In [2]:
```
import pandas as pd
```

In [123]:
```
df=pd.read_csv('scaler_apollo_hospitals.csv')
```

In [124]:
```
df.head()
```

Out[124]:

	Unnamed: 0	age	sex	smoker	region	viral load	severity level	hospitalization charges
0	0	19	female	yes	southwest	9.30	0	42212
1	1	18	male	no	southeast	11.26	1	4314
2	2	28	male	no	southeast	11.00	3	11124
3	3	33	male	no	northwest	7.57	0	54961
4	4	32	male	no	northwest	9.63	0	9667

In [11]:
```
df.dtypes
```

Out[11]:
```
Unnamed: 0                  int64
age                         int64
sex                        object
smoker                     object
region                     object
viral load                float64
severity level              int64
hospitalization charges     int64
dtype: object
```

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
Unnamed: 0                  0
age                         0
sex                         0
smoker                      0
region                      0
viral load                  0
severity level              0
hospitalization charges     0
dtype: int64
```

In [14]:
```
df.describe()
```

Out[14]:

	Unnamed: 0	age	viral load	severity level	hospitalization charges
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	668.500000	39.207025	10.221233	1.094918	33176.058296
std	386.391641	14.049960	2.032796	1.205493	30275.029296
min	0.000000	18.000000	5.320000	0.000000	2805.000000
25%	334.250000	27.000000	8.762500	0.000000	11851.000000
50%	668.500000	39.000000	10.130000	1.000000	23455.000000
75%	1002.750000	51.000000	11.567500	2.000000	41599.500000
max	1337.000000	64.000000	17.710000	5.000000	159426.000000

In [15]:
```
df.shape
```

Out[15]:
```
(1338, 8)
```

In [20]:
```
df.sex.value_counts(sort=True)
```

Out[20]:
```
male      676
female    662
Name: sex, dtype: int64
```

In [21]:
```
df.smoker.value_counts(sort=True)
```

Out[21]:
```
no     1064
yes     274
Name: smoker, dtype: int64
```

In [22]:
```
df.region.value_counts(sort=True)
```

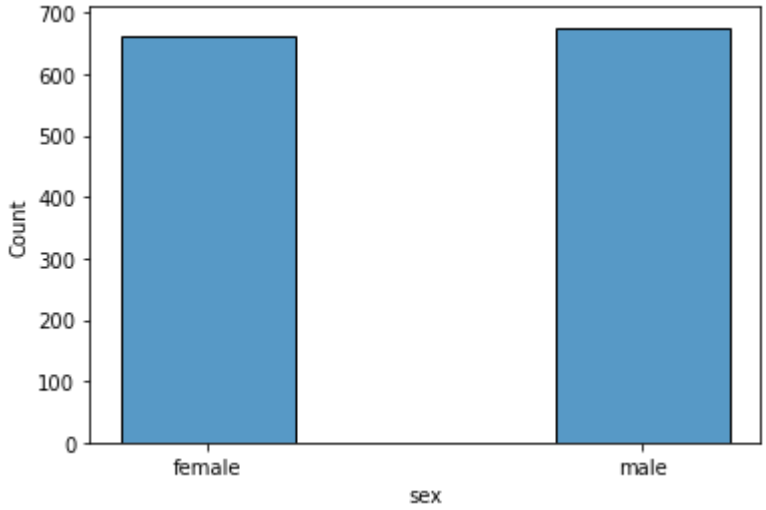
Out[22]:
```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

In [25]:
```
df['severity level'].value_counts(sort=True)
```

Out[25]:
```
0    574
1    324
2    240
3    157
4     25
5     18
Name: severity level, dtype: int64
```

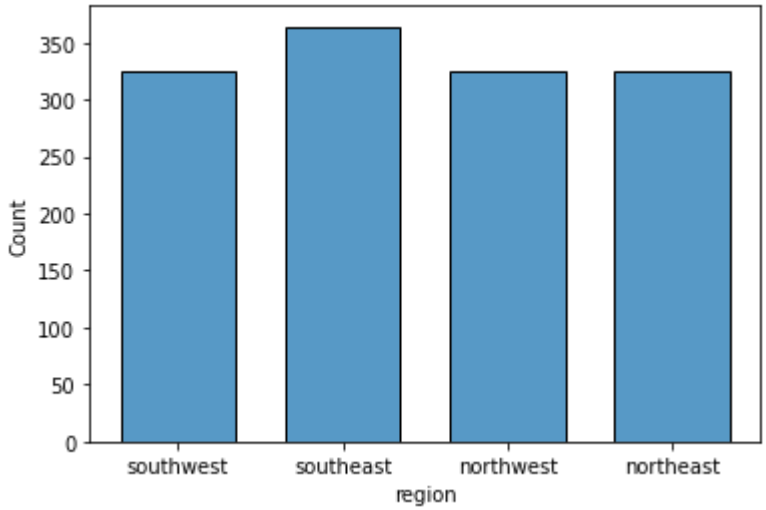
In [47]:
```
sns.histplot(data=df,

    x='sex', shrink=.4)
    #y=None)
```

Out[47]: <AxesSubplot:xlabel='sex', ylabel='Count'>



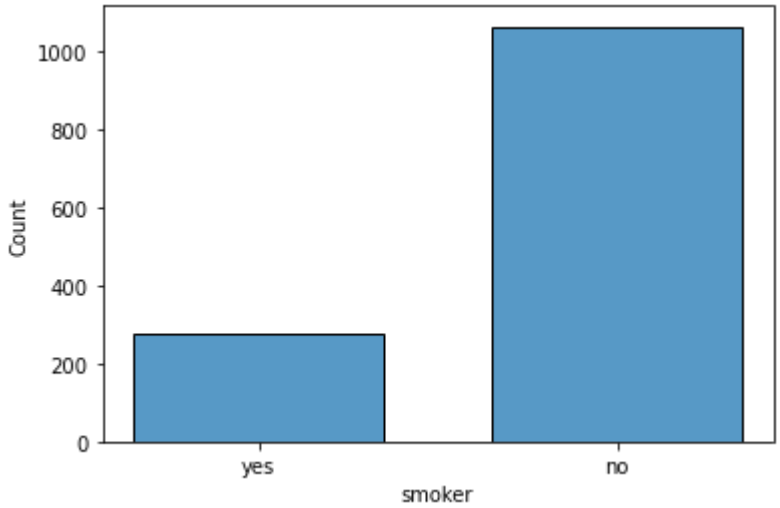
In [48]:
```
sns.histplot(data=df, x="region", shrink=.7, )
```

Out[48]:
```
<AxesSubplot:xlabel='region', ylabel='Count'>
```



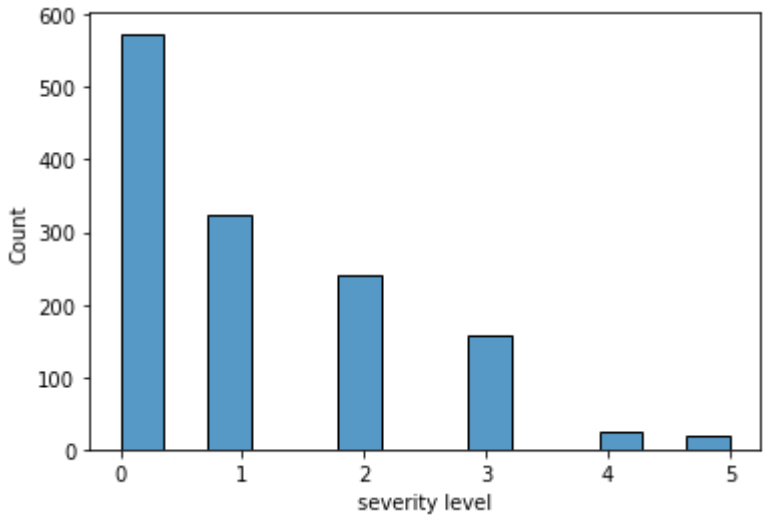
In [49]:
```
sns.histplot(data=df, x="smoker", shrink=.7, )
```

Out[49]: <AxesSubplot:xlabel='smoker', ylabel='Count'>



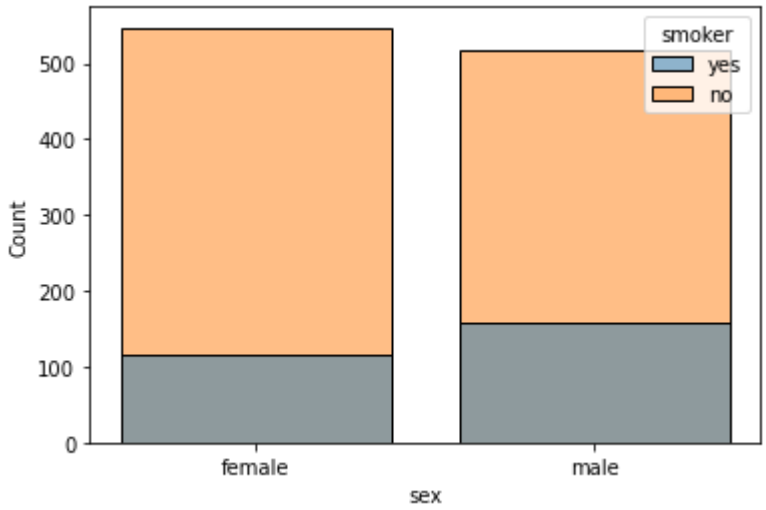
In [57]:
```
sns.histplot(data=df, x="severity level", shrink=1
                )
```

Out[57]: <AxesSubplot:xlabel='severity level', ylabel='Count'>



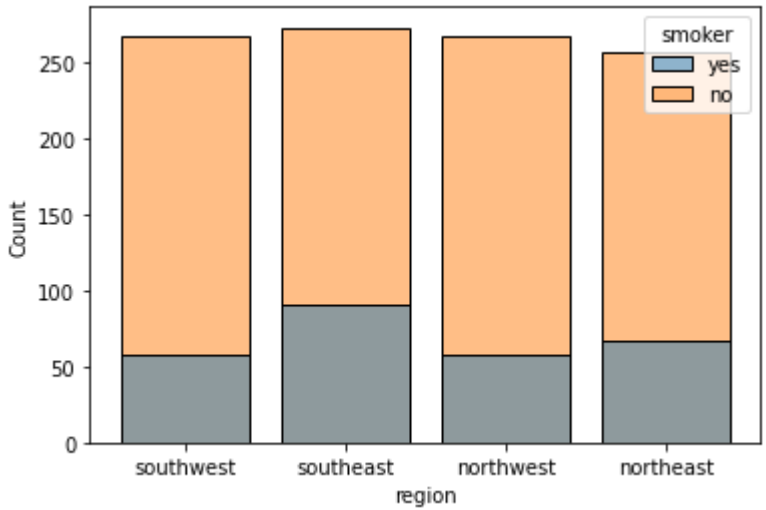
In [157]:
```
sns.histplot(data=df,

    x='sex', hue='smoker', shrink=.8)
    #y=None)
```

Out[157]: <AxesSubplot:xlabel='sex', ylabel='Count'>



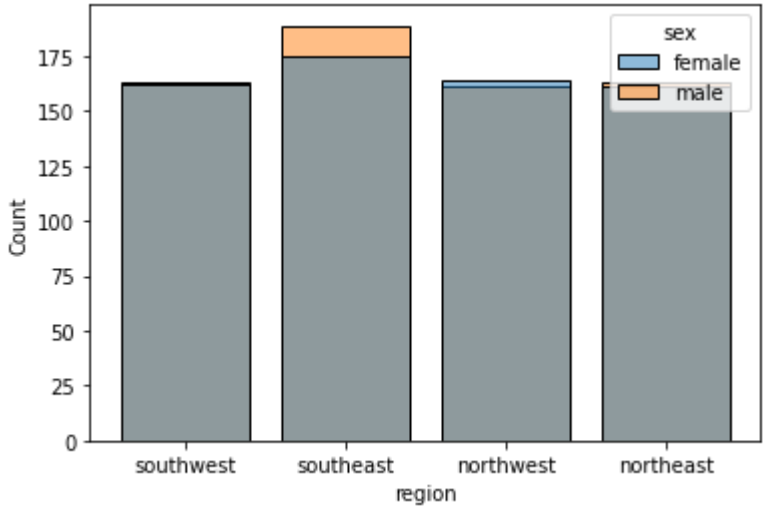
In [71]:
```
sns.histplot(data=df,

    x='region', hue='smoker', shrink=.8, legend=True)
    #y=None)
```

Out[71]: <AxesSubplot:xlabel='region', ylabel='Count'>



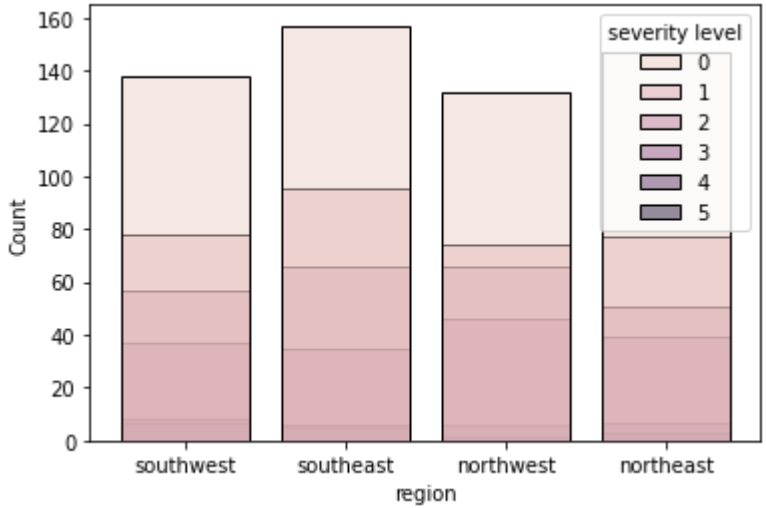
In [159]:
```
sns.histplot(data=df,
x='region', hue='sex', shrink=.8)
```

Out[159]: <AxesSubplot:xlabel='region', ylabel='Count'>



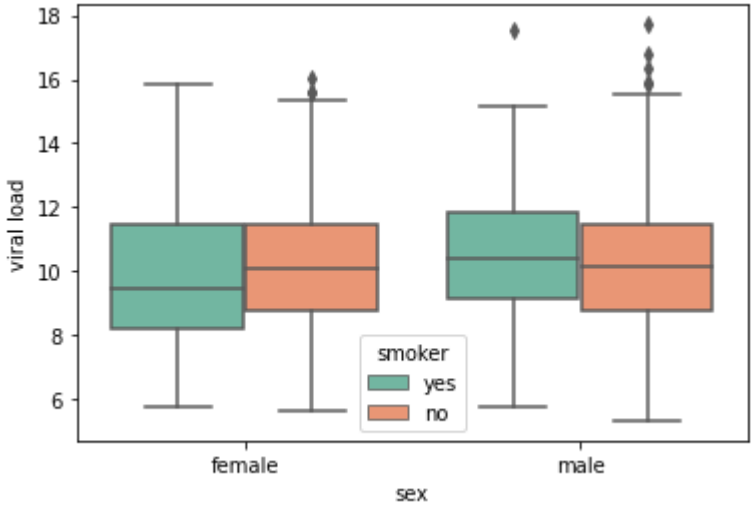
In [76]:
```
sns.histplot(data=df,
x='region', hue='severity level', shrink=.8)
```

Out[76]: <AxesSubplot:xlabel='region', ylabel='Count'>



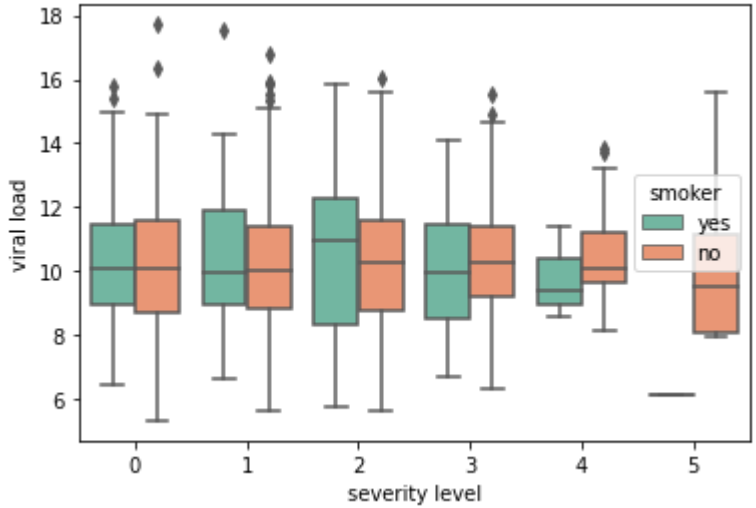
In [77]:
```
sns.boxplot(x = df['sex'],
            y = df['viral load'],
            hue = df['smoker'],
            palette = 'Set2')
```

Out[77]: <AxesSubplot:xlabel='sex', ylabel='viral load'>



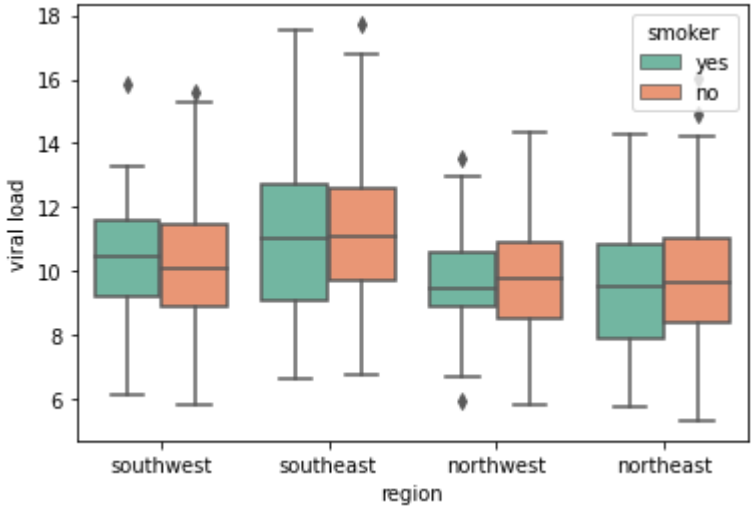
In [80]:
```
sns.boxplot(x = df['severity level'],
            y = df['viral load'],
            hue = df['smoker'],
            palette = 'Set2')
```

Out[80]: <AxesSubplot:xlabel='severity level', ylabel='viral load'>



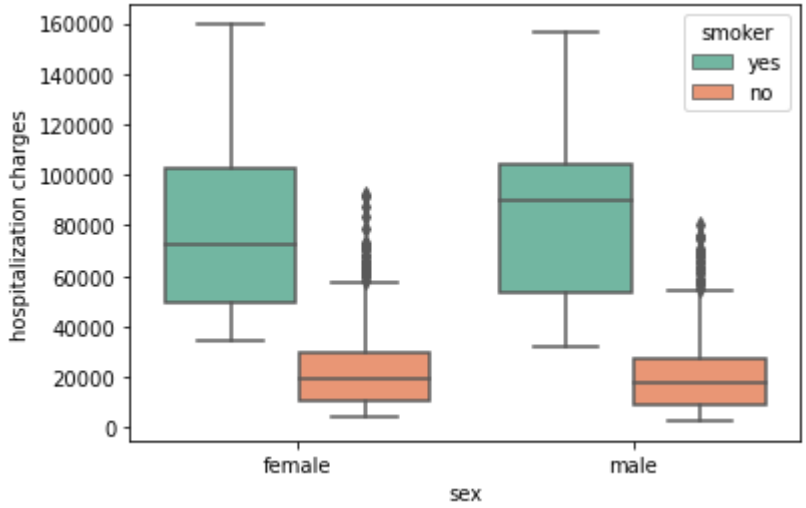
In [79]:
```
sns.boxplot(x = df['region'],
            y = df['viral load'],
            hue = df['smoker'],
            palette = 'Set2')
```

Out[79]: <AxesSubplot:xlabel='region', ylabel='viral load'>



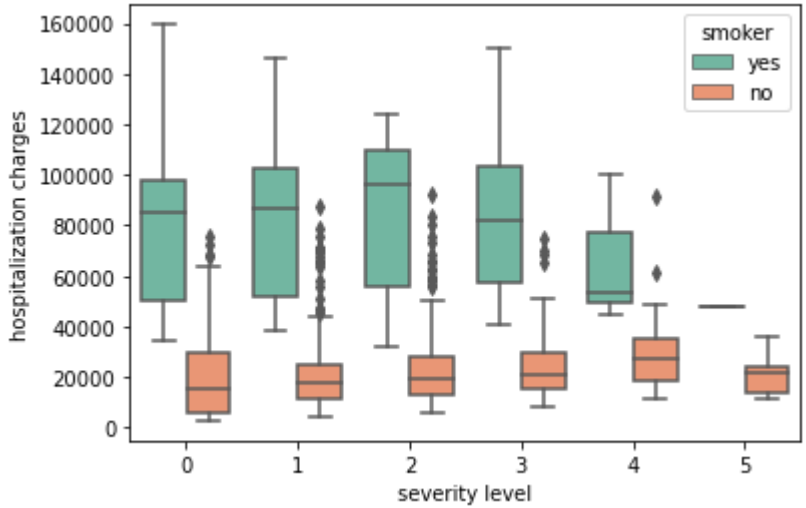
In [78]:
```
sns.boxplot(x = df['sex'],
            y = df['hospitalization charges'],
            hue = df['smoker'],
            palette = 'Set2')
```

Out[78]: <AxesSubplot:xlabel='sex', ylabel='hospitalization charges'>



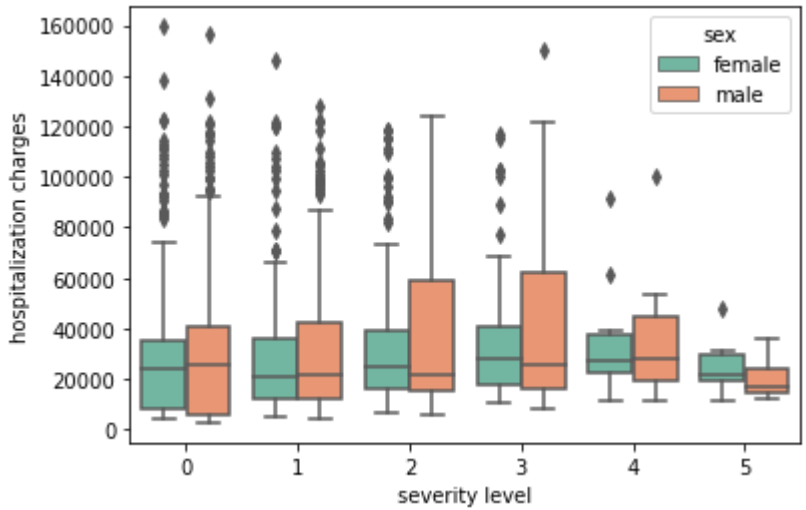
In [82]:
```
sns.boxplot(x = df['severity level'],
            y = df['hospitalization charges'],
            hue = df['smoker'],
            palette = 'Set2')
```

Out[82]: <AxesSubplot:xlabel='severity level', ylabel='hospitalization charges'>



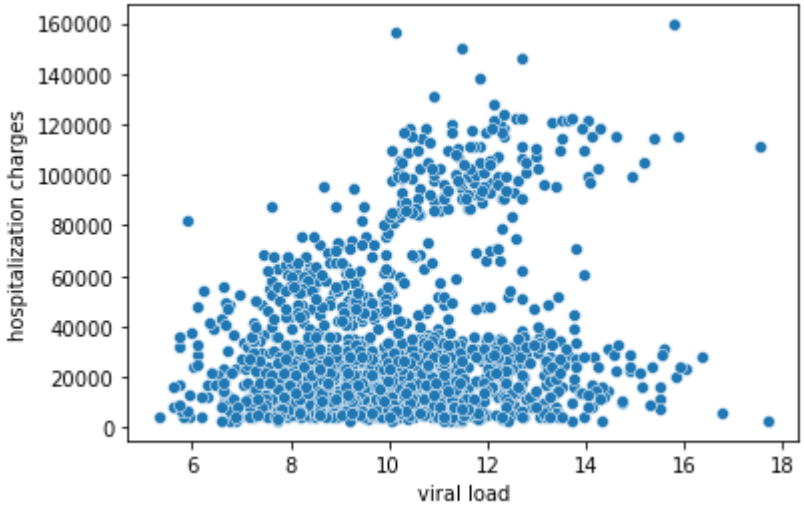
In [83]:
```
sns.boxplot(x = df['severity level'],
            y = df['hospitalization charges'],
            hue = df['sex'],
            palette = 'Set2')
```

Out[83]: <AxesSubplot:xlabel='severity level', ylabel='hospitalization charges'>



In [27]:
```
import seaborn as sns
sns.scatterplot(data=df, x="viral load", y="hospitalization charges")
```

Out[27]:
```
<AxesSubplot:xlabel='viral load', ylabel='hospitalization charges'>
```



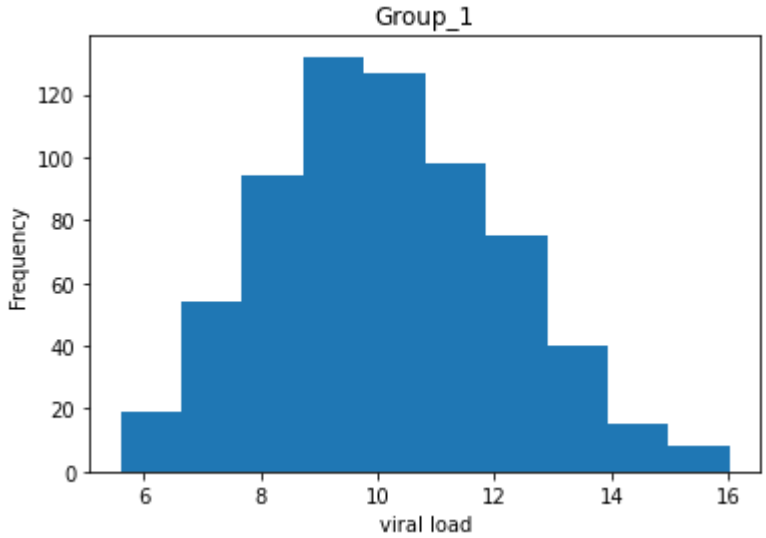
Statistical Analysis :

Prove (or disprove) with statistical evidence that the viral load of females is different from that of males (T-test Two tailed)

In [89]:
```
Group_1=df[df['sex']=='female']['viral load']
Group_2=df[df['sex']=='male']['viral load']
```

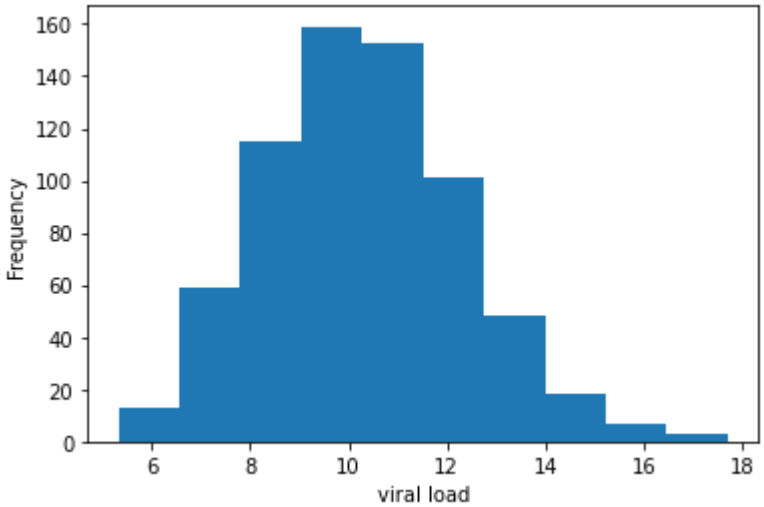
In [98]:
```
Group_1.plot(kind="hist", title="Group_1")
plt.xlabel("viral load")
```

Out[98]:
```
Text(0.5, 0, 'viral load')
```



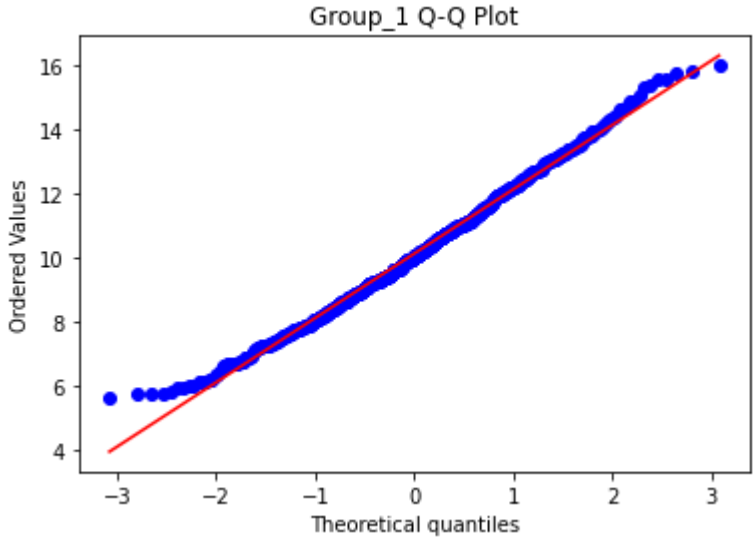
In [97]:
```
Group_2.plot(kind="hist", title="")
plt.xlabel("viral load")
```

Out[97]:
```
Text(0.5, 0, 'viral load')
```



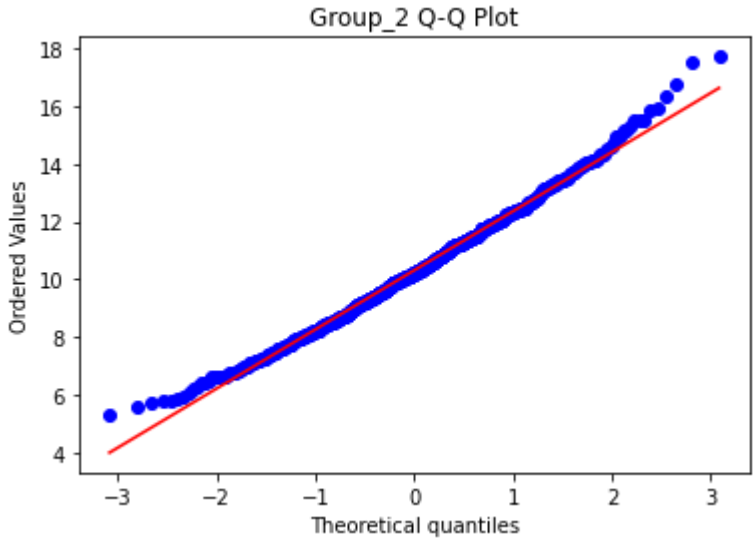
In [99]:
```
from scipy import stats
import matplotlib.pyplot as plt
stats.probplot(Group_1, dist="norm", plot= plt)
plt.title("Group_1 Q-Q Plot")
```

Out[99]: Text(0.5, 1.0, 'Group_1 Q-Q Plot')



In [100]:
```
stats.probplot(Group_2, dist="norm", plot= plt)
plt.title("Group_2 Q-Q Plot")
```

Out[100]:
```
Text(0.5, 1.0, 'Group_2 Q-Q Plot')
```



In [101]:
```
stats.shapiro(Group_1)
```

Out[101]:
```
ShapiroResult(statistic=0.9930474162101746, pvalue=0.003624602919444442)
```

In [102]:
```
stats.shapiro(Group_2)
```

Out[102]:
```
ShapiroResult(statistic=0.9930650591850281, pvalue=0.003189612179994583)
```

In [104]:
```
import numpy as np
print(np.var(Group_1), np.var(Group_2))
```

```
4.055708441872559 4.183557507396447
```

In [106]:
```
#Null Hypothesis (H0) Mu1=Mu2
#alternate hypothesis (H1) Mu1=!Mu2
```

In [107]:
```
# Set a significance level (alpha) = 0.05
```

In [108]:
```
stats.ttest_ind(Group_1, Group_2)
```

Out[108]:
```
Ttest_indResult(statistic=-1.695711164450323, pvalue=0.0901735841670204)
```

```
# Sinece P-value is greater than significance level we can not reject the null hypothesis mean we are
accepting the
# null hypothesis
```

Is the proportion of smoking significantly different across different regions? (Chi-square)

In [112]:
```
pd.crosstab([df['region']], [df['smoker']], df['smoker'], aggfunc='count')
```

Out[112]:

smoker	no	yes
region		
northeast	257	67
northwest	267	58
southeast	273	91
southwest	267	58

In [115]:
```
#Null Hypothesis (H0) Mu1=Mu2=mu3=mu4
#alternate hypothesis (H1) Mu1=!Mu2=!mu3=!Mu4
```

In [116]:
```
# Set a significance level (alpha) = 0.05
```

In [113]:
```
from scipy.stats import chi2_contingency

# defining the table
data = [[257, 267, 273, 267], [67, 58, 91, 58]]
stat, p, dof, expected = chi2_contingency(data)
```

In [114]:
```
stat, p, dof, expected
```

Out[114]:
```
(7.34347776140707,
 0.06171954839170547,
 3,
 array([[257.65022422, 258.44544096, 289.45889387, 258.44544096],
        [ 66.34977578,  66.55455904,  74.54110613,  66.55455904]]))
```

In [117]:
```
# Sinece P-value is greater than significance level we can not reject the null hypothesis mean we are accepting
# null hypothesis
```

Is the mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level the same?

#Explain your answer with statistical evidence (One way Anova)

In [118]:
```
#Null Hypothesis (H0) Mu1=Mu2=mu3
#alternate hypothesis (H1) Mu1=!Mu2=!mu3
```

In [151]:
```
Group11=df[(df.sex == "female") & (df['severity level'] == 0)]['viral load']
Group21=df[(df.sex == "female") & (df['severity level'] == 1)]['viral load']
Group31=df[(df.sex == "female") & (df['severity level'] == 2)]['viral load']
```

In [152]:
```
from scipy.stats import f_oneway
```

In [153]:
```
f_oneway(Group11,Group21,Group31)
```

Out[153]:
```
F_onewayResult(statistic=0.3355061434584082, pvalue=0.7151189650367746)
```

In [154]:
```
# Set a significance level (alpha) = 0.05
```

In [155]:
```
# Sinece P-value is greater than significance level we can not reject the null hypothesis mean we are accepting
# null hypothesis
```

Business Insights

In [ ]:
```
# 1 we have 25% of smoker in the dataset
# 2 smoker are paying more hospitalization charges
# 3 male smoke more than famale
# 4 female smoker have more viral load as compare to male viral load
# 5 the proportion of smoking is same across different regions.
```

Recommendations

In [ ]:
```
# 1 we have 25% of smoker in the dataset
# 2 smoker are paying more hospitalization charges
# 3 male smoke more than famale
# 4 female smoker have more viral load as compare to male viral load
# 5 the proportion of smoking is same across different regions.
```