Liana Cramb

ECO 520

**Midterm Project**

Table of Contents

[Descriptive Analytics 2](#_Toc190473148)

[Top 5 Counties with Highest Average Percentages in Positive Cases, Deaths, and Hospitalizations: 2](#_Toc190473149)

[Size Class Analysis 3](#_Toc190473150)

[Descriptive Statistics 3](#_Toc190473151)

[Scatter Plots Displaying Relationships 4](#_Toc190473152)

[Correlation with Social and Health Variables 6](#_Toc190473153)

[Consultation Recommendations 8](#_Toc190473154)

[Cluster Analysis 9](#_Toc190473155)

[Hierarchical Clustering 9](#_Toc190473156)

[Cluster Analysis 9](#_Toc190473157)

[Pseudo T-Squared Line Plot 10](#_Toc190473158)

[Dendrogram 11](#_Toc190473159)

[Four Cluster Plots 11](#_Toc190473160)

[Non-Hierarchical Cluster (K-Means) Analysis: 14](#_Toc190473161)

[Four Cluster Analysis 14](#_Toc190473162)

[Four Cluster Plots 15](#_Toc190473163)

[Why Not Three Clusters? 18](#_Toc190473164)

[ANOVA Test on Percentage of Positive Hospitalizations 20](#_Toc190473165)

[ANOVA Test on Percentage of Positive Test Results 21](#_Toc190473166)

[ANOVA Test Box Plots 22](#_Toc190473167)

[Predictive Analytics Using Regression 24](#_Toc190473168)

[Regression Models 24](#_Toc190473169)

[Model 1: Simple Regression Model 24](#_Toc190473170)

[Model 2: Custom Model 25](#_Toc190473171)

[Model 3: Full Model 28](#_Toc190473172)

[Model 4: Stepwise Regression 30](#_Toc190473173)

[Model 4: Adjusted R-Square 31](#_Toc190473174)

[Performance of the Models 31](#_Toc190473175)

[Plotting Models 2-5 to Compare Actual v. Predicted Values 34](#_Toc190473176)

# Descriptive Analytics

## Top 5 Counties with Highest Average Percentages in Positive Cases, Deaths, and Hospitalizations:

**Top 5 in Positive Cases**

| **Obs** | **COUNTY** | **avg\_PCT\_POSITIVE** |
| --- | --- | --- |
| **1** | Brown | 30.8299 |
| **2** | Barron | 29.9677 |
| **3** | Menominee | 29.8235 |
| **4** | Juneau | 29.5407 |
| **5** | Winnebago | 29.2541 |

The top five counties with the highest average percentages in positive Covid cases are Brown, Barron, Menominee, Juneau and Winnebago.

**Top 5 in Deaths**

| **Obs** | **COUNTY** | **avg\_PCT\_DEATH** |
| --- | --- | --- |
| **1** | Barron | 0.39693 |
| **2** | Forest | 0.39086 |
| **3** | Waupaca | 0.38021 |
| **4** | Lincoln | 0.36211 |
| **5** | Oneida | 0.35780 |

The top five counties with the highest average percentages in Covid deaths are Barron, Forest, Waupaca, Lincoln and Oneida.

**Top 5 in Hospitalizations**

| **Obs** | **COUNTY** | **avg\_PCT\_HOSP\_POSITIVE** |
| --- | --- | --- |
| **1** | Price | 7.82998 |
| **2** | Clark | 7.58762 |
| **3** | Iron | 7.13538 |
| **4** | Menominee | 6.55391 |
| **5** | Burnett | 6.32084 |

The top five counties with the highest average percentages in Covid hospitalizations are Price, Clark, Iron, Menominee and Burnett.

## Size Class Analysis

### Descriptive Statistics

**The MEANS Procedure**

| **SIZE\_CLASS** | **N Obs** | **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 87 | Pct\_POSITIVE  Pct\_DEATH  Pct\_POSITIVE\_TEST  Pct\_HOSP\_POSITIVE | 87  87  87  87 | 24.8035803  0.1822981  27.0168441  5.2415768 | 4.9904615  0.1389394  3.7446948  2.1594431 | 12.1069920  0  11.9499768  1.2578616 | 36.9727047  0.6670087  32.8817734  11.7768595 |
| 2 | 555 | Pct\_POSITIVE  Pct\_DEATH  Pct\_POSITIVE\_TEST  Pct\_HOSP\_POSITIVE | 555  555  555  555 | 25.8277522  0.1974967  28.0532400  4.2274129 | 4.7286878  0.1445652  3.9809773  1.6811491 | 9.6227868  0  10.3541013  0.1486989 | 51.6015899  1.0078740  42.3460899  15.7894737 |
| 3 | 158 | Pct\_POSITIVE  Pct\_DEATH  Pct\_POSITIVE\_TEST  Pct\_HOSP\_POSITIVE | 158  158  158  158 | 26.2077559  0.1851419  28.2145248  3.4651017 | 4.4039589  0.1319402  4.0392108  1.3157242 | 14.6806319  0  14.9432437  0.1288660 | 46.8651444  0.7717157  36.9835739  9.0641120 |

* **Class Size 1:** This class shows an average of 24.80% positive cases, as well as 27.02% of the total COVID tests taken being positive. It also shows a very small percentage of deaths, with 0.18%, and a small percentage of positive COVID cases that led to hospitalization, at 5.24%. The death percentage has a very low standard deviation of 0.14%, indicating that the death rate is relatively consistent across the observations. In contrast, the standard deviations for the other three variables are slightly higher, suggesting more variability in their percentages, but still relatively small: 5% for positive cases, 3.75% for positive tests, and 2.16% for positive hospitalizations.
* **Class Size 2:** This class shows an average of 25.83% positive cases, as well as 28.05% of positive COVID tests. Like Class Size 1, this class also shows a very small percentage of deaths, with 0.20%, and a small percentage of positive COVID hospitalizations at 4.23%. Similar to Class Size 1, this class has similar standard deviations for each variable, with the death percentage having a very low standard deviation of 0.14%, indicating that the percentages are relatively consistent. Meanwhile, the standard deviations of the other three variables are higher but still relatively small: 4.73% for positive cases, 3.98% for positive tests, and 1.68% for positive hospitalizations.
* **Class Size 3:** Like Class Sizes 1 and 2, this class shows an average of 26.21% positive cases, as well as 28.21% of positive COVID tests. The death percentage is very small at 0.19%, and the percentage of positive hospitalizations is small at 3.47%. The standard deviations follow a similar pattern, with the death percentage having a standard deviation of 0.13%, and the other variables having higher but still relatively small standard deviations: 4.40% for positive cases, 4.04% for positive tests, and 1.32% for positive hospitalizations.

### Scatter Plots Displaying Relationships

I chose to compare the variables by splitting them into two groups: Positive Cases vs. Positive Test Results and Covid Deaths vs. Positive Covid Hospitalizations. These variables were selected because they had similar average percentages and naturally correlate with each other.

A diagram of a positive test results

AI-generated content may be incorrect.

The scatter plot shows that the three size class groups have similar distributions and show a somewhat linear relationship, where an increase in positive cases is closely associated with an increase in positive tests. This strong correlation makes sense, as people likely confirmed their positive cases through testing, which helps explain the similar average percentage values for both variables.

A diagram of a number of covid-19 patients

AI-generated content may be incorrect.

This scatter plot shows that the three size class groups have similar distributions and a small, but still linear, relationship where an increase in COVID deaths is associated with higher hospitalizations of positive cases. This makes sense, as deaths and hospitalizations are related in terms of severity. It’s likely that most of the deaths in Wisconsin occurred among those hospitalized, as these patients would be experiencing the worst effects of the virus.

## Correlation with Social and Health Variables

| **Pearson Correlation Coefficients, N = 800 Prob > |r| under H0: Rho=0** | |
| --- | --- |
|  | **Pct\_POSITIVE\_TEST** |
| **Pct\_POSITIVE\_TEST** | 1.00000 |
| **POP\_LT18** | 0.18130  <.0001 |
| **POP\_65P** | -0.13439  0.0001 |
| **HOUS\_NO\_VEH** | -0.07661  0.0303 |
| **POP\_MEDICAD** | 0.12941  0.0002 |
| **POP\_MEDICARE** | 0.08601  0.0150 |
| **POP\_HEALTHINS** | 0.20885  <.0001 |
| **ADULT\_LIMITED\_ENGLISH** | 0.12772  0.0003 |
| **ADULT\_SPANISH\_LENG** | 0.18050  <.0001 |
| **POP\_BELOWPOV** | 0.05669  0.1091 |
| **POP\_DISABILITY** | 0.12473  0.0004 |
| **HOUS\_NOSMARTPHN** | 0.25752  <.0001 |
| **HOUS\_NOINTERNET** | 0.24109  <.0001 |

After running the correlation analysis on all variables that I found to fall under the demographic, social and health variables, the variable that had the highest correlation with Pct\_POSITIVE\_TEST was HOUS\_NOSMARTPHN.

A diagram of a number of blue dots

AI-generated content may be incorrect.

Based on the scatter plot, we can see a relationship between the two variables in which as the percentage of positive tests increases, so does the percentage of households without smartphones. One possible reason is that households without smartphones have less direct access to online information about the virus, which may have led to increased exposure and fewer preventative measures. Also, a lack of smartphones can be an indicator of lower-income households, where individuals are more likely to work in in-person jobs, increasing their risk of exposure and contributing to higher transmission rates.

## Consultation Recommendations

**The MEANS Procedure**

| **COUNTY** | **N Obs** | **Variable** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dane | 56 | Pct\_POSITIVE  Pct\_DEATH  Pct\_POSITIVE\_TEST  Pct\_HOSP\_POSITIVE | 56  56  56  56 | 21.2296693  0.0890481  20.1669405  2.2503174 | 6.6730165  0.1074878  5.0346982  1.3944866 | 9.6227868  0  10.3541013  0.1288660 | 39.4440874  0.6670087  31.9956020  8.8235294 |
| Milwaukee | 168 | Pct\_POSITIVE  Pct\_DEATH  Pct\_POSITIVE\_TEST  Pct\_HOSP\_POSITIVE | 168  168  168  168 | 26.8767364  0.1922156  27.1485316  5.1699185 | 4.7946451  0.1550684  3.5727682  2.2950223 | 16.0728018  0  12.9877008  0.6379585 | 51.6015899  1.0078740  34.5066667  15.7894737 |
| Waukesha | 53 | Pct\_POSITIVE  Pct\_DEATH  Pct\_POSITIVE\_TEST  Pct\_HOSP\_POSITIVE | 53  53  53  53 | 26.1619154  0.2295308  28.4693434  4.1977820 | 2.1682963  0.1816584  2.1920557  1.1401937 | 21.0400000  0.0364033  23.4843301  2.3622047 | 33.5015940  0.8766234  33.7707182  7.3170732 |

I recommend that the State of Wisconsin focus on the following three counties: Dane, Milwaukee, and Waukesha. These counties had the highest number of observations, as well as relatively high percentages of positive confirmed cases and positive tests. While these three counties were not among the top five for deaths and positive hospitalizations (which were mostly counties with fewer than 10 observations), they account for a combined total of 277 observations, making up approximately 34.63% of the data from all 72 counties. Given their substantial representation in the dataset, focusing on these three counties would provide valuable insights into how COVID is impacting Wisconsin.

# Cluster Analysis

## Hierarchical Clustering

### Cluster Analysis

| **Cluster History** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Number of Clusters** | **Clusters Joined** | | **Freq** | **Semipartial R-Square** | **R-Square** | **Approximate Expected R-Square** | **Cubic Clustering Criterion** | **Pseudo F Statistic** | **Pseudo t-Squared** | **Tie** |
| **15** | CL34 | CL28 | 55 | 0.0012 | .990 | .980 | 20.4 | 5590 | 95.6 |  |
| **14** | CL25 | CL36 | 27 | 0.0018 | .988 | .978 | 17.7 | 5097 | 78.0 |  |
| **13** | CL14 | OB9 | 28 | 0.0019 | .986 | .977 | 15.5 | 4731 | 21.4 |  |
| **12** | OB138 | CL24 | 4 | 0.0024 | .984 | .975 | 13.3 | 4392 | 13.6 |  |
| **11** | CL22 | CL29 | 113 | 0.0026 | .981 | .972 | 11.6 | 4162 | 223 |  |
| **10** | CL16 | CL32 | 10 | 0.0033 | .978 | .969 | 9.82 | 3926 | 16.3 |  |
| **9** | CL15 | CL19 | 102 | 0.0060 | .972 | .966 | 5.88 | 3445 | 198 |  |
| **8** | CL13 | CL18 | 38 | 0.0073 | .965 | .962 | 2.63 | 3105 | 48.8 |  |
| **7** | CL10 | CL31 | 13 | 0.0124 | .952 | .956 | -2.3 | 2647 | 27.1 |  |
| **6** | CL11 | CL20 | 611 | 0.0133 | .939 | .948 | -3.5 | 2450 | 1774 |  |
| **5** | CL9 | CL12 | 106 | 0.0170 | .922 | .937 | -4.4 | 2354 | 150 |  |
| **4** | CL5 | CL17 | 138 | 0.0179 | .904 | .915 | -2.6 | 2504 | 81.6 |  |
| **3** | CL8 | CL7 | 51 | 0.0852 | .819 | .867 | -7.3 | 1804 | 139 |  |
| **2** | CL4 | CL6 | 749 | 0.2115 | .608 | .732 | -11 | 1235 | 2402 |  |
| **1** | CL2 | CL3 | 800 | 0.6076 | .000 | .000 | 0.00 | . | 1235 |  |

Based on the cluster analysis, I found that the R-Square value improved significantly the most when it hit around 4 cluster groups. Since we are working with three variables, I chose to use the 'Ward's procedure' because this method provided better results based on the R-Square values versus the centroid method. This is because the ward procedure focuses on minimizing variance within clusters, leading to compact and well-separated groups.

### Pseudo T-Squared Line Plot

A graph of a number of clusters

AI-generated content may be incorrect.



As displayed by the Pseudo T-squared line plot, we observe an 'elbow' at around 4 clusters because this is where we see the plot first show the lowest value before increasing. This suggests that 4 clusters might be the best choice as it represents a natural division in the data.

### Dendrogram

A drawing of a whiteboard

AI-generated content may be incorrect.



The dendrogram also supports the choice of 4 clusters, as it reveals four distinct groupings. This visualization highlights the natural divisions within the data based on the three variables.

### Four Cluster Plots

A graph of different colored circles

AI-generated content may be incorrect.

In the Land Area vs. Population scatter plot, the first cluster group consists of observations with land area sizes ranging from 0 to around 150 million, with populations ranging from just over 900 to 12,500. The second cluster group includes land areas between 150 million and 375 million, with most populations falling between 1,000 and 12,500, except for one observation with a population just under 20,000. The third cluster spans land areas from approximately 400 million to 750 million, with populations ranging from just over 900 to 7,500. Lastly, the fourth cluster group consists of land areas over 750 million up to a little over 1.3 billion, with populations between just over 900 and 5,000.

A chart of different colored circles

AI-generated content may be incorrect.

In the Land Area vs. Water Area scatter plot, the first cluster group consists of observations with land areas ranging from 0 to around 150 million and water areas from 0 to just under 100 million. The second cluster group includes land areas between 150 million and 375 million, with most water areas falling between 0 and just under 100 million, except for four observations ranging from just over 100 million to 350 million and one observation with the highest water area size of just under 600 million. The third cluster spans land areas from approximately 400 million to 750 million, with water areas ranging from around 100,000 to just under 100 million, except for one observation with a water area of around 250 million. Lastly, the fourth cluster group consists of land areas over 750 million up to just over 1.3 billion, with water areas generally between 100,000 and just over 100 million, except for one observation with a water area of around 200 million.

A diagram of water area

AI-generated content may be incorrect.

The Water Area vs. Population scatter plot shows that the clusters are predominantly clumped together, indicating minimal variation or distinction between the two variables. While this suggests that water area size may not be a strong standalone predictor of population, its inclusion in the previous scatter plot helped form meaningful and distinct cluster groups. Overall, this graph indicates that population and water area most likely have limited variation when examined together.

## Non-Hierarchical Cluster (K-Means) Analysis:

### Four Cluster Analysis

| **Cluster Means** | | | |
| --- | --- | --- | --- |
| **Cluster** | **AREA\_LAND** | **AREA\_WATER** | **POPULATION** |
| **1** | 39,543,821 | 2,285,643 | 4,241 |
| **2** | 426,476,333 | 14,614,458 | 3,974 |
| **3** | 1,071,799,551 | 51,016,700 | 2,702 |
| **4** | 157,462,214 | 358,422,622 | 2,399 |

\*Added commas to make it easier to read the values

The means of the clusters provides insights into the central tendencies of the three variables for each cluster:

* **Cluster 1:** Has an average of land area size of 39.5 million, water area size of 2.3 million and total population of 4,241.
* **Cluster 2:** Has an average of land area size of 426.5 million, water area size of 14.6 million and total population of 3,974.
* **Cluster 3:** Has an average of land area size of 1.07 billion, water area size of 51 million and total population of 2,702.
* **Cluster 4:** Has an average of land area size of 157.5 million, water area size of 358.4 million and total population of 2,399.

| **Statistics for Variables** | | | | |
| --- | --- | --- | --- | --- |
| **Variable** | **Total STD** | **Within STD** | **R-Square** | **RSQ/(1-RSQ)** |
| **AREA\_LAND** | 186294937 | 80002050 | 0.816276 | 4.442940 |
| **AREA\_WATER** | 31070340 | 17215627 | 0.694142 | 2.269490 |
| **POPULATION** | 1784 | 1772 | 0.016961 | 0.017253 |
| **OVER-ALL** | 109043064 | 47246537 | 0.812970 | 4.34674 |

The R-squared values in this table represent the proportion of variance in each variable explained by the clustering. Land area size has an R-squared value of 0.82 (82%), indicating that the clustering explains most of the variance in land area, with clusters being largely differentiated by land area size. Water area size also shows a high R-squared value of 0.69 (69%), meaning that the clustering explains a substantial portion of the variance in water area, just not as strongly as land area. In contrast, population has a very low R-squared value of 0.02 (2%), suggesting that population size is not a major factor in defining the clusters, as it contributes very little to the variance within the cluster groups. Finally, the overall R-squared value of 0.81 (81%) indicates that the clustering explains a significant portion of the combined variance across land area, water area, and population, suggesting that the clustering model is a good fit for the data and captures meaningful patterns.

### Four Cluster Plots

A graph of different colored circles

AI-generated content may be incorrect.

Compared to the hierarchical clustering scatter plots shown earlier in this report, the k-means scatter plot of Land Area vs. Population shows slight differences in cluster groupings. Cluster group one remains largely the same but now includes part of the left side of cluster group two from the hierarchical plot. Cluster group two in k-means consists of mostly the same observations but merges the remainder of hierarchical cluster group two with cluster group three. K-means cluster group three uses the same observations to hierarchical cluster group four. Lastly, k-means cluster group four appears as a small set of observations mixed within cluster group one, particularly where land area is around 125 million and population is approximately around 2,500.

A graph of different colored circles

AI-generated content may be incorrect.

The k-means scatter plot of Land Area vs. Water Area shows only minor differences compared to the hierarchical plot. Cluster group one remains largely the same but now includes part of the left side of cluster group two from the hierarchical plot, with land area ranging from 0 to 250 million. Cluster group two in k-means combines the remainder of hierarchical cluster group two with cluster group three. K-means cluster group three closely uses the same observations to hierarchical cluster group four. Lastly, k-means cluster group four consists of only four observations, with land area sizes just under 250 million and water area sizes ranging from 250 million to just under 600 million.

A chart with different colored dots

AI-generated content may be incorrect.

Similarly, as seen in the hierarchical scatter plot for Water Area vs. Population, the k-means scatter plot shows that data points are mostly clustered together, with no distinct groups. This reinforces the observation that water area size and population exhibit minimal variability when analyzed together. This finding aligns with the earlier R-squared values, where water area size had an R-squared of 0.69, meaning it explains 69% of the variance in the clustering. This suggests that water area size plays a significant role in differentiating the clusters. Additionally, the scatter plots confirm that clear cluster groupings emerge only when either water area size or land area size is used as a variable.

## Why Not Three Clusters?

**Hierarchical Plots**

A graph of different colored circles

AI-generated content may be incorrect.

A graph of different colored circles

AI-generated content may be incorrect.

**K-Means Plots**

A graph of different colored circles

AI-generated content may be incorrect.

A graph of different colored circles

AI-generated content may be incorrect.

When I first looked at the cluster analysis and dendrogram, I initially considered three clusters to be a better fit. However, after testing both hierarchical and k-means clustering, I am confident that four clusters provide a more accurate grouping. As shown in the plots above, cluster group three in the hierarchical plot and cluster group one in the k-means plot are too widely spread along the x-axis. In the hierarchical plot, cluster group three spans land area sizes from approximately 500 million to over 1.3 billion, while in the k-means plot, cluster group one ranges from around 100 million to 650 million. These large ranges indicate significant variation within a single cluster, suggesting that splitting them would be better. By setting the number of clusters to four, the analysis effectively splits these broad groups into more distinct clusters, resulting in a better fit.

## ANOVA Test on Percentage of Positive Hospitalizations

**The ANOVA Procedure**

**Dependent Variable: Pct\_HOSP\_POSITIVE**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 3 | 76.366753 | 25.455584 | 8.65 | <.0001 |
| **Error** | 796 | 2342.206241 | 2.942470 |  |  |
| **Corrected Total** | 799 | 2418.572994 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **Pct\_HOSP\_POSITIVE Mean** |
| --- | --- | --- | --- |
| 0.031575 | 40.96735 | 1.715363 | 4.187147 |

| **Source** | **DF** | **Anova SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **CLUSTER** | 3 | 76.36675267 | 25.45558422 | 8.65 | <.0001 |

* **The null hypothesis (H0):** The clusters are **not significantly associated** with the percentage of positive hospitalizations, meaning the **mean percentages of positive hospitalizations are the same across all clusters.**
* **The alternative hypothesis (HA):** The clusters **are significantly associated** with the percentage of positive hospitalizations, meaning **at least one cluster has a significantly different mean percentage** compared to the others.

Since the p-value of <.0001 is smaller than 0.05, we reject the null hypothesis and conclude that the clusters are significantly associated with the percentage of positive hospitalizations. The ANOVA sum of squares for the clusters (76.37) represents the portion of the variance in positive hospitalizations explained by the clusters. While this value is relatively small compared to the total sum of squares (2418.57), it still indicates that the clusters explain some variability in the percentage of positive hospitalizations. The R-squared value of 0.031575 (or 3.16%) shows that the clusters account for a small portion of the variance in the percentage of positive hospitalizations. Although this is a small percentage, it suggests that there is a statistically significant association between clusters and positive hospitalizations. This also indicates that other factors, not captured by the clustering, contribute to the variability in the percentage of positive hospitalizations.

## ANOVA Test on Percentage of Positive Test Results

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **The ANOVA Procedure**    **Dependent Variable: Pct\_POSITIVE\_TEST**   | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | --- | --- | --- | --- | --- | --- | | **Model** | 3 | 332.07072 | 110.69024 | 7.16 | <.0001 | | **Error** | 796 | 12307.59528 | 15.46180 |  |  | | **Corrected Total** | 799 | 12639.66600 |  |  |  | |

| **R-Square** | **Coeff Var** | **Root MSE** | **Pct\_POSITIVE\_TEST Mean** |
| --- | --- | --- | --- |
| 0.026272 | 14.05726 | 3.932150 | 27.97239 |

| **Source** | **DF** | **Anova SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **CLUSTER** | 3 | 332.0707243 | 110.6902414 | 7.16 | <.0001 |

* **The null hypothesis (H0):** The clusters are **not significantly associated** with the percentage of positive test results, meaning the **mean percentages of positive tests are the same across all clusters.**
* **The alternative hypothesis (HA):** The clusters **are significantly associated** with the percentage of positive test results, meaning **at least one cluster has a significantly different mean percentage** compared to the others.

Since the p-value of <.0001 is smaller than 0.05, we reject the null hypothesis and conclude that the clusters are significantly associated with the percentage of positive test results. The ANOVA sum of squares for the clusters (332.07) represents the portion of the variance in positive test results explained by the clusters. While this value is relatively small compared to the total sum of squares (12639.67), it still indicates that the clusters explain some variability in the percentage of positive test results. The R-squared value of 0.026272 (or 2.63%) shows that the clusters account for a small portion of the variance in the percentage of positive test results. Although this is a small percentage, it suggests that there is a statistically significant association between clusters and positive test results. This also indicates that other factors, not captured by the clustering, contribute to the variability in the percentage of positive test results.

### ANOVA Test Box Plots

A diagram of a number of boxes

AI-generated content may be incorrect.

A diagram of a cluster

AI-generated content may be incorrect.

The box plots for both variables further support the conclusion that the percentage of positive hospitalizations and the percentage of positive test results are significantly associated with the clusters. If the clusters were not significantly associated with either variable, we would expect to see similar distributions, medians, and variability across all the clusters. However, as shown, each cluster group display differences.

# Predictive Analytics Using Regression

## Regression Models

### Model 1: Simple Regression Model

**The REG Procedure**

**Model: MODEL1**

**Dependent Variable: y**

|  |  |
| --- | --- |
| **Number of Observations Read** | 800 |
| **Number of Observations Used** | 560 |
| **Number of Observations with Missing Values** | 240 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 1 | 329.54048 | 329.54048 | 145.28 | <.0001 |
| **Error** | 558 | 1265.73701 | 2.26835 |  |  |
| **Corrected Total** | 559 | 1595.27749 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1.50610 | **R-Square** | 0.2066 |
| **Dependent Mean** | 4.13759 | **Adj R-Sq** | 0.2052 |
| **Coeff Var** | 36.40053 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 1.85716 | 0.19962 | 9.30 | <.0001 |
| **POP\_DISABILITY** | 1 | 0.18631 | 0.01546 | 12.05 | <.0001 |

For the simple regression model with only one independent variable, I selected POP\_DISABILITY (the percentage of the population with a disability) to predict Pct\_HOSP\_POSITIVE (the percentage of hospitalized COVID-19 patients who tested positive). This choice is based on the assumption that individuals with disabilities may have pre-existing health conditions that increase their risk of severe illness, making hospitalization more likely. The results indicate that POP\_DISABILITY has a statistically significant effect on hospitalization rates, with a p-value of <.0001, indicating a relationship between the two variables. The parameter coefficient of 0.18631 suggests that for each 1 percentage point increase in the disability population, the hospitalization rate increases by 0.18631 percentage points.

The adjusted r-squared of .2052 indicates that POP\_DISABILITY explains approximately 20.52% of the variation in hospitalization rates. While this suggests other factors also contribute to hospitalization rates, it demonstrates a meaningful relationship. Additionally, the F-value of 145.28 confirms the overall model is statistically significant. These findings support the expectation that individuals with disabilities, who may have underlying health conditions, are at higher risk for severe COVID-19 outcomes, leading to an increased likelihood of hospitalization.

### Model 2: Custom Model

**The REG Procedure**

**Model: MODEL2**

**Dependent Variable: y**

|  |  |
| --- | --- |
| **Number of Observations Read** | 800 |
| **Number of Observations Used** | 560 |
| **Number of Observations with Missing Values** | 240 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 7 | 608.34982 | 86.90712 | 48.61 | <.0001 |
| **Error** | 552 | 986.92767 | 1.78791 |  |  |
| **Corrected Total** | 559 | 1595.27749 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1.33713 | **R-Square** | 0.3813 |
| **Dependent Mean** | 4.13759 | **Adj R-Sq** | 0.3735 |
| **Coeff Var** | 32.31664 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 1.52077 | 0.39728 | 3.83 | 0.0001 |
| **SIZE\_CLASS** | 1 | -0.20327 | 0.11244 | -1.81 | 0.0712 |
| **POP\_65P** | 1 | 0.07744 | 0.01475 | 5.25 | <.0001 |
| **POP\_BELOWPOV** | 1 | -0.00811 | 0.00660 | -1.23 | 0.2196 |
| **POP\_DISABILITY** | 1 | 0.03899 | 0.01899 | 2.05 | 0.0405 |
| **POP\_MEDICAD** | 1 | 0.08030 | 0.00817 | 9.83 | <.0001 |
| **POP\_MEDICARE** | 1 | 0.04075 | 0.03519 | 1.16 | 0.2474 |
| **POP\_HEALTHINS** | 1 | 0.02184 | 0.01553 | 1.41 | 0.1601 |

For the custom regression model, I selected variables that I believed were logical predictors of why individuals who tested positive for COVID-19 would require hospitalization. One of the key variables I included was the dummy variable SIZE\_CLASS, which categorizes areas into three population groups: 0–2,199, 2,200–5,499, and 5,500 and above. Given the highly contagious nature of the virus, I wanted to examine whether population size played a role in hospitalization rates.

I also considered demographic variables that would logically predict hospitalization rates and selected POP\_65P (percentage of the population aged 65 and older), POP\_BELOWPOV (percentage of the population below the poverty level), and POP\_DISABILITY (percentage of the population with a disability). When COVID-19 became a global pandemic, health officials identified senior citizens and individuals with pre-existing conditions as some of the most vulnerable groups. Given their higher risk of severe illness, it is reasonable to expect that these demographics would account for a significant portion of hospitalizations. For individuals living below the poverty level, I considered the potential impact of workplace exposure. While many transitioned to remote work during the pandemic, lower-income individuals were more likely to hold essential or minimum-wage jobs that required in-person work. As a result, they may have faced a greater risk of exposure due to frequent interactions in public spaces, making them more susceptible to testing positive for covid.

I then considered healthcare-related variables that would logically predict hospitalization rates and selected POP\_MEDICAD (percentage of the population with Medicaid/Public Assistance), POP\_MEDICARE (percentage of the population with Medicare coverage), and POP\_HEALTHINS (percentage of the population without health insurance coverage). As with any hospital admission, one of the first things assessed is a patient’s health insurance coverage. Therefore, I thought it would be interesting to examine whether a patient’s insurance status had any effect on the rate of positive hospitalizations.

After running the custom model regression analysis, the adjusted R-squared of 0.3735 indicates that the selected variables explain 37.35% of the variation in positive hospitalization rates. Additionally, the F-value of 48.61 confirms that the overall model is statistically significant. Based on the p-values, only three variables were statistically significant at the 0.05 level: POP\_65P, POP\_DISABILITY, and POP\_MEDICAD.

POP\_65P has a positive impact on positive hospitalization rates, with a parameter coefficient of 0.07744. This suggests that for each 1 percentage point increase in the population aged 65 and older, the hospitalization rate increases by 0.07744 percentage points. POP\_DISABILITY also has a positive effect, with a coefficient of 0.03899, meaning that for each 1 percentage point increase in the population with a disability, the hospitalization rate increases by 0.03899 percentage points. Lastly, POP\_MEDICAD shows a positive relationship with hospitalization rates, with a coefficient of 0.08030, indicating that for each 1 percentage point increase in the population with Medicaid/public assistance, the hospitalization rate increases by 0.08030 percentage points.

### Model 3: Full Model

**The REG Procedure**

**Model: MODEL3**

**Dependent Variable: y**

|  |  |
| --- | --- |
| **Number of Observations Read** | 800 |
| **Number of Observations Used** | 560 |
| **Number of Observations with Missing Values** | 240 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 15 | 738.16007 | 49.21067 | 31.23 | <.0001 |
| **Error** | 544 | 857.11742 | 1.57558 |  |  |
| **Corrected Total** | 559 | 1595.27749 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 1.25522 | **R-Square** | 0.4627 |
| **Dependent Mean** | 4.13759 | **Adj R-Sq** | 0.4479 |
| **Coeff Var** | 30.33708 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 0.64686 | 0.50599 | 1.28 | 0.2016 |
| **SIZE\_CLASS** | 1 | -0.14648 | 0.10696 | -1.37 | 0.1714 |
| **AREA\_LAND** | 1 | 4.66965E-10 | 3.35258E-10 | 1.39 | 0.1642 |
| **AREA\_WATER** | 1 | -8.82449E-9 | 1.767553E-9 | -4.99 | <.0001 |
| **POP\_LT18** | 1 | 0.01669 | 0.01376 | 1.21 | 0.2256 |
| **POP\_65P** | 1 | 0.09495 | 0.01659 | 5.72 | <.0001 |
| **HOUS\_NO\_VEH** | 1 | 0.03104 | 0.01183 | 2.62 | 0.0089 |
| **ADULT\_LIMITED\_ENGLISH** | 1 | 0.03980 | 0.02159 | 1.84 | 0.0659 |
| **ADULT\_SPANISH\_LENG** | 1 | -0.07325 | 0.02439 | -3.00 | 0.0028 |
| **POP\_BELOWPOV** | 1 | -0.01731 | 0.00877 | -1.97 | 0.0490 |
| **POP\_DISABILITY** | 1 | -0.00047369 | 0.02079 | -0.02 | 0.9818 |
| **POP\_MEDICAD** | 1 | 0.06279 | 0.01104 | 5.69 | <.0001 |
| **POP\_MEDICARE** | 1 | 0.02775 | 0.03334 | 0.83 | 0.4055 |
| **POP\_HEALTHINS** | 1 | 0.02404 | 0.01888 | 1.27 | 0.2034 |
| **HOUS\_NOSMARTPHN** | 1 | 0.00669 | 0.01457 | 0.46 | 0.6463 |
| **HOUS\_NOINTERNET** | 1 | 0.04169 | 0.01671 | 2.49 | 0.0129 |

For the full model regression analysis, the variables included are all demographic, social, and healthcare related. The adjusted R-squared of 0.4479 indicates that the selected variables explain 44.79% of the variation in positive hospitalization rates. Additionally, the F-value of 31.23 confirms that the overall model is statistically significant. Based on the p-values, the variables that were statistically significant at the 0.05 level include AREA\_WATER, POP\_65P, HOUS\_NO\_VEH, ADULT\_SPANISH\_LENG, POP\_BELOWPOV, POP\_MEDICAD, and HOUS\_NOINTERNET.

The variables POP\_65P and POP\_MEDICAD show similar positive effects as in the custom model regression analysis. AREA\_WATER has a negative impact on positive hospitalization rates; the coefficient of -8.82449E-9 suggests that for each unit increase in water area size, there is a very small decrease in hospitalization rates. HOUS\_NO\_VEH has a positive effect on hospitalization rates, with a coefficient of 0.03104, indicating that a 1 percentage point increase in the population of households without vehicles results in an increase of 0.03104 percentage points in hospitalization rates. ADULT\_SPANISH\_LENG has a negative effect, with a coefficient of -0.07325, meaning that a 1 percentage point increase in adults who speak Spanish with limited English proficiency leads to a decrease of 0.07325 percentage points in the hospitalization rate. POP\_BELOWPOV also shows a negative effect, with a coefficient of -0.01731, meaning that a 1 percentage point increase in the population below the poverty level results in a slight decrease in hospitalization rates by 0.01731 percentage points. Finally, HOUS\_NOINTERNET has a positive effect on hospitalization rates, with a coefficient of 0.04169, indicating that a 1 percentage point increase in the population of households without internet access results in a 0.04169 percentage point increase in hospitalization rates.

### Model 4: Stepwise Regression

**The REG Procedure**

**Model: MODEL4**

**Dependent Variable: y**

|  |  |
| --- | --- |
| **Number of Observations Read** | 800 |
| **Number of Observations Used** | 560 |
| **Number of Observations with Missing Values** | 240 |

**All variables left in the model are significant at the 0.1500 level.**

**No other variable met the 0.1500 significance level for entry into the model.**

| **Summary of Stepwise Selection** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Variable Entered** | **Variable Removed** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | HOUS\_NOINTERNET |  | 1 | 0.2943 | 0.2943 | 158.554 | 232.67 | <.0001 |
| **2** | POP\_65P |  | 2 | 0.0384 | 0.3327 | 121.660 | 32.06 | <.0001 |
| **3** | POP\_MEDICAD |  | 3 | 0.0738 | 0.4065 | 48.8962 | 69.18 | <.0001 |
| **4** | AREA\_WATER |  | 4 | 0.0315 | 0.4380 | 18.9782 | 31.13 | <.0001 |
| **5** | ADULT\_SPANISH\_LENG |  | 5 | 0.0090 | 0.4471 | 11.8507 | 9.03 | 0.0028 |

The stepwise model resulted in five independent variables, which was a bit surprising at first, as the full regression model indicated that seven variables were statistically significant. The two variables excluded in this model were HOUS\_NO\_VEH and POP\_BELOWPOV. This is likely due to the stepwise analysis determining that, compared to the other five variables, these two were not as strong predictors of the dependent variable. The stepwise procedure likely removed them in favor of the variables with stronger relationships.

### Model 4: Adjusted R-Square

**The REG Procedure**

**Model: MODEL5**

**Dependent Variable: y**

**Adjusted R-Square Selection Method**

|  |  |
| --- | --- |
| **Number of Observations Read** | 800 |
| **Number of Observations Used** | 560 |
| **Number of Observations with Missing Values** | 240 |

| **Number in Model** | **Adjusted R-Square** | **R-Square** | **Variables in Model** |
| --- | --- | --- | --- |
| **12** | 0.4500 | 0.4618 | SIZE\_CLASS AREA\_LAND AREA\_WATER POP\_LT18 POP\_65P HOUS\_NO\_VEH ADULT\_LIMITED\_ENGLISH ADULT\_SPANISH\_LENG POP\_BELOWPOV POP\_MEDICAD POP\_HEALTHINS HOUS\_NOINTERNET |

Based on the Adjusted R-Square regression model, after evaluating all possible variable combinations for the best fit, the model concluded that using twelve variables resulted in the highest Adjusted R-Square of 0.4500, or 45%.

## Performance of the Models

**The MEANS Procedure**

| **Variable** | **N** | **Mean** |
| --- | --- | --- |
| mse1  mse2  mse3  mse4  mse5  rmse1  rmse2  rmse3  rmse4  rmse5  mpe1  mpe2  mpe3  mpe4  mpe5  mae1  mae2  mae3  mae4  mae5 | 240  240  240  240  240  240  240  240  240  240  240  240  240  240  240  240  240  240  240  240 | 2.5433218  2.0283016  1.6712778  1.6778965  1.6820507  1.1749947  1.0569726  0.9468964  0.9571609  0.9497483  0.4246203  0.3370744  0.3178406  0.3262297  0.3211129  1.1749947  1.0569726  0.9468964  0.9571609  0.9497483 |

\* Model 1 = Simple Regression, Model 2 = Custom, Model 3 = Full, Model 4 = Stepwise, and Model 5 = Adjusted R-Square

* **Mean Square Error (MSE):** Model 3 was the best with the lowest value of 1.6712778.
* **Root Mean Square Error (RMSE):** Model 3 was the best with the lowest value of 0.9468964.
* **Mean Percentage Error (MPE):** Model 3 was the best with the lowest value of 0.3178406.
* **Mean Absolute Error (MAE):** Model 3 was the best with the lowest value of 0.9468964.

Based on the statistics we measured, the overall best model was Model 3, the full regression model. The model’s performance can be explained by several factors. The rate of positive hospitalizations is influenced by a wide range of variables. For example, while one person may have little to no pre-existing conditions and be heavily affected by COVID-19, another person with a disability or a senior citizen might experience less severe effects. This demonstrates that an individual’s health history, age, and overall condition can influence how they respond to the virus. Additionally, socioeconomic status, access to healthcare, and the distinction between those who voluntarily seek medical help versus those admitted due to severe symptoms all play a role.

As we saw in the full regression model, the adjusted R-squared of around 45% indicates that nearly half of the variation in positive hospitalization rates can be explained by the selected independent variables. While this R-squared value may seem lower than expected, it is not unusual given the complexity of hospitalization rates, which are influenced by a variety of interconnected factors. In this case, it's better not to remove variables in the analysis, as it could result in losing critical insights. Instead, including a broad range of potential variables provides a more comprehensive understanding of what influences hospitalization rates in positive covid cases.

## Plotting Models 2-5 to Compare Actual v. Predicted Values

**Model 2: Simple Regression**

A graph with blue dots

AI-generated content may be incorrect.

**Model 3: Full Model**

A graph with blue dots

AI-generated content may be incorrect.

**Model 4: Stepwise**

A graph with blue dots

AI-generated content may be incorrect.

**Model 5: Adjusted R-Square**

A graph of blue dots

AI-generated content may be incorrect.

After plotting models 2-5 to compare the performance of the actual vs. predicted values, all four models show relatively linear relationships, indicating that each model performs well. We can see as the actual values increase; the predicted values follow a similar trend. While the plots appear similar at first glance, this is expected, as the models did not differ significantly from each other in the performance metrics of MSE, RMSE, MPE, and MAE. Based on our previous analysis, we determined that Model 3, the full regression model, had the best performance. This is reflected in its scatter plot, where we observe a slightly more linear alignment with points that are closer to the line compared to the other models.