An Analysis on Increasing the Sales of Orthopedic Equipment to High-Consuming Potential Customers

Name: Annapurna Atkuru

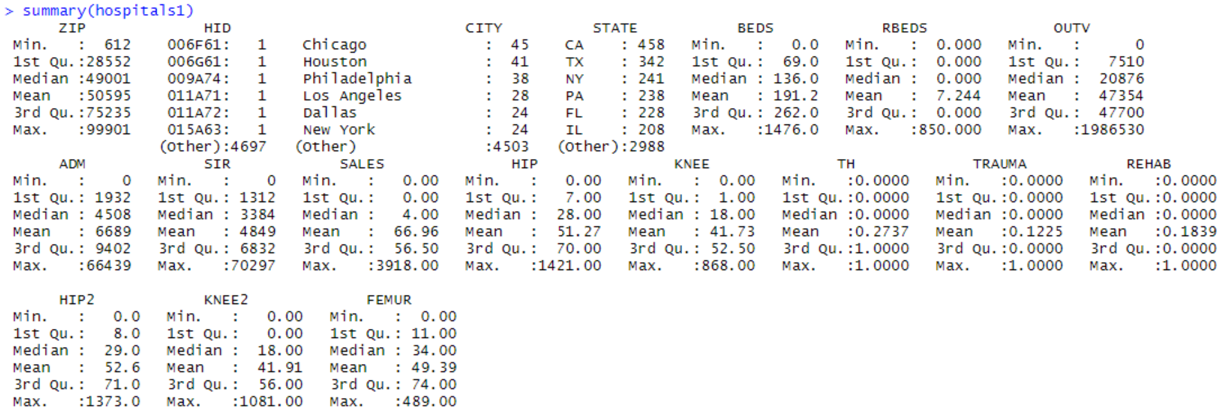
Executive Summary:

After examining the dataset containing information on hospitals across the United States that consume orthopedic equipment (but have $0 in sales), and performing a clustering analysis and k-medoids algorithm, I found that the highest potential sales gain that can be made by selling to hospitals in cluster 13 is approximately $5,010 per year. Hospitals in this cluster are expected to generate on average the highest amount of sales, and also on average have a high number of hospital operations and need for equipment that would potentially make them a profitable customer (or set of customers). Hospitals in cluster 13 were shown on average to require a higher number of beds, have a higher number of outpatient visits, higher revenue from inpatients, and a greater number of operations on hip, knee and femur (**Table 12**)

Introduction

In this case study, I analyze a dataset on hospitals across the United States to find ways to increase sales of orthopedic products to potential customers who have high consumption of such equipment but currently do not purchase this equipment from our company. The objective of this study is to identify a few hospitals where the potential gains in revenues from these hospitals would be advantageous to our company. In order to do so, I needed to identify hospitals that potentially generate a high amount of sales, but would also have a high demand for the orthopedic products our company specializes in.

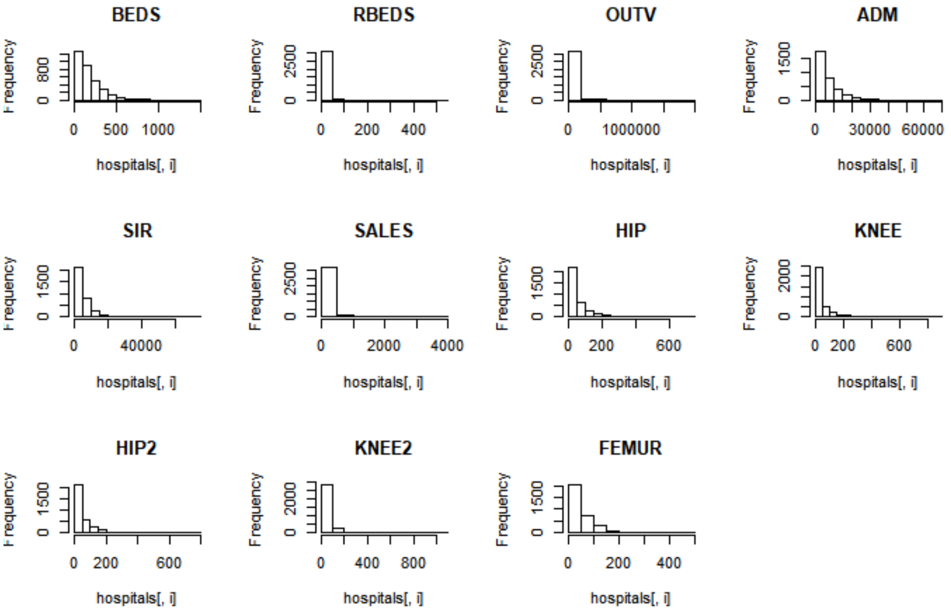
To begin this analysis, I perform a simple statistical summary of the hospitalUSA data to gain a better understanding of what the distribution of the data looks like (**Table 1**). From this summary, we understand that across the pertinent variables (i.e. bed, rbeds, outv, adm, sir, sales, hip, knee, trauma, rehab, hip2, knee and femur), we need to identify where the level of operations performed is high, but where our sales are currently zero. In order to do this, I created a subsection ("subset") of the larger hospitalUSA dataset into a smaller dataset that contained approximately 70% of the observations from the original dataset. The purpose of creating this subsection is to be able to use the estimates and conclusions from the subset to infer the parent population. Once the data was sub sectioned, I separated the variables of interest in two groups: demographics and operation numbers. Upon separation, I analyzed the distributions of these two groups of variables to observe abnormalities or skewness in the data (**Figure 1, Figure 2**)



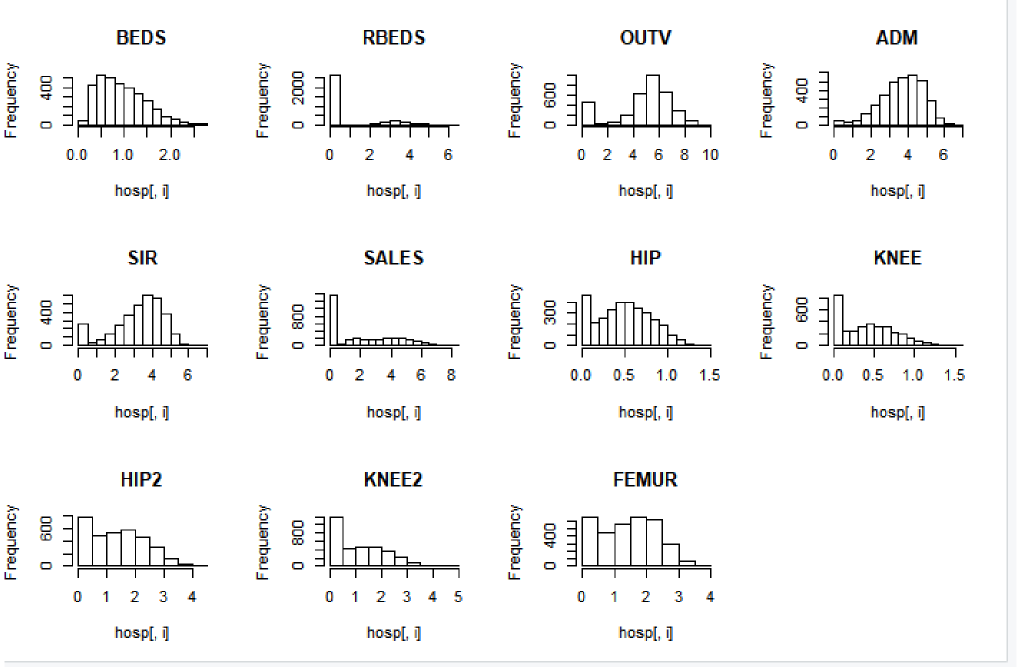
```
> summary(hospitals1)
      ZIP                HID                               CITY            STATE            BEDS            RBEDS               OUTV        
 Min.   :  612   006F61:   1   Chicago                    :  45   CA     : 458   Min.   :   0.0   Min.   :  0.000   Min.   :      0  
 1st Qu.:28552   006G61:   1   Houston                    :  41   TX     : 342   1st Qu.:  69.0   1st Qu.:  0.000   1st Qu.:   7510  
 Median :49001   009A74:   1   Philadelphia               :  38   NY     : 241   Median : 136.0   Median :  0.000   Median :  20876  
 Mean   :50595   011A71:   1   Los Angeles                :  28   PA     : 238   Mean   : 191.2   Mean   :  7.244   Mean   :  47354  
 3rd Qu.:75235   011A72:   1   Dallas                     :  24   FL     : 228   3rd Qu.: 262.0   3rd Qu.:  0.000   3rd Qu.:  47700  
 Max.   :99901   015A63:   1   New York                   :  24   IL     : 208   Max.   :1476.0   Max.   :850.000   Max.   :1986530  
                 (Other):4697  (Other)                    :4503   (Other):2988                                                       
      ADM             SIR             SALES               HIP               KNEE              TH             TRAUMA           REHAB       
 Min.   :    0   Min.   :    0   Min.   :   0.00   Min.   :   0.00   Min.   :  0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.: 1932   1st Qu.: 1312   1st Qu.:   0.00   1st Qu.:   7.00   1st Qu.:  1.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median : 4508   Median : 3384   Median :   4.00   Median :  28.00   Median : 18.00   Median :0.0000   Median :0.0000   Median :0.0000  
 Mean   : 6689   Mean   : 4849   Mean   :  66.96   Mean   :  51.27   Mean   : 41.73   Mean   :0.2737   Mean   :0.1225   Mean   :0.1839  
 3rd Qu.: 9402   3rd Qu.: 6832   3rd Qu.:  56.50   3rd Qu.:  70.00   3rd Qu.: 52.50   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
 Max.   :66439   Max.   :70297   Max.   :3918.00   Max.   :1421.00   Max.   :868.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
                                                                                                                                         
      HIP2             KNEE2              FEMUR       
 Min.   :   0.0   Min.   :   0.00   Min.   :  0.00  
 1st Qu.:   8.0   1st Qu.:   0.00   1st Qu.: 11.00  
 Median :  29.0   Median :  18.00   Median : 34.00  
 Mean   :  52.6   Mean   :  41.91   Mean   : 49.39  
 3rd Qu.:  71.0   3rd Qu.:  56.00   3rd Qu.: 74.00  
 Max.   :1373.0   Max.   :1081.00   Max.   :489.00  
```

The distribution of the variables were revealed to be heavily skewed, so a corresponding transformation needed to be performed to normalize the variables and gain better visualization (**Figure 2**).

Distribution of Variables Pre-Transformation



Distribution of Variables Post-Transformation



After transforming the variables of interest, I summarize the demographic and operation number variables through dimension reduction and factor analysis to find a reduced list of factor variables that would be useful to summarize and explain the model we will extract to make predictions on the sales variable (dependent variable of interest). I removed the dependent variable in our dataset to ensure we would only be able to reduce our independent variables to factors.The results of the factor analysis will be especially important to us in regards to interpreting the output of the clustering analysis. Once the optimal number of factors are found, we divide the list of hospitals into market subsets through clustering. We perform hierarchical clustering on a combined dataset of our factor scores and the dataset with the independent variables (hosp). By doing so, we are able to group the hospitals based on their similar characteristics, and find an optimal number of clusters which adequately characterize all of our observations using the factor scores we calculated previously. This will help us to further visualize which hospitals would be profitable potential customers based on their cluster characteristics.

Once we find the optimal number of clusters, we can then identify which clusters would contain hospitals that would be the most profitable to us. We then recluster our data to find the highest potential gain in sales using the k-means and k-medoids algorithm respectively. The k-means clustering algorithm helps us identify the optimal number of clusters, upon which the k-medoids algorithm is performed to find the means of all of the variables within a specific cluster. We analyze these values to understand where our potential sales gain could be the greatest and which cluster of hospitals to focus on.

Analysis and Results

Summary of Analysis Procedure

Step 1: Calculate and display descriptive statistics and perform a simple data analysis to gain a simple understanding of sample population.
Step 2: Perform variable transformation to gain better visualization of our independent variables and understand it's distribution.
Step 3: Perform dimension reduction and factor analysis to summarize independent variables of interest into factor variables.
Step 4: Perform hierarchical clustering to identify the optimal number of market segments to analyze, and perform a basic analysis of cluster performance.
Step 5: Recluster and perform predictive analysis to identify and estimate potential sales gain.
Step 6: Summarize results and findings in your conclusion.

Step 1: We first begin our analysis of the hospital dataset by analyzing a summary of the dataset to identify the important variables of interest and the dependent variable. Of the variables we are provided in the dataset, we can hypothesize that many of them are related or behave in similar ways based on the distribution of their values (**Table 1)**. At this stage, it is imperative to understand the similarity of these variables in order to identify the characteristics by which we can evaluate the

level of operations or demand for orthopedic equipment a potential customer may exhibit or need, so we divide the variables into their respective variable groups (**Table 3, Table 4**).

Step 2: Our assumption of related variables behaving similarly is validated through the initial histogram and summary of the variables shown in (**Figure 1**). Although all of the variables are heavily skewed, we see the magnitude of skewness to be greater for the demographic variables compared to the operation number variables. Our hypothesis at this stage is that we can gain better visualization of the variables if they were normalized. This not only allows us to standardize the distribution of the variables, but also allows us to make more accurate comparisons and understand the behavior of each independent variable. We normalize the variable through a series of logarithmic and square root transformations to obtain a bell-shaped distribution that is approximately normal. (**Figure 2**).

Step 3: Once we have transformed our variables of interest, we will need to identify which of them are most important for predicting the potential sales gain of a customer. Since we have approximately 10 variables of interest (excluding sales which is our dependent variable), it is safe to assume that our data, as it is, would be very difficult to visualize and extract meaningful information from since it is highly dimensional. Ultimately we would like to analyze these variables to help us interpret which potential customers have large scale operations and would therefore demand more orthopedic equipment. At this stage we hypothesize that all of these variables do not necessarily contribute explanatory power to our model, and can therefore be reduced into significant factor variables, which would effectively summarize these variables. To validate this hypothesis, we perform factor analysis (**Table 5**) to reduce the dimensions for our model. It is important to understand that factor analysis is a measurement of the latent variable within the model.

Since we want to summarize our variables into meaningful factor variables, it is important to reduce the number of factor variables we extract from our factor analysis. By default, we therefore choose to reduce to 3 factor variables.We can verify that this is the correct choice of factor variables, because when we perform the factor analysis, we see that the hypothesis tests produce a very low p-value (0). This indicates that 3 factor variables are statistically significant. When we perform this factor analysis, we want to exclude the sales variable, which will act as the independent variable for potential sales gain. We also perform a varimax rotation to create a regression factor analysis. We visualize the grouping of our independent variables into factors through (**Figure 3**) a factor plot. Ultimately, we want to see low uniqueness values when interpreting our factors - this indicates how well our variables were able to be "summarized" by the factor. Our factor analysis also produces factor scores, which will help us find the optimal number of clusters when we perform our clustering analysis. When we analyze our factor scores, we see that factors 1 and 2 are larger than factor 3 (**Table 7**). This will prove important in understanding how our hospitals are grouped in clustering based on their factor scores, and other similar characteristics. We can visualize the value of these factors through this plot (**Figure 3, Figure 4**), where we see higher values for factor 1 and 2.

Step 4: After we find the factor scores from our factor analysis, we need to use them to group observations by clusters and determine similarity between hospitals (i.e. market segments). Finding the ideal number of clusters is imperative to developing predictive models because models need to be able to accurately group observations based on certain characteristics into their appropriate clusters. This allows us to better analyze which clusters would potentially generate the highest amount of sales, and which clusters we can ignore in our analysis.

In order to find the ideal number of clusters, we first need to combine our hospital dataset (excluding the independent variable of sales), along with our factor scores from our factor analysis. If we assume that we will cluster our observations using factor scores, we cannot solely cluster factor scores as this will not cluster the entire dataset, which we are using to analyze which hospitals will have the highest sales potential. Once we group these two datasets, we perform hierarchical clustering using Ward's method to identify the optimal number of clusters. We use the second derivative to identify the optimal value for k, which from (**Table 9),** we see is 5.

How can we be sure that 5 is the optimal number of clusters for our dataset? At this point, in order to verify that we have the optimal number of clusters, we should perform another clustering analysis to see if our clustering methods provide different values for k. We use the fviz_nbclust() method to visualize the optimal number of clusters (**Figure 6**). We see that this method, using the silhouette statistic, also produces 5 as the optimal number of clusters. From this analysis, we can be sure that we have selected the optimal number of clusters to group our data by.

Once we have found the optimal number of clusters, it is imperative to understand the content of each cluster to interpret how each cluster behaves, and identify the cluster with the highest sales, so we can focus on that cluster for our analysis. In order to do so, we "re-cluster" with the specified optimal number of clusters, and produce a boxplot to show the highest sales per cluster. We find cluster 3 to have the highest average amount of sales (**Figure 7**), thereby making it the most optimal cluster. We can verify these results numerically by using the describeBy function to obtain a summary of the different clusters. This summary shows that cluster 3 has a higher mean value of sales compared to the other clusters. We can also see the distribution of observations in this cluster, which compared to other clusters, is lower and therefore more ideal.

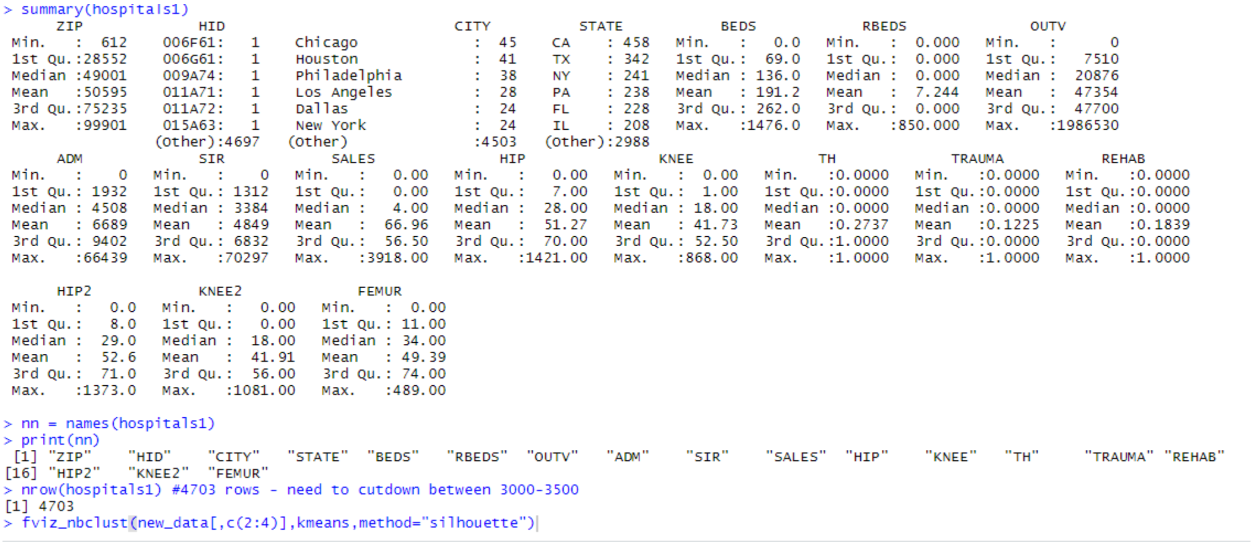
Step 5: Once we have determined the optimal level of clustering from our hospital dataset using the factor scores, we want to extrapolate this for our final analysis by looking between 15 to 30 clusters to identify the optimal number of clusters. If our clusters are inherently small (n<100), we may assume that the average sales of similarly performing clusters is just the average of each cluster itself. In order to test this assumption we perform clustering analysis again using the NbClust() function and k-means method, we find the next optimal number of clusters to be 16, as evidenced by (**Table 12**). We next need to determine how to find the potential gain in sales, which is defined as the difference between current sales and average sales. Because we need to identify customers who do not have a previous sales record with us, we need a testing data set to train a predictive model to determine the average amount of sales per cluster, which can then be used to determine which cluster has the more “rewarding” hospitals. We create a testing data set from our original dataset, hospitals, where hospital sales are equal to 0. We then use PAM, a robust clustering technique, to estimate and predict clusters with the highest average sales using our testing dataset. We then perform k-medoids to then find the medioid of each cluster for all of our independent variables, which we can use to determine the highest average sales per cluster. After viewing the output from k-medoids, we can see that cluster 13 (**Table 12**) had the highest sales mean. This is also verified by the boxplot produced by k-means (**Table 12**). From this analysis, we understand that cluster 13 would produce the highest average potential sales gain, and that it would be an advantageous decision for our company to prioritize hospitals within this cluster.

Conclusion

Based on the results from our k-medoids algorithm on our testing dataset, as well as the boxplot from the k-means clustering procedure, we see that it would be advantageous to focus on selling orthopedic equipment to hospitals in cluster 13, which had the overall highest average sales. By limiting the number of clusters within which we analyze the average amount of sales, we can reduce the number of hospitals per cluster to obtain more accurate results of which hospitals would be interested in purchasing our products versus which hospitals wouldn’t. Cluster 13 has the highest potential for sales of our products as the expected sales potential for hospitals in this cluster would be around $5,010. The pam statistic and the k-medoids algorithm also calculated the average consumption/usage rates for the other independent variables in the cluster. Compared to the other clusters we observed in our testing set, we also see that cluster 13 has overall higher mean values for all of the independent variables, which can indicate that hospitals in this cluster are also operating on a larger scale, therefore making them more profitable or advantageous to sell to.

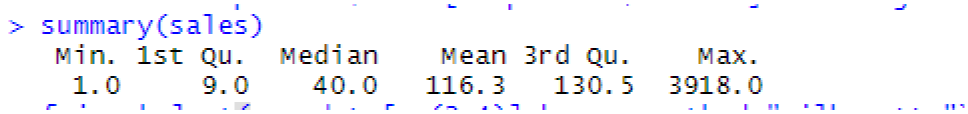
Appendix

Table 1: Summary of Hospital Data



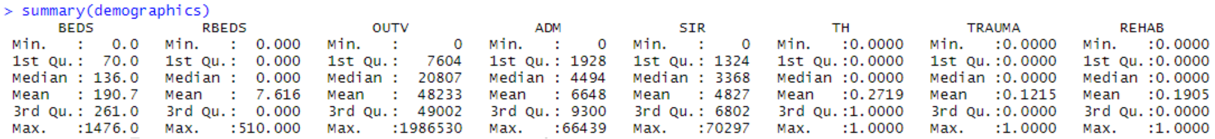
```
> summary(hospitals1)
      ZIP              HID                                    CITY           STATE            BEDS              RBEDS               OUTV        
 Min.   :  612   006F61:   1    Chicago                     :  45    CA     : 458   Min.   :   0.0   Min.   :  0.000   Min.   :      0  
 1st Qu.:28552   006G61:   1    Houston                     :  41    TX     : 342   1st Qu.:  69.0   1st Qu.:  0.000   1st Qu.:   7510  
 Median :49001   009A74:   1    Philadelphia                :  38    NY     : 241   Median : 136.0   Median :  0.000   Median :  20876  
 Mean   :50595   011A71:   1    Los Angeles                 :  28    PA     : 238   Mean   : 191.2   Mean   :  7.244   Mean   :  47354  
 3rd Qu.:75235   011A72:   1    Dallas                      :  24    FL     : 228   3rd Qu.: 262.0   3rd Qu.:  0.000   3rd Qu.:  47700  
 Max.   :99901   015A63:   1    New York                    :  24    IL     : 208   Max.   :1476.0   Max.   :850.000   Max.   :1986530  
                (Other):4697   (Other)                      :4503   (Other):2988                                                         
      ADM              SIR              SALES               HIP                KNEE               TH             TRAUMA            REHAB       
 Min.   :    0   Min.   :    0   Min.   :   0.00   Min.   :   0.00   Min.   :  0.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.: 1932   1st Qu.: 1312   1st Qu.:   0.00   1st Qu.:   7.00   1st Qu.:  1.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median : 4508   Median : 3384   Median :   4.00   Median :  28.00   Median : 18.00   Median :0.0000   Median :0.0000   Median :0.0000  
 Mean   : 6689   Mean   : 4849   Mean   :  66.96   Mean   :  51.27   Mean   : 41.73   Mean   :0.2737   Mean   :0.1225   Mean   :0.1839  
 3rd Qu.: 9402   3rd Qu.: 6832   3rd Qu.:  56.50   3rd Qu.:  70.00   3rd Qu.: 52.50   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
 Max.   :66439   Max.   :70297   Max.   :3918.00   Max.   :1421.00   Max.   :868.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
                                                                                                                                             
      HIP2              KNEE2              FEMUR       
 Min.   :   0.0   Min.   :   0.00   Min.   :  0.00  
 1st Qu.:   8.0   1st Qu.:   0.00   1st Qu.: 11.00  
 Median :  29.0   Median :  18.00   Median : 34.00  
 Mean   :  52.6   Mean   :  41.91   Mean   : 49.39  
 3rd Qu.:  71.0   3rd Qu.:  56.00   3rd Qu.: 74.00  
 Max.   :1373.0   Max.   :1081.00   Max.   :489.00  
                                                    
> nn = names(hospitals1)
> print(nn)
 [1] "ZIP"    "HID"    "CITY"   "STATE"  "BEDS"   "RBEDS"  "OUTV"   "ADM"    "SIR"    "SALES"  "HIP"    "KNEE"   "TH"     "TRAUMA" "REHAB" 
[16] "HIP2"   "KNEE2"  "FEMUR" 
> nrow(hospitals1) #4703 rows - need to cutdown between 3000-3500
[1] 4703
> fviz_nbclust(new_data[,c(2:4)],kmeans,method="silhouette")
```

Table 2: Summary of Hospital Sales



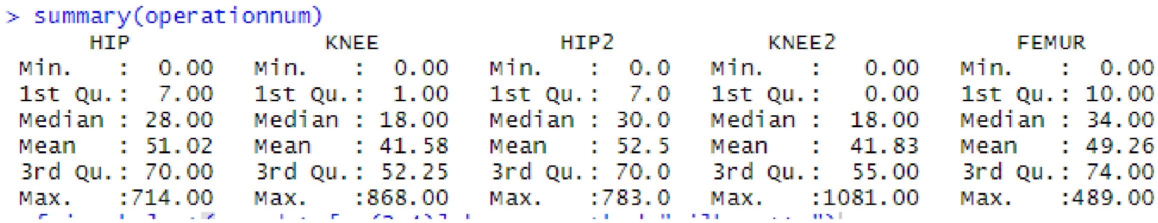
```
> summary(sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    1.0     9.0    40.0   116.3   130.5  3918.0 
```

Table 3: Demographics Summary Data



```
> summary(demographics)
      BEDS            RBEDS              OUTV              ADM              SIR              TH             TRAUMA            REHAB       
 Min.   :   0.0   Min.   :  0.000   Min.   :      0   Min.   :    0   Min.   :    0   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:  70.0   1st Qu.:  0.000   1st Qu.:   7604   1st Qu.: 1928   1st Qu.: 1324   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median : 136.0   Median :  0.000   Median :  20807   Median : 4494   Median : 3368   Median :0.0000   Median :0.0000   Median :0.0000  
 Mean   : 190.7   Mean   :  7.616   Mean   :  48233   Mean   : 6648   Mean   : 4827   Mean   :0.2719   Mean   :0.1215   Mean   :0.1905  
 3rd Qu.: 261.0   3rd Qu.:  0.000   3rd Qu.:  49002   3rd Qu.: 9300   3rd Qu.: 6802   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
 Max.   :1476.0   Max.   :510.000   Max.   :1986530   Max.   :66439   Max.   :70297   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
```

Table 4: Operation Number Summary Data



```
> summary(operationnum)
      HIP              KNEE             HIP2            KNEE2              FEMUR
 Min.   :  0.00   Min.   :  0.00   Min.   :  0.0   Min.   :   0.00   Min.   :  0.00
 1st Qu.:  7.00   1st Qu.:  1.00   1st Qu.:  7.0   1st Qu.:   0.00   1st Qu.: 10.00
 Median : 28.00   Median : 18.00   Median : 30.0   Median :  18.00   Median : 34.00
 Mean   : 51.02   Mean   : 41.58   Mean   : 52.5   Mean   :  41.83   Mean   : 49.26
 3rd Qu.: 70.00   3rd Qu.: 52.25   3rd Qu.: 70.0   3rd Qu.:  55.00   3rd Qu.: 74.00
 Max.   :714.00   Max.   :868.00   Max.   :783.0   Max.   :1081.00   Max.   :489.00
```

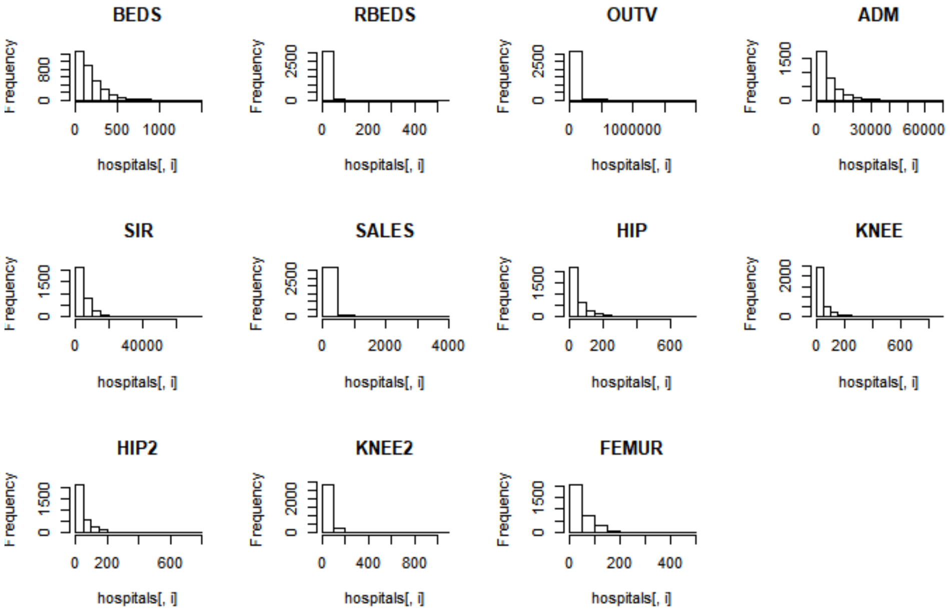


Figure 1: Variable Histogram

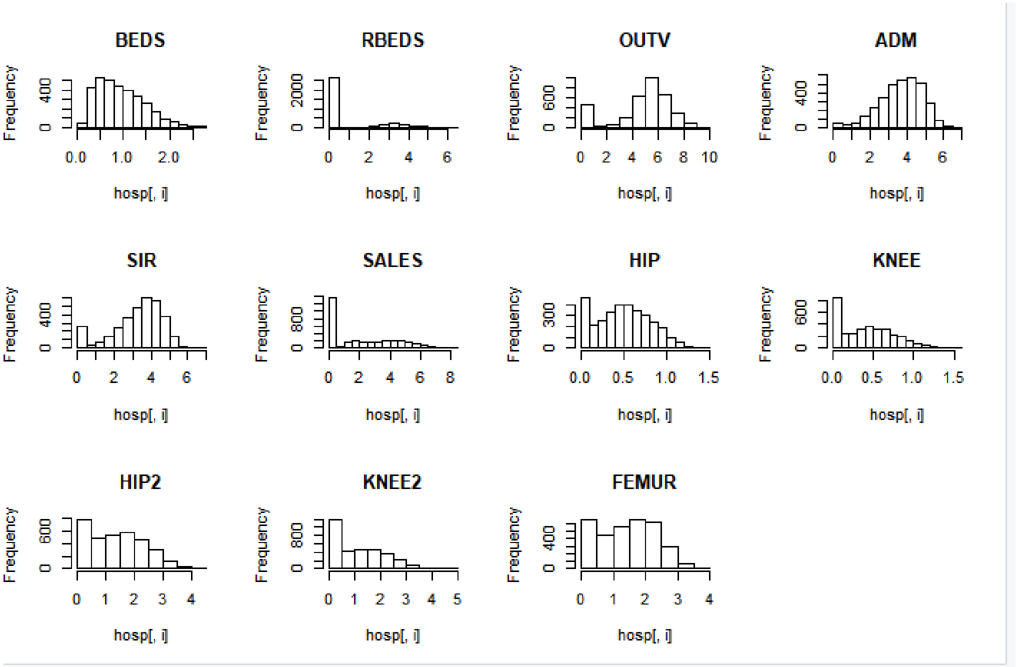
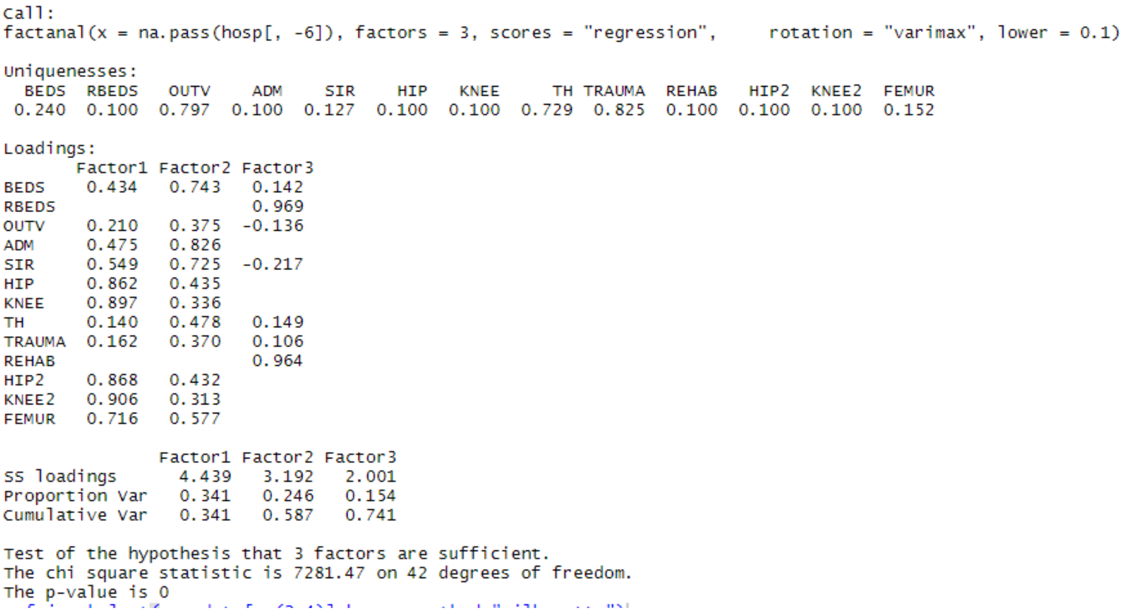


Figure 2: Variable Transformation Histogram

Table 5: Summary Results of Factor Analysis



```
Call:
factanal(x = na.pass(hosp[, -6]), factors = 3, scores = "regression",     rotation = "varimax", lower = 0.1)

Uniquenesses:
  BEDS  RBEDS   OUTV    ADM    SIR    HIP   KNEE     TH TRAUMA  REHAB   HIP2  KNEE2  FEMUR 
 0.240  0.100  0.797  0.100  0.127  0.100  0.100  0.729  0.825  0.100  0.100  0.100  0.152 

Loadings:
       Factor1 Factor2 Factor3
BEDS    0.434   0.743   0.142 
RBEDS                   0.969 
OUTV    0.210   0.375  -0.136 
ADM     0.475   0.826         
SIR     0.549   0.725  -0.217 
HIP     0.862   0.435         
KNEE    0.897   0.336         
TH      0.140   0.478   0.149 
TRAUMA  0.162   0.370   0.106 
REHAB                   0.964 
HIP2    0.868   0.432         
KNEE2   0.906   0.313         
FEMUR   0.716   0.577         

               Factor1 Factor2 Factor3
SS loadings      4.439   3.192   2.001
Proportion Var   0.341   0.246   0.154
Cumulative Var   0.341   0.587   0.741

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 7281.47 on 42 degrees of freedom.
The p-value is 0 
```

Figure 3: Factor Analysis Plot

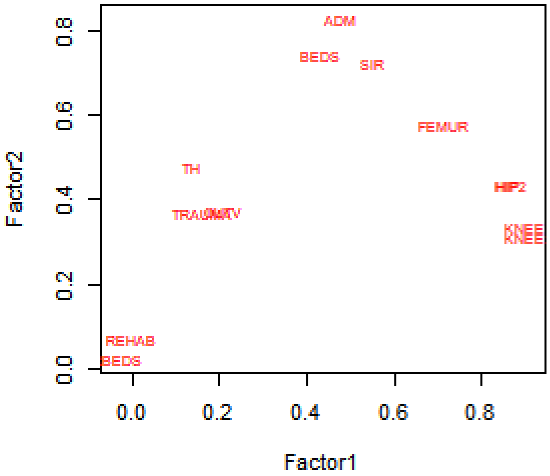
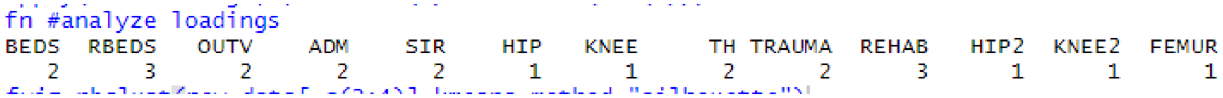
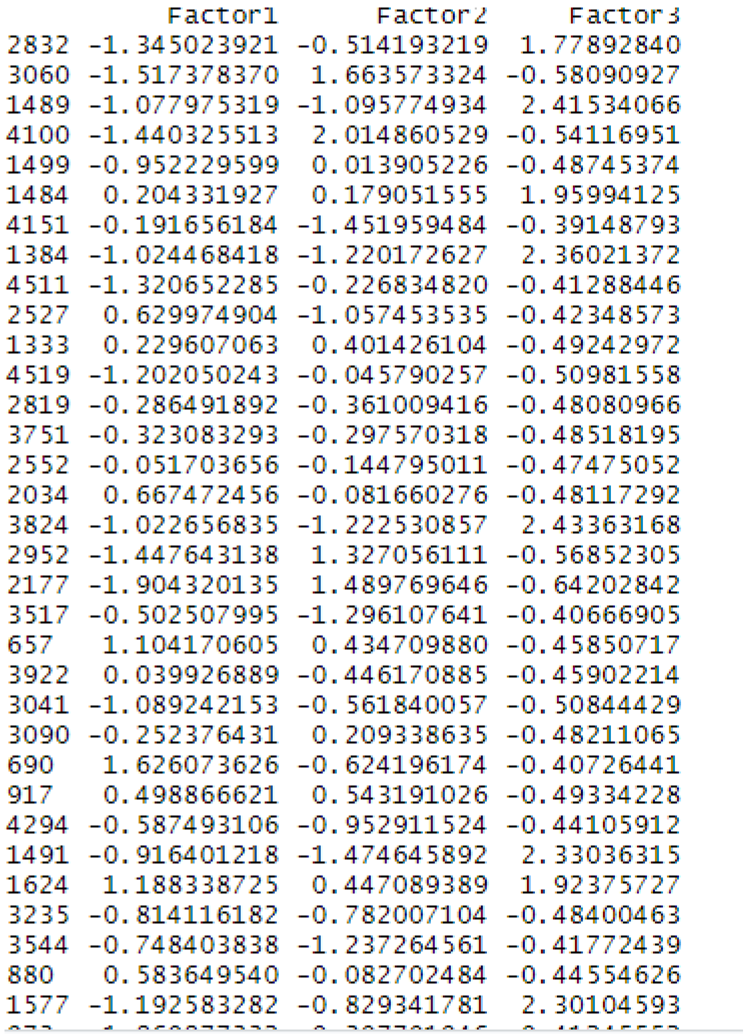


Table 6: Analysis of Loading Data



```
fn #analyze loadings
BEDS   RBEDS    OUTV     ADM     SIR     HIP    KNEE      TH  TRAUMA   REHAB    HIP2   KNEE2   FEMUR
    2       3       2       2       2       1       1       2       2       3       1       1       1
```

Table 7: Preview of Factor Scores



```
         Factor1      Factor2     Factor3
2832 -1.345023921 -0.514193219  1.77892840
3060 -1.517378370  1.663573324 -0.58090927
1489 -1.077975319 -1.095774934  2.41534066
4100 -1.440325513  2.014860529 -0.54116951
1499 -0.952229599  0.013905226 -0.48745374
1484  0.204331927  0.179051555  1.95994125
4151 -0.191656184 -1.451959484 -0.39148793
1384 -1.024468418 -1.220172627  2.36021372
4511 -1.320652285 -0.226834820 -0.41288446
2527  0.629974904 -1.057453535 -0.42348573
1333  0.229607063  0.401426104 -0.49242972
4519 -1.202050243 -0.045790257 -0.50981558
2819 -0.286491892 -0.361009416 -0.48080966
3751 -0.323083293 -0.297570318 -0.48518195
2552 -0.051703656 -0.144795011 -0.47475052
2034  0.667472456 -0.081660276 -0.48117292
3824 -1.022656835 -1.222530857  2.43363168
2952 -1.447643138  1.327056111 -0.56852305
2177 -1.904320135  1.489769646 -0.64202842
3517 -0.502507995 -1.296107641 -0.40666905
657   1.104170605  0.434709880 -0.45850717
3922  0.039926889 -0.446170885 -0.45902214
3041 -1.089242153 -0.561840057 -0.50844429
3090 -0.252376431  0.209338635 -0.48211065
690   1.626073626 -0.624196174 -0.40726441
917   0.498866621  0.543191026 -0.49334228
4294 -0.587493106 -0.952911524 -0.44105912
1491 -0.916401218 -1.474645892  2.33036315
1624  1.188338725  0.447089389  1.92375727
3235 -0.814116182 -0.782007104 -0.48400463
3544 -0.748403838 -1.237264561 -0.41772439
880   0.583649540 -0.082702484 -0.44554626
1577 -1.192583282 -0.829341781  2.30104593
```

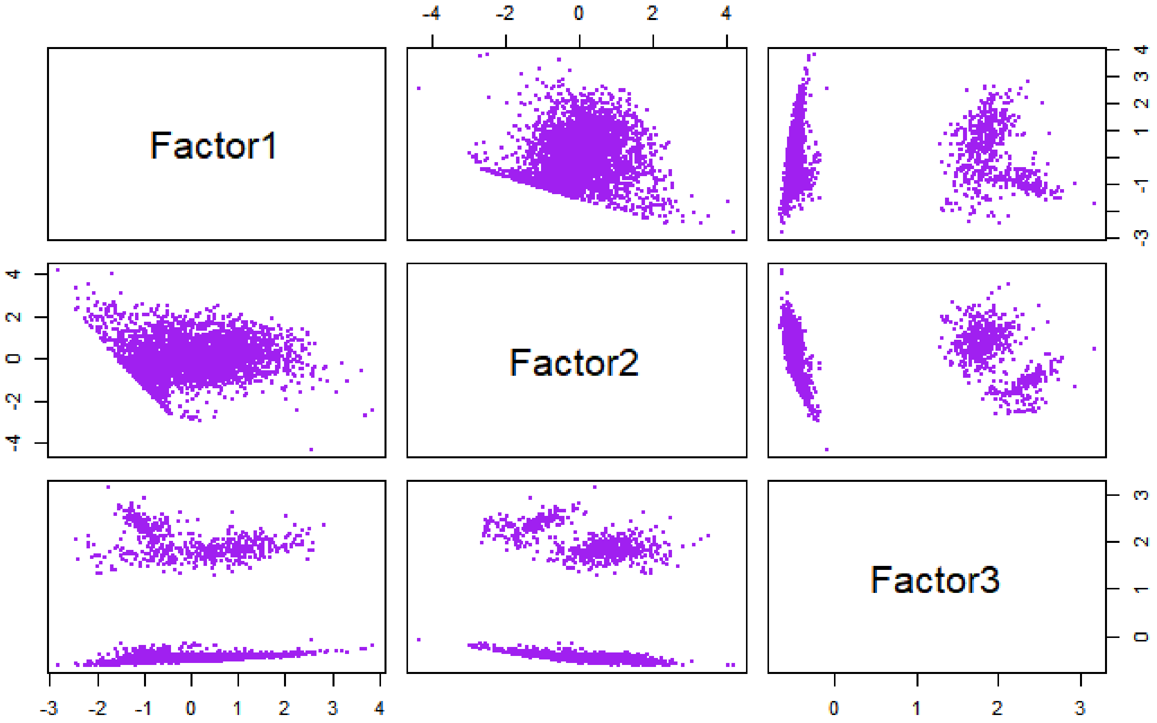
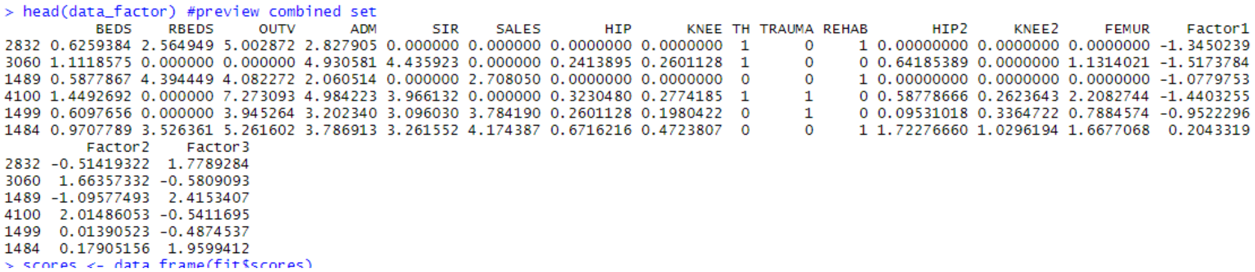


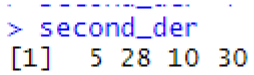
Figure 4: Factor Plot

Table 8: Combined Clustering Dataset (hosp,fit$scores)



```
> head(data_factor) #preview combined set
          BEDS    RBEDS     OUTV      ADM      SIR    SALES       HIP      KNEE TH TRAUMA REHAB       HIP2     KNEE2     FEMUR    Factor1
2832 0.6259384 2.564949 5.002872 2.827905 0.000000 0.000000 0.0000000 0.0000000  1      0     1 0.00000000 0.0000000 0.0000000 -1.3450239
3060 1.1118575 0.000000 0.000000 4.930581 4.435923 0.000000 0.2413895 0.2601128  1      0     0 0.64185389 0.0000000 1.1314021 -1.5173784
1489 0.5877867 4.394449 4.082272 2.060514 0.000000 2.708050 0.0000000 0.0000000  0      0     1 0.00000000 0.0000000 0.0000000 -1.0779753
4100 1.4492692 0.000000 7.273093 4.984223 3.966132 0.000000 0.3230480 0.2774185  1      1     0 0.58778666 0.2623643 2.2082744 -1.4403255
1499 0.6097656 0.000000 3.945264 3.202340 3.096030 3.784190 0.2601128 0.1980422  0      1     0 0.09531018 0.3364722 0.7884574 -0.9522296
1484 0.9707789 3.526361 5.261602 3.786913 3.261552 4.174387 0.6716216 0.4723807  0      0     1 1.72276660 1.0296194 1.6677068  0.2043319
         Factor2    Factor3
2832 -0.51419322  1.7789284
3060  1.66357332 -0.5809093
1489 -1.09577493  2.4153407
4100  2.01486053 -0.5411695
1499  0.01390523 -0.4874537
1484  0.17905156  1.9599412
> scores <- data.frame(fit$scores)
```

Table 9: Results of Clustering Procedure (optimal k using hclust)



```
> second_der
[1]  5 28 10 30
```

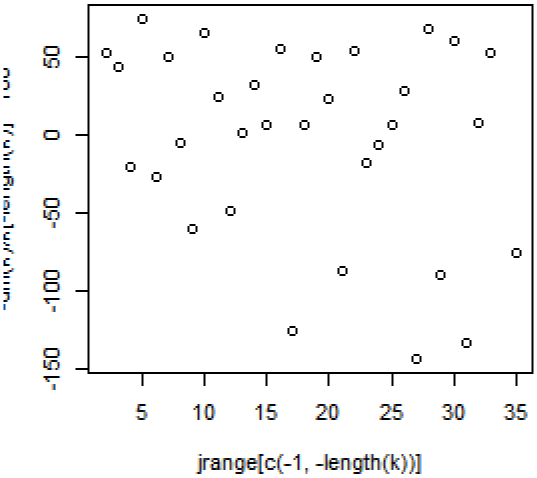


Figure 5: Clustering Analysis Plot (hclust)

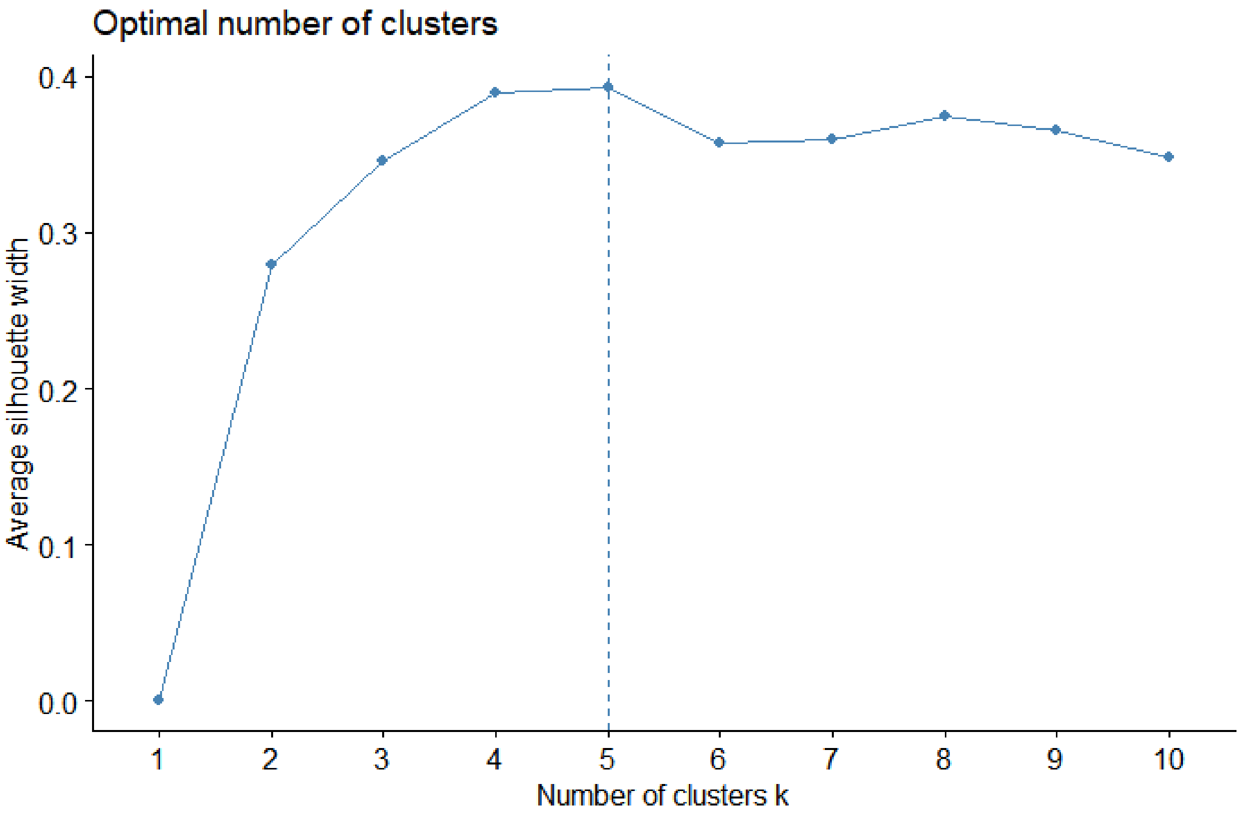
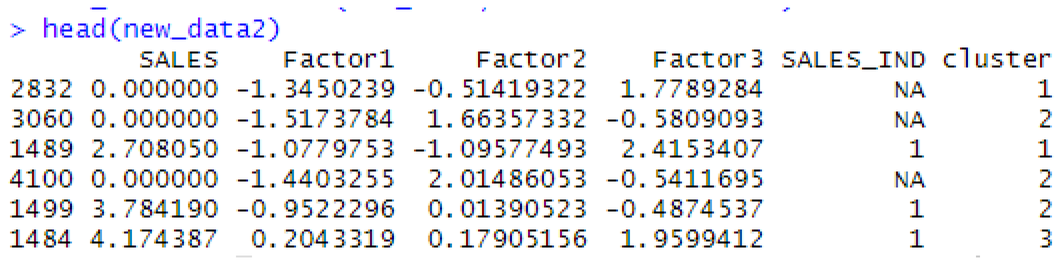


Figure 6: Silhouette statistic - Optimal Number of Clusters

Table 10: Preview of Factor Scores for Optimal Cluster of 5



```
> head(new_data2)
        SALES    Factor1     Factor2    Factor3 SALES_IND cluster
2832 0.000000 -1.3450239 -0.51419322  1.7789284        NA       1
3060 0.000000 -1.5173784  1.66357332 -0.5809093        NA       2
1489 2.708050 -1.0779753 -1.09577493  2.4153407         1       1
4100 0.000000 -1.4403255  2.01486053 -0.5411695        NA       2
1499 3.784190 -0.9522296  0.01390523 -0.4874537         1       2
1484 4.174387  0.2043319  0.17905156  1.9599412         1       3
```

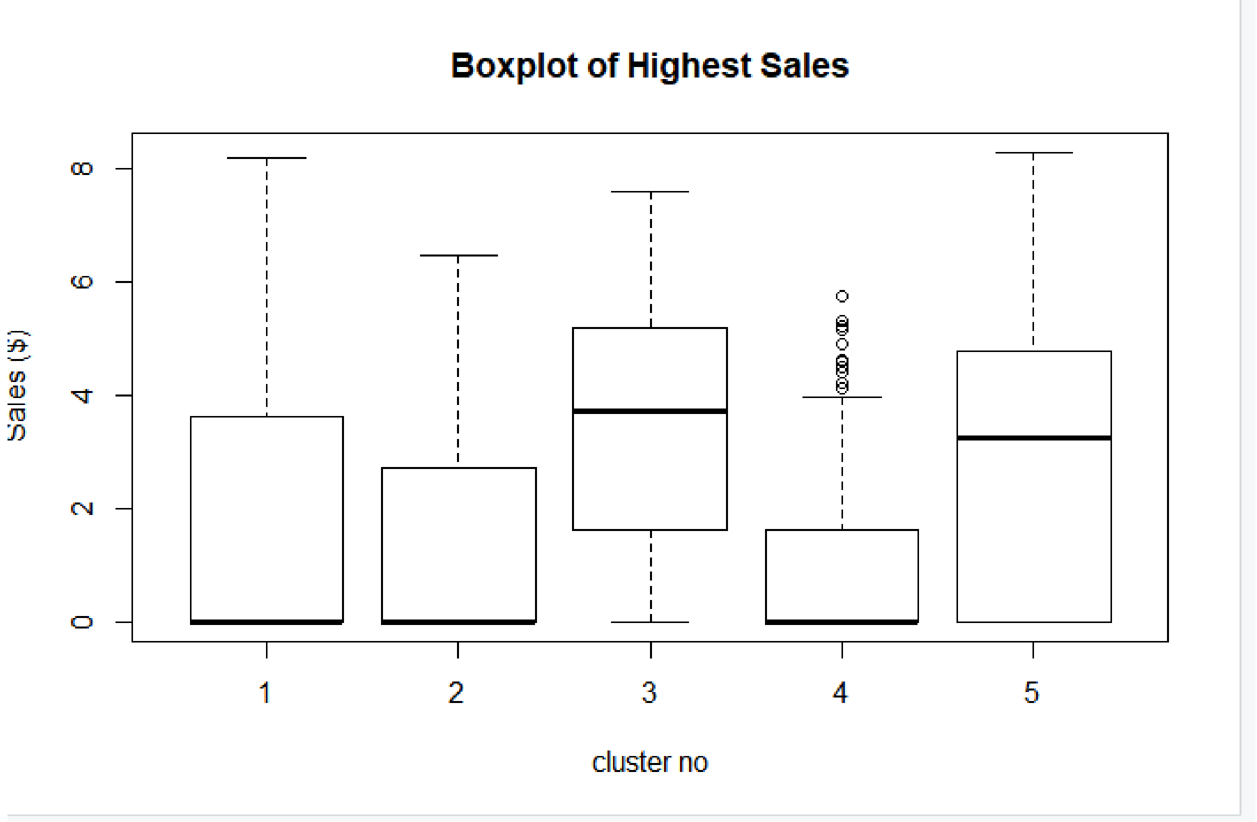
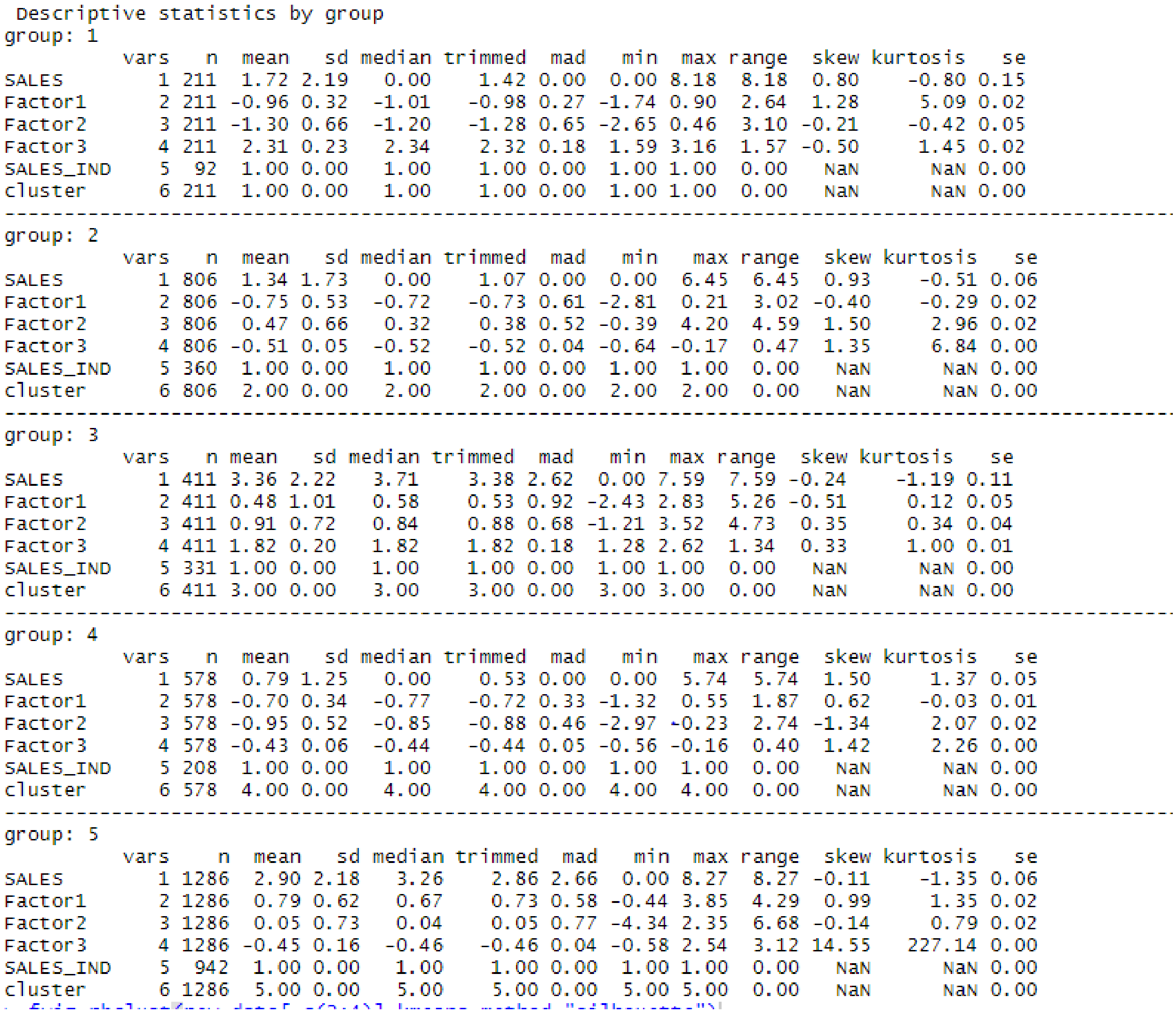


Figure 7: Boxplot of Highest Sales by Cluster



```
 Descriptive statistics by group 
group: 1
          vars   n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
SALES        1 211  1.72 2.19   0.00    1.42 0.00  0.00 8.18  8.18  0.80    -0.80 0.15
Factor1      2 211 -0.96 0.32  -1.01   -0.98 0.27 -1.74 0.90  2.64  1.28     5.09 0.02
Factor2      3 211 -1.30 0.66  -1.20   -1.28 0.65 -2.65 0.46  3.10 -0.21    -0.42 0.05
Factor3      4 211  2.31 0.23   2.34    2.32 0.18  1.59 3.16  1.57 -0.50     1.45 0.02
SALES_IND    5  92  1.00 0.00   1.00    1.00 0.00  1.00 1.00  0.00   NaN      NaN 0.00
cluster      6 211  1.00 0.00   1.00    1.00 0.00  1.00 1.00  0.00   NaN      NaN 0.00
-----------------------------------------------------------------------------------------------
group: 2
          vars   n  mean   sd median trimmed  mad   min   max range  skew kurtosis   se
SALES        1 806  1.34 1.73   0.00    1.07 0.00  0.00  6.45  6.45  0.93    -0.51 0.06
Factor1      2 806 -0.75 0.53  -0.72   -0.73 0.61 -2.81  0.21  3.02 -0.40    -0.29 0.02
Factor2      3 806  0.47 0.66   0.32    0.38 0.52 -0.39  4.20  4.59  1.50     2.96 0.02
Factor3      4 806 -0.51 0.05  -0.52   -0.52 0.04 -0.64 -0.17  0.47  1.35     6.84 0.00
SALES_IND    5 360  1.00 0.00   1.00    1.00 0.00  1.00  1.00  0.00   NaN      NaN 0.00
cluster      6 806  2.00 0.00   2.00    2.00 0.00  2.00  2.00  0.00   NaN      NaN 0.00
-----------------------------------------------------------------------------------------------
group: 3
          vars   n mean   sd median trimmed  mad   min  max range  skew kurtosis   se
SALES        1 411 3.36 2.22   3.71    3.38 2.62  0.00 7.59  7.59 -0.24    -1.19 0.11
Factor1      2 411 0.48 1.01   0.58    0.53 0.92 -2.43 2.83  5.26 -0.51     0.12 0.05
Factor2      3 411 0.91 0.72   0.84    0.88 0.68 -1.21 3.52  4.73  0.35     0.34 0.04
Factor3      4 411 1.82 0.20   1.82    1.82 0.18  1.28 2.62  1.34  0.33     1.00 0.01
SALES_IND    5 331 1.00 0.00   1.00    1.00 0.00  1.00 1.00  0.00   NaN      NaN 0.00
cluster      6 411 3.00 0.00   3.00    3.00 0.00  3.00 3.00  0.00   NaN      NaN 0.00
-----------------------------------------------------------------------------------------------
group: 4
          vars   n  mean   sd median trimmed  mad   min   max range  skew kurtosis   se
SALES        1 578  0.79 1.25   0.00    0.53 0.00  0.00  5.74  5.74  1.50     1.37 0.05
Factor1      2 578 -0.70 0.34  -0.77   -0.72 0.33 -1.32  0.55  1.87  0.62    -0.03 0.01
Factor2      3 578 -0.95 0.52  -0.85   -0.88 0.46 -2.97 -0.23  2.74 -1.34     2.07 0.02
Factor3      4 578 -0.43 0.06  -0.44   -0.44 0.05 -0.56 -0.16  0.40  1.42     2.26 0.00
SALES_IND    5 208  1.00 0.00   1.00    1.00 0.00  1.00  1.00  0.00   NaN      NaN 0.00
cluster      6 578  4.00 0.00   4.00    4.00 0.00  4.00  4.00  0.00   NaN      NaN 0.00
-----------------------------------------------------------------------------------------------
group: 5
          vars    n  mean   sd median trimmed  mad   min  max range  skew kurtosis   se
SALES        1 1286  2.90 2.18   3.26    2.86 2.66  0.00 8.27  8.27 -0.11    -1.35 0.06
Factor1      2 1286  0.79 0.62   0.67    0.73 0.58 -0.44 3.85  4.29  0.99     1.35 0.02
Factor2      3 1286  0.05 0.73   0.04    0.05 0.77 -4.34 2.35  6.68 -0.14     0.79 0.02
Factor3      4 1286 -0.45 0.16  -0.46   -0.46 0.04 -0.58 2.54  3.12 14.55   227.14 0.00
SALES_IND    5  942  1.00 0.00   1.00    1.00 0.00  1.00 1.00  0.00   NaN      NaN 0.00
cluster      6 1286  5.00 0.00   5.00    5.00 0.00  5.00 5.00  0.00   NaN      NaN 0.00
```

Table 11: Summary of Clustering Procedure

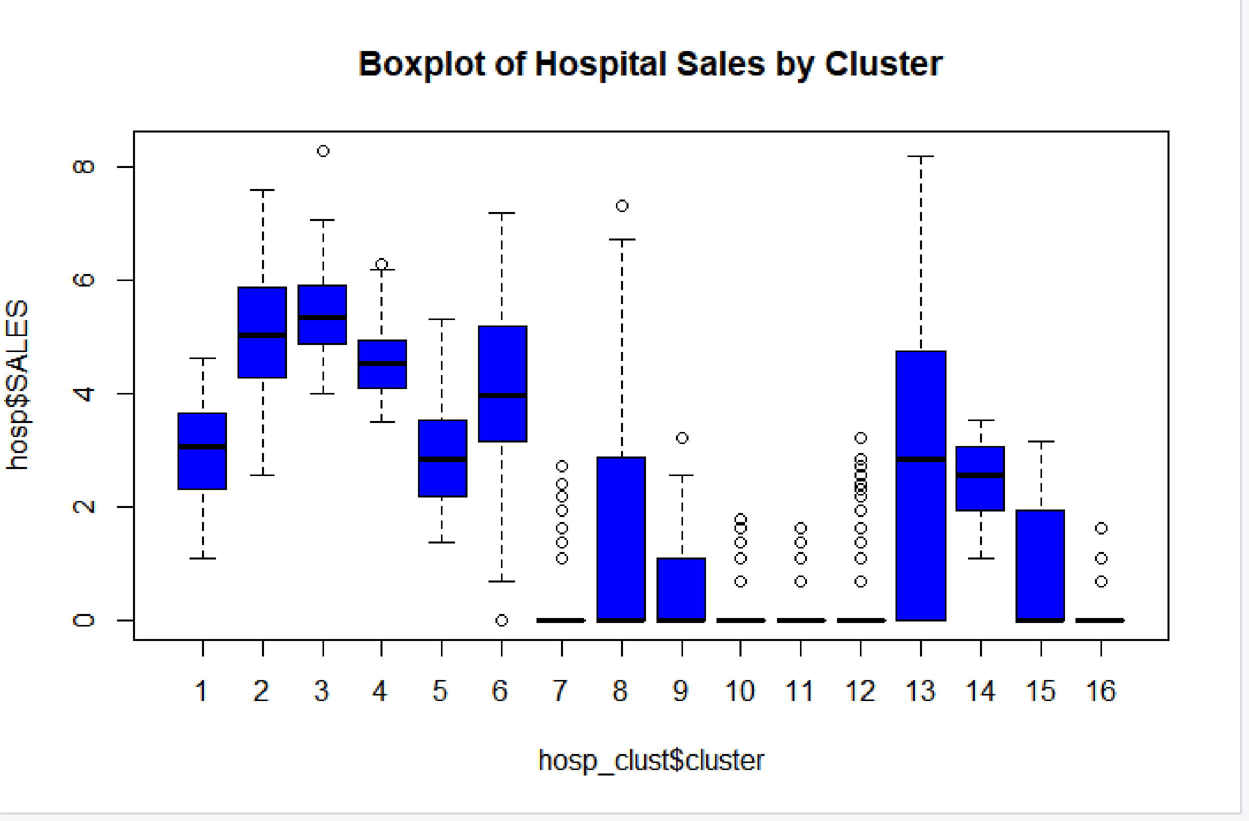
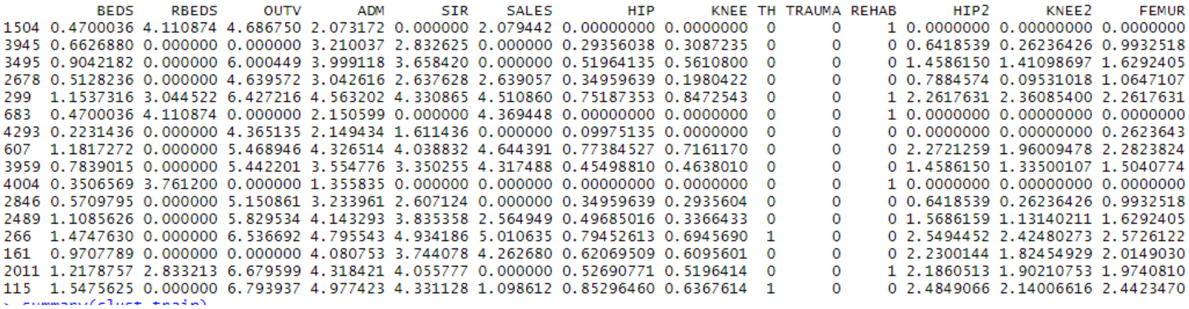


Figure 8: Boxplot of Hospital Sales By Cluster

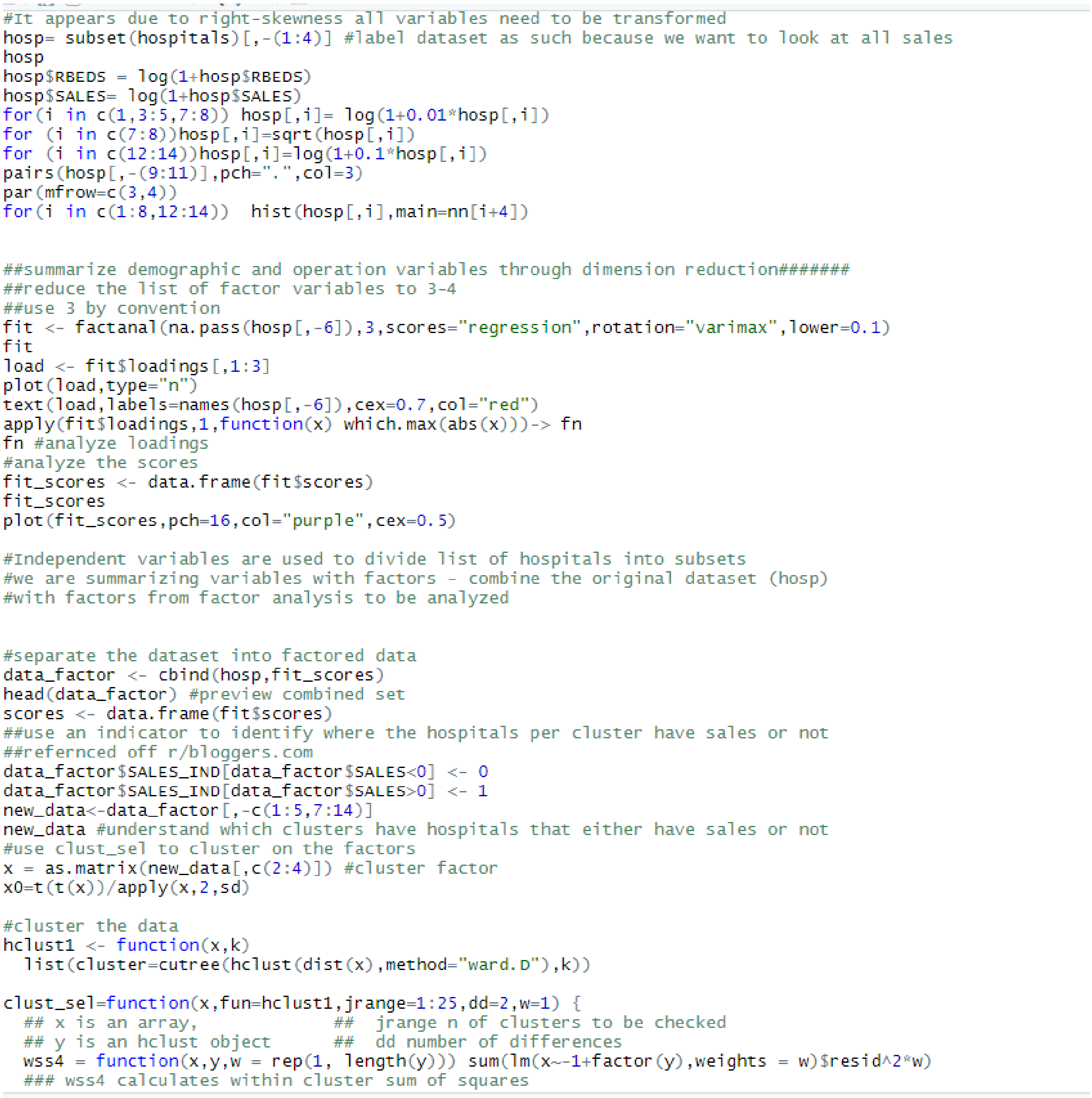


```
            BEDS    RBEDS     OUTV      ADM      SIR    SALES        HIP      KNEE TH TRAUMA REHAB      HIP2      KNEE2     FEMUR
1504 0.4700036 4.110874 4.686750 2.073172 0.000000 2.079442 0.00000000 0.0000000  0      0     1 0.0000000 0.00000000 0.0000000
3945 0.6626880 0.000000 0.000000 3.210037 2.832625 0.000000 0.29356038 0.3087235  0      0     0 0.6418539 0.26236426 0.9932518
3495 0.9042182 0.000000 6.000449 3.999118 3.658420 0.000000 0.51964135 0.5610800  0      0     0 1.4586150 1.41098697 1.6292405
2678 0.5128236 0.000000 4.639572 3.042616 2.637628 2.639057 0.34959639 0.1980422  0      0     0 0.7884574 0.09531018 1.0647107
299  1.1537316 3.044522 6.427216 4.563202 4.330865 4.510860 0.75187353 0.8472543  0      0     1 2.2617631 2.36085400 2.2617631
683  0.4700036 4.110874 0.000000 2.150599 0.000000 4.369448 0.00000000 0.0000000  0      0     1 0.0000000 0.00000000 0.0000000
4293 0.2231436 0.000000 4.365135 2.149434 1.611436 0.000000 0.09975135 0.0000000  0      0     0 0.0000000 0.00000000 0.2623643
607  1.1817272 0.000000 5.468946 4.326514 4.038832 4.644391 0.77384527 0.7161170  0      0     0 2.2721259 1.96009478 2.2823824
3959 0.7839015 0.000000 5.442201 3.554776 3.350255 4.317488 0.45498810 0.4638010  0      0     0 1.4586150 1.33500107 1.5040774
4004 0.3506569 3.761200 0.000000 1.355835 0.000000 0.000000 0.00000000 0.0000000  0      0     1 0.0000000 0.00000000 0.0000000
2846 0.5709795 0.000000 5.150861 3.233961 2.607124 0.000000 0.34959639 0.2935604  0      0     0 0.6418539 0.26236426 0.9932518
2489 1.1085626 0.000000 5.829534 4.143293 3.835358 2.564949 0.49685016 0.3366433  0      0     0 1.5686159 1.13140211 1.6292405
266  1.4747630 0.000000 6.536692 4.795543 4.934186 5.010635 0.79452613 0.6945690  1      0     0 2.5494452 2.42480273 2.5726122
161  0.9707789 0.000000 0.000000 4.080753 3.744078 4.262680 0.62069509 0.6095601  0      0     0 2.2300144 1.82454929 2.0149030
2011 1.2178757 2.833213 6.679599 4.318421 4.055777 0.000000 0.52690771 0.5196414  0      0     1 2.1860513 1.90210753 1.9740810
115  1.5475625 0.000000 6.793937 4.977423 4.331128 1.098612 0.85296460 0.6367614  1      0     0 2.4849066 2.14006616 2.4423470
> summary(clust_train)
```

Table 12: K-Medoids Analysis

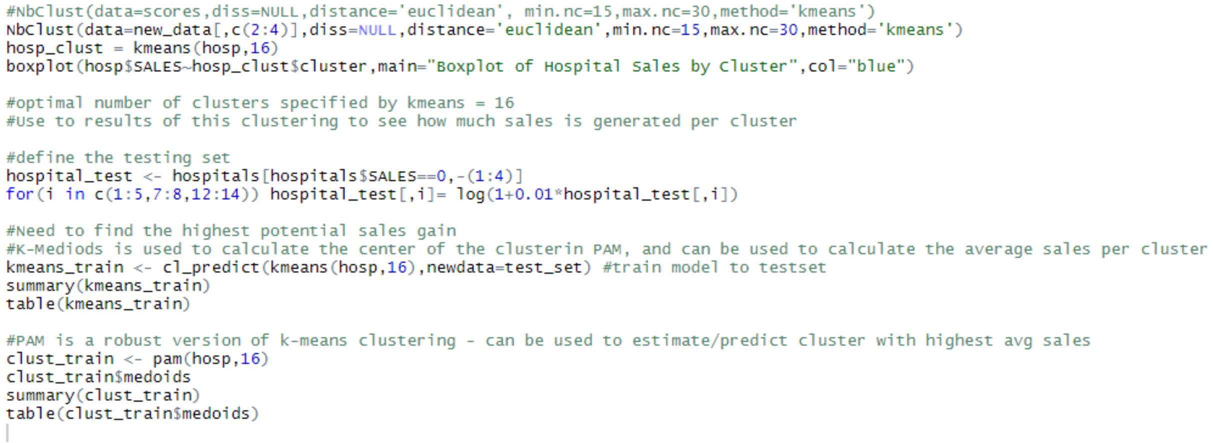
Code

```r
#############R Code for Final Project #########################################
#objective - find ways to increase sales of orthopedic products to all hospitals in US
#find those who have highest consumption of such equipment, but where our sales are = 0
#Come up with a selected group where you think our efforts will be rewarded. (a few hospitals 5 or 10 or 15).
#Estimate the potential or expected sales on those hospitals.
##############################################################################

#Download necessary libraries
library(caret)
library(MASS)
library(tidyverse)
library(tree)
library(nnet)
library(rpart)
library(naniar)
library(pysch)
library(clue)
library(cluster)
library(factoextra)
library(NbClust)
library(psych)

#Read in data + perform simple data analysis
hospitals1 <- read.csv(file='C:/Users/aatkuru/Desktop/hospitalUSA.csv')
hospitals1
summary(hospitals1)
nn = names(hospitals1)
print(nn)
nrow(hospitals1) #4703 rows - need to cutdown between 3000-3500
#Sales Data
sales <- hospitals1$SALES[hospitals1$SALES>0] #sales greater than 0
summary(sales)
#subset the data to get 3,292 observations (between 3000-3500)
set.seed(04248) #last 4 digits of RUID
n <- nrow(hospitals1)
n
subset = sample(n,0.7*n,replace=FALSE)
hospitals = hospitals1[subset,]

#separate response variable sales - set all 0 sales to NA
hospitals$SALES[hospitals$Sales==0]= NA

#separate demographics variables - BEDS, RBEDS, OUTV, ADM, SIR, TH, TRAUMA, REHAB
demographics <- hospitals[,c(5:9,13:15)]
summary(demographics)

#separate operation numbers - HIP, KNEE, HIP2, KNEE2, FEMUR2
operationnum <- hospitals[,c(11:12,16:18)]
summary(operationnum)

#analyze the variables to determine the required transformations
pairs(hospitals[,-c(1:4,13:15)],pch=".",col=2)
summary(hospitals)
par(mfrow=c(3,4))
for(i in c(5:12,16:18))  hist(hospitals[,i],main=nn[i])
```



```
#It appears due to right-skewness all variables need to be transformed
hosp= subset(hospitals)[,-(1:4)] #label dataset as such because we want to look at all sales
hosp
hosp$RBEDS = log(1+hosp$RBEDS)
hosp$SALES= log(1+hosp$SALES)
for(i in c(1,3:5,7:8)) hosp[,i]= log(1+0.01*hosp[,i])
for (i in c(7:8))hosp[,i]=sqrt(hosp[,i])
for (i in c(12:14))hosp[,i]=log(1+0.1*hosp[,i])
pairs(hosp[,-(9:11)],pch=".",col=3)
par(mfrow=c(3,4))
for(i in c(1:8,12:14))  hist(hosp[,i],main=nn[i+4])


##summarize demographic and operation variables through dimension reduction#######
##reduce the list of factor variables to 3-4
##use 3 by convention
fit <- factanal(na.pass(hosp[,-6]),3,scores="regression",rotation="varimax",lower=0.1)
fit
load <- fit$loadings[,1:3]
plot(load,type="n")
text(load,labels=names(hosp[,-6]),cex=0.7,col="red")
apply(fit$loadings,1,function(x) which.max(abs(x)))-> fn
fn #analyze loadings
#analyze the scores
fit_scores <- data.frame(fit$scores)
fit_scores
plot(fit_scores,pch=16,col="purple",cex=0.5)

#Independent variables are used to divide list of hospitals into subsets
#we are summarizing variables with factors - combine the original dataset (hosp)
#with factors from factor analysis to be analyzed


#separate the dataset into factored data
data_factor <- cbind(hosp,fit_scores)
head(data_factor) #preview combined set
scores <- data.frame(fit$scores)
##use an indicator to identify where the hospitals per cluster have sales or not
##refernced off r/bloggers.com
data_factor$SALES_IND[data_factor$SALES<0] <- 0
data_factor$SALES_IND[data_factor$SALES>0] <- 1
new_data<-data_factor[,-c(1:5,7:14)]
new_data #understand which clusters have hospitals that either have sales or not
#use clust_sel to cluster on the factors
x = as.matrix(new_data[,c(2:4)]) #cluster factor
x0=t(t(x))/apply(x,2,sd)

#cluster the data
hclust1 <- function(x,k)
  list(cluster=cutree(hclust(dist(x),method="ward.D"),k))

clust_sel=function(x,fun=hclust1,jrange=1:25,dd=2,w=1) {
  ## x is an array,             ##  jrange n of clusters to be checked
  ## y is an hclust object      ##  dd number of differences
  wss4 = function(x,y,w = rep(1, length(y))) sum(lm(x~-1+factor(y),weights = w)$resid^2*w)
  ### wss4 calculates within cluster sum of squares
```

```
clust_sel=function(x,fun=hclust1,jrange=1:25,dd=2,w=1) {
  ## x is an array,              ##  jrange n of clusters to be checked
  ## y is an hclust object       ##  dd number of differences
  wss4 = function(x,y,w = rep(1, length(y))) sum(lm(x~-1+factor(y),weights = w)$resid^2*w)
  ### wss4 calculates within cluster sum of squares
  sm1 = NULL
  for(i in jrange) sm1[i] = if(i==1) sum(lm(x~1)$resid^2) else wss4(x,fun(x,i)$cluster)
  sm1=sm1[jrange]
  k = if(dd==1) sm1[-1] else -diff(sm1)
  plot(jrange[c(-1,-length(k))], -diff(k)/k[-length(k)]*100)
  jrange[c(-1,-length(k))] [sort.list(diff(k)/k[-length(k)]*100)[1:4]]
}#use the second derivative method
second_der <- clust_sel(x0,hclust1,jrange=1:35)
second_der

#can verify 5 clusters with another method
fviz_nbclust(new_data[,c(2:4)],kmeans,method="silhouette")
#also got k=5

#once the clusters are chosen, we must specify summary statistics for e/a cluster
#you can use a boxplot of sales or transformed sales vs. your cluster number
#choose number of clusters with highest sales
reclust<- cutree(hclust(dist(x),method="ward.D"),5)
reclust #cluster now with specified optimal number of clusters
new_data2 <- cbind(new_data, cluster = reclust)
head(new_data2)
#use a boxplot to find highest sales
boxplot(new_data2$SALES~reclust, main = "Boxplot of Highest Sales",xlab='cluster no', ylab='Sales ($)')
abline(h=mean(x),col="red") #used to get an understanding of avg sales
z = mean(x)
z #sales for the highest cluster (per 1000's)


#cluster with highest sales is = 3
#still it is good to do further analysis to see which cluster agrees with our objectives
#but where there are hospitals were the company's sales is NA so they are not yet our customers
describeBy(new_data2,reclust)
#group 1 has a smaller variance/sd between the different features, so therefore it
#overall this cluster meets our objectives


#estimate potential gain in sales
#average sale to similar hospital - average sale to that cluster
#interested in hospitals where current sales = NA
#Find optimal number of clusters - redo clustering process in case N>100
#can use the NbClust function with kmeans to find optimal number of clusters
n=nrow(hospitals)
ni=sample(n,3292,rep=F)
hospital_train <- hospitals[ni,]
hospital_test <- hospitals[-ni,]
test_set <- hospital_test[,-6] #remove indep variable


#NbClust(data=scores,diss=NULL,distance='euclidean', min.nc=15,max.nc=30,method='kmeans')
NbClust(data=new_data[,c(2:4)],diss=NULL,distance='euclidean',min.nc=15,max.nc=30,method='kmeans')
hosp_clust = kmeans(hosp,16)
```



```
#NbClust(data=scores,diss=NULL,distance='euclidean', min.nc=15,max.nc=30,method='kmeans')
NbClust(data=new_data[,c(2:4)],diss=NULL,distance='euclidean',min.nc=15,max.nc=30,method='kmeans')
hosp_clust = kmeans(hosp,16)
boxplot(hosp$SALES~hosp_clust$cluster,main="Boxplot of Hospital Sales by Cluster",col="blue")

#optimal number of clusters specified by kmeans = 16
#Use to results of this clustering to see how much sales is generated per cluster

#define the testing set
hospital_test <- hospitals[hospitals$SALES==0,-(1:4)]
for(i in c(1:5,7:8,12:14)) hospital_test[,i]= log(1+0.01*hospital_test[,i])

#Need to find the highest potential sales gain
#K-Mediods is used to calculate the center of the clusterin PAM, and can be used to calculate the average sales per cluster
kmeans_train <- cl_predict(kmeans(hosp,16),newdata=test_set) #train model to testset
summary(kmeans_train)
table(kmeans_train)

#PAM is a robust version of k-means clustering - can be used to estimate/predict cluster with highest avg sales
clust_train <- pam(hosp,16)
clust_train$medoids
summary(clust_train)
table(clust_train$medoids)
```