Healthcare

Tori Green

11/27/2019

```
library(readxl)
hospital1 <- read_excel("~/Desktop/Simplilearn/Data Science with R/1555054100_hospitalcosts (1).xlsx")
```

#View the structure of the dataset

```
str(hospital1)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    500 obs. of  6 variables:
##  $ AGE   : num  17 17 17 17 17 17 17 16 16 17 ...
##  $ FEMALE: num  1 0 1 1 1 0 1 1 1 1 ...
##  $ LOS   : num  2 2 7 1 1 0 4 2 1 2 ...
##  $ RACE  : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ TOTCHG: num  2660 1689 20060 736 1194 ...
##  $ APRDRG: num  560 753 930 758 754 347 754 754 753 758 ...
```

#View the first few rows of dataset

```
head(hospital1)
```

```
## # 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
```

#Descriptive statistics of each variable.

```
summary(hospital1)
```

```
##       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
##
```

```
library(psych)
describe(hospital1)
```

```
##         vars   n    mean      sd median trimmed    mad min   max range
## AGE        1 500    5.09    6.95    0.0    4.26   0.00   0    17    17
## FEMALE     2 500    0.51    0.50    1.0    0.52   0.00   0     1     1
## LOS        3 500    2.83    3.36    2.0    2.31   1.48   0    41    41
## RACE       4 499    1.08    0.51    1.0    1.00   0.00   1     6     5
## TOTCHG     5 500 2774.39 3888.41 1536.5 1894.24 646.41 532 48388 47856
## APRDRG     6 500  616.36  178.32  640.0  656.37   0.00  21   952   931
##          skew kurtosis     se
## AGE      0.73    -1.34   0.31
## FEMALE -0.05    -2.00   0.02
## LOS      7.09    65.68   0.15
## RACE     7.43    57.52   0.02
## TOTCHG   5.53    45.97 173.89
## APRDRG -2.02     3.55   7.97
```

1.1 Mean total hospital discharge costs by age

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------ tidyvers
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

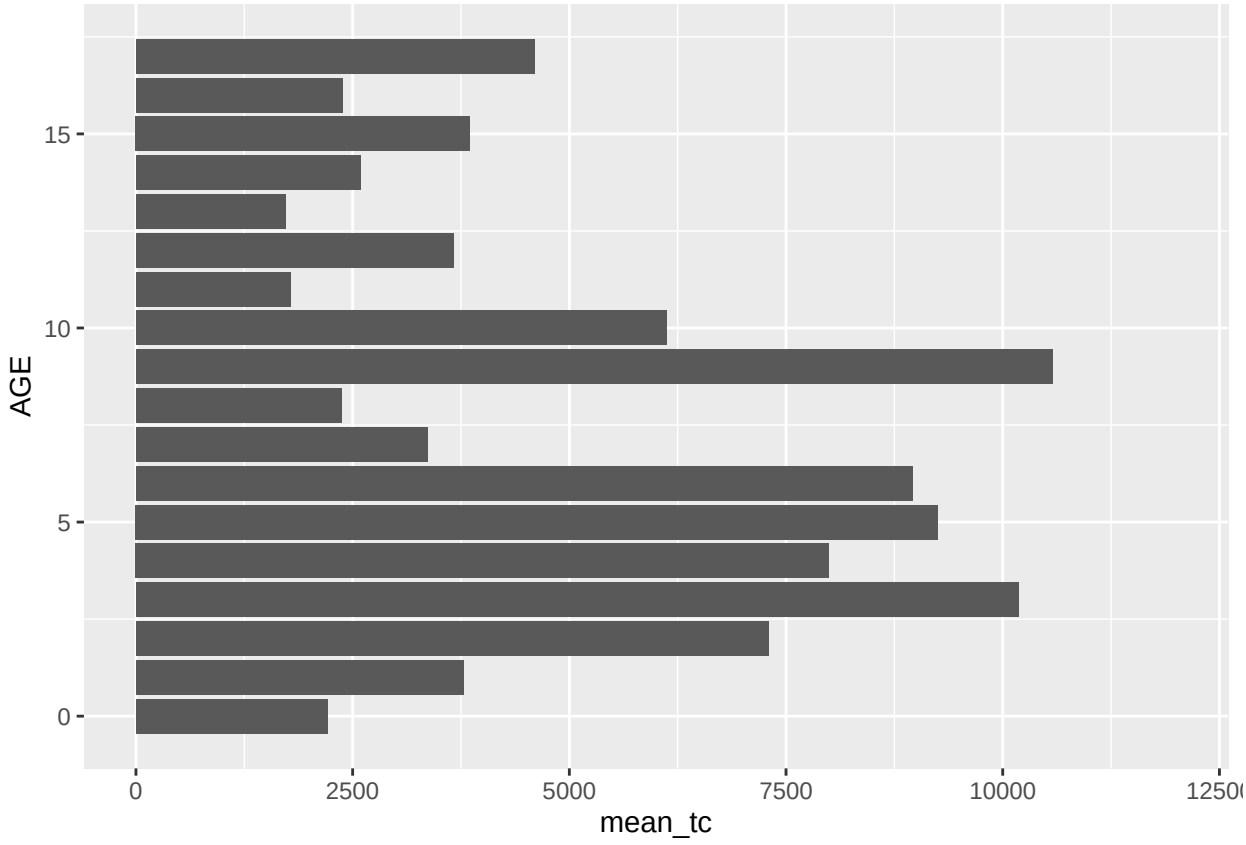
```
## -- Conflicts --------------------------------------------------------------------------- tidyverse_con
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
data2 <-hospital1  %>%
    group_by(AGE) %>%
    summarise(
        n = n(),
        mean_tc = mean(TOTCHG, na.rm=T))
print(data2)
```

```
## # A tibble: 18 x 3
##      AGE     n mean_tc
##    <dbl> <int>   <dbl>
##  1     0   307   2209.
##  2     1    10   3774.
##  3     2     1   7298
##  4     3     3  10183.
##  5     4     2   7996
##  6     5     2   9254.
##  7     6     2   8964
##  8     7     3   3362.
##  9     8     2   2370.
## 10     9     2  10574.
## 11    10     4   6117.
## 12    11     8   1781.
## 13    12    15   3661.
## 14    13    18   1730.
## 15    14    25   2586.
## 16    15    29   3853.
## 17    16    29   2384.
## 18    17    38   4599.
```

#Graph 1.1 Total Costs (Mean Value) By Age of Patient

```
library(ggplot2)
ggplot(data = data2, aes(x=AGE, y= mean_tc)) +geom_bar(stat = "identity", position="dodge") + scale_y_c
```



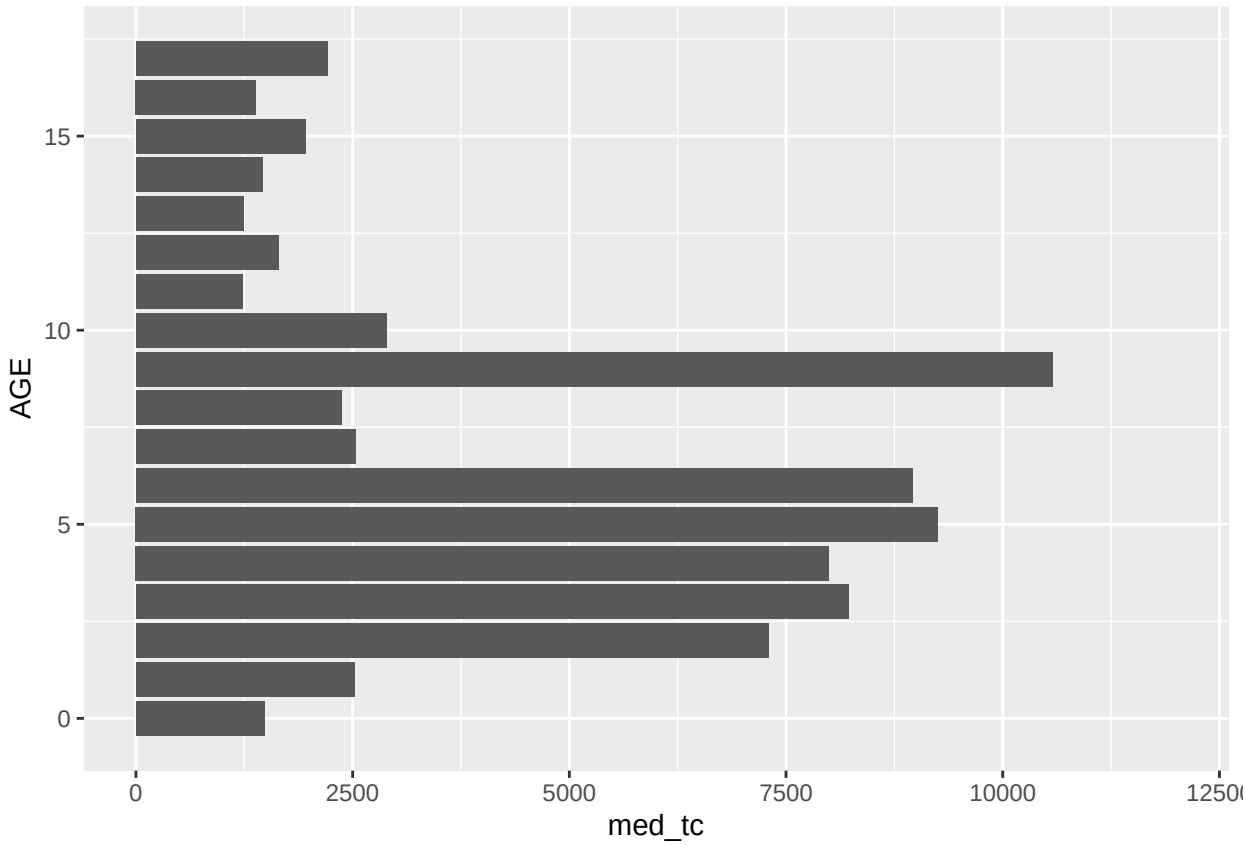
#1.2 Median total hospital discharge costs by age

```
data1 <-hospital1  %>%
    group_by(AGE) %>%
    summarise(
        n = n(),
        med_tc = median(TOTCHG, na.rm=T))
print(data1)
```

```
## # A tibble: 18 x 3
##      AGE     n med_tc
##    <dbl> <int>  <dbl>
##  1     0   307  1483
##  2     1    10  2524.
##  3     2     1  7298
##  4     3     3  8223
##  5     4     2  7996
##  6     5     2  9254.
##  7     6     2  8964
##  8     7     3  2530
##  9     8     2  2370.
## 10     9     2 10574.
## 11    10     4  2892.
## 12    11     8  1232.
## 13    12    15  1647
## 14    13    18  1242
## 15    14    25  1463
## 16    15    29  1956
## 17    16    29  1385
## 18    17    38  2211
```

#Graph 1.2 Total Costs (Median Value) By Age of Patient

```
library(ggplot2)
ggplot(data = data1, aes(x=AGE, y= med_tc)) +geom_bar(stat = "identity", position="dodge") + scale_y_co
```



#1.3 Total Costs (Total Sum) By Age of Patient

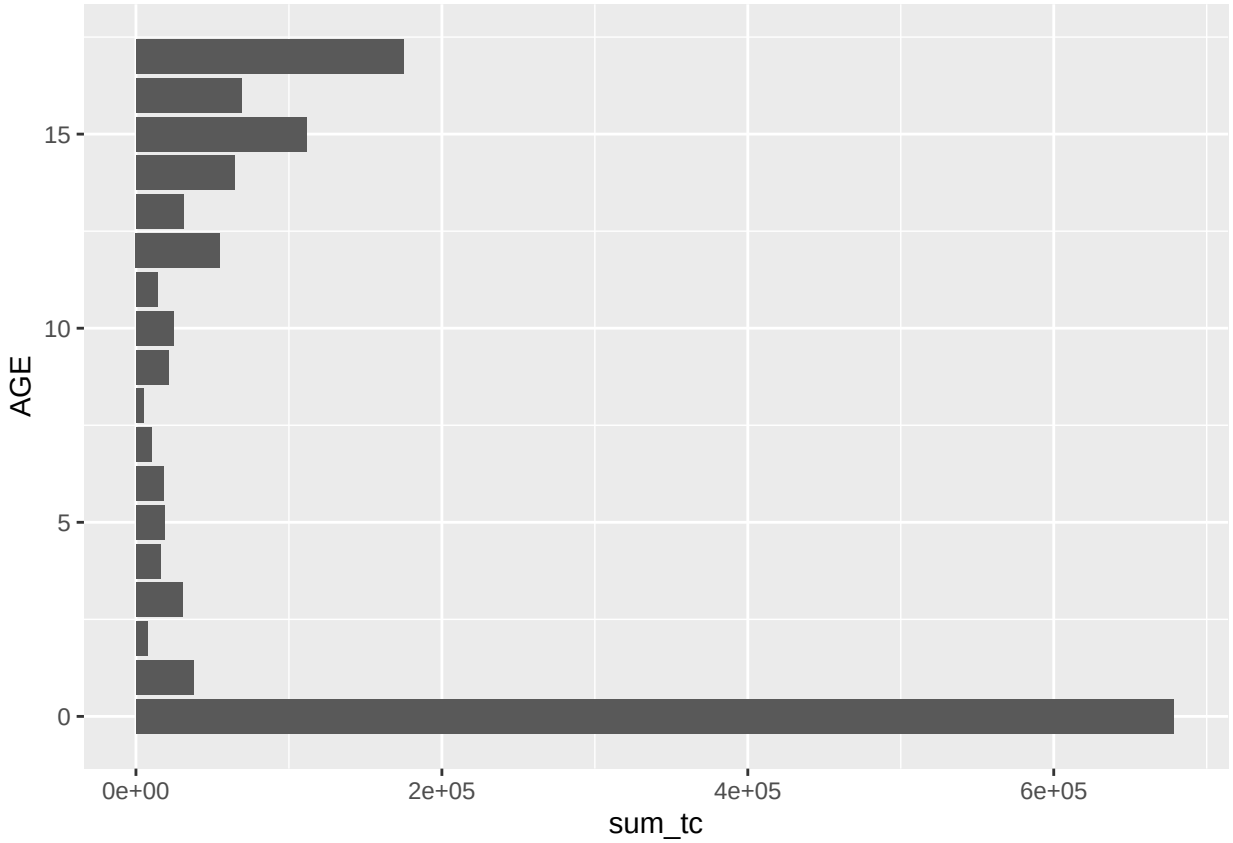
```
data3 <-hospital1  %>%
    group_by(AGE) %>%
    summarise(
        n = n(),
        sum_tc = sum(TOTCHG, na.rm=T))
print(data3)
```

```
## # A tibble: 18 x 3
##      AGE     n sum_tc
##    <dbl> <int>  <dbl>
##  1     0   307 678118
##  2     1    10  37744
##  3     2     1   7298
##  4     3     3  30550
##  5     4     2  15992
##  6     5     2  18507
##  7     6     2  17928
##  8     7     3  10087
##  9     8     2   4741
## 10     9     2  21147
## 11    10     4  24469
## 12    11     8  14250
## 13    12    15  54912
## 14    13    18  31135
## 15    14    25  64643
```

```
## 16    15    29 111747
## 17    16    29  69149
## 18    17    38 174777
```

#Graph 1.3 Total Costs (Total Sum) By Age of Patient

```
library(ggplot2)
ggplot(data = data3, aes(x=AGE, y= sum_tc)) +geom_bar(stat = "identity", position="dodge") + scale_y_co
```



#2

#2 Diagnosis-related group that has maximum average total hospital cost

```
totdata1 <-hospital1  %>%
    group_by(APRDRG) %>%
    summarise(
        n = n(),
        mean_tcexp = mean(TOTCHG, na.rm=T))
# Sort by mean_tcexp
data.frame(totdata1)
```

```
##    APRDRG   n mean_tcexp
## 1      21   1 10002.0000
## 2      23   1 14174.0000
## 3      49   1 20195.0000
## 4      50   1  3908.0000
## 5      51   1  3023.0000
```

```
## 6      53  10  8227.1000
## 7      54   1   851.0000
## 8      57   2  7254.5000
## 9      58   1  2117.0000
## 10     92   1 12024.0000
## 11     97   1  9530.0000
## 12    114   1 10562.0000
## 13    115   2 12916.0000
## 14    137   1 15129.0000
## 15    138   4  3405.5000
## 16    139   5  3553.2000
## 17    141   1  2860.0000
## 18    143   1  1393.0000
## 19    204   1  8439.0000
## 20    206   1  9230.0000
## 21    225   2 12824.5000
## 22    249   6  2773.6667
## 23    254   1   615.0000
## 24    308   1 10585.0000
## 25    313   1  8159.0000
## 26    317   1 17524.0000
## 27    344   2  7401.0000
## 28    347   3  4199.0000
## 29    420   2  3178.5000
## 30    421   1 26356.0000
## 31    422   3  1725.6667
## 32    560   2  2438.5000
## 33    561   1  2296.0000
## 34    566   1  2129.0000
## 35    580   1  2825.0000
## 36    581   3  2484.3333
## 37    602   1 29188.0000
## 38    614   3  9177.0000
## 39    626   6  3881.5000
## 40    633   4  4397.7500
## 41    634   2  4976.0000
## 42    636   3  7741.3333
## 43    639   4  3153.0000
## 44    640 267  1640.3670
## 45    710   1  8223.0000
## 46    720   1 14243.0000
## 47    723   2  2644.5000
## 48    740   1 11125.0000
## 49    750   1  1753.0000
## 50    751  14  1547.5714
## 51    753  36  2209.5000
## 52    754  37  1598.6486
## 53    755  13   859.0769
## 54    756   2   747.0000
## 55    758  20  1747.6500
## 56    760   2  4136.5000
## 57    776   1  1193.0000
## 58    811   2  1919.0000
## 59    812   3  3174.6667
```

```
## 60    863   1 13040.0000
## 61    911   1 48388.0000
## 62    930   2 13327.0000
## 63    952   1  4833.0000
```

```
totdata1[order(-totdata1$mean_tcexp),]
```

```
## # A tibble: 63 x 3
##    APRDRG     n mean_tcexp
##     <dbl> <int>      <dbl>
##  1    911     1      48388
##  2    602     1      29188
##  3    421     1      26356
##  4     49     1      20195
##  5    317     1      17524
##  6    137     1      15129
##  7    720     1      14243
##  8     23     1      14174
##  9    930     2      13327
## 10    863     1      13040
## # ... with 53 more rows
```

#2.2 Diagnosis-related group that has maximum hospitalization

```
losdata1 <-hospital1  %>%
    group_by(APRDRG) %>%
    summarise(
        n = n(),
        mean_los = mean(LOS, na.rm=T))
# Sort by mean_los
data.frame(losdata1)
```

```
##    APRDRG   n   mean_los
## 1      21   1  2.0000000
## 2      23   1  2.0000000
## 3      49   1  6.0000000
## 4      50   1  2.0000000
## 5      51   1  3.0000000
## 6      53  10  2.9000000
## 7      54   1  1.0000000
## 8      57   2  1.0000000
## 9      58   1  1.0000000
## 10     92   1  1.0000000
## 11     97   1  3.0000000
## 12    114   1  3.0000000
## 13    115   2  4.0000000
## 14    137   1 12.0000000
## 15    138   4  1.5000000
## 16    139   5  1.4000000
## 17    141   1  2.0000000
## 18    143   1  2.0000000
## 19    204   1  3.0000000
## 20    206   1  3.0000000
```

```
## 21    225   2  5.5000000
## 22    249   6  1.3333333
## 23    254   1  0.0000000
## 24    308   1  1.0000000
## 25    313   1  1.0000000
## 26    317   1  7.0000000
## 27    344   2  2.5000000
## 28    347   3  0.3333333
## 29    420   2  1.5000000
## 30    421   1 39.0000000
## 31    422   3  1.6666667
## 32    560   2  2.0000000
## 33    561   1  4.0000000
## 34    566   1  2.0000000
## 35    580   1  1.0000000
## 36    581   3  0.3333333
## 37    602   1 41.0000000
## 38    614   3 15.0000000
## 39    626   6  6.0000000
## 40    633   4  6.2500000
## 41    634   2  4.5000000
## 42    636   3  6.6666667
## 43    639   4  4.5000000
## 44    640 267  2.4419476
## 45    710   1  4.0000000
## 46    720   1  5.0000000
## 47    723   2  1.0000000
## 48    740   1  5.0000000
## 49    750   1  3.0000000
## 50    751  14  2.2857143
## 51    753  36  3.0277778
## 52    754  37  2.2972973
## 53    755  13  1.0000000
## 54    756   2  1.0000000
## 55    758  20  2.6500000
## 56    760   2  7.0000000
## 57    776   1  2.0000000
## 58    811   2  0.5000000
## 59    812   3  1.6666667
## 60    863   1 24.0000000
## 61    911   1  7.0000000
## 62    930   2  5.0000000
## 63    952   1  1.0000000
```

```
losdata1[order(-losdata1$mean_los),]
```

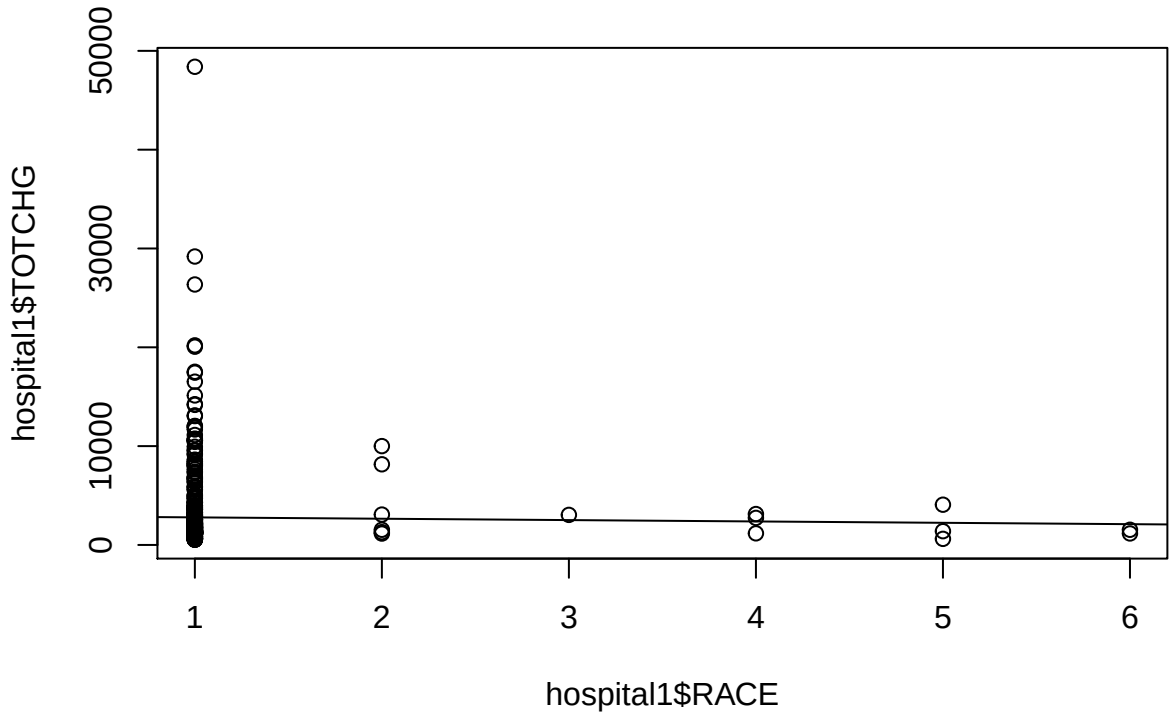
```
## # A tibble: 63 x 3
##    APRDRG     n mean_los
##     <dbl> <int>    <dbl>
##  1    602     1    41
##  2    421     1    39
##  3    863     1    24
##  4    614     3    15
##  5    137     1    12
```

```
##  6    317     1     7
##  7    760     2     7
##  8    911     1     7
##  9    636     3     6.67
## 10    633     4     6.25
## # ... with 53 more rows
```

#3

#3.1 #Create the scatter plot with RACE on the x-axis and TOTCHG on the y-axis #Add a regression linew ith the form abline(lm(y ~ x))

```
plot(hospital1$RACE, hospital1$TOTCHG)
abline (lm(hospital1$TOTCHG ~ hospital1$RACE))
```



#3.2 Calculate the correlation between RACE AND TOTCHG

```
cor(hospital1$TOTCHG, hospital1$RACE)
```

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

#3.3 Linear regression modeling for race and total charge

```
lrRACE <-lm(TOTCHG ~ RACE, data=hospital1)
summary(lrRACE)
```

```
##
## Call:
```

```
## lm(formula = TOTCHG ~ RACE, data = hospital1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2256  -1560  -1227   -258  45600
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2925.7      405.0   7.224 1.92e-12 ***
## RACE          -137.3      339.1  -0.405    0.686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3895 on 497 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.0003299,  Adjusted R-squared:  -0.001681
## F-statistic: 0.164 on 1 and 497 DF,  p-value: 0.6856
```

#4

#4.1 Determine the correlation among each of the variables.

```
cor_data <- hospital1[,1:length(hospital1)]
round(cor(cor_data),2)
```

```
##          AGE FEMALE   LOS RACE TOTCHG APRDRG
## AGE     1.00   0.23 -0.07   NA   0.13   0.15
## FEMALE  0.23   1.00  0.04   NA  -0.06   0.25
## LOS    -0.07   0.04  1.00   NA   0.62   0.01
## RACE      NA     NA    NA    1     NA     NA
## TOTCHG  0.13  -0.06  0.62   NA   1.00  -0.33
## APRDRG  0.15   0.25  0.01   NA  -0.33   1.00
```

#4.2 Multiple Regression with total costs as the dependent variable and Age + Gender as the independent variables

```
lr4 <-lm(TOTCHG ~ AGE + FEMALE, data=hospital1)
summary(lr4)
```

```
##
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE, data = hospital1)
##
## 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  < 2e-16 ***
## 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
```

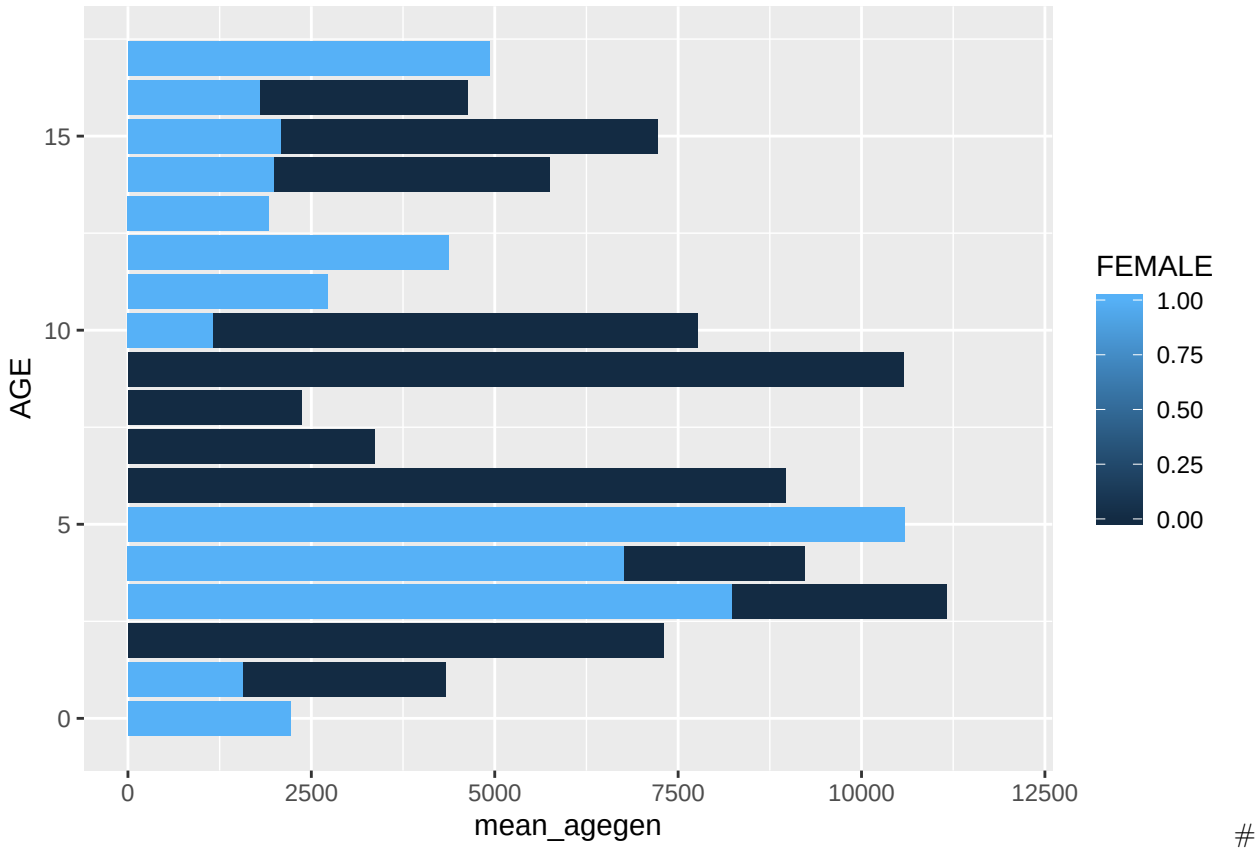
#Table 4.1 Mean of total charges subset by age and gender

```
data4 <-hospital1  %>%
    group_by(AGE,FEMALE) %>%
    summarise(
        n = n(),
        mean_agegen = mean(TOTCHG, na.rm=T))
print(data4)
```

```
## # A tibble: 31 x 4
## # Groups:   AGE [18]
##       AGE FEMALE     n mean_agegen
##     <dbl>  <dbl> <int>       <dbl>
##  1     0      0   170       2198.
##  2     0      1   137       2222.
##  3     1      0     8       4328.
##  4     1      1     2       1561
##  5     2      0     1       7298
##  6     3      0     2      11164.
##  7     3      1     1       8223
##  8     4      0     1       9230
##  9     4      1     1       6762
## 10     5      0     1       7923
## # ... with 21 more rows
```

#Graph 4.1 Mean of total charges subset by age and gender

```
library(ggplot2)
ggplot(data = data4, aes(x=AGE, y= mean_agegen, fill = FEMALE)) +geom_bar(stat = "identity", position="
```



#

#5

#5 Multiple linear regression with age, gender, and race as the independent/explanatory variables and length of stay as the dependent/response variable.

```
model <- lm(LOS ~ AGE + FEMALE + RACE, data = hospital1)
summary(model)
```

```
##
## Call:
## lm(formula = LOS ~ AGE + FEMALE + RACE, data = hospital1)
##
## 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 3.25e-13 ***
## 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
```

#6

#Model 6a. Multiple Regression Analysis with total charge as the dependent/response variable and the remaining variables as the independent or explanatory variables

```
model6a <- lm(TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = hospital1)
summary(model6a)
```

```
##
## Call:
## lm(formula = TOTCHG ~ AGE + FEMALE + RACE + LOS + APRDRG, data = hospital1)
##
## 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  < 2e-16 ***
## AGE          134.6949    17.4711   7.710 7.02e-14 ***
## FEMALE      -390.6924   247.7390  -1.577    0.115
## RACE        -212.4291   227.9326  -0.932    0.352
## LOS          743.1521    34.9225  21.280  < 2e-16 ***
## APRDRG        -7.7909     0.6816 -11.430  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2613 on 493 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5536, Adjusted R-squared:  0.5491
## F-statistic: 122.3 on 5 and 493 DF,  p-value: < 2.2e-16
```

#Figure 6.1 Relative Importance of Explanatory Variables

```
library(relaimpo)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: boot

##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:psych':
##
##     logit

## Loading required package: survey

## Loading required package: grid

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':
##
##     expand

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:boot':
##
##     aml

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart

## Loading required package: mitools

## This is the global version of package relaimpo.

## If you are a non-US user, a version with the interesting additional metric pmvd is available

## from Ulrike Groempings web site at prof.beuth-hochschule.de/groemping.
```

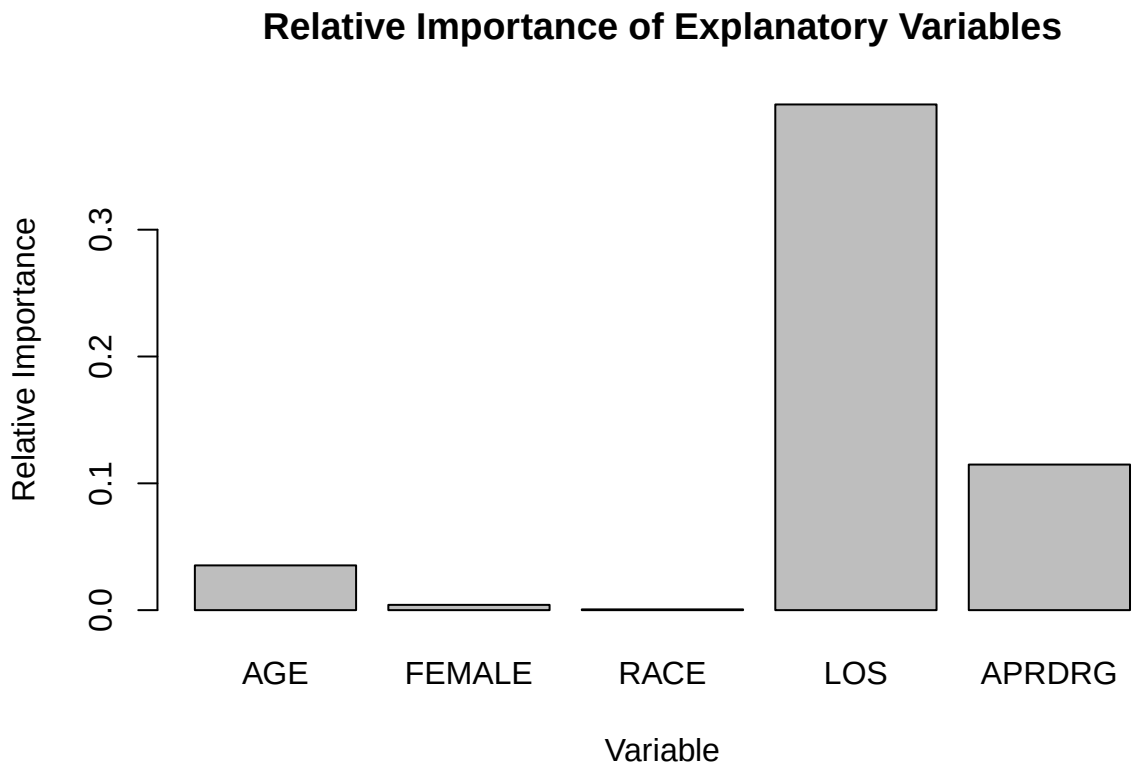
```
hospri = calc.relimp(model6a)
print(hospri)
```

```
## Response variable: TOTCHG
## Total response variance: 15144803
## Analysis based on 499 observations
##
## 5 Regressors:
## AGE FEMALE RACE LOS APRDRG
```

```
## Proportion of variance explained by model: 55.36%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##                 lmg
## AGE    0.0352884686
## FEMALE 0.0041653940
## RACE   0.0005803266
## LOS    0.3987475764
## APRDRG 0.1147960900
##
## Average coefficients for different model sizes:
##
##                 1X         2Xs         3Xs         4Xs         5Xs
## AGE      73.414846   90.138622  106.089194  121.017331  134.694873
## FEMALE -462.340678 -413.107026 -385.968482 -379.122207 -390.692432
## RACE   -137.327456 -152.082696 -169.607724 -190.002322 -212.429057
## LOS     720.301090  725.111508  730.791744  736.934830  743.152072
## APRDRG   -7.193932   -7.402635   -7.578275   -7.710881   -7.790948
```

#Graph 6.1 for Figure 6.1

```
barplot(hospri$lmg, ylab="Relative Importance", xlab="Variable", main = "Relative Importance of Explana
```



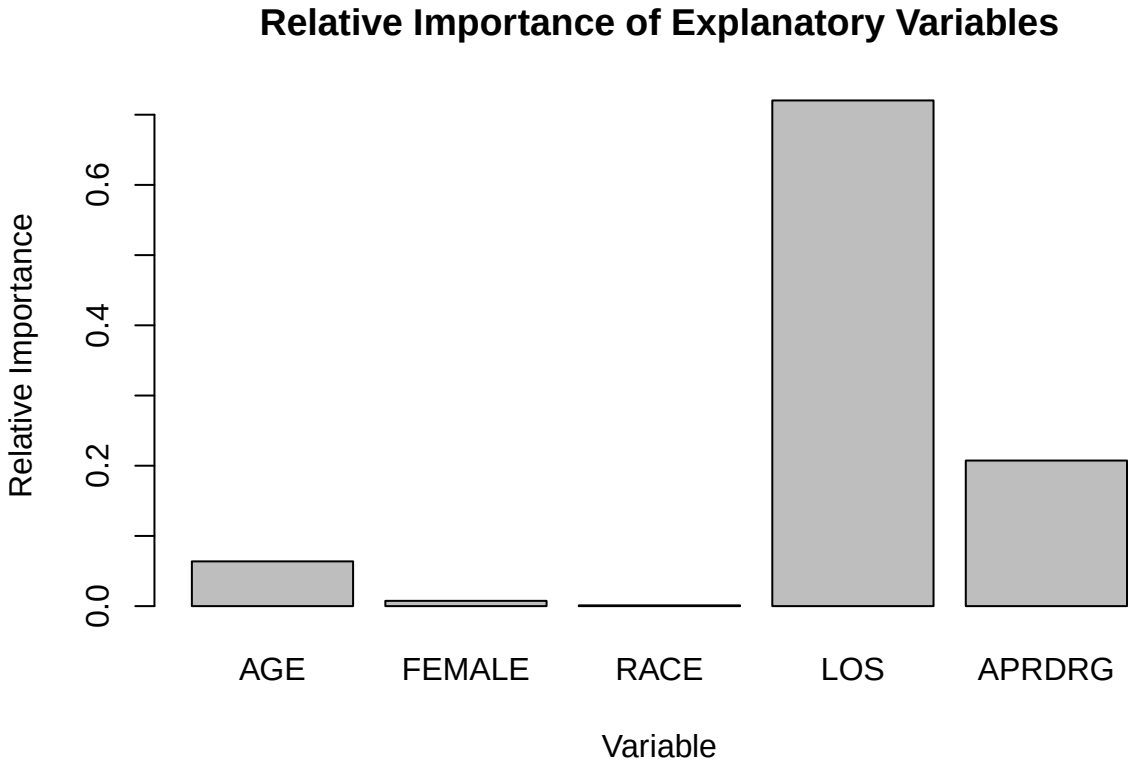
#6 Determine which variable mainly affects total costs. #Figure 6.2 Relative Importance of Explanatory Variables (Total sums to 100)

```r
hospri2 = calc.relimp(model6a, rela = TRUE)
print(hospri2)
```

```
## Response variable: TOTCHG
## Total response variance: 15144803
## Analysis based on 499 observations
##
## 5 Regressors:
## AGE FEMALE RACE LOS APRDRG
## Proportion of variance explained by model: 55.36%
## Metrics are normalized to sum to 100% (rela=TRUE).
##
## Relative importance metrics:
##
##                lmg
## AGE    0.063746171
## FEMALE 0.007524495
## RACE   0.001048320
## LOS    0.720309840
## APRDRG 0.207371174
##
## Average coefficients for different model sizes:
##
##                 1X         2Xs         3Xs         4Xs         5Xs
## AGE      73.414846   90.138622  106.089194  121.017331  134.694873
## FEMALE -462.340678 -413.107026 -385.968482 -379.122207 -390.692432
## RACE   -137.327456 -152.082696 -169.607724 -190.002322 -212.429057
## LOS     720.301090  725.111508  730.791744  736.934830  743.152072
## APRDRG   -7.193932   -7.402635   -7.578275   -7.710881   -7.790948
```

#Graph 6.2 for Figure 6.2

```r
barplot(hospri2$lmg, ylab="Relative Importance", xlab="Variable", main = "Relative Importance of Explana
```


Relative Importance of Explanatory Variables
Relative Importance
0.0
0.2
0.4
0.6
AGE
FEMALE
RACE
LOS
APRDRG
Variable