IT599E Research Project

Chiranjibi Ghimire

5/03/2019

Data Description and Introduction

This data set is derived from the Vermont Uniform Hospital Discharge Data Set for 2017. The files include hospital records for inpatient, outpatient, and emergency department patient discharges. For my research purpose, I only use inpatient data set. These are limited use/de-identified data sets containing a set of fields common in discharge data files, such as codes for hospitals, diagnosis and procedures, age group, sex, hospital service area, and non-professional charges. The files are formatted as comma delimited text files.

I choose inpatient data set for the research purpose, which contain 53686 rows and 79 columns. For the simplicity, I choose only 14 columns which are more interesting for the analysis.

Below are the questions I will try to answer using some data analysis techniques.

1. What is the highest and lowest cost DRG (Diagnostic Related Group)?
2. What primary procedure is performed the most?
3. What is the average hospital stay (in days) for inpatients?
4. What is the most common primary diagnosis?
5. What is the common principal payment source to pay the medical expenses?

Data Preprocessing and Cleaning

First, I Loaded the Inpatient data set and check if there are any missing values such as NAs, and empty cells. For simplicity, I converted the original text file into csv format.

Loading the data set

```
mydata_Inpat=read.csv("/Users/chiranjibighimire/Desktop/Spring\
2019_Courses/IT\ 599/Research\ Project/Assignment\ Files/Inpatient\
Data.csv")

dim(mydata_Inpat)

## [1] 53686    79
```

This data set contain 53686 rows and 79 columns.

Since there are so many columns, I decided to select only 14 interesting columns.

```
newdata_Inpat <- mydata_Inpat[c(1,2,4,6,8,9,10,30,54,55,56,61,76,77)]
head(newdata_Inpat, n=5)
```

```
##   hnum2 ATYPE intage sex PPAY    CHRGS   DX1     PX1 pdays ccsdx ccsdxgrp
## 1     1     4      1   2    6  2337.90 Z3800 3E0234Z     2   218       15
## 2     1     1     14   2    1  7146.68  J440    <NA>     3   127        8
## 3     1     1     11   1    7 12110.28  F449    <NA>     3   670        5
## 4     1     3      4   2    2  3705.35 O4202 10E0XZZ     1   191       11
## 5     1     4      1   1    2  2127.31 Z3800 0VTTXZZ     1   218       15
##   ccsproc DRG MDC
## 1      NA 794  15
## 2      NA 191   4
## 3      NA 880  19
## 4     137 775  14
## 5     115 794  15
```

The data contains so many blank cells and NAs. I tried to fix those missing values, but I was not successful. Initially, I wanted to fill up the empty cells by NAs and then replace those NAs by 0. I was able to fill up empty cells by NAs using excel, but when I load the data it displays <NA>instead of NA as shown in above. The R program could not recognize <NA> as a NA. When I try to replace all the NAs including <NAs>, it gives an error. I spent plenty of time, but I couldn't able to fix it. Finally, I decided to remove all the rows with NA and <NA>, which reduced the rows from 53686 to 23960.

```
cleaned_data<- na.omit(newdata_Inpat)
dim(cleaned_data)

## [1] 23960    14

head(cleaned_data, n=5)

##   hnum2 ATYPE intage sex PPAY    CHRGS    DX1     PX1 pdays ccsdx ccsdxgrp
## 4     1     3      4   2    2  3705.35  O4202 10E0XZZ     1   191       11
## 5     1     4      1   1    2  2127.31  Z3800 0VTTXZZ     1   218       15
## 7     1     3     11   2    5 29919.35  M4806 0SG00A1     2   205       13
## 8     1     3      4   2    7 11311.82 O34211 10D00Z1     3   189       11
## 9     1     3     14   1    1 37833.90  M1711 0SRC0J9     1   203       13
##   ccsproc DRG MDC
## 4     137 775  14
## 5     115 794  15
## 7     158 460   8
## 8     134 766  14
## 9     152 470   8
```

Now, I got clean data without any missing values and empty cells.

Data Visualization

I am using ggplot package to visualize the Inpatient data set. I will need to split the data later on to create a decision tree, so I also installed plyr package.

```
installed.packages("ggplot2")
```

```
library("ggplot2")
installed.packages('plyr')

library(plyr)
installed.packages("gridExtra")

library(gridExtra)
```

First, I want to visualize the data to get the basic idea of metadata.

```
p1=ggplot(data = cleaned_data, mapping = aes(x = DRG, y = CHRGS)) +
geom_point(alpha = 0.1, aes(color = intage))  + ggtitle("DRG Vs Charges Plot
Based on Age")

p2=ggplot(data = cleaned_data, mapping = aes(x = DRG, y = CHRGS)) +
geom_point(alpha = 0.1, aes(color = sex))  + ggtitle("DRG Vs Charges Plot
Based on Sex")

p3=ggplot(data = cleaned_data, mapping = aes(x = DRG, y = pdays)) +
geom_point(alpha = 0.1, aes(color = intage))  + ggtitle("DRG Vs pdays Plot
Based on Age")

p4=ggplot(data = cleaned_data, mapping = aes(x = DRG, y = pdays)) +
geom_point(alpha = 0.1, aes(color = sex))  + ggtitle("DRG Vs pdays Plot Based
on Sex")

grid.arrange(p1,p2,p3,p4, ncol = 2, top = "Visualization of the Inpatient
Data Set")
```

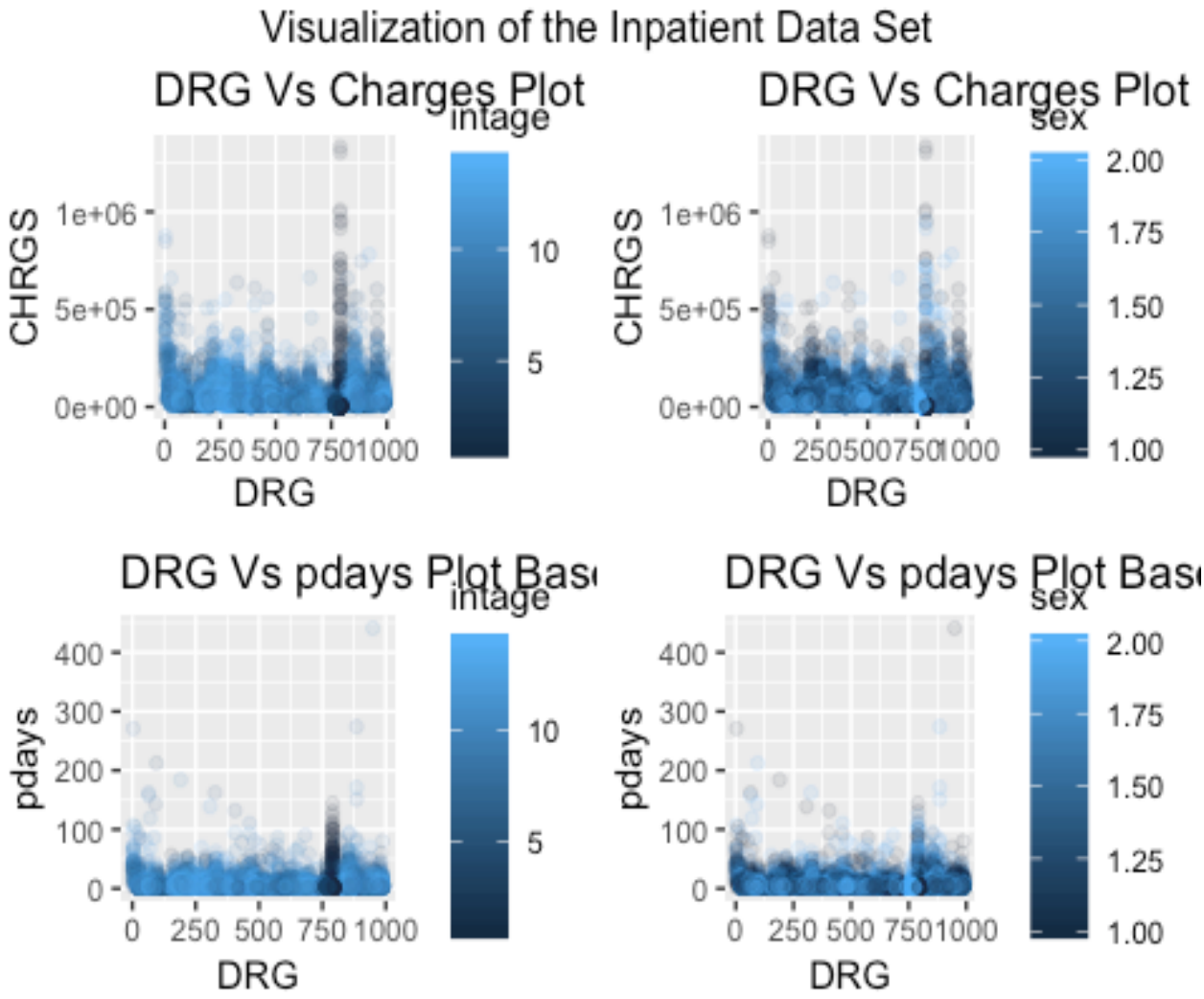

Fig 1: This plot describes the DRG Vs Charges and DRG Vs pdays based on age group and sex. The upper plots show that the DRG cost is high for age group less than 5 (<30-34 years) and most of them are male. The lower plots show that the patient, with most common DRG, stays longer (more than 100 days) in hospital with age group below 30-34 years and they are mostly male.

Also, the bar plots of Age group and Sex for the inpatient data set is shown in below.

```
barplot(table(cleaned_data$intage),col="red",ylab="Counts",las=2,main="Age
Group of Inpatient",cex.names = 0.8,cex.axis = 0.8)
```

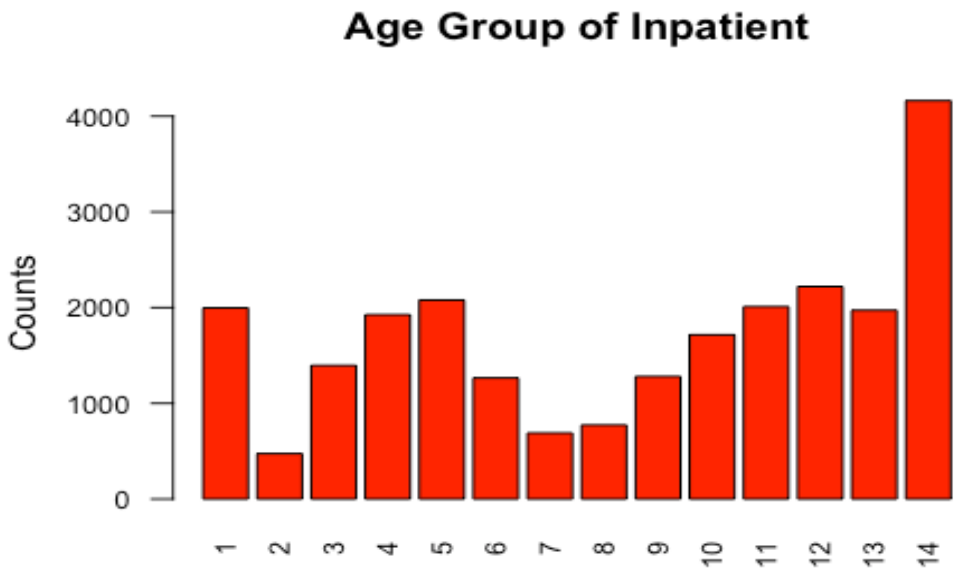

Fig 2: The bar plot shows that age group of 14 (Age of 75 or older) has majority in the inpatient data: there are more older people stay in the hospital while under treatment.

```
barplot(table(cleaned_data$sex),col="red",ylab="Counts",las=2,main=" Sex of
Inpatient",cex.names = 0.8,cex.axis = 0.8)+ geom_point(alpha = 0.1, aes(color
= sex))
```

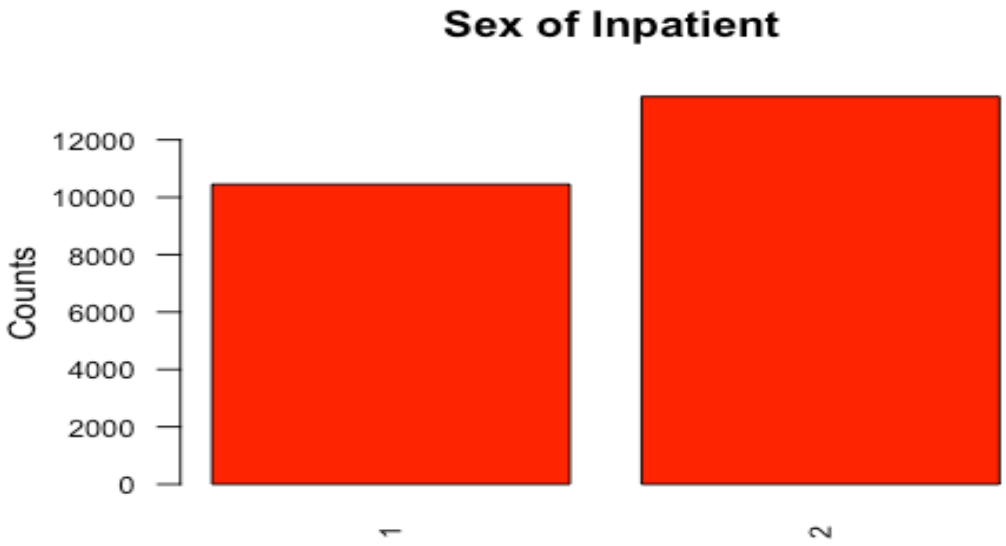

Fig 3: This bar plot illustrates that there is more female patient than male in the inpatient data set.

Now, I would like to answer some research questions that I mentioned above.

1. What is the highest and lowest cost DRG (Diagnostic Related Group)?

let's first make a bar plot of charges Vs DRG (Diagnostic Related Group).

```
plot1= ggplot(cleaned_data, aes(x =cleaned_data$DRG, y =
cleaned_data$CHRGS))+ geom_bar(aes(fill = cleaned_data$CHRGS), stat =
"identity" , color = "black", position = position_dodge(0.9))

plot1 + ggtitle("Plot of Charges by Diagnosis Related Group (DRG)") +
xlab("DRG") + ylab("Charges Amount (dollar)")
```

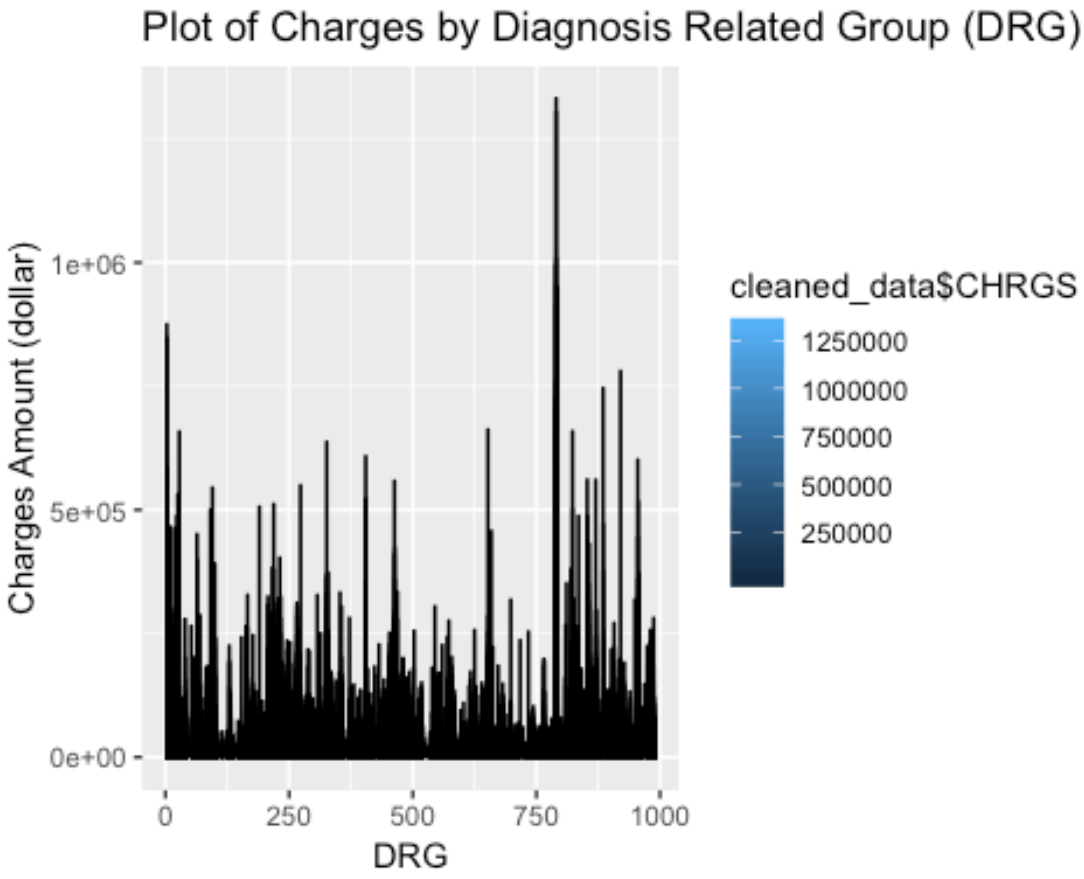

Fig 4: The above plot shows the Diagnostic Related Group (DRG) Vs charges amount (cost) of the patient in the hospital. It looks like there is one high peak after DRG 750 but couldn't tell what is the exact high cost DRG and lowest cost DRG. So, I need to find maximum and minimum DRG associated with Charges.

First, let's calculte minimum and maximum charges (cost) amount.

```
max(cleaned_data$CHRGS, na.rm=TRUE)
```

```
## [1] 1332080
```

```
min(cleaned_data$CHRGS, na.rm = TRUE)
```

```
## [1] 1272.35
```

The highest total cost for the patient is 1332080 and lowest cost is 1272.35 (in dollar).

Now, let's calculate highest and lowest cost DRG.

```
cleaned_data[which.max(cleaned_data[,6]),13]
```

```
## [1] 790
```

```
cleaned_data[which.min(cleaned_data[,6]),13]
```

```
## [1] 795
```

The highest cost DRG is 790: the DRG called "extreme immaturity or respiratory distress syndrome" cost more than other DRG. On the other hand, the lowest cost DRG is 795: the DRG called "Normal newborn" cost less than other DRG.

2. What primary procedure is performed the most?

Let's make a barplot of primary procedure to see how it looks like.

```
barplot(prop.table(table(cleaned_data$PX1)))
```

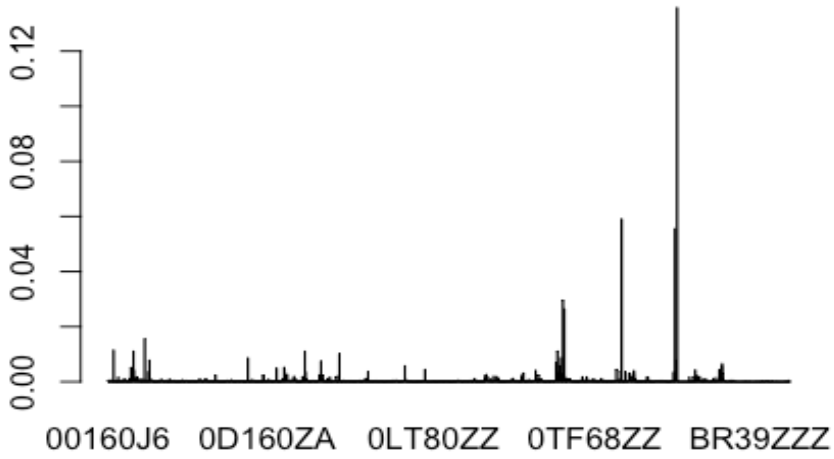

Fig 5: In the above barplot, there is a one high peak, but we need to calculate the frequency to know the most performed procedure.

```
sort(table(cleaned_data$PX1),decreasing=TRUE)[1]
```

```
## 10E0XZZ
##    3255
```

The primary procedure (PX1) code '10E0XZZ' performed the most with frequency of 3255. The primary procedure code '10E0XZZ' called "Delivery of Products of Conception" performed the most in inpatient which is done by External Approach. This procedure is intended for females as it is clinically and virtually impossible to be applicable to a male.

3. What is the average hospital stay in days for inpatients?

In order to know the average length of stay, we need to calculate mean (average) of the length of stay in days (pdays).

```
result.mean <- mean(cleaned_data$pdays)
print(result.mean)
```

```
## [1] 4.992821
```

The average stay for inpatients is 4.99 days, which is about 5 days.

4. What is the most common primary diagnosis?

```
sort(table(cleaned_data$DX1),decreasing=TRUE)[1]
```

```
## Z3800
##  1237
```

The most frequent primary diagnosis is Z3800 with frequency of 1237. The primary diagnosis 'Z3800' is a Single liveborn infant, delivered vaginally. The primary diagnosis "Single liveborn infant" is described as a product of a livebirth; an infant who shows evidence of life after birth; life is considered to be present after birth if any one of the following is observed: 1) the infant breathes; 2) the infant shows beating of the heart; 3) pulsation of the umbilical cord occurs; or 4) definite movement of voluntary muscles occurs.

5. What is the common principle payment source to pay the medical expenses?

In order to get an idea of principle payment source of inpatient, let's plot a histogram of principal payment source (PPAY).

```
ggplot(cleaned_data) + geom_histogram(aes(x = PPAY))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

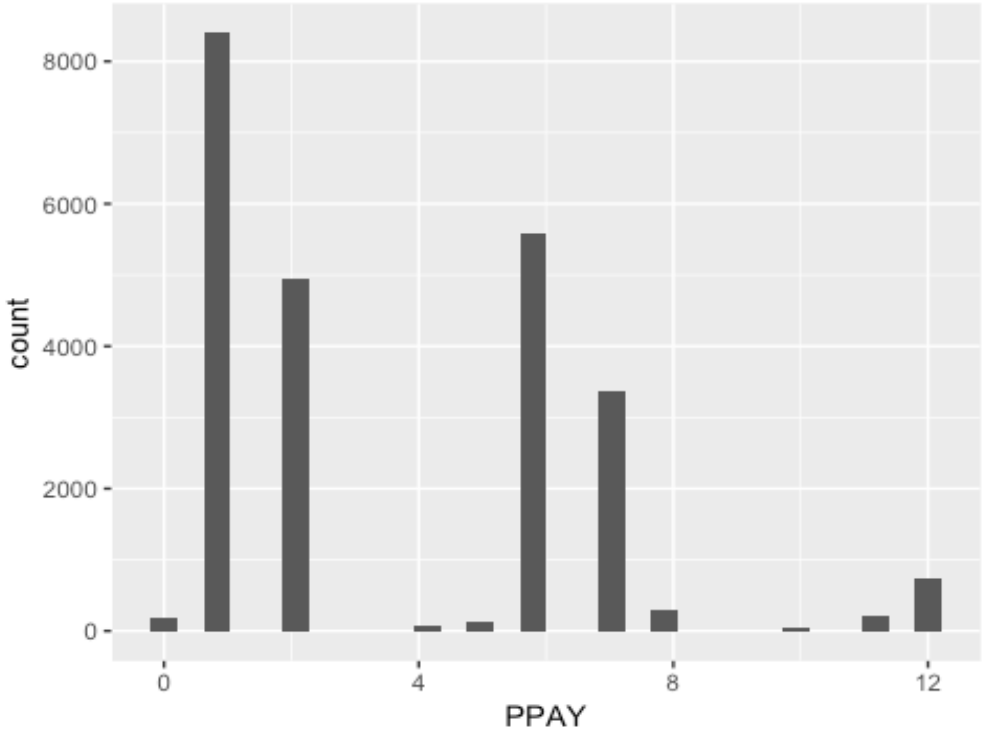

Fig 6: The histogram shows that the principal payment source '1' has a highest frequency. let's calculate exact frequency of that PPAY.

```
sort(table(cleaned_data$PPAY),decreasing=TRUE)[1]
```

```
##    1
## 8395
```

Form the calculation, the principal payment source (PPAY) 1 (Medicare) is the most common payment source with a frequency of 8395. That means Medicare is the most common principal payment source for a patient who stays in a hospital while under treatment: most of the inpatient's medical bills is covered by Medicare.

I would like to do more research on the data set to see if I can get more interesting results. I am applying a decision tree model to extract more interesting information from the Inpatient data set.

Decision Tree

I would like to split the data and to make a decision tree. Let's split the data with 80 % of training set and 20 % of testing (validate) data set.

```
library(rpart.plot)

## Loading required package: rpart

library(rpart)

set.seed(1234)
TrainingDataIndex= sample(2,nrow(cleaned_data), replace=TRUE, prob
=c(0.8,0.2))
train_1 = cleaned_data[TrainingDataIndex==1,]
validate= cleaned_data[TrainingDataIndex==2,]
nrow(train_1)

## [1] 19140

nrow(validate)

## [1] 4820
```

I got training set with 19140 rows and testing set with 4820 rows.

Let's create a Decision Tree using rpart. I would like to use DX1 as a response variable.

```
tree1 =rpart(DX1~intage+sex+pdays+CHRGS, data=train_1)
rpart.plot(tree1, extra=8)
```

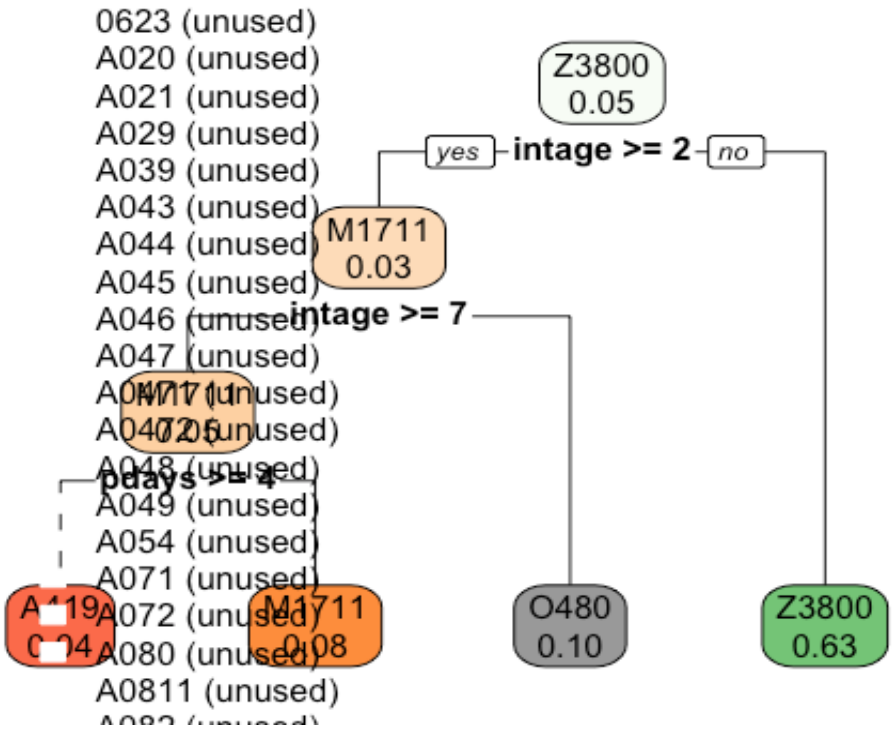

Fig 7: This is a decision tree for response variable DX1. There are lots of unused DX1 values. This tree looks messy, because there are lot of categorical variables that the tree can't capture.

Lastly, I would like to make another decision tree. The variable length of stay (pdays) is used as a response variable.

```
tree2 =rpart(pdays~ sex+intage+CHRGS+DRG, train_1)
rpart.plot(tree2)
```

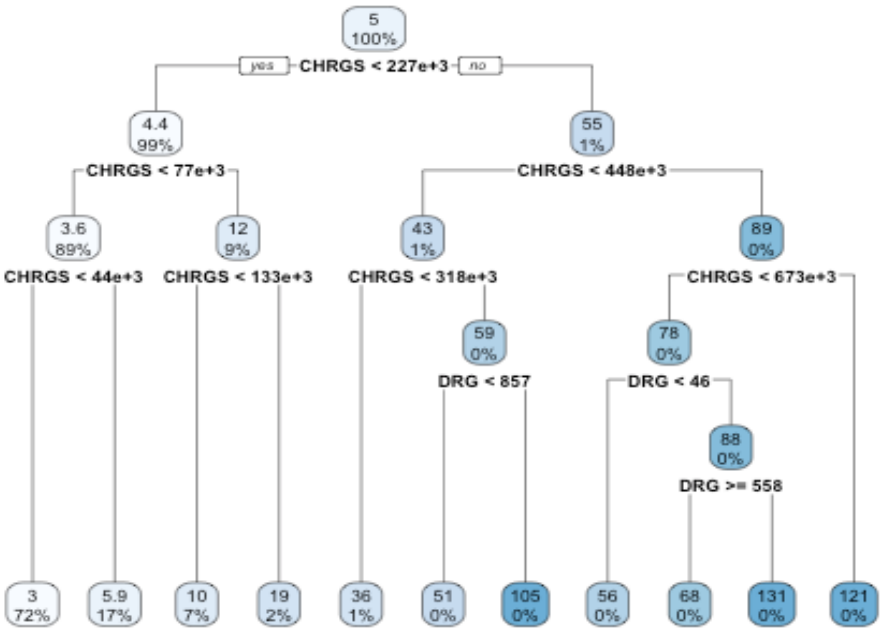

Fig 8: The decision tree demonstrates the predictive chart of length of stay of the patients with probabilities and percentage of patient in the leaf. For example, if the total cost is less than 227e+3 (227000), the patient is more likely to stay almost 5 days. On the other hand, if the charge is greater than 227000, the patient is more likely to stay 55 days and so on.

Conclusion

In summary, the given data set is very raw, and It needs more effort to get better findings, however I got some interesting results by utilizing my data analysis skill.

The patients, who stays in a hospital while under treatment, are mostly older (over 75 years old) and are more female in comparison to male. The DRG called “extreme immaturity or respiratory distress syndrome” cost more and DRG called “Normal newborn” cost less than other DRG. The most common primary diagnosis is ‘Single liveborn infant, delivered vaginally’.

The highest cost of diagnosis group is pretty high (1,332,080 USD), however most patients are covered under Medicare. The average length of stay in the hospital is about 5 days and the patient need to stay longer up to 441 days based on the nature of the diagnosis.

The patient, who belongs to DRG called ‘After care with CC/MCC’, discharged after 441 days of admission: which means a patient requires more resources; therefore, hospitals are paid more to care. The patient, who belongs to DRG called ‘Vaginal delivery w/o complicating diagnosis’, discharged on date of admission or on day after admission, which makes perfectly sense.