Problem Statement : Apollo Hospitals was established in 1983, renowned as the architect of modern healthcare in India. As the nation's first corporate hospital, Apollo Hospitals is acclaimed for pioneering the private healthcare revolution in the country.

In this case study we are going to use the dataset given to us for predicting the reason for hospitalization for different regions. This can be done using Hypothesis testing using various tests where different categorical and numerical variables can be test against each other and draw meaningful insights. The main motive as data scientists is to draw insights and make recommendations which will help Apollo build a robust response system which can withstand any critical times in the country.

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import copy
```

In [2]:
```
from scipy.stats import shapiro
from scipy.stats import levene
from scipy.stats import ttest_ind
from scipy.stats import chi2_contingency
from scipy.stats import f_oneway
from statsmodels.graphics.gofplots import qqplot
```

In [3]:
```
apollo = pd.read_csv('C:/DSML/Apollo - case study/scaler_apollo_hospitals.csv')
```

In [4]:
```
apollo.shape
```

Out[4]:
```
(1338, 8)
```

In [5]:
```
apollo.columns
```

Out[5]:
```
Index(['Unnamed: 0', 'age', 'sex', 'smoker', 'region', 'viral load',
       'severity level', 'hospitalization charges'],
      dtype='object')
```

In [6]:
```
print(apollo['sex'].value_counts())
print(apollo['smoker'].value_counts())
print(apollo['region'].value_counts())
print(apollo['severity level'].value_counts())
```

```
male      676
female    662
Name: sex, dtype: int64
no     1064
yes     274
Name: smoker, dtype: int64
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
0    574
1    324
2    240
3    157
4     25
5     18
Name: severity level, dtype: int64
```

Basic Metrics: We have a total of 1338 rows which represent the number of patients and 8 columns which indicate the attributes/columns we will use for drawing insights.

1. We have 676 Male and 662 female patients in the dataset
2. There are 1064 non smokers and 274 smokers among the patients
3. The regions are divided into southeast(364), southwest(325), northwest(325) and northeast(324).

4. The severity levels ranging from 0 to 5 where severity 0 has the highest no. of patients and severity 5 has the lowest no. of patients

In [7]:
```
apollo.head()
```

Out[7]:

	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 [8]:
```
apollo.dtypes
```

Out[8]:
```
Unnamed: 0                  int64
age                         int64
sex                        object
smoker                     object
region                     object
viral load                float64
severity level              int64
hospitalization charges     int64
dtype: object
```

In [9]:
```
#converting severity level to object since this will be treated as a categorical variable
apollo['severity level'] = apollo['severity level'].astype('object')
```

Missing value detection

In [10]:
```
apollo.isna().sum()
```

Out[10]:
```
Unnamed: 0                0
age                       0
sex                       0
smoker                    0
region                    0
viral load                0
severity level            0
hospitalization charges   0
dtype: int64
```

Since there are no missing values, we can continue

Statistical Summary

In [11]:
```
apollo[['age','viral load','hospitalization charges']].describe(include='all')
```

Out[11]:

	age	viral load	hospitalization charges
count	1338.000000	1338.000000	1338.000000
mean	39.207025	10.221233	33176.058296
std	14.049960	2.032796	30275.029296
min	18.000000	5.320000	2805.000000
25%	27.000000	8.762500	11851.000000
50%	39.000000	10.130000	23455.000000
75%	51.000000	11.567500	41599.500000
max	64.000000	17.710000	159426.000000

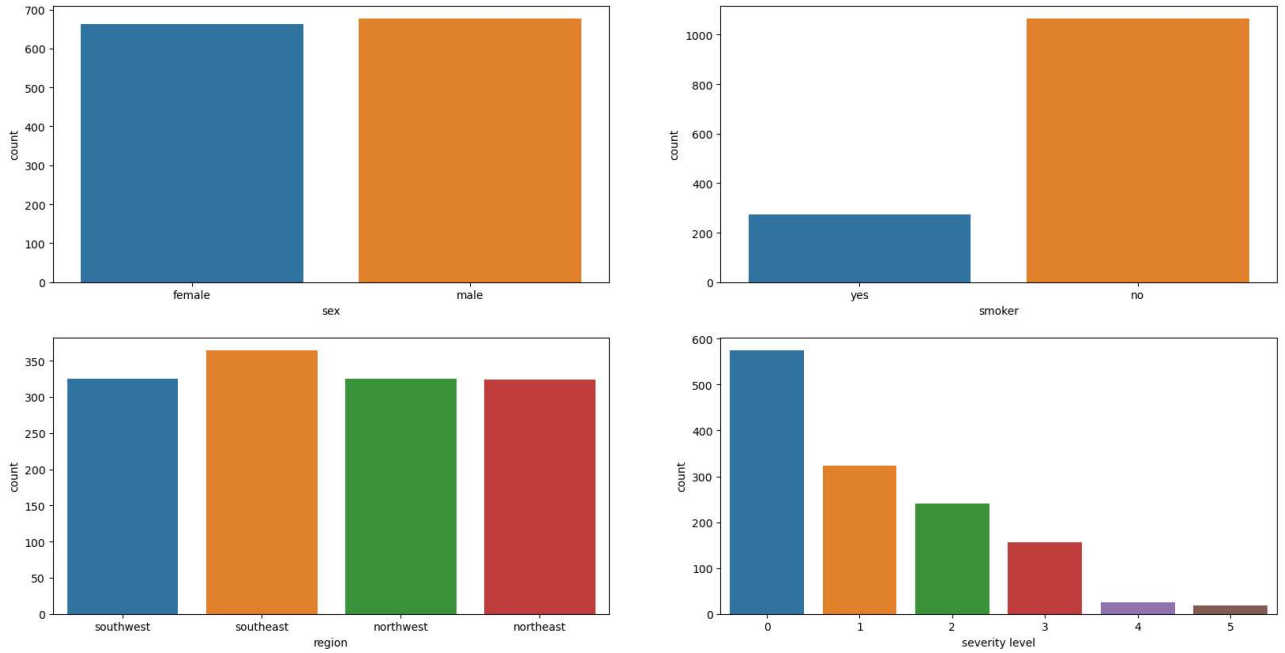
```
Observations on the above statistical summary
1. The minimum age range of the patient is 18 years and maximum is 64 years. Mean age of the patients in
the dataset is around 39 years
2. Viral load does not have a lot of range with minimum viral load being 5.32 and maxinum being 17.71
3. Hospitalization charges has a very big range with minimum being Rs.2805 and maximum being Rs.159426.
Also by looking at the 75th percentile there seems to be outliers in hospitalization charges
```

Visual Analysis

Univariate Analysis

In [12]:
```
fig, axis = plt.subplots(nrows=2, ncols=2, figsize=(20, 10))
sns.countplot(x = 'sex',data = apollo,ax=axis[0,0])
sns.countplot(x = 'smoker',data = apollo,ax=axis[0,1])
sns.countplot(x = 'region',data = apollo,ax=axis[1,0])
sns.countplot(x = 'severity level',data = apollo,ax=axis[1,1])
```

Out[12]: <AxesSubplot:xlabel='severity level', ylabel='count'>



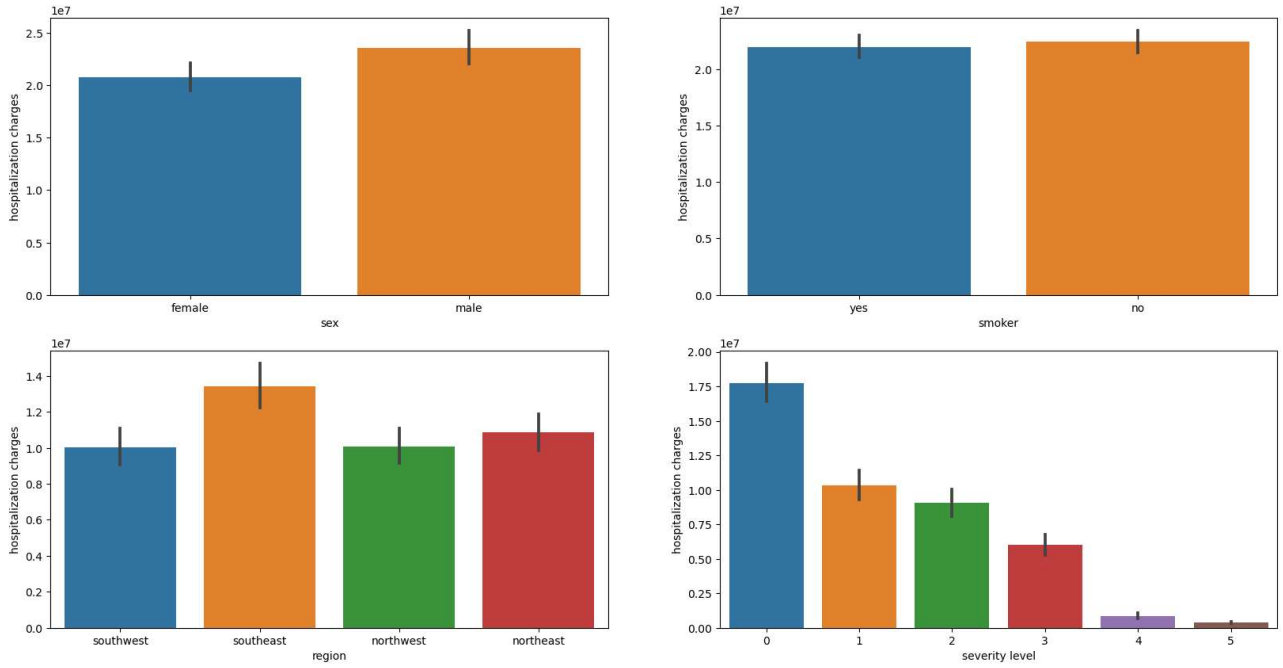

```
Observations based on the univariate analysis of the categorical variables

1. The number of male and  female patients are almost the same
2. The number of non smokers are more than 4 times the smokers in the dataset
```

3. All 4 regions have almost the same number of patients with southeast region having a slightly higher count
4. The number of patients with severity 0 is the highest and with severity 5 is the lowest

Bivariate Analysis

In [13]:
```
fig, axis = plt.subplots(nrows=2, ncols=2, figsize=(20, 10))
sns.barplot(x = 'sex',y = 'hospitalization charges',data = apollo,estimator = np.sum,ax=axis[0,0])
sns.barplot(x = 'smoker',y = 'hospitalization charges',data = apollo,estimator = np.sum,ax=axis[0,1])
sns.barplot(x = 'region',y = 'hospitalization charges',data = apollo,estimator = np.sum,ax=axis[1,0])
sns.barplot(x = 'severity level',y = 'hospitalization charges',data = apollo,estimator = np.sum,ax=axis[1,1
```

Out[13]: <AxesSubplot:xlabel='severity level', ylabel='hospitalization charges'>



Some observations based on the bivariate analysis of categorical variables vs the hospitalization charges

1. The hospitalization charges is higher in males when compared to females. Although we have seen previously that the no. of male and female patients in the dataset are almost the same
2. The hospitalization charges are same for smokers and non smokers. Although previously we have seen that the non smokers count is more than 4 times than that of smokers. This means smokers have much higher hospitalization charges than non smokers.
3. The hospitalization charges for southeast region is higher than the other 3 with other 3 having more or less the same charges
4. The hospitalization charges for patients with severity 0 is the highest and severity 5 is the lowest. Although, this could be since we have previously seen that number of patients with severity 0 is the highest and severity 5 is the lowest

Soon, we are going to visualize what is the range of hospitalization charges for all these categories to compare if some of these categories are actually similar or if there is a difference

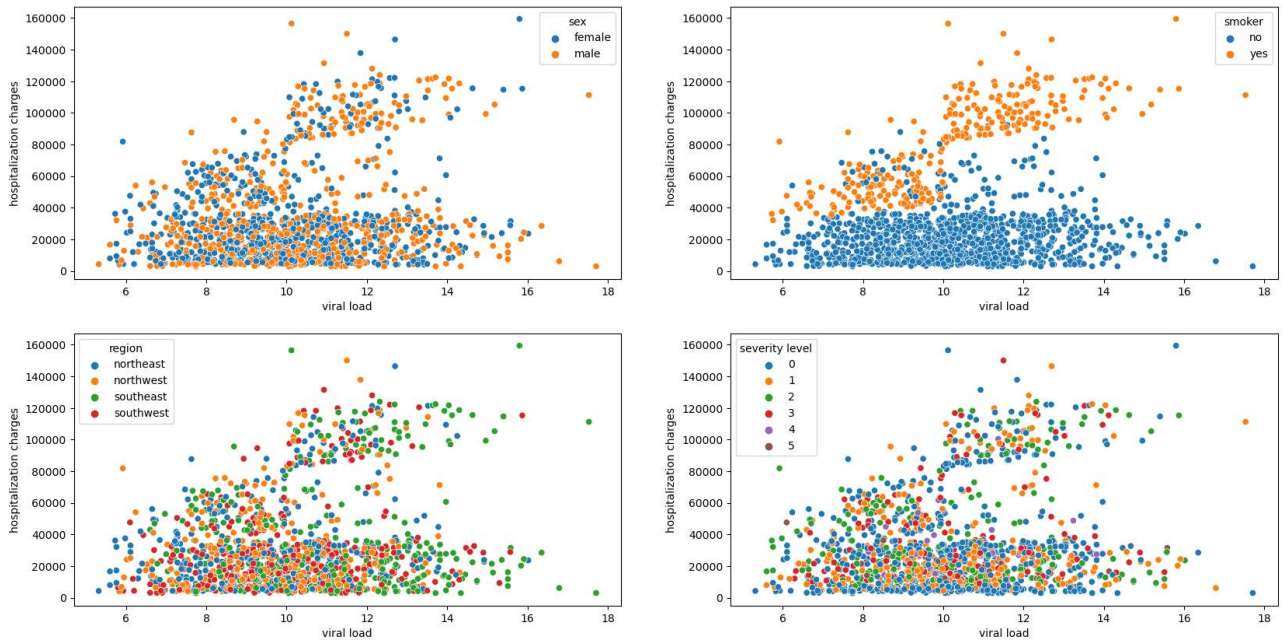
In [14]:
```
apollo['sex'] = apollo['sex'].astype('category')
apollo['smoker'] = apollo['smoker'].astype('category')
apollo['region'] = apollo['region'].astype('category')
apollo['severity level'] = apollo['severity level'].astype('category')
```

In [15]:
```
apollo.dtypes
```

Out[15]:
```
Unnamed: 0                  int64
age                         int64
sex                      category
smoker                   category
region                   category
viral load                float64
severity level           category
hospitalization charges     int64
dtype: object
```

In [16]:
```
fig, axis = plt.subplots(nrows=2, ncols=2, figsize=(20, 10))
sns.scatterplot(data=apollo, y='hospitalization charges',x='viral load',hue='sex',ax=axis[0,0])
sns.scatterplot(data=apollo, y='hospitalization charges',x='viral load',hue='smoker',ax=axis[0,1])
sns.scatterplot(data=apollo, y='hospitalization charges',x='viral load',hue='region',ax=axis[1,0])
sns.scatterplot(data=apollo, y='hospitalization charges',x='viral load',hue='severity level',ax=axis[1,1])
```

Out[16]:
```
<AxesSubplot:xlabel='viral load', ylabel='hospitalization charges'>
```

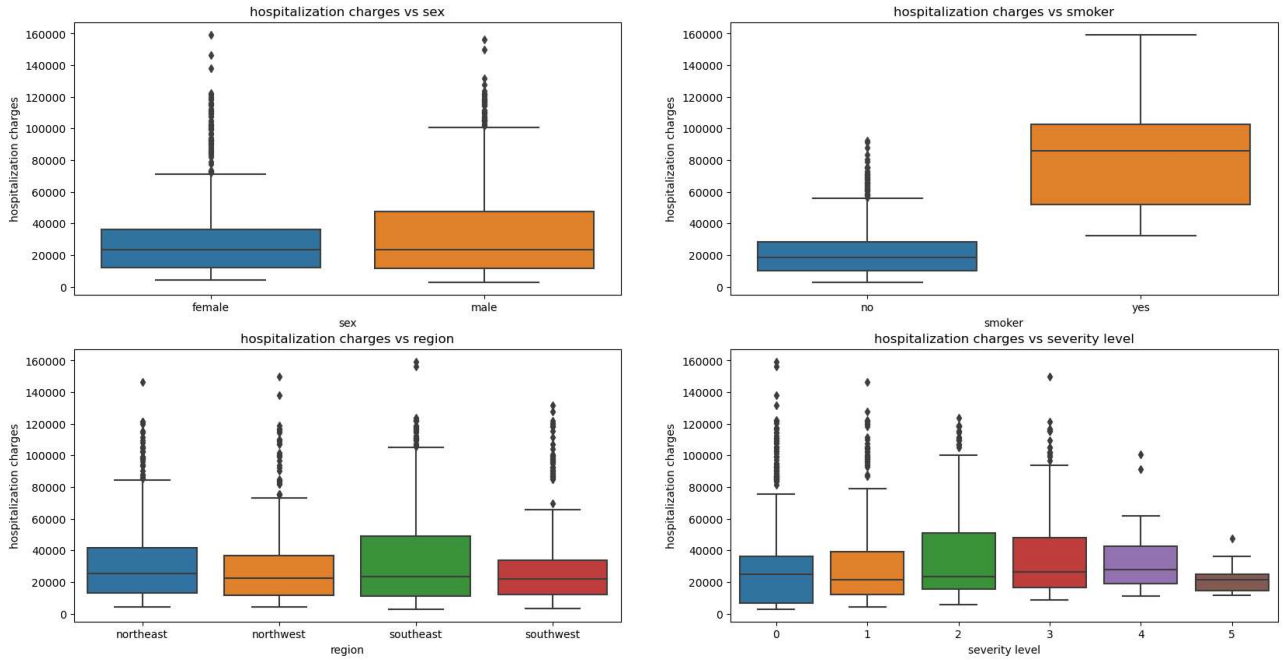


Some observations based on the analysis of viral load vs hospitalization charges

1. Majority of patients lie between the viral load around 7 to 14. Although its interesting to note that increase in the viral load has not directly impacted the increase in hospitalization charges. Also, the distribution is similar for both males and females
2. Majority of non smoker patients lie between the viral load around 7 to 14. But the smoker patients are distributed across all levels of viral load. Also, we can note here that for smokers, the hospitalization charges have increased with the viral load. This is not the case for non smokers.
3. The viral load's impact on hospitalization charges for different regions is similar to males and females. There is no noticable distinction for any particular region
4. The viral load's impact on hospitalization charges for severity levels to males and females. There is no noticable distinction for any particular severity level

In [17]:
```
plt.subplots(nrows=2, ncols=2, figsize=(20, 10))
(data=apollo, y='hospitalization charges',x='sex',ax=axis[0,0]).set_title('hospitalization charges vs sex')
(data=apollo, y='hospitalization charges',x='smoker',ax=axis[0,1]).set_title('hospitalization charges vs sm
(data=apollo, y='hospitalization charges',x='region',ax=axis[1,0]).set_title('hospitalization charges vs re
(data=apollo, y='hospitalization charges',x='severity level',ax=axis[1,1]).set_title('hospitalization charg
```

Out[17]: Text(0.5, 1.0, 'hospitalization charges vs severity level')

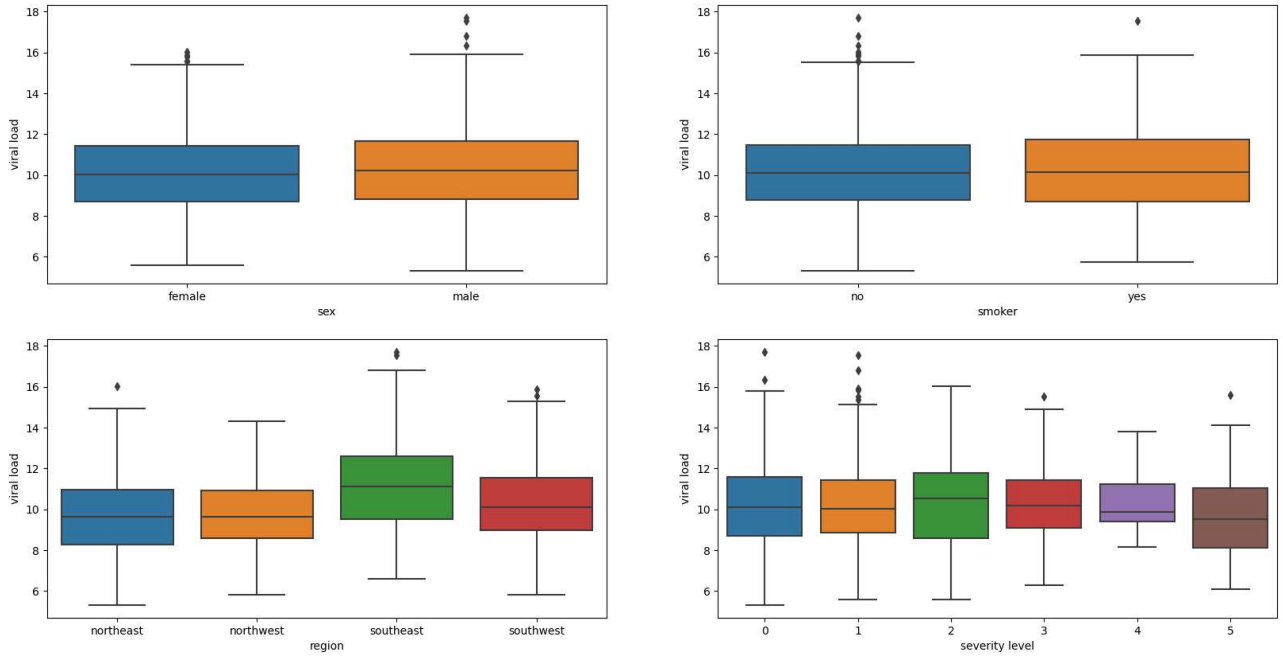


Some observations based on the analysis of different categories vs hospitalization charges(outliers visualization)

1. The hospitalization charges is higher in males when compared to females. Although the outliers are much more for females than males
2. The hospitalization charges for smokers is much more as expected than non smokers. We have some outliers only for non smokers
3. The hospitalization charges for southeast region is higher than the other 3 with other 3 having more or less the same charges as seen before from the bar plots.
4. The range hospitalization charges for patients with all severities are more or less similar with only patients with severity levl 5 being an exception. Severity levels 0 and 1 have more outliers than the rest

In [18]:
```
fig, axis = plt.subplots(nrows=2, ncols=2, figsize=(20, 10))
sns.boxplot(data=apollo, y='viral load',x='sex',ax=axis[0,0])
sns.boxplot(data=apollo, y='viral load',x='smoker',ax=axis[0,1])
sns.boxplot(data=apollo, y='viral load',x='region',ax=axis[1,0])
sns.boxplot(data=apollo, y='viral load',x='severity level',ax=axis[1,1])
```

Out[18]: <AxesSubplot:xlabel='severity level', ylabel='viral load'>



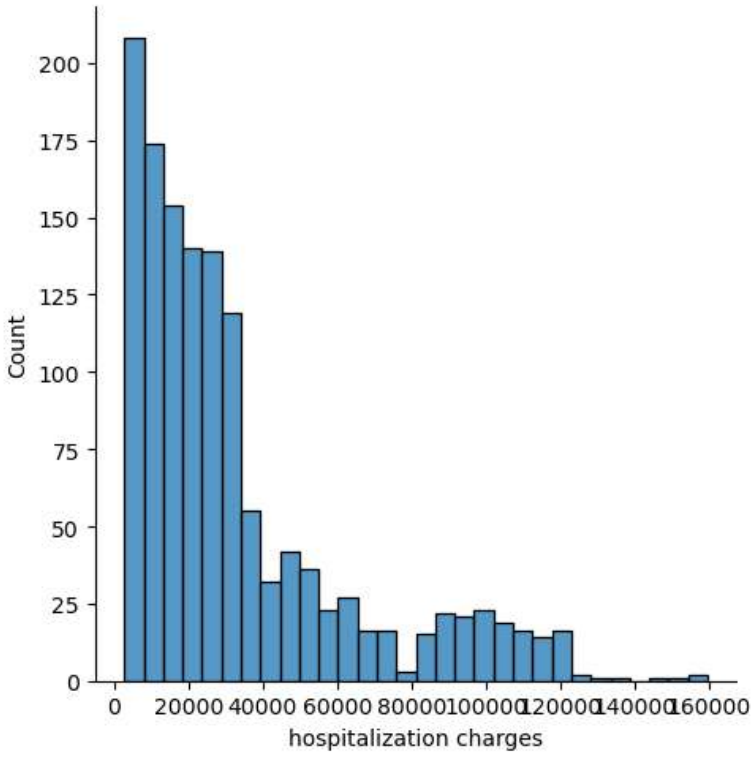
Some observations based on the analysis of different categories vs viral load

1. Viral load between males and females are similar
2. Viral load between smokers and non smokers are similar
3. Viral load in southeast region is slightly higher than the other 3 regions
4. Viral load between different severity levels are more or less similar except for severity level 4

Let us check the distribution of continuos variables i.e Hospitalization charges and Viral load since this is crucial for our testing

In [19]:
```
sns.displot(apollo['hospitalization charges'])
```

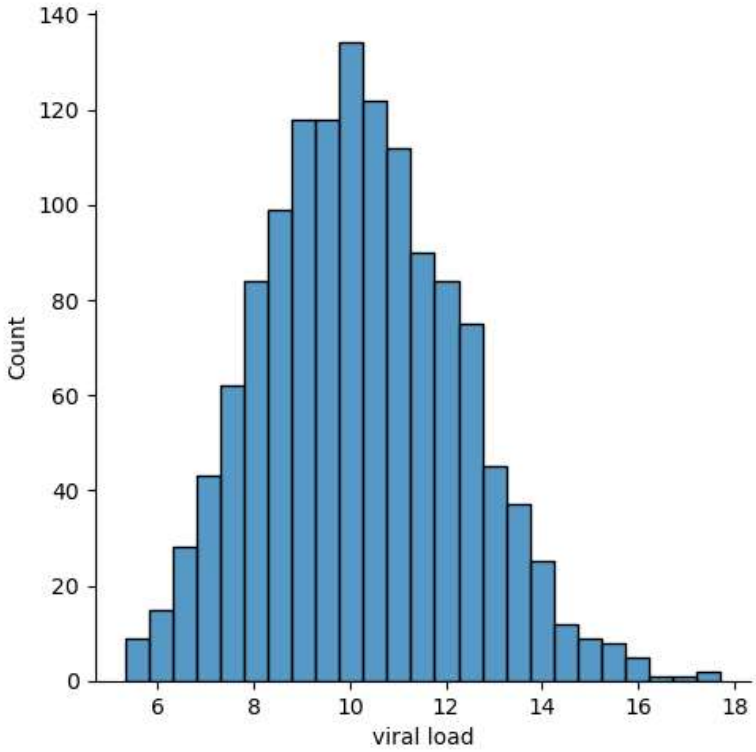
Out[19]:
```
<seaborn.axisgrid.FacetGrid at 0x1e906b12a58>
```



The distribution of the hospitalization charges is heavily right-skewed

In [20]:
```
sns.displot(apollo['viral load'])
```

Out[20]:
```
<seaborn.axisgrid.FacetGrid at 0x1e906b12710>
```

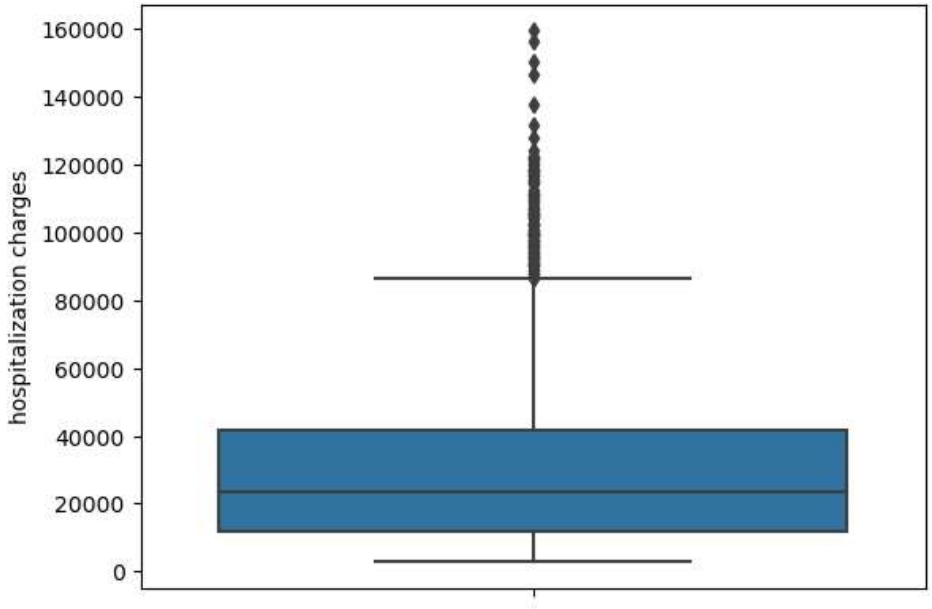


The viral load follows almost a normal distribution as per the visualization above

Outlier treatment

In [21]:
```
sns.boxplot(data=apollo, y='hospitalization charges')
```

Out[21]: <AxesSubplot:ylabel='hospitalization charges'>



The hospitalization charges has a lot outliers. Let us see how to handle them

In [22]:
```
def outlier_treatment(datacolumn):
    sorted(datacolumn)
    Q1,Q3 = np.percentile(datacolumn , [25,75])
    IQR = Q3 - Q1
    lower_range = Q1 - (1.5 * IQR)
    upper_range = Q3 + (1.5 * IQR)
    return lower_range,upper_range
```

In [23]:
```
apollo1 = copy.deepcopy(apollo)
```

In [24]:
```
lower_range, upper_range = outlier_treatment(apollo1['hospitalization charges'])
```

In [25]:
```
percentage_outliers = 1 - len(apollo1[apollo1['hospitalization charges'] <= upper_range ]) / len(apollo1)
percentage_outliers
```

Out[25]: 0.1038863976083707

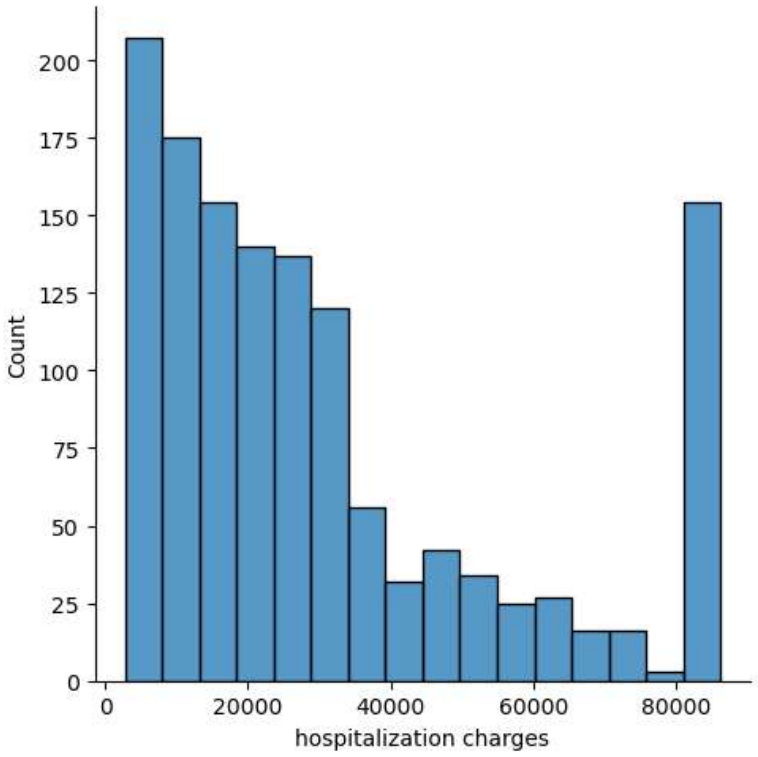
We cannot drop all the outliers since they constitute 10% of the data column

Let us use the capping method to check if the outliers can be handled effectively

In [26]:
```
apollo1.loc[apollo1['hospitalization charges'] > upper_range,['hospitalization charges']] = upper_range
```

In [27]:
```
sns.displot(apollo1['hospitalization charges'])
```

Out[27]:
```
<seaborn.axisgrid.FacetGrid at 0x1e908e5a9e8>
```




```
Although we can see that the capping has removed the outliers, it made the data distribution not fit for analysis. Hence we decide to go ahead with the original data
```

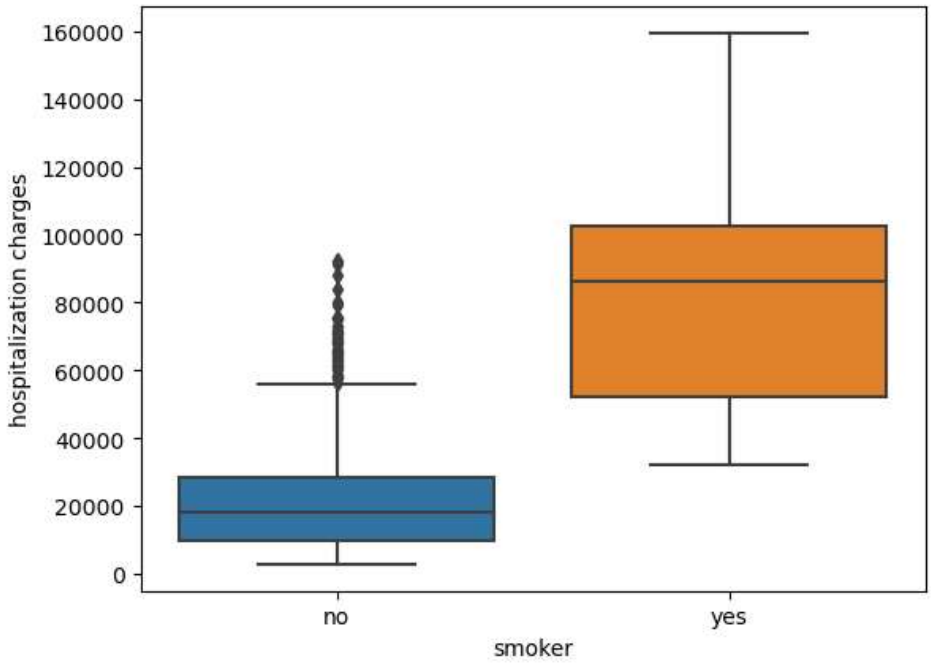
Hypothesis Testing

Lets do the statistical analysis using the Hypothesis testing one by one for each case

Prove (or disprove) that the hospitalization of people who do smoking is greater than those who don't? (T-test Right tailed)

In [28]:
```
sns.boxplot(data=apollo, y='hospitalization charges',x='smoker')
```

Out[28]: <AxesSubplot:xlabel='smoker', ylabel='hospitalization charges'>



From the visual analysis, it is evident that the hospitalization charges of the smoker group is very much higher than the non smoker group.
Now let us do some hypothesis testing to investigate this further

Before using the T-test Right tailed directly, let us test the normality and variances of the two categories using Shapiro-wilk test and Levene's test respectively as these are the crucial assumptions of a t-test

Levene's test to check variances between smoker and non-smoker category

In [29]:
```
# H0: The variance of smoker group is equal to non-smoker group
# Ha: The variance of smoker and non-smoker groups are not equal
smoker_group = apollo[apollo['smoker'] == 'yes']['hospitalization charges']
non_smoker_group = apollo[apollo['smoker'] == 'no']['hospitalization charges']
alpha = 0.05 # testing the null hypothesis at 95% confidence level
test_stat, p_value = levene(smoker_group,non_smoker_group)
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say the variances are not equal')
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say the variances are equal')
```

p_value: 1.5595259401311176e-66 Hence we reject the H0 and say the variances are not equal

Since we can clearly see that the p value from the Levene's test is very much lower than alpha value, we can Reject the null hypothesis and conclude that the variances of smoker and non smoker groups are not equal

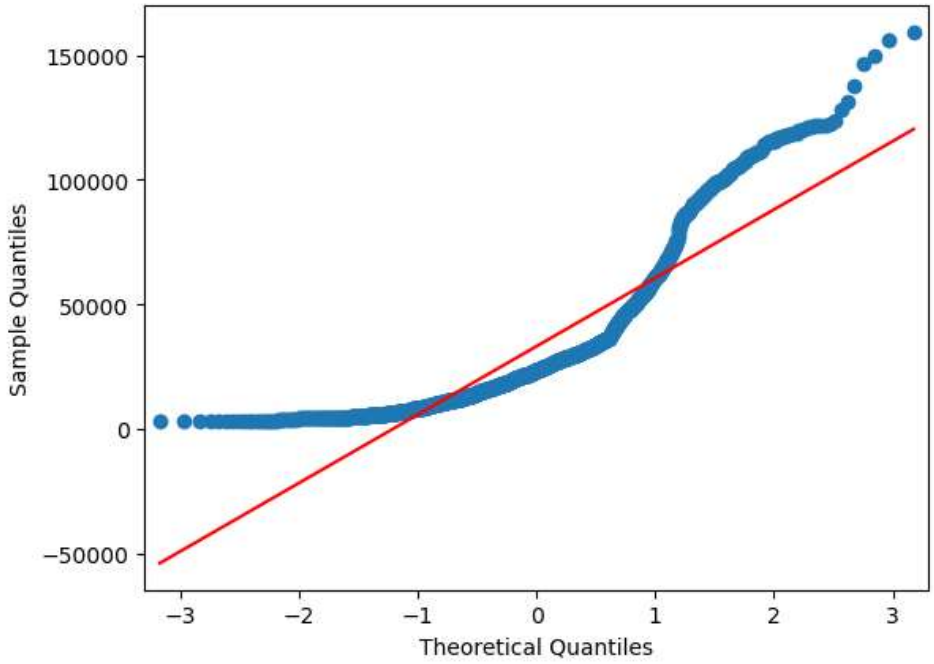
Shapiro-wilk test to check the normality of hospitalization charges

In [30]:
```
# H0 : Hospitalization charges data is normally distributed
# Ha : Hospitalization charges data is not normally distributed
hosp_charges = apollo['hospitalization charges']
alpha = 0.05 # testing the null hypothesis at 95% confidence level
test_stat, p_value = shapiro(hosp_charges)
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say the data is not normally distributed')
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say the data is not normally distributed'
```

```
p_value: 1.1505333015369624e-36 Hence we reject the H0 and say the data is not normally distributed
```

Since we can clearly see that the p value from the Shapiro-wilk test is very much lower than alpha value, we can Reject the null hypothesis and conclude that the hospitalization charges data is not normally distributed

In [31]:
```
#We can also visualize the same using QQ-plot to visually represent what the data looks like in comparison
#to normal distribution

qqplot(hosp_charges,line='r')
plt.show()
```

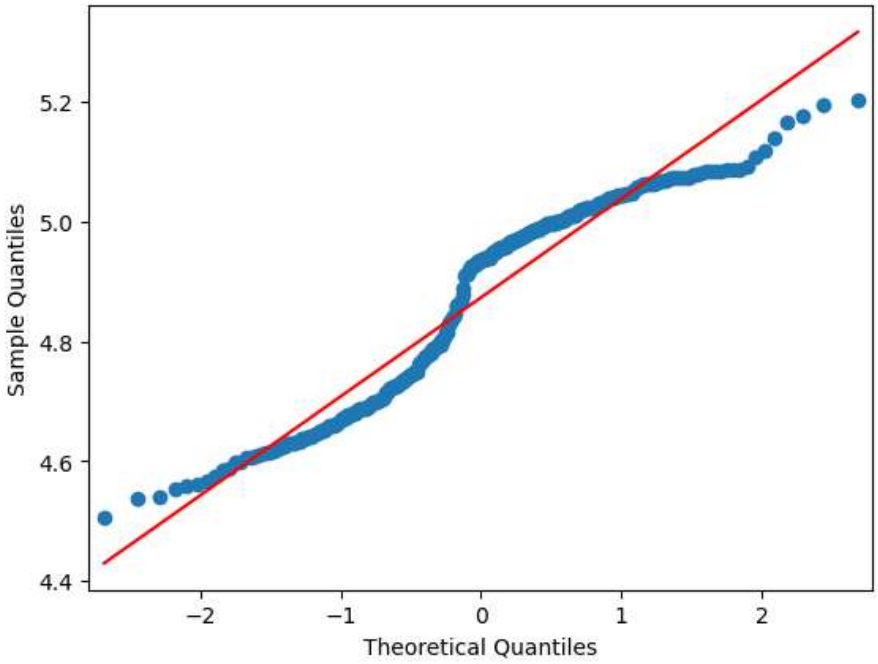


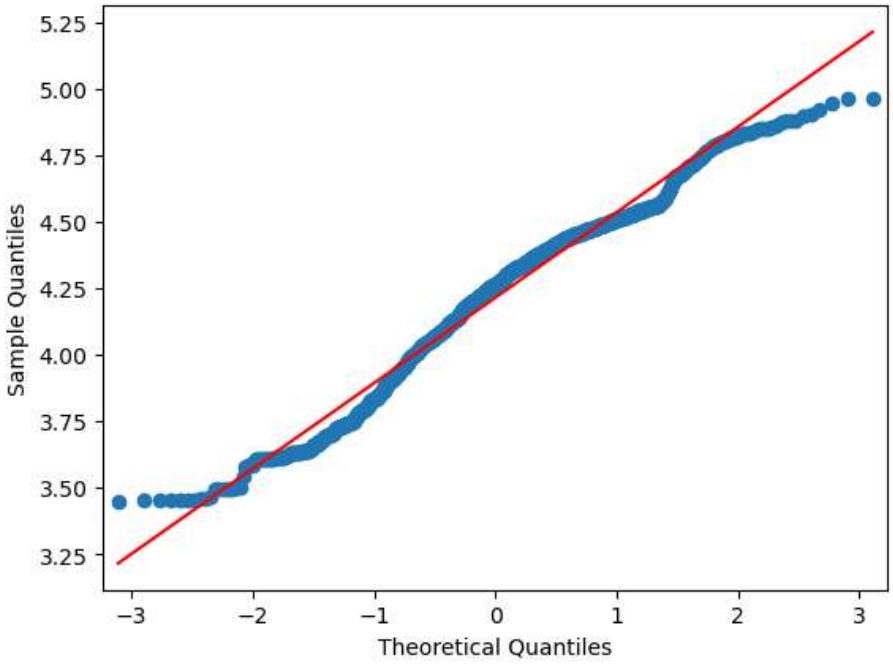
From the QQ-plot we can very clearly visualize how the hospitalization charges data is distributes compared to a normal distribution

From our above tests, we can clearly see that the data is neither normal nor does the groups have equal variances. Hence we cannot use a parametric test like t-test directly on this data. Hence lets transform the data using log10 function and then perform the t-test on the same

In [32]:
```
log_smoker_data = np.log10(smoker_group)
log_non_smoker_data = np.log10(non_smoker_group)
```

In [33]:
```
qqplot(log_smoker_data,line='r')
plt.show()
qqplot(log_non_smoker_data,line='r')
plt.show()
```





The data seem to be slightly better after the transformation. So let us go ahead and conduct the right tailed T-test

In [34]:
```
# H0: The smoker and non smoker groups have similar hospitalization charges
# Ha: The smoker group has more hospitaliization charges compared to non smoker group
test_stat, p_value = ttest_ind(log_smoker_data,log_non_smoker_data,alternative='greater')
alpha = 0.05 # testing the null hypothesis at 95% confidence level
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say that the smoker group has higher hospitalizati
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say both groups have similar hospitalizati
```

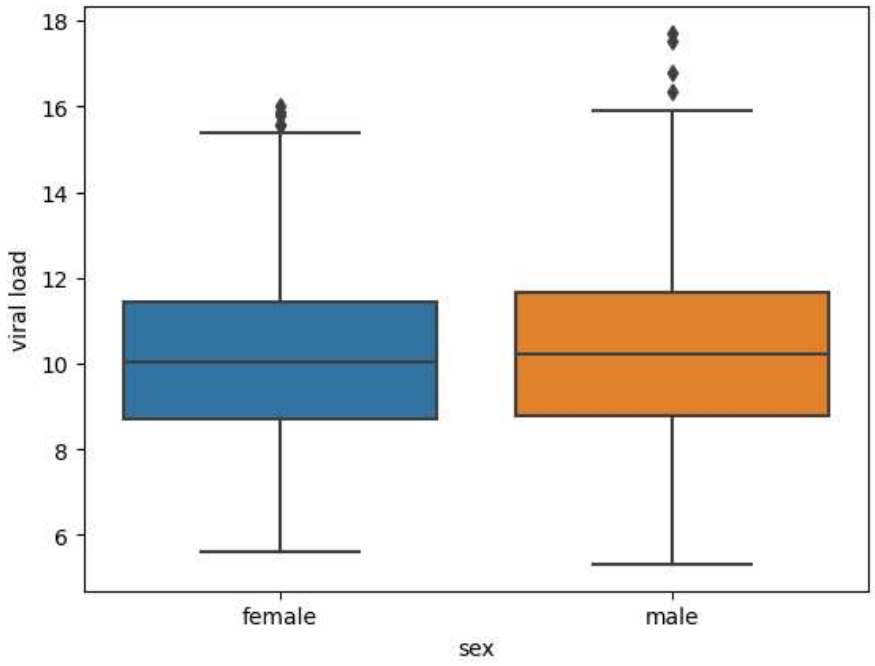
```
p_value: 3.152579721103535e-172 Hence we reject the H0 and say that the smoker group has higher hospitali
zation charges
```

Since we can clearly see that the p value from the t-test is very much lower than alpha value, we can Reject the null hypothesis and conclude that the smoker group has higher hospitalization charges

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

In [35]:
```
sns.boxplot(data=apollo, y='viral load',x='sex')
```

Out[35]:
```
<AxesSubplot:xlabel='sex', ylabel='viral load'>
```



From the above box plot it is clear that the viral load is similar between males and females. We can test this further using hypothesis testing

Before using the T-test Right tailed directly, let us test the normality and variances of the two categories using Shapiro-wilk test and Levene's test respectively as these are the crucial assumptions of a t-test

Levene's test to check variances between males and females

In [36]:
```
# H0: The variance of males is equal to females
# Ha: The variance of males and females are different
male = apollo[apollo['sex'] == 'male']['viral load']
female = apollo[apollo['sex'] == 'female']['viral load']
test_stat, p_value = levene(male,female)
alpha = 0.05 # testing the null hypothesis at 95% confidence level
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say the variances are not equal')
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say the variances are equal')
```

```
p_value: 0.9503708012456551 Hence we fail to reject the H0 and say the variances are equal
```

Since we can clearly see that the p value from the Levene's test is very much higher than alpha value, we fail to reject the null hypothesis and conclude that the variances of males and females are equal

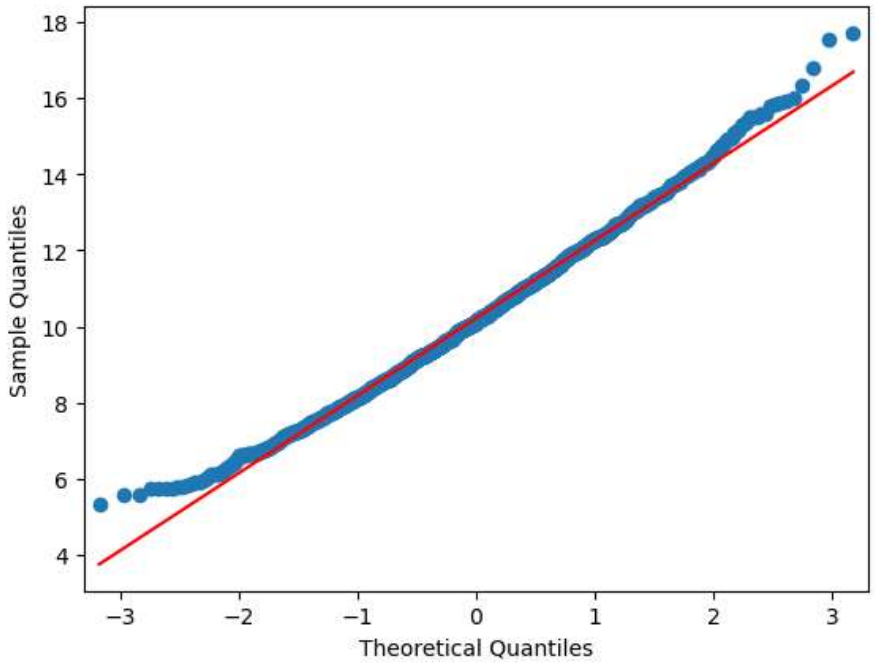
Shapiro-wilk test to check the normality of hospitalization charges

In [37]:
```
# H0 : Viral Load data is normally distributed
# Ha : Viral Load data is not normally distributed
viral_load = apollo['viral load']
test_stat, p_value = shapiro(viral_load)
alpha = 0.05 # testing the null hypothesis at 95% confidence level
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say the data is not normally distributed')
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say the data is not normally distributed'
```

```
p_value: 2.6902040190179832e-05 Hence we reject the H0 and say the data is not normally distributed
```

Since we can clearly see that the p value from the Shapiro-wilk test is very much lower than alpha value, we can Reject the null hypothesis and conclude that the viral load data is not normally distributed

In [38]:
```
#We can also visualize the same using QQ-plot to visually represent what the data looks like in comparison
#to normal distribution

qqplot(viral_load,line='r')
plt.show()
```



From the QQ-plot we can visualize how the viral load data is distributed compared to a normal distribution. It is very close to a normal distribution

From the above test we can fairly assume that we can use a parametric test like t-test to compare the difference in viral load between Males and Females

In [39]:
```
# H0: Males and Females have similar viral load
# Ha: Males and Females have different viral load
test_stat, p_value = ttest_ind(male,female)
alpha = 0.05 # testing the null hypothesis at 95% confidence level
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say that the viral load in females is different fr
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say both groups have similar viral load')
```

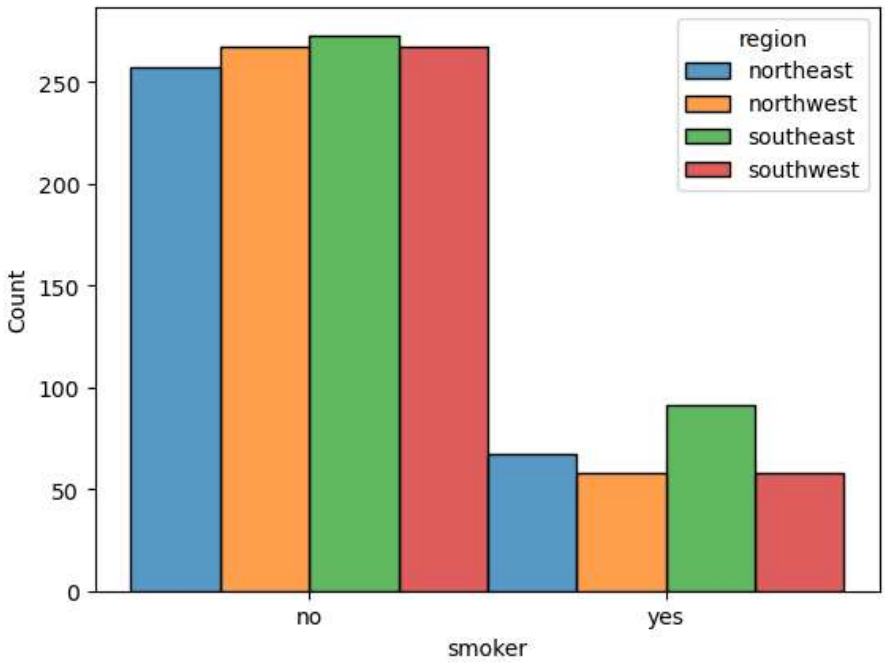
```
p_value: 0.0901735841670204 Hence we fail to reject the H0 and say both groups have similar viral load
```

Since we can clearly see that the p value from the t-test is more than alpha value, we Fail to reject the null hypothesis and conclude that males and females have a similar viral load

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

In [40]:
```
sns.histplot(binwidth=1,
             x='smoker',
             hue='region',
             data=apollo,
             stat="count",
             multiple="dodge")
```

Out[40]: <AxesSubplot:xlabel='smoker', ylabel='Count'>



From the above plot we can see that the proportion of smoking is similar across all 4 regions. Let us further do hypothesis testing to validate our intution

Since Chi-square is a non-parametric test, we will not be doing any assumptions check as we did before using t-test

In [41]:
```
smoking_region_data = pd.crosstab(index=apollo['smoker'],columns=apollo['region'])
smoking_region_data
```

Out[41]:

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

In [42]:
```
# H0 : Proportion of smoking is similar across all the regions
# Ha : Proportion of smoking is significantly different across all regions
chi_stat, p_value, dof, expected = chi2_contingency(smoking_region_data)
alpha = 0.05 # testing the null hypothesis at 95% confidence level
print("chi_stat:",chi_stat)
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say that roportion of smoking is significantly di
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say Proportion of smoking is similar acros
```

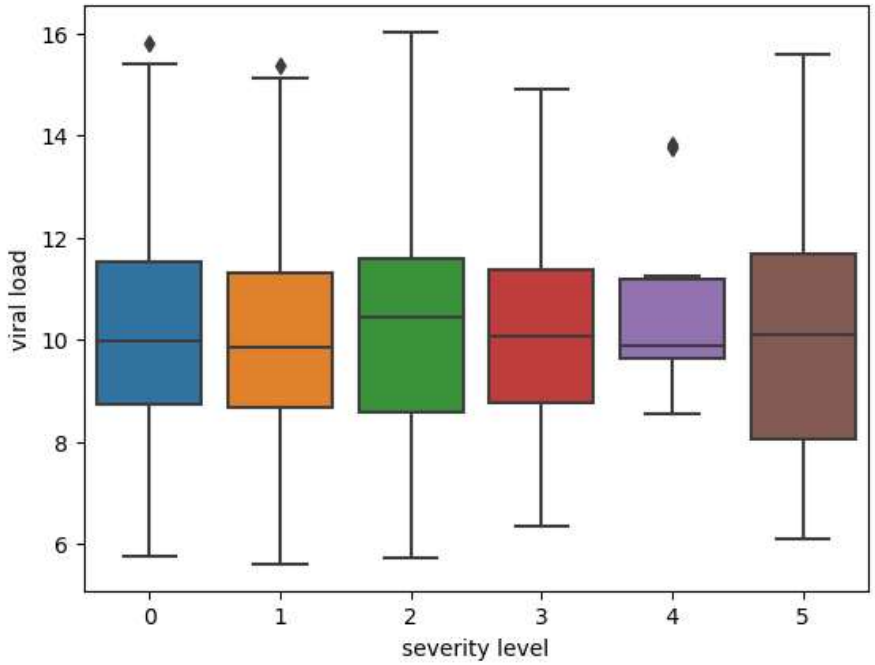
```
chi_stat: 7.34347776140707
p_value: 0.06171954839170547 Hence we fail to reject the H0 and say Proportion of smoking is similar acro
ss all the regions
```

Since we can clearly see that the p value from the chi-square test is more than alpha value, we fail to reject the null hypothesis and conclude that the smoking proportion is similar across all the regions

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 [43]:
```
sns.boxplot(data=apollo[apollo['sex'] == 'female'], y='viral load',x='severity level')
```

Out[43]: <AxesSubplot:xlabel='severity level', ylabel='viral load'>



From the above plot we can see that the viral load between females of severity 0,1 and 2 are similar. We can further test this using hypothesis testing

Before using One way ANOVA, let us test the normality and variances of the two categories using Shapiro-wilk test and Levene's test respectively as these are the crucial assumptions of ANOVA since its a parametric test

Levene's test to check variances in viral load between women with with 0 Severity level , 1 Severity level, and 2 Severity level

In [44]:
```
# H0: The variance in viral load between women with with 0 Severity level , 1 Severity level, and 2 Severit
# Ha: The variance in viral load between women with with 0 Severity level , 1 Severity level, and 2 Severit
women_severity_0 = apollo[(apollo['severity level'] == 0) & (apollo['sex'] == 'female')]['viral load']
women_severity_1 = apollo[(apollo['severity level'] == 1) & (apollo['sex'] == 'female')]['viral load']
women_severity_2 = apollo[(apollo['severity level'] == 2) & (apollo['sex'] == 'female')]['viral load']
alpha = 0.05 # testing the null hypothesis at 95% confidence level
test_stat, p_value = levene(women_severity_0,women_severity_1,women_severity_2)
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say the variances are not equal')
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say the variances are equal')
```

p_value: 0.38987253596513605 Hence we fail to reject the H0 and say the variances are equal

Since we can clearly see that the p value from the Levene's test is higher than alpha value, we fail to reject the null hypothesis and conclude that the variances between the 3 groups are equal

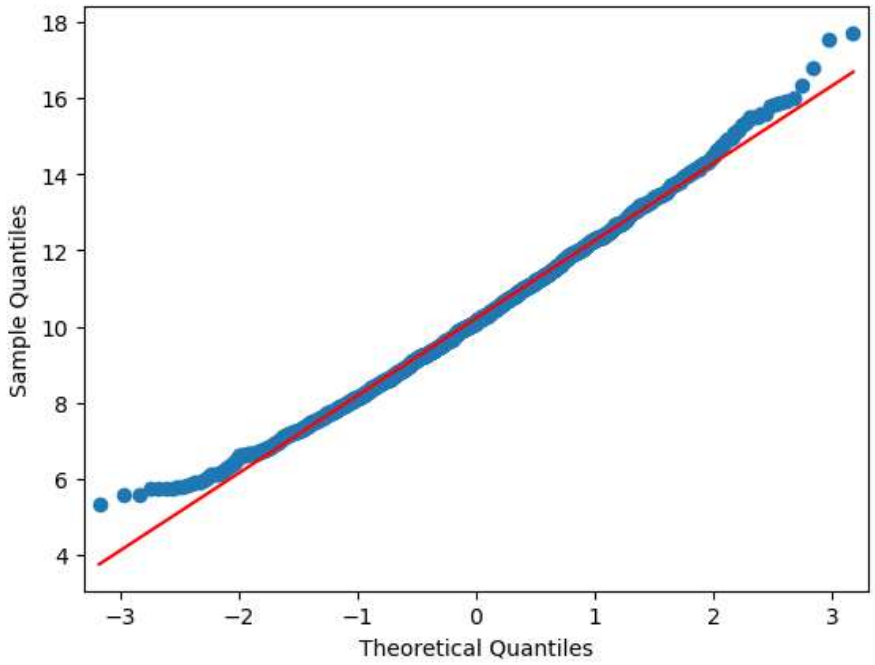
Shapiro-wilk test to check the normality of viral load

In [45]:
```
# H0 : Viral load data is normally distributed
# Ha : Viral load data is not normally distributed
viral_load = apollo['viral load']
alpha = 0.05 # testing the null hypothesis at 95% confidence level
test_stat, p_value = shapiro(viral_load)
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say the data is not normally distributed')
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say the data is not normally distributed'
```

p_value: 2.6902040190179832e-05 Hence we reject the H0 and say the data is not normally distributed

Since we can clearly see that the p value from the Shapiro-wilk test is much lower than alpha value, we reject the null hypothesis and conclude that the data viral load is not normally distributed

In [46]:
```
#We can also visualize the same using QQ-plot to visually represent what the data looks like in comparison
#to normal distribution

qqplot(viral_load,line='r')
plt.show()
```



From the QQ-plot we can visualize how the viral load data is distributed compared to a normal distribution. It is very close to a normal distribution

From the above test, we can conclude that we can use ANOVA test to compare the difference in viral load between females of severity levels 0,1 and 2

Since we also need to have the same number of samples across the groups we need to first check how many samples we need to use

In [47]:
```
print('no of women with severity 0 : ',len(women_severity_0))
print('no of women with severity 1 : ',len(women_severity_1))
print('no of women with severity 2 : ',len(women_severity_2))
```

```
no of women with severity 0 :  289
no of women with severity 1 :  158
no of women with severity 2 :  119
```

We are going to use 119 samples since that is the least number we have among the 3 categories

In [48]:
```
# H0: Viral load between females of severity levels 0, 1 and 2 are the same
# Ha: Viral load between females of severity levels 0, 1 and 2 are different
f_stat, p_value = f_oneway(women_severity_0.sample(119),women_severity_1.sample(119),women_severity_2.samp
alpha = 0.05 # testing the null hypothesis at 95% confidence level
if p_value < alpha:
    print('p_value:',p_value,'Hence we reject the H0 and say that the Viral load between females of severit
else:
    print('p_value:',p_value,'Hence we fail to reject the H0 and say Viral load between females of severity
```

```
p_value: 0.48352764350413124 Hence we fail to reject the H0 and say Viral load between females of severit
y levels 0, 1 and 2 are the same
```

```
Since the p value is much higher than the alpha value, we fail to reject the H0 and say Viral load
between females of severity levels 0, 1 and 2 are similar
```

In [ ]: