In [1]:
```
import pandas as pd
pd.plotting.register_matplotlib_converters()
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import numpy as np
print("Setup Complete")
```

```
Setup Complete
```

In [2]:
```
#running my data

dt = pd.read_csv('insurance.csv')
dt
```

Out[2]:

	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 [3]:
```
#checking out my dataset

dt.head(5)
```

Out[3]:

	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

In [4]:
```
dt.tail(5)
```

Out[4]:

	age	sex	bmi	children	smoker	region	charges
1333	50	male	30.97	3	no	northwest	10600.5483
1334	18	female	31.92	0	no	northeast	2205.9808
1335	18	female	36.85	0	no	southeast	1629.8335
1336	21	female	25.80	0	no	southwest	2007.9450
1337	61	female	29.07	0	yes	northwest	29141.3603

In [5]:
```
#getting a random sample of my data set
dt.sample(n = 5)
```

Out[5]:

	age	sex	bmi	children	smoker	region	charges
971	34	female	23.56	0	no	northeast	4992.3764
297	47	male	25.41	1	yes	southeast	21978.6769
870	50	male	36.20	0	no	southwest	8457.8180
189	29	female	32.11	2	no	northwest	4922.9159
13	56	female	39.82	0	no	southeast	11090.7178

In [6]:
```
dt.charges.unique()
```

Out[6]:
```
array([16884.924 ,  1725.5523,  4449.462 , ...,  1629.8335,  2007.945 ,
       29141.3603])
```

In [7]:
```
# checking for missing values in my data set

dt.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 [8]:
```
dt.isna().sum()
```

Out[8]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

In [9]:
```
#setting my data set columns to_list

dt.columns.to_list()
```

Out[9]:
```
['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges']
```

In [10]:
```
dt.charges.mean()
```

Out[10]:
```
13270.422265141257
```

In [11]:
```
dt.charges.median()
```

Out[11]:
```
9382.033
```

In [12]:
```
dt.charges.mode()
```

Out[12]:
```
0    1639.5631
Name: charges, dtype: float64
```

In [15]:
```
dt.charges.max()
```

Out[15]:
```
63770.42801
```

In [13]:
```
dt.bmi.max()
```

Out[13]:
```
53.13
```

In [16]:
```
dt.charges.min()
```

Out[16]:
```
1121.8739
```

In [14]:
```
dt.bmi.min()
```

Out[14]:
```
15.96
```

In [17]:
```
#describing my data set

dt.describe()
```

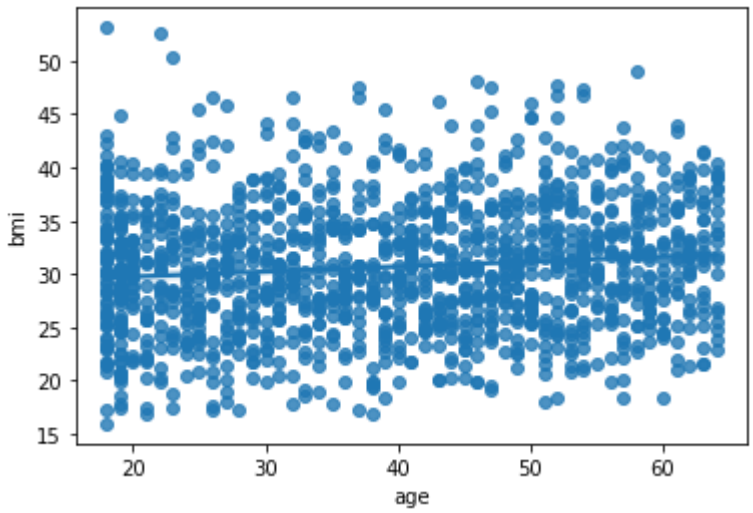
Out[17]:

	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

Question

- check the correlation between bmi and age. i.e
  - does the increase in age cause the increase in bmi('body mass index')
- check the relationship between charges and bmi. i.e
  - does the increase in bmi('body mass index') cause the increase in charges?
- check the relationship between region and bmi. i.e -does the region affect the charges to be paid

In [18]:
```
#using the scatter plot to check the correlation between individual age and their
ans = sns.regplot(x = 'age', y = 'bmi', data = dt)
```



- the above graph shows a non-correlation relationship between age and bmi as increase in bmi is not dependent on age.

Type *Markdown* and LaTeX: α^2

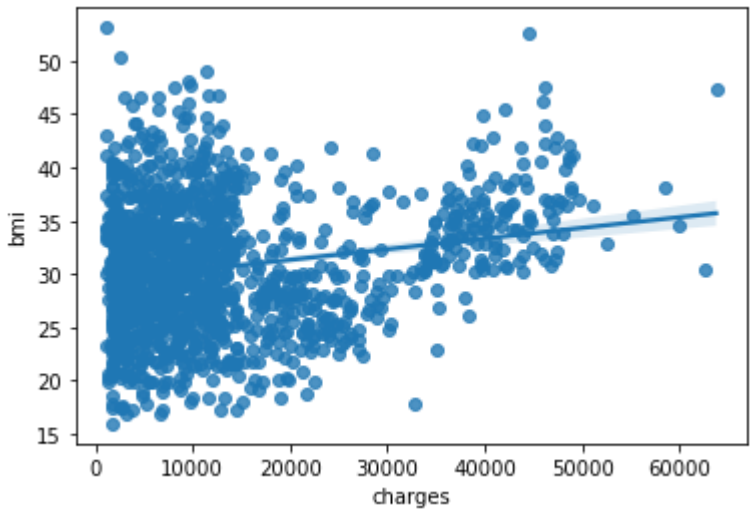
In [19]:
```
cols = ['charges', 'bmi', 'sex']
new_dt = dt[cols]
new_dt
```

Out[19]:

	charges	bmi	sex
0	16884.92400	27.900	female
1	1725.55230	33.770	male
2	4449.46200	33.000	male
3	21984.47061	22.705	male
4	3866.85520	28.880	male
...	...	...	...
1333	10600.54830	30.970	male
1334	2205.98080	31.920	female
1335	1629.83350	36.850	female
1336	2007.94500	25.800	female
1337	29141.36030	29.070	female

1338 rows × 3 columns

In [20]:
```
#using the scatter plot to check the correlation between individual bmi and their
ans = sns.regplot(x = 'charges', y = 'bmi', data = dt)
```

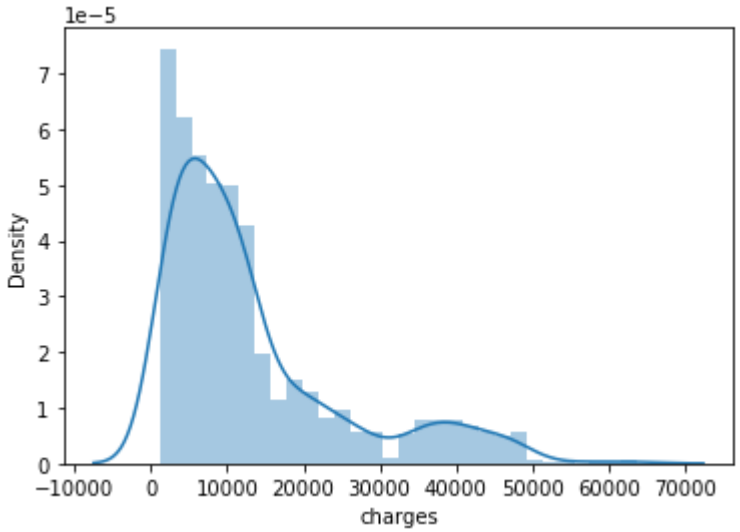


- the above graph shows a non-linear relationship between charges and bmi as increase in bmi causes a slight increase in charges.

In [21]:
```
#using the histogram plot to check for correlation between
sns.distplot(dt.charges.dropna(axis=0))
```

```
C:\Users\user\anaconda3\lib\site-packages\seaborn\distributions.py:2619: Future
Warning: `distplot` is a deprecated function and will be removed in a future ve
rsion. Please adapt your code to use either `displot` (a figure-level function
with similar flexibility) or `histplot` (an axes-level function for histogram
s).
  warnings.warn(msg, FutureWarning)
```

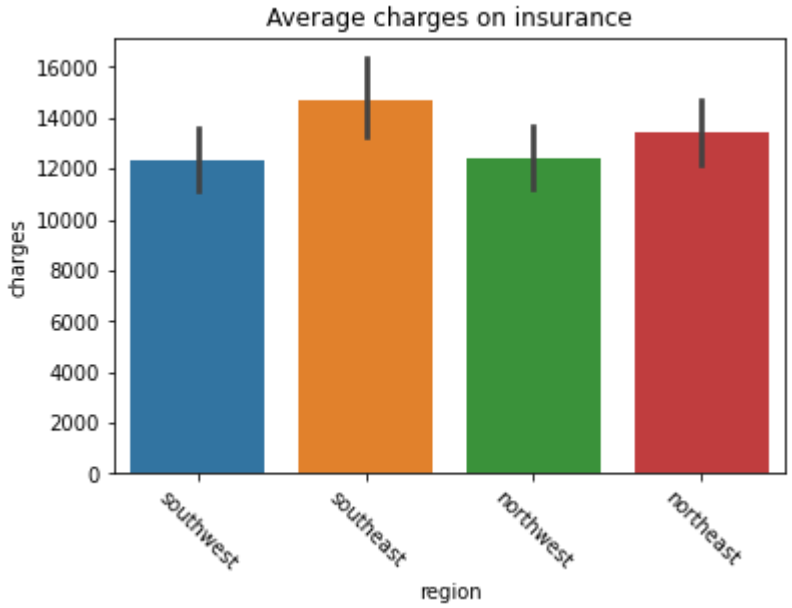
Out[21]: <AxesSubplot:xlabel='charges', ylabel='Density'>



- chart shows skewness to the right side

In [22]:
```
p = sns.barplot(x ='region', y = 'charges',
                data = dt, estimator = np.mean);
p.set(title = 'Average charges on insurance')
p.set_xticklabels(p.get_xticklabels(), rotation = -45)
```

Out[22]:
```
[Text(0, 0, 'southwest'),
 Text(1, 0, 'southeast'),
 Text(2, 0, 'northwest'),
 Text(3, 0, 'northeast')]
```



- the above chart shows that the location of an insurer affects the charges giving as
- 'Northwest' has a smaller percent of insurer with lower charges with respect to their bmi. while
- 'Southeast' has a higher percentage of insurer with the highest charges with respect to their bmi.

In [23]:
```
# lets look at bmi greater than 30

large_bmi = dt[dt['bmi']>30]
large_bmi
```

Out[23]:

	age	sex	bmi	children	smoker	region	charges
1	18	male	33.77	1	no	southeast	1725.55230
2	28	male	33.00	3	no	southeast	4449.46200
6	46	female	33.44	1	no	southeast	8240.58960
12	23	male	34.40	0	no	southwest	1826.84300
13	56	female	39.82	0	no	southeast	11090.71780
...	...	...	...	...	...	...	...
1331	23	female	33.40	0	no	southwest	10795.93733
1332	52	female	44.70	3	no	southwest	11411.68500
1333	50	male	30.97	3	no	northwest	10600.54830
1334	18	female	31.92	0	no	northeast	2205.98080
1335	18	female	36.85	0	no	southeast	1629.83350

705 rows × 7 columns

In [24]:
```
sns.displot(large_bmi.charges)
```

Out[24]:
```
<seaborn.axisgrid.FacetGrid at 0x1f5c3cabbb0>
```