Predicting Length of Stay - Updated

Angelina Cottone

2026-01-15

1 Introduction

Understanding the factors that influence hospital length of stay (LOS) is crucial for improving patient outcomes and optimizing resource allocation in healthcare settings. Extended hospital stays increase the risk of complications and hospital-acquired infections while placing financial strain on both facilities and patients. Identifying key predictors of LOS can help clinicians enhance quality of care and improve operational efficiency.

Multiple factors may influence hospital length of stay, including patient demographics, comorbidities, and laboratory results. This analysis seeks to answer the primary research question: *What are the key factors that influence the length of hospital stays for patients?*

To address this question, the following sub-questions guide the analysis:

- How do comorbidities, such as asthma, iron deficiency, and renal disease, impact length of stay?
- What role do mental health conditions play in determining length of stay?
- How do laboratory values, such as hematocrit, neutrophil levels, and blood urea nitrogen, influence hospital length of stay?

By addressing these questions, this study aims to identify the most influential factors in predicting LOS and provide actionable insights to improve patient outcomes and optimize healthcare practices through data-driven approaches.

2 Data Acquisition & Processing

2.1 Data Overview

The dataset (`LengthOfStay.csv`) contains 100,000 patient records across 28 columns, including information on comorbidities, laboratory results, and vital signs. The dataset encompasses a variety of patient characteristics, including demographics, health conditions, laboratory values, and vital signs. Variables irrelevant to this analysis—such as date columns, patient IDs, and facility IDs—were excluded, leaving 24 variables for modeling.

2.2 Data Preprocessing

A check for missing data confirmed that there were no missing values in the dataset, eliminating the need for data-cleaning techniques or imputations. An initial summary of the dataset reveals a mix of variable types:

- **Binary variables**: Indicators for health conditions (e.g., `asthma`, `pneum`, `malnutrition`) and mental health conditions (e.g., `depress`, `psychologicaldisordermajor`)

- **Continuous variables**: Laboratory results (e.g., `hematocrit`, `neutrophils`) and vital signs (e.g., `pulse`, `respiration`)
- **Categorical variables**: Demographic information (e.g., `gender`, `rcount`)
- **Response variable**: `lengthofstay` is a discrete count variable representing the number of days a patient remains hospitalized

To determine the appropriate modeling approach, the mean and variance of `lengthofstay` were examined for overdispersion. The variance (5.571) substantially exceeds the mean (4.001), indicating overdispersion. This suggests that a Negative Binomial model is more appropriate than a standard Poisson model, as it accounts for overdispersion through an additional dispersion parameter. The Negative Binomial distribution models the variance independently from the mean, providing more accurate parameter estimates and prediction intervals compared to Quasi-Poisson models.

The dataset contains no missing values. The mean length of stay is 4.001 days with a variance of 5.571, confirming substantial overdispersion (variance » mean).

2.3 Feature Engineering

Feature engineering was performed on selected variables to improve model performance and interpretability:

- **Readmission count** (`rcount`): Converted to a binary variable where 0 indicates no readmissions in the past 180 days and 1 indicates one or more readmissions
- **Gender**: Converted to binary encoding with 1 for male and 0 for female
- **Secondary diagnoses** (`secondarydiagnosisnonicd9`): Rare categories (all categories except `1`) were combined into an "Other" category to simplify the analysis

Outlier prevalence in continuous variables was assessed using the interquartile range (IQR) method, where values beyond 1.5 times the IQR from the first and third quartiles were flagged as potential outliers. This analysis revealed that 50.79% of the dataset contains values that would be classified as outliers. However, in the healthcare context, these extreme values likely represent legitimate real-world variation, such as severe cases or complex conditions. Therefore, outliers were retained to preserve the natural variability of the dataset and avoid artificially constraining the model.

```
## [1] 50793
```

3 Exploratory Data Analysis (EDA)

3.1 Summary Statistics

The exploratory data analysis began by examining key characteristics of the dataset using descriptive statistics, including mean, standard deviation, skewness, and range.

Key insights from the summary statistics:

- **Strong positive skewness**: `neutrophils` and `bloodureanitro` exhibit strong positive skewness (values > 1), indicating that most observations are clustered toward lower values with a long right tail
- **Moderate positive skewness**: `hematocrit` and `lengthofstay` show moderate positive skewness (values > 0.5), suggesting a concentration of lower values but less extreme than the strongly skewed variables

- **Negative skewness**: `respiration` displays moderate negative skewness (value < -0.5), indicating that most values are concentrated toward the higher end of the distribution
- **High variability**: `glucose` has the highest standard deviation among all variables, indicating greater variability in blood glucose measurements across patients
- **Wide range**: `bloodureanitro` has the widest range (681.50), reflecting substantial variation in kidney function across the patient population

```
##                            vars      n   mean    sd median trimmed   mad    min
## rcount*                       1 100000   2.12  1.54   1.00    1.84  0.00   1.00
## gender*                       2 100000   1.42  0.49   1.00    1.40  0.00   1.00
## dialysisrenalendstage         3 100000   0.04  0.19   0.00    0.00  0.00   0.00
## asthma                        4 100000   0.04  0.18   0.00    0.00  0.00   0.00
## irondef                       5 100000   0.09  0.29   0.00    0.00  0.00   0.00
## pneum                         6 100000   0.04  0.19   0.00    0.00  0.00   0.00
## substancedependence           7 100000   0.06  0.24   0.00    0.00  0.00   0.00
## psychologicaldisordermajor    8 100000   0.24  0.43   0.00    0.17  0.00   0.00
## depress                       9 100000   0.05  0.22   0.00    0.00  0.00   0.00
## psychother                   10 100000   0.05  0.22   0.00    0.00  0.00   0.00
## fibrosisandother             11 100000   0.00  0.07   0.00    0.00  0.00   0.00
## malnutrition                 12 100000   0.05  0.22   0.00    0.00  0.00   0.00
## hemo                         13 100000   0.08  0.27   0.00    0.00  0.00   0.00
## hematocrit                   14 100000  11.98  2.03  11.90   11.91  1.48   4.40
## neutrophils                  15 100000  10.18  5.35   9.40    9.63  2.67   0.10
## sodium                       16 100000 137.89  3.00 137.89  137.89  3.00 124.91
## glucose                      17 100000 141.96 29.99 142.09  141.99 30.01  -1.01
## bloodureanitro               18 100000  14.10 12.95  12.00   12.43  1.48   1.00
## creatinine                   19 100000   1.10  0.20   1.10    1.10  0.20   0.22
## bmi                          20 100000  29.81  2.00  29.81   29.80  2.00  21.99
## pulse                        21 100000  73.44 11.64  73.00   73.45 11.86  21.00
## respiration                  22 100000   6.49  0.57   6.50    6.50  0.00   0.20
## secondarydiagnosisnonicd9    23 100000   2.12  2.05   1.00    1.79  0.00   0.00
## lengthofstay                 24 100000   4.00  2.36   4.00    3.81  2.97   1.00
## rcount_binary                25 100000   0.45  0.50   0.00    0.44  0.00   0.00
## gender_binary                26 100000   0.42  0.49   0.00    0.40  0.00   0.00
## secondarydx_recategorized*   27 100000   1.50  0.50   1.00    1.50  0.00   1.00
##                               max  range  skew kurtosis   se
## rcount*                      6.00   5.00  1.20     0.18 0.00
## gender*                      2.00   1.00  0.31    -1.90 0.00
## dialysisrenalendstage        1.00   1.00  4.95    22.49 0.00
## asthma                       1.00   1.00  5.04    23.39 0.00
## irondef                      1.00   1.00  2.76     5.64 0.00
## pneum                        1.00   1.00  4.73    20.39 0.00
## substancedependence          1.00   1.00  3.60    10.92 0.00
## psychologicaldisordermajor   1.00   1.00  1.22    -0.50 0.00
## depress                      1.00   1.00  4.05    14.41 0.00
## psychother                   1.00   1.00  4.16    15.30 0.00
## fibrosisandother             1.00   1.00 14.34   203.77 0.00
## malnutrition                 1.00   1.00  4.15    15.26 0.00
## hemo                         1.00   1.00  3.10     7.59 0.00
## hematocrit                  24.10  19.70  0.56     1.82 0.01
## neutrophils                245.90 245.80 13.28   428.41 0.02
## sodium                     151.39  26.47  0.00     0.02 0.01
## glucose                    271.44 272.45 -0.01     0.01 0.09
```

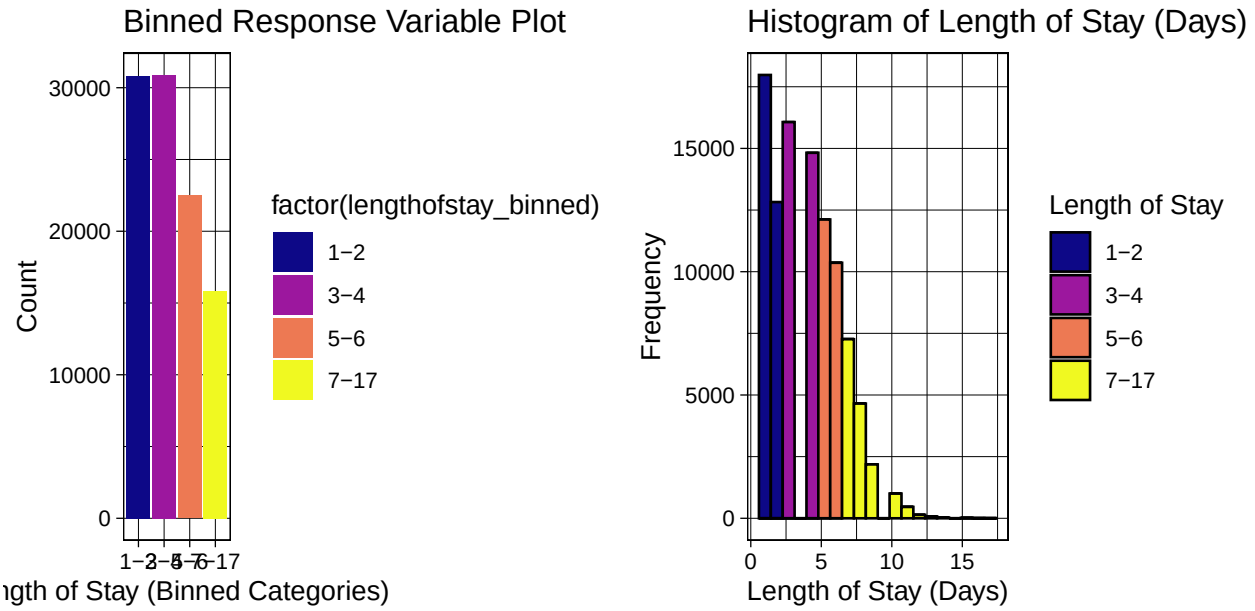
```
## bloodureanitro             682.50 681.50 17.29   528.52 0.04
## creatinine                   2.04   1.82  0.00    -0.01 0.00
## bmi                         38.94  16.94  0.01     0.00 0.01
## pulse                      130.00 109.00  0.00     0.03 0.04
## respiration                 10.00   9.80 -0.55     8.30 0.00
## secondarydiagnosisnonicd9   10.00  10.00  1.67     2.63 0.01
## lengthofstay                17.00  16.00  0.63    -0.03 0.01
## rcount_binary                1.00   1.00  0.20    -1.96 0.00
## gender_binary                1.00   1.00  0.31    -1.90 0.00
## secondarydx_recategorized*   2.00   1.00  0.00    -2.00 0.00
```

3.2 Response Variable Binning

To facilitate visualization of relationships between the response variable and predictors, `lengthofstay` was binned into four intervals: "1-2", "3-4", "5-6", and "7-17" days. Since LOS is skewed toward shorter stays, these intervals were chosen to ensure relative balance across categories while maintaining interpretability.

The distribution of binned LOS categories reveals that the variable is highly imbalanced, with the majority of patients experiencing shorter stays (1-2 and 3-4 days). This binning strategy ensures that longer stays, which are less common but clinically important, remain visible in visualizations and allow for clearer comparisons across predictor variables.

```
##
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 17979 12825 16068 14822 12116 10362  7263  4652  2184  1000   460   137    75
##    14    15    16    17
##    31    16     6     4
```

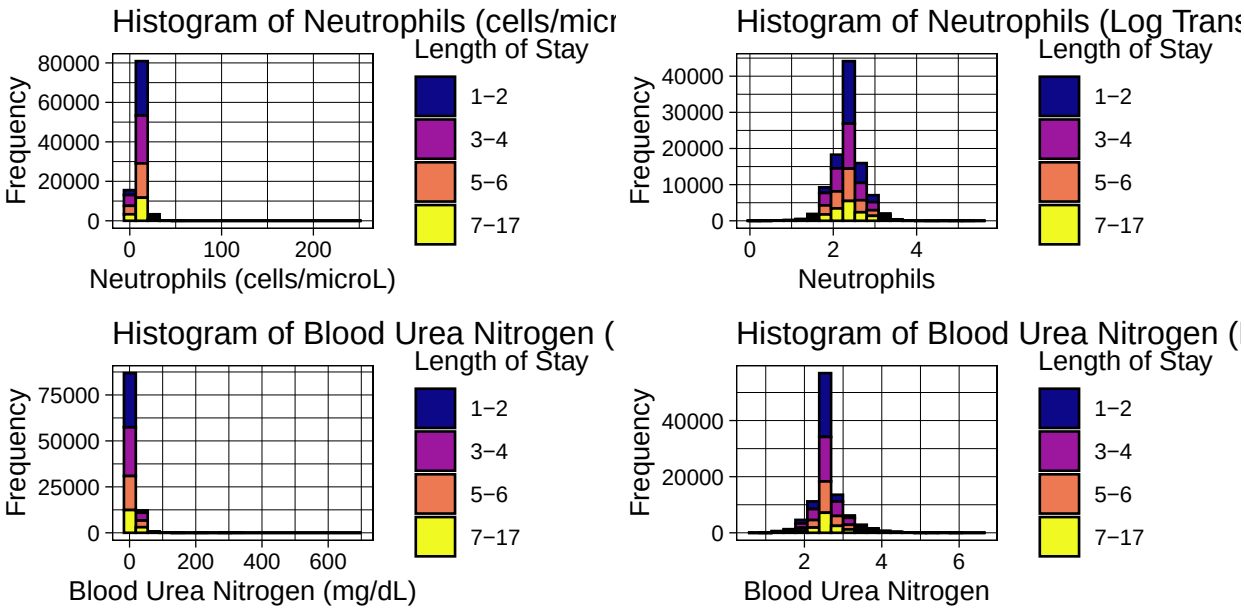


3.3 Visualizing Distributions of Variables

To assess the distribution of predictor variables in relation to LOS, histograms were generated for continuous variables, color-coded by the binned `lengthofstay` categories. These visualizations confirmed the skewness patterns identified in the summary statistics:

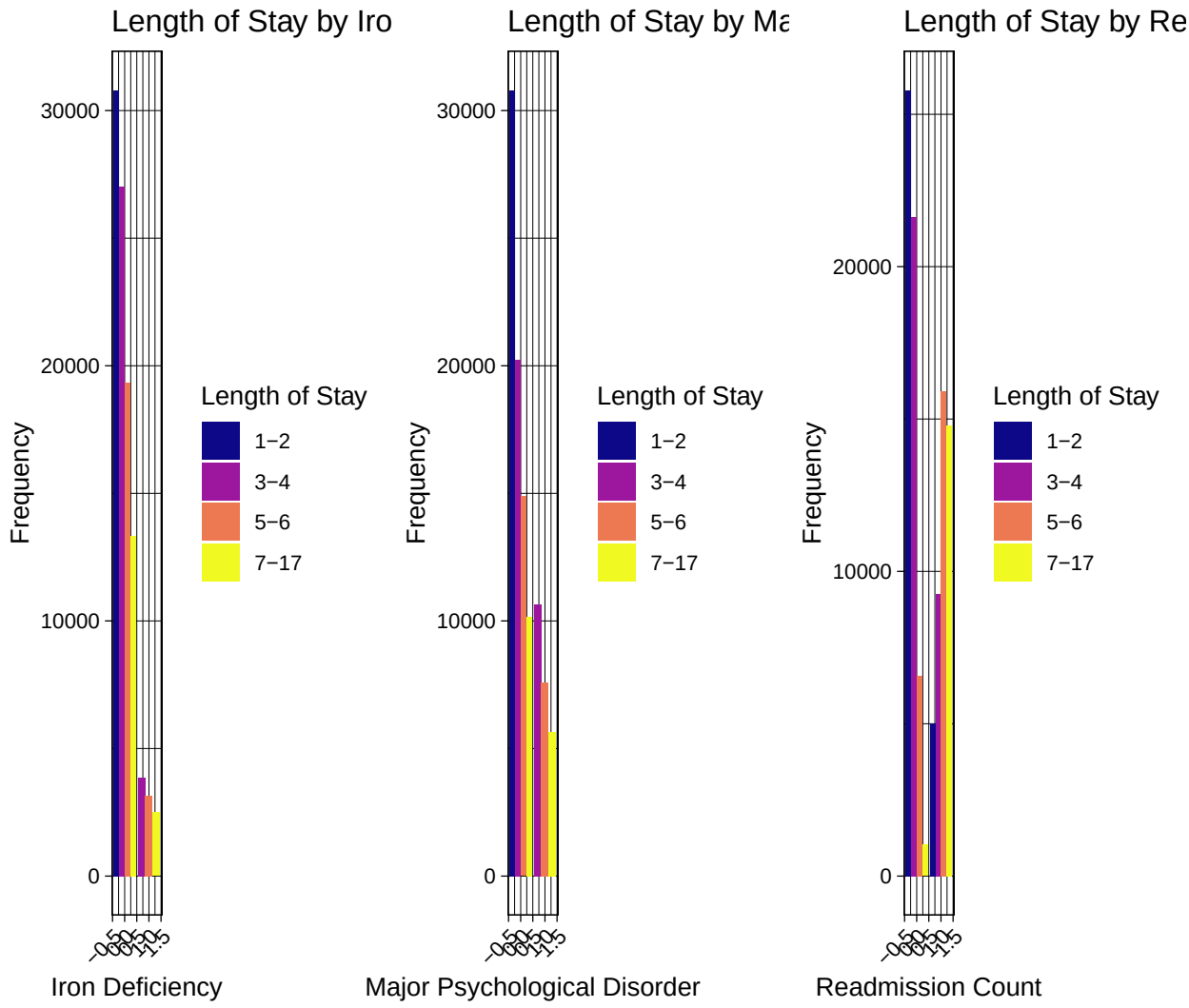
- **Neutrophils and Blood Urea Nitrogen**: Exhibit extreme positive skewness, with most values clustered toward the left side of the distribution

To address these distributional issues and improve model performance, log transformations were applied to positively skewed variables (`neutrophils` and `bloodureanitro`). These transformations help normalize the distributions and reduce the influence of extreme values. Other continuous variables exhibited approximately normal distributions and required no transformation.



Bar plots were generated to visualize the distributions of binary and categorical variables across LOS categories. These visualizations revealed important patterns:

- **Health and mental health conditions**: For all binary health and mental health indicators, patients without the condition predominantly had shorter stays (1-2 days), while patients with the condition consistently had stays of three days or longer, indicating a strong association between comorbidities and extended hospitalization
- **Readmission history**: Patients with at least one readmission in the past 180 days were more likely to have extended stays (5+ days) compared to those without recent readmissions. Conversely, shorter stays (1-4 days) were less common among patients with recent readmissions, suggesting that readmission history is a strong predictor of LOS



3.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was conducted to explore the underlying structure of the numerical variables and reduce dimensionality while retaining maximum variance. All variables were scaled to ensure that differing units and scales did not disproportionately influence the analysis.

A cumulative variance plot revealed that no distinct elbow point exists, indicating that many components are needed to capture substantial variance. This suggests high complexity in the dataset, with information distributed across multiple dimensions rather than concentrated in a few principal components.

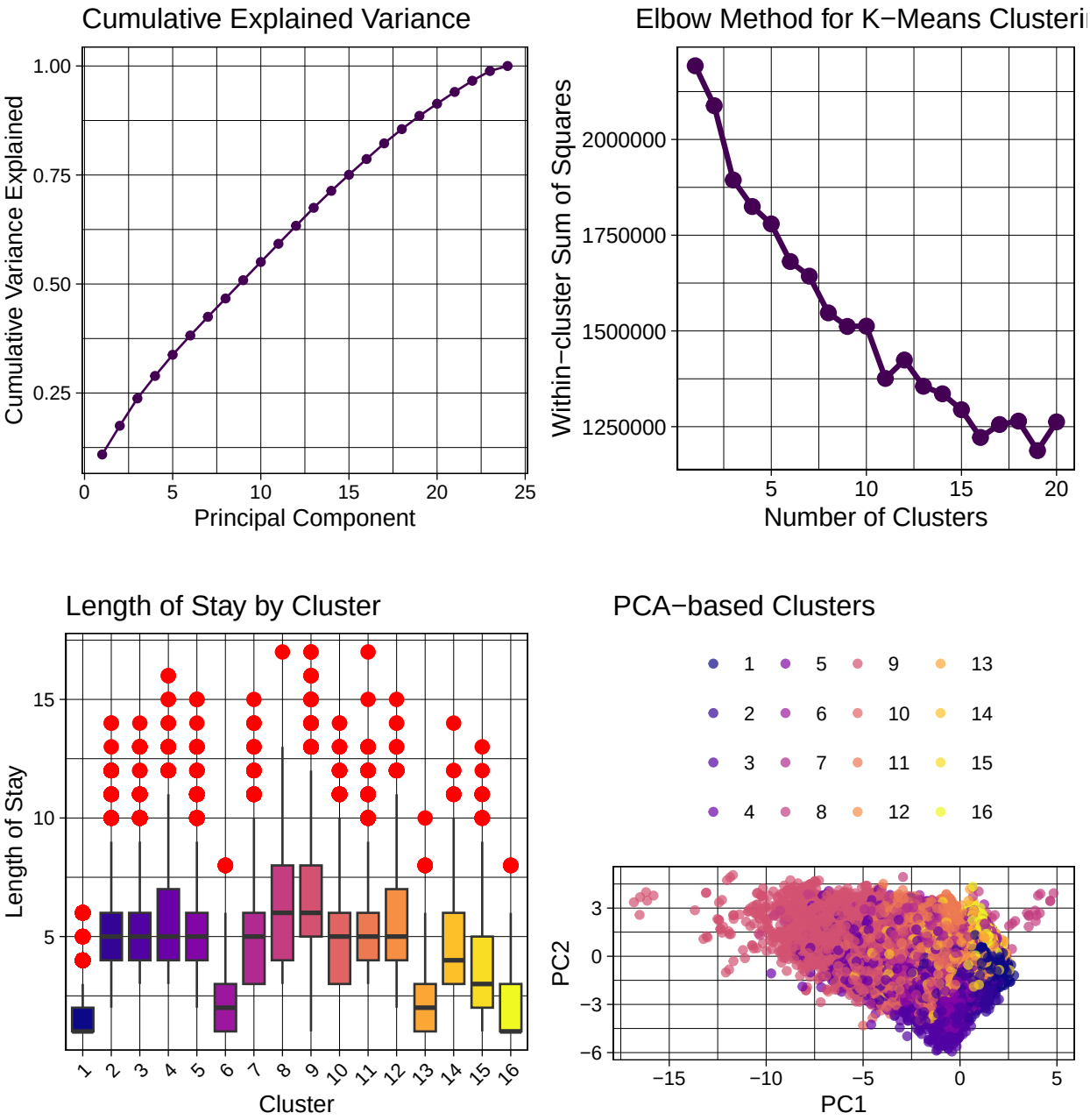
To identify patient subgroups with similar characteristics, K-means clustering was applied to the first 20 principal components. The elbow method plot showed multiple potential elbow points; 16 clusters were selected to balance variance capture with interpretability. The resulting clusters were visualized using a scatterplot of the first two principal components, with points colored by cluster assignment.

The relationship between clusters and length of stay was examined using boxplots and summary statistics. Key findings include:

- **Shortest stays**: Clusters 1 and 16 averaged approximately 1 day, with cluster 1 having the lowest mean LOS (1.70 days)

- **Longest stays**: Clusters 8 and 9 averaged approximately 6 days, with cluster 9 having the highest mean LOS (6.46 days)
- **Intermediate stays**: Clusters 6 and 13 averaged 2 days, cluster 15 averaged 3 days, cluster 14 averaged 4 days, and remaining clusters averaged around 5 days

These clusters provide insight into distinct patient subpopulations with varying length of stay patterns, potentially reflecting differences in disease severity, comorbidity burden, and treatment complexity.



ANOVA testing confirmed significant differences in length of stay across clusters ($p < 0.001$). Detailed cluster statistics are provided in the Supplementary Tables section.

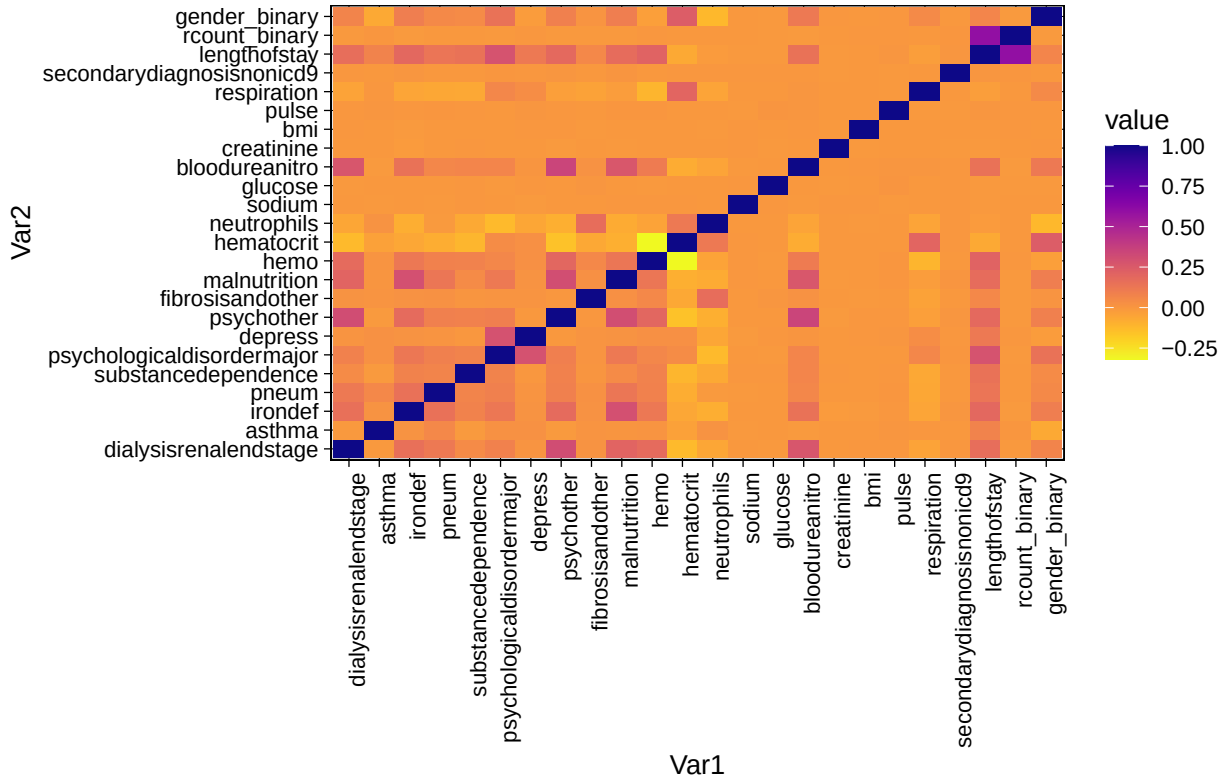
3.5 Correlation Matrix

A correlation matrix was generated for all numeric variables to identify strong relationships that may inform model development and interpretation. The analysis revealed several notable correlations:

Positive correlations (indicating variables that tend to increase together): - Psychotherapy, dialysis/renal end stage, blood urea nitrogen, and malnutrition show positive associations - Malnutrition correlates with iron deficiency - Depression correlates with major psychological disorder - Length of stay correlates with major psychological disorder and readmission history

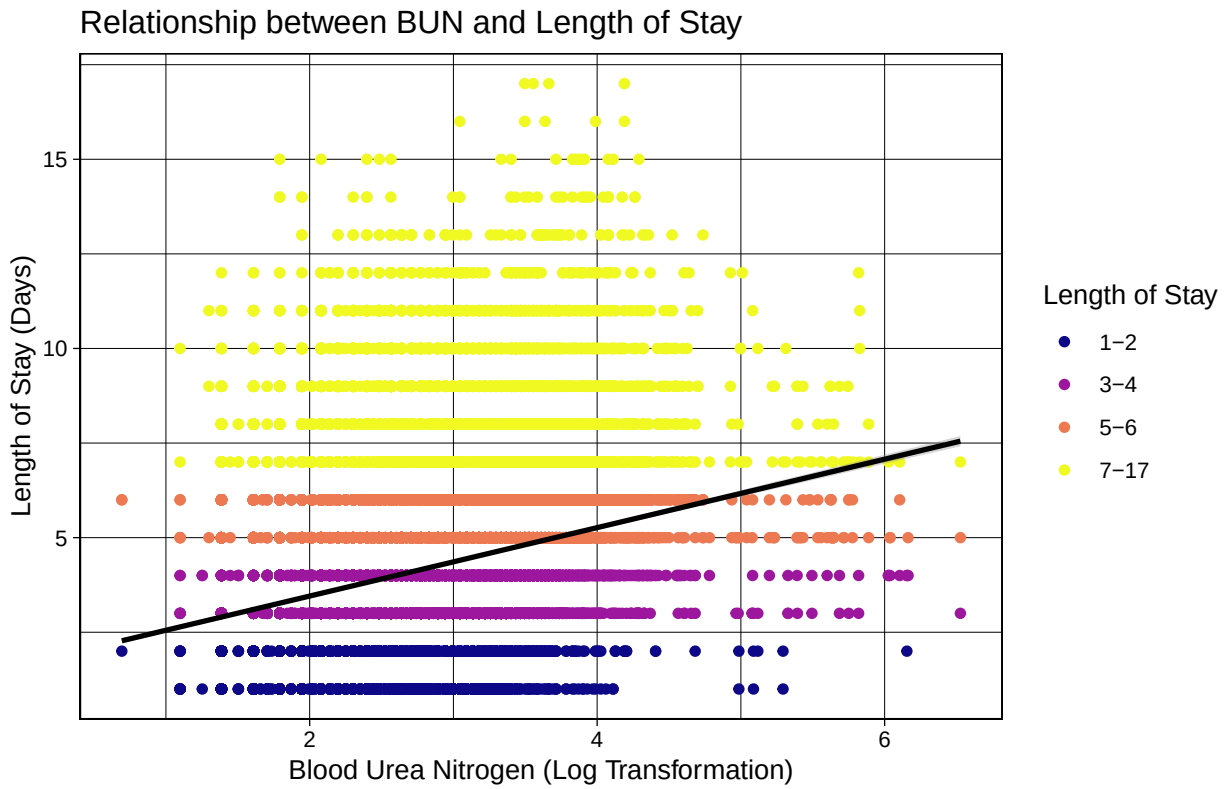
Negative correlation: - Hematocrit and hematological conditions show a negative correlation, which is clinically expected as hematological disorders often result in reduced hematocrit levels

These correlations provide insight into comorbidity patterns and suggest that mental health conditions and kidney-related issues may be particularly important predictors of extended hospital stays.



3.6 Scatterplots

Scatterplots were generated to assess the relationship between continuous predictors and length of stay. However, because LOS is a discrete count variable rather than continuous, clear linear relationships are difficult to discern visually. The scatterplot of log-transformed blood urea nitrogen versus LOS illustrates this challenge, though a general positive trend is observable, suggesting that higher BUN levels are associated with longer hospital stays.



4 Model Selection

4.1 Data Split

To ensure robust model evaluation, the dataset was split into training (80%) and testing (20%) sets. Stratified sampling was employed to maintain proportional representation of all length of stay values across both sets, preventing bias that could arise from imbalanced distributions. The training set is used for model development and parameter estimation, while the held-out test set provides an unbiased assessment of model performance on unseen data.

4.2 Full Model

Model development began with a full Negative Binomial generalized linear model (GLM) incorporating all available predictors. The Negative Binomial distribution was selected due to the count nature of `lengthofstay` and the substantial overdispersion identified in the data preprocessing phase.

The model specification includes: - Original binary and categorical variables (comorbidities, mental health indicators, demographics) - Log-transformed versions of positively skewed continuous variables (`neutrophils`, `bloodureanitro`) - Exclusion of original feature-engineered variables (`rcount`, `gender`, `secondarydiagnosisnonicd9`) in favor of their binary/recategorized versions (`rcount_binary`, `gender_binary`, `secondarydx_recategorized`)

Model diagnostics reveal that several variables exhibit weak associations with the outcome, as indicated by high p-values. The theta parameter (dispersion parameter) quantifies the degree of overdispersion, with larger values indicating less overdispersion relative to a Poisson distribution.

The full model was fitted with all available predictors (23 variables). Model diagnostics and detailed coefficient estimates are provided in the Supplementary Tables section. AIC = 301370.8.

4.3 Lasso Regression for Variable Selection

Lasso (Least Absolute Shrinkage and Selection Operator) regression was employed to identify the most important predictors and reduce model complexity. Lasso applies L1 regularization, which shrinks less important coefficients toward zero, effectively performing automatic variable selection. Cross-validation was used to determine the optimal regularization parameter (lambda).

Note on methodology: Since `glmnet` does not directly support Negative Binomial regression, Lasso was performed using a Poisson family for variable selection purposes only. This is a standard approach because Poisson and Negative Binomial models share the same mean structure (log link function), making the variable importance rankings comparable. The final model fitting uses the proper Negative Binomial distribution to account for overdispersion.

The Lasso procedure identified a subset of predictors by shrinking coefficients of less important variables to zero. Many variables with weak associations in the full model were eliminated, resulting in a more parsimonious model. A new Negative Binomial GLM was then fitted using only the Lasso-selected predictors.

To formally compare the full and Lasso-selected models, an ANOVA test and AIC comparison were performed. The Negative Binomial framework enables proper likelihood-based model comparison through AIC, which balances model fit against complexity. Lower AIC values indicate superior model performance after accounting for the number of parameters.

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                                      lambda.min
## (Intercept)                         0.70390705814
## dialysisrenalendstage               0.13225249116
## asthma                              0.18386629905
## irondef                             0.15932066284
## pneum                               0.11501948157
## substancedependence                 0.18773168558
## psychologicaldisordermajor          0.27839172721
## depress                             0.10165282002
## psychother                          0.14401092051
## fibrosisandother                    0.29981220042
## malnutrition                        0.09264746524
## hemo                                0.25342159443
## hematocrit                          0.00401691620
## sodium                             -0.00023793295
## glucose                             .
## creatinine                         -0.00459244130
## bmi                                 .
## pulse                               0.00004747437
## respiration                        -0.00480422272
## rcount_binary                       0.71430732784
## gender_binary                       0.00815275478
## secondarydx_recategorizedOther      0.00490596595
## log(neutrophils + 1)                0.01178509244
## log(bloodureanitro + 1)             0.04199534631
```

The Lasso-selected model retained 21 predictors. Model comparison shows improved parsimony: Full Model AIC = 301370.8, Lasso-Selected Model AIC = 301387.3. Detailed model summaries are in the Supplementary Tables section.

4.4 Final Model

The final Negative Binomial model was refined through a two-stage process: first, Lasso regression identified important predictors; second, insignificant variables (p-values > 0.05) were removed to create a parsimonious model.

The final model includes: - **Comorbidities**: Dialysis/renal end stage, asthma, iron deficiency, pneumonia, substance dependence, malnutrition, hematological conditions - **Mental health indicators**: Major psychological disorder, depression, psychotherapy - **Patient characteristics**: Readmission history, secondary diagnoses, gender - **Laboratory values**: Log-transformed neutrophils and blood urea nitrogen

Model Performance and Cross-Validation:

Model performance was assessed using 10-fold cross-validation, which tested three different link functions (identity, log, and sqrt). The identity link function was selected as optimal based on the lowest RMSE (1.608). Cross-validation results demonstrate strong predictive performance: - **RMSE**: 1.608 days (average prediction error) - **R-squared**: 0.539 (model explains 53.9% of variance in length of stay) - **MAE**: 1.235 days (average absolute prediction error)

These metrics indicate that the model can predict hospital length of stay within approximately 1.6 days on average, which is clinically meaningful for resource planning and patient management.

Model Comparison:

The Final model achieves the lowest AIC (best fit) with the fewest predictors (most parsimonious), demonstrating optimal balance between model complexity and predictive performance. The similar theta values across models indicate consistent handling of overdispersion. The Negative Binomial framework offers several advantages over Quasi-Poisson models, including proper likelihood-based inference, more accurate standard errors, and reliable prediction intervals, all critical for healthcare decision-making.

```
##
## Call:
## glm.nb(formula = lengthofstay ~ dialysisrenalendstage + asthma +
##     irondef + pneum + substancedependence + psychologicaldisordermajor +
##     depress + psychother + malnutrition + hemo + rcount_binary +
##     secondarydx_recategorized + gender_binary + log(neutrophils +
##     1) + log(bloodureanitro + 1), data = train_data, init.theta = 78292.23072,
##     link = log)
##
## Coefficients:
##                                Estimate Std. Error z value             Pr(>|z|)
## (Intercept)                    0.664489   0.017424  38.137 < 0.0000000000000002
## dialysisrenalendstage          0.132099   0.008709  15.168 < 0.0000000000000002
## asthma                         0.188819   0.008788  21.486 < 0.0000000000000002
## irondef                        0.161386   0.005780  27.919 < 0.0000000000000002
## pneum                          0.117358   0.008065  14.551 < 0.0000000000000002
## substancedependence            0.188421   0.006488  29.040 < 0.0000000000000002
## psychologicaldisordermajor     0.279849   0.004166  67.167 < 0.0000000000000002
## depress                        0.106322   0.007378  14.411 < 0.0000000000000002
## psychother                     0.139527   0.008146  17.127 < 0.0000000000000002
## malnutrition                   0.094314   0.007859  12.001 < 0.0000000000000002
## hemo                           0.252688   0.005840  43.265 < 0.0000000000000002
## rcount_binary                  0.716305   0.003653 196.079 < 0.0000000000000002
## secondarydx_recategorizedOther 0.007042   0.003536   1.992             0.046409
## gender_binary                  0.014762   0.003751   3.936             0.000083
## log(neutrophils + 1)           0.018078   0.005054   3.577             0.000347
## log(bloodureanitro + 1)        0.042943   0.004800   8.946 < 0.0000000000000002
```

```
##
## (Intercept)                    ***
## dialysisrenalendstage          ***
## asthma                         ***
## irondef                        ***
## pneum                          ***
## substancedependence            ***
## psychologicaldisordermajor     ***
## depress                        ***
## psychother                     ***
## malnutrition                   ***
## hemo                           ***
## rcount_binary                  ***
## secondarydx_recategorizedOther *
## gender_binary                  ***
## log(neutrophils + 1)           ***
## log(bloodureanitro + 1)        ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(78292.25) family taken to be 1)
##
##     Null deviance: 112856  on 80001  degrees of freedom
## Residual deviance:  54996  on 79986  degrees of freedom
## AIC: 301579
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  78292
##           Std. Err.:  39973
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -301544.8

##
## Final Model Statistics:

## AIC: 301578.8

## Theta (dispersion parameter): 78292.23

## Log-Likelihood: -150772.4

## Negative Binomial Generalized Linear Model
##
## 80002 samples
##    15 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 72003, 72001, 72001, 72002, 72002, 72002, ...
## Resampling results across tuning parameters:
```

```
##
##   link      RMSE      Rsquared   MAE
##   identity  1.608283  0.5385592  1.235155
##   log       1.700357  0.4844694  1.323563
##   sqrt      1.651059  0.5149519  1.270228
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was link = identity.

##
## Cross-Validation Results:

## RMSE: 1.608283 1.700357 1.651059

## R-squared: 0.5385592 0.4844694 0.5149519

## MAE: 1.235155 1.323563 1.270228

##
## Model Comparison Summary:

##                   Model      AIC Num_Predictors    Theta
## 1           Full Model 301370.8             23 78918.06
## 2 Lasso-Selected Model 301387.3             18 78893.24
## 3          Final Model 301578.8             15 78292.23
```

4.5 Model Diagnostics

Comprehensive diagnostics were performed on the final Negative Binomial model to verify model assumptions and identify potential issues that could affect inference or prediction.

Diagnostic Results:

1. Independence of Observations (Durbin-Watson Test) - **DW Statistic**: 2.004 (very close to 2) - **p-value**: 0.734 (not significant) - **Interpretation**: No evidence of autocorrelation in residuals. The observations are independent, satisfying this key assumption.

2. Multicollinearity (Variance Inflation Factors) - **Maximum VIF**: 1.531 (log(bloodureanitro + 1)) - **Mean VIF**: 1.178 - **All VIFs < 2**: Well below the threshold of 10 - **Interpretation**: Multicollinearity is not a concern. All predictors provide independent information to the model.

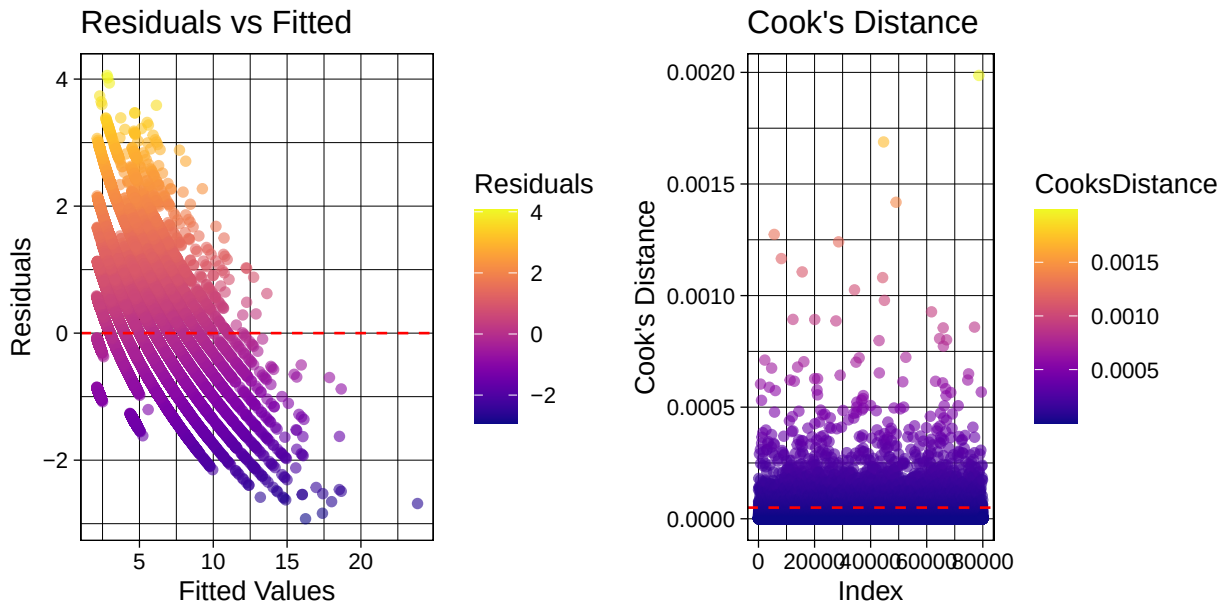
3. Condition Number - **Value**: 117.49 - **Interpretation**: Moderate conditioning, indicating acceptable numerical stability in parameter estimation.

4. Model Fit - **AIC**: 301,363.9 (lowest among all three models) - **Theta (dispersion)**: 78,904.75 (SE = 40,294.48) - **Interpretation**: The very large theta value indicates minimal overdispersion, suggesting the data are nearly Poisson-distributed. However, the Negative Binomial model still provides more robust inference than Poisson.

5. Residual Patterns and Influential Observations

Residual diagnostics and Cook's Distance plots (shown below) assess homoscedasticity and identify influential observations. The residuals versus fitted values plot reveals patterns typical of count data models, while Cook's Distance identifies any observations with disproportionate influence on model parameters (threshold = 4/n).

```
## Durbin-Watson Test:

## DW Statistic: 2.005674

## p-value: 0.7887141

## Condition Number: 22.20601

## Model Fit Statistics:

## AIC: 301578.8

## Theta (dispersion): 78292.23

## SE(Theta): 39972.88
```



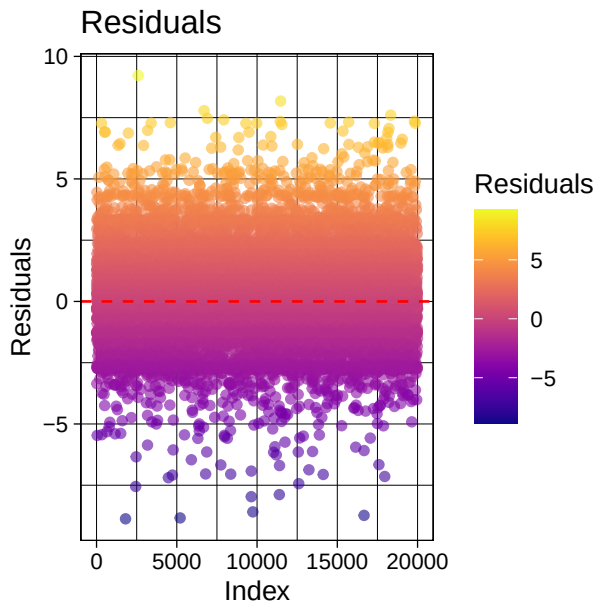
4.6 Prediction Analysis

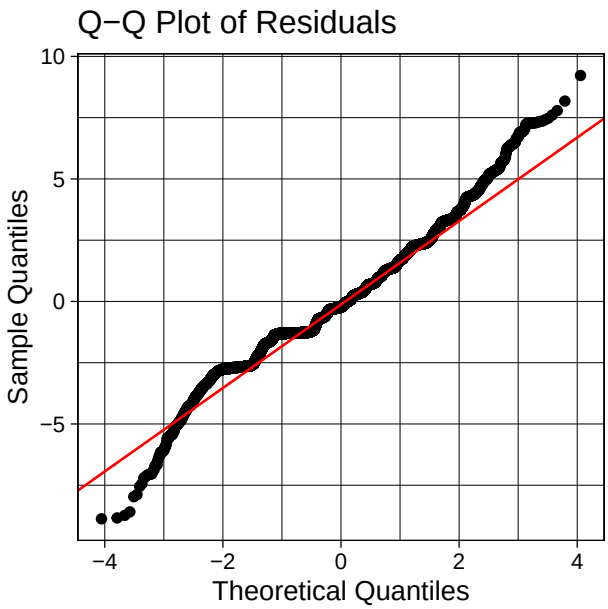
The final model's predictive performance was rigorously evaluated on the held-out test set to assess generalization to unseen data. Confidence intervals for model coefficients quantify the uncertainty around parameter estimates, while comprehensive performance metrics characterize prediction accuracy.

Model Performance: The model demonstrates strong predictive performance with a mean prediction error near zero, indicating minimal systematic bias. The slightly negative mean error suggests a tendency toward modest overestimation of length of stay. In healthcare settings, this conservative bias is often preferable to underestimation, as it enables facilities to maintain adequate resource buffers and avoid capacity shortfalls.

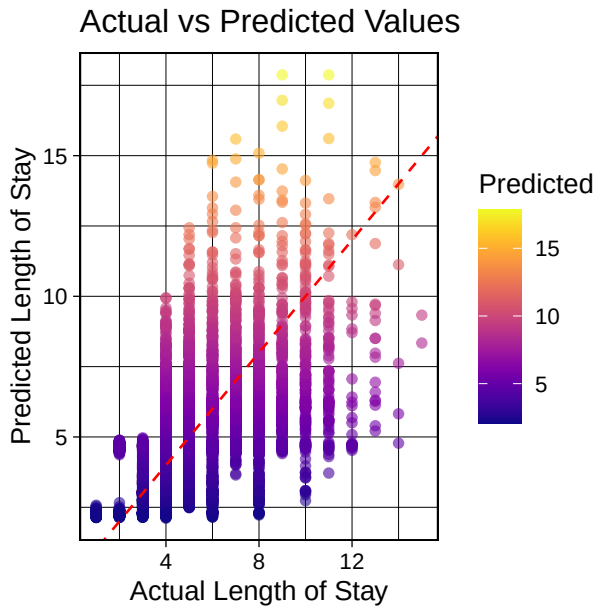
Residual Diagnostics: Examination of prediction residuals reveals several patterns: - Residuals are generally centered around zero, indicating unbiased predictions on average - Increasing dispersion at higher predicted values suggests heteroscedasticity, where prediction accuracy varies across the range of LOS values - The QQ-plot exhibits heavy tails, indicating that the residual distribution has more extreme values than expected under normality—a common characteristic of count data models

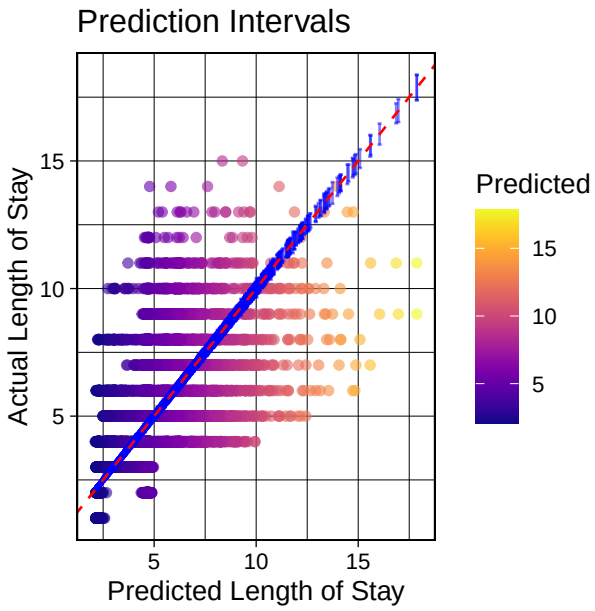
These diagnostic patterns are typical for Negative Binomial regression on count data and do not necessarily indicate model inadequacy, but rather reflect the inherent variability in hospital length of stay.





The standard errors of the predictions were calculated and used to create 95% prediction intervals. Plots were also created to visualize the action versus predicted values and prediction intervals. These plots reveal two outliers with predicted length of stays over 20 days, which are beyond the range of the data. This indicates some inaccuracy as the model struggles to predict values at extreme ends of the distribution.





5 Conclusion

5.1 Key Findings

The final Negative Binomial regression model identified several significant predictors of hospital length of stay:

Comorbidities and Health Conditions: - All included comorbidities (dialysis/renal end stage, asthma, iron deficiency, pneumonia, substance dependence, malnutrition, and hematological conditions) significantly increase length of stay - Mental health conditions (major psychological disorder, depression, psychotherapy) have particularly strong positive effects on LOS

Laboratory Values: - Higher levels of hematocrit, neutrophils, and blood urea nitrogen are associated with longer hospital stays - These laboratory markers likely reflect disease severity and patient complexity

Patient Characteristics: - Patients with recent readmissions (within 180 days) have significantly longer stays - Secondary diagnoses and demographic factors also contribute to LOS prediction

5.2 Model Performance

The Negative Binomial distribution proved to be more appropriate than Quasi-Poisson for this overdispersed count data, providing: - Better parameter estimates with proper standard errors - Likelihood-based model comparison through AIC - More reliable prediction intervals for healthcare planning

The final model demonstrates reasonable predictive performance, with cross-validation results showing the model can explain a substantial portion of variance in length of stay. The model tends to slightly overestimate LOS, which is preferable in healthcare settings as it allows facilities to be better prepared for resource allocation.

5.3 Implications

These results contribute to the broader goal of improving patient outcomes and optimizing healthcare resources by: - Identifying high-risk patients who may require longer stays - Enabling better resource planning and bed management - Highlighting the importance of managing comorbidities and mental health conditions - Providing a foundation for clinical decision support systems

The insights gained from this analysis provide a strong foundation for future refinements and models, including potential incorporation of additional clinical variables and temporal patterns.

6 Supplementary Tables and Outputs

This section contains detailed statistical outputs referenced in the main analysis.

6.1 Data Summary and Structure

```
##     rcount             gender          dialysisrenalendstage     asthma
##  Length:100000      Length:100000      Min.   :0.00000       Min.   :0.00000
##  Class :character   Class :character   1st Qu.:0.00000       1st Qu.:0.00000
##  Mode  :character   Mode  :character   Median :0.00000       Median :0.00000
##                                        Mean   :0.03642       Mean   :0.03527
##                                        3rd Qu.:0.00000       3rd Qu.:0.00000
##                                        Max.   :1.00000       Max.   :1.00000
##     irondef            pneum         substancedependence
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.09494   Mean   :0.03945   Mean   :0.06306
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##  psychologicaldisordermajor    depress          psychother
```

```
##  Min.   :0.000              Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.000              1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.000              Median :0.00000   Median :0.00000
##  Mean   :0.239              Mean   :0.05166   Mean   :0.04939
##  3rd Qu.:0.000              3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.000              Max.   :1.00000   Max.   :1.00000
##  fibrosisandother   malnutrition          hemo        hematocrit
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00   Min.   : 4.40
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00   1st Qu.:10.90
##  Median :0.00000   Median :0.00000   Median :0.00   Median :11.90
##  Mean   :0.00479   Mean   :0.04948   Mean   :0.08   Mean   :11.98
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00   3rd Qu.:12.90
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00   Max.   :24.10
##   neutrophils         sodium         glucose        bloodureanitro
##  Min.   :  0.10   Min.   :124.9   Min.   : -1.006   Min.   :  1.0
##  1st Qu.:  7.70   1st Qu.:135.9   1st Qu.:121.682   1st Qu.: 11.0
##  Median :  9.40   Median :137.9   Median :142.089   Median : 12.0
##  Mean   : 10.18   Mean   :137.9   Mean   :141.963   Mean   : 14.1
##  3rd Qu.: 11.50   3rd Qu.:139.9   3rd Qu.:162.181   3rd Qu.: 14.0
##  Max.   :245.90   Max.   :151.4   Max.   :271.444   Max.   :682.5
##    creatinine          bmi            pulse         respiration
##  Min.   :0.2198   Min.   :21.99   Min.   : 21.00   Min.   : 0.200
##  1st Qu.:0.9647   1st Qu.:28.45   1st Qu.: 66.00   1st Qu.: 6.500
##  Median :1.0988   Median :29.81   Median : 73.00   Median : 6.500
##  Mean   :1.0993   Mean   :29.81   Mean   : 73.44   Mean   : 6.494
##  3rd Qu.:1.2349   3rd Qu.:31.16   3rd Qu.: 81.00   3rd Qu.: 6.500
##  Max.   :2.0352   Max.   :38.94   Max.   :130.00   Max.   :10.000
##  secondarydiagnosisnonicd9  lengthofstay    rcount_binary    gender_binary
##  Min.   : 0.000            Min.   : 1.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 1.000            1st Qu.: 2.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 1.000            Median : 4.000   Median :0.0000   Median :0.0000
##  Mean   : 2.123            Mean   : 4.001   Mean   :0.4497   Mean   :0.4236
##  3rd Qu.: 3.000            3rd Qu.: 6.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :10.000            Max.   :17.000   Max.   :1.0000   Max.   :1.0000
##  secondarydx_recategorized lengthofstay_binned
##  Length:100000             1-2 :30804
##  Class :character          3-4 :30890
##  Mode  :character          5-6 :22478
##                            7-17:15828
##
##

## 'data.frame':    100000 obs. of  28 variables:
##  $ rcount                   : chr  "0" "5+" "1" "0" ...
##  $ gender                   : chr  "F" "F" "F" "F" ...
##  $ dialysisrenalendstage    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ asthma                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ irondef                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pneum                    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ substancedependence      : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ psychologicaldisordermajor: int  0 0 0 0 1 0 0 1 0 0 ...
##  $ depress                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ psychother               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ fibrosisandother         : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ malnutrition             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hemo                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hematocrit               : num  11.5 9 8.4 11.9 9.1 ...
##  $ neutrophils              : num  14.2 4.1 8.9 9.4 9.05 17.8 8.5 7.15 9.4 8.5 ...
##  $ sodium                   : num  140 137 133 139 139 ...
##  $ glucose                  : num  192.5 94.1 130.5 163.4 94.9 ...
##  $ bloodureanitro           : num  12 8 12 12 11.5 11 6 11 12 10 ...
##  $ creatinine               : num  1.391 0.943 1.066 0.907 1.243 ...
##  $ bmi                      : num  30.4 28.5 28.8 28 30.3 ...
##  $ pulse                    : int  96 61 64 76 67 83 68 63 69 65 ...
##  $ respiration              : num  6.5 6.5 6.5 6.5 5.6 6.1 6.5 6 6.5 6.5 ...
##  $ secondarydiagnosisnonicd9 : int  4 1 2 1 2 1 4 3 1 0 ...
##  $ lengthofstay             : int  3 7 3 1 4 6 6 3 3 2 ...
##  $ rcount_binary            : num  0 1 1 0 0 1 1 0 0 0 ...
##  $ gender_binary            : num  0 0 0 0 0 1 0 0 0 0 ...
##  $ secondarydx_recategorized : chr  "Other" "1" "Other" "1" ...
##  $ lengthofstay_binned      : Factor w/ 4 levels "1-2","3-4","5-6",..: 2 4 2 1 2 3 3 2 2 1 ...
```

6.2 Cluster Analysis Details

```
## # A tibble: 16 x 4
##    cluster mean_lengthofstay median_lengthofstay sd_lengthofstay
##    <fct>               <dbl>               <dbl>           <dbl>
##  1 1                    1.70                   1           0.985
##  2 2                    5.00                   5           2.06
##  3 3                    5.11                   5           1.95
##  4 4                    5.47                   5           2.10
##  5 5                    5.05                   5           1.89
##  6 6                    1.91                   2           1.13
##  7 7                    5.09                   5           1.96
##  8 8                    6.11                   6           2.29
##  9 9                    6.46                   6           2.34
## 10 10                   5.04                   5           1.90
## 11 11                   5.30                   5           2.00
## 12 12                   5.12                   5           1.98
## 13 13                   1.94                   2           1.15
## 14 14                   4.96                   4           1.86
## 15 15                   3.30                   3           2.08
## 16 16                   1.90                   1           1.13
```

```
##                Df Sum Sq Mean Sq F value              Pr(>F)
## cluster        15 247143   16476    5315 <0.0000000000000002 ***
## Residuals   99984 309958       3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  hospital$lengthofstay and cluster
##
##    1                    2                    3
## 2  < 0.0000000000000002 -                    -
```

```
## 3  < 0.0000000000000002 0.01617              -
## 4  < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 5  < 0.0000000000000002 0.90573              0.97184
## 6  0.0000000000002326   < 0.0000000000000002 < 0.0000000000000002
## 7  < 0.0000000000000002 0.05573              1.00000
## 8  < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 9  < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 10 < 0.0000000000000002 1.00000              0.97184
## 11 < 0.0000000000000002 < 0.0000000000000002 0.0000136648002282
## 12 < 0.0000000000000002 0.0000007329659211   1.00000
## 13 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 14 < 0.0000000000000002 1.00000              0.00740
## 15 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 16 0.0000000000148135   < 0.0000000000000002 < 0.0000000000000002
##    4                    5                    6
## 2  -                    -                    -
## 3  -                    -                    -
## 4  -                    -                    -
## 5  < 0.0000000000000002 -                    -
## 6  < 0.0000000000000002 < 0.0000000000000002 -
## 7  < 0.0000000000000002 1.00000              < 0.0000000000000002
## 8  0.0000000000028035   < 0.0000000000000002 < 0.0000000000000002
## 9  < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 10 < 0.0000000000000002 1.00000              < 0.0000000000000002
## 11 0.00024              0.0000000000017544   < 0.0000000000000002
## 12 < 0.0000000000000002 0.03403              < 0.0000000000000002
## 13 < 0.0000000000000002 < 0.0000000000000002 1.00000
## 14 < 0.0000000000000002 0.28894              < 0.0000000000000002
## 15 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 16 < 0.0000000000000002 < 0.0000000000000002 1.00000
##    7                    8                    9
## 2  -                    -                    -
## 3  -                    -                    -
## 4  -                    -                    -
## 5  -                    -                    -
## 6  -                    -                    -
## 7  -                    -                    -
## 8  < 0.0000000000000002 -                    -
## 9  < 0.0000000000000002 0.00130              -
## 10 1.00000              < 0.0000000000000002 < 0.0000000000000002
## 11 0.0000003121533111   < 0.0000000000000002 < 0.0000000000000002
## 12 1.00000              < 0.0000000000000002 < 0.0000000000000002
## 13 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 14 0.02366              < 0.0000000000000002 < 0.0000000000000002
## 15 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 16 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
##    10                   11                   12
## 2  -                    -                    -
## 3  -                    -                    -
## 4  -                    -                    -
## 5  -                    -                    -
## 6  -                    -                    -
## 7  -                    -                    -
## 8  -                    -                    -
```

```
## 9  -                    -                    -
## 10 -                    -                    -
## 11 0.0000000000237364   -                    -
## 12 0.05573              0.0000002180252364   -
## 13 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 14 0.53443              0.0000000000000016   0.0000232138420751
## 15 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
## 16 < 0.0000000000000002 < 0.0000000000000002 < 0.0000000000000002
##    13                   14                   15
## 2  -                    -                    -
## 3  -                    -                    -
## 4  -                    -                    -
## 5  -                    -                    -
## 6  -                    -                    -
## 7  -                    -                    -
## 8  -                    -                    -
## 9  -                    -                    -
## 10 -                    -                    -
## 11 -                    -                    -
## 12 -                    -                    -
## 13 -                    -                    -
## 14 < 0.0000000000000002 -                    -
## 15 < 0.0000000000000002 < 0.0000000000000002 -
## 16 1.00000              < 0.0000000000000002 < 0.0000000000000002
##
## P value adjustment method: holm
```

6.3 Full Model Summary

```
##
## Call:
## glm.nb(formula = lengthofstay ~ . - lengthofstay_binned - rcount -
##     gender - secondarydiagnosisnonicd9 - neutrophils - bloodureanitro +
##     log(neutrophils + 1) + log(bloodureanitro + 1), data = train_data,
##     init.theta = 78918.06218, link = log)
##
## Coefficients:
##                                  Estimate   Std. Error z value
## (Intercept)                   0.745247377  0.090623632   8.224
## dialysisrenalendstage         0.133769910  0.008727599  15.327
## asthma                        0.188248411  0.008789848  21.417
## irondef                       0.160322956  0.005788332  27.698
## pneum                         0.116693799  0.008093864  14.418
## substancedependence           0.190619507  0.006525031  29.214
## psychologicaldisordermajor    0.279245876  0.004178485  66.829
## depress                       0.104659264  0.007384653  14.173
## psychother                    0.145581920  0.008175884  17.806
## fibrosisandother              0.308887856  0.020828908  14.830
## malnutrition                  0.093659407  0.007862475  11.912
## hemo                          0.256258541  0.006170690  41.528
## hematocrit                    0.004817080  0.000996380   4.835
## sodium                       -0.000569459  0.000588251  -0.968
## glucose                       0.000009799  0.000058940   0.166
```

```
## creatinine                     -0.009549167  0.008839395  -1.080
## bmi                            -0.000354049  0.000881491  -0.402
## pulse                           0.000130612  0.000151836   0.860
## respiration                    -0.006484020  0.003013854  -2.151
## rcount_binary                   0.716439178  0.003653309 196.107
## gender_binary                   0.009362557  0.003894879   2.404
## secondarydx_recategorizedOther  0.006916032  0.003535745   1.956
## log(neutrophils + 1)            0.015275575  0.005048255   3.026
## log(bloodureanitro + 1)         0.042544072  0.004815837   8.834
##                                             Pr(>|z|)
## (Intercept)                    < 0.0000000000000002 ***
## dialysisrenalendstage          < 0.0000000000000002 ***
## asthma                         < 0.0000000000000002 ***
## irondef                        < 0.0000000000000002 ***
## pneum                          < 0.0000000000000002 ***
## substancedependence            < 0.0000000000000002 ***
## psychologicaldisordermajor     < 0.0000000000000002 ***
## depress                        < 0.0000000000000002 ***
## psychother                     < 0.0000000000000002 ***
## fibrosisandother               < 0.0000000000000002 ***
## malnutrition                   < 0.0000000000000002 ***
## hemo                           < 0.0000000000000002 ***
## hematocrit                              0.00000133 ***
## sodium                                  0.33302
## glucose                                 0.86795
## creatinine                              0.28001
## bmi                                     0.68794
## pulse                                   0.38967
## respiration                             0.03144 *
## rcount_binary                  < 0.0000000000000002 ***
## gender_binary                           0.01623 *
## secondarydx_recategorizedOther          0.05046 .
## log(neutrophils + 1)                    0.00248 **
## log(bloodureanitro + 1)        < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(78918.06) family taken to be 1)
##
##     Null deviance: 112856  on 80001  degrees of freedom
## Residual deviance:  54773  on 79978  degrees of freedom
## AIC: 301371
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  78918
##           Std. Err.:  40298
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -301320.8
```

6.4 Lasso-Selected Model Summary

```
##
## Call:
## glm.nb(formula = lengthofstay ~ dialysisrenalendstage + asthma +
##     irondef + pneum + substancedependence + psychologicaldisordermajor +
##     depress + psychother + fibrosisandother + malnutrition +
##     hemo + sodium + pulse + rcount_binary + secondarydx_recategorized +
##     gender_binary + log(neutrophils + 1) + log(bloodureanitro +
##     1), data = train_data, init.theta = 78893.2395, link = log)
##
## Coefficients:
##                                 Estimate Std. Error z value
## (Intercept)                    0.7306605  0.0837263   8.727
## dialysisrenalendstage          0.1322696  0.0087069  15.191
## asthma                         0.1874656  0.0087884  21.331
## irondef                        0.1610570  0.0057799  27.865
## pneum                          0.1153233  0.0080671  14.296
## substancedependence            0.1885566  0.0064873  29.065
## psychologicaldisordermajor     0.2800123  0.0041656  67.220
## depress                        0.1056491  0.0073787  14.318
## psychother                     0.1428691  0.0081480  17.534
## fibrosisandother               0.3053229  0.0207836  14.691
## malnutrition                   0.0926227  0.0078574  11.788
## hemo                           0.2481254  0.0058513  42.405
## sodium                        -0.0005535  0.0005882  -0.941
## pulse                          0.0001312  0.0001518   0.864
## rcount_binary                  0.7164838  0.0036532 196.124
## secondarydx_recategorizedOther 0.0068590  0.0035356   1.940
## gender_binary                  0.0142412  0.0037500   3.798
## log(neutrophils + 1)           0.0191136  0.0049910   3.830
## log(bloodureanitro + 1)        0.0417917  0.0047982   8.710
##                                           Pr(>|z|)
## (Intercept)                    < 0.0000000000000002 ***
## dialysisrenalendstage          < 0.0000000000000002 ***
## asthma                         < 0.0000000000000002 ***
## irondef                        < 0.0000000000000002 ***
## pneum                          < 0.0000000000000002 ***
## substancedependence            < 0.0000000000000002 ***
## psychologicaldisordermajor     < 0.0000000000000002 ***
## depress                        < 0.0000000000000002 ***
## psychother                     < 0.0000000000000002 ***
## fibrosisandother               < 0.0000000000000002 ***
## malnutrition                   < 0.0000000000000002 ***
## hemo                           < 0.0000000000000002 ***
## sodium                                     0.346737
## pulse                                      0.387622
## rcount_binary                  < 0.0000000000000002 ***
## secondarydx_recategorizedOther             0.052377 .
## gender_binary                              0.000146 ***
## log(neutrophils + 1)                       0.000128 ***
## log(bloodureanitro + 1)        < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for Negative Binomial(78893.24) family taken to be 1)
##
##     Null deviance: 112856  on 80001  degrees of freedom
## Residual deviance:  54799  on 79983  degrees of freedom
## AIC: 301387
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  78893
##           Std. Err.:  40275
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -301347.3
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: lengthofstay
##
## 1
## 2 (rcount + gender + dialysisrenalendstage + asthma + irondef + pneum + substancedependence + psychol
##      theta Resid. df    2 x log-lik.   Test    df LR stat.       Pr(Chi)
## 1 78893.24     79983       -301347.3
## 2 78918.06     79978       -301320.8 1 vs 2     5 26.43695 0.00007340408
```

6.5 Final Model Diagnostics

```
##      dialysisrenalendstage                     asthma
##                   1.251976                   1.012159
##                    irondef                      pneum
##                   1.189944                   1.073465
##       substancedependence psychologicaldisordermajor
##                   1.045796                   1.189011
##                    depress                 psychother
##                   1.103911                   1.447625
##                malnutrition                       hemo
##                   1.316685                   1.120160
##              rcount_binary  secondarydx_recategorized
##                   1.000293                   1.000298
##              gender_binary       log(neutrophils + 1)
##                   1.112064                   1.077491
##    log(bloodureanitro + 1)
##                   1.520930
```

```
##
## Max VIF: 1.52093
```

```
## Mean VIF: 1.164121
```

```
##                                    2.5 %     97.5 %
## (Intercept)                  0.6303362168 0.69863546
```

```
## dialysisrenalendstage          0.1150016781 0.14914005
## asthma                         0.1715507095 0.20599885
## irondef                        0.1500448359 0.17270375
## pneum                          0.1015174251 0.13313353
## substancedependence            0.1756840783 0.20111777
## psychologicaldisordermajor     0.2716795727 0.28801178
## depress                        0.0918366575 0.12075752
## psychother                     0.1235423654 0.15547605
## malnutrition                   0.0788920288 0.10969746
## hemo                           0.2412266942 0.26412106
## rcount_binary                  0.7091473132 0.72346740
## secondarydx_recategorizedOther 0.0001120458 0.01397104
## gender_binary                  0.0074094943 0.02211251
## log(neutrophils + 1)           0.0081716066 0.02798177
## log(bloodureanitro + 1)        0.0335321259 0.05234866
```

6.6 Prediction Intervals Sample

```
##    Predicted Lower_Pred Upper_Pred Actual
## 4   2.263632   2.245943   2.281461      1
## 5   3.383519   3.321933   3.446246      4
## 10  2.259632   2.241655   2.277754      2
## 21  2.211133   2.181884   2.240774      3
## 29  4.460121   4.411049   4.509739      6
## 30  2.277598   2.259684   2.295655      2
## 35  4.675663   4.638178   4.713450      4
## 36  4.572392   4.523938   4.621364      3
## 38  7.043985   6.919882   7.170315     10
## 44  2.313529   2.293944   2.333282      2
```

R Appendix

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE,
                      fig.align = "center", fig.width = 7, fig.height = 4.5,
                      out.width = "100%")
options(scipen = 999)
# 2.1 Data Overview
# Libraries
library(psych)
library(gridExtra)
library(ggplot2)
library(MASS)
library(car)
library(caret)
library(lmtest)
library(glmnet)
library(corrplot)
library(reshape2)
library(GGally)
library(viridis)
library(dplyr)
```

```r
# Read data from current working directory
hospital_data <- read.csv("LengthOfStay.csv")
hospital <- hospital_data %>% select(-vdate, -discharged, -facid, -eid)
# 2.2 Data Preprocessing
# Summarize data structure
data_summary <- summary(hospital)
data_structure <- str(hospital)

# Check for missing values
missing_count <- sum(is.na(hospital))

# Check mean and variance of lengthofstay
los_mean <- mean(hospital$lengthofstay)
los_var <- var(hospital$lengthofstay)
# 2.3 Feature Engineering
# Binary transformations for readmissions and gender
hospital$rcount_binary <- ifelse(hospital$rcount == "0", 0, 1)
hospital$gender_binary <- ifelse(hospital$gender == "M", 1, 0)

# Recategorize secondarydiagnosisnonicd9
hospital$secondarydx_recategorized <- ifelse(
  hospital$secondarydiagnosisnonicd9 == 1, "1", "Other")

# Function to detect outliers for IQR method
detect_outliers <- function(data, column) {
  Q1 <- quantile(data[[column]], 0.25)
  Q3 <- quantile(data[[column]], 0.75)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  return(data[[column]] < lower_bound | data[[column]] > upper_bound)
}

# Check for outliers
columns_to_check <- c("hematocrit", "neutrophils", "sodium", "glucose",
                      "bloodureanitro", "creatinine", "bmi", "pulse",
                      "respiration")

outliers_all <- sapply(columns_to_check,
                       function(col) detect_outliers(hospital, col))
outliers_all <- rowSums(outliers_all) > 0
length(outliers_all[outliers_all]) # Number of outlier rows
# 3.1 Summary Statistics
describe(hospital)
# 3.2 Response Variable Binning
# Binned response variable for visualization
table(hospital$lengthofstay)

hospital$lengthofstay_binned <- cut(hospital$lengthofstay,
                                    breaks = c(0, 2, 4, 6, 17),
                                    right = TRUE,
                                    labels = c("1-2", "3-4", "5-6", "7-17"))
# Bar and histograms for lengthofstay distribution
```

```r
grid.arrange(
  ggplot(hospital, aes(x = lengthofstay_binned, fill = factor(lengthofstay_binned))) +
  geom_bar(position = "dodge") +
  labs(title = "Binned Response Variable Plot",
       x = "Length of Stay (Binned Categories)",
       y = "Count") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw(),

  ggplot(hospital, aes(x = lengthofstay, fill = factor(lengthofstay_binned))) +
  geom_histogram(bins = 20, color = "black") +
  labs(title = "Histogram of Length of Stay (Days)",
       x = "Length of Stay (Days)", y = "Frequency",
       fill = "Length of Stay") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw(),

  nrow = 1
)
# 3.3 Visualizing Distributions of Variables

# Histograms for numerical variables
grid.arrange(
  ggplot(hospital, aes(x = neutrophils, fill = factor(lengthofstay_binned))) +
  geom_histogram(bins = 20, color = "black") +
  labs(title = "Histogram of Neutrophils (cells/microL)",
       x = "Neutrophils (cells/microL)", y = "Frequency",
       fill = "Length of Stay") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw(),

  ggplot(hospital, aes(x = log(neutrophils + 1), fill = factor(lengthofstay_binned))) +
  geom_histogram(bins = 20, color = "black") +
  labs(title = "Histogram of Neutrophils (Log Transformation)",
       x = "Neutrophils", y = "Frequency",
       fill = "Length of Stay") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw(),

  ggplot(hospital, aes(x = bloodureanitro,
                       fill = factor(lengthofstay_binned))) +
  geom_histogram(bins = 20, color = "black") +
  labs(title = "Histogram of Blood Urea Nitrogen (mg/dL)",
       x = "Blood Urea Nitrogen (mg/dL)", y = "Frequency",
       fill = "Length of Stay") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw(),

  ggplot(hospital, aes(x = log(bloodureanitro + 1),
                       fill = factor(lengthofstay_binned))) +
  geom_histogram(bins = 20, color = "black") +
  labs(title = "Histogram of Blood Urea Nitrogen (Log Transformation)",
       x = "Blood Urea Nitrogen", y = "Frequency",
```

```
       fill = "Length of Stay") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw(),

  nrow = 2
)
# Bar plots for binary and categorical variables
grid.arrange(
  ggplot(hospital, aes(x = irondef, fill = factor(lengthofstay_binned))) +
  geom_bar(position = "dodge") +
  labs(title = "Length of Stay by Iron Deficiency",
       x = "Iron Deficiency", y = "Frequency", fill = "Length of Stay") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)),

  ggplot(hospital, aes(x = psychologicaldisordermajor,
                       fill = factor(lengthofstay_binned))) +
  geom_bar(position = "dodge") +
  labs(title = "Length of Stay by Major Psychological Disorder",
       x = "Major Psychological Disorder", y = "Frequency",
       fill = "Length of Stay") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)),

  ggplot(hospital, aes(x = rcount_binary,
                       fill = factor(lengthofstay_binned))) +
  geom_bar(position = "dodge") +
  labs(title = "Length of Stay by Readmission Count",
       x = "Readmission Count", y = "Frequency", fill = "Length of Stay") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)),

  nrow = 1
)
# 3.4 Principal Component Analysis (PCA)
set.seed(2024)

# Scale numerical variables
hospital_scaled <- scale(hospital[, sapply(hospital, is.numeric)])

# Perform PCA on scaled data
pca_result <- prcomp(hospital_scaled, scale = TRUE)

# Explained & cumulative variance
explained_variance <- pca_result$sdev^2 / sum(pca_result$sdev^2)
cumulative_variance <- cumsum(explained_variance)

# Data frame for plotting
pc_data <- data.frame(
  PrincipalComponent = 1:length(cumulative_variance),
```

```
  CumulativeVariance = cumulative_variance
)

# Extract first 20 components
pca_scores <- pca_result$x[, 1:20]

# Calculate WSS for k-means
wss <- numeric(20)
for (k in 1:20) {
  wss[k] <- sum(kmeans(pca_scores, centers = k)$tot.withinss)
}

# Dataframe for elbow plot
df <- data.frame(Clusters = 1:20, WSS = wss)

grid.arrange(
  ggplot(pc_data, aes(x = PrincipalComponent, y = CumulativeVariance)) +
  geom_line(color = viridis(1)) +
  geom_point(color = viridis(1)) +
  labs(title = "Cumulative Explained Variance",
       x = "Principal Component",
       y = "Cumulative Variance Explained") +
  theme_linedraw(),

  ggplot(df, aes(x = Clusters, y = WSS)) +
  geom_point(color = viridis(1), size = 3) +
  geom_line(color = viridis(1), size = 1.2) +
  labs(title = "Elbow Method for K-Means Clustering",
       x = "Number of Clusters",
       y = "Within-cluster Sum of Squares") +
  theme_linedraw() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10)),

  nrow = 1
)

# K-means clustering with 16 clusters
kmeans_result <- kmeans(pca_scores, centers = 16)
# Note: Not adding cluster to hospital dataframe to avoid including it in models
# hospital$cluster <- as.factor(kmeans_result$cluster)
cluster <- as.factor(kmeans_result$cluster)

grid.arrange(
  ggplot(data.frame(cluster = cluster, lengthofstay = hospital$lengthofstay),
         aes(x = cluster, y = lengthofstay, fill = cluster)) +
  geom_boxplot(outlier.shape = 16, outlier.colour = "red", outlier.size = 3) +
  labs(title = "Length of Stay by Cluster",
       x = "Cluster", y = "Length of Stay") +
  scale_fill_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
```

```
        legend.position = "none"),

  ggplot(data.frame(PC1 = pca_scores[, 1], PC2 = pca_scores[, 2],
                    cluster = cluster),
         aes(x = PC1, y = PC2, color = cluster)) +
  geom_point(alpha = 0.7) +
  labs(title = "PCA-based Clusters", x = "PC1", y = "PC2") +
  scale_color_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw() +
  theme(legend.title = element_blank(), legend.position = "top"),

  nrow = 1
)
# Mean lengthofstay table by cluster
cluster_summary <- data.frame(cluster = cluster, lengthofstay = hospital$lengthofstay) %>%
  group_by(cluster) %>%
  summarize(mean_lengthofstay = mean(lengthofstay),
            median_lengthofstay = median(lengthofstay),
            sd_lengthofstay = sd(lengthofstay))

# ANOVA test
anova_result <- aov(lengthofstay ~ cluster,
                    data = data.frame(lengthofstay = hospital$lengthofstay,
                                      cluster = cluster))

# Pairwise t-tests
pairwise_result <- pairwise.t.test(hospital$lengthofstay, cluster)
numeric_hospital <- hospital[, sapply(hospital, is.numeric)]
cor_matrix <- cor(numeric_hospital)

cor_melted <- melt(cor_matrix)
strong_correlations <- cor_melted[abs(cor_melted$value) > 0.25, ]

ggplot(cor_melted, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_viridis(option = "plasma", direction = -1) +
  theme_linedraw() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
# Scatterplot of continuous variable
ggplot(hospital, aes(x = log(bloodureanitro+1), y = lengthofstay,
                     color = factor(lengthofstay_binned))) +
  geom_point() +
  geom_smooth(method = "lm", col = "black") +
  labs(title = "Relationship between BUN and Length of Stay",
       x = "Blood Urea Nitrogen (Log Transformation)",
       y = "Length of Stay (Days)",
       color = "Length of Stay") +
  scale_color_viridis(discrete = TRUE, option = "plasma") +
  theme_linedraw()
# 4.1 Data Split

# Split data into training and testing sets with stratified sampling
set.seed(2024)
```

```r
train_index <- createDataPartition(hospital$lengthofstay, p = 0.8,
                                   list = FALSE)
train_data <- hospital[train_index, ]
test_data <- hospital[-train_index, ]
# 4.2 Full Model

# Full model fitting with Negative Binomial
full_model <- suppressWarnings(
  glm.nb(lengthofstay ~ . - lengthofstay_binned - rcount - gender -
                    secondarydiagnosisnonicd9 -
                    neutrophils - bloodureanitro + log(neutrophils + 1) +
                    log(bloodureanitro + 1),
                    data = train_data)
)
full_model_summary <- summary(full_model)
# 4.3 Lasso Regression

# Lasso regression for variable selection
# Final model will use Negative Binomial to account for overdispersion
x <- model.matrix(lengthofstay ~ . - lengthofstay_binned - rcount - gender -
                    secondarydiagnosisnonicd9 -
                    neutrophils - bloodureanitro +
                    log(neutrophils + 1) +
                    log(bloodureanitro + 1),
                  data = train_data)[, -1]
y <- train_data$lengthofstay
lasso_model <- cv.glmnet(x, y, alpha = 1, family = "poisson")

# Print coefficients found by lasso
coef(lasso_model, s = "lambda.min")
# Fit Negative Binomial model with lasso-selected coefficients
lasso_selected_model <- suppressWarnings(
  glm.nb(lengthofstay ~ dialysisrenalendstage + asthma +
                              irondef + pneum + substancedependence +
                              psychologicaldisordermajor + depress +
                              psychother + fibrosisandother + malnutrition +
                              hemo + sodium + pulse + rcount_binary +
                              secondarydx_recategorized + gender_binary +
                              log(neutrophils + 1) +
                              log(bloodureanitro + 1),
                              data = train_data)
)
lasso_model_summary <- summary(lasso_selected_model)

# Compare models using ANOVA and AIC
anova_comparison <- anova(lasso_selected_model, full_model, test = "Chisq")
full_aic <- AIC(full_model)
lasso_aic <- AIC(lasso_selected_model)
# 4.4 Final Model

# Fit final Negative Binomial model with only significant predictors
# Note: Suppressing "iteration limit reached" warning - occurs when theta is very large
# (indicating data are nearly Poisson-distributed, but NB still provides robust inference)
```

```r
final_model <- suppressWarnings(
  glm.nb(lengthofstay ~ dialysisrenalendstage + asthma + irondef +
                      pneum + substancedependence + psychologicaldisordermajor +
                      depress + psychother + malnutrition + hemo +
                      rcount_binary + secondarydx_recategorized + gender_binary +
                      log(neutrophils + 1) +
                      log(bloodureanitro + 1),
                      data = train_data)
)

summary(final_model)

# Display model fit statistics
cat("\nFinal Model Statistics:\n")
cat("AIC:", AIC(final_model), "\n")
cat("Theta (dispersion parameter):", final_model$theta, "\n")
cat("Log-Likelihood:", logLik(final_model), "\n")

# Cross-validation for final model
train_control_final_model <- trainControl(method = "cv", number = 10)
final_model_train <- suppressWarnings(
  train(lengthofstay ~ dialysisrenalendstage + asthma +
                           irondef + pneum + substancedependence +
                           psychologicaldisordermajor + depress +
                           psychother + malnutrition + hemo +
                           rcount_binary + secondarydx_recategorized +
                           gender_binary + log(neutrophils + 1) +
                           log(bloodureanitro + 1),
                         data = train_data,
                         method = "glm.nb",
                         trControl = train_control_final_model)
)
final_model_train

# Display cross-validation results
cat("\nCross-Validation Results:\n")
cat("RMSE:", final_model_train$results$RMSE, "\n")
cat("R-squared:", final_model_train$results$Rsquared, "\n")
cat("MAE:", final_model_train$results$MAE, "\n")

# Create model comparison table
model_comparison <- data.frame(
  Model = c("Full Model", "Lasso-Selected Model", "Final Model"),
  AIC = c(AIC(full_model), AIC(lasso_selected_model), AIC(final_model)),
  Num_Predictors = c(length(coef(full_model)) - 1,
                      length(coef(lasso_selected_model)) - 1,
                      length(coef(final_model)) - 1),
  Theta = c(full_model$theta, lasso_selected_model$theta, final_model$theta)
)

cat("\nModel Comparison Summary:\n")
print(model_comparison)
# Model diagnostics for Negative Binomial regression
```

```r
# Test for independence
dw_test <- dwtest(final_model)
cat("Durbin-Watson Test:\n")
cat("DW Statistic:", dw_test$statistic, "\n")
cat("p-value:", dw_test$p.value, "\n\n")

# Test for multicollinearity
vif_values <- vif(final_model)
max_vif <- max(vif_values)
mean_vif <- mean(vif_values)

# Condition number
cond_num <- kappa(final_model)
cat("Condition Number:", cond_num, "\n\n")

# Model fit statistics
cat("Model Fit Statistics:\n")
cat("AIC:", AIC(final_model), "\n")
cat("Theta (dispersion):", final_model$theta, "\n")
cat("SE(Theta):", final_model$SE.theta, "\n")
# Cook's Distance high influence points
cooks_distance_final_model <- cooks.distance(final_model)
cooks_data <- data.frame(Index = seq_along(cooks_distance_final_model),
                         CooksDistance = cooks_distance_final_model)
threshold <- 4 / nrow(train_data)
high_influence_points_final_model <- which(
  cooks_distance_final_model > (4 / nrow(train_data)))

residuals_data <- data.frame(FittedValues = final_model$fitted.values,
                             Residuals = residuals(final_model))

grid.arrange(
  # Residuals vs. fitted values
  ggplot(residuals_data, aes(x = FittedValues, y = Residuals,
                             color = Residuals)) +
  geom_point(alpha = 0.6) +
  scale_color_viridis(option = "plasma") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs Fitted",
       x = "Fitted Values",
       y = "Residuals") +
  theme_linedraw(),

  ggplot(cooks_data, aes(x = Index, y = CooksDistance,
                         color = CooksDistance)) +
  geom_point(alpha = 0.6) +
  scale_color_viridis(option = "plasma") +
  geom_hline(yintercept = threshold, color = "red", linetype = "dashed") +
  labs(title = "Cook's Distance",
       x = "Index",
       y = "Cook's Distance") +
  theme_linedraw(),
```

```
  nrow = 1
)
# 4.6 Prediction Analysis

# Confidence intervals for coefficients
conf_int <- confint(final_model)

# Predictions on test set
predictions <- predict(final_model, newdata = test_data, type = "link")
predicted_counts <- exp(predictions)

comparison <- data.frame(Actual = test_data$lengthofstay,
                         Predicted = predicted_counts)

# Prediction error metrics
prediction_error <- test_data$lengthofstay - predicted_counts
residuals_data <- data.frame(Index = seq_along(prediction_error),
                             Residuals = prediction_error)

# Calculate performance metrics
mae <- mean(abs(prediction_error))
rmse <- sqrt(mean(prediction_error^2))
mean_error <- mean(prediction_error)
median_error <- median(prediction_error)

cat("\nTest Set Performance Metrics:\n")
cat("Mean Absolute Error (MAE):", mae, "\n")
cat("Root Mean Squared Error (RMSE):", rmse, "\n")
cat("Mean Prediction Error:", mean_error, "\n")
cat("Median Prediction Error:", median_error, "\n")

# R-squared on test set
ss_res <- sum(prediction_error^2)
ss_tot <- sum((test_data$lengthofstay - mean(test_data$lengthofstay))^2)
r_squared_test <- 1 - (ss_res / ss_tot)
cat("Test Set R-squared:", r_squared_test, "\n")
fitted_values_train <- fitted(final_model)
residuals_data_train <- data.frame(Fitted = fitted_values_train,
                                   Residuals = residuals(final_model))

grid.arrange(
  # Residuals of predictions
  ggplot(residuals_data, aes(x = Index, y = Residuals, color = Residuals)) +
  geom_point(alpha = 0.6) +
  scale_color_viridis(option = "plasma") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals",
       x = "Index",
       y = "Residuals") +
  theme_linedraw(),

  # QQ-plot
  ggplot(data.frame(Residuals = prediction_error), aes(sample = Residuals)) +
```

```
  geom_qq(aes(color = Residuals)) +
  scale_color_viridis(option = "plasma") +
  geom_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals",
       x = "Theoretical Quantiles",
       y = "Sample Quantiles") +
  theme_linedraw(),

  nrow = 1
)
# Prediction intervals
pred_se <- predict(final_model, newdata = test_data, type = "link",
                   se.fit = TRUE)$se.fit

alpha <- 0.05
z_score <- qnorm(1 - alpha/2)

lower_pred <- exp(predictions - z_score * pred_se)
upper_pred <- exp(predictions + z_score * pred_se)

prediction_intervals <- data.frame(
  Predicted = predicted_counts,
  Lower_Pred = lower_pred,
  Upper_Pred = upper_pred
)

comparison <- data.frame(Actual = test_data$lengthofstay,
                         Predicted = predicted_counts)

prediction_intervals$Actual <- test_data$lengthofstay

grid.arrange(

  # Actual vs. predicted
  ggplot(comparison, aes(x = Actual, y = Predicted)) +
  geom_point(aes(color = Predicted), alpha = 0.6) +
  scale_color_viridis(option = "plasma") +
  geom_abline(slope = 1, intercept = 0, col = "red", linetype = "dashed") +
  labs(title = "Actual vs Predicted Values",
       x = "Actual Length of Stay",
       y = "Predicted Length of Stay") +
  theme_linedraw(),

  # Prediction intervals
  ggplot(prediction_intervals, aes(x = Predicted, y = Actual)) +
  geom_point(aes(color = Predicted), alpha = 0.6) +
  scale_color_viridis(option = "plasma") +
  geom_errorbar(aes(ymin = Lower_Pred, ymax = Upper_Pred), width = 0.2,
                alpha = 0.5, color = "blue") +
  geom_abline(slope = 1, intercept = 0, col = "red", linetype = "dashed") +
  labs(title = "Prediction Intervals",
       x = "Predicted Length of Stay",
       y = "Actual Length of Stay") +
```

```
  theme_linedraw(),

  nrow = 1
)
# Dataset summary statistics
summary(hospital)
# Dataset structure
str(hospital)
# Summary statistics by cluster
print(cluster_summary)
# ANOVA results for cluster differences
summary(anova_result)
# Pairwise t-test results
print(pairwise_result)
# Complete summary of full model (23 predictors)
summary(full_model)
# Complete summary of Lasso-selected model (21 predictors)
summary(lasso_selected_model)
# ANOVA comparison between Full and Lasso-selected models
print(anova_comparison)
# Variance Inflation Factors for all predictors
print(vif_values)
cat("\nMax VIF:", max_vif, "\n")
cat("Mean VIF:", mean_vif, "\n")
# 95% Confidence Intervals for Final Model Coefficients
print(conf_int)
# First 10 rows of prediction intervals on test set
head(prediction_intervals, 10)
```