In [1]:

```
#import libraries
import pandas as pd
import numpy as np
import matplotlib as plt
```

In [2]:

```
# data importing and reading in of Data sets
df = pd.read_csv('plateau_Insurance.csv')
```

In [3]:

```
df
```

Out[3]:

```
      age     sex     bmi  children  smoker     region      charges
   0   21  female  16.000         1      no  northeast   3167.45585
   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   52  female  46.750         5      no  southeast  12592.53450
1335   54  female  47.410         0     yes  southeast  63770.42801
1336   37  female  47.600         2     yes  southwest  46113.51100
1337   46  female  48.070         2      no  northeast   9432.92530

1338 rows × 7 columns
```

In [4]:

```
#assess the dataset's description
df.describe()
```

Out[4]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.682687	1.094918	13270.422265
std	14.049960	6.145674	1.205493	12110.011237
min	18.000000	14.000000	0.000000	1121.873900
25%	27.000000	26.315000	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.700000	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

In [5]:

```
#assess the dataset for data duplicates
df.duplicated().sum()
```

Out[5]:

```
1
```

In [6]:

```
#assess the dataset for missing data
df.isnull().sum()
```

Out[6]:

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

EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

In [7]:

```
#importing visualization libraries
import matplotlib as plt
import seaborn as sns
%matplotlib inline
```

In [8]:

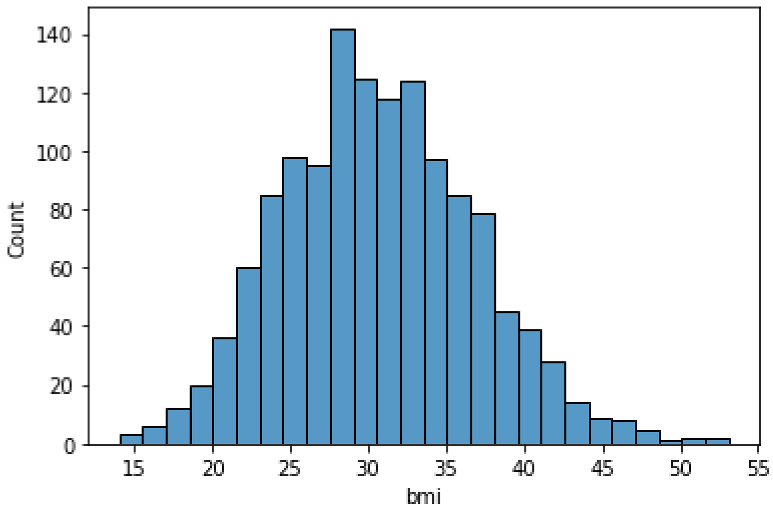
```
#univariate analysis for the body mass index feature
df['bmi'].describe()
```

Out[8]:

```
count    1338.000000
mean       30.682687
std         6.145674
min        14.000000
25%        26.315000
50%        30.400000
75%        34.700000
max        53.130000
Name: bmi, dtype: float64
```

In [9]:

```
sns.histplot(df['bmi'])
plt.style.use('bmh')
```



In [10]:

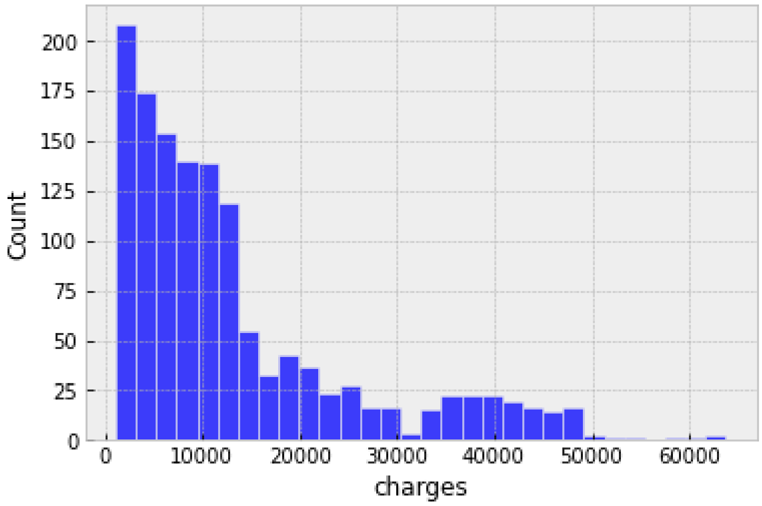
```
#univariate analysis for the charges feature
df['charges'].describe()
```

Out[10]:

```
count     1338.000000
mean     13270.422265
std      12110.011237
min       1121.873900
25%       4740.287150
50%       9382.033000
75%      16639.912515
max      63770.428010
Name: charges, dtype: float64
```

In [11]:

```
sns.histplot(df['charges'])
plt.style.use('bmh')
```



In [12]:

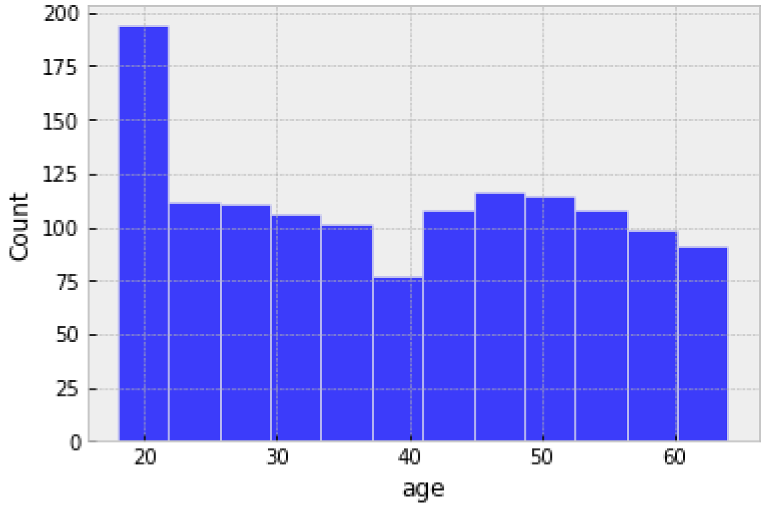
```
#univariate analysis for the age feature
df['age'].describe()
```

Out[12]:

```
count    1338.000000
mean       39.207025
std        14.049960
min        18.000000
25%        27.000000
50%        39.000000
75%        51.000000
max        64.000000
Name: age, dtype: float64
```

In [13]:

```
sns.histplot(df['age'])
plt.style.use('bmh')
```



In [14]:

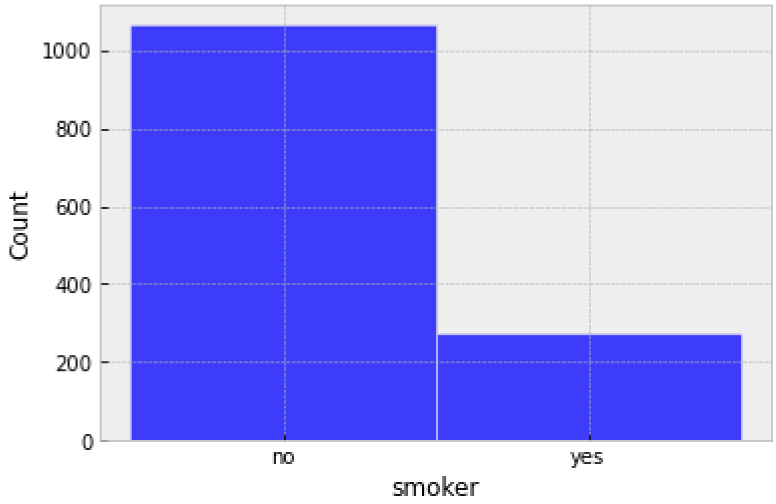
```
#univariate analysis for the smoker feature
df['smoker'].describe()
```

Out[14]:

```
count      1338
unique        2
top          no
freq       1064
Name: smoker, dtype: object
```

In [15]:

```
sns.histplot(df['smoker'])
plt.style.use('bmh')
```



OBSERVATIONS

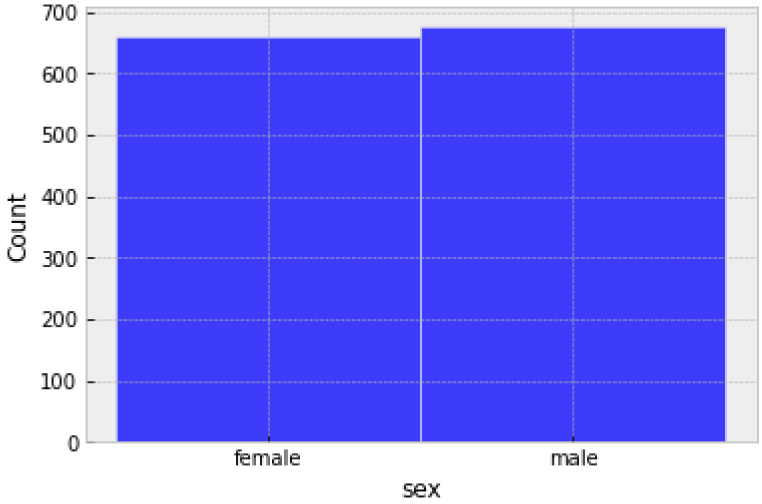
Highlight observations from the univariate analysis.

The univariate analysis is conducted for the purpose of making data easier to interpret and to understand how data is distributed within a sample of population being studied.

1. In this analysis, you can see that the number of smokers are more in the southeasthern part than the other region stated.

In [16]:

```
#graphical visualiation of the sex feature
sns.histplot(df['sex'])
plt.style.use('bmh')
```

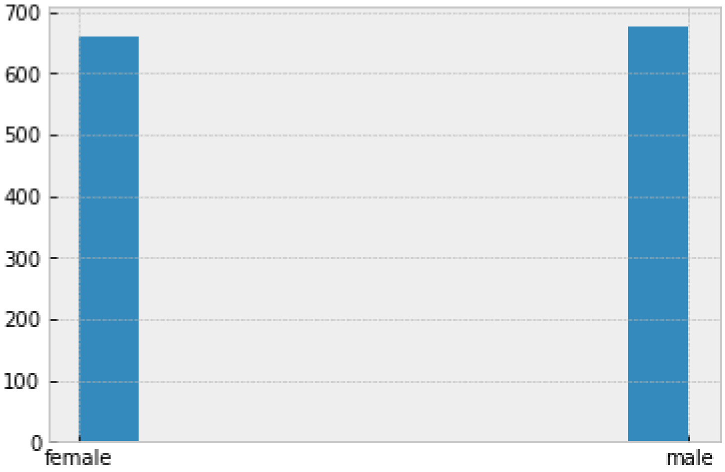


In [17]:

```
#OR WE CAN HAVE THIS
df['sex'].hist()
```

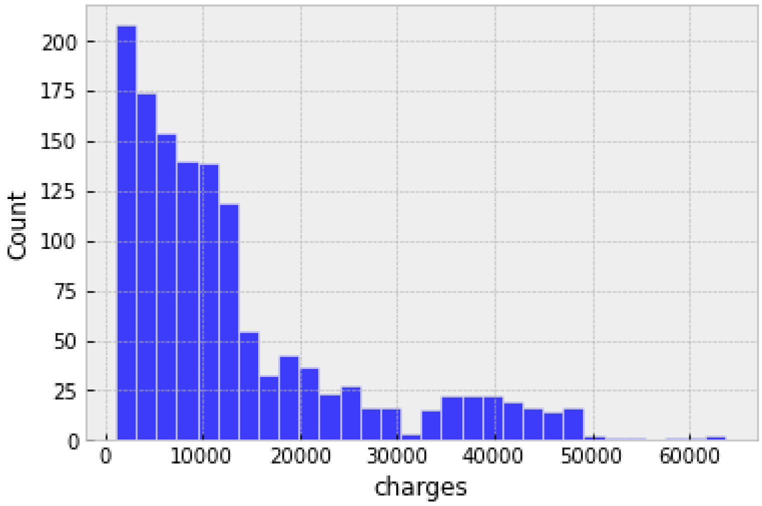
Out[17]:

```
<AxesSubplot:>
```



In [18]:

```
#graphical visualiation of the target label(charges feature)
sns.histplot(df['charges'])
plt.style.use('bmh')
```

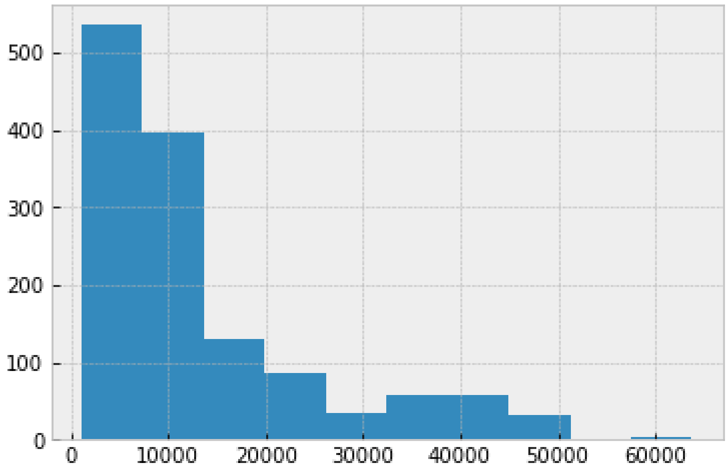


In [19]:

```
#OR WE CAN HAVE THIS
df['charges'].hist()
```

Out[19]:

```
<AxesSubplot:>
```



In [20]:

```
import seaborn as sns
%matplotlib inline
import matplotlib as plt
df = pd.read_csv('plateau_Insurance.csv')
```

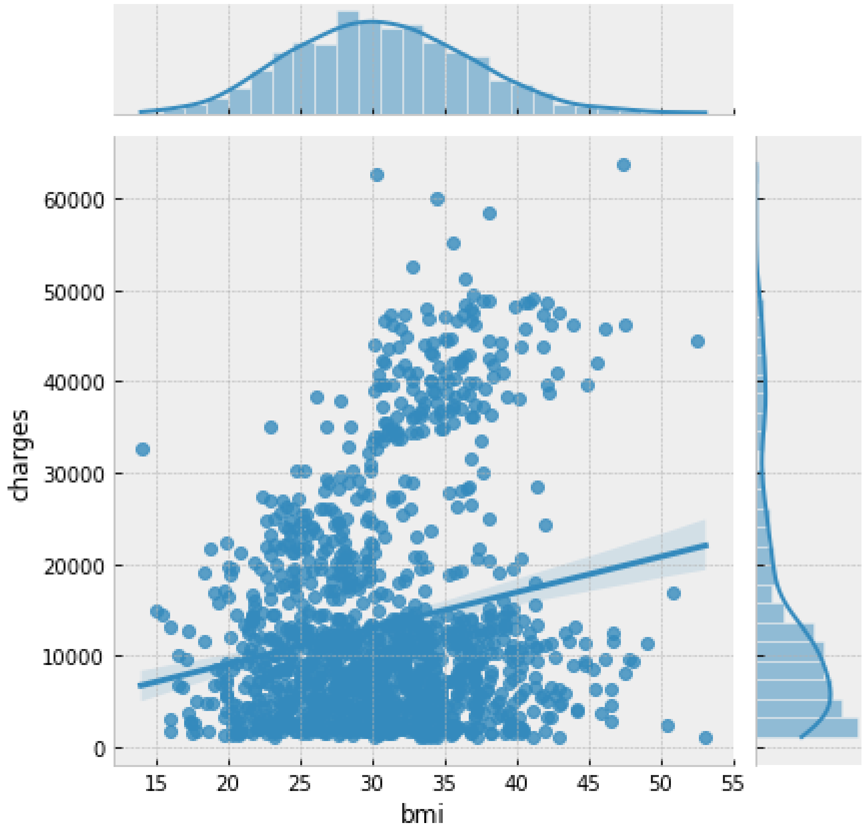
BIVARIATE ANANLYSIS

In [21]:

```
#jointplot exploring the relationship between the bmi vs charges features
sns.jointplot(data=df,x='bmi',y='charges',kind='reg',palette='Greens')
```

Out[21]:

```
<seaborn.axisgrid.JointGrid at 0x21de2488a90>
```

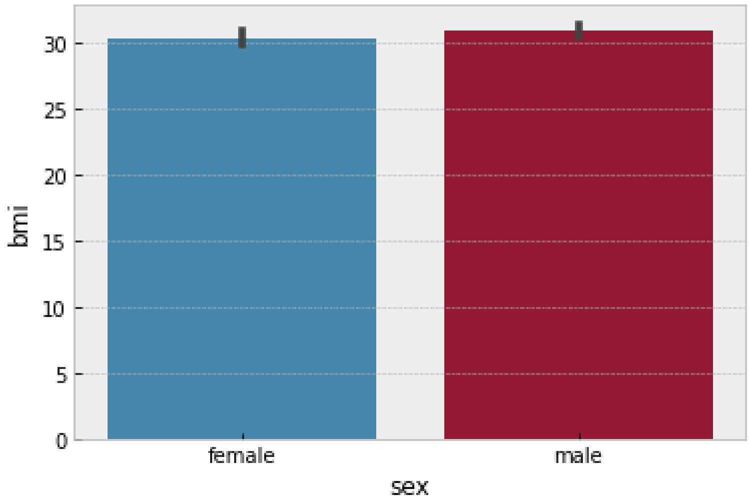


In [22]:

```
#exploring the distributions and the relationships between the bmi vs the sex/gender featur
sns.barplot(data=df, x='sex',y='bmi')
```

Out[22]:

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



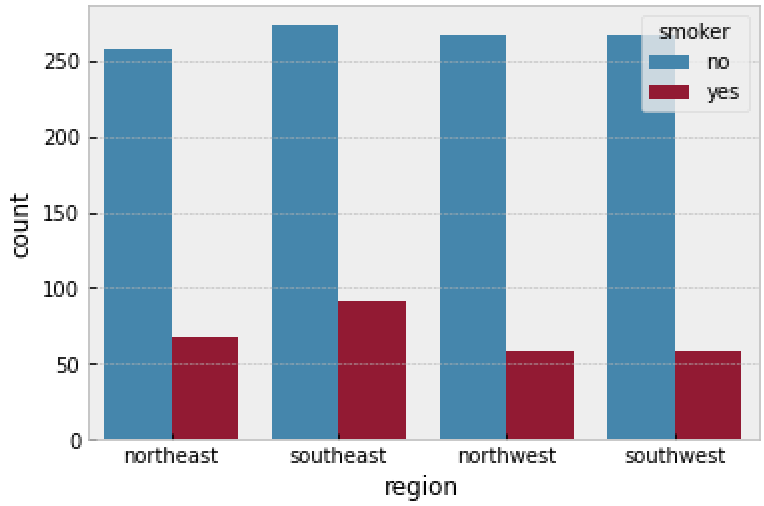
#2. Prove (or disprove) with that the BMI of females is different from that of males This satisfies the second objective proving that the bmi of female is different from men as we can see from the barplot above.

In [23]:

```
#bivariate analysis of the region vs smoker features
sns.countplot(data=df,x='region',hue='smoker')
```

Out[23]:

```
<AxesSubplot:xlabel='region', ylabel='count'>
```



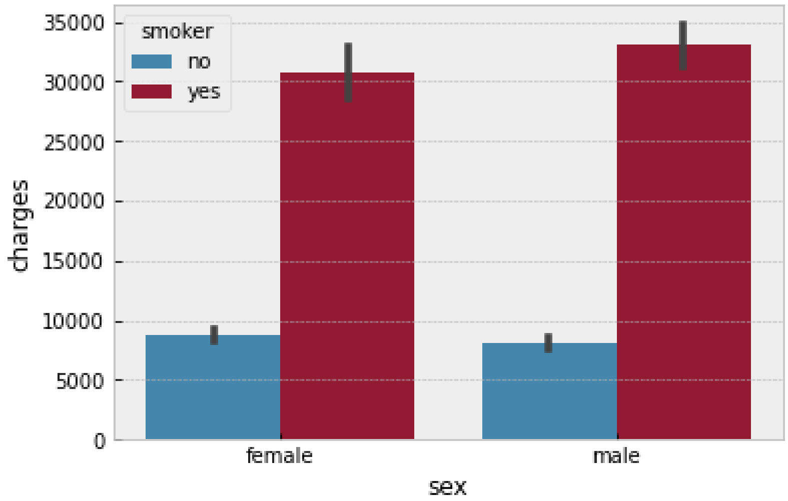
#3. Does the smoking habit of customers depend on their region? Yes it does, looking at the southesthern region we can see that the level of smokers is higher compare to the rest.

In [24]:

```
#exploring the relationship between the sex vs smoker vs charges features
sns.barplot(data=df,x='sex',y='charges',hue='smoker')
```

Out[24]:

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



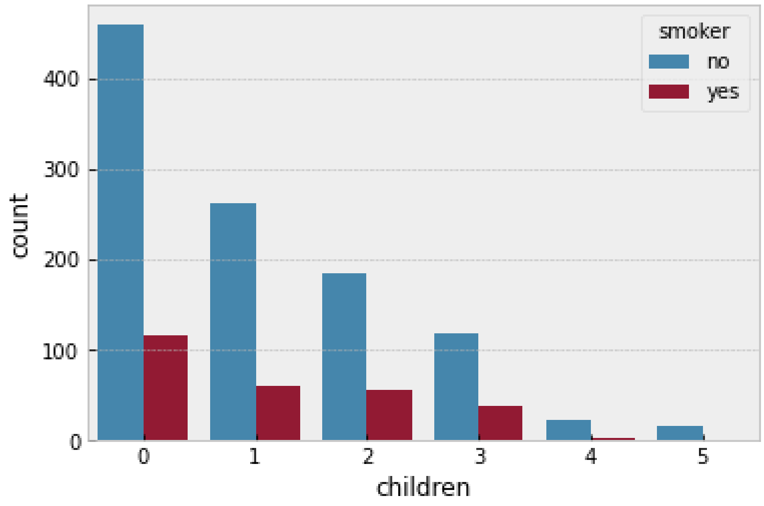
#1.Prove(or disprove) that the medical claims made by the people who smoke are greater than those who don't? This satisfies the first objectives showing that the medical claim of smokers is greater than non smokers.

In [25]:

```
#exploring the relationship between the number of children feature and the smoker feature
sns.countplot(data=df,x='children',hue='smoker')
```

Out[25]:

```
<AxesSubplot:xlabel='children', ylabel='count'>
```

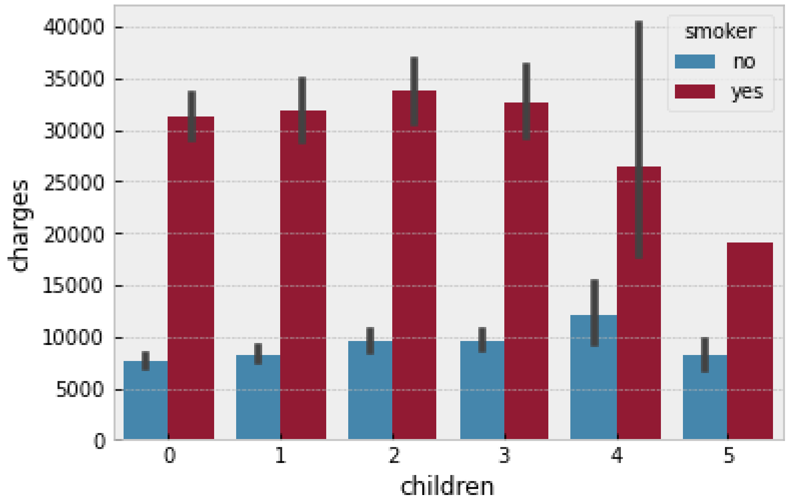


In [26]:

```
#bivariate analysis of the number of children feature Vs the medical claim charges feature
sns.barplot(data=df,x='children', y='charges',hue='smoker')
```

Out[26]:

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



Observation

state down your observtaions from the bivariate analysis

The bivariate analysis is usually conducted to determine whether a statistical association exist between two variables. Now In this analysis, I observed that:

1. In the jointplot exploring the relationship between 'bmi' and 'charges', the correlation between the two data is more concentrated at the point 35(bmi) and 1500(charges).
2. In the barplot between 'sex' and 'bmi', we can see that the bmi of female is different from men, the men bmi is higher than females.
3. looking at the southesthern region from the countplot, we can see that the level of smokers is higher compare to the other region.

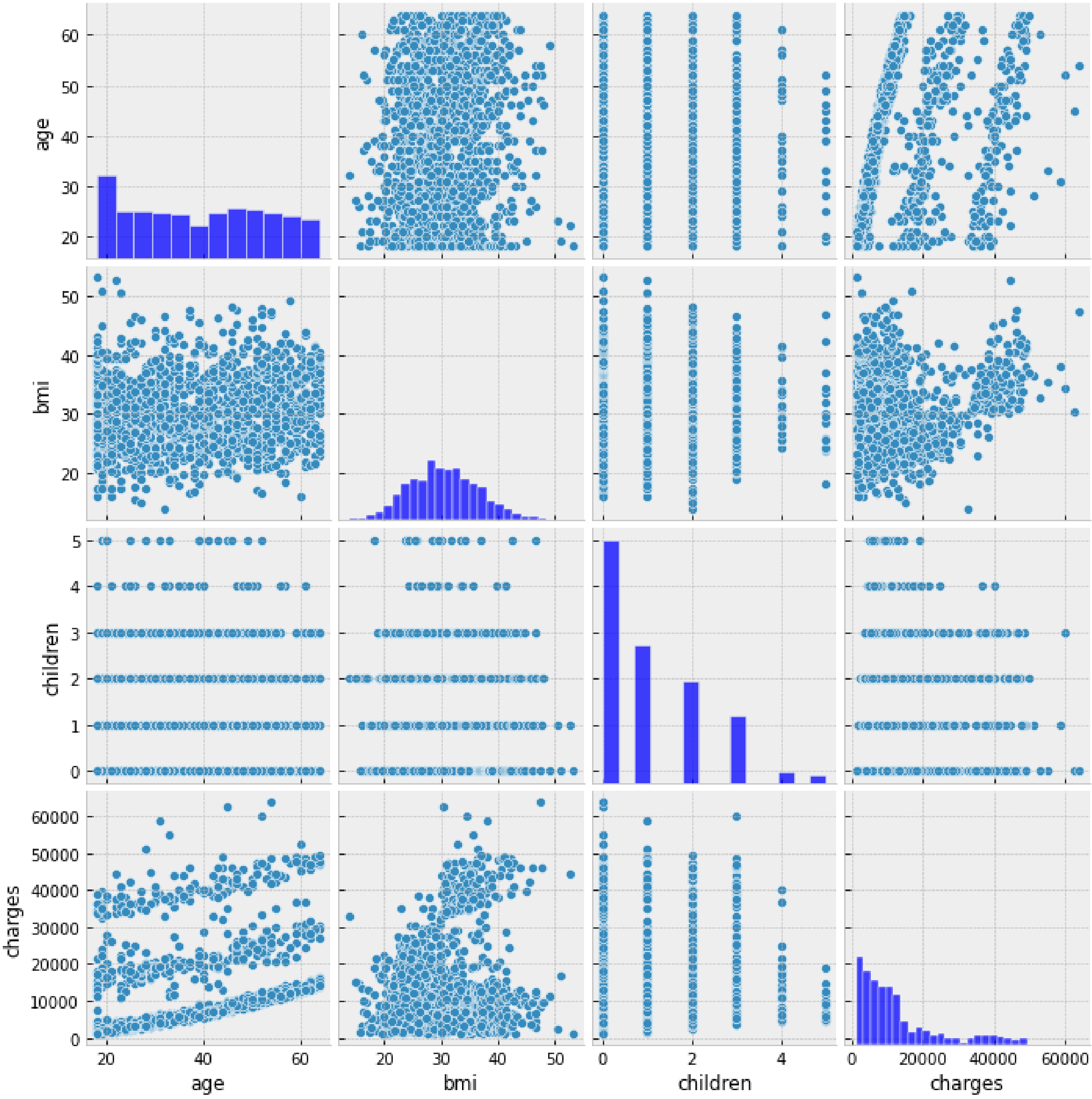
Multivariate Analaysis

In [27]:

```
#use a pairplot to explore the relationship between the nummerical features in the dataset
sns.pairplot(df)
```

Out[27]:

```
<seaborn.axisgrid.PairGrid at 0x21de4ef4c70>
```

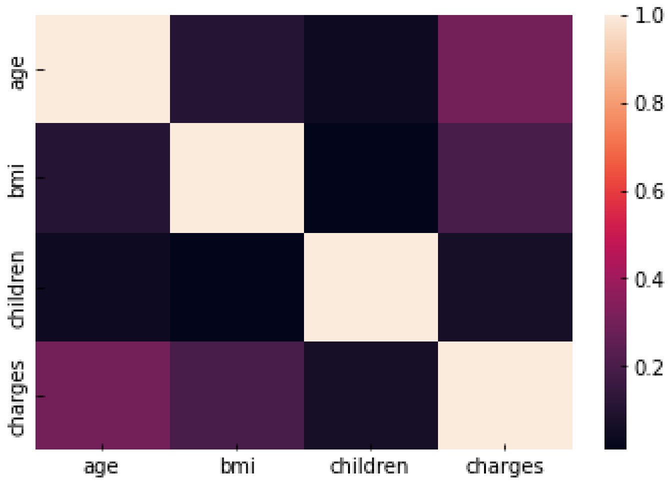


In [28]:

```
## heatmeap to see the correlation between features.
sns.heatmap(df.corr())
```

Out[28]:

```
<AxesSubplot:>
```



Note: the following dataset below satisfies the fourth objectives question that asked if the mean bmi of women with no children, one child or two children are the same.

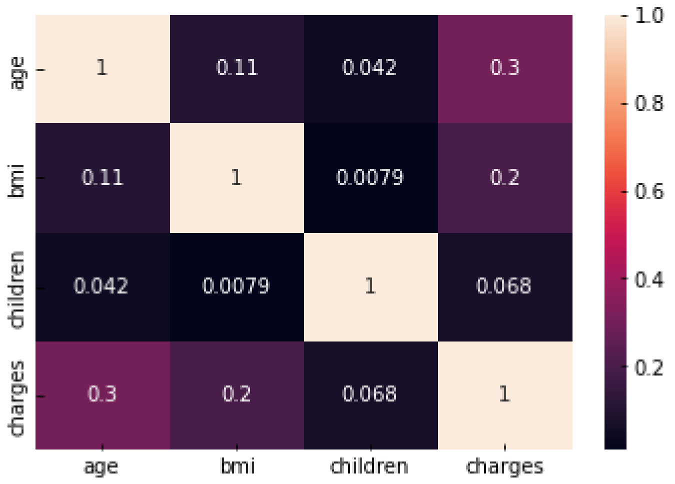
This shows that their respective mean is not the same as we can see from the zero child, one child and even two children, there is difference in the mean.

In [29]:

```
sns.heatmap(df.corr(),annot=True)
```

Out[29]:

```
<AxesSubplot:>
```



In [30]:

```
df.groupby(['sex','children'])['bmi'].mean()
```

Out[30]:

```
sex     children
female  0           30.485502
        1           30.047500
        2           30.572437
        3           30.436429
        4           31.943182
        5           30.620625
male    0           30.741719
        1           31.166145
        2           31.300992
        3           30.922937
        4           30.957500
        5           28.792500
Name: bmi, dtype: float64
```

Observation

state down your observtaions from the multivariate analysis

The multivariate analysis is a statistical study of the data where multiple measurements are made on each experimental unit and where the relationships among multivariate measurements and their structure are important. It is also used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest. In this analysis;

1. From the heatmap, we can viusalize the strength of relationships between the numerical variables and the variables that are related to each other.
2. You can see that the respective mean of the bmi of women with no children, one child or two children are not the same.

3. Also from the pairplot, we can observe the distribution of each variable as a shown histogram along the diagonal boxes and all other boxes display a scatterplot of the relationship between each pairwise combination of variables.

In [31]:

```
pip install -U pandas-profiling[notebook]
```

```
Requirement already satisfied: ipykernel>=4.5.1 in c:\users\peter\anaconda
3\lib\site-packages (from ipywidgets>=7.5.1->pandas-profiling[notebook])
(6.9.1)
Requirement already satisfied: ipython-genutils~=0.2.0 in c:\users\peter\a
naconda3\lib\site-packages (from ipywidgets>=7.5.1->pandas-profiling[noteb
ook]) (0.2.0)
Requirement already satisfied: widgetsnbextension~=3.5.0 in c:\users\peter
\anaconda3\lib\site-packages (from ipywidgets>=7.5.1->pandas-profiling[not
ebook]) (3.5.2)
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in c:\users\peter
\anaconda3\lib\site-packages (from ipywidgets>=7.5.1->pandas-profiling[not
ebook]) (1.0.0)
Requirement already satisfied: traitlets>=4.3.1 in c:\users\peter\anaconda
3\lib\site-packages (from ipywidgets>=7.5.1->pandas-profiling[notebook])
(5.1.1)
Requirement already satisfied: ipython>=4.0.0 in c:\users\peter\anaconda3
\lib\site-packages (from ipywidgets>=7.5.1->pandas-profiling[notebook])
(8.2.0)
Requirement already satisfied: tornado<7.0,>=4.2 in c:\users\peter\anacond
a3\lib\site-packages (from ipykernel>=4.5.1->ipywidgets>=7.5.1->pandas-pro
filing[notebook]) (6.1)
```

In [32]:

```
import seaborn as sns
import pandas as pd
import numpy as np
```

In [33]:

```
from pandas_profiling import ProfileReport
```

In [34]:

```python
df.profile_report()
```

```
Summarize dataset:                                   36/36 [00:17<00:00, 3.30it/s,
100%                                                 Completed]

Generate report structure:                                  1/1 [00:06<00:00,
100%                                                        6.47s/it]

Render HTML: 100%                                         1/1 [00:01<00:00, 1.94s/it]
```

Overview

Dataset statistics

Number of variables	7
Number of observations	1338
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	1
Duplicate rows (%)	0.1%
Total size in memory	73.3 KiB
Average record size in memory	56.1 B

Variable types

Numeric	4
Categorical	2
Boolean	1

Alerts

Dataset has 1 (0.1%) duplicate rows	**Duplicates**
`age` is highly correlated with `charges`	**High correlation**
`charges` is highly correlated with `age` and 1 other fields (age, smoker)	**High correlation**
`smoker` is highly correlated with `charges`	**High correlation**

```
Out[34]:
```

Thanks!

KARNAP BINSAK RIMVEN.