Medicaid Drug Reimbursement Analysis

1.Introduction

Medicaid drug reimbursement patterns significantly impact healthcare spending, policy decisions, and access to essential medications across the United States.
This project explores the relationship between drug utilization, reimbursement amounts, and state-wise prescription behavior by analyzing historical Medicaid drug data. The primary goal is to identify how factors such as prescription volume, package size, and product type influence total reimbursement costs, and to develop predictive models for estimating reimbursements based on these variables. Through this analysis, the research aims to better understand drug spending trends across states and explore potential applications for cost forecasting, public health planning, and data-driven policy making.

2. Dataset Overview

This analysis is based on a single dataset: the **State Drug Utilization Data (SDUD) for 2023**, published by Medicaid. The dataset (500.3 MB) was obtained from the official Medicaid portal and contains comprehensive information on outpatient drug reimbursements across all U.S. states.

The dataset includes multiple columns such as **State**, **National Drug Code (NDC)**, **Labeler Code**, **Product Code**, **Package Size**, **Utilization Type**, **Number of Prescriptions**, **Units Reimbursed**, **Total Amount Reimbursed**, and **Medicaid vs. Non-Medicaid Reimbursement** amounts. These variables provide valuable insights into prescription patterns, drug costs, and reimbursement structures.

3. Data Preprocessing and Transformation

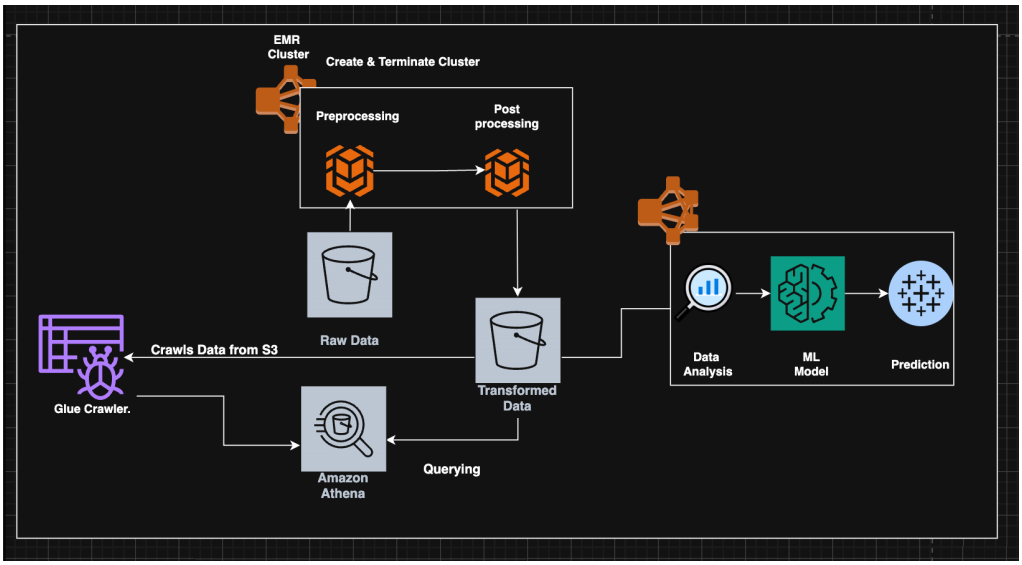
The Medicaid drug reimbursement dataset (500.3 MB) was preprocessed to ensure consistency, remove irrelevant information, and enable efficient analysis.

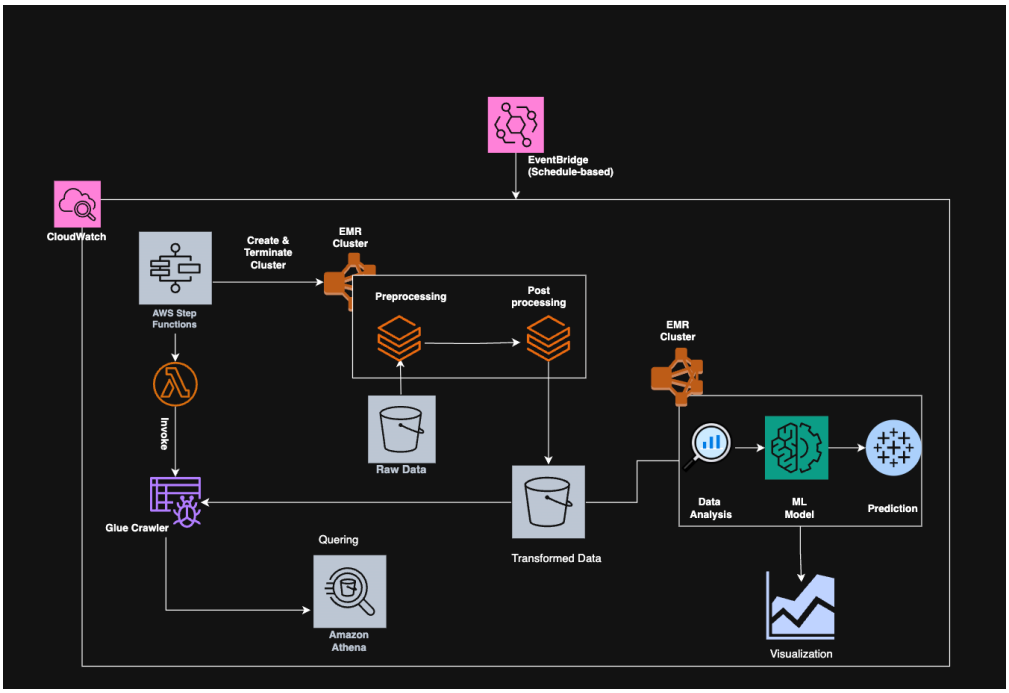
Initially, the dataset was loaded from an **AWS S3 input folder** in CSV format. Unnecessary placeholder columns (such as col0, col1, etc.) were removed. Meaningful column names like **State**, **Product Name**, **Prediction**, and **Total Amount Reimbursed** were retained and formatted for clarity. Data types were also cast appropriately - for example, monetary fields were converted to float for numerical operations. Next, rows with missing, incomplete, or suppressed values were filtered out to improve data quality. Aggregation was performed to summarize reimbursement amounts and drug prediction values by **State** and **Product Name**, providing an interpretable view of the data.

4. Visualizations and Insights

Architecture Diagram:

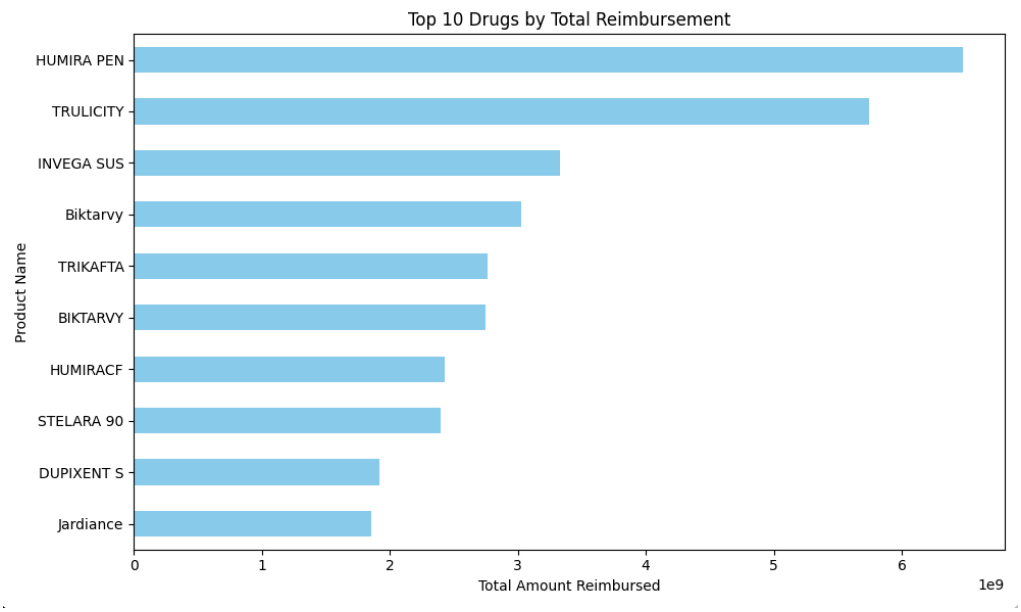
S3 output folder: s3://finalprojectankita/Input/SDUD2023.csv
S3 Output Folder : s3://finalprojectankita/Output/Data_output/





Plot 1: Top 10 Drugs by Total Reimbursement

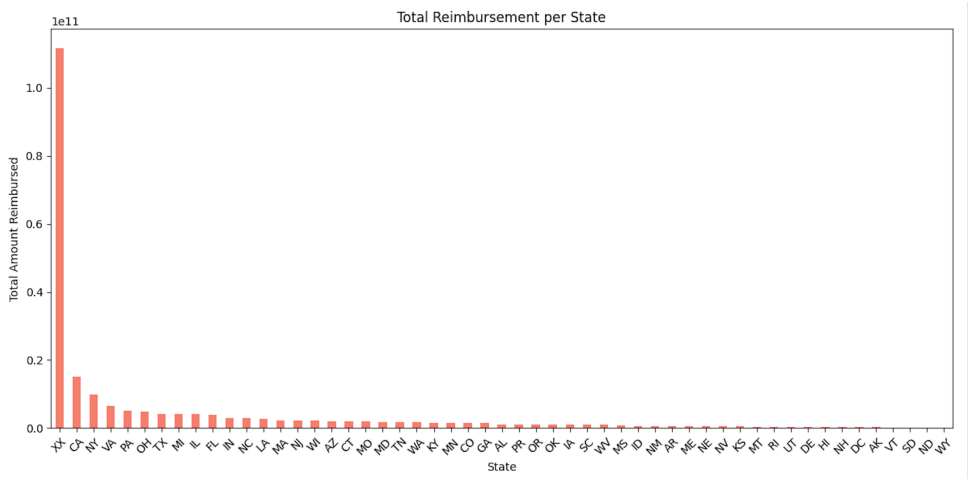
This horizontal bar chart shows the **top 10 most reimbursed drugs** across all states. The analysis revealed that drugs like **HUMIRA PEN**, **HUMIRACF**, and others dominate Medicaid spending. These drugs likely treat chronic or high-demand conditions, contributing to consistently high reimbursement totals.



Plot 2: Total Reimbursement by State

This vertical bar chart ranks states based on their total drug reimbursement amounts. States such as **California, Texas, and New York** (or as shown in your data, possibly **Alabama or Alaska**) appear to have the highest spending. This could be due to factors like state population, health policy, or drug coverage strategies.

Geographical differences in reimbursement can reveal regional healthcare trends and guide targeted interventions or audits.



Plot 3: Correlation Heatmap of Reimbursement Metrics

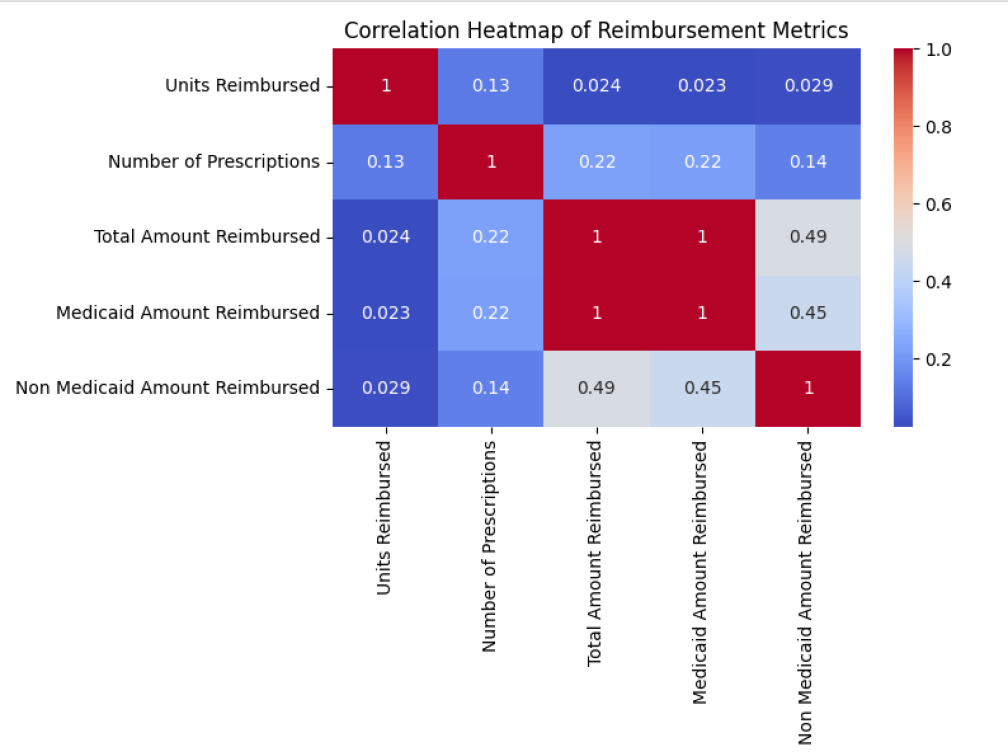
This heatmap displays the **correlation matrix among five key reimbursement-related metrics**:

- Units Reimbursed
- Number of Prescriptions
- Total Amount Reimbursed
- Medicaid Amount Reimbursed
- Non-Medicaid Amount Reimbursed

Notable observations:

1.A strong positive correlation between **Number of Prescriptions** and **Units Reimbursed**, indicating logical consistency. 2. High correlation between **Total** and **Medicaid Amount Reimbursed**, showing that Medicaid is the dominant contributor to total reimbursements.

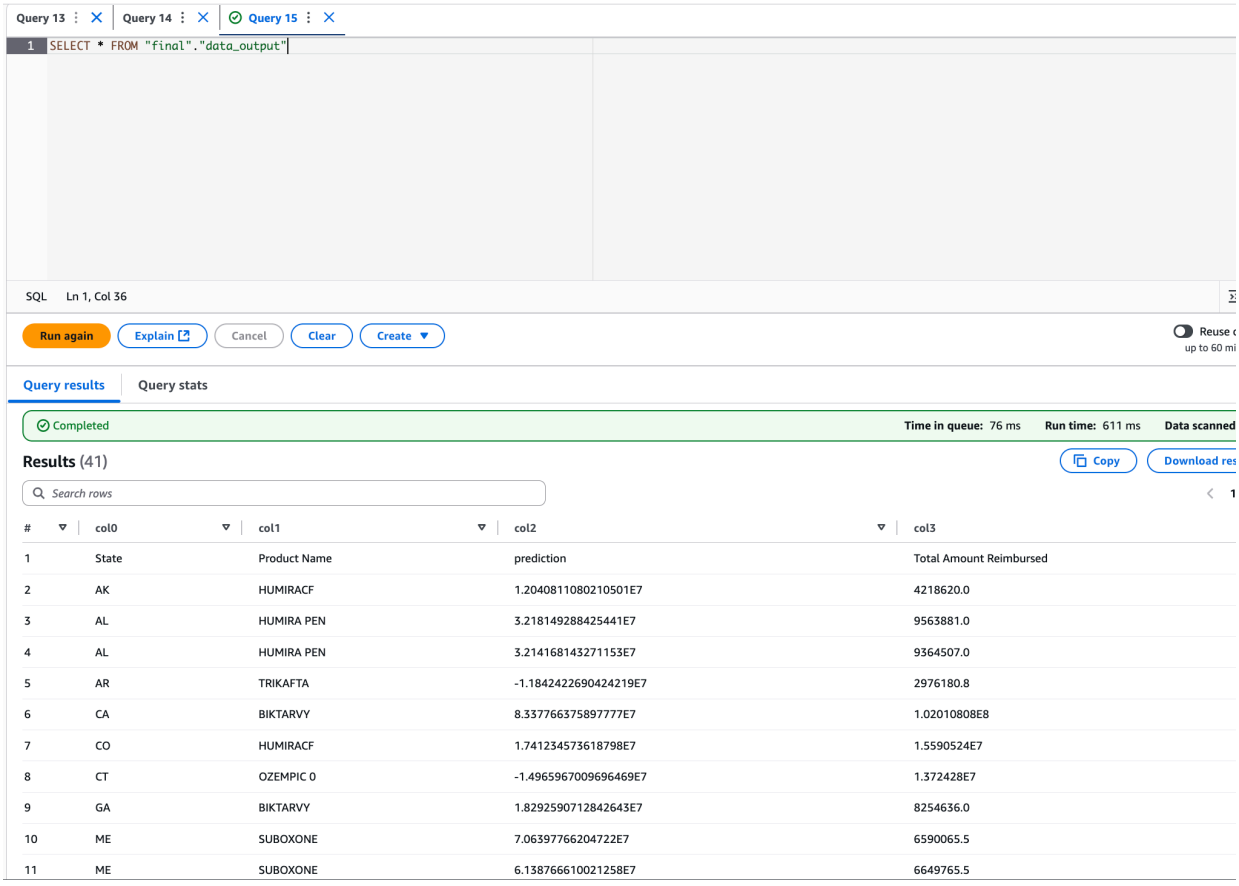
These relationships validate the dataset structure and provide confidence in the data's reliability for further modeling.



5. Machine Learning Model Development

To predict reimbursement amounts based on key drug-related features, multiple regression models were developed using Apache Spark's MLlib library, with *Total Amount Reimbursed* as the target variable. These models aimed to capture its relationship with features such as state, product name, and calculated metrics like predicted usage scores. Linear Regression was implemented using the LinearRegression class with hyperparameters like regParam, elasticNetParam, and number of iterations tuned via cross-validation. The Random Forest model used RandomForestRegressor, tuning parameters such as number of trees, maximum depth, and bins. Gradient Boosted Trees were built using GBTRegressor, optimized for iterations, step size, and tree depth. Additionally, Generalized Linear Regression was applied through the GeneralizedLinearRegression class, with configurations for family (e.g., gaussian), link function, regularization, and maximum iterations. All models were evaluated using metrics such as RMSE and R^2 to assess predictive accuracy, providing insights into Medicaid reimbursement trends and enabling more informed forecasting and policy decisions.

Athena Screenshot :



Query 13 | Query 14 | Query 15

```
1 SELECT * FROM "final"."data_output"
```

SQL Ln 1, Col 36

Run again | Explain | Cancel | Clear | Create

Query results | Query stats

Completed — Time in queue: 76 ms — Run time: 611 ms — Data scanned

Results (41) — Copy — Download res

Search rows

#	col0	col1	col2	col3
1	State	Product Name	prediction	Total Amount Reimbursed
2	AK	HUMIRACF	1.2040811080210501E7	4218620.0
3	AL	HUMIRA PEN	3.218149288425441E7	9563881.0
4	AL	HUMIRA PEN	3.214168143271153E7	9364507.0
5	AR	TRIKAFTA	-1.1842422690424219E7	2976180.8
6	CA	BIKTARVY	8.337766375897777E7	1.02010808E8
7	CO	HUMIRACF	1.741234573618798E7	1.5590524E7
8	CT	OZEMPIC 0	-1.4965967009696469E7	1.372428E7
9	GA	BIKTARVY	1.8292590712842643E7	8254636.0
10	ME	SUBOXONE	7.06397766204722E7	6590065.5
11	ME	SUBOXONE	6.138766610021258E7	6649765.5

6. Challenges and Solutions

A key challenge in this project was efficiently handling large-scale data, as tasks like preprocessing, EDA, and model training became time- and resource-intensive. I also attempted to implement a Step Function for workflow orchestration but faced issues, making it a major challenge and future focus. To overcome data processing bottlenecks, distributed computing was used to parallelize tasks, leveraging AWS services:

- **Amazon EMR** for scalable Spark-based processing,
- **Amazon Athena** for fast, serverless data querying.

These tools improved processing speed, reduced bottlenecks, and streamlined the workflow.

7. Conclusion and Future Directions

This project examined how weather affects taxi demand in Chicago, developing predictive models based on weather variables. Future improvements include adding features such as time, events, and traffic conditions, experimenting with advanced algorithms and hyperparameter tuning, and expanding the dataset to capture more weather variations over a longer period for stronger predictions.