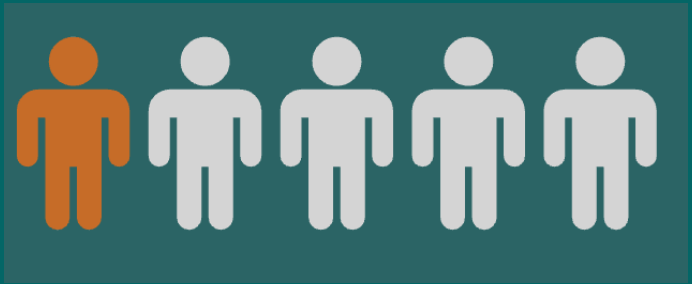
Dosage Disparities: Medicaid Reimbursement Trends and Predictive Insights

Raheela Charania
Prabhu Shankar
Rajivini Tiruveedhula

Background and Motivation

Medicaid covers 1 in 5 Americans.



Significant state-by-state differences exist in Medicaid drug reimbursements.

Reimbursement Insights Drive Better Allocation and Customer Transparency.

Exploring Medicaid reimbursement trends and identifying the key factors that influence them can lead to more efficient and equitable healthcare resource distribution.

Business Problem / Research Question

Main Question

What factors drive Medicaid drug reimbursement variations across U.S. regions?

Secondary Question

Can socioeconomic factors like GDP, poverty and weather conditions predict dosage disparities?

Goals and Objectives

Explore state-level Medicaid reimbursement trends (2019–2024), build predictive models for units reimbursed and test if socioeconomic factors significantly explain variations.

Data Sources

Medicaid Drug Utilization Data

2019-2024 data from the Centers for Medicare & Medicaid Services (CMS)

Massive Dataset

~5 million records per year, requiring efficient data processing and handling

State-level Socioeconomic Data

GDP, poverty and weather conditions from various government sources

The team leveraged a comprehensive set of Medicaid reimbursement data and socioeconomic indicators to explore the factors driving drug reimbursement disparities across the U.S.

Dataset Description

Identifiers & Keys

- Product Name
- Product Code
- Package Size
- NDC

Drug Details

- Units of Drug Reimbursed (# of tablet, patches, vials, etc.)
- # of Prescriptions Reimbursed
- Manufacturer/ Distributor of Drug
- Total Amount Reimbursed
- Medicaid & Non-Medicaid Amount Reimbursed

Date/Time Info

- Year & Quarter of Dataset

Utilization Type

- Fee-For-Service Utilization
- Managed Care Organization Utilization

Macro-Level State Indicators

- Per Capita Personal Income
- % of Population below Poverty Line
- Average Annual Temperature per State
- State-level Real GDP
- Total State Population

Data Cleaning

Data Filtering:
FFSU vs. **MCOU**

75% of states enrolled in MCOU; trend continues upward.

Removed outliers

Outliers were removed due to inconsistencies.

Variable Reduction

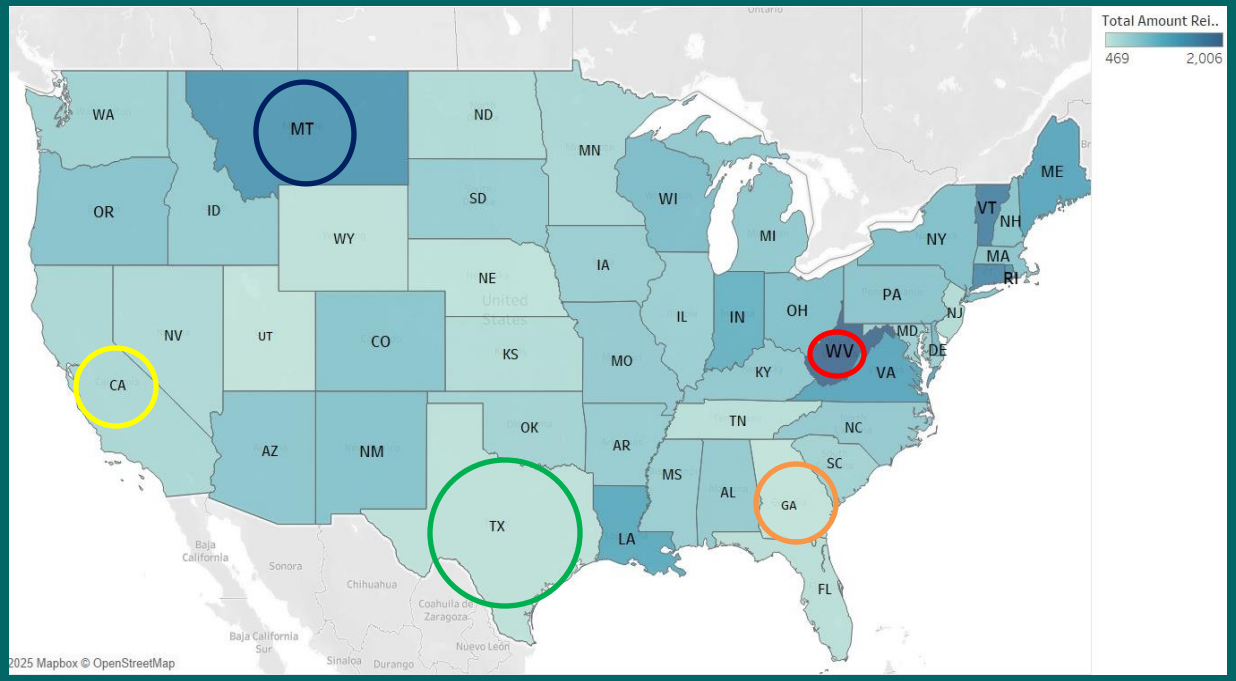
Product & labeler codes, package size dropped—already captured by NDC.

Numeric Standardization

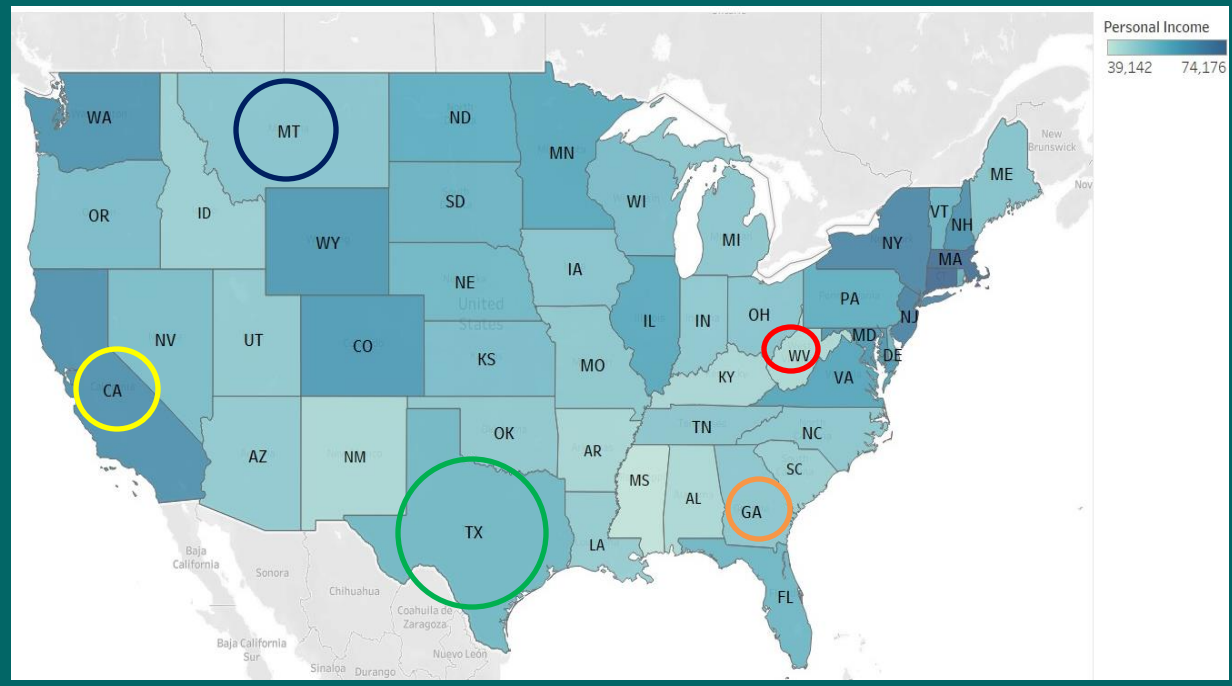
Numeric predictors were Min-Max scaled to ensure consistent ranges.

The data was thoroughly cleaned and prepared for further analysis, ensuring a robust foundation for the upcoming modeling and insights.

Exploratory Data Analysis



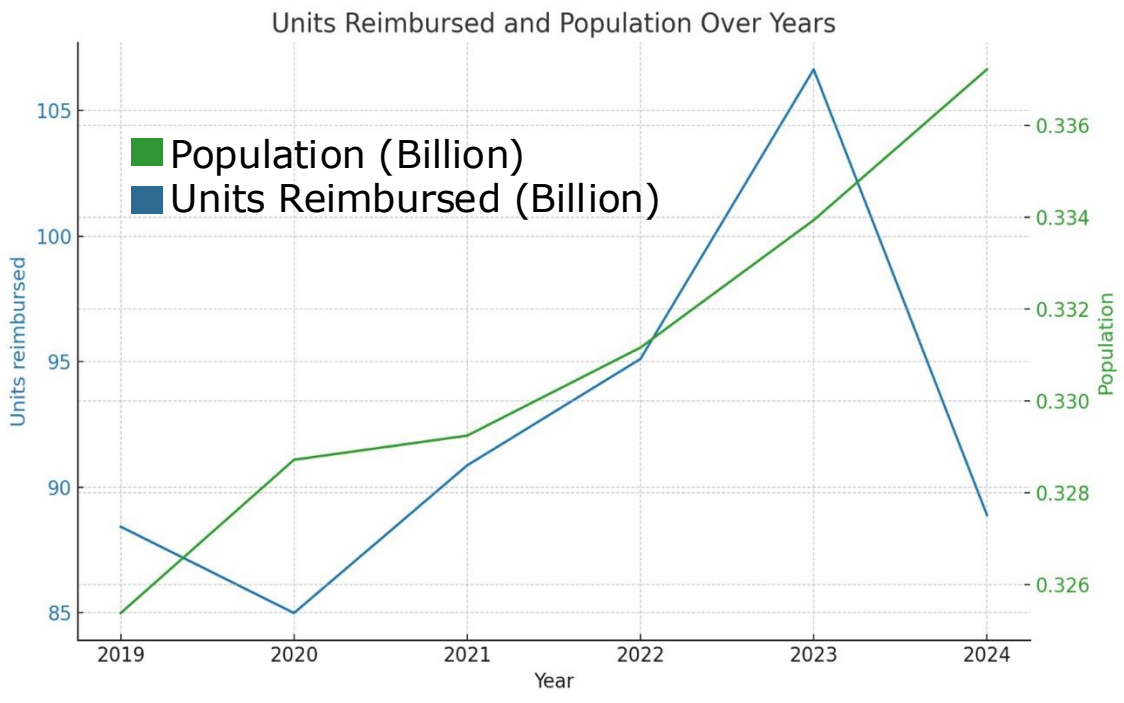
Total amount reimbursed per 100K individuals (2019)



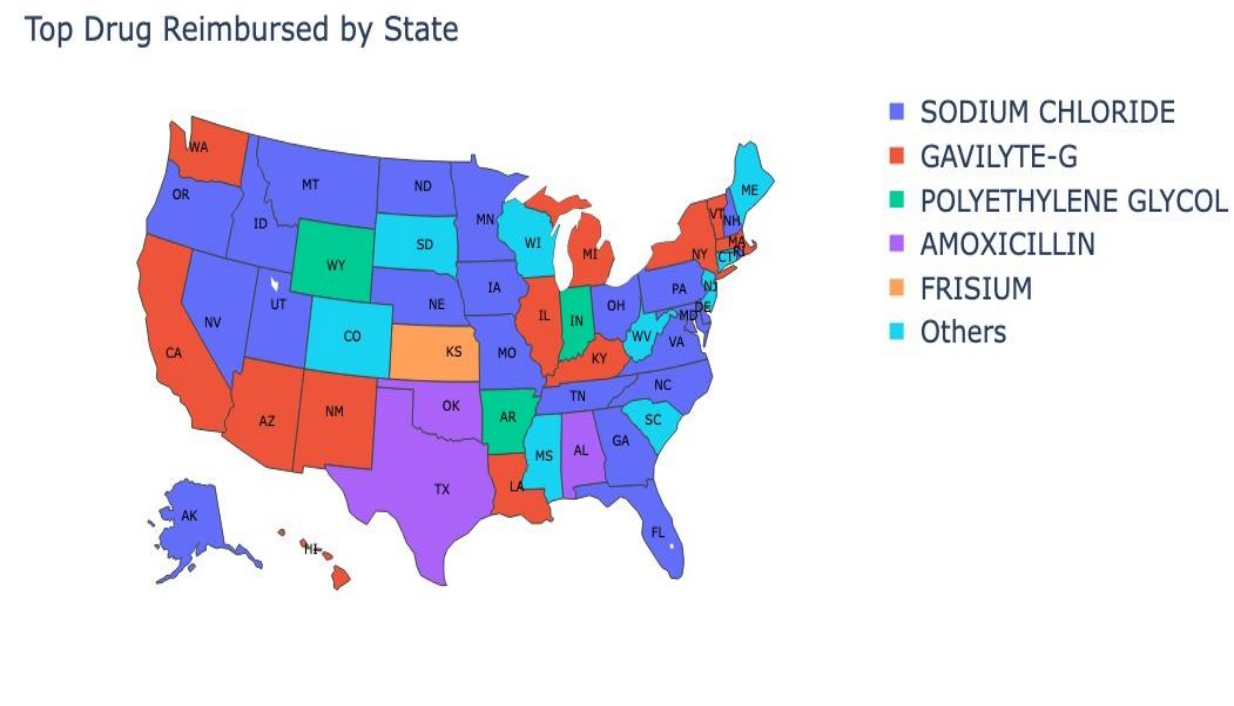
Per Capita Personal Income (2019)

A negative correlation is observed between Per Capita Personal Income and the total amount reimbursed per 100K individuals.

Exploratory Data Analysis



Units Reimbursed vs. Population over the years

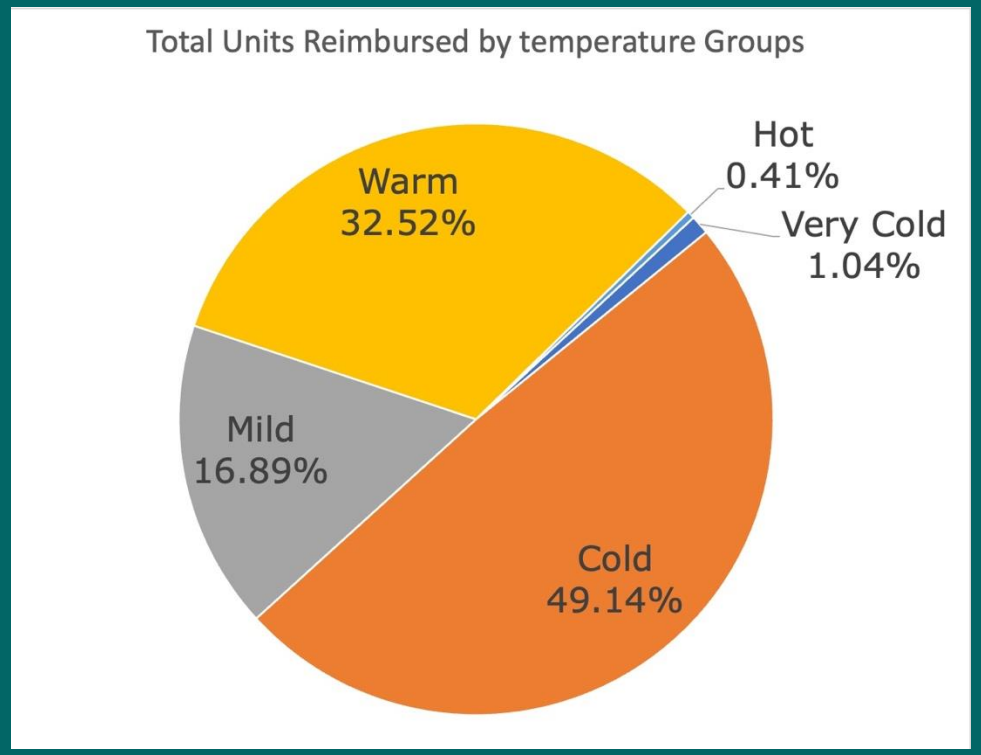


Top Drug by Units Reimbursed (2024)

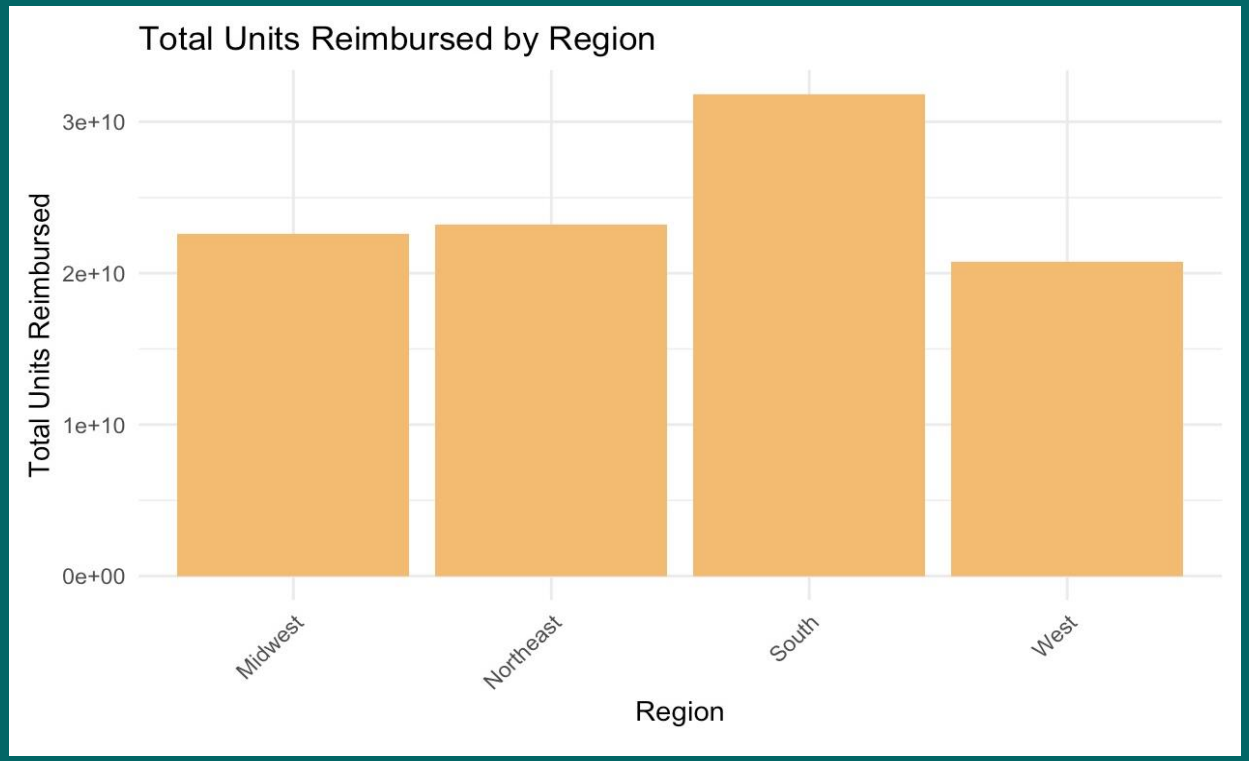
On average, as population increases, so does units reimbursed.

U.S. map showing top drug reimbursed in the U.S. is Sodium Chloride.

Exploratory Data Analysis



Units Reimbursed Across Different Temperature

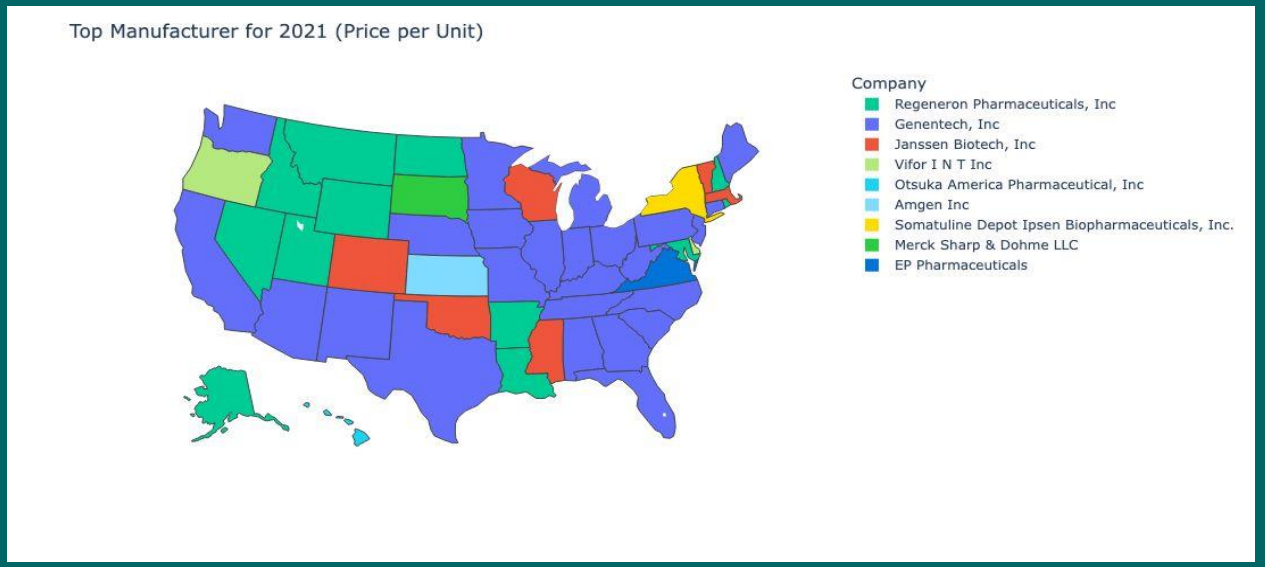


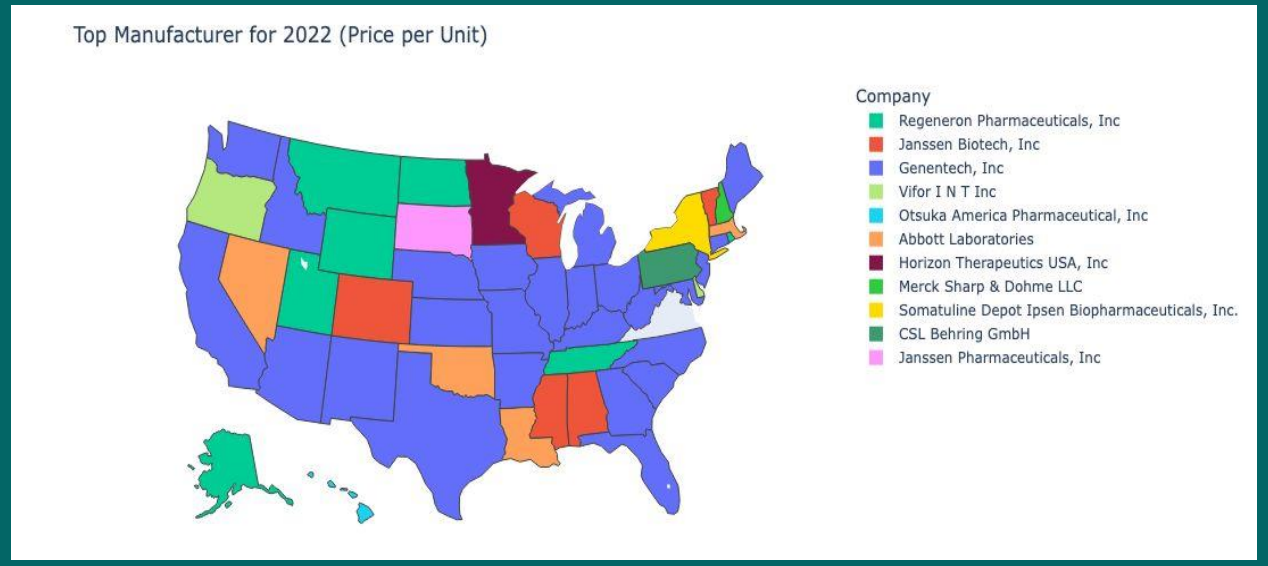
Units Reimbursed by Region

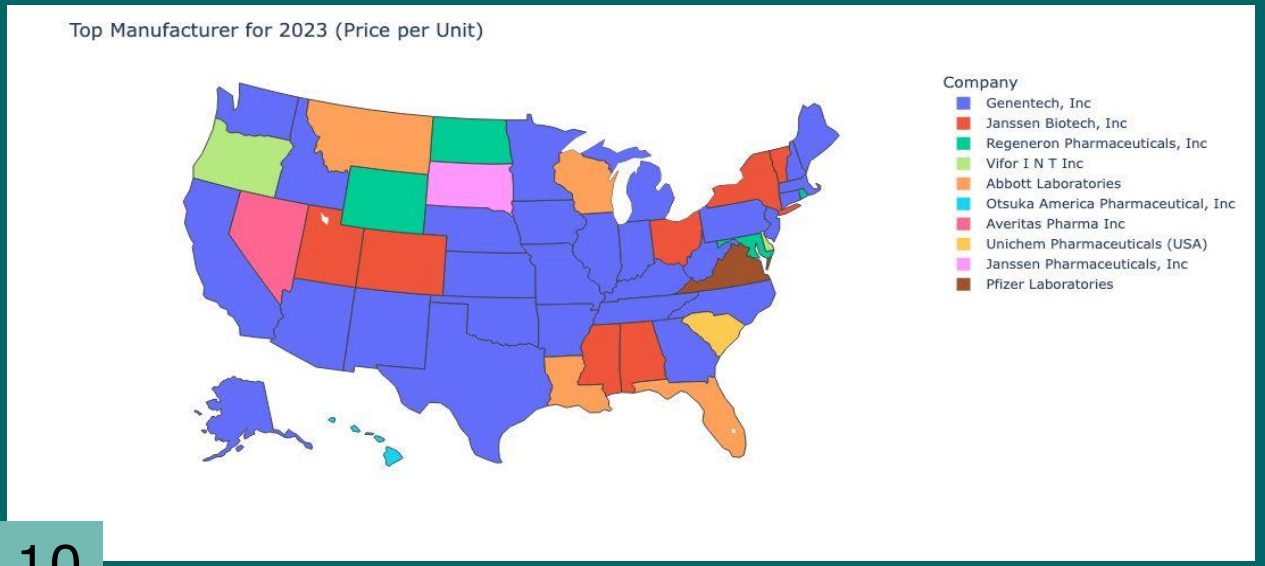
As expected, cold weathers have the highest reimbursements.

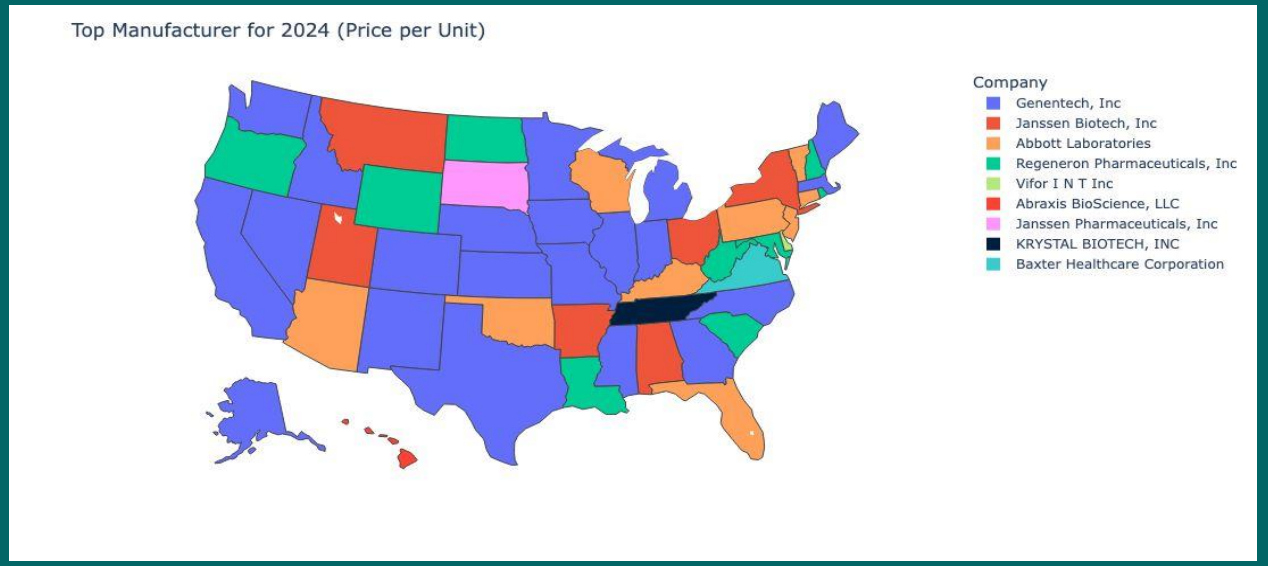
Units reimbursed according to population, grouped by regions

Top Manufacturer Trends (2021-2024)









Preprocessing for Model Creation

Combined Medicaid data with Socioeconomic data

Merged Medicaid drug utilization data with external metrics including Real GDP, average temperature, state poverty rates, population estimates and per capita income.

ARIMA-Based Index Projection

Applied ARIMA to forecast socioeconomic index trends to predict future trends

Target Encoding

Categorical variables were encoded using target mean encoding with 5-fold cross-validation, mapping category-wise training means onto validation folds.

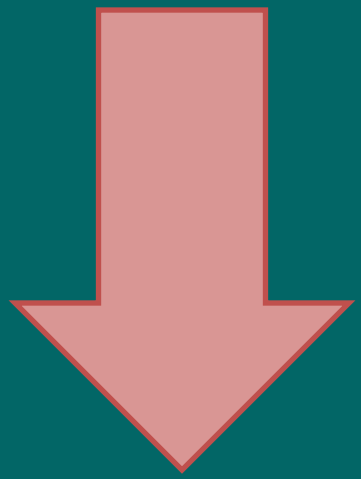
Data Partitioning

80% Training Dataset

20% Test Dataset

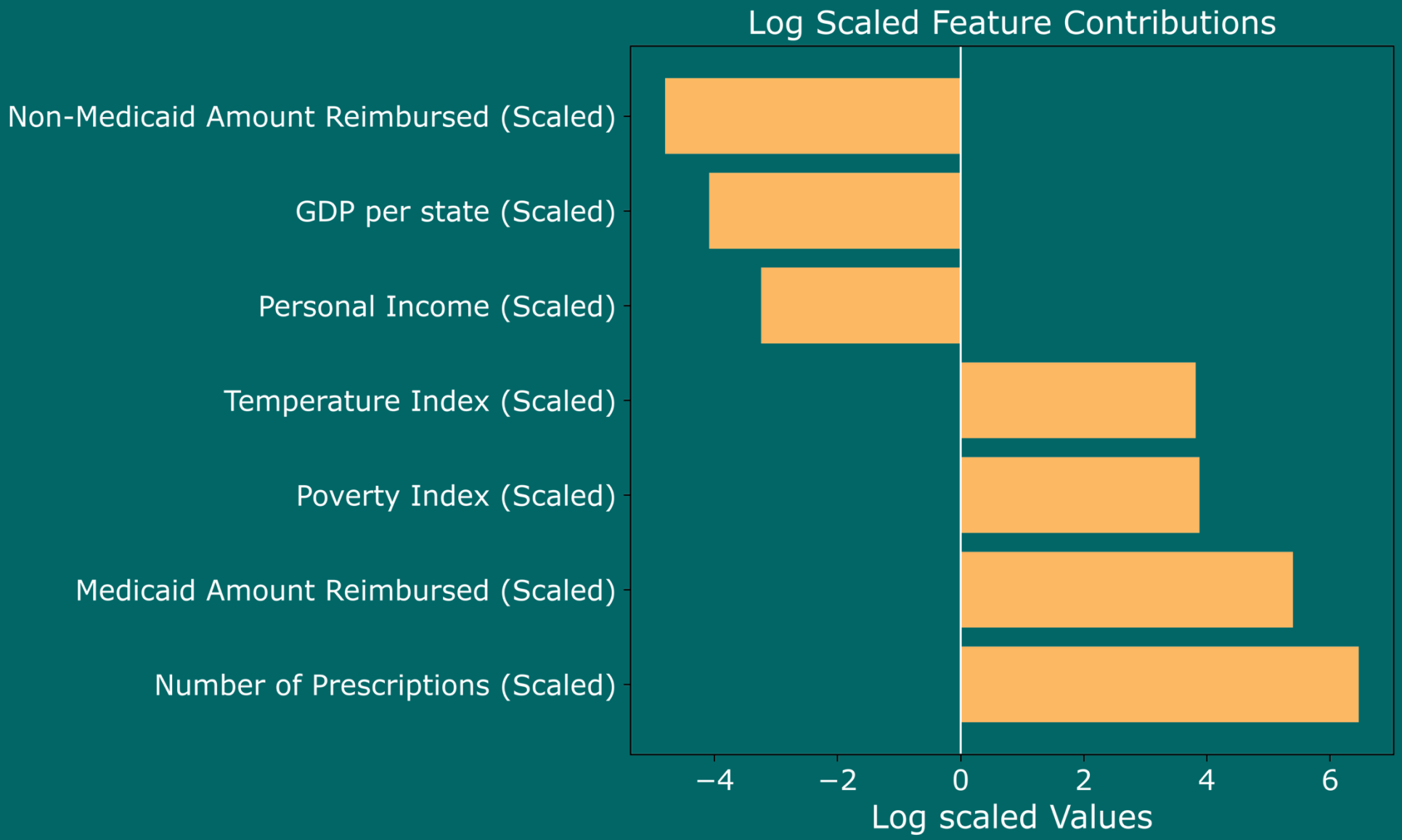
Model 1: Multiple Linear Regression

Variable	Coefficient
Temperature Index (Scaled)	6653.0
Poverty Index (Scaled)	7616.1
Medicaid Amount Reimbursed (Scaled)	255289.5
Number of Prescriptions (Scaled)	2963643.9

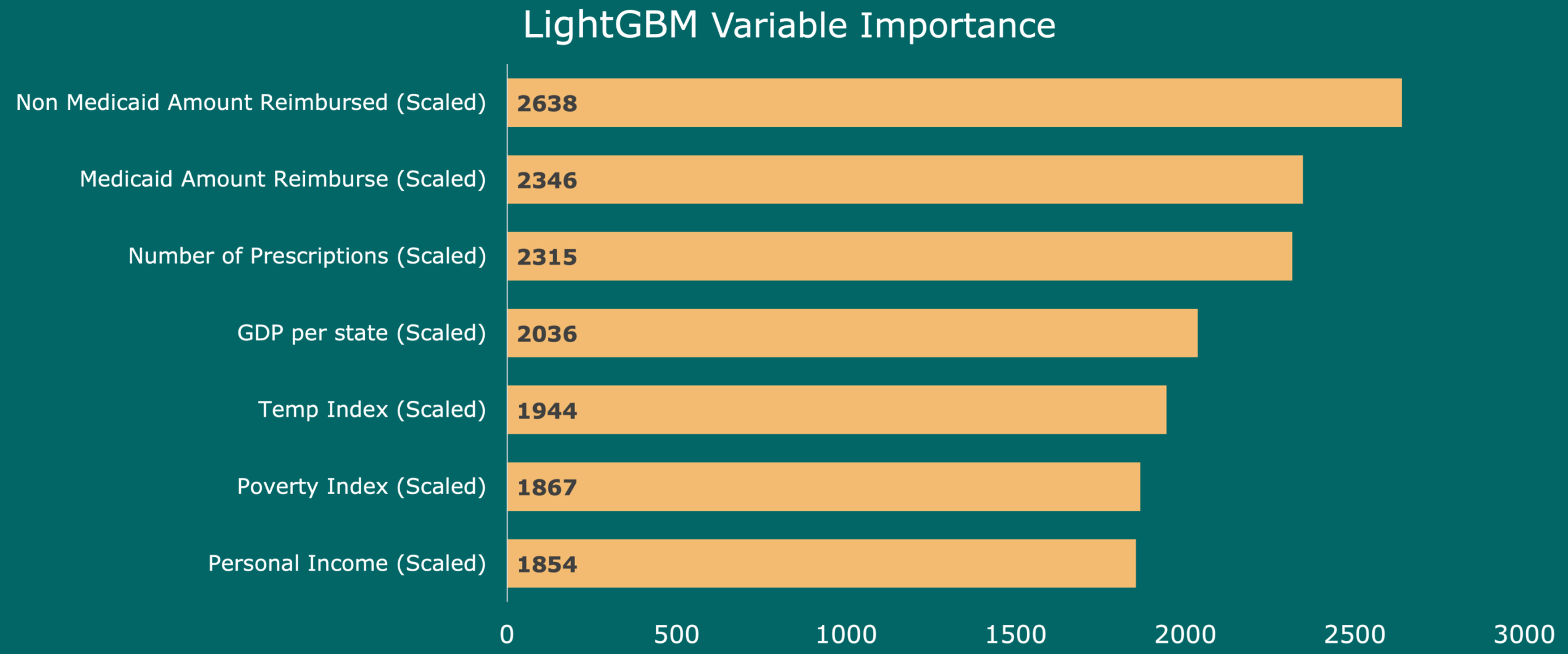


Variable	Coefficients
Personal Income (Scaled)	-1734.40
GDP per state (Scaled)	-12113.50
Non-Medicaid Amount Reimbursed (Scaled)	-64398.60

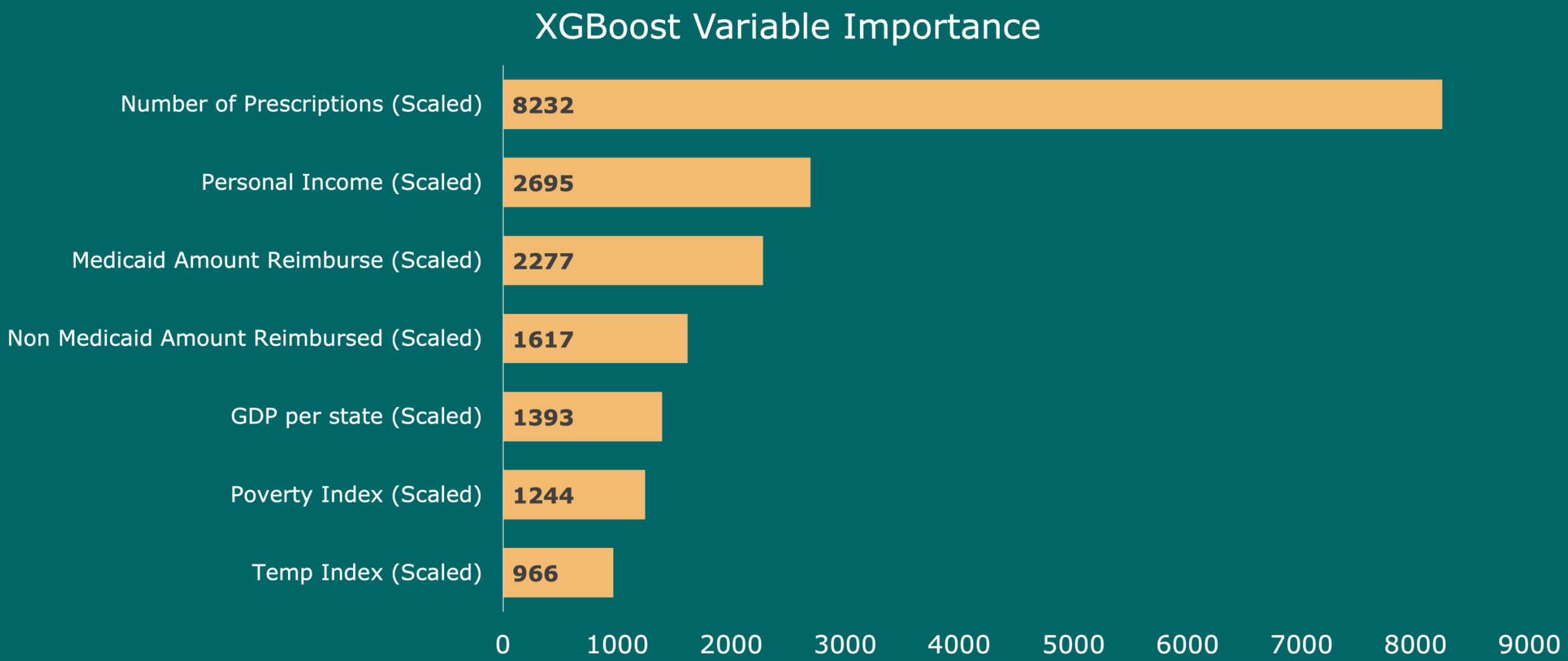
Model 1: Multiple Linear Regression



Model 2: Light GBM Method

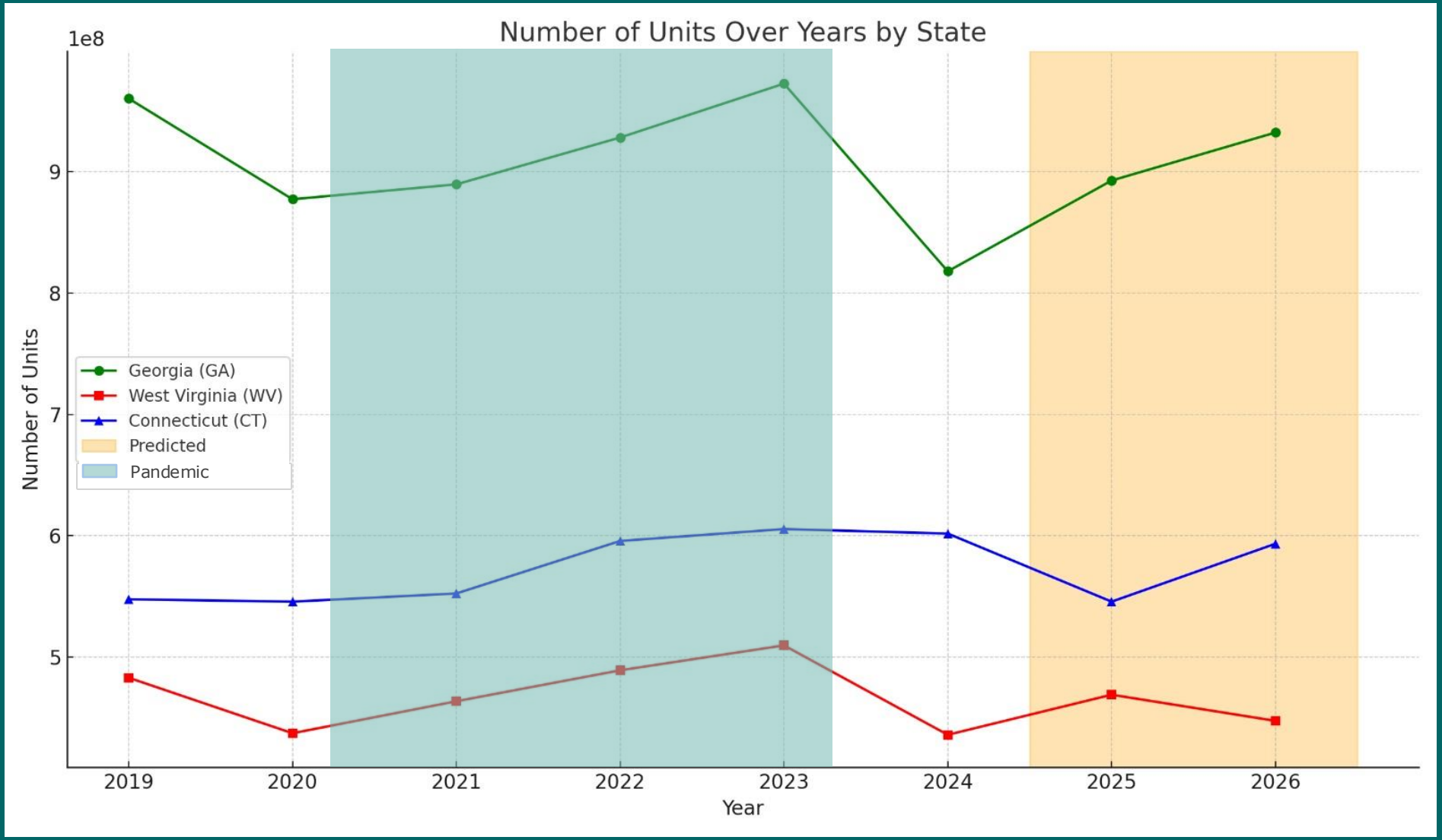


Model 3: XGBoost Method



Predictive Performance Comparison (R^2)

Predictions via XGBoost Model



Cost-Benefit Analysis of Predictive Modelling

Item	Value
Baseline error cost (8% RMSE)	$274,424,620
Model error cost (1% RMSE)	$34,303,080
Variance Reduced	$240,121,540
Estimated realized savings (10%)	$24,012,154

Model improvements reduced error variance by $240M, enabling an estimated **$24M in realized savings**.

Key Findings

Medicaid reimbursement driven by prescription volume

Analysis shows the number of prescriptions and personal income are the most important variables.

Consistency in reimbursed drug types across states

The top reimbursed drugs, such as Sodium Chloride, are consistent across U.S. states (22).

Inverse Link: Income vs. Reimbursements

States with higher Per capita Personal Income have lower reimbursements.

The analysis highlights the need to focus Medicaid forecasting and monitoring efforts on prescription volume trends and drug-specific patterns, rather than relying heavily on state-level economic factors.

Recommendations

Focus predictive monitoring on prescription behavior as well as state economics

Analysis shows the number of prescriptions and amount reimbursed are highly significant whereas Temperature and poverty index are significant.

Prioritize volume patterns and drug-specific trends in future Medicaid forecasting models

Socioeconomic data offers limited incremental value for short-term forecasting

Future Work

Add Drug Class Categories

Explore grouping drugs by class (ex- painkillers, antibiotics) to uncover more nuanced trends.

Prescription Practices Not Captured

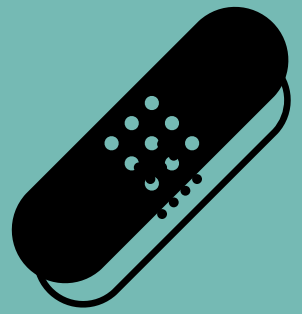
Factors like hospital protocols that influence prescription behavior are not widely available.

Explore Clustering by Policy vs. Geography

Group states based on shared Medicaid policies to reveal policy-driven insights.

The analysis was constrained by data availability and scope, offering opportunities for future work to enhance the understanding of Medicaid reimbursement disparities.

Thank you!
Questions?



Appendix Table of Contents

Additional Macrovariables Data Dictionary

Name	Type	Description
Total_Amount_Reimbursed	Numeric	Total amount paid
Medicaid_Amount_Reimbursed	Numeric	Paid by Medicaid
State_GDP	Numeric	State-level Real GDP
Poverty_Index	Numeric	Percentage of population below poverty line
Temperature_Index	Numeric	Average annual temperature per state
Per_Capita_Income	Numeric	Average income per person in the state
Population_Estimates	Numeric	Total population of the state

Medicaid Dataset Description

Name	Type	Description
utilization_type	string	Fee-For-Service Utilization & Managed Care Organization Utilization
state	string	State Name
ndc	string	LabelerCode + ProductCode + PackageSize
labeler_code	string	Name of Manufacturer/ Supplier
product_code	string	Manufacturer, labeler, packager or distributor of the Drug
package_size	string	Size of the Package of the Drug
year	integer	Year of Dataset
quarter	integer	Quarter of Dataset
product_name	string	Name of Product
units_reimbursed	number	Units of Drug Reimbursed
number_of_prescriptions	integer	Number of Prescriptions Reimbursed
total_amount_reimbursed	number	Total Amount Reimbursed
medicaid_amount_reimbursed	number	Medicaid Amount Reimbursed
non_medicaid_amount_reimbursed	number	Non-Medicaid Amount Reimbursed

*Medicaid Drug Utilization Dataset (CMS), 2019-2024

Model 1: Multiple Linear Regression Results

Variable	Estimate	Std. Error	t value	p-value
(Intercept)	3217	808	4	6.89E-05
Personal Income (Scaled)	-1734	1206	-1	0.15
GDP per state (Scaled)	-12114	855	-14	< 2e-16
Poverty Index (Scaled)	7617	1250	6	1.12E-09
Temp Index (Scaled)	6653	862	8	1.15E-14
Non-Medicaid Amt Reimbursed (Scaled)	-64399	2751	-23	< 2e-16
Medicaid Amt Reimbursed (Scaled)	255290	1799	142	< 2e-16
Number of Prescriptions (Scaled)	2963644	4013	739	< 2e-16

Units Reimbursed = 13836 − 17925×Personal Income(Scaled) − 16587×GDP per State(Scaled) + 8453×Poverty Index(Scaled) + 15508×Temp Index(Scaled) + 80725×Non-Medicaid Amount Reimbursed(Scaled) + 3142535xNumber of Prescriptions(Scaled)

Model 1: Multiple Linear Regression Results

Metric	Value
Multiple R-squared	0.6856
Adjusted R-squared	0.6856
F-statistic	1.11e+05
F-statistic p-value	<2.2e-16
Test Set R-squared	0.6737

Variable	VIF Value
GDP per State (Scaled)	1.210120
Poverty Index (Scaled)	2.523569
Temperature Index (Scaled)	1.366394
NonMedicaid Amount Reimbursed (Scaled)	1.008370
Number of Prescriptions (Scaled)	1.016373
Personal Income (Scaled)	2.541849

Units Reimbursed = 13836 − 17925×Personal Income(Scaled) − 16587×GDP per State(Scaled) + 8453×Poverty Index(Scaled) + 15508×Temp Index(Scaled) + 80725×Non-Medicaid Amount Reimbursed(Scaled) + 3142535xNumber of Prescriptions(Scaled)

Predictive Performance Comparison

Model	R^2 Train	R^2 Test
Linear (Full) Model	0.6856	0.6739
Linear (Stepwise) Model	0.6856	0.6739
Light Gradient Boost Model	0.8932	0.88
eXtreme Gradient Boost Model	0.9345	0.8932

Correlation Matrix