HEALTH INSURANCE COST PREDICTION

OBJECTIVE: TO CREATE A MODEL THAT PREDICTS THE COST OF AN INDIVIDUAL'S INSURANCE.

DATA SUPPLIED

- There is a CSV data file which contains 7 features out of which Charges is the target feature:

1. Age
2. Sex
3. BMI
4. Children
5. Smoker
6. Region
7. Charges

Software languages and libraries used

- Python programming language
- Google colab for a notebook environment
- Pandas
- NumPy
- Scikit-learn
- Seaborn
- matplotlib

Steps taken for this project

Part 1: Defined the problem

- Outlined what are the features, the target variable?
- Is it a regression problem or classification? Then decided the metric to optimize.

Part 2: Discovered the data

- Checked missing, duplicate data, and outliers and summarized the data.
- Visualized the features with the target to check their impact and relationship.

Cont…

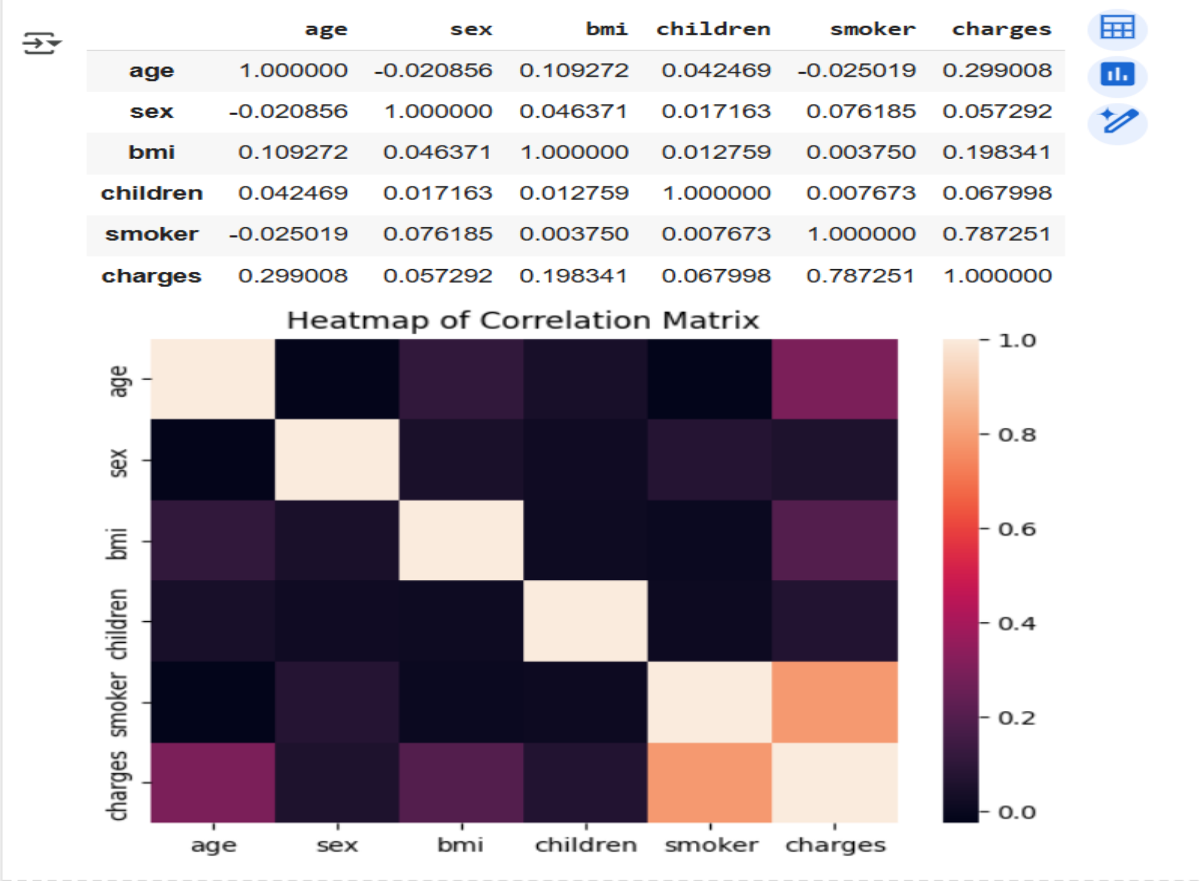
Part 3: Developed the model

- Built linear regression, support vector regressor, gradient boosting, and random forest regression model.
- Fine-tuned them by hand, and fit them, selected the best one, fit and checked the prediction.

Part 4: Deployed the model on a web-app

- Dumped the final model using Joblib library from Python and created app.py, html, css, JavaScript, requirements.txt, and app.yaml files.
- Deployed the entire model on Google Cloud Platform App Engine to predict the insurance cost for any new prediction.

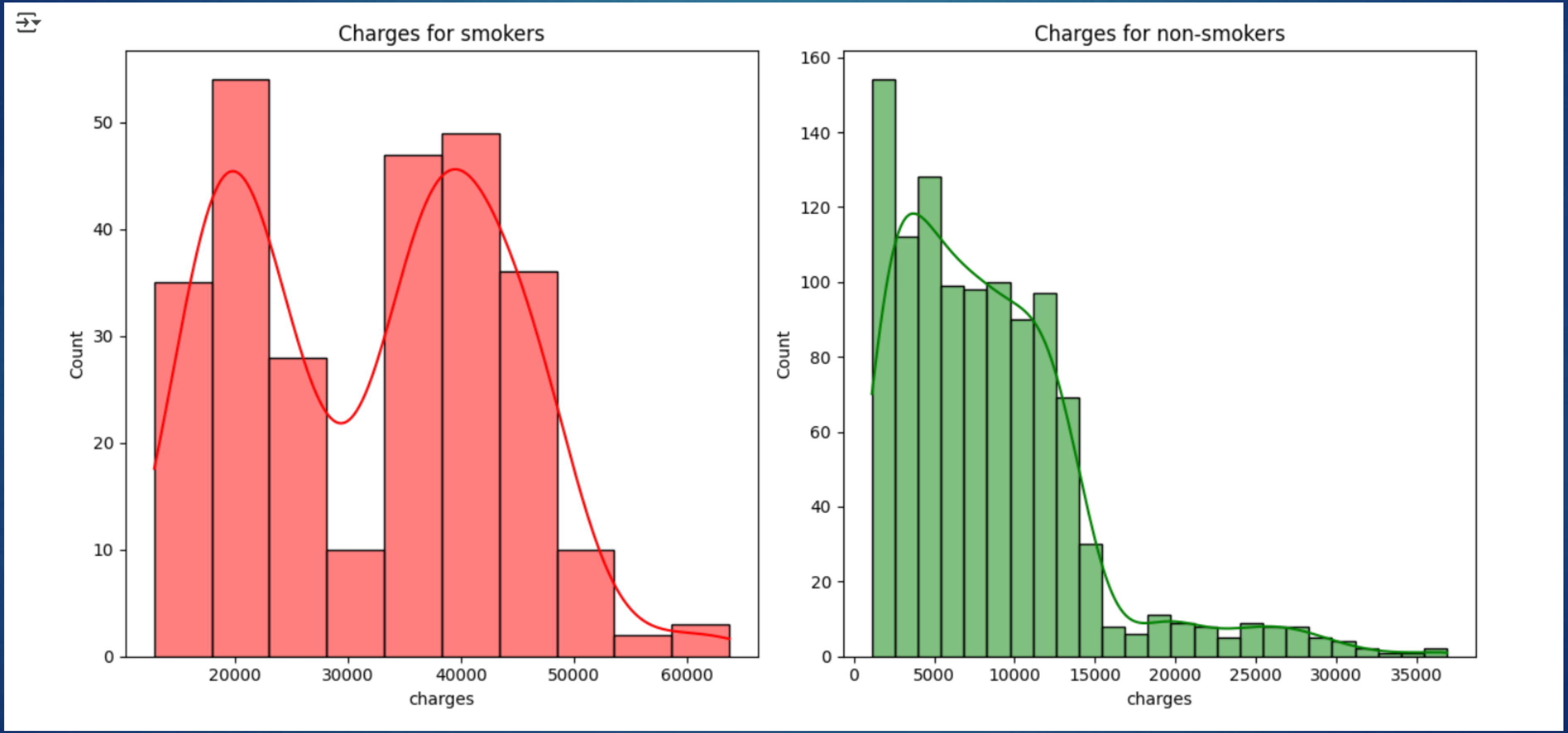
Chart shows that Smoker has the highest correlation with Charges



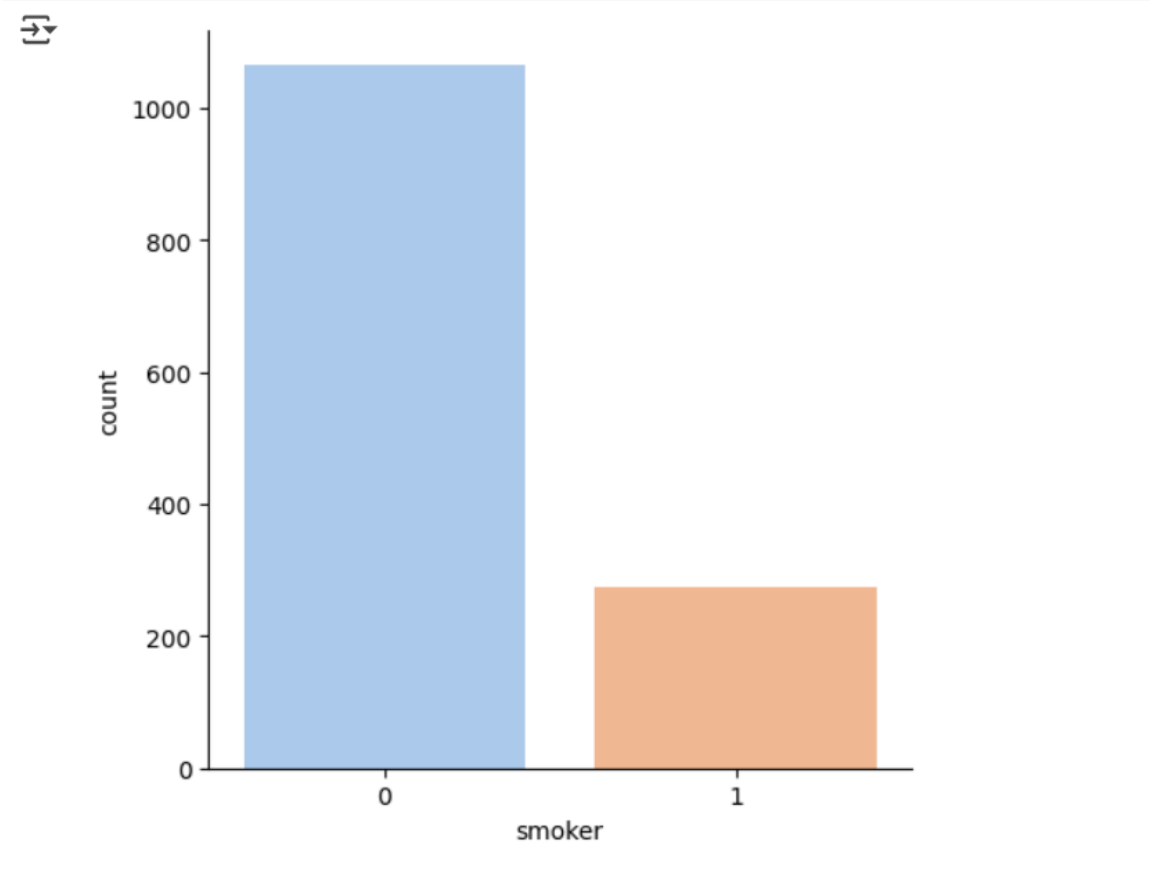
	age	sex	bmi	children	smoker	charges
age	1.000000	-0.020856	0.109272	0.042469	-0.025019	0.299008
sex	-0.020856	1.000000	0.046371	0.017163	0.076185	0.057292
bmi	0.109272	0.046371	1.000000	0.012759	0.003750	0.198341
children	0.042469	0.017163	0.012759	1.000000	0.007673	0.067998
smoker	-0.025019	0.076185	0.003750	0.007673	1.000000	0.787251
charges	0.299008	0.057292	0.198341	0.067998	0.787251	1.000000

Heatmap of Correlation Matrix

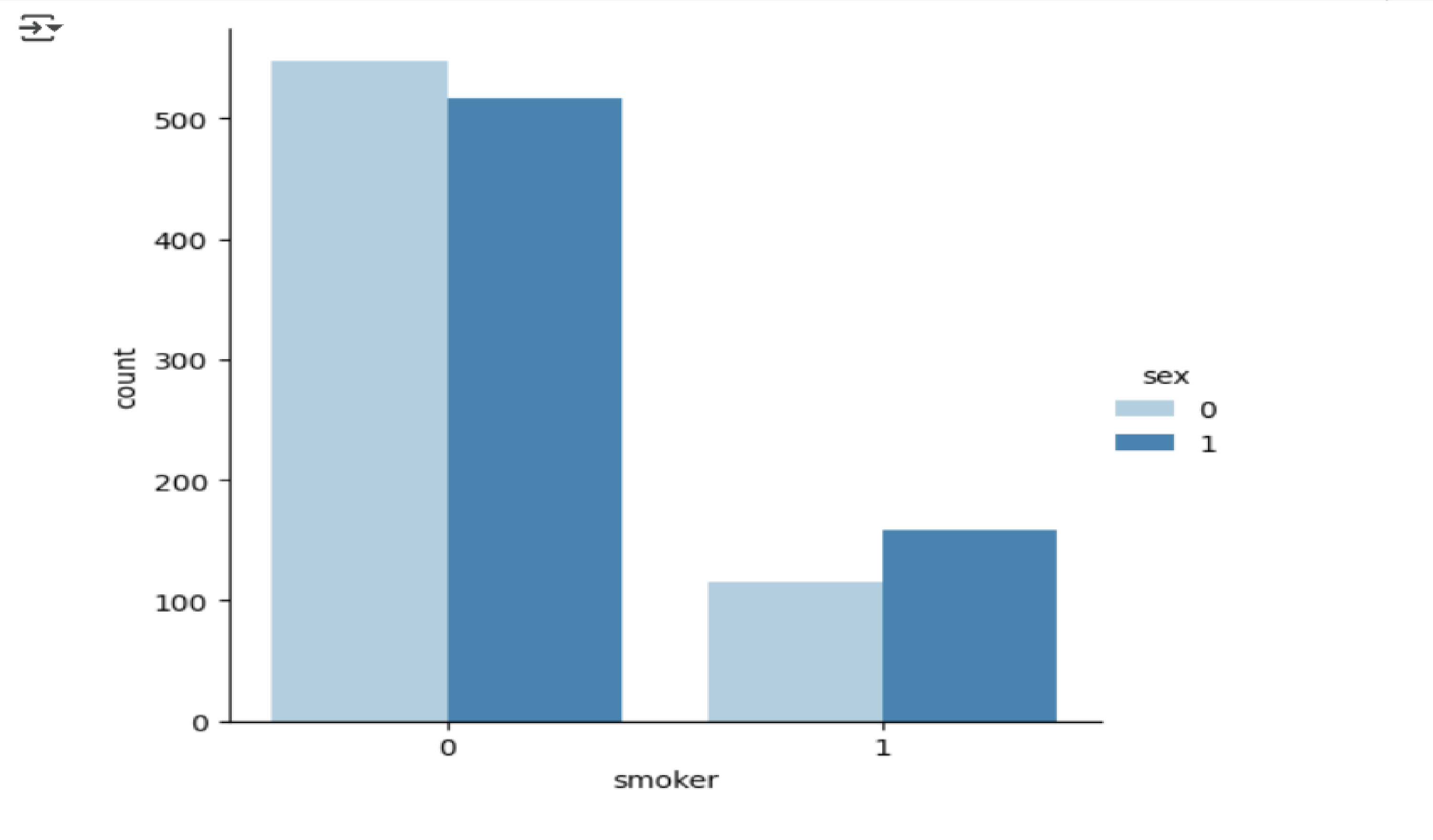
Charges for Smokers are higher than Charges for Non-smokers



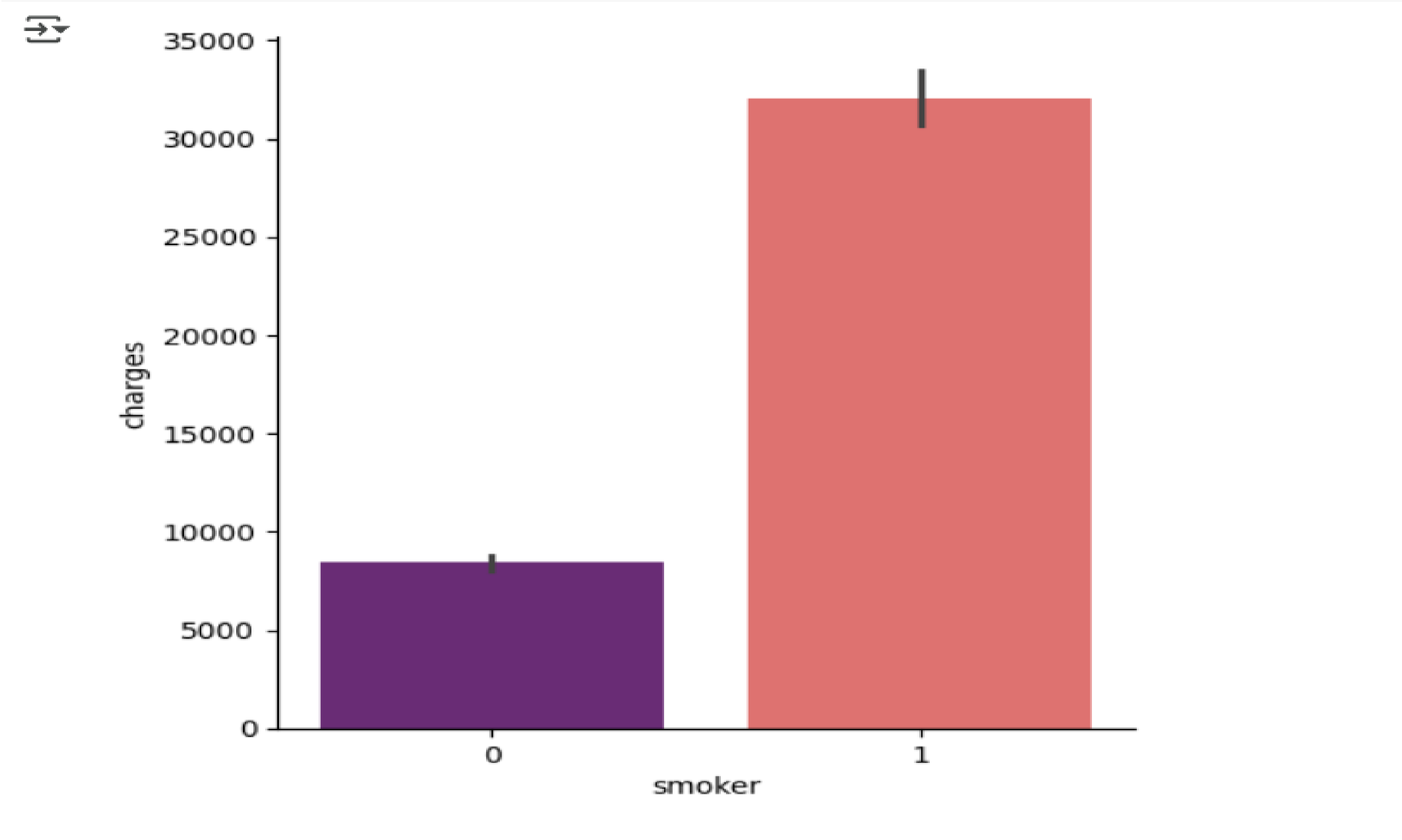
Non-smokers are higher than the smokers



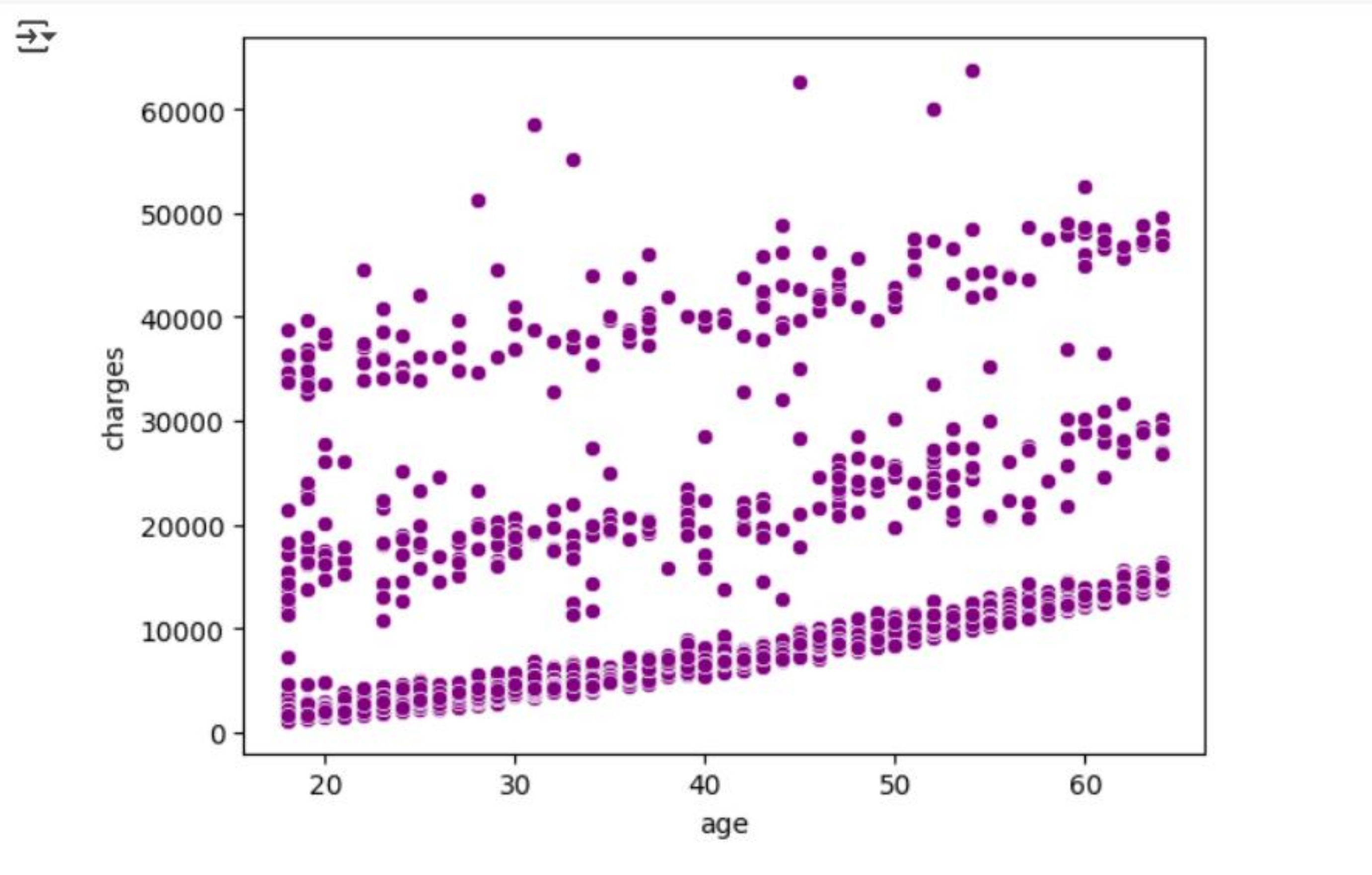
The chart shows the distribution of the smokers or non-smokers by Sex



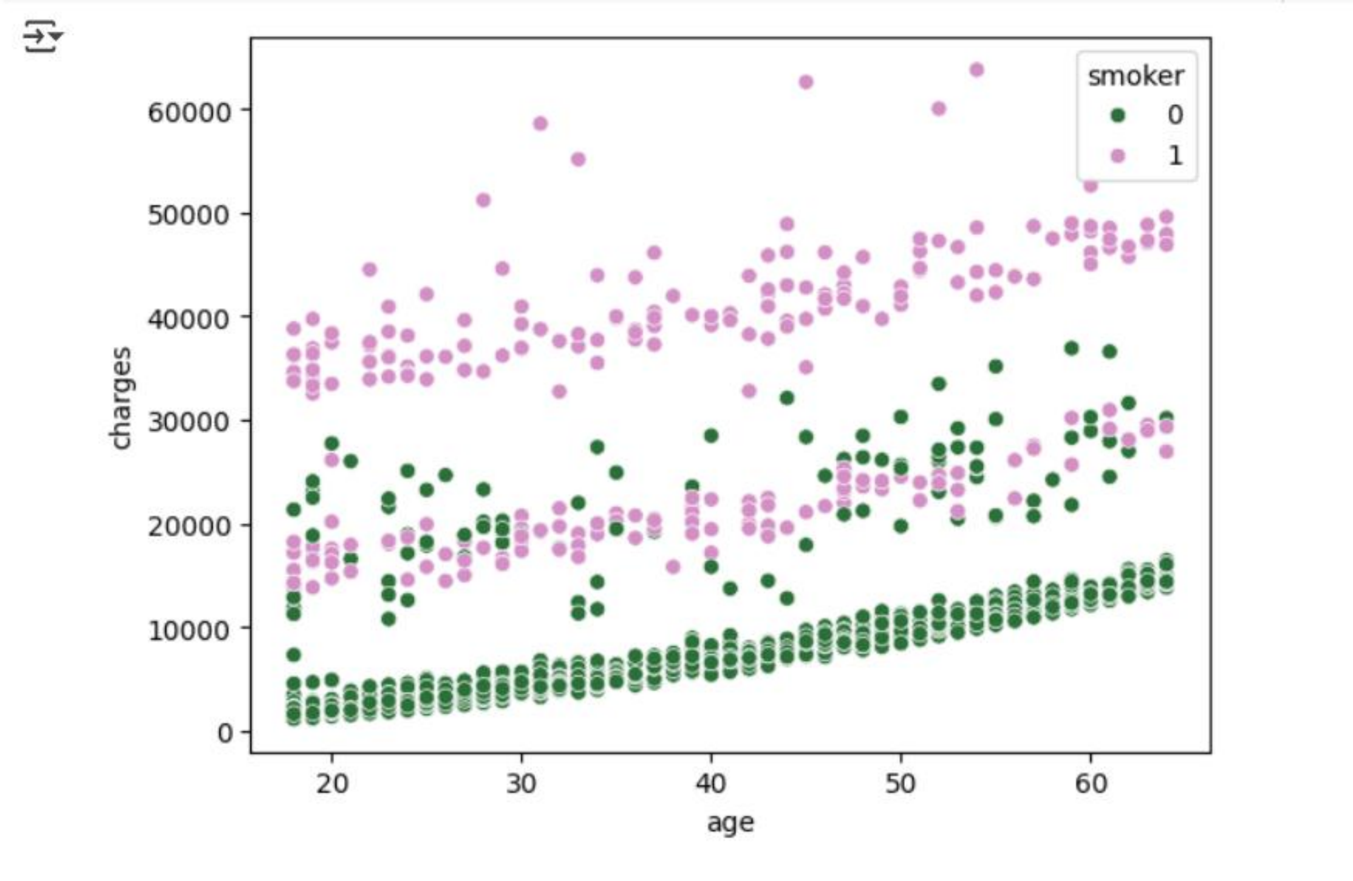
Charges are higher for Smokers compared to Non-smokers



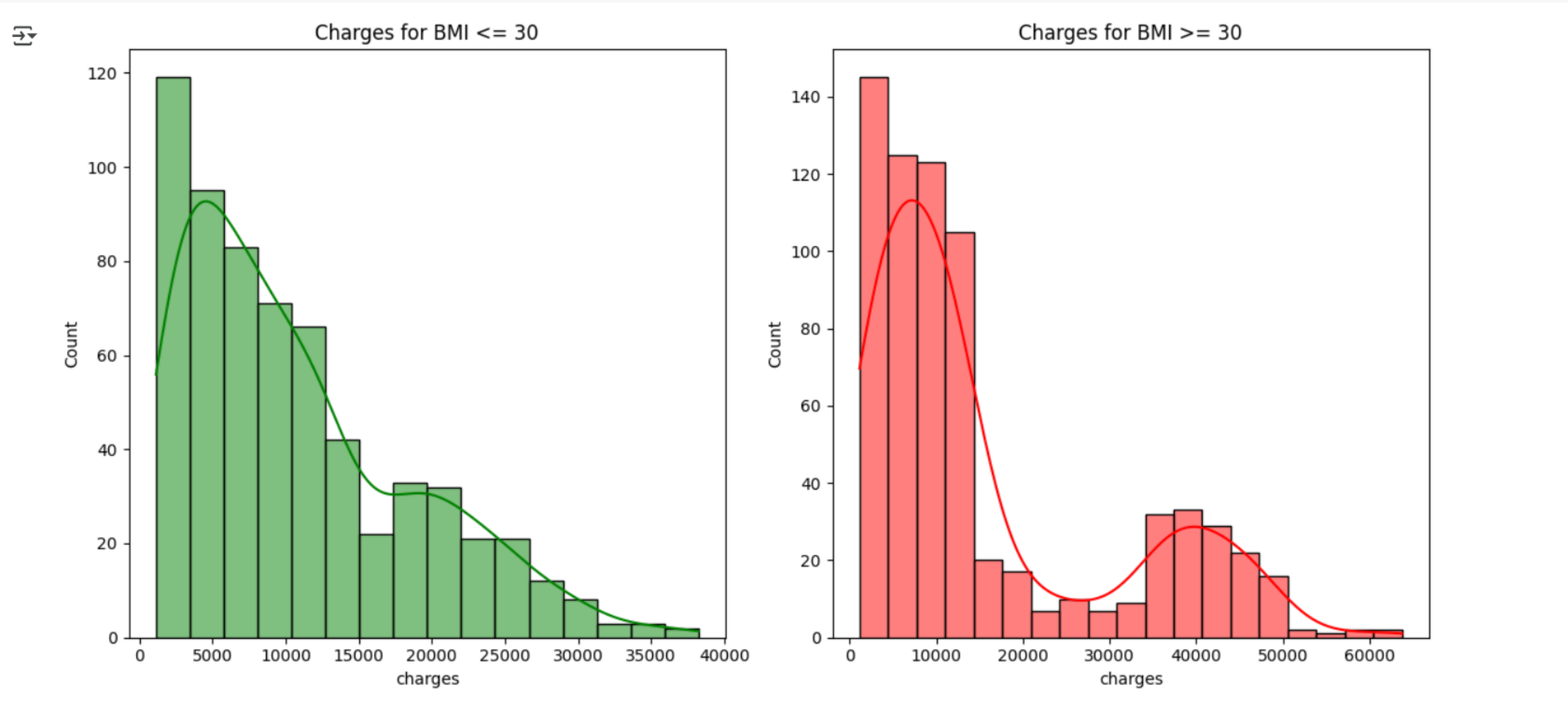
As the Age increases, Charges also increase.



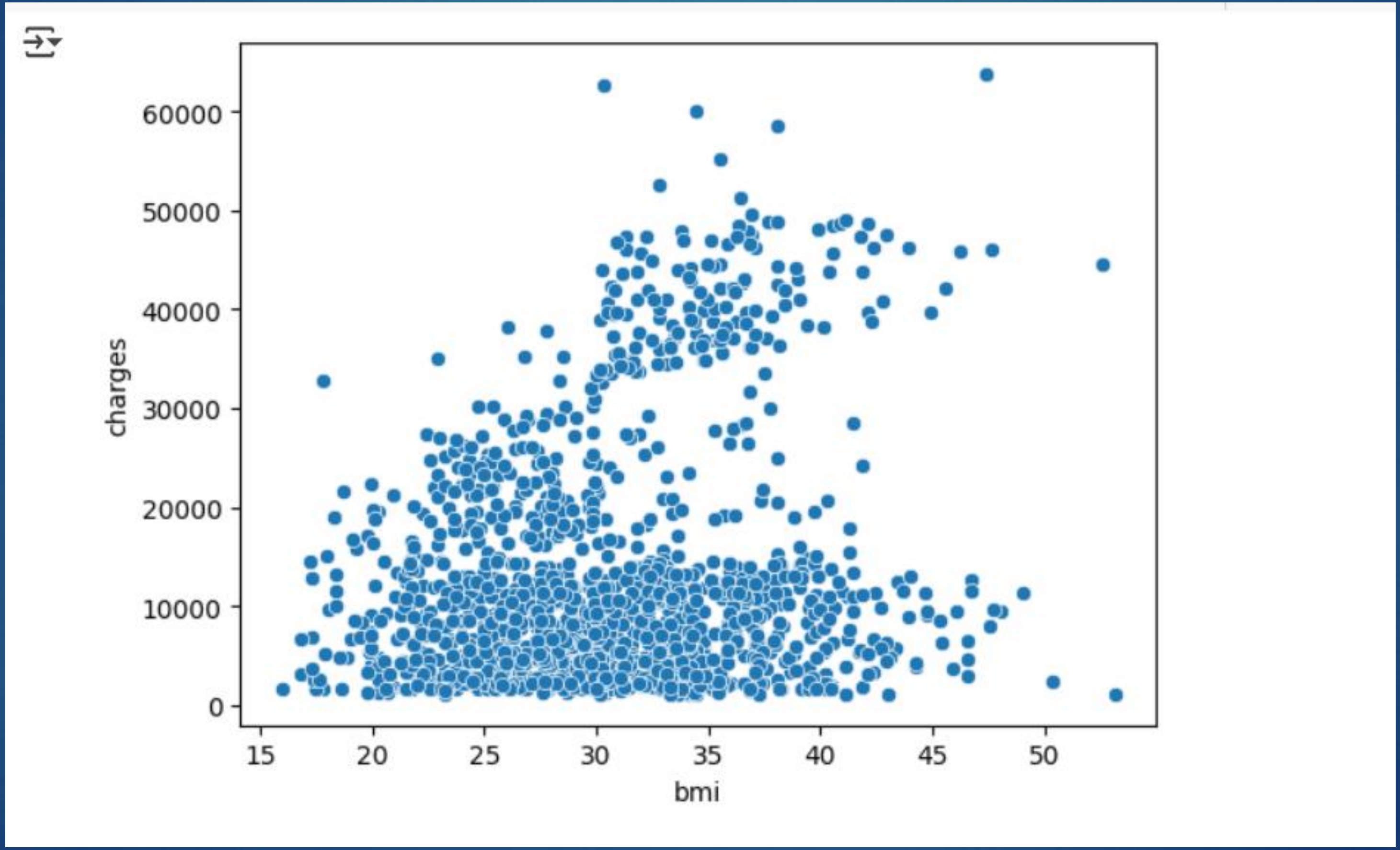
But if you smoke, then you pay high charges even if you are younger.



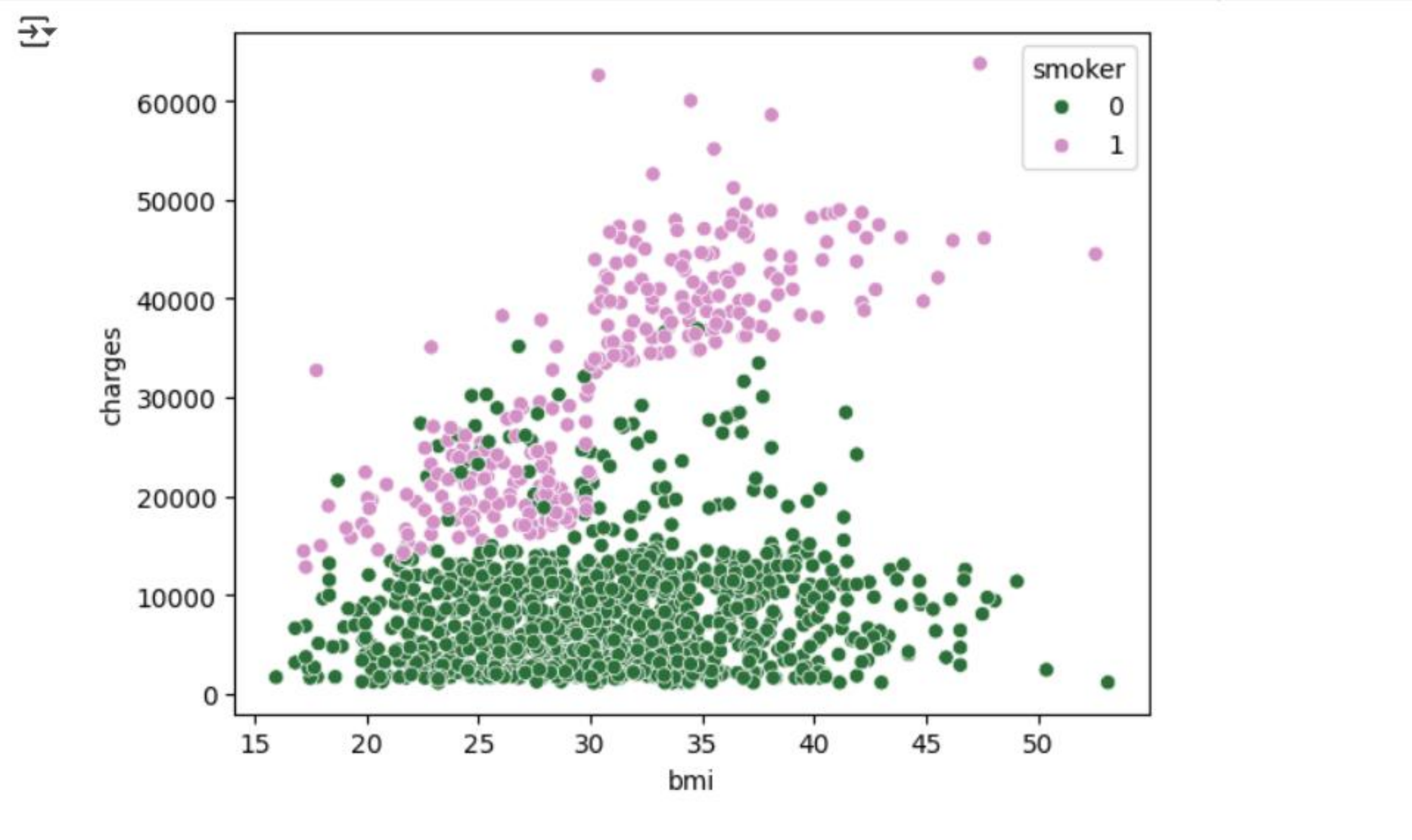
Generally, lower BMI means lower charges



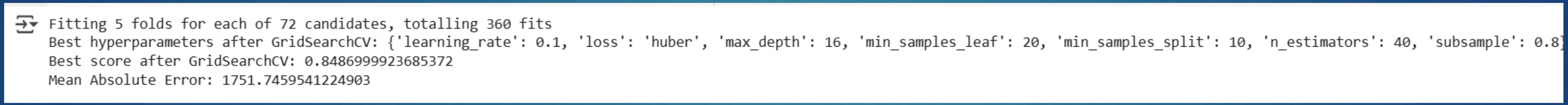
BMI does not have clear linear relationship with charges



But those who smoke pay high despite having lower BMI

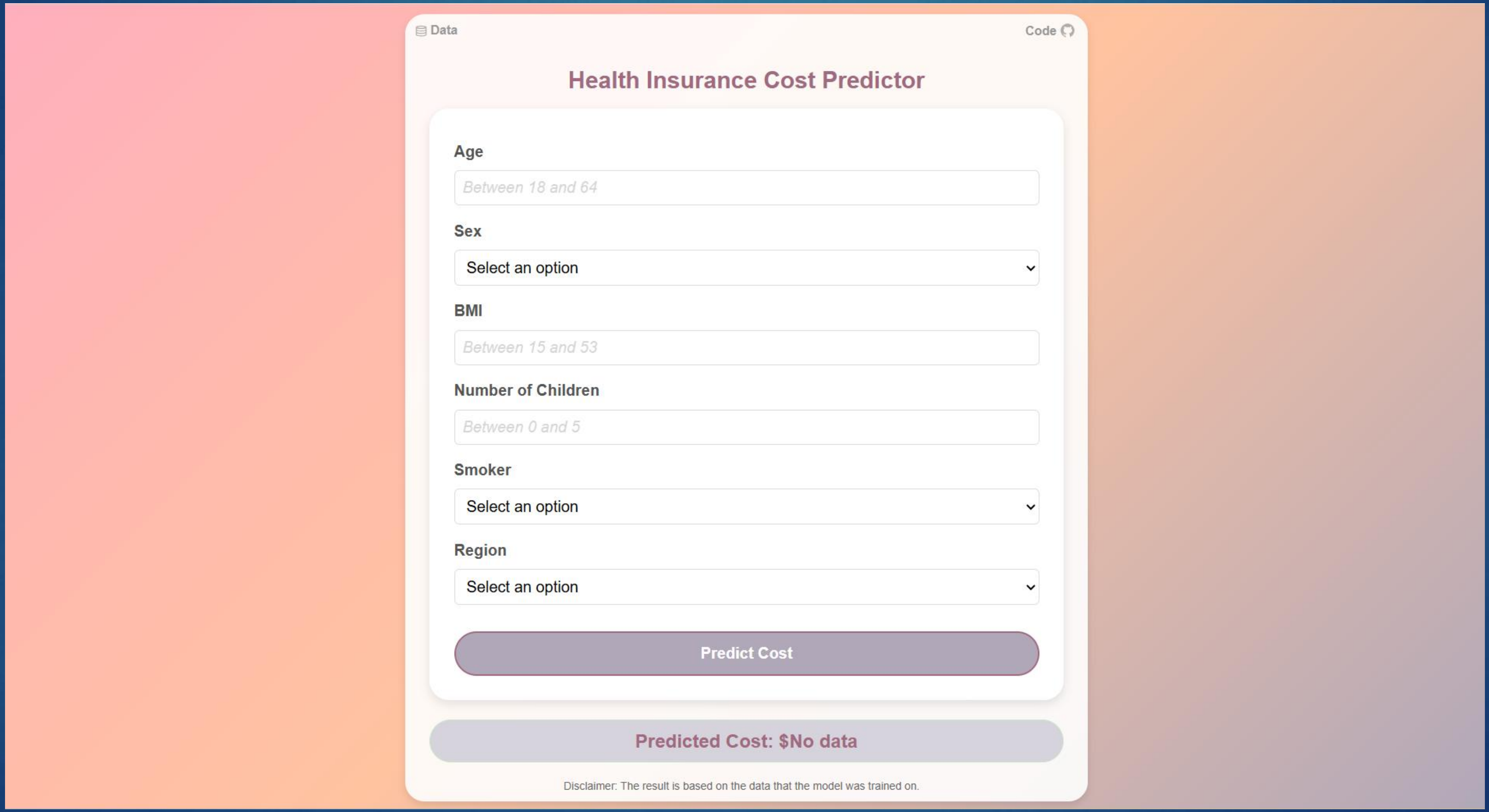


Best Model Parameters: GBM



```
Fitting 5 folds for each of 72 candidates, totalling 360 fits
Best hyperparameters after GridSearchCV: {'learning_rate': 0.1, 'loss': 'huber', 'max_depth': 16, 'min_samples_leaf': 20, 'min_samples_split': 10, 'n_estimators': 40, 'subsample': 0.8}
Best score after GridSearchCV: 0.8486999923685372
Mean Absolute Error: 1751.7459541224903
```

Web-App: https://insurance-cost-prediction-app.nn.r.appspot.com/



Health Insurance Cost Predictor

Age

Between 18 and 64

Sex

Select an option

BMI

Between 15 and 53

Number of Children

Between 0 and 5

Smoker

Select an option

Region

Select an option

Predicted Cost: $No data

Disclaimer: The result is based on the data that the model was trained on.