# Take-home assignment for: Data Analyst/Programmer - Thera Business Inc

Jiahao Ye

2023-02-02

## Task#1: Load the dataset in R and label the dataset as “test_thera”

```
# Load readxl package
library("readxl")
# Load the stringr package
library(stringr)

# Load the data and rename
test_thera = read_excel("D:/dataset/Research-Analyst_16SEPT2021.xlsx")
```

## Task#2: Present the names of variables

```
# present a list of variable names in the form
names(test_thera)
```

```
##  [1] "patient #"                  "days in the NICU"
##  [3] "age"                        "gender"
##  [5] "hospital code"              "type of surgery"
##  [7] "use of postoperative drain" "entry of paranasal sinus"
##  [9] "CSF leak"                   "duration of operation"
## [11] "diabetes mellitus"          "GCS"
## [13] "SSI"                        "discharge status"
## [15] "glucorticoids"              "lumbar drainage"
## [17] "income bracket"             "systolic bp preoperative"
```

The variables are shown in the above. There are in total 18 columns/variables

## Task#3: Find the number of rows and columns in this dataframe

```
number_of_rows = nrow(test_thera)
number_of_cols = ncol(test_thera)

str_glue("This dataframe has {number_of_rows} number of rows and {number_of_cols} number of columns.")
```

```
## This dataframe has 1079 number of rows and 18 number of columns.
```

## Task#4: Show the last 6 rows of “age”:

```
tail(test_thera[["age"]],6)
```

```
## [1] 21 43 39 37 45 62
```

The last 6 rows of column ‘age’ are: 21,43,39,37,45,62

## Task#5: Replace column name from “gender” to “sex”; replace last column to “blood pressure”

```
# replace gender
colnames(test_thera)[which(names(test_thera) == "gender")] = "sex"
# replace 'blood pressure'
colnames(test_thera)[which(names(test_thera) == "systolic bp preoperative")] = "blood pressure"
```

## Task#6: Replace the values in ‘income bracket’: 1-> <10,000….

```
# first, modify the variable type in column 'income bracket'
test_thera$`income bracket` = as.character(test_thera$`income bracket`)

# loop through the age column from top to bottom and replace value
for (i in 1:nrow(test_thera)) {

  if (test_thera[i,"income bracket"] == 1) {test_thera[i,"income bracket"] =  "<10,000"}
  else if (test_thera[i,"income bracket" ] == 2) {test_thera[i,"income bracket"] = "10,000 to 20,000"}
  else if (test_thera[i,"income bracket" ] == 3) {test_thera[i,"income bracket"] = "20,001 to 30,000"}
  else if (test_thera[i,"income bracket" ] == 4) {test_thera[i,"income bracket"] = "30,001 to 40,000"}
  else if (test_thera[i,"income bracket" ] == 5) {test_thera[i,"income bracket"] = ">40,001"}
}
```

Now the values in ‘income bracket’ has been replaced

## Task#7: Run the first 6 rows of “income bracket”

```
head(test_thera$`income bracket`,6)
```

```
## [1] ">40,001"          "10,000 to 20,000" ">40,001"          "10,000 to 20,000"
## [5] "<10,000"          "<10,000"
```

The first 6 values from income column are shown as above

## Task#8: Do a descriptive analysis on: “duration of operation” and “systolic bp preoperative”. Run an appropriate MICE procedure to impute for missing values

first check if there exists any missing values in these 2 columns

```
sum(is.na(test_thera$`duration of operation`))
```

```
## [1] 12
```

```
sum(is.na(test_thera$`blood pressure`))
```

```
## [1] 34
```

there are 12 missing values for ‘duration of opeartion’ and 34 missing values for ‘systolic bp preoperative’

use MICE procedure to impute for missing value: MICE stands for Multivariate Imputation by Chained Equation algorithm and it is an algorithm used to fill in the blanks. It simply uses values in other columns to predict the missing value

```
# first import the mice package
library(VIM)
```

```
## Warning: package 'VIM' was built under R version 4.2.2
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## VIM is ready to use.
```

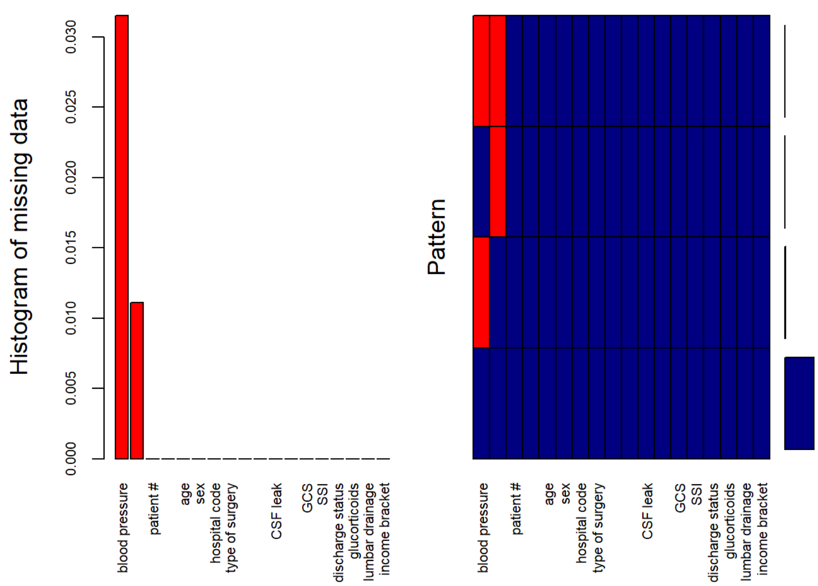
```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
# analyze the missing values:
aggr_plot <- aggr(test_thera, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(test_thera), cex.axis=.7, g
ap=3, ylab=c("Histogram of missing data","Pattern"))
```

```
## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```



```
##
##  Variables sorted by number of missings:
##                   Variable      Count
##             blood pressure 0.03151066
##      duration of operation 0.01112141
##                  patient # 0.00000000
##           days in the NICU 0.00000000
##                        age 0.00000000
##                        sex 0.00000000
##              hospital code 0.00000000
##            type of surgery 0.00000000
##  use of postoperative drain 0.00000000
##    entry of paranasal sinus 0.00000000
##                   CSF leak 0.00000000
##          diabetes mellitus 0.00000000
##                        GCS 0.00000000
##                        SSI 0.00000000
##           discharge status 0.00000000
##              glucorticoids 0.00000000
##            lumbar drainage 0.00000000
##             income bracket 0.00000000
```

As can be seen from the visualization above, only ‘duration of opeartion’ and ‘systolic bp preoperative’ has missing values. Missing values are of about 3.3% and 1.1% respectively.

After knowing the nature of the missing data, then the built-in mice() function should be used to compute the missing value

```
# there is an error when I try to run mice() on the test_thera dataframe; so I just create a deep copy(with different memory location) of the test_thera dataframe and run mice() on it.
test_thera_copy = data.frame(test_thera);

# turn the non-continuous variable into categorical variables:
names <- c("type.of.surgery","use.of.postoperative.drain", "entry.of.paranasal.sinus","CSF.leak","diabetes.mellitus" ,"SSI","discharge.status","glucorticoids","lumbar.drainage","income.bracket","hospital.code","sex")
test_thera_copy[,names] <- lapply(test_thera_copy[,names] , factor)
str(test_thera_copy)
```

```
## 'data.frame':    1079 obs. of  18 variables:
##  $ patient..                 : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ days.in.the.NICU          : num  3 7 7 7 9 2 6 4 9 5 ...
##  $ age                       : num  49 41 46 63 24 23 46 67 59 46 ...
##  $ sex                       : Factor w/ 2 levels "female","male": 2 2 1 2 1 1 2 1 2 1 ...
##  $ hospital.code             : Factor w/ 10 levels "1","2","3","4",..: 5 3 7 5 2 3 1 7 4 9 ...
##  $ type.of.surgery           : Factor w/ 4 levels "Burr hole operation",..: 1 1 1 3 4 3 1 1 1 1 ...
##  $ use.of.postoperative.drain: Factor w/ 2 levels "not used","used": 1 1 1 2 2 2 1 1 1 1 ...
##  $ entry.of.paranasal.sinus  : Factor w/ 2 levels "not present",..: 1 2 1 2 1 1 1 2 2 1 ...
##  $ CSF.leak                  : Factor w/ 2 levels "not present",..: 2 2 2 2 2 1 2 2 2 2 ...
##  $ duration.of.operation     : num  72 33 83 76 32 48 66 52 82 55 ...
##  $ diabetes.mellitus         : Factor w/ 2 levels "not present",..: 2 1 2 2 2 1 2 2 2 1 ...
##  $ GCS                       : num  15 2 8 5 7 15 4 14 13 2 ...
##  $ SSI                       : Factor w/ 2 levels "negative","postive": 2 1 1 2 1 1 1 2 2 1 ...
##  $ discharge.status          : Factor w/ 2 levels "alive","dead": 2 2 2 1 2 1 1 1 1 1 ...
##  $ glucorticoids             : Factor w/ 2 levels "not used","used": 2 2 2 2 2 1 1 2 2 1 ...
##  $ lumbar.drainage           : Factor w/ 2 levels "not used","used": 2 1 2 2 2 1 1 2 1 2 ...
##  $ income.bracket            : Factor w/ 5 levels "<10,000",">40,001",..: 2 3 2 3 1 1 2 4 2 3 ...
##  $ blood.pressure            : num  134 158 158 157 157 157 157 118 160 157 ...
```

It can be confirmed that variables such as gender, type of surgery etc has been converted into categorical variables

We can start the MICE imputation process now

```
library(mice);
```

```
## Warning: package 'mice' was built under R version 4.2.2
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
init = mice(test_thera_copy, maxit=0)
meth = init$method
predM = init$predictorMatrix

# remove the NA variable and not include that as a predictor
predM[, c("blood.pressure","duration.of.operation")]=0

# specify the method for imputation
meth[ c("blood.pressure","duration.of.operation")]="norm"
```

start the imputation process

```
set.seed(103)
imputed = mice(test_thera_copy, method=meth, predictorMatrix=predM, m=5)
```

```
##
##  iter imp variable
##   1   1  duration.of.operation  blood.pressure
##   1   2  duration.of.operation  blood.pressure
##   1   3  duration.of.operation  blood.pressure
##   1   4  duration.of.operation  blood.pressure
##   1   5  duration.of.operation  blood.pressure
##   2   1  duration.of.operation  blood.pressure
##   2   2  duration.of.operation  blood.pressure
##   2   3  duration.of.operation  blood.pressure
##   2   4  duration.of.operation  blood.pressure
##   2   5  duration.of.operation  blood.pressure
##   3   1  duration.of.operation  blood.pressure
##   3   2  duration.of.operation  blood.pressure
##   3   3  duration.of.operation  blood.pressure
##   3   4  duration.of.operation  blood.pressure
##   3   5  duration.of.operation  blood.pressure
##   4   1  duration.of.operation  blood.pressure
##   4   2  duration.of.operation  blood.pressure
##   4   3  duration.of.operation  blood.pressure
##   4   4  duration.of.operation  blood.pressure
##   4   5  duration.of.operation  blood.pressure
##   5   1  duration.of.operation  blood.pressure
##   5   2  duration.of.operation  blood.pressure
##   5   3  duration.of.operation  blood.pressure
##   5   4  duration.of.operation  blood.pressure
##   5   5  duration.of.operation  blood.pressure
```

```
imputed <- complete(imputed)

sapply(imputed, function(x) sum(is.na(x)))
```

```
##                  patient..           days.in.the.NICU
##                          0                          0
##                        age                        sex
##                          0                          0
##              hospital.code            type.of.surgery
##                          0                          0
## use.of.postoperative.drain   entry.of.paranasal.sinus
##                          0                          0
##                   CSF.leak      duration.of.operation
##                          0                          0
##          diabetes.mellitus                        GCS
##                          0                          0
##                        SSI           discharge.status
##                          0                          0
##              glucorticoids            lumbar.drainage
##                          0                          0
##             income.bracket             blood.pressure
##                          0                          0
```

It can be confirmed that all the missing values has been filled. The dataframe that I will be using for the next questions will be 'imputed'

Perform a descriptive analysis for: 'duration of operation' and 'systolic bp preoperative'

```
# use the summary function to find some descriptive data
print("-------duration of operation-----")
```

```
## [1] "-------duration of operation-----"
```

```
summary(imputed$duration.of.operation)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   9.752  44.000  59.000  58.056  72.000  83.000
```

```
print("-------blood pressure-----")
```

```
## [1] "-------blood pressure-----"
```

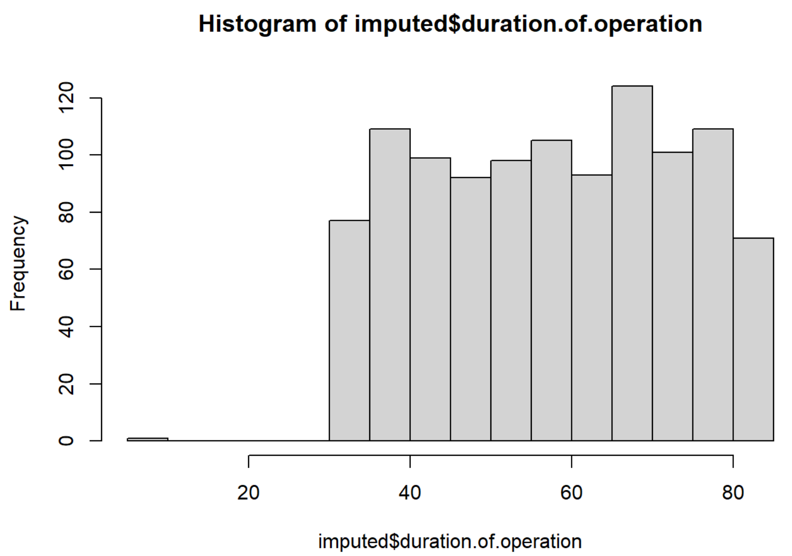
```
summary(imputed$blood.pressure)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   91.45  119.50  137.00  137.80  155.00  195.20
```

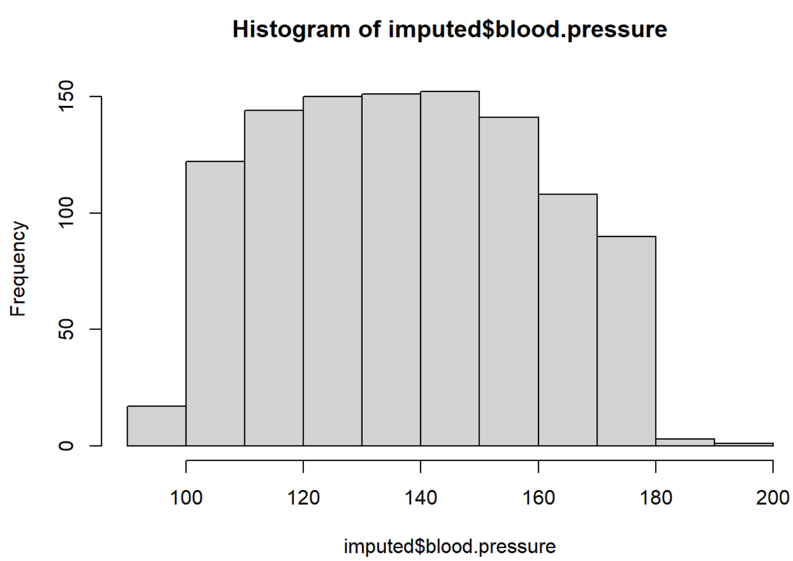
Draw histograms and boxplots: find out how the data is distributed

- Observation: not very close to a normal distribution, more like a uniform distribution

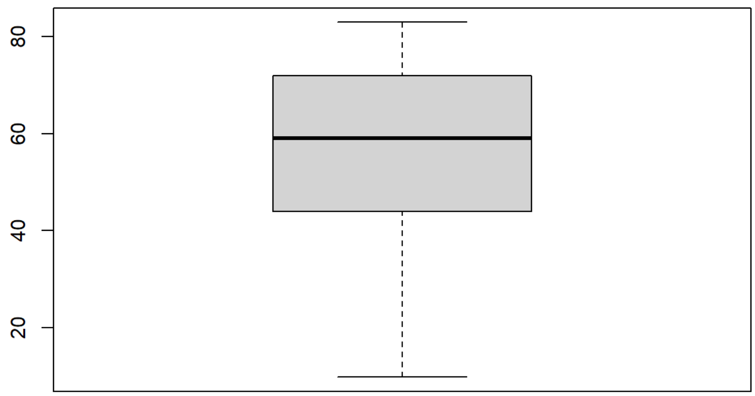
```
library('ggplot2')
# histogram:
hist(imputed$duration.of.operation)
```



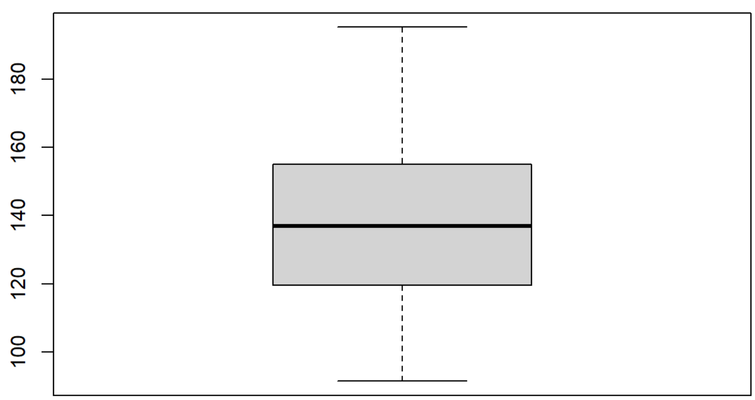

```
hist(imputed$blood.pressure)
```



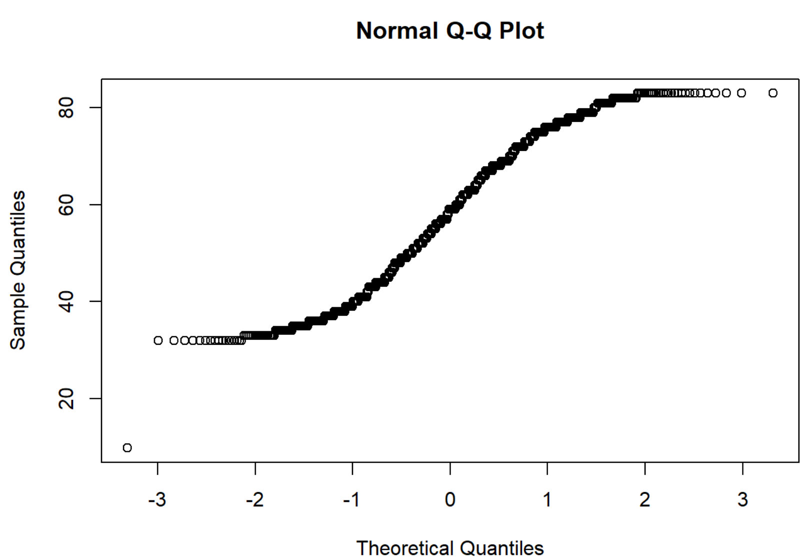

```
# boxplot:
boxplot(imputed$duration.of.operation)
```



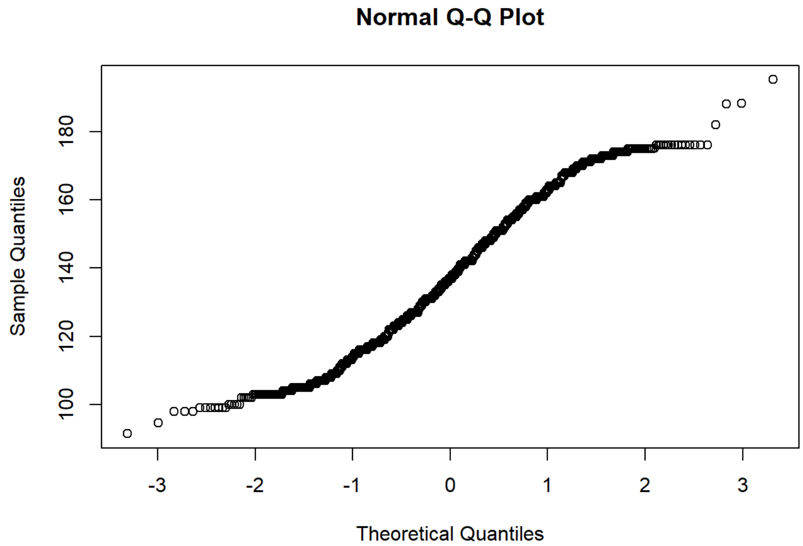

```
boxplot(imputed$blood.pressure)
```




```
# QQPlot
qqnorm(imputed$duration.of.operation)
```




```
qqnorm(imputed$blood.pressure)
```




## Task#9: Perform descriptive analysis on 5 other variables of your choice, and develop some graph

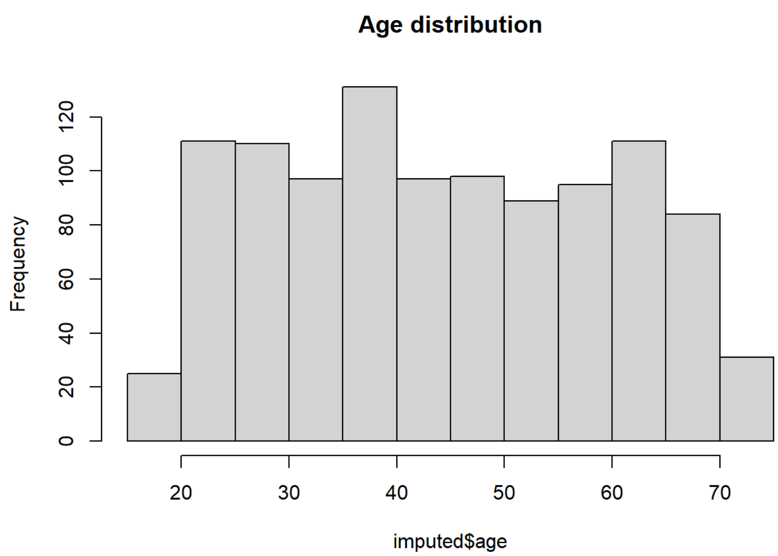
I am interested in seeing the following:

- age distribution
- income group distribution
- the discharge.status distribution across different income group
- numberofdays in ICU distirbution across different hispital
- how does blood pressure differs if lumbar.draingage exists or not

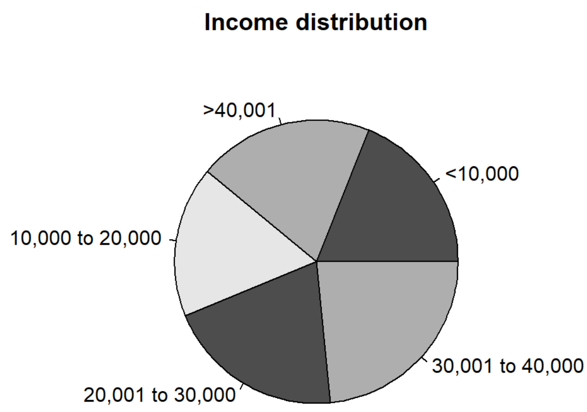
First find out the age and income group distribution

- observation: age range is mostly 20-70 years old; income distribution is quite even.

```
hist(imputed$age, main="Age distribution")
```



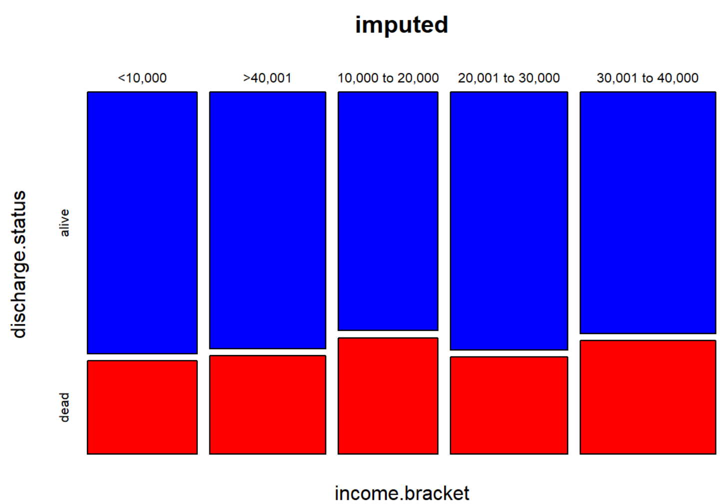

```
pie(table(imputed$income.bracket), col=grey.colors(3), main="Income distribution")
```




Find out the discharge.status distribution across different income group

- observation: there is no significant difference of alive-dead status across different income group

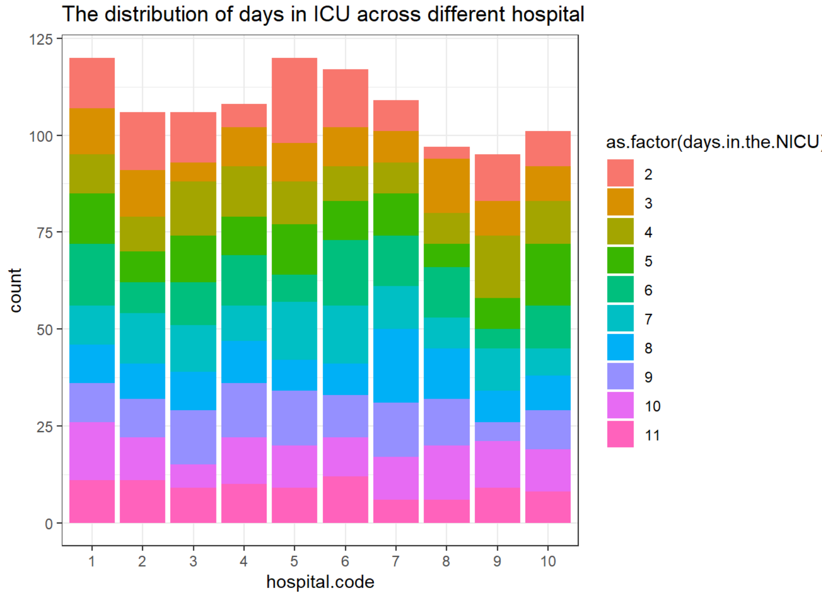
```
# I will simply use a mosaicplot to show
mosaicplot(income.bracket~discharge.status,data=imputed,col=c("Blue","Red"))
```




Find out numberofdays in ICU distirbution across different hispital

- observation: At hospital#5, the count of 2-day.in.ICU is a little higher compared to other hospitals. The occupations of ICUs in different hospitals are about the same.

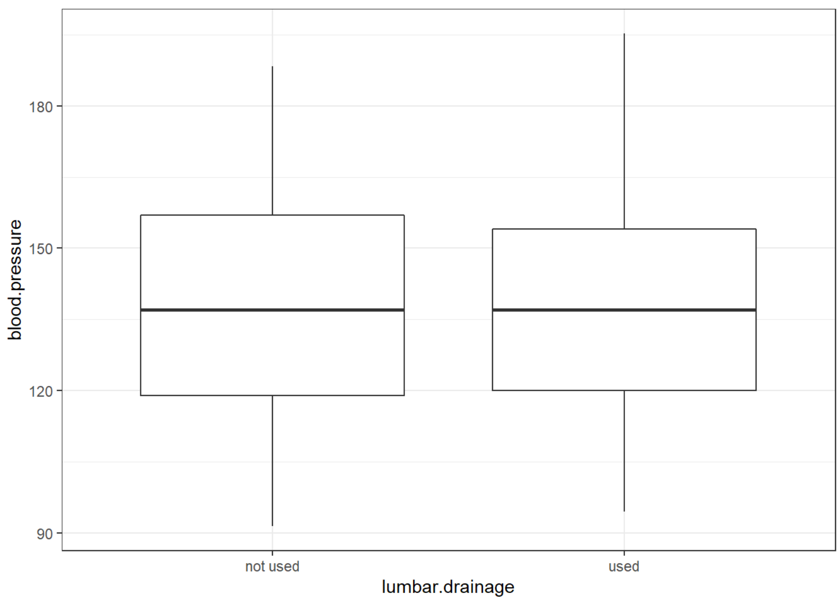
```
# mosaicplot(hospital.code~days.in.the.NICU,data=imputed,col=c("Blue","Red"))
ggplot(data = imputed)+geom_bar(aes(x=hospital.code, fill=as.factor(days.in.the.NICU))) + ggtitle(label="The distribution of days in ICU across different hospital")+ theme_bw()
```




Find lumbar.drainage vs blood pressure

- Looking from the graph, the use of lumbar drainage does not have a major impact on blood pressure

```
# draw 2 boxplots side by side
library(ggplot2)
ggplot(imputed, aes(lumbar.drainage, y=blood.pressure)) + geom_boxplot() + theme_bw()
```




Task#10: Low-risk versus High-risk

- 1~2 days: low-risk, more than 2 days: high-risk
- compare the characteristics btw patients who are “low-risk” compared to “high-risk”

First prepare the data by creating a new column ‘risk.level’, with categoraical variable:0 and 1

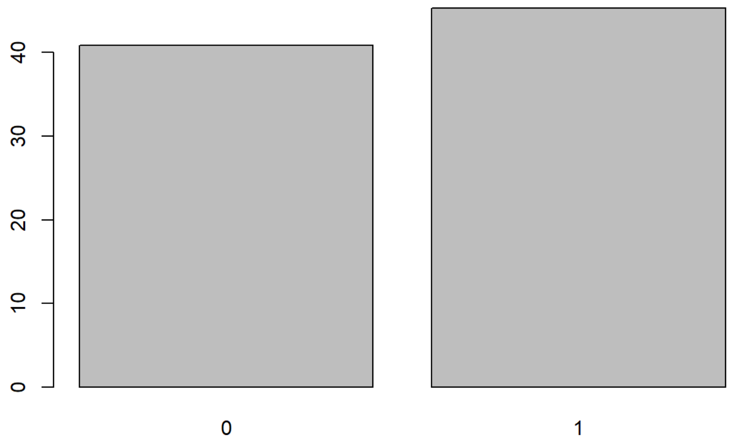
```
# create a new column called risk level; 0->'low-risk' 1->'high-risk'
imputed$'risk.level' = as.factor(ifelse(imputed$days.in.the.NICU <= 2, 0, 1))

# find some descriptive statistics about the risk-lvel column
print("---count---")
```

```
## [1] "---count---"
```

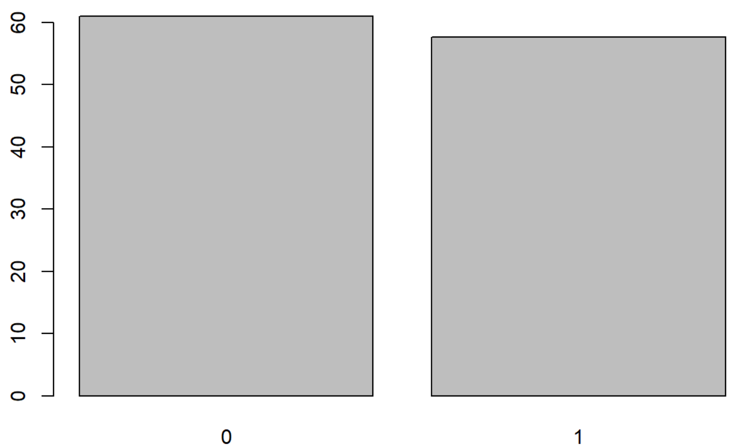
```
summary(imputed$risk.level)
```

```
##   0   1
## 116 963
```

```
# use the aggregate function to find the mean of continuous variable for each categorical variable
a <- aggregate(age ~ risk.level, data=imputed, FUN=mean)
barplot(a$age, names.arg=a$risk.level)
```




```
b <- aggregate(duration.of.operation ~ risk.level, data=imputed, FUN=mean)
barplot(b$duration.of.operation, names.arg=b$risk.level)
```




As can be seen from the 2 aggregate plot from above, the average age for high-risky group tends to be slightly higher; the average duration.of.operation for high-risky group tends to be lower

## Task#11: Esitimate the likelihood for being low-risk or high-risk

- Perform some univariate regression first: run univariate regression on: days.in.ICU ~ many other variable, and check the R value for statistical significance
- Then run a multivariate regression; add the categorical variables into the regression model as well

```
lm1 <- lm(days.in.the.NICU ~ age, data = imputed)
lm2 <- lm(days.in.the.NICU ~ duration.of.operation, data = imputed)
lm3 <- lm(days.in.the.NICU ~ blood.pressure, data = imputed)

# print the linear regression model result
summary(lm1)
```

```
##
## Call:
## lm(formula = days.in.the.NICU ~ age, data = imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7653 -2.4372 -0.1364  2.4534  4.8909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.835676   0.269820  21.628   <2e-16 ***
## age         0.013671   0.005702   2.398   0.0167 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.837 on 1077 degrees of freedom
## Multiple R-squared:  0.005309,   Adjusted R-squared:  0.004386
## F-statistic: 5.748 on 1 and 1077 DF,  p-value: 0.01667
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = days.in.the.NICU ~ duration.of.operation, data = imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5353 -2.4453 -0.3692  2.5426  4.6378
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.649503   0.342322  19.425   <2e-16 ***
## duration.of.operation -0.003461   0.005705  -0.607    0.544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.844 on 1077 degrees of freedom
## Multiple R-squared:  0.0003417,  Adjusted R-squared:  -0.0005865
## F-statistic: 0.3681 on 1 and 1077 DF,  p-value: 0.5442
```

```
summary(lm3)
```

```
##
## Call:
## lm(formula = days.in.the.NICU ~ blood.pressure, data = imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8851 -2.4509 -0.0624  2.4692  4.9491
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.874025   0.553957   8.799  < 2e-16 ***
## blood.pressure 0.011426   0.003971   2.877  0.00409 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 1077 degrees of freedom
## Multiple R-squared:  0.007629,   Adjusted R-squared:  0.006708
## F-statistic:  8.28 on 1 and 1077 DF,  p-value: 0.004088
```

As can be seen from the regression analysis above, the variable:duration.of.operation is statistically insignificant (p value > 0.1, not significant even at 10% level); the other 2 varibales: age & blood.pressure are significant, and I will incorporate these 2 variables into the multivariate regression model

Run a multivariate regression

```
multivariate.lm = lm(formula = days.in.the.NICU~age+blood.pressure+sex+glucorticoids+diabetes.mellitus+CSF.leak+SSI+use.of.p
ostoperative.drain+lumbar.drainage, data=imputed)

summary(multivariate.lm)
```

```
##
## Call:
## lm(formula = days.in.the.NICU ~ age + blood.pressure + sex +
##     glucorticoids + diabetes.mellitus + CSF.leak + SSI + use.of.postoperative.drain +
##     lumbar.drainage, data = imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6432 -2.4582 -0.0238  2.4430  5.6322
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.853413   0.631660   6.100 1.48e-09 ***
## age                            0.012902   0.005667   2.277  0.02299 *
## blood.pressure                 0.011333   0.004005   2.829  0.00475 **
## sexmale                        0.438485   0.172464   2.542  0.01115 *
## glucorticoidsused             -0.255485   0.171909  -1.486  0.13753
## diabetes.mellituspresent      -0.251992   0.176823  -1.425  0.15442
## CSF.leakpresent                0.184477   0.190620   0.968  0.33338
## SSIpostive                     0.374940   0.173338   2.163  0.03076 *
## use.of.postoperative.drainused  0.197571   0.198941   0.993  0.32088
## lumbar.drainageused            0.281471   0.173178   1.625  0.10439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.81 on 1069 degrees of freedom
## Multiple R-squared:  0.03149,    Adjusted R-squared:  0.02334
## F-statistic: 3.862 on 9 and 1069 DF,  p-value: 7.863e-05
```

As can be seen from the summary above, in the regression model, the following variables are statistically significant

- age - <0.001
- blood.pressure - 0.01
- male - 0.01
- SSI - 0.01

Incorporate some other categorical variables into the multivariate regression model:

```
multivariate.lm2 = lm(formula = days.in.the.NICU~age+blood.pressure+sex+SSI+hospital.code+type.of.surgery+income.bracket, da
ta=imputed)

summary(multivariate.lm2)
```

```
##
## Call:
## lm(formula = days.in.the.NICU ~ age + blood.pressure + sex +
##     SSI + hospital.code + type.of.surgery + income.bracket, data = imputed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6629 -2.3321  0.0547  2.3754  5.8320
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         3.861829   0.705858   5.471 5.58e-08 ***
## age                                 0.012928   0.005730   2.256  0.02427 *
## blood.pressure                      0.011783   0.003985   2.957  0.00318 **
## sexmale                             0.472632   0.175110   2.699  0.00706 **
## SSIpostive                          0.392434   0.172882   2.270  0.02341 *
## hospital.code2                     -0.084119   0.377243  -0.223  0.82359
## hospital.code3                     -0.014867   0.376756  -0.039  0.96853
## hospital.code4                      0.288690   0.375575   0.769  0.44227
## hospital.code5                     -0.390473   0.365072  -1.070  0.28505
## hospital.code6                     -0.055819   0.367579  -0.152  0.87933
## hospital.code7                      0.292832   0.375096   0.781  0.43516
## hospital.code8                      0.380088   0.386511   0.983  0.32565
## hospital.code9                     -0.175996   0.390074  -0.451  0.65195
## hospital.code10                    -0.116754   0.381458  -0.306  0.75961
## type.of.surgeryCraniotomy operation  0.314387   0.256392   1.226  0.22040
## type.of.surgeryShunt operation      0.038139   0.256576   0.149  0.88186
## type.of.surgerySpinal operation     0.082602   0.220567   0.374  0.70811
## income.bracket>40,001              -0.485599   0.277031  -1.753  0.07991 .
## income.bracket10,000 to 20,000     -0.142207   0.288509  -0.493  0.62218
## income.bracket20,001 to 30,000      0.012017   0.275662   0.044  0.96524
## income.bracket30,001 to 40,000     -0.059078   0.269360  -0.219  0.82644
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.818 on 1058 degrees of freedom
## Multiple R-squared:  0.03559,    Adjusted R-squared:  0.01736
## F-statistic: 1.952 on 20 and 1058 DF,  p-value: 0.007334
```

As can be seen from the summary above, there are a couple observations:

- the hospital that the patient is in have no impact on his/her number of days in ICU
- the type of surgery that the patient has have no impact on his/her number of days in ICU
- A higher income level (>40,001) has negative correlation to his/her number of days in ICU

Run a logistic regression(0 or 1) to find the Likelihood of being low-risk and high-risk

```r
## 'risk level' is a categorical variable with 0 and 1
multi_logit=glm(`risk.level`~age+sex+`diabetes.mellitus`+`CSF.leak`+`duration.of.operation`+`income.bracket`+`blood.pressure
`+`type.of.surgery`,data=imputed,family=binomial(link="logit"))
summary(multi_logit)
```

```
##
## Call:
## glm(formula = risk.level ~ age + sex + diabetes.mellitus + CSF.leak +
##     duration.of.operation + income.bracket + blood.pressure +
##     type.of.surgery, family = binomial(link = "logit"), data = imputed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6587   0.3173   0.4111   0.5235   0.9393
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -0.2016114  0.8727981  -0.231   0.8173
## age                                 0.0201406  0.0069389   2.903   0.0037 **
## sexmale                             0.5223534  0.2111800   2.473   0.0134 *
## diabetes.mellituspresent           -0.2740973  0.2119909  -1.293   0.1960
## CSF.leakpresent                     0.3120019  0.2174034   1.435   0.1513
## duration.of.operation              -0.0122093  0.0066842  -1.827   0.0678 .
## income.bracket>40,001               0.1222401  0.3333676   0.367   0.7139
## income.bracket10,000 to 20,000      0.2443661  0.3624029   0.674   0.5001
## income.bracket20,001 to 30,000     -0.0433057  0.3182858  -0.136   0.8918
## income.bracket30,001 to 40,000     -0.2320515  0.3064154  -0.757   0.4489
## blood.pressure                      0.0140955  0.0047821   2.948   0.0032 **
## type.of.surgeryCraniotomy operation  0.0004985  0.3020551   0.002   0.9987
## type.of.surgeryShunt operation      0.1323317  0.3107610   0.426   0.6702
## type.of.surgerySpinal operation    -0.0132789  0.2601419  -0.051   0.9593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 736.46  on 1078  degrees of freedom
## Residual deviance: 699.00  on 1065  degrees of freedom
## AIC: 727
##
## Number of Fisher Scoring iterations: 5
```

Observations:

- an older age && being a male && having higher blood pressure increase the chance of being ‘high-risk’
- Other variables does not provide any predictive power because they are statistically insignificant

# END OF DOCUMENT