Predicting Length Of Stay For Spinal Cord Injuries In The ICU

By:
Stefan Jafry

Introduction/Executive Summary:

Our Project aimed to create a better machine learning model for ICU Length of Stay (LOS) prediction of spinal cord injury (SCI) patients to surpass the standard APACHE IV benchmark model ($R^2 = 0.21$). Based on the assistance of the MIMIC-IV database, we constructed a specialized dataset, ICU_SPINAL_PATIENT_FEATURES_VITALS_EXT, which integrated clinical, demographic, physiological, and laboratory data distributed across 26 interdependent tables. Strict SQL-based feature extraction, cleaning, and aggregation permitted the creation of a high-fidelity dataset thoughtfully prepared to meet SCI patients. Our exploratory data analysis (EDA) identified key clinical and socioeconomic factors influencing LOS, such as injury type, age, mortality, admission type, and diagnosis number. Skewness to the right, outliers, and non-linear associations made it essential to utilize CatBoost and LightGBM—two gradient-boosting models that are particularly suited to handle structured, heterogeneous data with high cardinality and non-normality. Data leakage from patients with multiple occurrences was prevented using GroupKFold cross-validation with subject_id as the group variable.

Our baseline models performed very well, substantially surpassing the APACHE IV benchmark. LightGBM reached an R^2 of 0.5661 and CatBoost 0.5523, which fared well on MAE, MedAE, and RMSE. We enhanced model stability using log transformations, Huber loss functions, and SHAPE-based residual analysis. Residual heteroskedasticity and outliers still lingered due to the irreducible variance of LOS motivated by mortality and non-clinical delay.

We built cluster-specific niche models from PCA-informed K-means clustering to reduce heterogeneity between patient subgroups. These models, on surgical referrals, emergency admissions, and inter-hospital transfers, had improved R^2 values (up to 0.70 for Cluster 0) after the same transformations.

Data Analysis:

SQL/DataCleaning:

The original MIMIC IV database contained 26 interlinked tables that range from diagnostic records and clinical events to microbiological observations. To build a predictive model, extracting, joining, and transforming relevant data into a single cohesive tabular format was necessary, resulting in ICU_SPINAL_PATIENT_FEATURES_VITALS_EXT. This table only included ICU patients with spinal injuries and descriptive data such as gender, race, LOS, age, and insurance, just to name a few. First, identifiers and time anchors were extracted (subject_id,stay_id, hadm_id, los, intime, outtime). Identifier columns are required for tracking individual patients, while time anchors are our target variable for prediction. Next, spinal injury type was extracted, as certain injuries may affect LOS. Then, demographic and socioeconomic features were extracted, such as gender, anchor age, date of death, insurance, marital status, and race. Many of these columns, such as anchorage, are strong predictors of length of stay, so including them was a necessity. Vital signs such as heart rate and respiratory rate were aggregated into an average, minimum, and maximum as they may reflect important predictive factors such as heart complications or how SCI may impair respiratory functions, which can help indicate severity. Lastly, laboratory features such as creatinine and dx count were included, as more diagnoses usually indicate complex cases and longer LOS. Next, Data cleaning needs to be conducted. WBC has <1% coverage, making it an extremely sparse feature. Therefore, we will drop these rows. Next, numerical features are overall extremely clean, with 2-36 rows missing in columns, so inverse transformation was used to impute the data. We used inverse transformation because some numerical columns did not exhibit a normal distribution, so imputing the Mean, Median and Mode would introduce bias. Furthermore, the date of death will be turned into a binary indicator where 1=death and 0=survived so the model can properly interpret this feature. Next, extremely rare injuries that only happen to 1 or 2 patients will be aggregated into a rare injury column. This is done because the plots were practically horizontal lines and extremely high in our EDA (the box plot shows the bottom 10 injury types, with LOS as the Y-axis). This will reduce overfitting to those variables. The race columns are extremely specific, so they will be aggregated as well. Last admission type and admission location will be consolidated into broader more medically meaningful categories.

EDA:

(Figure 1): The length of stay graph exhibits positive skewness, with most stay time concentrated on the left. Furthermore, there is a long tail of high individual LOS values that reach over 2,000 hours. This skewness is further reflected in the standard deviation (144.29). Additionally, the mean (99.88) is much higher than the median (52.20). This means that outliers can affect our predictive model, and tree-based models are likely going to be our best option because of non-normality and outliers.

(Figure 2): For the top 10 injury types, the median LOS ranges between 50 to 120 hours (~2 to 5 days). Furthermore, there are many cases where LOS exceeds 200 hours, and some even reach 600 to 700 hours. Injuries such as spinal stenosis, lumbar region, and cervical region show wide IQR and more outliers. Injuries, such as Secondary malignant neoplasm of the brain and spinal cord, show a tighter IQR. Clinically speaking, injuries involving hardware complications may require more extensive ICU monitoring, which may help explain the longer and more variable LOS. The box plots for the bottom 10 injury types are almost non-existent, indicating how these injury types have very few patients and may indicate rare injuries. Overall, injury type has high predictive potential; however, rare injury types and categories may introduce noise, which must be handled. (another/rare category using frequency thresholds).

(Figure 3): In our first boxplot, a modest gender-based difference exists in ICU stay, with males having a slightly higher variance and median LOS. In our second boxplot, marital status is largely the same, with only subtle differences, such as a

slightly higher max for singles compared to widowed. For our third boxplot, Medicaid patients exhibit a higher median LOS, wider IQR, and more patients with extended ICU stays. This could be because Medicaid patients represent lower-income groups, although this is speculative.

(Figure 4): For the mortality bar plot, mortality increases steadily with age, especially after 60, where the slope becomes steeper, indicating age-related vulnerability. Mortality peaked sharply at age 91 (ages >89 are all listed as 91 in the MIMIC database). Our boxplot shows an overall decrease in LOS, variability, and max values as age increases. Combining both graphs, the observed lower length of stay (LOS) in older spinal cord injury patients is likely influenced by their significantly higher mortality rates, as evidenced by the steep increase in deaths with age in your dataset. Older patients often experience early ICU mortality due to greater frailty and comorbidities, which artificially truncates their LOS and skews the distribution downward. This creates a statistical bias where shorter stays do not reflect faster recovery but rather early death, negatively influencing predictive modeling.

(Figure 5): LOS varies significantly across admission types and race/ethnicity groups in the data set. Patients admitted under categories such as "SURGICAL SAME DAY ADMISSION," "EMER," and "OBSERVATION ADMIT" show the most prolonged and variable LOS. At the same time, those under "AMBULATORY OBSERVATION" and "DIRECT OBSERVATION" tend to have short, consistent stays, suggesting that the admission type captures procedural intensity and care duration well. Similarly, the race/ethnicity boxplot reveals substantial variability. At the same time, WHITE and BLACK/AFRICAN AMERICAN patients show relatively consistent LOS distributions, and specific subgroups like "HISPANIC OR LATINO – PUERTO RICAN" display unusually high LOS and outlier presence. This is because rare categories with few instances can skew model training if outliers dominate.

(Figure 6): Our correlation matrix shows no multicollinearity, which is a positive. Furthermore, lab results, vital signs, and diagnosis count demonstrate some correlation and predictive power; however, several negatively correlated columns, such as 'MIN_SP02', signify early mortality and the need to adjust for a more accurate model.

(**Figure 7):** For average heart rate, a very low heart rate equates to a longer LOS, while conversely, a very high average heart rate equates to a lower average LOS. The longest LOS is at extremely high values for average respiratory rate, indicating respiratory distress. This nonlinearity, combined with other nonlinear effects listed above, makes a linear regression model unfeasible. For average creatinine, high levels indicate complications and more extended ICU management. However, they drop after a 6.5+ threshold, indicating high mortality at severe renal dysfunction stages or a transition into palliative care or dialysis. In the box plot, avg_SpO_2 shows that patients with very low oxygen saturation (<85%) have significantly longer lengths of stay, indicating severe respiratory compromise. The rest of the distribution is relatively flat, suggesting that only critically low SpO_2 levels are strongly associated with extended ICU stays.

Clustering:

Clustering can be a challenge due to the high dimensionality of the dataset. In order to reduce dimensionality, principal component analysis(PCA) was conducted using k-mean clustering (see figure 8), and then kruskal-wallis tests were applied to see if the clusters significantly affect LOS. The main clusters we observed were: married patients with referrals for surgery, male patients who were admitted for emergencies, and hospital transfer patients with high diagnosis counts. However LOS was quite similar for all clusters with exception to outliers caused by our inverse transform sampling. We thought this could also be due to our inverse sample transform on white blood cell features, as they had the most missing values, however the results were similar to create some distrust in PCA. Pivoting, we decided to focus on pair clustering and analysing each centroid(see figure 9). We can also see that dx_count has the greatest difference for each cluster, causing LOS to be high for cluster 1, but this feature was also highly common amongst all clusters in PCA, thus we opted for using the PCA clusters for better comprehension and more complete/accurate clusters to our dataset.

Machine Learning Model Creation:

First we made a general machine learning model that was trained on the entire dataset. We are using CatBoost and LightGBM on this dataset because they are highly optimized gradient-boosting libraries that deliver state-of-the-art performance on structured, tabular data—precisely the kind of data we have in this spinal injury dataset. Both models handle mixed feature types very nicely. CatBoost offers native, automatic support for high-cardinality categorical features like diagnosis codes (long_title_grouped), making preprocessing easier while preventing overfitting through ordered boosting. LightGBM adds to this with swift training, tunable tuning parameters, and efficient handling of large quantities of features when categorical variables are properly label-encoded. Together, they form a powerful, accurate modeling method that is well suited to the complexity of Length of Stay (LOS) prediction while being interpretable using SHAPE-based interpretability. We used GroupKFold cross-validation because each patient (subject_id) can appear multiple times in the dataset due to multiple injuries. This ensures that all data from a given patient is assigned to either the training or validation set within a fold, preventing data leakage and producing a more realistic estimate of model performance on unseen patients. Last we analyzed residuals and performed a log transformation, feature engineering and a huber loss transformation to both models.

Evaluation:
CatBoost: RMSE= 3.9935, MAE=1.8380, MedAE=0.7686, R^2=0.5523, MAPE= 57.76%
LightGBM: RMSE=3.9214, MAE=1.8215, MedAE=0.7352, R^2=0.5661, MAPE= 57.53%
Overall both models outperformed the APACHE IV model(R^2=0.21). Overall the LightGBM model demonstrates a strong performance with especially low MedAE (0.7352) suggesting strong robustness to outliers and consistent accuracy across typical LOS cases. The MAPE of 57.53% indicates moderate reliability in relative error, with potential sensitivity to shorter LOS outliers. The CatBoost model achieved similar overall performance; it matched LightGBM in relative error, while offering the advantage of native handling of categorical variables. Overall, LightGBM outperformed CatBoost slightly across all metrics, however the difference is negligible. For the negatives both models exhibited high heteroscedasticity, extreme outliers in residuals, and several over and underprediction (figure 10).

Niche Model:
We again used LightGBM and CatBoost, but we created independent models for each cluster we discovered in our PCA, namely for referral, emergency, and transfer patients. We also performed residual analysis, which we did for our general model. (cluster 0: population of patients with referrals for surgery, Cluster 1: Emergency or urgent admission patients, Cluster 2: Transfer patients)

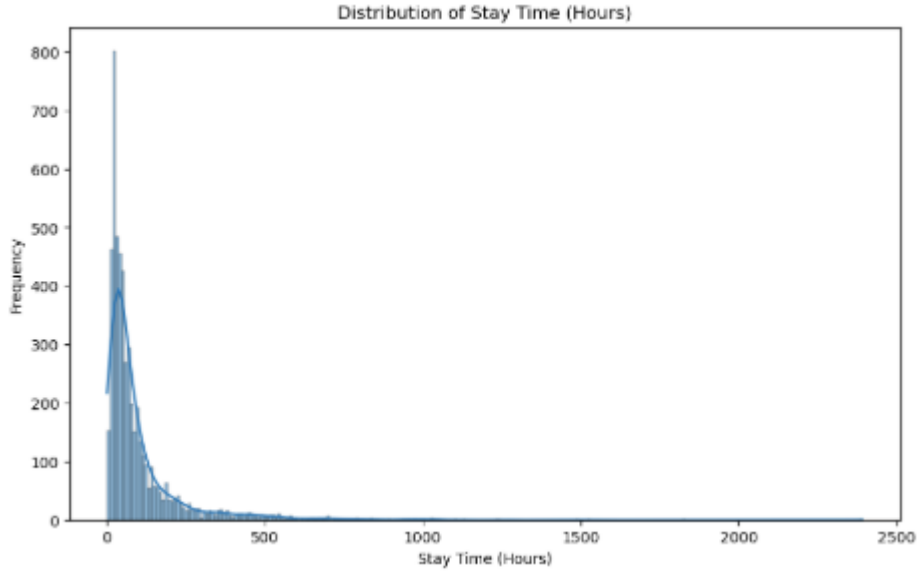
Niche Model Evaluation:
The main issue with our initial niche models was that our clusters had a lot of visible heteroscedasticity, non-normality, and extreme outliers. Furthermore, all models have poor predictive performance based on r-squared values of cluster 1: 0.42, cluster 2: 0.46, cluster 0: 0.62 for LightGBM and cluster 1: 0.41, cluster 2: 0.46, cluster 0: 0.61 for CatBoost. We used log transformations, same-feature engineering, and Hubert techniques to improve our performance and fix residuals. The result was significantly improved r-squared values of cluster 1: 0.54, cluster 2: 0.59, cluster 0: 0.70 for LightGBM, and cluster 1: 0.59, cluster 2; 0.62, cluster 0: 0. for CatBoost, as well as more homoscedasticity, normality, and fewer outliers across all models(figure 11). Since our threshold for model performance was an r-squared of 0.5, all of our niche models performed according to these expectations, with cluster 0 performing the best.

Drawbacks:
Our general machine-learning model still has heteroscedasticity, nonnormality, and outliers. Our model does not generalize well for long or short LOS and can over and under-predict values. LOS is inherently noisy and hard to predict; even with perfect vitals and labs, non-clinical delays also introduce irreducible variance. The MIMIC IV database is also missing some potential high-impact features, such as insurance approval delays, a substantial economic factor in the USA. If we had more time to perform more SQL and add more columns, our transformations would work much better in reducing these issues. Overall, our general model is under-informed due to data limitations. For our niche models, cluster 1's light GBM model underperformed compared to our general model, meaning that only CatBoost for cluster 1 would provide better predictions for that niche population. Moreover, certain factors, such as mortality, can skew our model as old patients with multiple injuries may pass away within a short time, which the model will interpret as a short LOS. Lastly, our model is trained on only data from one hospital, which can reduce generalization and performance when applied to other hospitals.

Conclusions:
Our project demonstrated convincingly that advanced machine learning models, i.e., CatBoost and LightGBM, could significantly outperform the standard APACHE IV scoring system in predicting ICU Length of Stay in spinal cord injury patients. Through rigorous SQL data integration, robust EDA, and intelligent feature engineering, we built a solid modeling pipeline to handle rich, high-cardinality clinical data. While our overall model was a good predictor ($R^2 > 0.55$), niche models for clusters better-enhanced accuracy (to $R^2 = 0.70$) by reflecting processes specific to subgroups. Still, issues like heteroscedasticity, outliers, and irreducible variance are due to variables not being measured, such as mortality and administrative delays. These findings highlight the strengths and limitations of data-driven modeling in critical care and stress the need for greater feature integration and external validation in future research.

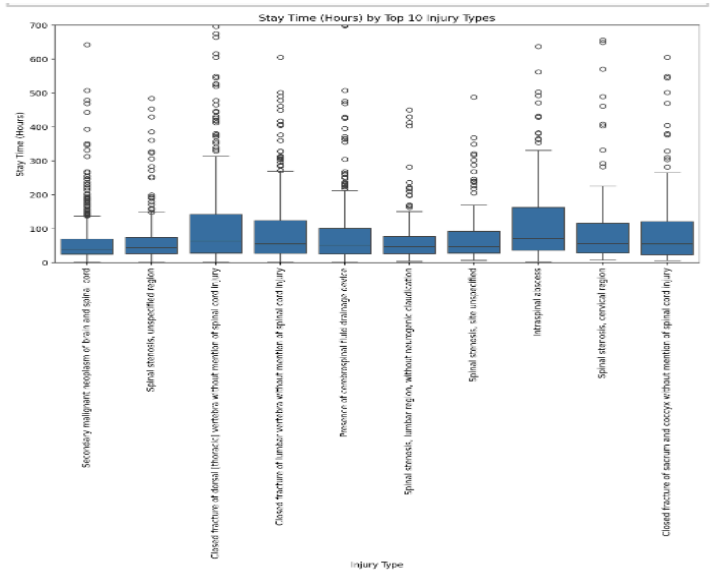
Figures:

Figure (1):




```
Stay Time (in hours):
count    5290.000000
mean       99.881878
std       144.280273
min         0.498889
25%        26.698056
50%        52.195972
75%       106.611458
max      2391.322778
Name: stay_time_hours, dtype: float64
```

Figure (2):


Stay Time (Hours) by Top 10 Injury Types
Stay Time (Hours)
Injury Type

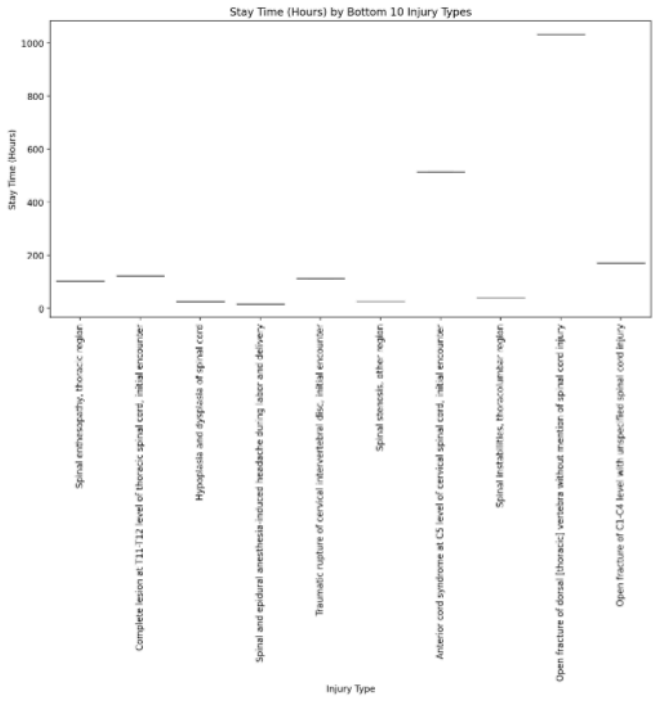
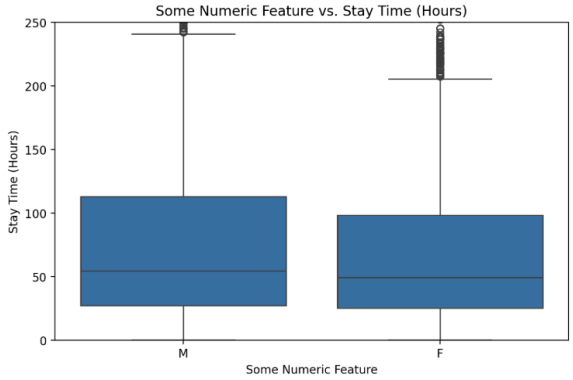
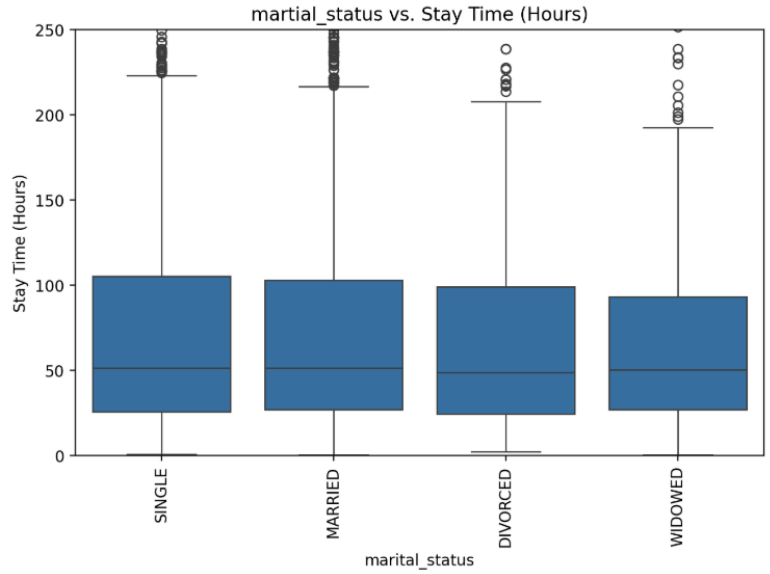

Stay Time (Hours) by Bottom 10 Injury Types
Stay Time (Hours)
Injury Type

Figure 3:


Some Numeric Feature vs. Stay Time (Hours)
Stay Time (Hours)
250
200
150
100
50
0
M
F
Some Numeric Feature


martial_status vs. Stay Time (Hours)
Stay Time (Hours)
250
200
150
100
50
0
SINGLE
MARRIED
DIVORCED
WIDOWED
marital_status

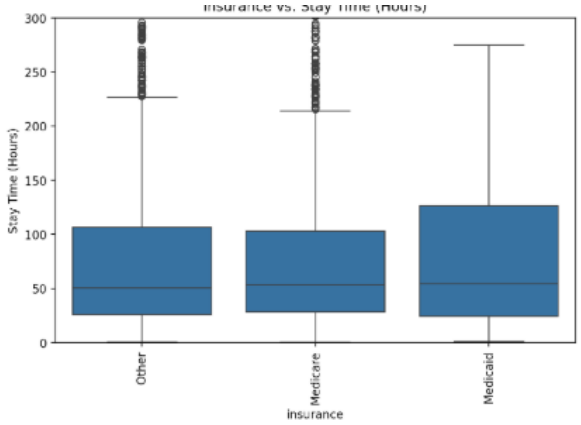
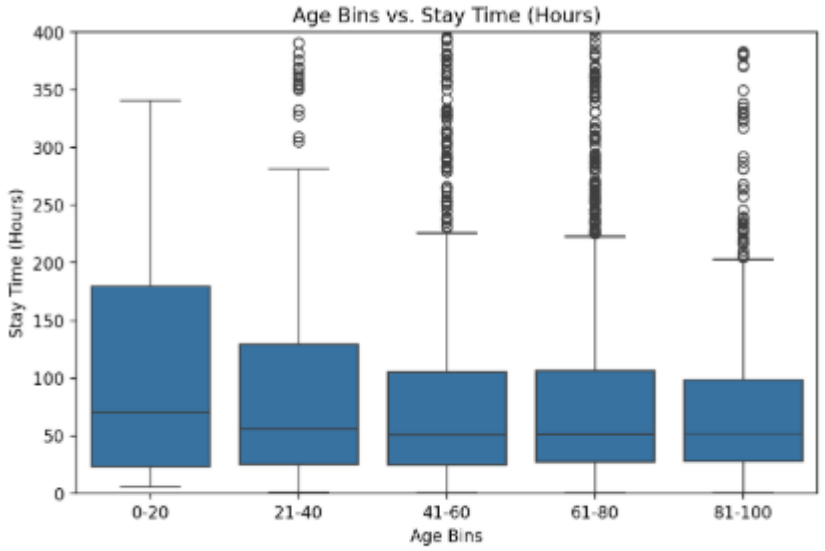

Stay Time (Hours)
300
250
200
150
100
50
0
Other
Medicare
Medicaid
insurance

Figure (4):



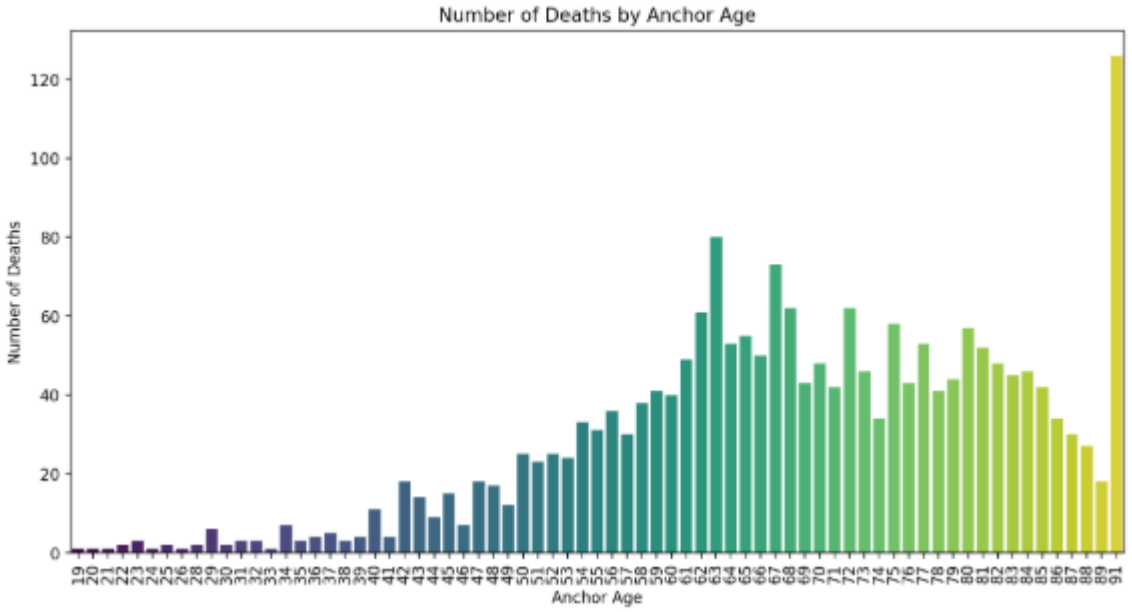
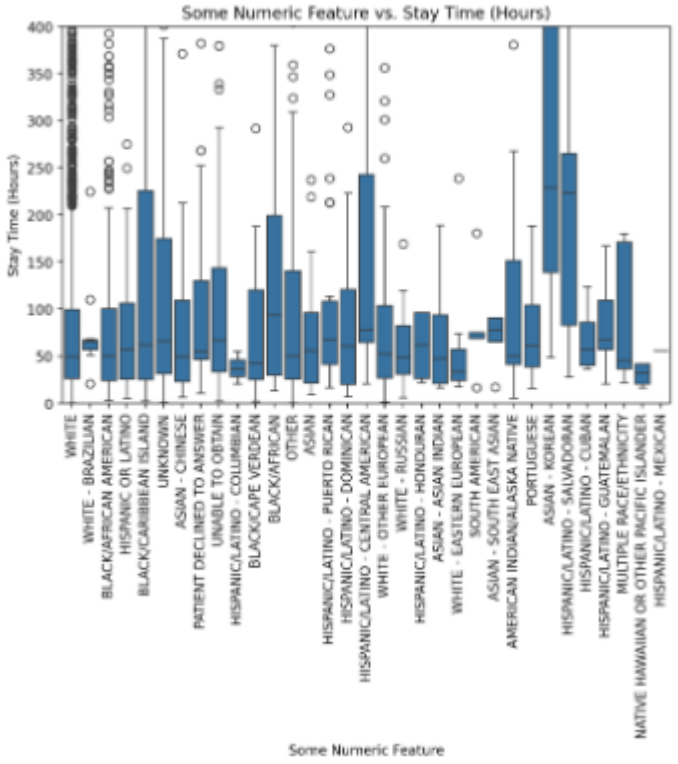


Figure (5):


Some Numeric Feature vs. Stay Time (Hours)
Stay Time (Hours)
0
50
100
150
200
250
300
350
400
WHITE
WHITE - BRAZILIAN
BLACK/AFRICAN AMERICAN
HISPANIC OR LATINO
BLACK/CARIBBEAN ISLAND
UNKNOWN
ASIAN - CHINESE
PATIENT DECLINED TO ANSWER
UNABLE TO OBTAIN
HISPANIC/LATINO - COLUMBIAN
BLACK/CAPE VERDEAN
BLACK/AFRICAN
OTHER
ASIAN
HISPANIC/LATINO - PUERTO RICAN
HISPANIC/LATINO - DOMINICAN
HISPANIC/LATINO - CENTRAL AMERICAN
WHITE - OTHER EUROPEAN
WHITE - RUSSIAN
HISPANIC/LATINO - HONDURAN
ASIAN - ASIAN INDIAN
WHITE - EASTERN EUROPEAN
SOUTH AMERICAN
ASIAN - SOUTH EAST ASIAN
AMERICAN INDIAN/ALASKA NATIVE
PORTUGUESE
ASIAN - KOREAN
HISPANIC/LATINO - SALVADORAN
HISPANIC/LATINO - CUBAN
HISPANIC/LATINO - GUATEMALAN
MULTIPLE RACE/ETHNICITY
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER
HISPANIC/LATINO - MEXICAN
Some Numeric Feature

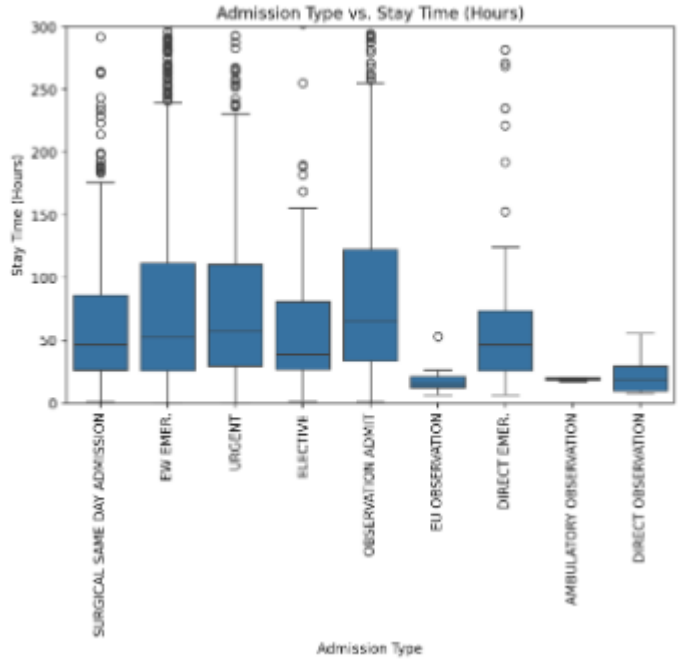
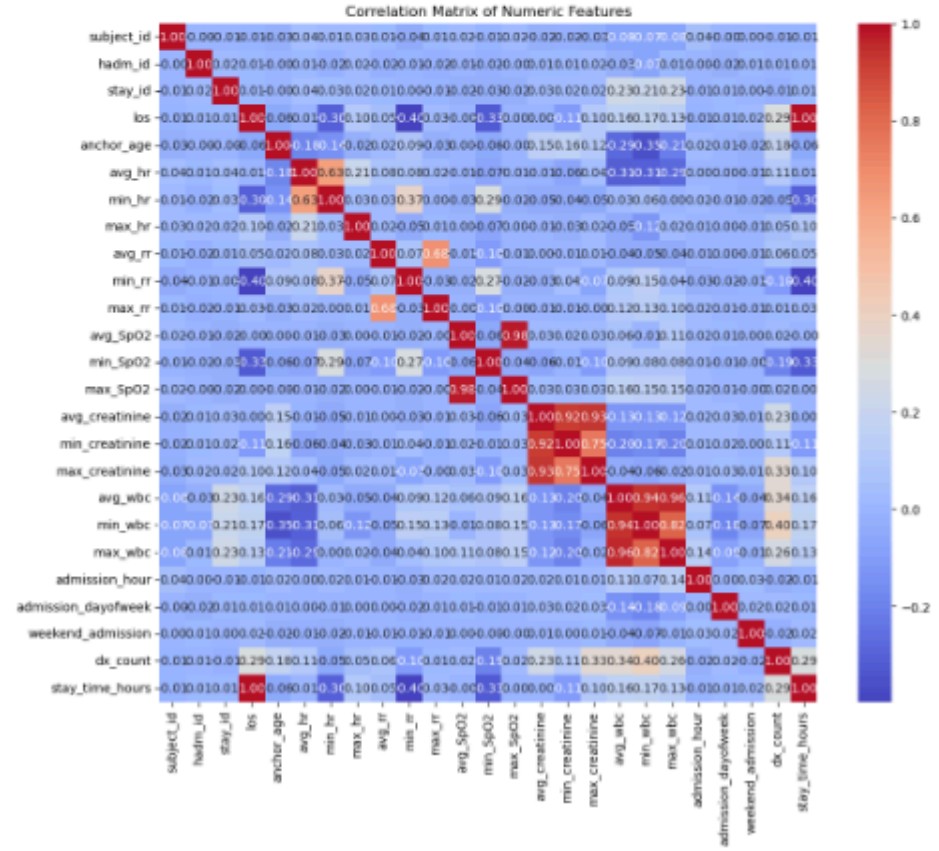

Admission Type vs. Stay Time (Hours)
Stay Time (Hours)
0
50
100
150
200
250
300
SURGICAL SAME DAY ADMISSION
EW EMER.
URGENT
ELECTIVE
OBSERVATION ADMIT
EU OBSERVATION
DIRECT EMER.
AMBULATORY OBSERVATION
DIRECT OBSERVATION
Admission Type

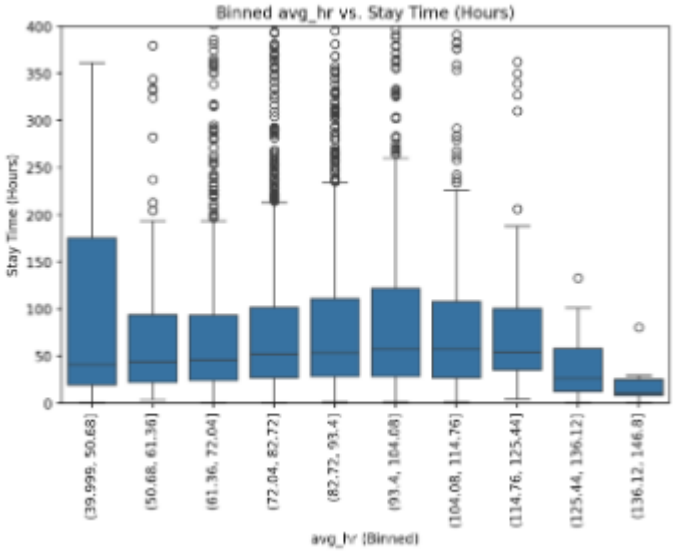
Figure 6:

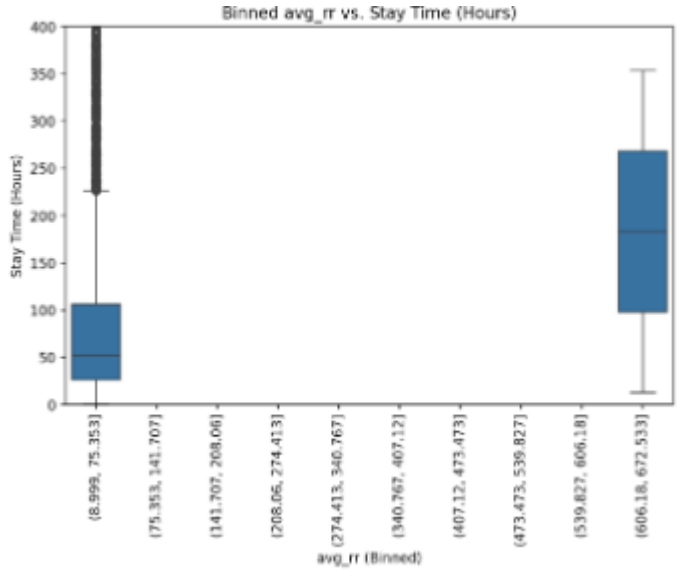
```
Correlation of features with stay_time_hours:
stay_time_hours          1.000000
los                      1.000000
dx_count                 0.292136
min_wbc                  0.171005
avg_wbc                  0.158115
max_wbc                  0.133161
max_creatinine           0.103625
max_hr                   0.101546
avg_rr                   0.052692
max_rr                   0.025719
weekend_admission        0.021226
admission_dayofweek      0.012757
hadm_id                  0.008641
stay_id                  0.007932
avg_hr                   0.005349
avg_creatinine           0.001863
max_SpO2                 0.000608
avg_SpO2                -0.000864
subject_id              -0.009703
admission_hour          -0.011398
anchor_age              -0.056567
min_creatinine          -0.105735
min_hr                  -0.300813
min_SpO2                -0.327704
min_rr                  -0.395968
Name: stay_time_hours, dtype: float64
```

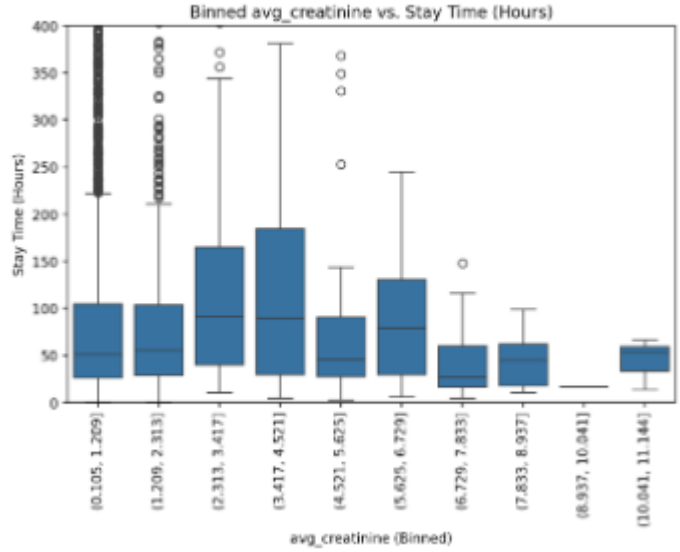
Out[20]:



Figure(7):


Binned avg_hr vs. Stay Time (Hours)
Stay Time (Hours)
0
50
100
150
200
250
300
350
400
(39.999, 50.68]
(50.68, 61.36]
(61.36, 72.04]
(72.04, 82.72]
(82.72, 93.4]
(93.4, 104.08]
(104.08, 114.76]
(114.76, 125.44]
(125.44, 136.12]
(136.12, 146.8]
avg_hr (Binned)


Binned avg_rr vs. Stay Time (Hours)
Stay Time (Hours)
0
50
100
150
200
250
300
350
400
(8.999, 75.353]
(75.353, 141.707]
(141.707, 208.06]
(208.06, 274.413]
(274.413, 340.767]
(340.767, 407.12]
(407.12, 473.473]
(473.473, 539.827]
(539.827, 606.18]
(606.18, 672.533]
avg_rr (Binned)


Binned avg_creatinine vs. Stay Time (Hours)
Stay Time (Hours)
0
50
100
150
200
250
300
350
400
(0.105, 1.209]
(1.209, 2.313]
(2.313, 3.417]
(3.417, 4.521]
(4.521, 5.625]
(5.625, 6.729]
(6.729, 7.833]
(7.833, 8.937]
(8.937, 10.041]
(10.041, 11.144]
avg_creatinine (Binned)

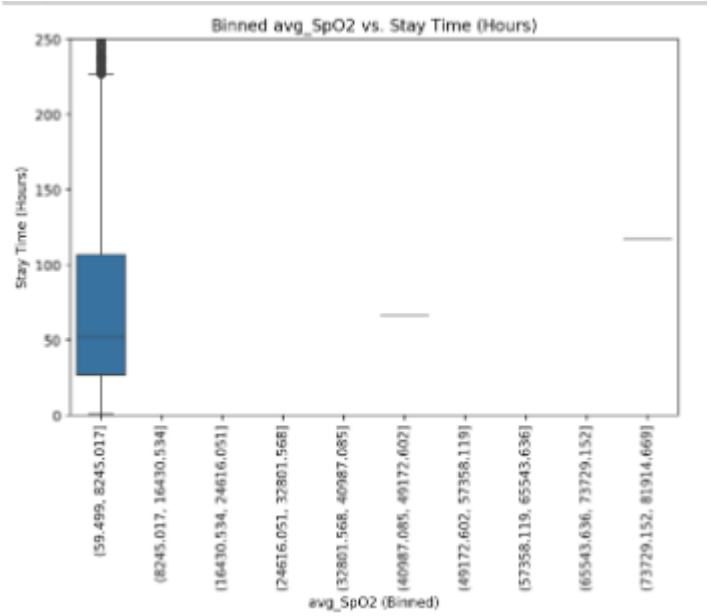
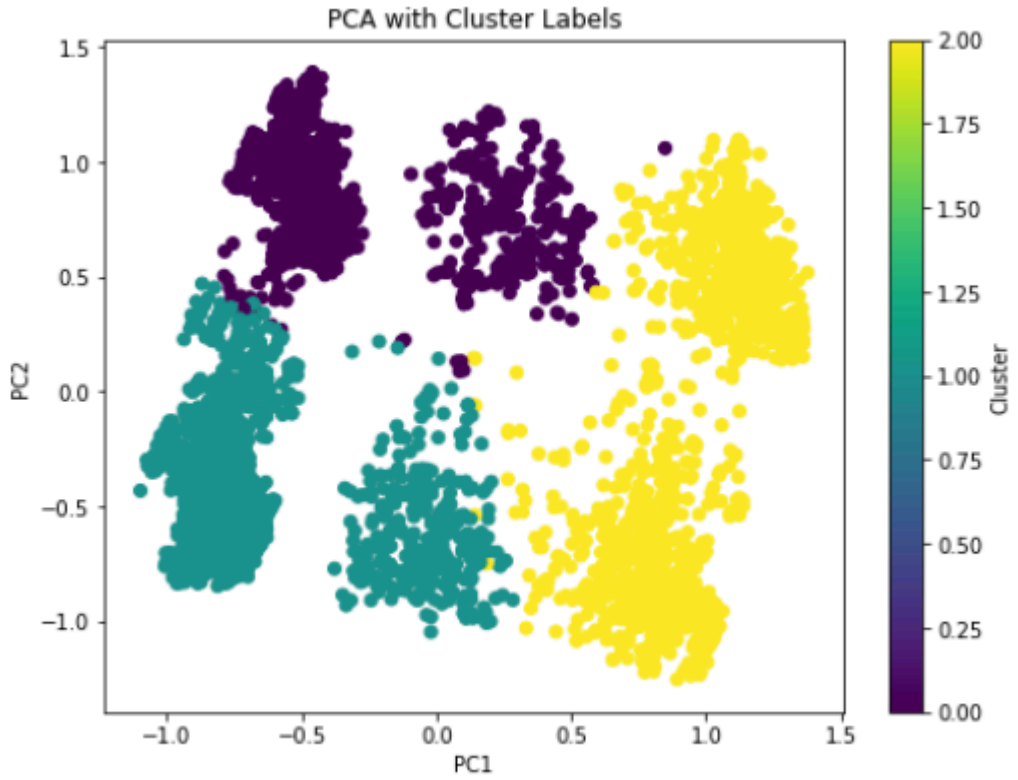

Binned avg_SpO2 vs. Stay Time (Hours)
Stay Time (Hours)
250
200
150
100
50
0
(59.499, 8245.017]
(8245.017, 16430.534]
(16430.534, 24616.051]
(24616.051, 32801.568]
(32801.568, 40987.085]
(40987.085, 49172.602]
(49172.602, 57358.119]
(57358.119, 65543.636]
(65543.636, 73729.152]
(73729.152, 81914.669]
avg_SpO2 (Binned)

Figure 8:



pca_cluster	0	1	2
los	0.041777	0.042609	0.040301
anchor_age	0.716175	0.503726	0.636994
avg_hr	0.401920	0.433967	0.408906
min_hr	0.473432	0.503228	0.486123
max_hr	0.009665	0.010152	0.010172
...	...	...	...
admission_dayofweek_4	0.131883	0.158779	0.149946
admission_dayofweek_5	0.131883	0.130789	0.154800
admission_dayofweek_6	0.153637	0.141476	0.142395
weekend_admission_1	0.288919	0.283969	0.279396
cluster	0.477226	0.286005	0.407767

[331 rows x 3 columns]

Top contributing features to PC1:

pca_cluster	0.726034
admission_location_EMERGENCY ROOM	0.389490
admission_type_EW EMER.	0.381932
admission_location_PHYSICIAN REFERRAL	0.219179
admission_location_TRANSFER FROM HOSPITAL	0.156144
admission_type_SURGICAL SAME DAY ADMISSION	0.143725
insurance_Other	0.141055
admission_type_URGENT	0.139566
insurance_Medicare	0.131553
died	0.103552

Name: PC1, dtype: float64

```
Top contributing features to PC2:
 insurance_Medicare                          0.582059
insurance_Other                              0.568532
cluster                                      0.244444
admission_type_EW EMER.                      0.234784
admission_location_EMERGENCY ROOM            0.230924
died                                         0.177118
anchor_age                                   0.156619
admission_location_TRANSFER FROM HOSPITAL    0.142971
dx_count                                     0.125202
admission_type_URGENT                        0.124726
Name: PC2, dtype: float64
```

Figure 9:

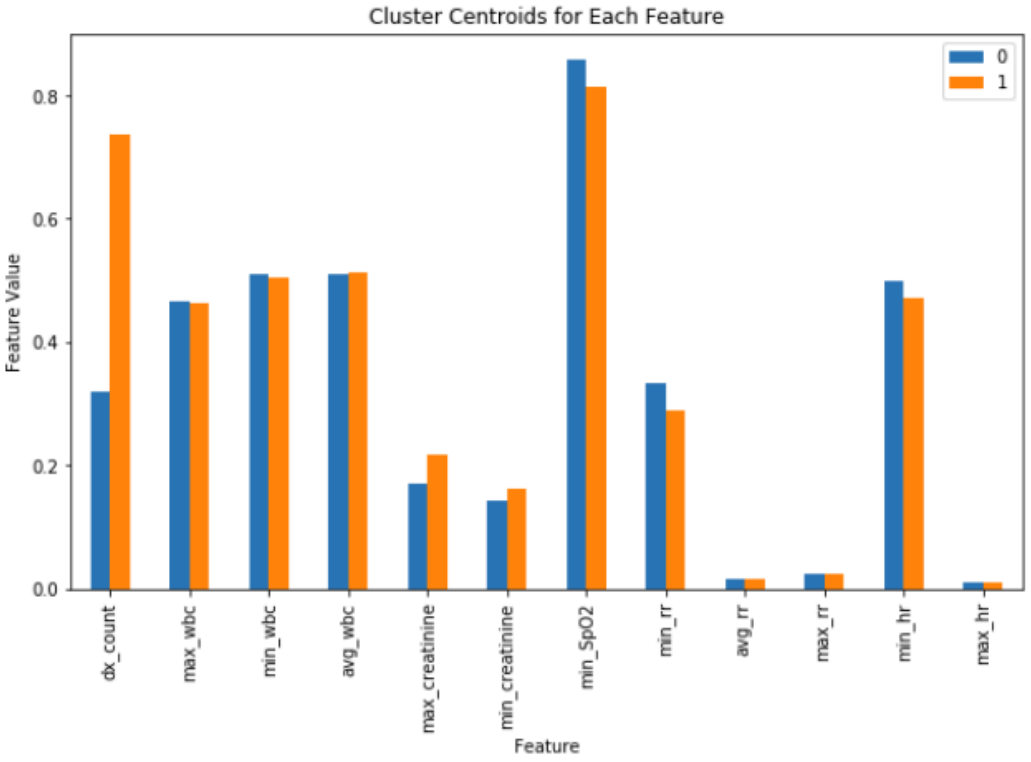


Figure 10:

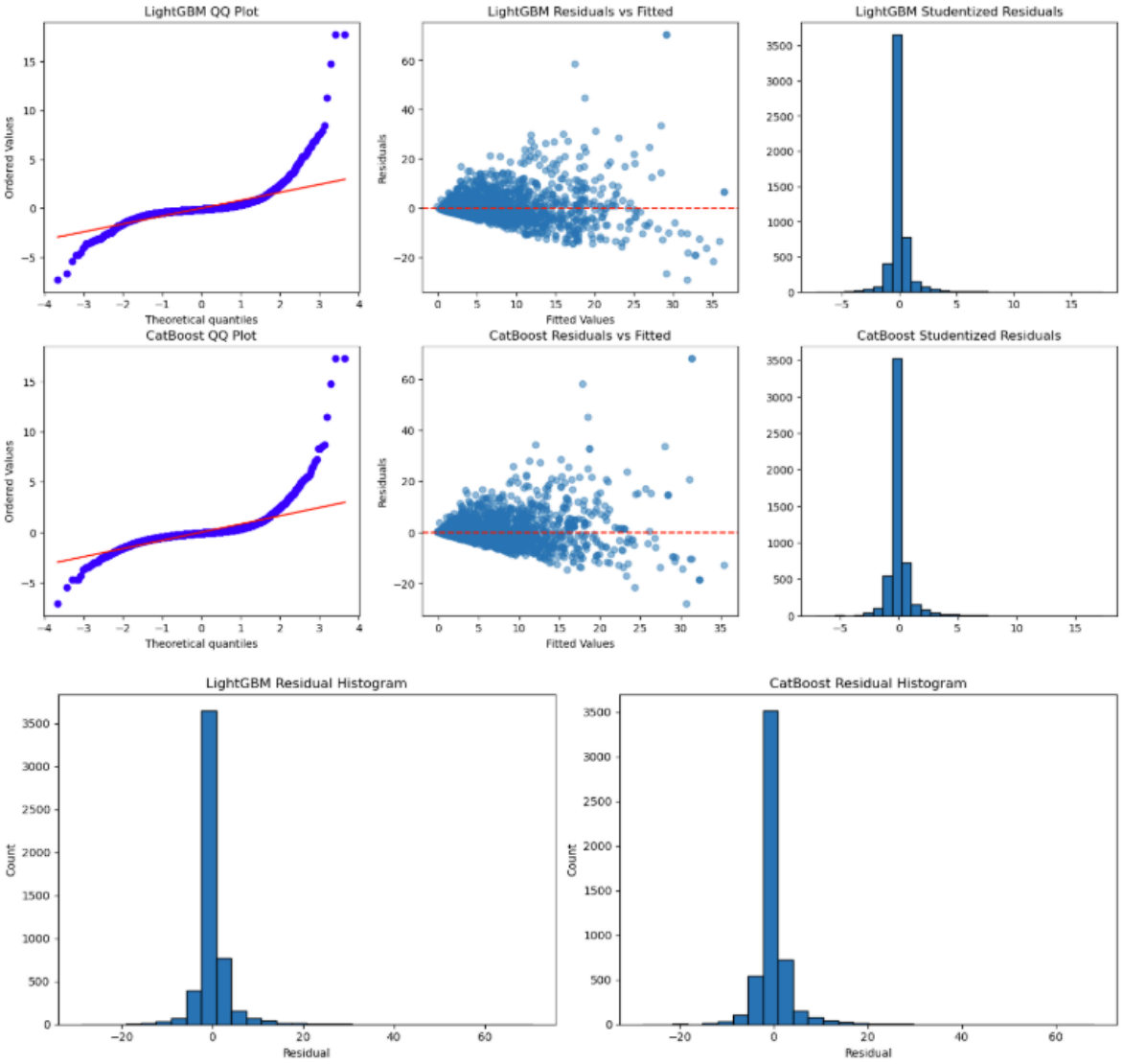


Figure 11 (other clusters and model looked like this too):

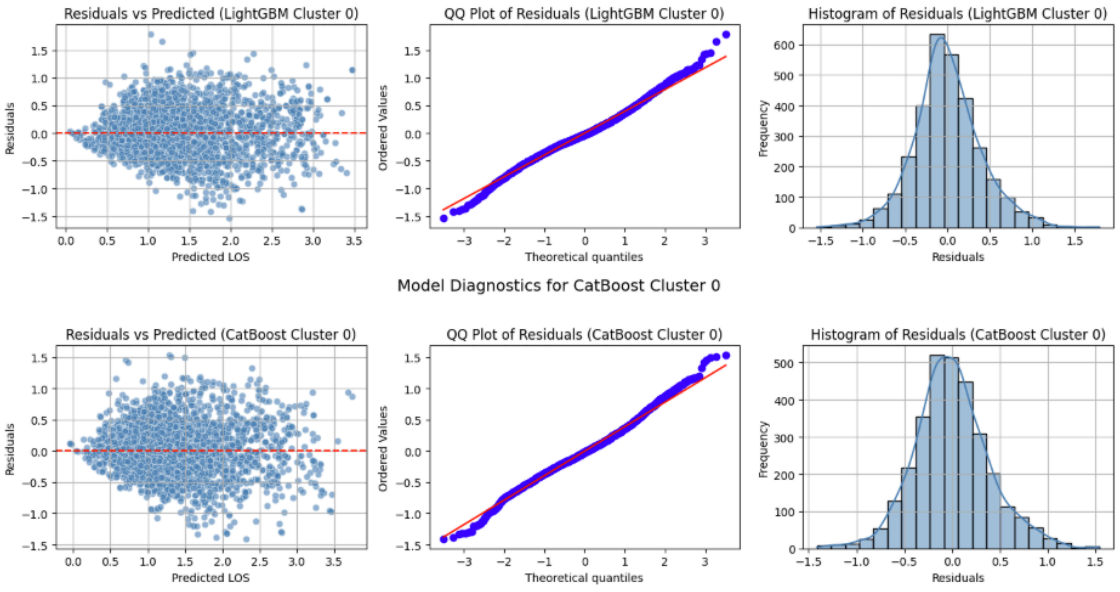


Figure 12:

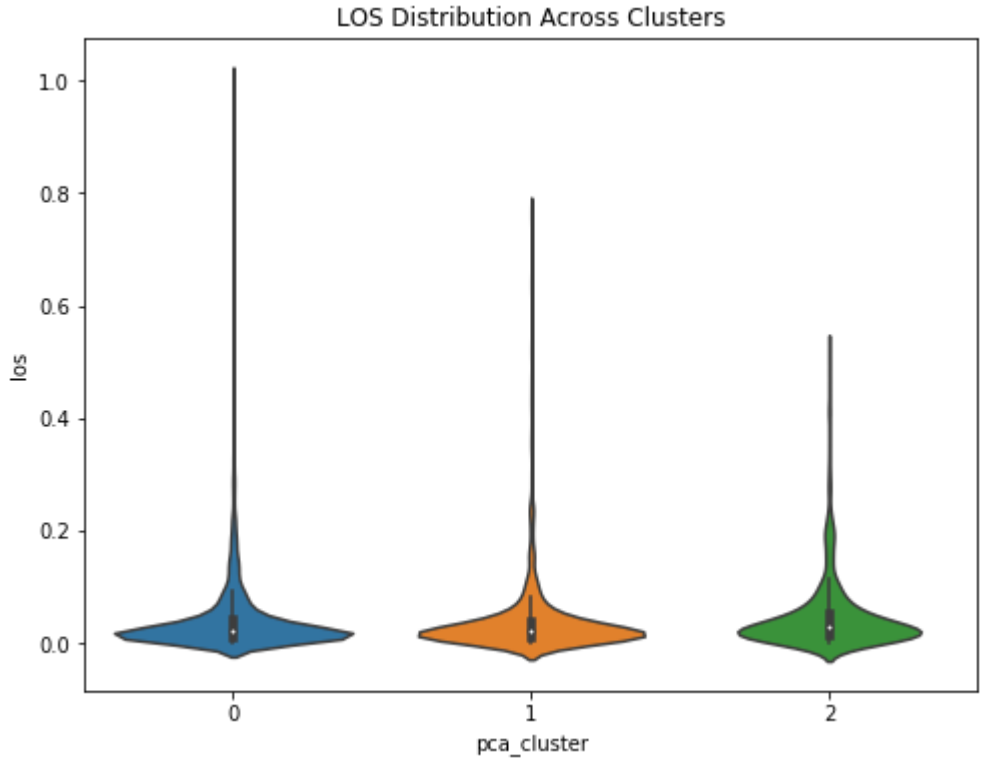


Figure 13:

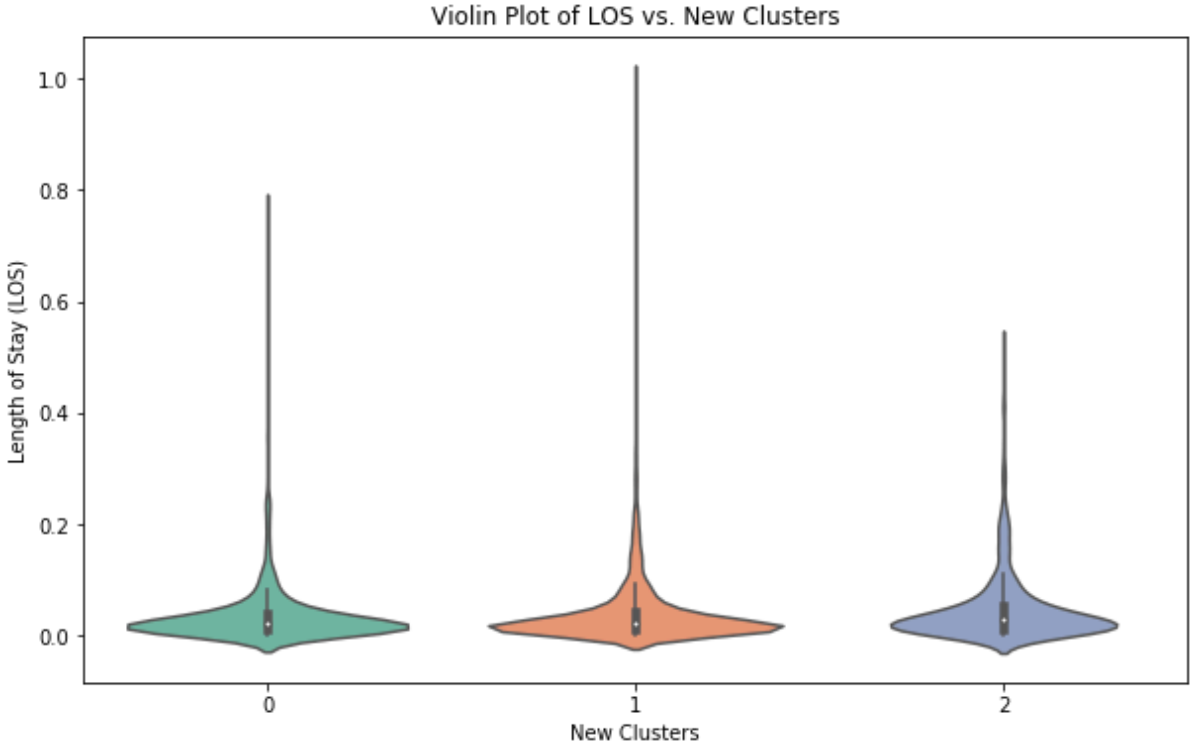


Figure 14:

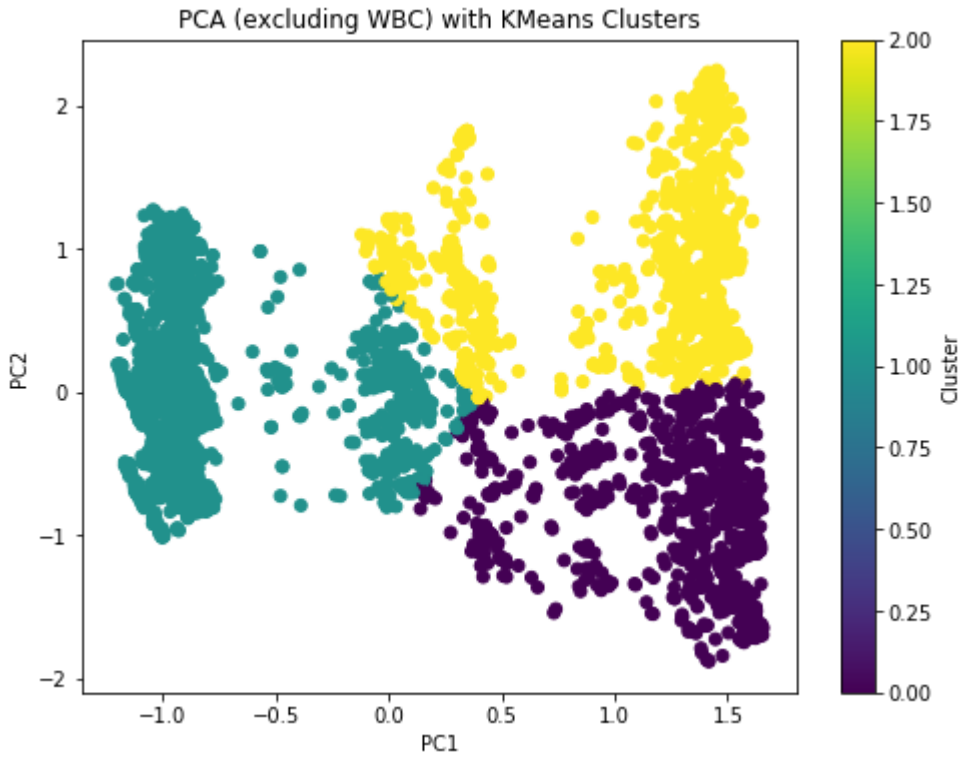


Figure 15:

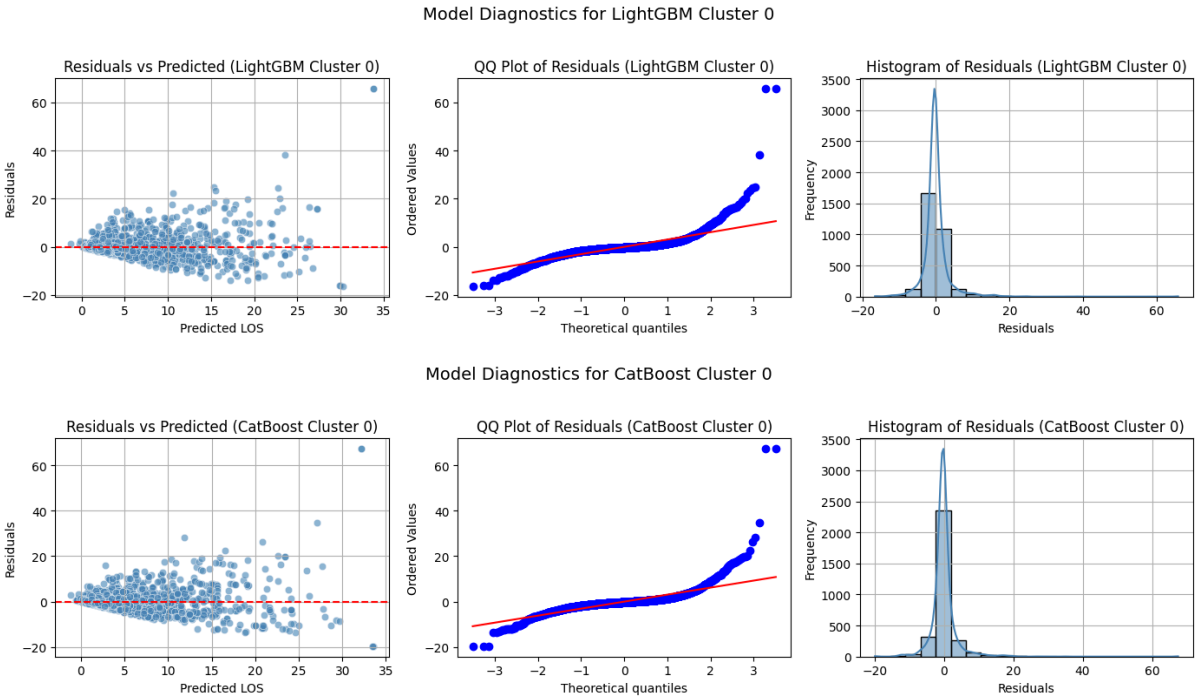


Figure 16:

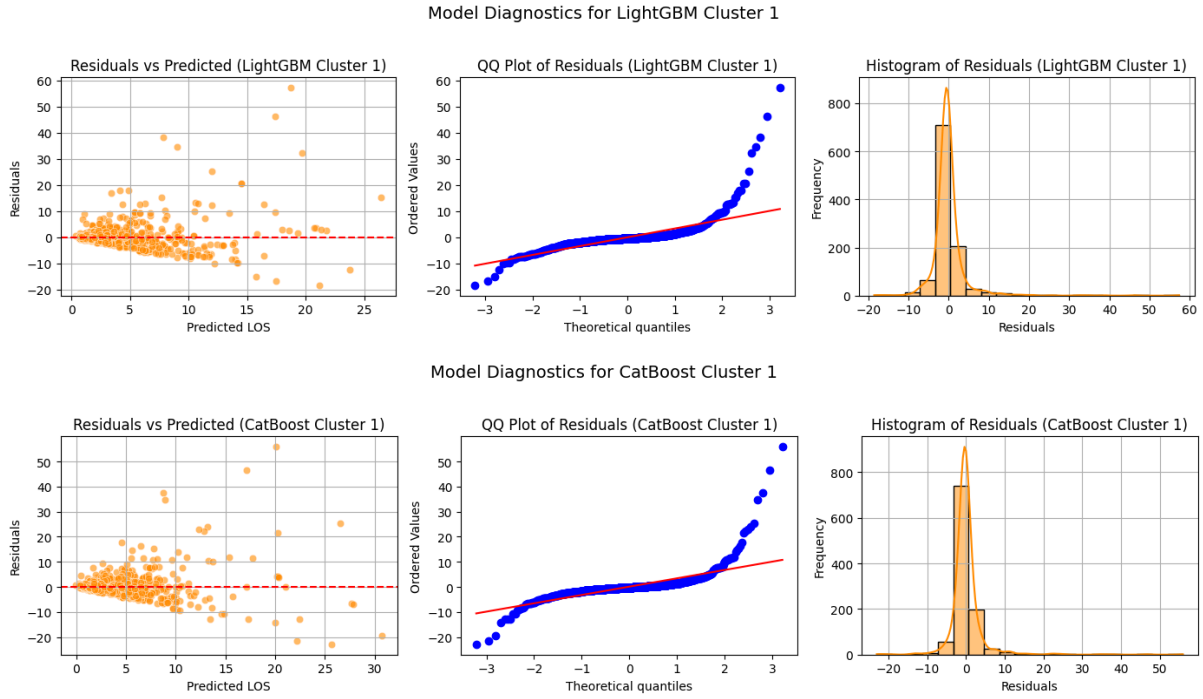


Figure 17:

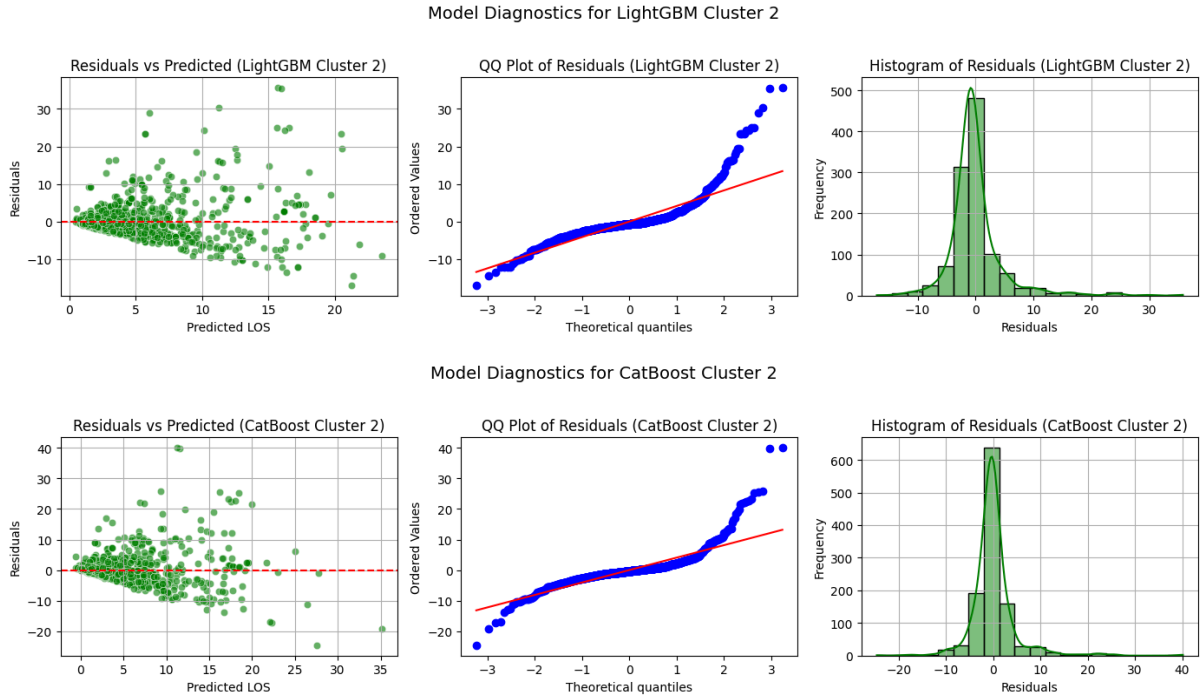


Figure 18:

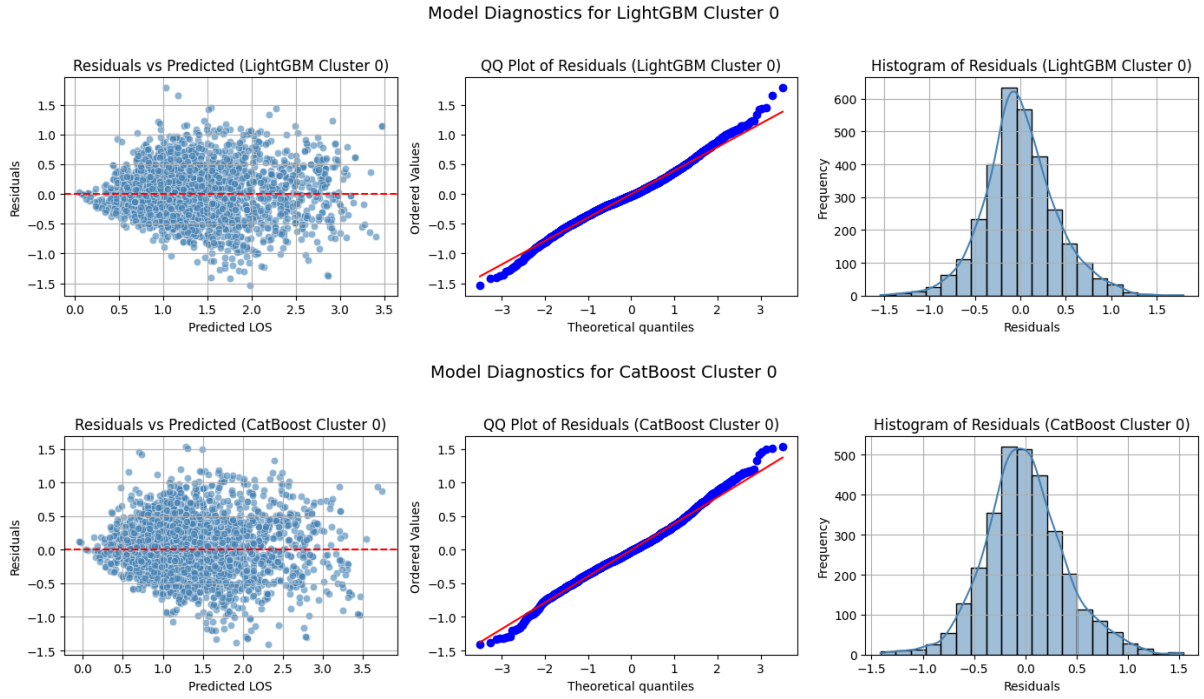


Figure 19:

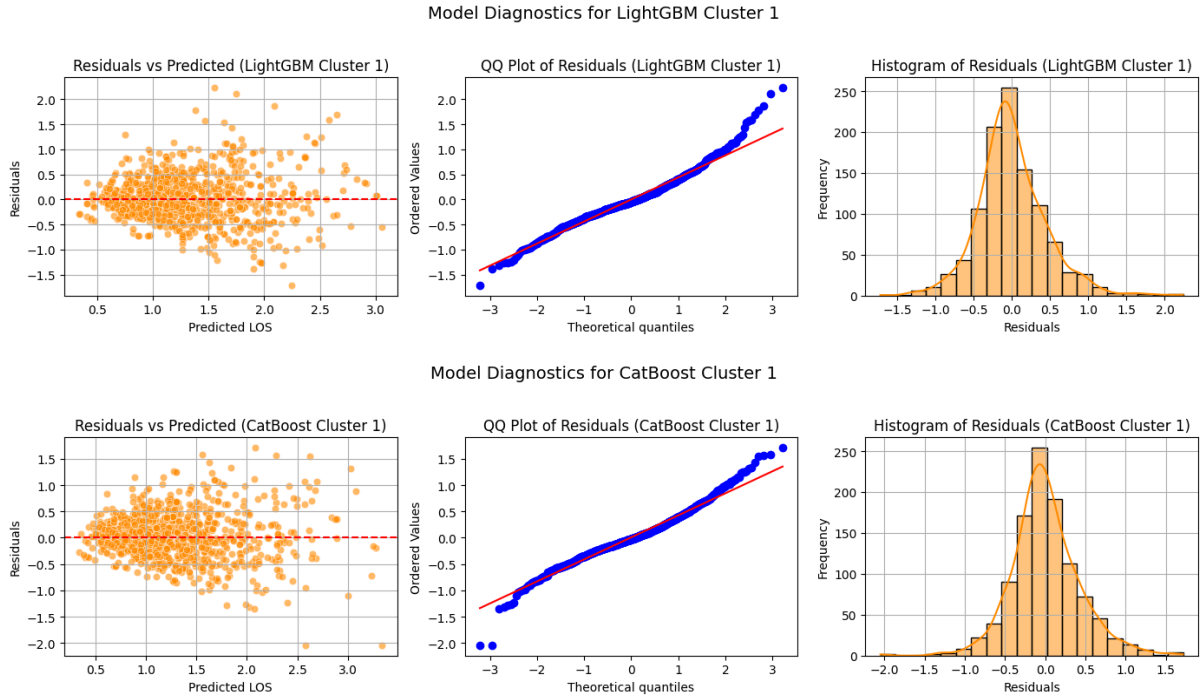
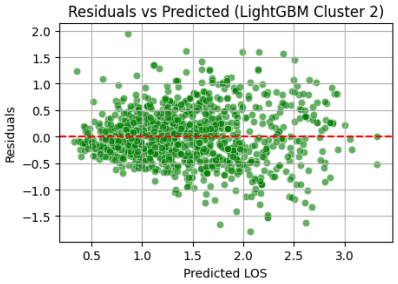
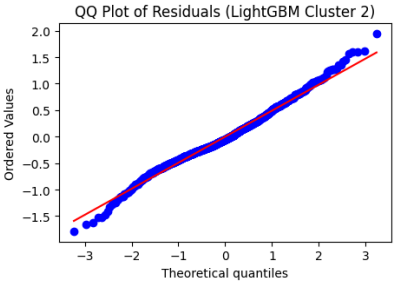
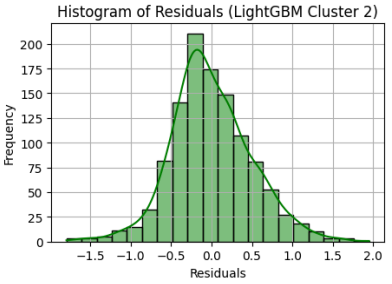


Figure 20:

Model Diagnostics for LightGBM Cluster 2







Model Diagnostics for CatBoost Cluster 2