

Skilled Nursing Facilities Performance Analysis

Measurement Period: 2015 – 2021



Ashley Cortez, Angelica Verduzco

BANA 620

May 7, 2024

Table of Contents

Executive Summary.....	3
Introduction.....	3
Methodology.....	4
Data Description.....	5
Analysis and Findings.....	6
Analytical Question 0: Evaluate the overall financial performance of nursing homes.....	7
Analytical Question 1: Are there specific types of facilities that are more profitable/tend to have a higher net income?	9
Rural vs. Urban:.....	9
Type of Ownership.....	11
Geographic Location.....	15
Analytical Question 2: Is there a relationship between facility profitability and Overall Rating?.....	18
Analytical Question 3: Is there a relationship between facility profitability and penalties?..	20
Analytical Question 4: Was there an impact on overall profitability for skilled nursing facilities during COVID?.....	22
COVID Stats – Rural vs Urban (2021):.....	23
COVID Stats – Ownership Types (2021):.....	25
COVID Stats by Region (2021):.....	27
Models Tested.....	29
Linear Regression Models:.....	29
Linear Regression 1:.....	29
Linear Regression 2:.....	29
Decision Tree:.....	30
Random Forest:.....	30
Discussion.....	31
Recommendations and Conclusion.....	34
Appendices.....	35
Final Google Colab Notebook:.....	35
References.....	35

Executive Summary

For this project, we were assigned to analyze the financial performance of skilled nursing facilities to determine whether it is advisable for our client to invest in these types of organizations.

Throughout this report we will cover nursing facility profitability (defined as the net income) as it relates to various nursing facility attributes. These attributes include rural versus urban locations, types of ownership and geographic location. We also cover an analysis of the relationship between profitability and overall facility rating, the relationship between fines and penalties with profitability, and lastly the effects the COVID-19 pandemic had on nursing home profitability.

From the data provided and analysis conducted, our team is able to recommend the investment of highly rated, for-profit facilities located in Urban areas that also maintain a strong regulatory compliance.

We've also identified areas of opportunity to further research the financial performance of skilled nursing facilities. These areas include extending the measurement period, investigating the distribution of facility resident demographics, and investigating the impact marketing efforts have on facility profitability.

Introduction

As the population continues to age, the demand for nursing home facilities has increased. In response to the increase in demand, there is a rising interest in investing in nursing homes. The data analysis team has been asked to investigate if nursing home investment is advisable to our client. After acquiring U.S. nursing home datasets from 2015–2021 we aim to analyze the key factors that contribute to financial performance. Given the impact of the COVID-19 pandemic on the healthcare system, our analysis also explores how the pandemic influenced various factors within nursing homes. Our analysis provides an understanding of the financial performance of the U.S nursing homes, from identifying influential factors, trends, and understanding the impact of the COVID-19 pandemic. With the insights from the analysis we can provide recommendations to our client regarding the investment in nursing homes.

Methodology

For our data analysis methodology, we took the following approach in order to ensure proper and accurate analytical results:

1. **Data Collection:** We used the provided dataset that were provided with the assignment. Specifically, we included data from the Cost Reports, Provider Info Reports, Penalty Reports and COVID Vax Reports.
2. **Data Analysis Techniques:** We completed the analysis for this project by using Python in Google Colab. Throughout our analysis, we utilized various techniques that ultimately aided us in completing any predictive models. Some of these techniques include:
 - a. T-tests/ANOVA Tests
 - b. Model Evaluation Techniques such as R-Square and MSE
 - c. Linear Regression Models
 - d. Decision Tree Models
 - e. Random Forest Models

Data Description

1. Dataset Descriptions:

- a. **Cost Report Dataset (2015–2021):** Before cleaning, the cost reports totaled over 106K rows and 137 columns. The reports included in this dataset included records of each Skilled Nursing Facility's cost report by year. For each cost report, there was provider information such as Provider Number, Address, and a few columns that describe the type of facility. There were also many columns that included numerical data to describe the facility's income, assets, and annual fees.
 - b. **Provider Info Dataset(2015–2021):** The Provider Info reports included more detailed information as it pertained to each provider. These reports were collected for each year throughout the measurement period. Some provider information included their ownership type (i.e., for-profit, non-profit, or government owned), number of certified beds, average number of residents per day, various facility ratings, incident counts, fine amounts and more. Before cleaning, the provider info reports totaled over 108K rows and 100 columns.
 - c. **Penalty Dataset(2015–2021):** The Penalty reports included records for whenever a provider had a penalty for that year. Penalty types included Fines and Payment Denials. If the penalty was a fine, there would be a fine amount associated with it. If the penalty was a payment denial, there would be a value that provided the length of the denial in days. Before cleaning, the penalty reports included over 71.4K rows and 14 columns.
 - d. **COVID Vax Dataset (2020–2021):** The 2021 COVID Vax report included the provider number, state and percentage of residents vaccinated and percentage of staff vaccinated. The 2022 report included metrics for the percentage of residents/staff that completed primary vaccinations and the percentage of residents/staff that completed the most recent round of vaccinations.
2. **Data Preprocessing and Preparation:** For each report type (Cost, Provider Info and Penalty), there was a report for each year from 2015 to 2021. Upon initial investigation, we found that there were varying column names within each report type. Some reports even had columns that did not exist in

previous year reports. We decided to merge the yearly reports for each report type. This involved renaming columns to be consistent across each year. In a newly merged dataframe, we also added a column for Year so that we did not lose the time series data within each report type. In order to further clean each dataset, we also addressed variables that had more than 50% null values. Since certain columns were missing data for over 50% of the dataset, we decided to drop these from the final data frames.

Analysis and Findings

For our analysis, we developed analytical questions (AQs) to help guide our approach for predictive models. The initial questions can be found below:

- AQ0: Evaluate the overall financial performance of nursing homes.
- AQ1: Are there specific types of facilities that are more profitable/tend to have a higher net income? Facility features analyzed:
 - Rural vs. Urban
 - Type of Ownership
 - Geographic Location (national regions)
- AQ2: Is there a relationship between facility profitability and Overall Rating?
- AQ3: Is there a relationship between facility profitability and penalties?
- AQ4: Was there an impact on overall profitability for skilled nursing facilities before/during COVID?

Analytical Question 0: Evaluate the overall financial performance of nursing homes.

Our team used the cost report's Net Income variable to assess nursing home profitability. To get a better understanding of financial performance, we first developed a statistical summary of the variable in the cost report data:

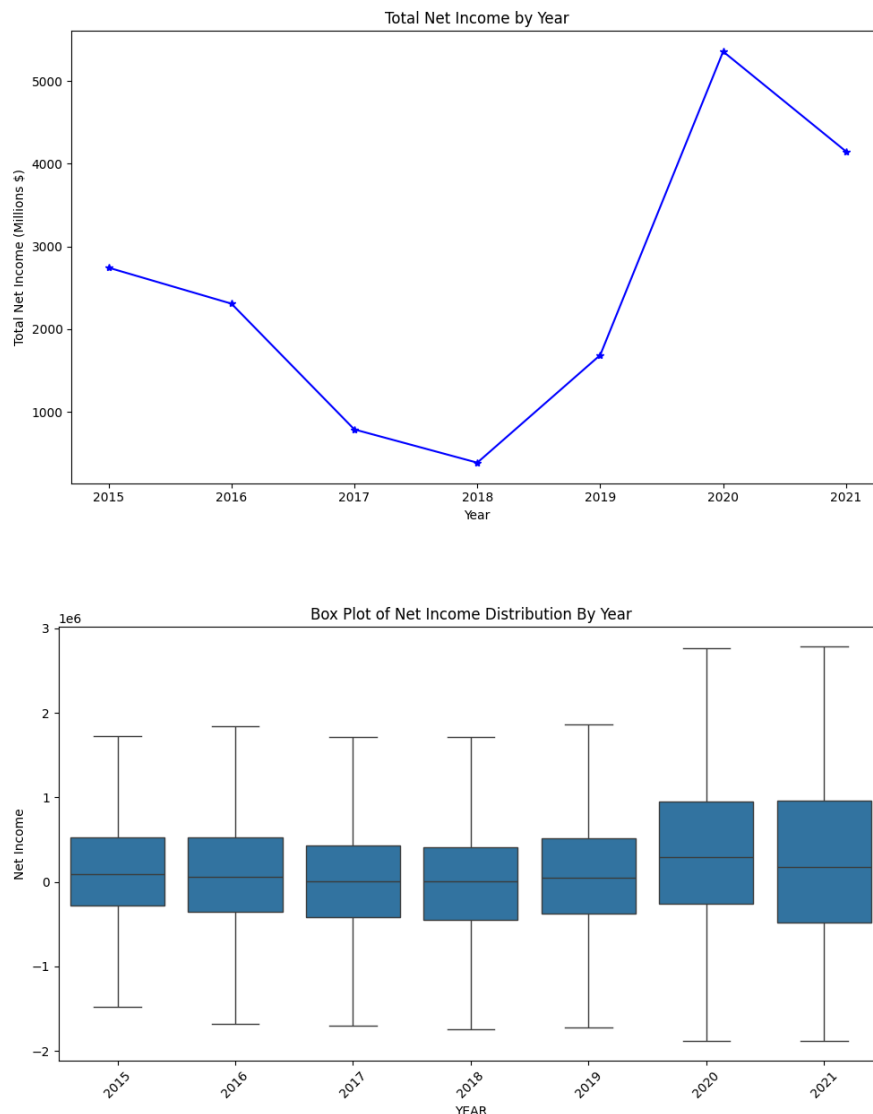
Summary Stats for Cost Reports from 2015 - 2021	Overall
Count	106,269
Mean	\$163,879
Standard Deviation	\$1,050,338

This gives us a general idea of what the average net income is across all providers from 2015–2021.

We also included the summary statistics for net income by year. From the code below, we can see that 2017 and 2018 had a lower average net income than other years. Another notable insight was that in 2019, the average net income began to increase and average net income ultimately reached a significantly higher average in 2020 and 2021.

Summary Stats for Cost Reports by Year	Count	%	Mean
2015	15,402	14.49%	\$178,122
2016	15,104	14.21%	\$152,804
2017	15,433	14.52%	\$51,010
2018	15,142	14.25%	\$25,449
2019	15,182	14.29%	\$111,074
2020	14,949	14.07%	\$358,325
2021	15,057	14.17%	\$275,506

Below we've provided a visual representation of how overall cost fluctuated for nursing facilities throughout the measurement period as well as the distribution of net income by year.



As shown in the previous summary statistics outputs, we can see a dip in net income in 2018 and a spike in 2020. Net income distribution by year shows that 2020 and 2021 had a higher median net income than previous years. From this high-level analysis of the cost report data, we hypothesize that there was a significant fluctuation in net income from 2019–2021 primarily due to the COVID pandemic.

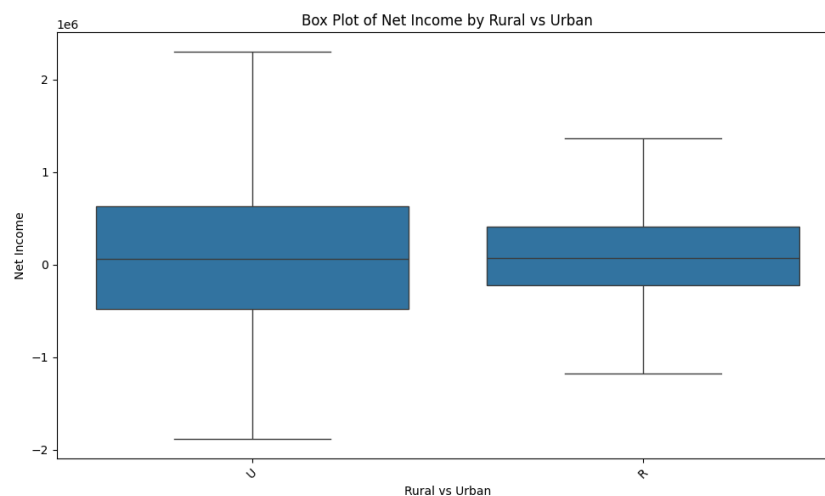
Analytical Question 1: Are there specific types of facilities that are more profitable/tend to have a higher net income?

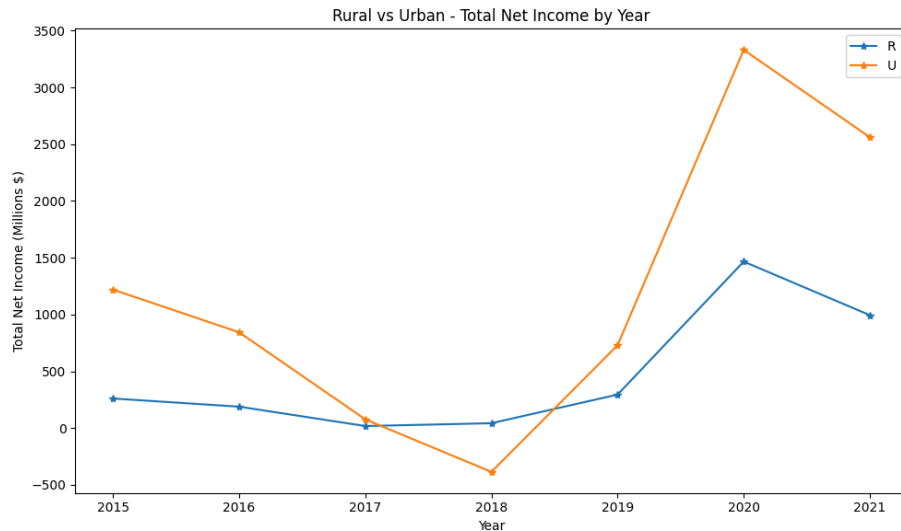
Rural vs. Urban:

The first facility attribute we will look at is the difference between Rural and Urban identified facilities. Below are the summary statistics for net income for both facility types. We can see that the cost report dataset primarily consists of 'Urban' facilities. Although this is the majority, the average net income is higher for 'Rural' facilities.

Summary Statistics Rural vs Urban	Count	%	Mean
Rural	28,306	27.2%	\$115,350
Urban	75,776	72.8%	\$110,527

Below, we've included a visual representation that displays the total Net Income distribution for facility types Urban and Rural as well as the distribution of Net Income by year per facility type. The median net income for both facility types are very similar, however 'Urban' facilities have a larger range distribution of net income, indicating that there is potential for more fluctuation in net income than for 'Rural' facility types.





Rural vs Urban t-test	Results
T-statistic	-3.867
P-value	0.00011

From the second visual above, we see a decreasing trend from the years 2015 to 2017 for 'Urban' facilities, where it drops below zero – this indicates a net loss for that facility type. However, after 2018 we see a spike in net income where it peaks in 2020. Despite a decline in 2021, net income for 'Urban' facilities still remains relatively high. In comparison, we have net income for Rural facilities which show a continuous decrease until 2020 where there is a significant increase year-over-year. Although, initially Urban and Rural facilities showed similar trends after the year 2017 their financial paths diverged.

We've also included a one-way t-test result to assess the significance in the difference between Urban and Rural facilities. The low p-value allows us to determine that there is a significant difference between the net income levels of these two facility types.

From this analysis we can hypothesize that the spike in Net Income for both Urban and Rural facilities in 2020 is associated with the COVID-19 pandemic.

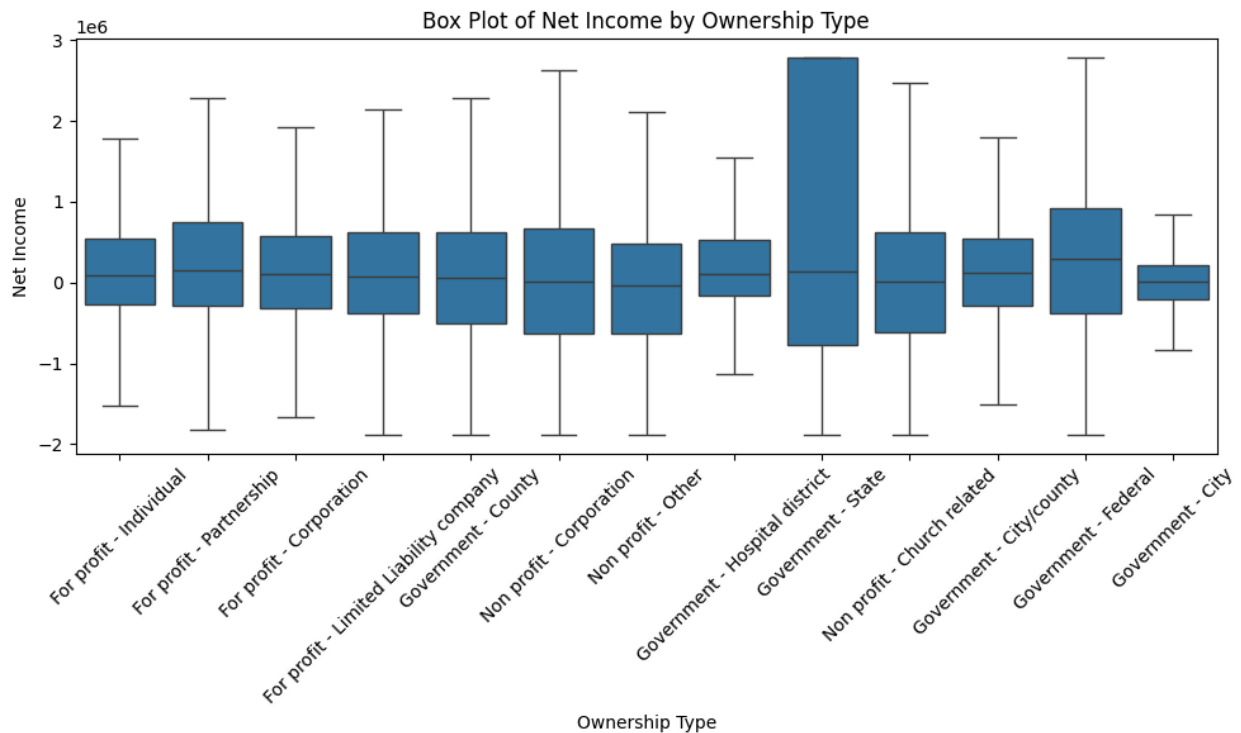
Type of Ownership

The next facility attribute we investigated was ownership type. Below we've included the summary statistics for net income for each ownership type. From the outputs below, we can see that 'For Profit - Corporation' accounted for 50% of the cost report records. However, 'Government - State' had the highest average net income, while the lowest (and negative) average net income was attributed to 'Non-Profit - Other' during the measurement period.

It's important to note that although the highest average net income and lowest net income callouts were made, both ownership types have a minimal presence in the cost report dataset.

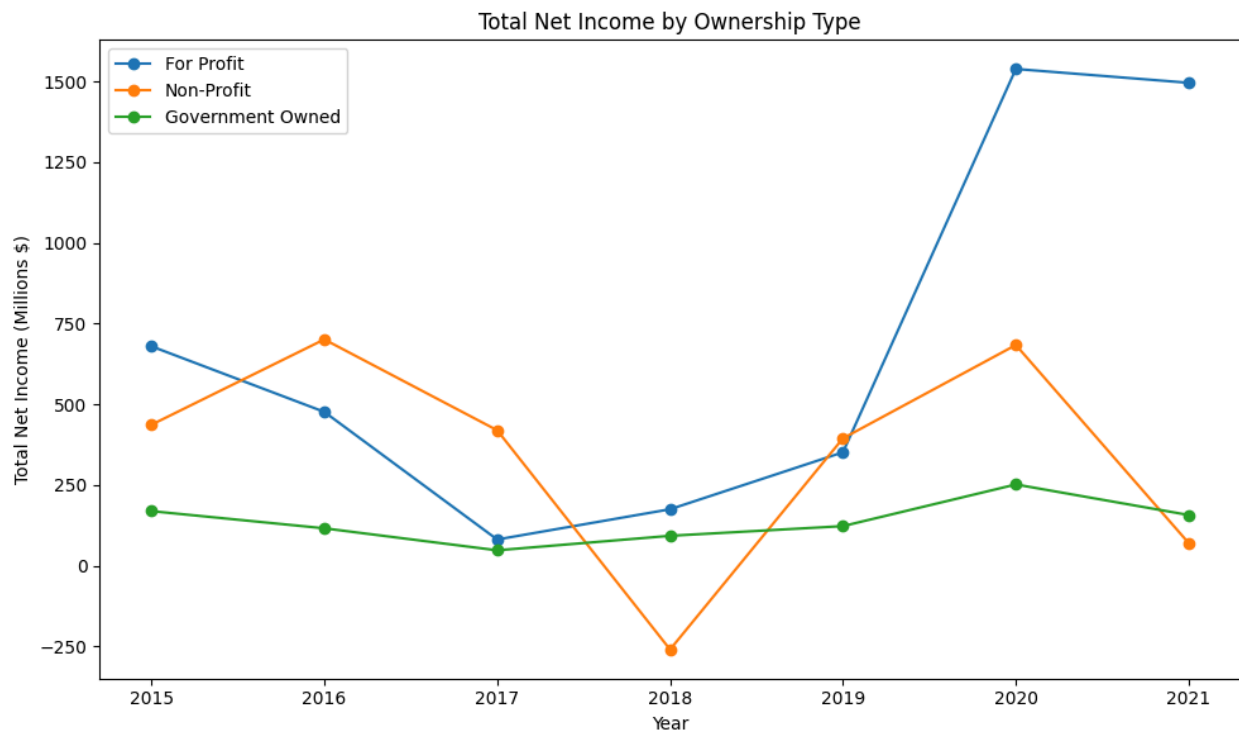
Ownership Type	Count	%	Mean
For Profit - Corporation	53,861	50.72%	\$169,008
For Profit - Individual	5,966	5.62%	\$189,996
For Profit - Limited Liability	13,541	12.75%	\$161,990
For Profit - Partnership	5,416	5.10%	\$276,688
Government - City	512	0.48%	\$68,535
Government - City/County	414	0.39%	\$146,740
Government - County	2,334	2.20%	\$116,848
Government - Federal	86	0.08%	\$349,183
Government - Hospital district	1,306	1.23%	\$186,492
Government - State	729	0.69%	\$450,847
Non-Profit - Church Related	2,577	2.43%	\$114,102
Non-Profit - Corporation	17,164	16.16%	\$126,345
Non-Profit - Other	2,295	2.16%	-\$7,885

When looking at the boxplot below, we found that the 'Government - Hospital District' had the largest range in distribution of net income and 'Government - Federal' had the highest median net income when compared to other ownership types.



Since there are many different types of ownership for nursing facilities, we decided to analyze ownership net income trends at a higher level. We grouped the ownership types by 'For Profit', 'Non-Profit', and 'Government Owned' to see if there were any noteworthy differences.

With the chart below, we observed that the decline in net income in 2018 (that was seen in AQO) was primarily driven by 'Non-Profit' type facilities. 'For Profit' facilities saw the greatest spike in net income from 2019 to 2020. We hypothesize that the need for skilled nursing facilities increased exponentially during the pandemic. As more people became ill, more people were likely urgently searching for accommodations for their older family members. As 'For Profit' facilities encompass the majority of the cost report dataset, we can assume that there were more of these facilities accessible and able to accommodate the influx in patients.



In order to confidently say that there is a significant difference in net income between ownership types, we conducted an ANOVA test that provided the results below:

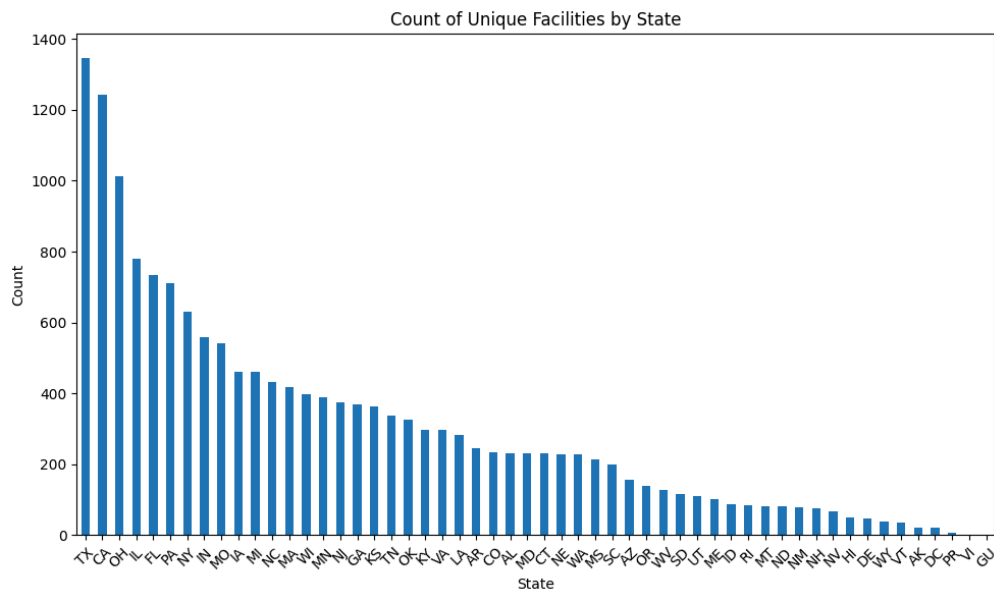
ANOVA Test – Ownership Types								
	Overall	2015	2016	2017	2018	2019	2020	2021
ANOVA F-Statistic	31.053	2.673	5.208	8.446	18.599	1.341	27.049	74.825
P-Value	3.323^{-14}	0.0691	0.0055	0.0002	8.761^{-09}	0.2617	1.975^{-12}	6.695^{-33}

The overall p-value given in the ANOVA test indicates that the net income differences between ownership types is statistically significant. Since this facility attribute is statistically significant, we will identify it as an influential factor that affects nursing facility performance.

We also included ANOVA tests on ownership type for each year. 2015 and 2019 were the only years that did not have statistical significance, according to their p-values.

Geographic Location

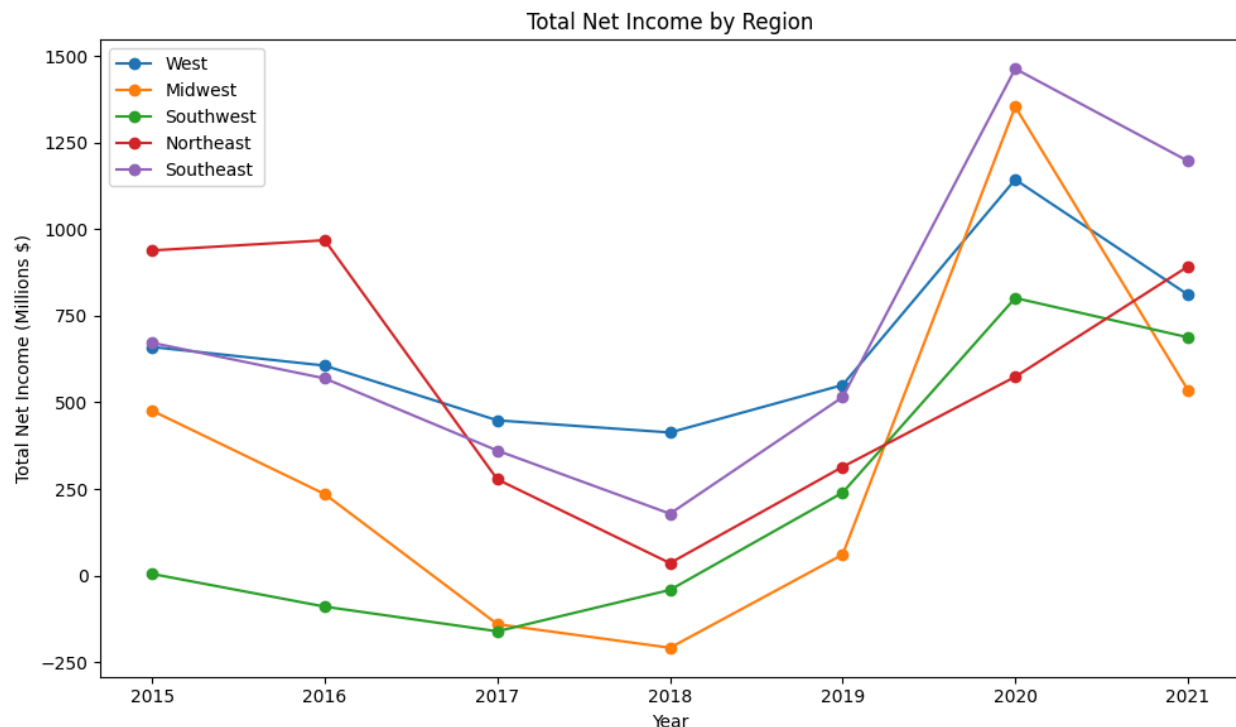
As a first step to analyze geographic location, we visualized the count of unique facilities by state. From the graph below, we can see that the majority of records in the cost report dataset can be primarily attributed to Texas, California, Ohio, Illinois, and Florida. As most of the top 10 states are known to have high population density, it makes sense that there are more facilities to accommodate appropriately.



There are many states included in the dataset, so instead of dropping states from the analysis, our team decided to group the states by region. The regions include Western states, South Western states, North Eastern states and South Eastern states. Once we grouped the states in this way, we were able to conduct an ANOVA test to assess whether the difference in net income between regions was statistically significant.

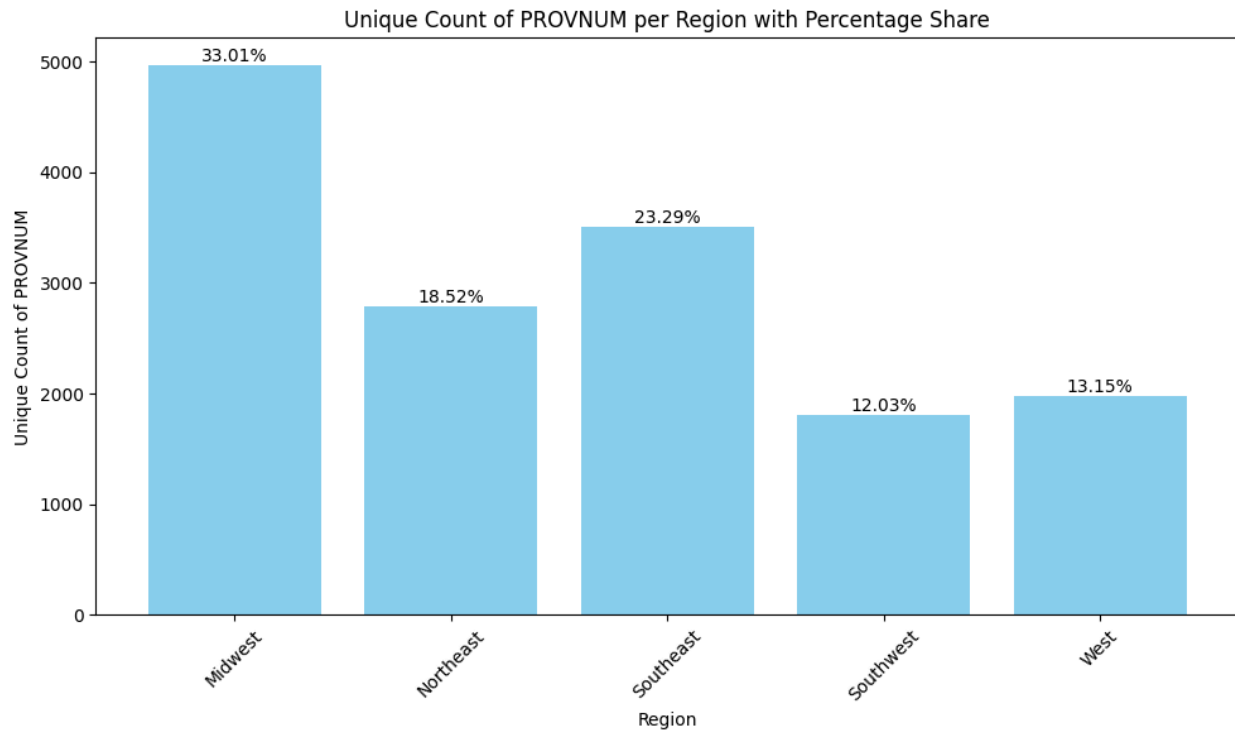
ANOVA Test – State Regions								
	Overall	2015	2016	2017	2018	2019	2020	2021
ANOVA F-Statistic	187.065	51.462	56.464	29.552	20.591	11.128	39.956	2.068
P-Value	4.550 ⁻¹⁶⁰	5.245 ⁻³³	3.442 ⁻³⁶	5.177 ⁻¹⁹	2.689 ⁻¹³	2.744 ⁻⁰⁷	1.168 ⁻²⁵	0.102

From an overall standpoint, the difference between each region's net income is statistically significant. 2021 was the only year that did not have statistical significance, according to its p-value.



We also visualized net income per year for each national region. From the line graph above, we can see that the West and South East regions typically had a higher net income than other regions from 2017 to 2021 (Note: 35% of the dataset was Southeastern region facilities). The North Eastern states started 2015 with the highest net income, but significantly decreased from 2016 to 2017. This region was also the only region to maintain a steady increase in net income after 2020. South Western states continuously had the lowest net income when compared to other regions.

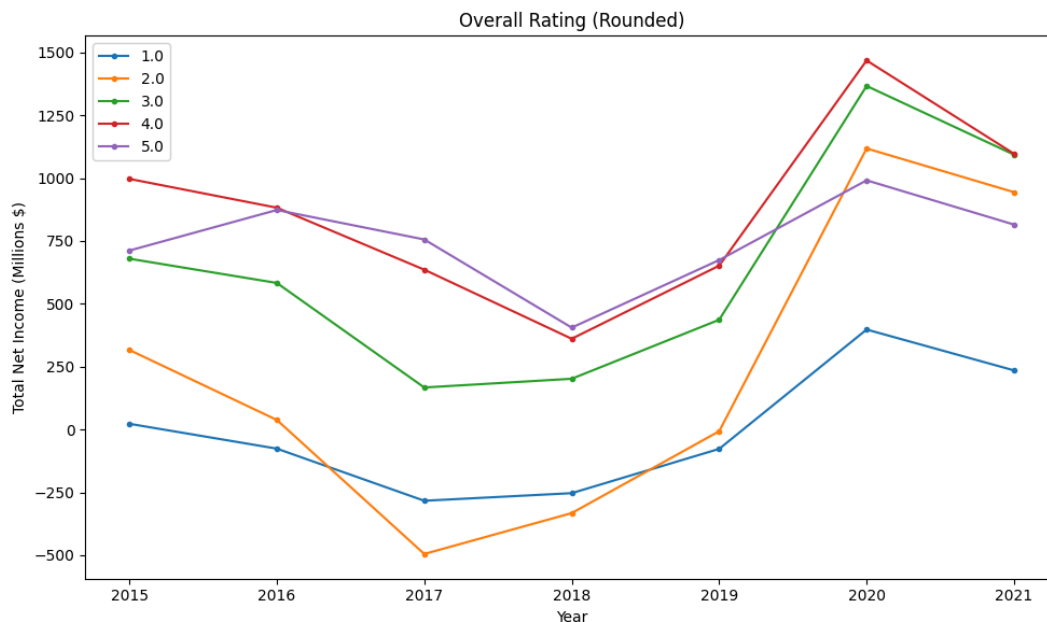
We've also included the chart below as a consideration that the amount of facilities are not evenly distributed across regions.



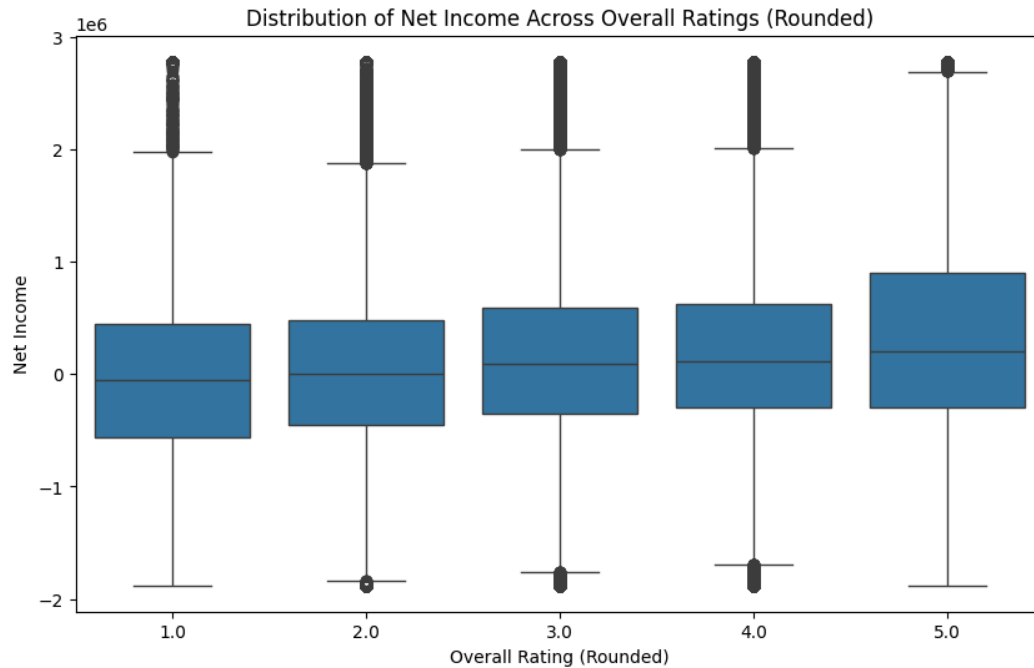
This distribution could explain why Midwestern states had a significantly higher upward trend in net income from 2020 to 2021. More facilities in the region likely received more patients with COVID, therefore increasing their overall net income.

Analytical Question 2: Is there a relationship between facility profitability and Overall Rating?

In our analysis we next focused on identifying if there is a relationship between facility profitability and overall rating. We visualized and captured the trend of the Overall Rating score distribution from facilities across Net Income for each year. From the visual below we can identify that facilities with an overall rating of 4.0 and 5.0 tend to achieve a higher Net Income specifically between 2018–2021. Lower overall rating scored facilities have a lower Net Income, however, Net Income increases for all facilities during 2020–2021. We can hypothesize that this general increase in Net Income for all facilities regardless of Overall Rating score is due to the COVID pandemic.



With the boxplot below, we can identify the distribution of Overall Rating scores from facilities across Net Income. The graph reveals a general trend of facilities with a higher overall rating score having a higher Net Income. The boxplot also highlights the median of facilities demonstrating how those with a higher Overall Rating scores generally have a higher median income. In addition, facilities with an Overall Rating of 5.0 have a larger range compared to those with lower scores.



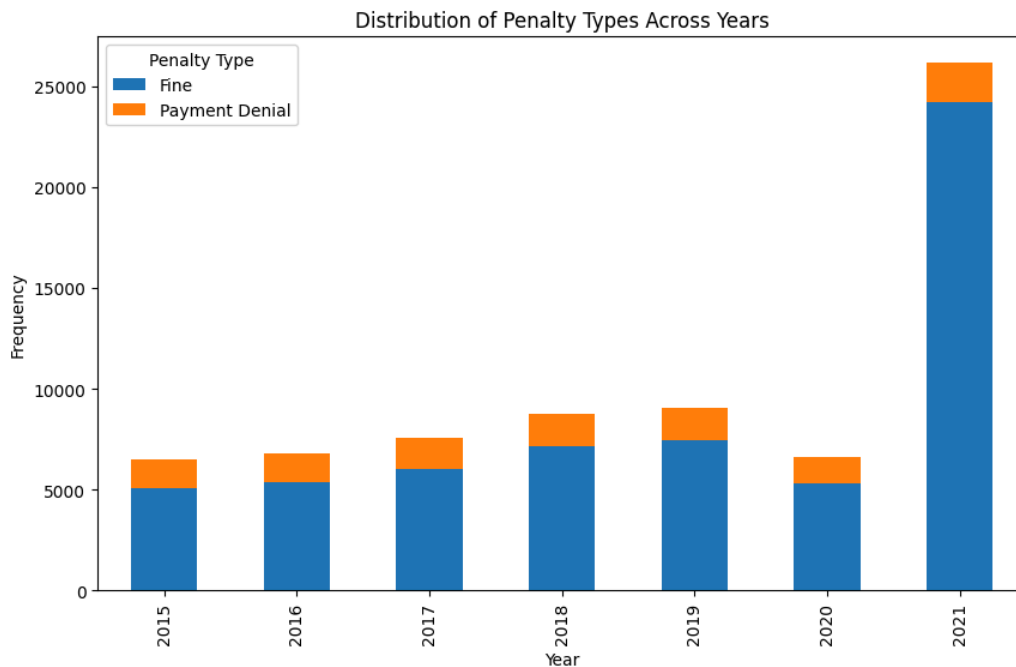
To confidently state that there is a relationship between Overall Rating and Net Income, we conducted an ANOVA. This test was implemented to evaluate if the rating scores in Overall Rating is significant and determines the Net Income.

Ratings ANOVA test	Overall
ANOVA F-Statistic	6.015
P-value	0.0011

Overall, from the results the F-statistic is significantly high which would indicate a strong significance within the groups. The P-Value given by the test is extremely small which indicates strong evidence, suggesting that the Overall Rating of a facility influences the Net Income.

Analytical Question 3: Is there a relationship between facility profitability and penalties?

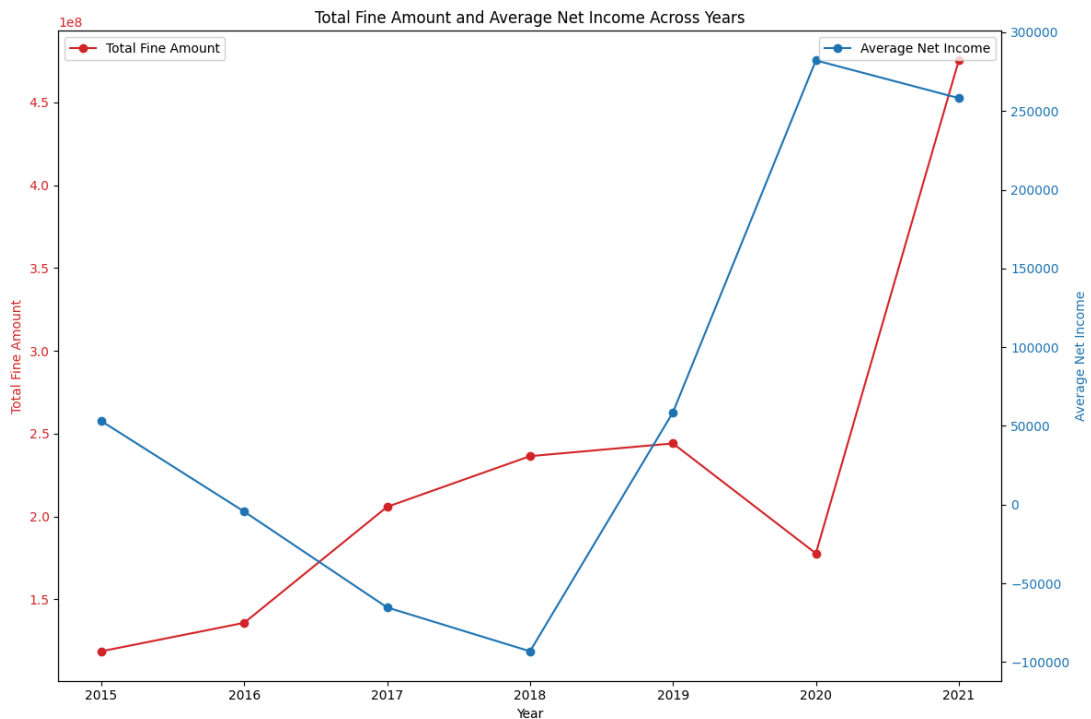
Our next focus was to determine if there is a relationship between a facility's Net Income and the penalties they receive. We visualized the distribution of penalty type across 2015–2021. The bar chart demonstrates how fine and payment denial are distributed across every year. Fines have a significantly higher frequency in comparison to payment denials over time. Specifically in 2021, the amount of fines increased significantly. This is likely related to the COVID-19 pandemic.



An ANOVA test was conducted to confidently state and identify that there is an overall correlation between penalty type and net income. From the results we can identify a small p-value score and a high f-statistic score, which would indicate that there is statistical significance between the penalty types and net income, thus supporting our hypothesis.

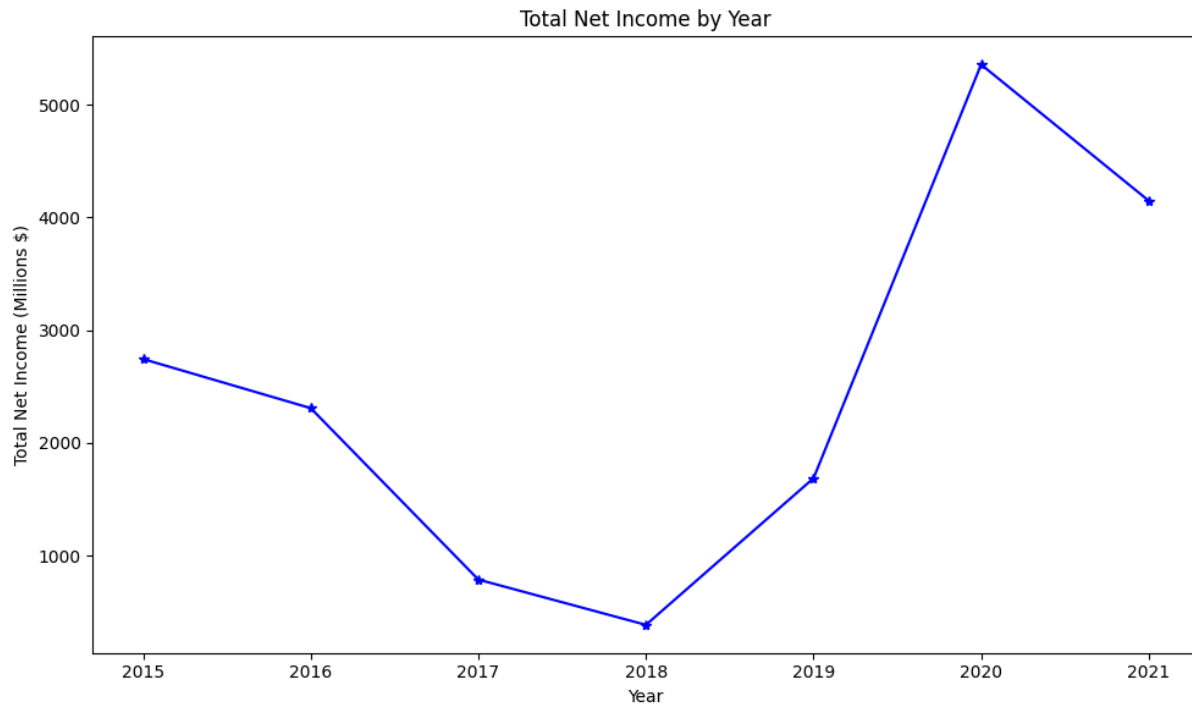
Penalty ANOVA test	Overall
ANOVA f-statistic	59.993
P-Value	9.535 ⁻¹⁵

Below, we have a graph that illustrates the trends of the total fine amount and average net income across the years 2015–2021. From this we can identify how as the fine amounts increase for every year, the net income starts to decrease over time. This graph also shows a potential correlation between the variables, it is specifically noticeable in the years 2017–2018 as the fines start to decrease and the net income increases significantly in 2019–2020. This would support our hypothesis that there is a relationship between a facility's net income and their penalties.



Analytical Question 4: Was there an impact on overall profitability for skilled nursing facilities during COVID?

**Resurfacing the net income trend for nursing facilities below.*

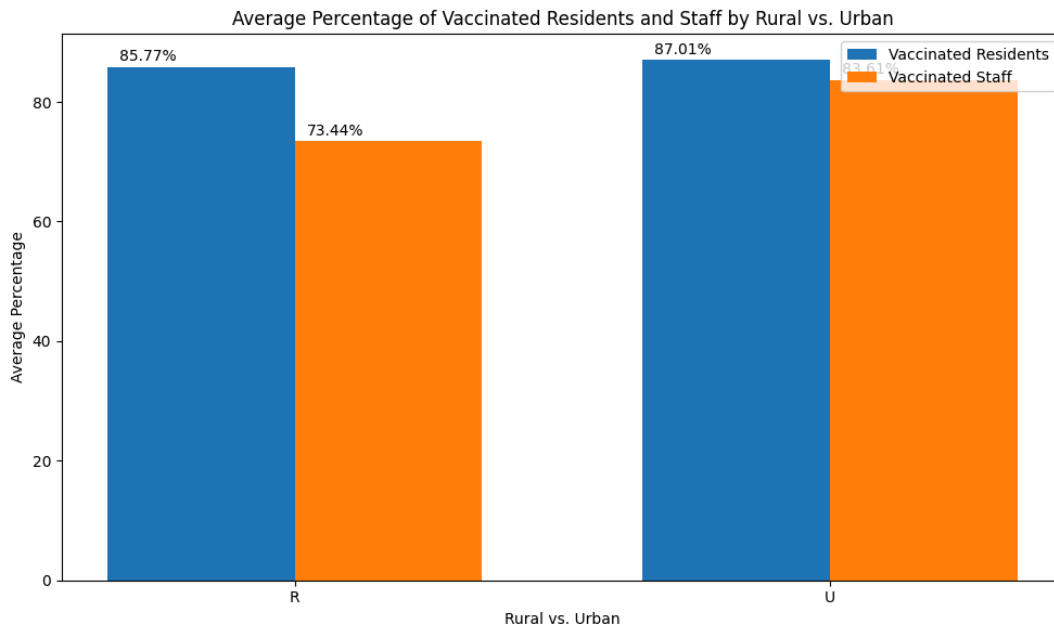


The first confirmed COVID case in the United States was in January 2020. As COVID cases in the U.S. increased, the total net income for skilled nursing facilities also increased from 2019 to 2020. There was also an overall decrease in net income from 2020 to 2021.

Our initial hypothesis for the slight decrease in 2021 is that it can be attributed to certain facility types and their residents/staff willingness to receive the COVID vaccine. We will take a deeper look into the different attribute types analyzed earlier in this report as they relate to our COVID dataset. Also, considering COVID was a new virus at the time, there were likely no proper regulations in place to guide nursing facility staff on proper protocol for handling COVID patients. This could explain the increase in fines/fine amounts and slight decrease in net income starting in 2021. As more rules and regulations were introduced after 2020, more fines/penalties could have been distributed as a result.

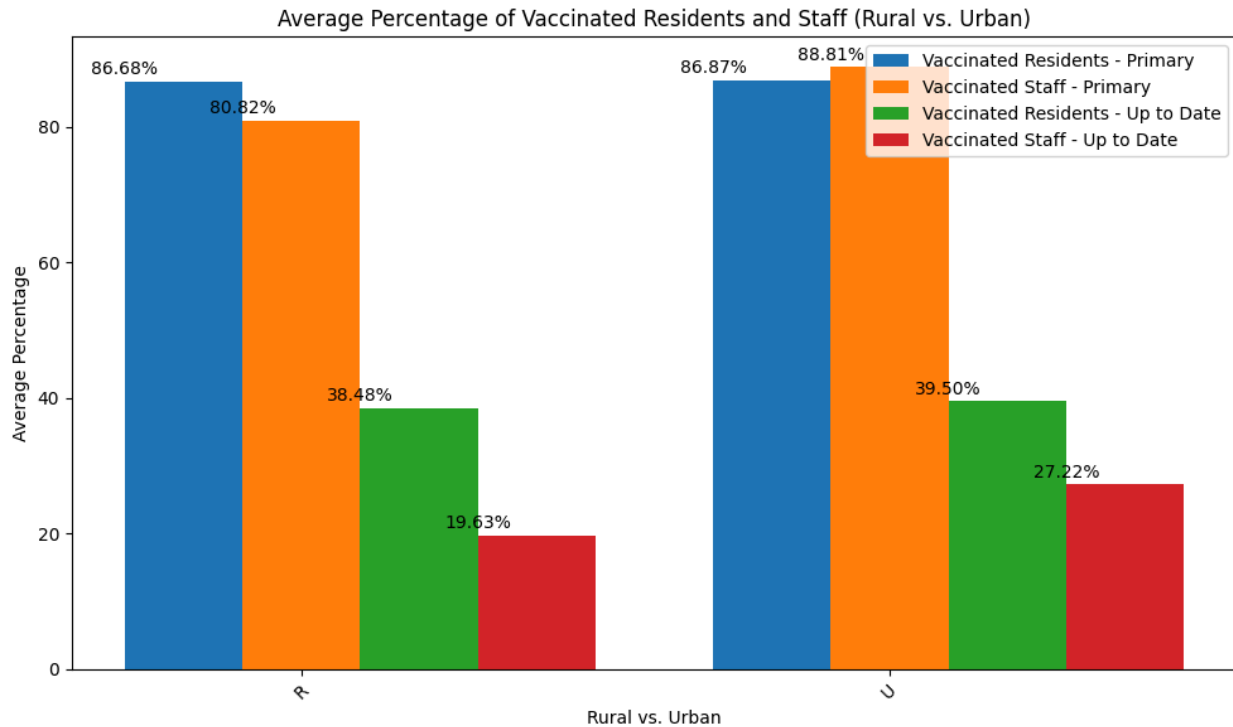
COVID Stats – Rural vs Urban (2021):

When we reference the previous analysis for [Rural vs Urban](#), we can recall that from 2020 to 2021, net income decreased for both Urban and Rural facility types. The graph below shows that the average percentage of vaccinated residents and staff for 'Urban' facilities was greater than the averages for 'Rural' facilities.



Rural vs Urban (Vaccination Rate) t-test	Residents	Staff
T-statistic	-1.199	29.821
P-value	0.02305	1.8282 ⁻¹⁸⁹

Although there is a visual difference between vaccination rates between 'Urban' and 'Rural' facility types, the above table shows that the t-test results do not indicate a significant difference between facility type vaccination rates for residents. Alternatively, there is a significant difference in vaccination rate for staff.

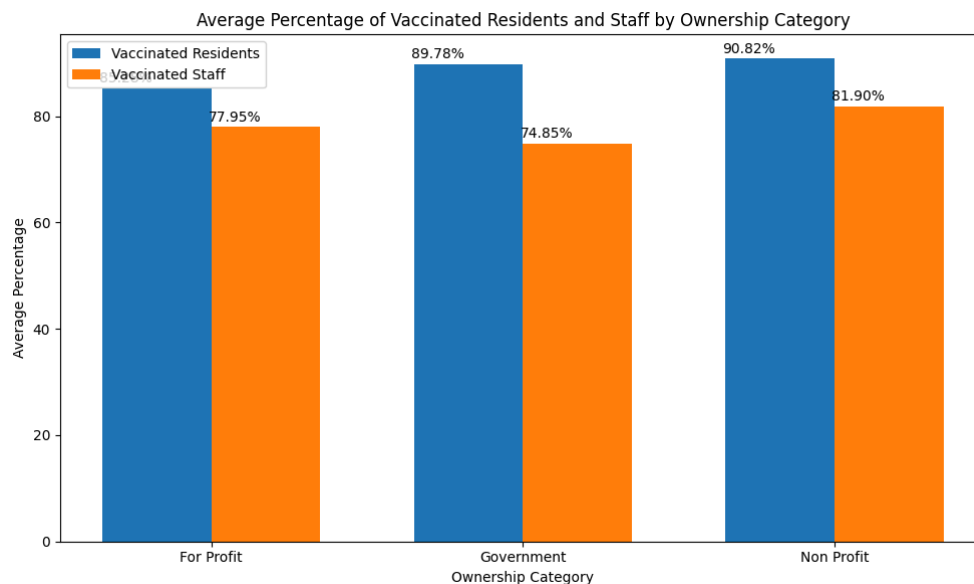


Rural vs Urban (Vaccination Rate) t-test	Residents Primary	Staff Primary	Residents Up To Date	Staff Up to Date
T-statistic	0.8396	31.416	1.5880	13.705
P-value	0.401	2.694 ⁻²⁰⁹	0.112	1.807 ⁻⁴²

When we look at the visual representation of 2022 vaccination averages in the chart above, it seems that there is somewhat of a difference in averages between vaccinated staff in urban vs rural areas. We conducted t-tests for each measure to compare significance between urban and rural facilities. Similarly to the initial t-test, the difference between resident vaccination status in urban and rural facilities was not statistically significant. However the differences for staff averages were statistically significant.

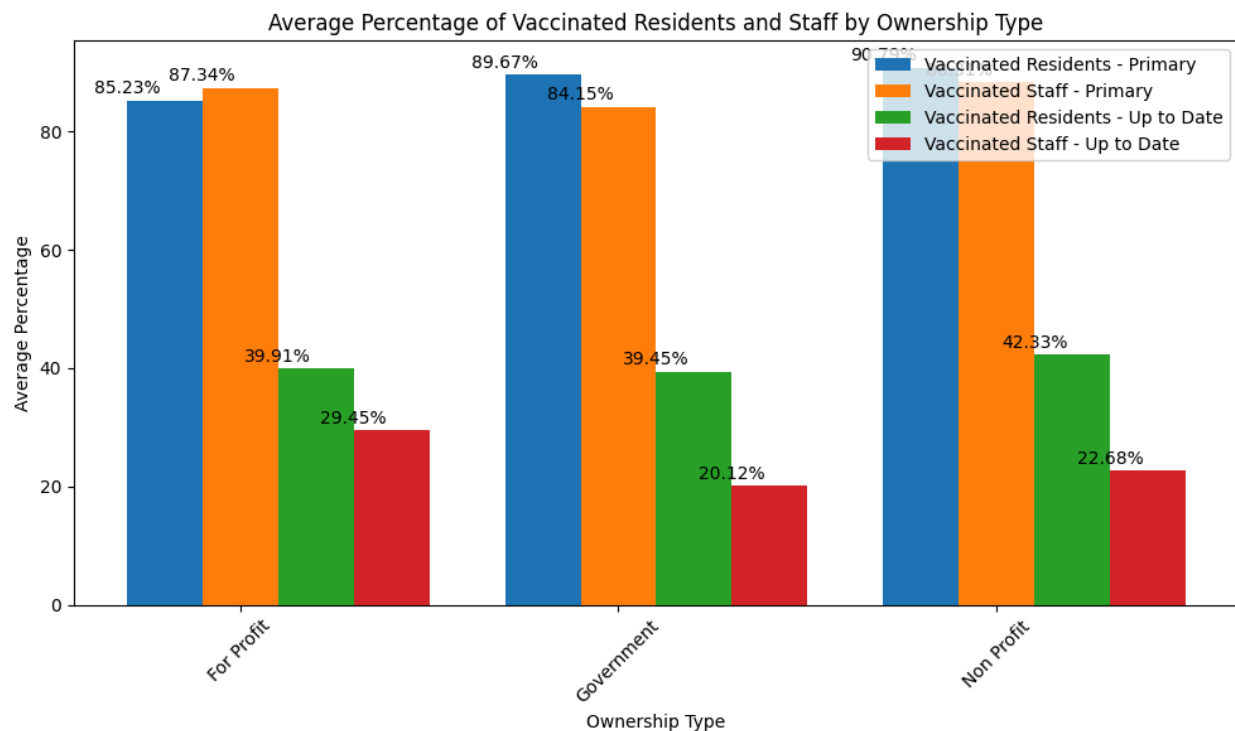
COVID Stats – Ownership Types (2021):

When we reference the previous analysis for [Ownership Type](#), we can recall that from 2020 to 2021, 'For Profit' facilities had the highest net income. However, the graph below shows that the average percentage of vaccinated residents was the lowest for 'For Profit' facilities and was the highest for 'Non-Profit' facilities. It is important to note that 'Non-Profit' facilities only included about 20% of the cost report dataset. Considering that most of the dataset is attributable to 'For Profit' facilities, we can infer that there is a higher volume of residents and staff that were vaccinated as soon as the vaccine was available. We assume that 'For Profit' facilities have increased accessibility to vaccines and can therefore administer vaccinations to their residents at a higher rate than facilities not identified as 'For Profit'.



Ownership (Vaccination Rate) ANOVA test	Residents	Staff
ANOVA F-Statistic	196.414	77.851
P-value	6.781 ⁻⁸⁴	3.398 ⁻³⁴

Similarly to the initial ownership type analysis, we can see by the graph and table above that there are some ownership types that have a higher average percentage of vaccinated residents and/or staff. We assessed these differences with an ANOVA test, which indicate significant differences between ownership types and their percentage of vaccinated residents and staff.

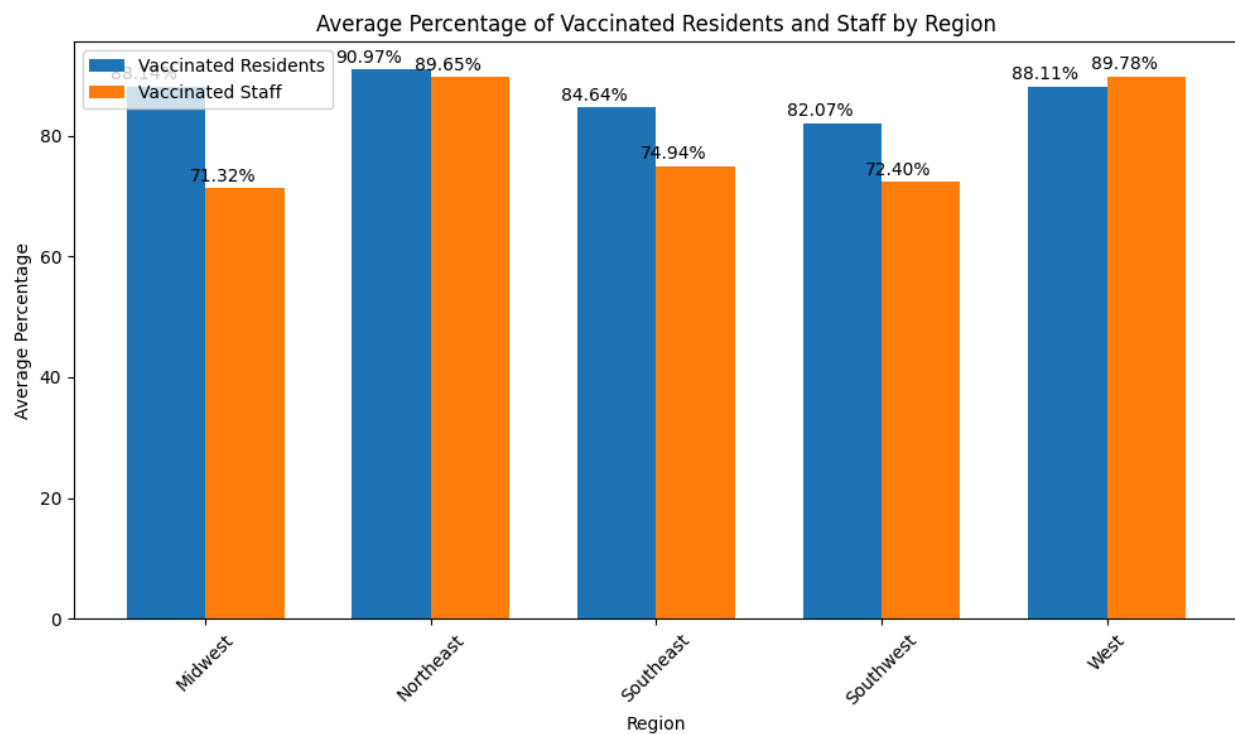


Ownership (Vaccination Rate) ANOVA test	Residents Primary	Staff Primary	Residents Up To Date	Staff Up to Date
ANOVA F-Statistic	212.34	34.163	5.112	63.382
P-value	1.969 ⁻⁹⁰	1.700 ⁻¹⁵	0.006	5.049 ⁻²⁸

In the above graph, we broke out 2022 vaccination averages for each category and tested their statistical significance between ownership types. All vaccinated types had a statistically significant difference between their ownership types. Residents up to date hit stat sig, but had the largest p-value in comparison to other vaccinated types.

COVID Stats by Region (2021):

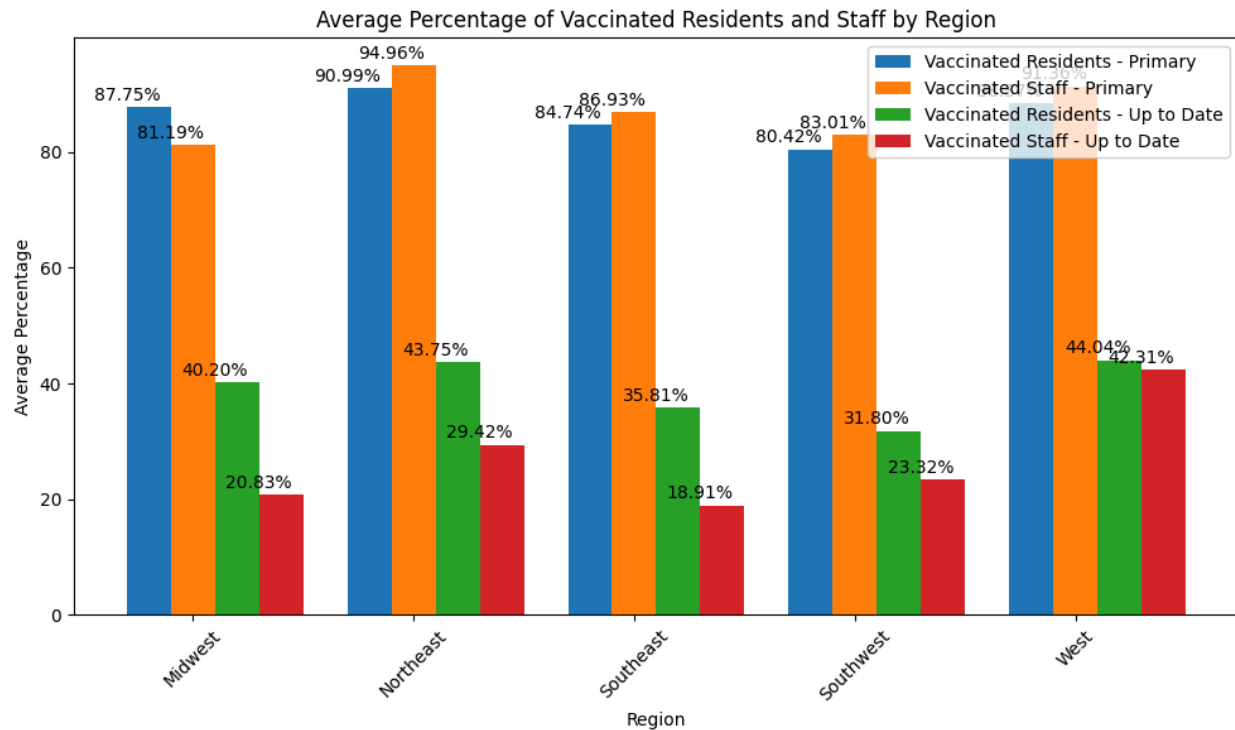
When we reference the previous analysis for [Geographic Location](#), we can recall that there was a significant difference in net income between the state regions. We can also recall that when looking at the net income trends for each region from 2015 – 2021, Southwestern states typically had an overall lower net income than other regions. Southeastern and Western states experienced spikes in net income during COVID years (2020–2021).



Region (Vaccination Rate) ANOVA test	Residents	Staff
ANOVA F-Statistic	259.554	844.193
P-value	8.831 ⁻¹⁶²	0.00

In the bar graph above, we can see that when looking at average percentages of vaccinations, Southeastern and Southwestern regions have lower average rates than Northeastern and Western states. This is interesting because although Northeastern states have a higher percentage of vaccinated residents and staff, the previous geographic analysis shows that Northeastern states had a lower net income

throughout the years leading up to the pandemic. There may be some correlation between increased vaccinations in these regions and increased net income for 2021. We've also included ANOVA test results to confirm that the differences in vaccination rates are significant between regions.



Region (Vaccination Rate) ANOVA test	Residents Primary	Staff Primary	Residents Up To Date	Staff Up to Date
ANOVA F-Statistic	266.651	609.186	52.859	271.094
P-value	4.589 ⁻²²¹	0.0	2.882 ⁻⁴⁴	1.230 ⁻²²⁴

The chart above represents 2022 COVID Vaccination data by region. Here we can see that by 2022, in terms of primary vaccinations, Northeast and Western regions are in leading positions with the greatest average percentage of vaccinated staff and residents. Contrasting with 2021 data, the average percentage of vaccinated staff exceeds the average percentage of vaccinated residents, except for in the Midwest region. However, as the vaccination is updated, residents are exceeding the average percentage of vaccinated people when compared to staff. There could be a prioritization here, where elderly patients are distributed updated vaccinations before staff. The different average percentages of vaccinated between regions are statistically significant for each group, as shown in the table above with ANOVA test results.

Models Tested

We tested multiple predictive models to see which would best determine profitability. Within each model type that we tried, we also tried altering our selection of independent variables to analyze the effects they had on the model evaluation metrics. Based on the exploratory analysis we've done throughout this report, we were able to identify key influencing factors that affected overall nursing facility financial performance.

Linear Regression Models:

Linear Regression 1:

First we tested a linear regression model using the following independent variables:

- Ownership Type, Rural vs. Urban, Overall Rating, Fine Amount, Penalty Count, and Year

With all of these variables included, the model returned the following results:

Model Evaluation Method	Result
Mean Squared Error	990,559,039,156
R-Squared	0.048495

Both the MSE and the R-Squared values tell us that the model in this setup did not result in accurate predictions. This model resulted in underfitting. We hypothesized that we would need to test different variables that might provide a better indication for the dependent variable, net income.

Linear Regression 2:

For the next linear regression, we decided to change the variables to see how it affected the model evaluation metrics. For this iteration, we used the following independent variables:

- State, Year, Penalty Type, Fine Amount, and Rural vs. Urban

With these variables included, the model returned the following results:

Model Evaluation Method	Result
Mean Squared Error	917,056,694,236
R-Squared	0.128079

Through the model evaluation metrics, we saw a slight improvement in both scores. However, both values were not ideal to confidently say the model could accurately predict

net income. The cons of using a linear regression model is the assumption that the relationship between the independent and dependent variables is linear. We initially began our process with these models since they are widely used and easily understood, however it might not be the best choice to assess our current dataset. In order to further assess the data, we decided to try a different type of predictive model, a decision tree.

Decision Tree:

Our next approach was to use a decision tree model. The decision tree can capture nonlinear relationships between variables. For our decision tree, we included the following variables:

- State, Year, Penalty Type, Fine Amount and Rural vs. Urban

Model Evaluation Method	Result
Mean Squared Error	919,214,537,041
R-Squared	0.126028

Unexpectedly, the MSE and R-squared results were slightly worse than the last linear regression model's results. They were somewhat close to each other, but the linear regression had slightly better results.

Random Forest:

Our final approach was to use a random forest model. The random forest model can have a higher accuracy through ensemble methods/collecting learnings from many different trees. For our random forest, we included the following variables:

- State, Year, Penalty Type, Fine Amount and Rural vs. Urban

Model Evaluation Method	Result
Mean Squared Error	774,285,438,634
R-Squared	0.263824

In comparison to other models that we tried and their evaluation metrics, the random forest model had the most positive results. We were able to minimize the MSE and get the R-Square closer to 1. The MSE and R-Square are still not at ideal levels, however we can take this as an indication that the random forest would be a better predictor for this project.

Discussion

Throughout the report, we addressed and analyzed many portions of the Skilled Nursing Facilities datasets and the variables we deemed influential on net income. In the beginning we analyzed the overall financial performance of nursing facilities from 2015 to 2021 at a high level. Initially, we made the observation that nursing facilities had shown to be profitable throughout the measurement period. Although there was a minimal decrease in overall total net income from 2015 to 2018, the trend took a swift turn upward especially as we entered the beginning of the COVID-19 pandemic. This caused us to hypothesize that nursing facilities were beginning to see an increase in profitability because baby boomers began to explore nursing home options AND because of the pandemic.

In order to identify the influential factors that could be affecting financial performance of nursing homes, we decided to curate our analytical questions that would help us investigate other underlying factors that were not obvious from the high level view. We analyzed various facility attributes such as Rural versus Urban types of facilities, Ownership types that consisted of for-profit, nonprofit and government owned facilities, and the geographical location of all of the facilities included in the data.

For rural versus urban facilities, we initially thought that rural facilities would not have a higher net income than urban facilities, therefore indicating that they would be less profitable. Through our analysis of this attribute, we found that over the years Urban facilities were typically more profitable than rural facilities (based on the data provided). However, it is important to note that over 70% of the cost report data was attributed to Urban facilities. As the first attribute, this could skew analysis, so that prompted us to investigate further.

In addition to Rural versus Urban, we analyzed the type of ownership of the facilities in the dataset. There were a total of 13 differently identified ownership types for skilled nursing facilities, all with various percent shares of the dataset. We decided to group the ownership types by for-profit, nonprofit and government owned in order to reduce variance. This allowed us to reveal better insights on skilled nursing facility financial performance. We revealed that government owned facilities' net income trend typically remained flat (even through the beginning of the pandemic) in comparison to the other two ownership types. Nonprofit organizations were the main contributors behind the decrease in net income that we identified in the overall view. Lastly, for profit facilities saw the greatest increase in net income from 2019 - 2021 in comparison to other facility types. Ultimately, these findings are beginning to guide us in the right direction on which facilities could be worth investing in.

With the geographical location attribute, we grouped the states by region to get a better understanding on if there were particular portions of the U.S. that had nursing facilities with better financial performance than others. Regions such as the Southeast, Midwest and West had significant improvements in net income in 2020. There were also some regions that had a significant decline in net income in 2021. We hypothesized that this might be due to new protocols and regulations that were implemented to accommodate the pandemic. Some regions may have been more willing/strict to follow guidelines where others may not have been.

We also analyzed overall facility ratings as they related to net income. Through this analysis we were able to identify that facilities with higher ratings typically had higher net income throughout the measurement period. Alternatively, facilities with 2 stars or less typically had a negative net income throughout the years. This indicates that customer satisfaction is highly influential in determining whether a facility will be profitable.

Our last facility attribute involved penalty amounts and fines associated with penalties. We found that the amount of penalties significantly increased in 2021, which would explain the overall decrease in profitability/net income in 2021. As mentioned above, this could be due to newly developed COVID protocols where facilities were learning to adhere to guidelines.

For our last analytical question, we closely analyzed the available COVID vaccine reports to assess the significance of average percentages of vaccinated residents and staff. We merged the data to include the attributes that we previously analyzed. We found that there were significant differences between each attribute and their rate of vaccination. This indicates that for 2021, COVID did have an influence on overall nursing facility profitability depending on the location and type of ownership.

After analyzing facility attributes and COVID data, we took what we learned about nursing facilities and applied them to developing our models. We tested three different types of models: Linear Regression, Decision Tree and Random Forest. All three models did not have significantly strong results in terms of their model evaluation metrics, however we ultimately decided to prioritize the Random Forest model. Relatively, it had stronger results than the other two methods. We hypothesized that since there are so many different variables and attributes associated with nursing facilities, that further testing would be necessary in order to have a strong random forest model.

The data provided for this assignment was split into four different categories, and within each category there was a report for each year. In order to assess every attribute and every variable that should be considered in predictive modeling, more time and detail would be

required. One limitation with this project was that there were many inconsistencies with the data and there was also a lack of context. Although a data dictionary was provided, it primarily consisted of jargon related to nursing facilities. We believe that if there was more context around the cost report, such as an example of the form filled out for each cost report, as well as more detailed descriptions of what each variable signified then there would be an easier understanding of how to handle certain variables to be able to confidently decide whether or not they should be included in the analysis.

In the future, we'd recommend additional research around certain details of nursing facilities. Some of which could include typical demographics of residents. If the residents are the primary drivers in determining nursing facility profitability, then details about the residents should be taken into consideration. Marketing efforts for nursing facilities should also be included in further research. There could be facilities that are not in urban or metropolitan areas that are not reaching their entire target audience in their region. Considering that older generations are the primary residents of nursing facilities, it is important to consider that they might not be easily reached through digital marketing means. If a facility does not have many marketing efforts through the ideal channels, that could affect their overall profitability in the long run.

Recommendations and Conclusion

Based on our detailed analysis we can advise our client to invest in certain types of nursing facilities. The data reveals that nursing facilities in general have maintained an increase in net income, especially during the COVID-19 pandemic.

Our comprehensive analysis has identified specific attributes that suggest investments in nursing facilities to be advisable. Our analysis that focused on Urban and Rural facilities indicated a significant difference of net income between the facility types. We recommend the investment of nursing facilities in Urban locations. Urban facilities have showcased an increase in revenue throughout the years and going into the pandemic when compared to Rural facilities.

Additionally, for-profit ownership types consistently maintained a higher net income level making them a more reliable and profitable investment option, even during challenging economic conditions such as the pandemic. After analyzing locations based on region, we can recommend Western and Southeastern regions as advisable locations to prioritize investment. These regions maintained a higher net income when compared to other regions.

Investing in nursing facilities that are maintaining a high overall rating score is also advisable. Facilities that had a high overall rating score had significantly higher net incomes compared to lower-scored facilities.


With fewer regulations before the pandemic, penalties were not of much concern. However once the CDC began implementing guidelines there was a notable spike in penalties and fines. It is advisable to invest in facilities that follow all regulations to maintain financial stability. Regular training and compliance would also aid in the profitability of the investment. We also recommend extending the measurement period through 2023 to analyze whether net income levels have evened out or if penalties and fines are an ongoing issue.


From the data provided and analysis conducted, we can ultimately recommend the investment of highly rated, for-profit facilities located in Urban areas that also maintain a strong regulatory compliance.

Also as mentioned in the Discussion, we also recommend further investigation of target resident demographics and facility marketing efforts. Both variables could provide additional context that could potentially help further inform a predictive model.

Appendices

Final Google Colab Notebook:

 BANA620_Analysis (CLEANED).ipynb

 BANA620_PREPROCESSING.ipynb

Included in the notebooks:

Code with Preprocessing, Data Cleaning and Visualizations used throughout the report

(.ipynb and PDF versions are also included in the final submission for this project)

References

U.S Centers for Medicare & Medicaid Services. "Skilled Nursing Facility Cost Reports." Nursing Homes Including Rehab Services, U.S Centers for Medicare & Medicaid Services, 2015–2021, data.cms.gov/provider-data/topics/nursing-homes.

U.S Centers for Medicare & Medicaid Services. "Skilled Nursing Facility Penalties Reports." Nursing Homes Including Rehab Services, U.S Centers for Medicare & Medicaid Services, 2015–2021, data.cms.gov/provider-data/topics/nursing-homes.

U.S Centers for Medicare & Medicaid Services. "Skilled Nursing Facility Provider Info Reports." Nursing Homes Including Rehab Services, U.S Centers for Medicare & Medicaid Services, 2015–2021, data.cms.gov/provider-data/topics/nursing-homes.