MathWorks Math Modeling Challenge 2024

The Pingry School

Team #17773, Basking Ridge, New Jersey
Coach: Bradford Poprik
Students: Elbert Ho, Laura Liu, Annabelle Shilling, Evan Xie, Alan Zhong

M3 Challenge FINALIST—$5,000 Team Prize

JUDGE COMMENTS

Specifically for Team #17773—Submitted at the close of triage judging

COMMENT 1: Your summary is present and adequate, assumption statements are provided with detailed justification, and models are presented and well explained. The paper is also well organized. Citing valid sources would help strengthen the paper.

COMMENT 2: Your team has put lot of effort in modeling. In Question 1, the use of a model with upper cap may produce more realistic predictions. In the same question, a strong assumption of independency in Regression analysis was used while in Question 2 that assumption was not used. A reader may expect more explanation on this.

COMMENT 3: Good executive summary. I found your use of housing permits to be creative and an interesting approach. You did a good job explaining the nuances of each model. An enjoyable read! Note: throughout the paper you refer to 2084 as the 50-year point but that would be 2074.

COMMENT 4: I liked that you tried to find mels that would fit the data give: such as ARIMA and Random Forests. Your model could be made a little stronger by giving a description of those models and what they do in terms that would be understandable to someone not mathematically familiar with them.

M3 Challenge 2024

It Was the Age of Wisdom: To Resolve the Coupled Crises of Homelessness and Housing Shortages

Team #17773
March 4, 2024

Executive Summary

Dear U.S. Department of Housing and Urban Development Secretary Marcia L. Fudge,

Homelessness manifests as a cruel stain on people's quality of life. The United States, despite being one of the wealthiest countries in the world, is no exception to this unfortunate reality. In an effort to reduce or altogether eliminate homelessness, a developed understanding of the variables (eg. housing availability, inflation) which correlate or contribute to it is requisite.

In the first section of our report, we predict the future of housing supply in Seattle and Albuquerque using simple linear regression. In order to improve accuracy, we found datasets outside those provided that contained population statistics monthly instead of annually, allowing us to extrapolate from 12 times as much information. Our model predicts that there will be 439,932 housing units, 499,491 housing units, and 678,169 units in Seattle in 2034, 2044, and 2084 respectively. In Albuquerque, there will be 275,235 housing units, 291,931 housing units, and 342,017 housing units respectively in 2034, 2044, and 2084 respectively.

In the second section of our report, we predict the future of the homeless population in Seattle and Albuquerque. We separated the homeless population into three components—unsheltered, transitional housing, and emergency shelter—as we observed more clear trends in those components than the overall homeless population. In order to accurately represent cyclical patterns in our data, we decided to use ARIMA models for our simulations, allowing us to achieve realistic, oscillating predictions. Our model predicts that the homeless population of Seattle will be 10603, 11707, and 8180 for the next 10, 20, and 50 years, respectively. We also predict that the homeless population of Albuquerque will be 1188, 1249, and 1400 in the next 10, 20, and 50 years, respectively.

In the third section of our report, we studied the various correlations between homelessness and statistics like the Consumer Price Index (CPI) and housing prices to determine how policymakers can reduce the rate of homelessness. We used linear regression to determine positive or negative correlation between factors, and then a random forest to rank their correlations. We found housing prices contributed the most to the rate of homelessness. Thus, we advise policy makers to create policies to lower housing prices and create affordable housing in order to decrease rates of homelessness.

Table of Contents

Question 1: It Was the Best of Times

1.1 Defining the Problem

The problem asks us to develop a model to predict the housing supply in two U.S. cities: Seattle, Washington and Albuquerque, New Mexico. Our model will take into consideration the number of building permits to determine total housing supply.

1.1 Assumptions

1-1 **The growth of private building permits and the total housing supply will be adjustable by a multiplier.**

- Justification: Most buildings in a city require a building permit in order to be built. However, not all of these permits eventually result in a housing unit being built. In addition, not every building is private. Still, the ratio of private permits issued compared to total housing should be approximately constant, so we can extrapolate the number of private permits to the total housing supply.

1-2 **There are no space considerations for housing in a city. In other words, the total number of housing units is unbounded.**

- Justification: Due to zoning laws or other considerations, the total number of housing units may eventually reach an asymptote. We elected to ignore this constraint since cities can possess millions of housing units. This is the case for New York City, which has far more housing units than either Seattle or Alburquerque. These large cities also exhibit a tendency to build upwards upon reaching a land constraint, where there exist no more plots of land to construct housing units upon. Therefore, both Seattle and Albuquerque have time to expand the number of their housing units.

1-3 **There will not be a major change in the housing market in the next 50 years.**

- Justification: The housing market is not easy to predict and it heavily influences the housing supply. For example, if there are a larger number of houses on the market, there will probably be few houses built. However, to keep this model simple, we decided to ignore these considerations.

1.3 Variables

For the purposes of this table, each symbol will be subscripted for the cities. A subscript of "SEA" corresponds to Seattle and a subscript of "ALB" corresponds to Albuquerque.

Symbol	Definition	Units

$H_{SEA\text{-}Y}$, $H_{ALB\text{-}Y}$	Number of total housing units in the city in year Y	units
$P_{SEA\text{-}Y}$, $P_{ALB\text{-}Y}$	Number of private building permits in the city in year Y	units
$U_{SEA\text{-}M}$, $U_{ALB\text{-}M}$	Number of total housing units in the city M months after 2010	units
M_{SEA}, M_{ALB}	Ratio of private building permits to housing units completed (adjustable multiplier)	unitless

Table 1: Variable definitions for Problem 1

1.4 Model

1.4.1 Developing the Model

We chose a simple linear regression to predict the growth of the total number of available houses. However, there were only 13 data points provided, and although the R^2 was 0.905, we felt that the model was overfit due to the lack of data. Still, we believed that a regression would work because the number of housing units seemed to depend exclusively on time. Due to a lack of data on housing units, we decided to use building permits instead, as these were readily available on U.S. census data.

1.4.1 Executing the Model

Since we wanted to find more data points, we looked for another dataset that would give us this data. In the end, we found the U.S. Census Bureau Survey Building Permits Survey which gave month by month data for housing permits for major U.S. regions and cities. Since this only gave the **change** in total number of houses every month, we needed to take a cumulative sum for our total houses.

After doing this, we ran a simple linear regression on this data, we found that the model had an R^2 of about 0.96 (for Seattle) and an R^2 of only 0.88 for Alburquerque. However, the line had a major drop around the middle and we believed that this was due to the 2008 Financial Crisis which severely affected the housing market and thus led to fewer housing units being built. We decided to write the code in Python for flexibility, however this could have also been potentially executed in Excel or a spreadsheet.

Therefore, we decided to truncate our data to only after 2010. This also lined up with the given data, so it gave us a better way to eventually calculate our multiplier. After we did this, we found that the R^2 improved to 0.990 (Seattle) and 0.998 (Albuquerque) and also looked like almost a straight line for both cities.

We then needed to extrapolate our data to housing while also making predictions for the future. To transform *P* into *H*, we decided to calculate the change in housing units between January 2010 and December 2022 using the provided data as well as the change in building permits between January 2010 and December 2022 from the original dataset. The ratio between these differences gave us our multiplier.

$$M_{ALB} = \frac{P_{ALB-2022} - P_{ALB-2010}}{H_{ALB-2022} - H_{ALB-2020}}$$

$$M_{SEA} = \frac{P_{SEA-2022} - P_{SEA-2010}}{H_{SEA-2022} - H_{SEA-2020}}$$

We found that $M_{ALB} \approx 1.33$ and $M_{SEA} \approx 0.69$. We would have expected for the multiplier to be below 1 because each housing unit should have a permit. However, the housing permits in the dataset only included private permits and did not include houses built with public funds. Therefore, it is possible that Albuquerque just built many houses using public money and thus did not have as many private permits. Finally, since our model gave the change in housing units, we added the number of starting housing units from the original provided dataset for each city (234,891 and 302,465 for Albuquerque and Seattle respectively).

1.5 Results

Below is a graph for our linear regression model for Projected number of housing units in Seattle. Training data was between 0 and 171 months after 2010. Predictions go for 5 years (600) months afterwards. The equation of the linear regression model was:

$$U_{SEA-M} = 496.328M + 295003$$

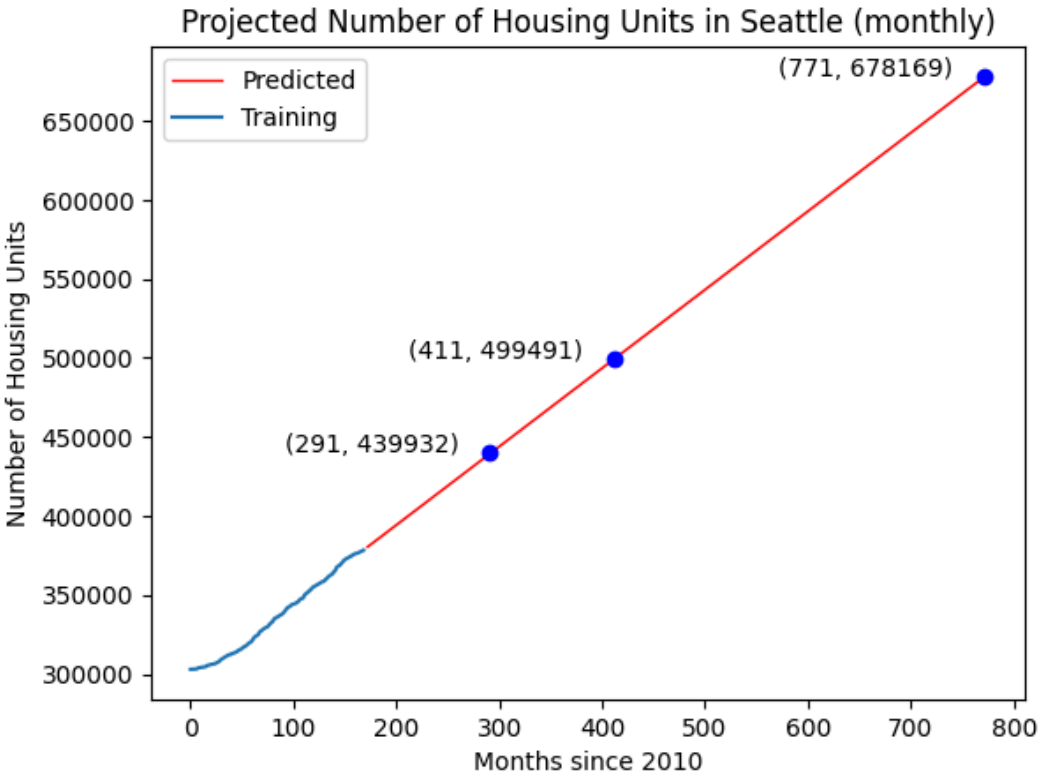

Figure 1: Graph of training and predicted housing units through 771 months in Seattle after January 2010 (50 years after March of 2024)

Seattle projected Housing Units by Month (10, 20, 50 year predictions)

Month	Estimated Number of Housing Units
January 2034	439,932
January 2044	499,491
January 2084	678,169

Table 2: Estimation of housing units in Seattle 10, 20, and 50 years in the future

Below is a graph for our linear regression model for Projected number of housing units in Albuquerque. Training data was between 0 and 171 months after 2010. Predictions go for 5 years (600) months afterwards. The equation of the linear regression model was:

$$U_{ALB-M} = 139.128M + 234610$$

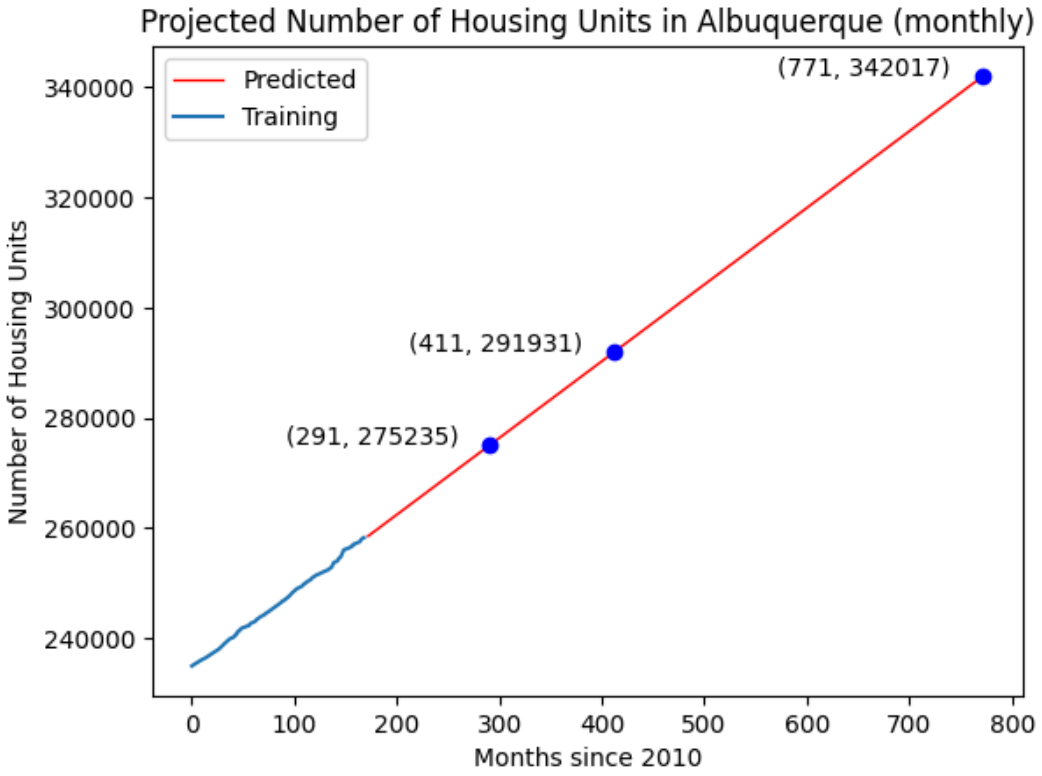

Figure 2: Graph of training and predicted housing units in Albuquerque through 771 months after January 2010 (50 years after March of 2024)

Albuquerque projected Housing Units by Month (10, 20, 50 year predictions)

Month	Estimated Number of Housing Units
January 2034	275,235
January 2044	291,931
January 2084	342,017

Table 3: Estimation of housing units in Albuquerque 10, 20, and 50 years in the future

1.6 Discussion

In summary, our model predicts that there will be 439,932 housing units, 499,491 housing units, and 678,169 units in Seattle in 2034, 2044, and 2084 respectively. In Albuquerque, there will be 275,235 housing units, 291,931 housing units, and 342,017 housing units respectively in 2034, 2044, and 2084 respectively. From this, we conclude that housing units will continue to increase in the U.S. in the next few decades. Given that the population in the U.S. continues to grow, this seems reasonable as more people need more houses to live in. In particular, Seattle was the fastest growing big-city in America between 2021 and 2022[2]. Albuquerque is also growing relatively quickly, being the sixth-fastest growing mid-size city since 2000[3].

Strengths: A strength of our model is that the linear regression is able to take advantage of many more data points since we are using a dataset that includes monthly changes in housing permits (12 times as many data points). In addition, our model is extremely simple and easily understood. For example, the Seattle model's slope (439.328) means that we expect about 439 new housing units per month. Similarly, the Albuquerque model expects about 139 new housing units per month.

Weaknesses: Some weaknesses of this model are that we are extrapolating housing units from private housing permits. This might not be completely accurate, especially because the relationship might change depending on public policy. In addition, the 50 year prediction is probably not very accurate as the data is only from the last 10 years, so it will probably not extrapolate very well to 50 years.

1.8 Sensitivity Analysis

We performed a sensitivity analysis by jittering our data and recomputing the 10, 20, and 50 year predictions. We jittered the data by adding a small amount of Gaussian noise (standard deviation 5%) and reran it 5 times, taking the average absolute deviation from our original estimate. We found that the 10 year jitter error was 8.8%, 20 year was 12.2% and 50 year was 18.9%. Intuitively, we would expect the error to increase as years go on as this is one of linear regressions weaknesses (errors compound together). Therefore, the Seattle model is somewhat sensitive to small changes in noise. For the Albuquerque data, we actually found that it was extremely resistant to noise. This may be due to the general stability in the Albuquerque housing unit prediction as its R^2 was significantly higher than that of Seattle's in the first place. In particular, the 10 year jitter error was 1.1%, the 20 year was 1.5% and the 50 year was 2.8%.

Question 2: It Was the Worst of Times

2.1 Defining the Problem

The problem asks us to predict the changes in the homeless population in the US cities of Seattle, Washington, and Albuquerque, New Mexico, for the next 10, 20, and 50 years.

2.2 Assumptions

2-1 **There will not be a major world event impacting cost of living or housing affordability in the next 50 years.**

- Justification: The housing market is not easy to predict and it heavily influences the housing supply. For example, if there are a larger number of houses on the market, there will probably be few houses built. However, to keep this model simple, we decided to ignore these considerations.

2-2 **Recent data in homeless populations accurately represents general trends in homelessness.**

- Justification: The model assumes that we can use data from the past 10-15 years to estimate homeless populations for the next 50 years. We made this assumption in order to improve data quality, and we believe that it won't significantly affect our model due to consistency in the homeless population over time.

2.3 Variables

For the purposes of this table, each symbol will be subscripted for the cities. A subscript of "SEA" corresponds to Seattle and a subscript of "ALB" corresponds to Albuquerque.

Symbol	Definition	Units
U_{SEA}	Annual unsheltered homeless population in Seattle	Persons (population in a given year)
T_{SEA}	Annual transitional housing homeless population in Seattle	Persons (population in a given year)
E_{SEA}	Annual emergency sheltered homeless population in	Persons (population in a given year)

	Seattle	
H_{SEA}	Annual homeless population in Seattle	Persons (population in a given year)
U_{ALB}	Annual unsheltered homeless population in Albuquerque	Persons (population in a given year)
T_{ALB}	Annual transitional housing homeless population in Albuquerque	Persons (population in a given year)
E_{ALB}	Annual emergency sheltered homeless population in Albuquerque	Persons (population in a given year)
H_{SEA}	Annual homeless population in Albuquerque	Persons (population in a given year)

Table 4: Variable definitions for Problem 2

2.4 Model

2.4.1 Developing the Model

We used data provided by the US Dept. of Housing and Urban Development[1], which contain Point-in-Time (PIT) estimates of homelessness by Continuum of Care (CoC) geographic service area. We sampled annual data points for the variables described in Table 4. We also decided to drop data points before 2011 and after 2019. We believed that homeless populations during these periods did not accurately represent homeless population trends over time, as they were significantly impacted by world events (the 2007-2008 Financial Crisis and the COVID-19 pandemic). As we have no way of predicting these major world events without significant unrelated additional data (consistent with assumption 1-3), we dropped these data points to help our model predict generalized trends.

Our preliminary exploratory data analysis revealed two main conclusions:

1. Time series for the general homeless population exhibits randomness, and would not result in accurate estimates of the future homeless population. Thus, we moved to analyzing the components that make up the general homeless population: unsheltered homeless, transitional housing, and emergency sheltered populations. Our goal was to

predict future data for each of these components, and then sum them up together to obtain a cumulative homeless population.

2. Linear regression models could occasionally capture trends in these components, but we found that they would often predict clearly erroneous homeless populations (e.g. negative or upwards of 50% of the total city population), and were not able to capture more cyclic trends that we observed in a few variables.

Thus, we wanted to find a model that would account for cyclical patterns in homeless populations over time. We noticed that by using a cyclical model, we could accurately predict all variables involved in the total homeless population, and we could thus predict the total homeless population over time. Our rationale was both quantitative and qualitative: we observed wave-like patterns in the data that we plotted, and we also hypothesized that homeless populations would follow the fluctuations that are innate within the economy.

To address these challenges and to accurately model the cyclical patterns observed in the homeless populations, we turned to the ARIMA (AutoRegressive Integrated Moving Average) model[4]. The ARIMA model is particularly suited for time series data that shows evidence of non-stationarity, where data show trends or seasonality. The 'AR' part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., past) values. The 'I' (for 'integrated') part indicates that the data values have been replaced with the difference between their values and the previous values in order to make the series stationary. Finally, the 'MA' part involves modeling the error term as a linear combination of error terms that occurred contemporaneously and at various times in the past.

We chose the ARIMA model for each component of the homeless population (unsheltered, transitional housing, and emergency sheltered) for Seattle and Albuquerque because it allows us to incorporate both the trend and cyclicality in the data. This is crucial for capturing the inherent fluctuations in homeless populations that are influenced by economic cycles, policy changes, and other external factors.

2.4.2 Executing the Model

To execute the ARIMA model, we first performed a stationarity test on each time series to determine the degree of differencing (d) needed to make the series stationary. We then used plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF) to help identify the order of the AR (p) and MA (q) components. We repeated these tests to determine different p, d, and q values for each city, as the homeless populations depend on different factors based on region; we also adjusted the degree of differencing based on our observations of stationarity in the data. Our final parameters are summarized in the table below.

ARIMA Model for Components of Homeless Population	AR (p)	Degree of Differencing (d)	MA (q)
Unsheltered Seattle	4	0	0
Transitional Housing Seattle	3	0	0
Emergency Sheltered Seattle	5	0	0
Unsheltered Albuquerque	5	0	0
Transitional Housing Albuquerque	5	1	0
Emergency Sheltered Albuquerque	5	1	0

Table 5: ARIMA Parameters by City and Model

Finally, to predict the total homeless population, we summed the forecasts for the unsheltered, transitional housing, and emergency sheltered components. This approach provided us with a comprehensive view of the future homeless population in Seattle and Albuquerque, taking into account the cyclicality and trends observed in the historical data.

2.5 Results

Graphs of the three homeless population components and the total homeless population are shown below. The blue lines represent the data from our dataset, and the orange lines represent the ARIMA model's predicted data for the next 50 years. The x-axes represent the year numbers; the y-axes represent the homeless populations (and their components); and the blue points represent the predicted results in 10, 20, and 50 years.

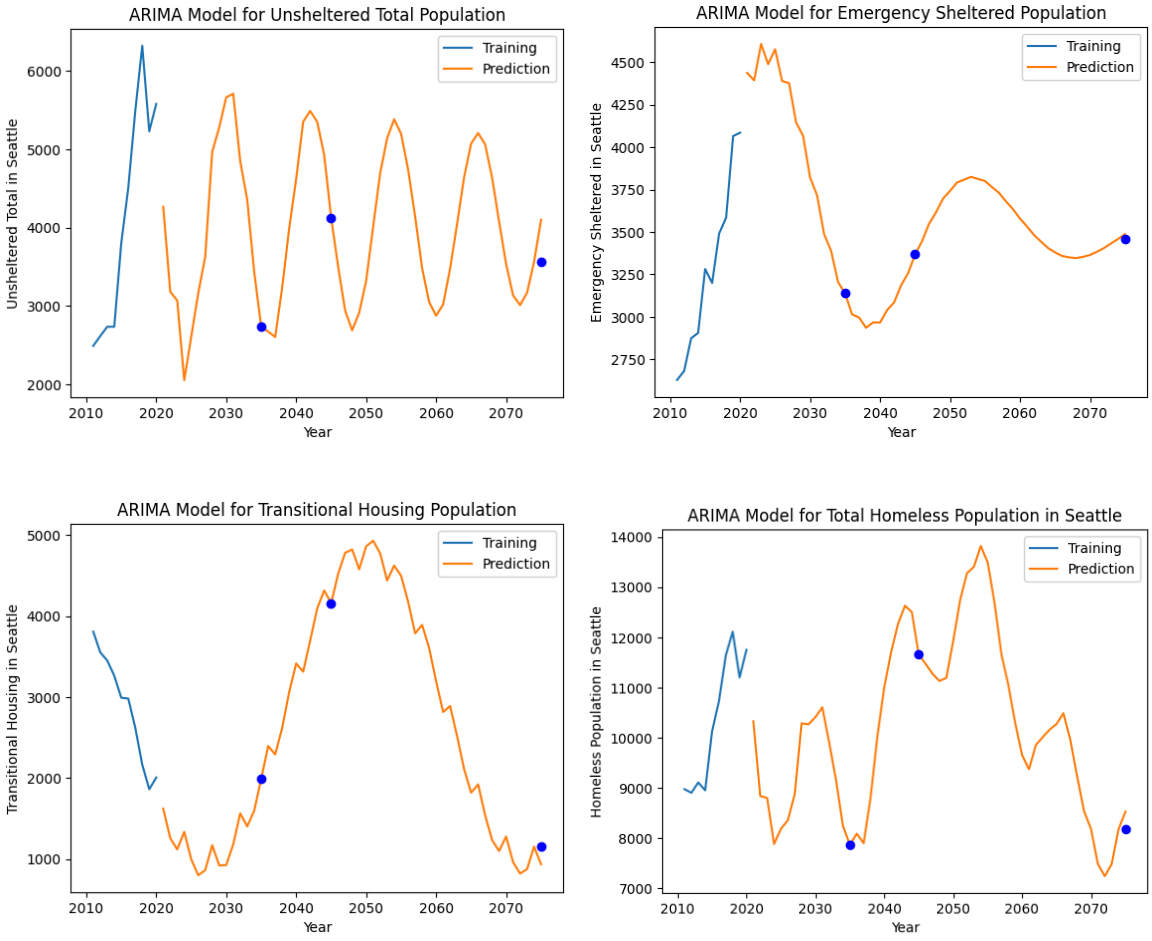

Figure 3: US Housing and ARIMA Prediction Data for Seattle, By Year. Top Left is U_{SEA}; Top Right is E_{SEA}; Bottom Left is T_{SEA}; Bottom Right is H_{SEA}. Note that H_{SEA} is the sum of all other variables, not a separate ARIMA model.

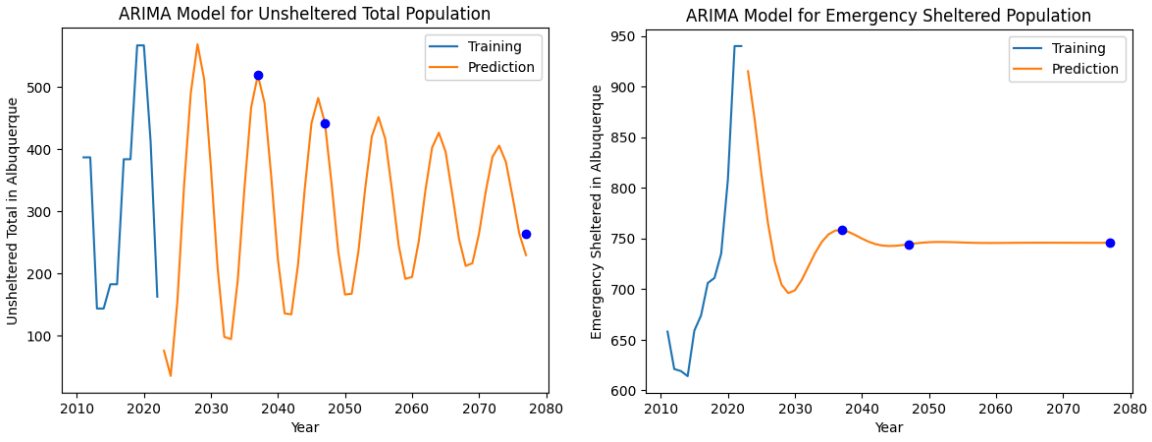

Figure 4: US Housing and ARIMA Prediction Data for Albuquerque, By Year. Top Left is U_{ALB}; Top Right is E_{ALB}; Bottom Left is T_{ALB}; Bottom Right is H_{ALB}. Note that H_{ALB} is the sum of all other variables, not a separate ARIMA model.

# Years in Future	Seattle Homeless Population	Albuquerque Homeless Population
10	10603	1188
20	11707	1249
50	8180	1400

Table 6: Forecasted Homeless Populations for Seattle and Albuquerque

2.6 Discussion

We predict that the homeless population of Seattle will be 10603, 11707, and 8180 persons after time periods of 10, 20, and 50 years respectively, from 2024. We predict that the homeless population of Albuquerque will be 1188, 1249, and 1400 persons after time periods of 10, 20, and 50 years respectively, from 2024.

Strengths: A strength of our model is that the ARIMA model is able to take advantage of yearly changes in homelessness that are potentially cyclical. For example, cities could create policies that decrease homelessness in one year and these results could lead to other policies that potentially increase homelessness. Although the general homeless population doesn't follow a clear trend, we were able to find clear cyclical trends by breaking the population down into

segments. This ensemble method—combining 3 different models of homeless populations—also increases model robustness.

Weaknesses: A weakness of this model is that there was very little initial starting data. Therefore, our model is very sensitive to small fluctuations in the initial data. For example, the 2008 financial crisis led to a large change in the homeless population because of the economic downturn.

Question 3: Rising from This Abyss

3.1 Defining the Problem

For the third problem, we created a model to help a city determine a long-term plan for homelessness, focusing on Seattle. To do this, we studied the most important factors that are controllable and correlate to homelessness, and researched how adjusting those factors lowers the homeless population.

3.2 Variables

Variable	Description	Units
Housing Prices	Average price of housing	Dollars
CPI	Consumer Price Index (a common measure of inflation)	Unitless
Income	Median household income (inflation adjusted)	Dollars
Housing Units	Number of available housing units	Units

Table 7: Variables for Problem 3

3.3 The Model

We determined the factors that correlate with homelessness using two models. The first model, the correlation model, determined the correlation between each factor and the size of the homeless population. The second model, the random forest model, determined the relative order of importance between these different factors.

3.3.1 Developing the Correlation Model

First, we looked at population statistics which potentially have correlations with homelessness. We studied the Consumer Price Index (CPI)[5], the number of housing units, house prices, population, and income/inflation.

From the trendlines, we determined the positive correlations between CPI, the number of housing units, house prices, population, and income/inflation. However, because all the R^2 values were about equal, we decided to use a random forest approach to rank the importance of the variables, allowing us to best inform policymakers on what to tackle to decrease homelessness.

3.3.2 Developing the Random Forest Model

The random forest model considers four of the aforementioned factors—CPI, housing prices, inflation-adjusted income, and number of housing units—to predict the percentage of the Seattle population that will be homeless over the span of the next 50 years.

First, models were used to predict values for CPI, housing prices, inflation-adjusted income, number of housing units, total population size, and homeless population size. Linear regression models were developed for CPI, housing prices, inflation-adjusted income, and total population size with r-squared values of 0.93, 0.89, 0.93, and 0.97 respectively. The previously developed models from parts 1 and 2 were used to predict the number of housing units and the size of the homeless population.

Second, the size of the total population and the size of the homeless population were synthesized into a single value—the percentage of the population that is homeless. The synthesis was performed to reduce the number of input and output variables in the random forest model, simplifying the modeling task and improving the performance of the random forest model.

Third, for each metric—CPI, housing prices, inflation-adjusted income, number of housing units, and homeless percentage of total population—the predicted data points across the next 50 years were scaled between 0 and 1 to simplify the modeling task for the random forest model.

Fourth, the random forest model was trained to predict the homeless percentage of the population based on input values of CPI, housing prices, inflation-adjusted income, and number of housing units. Importantly, for each of these four input features, the random forest creates a feature importance value throughout the training process. These importance values reflect the importance of each factor in contributing to the size of the homeless population.

3.4 Results

Our graphs and trendlines for Seattle are below:

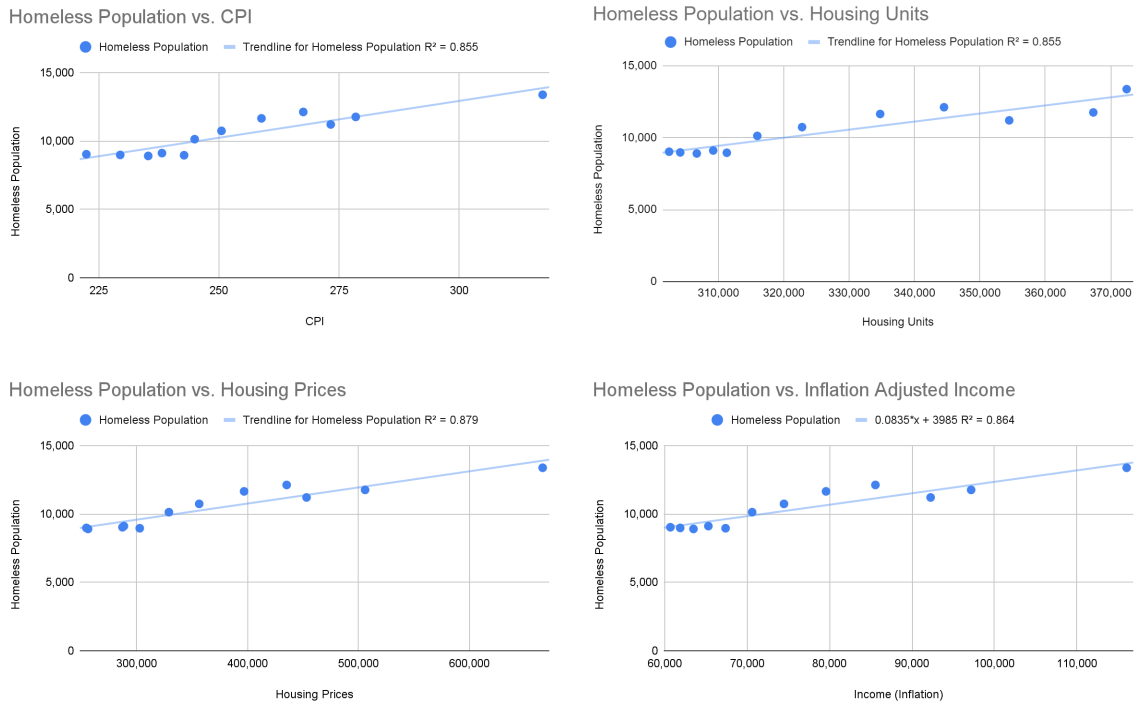

Figure 5: Trendlines to show correlation for Homeless Population vs. CPI (top left), Housing Units (top right), Housing Prices (bottom left), and Inflation-Adjusted Income (bottom right)

Factor	Importance (unitless)
Housing Prices	0.316
CPI	0.297
Income (Inflation)	0.194
Housing Units	0.193

Table 8: Factor Importance

3.5 Discussion

As we discovered from performing linear regression on multiple variables, homelessness is a complex problem which does not have just one cause. A better model would take into account more variables and come up with a more nuanced solution.

We found that increasing housing prices increases homelessness which makes sense as housing costing more would make it harder for homeless people to afford homes. For lawmakers, this may mean that more affordable housing should be available to help fight homelessness.

For CPI, which is a measure of inflation, we found that it is positively correlated with homelessness. This also makes sense because inflation creates an increased risk for homelessness as people have less spending power. For lawmakers, they could encourage the Federal Reserve to increase interest rates, which would help to cool inflation.

It was interesting that an increase in inflation-adjusted income is positively correlated to homelessness. This could possibly be attributed to the wealth gap; in particular, though the median amount of income may be increasing, the amount of income at the **bottom** may not be increasing. Therefore, there may still be more homeless people even if the median amount of income increases. For lawmakers, we advise that they attempt to decrease the wealth gap, potentially by expanding on wealth taxes and using that income on welfare and other policies to combat homelessness.

Finally, it was also quite shocking to find that increasing housing units would increase the rate of homelessness. This might be due to the fact that many housing units built are not affordable housing and are rather for the wealthy. Therefore, like in the inflation-adjusted income, there is little correlation between the two and that is why we found that it had little effect in our random forest model. For lawmakers, we would similarly encourage affordable housing or an increase in shelters instead of just awarding building contracts for housing units in general.

Given a natural disaster or economic recession, our models are probably not very adaptable in predicting homelessness because they are almost all linear which are not able to predict outliers very well. On the other hand, we still believe that the order of importance as generated in our random forest is robust to these disasters. This is because CPI and Housing Prices would still go down in a natural disaster and we believe that their relationship with homelessness would stay approximately the same.

Strengths: This model takes into account multiple factors such as CPI, Housing Prices, Median Income, and Housing Units. Therefore, it is more resistant to changes in any one of them.

Weaknesses: The model assumes linear trends for calculating the positive or negative correlation. In addition, the random forest data was extrapolated from linear trends which may not be true.

Conclusion

Ultimately, housing prices were the most important factor when evaluated against CPI, income accounting for inflation, and housing units. Consequently, a good first step when creating policies to lessen the effects of homelessness is lowering the price of the average home. When living costs continue to skyrocket, it becomes increasingly difficult for homeless individuals to bounce back. A second step is reducing the wage gap, the phenomenon which is likely

responsible for increases in income which drive increases in homelessness. Without intervention, we predict that the homeless population of Seattle will increase to 11707 persons by 2044, and the homeless population of Albuquerque will increase to 1400 persons by 2074. So, we implore the U.S. Department of Housing and Urban Development to take action against homelessness and acknowledge the severity of the housing crisis through our results.

References

1. A Tale of Two Crises, MathWorks Math Modeling Challenge 2024, curated data, https://m3challenge.siam.org/kdfrldh/.
2. https://www.seattletimes.com/seattle-news/data/seattle-is-once-again-the-fastest-growing-big-city-census-data-shows/
3. https://www.bizjournals.com/albuquerque/news/2014/12/04/albuquerque-sixth-fastest-growing-mid-size-city-in.html
4. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
5. https://www.seattle.gov/economic-and-revenue-forecasts/consumer-price-index-inflation/historical-data-and-forecasts
6. https://www.census.gov/library/visualizations/interactive/bps-new-privately-owned-housing-unit-authorizations.html
7. https://www.huduser.gov/portal/pdrdatas_landing.html
8. https://www.usgs.gov/regions/rocky-mountain/data
9. https://www.huduser.gov/portal/datasets/hpmd.html?q=datasets%2Fhpmd.html
10. https://www.cabq.gov/gis/geographic-information-systems-data
11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10574586
12. https://zenodo.org/records/8361378
13. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
14. https://fred.stlouisfed.org/series/PERMIT

Code Appendix

Question 1: It Was the Best of Times

```python
import pandas as pd
import matplotlib.pyplot as plt
import math
import numpy as np
from sklearn.linear_model import LinearRegression

# convert strings to numbers
def string_to_num(str):
  try:
    vals = str.split(',')
    output = ''
    for v in vals:
      output += v
    return int(output)
  except AttributeError:
    return str

# do the linear regression and print out r squared, coefficient, and intercept
def linear_fit(y):
  x = [[i] for i in range(len(y))]
  model = LinearRegression()
  model.fit(x, y)
  r_sq = model.score(x, y)
  print(r_sq)
  print(model.coef_)
  print(model.intercept_)
  return model

# Clean our data by looking only at the survey date (months) and total units
# We also truncate until after 2010
def census_survey(df):
  dates = []
  years = []
  units = []
  for i in range(len(df)):
    date = int(df.at[i, "Survey Date"])
    year = int(date / 100)
    if year >= 2010:
      dates.append(date)
      years.append(year)
      units.append(string_to_num(df.at[i, 'Total units']))
  totals = []
  for i in range(len(units)):
    if i == 0:
      totals.append(units[i])
    else:
      totals.append(totals[i-1]+units[i])

    return totals


# CODE FOR SEATTLE
# Seattle without cutting off years
df_sea = pd.read_csv('/content/Seattle_Survey_Final.csv')
```

```
# Seattle after cutting off years and totaling (cumulative sum)
totals_sea = census_survey(df_sea)

# Manually calculated multiplier and baseilne
multiplier_sea = 0.6963397158
baseline_sea = 302465
totals_sea = [t*multiplier_sea+baseline_sea for t in totals_sea]
model_sea = linear_fit(totals_sea)

sea_X_prediction = np.arange(172, 773).reshape(-1, 1)
sea_Y_prediction = model_sea.predict(sea_X_prediction)

# Original line
plt.plot(sea_X_prediction, sea_Y_prediction, color="red", linewidth=1)
sea_X_original = np.arange(0, 172).reshape(-1, 1)
sea_Y_original = df_sea['Total units']

# Our training data
plt.plot(totals_sea)

# Add in our labels and plotting
plt.plot(291, sea_Y_prediction[291 - 171], 'bo')
plt.annotate("(291, " + str(round(sea_Y_prediction[291 - 171])) + ")", (291 - 200, sea_Y_prediction[291 - 171]))
plt.plot(411, sea_Y_prediction[411 - 171], 'bo')
plt.annotate("(411, " + str(round(sea_Y_prediction[411 - 171])) + ")", (411 - 200, sea_Y_prediction[411 - 171]))
plt.plot(771, sea_Y_prediction[771 - 171], 'bo')
plt.annotate("(771, " + str(round(sea_Y_prediction[771 - 171])) + ")", (771 - 200, sea_Y_prediction[771 - 171]))
plt.xticks()
plt.xlabel("Months since 2010")
plt.yticks()
plt.ylabel("Number of Housing Units")
plt.legend(['Predicted', 'Training'])
plt.title("Projected Number of Housing Units in Seattle (monthly)")
plt.show()

# Jittering for Seattle
jittered_variation_10 = []
jittered_variation_20 = []
jittered_variation_50 = []
year_10_expected = 439932
year_20_expected = 499491
year_50_expected = 678169
for i in range(5):
  totals_sea_jitter = np.zeros(np.shape(totals_sea))
  for j in range(len(totals_sea)):
    eta = np.random.normal(scale=totals_sea[j] * .05)
    totals_sea_jitter[j] = totals_sea[j] + eta

  model_sea = linear_fit(totals_sea_jitter)
  sea_X_prediction = np.arange(172, 773).reshape(-1, 1)
  sea_Y_prediction = model_sea.predict(sea_X_prediction)

  jittered_variation_10.append(abs((year_10_expected - sea_Y_prediction[291 - 171]) / (year_10_expected)))
  jittered_variation_20.append(abs((year_20_expected - sea_Y_prediction[411 - 171]) / (year_20_expected)))
```

~/Desktop/School23-24/M3/buildingpermits

```
  alb_X_prediction = np.arange(172, 773).reshape(-1, 1)
  alb_Y_prediction = model_alb.predict(alb_X_prediction)

  jittered_variation_10.append(abs((year_10_expected - alb_Y_prediction[291 - 171]) / (year_10_expected)))
  jittered_variation_20.append(abs((year_20_expected - alb_Y_prediction[411 - 171]) / (year_20_expected)))
  jittered_variation_50.append(abs((year_50_expected - alb_Y_prediction[771 - 171]) / (year_50_expected)))

print("10 years: " + str(np.mean(jittered_variation_10)))
print("20 years: " + str(np.mean(jittered_variation_20)))
print("50 years: " + str(np.mean(jittered_variation_50)))
```

Question 2: It Was the Worst of Times

```
import math
import numpy as np
from sklearn.linear_model import LinearRegression
import pandas as pd

def process_years(df):
  # CompletedDate, StatusCurrent, HousingUnitsRemoved, HousingUnitsAdded

  # loop through years (rows)
  for year in range(0, len(df)):
    # loop through columns
    for data in range(0, len(df.columns)):
      # if the value is a string
      if type(df.iloc[year, data]) == str:
        # remove commas
        df.iloc[year, data] = df.iloc[year, data].replace(',', '')
        # convert to int
        df.iloc[year, data] = int(df.iloc[year, data])

  return df

def get_column(df, column):
  totals = df[column].tolist()
  diffs = []
  for i in range(1, len(totals)):
    diffs.append(totals[i] - totals[i-1])
  return totals, diffs

df = pd.read_csv('/content/albuquerque_trimmed.csv')
df = df.dropna()
processed_df = process_years(df)

totalsUS, diffsUS = get_column(processed_df, 'Unsheltered Total')
totalsES, diffsES = get_column(processed_df, 'Emergency Sheltered')
totalsTH, diffsTH = get_column(processed_df, 'Transitional Housing')

# plt.plot(totals)
# plt.plot(diffs)
# plt.show()
```

```python
from statsmodels.tsa.arima.model import ARIMA
import matplotlib.pyplot as plt

allTotals = [(totalsUS, 'Unsheltered Total', (4,0,0)), (totalsES, 'Emergency Sheltered', (3,0,0)), (totalsTH, 'Transitional Housing', (5,0,0))]

years = 55

total10 = 0
total20 = 0
total50 = 0

summed_forecast = None

for ind, (totals, name, order) in enumerate(allTotals):
    future_model = ARIMA(totals, order=order)
    future_model_fitted = future_model.fit()
    fifty_year_forecast = future_model_fitted.forecast(years)

    plt.plot(np.arange(2011, 2023), totals, label='Training')
    plt.title(f'ARIMA Model for {name} Population')
    plt.xlabel('Year')
    plt.ylabel(f'{name} in Albuquerque')
    plt.plot(np.arange(2011 + len(totals), 2011 + len(totals) + years), fifty_year_forecast, label='Prediction')
    pt1 = (10 + len(totals) + 4, fifty_year_forecast[14])
    plt.plot(pt1[0] + 2011, pt1[1], 'bo')
    print(f"Population in 10 years: {pt1[1]}")
    total10 += pt1[1]

    pt2 = (20 + len(totals) + 4, fifty_year_forecast[24])
    plt.plot(pt2[0] + 2011, pt2[1], 'bo')
    print(f"Population in 20 years: {pt2[1]}")
    total20 += pt2[1]

    pt3 = (50 + len(totals) + 4, fifty_year_forecast[53])
    plt.plot(pt3[0] + 2011, pt3[1], 'bo')
    print(f"Population in 30 years: {pt3[1]}")
    total50 += pt3[1]

    plt.legend(loc='upper right')
    plt.show()

    if ind == 0:
        summed_forecast = fifty_year_forecast
    else:
        summed_forecast += fifty_year_forecast

totalsAll, diffsAll = get_column(processed_df, 'Homeless Total')

print(totalsAll)

plt.plot(np.arange(2011, 2023), totalsAll, label='Training')
plt.plot(np.arange(2011 + len(totalsAll), 2011 + len(totalsAll) + years), summed_forecast, label='Prediction')
pt3 = (10 + len(totals) + 4, summed_forecast[14])
pt3 = (20 + len(totals) + 4, summed_forecast[24])
pt3 = (50 + len(totals) + 4, summed_forecast[53])
plt.plot(pt1[0] + 2011, total10, 'bo')
plt.plot(pt2[0] + 2011, total20, 'bo')
plt.plot(pt3[0] + 2011, total50, 'bo')
plt.title(f'ARIMA Model for Total Homeless Population in Albuquerque')
plt.xlabel('Year')
plt.ylabel(f'Homeless Population in Albuquerque')
plt.legend(loc='upper right')
plt.show()

print("Total homeless population in 10 years: " + str(int(total10)))
print("Total homeless population in 20 years: " + str(int(total20)))
print("Total homeless population in 50 years: " + str(int(total50)))
```

Question 3: Rising from This Abyss

```
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
import matplotlib.pyplot as plt

def string_to_num(str):
  try:
    vals = str.split(',')
    output = ''
    for v in vals:
      output += v
    return int(output)
  except AttributeError:
    return str

def linear_fit(metric):
    x = list(df['Year'])
    y = list(df[metric])
    x = [[v] for v in x]
    y = [string_to_num(v) for v in y]

    model = LinearRegression()
    model.fit(x, y)
    r_sq = model.score(x, y)
    print(r_sq)

    return model

def transform_datapoints(datapoints):
    output = []
    for j in range(len(datapoints[0])):
        cur = []
        for i in range(len(datapoints)):
            cur.append(datapoints[i][j])
        output.append(cur)
    return output
```

```python
def scale(vals):
    mi = min(vals)
    ma = max(vals)
    return [(v-mi)/(ma-mi) for v in vals]

def create_datapoints(metrics, years, dataframe):
    # X
    datapoints = []
    for m in metrics:
        model = linear_fit(m)
        pred = model.predict(years)
        datapoints.append(scale(pred))
    datapoints = transform_datapoints(datapoints)

    model = linear_fit('Homeless Total')
    homeless = model.predict(years)
    model = linear_fit('Total Population')
    total = model.predict(years)
    y = [homeless[i]/total[i] for i in range(len(homeless))]
    y = scale(y)

    rf_model = RandomForestRegressor(random_state=1)
    rf_model.fit(datapoints, y)
    importances = rf_model.feature_importances_

    print('Random Forest')
    print(metrics)
    print(importances)


df = pd.read_csv('Data/M3 Q3 - Seatle.csv')
df_homeless = pd.read_csv('Data/homeless_forecast.csv')
#metrics = ['CPI', 'Housing Prices', 'Total Population', 'Income (Inflation)', 'Housing Units']
metrics = ['CPI', 'Housing Prices', 'Income (Inflation)', 'Housing Units']
years = [[2023+i] for i in range(50)]
create_datapoints(metrics, years, df_homeless)
```