In [2]:

```
#importing python libraries

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [20]:

```
#importing my dataset
df = pd.read_csv('insurance.csv')
df
```

Out[20]:

	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

In [39]:

```
#ordering my data to see a few observations

df1 = df.sort_values(by=['age'], ascending=False)
df1
```

Out[39]:

	age	sex	bmi	children	smoker	region	charges
335	64	male	34.500	0	no	southwest	13822.80300
603	64	female	39.050	3	no	southeast	16085.12750
752	64	male	37.905	0	no	northwest	14210.53595
1265	64	male	23.760	0	yes	southeast	26926.51440
534	64	male	40.480	0	no	southeast	13831.11520
...	...	...	...	...	...	...	...
942	18	female	40.185	0	no	northeast	2217.46915
46	18	female	38.665	2	no	northeast	3393.35635
295	18	male	22.990	0	no	northeast	1704.56810
50	18	female	35.625	0	no	northeast	2211.13075
648	18	male	28.500	0	no	northeast	1712.22700

1338 rows × 7 columns

In [41]:

```
df1.head(10)
```

Out[41]:

	age	sex	bmi	children	smoker	region	charges
335	64	male	34.500	0	no	southwest	13822.80300
603	64	female	39.050	3	no	southeast	16085.12750
752	64	male	37.905	0	no	northwest	14210.53595
1265	64	male	23.760	0	yes	southeast	26926.51440
534	64	male	40.480	0	no	southeast	13831.11520
328	64	female	33.800	1	yes	southwest	47928.03000
768	64	female	39.700	0	no	southwest	14319.03100
1241	64	male	36.960	2	yes	southeast	49577.66240
62	64	male	24.700	1	no	northwest	30166.61817
801	64	female	35.970	0	no	southeast	14313.84630

In [13]:

```
#cleaning my data using the .isna() method

df.isna().sum()
```

Out[13]:

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

In [14]:

```
df.info()
```

```
<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 [15]:

```
df.describe()
```

Out[15]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

In [16]:

```
#checking for null values to clean my data

df.isnull()
```

Out[16]:

	age	sex	bmi	children	smoker	region	charges
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...
1333	False	False	False	False	False	False	False
1334	False	False	False	False	False	False	False
1335	False	False	False	False	False	False	False
1336	False	False	False	False	False	False	False
1337	False	False	False	False	False	False	False

1338 rows × 7 columns

In [19]:

```
#checking my columns number

df.columns
```

Out[19]:

```
Index(['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges'], dtyp
e='object')
```

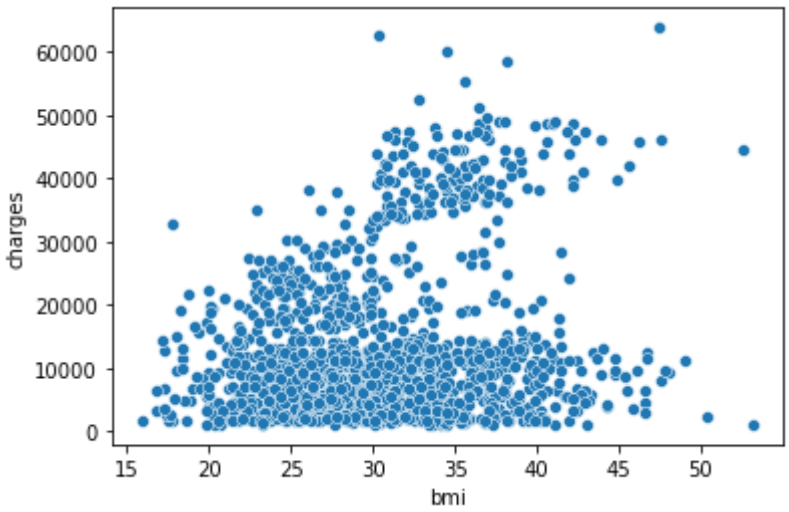
In [12]:

```
#finding the relationship between bmi and insurance charges

sns.scatterplot(x='bmi', y='charges', data=df)
```

Out[12]:

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



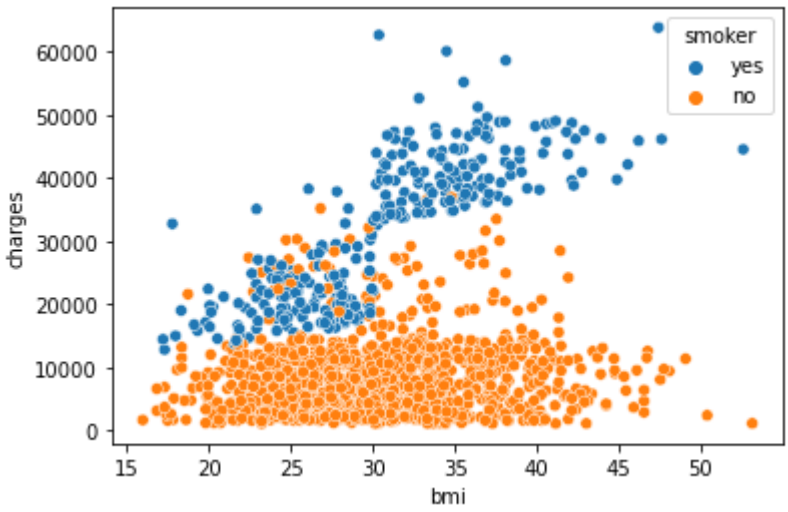
In [10]:

```
#relating to a categorical variable using color coding
# to see how it affects insurance charges

sns.scatterplot(x=df['bmi'], y=df['charges'],
                hue=df['smoker'])
```

Out[10]:

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



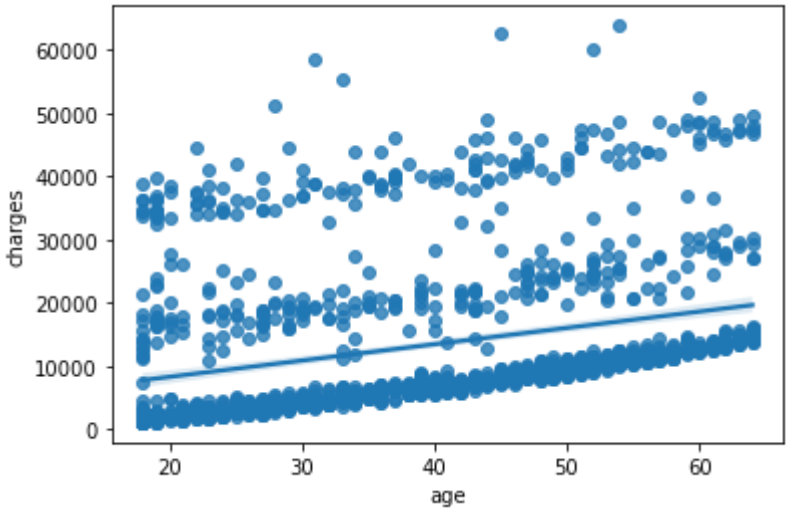
In [43]:

```
#adding a regression line to see the linear r/ship
#between age and insurance charges

sns.regplot(x=df['age'], y=df['charges'])
```

Out[43]:

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



Insight from the chart

This plot shows that the charges tend to increase with increase in age,

the regression line indicates that there's a direct relationship between

age of client and likely insurance charges.

There's equally a possibility that insurance charges could also depend on a list of other factors such as

bmi and if the client smokes or not.

Secondly, a few outliers could be sighted in the data because of their

deviation from other data points.