T.E DITHATE

Insurance Exploratory Data Analysis

In [1]:

```
#IMPORTING LIBRARIES
import pandas as pd
import numpy as np
import matplotlib.pylab as plt
import seaborn as sns
```

In [2]:

```
#LOADING THE DATASET
df = pd.read_csv(r"C:\Users\Goitsimodimo Madue\Downloads\insurance.csv")
df.head()
```

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

In [3]:

```
#THE FOLLOWING DISPLAYS THE SHAPE OF THE DATASET
df.shape
```

Out[3]:

```
(1338, 7)
```

In [4]:

```
#THE FOLLOWING SHOWS THE DATATYPES OF THE COLOUMN
df.dtypes
```

Out[4]:

```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

In [5]:

```
#CHECKING MISSING VALUES
missing_data = df.isnull()
missing_data.head()
```

Out[5]:

	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

In [6]:

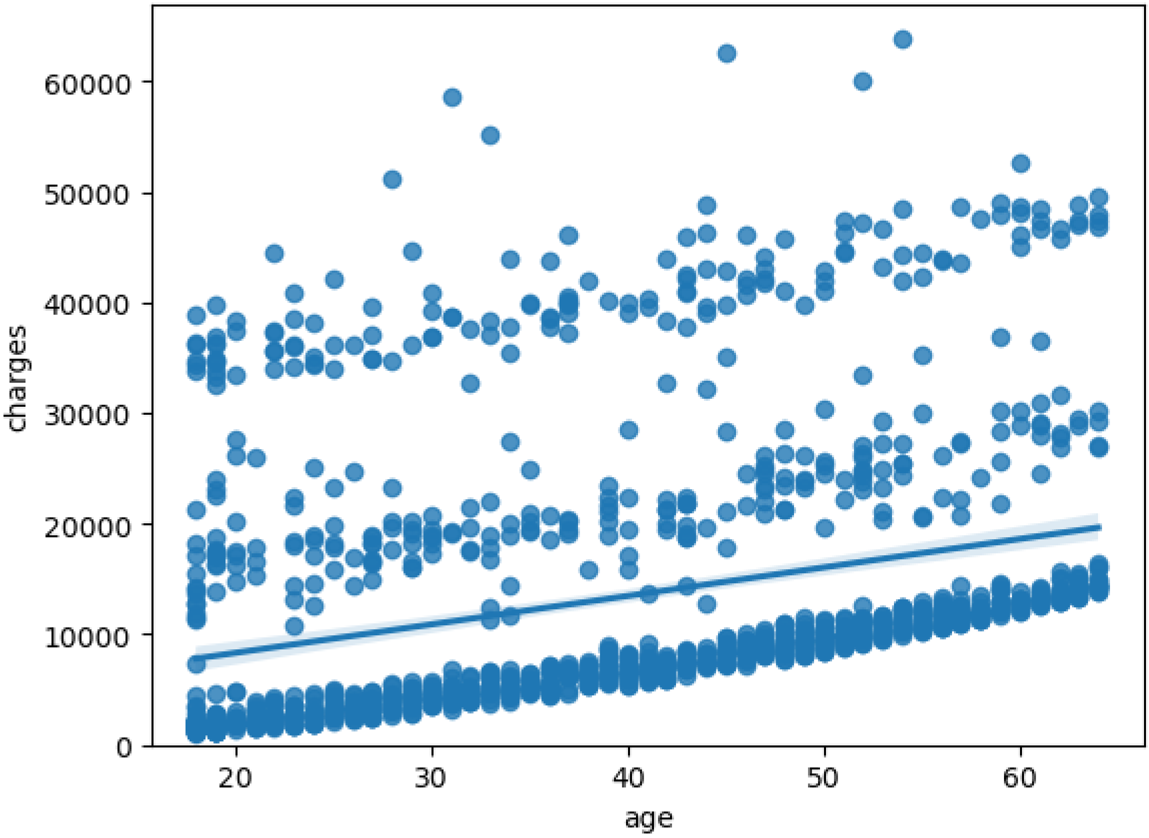
```
for column in missing_data.columns.values.tolist():
    print(column)
    print (missing_data[column].value_counts())
    print("")
```

```
age
False    1338
Name: age, dtype: int64

sex
False    1338
Name: sex, dtype: int64

bmi
False    1338
Name: bmi, dtype: int64

children
False    1338
Name: children, dtype: int64

smoker
False    1338
Name: smoker, dtype: int64

region
False    1338
Name: region, dtype: int64

charges
False    1338
Name: charges, dtype: int64
```

There is no missing data in any column.

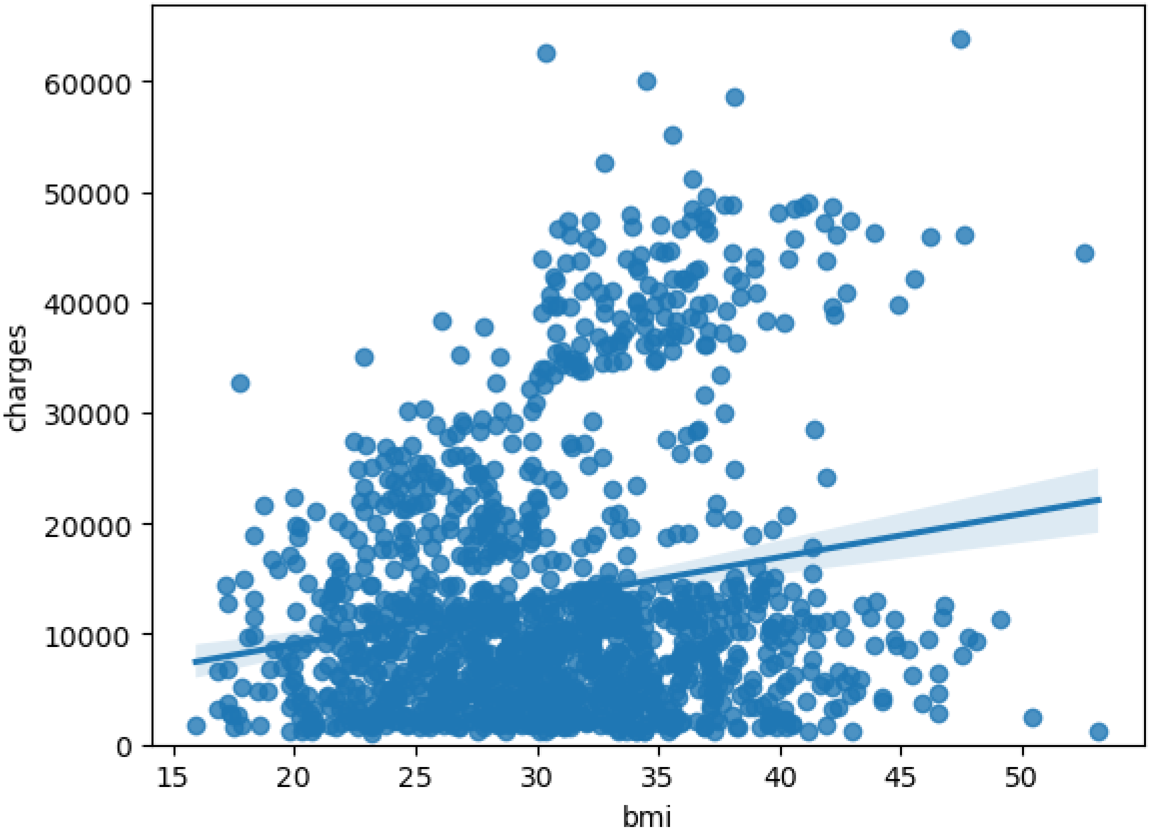
In [7]:

```
# age as potential predictor variable of charges
sns.regplot(x="age", y="charges", data=df)
plt.ylim(0,)
plt.show()
```



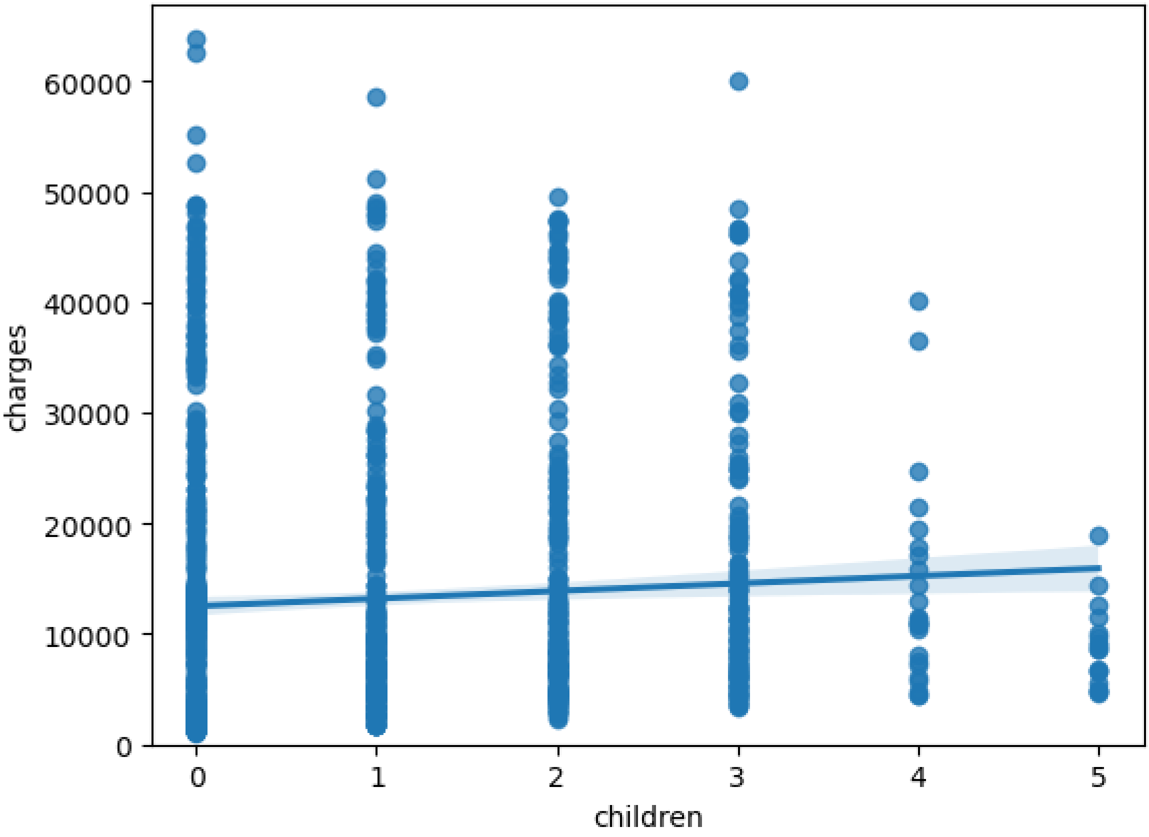
In [8]:

```
# bmi as potential predictor variable of charges
sns.regplot(x="bmi", y="charges", data=df)
plt.ylim(0,)
plt.show()
```

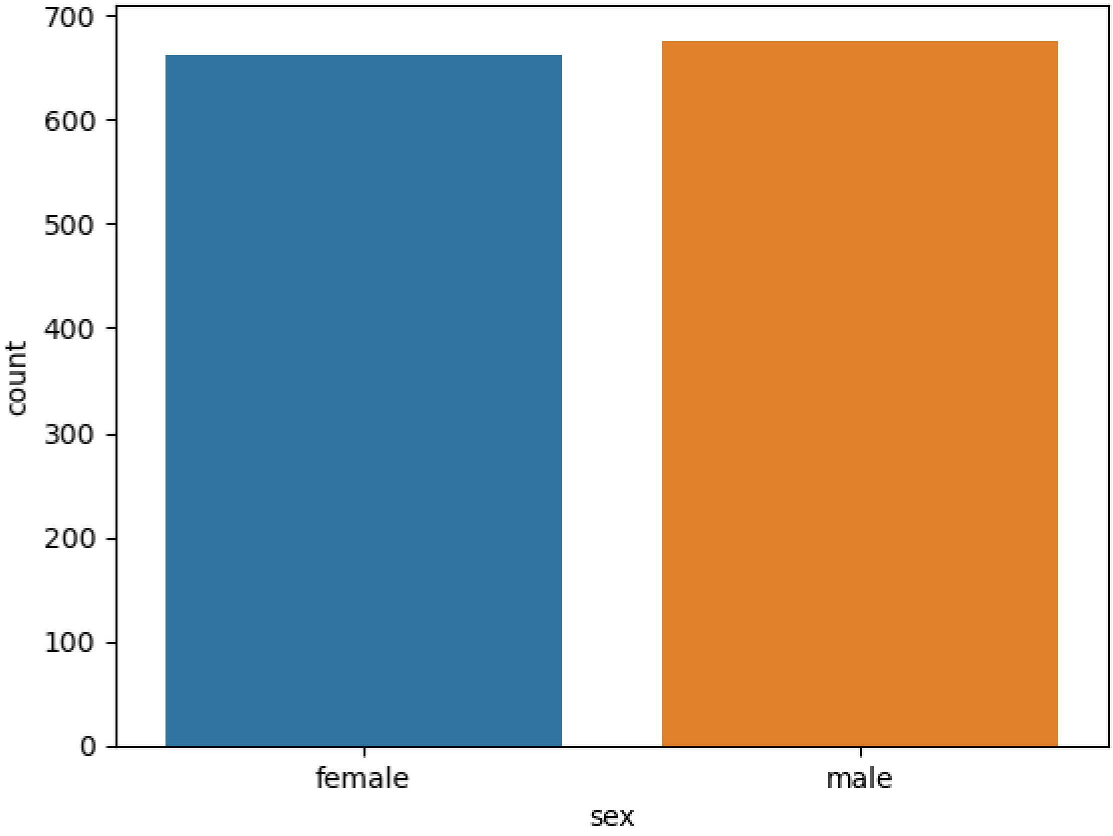


In [9]:

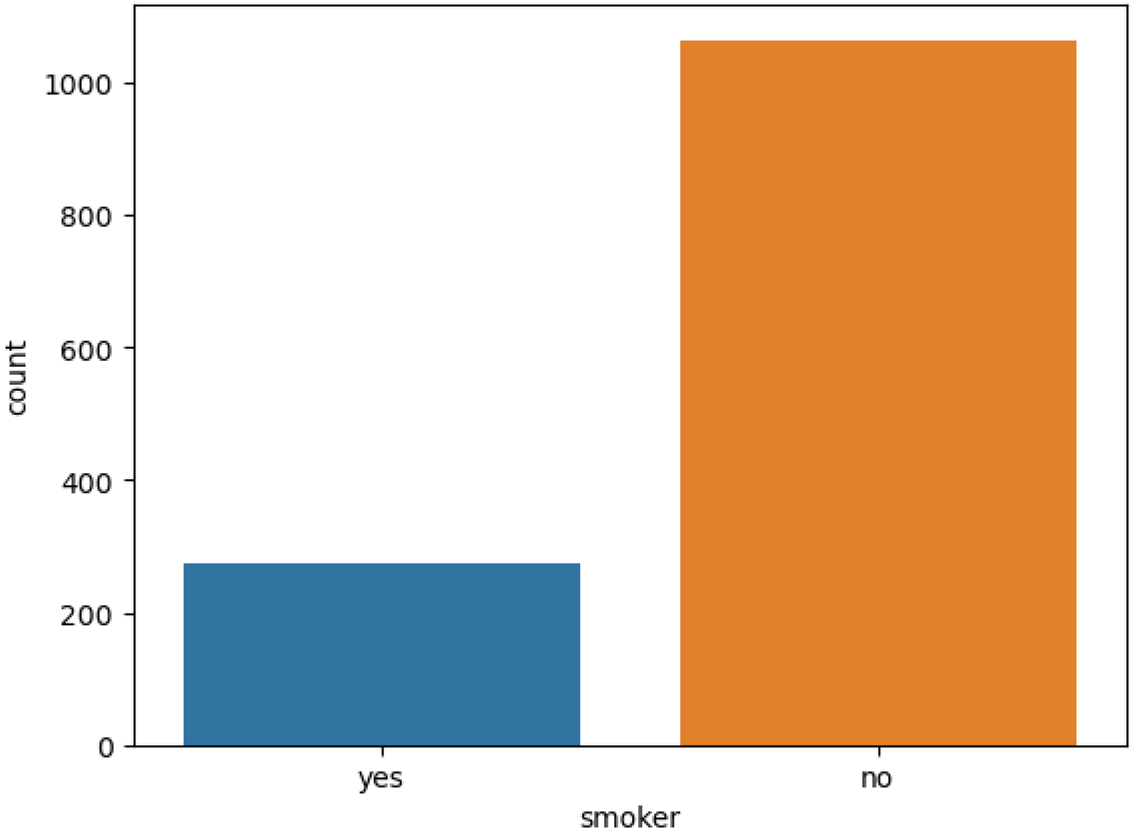
```
# number of childern as potential predictor variable of charges
sns.regplot(x="children", y="charges", data=df)
plt.ylim(0,)
plt.show()
```



In [10]:
```
sns.countplot(x='sex', data=df)
plt.show()
```

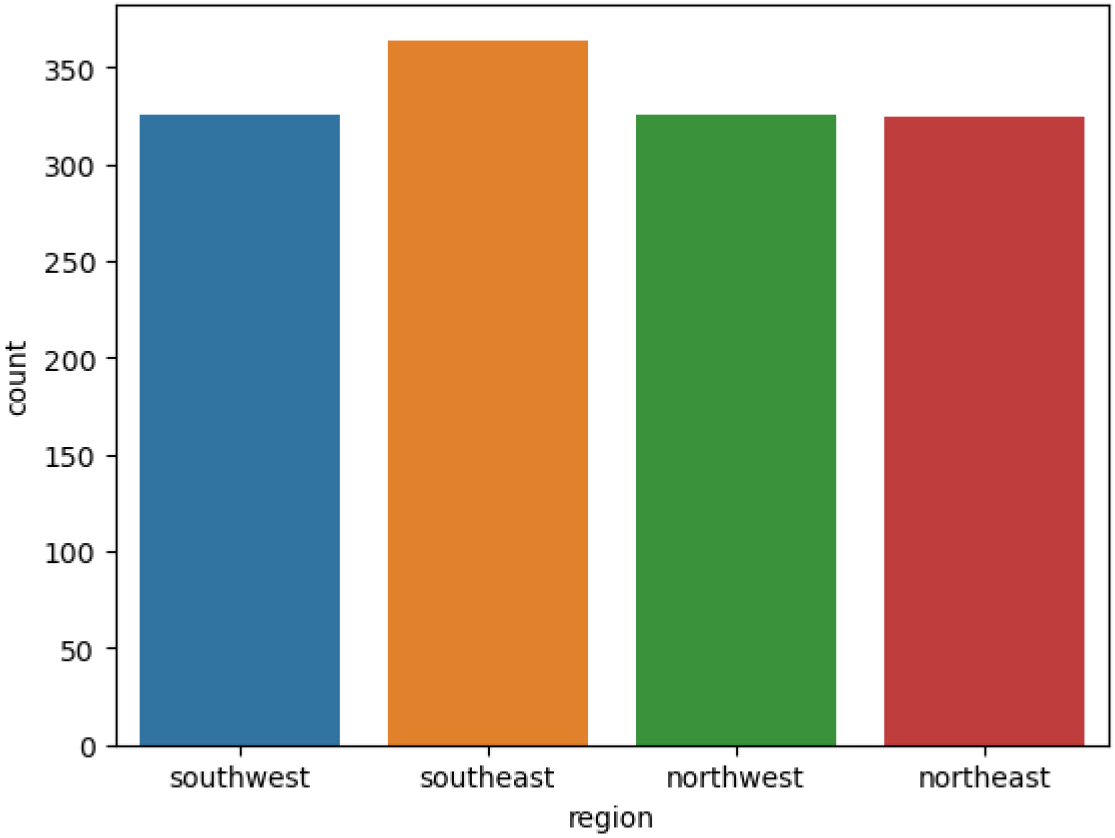


In [11]:
```
sns.countplot(x='smoker', data=df)
plt.show()
```



In [12]:

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



In [ ]:

```
df.head()
```

In [ ]:

```
```