FINAL PROJECT DISSERTATION

Analyze the Healthcare cost and Utilization in Wisconsin hospitals

Submitted By:

Amanjeet Sahu

Overview:

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 the healthcare costs and their utilization.

Here is a detailed description of the features in the dataset:

AGE	: Age of the patient discharged
FEMALE	: 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

Profiling the Data:

```
setwd("\\Project\\Projects for Submission\\Healthcare\\Healthcare")
df<- read.csv(file = "HospitalCosts.csv" , sep = ",")
View(df)
head(x=df, n=10)
```

```
   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
7   17      1   4    1   2205    754
8   16      1   2    1   1167    754
9   16      1   1    1    532    753
10  17      1   2    1   1363    758
> |
```

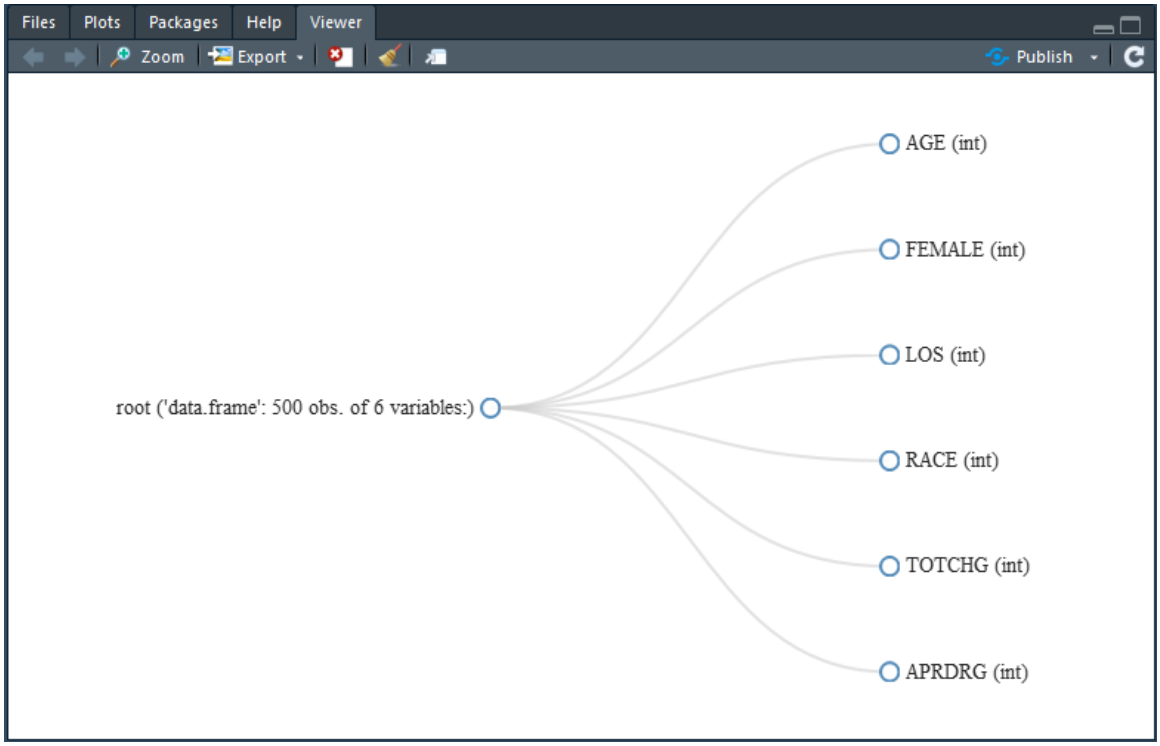
Exploratory Data Analysis:

Variables:

The first that we need to do in EDA is to check the dimension of our dataset and types of features.

```
#install.packages('DataExplorer)
library(DataExplorer)
plot_str(df)
```

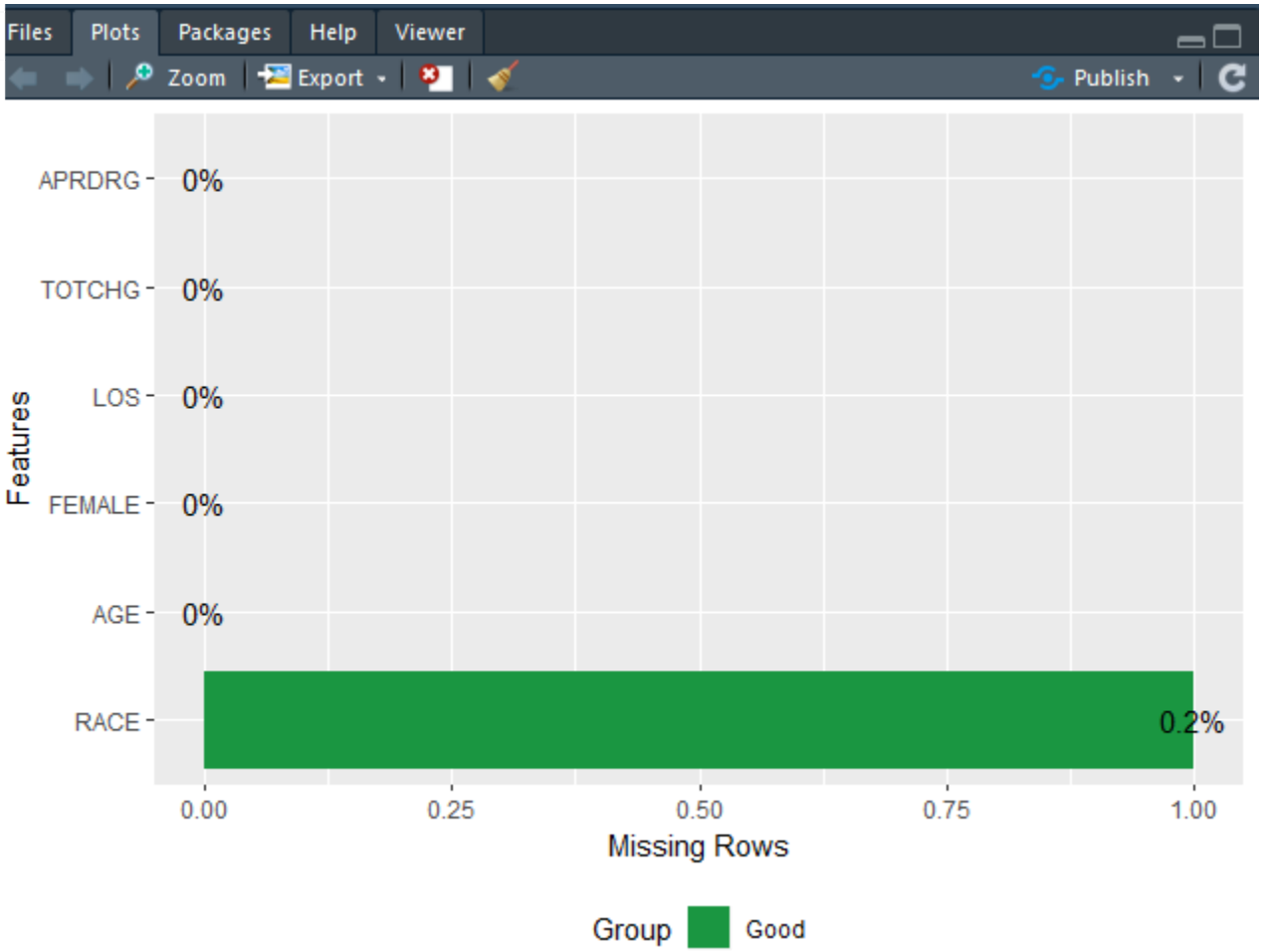
This plot describes #that is (500 x 6) and it also shows the type of each feature that we have in our dataset.



Search for Missing Variables:

```
plot_missing(df)
```

For the above code you need to download DataExplorer Package which we installed in the first step of EDA, Next you will see the plot that will show which variable has missing values and how many missing values are there.



This above graph clearly shows that Race feature has a single missing or NA value which is 0.2% of total number of rows. Now Let's find out whether this graph is telling the truth or not using our R code.

```
sum(is.na(df$RACE))
```

```
> sum(is.na(df$RACE))
[1] 1
>
```

That is a great insight, we got what we expected Race feature has 1 missing row, now let's see how to deal with this

As we can see that the RACE column is categorical variable so the best practice is to impute the missing data with a mode value or we can say most frequent category in the column.

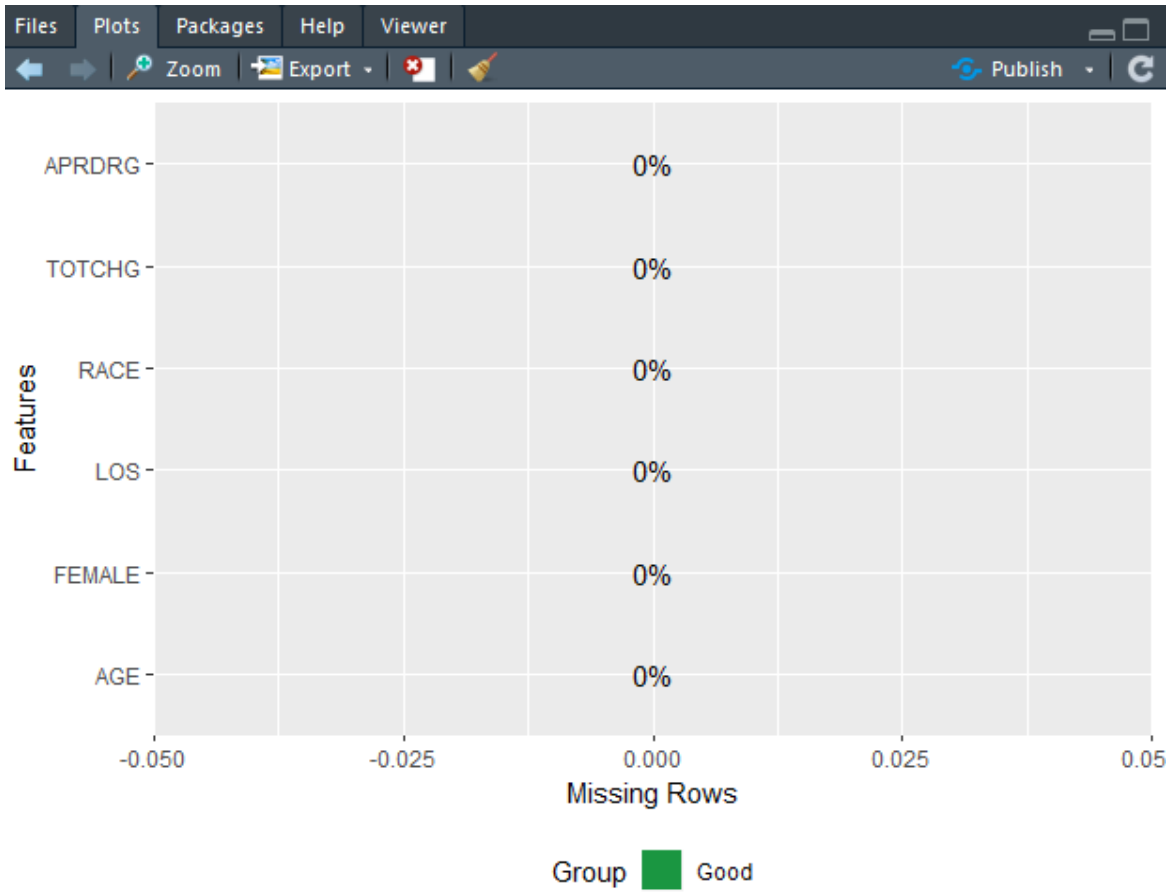
```
head(df$RACE, n=30)
table(df$RACE)
df$RACE = ifelse(test=is.na(df$RACE),yes=1,no=df$RACE)
sum(is.na(df$RACE))
```

```
> head(df$RACE, n=30)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 1 1 1 1 1
> View(df)
> table(df$RACE)

  1   2   3   4   5   6
484   6   1   3   3   2
> df$RACE = ifelse(test=is.na(df$RACE),yes=1,no=df$RACE)
> sum(is.na(df$RACE))
[1] 0
>
```

Now let's plot the missing value plot again.

```
plot_missing(df)
```

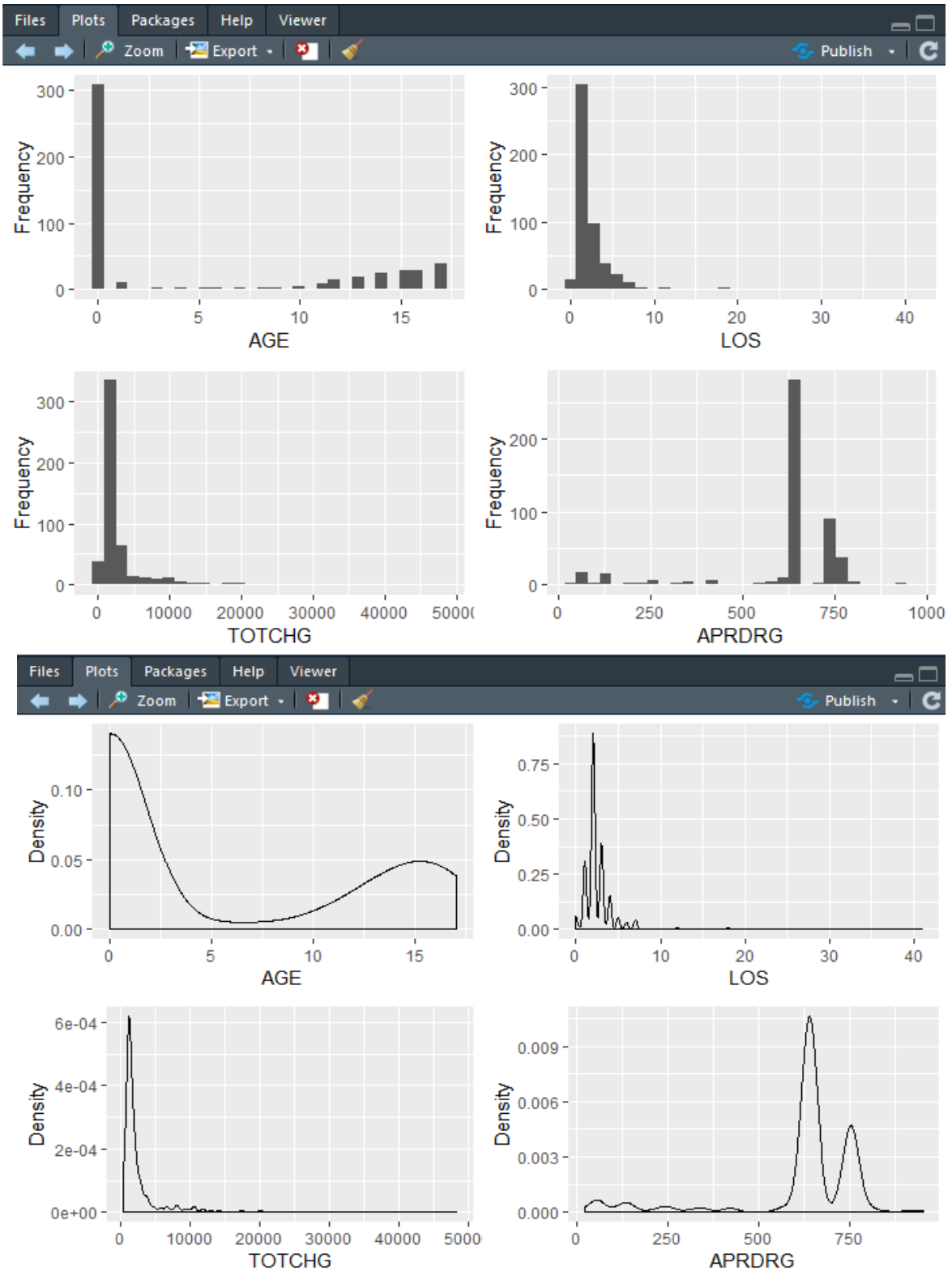


Great, now we have no missing values in our dataset.

Continuous Variable:

Histogram and density plots are best options to analyse continuous data.

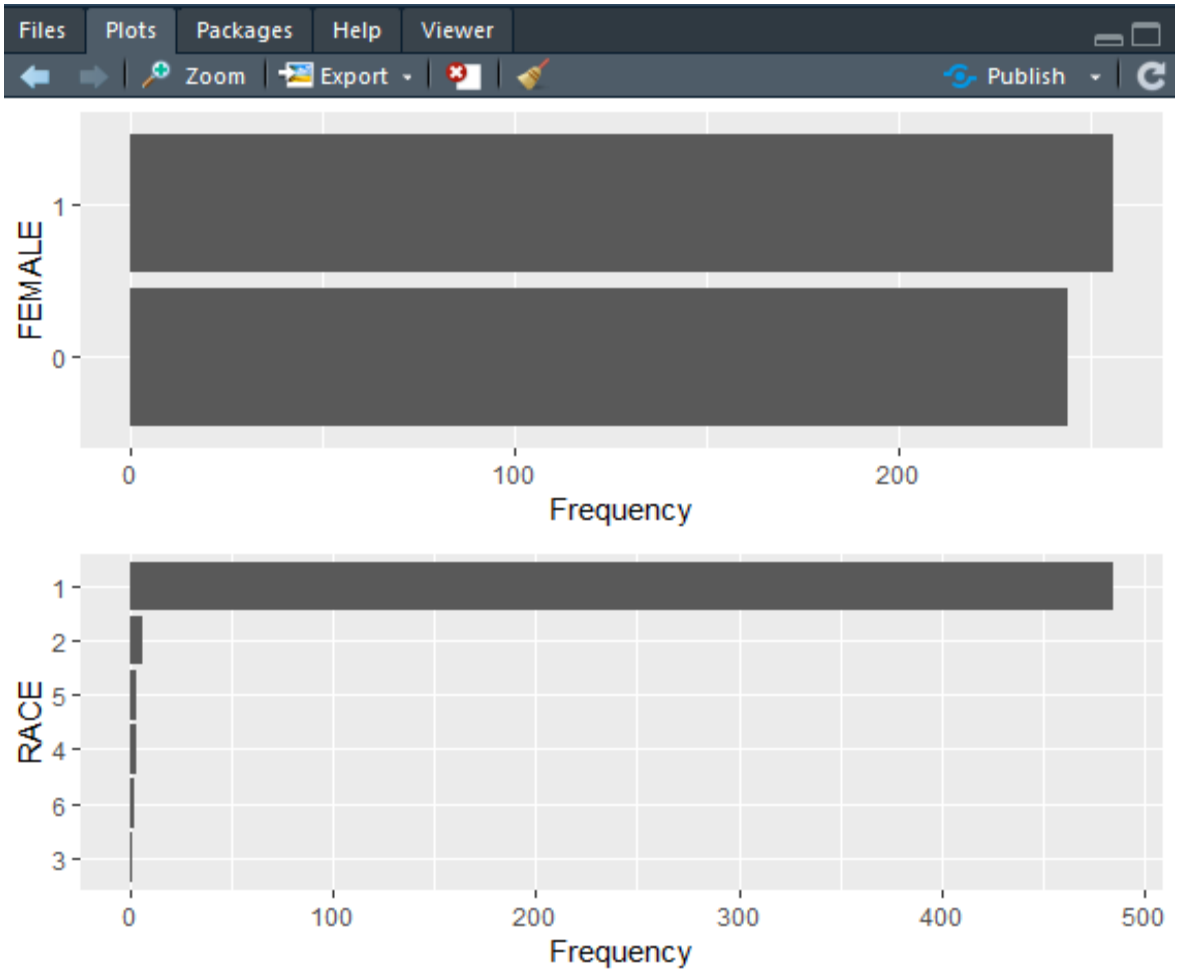
```
plot_histogram(df[,-c(2,4)])
plot_density(df[,-c(2,4)])
```



Discrete Variable:

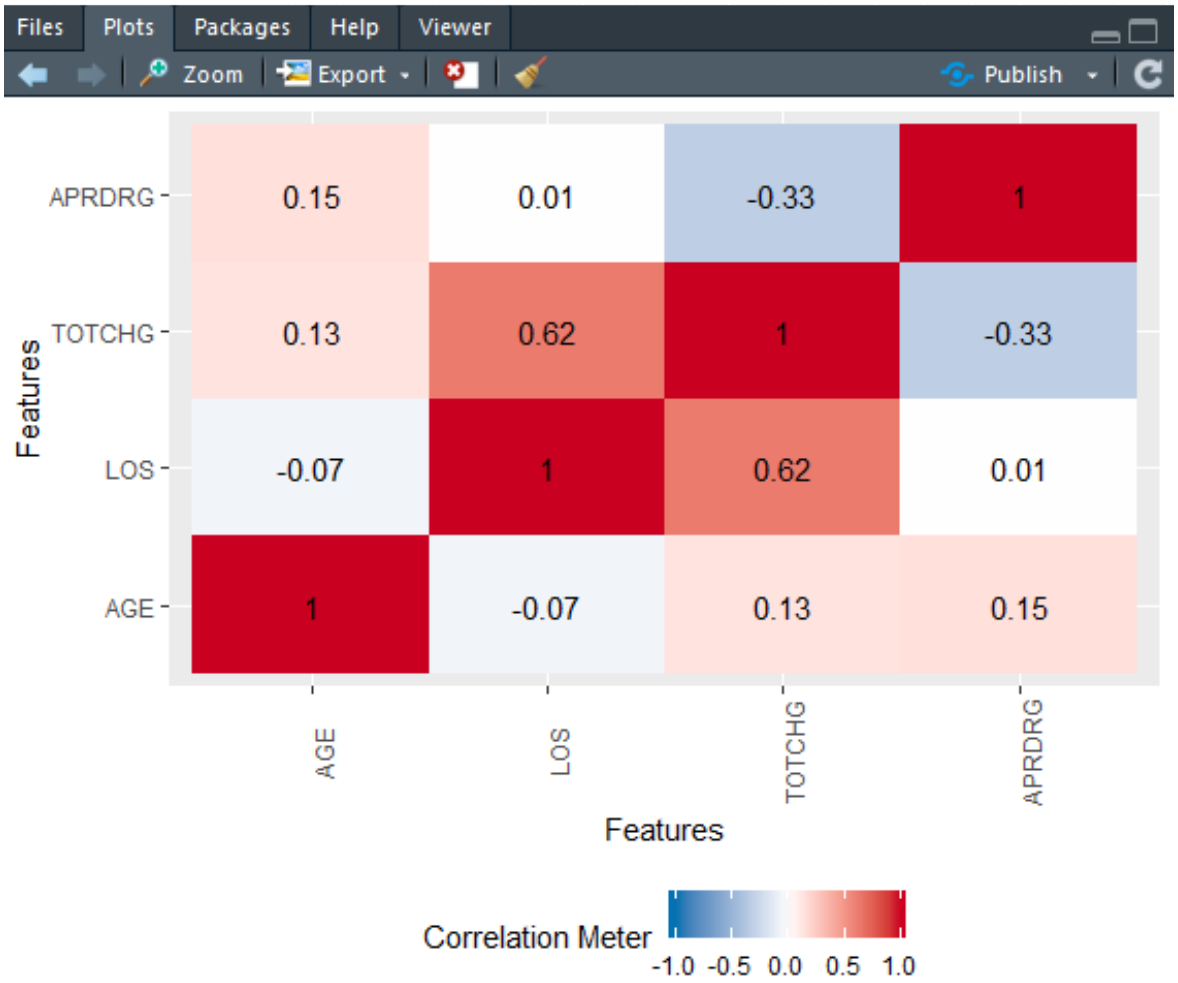
Bar Graph or a pie chart will be best option for discrete data.

```
df$FEMALE<- as.factor(df$FEMALE)
df$RACE<- as.factor(df$RACE)
plot_bar(df)
```



Multivariate Analysis:

```
plot_correlation(df, type = 'continuous')
```

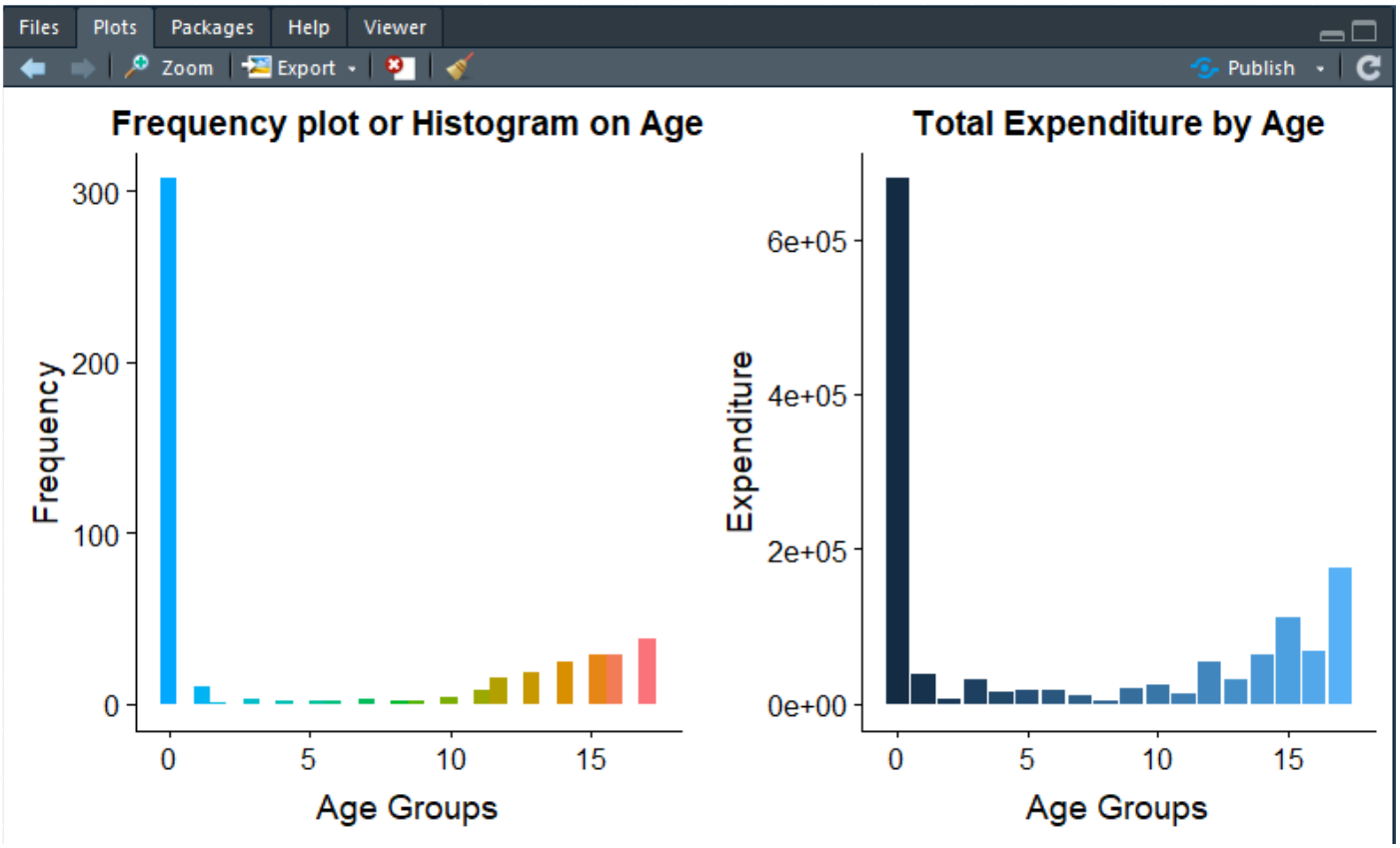


The goals of the project are:

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

Now to tackle this we can draw we can visualise the results using a bar plot and histogram because *"A picture is worth a thousand word"*.

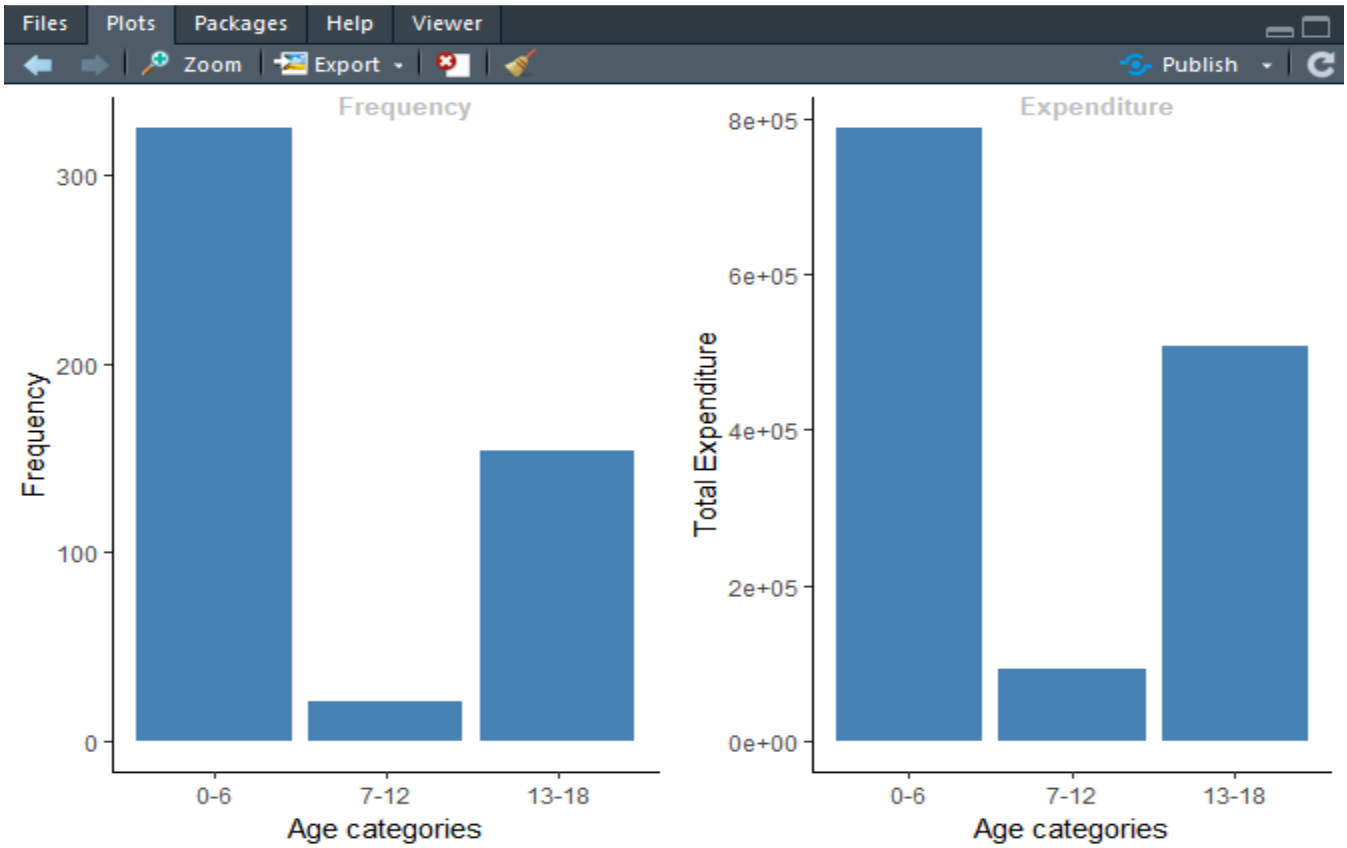
```
library(ggplot2)
b1<- ggplot(df, aes(AGE, fill = cut(AGE, 30))) +
  geom_histogram(show.legend = FALSE) +
  scale_fill_discrete(h = c(240, 10))+
  labs(x="Age Groups",y="Frequency",title="Frequency plot or Histogram on Age")
df1<- aggregate(TOTCHG ~ AGE, data = df, sum)
b2<- ggplot(data=df, aes(x=AGE, y=TOTCHG, fill=AGE)) +
  geom_bar(stat="identity", show.legend = FALSE)+
  labs(x="Age Groups",y="Expenditure",title="Total Expenditure by Age")
  theme_minimal()
#install.packages("cowplot") to plot multiple plots in a single line
library(cowplot)
plot_grid(b1, b2) # function from cowplot library
```



From the above graph one can clearly see that the age group of 0 has been frequent to the hospital and from the second graph expenditure of 0 age group is highest that expenditure on new born Babies is highest or it can be that in Wisconsin city new born babies are falling ill very often.

Now let's create 3 age groups and do the analysis

```
df$AGE_CATE<-cut(df$AGE, seq(0,18,6), right=FALSE,
                 labels=c("0-6", "7-12", "13-18"))# creating age categories with 3 cuts
freq<-table(df$AGE_CATE)# calculating frequency of each category
df1<- aggregate(TOTCHG ~ AGE_CATE, data = df, sum) # with aggregate function calculating total expenditure
df1$freq<- freq
p1<-ggplot(data=df1, aes(x=AGE_CATE, y=freq)) +
  geom_bar(stat="identity", fill="steelblue") + theme_classic() +
  labs(y="Frequency", x= "Age categories")
p2<-ggplot(data=df1, aes(x=AGE_CATE, y=TOTCHG)) +
  geom_bar(stat="identity", fill="steelblue") + theme_classic() +
  labs(y="Total Expenditure", x= "Age categories")
#install.packages("cowplot") to plot multiple plots in a single line
library(cowplot)
plot_grid(p1, p2, labels = c("Frequency","Expenditure"), label_size = 10,
          label_x = 0.4, label_colour = "grey") # function from cowplot library
```



From the above graph one can clearly see that the age group of 0-6 that is the children are most frequent to the hospital and they also have the highest expenditures. The first age group is from 0-6 that means new born babies expenditure is also included in it and from this analysis we can clearly say that children in Wisconsin city of United States are more likely to fall ill.

- 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.

```
table(df$APRDRG)
freq1<-as.data.frame(table(df$APRDRG))
freq1$Var1[which.max(freq1$Freq)]
```

```
> table(df$APRDRG)

 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
> freq1<-as.data.frame(table(df$APRDRG))
> freq1$Var1[which.max(freq1$Freq)]
[1] 640
```

From this chunk of code we can say that All Patient Refined Diagnosis Related Groups with a group id of 640 are most frequent to the hospital and now let's find which group has highest expenditure.

```
freq2<- aggregate(TOTCHG ~ APRDRG, data = df, sum)
head(freq2)
freq2$APRDRG[which.max(freq2$TOTCHG)]
```

```
> freq2<- aggregate(TOTCHG ~ APRDRG, data = df, sum)
> head(freq2)
  APRDRG TOTCHG
1     21  10002
2     23  14174
3     49  20195
4     50   3908
5     51   3023
6     53  82271
> freq2$APRDRG[which.max(freq2$TOTCHG)]
[1] 640
>
```

Great, From above we can see that all patient refined diagnosis related groups with a group id 640 has the highest expenditure.

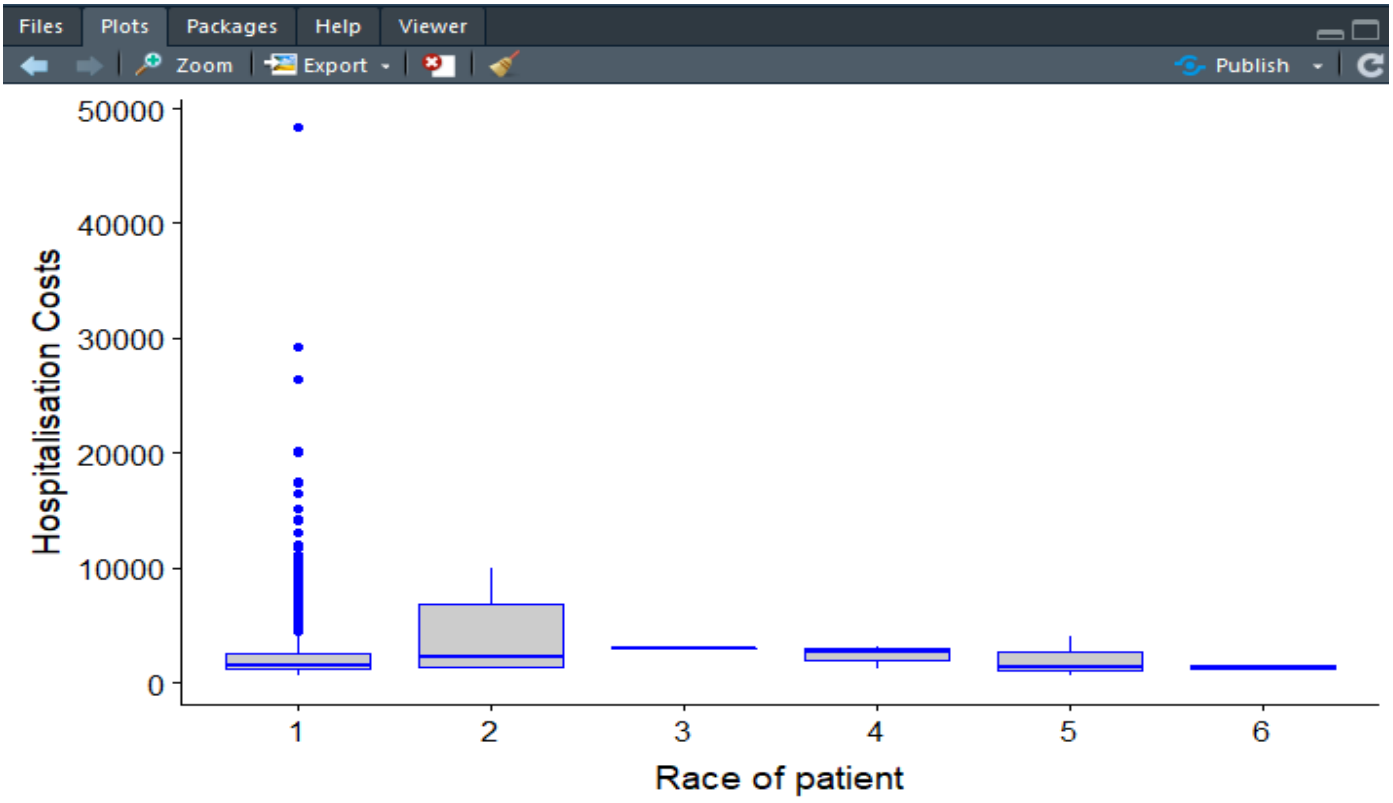
So, Group number 640 has the highest expenditure and visits to the hospital most frequently.

- 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.

For this case there will be no better choice than implementing one way ANOVA. It compares the means between the groups you are interested in and determines whether any of those means are statistically significantly different from each other.

```
head(df$RACE)
summary(df$RACE)
ggplot(df, aes(x = RACE, y = TOTCHG)) +
  geom_boxplot(fill = "grey80", colour = "blue") +
  scale_x_discrete() + xlab("Race of patient") +
  ylab("Hospitalisation Costs")
```

```
> head(df$RACE)
[1] 1 1 1 1 1 1
Levels: 1 2 3 4 5 6
> summary(df$RACE)
  1   2   3   4   5   6
485   6   1   3   3   2
> ggplot(df, aes(x = RACE, y = TOTCHG)) +
+   geom_boxplot(fill = "grey80", colour = "blue") +
+   scale_x_discrete() + xlab("Race of patient") +
+   ylab("Hospitalisation Costs")
>
```




```
model<- lm(TOTCHG~RACE, data = df)
summary(model)
```

```
> model<- lm(TOTCHG~RACE, data = df)
> summary(model)

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

Residuals:
   Min     1Q Median     3Q    Max
 -3049  -1550  -1223   -236  45619

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2769.3      177.2  15.625   <2e-16 ***
RACE2         1432.8     1603.3   0.894    0.372
RACE3          271.7     3907.2   0.070    0.945
RACE4         -424.7     2260.5  -0.188    0.851
RACE5         -742.7     2260.5  -0.329    0.743
RACE6        -1420.3     2765.7  -0.514    0.608
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3903 on 494 degrees of freedom
Multiple R-squared:  0.002467,  Adjusted R-squared:  -0.00763
F-statistic: 0.2443 on 5 and 494 DF,  p-value: 0.9426

>
```

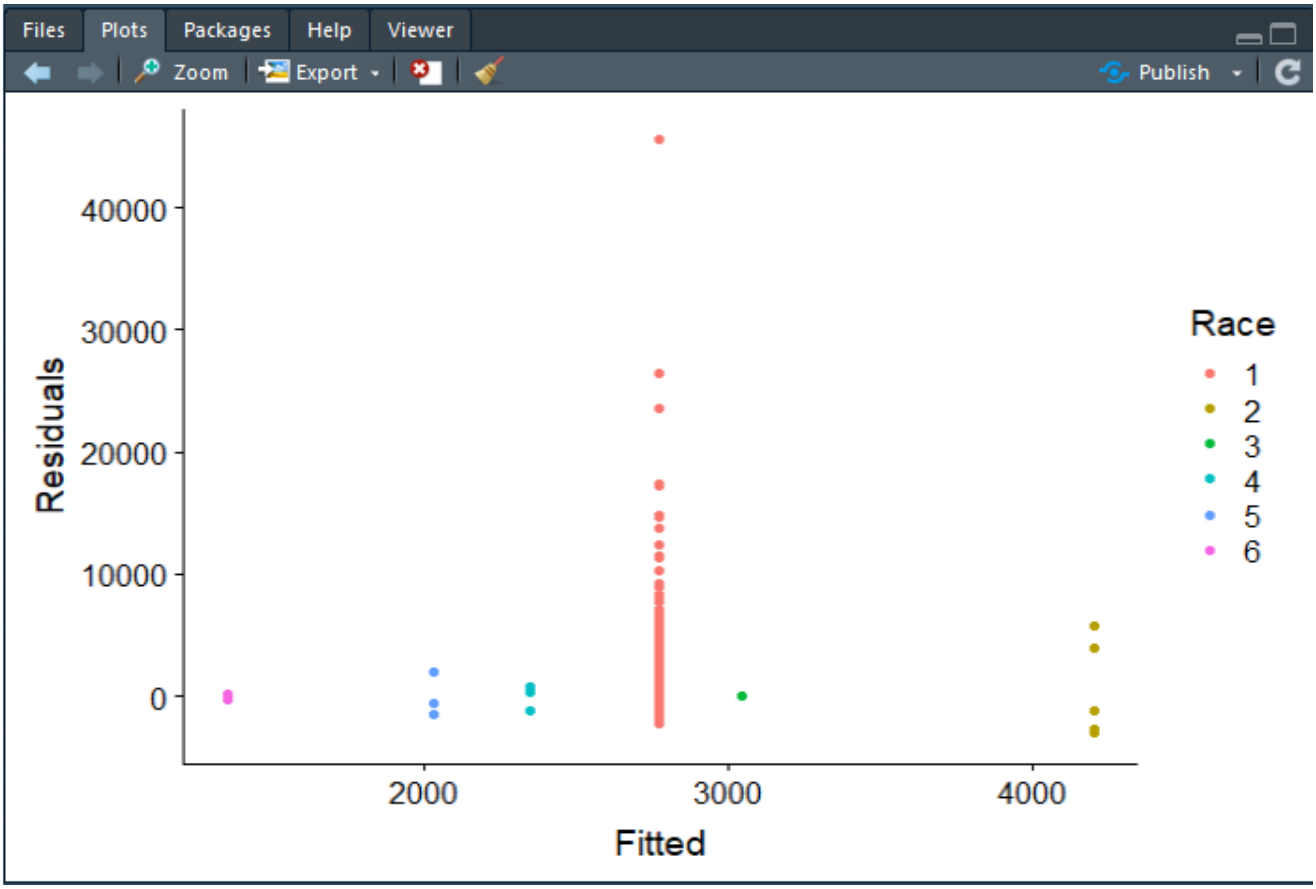
The model output indicates some evidence of a high residual standard error and a very high P-Value which indicates that there is no relation between Race and Hospitalization Charges. An analysis of variance table for this model can be produced via the anova command.

```
anova(model)
confint(model)
mod = data.frame(Fitted = fitted(model),
                         Residuals = resid(model), Race = df$RACE)
ggplot(mod, aes(Fitted, Residuals, colour = Race)) + geom_point()
```

```
> anova(model)
Analysis of Variance Table

Response: TOTCHG
           Df     Sum Sq  Mean Sq F value Pr(>F)
RACE        5   18609476  3721895  0.2443 0.9426
Residuals 494 7526126736 15235074
> confint(model)
                2.5 %   97.5 %
(Intercept)  2421.107 3117.565
RACE2       -1717.310 4582.971
RACE3       -7405.185 7948.513
RACE4       -4866.011 4016.672
RACE5       -5184.011 3698.672
RACE6       -6854.270 4013.598
> mod = data.frame(Fitted = fitted(model),
+                          Residuals = resid(model), Race = df$RACE)
> ggplot(mod, aes(Fitted, Residuals, colour = Race)) + geom_point()
>
```

This table confirms that there are differences between the RACES which were highlighted in the model summary. The function confint is used to calculate confidence intervals on the treatment parameters, by default 95% confidence intervals. The model residuals can be plotted against the fitted values to investigate the model assumptions.



So, From this graph we see can residuals are spread and with this evidence and confidence levels we can say that there is no malpractice and RACE of the patient is not related to the hospitalisation costs

- 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.

There will be no better option than building a linear model.

```
lin_Mod<- lm(TOTCHG~AGE+FEMALE, data= df)
summary(lin_Mod)
```

```
> lin_Mod<- lm(TOTCHG~AGE+FEMALE, data= df)
> summary(lin_Mod)

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

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 ***
FEMALE1      -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

>
```

From above we can say that AGE is highly statistically significant for the analysis and it increases with total charges with a positive coefficient. And a very low P-Value also depicts that how significant this variable.

We can clearly see that Gender is also statistically significant which is impacting our analysis but we can see a negative coefficient for Female1 which says hospitalisation costs reduces in case of female and in case of males is quite high.

- 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.

Again a linear model will help.

```
lin_Mod1<- lm(LOS~AGE+FEMALE+RACE, data= df)
summary(lin_Mod)
```

```
> lin_Mod<- lm(LOS~AGE+FEMALE+RACE, data= df)
> lin_Mod1<- lm(LOS~AGE+FEMALE+RACE, data= df)
> summary(lin_Mod)

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

Residuals:
   Min     1Q Median     3Q    Max
-3.204 -1.204 -0.856  0.144 37.796

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.85563    0.23137  12.342   <2e-16 ***
AGE         -0.03902    0.02254  -1.731    0.084 .
FEMALE1      0.34799    0.31221   1.115    0.266
RACE2       -0.37573    1.39444  -0.269    0.788
RACE3        0.79638    3.38275   0.235    0.814
RACE4        0.59690    1.95542   0.305    0.760
RACE5       -0.85563    1.96098  -0.436    0.663
RACE6       -0.71745    2.39082  -0.300    0.764
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.373 on 492 degrees of freedom
Multiple R-squared:  0.008562,  Adjusted R-squared:  -0.005544
F-statistic: 0.607 on 7 and 492 DF,  p-value: 0.7503

> |
```

We can clearly see that gender i.e. FEMALE1 and RACE are not totally statistically significant that means it will not contribute towards the prediction. But in case of AGE you can see that 0.1 significance code which says it is somewhat significant. If we take a look at p value for AGE which is 0.084 which gives an alpha=8.4% and we assume that threshold is 5% then our Null Hypothesis will be accepted and Hence we cannot keep AGE in our analysis.

So, we can say none of the AGE, Gender and Race will affect the analysis and hence we cannot predict length of stay with respect to these variable.

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

```
lin_Mod3<- lm(TOTCHG~., data= df)
summary(lin_Mod3)
```

```
> lin_Mod3<- lm(TOTCHG~., data= df)
> summary(lin_Mod3)

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

Residuals:
   Min     1Q Median     3Q    Max
 -6367   -690   -185    121  43412

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5025.0347   439.6618  11.429  < 2e-16 ***
AGE           133.2083    17.6303   7.556 2.06e-13 ***
FEMALE1      -392.3740   248.6992  -1.578    0.115
LOS           742.9549    35.0060  21.224  < 2e-16 ***
RACE2         458.2628  1084.1236   0.423    0.673
RACE3         330.2836  2626.7849   0.126    0.900
RACE4        -499.4421  1519.3720  -0.329    0.743
RACE5       -1784.6296  1530.4375  -1.166    0.244
RACE6        -594.3428  1857.2266  -0.320    0.749
APRDRG         -7.8175     0.6874 -11.372  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2619 on 490 degrees of freedom
Multiple R-squared:  0.5546,    Adjusted R-squared:  0.5464
F-statistic: 67.78 on 9 and 490 DF,  p-value: < 2.2e-16

> |
```

From the significance codes we can see RACE and Gender i.e. are clearly not affecting the hospital costs.

While the Age and length of stay in days are statistically significant with a very low P-Value. Length of stay is obvious to be significant because the longer you stay higher will be your hospitalisation costs and we can see from the Estimate that for 1 day of stay hospitalisation costs increases by 743.

All Patient Refined Diagnosis Related Groups are also affecting the hospitalization costs as we have seen from one of our prior analysis that diagnosis group 640 have high expenditure. So, May be people with diagnosis group 640 are more likely to hospitalized and this will increase the hospitalization costs.