Predicting Yearly Medical Costs Using Synthetic Patient Data

Stony Brook University Data Science Bootcamp Capstone Project

Diana Kulawiec

Introduction

- Can a machine learning model be developed to predict yearly medical encounter costs from synthetic patient data?
- Which factors have the most important impact on healthcare expenses?

The Process

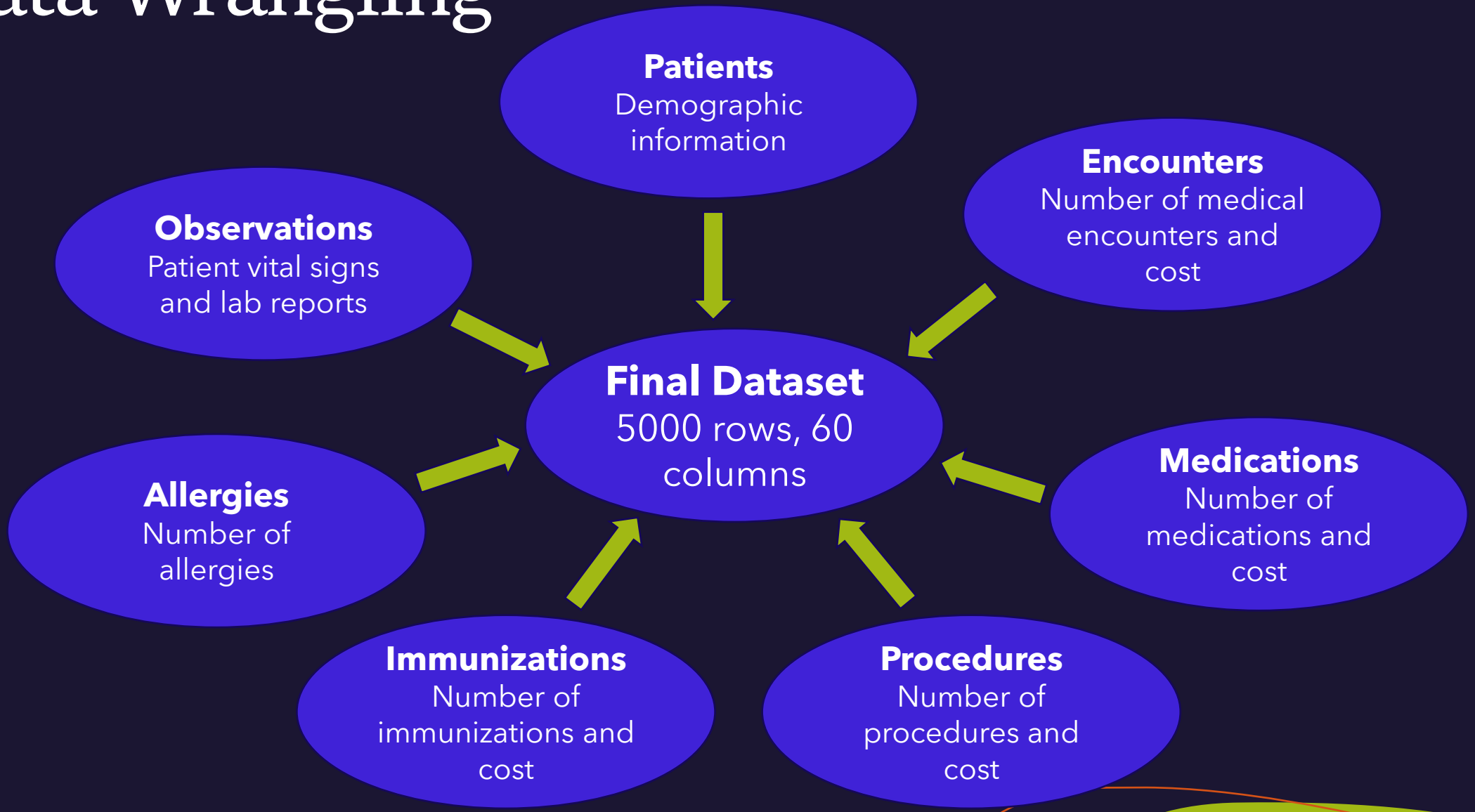

Data Generation
Data Wrangling
Exploratory Data Analysis
Data Imputation
Model Training and Development
Final Model Selection

SYNTHEA EMPOWERS
DATA-DRIVEN HEALTH IT

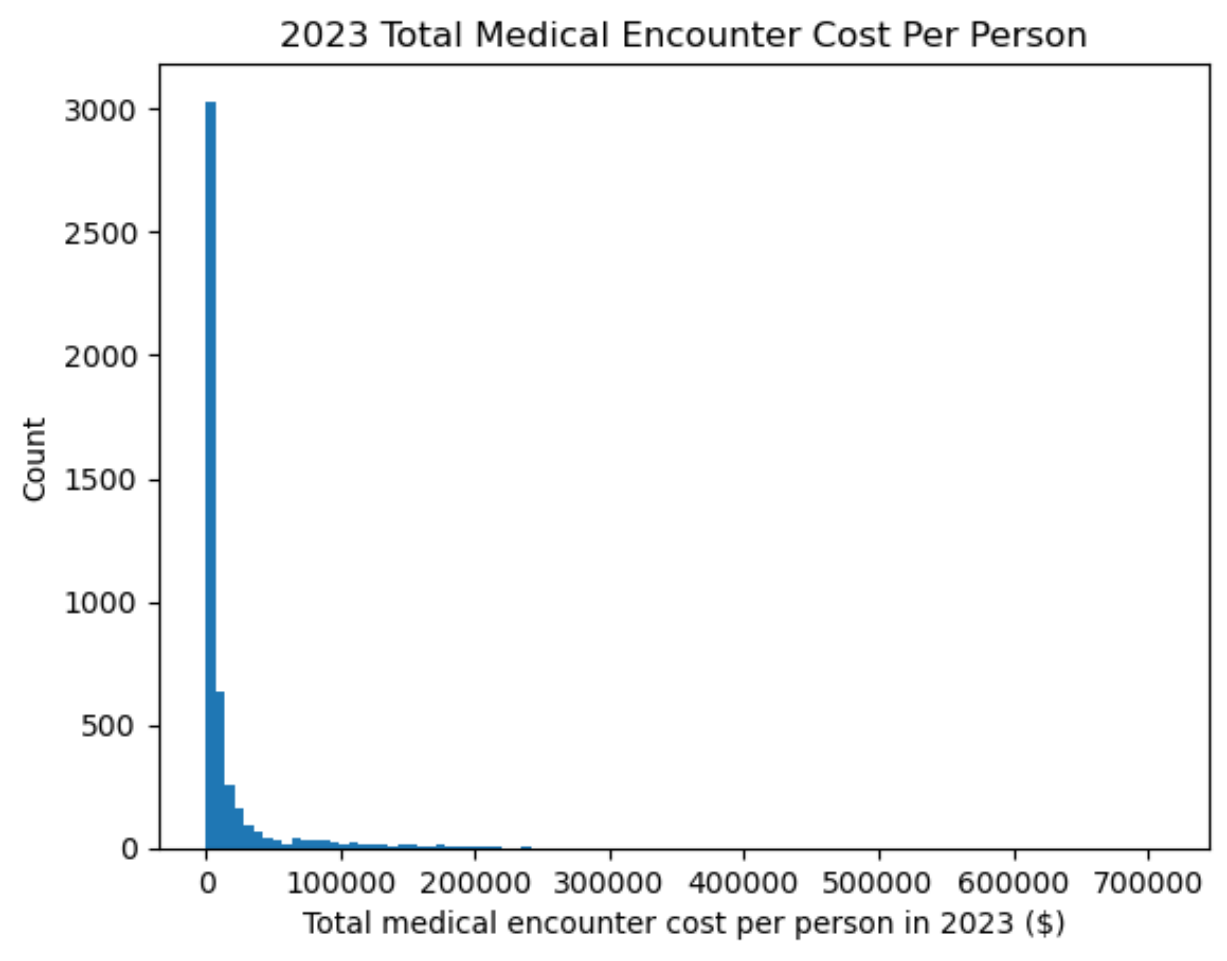
Data Generation

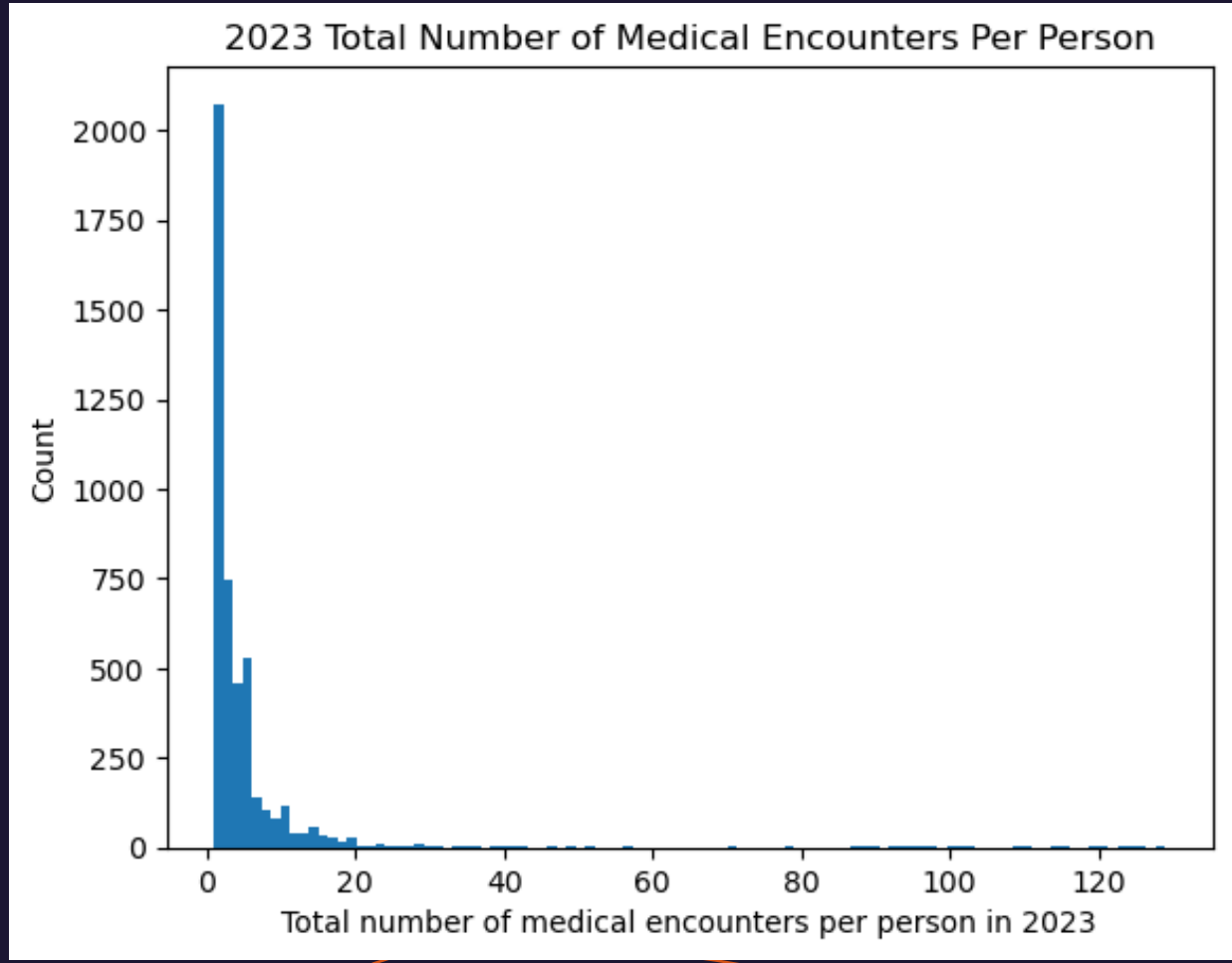
- Downloaded synthetic patient data from Synthea for 100 living patients from each of the 50 states
- CSV files:
  - Patients
  - Encounters
  - Medications
  - Procedures
  - Immunizations
  - Allergies
  - Observations

Data Wrangling

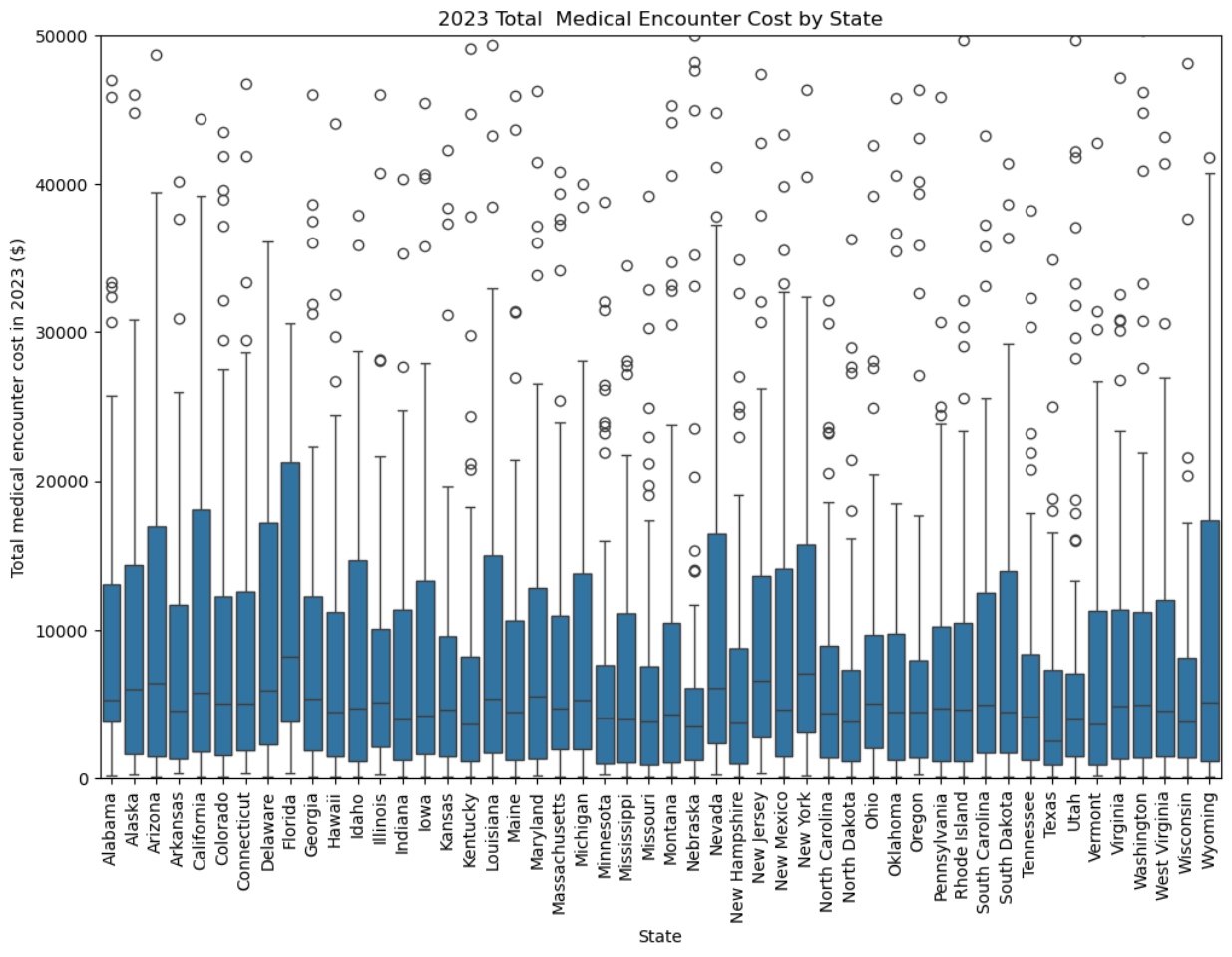


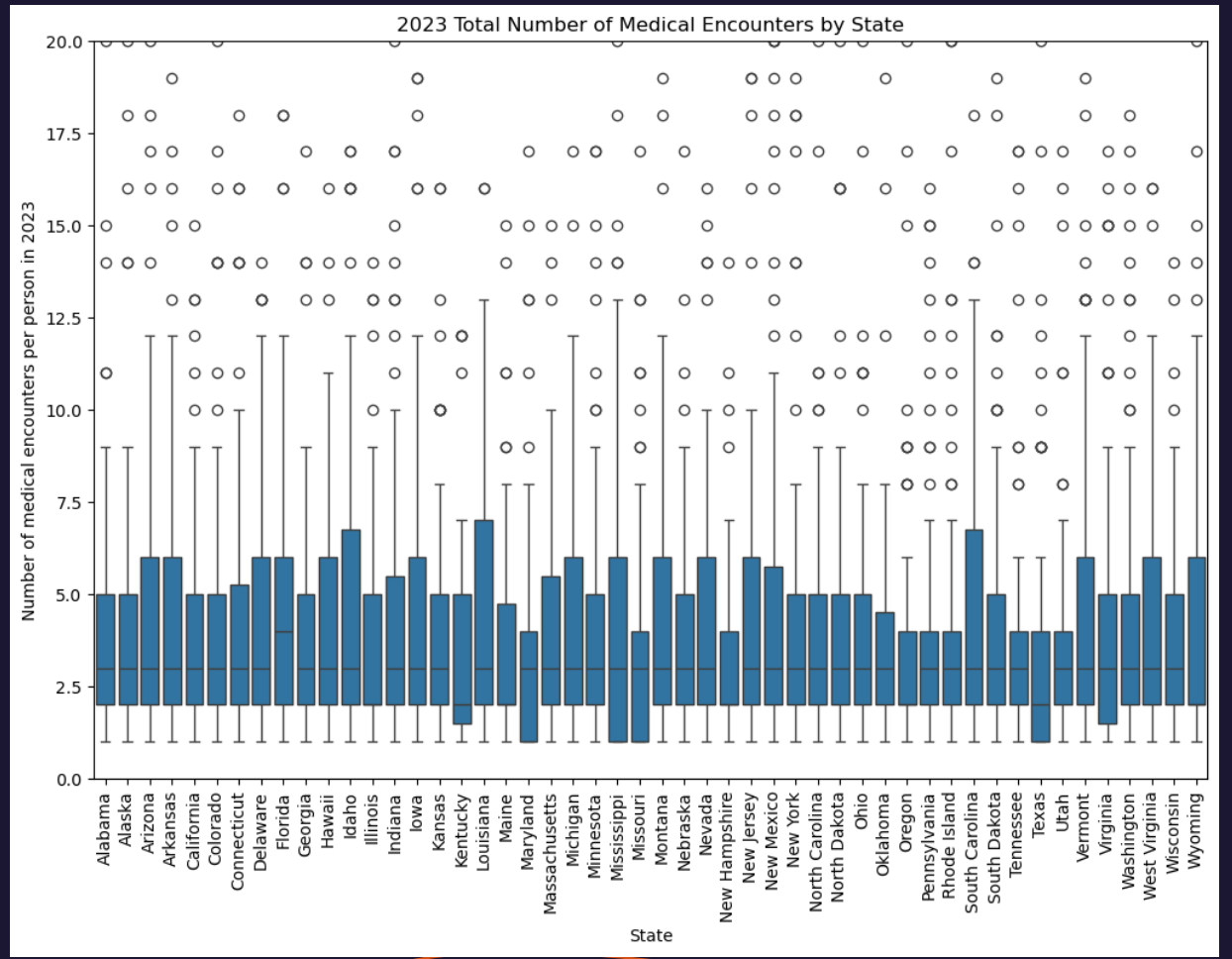
Medical Encounters Distributions



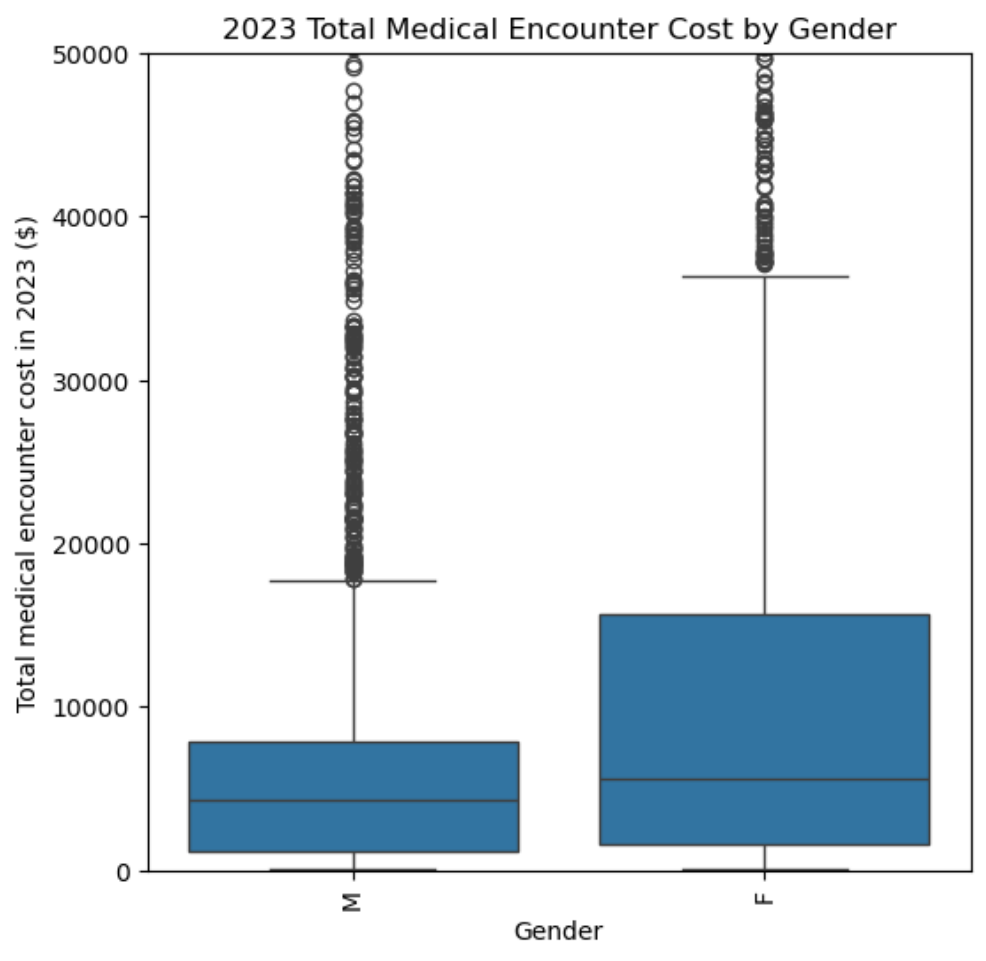


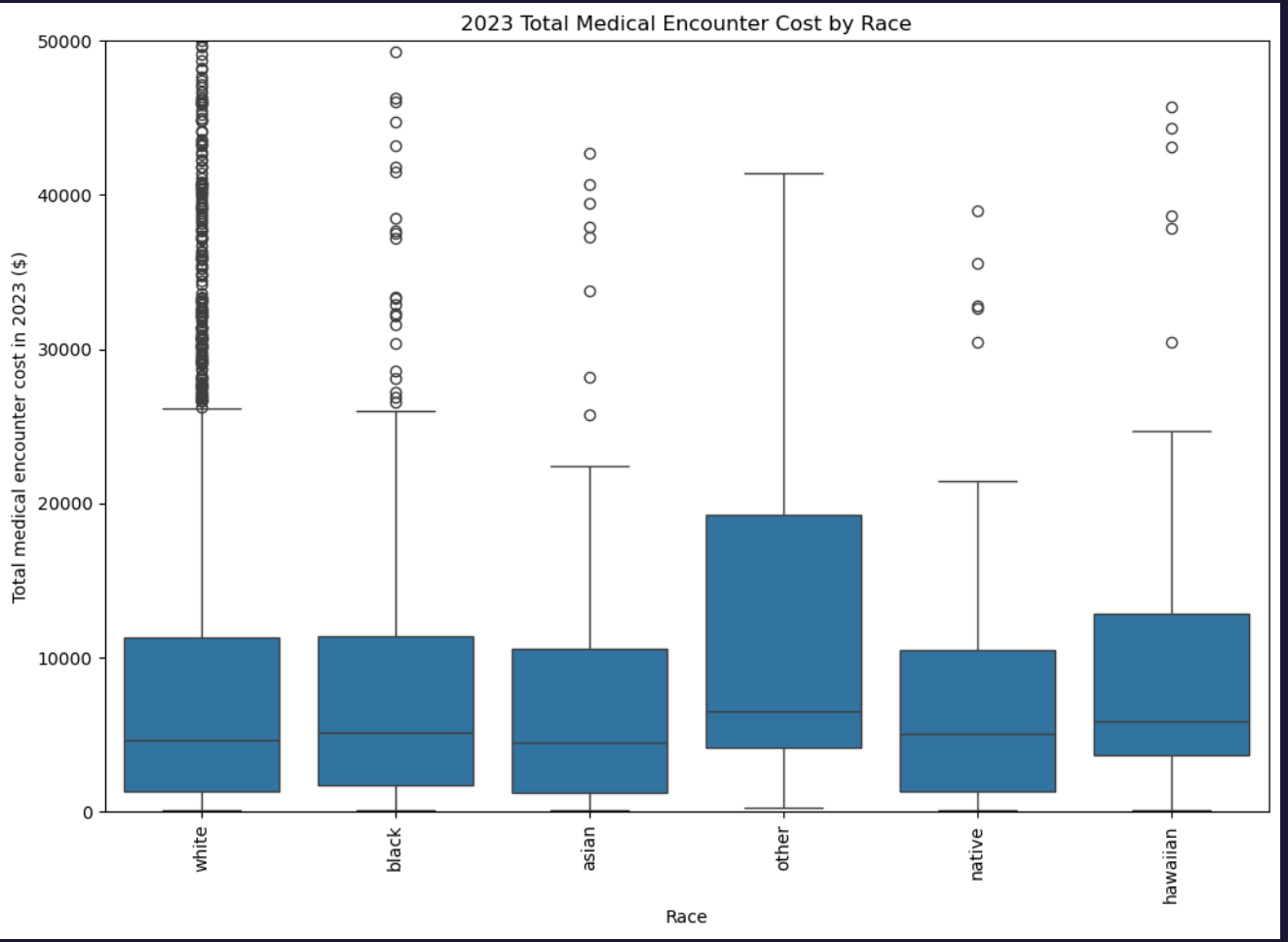
Medical Encounters by State



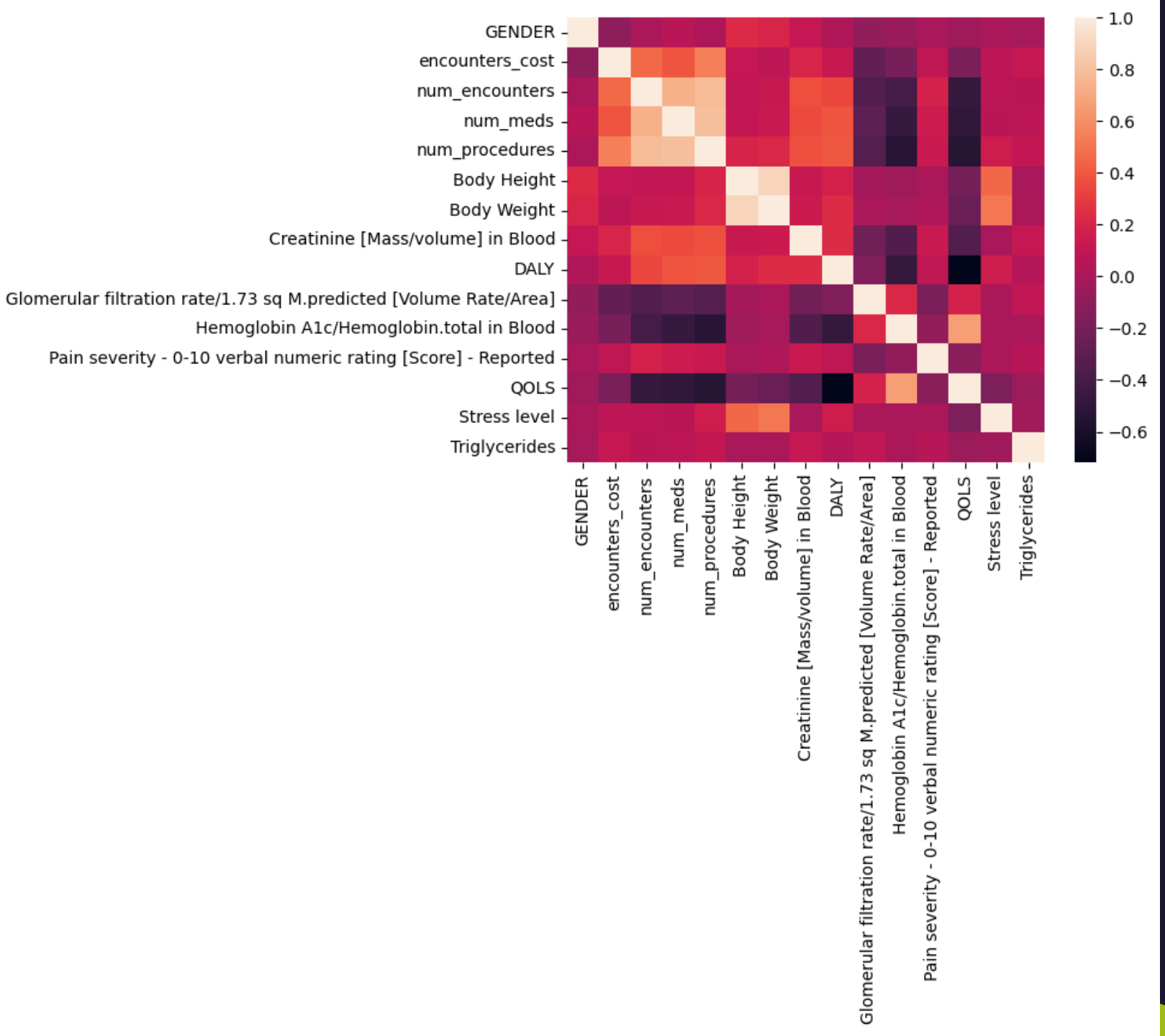


Medical Encounters by Gender and Race





Exploratory Data Analysis



Data Imputation

- Tested 4 different imputation techniques to fill in missing values
- Assessed R-squared values and distribution shape
- Selected K Nearest Neighbor

Imputation Technique	R-Squared Value
Mean	0.4934
Median	0.4941
K Nearest Neighbor (KNN)	0.5313
Multivariate Imputation by Chained Equations (MICE)	0.5139

Model Training and Development

- Split data into training (75%) and testing (25%) sets
- Baseline model – mean value of the training set (dummy regression)
- Evaluated R-squared and mean absolute error (MAE)

Mean value of training data	15,040.24
Training R-squared	0.0000
Testing R-squared	-0.0006
Training MAE	18,029.71
Testing MAE	17,276.07

Linear Regression Models

	Linear Regression	Ridge Regression	Lasso Regression
Training R-squared	0.4464	0.4073	0.4223
Testing R-squared	-0.1396	0.2567	0.4184
Training MAE	13,050.87	12,431.39	12,130.95
Testing MAE	12,799.56	12,021.69	11,354.02

Ensemble Models

	Random Forest	Gradient Boosting
Training R-squared	0.9426	0.9475
Testing R-squared	0.6350	0.6404
Training MAE	2,519.64	4,236.62
Testing MAE	6,167.80	6,439.04

Final Model Selection - Random Forest

- Random forest model was selected
- Included the best 45 features and 80 trees in the forest
- Cross validation R-squared: 0.6649
- Cross validation mean absolute error: 6,583.35

Final Model Selection - Random Forest

Random Forest Model vs. Dummy Regression Model	
Percent change training R-squared	100.00%
Percent change testing R-squared	100.10%
Percent change training MAE	615.57%
Percent change testing MAE	180.10%

Final Model Selection – Top Features

1. Number of medical encounters
2. Number of medical procedures
3. DALY (disability-adjusted life years)
4. Glomerular filtration rate/1.73 sq M
5. Leukocytes [#/volume] in Blood
6. Hematocrit [Volume Fraction] of Blood
7. Body mass index (BMI)
8. Pain severity - 0-10
9. Age
10. Cost of medications
11. Urea nitrogen [Mass/volume] in Blood
12. Chloride [Moles/volume] in Blood
13. Cholesterol in HDL [Mass/volume] in Serum or Plasma
14. Potassium [Moles/volume] in Blood
15. Number of medications
16. Triglycerides
17. Carbon dioxide total [Moles/volume] in Blood
18. QALY (quality adjusted life years)
19. Creatinine [Mass/volume] in Blood
20. State population

Conclusion

- Developed a machine learning model to predict yearly medical encounter costs from synthetic patient data
- On average, this model is expected to estimate a patient's yearly medical encounters cost within about $6,500
- Future work:
  - Include different types of data
  - Test multiple years