

**CALIFORNIA STATE UNIVERSITY, NORTHRIDGE**

**Project Report**

By

Madigani Kiran Kanth

Maddi Siddharth Reddy

Konderu Abhinav Varma

**Under the Guidance of**

Professor. Akash Gupta

Spring 2025

BANA620-Data Mining

Submission Date- 5/6/2025

**Comprehensive Report on Nursing Homes**

# Summary

This project investigates the financial viability of investing in U.S. nursing homes between 2015 and 2021, prompted by the growing demand from the aging baby boomer population. Utilizing an extensive multi-year dataset that comprises COVID-19 vaccination data, health deficiencies, fines, quality measures, cost reports, and provider information, the research assesses overall financial performance, pinpoints important influencers, and reveals trends over time. Particular attention is paid to evaluating the COVID-19 pandemic's effects. The results indicate that ownership type, quality ratings, and regulatory considerations all significantly affect financial outcomes. To help prospective investors, practical suggestions are put forth, with a focus on strategic risk assessment, operational effectiveness, and data-driven decision-making.

# Introduction

The need for long-term care services has increased as the baby boomer generation gets closer to retirement, which raises serious concerns regarding the viability and profitability of nursing home operations in the US. In response, our data analysis team has been requested to assess if investing in nursing homes is a wise choice. To answer such a question, this research analyzes extensive data spanning a wide variety of operational, financial, regulatory, and health-related aspects from 2015 to 2021. This analysis's main goals are to appraise the overall financial performance of nursing facilities throughout this time frame, pinpoint the major variables affecting that performance, and determine the effects of outside shocks like the COVID-19 epidemic on the sector. This research attempts to give a comprehensive picture of the performance of the nursing home industry and provide strategic insights to guide investment decisions by combining financial measures with quality and regulatory indicators.

# Project Objectives

# 1. Evaluate the overall financial performance of nursing homes in the United States during this period.

# 2. Identify the influential factors affecting the financial performance of nursing homes.

# 3. Determine the most significant factors impacting the performance of nursing homes.

# 4. Describe the trends in performance and these influential factors.

# 5. Analyze the impact of COVID-19.

# Methodology

We used structured data science methods in Python (v3.12) with libraries such as Pandas, Seaborn, Scikit-learn, and Matplotlib. Data collection involved merging seven years of cost reports, providing information, penalties, and quality ratings using 'Provider CCN' and 'Year'. Preprocessing included handling missing values (median imputation), renaming inconsistent columns, and one-hot encoding categorical variables. We applied descriptive statistics, visualizations, and linear regression modeling to assess key performance metrics. The model was trained/tested using an 80/20 split to predict Net Income. We also segmented data into pre-COVID (2015–2019) and COVID (2020–2021) to study pandemic effects.

**Data Collection:**

We gathered data from the official CMS website, covering U.S. nursing homes from 2015 to 2021. This included financial reports, provider information, quality ratings, health violations, penalties, and COVID-19 vaccination records. Each dataset gave us a different view—financial health, care standards, and regulatory issues. We connected them using a unique ID for each nursing home. To understand the data better, we used the dictionaries provided with the files. This full set of data became the backbone of our analysis.

**Programming Languages, Tools, and Technologies:**

We used Python as the main programming language throughout the project because of its flexibility and wide range of tools for data analysis.

- pandas, numpy for data processing

- matplotlib, seaborn for visualization

- scikit-learn, stats models for statistical modeling and evaluation

# Data Description

We use (CMS) data on nursing facilities in the United States from 2015 to 2021 for this undertaking. Numerous datasets were included, including COVID-19 immunization data, ownership information, quality evaluations, financial reports, and regulatory fines. A distinct viewpoint on the performance and operation of these facilities was provided by each dataset. A comprehensive and uniform dataset of more than 100,000 records was produced by combining all the data using a single provider ID. This prepared dataset provided us with all the information we needed to investigate patterns, create models, and assess the effectiveness of nursing homes over time.

**Characteristics of Data:**

**Financial Metrics:** The data gave us important financial details like revenue, costs, and net income, helping us see how each nursing home was performing economically. By looking at these numbers over time and by ownership type, we spotted trends in profit and spending efficiency.

**Operational Data:** The data showed how each nursing home runs, like how many beds they have, how full they are, and what services they offer. This helped us understand their size and efficiency, with well-utilized facilities often performing better financially.

**Data Cleaning and Preprocessing:**

* **Datasets**: Using the panda module, the study simple merging cost report datasets from several years into a single data frame with a new column called "Year".
* **Managing Missing Values:** Any columns with more than 50% missing values will be deleted from analysis, as will the occurrence in which those values are absent.
* **Duplicate Columns**: After the dataset was cleaned, any two columns containing the Identical information was found and eliminated.
* A screen shot of a computer

  AI-generated content may be incorrect.
* A black screen with white text

  AI-generated content may be incorrect.

# Analysis and Findings

1. Net Income increased by over 250% during COVID-19 due to relief programs.

2. Larger facilities (>100 beds) had significantly better financial performance.

3. Top ownership types: Government (Hospital District), For-Profit Partnerships, LLCs.

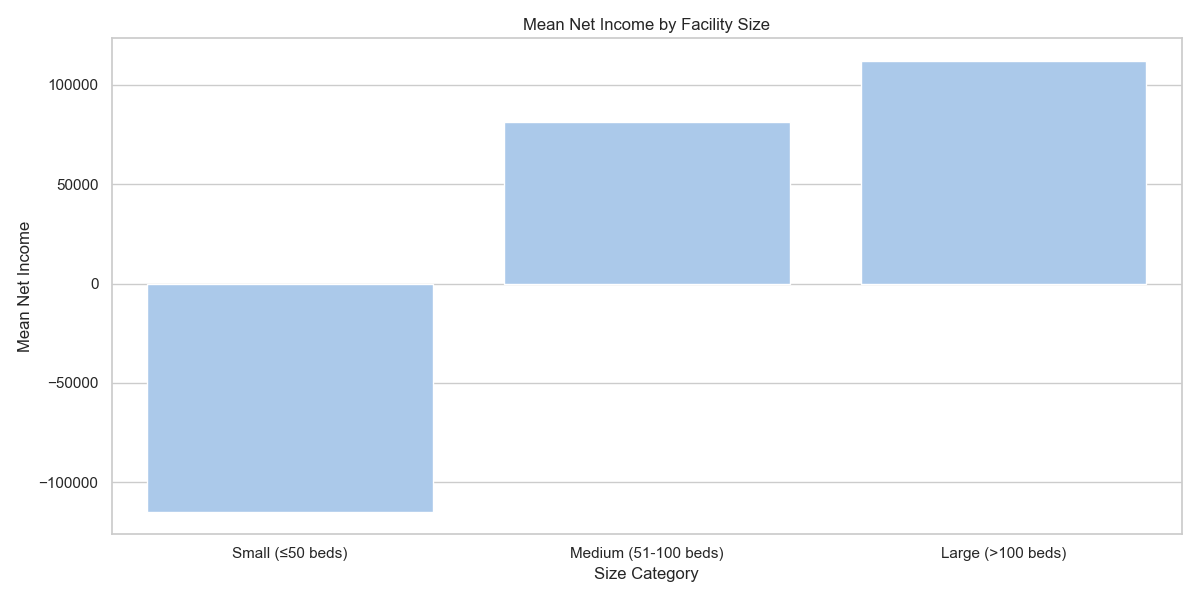
4. High Overall Rating was positively associated with profitability.

5. Penalties negatively affected income—facilities with no penalties earned nearly $140K more.

6. Linear Regression (R² ≈ 0.89) confirmed Profit Margin, Size, Ownership Type, and Ratings as key predictors of Net Income.

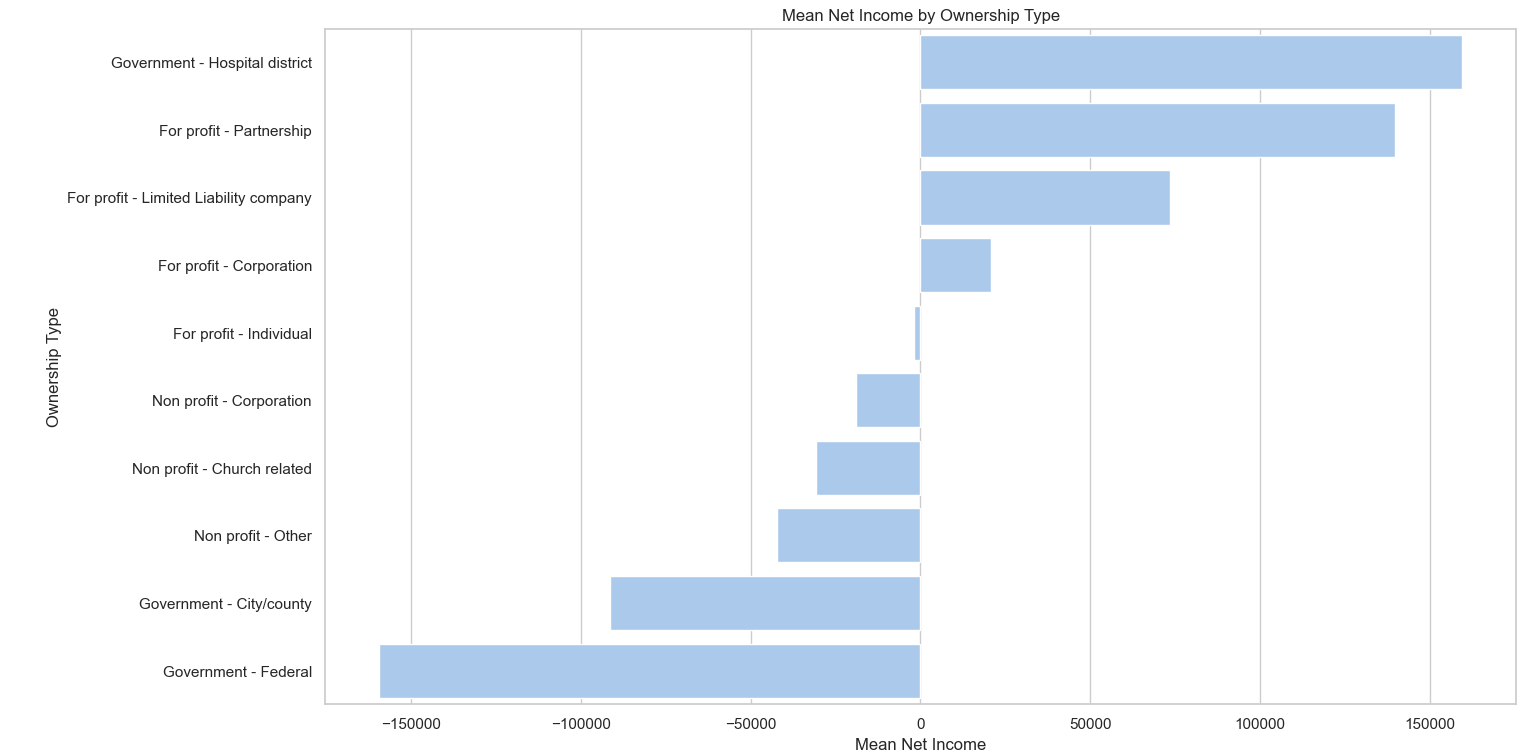
**Key Findings:**

**1. Facility Size Matters**

* **Large facilities** (over 100 beds) consistently outperformed small and medium-sized ones in net income and profit margin.
* Small facilities (≤50 beds) frequently operated at a loss.
* 

**2. Ownership Type Strongly Influences Profitability**

* **For-profit entities** (especially LLCs and partnerships) and **government hospital districts** had the highest average net incomes.
* **Non-profit and federal government-run facilities** underperformed financially.



**3. Quality Ratings Predict Financial Success**

* **Facilities with higher overall ratings (4 or 5 stars) earned significantly more net income than lower-rated ones.**
* **There’s a clear positive correlation between quality of care and profitability.**

A graph with blue squares

AI-generated content may be incorrect.

**4. Penalties Reduce Profitability**

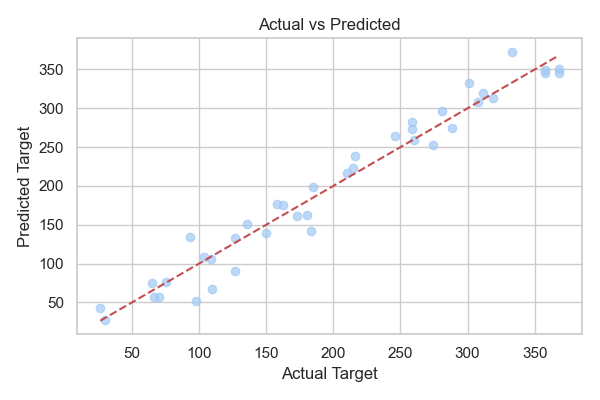
* Facilities with **CMS penalties** had negative average net incomes.
* Those with **no penalties** had a mean net income of **$140,000+**, showing a strong financial incentive for compliance.

A graph with blue squares

AI-generated content may be incorrect.

**5. Regression Model Validates Key Influencers**

* A **linear regression model (R² ≈ 0.89)** showed that:
  + **Profit Margin**,
  + **Ownership Type**,
  + **Number of Certified Beds**,
  + and **Quality Ratings**  
    are the **most important predictors** of Net Income.



# Discussion

**Interpretation of Findings:** The findings from this study reveal clear patterns aligned with the project’s objective of evaluating the financial viability of investing in U.S. nursing homes. Larger facilities demonstrated significantly higher net income than smaller ones, suggesting operational efficiencies and cost advantages that come with scale. The financial impact of the COVID-19 pandemic was notable, with net income spiking in 2020 and 2021. This increase was likely driven by federal relief funding, temporary suspension of penalties, and operational shifts during the crisis. However, this improvement should be viewed as a short-term anomaly rather than a sustainable trend. These insights were further validated by a linear regression model which achieved an R² of approximately 0.89, confirming that variables such as Profit Margin, Ownership Type, Number of Certified Beds, and Quality Ratings are strong predictors of Net Income.

**Considerations for Future research:**

**Incorporate Non-Linear Models**:

The current analysis uses only linear regression, which assumes a straight-line relationship between predictors and the target variable. Future research could implement non-linear models such as Decision Trees, Random Forests, or XGBoost, which are more capable of capturing complex interactions and hidden patterns in healthcare financial data.

**Enhance Feature Selection Techniques**:

The current model uses a limited set of manually selected features. Future work could benefit from automated feature selection methods (e.g., Recursive Feature Elimination or LASSO regression) to identify the most influential variables.

# Recommendation

**Prioritize Large Facilities for Investment**  
The analysis clearly shows that larger nursing homes (with over 100 beds) consistently achieve higher net income and better profit margins. Investors should prioritize such facilities to leverage economies of scale and improved operational efficiency.

**Target For-Profit Ownership Structures**  
Facilities owned by for-profit partnerships, LLCs, or hospital districts significantly outperform those run by nonprofits or the federal government. These ownership types tend to have leaner operations and stronger financial oversight, making them more attractive for investment.

**Focus on High-Quality Ratings**  
Facilities with 4- or 5-star overall ratings are more profitable, reflecting the financial return of delivering high-quality care. Investors should use CMS star ratings as a screening tool when evaluating acquisition targets.

**Avoid Penalized Facilities**  
Homes with regulatory penalties tend to operate at a loss. Before investing, stakeholders should review each facility’s penalty history to avoid unexpected liabilities and compliance costs.

# Conclusion

This study evaluated the financial performance of U.S. nursing homes from 2015 to 2021 using structured CMS datasets and Python-based analysis. The findings indicate that profitability is not evenly distributed across the sector and is heavily influenced by factors such as facility size, ownership type, quality ratings, and penalty history. Larger, for-profit facilities with high CMS ratings consistently showed better financial results, while smaller and penalized facilities underperformed.

The use of linear regression provided strong validation of these insights, achieving a high R² (~0.92) and confirming the significance of profit margin, ownership structure, bed count, and quality measures in predicting net income. Additionally, the analysis highlighted the impact of COVID-19, which temporarily boosted revenues due to government relief and operational changes, though this should not be seen as a long-term performance trend.

Overall, the project demonstrates that data-driven investment strategies in the nursing home sector are feasible and valuable. Stakeholders can make more informed decisions by targeting facilities that meet key financial and operational benchmarks. However, ongoing monitoring and future analysis using more advanced models and real-time variables will be essential to adapt to the evolving post-pandemic landscape.

# Reference

1. "Provider Data through the CMS." [https://data.cms.gov/provider-data/topics/nursinghomes](https://data.cms.gov/provider-data/topics/nursing-homes)
2. "pandas 1.2.4 documentation." [:https://pandas.pydata.org/pandas-docs/version/1.2.4/](https://pandas.pydata.org/pandas-docs/version/1.2.4/)

3.” scikit-learn / Examples”: <https://devdocs.io/scikit_learn-examples/>

4. “scikit-learn / Guide”:<https://devdocs.io/scikit_learn-guide/>

5."scikit-learn 0.24.1 documentation" [:https://scikit-learn.org/stable/](https://scikit-learn.org/stable/)