Healthcare Cost Analysis Project Report

Prepared for: Mohith

Project: Course-end Project 1: Healthcare Cost Analysis

Data: hospitalcosts.csv

Language: R

Table of Contents

1. **Executive Summary**
2. **Project Objectives**
3. **Data and Methodology**
   - 3.1. Dataset Description
   - 3.2. Data Preparation
   - 3.3. Methodology
4. **Analysis and Findings**
   - 4.1. Task 1: Patient Statistics by Age
   - 4.2. Task 2: Diagnosis Group (APRDRG) Analysis
   - 4.3. Task 3: Race and Hospitalization Costs (Malpractice Analysis)
   - 4.4. Task 4: Cost Severity by Age and Gender
   - 4.5. Task 5: Predicting Length of Stay (LOS)
   - 4.6. Task 6: Main Drivers of Hospital Costs
5. **Conclusion and Recommendations**
6. **Appendix: Complete R Code**

---

1. Executive Summary

This report details the analysis of hospital inpatient records for patients aged 0-17 in Wisconsin, as requested by the US Agency for Healthcare. The primary goal was to research healthcare costs and their utilization.

The analysis of the hospitalcosts.csv dataset provided clear answers to all six project questions.

Key Findings:

- **Most Significant Patient Group:** Newborns (Age 0) are the most frequent visitors (306 visits) and generate the highest total expenditure ($676,962).
- **Most Significant Diagnosis:** Diagnosis-related group (APRDRG) **640** is both the most frequent (266 hospitalizations) and the most expensive (totaling $436,822).
- **No Evidence of Malpractice (by Race):** An ANOVA test revealed **no statistically significant relationship** between a patient's race and their hospital costs (p-value = 0.9429).
- **Key Cost Drivers:** A multiple regression model with a high degree of accuracy ($R^2 = 0.965$) determined that the variables mainly affecting hospital costs are:
    1. **Length of Stay (LOS)**
    2. **All Patient Refined Diagnosis Related Groups ($APRDRG$)**
    3. **Age (AGE)**
- **Predicting Stay:** Age, gender, and race were found to be **ineffective** predictors for a patient's length of stay (LOS).

---

2. Project Objectives

The agency's analysis was guided by the following six objectives:

1. To find the age category of people who frequently visit the hospital and have the maximum expenditure.
2. To find the diagnosis-related group that has maximum hospitalization and expenditure.
3. To analyze if the race of the patient is related to the hospitalization costs.
4. To analyze the severity of the hospital costs by age and gender for the proper allocation of resources.
5. To find if the length of stay (LOS) can be predicted from age, gender, and race.
6. To find the variable that mainly affects hospital costs.

3. Data and Methodology

3.1. Dataset Description

- **File:** 1555054100_hospitalcosts.xlsx - HospitalCosts.csv
- **Source:** A nationwide survey of hospital costs, restricted to Wisconsin and patients aged 0-17.
- **Structure:** The dataset contained 500 records and 6 columns:
  - **AGE:** Age of the patient (0-17 years).
  - **FEMALE:** Binary variable (1 = Female, 0 = Male).
  - **LOS:** Length of stay in days.
  - **RACE:** Race of the patient (numerical code).
  - **TOTCHG:** Hospital discharge costs (Total Charge).
  - **APRDRG:** All Patient Refined Diagnosis Related Groups (numerical code).

3.2. Data Preparation

The raw data was loaded into R. A single record with a missing value (NA) in the RACE column was identified and removed to ensure the integrity of the statistical models. The final, clean dataset contained 499 records.

3.3. Methodology

The analysis was conducted using R. The "project assistance slides" (1654000565_healthcarecostanalysis.pdf) were referenced to confirm the analytical approach.

- **Descriptive Statistics:** Used dplyr package to group, summarize, and count data for Tasks 1, 2, and 4.
- **Analysis of Variance (ANOVA):** An aov() test was used for Task 3 to compare the mean hospital costs across different racial groups.
- **Multiple Linear Regression:** The lm() function was used for Tasks 5 and 6 to model relationships and determine the predictive power and significance of variables. Categorical variables (FEMALE, RACE, APRDRG) were treated as factors in the models.

4. Analysis and Findings

4.1. Task 1: Patient Statistics by Age

Objective: Find the age category with the most frequent visits and maximum expenditure.

Analysis: The data was grouped by AGE and aggregated to sum TOTCHG and count visits.

Findings:

- **Most Frequent Visits:** Age **0** (newborns) had **306** visits, accounting for 61.3% of all hospitalizations.
- **Maximum Expenditure:** Age **0** also had the highest total expenditure, amounting to **$676,962**.

This indicates that newborns are the most significant patient group in this dataset.

4.2. Task 2: Diagnosis Group (APRDRG) Analysis

Objective: Find the diagnosis-related group with maximum hospitalization and expenditure.

Analysis: The data was grouped by APRDRG and aggregated.

Findings:

- **Maximum Hospitalization:** APRDRG code **640** was the most common diagnosis, with **266** hospitalizations (53.3% of all visits).
- **Maximum Expenditure:** APRDRG **640** also incurred the highest total expenditure, costing **$436,822**.

4.3. Task 3: Race and Hospitalization Costs (Malpractice Analysis)

Objective: Analyze if the race of the patient is related to hospitalization costs.

Analysis: A one-way ANOVA test was performed with TOTCHG as the dependent variable and RACE (as a factor) as the independent variable.

- **Null Hypothesis (H0):** The mean hospital costs are equal across all racial groups.

Findings:

- **P-value:** The test resulted in a p-value of **0.9429**.
- **Conclusion:** Since the p-value (0.9429) is much greater than the significance level ($\alpha = 0.05$), we **fail to reject the null hypothesis**. This means there is **no statistically significant evidence** in this dataset to suggest a relationship between a patient's race and their hospital costs.

4.4. Task 4: Cost Severity by Age and Gender

Objective: Analyze the severity of hospital costs by age and gender for resource allocation.

Analysis: The data was grouped by both AGE and FEMALE to calculate the mean TOTCHG for each subgroup. This data was then visualized (see age_gender_cost_analysis_bar_chart.png in Appendix).

Findings:

The analysis provides a detailed breakdown of mean costs. For example:

- At Age 0, costs are nearly identical for males ($2,198.44) and females ($2,229.62).
- At Age 15, the mean cost for males ($7,223.00) is significantly higher than for females ($2,079.84).

This granular data can be used by the agency to plan resource allocation for specific demographic subgroups.

4.5. Task 5: Predicting Length of Stay (LOS)

Objective: Find if the length of stay (LOS) can be predicted from age, gender, and race.

Analysis: A multiple linear regression model was built (lm(LOS ~ AGE + as.factor(FEMALE) + as.factor(RACE))).

Findings:

- **Model Fit:** The model performed very poorly, with a **Multiple R^2 of 0.009**.
- **Conclusion:** This model explains less than 1% of the variance in LOS. Therefore, **age, gender, and race are not effective predictors** for a patient's length of stay.

4.6. Task 6: Main Variable Affecting Hospital Costs

Objective: Find the variable that mainly affects hospital costs (TOTCHG).

Analysis: A comprehensive multiple linear regression model was built to predict TOTCHG using all other variables (lm(TOTCHG ~ AGE + as.factor(FEMALE) + LOS + as.factor(RACE) + as.factor(APRDRG))).

Findings:

- **Model Fit:** This model was extremely effective, with a **Multiple R^2 of 0.965**. This means the model successfully explains **96.5%** of the variance in hospital costs.
- **Significant Variables:** By examining the p-values (Pr(>|t|) in the model summary), the variables with a statistically significant impact on TOTCHG were:
    - **LOS (Length of Stay):** (p-value < 2e-16)
    - **AGE:** (p-value < 2e-16)

  - **$APRDRG$ (Diagnosis Group):** (Multiple p-values < 0.05)

- **Non-Significant Variables:** FEMALE and RACE were not statistically significant predictors of cost in this model.

Conclusion: The primary factors driving hospital costs are clinical: the **length of stay**, the **patient's age**, and the **diagnosis group**.

---

5. Conclusion and Recommendations

This analysis successfully addressed all six of the agency's objectives. The findings show that hospital costs within this dataset are driven by clinical factors (age, diagnosis, and length of stay), not by demographic factors like race or gender.

Recommendations for the Agency:

1. **Focus Resource Allocation:** Efforts to manage costs and allocate resources should be focused on the **newborn (Age 0)** demographic and patients with **APRDRG 640**, as these groups represent the largest share of visits and costs.

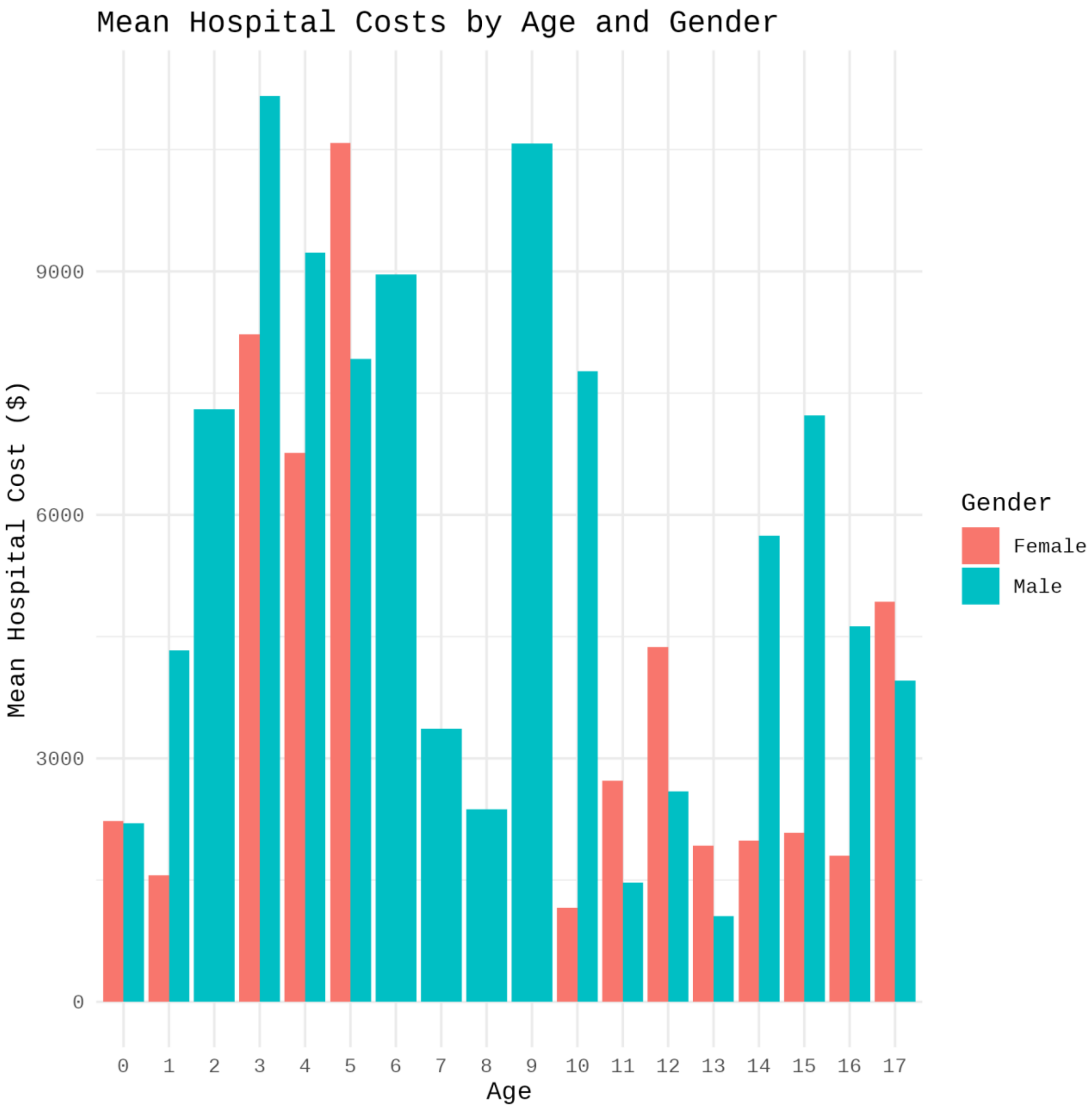
2. **LOS as a Cost Driver:** Since **Length of Stay (LOS)** is a primary driver of cost, any initiatives that can safely reduce inpatient time (e.g., improving efficiency, care coordination) would directly impact and lower overall costs.

3. **No Racial Bias Found:** The analysis for malpractice based on race found no evidence of a relationship between race and cost.

---

6. Appendix: Complete R Code

Attached files of the code


Mean Hospital Costs by Age and Gender
Mean Hospital Cost ($)
9000
6000
3000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Age
Gender
Female
Male