Import data into R environment.

The Excel File attached in Project Requirement is converted into CSV and then used.

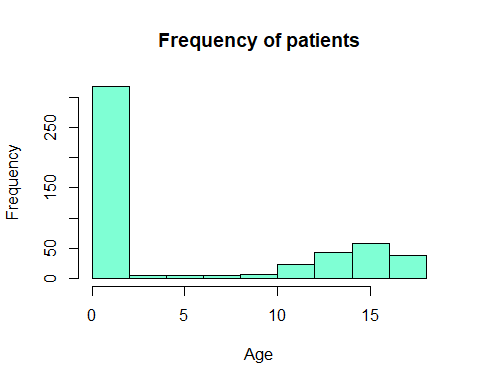
hosp<-read.csv("HospitalCosts.csv")

head(hosp, n=3)

## AGE FEMALE LOS RACE TOTCHG APRDRG  
## 1 17 1 2 1 2660 560  
## 2 17 0 2 1 1689 753  
## 3 17 1 7 1 20060 930

1. **To record the patient statistics, the agency wants to find the age category of people who frequent the hospital and has the maximum expenditure.**

hist(hosp$AGE,main = "Frequency of patients",col = "aquamarine",xlab = "Age")



attach(hosp)   
AGE<-as.factor(AGE)   
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

Conclusion 1: From the above results we conclude that infant category h as the max hospital visits (above 300). The summary of Age gives us the exact numerical output showing that Age 0 patients have the max visits followed by Ages 15-17.

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

## AGE TOTCHG  
## 1 0 678118  
## 2 1 37744  
## 3 2 7298  
## 4 3 30550  
## 5 4 15992  
## 6 5 18507  
## 7 6 17928  
## 8 7 10087  
## 9 8 4741  
## 10 9 21147  
## 11 10 24469  
## 12 11 14250  
## 13 12 54912  
## 14 13 31135  
## 15 14 64643  
## 16 15 111747  
## 17 16 69149  
## 18 17 174777

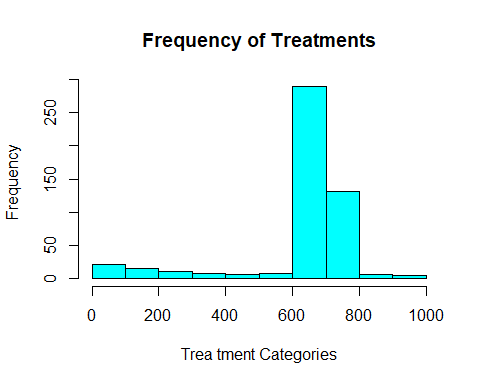
max(aggregate(TOTCHG~AGE,FUN=sum,data=hosp))

## [1] 678118

Conclusion 2: Thus, we can conclude that the infants also have the maximum hospital costs followed by Age groups 15 to 17, additionally we can say confidently that number of hospital visits are proportional to hospital costs.

1. **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 = "cyan1",main = "Frequency of Treatments",xlab = "Trea tment Categories")



APRDRG\_fact<-as.factor(hosp$APRDRG)   
summary(APRDRG\_fact)

## 21 23 49 50 51 53 54 57 58 92 97 114 115 137 138 139 141 143 204 206   
## 1 1 1 1 1 10 1 2 1 1 1 1 2 1 4 5 1 1 1 1   
## 225 249 254 308 313 317 344 347 420 421 422 560 561 566 580 581 602 614 626 633   
## 2 6 1 1 1 1 2 3 2 1 3 2 1 1 1 3 1 3 6 4   
## 634 636 639 640 710 720 723 740 750 751 753 754 755 756 758 760 776 811 812 863   
## 2 3 4 267 1 1 2 1 1 14 36 37 13 2 20 2 1 2 3 1   
## 911 930 952   
## 1 2 1

which.max(summary(APRDRG\_fact))

## 640   
## 44

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

## APRDRG TOTCHG  
## 1 21 10002  
## 2 23 14174  
## 3 49 20195  
## 4 50 3908  
## 5 51 3023  
## 6 53 82271  
## 7 54 851  
## 8 57 14509  
## 9 58 2117  
## 10 92 12024  
## 11 97 9530  
## 12 114 10562  
## 13 115 25832  
## 14 137 15129  
## 15 138 13622  
## 16 139 17766  
## 17 141 2860  
## 18 143 1393  
## 19 204 8439  
## 20 206 9230  
## 21 225 25649  
## 22 249 16642  
## 23 254 615  
## 24 308 10585  
## 25 313 8159  
## 26 317 17524  
## 27 344 14802  
## 28 347 12597  
## 29 420 6357  
## 30 421 26356  
## 31 422 5177  
## 32 560 4877  
## 33 561 2296  
## 34 566 2129  
## 35 580 2825  
## 36 581 7453  
## 37 602 29188  
## 38 614 27531  
## 39 626 23289  
## 40 633 17591  
## 41 634 9952  
## 42 636 23224  
## 43 639 12612  
## 44 640 437978  
## 45 710 8223  
## 46 720 14243  
## 47 723 5289  
## 48 740 11125  
## 49 750 1753  
## 50 751 21666  
## 51 753 79542  
## 52 754 59150  
## 53 755 11168  
## 54 756 1494  
## 55 758 34953  
## 56 760 8273  
## 57 776 1193  
## 58 811 3838  
## 59 812 9524  
## 60 863 13040  
## 61 911 48388  
## 62 930 26654  
## 63 952 4833

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

## APRDRG TOTCHG  
## 44 640 437978

Conclusion: Hence can conclude that category 640 has the maximum hospitalizations by a huge number (267 out of 500), along with this it also has the highest hospitalization cost.

1. **To make sure that there is no malpractice, the agency needs to analyze if the race of the patient is related to the hospitalization costs.**

hosp<-na.omit(hosp)  
  
#first we remove "NA"values   
hosp$RACE<-as.factor(hosp$RACE)   
model\_aov<-aov(TOTCHG~RACE,data = hosp)   
model\_aov#ANOVA RESULTS

## Call:  
## aov(formula = TOTCHG ~ RACE, data = hosp)  
##   
## Terms:  
## RACE Residuals  
## Sum of Squares 18593279 7523518505  
## Deg. of Freedom 5 493  
##   
## Residual standard error: 3906.493  
## Estimated effects may be unbalanced

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

summary(hosp$RACE)#getting max hospital cost per race

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

Conclusion: F value is quite low, which means that variation between hospital costs among different races is much smaller than the variation of hospital costs within each race, and P value being quite high shows that there is no relationship between race and hospital costs, thereby accepting the Null hypothesis. Additionally, we have more data for Race 1 in comparison to other races (484 out of 500 patients) which make the observations skewed and thus all we can say is that there isn’t enough data to verify whether race of a patient affects hospital costs.

1. **To properly utilize the costs, the agency has to analyze the severity of the hospital costs by age and gender for proper allocation of resources.**

hosp$FEMALE<-as.factor(hosp$FEMALE)   
model\_lm4<-lm(TOTCHG~AGE+FEMALE,data = hosp)  
#calling Regression funtion   
summary(model\_lm4)

##   
## Call:  
## lm(formula = TOTCHG ~ AGE + FEMALE, data = hosp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3403 -1444 -873 -156 44950   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2719.45 261.42 10.403 < 2e-16 \*\*\*  
## AGE 86.04 25.53 3.371 0.000808 \*\*\*  
## FEMALE1 -744.21 354.67 -2.098 0.036382 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3849 on 496 degrees of freedom  
## Multiple R-squared: 0.02585, Adjusted R-squared: 0.02192   
## F-statistic: 6.581 on 2 and 496 DF, p-value: 0.001511

summary(hosp$FEMALE) #comapring genders

## 0 1   
## 244 255

Conclusion-Age has more impact than gender according to the P-values and significant levels, also there are equal number of Females and Males and on an average (based on the negative coefficient values) females incur lesser hospital costs than males.

1. **Since the length of stay is the crucial factor for inpatients, the agency wants to find if the length of stay can be predicted from age, gender, and race.**

hosp$RACE<-as.factor(hosp$RACE)   
model\_lm5<-lm(LOS~AGE+FEMALE+RACE,data = hosp)   
summary(model\_lm5)

##   
## Call:  
## lm(formula = LOS ~ AGE + FEMALE + RACE, data = hosp)  
##   
## 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

Conclusion-p-values for all independent variables are quite high thus signifying that there is no linear relationship between the given variables, finally concluding the fact that we can’t predict length of stay of a patient based on age, gender and race.

1. **To perform a complete analysis, the agency wants to find the variable that mainly affects the hospital costs.**

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

##   
## Call:  
## lm(formula = TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = hosp)  
##   
## 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

Conclusion-Age and length of stay affect the total hospital costs. Additionally, there is positive relationship between length of stay to the cost, so with an increase of 1 day there is an addition of a value of 742 to the cost.