

PROJECT OWNER:
AMADO DE JESÚS VÁZQUEZ ACUÑAS

INSURENCE

[HTTPS://GITHUB.COM/JESUS-
VAZQUEZ-A/INSURENCE](https://github.com/jesus-vazquez-a/insurence)

BY: İCLAL ŞEYMA KOCA

COPYRIGHT

16:17 4 Mar, Cts

Amado de Jesús Vázquez Acuña
Mobile • 10sa önce

Amado de Jesús Vázquez Acuña
Estudiante en UVEG Guanajuato

PERŞEMBE

Iclal Seyma Koca • 23:27
Hello Amado de Jesús Vázquez Acuña,
I am working for a boot camp. While working on data analysis
projects, we saw your project on GitHub. ([https://github.com/
Jesus-Vazquez-A/Insurence](https://github.com/Jesus-Vazquez-A/Insurence)) And worked on it. Can I get your
permission to publish the subject, content, and outputs of this
project, together with your name, via Bootcamp?

**Jesus-Vazquez-A/
Insurance**

We solve a regression problem in which it consists
of calculating the health insurance charge in the
United States Where...
A 1 0 3 3
Contributor Issues Stars Forks

GitHub - Jesus-Vazquez-A/Insurance: We solve a
regression problem in which it consists of calculating t...
github.com • 4 dakikalık okuma

CUMA

Amado de Jesús Vázquez Acuña • 05:30
Hello Iclal , of course you can use it would be a pleasure
(Düzenlendi)

Bir mesaj yazın...

0



INSURENCE

THE PROJECT AIMS TO DETERMINE,
ESTIMATE
THE INSURANCE PRICE FOR THIS
PERSON, CONSIDERING DIFFERENT
PARAMETERS SUCH AS PEOPLE'S AGE,
BODY MASS INDEX, ETC.

LIBRARIES USED FOR THE PROJECT

Pandas

Matplotlib

Numpy

Shap

Numpy

Warnings

Seaborn

CONTENT

EDA

FEATURE
ENGINEERING

SELECTION
OF MODEL
IDEAL

DEFINITIVE
MODEL

1.EDA

IMPORTANT HEADINGS IN DATA

Age

Smoker

Sex

Region

BMI Body Mass Index Charges

Medical Problem

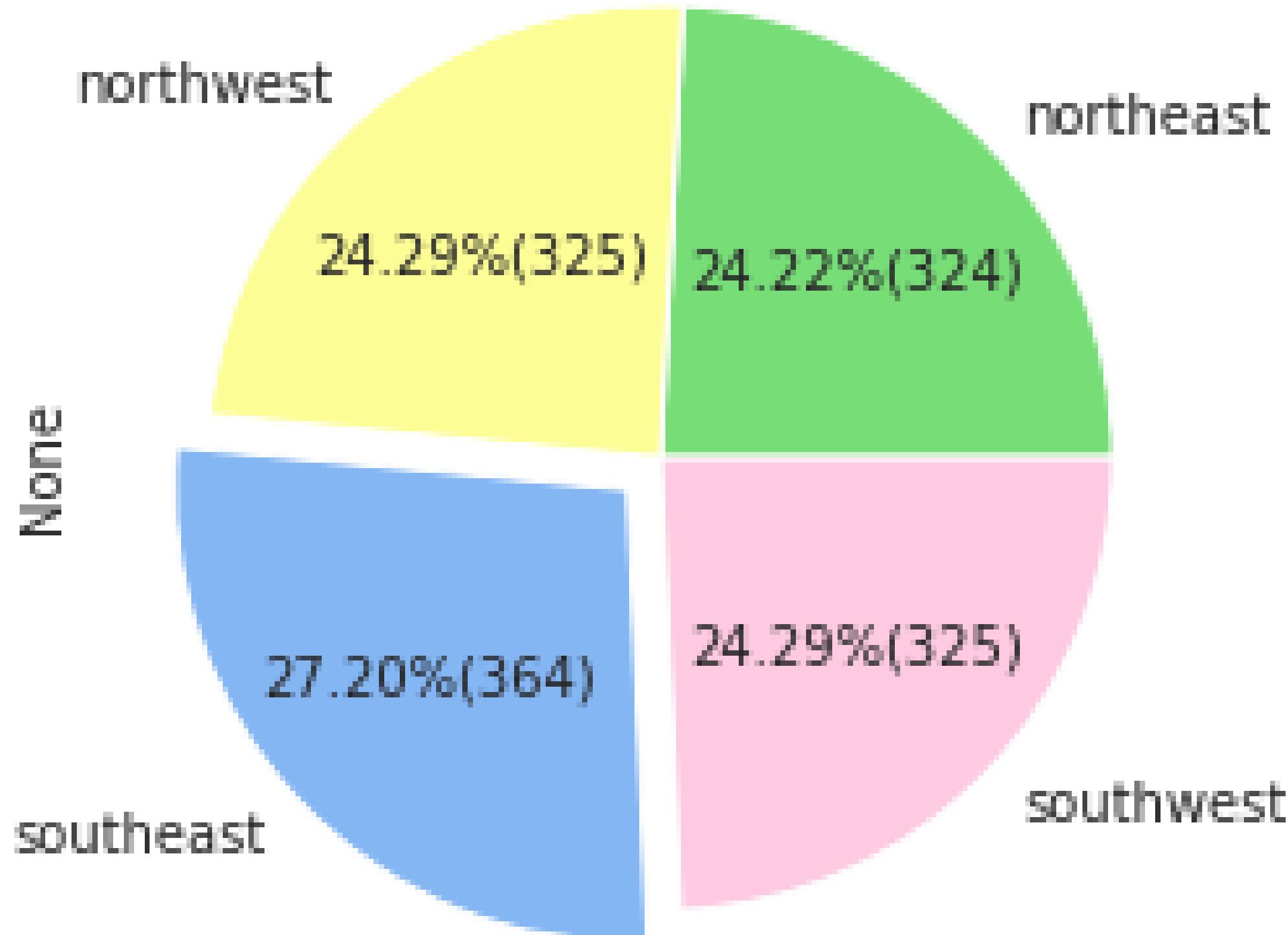
Children

EXPLORATORY DATA ANALYSIS (EDA)

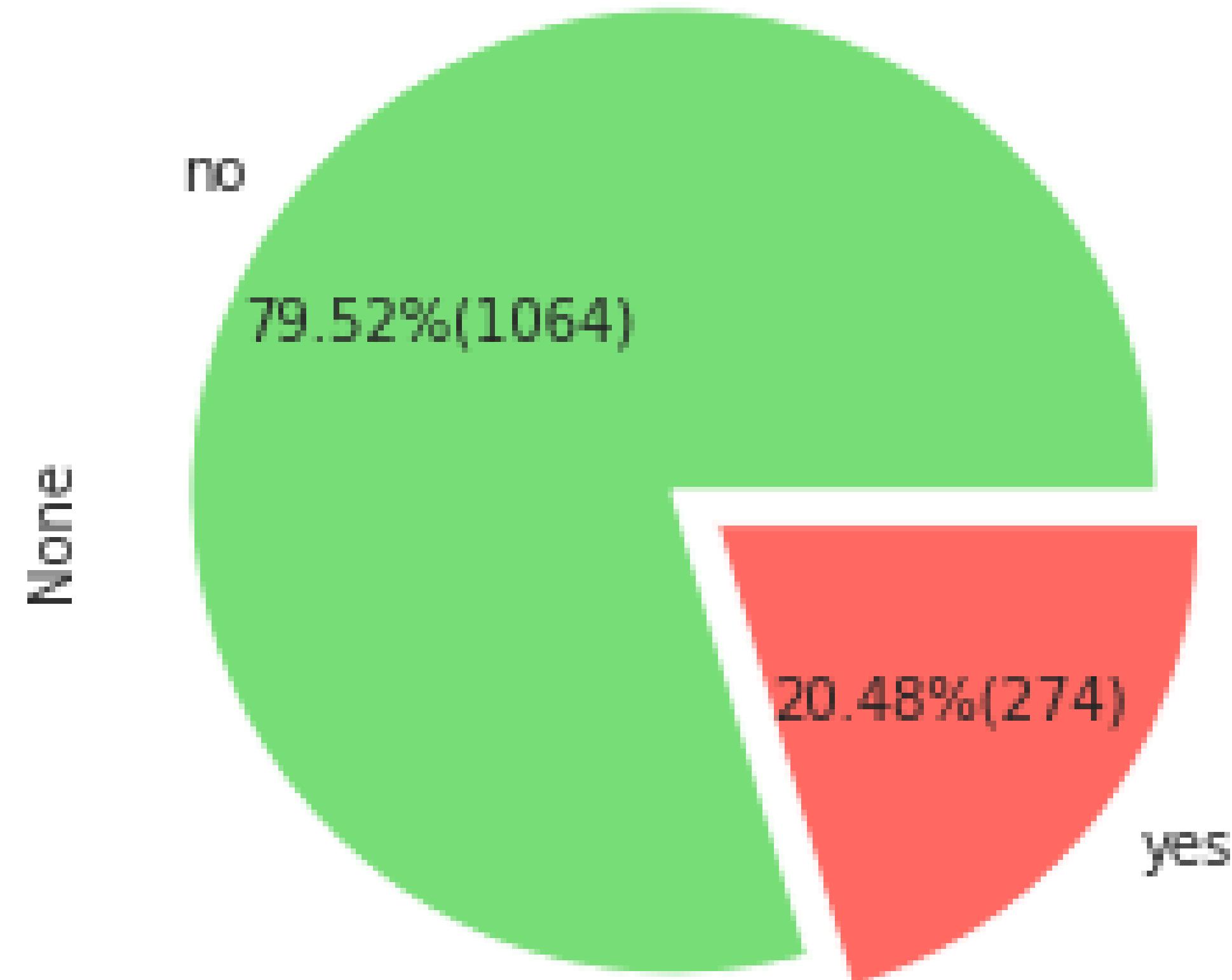


Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies, test hypotheses and check assumptions with the help of summary statistics and graphical representations.

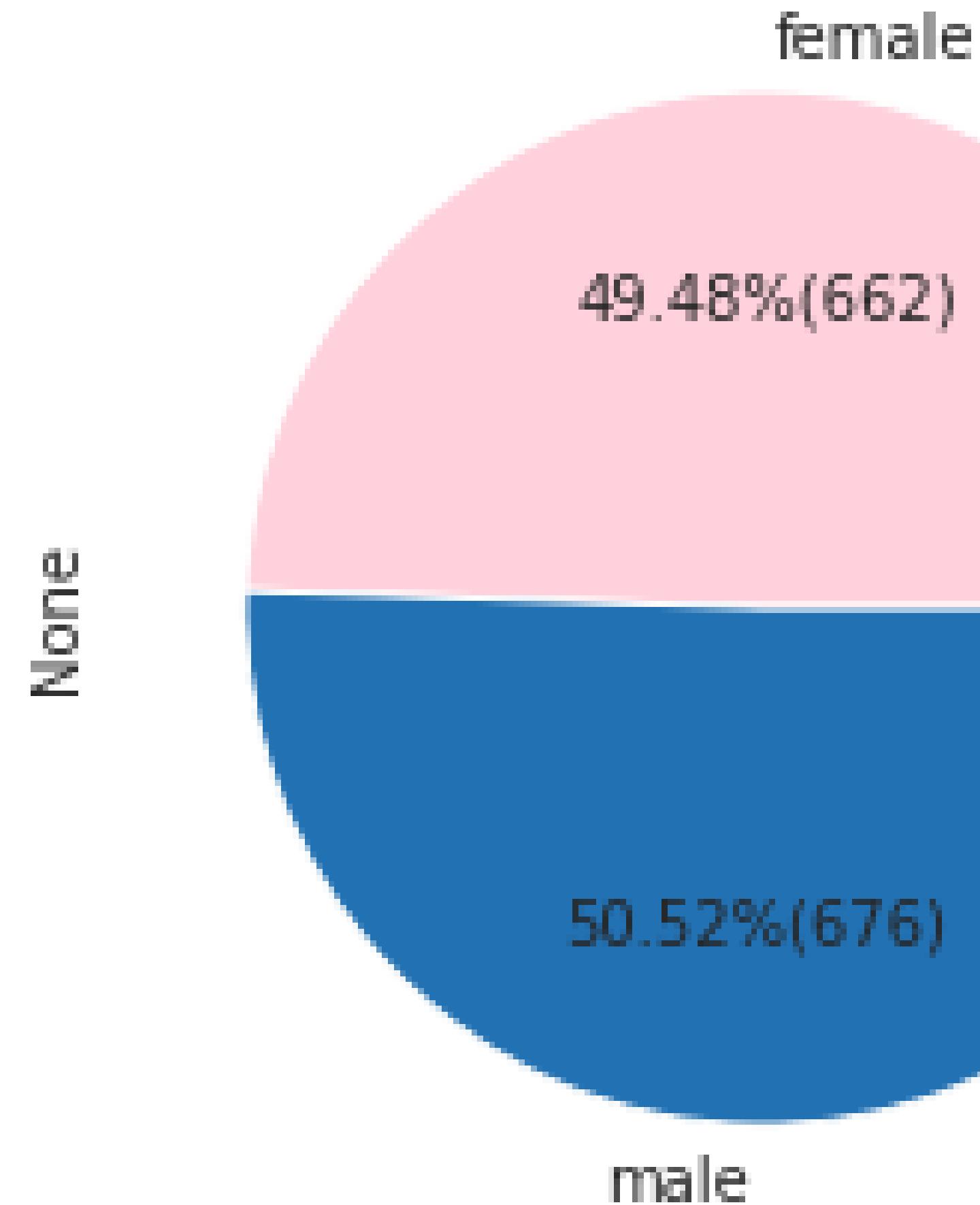
Region Percent



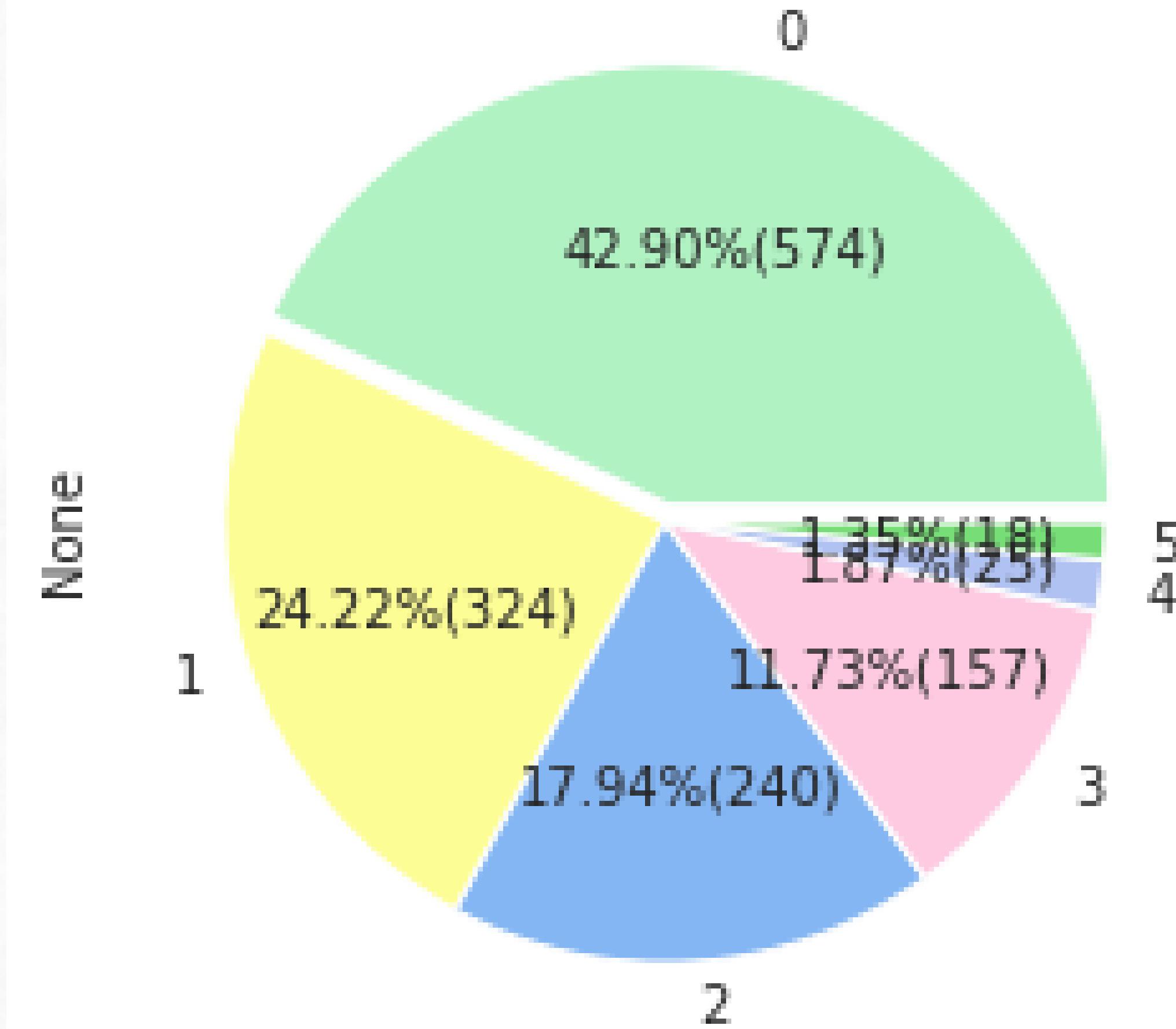
Smoker Percent



Sex Percent



Children Percent



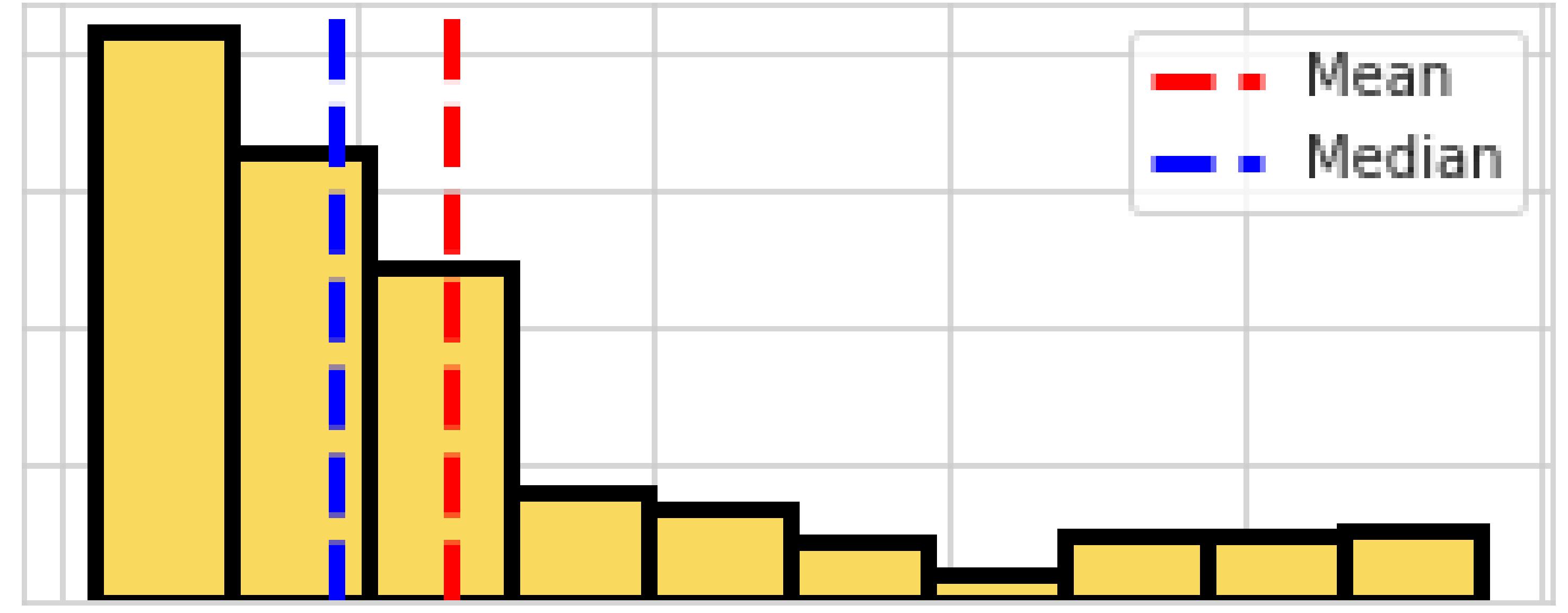
Charges

NUMBER OF PEOPLE

400
300
200
100
0

0 10000 20000 30000 40000 50000

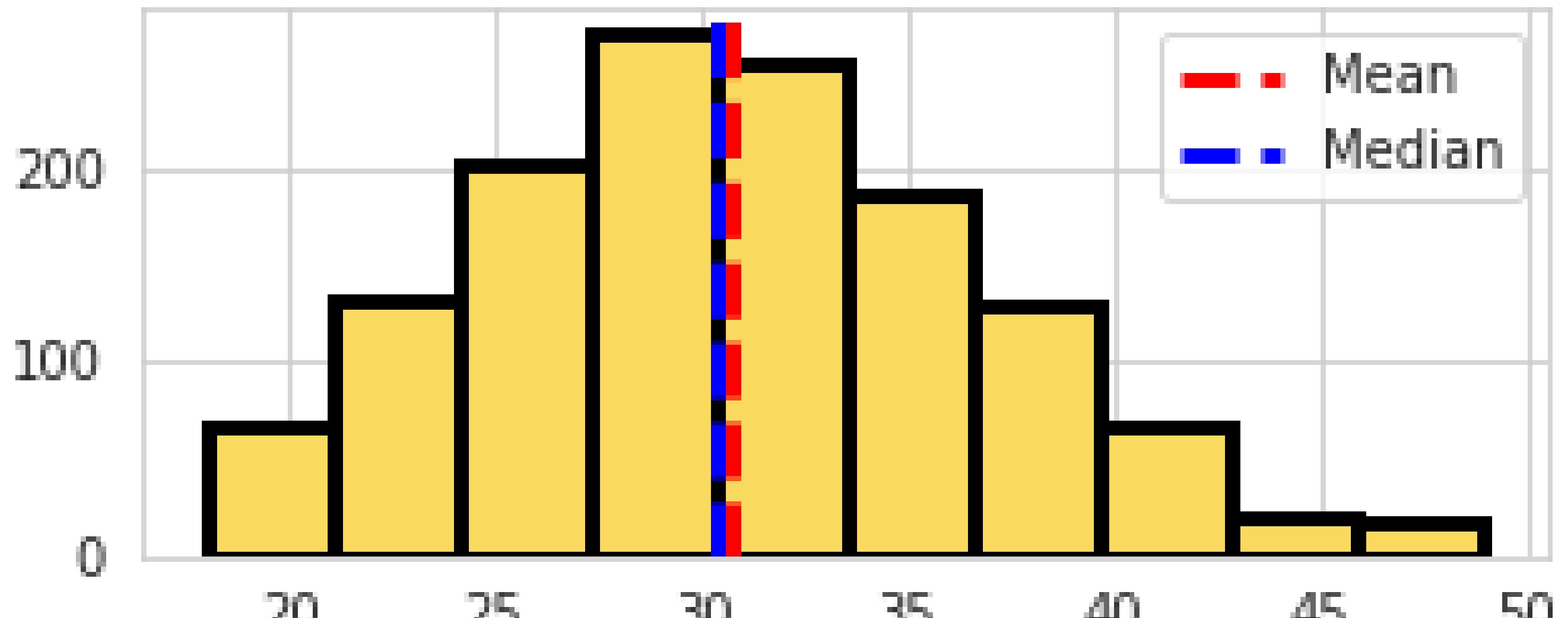
CHARGES



Mean
Median

NUMBER OF PEOPLE

BMI



BODY MASS INDEX

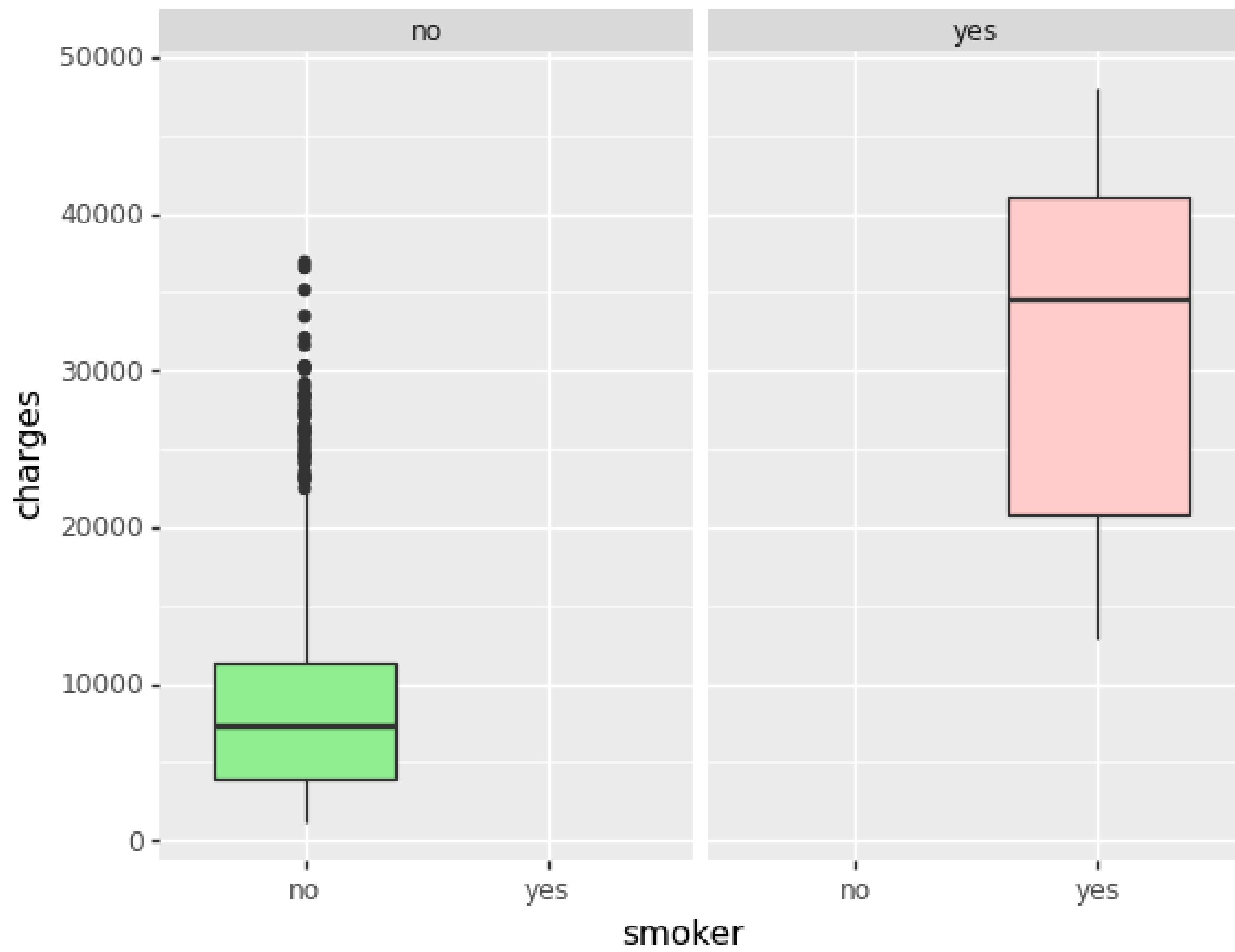
HISTOGRAM



A histogram is a graph showing frequency distributions.

It is a graph showing the number of observations within each given interval.

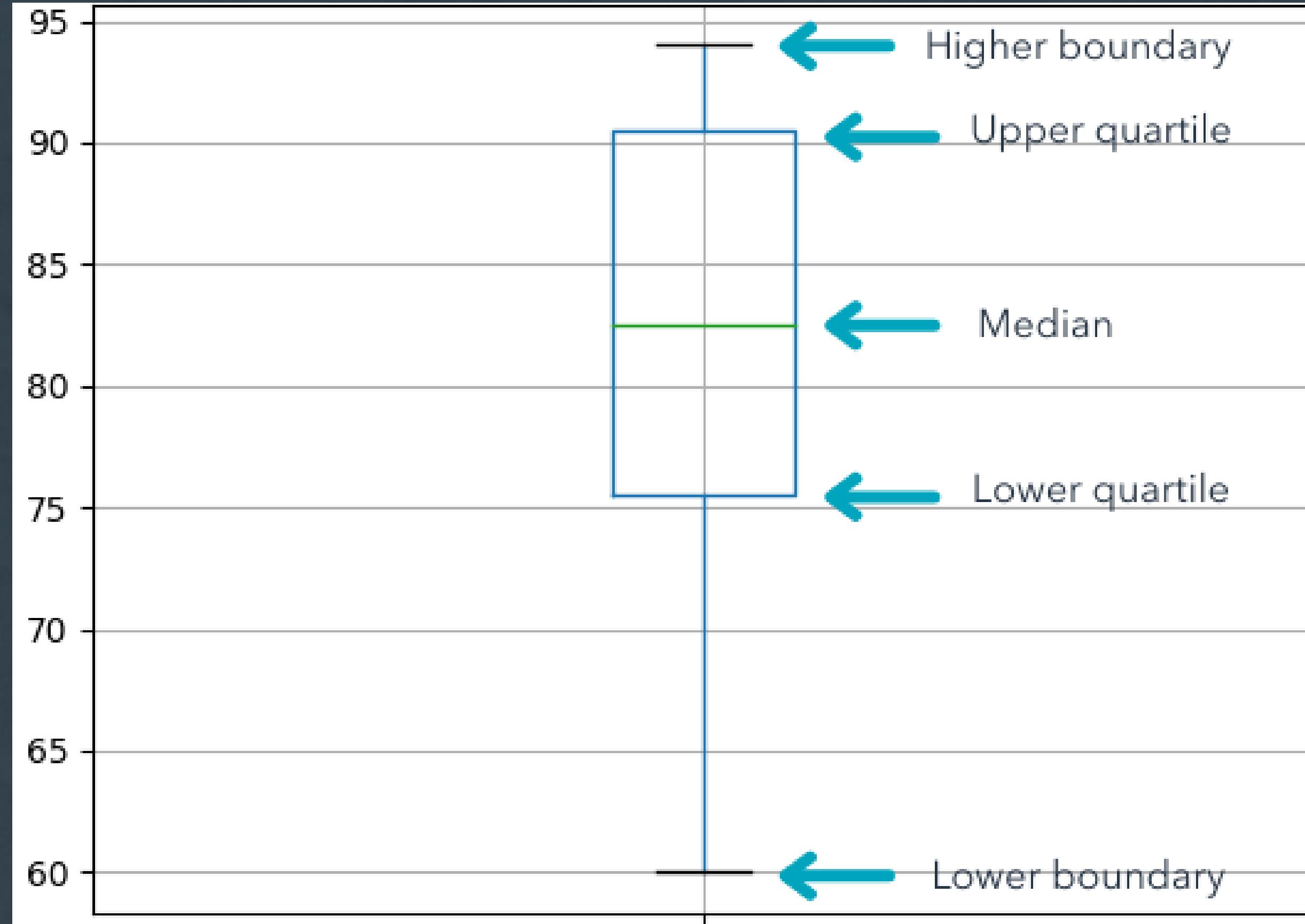
Smoker vs Charges



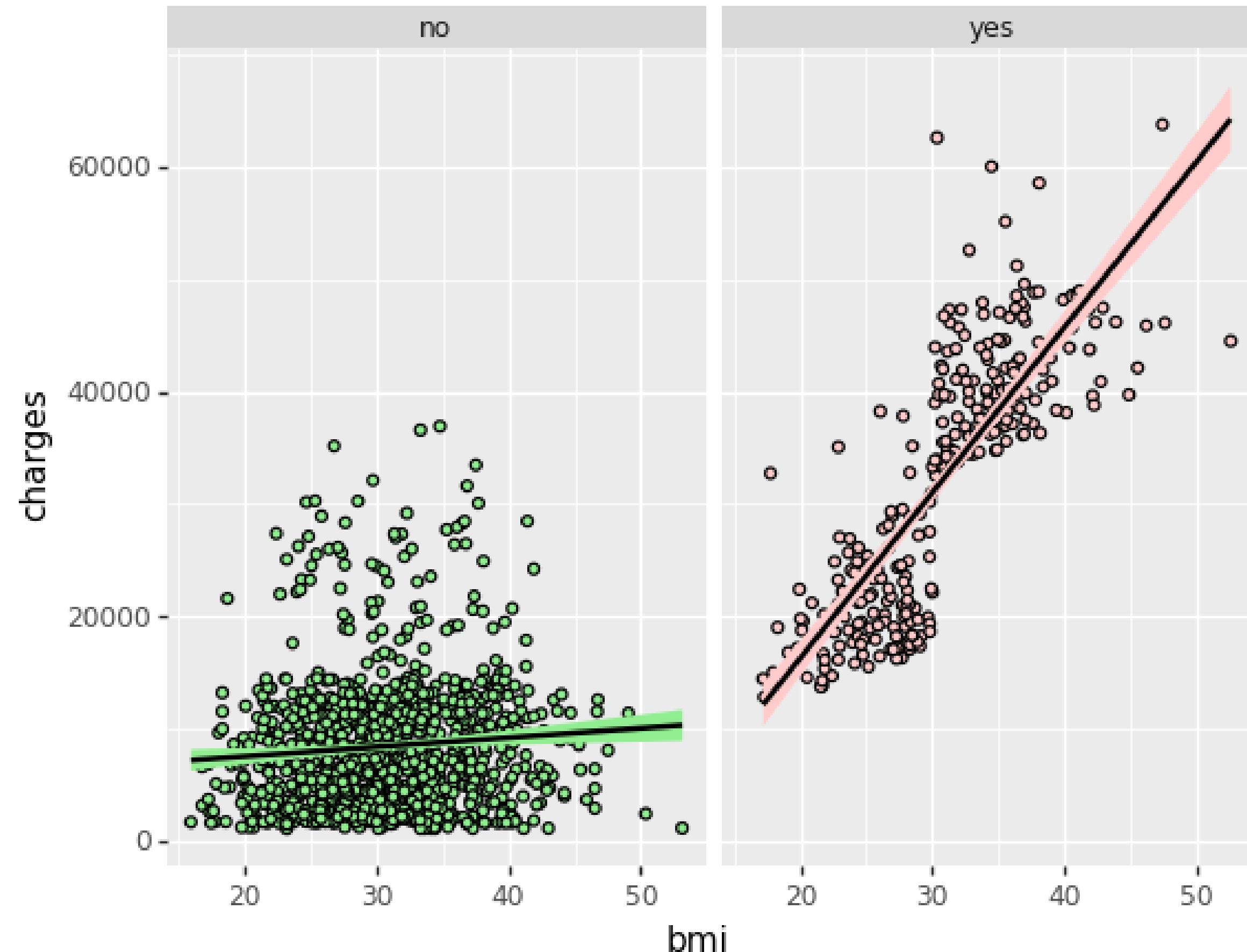
BOXPLOT



Boxplots are a measure of how well distributed the data in a data set is. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile, and third quartile in the data set. It is also useful in comparing the distribution of data across data sets by drawing boxplots for each of them.



BMI vs Charges



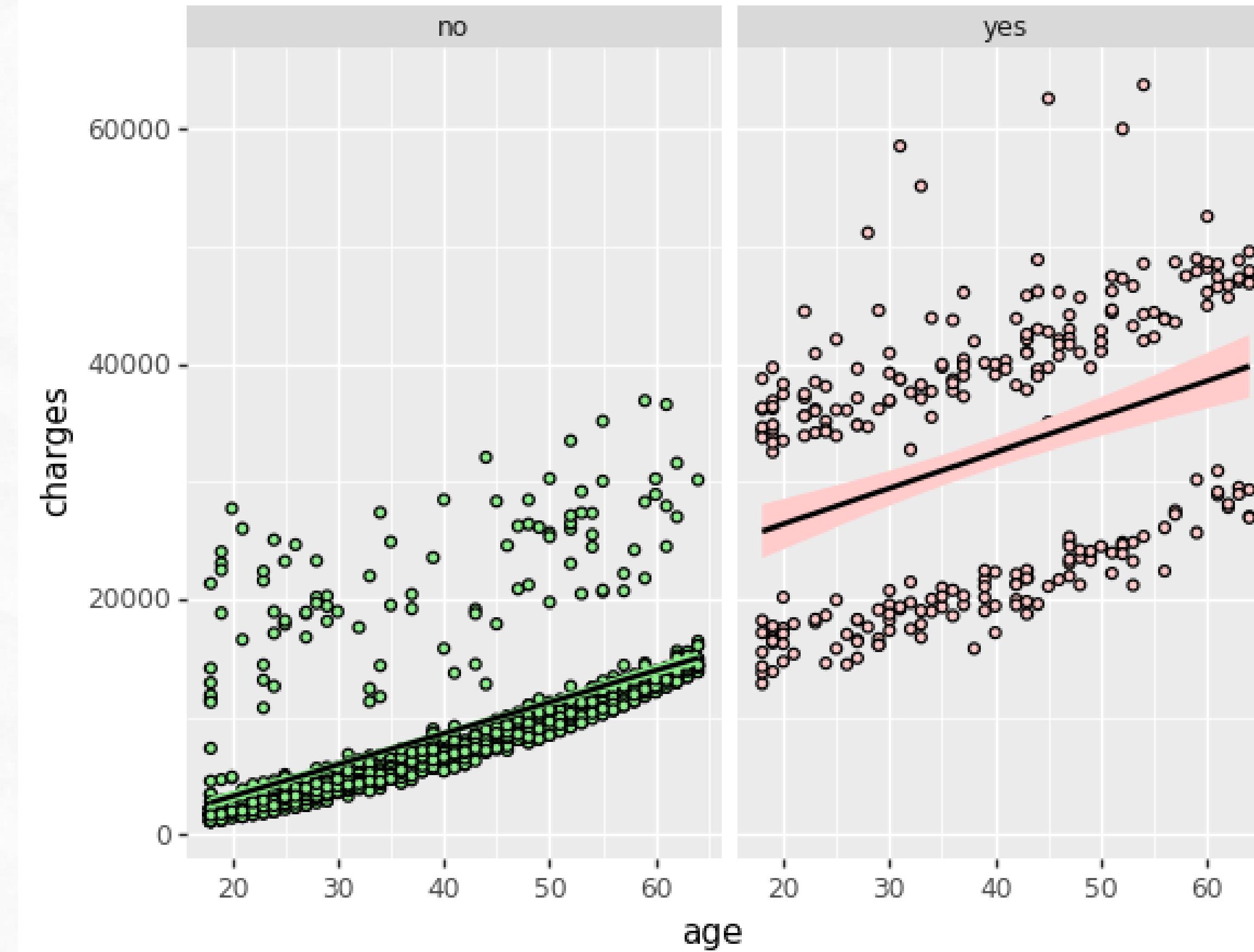
YES AND NO FOR SMOKERS OR NON-SMOKERS

OUTPUTS:

DATA TENDS
TO REMAIN
STABLE FOR
NON-SMOKERS

DATA TENDS
TO INCREASE
LINEARLY FOR
SMOKERS.

Age vs Charges



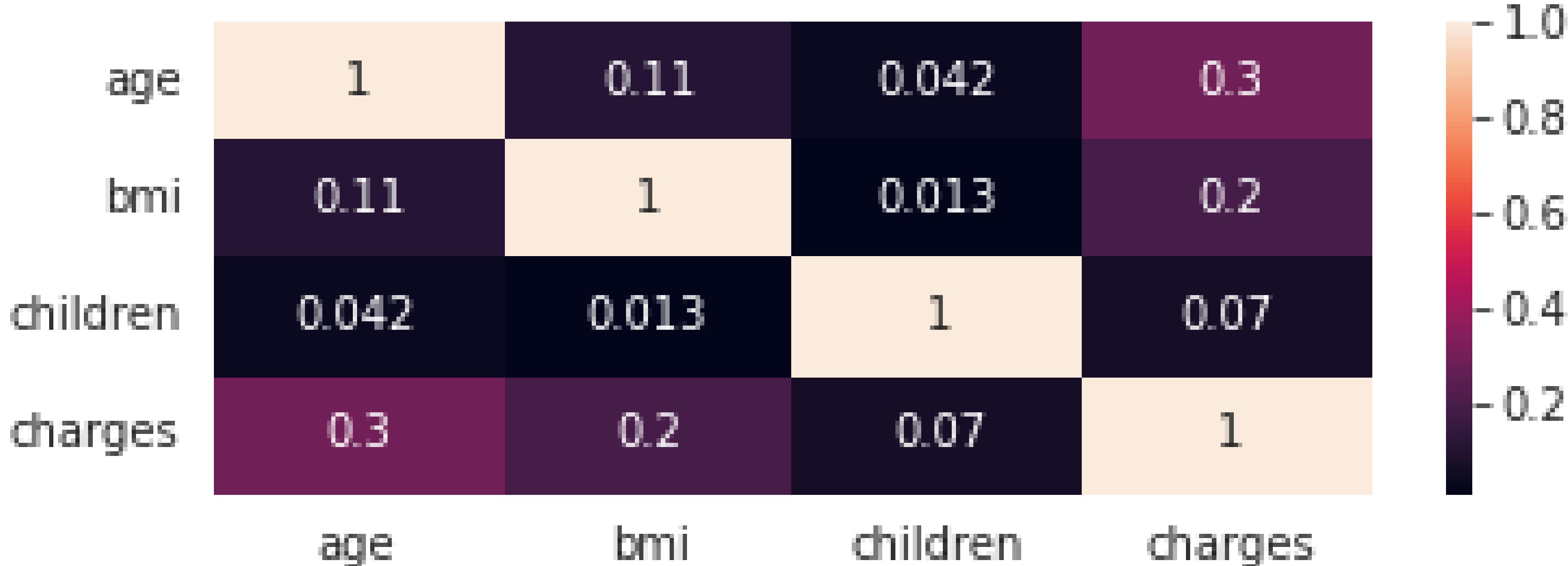
YES AND NO FOR SMOKERS OR NON-SMOKERS

SCATTER PLOTS

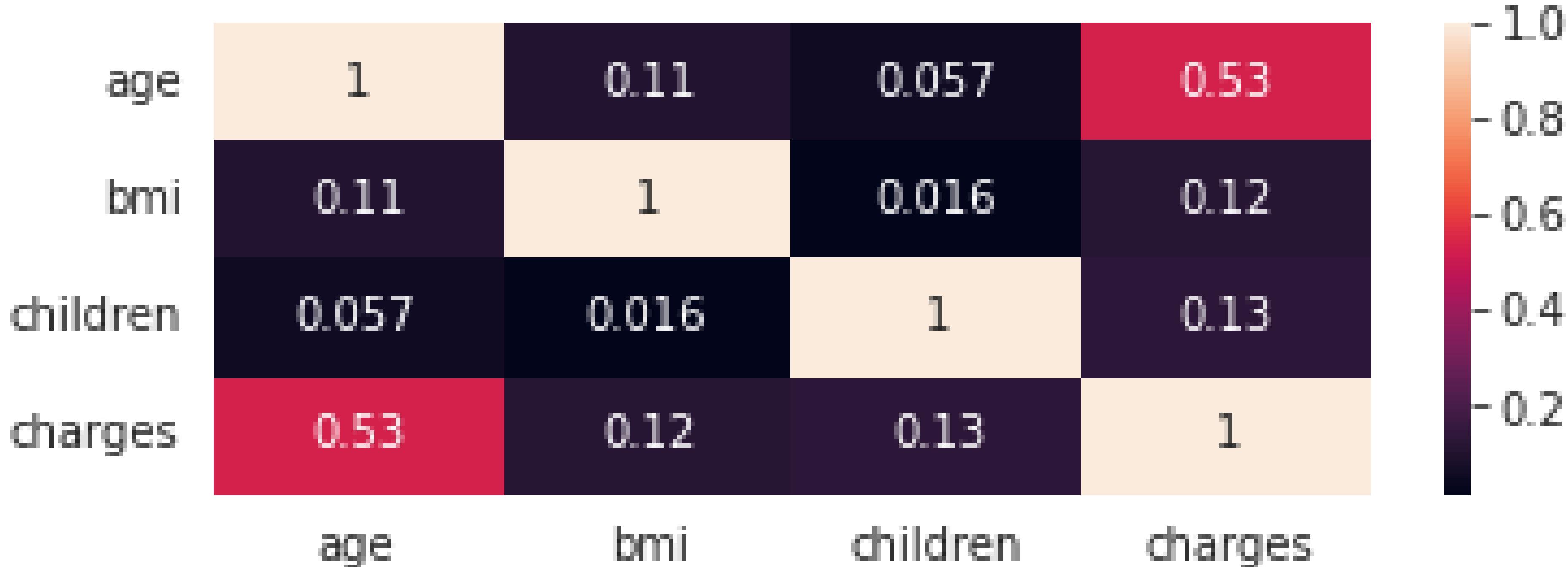
A scatter plot is a diagram where each value in the data set is represented by a dot.



CORRELATION MATRIX PEARSON



CORRELATION MATRIX SPEARMAN



CORRELATION MATRIX



A correlation is a statistical measure of the relationship between two variables. A correlation matrix is a table containing correlation coefficients between variables. Each cell in the table represents the correlation between two variables. The value lies between -1 and 1.

Outputs:

Four clusters are observed.

- Healthy people who do not smoke.
- People who do not smoke but have health problems.
- People who smoke but are healthy.
- People who smoke and have health problems.

Depending on the age of the users or not, their health status and smoking habits affect the insurance fee.

2. FEATURE ENGINEERING



Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.

Feature engineering, in simple terms, is the act of converting raw observations into desired features using statistical or machine-learning approaches.

2. FEATURE ENGINEERING



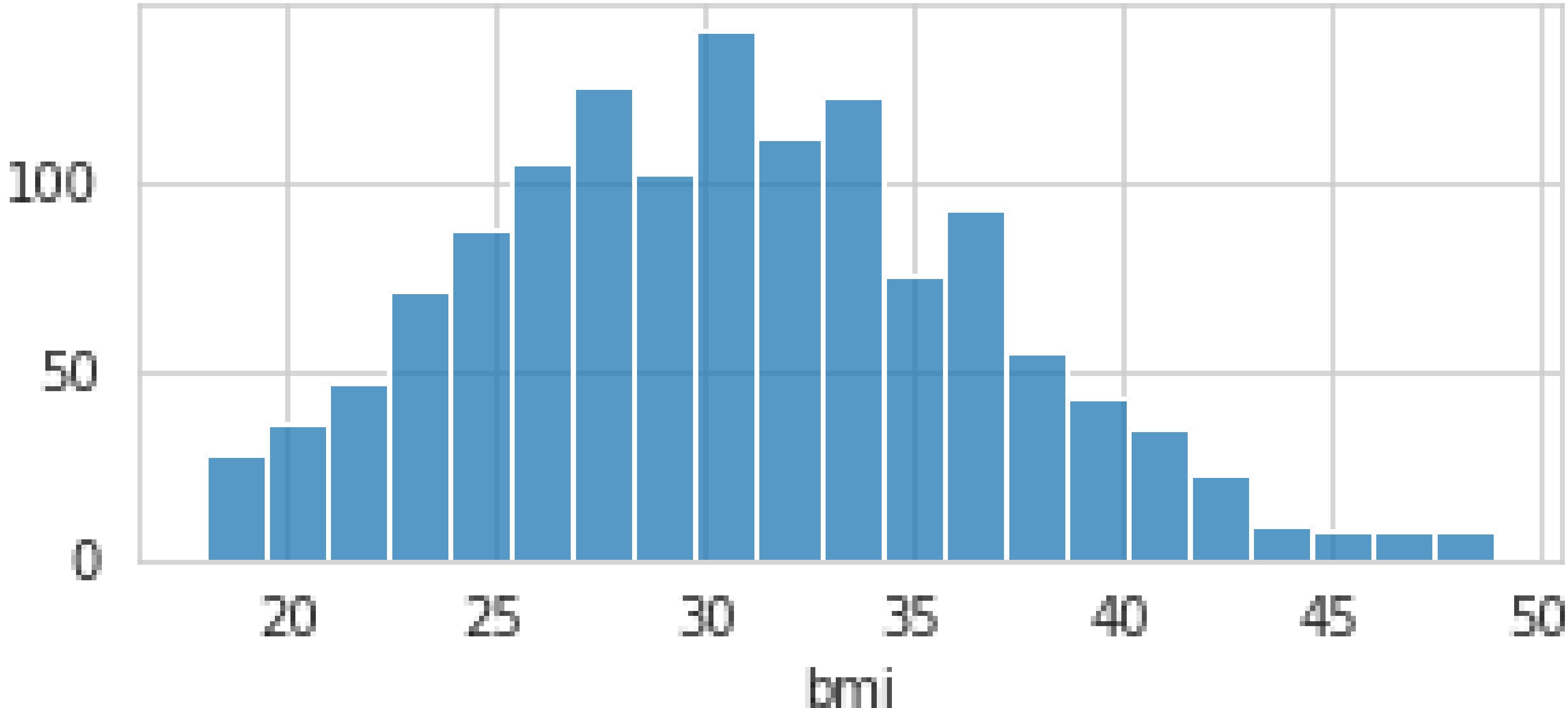
What to do in preparation for this part:

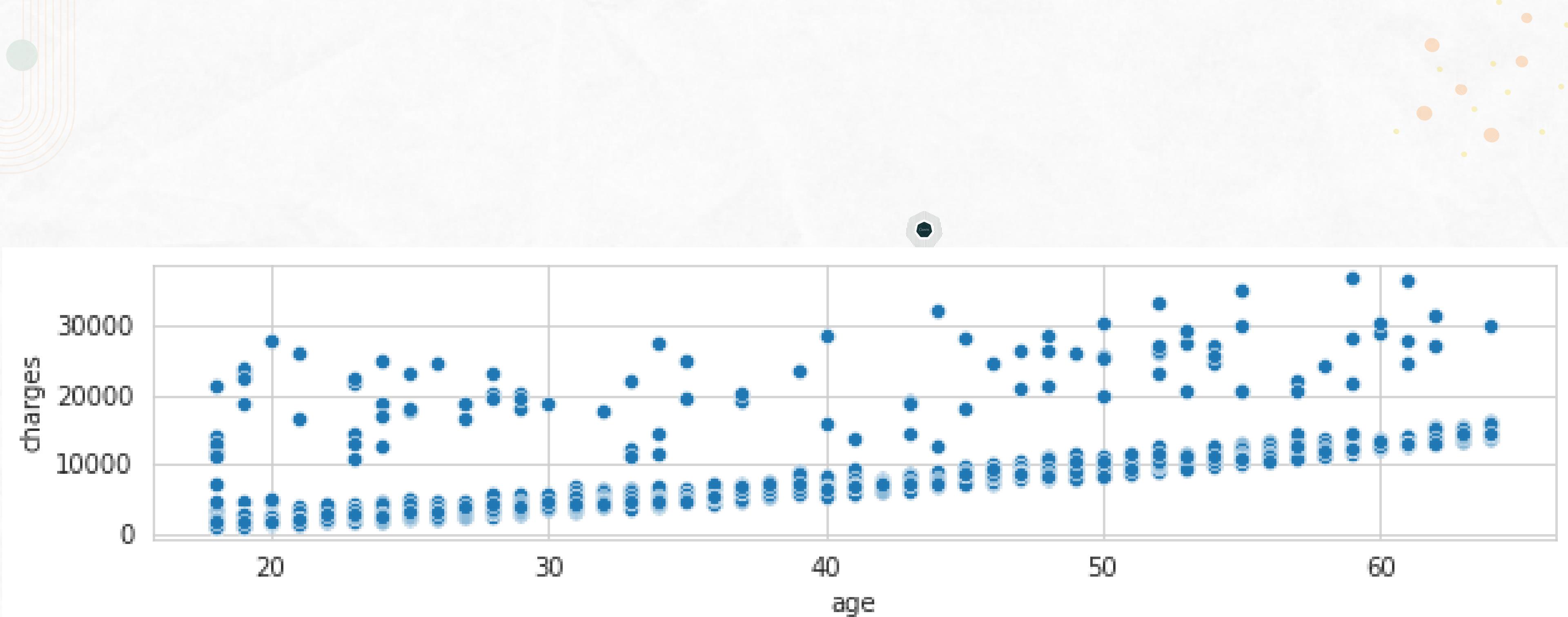
Deleting any null values found and replacing them with logical and appropriate data.

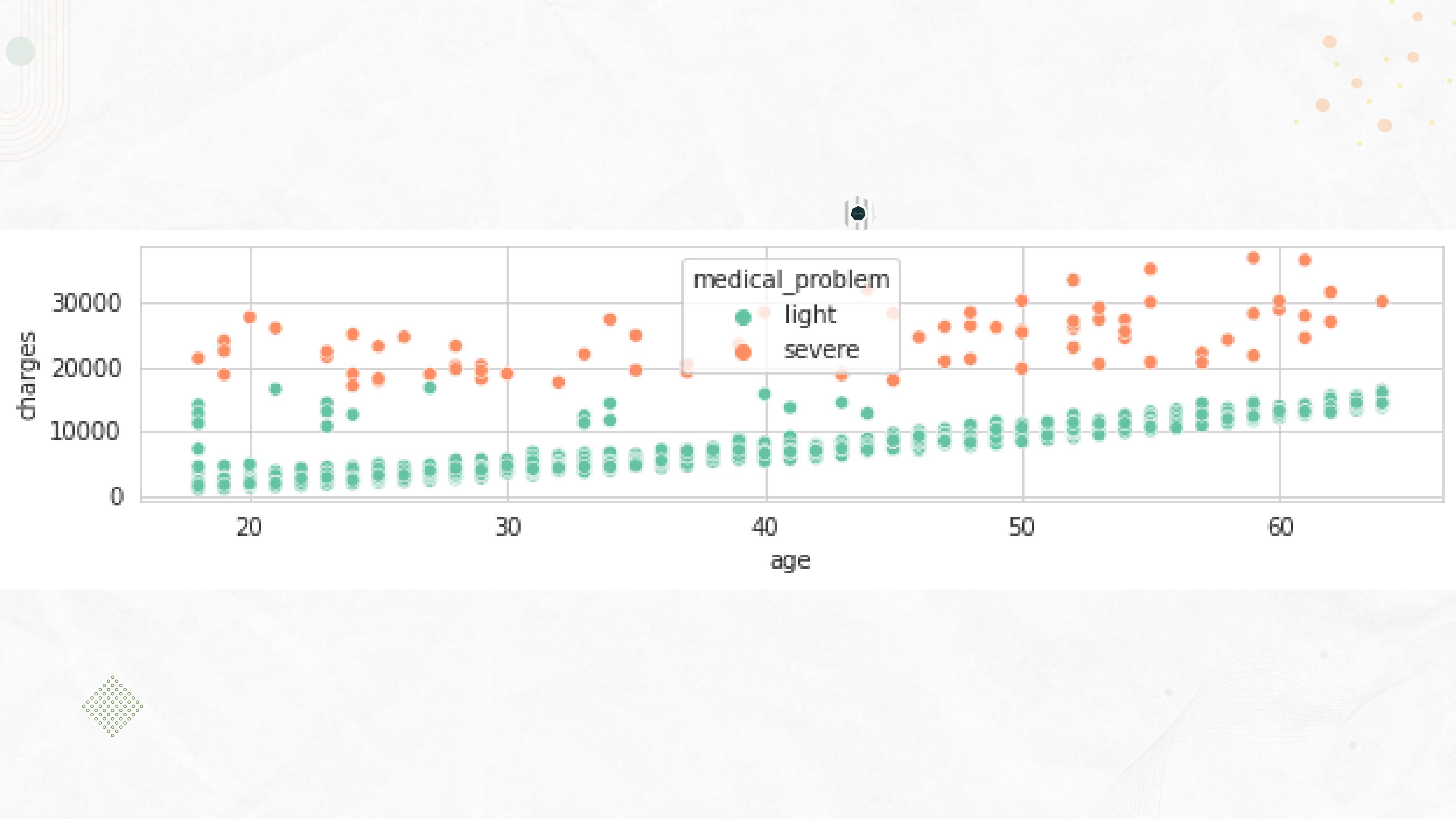
Adding the missing medical problems in the data set.

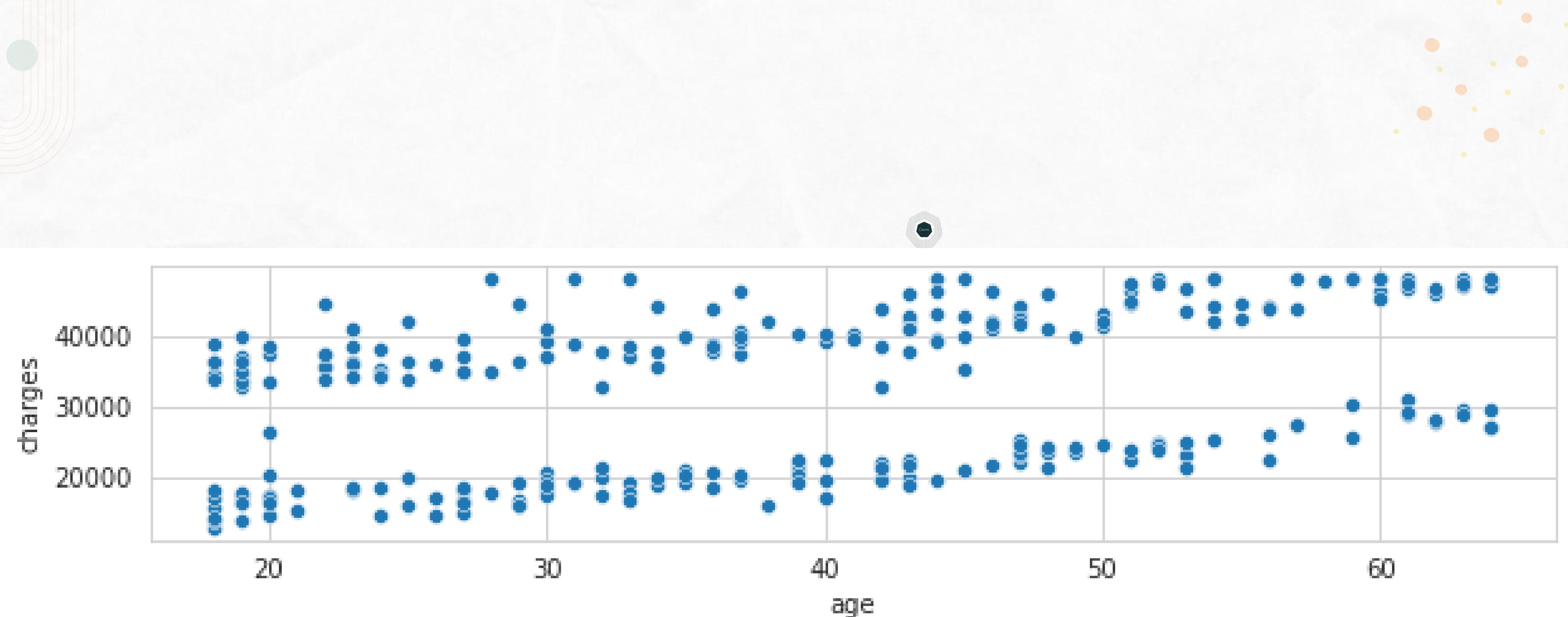
Intervals are immutable objects that can be created by specifying their connected components. Studies will be made to find the best interval.

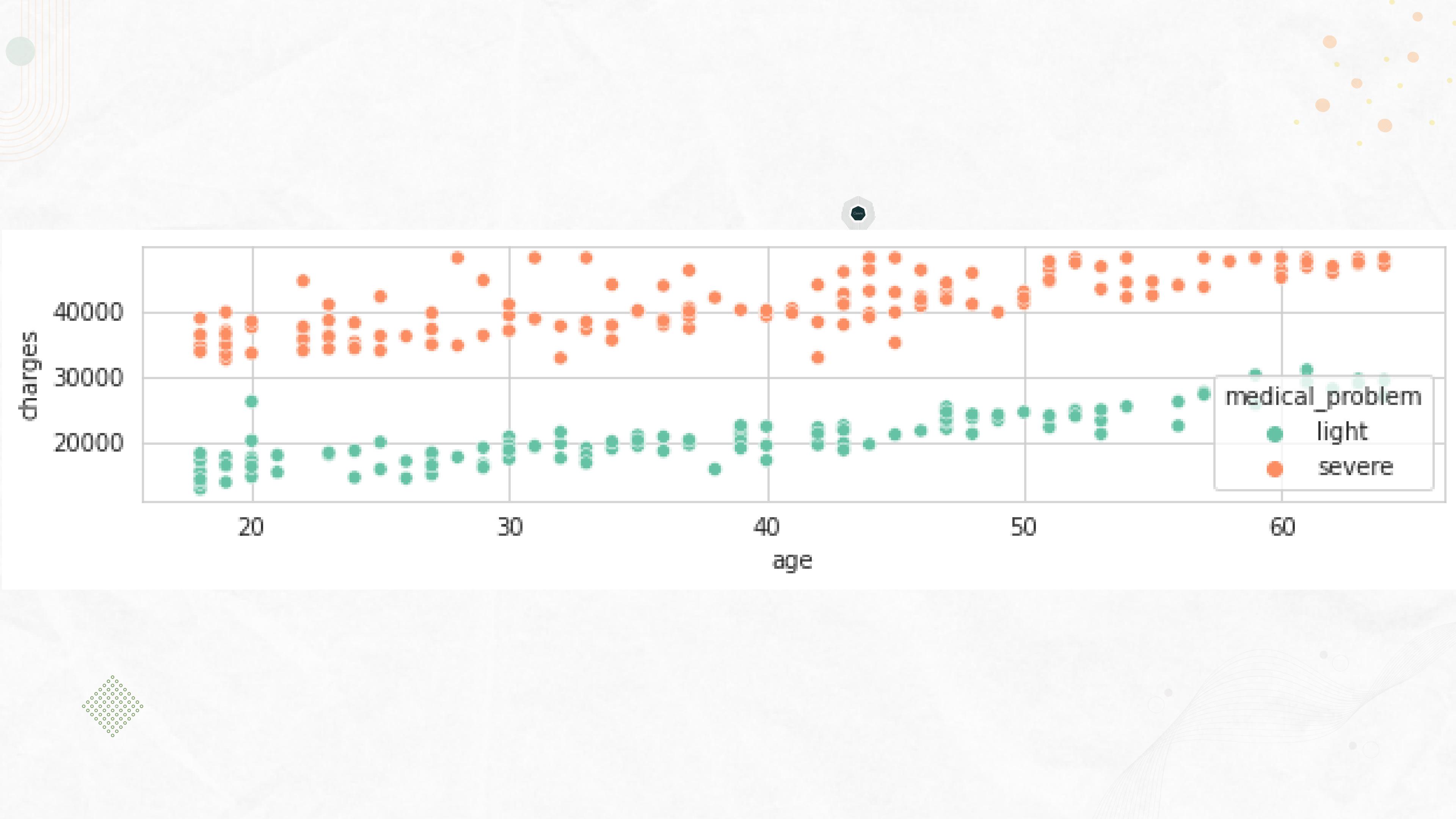
After finding the upper and lower ranges through the code, the author repeats the EDA after determining the reasonable upper and lower limits.



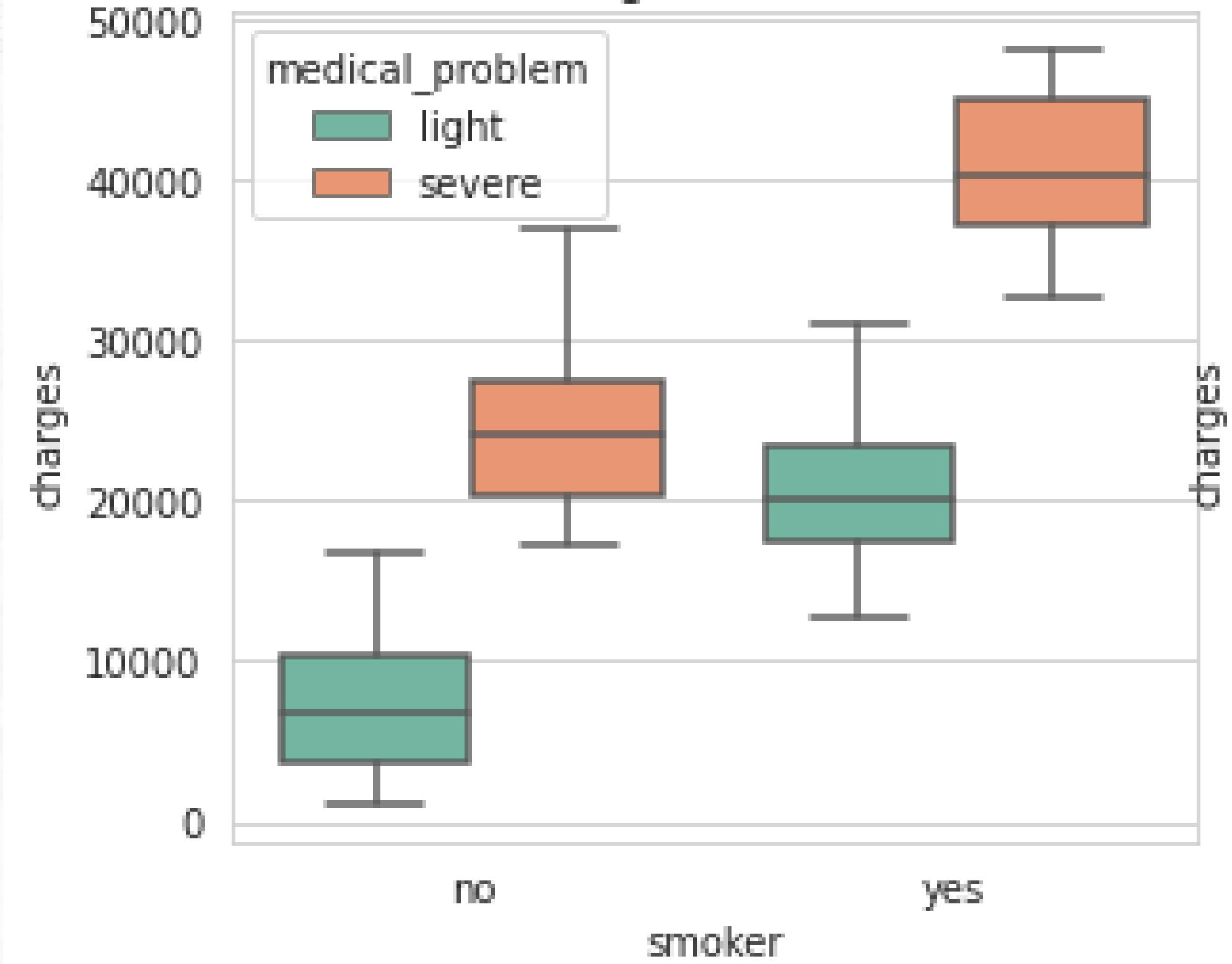




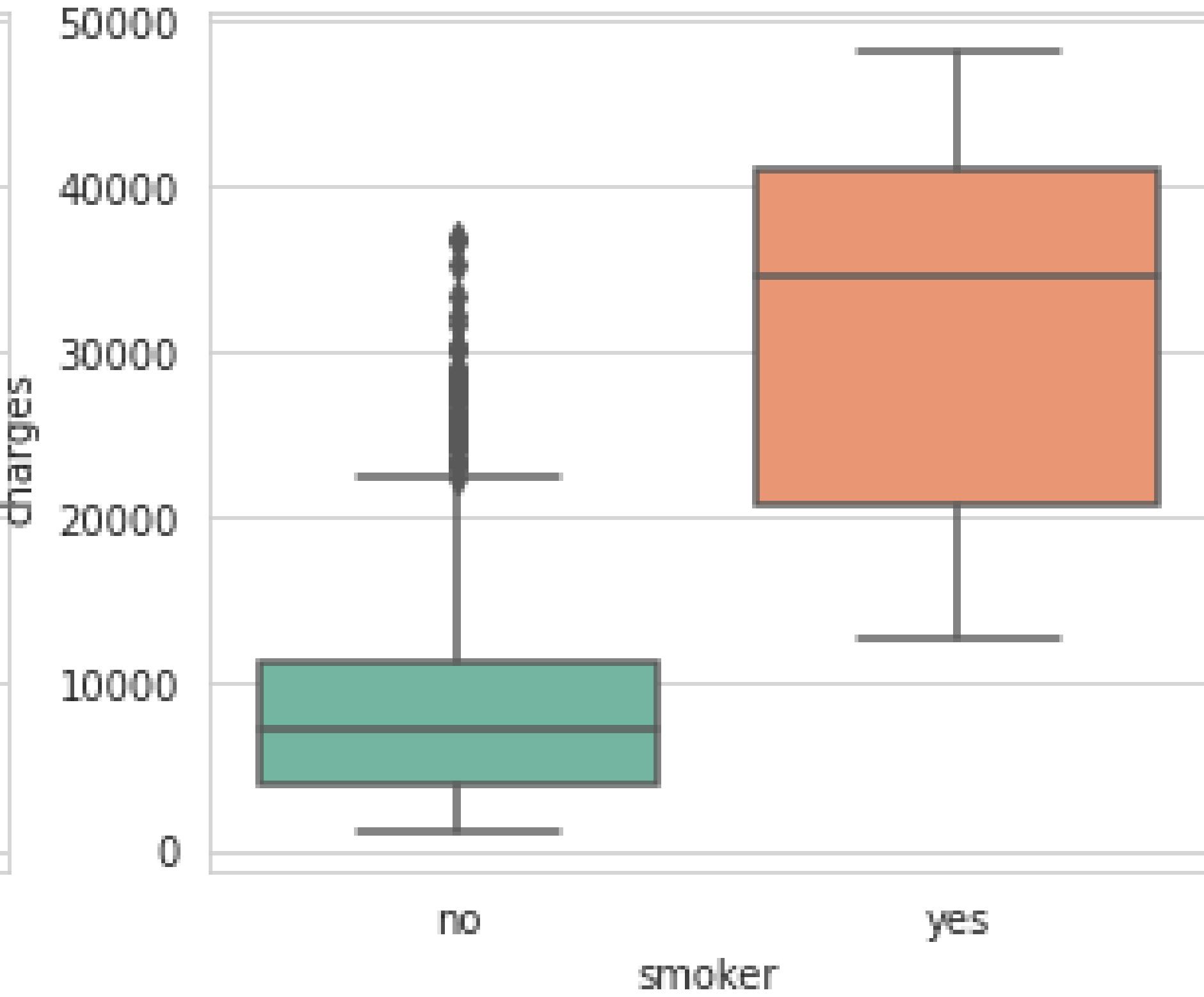




Adding new feature



Without the new feature



3. SELECTION OF MODEL IDEAL

Considering the purpose of the project, it is the part where the models to be used for problem-solving will be tested and the most suitable model for the problem will be selected.

MODELS TO BE USED:



- Linear Regression
- Gradient Boosting
- XGBOOST

DATA PREPROCESSING AND TRANSFORMING

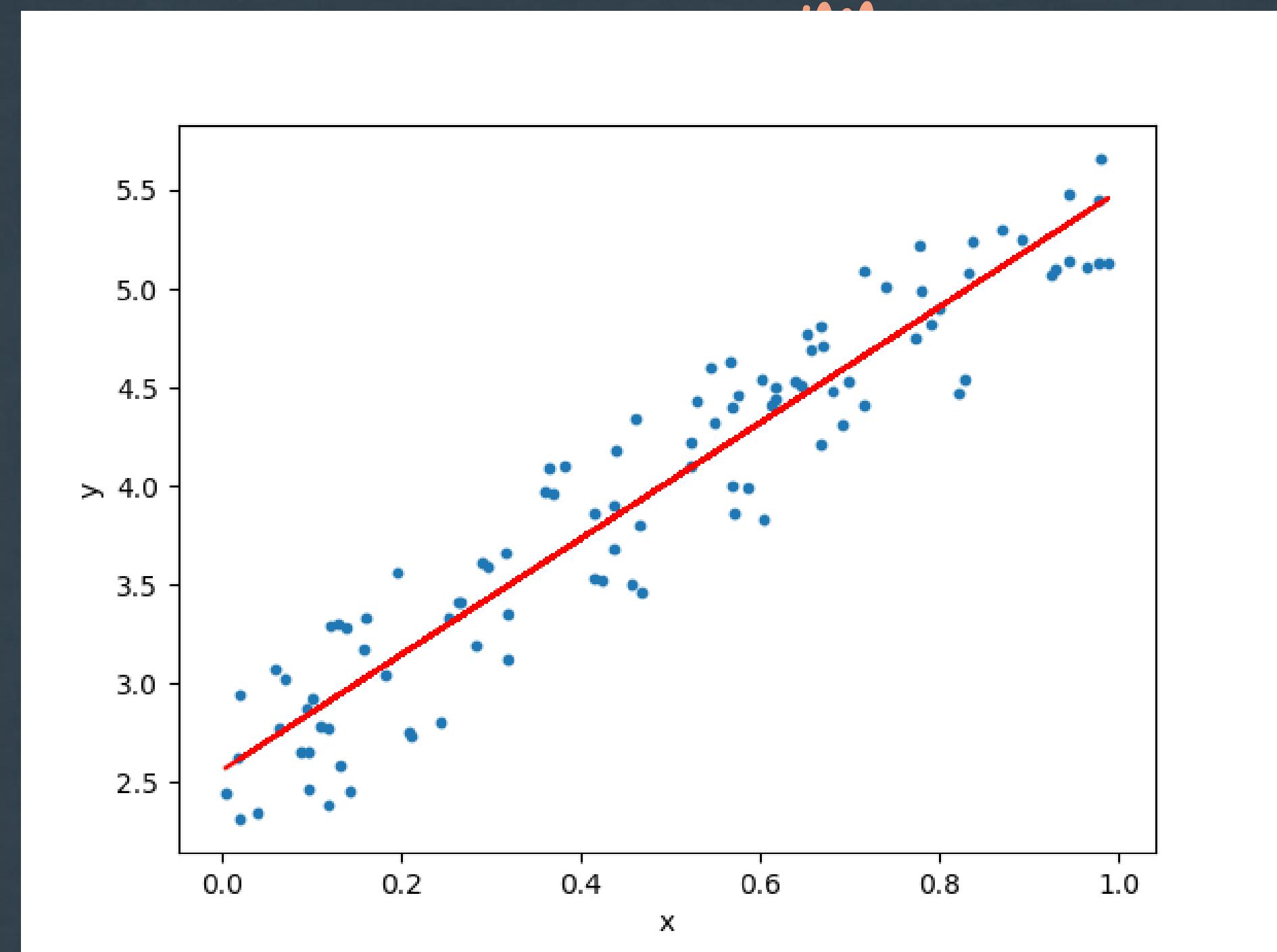
Split Data

MinMax
Scaler

One Hot
Encoder

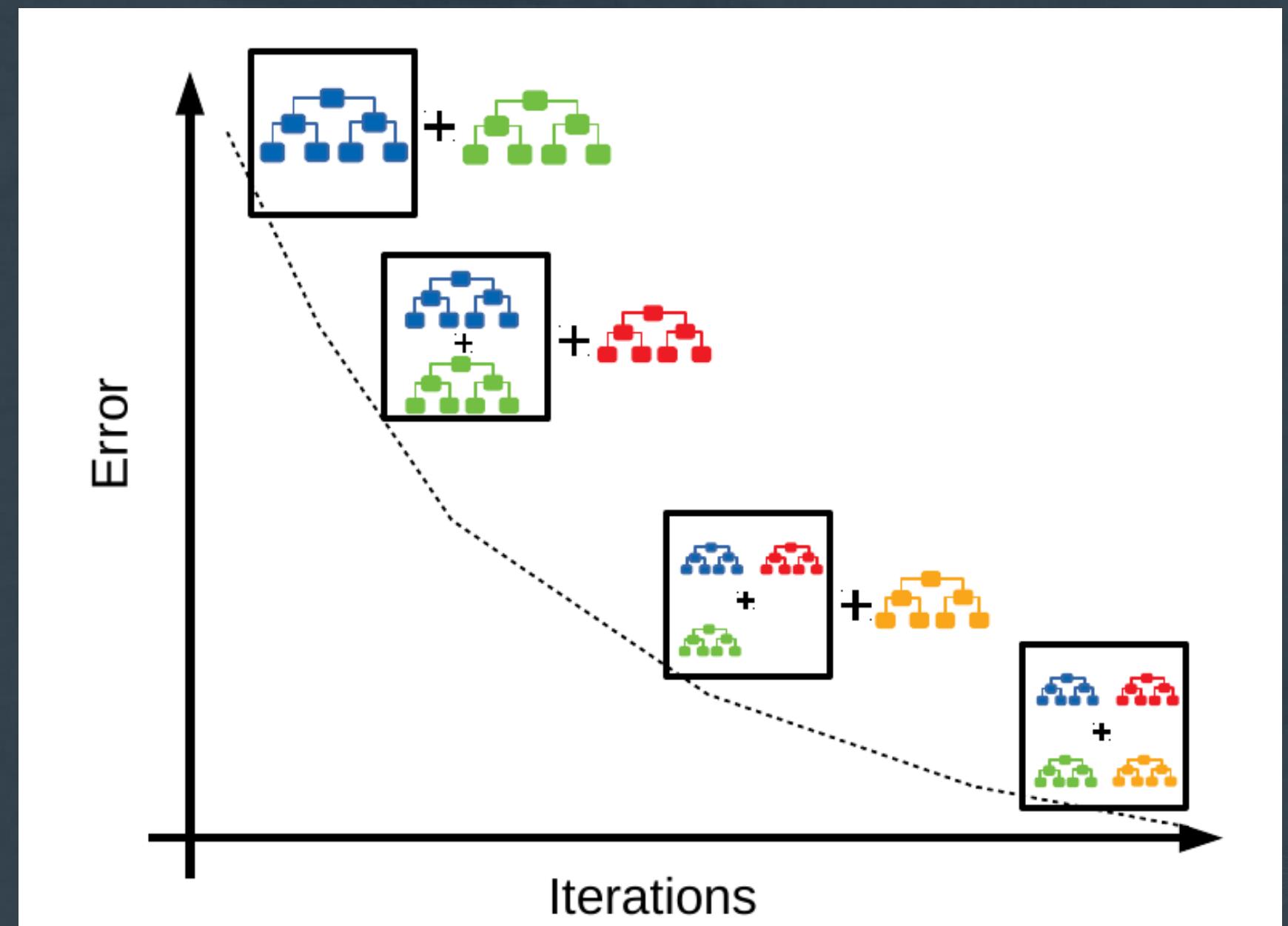
LINEAR REGRESSION

Linear regression uses the relationship between the data points to draw a straight line through all of them. This line can be used to predict future values. In Machine Learning, predicting the future is quite used.



GRADIENT BOOSTING

Gradient Boosting is a functional gradient algorithm that repeatedly selects a function that leads in the direction of a weak hypothesis or negative gradient so that it can minimize a loss function.



XGBOOST



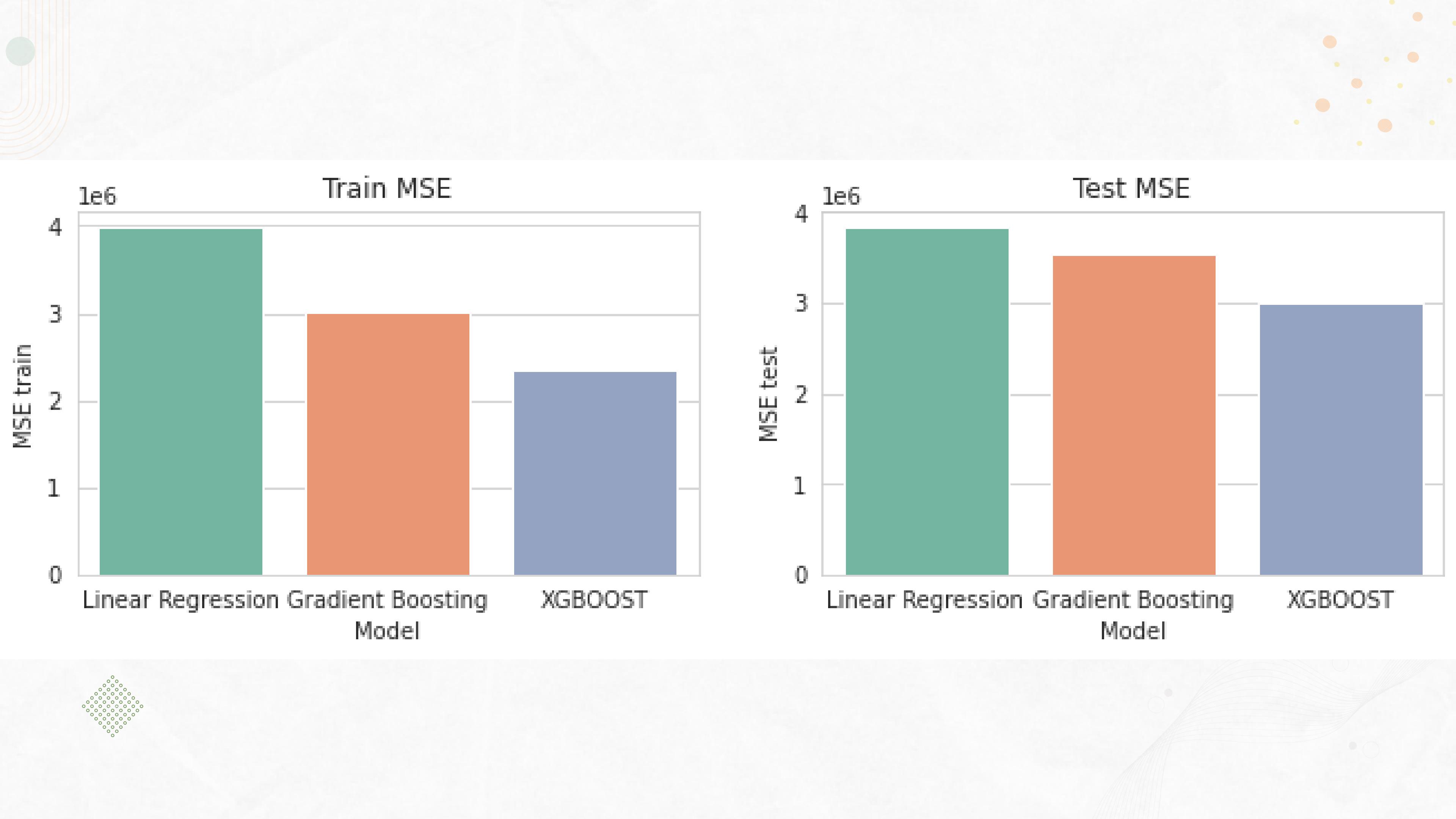
XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine-learning library for regression, classification, and ranking problems.

TO THE CONCLUSION

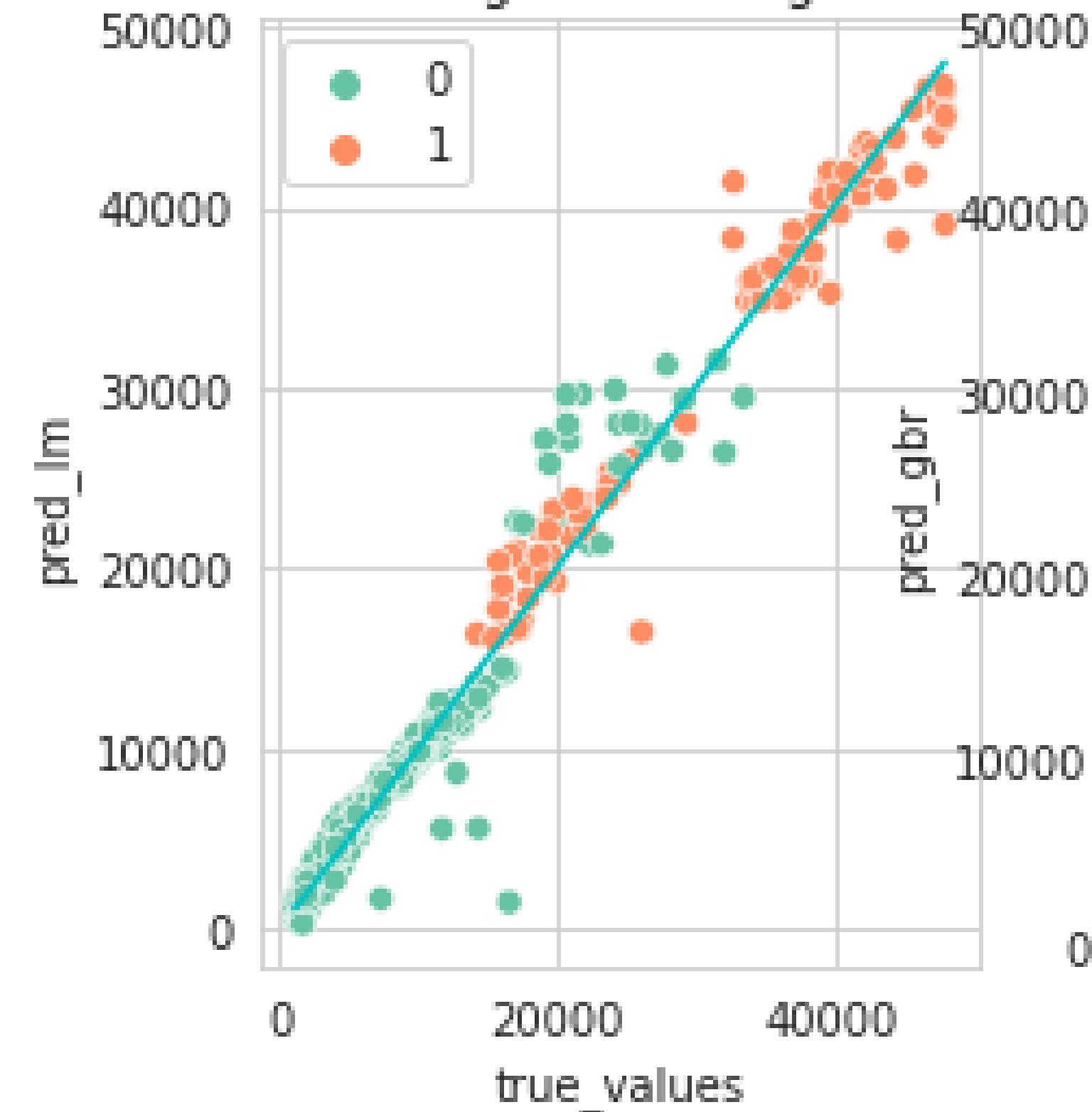
Mean
Squared
Error

Cross
Validation

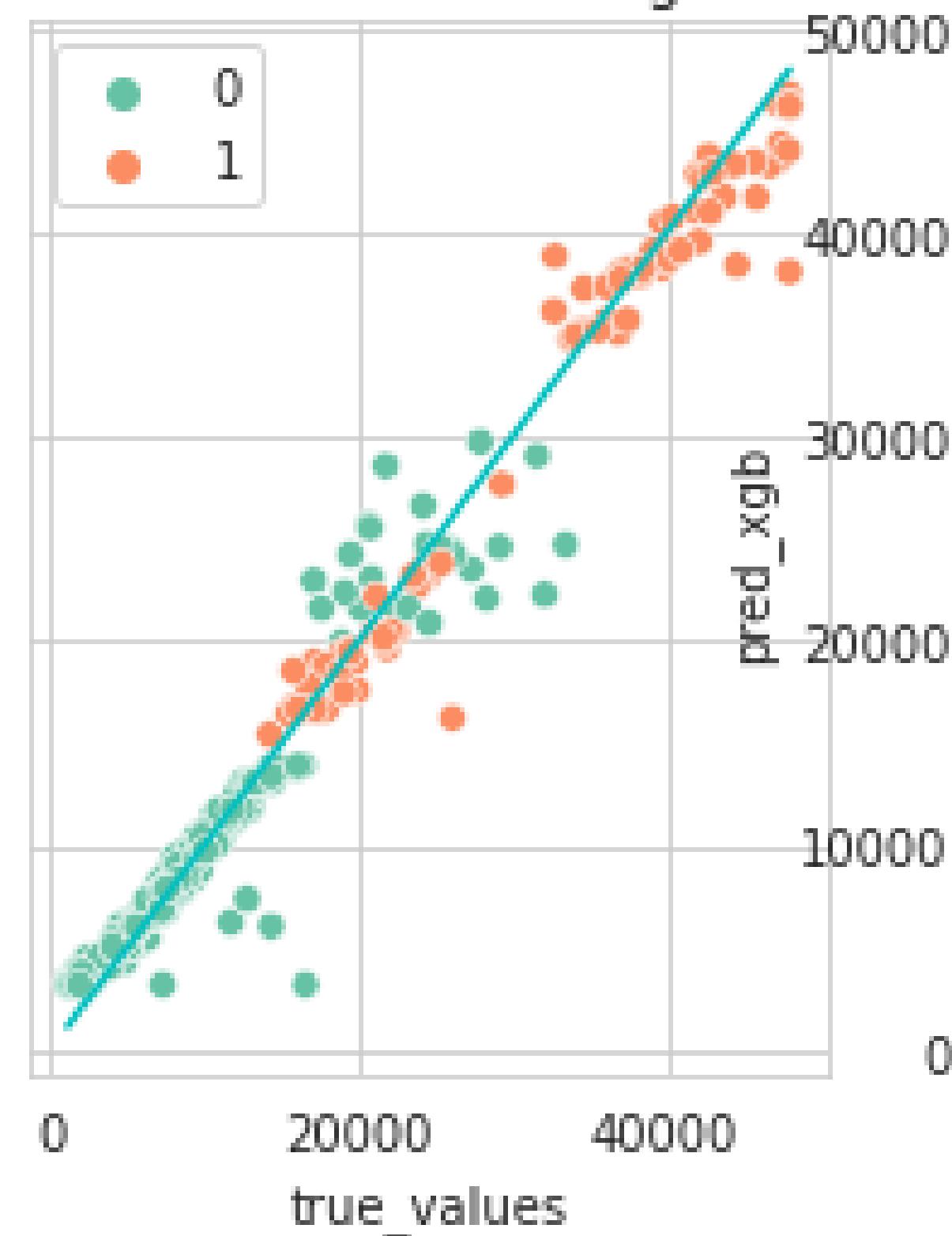
Dataframe
Evaluation



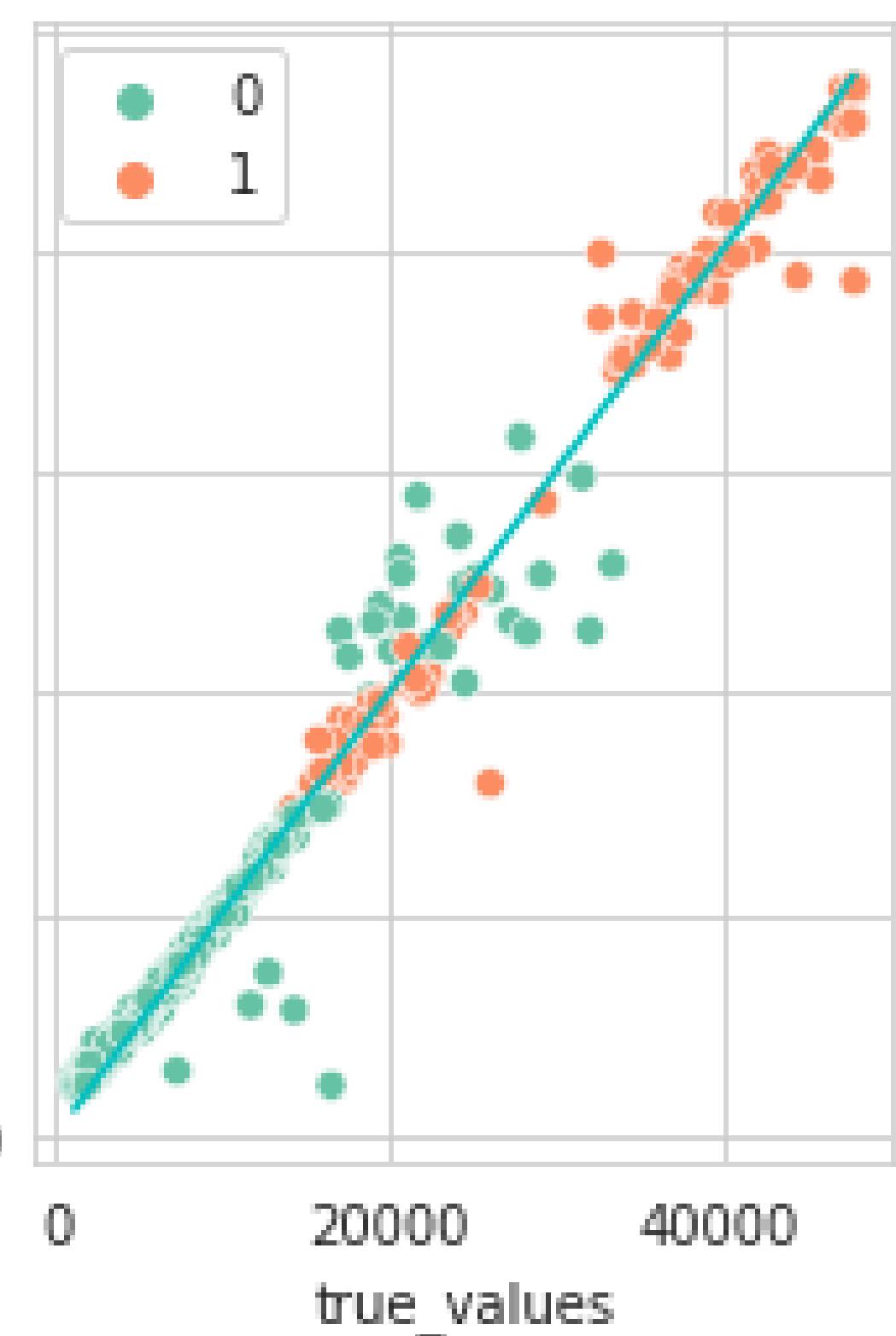
Linear Regression Regression



GradientBoosting

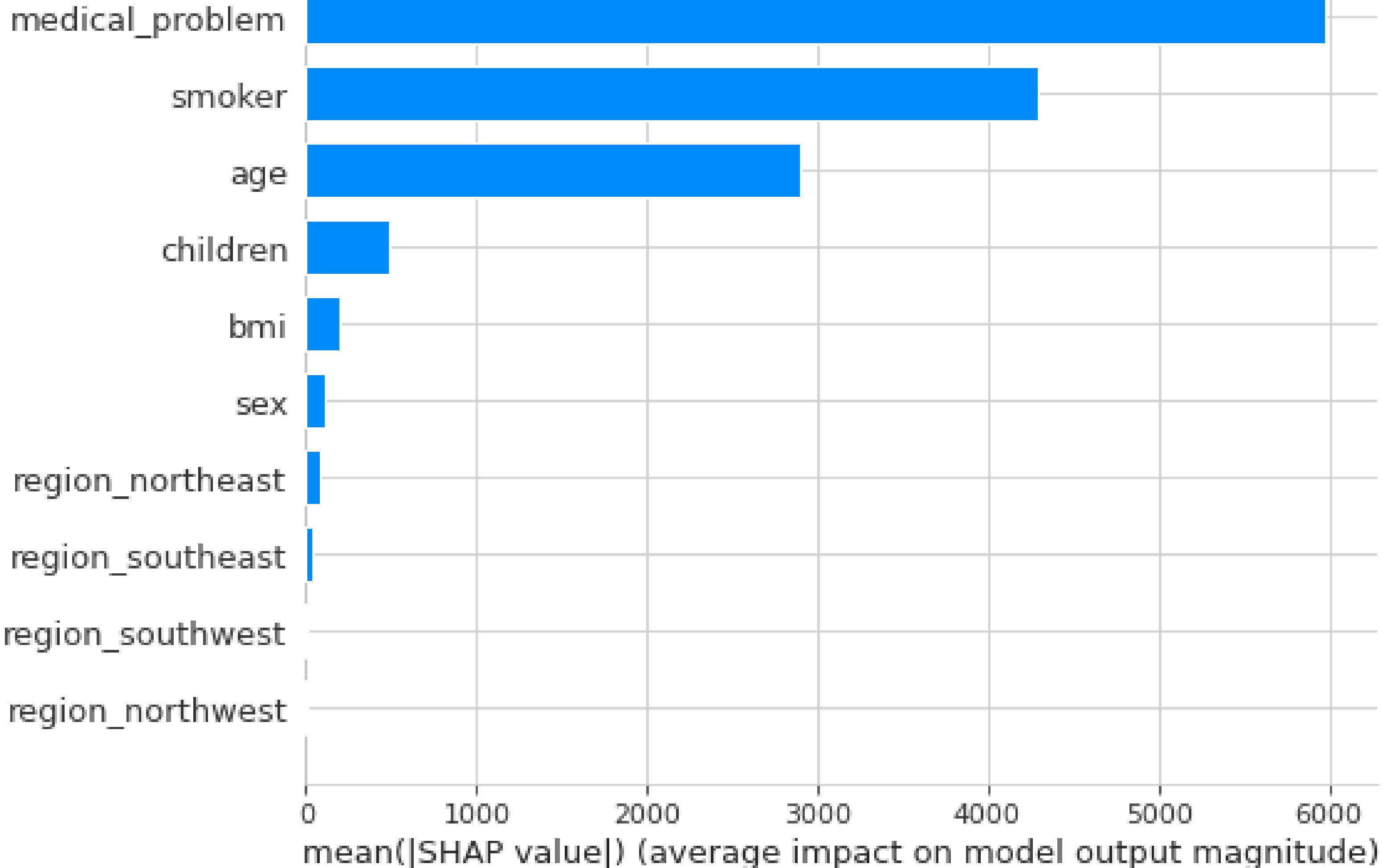


XGBOOST



4.DEFINITIVE MODEL

With the result method, the transactions are arranged and the final code is found.



medical_problem

smoker

age

children

bmi

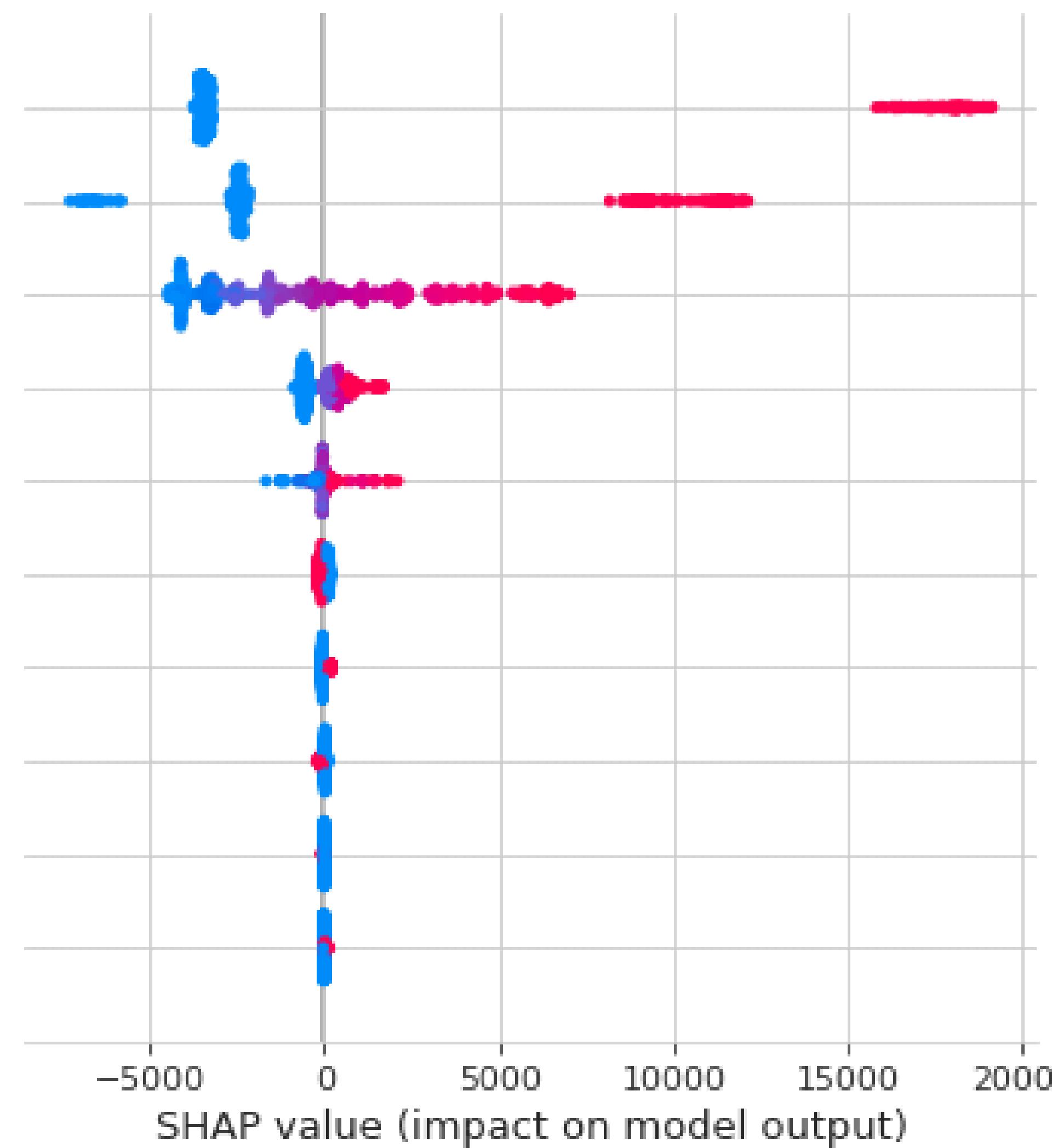
sex

region_northeast

region_southeast

region_southwest

region_northwest



High

Low

Feature value

True Values vs Predictions

