
# Predicting Malpractice Premiums: Analyzing the Impact of Hospital Costs and Other Predictors

Ismat Ara Khan

18.12.24

# 1 Abstract

This report investigates the prediction of malpractice costs in hospitals and examines the factors that influence these costs. Using a variety of hospital-level data, including both cost-related and non-cost-related variables, the study aims to identify the most significant predictors of malpractice premiums per inpatient day. A comprehensive exploratory data analysis (EDA) is performed to uncover underlying relationships between the response variable and potential predictors. The report employs multiple predictive modeling techniques, including decision trees, random forests, and neural networks, to evaluate the effectiveness of these models in forecasting malpractice costs. The results provide insights into the importance of various factors, such as hospital characteristics, care metrics, and population-specific factors, in predicting malpractice premiums. The findings suggest that a combination of these variables can significantly improve the prediction of malpractice costs, which may assist hospitals in managing financial risk. The report concludes by comparing the performance of different models and highlighting the best approach for predicting malpractice costs in the context of hospital management.

# 2 Introduction

Malpractice costs are a significant financial burden for hospitals and healthcare providers, impacting their overall financial health and patient care outcomes. These costs, primarily associated with insurance premiums, can vary significantly based on a range of hospital and patient-related factors. Predicting malpractice premiums accurately is crucial for effective financial planning and resource allocation within healthcare institutions. However, the complexity of the factors involved in malpractice costs—such as hospital characteristics, care quality metrics, and patient demographics—makes prediction a challenging task.

This report focuses on predicting malpractice premiums per inpatient day, using a range of hospital-level data. By identifying the most significant predictors of malpractice costs, this study aims to provide insights that can assist healthcare administrators in better understanding and managing the financial risks associated with malpractice insurance. The study incorporates exploratory data analysis (EDA) to uncover key relationships between the response variable—malpractice premiums and various predictor variables, both cost-related and non-cost-related.

The report employs several machine learning models, including decision trees, random forests, and neural networks, to compare their predictive performance. These models are evaluated based on their accuracy, interpretability, and ability to generalize to new data. The ultimate goal of this analysis is to identify the best predictive model for estimating malpractice costs and to highlight the key factors that influence these costs, providing valuable insights for hospital decision-making and policy development.

# 3 Methodology

The methodology for this report involves several key steps: data collection and pre-processing, exploratory data analysis (EDA), model selection and evaluation, and comparison of predictive models. Each of these steps contributes to building a reliable model for predicting malpractice costs and understanding the factors that drive these costs.

## 3.1 Data Collection

The data set used in this study includes hospital-level data containing various continuous and categorical variables. The data set consists of 2616 observations from USA Hospital Data collected in 2021 with 74 variables. The response variable of interest is the malpractice premiums per inpatient day. The data set includes hospital characteristics such as hospital size, SDI score, charity care burden, salary costs per inpatient day, readmission rates, and several other hospital performance metrics.

## 3.2 Exploratory Data Analysis (EDA)

EDA is performed to uncover the underlying relationships between the response variable (malpractice costs) and potential predictors. Key EDA steps include:

### 3.2.1 Correlation Analysis

A correlation matrix is calculated to assess the strength and direction of the relationships between continuous variables and malpractice costs.

### 3.2.2 Heatmap:

A correlation heatmap for some significant variables with respect to the response variable, malpractice premiums, provides valuable insights into the relationships between the predictors and the response. It helps with model selection, identifying multicollinearity, and feature selection, which ultimately improves the model-building process and prediction accuracy.

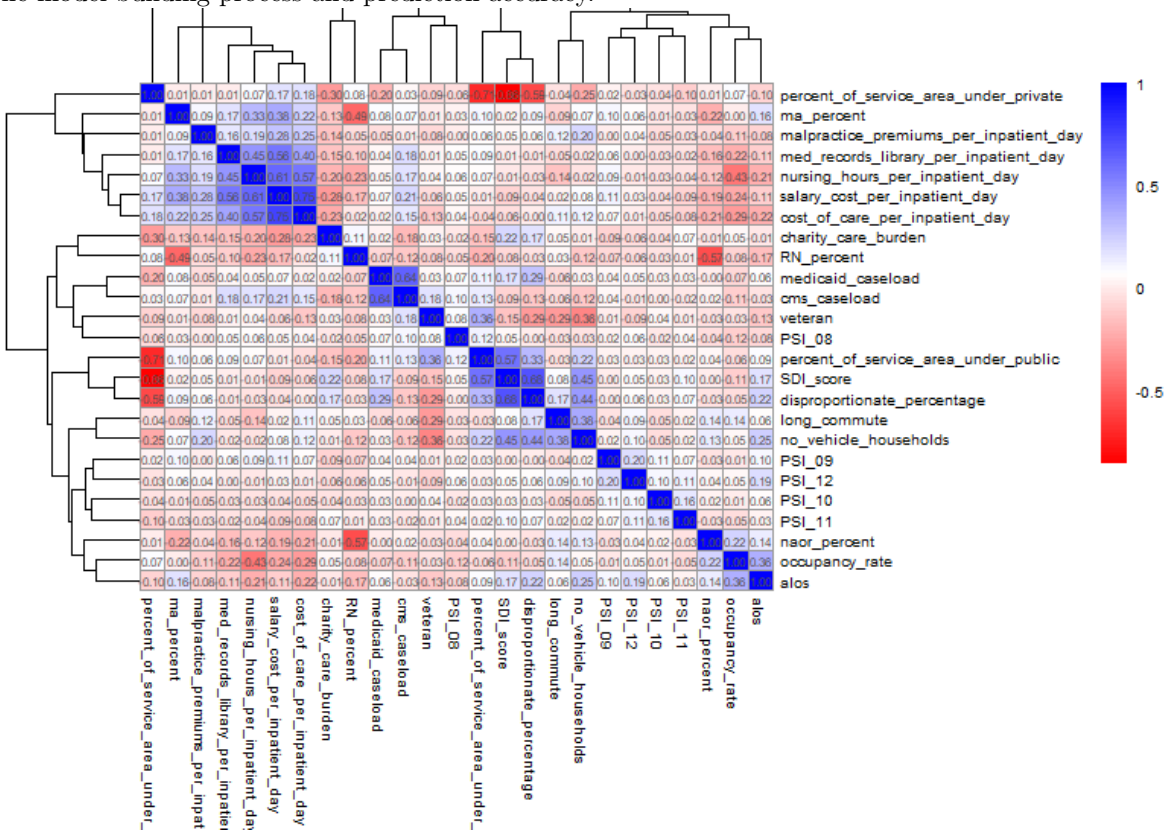


Figure 1: Correlation matrix between potential variables

This Correlation heatmap describe some important insights as follows,

- Positive Correlations: Variables like "salary cost per inpatient day" and "nursing hours per inpatient day" show strong positive relationships with "malpractice premiums per inpatient day", indicating higher operational costs are linked to higher premiums.
- Negative Correlations: Variables like "percent of service area under public" show negative correlations, suggesting hospitals serving public areas may face lower malpractice premiums.
- Operational Scale: Variables like "med records library per inpatient day" may reflect workload or hospital scale, influencing malpractice premiums positively.
- Weak Relationships: Some variables show little to no correlation with malpractice premiums, implying minimal impact or indirect associations.
- Key Focus Areas: Cost-related variables and service area characteristics appear to be significant predictors of malpractice premiums.

### 3.2.3 Visualization:

Various visualizations are created to explore the data for further analysis:

### 3.2.4 Histogram:

Understanding the distribution of the response variable is crucial for selecting the right model. The following plot of the histogram is showing the distribution of the response variable "malpractice premiums per inpatient day".

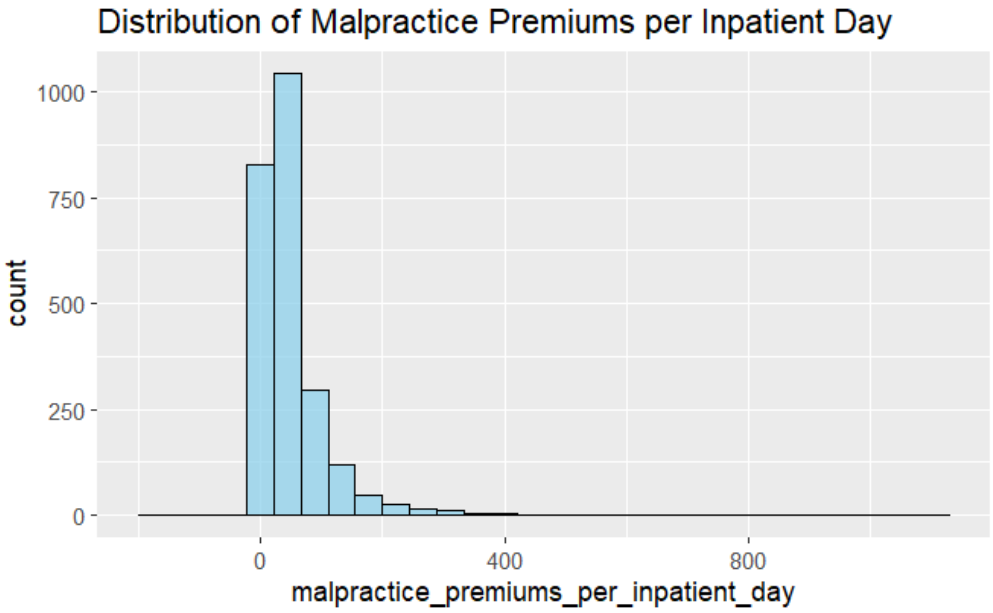


Figure 2: Histogram of response variables

From this histogram we can visualize that,

1. The distribution is likely positively skewed (right-skewed), as most of the values lie in the lower range, but the tail extends to higher values.

2. The highest frequency small range, indicating that most observations are concentrated in this range.

3. The distribution has a long tail to the right, with fewer observations in the higher ranges.

which indicates that, since the data is highly skewed or follows a non-normal distribution, certain models (e.g., linear regression) may not perform well. We have to choose models that can handle skewed data, such as Decision Tree, Random Forest etc.

### 3.2.5 Bar Plot:

To visualize the frequency distributions of categorical variables and their relationship with malpractice costs.

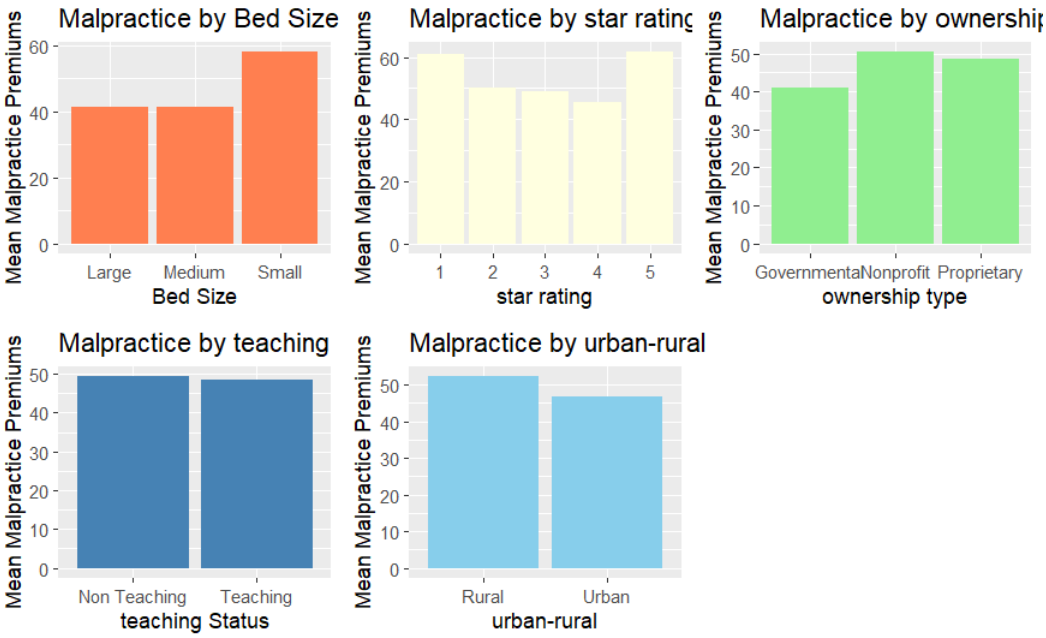


Figure 3: Bar plots of Categorical Variables with Response Variables

These bar plots display the mean malpractice premiums for various Categorical Variables:

- Malpractice by Bed Size: Small and Medium bed size hospitals have higher mean malpractice premiums than Large bed size hospitals.
- Malpractice by Star Rating: Mean malpractice premiums appear fairly uniform across star ratings (1–5). There doesn't seem to be much difference among the groups.
- Malpractice by Ownership: Governmental/Nonprofit hospitals and Proprietary hospitals have similar mean malpractice premiums.
- Malpractice by Teaching Status: Teaching hospitals and non-teaching hospitals have nearly equal mean malpractice premiums.
- Malpractice by Urban-Rural: Urban and rural hospitals show similar mean malpractice premiums.

### 3.2.6 Box Plot:

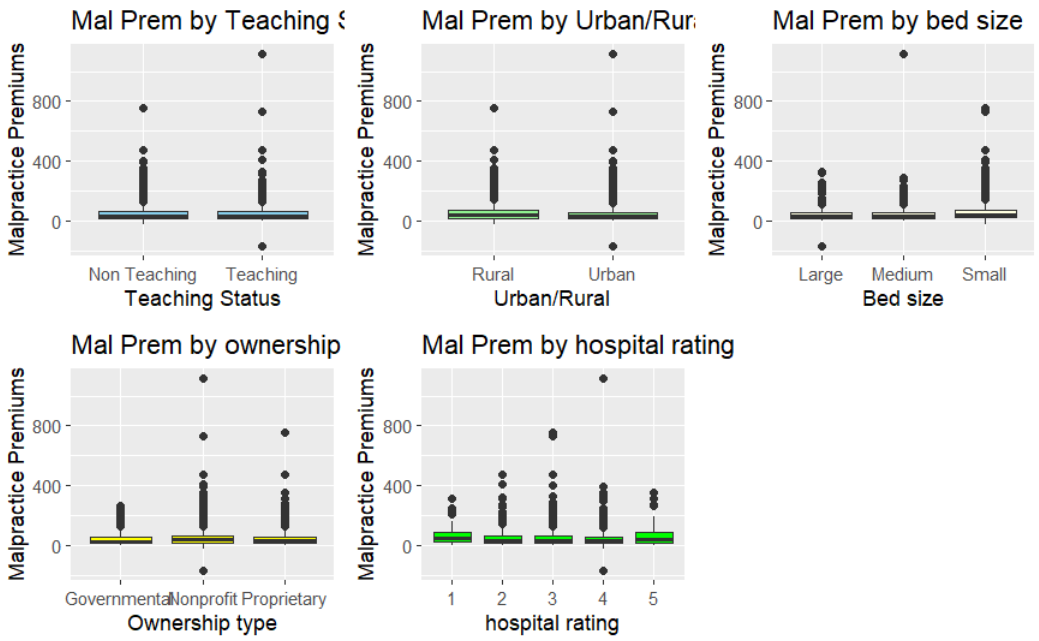


Figure 4: Box plots of categorical variables with response variables

Box plots highlight variability in malpractice premiums, particularly among bed size categories and ownership types, with visible outliers in all cases.

### 3.2.7 Scatter Plot:

To visualize bivariate relationships between continuous predictors and the response variable.

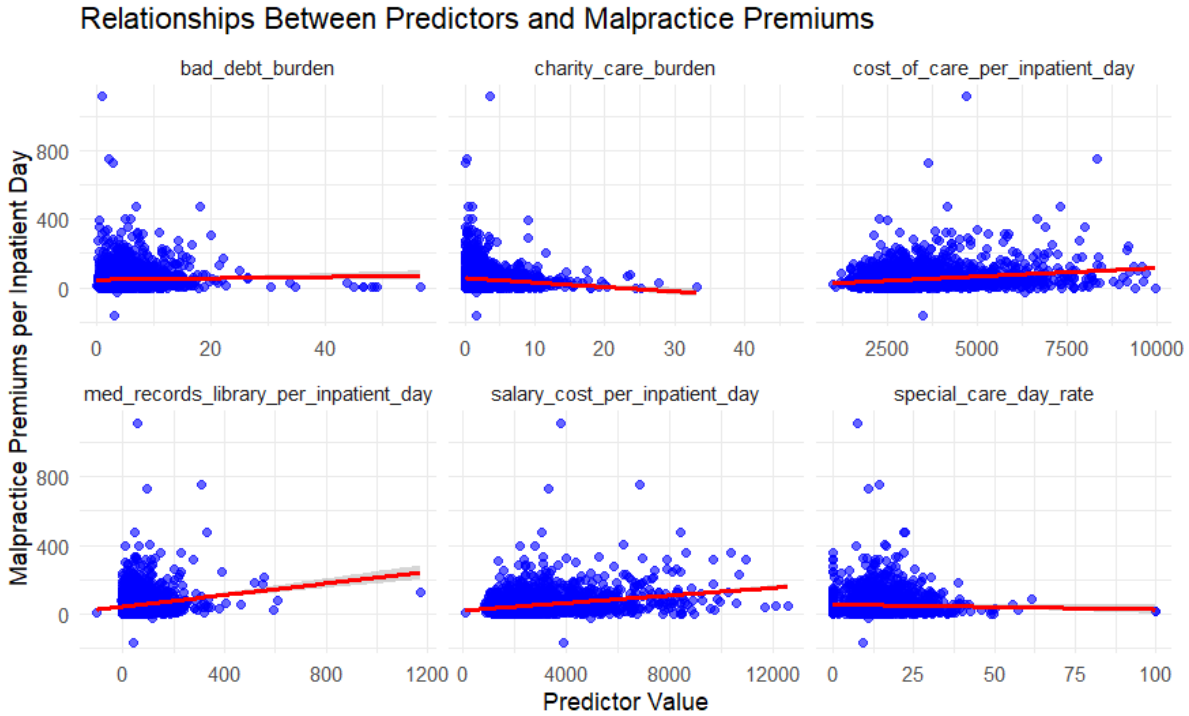


Figure 5: scatter plots of cost related variables with response variables

The scatter plots reveal that certain predictors like med records library per inpatient day and salary cost per inpatient day show positive trends, while others, such as bad debt burden and special care day rate, exhibit weak or negative trends.

For most predictors, data points are densely concentrated at lower values, indicating skewed distributions. The fitted trend lines suggest that the relationships between the predictors and malpractice premiums are relatively weak overall, but some variables hint at positive or negative associations.

This visualization effectively highlights how these cost related predictor variables influence malpractice premiums and helps identify patterns for further analysis.

#### 3.2.8 Descriptive Statistics

Summary statistics such as mean, median, standard deviation, and percentiles are calculated for the continuous variables.

| | Min | Q1 | Median | Mean | Q3 | Max | SD |
|---|---|---|---|---|---|---|---|
| malpractice_premiums_per_inpatient_day | -167.19 | 16.05 | 33.01 | 49.05 | 61.04 | 1118.79 | 60.29 |
| cost_of_care_per_inpatient_day | 970.29 | 2317.09 | 2878.84 | 3255.66 | 3763.20 | 9982.50 | 1386.62 |
| charity_care_burden | 0.00 | 0.92 | 1.92 | 2.99 | 4.03 | 46.19 | 3.25 |
| bad_debt_burden | 0.01 | 2.30 | 4.12 | 5.09 | 6.81 | 56.64 | 4.42 |
| salary_cost_per_inpatient_day | 107.21 | 1750.81 | 2350.55 | 2800.58 | 3323.95 | 12572.33 | 1569.68 |
| alos | 1.53 | 4.68 | 5.29 | 5.38 | 5.95 | 17.36 | 1.12 |
| occupancy_rate | 5.57 | 40.00 | 53.80 | 52.55 | 65.43 | 152.18 | 18.11 |
| private_room_percentage | 0.00 | 0.00 | 0.00 | 1.70 | 0.00 | 96.19 | 10.12 |
| RN_percent | 36.40 | 66.07 | 71.28 | 71.28 | 76.88 | 96.73 | 8.31 |
| LPN_percent | 0.00 | 3.98 | 5.77 | 7.34 | 9.08 | 52.54 | 5.26 |
| naor_percent | 0.02 | 13.62 | 18.35 | 18.41 | 23.21 | 39.92 | 7.10 |
| ma_percent | 0.00 | 0.00 | 0.77 | 3.09 | 3.90 | 39.44 | 5.16 |
| nursing_hours_per_inpatient_day | 2.84 | 18.19 | 21.72 | 24.66 | 27.03 | 153.00 | 11.84 |
| special_care_day_rate | 0.00 | 9.82 | 13.72 | 14.39 | 18.26 | 100.00 | 7.80 |
| SDI_score | 5.33 | 40.37 | 52.16 | 52.20 | 63.21 | 99.29 | 16.97 |
| veteran | 1.07 | 5.89 | 7.37 | 7.43 | 8.69 | 25.20 | 2.71 |
| Medicaid_Caseload | 0.00 | 1.94 | 4.19 | 7.75 | 9.76 | 66.10 | 8.92 |
| disproportionate_percentage | 0.00 | 0.23 | 0.30 | 0.32 | 0.39 | 1.16 | 0.15 |
| cms_caseload | 1.11 | 25.32 | 32.02 | 34.12 | 41.56 | 89.65 | 12.55 |
| percent_of_service_area_under_public | 13.87 | 33.53 | 38.86 | 39.10 | 44.38 | 74.31 | 8.19 |
| percent_of_service_area_under_private | 22.55 | 59.42 | 65.80 | 65.21 | 71.85 | 88.03 | 9.22 |
| hospital_wide_readmission_rate | 11.60 | 14.00 | 14.50 | 14.53 | 15.10 | 20.70 | 0.88 |
| heart_failure_readmission_rate | 14.30 | 19.40 | 20.20 | 20.27 | 21.10 | 27.40 | 1.37 |
| acute_myocardial_infarction_readmission_rate | 11.00 | 13.40 | 14.00 | 14.03 | 14.60 | 17.70 | 0.93 |
| copd_readmission_rate | 15.50 | 18.60 | 19.20 | 19.32 | 20.00 | 24.10 | 1.14 |
| hip_knee_replacement_readmission_rate | 2.60 | 3.90 | 4.20 | 4.28 | 4.60 | 6.30 | 0.52 |
| pneumonia_readmission_rate | 13.60 | 16.30 | 16.90 | 16.95 | 17.60 | 24.30 | 1.04 |
| clabsi_rate | 0.00 | 0.23 | 0.75 | 0.92 | 1.30 | 17.39 | 0.99 |
| cauti_rate | 0.00 | 0.29 | 0.69 | 0.81 | 1.17 | 7.83 | 0.72 |
| ssi_colon_rate | 0.00 | 0.00 | 15.04 | 21.24 | 30.30 | 857.14 | 33.17 |
| ssi_hyster_rate | 0.00 | 0.00 | 0.00 | 9.11 | 10.92 | 1000.00 | 35.41 |
| mrsa_rate | 0.00 | 0.00 | 0.05 | 0.06 | 0.09 | 0.50 | 0.06 |
| cdi_rate | 0.00 | 0.11 | 0.21 | 0.27 | 0.37 | 1.85 | 0.22 |
| PSI_03 | 0.05 | 0.29 | 0.44 | 0.57 | 0.69 | 6.31 | 0.49 |
| PSI_04 | 86.68 | 154.64 | 167.86 | 168.68 | 182.77 | 241.81 | 21.73 |
| PSI_06 | 0.12 | 0.22 | 0.24 | 0.25 | 0.27 | 0.51 | 0.04 |
| PSI_08 | 0.06 | 0.09 | 0.09 | 0.09 | 0.09 | 0.13 | 0.01 |
| PSI_09 | 1.10 | 2.28 | 2.45 | 2.51 | 2.68 | 6.10 | 0.45 |
| PSI_10 | 0.47 | 1.43 | 1.53 | 1.58 | 1.57 | 4.55 | 0.36 |

Table 1: Descriptive Statistics for Selected Variables

This table provides a comprehensive summary of descriptive statistics for the variables analyzed in the study. Key observations are outlined below:

- Malpractice Premiums per Inpatient Day:

The variable shows a wide range, with values spanning from a minimum of -167.19 to a maximum of 1118.79. The mean is 49.05, and the standard deviation is relatively high (60.29), indicating notable variability.

- Cost of Care per Inpatient Day:

  This variable has the highest mean (3255.66) and a maximum of 9982.50, reflecting the significant cost burden associated with inpatient care. The standard deviation (1386.62) suggests substantial variability among observations.

- Charity and Bad Debt Burden:

  Charity care burden ranges from 0.00 to 46.19, with a mean of 2.99, showing skewness due to outliers. Bad debt burden also varies widely (0.01–56.64), with a mean of 5.09. Labor-Related Costs:

- Salary cost per inpatient day ranges dramatically from 107.21 to 12572.33, with a mean of 2800.58. Nursing hours per inpatient day also exhibit variability, with a mean of 24.66 and a standard deviation of 11.84.

- Hospital Characteristics:

  The average length of stay (ALOS) has a relatively narrow range (1.53–17.36) and low standard deviation (1.12). Occupancy rates and private room percentages demonstrate wide variability, with means of 52.55Staffing Mix:

- Registered Nurse (RN) percentage has a mean of 71.28Licensed Practical Nurse (LPN) percentage is much lower, with a mean of 7.34

- Clinical and Quality Metrics:

  Readmission rates (e.g., hospital-wide, heart failure, and pneumonia) are relatively consistent, with means close to standard thresholds (e.g., 14.53Infection rates (e.g., CLABSI and CAUTI) show variability but are generally low in magnitude.

- Social Deprivation Index (SDI):

  SDI scores range from 5.33 to 99.29, with a mean of 52.20, suggesting significant disparities across regions.

- Service Area and Payer Metrics:

  Percentages of service areas under public and private payers exhibit complementary distributions, with means of 39.10

- Patient Safety Indicators (PSIs):

  Indicators such as $PSI_03$ (pressure ulcers) and $PSI_04$ (death among surgical inpatients) have low mean rates but substantial variability, particularly $PSI_04$ (mean = 168.68, SD = 21.73).

# 4 Selection of significant variables by using Lasso

LASSO (Least Absolute Shrinkage and Selection Operator) is a regression technique that is particularly useful for variable selection and regularization. It works by adding a penalty term to the ordinary least squares (OLS) loss function. This penalty is proportional to the absolute value of the coefficients:

$$\min_{\beta}(\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|)$$

where,
$y_i$ is the response variable,
$x_{ij}$ represents the predictor variables,
$\beta_j$are the coefficients for predictors,
$\lambda$ is the tuning parameter that controls the strength of the penalty.

| Variable | Coefficient |
|---|---|
| (Intercept) | 18.68 |
| CCN | -0.00 |
| bad_debt_burden | 0.01 |
| salary_cost_per_inpatient_day | 0.01 |
| alos | -3.78 |
| private_room_percentage | 0.01 |
| semi_private_room_percentage | -0.01 |
| LPN_percent | -0.02 |
| ma_percent | 0.14 |
| discharge_information_star_rating | -2.72 |
| quietness_star_rating | -0.61 |
| recommend_hospital_star_rating | -1.47 |
| crowded | -0.54 |
| unemployment | 1.02 |
| no_vehicle_households | 2.55 |
| long_commute | 0.76 |
| medicaid_caseload | -0.10 |
| percent_of_service_area_under_public | 0.04 |
| mrsa_rate | -34.50 |
| PSI_08 | -53.30 |
| PSI_09 | -1.07 |
| PSI_10 | -2.61 |
| PSI_12 | 1.95 |
| hospital_wide_readmission_rate | 1.49 |
| pneumonia_readmission_rate | 0.51 |

Table 2: Selection of significant variables by using Lasso

The variables selected by LASSO capture various factors affecting malpractice premiums, ranging from socioeconomic variables (e.g., unemployment, no vehicle households) to hospital operational metrics (e.g., ALOS, readmission rates, PSI indicators). These insights allow hospitals to focus on key factors influencing costs and improve risk management.

# 5 Model Selection and Evaluation:

Several machine learning models are selected to predict malpractice costs based on some important factor of this data.

- Complex Relationships: Malpractice costs could depend on non-linear interactions (e.g., a combination of hospital size, patient satisfaction, and regional factors might influence costs).
- Interactions Between Variables: Models like Random Forests and Neural Networks can automatically capture interactions, whereas linear models require you to specify these interactions manually.
- Large Dataset Handling: If your dataset contains many observations and variables, tree-based models (especially Random Forests) and Neural Networks can leverage this scale for better predictions.
- Feature Importance: Random Forests provide feature importance scores, which are useful for understanding the key drivers of malpractice costs.
- Predictive Power: Tree-based models and neural networks often outperform traditional linear models in capturing complex patterns in data, leading to better predictions.

## 5.1 Decision Tree:

A non-linear model that splits the data into subsets based on feature values, aiming to minimize variance in the target variable within each subset. Decision trees provide interpretable results, making them useful for identifying key factors influencing malpractice costs.

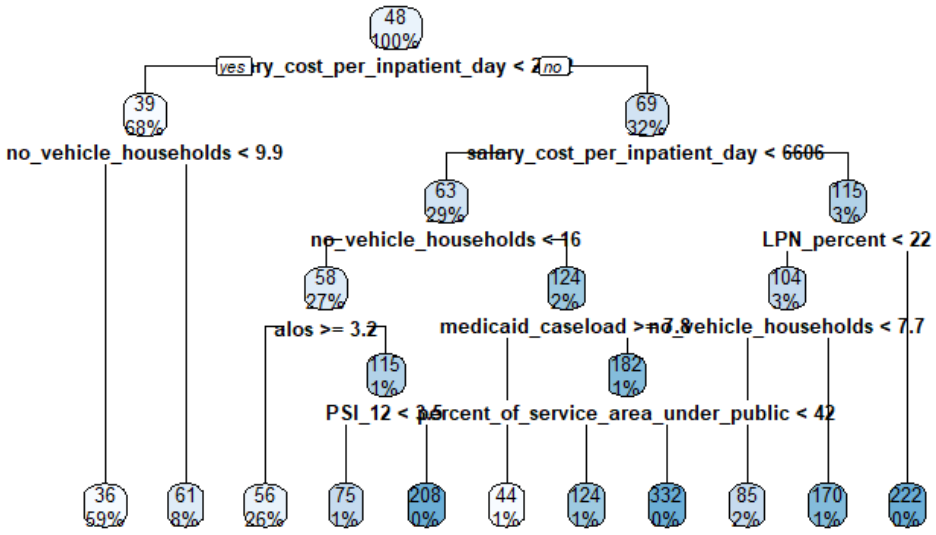


Figure 6: Decision Tree Plot

Key observations that we can interpret from this plot that salary cost per inpatient day is the strongest predictor, forming the root of the decision tree. Variables like no vehicle households, LPN percent, and percent of service area under public play key roles in subsequent splits.

The tree captures non-linear relationships effectively and shows how different thresholds for predictors divide the data into subsets with varying malpractice costs.

## 5.2 Random Forest:

An ensemble method based on decision trees, random forests aggregate predictions from multiple decision trees to improve accuracy and reduce over-fitting. Random forests are robust to noisy data and are useful for handling both linear and nonlinear relationships.

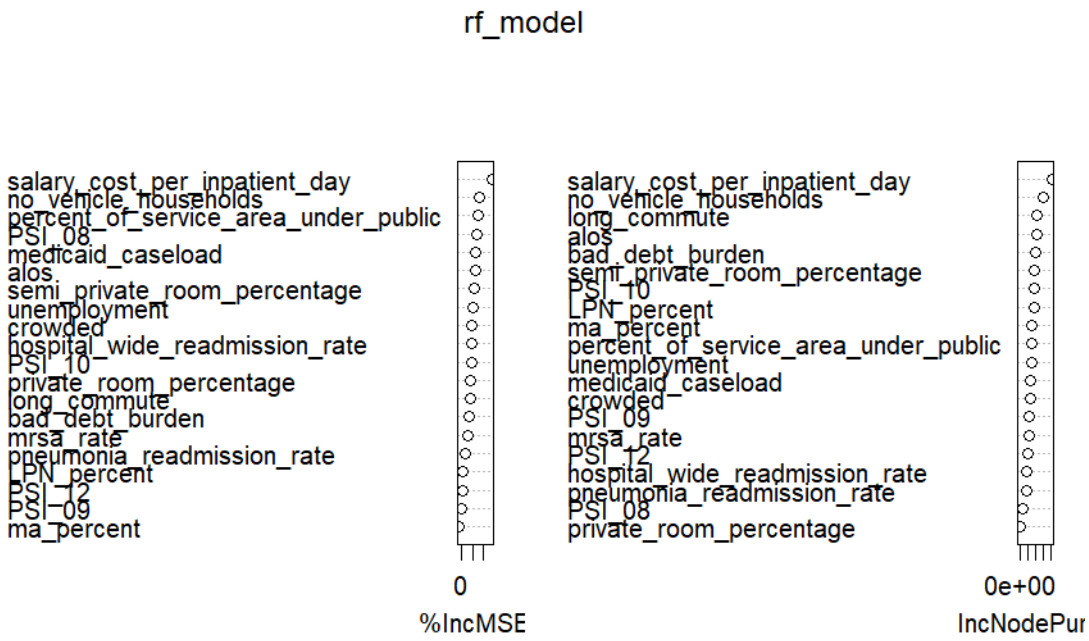


Figure 7: Random Forest Plot

This model provides interpretability and insights that salary cost per inpatient day consistently appears as the most important predictor, aligning with its role in the decision tree. Variables like no vehicle households and percent of service area under public are also significant, reinforcing their importance. Random forests handle complex relationships better than individual decision trees by combining predictions across multiple trees.

### 5.3 Neural Networks:

A more complex, non-linear model that is capable of capturing intricate patterns in the data. Neural networks are used to explore whether deep learning techniques can outperform traditional machine learning methods to predict malpractice costs.

Each model is trained using the full dataset, with cross-validation used to estimate generalization performance. Hyper-parameter tuning is performed to optimize the model parameters.

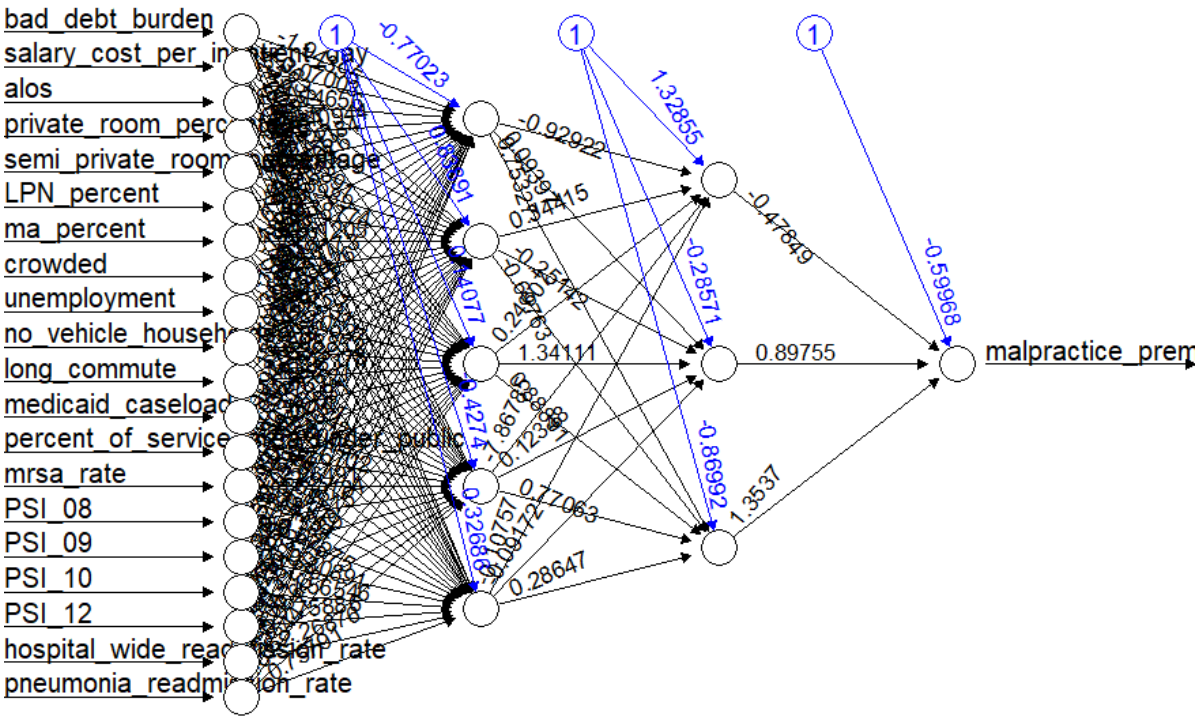


Figure 8: Neural Network Plot

- This neural network identifies how input features like salary cost per inpatient day, alos, PSI 08, and others contribute to predicting cost of malpractice prem.
- Some inputs have a stronger influence (high weights, either positive or negative), while others contribute less.
- The hidden layer processes these weighted inputs and combines them to predict the malpractice premiums.

# 6 Model Comparison and Evaluation Metrics:

The performance of the models is compared using the following evaluation metric.

Root Mean Squared Error (RMSE): A common metric for regression tasks, RMSE measures the average magnitude of error between predicted and actual values. It penalizes larger errors more than smaller ones.

| **Model** | **RMSE** |
|---|---|
| Decision Tree | 64.87 |
| Random Forest | 63.04 |
| Neural Network | 133.64 |

Table 3: RMSE Values for Different Models

From this table we can conclude as the Random Forest model is the most appropriate choice for predicting malpractice costs due to its balance of predictive accuracy and ability to handle complex relationships in the data. However, if model interpretability is crucial for stakeholders, the Decision Tree model can be considered as an alternative with only slightly lower performance, but the Neural Network model is not suitable for this specific task given its significantly higher RMSE, and further tuning or adjustments would be needed if it were to be reconsidered.

# 7 Key Predictor Variables:

Through the modeling process, the most important predictors of malpractice costs are identified.In this case, "Salary cost per inpatient day" variable is the most significant predictor among all cost related variable and "no vehicle households", and "alos" were also identified as the most significant predictors among non cost variables contributing substantially to the model's accuracy and node purity. These predictors are selected based on their ability to explain the variance in the response variable. This allows healthcare administrators to gain insights into which factors most significantly influence malpractice premiums.

# 8 Conclusion:

This process of analyzing the correlation matrix and refining the list of significant variables is an essential part of building an accurate model to predict malpractice costs. By using both statistical methods and domain knowledge, we were able to identify the most relevant predictors and ensure that the selected variables provided a strong foundation for the final predictive model.

# References


[1] R. Kumar and A. Indrayan, "Receiver Operating Characteristic (ROC) Curve for Medical Researchers," *Indian Pediatrics*, vol. 48, no. 4, pp. 277–287, 2011.

[2] K. Topuz, T. Van Gestel, and J. M. Garibaldi, "Predictive Analytics in Healthcare: Applications, Challenges, and Future Perspectives," *International Journal of Health Care Management*, vol. 12, no. 1, pp. 30–40, 2019.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.


**R Code**

```
library(ggplot2)
library(gridExtra)
library(GGally)
library(pheatmap)
library(ggplot2)
library(patchwork)
library(dplyr)
library(knitr)
library(kableExtra)
library(glmnet)
library(tidyselect)
library(caret)

# Load the dataset
h_data <- read.csv("C:/Users/Ismat/OneDrive/Desktop/fall24/Stat consulting/Final Project/2021_USA_HOS
colnames(h_data)
str(h_data)


# Box Plot
# Change the class of the variable from integer to character
h_data$overall_hospital_rating_star_rating <- as.character(h_data$overall_hospital_rating_star_rating


plot1 = ggplot(h_data, aes(x = teaching_status, y = malpractice_premiums_per_inpatient_day)) +
```

```
  geom_boxplot(fill = "skyblue") +
  labs(title = "Mal Prem by Teaching Status", x = "Teaching Status", y = "Malpractice Premiums")

plot2 = ggplot(h_data, aes(x = urban_rural, y = malpractice_premiums_per_inpatient_day)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Mal Prem by Urban/Rural", x = "Urban/Rural", y = "Malpractice Premiums")

plot3 = ggplot(h_data, aes(x = bed_size_simple, y = malpractice_premiums_per_inpatient_day)) +
  geom_boxplot(fill = "lightyellow") +
  labs(title = "Mal Prem by bed size", x = "Bed size", y = "Malpractice Premiums")

plot4 = ggplot(h_data, aes(x = ownership_type_simple, y = malpractice_premiums_per_inpatient_day)) +
  geom_boxplot(fill = "yellow") +
  labs(title = "Mal Prem by ownership type", x = "Ownership type", y = "Malpractice Premiums")

plot5 = ggplot(h_data, aes(x = overall_hospital_rating_star_rating, y = malpractice_premiums_per_inpa
  geom_boxplot(fill = "green") +
  labs(title = "Mal Prem by hospital rating", x = "hospital rating", y = "Malpractice Premiums")
grid.arrange(plot1, plot2, plot3, plot4, plot5, ncol = 3, nrow = 2)



#Bar Plot
# Change the class of the variable from integer to character
h_data$overall_hospital_rating_star_rating <- as.character(h_data$overall_hospital_rating_star_rating

p1 = h_data %>%
  group_by(bed_size_simple) %>%
  summarise(mean_premiums = mean(malpractice_premiums_per_inpatient_day, na.rm = TRUE)) %>%
  ggplot(aes(x = bed_size_simple, y = mean_premiums)) +
  geom_bar(stat = "identity", fill = "coral") +
  labs(title = "Malpractice by Bed Size", x = "Bed Size", y = "Mean Malpractice Premiums")
p2 = h_data %>%
  group_by(overall_hospital_rating_star_rating) %>%
  summarise(mean_premiums = mean(malpractice_premiums_per_inpatient_day, na.rm = TRUE)) %>%
  ggplot(aes(x = overall_hospital_rating_star_rating, y = mean_premiums)) +
  geom_bar(stat = "identity", fill = "lightyellow") +
  labs(title = "Malpractice by star rating", x = "star rating", y = "Mean Malpractice Premiums")
p3 = h_data %>%
  group_by(ownership_type_simple) %>%
  summarise(mean_premiums = mean(malpractice_premiums_per_inpatient_day, na.rm = TRUE)) %>%
  ggplot(aes(x = ownership_type_simple, y = mean_premiums)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  labs(title = "Malpractice by ownership type", x = "ownership type", y = "Mean Malpractice Premiums"
p4 =  h_data %>%
  group_by(teaching_status) %>%
  summarise(mean_premiums = mean(malpractice_premiums_per_inpatient_day, na.rm = TRUE)) %>%
  ggplot(aes(x = teaching_status, y = mean_premiums)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Malpractice by teaching", x = "teaching Status", y = "Mean Malpractice Premiums")
p5 =  h_data %>%
  group_by(urban_rural) %>%
  summarise(mean_premiums = mean(malpractice_premiums_per_inpatient_day, na.rm = TRUE)) %>%
  ggplot(aes(x = urban_rural, y = mean_premiums)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Malpractice by urban-rural", x = "urban-rural", y = "Mean Malpractice Premiums")
grid.arrange(p1, p2, p3, p4, p5, ncol = 3, nrow = 2)
```

```
# Ensure required libraries are loaded
library(dplyr)
library(ggplot2)
library(tidyr)

# Reshape data to long format using pivot_longer
h_data_long <- h_data %>%
  pivot_longer(cols = c("bed_size_simple", "overall_hospital_rating_star_rating",
                        "ownership_type_simple", "teaching_status", "urban_rural"),
               names_to = "Variable",
               values_to = "Category")

# Create a bar plot with facet_wrap
facet_plot <- ggplot(h_data_long, aes(x = Category, y = malpractice_premiums_per_inpatient_day)) +
  geom_bar(stat = "summary", fun = "mean", fill = "lightblue") +  # Use mean for each category
  facet_wrap(~ Variable, scales = "free_x") +                      # Facet by the categorical variabl
  labs(title = "Malpractice Premiums by Different Categories",
       x = "Category",
       y = "Mean Malpractice Premiums") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabi

# Display the plot
print(facet_plot)

# correlation between the response and all potential variables
vars_with_potential <- h_data %>%
  select(malpractice_premiums_per_inpatient_day,cost_of_care_per_inpatient_day,charity_care_burden,ba
cms_caseload,percent_of_service_area_under_public,
percent_of_service_area_under_private,hospital_wide_readmission_rate,heart_failure_readmission_rate,a
ssi_hyster_rate,mrsa_rate,cdi_rate,PSI_03,PSI_04,PSI_06,PSI_08,PSI_09,PSI_10,PSI_11,PSI_12)
# Compute the correlation matrix
cor_matrix <- cor(vars_with_potential, use = "complete.obs")


# Plot the heatmap
pheatmap(cor_matrix,
         display_numbers = TRUE,  # Show correlation values in cells
         color = colorRampPalette(c("red", "white", "blue"))(100),  # Color scale
         main = "Correlation Heatmap of Continuous Variables with Response Variable",fontsize = ,cell



summary(h_data$malpractice_premiums_per_inpatient_day)
#Distribution Plot: Plot a histogram or density plot to visualize the distribution of malpractice_pre

ggplot(h_data, aes(x = malpractice_premiums_per_inpatient_day)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Malpractice Premiums per Inpatient Day")

ggplot(h_data, aes(y = malpractice_premiums_per_inpatient_day)) +
  geom_boxplot(fill = "skyblue", color = "black") +
  labs(title = "Boxplot of Malpractice Premiums per Inpatient Day",
       y = "Malpractice Premiums per Inpatient Day") +
  theme_minimal()

# Save the plot
```

```
ggsave("malpractice_boxplot.png", width = 8, height = 6)


# Reshape data to long format using tidyr
library(ggplot2)
library(tidyr)

# Gather predictors into a long format
h_data_long <- h_data %>%
  pivot_longer(cols = c("cost_of_care_per_inpatient_day", "salary_cost_per_inpatient_day",
                        "med_records_library_per_inpatient_day", "special_care_day_rate",
                        "charity_care_burden", "bad_debt_burden"),
               names_to = "Predictor",
               values_to = "Value")

# Create scatter plot with facet_wrap
facet_plot <- ggplot(h_data_long, aes(x = Value, y = malpractice_premiums_per_inpatient_day)) +
  geom_point(alpha = 0.6, color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +  # Add trendline
  facet_wrap(~ Predictor, scales = "free_x") +            # Facet by predictor
  labs(title = "Relationships Between Predictors and Malpractice Premiums",
       x = "Predictor Value",
       y = "Malpractice Premiums per Inpatient Day") +
  theme_minimal()

# Display the plot
print(facet_plot)

# Potential variables
selected_vars <- h_data %>%
  select(malpractice_premiums_per_inpatient_day,cost_of_care_per_inpatient_day,charity_care_burden,ba
cms_caseload,percent_of_service_area_under_public,
percent_of_service_area_under_private,hospital_wide_readmission_rate,heart_failure_readmission_rate,a
ssi_hyster_rate,mrsa_rate,cdi_rate,PSI_03,PSI_04,PSI_06,PSI_08,PSI_09,PSI_10,PSI_11,PSI_12)

# Calculate descriptive statistics
descriptive_stats <- selected_vars %>%
  summarise(
    Min = sapply(selected_vars, min, na.rm = TRUE),
    Q1 = sapply(selected_vars, quantile, 0.25, na.rm = TRUE),
    Median = sapply(selected_vars, median, na.rm = TRUE),
    Mean = sapply(selected_vars, mean, na.rm = TRUE),
    Q3 = sapply(selected_vars, quantile, 0.75, na.rm = TRUE),
    Max = sapply(selected_vars, max, na.rm = TRUE),
    SD = sapply(selected_vars, sd, na.rm = TRUE)
  ) %>%

  as.data.frame() %>%
  `rownames<-`(c( "malpractice_premiums_per_inpatient_day","cost_of_care_per_inpatient_day","charity_

# Format and display the table
descriptive_stats %>%
  kbl(caption = "Descriptive Statistics for Selected Variables", digits = 2) %>%
  kable_styling(full_width = FALSE, bootstrap_options = c("striped", "hover"))
# Save as LaTeX table
descriptive_stats %>%
  kbl(format = "latex", caption = "Descriptive Statistics for Selected Variables", digits = 2) %>%
```

```
  save_kable("descriptive_stats.tex")
# Save descriptive statistics as a CSV file
#write.csv(descriptive_stats, "descriptive_stats.csv", row.names = TRUE)


# Select numeric columns
data_numeric <- h_data %>%
  select_if(is.numeric)

# Remove rows with missing values
data_numeric <- na.omit(data_numeric)

# Prepare the data for Lasso regression
# Set 'malpractice_premiums_per_inpatient_day' as the response variable (y) and the rest as predictor
y <- data_numeric$malpractice_premiums_per_inpatient_day
X <- data_numeric %>% select(-malpractice_premiums_per_inpatient_day)

# Fit the Lasso model
lasso_model <- cv.glmnet(as.matrix(X), y, alpha = 1)

# Get the coefficients from the Lasso model
lasso_coefficients <- coef(lasso_model, s = "lambda.min")

# Print the coefficients
print(lasso_coefficients)

# Identify which variables are selected by Lasso (non-zero coefficients)
selected_variables <- rownames(lasso_coefficients)[lasso_coefficients[,1] != 0]
selected_variables

# Load necessary libraries
library(dplyr)
library(tidyr)

# Calculate missing values for each variable
missing_values <- h_data %>%
  summarise_all(~sum(is.na(.))) %>%
  pivot_longer(cols = everything(),
               names_to = "Variable",
               values_to = "Missing Values")

# View the missing values table
missing_values


# Clean the Data
# List of variables selected by the Lasso model
selected_variables <- c("bad_debt_burden" ,"salary_cost_per_inpatient_day" ,"alos", "private_room_per

# Select only the relevant columns from your data
data_selected <- h_data %>%select(all_of(selected_variables))
# Check for missing values
sum(is.na(h_data))

# Impute missing values
library(mice)
set.seed(123)  # For reproducibility
```

```
imputed_data <- mice(data_selected, m = 1, method = 'pmm', maxit = 5)
data_cleaned <- complete(imputed_data)

# Verify that there are no missing values
sum(is.na(data_cleaned))


#  Split the Data


# Set seed for reproducibility
set.seed(123)

# Split data into training (70%) and testing (30%) sets
trainIndex <- createDataPartition(data_cleaned$malpractice_premiums_per_inpatient_day, p = 0.8, list
train_data <- data_cleaned[trainIndex, ]
test_data <- data_cleaned[-trainIndex, ]


# Load necessary library
library(rpart)
library(rpart.plot)

# Train a Decision Tree
decision_tree_model <- rpart(malpractice_premiums_per_inpatient_day ~ ., data = train_data, method =

# Plot the tree
rpart.plot(decision_tree_model)

# Predict and evaluate
predictions_tree <- predict(decision_tree_model, newdata = test_data)
RMSE_tree <- sqrt(mean((test_data$malpractice_premiums_per_inpatient_day - predictions_tree)^2))
print(paste("Decision Tree RMSE:", RMSE_tree))


# Load necessary library
library(randomForest)

# Train a Random Forest
set.seed(123)
rf_model <- randomForest(malpractice_premiums_per_inpatient_day ~ ., data = train_data, importance =

# Plot feature importance
varImpPlot(rf_model)

# Predict and evaluate
predictions_rf <- predict(rf_model, newdata = test_data)
RMSE_rf <- sqrt(mean((test_data$malpractice_premiums_per_inpatient_day - predictions_rf)^2))
print(paste("Random Forest RMSE:", RMSE_rf))

# Load necessary library
library(neuralnet)

# Normalize data
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}
data_normalized <- as.data.frame(lapply(data_cleaned, normalize))
```

```
# Split normalized data
train_data_norm <- data_normalized[trainIndex, ]
test_data_norm <- data_normalized[-trainIndex, ]

# Train Neural Network
nn_model <- neuralnet(malpractice_premiums_per_inpatient_day ~ ., data = train_data_norm, hidden = c(

# Plot the neural network
plot(nn_model)

# Predict and evaluate
predictions_nn <- compute(nn_model, test_data_norm[, -1])$net.result
predictions_nn <- predictions_nn * (max(data_cleaned$malpractice_premiums_per_inpatient_day) - min(da
RMSE_nn <- sqrt(mean((test_data$malpractice_premiums_per_inpatient_day - predictions_nn)^2))
print(paste("Neural Network RMSE:", RMSE_nn))

# Compare models
print(paste("Decision Tree RMSE:", RMSE_tree))
print(paste("Random Forest RMSE:", RMSE_rf))
print(paste("Neural Network RMSE:", RMSE_nn))
```