Study on ICU Stay Prediction in People with Acute Traumatic Spinal Cord Injury

MATH 2130: Final Project Report

By: Aneka Balakumar, Tien Hoang, Xidong Liu, Julianna Manalo, Xuan Quynh Vo

Overall Research Topic:
Enhancing ICU Length of Stay Prediction in People With Acute Traumatic Spinal Cord Injury At The Neck Through a Comparative Analysis with the APACHE-IV Score

SUBTOPIC 1 - factors that influence ICU stay duration

Background/introduction of problem:

In examining this subtopic, we attempted to determine which factors contribute the most to the length of ICU stays and which factors are seemingly irrelevant. Understanding the factors that impact ICU stay is greatly beneficial as it greatly improves length of stay estimations.

Data analysis steps:

Before using our data, we had to transform it since important variables are scattered across different csv files and our dataset included patients who had other injuries that were not spinal cord related. We began by only picking patients from "diagnoses_icd.csv" with spinal cord injuries among all the recorded patients by filtering for patients that have icd_codes related to spinal cord injuries at the neck using both ICD 9 and ICD 10 codes. Some patients with spinal cord injuries at the neck also have other kinds of spinal cord injuries (e.g. at other parts of their body), thus we created a new column called "others_icd" that stores patients' other spinal cord injuries except injuries at the neck. From there, we added the important identifying variables of each patient from the other csv files based on the patient unique identifier (subject_id) and hospitalization unique identifier (hadm_id) of patients with spinal cord injury. Filtering for unique patients, we found that there were 264 unique patients with spinal cord injury at the neck. This will be our dataset to work with.

We then conducted an exploratory data analysis. Using a correlation matrix which we visualized via a heatmap (Appendix A-1), we selected variables with meaningful predictive power to further visualize. Further exploring our dataset, we plotted some of the meaningful categorical variables via boxplot. Using boxplots we plotted the Length Of Stay (LOS) by gender, and the LOS by marital status (including "divorced", "married", "single", "widowed" and "0") (Appendix A-2). Adding onto that, we used histograms to visualize the distribution of patient race (Appendix A-3) and patient marital status (Appendix A-4). For numerical variables, such as age, we used a regression plot to visualize the relationship between it and LOS (Appendix A-5).

Main results/conclusions:

Although correlation analysis and visualizations such as heatmaps and boxplots provide valuable exploratory insights, they do not imply causation. More robust statistical modelling would be needed to more confidently identify predictors of ICU stay length and control for confounding variables. For this, we chose to use XGBoost, a tree-centered machine-learning algorithm. After training our XGBoost model with our dataset but changing to follow one hot encoding format, we extracted the feature importance scores from the trained model for the top 15 most important features and plotted them with a bar plot for easy visualization.

As a result, we found that important factors influencing ICU length of stay in spinal cord injury patients are as follows: complete spinal lesion, admitted to ICU for observation, closed fractured C5-C7 with injury, marital status married … (See Appendix A-6 for further variables). Interestingly, social variables like marital status proved to be significant alongside clinical indicators.

Furthermore, from our box plots, regression plots, and heatmap we discovered that age interestingly does not show any clear relationship with LOS, widowed and single patients have longer ICU stays, emergency cases exhibited more variability in LOS, Medical ICUs demonstrated higher LOS variance than Cardiac units, and gender and race only had mild variations.

Drawbacks of the Analysis Performed and Any Concerns:

One major drawback of our dataset is that it assumes the first recorded spinal cord injury is the most relevant or severe one, which may not accurately reflect the clinical reality. In many cases, patients suffer from multiple spinal cord injuries or related complications and relying solely on the first listed diagnosis could lead to misrepresentation of the patient's condition. This simplification may overlook the cumulative impact of multiple injuries that are recorded later in the diagnosis sequence but could be more influential in determining ICU length of stay.

SUBTOPIC 2 - Creating a predictive model for ICU stay duration

Background/introduction of problem:

This subtopic outlines the primary goal of our project: developing a model that can accurately predict a patient's length of stay in the ICU based on their pre-existing conditions. By understanding the factors that influence length of stay of hospital patients, it will allow us to make a prediction model that can prove to be better than current systems in place. An accurate model would be highly valuable to the healthcare system, enabling hospitals to better optimize capacity, estimate staffing, reduce unnecessary ICU stays, and enhance overall patient flow.

Data analysis steps:

Visualizing the distribution of the length of stay (LOS) among unique patients with spinal cord injuries with a histogram, we find that the length of stay is severely right-skewed (Appendix B-1) with the average LOS of 7.74 days, indicating that most patients had short stays of around this amount of days, but a rare number of them stayed significantly longer. The standard deviation of the distribution of LOS is 10.24 days. From this graph and other summary statistics, we noticed some extreme outliers in our data, like the max LOS being 99 days. Therefore, we applied log transformation to the LOS variable to minimize the effects of outliers, this became a new column "los_log", which we visualized through a histogram. This led to a graph that is less skewed with a much smaller range and mean of 1.72, standard deviation of 0.91, min of 0.21, max of 4.61 (Appendix B-2).

Plotting the distribution of the emergency department stay duration via the use of another histogram, we find that it is also highly right skewed (Appendix B-3) with average emergency department stay duration of 5.89 hours and standard deviation of 5.17 hours.

As most regression models cannot handle categorical values correctly, we transformed our dataset into one-hot encoding format through the use of pivot tables, leaving us with a final data frame with 264 rows × 1601 columns.

Lastly, because our final dataset is so small (264 rows), we decided to use data augmentation to double the size of the dataset from 264 to 528. We did this by duplicating our dataset and adding some noise to the "ed_duration_hours" column (5% noise), the ed_duration_hours column that tells us, in hours, a patient's emergency department stay duration.

Main results/conclusions:

After creating our final dataset with all categorical values following a one hot encoding format and enlarging the dataset to double its size, We implemented XGBoost, a tree-based machine learning algorithm, which uses decision trees as base learners and employs a gradient boosting framework to combine them, to predict the log-transformed ICU Length of Stay (LOS) for patients with spinal cord injuries. We configured the model to build 100 trees during training (n_estimators=100), with a maximum depth of 4 to reduce the risk of overfitting, and a learning rate

of 0.1 to improve generalization. Model performance was evaluated using R^2 and RMSE scores, calculated via 10-fold cross-validation to minimize overfitting.

The resultant average R^2 score for the model was 0.518, which means that the model can explain approximately 51.8% of the variance in the log-transformed LOS. This is considered a moderate score. While it indicates that the model captures a fair amount of the variability in LOS, it also shows there is still significant unexplained variance.

The average RMSE score of the model is 0.62, which represents the average error between the predicted and actual log-transformed LOS values. When back-transformed to the original LOS scale, this corresponds to an error of about 1.8 days.

Drawbacks of the Analysis Performed and Any Concerns:

The model's effectiveness may be limited by the dataset's size and variety. With just 264 rows and around 1,600 features, it might not be enough to fully reflect the complexity of predicting patient Length of Stay (LOS), particularly given how much hospital stays can vary. While data augmentation is applied by slightly altering one continuous variable to double the dataset, this method alone may not be enough to capture the full range of variability in patient data.

Additionally, the model's score remains low and does not consider patients with multiple coexisting conditions. Although we've grouped all diagnoses under a general category called 'other_icd', we were unable to expand the model to include more predictors.

SUBTOPIC 3 - Comparison with APACHE-IV

Background/introduction of problem:

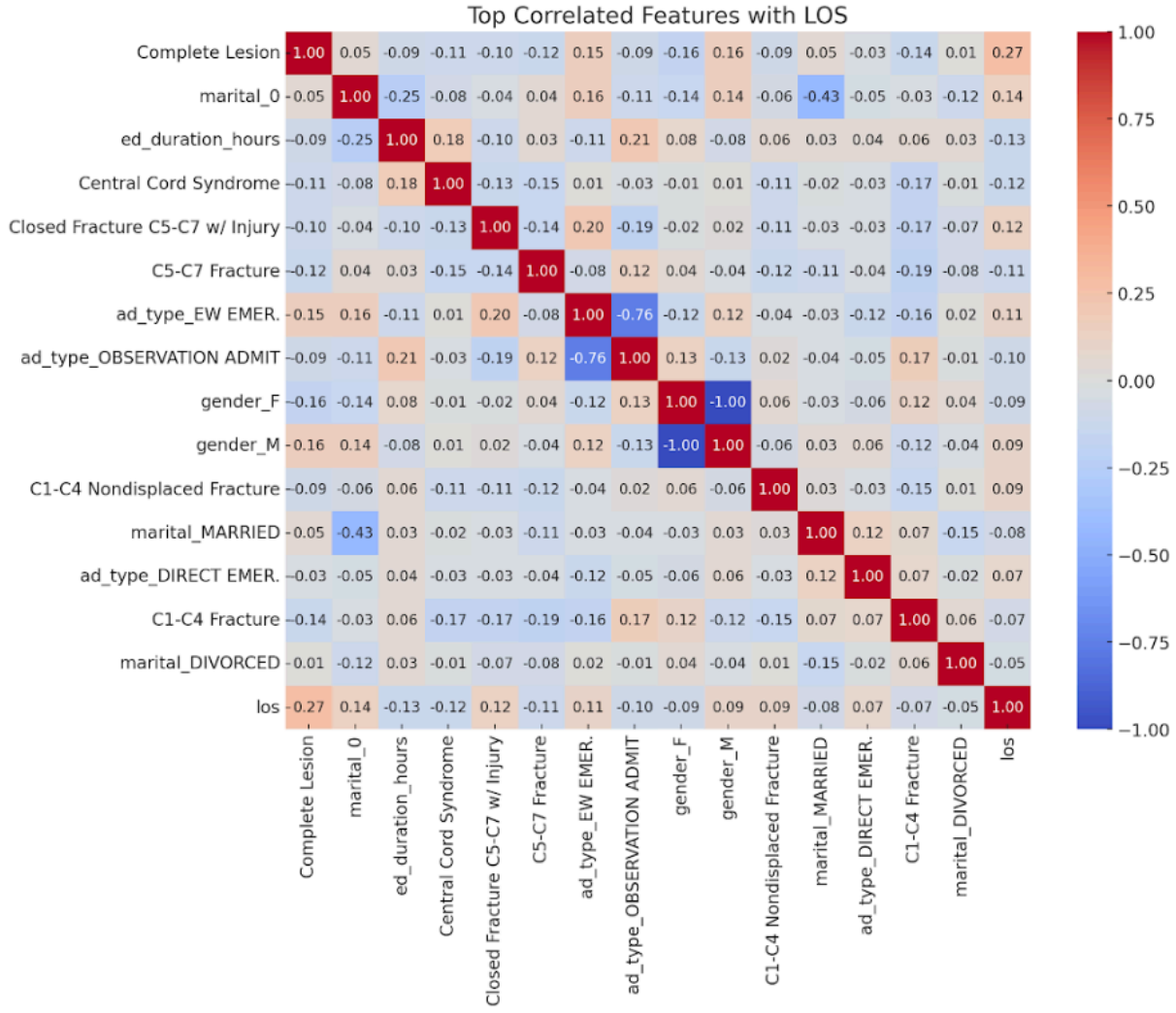
The APACHE IV, or Acute Physiology and Chronic Health Evaluation IV, is a model used in intensive care units (ICUs) to assess the severity of illness and predict mortality in critically ill patients. It's a scoring system that incorporates various physiological and demographic factors to estimate a patient's risk of death and length of stay in the ICU. Our last goal was to compare the scores of the model we have developed with the scores of the APACHE-IV model.

Why We Omitted This:

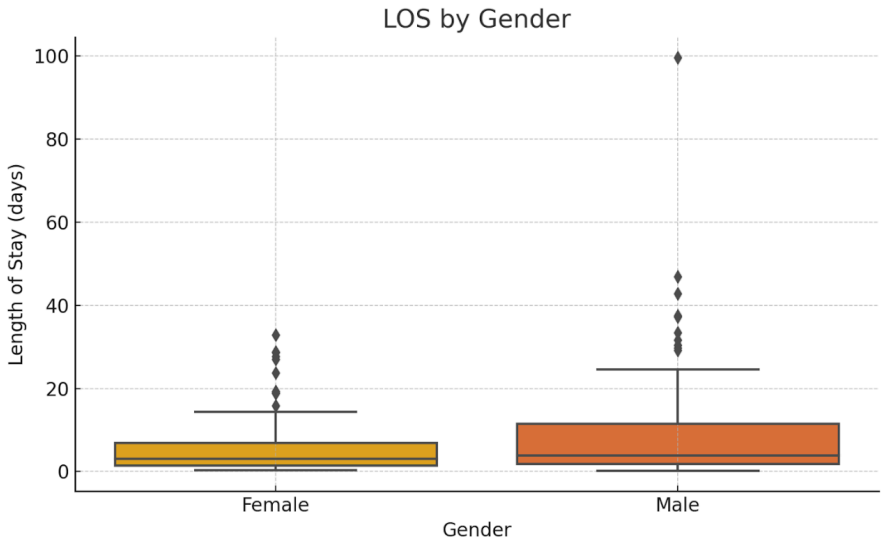
Although this was one of our goals in the midterm, we have decided against a comparison of our model with the APACHE-IV model. This is because the APACHE-IV score is a general ICU patient predictor, with a model specifically trained on spinal cord injury patients, especially those with neck injuries, we believe that the two may not be directly comparable. The APACHE-IV score does not capture the unique factors associated with spinal cord injuries, so concluding without accounting for differences in patient populations and relevant clinical variables could be misleading.

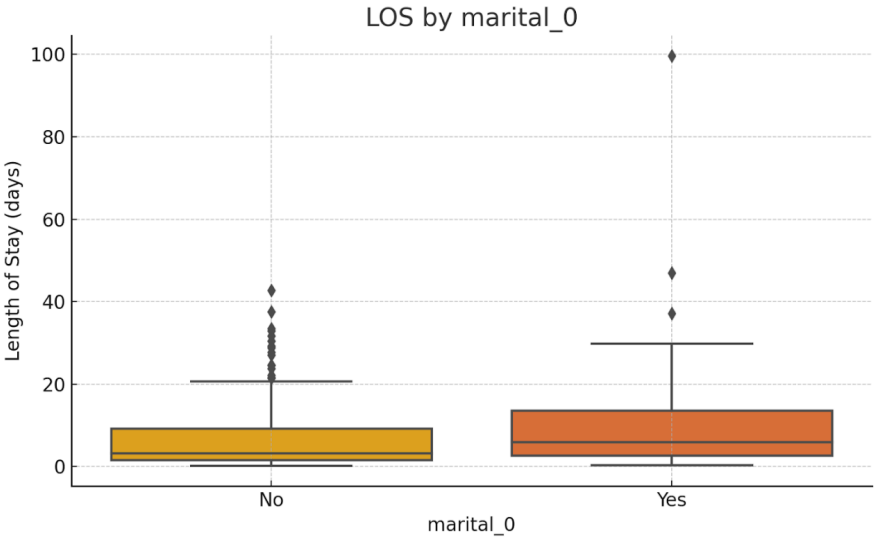
APPENDIX

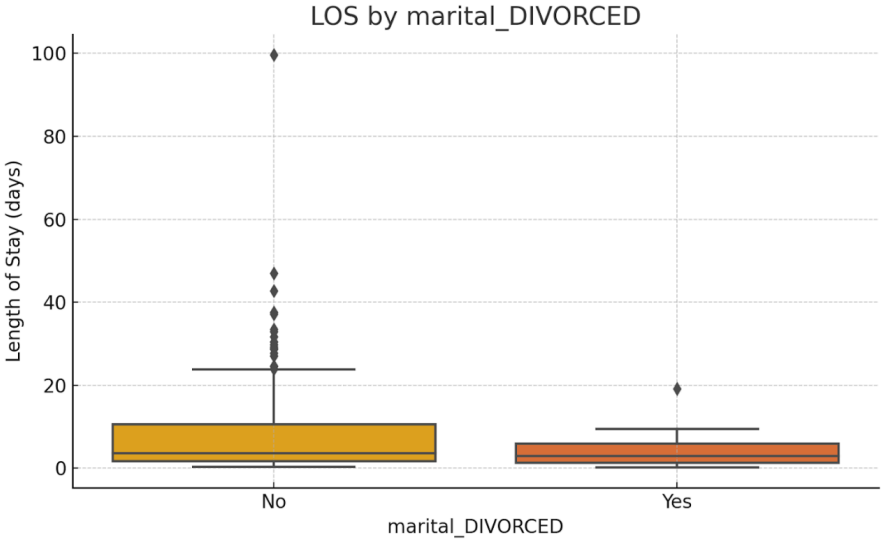
A-1

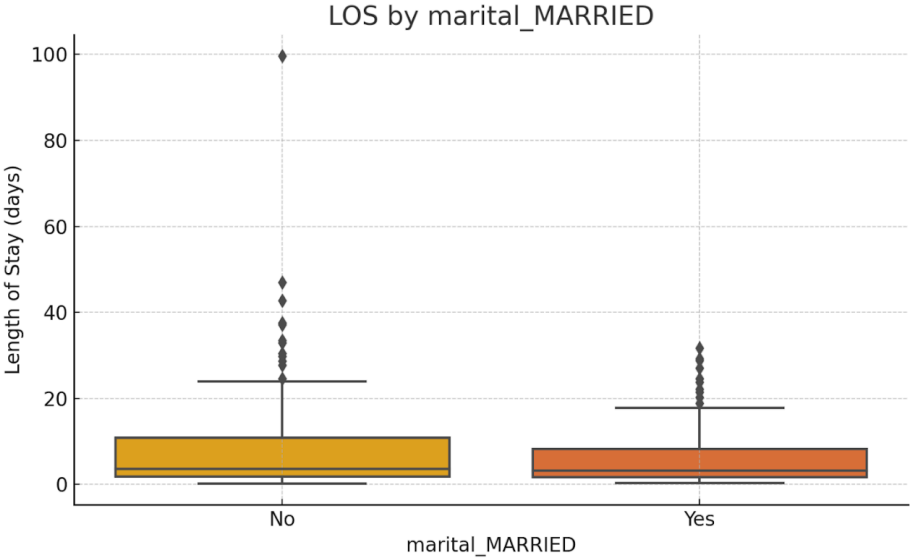

Top Correlated Features with LOS
Complete Lesion 1.00 0.05 -0.09 -0.11 -0.10 -0.12 0.15 -0.09 -0.16 0.16 -0.09 0.05 -0.03 -0.14 0.01 0.27
marital_0 0.05 1.00 -0.25 -0.08 -0.04 0.04 0.16 -0.11 -0.14 0.14 -0.06 -0.43 -0.05 -0.03 -0.12 0.14
ed_duration_hours -0.09 -0.25 1.00 0.18 -0.10 0.03 -0.11 0.21 0.08 -0.08 0.06 0.03 0.04 0.06 0.03 -0.13
Central Cord Syndrome -0.11 -0.08 0.18 1.00 -0.13 -0.15 0.01 -0.03 -0.01 0.01 -0.11 -0.02 -0.03 -0.17 -0.01 -0.12
Closed Fracture C5-C7 w/ Injury -0.10 -0.04 -0.10 -0.13 1.00 -0.14 0.20 -0.19 -0.02 0.02 -0.11 -0.03 -0.03 -0.17 -0.07 0.12
C5-C7 Fracture -0.12 0.04 0.03 -0.15 -0.14 1.00 -0.08 0.12 0.04 -0.04 -0.12 -0.11 -0.04 -0.19 -0.08 -0.11
ad_type_EW EMER. -0.15 0.16 -0.11 0.01 0.20 -0.08 1.00 -0.76 -0.12 0.12 -0.04 -0.03 -0.12 -0.16 0.02 0.11
ad_type_OBSERVATION ADMIT -0.09 -0.11 0.21 -0.03 -0.19 0.12 -0.76 1.00 0.13 -0.13 0.02 -0.04 -0.05 0.17 -0.01 -0.10
gender_F -0.16 -0.14 0.08 -0.01 -0.02 0.04 -0.12 0.13 1.00 -1.00 0.06 -0.03 -0.06 0.12 0.04 -0.09
gender_M -0.16 0.14 -0.08 0.01 0.02 -0.04 0.12 -0.13 -1.00 1.00 -0.06 0.03 0.06 -0.12 -0.04 0.09
C1-C4 Nondisplaced Fracture -0.09 -0.06 0.06 -0.11 -0.11 -0.12 -0.04 0.02 0.06 -0.06 1.00 0.03 -0.03 -0.15 0.01 0.09
marital_MARRIED -0.05 -0.43 0.03 -0.02 -0.03 -0.11 -0.03 -0.04 -0.03 0.03 0.03 1.00 0.12 0.07 -0.15 -0.08
ad_type_DIRECT EMER. -0.03 -0.05 0.04 -0.03 -0.03 -0.04 -0.12 -0.05 -0.06 0.06 -0.03 0.12 1.00 0.07 -0.02 0.07
C1-C4 Fracture -0.14 -0.03 0.06 -0.17 -0.17 -0.19 -0.16 0.17 0.12 -0.12 -0.15 0.07 0.07 1.00 0.06 -0.07
marital_DIVORCED 0.01 -0.12 0.03 -0.01 -0.07 -0.08 0.02 -0.01 0.04 -0.04 0.01 -0.15 -0.02 0.06 1.00 -0.05
los 0.27 0.14 -0.13 -0.12 0.12 -0.11 0.11 -0.10 -0.09 0.09 0.09 -0.08 0.07 -0.07 -0.05 1.00
Complete Lesion
marital_0
ed_duration_hours
Central Cord Syndrome
Closed Fracture C5-C7 w/ Injury
C5-C7 Fracture
ad_type_EW EMER.
ad_type_OBSERVATION ADMIT
gender_F
gender_M
C1-C4 Nondisplaced Fracture
marital_MARRIED
ad_type_DIRECT EMER.
C1-C4 Fracture
marital_DIVORCED
los
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

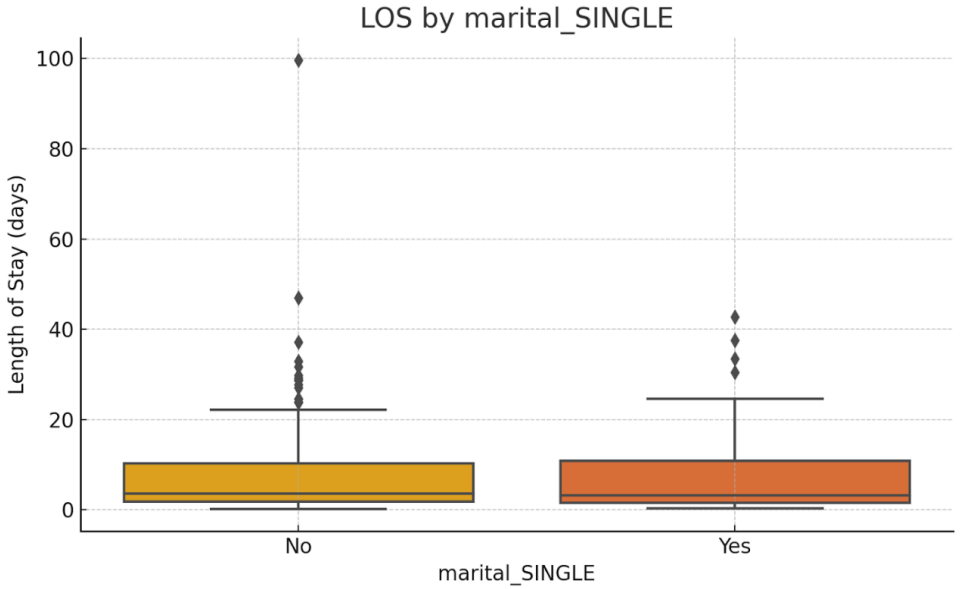
A-2

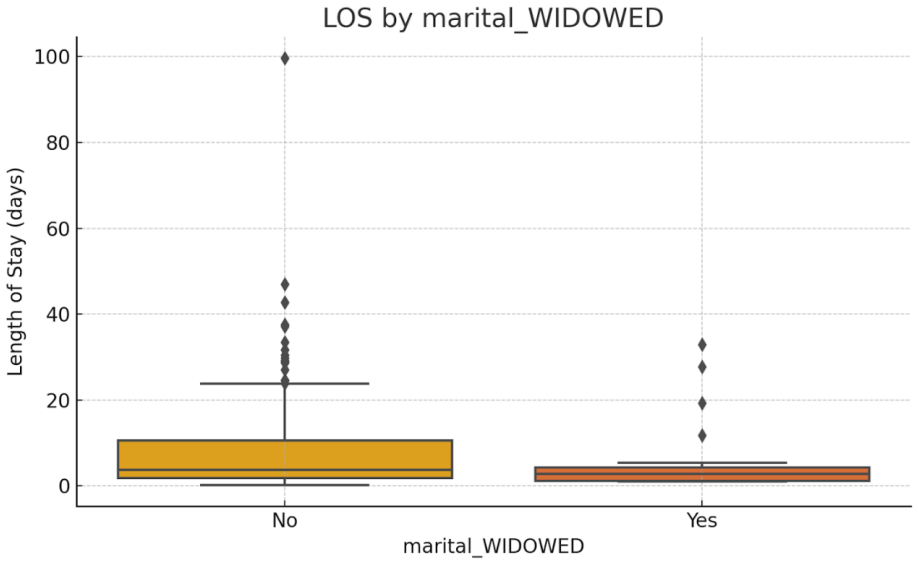

LOS by Gender
Length of Stay (days)
0
20
40
60
80
100
Female
Male
Gender


LOS by marital_0
Length of Stay (days)
0
20
40
60
80
100
No
Yes
marital_0

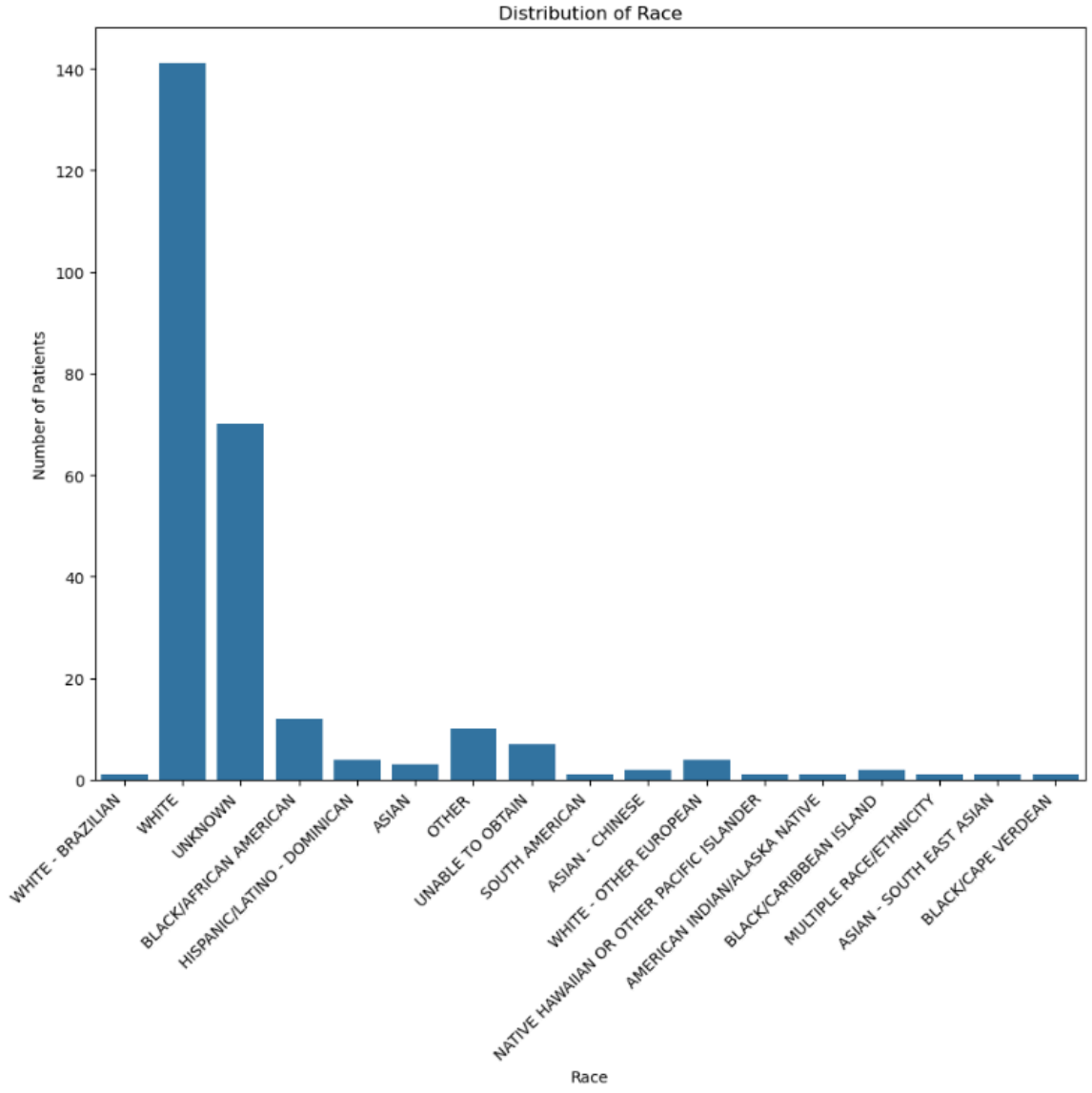

LOS by marital_DIVORCED
Length of Stay (days)
0
20
40
60
80
100
No
Yes
marital_DIVORCED


LOS by marital_MARRIED
Length of Stay (days)
0
20
40
60
80
100
No
Yes
marital_MARRIED

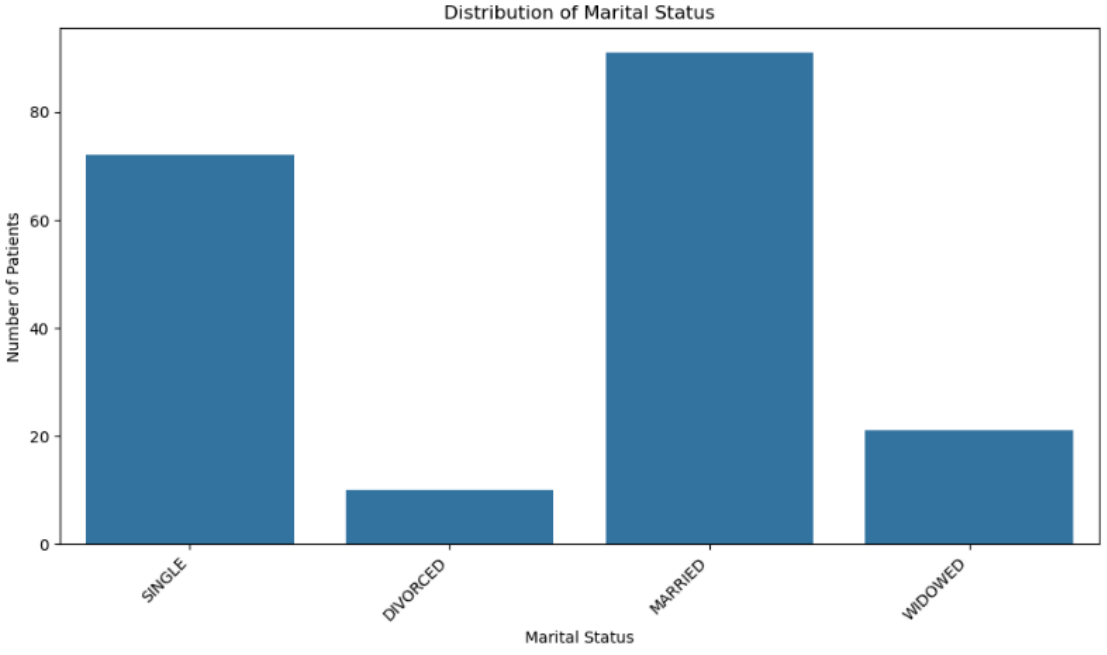

LOS by marital_SINGLE
Length of Stay (days)
100
80
60
40
20
0
No
Yes
marital_SINGLE


LOS by marital_WIDOWED
Length of Stay (days)
100
80
60
40
20
0
No
Yes
marital_WIDOWED

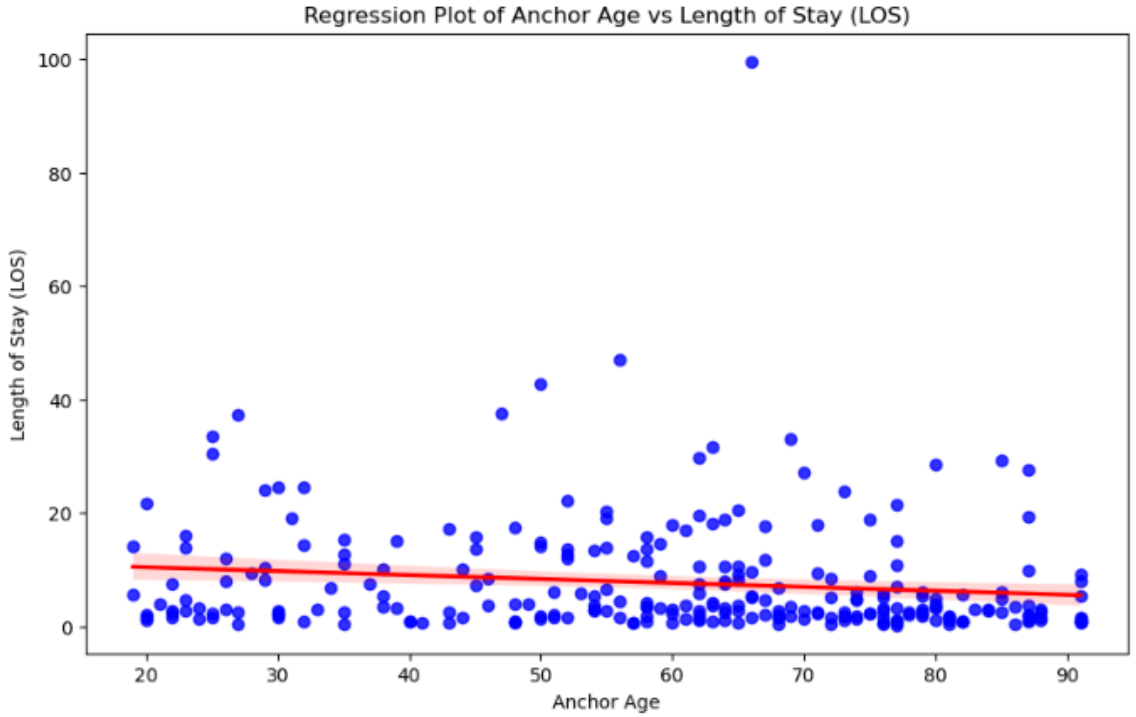
A-3


Distribution of Race
Number of Patients
0
20
40
60
80
100
120
140
WHITE - BRAZILIAN
WHITE
UNKNOWN
BLACK/AFRICAN AMERICAN
HISPANIC/LATINO - DOMINICAN
ASIAN
OTHER
UNABLE TO OBTAIN
SOUTH AMERICAN
ASIAN - CHINESE
WHITE - OTHER EUROPEAN
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER
AMERICAN INDIAN/ALASKA NATIVE
BLACK/CARIBBEAN ISLAND
MULTIPLE RACE/ETHNICITY
ASIAN - SOUTH EAST ASIAN
BLACK/CAPE VERDEAN
Race

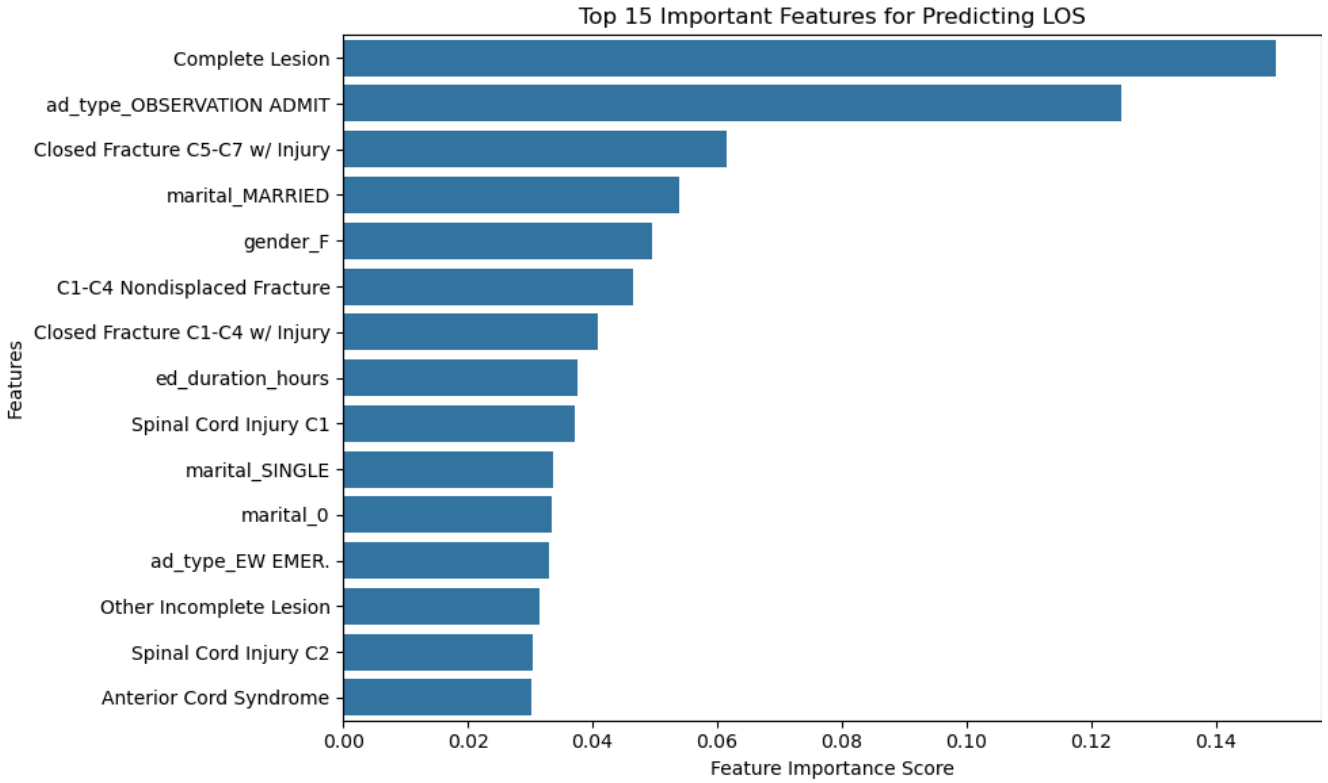
A-4


Distribution of Marital Status
Number of Patients
0
20
40
60
80
SINGLE
DIVORCED
MARRIED
WIDOWED
Marital Status

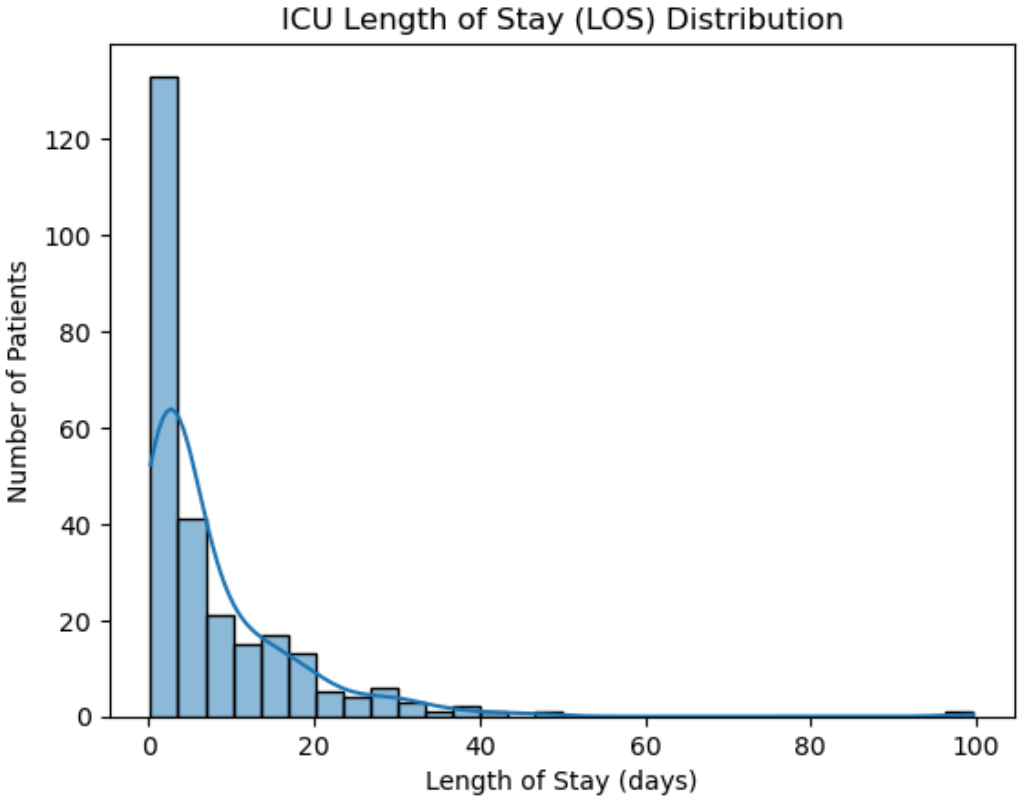
A-5


Regression Plot of Anchor Age vs Length of Stay (LOS)
Length of Stay (LOS)
Anchor Age
0
20
40
60
80
100
20
30
40
50
60
70
80
90

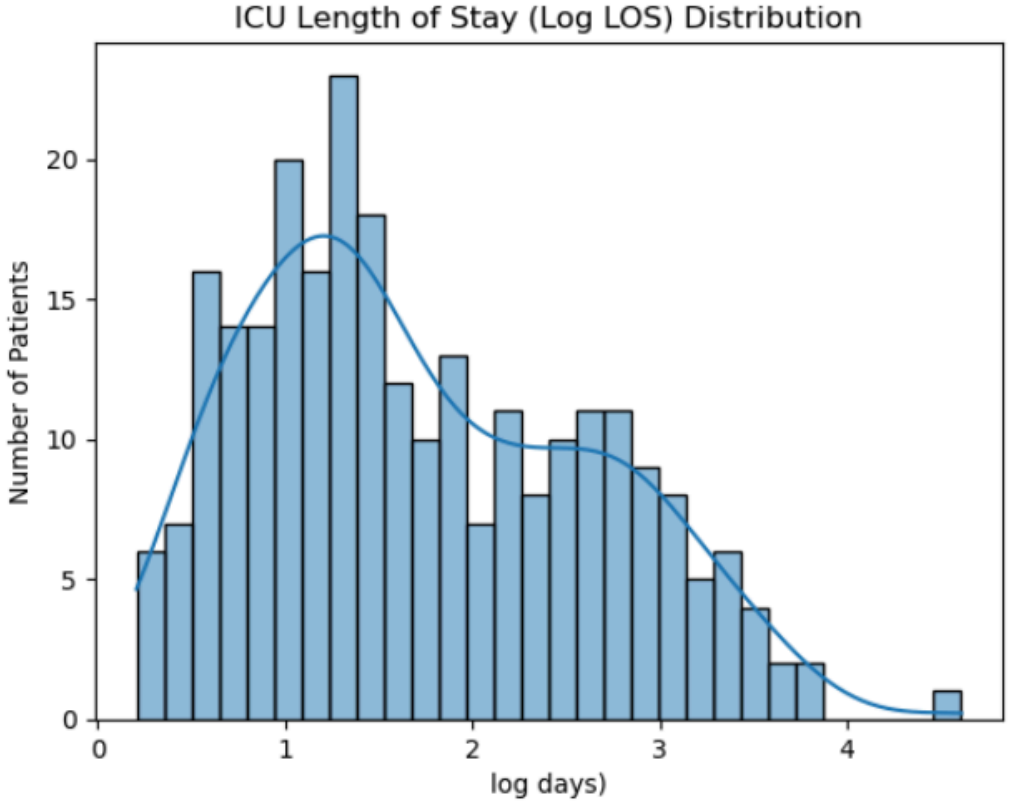
A-6


Top 15 Important Features for Predicting LOS
Features
Complete Lesion
ad_type_OBSERVATION ADMIT
Closed Fracture C5-C7 w/ Injury
marital_MARRIED
gender_F
C1-C4 Nondisplaced Fracture
Closed Fracture C1-C4 w/ Injury
ed_duration_hours
Spinal Cord Injury C1
marital_SINGLE
marital_0
ad_type_EW EMER.
Other Incomplete Lesion
Spinal Cord Injury C2
Anterior Cord Syndrome
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Feature Importance Score

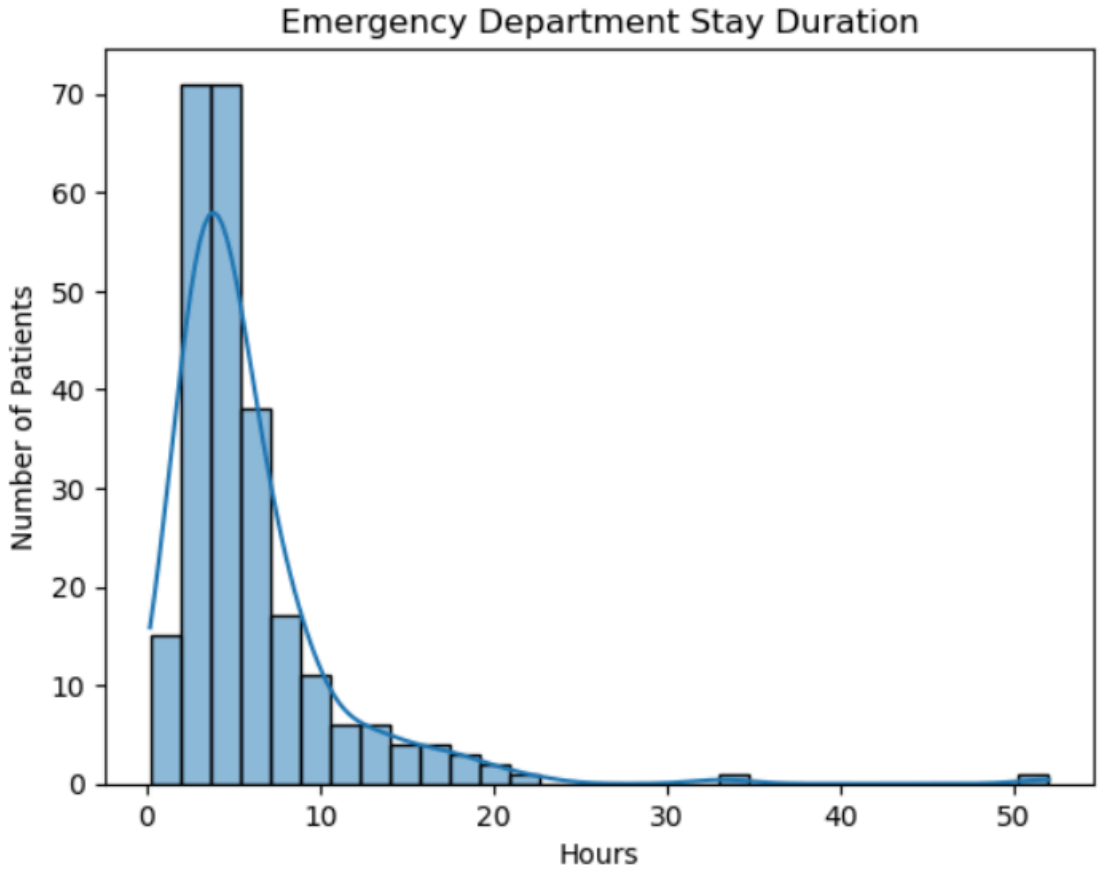
B-1


ICU Length of Stay (LOS) Distribution
Number of Patients
0
20
40
60
80
100
120
0
20
40
60
80
100
Length of Stay (days)

B-2


ICU Length of Stay (Log LOS) Distribution
Number of Patients
0
5
10
15
20
0
1
2
3
4
log days)

B-3


Emergency Department Stay Duration
Number of Patients
Hours
0
10
20
30
40
50
60
70