Predicting COVID-19 Case Concentration in Massachusetts Municipalities

IE 7280 Statistical Methods in Engineering

Professor Nizar Zaarour

Project 1: Final Report

Group Number-3

Members: Noah Chicoine, Prerit Samria, Hirendra Tandekar

Executive Summary

In this project, we will study how certain demographics of towns in Massachusetts predict that the total COVID-19 case counts in that town throughout the pandemic. By understanding which factors are predictive of disease spreading in communities, extra preventative measures can be put in place in communities that are predicted to be "hotspots" in future disease outbreaks and pandemics. In this study, we randomly selected 50 municipalities from 361 towns/cities in Massachusetts to represent the entire state. For each municipality, we will use **population density** (number of people per square mile) and **median household income** ($) to predict the town's **total COVID-19 standardized case count** (number of cases per 1000 people) as of February 2, 2021. Summary statistics and basic visualizations of the three variables are provided in the following sections. The sample of 50 municipalities was created from a list of municipalities provided by Mass.gov [1] and a random number generator. The data of COVID-19 case concentration was provided my Mass.gov as well in a Weekly Covid-19 Public Health Report [2]. Lastly, each town's population, median household income, and land area were gathered from sources who obtained the respective data from the 2019 US Census ASC database and the 2010 US Census [3,4,5].

We run the simple linear regression model with each of the independent variables separately. Through the simple linear regression, we concluded that there is a positive linear relationship between population density and COVID-19 case concentration within Massachusetts towns. Also, there is no correlation between median household income and COVID-19 case concentration within Massachusetts towns. Since most of the towns have low population density, we decided to see if a logarithmic transformation of the population density gives a better fit with the dependent variable (COVID-19 case concentration), as the scatter plots of these two variables showed a possibly logarithmic relationship. Using the log transformed population density in the simple regression model, we achieved a stronger positive relation between COVID-19 case concentration and logarithmic population.

The multiple linear regression model using both the independent variables resulted in the conclusion that only the population density variable is significant. This model is no better than the SLR model with population density as the independent variable. After analyzing both the SLR Logarithmic Transformation and MLR Logarithmic Transformation models, we conclude that the regression model improved compared the previous SLR and MLR models without the log transformation. The R^2 values increased after the logarithmic transformation of population density, which shows each of the models captures more variation in the data. Also, the standard error was reduced significantly. The R^2 value improved for the log transformed MLR model, however, it did not improve as much as we would like to choose this MLR model over the SLR model with the log transformation. Hence, we can conclude that the SLR Logarithmic Transformation model is the simplest and most effective model at predicting COVID-19 concentration given our data.

Intuitively, this linear relationship between the population density and COVID-19 case concentration makes sense. Since COVID-19 is an air-borne disease, there should be more COVID-19 cases in towns that have a higher population density than in towns that have a lower population density. Further, the non-existence of relationship between the median household income and COVID-19 case concentration makes sense. COVID-19 is affecting everyone equally and does not discriminate on the basis of economic prosperity. The rich and the poor are affected equally.

Data

Our hypothesis is that certain demographics of towns will help to predict that town's total COVID-19 standardized case count (number of cases per 1000 people as of February 2, 2021). To examine this, we randomly selected 50 municipalities from 361 towns and/or cities in Massachusetts to represent the entire state (see Table 1). The independent quantitative variables to help predict the total COVID-19 cases were chosen to be population density (number of people per square mile) and median household income ($). Data for each of these variables were collected from online state and federal resources (see references 1-5).

Table 1: 50 randomly selected Municipalities in Massachusetts

Acushnet	Andover	Ashby	Avon	Bernardston
Beverly	Boxborough	Boxford	Carver	Chicopee
Dartmouth	Great Barrington	Greenfield	Halifax	Hamilton
Hardwick	Hawley	Hubbardston	Hull	Lenox
Leverett	Lexington	Leyden	Lynn	Lynnfield
Manchester-by-the-sea	Marlborough	Maynard	Millbury	Norfolk
Oakham	Oxford	Pittsfield	Provincetown	Quincy
Rowe	Royalston	Sandisfield	Somerville	Southwick
Sterling	Sudbury	Taunton	Tewksbury	Tolland
Warren	Wellesley	West Bridgewater	Westwood	Worthington

The descriptive statistics and basic plots of the variables, i.e., population density (say, X_1), median household income (say, X_2), and COVID-19 Cases per 1,000 People (Y) are shown in below.

Table 2: Descriptive statistics of the dependent variable and predictors.

Descriptive Statistics (n = 50)	COVID-19 Cases per 1,000 People (Y)	Population Density (X_1)	Median Household Income (X_2)
Minimum	7.04	11.97	46871.00
Q1	29.75	127.68	68491.50
Median	44.20	455.88	76754.00
Q3	68.62	1192.19	97144.00
Maximum	153.28	19263.33	176852.00
Mean	49.37	1215.79	87492.14
Standard Deviation	28.21	2859.53	32761.41

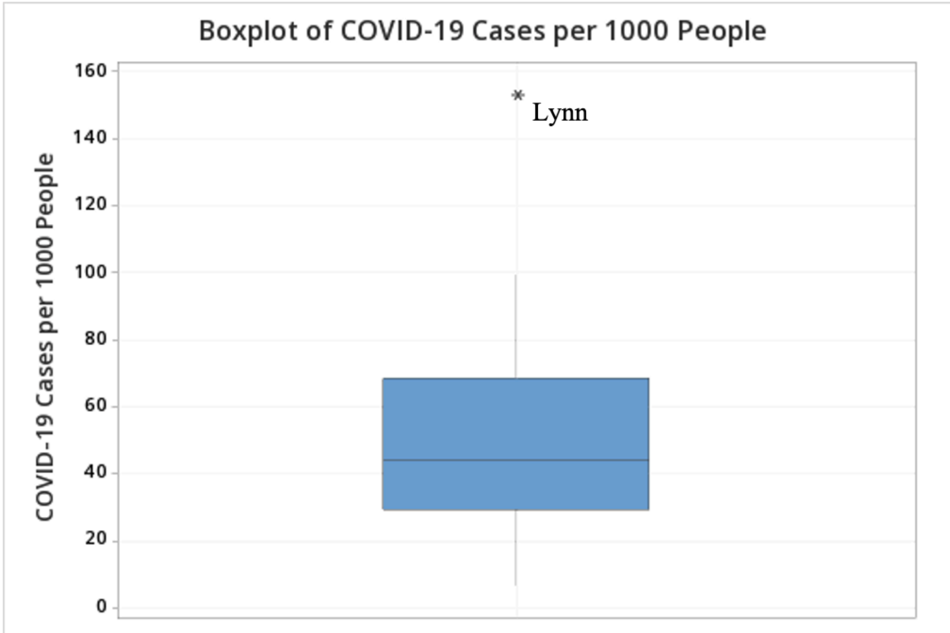

Figure 1: Boxplot of COVID-19 case concentration data (Y).

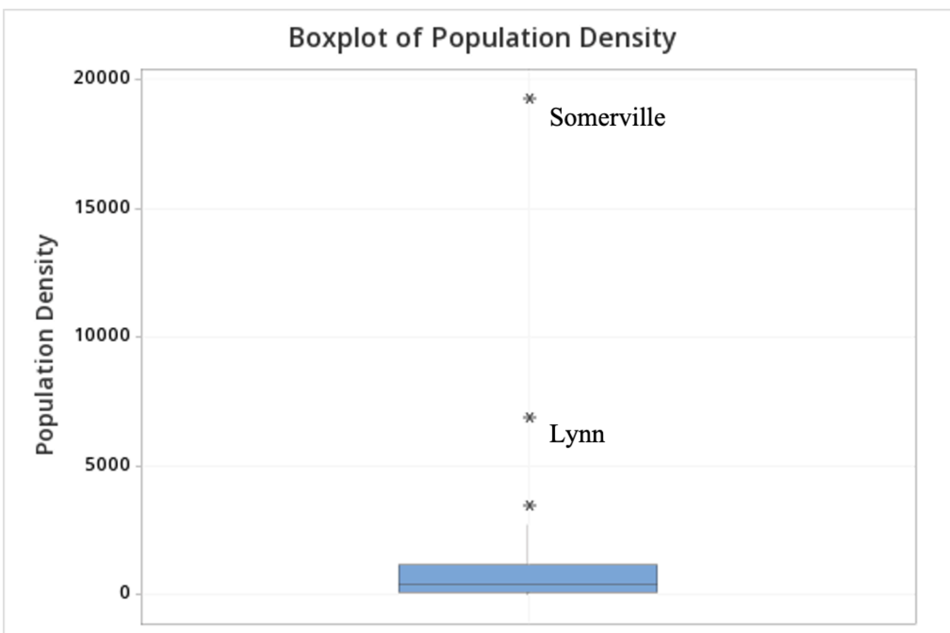

Figure 2: Boxplot of population density data (X_1).

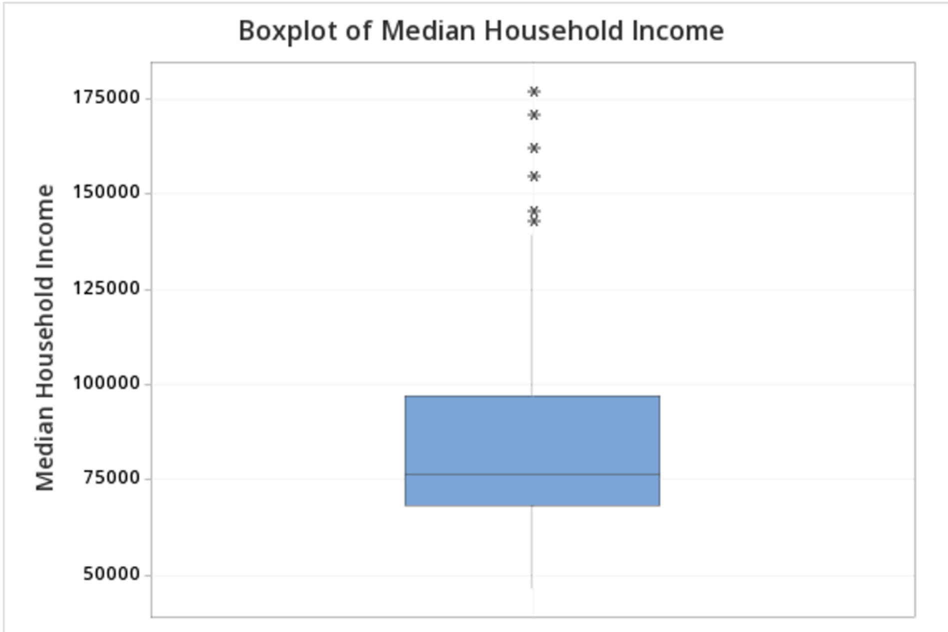

Figure 3: Boxplot of median household income (X_2).

Figure 2 shows that population density of two towns, Lynn and Somerville, are much higher than the rest of the sample. These towns are considered outliers to the data and hence are removed before proceeding further with the analysis. The resulting descriptive statistics and box plots are provided in the appendix.

Before conducting the regression analysis, we examined the correlation coefficients between the 3 variables to ensure none were highly correlated with one another. The table of correlation coefficients is shown below.

Table 3: Correlation coefficients of pair of variables (after Lynn and Somerville were removed from the dataset).

	Y	**X_1**	**X_2**
Y	-	0.4112	-0.0325
X_1	0.4112	-	0.2819
X_2	-0.0325	0.2819	-

From Table 3, we infer that none of the variables are highly correlated to each other. Hence, there should not be any multicollinearity between the variables when hypothesizing a multiple linear regression to predict the COVID-19 Cases per 1,000 People (Y) using both the independent variables (X_1 and X_2).

Simple Linear Regression (SLR) Models

We hypothesize that there is a simple linear relation between the dependent variable (Y) and each of the independent variables (X_1 and X_2). Simple linear regression analysis of COVID-19 Case concentration vs. Population density and vs. Median household income are conducted and discussed below.

1. **COVID-19 Case concentration (Y) vs. Population Density (X1)**

Table 4: Regression Analysis statistics of the COVID-19 case concentration data (Y) vs Population Density (X_1)

Regression Analysis	
Correlation coefficient	0.4112
R^2	0.1691
Adjusted R^2	0.1510
Standard error	22.4592
Observations	48

Table 5: ANOVA Table for Simple Linear Regression of COVID-19 case concentration data vs Population Density

	df	SS	MS	F	Significance F
Regression	1	4722.1013	4722.1014	9.3616	0.0037
Residual	46	23203.0470	504.4141		
Total	47	27925.1483			

Table 6: Simple Linear Regression of COVID-19 case concentration data vs Population Density: Values of Intercept and Slope of Line

	Intercept (b_0)	Slope (b_1, X_1)
Coefficients	38.1420	0.0125
Standard Error	4.3739	0.0041
t Stat	8.7204	3.0597
F Stat	-	9.3616
P-value	0.0000	0.0037
Lower 95%	29.3378	0.0043
Upper 95%	46.9461	0.0207

We can see from table 4 that the sample correlation coefficient is 0.41, which shows that there is some relationship between the two variables. Also, the coefficient of determination, R^2 is 0.17. This shows that about 17% of the total variability can be explained by this linear relationship. The standard error of the estimate is 22.46.

From table 6, we can see that the estimated standard deviation of the slope is 0.0041. Since we are assuming a significance level of 5%, and the P-value for the slope (0.0037) is less than the significance level, we have enough evidence to support the conclusion that there is a positive correlation between population density and COVID-19 case concentration within Massachusetts towns.

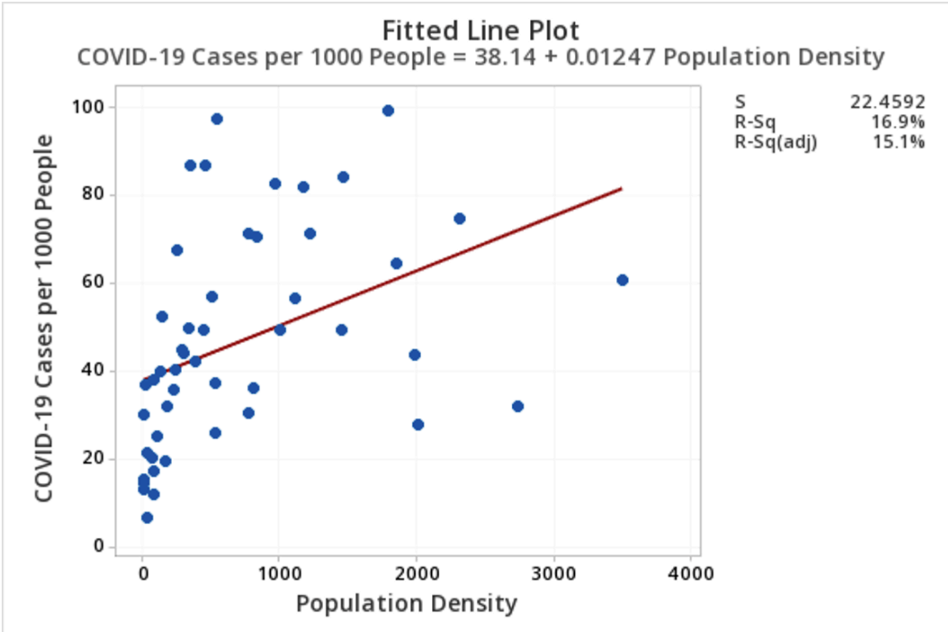

Figure 4: Fitted Line Plot of COVID-19 case concentration data (Y) vs Population Density (X_1)

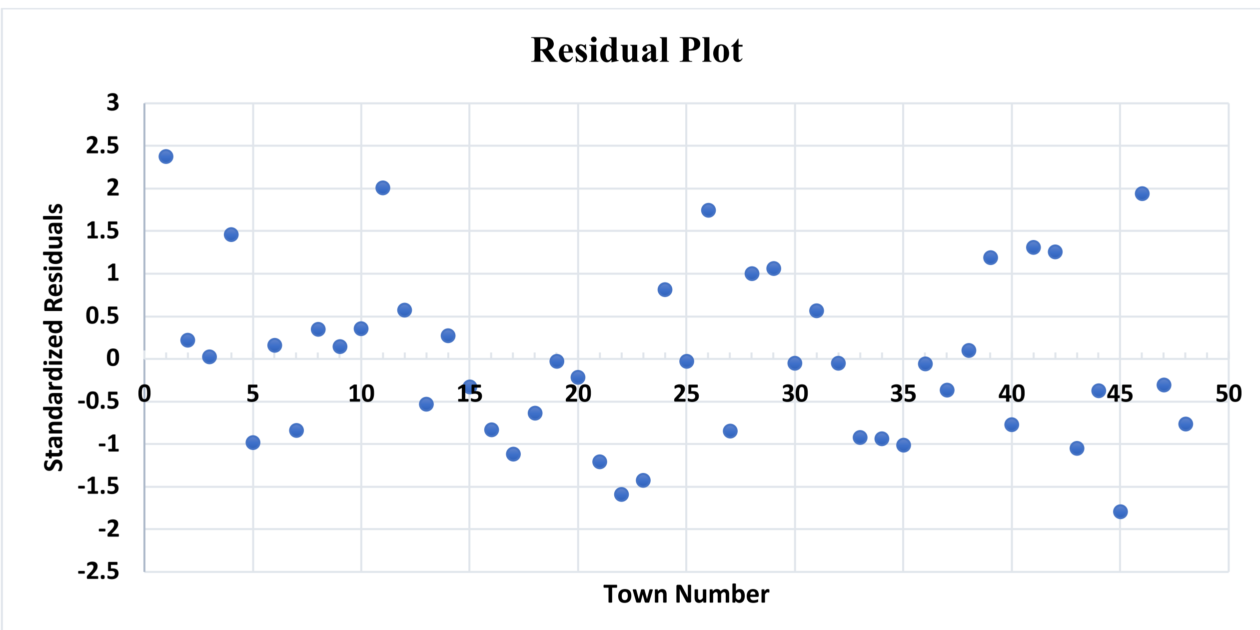

Figure 5: Standardized Residual Plot of SLR of COVID-19 case concentration data (Y) vs Population Density (X_1)

The fitted line can be seen in figure 4. We can see from figure 4 that there exists a linear relation between population density and COVID-19 case concentration. From figure 5, we can also see that the residuals are scattered around the X-axis and most of them have values near 0. The uniform distribution of the residuals suggests that a linear model may be a good model to capture the behavior of the data.

95% confidence and prediction intervals for $E(y_p)$ and y_p for a random town: (see Appendix Table A2)

Town = Boxborough

X_1 = 534.71 people per square mile, Actual Y = 26.25

Table 7: Simple Linear Regression of Y vs X_1: Confidence and Prediction Interval

Predicted Y	44.8098	
95% Confidence Interval	38.1090	51.5106
95% Prediction Interval	-0.8920	90.5117

2. COVID-19 Case concentration (Y) vs. Median Household Income (X_2)

Table 8: Regression Analysis statistics of the COVID-19 case concentration data (Y) vs Median Household Income (X_2)

Regression Analysis	
Correlation coefficient	0.0325
R^2	0.0010
Adjusted R^2	-0.0207
Standard error	24.6257
Observations	48

Table 9: ANOVA Table for Simple Linear Regression of COVID-19 case concentration data vs Median Household Income

	df	SS	MS	F	Significance F
Regression	1	29.4937	29.4937	0.0486	0.8264
Residual	46	27895.6550	606.4273		
Total	47	27925.1480			

Table 10: Simple Linear Regression of COVID-19 case concentration data vs Median Household Income: Values of Intercept and Slope of Line

	Intercept (b_0)	Slope (b_1)
Coefficients	49.2404	-2.39×10^{-5}
Standard Error	10.2241	0.0001
t Stat	4.8161	-0.2205
F Stat	-	0.0486
P-value	1.62×10^{-5}	0.8264
Lower 95%	28.6603	-0.0002
Upper 95%	69.8204	0.0002

We can see from table 8 that the sample correlation coefficient is 0.0325, which gives early signs that there might not be a relationship between the two variables. Also, the coefficient of determination, R^2 is very low (0.001). This shows that only 0.1% of the variability can be explained out of the total variability. The standard error of the estimate is 24.63.

From table 10, we can see that the estimated standard deviation of the slope is 0.0001. Since we are assuming a significance level of 5%, and the P-value for the slope (0.8264) is greater than the significance level, our data suggests that there is no correlation between median household income and COVID-19 case concentration within Massachusetts towns.

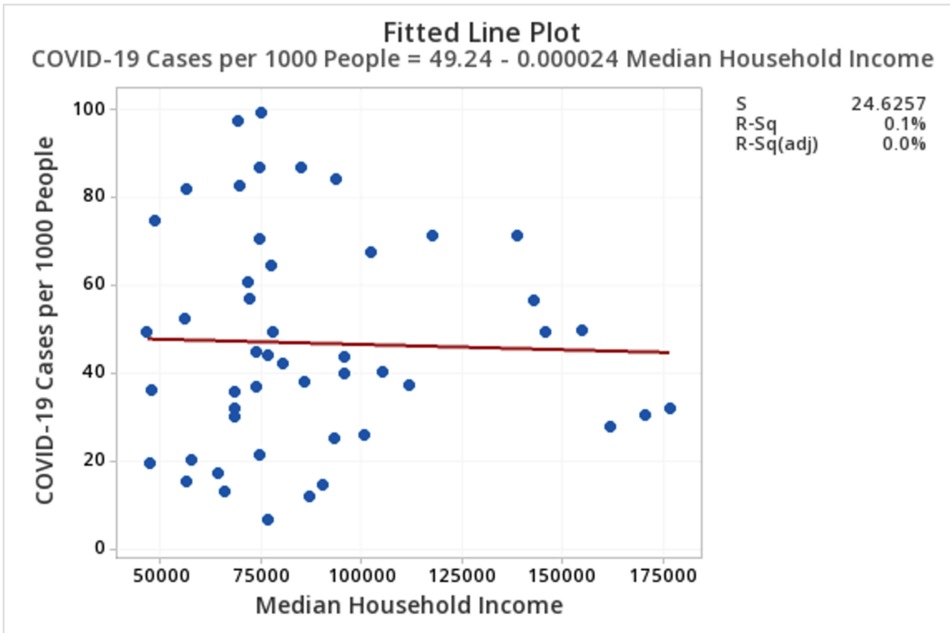

Figure 6: Fitted Line Plot of COVID-19 case concentration data (Y) vs Median Household Income (X_2)

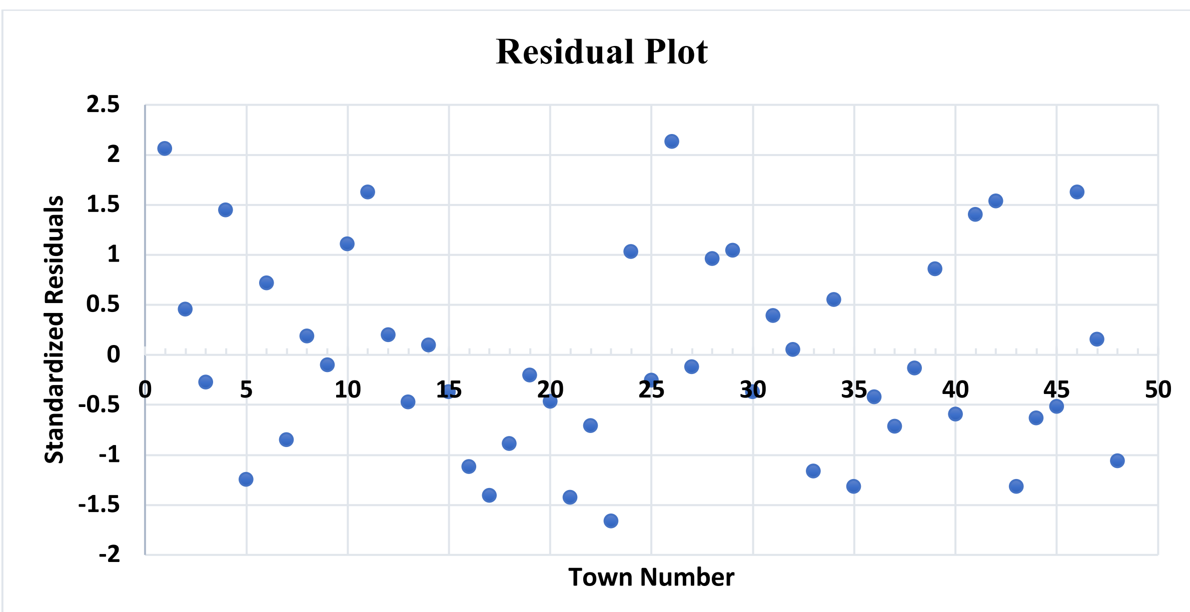

Figure 7: Standardized Residual Plot of SLR of COVID-19 case concentration data (Y) vs Median Household Income (X_2)

The fitted line can be seen in figure 6. We can see from figure 6 that the fitted line is almost parallel to the X-axis, which means that there is no relationship between the variables COVID-19 case concentration data (Y) and

Median Household Income (X_2). From figure 7, we can also see that the residuals are scattered around the X-axis and most of them do not have values near 0. This supports the conclusion that there is no relationship between the two variables.

95% confidence and prediction intervals for $E(y_p)$ and y_p for a random town: (see Appendix Table A3)

Town = Boxborough

X_2 = \$101077, Actual Y = 26.25

Table 11: Simple Linear Regression of Y vs X_2: Confidence and Prediction Interval

Predicted Y	46.8192	
95% Confidence Interval	39.1351	54.5032
95% Prediction Interval	-3.3419	96.9803

Conclusion of Simple Linear Regression

Through the simple linear regression, we conclude that there is a linear relationship between population density and COVID-19 case concentration within Massachusetts towns. Intuitively, this linear relationship between the two variables makes sense. Since COVID-19 is an air-borne disease, there should be more COVID-19 cases in towns that have a higher population density than in towns that have a lower population density. Thus, the slope of the fitted line is positive.

Also, we conclude that there is no correlation between median household income and COVID-19 case concentration within Massachusetts towns. Intuitively, the non-existence of relationship between the two variables makes sense. COVID-19 is affecting everyone equally and does not discriminate based on economic prosperity. The rich and the poor are getting affected equally. Thus, the slope of the fitted line is almost 0.

Although the results here are for Massachusetts, this can likely be extended most of the states in United States if not all. We obtained a positive correlation between population density and COVID-19 case concentration within Massachusetts towns.

Multiple Linear Regression (MLR)

COVID-19 Case concentration (Y) vs. Population Density (X_1) & Median Household Income (X_2)

We will hypothesize that there is a simple linear relation between the dependent variable (Y) and both the independent variables (X_1 & X_2). Then we will draw conclusions whether that relation is significant or not.

Table 12: Regression Analysis statistics of the COVID-19 case concentration data vs Population Density and Median Household Income

Regression Analysis	
Multiple R	0.4394
R^2	0.1930
Adjusted R^2	0.1572
Standard error	22.3780
Observations	48

Table 13: ANOVA Table for Simple Linear Regression of COVID-19 case concentration data vs Population Density and Median Household Income

	df	SS	MS	F	Significance F
Regression	2	5390.3698	2695.1849	5.3821	0.0080
Residual	45	22534.7790	500.77286		
Total	47	27925.1480			

Table 14: Simple Linear Regression of COVID-19 case concentration data vs Population Density and Median Household Income: Values of Intercept and Slope of Line

	Intercept (b_0)	Slope (b_1, X_1)	Slope (b_2, X_2)
Coefficients	47.6358	0.0139	-0.0001
Standard Error	9.3038	0.0042	0.0001
t Stat	5.1203	3.2719	-1.1152
P-value	0.0000	0.0021	0.2541
Lower 95%	28.8989	0.0053	-0.0003
Upper 95%	66.3763	0.0224	0.0000

We can see from table 12 that the sample correlation coefficient is 0.43, which suggests a moderate correlation between the two variables. Also, the coefficient of determination, R^2 is 0.1930. This shows that about 19.3% of the total variability can be explained by this linear relationship. The standard error of the estimate is 22.37. The P-value for the slope (X_1) is 0.0021 while P-value for the slope (X_2) is 0.2541.

Case 1: Significance level of 10%

We can see that the P-value for the slope of X_1 is less than the significance level and the P-value for the slope of X_2 is much higher than the significance level. Thus, we can drop the independent variable Median Household Income since it's insignificant. Hence, we have enough evidence to support the conclusion that there is a positive correlation between population density and COVID-19 case concentration within Massachusetts towns and there is no correlation between median household income and COVID-19 case concentration within Massachusetts towns.

Case 2: Significance level of 5%

The result of case 1 holds true for a significance level of 5%.

Case 3: Significance level of 1%

The result of case 1 holds true for a significance level of 1%.

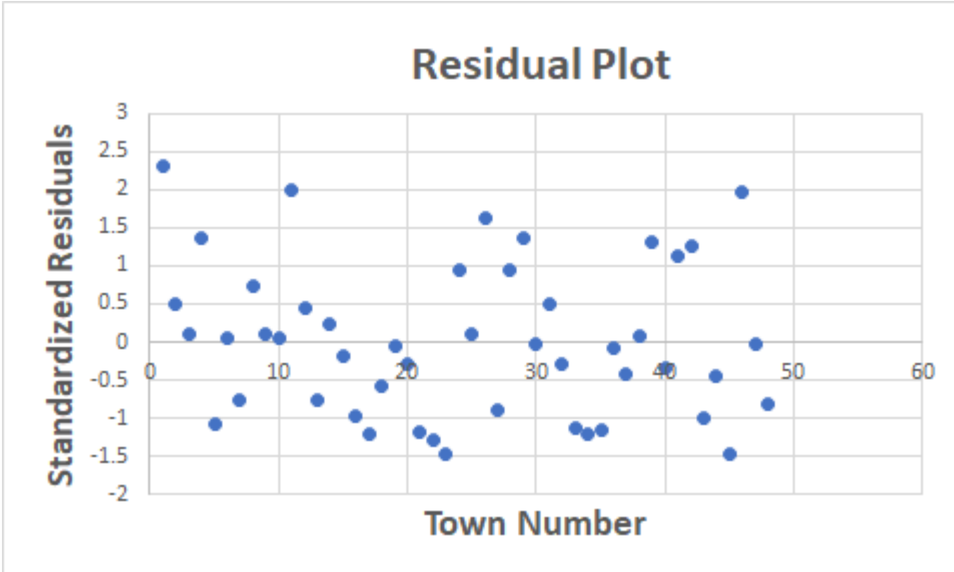

Figure 8: Standardized Residual Plot of MLR of COVID-19 case concentration data (Y) vs Population Density (X_1) & Median Household Income (X_2)

The regression analysis indicates that there is a positive relationship between variables COVID-19 case concentration data (Y) and Population density (X_1) and a negative relationship between variables COVID-19 case concentration data (Y) and Median Household Income (X_2). However, the relationship between COVID-19 case concentration and median household income is insignificant ($p = 0.25$) when both are used simultaneously as predictors. Population density, however, is still significant after the addition of median household income as a predictor.

Conclusion of Multiple Linear Regression

Through this Multiple Linear Regression analysis, we can once again see that population density and COVID-19 case concentration are positively correlated, and median household income and COVID-19 case concentration share no significant relationship.

Although the R^2 value of this MLR improves from both individual SLR models, the adjusted R^2 value does not significantly change from that calculated with just population density as a predictor. Thus, the MLR model is not significantly better than the SLR model of variables COVID-19 case concentration data (Y) vs Population density (X_1). This, combined with the fact that in the Multiple Linear Regression model, only X_1 is significant, leads us to conclude that the SLR model of variables COVID-19 case concentration data (Y) vs Population density (X_1) should be preferred over this MLR model.

Logarithmic Transformation of Population Density

Given the cluster of points near the Y-axis in figure 4, we conducted regression analysis on the transformation of the population density to see if it results in a model of better fit.

1. Simple Linear Regression

Table 15: Regression Analysis statistics of the COVID-19 case concentration data vs ln(Population Density)

Regression Analysis	
Correlation coefficient	0.6281
R^2	0.3945
Adjusted R^2	0.3813
Standard error	19.1722
Observations	48

Table 16: ANOVA Table for Simple Linear Regression of COVID-19 case concentration data vs ln(Population Density)

	df	SS	MS	F	Significance F
Regression	1	11016.7812	11016.7812	29.9717	1.76×10^{-6}
Residual	46	16908.3671	367.5732		
Total	47	27925.1483			

Table 17: Simple Linear Regression of COVID-19 case concentration data vs ln(Population Density): Values of Intercept and Slope of Line

	Intercept (b_0)	Slope (b_1, X_1)
Coefficients	-12.1626	10.2431
Standard Error	11.1777	1.8710
t Stat	-1.0881	5.4746
F Stat	-	29.9717
P-value	0.2822	1.76×10^{-6}
Lower 95%	-34.6621	6.4769
Upper 95%	10.3369	14.0092

We can see from table 15 that the sample correlation coefficient is 0.63, which shows that there is a better relationship between the two variables than when population density was used. Also, the adjusted coefficient of determination, R^2 adj. is 0.38. This shows that about 38% of the variability can be explained out of the total variability. The standard error of the estimate is 19.17.

From table 17, we can see that the estimated standard deviation of the slope is 1.87. Since we are assuming a significance level of 5%, and the P-value for the slope (1.76×10^{-6}) is less than the significance level, we have enough evidence to support the conclusion that there is a positive correlation between logarithmic population density and COVID-19 case concentration within Massachusetts towns.

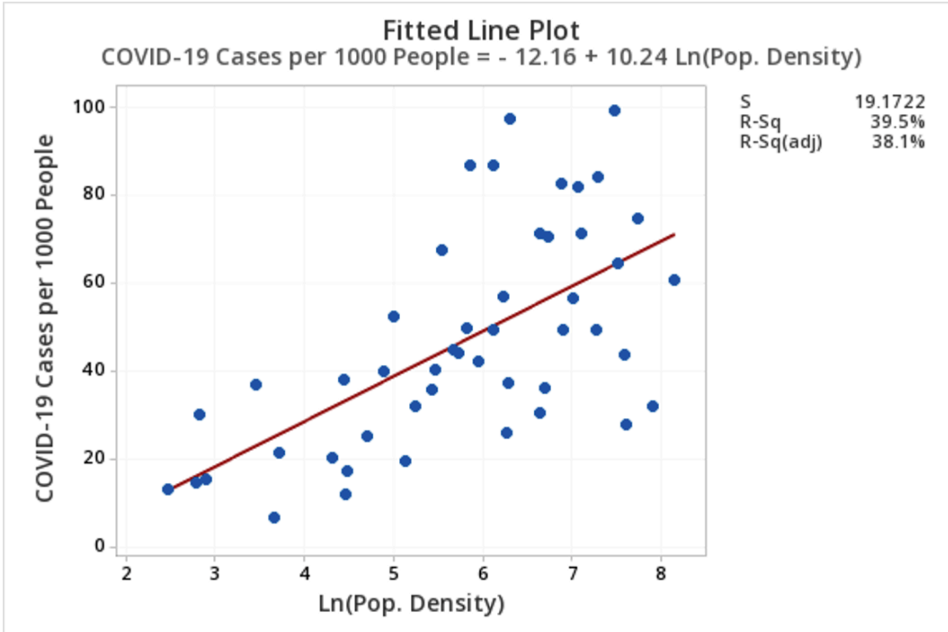

Figure 8: Fitted Line Plot of COVID-19 case concentration data (Y) vs ln(Population Density)

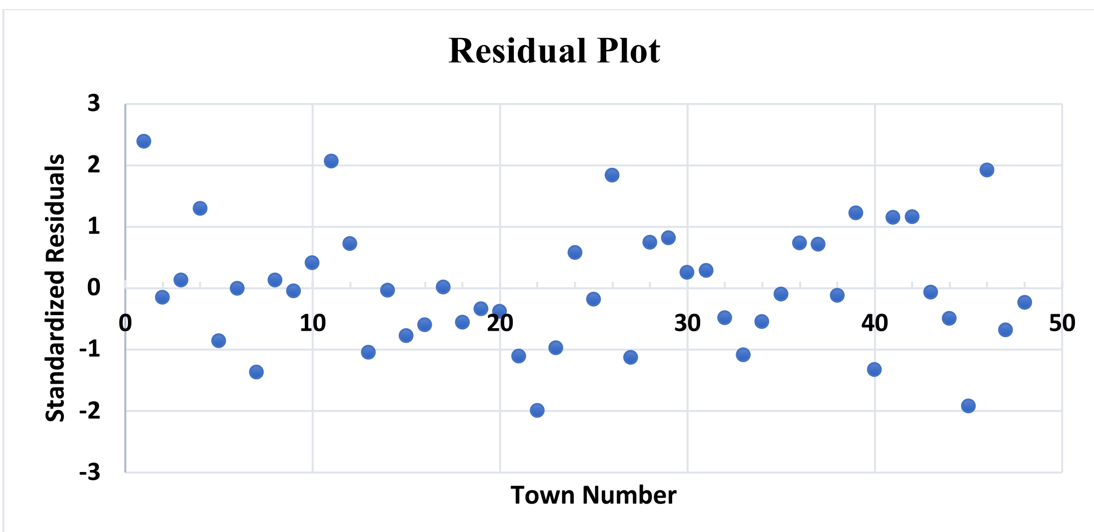

Figure 9: Standardized Residual Plot of SLR of COVID-19 case concentration data (Y) vs ln(Population Density)

The fitted line can be seen in figure 8. We can see from figure 8 that there exists a strong linear relation between logarithmic population density and COVID-19 case concentration. Also, the data points are scattered around and are not forming a cluster which was the case when we did not use logarithmic transformation (can be seen in figure 4). From figure 9, we can also see that the residuals are scattered around the X-axis and most of them have values near 0. This shows that a strong linear relationship exists between the two variables.

Our hypothesis that there should be a stronger relation between COVID-19 case concentration and logarithmic population density has been proven correct.

2. Multiple Linear Regression

Table 18: Regression Analysis statistics of the COVID-19 case concentration data vs ln(Population Density) and Median Household Income

Regression Analysis	
Multiple R	0.6777
R^2	0.4592
Adjusted R^2	0.4352
Standard error	18.3191
Observations	48

Table 19: ANOVA Table for Simple Linear Regression of COVID-19 case concentration data vs ln(Population Density) and Median Household Income

	df	SS	MS	F	Significance F
Regression	2	12823.6590	6411.8295	19.1062	9.87×10^{-7}
Residual	45	15101.4890	335.5887		
Total	47	27925.1480			

Table 20: Simple Linear Regression of COVID-19 case concentration data vs ln(Population Density) and Median Household Income: Values of Intercept and Slope of Line

	Intercept (b_0)	Slope (b_1, $\ln(X_1)$)	Slope (b_2, X_2)
Coefficients	-3.0378	11.6957	-0.0002
Standard Error	11.3813	1.8942	~ 0.0000
t Stat	-0.2669	6.1745	-2.3204
P-value	0.7907	1.72×10^{-7}	0.0249
Lower 95%	-25.9608	7.8806	-0.0004
Upper 95%	19.8852	15.5107	-0.0000

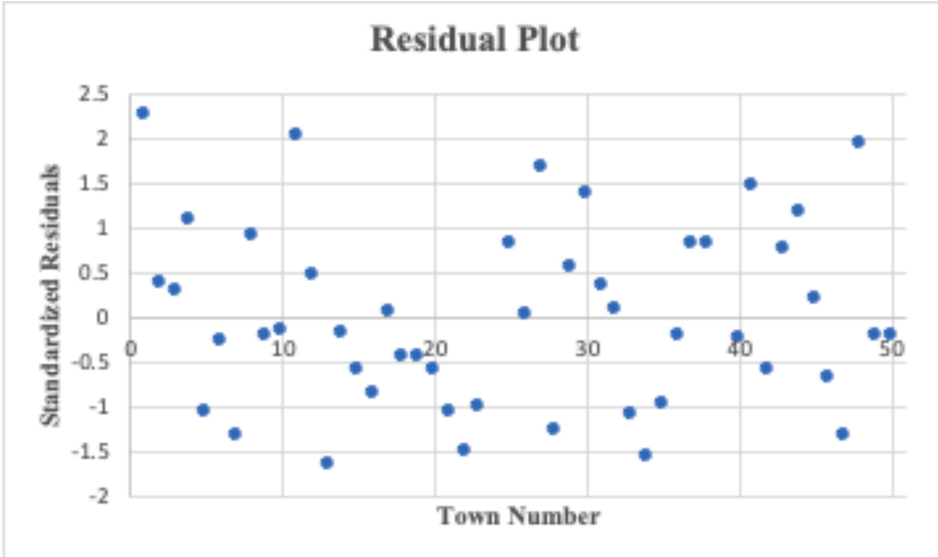

Figure 10: Standardized Residual Plot of MLR of COVID-19 case concentration data (Y) vs Population Density ln(X_1) & Median Household Income ln(X_2)

We can see from table 18 that the sample correlation coefficient is 0.68, which strongly shows that there is some relationship between the two variables. Also, the coefficient of determination, R^2 is 0.46. This shows that about 46% of the total variability can be explained by this logarithmic relationship. The standard error of the estimate is 18.31.

The P-value for the slope (X_1) is nearly 0 while P-value for the slope (X_2) is 0.0249.

Case 1: Significance level of 10%

We can see that the P-value for the slope of X_1 and slope of X_2 is less than the significance level. Hence, we have enough evidence to support the conclusion that there is a positive correlation between population density and COVID-19 case concentration within Massachusetts towns, and there is negative correlation between median household income and COVID-19 case concentration within Massachusetts towns.

Case 2: Significance level of 5%

The result of case 1 holds true for a significance level of 5%.

Case 3: Significance level of 1%

The P-value for the slope of X_1 is less than the significance level. However, the P-value for the slope of X_2 is higher than the significance level. Thus, we can drop the independent variable Median Household Income since it's insignificant. Hence, we have enough evidence to support the conclusion that there is a positive correlation between population density and COVID-19 case concentration within Massachusetts towns and there is no correlation between median household income and COVID-19 case concentration within Massachusetts towns.

This analysis suggests that there is a positive relationship between variables COVID-19 case concentration data (Y) and logarithmic population density ($Ln(X_1)$) and a negative relationship between variables COVID-19 case concentration data (Y) and Median Household Income (X_2).

The independent variable Median Household Income (X_2) is significant only for a high α. Based on this analysis, we can say that there is relatively weak relationship between the variables COVID-19 case concentration data (Y) and Median Household Income (X_2) in comparison to the relationship between variables COVID-19 case concentration data (Y) and Population density (X_1). Hence, Population density (X_1) is much more significant variable for COVID-19 case concentration (Y). It is interesting to note that despite median household income being insignificant in the first MLR model, is significant when a logarithm transform is applied to the population density data. Given its high significance value, and the fact that the adjusted R^2 value only increases slightly, median household income is proving to still be a negligible factor in predicting COVID-19 concentration.

Conclusion

Through this Multiple Linear Regression analysis, we can conclude that population density and COVID-19 case concentration are positively correlated, and median household income and COVID-19 case concentration are negatively correlated. Also, the significance of population density on the dependent variable COVID-19 case concentration is much higher in comparison to the significance of median household income on COVID-19 case concentration. This tells us that the population density is a more significant predictor of the spread of COVID-19.

After analyzing both the SLR Logarithmic Transformation and MLR Logarithmic Transformation models, we can say that regression model did not significantly improve as compared to the SLR model of logarithmic population density. Despite both predictors becoming significant in the logarithmic MRL model, the adjusted R^2 value only improved slightly from the logarithmic transformation SLR model which indicates that the median household income is only contributing slightly to accounting for variation in the data. In other words, it may be sufficient to use the logarithmic transformation of population data to predict COVID-19 cases in cities.

Though we discovered that population density is a significant predictor of COVID-19 concentration in towns, population density and median household income are only two geographic measures. Other measures such as location and accessibility measures can be examined to see if there are other, more powerful predictors of COVID-19 concentration. Testing other variables on other locations (besides Massachusetts towns) could uncover more easily attainable data that can predict the spread of infectious diseases in an area. These models could then be used to advocate for more proactive measures in locations that are especially susceptible to high rates of disease spreading.

References

[1] Galvin, William Galvin, Secretary of the Commonwealth of Massachusetts. (Feb. 2021). CIS: Massachusetts City and Town Incorporation and Settlement Dates <https://www.sec.state.ma.us/cis/cisctlist/ctlistalph.htm >

[2] Massachusetts Department of Public Health COVID-10 Dashboard. (Feb. 4, 2021) Weekly COVID-19 Public Health Report. <https://www.mass.gov/doc/weekly-covid-19-public-health-report - february-4-2021/download

[3] Cubit Planning Inc. (2021). Massachusetts Cities by Population. Massachusetts Demographics by Cubit. <https://www.massachusetts-demographics.com/cities_by_population>

[4] Reiss, J. & Rocheleau, M. (Dec. 11, 2018). Full list of Massachusetts median household incomes by town. The Boston Globe. <https://www.bostonglobe.com/metro/2018/12/11/full-list-massachusetts-median-household-incomes-town/eZpgJkpB1uF2FVmpM4O8XO/story.html>

[5] Wikipedia (Dec. 31, 2020). List of municipalities in Massachusetts. Wikipedia. <https://en.wikipedia.org/wiki/List_of_municipalities_in_Massachusetts>

Appendix

The descriptive statistics and boxplots the variables, i.e., population density (say, X_1), median household income (say, X_2), and COVID-19 Cases per 1,000 People (Y) after removing the outliers (Lynn and Somerville)

Table A1: Descriptive statistics after removing Lynn and Somerville form the sample data.

Descriptive Statistics n = 48	COVID-19 Cases per 1,000 People (Y)	Population Density (X_1)	Median Household Income (X_2)
Minimum	7.04	11.97	46871.00
Q1	28.65	116.94	68528.00
Median	43.32	420.98	76754.00
Q3	67.12	1088.38	99766.00
Maximum	99.46	3502.12	176852.00
Mean	47.13	720.47	88257.75
Standard Deviation	24.38	803.80	33070.46

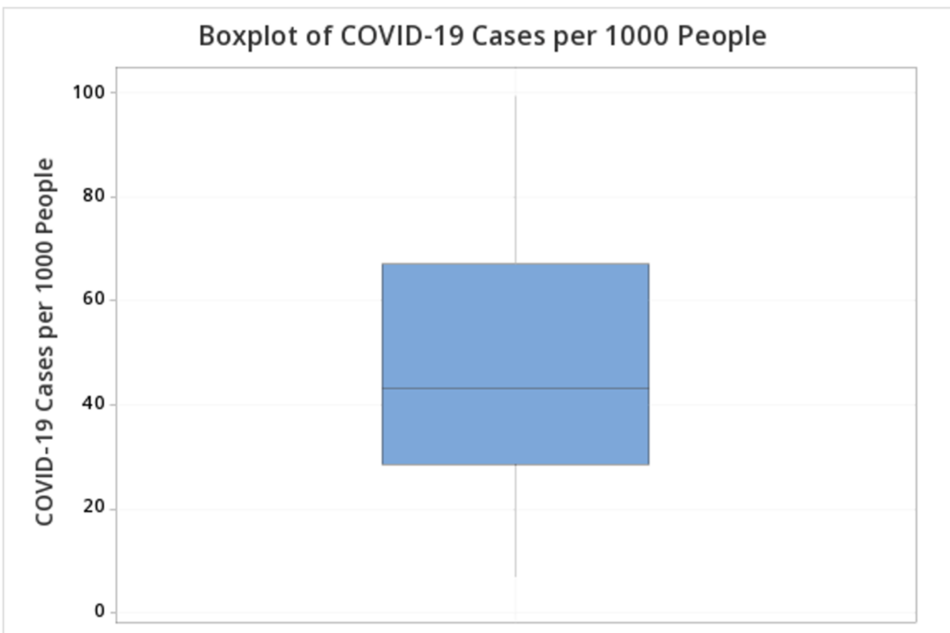

Figure A1: Boxplot of COVID-19 case concentration data (Y).

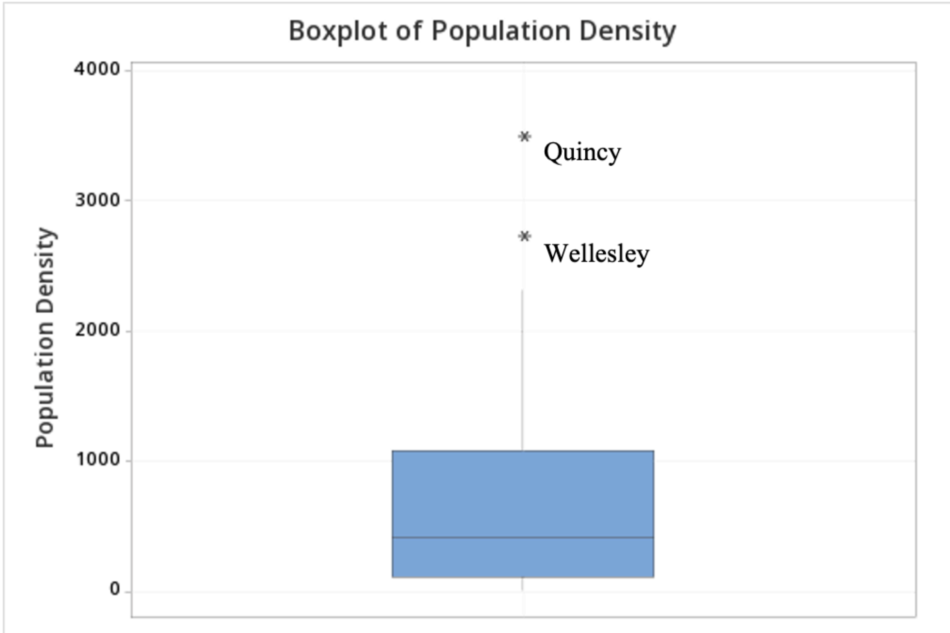

Figure A2: Boxplot of population density data (X_1).

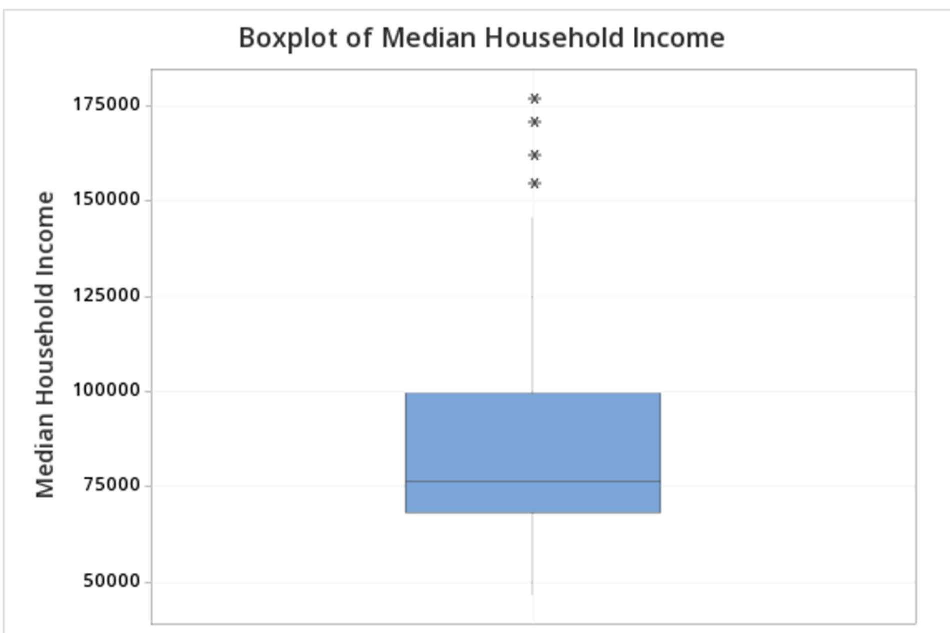

Figure A3: Boxplot of median household income (X_2).

Table A2: Simple Linear Regression of Y vs X_1: Predicted Y, 95% Confidence Interval, and 95% Prediction Interval

S. No.	Town	Y	X_1	Predicted Y	95% Confidence Interval		95% Prediction Interval	
1	Acushnet	97.83	554.16	45.0524	38.3861	51.7187	-0.6445	90.7492
2	Andover	56.96	1115.76	52.0556	44.7689	59.3422	6.2641	97.8470
3	Ashby	40.37	133.06	39.8012	31.6895	47.9130	-6.1287	85.7312
4	Avon	82.89	978.26	50.3409	43.4816	57.2003	4.6155	96.0663
5	Bernardston	17.44	88.21	39.2420	30.9063	47.5776	-6.7281	85.2120
6	Beverly	64.92	1853.32	61.2530	49.8973	72.6086	14.6406	107.8653
7	Boxborough	26.25	534.71	44.8098	38.1090	51.5106	-0.8920	90.5117
8	Boxford	50.11	339.43	42.3747	35.1394	49.6100	-3.4086	88.1580
9	Carver	45.14	295.21	41.8233	34.4240	49.2226	-3.9862	87.6327
10	Chicopee	75.08	2318.87	67.0584	52.4116	81.7052	19.5370	114.5798
11	Dartmouth	87.1	349.37	42.4986	35.2982	49.6991	-3.2792	88.2764
12	Great Barrington	52.89	150.68	40.0210	31.9942	48.0477	-5.8941	85.9360
13	Greenfield	36.6	811.92	48.2667	41.6985	54.8348	2.5841	93.9493
14	Halifax	49.86	453.29	43.7945	36.9110	50.6780	-1.9345	89.5235
15	Hamilton	37.6	538.99	44.8632	38.1703	51.5561	-0.8375	90.5639
16	Hardwick	20.67	74.71	39.0736	30.6686	47.4786	-6.9090	85.0562
17	Hawley	13.51	11.97	38.2912	29.5527	47.0298	-7.7535	84.3360
18	Hubbardston	25.49	111.56	39.5331	31.3153	47.7509	-6.4157	85.4819
19	Hull	42.47	388.66	42.9886	35.9184	50.0588	-2.7689	88.7461
20	Lenox	36.27	228.71	40.9940	33.3224	48.6656	-4.8603	86.8483
21	Leverett	12.49	87	39.2269	30.8851	47.5687	-6.7443	85.1980
22	Lexington	28.1	2020.61	63.3391	50.8353	75.8428	16.4338	110.2443
23	Leyden	7.04	39.44	38.6338	30.0435	47.2241	-7.3831	84.6507
25	Lynnfield	71.58	1228	53.4552	45.7147	61.1956	7.5894	99.3210
26	Manchester-by-the-Sea	40.61	239.51	41.1287	33.5033	48.7541	-4.7179	86.9752
27	Marlborough	99.46	1798.01	60.5632	49.5759	71.5506	14.0393	107.0872
28	Maynard	44.17	1991.48	62.9758	50.6753	75.2764	16.1243	109.8273
29	Millbury	70.93	842.45	48.6474	42.0459	55.2489	2.9600	94.3348
30	Norfolk	71.44	775.39	47.8111	41.2704	54.3519	2.1325	93.4898
31	Oakham	38.19	85.26	39.2052	30.8545	47.5559	-6.7676	85.1779
32	Oxford	57.11	508.15	44.4786	37.7249	51.2323	-1.2310	90.1883
33	Pittsfield	49.55	1006.26	50.6901	43.7564	57.6237	4.9535	96.4267
34	Provincetown	19.85	169.89	40.2605	32.3245	48.1965	-5.6387	86.1597
35	Quincy	61.11	3502.12	81.8136	58.0789	105.5483	30.7539	132.8733
36	Rowe	15.87	18.38	38.3712	29.6675	47.0748	-7.6670	84.4093

37	Royalston	37.34	32.14	38.5428	29.9134	47.1721	-7.4814	84.5669
38	Sandisfield	30.3	16.81	38.3516	29.6394	47.0638	-7.6882	84.3914
40	Southwick	44.24	306.62	41.9655	34.6099	49.3212	-3.8369	87.7680
41	Sterling	67.85	256.04	41.3348	33.7787	48.8909	-4.5003	87.1699
42	Sudbury	30.8	777.32	47.8352	41.2933	54.3770	2.1564	93.5140
43	Taunton	82.08	1180.25	52.8597	45.3228	60.3967	7.0278	98.6917
44	Tewksbury	84.54	1473.84	56.5208	47.5332	65.5084	10.4281	102.6135
45	Tolland	15.09	16.16	38.3435	29.6278	47.0592	-7.6970	84.3839
46	Warren	32.23	188.88	40.4973	32.6489	48.3457	-5.3869	86.3815
47	Wellesley	32.42	2737.81	72.2826	54.4928	90.0724	23.7003	120.8649
48	West Bridgewater	86.97	458.47	43.8591	36.9890	50.7292	-1.8679	89.5861
49	Westwood	49.58	1453.69	56.2696	47.3948	65.1443	10.1987	102.3404
50	Worthington	21.67	41.68	38.6617	30.0834	47.2401	-7.3529	84.6764

Table A3: Simple Linear Regression of Y vs X_2: Predicted Y, 95% Confidence Interval, and 95% Prediction Interval

S. No.	Town	Y	X2	Predicted Y	95% Confidence Interval		95% Prediction Interval	
1	Acushnet	97.83	69402	47.57792	39.32051	55.83532	-2.6742	97.83004
2	Andover	56.96	143292	45.80797	31.80906	59.80688	-5.6999	97.31584
3	Ashby	40.37	95833	46.94479	39.60092	54.28867	-3.16532	97.05491
4	Avon	82.89	69709	47.57056	39.34646	55.79467	-2.6761	97.81722
5	Bernardston	17.44	64647	47.69182	38.86928	56.51436	-2.65625	98.03989
6	Beverly	64.92	77893	47.37453	39.86955	54.8795	-2.75945	97.5085
7	Boxborough	26.25	101077	46.81918	39.13512	54.50324	-3.34192	96.98028
8	Boxford	50.11	155034	45.5267	29.26816	61.78524	-6.64063	97.69404
9	Carver	45.14	73904	47.47008	39.6574	55.28276	-2.71088	97.65104
10	Chicopee	75.08	48866	48.06983	36.87324	59.26643	-2.74802	98.88769
11	Dartmouth	87.1	74742	47.45	39.7091	55.19091	-2.71983	97.61984
12	Great Barrington	52.89	56124	47.89598	37.8686	57.92336	-2.67713	98.46908
13	Greenfield	36.6	47821	48.09487	36.72159	59.46814	-2.76221	98.95194
14	Halifax	49.86	77993	47.37213	39.87373	54.87053	-2.76086	97.50512
15	Hamilton	37.6	112250	46.55154	37.67994	55.42315	-3.80515	96.90823
16	Hardwick	20.67	57813	47.85552	38.08332	57.62772	-2.66761	98.37865
17	Hawley	13.51	66250	47.65342	39.03126	56.27558	-2.65992	97.96676
18	Hubbardston	25.49	93387	47.00338	39.76135	54.24542	-3.0919	97.09867
19	Hull	42.47	80584	47.31007	39.96131	54.65883	-2.80076	97.42089
20	Lenox	36.27	68492	47.59972	39.2412	55.95823	-2.66912	97.86855
21	Leverett	12.49	87174	47.15221	39.99361	54.31081	-2.93109	97.23551
22	Lexington	28.1	162083	45.35785	27.70236	63.01334	-7.26161	97.97731
23	Leyden	7.04	76771	47.4014	39.81875	54.98405	-2.74426	97.54706
25	Lynnfield	71.58	117706	46.42085	36.79573	56.04597	-4.07404	96.91574
26	Manchester-by-the-Sea	40.61	105500	46.71323	38.62617	54.8003	-3.51118	96.93764
27	Marlborough	99.46	75418	47.43381	39.74811	55.11951	-2.72754	97.59516
28	Maynard	44.17	95833	46.94479	39.60092	54.28867	-3.16532	97.05491
29	Millbury	70.93	74713	47.4507	39.70737	55.19402	-2.71951	97.62091
30	Norfolk	71.44	139137	45.9075	32.68126	59.13374	-5.39576	97.21076
31	Oakham	38.19	85938	47.18182	40.00919	54.35445	-2.90349	97.26712
32	Oxford	57.11	72563	47.5022	39.5672	55.43719	-2.69795	97.70235
33	Pittsfield	49.55	46871	48.11762	36.58215	59.65309	-2.77597	99.01122
34	Provincetown	19.85	47500	48.10255	36.67464	59.53047	-2.76677	98.97188
35	Quincy	61.11	71808	47.52028	39.51252	55.52805	-2.69142	97.73199
36	Rowe	15.87	56667	47.88297	37.93841	57.82753	-2.67378	98.43972

37	Royalston	37.34	74219	47.46253	39.67726	55.2478	-2.71417	97.63923
38	Sandisfield	30.3	68636	47.59627	39.25399	55.93855	-2.66987	97.8624
40	Southwick	44.24	76737	47.40222	39.8171	54.98733	-2.74382	97.54825
41	Sterling	67.85	102500	46.78509	38.98217	54.58801	-3.39435	96.96454
42	Sudbury	30.8	170945	45.14557	25.70289	64.58826	-8.10017	98.39131
43	Taunton	82.08	56797	47.87986	37.95502	57.8047	-2.67302	98.43273
44	Tewksbury	84.54	93817	46.99308	39.7359	54.25027	-3.1044	97.09057
45	Tolland	15.09	90417	47.07453	39.90429	54.24476	-3.01043	97.15949
46	Warren	32.23	68490	47.59976	39.24102	55.9585	-2.66911	97.86864
47	Wellesley	32.42	176852	45.00408	24.35506	65.65309	-8.6939	98.70205
48	West Bridgewater	86.97	85368	47.19547	40.01295	54.37799	-2.89125	97.28219
49	Westwood	49.58	145799	45.74792	31.27517	60.22066	-5.89075	97.38658
50	Worthington	21.67	75000	47.44382	39.72428	55.16337	-2.72272	97.61037