

A MULTIPLE LINEAR REGRESSION MODEL FOR PREDICTING INSURANCE PREMIUMS

GROUP MEMBERS



- Brenda Mwangi
- Brian Kipchumba
- Carol Kilonzo
- Catherine Chelaga
- Francis Kyalo
- Jean-Marie Wachira
- Kelvin Njenga
- Mitchell Chege
- Whitney Ngili

AGENDA

Introduction	4
Business Context	5
Exploratory Data Analysis	7
Data Modeling	16
Insights	23
Future Improvement Ideas	24


INTRODUCTION



Adverse selection is considered one of the major risks facing the insurance sector. This is where an applicant gains insurance at a cost that is below their true level of risk due to important variables not being taken into consideration while computing the premiums.

Developing a model factoring these variables would be useful for insurance companies in reducing the risk of adverse selection.

BUSINESS CONTEXT



Problem Statement	Insurance companies face several challenges when determining insurance premiums. They must assess the risk of an insurance event occurring for each policyholder, taking into account a wide range of factors which can be a complex and time-consuming process.
Main Objective	The goal of this project is to build a predictive model that can accurately estimate insurance premiums based on a variety of input features.

SPECIFIC OBJECTIVES

To identify the most significant features in determining insurance premiums

To build a multiple linear regression model that can accurately predict insurance premiums based on input features from the Kaggle insurance premium prediction dataset

To assess the performance of the predictive model and identify potential areas for improvement

METRIC FOR SUCCESS

The project will be considered a success if the developed predictive model is able to explain 80% of the variation of the target variable.

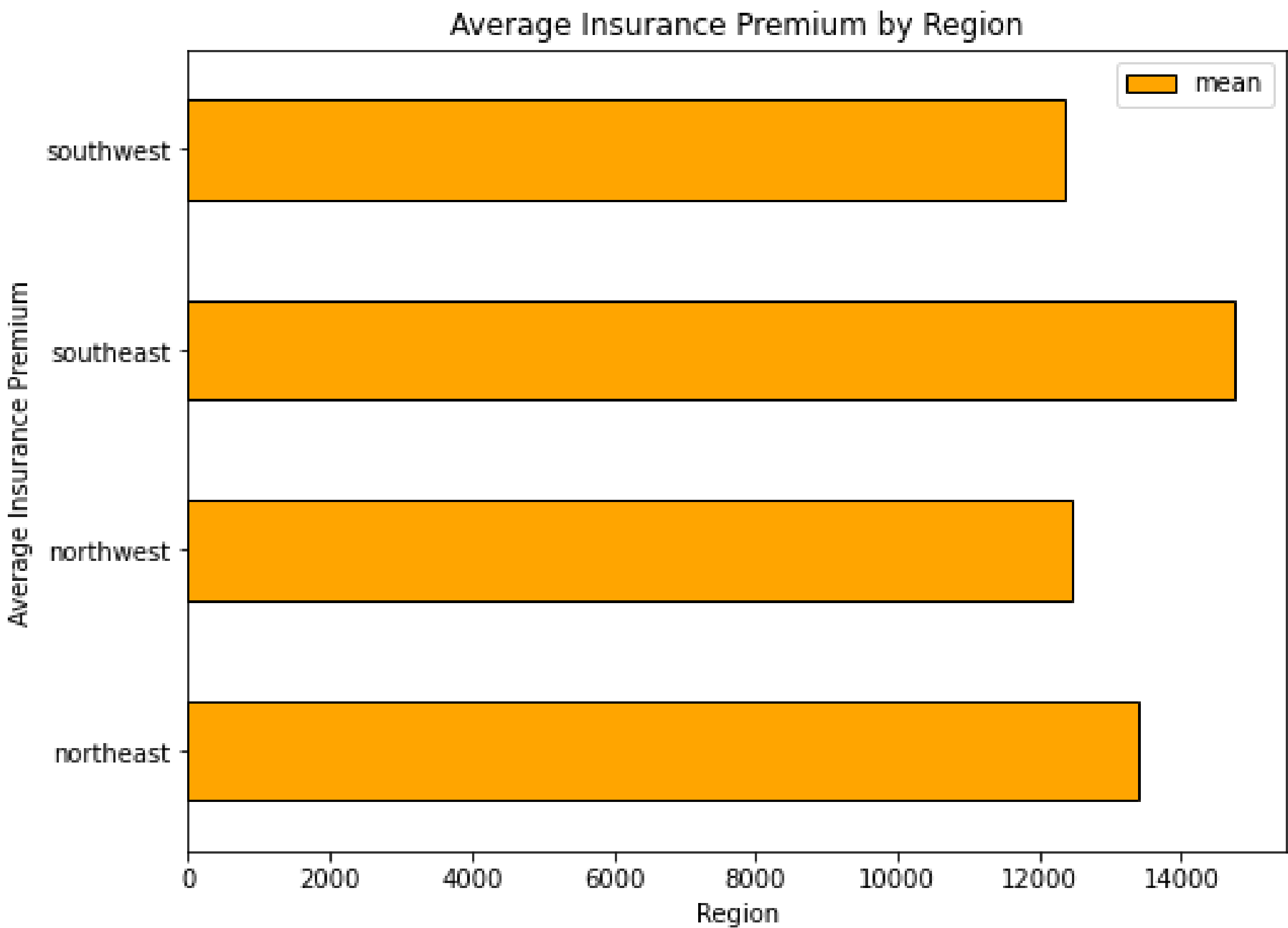
Have a Mean Absolute Percentage Error (MAPE) of not more than 10%.



EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

a. How does insurance expense change per region?

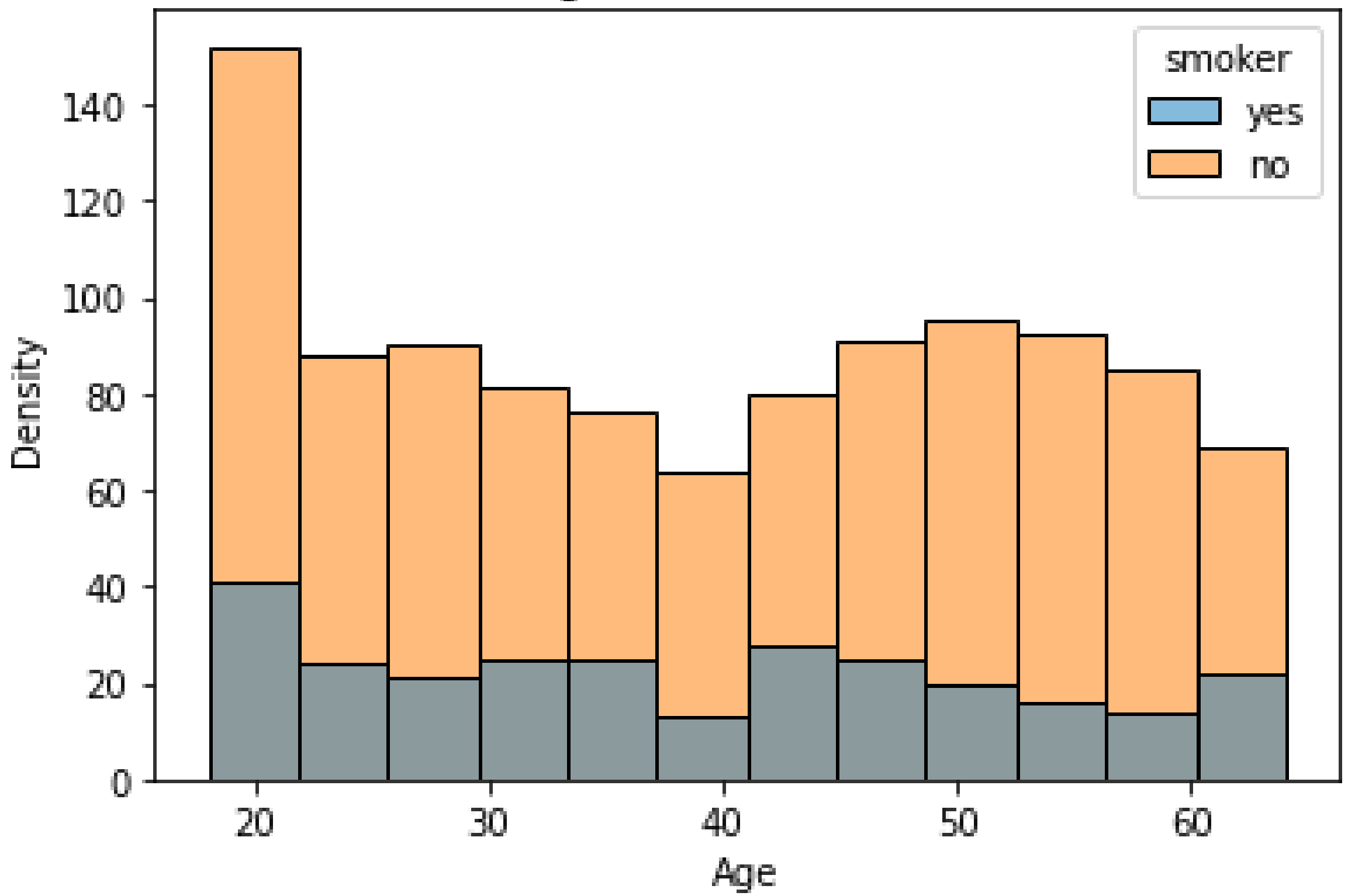


Individuals in the Southeast region make a higher contribution of about USD 14,500 while individuals in the Southwest region make contributions of USD 12,000.

BIVARIATE ANALYSIS

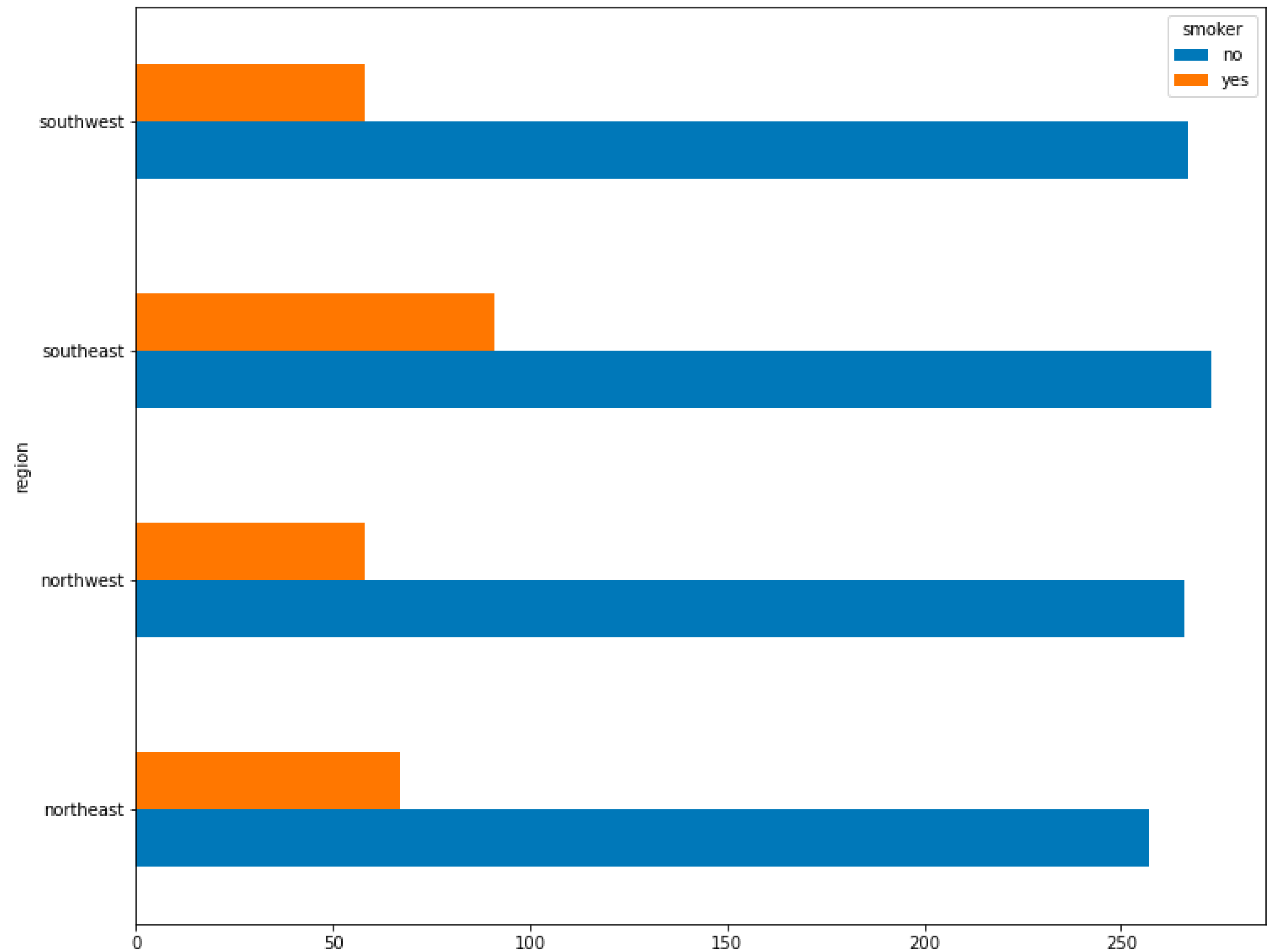
a. At what age is smoking prevalent?

Distribution of ages for smokers and non-smokers



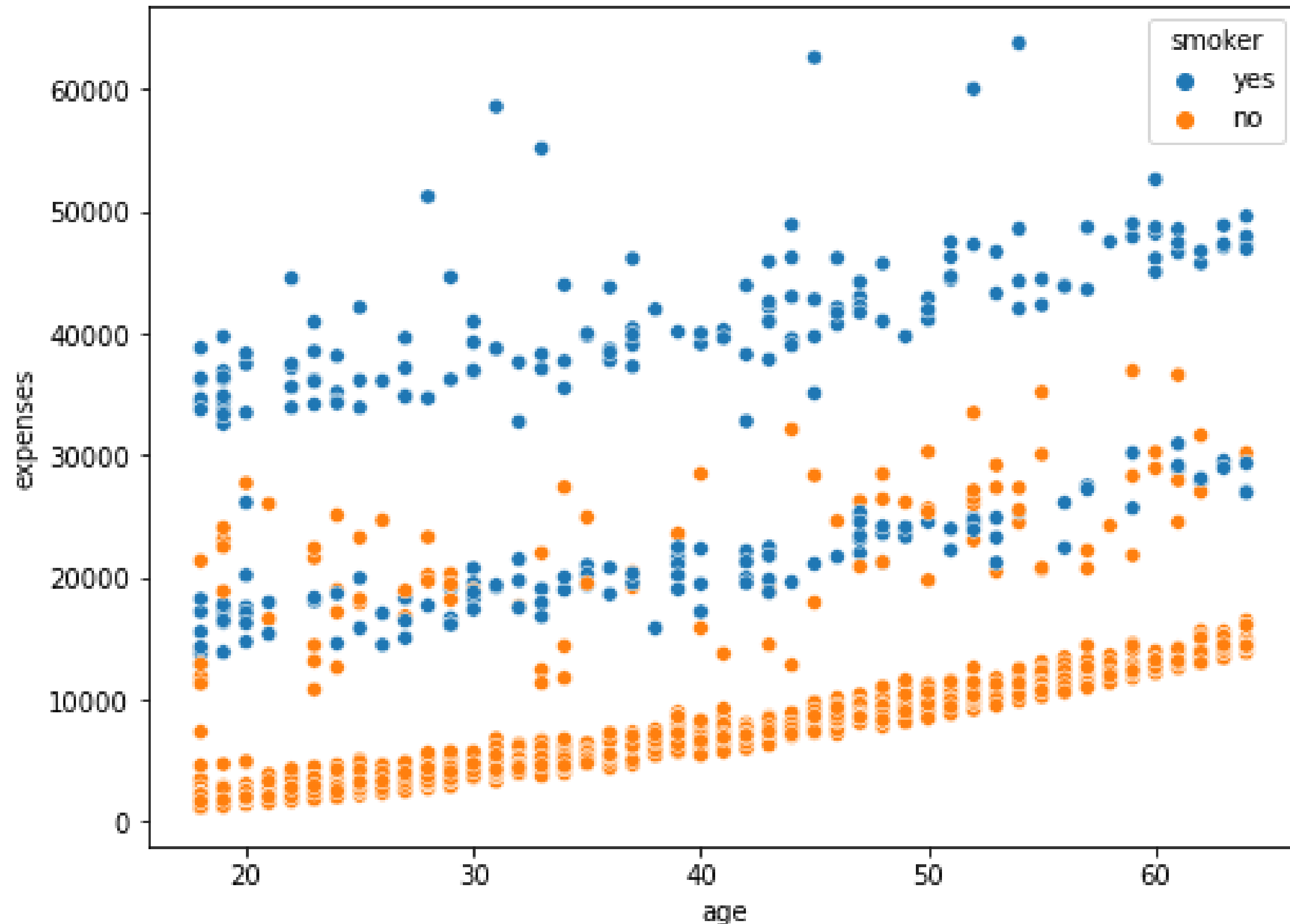
The graph shows that a larger number of smokers are 20 years old depicted with a density of 40. While the least prevalent age is 40 years old in this dataset.

b.How do smokers and Non-smokers compare to each of the regions?



The southeast region has the highest number of smokers, followed by the northeast, southwest, and lastly northwest.

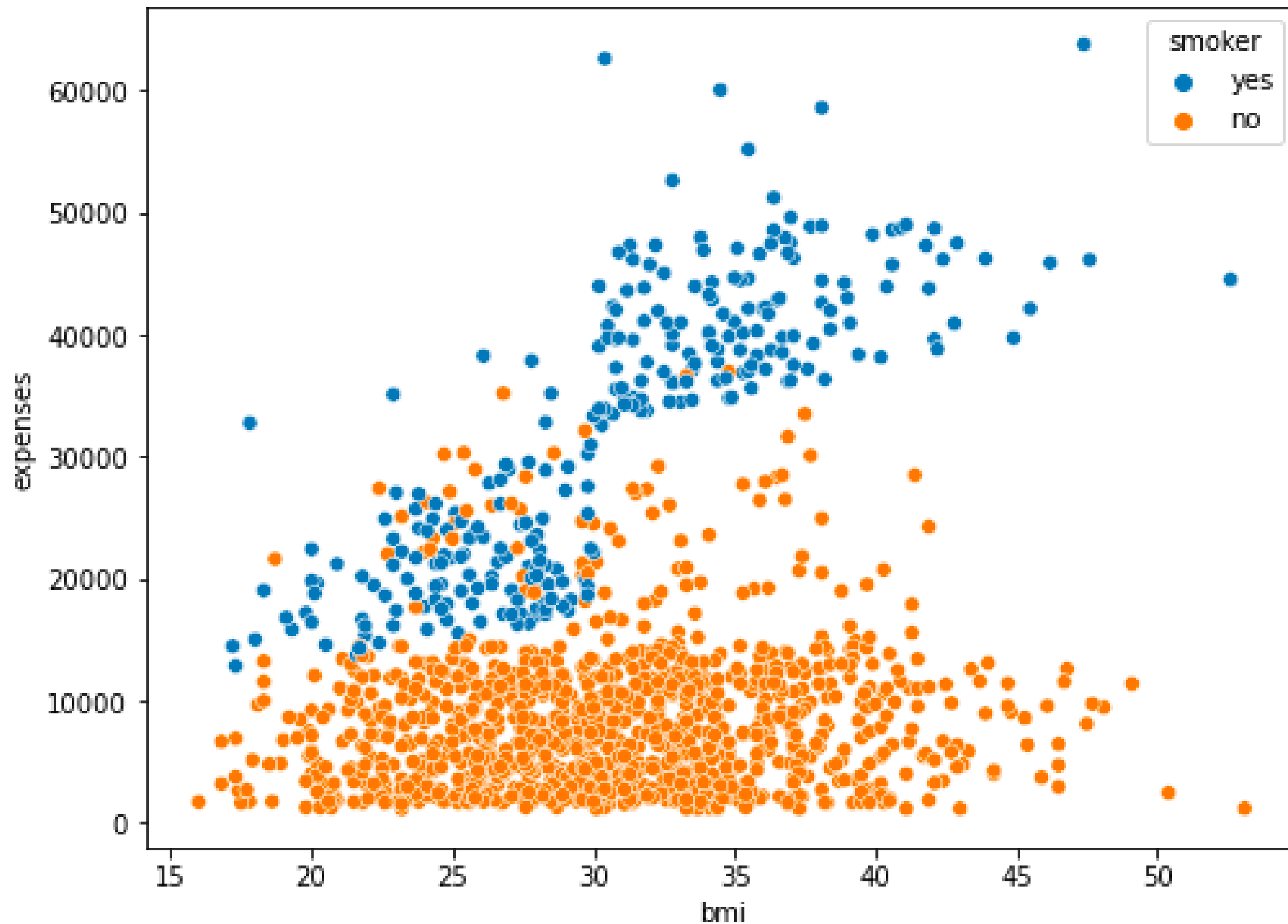
c. Depending on age, what premium will a smoker be charged?



The expenses billed to both smokers and non-smokers increases as the age increases. It is also noted that the expenses for smokers are higher when compare to non-smokers.

This can allude to the health risks associated with smoking.

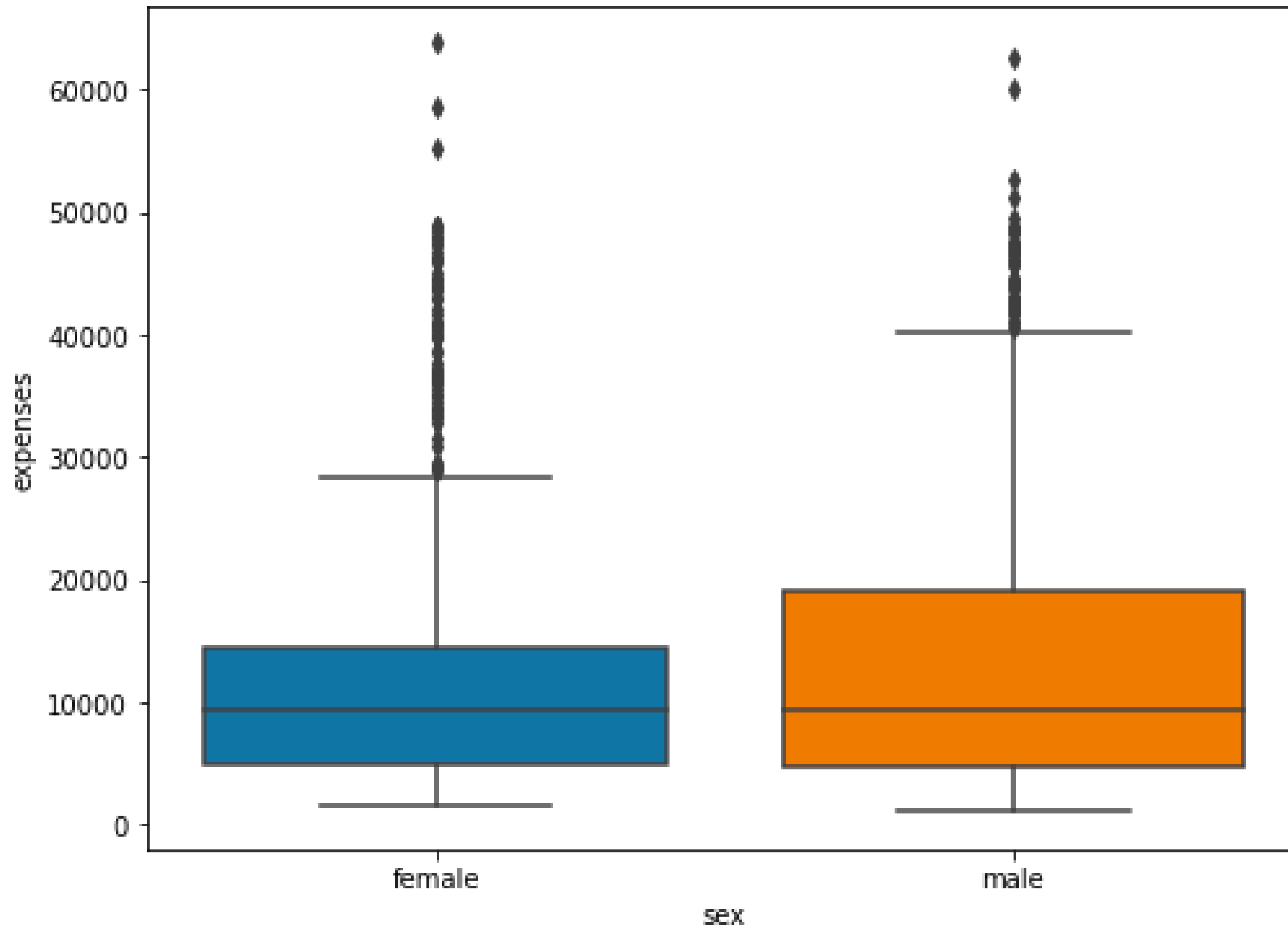
d. Does smoking and BMI have some sort of effect on how expenses are determined?



There is a positive relationship between BMI and expenses, meaning that as BMI increases, expenses also increase.

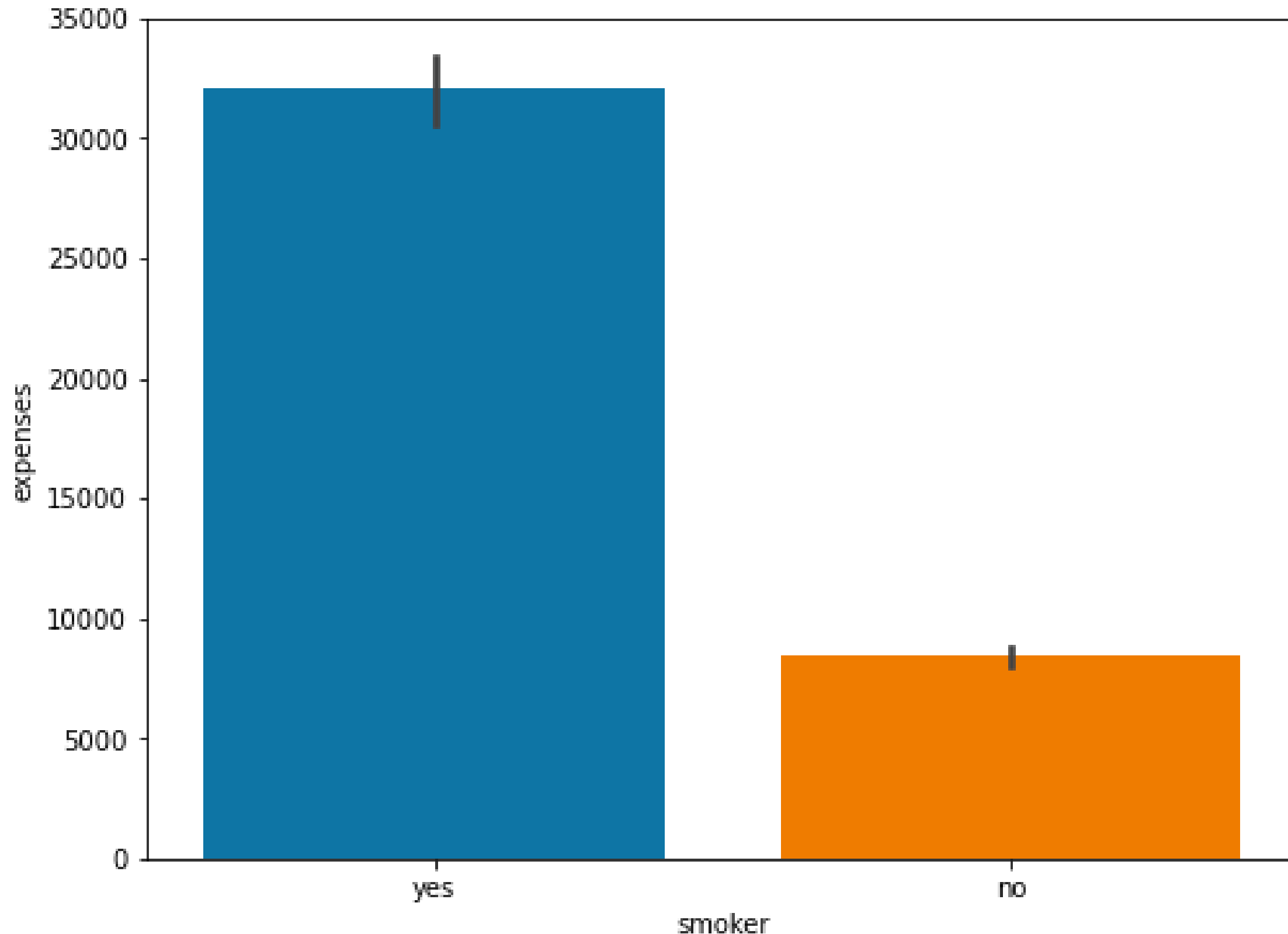
Moreover, individuals with a high BMI who smoke incur the highest expenses as shown above.

e. Does gender determine the amount of insurance expenses paid?



The box plot for males would likely have a higher median premium expense than the box plot for females, this indicates that males on average pay more than females.

f. How does being a smoker affect the amount of insurance expenses paid?

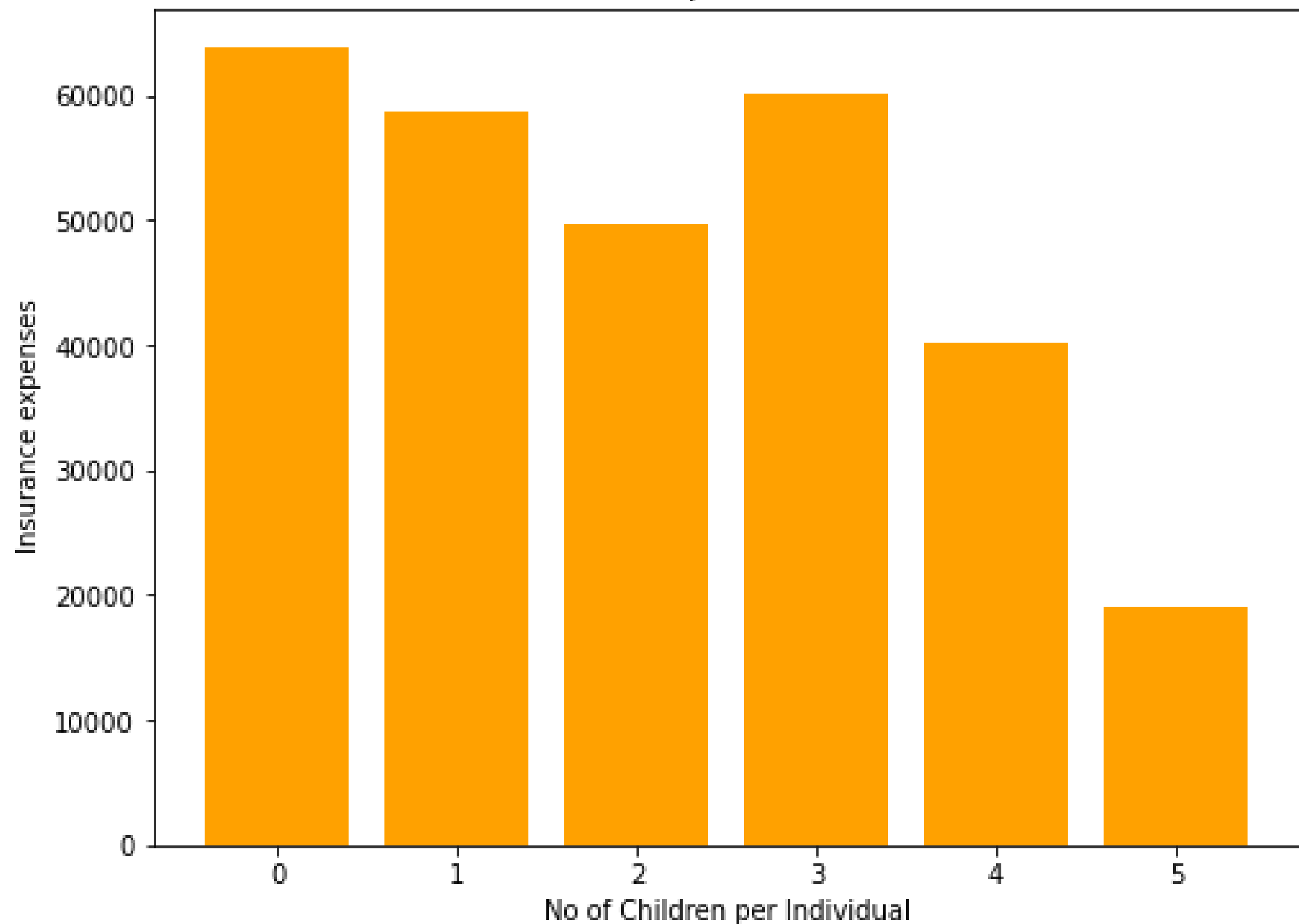


The insurance premium is relatively higher for policyholders who smoke as compared to those who do not smoke.

This can allude to the high health risks associated with smoking.

g. What is the relationship between the number of children and expenses paid?

Children vs Expenses Distribution

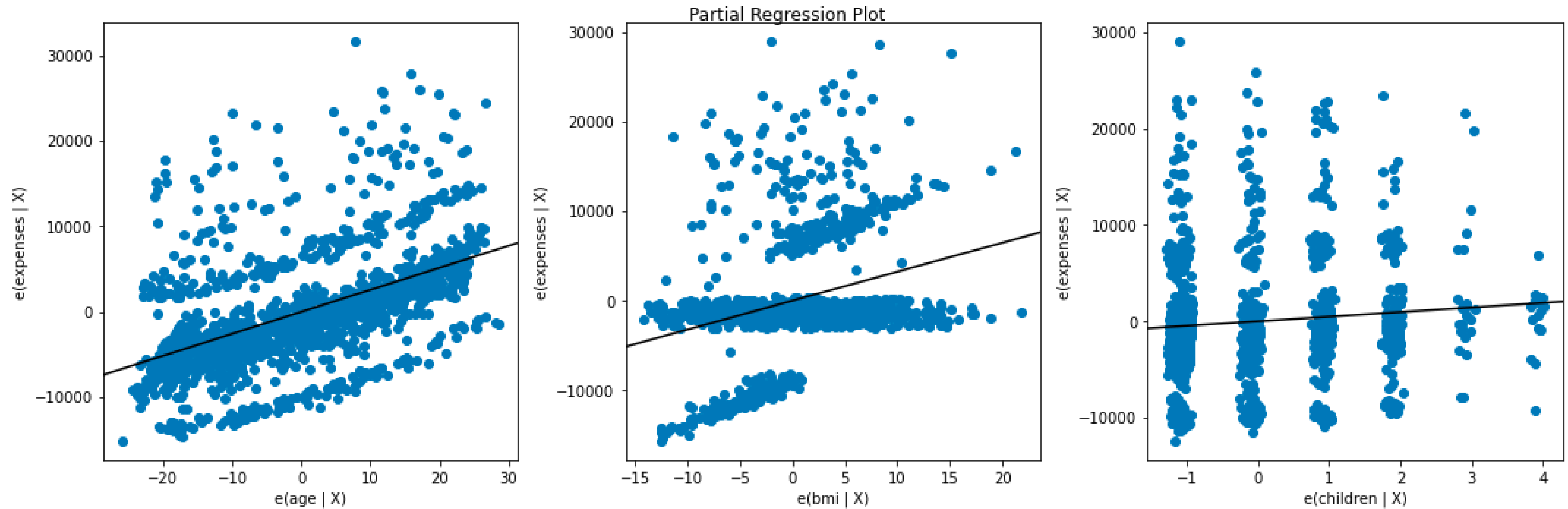


The mean amount of premium charged goes down as the number of children or dependents increases with the exception of 3 children.

DATA MODELING

MODEL 1- BASELINE

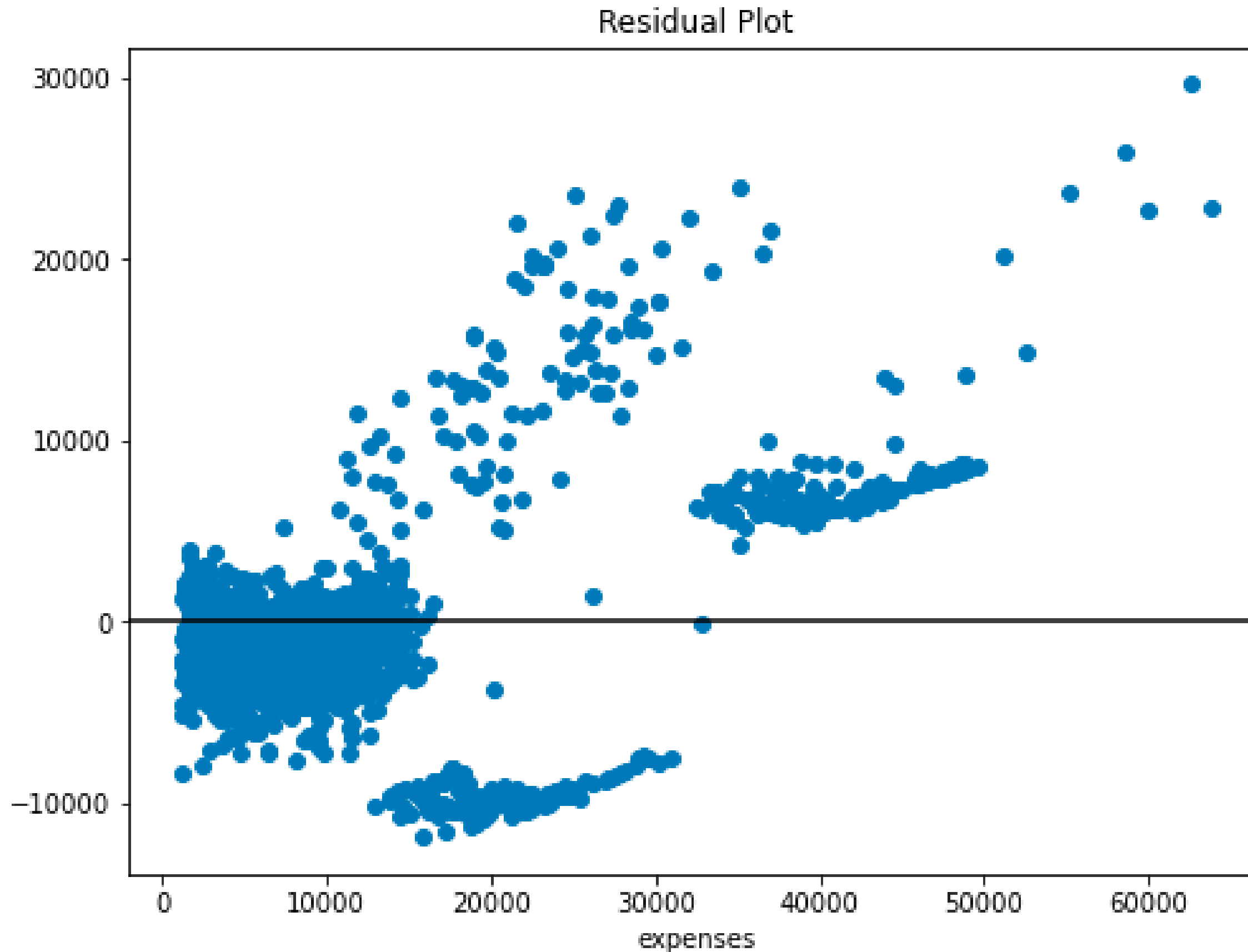
a. Model Diagnostics



The independent variables (age, bmi and children) show a linear relationship with the target variable (expenses).

This means that an increase in the independent variable will lead to an increase in the target variable and the vice versa is true.

b. Testing the linearity and heteroskedasticity

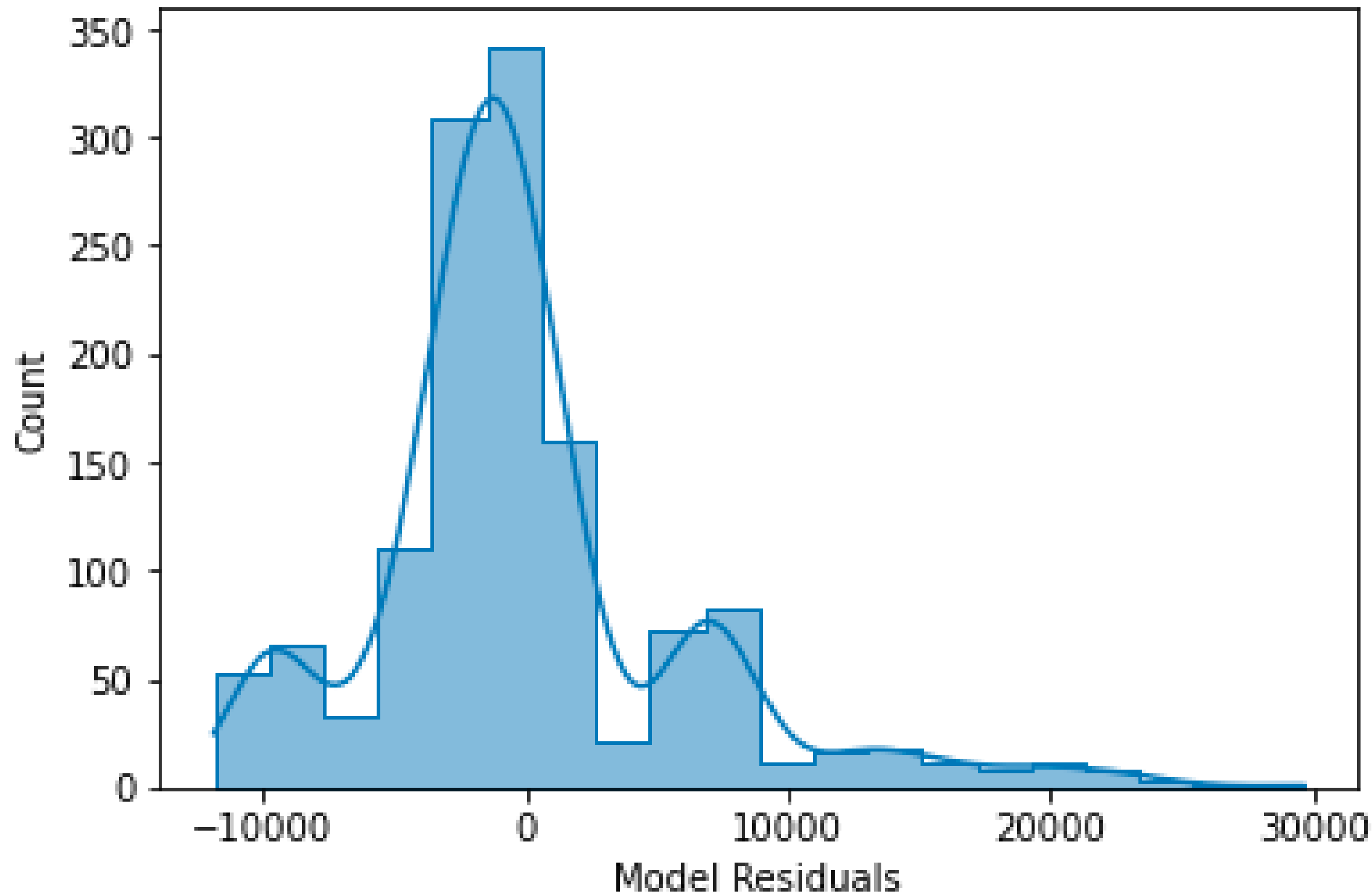


The residual plot shows that there is some level of non-linear relation between the predictors and the response variable.

The plot also shows that there is some evidence of heteroskedasticity as the residual display a funnel-shape.
The R-squared recorded was 0.748.

c. Normality assumption

Not So Normal Distribution

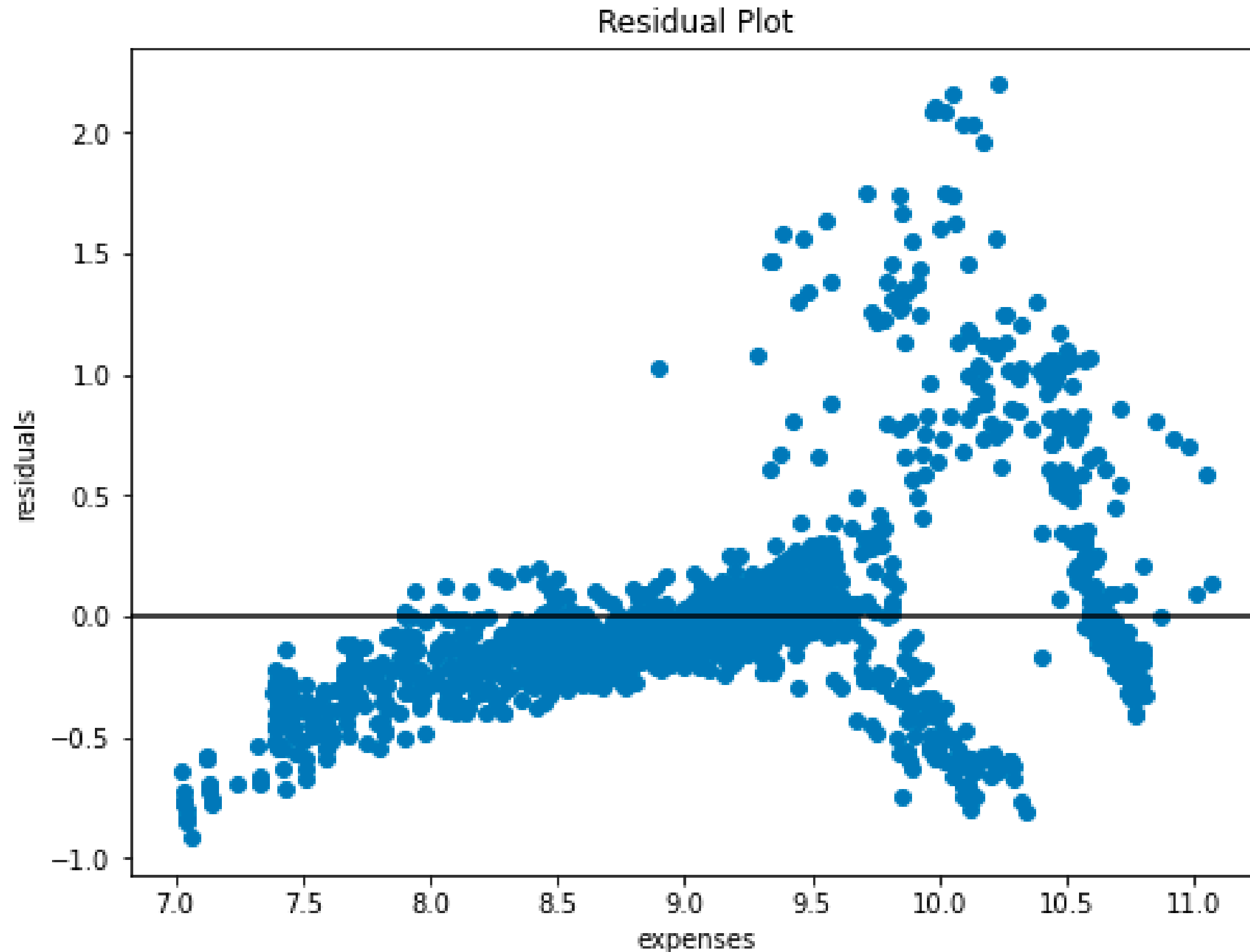


There exists a non-linear relationship between the predictor and target variables.

Therefore, to normalize this, log transformations are carried out in the next steps.

MODEL 2 - LOG-TRANSFORMED

a. Testing the linearity and heteroskedasticity

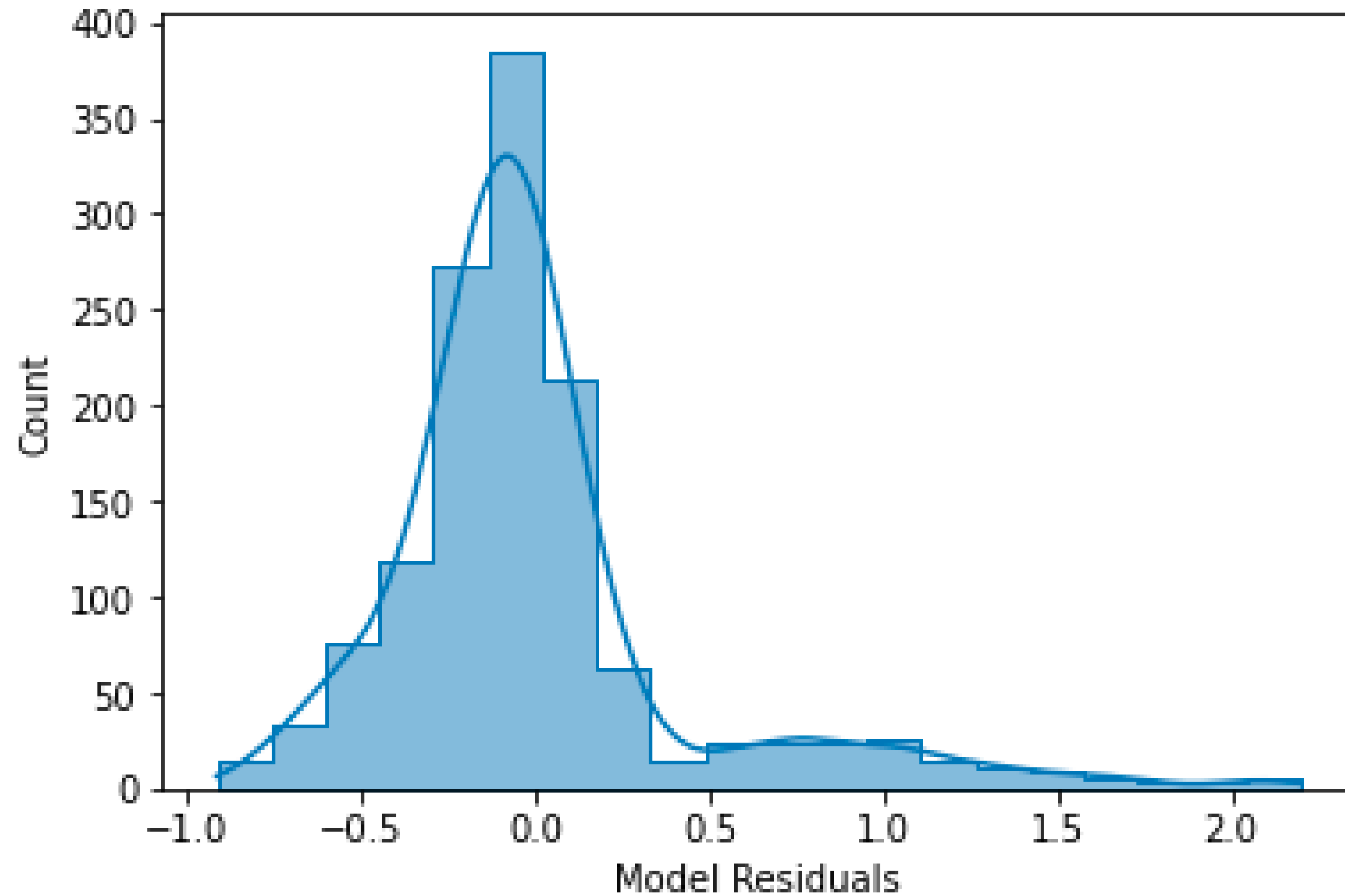


After conducting log transformations on the x and y variables, it was noted that the residual plot improved due to the reduction in the scatter of the residuals.

The R-squared increased to 0.764 from 0.748

b. Normality assumption

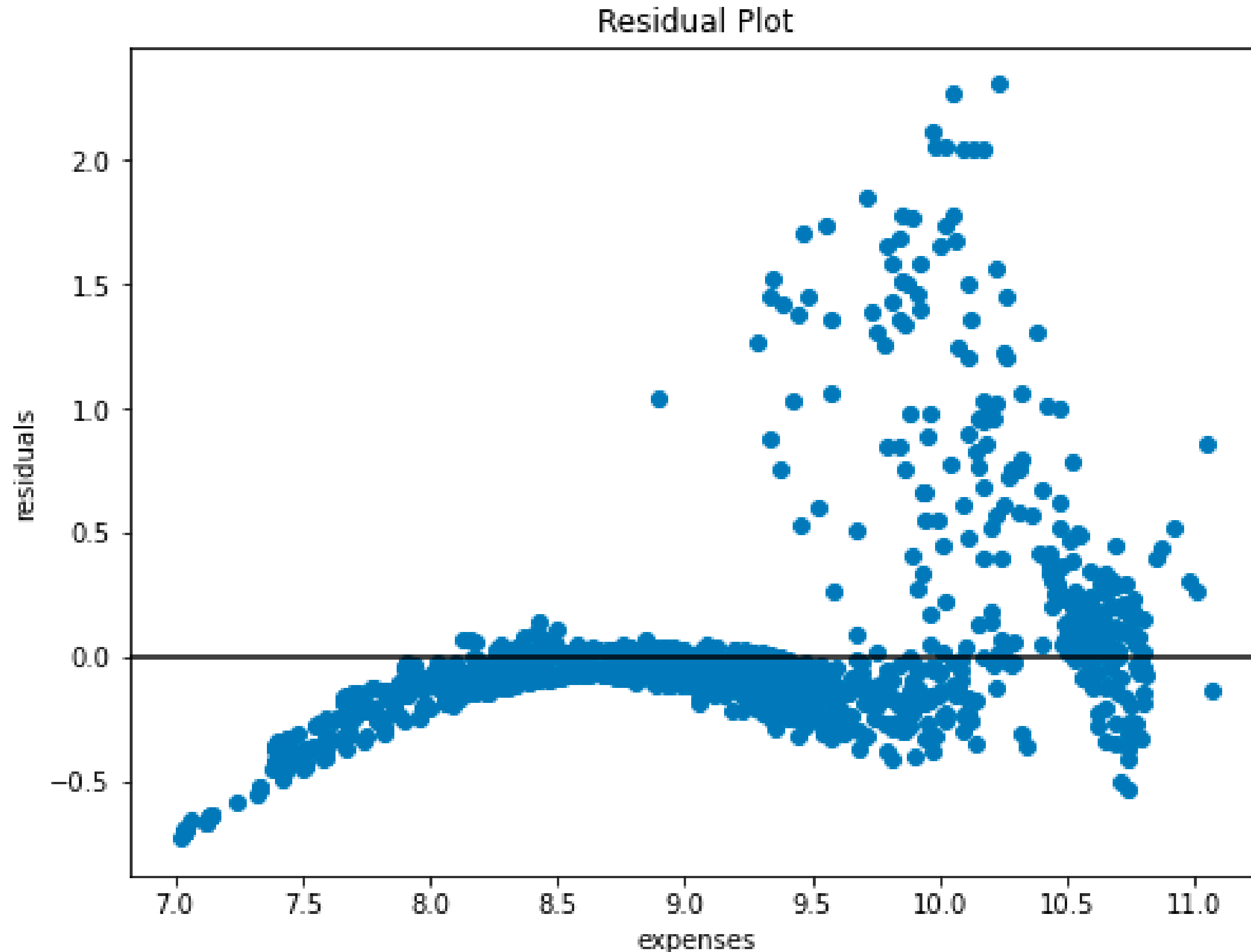
Normal Distribution



After conducting the log transformation, it can be noted that the residuals are now relatively normally distributed.

MODEL 3 - INTERACTION TERMS

a. Testing the linearity and heteroskedasticity

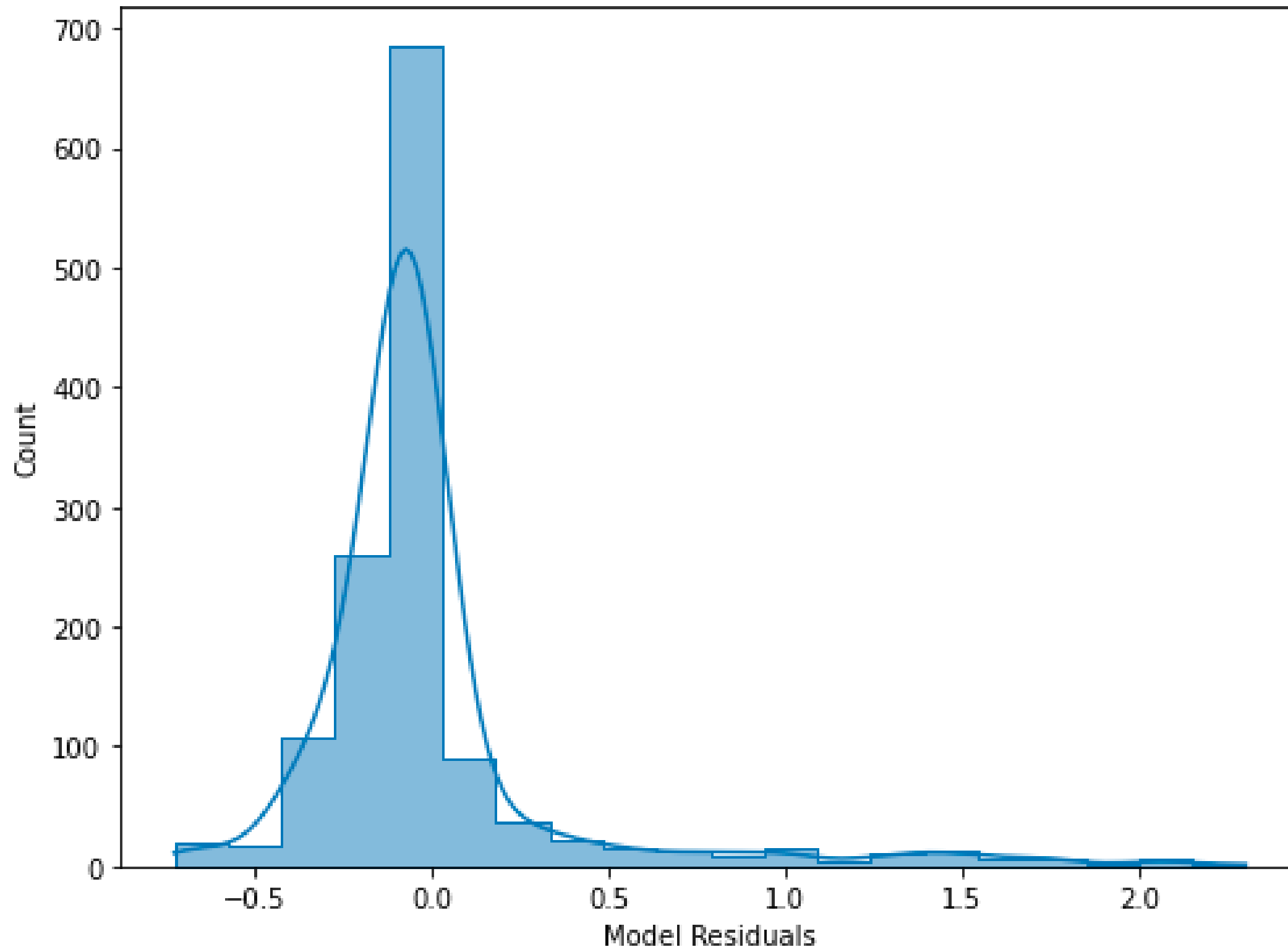


The heteroskedasticity has reduced significantly after performing log-transformation of the target and including interaction terms as the residuals seem to be concentrated at in the middle and the earlier funnel shape has also significantly reduced.

The R-squared increased to 0.821 from 0.764.

b. Normality assumption

Normal Distribution of Residuals



The histogram of the residual plot shows the residuals are normally distributed and the inclusion of a log-transformed target and interaction terms has provided insights into the non-linear relationships in the model.

INSIGHTS



MODEL 1

The Baseline Model has a MAPE of 42% and an accuracy of 75%, its performance is relatively low compared to the metrics set however, it does not overfit or underfit.

MODEL 2

The log-transformed model has an MAPE of 23% and an accuracy of 76.4% its performance is also low compared to the set metrics. This still wasn't a perfect fit.

MODEL 3

The interaction term model has a MAPE of 10% and an accuracy of 82.1% its performance is high as it meets the set metrics. Therefore it's the best one to use in the prediction of insurance premiums.

RECOMMENDATIONS & FUTURE IMPROVEMENT IDEAS

An insurance company should consider:

1

Implementing a higher premium for smokers, as the data indicates that they have a higher expected expense

2

Taking into account the number of children when setting premiums, as this variable had a positive correlation with expenses

3

Offering discounted premiums for policyholders from the SouthWest and Northwest regions, as they had lower expected expenses compared to those from the SouthEast region

4

Collecting more data on other factors such as medical conditions, previous claims, lifestyle, and occupation to improve the accuracy of their pricing model.

RECOMMENDATIONS & FUTURE IMPROVEMENT IDEAS

An insurance company should consider:

5

Offering discounts for individuals who maintain a healthy BMI, as this variable was negatively correlated with expenses

6

The age of a policyholder when setting premiums, as this variable had a positive correlation with expenses

7

The policyholders sex when setting premiums, as this variable had a positive correlation with expenses for males

8

Using the developed model as a guide for pricing premiums, as it had a low MAPE of 10% which indicates that it is a good model for predicting expense.

ANY QUESTIONS?



Presented By Group Naruto

THANK YOU!