MathWorks Math Modeling Challenge 2024

Phillips Academy Andover

Team #17645, Andover, Massachusetts

Coach: Heidi Wall

Students: Tianyi Evans Gu, Yifan Kang, Eric Wang, Anthony Yang, Angeline Zhao

M3 Challenge CHAMPION—$20,000 Team Prize

JUDGE COMMENTS

Specifically for Team #17645—Submitted at the close of triage judging

COMMENT 1: You presented a well-written executive summary providing some results and rationale. The assumptions and justifications are clearly described. Your choice of models was creative. In each part, the analysis is clearly presented and justified. Well done!

COMMENT 2: The executive summary can be improved by including the problem, and a clearer explanation of the modeling. Assumptions are very clear. The use of a single model (logistic model and multivariate linear regression) needs to be better justified—i.e., why not other models? Overall, the report is well written and the organization was clear.

COMMENT 3: The section of the report for Q3 would be improved by including more specific suggestions to the city for mitigating the conditions of the unhoused population.

COMMENT 4: You answered all questions with adequate details. Well done! Also, good job providing citations for your references.

MathWorks Math Modeling Challenge
A program of siam

***Note: This cover sheet has been added by SIAM to identify the winning team after judging was completed. Any identifying information other than team # on a MathWorks Math Modeling Challenge submission is a rules violation. Further, this paper is posted exactly as submitted to M3 Challenge. Typos, odd formatting, or other mistakes may be attributed to the 14-hour time constraint.

M3 Challenge 2024:
A Tale of Two Crises: *The Housing Shortage and Homelessness*

Team #17645

March 2nd, 2024

Executive Summary

To the Secretary of the U.S. Department of Housing and Urban Development,

In recent years, the housing crisis in major cities around the U.S. has become more significant. As incomes increase slower than housing prices, the number of affordable housing options has decreased significantly[1], leaving many individuals homeless due to an inability to pay rent or purchase homes. However, combating the issue of homelessness is difficult given the many limitations over potential government policies that could have a feasible impact[2]. As such, the housing crisis and homelessness, as well as related factors, are critical to understand in order to resolve the problem that leaves so many Americans out of a home every night.

We first predicted the number of housing units that will be in Seattle, Washington and Albuquerque, New Mexico in the next 10, 20, and 50 years using a logistic model. As the number of housing units that can be feasibly constructed in a given land area is limited, we used collected data on developable land in the two cities to determine a carrying capacity for the logistic models[3, 4]. Our model estimates 411,600, 453,000, and 513,000 housing units in Seattle and 261,000, 270,900, and 290,800 housing units in Albuquerque by 2030, 2040, and 2070, respectively.

We then determined the correlation of several factors with homelessness utilizing a multivariate regression model to determine significant factors that could be used to model homelessness. We found that the number of housing units and median listing price of housing units were highly correlated with homelessness, which also accounted for other similar factors such as total population shifts and median household income. We then used our previous logistic model for housing units and a linear model for median listing price together in a multiple regression model to estimate homelessness in Seattle and Albuquerque in the next 10, 20, and 50 years. Our model predicted that 14,400, 17,100, and 23,200 individuals in Seattle and 3,800, 5,100, and 10,400 individuals in Albuquerque would be homeless by 2030, 2040, and 2070, respectively.

Lastly, we selected four factors: housing units, median listing price, deaths by opioid use, and the unemployment rate to test causality with homelessness in Seattle through a Granger causality test. We calculated the F-statistics and p-values with lag values of 1 and 2 years and determined statistical significance with $\alpha = 0.05$. We found that housing units, median listing price, median household income, deaths by opioid use were significant with 1 year lag while unemployment was not significant. To then calculate the most impactful factors to target in policymaking, we adjusted each of the significant factors by 10% to see the impact they would have in reducing homelessness. We found that increasing the number of housing units and reducing deaths by opioid use had the most substantial effect on reducing homelessness by 2284 and 360 individuals, respectively.

We believe these results will assist policymakers in combating the housing crisis and homelessness by providing specific factors to target through new policies to aid the many unhoused individuals across the U.S. and provide them with homes.

Table of Contents

Q1: It Was the Best of Times

1.1 Defining the Problem

The first problem asks us to develop a model to predict the growth of the housing supply in two cities for the next 10, 20, and 50 years. We have selected Seattle, Washington, and Albuquerque, New Mexico as our cities. Our model will take into account previous housing unit data from the U.S. in the last decade.

1.2 Assumptions

1.2-1. **Current construction firms will not undergo any drastic changes in the near future.**

- **Justification:** It is difficult to predict the business decisions of any construction firms or other manufacturers. As such, we will exclude these changes from our model.

1.2-2. **Covid-19 has a negligible impact on the housing market after 2024 and did not significantly impact historical data.**

- **Justification:** Any short-term impacts of Covid-19 on the housing market have ended by 2024[5]. Our model will not consider any continued influence of Covid-19 to create accurate predictions for future decades.

1.2-3. **There will be no major policy changes regarding the housing market in Seattle or Albuquerque in the near future.**

- **Justification:** It is difficult to predict the policy changes that may be enacted in Seattle and Albuquerque, so we will exclude these changes from our model, assuming that any policy changes will not drastically change the way the housing market operates in the near future.

1.2-4. **Any detached or attached housing units, as well as any mobile homes, boats, RVs, vans, etc., used primarily for housing are considered housing units in our model.**

- **Justification:** Data collected on housing units in the 2020 Census includes mobile homes, boats, RVs, vans, etc[6].

1.2-5. **The number of housing units that can be built in Seattle or Albuquerque is finite.**

- **Justification:** There are strict limits on the number of housing units built in a city based on zoning laws and land area[7].

1.2-6 **The growth of the housing market follows a logistic curve.**

- **Justification:** Based on 1.2-5, there is a limit on the number of housing units, so we assume this is the capacity in a logistic model[8].

1.2-7 **The growth of the housing market is independent of which kinds of housing units are built.**

- **Justification:** We assume that houses are built according to market demands and focus only on the overall growth of the housing market.

1.2-8 **The growth of the housing market from 2010 to 2021 in the U.S. is sufficient to determine growth in the future.**

- **Justification:** The data provided by the Mathworks Math Modelling Challenge is only from 2010 to 2021, so we assume for simplicity that this data is sufficient.

1.2-9 **Rising sea levels caused by global warming will not impact the production of housing units in Seattle.**

- **Justification:** Seattle Public Utilities predicts that sea levels will rise by approximately one foot by 2050[9], but this pushes most coastlines inland by about 100 feet, which for Seattle will not significantly encroach on current or future housing units[10].

1.3 The Model

1.3.1 Model Development

We chose a logistic model for determining the growth of housing units Logistic models measure the growth of a variable that starts exponential and then approaches an upper limit. The production of housing units already experienced exponential growth during the 1990s[11], so our model mainly exhibits the latter logarithmic growth. As stated in 1.2-5, the number of housing units in a given city is finite, and we can calculate a limit on the number of housing units based on information regarding land use. Logistic models are also more accurate for long-term forecasting because the logistic model includes a carrying capacity as a limit[12].

Our team considered utilizing other models, particularly linear regression. Using the provided data points from 2010 to 2019 in Seattle, a linear model estimated that there would be 680,568 housing units by 2070. The prediction exceeds the estimated maximum number of housing units that can feasibly be constructed in Seattle, as described later in our model execution. Thus, a linear model could not accurately estimate the number of housing units because it completely disregarded any limits on the land area available. On the other hand, the logistic model does consider the carrying capacity of a given variable, making it far more accurate in the long term when the variable may be near the carrying capacity.

1.3.2 Model Execution

We used the provided data on housing units from 2010 to 2023 in Seattle, Washington, and Albuquerque, New Mexico to conduct our logistic model.

As the logistic model requires a carrying capacity that bounds the growth of the given variable, we used information on developed and undeveloped land areas to determine the limit on housing units in each city.

Calculations from the Seattle Development Capacity Report considered data on the amount of available land in Seattle based on vacant and undeveloped parcels to estimate future development on those land areas[13]. The final estimate on housing unit capacity in Seattle was 531,770 housing units, which we then used as our carrying capacity for the Seattle logistic model.

For Albuquerque, data from the City of Albuquerque Comprehensive Plan in 2017 shows the following data regarding developed and vacant land areas[14].

Developed Land (sq. mi.)	Vacant Land (sq. mi.)	Total Developable Land (sq. mi.)	Total Undevelopable Land (sq. mi.)
89	24	113	76

Table 1: Developed and Vacant Land Areas by Square Miles

Given that there were 243,402 housing units in Albuquerque in 2017, we calculated the maximum capacity for housing units in the city based on the remaining amount of vacant land. We assumed that any potential development of housing units in the currently vacant land would be similar to the level of development in the currently developed land. As such, we scaled the current level of housing development to include the currently vacant land. Given that there were 89 square miles of developed land and 243,402 housing units, if the entire 113 square miles of developable land in Albuquerque were filled, there could be a maximum of $243,402\times\frac{113}{89}\approx 309,038$ housing units, which we then used as our carrying capacity for the Albuquerque logistic model.

Utilizing the data on Seattle and Albuquerque, we used Python's scipy library and curvefit function to create logistic models to determine the projected number of housing units in the two cities. The logistic formula and variables are listed below:

$$f(x) = \frac{L}{1+e^{-k(x-x_0)}}$$

Symbol	Variable	Unit	Values for Seattle	Values for Albuquerque
L	Carrying Capacity	Housing Units	531,770	$309,038$
k	Relative Growth Rate Coefficient	N/A	0.04429	0.02686
x	Year	Year	N/A	N/A
x_0	Inflection Point	Year	2005	2019

Table 2: Variables in Logistic Function

1.4 Results

Using our logistic models developed in the previous section, we estimated the number of housing units in Seattle and Albuquerque for 10, 20, and 50 years into the future in Seattle and

Albuquerque in Python. Below are the graphs of our logistic model predictions for 2030, 2040, and 2070, as well as a table consisting of the estimated number of housing units.

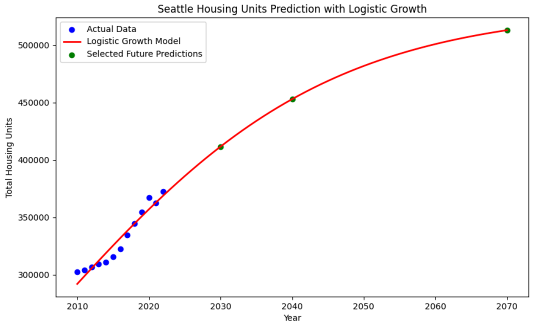


Figure 3: Logistic Model For Seattle, Washington

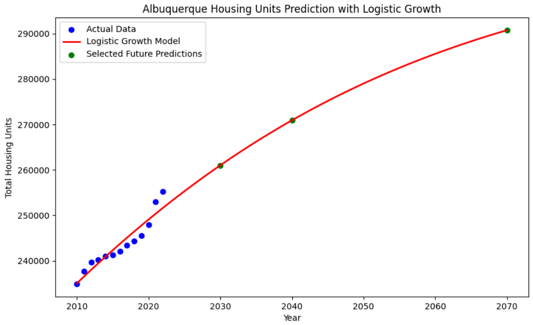


Figure 4: Logistic Model for Albuquerque, New Mexico

	2030	**2040**	**2070**
Seattle	411,600	453,000	513,000
Albuquerque	261,000	270,900	290,800

Table 5: Estimated Number of Housing Units (to the nearest hundred)

1.5 Discussion

For Seattle, our model estimates that there will be 411,600, 453,000, and 513,000 housing units in 2030, 2040, and 2070, respectively. For Albuquerque, our model estimates that there will be 261,000, 270,900, and 290,800 housing units in 2030, 2040, and 2070, respectively. From this, we can conclude that the housing market will continue to grow in Seattle and Albuquerque over the next fifty years, albeit at a decreased rate over time as it nears the carrying capacity.

1.6 Sensitivity Analysis

To determine the accuracy of our predictions, we randomly offset each data point by up to 5% and then ran the model again to get a new prediction. Then we calculated the percent difference between the original prediction and the new prediction. We repeated this process five times and averaged the percent differences. This process determined that our prediction of housing units had the following average jittered variations:

	2030	**2040**	**2070**
Seattle	2.117%	1.895%	0.750%
Albuquerque	2.380%	3.736%	5.138%

Table 6: Average Jittered Variations of Housing Units

The average jittered variations are all relatively low, so we are relatively confident in our model's resilience to random error.

1.7 Strengths & Weaknesses

The use of a logistic model is better for long-term predictions compared to other models[15], which is particularly important for our problem considering that we are making predictions up to fifty years in the future. Moreover, the logistic model uses a carrying capacity that takes into account the capacity limit of the number of housing units that can be built in a given city, based on the available land area. This is particularly important to consider for the housing market because we cannot expect the market to increase without limits given the physical and legal limitations on the number of housing units that can be constructed in a given area.

In the same vein, such a carrying capacity can be hard to estimate accurately, given that there exists minimal data on predictions of housing capacity in the two cities we are considering. Because the logistic model is based largely on the carrying capacity we calculated, our model could consequently be less accurate. Moreover, we assumed in 1.2-3 that there would be no major changes to zoning laws in either of the two cities in the near future. However, given the severity of the housing crises in recent years, there could be significant changes to legislation regarding zoning laws, particularly the allowed maximum capacity in land parcels, which would change our carrying capacity. Moreover, our exclusion of the data from Covid-19 in 2020 through 2022 eliminates the potential impact of Covid-19 on the housing market, although we expect that any such impact would be minimized for such long-term estimates.

Q2: It Was the Worst of Times

2.1 Defining the Problem

The second problem asks us to estimate the homeless population in Seattle and Albuquerque in the next 10, 20, and 50 years. We used a number of factors that were strongly correlated with shifts in the homeless population in order to generate an accurate estimate.

2.2 Assumptions

2.2-1. **There will be no natural disasters or other anomalies that cause significant unpredictable changes to the homeless population in Seattle and Albuquerque.**

- **Justification:** It is difficult to account for any unpredictable future events that may cause drastic changes to the number of homeless individuals in Seattle and Albuquerque. We will assume that any external factors remain the same for the near future.

2.2-2. **Examined historical trends over the last decade will remain similar in the near future.**

- **Justification:** It is difficult to account for unexpected changes to historical trends. Moreover, our data spans the last decade, so we can assume that such historical trends will remain the same for the near future.

2.2-3. **There are no other independent factors besides total housing units and median listing price that significantly impact the number of homeless people.**

- **Justification:** We cannot comprehensively include all potential parameters within the time allotted. Moreover, as discussed in 2.3.1 below, most other factors are strongly correlated with the two variables above, so we disregard them.

2.2-4 **The number of total housing units over time reflects the total population growth.**

- **Justification:** The number of housing units and population growth are each logistic over time[16]. Moreover, the number of occupied housing units remains linear with respect to total housing units, so we can then interpret that the growth of the total population is reflected by the growth of housing units. Detailed justification is provided in 2.3.1 below. Thus, we do not need to independently consider population growth.

2.2-6 **The median listing price of housing units reflects the median household income due to the way the housing market operates.**

- **Justification:** The median listing price of housing units must grow at a similar rate to the median household income, as the housing market will adjust to match the demand for housing units[17]. Detailed justification is provided in 2.3.1 below. Thus, we do not need to independently consider median household income.

2.2-7. **Staggering data values by year is negligible.**

- **Justification:** Although causal effects typically take some lag time to manifest, all variables most likely still reflect the change in homelessness within the same year the variable changes. Since our data is measured each year, the time lag will be trivial.

2.2-8 **The data for variables considered is homoscedastic.**

- **Justification:** We assume the error variance is constant or equal across the levels of independent variables. That is, our independent variables are good predictors of homelessness.

2.2-9 **There is a linear relationship between homelessness and our independent variables, and the residuals of the independent variables are normally distributed.**

- **Justification:** The multivariate regression we will use requires this assumption, and we can verify this assumption from the data we're given.

2.2-10 **Missing data can be extrapolated from the given data for Question 2.**

- **Justification:** Some of the data given for Question 2 of the Mathwork Math Modeling Challenge is missing, so we use imputing to fill in the missing values.

2.2-11 **Covid-19 was a significant anomaly in terms of homelessness in each city, so we do not consider the data from 2020-2022, instead assuming that homelessness returned to normal by 2023.**

- **Justification:** During 2020-2022, the number of homeless people for each city shifted significantly and data collection capabilities were diminished, leading to less accurate values[18]. As such, we assume that the number of homeless people from 2023 onwards will adopt the same trends as before Covid-19.

2.3 The Model

2.3.1 Model Development

We first examined various factors to determine correlations to the number of homeless individuals in each of the two cities through a multivariate linear regression. Our data for Question 2 spanned from 2010 to 2023, with some missing data values, so we used imputing to fill in the missing values. We used Python to import all the data and subset the portion we are analyzing.

We then chose multivariate linear regression as a model to determine the correlation for the various factors in each city. We used the following equation, where E is the total number of homeless people, b is an error term, a_i are coefficients of independent variables, and V_i are the values of the independent variables:

$$E = b + \sum_{i=1}^{n} a_i V_i$$

Symbol	**Variable**	**Unit**
E	Number of homeless people	People
b	Error term	N/A
a_i	Coefficients of independent variables	N/A
V_i	Values of independent variables	N/A

Table 7: Variables in a Multivariate Linear Regression Model

We then produced the following heat maps that indicate factors with high correlation using Python's *seaborn* library[19].

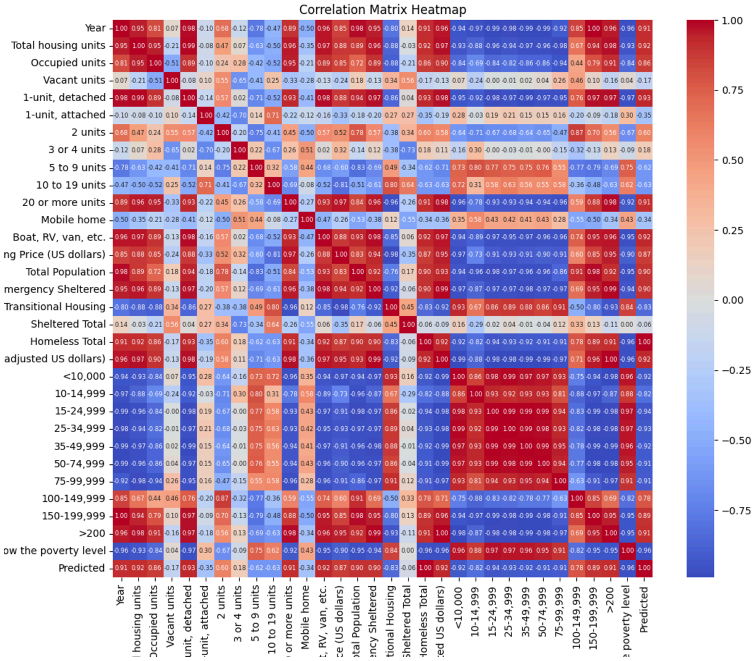


Figure 8: Variable Correlation Heat Map for Seattle

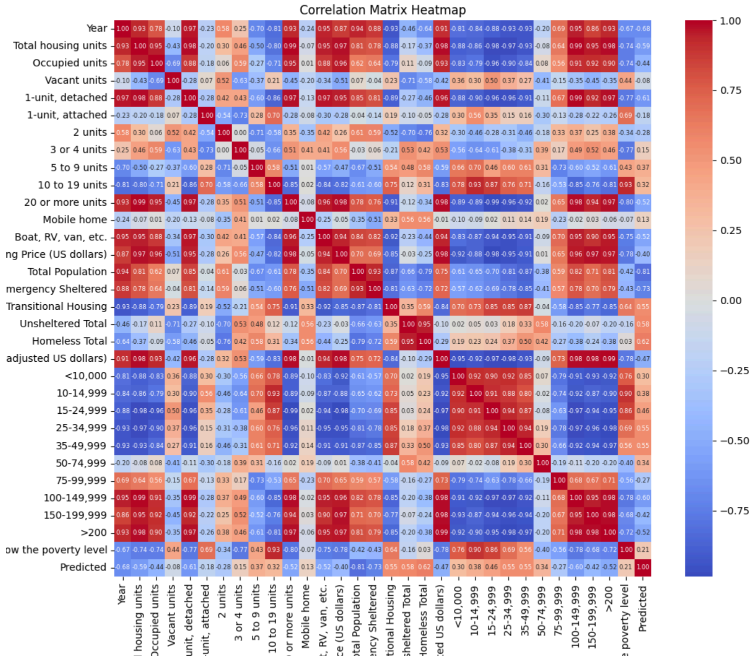


Figure 9: Variable Correlation Heat Map for Albuquerque

From the heat maps, we noticed that the number of housing units, total population, median household income, and median listing price of housing units were highly correlated with homelessness in each city. For Seattle, the four factors had 0.92, 0.90, 0.90, and 0.92 correlation values with homelessness, respectively. For Albuquerque, the four factors had -0.37, -0.79, -0.29, -0.25 correlation values with homelessness, respectively.

We then plotted the number of housing units versus the number of occupied housing units in each city as justification for 2.2-4, which allowed us to disregard total population change over time in our analysis of homelessness.

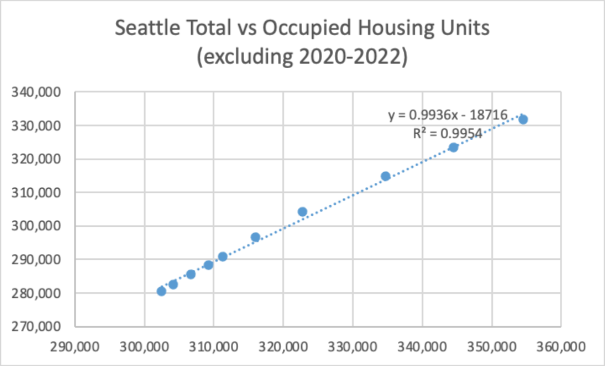


Figure 10: Seattle Total vs Occupied

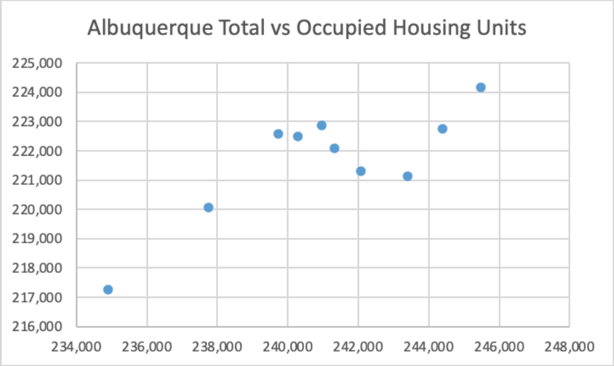


Figure 11: Albuquerque Total vs Occupied

In Seattle, the total vs occupied housing units is correlated linearly with $R^2 = 0.9954$. Thus, the total population must be increasing at a similar rate to the total number of housing units for the units to be occupied at a constant rate. In Albuquerque, the number of total vs occupied housing units is not correlated linearly for the entire period from 2010 to 2019 due to a significant increase in policy changes by the city to deal with the housing crisis[20]. Other than this anomaly, we can assume that these factors are still strongly correlated after the slight adjustment period. As a result, for simplicity, we will use total housing units in our estimate for homelessness, assuming that it will also reflect the impact of population growth.

We then plotted the median listing price of housing units versus the median household income as justification for 2.2-5, which allowed us to disregard median income.

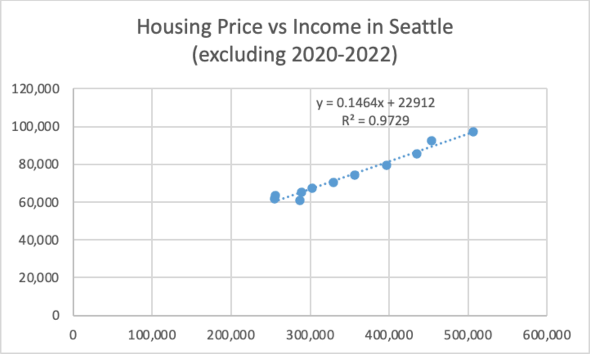


Figure 12: Seattle Price vs Income

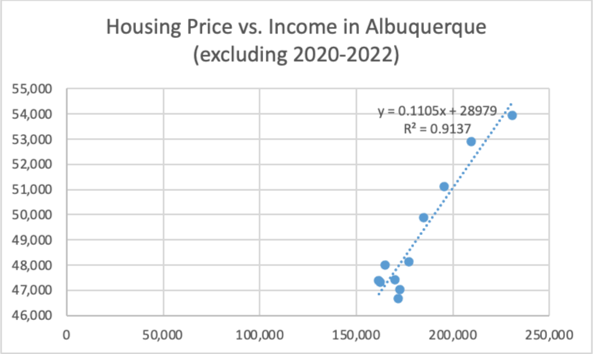


Figure 13: Albuquerque Price vs Income

Given that median housing price vs median household income is correlated linearly with high values of $R^2 = 0.9729$ and $R^2 = 0.9137$ for Seattle and Albuquerque, respectively, we can assume that the median housing price reflects the median household income, so we only need to consider the former in our predictions regarding homelessness numbers in the two cities.

Thus, the two factors we used in our model are the number of housing units and the median household price, as both included high correlations in the heat map with respect to the homeless population.

2.3.2 Model Execution

The predicted total housing units in Seattle and Albuquerque were calculated in Section 1 based on the logistic model in 1.3.2.

We then calculated the median listing price for housing units in Seattle and Albuquerque based on a linear model[21], included below with predictions from 2030 to 2070.

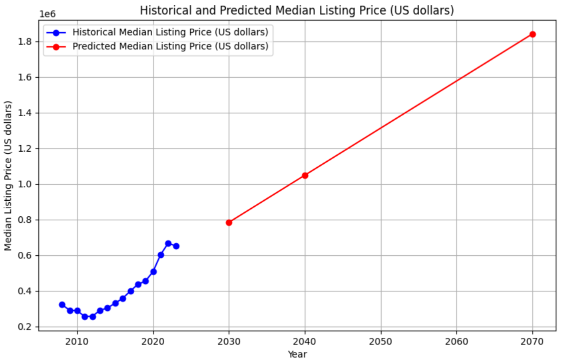


Figure 14: Seattle Listing Price (millions)

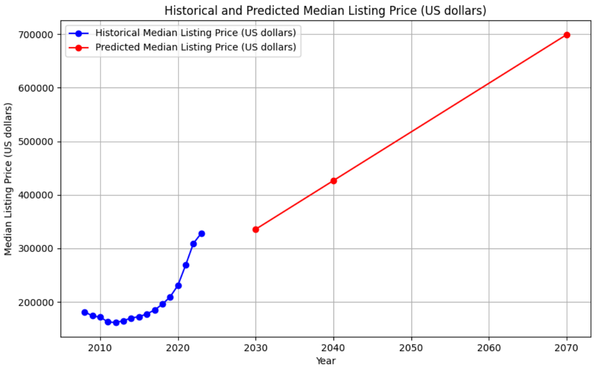


Figure 15: Albuquerque Listing Price

	2030	**2040**	**2070**
Seattle	782,900	1,047,100	1,839,700
Albuquerque	335,600	426,500	699,200

Table 16: Estimated Number of Housing Units (to the nearest hundred)

Given these data points, we can now calculate estimations of homelessness in the two cities.

2.4 Results

We used a multiple regression model to calculate the homelessness numbers in Seattle and Albuquerque based on total housing units and median housing prices. The results are included below.

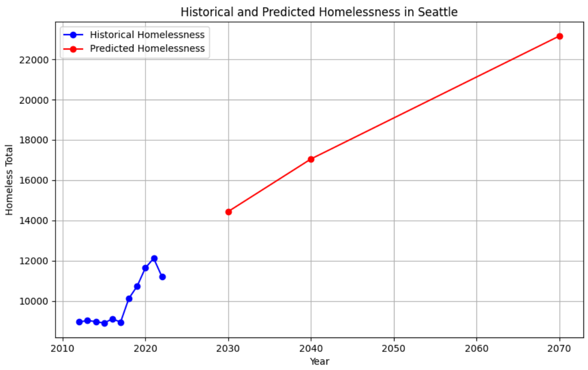


Figure 17: Homelessness in Seattle

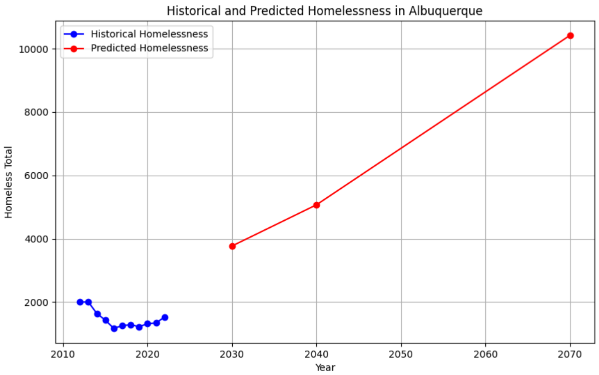


Figure 18: Homelessness in Albuquerque

	2030	**2040**	**2070**
Seattle	14,400	17,100	23,200
Albuquerque	3,800	5,100	10,400

Table 19: Estimated Number of Homeless Individuals (to the nearest hundred)

2.5 Discussion

In Seattle, homelessness will increase to 14,400, 17,100, and 23,200 individuals by 2030, 2040, and 2070, respectively. In Albuquerque, homelessness will increase to 3,800, 5,100, and 10,400, respectively. The results above demonstrate that without intervention through government policies or other means, the homeless populations in each city will increase significantly in the following decades. These results reflect the increase in the overall population expected in Seattle and Albuquerque, as well as the steadily increasing price of housing units in each city that contribute to greater values of homelessness.

2.6 Sensitivity Analysis

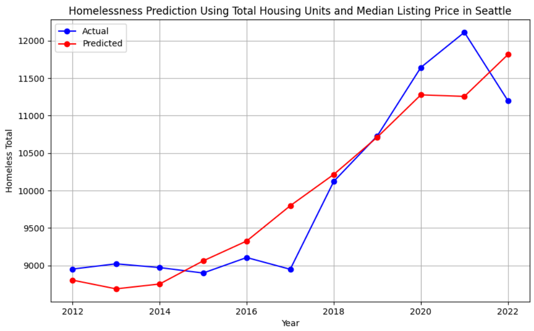


Figure 20: Homelessness in Seattle

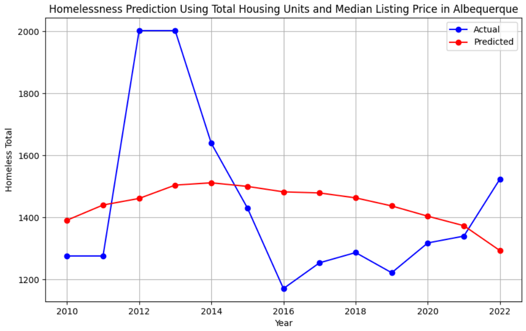


Figure 21: Homelessness in Albuquerque

To determine the accuracy of our predictions, we plotted our model's predicted values for homeless people for the years 2012 to 2022 against the actual data. Then we calculated the cumulative percentage error for each graph, which is the average error calculated for each data point, giving values of 3.482% for Seattle and 14.338% for Albuquerque. The percentage of error for Seattle is relatively low, so we are confident in our prediction for homelessness in this case. For Albuquerque, however, our error is higher, so we are less confident in our prediction here. This error stems from extreme spikes and dips in homelessness from 2011 to 2016, which do not reflect the normal rate of homelessness shifts during normal years, as seen in the data points from 2016 onward.

2.7 Strengths & Weaknesses

One major strength of our model is that we have taken into account the two distinct factors of total housing supply and median listing price that are each highly correlated with the homeless population in each city to perform our analysis. By using these two factors, we were able to achieve more accurate results than by considering just one. Also, our model for Seattle had a low cumulative error when compared to the actual data for the last decade, which shows that it has a high potential for accurate predictions in future years.

Our model was weaker in our prediction of the number of homeless people in Albuquerque due to the abnormal fluctuation from 2011 to 2016, which made it more difficult to get an accurate measure of homelessness over the last decade. As a result, the cumulative error was also higher, decreasing our confidence in our prediction for Albuquerque in comparison to Seattle. Our heat map showed that for Albuquerque, homelessness was generally less correlated to the two factors we used in comparison to Seattle, which made it harder to create an accurate prediction for Albuquerque, as a result. Another weakness is that although we tested high values of correlation for each of our factors, we can not assume causation stems from correlation, even though the logical assumption should indicate that housing units and prices likely influence homelessness.

Q3: Rising from This Abyss

3.1 Defining the Problem

The third problem asks us to create a model that would help determine a long-term plan to handle homelessness in at least one of the cities. We combined results from the previous two questions to create a model for Seattle, Washington.

3.2 Assumptions

3.2-1. **Certain factors, such as household units and unit prices, can be adjusted to model unforeseen circumstances such as natural disasters, economic recessions, or incoming migrant populations.**

- **Justification:** Factors such as household units and unit prices will naturally be affected by unforeseen circumstances, which we can adjust to properly mimic the effects of the event. For example, a reduction of household units in our model would mimic a flood.

3.2-2. **The government and social services are capable of adjusting the variables in our model.**

- **Justification:** We chose specific variables that the government and social services could feasibly adjust, such as unemployment rate and substance abuse.

3.2-3. **Monthly data for variables not influenced by seasons are reflective of the entire year.**

- **Justification:** We determined that variables such as unemployment rate and housing cost depend on the economic landscape rather than season. Thus, monthly data was adapted to be reflective of annual data.

3.3 The Model

3.3.1 Model Development

Symbol	Variable	Unit
u_t	Unemployment rate (%) in Seattle area for year *t*	N/A
d_t	Number of opioid-related deaths in Washington for year *t*	People
c_t	Median household listing price for year *t*	$
h_t	Number of housing units for year *t*	Housing Units

Table 22: Variables in Granger Causality Test

We first determined the variables we wanted to consider that may impact the homeless population in Seattle. We selected the number of housing units and the median listing price[22, 23] from Question 2. These two factors previously showed a high correlation with homelessness in Seattle, so we investigated the potential causation of homelessness as a result of those factors. We also selected unemployment rates, since it is likely correlated with factors like income, which may affect homelessness[24]. The final variable selected was deaths caused by opioid use, as a proxy for drug use in Seattle that is likely related to the increase in homelessness[25]. We were sure to select factors that the government in Seattle can adjust accordingly through policy changes to reduce rates of homelessness in the city, instead of factors that they cannot feasibly control. Potential policies that could be passed to deal with these factors will be detailed in the discussion.

To assess the significance of these factors, our team implemented the Granger causality test to determine if they cause homelessness in Seattle. The Granger causality test is a statistical test that helps determine if one single-time series can forecast another, providing useful evidence for causality between the two variables[26]. We used the F-statistic to determine significance, since our times-series data is continuous. We will be considering any α value below 0.05 as indicating significance. The Granger causality test also includes a parameter for lag, which indicates the number of periods that causality is measured over. For example, if the *p*-value for a lag of one year is less than the α value, then there is a significant statistical causation between the two variables given that the causation occurs after one year. The following equation is the general equation for Granger causality:

$$y_i = \alpha_0 + \sum_{j=1}^{m} \alpha_j y_{i-j} + \sum_{j=1}^{m} \beta_j x_{i-j} + \varepsilon_i$$

3.3.1 Model Execution

To use the datasets we collected for each factor, we first cleaned them so that they each only included years from 2010 through 2019, getting rid of any data points outside of that range to maintain consistency. We were then able to integrate the dataset for homelessness in Seattle with each of the datasets we collected and cleaned to perform a Granger causality test. We used the *statsmodel* Python library to do so, testing each of the four factors listed in 3.3.1 against the homelessness data.

After we obtained the F-statistics for different lag values, namely 1 and 2 years of lag, we conducted F-statistic tests to calculate the *p*-values for each test. We then compared these *p*-values with α, which was our benchmark for statistical significance. The results then showed evidence of causation between our selected factors and homelessness in Seattle over the corresponding lags.

After running our Granger causality tests, we then adjusted each of the significant factors: housing units, median listing price, and deaths by opioid use by 5% to see the impact it would have in reducing homelessness. Housing units were increased, median listing price was decreased, and deaths by opioid use were decreased. We did so by using a multiple regression model, as in our solution to Question 2, using the four factors from our Granger causality test as well as two other correlated factors for the sake of gaining more accurate estimates, total population and median household income, adjusting one factor at a time by 5% to see the impact it would have on homelessness.

3.4 Results

The table below lists the four factors we tested over the two lag values, as well as the F-statistics, *p*-values, and whether they were significant or not according to our benchmark of 0.05.

Factor	Lag (years)	F-statistic	*p*-value	Significance
Housing Units	1	11.3964	0.0149	**Yes**
	2	3.7176	0.1410	No
Median Housing Listing Prices	1	8.7505	0.0253	**Yes**
	2	4.5325	0.1240	No
Deaths by Opioid Use	1	9.6662	0.0209	**Yes**
	2	3.7176	0.1541	No
Unemployment Rate	1	0.0242	0.8815	No
	2	0.3654	0.7211	No

Table 23: Granger Causality Results with 1 and 2 Years Lag

Factor	Reduction in Homelessness (People)
Housing Units	2284
Median Listing Prices	85
Deaths by Opioid Use	360

Table 24: Reduction in Homelessness with Adjustment of Factors

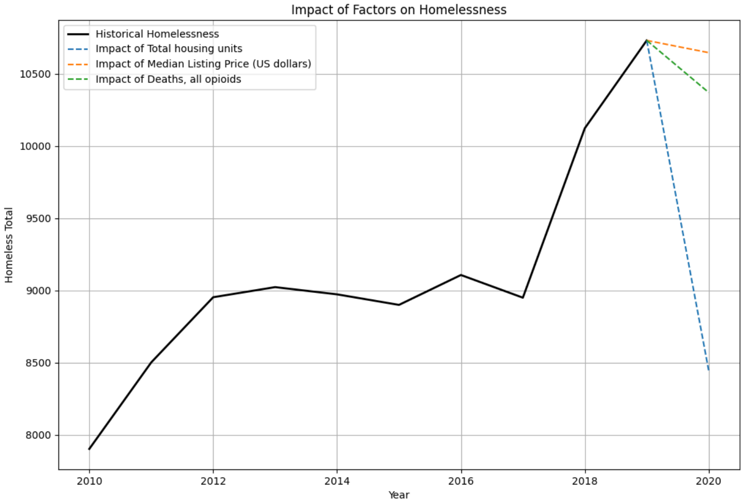


Figure 25: Impact of Factors on Homelessness

3.5 Discussion

Our Granger causality has provided evidence as to causation from our selected four factors to homelessness. The factors that produced significant evidence as to a causal relationship with homelessness are housing units, median housing listing prices, and deaths by opioid use. Since our lag was by year the causations all signal relatively long-term impacts. Interestingly, our analysis shows that the unemployment rate does not cause homelessness in Seattle. The results produced here allow policymakers in Seattle to consider the best factors they should target to reduce homelessness.

Our tests on the impact of the three significant factors showed that increasing housing units would have the greatest impact on reducing homelessness in Seattle, followed by reducing deaths caused by opioid use, and then decreasing the median listing price. This allows policymakers to recognize the best areas to concentrate work in reducing homelessness. For example, increasing housing units and decreasing median listing prices together imply a strong need for more affordable housing, which is expected. Moreover, the high impact of reducing opioid deaths signifies the importance of tackling substance abuse in Seattle to reduce homelessness.

Moreover, these models also allow policymakers to see the potential adverse impacts caused by anomalies like natural disasters or economic downturns. Natural disasters would likely reduce housing units, which would have a significant impact on increasing homelessness as shown in our model. Economic downturns, on the other hand, would likely increase median listing prices, also having a significant impact on increasing homelessness. Adjusting these values in our model would easily allow policymakers to consider the best steps to take and the best factors to target in such abnormal circumstances.

3.7 Strengths & Weaknesses

The primary strength of our model is that our usage of the Granger causality test allows us to test for causality of certain factors on homelessness in order to decide which factors policymakers should focus on to reduce homelessness. Moreover, our test on the impacts of significant factors when adjusted allows us to identify the factors that have the greater quantifiable impact on homelessness. That gives policymakers a clear direction to pursue active change that will have an impact. Moreover, by analyzing the impact of each factor separately, policymakers can decide which factors would be the easiest to implement in policies to reduce homelessness, as well as which would have the largest impact.

However, our model's usage of the Granger causality test assumes that the relationship between variables is linear, which may not always be true. Further, while the purpose of this test is to suggest causation, it is impossible to confidently deduce that one variable directly causes another. Thus, the significance levels found may be trivial, as correlation does not imply causation.

4 Conclusion

We examined the problem of homelessness and housing in the cities of Seattle and Albuquerque, and their predicted effect over the next 10, 20, and 50 years. To predict the number of housing units available, we used a logistic regression model. The model predicted a housing supply of 411.6, 453, and 513 thousand housing units for Seattle in the next 10, 20, and 50 years respectively. It also predicted a supply of 261, 270.9, 290.8 thousand housing units for Albuquerque in the next 10, 20, and 50 years respectively.

Meanwhile, for homelessness, we used multivariate linear regression to identify total housing units and median listing price as key factors that are correlated with homelessness and then used multivariable linear regression to predict how homelessness would change over the years. The model predicted 14.4, 17.1, and 23.2 thousand homeless people in Seattle in the next 10, 20, and 50 years respectively. It also predicted 3.8, 5.1, and 10.4 thousand homeless people in Albuquerque in the next 10, 20, and 50 years respectively.

Finally, we examined the causation of factors that are correlated with homelessness in Seattle that may be useful for government planning, namely housing units, median listing price, unemployment, and deaths by opioid use. We found that every factor but unemployment was significant over a 1 year lag time. The impact of the three factors was then modeled by adjusting the values by 5% each to see the quantifiable impact on homelessness numbers in Seattle. Housing units reduced homelessness numbers by 2284 individuals, deaths by opioid use, and listing prices decreased by 360 and 84 individuals, respectively.

In summary, our findings suggest the rapid growth of housing supply as well as homelessness in the next 10, 20, and 50 years. Homelessness is a pervasive problem around the globe, but there are some key factors such as housing units, median listing price, and drug use, which indicate causation of homelessness. These findings will be useful for policymakers and world leaders in determining some facets to aim for when tackling such a tremendous task.

5 References

1. https://www.cnbc.com/2021/11/10/home-prices-are-now-rising-much-faster-than-incomes-studies-show.html
2. https://endhomelessness.org/blog/what-can-and-cant-local-government-do-to-address-homelessness/
3. https://www.seattle.gov/documents/Departments/OPCD/Demographics/Development%20Capacity%20Report.pdf
4. https://documents.cabq.gov/planning/UDD/CompPlan2017/CompPlan-Chapter2.pdf
5. https://www.census.gov/library/stories/2021/10/zillow-and-census-bureau-data-show-pandemics-impact-on-housing-market.html
6. https://www.ffiec.gov/census/htm/2020CensusInfoSheet.htm
7. https://www.seattle.gov/sdci/codes/codes-we-enforce-(a-z)/zoning
8. https://www.sciencedirect.com/science/article/pii/S1877705815042745/pdf?md5=36e6c3dd93fc34a4e15eece53949a5ca&pid=1-s2.0-S1877705815042745-main.pdf
9. https://www.seattle.gov/utilities/protecting-our-environment/community-programs/climate-change/projections-and-maps
10. https://www.wcvb.com/article/sea-level-rise-forecasting-our-future/39135613
11. https://www.huduser.gov/periodicals/ushmc/summer2001/summary-2.html
12. https://shs.hal.science/halshs-00440438/document
13. https://www.seattle.gov/documents/Departments/OPCD/Demographics/Development%20Capacity%20Report.pdf
14. https://documents.cabq.gov/planning/UDD/CompPlan2017/CompPlan-Chapter2.pdf
15. https://shs.hal.science/halshs-00440438/document
16. https://www.britannica.com/science/population-ecology/Logistic-population-growth
17. https://iacis.org/iis/2018/2_iis_2018_109-118.pdf
18. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2795301
19. https://seaborn.pydata.org/
20. https://www.cabq.gov/health-housing-homelessness/documents/2018-whtf-report-final.pdf
21. https://iacis.org/iis/2018/2_iis_2018_109-118.pdf
22. A Tale of Two Crises, MathWorks Math Modeling Challenge 2024, curated data, https://m3challenge.siam.org/kdfrldh/.
23. A Tale of Two Crises, MathWorks Math Modeling Challenge 2024, curated data, https://m3challenge.siam.org/kdfrldh/.
24. https://fred.stlouisfed.org/series/SEAT653URN
25. https://adai.uw.edu/wadata/opiate_home.htm
26. https://towardsdatascience.com/a-quick-introduction-on-granger-causality-testing-for-time-series-analysis-7113dc9420d2

6 Code Appendix

6.1 Part 1: It Was the Best of Times

```
#The following code is used for answering Q1 in predicting the number of
housing units in 2030, 2040, and 2070 in Seattle and for sensitivity
analysis. The same code with adjusted file paths was used for the
calculations for Albuquerque.

import pandas as pd
import numpy as np
from scipy.optimize import curve_fit
import matplotlib.pyplot as plt

def logistic_growth(t, P0, r):
    #Carrying capacity for Seattle based off sources
    K = 532000
    return K / (1 + ((K - P0) / P0) * np.exp(-r * t))

def add_noise(y, sigma=0.05):
    #Adds a random 5% noise for jittering to test model sensitivity
    return y * (1 + np.random.normal(0, sigma, size=y.shape))

seattle_housing = pd.read_csv('seattle_housing_units.csv', converters={
    'Total housing units': lambda x: int(x.replace(',', ''))
})
seattle_housing.columns = seattle_housing.columns.str.strip()

# We chose to include Covid data, uncomment to not include covid data
# seattle_housing = seattle_housing[seattle_housing['Year'] < 2020]

X = seattle_housing['Year'] - seattle_housing['Year'].min()
y = seattle_housing['Total housing units']

params, _ = curve_fit(logistic_growth, X, y, p0=[y.iloc[0], 0.1])

start_year = seattle_housing['Year'].min()
end_year = 2070
future_years = np.arange(start_year, end_year + 1)
future_times = future_years - start_year
```

```
future_predictions = logistic_growth(future_times, *params)

selected_future_years = np.array([2030, 2040, 2070])
selected_predictions = logistic_growth(selected_future_years - start_year,
*params)
# Sensitivity analysis averaged over 5 trials
num_trials = 5
percent_differences = []

for _ in range(num_trials):
    y_jittered = add_noise(y)
    params_jittered, _ = curve_fit(logistic_growth, X, y_jittered,
p0=[y.iloc[0], 0.1])
    selected_predictions_jittered = logistic_growth(selected_future_years
- start_year, *params_jittered)
    percent_diff = 100 * np.abs(selected_predictions -
selected_predictions_jittered) / selected_predictions
    percent_differences.append(percent_diff)

average_percent_diff = np.mean(percent_differences, axis=0)
print("Average Percent Differences for Selected Future Predictions:")
for year, diff in zip(selected_future_years, average_percent_diff):
    print(f'Year: {year}, Average Percent Difference: {diff:.3f}%')

print("Estimated Number of Houses for Selected Future Years:")
for year, prediction in zip(selected_future_years, selected_predictions):
    print(f'Year: {year}, Estimated Total Housing Units:
{prediction:.0f}')

plt.figure(figsize=(10, 6))
plt.scatter(seattle_housing['Year'], y, color='blue', label='Actual Data')
plt.plot(future_years, logistic_growth(future_times, *params),
color='red', linewidth=2, label='Logistic Growth Model')
plt.scatter(selected_future_years, selected_predictions, color='green',
label='Selected Future Predictions')
plt.xlabel('Year')
plt.ylabel('Total Housing Units')
plt.title('Seattle Housing Units Prediction with Logistic Growth')
plt.legend()
plt.show()
```

6.2 Part 2: It Was the Worst of Times

```
#The following code is used for answering Q2 in predicting the
homelessness population in 2030, 2040, and 2070 in Seattle and for
sensitivity analysis. It includes the implementation to generate the
heatmap from the multivariate linear regression and the multiple linear
regression for projecting homelessness. The same code with adjusted file
paths was used for the calculations for Albuquerque.
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.impute import SimpleImputer
from sklearn.metrics import r2_score

def plot_variable_with_forecast(df, column_name, forecasts, future_years):
    plt.figure(figsize=(10, 6))
    plt.plot(df['Year'], df[column_name], marker='o', color='blue',
label=f'Historical {column_name}')
    plt.plot(future_years, forecasts, marker='o', color='red',
label=f'Predicted {column_name}')
    plt.xlabel('Year')
    plt.ylabel(column_name)
    plt.title(f'Historical and Predicted {column_name}')
    plt.legend()
    plt.grid(True)
    plt.show()

def convert_to_int(x):
    return int(x.replace(',', '')) if isinstance(x, str) and ',' in x else
x

housing_units = pd.read_csv('seattle_housing_units.csv')
housing_price = pd.read_csv('seattle_housing_price.csv')
total_population = pd.read_csv('seattle_total_population.csv')
homelessness = pd.read_csv('seattle_homelessness.csv')
income = pd.read_csv('seattle_income.csv')

cols_to_convert_housing_units =
housing_units.select_dtypes(include=['object']).columns
```

```
housing_units[cols_to_convert_housing_units] = housing_units[cols_to_convert_housing_units].applymap(convert_to_int)

cols_to_convert_housing_price = housing_price.select_dtypes(include=['object']).columns
housing_price[cols_to_convert_housing_price] = housing_price[cols_to_convert_housing_price].applymap(convert_to_int)

cols_to_convert_total_population = ['Total Population']
total_population[cols_to_convert_total_population] = total_population[cols_to_convert_total_population].applymap(convert_to_int)

cols_to_convert_homelessness = homelessness.select_dtypes(include=['object']).columns
homelessness[cols_to_convert_homelessness] = homelessness[cols_to_convert_homelessness].applymap(convert_to_int)

cols_to_convert_income = income.select_dtypes(include=['object']).columns
income[cols_to_convert_income] = income[cols_to_convert_income].applymap(convert_to_int)

housing_units.columns = housing_units.columns.str.strip()
housing_price.columns = housing_price.columns.str.strip()
total_population.columns = total_population.columns.str.strip()
homelessness.columns = homelessness.columns.str.strip()
income.columns = income.columns.str.strip()

data_frames = [housing_units, housing_price, total_population, homelessness, income]
merged_data = pd.concat(data_frames, axis=1).loc[:,~pd.concat(data_frames, axis=1).columns.duplicated()]
merged_data = merged_data[merged_data['Year'].between(2012, 2023)]
merged_data = merged_data.drop(columns=[col for col in merged_data.columns if 'Unnamed' in col], errors='ignore')

# Selecting predictor columns (choosing only total housing units and median listing price)
X = merged_data[['Total housing units', 'Median Listing Price (US dollars)']]
```

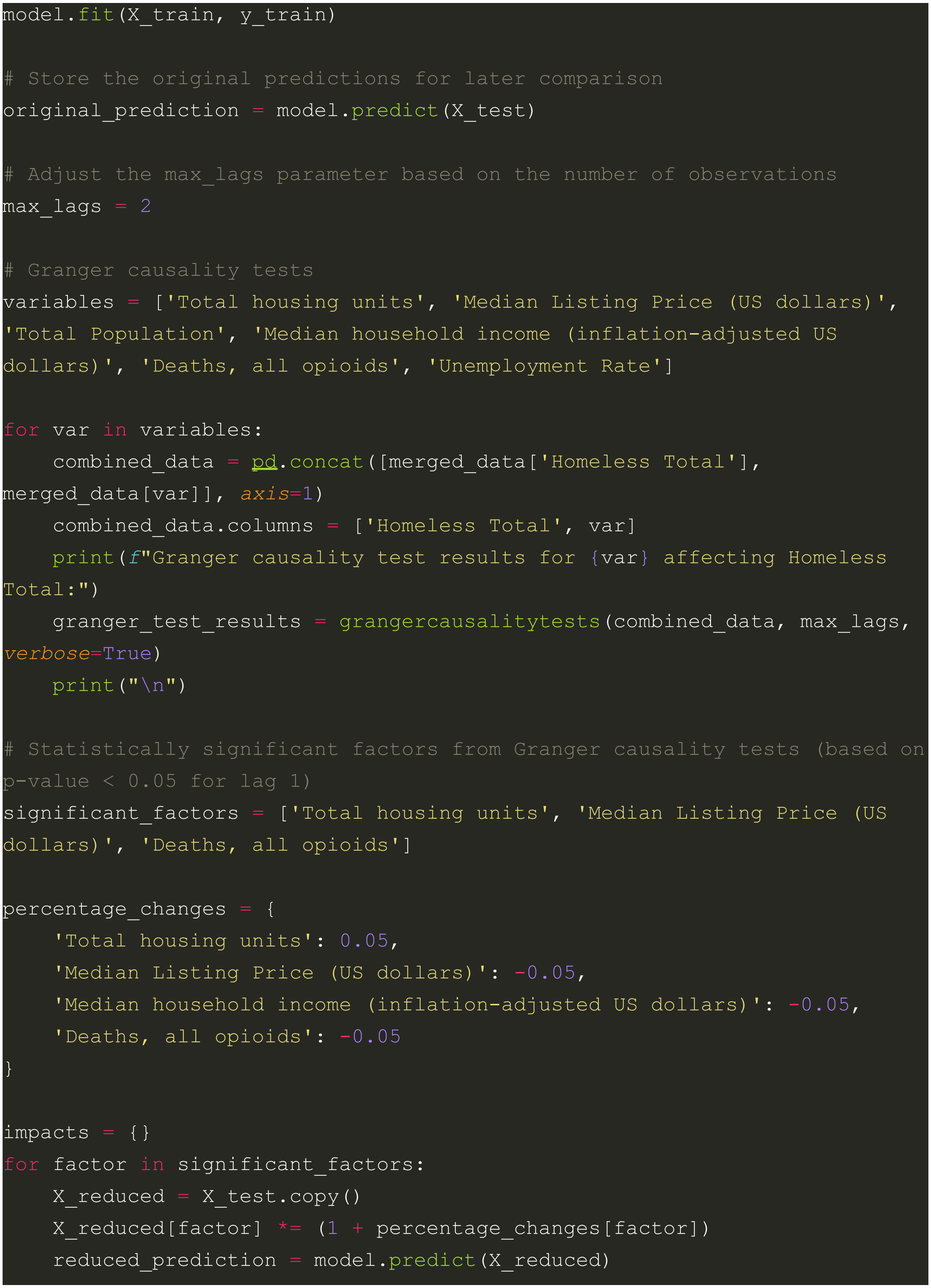
```
# Impute missing values because NaN is not allowed
imputer = SimpleImputer(strategy='mean')
X_imputed = imputer.fit_transform(X)

# Target variable to be projected
y = merged_data['Homeless Total']

model = LinearRegression()
model.fit(X_imputed, y)
merged_data['Predicted'] = model.predict(X_imputed)

# Create a heatmap of the correlation matrix
corr_matrix = merged_data.corr()
plt.figure(figsize=(12, 10))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f',
annot_kws={"size": 6})
plt.title('Correlation Matrix Heatmap')
plt.show()

# Forecasting housing price for 2030, 2040, and 2070 using linear
regression
housing_price_model = LinearRegression()
housing_price_model.fit(housing_price[['Year']], housing_price['Median
Listing Price (US dollars)'])
future_years = pd.DataFrame({'Year': [2030, 2040, 2070]})
housing_price_forecast = housing_price_model.predict(future_years)
# Forecasting predictor variables for 2030, 2040, and 2070 using previous
values for total housing units
future_years_df = pd.DataFrame({
    'Year': [2030, 2040, 2070],
    'Total housing units': [411600, 453000, 513000],
    'Median Listing Price (US dollars)': housing_price_forecast,
})
# Predicting homelessness for the future years using the fitted model from
projections of predictor variables
future_homelessness_predictions =
model.predict(future_years_df.drop(columns=['Year']))
# Adding the predictions to the DataFrame
```

```
future_years_df['Predicted Homeless Total'] =
future_homelessness_predictions

plt.figure(figsize=(10, 6))
plt.plot(merged_data['Year'], merged_data['Homeless Total'], marker='o',
color='blue', label='Historical Homelessness')
plt.plot(future_years_df['Year'], future_years_df['Predicted Homeless
Total'], marker='o', color='red', label='Predicted Homelessness')
plt.xlabel('Year')
plt.ylabel('Homeless Total')
plt.title('Historical and Predicted Homelessness in Seattle')
plt.legend()
plt.grid(True)
plt.show()

plot_variable_with_forecast(housing_units, 'Total housing units', [411600,
453000, 513000], future_years)
plot_variable_with_forecast(housing_price, 'Median Listing Price (US
dollars)', housing_price_forecast, future_years)
print("Predictions for Total Housing Units:")
print(f"2030: 411600")
print(f"2040: 453000")
print(f"2070: 513000")
print("\nPredictions for Median Listing Price (US dollars):")
print(f"2030: {housing_price_forecast[0]}")
print(f"2040: {housing_price_forecast[1]}")
print(f"2070: {housing_price_forecast[2]}")
print("\nPredictions for Homeless Total:")
print(f"2030: {future_homelessness_predictions[0]}")
print(f"2040: {future_homelessness_predictions[1]}")
print(f"2070: {future_homelessness_predictions[2]}")

percentage_error = ((merged_data['Homeless Total'] -
merged_data['Predicted']).abs() / merged_data['Homeless Total']) * 100

# Calculate the average percentage error of all the points over number of
points
cumulative_percentage_error = percentage_error.sum() /
len(percentage_error)
print('Cumulative Percentage Error:', cumulative_percentage_error)
```

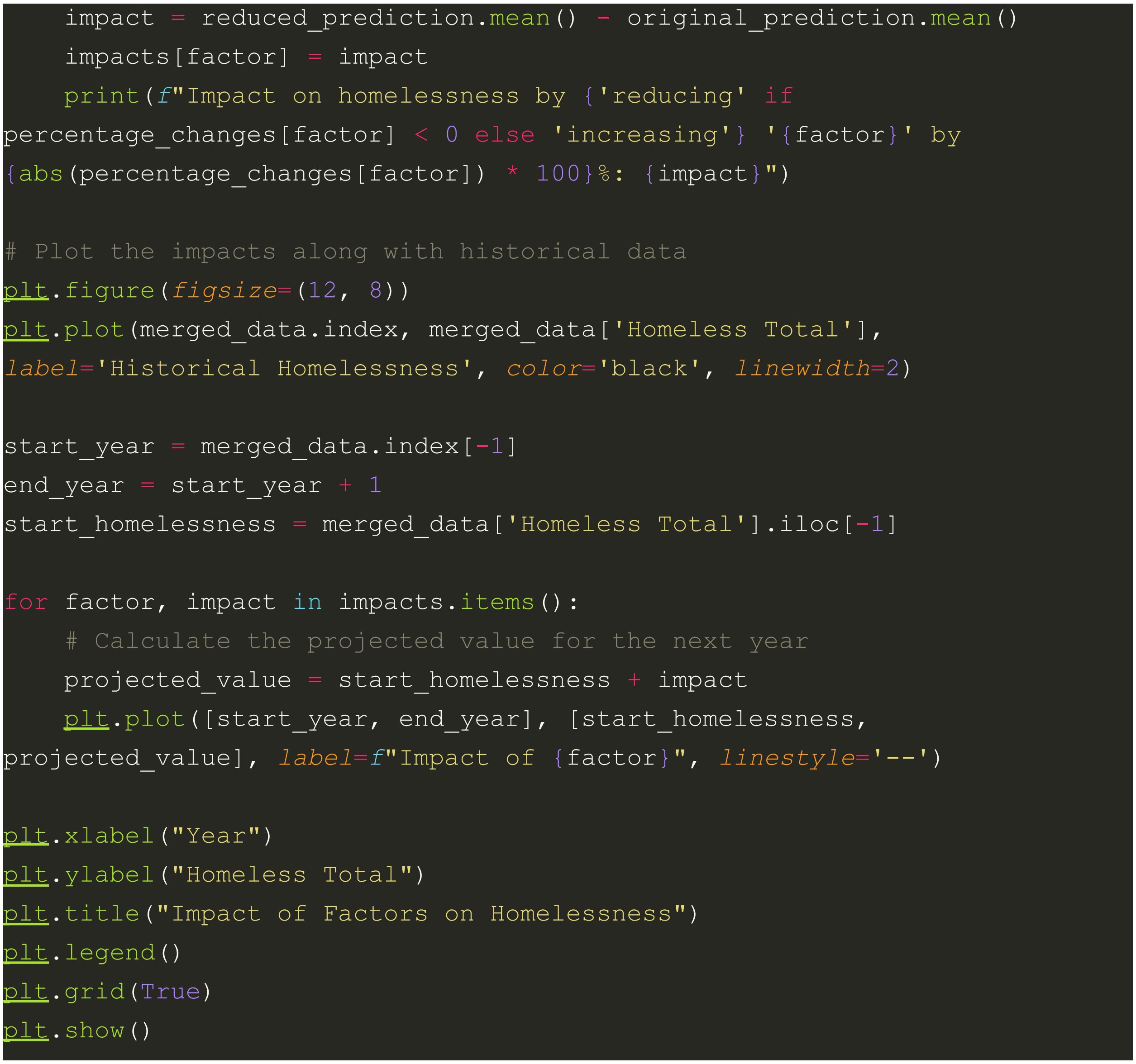
6.3 Part 3: Rising from This Abyss

```
#The following code was used to answer Q3, using Granger causality tests
to find statistically significant factors and then a multiple regression
to evaluate the impact of changing the statistically significant factors
by increasing or reducing 5%.
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from statsmodels.tsa.stattools import grangercausalitytests
import matplotlib.pyplot as plt

def convert_to_int(x):
    return int(x.replace(',', '')) if isinstance(x, str) and ',' in x else
x

housing_units = pd.read_csv('seattle_housing_units.csv')
housing_price = pd.read_csv('seattle_housing_price.csv')
total_population = pd.read_csv('seattle_total_population.csv')
homelessness = pd.read_csv('seattle_homelessness.csv')
income = pd.read_csv('seattle_income.csv')
opioid_deaths = pd.read_csv('opioid_deaths.csv')
unemployment = pd.read_csv('unemployment.csv', names=['Year',
'Unemployment Rate'], skiprows=1)

opioid_deaths['year'] = opioid_deaths['year'].apply(lambda x: int(x))
unemployment['Year'] = unemployment['Year'].apply(lambda x: int(x))

cols_to_convert_housing_units =
housing_units.select_dtypes(include=['object']).columns
housing_units[cols_to_convert_housing_units] =
housing_units[cols_to_convert_housing_units].applymap(convert_to_int)

cols_to_convert_housing_price =
housing_price.select_dtypes(include=['object']).columns
housing_price[cols_to_convert_housing_price] =
housing_price[cols_to_convert_housing_price].applymap(convert_to_int)

cols_to_convert_total_population = ['Total Population']
```

```
total_population[cols_to_convert_total_population] =
total_population[cols_to_convert_total_population].applymap(convert_to_int
)

cols_to_convert_homelessness =
homelessness.select_dtypes(include=['object']).columns
homelessness[cols_to_convert_homelessness] =
homelessness[cols_to_convert_homelessness].applymap(convert_to_int)

cols_to_convert_income = income.select_dtypes(include=['object']).columns
income[cols_to_convert_income] =
income[cols_to_convert_income].applymap(convert_to_int)

housing_units.columns = housing_units.columns.str.strip()
housing_price.columns = housing_price.columns.str.strip()
total_population.columns = total_population.columns.str.strip()
homelessness.columns = homelessness.columns.str.strip()
income.columns = income.columns.str.strip()

# Merge the dataframes by Year
data_frames = [housing_units, housing_price, total_population,
homelessness, income, opioid_deaths, unemployment]
merged_data = pd.concat(data_frames, axis=1).loc[:,
~pd.concat(data_frames, axis=1).columns.duplicated()]
merged_data = merged_data[merged_data['Year'].between(2010, 2019)]
merged_data.set_index('Year', inplace=True)

# Multiple Linear Regression model
X = merged_data[['Total housing units', 'Median Listing Price (US
dollars)', 'Total Population', 'Median household income
(inflation-adjusted US dollars)', 'Deaths, all opioids', 'Unemployment
Rate']]
y = merged_data['Homeless Total']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
random_state=42)

# Fit the model
model = LinearRegression()
```



```
model.fit(X_train, y_train)

# Store the original predictions for later comparison
original_prediction = model.predict(X_test)

# Adjust the max_lags parameter based on the number of observations
max_lags = 2

# Granger causality tests
variables = ['Total housing units', 'Median Listing Price (US dollars)',
'Total Population', 'Median household income (inflation-adjusted US
dollars)', 'Deaths, all opioids', 'Unemployment Rate']

for var in variables:
    combined_data = pd.concat([merged_data['Homeless Total'],
merged_data[var]], axis=1)
    combined_data.columns = ['Homeless Total', var]
    print(f"Granger causality test results for {var} affecting Homeless
Total:")
    granger_test_results = grangercausalitytests(combined_data, max_lags,
verbose=True)
    print("\n")

# Statistically significant factors from Granger causality tests (based on
p-value < 0.05 for lag 1)
significant_factors = ['Total housing units', 'Median Listing Price (US
dollars)', 'Deaths, all opioids']

percentage_changes = {
    'Total housing units': 0.05,
    'Median Listing Price (US dollars)': -0.05,
    'Median household income (inflation-adjusted US dollars)': -0.05,
    'Deaths, all opioids': -0.05
}

impacts = {}
for factor in significant_factors:
    X_reduced = X_test.copy()
    X_reduced[factor] *= (1 + percentage_changes[factor])
    reduced_prediction = model.predict(X_reduced)
```



```
    impact = reduced_prediction.mean() - original_prediction.mean()
    impacts[factor] = impact
    print(f"Impact on homelessness by {'reducing' if
percentage_changes[factor] < 0 else 'increasing'} '{factor}' by
{abs(percentage_changes[factor]) * 100}%: {impact}")

# Plot the impacts along with historical data
plt.figure(figsize=(12, 8))
plt.plot(merged_data.index, merged_data['Homeless Total'],
label='Historical Homelessness', color='black', linewidth=2)

start_year = merged_data.index[-1]
end_year = start_year + 1
start_homelessness = merged_data['Homeless Total'].iloc[-1]

for factor, impact in impacts.items():
    # Calculate the projected value for the next year
    projected_value = start_homelessness + impact
    plt.plot([start_year, end_year], [start_homelessness,
projected_value], label=f"Impact of {factor}", linestyle='--')

plt.xlabel("Year")
plt.ylabel("Homeless Total")
plt.title("Impact of Factors on Homelessness")
plt.legend()
plt.grid(True)
plt.show()
```