

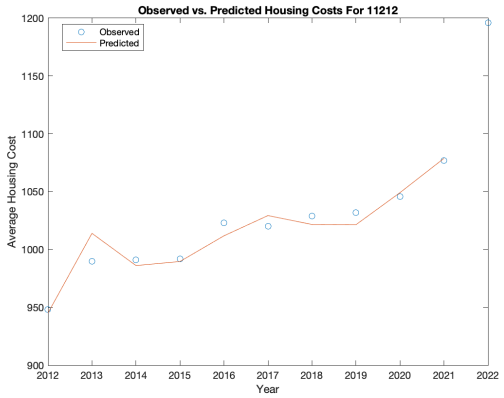
3.1 Qualitative Results

Below we have five graphs, one for each zip code, showing the linear regression model over the historical housing costs. As expected, housing prices became more expensive throughout all five zip codes, but to varying degrees. For instance, the graph for zip 11212 represents Brownsville, NY, one of the city's poorest areas. Here, housing costs between 2012 and 2022 increased much less than the other areas, which was predicted rather successfully by the model. We also see that the rate of increase between years is rather small, implying a steadier increase in prices in this area of the city. In contrast, Williamsburg, one of the most famously gentrified parts of the city saw the largest increase between 2012 and 2022. Notably, the rate of this increase is shown on the model to have increased between each year, implying a much more drastic growth of living costs in the community, which was expected.

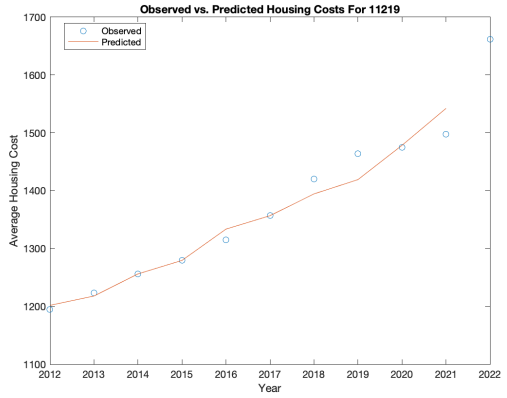
For the most part, prediction of each zip code's growth in costs was most strongly based on the change in median income for the area. This is particularly evident in the change in costs for zip 11235, which contains the Sheepshead Bay neighbourhood. Although not famed for gentrification, this neighbourhood saw an increase in prices on par with Williamsburg, which aligned with an almost 50% growth of median income.



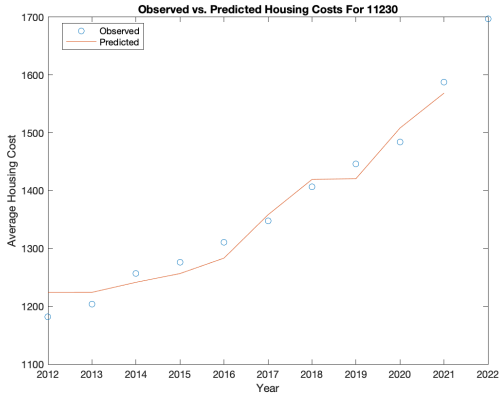
(a) 11206 Model



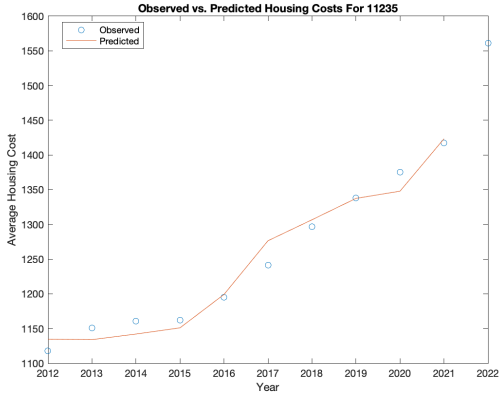
(b) 11212 Model



(a) 11219 Model



(b) 11230 Model



(a) 11235 Model

Figure 3: Linear Regression Models Compared to Actual Data for Each Zip Code

3.2 Quantitative Results

From a quantitative perspective, below are the linear regression equations for each of the zip codes. These are the same equations that are graphed above among the actual cost of living prices.

- **11206**

$$H(t) = 314.67 * P(t) + 961.41 * I(t) + 76.187 * E(t) + 66.447 * S(t) \tag{2}$$

- **11212**

$$H(t) = 230.87 * P(t) + 566.75 * I(t) - 108.07 * E(t) + 497.03 * S(t) \tag{3}$$

- **11219**

$$H(t) = 698.7 * P(t) + 1308.5 * I(t) + 232.93 * E(t) - 485.91 * S(t) \tag{4}$$

- **11230**

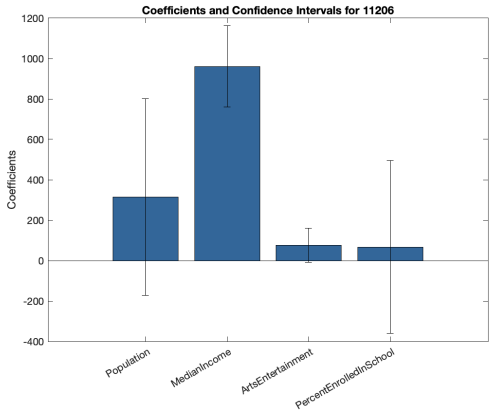
$$H(t) = 1021.2 * P(t) + 928.93 * I(t) + 106.26 * E(t) - 430.8 * S(t) \tag{5}$$

- **11235**

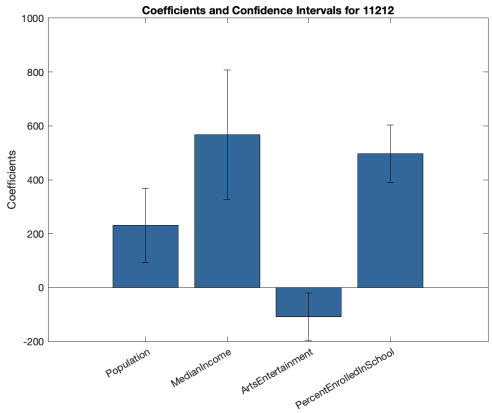
$$H(t) = 362.25 * P(t) + 1076.7 * I(t) - 63.234 * E(t) + 128.59 * S(t) \tag{6}$$

From these equations, we can analyze the coefficients to see which ones impact the model significantly and whether they have a positive or negative correlation. Shown below are bar charts showing the impact of each coefficient.

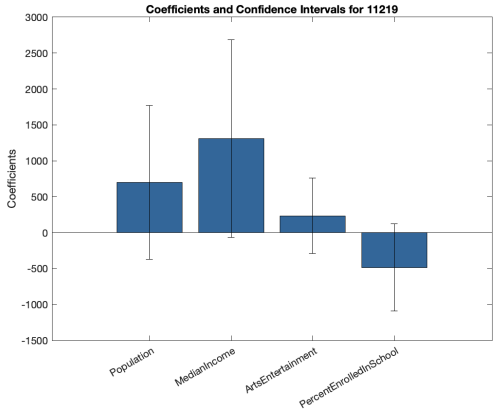
We can see for each zip code, that median income does have a very large correlation with housing price, as does population. So, we can expect that in other zip codes an increasing population that is getting increasingly wealthy will almost definitely drive up rent costs in the area. This makes sense with our initial assumptions for the model. However, the number of arts and entertainment venues and percent of adults enrolled in school is shown to be very inconsistently correlated with the housing costs of an area. In some zip codes, these two variables behaved exactly as we had initially predicted; more art galleries and more college students seemed to be tied to raising prices. Yet, in other neighborhoods, the inverse seemed to be true. This could be for a multitude of reasons but it is most likely that in reality, these variables have very little effect on rent prices of an area.



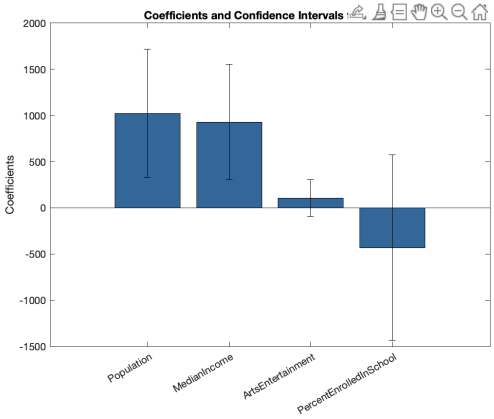
(a) 11206 Coefficients



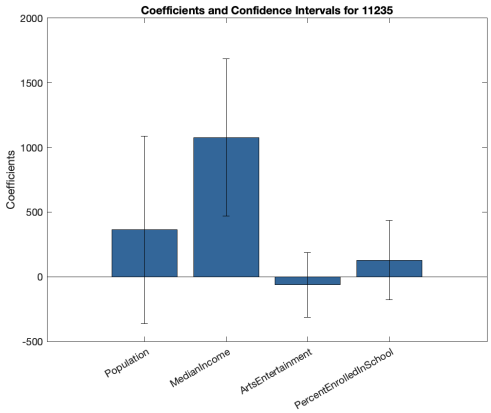
(b) 11212 Coefficients



(a) 11219 Coefficients



(b) 11230 Coefficients



(a) 11235 Coefficients

Figure 4: Variable Coefficient Significance for Each Zip Code

4 Discussion

4.1 Discussion of Results

The results of the model successfully highlight some of the factors that go into the pricing of homes in an area. Notably, median income and population seem to directly lead to a larger increase in costs, so one would expect similar outcomes to occur city-wide and even nationally. However, some of the results proved rather inconclusive. In regards to arts venues and university students, the model demonstrated very little effects. We had initially hoped that these variables could predict increasing costs, or at least showcase some effects of an increasingly affluent community. This did not prove to be the case. It is important to note that the COVID-19 pandemic could of had an effect on this data, with businesses shutting down and neighborhood residents withdrawing from higher education as a consequence of social distancing and online education. Furthermore, the relationships between arts spaces, universities, and their communities can be very different depending on the area and its economy.

4.2 Discussion of Model

Perhaps the biggest takeaway from our prediction model is that housing prices are inherently unpredictable; especially with a small number of variables. Each neighborhood can react to housing costs and trends incredibly differently. It would have made sense for us to include more variables in the model, such as employment statistics, local industry, and crime rate, as these are all factors that are expected to contribute to an area's desirability. It would also have been interesting to compare these variables to the national (or city-wide) average to determine how much these changes are indicative of the behavior of a particular neighborhood, as opposed to the behavior of the country as a whole.

Although very difficult to quantify, if we could track the migration patterns in and out of these areas of Brooklyn, the model could consider the interactions between long-time community residents and newcomers, and its effect on price. It is possible that in a city like New York, newcomers may be attracted to a group of particular areas, which may drive up costs, and price long-time residents out of the neighborhood, similar to a Lotka–Volterra system of equations. It would however be incredibly challenging to find such data.

5 Future Work

Our project attempted to model the cost of living in different neighborhoods of Brooklyn, New York, through the separation of zip codes. If given more resources and time, one could expand the chain of variables to account for more factors that could potentially influence the cost of living. This could help get a more holistic view of the problem at hand. Moreover, one could add more years to the data to track a longer period and have a more refined model.

Additionally, by targeting a wider range of zip codes and performing cross-analysis, one could figure out what variables are common amongst zip codes and what is specific to the zip codes'

environment. In our results, we saw that some variables' importance varied significantly, and with cross-examination one might be able to tell whether this is due to a local factor or a flaw in the model.

6 Computer Data

To generate our figures and model results we used Matlab. Below is the code that we developed to fit the linear regression model, output the results of the function, graph the results, and output the predictions the function made for each year 2012-2022. The code below is for zip code 11206 specifically , but we repeated this process with data from the other zip codes as well which have been omitted due to repetition.

```
% Data from census bureau
years = [2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020,
   2021, 2022];
population = [80086, 81525, 83180, 84806, 86486, 87767, 88349,
   88422, 87599, 90903, 89949];
median_income = [28584, 28559, 30686, 30779, 31549, 34122,
   36404, 39753, 43065, 49013, 51507];
average_housing_cost = [875, 908, 962, 976, 1002, 1039, 1117,
   1187, 1257, 1353, 1497];
percent_enrolled_in_school = [30.4, 31, 31, 29.9, 32, 32, 33.4,
    33.8, 32.7, 31.4, 29.1];
arts_entertainment = [14, 12, 18, 22, 26, 31, 36, 42, 43, 37,
   NaN];

% Normalize variables between 0 and 1
population_normalized = population / max(population);
median_income_normalized = median_income / max(median_income);
arts_entertainment_normalized = arts_entertainment / max(
   arts_entertainment(~isnan(arts_entertainment))); % Handle
   NaN for max calculation
percent_enrolled_in_school_normalized =
   percent_enrolled_in_school / max(percent_enrolled_in_school)
   ;

% Create a table with normalized data
data = table(years', population_normalized',
   median_income_normalized', arts_entertainment_normalized',
   percent_enrolled_in_school_normalized', average_housing_cost
   ',...
    'VariableNames', {'Year', 'Population', 'MedianIncome', '
       ArtsEntertainment', 'PercentEnrolledInSchool', '
       AverageHousingCost'});
```

```
% Run fitlm on data, predicting average housing cost using the
   other variables
mdl = fitlm(data, 'AverageHousingCost ~ -1 + Population +
   MedianIncome + ArtsEntertainment + PercentEnrolledInSchool')
   ;

% Output function
disp(mdl);

% Analysis of our Model
% Confidence Interval on Coefficients
disp('Confidence Intervals:');
disp(table(mdl.coefCI, mdl.CoefficientNames', 'VariableNames',
   {'LowerCI', 'UpperCI'}));

% Plot this
figure;
% Coefficients
bar(mdl.Coefficients.Estimate, 'FaceColor', [0.2, 0.4, 0.6]);
hold on;
% Extract confidence intervals
ci = coefCI(mdl);
lower_errors = mdl.Coefficients.Estimate - ci(:,1);
upper_errors = ci(:,2) - mdl.Coefficients.Estimate;
errorbar(1:numel(mdl.Coefficients.Estimate), mdl.Coefficients.
   Estimate, lower_errors, upper_errors, 'k', 'linestyle', '
   none');
xticks(1:numel(mdl.Coefficients.Estimate));
xticklabels(mdl.CoefficientNames);
ylabel('Coefficients');
title('Coefficients and Confidence Intervals for 11206');
hold off;
saveas(gcf, '11206coeff.png');

% Get predicted costs from function using data
predicted_costs = predict(mdl, data);

% Plot predicted against actual cost
figure;
plot(data.Year, data.AverageHousingCost, 'o', 'DisplayName', '
   Observed');
hold on;
plot(data.Year, predicted_costs, 'DisplayName', 'Predicted');
legend('Location', 'Best');
xlabel('Year');
```

```
ylabel('Average Housing Cost');
title('Observed vs. Predicted Housing Costs For 11206');
hold off;

saveas(gcf, '11206line.png');

% Display the predicted housing cost for each year 2012-2022
disp(table(data.Year, predicted_costs, 'VariableNames', {'Year'
   , 'Predicted Housing Cost'}));

% Graph again but scatter plot
figure;
scatter(data.Year, data.AverageHousingCost, 'o', 'DisplayName',
    'Observed');
hold on;
scatter(data.Year, predicted_costs, 'x', 'DisplayName', '
   Predicted');
legend('Location', 'Best');
xlabel('Year');
ylabel('Average Housing Cost');
title('Observed vs. Predicted Housing Costs For 11206');
hold off;

saveas(gcf, '11206plot.png');
```

References

[1] Clara H. Mulder. *The relationship between population and housing.* University of Amsterdam. Department of Geography, Planning and International Development Studies. https://unece.org/fileadmin/DAM/hlm/archive/Key%20note%20population%20and%20housing.pdf.

[2] Aamir Surani and John Michael Young. *Relationship between Income and Cost of Living in US Cities.* Georgia Institute of Technology, November 18, 2022. https://repository.gatech.edu/server/api/core/bitstreams/f731e89b-c79f-4ea3-809c-712cd65aec52/content.

[3] Josip Mikulić, Maruška Vizek, Nebojša Stojčić, James E. Payne, Anita Čeh Časni, Tajana Barbić. *The effect of tourism activity on housing affordability.* Published on ScienceDirect, September 2021, Article ID: 103264. https://www.sciencedirect.com/science/article/pii/S0160738321001420.

[4] Jonathan Rothwell. *What colleges do for local economies: A direct measure based on consumption.* Published on Brook-

ings, November 17, 2015. https://www.brookings.edu/articles/what-colleges-do-for-local-economies-a-direct-measure-based-on-consumption/.

[5] U.S. Census Bureau. *Data.census.gov*. Accessed on Feb 12, 2024. https://data.census.gov/.