CAPSTONE PROJECT

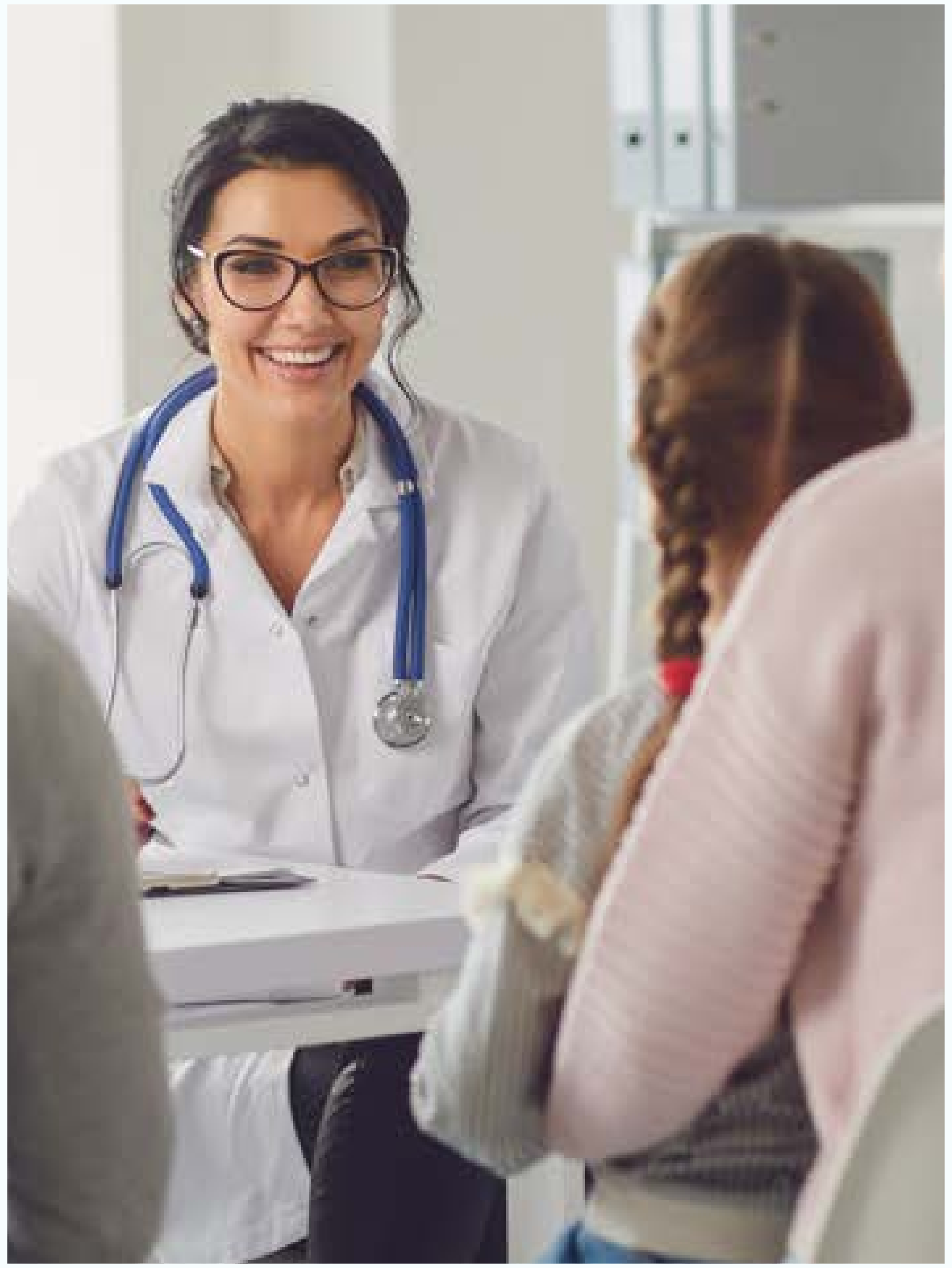
MEDICAL COST PREDICTION

Presented by
Parveen
Kanika
Meenu
Harsh

Instructor:- Jaspreet Gill

AGENDA

- Introduction
- Objective
- Data Exploration
- Data Visualization
- Model Building
- Results
- Conclusion



INTRODUCTION

- Our examination of healthcare expenditures utilizing the Medical Cost Personal Dataset sourced from Kaggle provides a valuable perspective on the determinants impacting healthcare expenses in the United States.
- The dataset contains **1338 observations and 7 variables**

  > Age

  > BMI

  > Children

  > Smoker

  > Region

  > Charges

OBJECTIVE

- Our objective is to analyze the correlation between features and construct a model capable of forecasting medical costs. This endeavor aims to provide proactive insights to the healthcare system, aiding in the formulation of policies and strategies.

DATA EXPLORATION

Rows & Columns

```
In [3]:  1 #number of rows and columns
         2 df.shape

Out[3]: (1338, 7)
```

Data Info

```
1 #Checking for missing values
2 df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
```

We have discovered that the variables "sex," "smoker," and "region" are categorical in nature. Consequently, we will proceed to convert them into numerical format.

DATA EXPLORATION

Categorical to Numerical conversion

```
#changing categorical variables to numerical
df['sex'] = df['sex'].map({'male':1,'female':0})
df['smoker'] = df['smoker'].map({'yes':1,'no':0})
df['region'] = df['region'].map({'southwest':0,'southeast':1,'northwest':2,'northeast':3})
```

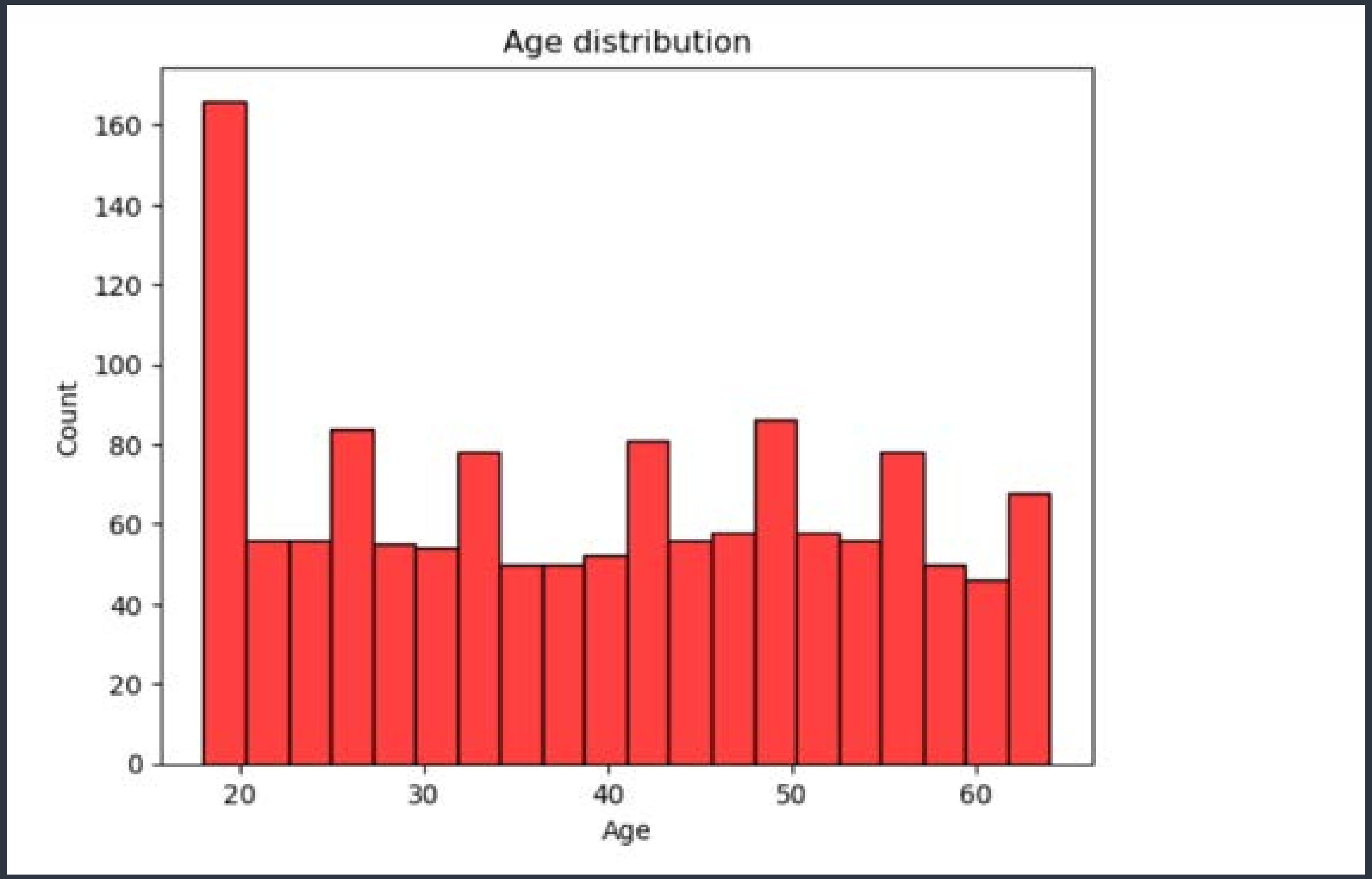
DATA EXPLORATION

```
#checking descriptive statistics
df.describe()
```

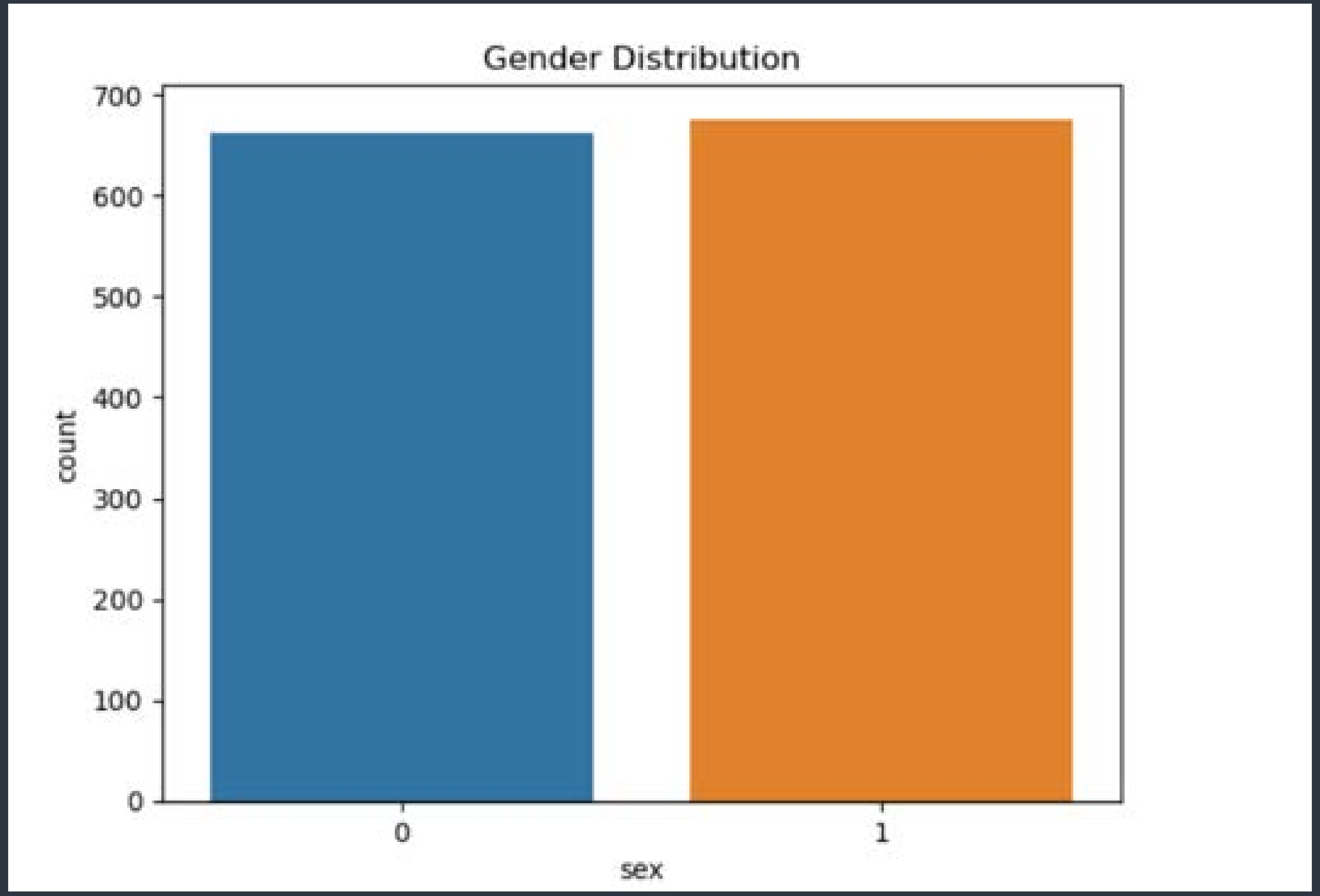
	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

5 Point Summary

EXPLORATORY DATA ANALYSIS

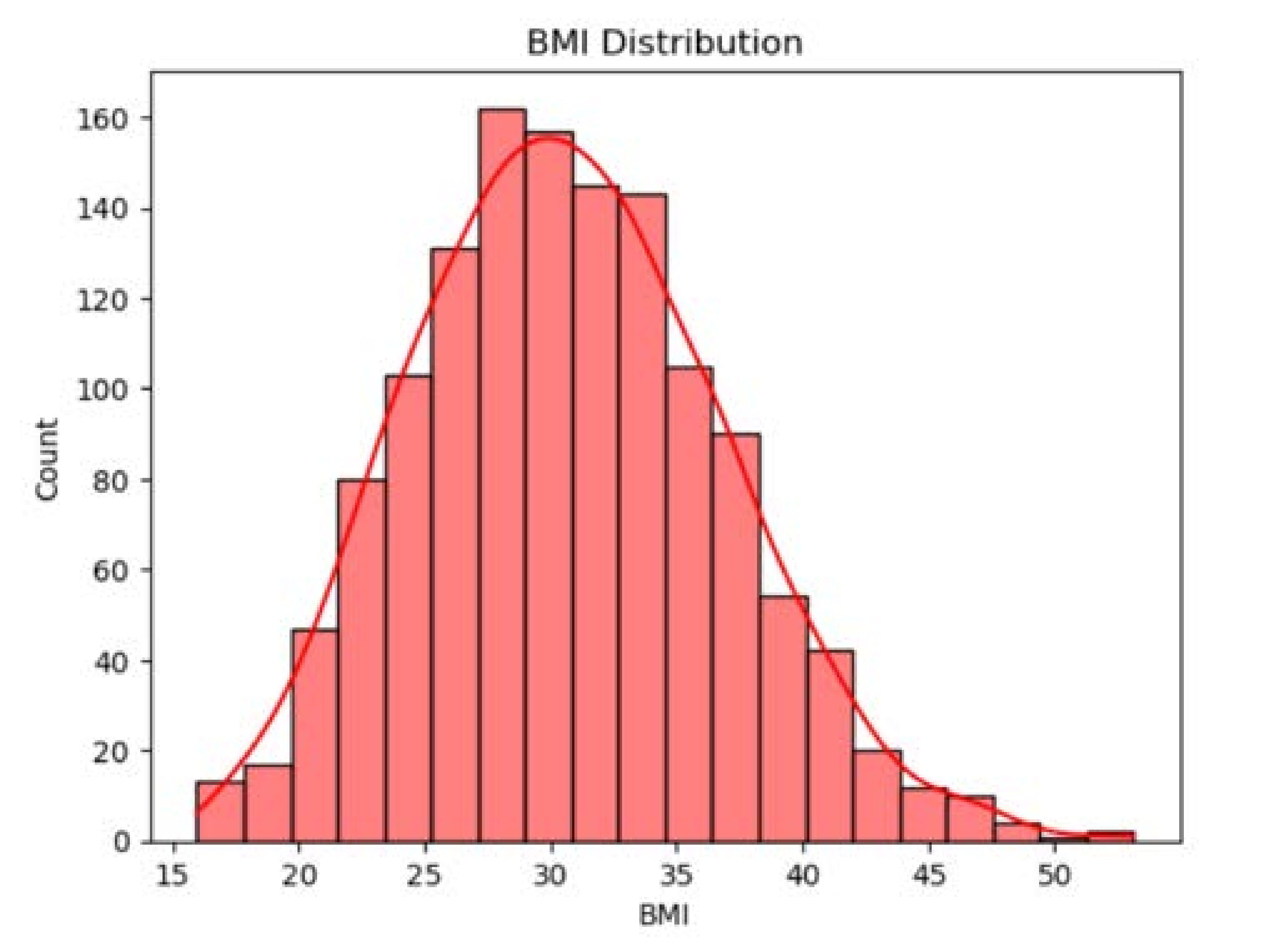


Gender Distribution (Almost Equal)



Age distribution (Max 19 year age)

EXPLORATORY DATA ANALYSIS

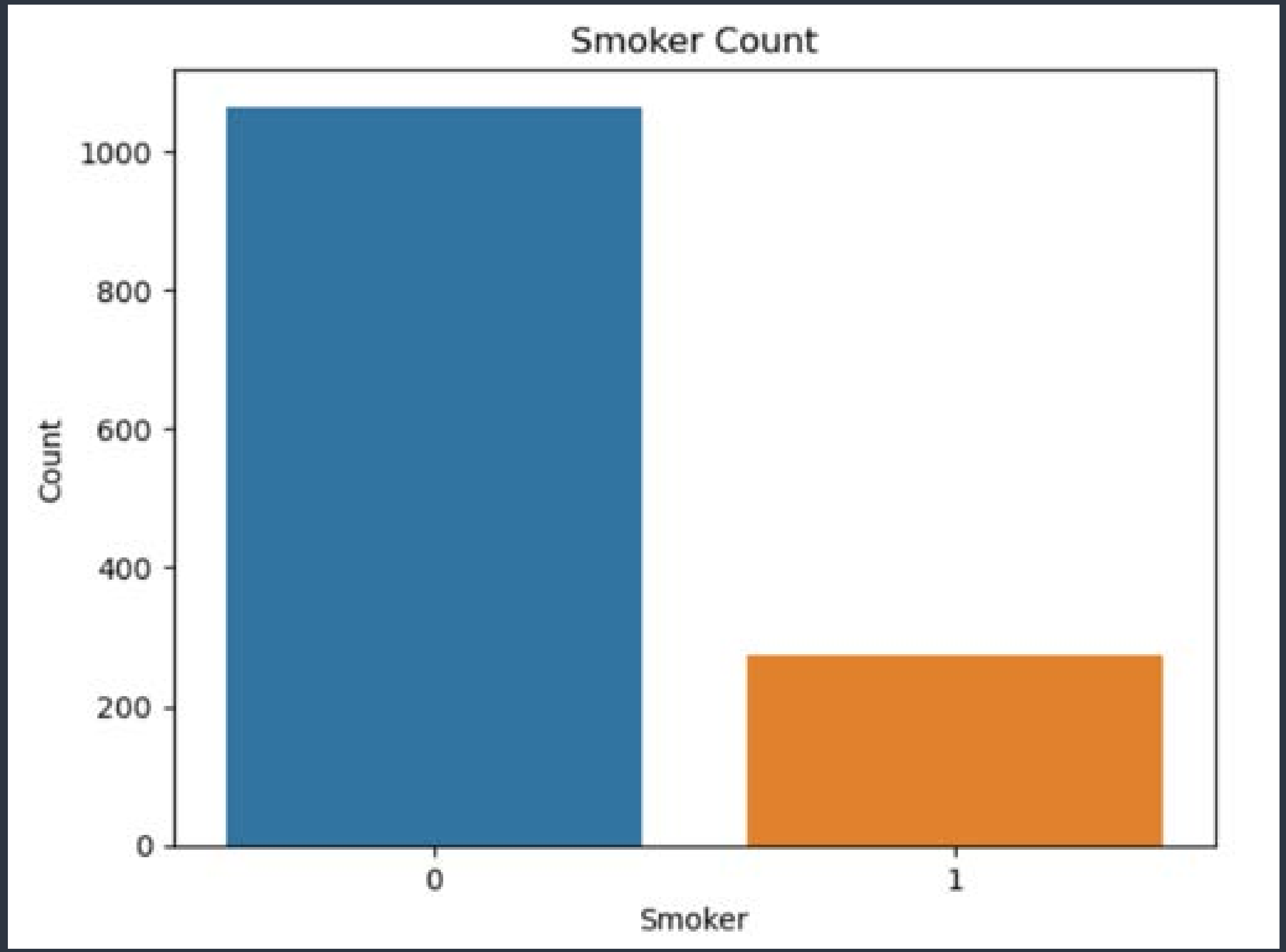


BMI distribution

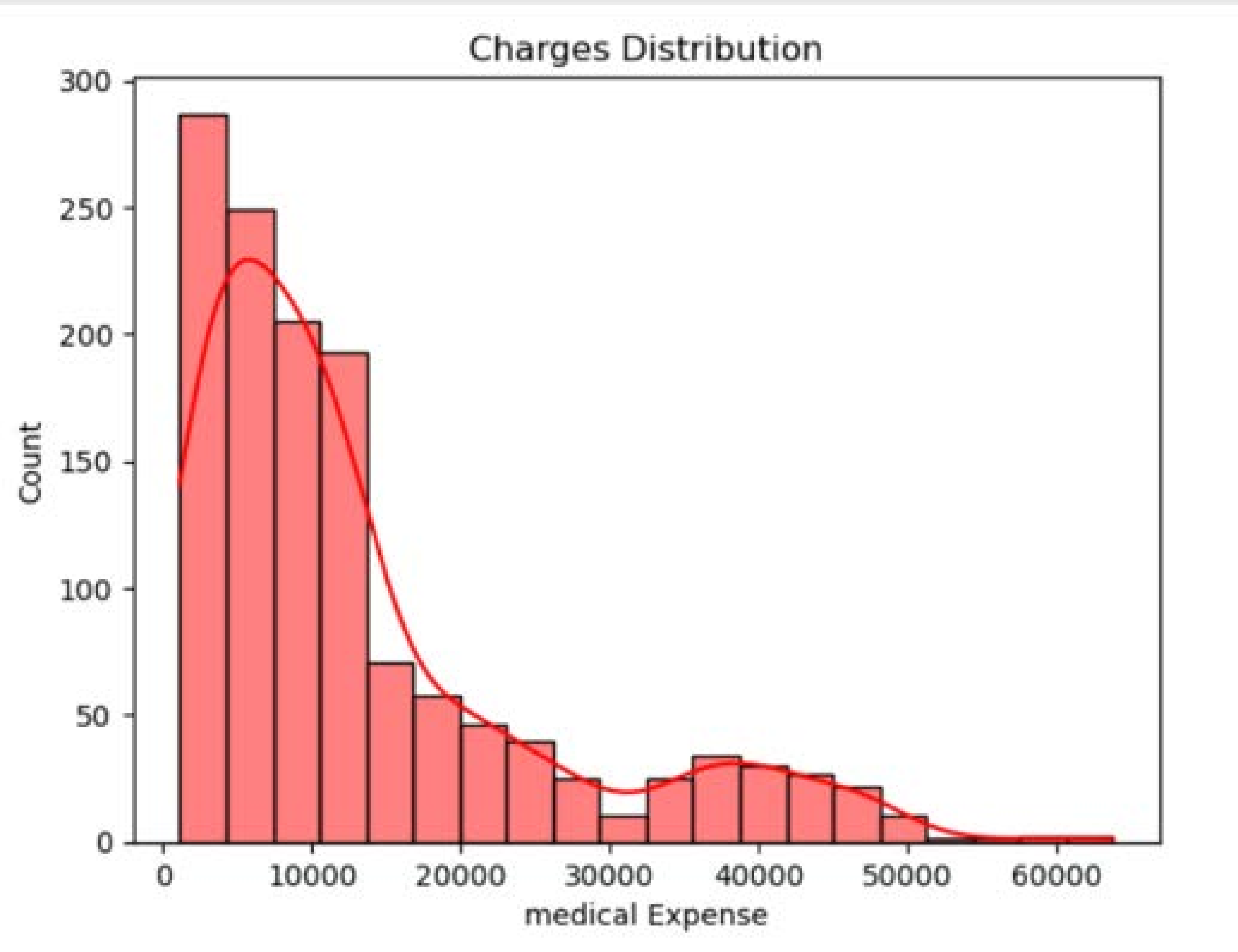
(Between 25 to 40, overweight)

Smoker Distribution

(Almost 80% are non smokers)



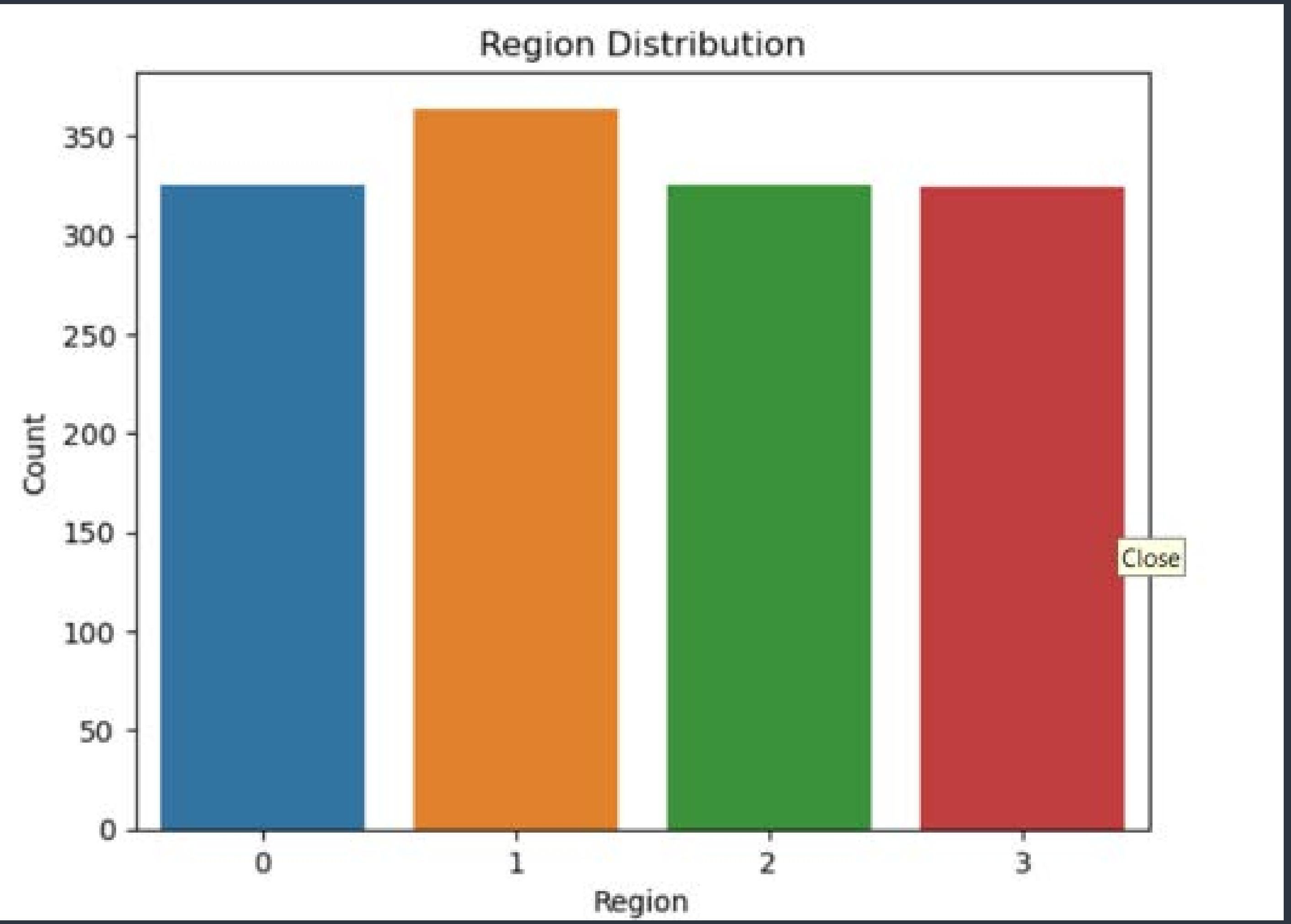
EXPLORATORY DATA ANALYSIS



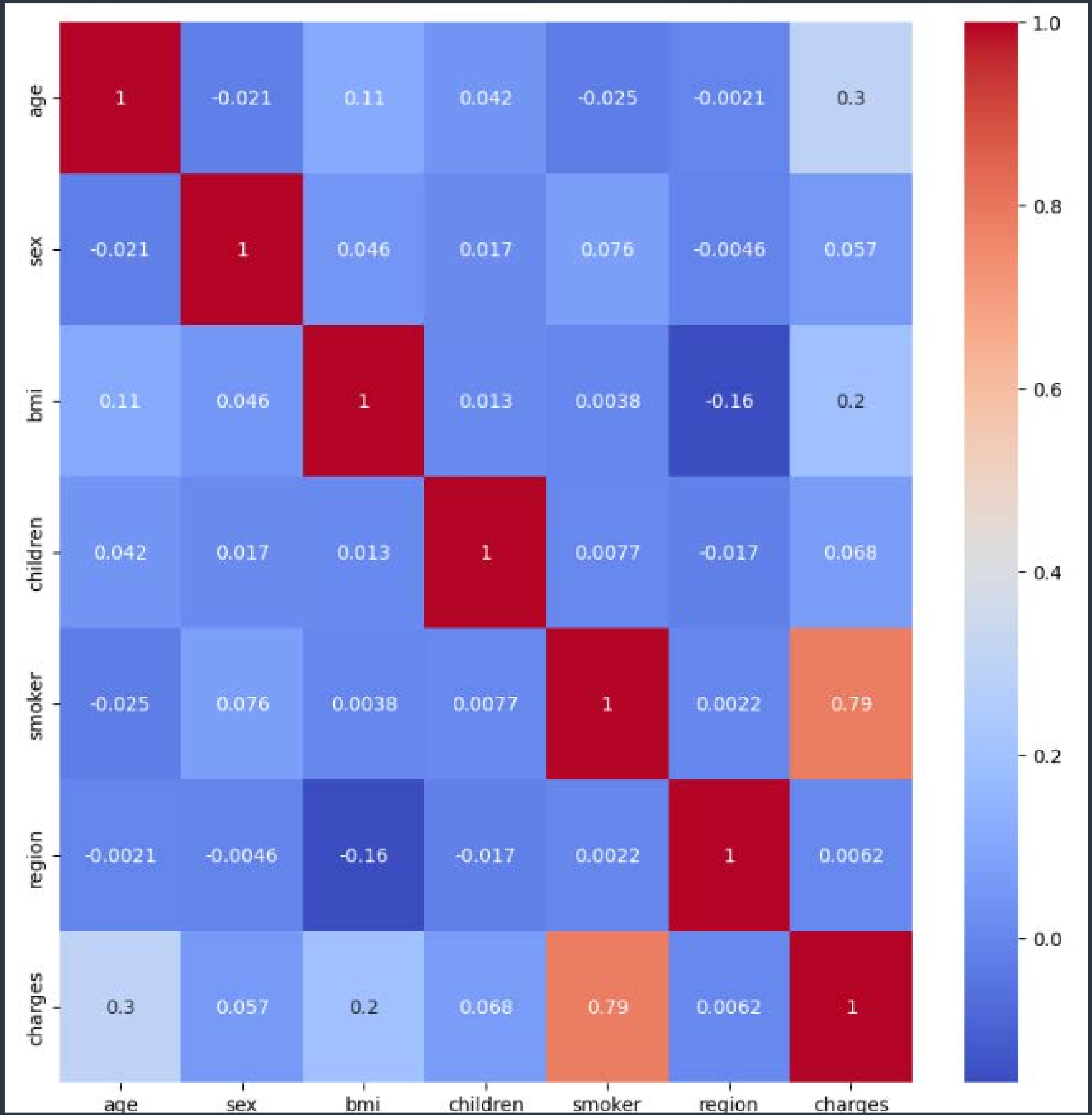
Charges distribution,

Most of the expenses below 20000

Region Distribution,SE have slightly higher no. of patients



CORRELATION

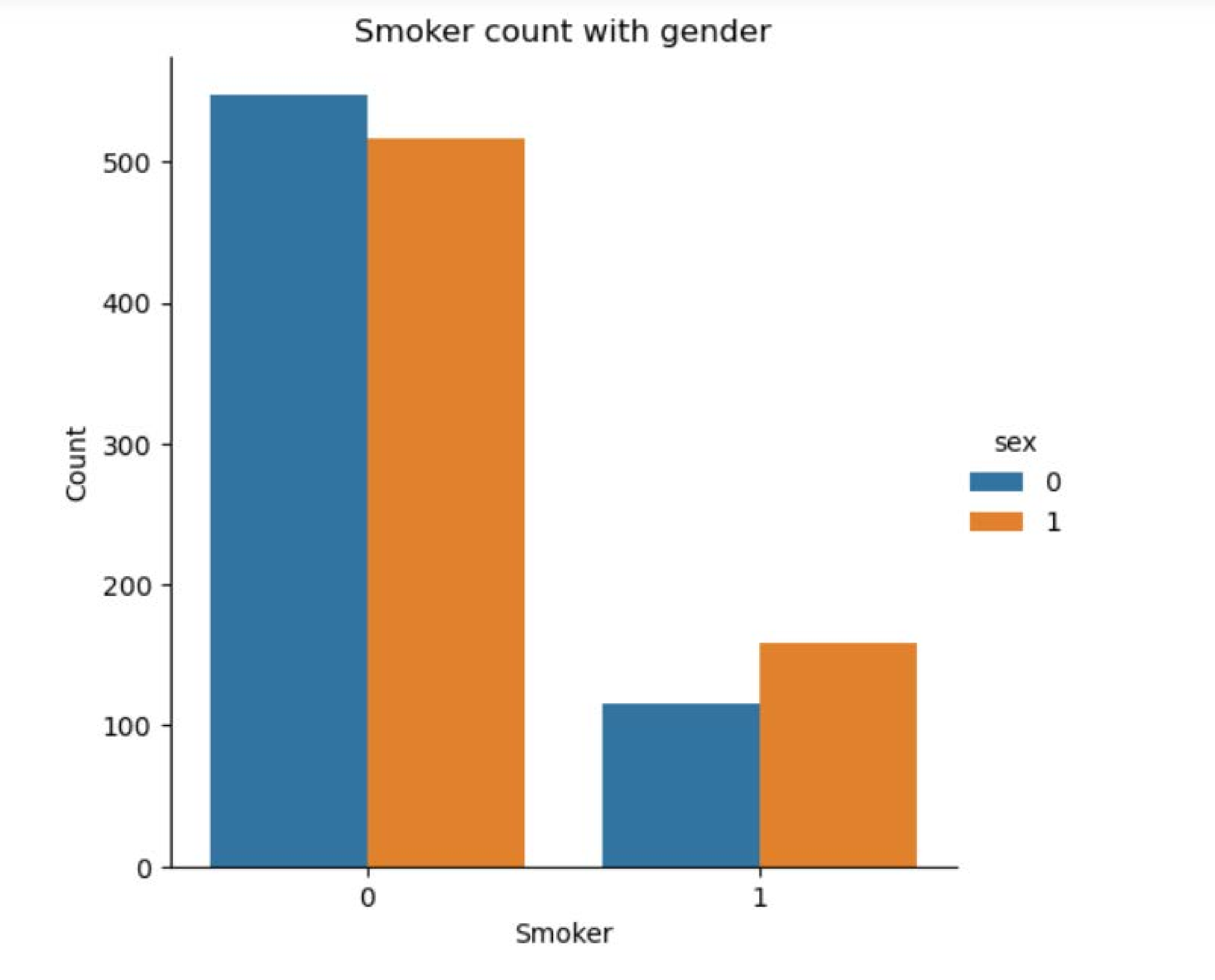


<u>**Correlation heat map**</u>**,**

The variable smoker shows a significant correlation with medical expenses.

DATA VISUALISATION

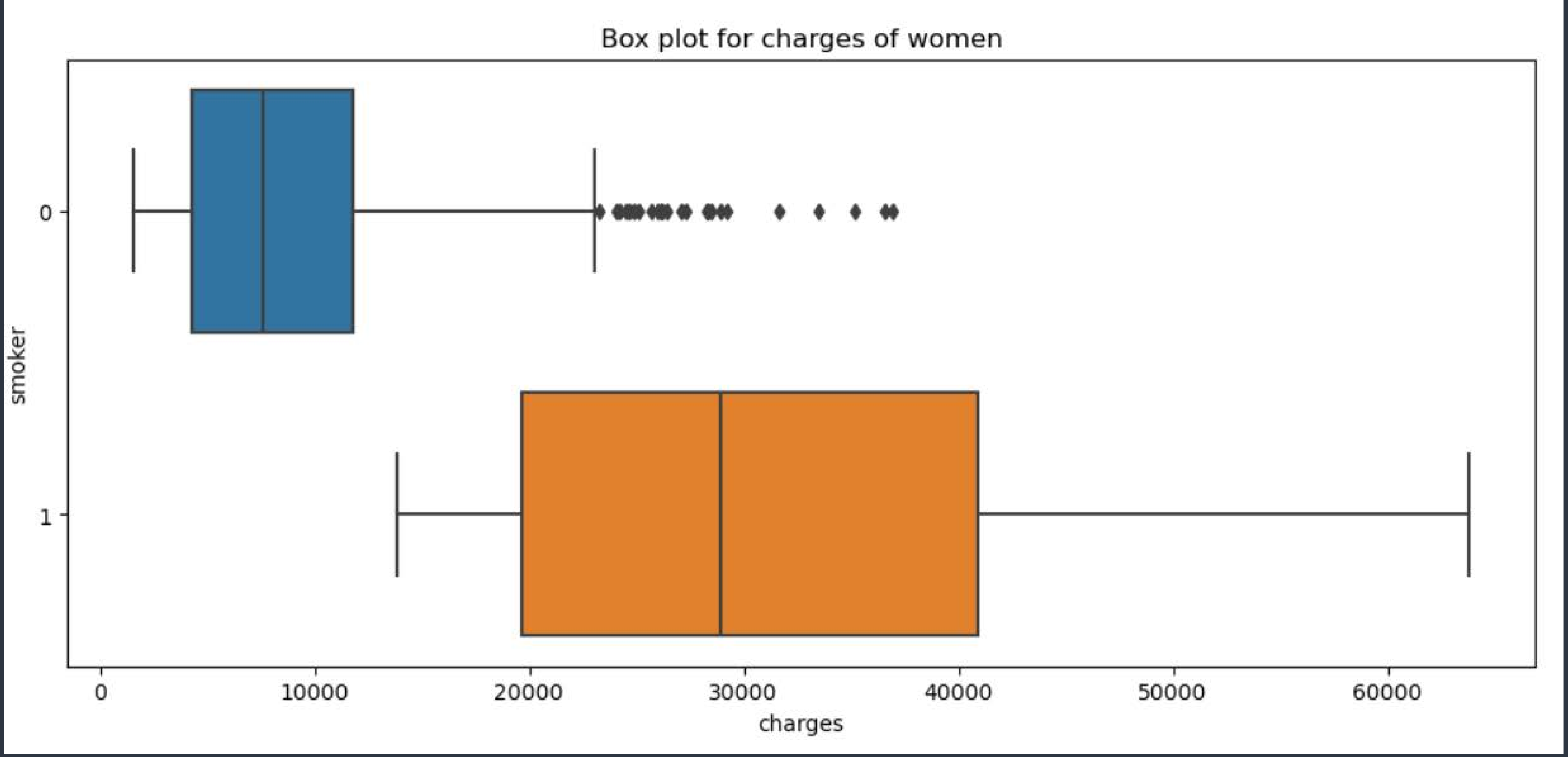
GENDER SMOKER COUNT



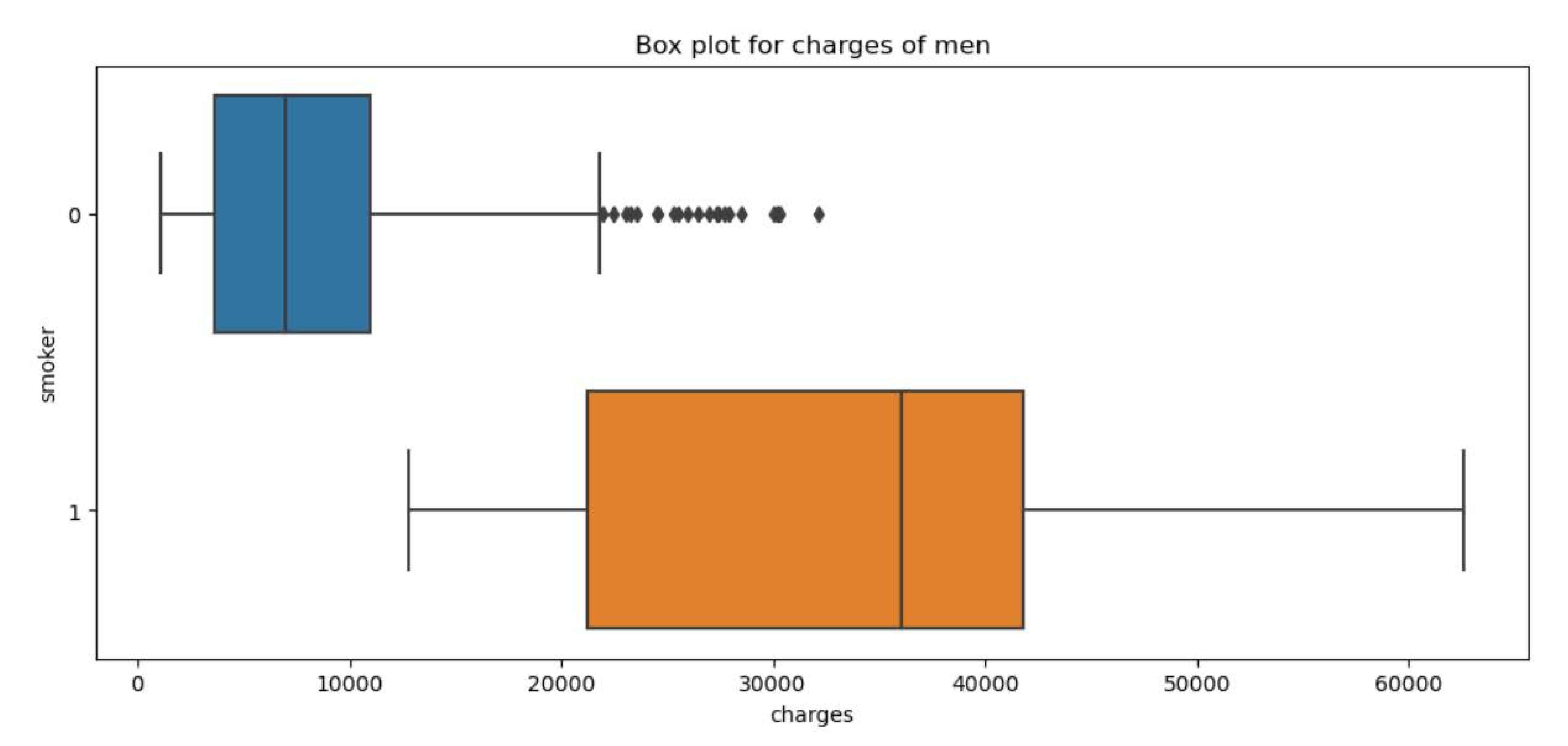
The gender smoker count Clearly describes more male smokers than female smokers.

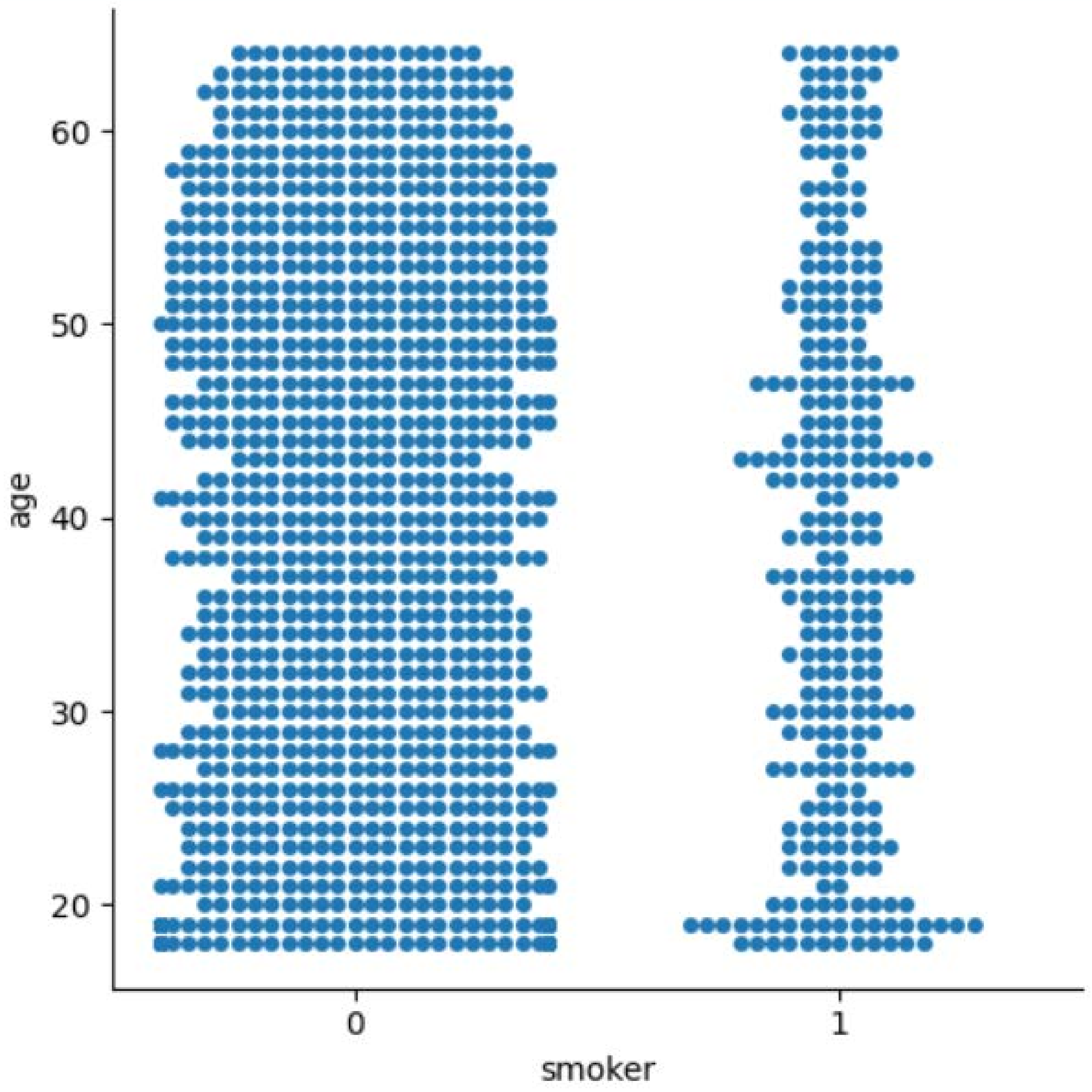
So we will ensure that more expenses will be for male patients than females.

BOX PLOTS FOR FEMALE EXPENSES



BOX PLOTS FOR MALE EXPENSES

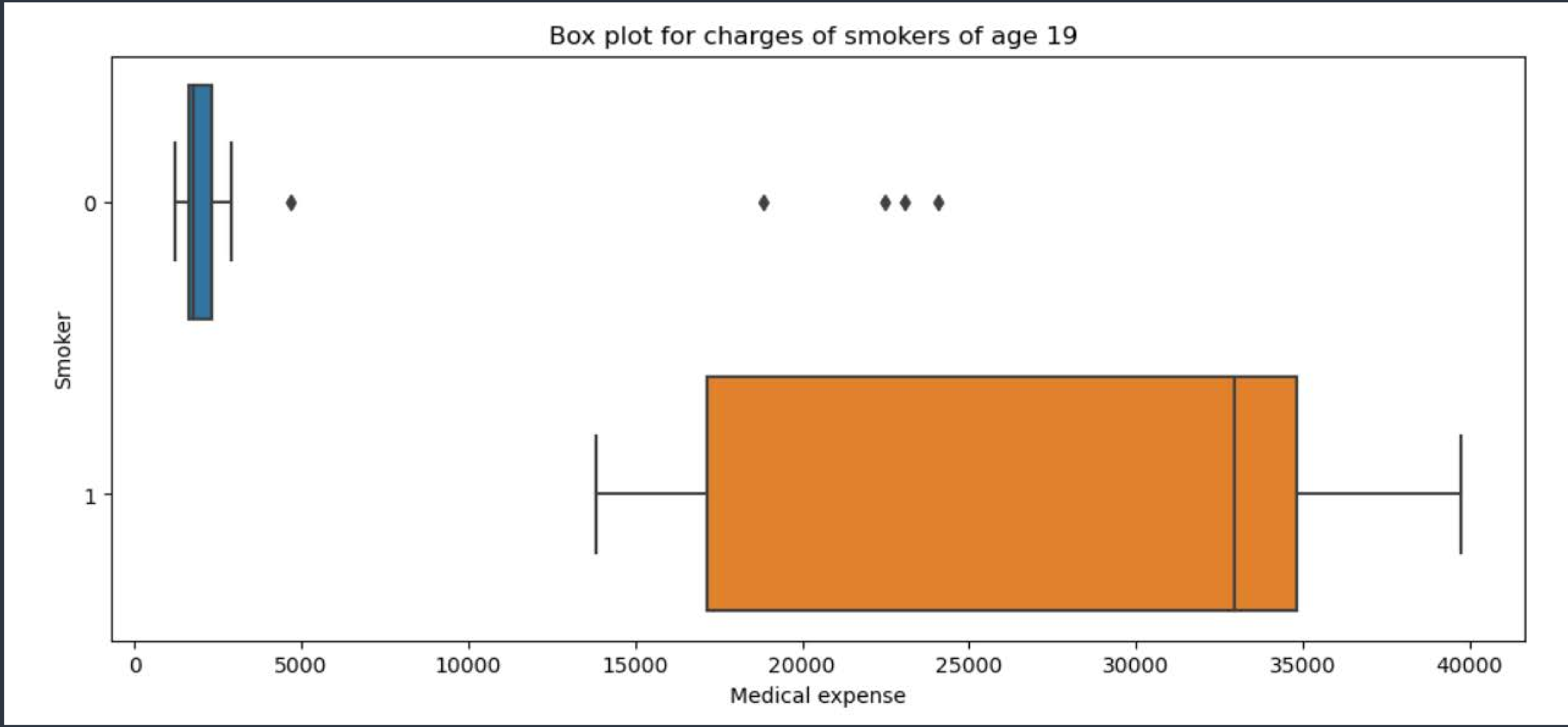




SMOKER AND AGE DISTRIBUTION

FROM THE GRAPH,
WE CAN SEE THAT THERE SIGNIFICANT NUMBER OF SMOKER OF AGE 19.

BOX PLOT FOR CHARGES OF SMOKERS OF AGE 19

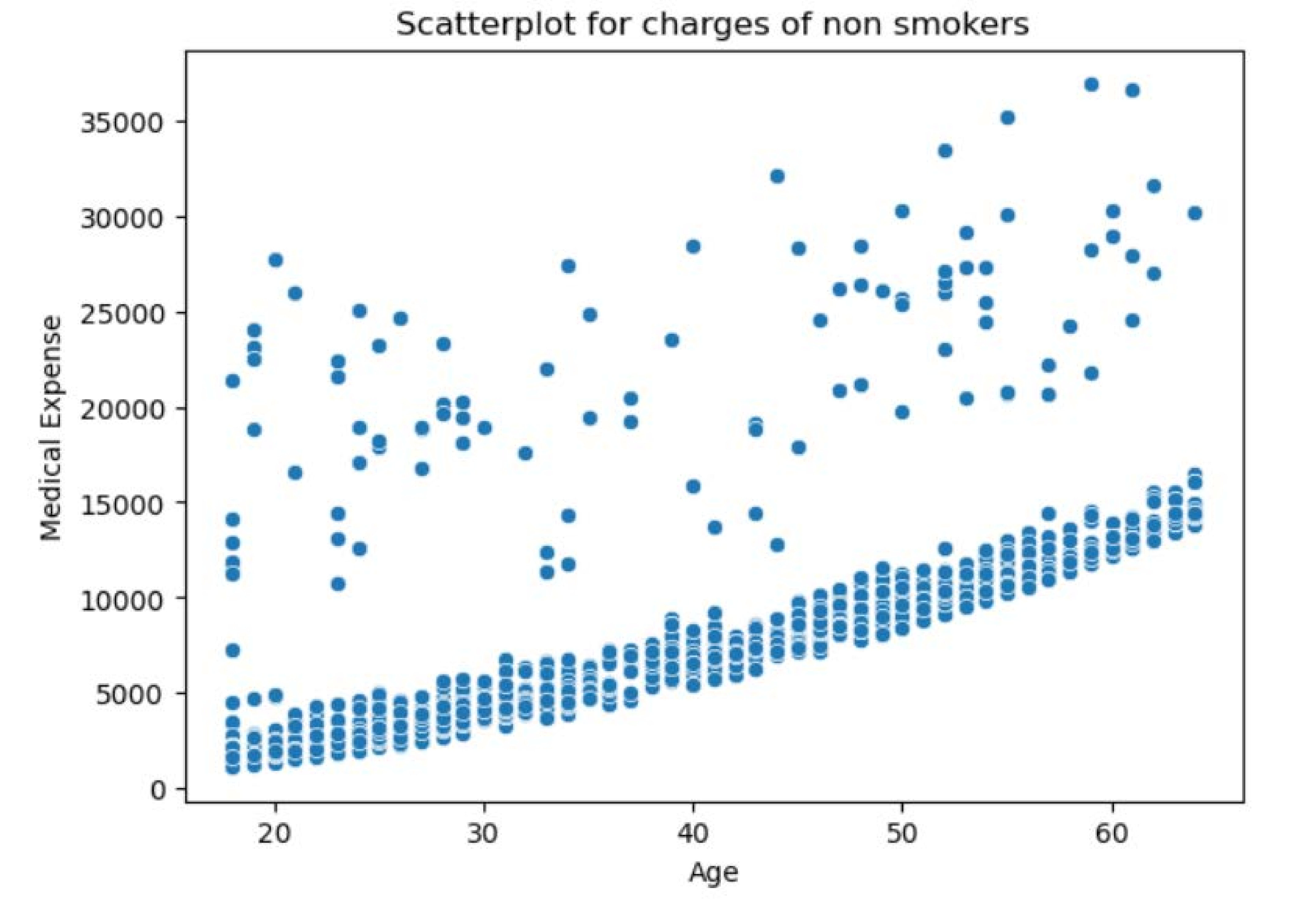


SURPRISINGLY THE MEDICAL EXPENSE OF SMOKERS OF AGE 19 IS VERY HIGH IN COMPARISON TO NON SMOKERS.

THE MEDICAL EXPENSE OF SMOKERS IS HIGHER THAN THAT OF NON-SMOKERS.

NOW LET'S PLOT THE CHARGES DISTRIBUTION CONCERNING PATIENT'S AGES OF SMOKERS AND NON-SMOKERS.

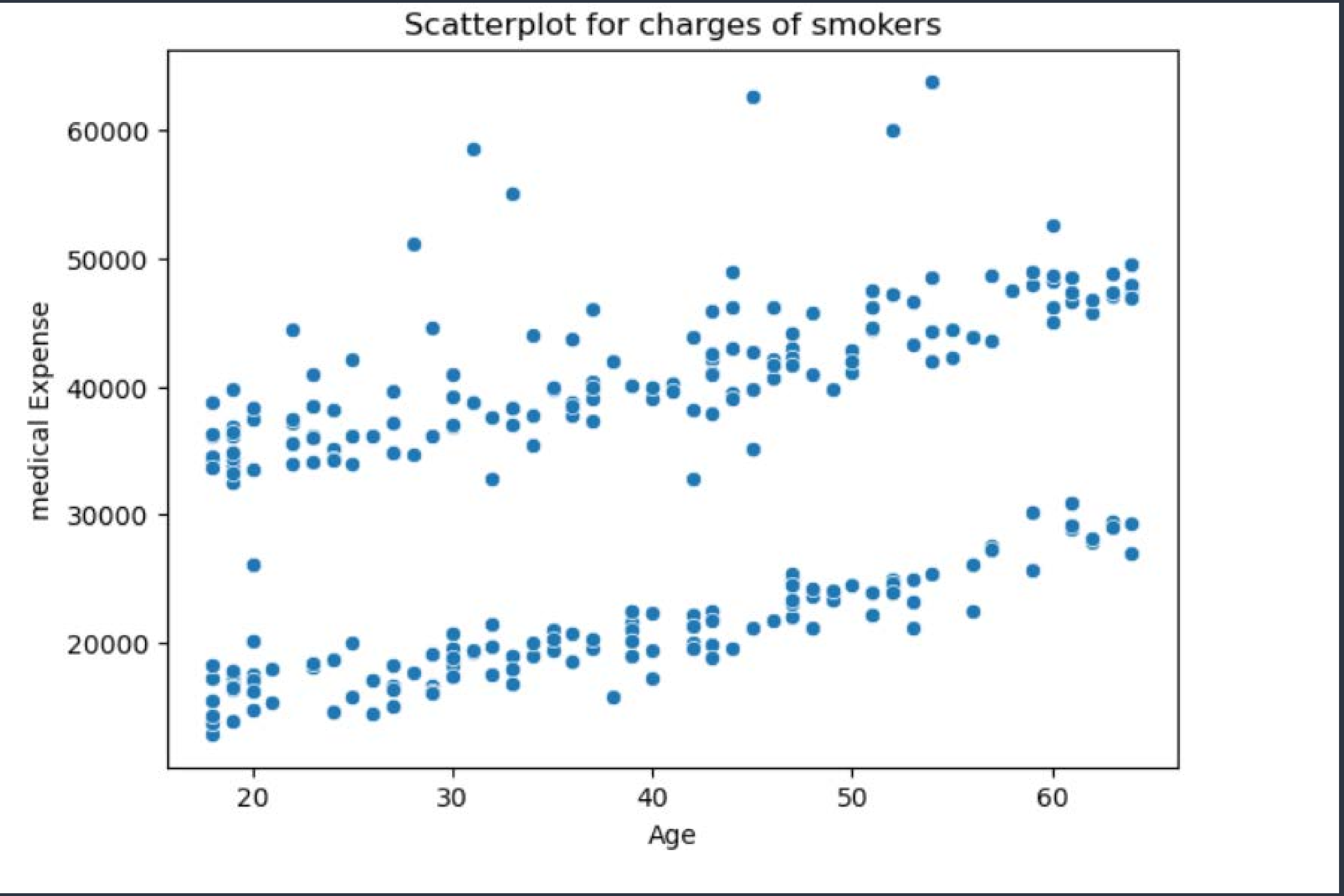
SCATTER PLOT FOR CHARGES OF NON SMOKERS



Majority of the points show that medical expense increases with age which may be due to the fact that older people are more prone to illness.

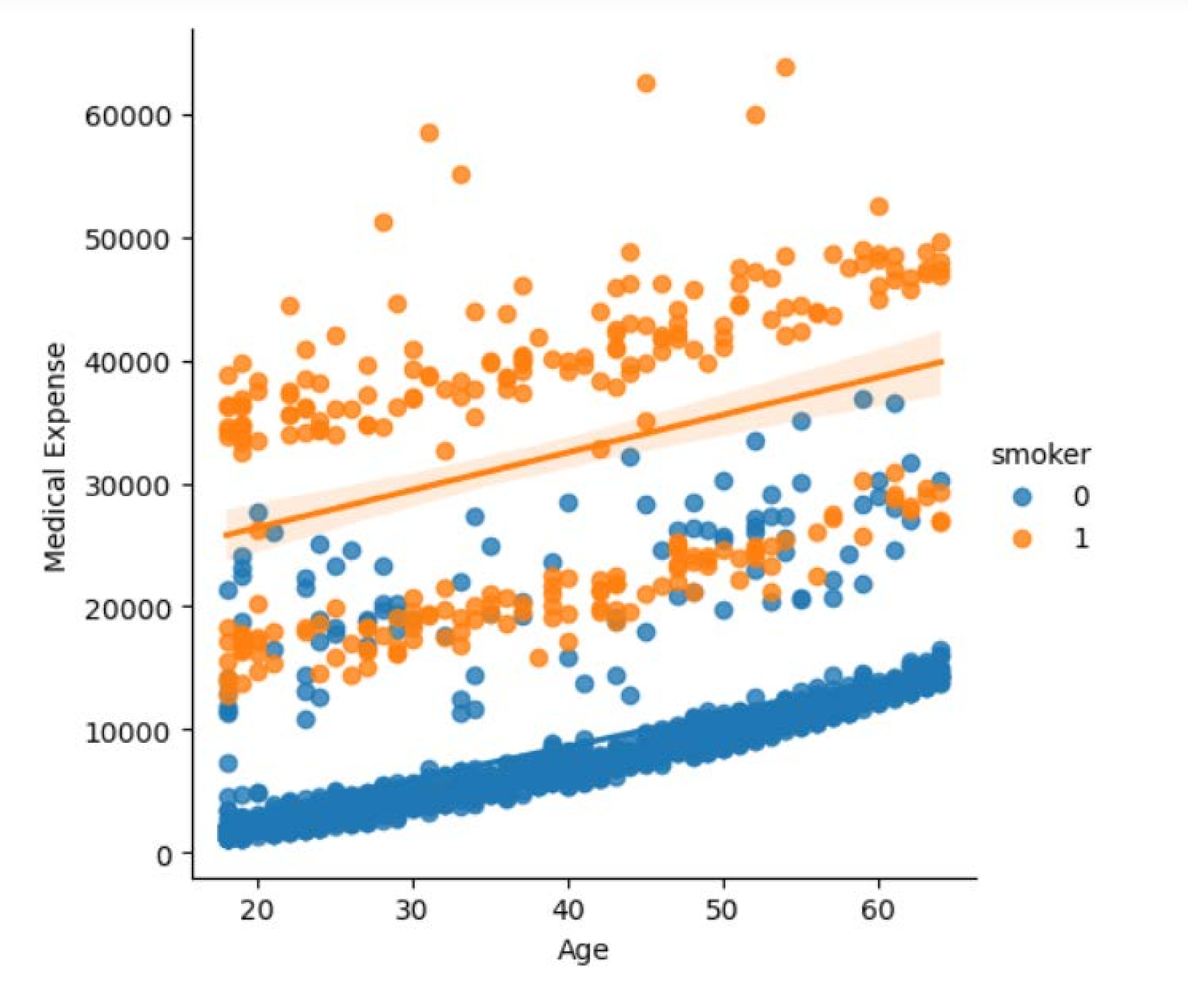
But there are some outliners which shows that there are other illness or accidents which may increase the medical expense.

SCATTER PLOT FOR CHARGES OF SMOKERS



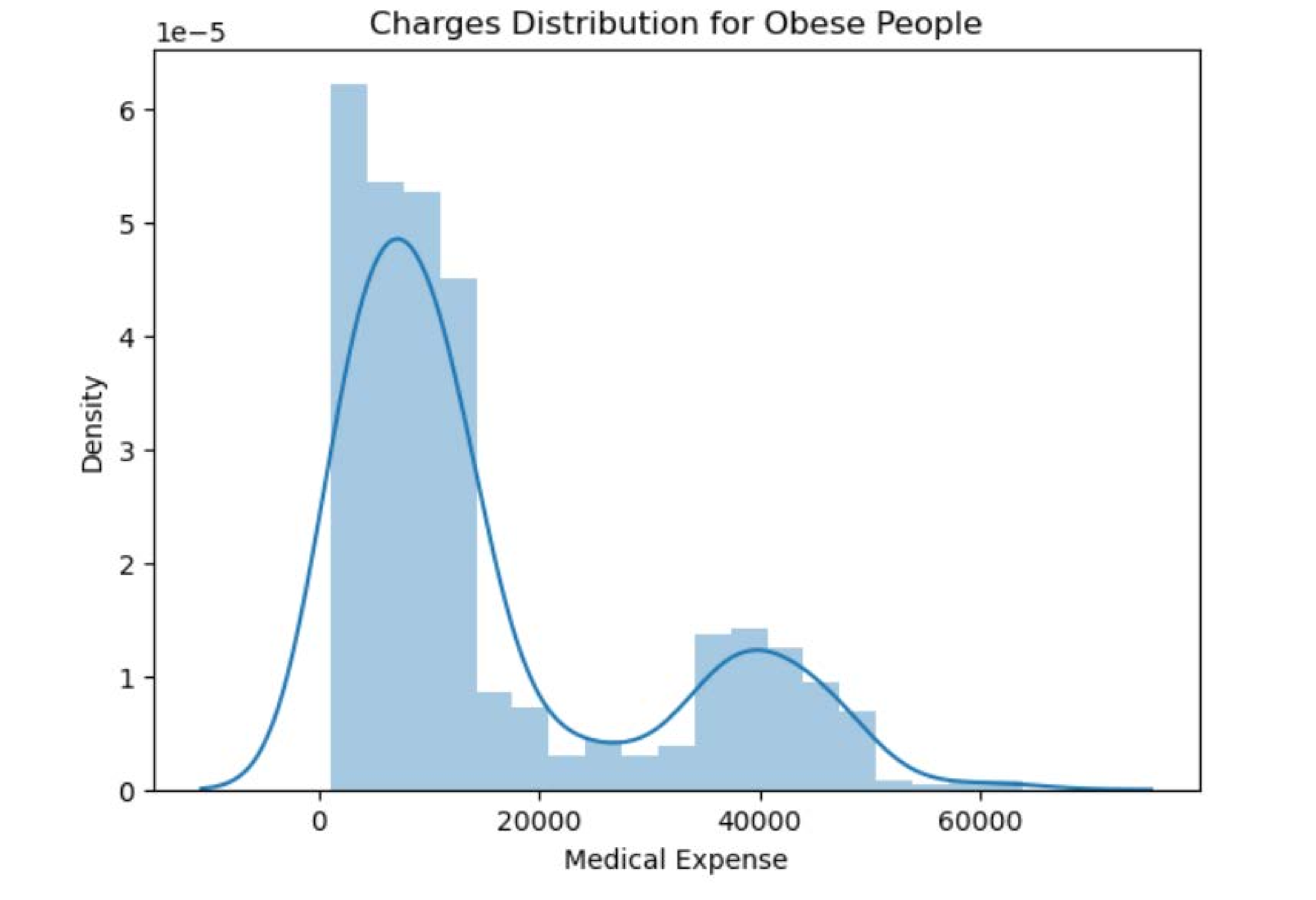
In the graph, there are two segments, one with high medical expenses which may be due to smoking related illness and the other with low medical expenses which may be due age related illness.

COMBINED GRAPH

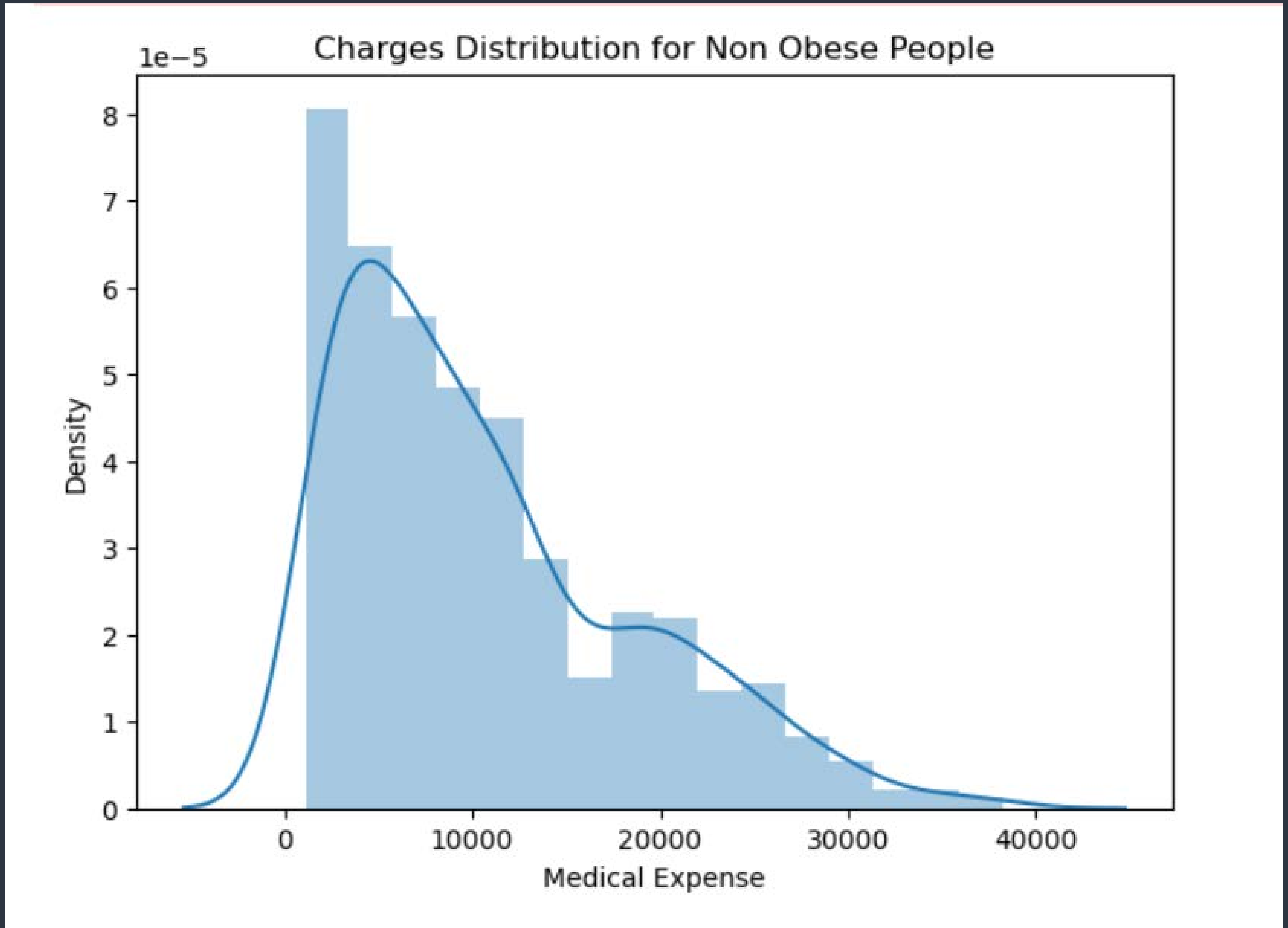


Now, we clearly understand the variation in charges with respect to age and smoking habit. The medical expense of smokers is higher than that of non - smokers. In non-smokers, the cost of treatment increase with age which is obvious. But in smokers, the cost of treatment is high even for younger patients, which means the smoking patients are spending upon their smoking related illness as well as age related illness.

CHARGES DISTRIBUTION FOR PATIENTS WITH BMI > 30 I.E. OBESE PATIENTS



CHARGES DISTRIBUTION FOR PATIENTS WITH BMI < 30 I.E. HEALTHY PATIENTS



Therefore, patients with BMI less than 30 are spending less on medical treatment than those with BMI greater than 30.

MODEL BUILDING

Linear Regression

In [31]:
```
#Linear Regression
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr
```

Out[31]:
```
▾ LinearRegression
LinearRegression()
```

In [32]:
```
#model training
lr.fit(x_train, y_train)
#model accuracy
lr.score(x_train,y_train)
```

Out[32]:
```
0.7368306228430945
```

In [33]:
```
#Model prediction
y_pred = lr.predict(x_test)
```

LINEAR REGRESSION

POLYNOMIAL REGRESSION

Polynomial Regression

[34]:
```
from sklearn.preprocessing import PolynomialFeatures
poly_reg = PolynomialFeatures(degree = 2)
poly_reg
```

t[34]:
```
▾ PolynomialFeatures
PolynomialFeatures()
```

[35]:
```
#transforming the features to higher degree
x_train_poly = poly_reg.fit_transform(x_train)
#splitting the data
x_train, x_test, y_train, y_test = train_test_split(x_train_poly, y_train, test_size = 0.2, random_state = 0)

```

[36]:
```
plr = LinearRegression()
#model training
plr.fit(x_train, y_train)
#model accuracy
plr.score(x_train, y_train)
```

t[36]:
```
0.8372892283870186
```

[49]:
```
#model prediction
y_pred = plr.predict(x_test)
```

DECISION TREE REGRESSION

Decision Tree Regressor

In [38]:
```
#decision tree regressor
from sklearn.tree import DecisionTreeRegressor
dtree = DecisionTreeRegressor()
dtree
```

Out[38]:
```
▾ DecisionTreeRegressor
DecisionTreeRegressor()
```

In [39]:
```
#model training
dtree.fit(x_train,y_train)
#model accuracy
dtree.score(x_train,y_train)
```

Out[39]: 0.9993688476658964

In [40]:
```
#model prediction
dtree_pred = dtree.predict(x_test)
```

RANDOM FOREST REGRESSOR

Randorm Forest Regressor

In [41]:

```python
#random forest regressor
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor(n_estimators=100)
rf
```

Out[41]:

```
▾ RandomForestRegressor
RandomForestRegressor()
```

In [42]:

```python
#model training
rf.fit(x_train,y_train)
#model accuracy
rf.score(x_train,y_train)
```

Out[42]:

```
0.9754114505615482
```

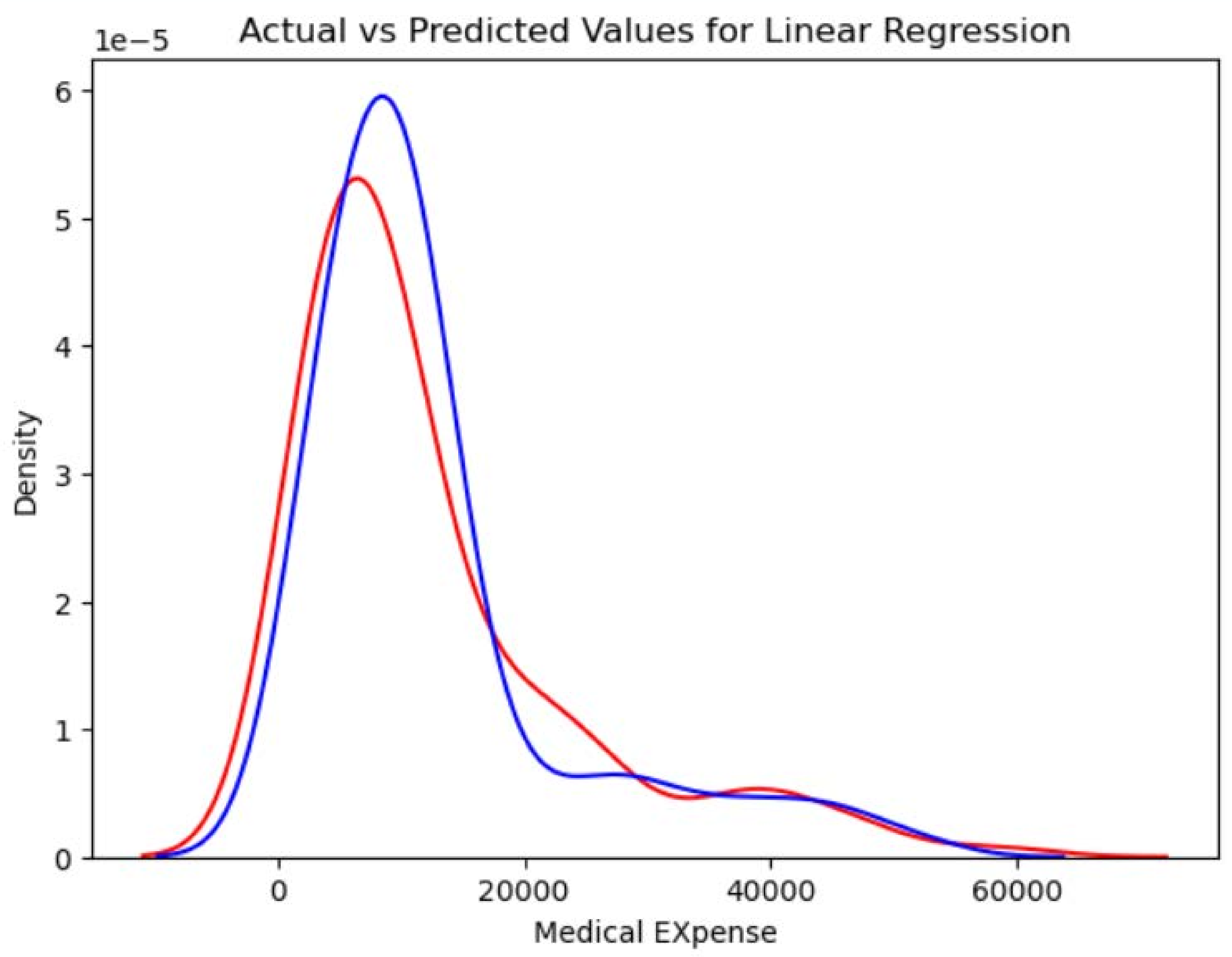
In [43]:

```python
#model Prediction
rf_pred = rf.predict(x_test)
```

RESULT

```
print('MAE:', mean_absolute_error(y_test, y_pred))
print('MSE:', mean_squared_error(y_test, y_pred))
print('RMSE:', np.sqrt(mean_squared_error(y_test, y_pred)))
print('R2 Score:', r2_score(y_test, y_pred))
```

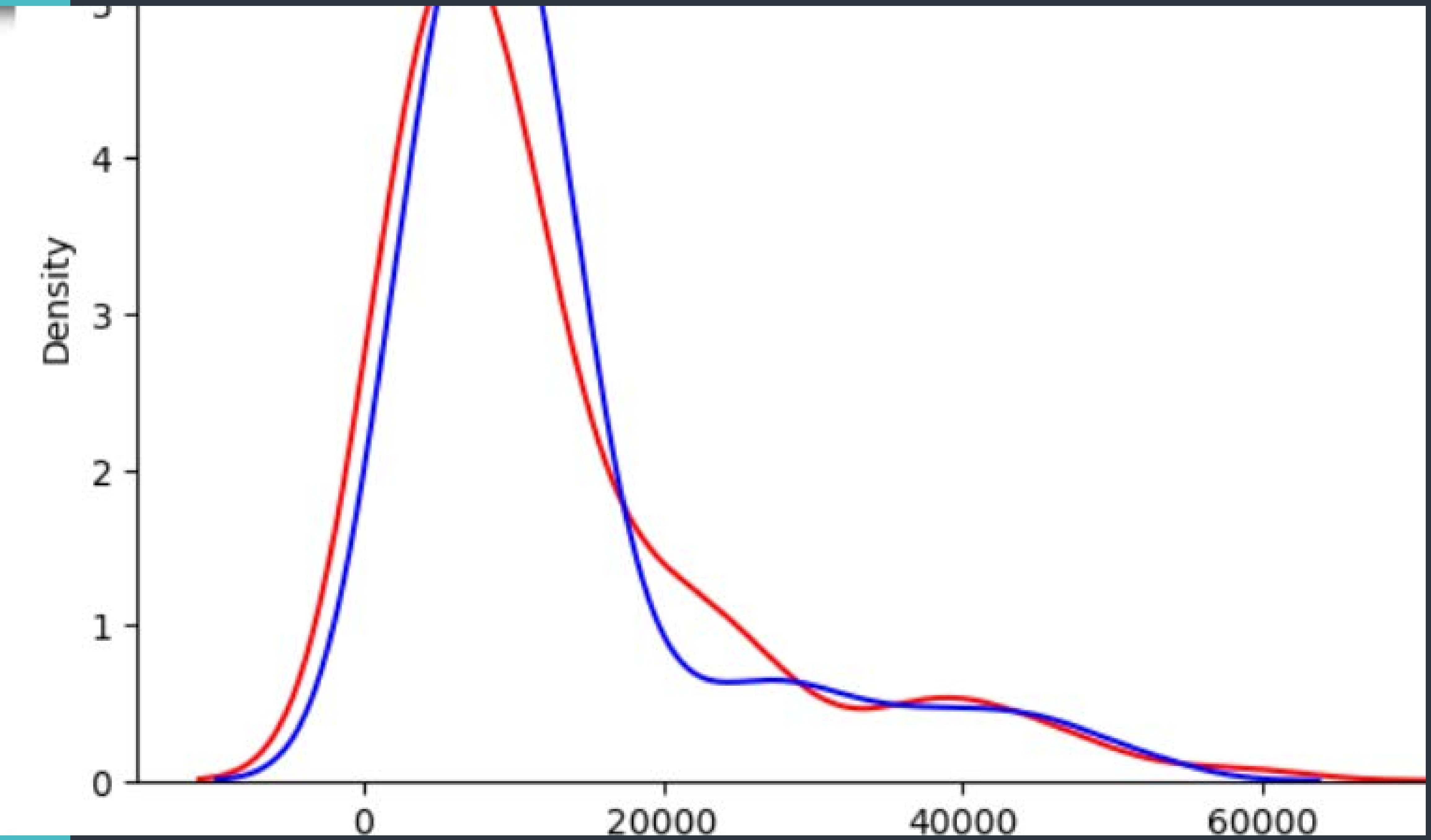
```
MAE: 2988.626627897196
MSE: 24512834.56541676
RMSE: 4951.043785447344
R2 Score: 0.8221477010678055
```



LINEAR REGRESSION

```
print('MAE:', mean_absolute_error(y_test, y_pred))
print('MSE:', mean_squared_error(y_test, y_pred))
print('RMSE:', np.sqrt(mean_squared_error(y_test, y_pred)))
print('R2 Score:', r2_score(y_test, y_pred))
```

```
MAE: 2988.626627897196
MSE: 24512834.56541676
RMSE: 4951.043785447344
R2 Score: 0.8221477010678055
```

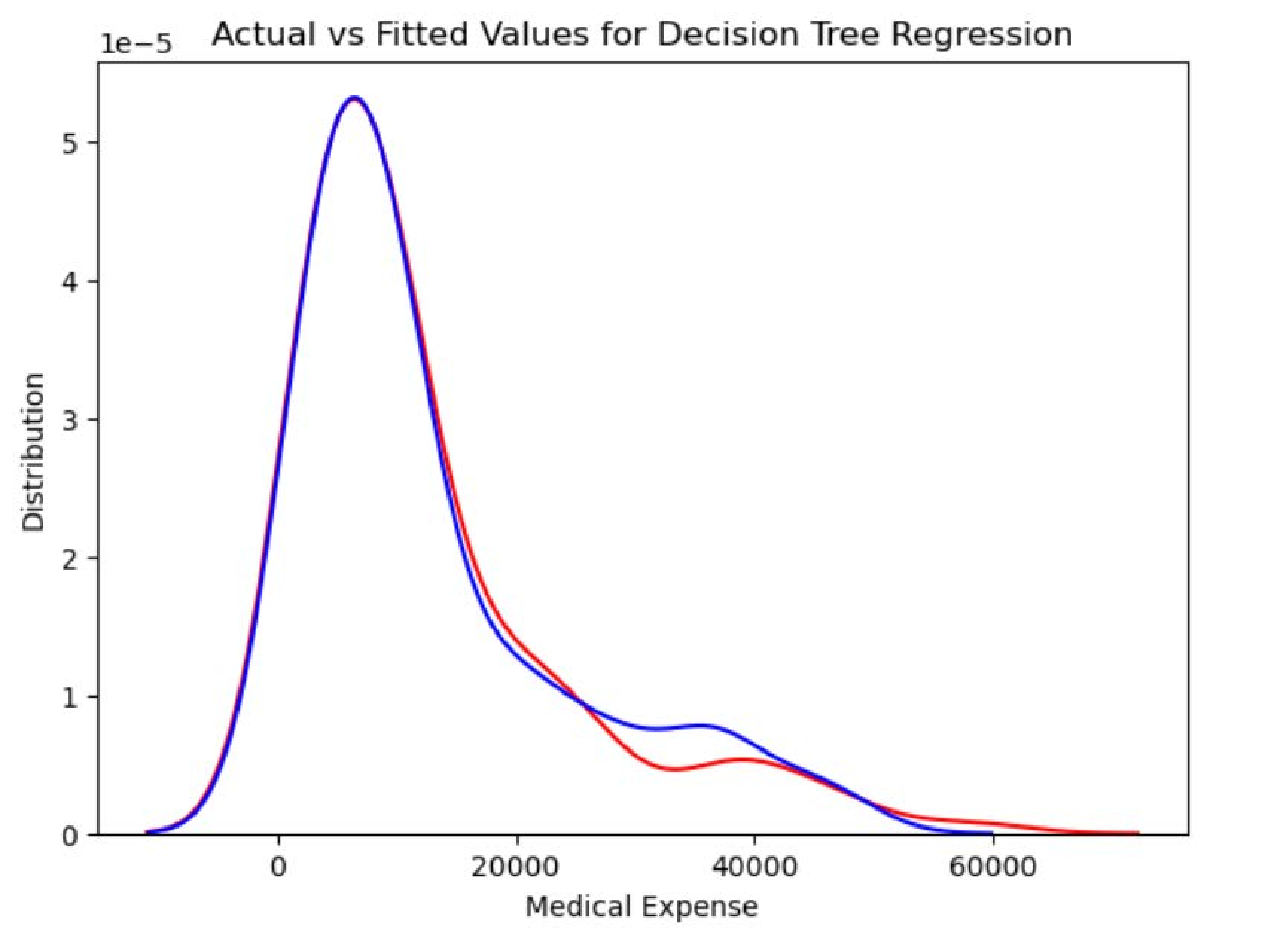


POLYNOMIAL REGRESSION

```
print('MAE:', mean_absolute_error(y_test, dtree_pred))
print('MSE:', mean_squared_error(y_test, dtree_pred))
print('RMSE:', np.sqrt(mean_squared_error(y_test, dtree_pred)))
print('Accuracy:', dtree.score(x_test,y_test))
print('R2 Score:', r2_score(y_test, dtree_pred))
```

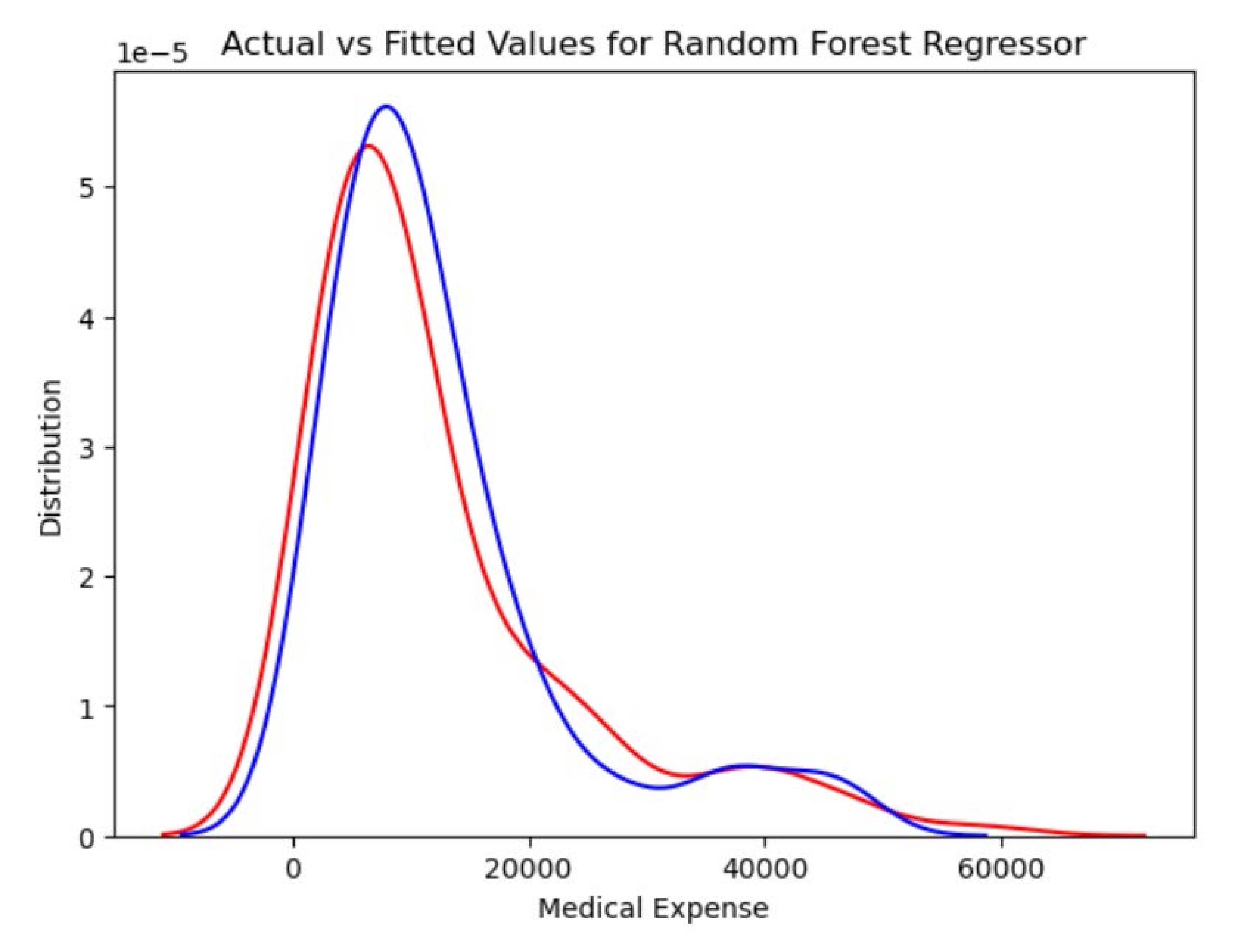
```
MAE: 3361.123098971962
MSE: 51166805.10356602
RMSE: 7153.0975880080105
Accuracy: 0.628760440070712
R2 Score: 0.628760440070712
```

DECISION TREE



```
print('MAE:', mean_absolute_error(y_test, rf_pred))
print('MSE:', mean_squared_error(y_test, rf_pred))
print('RMSE:', np.sqrt(mean_squared_error(y_test, rf_pred)))
print('Accuracy:', rf.score(x_test,y_test))
```

```
MAE: 2854.2702910822827
MSE: 26972820.815520305
RMSE: 5193.536445960527
Accuracy: 0.8042993282590665
```



RANDOM FOREST REGRESSOR

CONCLUSION

- From the above models, we can see that Decision Tree Regressor and Random Forest Regressor are giving the best results. However, Random Forest Regressor gives the best results with the least RMSE value. Therefore, I will use a Random Forest Regressor to predict the medical expenses of patients.

- Moreover, the medical expense of smokers is higher than that of non-smokers. The medical expense of patients with a BMI greater than 30 is higher than that of patients with a BMI less than 30. The medical expenses of older patients are higher than that of younger patients.

- Thus, from the overall analysis, we can conclude that the medical expense of patients depends on their age, BMI, and smoking habits.

THANK YOU