Visual Analysis of Hospital Operations

Project Overview

This hospital visit analysis project investigates over 63,000 anonymized patient records collected over a ten-year span, from January 2011 through December 2020. The primary objective is to uncover hidden patterns, identify inefficiencies, and generate actionable insights that healthcare providers can use to improve operational performance, resource planning, and patient outcomes. Using Python and widely adopted libraries like Pandas, Matplotlib, and Scikit-learn, the study focuses on profiling patients, understanding department-wise load, predicting hospital stay durations, and determining which factors contribute to costs and length of care. The results aim to serve both administrators and care teams with data-driven intelligence.

Data Preprocessing & Exploration

Before conducting the analysis, extensive preprocessing was required to clean and refine the data. This included removing or imputing missing values, encoding categorical variables into numerical formats, filtering out outliers, and generating derived metrics such as "Length of Stay." Descriptive statistics and visualizations were used to explore patient demographics and admission types. Most patients fell in the 31–50 age bracket, with private hospitals managing over 60% of all visits. Emergency room cases and high severity levels were linked to increased lengths of stay. Departments like Gynecology and Pediatrics received more visits compared to others, while Radiology saw fewer patients but longer visit durations, hinting at more complex or time-consuming cases.

Predictive Modeling

To support hospital planning and resource allocation, a Decision Tree Regressor was used to build a model for predicting patient stay duration. The model yielded an R^2 score of 0.65, suggesting moderate predictive power. The most impactful predictors included severity of illness, the department of admission, and the type of hospital admission (emergency or scheduled). These insights help with early triage, bed allocation, and scheduling staff appropriately. Although this is a foundational model, it provides a baseline for further improvement using advanced algorithms like Random Forest or Gradient Boosted Trees.

Key Visual Insights

- **Average Fee by Diagnosis:** A bar chart showcased the ten most expensive diagnoses, with each averaging close to $45,000. Understanding these cost drivers enables healthcare administrators to re-evaluate procedure costs and optimize billing strategies.

- **Monthly Revenue Trend:** The time-series plot illustrates revenue patterns over the 10-year period. It reveals cyclical trends and seasonal spikes, assisting hospitals in identifying high-demand periods. This knowledge supports financial planning, inventory management, and appropriate staff scheduling.
- **Most Active Doctors:** Doctors like Marie Curie and Jonas Salk had the highest patient loads, signaling potential overwork or superior efficiency. These statistics can be leveraged for performance reviews, rewards, and scheduling to distribute patient load more evenly.
- **Most Frequent Patients:** Repeat visits were tracked to uncover chronic patients or those needing long-term care. Patients such as Louis Parrish and Jenny Gold had unusually high visit counts, indicating the need for personalized care plans or targeted outreach programs.
- **Total Revenue by Department:** This chart broke down revenue generation by department. General Surgery and Cardiology emerged as top earners, whereas areas like Nutrition and Neonatology brought in less. Such information is invaluable for making informed budget decisions, expanding profitable departments, or investigating underperformance.

Limitations & Future Scope

While the analysis produced several key takeaways, it also came with a few constraints. The dataset lacked specific information such as time of admission and discharge, medication details, post-visit outcomes, and socioeconomic data. Incorporating this missing data in the future could vastly enhance model performance and interpretability. Additionally, applying more advanced models such as XGBoost or neural networks could improve prediction accuracy. Another valuable extension would be to integrate patient satisfaction data and real-time analytics dashboards for hospital management teams.

Conclusion

This project highlights how data science can be applied to healthcare systems to generate impactful insights. It enables hospitals to better anticipate patient needs, allocate resources more efficiently, and enhance the overall quality of care. By identifying patterns in cost, duration, and volume, this analysis supports evidence-based decision-making in hospital management. The predictive tools and visualizations developed here can be integrated into hospital information systems for day-to-day operational support.

All code, visualizations, and extended documentation are hosted in the GitHub repository:
https://github.com/Santhakumarramesh/Hospital_Visit_Analysis