Insurance Claim Data and Analytics

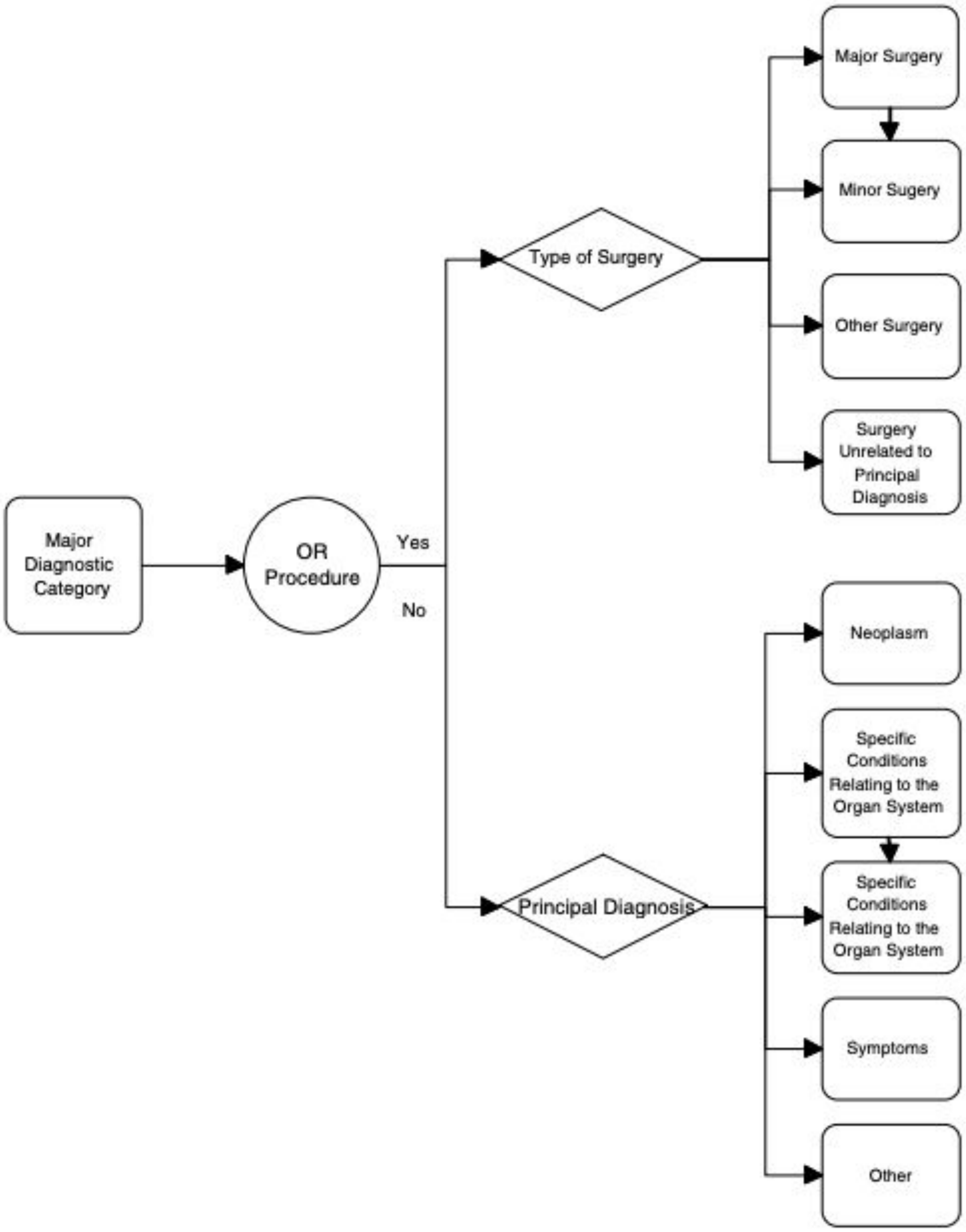
Team 1 - Zhiqi Chen, Nkosingiphile Shongwe, Salil Redkar,
Debarati Mazumdar, Lakshmi Malavika Andavilli

Introduction

We know that charges for hospital admissions with the same principal procedure cannot be expected to be identical, not only because different hospitals assign different costs, but also because different patients have different severities in their conditions. As stated in the lecture, in the past, hospital characteristics such as teaching status, and patients treated were more sick, have been used to explain the cost differences across hospitals. However, these explanations failed to adequately justify the cost impact of a hospital's case mix. Even though it makes sense that a complex case would result in higher costs, there has never been a precise definition. Case mix complexity refers to an interrelated and distinct set of patient attributes, which describes a particular aspect of a hospital's case mix in the following manner: severity of illness, prognosis, treatment difficulty, need for intervention and resource intensity (www.cms.gov).

In the late sixties, the Social Security Act created a system of payment for the operating costs associated with Medicare Part A hospital inpatient stays (https://www.ehealthmedicare.com and www.cms.gov), which is based on set rates and is referred to as the Inpatient Prospective Payment System (IPPS). In this system, each case is categorized into a Medicare Severity - Diagnosis Related Group (MS-DRG). Patients who have similar clinical characteristics and similar treatment costs are assigned to the same MS-DRG. The MS-DRG is a fixed payment amount based on the average treatment cost of patients in a certain group. Patients are assigned to a DRG based on their diagnosis, surgical procedures, age, and other information. Hospitals provide this information on their Medicare claim, and Medicare uses this information to decide how much the hospitals should be paid.

The development of the DRGs provided the first operational means of defining and measuring a hospital's case mix complexity. With the evolution of the DRGs and their use as the basic unit of payment represents a recognition of the fundamental role which a hospital's case mix plays in determining its costs. Please see example of structure of DRG (www.ncbi.nlm.nih.gov):


Major Diagnostic Category
OR Procedure
Yes
No
Type of Surgery
Major Surgery
Minor Sugery
Other Surgery
Surgery Unrelated to Principal Diagnosis
Principal Diagnosis
Neoplasm
Specific Conditions Relating to the Organ System
Specific Conditions Relating to the Organ System
Symptoms
Other

In this report we applied clustering to understand how costs are grouped together in relation to their DRGs. We want to determine if there are any similarities in patient admissions in the state of Vermont. Since clustering groups categories in accordance with their similarities, we expect to find DGRs in the same group to be grouped together in terms of how much they cost.

DRG Clustering

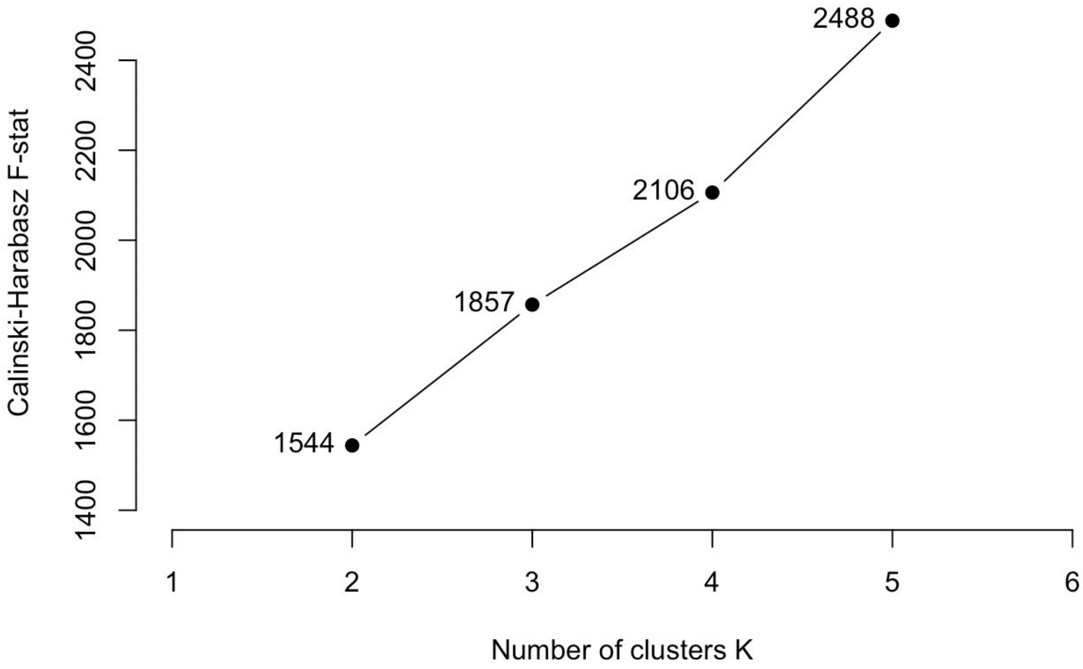
The objective was to use clustering to derive insights from the total operating room and anesthesiology costs of hospital visits for specific DRGs. The hypothesis was that using algorithmic division of the DRGs into clusters would reveal underlying patterns in the cost grouping of the DRGs. Clustering is a well-known technique in unsupervised learning. In this analysis, we will be using k-means clustering. K-means clustering algorithm is an unsupervised hard clustering method, which assigns the number of data objects to a predefined number of exactly k clusters.

Table 1

Number of clusters	**F-stat**
2	1544
3	1857
4	2106
5	2488

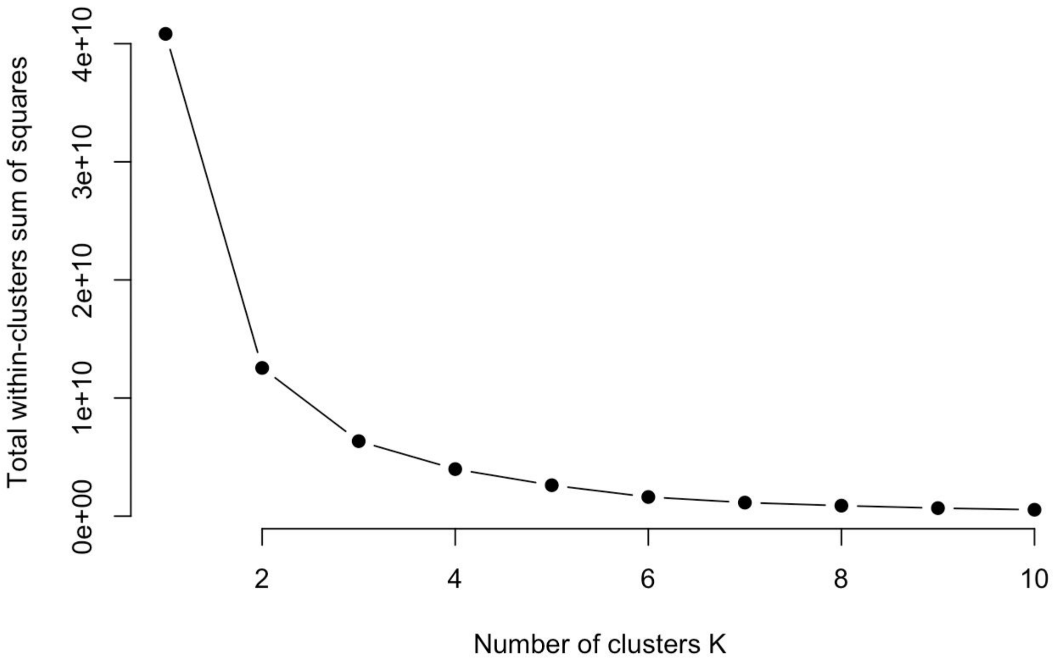
In Table 1, we observe an increasing F-stat value with increasing clusters, with the highest value at 5 clusters (2488). The scree-plot (Figure 2) shows that the steep rise at 3 clusters is indicative of it being the optimal number of clusters, and we decided to move ahead with the division of the data into 3 clusters.

Figure 1



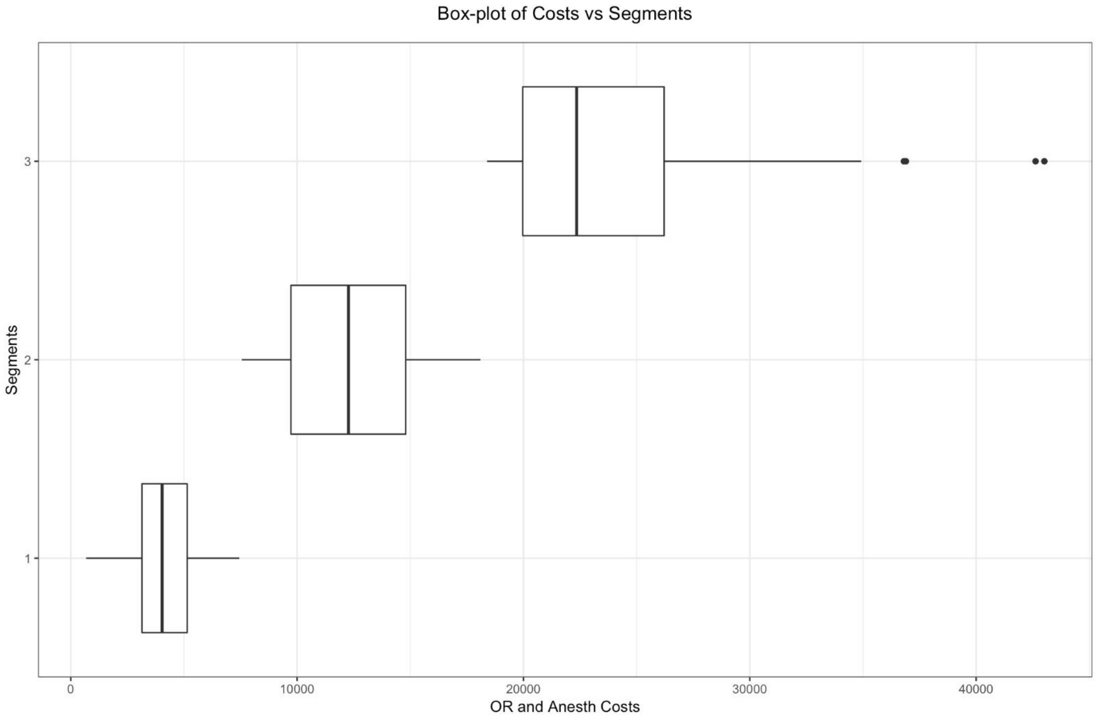
We used two methods to determine the optimal number of clusters for the given data; Calinski-Harabasz F-statistic and scree plot. The Calinski-Harabasz F-statistic (CH-stat) is a heuristic measure of determining the optimal number of clusters. CH-stat is a measure of the variance within the cluster against the variance across the different clusters. A high CH stat is indicative of “close-to-optimal” segmentation of the data.

Figure 2



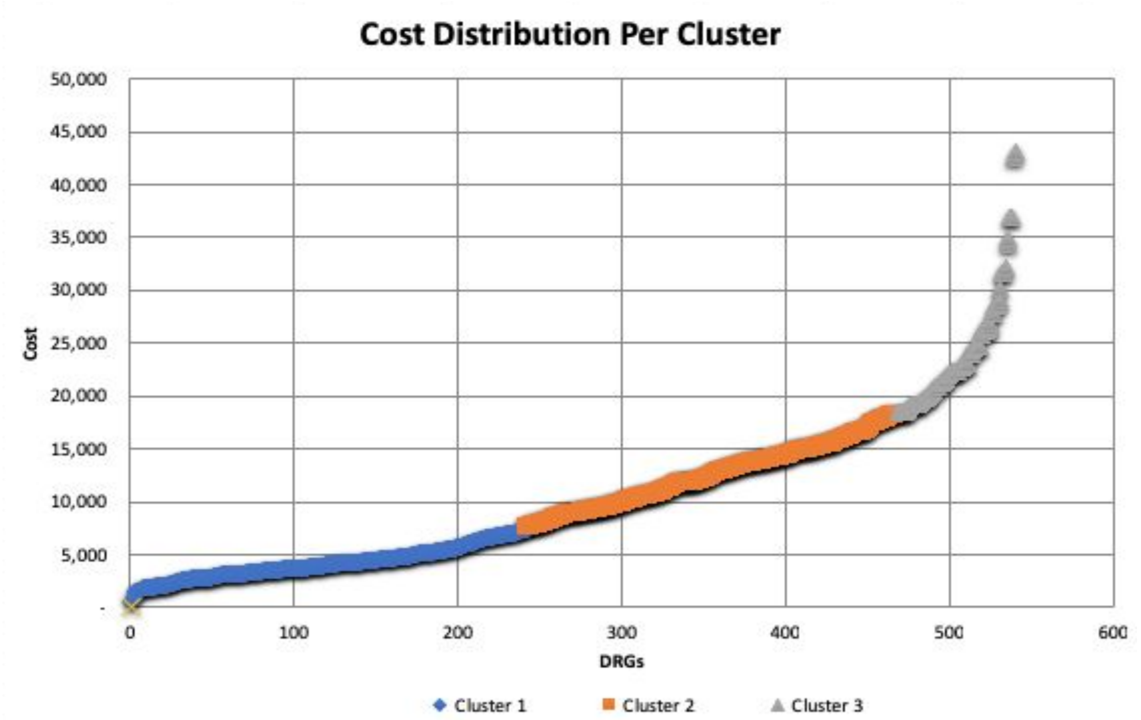
Another similar measure to determine the number of clusters is the scree plot. As the number of clusters increases, the variance (within-group sum of squares) decreases. A steep "elbow" at the optimal number of clusters represents the balance between minimizing the number of clusters and minimizing the variance within each cluster.

Figure 3



From the observation of the box-plot detailing the distribution of the costs of each of the segments, it was clear that the clusters showed a trend of increasing costs, with cluster 3 being the cluster with the highest costs. Interestingly, cluster 1, and 2 indicate a roughly symmetric distribution of the costs, while cluster 3 shows a right-sided skewness. The high costs of the these outliers skews the mean of the data in cluster 3. It may be worthwhile to determine the surgical procedures associated with these DRGs to identify diagnostic areas that cost the most.

Figure 4



The increasing costs associated with DRGs across the clusters becomes even more evident from Figure (4). The three colors represent the three distinct clusters. In accordance with the box-plot, clusters 1 and 2 show a gradual increase in costs. Cluster 1 shows a varied spread of cost from approximately $1000 to $8000. DRGs for this cluster include cardiac disorders, fractures and respiratory problems. Cluster 2 has a spread from about $8000 to $20,000. We predict that the higher costs associated with cluster 2 probably indicate more complex procedures compared to cluster 1. The sharp steep rise for cluster 3 indicates swiftly rising charges for complex procedures. This cluster has the highest costs and the spread ranges from $20,000 to $45,000. The outliers seen in the box-plot are also clearly visible in the "knee" portion of this plot. Thus, we can conclude that DRGs in Cluster 3 probably indicate complex and intricate surgical procedures that typically cost significantly higher than routine medical procedures. The higher cost may be an outcome of longer hours required for the surgery, requirement of more specialized doctors for the procedure and expensive equipments for the procedure. Traditionally, these procedures require the expertise and support from the anesthesiology department and the resources of the operations department. With an overall summary of the division of the DRGs, we decided to investigate which DRGs were representative of their cluster.

Percentage of Records with a Complication

DRG can be three different types for each description:

- Complications and Comorbidities (CC)
- Major Complications and Comorbidities (MCC)
- Without CC(Complications and Comorbidities)

Table 2

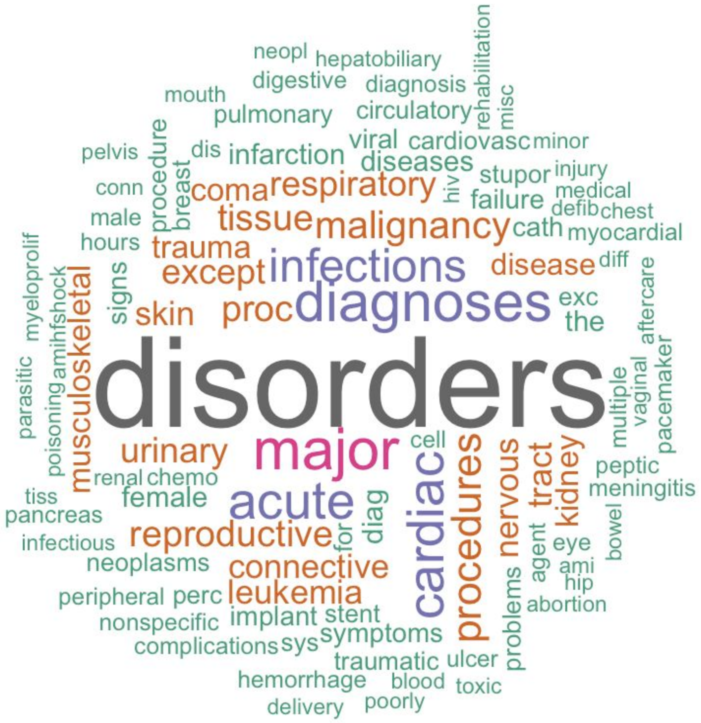
Cluster	Total Records	Total Records with CC/MCC	Percentage of records with CC or MCC
1	389	227	= (227/389) = 58.3%
2	225	147	= (147/225)= 65.3%
3	73	49	= (49/73) = 67.1%

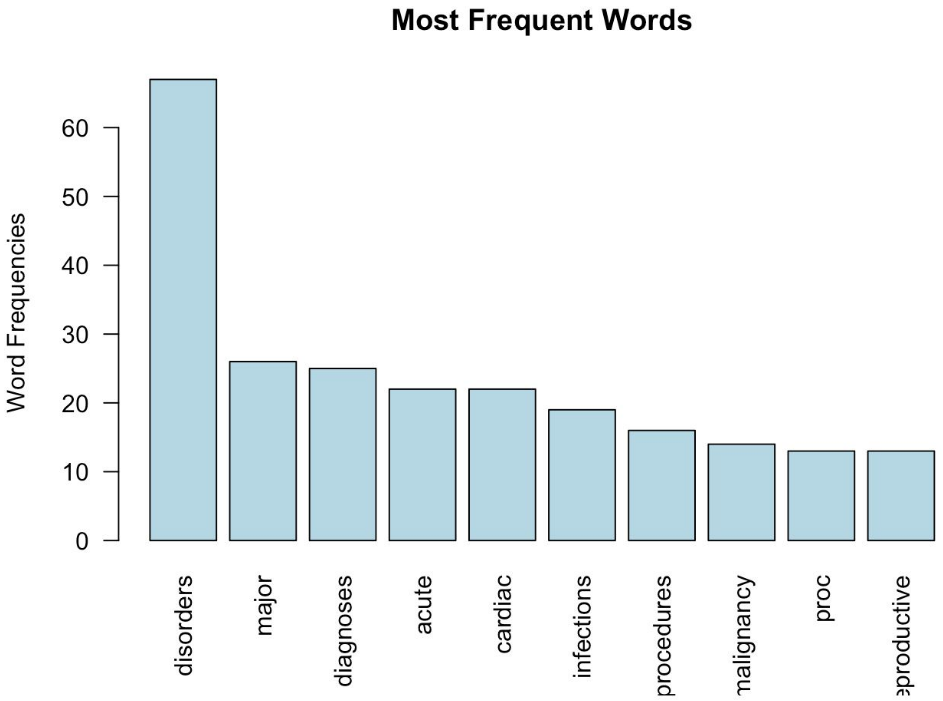
MCC and CC are coding related to Major (complications and comorbidities) for the procedures. If DRG is coded as MCC/CC those will have higher cost.

As per the table above we can see that a DRGs in cluster 1 have the least % hence this relates to the fact that cluster defines disorders which have fewer complications and hence cost is less. However cluster 2 and 3 have almost same % of DRGs with MCC/CC and these are mostly the ones with procedures which require more care and longer stay at hospitals hence cost is highers for these two clusters.

Frequent Words in Each Cluster

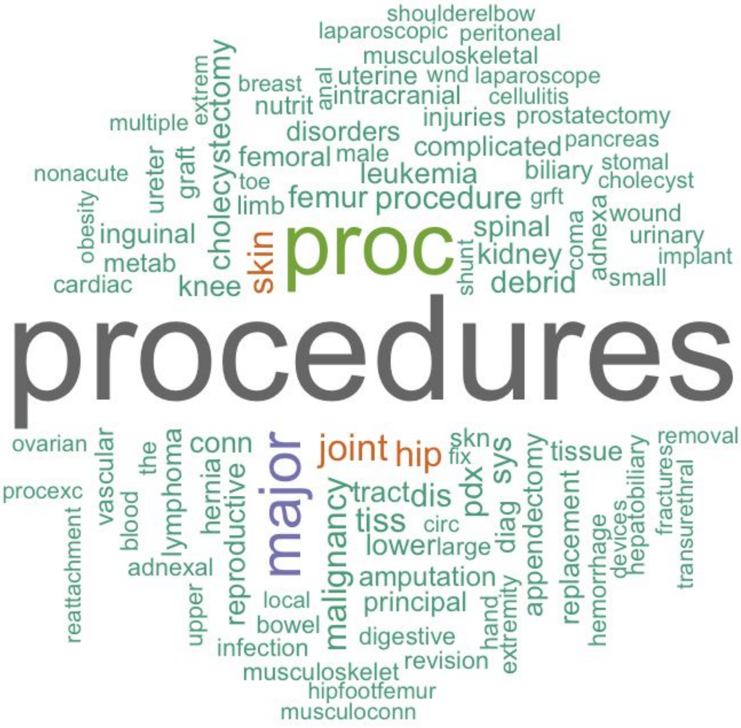
Cluster 1

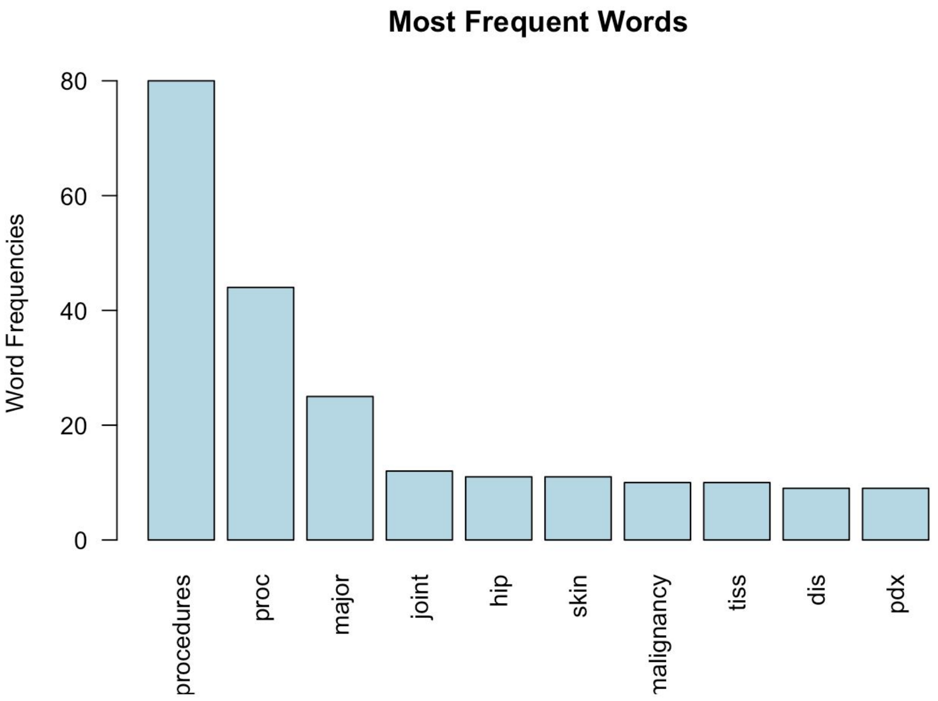




Mostly disorders, acute diseases and diagnosis (as shown in the cloud below) that need just medication, as opposed to any major procedures or surgeries thus a lower cost cluster. Disorders, Diagnosis, infections,malignancy suggest a very minimal requirement of operating room and anesthesiology. If we see the most frequent words graph we can say that disorders are mostly related to cardiac and malignancy.

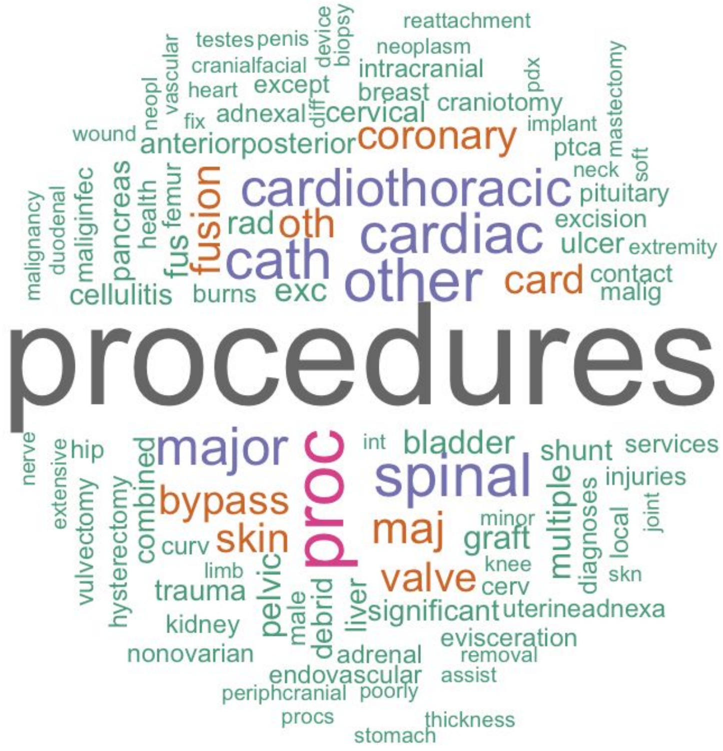
Cluster 2

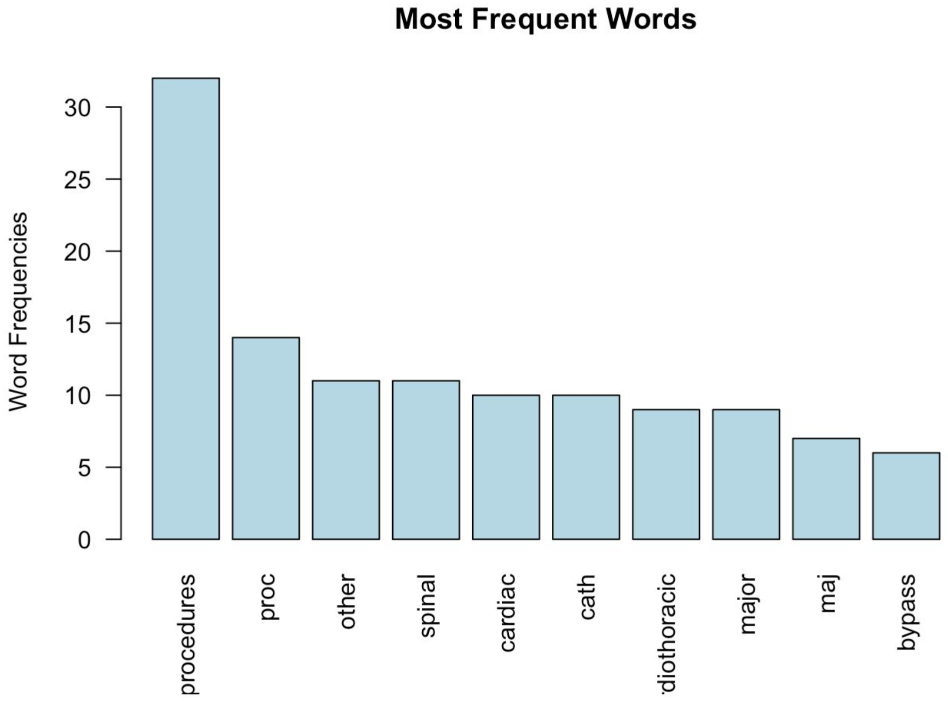




In this cluster, we can see that it has a medium cost structure, which implies that the DRGs don't cost as much. For example, we determined that in describing DRGs in this cluster, the most frequent words were procedures, proc, major, hip which we assume were not major such as joint surgeries for example. We can see a clear change from cluster 1 to cluster 2 in the most common words, like we have more procedures in cluster 2 which makes the cost higher and also most procedures are related to joint,hip or skin. Thus costs are higher in cluster 2 as compared to cluster.

Cluster 3


reattachment
testes
penis
device
biopsy
neoplasm
vascular
cranialfacial
intracranial
pdx
mastectomy
neopl
heart
except
breast
craniotomy
diff
fix
adnexal
cervical
implant
wound
anteriorposterior
coronary
ptca
soft
neck
malignancy
duodenal
maliginfec
pancreas
health
femur
fusion
cardiothoracic
pituitary
rad
oth
excision
cardiac
ulcer
extremity
fus
cath
other
card
contact
malig
cellulitis
burns
exc
procedures
nerve
extensive
hip
combined
major
int
bladder
shunt
services
proc
spinal
multiple
injuries
vulvectomy
hysterectomy
bypass
diagnoses
joint
minor
local
curv
skin
maj
graft
knee
limb
skn
trauma
valve
cerv
pelvic
male
debrid
liver
significant
uterineadnexa
kidney
evisceration
nonovarian
adrenal
removal
endovascular
assist
periphcranial
poorly
procs
thickness
stomach


Most Frequent Words
Word Frequencies
30
25
20
15
10
5
0
procedures
proc
other
spinal
cardiac
cath
diothoracic
major
maj
bypass

Cluster 3 shows a significantly high cost. Further examination of the cluster shows that the procedures might be high risk such as spinal, cardiac,cardiothoracic, spinal and other major surgeries. We can see a clear shift from cluster 2 to cluster 3 in the way DRGs are functional. This cluster has more surgery related DRGs and surgery requires more longer days of stay at the hospital hence higher costs as compared to the previous two clusters.

Contribution of OR and Anesthesiology Costs to Total Cost to Cure

All Vermont Acute Care Community Hospitals has published a report on Count of Top 2016 Inpatient Diagnoses.
(http://www.healthvermont.gov/sites/default/files/documents/pdf/2018_Table1B_0.pdf)

Cluster	DRG	Total Cost	Operating Room and Anesthesiology Cost	% of Op_Anesthin Total Cost
Cluster 1	Respiratory Infections and Inflammations with C	$16852	$3294	11%
Cluster 1	Disorders of pancreas except malignancy with C	$16852	$5227	31%
Cluster 2	Aftercare w/o CC/MCC	$22140	$9763	44%
Cluster 2	Fractures of hip & pelvis w/o MCC	$13,861	$9182	66%

Finally we understood that DRGs are clustered based on the percentage of total costs. As we move from cluster 1 to cluster 3 there is more cost weightage to operating and anesthesiology suggesting that DRGs are shifting to more surgeries as we move to cluster 3. For cluster 3, information about total cost is not available.

Conclusion

In this project, we used the 2016 Vermont Inpatient and Revenue data to analyze the cost profiles for different DRGs. We used the k-means clustering algorithm to classify DRGs against costs to see the differences in costs per DRG cluster. We used three clusters, whereby the three DRG clusters were divided into lower, medium and high cost.

We know that charges for hospital admissions with the same principal procedure cannot be expected to be identical, not only because different hospitals assign different costs, but also because different patients have different severities in their conditions. The development of the DRGs provided the first way on how we can define and measure a hospital's case mix complexity. DRGs and their use represents a fundamental role in which a hospital's case mix plays in determining treatment costs. In this project, we analyzed the different cluster costs for different DRGs using the 2016 Vermont Inpatient and Revenue files.

The DRGs form a manageable set of patient classes that relate to a hospital's case mix and costs experienced by that hospital. Since DRGs are defined based on diagnosis and procedures, it is easy to classify them into different groups based in costs. The results of the clusters show that the cost determination of each DGR is mainly based on the services the patient required, which we determined using frequent words used in that cluster. Specifically, DRGs clustered into the lower cost range show that there is mostly acute diseases and diagnosis, whereby in most cases, there is prescriptions given, and a patient is discharged. It is therefore easy to see in this case why the average cost is below $10,000. Cluster two was determined to be a medium cost structure whereby, there are procedures done but their severity is low such as joint procedures. Therefore, the cost in this cluster ranges from $8,000 to $20,000, which is reasonable. Lastly, the third cluster shows a significant increase in costs, whereby the DGRs are classified ans high costing procedures. These procedures are mainly risky and require a lot of expertise as well as a team of highly experienced physicians. It is reasonable then for this cluster to have a range of cost from $20,000 to approximately $45,000, with procedures for the main body parts such as spinal cord and the heart.

Though not stated directly, it is clear that cluster one has mostly non-surgical DGRs compared to cluster two and three where there are surgical procedures, whereby cluster three requires specialty physicians involved. It is important to note that according to DRG classifications, a surgical patient is classified as surgical if that patient also requires an operating room. For example, cluster two shows that there are skin related diseases where a skin suture wouldn't require an operating room but if a patient requires a biopsy, they would also require an operating room, and thus would be classified as a surgical patient. The two procedures are moderately priced, so they are both grouped into cluster two. From the results, we also determined that the complexity of treatment especially through surgical procedures increases with the clusters. Fro example, cluster one has less complicated diseases as classified by CC or MCC, while cluster three has the highest percentage of complications. This makes sense because if a patient is diagnosed with a disorder for example, chances that there are complications with that a lower compared to a patient who undergoes a spinal surgery.

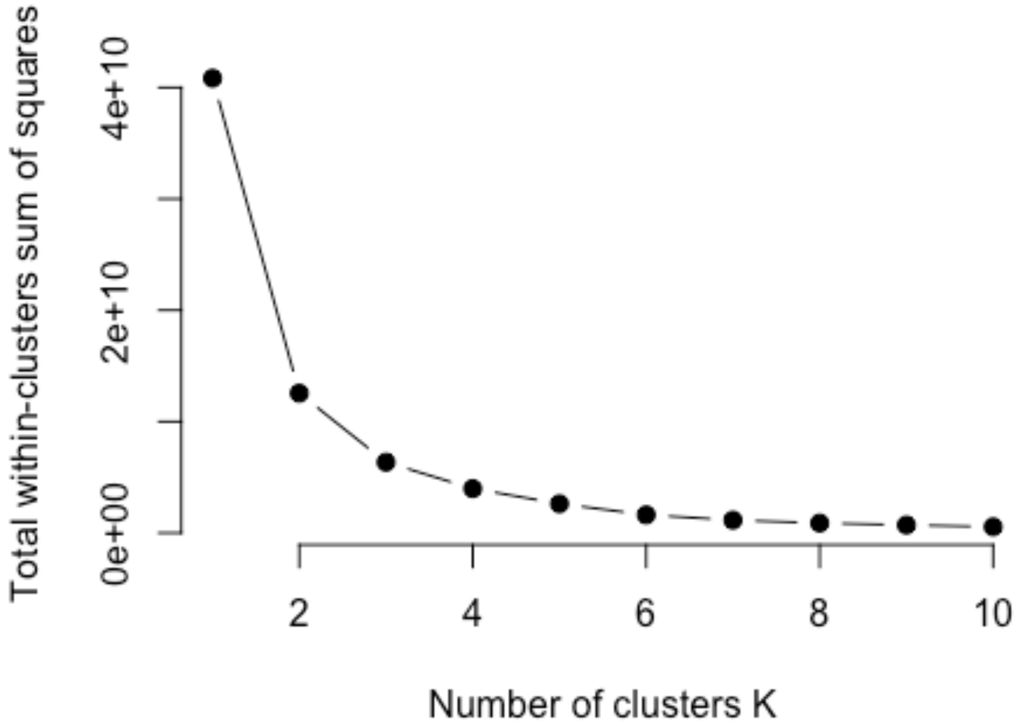
In this project, we learned that through DRGs, hospitals can gain an understanding of the patients being treated and the costs they incur while doing so. Classifying DRGs into different cost clusters through k-means clustering gave us useful insights into how DGRs are grouped together; range of cost in this case.

Appendix of Healthcare HW3

Written by Zhiqi Chen, Nkosingiphile Shongwe, Salil Redkar, Debarati Mazumdar, Lakshmi Malavika Andavilli

3/5/2019

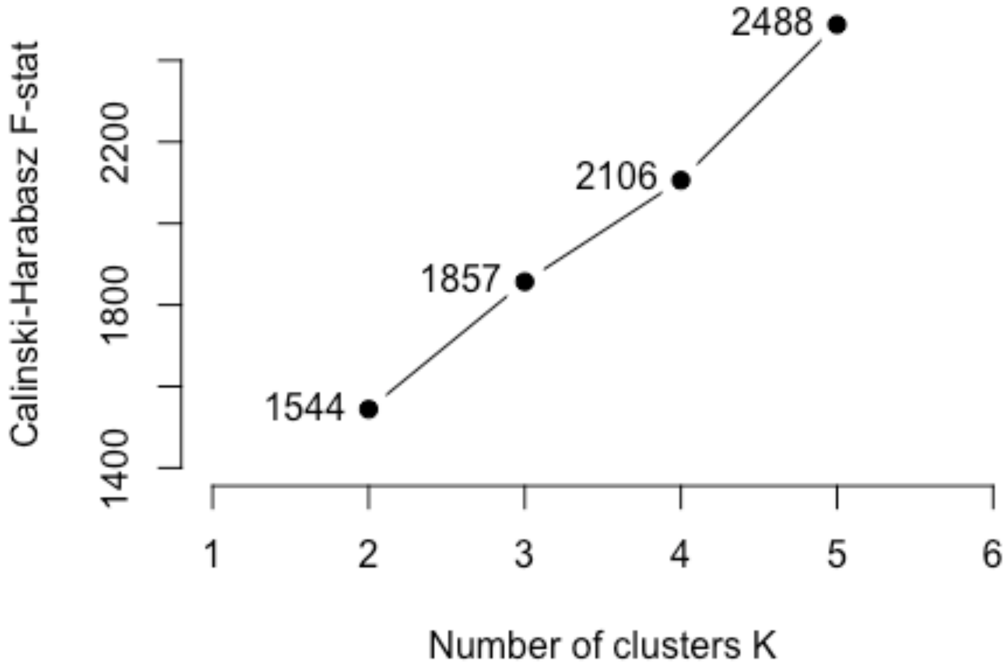
```r
library(readxl)
library(tidyverse)
library(clusterCrit)
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
library(readtext)

inpatient = read.delim("VTINP16_upd.TXT", sep=',')

revcode = read.delim("VTREVCODE16.TXT", sep=',')
colnames(revcode)[colnames(revcode) == 'Uniq'] = 'UNIQ'

drg_name = read_excel("HS_2016VT_PUF_FILE_LAYOUT_and_CODES.xls", sheet="MSDRG
2007 forward")
drg_name = drg_name %>%
  select(MSDRG, MSDRG_DESC)
colnames(drg_name)[colnames(drg_name) == 'MSDRG'] = 'DRG'

PCCR_name = read_excel("HS_2016VT_REVCODE_FILE_LAYOUT_and_CODES.xls",
sheet="PCCR")

inp_filter = inpatient %>%
  filter(DRG %in% (20:977)) %>%
  select(UNIQ, DRG)

revcode = revcode %>%
  filter(!REVCHRGS < 100) %>%
  select(UNIQ, PCCR, REVCHRGS)

inp_merge = merge(inp_filter, revcode, by = "UNIQ")

inp_merge = na.omit(inp_merge)

inp_merge = inp_merge %>%
  group_by(DRG, PCCR) %>%
  summarize(CHRGS = round(mean(REVCHRGS)))

inp_merge = merge(inp_merge, drg_name, by = "DRG") %>%
  select(c(4, 2, 3))

inp_merge = merge(inp_merge, PCCR_name, by = "PCCR") %>%
  select(c(2, 4, 3))

colnames(inp_merge)[colnames(inp_merge) == 'MSDRG_DESC'] = 'DRG'
colnames(inp_merge)[colnames(inp_merge) == 'PCCR_NAME'] = 'PCCR'

ctab1 = inp_merge %>%
  spread(PCCR, CHRGS)

ctab2 = ctab1 %>%
  mutate(PCCR_OR_and_Anesth_Costs = ctab1$`Operating Room` +
ctab1$`Anesthesiology`)

ctab3 = ctab2[,-1]
rownames(ctab3) = ctab2[,1]

ctab3[is.na(ctab3)] = 0

ctab4 = ctab3 %>%
  select(PCCR_OR_and_Anesth_Costs)

seg.summ <- function(data, groups) {
  aggregate(data, list(groups), function(x) mean(as.numeric(x)))
}

seg.df.num = model.matrix(~ PCCR_OR_and_Anesth_Costs, data = ctab4)

k.max <- 10
wss <- sapply(1:k.max,
              function(k){kmeans(seg.df.num, k, nstart = 50, iter.max =
20)$tot.withinss})

plot(1:k.max, wss,
     type="b", pch=19, frame=FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```



```
seg.k2 <- kmeans(seg.df.num, centers=2, nstart=50, iter.max=20)
seg.k3 <- kmeans(seg.df.num, centers=3, nstart=50, iter.max=20)
seg.k4 <- kmeans(seg.df.num, centers=4, nstart=50, iter.max=20)
seg.k5 <- kmeans(seg.df.num, centers=5, nstart=50, iter.max=20)

k2 = intCriteria(seg.df.num, seg.k2$cluster, "Calinski_Harabasz")
k3 = intCriteria(seg.df.num, seg.k3$cluster, "Calinski_Harabasz")
k4 = intCriteria(seg.df.num, seg.k4$cluster, "Calinski_Harabasz")
k5 = intCriteria(seg.df.num, seg.k5$cluster, "Calinski_Harabasz")

print(paste("F-stat for 2 clusters is", k2))

## [1] "F-stat for 2 clusters is 1543.97743118299"

print(paste("F-stat for 3 clusters is", k3))

## [1] "F-stat for 3 clusters is 1857.29665736858"

print(paste("F-stat for 4 clusters is", k4))

## [1] "F-stat for 4 clusters is 2106.24751450324"

print(paste("F-stat for 5 clusters is", k5))
```

```
## [1] "F-stat for 5 clusters is 2487.55745025304"
```

```r
f = c(1544, 1857, 2106, 2488)
plot(2:5, f,
     type="b", pch=19, frame=FALSE,
     xlab="Number of clusters K",
     ylab="Calinski-Harabasz F-stat",
     xlim=c(1,6),
     ylim=c(1400, 2500))
text(2:5, f, labels=f, pos=2)
```

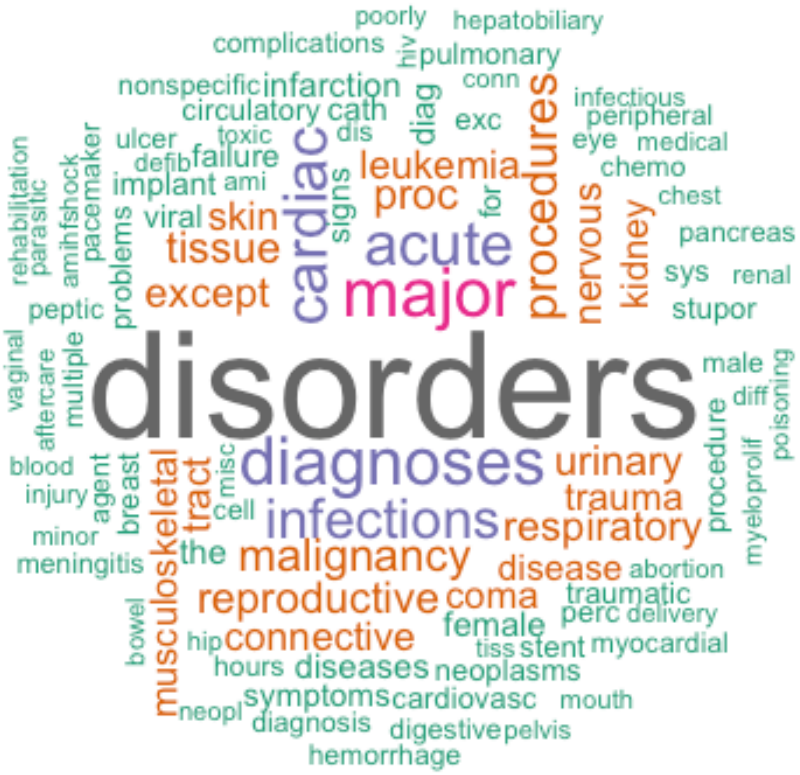


```r
#cluster_number = seg.k3$cluster
#ctab4$Sorted_Cluster = cluster_number

# Most Frequent Words in Cluster 1
text = readtext("cluster1.txt")
docs = Corpus(VectorSource(text))
docs = tm_map(docs, content_transformer(tolower))
dtm = TermDocumentMatrix(docs)
m = as.matrix(dtm)
v = sort(rowSums(m), decreasing = TRUE)
d = data.frame(word = names(v), freq = v)
d = d[c(-1, -2), ]
```

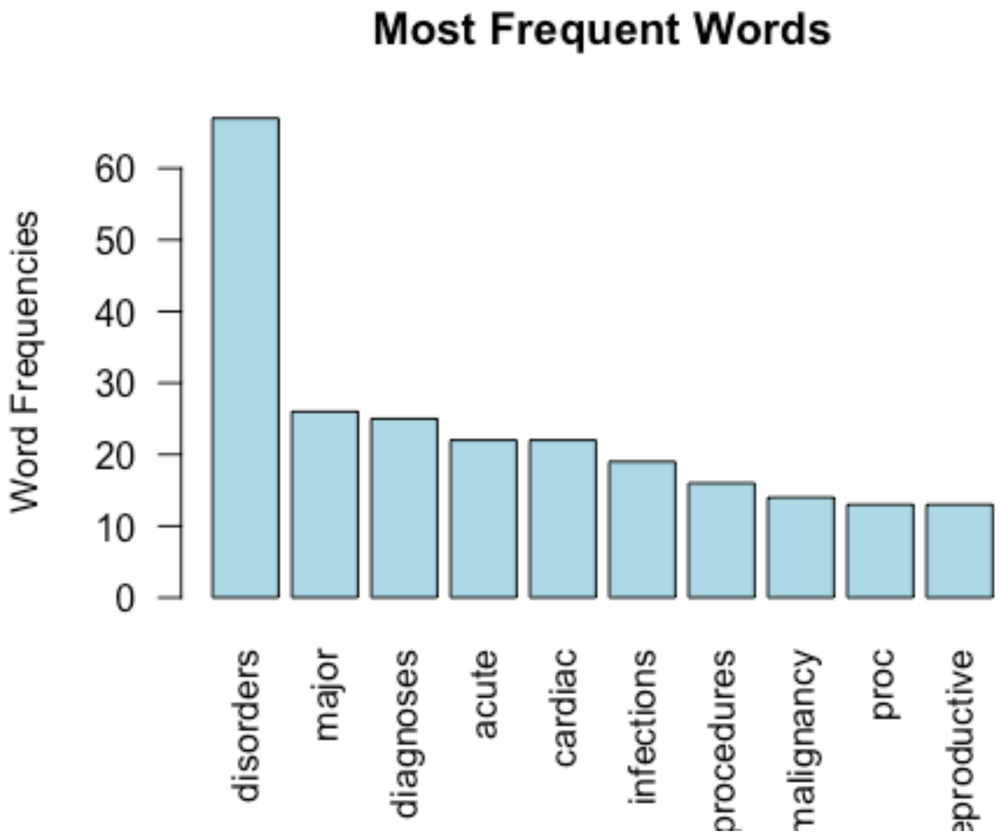
```
d = d[c(-2, -3), ]
head(d, 10)
```

```
##                    word freq
## disorders       disorders   67
## major               major   26
## diagnoses       diagnoses   25
## acute               acute   22
## cardiac           cardiac   22
## infections     infections   19
## procedures     procedures   16
## malignancy     malignancy   14
## proc                 proc   13
## reproductive reproductive   13
```

```
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words = 100, random.order = FALSE, rot.per = 0.35,
          colors = brewer.pal(8, "Dark2"))
```



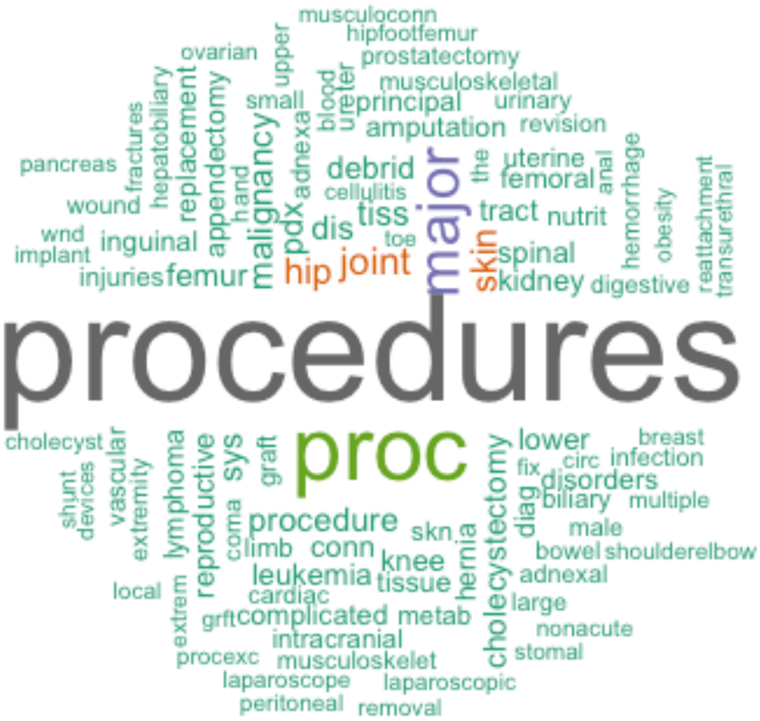
```
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,
        col = "lightblue", main ="Most Frequent Words",
        ylab = "Word Frequencies")
```



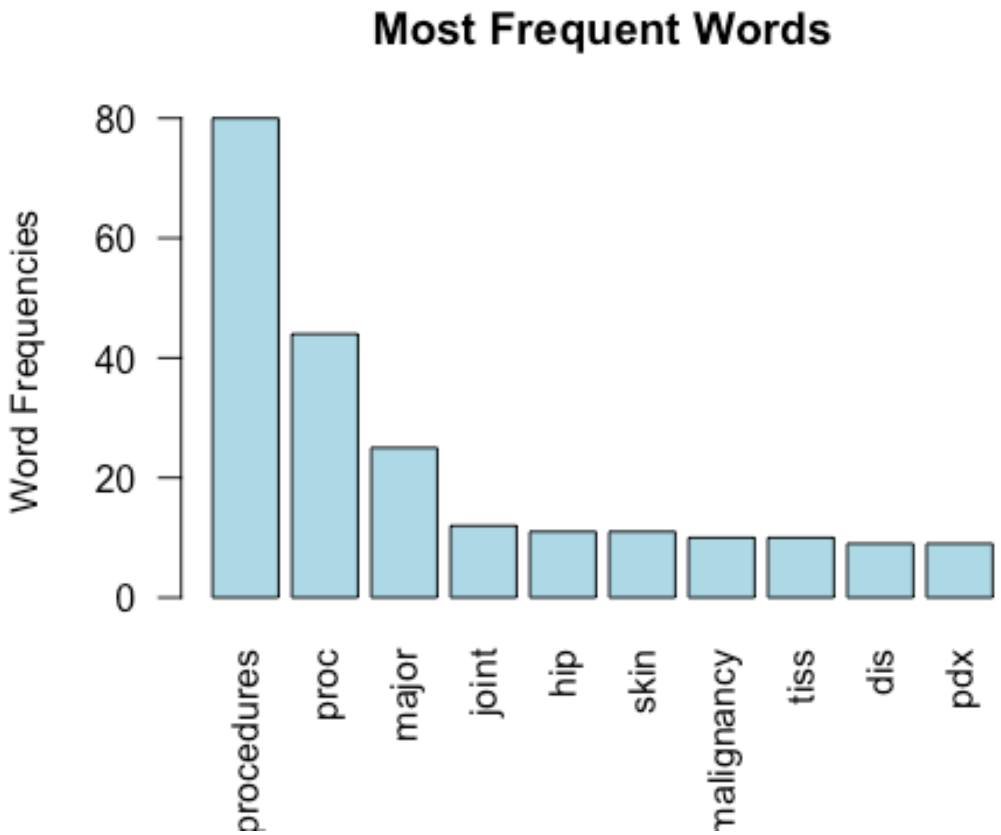
```
# Most Frequent Words in Cluster 2
text = readtext("cluster2.txt")
docs = Corpus(VectorSource(text))
docs = tm_map(docs, content_transformer(tolower))
dtm = TermDocumentMatrix(docs)
m = as.matrix(dtm)
v = sort(rowSums(m), decreasing = TRUE)
d = data.frame(word = names(v), freq = v)
d = d[c(-1, -3), ]
d = d[c(-3, -4, -6, -7, -8, -10), ]
head(d, 10)
```

```
##                   word freq
## procedures procedures   80
## proc              proc   44
## major            major   25
## joint            joint   12
## hip                hip   11
## skin              skin   11
## malignancy malignancy   10
## tiss              tiss   10
## dis                dis    9
## pdx                pdx    9
```

```
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words = 100, random.order = FALSE, rot.per = 0.35,
          colors = brewer.pal(8, "Dark2"))
```



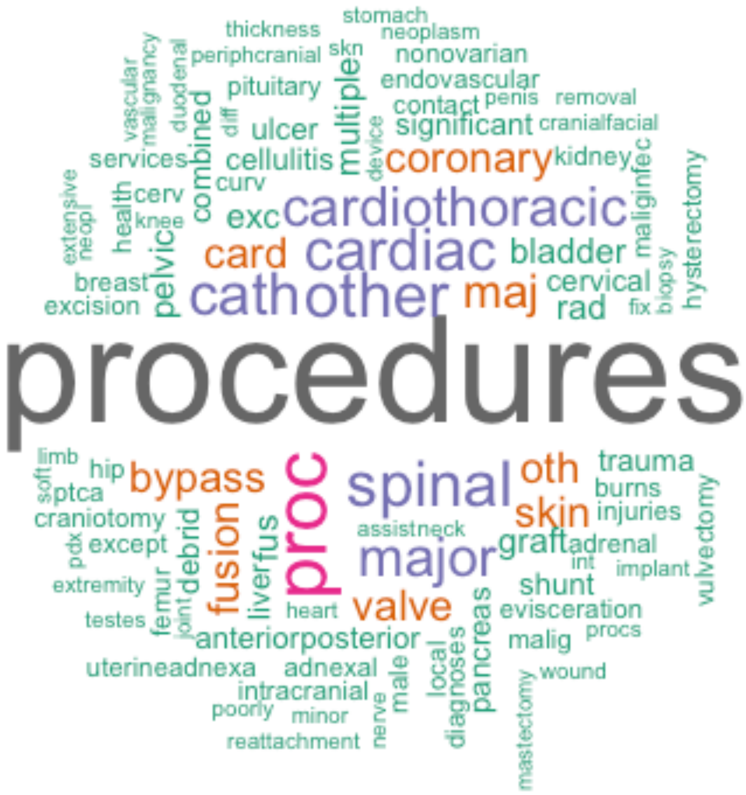
```
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,
        col = "lightblue", main ="Most Frequent Words",
        ylab = "Word Frequencies")
```



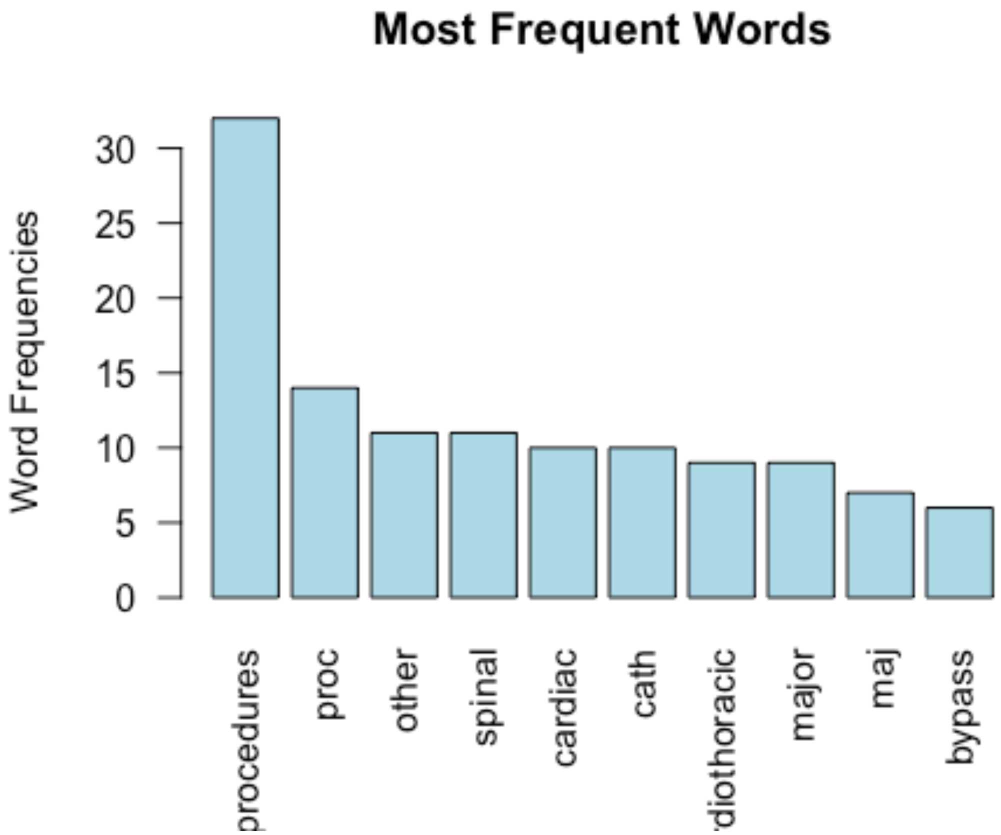
```
# Most Frequent Words in Cluster 3
text = readtext("cluster3.txt")
docs = Corpus(VectorSource(text))
docs = tm_map(docs, content_transformer(tolower))
dtm = TermDocumentMatrix(docs)
m = as.matrix(dtm)
v = sort(rowSums(m), decreasing = TRUE)
d = data.frame(word = names(v), freq = v)
d = d[c(-1, -3), ]
d = d[-3, ]
head(d, 10)
```

```
##                           word freq
## procedures          procedures   32
## proc                      proc   14
## other                    other   11
## spinal                  spinal   11
## cardiac                cardiac   10
## cath                      cath   10
## cardiothoracic cardiothoracic    9
## major                    major    9
## maj                        maj    7
## bypass                  bypass    6
```

```r
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words = 100, random.order = FALSE, rot.per = 0.35,
          colors = brewer.pal(8, "Dark2"))
```



```r
barplot(d[1:10,]$freq, las = 2, names.arg = d[1:10,]$word,
        col = "lightblue", main ="Most Frequent Words",
        ylab = "Word Frequencies")
```


Most Frequent Words
Word Frequencies
0
5
10
15
20
25
30
procedures
proc
other
spinal
cardiac
cath
diothoracic
major
maj
bypass