Analysis of Factors Influencing Length of Hospital Stays

Angelina Cottone

2024-12-11

Introduction

Understanding the factors that lead to longer hospital stays is crucial for improving patient outcomes and optimizing resource allocation in healthcare settings. Longer hospital stays can increase the risk of complications and hospital-acquired infections, and place financial strain on facilities and patients alike. Identifying factors that influence LOS can help clinicians enhance quality of care and improve hospital operations.

Several factors, such as patient demographics, comorbidities, and laboratory results can potentially influence the length of hospital stays (LOS). Identifying and understanding the key contributors can provide valuable insights for clinicians and healthcare facilities to enhance efficiency and quality of care for patients. The primary question this research aims to answer is: *What are the key factors that influence the length of hospital stays for patients?* Additional sub-questions have been identified to refine the focus of the research question. These sub-questions include:

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

By addressing these questions, this study aims to identify what the most influential factors are in predicting LOS 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 rows and 28 columns, with information on comorbidities, laboratory results, and vital signs. This datset contains information on a variety of patient characteristics, including demographics, health conditions, laboratory results, and vital signs. Variables irrelevant to this analysis, including date columns, patient IDs, and facility IDs were excluded, leaving 24 variables for analysis.

2.2 Data Preprocessing

A check for missing data confirmed that no there were no missing values in the dataset, so no data-cleaning techniques or imputations were required. An initial summary of the dataset shows a mix of binary, continuous, and categorical variables:

- **Binary**: Indicators for health conditions including `asthma`, `pneum`, and `malnutrition`, as well as mental health indications including `depress` and `psychologicaldisordermajor`.

- **Continuous**: Laboratory results such as `hematocrit`, `neutrophils`, and vital signs such as `pulse`.
- **Categorical**: Demographic information including `gender` and `rcount`.
- **Response variable**: `lengthofstay` is a **discrete count** variable.

Since `lengthofstay` is a discrete count variable, representing the number of days a patient stays in the hospital as an integer, its mean and variance were checked for over dispersion. The variance (5.571) was found to be greater than the mean (4.001), suggesting that a Quasi-Poisson model is more appropriate for this dataset than a standard Poisson model, since it accounts for over dispersion.

2.3 Feature Engineering

Feature engineering was performed for select variables to improve model performance and interpretability. The following changes were made:

- `rcount` (readmission count) was converted into a binary variable: 0 for no readmissions in the past 180 days, and 1 for one or more readmissions.
- `gender` was also converted to binary, with `1` for male and `0` for female.
- Rare categories of `secondarydiagnosisnonicd9` (all categories beside `1`) were combined into "Other" categories to simplify analysis.

The prevalence of outliers in continuous variables was assessed using the interquartile range (IQR) method, where points beyond 1.5 times the IQR from the first and third quartile were identified as potential outliers. Results from the IQR method also found that 50.79% of the the dataset contains values that would be considered outliers. These outliers, given the context of healthcare, likely represent real-world variation (extreme cases and conditions). Therefore, outliers were retained to preserve the natural variation of the dataset and to prevent overfitting.

3 Exploratory Data Analysis (EDA)

3.1 Summary Statistics

The exploratory data analysis (EDA) began by analyzing the key characteristics of the dataset using descriptive statistical summaries, which includes mean, standard deviation, skewness, and more.

Key insights from the summary statistics:

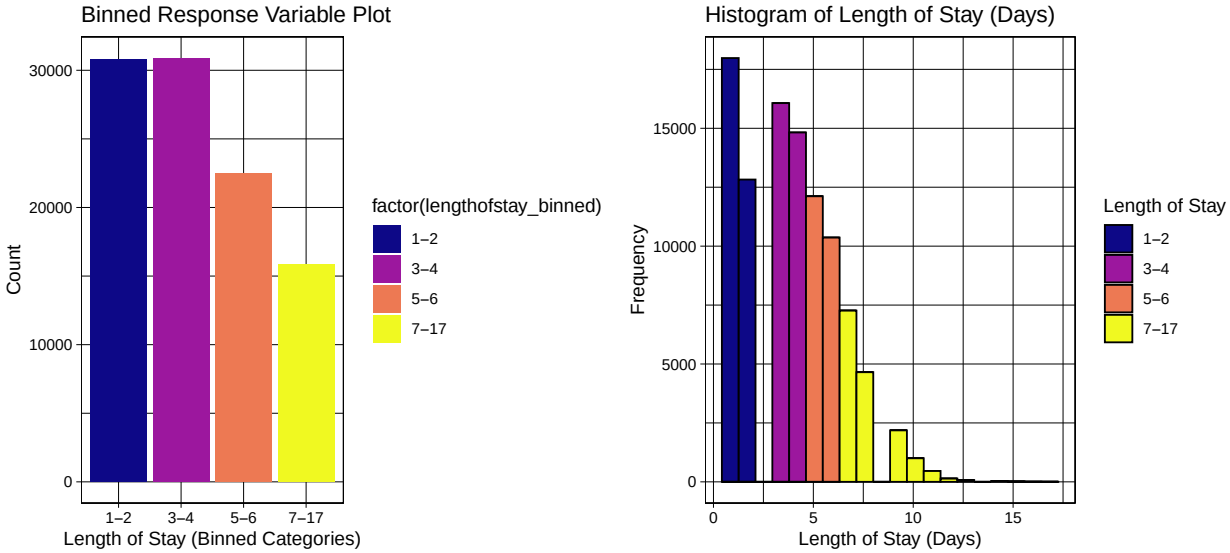
- `neutrophils` and `bloodureanitro` exhibit strong positive skewness, with values greater than 1, suggesting most observations are clustered towards the lower values.
- `hematocrit` and `lengthofstay` exhibit moderate positive skew with values greater than 0.5, suggesting most values are still on the lower side but not as extreme.
- `respiration` has a moderate negative skew with value less than -0.5, suggesting most values are on the higher side.
- `glucose` has the highest standard deviation of all the variables, meaning that data points for this variable display greater variability than those of other variables
- `bloodureanitro` has a biggest range of the variables, at 681.50.

3.2 Response Variable Binning

To facilitate visualization of the response and predictor variables, `lengthofstay` was binned into four intervals.

Since LOS is skewed toward shorter stays, the intervals ("1-2", "3-4", "5-6", and "7-17") were chosen to ensure relative balance of frequency across levels. This allows for clearer visual comparisons of LOS with other variables. By visualizing the distribution of the `lengthofstay` categories, it is evident that the variable is highly imbalanced, with a majority of patients have shorter stay ("1-2" and "3-4" days).

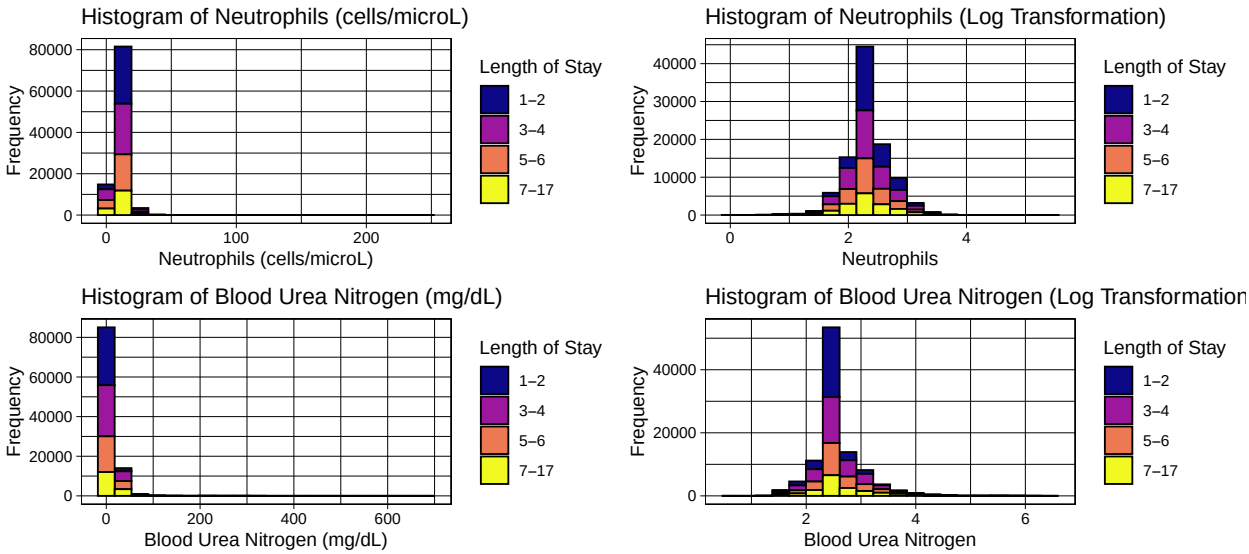
This binning strategy ensures that longer stays, which are less common, are still visible in the plots.



3.3 Visualizing Distributions of Variables

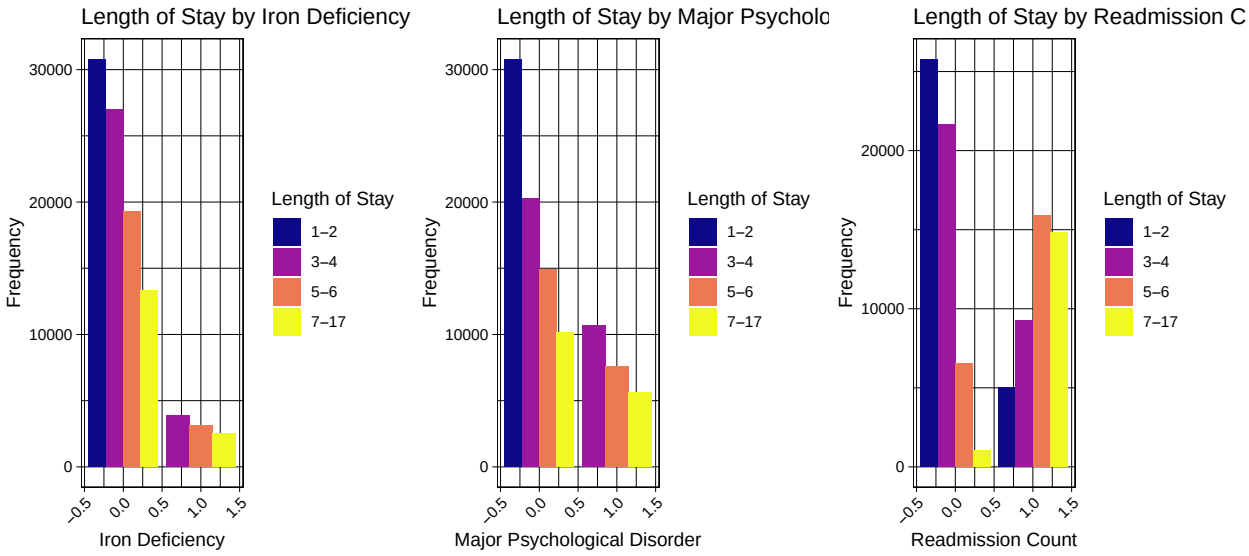
To further assess the distribution of predictor variables in relation to LOS, various visualization methods were used. For continuous variables, histograms were generated, filled by the binned `lengthofstay` variable. These plots confirmed the skewness of certain variables found previously:

- `hematocrit` shows a slight right skew, indicating a higher concentration of lower values.
- `neutrophils` and `bloodureanitro` exhibit more extreme positive skews, with most values being clusters to left of the plot.
- `respiration` exhibits a left skew, with most values lying on the higher end of the plot. To normalize skewed distributions, log transformations were applied to the positively skewed variables (`hematocrit`, `neutrophils`, and `bloodureanitro`), while `respiration` was squared to address its negative skew. All other continuous variables showed a normal distribution of values.



Bar plots were used to visualize the distributions of binary and categorical variables, also filled by the binned `lengthofstay` variable. These plots revealed several trends:

- For all health and mental health indicator (binary) variables, length of stays from 1-2 days were present if the patient did not have that condition, but all stays were three days or longer if the patient did have that condition.
- The bar plot for readmission count also show that length of stays for five days or more are most common in patients that have at least one readmission in the past 180 days compared to those that don't. Short stay lengths (1-4 days) are also less common in those with recent readmissions.



3.4 Principal Component Analysis (PCA)

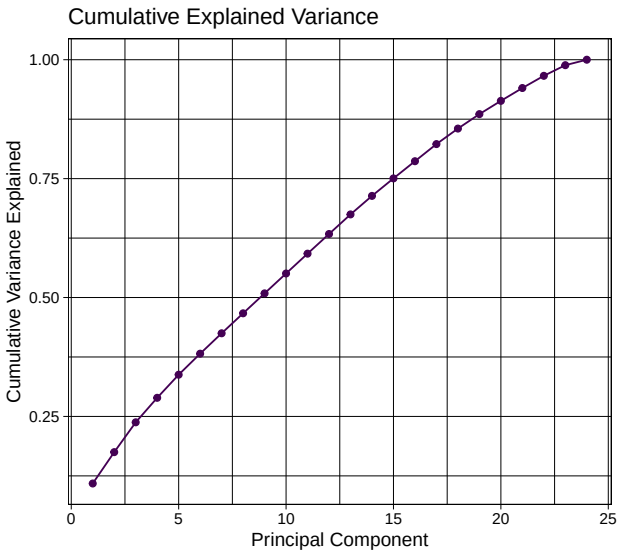
Principal Component Analysis (PCA) was conducted to explore the underlying structure of the numerical variables in the data. Scaling was applied to ensure that the differing units and scales of variables did

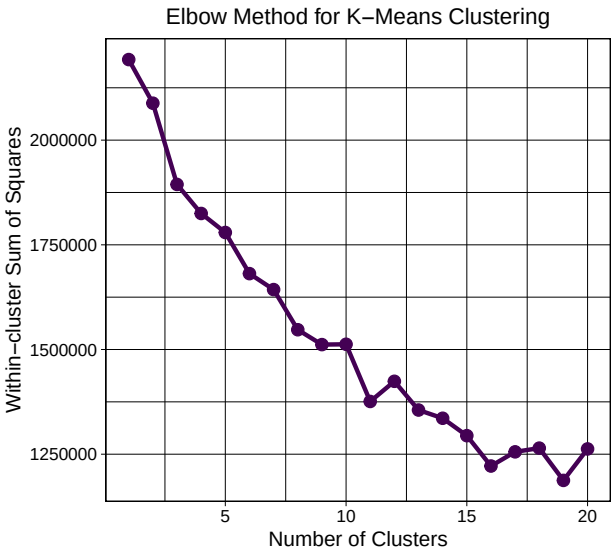
not influence the analysis disproportionately. The goal was to reduce the dimensionality of the data while retaining as much variance as possible. A cumulative variance plot was generated to see how much variance is explained by the first few principal components. The plot did not show a distinct increase or elbow point, indicating that many components would be needed to capture the variance. This suggests that the dataset is quite complex and several components would be needed.

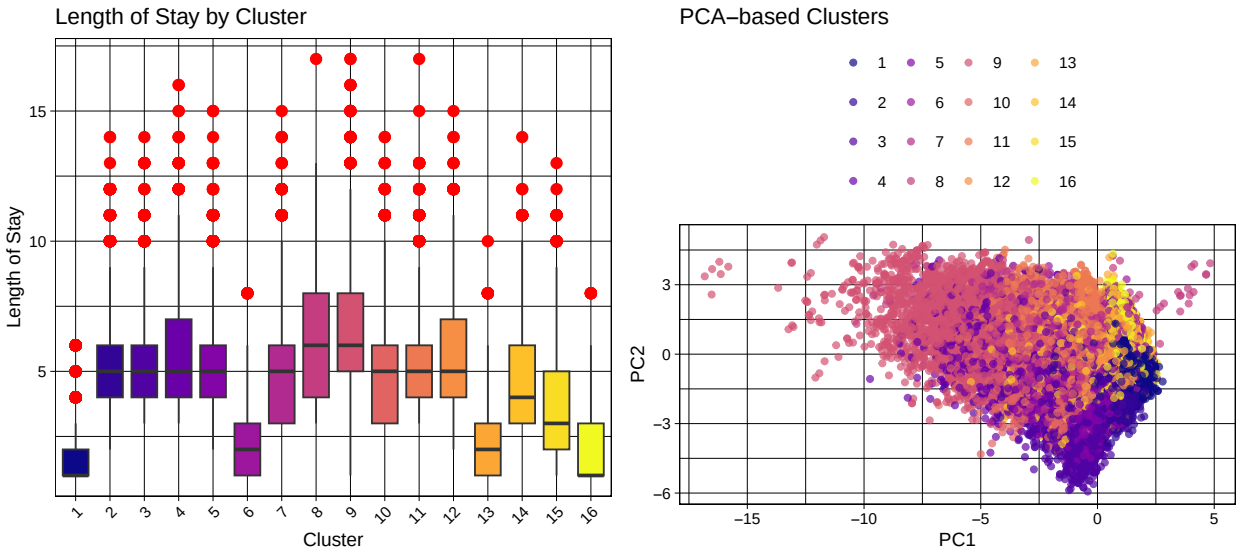
To identify potential clusters in the data, K-means clustering was applied using the first 20 principal components. The plot for the elbow method displayed more than one distinct 'elbow', it was decided that 16 clusters would be an adequate number to capture the most variance possible while still reducing dimensionality. After applying K-means clustering for 16 clusters, the results were visualized using a scatterplot of the first two principal components. The data was colored according to its cluster.

The relationship between the clusters and length of stay was assessed using boxplots, which helped identify which clusters corresponded to shorter or longer stays. Clusters 1 and 16 had the shortest average length of stay (around 1 day), and cluster 8 and 9 had the longest average stay (around 6 days). Clusters 6 and 13 had stays averaging around 2 days, while cluster 15 averaged 3 days and cluster 14 averaged 4 days. All other clusters had averages around 5 days.

To better understand the variation in `lengthofstay` for each cluster, summary statistics were generated. These summary statistics supported what was seen in the boxplot, but also shows that cluster 1 had the lowest average length of stay overall at 1.70 days, while cluster 9 had the highest average at 6.46 days.







3.5 Correlation Matrix

A correlation matrix was generated for all numeric variables to identify any strong relationships between them. Several positive correlations were identified, including:

- `psychother`, `dialysisrenalendstage`, `bloodureanitro`, and `malnutrition`
- `malnutrition` and `irondef`, `depress` and `psychologicaldisordermajor`
- `lengthofstay` and `psychologicaldisordermajor`, `rcount_binary` and `lengthofstay`

A negative correlation was also observed between `hematocrit` and `hemo`.