STAT 586 Final report

Sales prediction of orthopedic equipment in US hospitals

Lina Gao

Dec 20th, 2018

Summary

2500 hospitals in US were randomly selected in order to find the potential clients for a medical equipment company. 18 variables, which described the location, size, operation, function and sales of these hospitals, were provided. After proper data transformation, principle component analysis was applied to reduce the dimension of the data. Four PC, which can explain more than 75% variance in the data, were chosen. PC1 indicated the size of the hospital, PC2 for rehab, PC3 for trauma and PC4 for operations. Then, K-means, hierarchical and pam clustering were applied to segment these hospitals. Elbow, silhouette and gap statistics were used to determine the optimal cluster number. The hospitals with large size, high operation, rehab or trauma function were grouped into the selected clusters. At last, multivariate regression model was developed to predict the sales using PC 1 to 4. ~43 to 200 hospitals were selected as potential clients with estimated total revenues from 1204 to 4852 thousand dollars/year.

1. Introduction

This study investigated the potential markets for the orthopedic equipment from a medical device company. The objective is to increase sales by identifying the hospitals with high demands of these equipment but low sales from this company in US. 2500 hospitals were randomly selected. 18 variables were collected, which described hospital location, hospital size, related service size etc. The details of these variables were listed in table 1.

Table1. The description of variables in this study

Variables	description
ZIP	US POSTAL CODE
HID	HOSPITAL ID
CITY	CITY NAME
STATE	STATE NAME
BEDS	NUMBER OF HOSPITAL BEDS
RBEDS	NUMBER OF REHAB BEDS
OUT-V	NUMBER OF OUTPATIENT VISITS
ADM	ADMINISTRATIVE COST ($1000's per year)
SIR	REVENUE FROM INPATIENT
SALES	SALES OF REHAB. EQUIP ($1000's per year)
HIP	NUMBER OF HIP OPERATIONS
KNEE	NUMBER OF KNEE OPERATIONS
TH	TEACHING HOSPITAL (0, 1)
TRAUMA	DO THEY HAVE A TRAUMA UNIT (0, 1)
REHAB	DO THEY HAVE A REHAB UNIT (0, 1)
HIP2	NUMBER HIP OPERATIONS Year 2
KNEE2	NUMBER KNEE OPERATIONS Year 2
FEMUR2	NUMBER FEMUR OPERATIONS Year 2

Considering the large data size, several statistical analyses were performed to help us find the potential clients, principle component analysis (PCA), cluster analysis and multivariate regression. The entire analysis was described as the flow chart in figure 1. Using PCA and cluster analysis, large size hospitals with trauma or rehabilitation units or that performs orthopedic surgery were selected. Among them, the ones with low sales values from this company were identified as the potential groups. Finally, the expected revenue was calculated using the multivariate regression models.

2. Methods and results

2.1 Data transformation

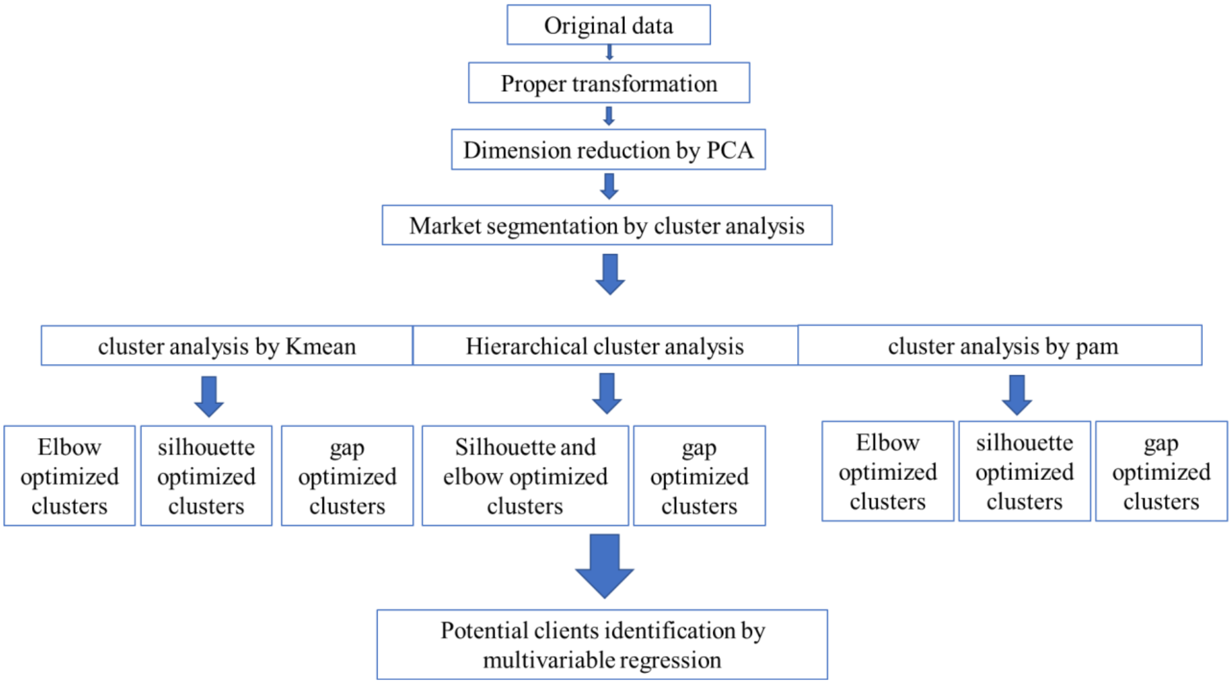
Original data
Proper transformation
Dimension reduction by PCA
Market segmentation by cluster analysis
cluster analysis by Kmean
Hierarchical cluster analysis
cluster analysis by pam
Elbow optimized clusters
silhouette optimized clusters
gap optimized clusters
Silhouette and elbow optimized clusters
gap optimized clusters
Elbow optimized clusters
silhouette optimized clusters
gap optimized clusters
Potential clients identification by multivariable regression

Figure 1. Pipeline of the statistical analysis

Table 2. Data transformation

Variable	Transformation
RBEDS	log(1+0.02*RBEDS)
BEDS	log(1+0.1*BEDS)
OUTV	log(1+0.0001*OUTV
ADM	log(1+0.001*ADM)
SIR	log(1+0.001*SIR)
HIP95	log(1+0.1*HIP95)
KNEE	log(1+0.1*KNEE95)
SALES	log(1+SALES)
HIP96	log(1+0.1*HIP96)
KNEE96	log(1+0.1*KNEE96)
FEMUR	log(1+0.1*FEMUR96)
REHAB	gr = TH*4 + TRAUMA*2 + REHAB
TH	gr = TH*4 + TRAUMA*2 + REHAB
TRAUMA	gr = TH*4 + TRAUMA*2 + REHAB

The single variable distribution showed most variables skewed right in this study. Therefore, proper data transformation was applied, as indicated in table 2. After proper transformation (table 2), the normality of most variables improved greatly (figure 3). The scatterplot between sales and other variables became clear (figure 2).

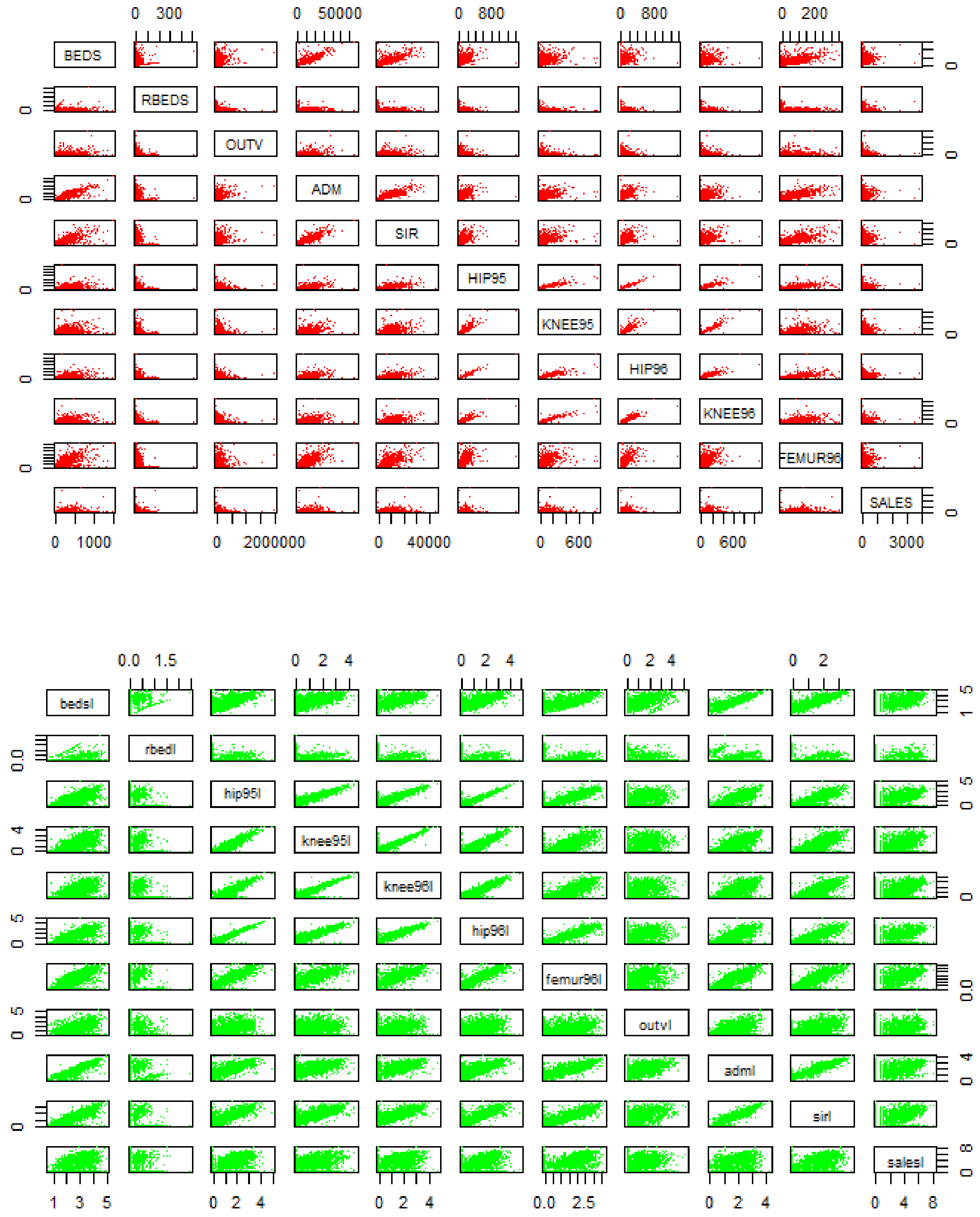

Figure 2. pairwise plot for continuous variables before and after transformation

(red for before transformation, green for after transformation)

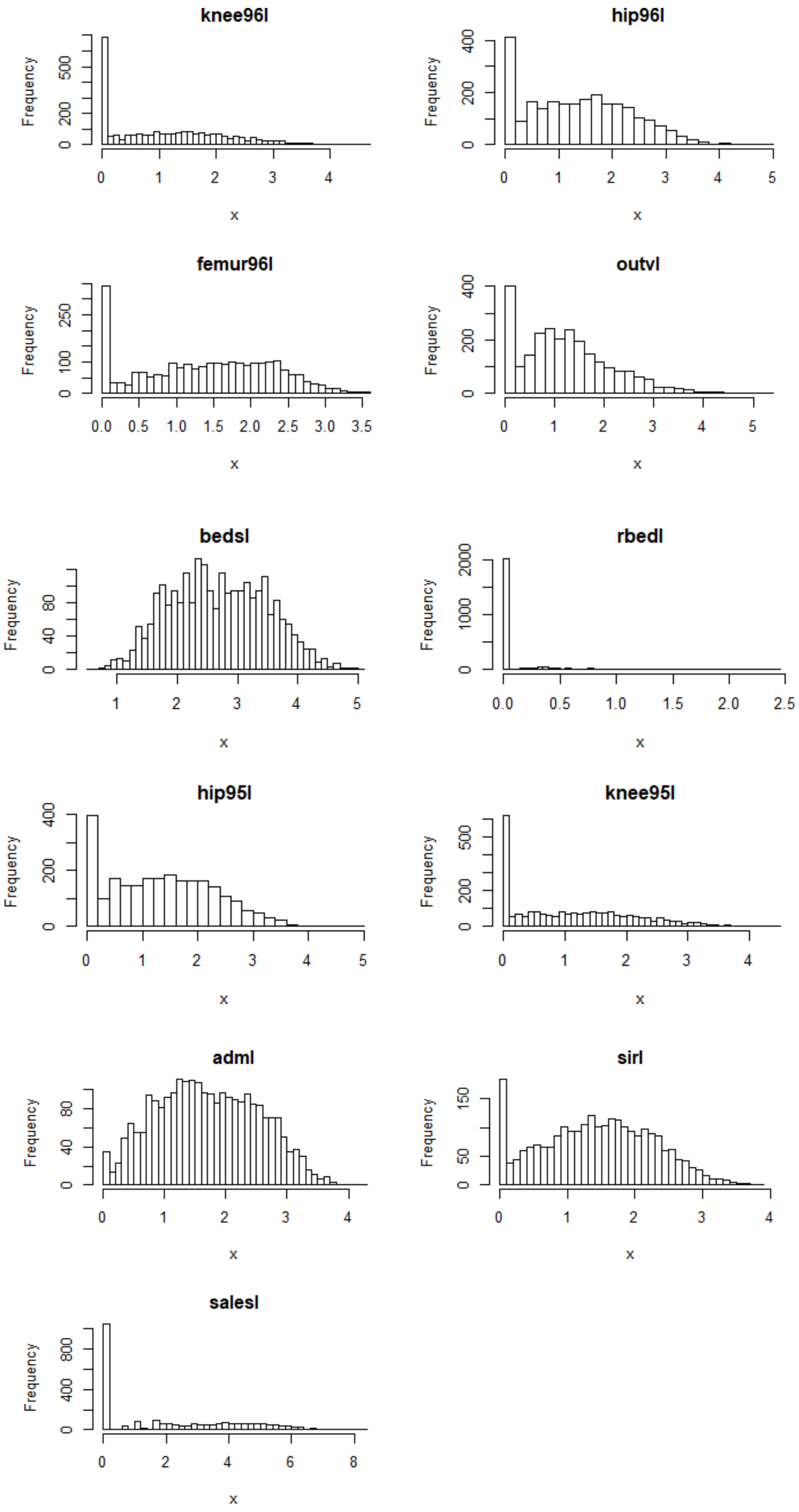

Figure 3. The histogram of the transformed data

2.2 PCA analysis

Except location variables, all the other independent variables can be divided into two categories: demographics and operation numbers. The demographic variables, group1, included BEDS, RBEDS, OUTV, ADM, SIR, TH, TRAUMA and REHAB. The operation variables, group2, included HIP95, KNEE95, HIP96, KNEE96 and FEMUR2. PCA analysis was performed in these two groups individually. The optimized PCA numbers were determined using scree and cumulative scree plots, as shown in figure 4. In group 1, only the first PC, PC1, was chosen because it alone explained more than 75% variance in the data.

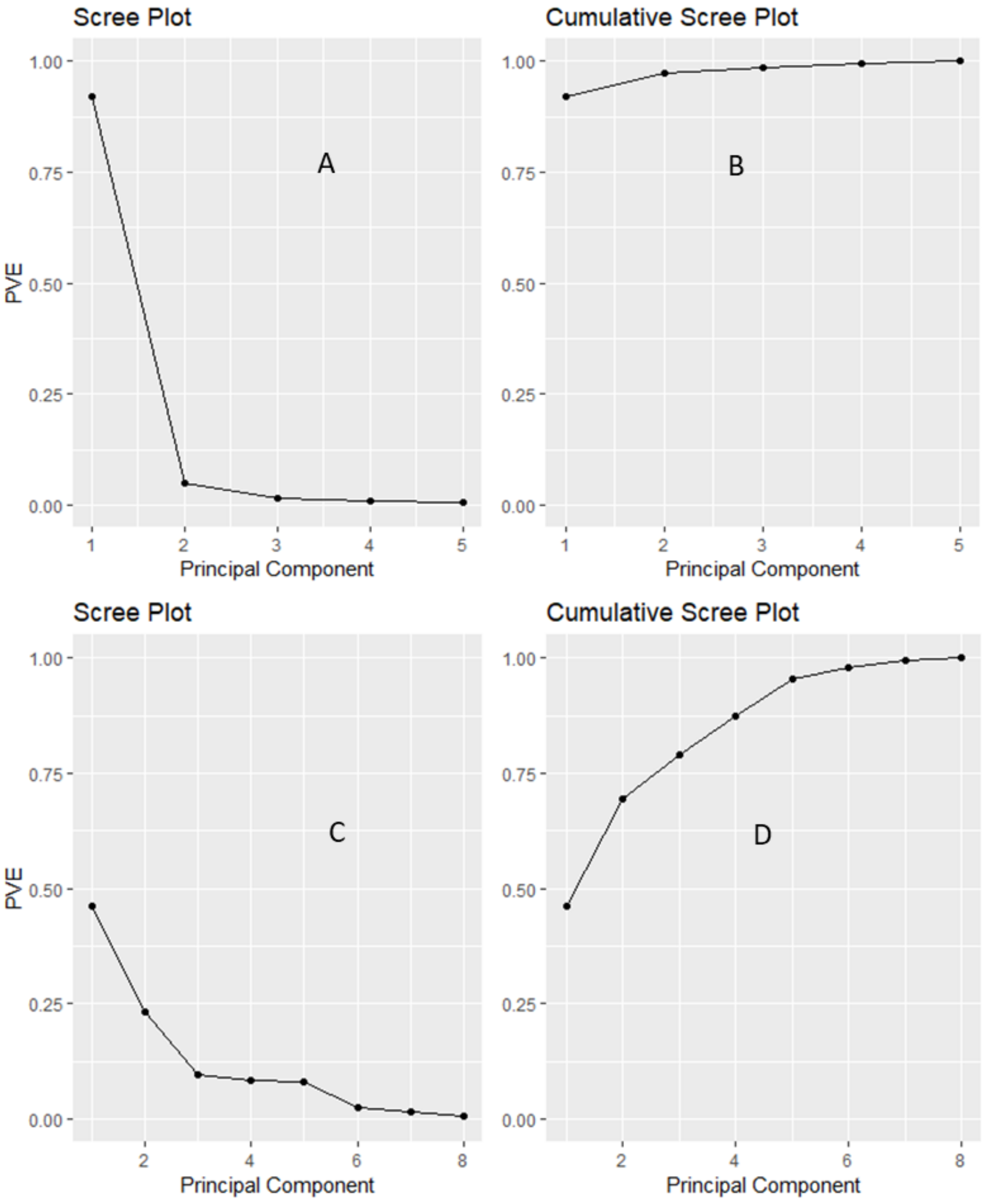

Figure 4. The scree and cumulative scree plot in PCA analysis

(A and B for group 1, C and D for group2)

In group 2, three PC (PC2, PC3, PC4) were chosen. Based on the loadings of these PCs in table 3, PC1 mainly described the average operations. PC2 indicated the hospital size. PC3 was primarily controlled by rehab related variables (rehab and rbedl). TRAUMA was the only dominant player in PC4.

Table3. The loadings of PC 1, 2, 3 and 4

GROUP2	TH	TRAUMA	REHAB	bedsl	rbedl	outvl	adml	sirl
PC1 (PC1)	**0.34**	**0.28**	0.12	**0.48**	0.05	**0.33**	**0.49**	**0.46**
PC2 (PC2)	0.09	0.05	**0.67**	0.01	**0.7**	-0.09	-0.13	-0.18
PC3 (PC3)	-0.01	**0.95**	-0.07	-0.2	-0.07	-0.02	-0.16	-0.15
GROUP1	hip95l	knee95l	knee96l	hip96l	femur96l			
PC1(PC4)	0.46	0.45	0.45	0.46	0.43			

2.3 Cluster analysis

Three cluster techniques were applied for market segmentation, K-means, hierarchical and Partitioning Around Medoids (PAM) clustering. As one of the oldest methods of cluster analysis, K-means is fast and doesn't require calculating all the distances between each observation and every other observation. But it can give different results each time if the order of the data changed. With different target number of clusters, the final clustering results can be totally different. Compared with K-means, hierarchical clustering examines all the distances between all the observations and pairs together the two closest ones to form a new cluster. Thus, observations don't "jump ship" as they do in K-means. PAM is the robust version of K-means using very similar algorithm. Considering their difference, all three methods were applied to comprehensively detect the potential clients for this company.

2.3.1 K-means clustering

With K-means clustering, three methods were applied to determine the optimal cluster number in this study, elbow, silhouette and gap statistic. Elbow uses the total within-cluster sum of square (wss) to measure the compactness of the clustering and plots the curve of was at different number of clusters k. The location of a “knee” in the plot is generally considered as the appropriate number of clusters. Silhouette computes the average silhouette of observations for different k and determines the optimal k value which maximizes the average silhouette over a range. Gap statistic compares the total intra-cluster variation for different k values with their expected values under null reference distribution of the data.

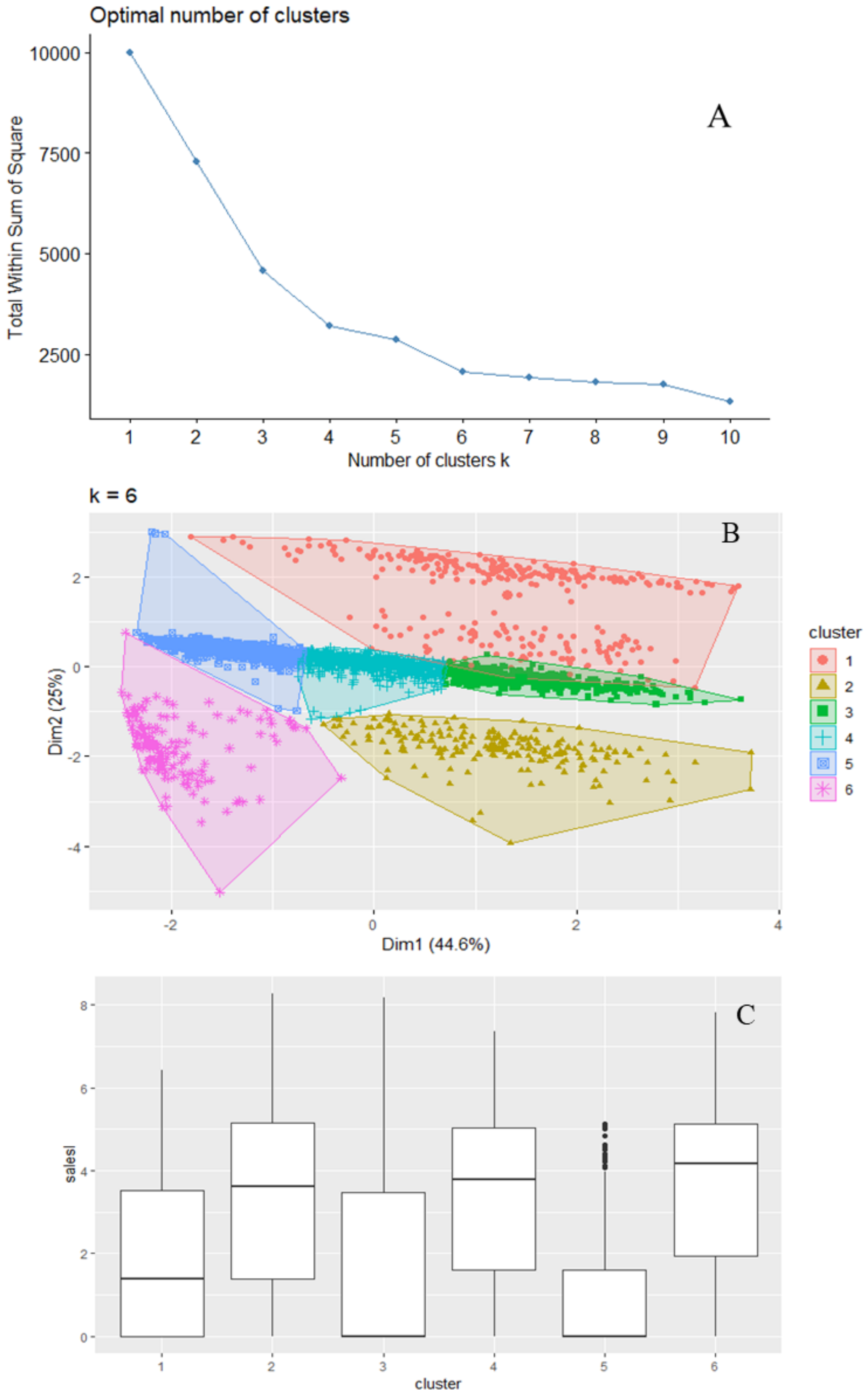

Figure 5. The optimal clusters determined by K-means elbow method

(A for elbow plot, b for cluster demonstration and c for distribution of sales in these clusters)

The reference dataset is generated using Monte Carlo simulations of the sampling process. Thus, different methods probably can give us totally different results.

2.3.1.1 Elbow method

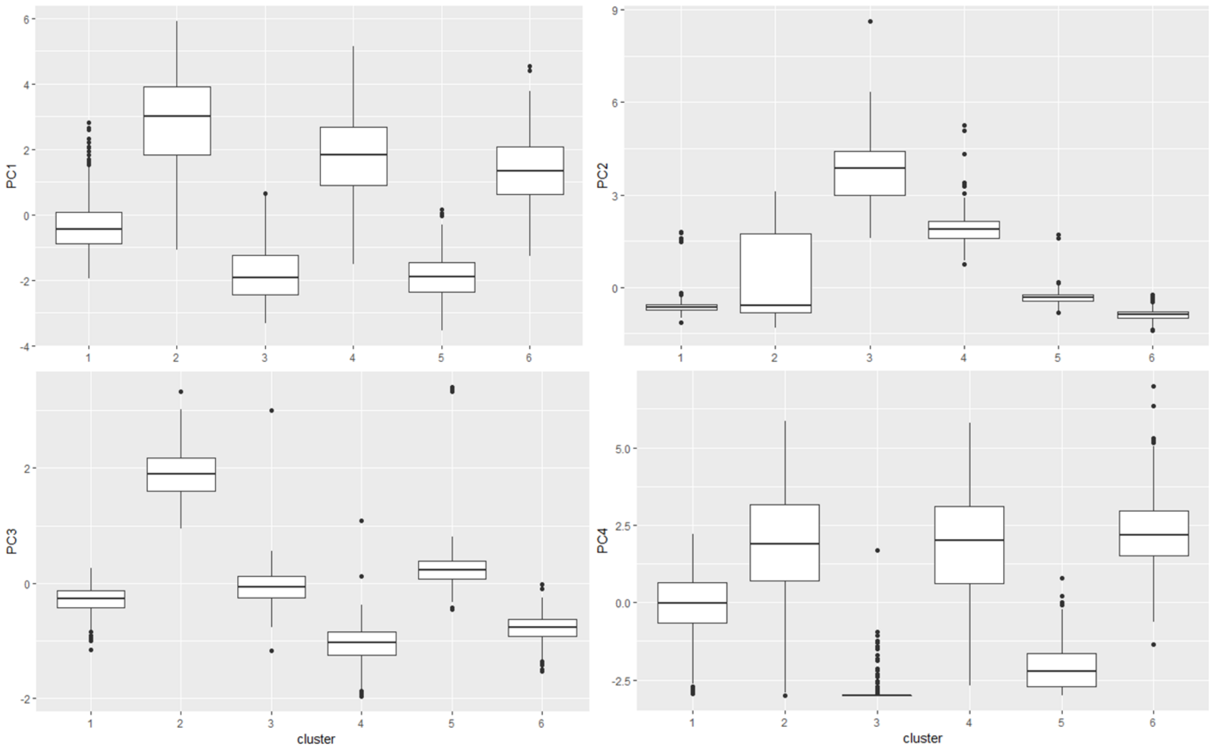

Figure 6. The box-plot of PC1, 2, 3 and 4 in the clusters 1~6 determined by K-means elbow

Six clusters were selected using K-means elbow method in figure 5. The sample size of each cluster was listed in table 4. The salesl distribution plot (figure 5C) indicated that 1, 2, 4 and 6 clusters were our potential target groups. But cluster 1 was removed from the candidate list considering its PC4 distribution in Figure 6. Low PC4 indicated low equipment demand because PC4 described the average related operations in the hospital. Cluster 2 was characterized by high operations (PC4), large size (PC1) and trauma (PC3) function. Cluster 4 was composed by high operation, large size and rehab hospitals. Cluster 6 was dominated by high operation and large size general hospitals.

Table 4. Sample numbers in selected clusters by K-means elbow method

cluster	1	**2**	3	**4**	5	**6**
N	709	306	157	202	690	436

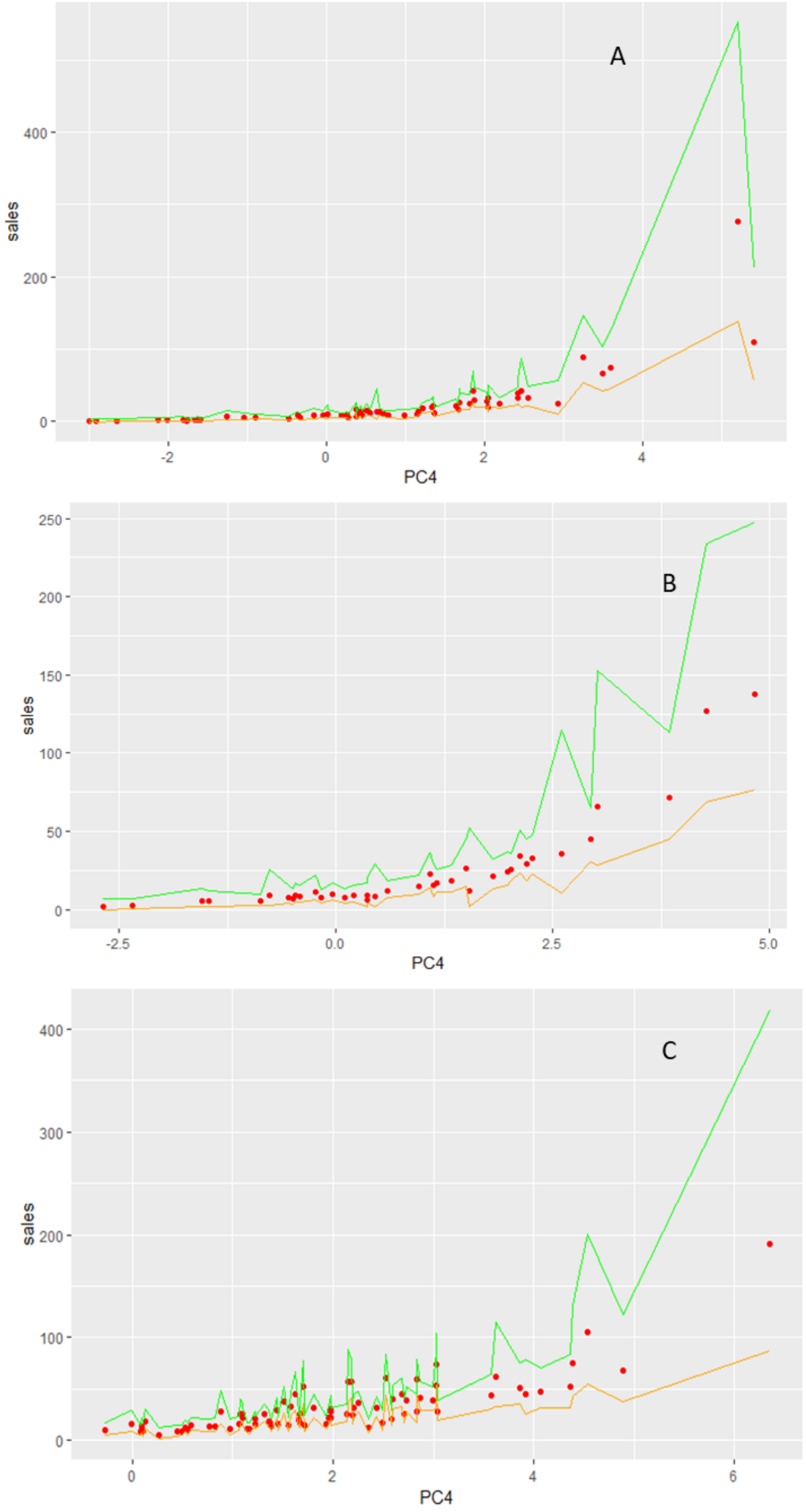

Figure 7. The predicted sales values and the confidence interval by regression models.

(A for cluster 2, B for cluster 4 and C for cluster 6)

Multivariate regression models were developed using PC1~4 as predictors and log transformed SALES as response for cluster 2, 4 and 6. PC4 was consistently listed as the significant predictors in all three models in table 5. Using these statistically significant models, the predicted sale values were shown in table 6 and figure 7 for hospitals with sales = 0.

Table 5. Multivariate regression results for optimal clusters determined by K-means elbow method

	cluster	coefficients of PC4	p-value for PC4 coefficient	R-square	ad-R-square	F	p
model1	2	0.536	1.52e-10	0.2025	0.1919	19.1(4, 301)	5.15E-14
model2	4	0.516	1.68e-07	0.1667	0.1498	9.855(4, 197)	2.74E-07
model3	6	0.225	0.0121	0.08648	0.078	10.2(4, 431)	6.73E-08

Table 6. The predicted revenue in US (1000 dollars/year)

cluster	N	sum	min	1st Q	median	mean	3stQ	max	Total
2	68	1468.33	0.3091	5.8913	10.9092	21.5931	24.6647	277.236	
4	38	921.849	1.744	8.06	12.127	24.259	25.935	137.991	
6	78	2462.62	4.875	15.264	25.239	31.572	40.577	190.898	4852

2.3.1.2 Silhouette method

Five clusters were identified using K-means silhouette method in figure 8. Based on sales distribution and sample size in table 7, cluster 4 and 5 were selected as potential candidates for further analysis. The same conclusion was obtained based on PC 1 and 4 distributions. Although cluster 1 was characterized by large general hospitals, its sample size, 784, was too big as a reasonable cluster compared with 2500 full sample size. So, it was not identified as a good candidate for further analysis. Large rehab hospitals were grouped in cluster 4 and large trauma hospitals in cluster 5. Multivariate regression models were developed for cluster 4 and 5 with log transformed sales as dependent variables and PC1, 2, 3 and 4 as predictors. The results were listed in table 8. PC4 was the only significant predictors in these two models indicating the importance of operation size as potential demand for the medical equipment. The predicted sales values were listed in table 9. The confidence interval and distribution of sales across PC4 values was described in figure 10. 109 hospitals were identified as potential clients for the medical equipment company using K-means silhouette. They are mainly large rehab or trauma hospitals with high average operations. The expected total revenue was 2392.78 thousand dollars/year.

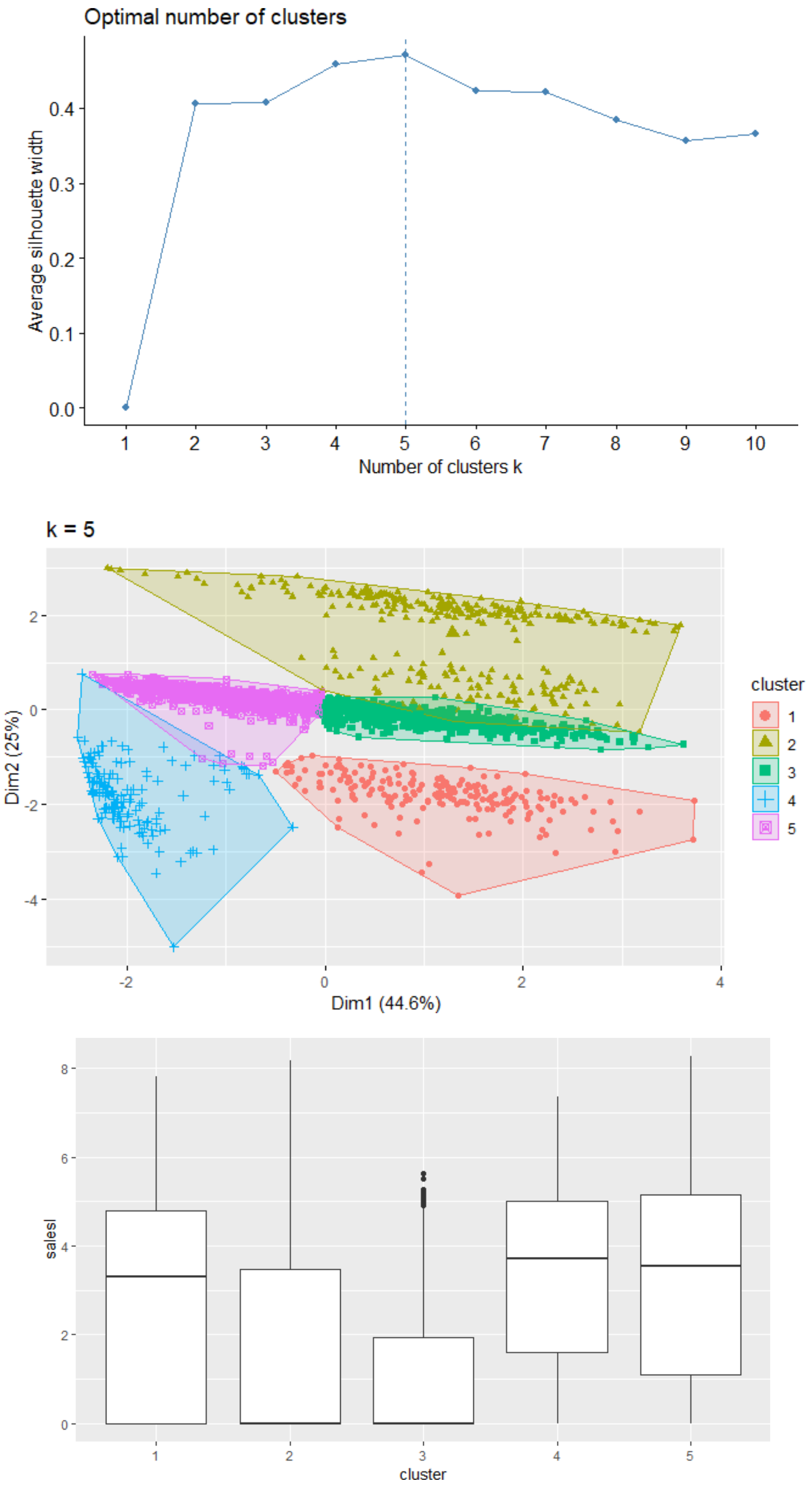

Figure 8. The optimal clusters determined by K-means silhouette

(A for elbow plot, b for cluster demonstration and c for distribution of sales in these clusters)

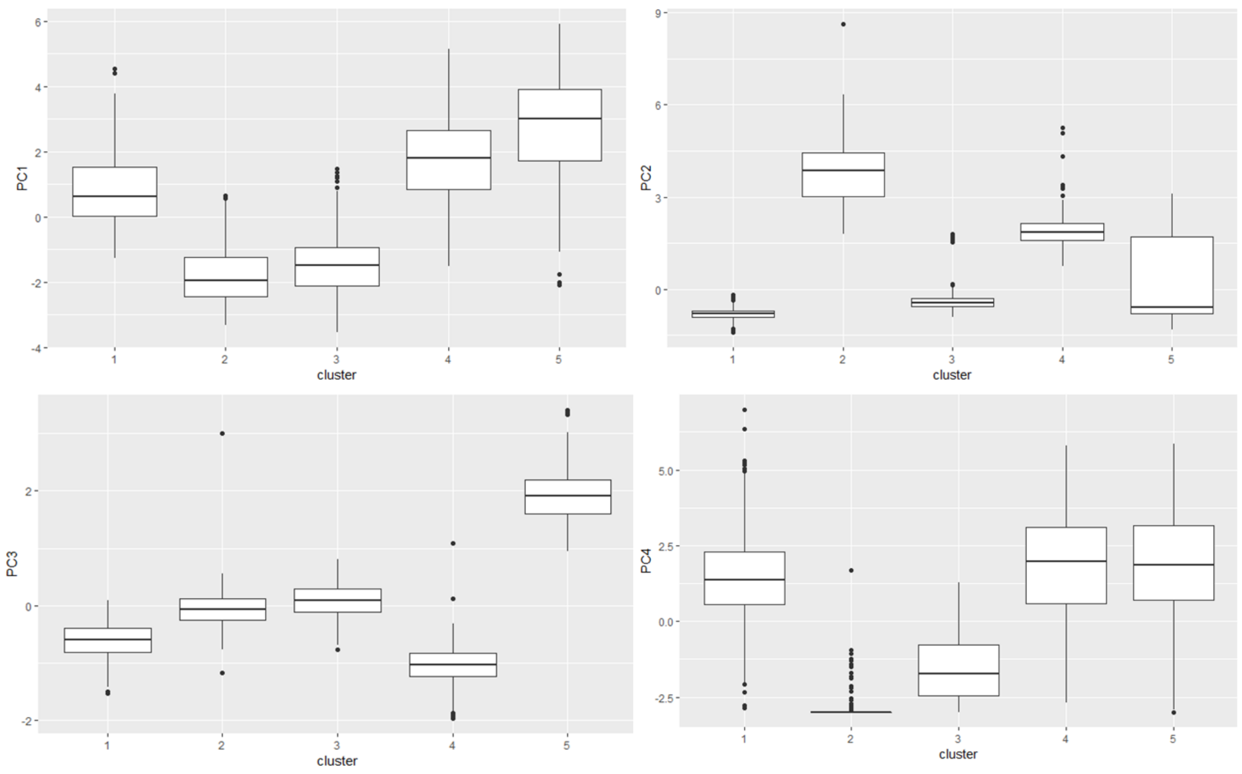

Figure 9. The box-plot of PC1, 2, 3 and 4 in clusters 1~5 determined by K-means silhouette

Table 7. Sample numbers in selected clusters determined by K-means silhouette method

Clusters by silhouette	1	2	3	**4**	**5**
N	784	155	1047	205	309

Table 8. Multivariate regression results for optimal clusters determined by K-means silhouette

	cluster	coefficients of PC4	p-value for PC4 coefficient	Df	R-square	ad-R-square	F	p
model4	4	0.5183	1.22E-07	200	0.171	0.1545	10.32 (4, 200)	1.29E-07
model5	5	0.54	9.18E-11	304	0.1667	0.1498	21.15 (4, 304)	2.11E-15

Table 9. The predicted revenue for potential target hospitals in US (1000 dollars/year)

cluster	N	sum	min	1st Q	median	mean	3stQ	max	total
4	38	919.92	1.732	8.046	12.105	24.209	25.595	139.638	
5	71	1472.86	0.2479	4.8102	10.054	20.745	23.32	287.707	2392.78

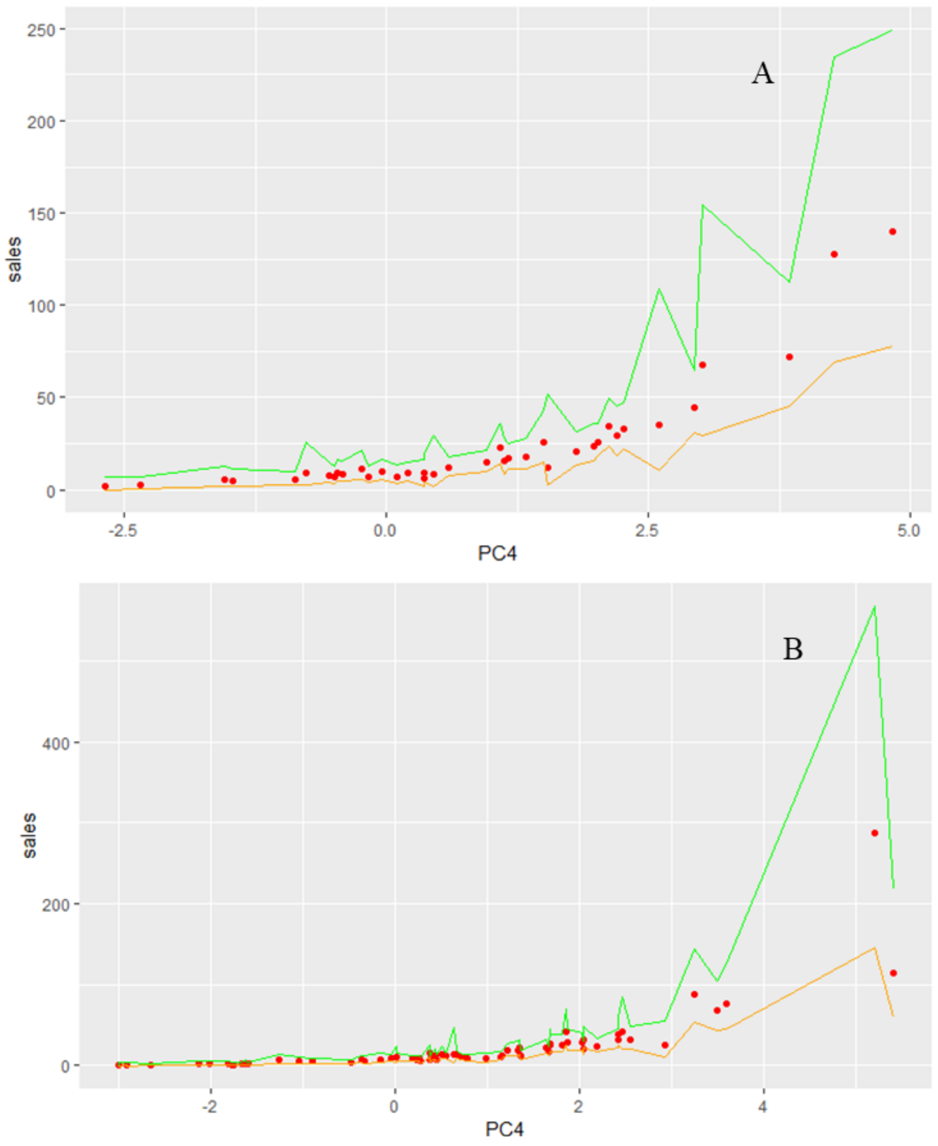

Figure 10 The predicted sales values and the confidence interval by regression models.

(A for cluster 4, B for cluster 5)

2.3.1.3 Gap statistics

7 clusters were determined by K-means gap. Based on sales distribution in figure 11, cluster1, 3, 4, 5, 6 were identified as potential candidates. Because cluster 3 had low PC1 and PC4 values in figure 12, it was not selected. Cluster 1, 4, 5 and 6 were the choices for further analysis, given their good PC1, PC4 and sales values. Cluster 1 is characterized by large general hospitals with high operations, cluster 4 by large rehab and trauma hospital with high operations, cluster 5 by large trauma hospitals with high operations and cluster 6 by large rehab hospitals with high operations.

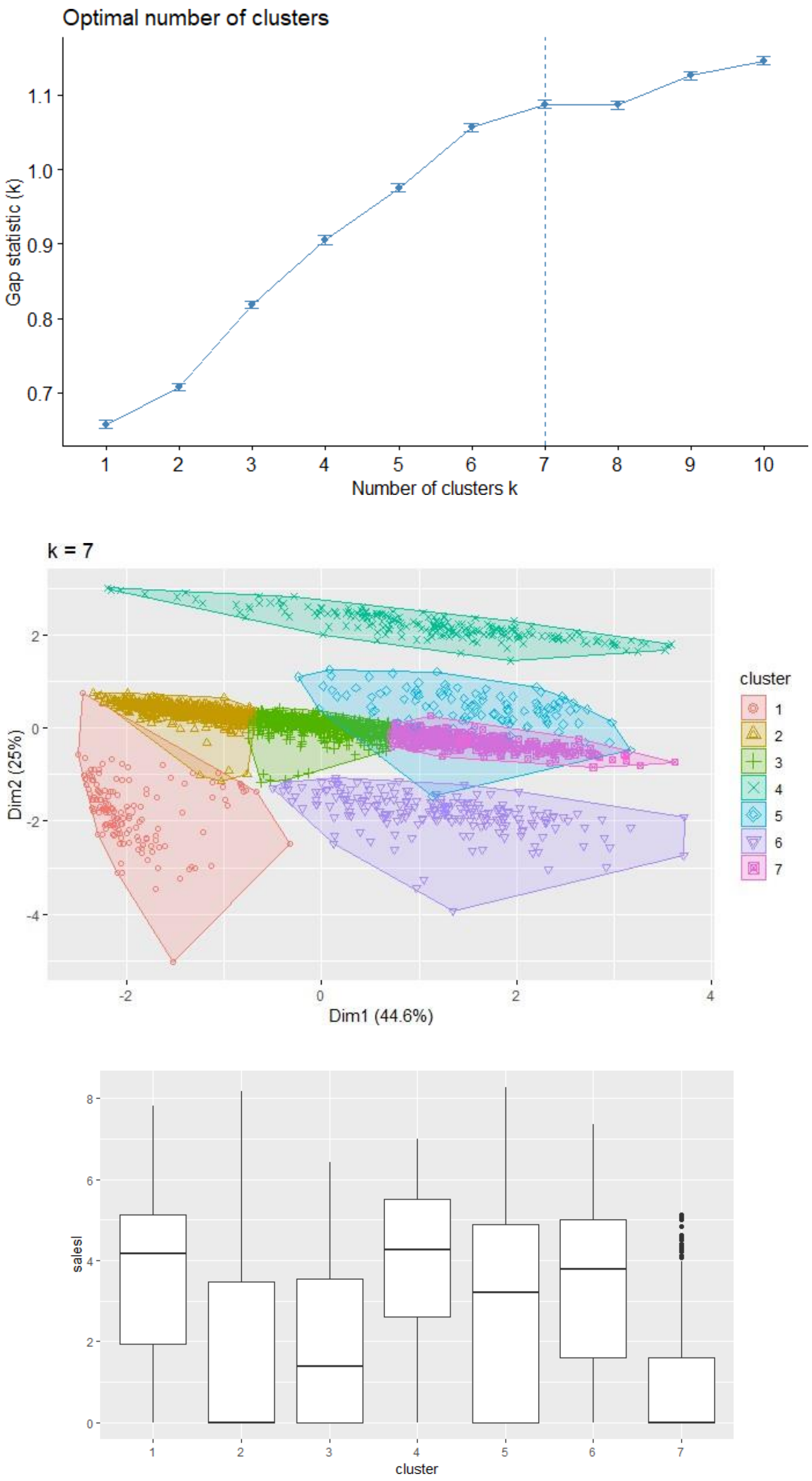

Figure 11. The optimal clusters determined by K-means gap method

(the top for elbow plot, middle for cluster demonstration and bottom for distribution of sales in these clusters)

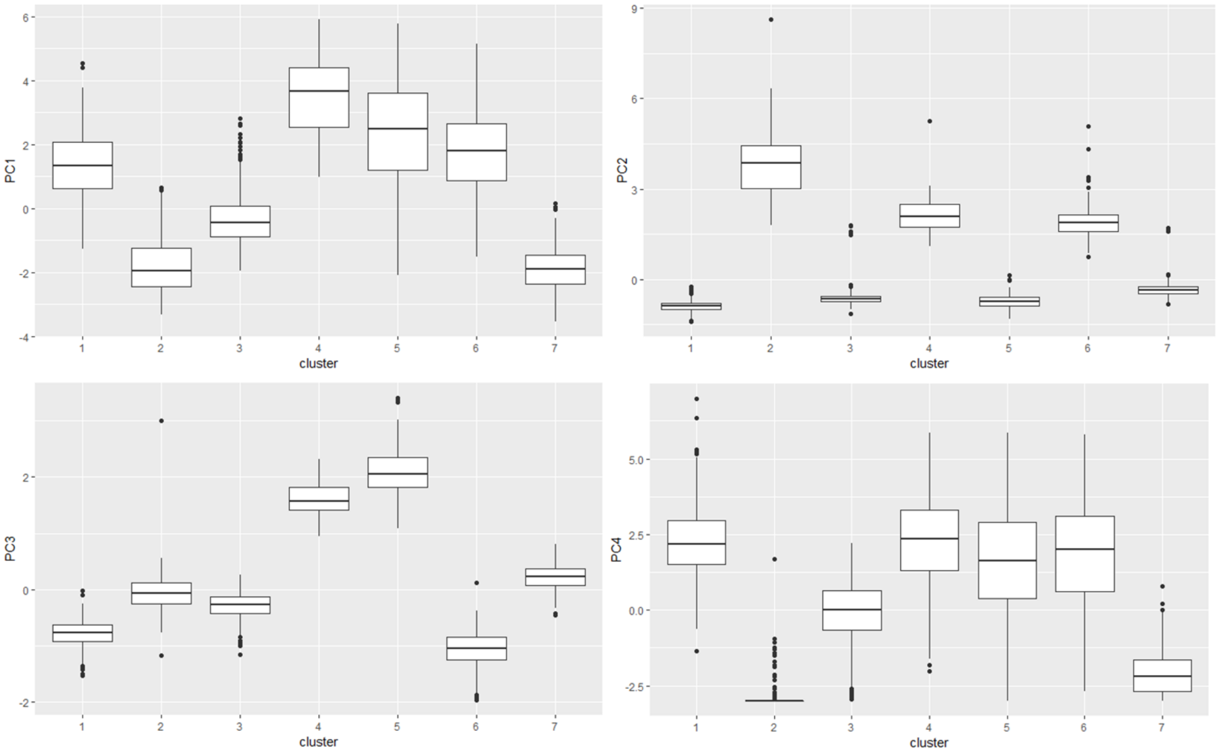

Figure 12. The box-plot of PC1, 2, 3 and 4 in clusters 1~7 determined by K-means gap

Table 10. Sample numbers in selected clusters determined by K-means gap method

Gap cluster	**1**	2	3	**4**	**5**	**6**	7
N	436	155	707	104	206	201	691

Multivariate regression analysis was performed for in cluster 1, 4, 5 and 6. The results were shown in table 11. PC4 was the only significant predictors in all four models indicating the importance of operation size as potential demand for the medical equipment.

Table 11. Multivariate regression results for selected optimal clusters determined by K-means gap

model	cluster	coefficients of PC4	p-value for PC4 coefficient	Df	R-square	ad-R-square	F	p
6	5	0.59153	2.33E-09	201	0.2439	0.2288	16.21 (4, 201)	1.60E-11
7	6	0.4529	6.35E-06	196	0.1793	0.1625	10.7 (4, 196)	7.27E-08
8	1	0.2252	0.0121	431	0.08648	0.078	10.2 (4, 431)	6.73E-08
9	4	0.3622	0.0212	99	0.09968	0.0633	2.74 (4, 99)	0.03282

Using these models, the predicted sales values for hospitals with sales = 0 were listed in table 12. The confidence interval and distribution of sales values was described in figure 13. Totally 187 hospitals were identified as potential clients for the medical equipment company. They are mainly large general, rehab or trauma hospitals with large operations. The expected total revenue was 4750.34 thousand dollars/year.

Table 12. The predicted revenue for potential target hospitals in US (1000 dollars/year)

cluster	N	sum	min	1st Q	median	mean	3stQ	max	total
1	78	2462.62	4.875	15.264	25.239	31.572	40.577	190.898	
4	14	602.492	5.893	11.464	25.71	43.035	62.542	203.529	
5	57	797.304	-0.038	3.4536	8.3063	13.9878	18.4416	124.058	
6	38	887.923	1.741	6.888	11.674	23.366	29.27	117.339	4750.34

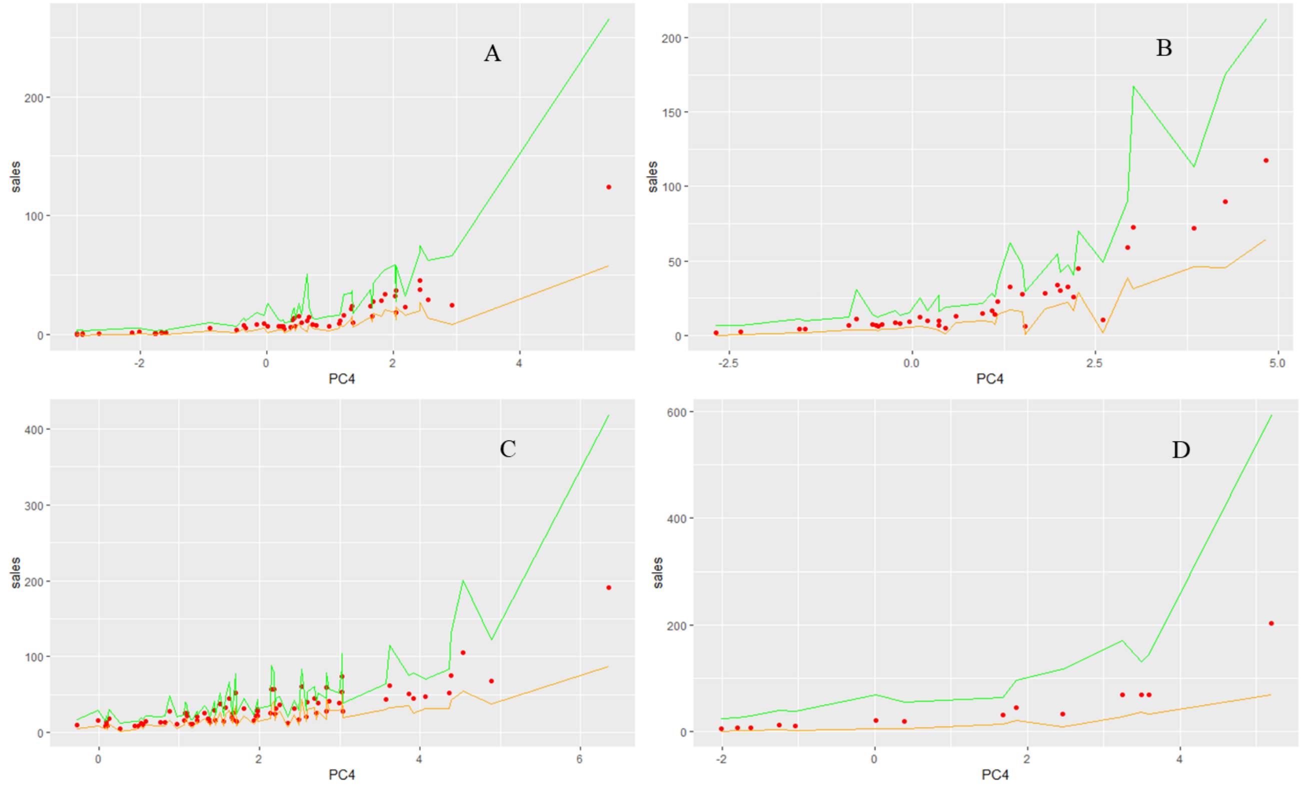

Figure 13 The predicted sales values and the confidence interval by regression models.

(A for cluster 5, B for cluster 6, C for cluster 1 and D for cluster 4)

2.3.2 Hierarchical clustering

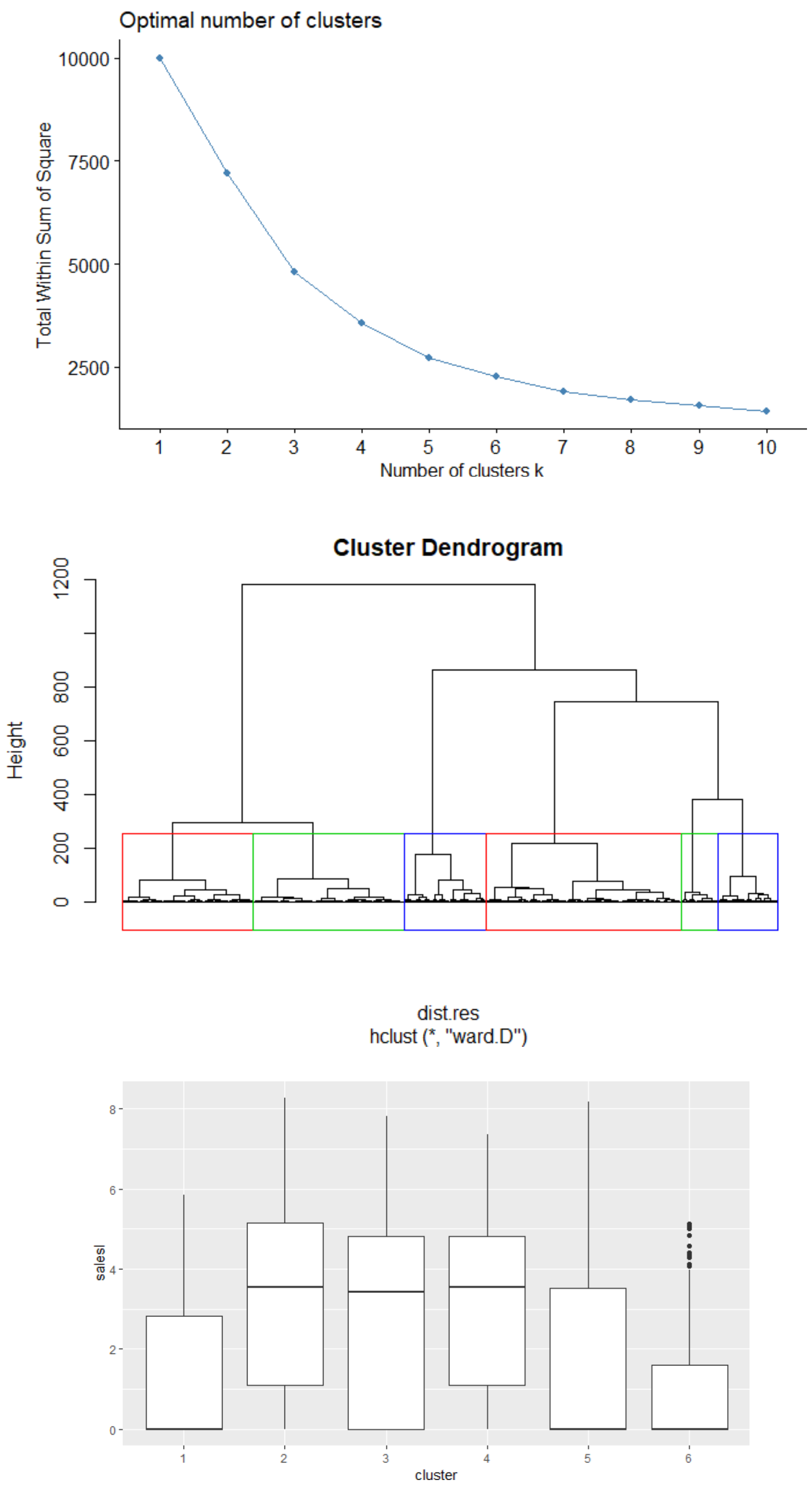

Figure 14. The optimal clusters determined by hierarchical elbow method

(the top for elbow plot, middle for cluster demonstration and bottom for distribution of sales in these clusters)

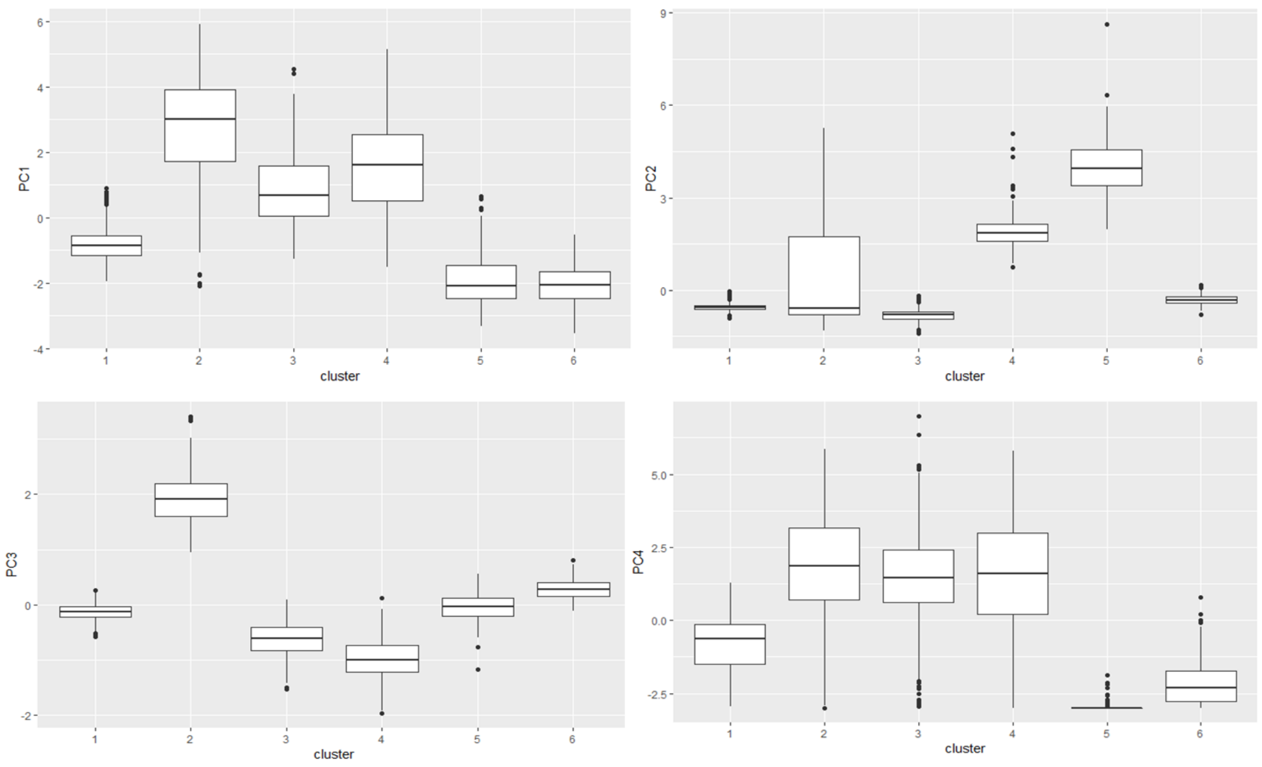

Figure 15. The box-plot of PC1, 2, 3 and 4 in clusters 1~6 determined by hierarchical elbow

Elbow, Silhouette and gap statistic were all applied to determine the optimal cluster number in hierarchical clustering in this study

2.3.2.1 Elbow method

6 clusters were determined by hierarchical elbow method. The results were listed in figure 14. Three clusters, 2, 3, and 4, were selected as good potential candidates based on sales, PC1 and PC4 values. However, there were more than 700 samples in cluster 3 in table 14. Thus, cluster 2 and 4 were selected for further analysis. Cluster 2 has high PC1, 3 and 4 values, indicating large trauma hospitals with high operations; cluster 4 has high PC1, PC4 and median high PC 2 values, indicating large rehab hospitals with high operations.

Table 14. Sample numbers in selected clusters determined by hierarchical elbow method

cluster	1	2	3	4	5	6
n	501	**311**	**748**	**225**	140	575

Multivariate regression models were developed to predict the revenues of hospitals with sales = 0 in cluster 2 and 4 with log transformed sales as response and PC 1~4 as predictors. The results were shown in table 15 and 16. PC4 was still the only statistically significant predictor in these models. The predicted revenue distribution was listed in table 16. The sum of the total revenue for all these hospitals was 2410.35 thousand dollars/year. The plot of predicted revenue and PC4 was shown in figure 16.

Table 15. Multivariate regression models for selected optimal clusters determined by hc elbow

	cluster	coefficients of PC4	p-value for PC4 coefficient	R-square	ad-R-squre	F	p
model10	2	0.5314	1.38E-10	0.2231	0.2129	21.96 (4, 306)	5.95E-16
model11	4	0.4559	2.98E-07	0.2077	0.1933	14.42 (4, 220)	1.79E-10

Table 16. The predicted revenue for potential target hospitals in US (1000 dollars/year)

cluster	N	sum	min	1st Q	median	mean	3stQ	max	total
2	72	1501.44	0.2139	4.2192	10.0619	20.8533	22.514	311.36	
4	44	908.914	1.458	6.954	10.576	20.657	28.437	112.956	2410.35

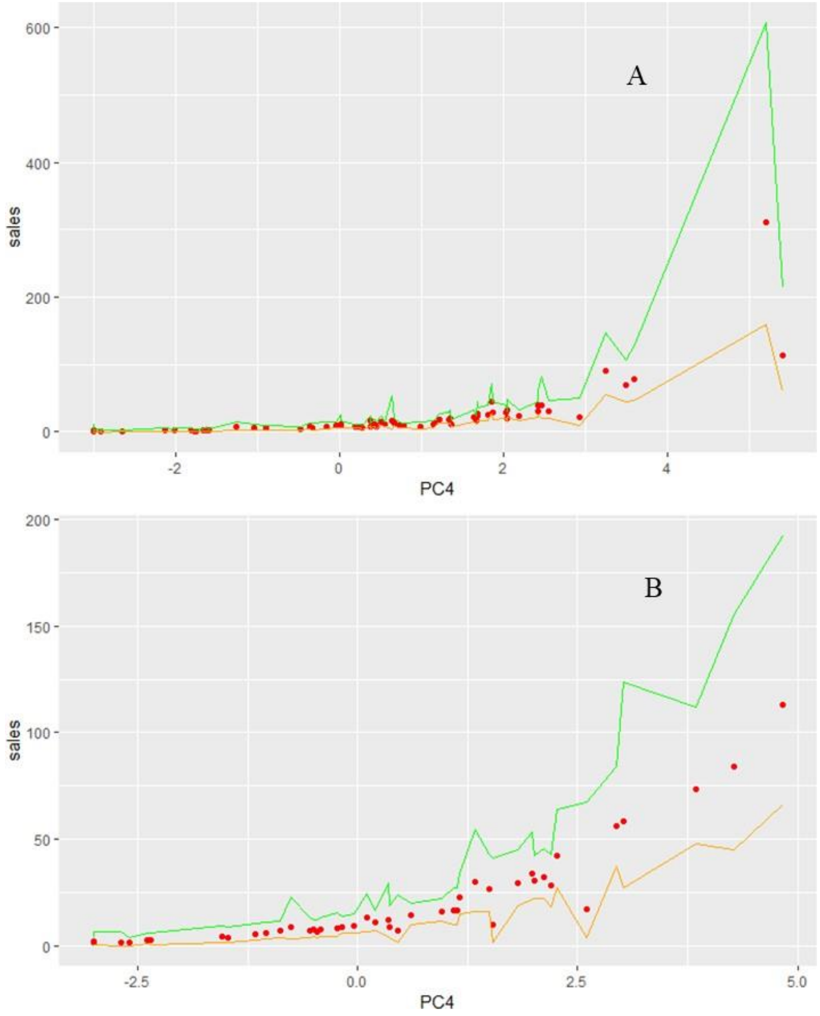

Figure 16 The predicted sales values and the confidence interval by regression models.

(A for cluster 2, B for cluster 4)

2.3.2.2 Silhouette method

Using hierarchical silhouette, five clusters were selected in figure 17. Based on sales, PC1 and PC4 values, clusters 2, 3 and 4 were determined as potential candidates. But there were more than 700 samples in cluster 3 indicating that it was not a good choice for further analysis. Thus, only cluster 2 and 4 were chosen. They were the same clusters we obtained by hierarchical elbow method. Therefore, we can just refer to all related results in 2.3.2.1, table 14-16 and figure 16.

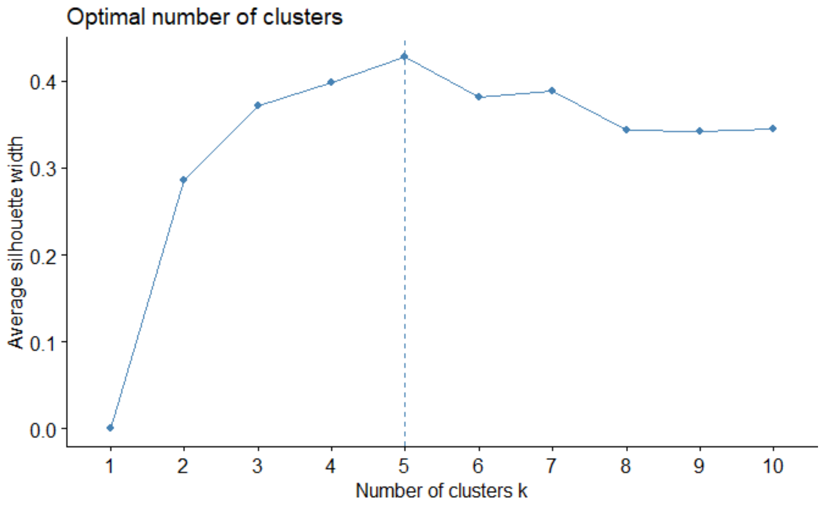

Figure 17. The optimal cluster number determined by hierarchical silhouette method

2.3.2.3 Gap statistic

24 clusters were selected using hierarchical gap in figure 18. Based on sales value, PC1, PC4 and sample size in each cluster in table 17, four clusters were selected as good candidates, 4, 5, 8 and 22. The clusters with sample size less than 90 were not considered for multivariate analysis, because small sample size is not for model development. Cluster 8 and 22 did not give good results with regression analysis at α = 0.05. Thus, only cluster 4 and 5 were used for prediction. Observing the PC and sales values, it was found that cluster 4 was the group with high sales, high operations and large size. Cluster 5 was for large rehab hospitals with high operations. The model fitting results were shown in table 18. The predicted revenue for potential clients were described in table 19 and figure 20. Totally 43 hospitals were selected with the sum of revenue 1204.16 thousand dollars/year.

Table 17. Sample size in selected clusters determined by hierarchical gap

Cluster	1	2	3	**4**	**5**	6	7	**8**	9	10	11	12
N	193	156	41	**147**	**104**	95	126	**106**	152	142	114	106
Cluster	13	14	15	16	17	18	19	20	21	**22**	23	24
N	234	225	68	48	29	116	33	64	64	**90**	26	21

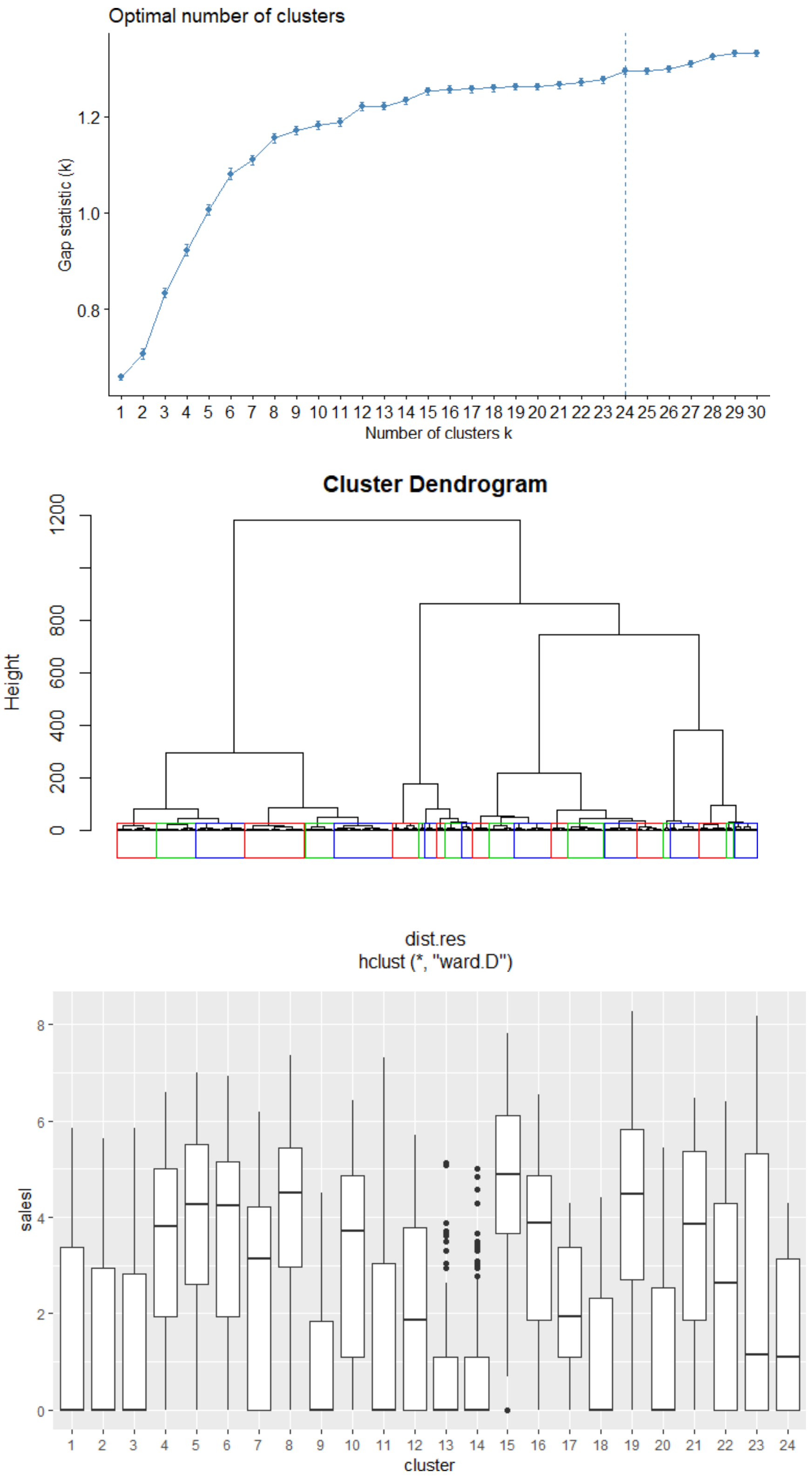

Figure 18. The optimal clusters determined by hierarchical gap

(the top for gap plot, middle for cluster demonstration and bottom for distribution of sales in these clusters)

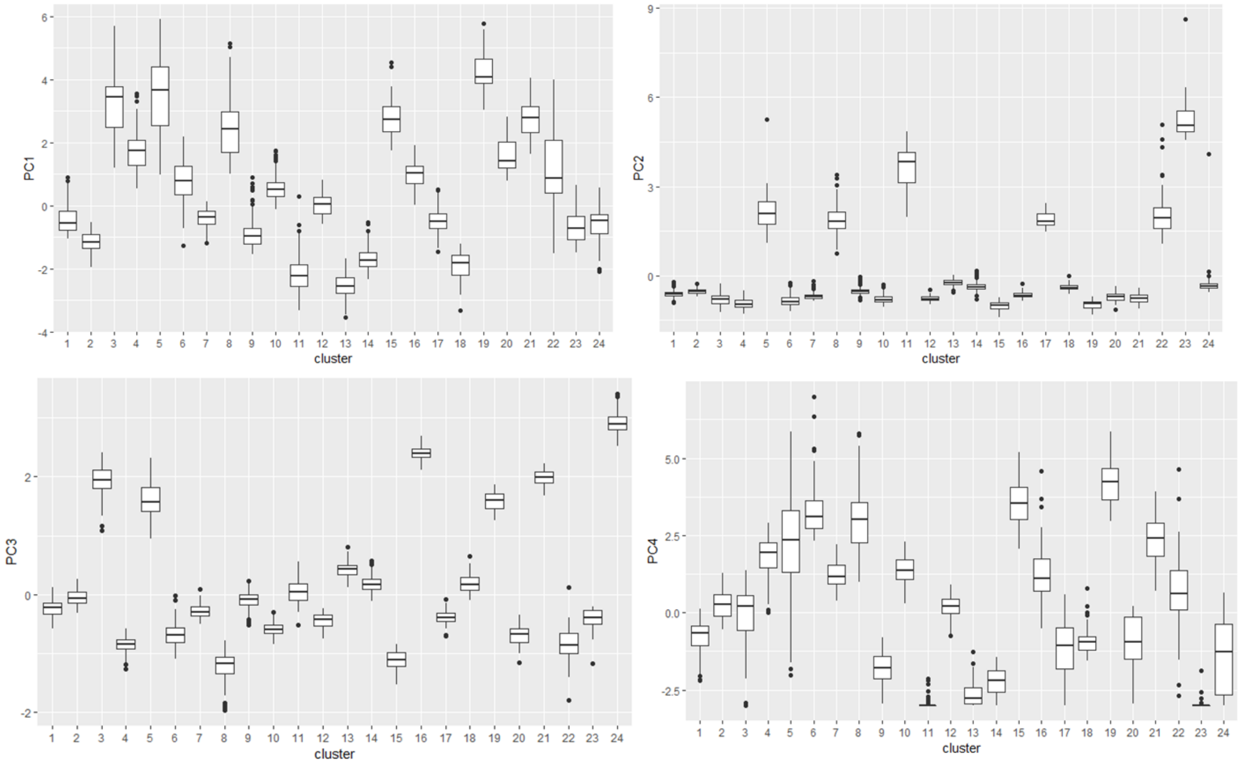

Figure 19. The box-plot of PC1, 2, 3 and 4 in clusters 1~24 determined by hierarchical gap

Table 18. Multivariate regression models for selected optimal clusters determined by hc gap

	cluster	coefficients of PC4	p-value for PC4 coefficient	Df	R-square	ad-R-square	F	p
model12	4	0.4267	1.46E-01	142	0.173	0.1497	7.428 (4, 142)	1.84E-05
model13	5	0.3622	2.12E-02	99	0.0997	0.0633	2.74 (4, 99)	3.28E-02

Table 19. The predicted revenue for potential target hospitals in US (1000 dollars/year)

cluster	N	sum	min	1st Q	median	mean	3stQ	max	total
4	29	601.57	1.635	6.804	12.138	20.744	27.447	83.74	
5	14	602.49	5.89	11.464	25.71	43.035	62.542	203.529	1204.06

2.3.3 PAM clustering

Elbow, Silhouette and gap statistic were also applied to determine the optimal cluster number in pam clustering in this study.

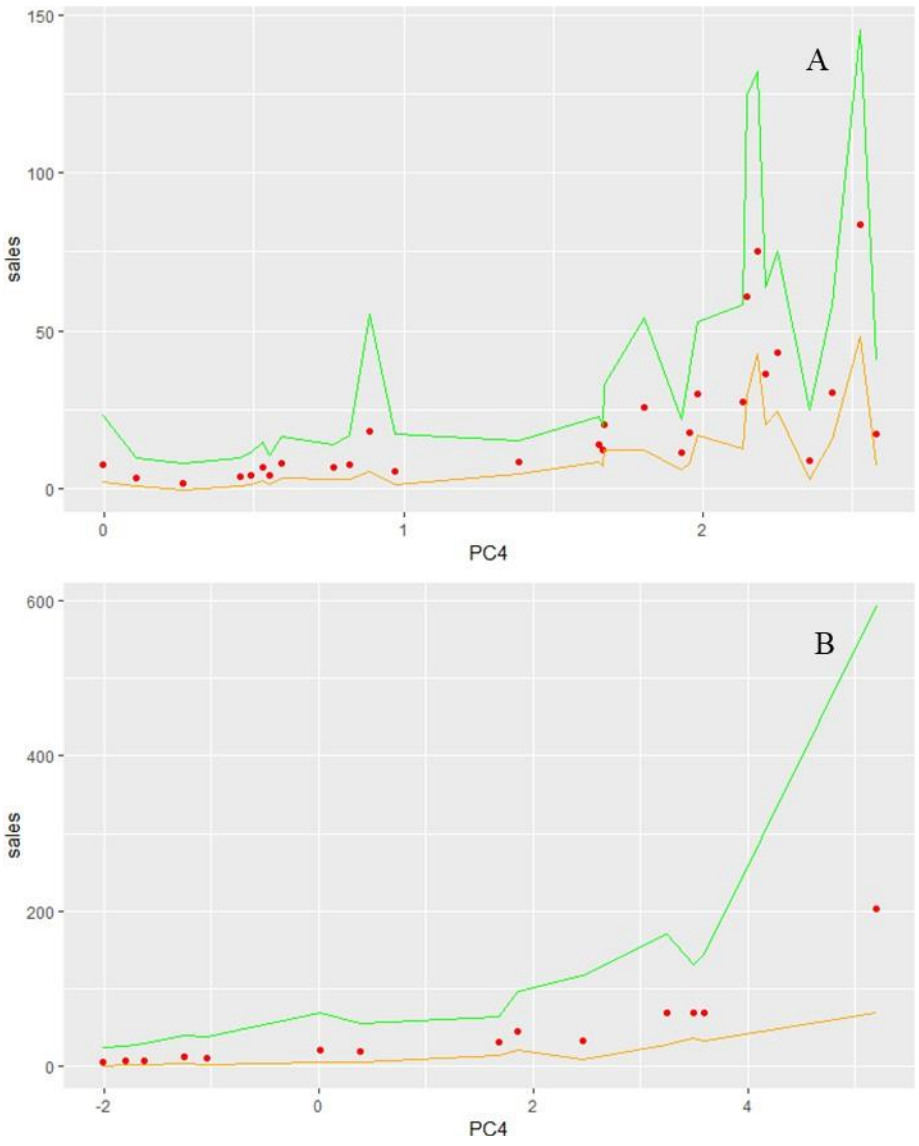

Figure 20 The predicted sales values and the confidence interval by regression models.

(A for cluster 4, B for cluster 5)

2.3.3.1 Elbow and Silhouette method

The same optimal cluster number was obtained using both elbow and silhouette method in pam clustering. The results were shown in Figure 21. The sample size in each cluster was shown in table 20.

Table 20. Sample size in the clusters determined by pam elbow and silhouette

cluster	1	**2**	**3**	**4**	**5**	6	7
n	687	**208**	**510**	**102**	**199**	639	155

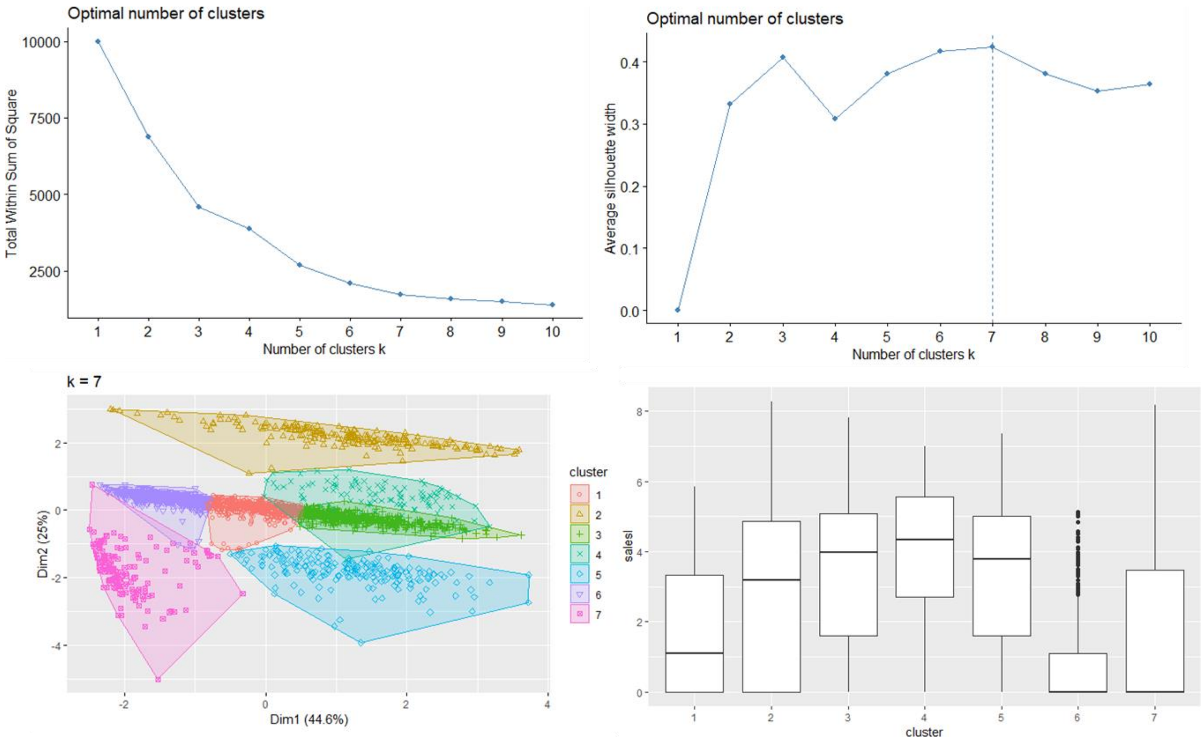

Figure 21. The optimal clusters determined by pam elbow and silhouette

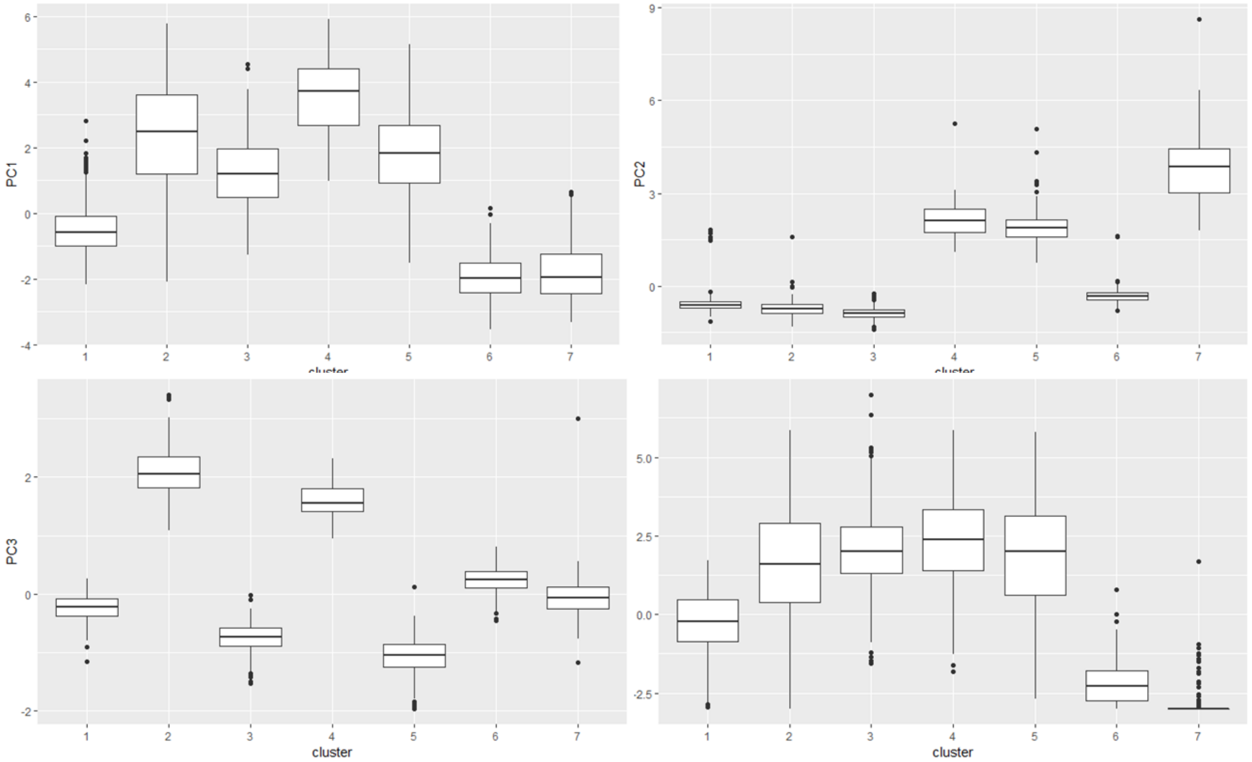

Figure 22. The box-plot of PC1, 2, 3 and 4 in clusters 1~7 determined by pam elbow and silhouette

Based on the boxplot of sales, PC1 and PC4, four clusters were selected as the good candidates for multivariate regression. They were cluster 2, 3, 4 and 5. Cluster 2 was for hospitals with high sales, PC1 and PC4 values. It indicated that they were large trauma hospitals with high operations. Cluster 3 was for large hospitals with high operations and high sales. Large rehab and trauma hospitals with high operations were grouped into cluster 4. Large rehab hospitals with high operations were grouped into cluster 5. After multivariate regression, it was found that the model for cluster 4 was not statistically significant. Thus, only cluster 2, 3 and 5 were applied for revenue estimation. The model fitting results were listed in table 21. The predicted revenue for potential clients were listed in table 22 and figure 23. Totally 200 hospitals were selected with the sum of revenue 4477.14 thousand dollars/year.

Table 21. Multivariate regression models for selected optimal clusters determined by pam elbow and silhouette method

	cluster	coefficients of PC4	p-value for PC4 coefficient	R-square	ad-R-squre	F	p
model 14	2	0.5864	2.49E-09	0.2543	0.2396	17.31 (4, 203)	3.11E-12
model 15	5	0.4592	4.98E-06	0.1864	0.1697	11.11 (4, 194)	3.88E-08
model 16	3	0.34394	1.50E-05	0.1074	0.1004	15.2 (4, 505)	9.66E-12

Table 22. The predicted revenue for potential target hospitals in US (1000 dollars/year)

cluster	N	sum	min	1st Q	median	mean	3stQ	max	total
3	103	2787.785	2.777	11.165	18.866	27.066	31.823	239.82	
2	59	807.16	0.5025	2.1917	7.9426	13.6807	17.0208	129.8195	
5	38	882.1957	1.69	6.778	11.414	23.216	28.814	121.513	4477.141

2.3.3.2 Gap statistic

Using gap statistic, the optimal cluster number was determined to be 13. Given k = 13, pam clustering produced 13 optimized clusters in figure 24. The sample size in each cluster was listed in table 23. Based on the value and distributions of sales, PC 1 and PC4 in figure 24~25, 5 clusters were selected for multivariate regression. They were cluster 3, 4, 5, 7 and 12. Cluster 3 has high sales, PC1 and PC4 values, which indicated large general hospitals with high operations. Cluster 4 has high sales, PC1, 2, 3 and 4, which grouped large hospitals with rehab, trauma function and high operations. Cluster 5 is characterized by high sales, median PC1 and high PC4 values. The median sized hospitals with high sales and operations contributed this cluster. Cluster 7 was noticed by high sales, PC1, PC4 with median high PC2 values. It indicated that large hospitals with high operations, high sales and rehab function were the dominant members inside. Cluster 12 was mainly composed by large trauma hospitals with high sales and operations.

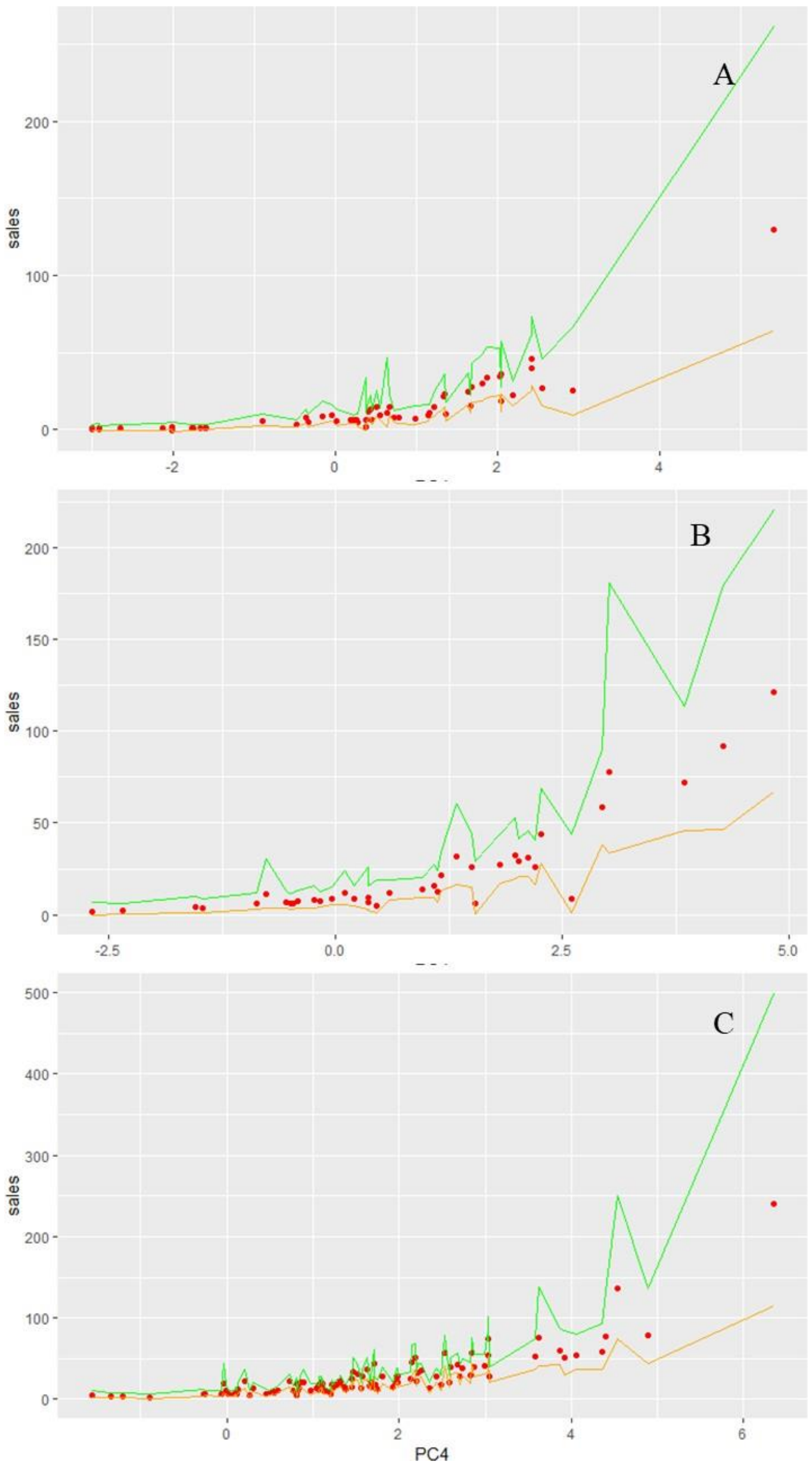

Figure 23 The predicted sales values and the confidence interval by regression models.

(A for cluster 2, B for cluster 5, C for cluster 3)

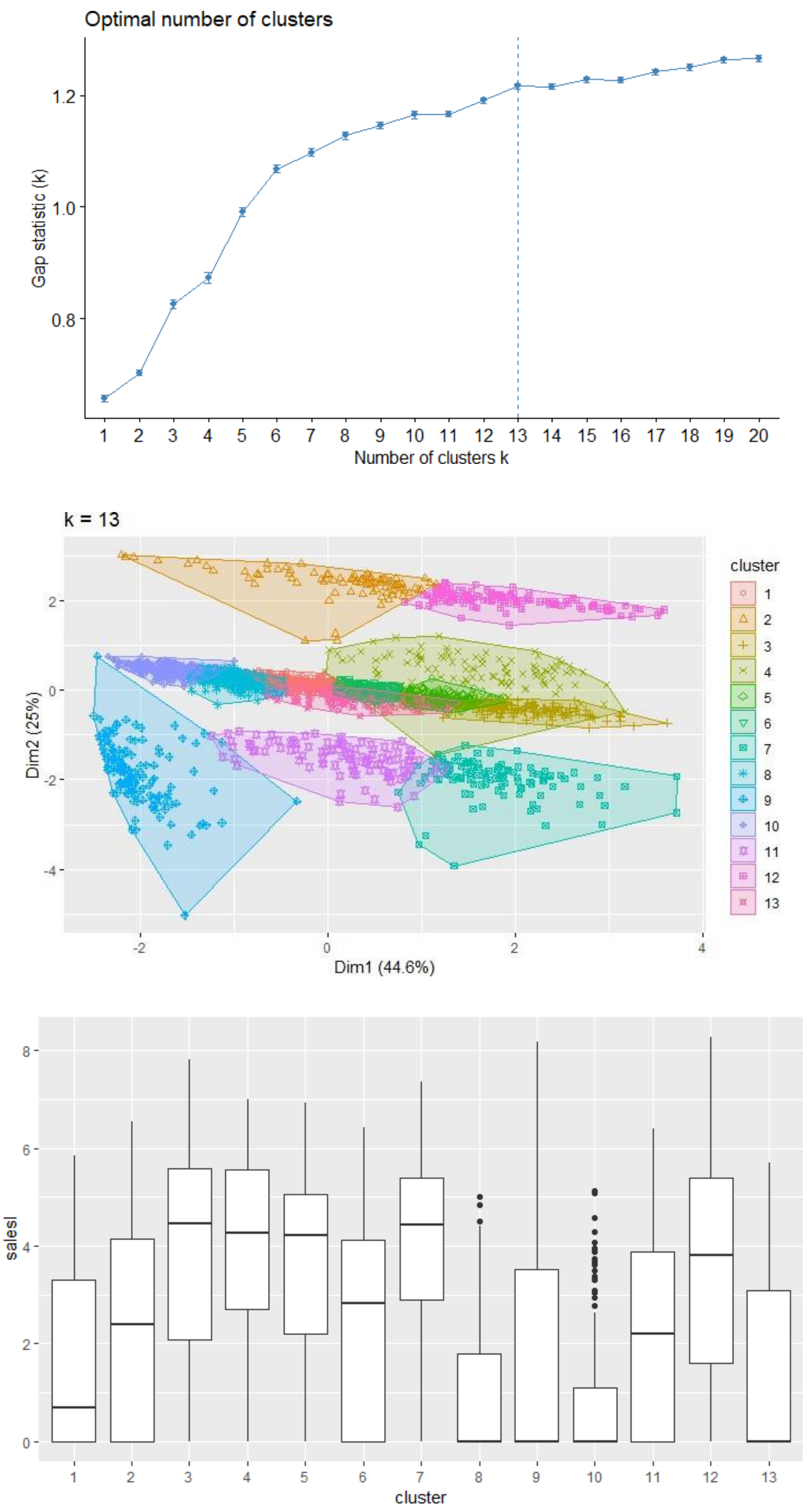

Figure 24. The optimal clusters determined by pam gap statistic

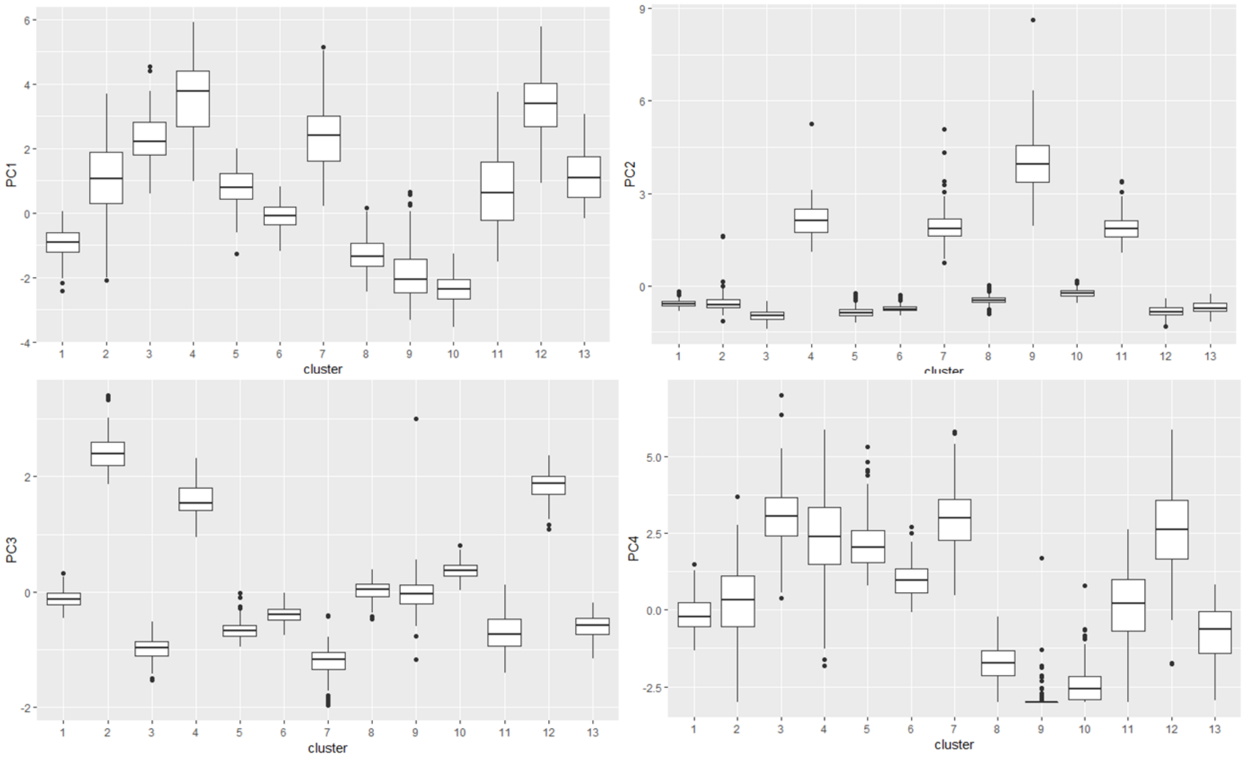

Figure 25. The box-plot of PC1, 2, 3 and 4 in clusters 1~13 determined by pam gap statistic

Table 23. Sample size in the clusters determined by pam gap statistic

cluster	1	2	**3**	**4**	**5**	6	**7**	8	9	10	11	**12**	13
n	338	91	**155**	**101**	**232**	236	**113**	362	144	358	109	**118**	143

Table 24. Multivariate regression models for selected clusters determined by pam gap statistic

	cluster	coefficients of PC4	p-value for PC4 coefficient	R-square	ad-R-square	F	p
model 17	3	0.04343	8.01E-01	0.1087	0.08496	4.575 (4, 150)	1.63E-03
model 18	4	0.3306	4.17E-02	0.0778	0.0393	2.024 (4, 96)	9.71E-02
model 19	7	0.3962	3.39E-02	0.1174	0.0847	3.591 (4, 108)	8.65E-03
model 20	12	0.5705	8.60E-05	0.1657	0.1361	5.61 (4, 113)	0.00037

Table 25. The predicted revenue for potential target hospitals in US (1000 dollars/year)

cluster	N	sum	min	1st Q	median	mean	3stQ	max	total
3	26	998.5037	7.442	15.651	27.267	38.404	43.946	125.465	
4	12	569.495	8.758	13.682	32.711	47.458	66.139	183.876	
7	12	570.5264	7.432	23.064	43.707	47.544	71.455	109.221	
12	25	462.3553	0.8017	7.6116	15.4149	18.4942	21.0218	98.0039	2600.88

Multivariate regression was performed for cluster 3, 4, 5, 7 and 12. The results were shown in table 24. All models were statistically significant. Model 17 is different from the others because the coefficient of PC1 and PC3, not PC4 were significant. Using these models, the predicted revenue was listed in table 25 for hospitals in these clusters with sales = 0. Totally 75 hospitals were selected as potential clients with the sum of revenue 2600.88 thousand dollars/year. The distribution of the predicted revenues was described in figure 26.

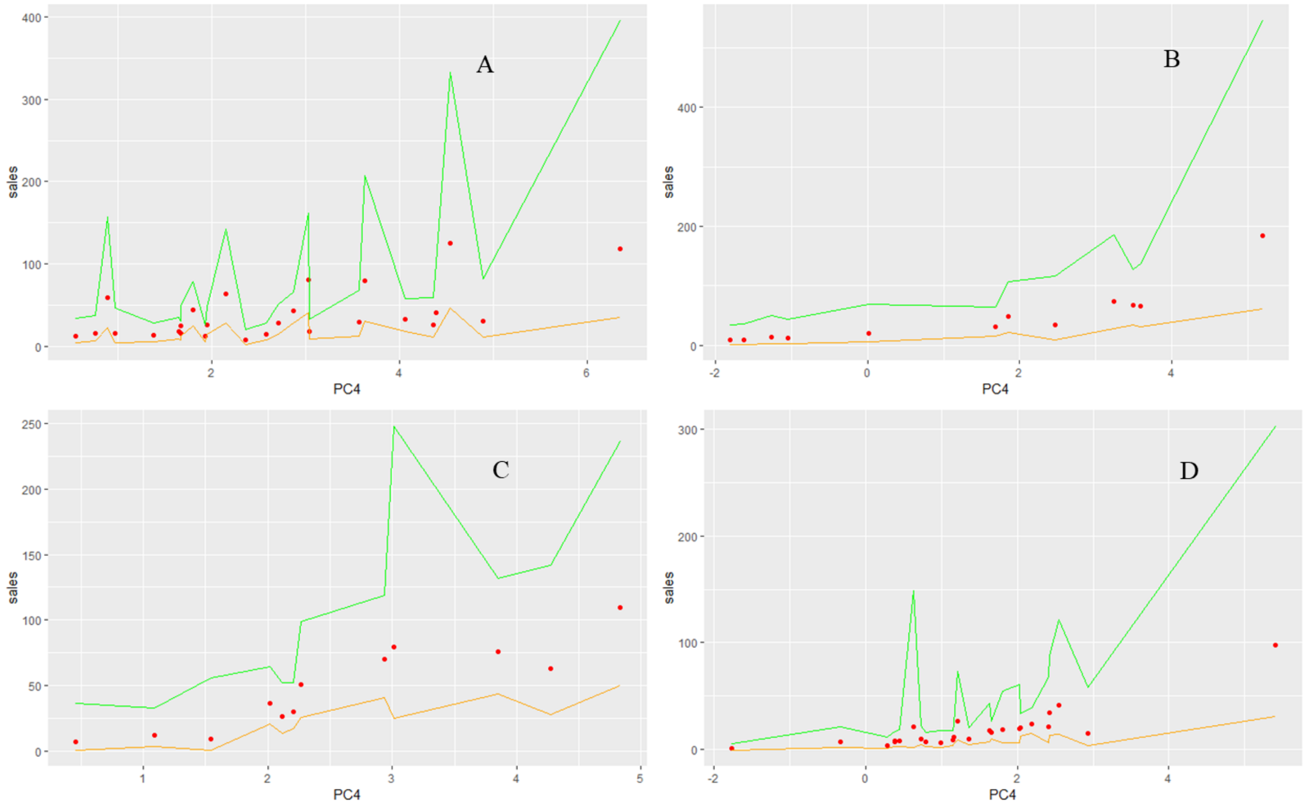

Figure 26 The predicted sales values and the confidence interval by regression models.

(A for cluster 3, B for cluster 4, C for cluster 7, D for cluster 12)

3. Conclusion

Three clustering methods, K-means, hierarchal and pam were applied to explore the potential clients in US for a medical equipment company. Because three different methods, elbow, silhouette and gap statistic, were used to determine the optimal cluster numbers with each clustering method, totally 9 methods were applied to detect the possible customers. Considering the different algorithms of each method, different results were obtained. Some produced high cluster numbers like pam gap, some with low cluster numbers like hierarchical silhouette. Some clusters were formed with good character, such as high sales and high PC4 values. But when multivariate regression was performed, no good model was developed. Thus, they must be removed from the list. Finally using the good models with good clusters, ~43 to 200 potential clients were selected with total revenues from 1204 to 4852 thousand dollars/year using different methods.