**Methodology**

**1. Data Collection and Description**

The dataset used in this project was provided in an Excel file, which contains multiple sheets with information related to patient hospitalization events.

The dataset includes various attributes relevant to analyzing patient re-hospitalization, ER visit patterns, and other demographic factors.

**2. Data Cleaning and Preparation**

Data cleaning was performed to handle missing values, outliers, and inconsistencies. For the general data table, specific tasks included:

* **Handling Missing Values:** Missing entries were identified and handled using suitable methods such as filling with mean/median values for continuous variables and mode for categorical variables.
* **Date-Time Conversion:** Ensured that all date-related columns were correctly formatted as datetime objects for proper computation and time series analysis.
* **Feature Engineering:** New columns, such as the number of days between hospitalizations and categorized time periods, were created to aid in further analysis.

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

EDA of the data was conducted to understand the distributions and relationships of various parameters. The steps included:

* **Histogram Plots:** Created histograms for continuous variables, such as the time between admissions and ER visit duration, to visualize their distributions.
* **Category Analysis:** Plots were generated for categorical variables like the admitting department, ER arrival method, and type of ER visit to understand their frequency distributions.
* **Unique Patient Analysis:** The number of unique patients was calculated to determine the dataset's breadth.

**4. Relationship Analysis Between Hospital Factors and Re-Hospitalization**

The relationship between re-hospitalization and various hospital factors was analyzed to understand which factors affect re-hospitalization and which of them has the strongest impact.

* **Variable Selection:** Focused on factors that we thought can lead to re-hospitalizations and that are related to the hospital such as: department, doctor, number of patients in department, number of patients per doctor, etc…
* **Visualization:** Generated histograms to visually inspect the relationship between these factors and the likelihood of re-hospitalization.
* **Statistical Analysis:** Conducted various statistical analysis and test to determine if there were significant associations between those factors and re-hospitalization rates.

**5.** **Relationship Analysis Between Demographic Factors and Re-Hospitalization, Clustering**

The relationship between re-hospitalization and various demographic factors was then analyzed. And then clustering re-hospitalization.

* **Variable Selection:** Focused on factors that are related to the patient’s demographic such as: age, gender, BMI, weight, height, education, number of children, etc…
* **Visualization:** Generated histograms to visually inspect the relationship between these factors and the likelihood of re-hospitalization.
* **Statistical Analysis:** Conducted various statistical analysis and test to determine if there were significant associations between those factors and re-hospitalization rates.
* **Clustering:** clustering re-hospitalization by various factors.

**6. Time Series Analysis of Departments**

Time series analysis was conducted for various departments to understand patterns over time: admissions and releases in first and second hospitalizations.

* **Grouping and Aggregation:** Data was grouped by department and resampled monthly to observe trends in admissions and discharges.
* **Visualization:** Plots were created to visualize trends over time.
* **Identify Trends:** Understand the trend in each department in the first and second hospitalizations.

**7. Dimension Reduction**

For better understanding of the data and reduce noise in the data dimension reduction was preformed for tables.

**8. Tools and Libraries Used**

* **Python Libraries:** The analysis and implementation were performed using Python, with key libraries including:
  + Pandas for data manipulation and cleaning.
  + Plotly for dynamic visualizations.
  + Matplotlib for static visualizations.
  + Statsmodels for time series analysis and seasonal decomposition.
  + SciPy for statistical tests such as Chi-Square.
* **Software Environment:** All coding and analysis were done in a python environment, allowing for interactive data exploration and visualization.
* **Architecture:** All code was managed using Git and GitHub.

**Conclusion**

The methodology described above combines data cleaning, exploratory data analysis, statistical testing, and time series analysis to provide a comprehensive understanding of factors influencing patient re-hospitalization and patterns in hospital admissions. This approach enables insights into potential areas for improving patient care and reducing re-hospitalization.