Insurance Data Analysis

Description

An insurance agency, ABC Insurance, has a large dataset containing information about their policyholders and claims. They want to perform exploratory data analysis (EDA) on this dataset to gain insights that can help them make better business decisions and improve their operations.

The agency wants to analyze the different body types and the environment that affect the premium. The disease's effect or the cost of treatment differs depending on the circumstances. For example, a smoker's medical insurance premium may be higher than that of a healthy person, because smokers are more likely to develop chronic diseases. The agency wants to analyze the data to research healthcare premium costs.

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

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

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline


data = pd.read_csv(r'C:\Users\mauur\Desktop\Data Analytics With R\Python Project\1705482784_insurance\insurance.csv')

print(data)
```

```
In [29]: data = pd.read_csv(r'C:\Users\mauur\Desktop\Data Analytics With R\Python Project\1705482784_insurance\insurance.csv')
         print(data)

      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
...   ...     ...     ...       ...    ...        ...          ...
1333   50    male  30.970         3     no  northwest  10600.54830
1334   18  female  31.920         0     no  northeast   2205.98080
1335   18  female  36.850         0     no  southeast   1629.83350
1336   21  female  25.800         0     no  southwest   2007.94500
1337   61  female  29.070         0    yes  northwest  29141.36030

[1338 rows x 7 columns]
```

Observations:

- **The dataset has 1338 rows and 7 columns**

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

data.shape

```
data.shape

(1338, 7)
```

data.dtypes

```
data.dtypes

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

Observation:

- **As we can see age, BMI, children, and charges are numerical columns and sex, smoker, and region are categorical columns.**

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

data.isna().sum()

```
data.isna().sum()

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

Observation:

- **As we can see there are no missing values present in the dataset**

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

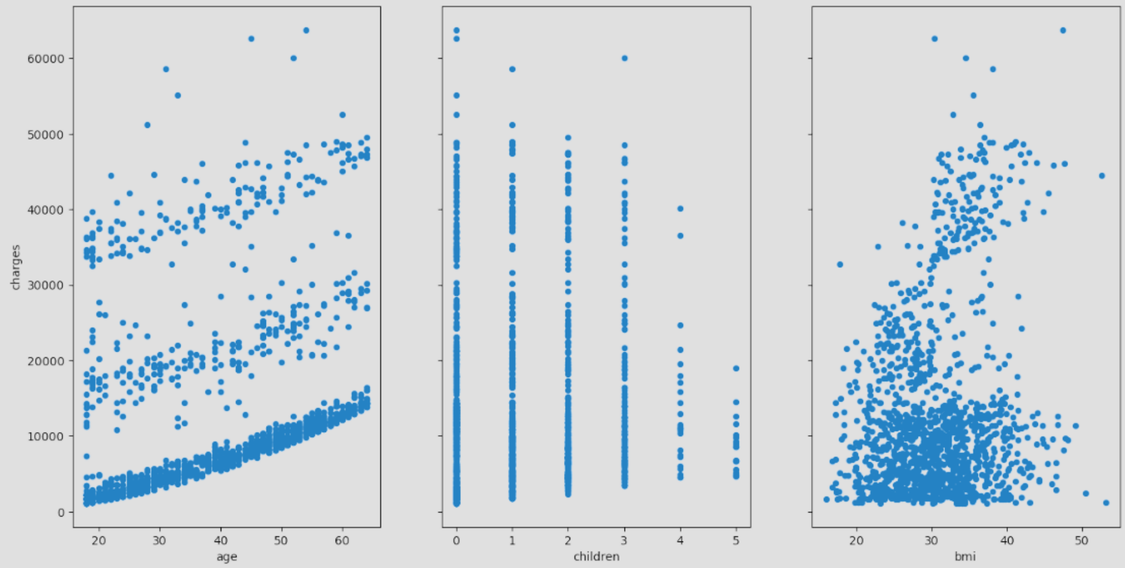
fig, axs = plt.subplots(1, 3, sharey=True)

data.plot(kind='scatter', x='age', y='charges', ax=axs[0], figsize=(16, 8))

data.plot(kind='scatter', x='children', y='charges', ax=axs[1])

data.plot(kind='scatter', x='bmi', y='charges', ax=axs[2])

```
Out[33]: <Axes: xlabel='bmi', ylabel='charges'>
```



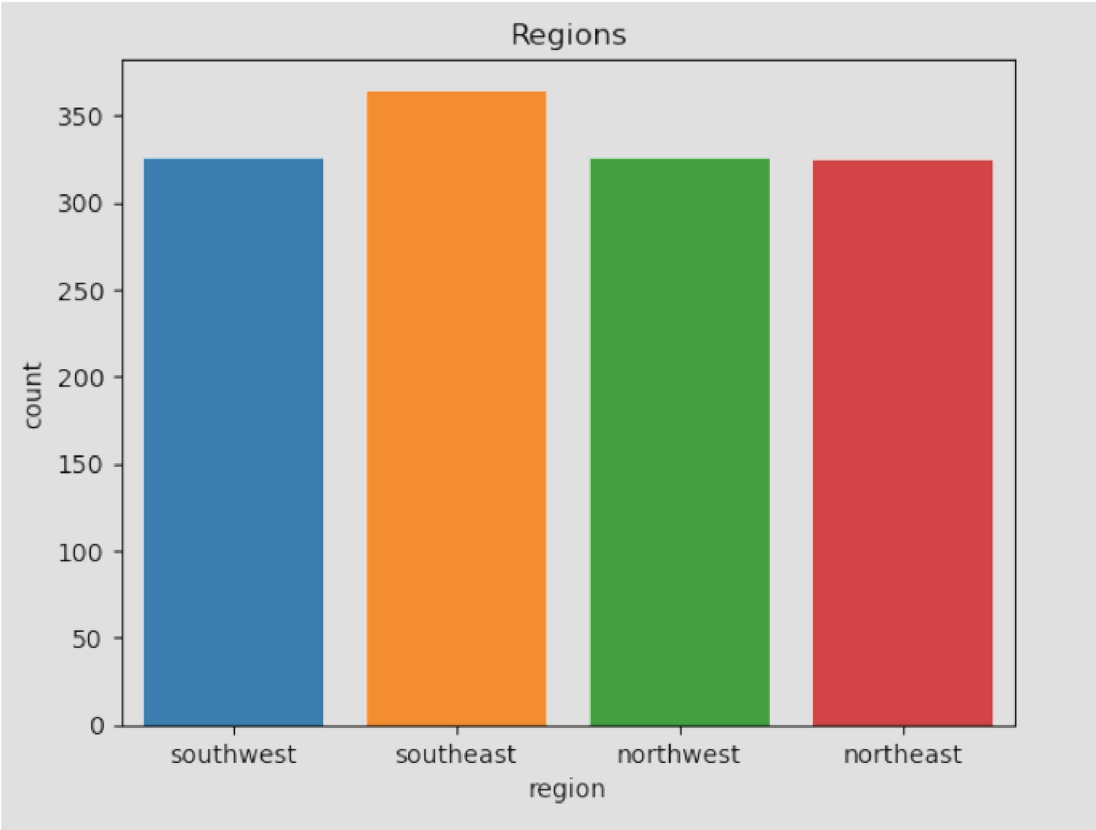
Observation:

- **As we can see in the first graph that if the age is increasing the insurance charges are also increasing.**
- **In the second graph we can see the majority of the customers do not have children.**
- **In the third graph there is no such inference found.**

sns.countplot(data=data, x='region')

plt.title('Regions')

plt.show()

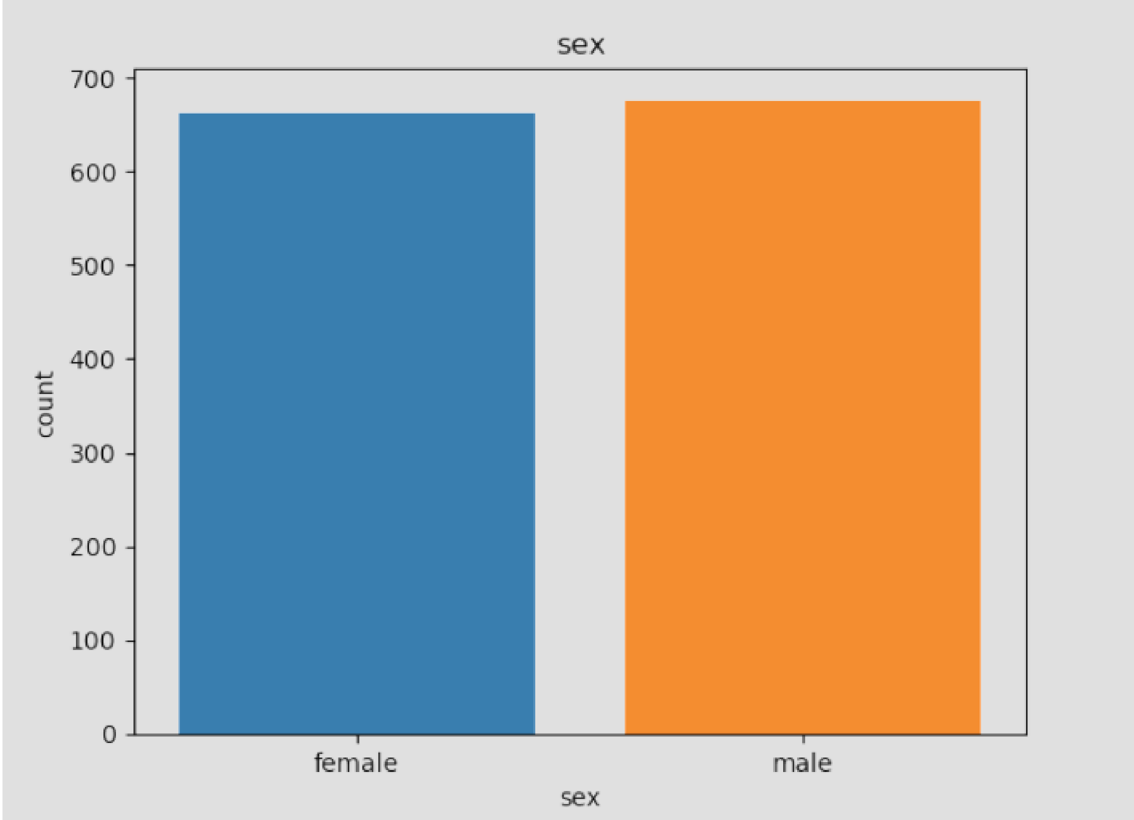


Observation:

- **You can see that the southeast region has the highest count.**
- **Plot a count plot for the sex column.**

```
sns.countplot(data=data, x='sex')

plt.title('sex')
```

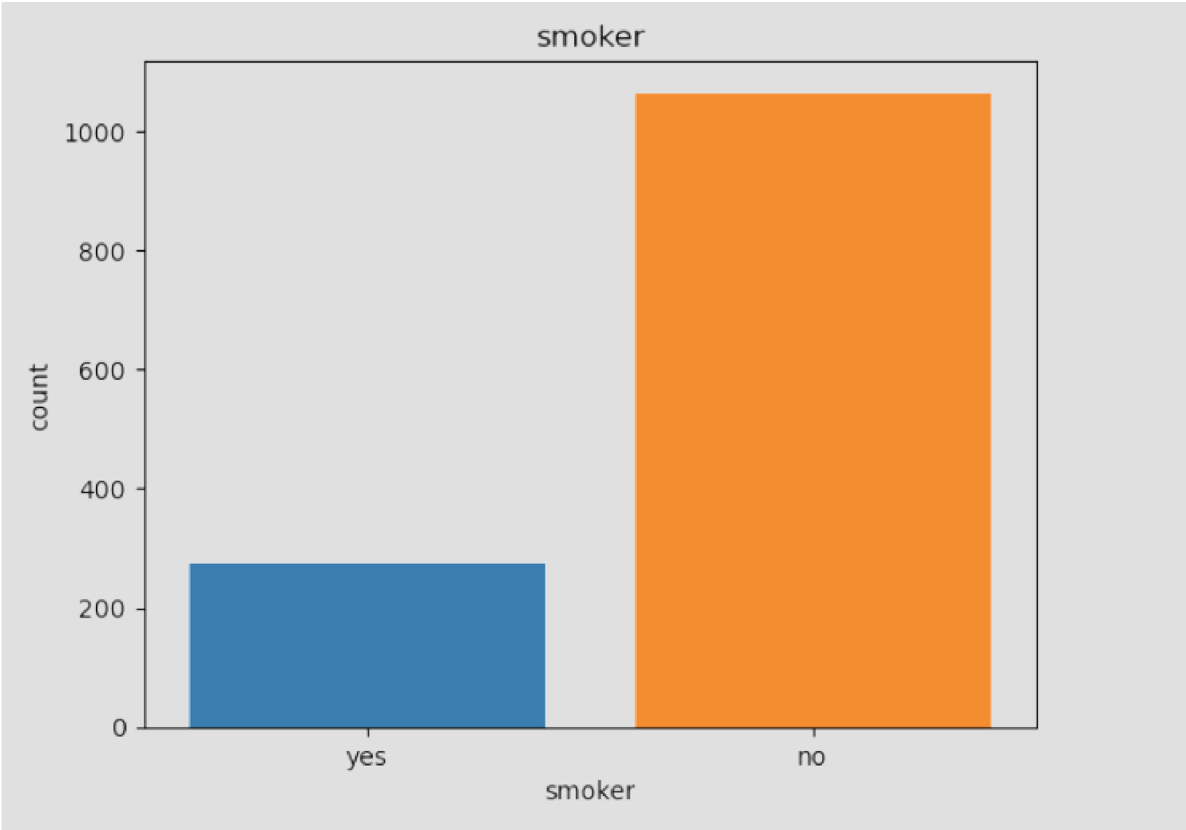
```
Text(0.5, 1.0, 'sex')
```



Observation:

- **The number of males and females is almost equal.**
- **Plot a count plot of the smoker column.**

```
sns.countplot(data=data, x='smoker')

plt.title('smoker')

plt.show()
```



Observation:

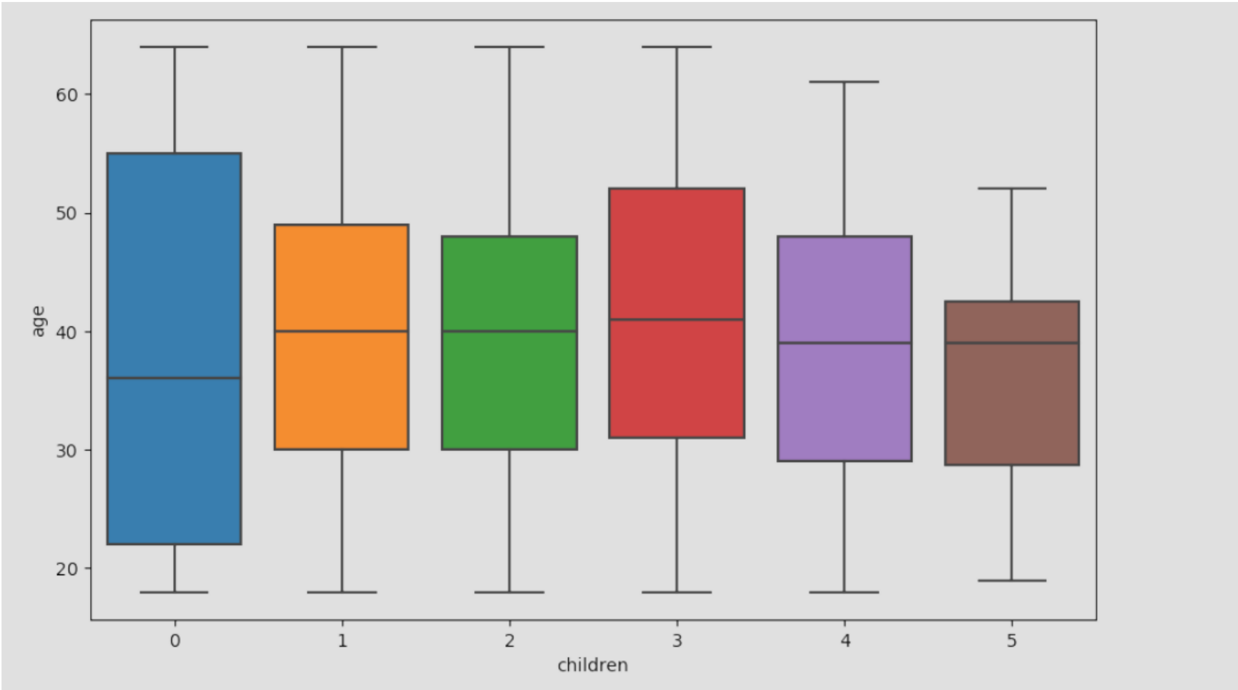
- **Most of the people who have taken insurance are not smokers.**

Perform data visualization using plots of feature vs feature

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

sns.boxplot(x='children',y='age',data=data)

plt.show()



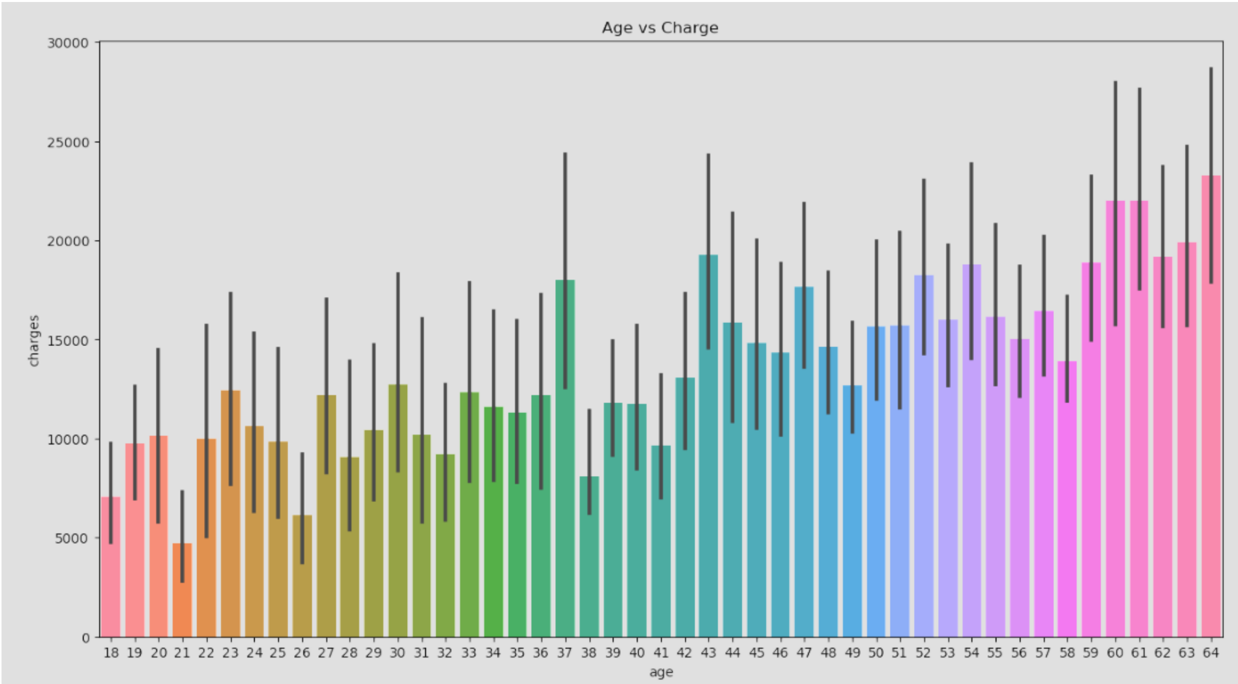
Observation:

- **Now we are confirmed that there are no other outliers in the above-pre-processed column, we can proceed with EDA.**

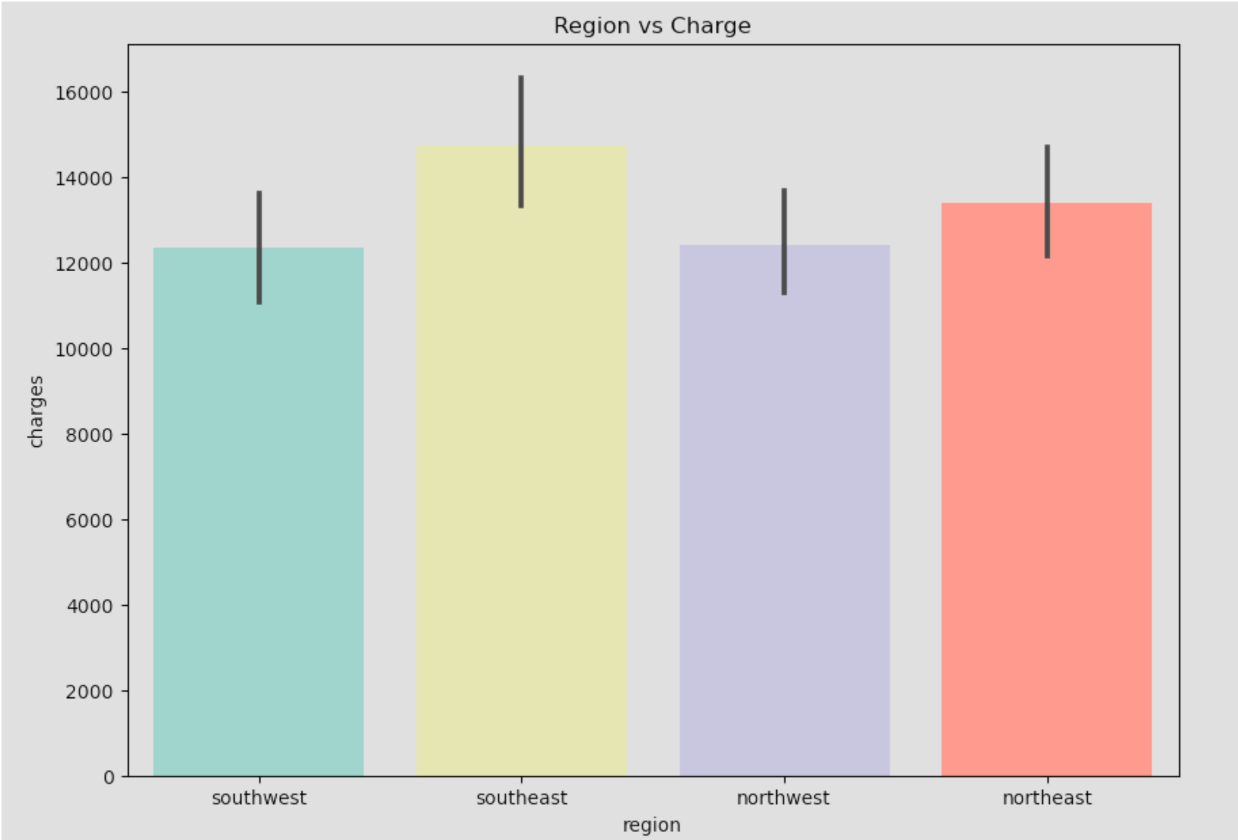
plt.figure(figsize=(15,8))

plt.title('Age vs Charge')

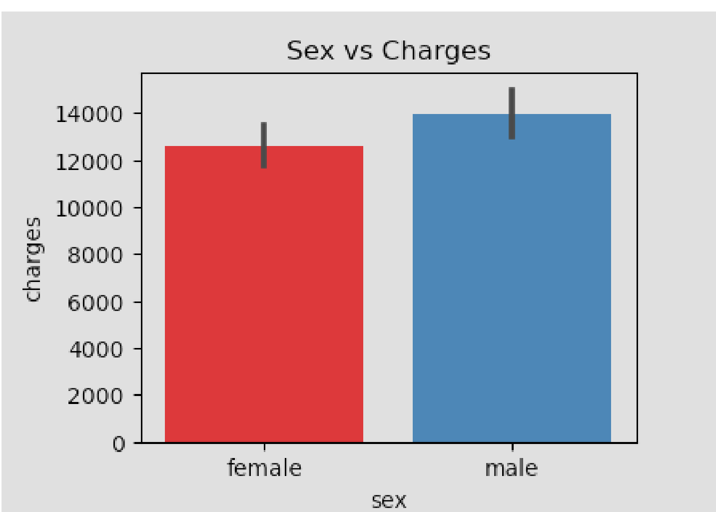
sns.barplot(x='age',y='charges',data=data,palette='husl')



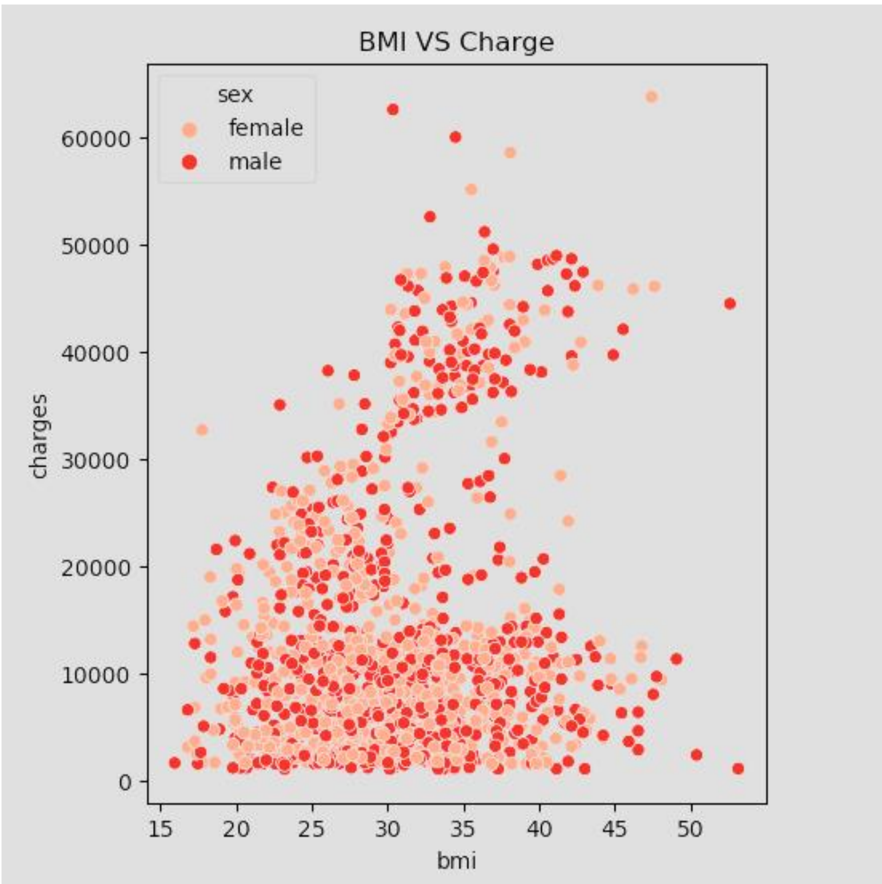

```
plt.figure(figsize=(10,7))
plt.title('Region vs Charge')
```




```
plt.figure(figsize=(4,3))

plt.title('Sex vs Charges')

sns.barplot(x='sex',y='charges',data=data,palette='Set1')
```

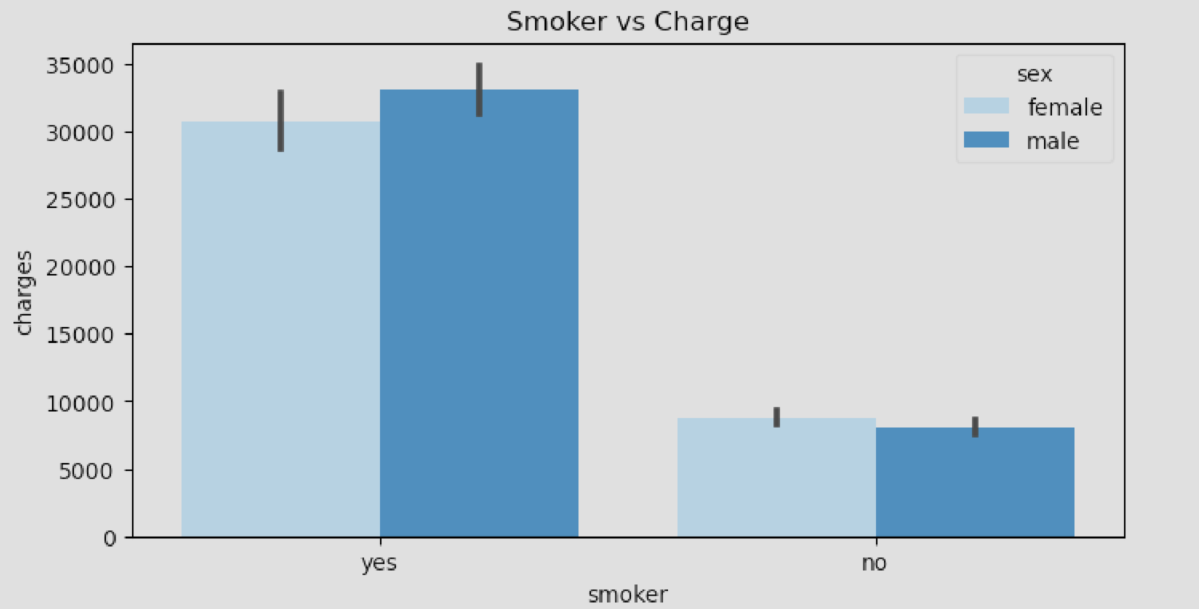


```
plt.figure(figsize=(5,6))

sns.scatterplot(x='bmi',y='charges',hue='sex',data=data,palette='Reds')

plt.title('BMI VS Charge')
```




```
plt.figure(figsize=(8,4))

plt.title('Smoker vs Charge')

sns.barplot(x='smoker',y='charges',data=data,palette='Blues',hue='sex')
```

<Axes: title={'center': 'Smoker vs Charge'}, xlabel='smoker', ylabel='charges'>



Observation:

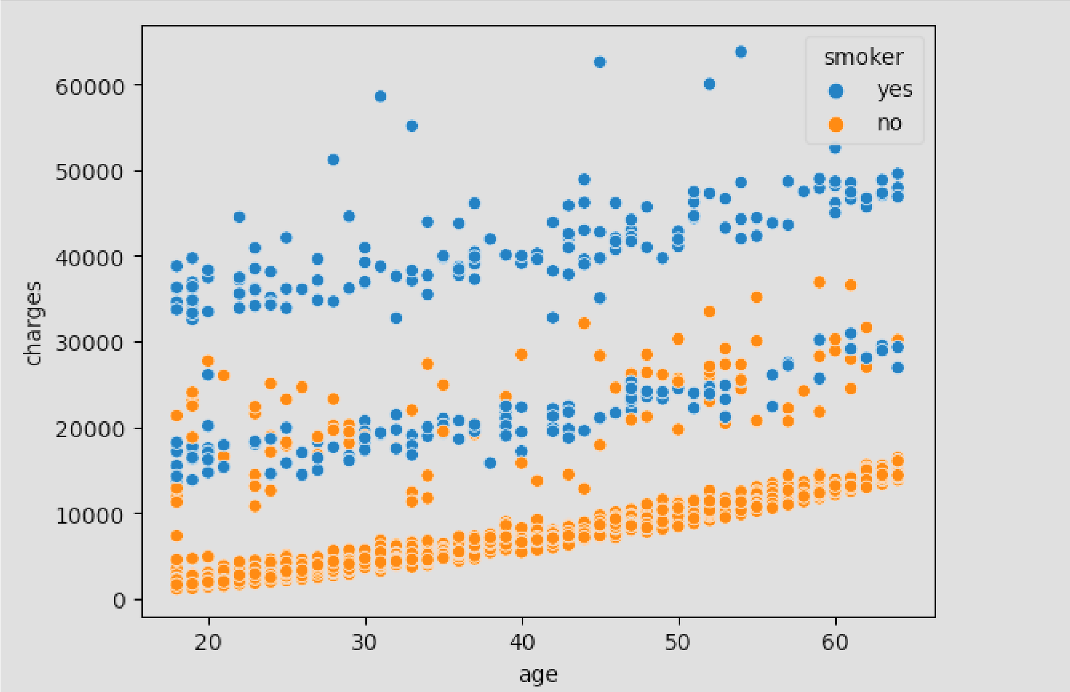
- **As we can see in the graph smokers are paying higher premiums compared to nonsmokers**

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

sns.scatterplot(x ="age", y="charges", hue='smoker', data=data)

plt.show()

```
sns.scatterplot(x ="age", y="charges", hue='smoker', data=data)
plt.show()
```



Observation:

- **As we can see the premium for nonsmokers remains constant with the increase of their age whereas, smokers pay a higher premium amount even at young age which increases with the increase in their age**

Project Submitted By

Mayur Nivadekar