Healthcare Cost Analysis

Business Scenario: A nationwide survey of hospital costs conducted by the US Agency of Healthcare consists of hospital records of inpatient samples. “HospitalCosts.csv” is the data set given here with a subset of the data – restricted to the city of Wisconsin and relating to patients in the age group 0-17 years.

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

   To find the category that has the highest number of hospital visits can be found by graphical analysis. To begin, we read in our dataset and plot a histogram of the age distribution. The as.factor() is called to make sure that the AGE categories are not interpreted as numbers in the data summary.

   Code:

   #Read in and explore Dataset
   hosp <- read.csv("HospitalCosts.csv")
   #Attach the dataframe so it is automatically searched when declaring a variable
   attach(hosp)
   #Plot Age distribution
   hist(AGE, xlab = "Age (years)")
   summary(as.factor(AGE))

   Result:

   From the graph below, we see that infants have the maximum frequency of hospital visits with over 300. The summary output of age when displayed as a factor shows that there are 307 entries for those in the range of 0-1 years.

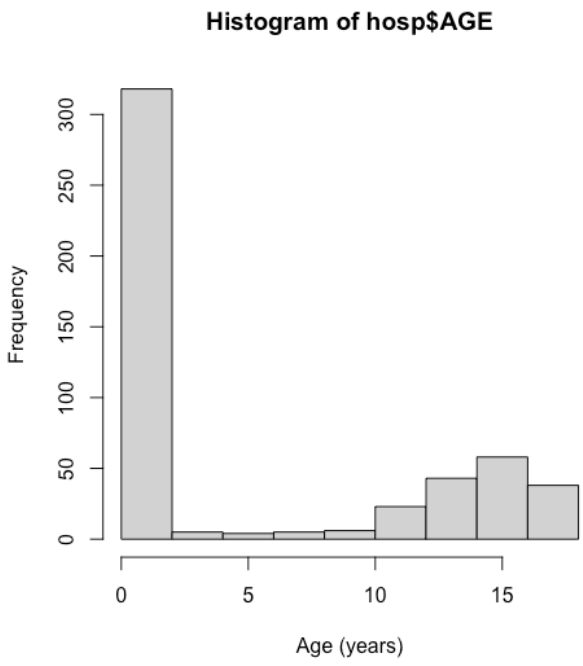
```
> hosp <- read.csv("HospitalCosts.csv")
> head(hosp)
  AGE FEMALE LOS RACE TOTCHG APRDRG
1  17      1   2    1   2660    560
2  17      0   2    1   1689    753
3  17      1   7    1  20060    930
4  17      1   1    1    736    758
5  17      1   1    1   1194    754
6  17      0   0    1   3305    347
> attach(hosp)
The following objects are masked _by_ .GlobalEnv:

    APRDRG, FEMALE, RACE

The following objects are masked from hosp (pos = 3):

    AGE, APRDRG, FEMALE, LOS, RACE, TOTCHG

> hist(AGE, xlab = "Age (years)")
> summary(as.factor(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
>
```



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.

   To find the diagnosis-related group that has the highest hospitalization, we can observe summary statistics. Since there are 63 different groups, graphing these is not ideal. Just like part (1), the as.factor() is called to make sure that the APRDRG categories are not interpreted as numbers in the data summary.

   Code:

   summary(as.factor(APRDRG))

   Result:

   From the summary output, we see that the diagnosis-related group 640 has the highest number of hospitalizations by a large margin (at 267), with the next highest being group 754 at 37 hospitalizations

```
> summary(as.factor(APRDRG))
 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
  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
344 347 420 421 422 560 561 566 580 581 602 614 626 633 634 636 639 640 710 720 723 740 750 751 753 754
  2   3   2   1   3   2   1   1   1   3   1   3   6   4   2   3   4 267   1   1   2   1   1  14  36  37
755 756 758 760 776 811 812 863 911 930 952
 13   2  20   2   1   2   3   1   1   2   1
>
```

3. 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.

   To try and determine if there is a relationship between the race of the patient and the cost of hospitalization, we will run a linear regression model and see if any of the categories are statistically significant to the model, which will indicate a correlation.

   Code:

   #Convert categorical variable from int type to Factor
   hosp$RACE <- factor(RACE)
   #Linear regression model to see the effect of race on total cost
   cost <- lm(TOTCHG ~ RACE, data = hosp)
   summary(cost)

   Result:

   According to the summary output from the linear regression, none of the race factors produced a p-value small enough to be significant enough to the model. However, our regression does show a large residual standard error, likely due to the fact that the largest race category was dropped in the model in order to prevent multicollinearity.

```
> cost <- lm(TOTCHG ~ RACE, data = hosp)
> summary(cost)

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

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
```

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

   To observe the severity of the hospital costs by age and gender, the analysis of variance method (ANOVA) is used. We compute the analysis of variance on the data print out a summary of the results.

   Code:

   #Compute the analysis of variance with hospital costs by age and gender
   res.aov <- aov(TOTCHG ~ AGE + FEMALE, data=hosp)
   summary(res.aov)

   Result:

   From the summary output below, we see that age has a greater correlation with the hospital costs than gender, so resource allocation should focus on age brackets that tend to have higher hospitalization costs.

```
> res.aov <- aov(TOTCHG ~ AGE + FEMALE, data=hosp)
> summary(res.aov)
             Df    Sum Sq   Mean Sq F value  Pr(>F)
AGE           1 1.308e+08 130822234   8.849 0.00308 **
FEMALE        1 6.610e+07  66104210   4.471 0.03497 *
Residuals   497 7.348e+09  14784325
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

5. 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.

   In order to determine whether length of stay can be predicted from age, gender, and race, we calculate an analysis of variance for each of these independent variables.

Code:

```
res2.aov <- aov(LOS ~ AGE + FEMALE + RACE, data=hosp)
summary(res2.aov)
```

Result:
From the result below, it would seem that age would be the greatest predictor of length of stay, followed by gender. Race seems to be the most insignificant when it comes to predicting the length of stay.

```
> res2.aov <- aov(LOS ~ AGE + FEMALE + RACE, data=hosp)
> summary(res2.aov)
             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
1 observation deleted due to missingness
```

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

   If we want to find the variable that mainly affects hospital costs, we run a regression model with all the variables and see the associated p-values generated. The lowest values reveal the variables that are the most significant to the model.

   Code:

   ```
   totcost <-lm(formula = TOTCHG ~ ., data=hosp)
   summary(totcost)
   ```

   Result:

   From the summary output on the following page, it appears that the length of stay has the smallest p-value and therefore is most likely correlated with hospital cost. Many of the diagnosis groups also have a low p-value, so it appears that the diagnosis group is somewhat correlated with hospital cost, but not as much as length of stay as some diagnosis groups still have a large p-value calculated with the regression.

```
> hosp$FEMALE <- factor(FEMALE)
> hosp$APRDRG <- factor(APRDRG)
> totcost <-lm(formula = TOTCHG ~ ., data=hosp)
> summary(totcost)

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

Residuals:
    Min      1Q  Median      3Q     Max
-5403.7  -188.8   -52.0   113.5  5403.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7017.4364   966.0317   7.264 1.79e-12 ***
AGE            86.5944    20.7881   4.166 3.76e-05 ***
FEMALE1      -136.8780    78.7821  -1.737 0.083032 .
LOS            664.6593    21.2924  31.216  < 2e-16 ***
RACE2          269.7343   408.6436   0.660 0.509563
RACE3          641.3334   862.2531   0.744 0.457413
RACE4          106.4079   458.4198   0.232 0.816557
RACE5         1577.1875   908.2736   1.736 0.083201 .
RACE6          -73.8266   566.3145  -0.130 0.896340
APRDRG23      4355.1399  1182.4224   3.683 0.000260 ***
APRDRG49      7890.6917  1187.2479   6.646 9.18e-11 ***
APRDRG50     -5254.4156  1194.6819  -4.398 1.38e-05 ***
APRDRG51     -7323.6414  1184.2871  -6.184 1.46e-09 ***
APRDRG53     -1199.9825   954.2018  -1.258 0.209230
APRDRG54     -8166.3229  1184.4591  -6.895 1.95e-11 ***
APRDRG57      -860.5678  1081.7666  -0.796 0.426752
APRDRG58     -5651.6901  1238.0309  -4.565 6.54e-06 ***
APRDRG92      3042.9880  1184.6409   2.569 0.010546 *
APRDRG97        -0.9807  1211.2219  -0.001 0.999354
APRDRG114      771.2360  1199.1537   0.643 0.520471
APRDRG115     2529.0158  1063.9012   2.377 0.017887 *
APRDRG137      135.6525  1262.5545   0.107 0.914488
APRDRG138    -4574.7058  1042.1335  -4.390 1.43e-05 ***
APRDRG139    -4931.6448   985.5923  -5.004 8.23e-07 ***
APRDRG141    -6352.6992  1195.6625  -5.313 1.74e-07 ***
APRDRG143    -8530.9425  1540.3888  -5.538 5.34e-08 ***
APRDRG204    -2044.5193  1182.5513  -1.729 0.084547 .
APRDRG206     -127.7919  1220.9720  -0.105 0.916691
APRDRG225      895.8186  1049.4715   0.854 0.393810
APRDRG249    -5315.2746   997.4554  -5.329 1.60e-07 ***
APRDRG254    -7979.6240  1540.5665  -5.180 3.43e-07 ***
APRDRG308     2123.5545  1199.1936   1.771 0.077303 .
APRDRG313    -1178.3407  1110.6267  -1.061 0.289302
APRDRG317     4988.0046  1200.2669   4.156 3.92e-05 ***
APRDRG344    -2162.1842  1056.0034  -2.048 0.041217 *
APRDRG347    -3802.5781  1012.6388  -3.755 0.000197 ***
APRDRG420    -6004.9500  1049.0367  -5.724 1.96e-08 ***
APRDRG421    -6583.1473  1473.4757  -4.468 1.01e-05 ***
APRDRG422    -7058.7682  1015.5830  -6.950 1.37e-11 ***
APRDRG560    -7243.4821  1045.9573  -6.925 1.60e-11 ***
APRDRG561    -8455.5174  1188.4307  -7.115 4.75e-12 ***
APRDRG566    -7552.9821  1184.1817  -6.378 4.65e-10 ***
APRDRG580    -4857.0957  1244.3640  -3.903 0.000110 ***
APRDRG581    -4663.4042  1068.0593  -4.366 1.59e-05 ***
APRDRG602    -4943.5879  1497.9440  -3.300 0.001047 **
APRDRG614    -7719.0733  1102.9638  -6.998 1.00e-11 ***
APRDRG626    -7139.5289  1032.3552  -6.916 1.70e-11 ***
APRDRG633    -6705.3678  1046.5992  -6.407 3.92e-10 ***
APRDRG634    -5032.4031  1114.8054  -4.514 8.23e-06 ***
APRDRG636    -3615.9128  1072.1686  -3.373 0.000813 ***
APRDRG639    -7181.2610  1069.0365  -6.718 5.91e-11 ***
APRDRG640    -6940.8612   966.2080  -7.184 3.03e-12 ***
APRDRG710    -1575.9787  1229.1167  -1.282 0.200465
APRDRG720     3642.4840  1227.8470   2.967 0.003180 **
APRDRG723    -5705.2093  1065.3976  -5.355 1.40e-07 ***
APRDRG740     -377.7710  1187.7288  -0.318 0.750593
APRDRG750    -8730.5193  1182.5513  -7.383 8.15e-13 ***
APRDRG751    -8155.6282   914.2108  -8.921  < 2e-16 ***
APRDRG753    -8003.8054   892.9601  -8.963  < 2e-16 ***
APRDRG754    -8103.7523   898.2150  -9.022  < 2e-16 ***
APRDRG755    -7940.5790   916.3521  -8.665  < 2e-16 ***
APRDRG756    -7949.0870  1051.9336  -7.557 2.53e-13 ***
APRDRG758    -8234.5317   887.8911  -9.274  < 2e-16 ***
APRDRG760    -8608.9951  1055.0501  -8.160 3.78e-15 ***
APRDRG776    -8625.8601  1182.4224  -7.295 1.46e-12 ***
APRDRG811    -6636.1016  1048.1659  -6.331 6.15e-10 ***
APRDRG812    -6042.7403   999.2659  -6.047 3.21e-09 ***
APRDRG863    -9792.3805  1331.7366  -7.353 9.94e-13 ***
APRDRG911    35382.7216  1188.5291  29.770  < 2e-16 ***
APRDRG930     1651.0401  1047.6143   1.576 0.115765
APRDRG952    -4321.2008  1182.6769  -3.654 0.000291 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 785.2 on 428 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.965,     Adjusted R-squared:  0.9593
F-statistic: 168.6 on 70 and 428 DF,  p-value: < 2.2e-16
```