**Project Write Up**

I will be giving a step by step explanation of my source code as well as a written analysis for each question.

Getting things started I imported and opened the required libraries, dplyr and ggplot. I then opened the required dataset and looked at the first few rows.

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.

For the first question I displayed the required data with a histogram as well as using table to look at each individual age and the corresponding age frequency. With both measures it shows that newborns or babies who are less than a year old is the age group/category who visit the hospital the most, and by a wide margin. If you disregard them, its teenagers who have a much higher frequency then those from 2 - 10 years old. For measuring the maximum expenditure, it is unsurprisingly the same, with babies under 1-year old costing almost half as much and the rest of the age groups in the data set. To calculate this, I used summarise and used ggplot for a visual representation.

1. 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.

For the second question I again used summarise to show the sum of all length of stay and expenditure for each diagnosis related group. Since there are so many different diagnosis related groups, finding the maximum values by table or graph would not be very easy or quick to notice. Instead I used which.max of the summarise dataframe. This is created to find the maximum values for both length of stay and expenditure. It showed that the group 640 has the maximum value for both, and thus that is the group that the agency wishes to find.

1. 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 find out if there is a relation between race of the patient and hospitalization costs, I used a simple linear regression with the lm function and displaying the summary. With the summary showing that the coefficient of race has a p-value of 0.686, it is not statistically significant. It also shows that the r-squared value is 0.0003, which is very small, showing that the model does not reliable. With these two statistics, we can safely say that the race of the patient is not related to hospitalization costs in our dataset.

1. 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.

To analyze the severity of hospital costs by age and gender, I decided to again use summarize and a bar graph to explain my results. I created a stacked bar graph with the variable for male or female. With this data it shows overall that the difference in cost between males and females in our dataset is not very large. But looking past babies who are under one year old, it shows that on average, males from ages 2-11 tend to have more hospital costs versus females from ages 12-17 having higher hospital costs.

1. 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.

Similar to question 3, I ran a regression for age, gender, and race on length of stay to see if a significant model could be created. But again, I found that the p-value for each variable was higher than 0.05, and the r-squared value was very low. With this analysis we can say that length of stay cannot be predicted from age, gender, and race.

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

For the final question I created another regression model for each variable against hospital costs to find out which variable mainly affects hospital costs. Analyzing the summary, I found that the variables age, length of stay, and the Diagnosis Related Groups were the only ones which are statistically significant with a very low p-value. After that I looked at the values of each coefficient, and it showed that length of stay has a value of 743.15 compared to 134.69 and –7.79 for age and diagnosis related groups respectively. This means that for every extra day spent for a patient at the hospital, the hospital charges increased by 743.15 units. With this value being much larger than both those for age and diagnosis related groups, we can say that the variable length of stay has the largest effect on hospital costs.