Fields:

AGE : Age of the patient discharged

FEMALE : Binary variable that indicates if the patient is female (1 = female, 0 =2)

LOS : Length of stay, in days

RACE : Race of the patient (specified numerically)

TOTCHG : Hospital discharge costs

Diagnosis Code : All Patient Refined Diagnosis Related Groups

The goals of this project are:

? To record the patient statistics, the agency wants to find the age category

of people who frequent the hospital and have the maximum expenditure.

? 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 properly utilize the costs, the agency has to analyze the severity of the

hospital costs by age and gender for proper allocation of resources.

? 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.

? To perform a complete analysis, the agency wants to find the variable that

mainly affects the hospital costs.

**Goal 1:** **To record the patient statistics, the agency wants to find the age category**

**of people who frequent the hospital and have the maximum expenditure.**

A screenshot of a computer

Description automatically generated with medium confidenceLet’s do a PIVOT table that shows the most common age group

A screenshot of a video game

Description automatically generated with medium confidence

Highlight the cell according to the value:

Home > Conditional Formatting > Data Bars > Gradient Fill

Chart, histogram

Description automatically generated

So this easily tells you that most patients at this hospital were age 0 so less than one years of age and then it kind of ramps up as you get up in your teenage and later years

A screenshot of a computer

Description automatically generated with medium confidenceNow let’s look at total charges by Age

Table

Description automatically generated

Table

Description automatically generatedA screenshot of a computer

Description automatically generated with medium confidenceWe see that Age 0 has the most charges because that age dominates the population of the hospital, so let’s bring in the average charge.

So every 3 year that came to this hospital, had a average charge of 10 grand. Every 9 year old that came to the hospital had the average total charge of 10 grand.

Question: To record the patient statistics, the agency wants to find the age category

of people who frequent the hospital and have the maximum expenditure.

Answer: Age category that is most frequent and the most total charges to the hospital is ages 0, however ages 3 and 9 have the most average expenditure which about $10k.

**Goal 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.**

We need a table that lists diagnosis group and we want maximum hospitalization which means length of stay and then the associate expenditure associated with that.

Make a new Pivot Table – let’s pull in Diagnosis Code as the rows, Count of Diagnosis Code, Average Length of Stay, Sum of Total Charge, Average of Total Charge

A screenshot of a computer screen

Description automatically generated with medium confidenceTable

Description automatically generated

Table

Description automatically generated

We now want to create a scatterplot of these values to better visualize.

First let’s make a scatch table off to the side that im going to use as the basis for my scatterplot.

Pull in length of Average Length of Stay and Average of TotalCharg2

Table

Description automatically generated

Chart, scatter chart

Description automatically generatedNow prepare the ScatterPlot

We see from the Graph that the Average Length of Stay is relatively short but has a high average cost for that service. For everything in this upper left quadrant is going to be high cost and low length of stay. On the otherhand, people that sit on the hospital 42 days or so and they get out paying 30 grand and it’s still a lot of money – 40 days of service. It’s nothing like being in the hospital for 40 grand.

Question 2: In order of severity of the diagnosis and treatments and to find out theexpensive treatments, the agency wants to find the diagnosis related group that has maximum hospitalization and expenditure.

Answer 2: The Highest amount of being paid about $48,333 which is diagnostic code 911.

**Goal 3: 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.**

A screenshot of a video game

Description automatically generated with medium confidenceLet’s make a new Pivot Table – Female, Charges twice

Table

Description automatically generated

From seeing the table and Conditionally format on each Quandrant. We see the charge basis across both sexes males and female that age zero is the highest ands we know why there were 267 patients of that age – it is no surprise. But now let’s look now, looking at averaghe charge. For males, it is age 3 year old, 11 grand. However, for females – it is females that paid roughly 10 grand. In the total basis, males pay 3 grand in average charges and females pay 25k which a little bit less in average charges.

**Goal 4: 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.**

A picture containing text, building, window blind

Description automatically generatedWhat they are talking about here is multiple regression! So want to predict length of stay, age, and female. For race remember they have the data is coded 1 through 6. When we’re looking at regression modeling, we don’t need like the number 1 is greater than the other numbers. So we need to code it out to a dumby variable to make it work. Use formula =IF(D2=1,1,0) (if race =#, then give us a 1, if not give us a 0). We use this new table for Multiple Regression!

Now go to Data > Data Analysis > Regression

Dependent Variable: Length of Stay ( What do want to predict)

Independent Variables: Age, Female,Race1, Race2,Race3,Race4,Race5,Race6

Table

Description automatically generated

We see that non of the variables are not significant of length of stay because of the p-values are not less than .05. We can conclude that no you cannot predict length of stay from of all these variables.

**Goal 5: To perform a complete analysis, the agency wants to find the variable that**

**mainly affects the hospital costs.**

Now we’re adding Total charges – because of hospital cost to our multiple regression analysis table.

Dependent variable: Total Charge

Independent Variable: Length of Stay, Age,Female,Race

Table

Description automatically generated

We see that Length of Stay, Age, and Sex is a good predictor of Total Charge that mainly affects the hospital costs. When we’re looking at these three statistically is this.

Looking at the coefficients, for intercept – someone is going to be charge $679 before how long they are there before the other coefficients. For Length of stay we have $742 increase per day of the length of stay. For Age we have a $114 increase per year of age increase – as you get older, you pay more. For FEMALE(sex), if you are a female – you will be charge on average $1000.00 less if you were male.