Table of Contents

In this Healthcare Project:

Code mentions the code as used

Output Each code is followed by the respective output. Taken as is from the console of RStudio

Observation Is provided in order to Analyze the outcomes of the output (wherever necessary)

Conclusion At the end of each Assignment.

Domain: Healthcare

Background

A nationwide survey of hospital costs conducted by the US Agency for Healthcare consists of hospital records of inpatient samples. The given data is restricted to the city of Wisconsin and relates to patients in the age group 0-17 years. The agency wants to analyze the data to research on healthcare costs and their utilization.

Dataset Description:

Here is a detailed description of the given dataset: (**HospitalCosts.xlsx**)

Attribute	Description
Age	Age of the patient discharged
Female	A binary variable that indicates if the patient is female
Los	Length of stay in days
Race	Race of the patient (specified numerically)
Totchg	Hospital discharge costs
Aprdrg	All Patient Refined Diagnosis Related Groups

Analysis with Answers

Q. 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.

Answer 1.

Code

install.packages("readxl",dependencies = T) **# install package readxl with dependencies = TRUE**

library (readxl) **# read the library readxl**

Output

```
> install.packages("readxl",dependencies = T)
package 'readxl' successfully unpacked and MD5 sums checked

> library(readxl)
```

Code

read_xlsx("F:/Simplilearn DataScience/DataScience with R/Project/HospitalCosts.xlsx")->hp

 # read the path of the file HospitalCosts.xlsx and assigned to hp

Output

```
> read_xlsx("F:/Simplilearn DataScience/DataScience with
R/Project/HospitalCosts.xlsx")->hp
```

Code

table(hp$AGE) **# table uses the cross-classifying factors to build a contingency table of the counts at each combination of factor levels for Age**

Output

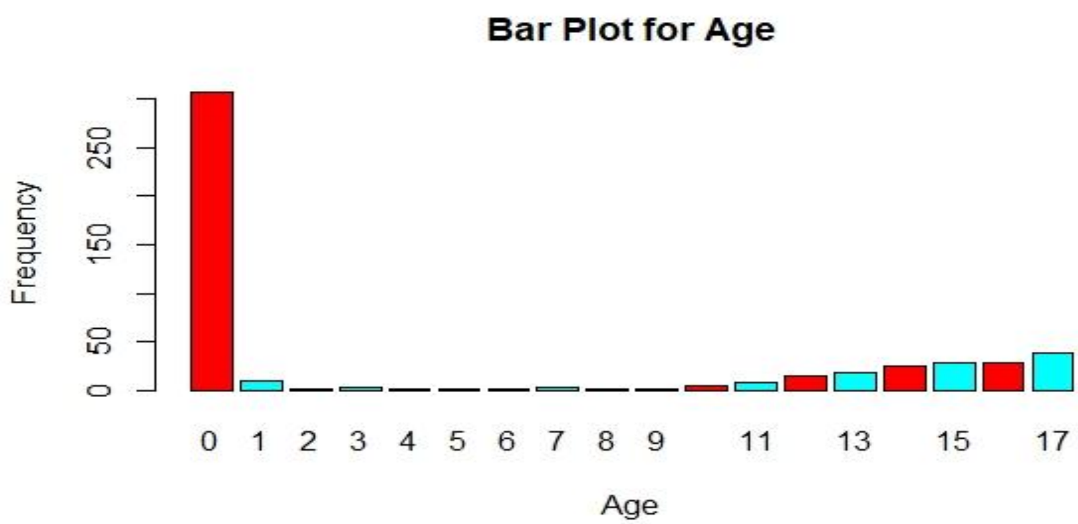
```
>  table(hp$AGE)
 0  1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17
307  10  1  3  2  2  2  3  2  2  4  8  15  18  25  29  29  38
```

Code

barplot(tab_age,col = rainbow(2), main= "Bar Plot for Age",xlab = "Age",ylab = "Frequency")

Output

```
> barplot(tab_age,col = rainbow(2), main= "Bar Plot for Age",xlab =
"Age",ylab = "Frequency")
```



Code

install.packages("dplyr",dependencies =T) **# install package dplyr with dependencies = TRUE**

library ("dplyr") **# read the library readxl**

Output

```
> install.packages("dplyr",dependencies = T)
package ‘dplyr’ successfully unpacked and MD5 sums checked
> library("dplyr")
```

Code

```
hp%>%group_by(hp$AGE)%>%summarise(no_of_visit =sum(LOS),totalexp =
    sum(TOTCHG))%>% mutate(myrank = rank(desc(totalexp)))%>% filter(myrank==1) -> cat1
```

cat1

Output

```
> hp%>%group_by(hp$AGE)%>%summarise(no_of_visit =sum(LOS),totalexp =
sum(TOTCHG))%>%
+   mutate(myrank = rank(desc(totalexp)))%>% filter(myrank==1) -> cat1
> cat1
# A tibble: 1 x 4
  `hp$AGE` no_of_visit totalexp myrank
     <dbl>       <dbl>    <dbl>  <dbl>
1        0         941   678118      1
```

Observation As per the Bar Plot and the above output **age is 0**

Conclusion **maximum expenditure is of 678118,**

Q. 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.

Answer 2.

Code

```
aggregate(hp[, c("LOS", "TOTCHG")], list(hp$APRDRG), sum)->df
head(df)
```

Output

```
> aggregate(hp[, c("LOS", "TOTCHG")], list(hp$APRDRG), sum)->df
> head(df)
  Group.1 LOS TOTCHG
1      21   2  10002
2      23   2  14174
3      49   6  20195
4      50   2   3908
5      51   3   3023
6      53  29  82271
```

Code

```
df$Group.1[df$LOS == max(df$LOS, na.rm = T)] #The group having max hospitalization
```

Output

```
> df$Group.1[df$LOS == max(df$LOS, na.rm = T)]
[1] 640
```

Code

```
aggregate(hp[,c("TOTCHG")],list(hp$APRDRG), sum)->df1# Group that has maximum expenditure
head(df1)
max(df1$TOTCHG)
```

Output

```
> aggregate(hp[, c("LOS", "TOTCHG")], list(hp$APRDRG), sum)->df1
```

```
> head(df1)
  Group.1 TOTCHG
1      21  10002
2      23  14174
3      49  20195
4      50   3908
5      51   3023
6      53  82271
>
> max(df1$TOTCHG)
[1] 437978
```

Conclusion **The diagnosis-related group that has maximum hospitalization of 640 and expenditure 437978.**

Q. 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.

Answer 3.

Code

```
hp$RACE  # Race of the patient (numerical)
```

Output

```
> hp$RACE
  [1]   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
4   1   1
 [28]   1   1   1   1   1   6   1   1   1   1   1   1   1   1   2   1   1   1   1   1   1   1   1   1
1   1   1
 [55]   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1
 [82]   1   1   1   1   1   1   1   1   5   5   5   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1
[109]   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1
[136]   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1
[163]   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1
[190]   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1   1
```

Code

```
hp$TOTCHG  # Hospital discharge costs (continuous variable)
```

Output

```
> hp$TOTCHG
  [1]  2660  1689 20060   736  1194  3305  2205  1167   532  1363  1245  1656
1379
 [14]  2346  4006  2181   628  2463  1956  1802  3188  2129  7421  1122  1173
3625
 [27]  3908  3994  1033  2860  3814  1132  1163   610  9530  1268  2582  1287
6594
 [40]   909  2530  1534 14243  1699  7298   636   626  3782  1444  1183  3045
3624
 [53]  6810  1409  1211  9606  1411   607  2932  5075   762  6329  1226  8223
1193
 [66]  1076 17434  1647  3865   628   806 29188  4717 15129  1085  1607  1499
7648
 [79]  1527  1483  2844  3124  1760  1278  1620  1220  1134  1235  1656  4072
1393
 [92]   615   779  1385  1224  1779  1526   882  2075 12042  1309  1290  1280
1719
```

Code

table(hp$RACE) #0 levels

Output

```
> table(hp$RACE)

  1   2   3   4   5   6
484   6   1   3   3   2
```

Code

aov(TOTCHG ~ RACE, data = hp)->model1

summary(model1)

Output

```
> aov(TOTCHG ~ RACE, data = hp)->model1
> summary(model1)
             Df    Sum Sq  Mean Sq F value Pr(>F)
RACE          1 2.488e+06  2488459   0.164  0.686
Residuals   497 7.540e+09 15170268
1 observation deleted due to missingness
```

Observation

Ho (Null Hypothesis) There is no significance in the mean difference between TOTCHG (Hospital discharge costs) with Race (Race of the patient)

Ha (Alternate hypothesis) There is difference between TOTCHG and Race

Conclusion

The Ho (Null Hypothesis) is retained as there is no significant difference between TOTCHG and Race as they are not varying. We therefore reject the Ho. No significance exists between these two (TOTCHG and Race). Race is not a significant variable. This means that it's highly unlikely that differences among the means are due to chance. It means that you reject Ho.

Also please note that as aov (Annova) is designed for balanced designs, and the results can be hard to interpret without balance: the **missing values** in the response will likely lose the balance. The methods used are statistically difficult without balance.

We therefore can say that," **The race of the patient is not related to the hospitalization costs**"

Q.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.

Answer 4.

Code

aov(TOTCHG ~ AGE + hp$FEMALE, data = hp)-> model4

summary(model4)

Output

```
> aov(TOTCHG ~ AGE + hp$FEMALE, data = hp)-> model4
> summary(model4)
              Df    Sum Sq   Mean Sq F value  Pr(>F)
AGE            1 1.308e+08 130822234   8.849 0.00308 **
hp$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
```

Observation

In this model (model4) both the variables are leading to significance varying in total costs. The probability is smaller for Age (0.00308) and for female (0.03497). Age is leading to higher variance.

In the above output

** means Ho (Null Hypothesis) is rejected at Alpha = 1% P value is lying between .1% to 1%

* means Ho (Null Hypothesis) is rejected at Alpha = 5% P value is lying between 1% to 5%

Conclusion **The hospital costs are more according to Age then Female for proper allocation of resources**

Q.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.

Answer 5.

Code

hp$LOS **# continuous variable**

Output

```
> hp$LOS
  [1]  2  2  7  1  1  0  4  2  1  2  2  2  2  4  7  4  1  4  3  3  1  2  1  1
2  2  2
 [28]  1  0  2  2  0  2  1  3  1  4  2  3  0  0  2  5  3  2  1  1  2  2  2  5
5 12  1
 [55]  2  4  1  0  1  3  1  6  1  4  2  2  6  2  7  1  1 41  2 12  2  3  3  3
2  2  4
 [82]  3  3  2  2  2  2  0  3  4  2  0  1  2  2  3  2  1  1 17  2  2  2  3  2
3  2  2
[109]  2  2  4  2  4  5  2  2  2  2  4  2  2  2 10  4  0  2  3  5  2  2  3  2
3  1  2
[136]  2  7  2  3  4  4  2  2  4  2  2  4  2  2  2  2  2  2  2  2  3  1  1  5
2  2  2
[163]  2  1  2  1  1  1 39  2  2  2  1  4  2  3  2  2  4  1  4  3  2  1  2  2
3  0  2
[190]  1  3  2  4  2  5  1  3  2  2  3  2  0  2  2  6  5  3  2  2  3  0  3  3
3  3  1
```

Code

lm(LOS ~ AGE + FEMALE + RACE, data = hp)->model2 **# Linear model to predict from Age, gender and race from length of stay**

summary(model2)

Output

```
>  lm(LOS ~ AGE + FEMALE + RACE, data = hp)->model2
>   summary(model2)
```

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

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
```

Observation

In this model,

Adjusted R –squared is 0.00186 i.e. .18% that is very low.

The probability model is not linear. Age is significant at alpha = 10%. Because of Age we have a very very bad performance. The model is not linear and not significant. Even if, it is not linear relationship, where probability values are less than 5%, that could lead to feature selection, but if variance is not significant at all. In that case there is not even a non linear relationship. Even a non linear relationship cannot be predicted.

Conclusion

In this case using the linear model (model2) found out that may be Age can have non linear relationship, but Female (Gender) and Race do not have a significant relationship. **Gender (Female) ,Race and Age is not leading to the prediction of length of stay.**

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

Answer 6.

Code

names(hp) **# get the names in hp**

Output

```
> names(hp)
[1] "AGE"    "FEMALE" "LOS"    "RACE"   "TOTCHG" "APRDRG"
```

Code

aov(TOTCHG ~., data = hp)->model3

summary(model3)

Output

```
> aov(TOTCHG ~., data = hp)->model3
> summary(model3)
             Df    Sum Sq   Mean Sq F value   Pr(>F)    
AGE           1 1.297e+08 1.297e+08  18.998 1.59e-05 ***
FEMALE        1 6.522e+07 6.522e+07   9.550  0.00211 ** 
LOS           1 3.086e+09 3.086e+09 451.889  < 2e-16 ***
RACE          1 1.715e+06 1.715e+06   0.251  0.61652    
APRDRG        1 8.923e+08 8.923e+08 130.648  < 2e-16 ***
Residuals   493 3.367e+09 6.830e+06                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

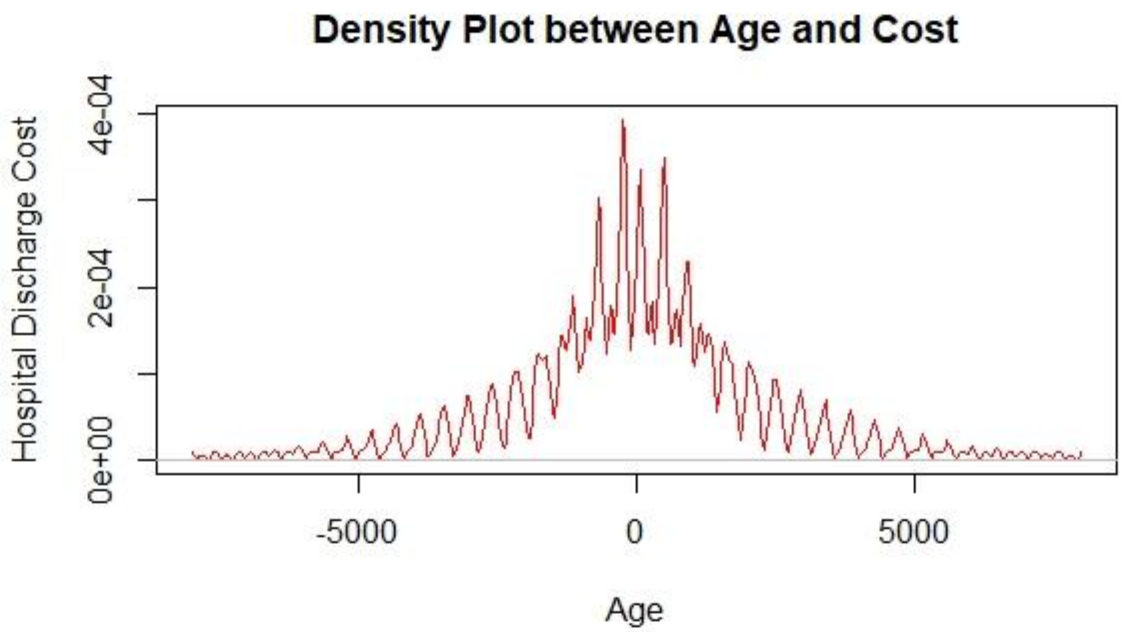
Code

```
plot(density((x=tp1$AGE),(y=tp1$TOTCHG)),
                  col="red",
                  xlab="Age & Diagnosis Related Group",
                  ylab="Hospital Discharge Cost",
```

main="Density Plot between Age and Cost")

Output for Density between Age and Hospital Costs

```
> plot(density((x=tp1$AGE),(y=tp1$TOTCHG)),col="red",
+      xlab="Age",
+      ylab="Hospital Discharge Cost",
+     main="Density Plot between Age and Cost")
```



Code

plot(density((x=tp1$APRDRG),(y=tp1$TOTCHG)),

col="purple",

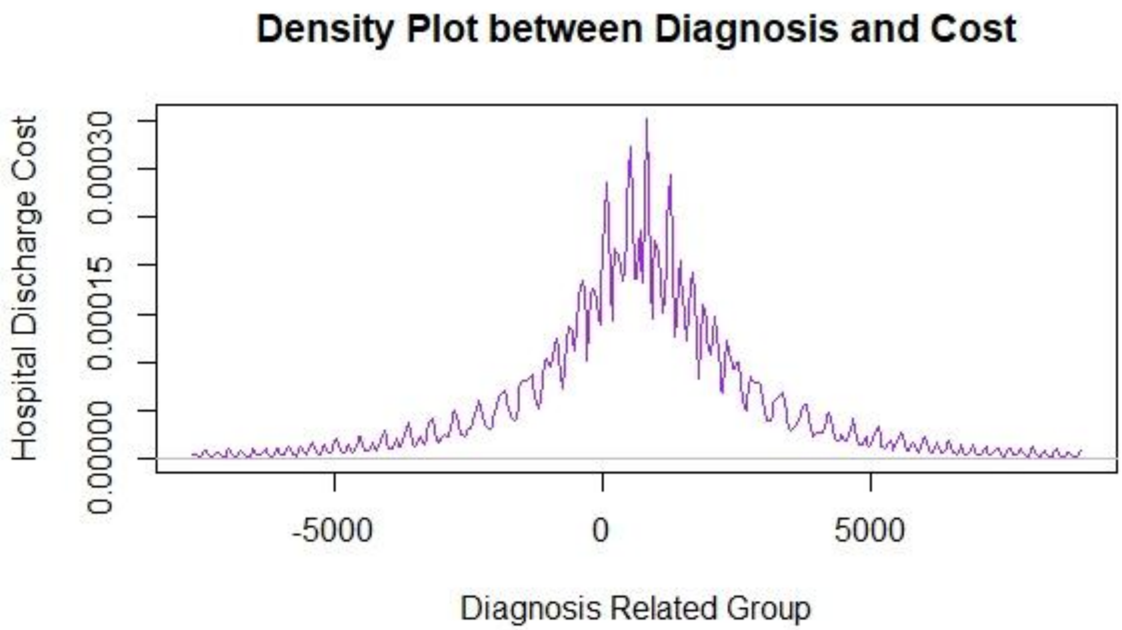
xlab="Diagnosis Related Group",

ylab="Hospital Discharge Cost",

main="Density Plot between Diagnosis and Cost")

Output for Density between Diagnosis Related Groups and Hospital Costs

```
> plot(density((x=tp1$AGE),(y=tp1$TOTCHG)),col="red",
+      xlab="Diagnosis Related Group",
+      ylab="Hospital Discharge Cost",
+     main="Density Plot between Diagnosis and Cost")
```



No table of contents entries found. Observation

AGE and APRDRG have the lowest probability leading to highest variance in Y

Also from the above 2 density plots we observe that Age affects Hospital costs and Diagnosis Related Group also affects Hospital Costs

Conclusion

AGE and APRDRG are the variable that mainly affects hospital costs.

<End of document>