Individual Predictive and Decision-Making Report

Prepared by: Sumiya Fatema Keya

Date: November 14, 2023

Table of Contents

List of Tables

List of Figures

Introduction

Good health is essential for everyone in all aspects, as everything evolves around good health. But with the fast pace of life gradually, we are adopting some bad habits that are affecting our life. Health insurance plays a vital role in reducing the financial burden associated with medical expenses. Insurance helps to provide quality medical services to people. There are different forms of healthcare insurance that vary due to various factors such as age, health condition, state or province, and coverage preferences. It is imperative for both policymakers and individuals to understand these variations in order to design and choose effective healthcare policies. The main objective of this analysis study is to predict the future medical charges of insurance policies based on the provided sample data. Secondly, the study also intends to identify multiple factors that affect the insurance charges of the sample's subjects based on different descriptive analyses.

Background

Health insurance is a multifaceted and complex landscape. In the backdrop of escalating healthcare costs, health insurance emerges as an important and critical financial safeguard against unforeseen medical expenses (Cutler & Zeckhauser, 2000). Health insurance is a financial system designed to cover medical expenses, and it was established in the early 20th century (Finkelstein, Baicker, & Malani, 2020). To talk about the population segment who greatly benefit from these health insurance services are older individuals and people with low income. Unbalanced lifestyle choices and adopting bad habits such as smoking, diet,exercise,e etc. are destroying human health, resulting rise in health expenses. So, understanding and analyzing several factors that affects the insurance charges to go up is crucial for all the people involved in this industry, starting from policymakers and individuals seeking for insurance coverage for themselves.

Literature Review

Health insurance is considered a complex aspect of financing, and understanding all the factors that has impacts on these expenses is vital for effective policy formulation and planning. In this analysis, both descriptive and predictive analysis plays a crucial role by highlighting the complexities related to health insurance expenditure. A comprehensive descriptive analysis has been conducted by Wong et al. (2018), focusing on demographic and regional variations in insurance spending. The analysis revealed disparities in expense patterns across different age groups and demographic locations. Smith and Johnson (2020) used machine learning algorithms to analyze and predict future expenses based on age, pre-existing conditions.

Furthermore, another study by Brown et al. (2019) focused on lifestyle factors in predicting health insurance costs. In his study, there were many variables incorporated, like smoking, physical activities etc. and their analysis provided valuable insights for insurance risk assessment. Overall, the combination of descriptive and predictive analysis in the field of health insurance expenditure understanding will help both policymakers and stakeholders to make informed decisions about healthcare finances.

Methodology

Data Collection

The provided dataset contains 1338 rows and 7 columns representing column -age, sex, bmi, children, smoker, region, and charges. Here, the ‘charge’ column is the target column, and others are being considered as independent data.

Data Diagnostics

To perform the descriptive and prediction analysis on the insurance charge data as an analytical tool, Exce v16 has been used here. There was no missing data found in the dataset. Three out of seven columns contain categorical data.

Analytics Strategies

To perform the descriptive analysis, different Excel function such as histogram, pivot chart, box plot, frequency distribution, scatter plot, and hypothesis analysis has been used. As the dataset contains categorical data. In order to normalize the data dummy variable has been used. Later on, to perform prediction analysis, piecewise regression was adopted for this study.

Results and Discussion

Descriptive Analysis

Age, sex, children, smoker, region, bmi, and charge data analysis

The min and max age of the dataset was calculated-

Min age =18

Max age= 64 and the Average Age = 39.2

Plotted a histogram with six bin to visualize data. The histogram seems asymmetric, and it has almost equal no of people from all age groups.

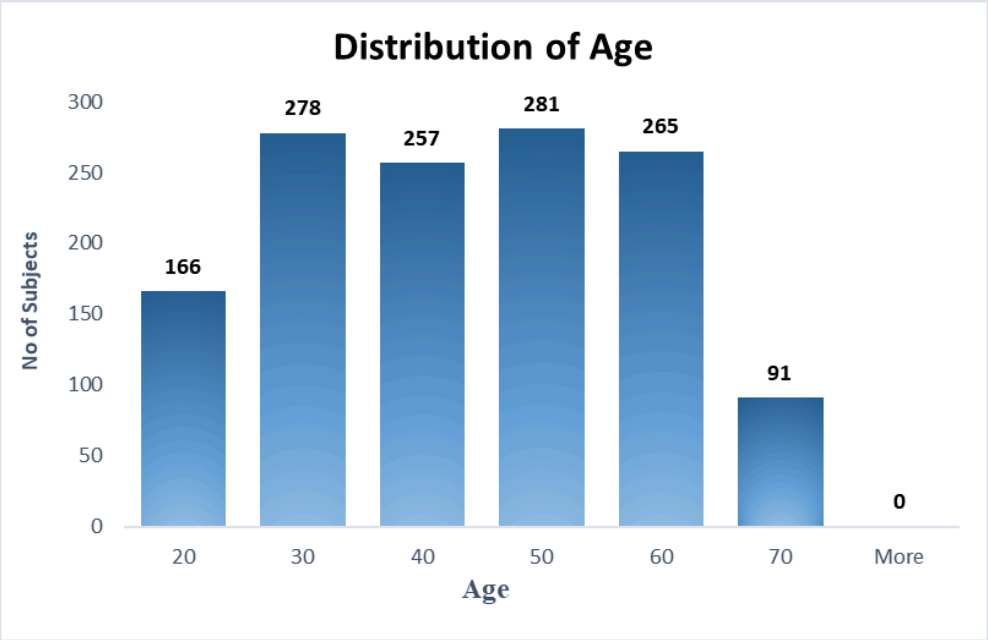

Fig.1: Distribution of age

To understand the ratio of male and female in the sample data, percent frequency and a pie chart were plotted to visualize the ratio.

male = 50.52%

female: 49.48%

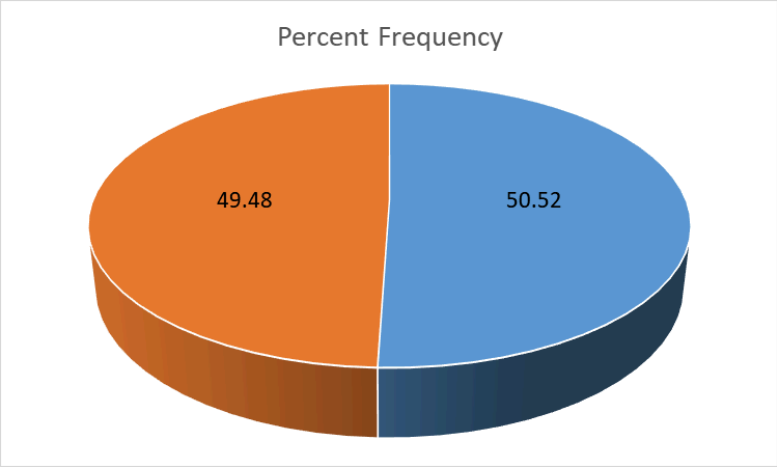

Fig.2: Ratio of male and female subjects

Calculated Mean. Median, Min and Max for the bmi data also calculated it's standard deviation and 25, 50 and 75 quartiles to understand the analysis of BMI.

bmi	
Mean	30.66339686
Median	30.4
Mode	32.3
Min	15.96
Max	53.13
Standard deviation	6.098186912
Quartile (25%)	26.2725
Quartile (50%)	30.4
Quartile (75%)	34.7

Table 1: Analysis of bmi data

Children data were also analyzed, where the subjects have the highest 5 children. The frequency analysis using 6 classes/bin shows-

- No of people have no children: 574 (Highest)
- No of people have 1 children: 324
- No of people have 2 children: 240
- No of people have 3 children: 157
- No of people have 4 children: 25
- No of people have 5 children: 18 (Lowest)

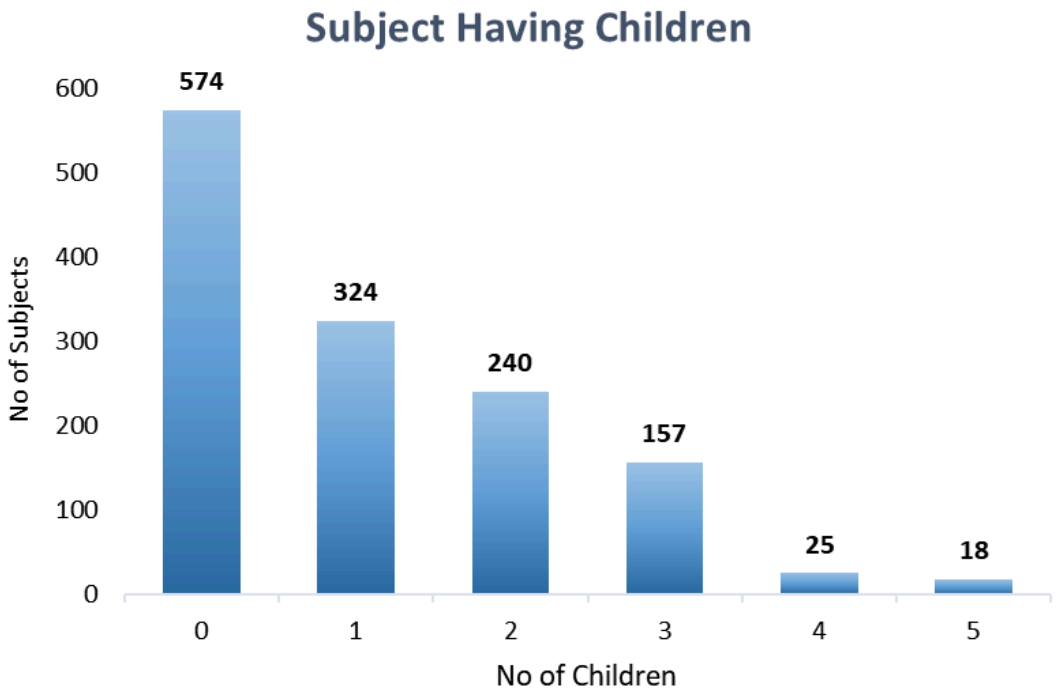

Fig.3: Distribution of subject with different numbers of children

For categorical 'Smoker' data, frequency =distribution percent frequency was calculated, which shows the below data-

smokers: 20.48~20%

non-smokers: 79.52 ~ 80%

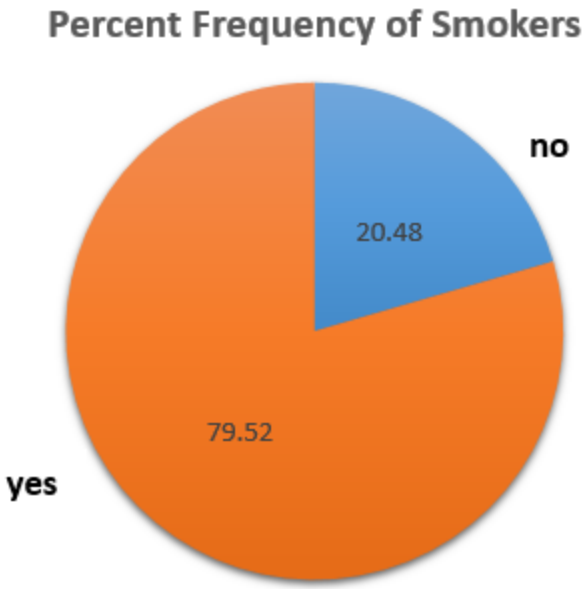

Fig. 4: Percentage of smokers and non-smokers

To analyze the region-wise inhabitants numbers frequency distribution process was used and % of inhabitants across the region was detected-

Region	No of Inhabitants
southeast	364
southwest	325
northeast	324
northwest	325

Table 2: Region-wise inhabitants no.

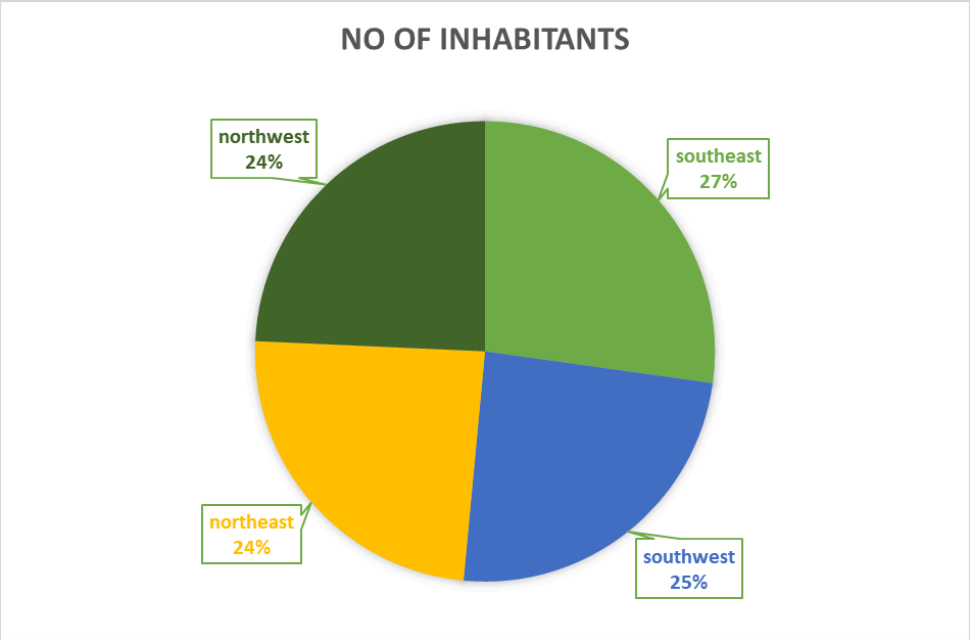

Fig.5: Percentage of inhabitants

Calculated Mean, Min, and Max for the charges data and calculated its standard deviation to understand the analysis of charges-

Charges	
Min	1121.87
Max	63770.43
Mean	13270.42
Standard Deviation	12110.01

Table 3: Analysis of charge

Relation of Independent Column with Charge Column

Utilized scatter plot to detect age vs charge relationship. From the scatter plot, is has been shown that for most of the observants, insurance charges are increasing with age, but some expectations are available where some subject with higher age has low insurance charges. But overall, it can be said that age and charges have a linear relationship.

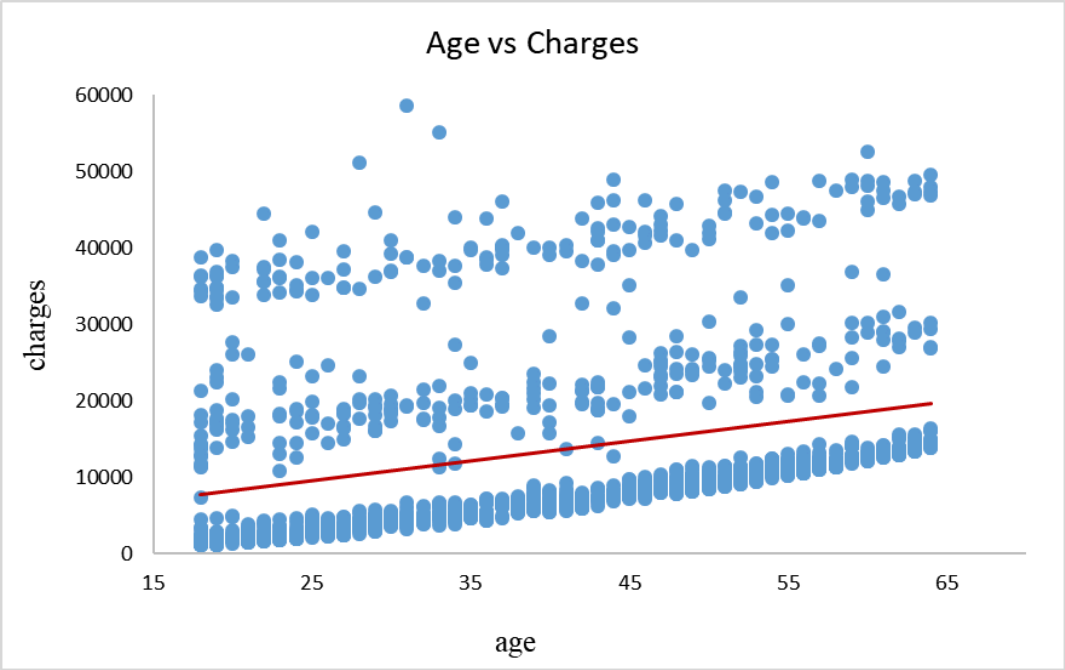

Fig. 6: Relation between age and charge

Similarly, the scatter plot for bmi and charges shows that many data fall below the trendline, which means that all the people with different bmi can have low insurance charges and the distribution is irregular. But if we follow the trendline, it seems that with the increase in bmi insurance charges increase.

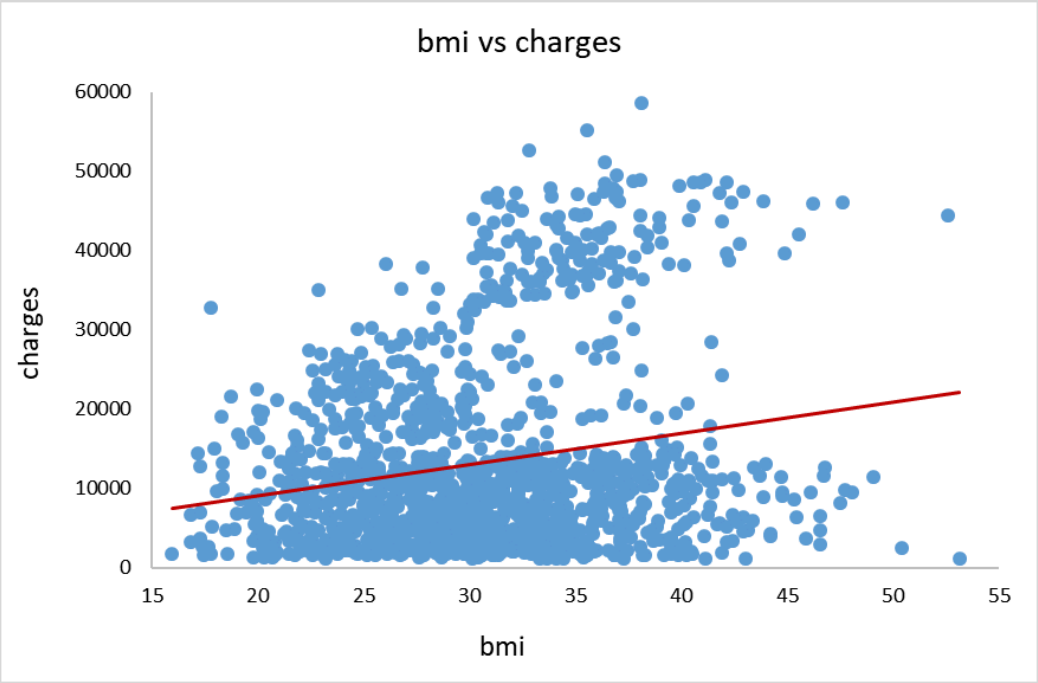

Fig 7: Relation between bmi and charge

To determine the possible relation between children and charges, here used the Correlation coefficient, and the result shows -

rxy = 0.07 and rxy>0

which indicates a positive relationship. So it can be said that with the increasing no of children, the insurance charges also increase.

By using box plot the relation between smokers and charges has been identified. it shows that the insurance charges for non-smokers is less than smoker. In fact, for smokers,s the mean is higher ($32050.23) than for non-smokers

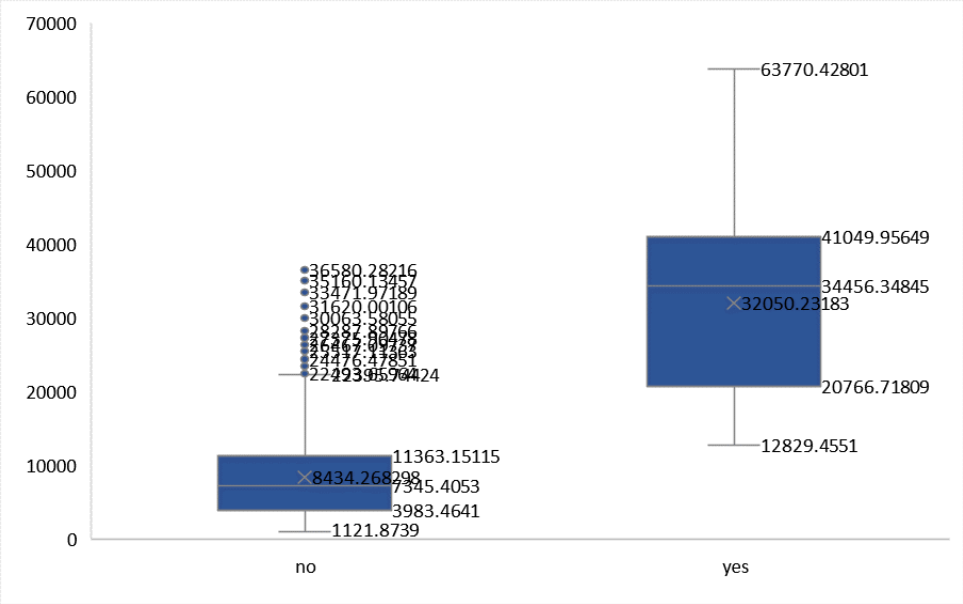

Fig.8: Box plot showing the relation between smokers and charge

Boxplot was used to determine region wise charge differences. The inhabitants of the southeast has the highest insurance charges whereas the inhabitants of southwest have the lowest charges among the 4 region.

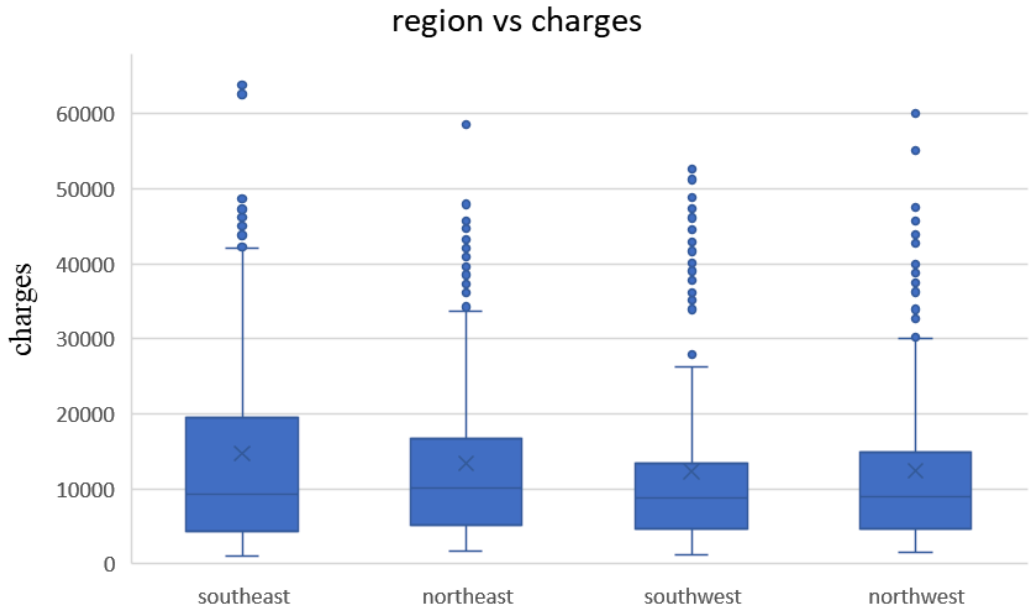

Fig.9: Relationship between the region and charges

Hypothesis testing

A hypothesis was conducted assuming that the mean charges for both male and female are same, but the test result shows that the p-value (Two-tailed) is less than the Alpha value (0.05). Hence we can reject H0 (Null hypothesis). So the analysis of sample data shows that the mean insurance charges for female subjects are different from male

Mean Charges (Male)	**13265.93**
Mean Charges (Female)	**13270.42**

Predictive Analysis

The Piecewise Regression model was developed here to predict insurance charges considering other data provided. To normalize the categorical data dummy variable was used to replace them. Using the 80/20 rule the total sample data was divided into train and test data. Developing the model on train data prediction was conducted on the train data.

Prediction Data Evaluation

To evaluate the prediction, a scatter was used to define the R^2 from the trendline.

R^2= 0.8229

It indicates that the 0.8229 or 82.29% variance of the target column charges can be well explained. The R^2 value indicates a relatively high value.

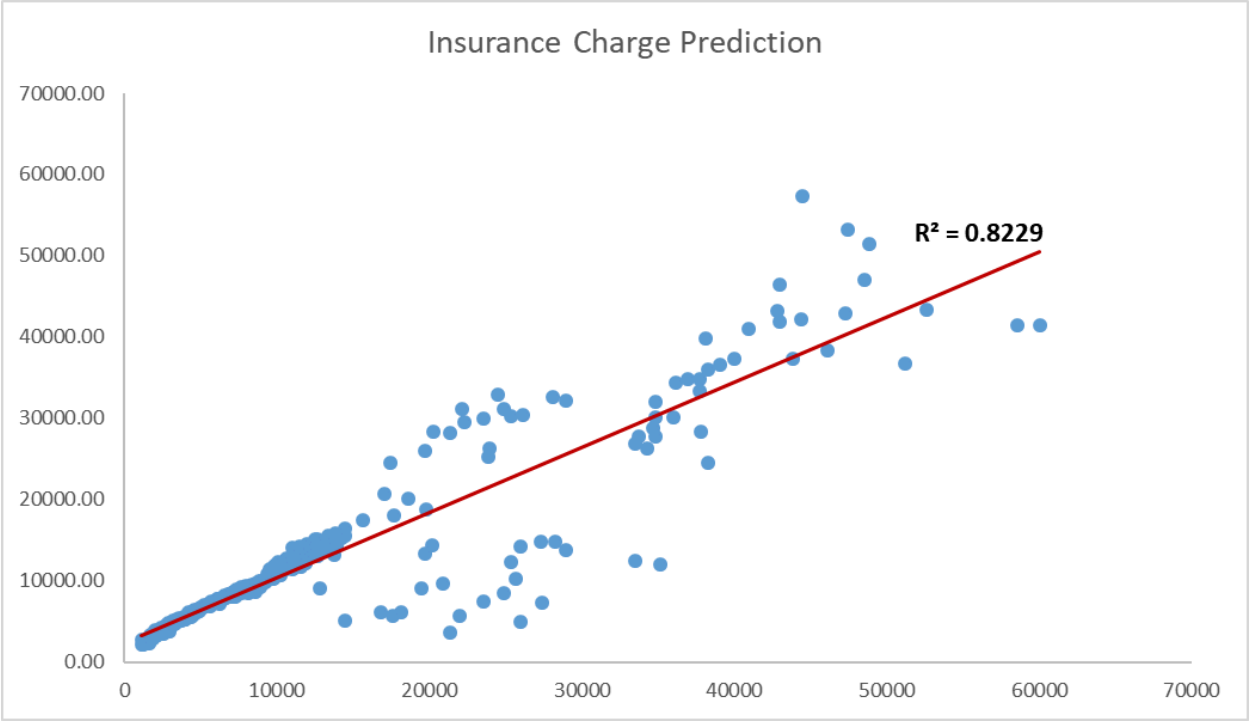

Fig.10: Insurance charge prediction for Test Dataset

To evaluate it more, calculated MSE, RMSE, MAE-

MSE	RMSE	MAE
28789222.04	5365.56	3179.85

Table 4: Evaluation of Prediction Data

Conclusion

In this study, descriptive analysis highlights age and smoking as significant factors in predicting health insurance charges. While smoking results health hazards, increasing age requires more take care of health. Other variables also has impact on charges. Finally teh piecewise regression has a R^2 of 82.29% implying the model as a good fit. The RMSE of 5365.56 suggests charges might vary around this value.

References

Brown, S., et al. (2019). Lifestyle Factors and Health Insurance Expenses: A Predictive Analysis. Journal of Risk Assessment, 25(4), 301-318.

Cutler, D. M., & Zeckhauser, R. J. (2000). *The anatomy of health insuranc*e. In A. J. Culyer & J. P. Newhouse (Eds.), Handbook of Health Economics (Vol. 1, pp. 563-643). Elsevier.

Finkelstein, A., Baicker, K., & Malani, J. (2020). *The Economics of Health Insurance. In Handbook of Health Economics* (Vol. 2, pp. 139-267). Elsevier.

Smith, J., Johnson, M., & Brown, A. (2020). *The Role of Health Insurance in Access to Healthcare*. Journal of Health Economics, 29(4), 532-545.

Wong, A., et al. (2018). Demographic and Regional Variations in Health Insurance Expenditure: A Descriptive Analysis. Journal of Health Economics, 20(3), 45-62.