Increasing Sales of Orthopedic Equipment

Written By: Amber Che, Lu Shi , Michael Wang , Zhiqian Li, and Zhixuan Cheng

December 22, 2018

Executive Summary

We were able to find that the following twelve hospitals (identified by ID number and location) could maximize potential sales gains totaling $858,196.77: 224093 (Oakland, CA); 160022 (Voorhees, NJ); 095593 (Hemet, CA); 024039 (Fort Mye, FL); 053039 (Melbourn, FL); 068091 (Seattle, WA); 060591 (Seattle, WA); 383021 (Oneonta, NY); 006839 (Cape Cor, FL); 151023 (New Cast, PA); 037091 (Longview, WA); 135222 (Stratfor, NJ).

I. Introduction

In this project, our job is to help a company that sells orthopedic equipment to increase sales. We did this by identifying a list of hospitals (that the company currently has zero sales in) that we believe will maximize gains if the company targeted them. First, we chose a subset of the states so that we got a subset of about 2000 hospitals from the 4703 total. Each hospital has various descriptive variables and our response variable, which is sales of rehab equipment. We then transformed the variables using either square root or logarithmic transformations. For factor analysis, we divided the variables into three groups: response, demographics, and operation numbers. Using all this, we performed cluster analysis to group hospitals with similar characteristics together. We identified two clusters that each satisfied two conditions: one, it had high average sales, and two, it contained hospitals with zero sales. Finally, we analyzed the "zero-sale hospitals" inside these "high-sales clusters" using regression analysis to calculate their predicted gains if targeted for sales; this is because these hospitals are predicted to have the highest possible gains in sales.

II. Results

A. Selecting a Market Subset

Based on the project guidelines, we selected a subset of about 2000 hospitals out of 4703 total hospitals; these hospitals were selected from the ten states of New Jersey, New York, Connecticut, Pennsylvania, Ohio, Virginia, Florida, Texas, California, and Washington. The exact total is 2019 different hospitals from these ten states.

B. Transformations

To make the relationship between the explanatory variables and the response variable (sales) about linear, we performed transformations. For the explanatory variables OUTV, ADM, and SIR, we used log transformations, while for the explanatory variables BEDS, RBEDS, HIP95, KNEE95, HIP96, KNEE96, and FEMUR96, we used square root transformations. The response variable was transformed using a log transformation. Comparing the scatterplots in graphs 1a (original) and 1b (transformed) show that the transformations helped the data take on a linear trend.

C. Dimension Reduction/Factor Analysis

Factor analysis was used to divide variables into three categories, namely Response, Demographics, and Operation Numbers. The first category is only one variable which does not need to be analyzed, and the second variable includes BEDS, RBEDS, OUTV, ADM, SIR, TH, TRAUMA, REHAB. After factor analysis, dimension reduction is reduced to three factors. The third variable Operation Numbers includes HIP95, KNEE95, HIP96, KNEE96, FEMUR96, and dimensionality reduction as one factor (as shown in table 2). In total, there are four factors. From this, we can see that the four factors screened out by factor analysis can well represent the characteristics of the original sequence and facilitate regression analysis by dimensionality reduction. The method of factor rotation is orthogonal rotation. As shown in table 3, by means of orthogonal rotation with the largest variance (ROTATE= VARIMAX), we found that factor 1, or the factor created based off operation variables, had the greatest effect on the variance with a value of 4.6559005. This suggests that the number of medical operations at a hospital influences our company's sales of orthopedic equipment the most compared to other non-operation variables such as the number of beds or number of outpatient visits.

D. Cluster Analysis

In table 4, Ward's Analysis shows that there is a big jump in the Semipartial R-square between clusters 13 and 14, going from 0.0150 to 0.0117. Thus, we chose 14 clusters for our analysis. Then, we analyzed the boxplots for sales by each cluster to identify which clusters have mostly high sales but also have some hospitals with zero sales. As graph 5 shows, clusters 8 and 13 appear to have a higher mean sales than most of the other clusters but also hospitals with low sales as well; thus, we chose clusters 8 and 13 for further analysis. This is corroborated with the exact mean sales shown in table 5, with the mean sales after transformation being 4.64804 and 4.86749, respectively. The cluster with the highest mean sales is actually cluster 14, but it is very small with only four hospitals, while clusters 8 and 13 consist of 120 and 37 hospitals, respectively.

E. Regression Analysis

Finally, we performed regression analysis for clusters 8 and 13. First, the backwards elimination procedure determined which factors were significant and eliminated the rest; as shown in table 6 and 7, factors 2-4 were eliminated for cluster 8 and all four factors were eliminated for cluster 13. When performing regression analysis, we sorted the hospitals by residual to look for hospitals with the largest negative residuals (because these hospitals have the lowest sales compared to the mean sales and thus have the most potential for increased sales) and took note of their predicted gains.

For cluster 8, we discovered multiple hospitals with no sales but large negative residuals, so we took the top eight hospitals from this cluster with the highest gains. For cluster 13, we discovered four hospitals with no sales. Together, we identified twelve different hospitals that

would maximize gains if the company concentrated their efforts on them. From cluster 8: 224093 (Oakland, CA); 160022 (Voorhees, NJ); 095593 (Hemet, CA); 383021 (Oneonta, NY); 006839 (Cape Cor, FL); 151023 (New Cast, PA); 037091 (Longview, WA); 135222 (Stratfor, NJ). From cluster 13: 024039 (Fort Mye, FL); 053039 (Melbourn, FL); 068091 (Seattle, WA); 060591 (Seattle, WA). The projected gain in all is $858,196.77 as shown in table 10.

F. (Extra Credit) R Analysis

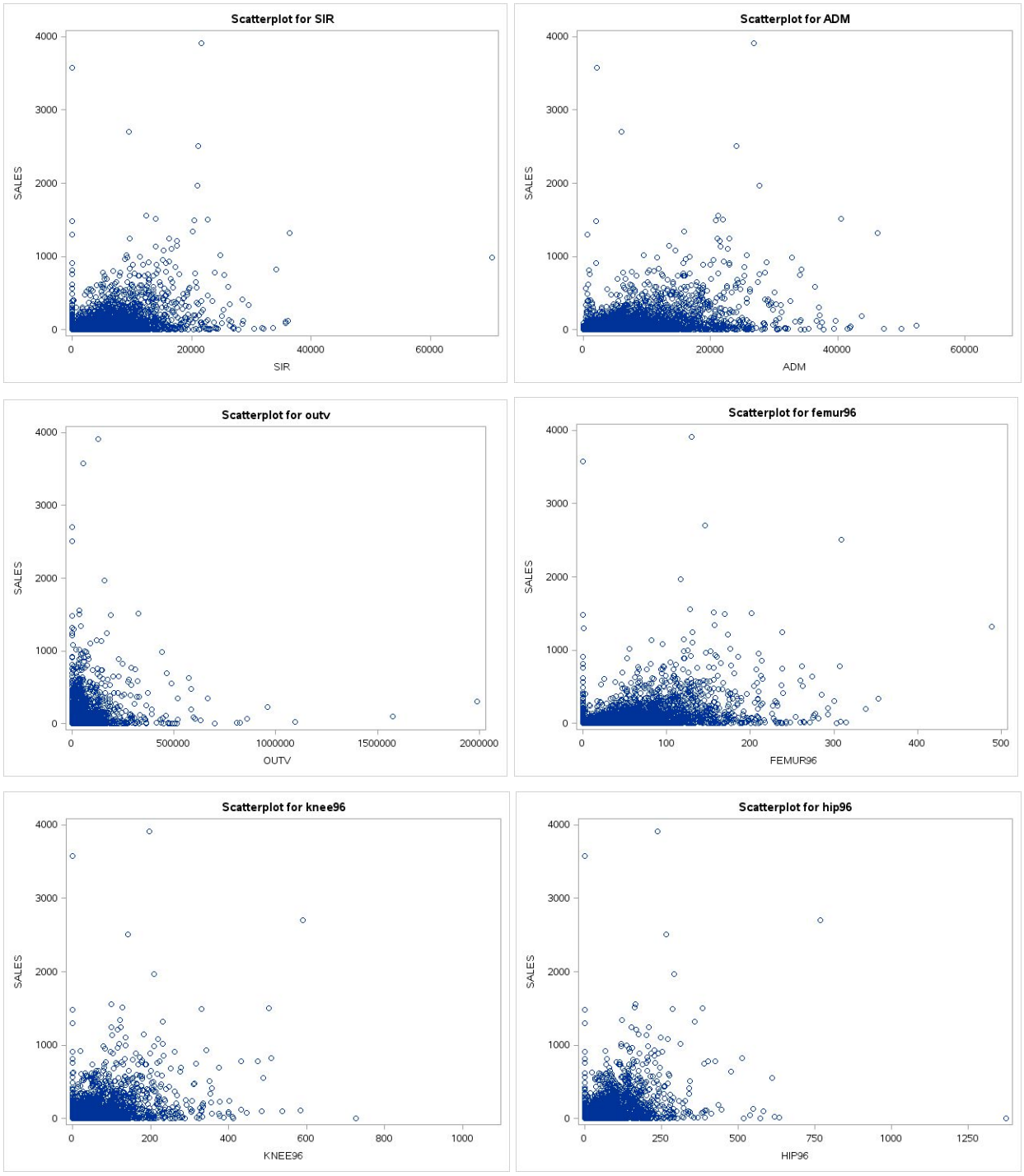
We repeated our analysis using R methods for robust clustering (pam) and for classification and regression trees (rpart). Please see the second submitted pdf.

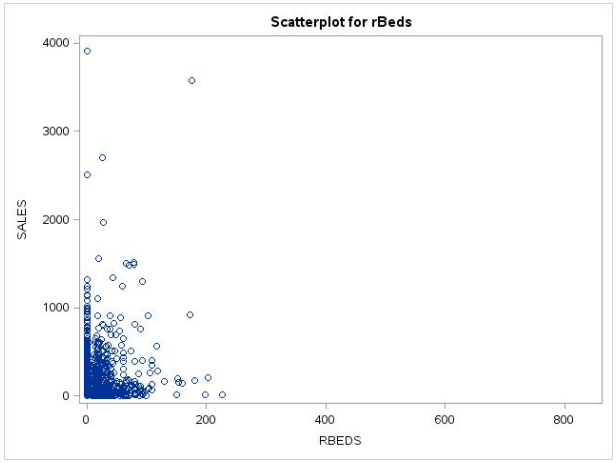
III. Conclusion

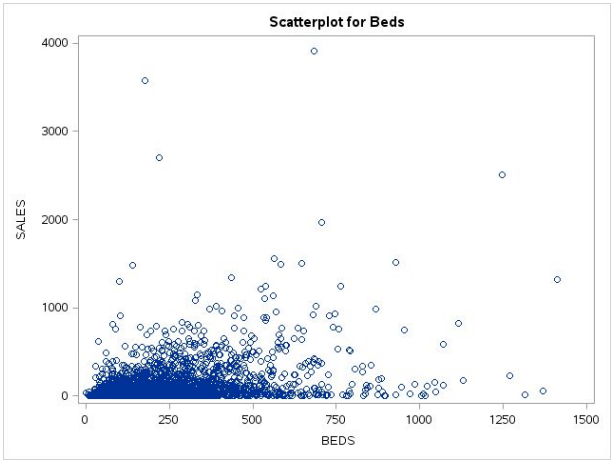
In this project, we used a ten-state subset of the list of hospitals from all over the United States to perform cluster analysis (to group similar hospitals together and identify high-sales clusters with zero-sale hospitals) and regression analysis (to determine predicted gains in sales if the company focused its efforts to sell orthopedic equipment to these zero-sale hospitals). Based on our results, we identify twelve hospitals that had the highest predicted gain in sales, as they would likely be the best possible customers for orthopedic equipment, for a gain of $858,196.77 in sales. We also found that the most important factor affecting sales is the number of operations at the hospital. Overall, this insightful information will help the company increase its orthopedic equipment sales.

Appendix

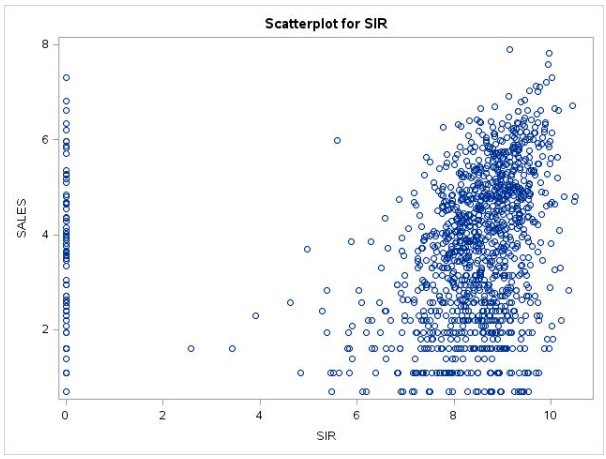
Graph 1a: Scatterplots for Original Variables

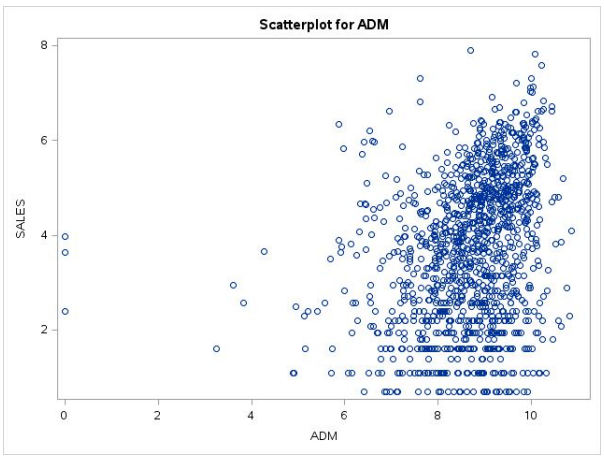


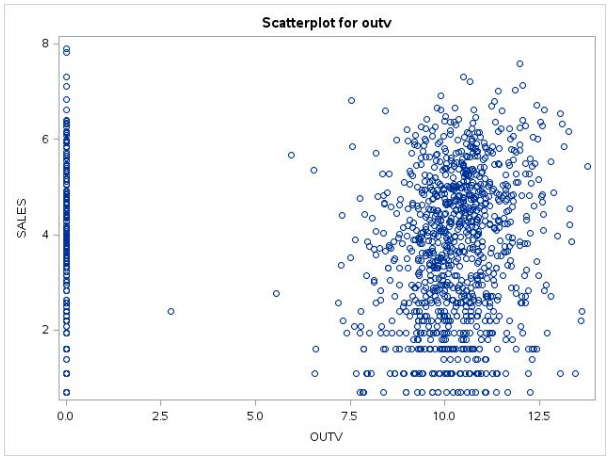


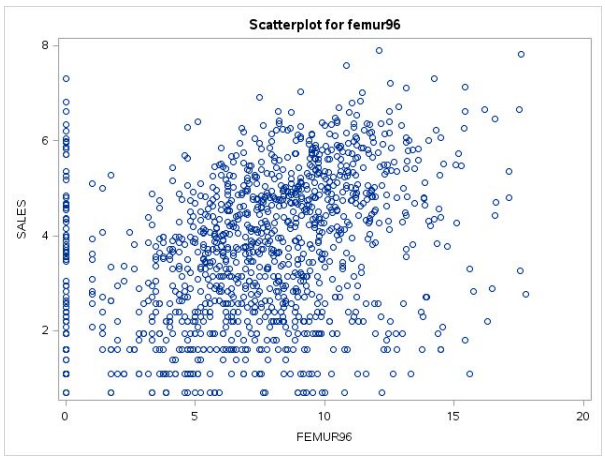


Graph 1b: Scatterplots for Transformed Variables









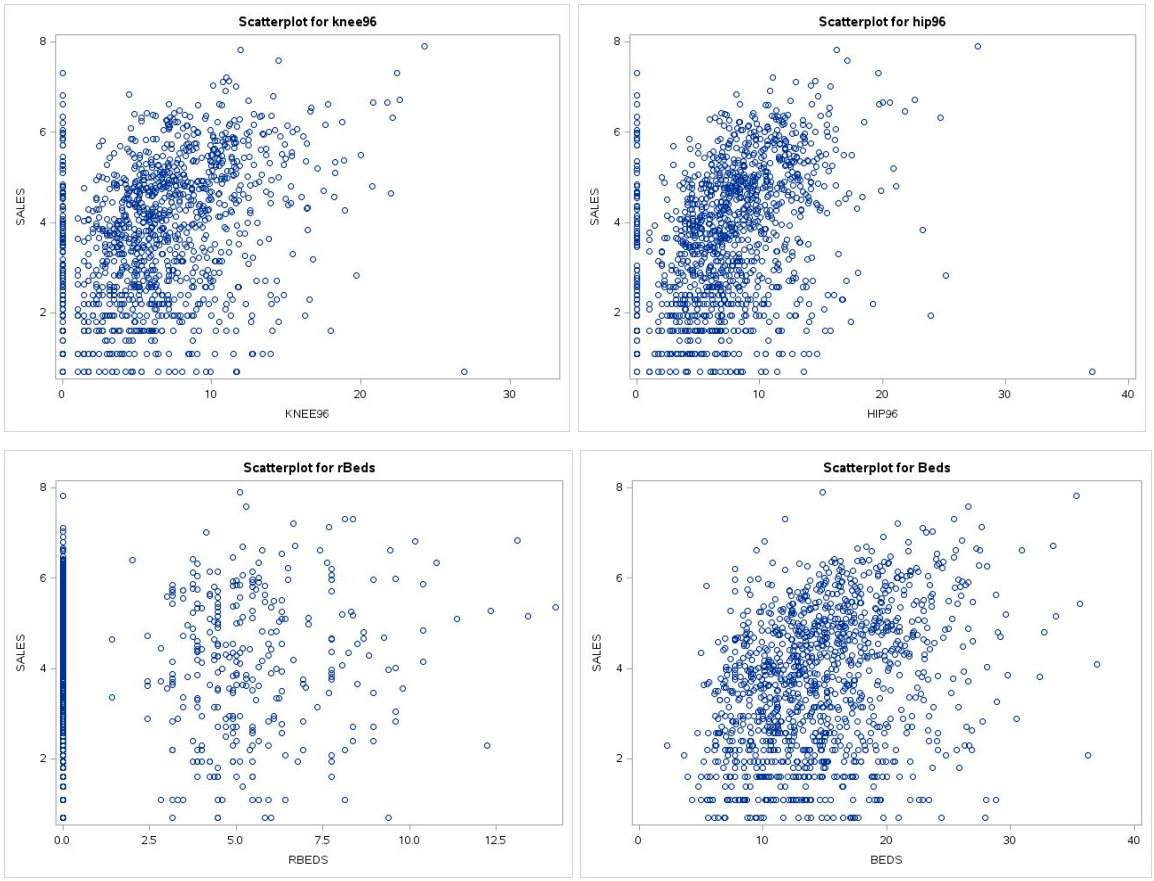


Table 2: Factor Analysis/Eigenvalues for Operation (left) and Demographic variables

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	2.84207847	1.74709750	0.5684	0.5684
2	1.09498097	0.27289109	0.2190	0.7874
3	0.82208987	0.63727562	0.1644	0.9518
4	0.18481425	0.12877781	0.0370	0.9888
5	0.05603644		0.0112	1.0000

1 factor will be retained by the NFACTOR criterion.

Factor Pattern

	Factor1
HIP95	0.39973
KNEE95	0.03508
HIP96	0.96494
KNEE96	0.93184
FEMUR96	0.93895

Variance Explained by Each Factor

Factor1
2.8420785

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 8 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	3.55135280	2.01961526	0.4439	0.4439
2	1.53173754	0.46096343	0.1915	0.6354
3	1.07077411	0.24986836	0.1338	0.7692
4	0.82090575	0.29216157	0.1026	0.8718
5	0.52874418	0.26927385	0.0661	0.9379
6	0.25947033	0.11339512	0.0324	0.9704
7	0.14607521	0.05513515	0.0183	0.9886
8	0.09094007		0.0114	1.0000

3 factors will be retained by the NFACTOR criterion.

Factor Pattern

	Factor1	Factor2	Factor3
BEDS	0.82117	0.21063	0.30422
RBEDS	-0.13851	0.80182	0.28053
OUTV	0.33589	-0.45936	0.24958
ADM	0.85275	-0.20341	0.20403
SIR	0.75954	-0.52179	0.01570
TH	0.78275	0.30481	-0.47988
TRAUMA	0.73861	0.31848	-0.52767
REHAB	0.53169	0.35423	0.53539

Variance Explained by Each Factor

Factor1	Factor2	Factor3
3.5513528	1.5317375	1.0707741

Table 3: Rotated Factor Analysis using all Four Factors

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix				
	1	2	3	4
1	0.84126	0.45915	-0.11446	0.26145
2	0.07277	0.24026	0.93613	-0.24626
3	-0.53562	0.74635	-0.04672	0.39229
4	0.00955	-0.41764	0.32922	0.84682

Rotated Factor Pattern				
	Factor1	Factor2	Factor3	Factor4
BEDS	0.49512	0.69402	0.07024	0.20712
RBEDS	-0.02586	0.09473	0.96713	-0.08455
OUTV	0.05148	-0.01253	-0.02041	0.92056
ADM	0.48526	0.56099	-0.22703	0.41832
SIR	0.45043	0.36369	-0.49275	0.47160
TH	0.93089	0.13341	0.04030	-0.02949
TRAUMA	0.91600	0.06776	0.06317	-0.04907
REHAB	0.13736	0.77565	0.10803	-0.07213
HIP95	0.05875	0.71370	0.08069	0.01938
KNEE95	0.04430	0.12990	0.94538	0.01618
HIP96	0.92132	0.23695	-0.08627	0.16863
KNEE96	0.92541	0.15470	-0.03116	0.14548
FEMUR96	0.73071	0.41955	-0.14790	0.28772

Variance Explained by Each Factor			
Factor1	Factor2	Factor3	Factor4
4.6559005	2.3441443	2.1828576	1.4363792

Table 4: Ward's Analysis to pick the number of clusters

20	CL45	CL53	120	0.0056	.865	
19	CL26	CL58	160	0.0066	.859	
18	CL48	CL50	72	0.0067	.852	
17	CL47	CL39	147	0.0079	.844	
16	CL25	CL28	86	0.0085	.835	
15	CL35	CL40	51	0.0110	.824	
14	CL17	CL27	217	0.0117	.813	
13	CL24	CL18	270	0.0150	.798	
12	CL29	CL38	132	0.0163	.781	
11	CL30	CL20	497	0.0170	.764	
10	CL15	CL19	211	0.0176	.747	

Graph 2: Boxplots of Sales by Cluster to identify clusters with mostly high sales

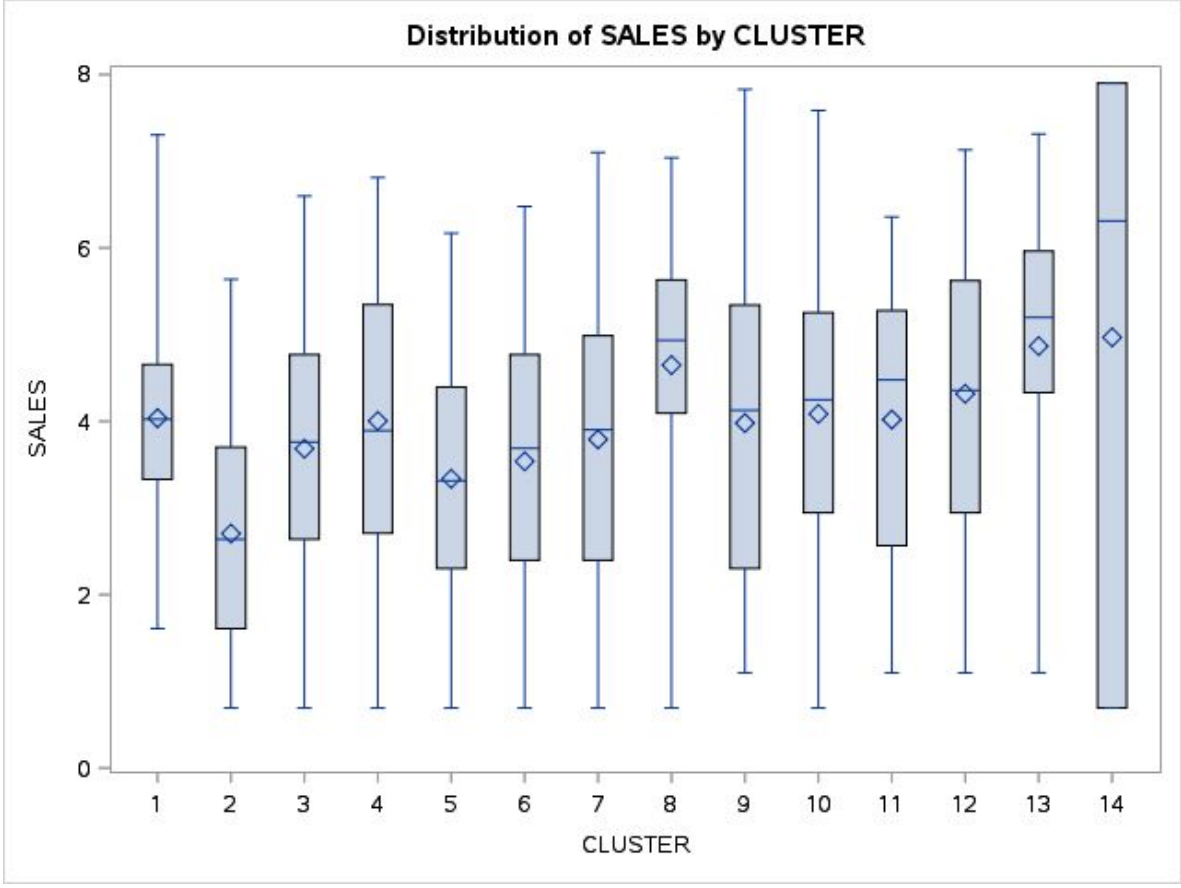


Table 5: Mean Sales by Cluster to identify best clusters for analysis

Obs	CLUSTER	_TYPE_	_FREQ_	msales	mf1	mf2	mf3	mf4
1	1	0	80	4.03546	-0.81184	-0.56457	2.33050	-2.17952
2	2	0	510	2.70467	-0.63797	-0.63007	-0.42412	0.23063
3	3	0	377	3.68345	0.18635	-0.47259	-0.48381	0.56834
4	4	0	52	4.00308	-1.00808	-0.77796	2.72685	-0.19661
5	5	0	160	3.33676	-0.10632	-0.22277	-0.74374	-1.72695
6	6	0	198	3.53928	-0.31459	0.91080	-0.52025	0.39028
7	7	0	86	3.79170	-0.16386	1.74372	-0.63815	-1.85572
8	8	0	120	4.64804	1.35936	-0.20684	-0.43200	0.33376
9	9	0	51	3.97926	1.28346	0.50858	0.32549	-1.46935
10	10	0	217	4.08478	0.32103	0.06808	1.45661	0.96184
11	11	0	72	4.01921	-0.05901	2.38873	-0.55387	0.17554
12	12	0	55	4.31720	0.28423	2.47783	1.52235	0.79522
13	13	0	37	4.86749	3.30229	-0.69025	0.54613	-0.02708
14	14	0	4	4.96912	8.51412	-2.03899	0.61246	-4.46951

CLUSTER 8

Table 6: Backward Substitution to eliminate factors for Cluster 8

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	30.08056	30.08056	6.59	0.0115
Error	118	539.00502	4.56784		
Corrected Total	119	569.08558			

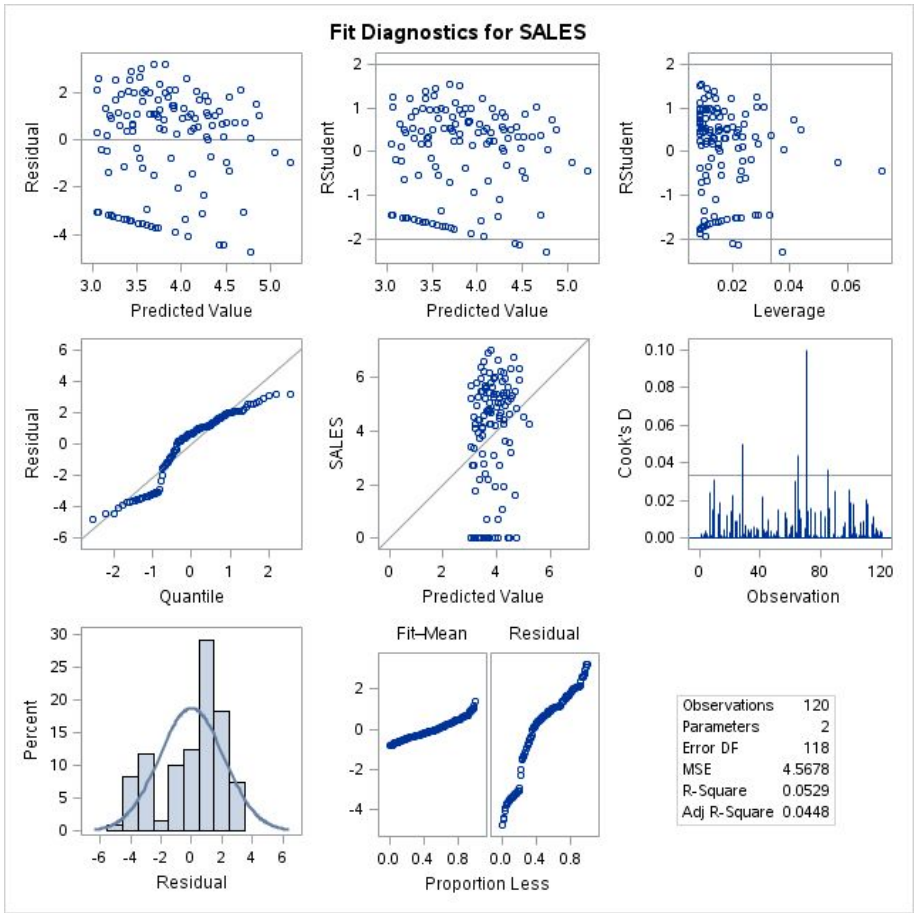
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.17370	0.67600	47.22951	10.34	0.0017
Factor1	1.22184	0.47613	30.08056	6.59	0.0115

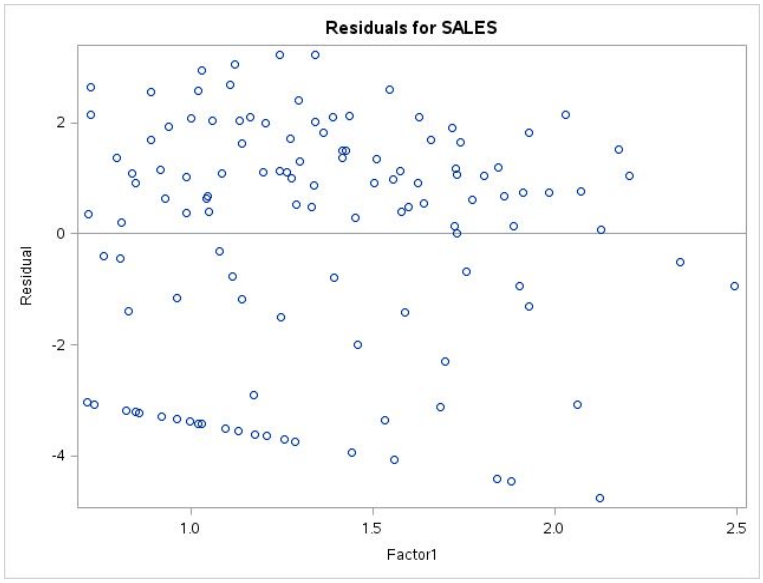
Bounds on condition number: 1, 1

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Factor2	3	0.0004	0.0591	3.0513	0.05	0.8212
2	Factor3	2	0.0021	0.0570	1.3116	0.26	0.6094
3	Factor4	1	0.0041	0.0529	-0.1810	0.51	0.4745

Graph 3: Fit Diagnostics for Cluster 8





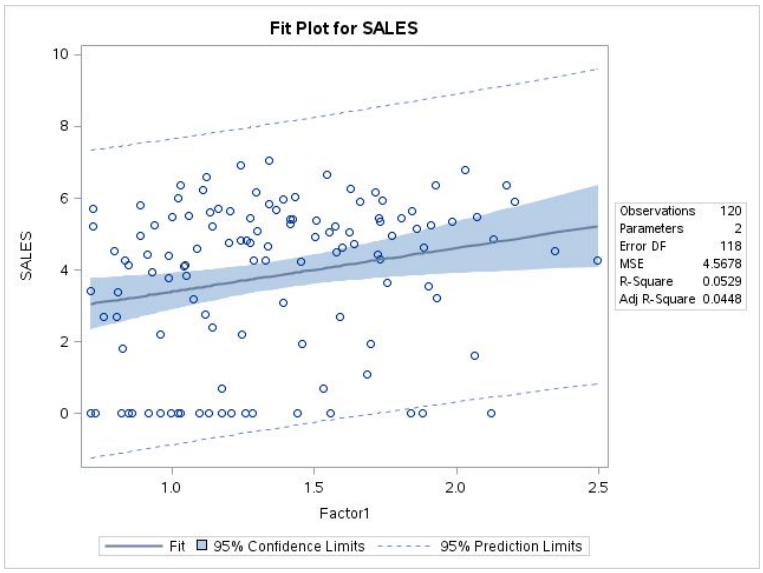


Table 7: Hospitals with largest negative residuals and their gains for Cluster 8

Obs	ZIP	CITY	STATE	HID	SALES	RESID	gain
1	94609	Oakland	CA	224093	0	-4.76769	127.10
2	08043	Voorhees	NJ	160022	0	-4.47208	91.02
3	92543	Hemet	CA	095593	0	-4.42252	86.16
4	13820	Oneonta	NY	383021	0	-4.07735	59.39
5	33990	Cape Cor	FL	006839	0	-3.93497	51.18
6	16105	New Cast	PA	151023	0	-3.74405	42.11
7	98632	Longview	WA	037091	0	-3.71059	40.71
8	08084	Stratfor	NJ	135222	0	-3.65069	38.34
9	33435	Boynton	FL	005039	0	-3.60961	36.81
10	33334	Fort Lau	FL	019539	0	-3.55709	34.94
11	15009	Beaver	PA	010F23	0	-3.51333	33.47

CLUSTER 13

Table 8: Backward Substitution to eliminate factors for Cluster 13

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	36	159.50074	4.43058		
Corrected Total	36	159.50074			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	4.34127	0.34604	697.32528	157.39	<.0001

Bounds on condition number: 0, 0

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Factor4	3	0.0002	0.0304	3.0080	0.01	0.9295
2	Factor3	2	0.0011	0.0294	1.0438	0.04	0.8488
3	Factor2	1	0.0025	0.0268	-0.8722	0.09	0.7670
4	Factor1	0	0.0268	0.0000	-1.9872	0.96	0.3329

Graph 4: Fit Diagnostics for Cluster 13

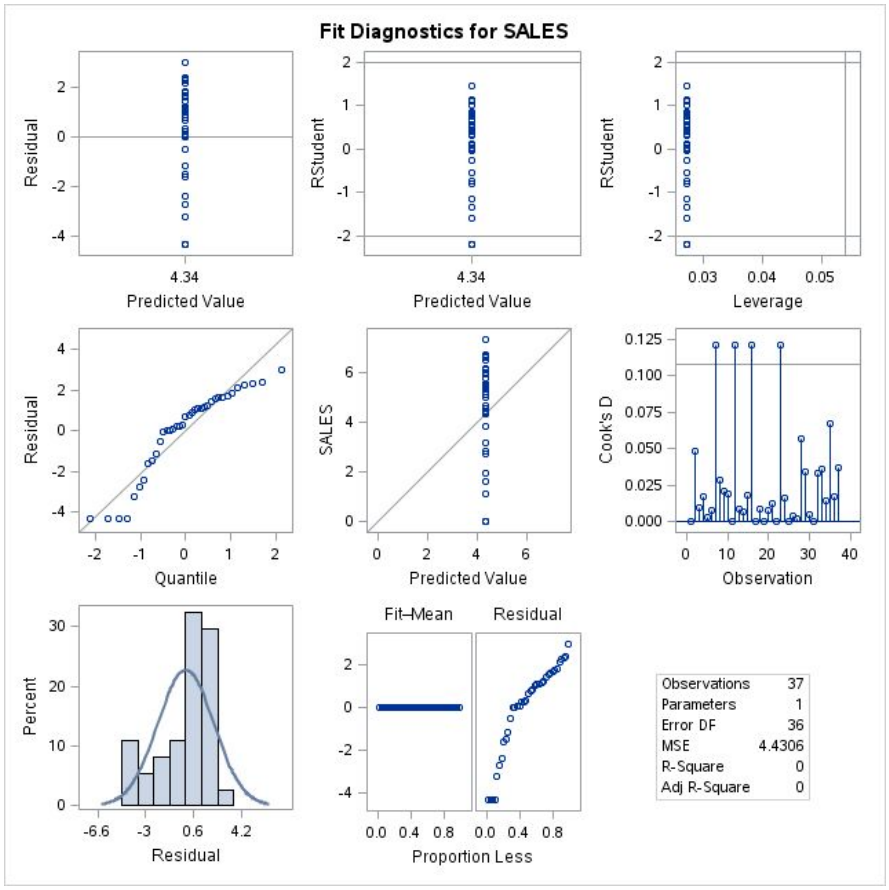


Table 9: Hospitals with largest negative residuals and their gains for Cluster 13

Obs	ZIP	CITY	STATE	HID	SALES	RESID	gain
1	33901	Fort Mye	FL	024039	0	-4.34127	80.54
2	32901	Melbourn	FL	053039	0	-4.34127	80.54
3	98122	Seattle	WA	068091	0	-4.34127	80.54
4	98112	Seattle	WA	060591	0	-4.34127	80.54
5	22206	Arlingto	VA	003534	2	-3.24266	78.54
6	33901	Fort Mye	FL	023839	4	-2.73183	76.54
7	33308	Fort Lau	FL	023039	6	-2.39536	74.54

Table 10: Final List of Hospitals with Potential Gains and Potential Gain Total

Obs	ZIP	CITY	STATE	HID	SALES	RESID	gain
1	94609	Oakland	CA	224093	0	-4.76769	127.102
2	08043	Voorhees	NJ	160022	0	-4.47208	91.016
3	92543	Hemet	CA	095593	0	-4.42252	86.165
4	33901	Fort Mye	FL	024039	0	-4.34127	80.544
5	32901	Melbourn	FL	053039	0	-4.34127	80.544
6	98122	Seattle	WA	068091	0	-4.34127	80.544
7	98112	Seattle	WA	060591	0	-4.34127	80.544
8	13820	Oneonta	NY	383021	0	-4.07735	59.392
9	33990	Cape Cor	FL	006839	0	-3.93497	51.184
10	16105	New Cast	PA	151023	0	-3.74405	42.108
11	98632	Longview	WA	037091	0	-3.71059	40.712
12	08084	Stratfor	NJ	135222	0	-3.65069	38.342

The MEANS Procedure

Analysis Variable : gain
Sum
858.1967675