Analysis of Profitability Potential for Low Profit Hospitals

Indrajit Choudhury

12/19/16

Executive Summary:

We have determined that our client can increase sales in their upcoming fiscal quarter by focusing a sales campaign on the following 4 hospitals, where potential gain in earnings by selling to each hospital is listed:

1. Sioux Falls SD (HID 43065)- Potential Gain of $173,019
2. Fort Myers FL (HID 24039)- Potential Gain of $156,599
3. Buffalo NY (HID 107021)- Potential Gain of $154,486
4. Peoria IL (HID 257043)- Potential Gain of $133,103

Introduction:

We are given a dataset consisting of 4,703 observations. Each observation refers to a specific hospital in the US with a series of identifier variables (State, Zip Code, Hospital Identification Number, etc), demographic variables, operation variables and sales variables. The demographic variables refer to specific traits of a hospital (such as number of beds or an indicator variable that specifies whether the hospital has a trauma ward or not). The operation variables specify the number of specific operations done during a certain time period (such as number of Hip operations performed in 1996). The sales variables (salesy and sales12) refer to the number of sales of rehabilitation equipment during a certain time period (in 1000s of dollars). In the context of this project, our client is a company that sells orthopedic equipment to these types of hospitals. The goal of this project is find ways to increase the sale of orthopedic equipment to hospitals in the US.

We hope to help our client increase sales by finding an optimal subset of hospitals within these 4,703 in which to focus their sales campaign on. We wish to find this subset by using dimension reduction, cluster analysis and regression methods to find hospitals that have low or nonexistent sales despite sharing characteristics with hospitals that have high sales numbers. It is our belief that if the company were to focus their sales on these types of hospitals, they would have a high chance of increasing their sales numbers.

Definitions and Transformations:

Rehabilitation sales are defined by two variables: SALESY and SALES12, which refer to the sales of rehabilitation equipment for a given hospital since January 1st and sales from the last 12 months. For simplicity, from this point forward, the response variable will now be called SALES, which we define as SALES = SALESY + SALES12.

Our 8 demographics variables are BEDS (Number of Hospital Beds), RBEDS (Number of Rehab Beds), OUTV (Number of Outpatient Visits), ADM (Administrative Costs), SIR (Revenue from Inpatient Procedures), TH (Binary Variable indicating if hospital is a teaching hospital), TRAUMA (binary variable indicating presence of a trauma unit) and REHAB (binary variable indicating presence of a rehab unit).

Our 5 Operation variables are HIP95 (Number of Hip Operations in 1995), KNEE95 (Number of Knee Operations in 1995), HIP96 (Number of Hip Operations in 1996), KNEE96 (Number of Knee Operations in 1996) and FEMUR96 (Number of Femur Operations in 1996).

In the interest of variance stabilization, the following set of transformations was made on these variables (Note that for simplicity we kept the same name for our transformed variables):

$BEDS = \sqrt{BEDS}$, $RBEDS = \sqrt{RBEDS}$, $HIP95 = HIP95^{1/4}$, $KNEE95 = KNEE95^{1/4}$, $HIP96 = HIP96^{1/4}$, $KNEE96 = KNEE96^{1/4}$, $FEMUR96 = FEMUR96^{1/4}$, $OUTV = ln\left(1 + \frac{OUTV}{500}\right)$, $ADM = ln\left(1 + \frac{ADM}{500}\right)$, $SIR = ln\left(1 + \frac{SIR}{500}\right)$ and $SALES = \ln(1 + SALES)$.

For each of these variables, the initial data appears skewed in one way or another while after the transformation was applied, the data then appeared to look approximately normal. This can be exhibited in the histograms for the variable ADM before transformation (Figure 1) and after transformation (Figure 2). The same pattern is exhibited in all of the other variables that were transformed as well.

Dimension Reduction:

We have one response variable and 13 explanatory variables split among demographics and operation variables. We wish to reduce these 13 explanatory variables to a smaller number of factors. We will accomplish this with the use of Principal Component Analysis and Factor Analysis. The main idea behind both Principal Component Analysis and Factor Analysis is that a certain percentage of the variance of the data can be explained by a series of linear combinations of our explanatory variables. In the case of our data, 13 of these linear combinations would explain

100% of the variation in the data, but typically, most the variance can be explained by far fewer of these linear combinations, which we call factors.

First, we subdivide our explanatory variables into the two groups we have defined earlier (demographics and operation), and apply dimension reduction techniques on them separately. We apply the principal components procedure in SAS to our dataset specifying our variables as our 8 demographic variables and receive as an output the SCREE Plot (Figure 3), which will tell us how much cumulative variance of the data is explained by each new principal component. The SCREE Plot shows us that around 90% of the variance in the data from the demographic variables can be explained by the first 2 principal components. Similarly, the SCREE Plot obtained when running the principal components procedure for our 5 operation variables (Figure 4) reveals that nearly 90% of the variance in the data from the operation variables can be explained by the first principal component. Therefore, we will use 2 principal components to represent our demographic variables and 1 principal component to represent our operation variables.

Now that we know how many principal components we wish to use for our dimension reduction, we can apply factor analysis to do just this. Factor analysis achieves dimension reduction by imagining a dataset to be completely explained by a series of unobservable factors and estimating these factors. Just like we did for Principal Component Analysis, we choose to apply Factor Analysis on our Demographic and Operation Variables. The number of factors we choose to use to explain the dataset is equal to the number of principal components we have decided upon above, so we will choose to use 2 factors to represent our demographic variables and 1 factor to represent our operation variables.

We first apply the factor analysis SAS procedure to our dataset, specifying our variables as our 8 demographic variables, specifying our method type as principal component, our number of factors as 2 and a varimax rotation. This gives us the correlation between our 2 factors and each of the demographic variables (Figure 5) as well as the same correlations after a varimax rotation of these factors (Figure 6) which can give us further interpretation on the meaning of the factors. If we consider an absolute correlation value of 0.85 or greater as high we note that Factor 1 has high positive correlation with BEDS, ADM, and SIR while Factor 2 has high positive correlation with RBEDS and REHAB. From this, we can say that our first Demographic Factor is positively associated with number of beds, Administrative Costs and Revenue from Inpatient Procedures. Though we would have to conduct further research on this hypothesis, we could attempt to make the claim that Factor 1 is related to the relative size of a hospital, as these 3 explanatory variables would all conceivably be expected to increase in bigger hospital and decrease in smaller ones. The second demographic factor appears positively associated with number of rehab beds and the presence of a rehab unit. We could reasonably make the claim that we believe this factor is heavily related to the operation of rehab units. To summarize, we have reduced our 8 demographic variables to 2 factors which we have defined above. Additionally, the SAS procedure has also provided us with the factor scores for both factors for all of our observations. For the remainder of this project, we rename these 2 factors as Demographic Factor 1 and Demographic Factor 2.

We now run this exact same analysis for our 5 Operation Variables. We apply the factor analysis SAS procedure to our dataset specifying our variables as our 5 operation variables, specifying our method type as principal component with our number of factors specified as 1. Once again, we are provided with a table that lists the correlations between each of our operation variables and our single factor (Figure 7). Since we only have a single factor, a varimax rotation is not possible. Since we cannot use a varimax interpretation, we can only make the claim that since our factor is heavily positively correlated with all 5 variables HIP95, KNEE95, HIP96, KNEE96 and FEMUR96, we can consider it an approximation of the averages of these variables. Therefore, we say that our factor approximates the averages of our 5 operation variables. To summarize, we have reduced our 5 operation variables to 1 factor, which is we have defined above. For the remainder of this project, we will rename our factor as Operation Factor 1.

To reiterate, we have taken our 13 explanatory variables and reduced them to 3 factors (Demographic Factor 1, Demographic Factor 2 and Operation Factor 1) using factor analysis. These 3 factors will now be used as the basis of our inferences.

Cluster Analysis:

Before we begin our cluster analysis, we first reiterate that our response variable of interest is SALES. We split our data (4,703 observations) into two groups: 1. Our training dataset, which only consists of hospitals that have SALES greater than 0 (2,707 observations) and 2. Our testing dataset, which only consists of hospitals that have SALES equal to 0 (1,996 observations).

We wish to use cluster analysis to divide our multivariate training dataset into a set of natural clusters that we can use to find a cluster of hospital that have the characteristics of profitable hospitals. We conduct our Cluster Analysis with an agglomerative hierarchical clustering approach that starts by defining each data point to be a cluster and then combining existing clusters at each iteration until we have our specified number of clusters. Specifically, we use Ward's Method which chooses clusters based on the minimum increase in error sum of squares from ANOVA analysis at each iteration. We keep this in mind when we choose k, the number of clusters we want to divide our training dataset into. Staring with the value k = 8, we run the cluster, tree and glm SAS procedures on our training dataset, specifying our four factors as our explanatory variables, SALES as our response variable and Ward's Method as our clustering method, where k is our desired number of clusters. We do this from k = 8 to k = 20. At each of these iterations, the glm procedure gives us an R^2 value for each of our factors. We recorded each of these (Figure 8) and used them to determine what our optimal cluster size would be. With each additional cluster, our average R^2 value for our 3 factors increased, but by noting the amount it increased at each step, we also found that it stagnated after the 12th cluster. Therefore, we decided that 12 would be our optimal number of clusters. We-reran the cluster, tree and glm procedures specifying our number of clusters as 12. This procedure gave us an output that contained our training dataset, but with each observation having its specific cluster (1-12) specified next to it.

Now, within our training dataset, we create a binary variable called HighSales, which has a value of 1 if the given observation has a SALES value at or above the 80th percentile value of SALES within the dataset (calculated to be 5.12396) and a value of 0 if its SALES value is below the 80th percentile value. Now, we conduct summary statistics for the response variable HighSales for each of our 12 clusters. This is summarized in Figure 9, where we show the number of observations, mean of HighSales, DemoFactor1, DemoFactor2 and OperationFactor1 for each individual cluster. Note that the mean of HighSales represents the percentage of observations in a cluster that have SALES at or above the 80th percentile value. We immediately see that Cluster 11 has the largest mean value of HighSales, as roughly 47% of the hospitals in Cluster 11 have sales at or above the 80th percentile for all hospitals in the training dataset. Therefore, we believe the hospitals in Cluster 11 are the appropriate group from which our client can find profitable opportunity.

Now that we know which cluster we are interested in, we must go back to our testing dataset and determine which cluster each observation must belong to. We do this by considering the combination of our 3 factors Demographic Factor 1, Demographic Factor 2 and Operation Factor 3 as a point in space-time. We consider the average of these three factors in our training dataset at each cluster as the centroid of each cluster. For our testing dataset, we simply look at the point (DemoFactor1, DemoFactor2, OperationFactor1) and find the Euclidean distance between this point and each of the centroids. The observation belongs to the cluster whose centroid has the shortest distance to the observation. We do this for all of the observations in the testing dataset and then we filter out all observations that do not belong to Cluster 11. We do the same filtering for our training dataset. We note that our training dataset has 93 observations and our testing dataset has 16 observations. These 16 observations are the hospitals that have zero sales, despite sharing the characteristics of the cluster of data with the highest percentage of hospitals with sales greater than the 80th percentile.

Now, we wish to quantify the potential sales the client would receive if they focused their sales on these 16 hospitals. This is done by running a linear regression on Cluster 11 of our training dataset using proc reg specifying our response variable as SALES and our explanatory variables as DemoFactor1, DemoFactor2 and OperationFactor1. Once we do this, we use the proc score procedure to use the same linear regression model we just computed as a predictor in Cluster 11 of our training dataset. The estimated linear regression model found by SAS was $\widehat{SALES} = 4.243678738 - 0.879532298 * DemoFactor1 + 1.0311611957 * DemoFactor2 + 0.2015656055 * OperationFactor1$. This model had an associated root mean squared error of 1.6998856648. Using this, we added a column to our Cluster 11 testing dataset specifying the predicted value of Sales based on the regression model produced by the training dataset. Note that since we initially used the transformation $SALES = \ln(1 + SALES)$, to have our predicted and actual sales values back to their normal scale, we must reapply the transformation $SALES = \exp(SALES) - 1$ for both values. We apply these transformations and now call these variables Predicted_Sales and Actual_Sales. We also define a new variable, Potential_Gain = Predicted_Sales – Actual_Sales. For each of these 16 hospitals, we calculate the potential gain and sort them in increasing order. We tabulate this in Figure 10.

Now, we wish to quantify the potential sales that the client would receive if they focus sales on these 5 hospitals. This is done by running a linear regression of the response variable SALES with explanatory variables DemoFactor1, DemoFactor2, DemoFactor3 and OperationFactor1 exclusively for Cluster 16. We do this with the proc reg SAS procedure and we also specify that we wish to have a predicted value for each observation using the regression model. We then convert the SALES values back to their original values by applying the inverse of the initial transformation done to it. This means that we apply an exponential transformation to the SALES variables since a logarithmic transformation was initially done to them. This transformation is done both for the actual SALES values and the predicted SALES values. Then, we define the potential gain with the variable GAIN, which is equal to the difference between the predicted sales and the actual sales. The potential gain for the 5 hospitals we specified are listed in Figure 8. If we use 130 (which would translate to $130,000) as a cutoff for potential gain desired, we would recommend that our client focus their next sales campaign on the following 4 hospitals (with respective potential gains in earnings listed next to them):

5. Sioux Falls SD (HID 43065)- Potential Gain of $173,019
6. Fort Myers FL (HID 24039)- Potential Gain of $156,599
7. Buffalo NY (HID 107021)- Potential Gain of $154,486
8. Peoria IL (HID 257043)- Potential Gain of $133,103

Conclusion:

The initial goal of this project was to find a way to increase the sales of our client by using historical hospital data that contained multivariate data. Using principal component analysis and factor analysis, we reduced the 13 explanatory variables (8 Demographic Variables and 5 Operation Variables) to 3 Factors (2 Demographic Factors and 1 Operation Factor). Then, we split our data into a subset with zero sales and a subset with positive sales. We conducted cluster analysis on the subset with positive sales using 12 clusters and isolated the cluster (Cluster 11) that had the highest percentage of observations that had sales values greater than the 80th percentile of all observations through all 12 Clusters. Then, after determining, which of the hospitals in the zero sales subset would belong to Cluster 11, we found a list of 16 hospitals. Then, by using a linear regression model conducted on the Cluster 11 data with sales, we found predicted sales values for the 16 hospitals without sales. With this, we were able to calculate potential gains in profit for these 16 hospitals and after sorting this in order, we found the 4 hospitals with the highest earnings potential for our client: Sioux Falls SD, Fort Myers FL, Buffalo NY and Peoria IL with potential gains in earnings of $173,019, $156,599, $154,486 and $133,103 respectively. Therefore, we recommend that our client focus a heavily sales campaign on the 4 hospitals listed above during their next fiscal period.

Appendix:

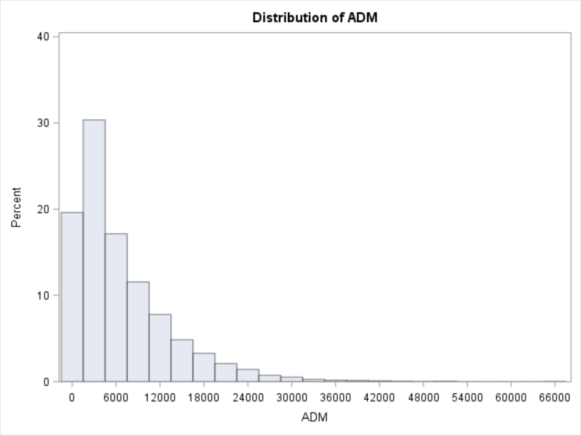


Figure 1

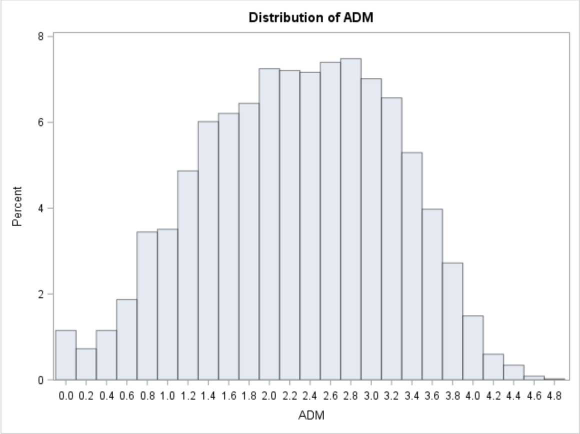


Figure 2

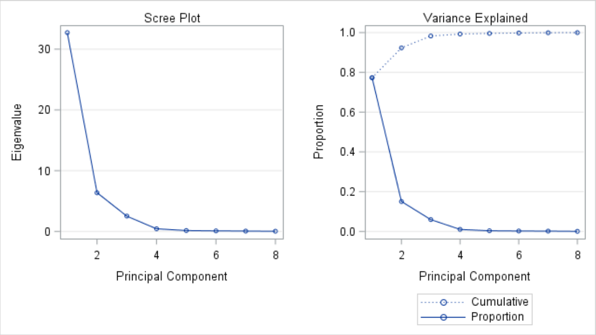


Figure 3

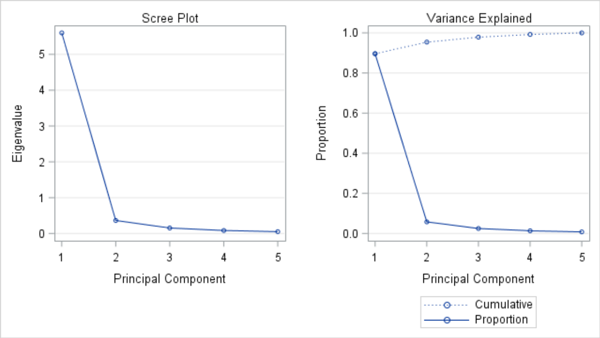
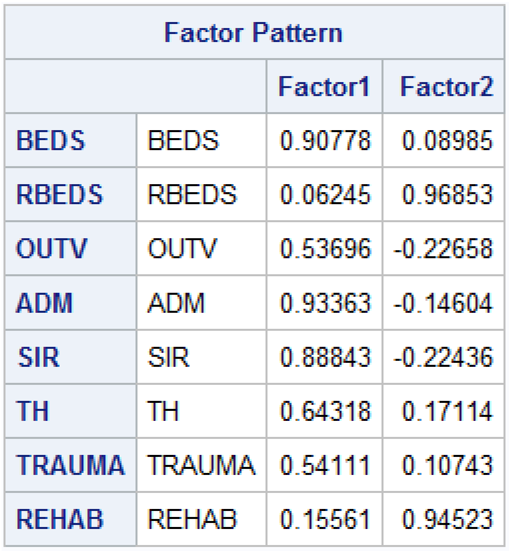
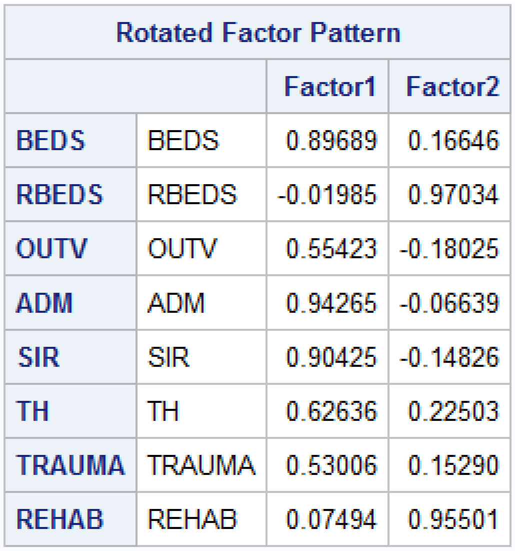


Figure 4



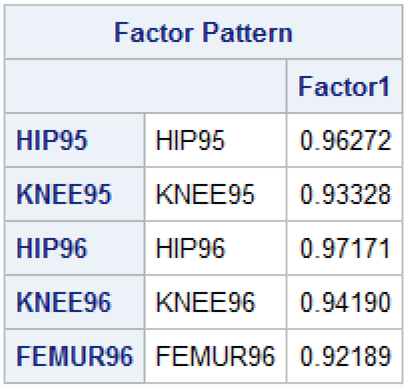
Factor Pattern			
		Factor1	Factor2
BEDS	BEDS	0.90778	0.08985
RBEDS	RBEDS	0.06245	0.96853
OUTV	OUTV	0.53696	-0.22658
ADM	ADM	0.93363	-0.14604
SIR	SIR	0.88843	-0.22436
TH	TH	0.64318	0.17114
TRAUMA	TRAUMA	0.54111	0.10743
REHAB	REHAB	0.15561	0.94523

Figure 5



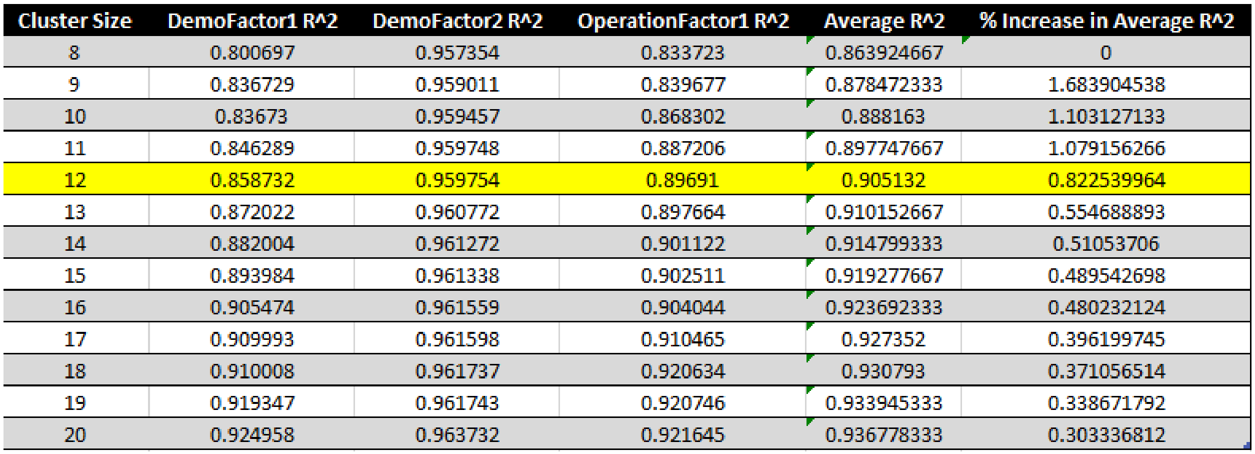
Rotated Factor Pattern			
		Factor1	Factor2
BEDS	BEDS	0.89689	0.16646
RBEDS	RBEDS	-0.01985	0.97034
OUTV	OUTV	0.55423	-0.18025
ADM	ADM	0.94265	-0.06639
SIR	SIR	0.90425	-0.14826
TH	TH	0.62636	0.22503
TRAUMA	TRAUMA	0.53006	0.15290
REHAB	REHAB	0.07494	0.95501

Figure 6



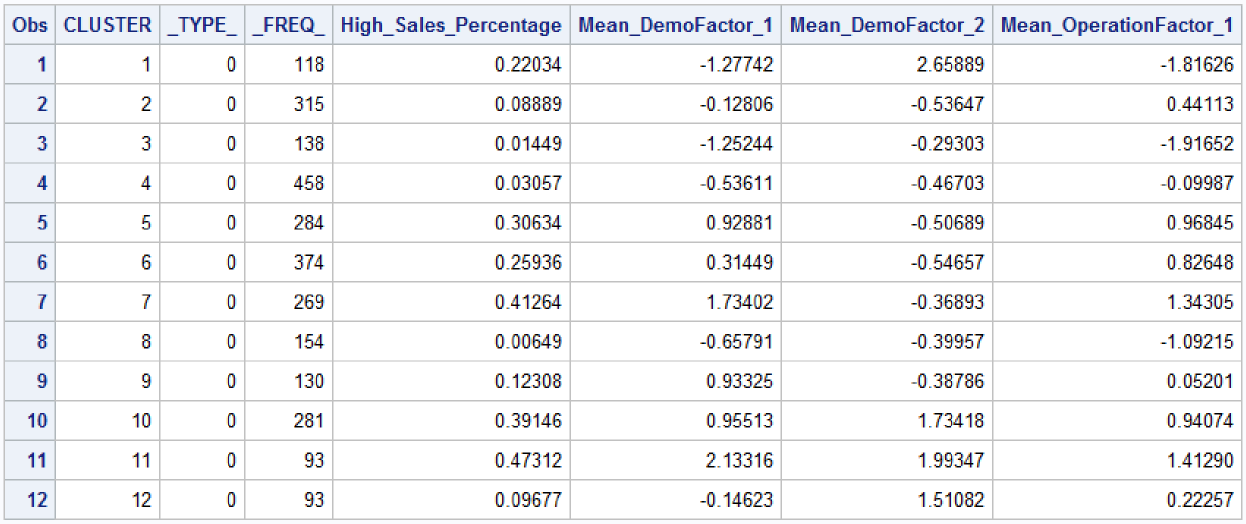
Factor Pattern		
		Factor1
HIP95	HIP95	0.96272
KNEE95	KNEE95	0.93328
HIP96	HIP96	0.97171
KNEE96	KNEE96	0.94190
FEMUR96	FEMUR96	0.92189

Figure 7



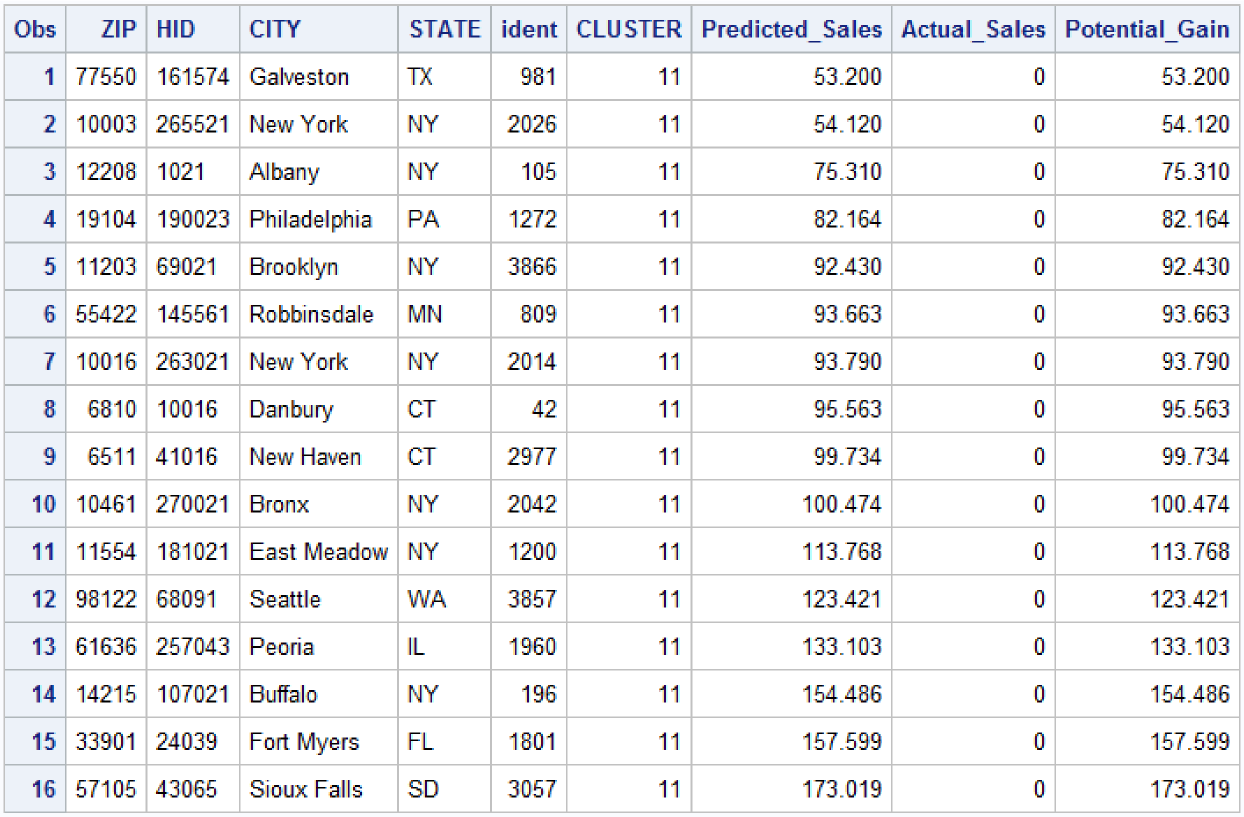
Cluster Size	DemoFactor1 R^2	DemoFactor2 R^2	OperationFactor1 R^2	Average R^2	% Increase in Average R^2
8	0.800697	0.957354	0.833723	0.863924667	0
9	0.836729	0.959011	0.839677	0.878472333	1.683904538
10	0.83673	0.959457	0.868302	0.888163	1.103127133
11	0.846289	0.959748	0.887206	0.897747667	1.079156266
12	0.858732	0.959754	0.89691	0.905132	0.822539964
13	0.872022	0.960772	0.897664	0.910152667	0.554688893
14	0.882004	0.961272	0.901122	0.914799333	0.51053706
15	0.893984	0.961338	0.902511	0.919277667	0.489542698
16	0.905474	0.961559	0.904044	0.923692333	0.480232124
17	0.909993	0.961598	0.910465	0.927352	0.396199745
18	0.910008	0.961737	0.920634	0.930793	0.371056514
19	0.919347	0.961743	0.920746	0.933945333	0.338671792
20	0.924958	0.963732	0.921645	0.936778333	0.303336812

Figure 8



Obs	CLUSTER	_TYPE_	_FREQ_	High_Sales_Percentage	Mean_DemoFactor_1	Mean_DemoFactor_2	Mean_OperationFactor_1
1	1	0	118	0.22034	-1.27742	2.65889	-1.81626
2	2	0	315	0.08889	-0.12806	-0.53647	0.44113
3	3	0	138	0.01449	-1.25244	-0.29303	-1.91652
4	4	0	458	0.03057	-0.53611	-0.46703	-0.09987
5	5	0	284	0.30634	0.92881	-0.50689	0.96845
6	6	0	374	0.25936	0.31449	-0.54657	0.82648
7	7	0	269	0.41264	1.73402	-0.36893	1.34305
8	8	0	154	0.00649	-0.65791	-0.39957	-1.09215
9	9	0	130	0.12308	0.93325	-0.38786	0.05201
10	10	0	281	0.39146	0.95513	1.73418	0.94074
11	11	0	93	0.47312	2.13316	1.99347	1.41290
12	12	0	93	0.09677	-0.14623	1.51082	0.22257

Figure 9



Obs	ZIP	HID	CITY	STATE	ident	CLUSTER	Predicted_Sales	Actual_Sales	Potential_Gain
1	77550	161574	Galveston	TX	981	11	53.200	0	53.200
2	10003	265521	New York	NY	2026	11	54.120	0	54.120
3	12208	1021	Albany	NY	105	11	75.310	0	75.310
4	19104	190023	Philadelphia	PA	1272	11	82.164	0	82.164
5	11203	69021	Brooklyn	NY	3866	11	92.430	0	92.430
6	55422	145561	Robbinsdale	MN	809	11	93.663	0	93.663
7	10016	263021	New York	NY	2014	11	93.790	0	93.790
8	6810	10016	Danbury	CT	42	11	95.563	0	95.563
9	6511	41016	New Haven	CT	2977	11	99.734	0	99.734
10	10461	270021	Bronx	NY	2042	11	100.474	0	100.474
11	11554	181021	East Meadow	NY	1200	11	113.768	0	113.768
12	98122	68091	Seattle	WA	3857	11	123.421	0	123.421
13	61636	257043	Peoria	IL	1960	11	133.103	0	133.103
14	14215	107021	Buffalo	NY	196	11	154.486	0	154.486
15	33901	24039	Fort Myers	FL	1801	11	157.599	0	157.599
16	57105	43065	Sioux Falls	SD	3057	11	173.019	0	173.019

Figure 10