Predictive Analysis on Medical Expenditure

FOR CHRONIC HEALTH DISEASE TREATMENT IN US (SOUTH REGION)

Prepared by Group 3

Aafrin Shehnaz Mohamed Sulaiman, Aiswarya Raghavadesikan, Aryama Ray, Dongmei Han

November 23, 2024

Introduction

Our project aims to **analyze the cost drivers** behind the hefty expenses of chronic health conditions like Blood Pressure, Arthritis, Cardio vascular disease and Cancer etc., to understand **how the healthcare resources are distributed** and utilized and how effectively the **policies are supporting patients** in managing their conditions. The Medical Expenditure Panel Survey (**MEPS**) has done a extensive research on this and provided **data with respect to various** demographics, insurance plans, treatment services related with these conditions.

Data processing

Original Dataset

2022: 22432 * 1420

2021: 28337 * 1488

2020: 27805 * 1452

Cleaned Dataset

15549 * 62

Input Datasets

Years considered: **2020**, **2021** and **2022**

Data cleaning

- Standardize and rename columns
- Handle negative values

Merge & Filter

- Merged datasets
- Filtered individuals with chronic diseases

Feature Selection

Select relevant features:

Demographics, Geography, Diseases, Expenses, Payment through Insurances

Handle Missing Data

- **Impute** missing values with **subgrouping and based on corresponding Median and Mode values**

Outlier Removal

- Applied outlier removal on the target variable using Z-score

Feature Engineering

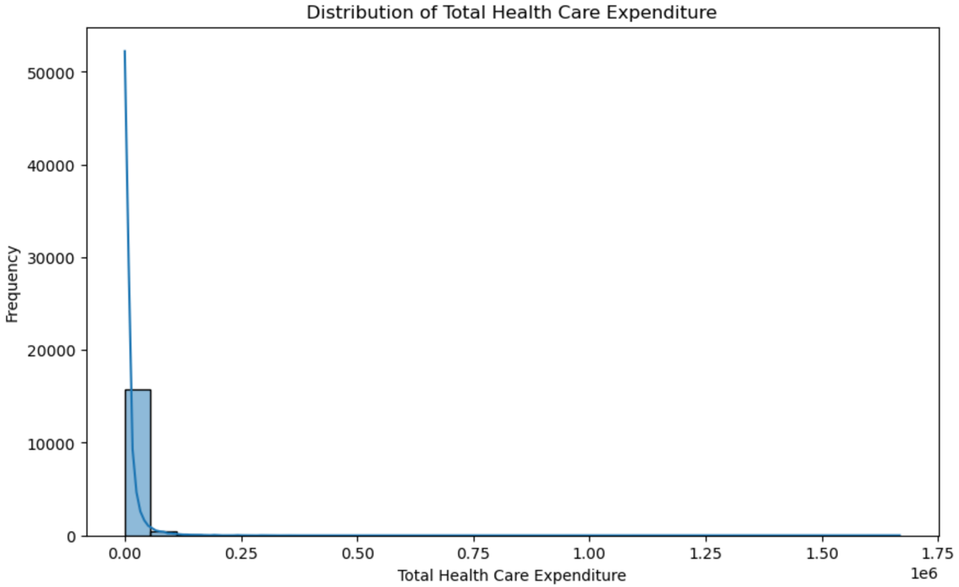
- Derived features:
- Sub-total Expenditure
- Critical Comorbidity

Output

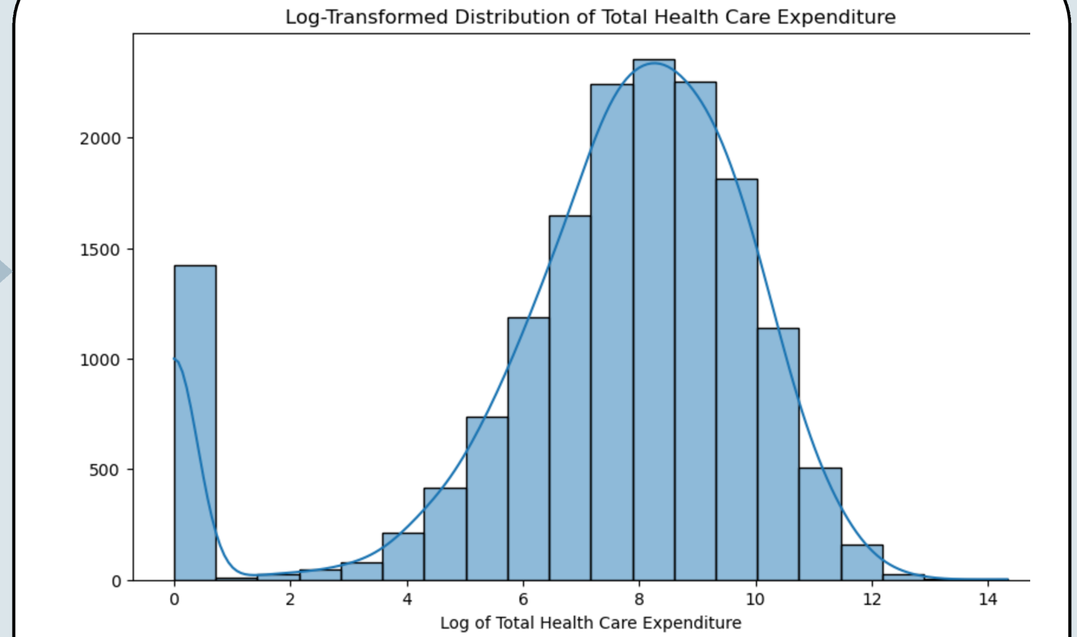
- Final preprocessed dataset for further analysis

EDA on Target Variable

Distrubution of **Total Health Care Expenditure**



Log transform



Original:

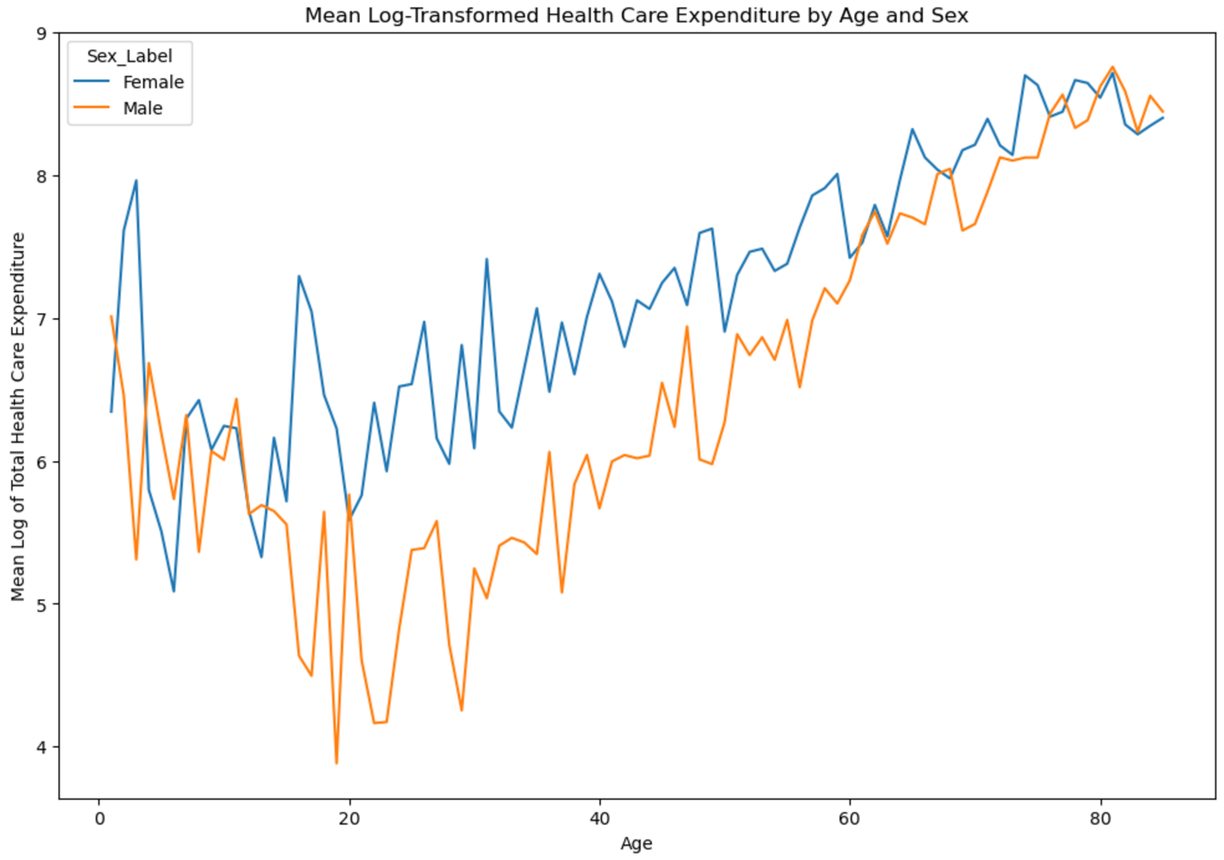
- Highly right skewed
- Hard to identify patterns

Log-transformed:

- Basically follows normal distribution, with outliers exist
- Remove outliers using Z-score technique

EDA on Demographic features

Total Health care expenditure distribution by Age and Sex



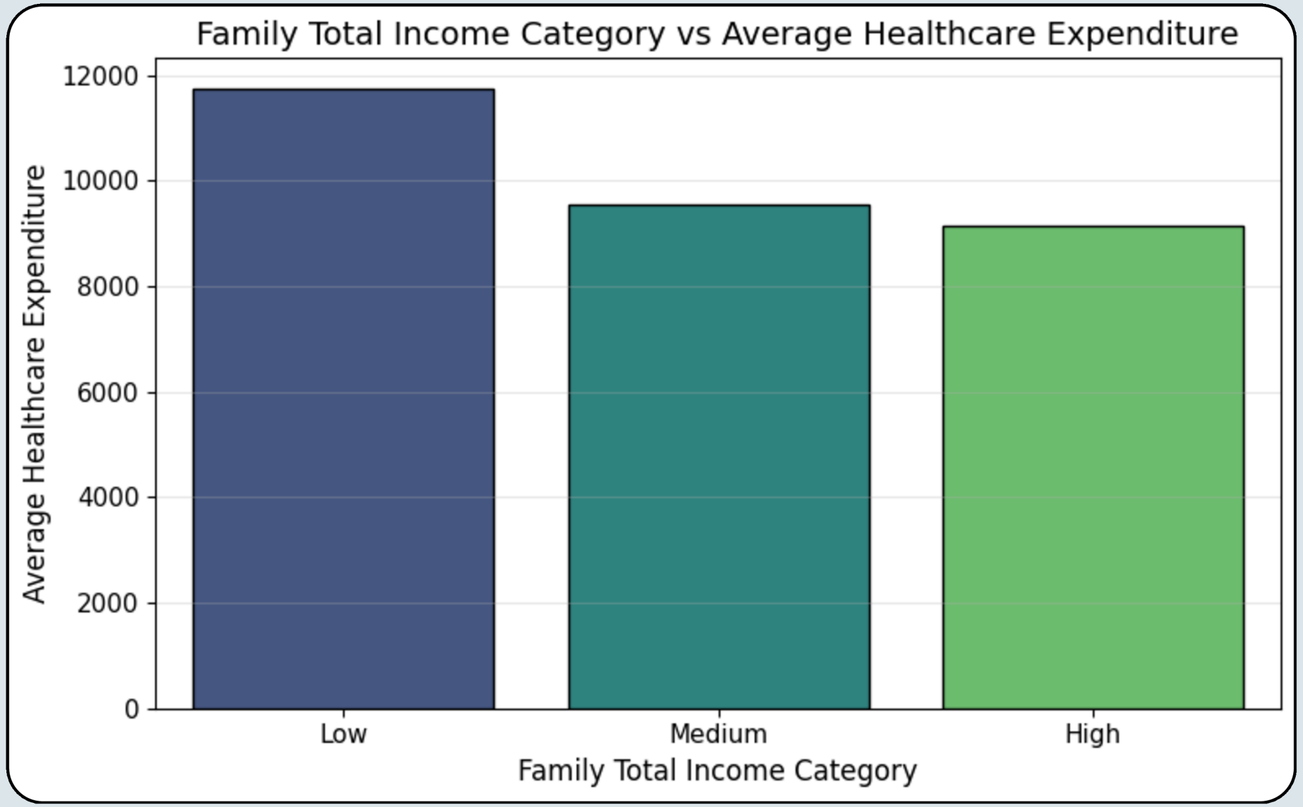
Age and Gender Impact:

- Expenditure **increases with age**, surging after 40.
- **Females** spend consistently more than males, likely due to gender-specific needs.

Unexpected Variability:

- **High fluctuations in male expenditure under 30** suggest a subset with chronic conditions driving costs.

EDA on Demographic features



Total Health care expenditure distribution by **Family Total Income**

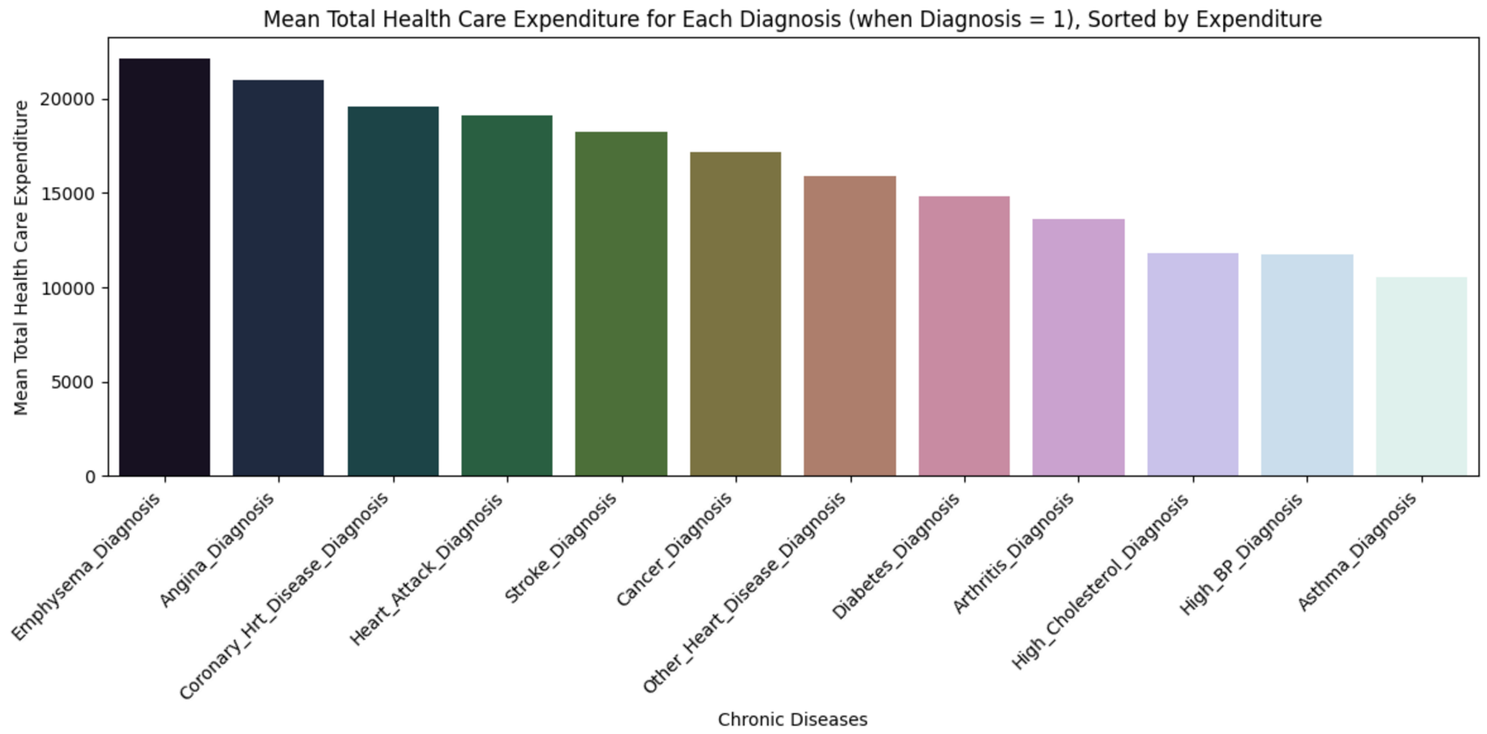
Income categories based on distribution:

- Low: below Q1
- Medium: Between Q1 and Q3
- High: Above Q3

Key Insights:

- Lower-income families spend proportionally more on healthcare.
- Likely due to less preventive care or greater out-of-pocket reliance.
- Highlights disparities in financial healthcare burden.

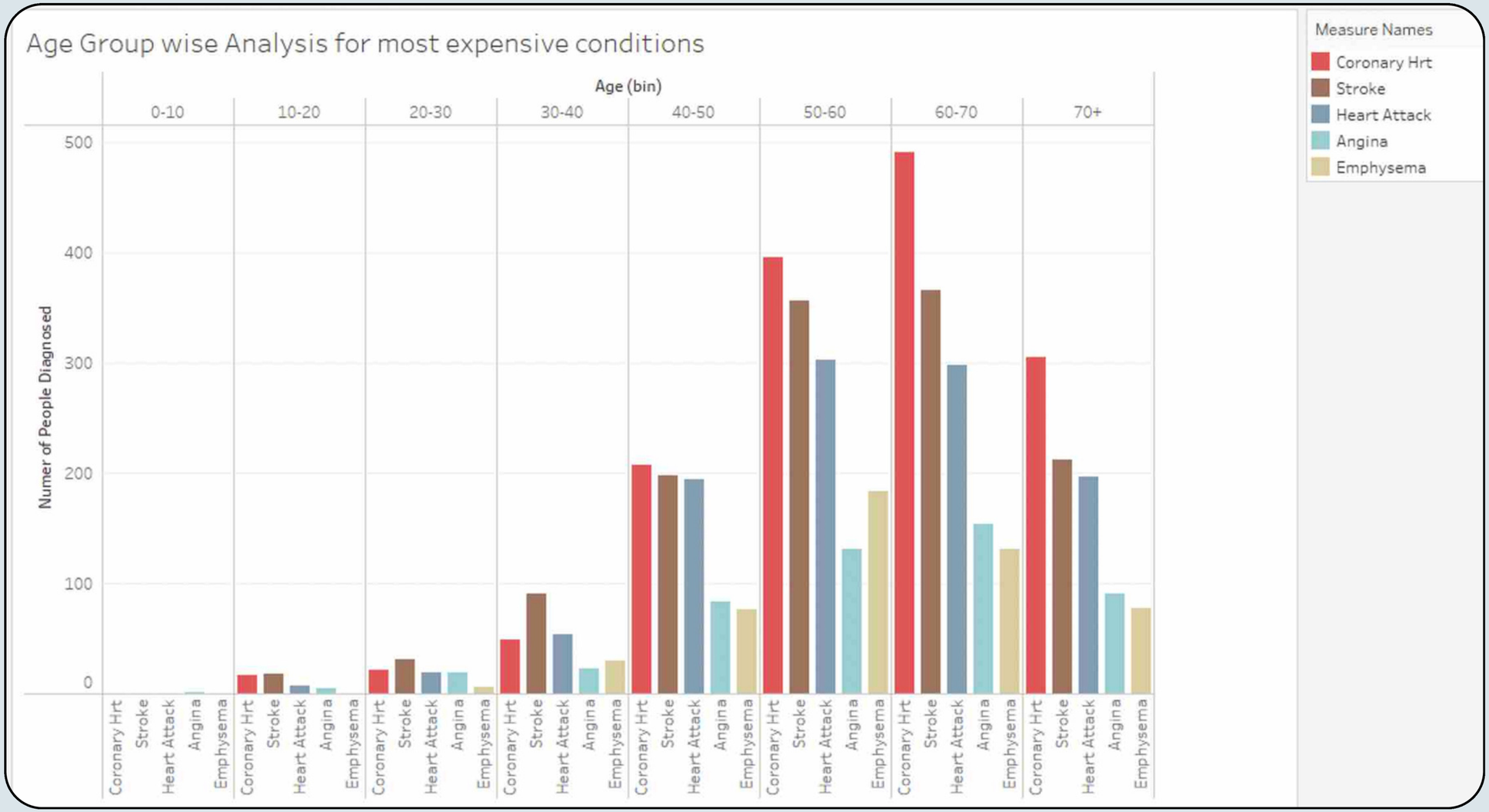
EDA on Demographic features



Since **2020** to **2022,** Patients diagnosed with **Emphysema, Angina,** and **Coronary heart disease** incurred the **highest average treatment costs** in the **Southern U.S.**

t

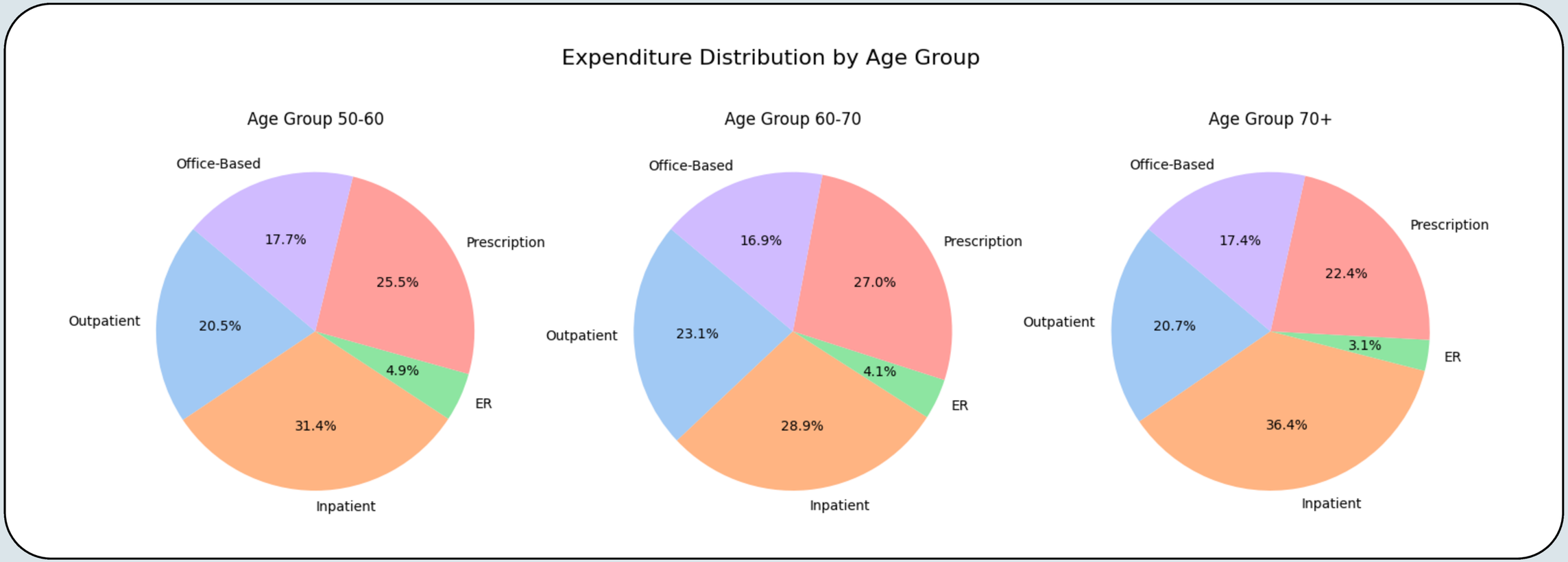
EDA on Medical Expenses



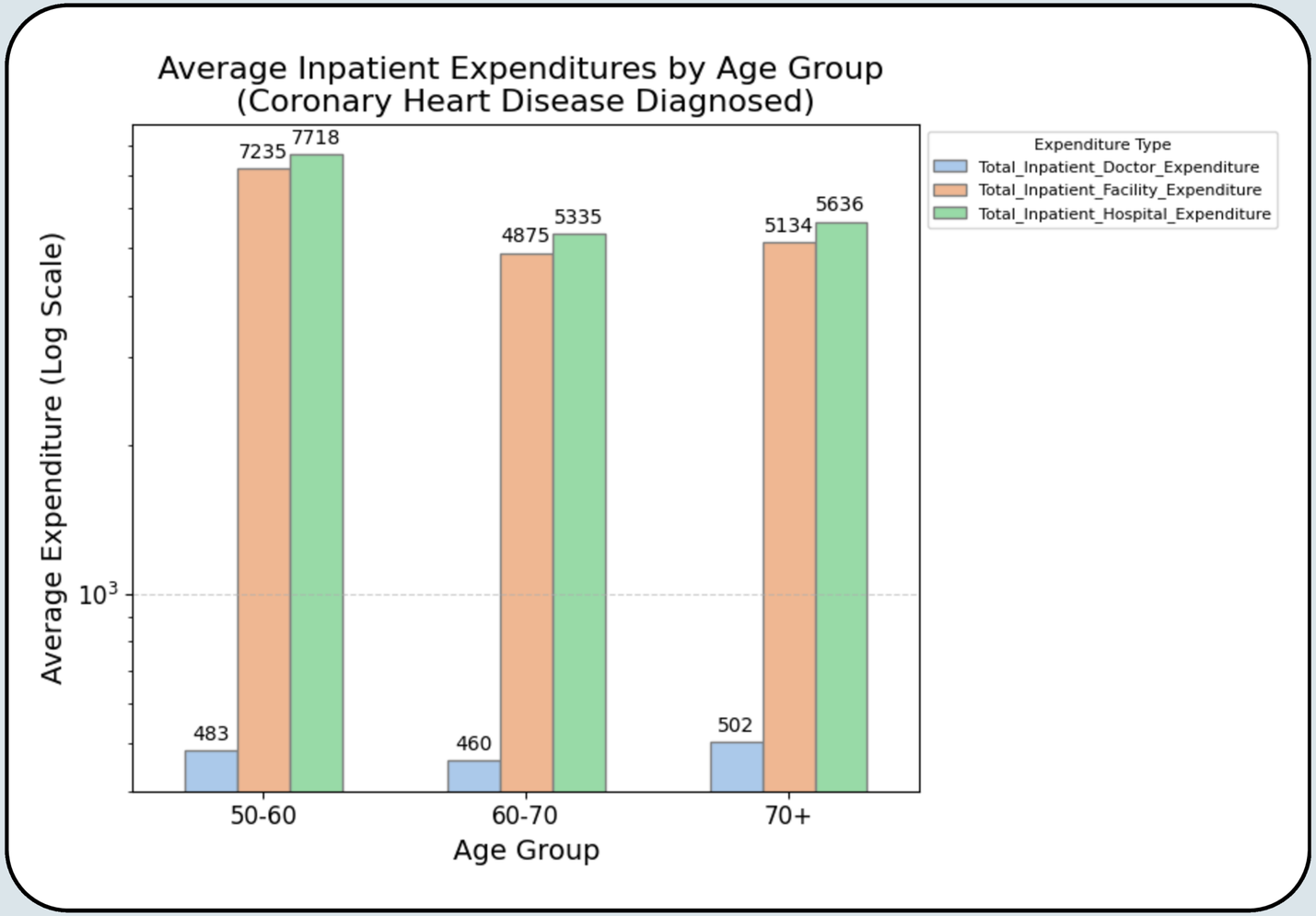
- Individuals aged 50 and above represent the most vulnerable segment of the population.
- The 50-70+ age group predominantly experiences Coronary Heart Disease, with Stroke being the second most common condition.

EDA on Medical Expenses

- For individuals aged 50 and above, medical expenses are primarily driven by inpatient care, outpatient services, and prescription costs.



EDA on Medical Expenses

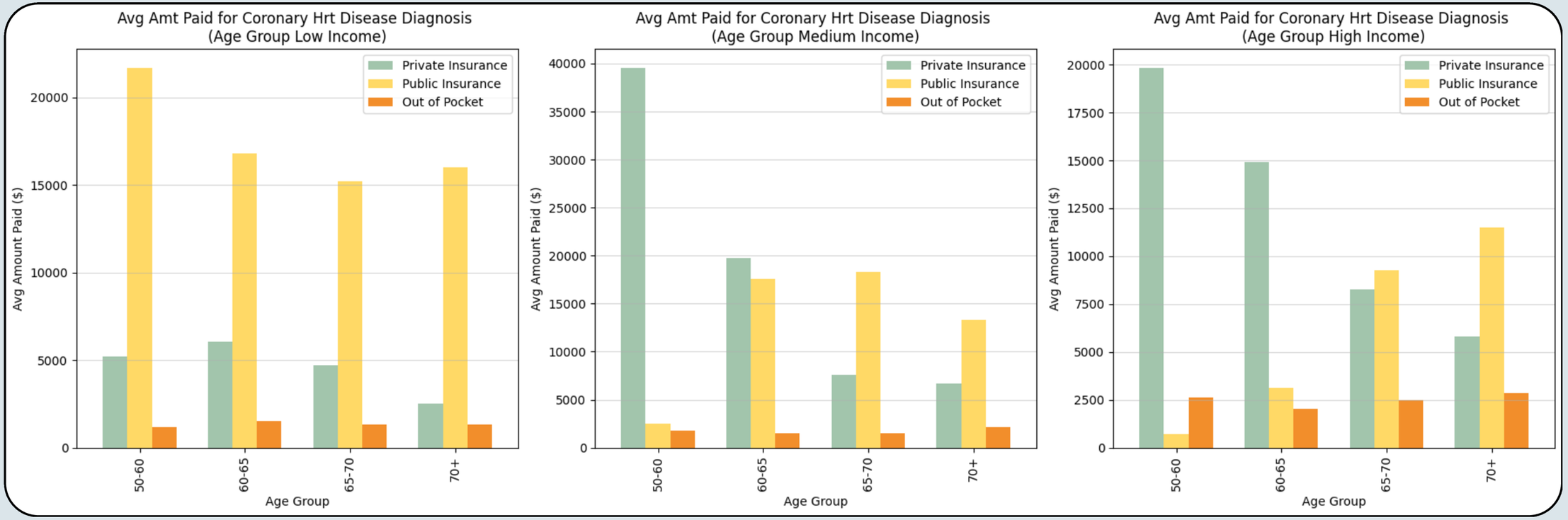


- Conducted an analysis of inpatient expenses for individuals aged 50 and above diagnosed with coronary heart disease.

- Findings reveal that for individuals over 50, hospital and facility charges constitute the largest portion of inpatient care expenses, significantly exceeding doctor fees.

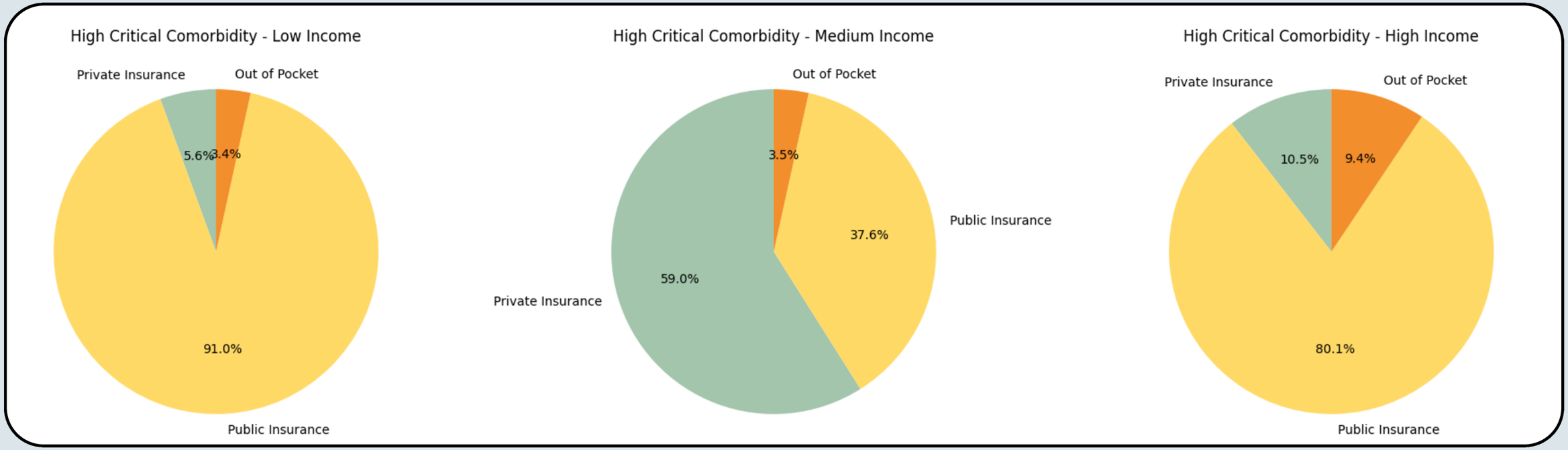
EDA on Payment Process for Medical Expenses

- Conducted an analysis of payment methods for individuals aged 50 and above across various family income groups.
- For the 50-60 age group with coronary heart disease, those from medium- and high-income families primarily rely on private insurance for medical expenses.

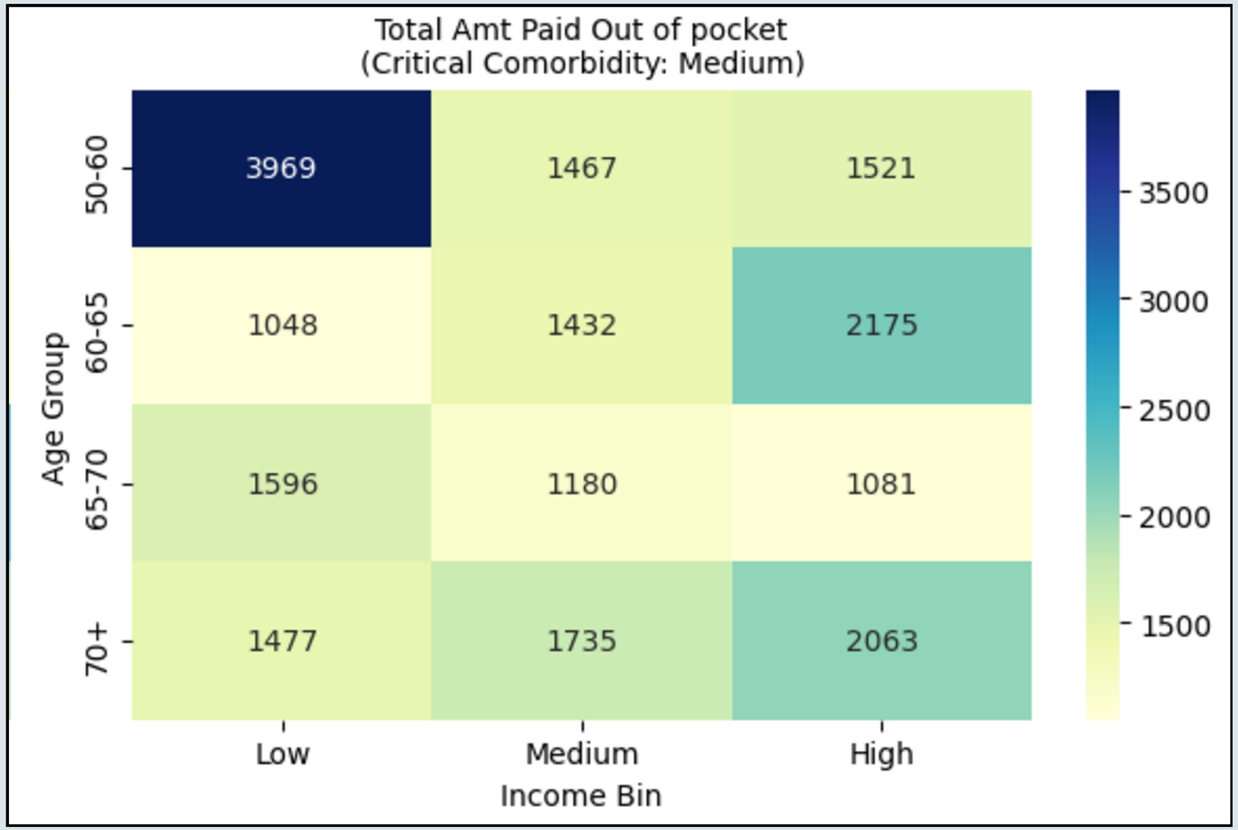


EDA on Payment Methods for High Comorbidity Expenses by Income

- Public insurance plays a significant role in covering healthcare costs across all income levels, especially for individuals from low-income and high-income family
- Private insurance covers majority(59%) of the medical expenses for individuals from medium-income family group compared to other two income groups, suggesting employer-sponsored insurance is a key driver.
- While out-of-pocket expenses are minimal for low and medium-income groups, they increase for high-income individuals.

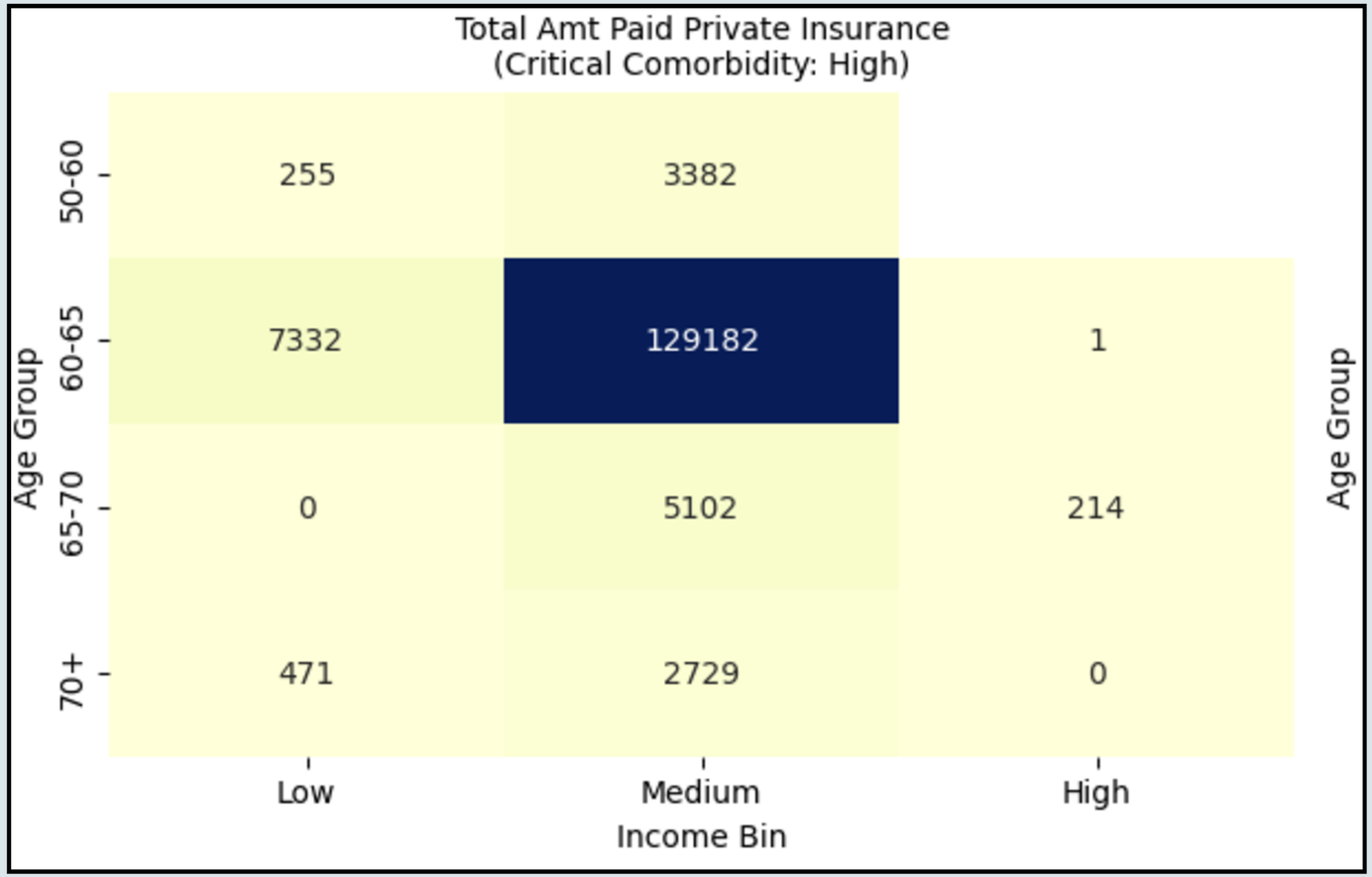


Anomaly in Payment Distribution : Out of Pocket Payment



- Individuals aged 50 to 60 from low-income families bear the highest out-of-pocket costs when managing atleast two critical chronic diseases.

Anomaly in Payment Distribution : Private Insurance



- Individuals aged 60 to 65 from medium-income families typically pay an average of $129,182 for medical expenses through private insurance when managing more than three critical chronic diseases.

- This amount is excessively high compared to other income groups.

Models proposed

Linear Regression

- Linear Regression with Lasso Regularization (L1)
- Linear Regression with Ridge Regularization (L2)
- Lasso Regularization with Hyperparameter tuning
- Ridge Regularization with Hyperparameter tuning

Random Forest

- Random Forest Baseline Model
- Random Forest with Hyperparameter tuning

Support vector Regressor

- SVR Baseline Model
- SVR with Hyperparameter Tuning

Gradient Boosting

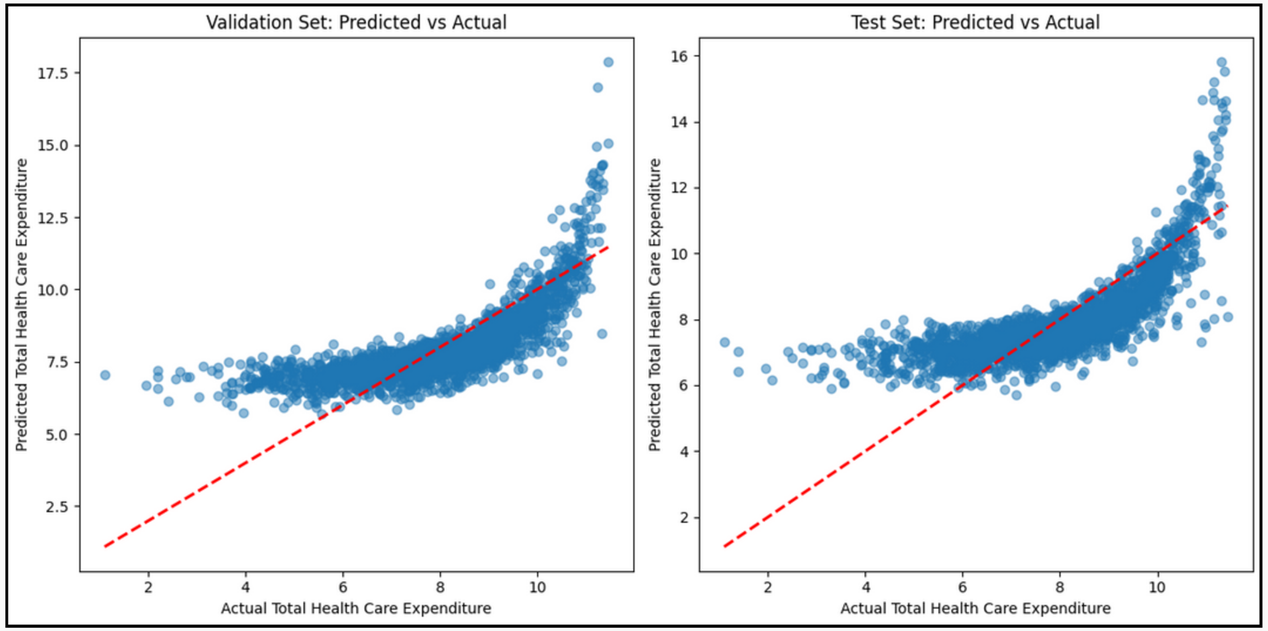
- Gradient Boosting Baseline Model
- Gradient Boosting with Hyperparameter tuning

XG Boost

Linear Regression Models and Evaluation

- Performed Basic Linear Regression, L1 and L2 Regularization, hyperparameter tuned linear regression.
- Split the dataset into training, validation, and test sets in a **70%-15%-15%** ratio
- Performed StandardScalar for standardizing features and removed multi-colinearity using **Variance Inflation Factor (VIF).**

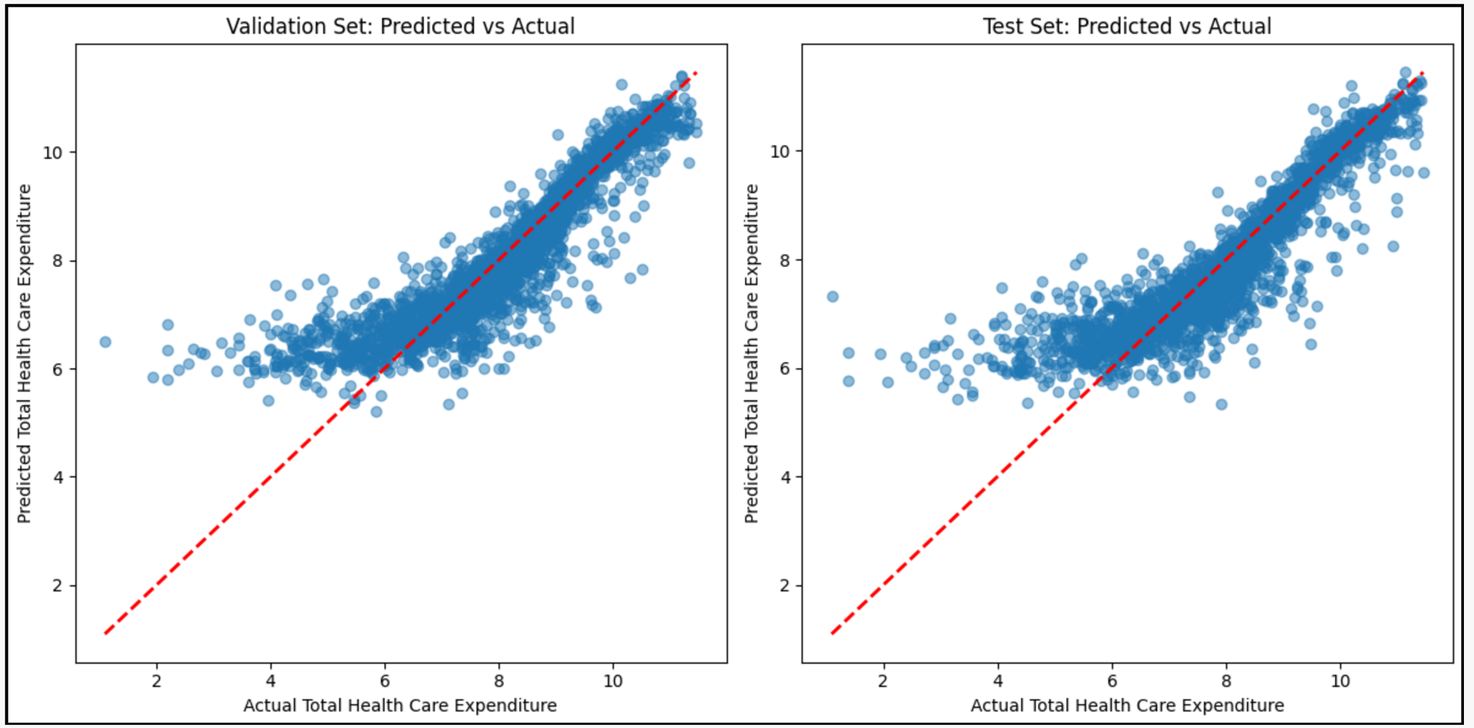
	Validation Set	Test Set
Base Linear Regression	MAE: 0.83 RMSE: 1.10 R-squared: 0.57	MAE: 0.84 RMSE: 1.11 R-squared: 0.56
Lasso	MAE: 0.83 RMSE: 1.10 R-squared: 0.57	MAE: 0.84 RMSE: 1.11 R-squared: 0.56
Lasso with GridSearchCV	MAE: 0.83 RMSE: 1.10 R-squared: 0.57	MAE: 0.84 RMSE: 1.11 R-squared: 0.56
Ridge	MAE: 0.82 RMSE: 1.09 R-squared: 0.57	MAE: 0.84 RMSE: 1.11 R-squared: 0.56
Ridge with GridSearchCV	MAE: 0.82 RMSE: 1.10 R-squared: 0.57	MAE: 0.84 RMSE: 1.11 R-squared: 0.56



- Base Linear regression model and linear regresion models with regularization could not achieve high accuracy on the current dataset.
- Curved pattern suggests these models could not capture the non-linear relationship between features and the target variable

Support Vector Regression Model and Evaluation

- SVR works well for **high dimensional feature space.**
- Model performed better than Linear regression for the majority of the data, especially for mid-range expenditures.
- With GridSearchCV SVR performed similar way with best Parameters: ‘C’: 100, 'epsilon': 0.1, 'gamma': 'scale', 'kernel': 'rbf'



	Validation Set	Test Set
Base SVR	MAE: 0.50 RMSE: 0.77 R-squared: 0.79	MAE: 0.50 RMSE: 0.78 R-squared: 0.78
SVR with GridSearchCV	MAE: 0.53 RMSE: 0.77 R-squared: 0.79	MAE: 0.56 RMSE: 0.79 R-squared: 0.78

Decision tree models

Random Forest

- **Robust to noisy data** and overfitting.
- Parameters:

n_estimators: 200,
min_samples_split: 2,
min_samples_leaf: 1,
max_features: 'sqrt',
max_depth: 20

Gradient Boosting

- Performs **better with large datasets** and complex data distributions.
- Parameters:

n_estimators: [200, 300]
learning_rate: [0.1, 0.2]
max_depth: [4, 5],
min_samples_split: [2, 5]
min_samples_leaf: [1, 2]

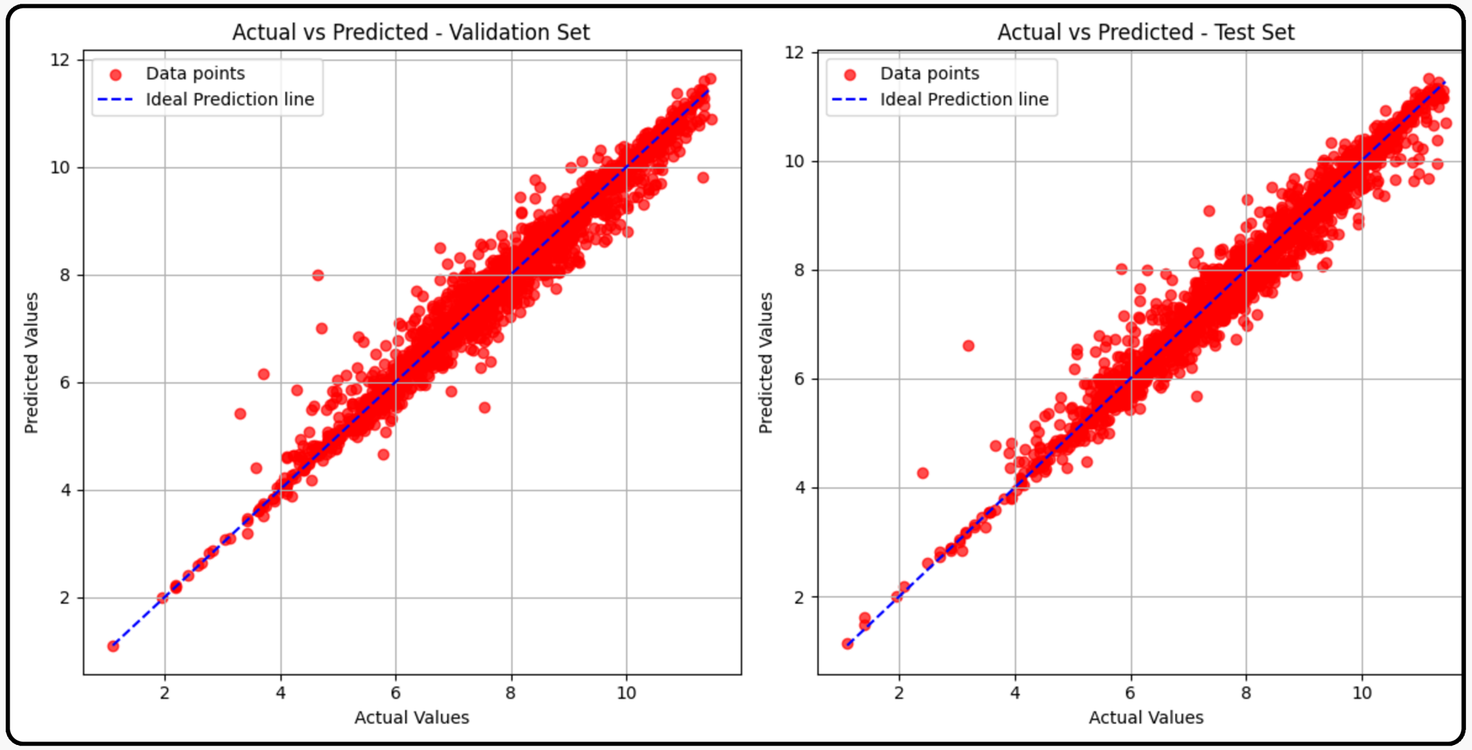
XG Boost

- Chosen for its **best predictive performance**, ability to handle non-linear relationships.
- Parameters:

n_estimators=100,
max_depth=6,
learning_rate=0.1,
random_state=42

Models/Metrics	Valid RMSE	Test RMSE	Difference	Test MAE	R-squared
Random Forest	0.36	0.37	0.01	0.24	0.95
Random Forest with Hyperparameter tuning	0.41	0.42	0.01	0.29	0.94
Gradient Boosting	0.37	0.38	0.01	0.27	0.95
Gradient Boosting with Hyperparameter tuning	**0.34**	**0.33**	**0.01**	**0.22**	**0.96**
XG Boost	0.28	0.34	0.06	0.23	0.96

Gradient Boosting Regressor (Hyperparameter Model)



Generalization:

- The consistency across the validation and test sets indicates the model is **generalizing well without overfitting** the training data.

Error Distribution:

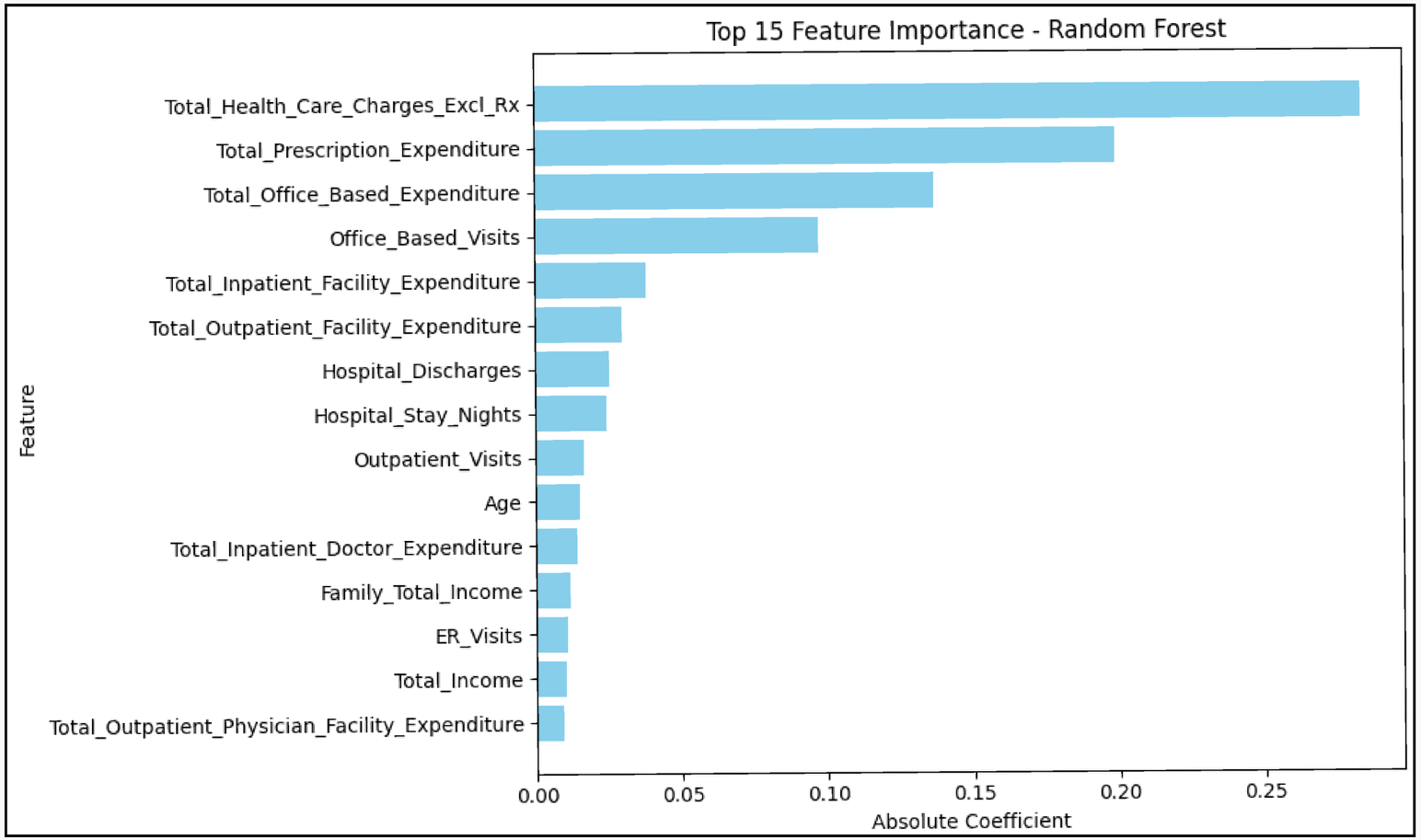
- Any deviations or scatter around the ideal prediction line represent areas where the model struggled slightly to predict accurately.

Bias vs. Variance:

- There is **no significant bias** (systematic underprediction or overprediction) in the predictions.

Feature impotance from Gradient Boosting Regressor (Hyperparameter Model)

- Total_**Health_Care_Charges**_Excl_Rx and Total_**Prescription_Expenditure**, which dominate the model, indicating that medication-related expenses are the **most significant** drivers of healthcare costs.
- Total_**Office_Based**_Expenditure and Office_Based_Visits, which also contribute significantly, reflecting the importance of **outpatient services**.
- While medication and outpatient services dominate, factors like age and family income also contribute to healthcare cost expenditure, consistent with EDA findings and reflecting the influence of demographic and socioeconomic factors.



Community Contributions

Improved Decision-Making in Healthcare:

- The model provides accurate predictions for Total Health Care Expenditure, which can help policymakers and healthcare providers **allocate resources more efficiently**, especially for individuals with chronic conditions.

Financial Planning for Families:

- The model offers insights that individuals and families can use for **financial planning**, particularly those managing chronic health conditions.

Conclusion

By implementing **Data mining** techniques and **Machine learning models**, the project effectively handles non-linear relationships between the features. **Key drivers** of healthcare costs identified include **income levels**, **insurance coverage**, and **comorbidity counts,** providing **actionable insights** for policymakers, healthcare providers, and insurers.

Thank you