Insurance Data Analysis

Objective:

To analyze the dataset that will help to create a model that will predict the cost of medical insurance based on various input features

1. Import libraries such as Pandas, matplotlib, NumPy, and seaborn and load the insurance dataset.

In [1]:

```python
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [2]:

```python
df = pd.read_csv('insurance.csv')
```

In [3]:

```python
df.head()
```

Out[3]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Observation:

- Successfully imported the required libraries and loaded the insurance dataset into a DataFrame named df.

2. Check the shape of the data along with the data types of the column

In [4]:

```
shape=df.shape
print('shape of data',shape)
```

```
shape of data (1338, 7)
```

In [5]:

```
data_type=df.dtypes
print(data_type)
```

```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

Observation:

- The dataset has a specific number of rows and columns (1338, 7).
- The columns age, bmi, children, and charges are numerical.
- The columns sex, smoker, and region are categorical.

3. Check missing values in the dataset and find the appropriate measures to fill in the missing values

In [6]:

```
df.isnull().sum()
```

Out[6]:

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Observation:

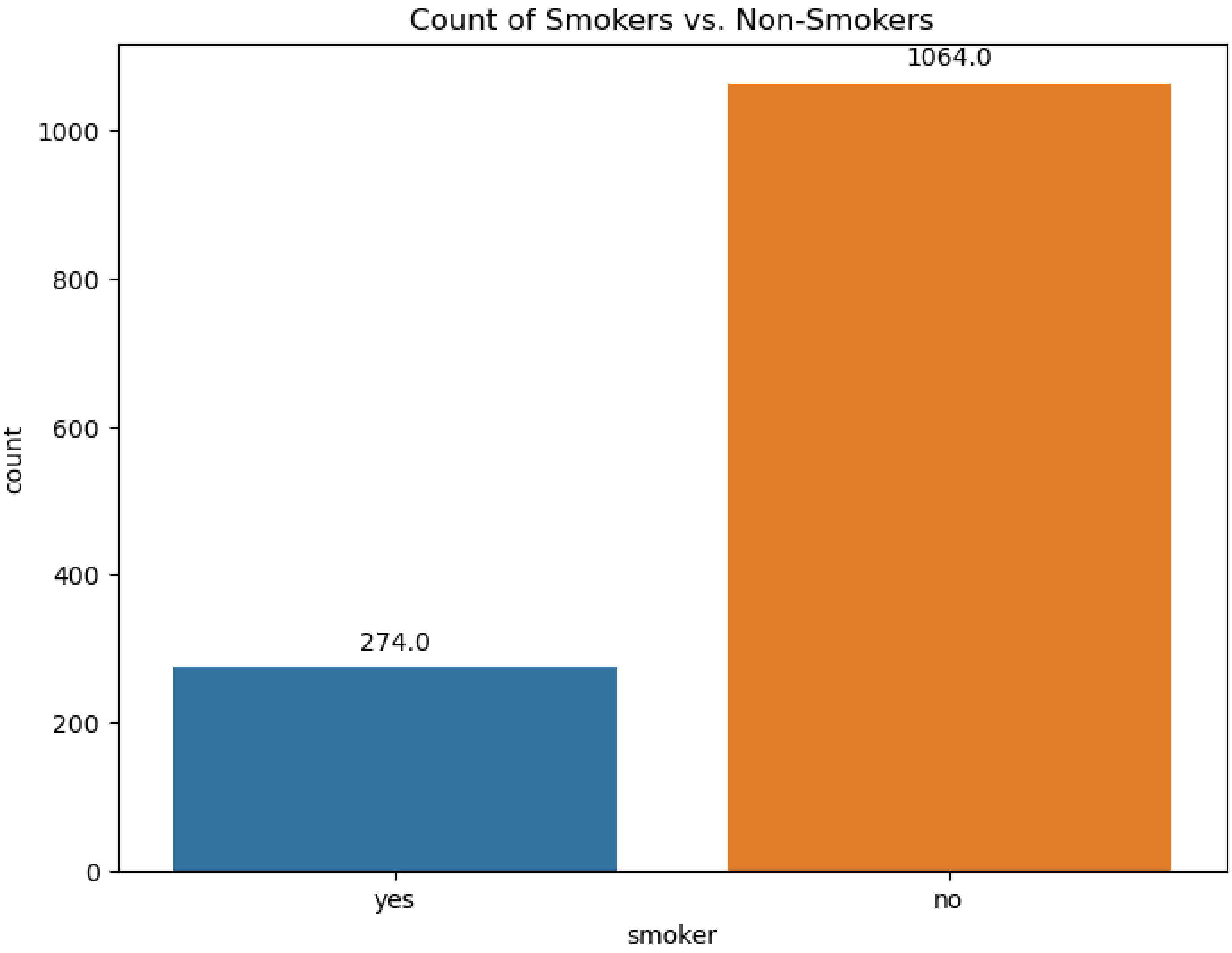
- There are no missing values in any of the columns of the dataset.

4. Explore the relationship between the feature and target column using a count plot of categorical columns and a scatter plot of numerical columns

Count Plot of Smokers vs. Non-Smokers

In [7]:

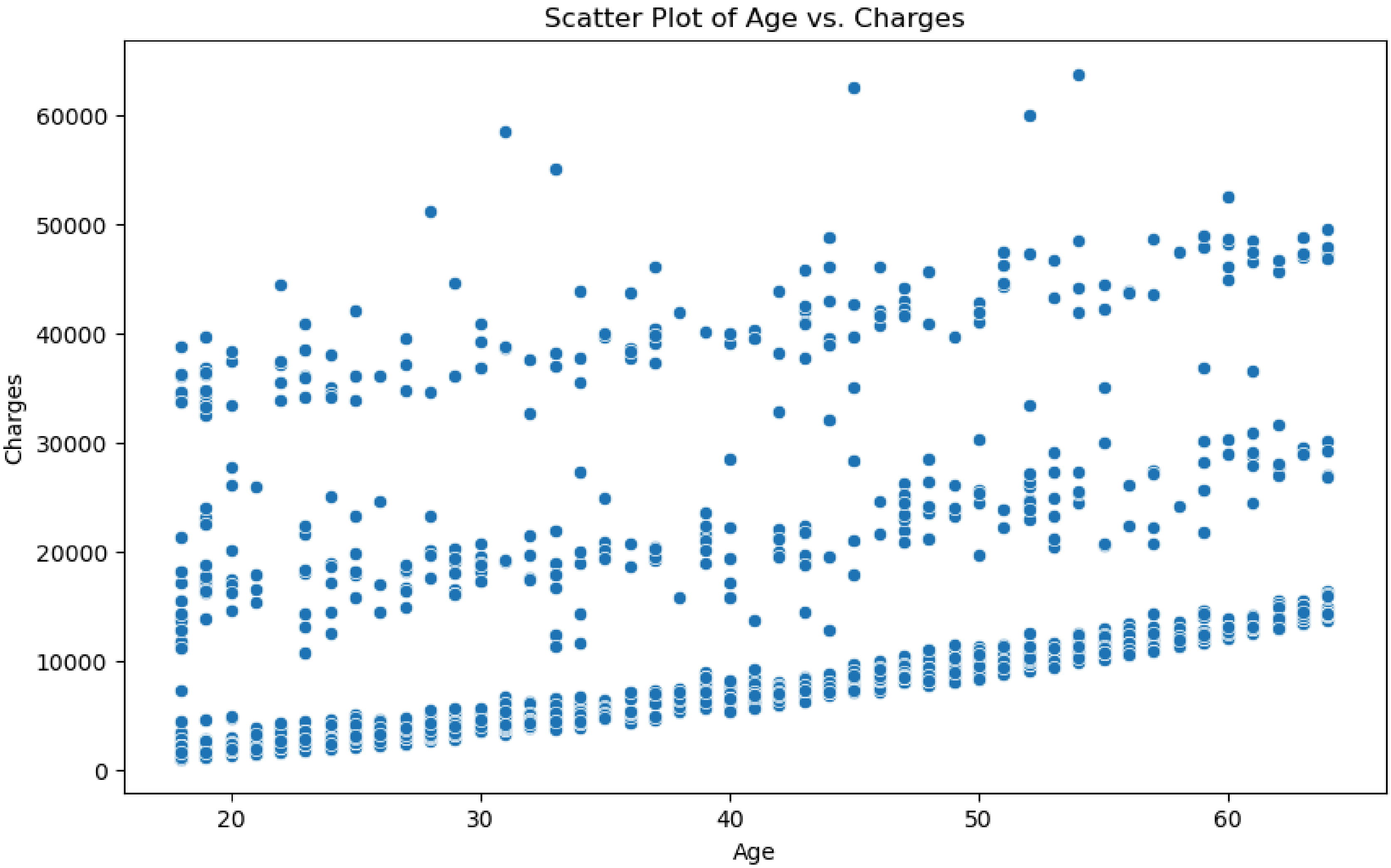
```
plt.figure(figsize=(8, 6))
sns.countplot(x='smoker', data=df)
plt.title('Count of Smokers vs. Non-Smokers')

for p in plt.gca().patches:
    plt.gca().annotate(f'{p.get_height()}', (p.get_x() + p.get_width() / 2., p.get_height()),
                       ha='center', va='center', xytext=(0, 10), textcoords='offset points')
plt.show()
```



Scatter Plot of Age vs. Charges

In [8]:

```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='age', y='charges', data=df)
plt.title('Scatter Plot of Age vs. Charges')
plt.xlabel('Age')
plt.ylabel('Charges')
plt.show()
```



Observation:

- The count plot shows the distribution of smokers and non-smokers.
- The scatter plot shows the relationship between age and charges, indicating that charges tend to increase with age.

5. Perform data visualization using plots of feature vs feature

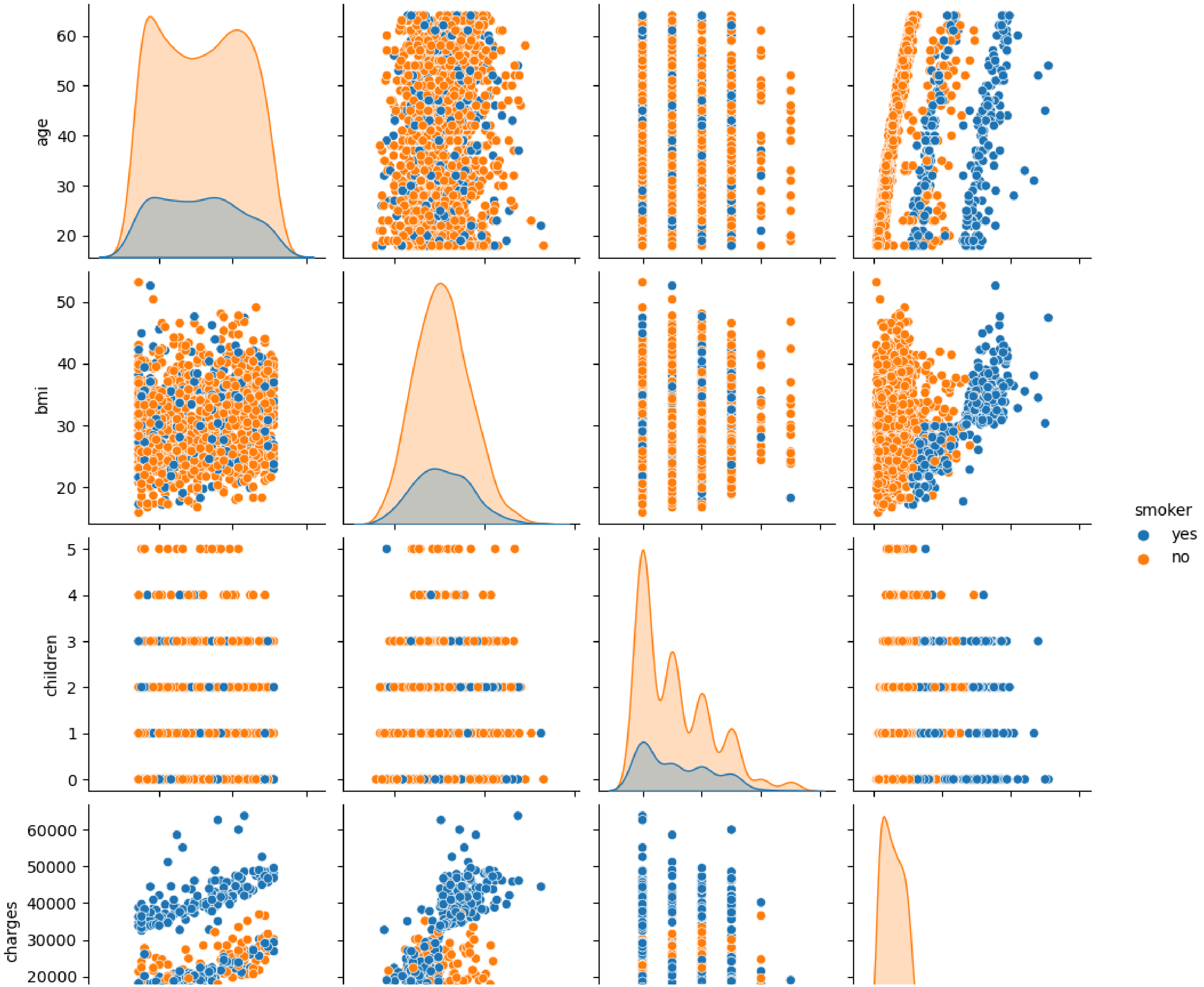
Pair Plot

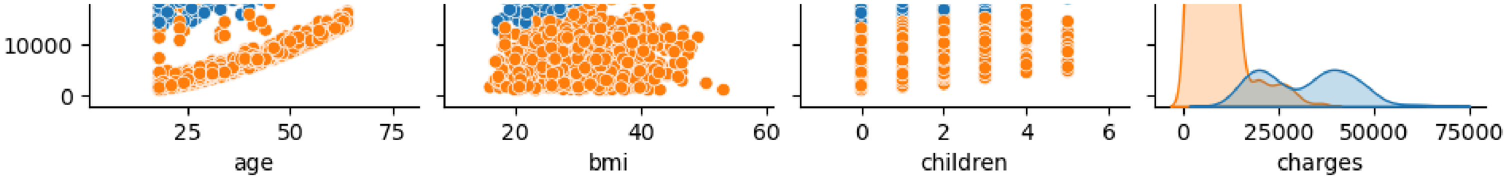
In [9]:

```
sns.pairplot(df, hue='smoker')
plt.suptitle('Pair Plot of All Numerical Features Colored by Smoker', y=1.02)
plt.show()
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecate
d and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecate
d and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecate
d and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecate
d and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

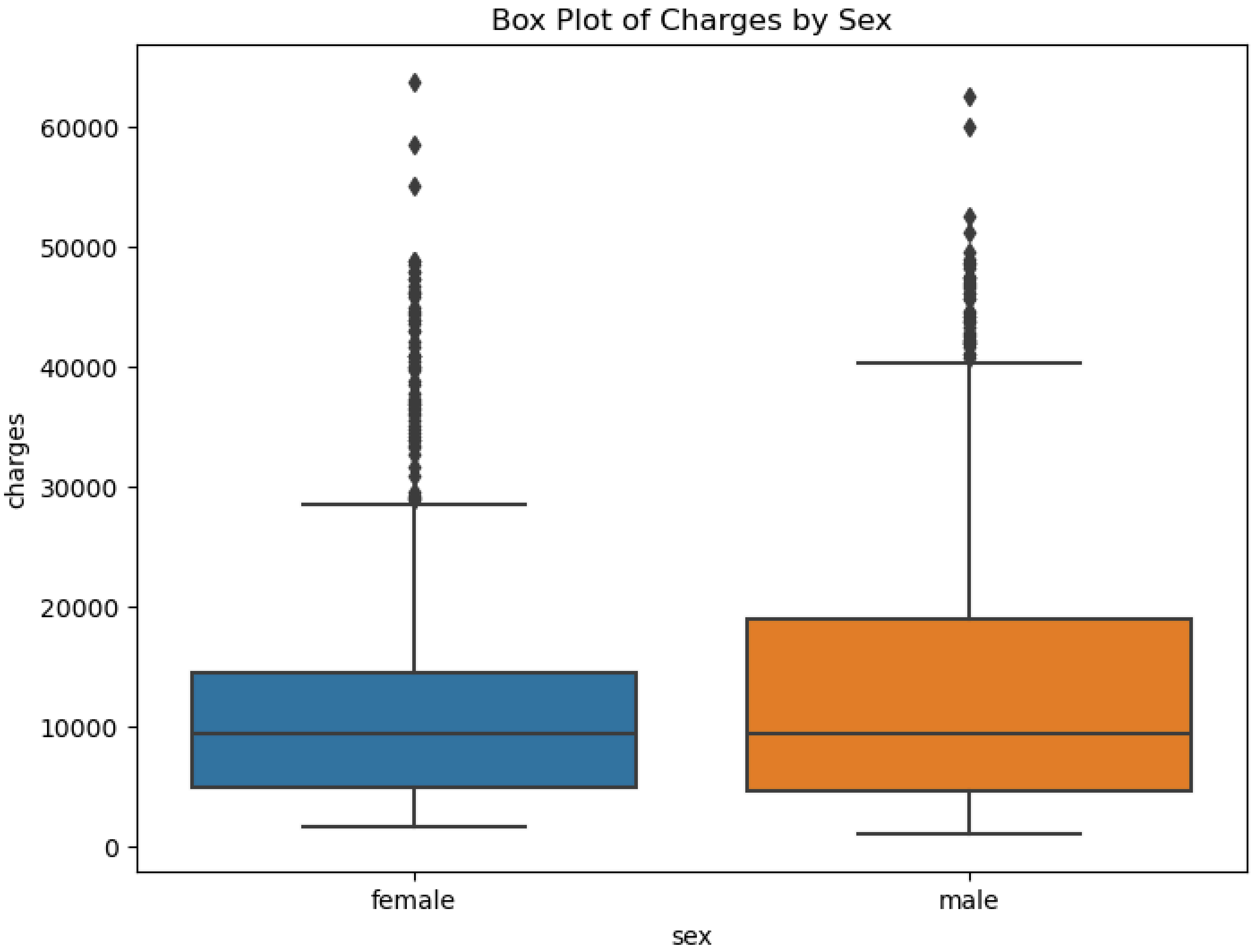
Pair Plot of All Numerical Features Colored by Smoker





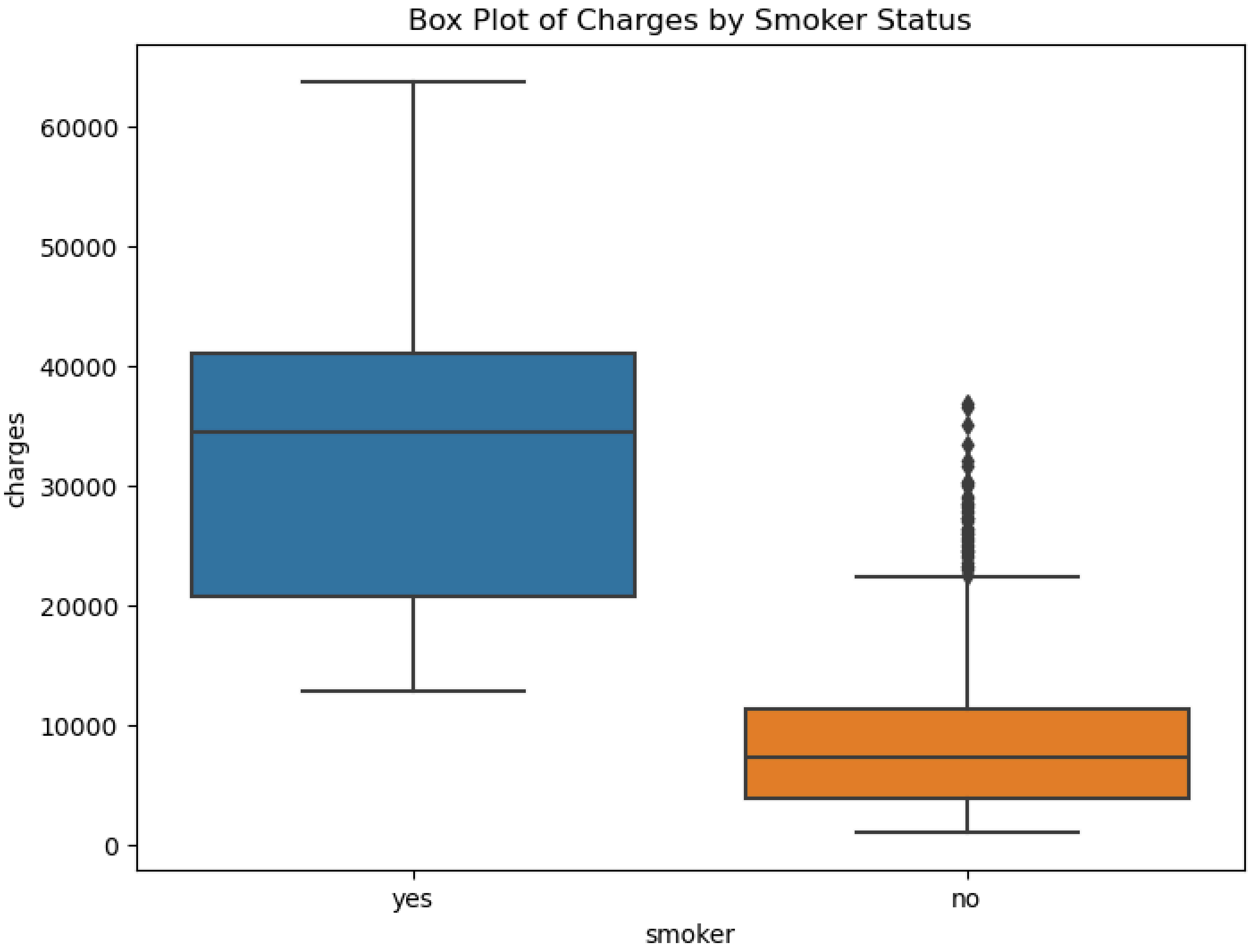
Box Plot of Charges by Sex

In [10]:
```
plt.figure(figsize=(8, 6))
sns.boxplot(x='sex', y='charges', data=df)
plt.title('Box Plot of Charges by Sex')
plt.show()
```



Box Plot of Charges by Smoker Status

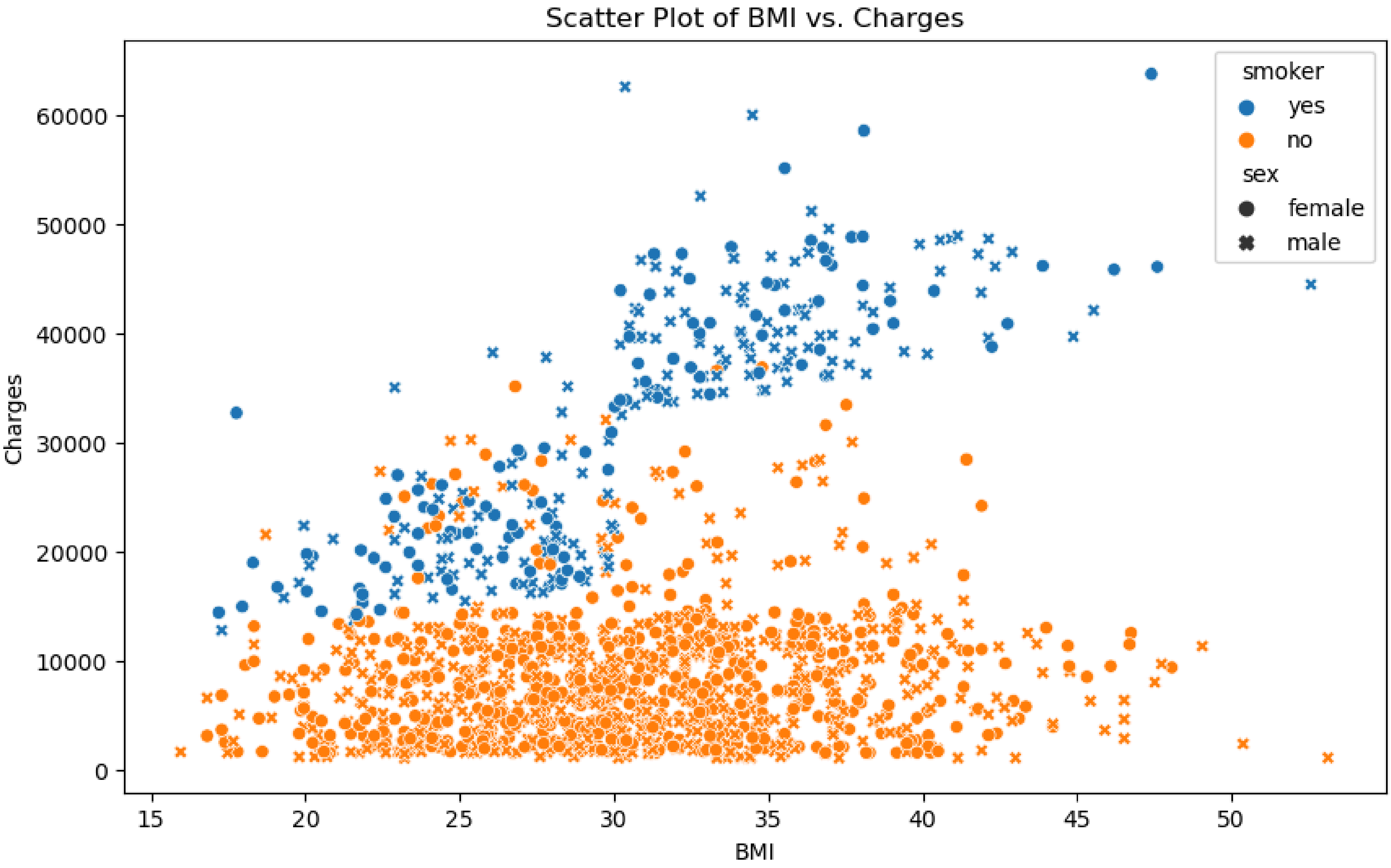
In [11]:
```
plt.figure(figsize=(8, 6))
sns.boxplot(x='smoker', y='charges', data=df)
plt.title('Box Plot of Charges by Smoker Status')
plt.show()
```



Scatter Plot of BMI vs. Charges

In [12]:

```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='bmi', y='charges', hue='smoker', style='sex', data=df)
plt.title('Scatter Plot of BMI vs. Charges')
plt.xlabel('BMI')
plt.ylabel('Charges')
plt.show()
```



Observation:

- Pair plots show relationships between all numerical features, colored by smoker status.
- Box plots display the distribution of charges across different categories (sex, smoker, region).
- Scatter plots show relationships between age/BMI and charges, with markers colored by smoker status and styled by sex.

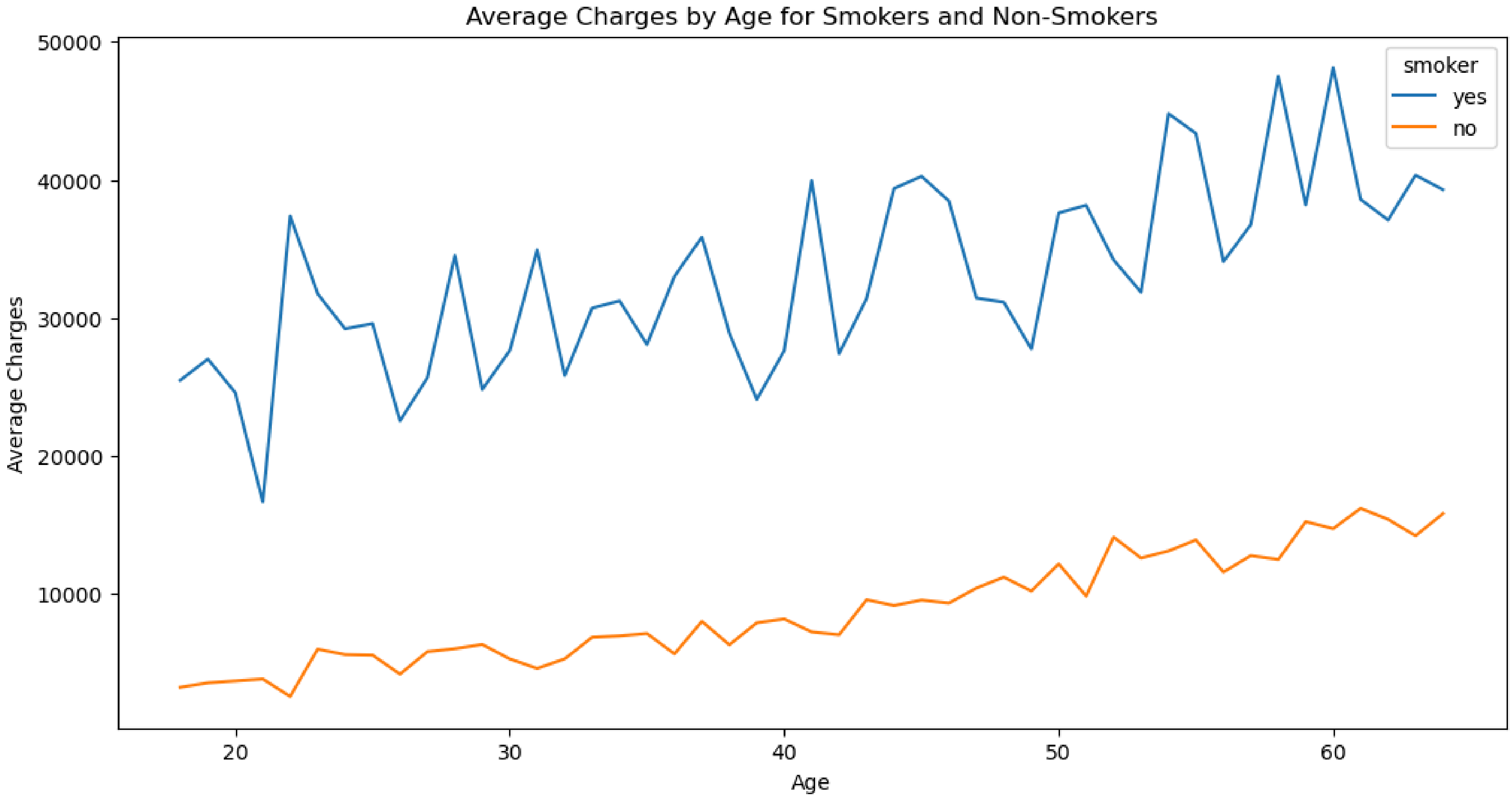
6. Check if the number of premium charges for smokers or non-smokers is increasing as they are aging

Line Plot: Average charges by age for smokers and non-smokers

In [13]:

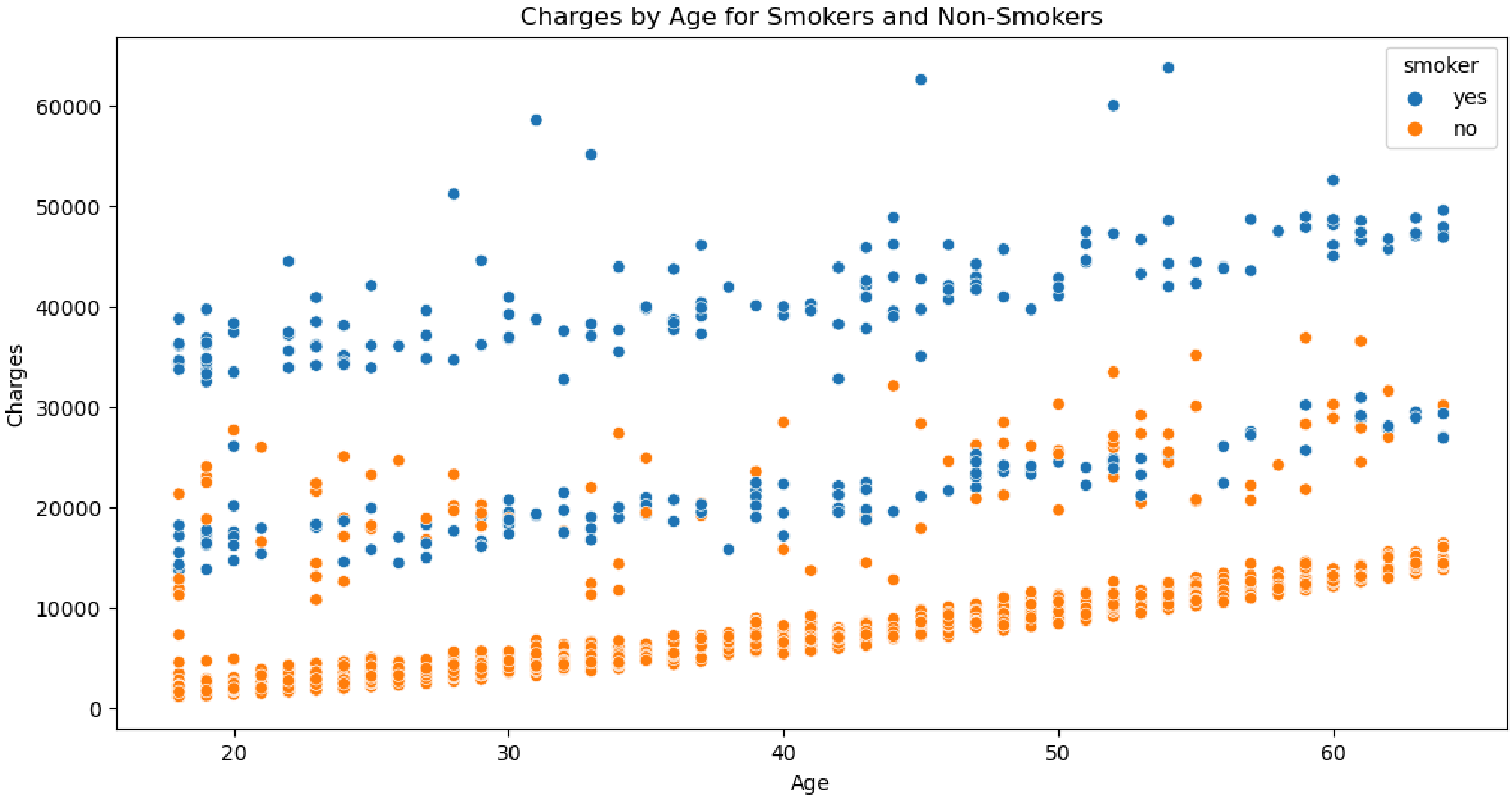
```
plt.figure(figsize=(12, 6))
sns.lineplot(x='age', y='charges', hue='smoker', data=df, ci=None)
plt.title('Average Charges by Age for Smokers and Non-Smokers')
plt.xlabel('Age')
plt.ylabel('Average Charges')
plt.show()
```

```
C:\Users\vinay\AppData\Local\Temp\ipykernel_20580\2803374483.py:2: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.

  sns.lineplot(x='age', y='charges', hue='smoker', data=df, ci=None)
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecate
d and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecate
d and will be removed in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



Scatter Plot: Charges by age for smokers and non-smokers

In [14]:
```
plt.figure(figsize=(12, 6))
sns.scatterplot(x='age', y='charges', hue='smoker', data=df)
plt.title('Charges by Age for Smokers and Non-Smokers')
plt.xlabel('Age')
plt.ylabel('Charges')
plt.show()
```



Observation:

- The line plot shows that average charges tend to increase with age for both smokers and non-smokers, but charges for smokers are consistently higher.
- The scatter plot illustrates that individual charges generally increase with age, with smokers having higher charges compared to non-smokers.

Contact Information

For any queries or further information, please feel free to reach out to me through the following platforms:

- **LinkedIn**: Vinay Kumar Panika
- **GitHub**: Vinaypanika