**HealthCare Data Analysis**

## Read data file

Hospital <- read.csv(file.choose(),header = T)

View(Hospital)

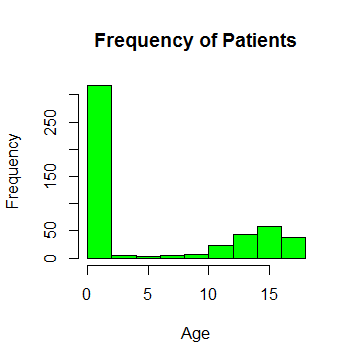
Answer-1:: Patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.

hist(Hospital$AGE,

main="Frequency of Patients",

col ="green",

xlab = "Age")



attach(Hospital)

AGE <-as.factor(AGE)

summary(AGE)

View(summary(AGE))

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

307 10 1 3 2 2 2 3 2 2 4 8 15 18 25 29 29 38

##Aggregate function is used to add the expenditure from each age

aggregate(TOTCHG~AGE,FUN = sum,data = Hospital)

##Max

max(aggregate(TOTCHG~AGE,FUN = sum,data = Hospital))

[1] 676962

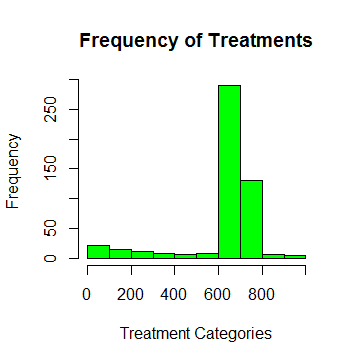
Answer-2::In order of severity of the diagnosis and treatments and to find out the expensive treatments ,the agency wants to find the diagnosis related group that has maximum hospitalization and expenditure.

hist(APRDRG,

col = "green",

main = "Frequency of Treatments",

xlab = "Treatment Categories")



APRDRG\_fact <-as.factor(Hospital$APRDRG)

summary(APRDRG\_fact)

21 23 49 50 51 53 54 57 58 92 97 114 115 137 138 139 141 143 204 206 225 249 254 308 313 317 344 347 420

1 1 1 1 1 10 1 2 1 1 1 1 2 1 4 5 1 1 1 1 2 6 1 1 1 1 2 3 2

421 422 560 561 566 580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 754 755 756 758 760 776 811

1 3 2 1 1 1 3 1 3 6 4 2 3 4 266 1 1 2 1 1 14 36 37 13 2 20 2 1 2

812 863 911 930 952

3 1 1 2 1

#Max

which.max(summary(APRDRG\_fact))

640

44

df<-aggregate(TOTCHG~APRDRG,FUN = sum,data = Hospital)

df

#Total charge

df[which.max(df$TOTCHG),]

APRDRG TOTCHG

44 640 436822

Answer-3::To make sure that there is no malpractice, the agency need to analyze if the race of the patient is related to the hospitalization costs.

# Remove "NA" value

Hospital<-na.omit(Hospital)

#Factorize the Race variable

Hospital$RACE<-as.factor(Hospital$RACE)

#ANOVA function with TOTCHG and RACE Variable.

model\_aov<-aov(TOTCHG~RACE,data = Hospital)

#ANOVA RESULTS

model\_aov

summary(model\_aov)

Df Sum Sq Mean Sq F value Pr(>F)

RACE 5 1.859e+07 3718656 0.244 0.943

Residuals 493 7.524e+09 15260687

#Getting max hospital cost per race

summary(Hospital$RACE)

1 2 3 4 5 6

484 6 1 3 3 2

Answer-4:: Utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for proper allocation of resources.

#Analyze the severity of costs

Hospital$FEMALE<-as.factor(Hospital$FEMALE)

#calling Regression function

model\_lm4 <-lm(TOTCHG~AGE+FEMALE,data = Hospital)

summary(model\_lm4)

#comparing genders

summary(Hospital$FEMALE)

0 1

244 255

Answer-5::Since the length of stay is the crucial factor for inpatients,the agency wants to find id the length of stay can be predicted from age,gender,and race.

#Linier Regression

Hospital$RACE<-as.factor(Hospital$RACE)

model\_lm5<-lm(LOS~AGE+FEMALE+RACE,data = Hospital)

summary(model\_lm5)

Call:

lm(formula = LOS ~ AGE + FEMALE + RACE, data = Hospital)

Residuals:

Min 1Q Median 3Q Max

-3.211 -1.211 -0.857 0.143 37.789

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.85687 0.23160 12.335 <2e-16 \*\*\*

AGE -0.03938 0.02258 -1.744 0.0818 .

FEMALE1 0.35391 0.31292 1.131 0.2586

RACE2 -0.37501 1.39568 -0.269 0.7883

RACE3 0.78922 3.38581 0.233 0.8158

RACE4 0.59493 1.95716 0.304 0.7613

RACE5 -0.85687 1.96273 -0.437 0.6626

RACE6 -0.71879 2.39295 -0.300 0.7640

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.376 on 491 degrees of freedom

Multiple R-squared: 0.008699, Adjusted R-squared: -0.005433

F-statistic: 0.6156 on 7 and 491 DF, p-value: 0.7432

Answer-6::TO perform a complete analysis, the agency wants to find the variable that mainly affects the hospital costs.

#Linier Regressin

model\_lm6<-lm(TOTCHG~AGE+FEMALE+RACE+LOS+APRDRG,data = Hospital)

summary(model\_lm6)

Call:

lm(formula = TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = Hospital)

Residuals:

Min 1Q Median 3Q Max

-6367 -691 -186 121 43412

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5024.9610 440.1366 11.417 < 2e-16 \*\*\*

AGE 133.2207 17.6662 7.541 2.29e-13 \*\*\*

FEMALE1 -392.5778 249.2981 -1.575 0.116

RACE2 458.2427 1085.2320 0.422 0.673

RACE3 330.5184 2629.5121 0.126 0.900

RACE4 -499.3818 1520.9293 -0.328 0.743

RACE5 -1784.5776 1532.0048 -1.165 0.245

RACE6 -594.2921 1859.1271 -0.320 0.749

LOS 742.9637 35.0464 21.199 < 2e-16 \*\*\*

APRDRG -7.8175 0.6881 -11.361 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2622 on 489 degrees of freedom

Multiple R-squared: 0.5544, Adjusted R-squared: 0.5462

F-statistic: 67.6 on 9 and 489 DF, p-value: < 2.2e-16