The Exploration and Analyzation of Healthcare in Hospitals within Wisconsin

Cost of Healthcare

Kileigh Boggess

A picture containing text, electronics, computer, computer

Description automatically generated

# Overview

This project has two purposes I wanted to investigate.

* First, I want to find the age category of people who most frequently made hospital visits and who also had the maximum overhead. Second, I wanted to find the group that displayed the maximum hospitalization and overhead costs. I also wanted to see if age and/or gender played a role in the hospitalization numbers and overall costs of those visits. I have done this by analyzing this data provided by Ravi Chaubey, via Kaggle.com.
* Second, I wanted to find the group that displayed the maximum hospitalization and overhead costs.
* I also wanted to see if age and/or gender played a role in the hospitalization numbers and overall costs of those visits.

## Data

This data was provided by Ravi Chaubey, via Kaggle.com. You can find the dataset, and more information about it here - <https://www.kaggle.com/ravichaubey1506/healthcare-cost>. The data consists of the patient’s age, gender, their length of stay, their race/ethnicity, their hospital discharge costs as well as all patient refined diagnosis related groups.

To perform a complete analysis on this matter, we need to find the specific variable that is affecting the hospital costs. According to the Healthcare Costs and Utilization Project, in their article written July 2020, by Lan Liang, Ph.D., Brian Moore, Ph.D., and Anita Soni, Ph.D., healthcare spending in the United States constituted to one-third of all healthcare expenditures. It is said that the five most expensive inpatient conditions were as followed: septicemia, osteoarthritis, liveborn (newborn) infants, acute myocardial infarction, and heart failure. If this article interests you, you can read the article by following this link: <https://www.hcup-us.ahrq.gov/reports/statbriefs/sb261-Most-Expensive-Hospital-Conditions-2017.jsp>.

## Patient Statistics

In this section, we will explore which age group of patients who visited the hospital had the max overhead. In the hospital realm, early patient classification systems like the one used in this dataset – All Patient (AP) DRGs, were created to provide these systems with the types of patients treated with the resources they consume. The focus of this is resource intensity. There are known drawbacks to this system, but APR-DRGs are more relevant than not – clinically they just make sense in the hospital setting.

Started by inserting a new column to calculate the cost of patient stay for each patient. Here is a boxplot and barplot displaying the comparison of the cost and age.

## Cost VS. Age

Chart, box and whisker chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

Here you’ll see that the age groups between three and six years old, had the most expenditure. Of course, this is without considering the potential for outliers. You can also see those ten years old patients displays the longest error bar this is represented by the black line attached to the bar. Error bars in data visualization represent the variability within the data, we use them to indicate error or uncertainty in the data.

Chart, bar chart, histogram

Description automatically generated

We discovered that the average cost for each age group in descending order was as followed:

A picture containing timeline

Description automatically generated

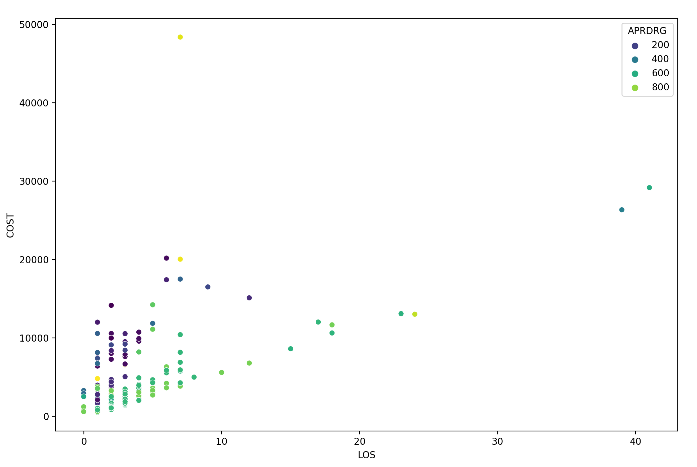
With this information, we were able to conclude that the age groups between zero and one, and twelve through seventeen had the most frequent hospital visits, and that the three through six age group had the max expenditure.

## Maximum Hospitalization and Expenditure

To determine the severity of the diagnosis and treatment methods, we need to focus on the most expensive cases. To do this, we needed to pinpoint the specific age group that had the max hospitalizations and expenditures.

Chart, histogram

Description automatically generated



What we discovered was: That despite the 911 treatment group had the highest expenditure (48388.00 cost), and that the 640 had the higher hospitalization (you can see this by looking at the 640 group and seeing it tower over the others in this graph). This led us to conclude that the 602 group had the highest value of both expenditure and hospitalization because the cost was 29188.00 and the LOS was 41.

## Eliminating Potential Malpractice

To ensure that there was no malpractice committed, we determined that an analysis needed to be done to determine whether the race of the patient was or wasn’t related to the hospitalization costs. From exploring the data, we discovered that based on the overall mean of cost, we’re able to see that the race 2 and 3 had the highest average cost – so, these were the races we focused on.

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

Text

Description automatically generated with medium confidence

As you can see, there were many outliers, most coming from race 1, but we also noticed that the mean for race one was almost identical to the overall mean. But when we viewed race 2, although it only had two outliers, it also had a very high mean.

This information led us to conclude that although the race 2 mean was below the superior limit, we believe it should’ve been investigated anyway because the value of the mean cost, was so much higher than that of the other groups.

## Utilizing Hospital Costs

To properly utilize the costs, we need to analyze the severity of these hospital costs by looking at the age and gender for proper allocation of resources.

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generated

As you can see, it’s possible to see those patients over the age of 12, the number of female patients is higher than the male patients. It was also clear to us that there is also a lack of female patients to be seen between the ages of six and nine. We were also able to conclude that the costs with male patients is larger than that of female patients, this includes almost all the ages, except that of those ages five, eleven, thirteen, and those greater than seventeen.

## The Length of Stay

Since the length of stay is the most crucial factor for inpatients, we knew that what we needed to explore next was the length of the hospital stay was something that could be predicted based on just the age, gender, and race of the patients.

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

As you can see, the regression line is almost parallel to the x-axis (age, gender, and race), this indicated to us that there is no correlation between the variables. When we chose this topic and looked at other research papers, we often saw the same results when others compared these variables.

## What Variable Mainly Affects Hospital Costs?

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

With these visualizations, we were able to easily conclude that the LOS variable and female variable affect the cost the most (slope 720.3 and -462.34).