



HEALTHCARE DATASET (2019-2024)

HIBA T M




Introduction

- This synthetic healthcare dataset is crafted to be a valuable tool for data science, machine learning, and data analysis enthusiasts. It aims to replicate real-world healthcare data, providing a platform for users to practice, develop, and showcase their data manipulation and analysis skills within the healthcare sector.
- The creation of this dataset stems from the need for practical and diverse healthcare data for educational and research purposes. Due to the sensitive nature and privacy regulations surrounding real healthcare data, access for learning and experimentation can be limited. To bridge this gap, I have utilized Python's Faker library to generate data that closely mirrors the structure and attributes of actual healthcare records. By offering this synthetic dataset, I aim to foster innovation, learning, and knowledge sharing in the field of healthcare analytics.



Data Overview

- The data used in this project has been downloaded from kaggle (<https://www.kaggle.com/code/hiba2004/healthcare-dataset-analysis>)
- This dataset includes 10,000 synthetic records, each representing a healthcare record for a fictional patient. It covers a range of attributes, including patient demographics, medical conditions, admission details, and billing information. The dataset is crafted for educational and non-commercial purposes, providing a valuable resource for practicing data analysis and machine learning techniques. It is completely synthetic, ensuring no real patient data is included or compromised. This makes it ideal for developing skills in healthcare data analysis without privacy concerns.



Methodology

- The classification algorithms that has been used in this project:
 - *Logistic Regression
 - *Decision Trees
 - *Random Forest
 - *Naive Bayes
 - *k-Nearest Neighbors (k-NN)



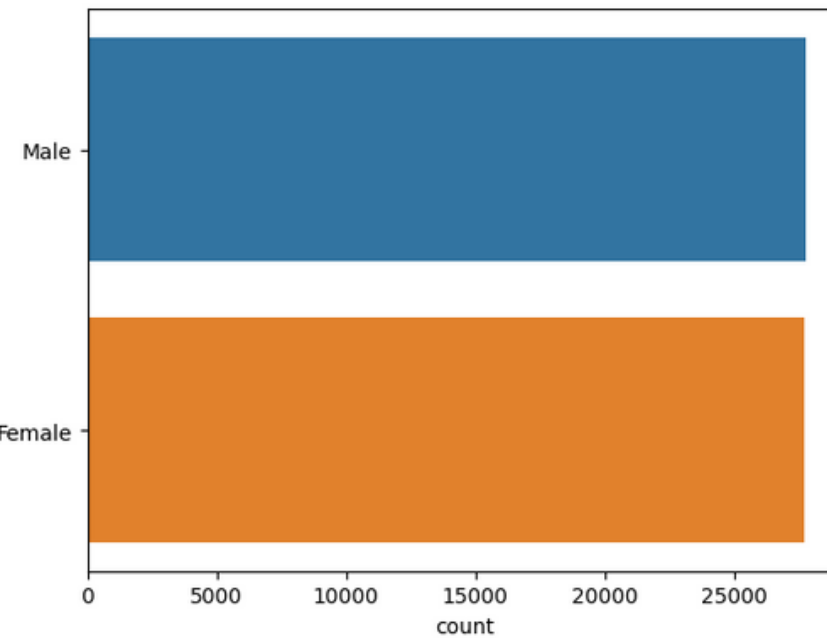
Implementation

TOOLS

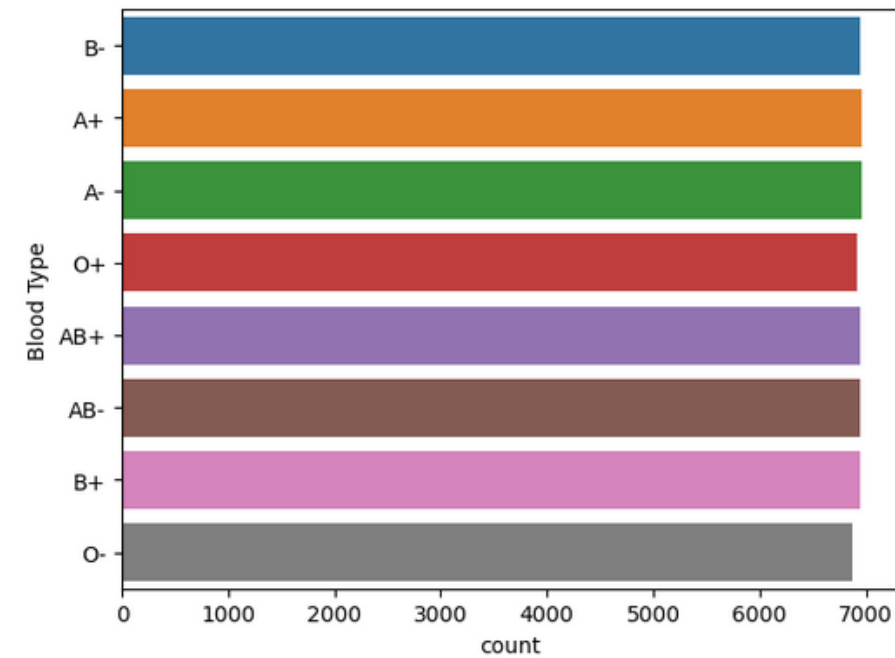
- *python and jupyter notebook
- *Electronic Health Records (EHR) Systems

Results

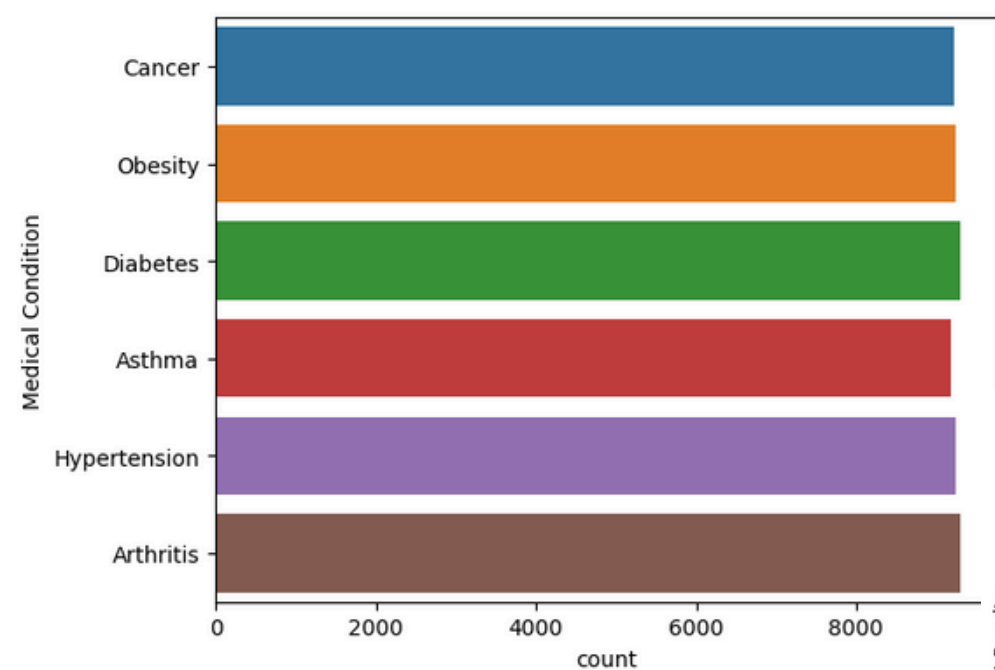
Gender



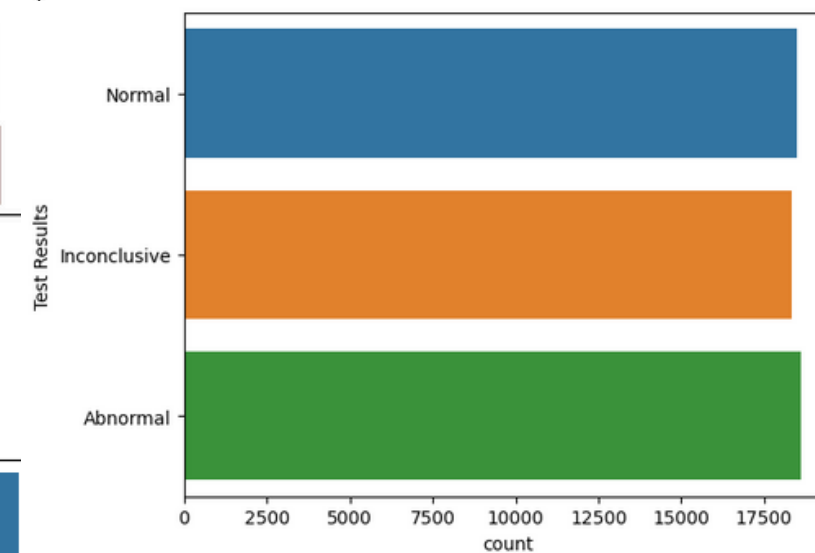
Blood Type



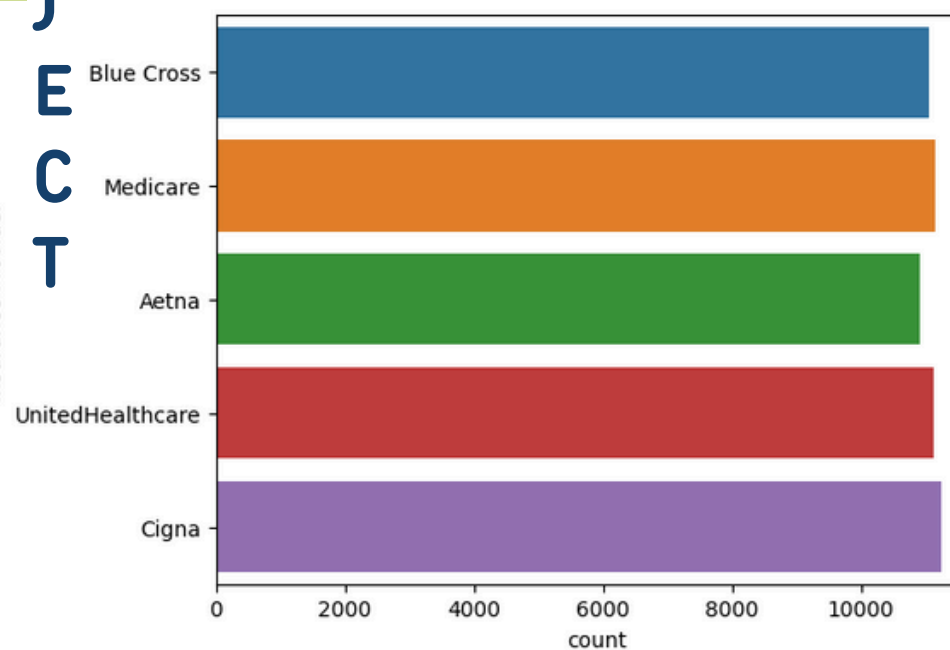
Medical Condition



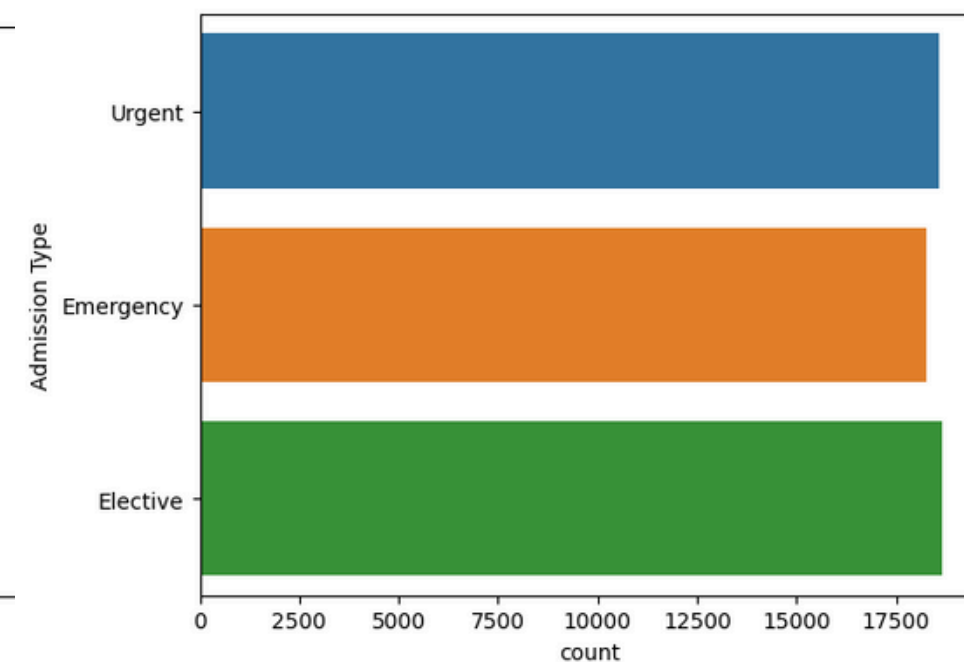
Test Results



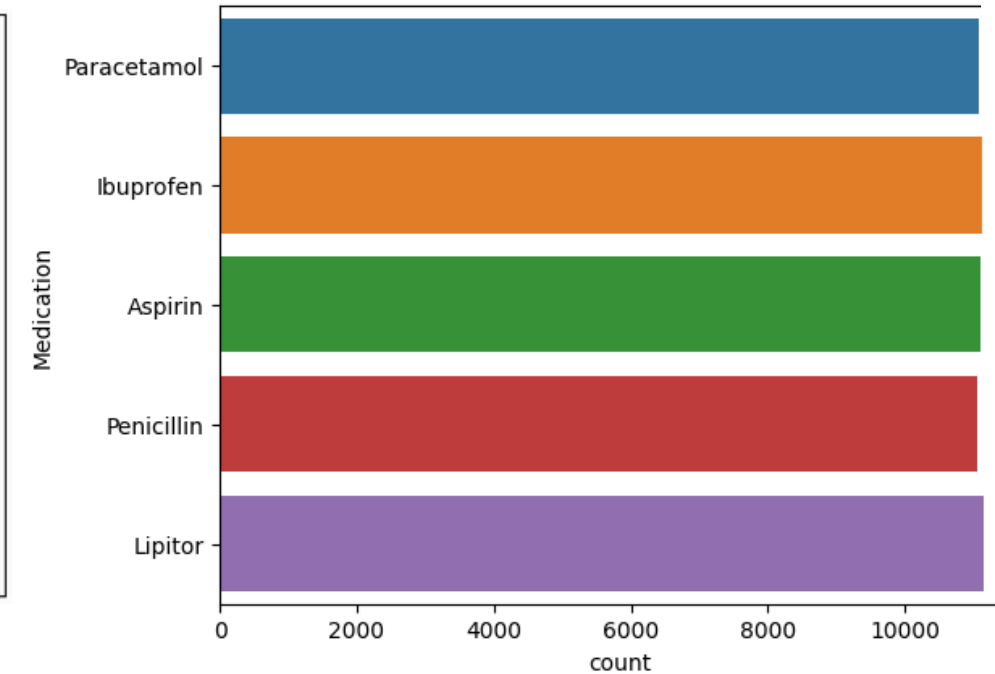
Insurance Provider



Admission Type

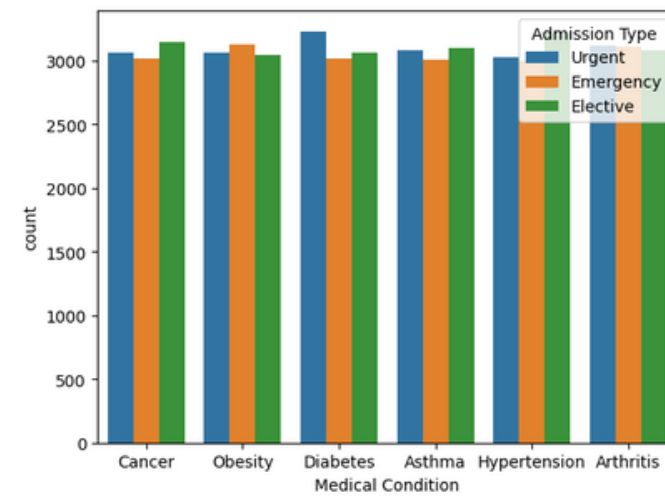


Medication

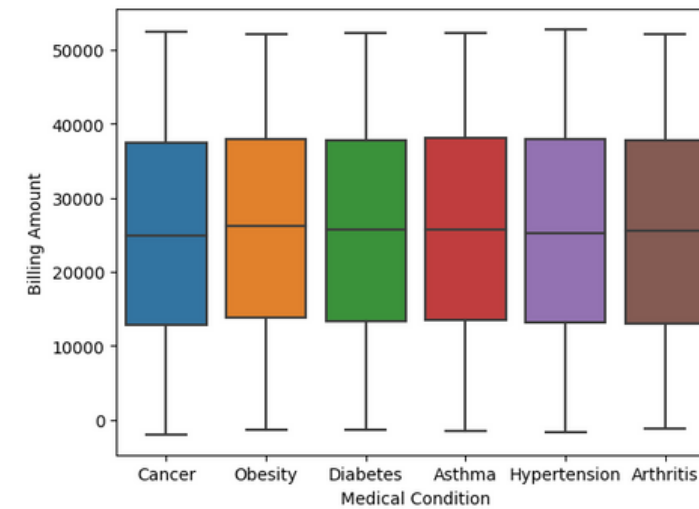


Results

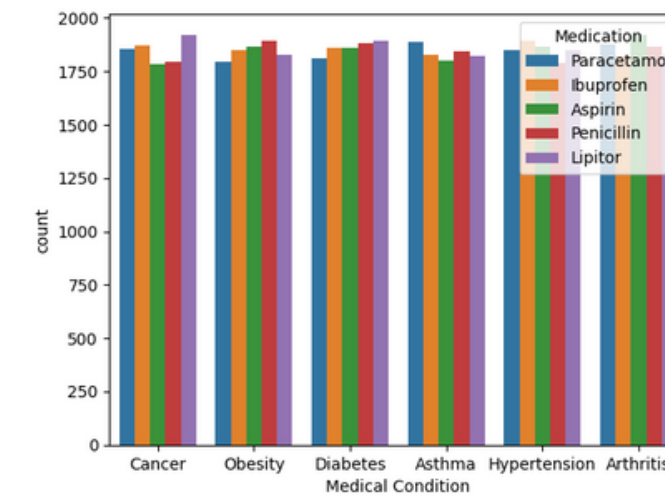
Medical condition and admission type



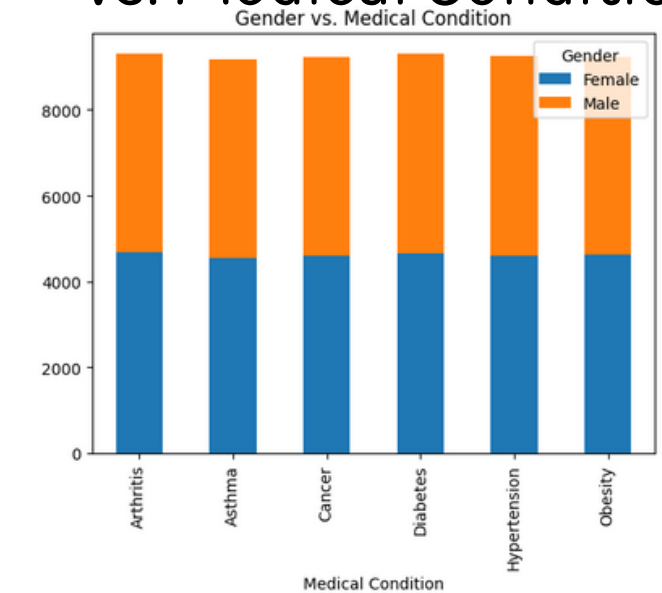
Billing amount range for each medical condition



Medications count for each medical condition

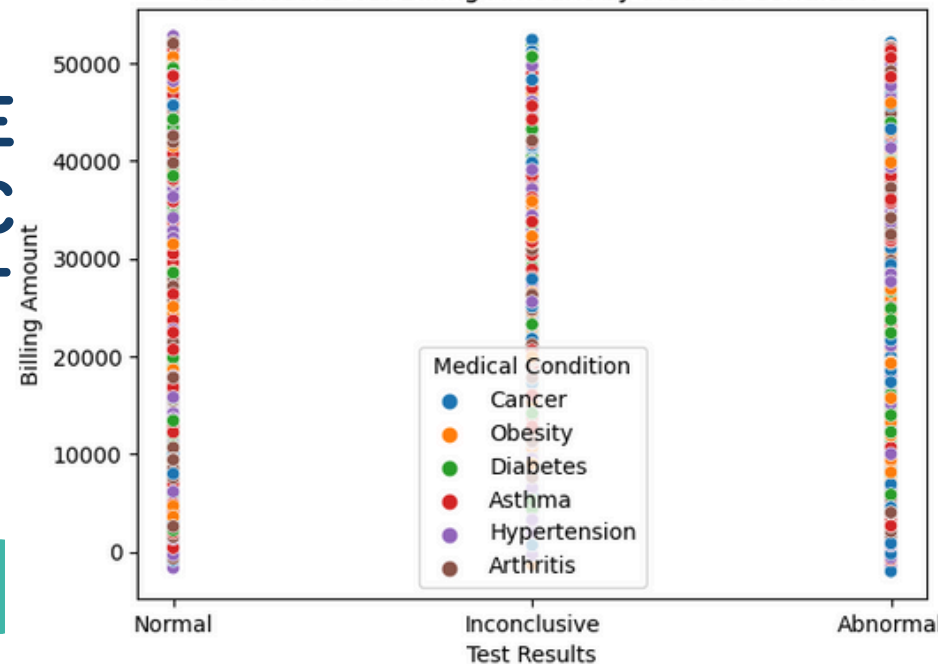


Stacked bar plot for Gender vs. Medical Condition

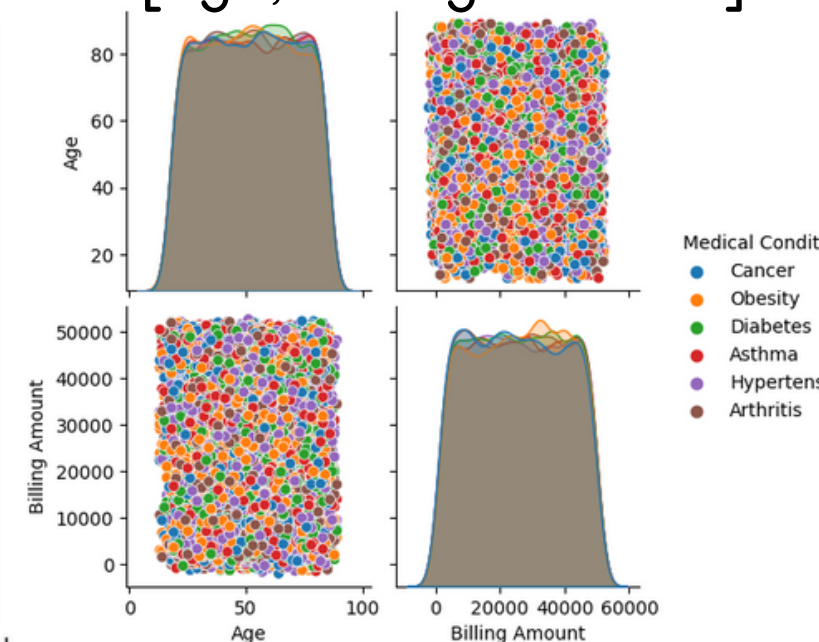


Scatterplot for test results with medical conditions and bills

Test Results vs. Billing Amount by Medical Condition

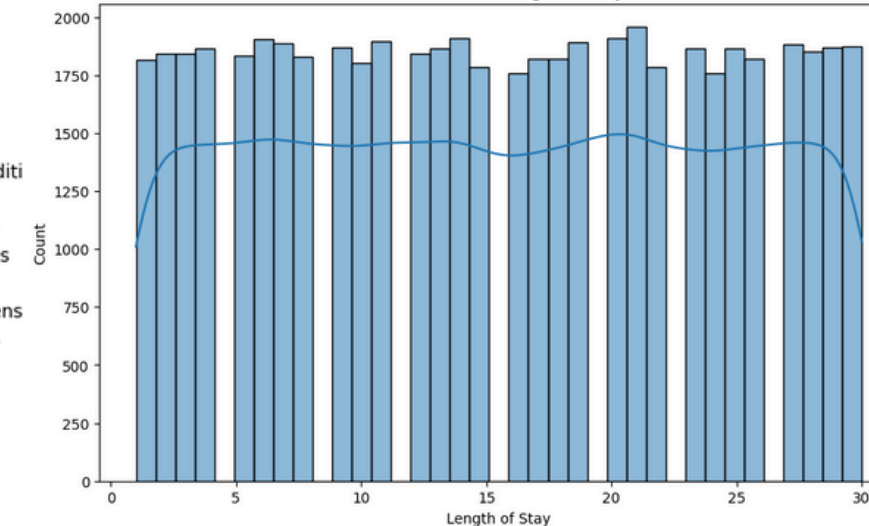


Pairplot between Medical condition and [age, billing amount]



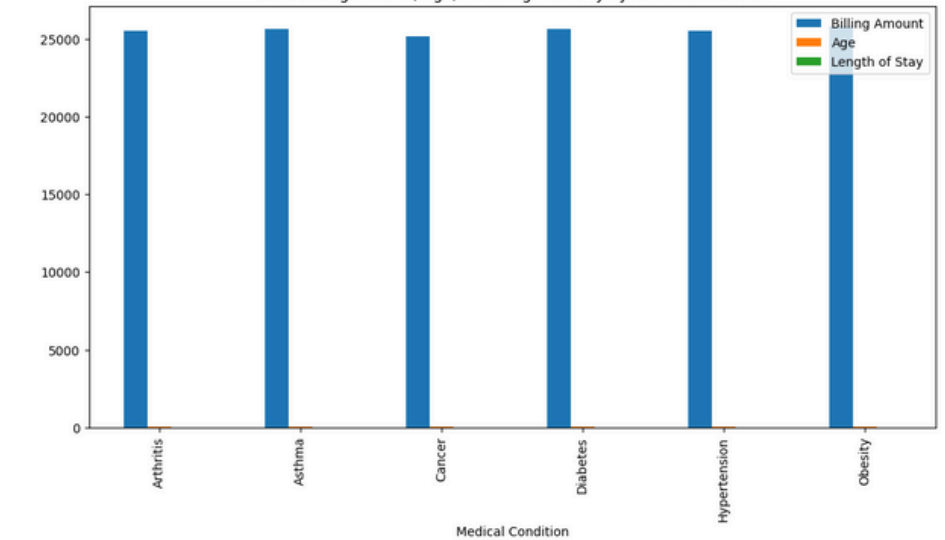
How long does patient stays

Distribution of Length of Stay



Billing Amount, Age, Length stay

Mean Billing Amount, Age, and Length of Stay by Medical Condition





Conclusion

This dataset can be used for various purposes, including:

- *Developing and testing healthcare predictive models.
- *Practicing data cleaning, transformation, and analysis techniques.
- *Creating data visualizations to understand healthcare trends.
- *Learning and teaching data science and machine learning in a healthcare context



References

- kaggle
- Github
- sklearn, numpy, matplotlib, pandas, seaborn documentation
- youtube