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

STEP-1 -- Importing the libraries and loading the dataset:

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

# Load the dataset
df = pd.read_csv('insurance.csv')

# Display the first few rows of the dataset
print(df.head())
```

```
   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
```

STEP-2 -- Checking the shape of the dataset and the data types of each column:

```
# Check the shape of the dataset
shape = df.shape
print("Shape of the dataset:", shape)

# Check the data types of the columns
data_types = df.dtypes
print("Data types of the columns:\n", data_types)
```

```
Shape of the dataset: (1338, 7)
Data types of the columns:
 age           int64
sex           object
bmi          float64
children       int64
smoker        object
region        object
charges      float64
dtype: object
```

Observation: The shape indicates the number of rows and columns, while data types help identify how to handle each column during analysis. The dataset contains several columns, including both categorical and numerical types.

STEP-3 -- Checking for Missing Values:

```
# Check for missing values
missing_values = df.isnull().sum()
print("Missing values in each column:\n", missing_values)
```

```
Missing values in each column:
 age         0
sex          0
bmi          0
children     0
smoker       0
region       0
charges      0
dtype: int64
```

Observation: There are no missing values.

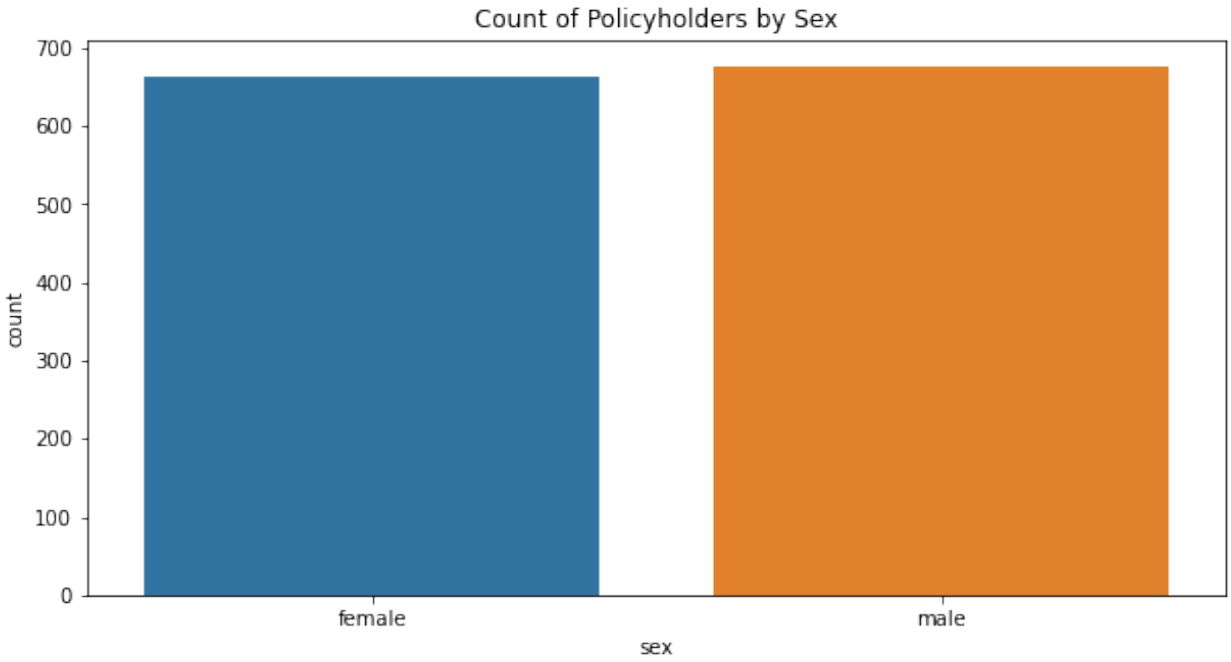
STEP-4 -- Exploring Relationships:

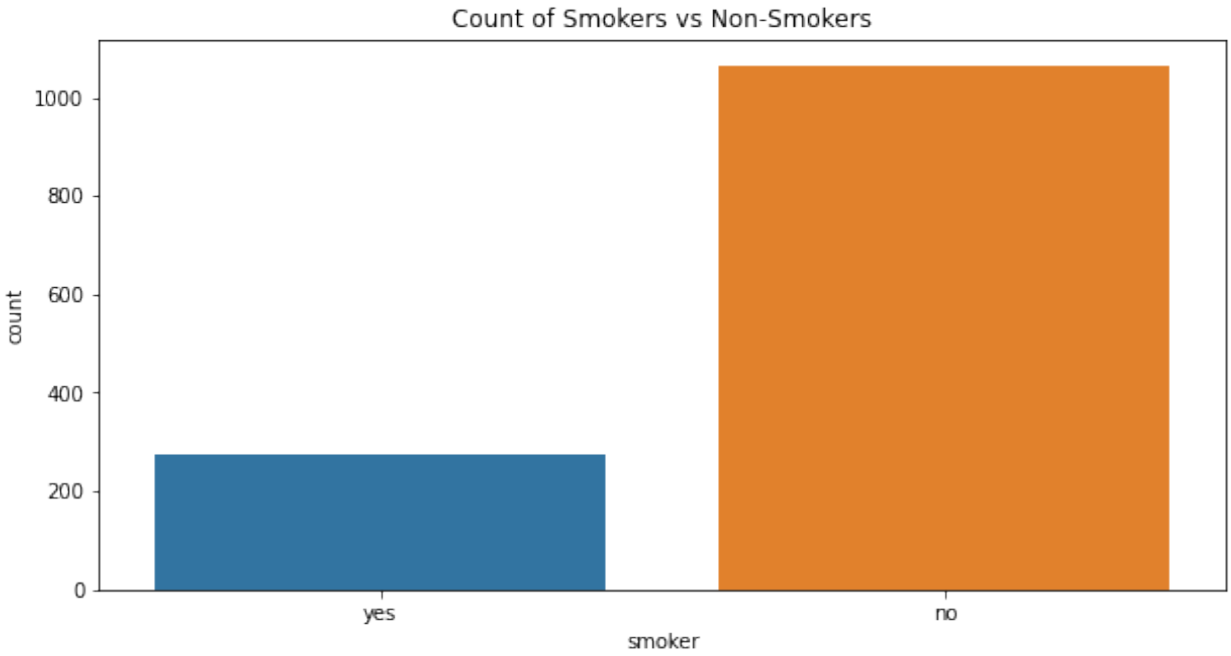
Exploring the relationship between features and the target column using visualizations:

Count Plot for Categorical columns:

```
# Count plot for categorical variables
plt.figure(figsize=(10, 5))
sns.countplot(x='sex', data=df)
plt.title('Count of Policyholders by Sex')
plt.show()

plt.figure(figsize=(10, 5))
sns.countplot(x='smoker', data=df)
plt.title('Count of Smokers vs Non-Smokers')
plt.show()
```



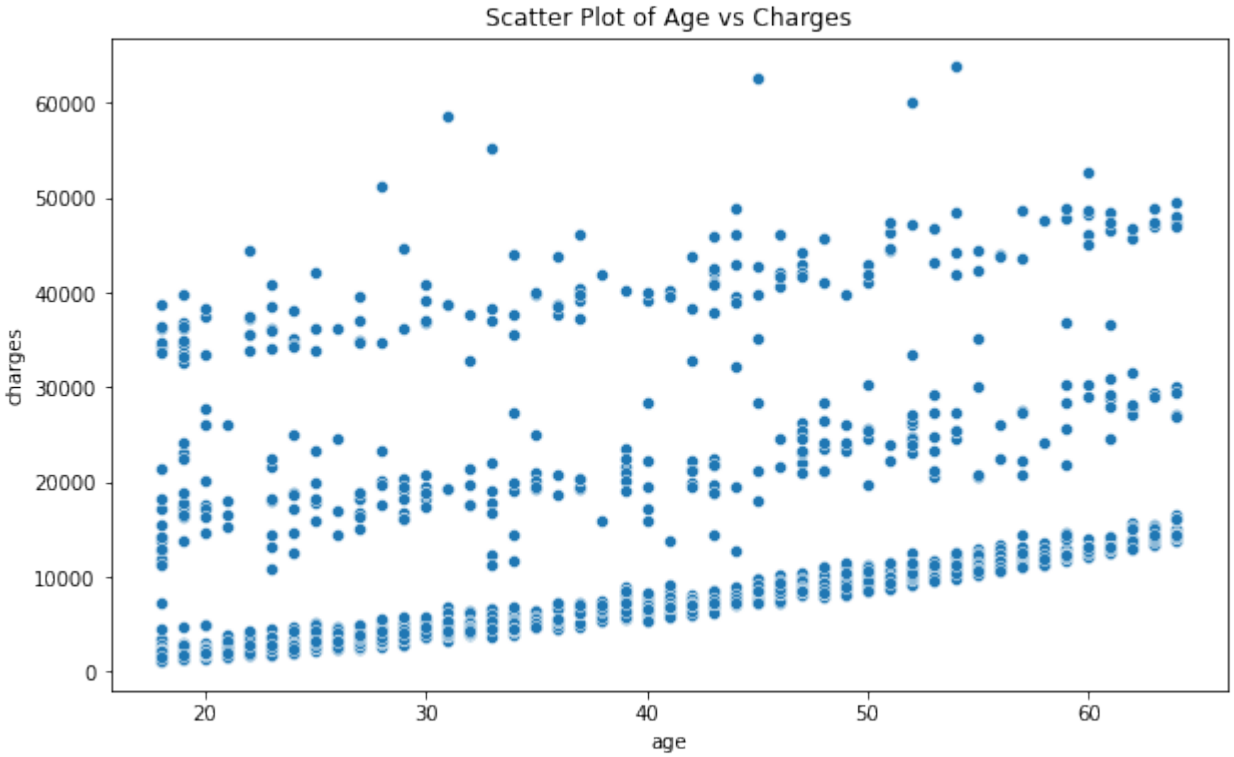


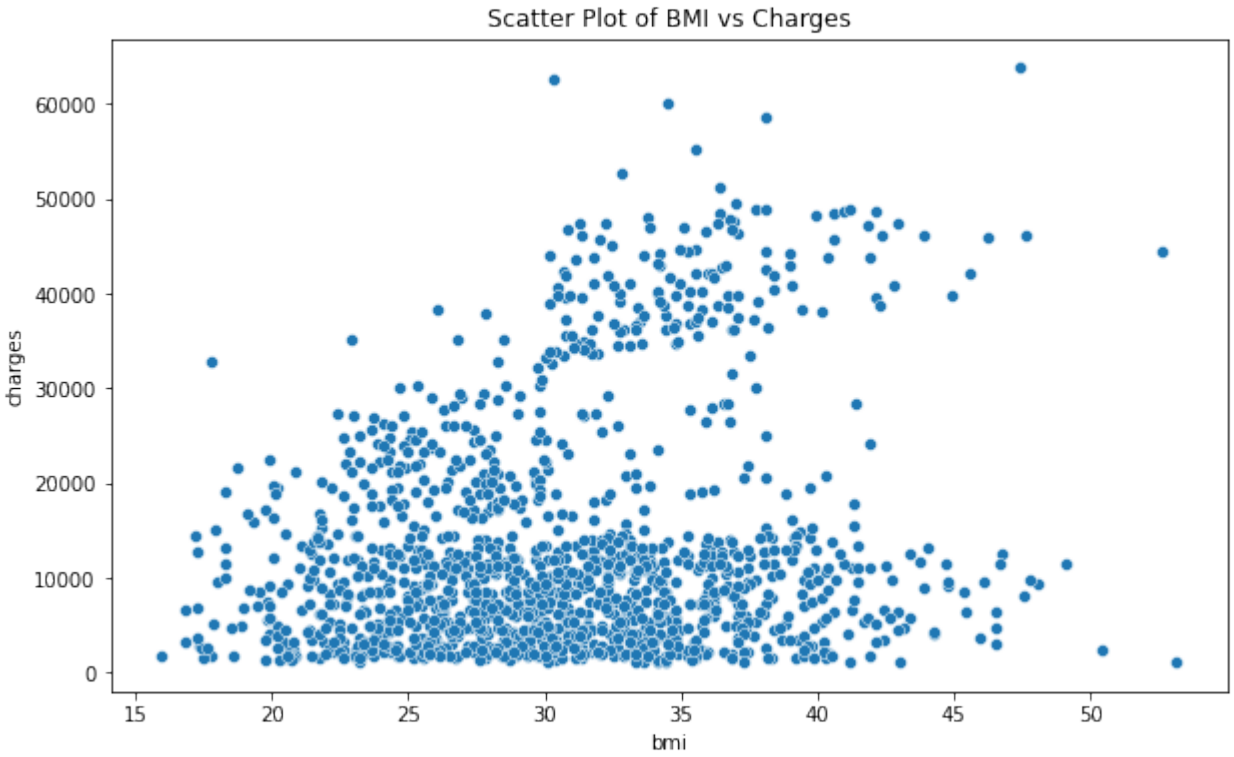
Scatter plot for Numerical columns:

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

plt.figure(figsize=(10, 6))
```

```
sns.scatterplot(x='bmi', y='charges', data=df)
plt.title('Scatter Plot of BMI vs Charges')
plt.show()
```



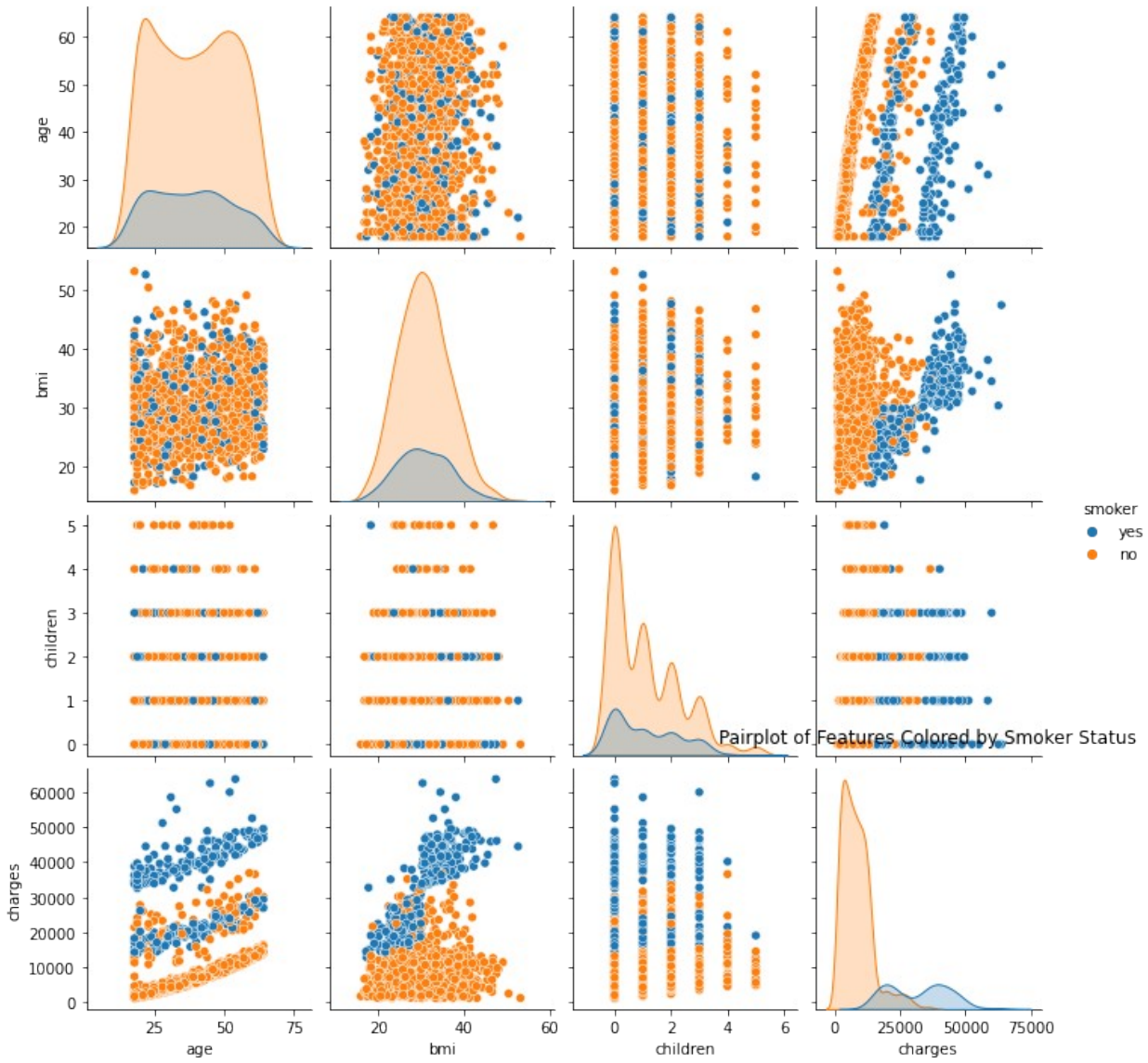


Observations: According to Count Plots: The count of policyholders is relatively balanced between genders, with a higher number of non-smokers compared to smokers.

According to Scatter Plots: The scatter plot of age vs. charges shows a positive correlation, suggesting that as age increases, the insurance charges tend to increase. The BMI vs. charges plot also indicates a similar trend.

STEP-5 -- Performing data visualization using plots of feature vs feature relationships:

```
# Pairplot to visualize the relationships between features
sns.pairplot(df, hue='smoker')
plt.title('Pairplot of Features Colored by Smoker Status')
plt.show()
```

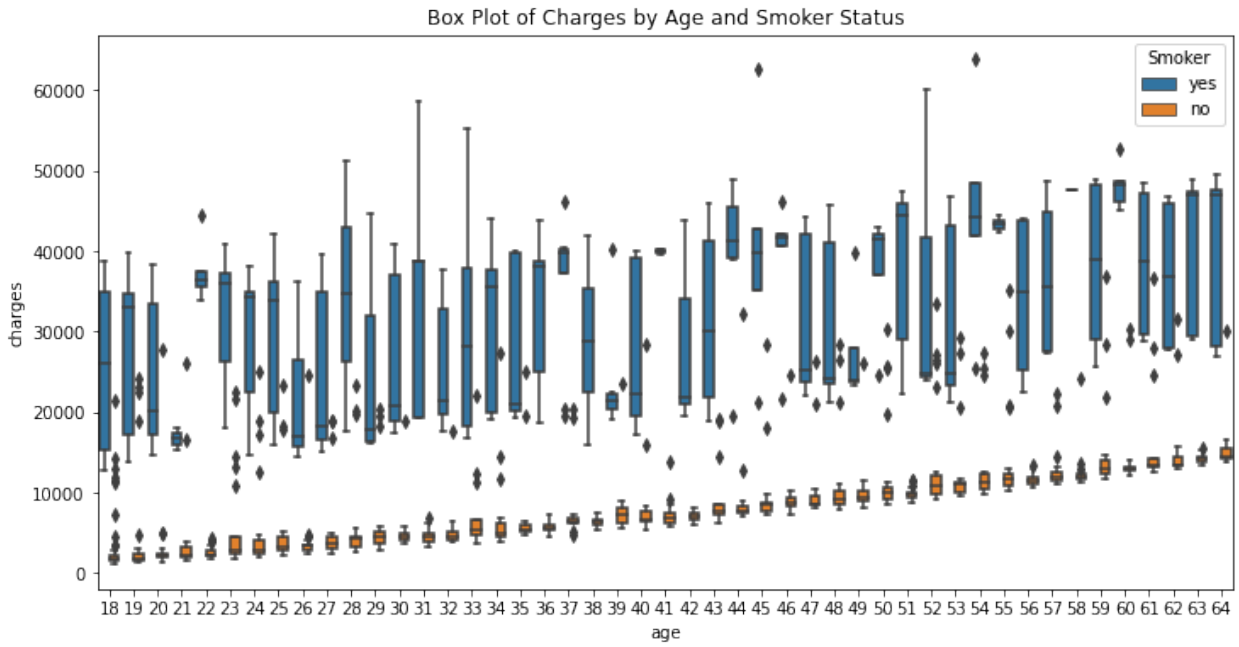


Observations: According to the Pairplot: The pairplot colored by smoker status shows that smokers tend to have higher insurance charges compared to non-smokers across various ages and BMI values.

STEP-6 -- Checking if the number of premium charges for smokers or non-smokers is increasing as they are aging:

```
# Box plot to visualize the distribution of charges by age and smoker status
plt.figure(figsize=(12, 6))
sns.boxplot(x='age', y='charges', hue='smoker', data=df)
plt.title('Box Plot of Charges by Age and Smoker Status')
```

```
plt.legend(title='Smoker', loc='upper right')
plt.show()
```



Observations: The boxplot illustrates that smokers generally have higher insurance charges than non-smokers, especially as they age. This reinforces the notion that smoking is a significant factor in determining health insurance premiums.

CONCLUSION

The exploratory data analysis reveals critical insights into the factors affecting healthcare premiums, particularly the impact of smoking and age on insurance charges. These insights can inform the development of predictive models to estimate insurance costs based on individual characteristics.