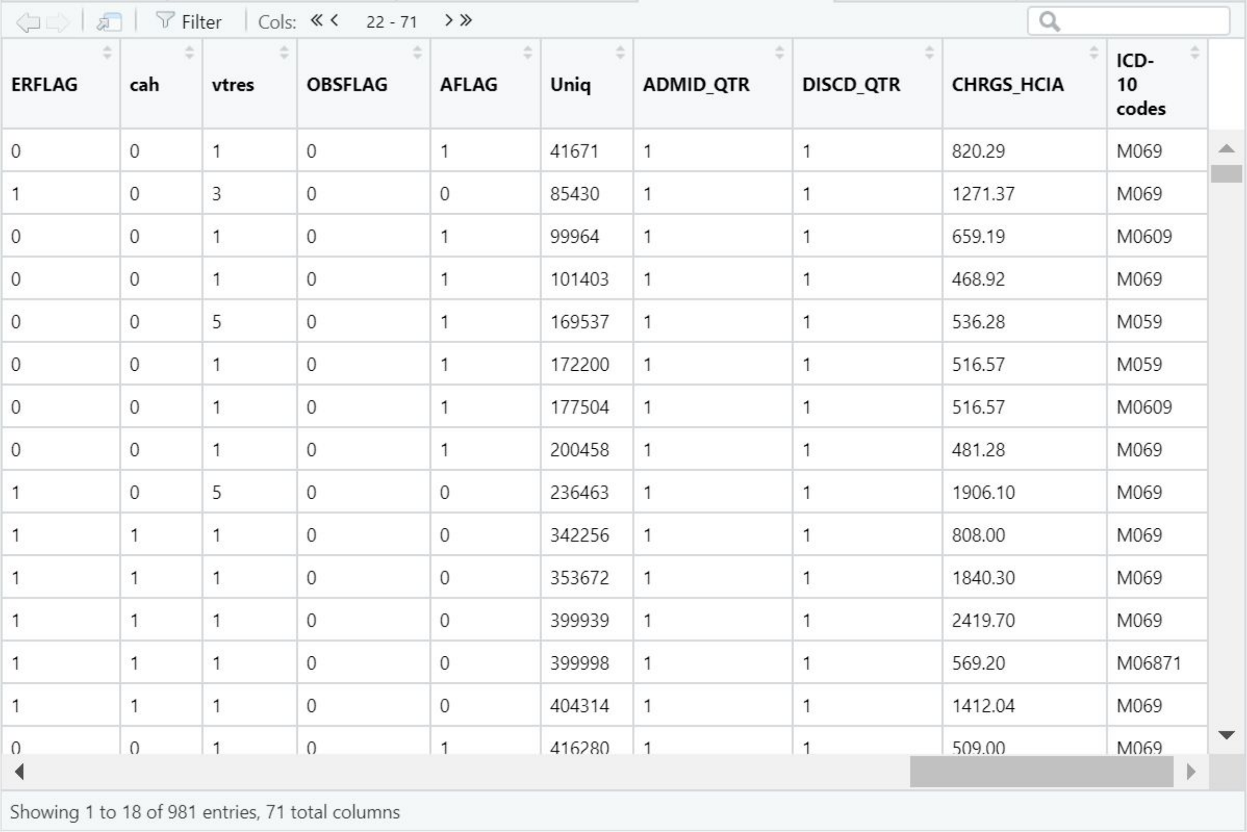
Analysis Report of Insurance Claim Data & Analytics

Group 2
Menghong Han, Xia Dai, Rui Cao, Mengchun Li, Xuhui Bai, Bo Chen

Question 1 Study of A Disease Cohort

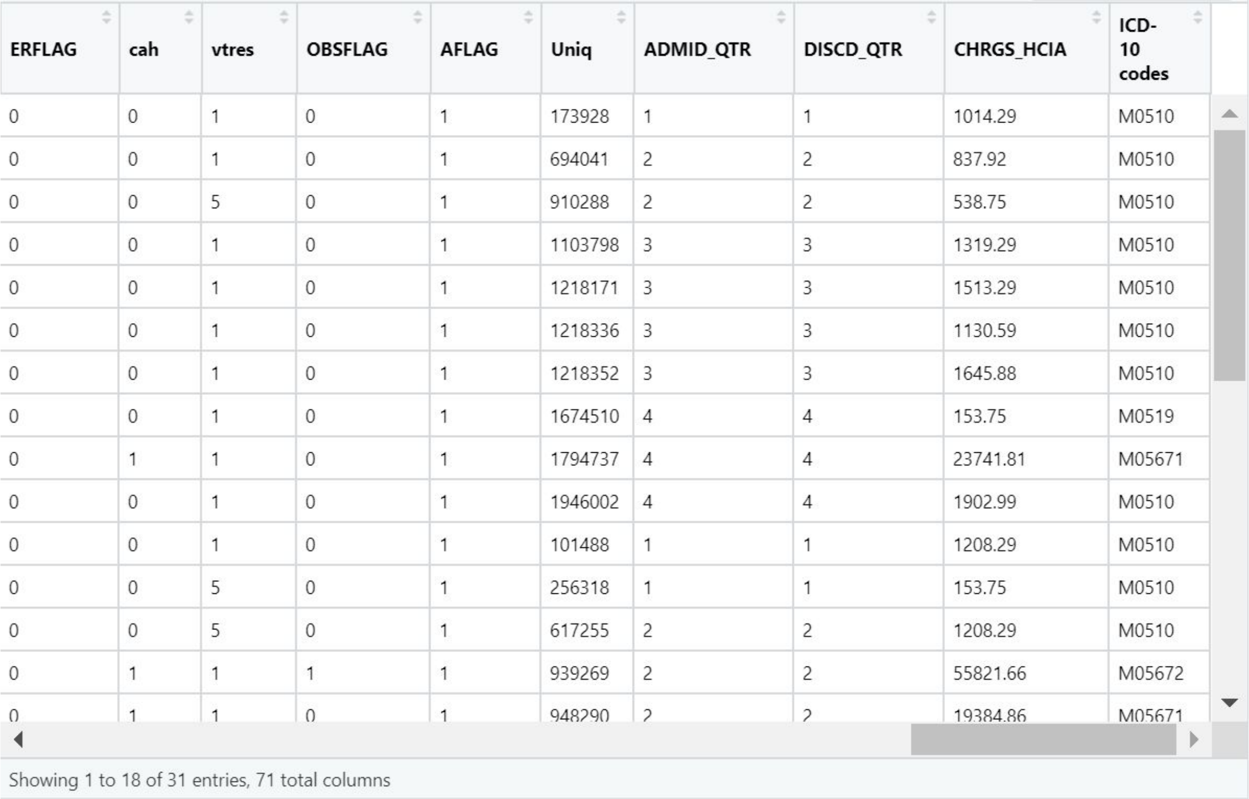
Step 1: Identify the RA cohort using the outpatient file

cohort for the first tab in the Excel file (chronic RA) :



ERFLAG	cah	vtres	OBSFLAG	AFLAG	Uniq	ADMID_QTR	DISCD_QTR	CHRGS_HCIA	ICD-10 codes
0	0	1	0	1	41671	1	1	820.29	M069
1	0	3	0	0	85430	1	1	1271.37	M069
0	0	1	0	1	99964	1	1	659.19	M0609
0	0	1	0	1	101403	1	1	468.92	M069
0	0	5	0	1	169537	1	1	536.28	M059
0	0	1	0	1	172200	1	1	516.57	M059
0	0	1	0	1	177504	1	1	516.57	M0609
0	0	1	0	1	200458	1	1	481.28	M069
1	0	5	0	0	236463	1	1	1906.10	M069
1	1	1	0	0	342256	1	1	808.00	M069
1	1	1	0	0	353672	1	1	1840.30	M069
1	1	1	0	0	399939	1	1	2419.70	M069
1	1	1	0	0	399998	1	1	569.20	M06871
1	1	1	0	0	404314	1	1	1412.04	M069
0	0	1	0	1	416280	1	1	509.00	M069

Showing 1 to 18 of 981 entries, 71 total columns the second cohort based on the Excel file tab “other RA with systemic involvement”

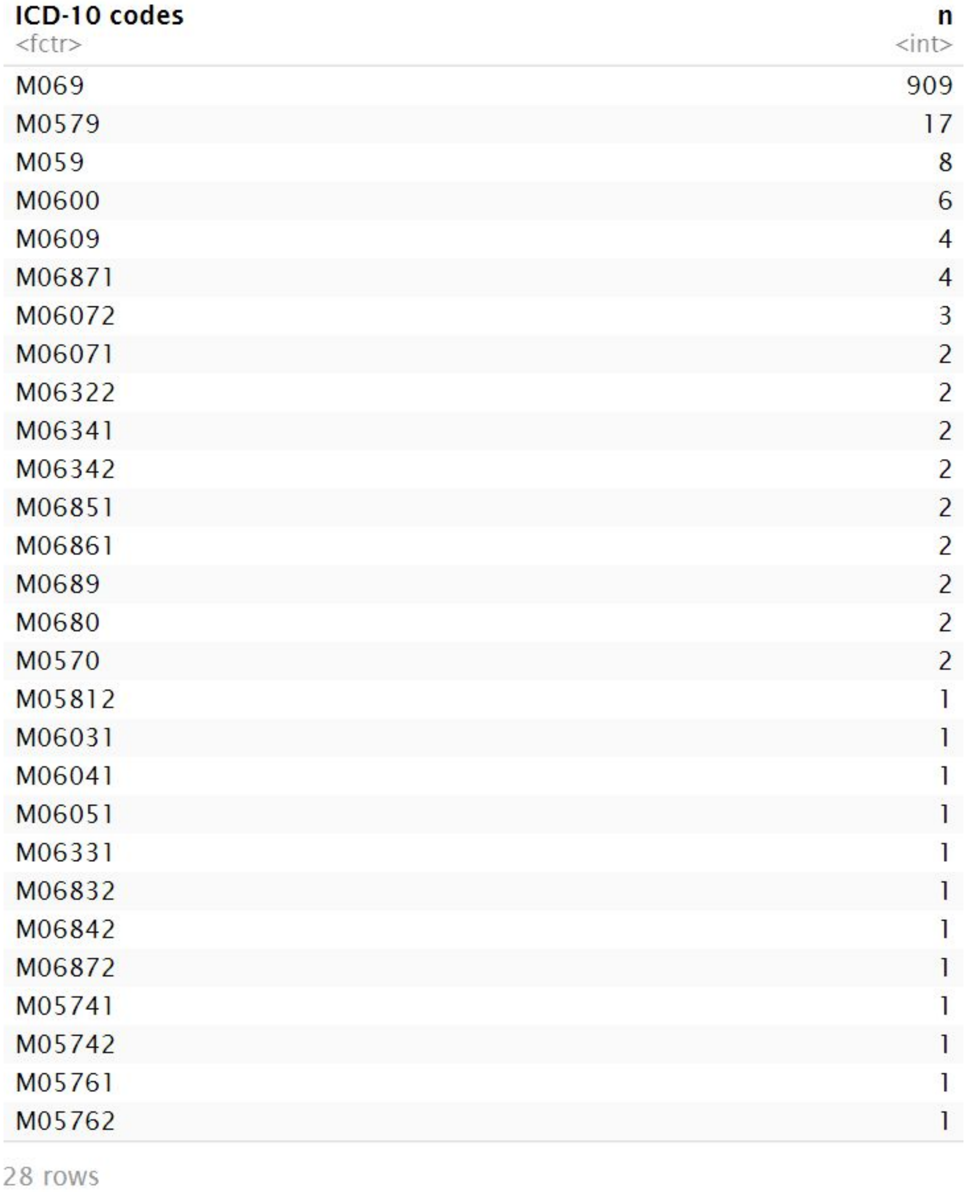


ERFLAG	cah	vtres	OBSFLAG	AFLAG	Uniq	ADMID_QTR	DISCD_QTR	CHRGS_HCIA	ICD-10 codes
0	0	1	0	1	173928	1	1	1014.29	M0510
0	0	1	0	1	694041	2	2	837.92	M0510
0	0	5	0	1	910288	2	2	538.75	M0510
0	0	1	0	1	1103798	3	3	1319.29	M0510
0	0	1	0	1	1218171	3	3	1513.29	M0510
0	0	1	0	1	1218336	3	3	1130.59	M0510
0	0	1	0	1	1218352	3	3	1645.88	M0510
0	0	1	0	1	1674510	4	4	153.75	M0519
0	1	1	0	1	1794737	4	4	23741.81	M05671
0	0	1	0	1	1946002	4	4	1902.99	M0510
0	0	1	0	1	101488	1	1	1208.29	M0510
0	0	5	0	1	256318	1	1	153.75	M0510
0	0	5	0	1	617255	2	2	1208.29	M0510
0	1	1	1	1	939269	2	2	55821.66	M05672
0	1	1	0	1	948290	2	2	19384.86	M05671

Showing 1 to 18 of 31 entries, 71 total columns

Step 2: Identify the most common types of RA

frequency of each of the ICD-10 codes for the cohort of chronic RA:



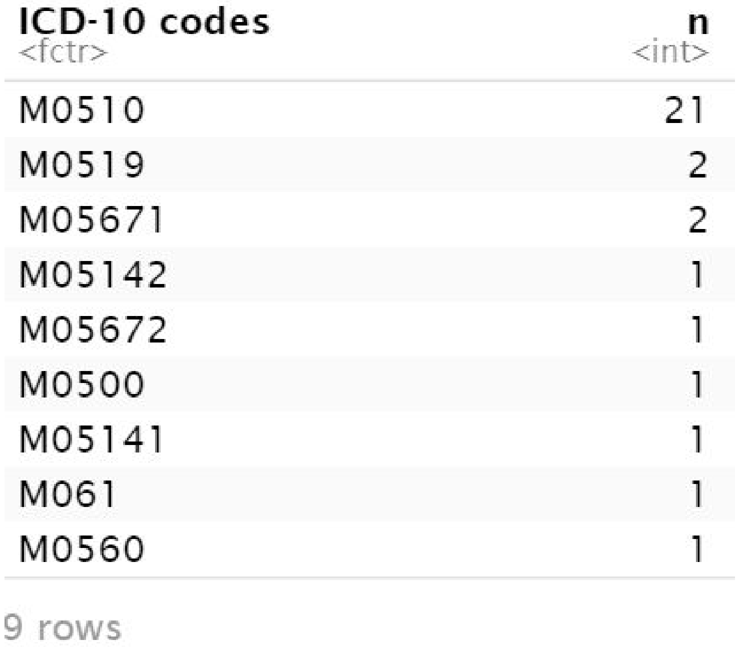
ICD-10 codes <fctr>	n <int>
M069	909
M0579	17
M059	8
M0600	6
M0609	4
M06871	4
M06072	3
M06071	2
M06322	2
M06341	2
M06342	2
M06851	2
M06861	2
M0689	2
M0680	2
M0570	2
M05812	1
M06031	1
M06041	1
M06051	1
M06331	1
M06832	1
M06842	1
M06872	1
M05741	1
M05742	1
M05761	1
M05762	1

28 rows

the top-3 ICD-10 codes for the cohort of chronic RA:

M069 M0579 M059

frequency of each of the ICD-10 codes for the cohort of “other RA with systemic involvement”:



ICD-10 codes <fctr>	n <int>
M0510	21
M0519	2
M05671	2
M05142	1
M05672	1
M0500	1
M05141	1
M061	1
M0560	1

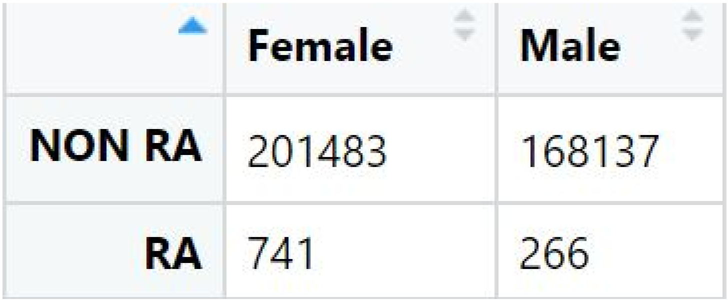
9 rows

the top-3 ICD-10 codes for the cohort of “other RA with systemic involvement”:

M0510 M0519 M05671

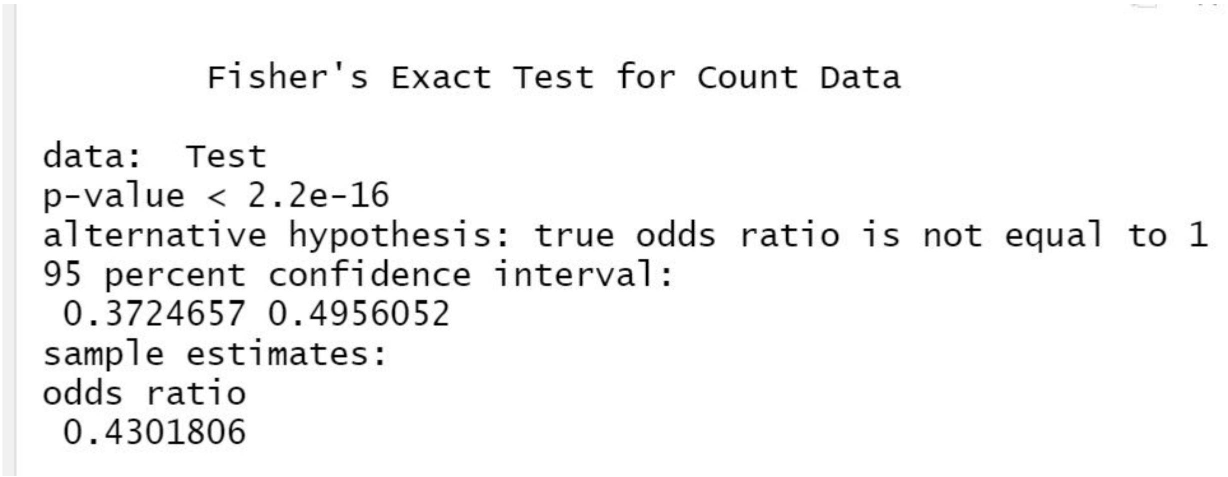
Step 3:

To conduct Fisher’s Exact Test, we should set up a 2*2 table on gender and RA. Firstly, we filter out the non-RA patients and get the specific number of male and female patients free of RA and then we calculate the number of male and female patients with RA. On top of that, we made a contingency table like this:



	Female	Male
NON RA	201483	168137
RA	741	266

We conduct the Fisher’s Test on this table, and the result is shown below:



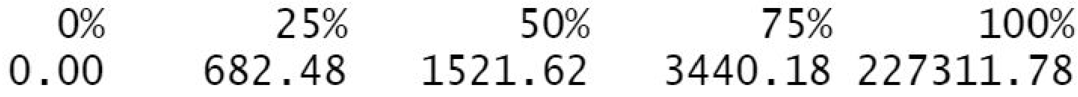
```
        Fisher's Exact Test for Count Data

data:  Test
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3724657 0.4956052
sample estimates:
odds ratio
 0.4301806
```

the p-value is so small that the result is statistically significant and we can reject the null hypothesis and claim that there exists a relationship of RA with gender.

Step 4:

By conducting the quantile of charge of outpatient file, we get:

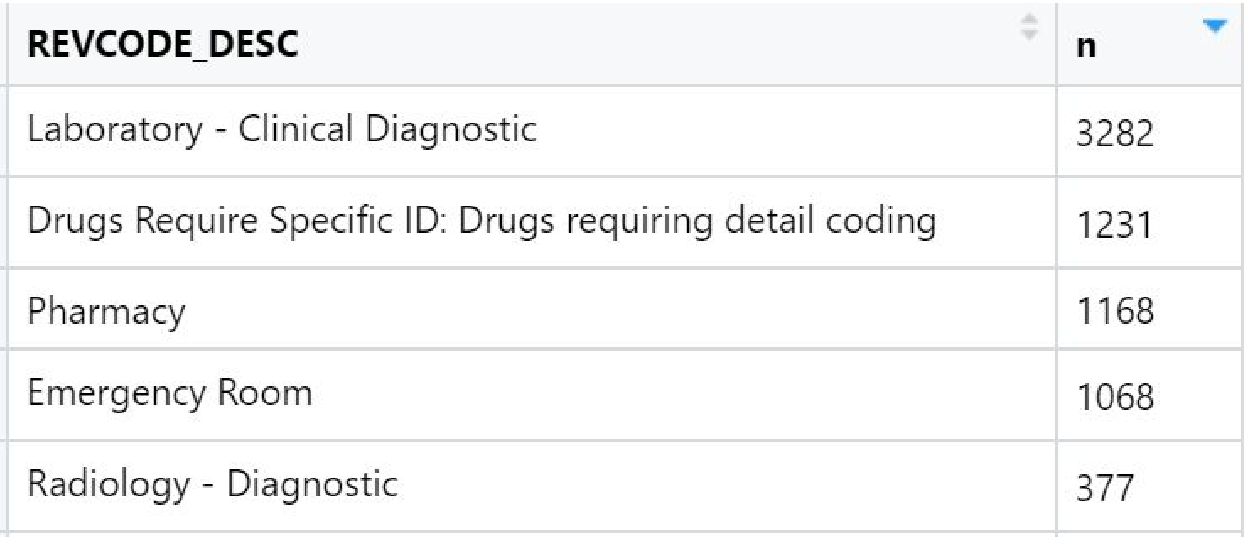


```
       0%       25%       50%       75%      100%
     0.00    682.48   1521.62   3440.18 227311.78
```

IQR= Q3-Q1=3440.18-682.48=2757.7

Step5:

To get the five most common services for treatment of RA, we merge the tables of revcode and rev with the key “REVCODE” and then merge this table with data of RA with the key”Uniq”, and finally we derive the following table which can perfectly explain the top five popular service of RA:



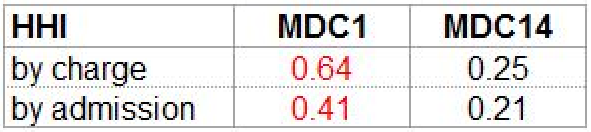
REVCODE_DESC	n
Laboratory - Clinical Diagnostic	3282
Drugs Require Specific ID: Drugs requiring detail coding	1231
Pharmacy	1168
Emergency Room	1068
Radiology - Diagnostic	377

Question 2 Which clinical chapter as defined by the Major Diagnostic Category MDC in inpatient care is more concentrated among a few big players?

Before the analytics, our team guessed MDC 14 : Pregnancy, Childbirth, And Puerperium would be done more generally by most of the hospitals as this is a common need for most families; and MDC 1: Diseases; and Disorders of the Brain tends to be highly concentrated among specialized high technology medical centers, as the brain is the most important and complex part of the body, treatments to it must be precise and high-tech.

Our analysis is as below:

Step One: Use HHI index to find out the concentration of the care for the two MDCs.

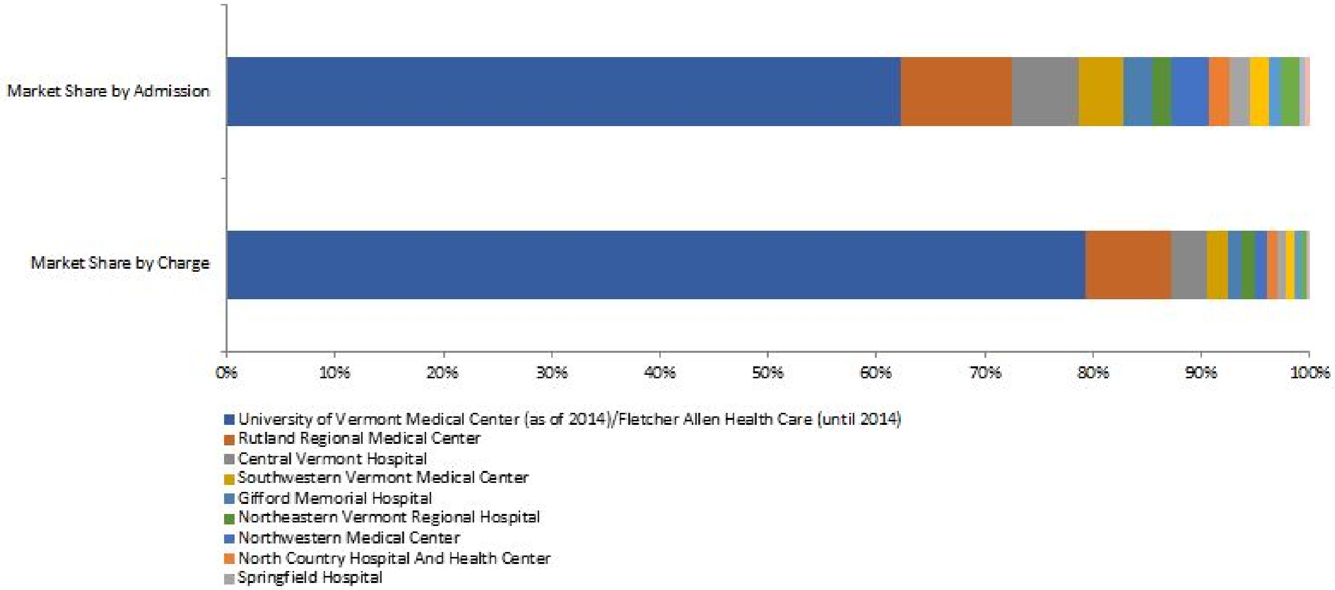


HHI	MDC1	MDC14
by charge	0.64	0.25
by admission	0.41	0.21

As shown above, MDC1 market is less competitive and players in it have higher market power.

Step Two: So we focus on the more concentrated MDC - MDC1 to find more information about the player in it

We calculated players' market share by charge and by admission in MDC1market as below:



Telling from the above chart, University of Vermont Medical Center holds the "lion share" of the market both from admission counts (62.24%)and for dollars(79.27%).

Learned from the website, University of Vermont Medical Center offers comprehensive, family-centered primary care, specialty care, and neonatal and pediatric intensive care for children across the region; The medical center has comprehensive surgical services (neurological, cardiac, pediatric) and imaging equipment. As the only Level I Trauma Center in Vermont, the hospital offers the region advanced technology and techniques to care for the most seriously ill and injured pediatric and adult patients.[1]

So compared to our original guess, we are right!

Q3: Clustering Costs

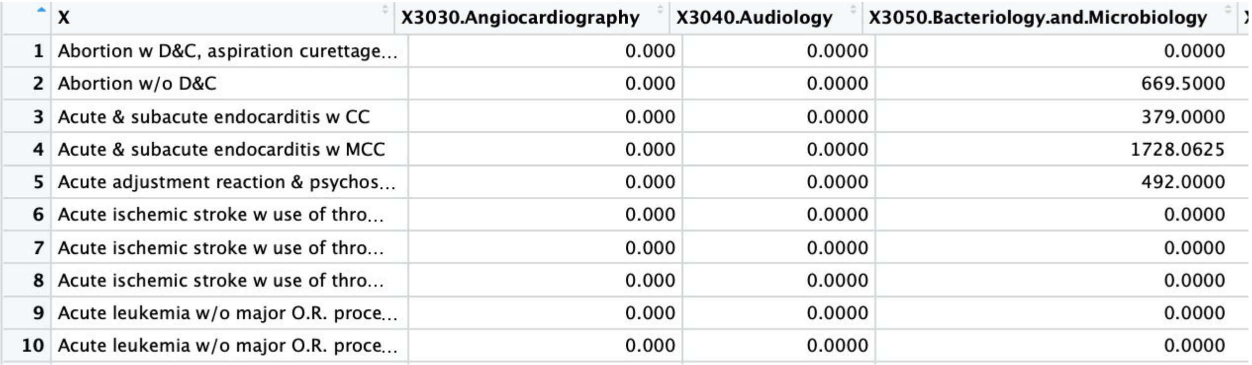
For this clustering, we firstly did data cleaning and preparation. After that, we loaded the data to R for clustering. We explored the cluster with two methods as following steps: 1)

[1] From Wikipedia： https://en.wikipedia.org/wiki/University_of_Vermont_Medical_Center

Normalize the data. 2) Calculate the distances to have an initial understanding. 3) Method 1: directly use classification k-means and choose different k to cluster the data. 4) Method 2: add self organizing maps (SOM) before k-means to cluster the data. 5) Analyze and evaluate the goodness of different clusters by f-statistics, silhouette value, between Cluster Sum of Squares and within Cluster Sum of Squares. 6) Choose the better clustering and verify the clustering by understanding the attributes of each cluster, such MDC group name, DRG name, etc.

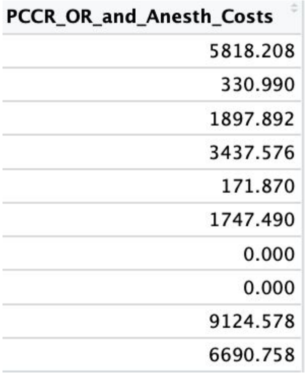
1. Data cleaning and preparation

a. Link DRG code and PCCR code with DRG and PCCR
b. Filter DRGs between 20 and 977
c. Merge filtered DRG to the Revenue Code file on UNIQ
d. Exclude the low dollar value services (less than $100)
e. Sum all charges group by DRG, PCCR categories
f. Cross tabulate with selected DRGs (in the row) and the mean value of the PCCR
g. Combine the PCCR 3700 Operating Room & PCCR 4000 Anesthesiology
h. Turn NA to 0



	X	X3030.Angiocardiography	X3040.Audiology	X3050.Bacteriology.and.Microbiology
1	Abortion w D&C, aspiration curettage...	0.000	0.0000	0.0000
2	Abortion w/o D&C	0.000	0.0000	669.5000
3	Acute & subacute endocarditis w CC	0.000	0.0000	379.0000
4	Acute & subacute endocarditis w MCC	0.000	0.0000	1728.0625
5	Acute adjustment reaction & psychos...	0.000	0.0000	492.0000
6	Acute ischemic stroke w use of thro...	0.000	0.0000	0.0000
7	Acute ischemic stroke w use of thro...	0.000	0.0000	0.0000
8	Acute ischemic stroke w use of thro...	0.000	0.0000	0.0000
9	Acute leukemia w/o major O.R. proce...	0.000	0.0000	0.0000
10	Acute leukemia w/o major O.R. proce...	0.000	0.0000	0.0000

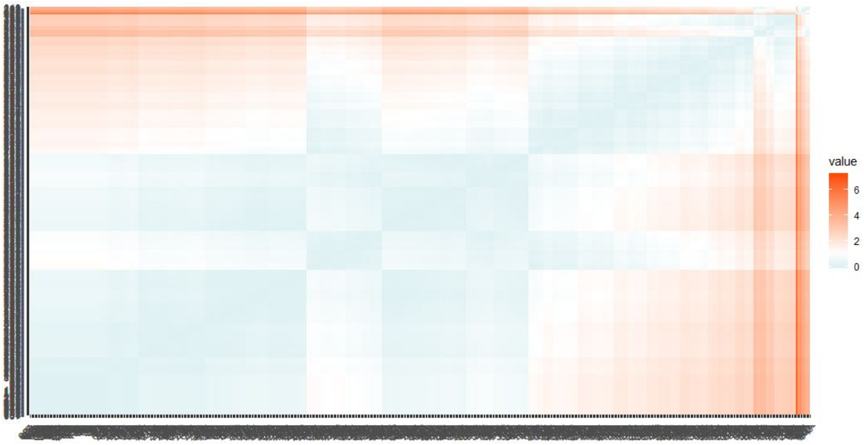
… ...



PCCR_OR_and_Anesth_Costs
5818.208
330.990
1897.892
3437.576
171.870
1747.490
0.000
0.000
9124.578
6690.758

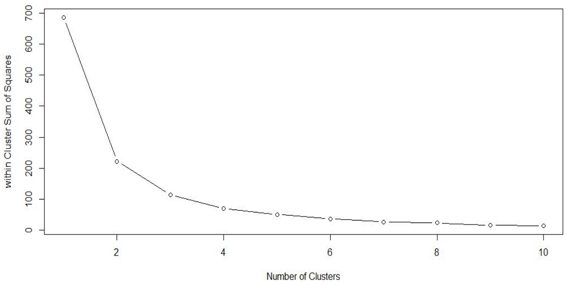
2. Cluster Exploration and Verification

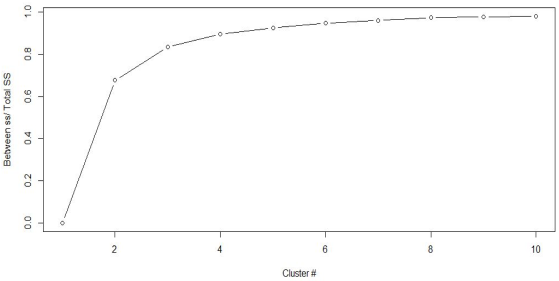
   After data normalization, we explored the PCCR Operating Room + Anesthesiology distance and visualize the distance as follows. We got a rough idea about the data that most of the values are close to each other, but still there are some values with larger distance.



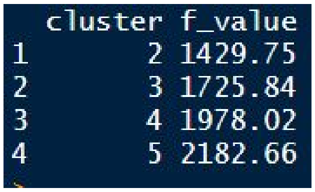
Method1: Only use k-means

We plot between Cluster Sum of Squares and within Cluster Sum of Squares under k from 1 to 10. From the plots we can infer that 3 or 4 is a better number for k because the relatively lower within and larger between Sum of Squares.



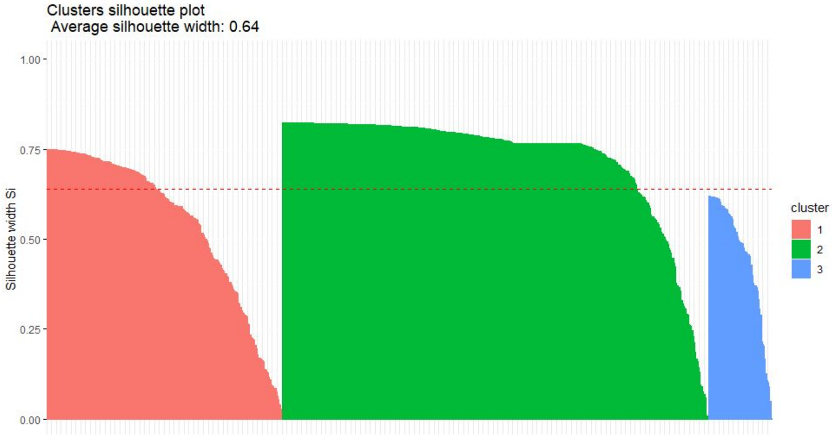


Then we tried 2 to 5 for k and obtained the f-statistics under each k number. We can see f-statistics value increases with cluster number increases.



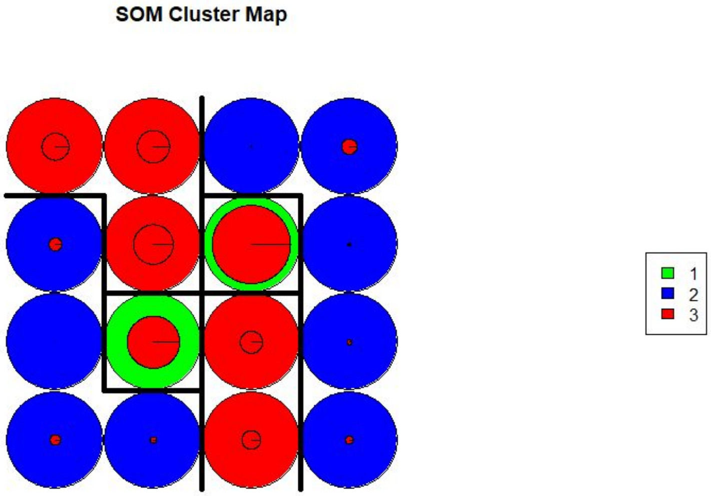
```
  cluster f_value
1       2 1429.75
2       3 1725.84
3       4 1978.02
4       5 2182.66
```

After that, we evaluated our cluster by Silhouette Index.The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). If silhouette value is close to 1, sample is well-clustered and already assigned to a very appropriate cluster. From the following plot, we can see the average silhouette value is above 0.5 and no negative value. Therefore, this is a pretty good clustering.

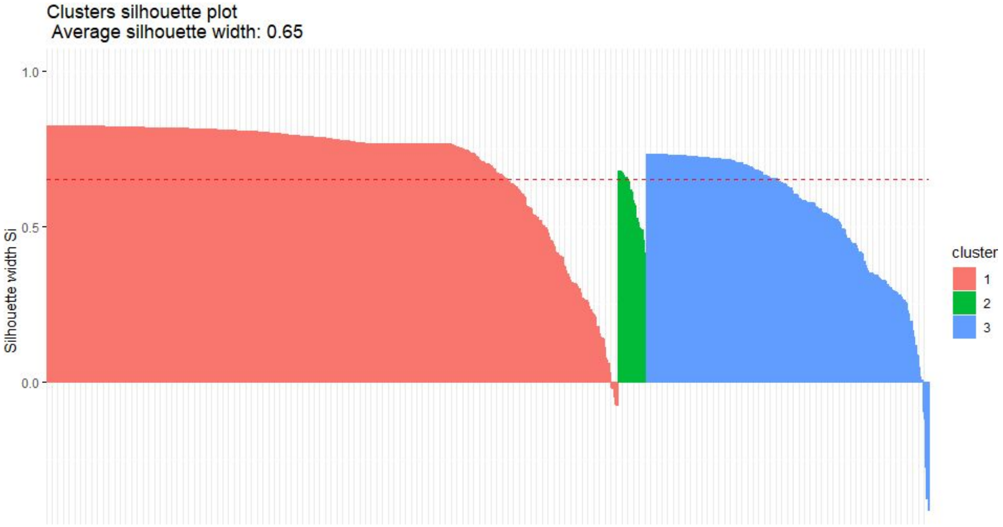


Method 2: Add self organizing map (SOM) before k-means

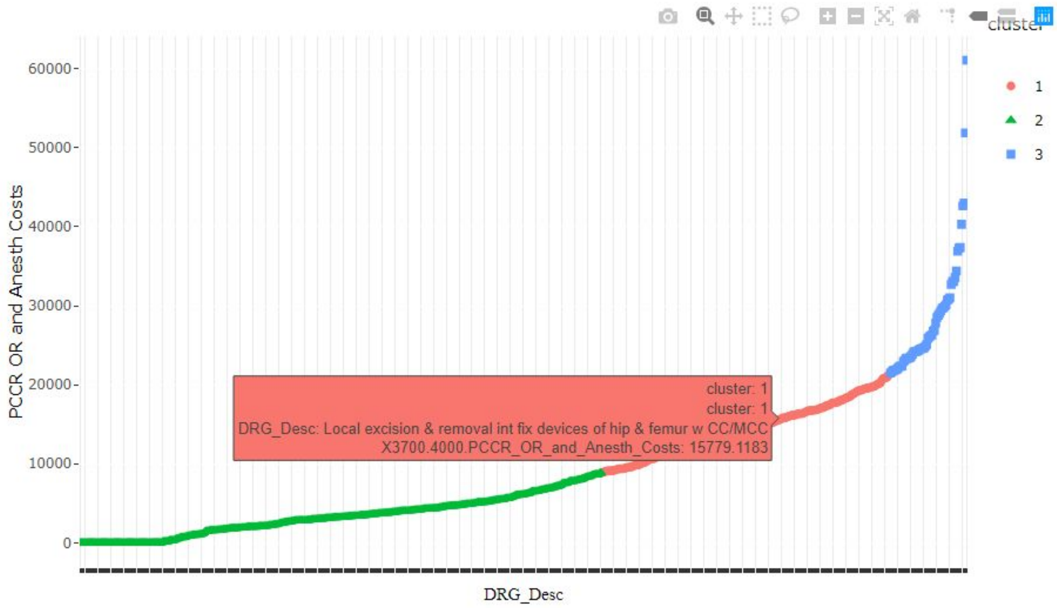
In addition, we also tried to add SOM step before k-means to see whether we can find better cluster. Firstly, we map the data to 4 by 4 mapping and obtained the cluster as follows with lower f-statistics value 1515.66 but higher between_SS / total_SS = 87.8 %.



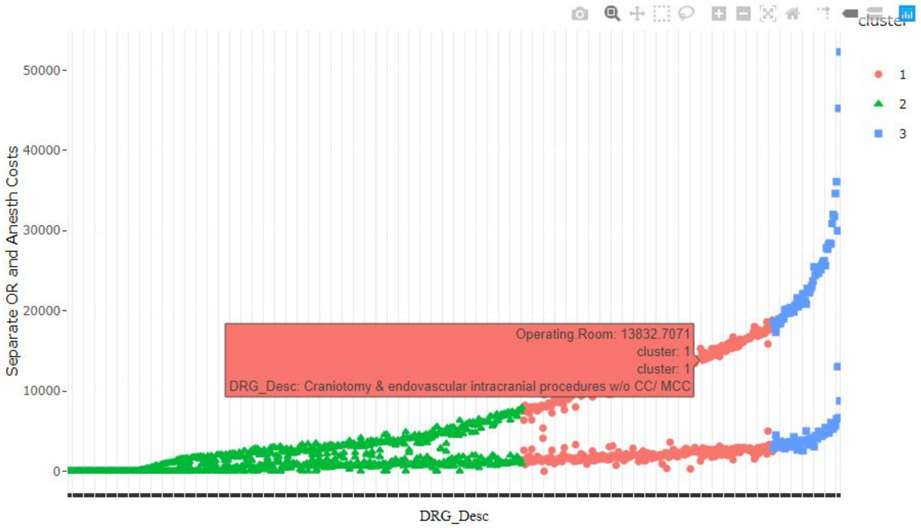
It seems like a good clustering too. Thus we evaluated the cluster by Silhouette Index and obtained the plot as follows. It surprised us that there is width less than 0 which means misclassified and is merely placed somewhere in between the clusters. Thus this method is not the best.



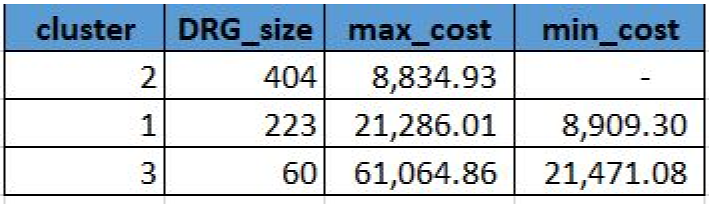
Finally, we come to choose the method 1 for our clustering and plot the cluster for DRG using interactive ggplotly as follows.



We also investigated the data by splitting back to the operating room and Anesthesiology, we can see that the cost of operating room is generally higher than Anesthesiology. Moreover, the cluster seems to work well for both indicators.



Summary of cluster:



cluster	DRG_size	max_cost	min_cost
2	404	8,834.93	-
1	223	21,286.01	8,909.30
3	60	61,064.86	21,471.08

3. Cluster Interpretation

Cluster 1:

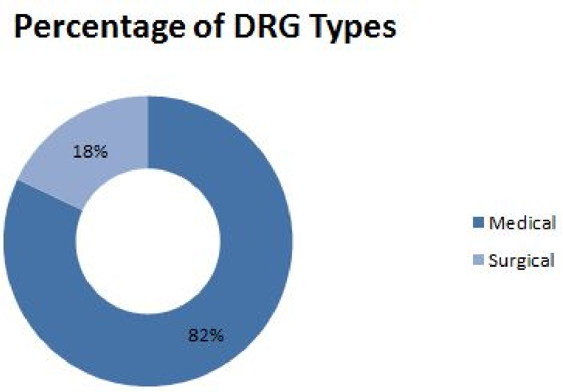
Firstly, we focused on the “Mid-cost Cluster”, and figured out some common features as below:

1. Most DRG in this cluster is surgical
   Surgical procedure means the treatment requires operating room and anesthesia, that is, the cost of OR and anesthesia can be generated.
2. Most DRG is related to the surgery about certain organ
   As we sorted and filtered, the most common organs related to this cluster are muscular system, reproductive system, and integumentary system, which means the surgery may not be as complicated as those on brain and heart. So, the time length of surgical operation is not so long as the most critical surgery. In addition, both general and local anesthesia are common among such surgery, given the surgery time is not super long, we believe that the cost of anesthesia is moderate.

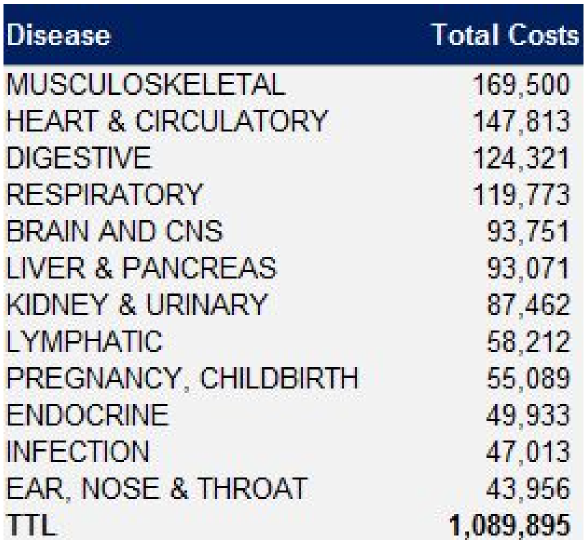
Therefore, the “Mid-cost Cluster” majorly contains the DRG of mid-level complicated surgical procedure about human body organs other than the brain and heart.

Cluster 2:

Through below charts and information search in the website, we find Cluster 2 mainly concentrated in Medical treatment:

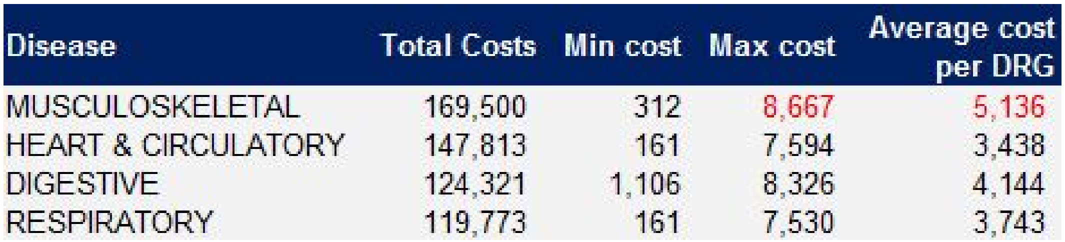


12 out 24 MDCs contributed 80% of the cluster’s total cost



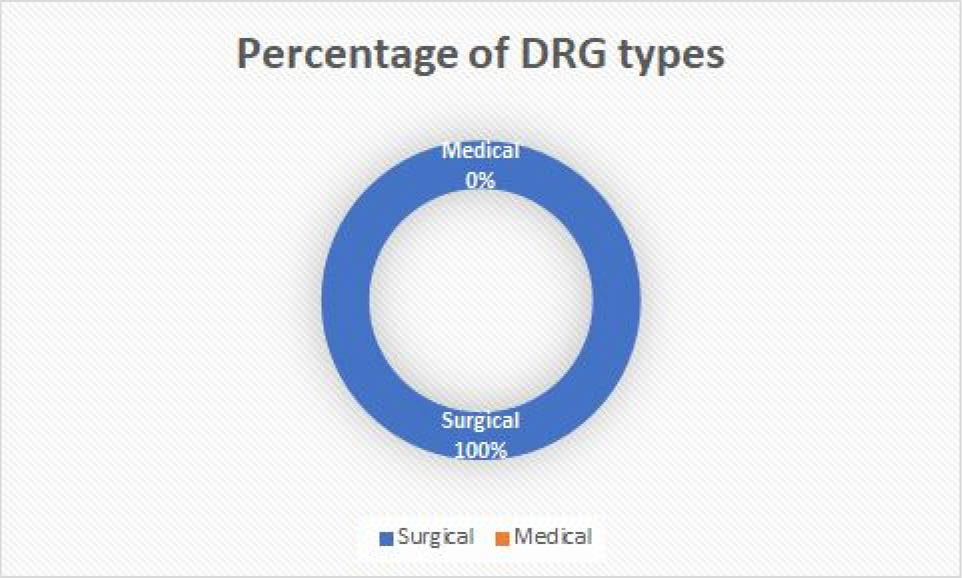
Disease	Total Costs
MUSCULOSKELETAL	169,500
HEART & CIRCULATORY	147,813
DIGESTIVE	124,321
RESPIRATORY	119,773
BRAIN AND CNS	93,751
LIVER & PANCREAS	93,071
KIDNEY & URINARY	87,462
LYMPHATIC	58,212
PREGNANCY, CHILDBIRTH	55,089
ENDOCRINE	49,933
INFECTION	47,013
EAR, NOSE & THROAT	43,956
TTL	**1,089,895**

4 MDCs costs above $100K, and are top Primary Care visits, if serious, then proposed to surgical hospitals

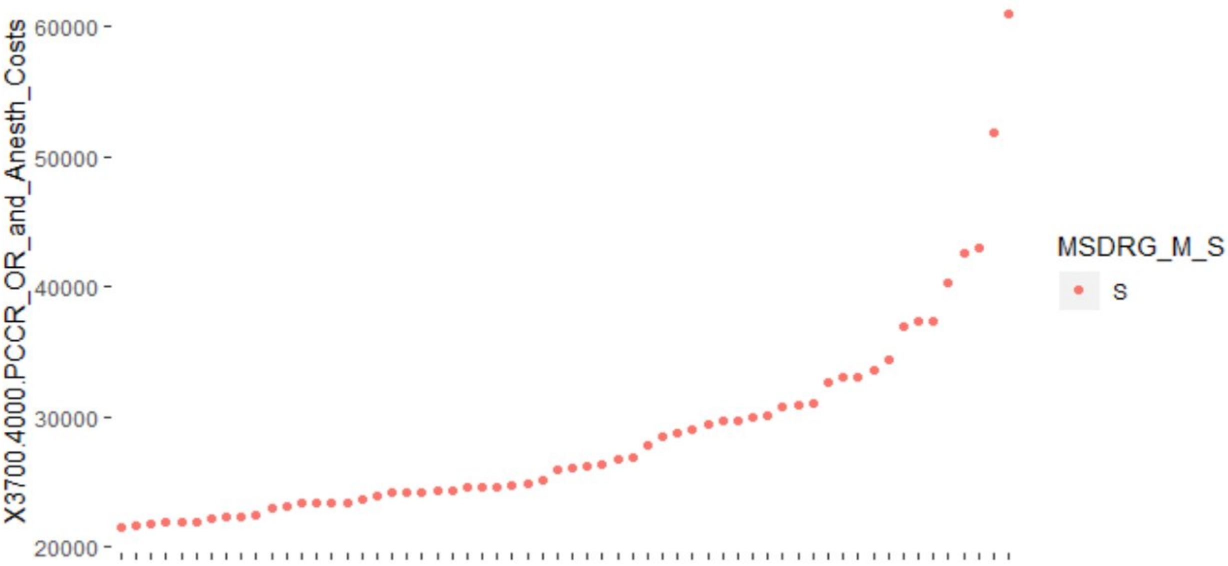


Disease	Total Costs	Min cost	Max cost	Average cost per DRG
MUSCULOSKELETAL	169,500	312	8,667	5,136
HEART & CIRCULATORY	147,813	161	7,594	3,438
DIGESTIVE	124,321	1,106	8,326	4,144
RESPIRATORY	119,773	161	7,530	3,743

Cluster 3:

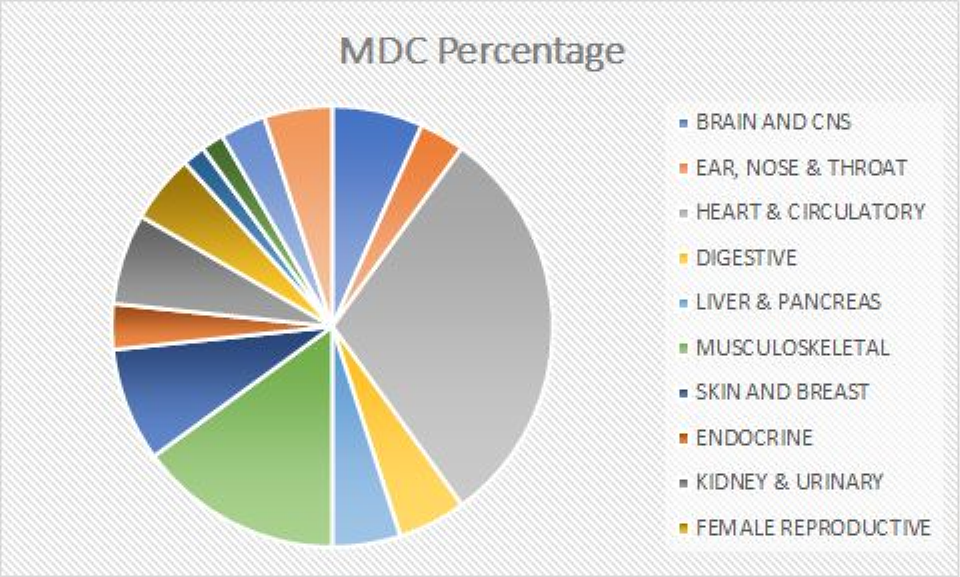


From the chart, we notice that all elements in this cluster are under the surgical category. This makes sense that surgical treatments are often more expensive than medical treatments and average cost in cluster 3 is $28456, which is the biggest among three clusters.



The cost in this category ranges from 20000 to 60000, and most of them concentrate on 20000 to 30000, which means that there are very few outliers(very expensive ones) in this cluster.

To further explore this cluster, we try to use MDC to dictate the category of each DRG.



From this pie chart, we are aware of the fact that heart & Circulatory and musculoskeletal treatments take up a major part of the whole cluster 3, which is larger than ½. It can be inferred that these two categories are the most difficult treatments so that doctors must put much more effort and time into these two kinds of diseases.

Some research:

- Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2015 were in men.
- About 630,000 Americans die from heart disease each year—that's 1 in every 4 deaths.
- In the United States, someone has a heart attack every 40 seconds. Each minute, more than one person in the United States dies from a heart disease-related event.
- Heart disease costs the United States about $200 billion each year.1 This total includes the cost of healthcare services, medications, and lost productivity.

Based on the information above, we are able to realize that heart disease has a high death rate in America, and it is hard to cure this kind of diseases, which means that it is reasonable that the treatment of heart disease should be comparatively expensive.

Appendix

Q1

Step 1

```
library(readxl)
library(tidyverse)
OUTP16 <- read.csv("E:/healthcare/assignment 4/VTOUTP16.TXT")
chronic_RA <- read_excel("E:/healthcare/assignment 4/RA_ICD10_Codes.xlsx",sheet
=1)
otherRA_systemic_involvement <-read_excel("E:/healthcare/assignment
4/RA_ICD10_Codes.xlsx",sheet =2)

#select one cohort for the first tab in the Excel file (chronic RA)
pt_chronic_RA<- OUTP16%>% filter(OUTP16[,10] %in% chronic_RA$`ICD-10 Codes`)
pt_chronic_RA<- cbind(pt_chronic_RA, pt_chronic_RA[,10])
colnames(pt_chronic_RA)[71] <- "ICD-10 codes"
for (i in c(11:29)) {
  pt<- OUTP16%>% filter(OUTP16[,i] %in% chronic_RA$`ICD-10 Codes`)
  pt<- cbind(pt, pt[,i])
  colnames(pt)[71] <- "ICD-10 codes"
  pt_chronic_RA<- rbind(pt_chronic_RA, pt)
}

#select unique patient in case of duplicate patients in table'pt_chronic_RA'
pt_chronic_RA<-pt_chronic_RA[!duplicated(pt_chronic_RA$Uniq), ]



#select second cohort based on the Excel file tab “other RA with systemic involvement”
pt_other_RA<- OUTP16%>% filter(OUTP16[,10] %in%
otherRA_systemic_involvement$`ICD-10 Codes`)
pt_other_RA<- cbind(pt_other_RA, pt_other_RA[,10])
colnames(pt_other_RA)[71] <- "ICD-10 codes"
for (i in c(11:29)) {
  pt<- OUTP16%>% filter(OUTP16[,i] %in% otherRA_systemic_involvement$`ICD-10
Codes`)
  pt<- cbind(pt, pt[,i])
  colnames(pt)[71] <- "ICD-10 codes"
  pt_other_RA<- rbind(pt_other_RA, pt)
}

#select unique patient in case of duplicate patients in table'pt_other_RA'
pt_other_RA<-pt_other_RA[!duplicated(pt_chronic_RA$Uniq), ]
```

Step 2

```
#frequency of each of the ICD-10 codes for the cohort of chronic RA
fre_chronic_RA<-count(pt_chronic_RA, `ICD-10 codes`)
fre_chronic_RA[order(fre_chronic_RA$n,decreasing = TRUE),]

#frequency of each of the ICD-10 codes for the cohort of  “other RA with systemic
involvement”
fre_other_RA<-count(pt_other_RA, `ICD-10 codes`)
fre_other_RA[order(fre_other_RA$n,decreasing = TRUE),]
```

Step3:

```
dup<-rbind(OUTP16,pt_chronic_RA[1:70],pt_other_RA[1:70])
dup<-dup[!duplicated(dup,fromLast = FALSE) & !duplicated(dup,fromLast = TRUE),]
nonRA_male<-dup %>% filter(sex==1)%>%tally()
nonRA_female<-dup %>% filter(sex==2)%>%tally()
RA_male<-pt_chronic_RA %>% filter(sex==1)%>% tally()+pt_other_RA %>%
filter(sex==1)%>%tally()
RA_female<-pt_chronic_RA %>% filter(sex==2)%>% tally()+pt_other_RA %>%
filter(sex==2)%>%tally()
Test <-
  matrix(c(nonRA_female$n,RA_female$n,nonRA_male$n, RA_male$n),
         nrow = 2,
         dimnames = list(c("NON RA", "RA"),
                         c("Female", "Male")))
fisher.test(Test)
```

Step4:

```
quantile(OUTP16$CHRGS,probs = c(0,0.25,0.5,0.75,1),na.rm = TRUE)
```

Step5:

```
total_RA<-rbind(pt_chronic_RA[1:70],pt_other_RA[1:70])
colnames(revcode)<-c("REVCODE","REVCODE_DESC")
rev<-merge(rev,revcode,by="REVCODE")
rev<-rev%>%select(REVCODE,Uniq,REVCHRGS,REVCODE_DESC)
RA_rev<-merge(total_RA,rev,by="Uniq")
fre_rev_RA<-count(RA_rev,REVCODE_DESC)
```

Q2

```
# MDC 1: By charge

SELECT
                HOSP_DESC,
                ROUND(a.hospital_total_charge/b.total, 4) AS
MDC1market_share_by_charge
        FROM
                (SELECT
                        HNUM2,
                        HOSP_DESC,
                        ROUND(SUM(CHRGS),2) AS hospital_total_charge
                FROM inpatient AS a, hospital AS b
                WHERE MDC = 1 AND a.HNUM2 = b.hnum
                GROUP BY HNUM2
      ) AS a,
                (SELECT
                        SUM(CHRGS) as TOTAL
                FROM inpatient
                WHERE MDC = 1
      ) AS b
        ORDER BY MDC1market_share_by_charge DESC
   INTO outfile '/Users/baixuhui/Desktop/q2.xlsx';

SELECT ROUND(SUM(POWER(MDC1market_share_by_charge,2)),2) AS
HHI_MDC1_by_charge FROM
(SELECT
                HOSP_DESC,
                ROUND(a.hospital_total_charge/b.total, 4) AS
MDC1market_share_by_charge
        FROM
                (SELECT
                        HNUM2,
                        HOSP_DESC,
                        ROUND(SUM(CHRGS),2) AS hospital_total_charge
                FROM inpatient AS a, hospital AS b
                WHERE MDC = 1 AND a.HNUM2 = b.hnum
                GROUP BY HNUM2
      ) AS a,
                (SELECT
            SUM(CHRGS) as TOTAL
        FROM inpatient
        WHERE MDC = 1
   ) AS b
    ORDER BY MDC1market_share_by_charge DESC) c;


# MDC1: By admission
SELECT
        HOSP_DESC,
        ROUND(a.hospital_total_admission/b.total, 4) AS
MDC1market_share_by_admisssion
    FROM
        (SELECT
            HNUM2,
            HOSP_DESC,
            COUNT(UNIQ) AS hospital_total_admission
        FROM inpatient AS a, hospital AS b
        WHERE MDC = 1 AND a.HNUM2 = b.hnum
        GROUP BY HNUM2
   ) AS a,
        (SELECT
            COUNT(UNIQ) as TOTAL
        FROM inpatient
        WHERE MDC = 1
   ) AS b
    ORDER BY MDC1market_share_by_admisssion DESC;

SELECT ROUND(SUM(POWER(MDC1market_share_by_admisssion,2)),2) AS
HHI_MDC1_by_admission FROM
(SELECT
        HOSP_DESC,
        ROUND(a.hospital_total_admission/b.total, 4) AS
MDC1market_share_by_admisssion
    FROM
        (SELECT
            HNUM2,
            HOSP_DESC,
            COUNT(UNIQ) AS hospital_total_admission
        FROM inpatient AS a, hospital AS b
        WHERE MDC = 1 AND a.HNUM2 = b.hnum
        GROUP BY HNUM2
   ) AS a,
        (SELECT
            COUNT(UNIQ) as TOTAL
        FROM inpatient
        WHERE MDC = 1
   ) AS b
    ORDER BY MDC1market_share_by_admisssion DESC)C;


# MDC14: By charge
SELECT
        HOSP_DESC,
        ROUND(a.hospital_total_charge/b.total, 4) AS
MDC14market_share_by_charge
    FROM
        (SELECT
            HNUM2,
            HOSP_DESC,
            ROUND(SUM(CHRGS),2) AS hospital_total_charge
        FROM inpatient AS a, hospital AS b
        WHERE MDC = 14 AND a.HNUM2 = b.hnum
        GROUP BY HNUM2
   ) AS a,
        (SELECT
            SUM(CHRGS) as TOTAL
        FROM inpatient
        WHERE MDC = 14
   ) AS b
    ORDER BY MDC14market_share_by_charge DESC;

SELECT ROUND(SUM(POWER(MDC14market_share_by_charge,2)),2) AS
HHI_MDC14_by_charge FROM
(SELECT
        HOSP_DESC,
        ROUND(a.hospital_total_charge/b.total, 4) AS
MDC14market_share_by_charge
```

```
    FROM
        (SELECT
            HNUM2,
            HOSP_DESC,
            ROUND(SUM(CHRGS),2) AS hospital_total_charge
        FROM inpatient AS a, hospital AS b
        WHERE MDC = 14 AND a.HNUM2 = b.hnum
        GROUP BY HNUM2
   ) AS a,
        (SELECT
            SUM(CHRGS) as TOTAL
        FROM inpatient
        WHERE MDC = 14
   ) AS b
    ORDER BY MDC14market_share_by_charge DESC)c;


# MDC14: By admission
SELECT
        HOSP_DESC,
        ROUND(a.hospital_total_admission/b.total, 4) AS
MDC14market_share_by_admission
FROM
        (SELECT
            HNUM2,
            HOSP_DESC,
            COUNT(UNIQ) AS hospital_total_admission
        FROM inpatient AS a, hospital AS b
        WHERE MDC = 14 AND a.HNUM2 = b.hnum
        GROUP BY HNUM2
   ) AS a,
        (SELECT
            COUNT(UNIQ) as TOTAL
        FROM inpatient
        WHERE MDC = 14
        ) AS b
ORDER BY MDC14market_share_by_admission DESC;
```

```
SELECT ROUND(SUM(POWER(MDC14market_share_by_admission,2)),2) AS
HHI_MDC14_by_admission FROM
(SELECT
        HOSP_DESC,
        ROUND(a.hospital_total_admission/b.total, 4) AS
MDC14market_share_by_admission
FROM
        (SELECT
            HNUM2,
            HOSP_DESC,
            COUNT(UNIQ) AS hospital_total_admission
        FROM inpatient AS a, hospital AS b
        WHERE MDC = 14 AND a.HNUM2 = b.hnum
        GROUP BY HNUM2
   ) AS a,
        (SELECT
            COUNT(UNIQ) as TOTAL
        FROM inpatient
        WHERE MDC = 14
        ) AS b
ORDER BY MDC14market_share_by_admission DESC)c;
```

Q3

```
library(dplyr)
library(sqldf)
library(reshape2)

VTREVCODE16 <- read.csv("~/Desktop/healthcare/03/VTREVCODE16.TXT")
VTINP16_upd <- read.csv("~/Desktop/healthcare/03/VTINP16_upd.TXT")
`FILE_LAYOUT_and_CODES_MSDRG2007+_20.977` <-
read.csv("~/Desktop/healthcare/03/FILE_LAYOUT_and_CODES_MSDRG2007+_20-97
7.csv", sep=";")
REVCODE_FILE_LAYOUT_and_CODES_PCCR <-
read.csv("~/Desktop/healthcare/03/REVCODE_FILE_LAYOUT_and_CODES_PCCR.cs
v", sep=";")

# Filter DRGs between 20 and 977
```

```
VTREVCODE16_new<-VTREVCODE16%>%select(Uniq,REVCODE,REVCHRGS,PCC
R)
DRG0 = subset(VTINP16_upd, VTINP16_upd$DRG >=20&VTINP16_upd$DRG
<=977,select=names(VTINP16_upd))

# Merge filtered DRG to the Revenue Code file on UNIQ
dt1<-merge(VTREVCODE16_new,DRG0, by.x = "Uniq", by.y = "UNIQ")
dt1_new<-dt1%>% select(Uniq,PCCR,REVCHRGS,DRG)

# Exclude the low dollar value services (less than $100)
dt2 = subset(dt1_new, dt1_new$REVCHRGS >=100,select=names(dt1_new))

# Sum all charges group by DRG, PCCR categories
PCCR_groupby <- dt2 %>% group_by(Uniq, DRG, PCCR) %>% summarise(revchrgs =
sum(REVCHRGS))
PCCR <- merge(PCCR_groupby, REVCODE_FILE_LAYOUT_and_CODES_PCCR, by
= "PCCR")
DRG<-merge(PCCR, `FILE_LAYOUT_and_CODES_MSDRG2007+_20.977`, by.x =
"DRG", by.y = "MSDRG")
dt3<-DRG%>% select(MSDRG_DESC,PCCR.NAME,revchrgs)

# Tabulate
tablulate <- dcast(dt3,MSDRG_DESC~PCCR.NAME,mean)

# Combining the PCCR 3700 Operating Room & PCCR 4000 Anesthesiology
tablulate$PCCR_OR_and_Anesth_Costs <- tablulate$`3700-Operating Room` +
tablulate$`4000-Anesthesiology`

rownames(tablulate) <- tablulate[,1]
tablulate <- tablulate[,-1]

# Turn NA to 0
tablulate[is.na(tablulate)] = 0
View(tablulate) #687 rows and 55 columns

setwd("/Users/tmh/desktop")
write.csv(tablulate,"PCCR_DRG.csv")
```

Cluster 3 analysis:

```
code=read.csv("FILE_LAYOUT_and_CODES.csv")
mdc<-read.csv("DRG_MDC_cluster_V2.csv")
colnames(code)[1]<-"MSDRG"
total=merge(mdc,code,by="MSDRG")
new=total%>%filter(cluster==3)%>%select(MSDRG,MSDRG_DESC,MSDRG_M_S,X37
00.4000.PCCR_OR_and_Anesth_Costs,MDC_CAT_NAME)
new%>%filter(MSDRG_M_S=="M")%>%count()
new%>%filter(MSDRG_M_S=="S")%>%count()
new%>%filter(
MSDRG_M_S=="S")%>%summarise(mean=mean(X3700.4000.PCCR_OR_and_Anest
h_Costs))
for (i in c(1:195)){
  if(new$MSDRG_M_S[i]=="S "){
    new$MSDRG_M_S[i]="S"
  }
}
mean(new$X3700.4000.PCCR_OR_and_Anesth_Costs)

new$MSDRG_M_S<-as.character(new$MSDRG_M_S)
new$MSDRG_DESC <- factor(new$MSDRG_DESC,
                              levels =
new$MSDRG_DESC[order(new$X3700.4000.PCCR_OR_and_Anesth_Costs)])
new%>%ggplot(aes(x=MSDRG_DESC,y=X3700.4000.PCCR_OR_and_Anesth_Costs))
+geom_point(aes(color=MSDRG_M_S))+
  theme(panel.background = element_blank())
```

Q3 Cluster Exploration and Validation

```
 setwd("E:/MBA@Brandeis/Syllabus/193HS-256F Healthcare Data Analytics and Data
Mining/Final")
getwd()

library(dplyr)  # data manipulation
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
library(fpc)
library(amap)
library(clusterSim)
library(ggplot2)
library(plotly)

drg_pccr <- read.csv("PCCR_DRG_DX.csv",header = T,fileEncoding = 'UTF-8-BOM')
summary(drg_pccr)
mdc <- read.csv("DRG_MDC.csv", header = T)
#Normalize
nmp <- drg_pccr$X3700.4000.PCCR_OR_and_Anesth_Costs
nmp_m <- mean(nmp)
nmp_sd <- sd(nmp)

drg_z <- scale(nmp, center = nmp_m, scale = nmp_sd)
drg_z

#Calculate distance
dist_eu <- get_dist(drg_z, method = "euclidean")
fviz_dist(dist_eu, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))

# Scree plot
wss <- (nrow(drg_z) -1) * sum(apply(drg_z, 2, var))
# var
for (i in 2:10) wss[i] <- sum(kmeans(drg_z, centers = i)$withinss)
plot(1:10,wss,
     type ='b',
     xlab = "Number of Clusters",
     ylab = "within Cluster Sum of Squares")

# Choosing K ....
set.seed(200)
k <- list()
for (i in 1:10){
  k[[i]] <- kmeans(drg_z, i,nstart =15)
}

between_totss <- list()
for (i in 1:10) {
  between_totss[[i]] <- k[[i]]$betweenss/k[[i]]$totss
}

plot(1:10,
    between_totss,
    type = 'b',
    ylab = "Between ss/ Total SS",
    xlab = "Cluster #")

plot(drg_z, col = k[[2]]$cluster)
# K-means clustering
## k=2
#fviz_cluster(k[[2]], data = drg_z)
between_totss
index.G1(drg_z,k[[2]]$cluster,d=NULL,centrotypes="centroids")

f_stat <- list()
for (i in 2:5) {
  f_stat[i] <-  round(calinhara(drg_z,k[[i]]$cluster),digits=2) ## f-stat
}

f_stat_cluster <- cbind(cluster = 2:5, f_value = f_stat[2:5])
f_stat_cluster <- as.data.frame(f_stat_cluster)
f_stat_cluster$cluster <- as.character(f_stat_cluster$cluster)

f_stat_cluster %>% ggplot(aes(x= cluster, y = f_value)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = f_value))

## k=3
col_pl <- c('green','blue','red')
plot(drg_z, col = k[[3]]$cluster)
legend("right",
      legend = levels(as.factor(k[[3]]$cluster)),
      fill = col_pl)

drg.cluster <- cbind(drg_pccr[,c(1:2,57)],cluster = k[[3]]$cluster)

drg.cluster <- cbind(drg.cluster,Operating.Room=drg_pccr$X3700.Operating.Room,
                 Anesthesiology= drg_pccr$X4000.Anesthesiology)
drg.cluster$DRG_Desc <- factor(drg.cluster$DRG_Desc,
                 levels =
drg.cluster$DRG_Desc[order(drg.cluster$X3700.4000.PCCR_OR_and_Anesth_Costs)])
drg.cluster$cluster <- as.character(drg.cluster$cluster)
q <- drg.cluster %>% ggplot(aes(x= DRG_Desc, y
=X3700.4000.PCCR_OR_and_Anesth_Costs)) +
  geom_point(mapping = aes(shape = cluster,col = cluster)) +
  scale_y_continuous(name = 'PCCR OR and Anesth Costs', breaks =
seq(0,65000,10000)) +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.title.x = element_text("MSDRG_DESC"))
ggplotly(q)

Anesth <- drg.cluster %>% ggplot(aes(x= DRG_Desc )) +
  geom_point(mapping = aes(y =Anesthesiology,shape = cluster,col = cluster)) +
  geom_point(aes(y=Operating.Room, shape = cluster,col = cluster))+
  scale_y_continuous(name = 'Separate OR and Anesth Costs', breaks =
seq(0,65000,10000)) +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.title.x = element_text("MSDRG_DESC"))
ggplotly(Anesth)

drg_mdc.cluster <- merge(drg.cluster,mdc, by.x = "MSDRG", by.y = "DRG" )

DRG_mdc_k <- drg_mdc.cluster %>% group_by(cluster, MDC,MDC_CAT_NAME)
%>% summarise(n = n())
drg_mdc.cluster %>% group_by(cluster, MDC,MDC_CAT_NAME) %>% summarise(n =
n()) %>%
  ggplot(aes(x = MDC_CAT_NAME, y = n, fill = cluster)) +
  geom_bar(stat = "identity", ) +
  theme(legend.position="right",
        axis.text.x = element_text(angle = 90, hjust = 1))

## summary of cluster
smry_cluster <- drg_mdc.cluster %>% group_by(cluster) %>% summarise(DRG_size
=n(),
                                    max_cost =
max(X3700.4000.PCCR_OR_and_Anesth_Costs),
                                    min_cost =
min(X3700.4000.PCCR_OR_and_Anesth_Costs)) %>%
  arrange(max_cost)

# SOM
library(kohonen)
set.seed(222)

drg_g <- somgrid(xdim = 4, ydim = 4, topo = "rectangular")
map <- som(drg_z,
         grid = drg_g,
         alpha = c(0.05,0.01),
         radius = 1)
plot(map,
    type = 'codes',
    palette.name = rainbow,
    main = "4 by 4  Mapping of DRG")

#### option 2
set.seed(100)
#kmeans
clust <- kmeans(map$codes[[1]], 3)

round(calinhara(drg_z,clust$cluster[map$unit.classif]),digits=2)
plot(map, type = "codes",
    palette.name = rainbow,
    bgcol = col_pl[clust$cluster],
    main = "SOM Cluster Map"
)
add.cluster.boundaries(map, clust$cluster)
legend("right",
      legend = levels(as.factor(clust$cluster)),
      fill = col_pl)
drg_som <-data.frame(drg_pccr[,c(1:2,57)], cluster = clust$cluster[map$unit.classif])

drg_som$DRG_Desc <- factor(drg_som$DRG_Desc,
                          levels =
drg_som$DRG_Desc[order(drg_som$X3700.4000.PCCR_OR_and_Anesth_Costs)])
drg_som$cluster <- as.character(drg_som$cluster)
p <- drg_som %>% ggplot(aes(x= DRG_Desc, y
=X3700.4000.PCCR_OR_and_Anesth_Costs)) +
  geom_point(aes(shape = cluster,color = cluster))+
  scale_y_continuous(name = 'PCCR_OR_and_Anesth_Costs', breaks =
seq(0,65000,10000)) +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank(),
        axis.title.x = element_text("MSDRG_DESC"),
        legend.position = "left")

ggplotly(p)
sil <- silhouette(k[[3]]$cluster, dist(drg_z))
fviz_silhouette(sil)

sil2 <- silhouette(clust$cluster[map$unit.classif], dist(drg_z))
fviz_silhouette(sil2)
```