Project 2

Huiyue Li, Yujia Wang, Lujun Zhang<- Huiyue Li

```
library(tidyverse)
library(knitr)
```

Import both datasets

```
# input the two datasets
encounter<- read_csv("Encounter Level Data.csv")
patient <- read_csv("Patient Level Data.csv")
```

Merge the patient level data into the encounter level data

```
# before merging the two datasets, we need to identify the key variable
 unique in these two tibbles
# look at the two datasets firstly
head(encounter)
```

```
## # A tibble: 6 x 7
##   MRN    contact_date enc_type      temp distress_score   WBC   BMI
##   <chr>  <chr>        <chr>        <dbl>          <dbl> <dbl> <dbl>
## 1 HJ9754 6/25/2016    Office visit  97.7              2 15.1   28.3
## 2 GE5166 8/7/2016     Office visit  97.8              2  6.86  38.2
## 3 XV9573 1/19/2018    Office visit  96.5              2  5.48  32.1
## 4 CQ9338 7/4/2015     Office visit  96.4              3 15.1   25.1
## 5 DH1301 3/24/2018    Office visit  97.4              3  3.4   33.4
## 6 WQ8508 8/24/2019    Office visit  96.4              1  5.04  21.3
```

```
head(patient)
```

```
## # A tibble: 6 x 8
##   MRN    DOB       race  financialclass ethnicity    hypertension CH
F   diabetes
##   <chr>  <chr>     <chr> <chr>          <chr>        <chr>        <c
hr> <chr>
## 1 DH1301 9/25/1971 Other Private        non-Hispanic N            N
    N
## 2 JV9469 4/28/1962 White Private        non-Hispanic Y            N
    N
## 3 TH8119 5/15/1942 White Medicare       non-Hispanic N            N
    N
## 4 TJ3799 9/7/1934  White Medicare       non-Hispanic Y            N
    Y
## 5 HP1319 4/30/1973 White Private        non-Hispanic Y            N
    N
```

```
## 6 KR5834 7/15/1982 White Private        non-Hispanic N            N
    N

# from the result, we can identify the variable MRN is the key variable
# merge the patient data into the encounter data using the key variable
encounter_patient<-merge(encounter,patient,by=c("MRN"))
```

Re-categorize WBC into a categorical variable

```
# categorize WBC
WBC1<-"Not Taken"
WBC1[encounter_patient$WBC<3.2]="Low"
WBC1[3.2<=encounter_patient$WBC & encounter_patient$WBC<=9.8]="Normal"
WBC1[9.8<encounter_patient$WBC]="High"
encounter_patient$WBC<-WBC1
# we still have NA in WBC, then turn the NA into "Not Taken"
# use which[] to identify the position of NA in WBC
NT=which(is.na(encounter_patient$WBC))
# replace the NA with "Not Taken"
encounter_patient$WBC[NT]="Not Taken"
```

print a table of the categorical WBC variable

```
# obtain the dataframes of the counts and percentages for each category
count<-data.frame(table(encounter_patient$WBC))
proportion<-data.frame(prop.table(table(encounter_patient$WBC))*100)
# combine dataframes and print the table
WBC_table<-merge(count,proportion,by=c("Var1"))
kable(WBC_table,col.names = c("WBC","Count","Percentage (%)"),caption =
 "Table1: for categorical WBC variable",digits = 3,align = "c")
```

Table1: for categorical WBC variable

WBC	Count	Percentage (%)
High	113	20.545
Low	169	30.727
Normal	196	35.636
Not Taken	72	13.091

Calculate & print a table of the mean BMI for the following MRNs: CI6950, IW9164, HJ8458, & XE4615

```
MRNs<-encounter_patient%>%
# filter the rows required
  filter(MRN=="CI6950"|MRN=="IW9164"|MRN=="HJ8458"|MRN=="XE4615")%>%
  group_by(MRN)%>%
```

```
  summarise(mean=mean(BMI,na.rm = T),.groups="drop_last")
# print the table
kable(MRNs,col.names = c("MRN","Mean of BMI"),caption = "Table 2: the m
ean of BMI for the MRN",digits = 3,align = "c")
```

Table 2: the mean of BMI for the MRN

MRN	Mean of BMI
CI6950	25.842
HJ8458	28.948
IW9164	29.435
XE4615	29.755

Create a table showing how many hospital encounters occurred each year

```
library(lubridate)
# convert the date into the standard format, and add the year to a new
column of the dataframe
Date<-as.Date(encounter_patient$contact_date,format = '%m/%d/%Y')
encounter_patient$Year<-year(Date)
# print the table
Year_en<-encounter_patient%>%
  filter(encounter_patient$enc_type=="Hospital Encounter")%>%
  group_by(Year)%>%
  summarise(n=n(),.groups="drop_last")
kable(Year_en, col.names = c("Year","Count"),caption = "Table 3: the nu
mber of hospital encounters each year",align = "c")
```

Table 3: the number of hospital encounters each year

Year	Count
2014	12
2015	9
2016	9
2017	7
2018	8
2019	8

Create & print a table of the counts & percentages of race, financial class, hypertension, congestive heart failure, and diabetes

```
# since the five variables are all from the 'patient' dataframe, then u
se the 'patient' dataframe to print table (to avoid the replicate rows)
# for race
race<-patient%>%
```

```
  group_by(race)%>%
  summarise(Count1=n(),.groups="drop_last")%>%
  mutate(Proportion1=prop.table(Count1)*100)
  colnames(race)<-c("**Classification of Race**","**Count**","**Percent
age (%)**")

# for financial class
financial<-patient%>%
  group_by(financialclass)%>%
  summarise(Count2=n(),.groups="drop_last")%>%
  mutate(Proportion2=prop.table(Count2)*100)
 colnames(financial)<-c("**Classification of Financialclass**","**Count
**","**Percentage (%)**")

# for hypertension
hyper<-patient%>%
  group_by(hypertension)%>%
  summarise(Count3=n(),.groups="drop_last")%>%
  mutate(Proportion3=prop.table(Count3)*100)
 colnames(hyper)<-c("**Classification of Hypertension**","**Count**","*
*Percentage (%)**")

# for congestive heart failure (CHF)
CHF<-patient%>%
  group_by(CHF)%>%
  summarise(Count4=n(),.groups="drop_last")%>%
  mutate(Proportion4=prop.table(Count4)*100)
 colnames(CHF)<-c("**Classification of CHF**","**Count**","**Percentage
 (%)**")

# for diabetes
diabetes<-patient%>%
  group_by(diabetes)%>%
  summarise(Count5=n(),.groups="drop_last")%>%
  mutate(Proportion5=prop.table(Count5)*100)
 colnames(diabetes)<-c("**Classification of Diabetes**","**Count**","**
Percentage (%)**")

# print the table
kable(list(race,financial,hyper,CHF,diabetes),caption = "Table 4: for t
he following five variables",align = "c",digits = 3)
```

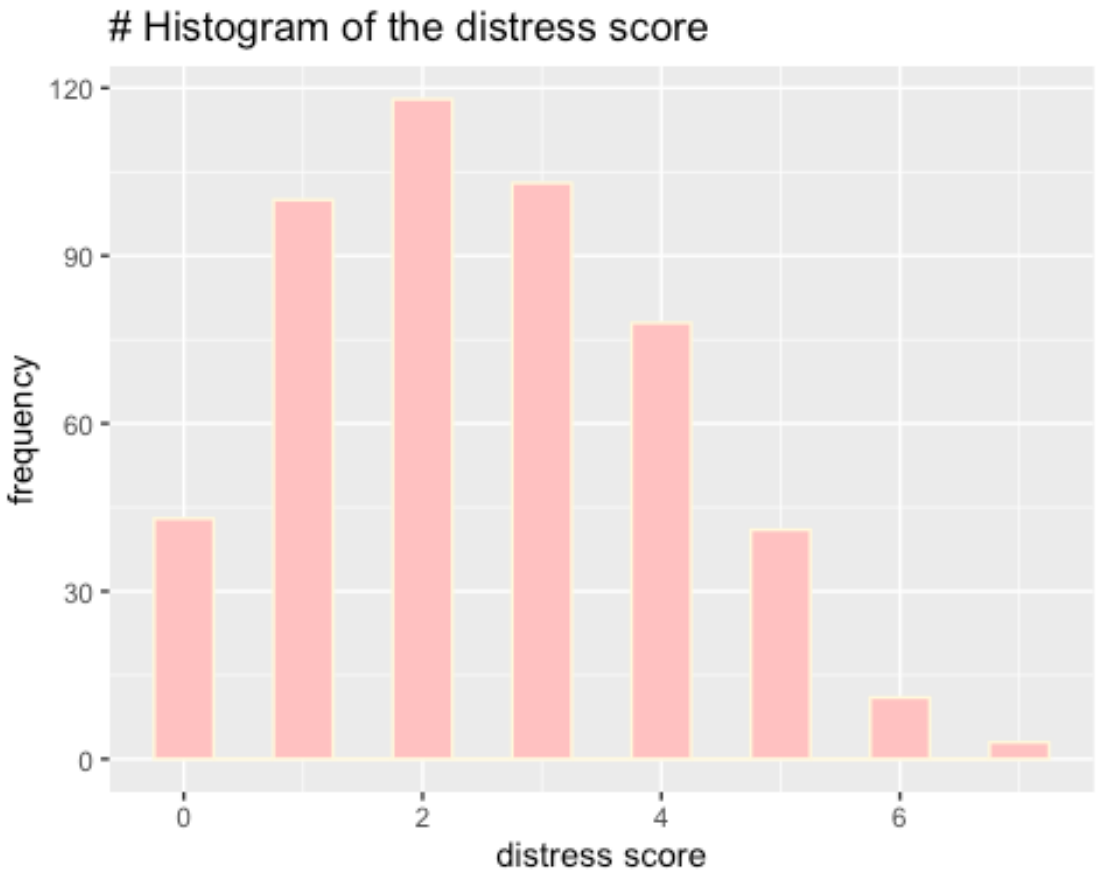
Table 4: for the following five variables

Classification of Race	Count	Percentage (%)
Black	10	20
Other	3	6
White	37	74

Classification of Financialclass	Count	Percentage (%)
Medicare	29	58
Private	21	42
Classification of Hypertension	**Count**	**Percentage (%)**
N	30	60
Y	20	40
Classification of CHF	**Count**	**Percentage (%)**
N	45	90
Y	5	10
Classification of Diabetes	**Count**	**Percentage (%)**
N	48	96
Y	2	4

Create a histogram of the distress score

```
# create the histogram
qplot(encounter_patient$distress,main = "# Histogram of the distress sc
ore",geom = "histogram",binwidth=0.5,fill=I("rosybrown1"),col=I("cornsi
lk"),xlab = "distress score",ylab = "frequency")
```


Histogram of the distress score
120
90
60
30
0
frequency
0
2
4
6
distress score