Introduction

Health insurance in India is a growing segment of India's economy. The Indian health system is one of the largest in the world, with the number of people it concerns: nearly 1.3 billion potential beneficiaries. The health industry in India has rapidly become one of the most important sectors in the country in terms of income and job creation. In 2018, one hundred million Indian households (500 million people) do not benefit from health coverage. In 2011, 3.9%[1] of India's gross domestic product was spent in the health sector.

According to the World Health Organization (WHO), this is among the lowest of the BRICS (Brazil, Russia, India, China, South Africa) economies. Policies are available that offer both individual and family cover. Out of this 3.9%, health insurance accounts for 5-10% of expenditure, employers account for around 9% while personal expenditure amounts to an astounding 82%.

In the year 2016, the NSSO released the report “Key Indicators of Social Consumption in India: Health” based on its 71st round of surveys. The survey carried out in the year 2014 found out that, more than 80% of Indians are not covered under any health insurance plan, and only 18% (government funded 12%) of the urban population and 14% (government funded 13%) of the rural population was covered under any form of health insurance.

India's public health expenditures are lower than those of other middle-income countries. In 2012, they accounted for 4% of GDP, which is half as much as in China with 5.1%. In terms of public health spending per capita, India ranks 184th out of 191 countries in 2012. Patients' remaining costs represent about 58% of the total.[4] The remaining costs borne by the patient represent an increasing share of the household budget, from 5% of this budget in 2000 to over 11% in 2004-2005.[5] On average, the remaining costs of poor households as a result of hospitalization accounted for 140% of their annual income in rural areas and 90% in urban areas.

This financial burden has been one of the main reasons for the introduction of health insurance covering the hospital costs of the poorest.

Data Description:

The data at hand contains medical costs of people characterized by certain attributes.

Domain:

Healthcare

Context:

Leveraging customer information is paramount for most businesses. In the case of an insurance company, attributes of customers like the ones mentioned below can be crucial in making business decisions. Hence, knowing to explore and generate value out of such data can be an invaluable skill to have.

Attribute Information

age : age of primary beneficiary

sex : insurance contractor gender, female, male

bmi : Body mass index, providing an understanding of body,

weights that are relatively high or low relative to height,

objective index of body weight (kg / m ^ 2) using the ratio of

height to weight, ideally 18.5 to 24.9

children : Number of children covered by health insurance /

Number of dependents

smoker : Smoking

region : the beneficiary's residential area in the US, northeast,southeast, southwest, northwest.

charges : Individual medical costs billed by health insurance.

Import all the necessary libraries

In [1]:
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
import seaborn as sns
import statsmodels.api as sm
import scipy.stats as stats
import copy
import os
```

In [2]:
```
sns.set() #setting the default seaborn style for our plots
```

In [3]:
```
os.getcwd() # Checking Working directory
```

Out[3]:
```
'C:\\Users\\PC'
```

In [4]:
```
import os

# Get the current working directory
current_directory = os.getcwd()
print(f"Current Workng Directory: {current_directory}")

# Change the current working directory to a new path
new_directory = "C:\\Users\\PC\\Downloads\\GL_P"
os.chdir(new_directory)

# verify the change
updated_directory = os.getcwd()
print(f"Updated Working Directory: {updated_directory}")
```

```
Current Workng Directory: C:\Users\PC
Updated Working Directory: C:\Users\PC\Downloads\GL_P
```

Read the data into the notebook

In [5]:
```
df = pd.read_csv('insurance.csv') # read the data as a data frame
```

In [6]:
```
df.head() # checking the head of the data frame
```

Out[6]:

	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

Basic EDA

Find the shape of the data,data type of individual columns

Check the presence of missing values

Descriptive stats of numerical columns

Find the distribution of numerical columns and the asssociated skeweness and presence of outliers

Distribution of categorical columns

In [7]:
```
df.info() # info about the data
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

In [10]:
```
df.shape # The dataset contains 1338 observation of data and 7 variables.
```

Out[10]:
```
(1338, 7)
```

Check for missing value

In [11]:
```
## Check for missing value in any colum
df.isnull().sum()
```

Out[11]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

In [12]:
```
## There are no null values in any of the column
```

In [17]:
```
## Checking the summary of the dataset
df.describe().T # five point summary of the continuous attributes
```

Out[17]:

	count	mean	std	min	25%	50%	75%	
age	1338.0	39.207025	14.049960	18.0000	27.00000	39.000	51.000000	
bmi	1338.0	30.663397	6.098187	15.9600	26.29625	30.400	34.693750	
children	1338.0	1.094918	1.205493	0.0000	0.00000	1.000	2.000000	
charges	1338.0	13270.422265	12110.011237	1121.8739	4740.28715	9382.033	16639.912515	637

Data looks legit as all the statistics seem reasonable -Looking at the age column, data looks representative of the true age distribution of the adult population -Very few people have more than 2 children. 75% of the people have 2 or less children -The claimed amount is higly skewed as most people would require basic medi-care and only few suffer from diseases which cost more to get rid of

In [18]:
```
df.describe(include='all').T # include object column as also
```

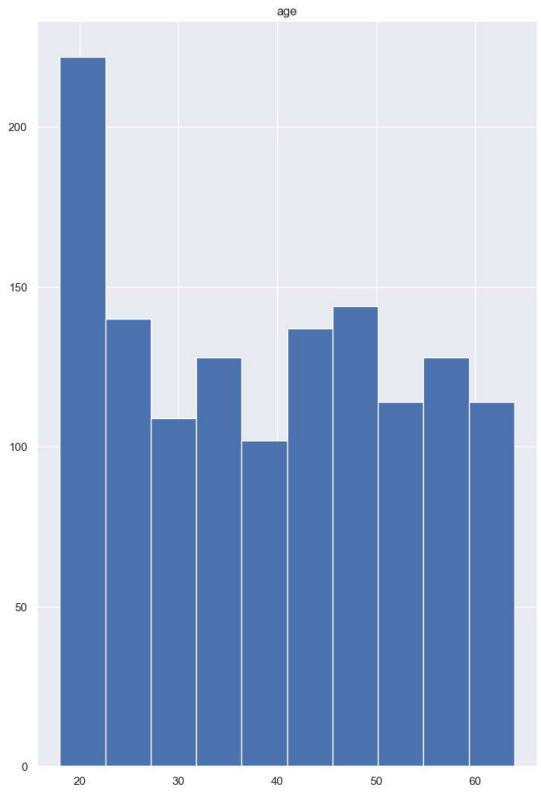
Out[18]:

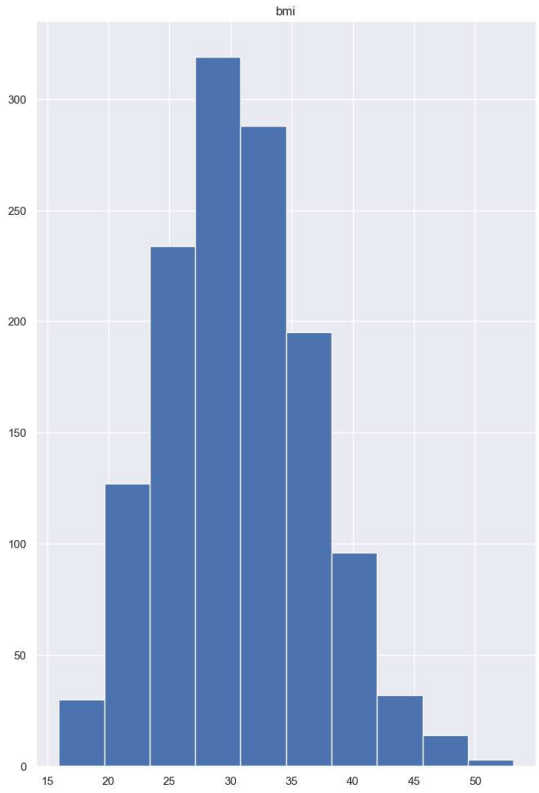
	count	unique	top	freq	mean	std	min	25%
age	1338.0	NaN	NaN	NaN	39.207025	14.04996	18.0	27.0
sex	1338	2	male	676	NaN	NaN	NaN	NaN
bmi	1338.0	NaN	NaN	NaN	30.663397	6.098187	15.96	26.29625
children	1338.0	NaN	NaN	NaN	1.094918	1.205493	0.0	0.0
smoker	1338	2	no	1064	NaN	NaN	NaN	NaN
region	1338	4	southeast	364	NaN	NaN	NaN	NaN
charges	1338.0	NaN	NaN	NaN	13270.422265	12110.011237	1121.8739	4740.28715

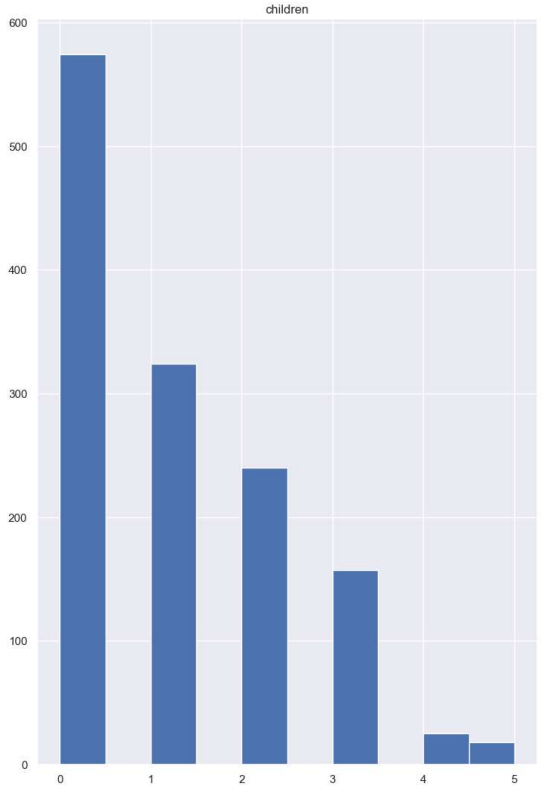
Plot the Histograms

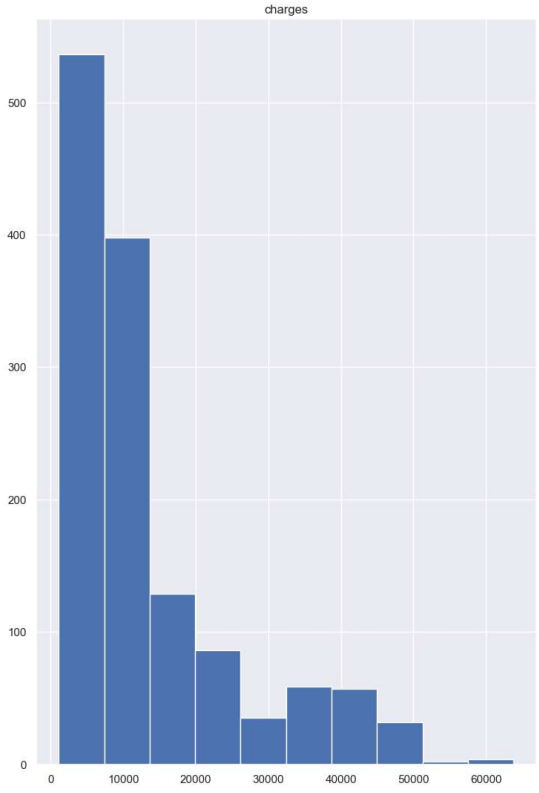
In [23]:
```
# Plots to see the distribution of the continuous features
df.hist(figsize=(20,30))
```

Out[23]:
```
array([[<Axes: title={'center': 'age'}>, <Axes: title={'center': 'bmi'}>],
       [<Axes: title={'center': 'children'}>,
        <Axes: title={'center': 'charges'}>]], dtype=object)
```









bmi looks quiet normally distributed

Age seems be be distributed quiet uniformly

As seen in the previous step, charges are highly skewed

In [24]:
```
Skewness = pd.DataFrame({'Skewness' : [stats.skew(df.bmi),stats.skew(df.age),s
                        index=['bmi','age','charges'])  # Measure the skewenes
Skewness
```

Out[24]:

	Skewness
bmi	0.283729
age	0.055610
charges	1.514180

Skew of bmi is very less as seen in the previous step

age is uniformly distributed and there's hardly any skew

charges are highly skewed

Check Outliers

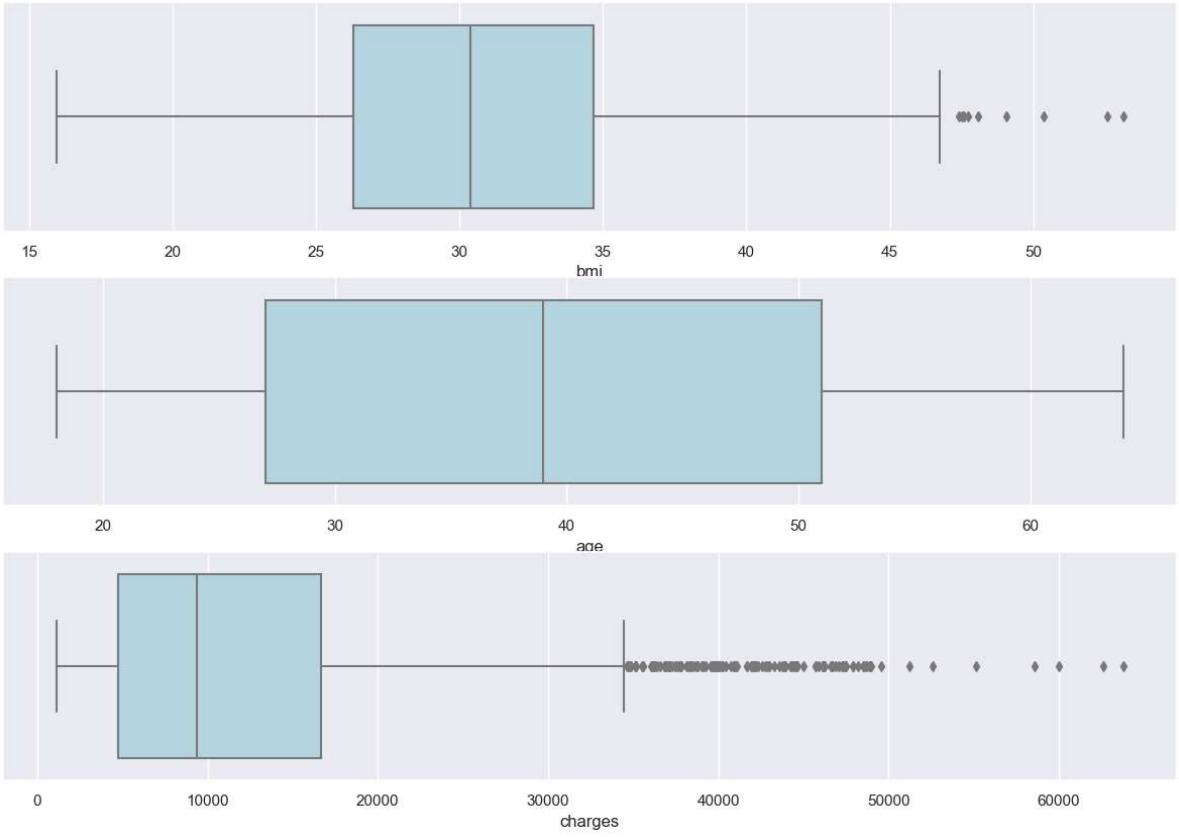
In [28]:
```
## Box plot will be plotted to check for outliers

plt.figure(figsize= (15,10))
plt.subplot(3,1,1)
sns.boxplot(x= df.bmi, color = 'lightblue')

plt.subplot(3,1,2)
sns.boxplot(x= df.age, color='lightblue')

plt.subplot(3,1,3)
sns.boxplot(x=df.charges, color='lightblue')

plt.show()
```



There are no outliers present in the age variable.

bmi variable shows presence of few extreme values

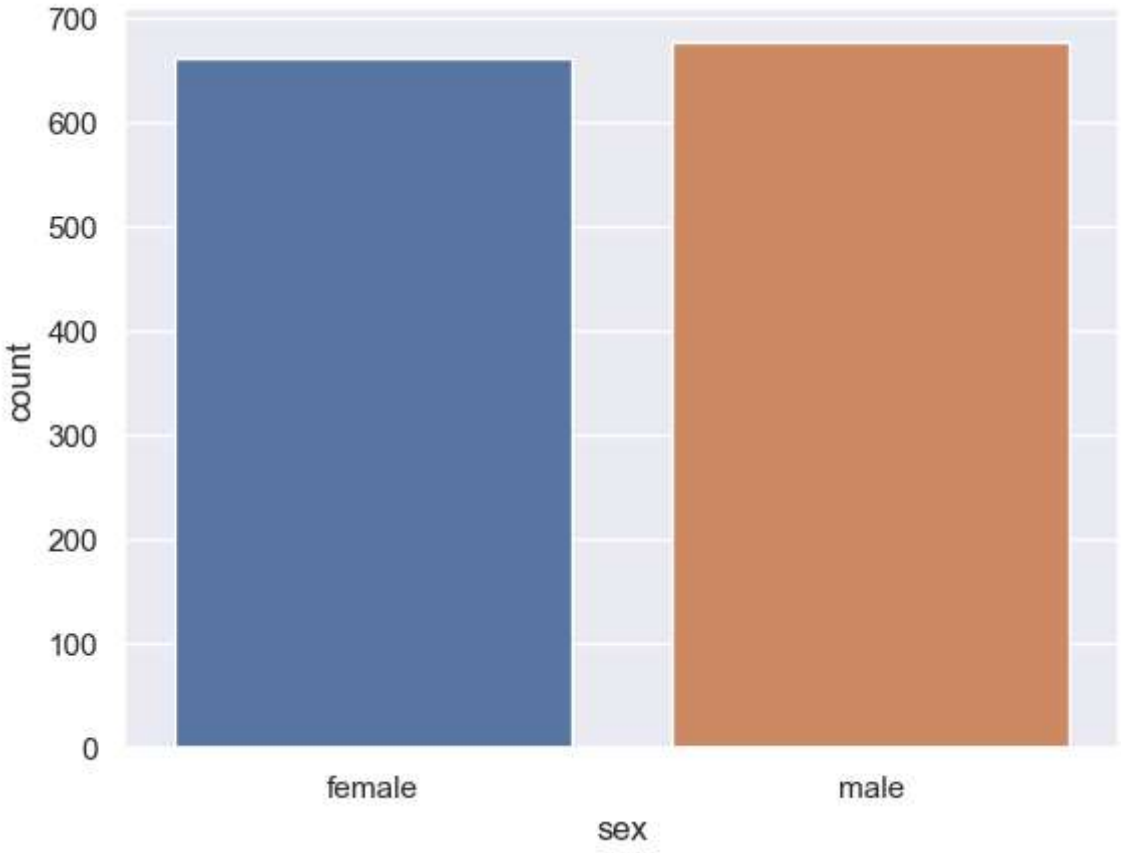
charges as it is highly skewed, there are quiet a lot of extreme values.

Plot Count Plot

In [30]:
```
## We will plot various count plot to see how the variable has been distribute
```

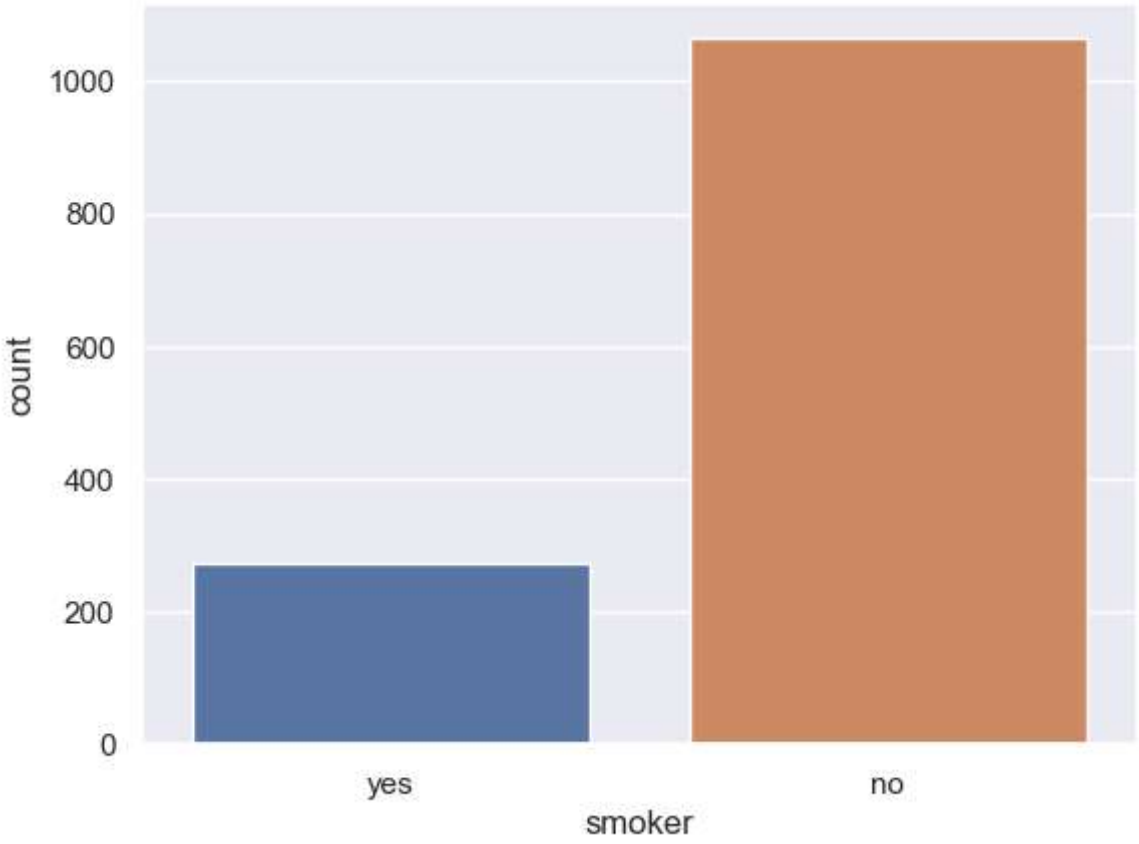
In [31]:
```
sns.countplot(x=df['sex'])
```

Out[31]:
```
<Axes: xlabel='sex', ylabel='count'>
```



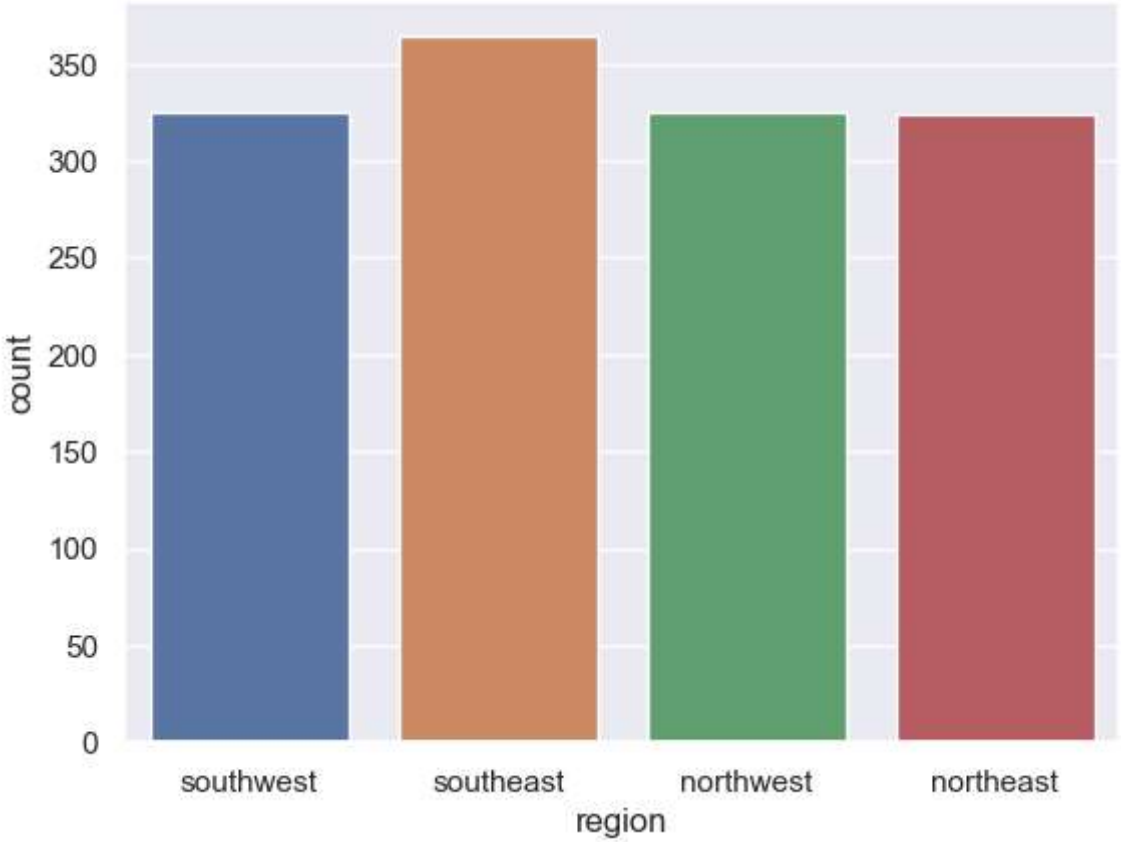
In [32]:
```
sns.countplot(x=df['smoker'])
```

Out[32]:
```
<Axes: xlabel='smoker', ylabel='count'>
```



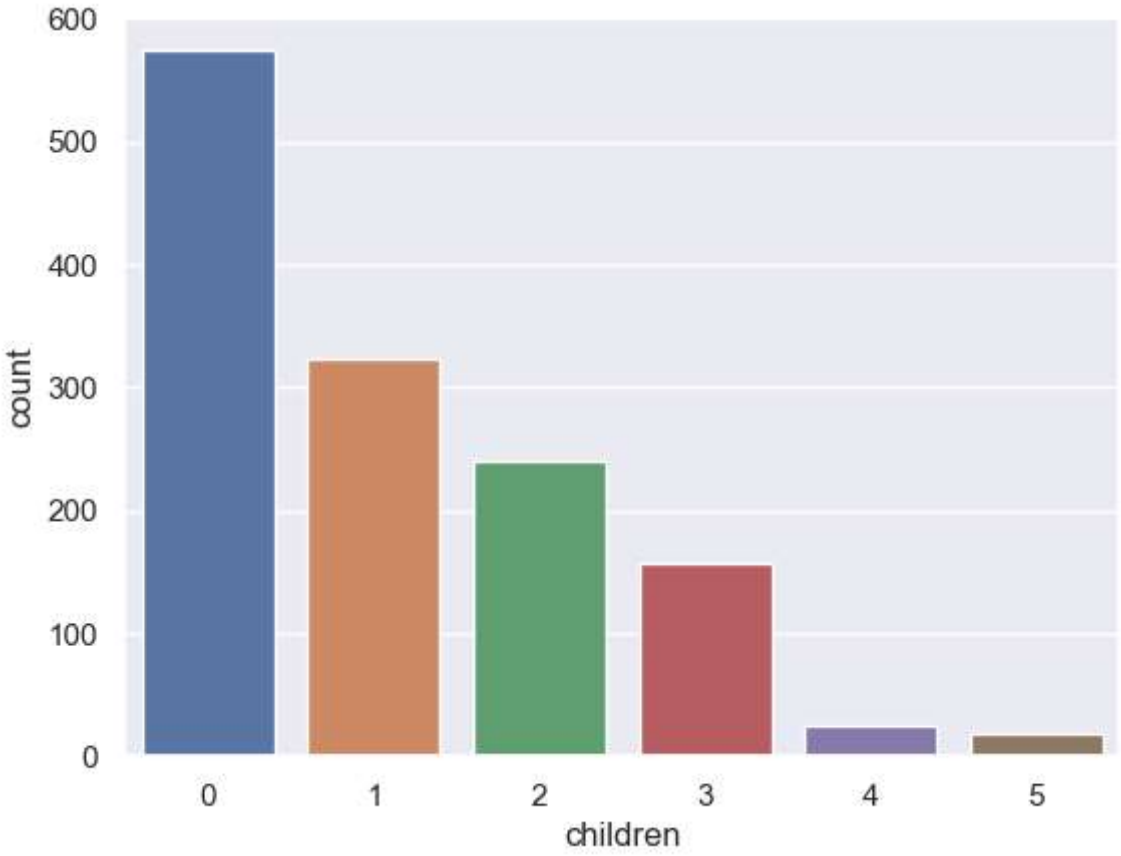
In [33]:
```
sns.countplot(x=df['region'])
```

Out[33]:
```
<Axes: xlabel='region', ylabel='count'>
```



In [34]:
```
sns.countplot(x=df['children'])
```

Out[34]:
```
<Axes: xlabel='children', ylabel='count'>
```



There are a lot more non-smokers than there are smokers in the data

Instances are distributed evenly accross all regions

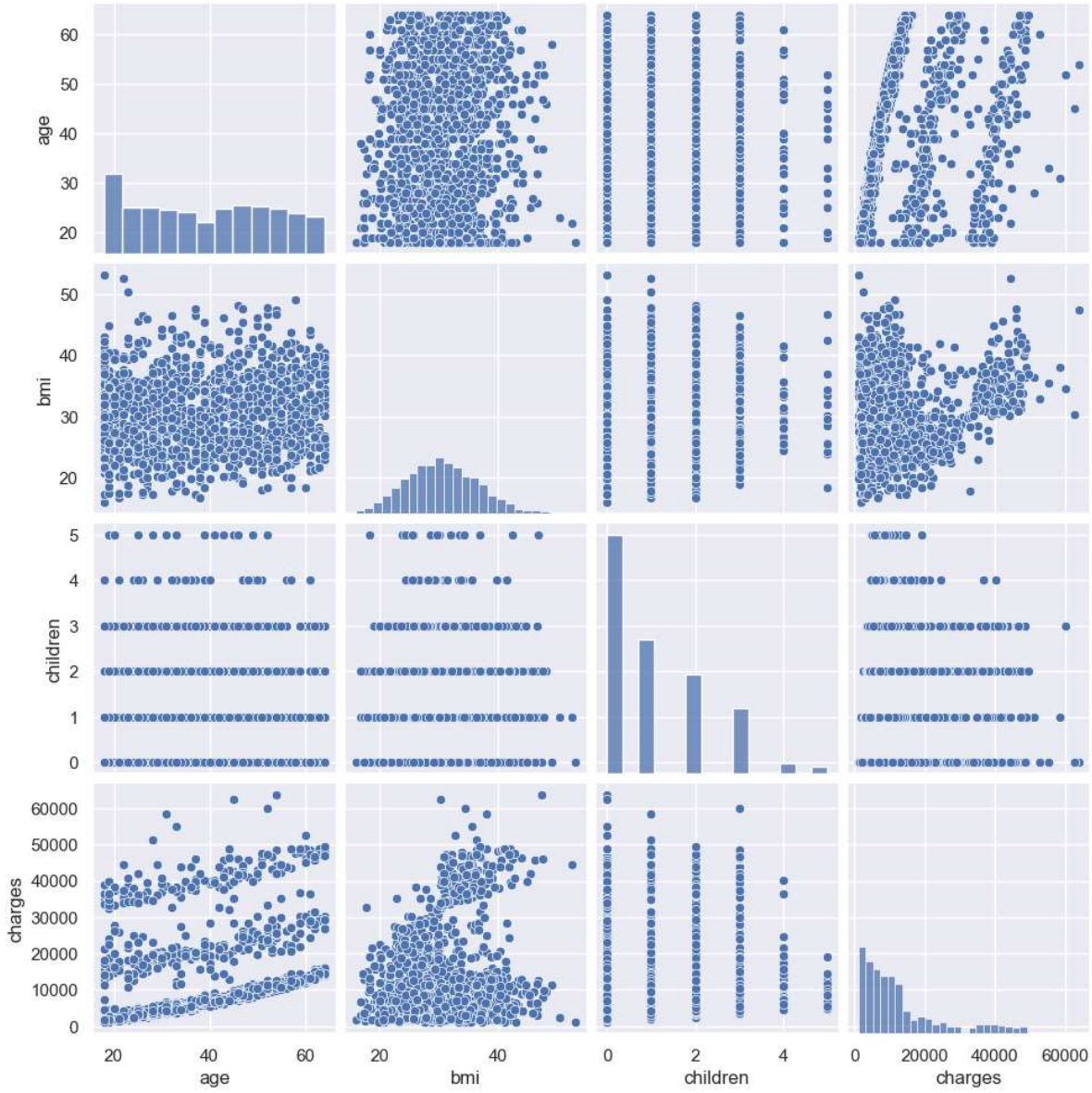
Gender is also distributed evenly

Most instances have less than 2 children and very few have 4 or 5 children

Bi-variate distribution of every possible attribute pair

In [36]:
```
sns.pairplot(df) # pairplot
plt.show()
```

```
C:\Users\PC\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning:
The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```



The only obvious correlation of 'charges' is with 'smoker'

Looks like smokers claimed more money than non-smokers

There's an interesting pattern between 'age' and 'charges. Could be because for the same ailment, older people are charged more than the younger ones

Check Correlation

To find out the correlation we will use the corr function and also we will plot a heatmap to visualise this correlation.

In [50]:
```
numeric_columns = df.select_dtypes(include=['number'])
```

In [51]:
```
corr = numeric_columns.corr()
corr
```

Out[51]:

	age	bmi	children	charges
age	1.000000	0.109272	0.042469	0.299008
bmi	0.109272	1.000000	0.012759	0.198341
children	0.042469	0.012759	1.000000	0.067998
charges	0.299008	0.198341	0.067998	1.000000

In [52]:
```
sns.heatmap(corr, annot = True)
```

Out[52]: <Axes: >

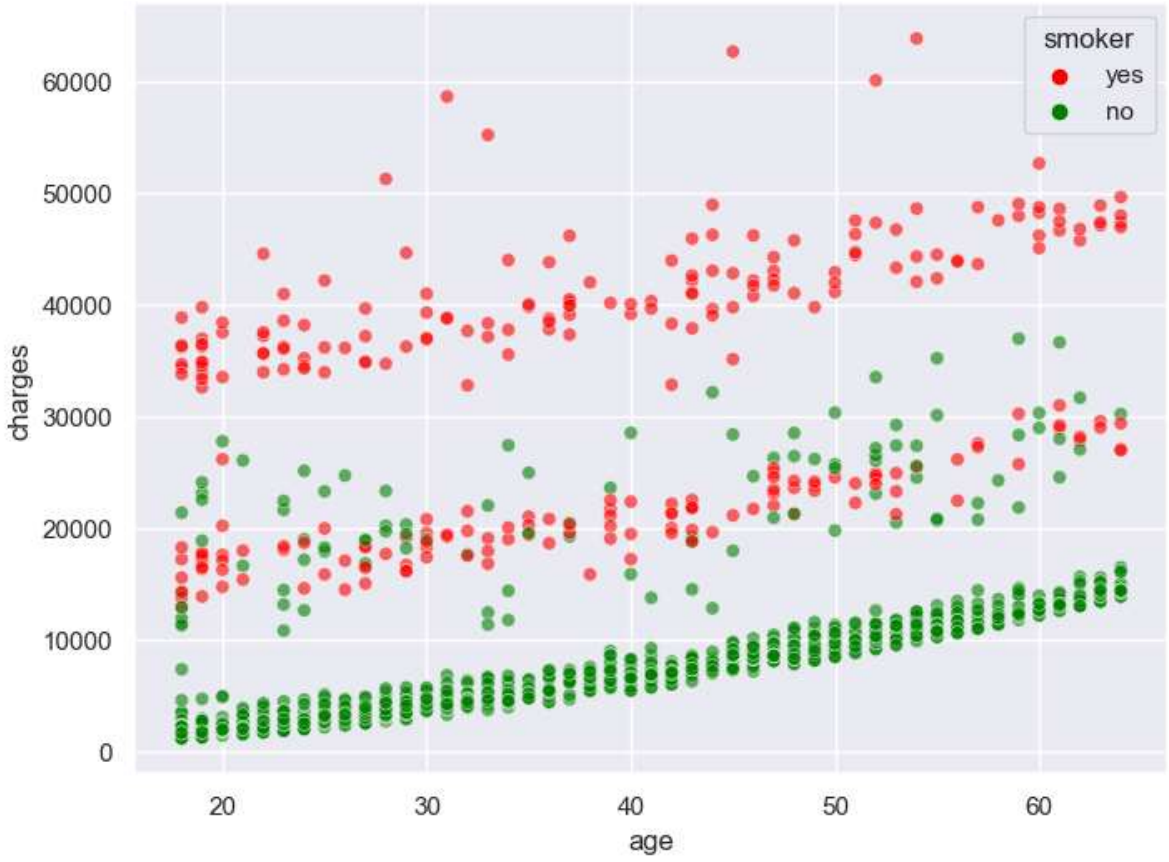
Do charges of people who smoke differ significantly from the people who don't?

In [53]:
```
df.smoker.value_counts()
```

Out[53]:
```
smoker
no     1064
yes     274
Name: count, dtype: int64
```

In [57]:
```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(8, 6))
sns.scatterplot(x='age', y='charges', hue='smoker', data=df, palette=['red', '
plt.show()
```



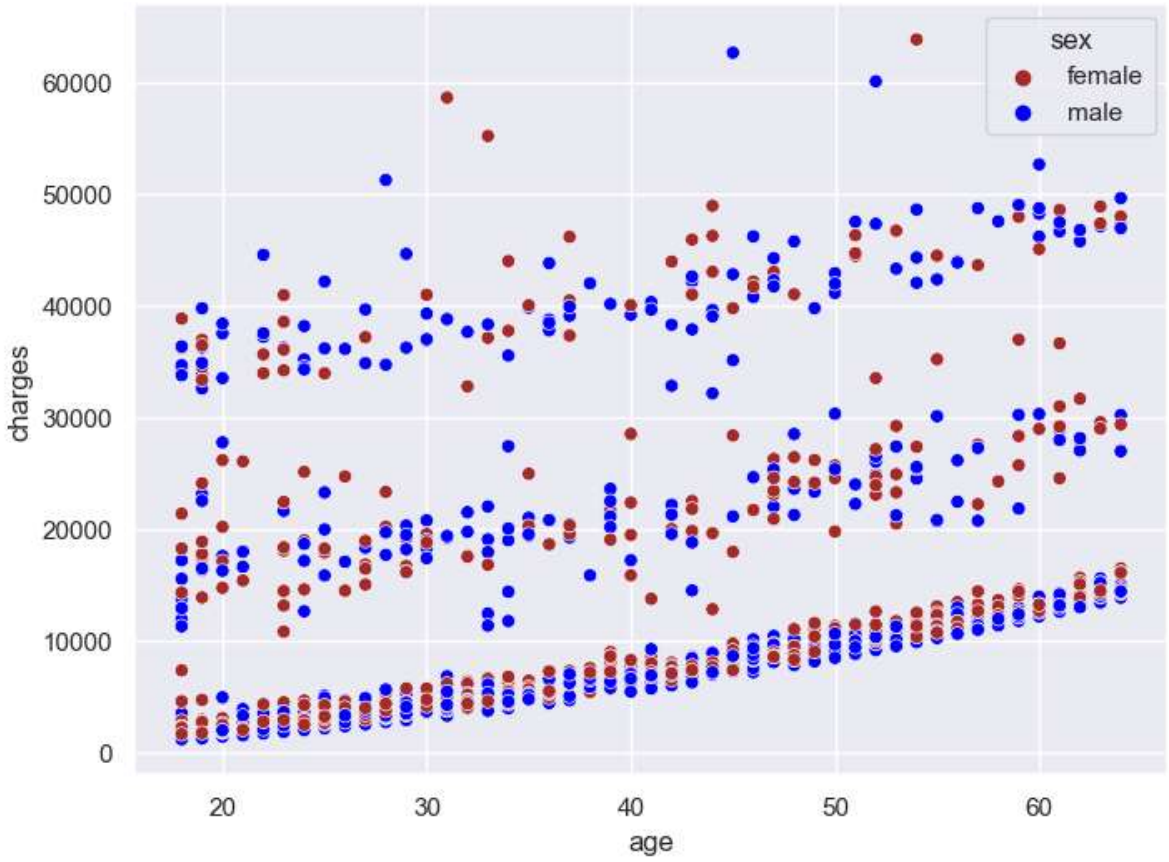
Visually the difference between charges of smokers and charges of non-smokers is apparent

Does bmi of males differ significantly from that of females?

In [58]:
```
df.sex.value_counts()   #Checking the distribution of males and females
```

Out[58]:
```
sex
male      676
female    662
Name: count, dtype: int64
```

In [62]:
```
plt.figure(figsize=(8,6))
sns.scatterplot(x='age', y = 'charges',data=df, hue = 'sex',palette= ['brown',
plt.show()
```



In [ ]:
```
```