Project-Healthcare Cost Analysis

R Project

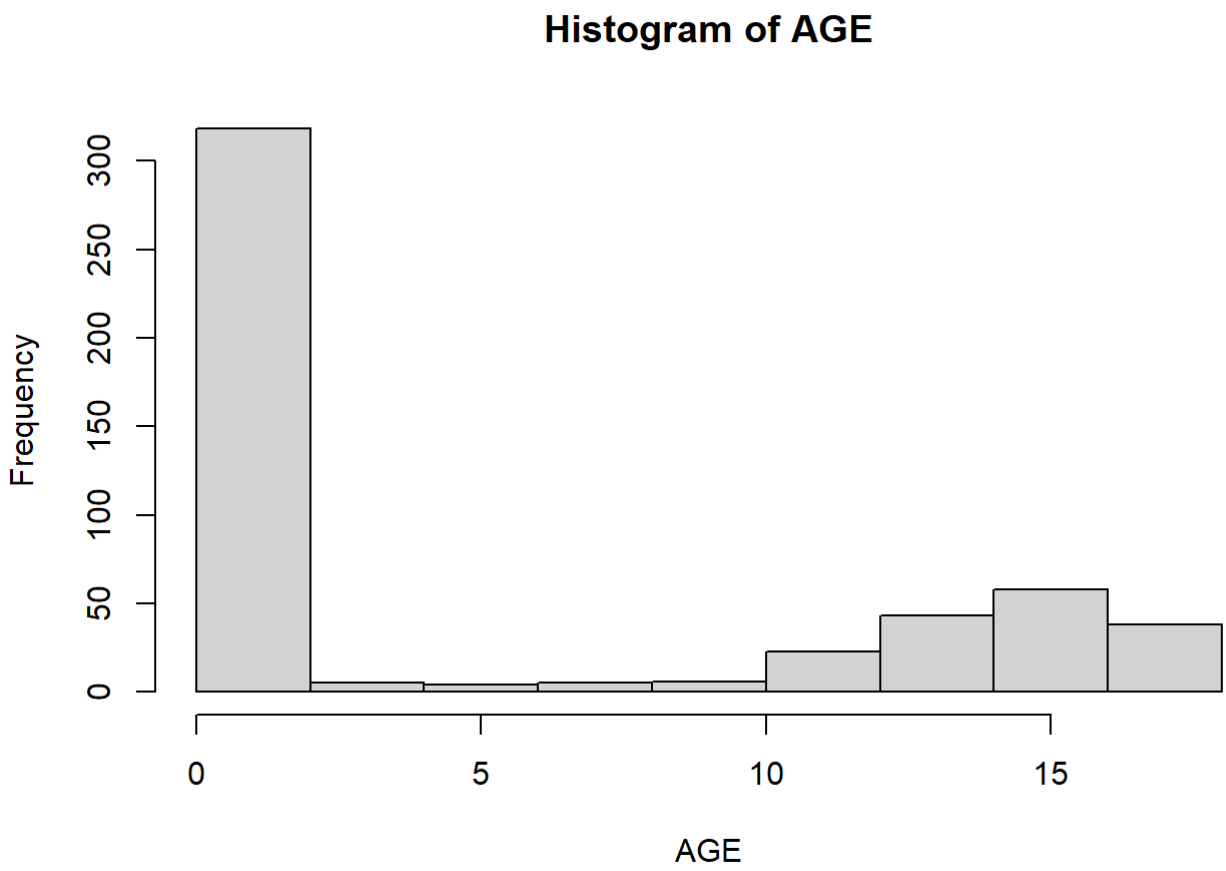
```
#importing data sets
library(readxl)
library(plyr)
library(rmarkdown)
hosp<-read_excel("C:\\Users\\HP\\Desktop\\R Projects\\7 Healthcare cost analysis\\hospitalcos
ts.xlsx")
head(hosp)
```

```
## # A tibble: 6 x 6
##     AGE FEMALE   LOS  RACE TOTCHG APRDRG
##   <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>
## 1    17      1     2     1   2660    560
## 2    17      0     2     1   1689    753
## 3    17      1     7     1  20060    930
## 4    17      1     1     1    736    758
## 5    17      1     1     1   1194    754
## 6    17      0     0     1   3305    347
```

```
summary(hosp)
```

```
##       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
##
```

```
attach(hosp)
#1 insight
hist(AGE)
```




```
#to see the value of category of infants
ag<-as.factor(AGE)
summary(ag)
```

```
##   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
```

```
#age category of 0 seems to be  frequently using the hospital

tapply(TOTCHG,AGE,sum)
```

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

```
which.max(tapply(TOTCHG,AGE,sum))
```

```
## 0
## 1
```

```
#max expenditure also by infant of 0 age =678118, 15=111747 17=174777

#2insight
diagg<-as.factor(APRDRG)
summary(diagg)
```

```
##  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(diagg))
```

```
## 640
##  44
```

```
tapply(TOTCHG,diagg,sum)
```

```
##     21     23     49     50     51     53     54     57     58     92     97
##  10002  14174  20195   3908   3023  82271    851  14509   2117  12024   9530
##    114    115    137    138    139    141    143    204    206    225    249
##  10562  25832  15129  13622  17766   2860   1393   8439   9230  25649  16642
##    254    308    313    317    344    347    420    421    422    560    561
##    615  10585   8159  17524  14802  12597   6357  26356   5177   4877   2296
##    566    580    581    602    614    626    633    634    636    639    640
##   2129   2825   7453  29188  27531  23289  17591   9952  23224  12612 437978
##    710    720    723    740    750    751    753    754    755    756    758
##   8223  14243   5289  11125   1753  21666  79542  59150  11168   1494  34953
##    760    776    811    812    863    911    930    952
##   8273   1193   3838   9524  13040  48388  26654   4833
```

```
which.max(tapply(TOTCHG,diagg,sum))
```

```
## 640
##  44
```

```
max(tapply(TOTCHG,diagg,sum))
```

```
## [1] 437978
```

```
#From the results we can see that the category 640 has the maximum entries of hospitalization
#and also has the highest total hospitalization cost (437978).

#h0:The race of the patient is related to the hospitalization costs.
#ha:no relation

rc<-as.factor(RACE)
summary(rc)
```

```
##    1    2    3    4    5    6 NA's
##  484    6    1    3    3    2    1
```

```
#now to omit na values from data set
hospna<-na.omit(hosp)

modelannova<-aov(TOTCHG~RACE)
options(scipen = 999)
summary(modelannova)
```

```
##              Df     Sum Sq  Mean Sq F value Pr(>F)
## RACE          1    2488459  2488459   0.164  0.686
## Residuals   497 7539623326 15170268
## 1 observation deleted due to missingness
```

```
#p-value comes out to be very high 68% this means we can take risk and reject the null hypoth
esis
#this means  there is no relation between the race of patient and the hospital cost.

modelm1<-lm(TOTCHG~AGE+FEMALE)
summary(modelm1)
```

```
##
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3406  -1443   -869   -152  44951
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  2718.63     261.14  10.411 < 0.0000000000000002 ***
## AGE            86.28      25.48   3.387             0.000763 ***
## FEMALE       -748.19     353.83  -2.115             0.034967 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3845 on 497 degrees of freedom
## Multiple R-squared:  0.0261, Adjusted R-squared:  0.02218
## F-statistic:  6.66 on 2 and 497 DF,  p-value: 0.001399
```

```
#p-value for age is very less this means it is a  important factor in the hospital costs as s
een by the significance levels and p-values
#gender has also less p value means it is also having the impact on cost and same with interc
ept

modelm2<-lm(LOS~AGE+FEMALE+RACE)
summary(modelm2)
```

```
##
## Call:
## lm(formula = LOS ~ AGE + FEMALE + RACE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3.22  -1.22  -0.85   0.15  37.78
##
## Coefficients:
##             Estimate Std. Error t value          Pr(>|t|)
## (Intercept)  2.94377    0.39318   7.487 0.000000000000325 ***
## AGE         -0.03960    0.02231  -1.775            0.0766 .
## FEMALE       0.37011    0.31024   1.193            0.2334
## RACE        -0.09408    0.29312  -0.321            0.7484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.363 on 495 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.007898,   Adjusted R-squared:  0.001886
## F-statistic: 1.314 on 3 and 495 DF,  p-value: 0.2692
```

```
#except for the intercept.
#The very high p-value signifies that there is no linear relationship between the given varia
bles.
#That is, with just the age, gender, and race, it is not possible to predict the los of a pat
ient

modelm3<-lm(TOTCHG~ .,data=hospna)
summary(modelm3)
```

```
##
## Call:
## lm(formula = TOTCHG ~ ., data = hospna)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6377   -700   -174    122  43378
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 5218.6769   507.6475  10.280 < 0.0000000000000002 ***
## AGE          134.6949    17.4711   7.710   0.0000000000000702 ***
## FEMALE      -390.6924   247.7390  -1.577                0.115
## LOS          743.1521    34.9225  21.280 < 0.0000000000000002 ***
## RACE        -212.4291   227.9326  -0.932                0.352
## APRDRG        -7.7909     0.6816 -11.430 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2613 on 493 degrees of freedom
## Multiple R-squared:  0.5536, Adjusted R-squared:  0.5491
## F-statistic: 122.3 on 5 and 493 DF,  p-value: < 0.00000000000000022
```

```
#APRDRG also affect
#We can see that age and length of stay affect the total hospital cost.
```