Healthcare cost analysis

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

```
hosp_cost<-read.table(file.choose(),sep=",",header=TRUE)

head(hosp_cost)

summary(hosp_cost)


head(hosp_cost$AGE)

summary(hosp_cost$AGE)

table(hosp_cost$AGE)

hist(hosp_cost$AGE)

summary(as.factor(hosp_cost$AGE))


max(table(hosp_cost$AGE))

max(summary(as.factor(hosp_cost$AGE)))


which.max(table(hosp_cost$AGE))

age <- aggregate(TOTCHG ~ AGE, data = hosp_cost, sum)

max(age)
```

#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.

```
t <- table(hosp_cost$APRDRG)
```

```
d <- as.data.frame(t)

names(d)[1] = 'Diagnosis Group'

d


which.max(table(hosp_cost$APRDRG))

which.max(t)

which.max(d)


res <- aggregate(TOTCHG ~ APRDRG, data = hosp_cost, sum)

res

which.max(res$TOTCHG)

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


#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

table(hosp_cost$RACE)

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

fit <- lm(TOTCHG ~ RACE,data=hosp_cost)

fit

summary(fit)


fit1 <- aov(TOTCHG ~ RACE,data=hosp_cost)

summary(fit1)

hosp_cost <- na.omit(hosp_cost)
```

#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.

```
table(hosp_cost$FEMALE)

a <- aov(TOTCHG ~ AGE+FEMALE,data=hosp_cost)

summary(a)

b <- lm(TOTCHG ~ AGE+FEMALE,data=hosp_cost)

summary(b)
```

#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.

```
table(hosp_cost$LOS)

cat <- aov(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)

summary(cat)

cat <- lm(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)

summary(cat)
```

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

```
aov(TOTCHG ~.,data=hosp_cost)

mod <- lm(TOTCHG ~ .,data=hosp_cost)

summary(mod)
```

```
#plot of residual error between observed values and fitted regression line

#To see how AGE,LOS, and APRDRG affects Total charges of discharge.


library(tidyverse)

library(broom)

m<-augment(mod)

head(m)

library(ggplot2)


ggplot(m,aes(LOS,TOTCHG))+

 geom_point()+

 stat_smooth(method = lm,se=FALSE)+

 geom_segment(aes(xend=LOS,yend=.fitted),color="red",size=0.3)



ggplot(m,aes(AGE,TOTCHG))+

 geom_point()+

 stat_smooth(method = lm,se=FALSE)+

 geom_segment(aes(xend=AGE,yend=.fitted),color="red",size=0.3)



ggplot(m,aes(APRDRG,TOTCHG))+

 geom_point()+

 stat_smooth(method = lm,se=FALSE)+

 geom_segment(aes(xend=APRDRG,yend=.fitted),color="red",size=0.3)
```

……OUTPUT…….

```
hosp_cost <- read.csv("C:/Users/Harshada/Downloads/Healthcare-Cost-Analysis-of-Wisconsin-Hospital-master/Healthcare-Cost-Analysis-of-Wisconsin-Hospital-master/HospitalCosts.csv")

>  View(hosp_cost)

> head(hosp_cost)

  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

4  17     1   1    1    736    758

5  17     1   1    1   1194    754

6  17     0   0    1   3305    347

> summary(hosp_cost)

     AGE          FEMALE         LOS          RACE         TOTCHG

 Min.   : 0.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000   Min.   :  532

 1st Qu.: 0.000   1st Qu.:0.000   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 1216

 Median : 0.000   Median :1.000   Median : 2.000   Median :1.000   Median : 1536

 Mean   : 5.086   Mean   :0.512   Mean   : 2.828   Mean   :1.078   Mean   : 2774

 3rd Qu.:13.000   3rd Qu.:1.000   3rd Qu.: 3.000   3rd Qu.:1.000   3rd Qu.: 2530

 Max.   :17.000   Max.   :1.000   Max.   :41.000   Max.   :6.000   Max.   :48388

                                                   NA's   :1

    APRDRG

 Min.   : 21.0

 1st Qu.:640.0

 Median :640.0

 Mean   :616.4
```

```
 3rd Qu.:751.0

 Max.   :952.0


> head(hosp_cost$AGE)

[1] 17 17 17 17 17 17

> summary(hosp_cost$AGE)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

 0.000   0.000   0.000   5.086  13.000  17.000

> table(hosp_cost$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

> hist(hosp_cost$AGE)

> summary(as.factor(hosp_cost$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

> max(table(hosp_cost$AGE))

[1] 307

> max(summary(as.factor(hosp_cost$AGE)))

[1] 307

> which.max(table(hosp_cost$AGE))

0

1

> age <- aggregate(TOTCHG ~ AGE, data = hosp_cost, sum)

> max(age)
```

```
[1] 678118

> t <- table(hosp_cost$APRDRG)

> d <- as.data.frame(t)

> names(d)[1] = 'Diagnosis Group'

> d

   Diagnosis Group Freq
1               21    1
2               23    1
3               49    1
4               50    1
5               51    1
6               53   10
7               54    1
8               57    2
9               58    1
10              92    1
11              97    1
12             114    1
13             115    2
14             137    1
15             138    4
16             139    5
17             141    1
18             143    1
19             204    1
```

20	206	1
21	225	2
22	249	6
23	254	1
24	308	1
25	313	1
26	317	1
27	344	2
28	347	3
29	420	2
30	421	1
31	422	3
32	560	2
33	561	1
34	566	1
35	580	1
36	581	3
37	602	1
38	614	3
39	626	6
40	633	4
41	634	2
42	636	3
43	639	4
44	640	267

```
45          710   1
46          720   1
47          723   2
48          740   1
49          750   1
50          751  14
51          753  36
52          754  37
53          755  13
54          756   2
55          758  20
56          760   2
57          776   1
58          811   2
59          812   3
60          863   1
61          911   1
62          930   2
63          952   1
> which.max(table(hosp_cost$APRDRG))
640
 44
> which.max(t)
640
 44
```

```
> which.max(d)
Error in which.max(d) : 'list' object cannot be coerced to type 'double'
> res <- aggregate(TOTCHG ~ APRDRG, data = hosp_cost, sum)
> res
   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

> which.max(res$TOTCHG)

[1] 44

> res[which.max(res$TOTCHG),]

APRDRG TOTCHG

44 640 437978

> #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

> table(hosp_cost$RACE)

```
 1  2  3  4  5  6
484  6  1  3  3  2
> hosp_cost$RACE <- as.factor(hosp_cost$RACE)
> fit <- lm(TOTCHG ~ RACE,data=hosp_cost)
> fit


Call:
lm(formula = TOTCHG ~ RACE, data = hosp_cost)


Coefficients:
(Intercept)       RACE2       RACE3       RACE4       RACE5       RACE6
    2772.7      1429.5       268.3      -428.0      -746.0     -1423.7


> summary(fit)


Call:
lm(formula = TOTCHG ~ RACE, data = hosp_cost)


Residuals:
   Min     1Q Median     3Q    Max
 -3049  -1551  -1223   -238  45615


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2772.7     177.6  15.615   <2e-16 ***
RACE2        1429.5    1604.7   0.891    0.373
RACE3         268.3    3910.5   0.069    0.945
RACE4        -428.0    2262.4  -0.189    0.850
RACE5        -746.0    2262.4  -0.330    0.742
RACE6       -1423.7    2768.0  -0.514    0.607
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3906 on 493 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.002465,  Adjusted R-squared:  -0.007652
F-statistic: 0.2437 on 5 and 493 DF,  p-value: 0.9429

> fit1 <- aov(TOTCHG ~ RACE,data=hosp_cost)
> summary(fit1)
             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
1 observation deleted due to missingness
> hosp_cost <- na.omit(hosp_cost)
> table(hosp_cost$FEMALE)

  0   1
244 255
```

```
> a <- aov(TOTCHG ~ AGE+FEMALE,data=hosp_cost)
> summary(a)
             Df    Sum Sq   Mean Sq F value  Pr(>F)
AGE           1 1.297e+08 129749266   8.759 0.00323 **
FEMALE        1 6.522e+07  65219972   4.403 0.03638 *
Residuals   496 7.347e+09  14812787
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> b <- lm(TOTCHG ~ AGE+FEMALE,data=hosp_cost)
> summary(b)

Call:
lm(formula = TOTCHG ~ AGE + FEMALE, data = hosp_cost)

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 ***
FEMALE       -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

> table(hosp_cost$LOS)

  0   1   2   3   4   5   6   7   8   9  10  12  15  17  18  23  24  39  41
 15  79 223  98  38  14   8  11   1   1   1   2   1   1   2   1   1   1   1
> cat <- aov(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)
> summary(cat)
             Df Sum Sq Mean Sq F value Pr(>F)
AGE           1     27  26.907   2.361  0.125
FEMALE        1     17  16.510   1.449  0.229
RACE          5      6   1.138   0.100  0.992
Residuals   491   5595  11.396
> cat <- lm(LOS ~ AGE+FEMALE+RACE,data=hosp_cost)
> summary(cat)

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

Residuals:
    Min      1Q  Median      3Q     Max
-3.2110 -1.2110 -0.8570  0.1430 37.7890
```

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	.
FEMALE	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

```
> aov(TOTCHG ~.,data=hosp_cost)
Call:
  aov(formula = TOTCHG ~ ., data = hosp_cost)

Terms:
                      AGE    FEMALE       LOS      RACE    APRDRG  Residuals
Sum of Squares  129749266  65219972 3086194093  13244291 887028136 3360676025
```

```
Deg. of Freedom        1        1        1        5        1      489

Residual standard error: 2621.555
Estimated effects may be unbalanced
> mod <- lm(TOTCHG ~ .,data=hosp_cost)
> summary(mod)

Call:
lm(formula = TOTCHG ~ ., data = hosp_cost)

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 ***
FEMALE       -392.5778   249.2981  -1.575    0.116
LOS           742.9637    35.0464  21.199  < 2e-16 ***
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
```

```
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

> library(tidyverse)
> library(broom)
> m<-augment(mod)
> head(m)
# A tibble: 6 x 13
  .rownames TOTCHG   AGE FEMALE   LOS RACE  APRDRG .fitted .resid .std.resid    .hat
  <chr>     <int> <int>  <int> <int> <fct>  <int>   <dbl>  <dbl>     <dbl>   <dbl>
1 1          2660    17     1     2 1        560   4005. -1345.     -0.516 0.00991
2 2          1689    17     0     2 1        753   2889. -1200.     -0.461 0.0141
3 3         20060    17     1     7 1        930   4828. 15232.      5.86  0.0165
4 4           736    17     1     1 1        758   1714.  -978.     -0.375 0.00958
5 5          1194    17     1     1 1        754   1746.  -552.     -0.211 0.00954
6 6          3305    17     0     0 1        347   4577. -1272.     -0.490 0.0180
# ... with 2 more variables: .sigma <dbl>, .cooksd <dbl>
> library(ggplot2)
> ggplot(m,aes(LOS,TOTCHG))+
+   geom_point()+
```

```
+   stat_smooth(method = lm,se=FALSE)+
+   geom_segment(aes(xend=LOS,yend=.fitted),color="red",size=0.3)
`geom_smooth()` using formula 'y ~ x'
> ggplot(m,aes(AGE,TOTCHG))+
+   geom_point()+
+   stat_smooth(method = lm,se=FALSE)+
+   geom_segment(aes(xend=AGE,yend=.fitted),color="red",size=0.3)
`geom_smooth()` using formula 'y ~ x'
> ggplot(m,aes(APRDRG,TOTCHG))+
+   geom_point()+
+   stat_smooth(method = lm,se=FALSE)+
+   geom_segment(aes(xend=APRDRG,yend=.fitted),color="red",size=0.3)
`geom_smooth()` using formula 'y ~ x'
>
```