A Project Report on

Healthcare Forecasting and Resource Allocation

Capstone Project Final Report

Submitted in partial fulfilment of requirement of

Master of Information Technology and Analytics (MITA)

Semester: Fall 2024

Batch September 2023 – December 2024

Submitted to:

Professor Michail Xyntarakis

Submitted by:

Parul Ghai

pg611

RUTGERS BUSINESS SCHOOL, NEWARK

OVERVIEW

Hospitals are constantly faced with the challenge of optimizing their resources to meet the demands of patient care. Accurate prediction of patient stay durations plays a critical role in resource planning, ensuring that hospitals are neither under-prepared nor over-extended. Hospital administrators need reliable data-driven solutions to balance the allocation of rooms, staff, and other critical resources.

Recent Covid-19 Pandemic has raised alarms over one of the most overlooked areas to focus: Healthcare Management. While healthcare management has various use cases for using data science, patient length of stay (LOS) is one critical parameter to observe and predict if one wants to improve the efficiency of healthcare management in a hospital. This parameter helps hospitals to identify patients of high LOS risk (patients who will stay longer) at the time of admission. Once identified, patients with high LOS risk can have their treatment plan optimized to minimize LOS and lower the chance of staff/visitor infection. Additionally, prior knowledge of LOS can aid in logistics such as room and bed allocation planning.

This project leverages historical patient data to analyze patterns in hospital stays and develop predictive models using machine learning. By understanding the underlying factors that influence patient stay durations, this study aims to provide actionable insights to enhance operational efficiency, improve patient outcomes, and reduce costs associated with underutilization or overburdening of resources. The project's focus is not only on creating accurate predictions but also on transforming those predictions into practical strategies for resource allocation.

Healthcare institutions, especially in urban settings, often grapple with fluctuating patient loads, making predictive modeling an invaluable tool for strategic decision-making. By integrating advanced data analytics with hospital operations, this study aspires to bridge the gap between predictive insights and actionable resource management.

This project embarks on an intricate analysis of hospital data to predict patient stay durations and optimize resource allocation. It integrates data preprocessing, exploratory data analysis (EDA), and machine learning techniques to derive actionable insights. By cleaning and restructuring the dataset, missing values were treated, and categorical variables were encoded to ensure compatibility with predictive models. The study conducted univariate and bivariate analyses to uncover patterns, such as the prevalence of moderate-severity cases in extended stays and the dominance of specific age groups in admissions. Key trends, like high patient counts in Gynecology and the need for additional beds in certain wards, were visualized to guide resource planning.

Feature engineering introduced enriched attributes like hospital interaction patterns and department transitions, significantly enhancing model performance. A LightGBM classifier, tuned for multiclass predictions, was used within a StratifiedKFold framework to balance class distributions. Insights from predicted patient stays were applied to calculate room, nurse, and doctor requirements for hospitals and departments. The project's comprehensive approach combines statistical rigor with real-world applicability, delivering a transformative methodology for hospital operations. By bridging predictive analytics and resource management, it supports efficient decision-making and optimizes healthcare logistics, aligning data-driven insights with patient care excellence.

PROBLEM STATEMENT

Hospitals face significant challenges in managing patient inflow and resource allocation effectively, especially in high-demand scenarios. Predicting the length of patient stays is critical for optimizing bed utilization, staffing, and medical resources. However, the variability in patient demographics, admission types, severity levels, and hospital capabilities complicates this task. Inaccurate predictions often result in inefficiencies, such as underutilized resources or patient overcrowding. Moreover, hospitals struggle to balance patient care quality with operational efficiency, as misaligned resource allocation can compromise outcomes. This complexity underscores the need for a robust, data-driven approach to model patient stay durations and support proactive healthcare planning.

The absence of real-time insights into patient stay patterns leads to bottlenecks in key hospital departments. Departments like surgery or gynecology may face unexpected surges, straining staff and infrastructure, while other areas remain underutilized. Additionally, hospitals often lack the analytical tools to integrate historical data with patient behavior insights, hindering their ability to forecast resource needs accurately. Beyond logistics, the financial implications are significant; mismanagement can lead to lost revenue, increased operational costs, and decreased patient satisfaction. Addressing this multifaceted problem requires a comprehensive framework that incorporates predictive analytics and operational optimization for real-time and actionable insights.

Patient's diverse medical conditions and admission types further add to the complexity of predicting stay durations. Emergency admissions or trauma cases, for example, have unpredictable lengths of stay compared to routine procedures. The interplay between clinical, demographic, and operational factors makes it challenging for hospitals to generalize stay patterns across patients. Furthermore, external factors like regional infrastructure or department-specific dynamics exacerbate the difficulty of resource planning. Without tailored strategies, hospitals cannot effectively allocate rooms, doctors, and nurses, leading to operational inefficiencies and adverse patient experiences. Addressing these challenges requires an integrative approach combining advanced data processing and predictive modeling.

In addition to operational inefficiencies, the inability to predict stay durations impacts broader healthcare goals, such as equitable patient care and cost-effective resource use. Inconsistent allocation of resources may delay care for critical patients or overburden specific departments. These imbalances also hinder hospitals' ability to respond to peak demand periods, such as seasonal flu outbreaks or emergencies. The downstream effects include patient dissatisfaction, overworked staff, and strained hospital systems. This complex problem necessitates a scalable and data-driven solution to optimize hospital operations while maintaining the highest standards of patient care and resource equity.

Key questions include:

- How can hospitals predict the length of stay for incoming patients?
- What factors influence the length of stay?
- How can predictive insights be used to improve resource allocation and operational efficiency?

This project aims to address these challenges by developing a machine learning model to predict patient stay durations and provide actionable recommendations for resource planning.

OBJECTIVE

The primary objective of this project is to develop a data-driven framework to predict patient stay durations and optimize hospital resource allocation. By leveraging advanced machine learning models, the project aims to provide accurate, actionable insights into patient inflow dynamics. This involves transforming raw hospital data into a robust pipeline that facilitates exploratory analysis, feature engineering, and predictive modeling. The ultimate goal is to equip hospitals with tools that enable them to align resources, such as beds, nurses, and doctors, with anticipated patient needs, ensuring operational efficiency and improving patient care outcomes.

A key focus of this project is on feature enrichment and engineering to capture the nuanced behavior of patients and their interactions with hospitals. Features such as the number of departments visited, admission types, and regional variations are incorporated to enhance prediction accuracy. Additionally, interaction variables are created to analyze complex relationships, such as how admission type correlates with patient severity. These insights are visualized using intuitive plots and dashboards, enabling hospital administrators to identify trends and optimize resource allocation strategies at both departmental and hospital-wide levels.

Another critical objective is to bridge the gap between analytics and decision-making by creating models that integrate directly into hospital operations. Through LightGBM, a multiclass classifier optimized with StratifiedKFold, the project ensures reliable predictions across all patient stay categories. These predictions are further extended to resource planning by estimating room, nurse, and doctor requirements for different hospital units. The integration of predictive analytics with operational logistics ensures that hospitals can make informed, data-driven decisions, reducing inefficiencies and enhancing patient satisfaction through timely and adequate care delivery.

Beyond operational goals, the project aspires to drive a broader transformation in healthcare analytics. By demonstrating the value of predictive modeling and feature engineering, it sets the stage for scalable solutions in hospital management. The project's approach aligns predictive insights with real-world applicability, supporting hospitals in adapting to dynamic patient inflow scenarios. Ultimately, this framework provides a foundation for building smarter, more efficient healthcare systems that prioritize both operational excellence and patient-centric care.

The primary objectives of this project are:

1. To identify and analyze the factors that influence patient length of stay (LOS) in hospitals.
2. To develop a machine learning model that accurately predicts LOS based on patient and hospital data.
3. To use the predictions to recommend resource allocation strategies, ensuring optimal utilization of hospital resources like rooms, beds, and staff.
4. To enhance healthcare efficiency by minimizing LOS for high-risk patients while maintaining quality care.

DATASET OVERVIEW

The dataset for this project was accessed from the Kaggle JanataHack Healthcare Analytics competition. It comprises two primary files: one for training (train.csv) and another for testing (test.csv). Alongside these, a data dictionary file (train_data_dict.csv) provides definitions and descriptions for each variable present in the dataset.

Train Set:
train.csv: This file contains historical patient data, including features related to hospitals, patients, and admission details, along with the target variable Stay, which represents the length of stay for each case.
train_data_dict.csv: A data dictionary file that describes each column present in the train dataset.

Test Set:
test.csv: This file contains the same features as the train dataset, excluding the target variable Stay. The goal is to predict the length of stay for each case in the test set.

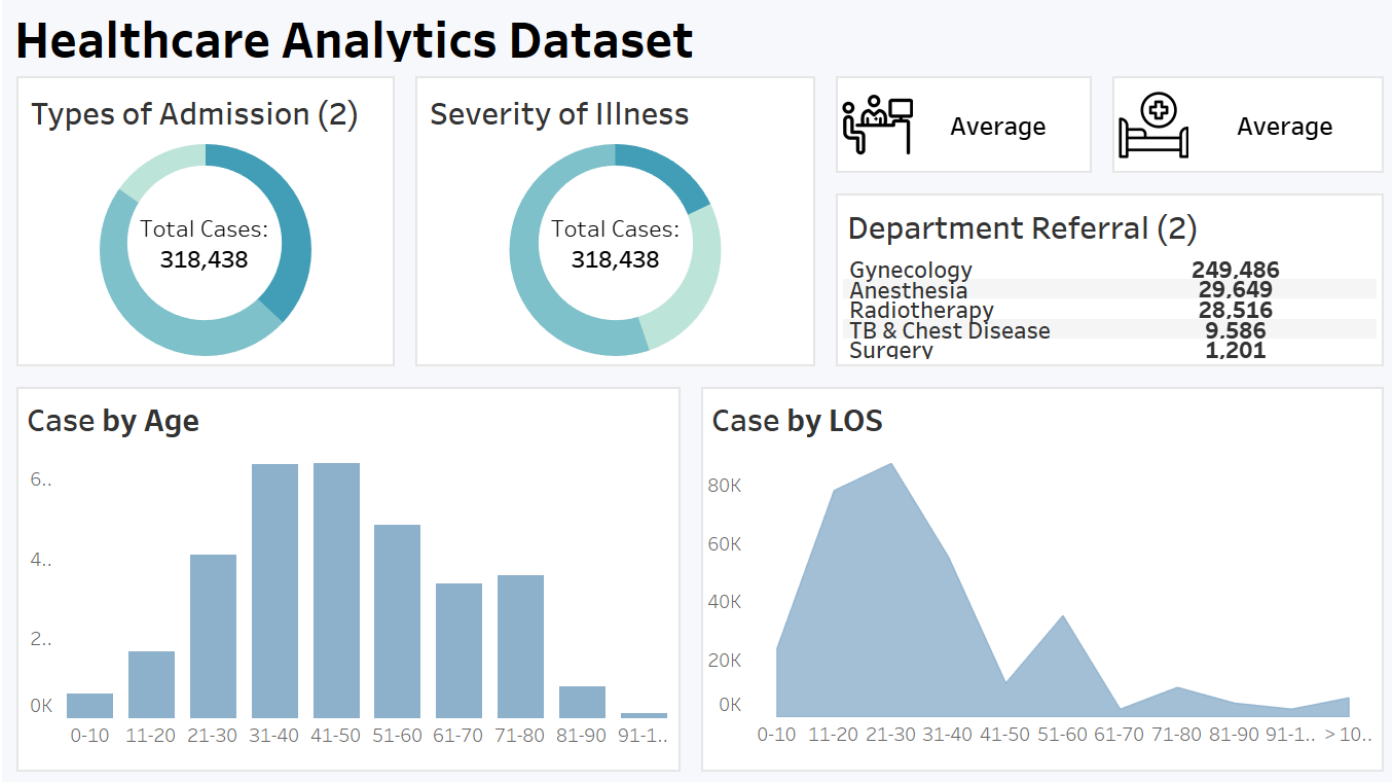
Below is a detailed breakdown of the dataset columns:

Column	Description
case_id	Case_ID registered in Hospital
Hospital_code	Unique code for the Hospital
Hospital_type_code	Unique code for the type of Hospital
City_Code_Hospital	City Code of the Hospital
Hospital_region_code	Region Code of the Hospital
Available Extra Rooms in Hospital	Number of Extra rooms available in the Hospital
Department	Department overlooking the case
Ward_Type	Code for the Ward type
Ward_Facility_Code	Code for the Ward Facility
Bed Grade	Condition of Bed in the Ward
patientid	Unique Patient Id
City_Code_Patient	City Code for the patient
Type of Admission	Admission Type registered by the Hospital
Severity of Illness	Severity of the illness recorded at the time of admission
Visitors with Patient	Number of Visitors with the patient
Age	Age of the patient
Admission_Deposit	Deposit at the Admission Time
Stay	Stay Days by the patient (Target Variable in Train Set)

EXPLORATORY DATA ANALYSIS

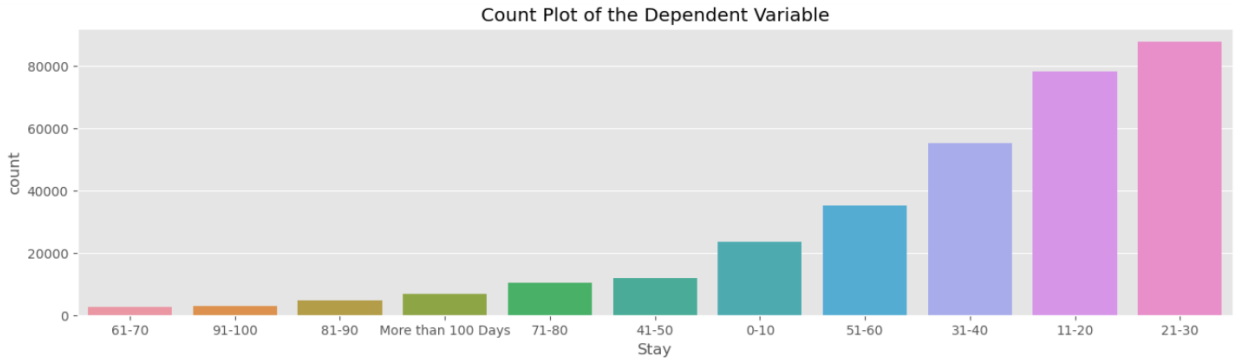
Extensive EDA was performed to understand the characteristics of the dataset, the relationships between features, and patterns influencing the target variable Stay. Below are the key insights gathered during the analysis:

Dataset Analysis



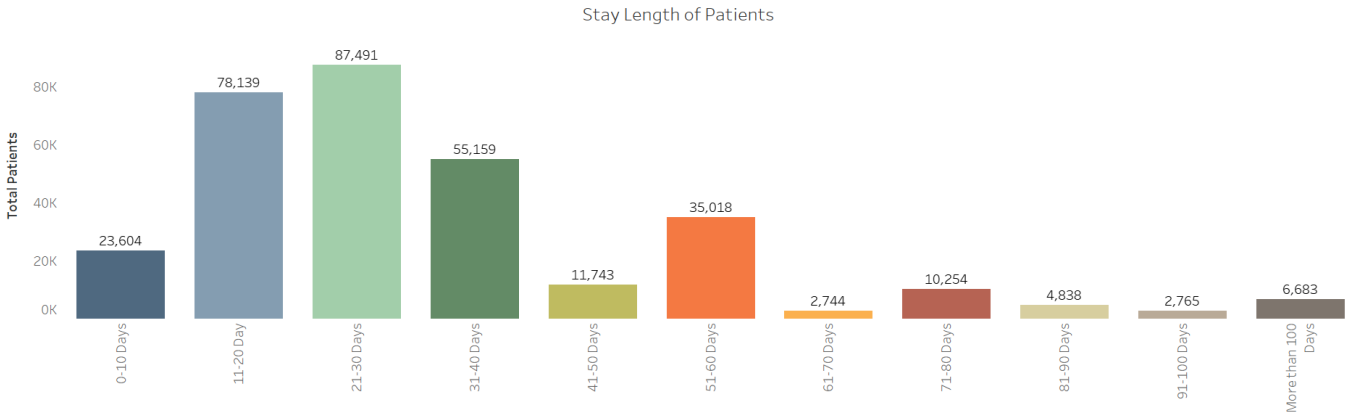
Dependent Variable

Stay Distribution: The Stay column was analyzed to understand the frequency of various stay durations.



1. We can see that the highest number of patients are from the Age range 21-30, followed by age ranges 11-20 and 3-40.
2. We can also see that the minimum number of patients are from the Age range 61-70.
3. We can also see that there are good number of patients who stay for more than 100 days.

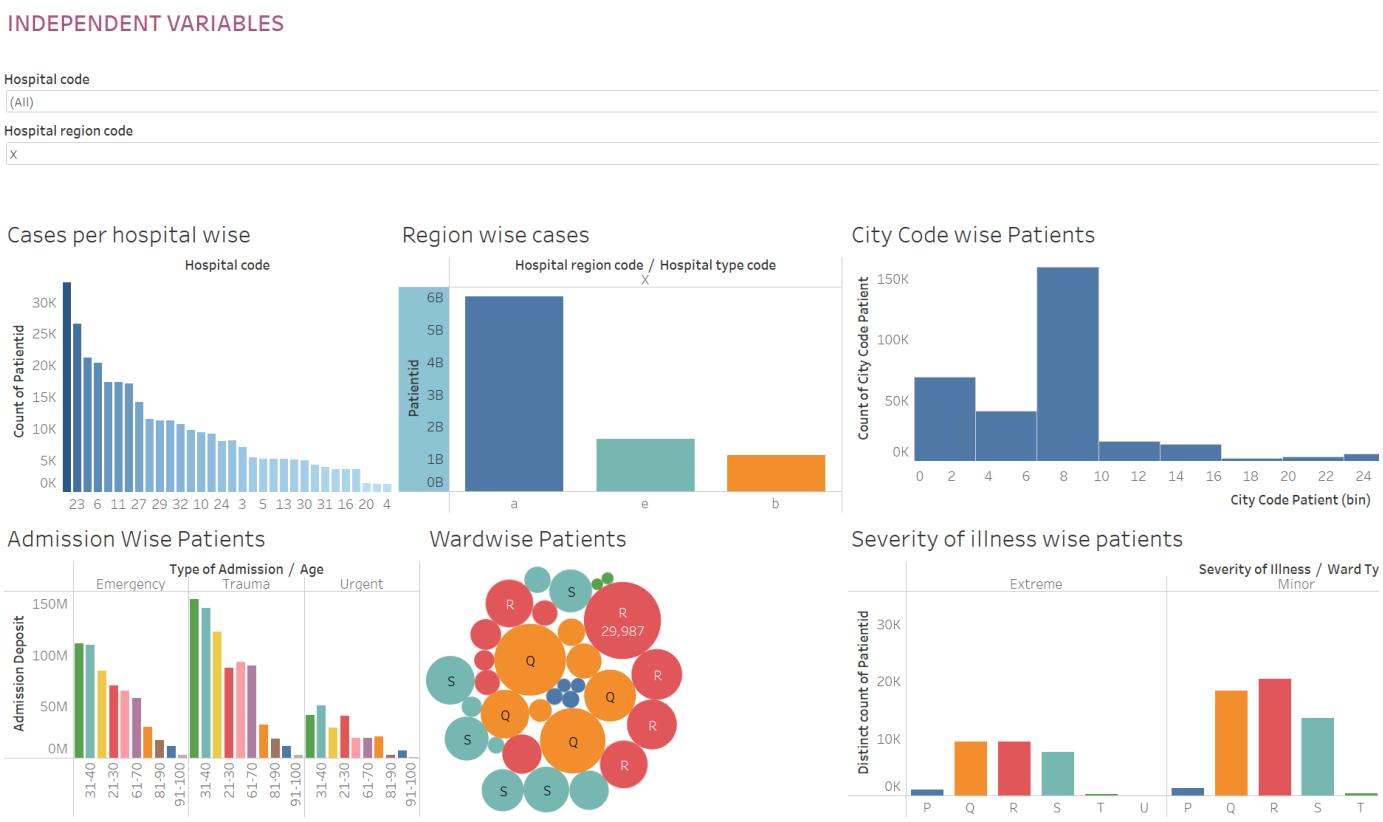
Stay Lengths of **11-20 and 21-30** Days were **most common** among all patients, while Stay Lengths suddenly **increase at 51-60** Days.



➤ **11-20 Days:** 78,139 patients, an increase of ↑ 54,525 compared to 0-10 Days
➤ **21-30 Days:** 87,491 patients, an increase of ↑ 9,352 compared to 11-20 Days
➤ **51-60 Days:** 35,018 patients, an increase of ↑ 23,275 compared to 41-50 Days, and ↑ 32,274 compared to 61-70 Days

Source - AV : Healthcare Analytics II

Independent Variable



Hospital Characteristics

- Hospital Code: A count plot of Hospital_code showed that certain hospitals (e.g., Hospital 26) admit a significantly higher number of patients compared to others.
- Hospital Region: Analysis revealed that Region X hospitals have the highest number of patients, followed by Region Y and Region Z.
- Available Extra Rooms: Most hospitals have between 2-4 extra rooms, but some lack additional capacity, which could impact resource management.

Patient Demographics

- Age Group:

  Patients aged 21-30 formed the largest group, followed closely by 31-40 and 41-50 age groups.

  Very few patients fell in the 0-10 and 91-100 age brackets, which aligns with expected trends.
- City_Code_Patient: Missing values were observed and imputed using the most frequent city code.

Severity and Admission Type

- Severity of Illness:
    - Most patients were categorized as having Moderate severity, followed by Minor severity and a smaller proportion of Extreme severity cases.
- Type of Admission:
    - Emergency admissions were the most common, followed by Trauma and Urgent cases.
    - Patients with longer stays often belonged to the Trauma and Emergency categories.

Financial Features

- Admission Deposit:
    - A box plot revealed that the deposit amount is fairly distributed, with no significant outliers.
    - The average deposit amount varied across different departments, with Anesthesia and TB requiring higher average deposits.

Ward and Department Analysis

- Ward Type and Facility:

  Ward Type R was the most common, followed by Types Q and S.

  Facilities in certain ward types had limited capacity, which could impact LOS.
- Departments:

  Gynecology had the highest number of patients, while Surgery had the least.

  Cross-tabulation showed that patients with longer stays were often admitted to the TB and Anesthesia departments.

BI VARIATE ANALYSIS :

Stay	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	More than 100 Days
Age											
0-10	615	1959	1489	1014	187	582	26	153	84	35	110
11-20	1552	5343	4312	2681	510	1429	89	350	223	71	208
21-30	3467	11272	11394	6912	1398	3793	263	1026	546	231	541
31-40	4916	15792	18550	10912	2373	6517	509	1807	801	484	978
41-50	4727	14959	17906	10983	2507	7189	562	2146	885	578	1307
51-60	3427	11346	13058	8569	1735	5739	448	1710	784	499	1199
61-70	2194	7870	9033	5930	1205	4081	325	1230	600	330	889
71-80	2201	7958	9534	6420	1383	4433	378	1367	670	386	1062
81-90	422	1392	1920	1504	379	1082	115	402	216	132	326
91-100	83	248	295	234	66	173	29	63	29	19	63

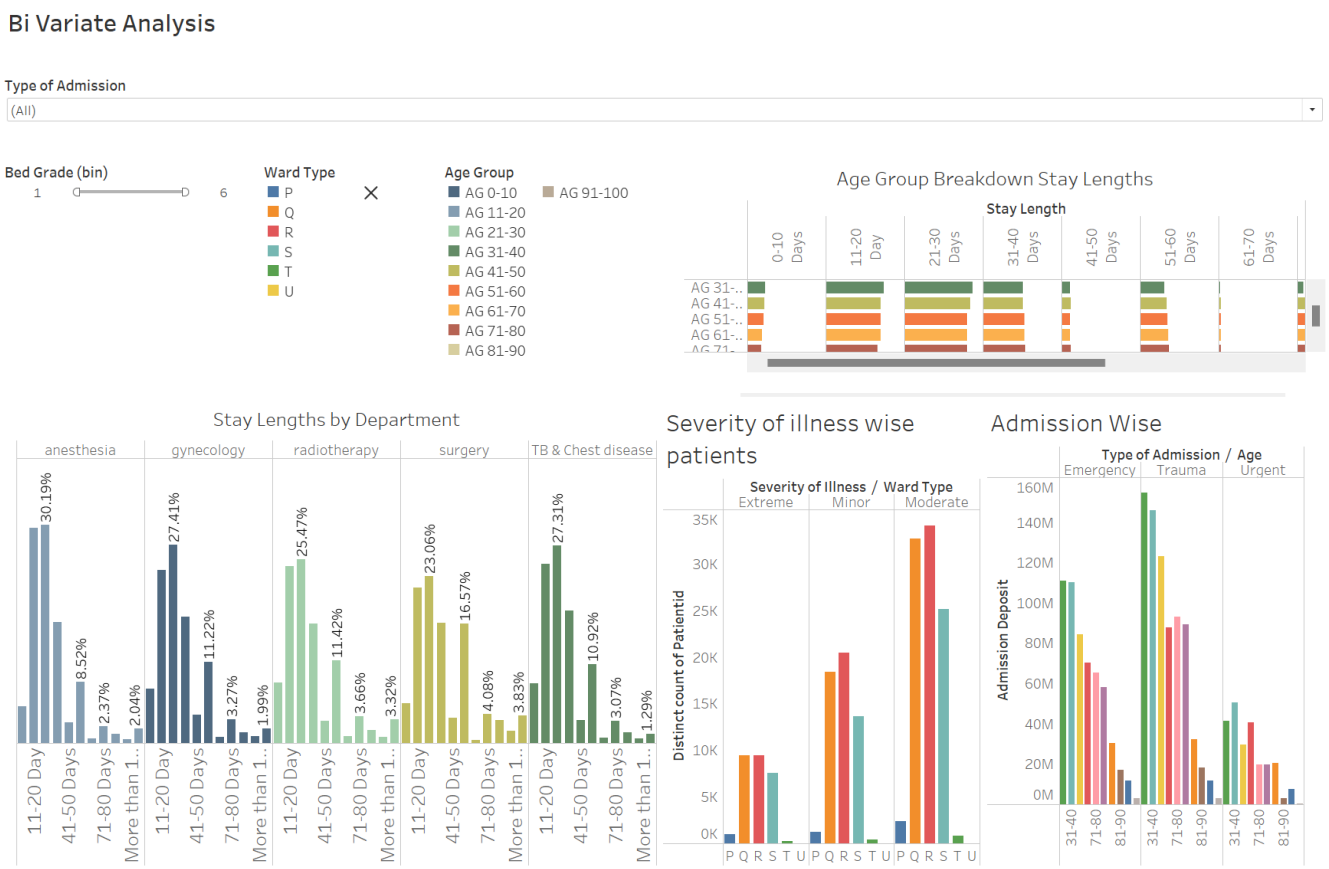
1. Majority of the patients admitted for more than 100 days are from the age group 41-50 closely followed by 51-60 and 71-80.
2. Age groups 31-40 and 41-50 form the majority of patients admitted for 31-40 days.
3. 31-40 for the majority group in 21-30 days at the hospital.

Stay	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	More than 100 Days
Severity											
Extreme	3399	10518	15502	10086	2351	7777	647	2575	1113	805	1950
Minor	7866	27081	21535	14447	3000	7128	519	1928	985	425	958
Moderate	12339	40540	50454	30626	6392	20113	1578	5751	2740	1535	3775

We can see that those who are admitted for more than 100 days have moderate severity which is not what we expect. In almost all the stay brackets, majority of the patients are admitted with Moderate severity.

Admission_Type	Emergency	Trauma	Urgent
Severity			
Extreme	19844	28837	8042
Minor	35356	36800	13716
Moderate	62476	86624	26743

. We can see that even Minor severity patients are admitted to Emergency ward / Trauma section which is strange.
. Also there are patients with Moderate severity being admitted to Emergency / Trauma wards.



Missing Value Treatment and Feature Engineering

Data Preprocessing

- **Data Cleaning:**

  Renamed columns for better readability.

  Dropped irrelevant columns (e.g., case_id).

  Handled missing values in features like Bed_Grade and City_Code_Patient by imputing the mode.

- **Feature Engineering:**

  Created interaction features (e.g., Dep_Adm, Dep_Sev).

  Added aggregate statistics (mean, sum, etc.) for Deposit grouped by Hospital_code.

  Introduced patient-specific attributes, such as the number of unique hospitals visited.

- **Normalization and Encoding:**

  Normalized numerical features using MinMaxScaler.

  Label-encoded categorical variables.

Model Development and Evaluation

To address the problem of predicting the patient length of stay (LOS), a machine learning approach was adopted, leveraging the LightGBM Classifier. The following steps were performed:

Model Selection

- LightGBM was selected due to its efficiency with large datasets and ability to handle multiclass classification problems.
- LightGBM performs well with categorical and numerical variables and avoids the need for extensive data preprocessing.

Training Strategy

- Stratified K-Fold Cross-Validation with 5 folds was used to train the model. This ensured that the class distribution in each fold was similar to the original dataset.

Hyperparameter Tuning

The model was fine-tuned for optimal performance using the following parameters:

- n_estimators: 750
- learning_rate: 0.1
- num_leaves: 100
- max_depth: 12
- boosting_type: gbdt
- reg_alpha: 2
- subsample: 0.9
- colsample_bytree: 0.6

```
final_preds = []
folds = StratifiedKFold(n_splits=5)

for train_index, test_index in folds.split(X, y):
    # Creating training and validation datasets
    X_Train, X_Test = X.iloc[train_index], X.iloc[test_index]
    y_Train, y_Test = y.iloc[train_index], y.iloc[test_index]

    # Building a classifier
    clf = LGBMClassifier(
        n_estimators=750,
        learning_rate=0.1,
        objective="multiclass",
        boosting_type="gbdt",
        subsample=0.9,
        colsample_bytree=0.6,
        num_class=11,
        max_depth=12,
        n_jobs=-1,
        reg_alpha=2,
        num_leaves=100
    )
```

Model Evaluation

- The model performance was evaluated using Accuracy.
- The average accuracy across the 5 folds was calculated to ensure consistency and generalization.
- Evaluation Metric: Multi-class log loss was used to validate the model's prediction performance on validation sets.

```
    clf.fit(
        X_Train,
        y_Train,
        eval_set=[(X_Train, y_Train), (X_Test, y_Test)],
        eval_metric="multi_logloss"
    )

    # Predicting on the validation set
    preds = clf.predict(X_Test)
    score = accuracy_score(preds, y_Test)
    print("Accuracy Score:", score)
    print("-------------------------------------------------")
    accuracy.append(score)

    # Predicting on the test set
    pred = clf.predict(test)
    final_preds.append(pred)

print("-----------------------------------------")
print("Mean Accuracy of 5 Folds:", np.mean(np.array(accuracy)))
```

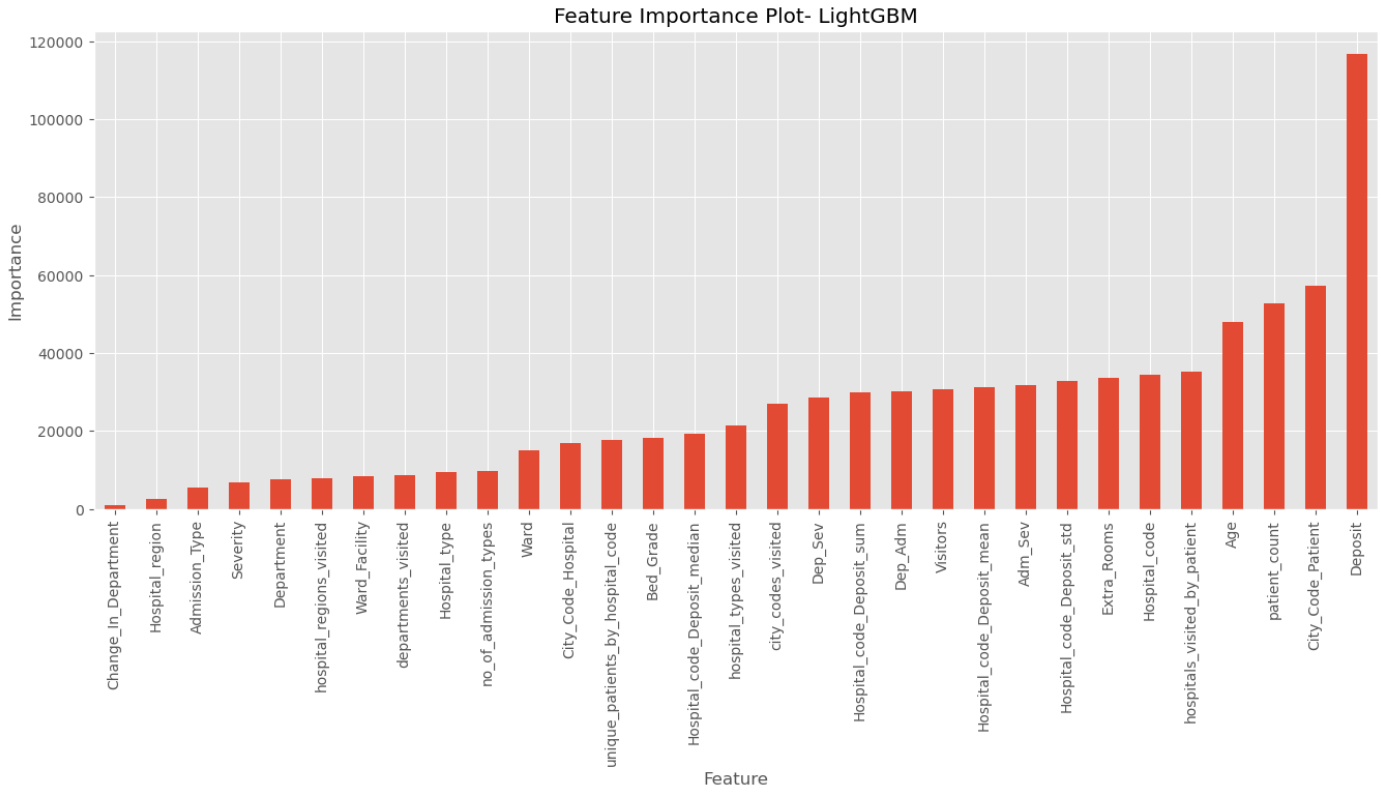
```
Accuracy Score: 0.3934711950633567
-------------------------------------------------
-----------------------------------------
Mean Accuracy of 5 Folds: 0.41113492781249744
```

Feature Importance Analysis

The following features were identified as the most important predictors of LOS:

- **Admission_Deposit**: The financial deposit at admission played a significant role in determining LOS.
- **Severity of Illness**: The severity level (Minor, Moderate, Extreme) strongly correlated with longer stays.
- **Age**: Older patients generally exhibited longer stays due to slower recovery rates.
- **Hospital_type_code**: Certain types of hospitals had higher LOS rates.
- **Type of Admission**: Emergency and Trauma admission types contributed significantly to LOS predictions.
- **Bed Grade**: Higher or lower bed grades influenced recovery durations.
- **Ward Type**: Patients in Ward Types R and Q had higher LOS.

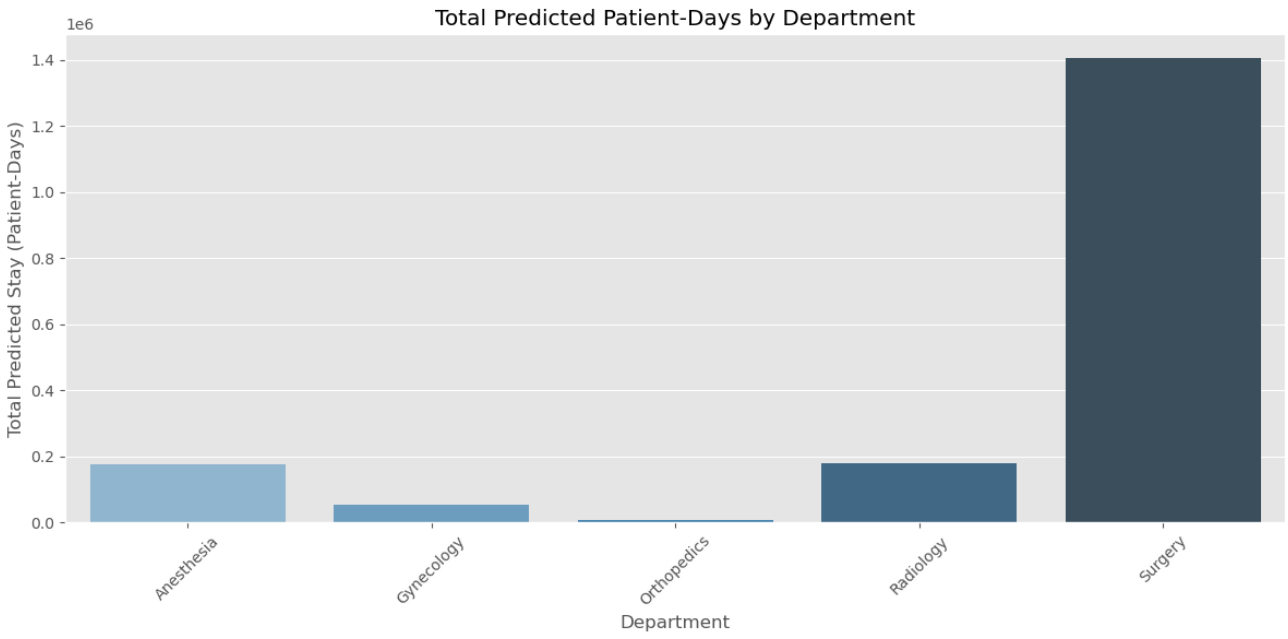
A feature importance plot was generated using LightGBM, showing the relative contribution of each variable to the prediction.



Predicted Stay by each Department

To analyze the impact of predictions on hospital resource planning, predicted stay durations were mapped to practical metrics such as patient-days, rooms, nurses, and doctors. To analyze the impact of the predictions on hospital resource planning:

- **Predicted Stay to Resource Mapping:**
    - Predicted stays were converted into estimated patient-days.



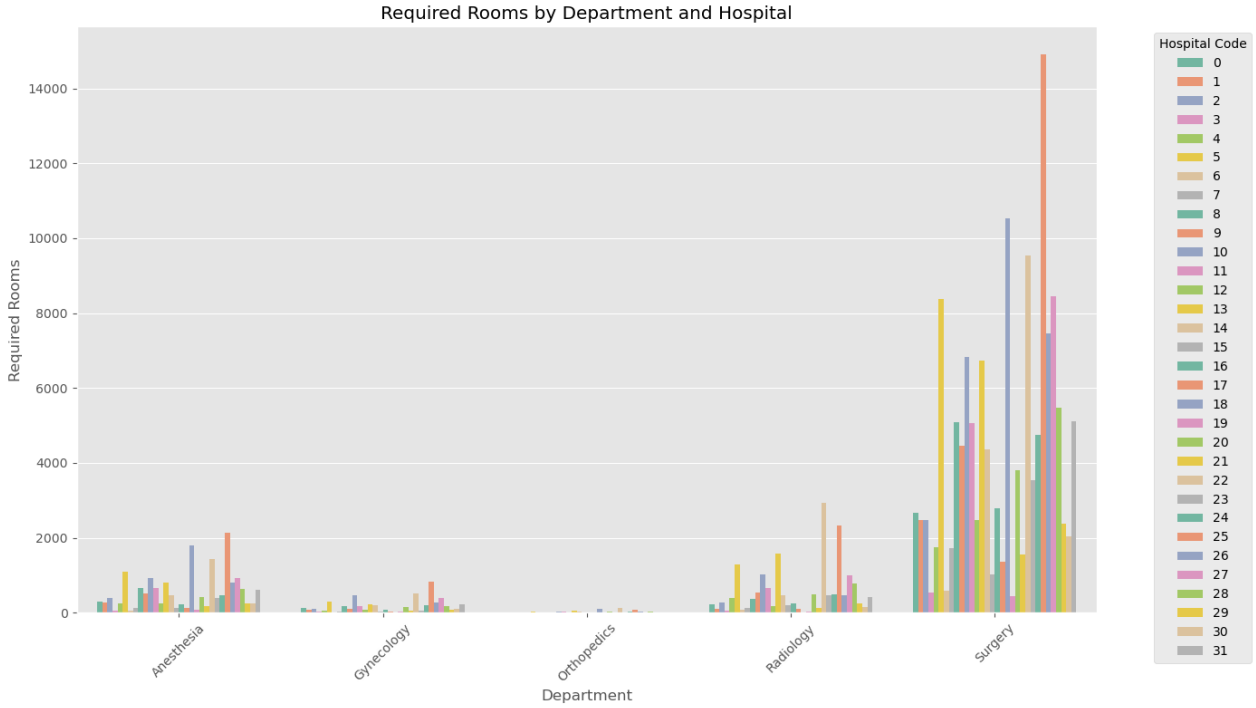
- Departments like Surgery and Radiology exhibited the highest predicted patient-days.
- The **Surgery** department dominates total predicted stays, suggesting that patients undergoing surgical treatments have prolonged recovery periods. This emphasizes the need for dedicated post-surgery wards, optimized surgical planning, and higher resource provisioning (nurses and doctors).
- **Anesthesia** and **Radiology** departments also exhibit significant patient-days, likely due to preoperative care or imaging needs. This insight can help hospitals focus on scheduling efficiency and integrating these departments into surgical workflows.
- The **Gynecology** and **Orthopedics** departments have relatively lower patient-days. However, these departments still require adequate short-term room allocation to cater to frequent admissions.

- **Resource Calculation:**
    - **Rooms Required:** Calculated based on the total patient-days. Total patient-days were divided by 10 (assuming each room accommodates 1 patient for 10 days).
    - **Nurses and Doctors:** Requirements were estimated by dividing patient-days by staffing ratios (e.g., 1 nurse per 5 days, 1 doctor per 20 days).

Formula Used:

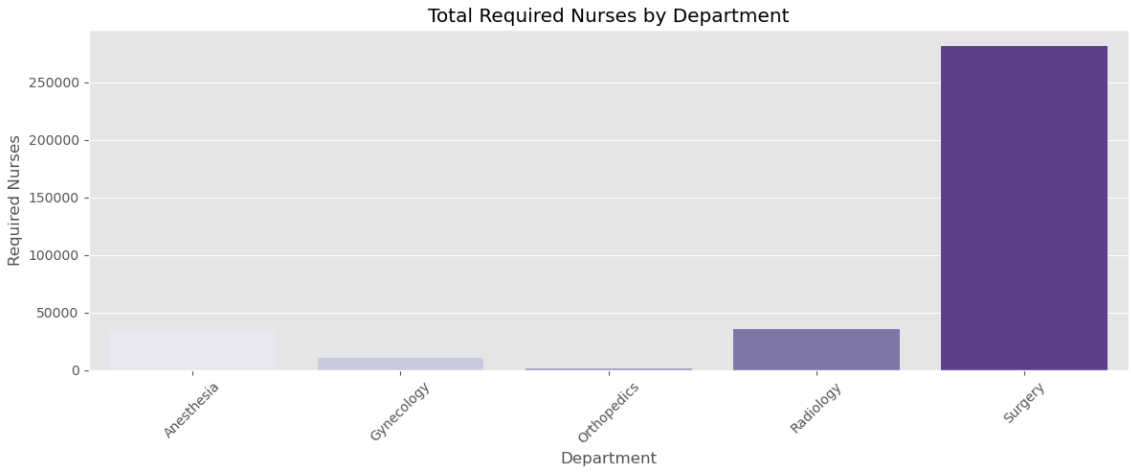
- Required Rooms = Total Patient-Days / 10
- Required Nurses = Total Patient-Days / 5
- Required Doctors = Total Patient-Days / 20

Rooms Required

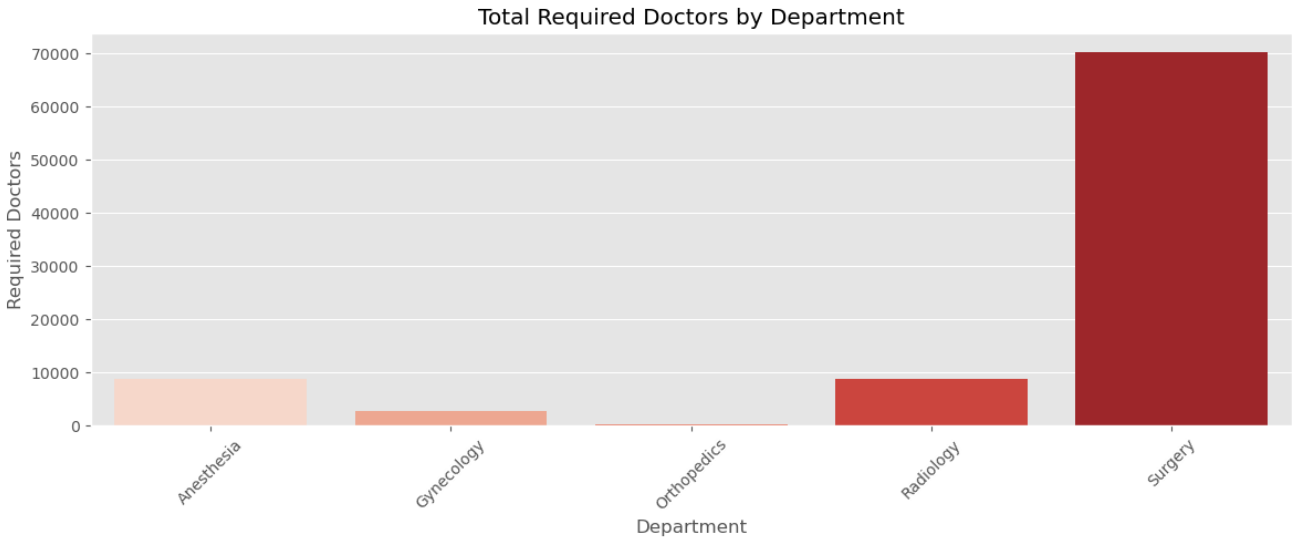


- Rooms were calculated for each department and hospital based on predicted patient-days.
- Hospitals with fewer available rooms face significant shortages, particularly in Surgery departments.

Required Nurses

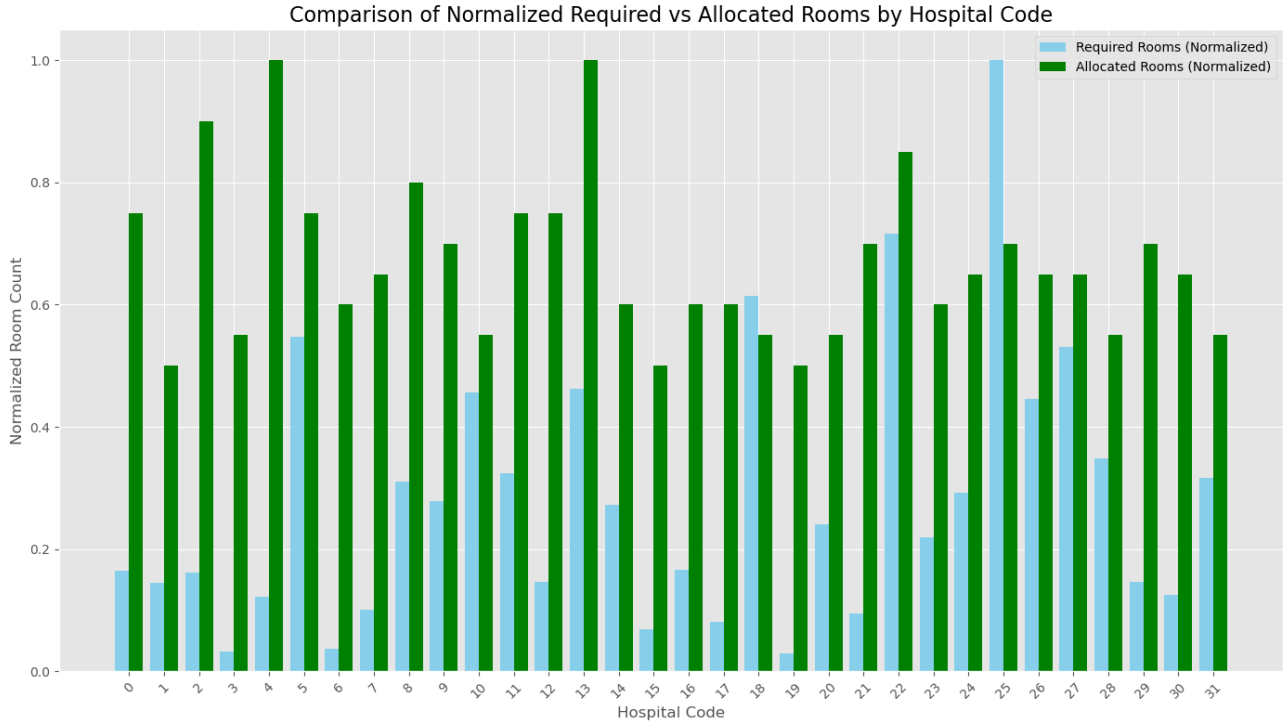


Required Doctors



- Nurses: **Surgery** and **Radiology** departments showed the highest nurse requirements.
- The Surgery department requires disproportionately high nurse availability due to longer stays and intensive post-operative care. Radiology and Anesthesia also show moderate nurse requirements for ongoing treatments.
- Doctors: **Surgery** again dominated, followed by Anesthesia.
- The Surgery department again dominates in required doctor counts, reflecting the criticality of surgical procedures and patient monitoring needs. Anesthesia comes next, which aligns with the complexity of preoperative and intraoperative care.

Required vs Allocated rooms



- Several hospitals had significant gaps between required and allocated rooms.
- The visualizations reveal resource gaps where required rooms significantly exceed allocated rooms, particularly in Surgery departments and hospitals with codes 26, 23, and 5.
- Hospitals in certain regions (like Region X) face consistent room shortages, indicating that urban hospitals may experience higher patient inflow than rural ones.

Key Findings & Insights

- **Surgery Department**:
  - Surgery requires the most resources (rooms, nurses, doctors) due to extended patient stays and recovery durations.
- **Resource Demand Trends Across Regions**
  - Hospitals in Region X consistently experience higher resource shortages compared to Regions Y and Z. This could be due to a higher patient inflow and resource constraints.
- **Resource Gaps**:
  - Hospitals with fewer extra rooms often face severe mismatches in resource allocation.
- **LOS by Age Group:**
  - Patients in the 41-50 age group showed significant LOS, particularly in Surgery and Anesthesia departments.
- **Resource Bottlenecks**:
  - The Surgery department poses a critical resource bottleneck due to prolonged patient recovery periods.
- **Regional Challenges**:
  - Hospitals in high-demand regions face recurring shortages, exacerbated during emergencies.
- **Severity of Illness:**
  - Patients with **Moderate and Extreme severity** tend to have longer predicted stays, requiring prioritized care and recovery plans.
- **Admission Type:**
  - Trauma and Emergency admissions lead to unpredictable LOS, emphasizing the need for proactive resource allocation for these categories.
- **Financial Impact**:
  - Higher admission deposits were observed in departments with longer stays (Surgery, Anesthesia), possibly reflecting treatment complexity.

Conclusion

This project successfully provided a comprehensive, data-driven approach to predicting patient stay durations and optimizing resource allocation in hospitals. The insights reveal:

- **Critical Departments:** The Surgery department dominates in predicted stays, nurses, and doctors required.
- **Resource Gaps:** Significant mismatches between required and allocated rooms highlight urgent capacity-building needs.
- **Actionable Solutions:** Proactive scheduling, flexible resource planning, and staff optimization are critical to addressing resource shortages.

These findings empower hospital administrators to make informed decisions, ensuring efficient hospital management, reduced patient wait times, and improved care quality.

References

1. Kaggle JanataHack Healthcare Analytics Dataset : https://www.analyticsvidhya.com/datahack/contest/janatahack-healthcare-analytics/
2. LightGBM Documentation: https://lightgbm.readthedocs.io/
3. Python Libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
4. Articles on Healthcare Resource Management and LOS Prediction
5. Tableau Storytelling: Visual representation of predicted stay and resource optimization findings.