# Insurance_project

June 16, 2024

[17]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

data = pd.read_csv("insurance.csv")
df = pd.DataFrame(data)
df.head()
df.info()
```

[17]:

| | 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 |

[11]:
```
df.isna().sum()##null value check
```

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

[ ]:
```
##no null values
```

[12]:
```
df.corr()
```

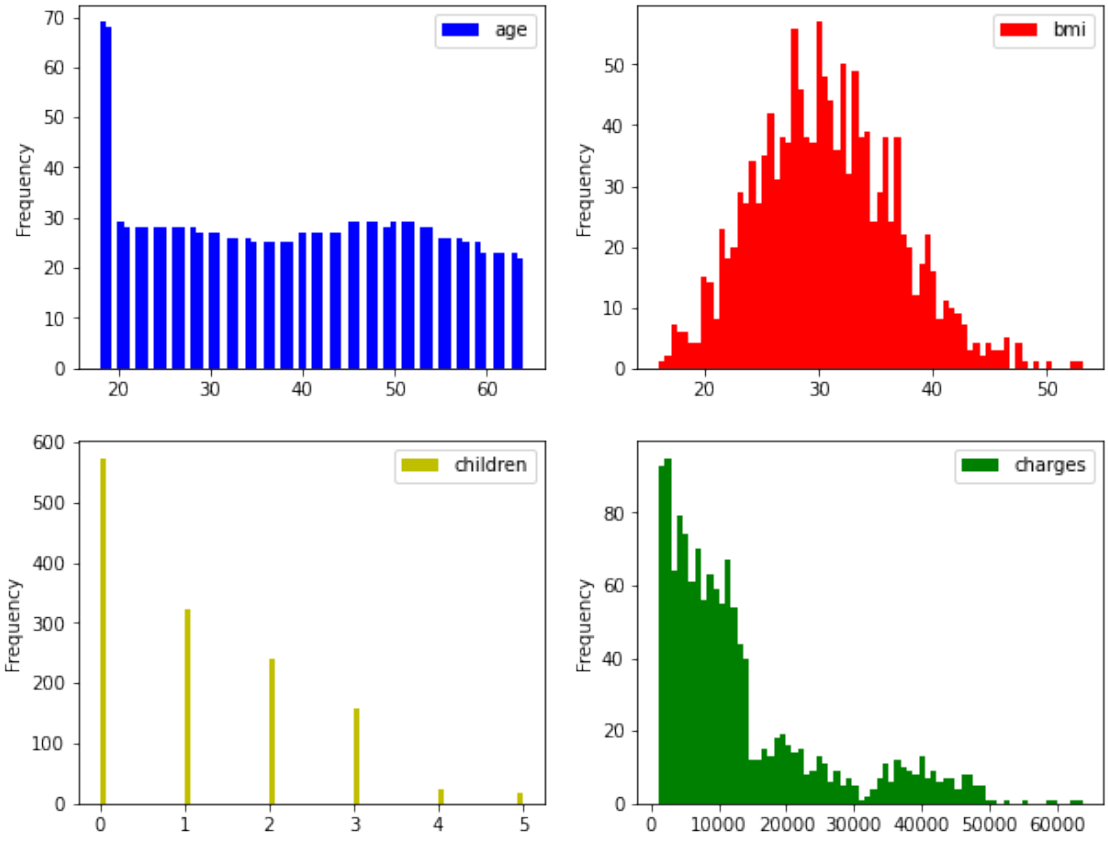
```
/tmp/ipykernel_154/1532097083.py:1: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it will
default to False. Select only valid columns or specify the value of numeric_only
to silence this warning.
  df.corr() ## age and charges have strongest relationship followed by bmi and
charges
```

[12]:

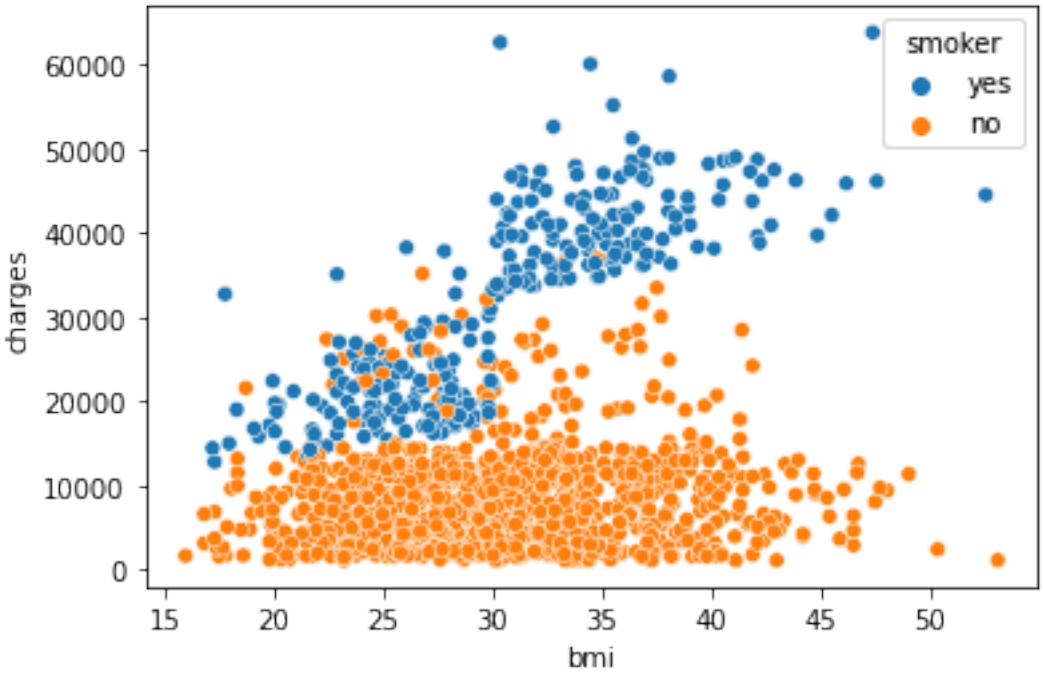
| | age | bmi | children | charges |
|---|---|---|---|---|
| age | 1.000000 | 0.109272 | 0.042469 | 0.299008 |
| bmi | 0.109272 | 1.000000 | 0.012759 | 0.198341 |
| children | 0.042469 | 0.012759 | 1.000000 | 0.067998 |
| charges | 0.299008 | 0.198341 | 0.067998 | 1.000000 |

[ ]:
```
## age and charges have strongest relationship followed by bmi and charges
```

[14]:
```
fig,axes = plt.subplots(nrows=2,ncols=2,figsize=(10,8))
df.plot(kind='hist', y = 'age', bins = 70, color = 'b', ax=axes[0][0])
df.plot(kind='hist', y = 'bmi', bins = 70, color = 'r', ax=axes[0][1])
df.plot(kind='hist', y = 'children', bins = 70, color = 'y', ax=axes[1][0])
df.plot(kind='hist', y = 'charges', bins = 70, color = 'g', ax=axes[1][1])
plt.show()
```

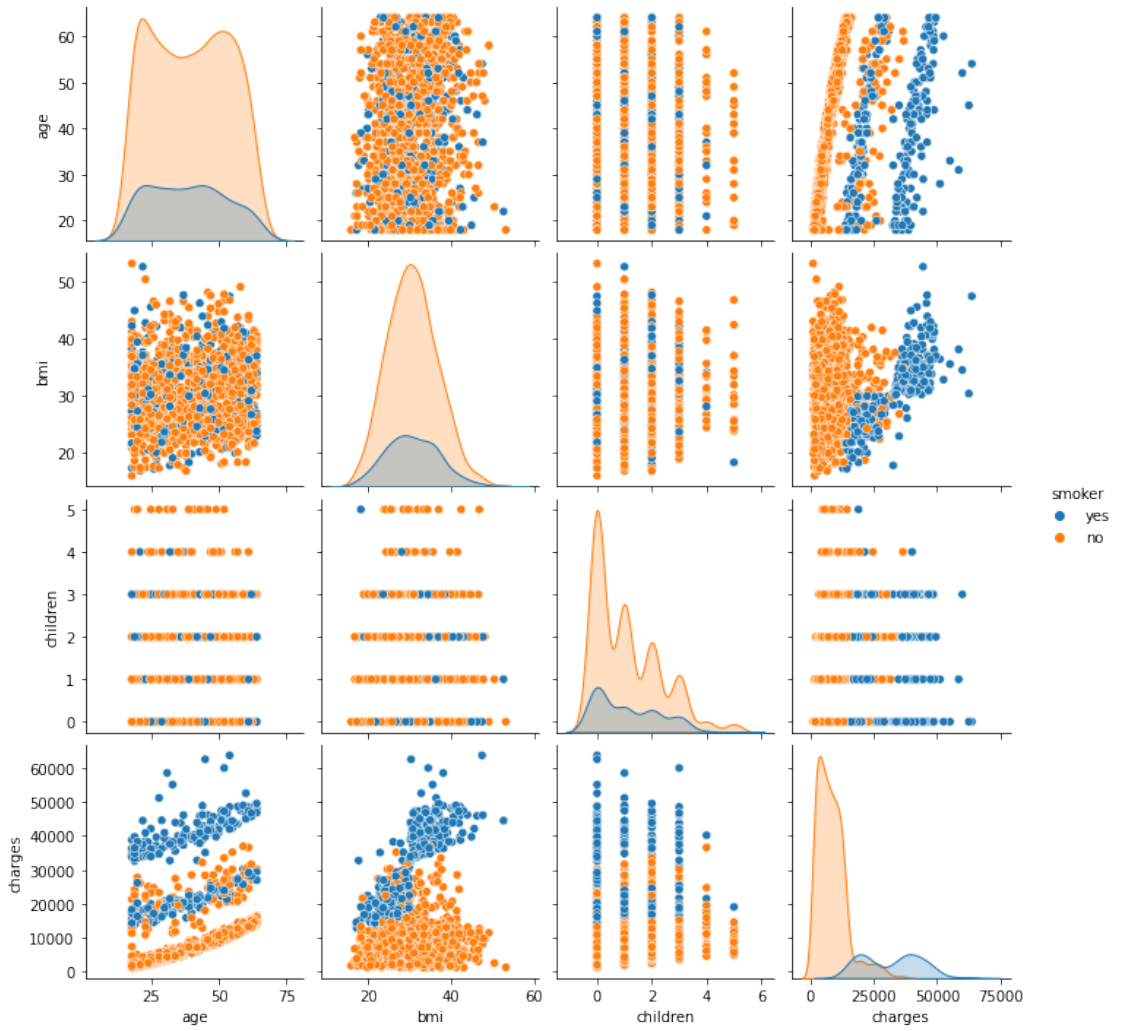



[19]:
```
##df.plot(kind='scatter',x='bmi',y='charges')
sns.scatterplot(data=df,x='bmi',y='charges',hue='smoker')
plt.show()
```



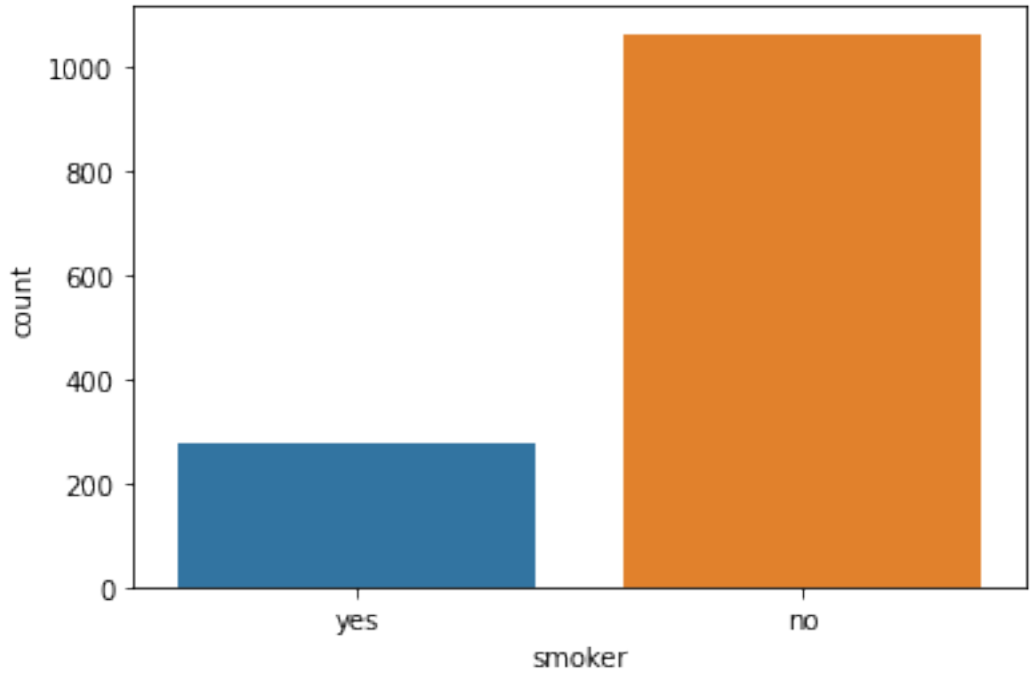

[ ]:
```
## indicates premium charges of smokers increases with bmi
```

[40]:
```
sns.pairplot(df,hue='smoker')
plt.show()
```

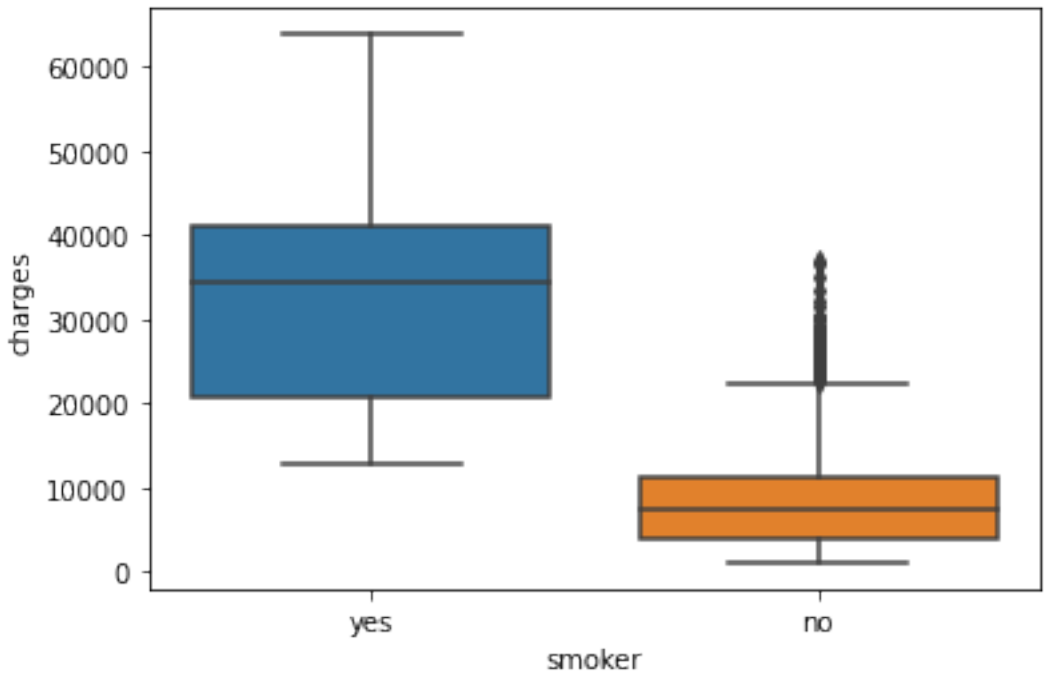



[ ]:
```
## indicates age and bmi contribute to increases premium charges for smokers
↪than non smokers
```

[31]:
```
pd.get_dummies(data=df,columns=['sex','smoker','region'])
sns.countplot(x=df['smoker'])
plt.show()
```

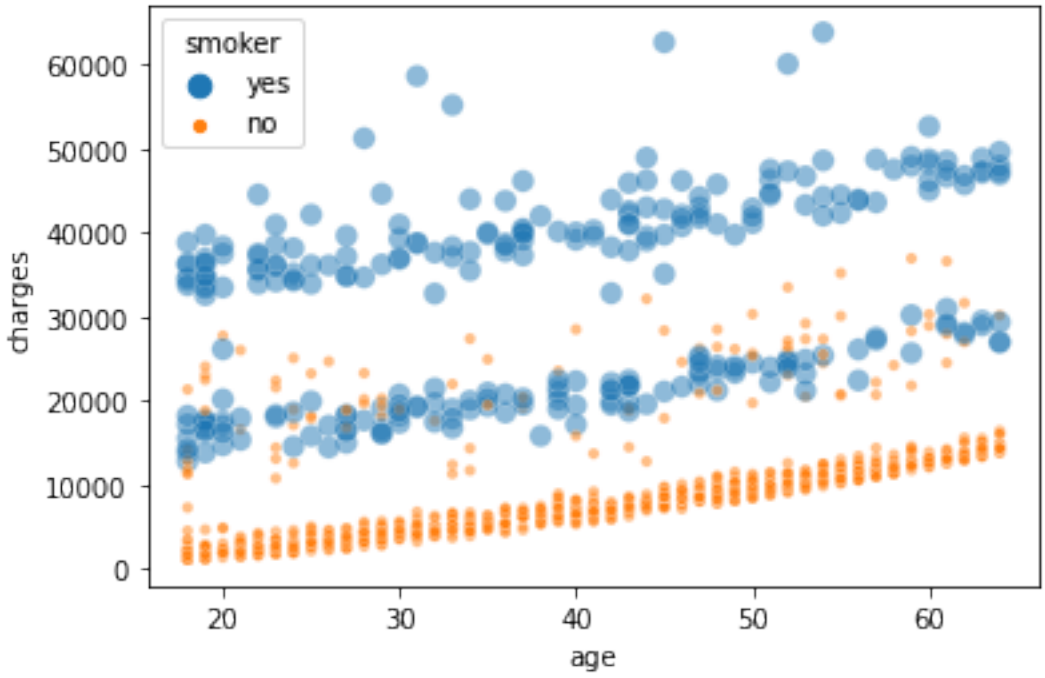



[29]:
```
sns.boxplot(x='smoker',y='charges',data=df)
plt.show()
```



[ ]:
```
## Above data clearly represents that the premium charges for smokers is␣
↪greater than non-smokers
```

[37]:
```
sns.scatterplot(data=df,x='age',y='charges',size='smoker',hue='smoker',alpha=0.
↪5)
plt.show()
```




[ ]:
```
## Above data clearly represents that the premium charges increases with age␣
↪and charges for smokers is greater than non-smokers
```