AMITY UNIVERSITY

**RAJASTHAN**

AMITY INSTITUTE OF INFORMATION TECHNOLOGY

**NATURAL LANGUAGE PROCESSING**

**M.Sc. (Data Science)**

|  |  |
| --- | --- |
| **Student Name**  Sourabh Sharma  A217117723010 | **Course Instructor**  Mr. Amit Kumar Singh |

**REPORT**

**Using Predictive Analytics for Healthcare Operations Analysis:**

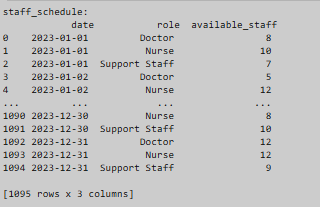
**1. Overview:**

Give a brief overview of the report's objective, which is to analyze operational data from the hospital to find inefficiencies and maximize resource use. Talk about the important areas of study and the datasets that were used, such as staff scheduling, patient wait times, and facility utilization. Using analytics to healthcare data improves both operational and quality outcomes. Forecasting equipment maintenance with the use of healthcare data analytics

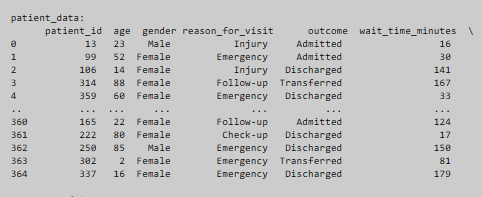
Most likely, the notebook opens by importing the required libraries, which include seaborn, matplotlib.pyplot, pandas, and numpy. For statistical analysis, data processing, and visualization, these libraries are indispensable.

**2. Initial Inspection and Data Loading:**

using pandas to load the "patient\_data," "staff\_schedule," and "facility\_usage" datasets. Next, it uses head() to show the top few rows of each dataset.

 A screenshot of a computer

Description automatically generated

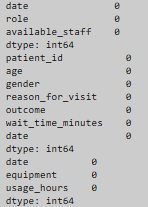
A screenshot of a computer

Description automatically generated

This step ensures the datasets are loaded correctly and provides an initial look at the structure and content of the data.

1. **Data Preprocessing:**

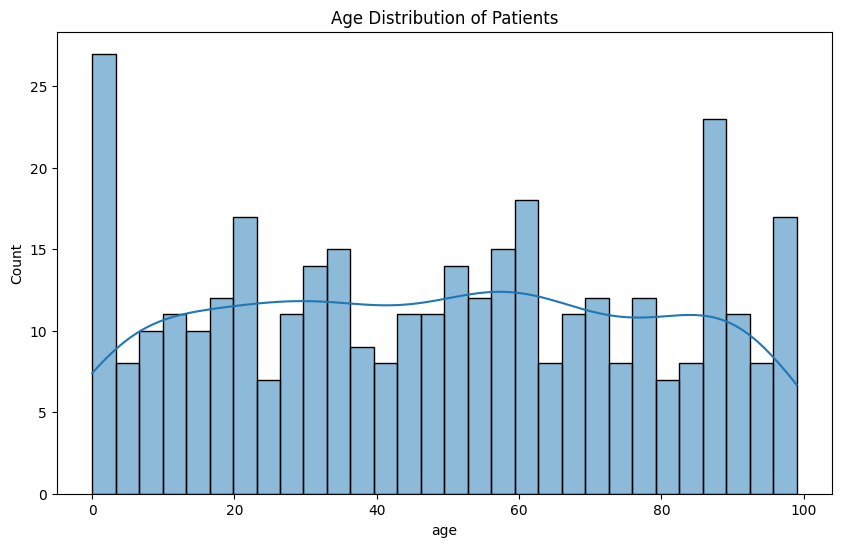
• Date Conversion: Use pd.to\_datetime() to convert date columns from text to datetime format. It sends two months' worth of data (Jan & Feb 2023).   
• Neglecting Principles In order to guarantee data quality, it may additionally verify the datasets for missing values.   
• Interpretation: Correct execution of time-based analysis is ensured by the conversion of dates.   
It's critical to look for missing values in order to spot possible problems that could interfere with analysis. The datasets don't contain any missing values.



1. **Exploratory Data Analysis (EDA)**

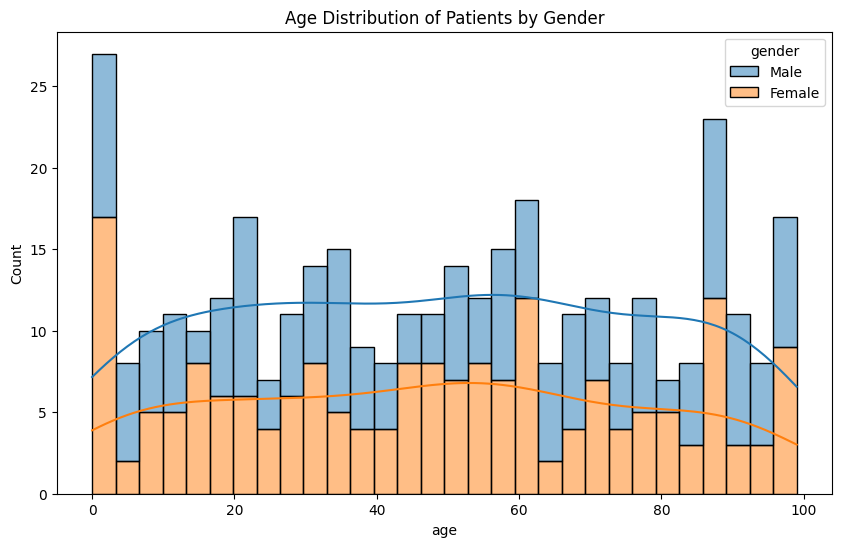
**Patient Data Analysis**:

* **Age Distribution**: A distribution of patient ages represented by a histogram. The age distribution of the patients is not very normal; the majority of the patients fall within the ranges of 0 to 10 and 80 to 100. It indicates that a significant portion of the patients are either elderly or youngsters.

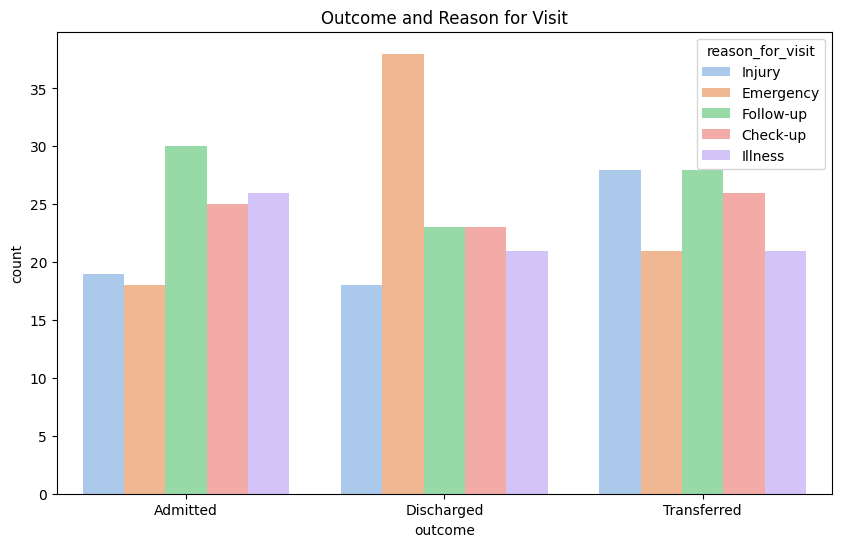
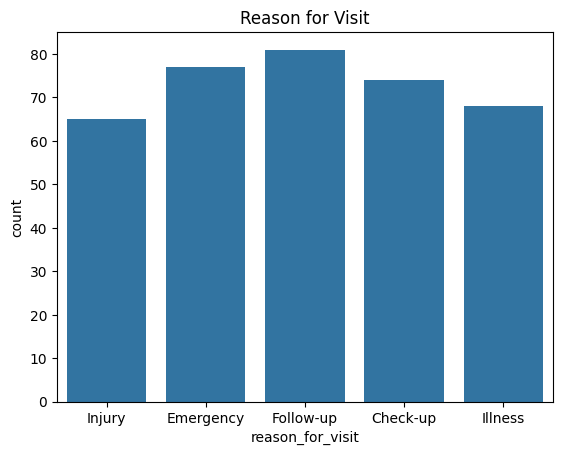


* **Gender Distribution**: a bar chart to show the patients' gender distribution. The proportion of female patients is slightly higher than that of male patients, but overall there aren't many distinctions between male and female patients.

A graph of a number of patients

Description automatically generated 

* **Reason for Visit with Outcome**: An additional bar plot illustrating the prevalence of various visitation causes.
* Admitted: Illness is the primary cause of admissions, followed by emergencies and injuries.
* Discharged: Follow-up visits account for the majority of discharges; check-ups and illness also play a big role.
* Transferred: Emergency cases account for the majority of transfers, although illness and injury also play a role.

The "Reason for Visit" bar chart indicates that the most frequent visits are for illness, follow-up, and check-up, with high counts approaching 80 for each. Significantly fewer injuries and ER visits occur. This implies that the facility should prioritize routine health management and preventative treatment.

* **Daily Patient Inflow:**

The number of patients fluctuates greatly, ranging from two to more than ten on any given day. When the number of patients surpasses 10, there are discernible peaks. These peaks can point to certain occurrences or circumstances that resulted in increased admissions. Additionally, there are times when there are only two patients, indicating days when there are fewer admissions.

A graph showing a patient inflow

Description automatically generated

* **Day-by-Day Analysis:**
* Peak Days: Based on median patient counts, Mondays and Tuesdays appear to be the busiest days.
* Low Days: Thursdays and Sundays are the least busy days, as seen by their lowest median numbers.
* Consistency: There is less fluctuation in the patient inflow on Wednesdays and Saturdays. A graph of blue boxes

  Description automatically generated
* **Staffing**: Allocate more staff on Mondays and Tuesdays to handle the higher patient inflow.
* **Resource Management**: Plan maintenance or training sessions on Thursdays and Sundays when patient inflow is lower.

1. **Predictive Modeling:**

**ARIMA Model Implementation**

The traditional time series forecasting technique is ARIMA. Autoregressive Integrated Moving Average is what it stands for. Three parameters make up the model: d (degree of differencing), q (moving average order), and p (autoregressive order).

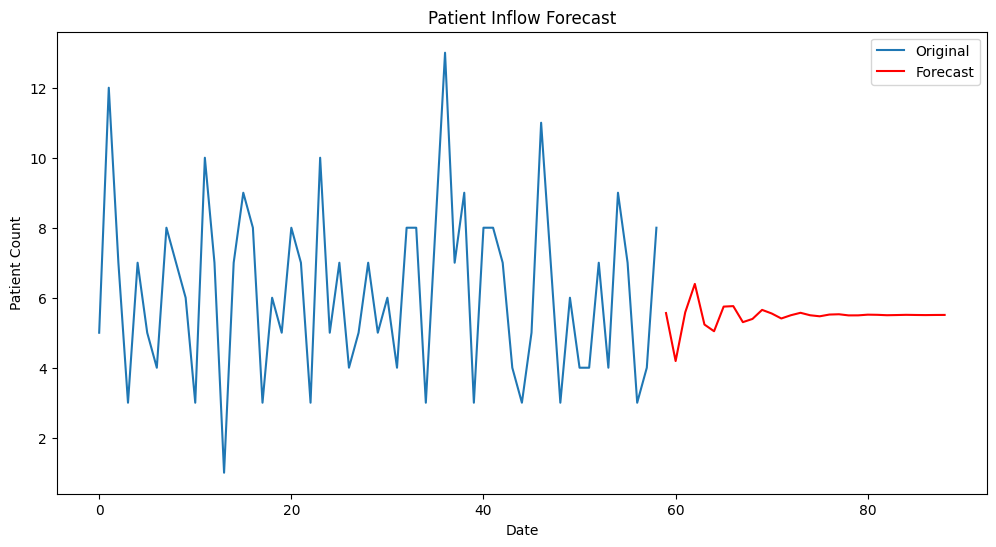
**Model Initialization**: The ARIMA model is initialized with the parameters (3, 1, 0), which represent the order of the model. These parameters are:

* + **p**: The number of lag observations included in the model (3).
  + **d**: The number of times that the raw observations are differenced (1).
  + **q**: The size of the moving average window (0).

**Fitting the Model**: The model is fitted to the daily\_patient\_inflow['Patient\_Count'] data.

**Forecasting**: The model is used to forecast patient inflow for the next 30 days (March 2023).

**Outcome**:



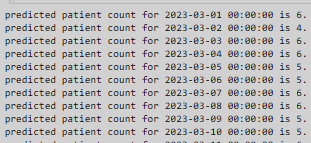
The anticipated data displays a smoother trend with less unpredictability than the original data, which displays notable swings in the patient count. This suggests that the ARIMA model has been successful in capturing the general patient inflow trend, offering a more reliable forecast for upcoming patient numbers. Planning and resource allocation in healthcare settings may benefit from this.

**Evaluation:**

MAE: 2.2122958097308296

RMSE: 2.5933130727549094

The forecast accuracy of the ARIMA model is demonstrated by the Mean Absolute Error (MAE) of 2.21 and the Root Mean Squared Error (RMSE) of 2.59. These numbers imply that there is an average deviation of 2 to 2.5 patients between the predicted and actual patient counts. Even if the model makes pretty accurate predictions, there is still space for development to further minimize these inaccuracies.



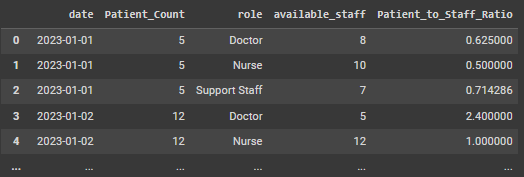
1. **Analyzing Staff Availability vs. Patient Inflow:**

**Goal:** Identify patterns where staff availability might be insufficient based on patient inflow.

**Method:**

Merge ‘patient\_data’ with ‘staff\_schedule’ on the date column.

Calculate the average number of patients per staff member for each day and identify days with high patient-to-staff ratios.



"**Patient-to-Staff Ratio Over Time**" is the title of the line graph in the image. It illustrates how the ratio of patients to hospital employees has evolved over time for three major staff categories: physicians, nurses, and support personnel. This is an explanation:

• **Physician (blue line):** Although it varies, the ratio often remains below the high ratio criterion.

• **Nurse (green line):** This line fluctuates a little bit as well, but it stays below the cutoff.

• **Support Staff (orange line):** This line is more erratic but generally remains below the threshold.

• **High Ratio Threshold (red dashed line):** This is a constant number that shows the highest ratio of patients to staff that is appropriate for the best possible outcome.

A graph showing a line of different colored lines

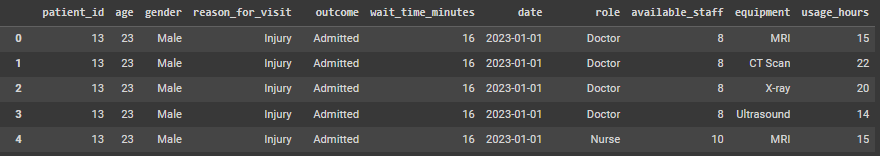
Description automatically generated with medium confidence

1. **Wait Time Analysis:**

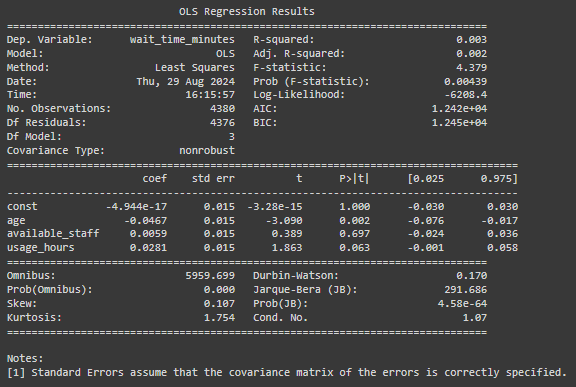
**Goal:** Determine factors that contribute to longer wait times.

**Method:**

* Examine wait\_time\_minutes in relation to several variables, including visitation purpose, personnel availability, and facility usage.
* To identify important factors influencing wait times, use regression analysis or correlation matrices.



the output of an Ordinary Least Squares (OLS) regression analysis, which is used to estimate the relationship between a dependent variable and one or more independent variables

****

* R-squared and Adjusted R-squared: Both values are very low (0.003 and 0.002), indicating that the model explains only a small fraction of the variability in wait times.
* Significant Variables: The variable ‘age’ has a statistically significant negative coefficient (P-value = 0.002), suggesting that as age increases, wait times slightly decrease.
* Other Variables The coefficients for ‘available\_staff’ and ‘usage\_hours’ are not statistically significant, indicating that these variables do not have a strong impact on wait times in this model.

This regression output provides insights into factors that might influence waiting times at a medical facility, with ‘age’ being the most significant predictor among the variables considered.

**Wait Time by Reason for Visit:**

It shows the distribution of wait times for different reasons for visiting a healthcare facility.

A diagram of a graph

Description automatically generated with medium confidence

* **Injury**: The box plot for injuries shows a relatively wide range of wait times, indicating variability in how long patients wait.
* **Emergency**: There appears to be more consistency in wait times since the wait times for emergencies are more closely packed.
* **Follow-Up**: The wait times in this category fluctuate moderately, with a few exceptions.
* **Check-Up:** Just like with injuries, wait times for check-ups can vary greatly.
* **Illness**: There are certain outliers in the moderate range of wait times for illnesses In brief:
* **Variability**: Depending on the purpose of the visit, wait times differ considerably. Wait times for emergencies are typically more consistent than those for checks-ups and injuries.
* **Standardization**: By standardizing the wait durations, it is now simpler to compare items across various categories.

1. **Facility Usage Optimization:**

**Goal:** Ensure that critical equipment is neither overutilized nor underutilized.

**Method:**

Examine usage\_hours for each equipment type against the number of patients and the types of visits (e.g., Emergency, Injury). Identify patterns of underutilization or overutilization and recommend adjustments to the facility schedule or staffing.

To identify underutilization or overutilization, we can compare the usage hours against the number of patients and types of visits. We will calculate the usage per patient and analyze the results.

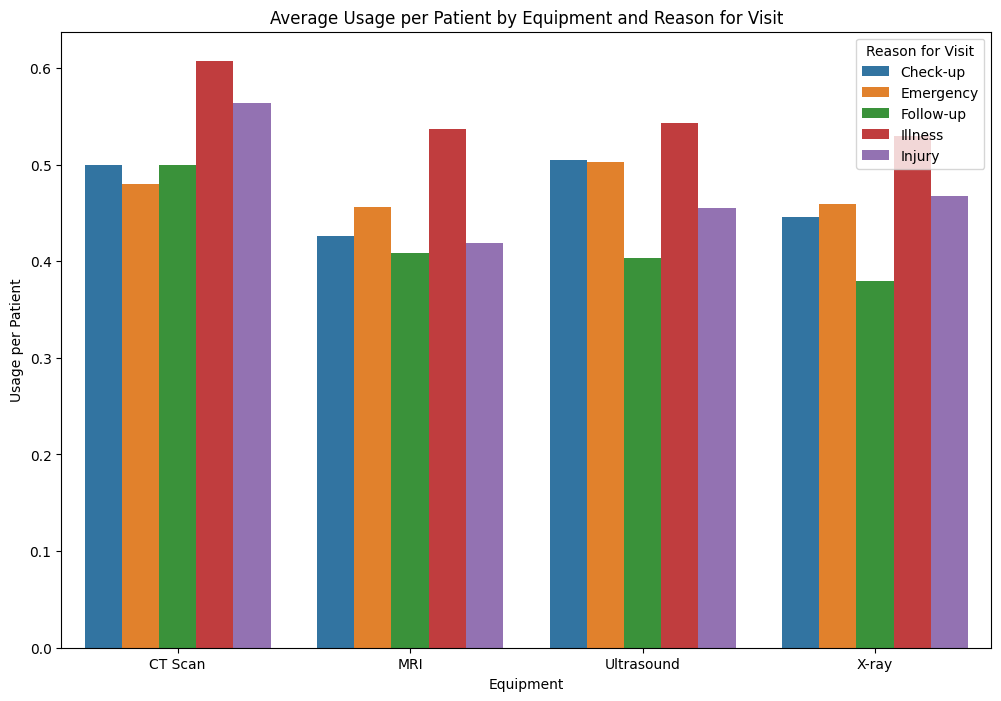
The ‘merge’ function is used to combine ‘patient\_data’ and ‘facility\_usage’ on the date column. We calculate ‘usage\_per\_patient’ by dividing the total ‘usage\_hours’ by the number of patients on that date. The grouped data shows the average usage per patient for each equipment type and reason for visit. This will help identify if certain types of equipment are overused or underused.

A screen shot of a computer

Description automatically generated

* **Average Usage per Patient by Equipment and Reason for Visit:**

It shows the average usage of different types of medical equipment for various reasons patients visit a healthcare facility.



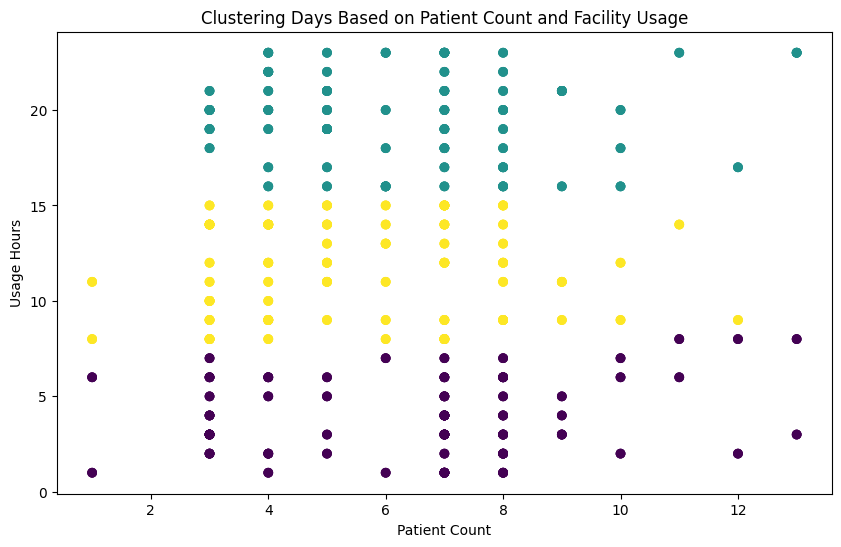
* **Usage per Patient**: Ranges from 0 to 0.6, indicating the average usage of each equipment type per patient is only 60% maximum.
* **Key Observations:**
  + **CT scan**: Highest usage for emergencies, followed by injuries and illnesses. Lower usage for check-ups and follow-ups.
  + **MRI:** Similar pattern to CT scans, with the highest usage for emergencies and injuries.
  + **Ultrasound**: More evenly distributed usage across different reasons, with slightly higher usage for emergencies and illnesses.
  + **X-ray:** Highest usage for injuries, followed by emergencies and illnesses. Lower usage for check-ups and follow-ups.
* **Summary:**
  + **Emergencies and Injuries**: These reasons generally have the highest usage of medical equipment, particularly CT scans, MRIs, and X-rays.
  + **Check-ups and Follow-ups:** These reasons have lower usage across all equipment types.
  + Illnesses: Moderate usage, with ultrasounds being used more frequently compared to other equipment types.

1. **Resource Allocation Strategy:**

**Goal**: Develop strategies for better resource allocation during peak times.

**Method**:

In order to decrease wait times and enhance patient outcomes, assign more staff or increase facility usage during high-demand periods, which can be identified by using segmentation or clustering techniques such as K-Means to segment days based on patient inflow and facility usage.



**Explanation:**

**• K-Means Clustering: This algorithm groups days into clusters based on similar patterns in patient inflow and facility usage. We assume 3 clusters (low, medium, and high demand) for simplicity. • Visualization: The scatter plot helps us visualize the different clusters and identify which days fall into high-demand periods. • Features for Clustering: We use "patient\_count" and "usage\_hours" as features to identify periods of high demand.**

* **K-Means Clustering:** This algorithm groups days into clusters based on similar patterns in patient inflow and facility usage. We assume 3 clusters (low, medium, and high demand) for simplicity.
* **Visualization:** The scatter plot helps us visualize the different clusters and identify which days fall into high-demand periods.
* **Features for Clustering:** We use "patient\_count" and "usage\_hours" as features to identify periods of high demand**.**

A screen shot of a computer

Description automatically generated

The three separate clusters' average patient counts, personnel availability, and equipment usage hours are displayed in the cluster summary. Cluster 0 has minimal facility demand, with an average of 7 patients, 10 staff members, and approximately 3.88 utilization hours. With six patients and nine employees, Cluster 1 had the highest average equipment usage (19.77 hours), indicating a high level of facility demand. Six patients and nine staff members make up Cluster 2, which has a modest consumption of roughly 11.57 hours. Understanding the disparities in demand and resource distribution among various clusters is made easier by this analysis.

**Explanation:**

* **Cluster Summary**: For each cluster, we compute the average patient population, available personnel, and usage hours to gain insight into how resources are used during varying demand times.
* **Adjusting Staff**: During times of strong demand, we boosted personnel availability by twenty percent. This is merely an illustration; the modification may be more intricate based on the hospital's particular requirements and limitations.
* For the sake of simplicity, let's say that we increase staffing by 20% when there is a strong demand (cluster with highest patient count).

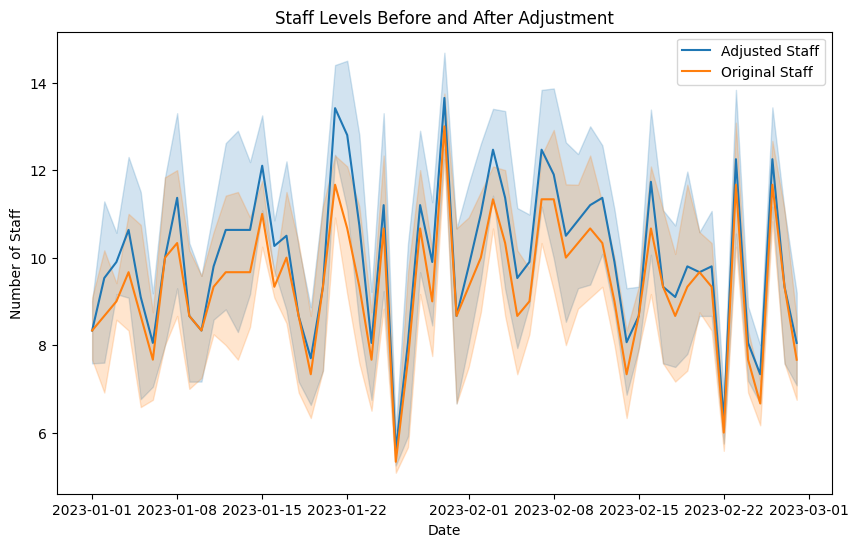
A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

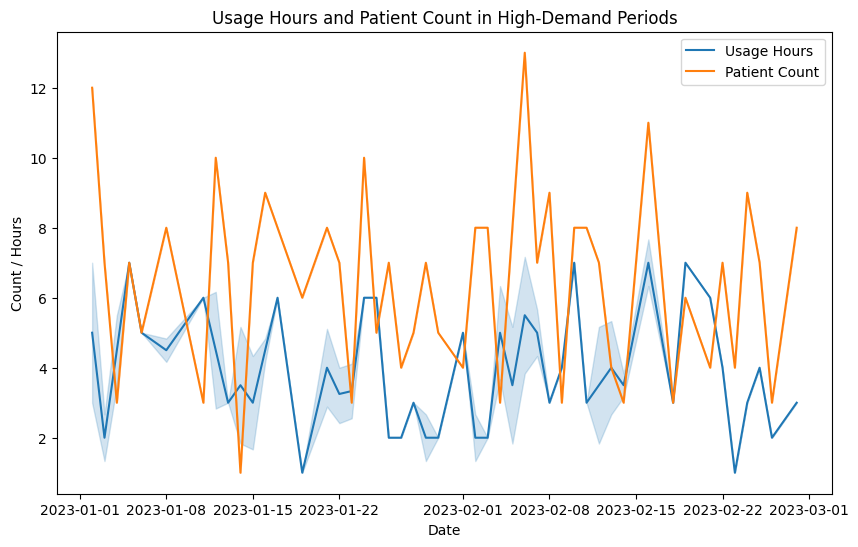
Description automatically generated

**Evaluation the Impact of the Strategy:**

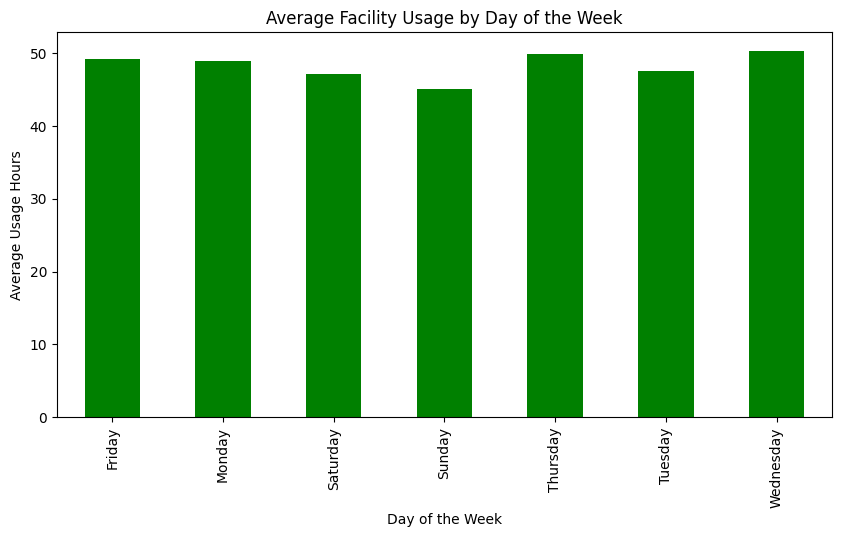
* **Staff Levels Visualization:** We make sure that enough resources are assigned during peak hours by comparing the initial and modified staffing levels.
* The following graph shows the difference between the original staff and the adjusted staff after a 20% increase in staff.



* **High-Demand Analysis:** To determine whether the changes had the intended impact, we examine how patient counts and usage hours react during times of high demand.
* The following graph shows that, with a 20% adjustment in staff, patient counts are rising while usage hours are down.



* As we can see, Monday, Thursday, Wednesday and Friday consume the highest usage hours so hospital management should increase their staff on these peak days.



1. **Facility Utilization Efficiency:**

Examine the utilization of the hospital's equipment (MRI, CT scan, X-ray, etc.) in relation to the staff members on hand.

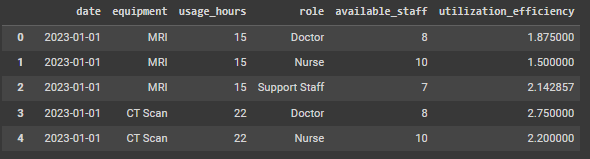
**Steps:**

* **Calculate Usage Efficiency:** Determine the ratio of personnel availability to facility utilization hours for each day.
* **Identify Under/Over-Utilized Facilities:** Examine the relationship between various jobs (e.g., nurses vs. doctors) and facility usage to see if some staff members are more essential for effective facility usage.
* **Compare Across Different Roles:** In order to combine the "staff availability" and "facility usage" data, merge the two databases on the date column. To find the utilization efficiency for each row, divide the usage\_hours by the available\_staff.

Merge the two datasets on the date column to combine ‘facility\_usage’ data with ‘staff\_availability’.

For each row, calculate the utilization efficiency by dividing the usage\_hours by the available\_staff.

Now that we have the utilization efficiency, we can analyze it across different facilities or staff roles.



**Underutilization and Overutilization Summary:**

* **MRI:**

**Doctors, Nurses, and Support Staff** all show high utilization efficiency, with bars extending beyond the ‘1’ mark. This indicates potential **overutilization**.

* **Ultrasound:**

**Doctors** have slightly higher utilization efficiency compared to Nurses and Support Staff, but none of the bars extend beyond ‘1’. This suggests **moderate utilization**.

* **X-ray:**

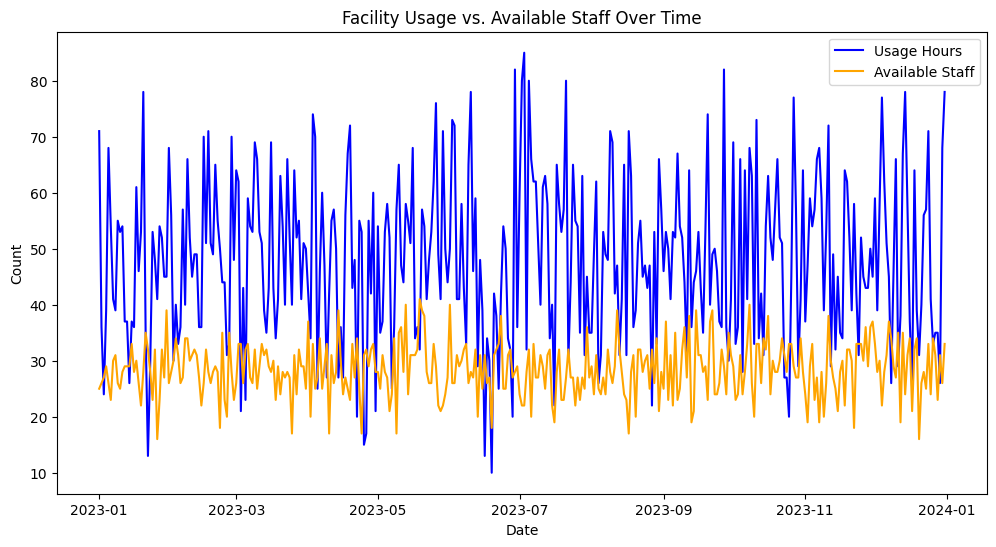
**Doctors** show higher utilization efficiency compared to Nurses and Support Staff. The bars for Nurses and Support Staff are closer to ‘0’, indicating **underutilization** for these roles.

* **CT Scan:**

**Doctors** show the highest utilization efficiency, with their bar extending furthest along the x-axis, indicating **overutilization**. Nurses and Support Staff have moderate utilization.

A graph of a number of colored bars

Description automatically generated with medium confidence



**SUMMARY:**

In order to comprehend relationships and conduct predictive analysis, we began by exploring three datasets: patient\_data, staff\_schedule, and facility\_usage. Our initial discussions centered on merging datasets in order to compute unique patient metrics and running ARIMA in order to predict patient wait times based on variables such as age, available staff, and equipment usage hours. Despite some technical difficulties with the regression model and data merging process, we improved the analysis by making sure calculations were performed on unique patient IDs and experimenting with different features for more precise predictions.

Then, using only the facility\_usage and staff\_schedule datasets, we looked at different analysis options, such as clustering techniques to find patterns in facility usage in relation to staff availability. We also talked about report preparation, where you were supposed to record the entire analysis process, including the insights gained from the data, the steps taken for model building, and the interpretation of results.

After this analysis, we suggest using different facilities, rather than under- or overusing them. Throughout, the emphasis remained on comprehending the data, using the proper modeling techniques, and interpreting the results for actionable insights, culminating in a thorough report for your assignment.

**THANK YOU :)**

**SOURABH SHARMA  
A217117723010**