OSTIM TECHNICAL UNIVERSITY
FACULTY OF ENGINEERING
DATA MINING COURSE
FINAL PROJECT
Hospital Data Warehouse

Student Name Surname: Moavia Mahmood
Student No: 210208693

Project Consultant:
Doç. Dr. Meltem Eryılmaz

Data Mining – SENG 312

Ankara - 2024

Index

1. Introduction

Hospital Name: Services Hospital

Address: Ghaus-ul-Azam (Jail) Road, Shadman, Lahore, Pakistan

Contact Person: Tahira Noreen (Head Nurse)

1.1. Hospital Information:

Services Hospital is a major public hospital in Lahore, Pakistan, and one of the largest healthcare facilities in the city. It offers comprehensive medical services including emergency care, inpatient and outpatient services, surgical procedures, diagnostic services, and specialized medical treatments.

The hospital serves thousands of patients daily and is known for its commitment to providing affordable healthcare. It is also a teaching hospital associated with the **Services Institute of Medical Sciences (SIMS)**, training future doctors and nurses.

The hospital is equipped with over 2,000 beds and provides services across various specialties, including cardiology, neurology, orthopedics, obstetrics, gynecology, pediatrics, and oncology.

2. Problems faced by nursing staff

Services Hospital faces challenges that are common in large public healthcare institutions, particularly those related to data management and hospital resource optimization. Key problems include:

2.1. Lack of Integrated Data Systems:

The hospital departments (such as emergency, surgery, diagnostics, and wards) operate on separate information systems. As a result, patient data is often fragmented, leading to delays in accessing vital information and inefficient care coordination.

For instance, the surgical department may not have immediate access to a patient's prior medical records from the diagnostic department, causing delays in surgical decisions.

2.2. Overburdened Resources and Staff:

Due to the hospital's high patient load, resources like beds, equipment, and staff (especially nursing staff) are often overextended. This has resulted in long wait times for certain treatments, especially in the emergency department.

Nurse Tahira Noreen, as head nurse, struggles to manage nurse scheduling efficiently due to a lack of comprehensive data on patient volumes, staff availability, and bed occupancy. The absence of real-time data makes it difficult to predict when additional nursing resources will be required.

2.3. Manual Reporting Processes:

Hospital staff, including administrative and medical personnel, rely on manual processes to create reports on patient admissions, discharges, treatment outcomes, and resource utilization. This slows down decision-making and leads to a higher risk of errors.

2.4. Inefficient Patient Flow:

The hospital experiences overcrowding in certain departments, particularly emergency care, due to inefficient patient flow management. Patients often face delays in being transferred from the emergency department to specialized wards or surgery due to a lack of coordination between departments.

3. Current Situation of problem

The hospital has been providing essential medical services to the local community for decades, but its operational efficiency is being compromised due to its outdated information systems and lack of integrated data across departments. Some key issues with the current situation include:

3.1. Fragmented Systems:

Each department uses a separate database or system for maintaining patient records, resource allocation, and scheduling. This makes it difficult to have a unified view of patient care across the hospital.

3.2. High Patient Load:

The hospital faces a high influx of patients, particularly in the emergency department, resulting in extended waiting times and increased pressure on staff, especially nurses. With limited data access, it becomes difficult to balance the staff workload efficiently.

3.3. Manual Workflows:

Due to a lack of digitization, many hospital functions are still handled manually, from generating reports to coordinating patient transfers between departments. This results in delays in decision-making and slower responses to emergencies.

The hospital administration, led by Tahira Noreen and other senior medical officers, is exploring the possibility of integrating a centralized data system to streamline operations, enhance patient care, and reduce resource strain.

4. Expectations from the Developed System

4.1. Quick Access to Patient Information:

Nurses expect a streamlined, single access point to view critical patient data such as treatment plans, medication schedules, and doctor notes. This ensures they can quickly access up-to-date information without navigating multiple systems.

4.2. Real-Time Resource and Bed Availability:

The system should provide real-time updates on bed occupancy, discharge schedules, and equipment availability. This allows nurses to manage patient admissions and resource allocation effectively, especially during peak times or emergencies.

4.3. Efficient Medication and Treatment Tracking:

Nurses need the ability to track medications, dosages, and treatment schedules accurately to reduce errors in medication administration. The system should help monitor patient responses to treatments, flagging any issues promptly.

4.4. Seamless Cross-Department Coordination:

Easy access to patient information from other departments (e.g., lab results from diagnostics, prescription data from pharmacy) would help nurses provide more coordinated care and reduce redundant tasks like reordering tests or medications.

4.5. Streamlined Communication and Updates:

The system should facilitate timely updates and communication among healthcare providers, alerting nurses to changes in treatment plans or patient status. This would reduce miscommunication and improve patient safety.

4.6. Data for Efficient Scheduling and Task Management:

Nurses expect the system to support workload management by helping supervisors plan shift schedules based on patient volumes and bed occupancy trends. Real-time data would help allocate resources efficiently, ensuring nurses are available where they’re most needed.

4.7. Enhanced Reporting and Documentation:

A simplified process for documenting patient interactions and treatment notes in the data warehouse would save time and improve data accuracy. Quick access to past documentation can also aid in patient assessments and ongoing care planning.

5. Questions the developed system should address

Here are key questions a nurse would need the developed hospital data warehouse system to answer:

5.1. Patient Care and Treatment:

- What is the patient's current treatment plan and medication schedule?
- Are there any recent updates or changes to the patient's treatment or medication?
- What are the patient's past medical history and previous diagnoses?

5.2. Resource Availability:

- Are there available beds for incoming patients in specific departments?
- Is the necessary equipment (e.g., IV pumps, monitors) available and ready for use?
- Are there sufficient medical supplies and medications for current patients?

5.3. Cross-Departmental Coordination:

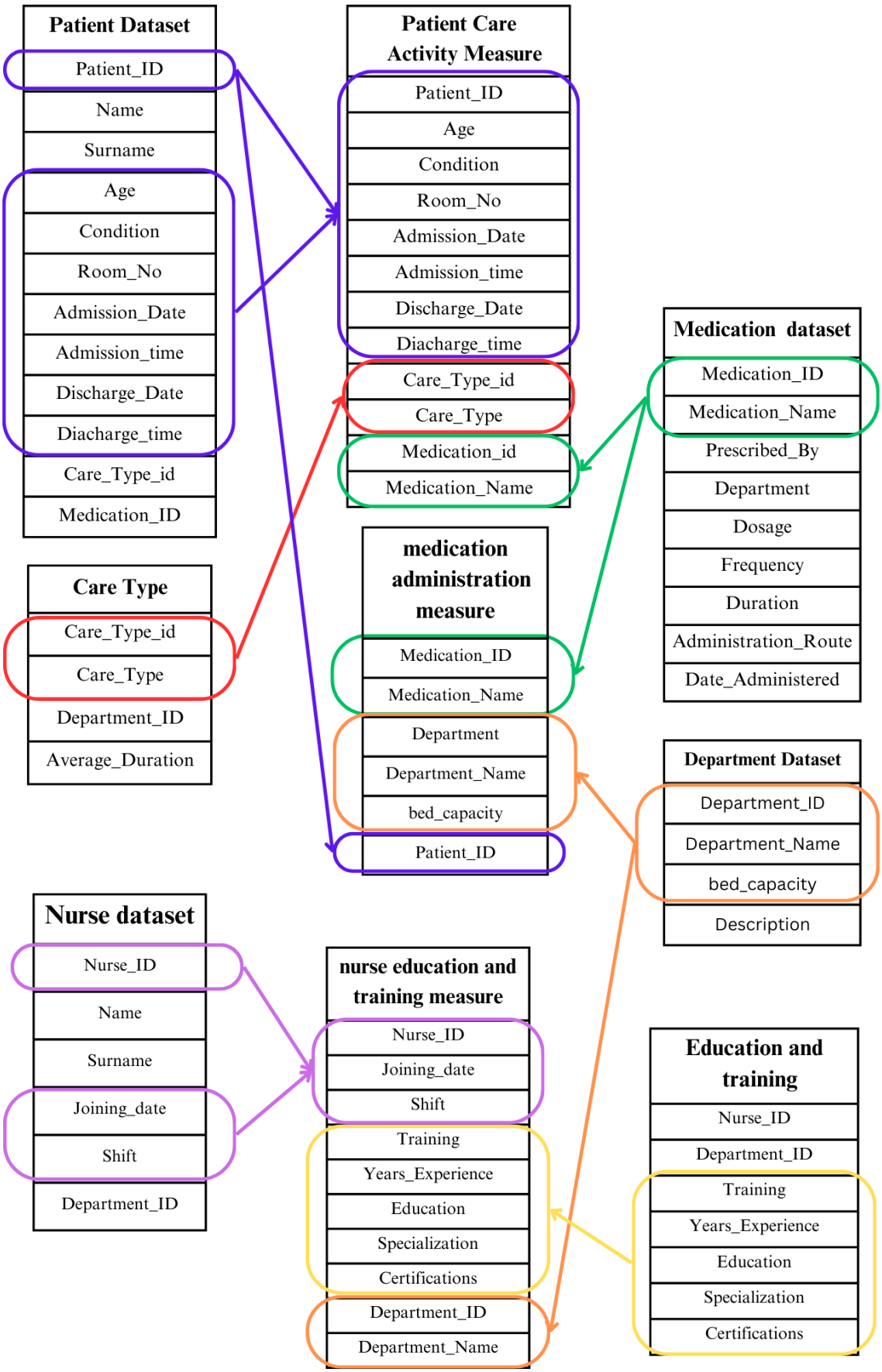
- What are the most recent lab results and diagnostic findings for my patients?
- Has the patient received necessary prescriptions, and are they ready for administration?
- Are there any pending test results that will impact ongoing treatment?

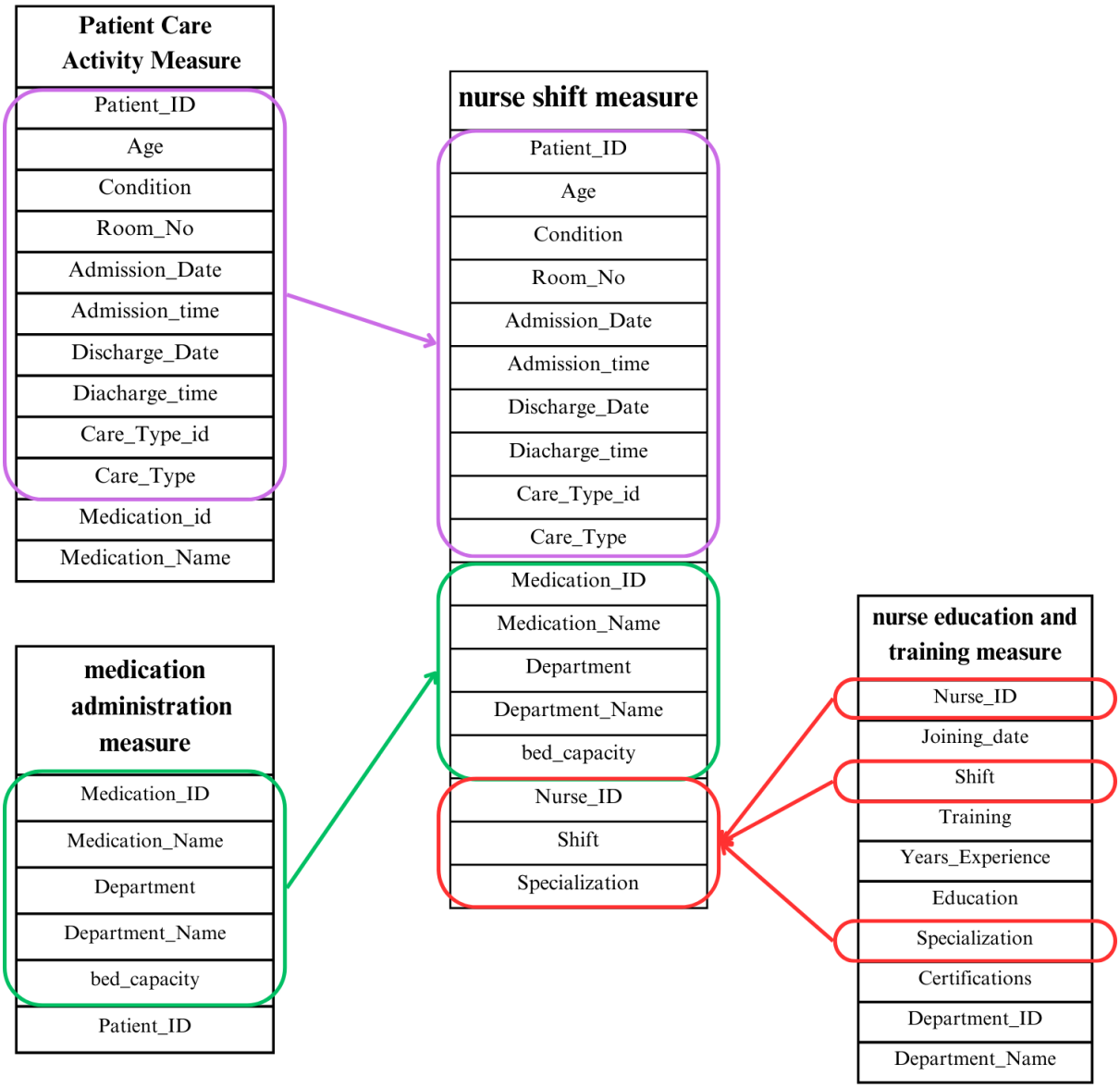
5.4. Scheduling and Workload Management:

- What are the patient volumes and staffing requirements for the upcoming shifts?
- Are there high-priority tasks or critical patients requiring immediate attention?
- Which departments are currently under high demand and may need additional nursing support?

These questions help's deliver efficient, well-coordinated, and safe patient care.

6. Conceptual Modeling

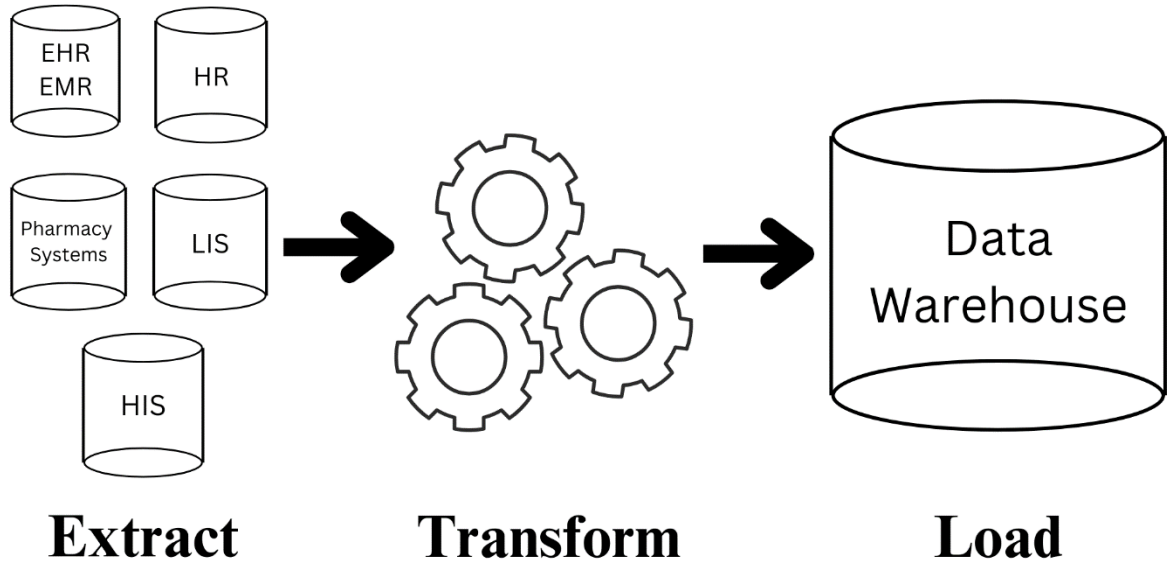




Fact tables store quantitative data that can be analyzed and aggregated.
Dimension tables represent descriptive attributes that give context to the fact data. These dimensions provide the "who," "what," "where," and "how" for the facts.

7. ETL (Extract, Transform, Load)

ETL is a process used in data management to move and process data from various sources into a data warehouse or other storage systems for analysis



7.1. Extract

The first step is to extract data from various source systems in the hospital. These could include:

- **Electronic Health Records (EHR/EMR)**: Patient data, medical history, diagnoses, treatments, etc.
- **Hospital Information System (HIS)**: Patient admission, discharge, and transfer details, billing, scheduling, etc.
- **Pharmacy Systems**: Medication data, prescriptions, dosage information.
- **Laboratory Information System (LIS)**: Test results, lab reports, diagnoses.
- **Nurse Scheduling System**: Information on nursing shifts, nurse assignments, and working hours.
- **HR and Payroll Systems**: Nurse information, salary, training records, etc.
- **External Data Sources**: For example, government health data or insurance records.

7.2. Transform

The transformation stage is where the data is cleaned, enriched, and structured for analysis. Key operations include:

- **Data Cleansing**: Remove errors, duplicates, and incomplete data. For example, fixing missing patient IDs or correcting mis-formatted dates.
- **Data Integration**: Combine data from different sources. For instance, patient data from EHR and billing data from HIS may need to be merged based on a unique patient ID.
- **Data Aggregation**: Summarize data where necessary. For instance, aggregating daily medication administration data to get weekly or monthly totals.
- **Data Enrichment**: Enhance the data by adding missing information. For example, enriching nurse data with additional information about their education or certifications.
- **Data Transformation**: Convert data into a format suitable for the data warehouse. This could involve normalizing data, creating derived fields (such as calculating the average length of stay for patients), or converting time-based data into time dimensions.
- **Data Mapping**: Mapping source data to the target data warehouse schema, which could involve creating surrogate keys (unique identifiers) and establishing relationships between different tables (fact and dimension tables).

7.3. Load

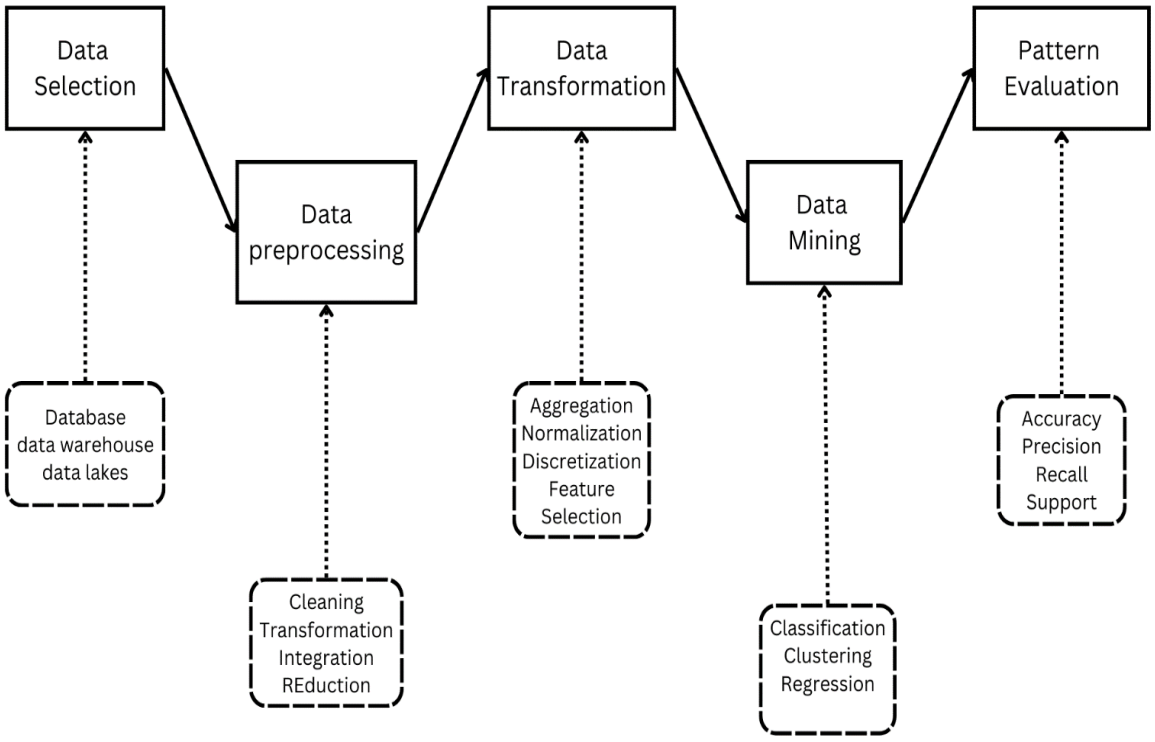
In the loading phase, the transformed data is loaded into the data warehouse.

Batch Load: Data is loaded in batches at specified intervals (e.g., daily, weekly).

We will use Batch Load in this process because data is generated and loaded at specific

Intervals.

8. KDD Process



8.1. Data Selection

Choosing the relevant data from different sources, such as databases, data warehouses, or data lakes. Identify the target data for analysis (e.g., hospital records, patient demographics, nursing schedules).

8.2. Data Preprocessing

Cleaning and preparing the data for analysis. Raw data can be messy, inconsistent, and incomplete, so preprocessing is essential.

- **Cleaning**: Handling missing data, correct errors, remove outliers, and fill missing values.
- **Transformation**: Converting data into a suitable format. For example, numerical values might need normalization, or categorical values may need encoding.
- **Integration**: Combining data from different sources into a unified format (e.g., merging patient records with billing or treatment data).
- **Reduction**: Reducing the data size by removing unnecessary features or aggregating data, which speeds up the analysis process.

8.3. Data Transformation

Transforming the data into a form that is suitable for the mining process.

- **Aggregation**: Summarizing data at a higher level, such as calculating daily medication totals or aggregating patient visits into monthly reports.
- **Normalization**: Scaling data so that values are within a certain range (e.g., converting patient age to a range from 0 to 1).
- **Discretization**: Grouping continuous data into discrete intervals (e.g., age groups such as 0-18, 19-35, etc.).
- **Feature Selection**: Identify the most relevant features for the analysis to reduce complexity and avoid overfitting.

8.4. Data Mining

Applying algorithms to extract patterns, trends, and relationships from the data.

- **Classification**: Assigning data to predefined categories (e.g., classifying patients as "high risk" or "low risk" based on their medical history).
- **Clustering**: Grouping similar data points together (e.g., clustering patients with similar health conditions or behaviors).
- **Association Rule Mining**: Finding relationships or associations between variables (e.g., identifying medications that are commonly prescribed together).
- **Regression**: Predicting a continuous outcome based on input variables (e.g., predicting patient recovery time based on treatment type and health condition).
- **Anomaly Detection**: Identifying unusual patterns or outliers in data (e.g., detecting errors in medical billing or fraudulent claims).

8.5. Pattern Evaluation

Evaluating the discovered patterns to identify the most interesting, valid, and useful ones.

- Evaluating the significance and relevance of discovered patterns using metrics such as **accuracy**, **precision**, **recall**, **support**, and **confidence** (for association rules).
- For instance, in a hospital setting, evaluate whether the discovered patterns (e.g., relationship between nurse shifts and patient care quality) are practically useful.
- Use **statistical validation** to ensure the patterns are not due to random chance.

9. Brief information about Input Data

9.1. Care Type

- **Number of Samples:** 14
- **Attributes**:
  - Care_Type_id
  - Care_Type
  - Department_ID
  - Average_Duration

9.2. Nurses

- **Number of Samples:** 33
- **Attributes**
  - Nurse_ID
  - Name
  - Surname
  - Joining_date
  - Shift
  - Department_id

9.3. Department

- **Number of Samples:** 36
- **Attributes:**
  - Department_ID
  - Department_Name
  - bed_capacity
  - Description

9.4. Medications

- **Number of Samples:** 10
- **Attributes:**
  - Medication_ID
  - Medication_Name
  - Prescribed_By
  - Department
  - Dosage
  - Frequency
  - Duration
  - Administration_Route

9.5. Patients Dataset

- **Number of Samples:** 220
- **Attributes:**
  - Patient_ID
  - Name
  - Surname
  - Age
  - Condition
  - Room_No
  - Admission_Date
  - Admission_time
  - Discharge_Date
  - Discharge_time
  - Care_type_id
  - Medication_id

9.6. Nurse Education & Training

- **Number of Samples:** 33
- **Attributes:**
  - Nurse_ID
  - Department_ID
  - Training
  - Years_Experience
  - Education
  - Specialization
  - Certifications

9.7. Patient Care Activity

- **Number of Samples:** 220
- **Attributes:**
  - patient_id
  - age
  - condition
  - room_no
  - Admission_Date
  - Admission_time
  - Discharge_Date
  - Discharge_time
  - Care_Type_id
  - Care_Type
  - Medication_ID
  - Medication_Name

9.8. Medication Administration

- **Number of Samples:** 36
- **Attributes:**
  - medication_ id
  - Medication_Name
  - Department_ID
  - Department_Name
  - bed_capacity
  - Patient_id

9.9. Nursing education and training Measure

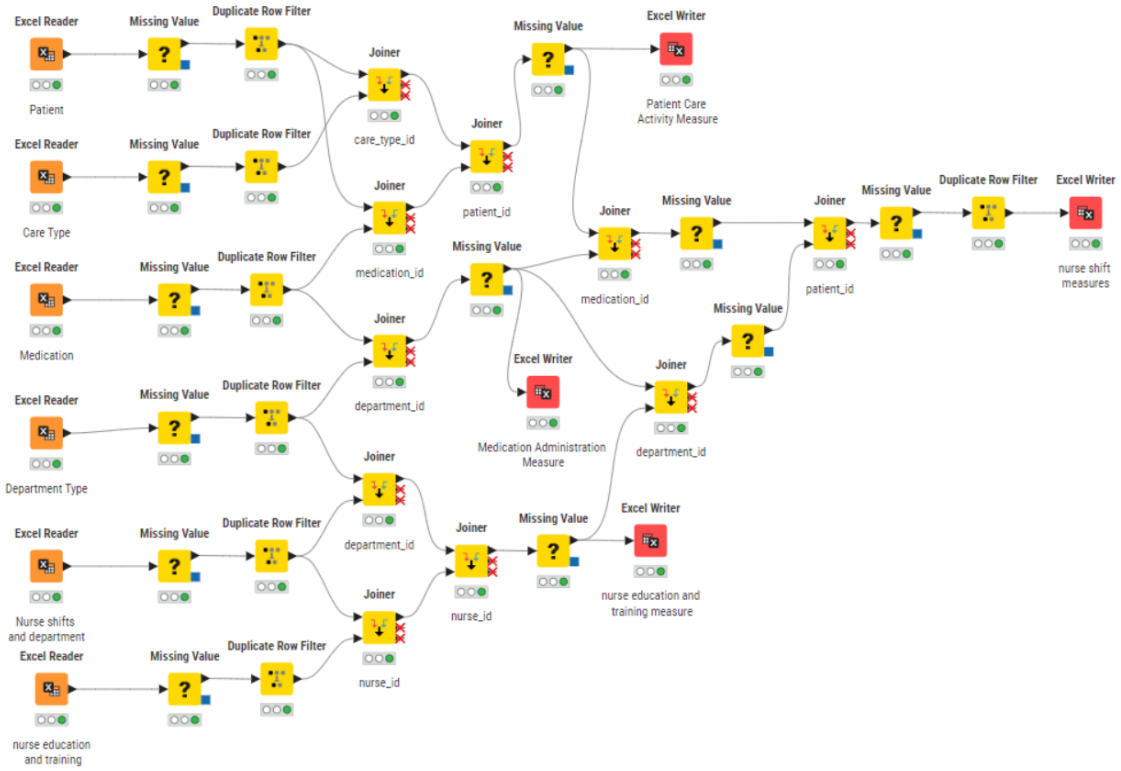
- **Number of Samples:** 36
- **Attributes**:
  - nurse_id
  - department_id
  - Joining_date
  - Shift
  - Training
  - Years_Experience
  - Education
  - Specialization
  - Certifications

9.10. Nurse Shift Measure

- **Number of Samples:** 905
- **Attributes:**
  - Patient_ID
  - Age
  - Condition
  - Room_No
  - Admission_Date
  - Admission_time
  - Discharge_Date
  - Diacharge_time
  - Care_Type_id
  - Care_Type
  - Medication_ID
  - Medication_Name
  - Department
  - Department_Name
  - bed_capacity
  - Nurse_ID
  - Shift
  - Years_Experience
  - Specialization

Nurse Shift Measure is the final table on which we will do our work of pattern recognition and data mining.

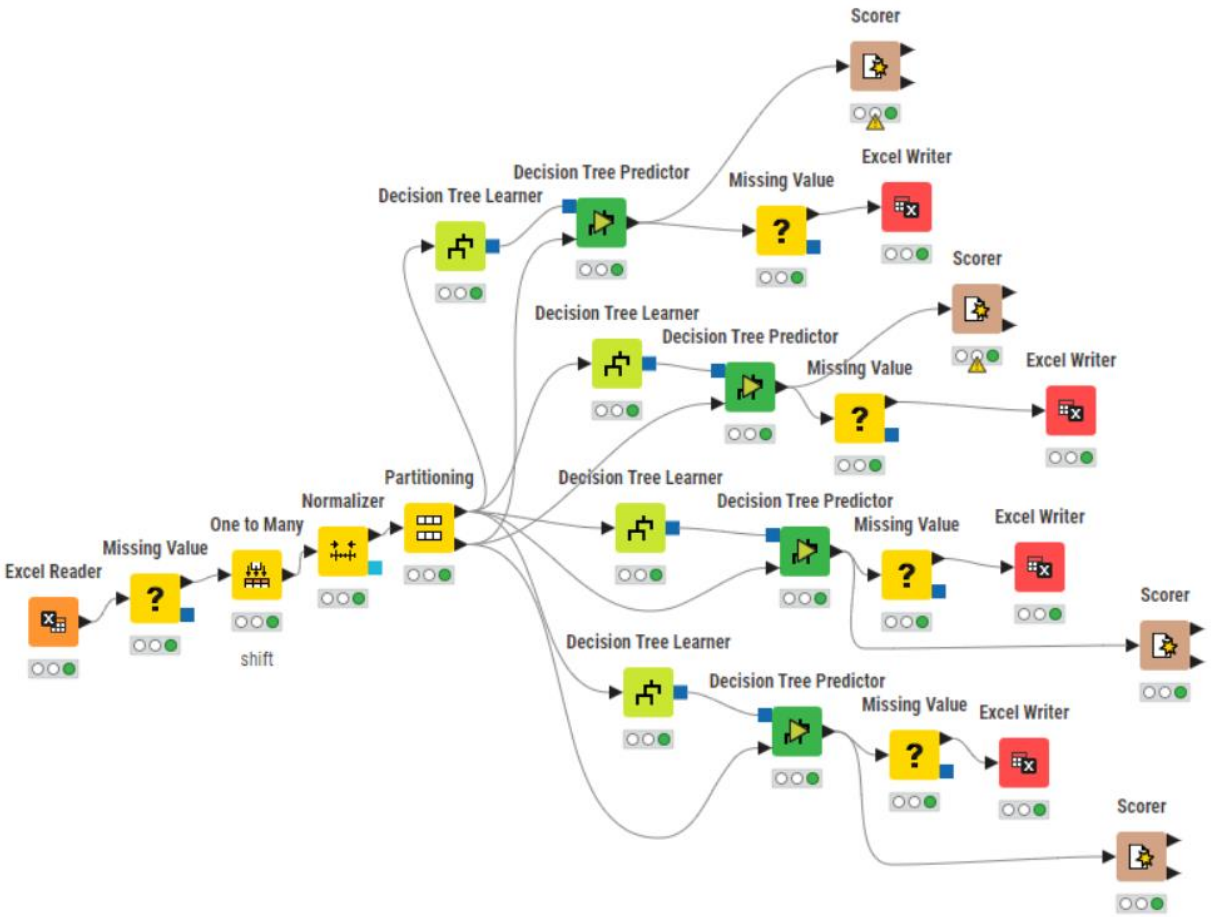
10. Knime Workflow for data cleaning



- **Data Input (Excel Readers):** Multiple Excel Reader nodes import datasets, including patients, care types, medications, department types, nurse shifts, and nurse education.
- **Data Cleaning (Missing Value and Duplicate Row Filters):** Nodes are used to handle missing values and remove duplicate rows in the datasets, ensuring data consistency and quality.
- **Data Integration (Joiners):** Joiner nodes merge datasets based on shared keys (e.g., patient_id, medication_id, department_id, etc.), creating unified datasets for further processing.
- **Derived Metrics (Missing Value Nodes):** These nodes are used to compute or impute missing data points for specific measures, such as patient care activity, medication administration, and nurse shift metrics.
- **Data Output (Excel Writers):** Cleaned and processed datasets are written to Excel files, categorized into measures like:

    - Patient Care Activity Measure
    - Medication Administration Measure
    - Nurse Shift Measures
    - Nurse Education and Training Measure

11. Knime workflow for pattern recognition



Here's a brief breakdown:

Data Input:

- The **Excel Reader** node imports the dataset from an Excel file.
- The **Missing Value** node handles missing data in the dataset.

Data Transformation:

- The **One to Many** node likely transforms categorical variables into one-hot encoded features.
- The **Normalizer** node scales or normalizes numerical features to standardize the data.

Partitioning:

- The **Partitioning** node splits the dataset into training and testing subsets for validation purposes.

Model Training and Prediction:

- Four **Decision Tree Learner** nodes train different decision tree models using subsets of the data.
- Each model's predictions are generated by corresponding **Decision Tree Predictor** nodes.

Post-Prediction Processing:

- **Missing Value** nodes ensure that missing values are handled appropriately in the predicted results.

Performance Evaluation:

- Each model's predictions are evaluated using the **Scorer** nodes, which likely calculate metrics like accuracy, precision, recall, etc.

Output:

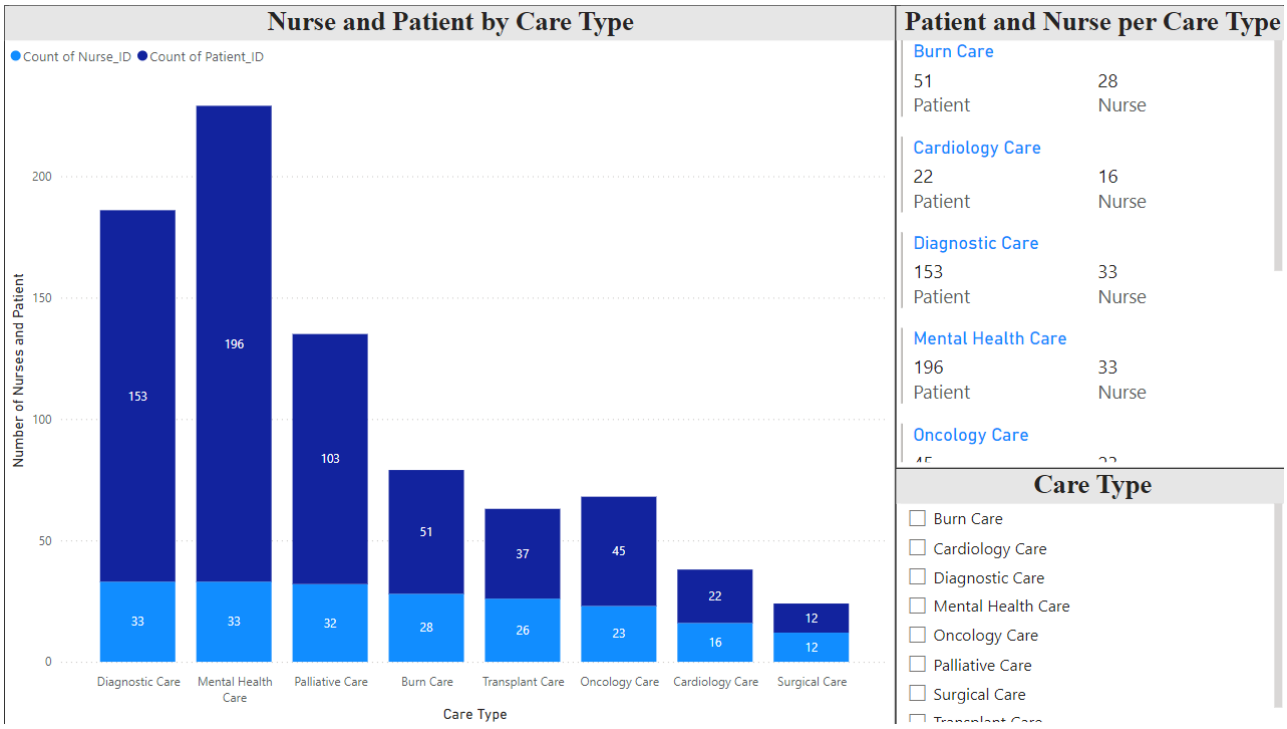
- The **Excel Writer** nodes save the evaluation results or predictions to Excel files for reporting or further analysis.

Data visualization

- I use **Power BI** for data visualization because KNIME lacks robust visualization capabilities.

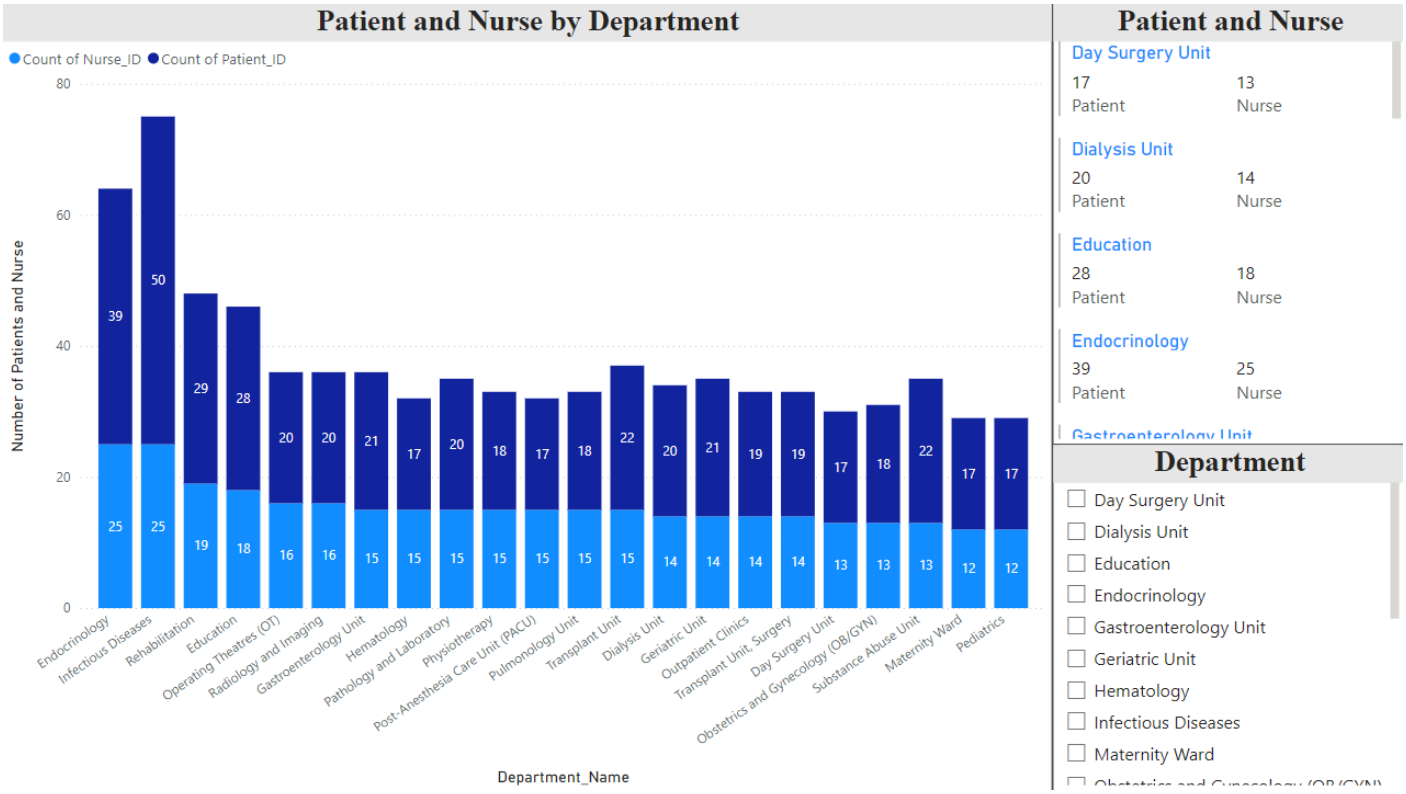
12. Patient and nurse volumes based on different factors

12.1. Patient and nurse by care type:



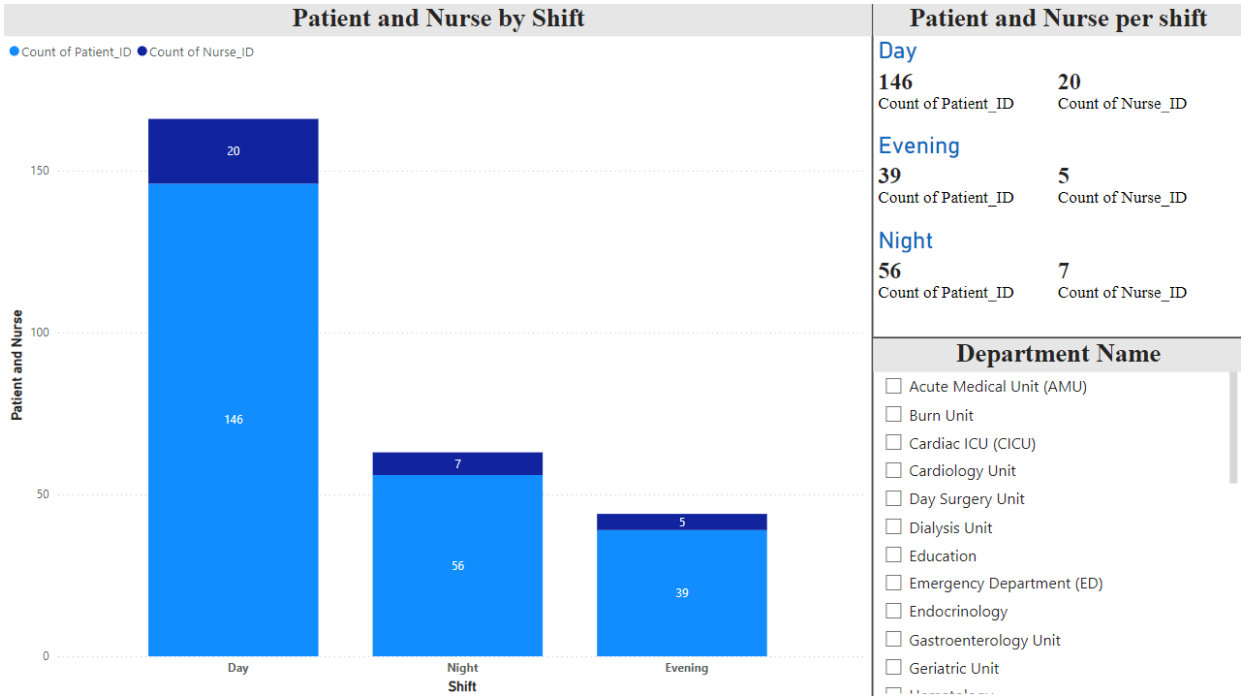
- **High demand areas** (e.g., Mental Health Care, Diagnostic Care) have a significant shortage of nurses' relative to patients, which may affect care quality.
- **Specialized care types** like Transplant and Surgical Care have better nurse-to-patient ratios but fewer resources overall.
- Mental Health and Diagnostic Care might need the most urgent staffing increases given their high patient loads.
- The interactive components (like the care type selector) allow users to focus on specific categories for deeper analysis.

12.2. Patient and Nurse by department



- Departments like **Infectious Diseases** and **Endocrinology** handle higher patient loads, suggesting these departments may need additional staffing or resource allocation.
- Smaller departments such as the **Day Surgery Unit** and **Dialysis Unit** show fewer patients and nurses, reflecting specialized or lower-capacity services.
- The relatively balanced nurse-to-patient ratios in some departments indicate efficient staffing levels, while others may need review to ensure optimal care.

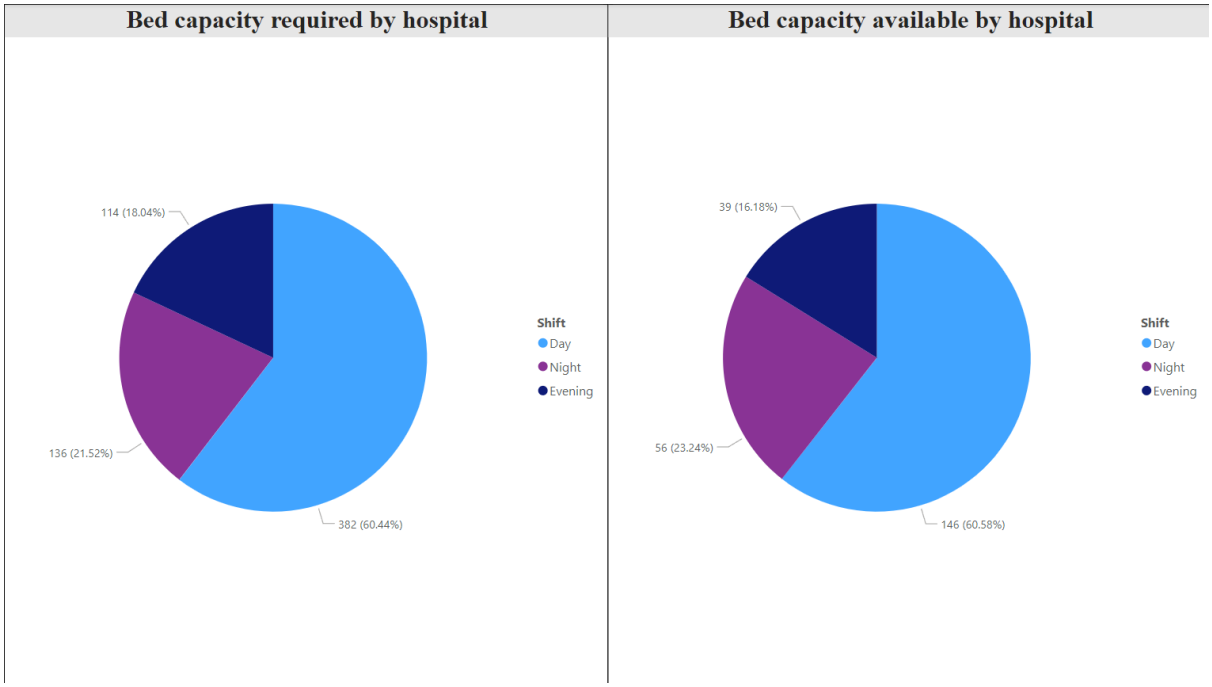
12.3. Patient per shift



- **Day Shift:** There are approximately 7.3 patients per nurse (146 patients / 20 nurses). This indicates the highest number of staff availability relative to patient count.
- **Night Shift:** The ratio is 8 patients per nurse (56 patients / 7 nurses), suggesting a moderate workload.
- **Evening Shift:** The ratio is 13 patients per nurse (39 patients / 3 nurses), indicating the highest workload for nurses during this shift.
- The **day shift** has the highest patient count (146), which is more than double the night shift (56) and nearly four times the evening shift (39). This aligns with typical hospital operations where the day is the busiest.

13. Recourse availability by hospital

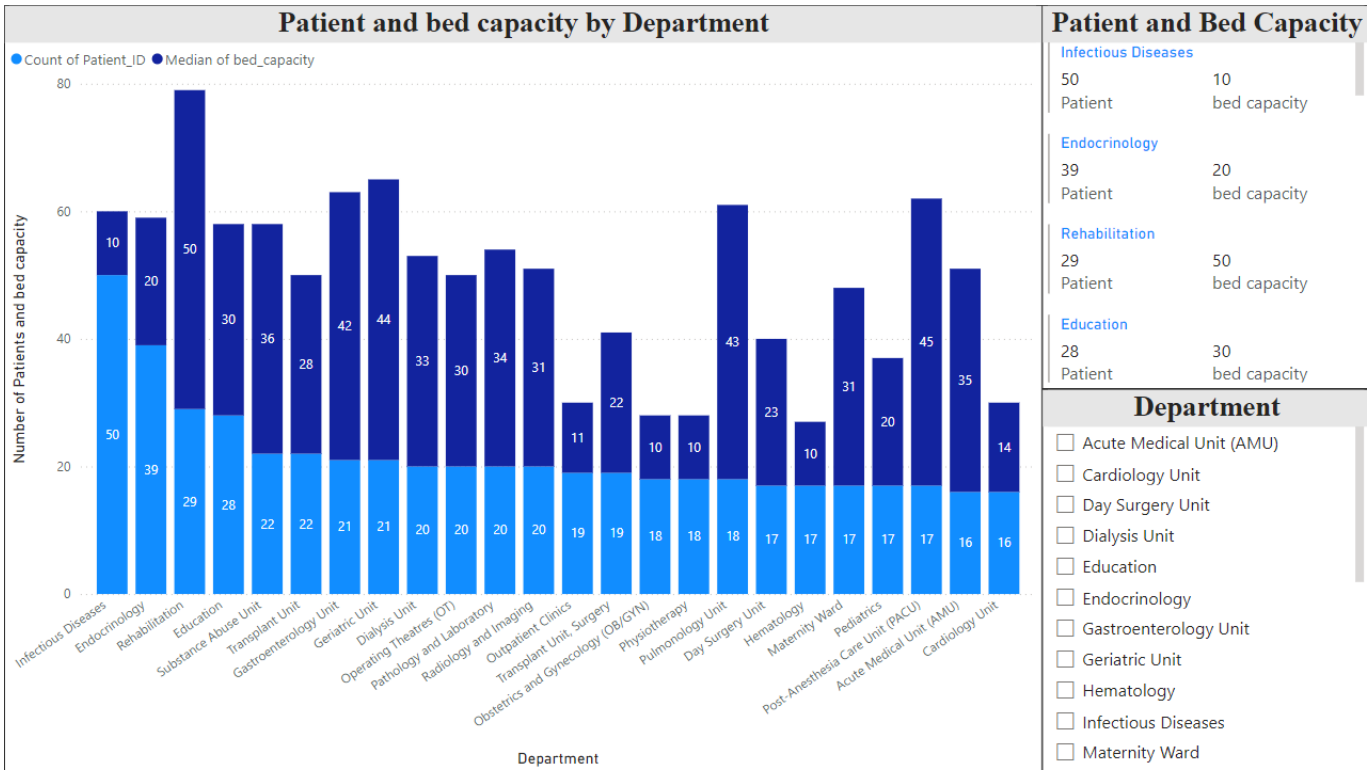
13.1. Bed capacity per shift



The pie chart illustrates the distribution of bed capacity utilization across three shifts: Day, Evening, and Night. Here's the breakdown:

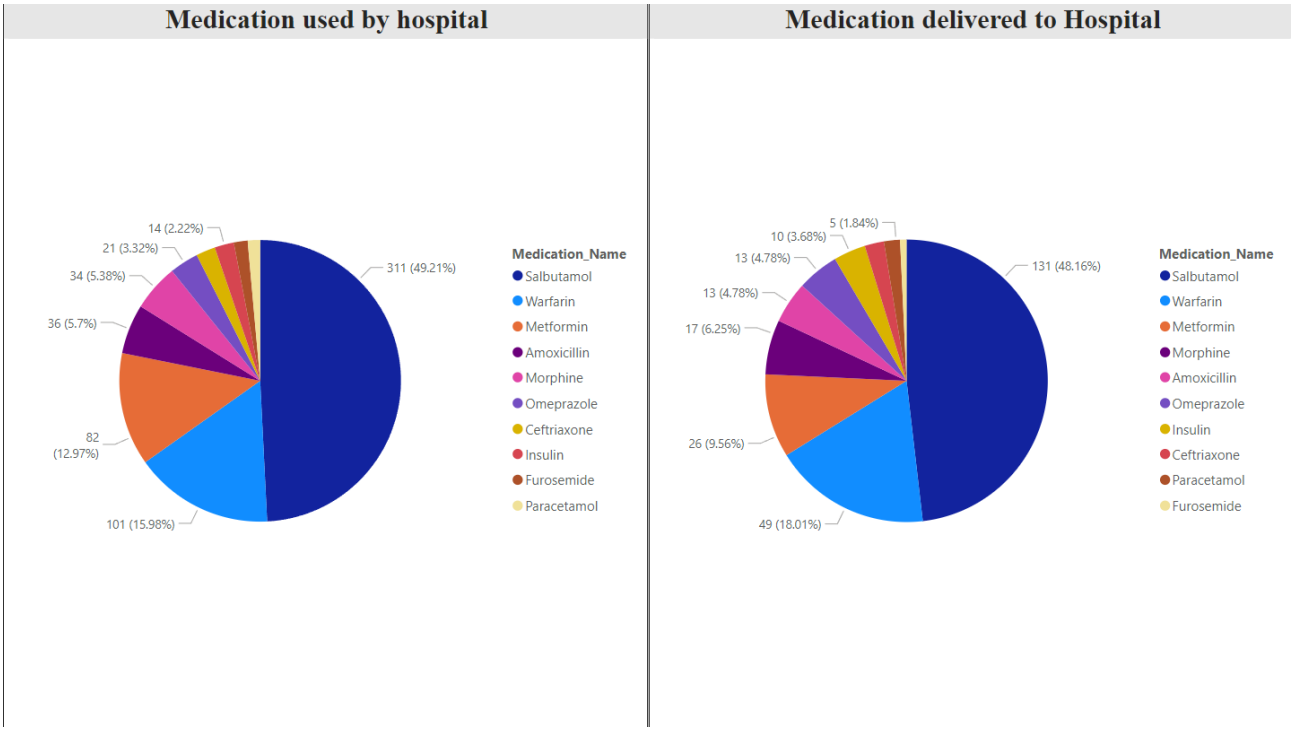
- **Day Shift:** The largest gap, with a **236-bed shortfall.** This requires urgent intervention, as daytime typically accommodates outpatient services, surgeries, and admissions.
- **Night Shift:** A **80-bed shortfall** indicates insufficient capacity for emergencies and overnight stays.
- **Evening Shift:** A **75-bed shortfall** presents a moderate but manageable challenge compared to other shifts.

13.2. Bed capacity per department



- **Over-utilized Departments**:
  - *Rehabilitation*: 50 patients vs. 50 beds (full capacity).
  - *Cardiology Unit*: 45 patients vs. 35 beds (overcapacity).
- **Underutilized Departments**:
  - *Infectious Diseases*: 10 patients vs. 50 beds (significant underutilization).
  - *Education* and *Geriatric Unit*: Low patient counts (28 and 20) compared to bed capacities (30 and 22).
- **Balanced Departments**:
  - *Endocrinology*: 39 patients vs. 20 beds (approaching capacity).
- **Highest Patient Load**: *Rehabilitation* (50 patients).
- **Lowest Patient Load**: *Infectious Diseases* (10 patients).
- **Resource Optimization**:
  - Overloaded departments need more resources or patient redistribution (*Rehabilitation, Cardiology*).
  - Underutilized departments require reevaluation of resource allocation (*Infectious Diseases*).

13.3. Medication



- **Salbutamol**: The most-used medication (311 used) is also adequately delivered (131 units). While there's a slight shortfall, it aligns with high usage.
- **Furosemide**: Usage (10 units) matches delivery (10 units), showing effective stock management.
- **Metformin**: 82 units used vs. 26 delivered — delivery significantly lags behind demand.
- **Paracetamol**: 34 units used vs. 13 delivered — under-delivery may lead to shortages.
- **Amoxicillin**: 21 units used vs. 13 delivered — slightly under-delivered.
- **Warfarin**: 101 units used vs. 49 delivered — deliveries exceed requirements, indicating potential overstocking.
- **Morphine**: Usage (14 units) remains low, and delivery (5 units) seems sufficient based on current demand.
- The hospital manages certain medications well (*Salbutamol*, *Furosemide*) but faces challenges with under-delivery for high-demand drugs (*Metformin*, *Paracetamol*).
- Over-delivery of some low-demand drugs (*Warfarin*) could lead to overstock and wasted resources.

14. Final Outcome:

14.1. Staff Allocation:

- Urgent hiring or reallocation of nurses for **Mental Health** and **Diagnostic Care**.
- Address evening shift understaffing to reduce nurse workload and improve care quality.

14.2. Resource Optimization:

- Reevaluate bed allocations for overburdened departments like **Rehabilitation** and redistribute resources from underutilized departments like **Infectious Diseases**.

14.3. Medication Management:

- Improve procurement and delivery for high-demand medications to avoid shortages.
- Adjust orders for overstocked items to prevent waste.

14.4. Long-Term Planning:

- Enhance scheduling systems to dynamically allocate staff and resources based on real-time data.
- Utilize predictive analytics to anticipate peak demand periods and plan resource availability accordingly.