Covid-19 Infection Rate & Hospitalization Rate Analysis

Results of Analysis

Based on the descriptive statistics below, I would suggest addressing the gap between the amount of infections in high-density areas and how few are hospitalized. Is this due to hospital capacity not meeting the needs of the population? Or is there a reason that high-density populations are not getting hit hard enough to require hospitalization? Further analysis is required, and additional measures might need to be taken to expand healthcare access in high-density areas.

Additionally, the Spanish-speaking population seems to have a high correlation to testing positive. This might be because they live in areas that got hit harder, or perhaps their weaker grasp of English means they missed out on early warnings or proper precautions to avoid covid-19. Their poverty correlation means they likely live in high-density groups, allowing easier spread of disease.

Correlation data suggests high Spanish population has a strong correlation to poverty, having no health insurance, and being on public healthcare programs. This suggests healthcare officials need to prioritize the uninsured or underinsured, making sure they don't get left behind.

Either way, the state needs to offer more Spanish-speaking support in hospitals and in health-related communications.

Descriptive Data Exploration

Variable	N	Mean	Std Dev	Minimum	Maximum
Pct_POSITIVE	800	25.5359203	4.8082086	9.3154552	50.7261411
Pct_HOSP_POSITIVE	800	4.1239758	1.6916608	0	11.4832536
Pct_DEATH	800	0.1950850	0.1619079	0	2.2821577
GEOID	800	55075600791	39438429.36	55001950201	55141011600
POPULATION	800	4139.07	1764.25	958.0000000	19084.00
SIZE_CLASS	800	1.9487500	0.7206057	1.0000000	3.0000000

Mean Percentage of Confirmed Cases (Top 5 Counties)

Obs	COUNTY	_TYPE_	_FREQ_	pos_mean
1	Brown	1	30	31.0681
2	Forest	1	2	29.1975
3	Fond du L	1	15	29.1847
4	Chippewa	1	6	28.8990
5	Barron	1	7	28.7104

Mean Percentage of Hospitalized (Top 5 Counties)

Obs	COUNTY	_TYPE_	_FREQ_	hosp_mean
1	Clark	1	4	7.66400
2	Marquette	1	3	6.75515
3	Burnett	1	3	6.44873
4	Rusk	1	2	6.04747
5	Green Lak	1	2	5.97553

Mean Percentage of Deaths (Top 5 Counties)

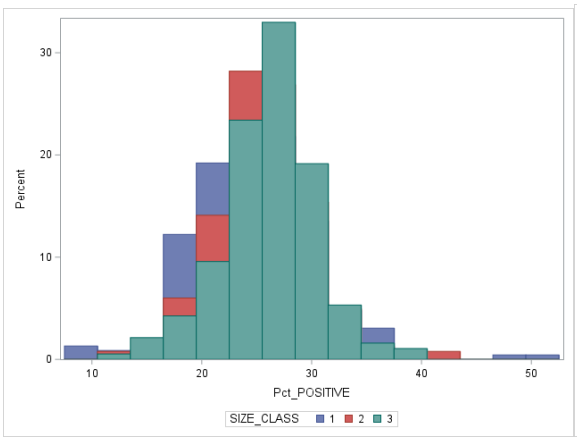
Obs	COUNTY	_TYPE_	_FREQ_	death_mean
1	Forest	1	2	0.46732
2	Iron	1	2	0.43553
3	Green Lak	1	2	0.35947
4	Kenosha	1	18	0.35250
5	Waupaca	1	8	0.33676

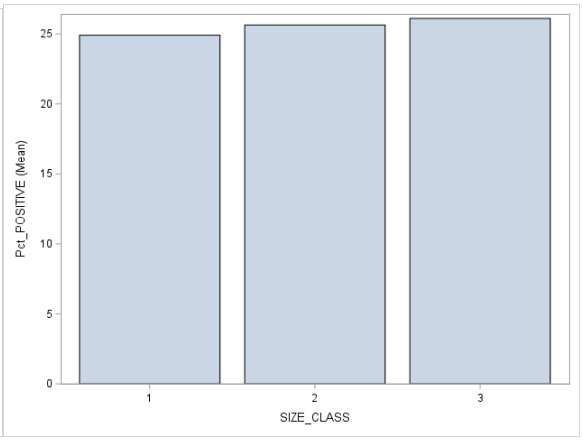
2)

The SUMMARY Procedure

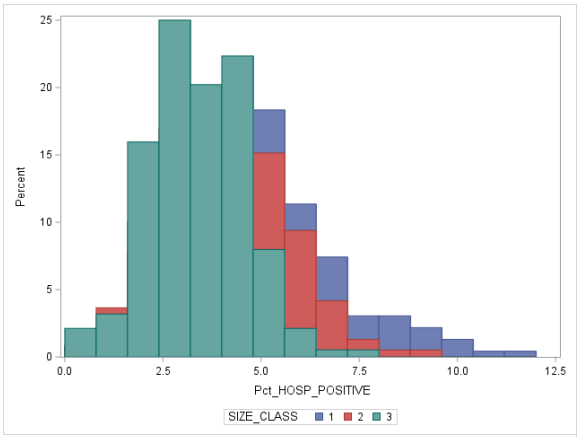
SIZE_CLASS	N Obs	Variable	N	Mean	Std Dev	Minimum	Maximum
1	**229**	**Pct_POSITIVE**	229	24.9036784	5.6006044	9.3154552	50.7261411
		Pct_HOSP_POSITIVE	229	4.7553398	1.9973605	0	11.4832536
		Pct_DEATH	229	0.1961914	0.2071032	0	2.2821577
2	**383**	**Pct_POSITIVE**	383	25.6348910	4.5160305	11.3748764	42.3578363
		Pct_HOSP_POSITIVE	383	4.1035701	1.5357857	0.1533742	9.3808630
		Pct_DEATH	383	0.2021873	0.1450072	0	0.8529997
3	**188**	**Pct_POSITIVE**	188	26.1044183	4.2474235	11.1092577	39.4440874
		Pct_HOSP_POSITIVE	188	3.3964919	1.2344825	0.1288660	7.3170732
		Pct_DEATH	188	0.1792685	0.1280631	0	0.7717157

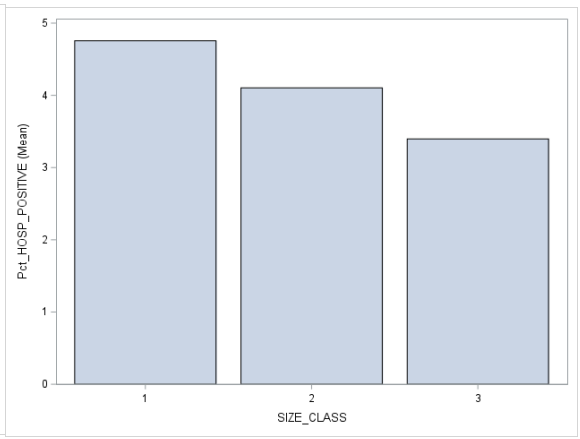
Pct_Positive Graphs



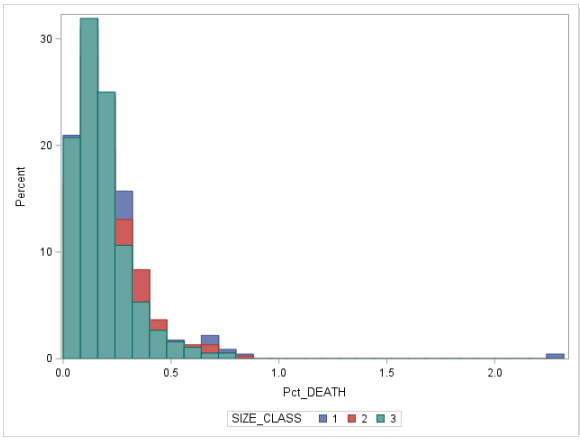


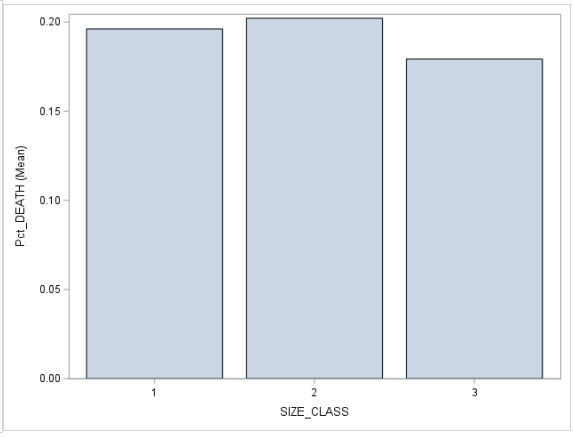
Pct_Hosp_Positive Graphs





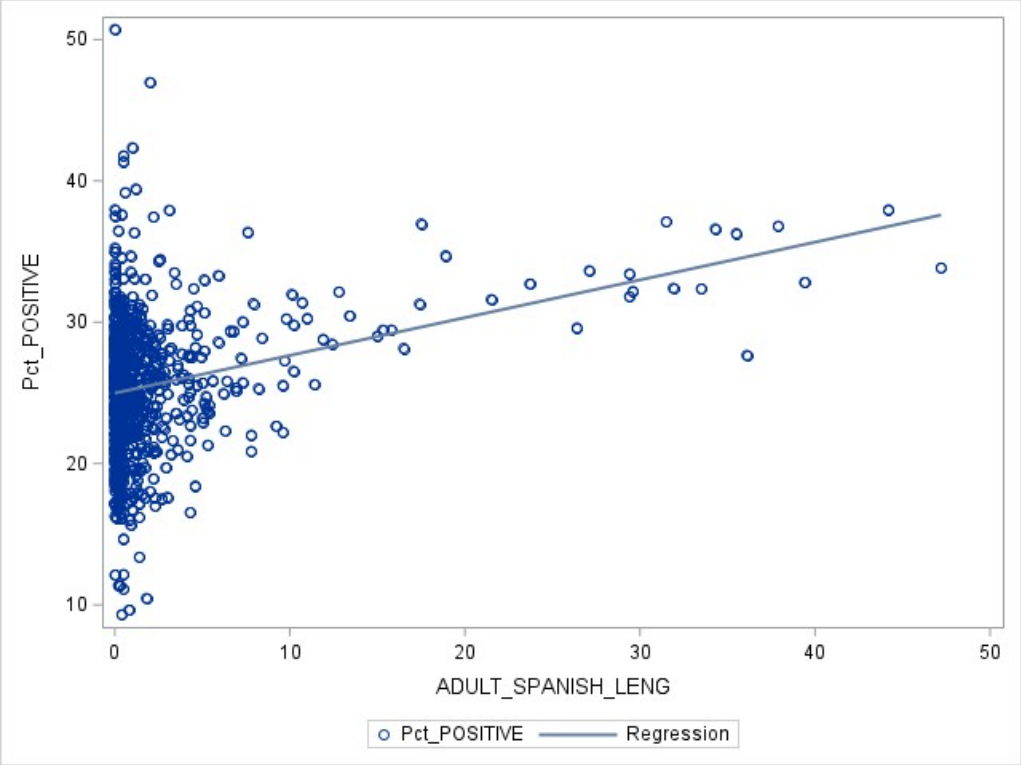
Pct_Death Graphs





Correlations with Pct_POSITIVE

	Pct_POSITIVE
Pct_POSITIVE	1.00000 800
ADULT_LIMITED_ENGLISH	0.23936 800
ADULT_SPANISH_LENG	0.29167 800
POP_LT18	0.09119 800
POP_65P	-0.18114 800
POP_BELOWPOV	0.13652 799
POP_DISABILITY	0.19314 799
POP_HEALTHINS	0.12168 799
POP_MEDICAD	0.17582 799
POP_MEDICARE	0.00304 799
SIZE_CLASS	0.09091 800
HOUS_NOINTERNET	0.11399 799
HOUS_NOSMARTPHN	0.00138 799
HOUS_NO_VEH	0.10040

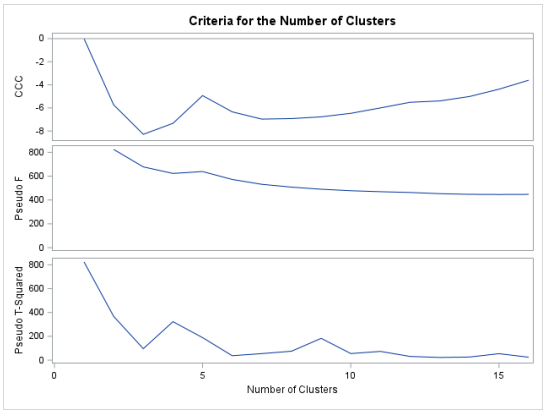

Pct_POSITIVE
50
40
30
20
10
0
10
20
30
40
50
ADULT_SPANISH_LENG
Pct_POSITIVE
Regression

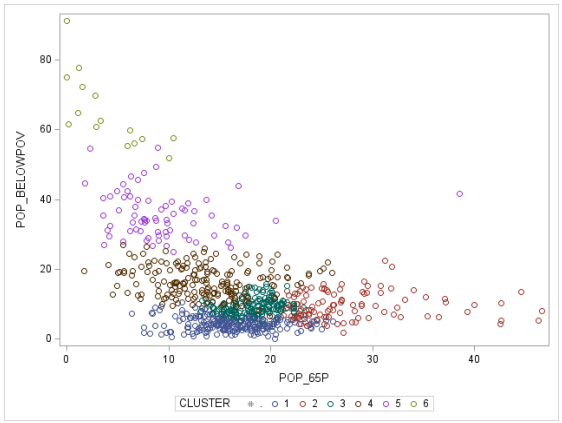
Hierarchal Clustering

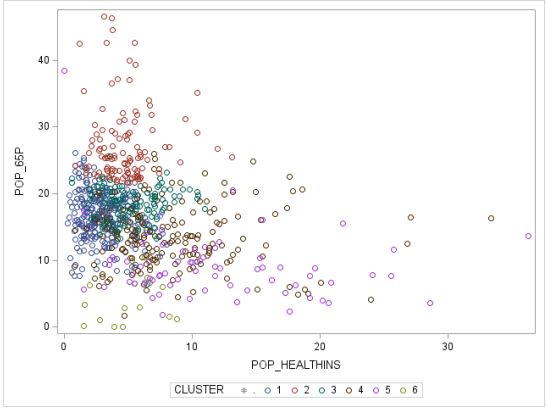
Cluster History

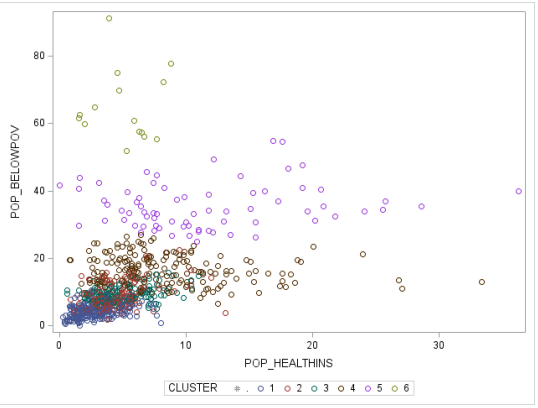
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic	Pseudo t-Squared	Tie
16	**OB1**	**CL31**	19	0.0053	.895	.905	-3.6	447	24.6	
15	**CL21**	**CL23**	104	0.0070	.888	.900	-4.4	446	54.3	
14	**CL32**	**CL37**	15	0.0074	.881	.895	-5.0	448	25.9	
13	**CL16**	**CL26**	53	0.0074	.874	.890	-5.4	453	22.6	
12	**CL24**	**CL36**	30	0.0075	.866	.883	-5.5	463	31.0	
11	**CL25**	**CL33**	91	0.0099	.856	.876	-6.0	469	73.4	
10	**CL18**	**CL11**	159	0.0113	.845	.868	-6.5	478	55.3	
9	**CL20**	**CL28**	233	0.0126	.832	.858	-6.8	490	183	
8	**CL15**	**CL53**	114	0.0143	.818	.846	-6.9	508	75.0	
7	**CL10**	**CL12**	189	0.0170	.801	.831	-7.0	531	55.0	
6	**CL13**	**CL17**	72	0.0181	.783	.812	-6.3	572	37.1	
5	**CL9**	**CL19**	409	0.0201	.763	.787	-4.9	639	189	
4	**CL5**	**CL8**	523	0.0614	.702	.752	-7.3	623	323	
3	**CL6**	**CL14**	87	0.0715	.630	.694	-8.3	678	96.1	
2	**CL4**	**CL7**	712	0.1216	.508	.576	-5.7	824	367	
1	**CL3**	**CL2**	799	0.5084	.000	.000	0.00	.	824	

Hierarchal Clustering

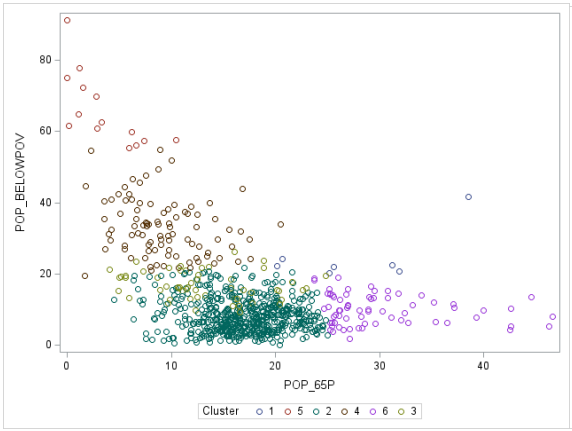


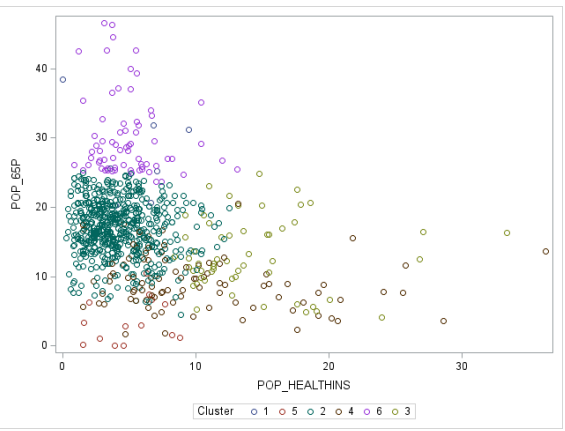


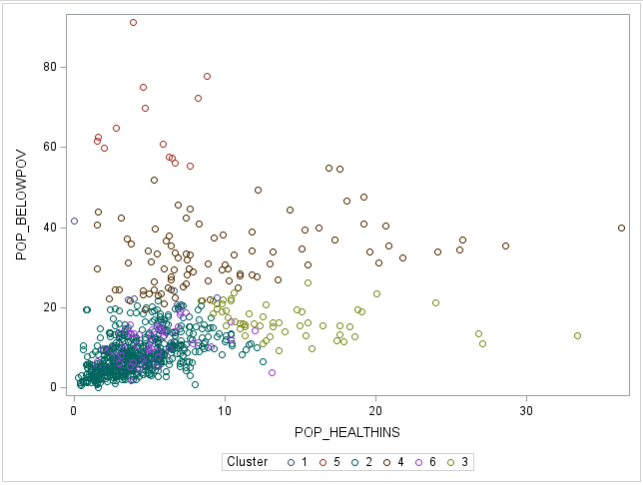




K-means Clustering







The K-cluster in this case seems to do a solid job of
creating clear regions across all three graphs, whereas the hierarchal cluster still intermingles significantly.

So we'll use the non-hierarchal clustering to define groups, and need to remember Pop_healthins represents the *uninsured.*

Using Correlation data to support the assumptions of wealth, the groups are:

Cluster 1— Mid poverty, high elderly, low % of uninsured.
- Retired_mid

Cluster 2— Low poverty, low elderly, low % of uninsured.
- Working_wealthy

Cluster 3— Low poverty, low elderly, large % of uninsured.
- Working_mid

Cluster 4— Mid poverty, low elderly, all ranges of health insurance.
- Working_poor

Cluster 5— High poverty, low elderly, low % of uninsured.
- Young_poor

Cluster 6— Low poverty, high elderly, low % of uninsured.
- Retired_wealthy

All clusters appear to be significant to positive confirmations.

Only clusters 1 (retired, midincome), 2(working wealthy), and 5(Young poor) appear to be significant to hospitalization rates.

Hospitalization Cluster Data

The GLM Procedure – Pct_POSITIVE

Class Level Information

Class	Levels	Values
CLUSTER	6	1 2 3 4 5 6

Number of Observations Read	800
Number of Observations Used	800

The GLM Procedure

Dependent Variable: Pct_POSITIVE

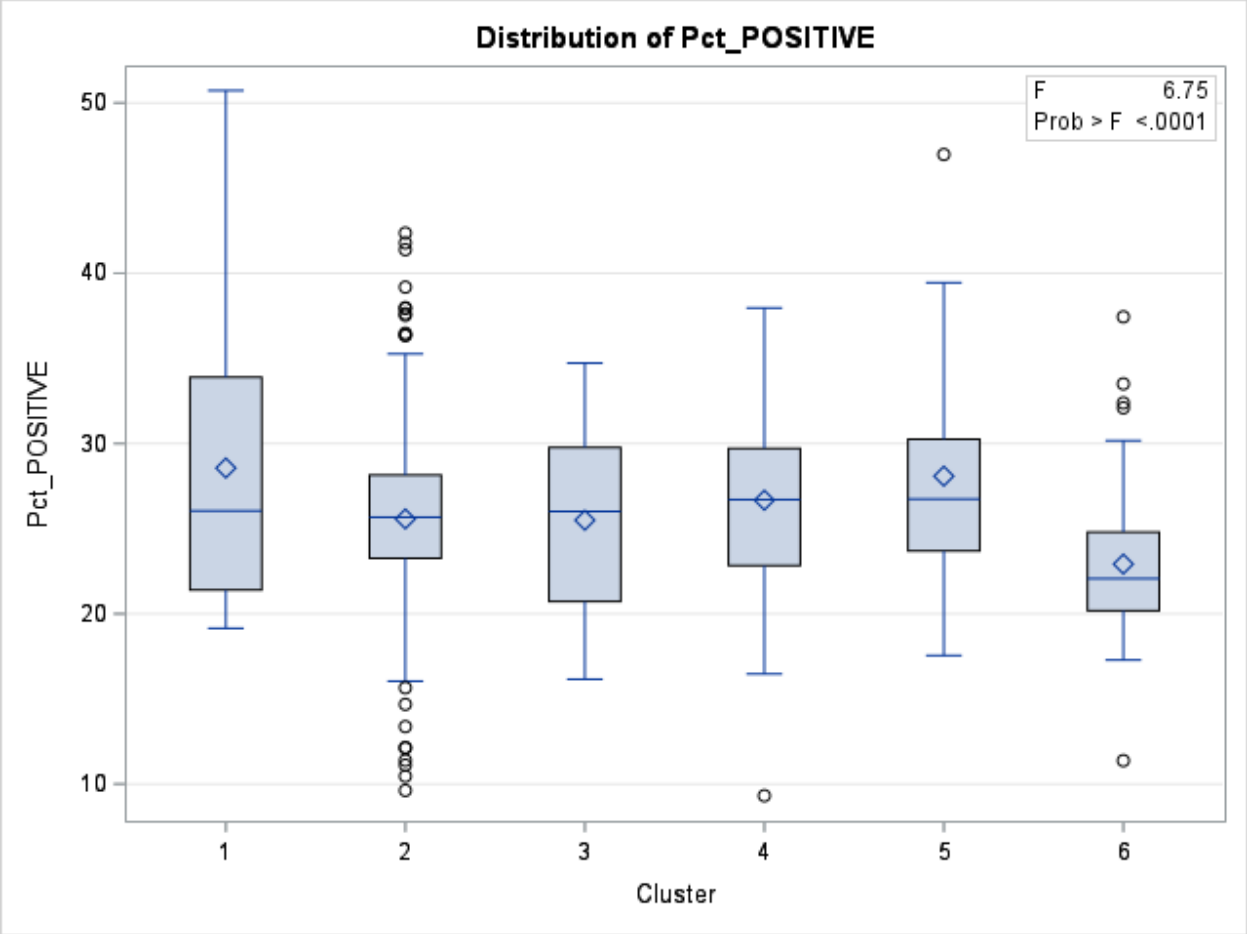
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	753.11263	150.62253	6.75	<.0001
Error	794	17718.86479	22.31595		
Corrected Total	799	18471.97742			

R-Square	Coeff Var	Root MSE	Pct_POSITIVE Mean
0.040771	18.49934	4.723976	25.53592

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CLUSTER	5	753.1126314	150.6225263	6.75	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CLUSTER	5	753.1126314	150.6225263	6.75	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	22.91886007	B	0.56869990	40.30	<.0001
CLUSTER 1	5.64117908	B	1.87387634	3.01	0.0027
CLUSTER 2	2.65105283	B	0.60260076	4.40	<.0001
CLUSTER 3	2.57528936	B	0.85391108	3.02	0.0026
CLUSTER 4	3.75182222	B	0.75231925	4.99	<.0001
CLUSTER 5	5.16397891	B	1.34578959	3.84	0.0001
CLUSTER 6	0.00000000	B	.	.	.


Distribution of Pct_POSITIVE
F 6.75
Prob > F <.0001
Pct_POSITIVE
50
40
30
20
10
1
2
3
4
5
6
Cluster

The GLM Procedure—Pct_HOSP_POSITIVE

Class Level Information

Class	Levels	Values
CLUSTER	6	1 2 3 4 5 6

Number of Observations Read	800
Number of Observations Used	800

Dependent Variable: Pct_HOSP_POSITIVE

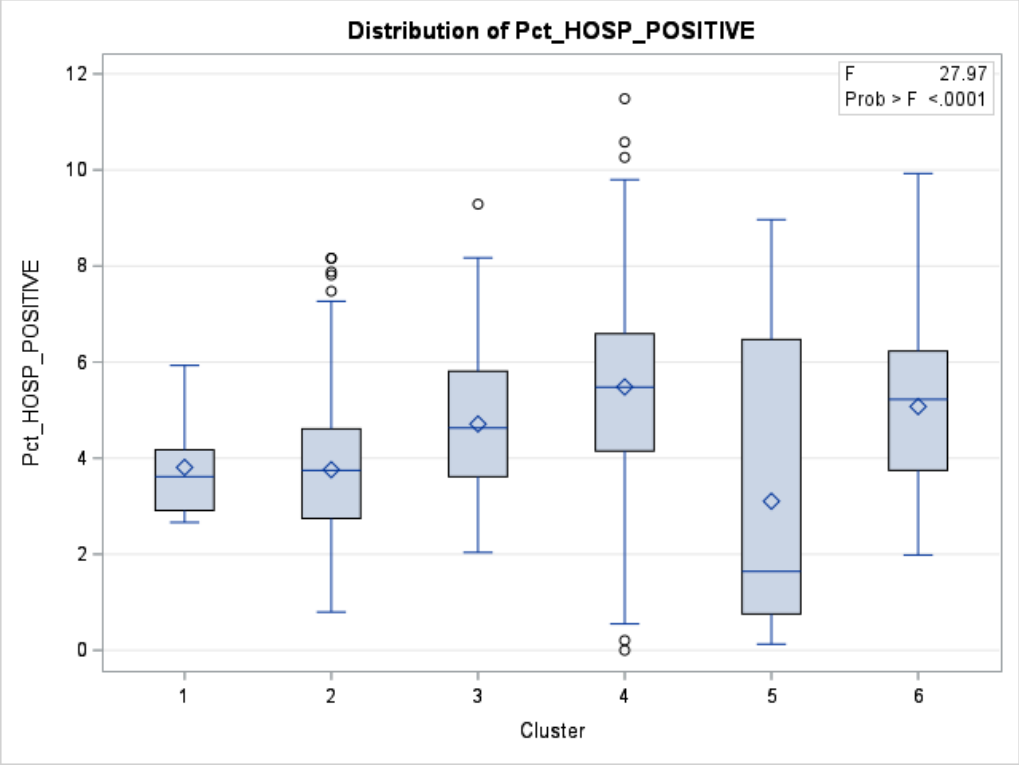
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	342.427771	68.485554	27.97	<.0001
Error	794	1944.083611	2.448468		
Corrected Total	799	2286.511382			

R-Square	Coeff Var	Root MSE	Pct_HOSP_POSITIVE Mean
0.149760	37.94295	1.564758	4.123976

Source	DF	Type I SS	Mean Square	F Value	Pr > F
CLUSTER	5	342.4277710	68.4855542	27.97	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
CLUSTER	5	342.4277710	68.4855542	27.97	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	5.074678206	B	0.18837474	26.94	<.0001
CLUSTER 1	-1.266949348	B	0.62069815	-2.04	0.0416
CLUSTER 2	-1.316076939	B	0.19960398	-6.59	<.0001
CLUSTER 3	-0.363434094	B	0.28284739	-1.28	0.1992
CLUSTER 4	0.407617380	B	0.24919636	1.64	0.1023
CLUSTER 5	-1.971331459	B	0.44577600	-4.42	<.0001
CLUSTER 6	0.000000000	B	.	.	.


Distribution of Pct_HOSP_POSITIVE
F 27.97
Prob > F <.0001
Pct_HOSP_POSITIVE
12
10
8
6
4
2
0
1
2
3
4
5
6
Cluster

1) Training Data Models

R^2 Comparisons

Model	R^2	Adj. R^2	Model
1	.023	.022	POP_HEALTHINS
2	.265	.255	Custom
3	.277	.252	All
4	.273	.261	Stepwise
5	.276	.262	Adj. R2

Custom Model 2=

Log(pct_positive) = b0 + b1(ADULT_LIMITED_ENGLISH) +b2(ADULT_SPANISH_LENG) +b3(POP_MEDICARE) +b4(POP_65P) +b5(POP_DISABILITY) +b6(AREA_LAND) +b7(cluster2) +b8(sizeclass1) +b9(sizeclass2) +e

The MEANS Procedure

Variable	N	Mean
rmse1	159	0.1452
rmse2	159	0.1265
rmse3	159	0.1291
rmse4	159	0.1282
rmse5	159	0.1283
mse1	159	0.0419
mse2	159	0.0332
mse3	159	0.0338
mse4	159	0.0341
mse5	159	0.0335
mae1	159	0.1452
mae2	159	0.1265
mae3	159	0.1291
mae4	159	0.1282
mae5	159	0.1283
mpe1	159	0.0472
mpe2	159	0.0409

Variable	N	Mean
mpe3	159	0.0416
mpe4	159	0.0414
mpe5	159	0.0414

Test Data Models

The MEANS Procedure

Variable	N	Mean
rmse1	640	0.1502
rmse2	640	0.1344
rmse3	640	0.1401
rmse4	640	0.1415
rmse5	640	0.1372
mse1	640	0.0395
mse2	640	0.0323
mse3	640	0.0351
mse4	640	0.0360
mse5	640	0.0336
mae1	640	0.1502
mae2	640	0.1344
mae3	640	0.1401
mae4	640	0.1415
mae5	640	0.1372
mpe1	640	0.0475
mpe2	640	0.0426
mpe3	640	0.0444
mpe4	640	0.0447
mpe5	640	0.0435

According to the means procedure of the testing data, the best model is Model 2, the custom model.

2)

Training Data

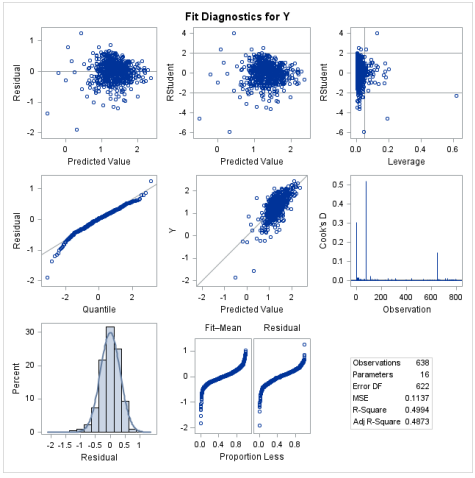
R^2 Comparisons

Model	R^2	Adj. R^2	Model
1	.011	.012	POP_BELOWPOV
2	.484	.477	Custom
3	.5	.483	All
4	.495	.486	Stepwise
5	.487	.499	Adj. R2

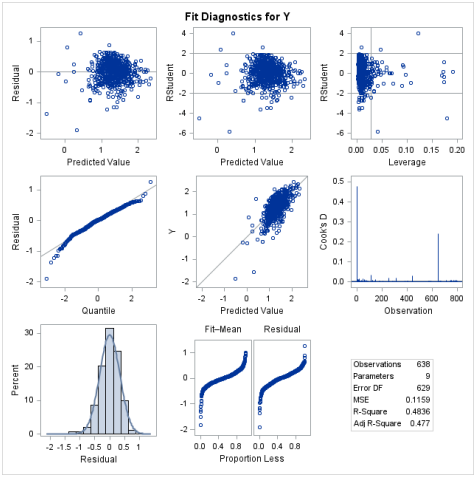
Custom Model =

log(pct_hosp) = b0 + b1(POP_MEDICARE) +b2(POP_DISABILITY) +b3(cluster1) +b4(HOUS_NOINTERNET) +b5(HOUS_NOSMARTPHN) b6(HOUS_NO_VEH) +b7(POP_BELOWPOV) +b8(POP_MEDICAD) + e

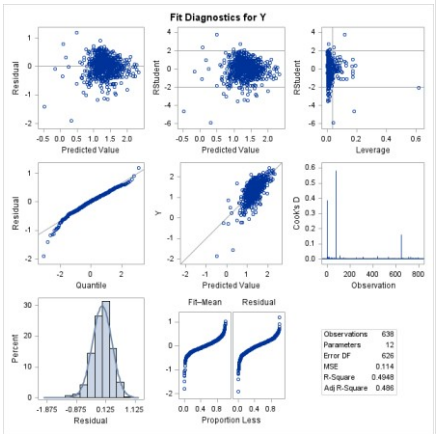
Model 1



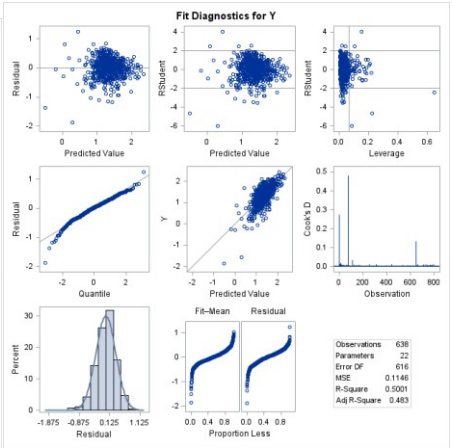
Model 2



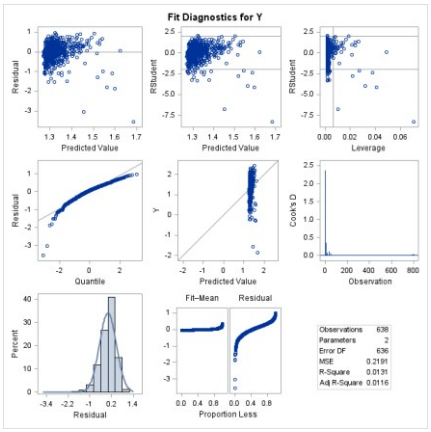
Model 3



Model 4



Model 5



Variable	N	Mean
rmse1	159	0.1452
rmse2	159	0.1265
rmse3	159	0.1291
rmse4	159	0.1282
rmse5	159	0.1283
mse1	159	0.0419
mse2	159	0.0332
mse3	159	0.0338
mse4	159	0.0341
mse5	159	0.0335
mae1	159	0.1452
mae2	159	0.1265
mae3	159	0.1291
mae4	159	0.1282
mae5	159	0.1283
mpe1	159	0.0472
mpe2	159	0.0409
mpe3	159	0.0416
mpe4	159	0.0414
mpe5	159	0.0414

Testing Data

Model	R^2	Adj. R^2	Model
1	.035	.029	POP_BELOWPOV
2	.632	.613	Custom
3	.695	.649	All
4	.672	.652	Stepwise
5	.686	.661	Adj. R2

The MEANS Procedure

Variable	N	Mean
rmse1	639	1.9198
rmse2	639	1.9009
rmse3	639	1.8728
rmse4	639	1.8946
rmse5	639	1.8784
mse1	639	3.7375
mse2	639	3.7758
mse3	639	3.6814
mse4	639	3.7424
mse5	639	3.7042
mae1	639	1.9198
mae2	639	1.9009
mae3	639	1.8728
mae4	639	1.8946
mae5	639	1.8784
mpe1	639	0.5942
mpe2	639	0.5897
mpe3	639	0.5814
mpe4	639	0.5881
mpe5	639	0.5832

According to the means procedure on the test data, model 3 (all variables) is the best fit.