Net Income Predictors in Nursing Homes Across America (2015-2021)

Emely Callejas and Robert Pimentel

Executive Summary

This report presents a comprehensive analysis of the financial performance of skilled nursing homes in the United States, focusing on the recent period and the significant impact of the COVID-19 pandemic. Our evaluation aimed to assess the financial health of these facilities, identify key factors influencing their performance, and understand the trends and impacts shaped by these elements. When performing exploratory data analysis, we found the data gives us insights on the skilled nursing homes in regards to their financial and operational trends over several years, focusing on the effects of the COVID-19 pandemic and regulatory changes.

Net income for skilled nursing homes has experienced significant fluctuations, with notable dips in 2017 and 2018 followed by a period of recovery. This financial variability may be attributed to changes in healthcare policies that impact reimbursements and operational costs, with recent improvements possibly resulting from enhanced management strategies or favorable policy adjustments. An analysis at the state level reveals considerable differences in financial performance, indicating that local policies, economic conditions, and demographic factors are key influencers. Additionally, there has been an increase in fines and penalties, with a sharp rise noted in 2021, which could suggest a stricter regulatory environment or increased non-compliance, potentially exacerbated by the pandemic. Furthermore, vaccination rates among residents and staff have been increasing, highlighting efforts to comply with public health mandates and mitigate risks, crucial for maintaining a good reputation and managing liabilities. The quarterly performance scores, though fluctuating, generally show an upward trend, suggesting continuous improvements in care quality and management practices in response to regulatory demands and internal evaluations. These findings underline the sector's resilience and adaptability in a complex mix of regulatory, economic, and public health challenges, emphasizing a sustained commitment to enhancing care standards and ensuring compliance for long-term success.

Introduction

The healthcare industry is facing an increasingly complex array of challenges that impact the operational and financial stability of skilled nursing homes across the United States. This project is centered on understanding the dynamics that shape the financial performance of these facilities, with a particular emphasis on recent years marked by significant regulatory changes and the unprecedented global impact of the COVID-19 pandemic. Skilled nursing homes are vital in the healthcare system, caring for the elderly and those needing long-term care. Despite their importance, these facilities face financial instability due to unpredictable net incomes, different levels of success across states, and constantly evolving regulations. This project focuses on pinpointing the factors that affect the net incomes of these nursing homes. Understanding these influences is crucial as it can help healthcare providers, policymakers, and investors grasp the financial health of these facilities better. With this knowledge, they can make informed decisions, tweak policies, and put in place strong support systems that boost the resilience and care quality of these homes. Ultimately, this analysis seeks to improve the healthcare outcomes for the vulnerable groups relying on skilled nursing facilities.

Methodology

The project used a mix of data collection methods, analytical techniques, and tools to explore the financial performance of skilled nursing homes. We collected data from several reliable sources that track healthcare performance, including government databases and financial records from the healthcare industry. This collection process was aimed at obtaining detailed financial information, regulatory compliance data, and operational statistics for skilled nursing homes spanning several years.

Our analysis utilized various statistical methods and machine learning algorithms such as bar charts and box plots to visualize financial trends and distributions, helping us to see changes over time and compare performance across states. Linear regression models helped us understand the relationship between various factors, like regulatory changes and economic conditions, and the net income of nursing homes. These models were crucial for identifying direct influences on financial outcomes. Decision tree models allowed us to examine complex interactions between multiple factors, offering a clearer view of how different variables impact net incomes. Python was the primary programming language for this analysis, supported by libraries such as Pandas and NumPy for data manipulation, and Matplotlib and Seaborn for creating visualizations. We also used Scikit-learn for implementing and evaluating the machine learning models, including linear regression, decision trees and random forest.

Data Description

For this project, we utilized datasets from the Centers for Medicare & Medicaid Services (CMS), which provide detailed information about skilled nursing homes. These datasets are publicly accessible via the CMS's provider data portal and cover various facets of nursing home operations, including financial performance, regulatory compliance, and patient outcomes. The datasets encompass records from the years 2015 to 2021, providing a rich historical context that allows for longitudinal studies of trends and patterns in the sector. This comprehensive data includes metrics such as net income, operational costs, patient demographics, staff details, and compliance with health regulations. The total dataset consists of thousands of records from nursing homes across the United States, making it a substantial resource for analysis.

The preparation of the data involved several critical steps. We removed duplicates to ensure the integrity of our analysis. Incomplete or missing data entries were addressed by imputation or deletion, based on their significance and the volume of missing data.We concatenated datasets from each year (2015-2021) to create a unified dataset that spans several years. This step was essential for analyzing trends over time and required careful alignment of data fields across different years. We transformed categorical data into numerical formats suitable for machine learning models and encoded variables as needed to facilitate analysis. Numerical data were normalized to standardize scales and reduce bias in the machine learning algorithms. We merged data from various tables into a single dataset by matching nursing home identifiers and ensuring consistency across the datasets.

These preprocessing steps ensured that the data was clean, consistent, and well-structured, setting a solid foundation for the subsequent analysis. This meticulous preparation was crucial for effectively identifying and understanding the factors influencing the financial health and operational efficacy of skilled nursing homes.

Analysis and Findings

When we performed an exploratory data analysis we went through the penalties, performance, and cost datasets to get a better idea of the financial health of the nursing homes.Net income for skilled nursing homes has seen significant ups and downs, with the lowest points in 2017 and 2018, followed by a recovery in the subsequent years. This could be tied to changes in healthcare policies affecting reimbursements or variations in operational costs. The recovery noted in the later years might reflect improvements in management strategies or favorable changes in funding.

When we look at the financial performance on a state level, there's a clear difference between states. This suggests that local policies, economic conditions, and demographic factors

play crucial roles in influencing outcomes. Our analysis also shows an increase in fines and penalties over the years, with a sharp rise in 2021. This may indicate a stricter regulatory environment or a rise in non-compliance among nursing homes, which could be related to the increased pressures of the pandemic. We noticed Texas, Michigan and Illinois were the states with the highest number of fines in 2021. We also looked at vaccination rates among residents and staff, which have been increasing. This is key for understanding how nursing homes are protecting their populations and complying with public health mandates, which affects both their risk management and reputation. Lastly, the quarterly performance scores by year show fluctuations but generally trend upwards, suggesting that nursing homes are continuously improving their care quality and management practices in response to both internal assessments and external pressures.

Based on the regression analysis conducted to understand factors affecting Net Income, the gained insight performing a linear regression on the cost dataset. The data includes variables like 'Less_Total_Operating_Expense', 'Total_Salaries_adjusted', and 'Net_Patient_Revenue', which are used to predict 'Net_Income'. The model has a Mean Squared Error of 109820516917.238, reflecting the average squared difference between estimated and actual values. The coefficients indicate how changes in these predictors influence Net Income. For example, an increase in 'Less_Total_Operating_Expense' is associated with an increase in Net Income.

The analysis highlights that the model explains only 14.32% of the variability in Net Income, suggesting it captures a limited range of factors influencing Net Income. The residual analysis shows a decrease in residuals as predicted values increase, indicating potential heteroscedasticity. This model limitation suggests that additional variables and factors are necessary to better predict Net Income. For future research, it would be beneficial to include more variables in the model and explore different model types or transformations to address the heteroscedasticity and improve the model's fit. Recommendations based on this analysis include focusing on increasing 'Net Patient Revenue' and managing 'Total Salaries' more effectively, as these variables significantly influence Net Income. Additionally, optimizing operational expenses could lead to better financial outcomes. Implementing these recommendations involves a detailed analysis of the components of salaries and expenses, developing targeted strategies to enhance patient revenue, and improving overall financial management. These actions are expected to improve profitability and operational efficiency, enhancing the organization's financial health.

We also ran a decision tree analysis using the cost dataset. The analysis involves using a Decision Tree Regressor to predict Net Income, focusing on handling missing data and evaluating the model's effectiveness. The initial steps include imputing missing values using the mean of the columns and splitting the data into features and a target variable, which is Net

Income in this scenario. The dataset was initially processed to handle missing values using an imputation method that replaces missing entries with the mean of each column. Some features were extracted, and the data was split into training and testing sets, maintaining a standard practice of testing size (20% of the data). The Decision Tree Regressor is trained on the dataset, and predictions are made on the test set. Model performance is quantified by calculating the Mean Squared Error (MSE) and the R-squared value. The MSE for this model is notably high, suggesting significant errors between the predicted and actual values. The R-squared value is 0.715, indicating that approximately 71.5% of the variance in Net Income is predictable from the independent variables used in the model. This implies that while the model has captured a substantial portion of the data's variability, there remains room for improvement, possibly by incorporating more features or using more complex modeling techniques. A visualization of the decision tree provides a graphical interpretation of how different features affect predictions. The tree splits represent decisions made based on feature values that lead to estimates of Net Income. This visual representation helps in understanding the logic behind the model's predictions, making it easier to explain the results and the model's decision-making process.

In terms of business application, these insights are important for identifying the key drivers of Net Income and can guide strategic decisions to optimize financial outcomes. The use of a Decision Tree allows for easy interpretation of these drivers and their thresholds, which can be particularly useful for setting targets and making informed business decisions. For future steps, exploring additional variables or employing a method like Random Forest might provide lower error metrics and higher explanatory power, further enhancing the predictive capability and business applications of the model.

We then perform a model comparison for three models to best predict “Net Income”: a Decision Tree Regressor, a Linear Regression model, and a Random Forest Regressor. Each model was trained using the training data. The data was preprocessed to handle missing values through mean imputation and split into training and testing sets. Each model was evaluated based on its Mean Squared Error and R-squared values, which measure the average errors and the proportion of variance in the dependent variable explained by the independent variables, respectively. The models demonstrate varying levels of effectiveness. The Decision Tree and Random Forest models provide insights into the data's structure and relationships, with the Random Forest model performing better in terms of explaining data variability. The Linear Regression model offers an understanding of the relationships between predictors and the target variable but may not capture complex patterns as effectively as the tree-based models, which could cause underfitting. The Random Forest model shows the highest R-squared value in both training and testing phases. This suggests that it is the best performer among the three in terms of explaining the variability of the dependent variable, 'Net Income', from the features used. It provides a balance between bias and variance, making it robust against overfitting while also capturing complex patterns in the data more effectively than the other two models. The Decision Tree model, while having a higher R-squared in training, shows a significant drop in

performance on the testing set, indicating that it might be overfitting the training data. Linear Regression, on the other hand, shows moderate performance in both training and testing but does not capture as much variability as the Random Forest. In summary, the Random Forest model is the most effective for this particular dataset and problem context, as it not only explains a higher proportion of variance but also keeps performance consistency from training to testing. This makes it a reliable choice for predicting Net Income based on the available features.

These models are crucial for financial analysis and planning, as they help identify key factors influencing Net Income. Insights from these models could guide strategic decisions aimed at enhancing business performance. For future analysis, incorporating more data, applying advanced modeling techniques, or using ensemble methods could further improve predictive accuracy and provide deeper insights into the financial dynamics of the business.

Discussion

When we performed exploratory data, we focused mainly on net income and states jurisdiction as variables to see if there were noticeable trends. We found that net income has fluctuated over the years. There was visible growth in net income after 2019 in some charts, suggesting a rebound from potential downturns, which could correlate with external economic factors or internal improvements. When looking at the dataset with state jurisdiction in mind, the average net income and fines imposed vary significantly by state, indicating that regional factors such as local regulations, market saturation, or operational efficiency could have substantial impacts on financial outcomes. The total fines by penalty type show a substantial increase in fines in 2021, suggesting either increased regulatory scrutiny or a higher incidence of regulatory breaches. In regards to vaccination rates, the heatmap of vaccination metrics by state demonstrates a varied response, which could impact public health measures, employee attendance, and operational capacity. The “State Comparison of Average Performance per Quarter” visualization shows that performance scores are relatively consistent across quarters for most states, except for the state of Massachusetts, suggesting stable operational conditions or management practices throughout the year.

The outcome of the linear regression model aimed at predicting Net_Income using the selected variables (Total_Salaries_adjusted, Net_Patient_Revenue, and 'Less_Total_Operating_Expense) was evaluated using the mean squared error (MSE) metric. The MSE was calculated as approximately 110244403671.04608, which provides a measure of the average squared difference between the predicted net income values and the actual net income values in the test data. This MSE value, being a large number, indicates that the model's predictions deviate considerably from the actual values. The high MSE suggests either the model's inability to capture all the nuances and factors affecting Net_Income or the presence of outliers or high variability in the data that are not adequately accounted for by the model. It might also imply that the selected features, despite their strong linear correlation with

Net_Income, do not encompass all the variables influencing the outcome, or that the linear model structure is too simple to capture the complex relationships in the data. The R-squared value of approximately 0.1399 derived from the linear regression model indicates that only about 14% of the variance in Net_Income is explained by the predictors. This low R-squared value suggests that the model has limited effectiveness in capturing the factors that influence Net_Income. Essentially, the model leaves a significant portion (around 86%) of the variance in Net_Income unexplained, pointing to its limited predictive power. When we ran a model comparison, the decision tree model demonstrated significantly better performance than the linear regression model, as indicated by both the Mean Squared Error (MSE) and the R-squared value. When running a classification model, we used states as the predictor in the decision tree, which achieved a lower MSE of approximately 87,768,271,287.66 and a higher R-squared value of 0.7158, compared to the linear regression model's MSE of approximately 110,244,403,671.05 and R-squared of 0.1399. Decision Trees can model non-linear relationships more effectively than linear regression, which might explain the improved performance. Decision Trees are particularly adept at handling complex datasets where interactions between variables can influence the outcome in a non-linear way. This allows them to capture more nuances in the data that linear models might miss.

For future research, given the substantial difference in performance, further validation and possibly tuning of both models would be important to ensure reliability of the predictions, especially in a production environment where decisions based on these models can have significant consequences because while Decision Trees generally perform well on complex datasets, they are also prone to overfitting, especially with very deep trees. We would suggest validating the model's performance using techniques like cross-validation or by adjusting the complexity parameters of the tree.

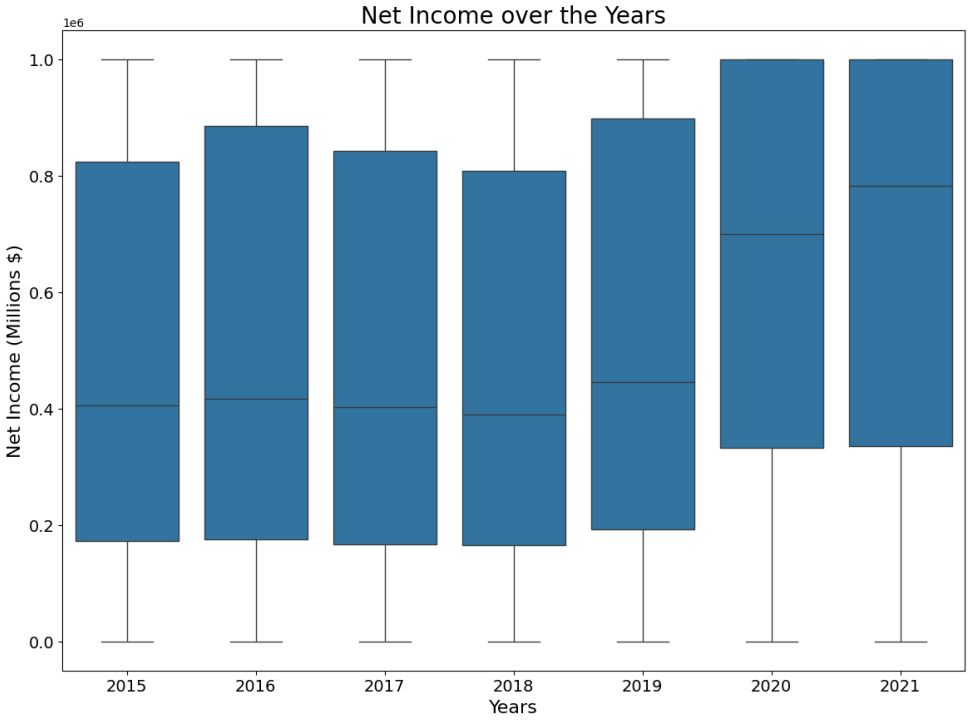
Recommendations

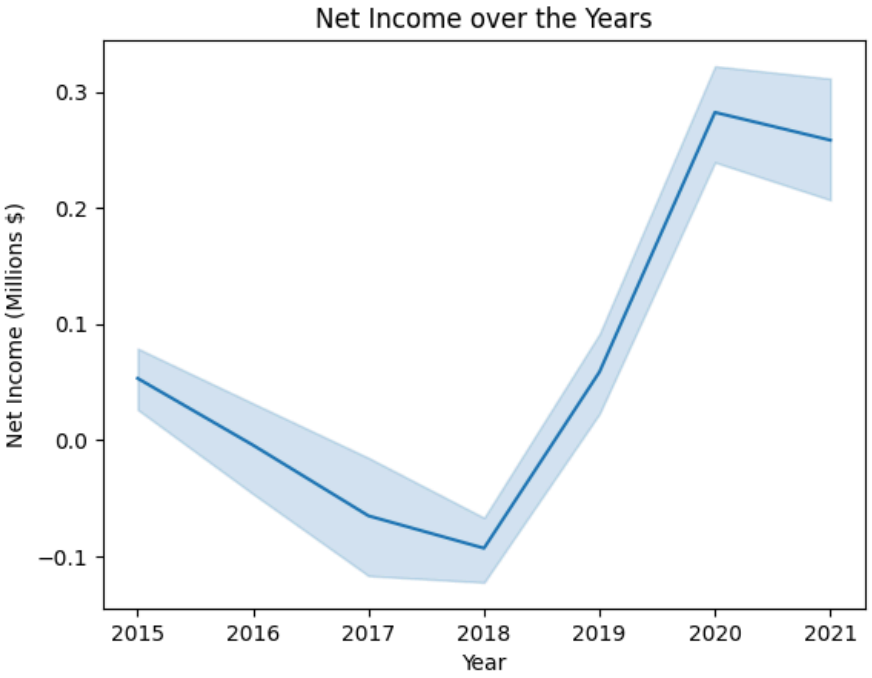
In regards to recommendations and given the better performance of the decision tree model in predicting Net_Income, integrating this model into the predictive analytics framework can significantly enhance the accuracy of financial forecasts. This improvement will aid in more effective budgeting, resource allocation, and investment decisions. Also to enhance data collection and engineering, increasing the quality of data collection and developing more informative features will likely boost model performance. This approach involves auditing existing data collection processes to identify and rectify gaps or inaccuracies and introducing new data points that could impact Net_Income. Techniques like principal component analysis can be used to reduce dimensionality and extract influential features, which will provide deeper insights and stronger predictive power. For now, we would recommend cutting down on operating costs to improve net income and focus more on net revenue from patients.

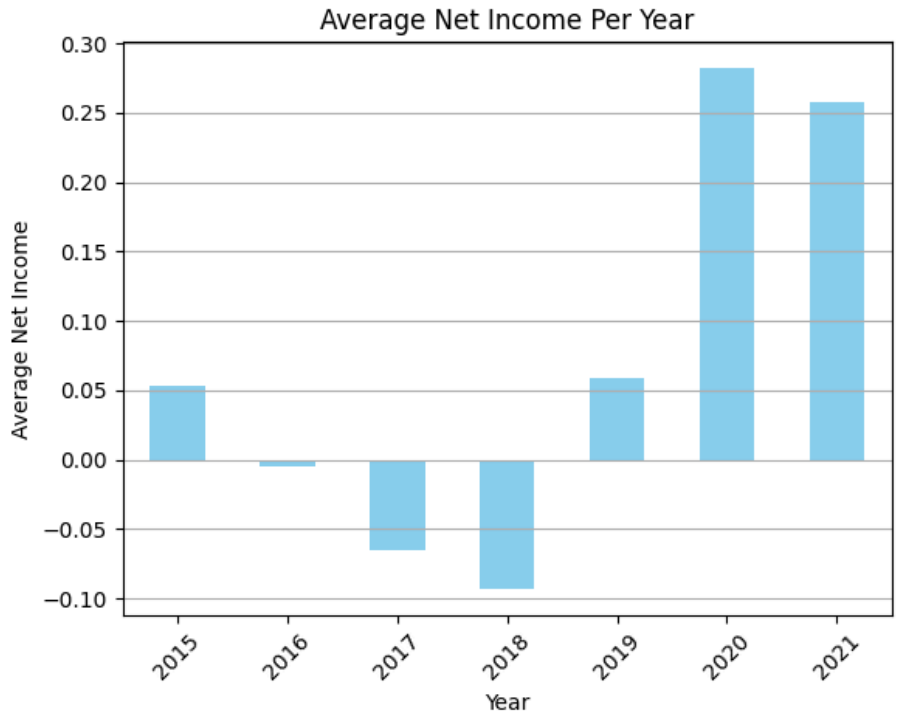
Conclusion

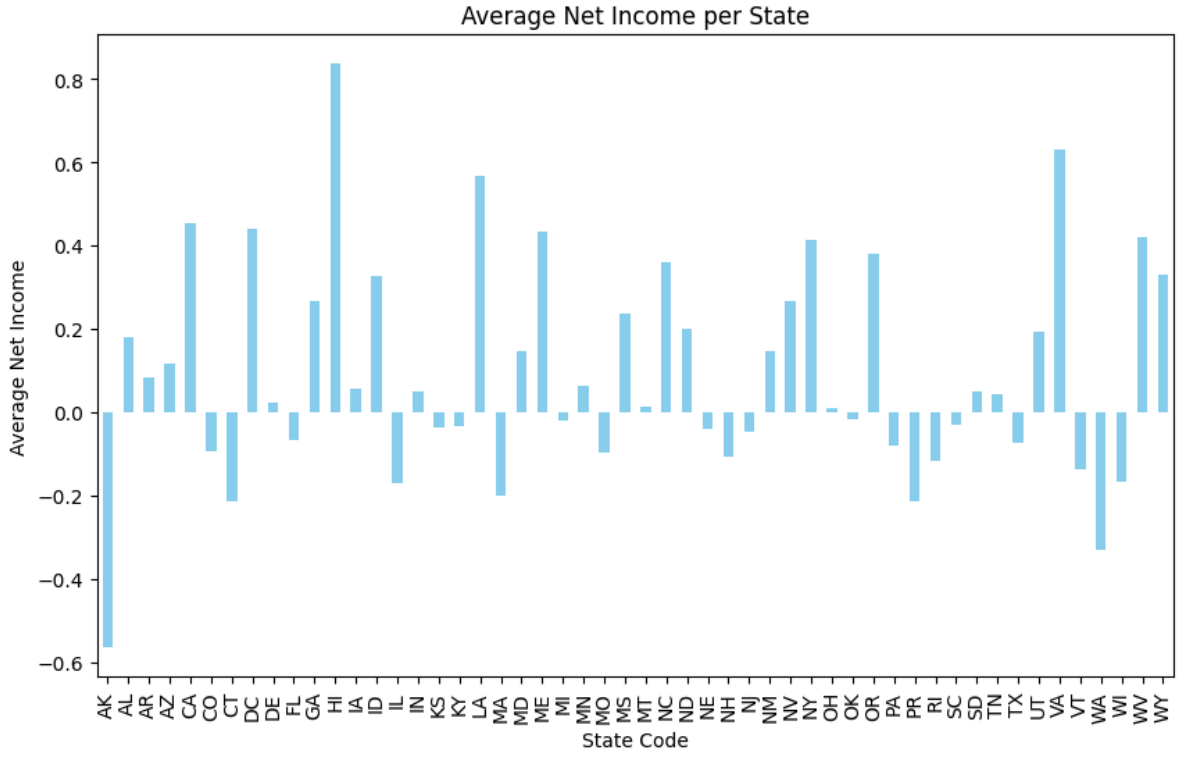
In conclusion, this report has provided a detailed analysis of the financial performance of skilled nursing homes in the United States, particularly during the challenging times brought by the COVID-19 pandemic. Our study revealed the financial fluctuations influenced by regulatory changes and varying performance across states, highlighting the importance of local economic conditions and policies. The report also noted an increase in fines and penalties alongside rising vaccination rates, indicating compliance with stricter regulations and public health mandates. Improved quarterly performance scores suggest ongoing enhancements in care quality and management practices. Utilizing statistical models and machine learning, the analysis offered insights for strategic decision-making. Recommendations include adopting predictive analytics for better financial forecasting, optimizing operational costs, and focusing on increasing net patient revenue to improve financial stability and care quality. These findings and recommendations are important for healthcare providers, policymakers, and investors, providing a clearer understanding of the financial health of these facilities and guiding improvements in their operations and services.

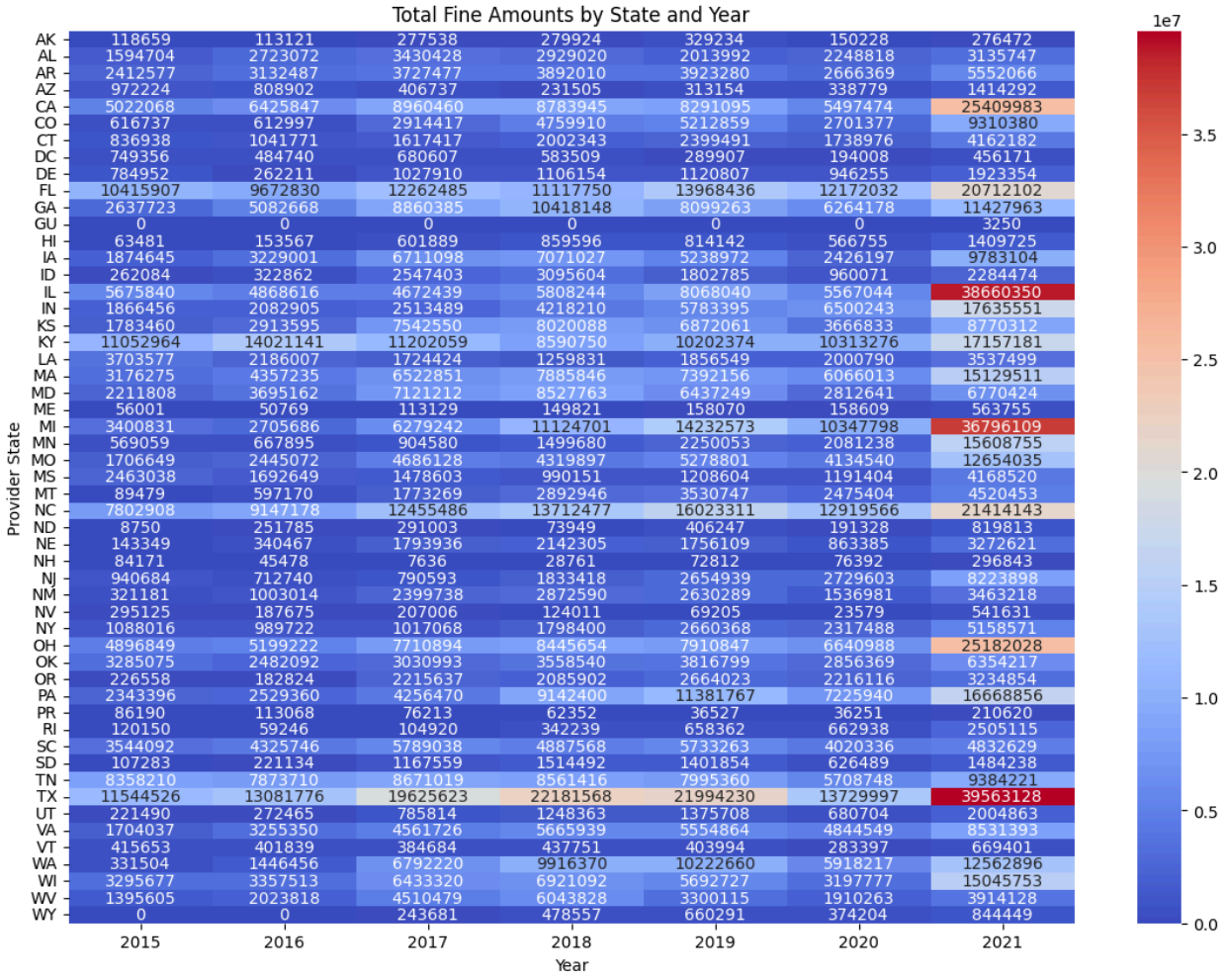
Appendices

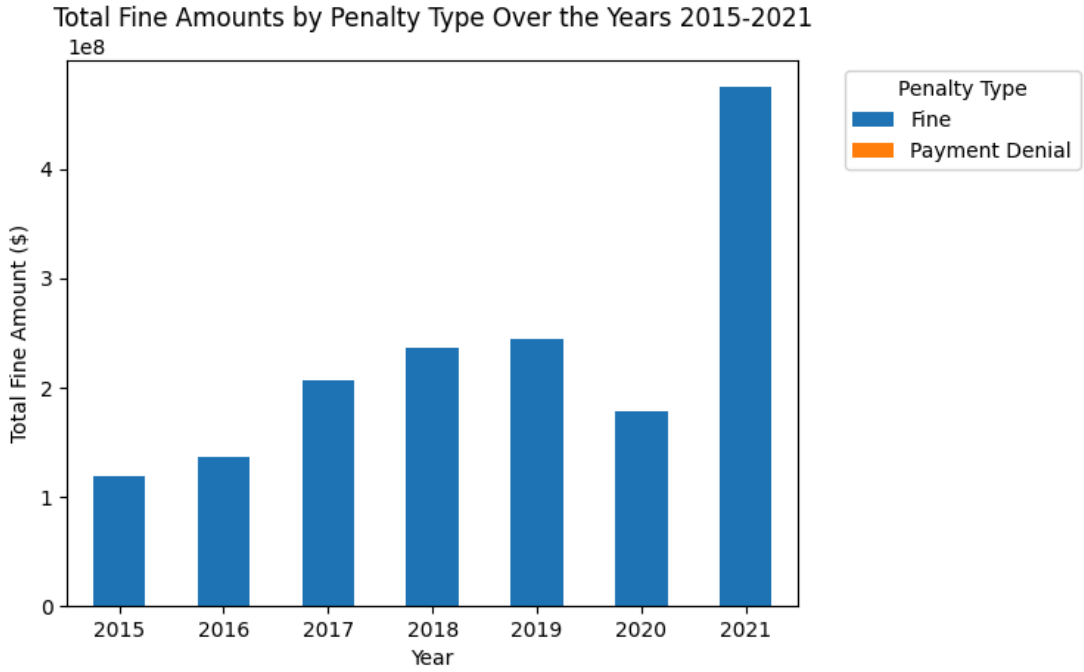

Net Income over the Years
1e6
1.0
0.8
0.6
0.4
0.2
0.0
Net Income (Millions $)
2015
2016
2017
2018
2019
2020
2021
Years

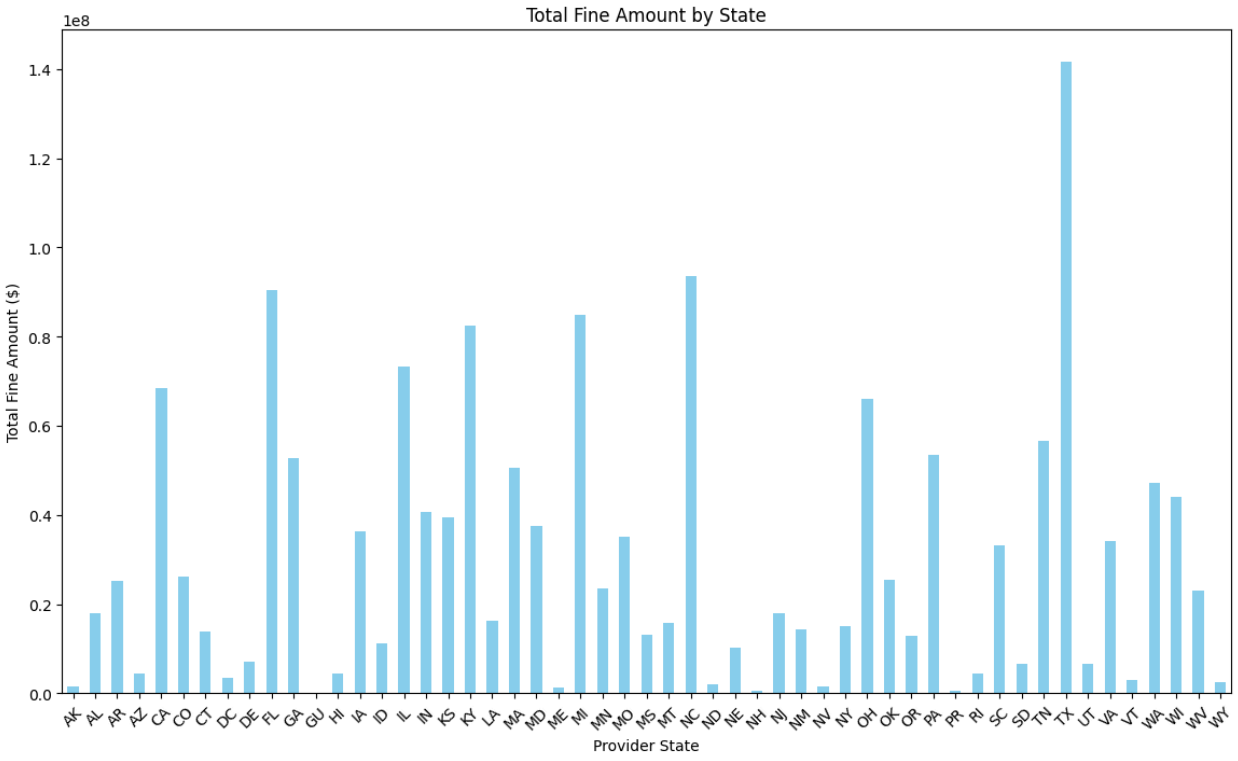

Net Income over the Years
0.3
0.2
0.1
0.0
−0.1
Net Income (Millions $)
2015
2016
2017
2018
2019
2020
2021
Year

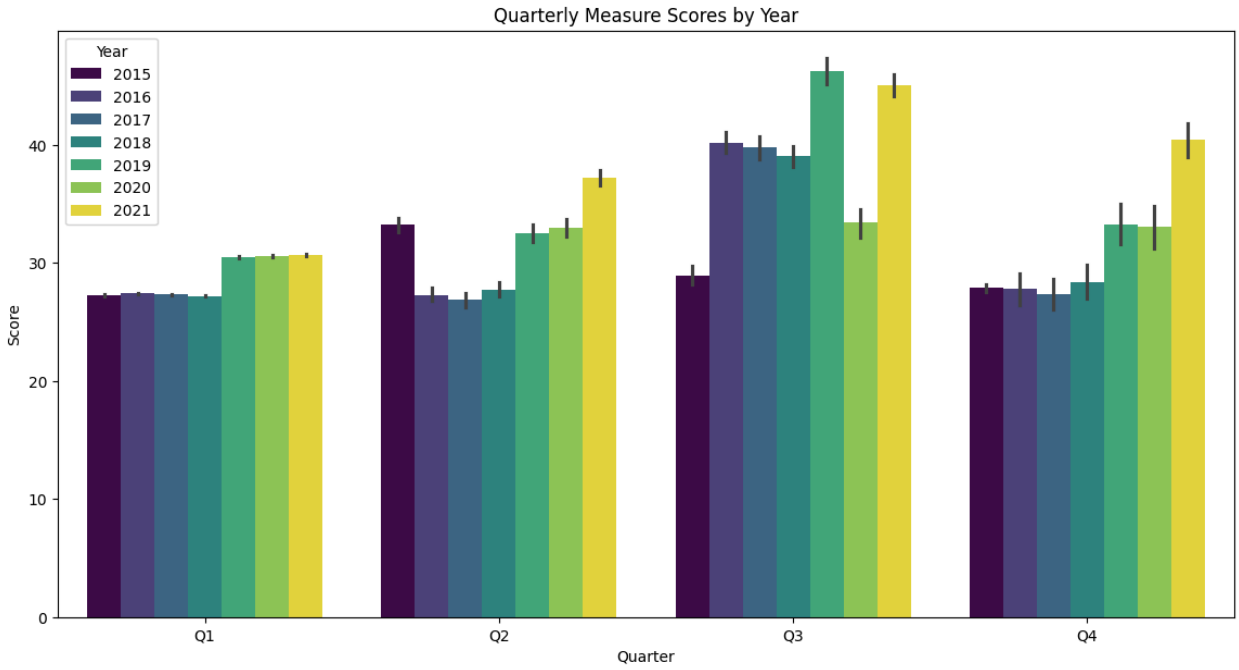

Average Net Income Per Year
Average Net Income
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
2015
2016
2017
2018
2019
2020
2021
Year

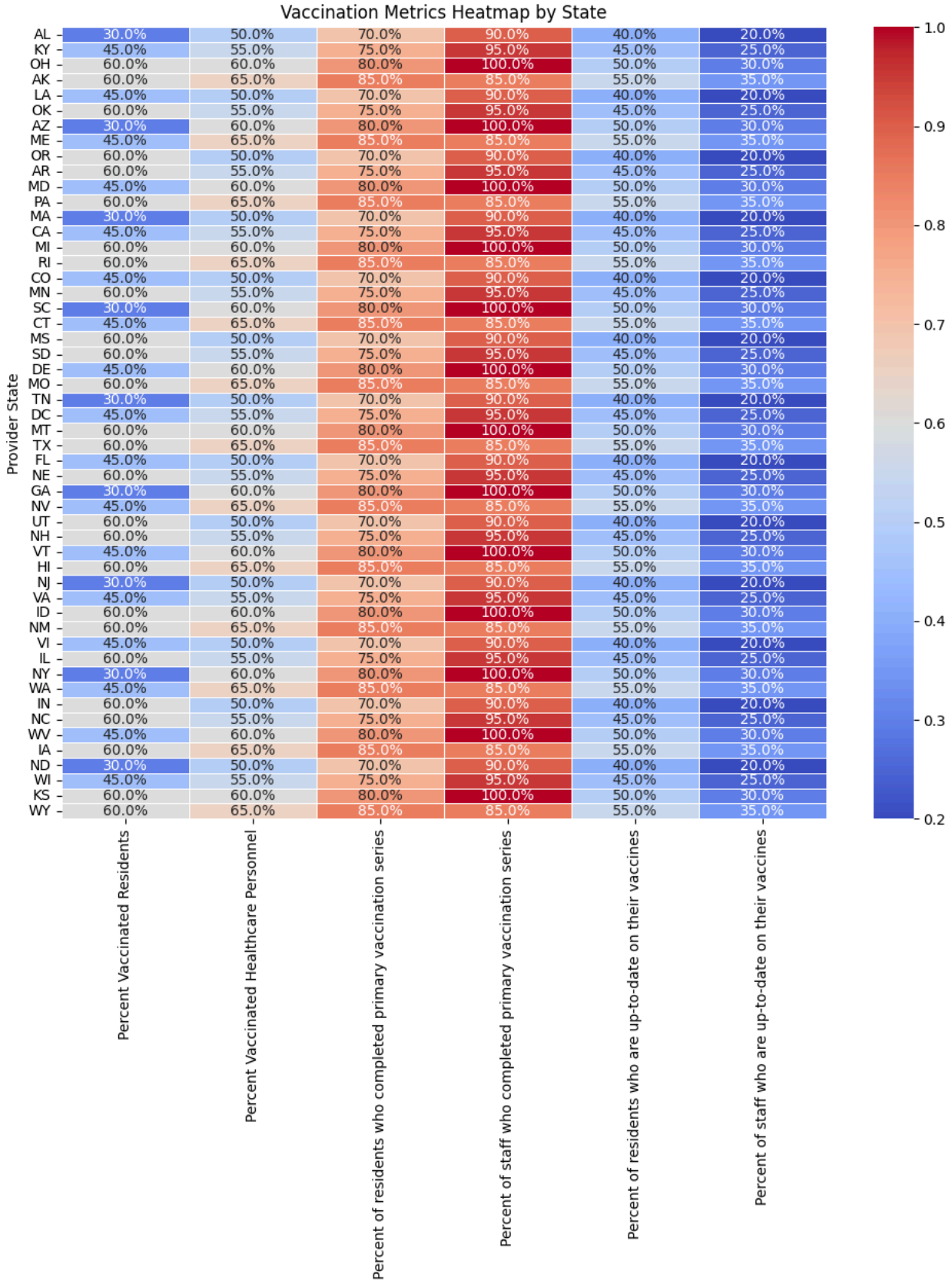

Average Net Income per State
Average Net Income
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA PR RI SC SD TN TX UT VA VT WA WI WV WY
State Code

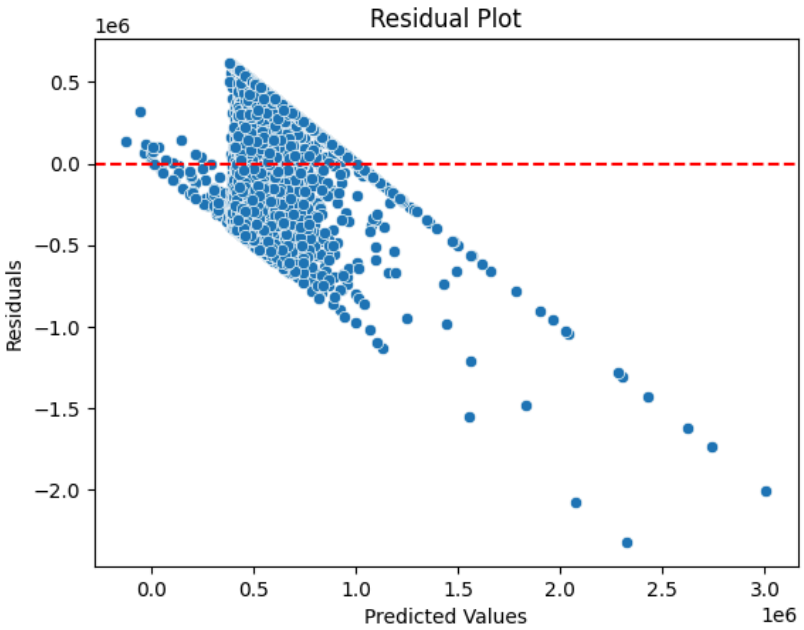

Total Fine Amounts by State and Year
Provider State
Year
1e7
AK 118659 113121 277538 279924 329234 150228 276472
AL 1594704 2723072 3430428 2929020 2013992 2248818 3135747
AR 2412577 3132487 3727477 3892010 3923280 2666369 5552066
AZ 972224 808902 406737 231505 313154 338779 1414292
CA 5022068 6425847 8960460 8783945 8291095 5497474 25409983
CO 616737 612997 2914417 4759910 5212859 2701377 9310380
CT 836938 1041771 1617417 2002343 2399491 1738976 4162182
DC 749356 484740 680607 583509 289907 194008 456171
DE 784952 262211 1027910 1106154 1120807 946255 1923354
FL 10415907 9672830 12262485 11117750 13968436 12172032 20712102
GA 2637723 5082668 8860385 10418148 8099263 6264178 11427963
GU 0 0 0 0 0 0 3250
HI 63481 153567 601889 859596 814142 566755 1409725
IA 1874645 3229001 6711098 7071027 5238972 2426197 9783104
ID 262084 322862 2547403 3095604 1802785 960071 2284474
IL 5675840 4868616 4672439 5808244 8068040 5567044 38660350
IN 1866456 2082905 2513489 4218210 5783395 6500243 17635551
KS 1783460 2913595 7542550 8020088 6872061 3666833 8770312
KY 11052964 14021141 11202059 8590750 10202374 10313276 17157181
LA 3703577 2186007 1724424 1259831 1856549 2000790 3537499
MA 3176275 4357235 6522851 7885846 7392156 6066013 15129511
MD 2211808 3695162 7121212 8527763 6437249 2812641 6770424
ME 56001 50769 113129 149821 158070 158609 563755
MI 3400831 2705686 6279242 11124701 14232573 10347798 36796109
MN 569059 667895 904580 1499680 2250053 2081238 15608755
MO 1706649 2445072 4686128 4319897 5278801 4134540 12654035
MS 2463038 1692649 1478603 990151 1208604 1191404 4168520
MT 89479 597170 1773269 2892946 3530747 2475404 4520453
NC 7802908 9147178 12455486 13712477 16023311 12919566 21414143
ND 8750 251785 291003 73949 406247 191328 819813
NE 143349 340467 1793936 2142305 1756109 863385 3272621
NH 84171 45478 7636 28761 72812 76392 296843
NJ 940684 712740 790593 1833418 2654939 2729603 8223898
NM 321181 1003014 2399738 2872590 2630289 1536981 3463218
NV 295125 187675 207006 124011 69205 23579 541631
NY 1088016 989722 1017068 1798400 2660368 2317488 5158571
OH 4896849 5199222 7710894 8445654 7910847 6640988 25182028
OK 3285075 2482092 3030993 3558540 3816799 2856369 6354217
OR 226558 182824 2215637 2085902 2664023 2216116 3234854
PA 2343396 2529360 4256470 9142400 11381767 7225940 16668856
PR 86190 113068 76213 62352 36527 36251 210620
RI 120150 59246 104920 342239 658362 662938 2505115
SC 3544092 4325746 5789038 4887568 5733263 4020336 4832629
SD 107283 221134 1167559 1514492 1401854 626489 1484238
TN 8358210 7873710 8671019 8561416 7995360 5708748 9384221
TX 11544526 13081776 19625623 22181568 21994230 13729997 39563128
UT 221490 272465 785814 1248363 1375708 680704 2004863
VA 1704037 3255350 4561726 5665939 5554864 4844549 8531393
VT 415653 401839 384684 437751 403994 283397 669401
WA 331504 1446456 6792220 9916370 10222660 5918217 12562896
WI 3295677 3357513 6433320 6921092 5692727 3197777 15045753
WV 1395605 2023818 4510479 6043828 3300115 1910263 3914128
WY 0 0 243681 478557 660291 374204 844449
2015 2016 2017 2018 2019 2020 2021
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5

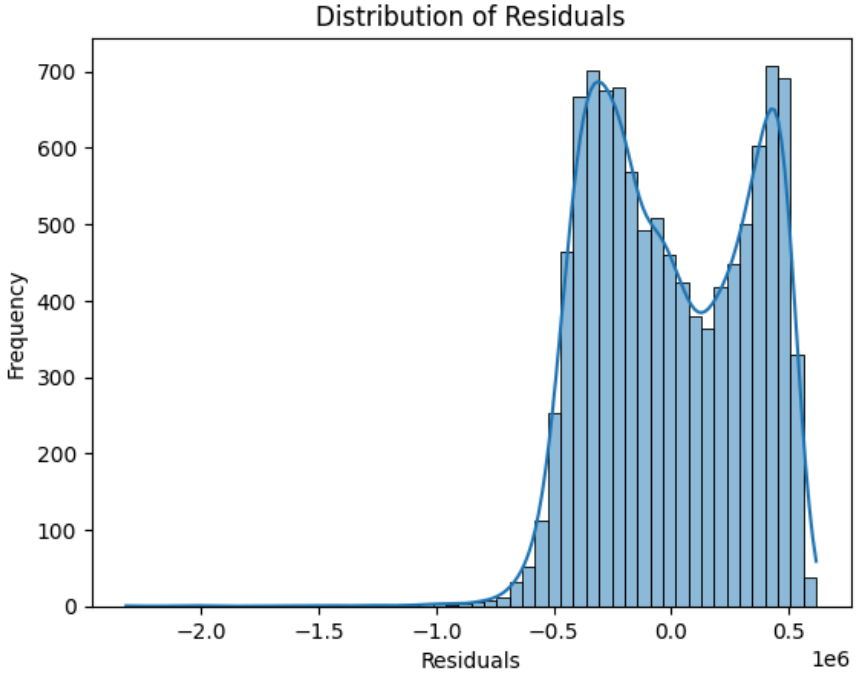

Total Fine Amounts by Penalty Type Over the Years 2015-2021
1e8
Total Fine Amount ($)
Year
Penalty Type
Fine
Payment Denial
0 1 2 3 4
2015 2016 2017 2018 2019 2020 2021


Total Fine Amount by State
1e8
Total Fine Amount ($)
Provider State
AK AL AR AZ CA CO CT DC DE FL GA GU HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA PR RI SC SD TN TX UT VA VT WA WI WV WY


Quarterly Measure Scores by Year
Year
2015
2016
2017
2018
2019
2020
2021
Score
Quarter
Q1
Q2
Q3
Q4

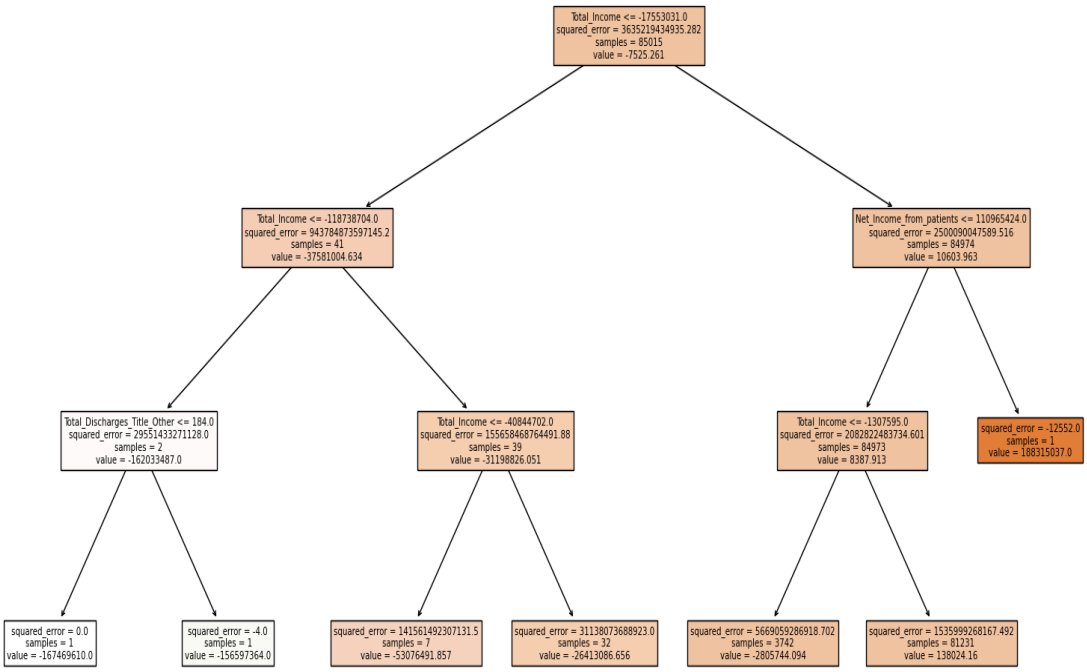

Vaccination Metrics Heatmap by State
Provider State
Percent Vaccinated Residents
Percent Vaccinated Healthcare Personnel
Percent of residents who completed primary vaccination series
Percent of staff who completed primary vaccination series
Percent of residents who are up-to-date on their vaccines
Percent of staff who are up-to-date on their vaccines
AL 30.0% 50.0% 70.0% 90.0% 40.0% 20.0%
KY 45.0% 55.0% 75.0% 95.0% 45.0% 25.0%
OH 60.0% 60.0% 80.0% 100.0% 50.0% 30.0%
AK 60.0% 65.0% 85.0% 85.0% 55.0% 35.0%
LA 45.0% 50.0% 70.0% 90.0% 40.0% 20.0%
OK 60.0% 55.0% 75.0% 95.0% 45.0% 25.0%
AZ 30.0% 60.0% 80.0% 100.0% 50.0% 30.0%
ME 45.0% 65.0% 85.0% 85.0% 55.0% 35.0%
OR 60.0% 50.0% 70.0% 90.0% 40.0% 20.0%
AR 60.0% 55.0% 75.0% 95.0% 45.0% 25.0%
MD 45.0% 60.0% 80.0% 100.0% 50.0% 30.0%
PA 60.0% 65.0% 85.0% 85.0% 55.0% 35.0%
MA 30.0% 50.0% 70.0% 90.0% 40.0% 20.0%
CA 45.0% 55.0% 75.0% 95.0% 45.0% 25.0%
MI 60.0% 60.0% 80.0% 100.0% 50.0% 30.0%
RI 60.0% 65.0% 85.0% 85.0% 55.0% 35.0%
CO 45.0% 50.0% 70.0% 90.0% 40.0% 20.0%
MN 60.0% 55.0% 75.0% 95.0% 45.0% 25.0%
SC 30.0% 60.0% 80.0% 100.0% 50.0% 30.0%
CT 45.0% 65.0% 85.0% 85.0% 55.0% 35.0%
MS 60.0% 50.0% 70.0% 90.0% 40.0% 20.0%
SD 60.0% 55.0% 75.0% 95.0% 45.0% 25.0%
DE 45.0% 60.0% 80.0% 100.0% 50.0% 30.0%
MO 60.0% 65.0% 85.0% 85.0% 55.0% 35.0%
TN 30.0% 50.0% 70.0% 90.0% 40.0% 20.0%
DC 45.0% 55.0% 75.0% 95.0% 45.0% 25.0%
MT 60.0% 60.0% 80.0% 100.0% 50.0% 30.0%
TX 60.0% 65.0% 85.0% 85.0% 55.0% 35.0%
FL 45.0% 50.0% 70.0% 90.0% 40.0% 20.0%
NE 60.0% 55.0% 75.0% 95.0% 45.0% 25.0%
GA 30.0% 60.0% 80.0% 100.0% 50.0% 30.0%
NV 45.0% 65.0% 85.0% 85.0% 55.0% 35.0%
UT 60.0% 50.0% 70.0% 90.0% 40.0% 20.0%
NH 60.0% 55.0% 75.0% 95.0% 45.0% 25.0%
VT 45.0% 60.0% 80.0% 100.0% 50.0% 30.0%
HI 60.0% 65.0% 85.0% 85.0% 55.0% 35.0%
NJ 30.0% 50.0% 70.0% 90.0% 40.0% 20.0%
VA 45.0% 55.0% 75.0% 95.0% 45.0% 25.0%
ID 60.0% 60.0% 80.0% 100.0% 50.0% 30.0%
NM 60.0% 65.0% 85.0% 85.0% 55.0% 35.0%
VI 45.0% 50.0% 70.0% 90.0% 40.0% 20.0%
IL 60.0% 55.0% 75.0% 95.0% 45.0% 25.0%
NY 30.0% 60.0% 80.0% 100.0% 50.0% 30.0%
WA 45.0% 65.0% 85.0% 85.0% 55.0% 35.0%
IN 60.0% 50.0% 70.0% 90.0% 40.0% 20.0%
NC 60.0% 55.0% 75.0% 95.0% 45.0% 25.0%
WV 45.0% 60.0% 80.0% 100.0% 50.0% 30.0%
IA 60.0% 65.0% 85.0% 85.0% 55.0% 35.0%
ND 30.0% 50.0% 70.0% 90.0% 40.0% 20.0%
WI 45.0% 55.0% 75.0% 95.0% 45.0% 25.0%
KS 60.0% 60.0% 80.0% 100.0% 50.0% 30.0%
WY 60.0% 65.0% 85.0% 85.0% 55.0% 35.0%
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

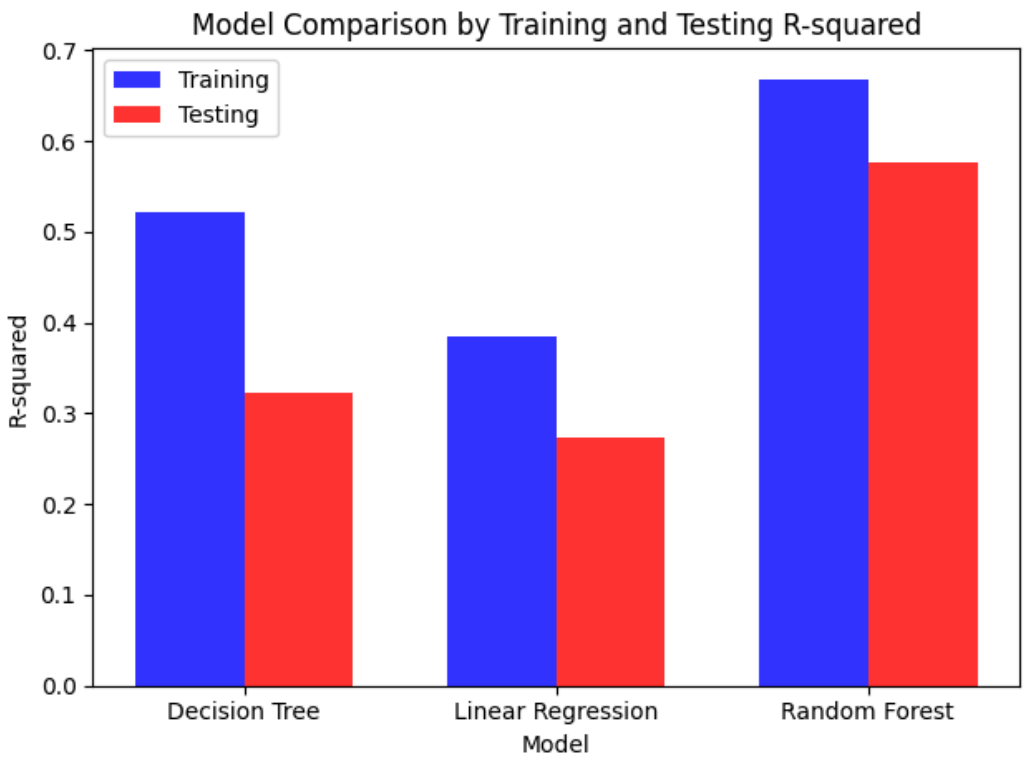




	Features	Coefficients	p-values
0	Less_Total_Operating_Expense	18095.876921	0.0
1	Total_Salaries_adjusted	67956.799082	0.0
2	Net_Patient_Revenue	75476.446800	0.0

```
Variables with high correlation with Net_Income (absolute correlation > 0.30):
Inpatient_PPS_Amount                                0.312355
Inpatient_Revenue                                   0.328968
Net_Income                                          1.000000
Net_Patient_Revenue                                 0.357998
Overhead_Non_Salary_Costs                           0.339591
SNF_Days_Title_XVIII                                0.340202
SNF_Days_Total                                      0.306341
SNF_bed_Days_Available                              0.338912
Total_Bed_Days_Available                            0.338912
Total_Days_Title_XVIII                              0.337167
Total_Days_Total                                    0.322214
Total_General_Inpatient_Revenue                     0.327974
Total_RUG_Days                                      0.325846
Total_Salaries_From_Worksheet_A                     0.349363
Total_Salaries_adjusted                             0.349314
Total Days Title V                                  0.410570
SNF Days Title V                                    0.361986
Nursing and Allied Health Education Activities      0.719044
Name: Net_Income, dtype: float64
```




Model Comparison by Training and Testing R-squared
Training
Testing
R-squared
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Decision Tree
Linear Regression
Random Forest
Model

References

https://data.cms.gov/provider-data/topics/nursing-homes

https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/#:~:text=The%20random%20forest%20has%20complex,is%20more%20accurate%20in%20predictions