Introductory Analysis on US Hospital Cost Report

May 27th, 2025

Table of Contents

1 Introduction

U.S. hospital costs represent a critical focus in public health management. The Centers for Medicare & Medicaid Services (CMS) requires all certified institutional providers to submit annual cost reports, which are published on the CMS website as CSV datasets. In this project, the author downloaded CMS hospital cost reports for the years 2017 through 2022. These datasets were preprocessed, analyzed, and visualized to identify introductory-level trends and patterns in U.S. hospital costs. Specifically, the study investigates (1) the overall trajectory of total and per-capita costs, (2) the distribution of costs across different hospital types of control, and (3) the behavior of the cost-to-charge ratio. To adjust for population differences, U.S. Census Bureau population data for the same period were obtained from census.gov. Data preprocessing was performed in SAS, while all subsequent analyses—such as trend modeling and cost-share calculations—were conducted in R.

2 Data Preprocessing

All six annual CSV files (2017–2022) were imported into SAS Studio. A SAS library was configured, and a macro was developed that (a) calls PROC IMPORT for each year—parameterized by the macro variable &yr—and (b) appends a new variable YEAR to each dataset reflecting its calendar year.

After importing, the individual year datasets were concatenated. SAS's automatic type inference generated inconsistencies across years, so variables flagged for type conflicts were identified and manually coerced to consistent types to ensure seamless merging.

To comply with SAS and R naming conventions, all variable names were standardized: spaces were replaced with underscores, and select names were abbreviated for readability. A mapping of original names to revised names, along with variable descriptions from the CMS data dictionary, was recorded in an Excel appendix.

Variables containing redundant information were dropped to streamline the dataset and eliminate potential ambiguities.

I outputted the preprocessed dataset in the format of sas7bdat.

3 Analysis

3.1 Initial Check

Prior to analysis, the proportion and count of missing values were computed for every variable. Branch-specific cost variables exhibited high missingness, whereas the total cost variable showed minimal gaps. Based on these findings, subsequent analyses focused on variables with robust data coverage.

3.2 Trend of Total Cost

A preliminary line plot of total cost for all 50 states proved overly dense. States were grouped into the four U.S. Census regions and nine divisions (per Wikipedia) by joining an auxiliary lookup table via dplyr. Total cost trends were then visualized at the regional, divisional, and individual-state levels. Overall, total costs rose modestly from 2017 to 2022, with the South region consistently incurring the highest aggregate cost, followed by the Midwest and West.

To account for population differences, Census Bureau data for 2017–2022 were merged with cost data. Per-capita total costs were computed for each state, region, and division. Choropleth maps were generated to display both absolute per-capita costs and average annual percent changes.

An R function was created to automate trend-plot generation: it reshapes input data via pivot_longer(), sets year as the x-axis, and calls ggplot() with a consistent aesthetic.

3.3 Cost Composition by Type of Control

Type of Control is the categories of hospital, there are three main types of control: 1) Voluntary; 2) Proprietary; 3) Governmental. The Voluntary is also considered as non-profit, which is organized by non-profit corporation, and it is exempt from paying federal income or state and local property taxes. The voluntary hospital is the most common type of hospital in US healthcare institution. Proprietary hospital is also referred as investor-owned hospitals, it runs like a large business and the objective is earning profit. The governmental hospital is also referred as public hospital, is a kind of hospital that funded and owned by government, generally federal government. The governmental hospital is a kind of public service and eligible for many healthcare functions. In the CMS Hospital Cost Report, the all hospitals are divided into 13 groups based on three main types as above. This introductory analysis re-grouped all types of controls into three main types for convenience.

In RStudio codebook, I created a new data frame by “dplyr” package to include all type of control, there are two variables, one is named “type_of_control”, it is a numeric index that assign values from 1 to 13, another is the “toc_name” that is the list of type of control based on data dictonary of the cost report. Then I created another variable named “toc_group”, which assigned the three main types based on the value of variable “type_of_control”. After that, I joined the type of control table with original dataset, then I got the main types of all records.

In the next step, I calculated the cost share of each type of control based on states, regions, and

divisions respectively, then I used charted bar chart to visualize how different types to share the total cost. I also focused on the states that has different trends in total per capita cost, to compare if there are different distribution in control types. Overall, I researched online that the voluntary hospital always has higher cost since it is non-profit healthcare institution.

3.4 Cost-to-charge Ratio

The cost-to-charge ratio is another variable in CMS Cost Report, it is defined as the total cost divided by total charges. The total cost and total charges are also variables in the dataset. Therefore, I tried to use total cost to divide the total charge, and I find that there are 117 observations have relatively big gap between calculated ratio and given ratio. I am not sure if it is the problem from raw data.

I used line plot to indicate the trend of cost-to-charge ratio, and I found that it varies and unstable. I tried to fit a series of regression model and Random Forest and set the Cost-to-Charge Ratio as outcome variable, the model does not show the significant fit to dataset.

4 Results

4.1 Total Per Capita Cost

The trend of total hospital per capita cost group by region is shown as figure 4-1. The Northeast region exhibited the highest per-capita cost, rising from roughly $3,000 in 2017 to about $3,800 in 2022. The Midwest and West followed closely, while the South ranked lower on a per-capita basis despite its high aggregate cost. Non-contiguous areas remained below $2,000 per person.

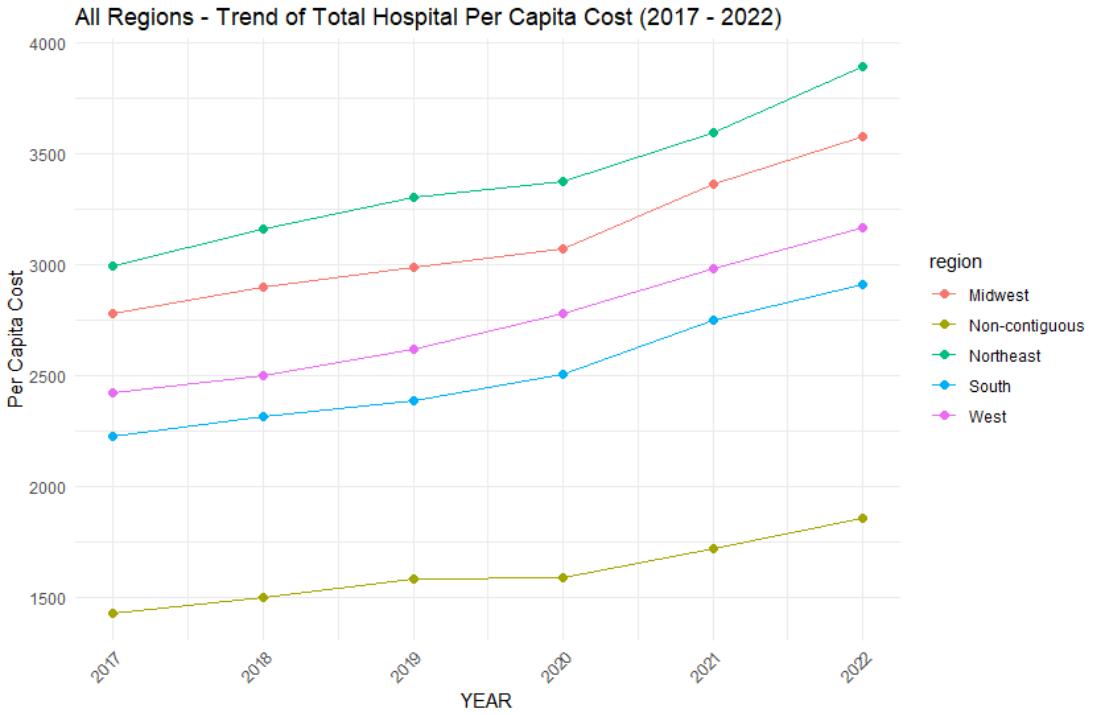

Figure 4-1 All Regions Trend of Total Hospital Per Capita Cost (2017-2022)

As for divisional breakdown, New England led all divisions—exceeding $4,000 per capita in 2022—followed by the Mid-Atlantic. West North Central, East South Central, and Pacific

divisions clustered between $2,000 and $2,500.

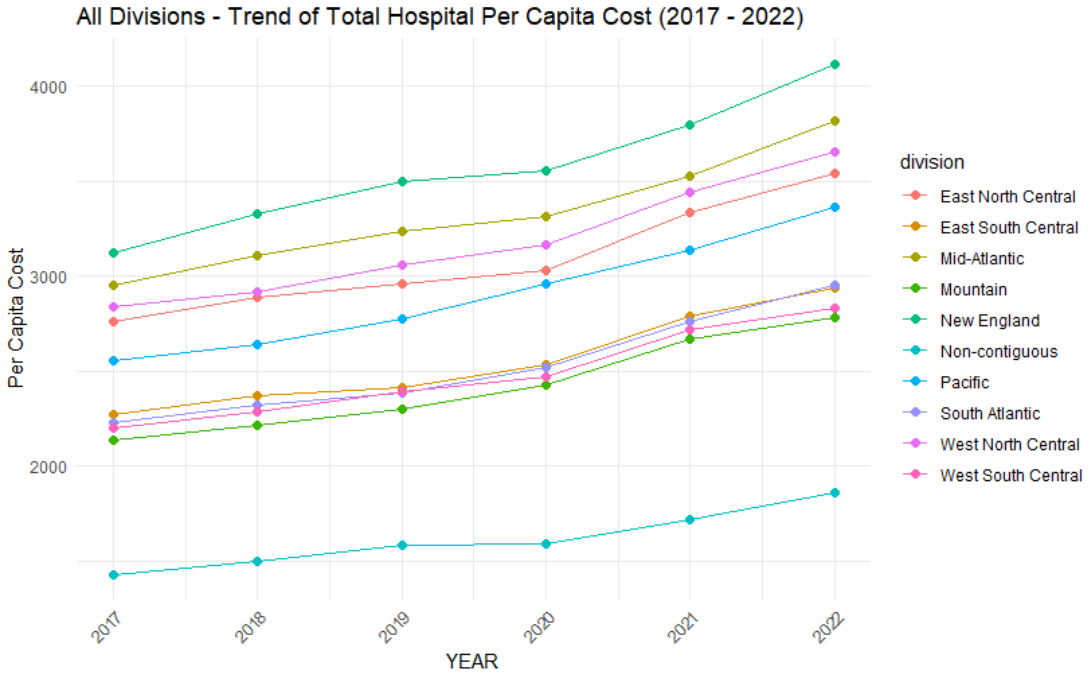

Figure 4-2 All Divisions Trend of Total Hospital Per Capita Cost (2017-2022)

Choropleth maps highlight state-level differences in average per-capita cost and mean annual growth. North Dakota showed irregular year-to-year changes, whereas Maine recorded the fastest growth rate.

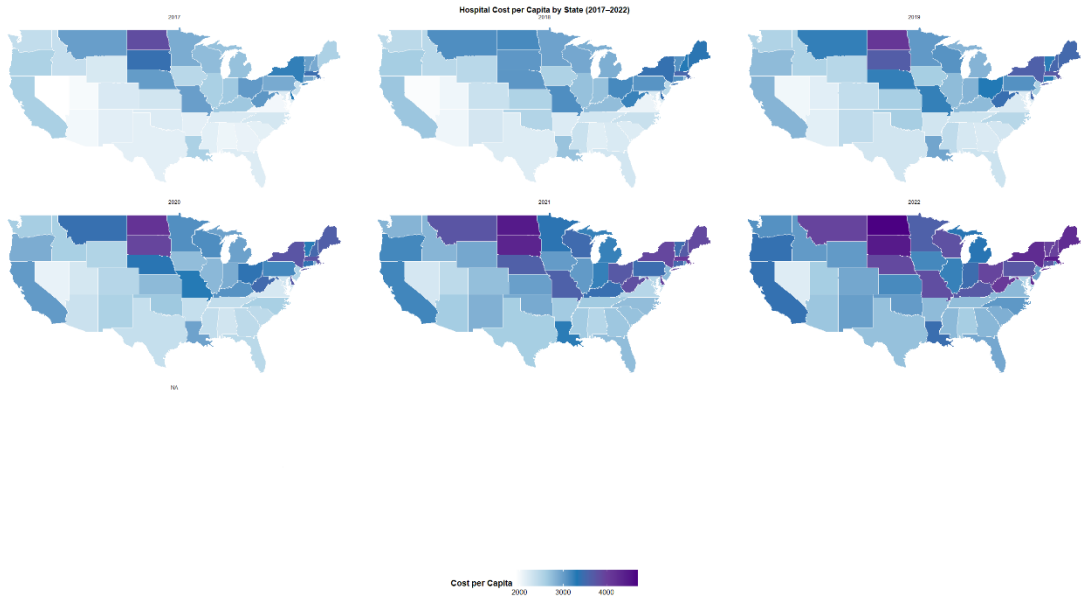

Figure 4-3 Hospital Per Capita Cost (2017-2022)

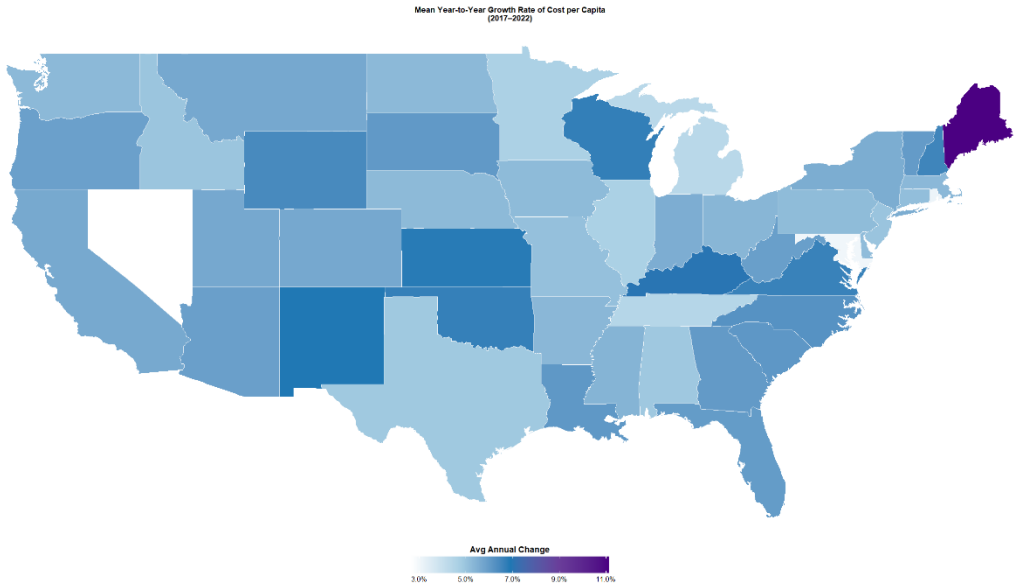

Figure 4-4 Mean Year-to-Year Growth Per Capita Cost (2017-2022)

4.2 Type of Control

I used charted-bar graph to visualize the cost share of three different types of control. Overall, the voluntary hospital accounted the highest percentage of cost in the US, and the Northeast and Midwest, which have higher per capita cost than other regions, displayed higher cost share in voluntary hospital cost.

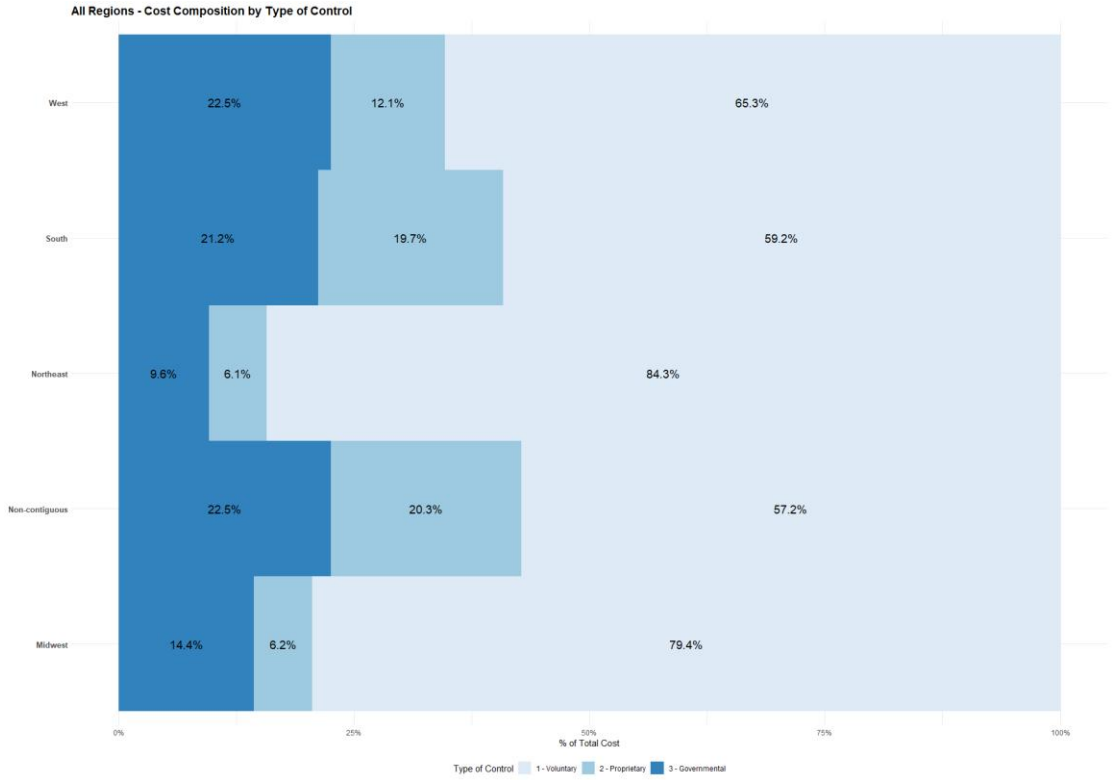

Figure 4-5 All Regions Cost Composition by Type of Control (2017-2022)

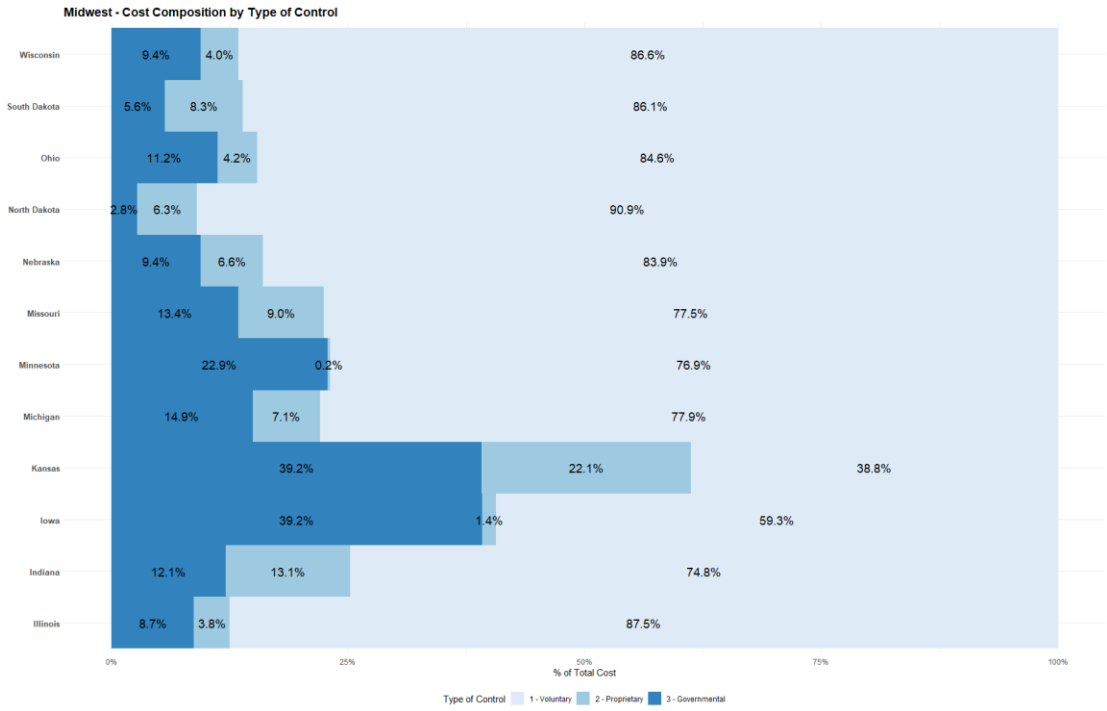

Figure 4-6 Mean Year-to-Year Growth Per Capita Cost (2017-2022)

Based on Figure 4-5 and Figure 4-6, it is cleared that North Dakota has relatively less cost share in governmental and proprietary hospitals, with the highest cost share in voluntary among all states in Midwest. Conversely, the Kansas has the lowest cost share in voluntary, as well as highest share in proprietary hospital. And compared with Figure 4-3, the North Dakota has the highest per capita cost among Midwest, and the Kansas has the lowest per-capita cost among Midwest. After further investigation, I found that the voluntary cost share on North Dakota went a drop in 2018 but rose back in 2019, then kept increasing until 2022, this is a good point to explain why North Dakota experienced fluctuations from 2017 to 2022.

4.3 Cost-to-Charge Ratio

There is not unique trend of cost-to-charge ratio (CCR) shown among all states, but the non-contiguous has the highest cost-to-charge ratio. This is because only few of voluntary or governmental hospitals are recorded in non-contiguous regions, and those hospitals always have high cost but less income. I also try to use CCR as outcome variables for model prediction, I tried several statistical algorithms but all performed bad. I assumed that this is because CCR is not the outcome variable in hospital management, thus the model does not show any input variables can significantly impact on result.

I decided to focus on CCR variable, I used box plot to check how CCR distribution in each year, after removing the outlier or some data points may be mistakenly recorded. The Figure 4-7 is the box plots of CCR.

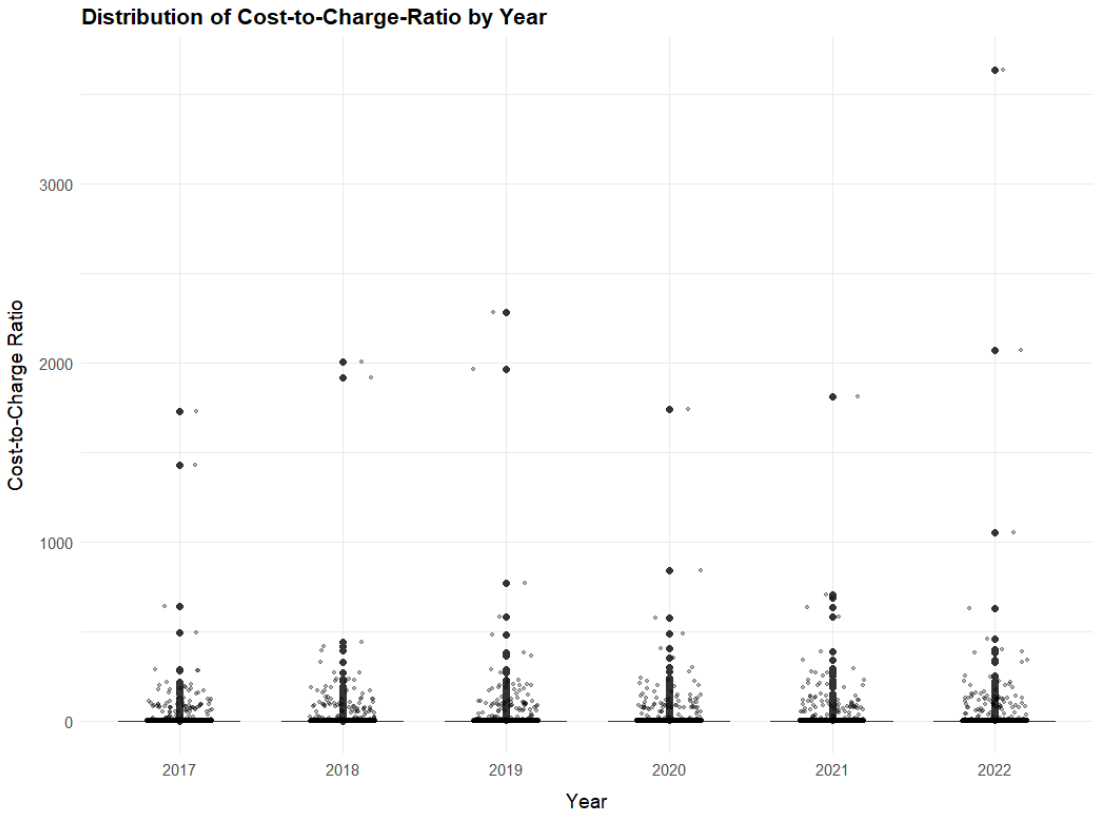

Figure 4-7 Distribution of Cost-to-Charge Ratio by Year

I decided to check the observations that CCR>=10, which has 517 observations, 1.41% of all records. I found that there are 507 records from governmental hospital, 9 records from voluntary hospitals, and only 1 from proprietary hospital. I assumed that the high outlier is highly related to the type of control, and the governmental hospital will have overall higher CCR than two other states, after checking the average CCR over years based on type of control as shown in figure 4-8, my assumption is confirmed.

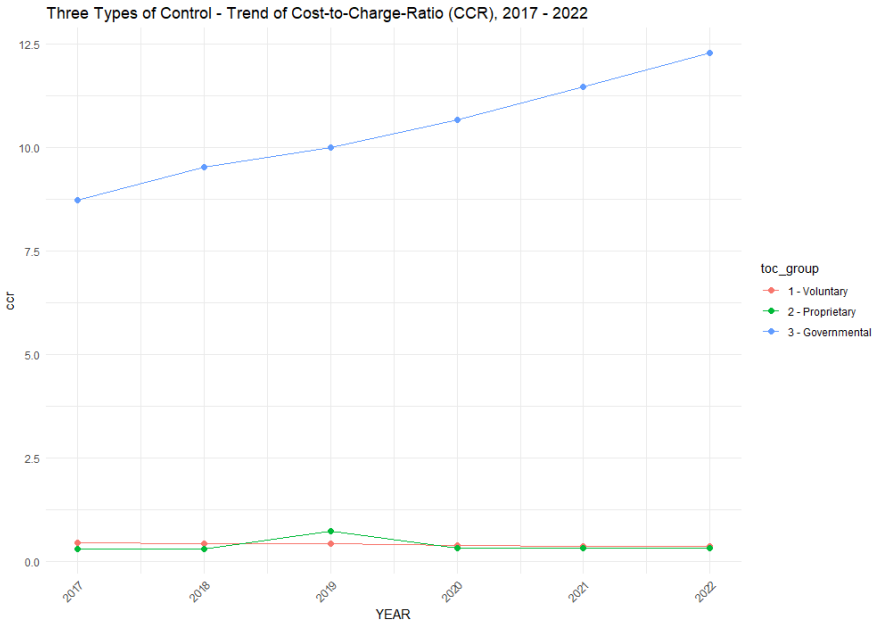

Figure 4-8 Type of Control and Trend of CCR (2017-2022)

5 Conclusions

This introductory analysis of CMS hospital cost reports (2017–2022) reveals several key insights:

5.1 Rising Costs with Population Adjustment

Although aggregate hospital costs grew modestly over six years, per-capita costs increased more substantially—especially in the Northeast and New England divisions—highlighting the importance of adjusting for population when assessing cost burdens.

5.2 Dominance of Non-Profit Sector

Voluntary (non-profit) hospitals consistently contribute the largest share of total costs, particularly in regions and divisions with higher per-capita expenditures.

5.3 Control-Type and Cost Dynamics

State-level examples (e.g., North Dakota vs. Kansas) demonstrate how variations in control-type composition align with per-capita cost differences, suggesting structural factors in hospital ownership influence financial outcomes.

6 Limitations

6.1 Variable Granularity

The CMS cost reports contain over 110 detailed financial variables. Due to missingness and accounting complexity, this analysis focused on a subset of high-coverage variables. Future work should leverage specialized accounting knowledge to interpret additional line items.

6.2 Model Specification

Attempts to model CCR were inconclusive, likely because the dataset lacks certain operational or reimbursement variables that drive cost structures. More comprehensive data (e.g., payer mix, case severity) may improve predictive insights.

Reference

1. https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States
2. https://en.wikipedia.org/wiki/Non-profit_hospital
3.https://data.cms.gov/sites/default/files/2022-12/47c231b2-3b8e-4b97-bbdd-d92e762330ff/Hospital%20Cost%20Report%20Data%20Dictionary_508.pdf
4. https://www.census.gov/data.html

Author Information

Ruifeng Wang
Drexel University, M.S. in Business Analytics, 2024
Email: ruifeng973@gmail.com