MEDICAL INSURANCE PREMIUM PREDICTION

Descriptive Analysis

GROUP 05

Kavindi Chamathka - s16367

Nethmi Sansala - s16252

Kavindu Weerasekara - s16076

Abstract

Medical insurance premium prediction is a critical aspect for both insurance providers and insurers, helping to determine fair pricing and manage financial risks. This report aims to present the findings of the exploratory data analysis conducted on '**Medical Insurance Premium Prediction**' dataset obtained from Kaggle. The key objective of this analysis is to identify the factors that have a huge impact on medical premium prices and predict the medical insurance premium. The findings from this analysis can serve as a foundation for the insurers to understand the factors affecting their premiums and make informed decisions about their insurance plans. Also, the insights of this analysis help insurance providers to improve their pricing strategies and evaluate risk for more accurate premium calculations.

Table of Contents

List of Figures

List of Tables

Introduction

Medical insurance offers medical coverage for expenses incurred by the insured in a medical emergency. As the costs of medical treatment rise, affording quality care becomes challenging. To manage this, people search for suitable medical insurance plans, paying premiums in exchange for benefits. The insurance premium may vary depending on many factors. Hence, determining the accurate insurance premium depending on the insured requirements is important to build stronger customer relationships, to customize medical insurance plans and to reduce the risk faced by the insurer. This report focuses on the descriptive analysis of medical insurance premium prediction, exploring several key variables.

Description of the question

In this descriptive analysis we focus on understanding the structure of the dataset, trends and relationships between variables. Through descriptive statistics and visualizations, this report aims to explore how different factors influence medical insurance premiums, answering the following questions:

- What are the main factors affecting medical insurance premiums?

- How well do descriptive statistical techniques help in understanding the distribution and relationships of those factors?

Description of the data set

The '**Medical Insurance Premium Prediction**' dataset obtained from Kaggle contains 986 observations of customers with 11 variables. A description of each variable can be found in the table below.

Variable Name	Description	Variable Type
Age	Age of customer	Quantitative
Diabetes	Whether The Person Has Abnormal blood sugar Levels	Categorical (Nominal)
BloodPressureProblems	Whether the person has abnormal blood pressure levels	Categorical (Nominal)
AnyTransplants	Any major organ transplants	Categorical (Nominal)
AnyChronicDiseases	Whether customer suffers from chronic ailments like asthma, etc.	Categorical (Nominal)
Height	Height of customer (in cm)	Quantitative
Weight	Weight of customer (in kg)	Quantitative
KnownAllergies	Whether the customer has any known allergies	Categorical (Nominal)
HistoryOfCancerInFamily	Whether any blood relative of the customer has had any form of cancer	Categorical (Nominal)
NumberOfMajorSurgeries	The number of major surgeries that the person has had	Categorical (Nominal)
PremiumPrice	Yearly premium price in INR(₹)	Quantitative

Table 1: Dataset Description

For categorical variables except *'NumberOfMajorSurgeries'*, 0 represents 'No' and a value of 1 represents 'Yes'

Data Pre-processing

- No duplicate records were found in the dataset.
- No missing values were found in the dataset.
- Checked for outliers. A very small number of outliers were identified only in weight and premium price, indicating that the majority of the data points fall within the expected range. Therefore, these outliers will not be removed.
- Then the dataset was split into training and test sets such that the training set consisted of 789 observations.
- Afterward, the descriptive analysis was conducted using the training set.

Main Results of the Descriptive Analysis

Univariate Analysis

Distribution of the Response Variable - Premium Price

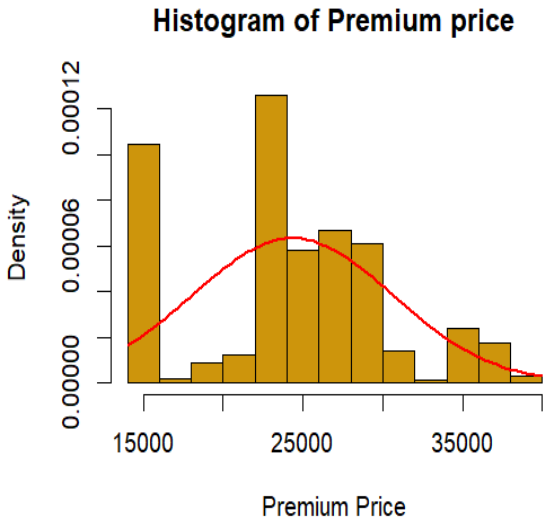


Figure 1: Histogram of Premium Price

The histogram in *Figure 01* indicates that the distribution of premium price appears to be right-skewed. There are some peaks around 15000, 25000 and 30000, suggesting that a significant number of premiums fall within these price ranges. The presence of multiple peaks suggests that the data may not follow a perfectly normal distribution. Therefore, we might need to apply some transformation technique like log transformation on premium price when building models. Also, some price ranges contain little or no data, which implies specific pricing structures that avoid certain ranges.

Bivariate Analysis

Premium Price by Several Categorical Variables

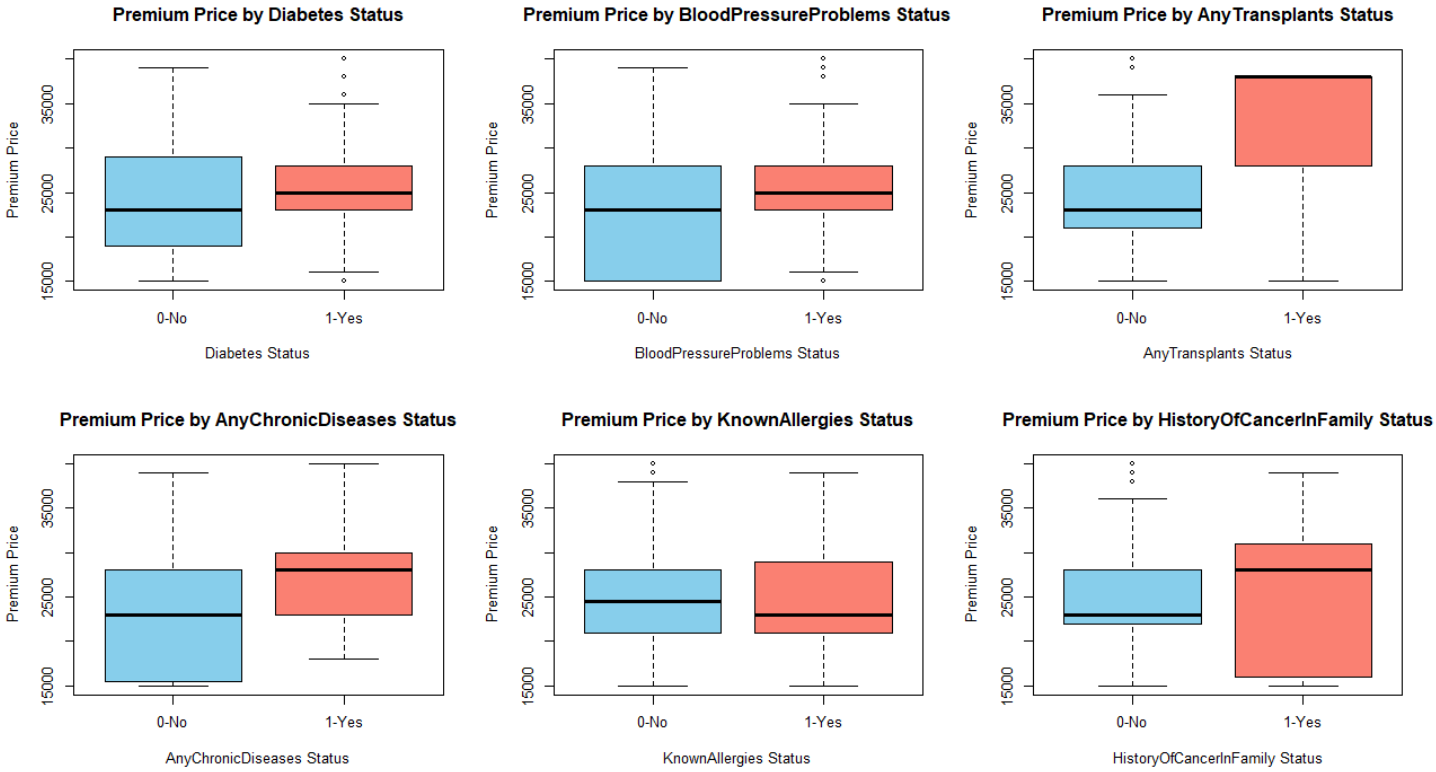


Figure 2: Boxplots for premium price by different categorical variables

The set of boxplots in *Figure 02* illustrates the relationship between Premium Price and six different health conditions: Diabetes, Blood Pressure Problems, whether the customer has any Chronic Diseases, History of Cancer in Family, and whether the customer has known allergies or not. Each boxplot compares premium prices for individuals with (1) and without (0) each condition.

1. **Premium Price comparison for people with and without Diabetes**: Individuals with diabetes tend to have a slightly higher median premium price compared to those without diabetes. This may be because a person with diabetes may require more frequent doctor visits, or medications. To account for this higher risk, insurers may raise the premium prices to ensure they can cover the additional medical expenses. The interquartile range (IQR) is larger for the non-diabetic group, indicating higher variability. The high-end outliers in the

affected group suggest that some individuals with diabetes are charged significantly higher premiums.

2. **Premium Price comparison for people with and without Blood Pressure**: Like diabetes, individuals having blood pressure problems also tend to have slightly higher median premium prices. The spread of premiums is relatively lower in the affected group. But here also there are some high-end outliers indicating that some individuals with blood pressure problems face significantly higher premiums.

3. **Premium Price comparison for people with and without a Transplant**: This plot shows that, people who has undergone a transplant have very high median premium price compared to those who haven't undergone any transplant. Also, a significant portion of the insured individuals with a history of transplants pay similar premium amounts (38,000), leading to a compressed upper range in the boxplot.

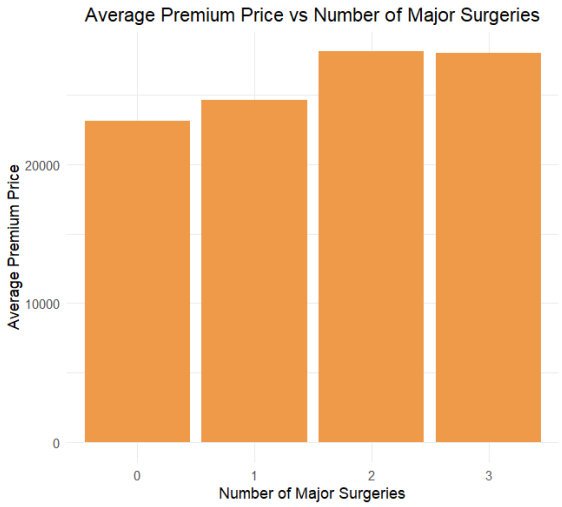
4. **Premium Price comparison for people with and without a Chronic Disease**: Having a chronic disease is also associated with higher premiums. The median premium price is significantly higher for affected individuals, and the overall spread of values is lower, compared to those without chronic diseases.

5. **Premium Price comparison for people with and without any Known Allergies:** The median premium price is slightly higher for individuals without known allergies compared to those with allergies. However, the overall distributions are quite similar, with a slightly larger spread in the allergy group. Therefore, the presence of allergies does not seem to have a strong effect on premium prices.

6. **Premium Price comparison for people with and without a Family History of Cancer**: Individuals with a family history of cancer generally have higher median premiums. The range of premiums is slightly larger for those with a family history of cancer.

Above analysis implies that having diabetes, blood pressure, a history of transplants, chronic diseases, and a family history of cancer have a significant impact on the insurance premium price. These health conditions seem to be closely linked to higher premium prices, indicating that individuals with such conditions are more likely to incur higher insurance costs due to the greater health risks they carry.

Number of Major Surgeries Vs. Premium Price



This shows that there is no strong variation in premium prices based on the number of major surgeries. However, we can observe that individuals who have undergone more surgeries tend to pay slightly higher premiums on average.

Figure 3: Bar chart of Average premium price vs no.of surgeries

Correlation Among Numerical Variables

This correlation heatmap represents the relationships between the numerical variables in the dataset. A strong positive correlation (0.69) exists between Age and Premium Price, which implies that older individuals tend to have higher premium prices. Weight and Premium Price show a weak positive correlation (0.16), while Height and Premium Price have a very low correlation (0.05), implying that these factors have a lower impact on the premium price. Similarly, the correlations between the predictor variables are also weak, indicating less multicollinearity.

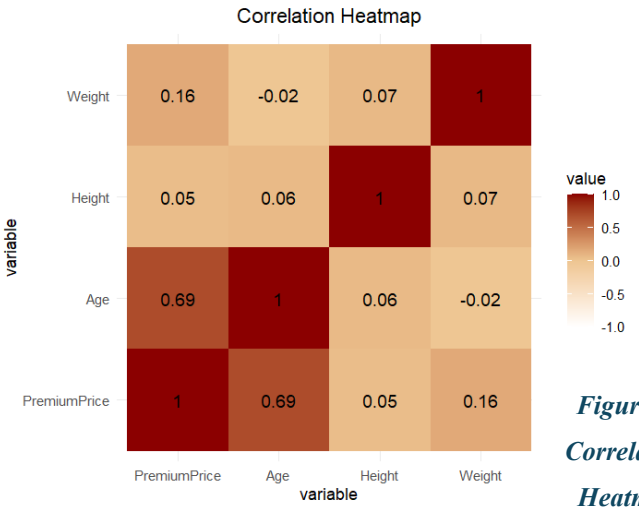


Figure 4: Correlation Heatmap

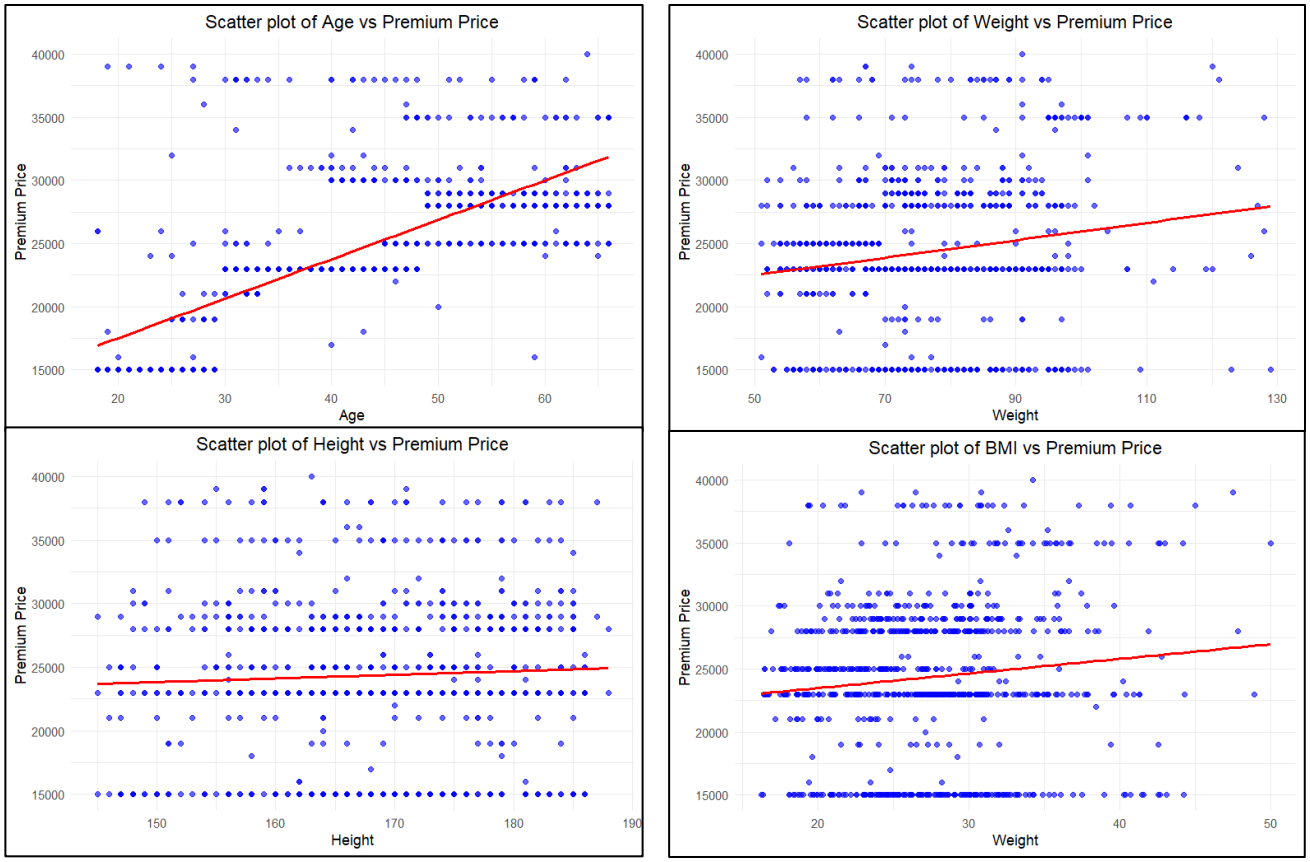


Figure 5: Scatter plots for premium price vs quantitative variables

The correlation heatmap and scatter plots provide a clear understanding of the relationships between variables and premium price. Among all of the above 3 factors analyzed, Age shows the strongest positive correlation with premium price, as reflected in both the heatmap and the scatter plot, where the points follow an upward trend. In contrast, height and weight exhibit minimal correlation with premium price, with their scatter plots showing a random distribution of points, confirming their weak relationships. To explore a possible combined effect, BMI was calculated using height and weight, but it also showed a weak correlation with premium price (0.1083). This indicates that BMI, like height and weight individually, is not a strong predictor of insurance premium costs. In conclusion, Age is the most significant factor influencing premium price, while height, weight and BMI have minimal impact.

Principal Component Analysis

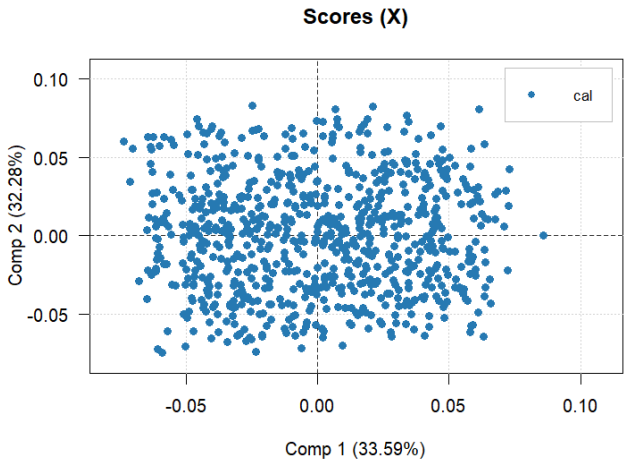


Figure 6: Score Plot of X

A Principal Component Analysis was conducted to examine the datasets for potential clusters. Since PCA is applicable only to numerical data, all categorical variables were removed before the analysis. To enable visualization in two dimensions, only the first two principal components were selected for representation. However, this dimensional reduction may limit the accuracy of the resulting score plot. This score plot does not indicate any well-defined clusters, as the points are spread out without distinct groupings.

Partial Least Square Regression

Further, PLS was applied to the training set to identify the relationships between the predictors and the premium price. The XY loadings plot also shows that age has the strongest connection to the premium price, followed by weight, which also has a moderate relationship with the response variable. Height, however, shows a weak connection, contributing little to explaining the response. The plot also indicates minimal multicollinearity between the predictors. However, the first two components explain only 33.59% of the variation in predictors and 32.28% in the response, suggesting that the model may need more components to provide a clearer picture.

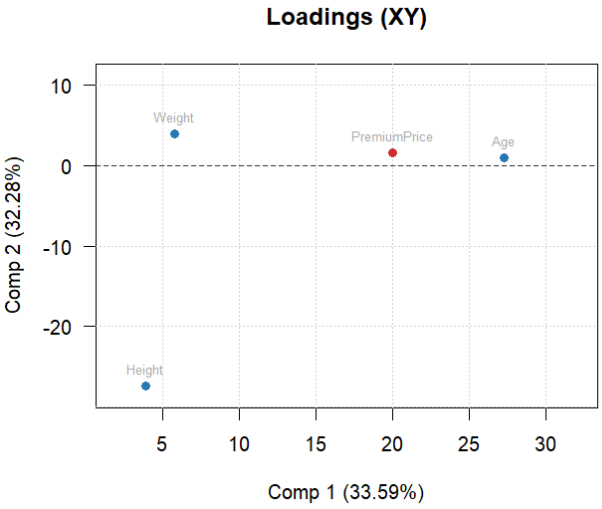


Figure 7: Loading Plot of XY

Conclusion

The descriptive analysis provided valuable insights into the factors influencing insurance premium prices. It indicated that individuals with diabetes, blood pressure, chronic diseases, history of transplants, and a family history of cancer tend to have higher premium prices, reflecting the increased health risks associated with these conditions. However, the presence of known allergies showed no significant impact on premium price. Among numerical variables, age had the strongest positive correlation with premium price, implying that older individuals generally pay higher premiums.

Suggestions for a Quality Advanced Analysis

As shown in *Figure 01,* since the premium price variable is right skewed, applying log transformation might help to normalize its distribution, improving the performance of regression-based models. As the number of outliers are minimal, it may not be necessary to remove them unless they significantly impact the model performance. Instead, using robust models like decision trees and random forests could naturally handle these outliers. To potentially enhance predictive accuracy, it might be beneficial to explore advanced models such as Gradient Boosting (XGBoost) or regularization techniques like Lasso regression. Lasso regression might be useful for feature selection, as it shrinks weak predictors like height to zero, reducing noise in the model.

Appendix

1. Link for the dataset: https://www.kaggle.com/tejashvi14/medical-insurance-premium-prediction
2. Click here for R codes

References

1. https://www.kaggle.com/code/smmomin/medical-premium-eda-93-prediction-model
2. https://www.kaggle.com/code/mohammedmunshif/eda-and-model-building-of-medical-premium-dataset#Description-of-the-question
3. https://www.kaggle.com/code/bidiptabikashgogoi/eda-medical-insurance-claim