CMS Hospital Rating Prediction (3-Class)

📦 Load Libraries

```r
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```r
library(readr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(randomForest)
```

```
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(smotefamily)
library(DALEX)
```

```
## Welcome to DALEX (version: 2.4.3).
## Find examples and detailed introduction at: http://ema.drwhy.ai/
## Additional features will be available after installation of: ggpubr.
## Use 'install_dependencies()' to get all suggested dependencies
##
## Attaching package: 'DALEX'
##
## The following object is masked from 'package:dplyr':
##
##     explain
```

```r
library(ggplot2)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
library(DT)
```

📥 Load and Clean Data

```
hospital_data <- read_csv("Hospital_General_Information.csv") %>%
  clean_names() %>%
  filter(hospital_type == "Acute Care Hospitals") %>%
  mutate(
    hospital_overall_rating = na_if(hospital_overall_rating, "Not Available"),
    hospital_overall_rating = as.numeric(hospital_overall_rating),
    rating_group = case_when(
      hospital_overall_rating %in% c(1, 2) ~ "Low",
      hospital_overall_rating == 3 ~ "Medium",
      hospital_overall_rating %in% c(4, 5) ~ "High"
    ),
    rating_group = as.factor(rating_group)
  ) %>%
  drop_na(rating_group)
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 5384 Columns: 38
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (32): Facility ID, Facility Name, Address, City/Town, State, ZIP Code, C...
## dbl  (6): Hospital overall rating footnote, MORT Group Footnote, Safety Grou...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Summary Statistics

```
glimpse(hospital_data)
```

```
## Rows: 2,537
## Columns: 39
## $ facility_id                                      <chr> "010001", "010005", "…
## $ facility_name                                    <chr> "SOUTHEAST HEALTH MED…
## $ address                                          <chr> "1108 ROSS CLARK CIRC…
## $ city_town                                        <chr> "DOTHAN", "BOAZ", "FL…
## $ state                                            <chr> "AL", "AL", "AL", "AL…
## $ zip_code                                         <chr> "36301", "35957", "35…
## $ county_parish                                    <chr> "HOUSTON", "MARSHALL"…
## $ telephone_number                                 <chr> "(334) 793-8701", "(2…
## $ hospital_type                                    <chr> "Acute Care Hospitals…
## $ hospital_ownership                               <chr> "Government - Hospita…
## $ emergency_services                               <chr> "Yes", "Yes", "Yes", …
## $ meets_criteria_for_birthing_friendly_designation <chr> "Y", NA, "Y", NA, NA,…
## $ hospital_overall_rating                          <dbl> 3, 2, 1, 1, 3, 2, 3, …
## $ hospital_overall_rating_footnote                 <dbl> NA, NA, NA, NA, NA, N…
## $ mort_group_measure_count                         <chr> "7", "7", "7", "7", "…
## $ count_of_facility_mort_measures                  <chr> "7", "6", "7", "3", "…
## $ count_of_mort_measures_better                    <chr> "1", "0", "0", "0", "…
## $ count_of_mort_measures_no_different              <chr> "6", "5", "6", "2", "…
## $ count_of_mort_measures_worse                     <chr> "0", "1", "1", "1", "…
## $ mort_group_footnote                              <dbl> NA, NA, NA, NA, NA, N…
## $ safety_group_measure_count                       <chr> "8", "8", "8", "8", "…
## $ count_of_facility_safety_measures                <chr> "7", "7", "7", "2", "…
## $ count_of_safety_measures_better                  <chr> "2", "0", "3", "0", "…
## $ count_of_safety_measures_no_different            <chr> "5", "7", "4", "2", "…
## $ count_of_safety_measures_worse                   <chr> "0", "0", "0", "0", "…
## $ safety_group_footnote                            <dbl> NA, NA, NA, NA, NA, N…
## $ readm_group_measure_count                        <chr> "11", "11", "11", "11…
## $ count_of_facility_readm_measures                 <chr> "11", "9", "9", "7", …
## $ count_of_readm_measures_better                   <chr> "1", "0", "0", "0", "…
## $ count_of_readm_measures_no_different             <chr> "8", "8", "7", "7", "…
## $ count_of_readm_measures_worse                    <chr> "2", "1", "2", "0", "…
## $ readm_group_footnote                             <dbl> NA, NA, NA, NA, NA, N…
## $ pt_exp_group_measure_count                       <chr> "8", "8", "8", "8", "…
## $ count_of_facility_pt_exp_measures                <chr> "8", "8", "8", "8", "…
## $ pt_exp_group_footnote                            <dbl> NA, NA, NA, NA, NA, N…
## $ te_group_measure_count                           <chr> "12", "12", "12", "12…
## $ count_of_facility_te_measures                    <chr> "10", "12", "11", "7"…
## $ te_group_footnote                                <dbl> NA, NA, NA, NA, NA, N…
## $ rating_group                                     <fct> Medium, Low, Low, Low…
```

```
summary(hospital_data)
```

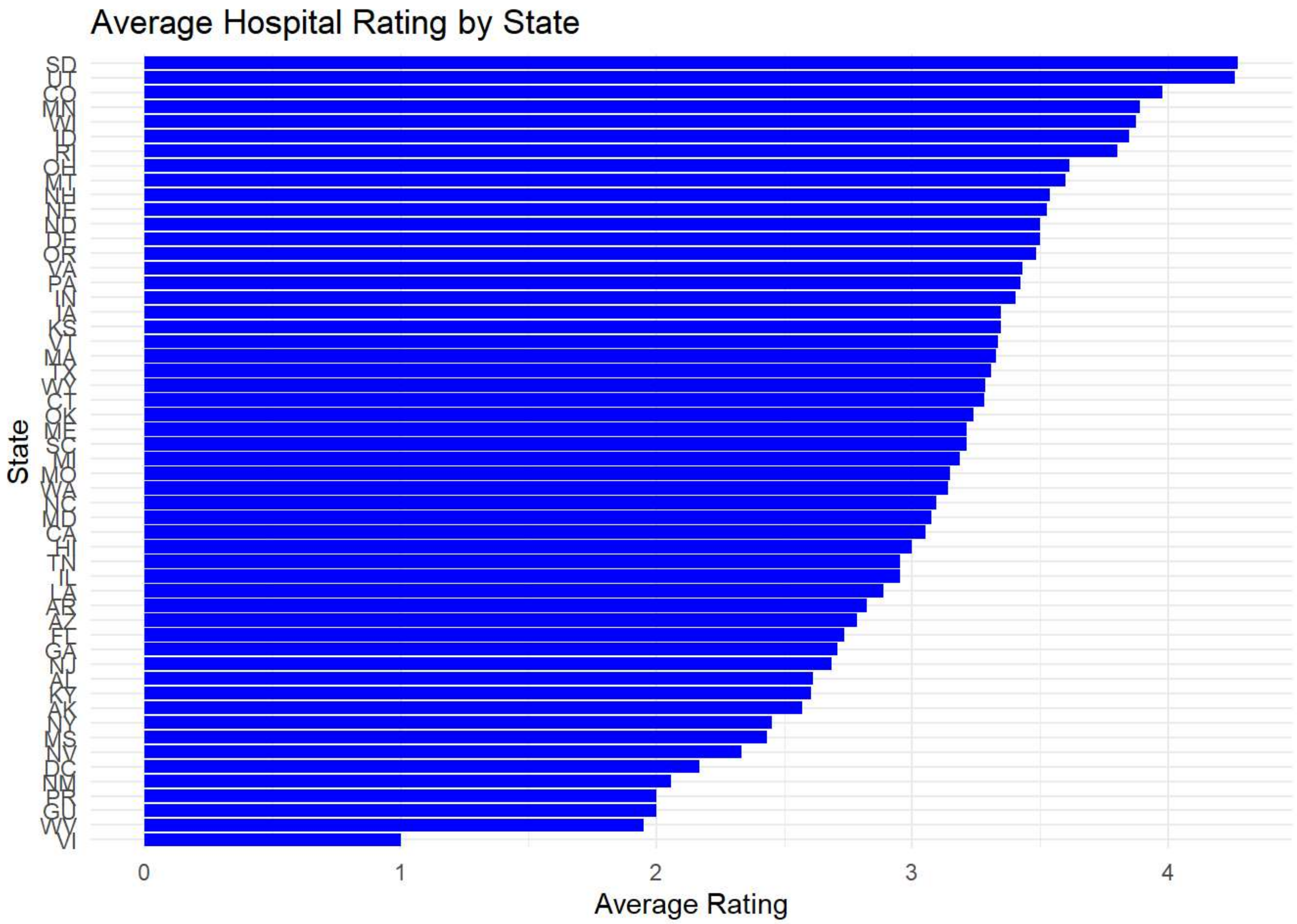
```
##  facility_id        facility_name        address            city_town
##  Length:2537        Length:2537        Length:2537        Length:2537
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     state             zip_code         county_parish      telephone_number
##  Length:2537        Length:2537        Length:2537        Length:2537
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  hospital_type      hospital_ownership emergency_services
##  Length:2537        Length:2537        Length:2537
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  meets_criteria_for_birthing_friendly_designation hospital_overall_rating
##  Length:2537                                      Min.   :1.000
##  Class :character                                 1st Qu.:2.000
##  Mode  :character                                 Median :3.000
##                                                   Mean   :3.106
##                                                   3rd Qu.:4.000
##                                                   Max.   :5.000
##
##  hospital_overall_rating_footnote mort_group_measure_count
##  Min.   :17.00                    Length:2537
##  1st Qu.:17.00                    Class :character
##  Median :17.00                    Mode  :character
##  Mean   :17.78
##  3rd Qu.:17.00
##  Max.   :23.00
##  NA's   :2491
##  count_of_facility_mort_measures count_of_mort_measures_better
##  Length:2537                     Length:2537
##  Class :character                Class :character
##  Mode  :character                Mode  :character
##
##
##
##
##  count_of_mort_measures_no_different count_of_mort_measures_worse
##  Length:2537                         Length:2537
##  Class :character                    Class :character
##  Mode  :character                    Mode  :character
```

```
##
##
##
##
##  mort_group_footnote safety_group_measure_count
##  Min.   : 5.000      Length:2537
##  1st Qu.: 5.000      Class :character
##  Median : 5.000      Mode  :character
##  Mean   : 5.806
##  3rd Qu.: 5.000
##  Max.   :23.000
##  NA's   :2470
##  count_of_facility_safety_measures count_of_safety_measures_better
##  Length:2537                       Length:2537
##  Class :character                  Class :character
##  Mode  :character                  Mode  :character
##
##
##
##
##  count_of_safety_measures_no_different count_of_safety_measures_worse
##  Length:2537                           Length:2537
##  Class :character                      Class :character
##  Mode  :character                      Mode  :character
##
##
##
##
##  safety_group_footnote readm_group_measure_count
##  Min.   : 5            Length:2537
##  1st Qu.: 5            Class :character
##  Median : 5            Mode  :character
##  Mean   :11
##  3rd Qu.:23
##  Max.   :23
##  NA's   :2525
##  count_of_facility_readm_measures count_of_readm_measures_better
##  Length:2537                      Length:2537
##  Class :character                 Class :character
##  Mode  :character                 Mode  :character
##
##
##
##
##  count_of_readm_measures_no_different count_of_readm_measures_worse
##  Length:2537                          Length:2537
##  Class :character                     Class :character
##  Mode  :character                     Mode  :character
##
##
##
##
```

```
##  readm_group_footnote pt_exp_group_measure_count
##  Min.   :5            Length:2537
##  1st Qu.:5            Class :character
##  Median :5            Mode  :character
##  Mean   :5
##  3rd Qu.:5
##  Max.   :5
##  NA's   :2536
##  count_of_facility_pt_exp_measures pt_exp_group_footnote te_group_measure_count
##  Length:2537                       Min.   :5             Length:2537
##  Class :character                  1st Qu.:5             Class :character
##  Mode  :character                  Median :5             Mode  :character
##                                    Mean   :5
##                                    3rd Qu.:5
##                                    Max.   :5
##                                    NA's   :2506
##  count_of_facility_te_measures te_group_footnote rating_group
##  Length:2537                   Min.   : NA       High  :995
##  Class :character              1st Qu.: NA       Low   :790
##  Mode  :character              Median : NA       Medium:752
##                                Mean   :NaN
##                                3rd Qu.: NA
##                                Max.   : NA
##                                NA's   :2537
```

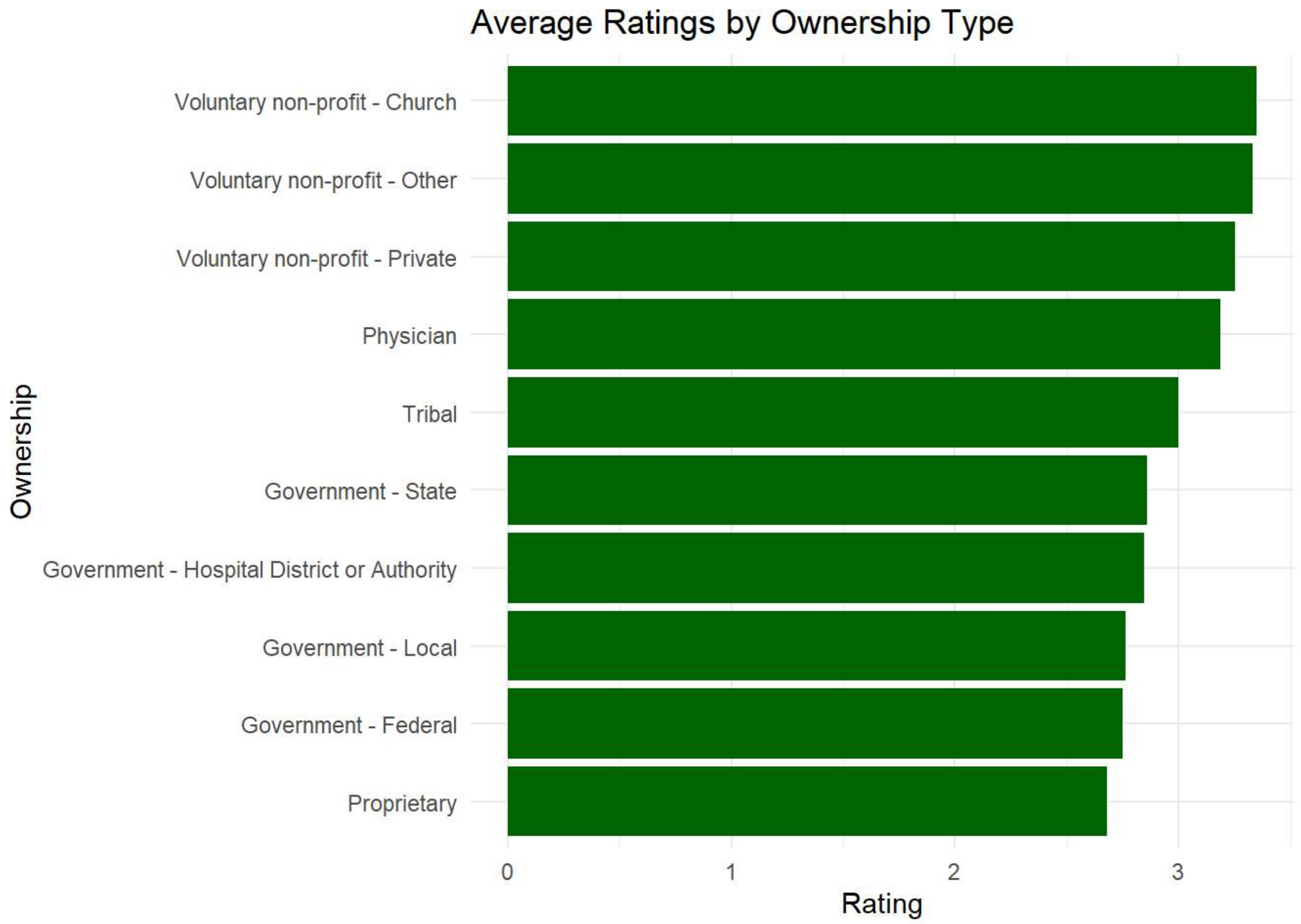
Average Rating by State

```
avg_rating_by_state <- hospital_data %>%
  group_by(state) %>%
  summarise(avg_rating = mean(hospital_overall_rating, na.rm = TRUE),
            count = n()) %>%
  arrange(desc(avg_rating))

ggplot(avg_rating_by_state, aes(x = reorder(state, avg_rating), y = avg_rating)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip() +
  labs(title = "Average Hospital Rating by State", x = "State", y = "Average Rating") +
  theme_minimal()
```



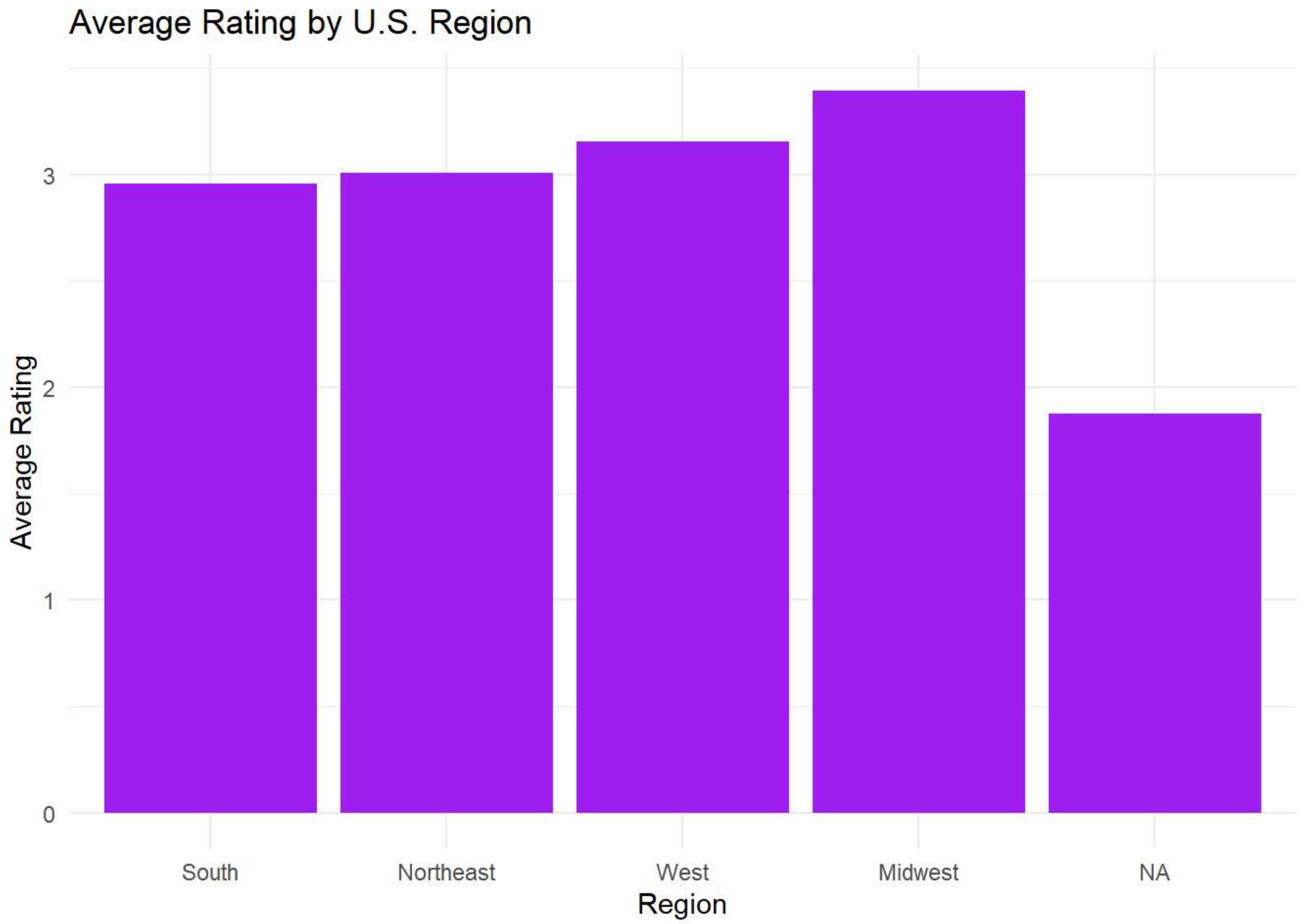
Ratings by Ownership

```
ownership_ratings <- hospital_data %>%
  filter(!is.na(hospital_overall_rating)) %>%
  group_by(hospital_ownership) %>%
  summarise(avg_rating = mean(hospital_overall_rating, na.rm = TRUE),
            count = n()) %>%
  arrange(desc(avg_rating))

ggplot(ownership_ratings, aes(x = reorder(hospital_ownership, avg_rating), y = avg_rating)) +
  geom_bar(stat = "identity", fill = "darkgreen") +
  coord_flip() +
  labs(title = "Average Ratings by Ownership Type", x = "Ownership", y = "Rating") +
  theme_minimal()
```


Average Ratings by Ownership Type
Ownership
Voluntary non-profit - Church
Voluntary non-profit - Other
Voluntary non-profit - Private
Physician
Tribal
Government - State
Government - Hospital District or Authority
Government - Local
Government - Federal
Proprietary
0
1
2
3
Rating

Regional Analysis

```r
state_region_map <- list(
  Northeast = c("CT", "ME", "MA", "NH", "RI", "VT", "NJ", "NY", "PA"),
  Midwest = c("IL", "IN", "IA", "KS", "MI", "MN", "MO", "NE", "ND", "OH", "SD", "WI"),
  South = c("DE", "FL", "GA", "MD", "NC", "SC", "VA", "DC", "WV", "AL", "KY", "MS", "TN", "AR",
"LA", "OK", "TX"),
  West = c("AZ", "CO", "ID", "MT", "NV", "NM", "UT", "WY", "AK", "CA", "HI", "OR", "WA")
)

get_region <- function(state) {
  for (r in names(state_region_map)) {
    if (state %in% state_region_map[[r]]) return(r)
  }
  return(NA)
}

hospital_data$region <- sapply(hospital_data$state, get_region)

region_avg <- hospital_data %>%
  group_by(region) %>%
  summarise(avg_rating = mean(hospital_overall_rating, na.rm = TRUE))

ggplot(region_avg, aes(x = reorder(region, avg_rating), y = avg_rating)) +
  geom_bar(stat = "identity", fill = "purple") +
  labs(title = "Average Rating by U.S. Region", x = "Region", y = "Average Rating") +
  theme_minimal()
```

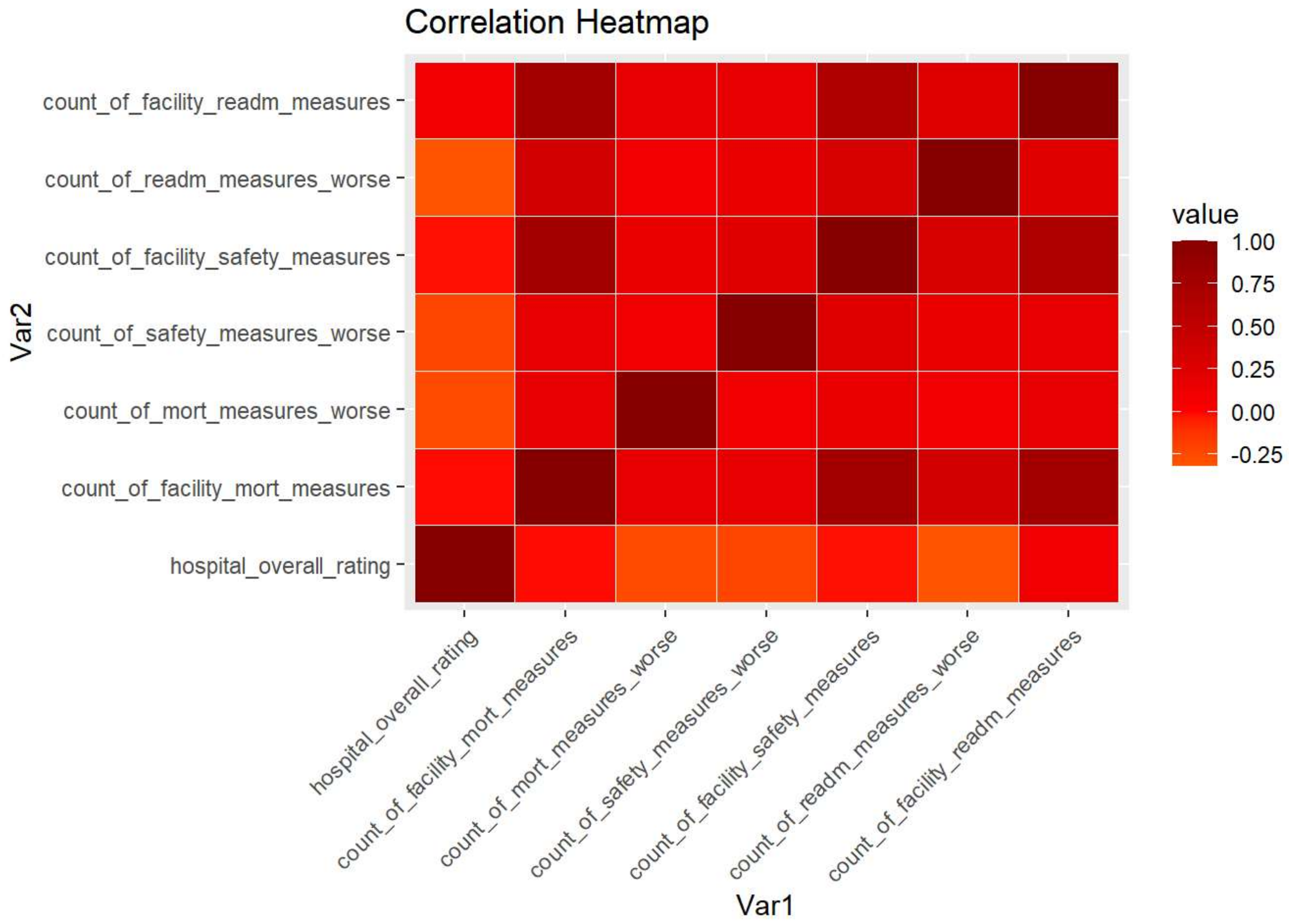


Correlation Heatmap

```
corr_data <- hospital_data %>%
  select(
    hospital_overall_rating,
    count_of_facility_mort_measures,
    count_of_mort_measures_worse,
    count_of_safety_measures_worse,
    count_of_facility_safety_measures,
    count_of_readm_measures_worse,
    count_of_facility_readm_measures
  ) %>%
  mutate(across(everything(), as.numeric)) %>%
  drop_na()
```

```
## Warning: There were 6 warnings in `mutate()`.
## The first warning was:
## ℹ In argument: `across(everything(), as.numeric)`.
## Caused by warning:
## ! NAs introduced by coercion
## ℹ Run `dplyr::last_dplyr_warnings()` to see the 5 remaining warnings.
```

```
corr_matrix <- cor(corr_data)
melted_corr <- melt(corr_matrix)

ggplot(melted_corr, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "orange", high = "darkred", mid = "red", midpoint = 0) +
  labs(title = "Correlation Heatmap") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Top Rated Hospitals

```
top_hospitals <- hospital_data %>%
  filter(hospital_overall_rating == 5) %>%
  group_by(hospital_ownership) %>%
  slice_max(order_by = hospital_overall_rating, n = 5, with_ties = FALSE) %>%
  select(facility_name, state, hospital_ownership, hospital_overall_rating)

datatable(top_hospitals, caption = "Top 5 Hospitals by Ownership Type")
```

Show 10 **entries** **Search:**

Top 5 Hospitals by Ownership Type

	facility_name	state	hospital_ownership	hospital_overall_rating
1	UNIVERSITY OF COLORADO HOSPITAL AUTHORITY	CO	Government - Hospital District or Authority	5
2	LEE MEMORIAL HOSPITAL	FL	Government - Hospital District or Authority	5
3	SARASOTA MEMORIAL HOSPITAL	FL	Government - Hospital District or Authority	5
4	UNIVERSITY OF KANSAS HOSPITAL	KS	Government - Hospital District or Authority	5
5	NORTHERN REGIONAL HOSPITAL	NC	Government - Hospital District or Authority	5
6	MADISON MEMORIAL HOSPITAL	ID	Government - Local	5
7	HENDRICKS REGIONAL HEALTH	IN	Government - Local	5
8	SCHNECK MEDICAL CENTER	IN	Government - Local	5
9	ST ELIZABETH DEARBORN HOSPITAL	IN	Government - Local	5
10	SPENCER MUNICIPAL HOSPITAL	IA	Government - Local	5

Showing 1 to 10 of 40 entries

Previous 1 2 3 4 Next

🧠 Feature Engineering

```
model_data <- hospital_data %>%
  mutate(
    readm_safety_gap = as.numeric(count_of_readm_measures_worse) - as.numeric(count_of_safety_measures_worse),
    mort_ratio = as.numeric(count_of_mort_measures_worse) / (as.numeric(count_of_facility_mort_measures) + 1),
    safety_ratio = as.numeric(count_of_safety_measures_worse) / (as.numeric(count_of_facility_safety_measures) + 1)
  ) %>%
  select(rating_group, hospital_ownership, state, readm_safety_gap, mort_ratio, safety_ratio) %>%
  mutate(across(c(hospital_ownership, state), as.factor)) %>%
  drop_na()
```

```
## Warning: There were 6 warnings in `mutate()`.
## The first warning was:
## ℹ In argument: `readm_safety_gap = as.numeric(count_of_readm_measures_worse) -
##   as.numeric(count_of_safety_measures_worse)`.
## Caused by warning:
## ! NAs introduced by coercion
## ℹ Run `dplyr::last_dplyr_warnings()` to see the 5 remaining warnings.
```

🔀 Train-Test Split

```
set.seed(123)
train_idx <- createDataPartition(model_data$rating_group, p = 0.8, list = FALSE)
train <- model_data[train_idx, ]
test <- model_data[-train_idx, ]
```

⚖️ Balance Classes with SMOTE

```
X <- train %>% select(-rating_group)
y <- train$rating_group

X_numeric <- data.frame(model.matrix(~ . - 1, data = X))
smote_result <- SMOTE(X_numeric, y, K = 5)

balanced_train <- smote_result$data
balanced_train$rating_group <- as.factor(smote_result$data$class)
balanced_train$class <- NULL
```

🧪 Train Random Forest Classifier

```
model <- train(rating_group ~ ., data = balanced_train, method = "rf")
```

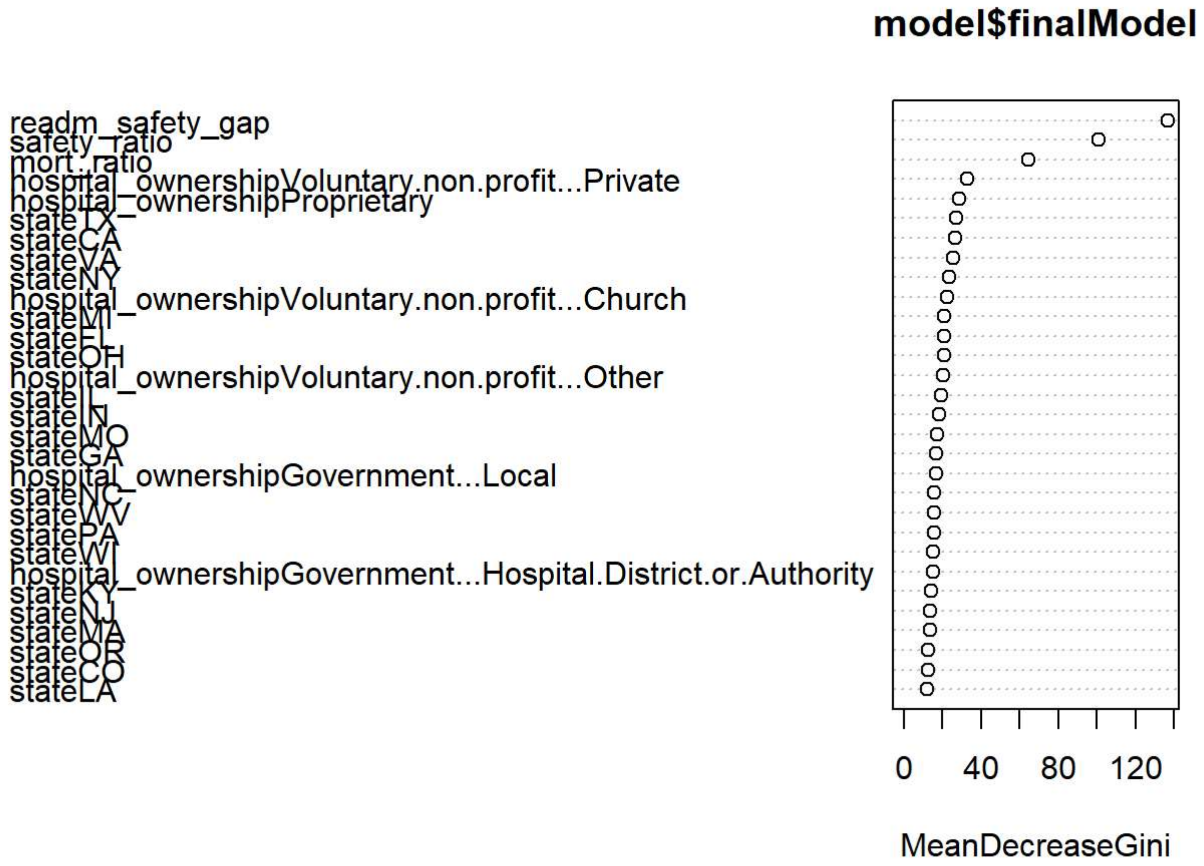
Evaluation

```
# Preprocess test
X_test <- data.frame(model.matrix(~ . - 1, data = test %>% select(-rating_group)))
y_test <- test$rating_group

# Align test to training columns
missing_cols <- setdiff(colnames(balanced_train)[-ncol(balanced_train)], colnames(X_test))
X_test[missing_cols] <- 0
X_test <- X_test[, colnames(balanced_train)[-ncol(balanced_train)]]

# Predict
preds <- predict(model, newdata = X_test)

# Confusion matrix
confusionMatrix(preds, y_test)
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction High Low Medium
##     High    106  33     50
##     Low      22  76     41
##     Medium   63  46     55
## 
## Overall Statistics
## 
##                Accuracy : 0.4817
##                  95% CI : (0.4368, 0.5269)
##     No Information Rate : 0.3882
##     P-Value [Acc > NIR] : 1.559e-05
## 
##                   Kappa : 0.2182
## 
##  Mcnemar's Test P-Value : 0.2633
## 
## Statistics by Class:
## 
##                      Class: High Class: Low Class: Medium
## Sensitivity               0.5550     0.4903        0.3767
## Specificity               0.7243     0.8131        0.6850
## Pos Pred Value            0.5608     0.5468        0.3354
## Neg Pred Value            0.7195     0.7762        0.7226
## Prevalence                0.3882     0.3150        0.2967
## Detection Rate            0.2154     0.1545        0.1118
## Detection Prevalence      0.3841     0.2825        0.3333
## Balanced Accuracy         0.6396     0.6517        0.5308
```

🔍 Feature Importance

```
varImpPlot(model$finalModel)
```

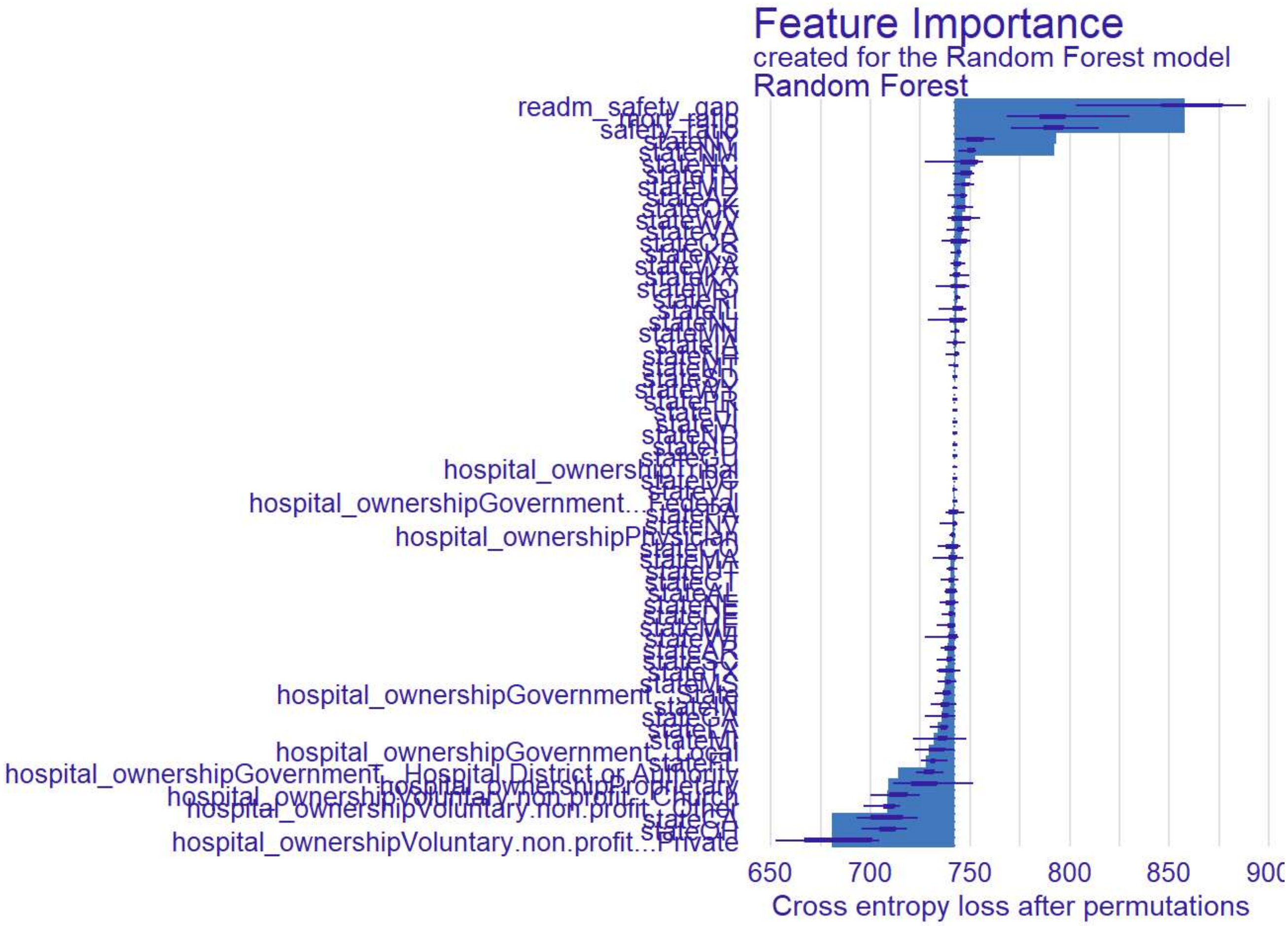


📊 SHAP/DALEX Model Explanation

```
explainer <- explain(model$finalModel, data = X_test, y = y_test, label = "Random Forest")
```

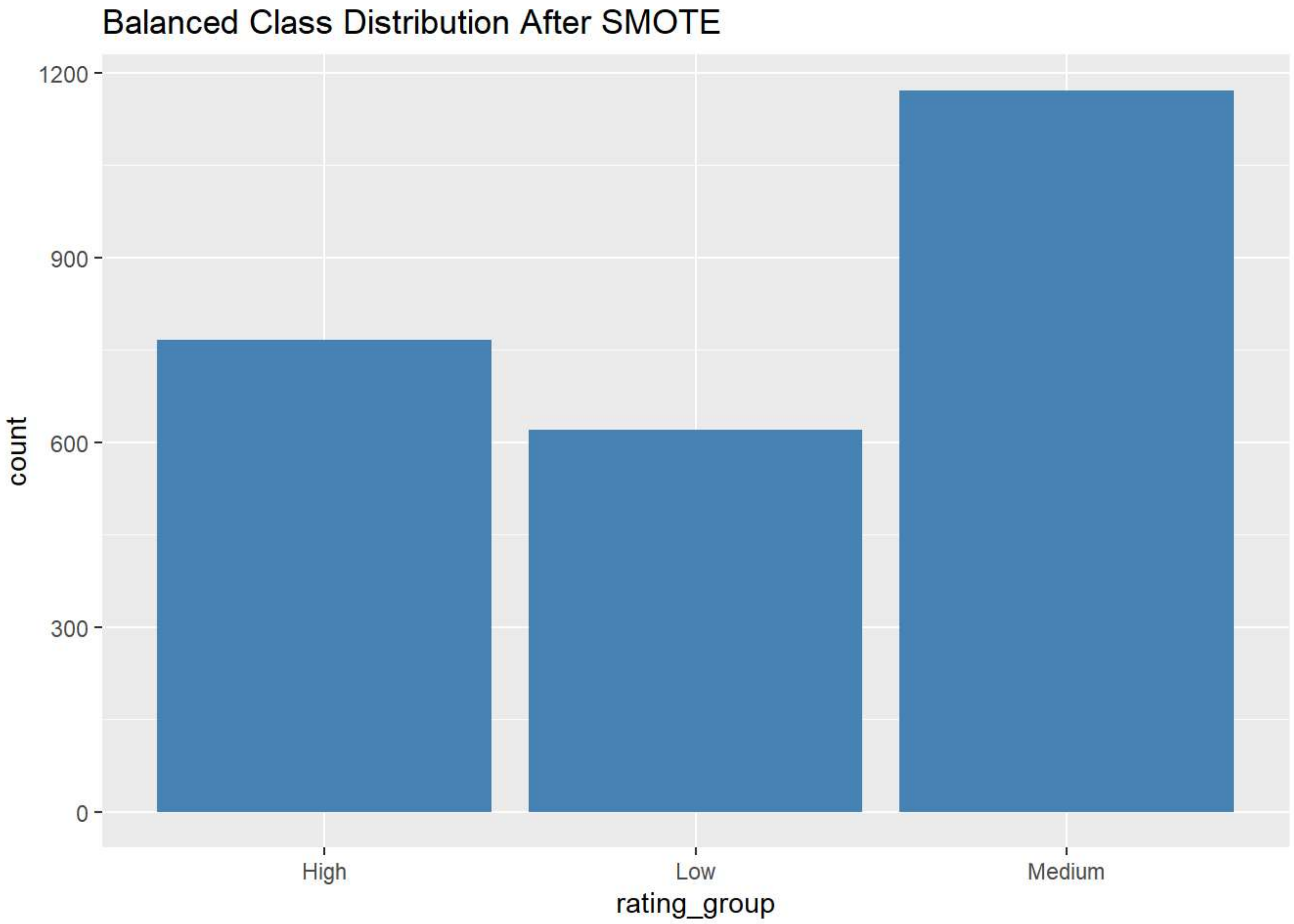
```
## Preparation of a new explainer is initiated
##   -> model label       :  Random Forest
##   -> data              :  492  rows  66  cols
##   -> target variable   :  492  values
##   -> predict function  :  yhat.randomForest  will be used (  default  )
##   -> predicted values  :  No value for predict function target column. (  default  )
##   -> model_info        :  package randomForest , ver. 4.7.1.2 , task multiclass (  default  )
##   -> predicted values  :  predict function returns multiple columns:  3  (  default  )
##   -> residual function :  difference between 1 and probability of true class (  default  )
##   -> residuals         :  numerical, min =  0 , mean =  0.5425528 , max =  1
##   A new explainer has been created!
```

```
model_parts(explainer) %>% plot()
```



📈 Class Distribution After SMOTE

```
ggplot(balanced_train, aes(x = rating_group)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Balanced Class Distribution After SMOTE")
```



✅ Summary

This analysis used CMS hospital data and predicted 3-level quality ratings using engineered features and SMOTE-balancing. Accuracy and feature interpretation help identify key drivers of hospital performance.