Course End Project Solution: Insurance Data Analysis

Steps to be followed:

1. Importing the necessary libraries
2. Check the Shape and the Data Types of the Dataset
3. Dealing with missing values
4. Exploring the relationship between the feature and target columns
5. Plotting of feature vs feature plots
6. Plotting the **age** vs. **charges** graph

Step 1: Importing the Necessary Libraries

- Import NumPy, Pandas, matplotlib, and seaborn libraries.
- Load the dataset **insurance.csv**

In [ ]:

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

%matplotlib inline
```

In [ ]:

```
data = pd.read_csv("insurance.csv")
data
```

Out[ ]:

	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 × 7 columns

Observations:

- The dataset has 1338 rows and 7 columns.

Step 2: Checking the Shape and the Data Types of the Dataset

- Check the data types of the column to know which columns are categorical and numerical.

In [ ]:

```
data.shape
```

Out[ ]:

```
(1338, 7)
```

In [ ]:

```
data.dtypes
```

Out[ ]:

```
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.

Step 3: Dealing with Missing Values

- Find the count of missing values from different columns.

In [ ]:

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

Out[ ]:

```
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

Step 4: Exploring the Relationship Between the Feature and Target Columns

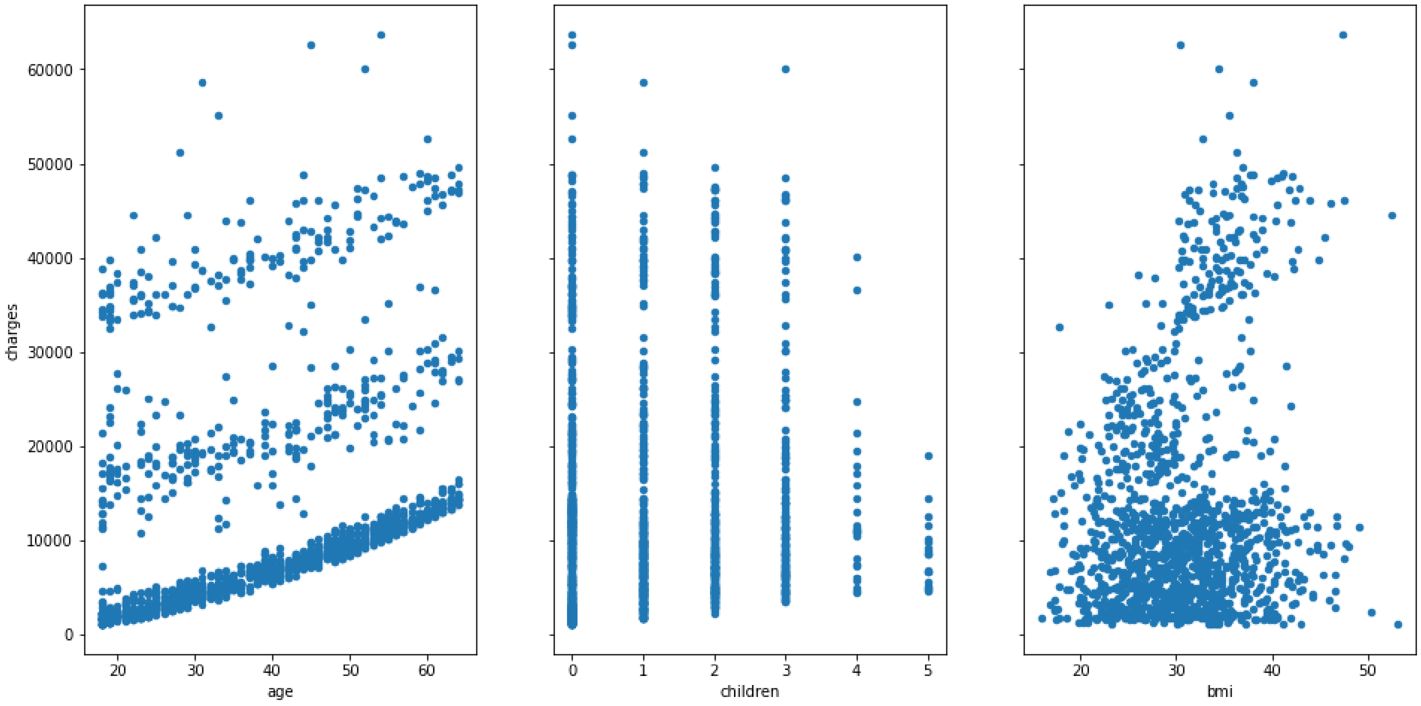
- Visualize the numerical columns using the scatter plot.
- Visualize the categorical columns using the count plot.

In [ ]:

```
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[ ]:

```
<AxesSubplot:xlabel='bmi', ylabel='charges'>
```

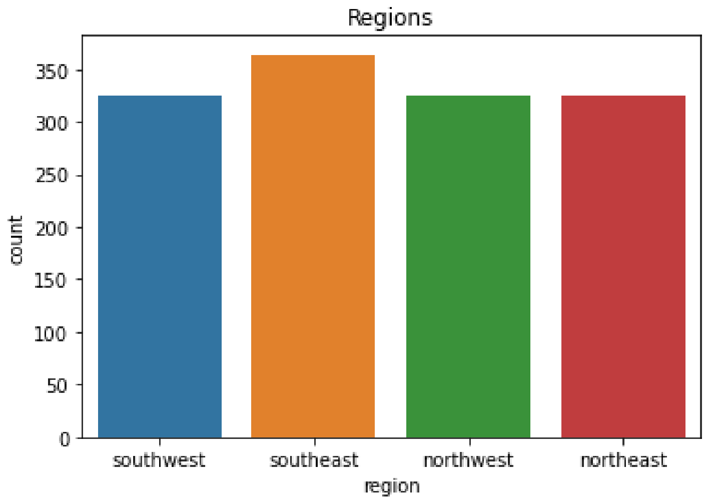


Observation:

- **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.**

In [ ]:

```
sns.countplot(data=data, x='region')
plt.title('Regions')
plt.show()
```



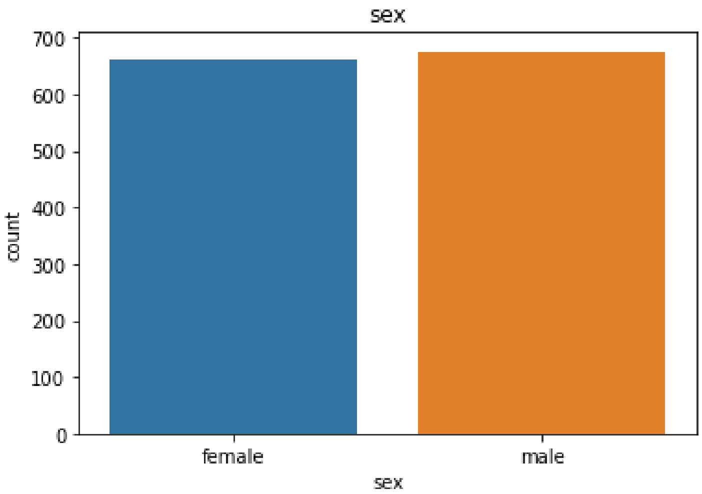
Observation:

- **You can see that the southeast region has the highest count.**

- **Plot a count plot for the sex column.**

In [ ]:

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



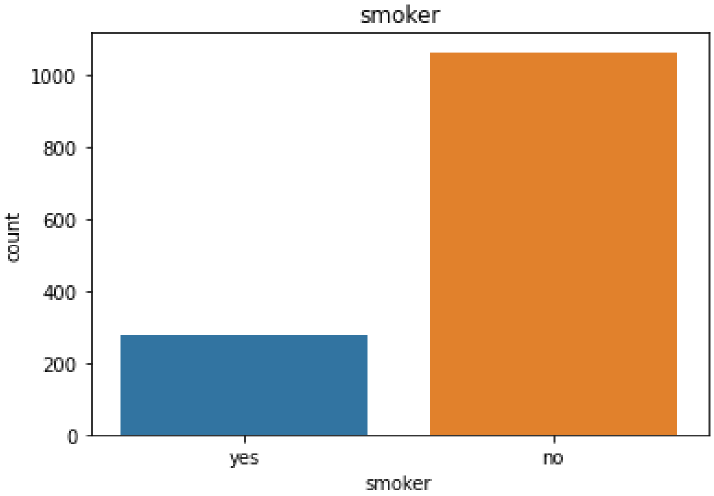
Observation:

- **The number of males and females is almost equal.**

- **Plot a count plot of the smoker column.**

In [ ]:

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



Observation:

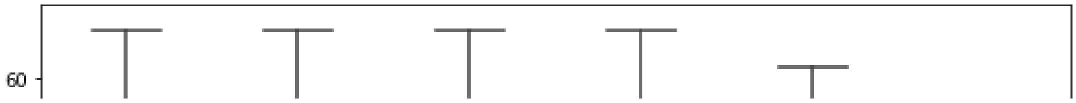
- **Most of the people who have taken insurance are not smokers.**

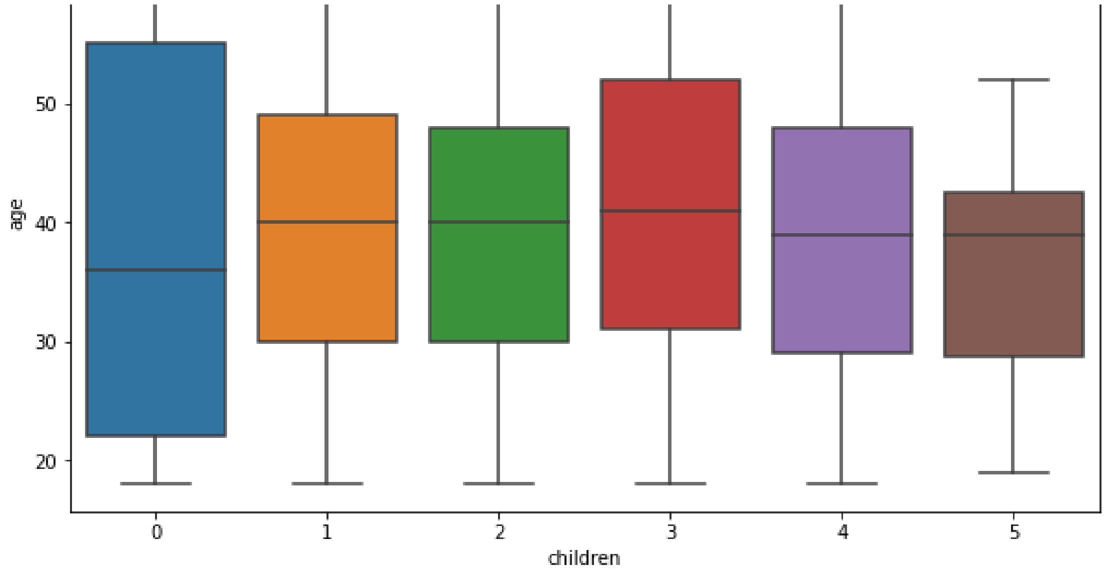
Step 5: Plotting of Feature vs Feature Plots

- **Data visualization using box plot**

In [ ]:

```
plt.figure(figsize=(10,6))
sns.boxplot(x='children',y='age',data=data)
plt.show()
```





Observation:

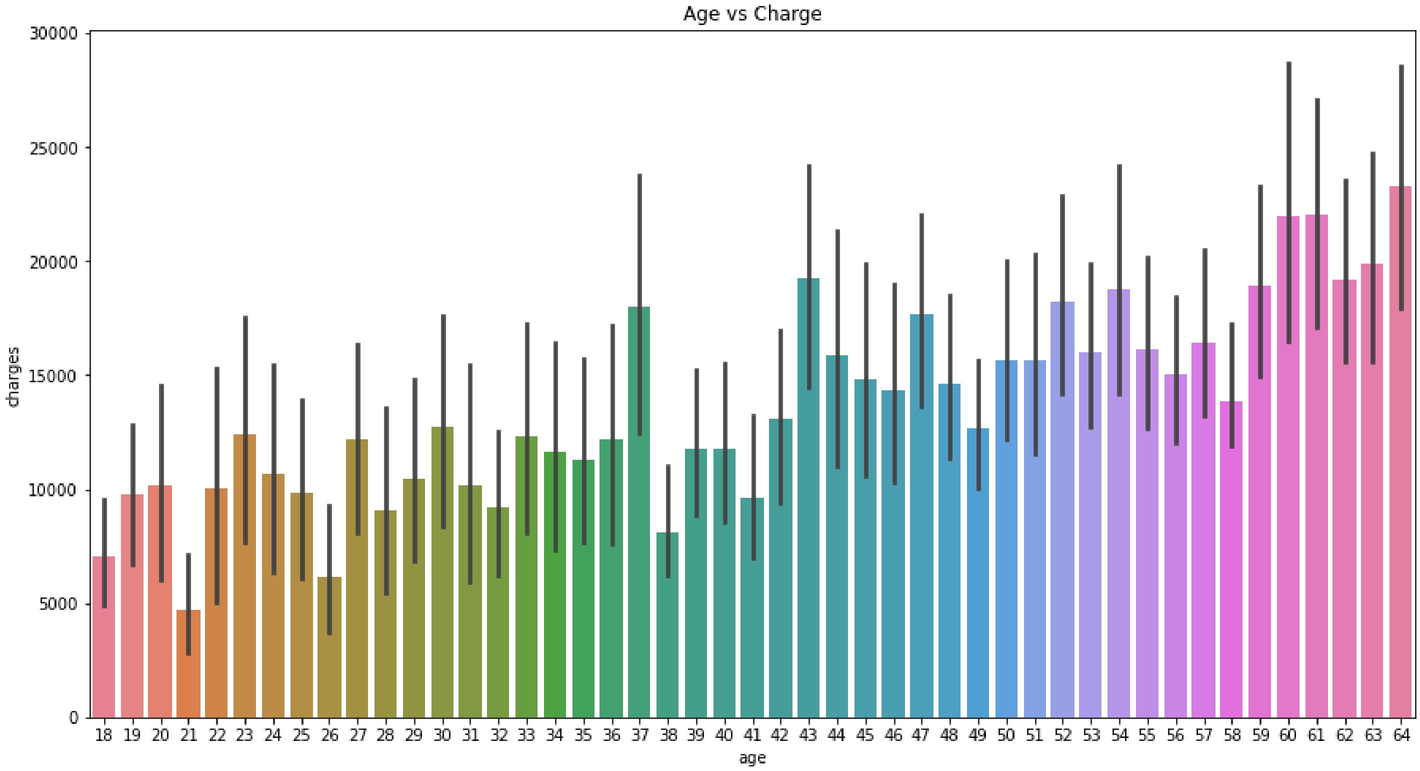
- **Now we are confirmed that there are no other outliers in the above-pre-processed column, we can proceed with EDA.**

In [ ]:

```
plt.figure(figsize=(15,8))
plt.title('Age vs Charge')
sns.barplot(x='age',y='charges',data=data,palette='husl')
```

Out[ ]:

```
<AxesSubplot:title={'center':'Age vs Charge'}, xlabel='age', ylabel='charges'>
```



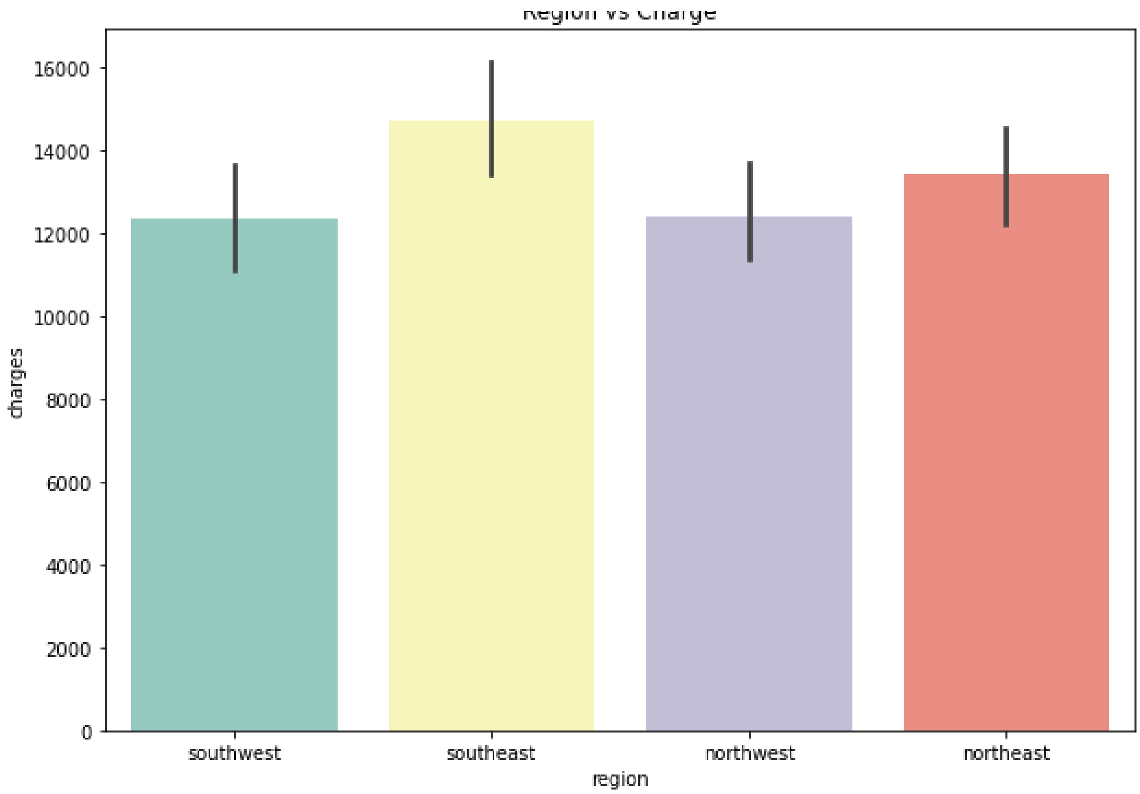
In [ ]:

```
plt.figure(figsize=(10,7))
plt.title('Region vs Charge')
sns.barplot(x='region',y='charges',data=data,palette='Set3')
```

Out[ ]:

```
<AxesSubplot:title={'center':'Region vs Charge'}, xlabel='region', ylabel='charges'>
```

Region vs Charge

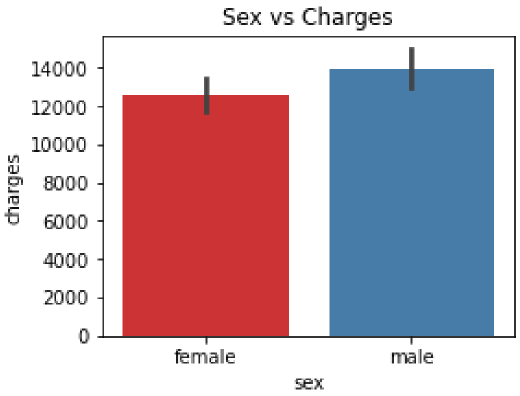


In [ ]:

```
plt.figure(figsize=(4,3))
plt.title('Sex vs Charges')
sns.barplot(x='sex',y='charges',data=data,palette='Set1')
```

Out[ ]:

```
<AxesSubplot:title={'center':'Sex vs Charges'}, xlabel='sex', ylabel='charges'>
```

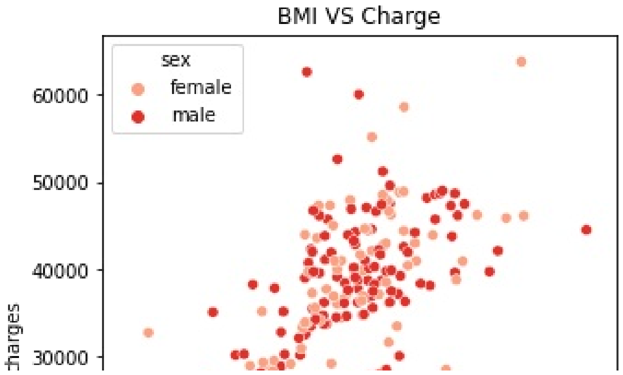


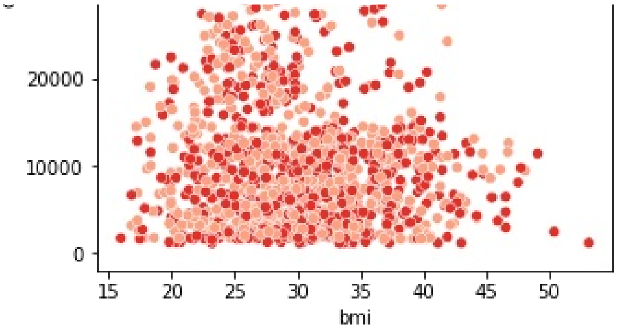
In [ ]:

```
plt.figure(figsize=(5,6))
sns.scatterplot(x='bmi',y='charges',hue='sex',data=data,palette='Reds')
plt.title('BMI VS Charge')
```

Out[ ]:

```
Text(0.5, 1.0, 'BMI VS Charge')
```



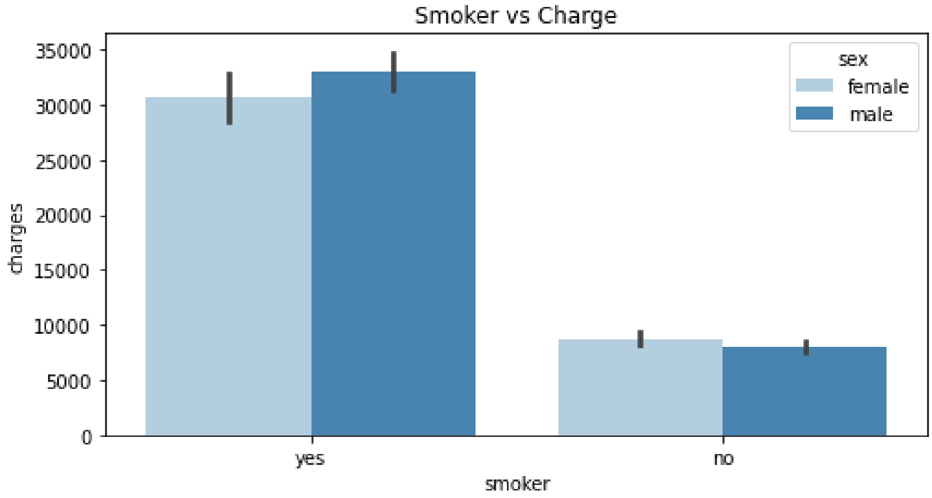


In [ ]:

```
plt.figure(figsize=(8,4))
plt.title('Smoker vs Charge')
sns.barplot(x='smoker',y='charges',data=data,palette='Blues',hue='sex')
```

Out[ ]:

```
<AxesSubplot: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.**

Step 6: Plotting the age vs. charges Graph

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

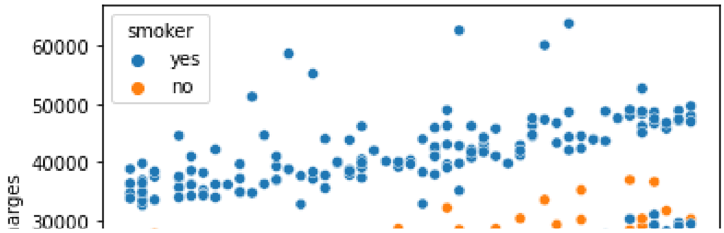
In [ ]:

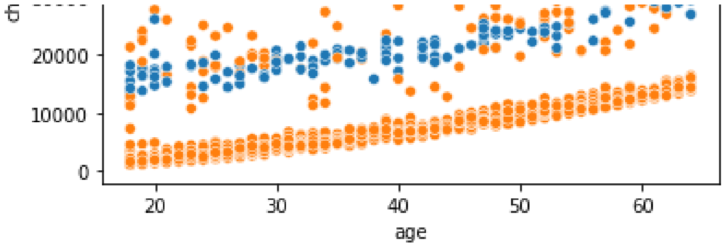
```
sns.scatterplot(x ="age", y="charges", hue='smoker', data=data)
plt.show()
```

```
C:\Users\alpika.gupta\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarnin
g: Pass the following variables as keyword args: x, y. From version 0.12, the only valid
positional argument will be `data`, and passing other arguments without an explicit keywo
rd will result in an error or misinterpretation.
  warnings.warn(
```

Out[ ]:

```
<AxesSubplot:xlabel='age', ylabel='charges'>
```





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.**