**Assignment 3: Orthopedic Materials Sales (100 points)**

**Student Name:**

**Purpose:** To perform cluster analysis to identify potential business for orthopedic material sales

**Description:** The objective of this study is to find ways to increase sales of orthopedic material from our company to hospitals in the United States. The data include information about over 4000 hospitals. Below is the data dictionary:

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 (In $1000's per year)

SIR: REVENUE FROM INPATIENT

SALESY: SALES OF REHABILITATION EQUIPMENT SINCE JAN 1

SALES12: SALES OF REHAB. EQUIP. FOR THE LAST 12 MO

HIP: NUMBER OF HIP OPERATIONS FOR TWO YEARS AGO

KNEE: NUMBER OF KNEE OPERATIONS FOR TWO YEARS AGO

TH: TEACHING HOSPITAL? 0, 1

TRAUMA: DO THEY HAVE A TRAUMA UNIT? 0, 1

REHAB: DO THEY HAVE A REHAB UNIT? 0, 1

HIP12: NUMBER HIP OPERATIONS FOR THE LAST 12 MO

KNEE12: NUMBER KNEE OPERATIONS FOR THE LAST 12 MO

FEMUR12: NUMBER FEMUR OPERATIONS FOR THE LAST 12 MO

**Instructions:** Please follow these steps:

1. In Canvas, navigate to Assignments and then Assignment4
2. Download and save the data set hospital\_ortho.csv
3. Read the file:

data <- fread("hospital\_ortho.csv", sep=",", header=T, strip.white = T, na.strings = c("NA","NaN","","?"))

1. The original data includes hospitals across the US. However, we can only sell our products in NC and the nearby states of SC, VA, GA, and TN. Use the following code to narrow down the data to hospitals in these states.

nc\_data <- data[(data$state == "NC") | (data$state == "SC") | (data$state == "VA") | (data$state == "GA") | (data$state == "TN")]

* 1. **(3 points)** Look at each individual variable and decide if it should be included in cluster analysis. For those variables that you decide not to include, give your reasons for exclusion.
  2. **(3 points)** Do you need to scale this data? Why?

1. Perform k-means clustering:
   1. **(3 points)** Use “Within Groups SSE” to determine the number of clusters for k-means. How many clusters you would like to create?
   2. **(3 points)** Paste the “Within Groups SSE” plot in the space below:
   3. **(3 points)** Perform k-means clustering using the number of clusters you recommended in 5.1. How many hospitals fall in each cluster?
   4. **(3 points)** Create a two-dimensional representation of the clusters and paste it below:
2. Perform Hierarchical clustering.
   1. **(4 points)** Try different hierarchical clustering and paste the dendrograms in the space below:
   2. **(3 points)** Determine which hierarchical clustering method would be more appropriate for this data. Why?
   3. **(3 points)** Based on hierarchical clustering results, how many clusters do you find in this data?
   4. **(3 points)** Paste the dendrogram that you chose with the red borders around the clusters in the space below:
3. Perform DBSCAN cluster analysis:
   1. **(7 points)** First, you need to determine minPts. The rule of thumb for minPts is the number of dimensions of the data + 1. Suggest a method to determine the number of dimensions of this data? Implement your method and suggest a good minPts.
   2. **(3 points)** Based on your suggested minPts, determine the eps. Explain your recommendation for eps.
   3. **(3 points)** Perform DBSCAN clustering using the minPts and eps that you recommended. How many clusters DBSCAN returns?
   4. **(3 points)** How many noise points it returns?
   5. **(3 points)** Create a two-dimensional representation of DBSCAN cluster(s) and paste it in the space below:
4. Perform principal component analysis on the original data (nc\_data). Then select the number of principal components based on PCs variance plot. Let’s call the number of PCs n\_pc. Then we can use the best PCs instead of the data to perform cluster analysis. To do this, run:

pca\_data <- predict(pca, newdata = nc\_data)

pc\_df <- as.data.frame(scale(pca\_data[,c(1:n\_pc)])) # replace n\_pc with the number of PCs you recommend.

* 1. **(10 points)** Repeat your analysis in question 5 using the new pc\_df. What is the best k? Paste the two-dimensional representation in the space below:
  2. **(10 points)** Repeat your analysis in question 6 using the new pc\_df. What is the best method? What is the best k? Paste the dendrogram in the space below:
  3. **(10 points)** Repeat your analysis in question 7 using the new pc\_df. What is the best minPts? What is the best eps? How many clusters DBSCAN returns? Perform the DBSCAN clustering and paste the two-dimensional representation in the space below:

1. For each hospital, determine the cluster (based on pc\_df) to which they belong. Then determine the value of "sales12","rbeds","hip12","knee12", and "femur12" for each cluster for each clustering method (e.g. k-means, hierarchical, DBSCAN). To do this, you need to run the following lines:

pc\_df$kmeans <- k.means.fit$cluster

pc\_df $hclust <- groups # these groups are created in hierarchical clustering

pc\_df $db <- db$cluster

pc\_df $hid <- nc\_data$hid # Add hospital id to pc\_df data

final\_data <- merge(x=pc\_df, y=nc\_data, key="hid")

aggregate(final\_data[,c("sales12","rbeds","hip12","knee12","femur12")], list(final\_data$kmeans), mean)

aggregate(final\_data[,c("sales12","rbeds","hip12","knee12","femur12")], list(final\_data$hclust), mean)

aggregate(final\_data[,c("sales12","rbeds","hip12","knee12","femur12")], list(final\_data$db), mean)

* 1. **(20 points)** Based on these results for each clustering method (e.g. k-means, hierartchical, and DBSCAN), recommend which cluster we should immediately reach out to. Give your reasons.