Analyzing Variations in Healthcare Plan Premiums, Benefits, and Regional Disparities in the United States

Akshaya Thangavel, Daksesh Kasumurthi, Imran Nawaz Shareef,

Vamshi Krishna Bairoju, Vinila Chowdary Potla

University of North Texas

Contents

List of Figures

Acknowledgment

We want to thank our professor Dr. Sameh Shamroukh, for their constant support and guidance throughout this project. Their advice and encouragement were key in helping us understand the concepts and improve our approach to studying healthcare plan premiums and variations.

We are also grateful to the University of North Texas for providing the tools, resources, and a great learning environment that made this research possible. A special thanks to our group members—Akshaya Thangavel, Daksesh Kasumurthi, Imran Nawaz Shareef, Vamshi Krishna Bairoju, and Vinila Chowdary Potla—for their hard work, teamwork, and dedication to completing this project.

This research involved using different techniques, like cleaning data, applying machine learning models, and conducting statistical analysis, to study and predict healthcare premium variations. The combined effort and collaboration of the group were essential in overcoming challenges and achieving our objectives.

Lastly, we appreciate the public datasets and open-source tools that played a big role in our research. This project would not have been possible without their contribution to the research community.

Abstract

This project looks at how healthcare plan premiums, benefits, and coverage differ across the United States. Using data from healthcare.gov, we studied how factors like metal levels (Bronze, Silver, Gold, Platinum), plan types (HMO, PPO), and geographic locations impact costs and coverage. To do this, we cleaned the data, used statistical methods, and applied machine learning models, such as Random Forest and XGBoost, to predict healthcare premiums based on plan features and regional factors.

Our findings show clear differences in premiums and coverage based on these factors. For instance, Platinum plans have the highest premiums, while bronze plans are the most affordable. Regional differences, confirmed by statistical tests, highlight that location plays a big role in healthcare affordability. Our predictive models were effective, with the Random Forest model explaining 90% of the variability in premiums.

This research is designed to help consumers make better decisions when choosing healthcare plans, assist insurance companies in improving their offerings, and guide policymakers in addressing regional inequalities in healthcare. By identifying these patterns, we aim to contribute to a better understanding of the healthcare insurance market and its challenges.

Introduction

The United States healthcare insurance sector is intricate and multifaceted where the clients are provided with so many choices of plans, coverage, and premiums. Plans are structured in metal levels consisting of Bronze, Silver, Gold, and Platinum having different cost sharing and benefits to the consumers. Besides these levels, insurance policies are also classified according to types, for instance, Health Maintenance Organizations (HMOs), Preferred Provider Organizations (PPOs) and so on. However, a diverse selection of this breadth is designed to address the needs and circumstances of many consumers; most of the consumers are left bewildered regarding the plan that best fits their specific or family needs out of the numerous options available.

One of the major difficulties in looking for a healthcare plan is linking the plan type to the premiums and benefits offered and its geographical distribution. Other aspects such as the residing state and county have a substantial impact on the access and cost of plans thus creating inequities in the healthcare system. In as much as these variations are detrimental to the ability of the consumers to search for affordable coverage, they are equally disturbing to the policymakers and the insurers who are wishing to provide healthcare that is equitable nationwide.

This project intends to examine the variations in premiums and benefits across healthcare plans about plan type, metal tier, and geographical location. By highlighting trends and inequities, we aspire to enable consumers to avoid making errors in decision making, help insurers in modifying their plans, and notify regulators about areas which may need regulatory controls.

Research Questions

1. How do healthcare plan premiums vary across different metal levels and plan types (PPO, HMO)

Metal Level	Plan Type	Premium Adult Individual Age 21	Premium Adult Individual Age 30	Premium Adult Individual Age 40	Premium Adult Individual Age 50	Premium Adult Individual Age 60
Bronze	EPO	211.939800	240.559415	270.865649	378.534201	575.219415
	HMO	279.425974	319.151973	358.682792	501.725217	760.596391
	POS	212.708805	242.214684	272.467630	380.959430	578.179056
	PPO	202.921460	230.418953	259.423812	362.553979	550.866199
Catastrophic	EPO	183.155125	207.881253	234.072574	327.115923	497.082096
	HMO	254.462457	290.020797	326.158189	456.087861	691.943181
	POS	149.322783	170.212654	191.416385	267.676772	406.086212
	PPO	168.561764	191.641878	215.673857	301.463167	457.829239
Gold	EPO	280.108935	317.931531	357.987226	500.285547	760.233040
	HMO	319.917374	366.326888	411.389214	575.668752	871.844834
	POS	293.207470	333.421087	375.224764	524.525307	796.483682
	PPO	281.322694	319.438831	359.645657	502.635301	763.685420
Platinum	EPO	302.335754	343.168784	386.402075	539.994571	820.574765
	HMO	314.972813	357.623727	402.639972	562.714388	854.998805
	POS	314.959477	357.463140	402.516953	562.513271	854.797888
	PPO	293.190108	332.767082	374.696894	523.636732	795.721253
Silver	EPO	247.716135	281.168699	316.590819	442.435591	672.323731
	HMO	264.574746	303.778428	340.875707	477.190913	721.953045
	POS	252.468309	287.006033	323.021812	451.529935	685.720334
	PPO	240.150321	272.653922	306.975956	429.018316	651.851563

Fig 1: Average premium rate table by Metal Level, Plan Type and Age Group

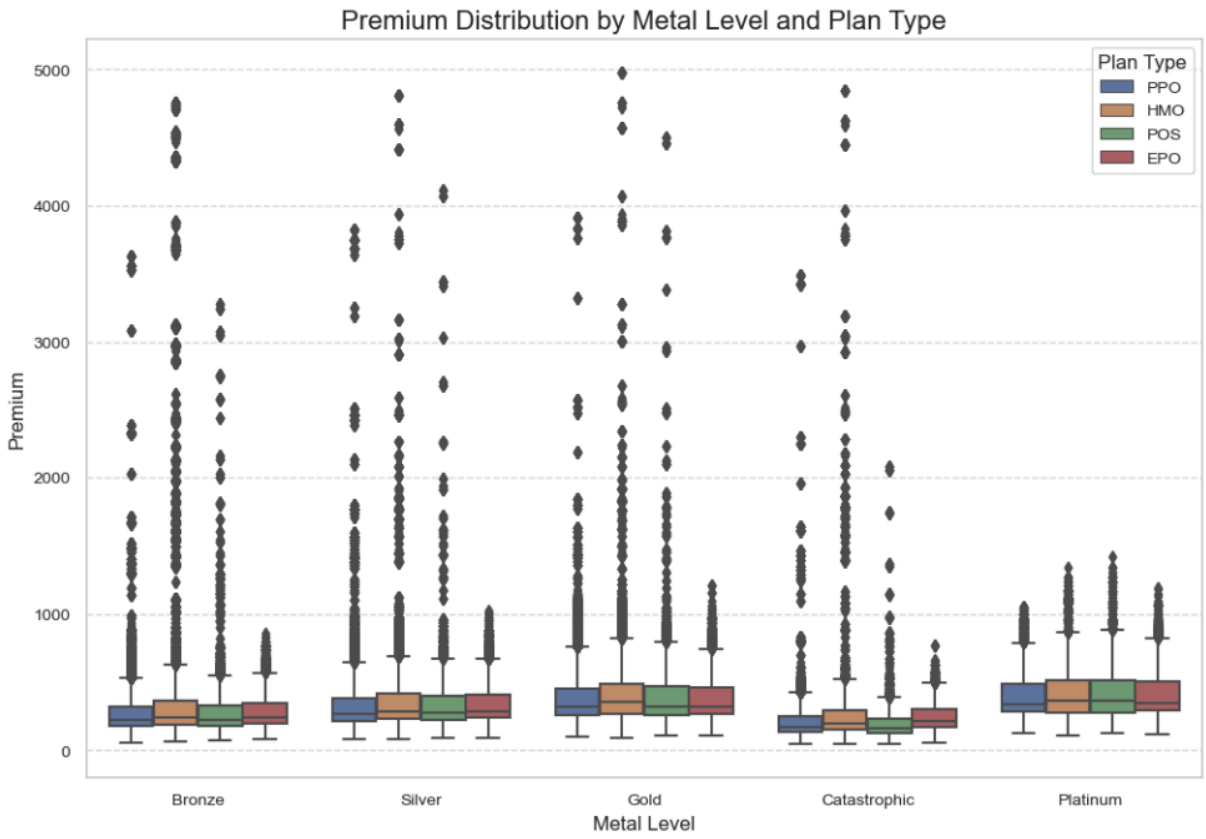


Fig 2: Premium Distribution by Metal Level and Plan Type

Observing the above plot, Platinum plans have the Highest premiums, while the premiums in the bronze level are the most affordable. Premium Increases as the metal level becomes more comprehensive. Across all the metal levels, plan types such as PPO and HMO exhibits the smaller trends, but PPO plans tend to have higher premiums in certain categories.

Model Performance:

The Random Forest model was applied to determine the effect of metal levels and plan types on healthcare premiums. Because of its effectiveness in representing non-linear relationships and interactions between various categorical variables, in this instance between plan type (PPO, HMO) and premium cost for different metal levels (Bronze, Silver, etc.), Random Forest is useful in this instance. Thanks to the aggregation of the results generated by several decision trees, Random Forest gives reliable and accurate results, reduces the effect of overfitting, and enhances predictive strength.

```
Plan Type Prediction:
Random Forest - MAE: 0.19216883461640566, RMSE: 0.5111523715053614, R-squared: 0.8990430464941704
```

The Random Forest model achieved an R^2 of 0.899, indicating that metal levels and plan types explain nearly 90% of the variability in premiums. This confirms that these factors are strong predictors of healthcare costs. The low MAE (0.192) and RMSE (0.511) further highlight the model's accuracy. The results validate observations from the EDA, such as higher premiums for Platinum plans and PPO types and provide a quantitative framework to predict premiums based on plan characteristics.

2. Are there regional differences by state in the types of healthcare plans offered and their affordability?

The **map visualization** of average premiums across states clearly highlights regional disparities. States such as **Virginia (VA)** show significantly higher average premiums compared to others.

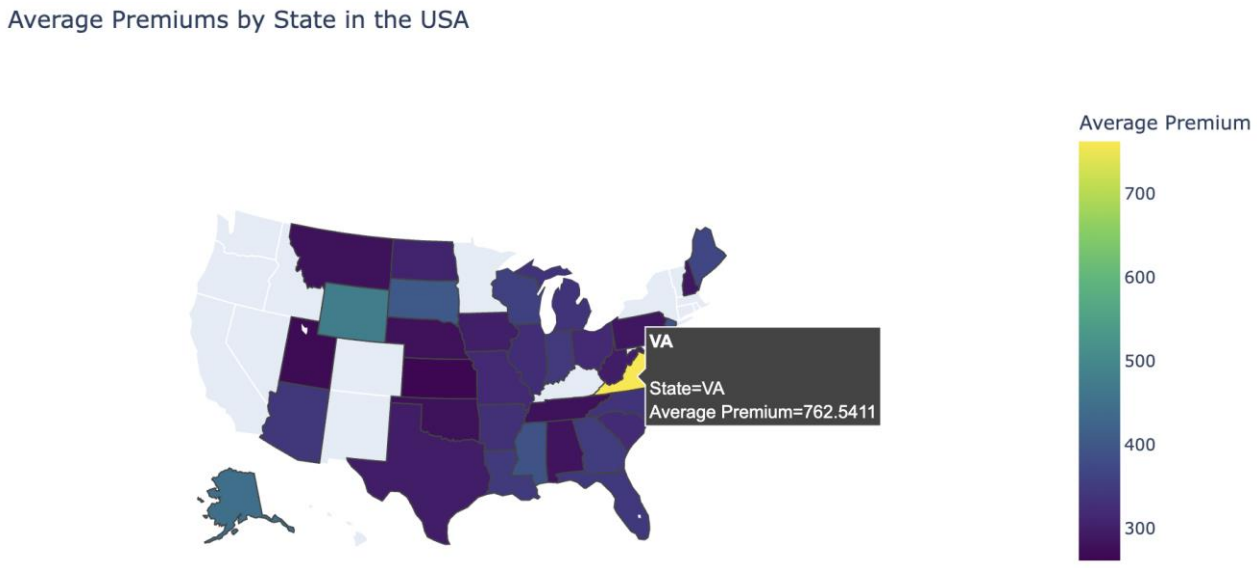
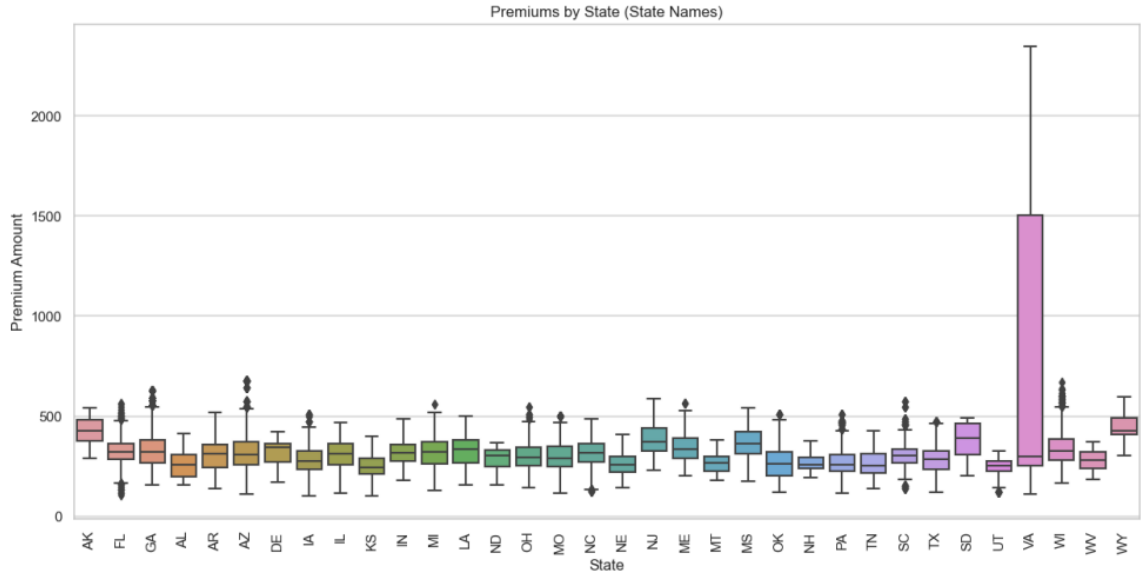


Fig 3: Average Premiums by state in USA

Wyoming (WY) ($477.39) and **Alaska (AK)** (282.69) shows lower premiums, indicating regional affordability disparities.



ANOVA Table:

	sum_sq	df	F	PR(>F)
C(State)	8.355112e+08	33.0	698.358506	0.0
Residual	2.804128e+09	77346.0	NaN	NaN

The state significantly impacts premiums ($p < 0.05$).

Fig 4: Boxplot of Health Insurance Premiums by State with ANOVA Results

The Analysis of Variance (ANOVA) test was conducted to evaluate whether there is a statistically significant difference in average premiums across states. A high F-value 698.35 indicates that the variation between states is much larger than the variation within states, suggesting a strong effect of geographic location on premiums. The **ANOVA** test confirmed that state significantly influences premium amounts ($p < 0.05$). This supports the hypothesis that regional disparities exist in healthcare affordability.

The reason why healthcare premiums vary from one region to another depends on several factors. Healthcare costs are region-based as states, which have higher provider charges or costs of prescription drugs, would have high premiums. Different states also set their premiums based on regionally specific requirements, for instance insurance mandates or the presence of Medicaid could increase the cost. Other such as such population demographic in terms of age structure or health risk also affect premiums as an older or less healthy population would push for more premiums. Market aspects such as the number of insurers and the level of competition have a bearing on the premiums affordability which is critical in less populated states where options are limited. Last but not least, factors such as the cost of living and

geographical factors of the region with respect to the availability of care also add to this regional variation especially in states where the level of healthcare resources is low, or the cost of healthcare is high.

3. **Can we predict healthcare plan premiums based on the plan's benefits, coverage options, and regional factors across the United States?**

In this analysis, we used coverage options like metal levels and plan types, State and county Information, Deductibles, out - of - pocket limits and drug coverage costs (Categorical columns were encoded).

Two regression models were implemented to predict healthcare premiums

1. **XGBoost Regression:**

```
XGBoost Regression Results:
R²: 0.7602
MAE: 22.0212
RMSE: 30.0511
```

The XGBoost regression model produced an R^2 value of 0.7602 suggesting that around 76% of health insurance premiums can be explained by the model. The Mean Absolute Error (MAE) has the value 22.0212 which suggests that the average predicted, and actual premium premiums are different on 22 units. The Root Mean Square Error of 30.0511 shows that the model's capability of reducing errors for large prediction values is quite good. It can be concluded from these metrics that the XGBoost model performs well in terms of prediction and is able to model relationships of premiums with that of determinants effectively.

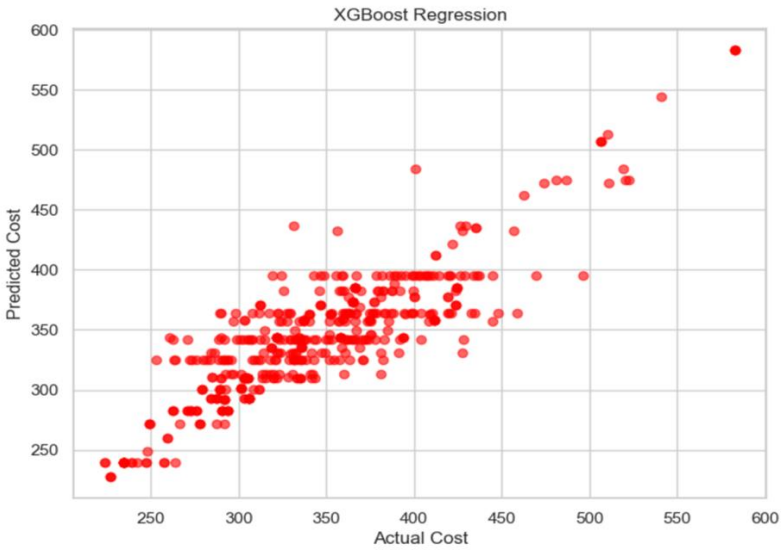


Fig 5: Scatter Plot of Actual vs. Predicted Costs Using XGBoost Regression

The scatter plot visually compares the actual healthcare premium costs (x-axis) with the predicted premium costs USING XGBoost.

2. K-NN Regression :

The second regression model that we run is K-Nearest neighbors, The K-NN Model explains 72.99% of the variance in healthcare premiums based on the selected features. Slightly lower (72.99%) compared to the XGBoost (76.02%). Higher MAE and RMSE Indicates that XGBoost Outperforms KNN in terms of accuracy.

```
KNN Regression Results:
R²: 0.7299
MAE: 23.6640
RMSE: 31.8911
```

	k	R^2	MAE	RMSE
0	1	0.410216	32.911943	47.126166
1	2	0.683266	24.927648	34.535287
2	3	0.716362	23.999966	32.681167
3	4	0.716980	24.459325	32.645565
4	5	0.729910	23.664020	31.891135
5	6	0.722687	23.995559	32.314723
6	7	0.721856	23.978829	32.363116
7	8	0.720626	24.289688	32.434578
8	9	0.717086	24.444915	32.639441
9	10	0.717913	24.492894	32.591677
10	11	0.722835	24.391156	32.306125
11	12	0.722273	24.490097	32.338872
12	13	0.728922	24.108919	31.949399
13	14	0.728662	24.234014	31.964727
14	15	0.725244	24.426822	32.165409
15	16	0.721963	24.628942	32.356921
16	17	0.720673	24.725260	32.431848
17	18	0.716422	24.875719	32.677722
18	19	0.714176	24.969950	32.806899
19	20	0.714115	24.990837	32.810377

The best performance is achieved when k = 5 as seen in the result

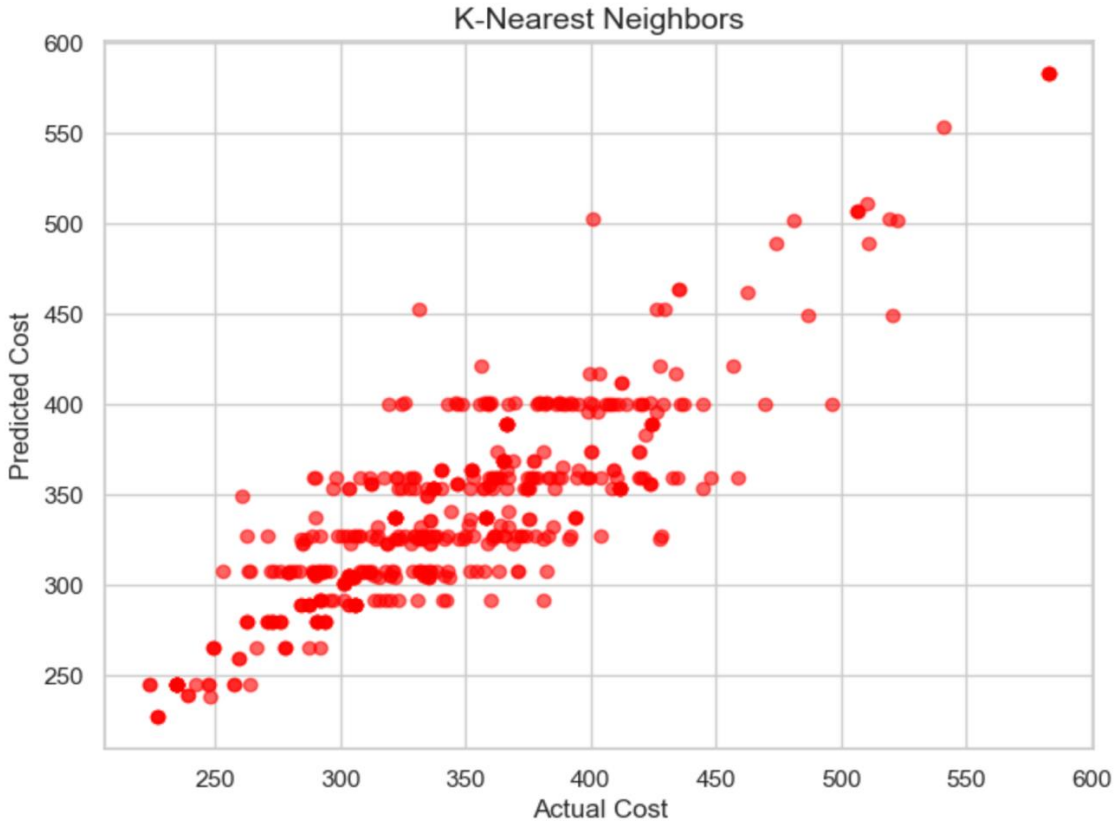


Fig 6: Scatter Plot of Actual vs. Predicted Costs Using K-Nearest Neighbors

The scatter plot visually compares the actual healthcare premium costs (x-axis) with the predicted premium costs using K-NN.

XGBoost is the better model for predicting healthcare costs due to its higher predictive power and lower error rates. This is due to XGBoost's advanced boosting mechanism, which effectively handles feature interactions and minimizes error iteratively, whereas KNN relies on simple distance-based predictions, which struggle with complex relationships in the data. Thus, XGBoost is more robust and better suited for this task.

Literature Review

For any research we do there needs to be a background that we can rely on to improve our research and overcome the limitation of previous research. This will help us to have a different approach to achieve our goals.

Related Works

Dafny et al (2012) and Ericson & Starc (2012) have shown the compaction in insurance plans regarding concentration based on the analysis which shows competition is inversely related to price meaning that in a less competitive market the health insurance plans will cost more in premiums. For this analysis which is designed according to the concentration of the insurers, the authors have calculated HHI based on those assumptions. The gap in this research is that the relationship and variation between premium prices in different geographical locations. We are correcting the model in relation to our analysis but in place of concentration of the insurers we are interested in the number of insurers to determine the trends in the metal levels as well as the premium plans.

In 2022, Sha Chen and Zhiye Lin conducted a detailed analysis of the existing health insurance plans in China which would help the individuals to maintain a healthy lifestyle. They have utilized a Two-Part model to examine health care utilization and employed Binary logistic regression in predicting the likelihood of out-of-pocket costs exceeding forty percent for households. As the data is self-reported there maybe falsification of data and may not contain all the factors affecting the premiums. We can incorporate

the use of Binary logistic regression in our model to find likelihood of an individual paying a high premium based solely on the state of residency.

Keshav Kaushik, Akashdeep Bhardwaj, Ashutosh Dhar Dwivedi, and Rajani Singh have developed a machine learning-based regression framework to forecast health insurance premium rates. They utilized an artificial neural network model where the co-related variables in the model are used in training the model, followed by linear regression and several evaluation metrics likes RMSE, MSE, and r^2. In their search paper, it was assumed that the relationship between the factors influencing the pricing with the cost of insurance based on premiums is linear. However, it is not the case in the real world, therefore linear regression can be replaced with Non-linear logistic regression to mitigate this concern. We are training our data set with the variables that are influencing the cost of the premium and predict the future price of the insurance premium.

As mentioned in the article by Schericki, Sukanya W. V. and V. M. Sivanand when predicting the price of insurance, different models are used , including Decision trees, support vector regression and Random Forest. But of all models, Random Forest test was the best model Because of the high R square value. It is also highly efficient in modelling of relations among the variable. In the study, researchers used mean to replace missing values caused by absence of subjects this method cannot be a good option considering that it is a method which can lead to wastage of valuable information within data instead we're trying to use median or mode so that the nonexistent data is represented as close as possible, outliers are directly removed in the study instead without proper analysis we are trying to assess the outliers influence on the data set and then take necessary actions them from the data set. This enhances our analysis to fully representant the dynamics present in our data.

The areas of interest of Kashish Bhatia, Shabeg Singh Gill focused on building a machine learning model, the aim of which is to predict the features of the insurance plans. Initially, they attempted to utilize linear regression, but they failed because the non-linear relationship was complicated. In the sept for reducing the complexity of the dataset they applied K-fold cross validation model by partitioning the data set into k subsets and k-1 is used as the training data set, while the other subsets are compared to the remaining

ones until all k subsets compared to k-1 have been tested. In the research, there are some outliers which were not identified and dealt with, which can bring down the accuracy and this affect the accuracy of the model which is clearly show in the results as being 81.3 percent the performance of the model can certainly be improved if outliers are dealt with properly. In our research we are focused on outliers and their proper management, which will raise the model's effectiveness. We will apply this same technique to our dataset in predicting premium cost forecast using various metal level, plan type and so on.

Bias and Limitations

While this research provides valuable insights into healthcare plan premiums, benefits, and regional differences, there are some limitations and biases to consider. First, the data we used was sourced from healthcare.gov, which focuses on public health insurance plans and does not include private insurance. This means our findings are limited to only part of the market. Additionally, the data only covers state and county levels, so smaller local differences are not captured. To address missing data, we used median imputation, which helps reduce skewed results but may not fully reflect the unique patterns in specific regions or plan categories. Outliers were removed using the IQR method, which improves data accuracy but might exclude extreme yet valid cases, like very high premiums, that could provide important insights.

On the methodological side, we used machine learning models like Random Forest and XGBoost. These models are good at finding patterns, but they rely on assumptions about the relationships between variables such as metal levels, plan types, and premiums. They may miss more complex, non-linear interactions. Like in similar studies, focusing on a few variables simplifies the analysis but leaves out other important factors like socioeconomic and demographic influences. We also applied log transformations to make the data more normal, which helps with analysis but might hide smaller premium variations that matter in low-cost areas.

In terms of model performance, while Random Forest worked well and explained about 90% of the premium variability, other models like KNN were less effective, showing that not all models are equally good at handling the data. Tests like ANOVA confirmed regional differences in premiums, but we didn't go into the deeper causes of these disparities, such as provider availability, state regulations, or income levels. Additionally, we focused mainly on premiums as the key measure, which means other important aspects like service quality, accessibility, or customer satisfaction were not explored.

Finally, because the data is from a single point in time, it doesn't account for changes over time, such as new healthcare policies or shifts in the market. We also didn't include factors like consumer preferences or employer contributions, which are important in real-life decision-making. Future research can address these issues by including more diverse data, looking at smaller geographic areas, and considering consumer behavior and market changes to give a more complete picture of healthcare plan dynamics.

Problem statement

The health insurance market in the USA is very complicated for consumers as it provides them with diversified plans that vary in terms of coverage and premiums. To achieve this, almost all health insurance plans in America are categorized into four metal levels; these include Bronze, Silver, Gold, and Platinum, which indicate an insurer's willingness to share costs with an insured. On top of this, there are also different plan types available such as Health Maintenance Organizations (HMOs) and Preferred Provider Organizations (PPOs), which only makes matters more complex for the clients.

For several consumers, choosing the relevant healthcare plan is a very difficult exercise owing to the absence of some clear information on the relations among the different factors affecting premiums and benefits. This variation in premiums is only worsened by the lack of regions that can provide a certain plan that is easily affordable. Such disparities mean that some parts of the country have easier access and are more likely to afford comprehensive coverage than other parts and this leads to unwanted inequalities in the access of healthcare.

There is consequently a clear relationship between the accessibility and affordability of healthcare providers in America depending on the region as mentioned above. This project focuses on these issues as it investigates how cost factors and premiums together with benefits available to subscribers vary with the metal levels and types of such plans. It is hoped that these analyses, will benefit the consumers, the insurers, and the policymakers.

Topic Backgrounds

In America, healthcare insurance takes the form of a framework of several, public and private, companies that have a wide variety of plans to cater for the myriad of the populations' needs. A structural element that typifies this system is the classification of the plans into so-called "metal levels", which assist in distinguishing the level of cost sharing associated with the plans. To give an example, bronze plans come with more affordable monthly premiums but higher out-of-pocket costs, and on the opposite end of the scheme, Platinum plans are the most expensive though the out-of-pocket costs are significantly reduced. Apart from the metal levels, healthcare plans are also classified into types, with Health Maintenance

Organizations (HMOs) and Preferred Provider Organizations (PPOs) being among the most common. HMOs are known to have lower premiums and there is a requirement to acquire service from specific provider network however PPOs provide more leeway by letting members see provider outside of the network albeit at a higher price.

Another factor that complicates the healthcare insurance market is its geographic aspect. Plans that can be sold in one region might be unsold in other regions or counties and differences in the amount paid can also depend greatly on which part of the country a client is located. Due to this situation, the opportunity to obtain medical services becomes unequal since there are areas where it is possible to get many reasonable and quite complete proposals and there are areas where it is not.

Understanding these variations in healthcare plans and premiums is crucial for consumers seeking to make informed decisions about their health coverage. It is also essential for insurers who must adjust their

offerings to remain competitive in different regions, as well as for policymakers who aim to address regional disparities and ensure more equitable access to healthcare across the United States.

Goals

The primary objectives of this research project are:

1. Analyze Premium Variations: Investigate how healthcare plan premiums vary across different metal levels and plan types (HMO, PPO, etc.).
2. Explore Regional Disparities: Identify and assess regional differences in plan affordability and availability across states and counties.
3. Develop Predictive Models: Create models to predict plan premiums based on coverage attributes and regional factors, allowing for a deeper understanding of price drivers.

Tools and Software

For this project, we will use the following tools and technologies:

1. Python for data cleaning, manipulation, and analysis. Libraries such as pandas and NumPy will be employed for data processing.

2. Visualization: matplotlib and seaborn will be used to generate detailed visualizations, such as heatmaps and distribution plots, that help illustrate trends in premiums and benefits.
3. Statistical Analysis: Regression models and hypothesis testing will be performed to quantify relationships between premiums, coverage attributes, and regional disparities.

Research Methods

The research is based on comparing the premium cost on health insurance based on the geographical location and type of insurance the insurers use for their family and the size of the family that is utilizing it. These points are needed to be remembered to be utilized in the research.

Logistic Regression

Logistic regression would be used in the classification of clients who are likely to incur high premiums. High-premium cases (top 10% of costs) are used to illustrate how societal factors and behavior impact upon the likelihood of having higher insurance premiums avoiding the confusion of risk factors in different groups.

Random Forest:

We can use the advantages provided by Random Forest Regression for predicting the health insurance premiums to illustrate high-cost premium cases. Because it is capable of modeling complex, non-linear interactions between variables such as family size, premium cost, and the location of the insurer. Random Forest offers a consistent, good prediction of premiums through the construction of several decision trees, predicting similar values, and all subsequent values are averaged. Further, the model will explain "feature importance" which is an understanding of which factors influence the premium costs to the highest extent. This two-way approach strategy makes it easy to analyze the situation, as Random Forest executes the explanation and prediction parts of the insurance pricing model while Logistic regression concerns with the risk patterns and associations among the high-cost categories in the model and provides an explanation to those policy issues.

DATA

Data Source: The data for this project is sourced from the publicly available dataset on health Insurance Plans from https://data.healthcare.gov

This Dataset was chosen because it provides comprehensive, up-to-date information health care plans across the United States. Including details of premium, meta levels, Plan Types, and regional factors.

Number of Records:

```
[25]: health_care_data = pd.read_csv('Health_Care_1.csv',low_memory=False)

      print(health_care_data.info())

      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 78379 entries, 0 to 78378
      Columns: 128 entries, State to Specialty Drugs - 94 percent
      dtypes: float64(41), object(87)
      memory usage: 50.5+ MB
      None
```

Fig 7: Raw data information

Approximately 78, 3799 records present in the dataset, capturing details about the plan's Premium, benefits, and other important attributes.

Number of Variables: We have 128 variables that are in the dataset we are using few of the important variables that are required for the research.

- **State:** The U.S. state where the healthcare plan is offered.
- **County:** The specific county within the state.
- **Metal Level:** The cost-sharing tier of the plan (e.g., Bronze, Silver, Gold, Platinum).
- **Plan Type:** The type of health insurance plan (e.g., Health Maintenance Organization (HMO), Preferred Provider Organization (PPO)).

- **Premiums:** The monthly cost consumers pay for the plan.
- **Deductibles:** The amount consumers must pay out of pocket before the insurance starts covering expenses.
- **Out-of-Pocket Maximums:** The maximum amount consumers are required to pay annually before the insurance covers 100% of expenses.
- **Co-payments:** Fixed amounts consumers pay for services like primary care visits or specialist consultations.
- **Co-insurance:** The percentage of costs consumers must pay for healthcare services after meeting the deductible.

Dependent Variables:

The healthcare plan premium is the main analytical outcome of interest. To determine how it differs by plan type, metal level, and region, this will be examined.

Independent Variables:

- **Metal Level:** Indicates the plan's cost-sharing tier (Bronze, Silver, Gold, Platinum, etc.), which has an impact on both premiums and out-of-pocket expenses.
- **Plan Type:** This refers to the insurance plan's classification (HMO, PPO, etc.), which affects the costs, and the way patients can obtain care (in-network versus out-of-network).
- **County and State:** geographic factors that represent regional variations in the accessibility and cost of healthcare plans in the United States.
- **Deductibles and Maximums for Outside Expenses:** measures the amount of financial responsibility that customers have both before and after insurance pays for all expenses, which affects the overall cost burden.

- **Co-payments and co-insurance:** These measures help to make healthcare more affordable by capturing the out-of-pocket expenses that patients must pay for things like doctor visits and specialized care.

Data Analysis :

Like any other dataset that needs to be cleaned of the errors, missing information and the Outliers our data set also needs to be cleaned of the errors and missing information. The values needed to be free of the missing values in the dataset.

First step of the cleaning process we have removed the variables that has no data ever recorded into it. The selected variables have some missing values that needed to be analyzed and needed to be modifies using mean, median or any other method.

Why Median Imputation?

Premium values across various plans for different age groups and regions can have missing values or outliers. The usage of median Imputation is insensitive to outliers compared to the mean; hence, it provides a stable and representative estimate for missing values in maintaining the integrity of data devoid of skewed results due to extremes.

The research we are trying to understand the disjointedness of health care premiums across different regions, plan types, and coverages. Missing values in key premium columns may skew such analyses by reducing the number of data points available for proper comparison. Median imputation provides a pragmatic middle ground that, while enabling wide-ranging insights, does not allow biases towards large or small premiums, which could distort averages.

Median Imputation Supports our Analysis Goals

The imputation of the missing values using median imputation furthers our goal of analyzing variation across different levels of metals and plan types, and regions. This ensures that any comparison reflects a good middle value for better reliability in case any region or categories of plans are identified with high or low premiums.

Filling missing values will prepare the dataset for predictive modeling, such as logistic regression, in which missing data could otherwise hurt model performance. The median acts like a neutral value that doesn't skew the model to head toward specific premium levels; thus, it keeps the support for a balanced and interpretable prediction.

Median imputation of data lets us study geographic variations in a more complete dataset. This helps in further information on affordability that might illustratively hint at regions where premiums are unaffordable or only very limitedly available.

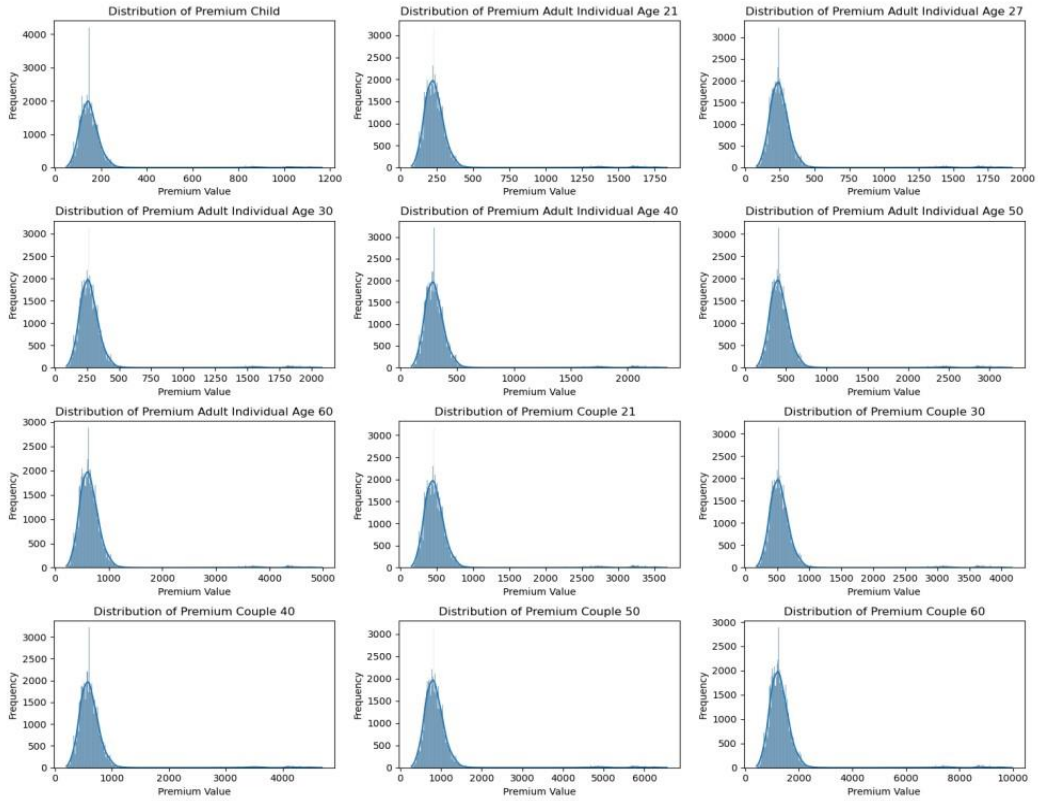


Fig 8. Distribution of Premium columns after Imputing Median

After the missing values are imputed using median the outliers still remines a part of the dataset that could result in poor model effectiveness.

```
# Creating boxplots for each premium column to visualize outliers
plt.figure(figsize=(14, 10))
healthcare_data_filtered[premium_columns_to_check].boxplot(rot=45)
plt.title("Boxplot of Premium Columns to Identify Outliers")
plt.ylabel("Premium Value")
plt.xlabel("Premium Categories")
plt.show()
```

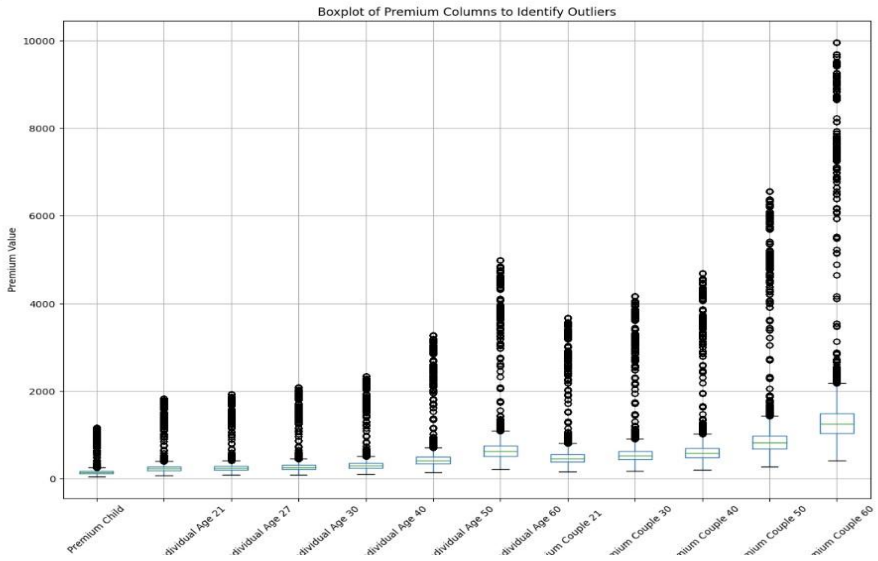


Fig 9. Box Plot of Premium Columns with Outliers

While analyzing the first boxplot, a range of outliers are observed across many of the premium categories, including premiums for children, different-aged adults, and couples. This is not uncommon with cost data, and certainly not in health insurance, where premium costs can be very different based on plan type, age, and region. This is important because outliers in the data related to premiums have a large impact on the mean and, therefore, could mislead the interpretation of central tendencies when we compare premiums across groups. If left untreated, they may skew statistical models and possibly produce biased results, most especially in models prone to extreme values.

Removing Outliers with IQR Method:

The outliers are removed using the IQR method, which excludes values that fall below Q1 - 1.5IQR or above Q3 + 1.5IQR. This removes extremely valued variables without distorting the rest of the data. E.g., I have added two columns that shows the normality after outlier removal.

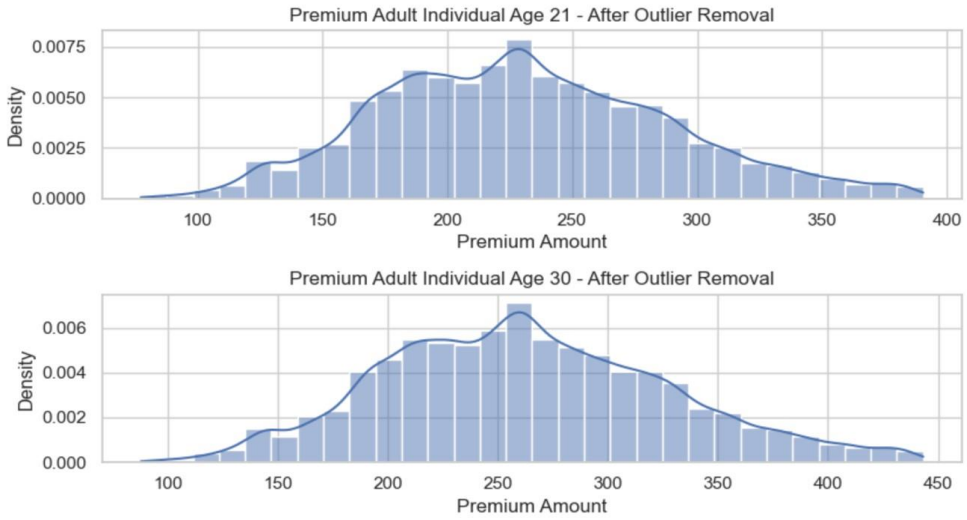


Fig 10. Normality of Two Premium columns after Removing Outliers with IQR Method

Removing outliers decreased the size of the dataset slightly, though this step was performed to provide a more stable dataset to make sure that extreme premiums did not disproportionately influence mean based analyses or predictive modeling. In other words, outlier removal assists in aligning data closer to a normal distribution. Since such distributions are more common, this step is useful for those analyses that assume a central tendency—a reduction in the effect caused by extreme values—such as regression. On the other hand, this might also cause the loss of some useful information if those extreme premiums were representative of actual, seldom-occurring cases in the data set.

Outlier removal was followed by applying the log transformation on the premium columns. It compresses the range of the data, especially for values high up, serving in skewness reduction.

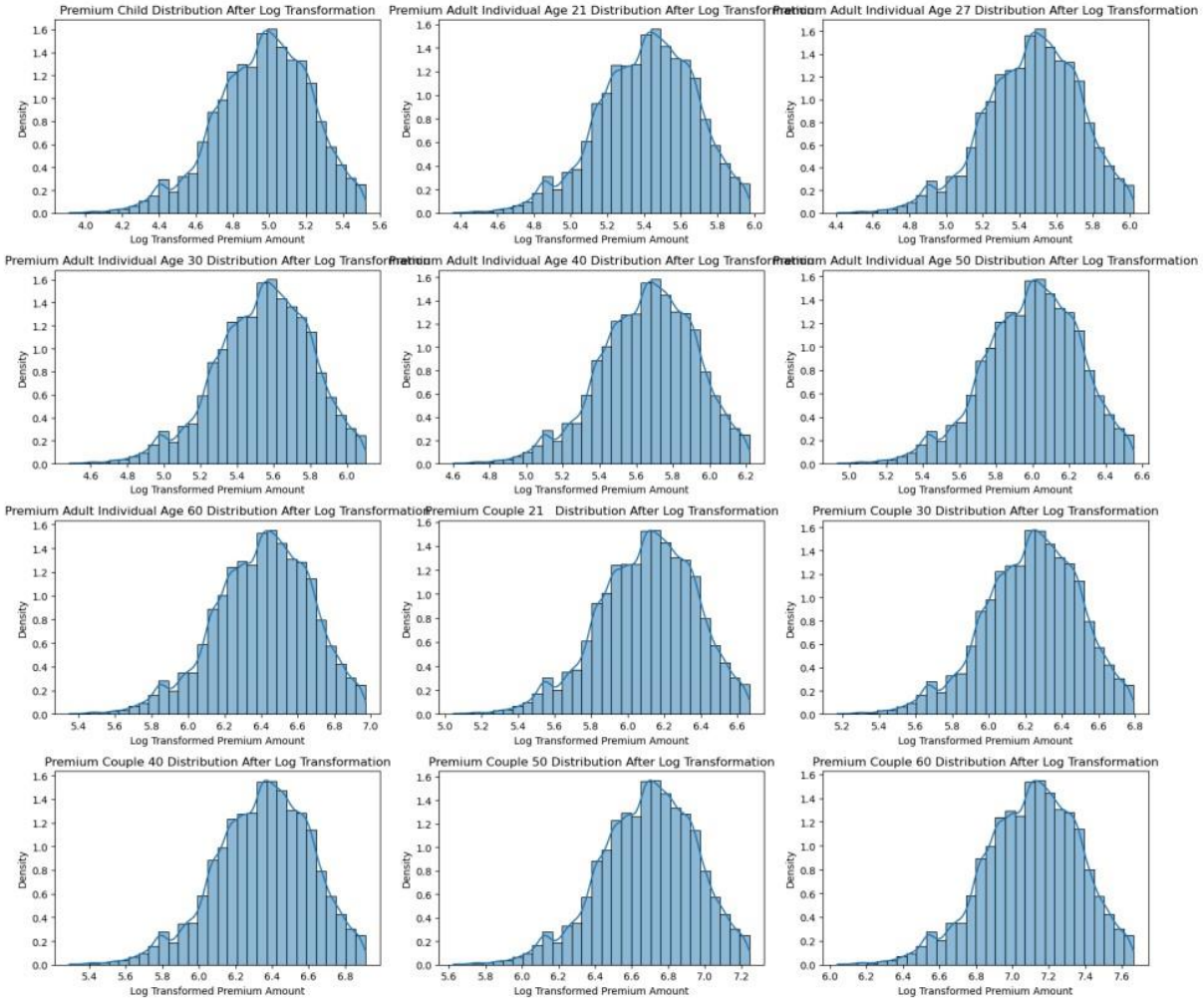


Fig 11. Normality of Premium columns after Log Transformation

Resulting in the log transformation that resulted in a significant reduction of skewness across the premium columns, with distributions much closer to normal. Q-Q plots supported this with the generally closer lying of the transformed data to a normal distribution, although there were still slight deviations at the tails for some columns. This would, therefore, imply that log transformations enhance the suitability of data for both parametric statistical tests and modeling techniques which linearity assume a normal distribution. By normalizing the distribution, we also achieve better interpretability, since it is much easier to compare premiums across categories.

Linearity of the Dataset : Since we have tried checking the linearity of the dataset as a part of assumption testing**.**

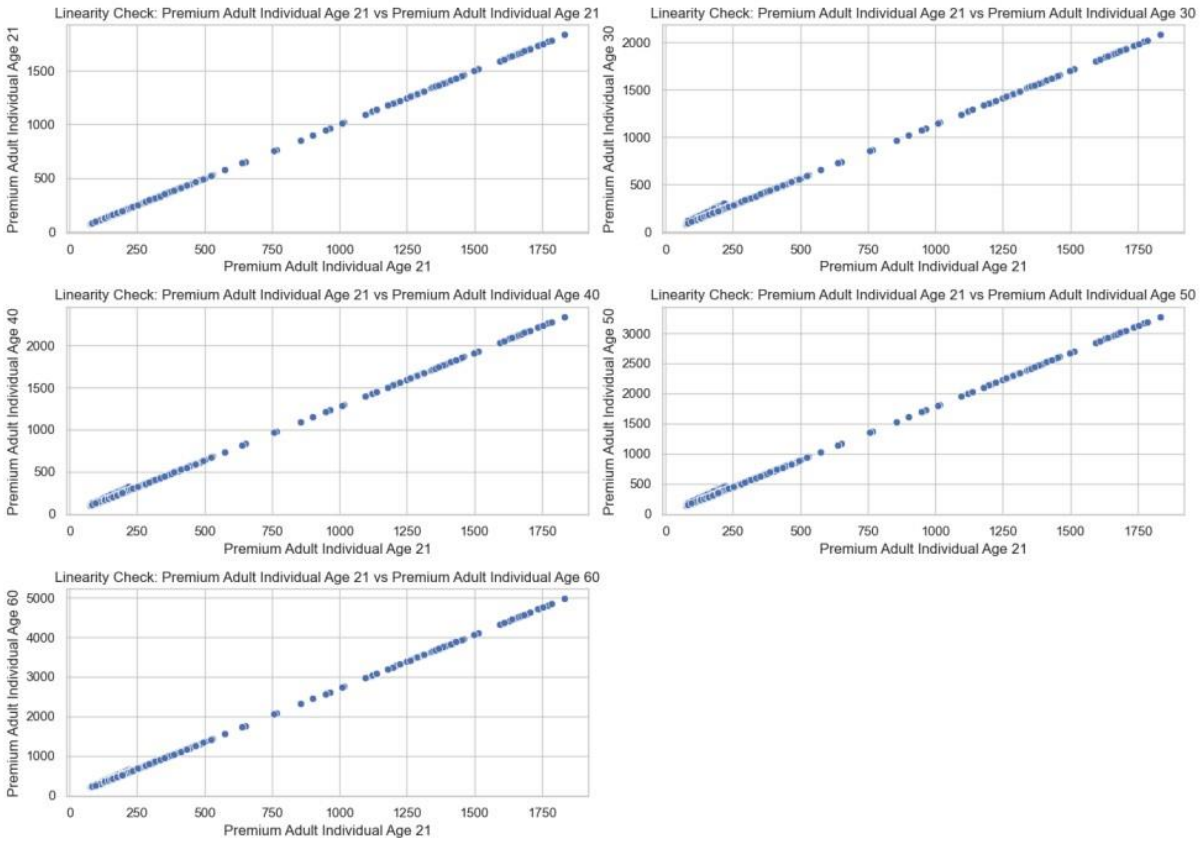


Fig 12. Linearity Check for the Premium Columns

The scatter plots above portray a high degree of linear relationship between the premiums for different ages. In each of the curves, the relationship between the premium of the adult individual who attained the age of 21 years, and the premiums for ages 30, 40 and 50, is almost linear. What this high degree of linearity means is that when premiums for a certain age group go up, premiums for other age groups tend to move up in about the same proportions.

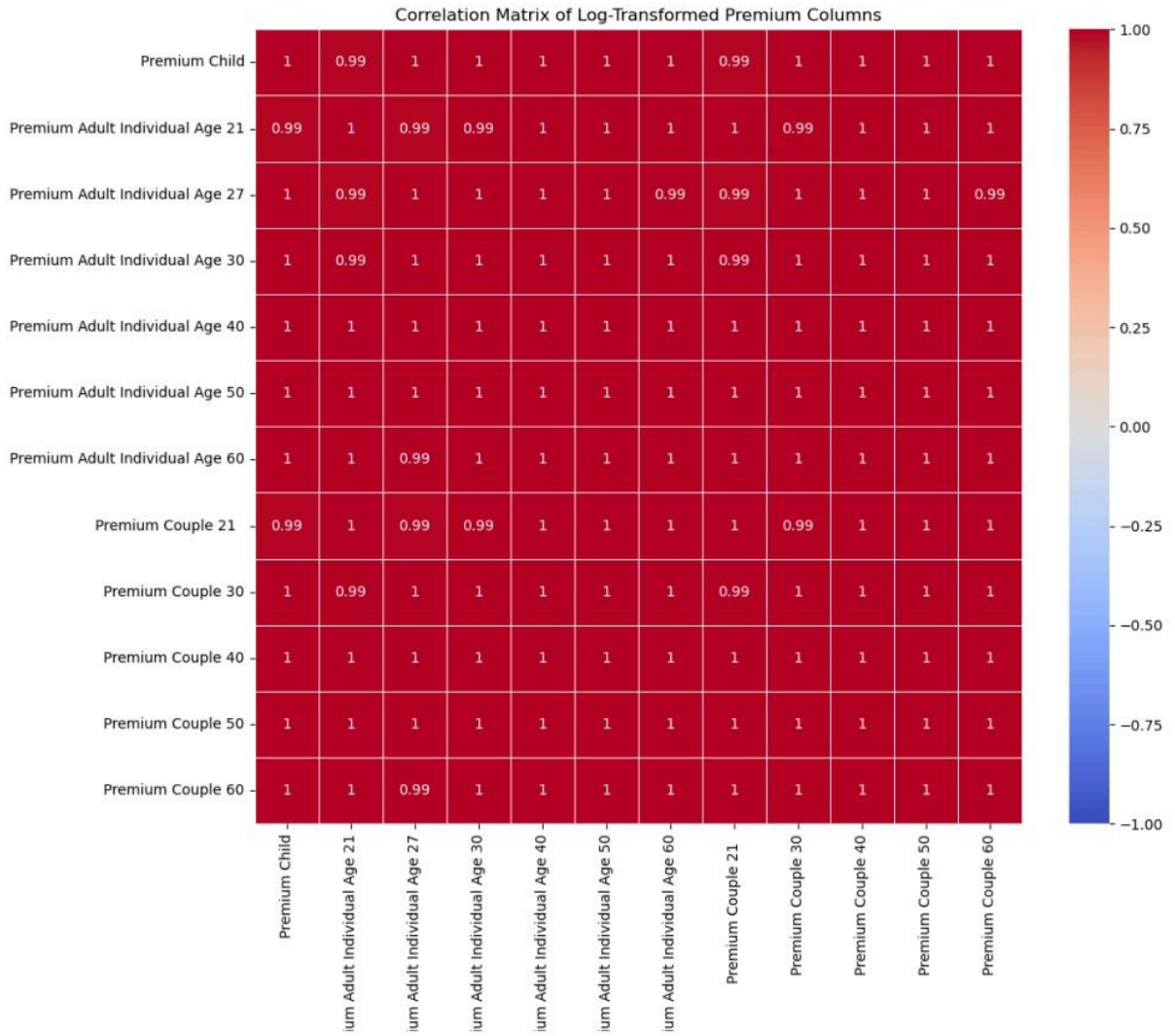


Fig 13. Correlation Matrix of Log-Transformed Premium Columns

By observing the structure of the correlation matrix, it could be seen that the correlations between different column premiums are very high with respect to each age class and each family composition, showing that premiums move practically proportionally. This is basically saying that there is a high level of multicollinearity; hence, several columns are carrying nearly identical information. This can also make the related regressions unstable and hence hard to comment on separate individual effects. Computing this correlation matrix aids in identifying these linear relationships and thereby helps us reduce redundancy. Toward this end, we could select one representative premium column, or apply dimensionality reduction to enable more parsimonious and interpretable analysis.

Overall Analysis

Improved Normality: Through removing outliers and log transformation, the data becomes more normal in distribution. This character makes the area of the figures more appropriate for further analysis – regression or predictive modeling, and so the assumptions of normality are needed.

Data Preprocessing and Trade-offs: Although the effect of outlier values is lowered, the data has been normalized, the data is most effectively ideal if outliers are present, but the removal of outliers means accepting a cost and missing out on data that may display real variability. This can be alarming in case extreme values are expected to be useful in analytics; for example, in a region consisting of high costs or low-cost regions.

Improved Baseline for Analysis: The dataset formed is more able to determine framework of the plans and age differences including region differences in terms of frontiers by limit more the extent of dispersion index. Low dispersion is then more appropriate to test the hypotheses, and even better, deep conclusions will allow and best create the models.

Discussion

This research looked at how healthcare plan premiums, benefits, and costs vary across the United States. We found that Platinum plans, which offer the most coverage, have the highest premiums, while bronze plans are the cheapest. PPO plans were generally more expensive than HMO plans because they offer more flexibility in choosing providers. These findings are helpful for consumers trying to decide between affordability and coverage.

There were also big differences in premiums based on geography. States like Wyoming and Alaska had very different average premiums compared to Virginia. These differences are likely due to factors like the number of healthcare providers, state regulations, population health, and competition among insurers. Our predictive models showed that features like metal levels, plan types, and location play an important role in determining premium costs. The Random Forest model performed the best, explaining about 90% of the variation in premiums, while XGBoost also provided useful predictions.

However, this study had some limitations. We didn't include private insurance data, and the methods used to handle missing data might have missed unique patterns in specific areas. Despite these challenges, the research gives important insights for consumers choosing plans, insurers creating better offerings, and policymakers addressing regional differences. In the future, adding more data, like private insurance and demographic information, and using time-based analysis could provide an even clearer picture of healthcare costs.

Overall, this research helps explain the factors that drive healthcare premiums and highlights areas where improvements could make coverage fairer and more affordable for everyone.

Conclusion

This research explored the factors that affect healthcare plan premiums, benefits, and regional differences across the United States. By analyzing data on metal levels, plan types, and geographic locations, we found clear patterns that impact affordability. Platinum plans, which offer the most coverage, had the highest premiums, while bronze plans were the most affordable. We also found significant regional differences in premiums, influenced by factors like the number of healthcare providers, state regulations, and competition among insurers.

Using predictive models like Random Forest and XGBoost, we showed how features like metal levels and plan types play an important role in determining premium costs. Random Forest performed the best, explaining about 90% of the variation in premiums. These findings provide helpful insights for consumers, insurers, and policymakers looking to understand and address healthcare costs.

However, there were some limitations, such as not including private insurance data and relying on methods to handle missing data that might have missed unique regional patterns. Future research could address these issues by adding more comprehensive datasets, demographic information, and looking at changes over time to provide a clearer picture of healthcare costs.

Overall, this research helps explain what drives healthcare premiums and highlights areas where improvements could make healthcare more affordable and fairer for everyone. It provides useful knowledge for better decision-making and supports efforts to create a more equitable healthcare system.

Future work

Including Private Insurance Data:

- Future studies should add data from private insurance providers to give a more complete picture of the healthcare market. This would allow us to compare public and private plans and understand overall affordability better.

Looking Beyond Premiums:

- Future work could explore other important factors like the quality of care, access to healthcare services, and customer satisfaction to get a full view of healthcare plans.

Adding Behavioral and Demographic Data:

- Including information like household income, age, health status, and consumer preferences could provide a deeper understanding of what influences premium costs and how people choose their plans.

Using Time-Series Analysis:

- Analyzing data over time would help us see how premiums and benefits change, especially with new policies, economic conditions, or shifts in the market.

Building Regional Models:

- Future research could create models specific to certain regions or states to give more targeted insights and recommendations for both policymakers and insurers.

Adding More Local Data:

- Including more detailed geographic information, like data at the city or ZIP code level, would provide a better understanding of healthcare affordability and access at a local level.

List of References

Chen, S., Lin, Z., Fan, X., Li, J., Xie, Y. J., & Hao, C. (2022). The comparison of various types of health insurance in the healthcare utilization, costs and catastrophic health expenditures among middle-aged and older Chinese adults. International Journal of Environmental Research and Public Health, 19(10), 5956.

Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. International journal of environmental research and public health, 19(13), 7898.

Bhatia, K., Gill, S. S., Kamboj, N., Kumar, M., & Bhatia, R. K. (2022, May). Health insurance cost prediction using machine learning. In 2022 3rd International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE. X

Dafny, L., Gruber, J., & Ody, C. (2015). More insurers lower premiums: Evidence from initial pricing in the health insurance marketplaces. American Journal of Health Economics, 1(1), 53-81.

Vijayalakshmi, V., Selvakumar, A., & Panimalar, K. (2023, January). Implementation of medical insurance price prediction system using regression algorithms. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1529-1534). IEEE

Authors' Contributions

This project was a true team effort, with everyone bringing their unique skills and ideas to the table:

Akshaya Thangavel: Kickstarted the project by crafting the introduction. Akshaya did an amazing job setting the stage by explaining the challenges of the U.S. healthcare system and why this research is so important.

Daksesh Kasumurthi: Took charge of the research questions, making sure we had clear goals to guide our work. Daksesh's questions really helped focus the project and gave us a solid direction.

Imran Nawaz Shareef: Dug deep into past studies for the literature review. Imran connected our work to existing research, showing where we could add something new and valuable to the conversation.

Vamshi Krishna Bairoju: Led the methodology section with expertise, walking us through the data cleaning, modeling, and analysis process step by step. Vamshi's approach made everything come together smoothly.

Vinila Chowdary Potla: Wrapped up the project by writing the discussion and conclusion. Vinila made sure our findings were clear and meaningful, tying everything back to how it can help consumers, insurers, and policymakers.