submission.R

hardey

2022-11-05

```
# loading required  libraries
library(ggplot2)
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(caTools)
# import the dataset
df <- read_excel('/Users/hardey/Desktop/Fortray/R/submission/1555054100_hospitalcosts.xlsx')
#View(df)

# shape of the dataset
dim(df)
```

```
## [1] 500   6
```

```
# basic statistics of the dataset
summary(df)
```

```
##       AGE             FEMALE           LOS              RACE      
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000  
##  1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000  
##  Median : 0.000   Median :1.000   Median : 2.000   Median :1.000  
##  Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078  
##  3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000  
##  Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000  
##                                                    NA's   :1      
##      TOTCHG          APRDRG
##  Min.   :  532   Min.   : 21.0
##  1st Qu.: 1216   1st Qu.:640.0
##  Median : 1536   Median :640.0
##  Mean   : 2774   Mean   :616.4
##  3rd Qu.: 2530   3rd Qu.:751.0
##  Max.   :48388   Max.   :952.0
##
```

```
# structure of the dataset
str(df)
```

```
## tibble [500 x 6] (S3: tbl_df/tbl/data.frame)
##  $ AGE   : num [1:500] 17 17 17 17 17 17 17 16 16 17 ...
##  $ FEMALE: num [1:500] 1 0 1 1 1 0 1 1 1 1 ...
##  $ LOS   : num [1:500] 2 2 7 1 1 0 4 2 1 2 ...
##  $ RACE  : num [1:500] 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTCHG: num [1:500] 2660 1689 20060 736 1194 ...
##  $ APRDRG: num [1:500] 560 753 930 758 754 347 754 754 753 758 ...
```

```
table(df$FEMALE)
```

```
##
##   0   1
## 244 256
```

```
table(df$RACE)
```

```
##
##   1   2   3   4   5   6
## 484   6   1   3   3   2
```

```
# check missing value
sum(is.na(df))
```

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

```
new_df <- na.omit(df)
attach(new_df)
# AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG
min(new_df$AGE)
```

```
## [1] 0
```

```
max(new_df$AGE)
```

```
## [1] 17
```

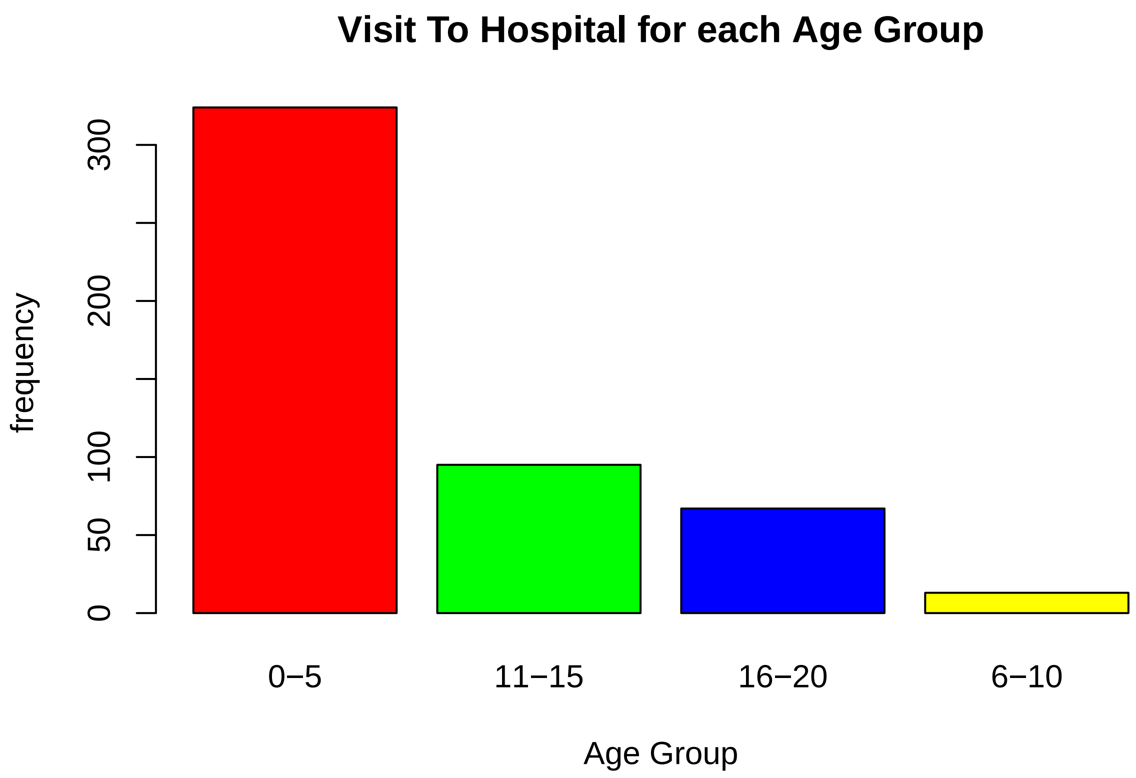
```
new_df$AGEGROUP <- ifelse (AGE <= 5,'0-5',
                ifelse (AGE <= 10,'6-10',
                ifelse (AGE <= 15,'11-15','16-20')))

new_df$GENDER <- ifelse(FEMALE==1,'Female','Male')

#View(new_df)
attach(new_df)
```

```
## The following objects are masked from new_df (pos = 3):
##
##     AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG
```

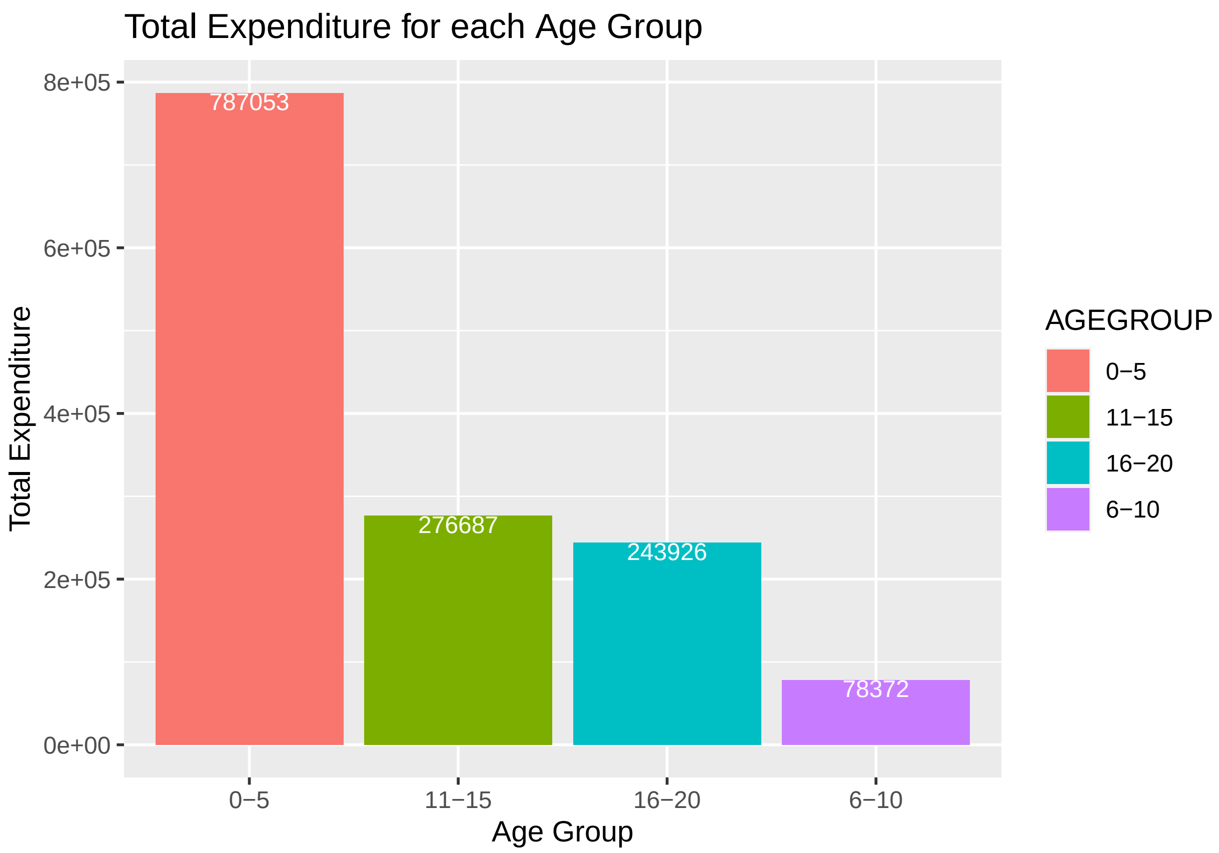
```
# convert the female and race column to factors
new_df$FEMALE  <-  as.factor(FEMALE)
new_df$RACE  <-  as.factor(RACE)
new_df$APRDRG  <-  as.factor(APRDRG)

#check the new structure to confirm if they have been converted
str(new_df)
```

```
## tibble [499 x 8] (S3: tbl_df/tbl/data.frame)
##  $ AGE     : num [1:499] 17 17 17 17 17 17 17 16 16 17 ...
##  $ FEMALE  : Factor w/ 2 levels "0","1": 2 1 2 2 2 1 2 2 2 2 ...
##  $ LOS     : num [1:499] 2 2 7 1 1 0 4 2 1 2 ...
##  $ RACE    : Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTCHG  : num [1:499] 2660 1689 20060 736 1194 ...
##  $ APRDRG  : Factor w/ 63 levels "21","23","49",..: 32 51 62 55 52 28 52 52 51 55 ...
##  $ AGEGROUP: chr [1:499] "16-20" "16-20" "16-20" "16-20" ...
##  $ GENDER  : chr [1:499] "Female" "Male" "Female" "Female" ...
##  - attr(*, "na.action")= 'omit' Named int 277
##   ..- attr(*, "names")= chr "277"
```

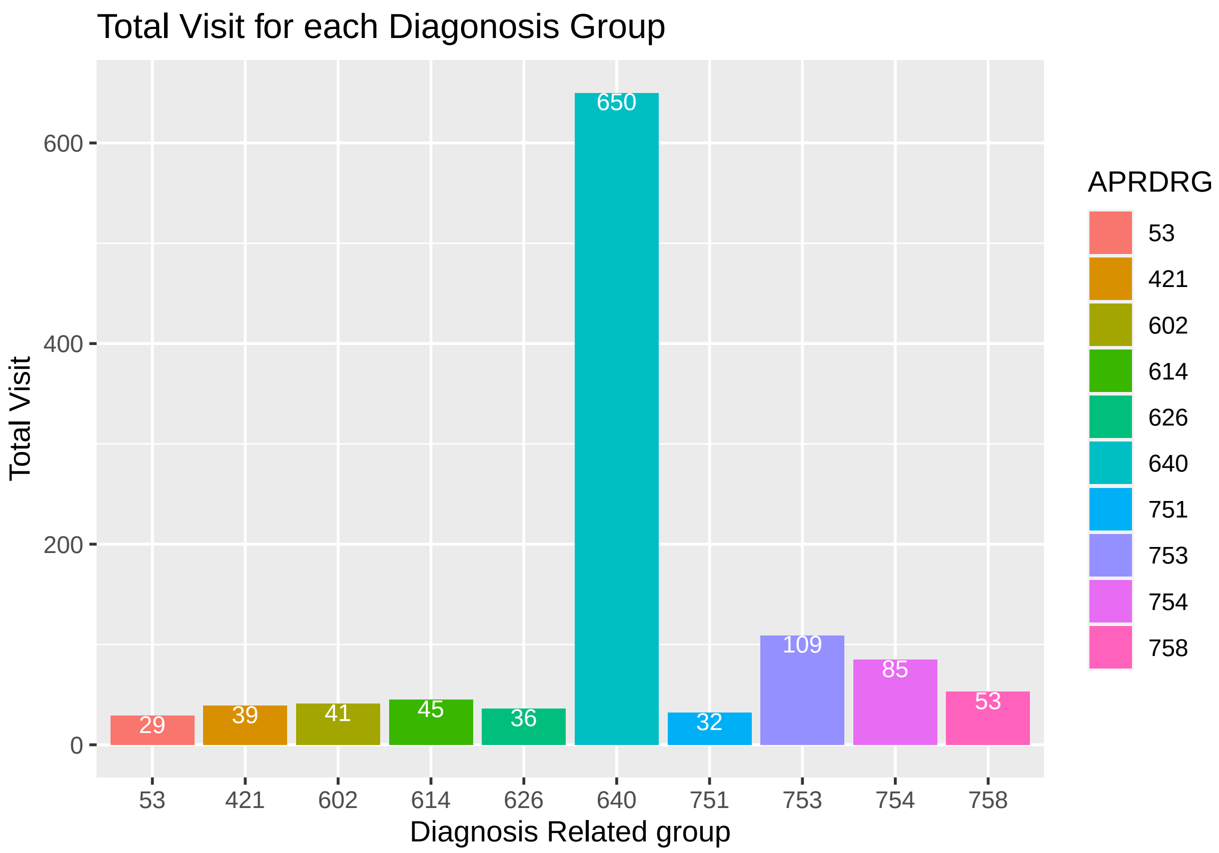
```
# Q1:

#Age category who frequently visited the hospital most
age_group<- table(AGEGROUP)
age_group_visit <-  barplot(age_group,
                  main='Visit To Hospital for each Age Group',
                  xlab='Age Group',
                  ylab='frequency',
                  col=c('red','green','blue','yellow'))
```



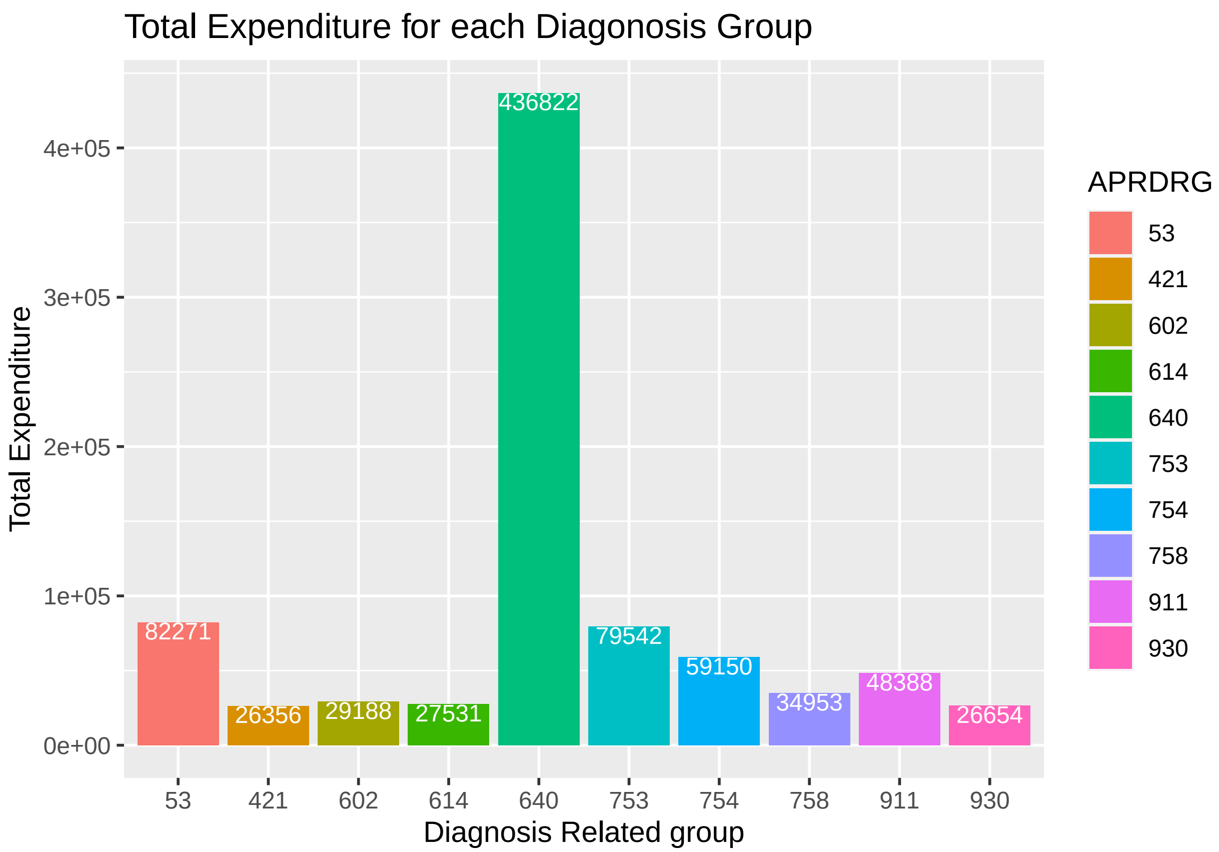
```r
# Age category with the maximum hospitalization and expenditure
new_df %>%
  group_by(AGEGROUP) %>%
  summarise(Total_Exp = sum(TOTCHG)) %>%
  ggplot(aes(x=AGEGROUP,y=Total_Exp,fill=AGEGROUP))+
  geom_bar(stat='identity')+
  geom_text(aes(label= (Total_Exp)),vjust=1.0,color="white",size=3.0)+
  ggtitle("Total Expenditure for each Age Group")+
  xlab("Age Group") + ylab('Total Expenditure')
```



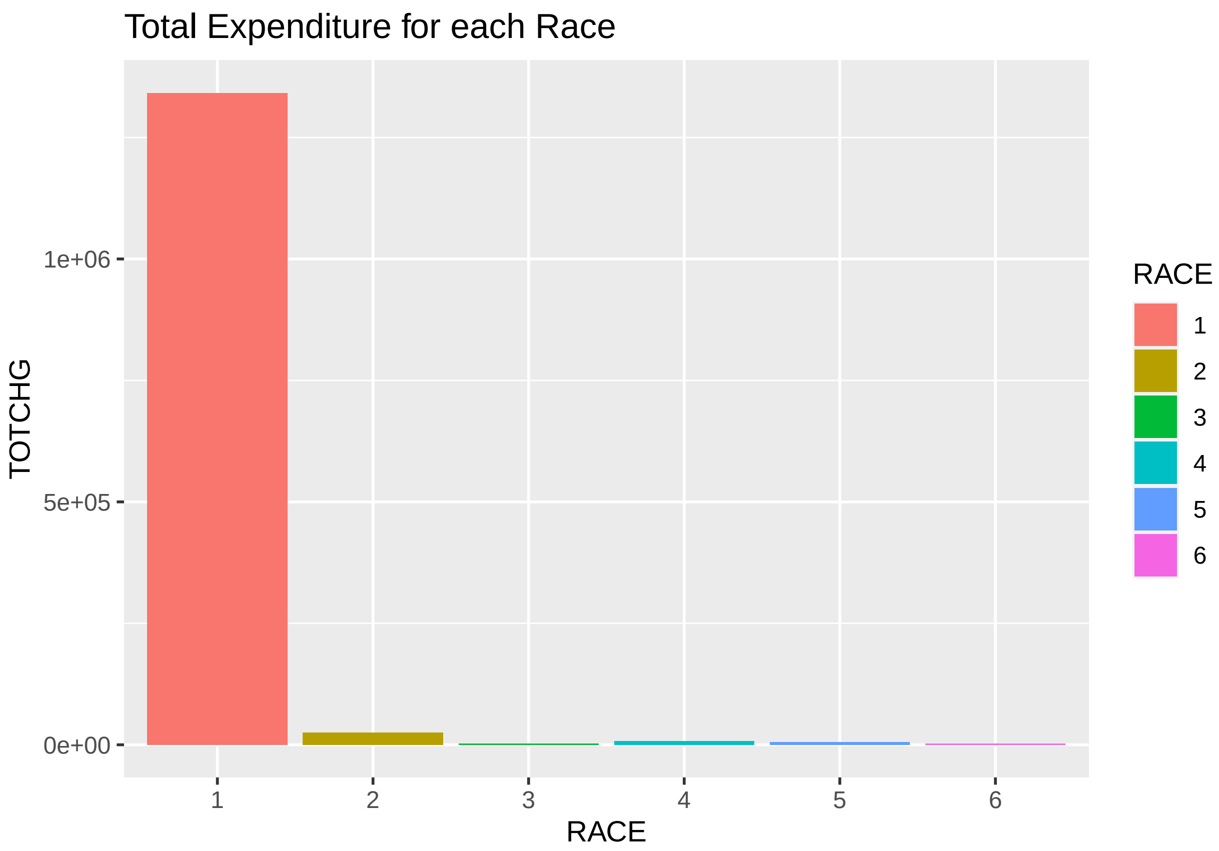
```
#Q2:
# Diagnosis Related group that has maximum hospitalization.
new_df %>%
  group_by(APRDRG) %>%
  summarise(Total_Visit = sum(LOS)) %>%
  arrange(desc(Total_Visit)) %>%
  slice(1:10) %>%
  ggplot(aes(x=APRDRG,y=Total_Visit,fill=APRDRG))+
  geom_bar(stat='identity')+
  geom_text(aes(label= (Total_Visit)),vjust=1.0,color="white",size=3.0)+
  ggtitle("Total Visit for each Diagonosis Group")+
  xlab("Diagnosis Related group") + ylab('Total Visit')
```



```
#Diagnosis Related group that has maximum expenditure.

new_df %>%
  group_by(APRDRG) %>%
  summarise(Total_Exp= sum(TOTCHG)) %>%
  arrange(desc(Total_Exp)) %>%
  slice(1:10) %>%
  ggplot(aes(x=APRDRG,y=Total_Exp,fill=APRDRG))+
  geom_bar(stat='identity')+
  geom_text(aes(label= (Total_Exp)),vjust=1.0,color="white",size=3.0)+
  ggtitle("Total Expenditure for each Diagonosis Group")+
  xlab("Diagnosis Related group") + ylab('Total Expenditure')
```



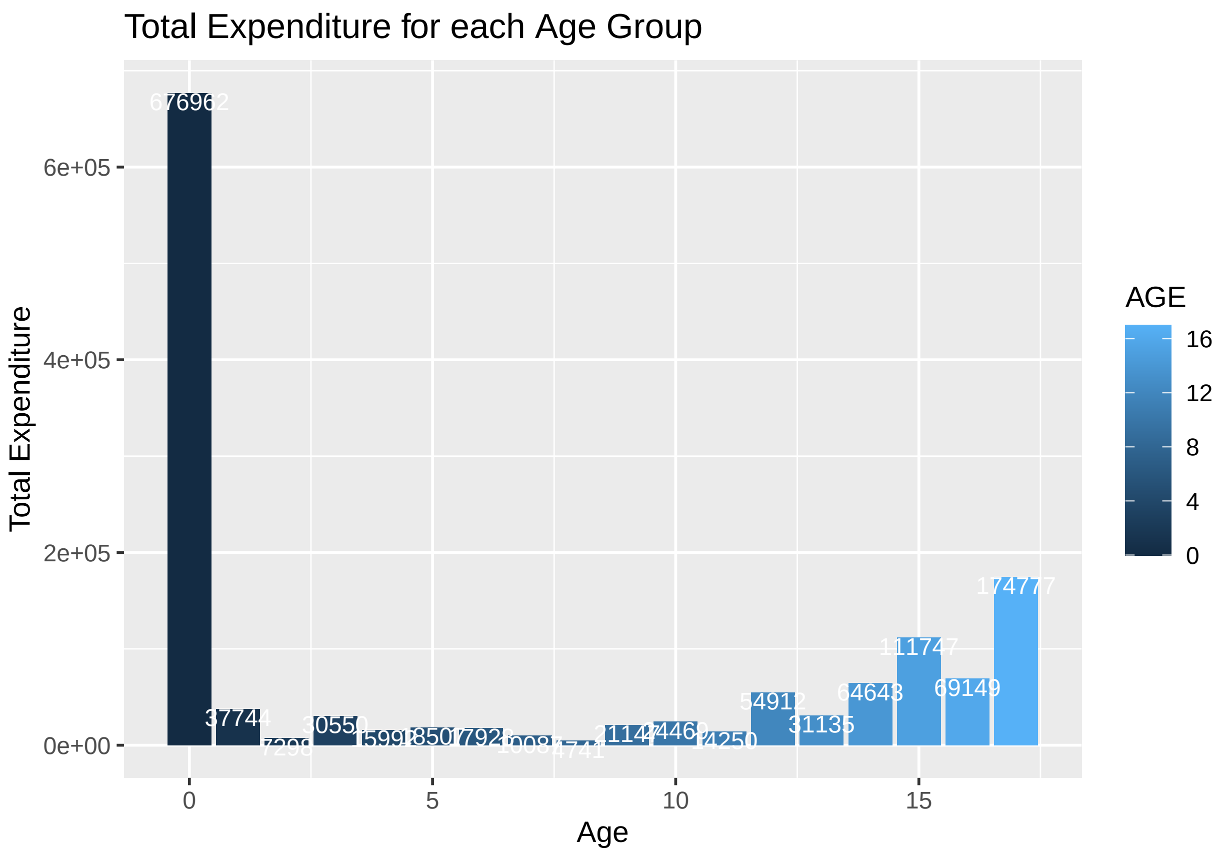
```
#Q3:Relationship between race of the patient and hospitalization costs.

ggplot(new_df, aes(x=RACE, y=TOTCHG)) + geom_col(aes(fill=RACE))+
  ggtitle("Total Expenditure for each Race")
```



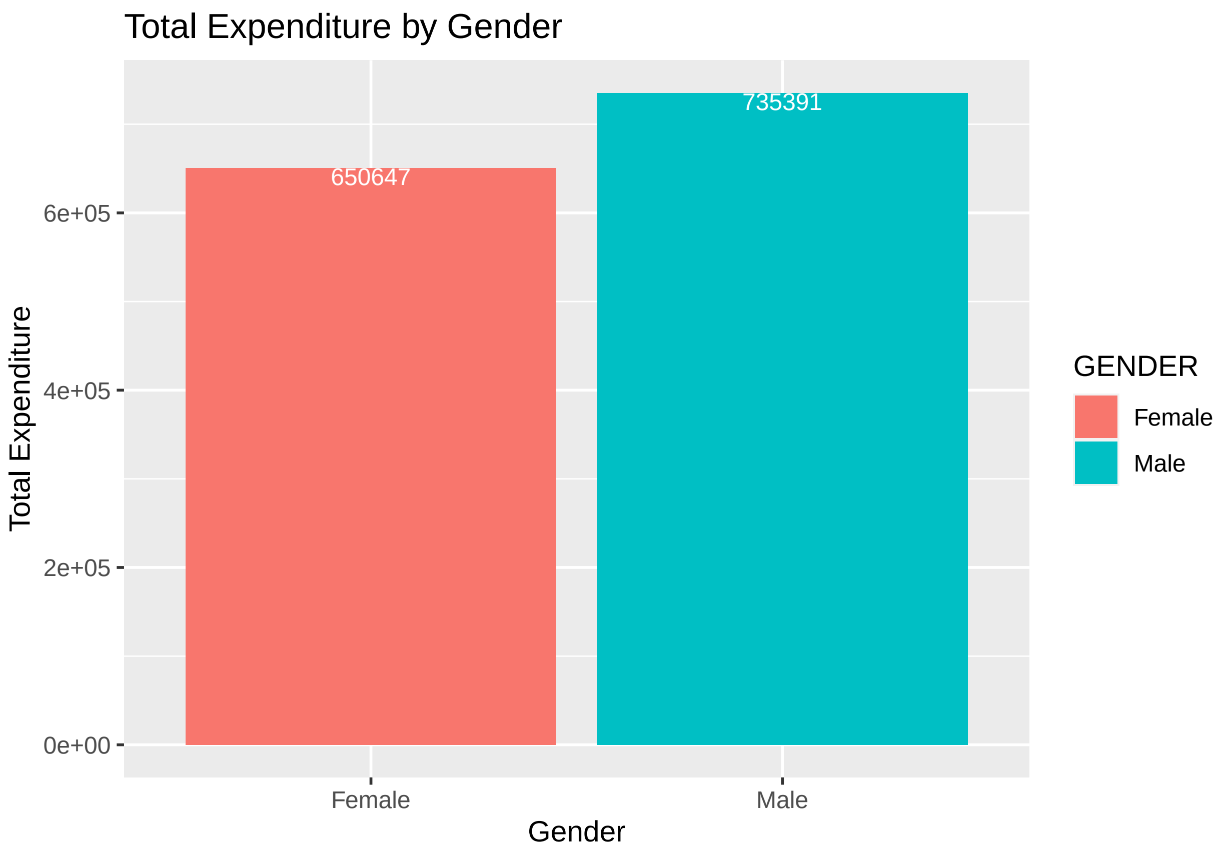
```
summary(aov(TOTCHG~RACE))
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## RACE          1 2.488e+06  2488459   0.164  0.686
## Residuals   497 7.540e+09 15170268
```

```
#Q4:

# severity of the hospital costs by age
new_df %>%
  group_by(AGE) %>%
  summarise(Total_Exp = sum(TOTCHG)) %>%
  ggplot(aes(x=AGE,y=Total_Exp,fill=AGE))+
  geom_bar(stat='identity')+
  geom_text(aes(label= (Total_Exp)),vjust=1.0,color="white",size=3.0)+
  ggtitle("Total Expenditure for each Age Group")+
  xlab("Age") + ylab('Total Expenditure')
```



```
# severity of the hospital costs by gender
new_df %>%
  group_by(GENDER) %>%
  summarise(Total_Exp= sum(TOTCHG)) %>%
  ggplot(aes(x=GENDER,y=Total_Exp,fill=GENDER))+
  geom_bar(stat='identity')+
  geom_text(aes(label= (Total_Exp)),vjust=1.0,color="white",size=3.0)+
  ggtitle("Total Expenditure by Gender")+
  xlab("Gender") + ylab('Total Expenditure')
```



```
# Q5:
# predicting length of stay from age, gender, and race.

# convert the female and race column back to numerical
new_df$FEMALE  <-  as.numeric(FEMALE)
new_df$RACE  <-  as.numeric(RACE)
new_df$APRDRG  <-  as.numeric(APRDRG)

# Finding the correlation between the variables
cor(new_df$AGE,new_df$RACE)
```

```
## [1] 0.01584962
```

```
cor(new_df$AGE,new_df$FEMALE)
```

```
## [1] 0.2357642
```

```
cor(new_df$RACE,new_df$FEMALE)
```

```
## [1] -0.03843368
```

```
# finding the covariance of the variables
cov(new_df$AGE,new_df$RACE)
```

```
## [1] 0.05672389
```

```
cov(new_df$AGE,new_df$FEMALE)
```

```
## [1] 0.8202228
```

```
cov(new_df$RACE,new_df$FEMALE)
```

```
## [1] -0.009899317
```

```
# set seed for reproducibility
set.seed(94)

# Train:Test split = 75:25
sample1 <- sample.split(new_df$LOS,
                       SplitRatio = 0.75)

train1 <- subset(new_df,sample1==TRUE)
test1 <- subset(new_df,sample1==FALSE)
model1 <- lm(LOS~AGE+FEMALE+RACE,data=train1)
prediction <- predict(model1,test1)
prediction
```

```
##        1        2        3        4        5        6        7        8
## 2.671643 2.315221 2.671643 2.711993 2.752343 2.792692 2.671643 2.516970
##        9       10       11       12       13       14       15       16
## 2.395920 2.671643 2.718720 2.880120 2.711993 2.920470 2.752343 2.960820
##       17       18       19       20       21       22       23       24
## 2.792692 2.671643 3.236542 2.833042 3.357592 3.357592 3.357592 3.001170
##       25       26       27       28       29       30       31       32
## 3.001170 3.001170 3.357592 3.357592 3.001170 3.001170 3.001170 3.357592
##       33       34       35       36       37       38       39       40
## 3.001170 3.001170 3.001170 3.357592 2.671643 3.001170 2.880120 3.001170
##       41       42       43       44       45       46       47       48
## 3.001170 2.752343 3.001170 3.001170 3.001170 3.001170 3.357592 2.711993
##       49       50       51       52       53       54       55       56
## 2.752343 3.001170 3.001170 2.395920 3.001170 2.315221 3.357592 3.357592
##       57       58       59       60       61       62       63       64
## 3.357592 3.357592 3.357592 3.001170 3.357592 3.357592 2.236895 3.001170
##       65       66       67       68       69       70       71       72
## 3.357592 3.357592 3.001170 3.001170 3.001170 3.001170 3.357592 3.357592
##       73       74       75       76       77       78       79       80
## 2.355570 2.711993 2.792692 2.671643 2.678370 2.792692 2.960820 3.001170
##       81       82       83       84       85       86       87       88
## 3.001170 3.001170 2.960820 3.001170 3.357592 3.001170 3.357592 3.357592
##       89       90       91       92       93       94       95       96
## 3.001170 3.357592 3.001170 3.001170 2.711993 3.001170 2.671643 3.001170
##       97       98       99      100      101      102      103      104
## 3.357592 3.001170 3.001170 2.407791 3.357592 3.001170 2.315221 3.357592
##      105      106      107      108      109      110      111      112
## 3.001170 3.357592 3.357592 3.357592 3.357592 2.315221 2.711993 2.792692
##      113      114      115      116      117      118      119      120
## 3.001170 3.001170 2.645142 3.357592 3.001170 2.516970 3.001170 3.001170
##      121      122      123
## 3.001170 3.357592 3.357592
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = LOS ~ AGE + FEMALE + RACE, data = train1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.358 -1.358 -1.001 -0.001 37.642
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.11985    0.51155   6.099 2.68e-09 ***
## AGE         -0.04035    0.02896  -1.393    0.164
## FEMALE       0.35642    0.40209   0.886    0.376
## RACE        -0.11868    0.38455  -0.309    0.758
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.796 on 372 degrees of freedom
## Multiple R-squared:  0.006486,   Adjusted R-squared:  -0.001527
## F-statistic: 0.8095 on 3 and 372 DF,  p-value: 0.4892
```

```
# Q6:

# Finding the variable which is significant to Hospital costs.

# set seed for reproducibility
set.seed(94)
# Train:Test split = 75:25
sample2 <- sample.split(new_df$APRDRG,
                       SplitRatio = 0.75)
train2 <- subset(new_df,sample2==TRUE)
test2 <- subset(new_df,sample2==FALSE)
model2 <- lm(TOTCHG~AGE+FEMALE+LOS+RACE+APRDRG,data=train2)
summary(model2)
```

```
##
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE + LOS + RACE + APRDRG, data = train2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6435   -872   -237    166  43123
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5136.1129   629.4866   8.159 4.96e-15 ***
## AGE          149.0945    21.9724   6.786 4.44e-11 ***
## FEMALE      -516.5136   317.7626  -1.625    0.105    
## LOS          755.4271    41.8556  18.048  < 2e-16 ***
## RACE        -210.7635   287.4179  -0.733    0.464    
## APRDRG        -7.6474     0.8245  -9.275  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2937 on 381 degrees of freedom
## Multiple R-squared:  0.5333, Adjusted R-squared:  0.5272 
## F-statistic: 87.08 on 5 and 381 DF,  p-value: < 2.2e-16
```