Predicting Yearly Medical Costs Using Synthetic Patient Data

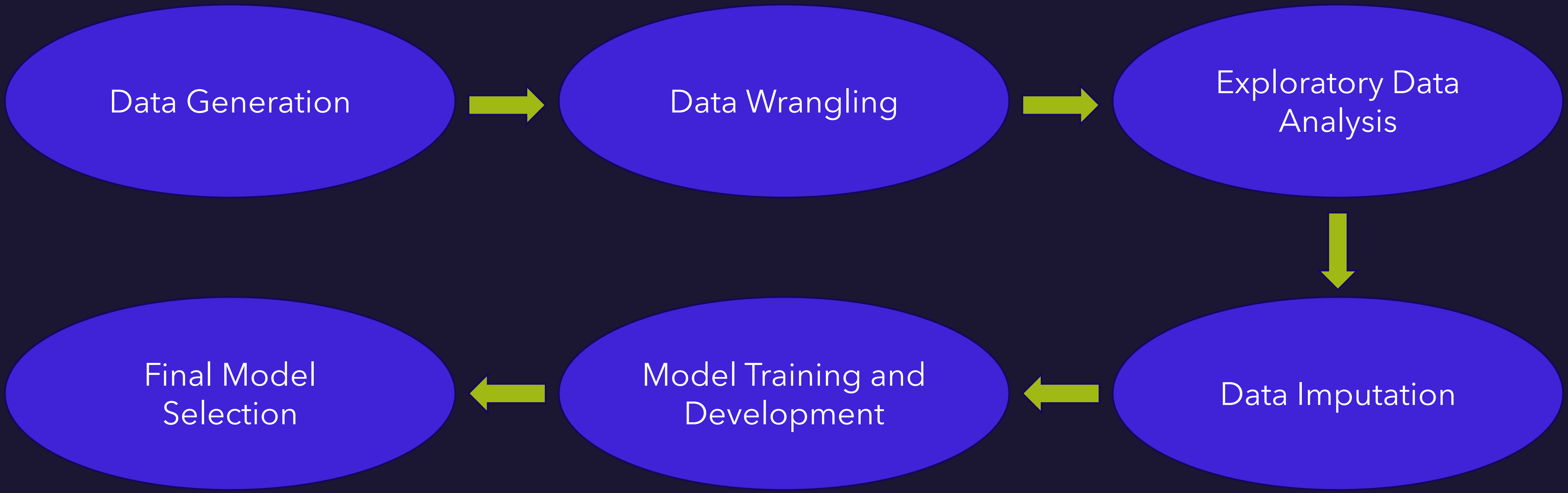
Stony Brook University Data Science Bootcamp Capstone Project

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

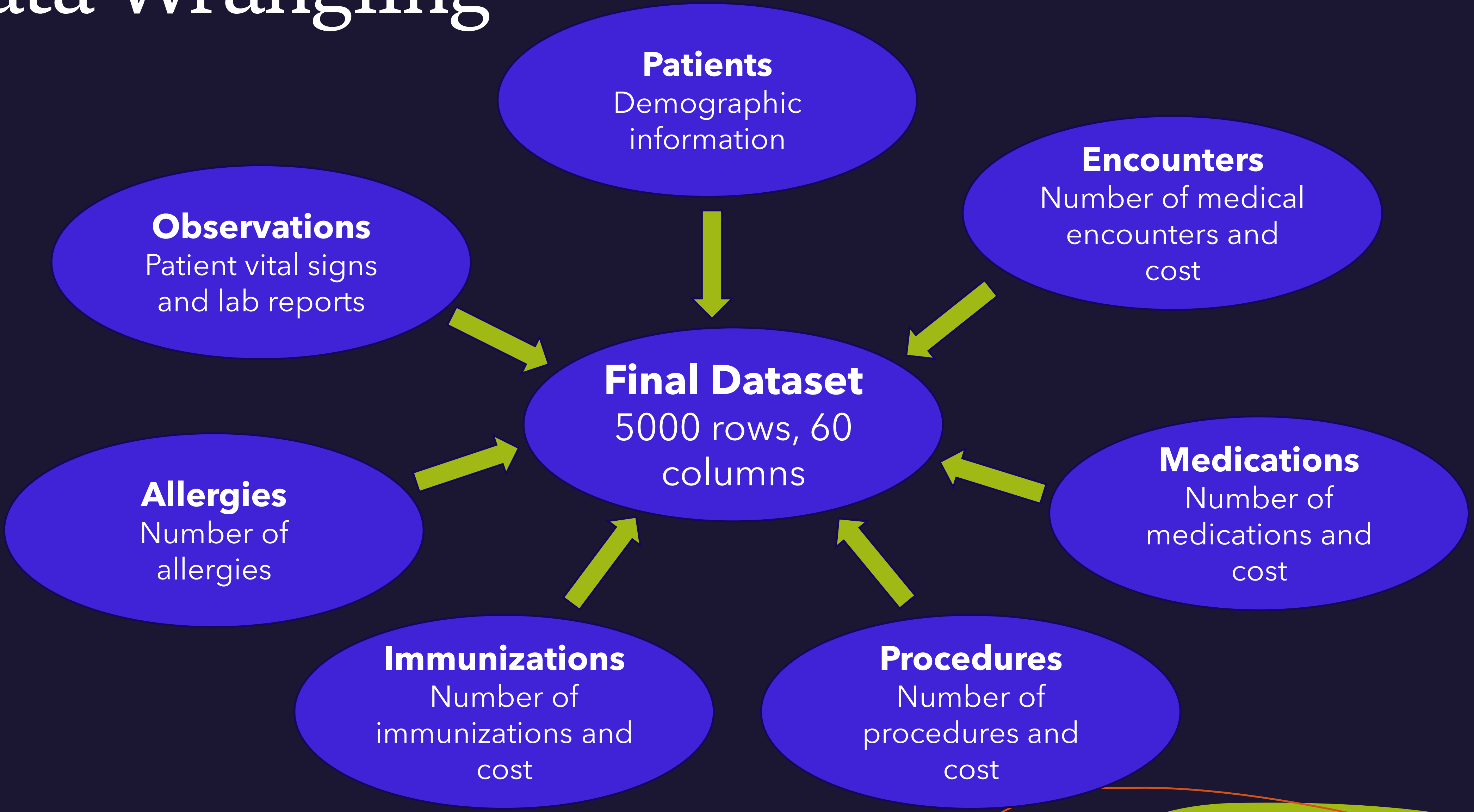
Model Training and Development

Final Model Selection

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



Patients
Demographic information

Encounters
Number of medical encounters and cost

Observations
Patient vital signs and lab reports

Final Dataset
5000 rows, 60 columns

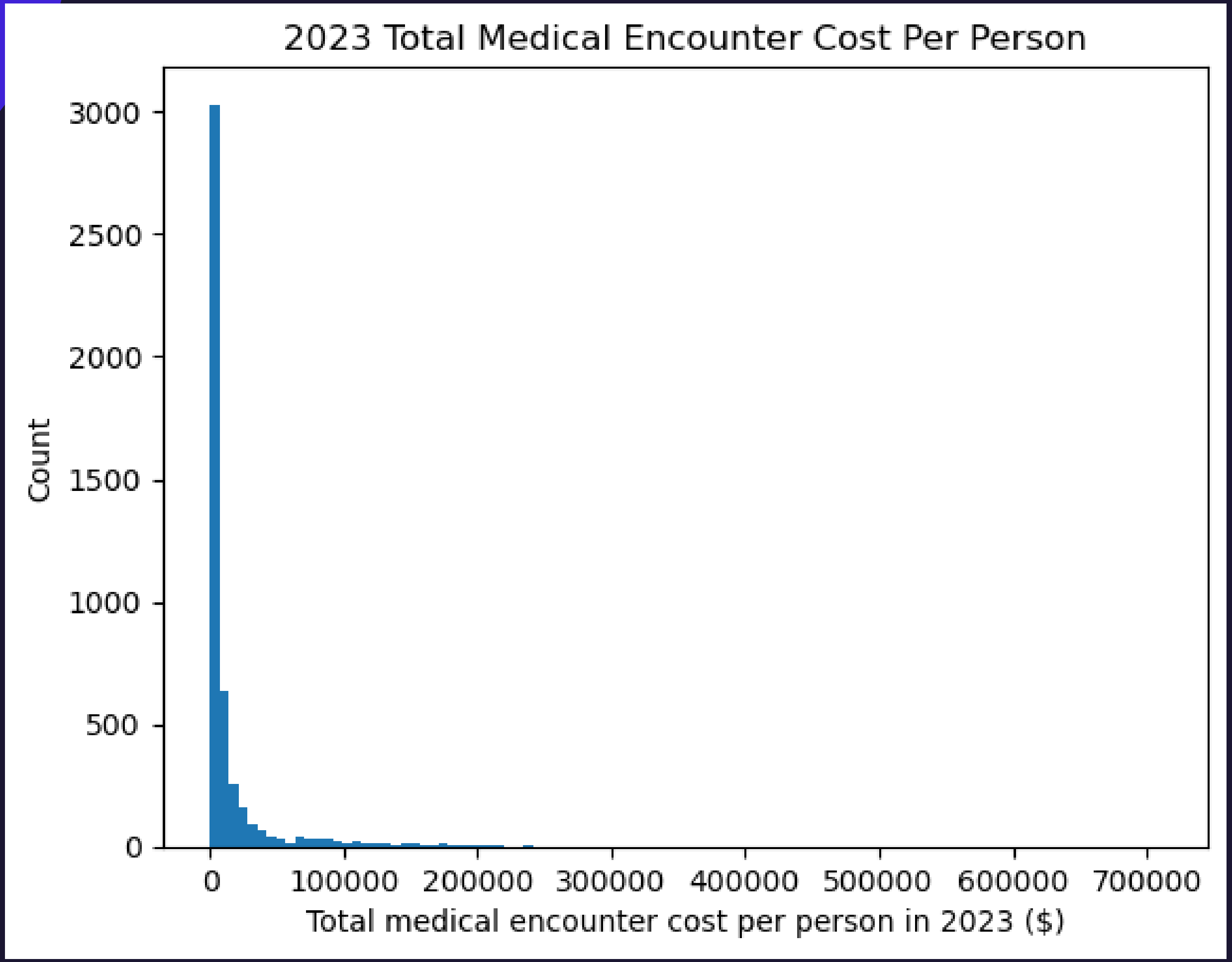
Allergies
Number of allergies

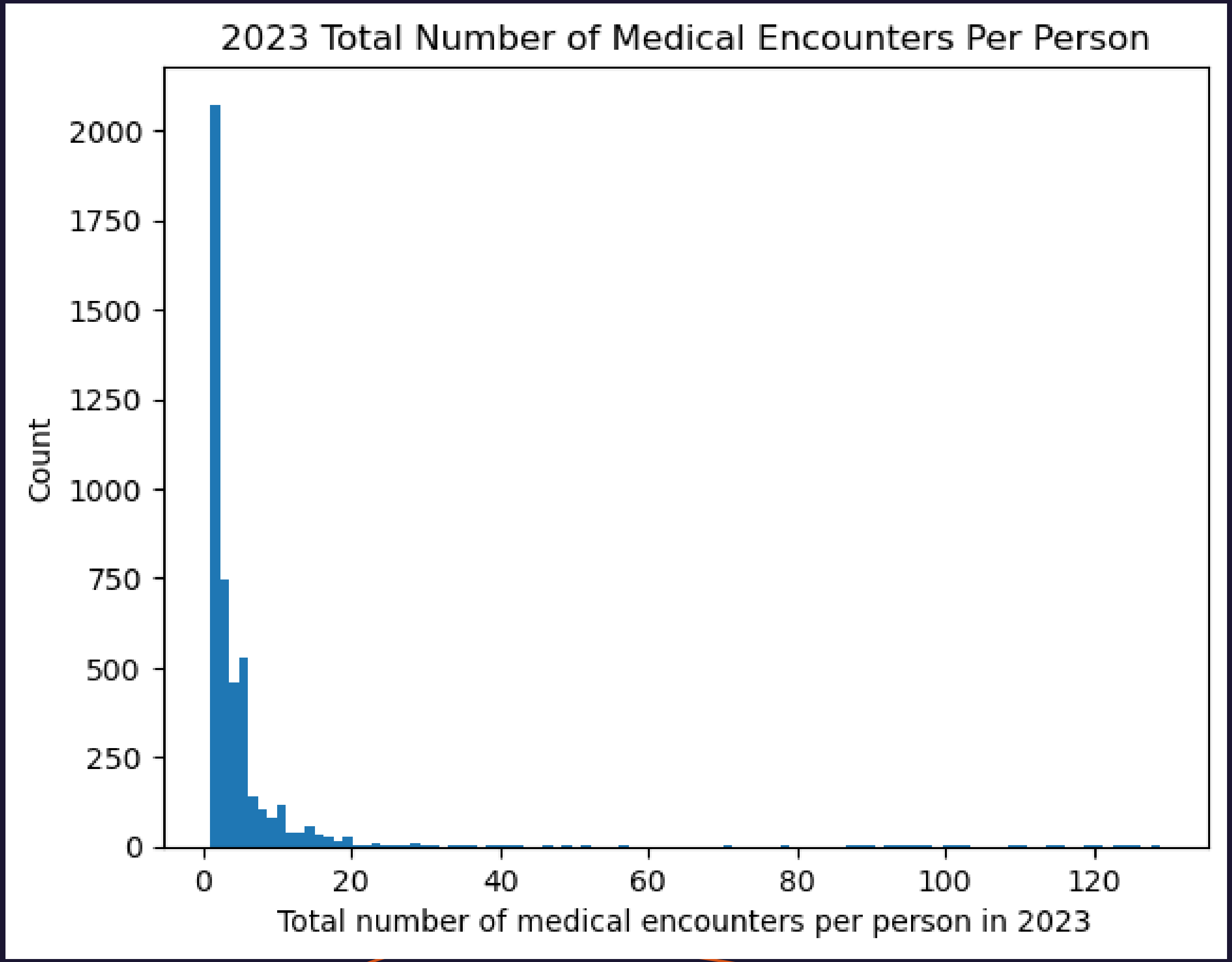
Medications
Number of medications and cost

Immunizations
Number of immunizations and cost

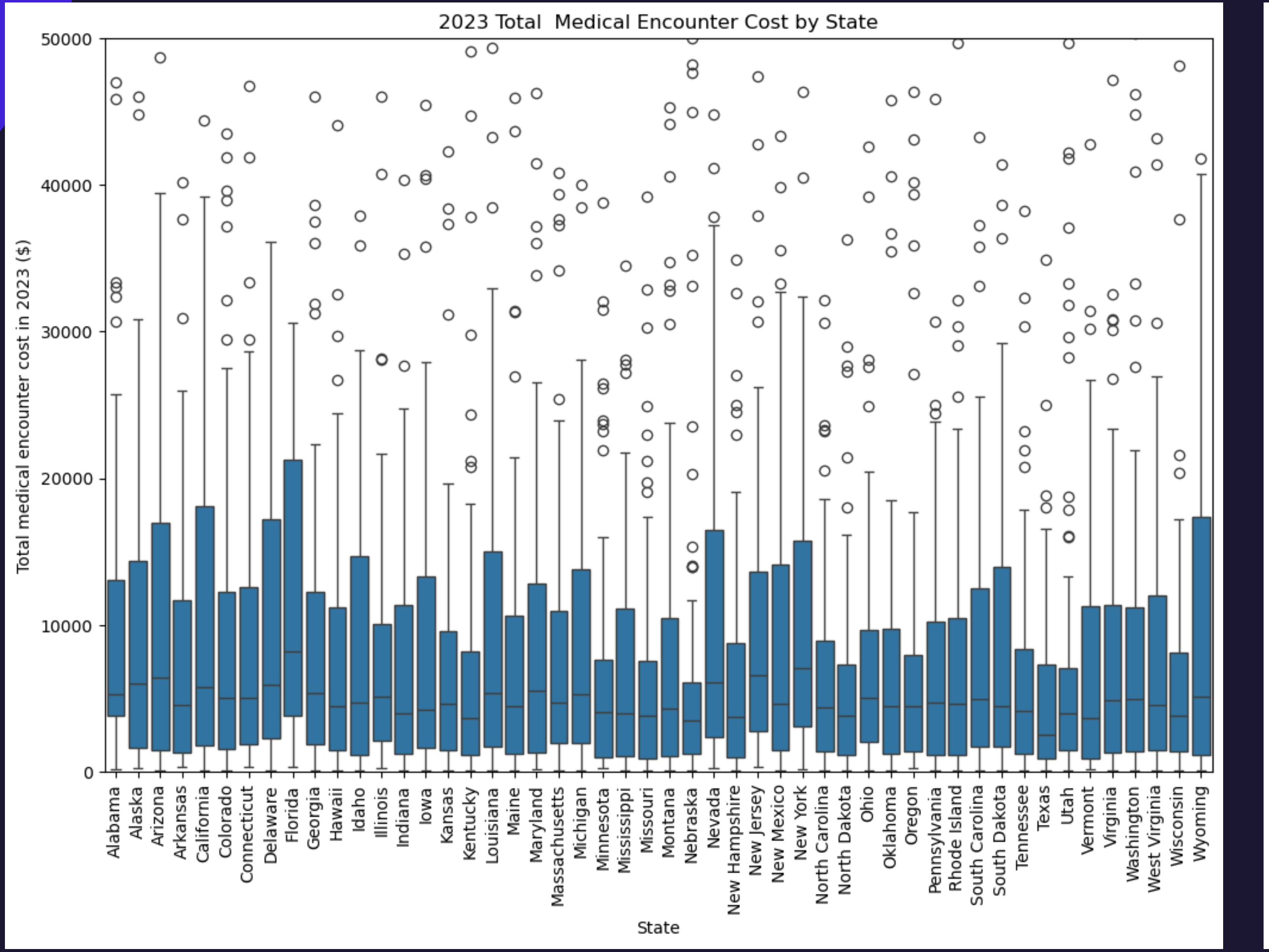
Procedures
Number of procedures and cost

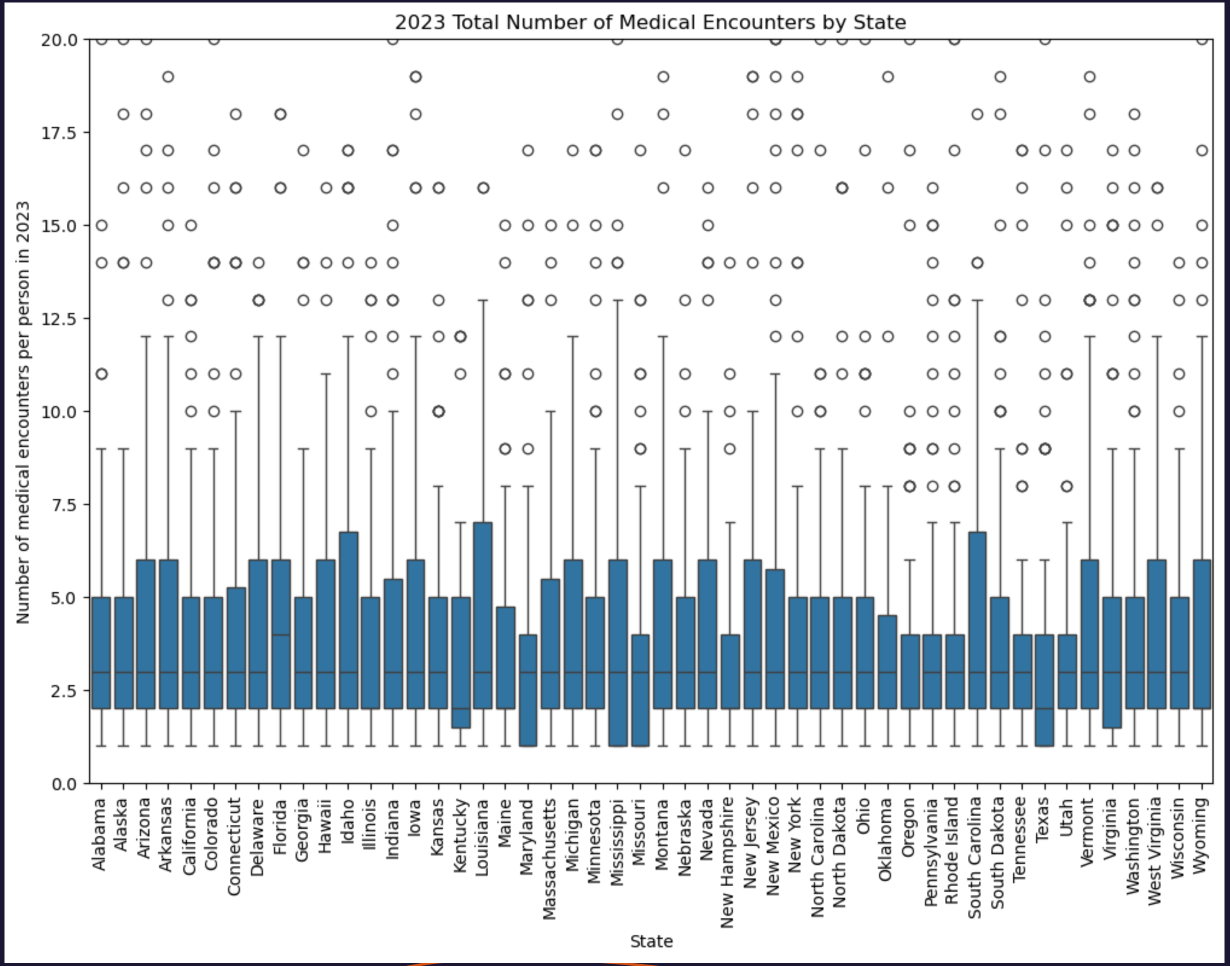
Medical Encounters Distributions



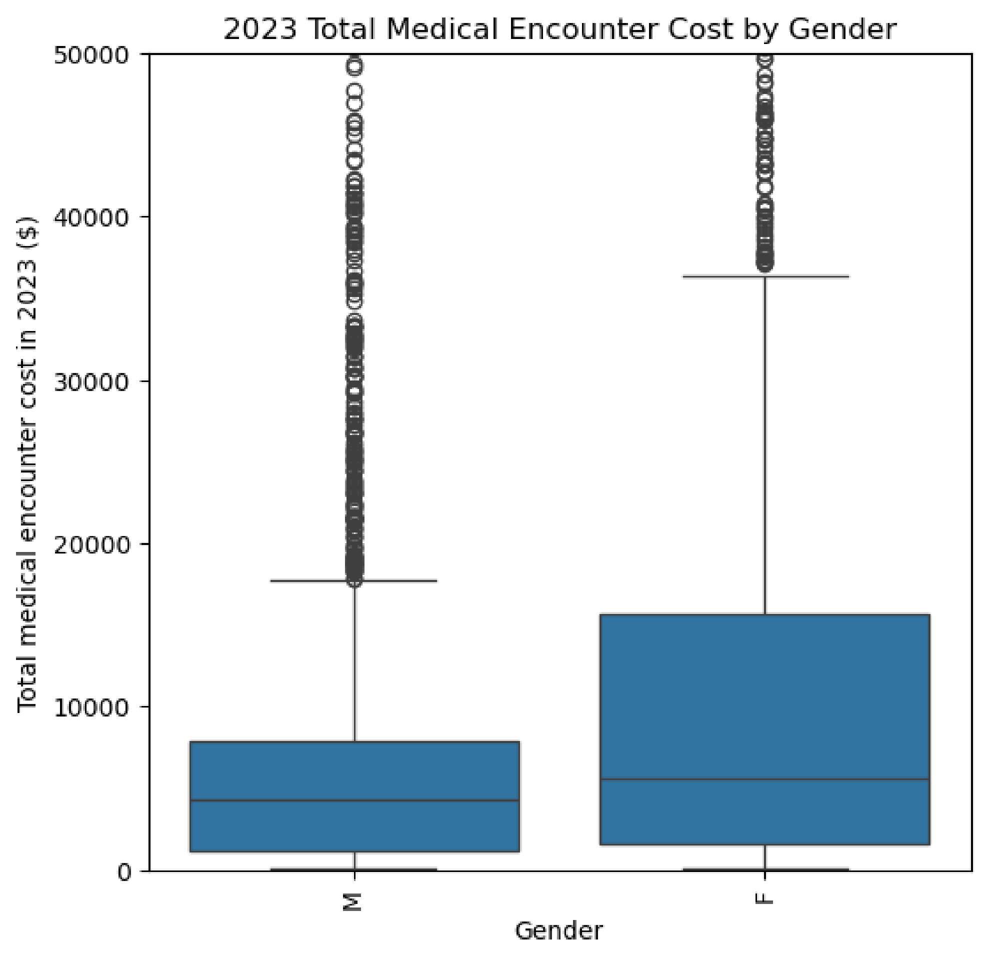


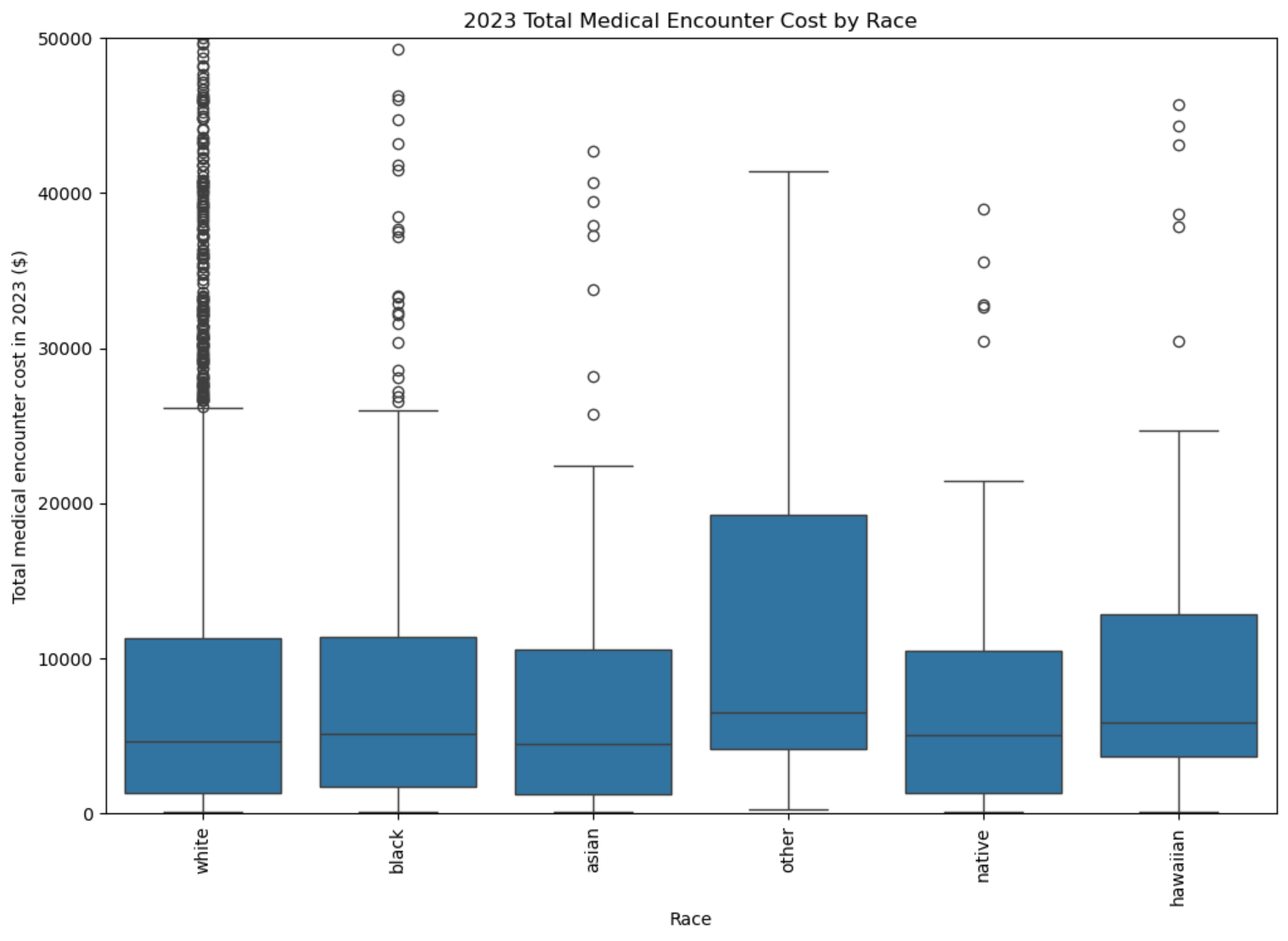
Medical Encounters by State



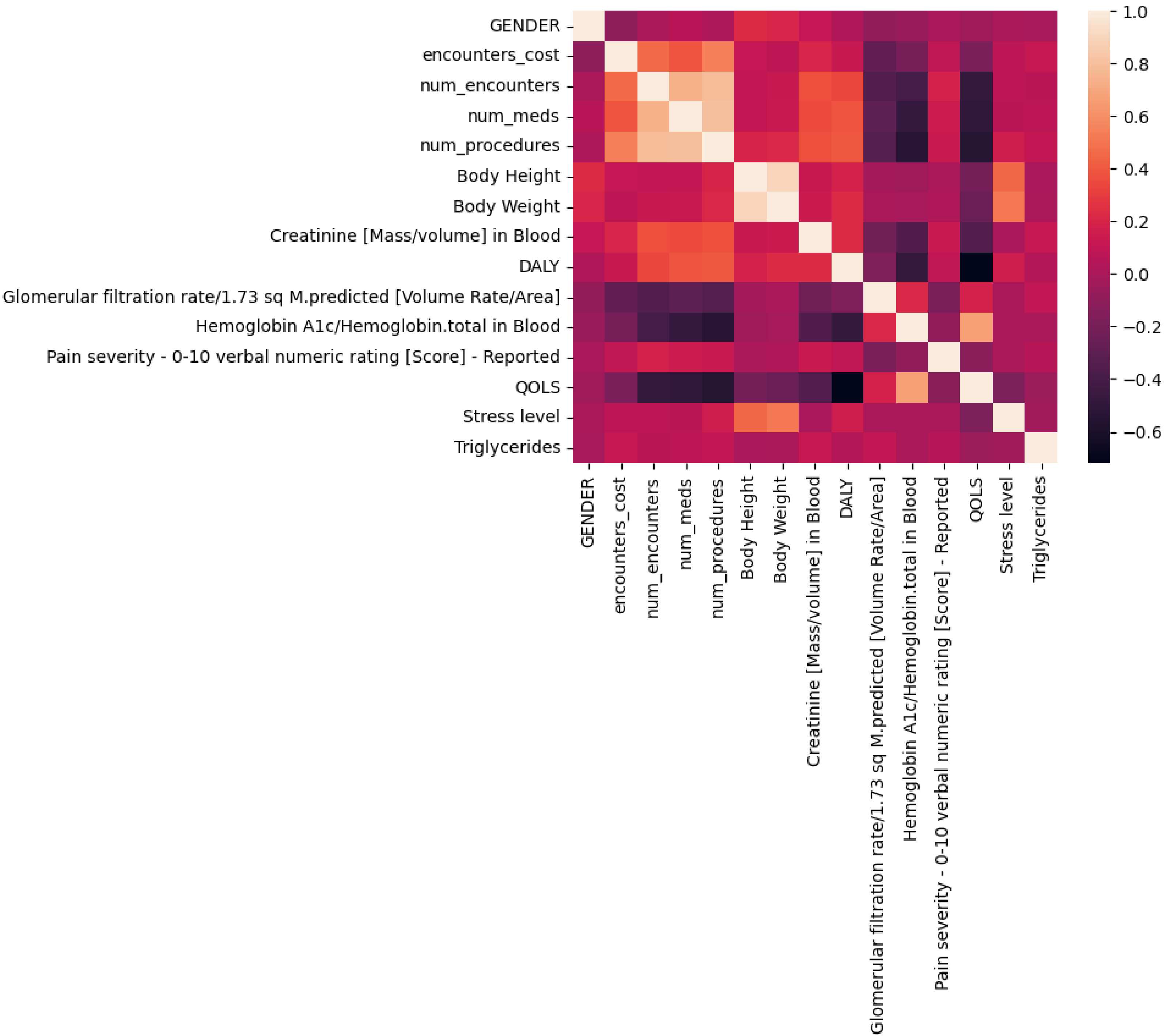


Medical Encounters by Gender and Race





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. Leukocytes [#/volume] in Blood
5. Glomerular filtration rate/1.73 sq M.predicted [Volume Rate/Area]
6. Age
7. Chloride [Moles/volume] in Blood
8. Cost of medications
9. Pain severity - 0-10
10. Number of medications
11. QALY (quality-adjusted life years)
12. Potassium [Moles/volume] in Blood
13. Body temperature
14. Body mass index (BMI) [Percentile] Per age and sex
15. Carbon dioxide total [Moles/volume] in Blood
16. Generalized anxiety disorder 7 item (GAD-7) total score [Reported.PHQ]
17. State population
18. Hemoglobin [Mass/volume] in Blood
19. Cholesterol in HDL [Mass/volume] in Serum or Plasma
20. Urea nitrogen [Mass/volume] in Blood

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