Apollo Business Case Study

Context

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.

As a data scientist working at Apollo 24/7, the ultimate goal is to tease out meaningful and actionable insights from Patient-level collected data.

You can help Apollo hospitals to be more efficient, to influence diagnostic and treatment processes, to map the spread of a pandemic.

One of the best examples of data scientists making a meaningful difference at a global level is in the response to the COVID-19 pandemic, where they have improved information collection, provided ongoing and accurate estimates of infection spread and health system demand, and assessed the effectiveness of government policies.

How can you help here?

The company wants to know:

• Which variables are significant in predicting the reason for hospitalization for different regions

• How well some variables like viral load, smoking, Severity Level describe the hospitalization charges

Dataset: **Click here (https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/681/original/scaler_apollo_hospitals.csv)**

Column Profiling

Age: This is an integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government). Sex: This is the policy holder's gender, either male or female Viral Load: Viral load refers to the amount of virus in an infected person's blood Severity Level: This is an integer indicating how severe the patient is Smoker: This is yes or no depending on whether the insured regularly smokes tobacco. Region: This is the beneficiary's place of residence in Delhi, divided into four geographic regions - northeast, southeast, southwest, or northwest Hospitalization charges: Individual medical costs billed to health insurance

Concept Used:

- Graphical and Non-Graphical Analysis

- Graphical and Non-Graphical Analysis
- 2-sample t-test: testing for difference across populations
- ANOVA
- Chi-square

In [941]:
```
# importing all the required packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from scipy.stats import shapiro, levene, ttest_ind, chi2_contingency, f_oneway

# ignore the warnings
import warnings
warnings.filterwarnings('ignore')

# spad the notebook accros the width of the screen
from IPython.core.display import display, HTML
display(HTML("<style>.container { width:100% !important; }</style>"))
from IPython.core.display import display, HTML
display(HTML("<style>.container { width:100% !important; }</style>"))
```

In [942]:
```
%%html
<!--# to avoid the scrooling if the output of viz is too large (optional)-->
<style>
    .output_scroll {
        height: unset !important;
        max-height: unset !important;
    }
</style>
```

Importing the dataset and performing usual data analysis steps like checking the structure & characteristics of the dataset

In [943]:
```
df = pd.read_csv('C:/Users/pshashank3/Desktop/Data Science/Scaler/Datasets/Projects/Apollo/scaler_apollo_hospitals
df.head()
```

Out[943]:

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

In [944]:
```
df.shape
```

Out[944]: (1338, 7)

In [945]:
```
print(f"Number of rows: {df.shape[0]:,} \nNumber of columns: {df.shape[1]:,}")
```

```
Number of rows: 1,338
Number of columns: 7
```

The datatype of all columns is shown below

In [946]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 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   smoker                  1338 non-null   object 
 3   region                  1338 non-null   object 
 4   viral load              1338 non-null   float64
 5   severity level          1338 non-null   int64  
 6   hospitalization charges  1338 non-null   int64  
dtypes: float64(1), int64(3), object(3)
memory usage: 83.6+ KB
```

Categorical columns Stats

In [947]:
```
df.describe(include = 'object').T
```

Out[947]:

	count	unique	top	freq
sex	1338	2	male	676
smoker	1338	2	no	1064
region	1338	4	southeast	364

Numerical columns Stats

In [948]:
```
df.describe().T
```

Out[948]:

	count	mean	std	min	25%	50%	75%	max
age	1338.0	39.207025	14.049960	18.00	27.0000	39.00	51.0000	64.00
viral load	1338.0	10.221233	2.032796	5.32	8.7625	10.13	11.5675	17.71
severity level	1338.0	1.094918	1.205493	0.00	0.0000	1.00	2.0000	5.00
hospitalization charges	1338.0	33176.058296	30275.029296	2805.00	11851.0000	23455.00	41599.5000	159426.00

Since there are no null values, we dont need to do any missing value treatement

In [949]:
```
df.isnull().sum()
```

Out[949]:
```
age                       0
sex                       0
smoker                    0
region                    0
viral load                0
severity level            0
hospitalization charges   0
dtype: int64
```

Feature analysis : Univariate and Bi-variate Analysis

Univariate Analysis

Categorical columns stats and vizualization.

In [950]:
```
df.describe(include = 'object').T
```

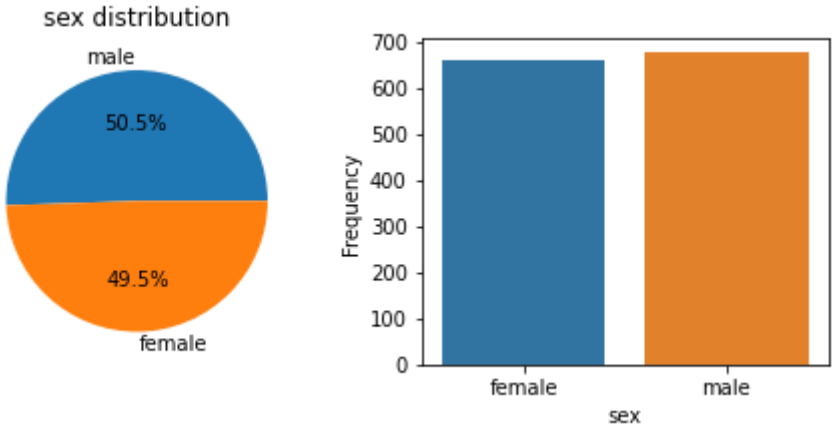
Out[950]:

	count	unique	top	freq
sex	1338	2	male	676
smoker	1338	2	no	1064
region	1338	4	southeast	364

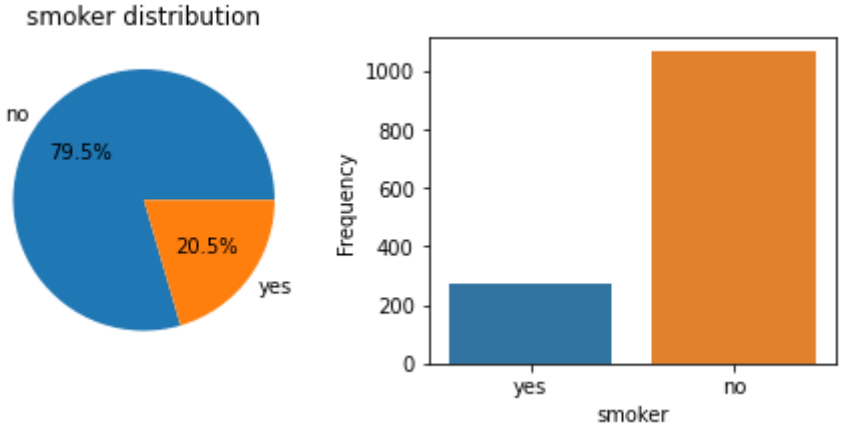
In [951]:

```
Categorical_Cols = df.select_dtypes(include='object').columns.to_list()

print('The value counts for all categorical columns: \n')
for col in Categorical_Cols:
    print('Column :- ', col)
    print(df[col].value_counts())

    data = df[col].value_counts().values.tolist()
    lbls = df[col].value_counts().index.tolist()

    plt.figure(figsize = (8,3))
    plt.subplot(121)
    plt.pie(data, labels = lbls, autopct='%1.1f%%')
    plt.title(col + ' distribution')

    plt.subplot(122)
    sns.countplot(df[col])
    plt.ylabel('Frequency')
    plt.show()
```

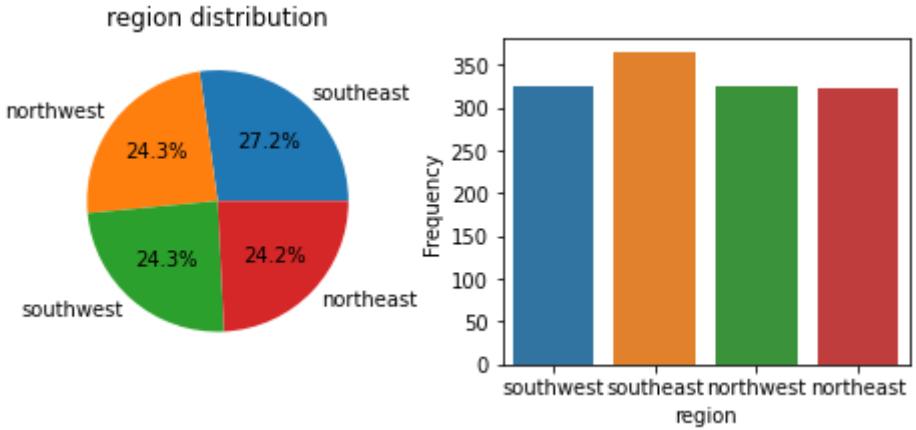
```
The value counts for all categorical columns:

Column :-  sex
male      676
female    662
Name: sex, dtype: int64
```



```
Column :-  smoker
no     1064
yes     274
Name: smoker, dtype: int64
```




```
Column :-  region
southeast    364
northwest    325
southwest    325
northeast    324
Name: region, dtype: int64
```



Insights

- Men and women are almost equal in number
- 20% of patients are smokers
- Patients are distributed across all the regions of Delhi(as said in data) but slightly higher in south-east part

Numerical columns stats and vizualization.

In [952]:
```
df.describe().T
```

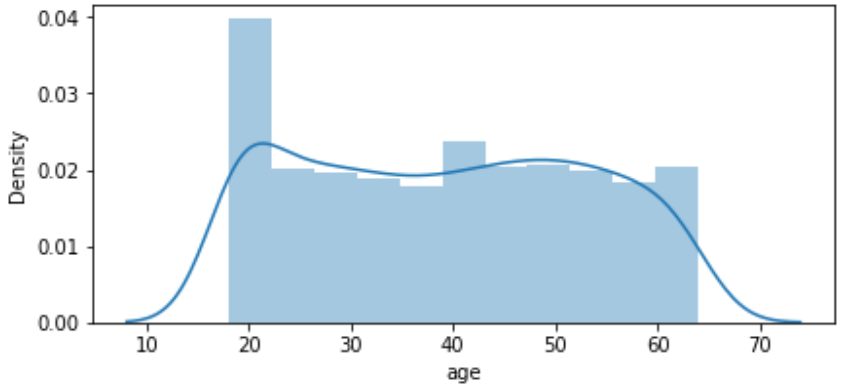
Out[952]:

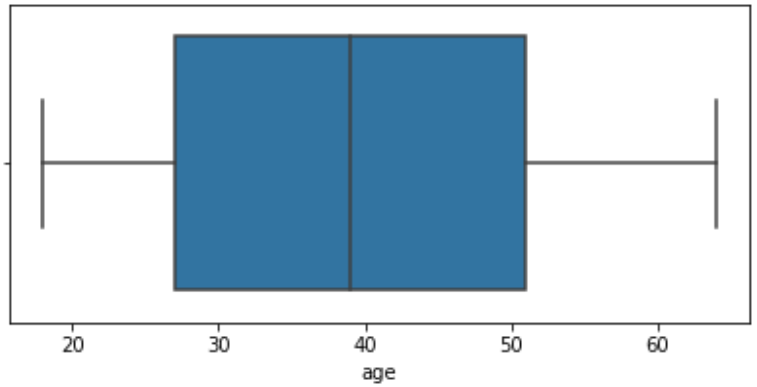
	count	mean	std	min	25%	50%	75%	max
age	1338.0	39.207025	14.049960	18.00	27.0000	39.00	51.0000	64.00
viral load	1338.0	10.221233	2.032796	5.32	8.7625	10.13	11.5675	17.71
severity level	1338.0	1.094918	1.205493	0.00	0.0000	1.00	2.0000	5.00
hospitalization charges	1338.0	33176.058296	30275.029296	2805.00	11851.0000	23455.00	41599.5000	159426.00

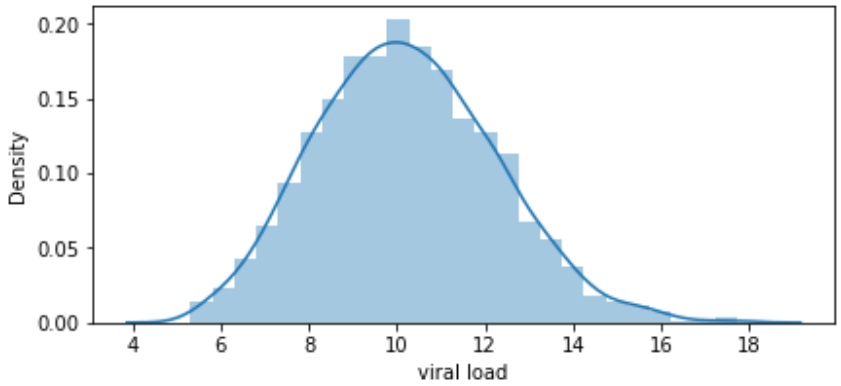
In [953]:

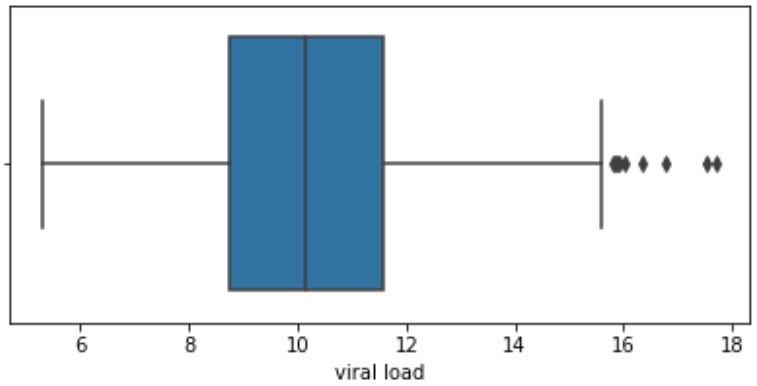
```
Numerical_Cols = df.select_dtypes(exclude = ['object','category']).columns.to_list()
print('Numerical Columns : ' + ', '.join(num_cols))

for col in num_cols:
    plt.figure(figsize = (15,3))
    plt.subplot(121)
    sns.distplot(df[col])
    plt.subplot(122)
    sns.boxplot(df[col], )
    plt.show()
```

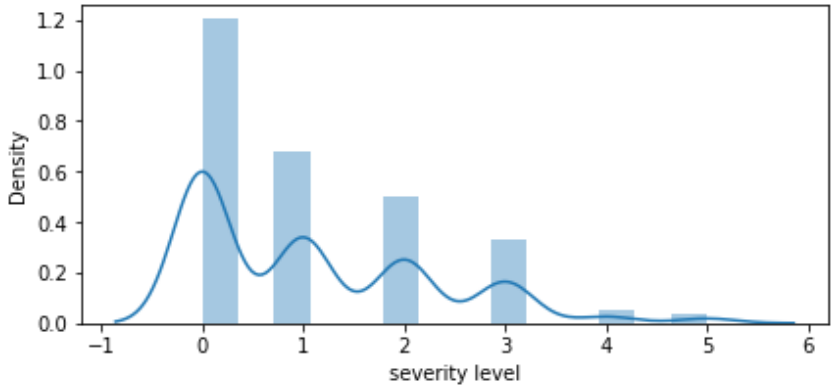
```
Numerical Columns : age, viral load, severity level, hospitalization charges
```

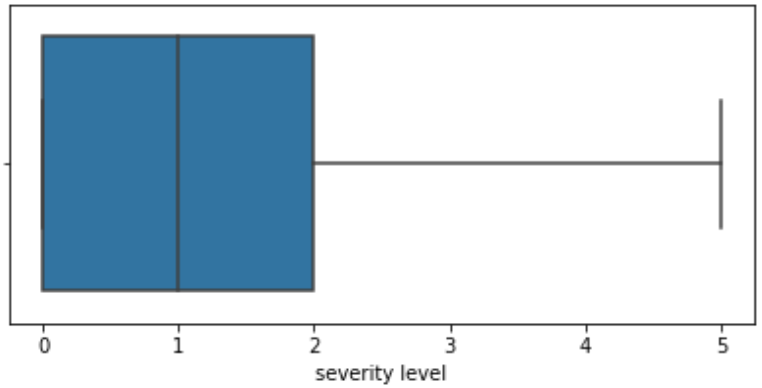


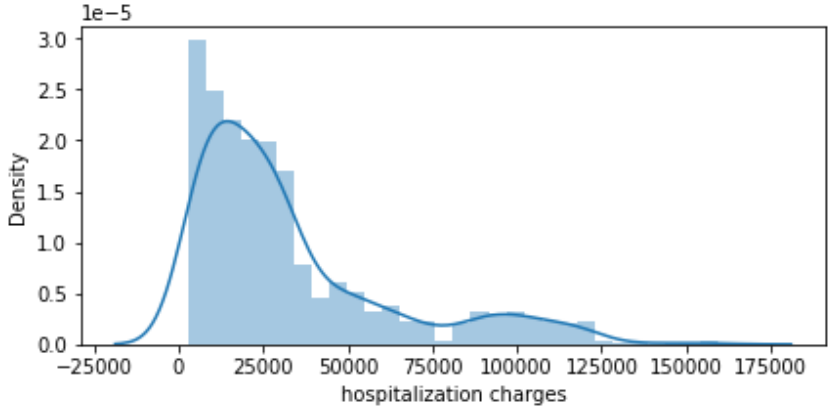


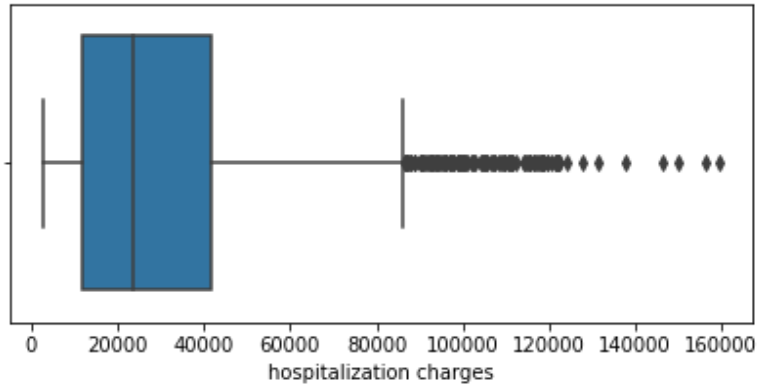












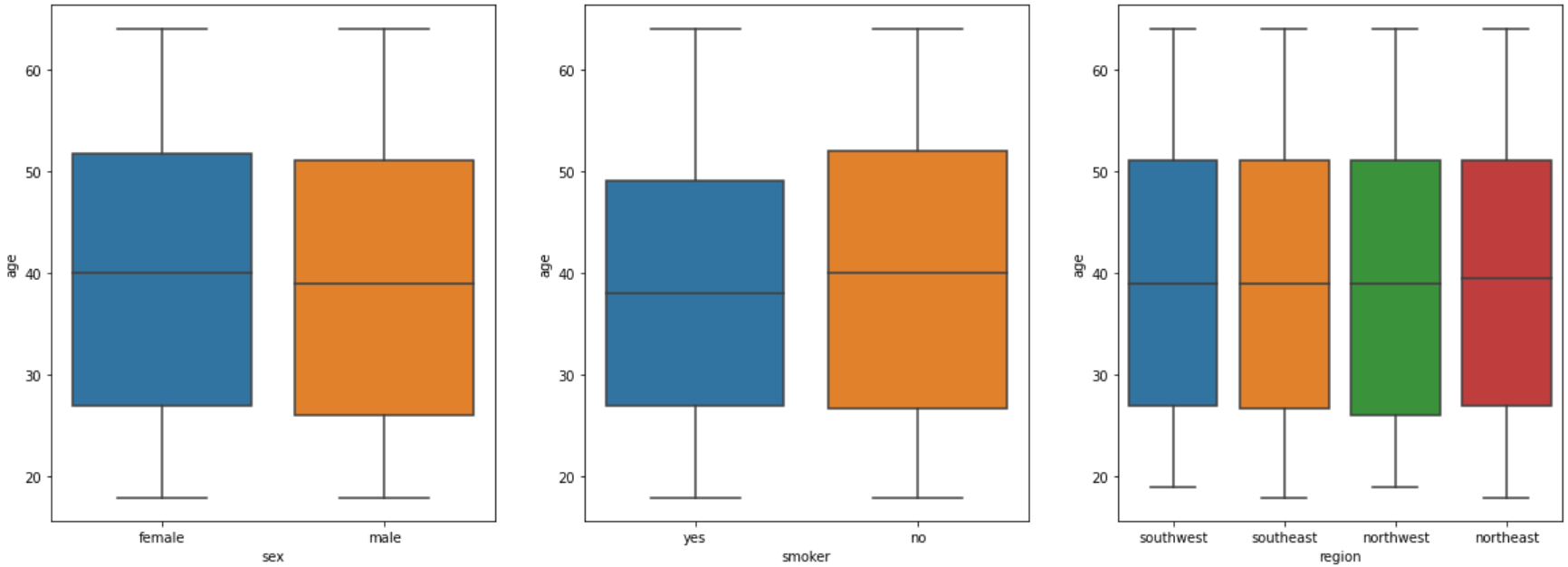
Insights

- The average `age` of the patient is ~ 39
- avg `viral load` is ~10 and `severity level` is 1
- `age` and `severity level` doesn't have outliers and avg age of people hospitalized is 39 years
- `viral load` and `hospitalization charges` have less and lot of outliers respectively

Bi-variate Analysis

In [954]:

```
plt.figure(figsize=(20,7))
for i, Cat_col in enumerate(Categorical_Cols):
    plt.subplot(1,3,i+1)
    sns.boxplot(x=Cat_col, y='age', data=df)
plt.show()
```

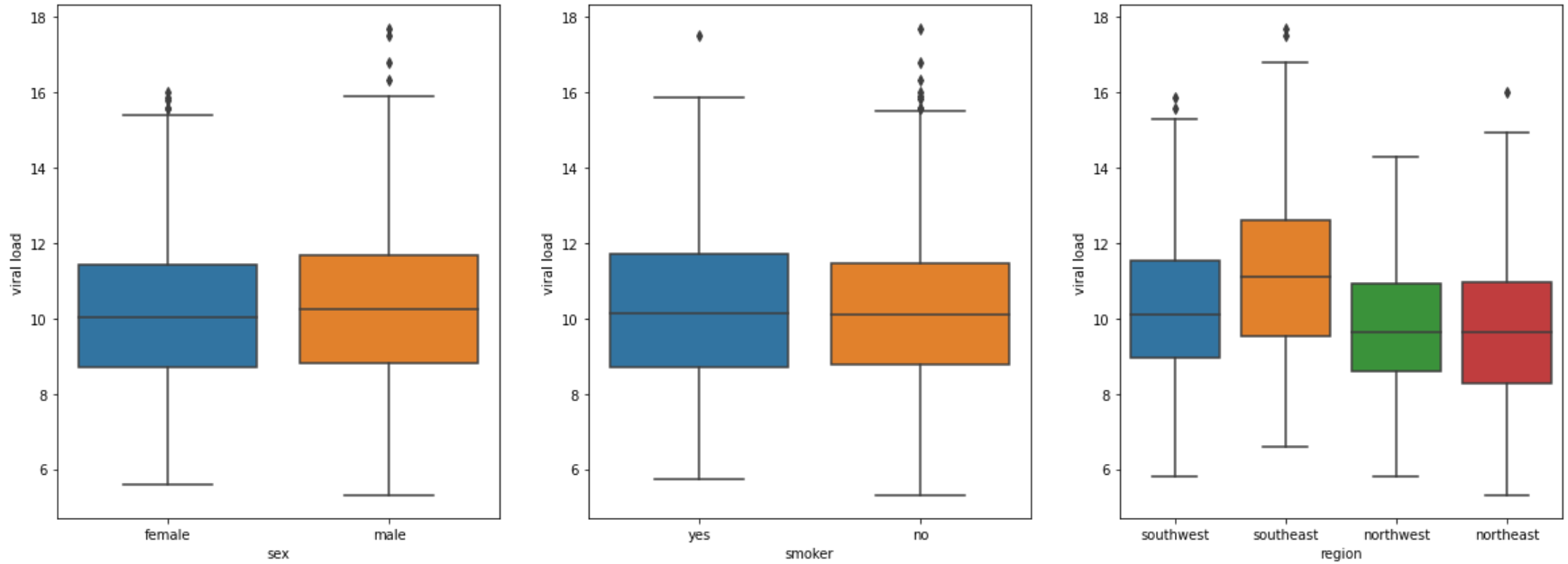


Insights

- The avg age of the male is ~39.5 and of female patients is ~39
- The avg age of smokers and non-smokers is ~40 and ~39 respectively
- The avg age of patients across all the regions is around 40

In [955]:

```
plt.figure(figsize=(20,7))
for i, Cat_col in enumerate(Categorical_Cols):
    plt.subplot(1,3,i+1)
    sns.boxplot(x=Cat_col, y='viral load', data=df)
plt.show()
```

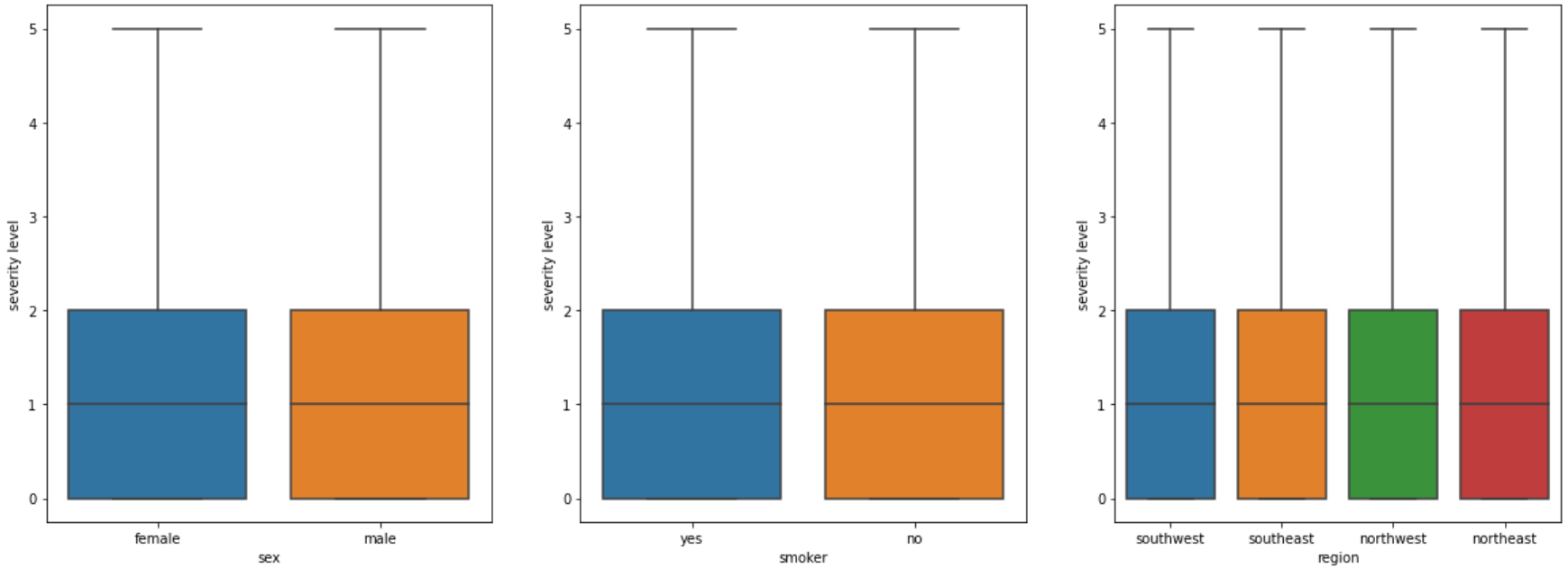


Insights

- `Viral load` between men and women patients are almost the same
- `Viral load` between smokers and non-smokers are also same
- Interestingly the `Viral load` in south-east patients is slightly high, followed by south-west region patients

In [956]:

```
plt.figure(figsize=(20,7))
for i, Cat_col in enumerate(Categorical_Cols):
    plt.subplot(1,3,i+1)
    sns.boxplot(x=Cat_col, y='severity level', data=df)
plt.show()
```

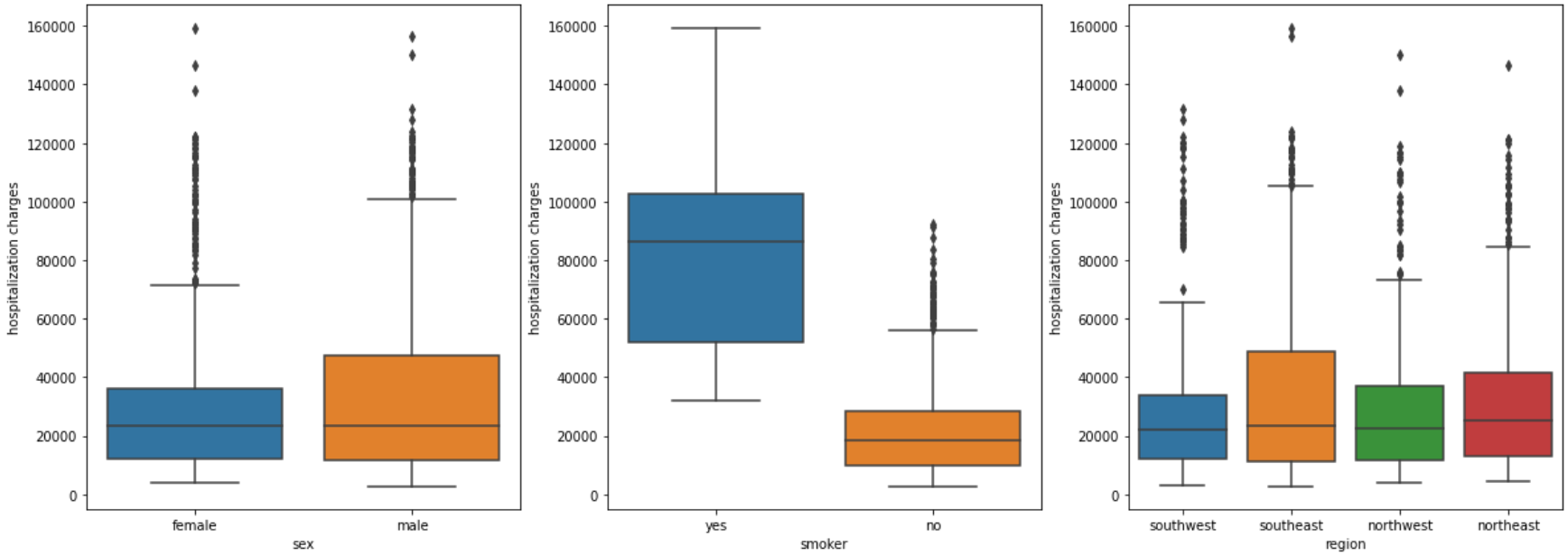


Insights

- `Severity level` seem independent of sex, smoking habits and regions as per the observations

In [957]:

```
plt.figure(figsize=(20,7))
for i, Cat_col in enumerate(Categorical_Cols):
    plt.subplot(1,3,i+1)
    sns.boxplot(x=Cat_col, y='hospitalization charges', data=df)
plt.show()
```



Insights

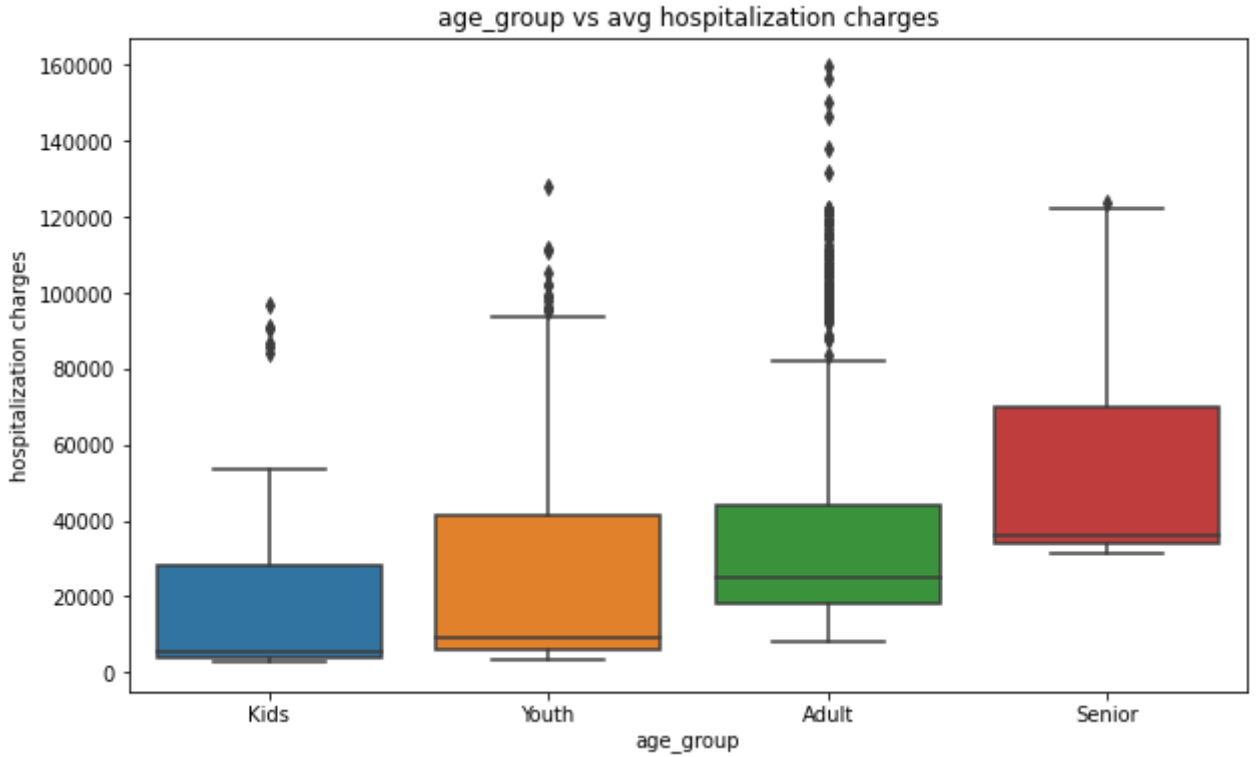
- Avg `hospitalization charges` of both men and women are almost the same
- Avg `hospitalization charges` is significantly greater in smokers compared to non-smokers
- avg `hospitalization charges` of patients from all the regions also almost the same

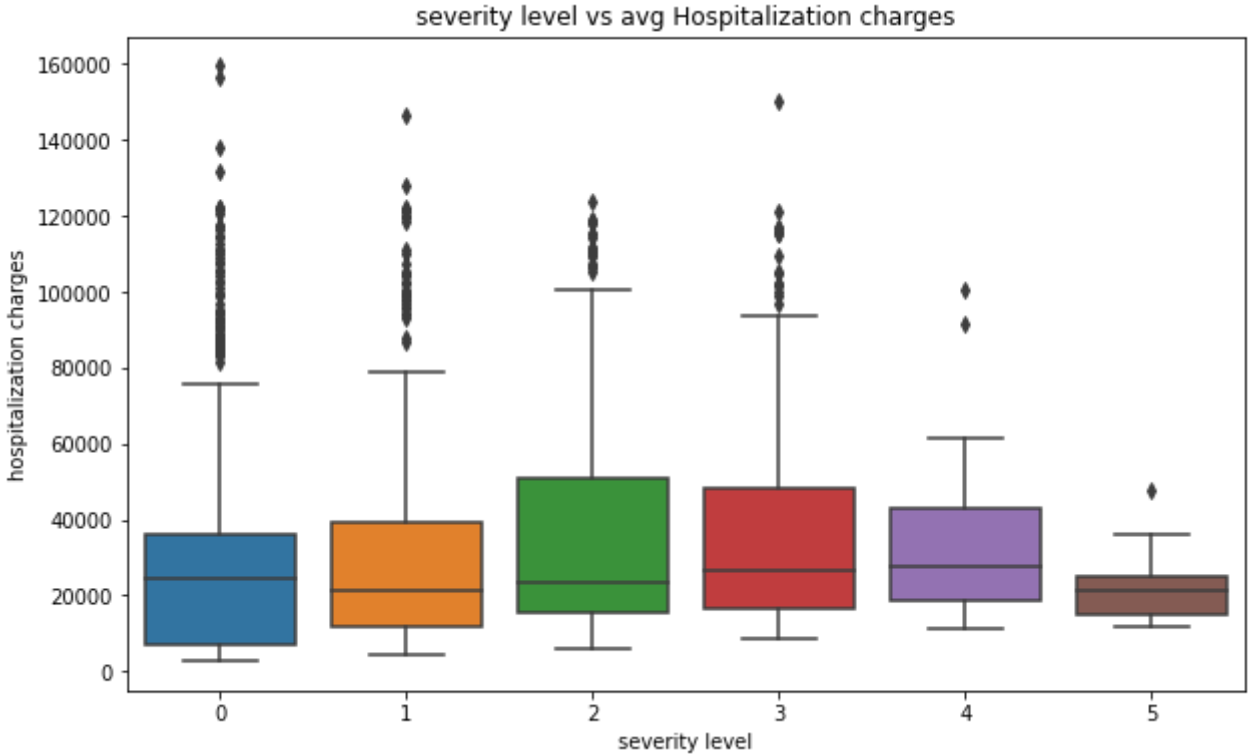
In [958]:

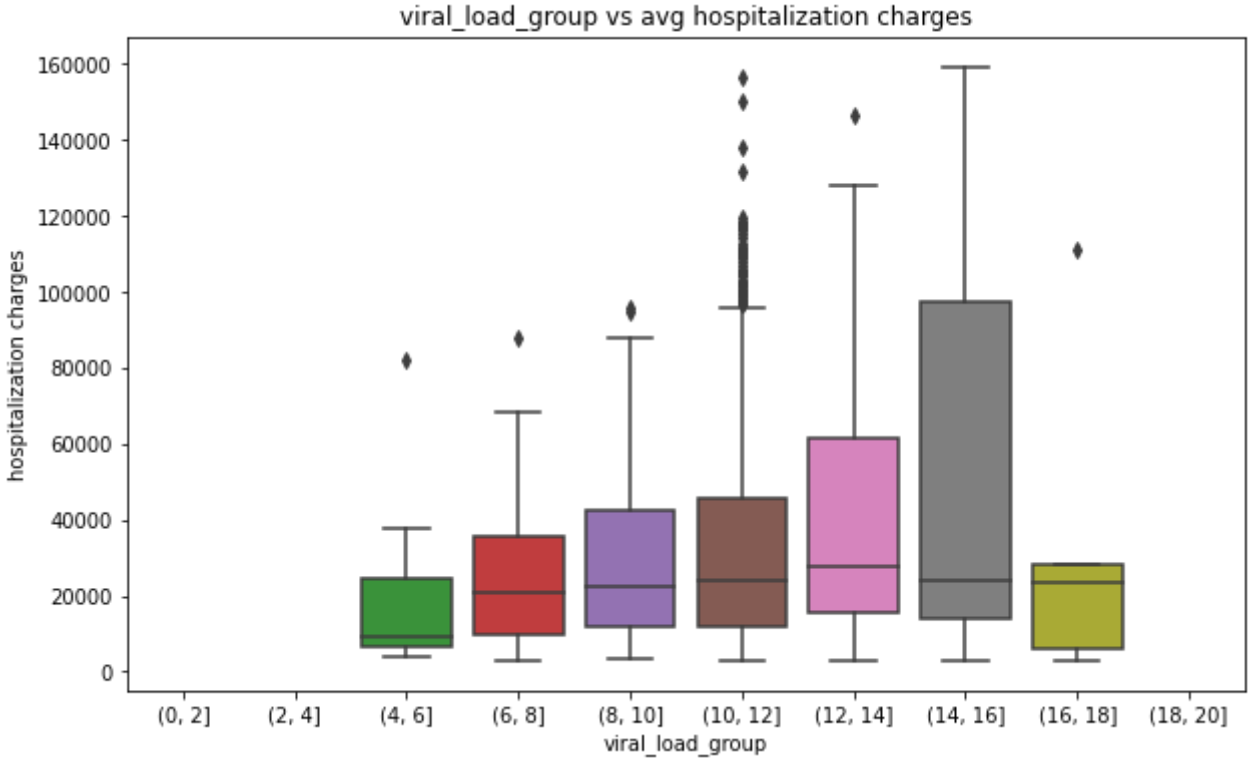
```
# for i, Num_col in enumerate(Numerical_Cols):
#     plt.figure(figsize=(20,8))
#     for j, Cat_col in enumerate(Categorical_Cols):
#         plt.subplot(1,3,j+1)
#         sns.boxplot(x=Cat_col, y=Num_col, data=df)
#     plt.show()
```

In [959]:

```
df['age_group'] = pd.cut(df['age'], bins=[0, 18, 30, 60, 100], labels=['Kids', 'Youth', 'Adult', 'Senior'])
df['viral_load_group'] = pd.cut(df['viral load'], bins=[i for i in range(0, 21, 2)])

plt.figure(figsize = (10,6))
sns.boxplot(data = df, x = 'age_group', y = 'hospitalization charges')
plt.title('age_group vs avg hospitalization charges')

plt.figure(figsize = (10,6))
sns.boxplot(data = df, x = 'severity level', y = 'hospitalization charges')
plt.title('severity level vs avg Hospitalization charges')

plt.figure(figsize = (10,6))
sns.boxplot(data = df, x = 'viral_load_group',  y = 'hospitalization charges')
plt.title('viral_load_group vs avg hospitalization charges')
plt.show()
```




severity level vs avg Hospitalization charges
hospitalization charges
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
2
3
4
5
severity level

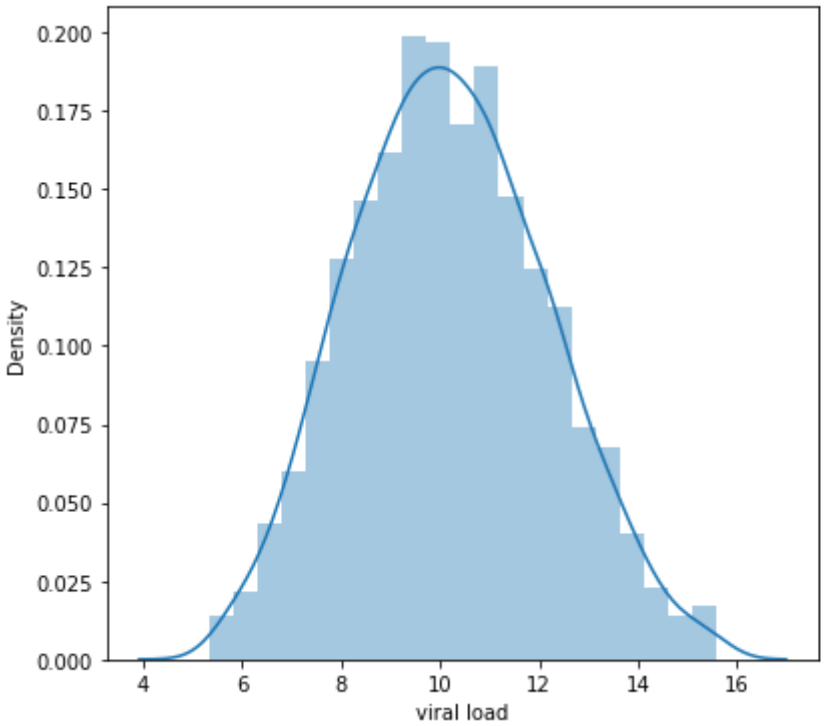


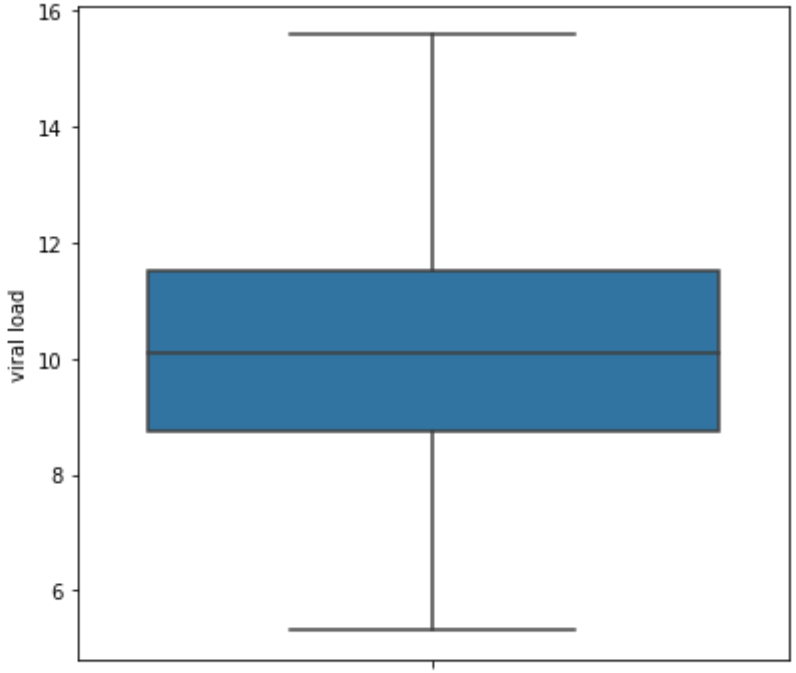
Insights

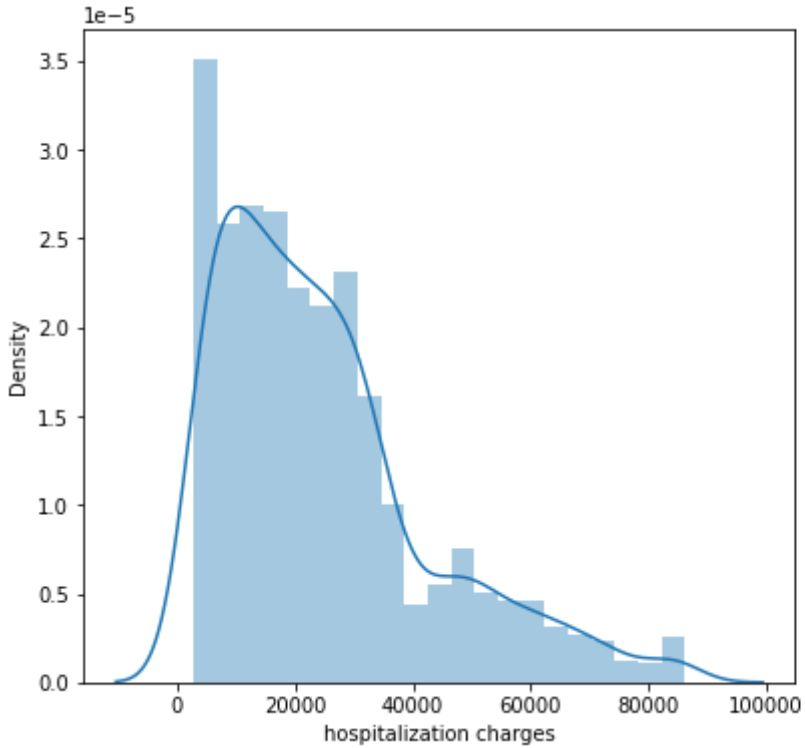
- Seniors are having highest hospitalization charges probably usual elderly health issue, followed by adults with lot of outliers could be because of accidents and sudden health diseases and injuries which can bill for huge hospitalization charges
- Interestingly the hospitalization charges is approximately same across all severity levels, slightly high for '3 Severity level', but significantly lot if outliers
- hospitalization charges are comparatively high in viral load between 10 and 16, but there is not significant difference

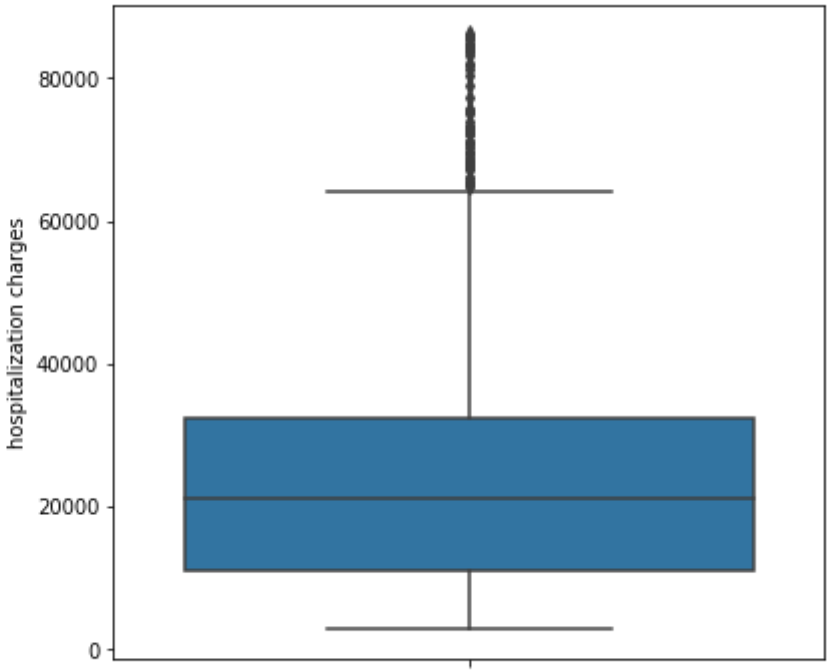
Outlier treatement for `viral load` and `hospitalization charges`

In [960]:
```
for col in ['viral load', 'hospitalization charges']:
    col_values = df[col].values
    q1 = np.quantile(col_values, 0.25)
    q3 = np.quantile(col_values, 0.75)
    IQR = q3 - q1
    without_outliers = list(filter(lambda val: (val>(q1-1.5*IQR)) & (val<(q3+1.5*IQR)), col_values))
    plt.figure(figsize = (14,6))
    plt.subplot(121)
    sns.distplot(without_outliers)
    plt.xlabel(col)
    plt.subplot(122)
    sns.boxplot(y= without_outliers)
    plt.ylabel(col)
    plt.show()
```









- After removing the outliers, we can see that the distribution looks somewhat even compared to what it was earlier.

Hypothesis testing

Q1. Prove (or disprove) that the hospitalization of people who do smoking is greater than those who don't? (T-test Right tailed)

Step 1 : Setting null and alternate hypothesis

H0 : Average hospitalization of smokers and non-smokers are same i.e., $\mu 1 = \mu 2$

H1 : Average hospitalization of smokers is greater than non-smokers i.e., $\mu 1 > \mu 2$

- Significance level = 0.05

In [961]:
```
alpha = 0.05
```

Step 2 : Select the appropriate test

Since we don't have the population SD, we can perform the T-test. and specially we will proceed with right tailed as deciding z-value of H0 would appear right side of sampling mean distribution.

In [962]:
```
# to perform the test on even number of samples
df.groupby('smoker')['hospitalization charges'].count()
```

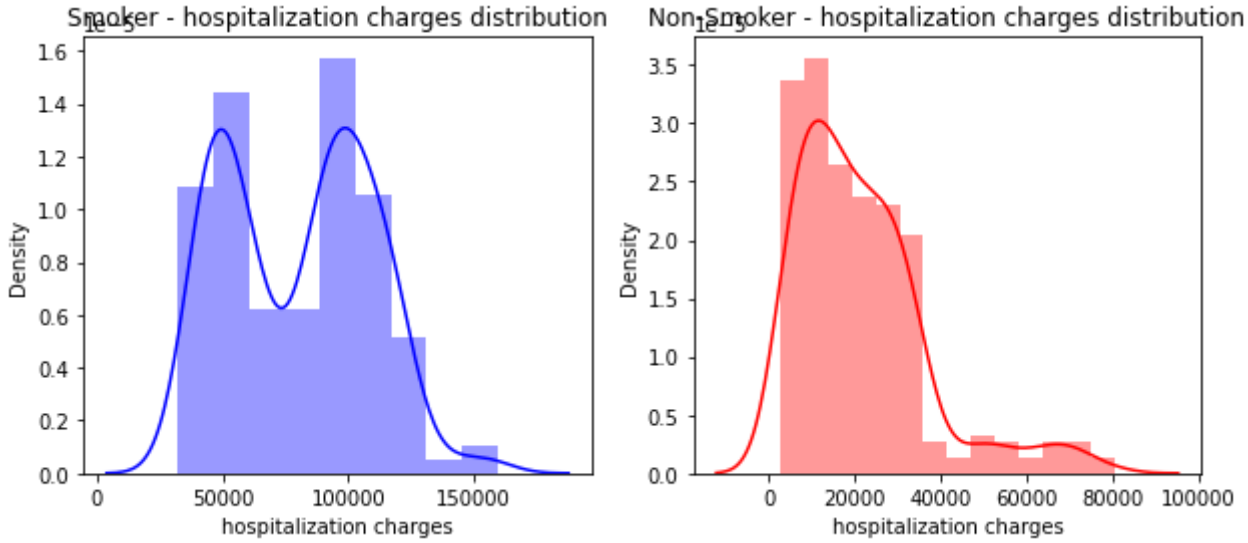
Out[962]:
```
smoker
no     1064
yes     274
Name: hospitalization charges, dtype: int64
```

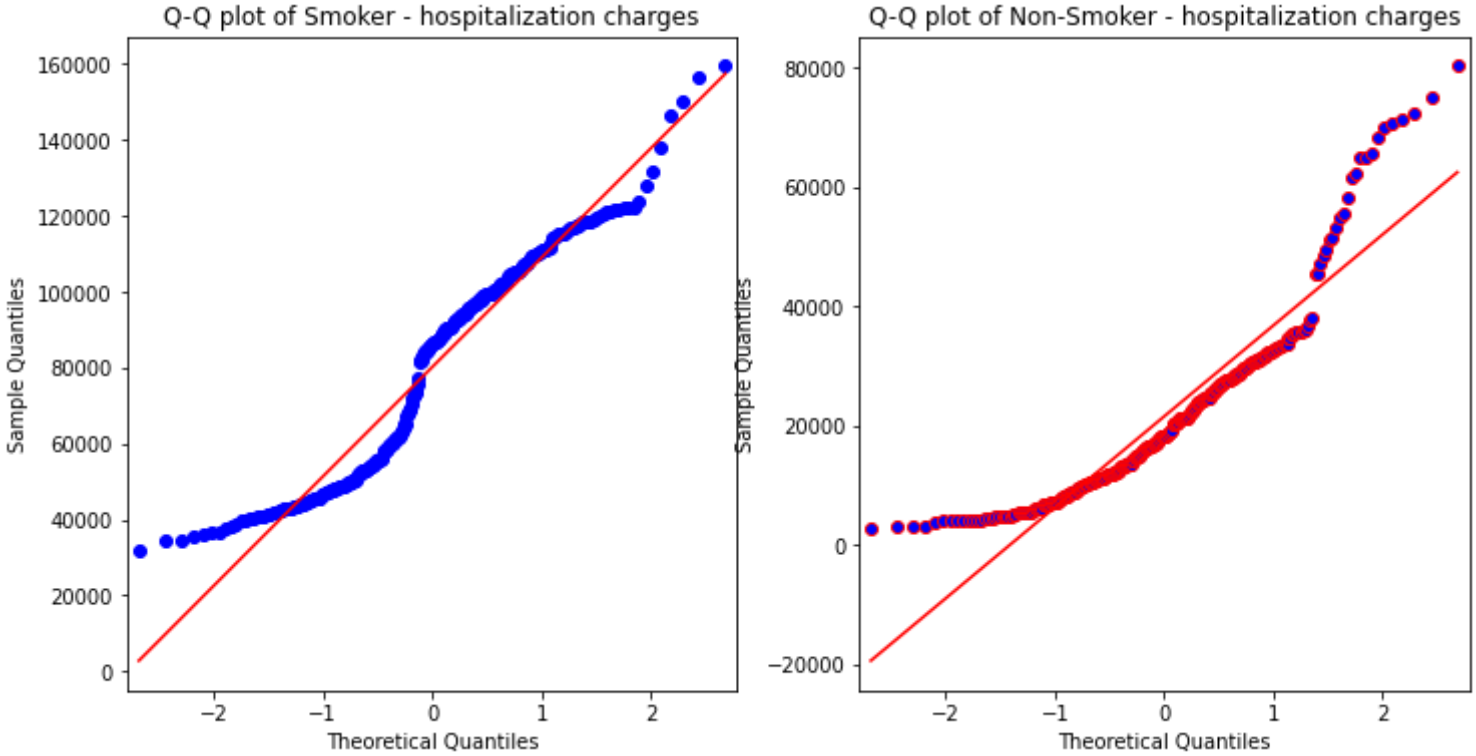
Step 3 : Checking the assumptions before doing the test

- Checking the assumptions like normality and equal variance through vizualization

In [963]:

```
smoker = df[df['smoker'] == 'yes']['hospitalization charges'].sample(274)
non_smoker = df[df['smoker'] == 'no']['hospitalization charges'].sample(274)

plt.figure(figsize=(10,4))
plt.subplot(1,2,1).set_title('Smoker - hospitalization charges distribution')
sns.distplot(smoker,  color = 'b')

plt.subplot(1,2,2).set_title('Non-Smoker - hospitalization charges distribution')
sns.distplot(non_smoker, color = 'r')

fig, ax = plt.subplots(1, 2, figsize=(12, 6))
sm.qqplot(smoker, line='s', ax=ax[0], color = 'b')
ax[0].set_title('Q-Q plot of Smoker - hospitalization charges')

sm.qqplot(non_smoker, line='s', ax=ax[1], color = 'r')
ax[1].set_title('Q-Q plot of Non-Smoker - hospitalization charges')

plt.show()
```





- Checking the assumptions like normality and equal variance through statistical methods

In [964]:

```
# Check normality using Shapiro-Wilk test
# H0 : data follows normal distribution
# H1 : data does not follow normal distribution
stat, p = shapiro(np.concatenate([smoker, non_smoker]))

print('Shapiro-Wilk test statistic: %.4f, p-value: %.4f' % (stat, p))
if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('Failed to reject H0. The data is normally distributed.')
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('Reject H0, hence accept H1. The data is not normally distributed.')

# Check equal variance using Levene's test
# H0 : variences of two sample data are same
# H1 : variences of two sample data are not same
stat, p = levene(smoker, non_smoker)
print('\nLevene test statistic: %.4f, p-value: %.4f' % (stat, p))
if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('Failed to reject H0. The variances are equal.')
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('Reject H0, hence accept H1. The variances are not equal.')
```

```
Shapiro-Wilk test statistic: 0.9162, p-value: 0.0000
Since p value is lesser than the significance level(alpha) ie., (0.0000 < 0.0500)
Reject H0, hence accept H1. The data is not normally distributed.

Levene test statistic: 158.5232, p-value: 0.0000
Since p value is lesser than the significance level(alpha) ie., (0.0000 < 0.0500)
Reject H0, hence accept H1. The variances are not equal.
```

Note : Though vizualizations and statistical methods saying the data samples are not normally distrubuted and their variences are non-homogenous, we are still proceding on the hypothesis testing

Step 4 : Find the p-value

In [965]:
```
stat, p = ttest_ind(non_smokers, smokers)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('We failed to reject H0 : Average hospitalization of smokers and non-smokers are same i.e., μ1 = μ2')
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('We reject H0, hence accept H1 : Average hospitalization of smokers is greater than non-smokers i.e., μ1
```

```
stat=-46.665, p=0.000
Since p value is lesser than the significance level(alpha) ie., (0.0000 < 0.0500)
We reject H0, hence accept H1 : Average hospitalization of smokers is greater than non-smokers i.e., μ1 > μ2
```

Step 5 : Conclusion

- We proved that Average hospitalization of smokers is greater than non-smokers

Q2. Prove (or disprove) with statistical evidence that the viral load of females is different from that of males (T-test Two tailed)

Step 1 : Setting null and alternate hypothesis

H0 : Average viral load between female patients and male patients is same i.e., μ1 = μ2

H1 : Average viral load between female patients and male patients is different i.e., μ1 ≠ μ2

- Significance level = 0.05

In [966]:
```
alpha = 0.05
```

Step 2 : Select the appropriate test

Since we don't have the population SD, we can perform the T-test. and specially we will proceed two tailed right tailed as deciding z-value of H0 would appear on either side of sampling mean distribution.

In [967]:
```
# to perform the test on even number of samples
df.groupby('sex')['viral load'].count()
```

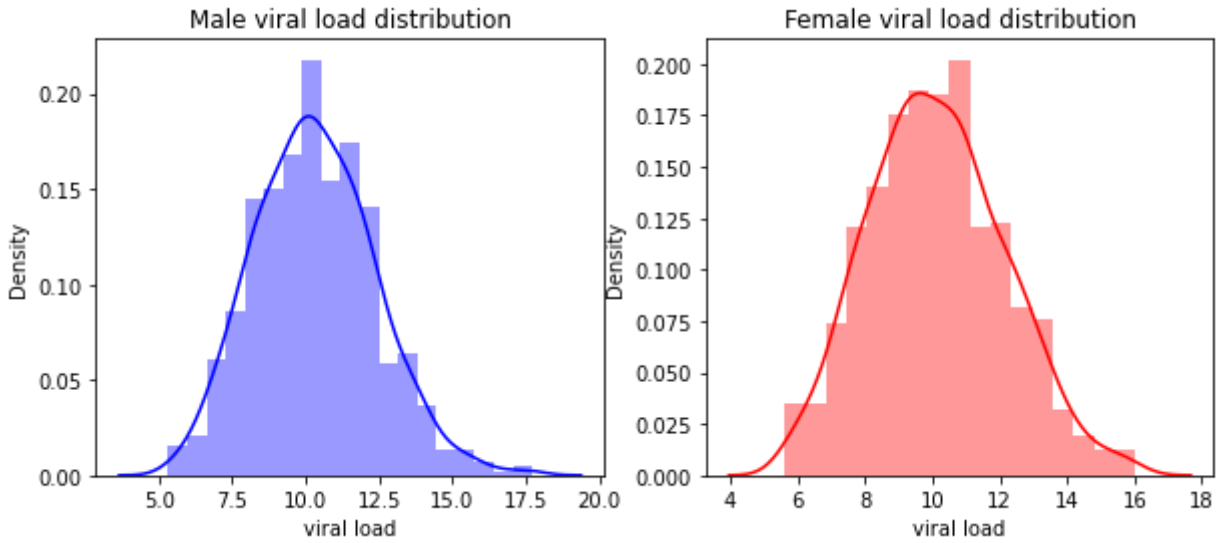
Out[967]:
```
sex
female    662
male      676
Name: viral load, dtype: int64
```

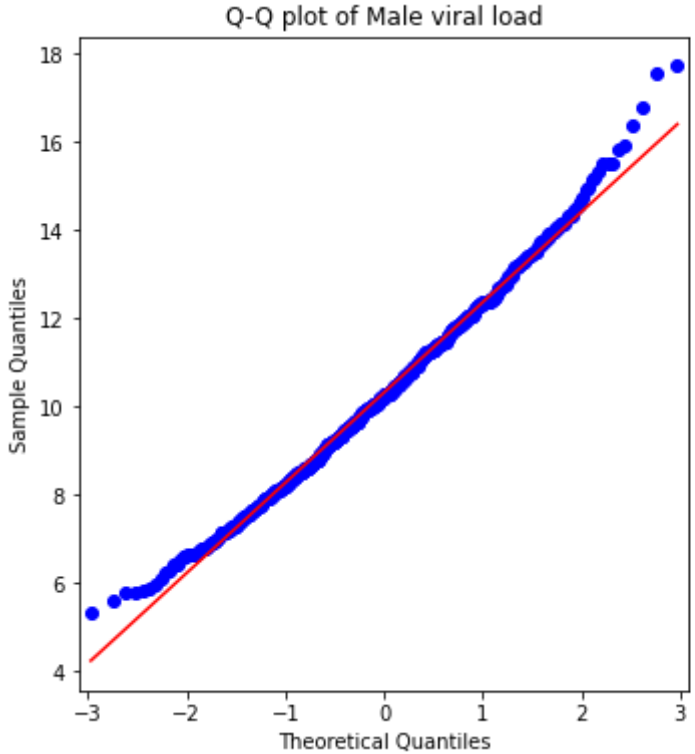
Step 3 : Checking the assumptions before doing the test

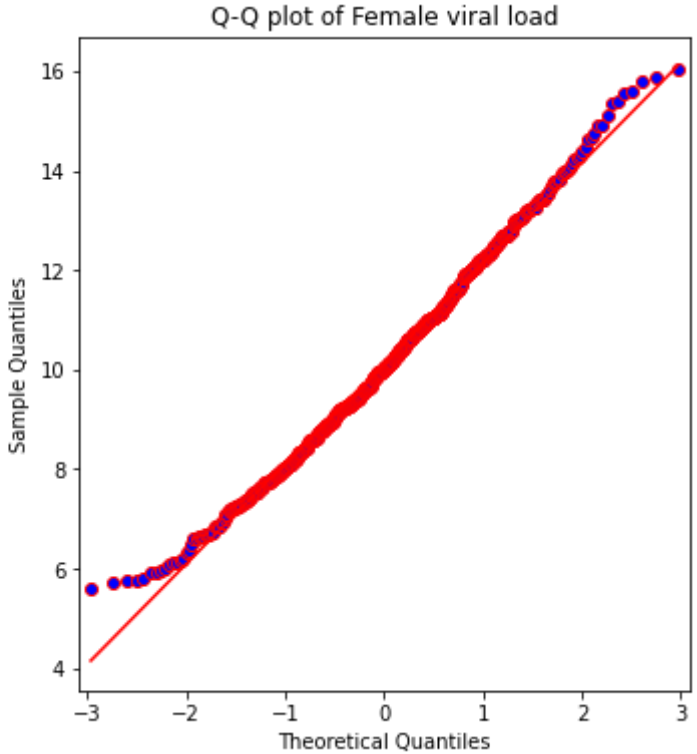
- Checking the assumptions like normality and equal variance through vizualization

In [968]:

```
male = df[df['sex'] == 'male']['viral load']
female = df[df['sex'] == 'female']['viral load']

plt.figure(figsize=(10,4))
plt.subplot(1,2,1)
sns.distplot(male,  color = 'b')
plt.title('Male viral load distribution')

plt.subplot(1,2,2)
sns.distplot(female, color = 'r')
plt.title('Female viral load distribution')

fig, ax = plt.subplots(1, 2, figsize=(12, 6))
sm.qqplot(male, line='s', ax=ax[0], color = 'b')
ax[0].set_title('Q-Q plot of Male viral load')

sm.qqplot(female, line='s', ax=ax[1], color = 'r')
ax[1].set_title('Q-Q plot of Female viral load')
plt.show()
```







- Checking the assumptions like normality and equal variance through statistical methods

In [969]:

```
# Check normality using Shapiro-Wilk test
# H0 : data follows normal distribution
# H1 : data does not follow normal distribution
stat, p = shapiro(np.concatenate([male, female]))

print('Shapiro-Wilk test statistic: %.4f, p-value: %.4f' % (stat, p))
if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('Failed to reject H0. The data is normally distributed.')
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('Reject H0, hence accept H1. The data is not normally distributed.')

# Check equal variance using Levene's test
# H0 : variences of two sample data are same
# H1 : variences of two sample data are not same
stat, p = levene(smoker, non_smoker)
print('\nLevene test statistic: %.4f, p-value: %.4f' % (stat, p))
if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('Failed to reject H0. The variances are equal.')
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('Reject H0, hence accept H1. The variances are not equal.')
```

```
Shapiro-Wilk test statistic: 0.9939, p-value: 0.0000
Since p value is lesser than the significance level(alpha) ie., (0.0000 < 0.0500)
Reject H0, hence accept H1. The data is not normally distributed.

Levene test statistic: 158.5232, p-value: 0.0000
Since p value is lesser than the significance level(alpha) ie., (0.0000 < 0.0500)
Reject H0, hence accept H1. The variances are not equal.
```

Note : Though vizualizations and statistical methods saying the data samples are not normally distrubuted and their variences are non-homogenous, we are still proceding on the hypothesis testing

Step 4 : Find the p-value

In [970]:
```
stat, p = ttest_ind(male, female)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('Failed to reject H0 : Average viral load between female patients and male patients is same i.e., μ1 = μ
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('Reject H0, hence accept H1 : Average viral load between female patients and male patients is different
```

```
stat=1.696, p=0.090
Since p value is greater than the significance level(alpha) ie., (0.0902 > 0.0500)
Failed to reject H0 : Average viral load between female patients and male patients is same i.e., μ1 = μ2
```

Step 5 : Conclusion

- We disproved that Average viral load between female patients and male patients are different. Hence they are similar

Q3. Is the proportion of smoking significantly different across different regions? (Chi-square)

Step 1 : Setting null and alternate hypothesis

H0 : proportion of smoking does not defers on the regions

H1 : proportion of smoking defers on the regions

- Significance level = 0.05

In [971]:
```
alpha = 0.05
```

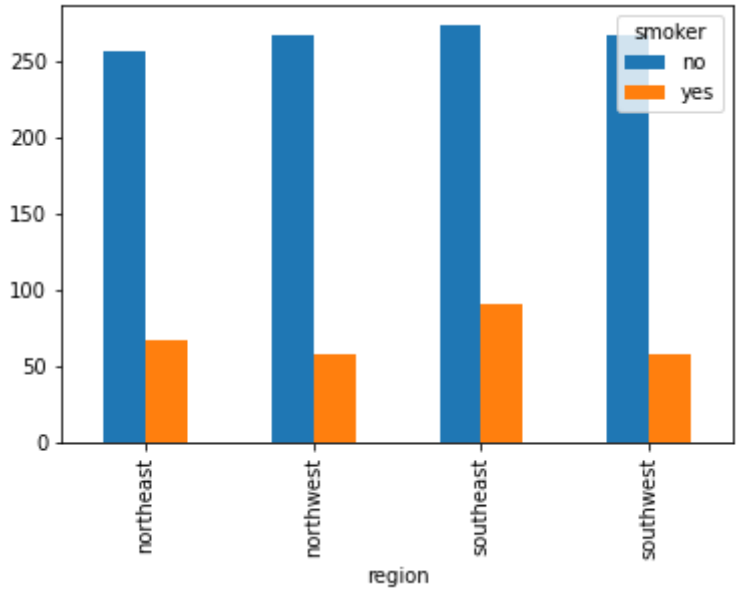
Step 2 : Select the appropriate test

Since smoke column and region are categorical column, and we wanted to check the spread or relativity among the regions we can perform the Chi-square teest

Step 2 : Checking the assumptions before doing the test

- data of contingency table should be independent and each cell should be more than 25

In [972]:
```
contingency_table = pd.crosstab(df['region'], df['smoker'])
contingency_table.plot(kind = 'bar')
plt.show()
contingency_table
```



Out[972]:

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

Step 4 : Find the p-value

In [973]:
```
chi2, p, dof, exp_freq = chi2_contingency(contingency_table)
print('chi-square statistic: %.4f, P-value: %.4f, Degree of freedom: %d, expected frequencies: \n%s \n' % (chi2, p

if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('Failed to reject H0 : proportion of smoking does not defers on the regions')
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('Reject H0, hence accept H1 : proportion of smoking does not defers on the regions')
```

```
chi-square statistic: 7.3435, P-value: 0.0617, Degree of freedom: 3, expected frequencies:
[[257.65022422  66.34977578]
 [258.44544096  66.55455904]
 [289.45889387  74.54110613]
 [258.44544096  66.55455904]]

Since p value is greater than the significance level(alpha) ie., (0.0617 > 0.0500)
Failed to reject H0 : proportion of smoking does not defers on the regions
```

Step 5 : Conclusion

- We disproved that the proportion of smoking is significantly different across different regions. Hence, proportion of smoking does not depend on the regions

Q4. 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)

Step 1 : Setting null and alternate hypothesis

H0 : Viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level are same i.e., $\mu 0 = \mu 1 = \mu 2$

H1 : Viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level are not same i.e., $\mu 0 \neq \mu 1 \neq \mu 2$

- Significance level = 0.05

In [974]:
```
alpha = 0.05
```

Step 2 : Select the appropriate test

Since we need to test the means of multiple groups, we can perform one way ANOVA

Step 3 : Checking the assumptions before doing the test

- Checking the assumptions like normality and equal variance through vizualization

In [975]:
```
female_sev_df = df[(df['sex'] == 'female') & (df['severity level'] <=2)]
female_sev_df
```

Out[975]:

	age	sex	smoker	region	viral load	severity level	hospitalization charges	age_group	viral_load_group
0	19	female	yes	southwest	9.30	0	42212	Youth	(8, 10]
5	31	female	no	southeast	8.58	0	9392	Adult	(8, 10]
6	46	female	no	southeast	11.15	1	20601	Adult	(10, 12]
9	60	female	no	northwest	8.61	0	72308	Adult	(8, 10]
11	62	female	yes	southeast	8.76	0	69522	Senior	(8, 10]
...	...	...	...	...	...	...	...	...	...
1331	23	female	no	southwest	11.13	0	26990	Youth	(10, 12]
1334	18	female	no	northeast	10.64	0	5515	Kids	(10, 12]
1335	18	female	no	southeast	12.28	0	4075	Kids	(12, 14]
1336	21	female	no	southwest	8.60	0	5020	Youth	(8, 10]
1337	61	female	yes	northwest	9.69	0	72853	Senior	(8, 10]

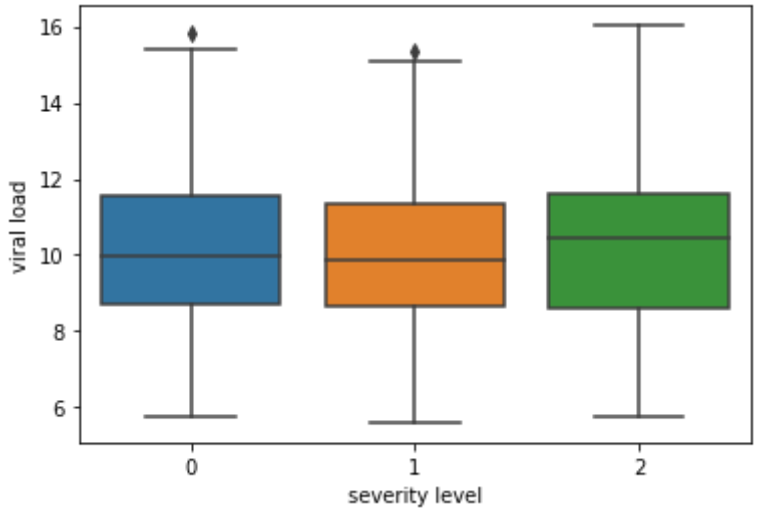
566 rows × 9 columns

In [976]:
```
female_sev_df.groupby('severity level')['viral load'].describe()
```

Out[976]:

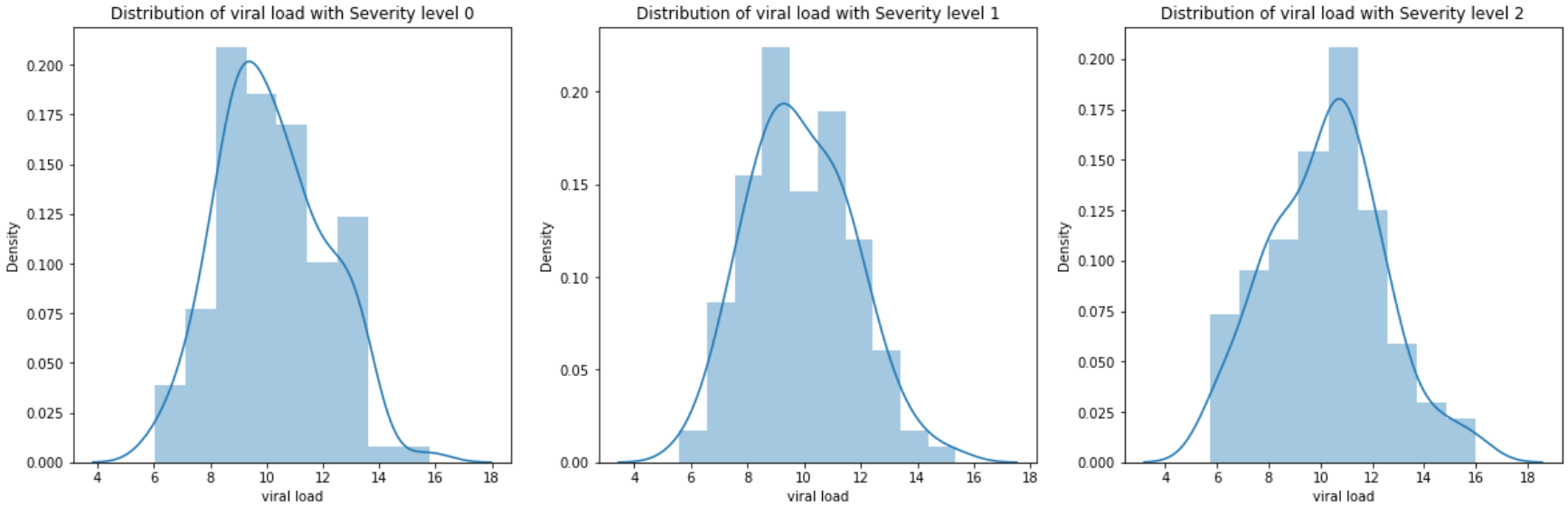
severity level	count	mean	std	min	25%	50%	75%	max
0	289.0	10.120727	1.989071	5.76	8.7300	9.980	11.530	15.80
1	158.0	10.017468	1.929065	5.60	8.6575	9.855	11.315	15.36
2	119.0	10.216807	2.209687	5.73	8.5900	10.430	11.585	16.02

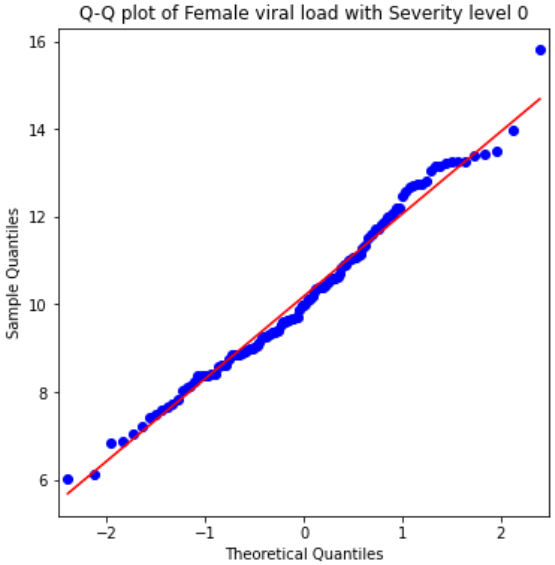
In [977]:
```
sns.boxplot(data = female_sev_df, x = 'severity level', y= 'viral load')
plt.show()
```

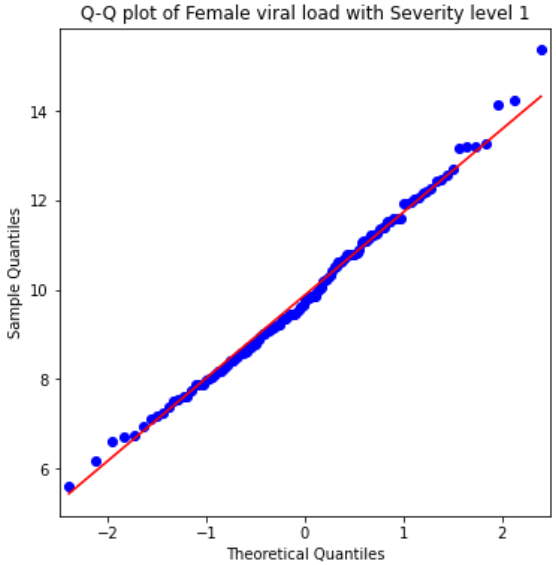


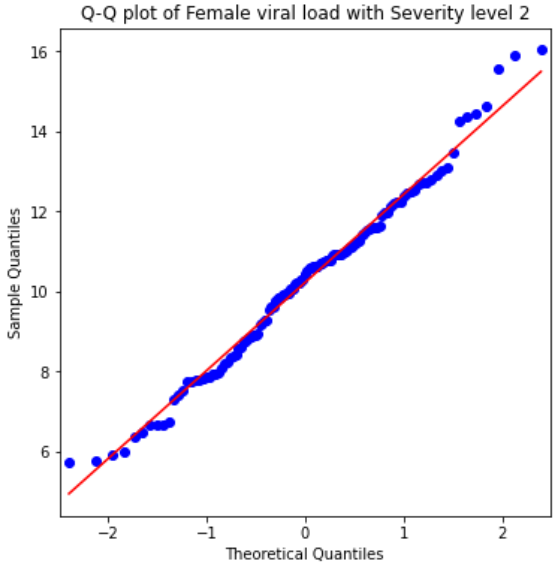
In [978]:

```
# female severity levels of severity of 0,1,2 => fs = [[viral-load where severity = 0], [viral-load where severity
fs = [female_sev_df[female_sev_df['severity level']==i]['viral load'].sample(119) for i in range(0,3)]
# print(np.array(fs).shape) # >> (3, 119)

plt.figure(figsize=(20,6))
for i in range(0,3):
    plt.subplot(1,3,i+1).set_title('Distribution of viral load with Severity level %d' % (i))
    sns.distplot(fs[i])
plt.show()

fig, ax = plt.subplots(1, 3, figsize=(20, 6))
for i in range(0,3):
    sm.qqplot(fs[i], line='s', ax=ax[i])
    ax[i].set_title('Q-Q plot of Female viral load with Severity level %d' % (i))
plt.show()
```




Q-Q plot of Female viral load with Severity level 0
Sample Quantiles
Theoretical Quantiles
16
14
12
10
8
6
−2
−1
0
1
2


Q-Q plot of Female viral load with Severity level 1
Sample Quantiles
Theoretical Quantiles
14
12
10
8
6
−2
−1
0
1
2


Q-Q plot of Female viral load with Severity level 2
Sample Quantiles
Theoretical Quantiles
16
14
12
10
8
6
−2
−1
0
1
2

In [979]:

```
# Check normality using Shapiro-Wilk test
# H0 : data follows normal distribution
# H1 : data does not follow normal distribution
stat, p = shapiro(np.concatenate(fs))

print('Shapiro-Wilk test statistic: %.4f, p-value: %.4f' % (stat, p))
if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('Failed to reject H0. The data is normally distributed.')
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('Reject H0, hence accept H1. The data is not normally distributed.')

# Check equal variance using Levene's test
# H0 : variences of two sample data are same
# H1 : variences of two sample data are not same
stat, p = levene(fs[0], fs[1], fs[2])
print('\nLevene test statistic: %.4f, p-value: %.4f' % (stat, p))
if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('Failed to reject H0. The variances are equal.')
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('Reject H0, hence accept H1. The variances are not equal.')
```

```
Shapiro-Wilk test statistic: 0.9928, p-value: 0.0863
Since p value is greater than the significance level(alpha) ie., (0.0863 > 0.0500)
Failed to reject H0. The data is normally distributed.

Levene test statistic: 1.4647, p-value: 0.2325
Since p value is greater than the significance level(alpha) ie., (0.2325 > 0.0500)
Failed to reject H0. The variances are equal.
```

The Assumptions are true, hence performing the ANOVA

Step 4 : Find the p-value

In [980]:
```
stat, p = f_oneway(fs[0], fs[1], fs[2])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > alpha:
    print('Since p value is greater than the significance level(alpha) ie., (%.4f > %.4f)' % (p, alpha))
    print('We failed to reject H0 : Viral load of women with 0 Severity level , 1 Severity level, and 2 Severity l
else:
    print('Since p value is lesser than the significance level(alpha) ie., (%.4f < %.4f)' % (p, alpha))
    print('We reject H0, hence accept H1 : Viral load of women with 0 Severity level , 1 Severity level, and 2 Sev
```

```
stat=1.017, p=0.363
Since p value is greater than the significance level(alpha) ie., (0.3629 > 0.0500)
We failed to reject H0 : Viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level are s
ame i.e., μ0 = μ1 = μ2
```

Step 5 : Conclusion

- We proved that the Viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level are same

Business Insights:

1. Through EDA

- Men and women are almost equal in number
- 20% of patients are smokers
- Patients are distributed across all the regions of Delhi(as said in data) but slightly higher in south-east part
- The average age of the patient is ~ 39
- avg viral load is ~10 and severity level is 1
- age and severity level doesn't have outliers and avg age of people hospitalized is 39 years
- viral load and hospitalization charges have less and lot of outliers respectively
- The avg age of the male is ~39.5 and of female patients is ~39
- The avg age of smokers and non-smokers is ~40 and ~39 respectively
- The avg age of patients across all the regions is around 40
- Viral load between men and women patients are almost the same
- Viral load between smokers and non-smokers is also the same
- Interestingly the Viral load in south-east patients is slightly high, followed by south-west region patients

- Severity level seem independent of sex, smoking habits, and regions as per the observations
- Avg hospitalization charges of both men and women are almost the same
- Avg hospitalization charges is significantly greater in smokers compared to non-smokers
- Avg hospitalization charges of patients from all the regions also almost the same
- Seniors are having highest hospitalization charges probably usual elderly health issues, followed by adults with a lot of outliers could be because of accidents and sudden health diseases and injuries which can bill for huge hospitalization charges
- Interestingly the hospitalization charges is approximately the same across all severity levels, slightly high for '3 Severity level', but significantly lot if outliers
- hospitalization charges are comparatively high in viral load between 10 and 16, but there is no significant difference

2. Through Hypothesis testing

- We proved that the Average hospitalization of smokers is greater than non-smokers
- We disproved that the Average viral load between female patients and male patients is different. Hence they are similar
- We disproved that the proportion of smoking is significantly different across different regions. Hence, the proportion of smoking does not depend on the regions
- We proved that the Viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level are same

Recomendations:

- Since hospitalization charges of smokers is high compared to non-smokers, and they are almost 20% of the patients, hospitals can focus on rehab centers and create complementary schemes for the smoking patients there by to grow the business.
- Since women smokers are also there, it is very important to take measures if they are new-age mothers and if we don't take corresponding actions as it could impact the newborn babies.
- Since the hospitalization charges of a severity level of 3 is slightly high, increasing efficiencies like making more accommodations of ICU and wards, and storing and tracking the necessary surgical instruments and medicines to treat such illness will be really helpful. Thereby reducing the cost and time for the treatment of such severity would be profitable.
- Hospitalizations of seniors is also high. This could be of elderly health issues. Making the treatments for treating seniors should be quick, cost-effective, and patient-friendly caring and nursing should be made. More doctors to treat such patients should be timely available.
- Also to treat the seniors, hospitals can also start the initiatives/programs like doctor and nursing facilities like door treatments, which makes it really convenient for the seniors to avail such treatment instead of visiting the hospitals.
- Since hospitalizations are high for higher viral load, necessary steps to increase immunity through supplements and medicines should be promoted. Also treating/nursing faculty treating such patients should be mandatorily skilled in virology and follow certain best protocols to avoid the spread as viral diseases are contagious.