Onyx Healthcare Data Analysis

Abstract

This project analyzes a rich healthcare dataset using Python and statistical methods to uncover trends in patient demographics, treatment costs, test outcomes, and hospital performance. We apply exploratory data analysis, hypothesis testing, and logistic regression to derive actionable insights. The results show that billing is largely unaffected by condition or insurance provider, while gender and length of stay do influence patient outcomes. This work highlights the power of data-driven healthcare evaluation.

Introduction

This project focuses on analyzing patient-level healthcare data to identify patterns in demographics, costs, and outcomes. With rising healthcare costs and an increasing focus on evidence-based care, such analysis can support hospital administrators and policymakers in making informed decisions. The project was meaningful as it offered hands-on experience in cleaning real-world datasets, applying statistical tests, and building interpretable models to extract insights. It also strengthened my skills in Python, data visualization, and healthcare data interpretation.

Project Details

Tools Used:

- Python (pandas, seaborn, matplotlib, statsmodels, patsy)
- Jupyter Notebook

Dataset Fields:

- Demographics: Age, Gender, Blood Type
- Hospital data: Name, Location, Admission/Discharge Dates
- Billing & Insurance: Billing Amount, Insurance Provider
- Clinical: Medical Condition, Test Results, Medication, Admission Type

Key Analysis Performed:

1. **Demographic Distribution**: Bar charts and counts of age groups, gender, and blood types.
2. **Cost Analysis**:
    - By condition, hospital, and insurance provider.
    - Mean/median comparisons showed negligible differences.
3. **Length of Stay**:
    - Calculated and compared across admission types.
4. **Statistical Testing**:
    - ANOVA: No cost difference across admission types.
    - Chi-square: Significant link between gender and test results.

Conclusions

- Billing amounts are flat across medical conditions, hospitals, and insurance providers.
- Gender significantly affects patient test outcomes, supported by chi-square and logistic regression.
- Admission type does not influence billing, but affects length of stay slightly.
- Medications are used consistently per condition, with some hospital-level variation.
- No regional health or billing pattern emerged from geographic data.

These findings reinforce the idea that patient demographics and treatment processes are the key levers in healthcare cost and outcome variation.

References

- Python Data Analysis Libraries: pandas, seaborn, statsmodels
- Statsmodels Documentation: https://www.statsmodels.org/