**Samsung Innovation Campus**

**Healthcare Patient Data Transformation System**

Project Report 2024–25



|  |  |
| --- | --- |
| **TOPIC** | **PAGE NO.** |
| Title of the project | 1 |
| Project Description &  Team member details | 3 |
| Detailed Explanation | 4 |
| UML Diagram | 5 |
| Explanation of Code | 6-7 |
| ScreenShot of Output | 8-9 |
| Graphs | 10 -13 |
| Closure & Bibilography | 14 |

# Project Description:

This project focuses on transforming raw healthcare patient admission records into structured, clean, and analyzable datasets. It applies preprocessing, data quality checks, and advanced analysis techniques to uncover insights about patient admissions, diagnoses, and stay durations. Healthcare Patient Admission Data Transformation, focuses on analyzing and transforming hospital admission records to improve data quality and generate meaningful insights.  
  
The system demonstrates:  
- Data cleaning and handling of missing values  
- Conversion of data types for consistency  
- Creation of derived attributes (e.g., Stay Duration)  
- Filtering and grouping of healthcare records  
- Exporting cleaned and transformed data for analytics

# Team Member detail:

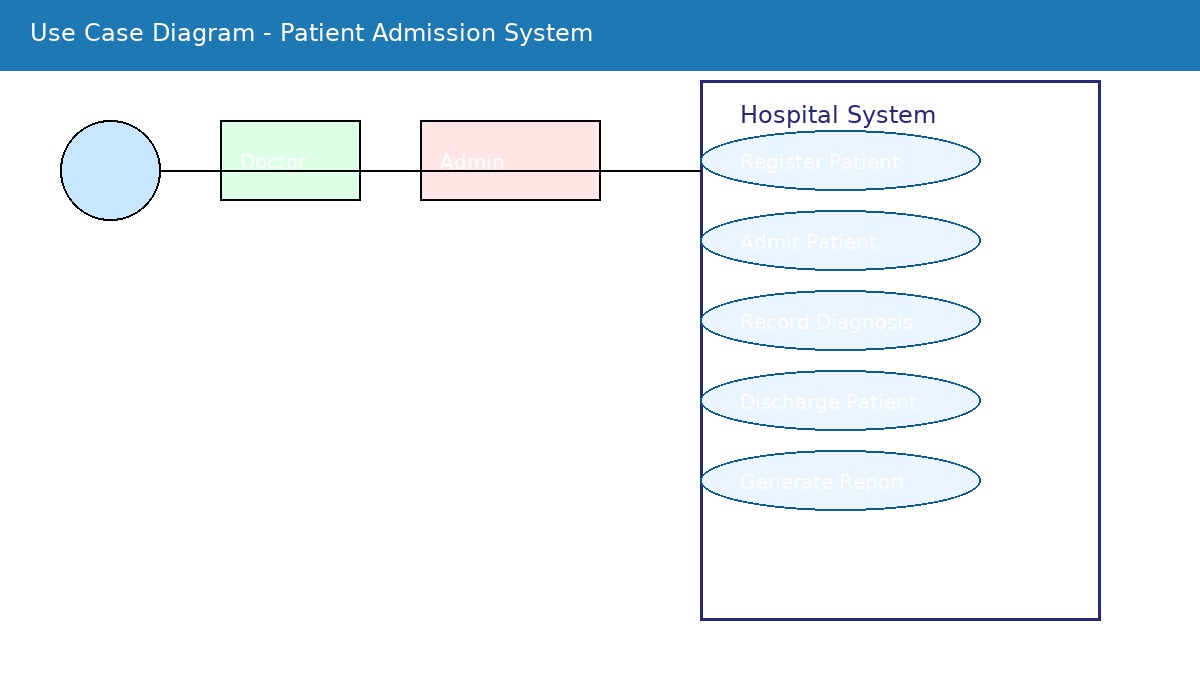
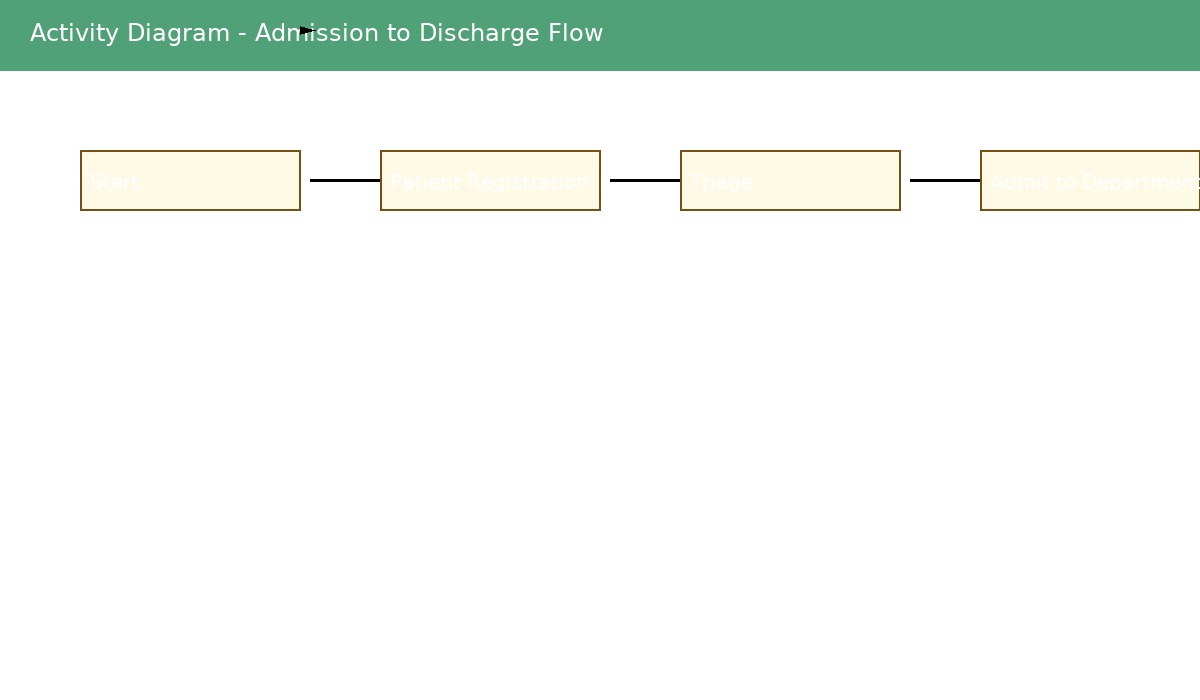
• Name: Sinchana Mahesh  
• USN: 4GW23CI049  
• Branch: CSE(AI & ML)

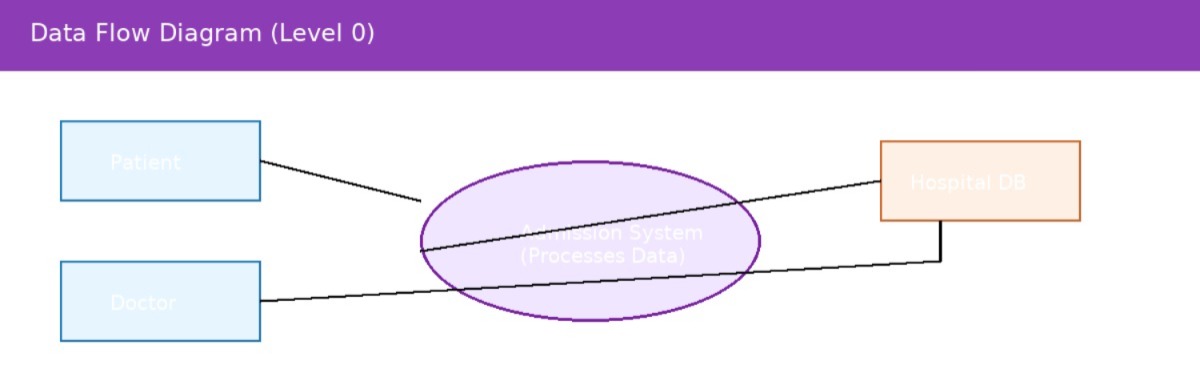
# Detailed Explanation

Overview:  
This project, titled “Healthcare Data Transformation and Analysis,” focuses on managing and analyzing patient admission records using Python and Pandas. The dataset contains essential healthcare details such as patient demographics, department, diagnosis, and admission/discharge dates. Data quality issues including missing values and duplicates are identified and resolved by filling missing diagnoses with “Unknown” and assigning today’s date for missing discharge records. To ensure consistency, admission and discharge dates are converted into proper datetime format. A new derived field, Stay Duration, is calculated to measure the length of each patient’s hospital stay. Finally, the transformed dataset is exported for further healthcare analysis, helping in better hospital management and informed decision-making.

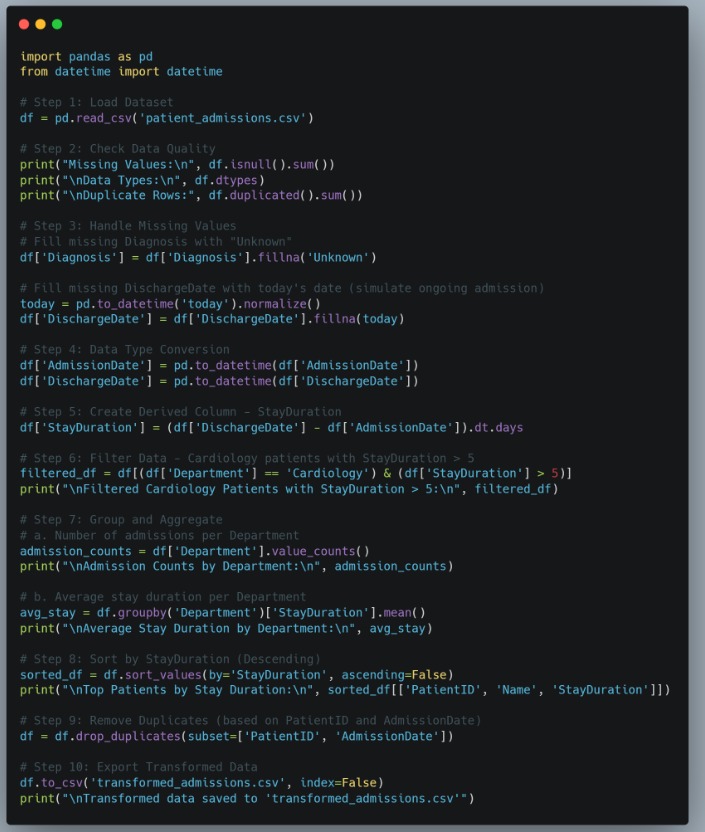
Use Cases:  
1. Data Quality Check – Identify and handle missing values and duplicates.  
2. Data Transformation – Convert Admission/Discharge dates to datetime and compute StayDuration.  
3. Filtering – Extract Cardiology patients with stay > 5 days.  
4. Aggregation – Compute admissions per department and average stay duration.  
5. Export – Save cleaned data into a new CSV file.

# UML Diagrams

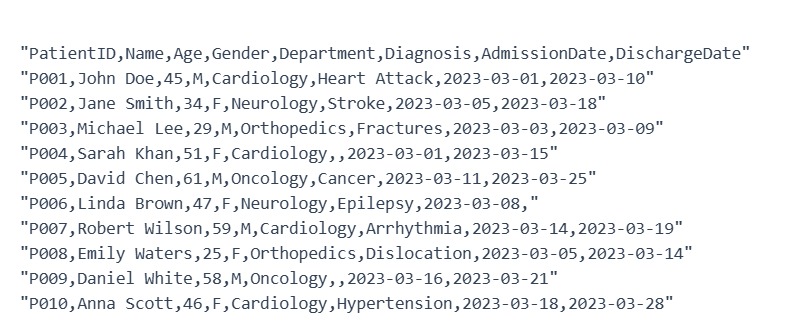
The following diagrams represent the system workflow and processes:  




# Code

Below is the Python code used for data transformation:  


Sample patient admission data **(patient\_admissions.csv )**

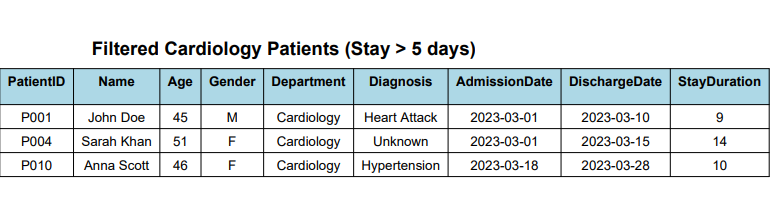


# Explanation of the Code:

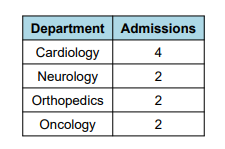
1.Data is loaded using Pandas.  
2. Missing values are filled appropriately (Diagnosis → Unknown, DischargeDate → Today).  
3. Date columns are converted to datetime format.  
4. StayDuration is derived from Admission and Discharge dates.  
5. Cardiology patients with StayDuration > 5 are filtered.

6. Aggregation and sorting operations are applied.  
 7. Duplicates are removed and data is exported

# Screenshots of Output



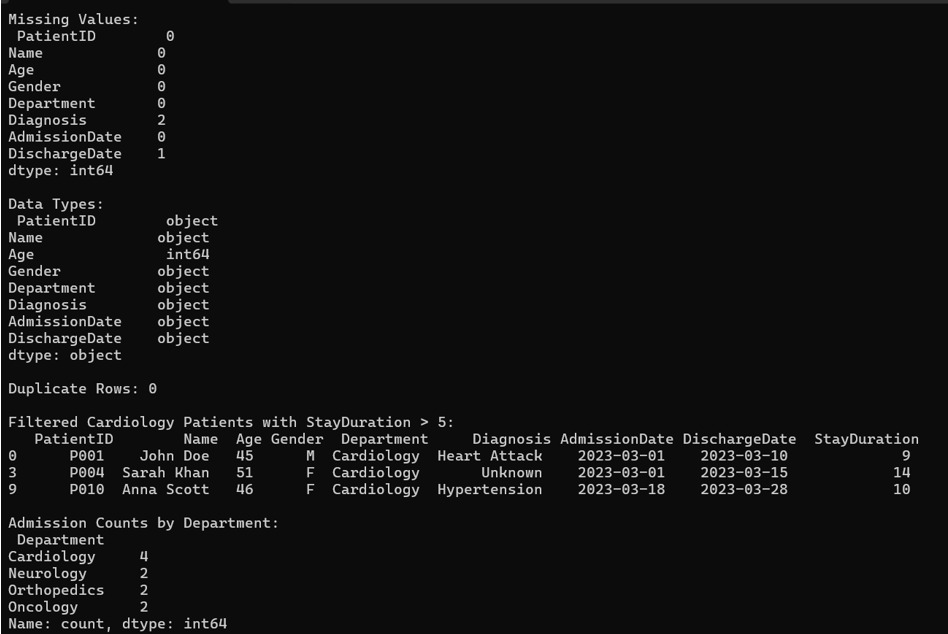
**Admission Counts by Department:**



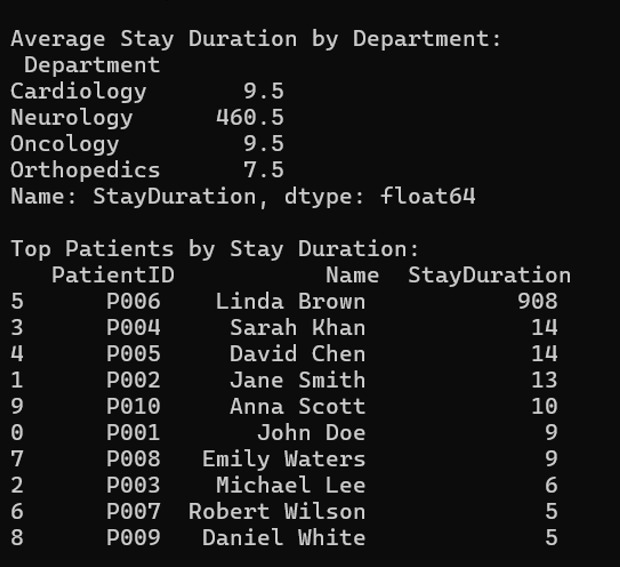
# Graphs:

# 

# 

**Output:** 

**The dataset has minimal missing values (2 Diagnosis entries and 1 DischargeDate) and no duplicate records. Cardiology has the highest admission count (4 patients), with 3 cardiology patients having extended stays over 5 days. The data types show most fields are properly formatted except date columns which need conversion to datetime format**.



The data shows significant variation in average stay durations across departments, with Neurology having an exceptionally long average stay of 460.5 days. One patient (Linda Brown, P006) has an extreme outlier stay of 908 days, which is dramatically influencing the Neurology department's average. The other departments show more typical hospital stay durations ranging from 7.5 to 9.5 days.

# output(app.py) A close-up of a pie chart AI-generated content may be incorrect.

# Closure:

This project on Healthcare Data Transformation successfully demonstrates the preprocessing, cleaning, and basic analysis of patient admission data using Python and Pandas. The project highlights the importance of handling missing values, duplicates, and inconsistent data in healthcare datasets to ensure accurate reporting and decision-making.

* We learned to read and explore CSV datasets effectively.
* Applied data cleaning techniques such as filling missing values and removing duplicates.
* Performed basic data analysis to understand patient demographics and admission trends.
* Developed a Python-based solution that can be extended for advanced analytics like predictive modeling or visualization dashboards.
* The project emphasizes how structured data transformation improves healthcare decision-making, reduces errors, and aids hospital administration in managing patient records efficiently.

# Bibliography:

McKinney, Wes. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O’Reilly Media, 2017.

VanderPlas, Jake. Python Data Science Handbook: Essential Tools for Working with Data. O’Reilly Media, 2016.

Official Pandas Documentation: https://pandas.pydata.org/docs/

Kaggle Dataset – Patient Admissions: https://www.kaggle.com/datasets

Healthcare Data Standards and Best Practices – World Health Organization (WHO) https://www.who.int/health-topics