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.

Problem Statement :

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

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

2-sample t-test: testing for difference across populations

ANOVA

Chi-square

In [2]:

```
import matplotlib.pyplot as plt
from matplotlib import figure
import seaborn as sns
import numpy as np
import statsmodels.api as sm
from scipy.stats import norm
from scipy.stats import t
from scipy.stats import binom

import warnings
warnings.filterwarnings('ignore')
```

In [3]:

```
import pandas as pd
df = pd.read_csv("scaler_apollo_hospitals.txt")
df.head()
```

Out[3]:

	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 [4]:

```
## removing Unnamed column
df = df.drop("Unnamed: 0",axis=1)
df.head()
```

Out[4]:

	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 [18]:

```
# Checking the shape of our data set
df.shape
```

Out[18]:

```
(1338, 7)
```

there are 1338 rows and 8 columns in the dataset.

In [19]:

```
# Checking the information of data
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 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: 73.3+ KB
```

we can see clearly there is no null values in any columns.

In [20]:

```
# checking again there is any null value or not with another code
df.isna().sum()
```

Out[20]:

```
age                         0
sex                         0
smoker                      0
region                      0
viral load                  0
severity level              0
hospitalization charges     0
dtype: int64
```

Not Detected null value

In [21]:

```
# Checking for the columns in this data set.
df.columns
```

Out[21]:

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

In [60]:

```
# Checking for the data types
df.dtypes
```

Out[60]:

```
age                          int64
sex                         object
smoker                      object
region                      object
viral load                 float64
severity level               int64
hospitalization charges      int64
dtype: object
```

In [22]:

```
# Checking the unique values in the columns
df.nunique()
```

Out[22]:

```
age                          47
sex                           2
smoker                        2
region                        4
viral load                  462
severity level                6
hospitalization charges    1320
dtype: int64
```

OR

In [23]:

```
df.apply(lambda x : x.nunique())
```

Out[23]:

```
age                          47
sex                           2
smoker                        2
region                        4
viral load                  462
severity level                6
hospitalization charges    1320
dtype: int64
```

here is the info or unique values

In [24]:

```
# Descriptive Statistics
df.describe()
```

Out[24]:

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

Observation -

50% of people are of age 39 and the mean also lies in this .

Severity level is increase as the age increase ,that means older people have the high sever level.and the viral load also increased as the age increase.

age ,viral load and severity level are the major reason for the high hospitalization charges.

mean of hospitalization charges is 33176,age mean is 39, viral load in people mean is 10 and severity is not much higher in mean perspective.

.

In [ ]:

```
1 # Statistical summary of categorical data in our dataset.
```

In [25]:

```
1 df.describe (include = ["object","category"])
```

Out[25]:

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

Observation -

male are given as the highest in the data with no smoking experience and also the frequency is in "Southeast area".

frequency of smokers is 1064.

Corelation metrix

In [26]:

```
# Corelation metrix :

df.corr()
```

Out[26]:

	age	viral load	severity level	hospitalization charges
age	1.000000	0.109300	0.042469	0.299008
viral load	0.109300	1.000000	0.012729	0.198388
severity level	0.042469	0.012729	1.000000	0.067998
hospitalization charges	0.299008	0.198388	0.067998	1.000000

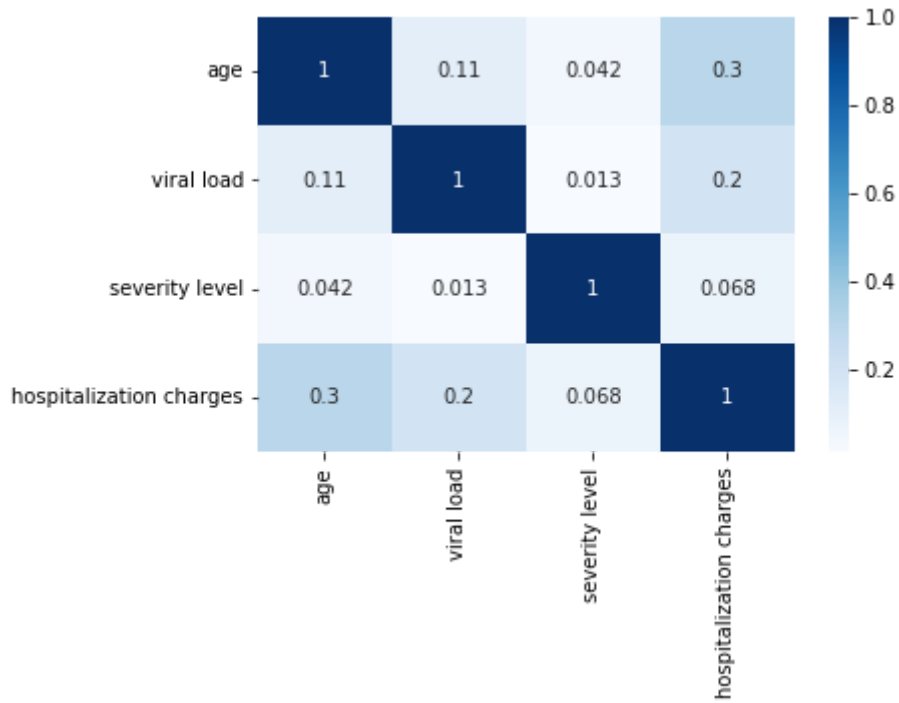
In [213]:

```
sns.heatmap(df.corr(),annot=True, cmap = "Blues")

```

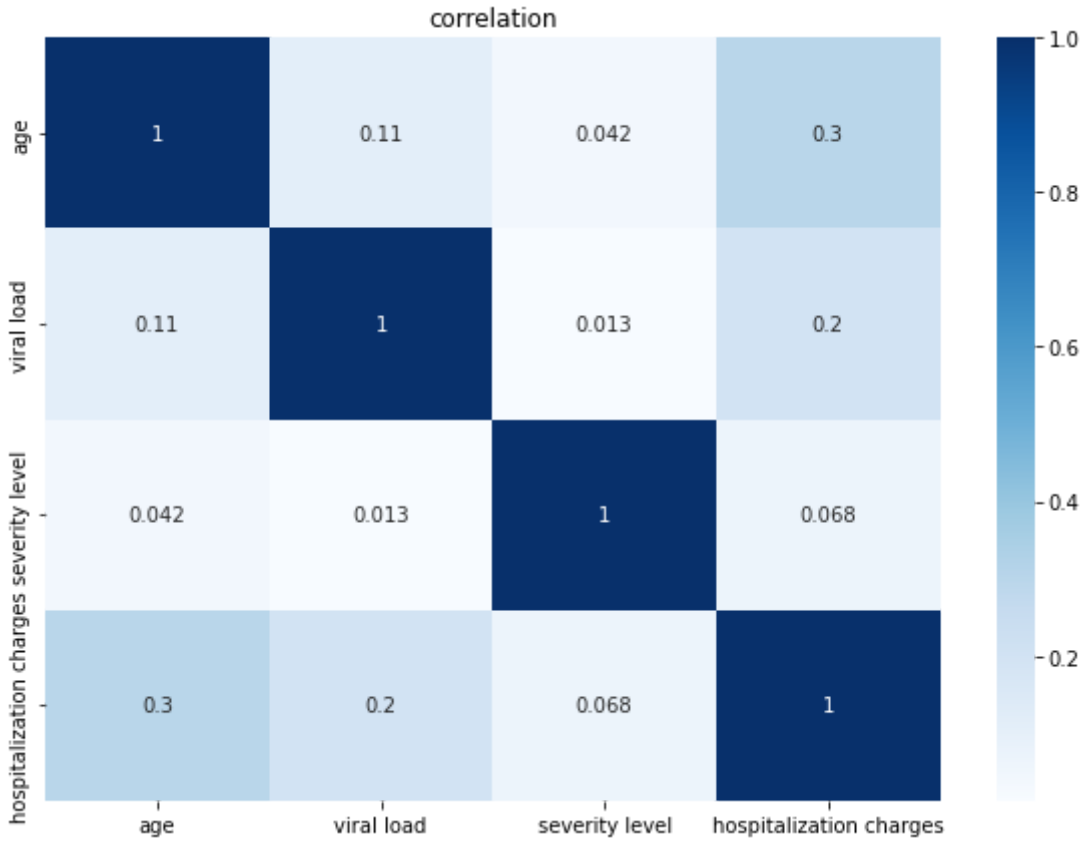
Out[213]:

<AxesSubplot:>



In [41]:

```
plt.figure(figsize=(10,7))
sns.heatmap(df[["age","sex","smoker","region","viral load","severity level","hospitaliz
plt.title("correlation")
plt.show()
```



Observation -

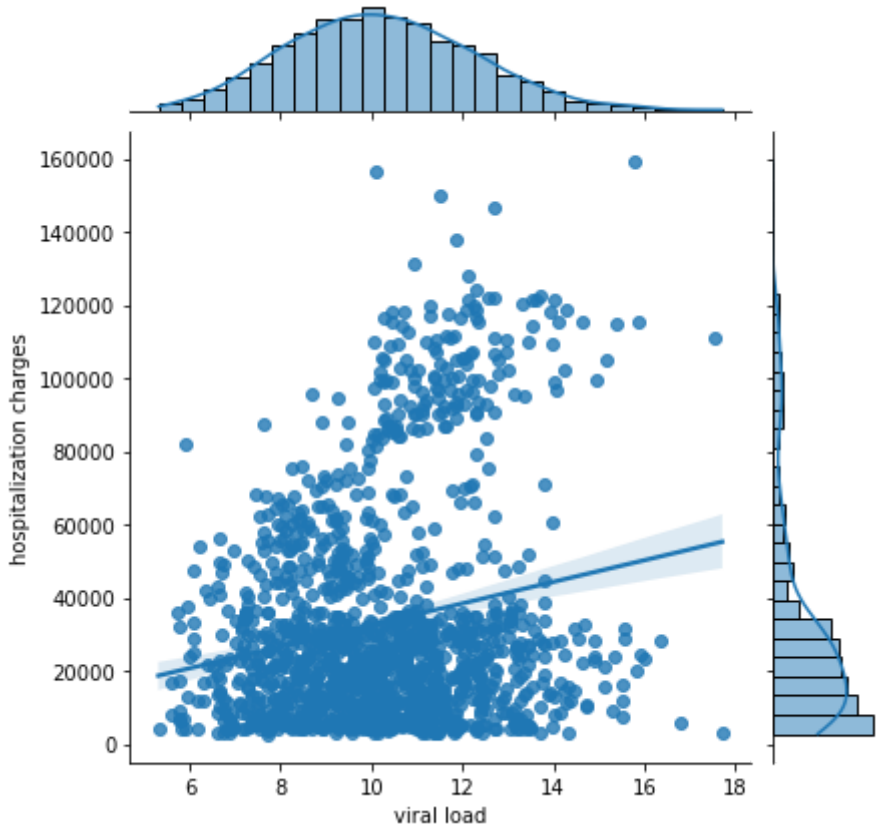
there seems to be a postitive correlation between hospitalization charges and age.

In [209]:

```
1 sns.jointplot(df["viral load"],df["hospitalization charges"],kind="reg")
```

Out[209]:

```
<seaborn.axisgrid.JointGrid at 0x1553857c670>
```



In [210]:

```
1 np.corrcoef(df["viral load"],df["hospitalization charges"])
```

Out[210]:

```
array([[1.        , 0.19838753],
       [0.19838753, 1.        ]])
```

Observation -

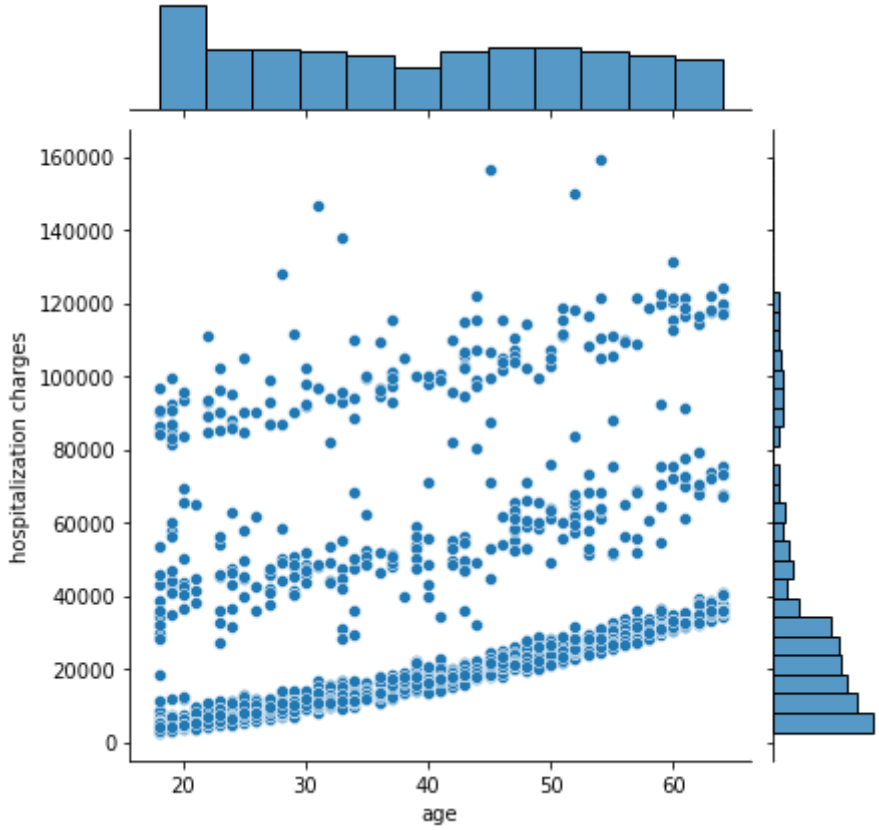
there's a very low correlation between viral load and hospitalization charges.

In [211]:

```
1 sns.jointplot(df["age"],df["hospitalization charges"])
```

Out[211]:

```
<seaborn.axisgrid.JointGrid at 0x15539b4dcd0>
```



In [212]:

```
1 df[["age","hospitalization charges"]].corr()
```

Out[212]:

	age	hospitalization charges
age	1.000000	0.299008
hospitalization charges	0.299008	1.000000

Observation -

it seems very slightly low correlation between age and the hospitalization charges

In [ ]:

```
1
```

Outlier Detection in Dataset

In [47]:

```
df1 = df.copy()
```

In [51]:

```
def Outliers(df):
    length_before = len(df)
    Q1 = np.percentile(df,25)
    Q3 = np.percentile(df,75)
    IQR = Q3-Q1
    upperbound = Q3 + 1.5*IQR
    lowerbound = Q1 - 1.5*IQR
    if lowerbound<0:
        lowerbound = 0

    length_after = len(df[(df>lowerbound)&(df<upperbound)])
    return f"{np.round((length_before-length_after)/length_before,4)} % Outliers data f
```

In [55]:

```
data = df1["viral load"]
Outliers(data)
```

Out[55]:

```
'0.0067 % Outliers data from input data found'
```

In [56]:

```
data = df1["hospitalization charges"]
Outliers(data)
```

Out[56]:

```
'0.1039 % Outliers data from input data found'
```

In [57]:

```
data = df1["severity level"]
Outliers(data)
```

Out[57]:

```
'0.4425 % Outliers data from input data found'
```

Observation -

outlier presence is not significant.

all the columns have outliers less than 5%.

In [ ]:

```
1
```

Range

Range is the simplest of the measurements but is very limited in its use, we calculate the range by taking the largest value of the dataset and subtract the smallest value from it, in other words, it is the difference of the maximum and minimum values of a dataset.

In [ ]:

```
1
```

Univariate analysis

In [89]:

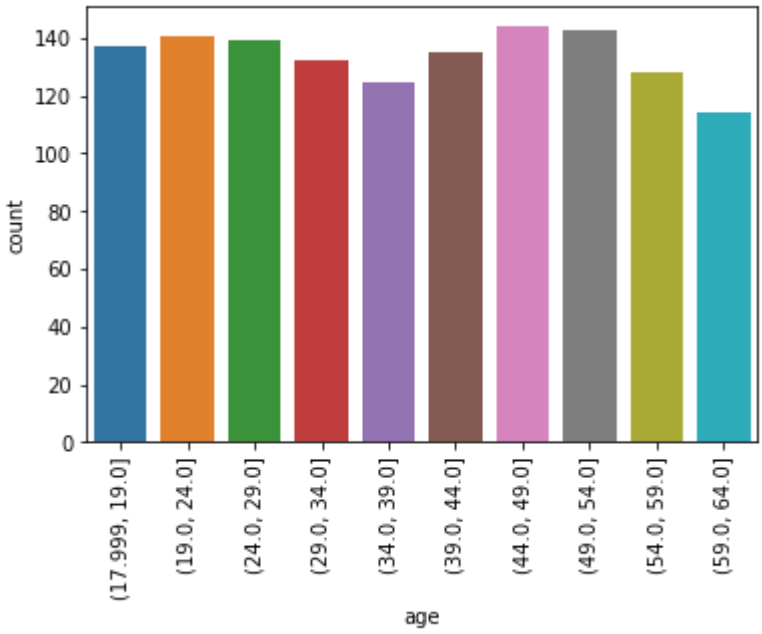
```
1 df["age_category"] = pd.qcut(df["age"],10)
2 df.head()
```

Out[89]:

	age	sex	smoker	region	viral load	severity level	hospitalization charges	age_category
0	19	female	yes	southwest	9.30	0	42212	(17.999, 19.0]
1	18	male	no	southeast	11.26	1	4314	(17.999, 19.0]
2	28	male	no	southeast	11.00	3	11124	(24.0, 29.0]
3	33	male	no	northwest	7.57	0	54961	(29.0, 34.0]
4	32	male	no	northwest	9.63	0	9667	(29.0, 34.0]

In [83]:

```
sns.countplot(pd.qcut(df["age"],10))
plt.xticks(rotation = 90)
plt.show()
```



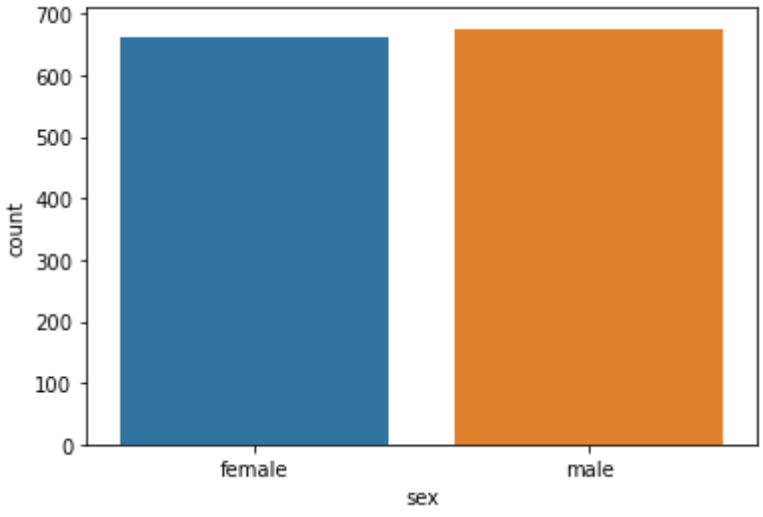
In [71]:

```
sns.countplot(df["sex"])
```

```
C:\Users\Shelendra\anaconda3\lib\site-packages\seaborn\_decorators.py:36: Fu
tureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other a
rguments without an explicit keyword will result in an error or misinterpret
ation.
  warnings.warn(
```

Out[71]:

```
<AxesSubplot:xlabel='sex', ylabel='count'>
```



There are most probably equal count of male and female in dataset

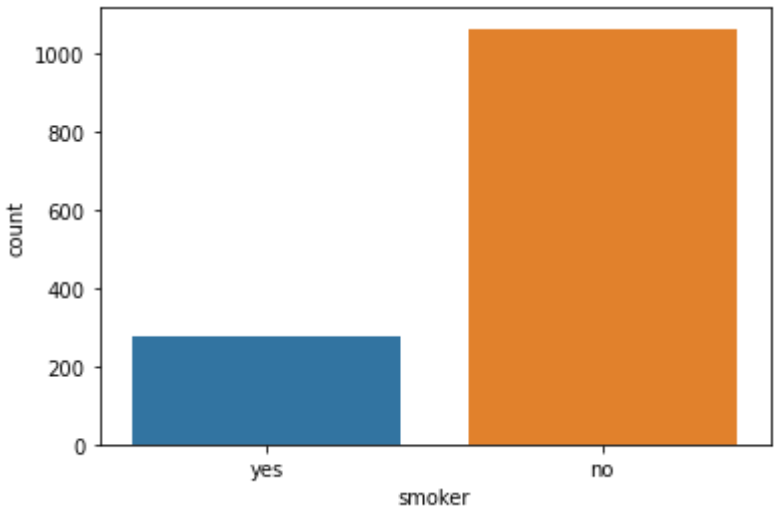
In [72]:

```
1 sns.countplot(df["smoker"])
```

```
C:\Users\Shelendra\anaconda3\lib\site-packages\seaborn\_decorators.py:36: Fu
tureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other a
rguments without an explicit keyword will result in an error or misinterpret
ation.
  warnings.warn(
```

Out[72]:

```
<AxesSubplot:xlabel='smoker', ylabel='count'>
```



So here are very less count of smokers in dataset who are diagnoses the diseases.

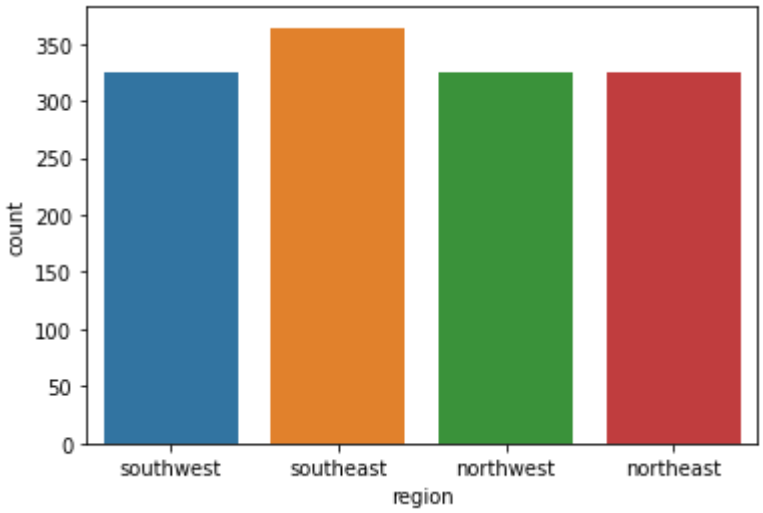
In [73]:

```
1 sns.countplot(df["region"])
```

```
C:\Users\Shelendra\anaconda3\lib\site-packages\seaborn\_decorators.py:36: Fu
tureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other a
rguments without an explicit keyword will result in an error or misinterpret
ation.
  warnings.warn(
```

Out[73]:

```
<AxesSubplot:xlabel='region', ylabel='count'>
```



In [ ]:

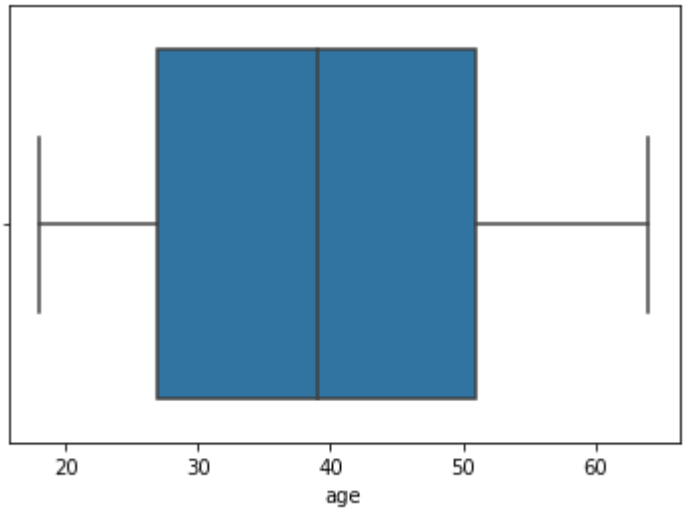
```
1 ## There is no such difference among the region wise disease .
```

In [74]:

```
1 sns.boxplot(x=df["age"])
```

Out[74]:

```
<AxesSubplot:xlabel='age'>
```



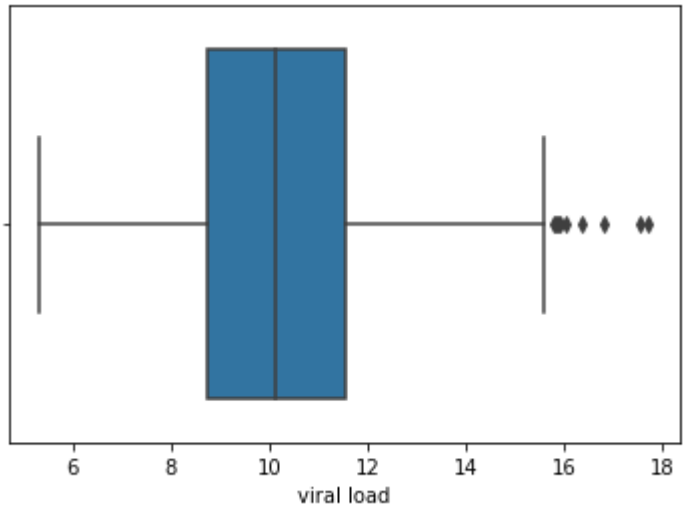
Here as we can see the people whose age are between 26 to 52 aprox are the most.

In [75]:

```
1 sns.boxplot(x=df["viral load"])
```

Out[75]:

```
<AxesSubplot:xlabel='viral load'>
```



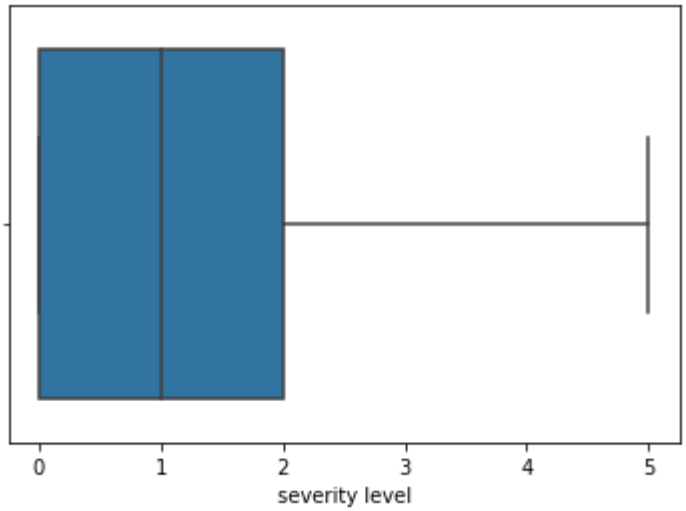
there are some outliers we have discussed in Detection Outliers previously.

In [76]:

```
1 sns.boxplot(x=df["severity level"])
```

Out[76]:

```
<AxesSubplot:xlabel='severity level'>
```

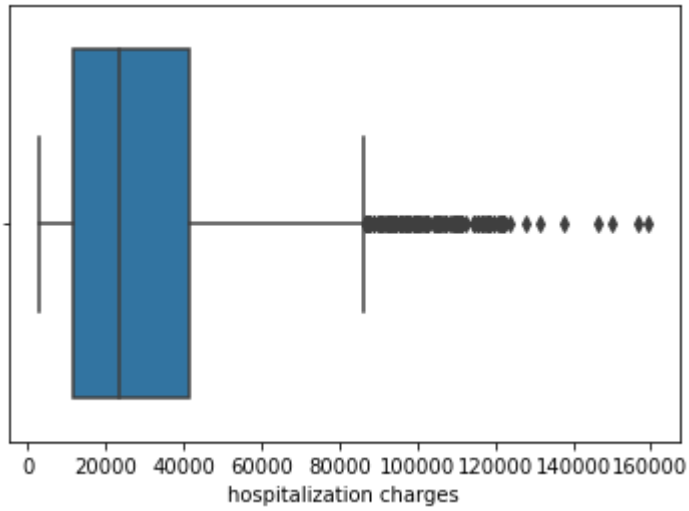


In [77]:

```
1 sns.boxplot(x=df["hospitalization charges"])
```

Out[77]:

```
<AxesSubplot:xlabel='hospitalization charges'>
```

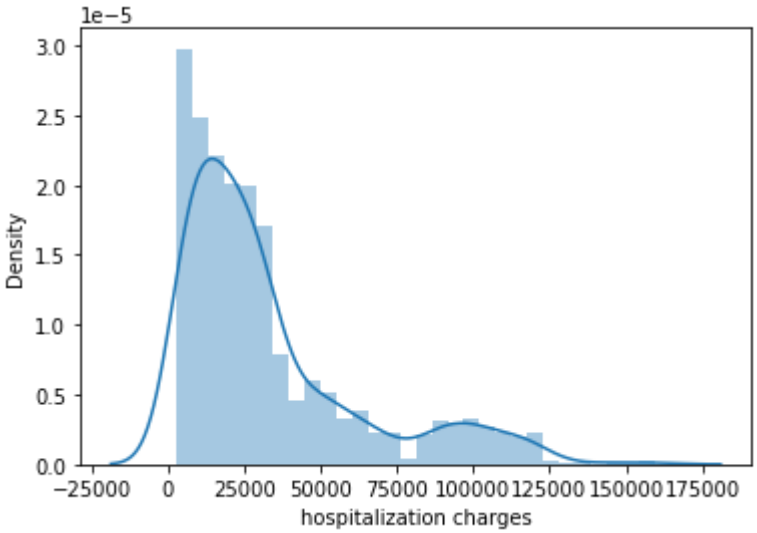


In [85]:

```
1 sns.distplot(df["hospitalization charges"])
```

Out[85]:

```
<AxesSubplot:xlabel='hospitalization charges', ylabel='Density'>
```



In [ ]:

```
1
```

In our dataset clearly written that "hospitalization charges" is a dependent variable and the other columns are independent So lets stablish a relation between them .

Bivariate Analysis

In [79]:

```
1 df.columns
```

Out[79]:

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

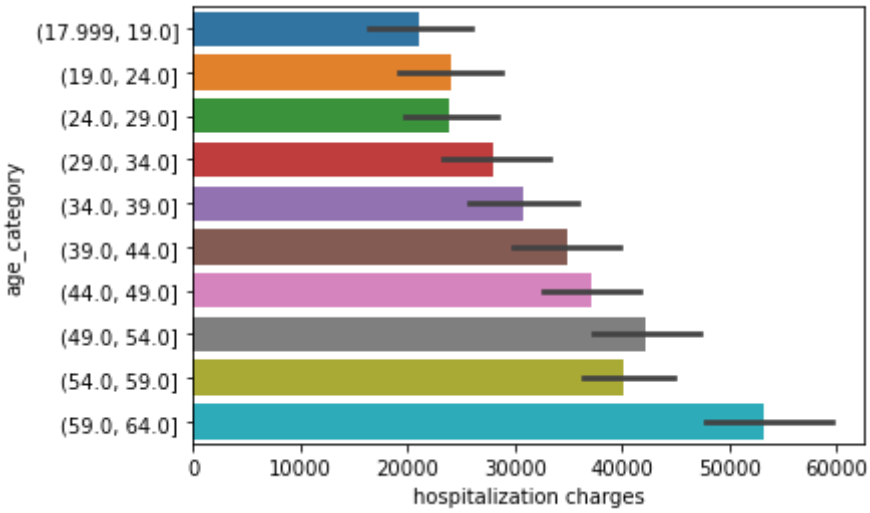
Age --->

In [117]:

```
1 sns.barplot(x = 'hospitalization charges',y =  "age_category",data = df)
```

Out[117]:

```
<AxesSubplot:xlabel='hospitalization charges', ylabel='age_category'>
```

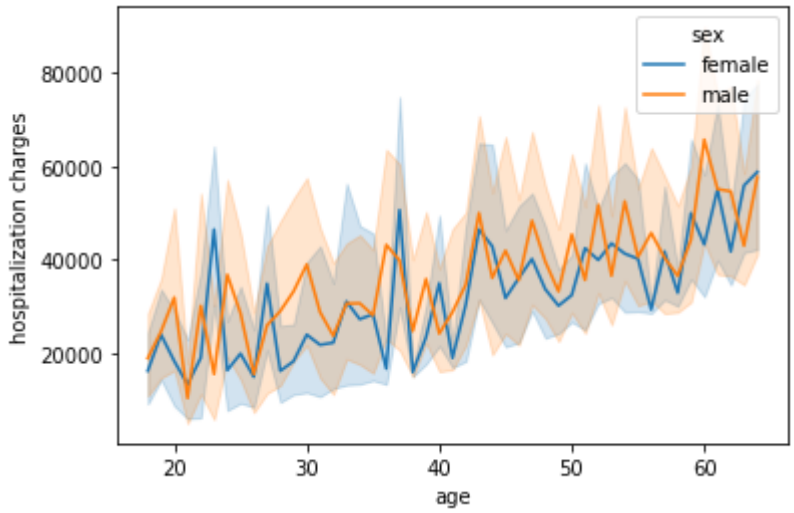


In [6]:

```
sns.lineplot(
    x="age",
    y="hospitalization charges",
    data=df,hue = "sex")
```

Out[6]:

```
<AxesSubplot:xlabel='age', ylabel='hospitalization charges'>
```



Observation -

As we can see the Hospitalization charges are increasing as the Age increase .

There are Hospitalization Charges are Dependent to the Age of the Person . It might be Posssible that Because of High viral load or week imunity system of person in High Age they Hospitalized for more days than Other Younger people.

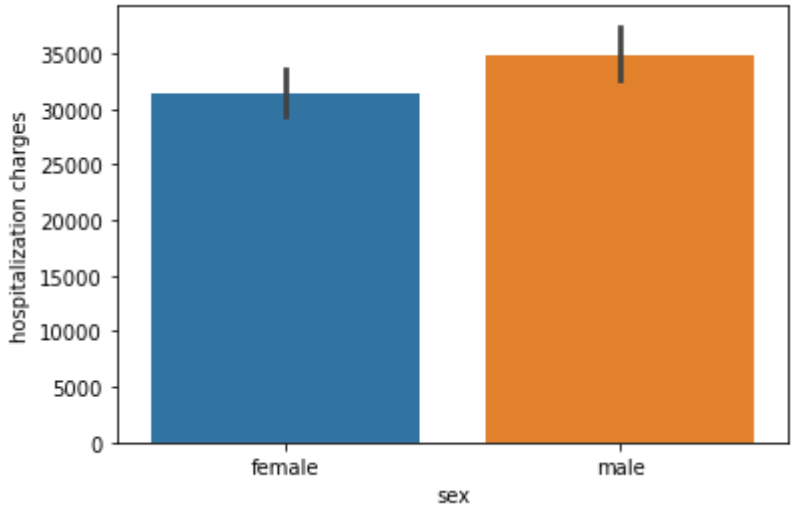
Sex -->

In [112]:

```
1 sns.barplot(df["sex"],df['hospitalization charges'])
```

Out[112]:

<AxesSubplot:xlabel='sex', ylabel='hospitalization charges'>



Observation -

As we can see that there are slighly Difference in Hospitalization charges between male and female .

Male are in the high side if we talk about the Hospitalization charges .

We will do some hypothesis test for more clearity on this ahead of this notebook.

Region ---->

In [9]:

```
1 df.groupby(["region","sex","smoker"]).mean()["hospitalization charges"].unstack()
```

Out[9]:

	smoker	no	yes
region	sex		
northeast	female	24101.090909	70080.068966
	male	21660.096000	77315.657895
northwest	female	21967.518519	74177.034483
	male	20801.734848	76782.862069
southeast	female	21100.525180	82587.111111
	male	19022.522388	90074.581818
southwest	female	20585.170213	79219.952381
	male	19447.293651	81497.189189

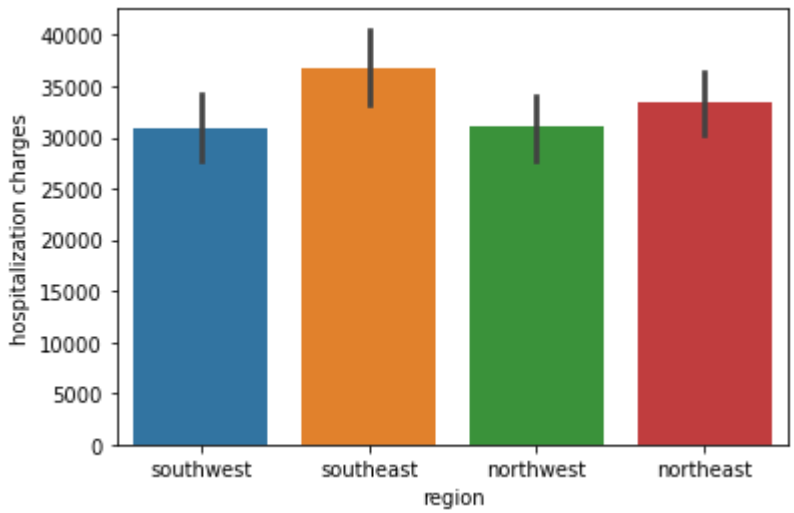
In every region for znon-smoker female having more Hospitalization charges as compare to male.

In [108]:

```
1 sns.barplot(df["region"],df['hospitalization charges'])
```

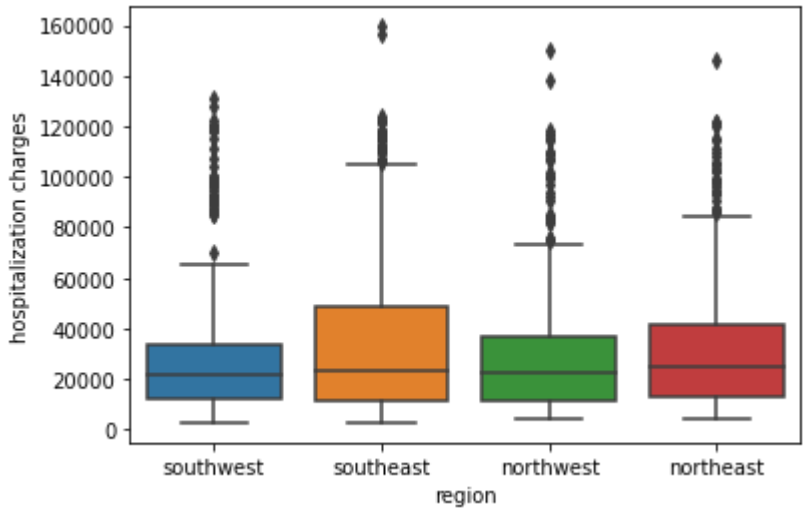
Out[108]:

```
<AxesSubplot:xlabel='region', ylabel='hospitalization charges'>
```



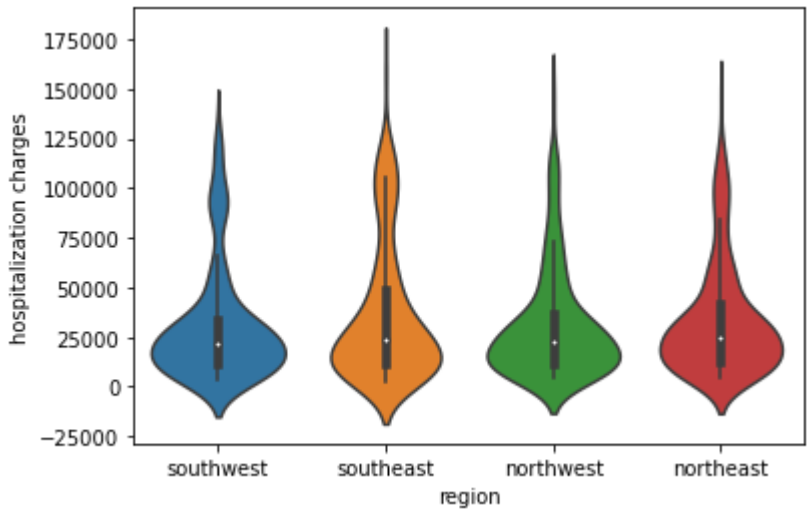
In [81]:

```
sns.boxplot(x ="region" , y = "hospitalization charges" , data = df)
plt.show()
```



In [82]:

```
sns.violinplot(x ="region" , y = "hospitalization charges" , data = df)
plt.show()
```




```
# Observation -
### According to the graphical visualization there is high chances of high Hospitalization charges in the southeast Area .
### In the second Number northeast area coming in this and left areas are same in the Hospitalization charges.
### Overall there is no such Difference among them .
```

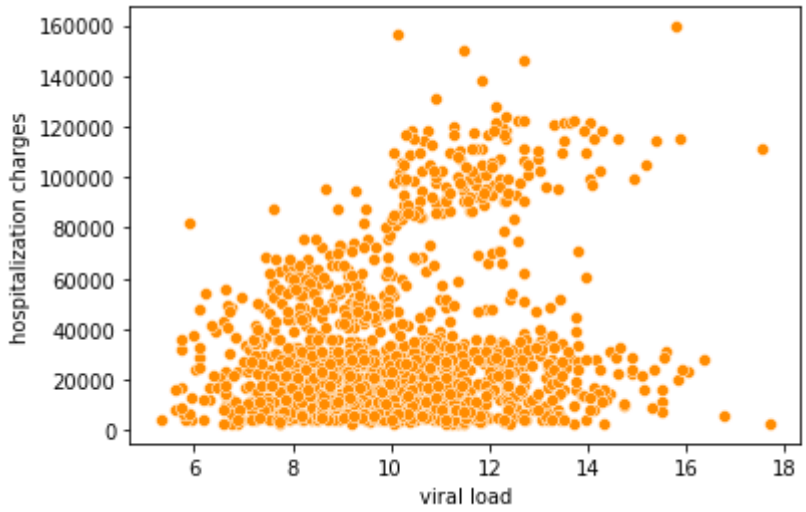
Viral load --->

In [94]:

```
1
2 sns.scatterplot(df["viral load"],df['hospitalization charges'],color='darkorange')
```

Out[94]:

<AxesSubplot:xlabel='viral load', ylabel='hospitalization charges'>

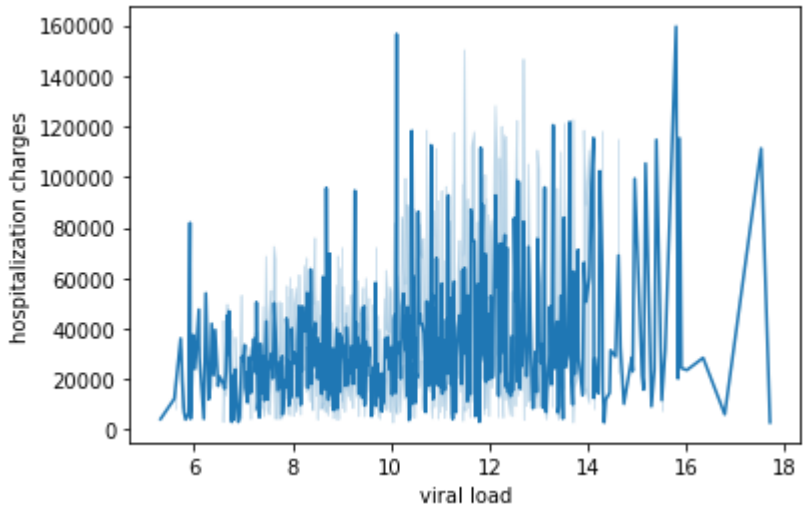


In [126]:

```
1 sns.lineplot(
2     x="viral load",
3     y="hospitalization charges",
4     data=df)
```

Out[126]:

<AxesSubplot:xlabel='viral load', ylabel='hospitalization charges'>



Observation -

It seems Hospitalization charges are not highly inceasing as viral load increase there is clear if viral load in between 6 to 15 then also

Hospitalization charges are almost same so that other factors are also possibly dependent for the Hospitalization charges.

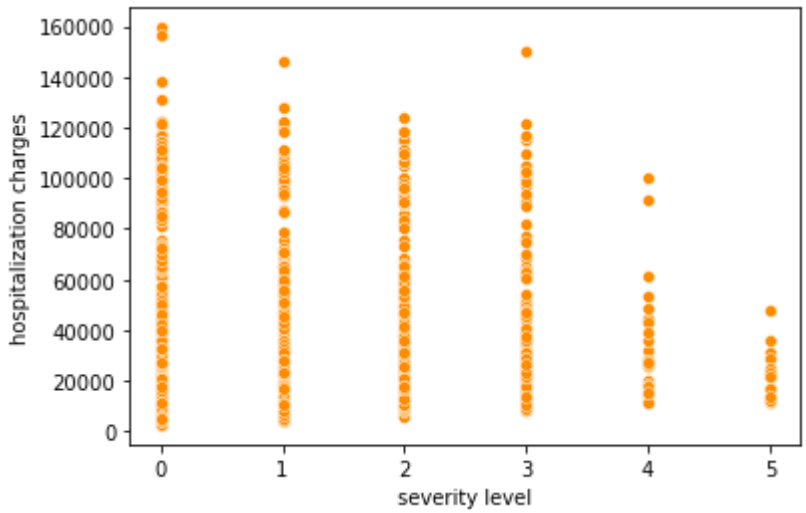
Severity level --->

In [95]:

```
1 sns.scatterplot(df["severity level"],df['hospitalization charges'],color='darkorange')
```

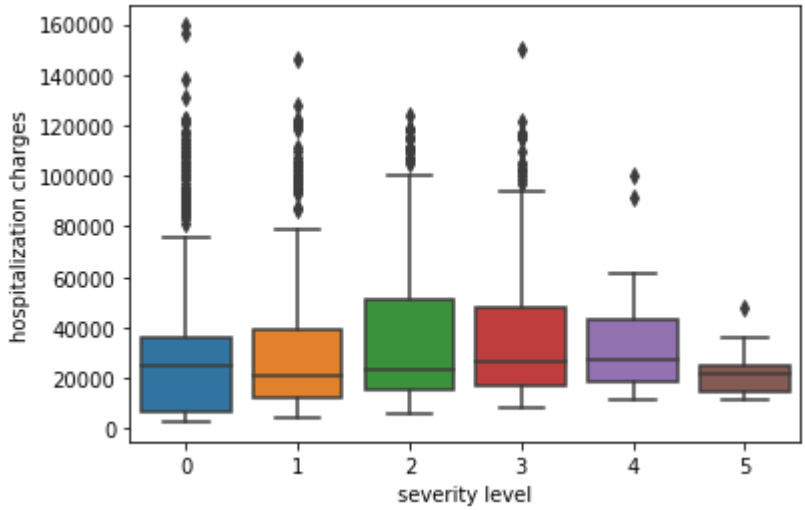
Out[95]:

```
<AxesSubplot:xlabel='severity level', ylabel='hospitalization charges'>
```



In [96]:

```
1 sns.boxplot(x ="severity level" , y = "hospitalization charges" , data = df)
2 plt.show()
```

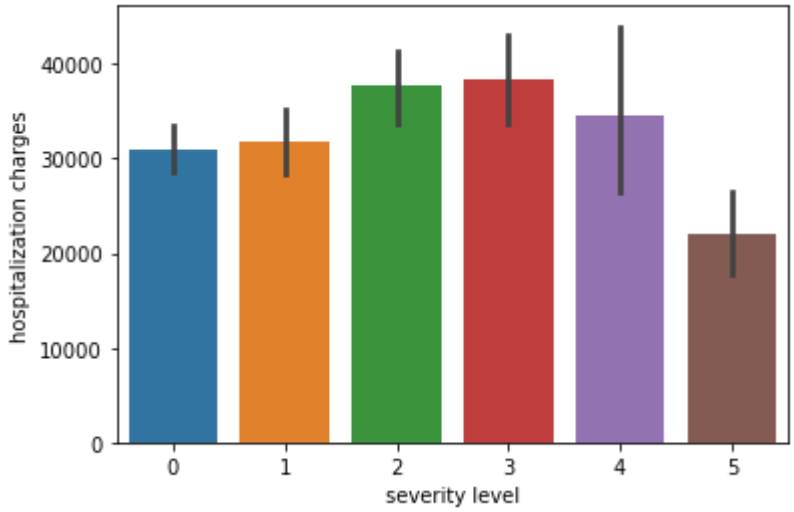


In [103]:

```
1 sns.barplot(x ="severity level" , y = "hospitalization charges" , data = df)
```

Out[103]:

```
<AxesSubplot:xlabel='severity level', ylabel='hospitalization charges'>
```



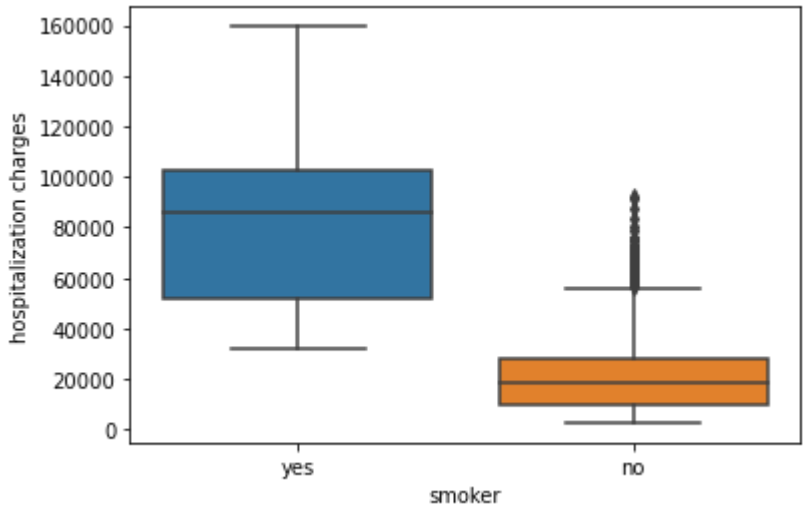
Observation -

Clearly can see severity level is normaly distributed with Hospitalization charges . Hospitalization charges are between 30000 to 35000 probabily there is if severity level is higher than hospitalization charges are less as comparitevely.

Smoker --->

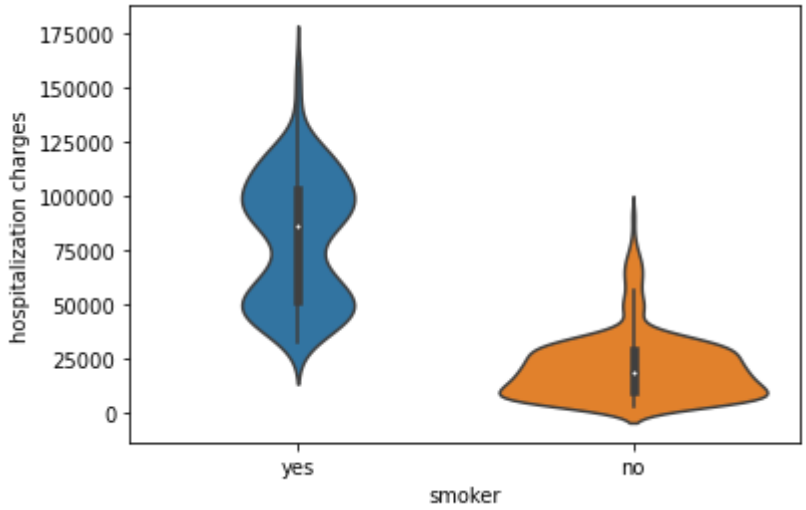
In [97]:

```
sns.boxplot(x ="smoker" , y = "hospitalization charges" , data = df)
plt.show()
```



In [100]:

```
sns.violinplot(x ="smoker" , y = "hospitalization charges" , data = df)
plt.show()
```

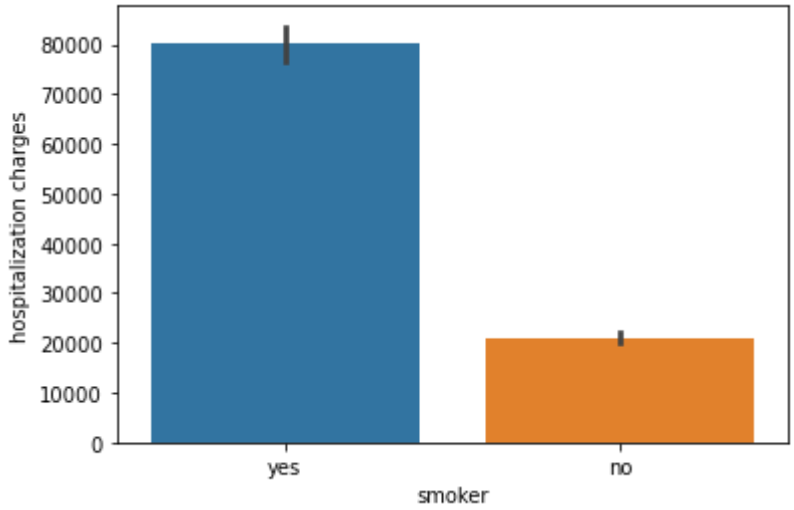


In [104]:

```
1 sns.barplot(x ="smoker" , y = "hospitalization charges" , data = df)
```

Out[104]:

```
<AxesSubplot:xlabel='smoker', ylabel='hospitalization charges'>
```



Observation -

There are the clear difference between smokers and non-smokers if person is a smoker then will hospitalized for more days or the expence would be more else less.

person who is non-smoker then the Hospitalization charges are less for those.

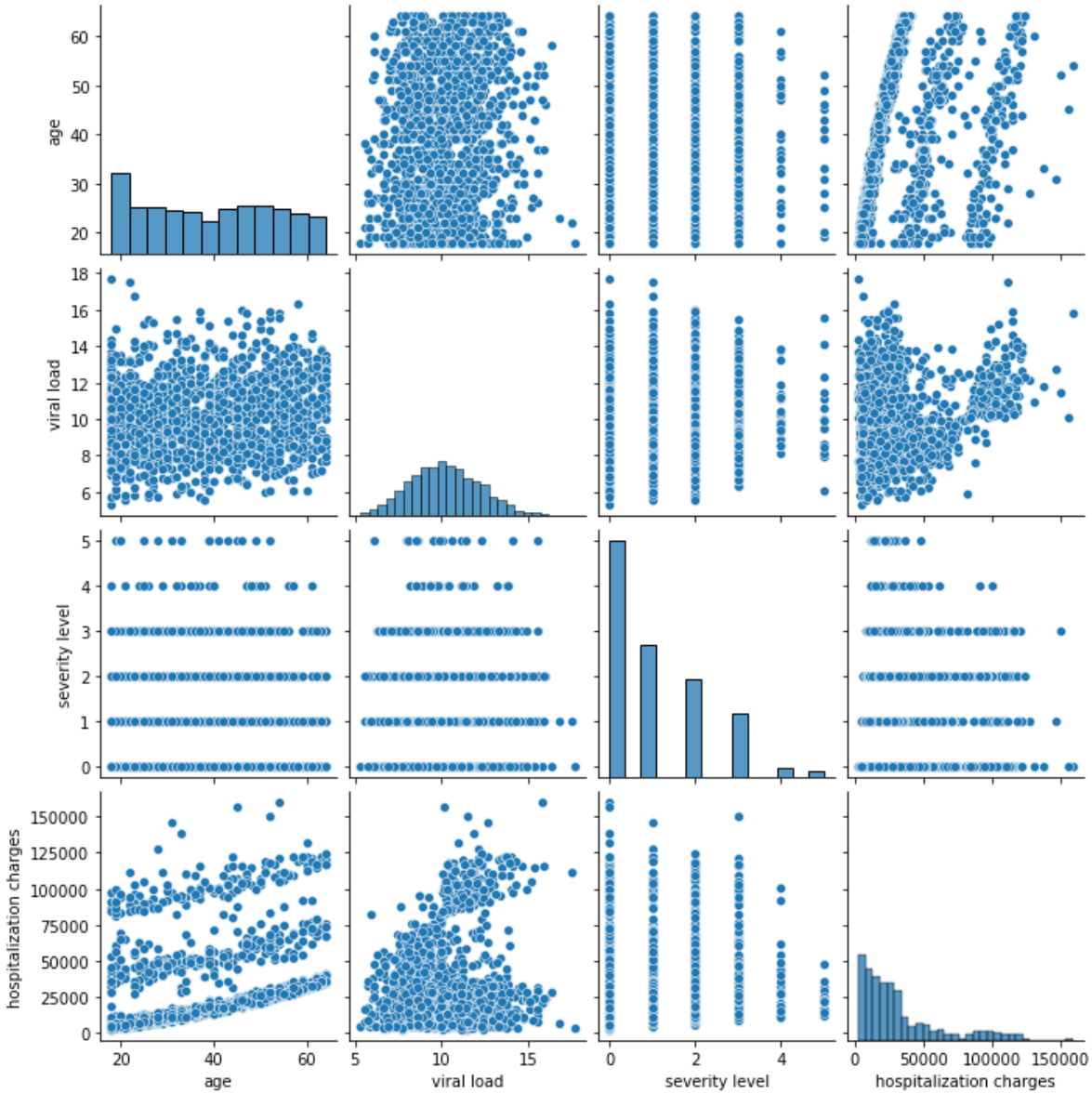
Multivariate Analysis

In [121]:

```
1 sns.pairplot(data=df)
```

Out[121]:

```
<seaborn.axisgrid.PairGrid at 0x15535d9dcd0>
```

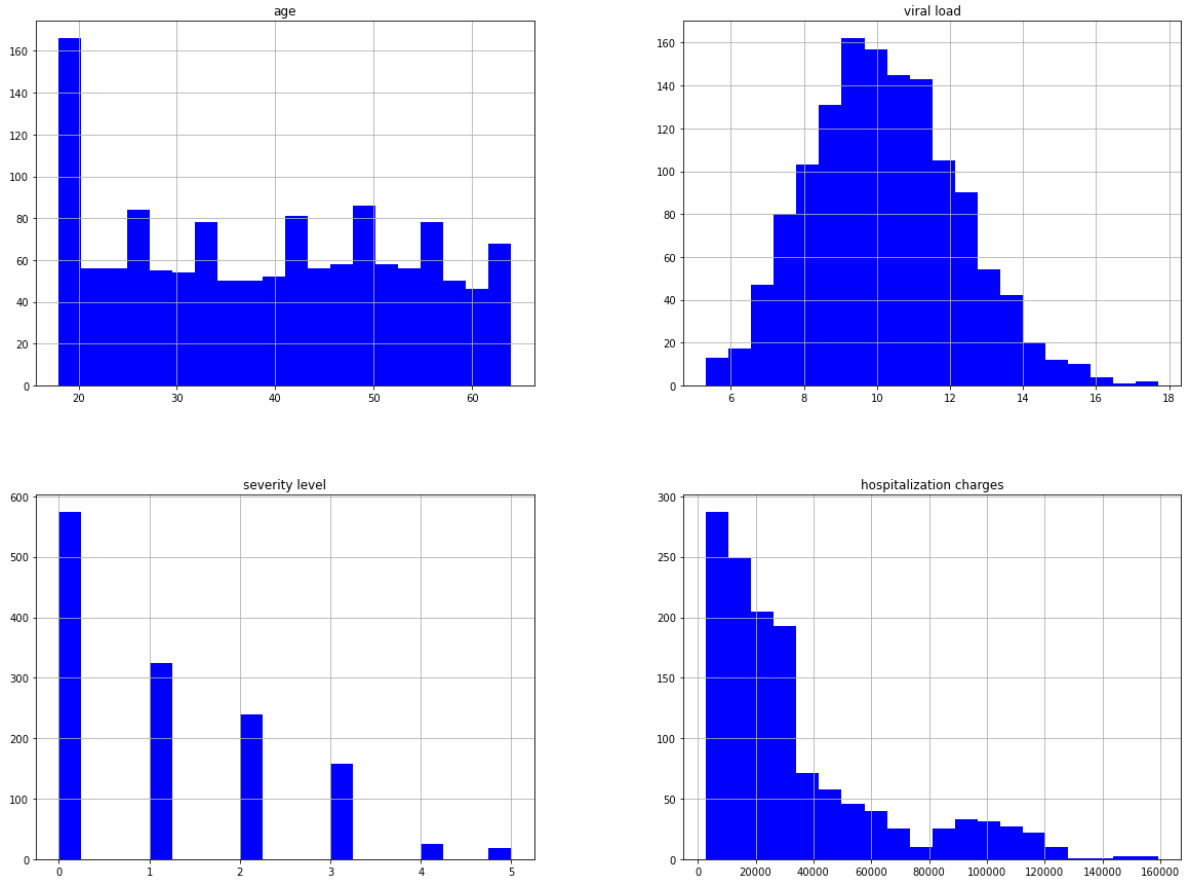


there are some positive corelated pairplots for dependent and independent variable.

overview on distributions of Numerical Features :

In [157]:

```
df.hist(bins=20,figsize=(20,15),color='blue')
plt.show()
```



Hypothesis Testing

Q1 - Prove (or disprove) that the hospitalization charges of people who do smoking are greater than those who don't?

(T-test Right tailed)

A right tailed test (sometimes called an upper test) is where your hypothesis statement contains a greater than (>) symbol. In other words, the inequality points to the right.

In [128]:

```
df.groupby("smoker")["hospitalization charges"].describe()
```

Out[128]:

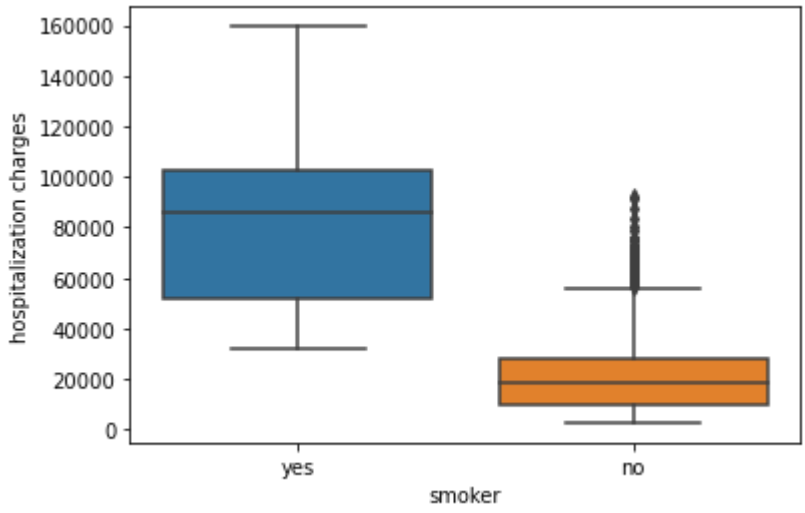
smoker	count	mean	std	min	25%	50%	75%	max
no	1064.0	21085.675752	14984.455500	2805.0	9966.25	18363.5	28407.25	92277.0
yes	274.0	80125.572993	28853.891137	32074.0	52065.50	86141.0	102548.25	159426.0

In [130]:

```
sns.boxplot(y = df["hospitalization charges"],
            x = df["smoker"])
```

Out[130]:

```
<AxesSubplot:xlabel='smoker', ylabel='hospitalization charges'>
```



Observation -

from above boxplot , hopitalization charges seems to be higher for smokers than who doent smoke.

t-test :

null hypotehsis : H0 : mean hospitalization charges for smokers and non smokers are same.

alternative hypothesis : Ha : Mean hospitalization charges for smokes is higher than non smokers for population .

H0 : smokers_charges <= non_smokers_charges

Ha : Smokers_charges > non_smokers_charges

alpha level = 0.05

In [136]:

```
# Differentiate the Smokers and the Non-Smokers
smokers = df[df["smoker"]=="yes"]["hospitalization charges"]
non_smokers = df[df["smoker"]=="no"]["hospitalization charges"]
```

In [ ]:

```

```

In [137]:

```
# Finding the mean of smokers and non-smokers
smokers.mean(),non_smokers.mean()
```

Out[137]:

```
(80125.57299270073, 21085.6757518797)
```

In [ ]:

```

```

In [138]:

```
# Checking the length of the smokers data and the Non-smokers data.
len(smokers),len(non_smokers)
```

Out[138]:

```
(274, 1064)
```

In [139]:

```
n1,n2 = len(smokers),len(non_smokers)
```

In [140]:

```
non_smokers = non_smokers.sample(274)
```

In [141]:

```
# Checking mean again
mean_smokers  = smokers.mean()
mean_non_smokers = non_smokers.mean()
mean_smokers,mean_non_smokers
```

Out[141]:

```
(80125.57299270073, 20298.846715328466)
```

In [142]:

```
std_smokers  = smokers.std()
std_non_smokers = non_smokers.std()
```

In [ ]:

```

```

In [145]:

```
test_statistic = (mean_smokers - mean_non_smokers)/(np.sqrt(((std_smokers**2)/(n1))+((s
```

In [146]:

```
test_statistic
```

Out[146]:

```
33.251524322494475
```

In [148]:

```
degreeOfFreedom = n1+n2-2
degreeOfFreedom
```

Out[148]:

```
1336
```

In [154]:

```
from scipy import stats
1-stats.t.cdf(test_statistic,degreeOfFreedom)
```

Out[154]:

```
0.0
```

In [155]:

```
stats.t.ppf(0.95,degreeOfFreedom)
```

Out[155]:

```
1.6459949688112576
```

In [156]:

```
stats.ttest_ind(smokers,non_smokers,    alternative='greater')
```

Out[156]:

```
Ttest_indResult(statistic=30.649951291948756, pvalue=4.67660393855921e-121)
```

Observation -

from the p-value we can observe the probability of having hospilization charges for smokers than non-smokers is is very high.

Thus from hypothesis test , we reject null hypothesis and conclude that hospitalization charges for Smokers are higher than Non Smokers.

In [ ]:

```
1
```

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

(T-test Two tailed)

A two-tailed test, in statistics, is a method in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.

In [160]:

```
1 df.columns
```

Out[160]:

```
Index(['age', 'sex', 'smoker', 'region', 'viral load', 'severity level',
       'hospitalization charges', 'age_category'],
      dtype='object')
```

In [161]:

```
1 df.groupby("sex")["viral load"].describe()
```

Out[161]:

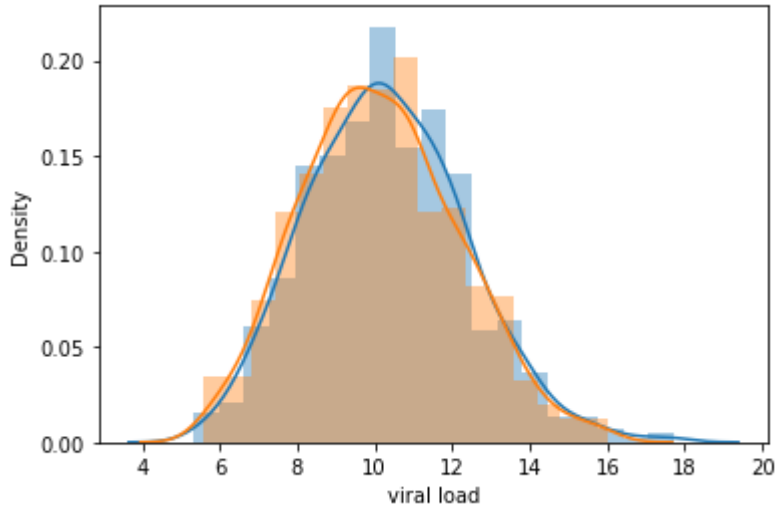
sex	count	mean	std	min	25%	50%	75%	max
female	662.0	10.126073	2.015402	5.60	8.71	10.035	11.4375	16.02
male	676.0	10.314423	2.046889	5.32	8.80	10.230	11.6625	17.71

In [176]:

```
sns.distplot(df.loc[df["sex"]=="male"]["viral load"])
sns.distplot(df.loc[df["sex"]=="female"]["viral load"])
```

Out[176]:

```
<AxesSubplot:xlabel='viral load', ylabel='Density'>
```

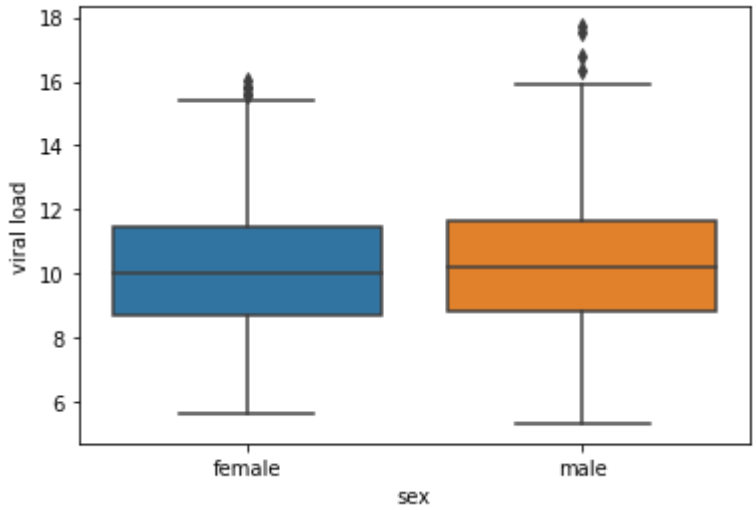


In [162]:

```
sns.boxplot(y = df["viral load"],
            x = df["sex"])
```

Out[162]:

```
<AxesSubplot:xlabel='sex', ylabel='viral load'>
```



Observation -

There are slightly difference between male and female .

t-test :¶

null hypotehsis : H0 : mean viral load for males and females are same.

alternative hypothesis : Ha : Mean viral load for males and mean viral load for females is not same .

H0 : viral_load_male = viral_load_female

Ha : viral_load_male != viral_load_female

level = 0.05

In [166]:

```
male_viral = df.loc[df["sex"]=="male"]["viral load"]
female_viral = df.loc[df["sex"]=="female"]["viral load"]

```

In [168]:

```
m1 = np.mean(male_viral)
n1 = len(male_viral)
s1 = np.std(male_viral,ddof = 1)
m2 = np.mean(female_viral)
n2 = len(female_viral)
s2 = np.std(female_viral,ddof = 1)
```

In [169]:

```
m1,n1,s1
```

Out[169]:

```
(10.314423076923074, 676, 2.0468891934763755)
```

In [170]:

```
m2,n2,s2
```

Out[170]:

```
(10.126072507552859, 662, 2.0154017361616767)
```

Test Statistic

In [171]:

```
T_observed =(m1-m2)/(np.sqrt(((s1**2)/n1)+((s2**2)/n2)))
T_observed
```

Out[171]:

1.6959864316229345

P-Value :

In [172]:

```
p_value = 2*(1-stats.t.cdf(T_observed,n1+n2-2))
p_value
```

Out[172]:

0.09012142591376415

Extream Critical Value

In [173]:

```
T_critical = stats.t.ppf(0.975,n1+n2-2)
T_critical
```

Out[173]:

1.9617412190546957

In [174]:

```
p_value > 0.05
```

Out[174]:

True

In [175]:

```
-T_critical < T_observed < T_critical
```

Out[175]:

True

Observation -

from the p-value we can observe the probability of having the viral load for male and female is almost same so

we failed to reject null hypothesis

mean viral load for males and females are same.

In [ ]:

```
1
```

Q3 - Is the proportion of smoking significantly different across different regions?

(Chi-square)-

A Pearson's chi-square test is a statistical test for categorical data. It is used to determine whether your data are significantly different from what you expected.

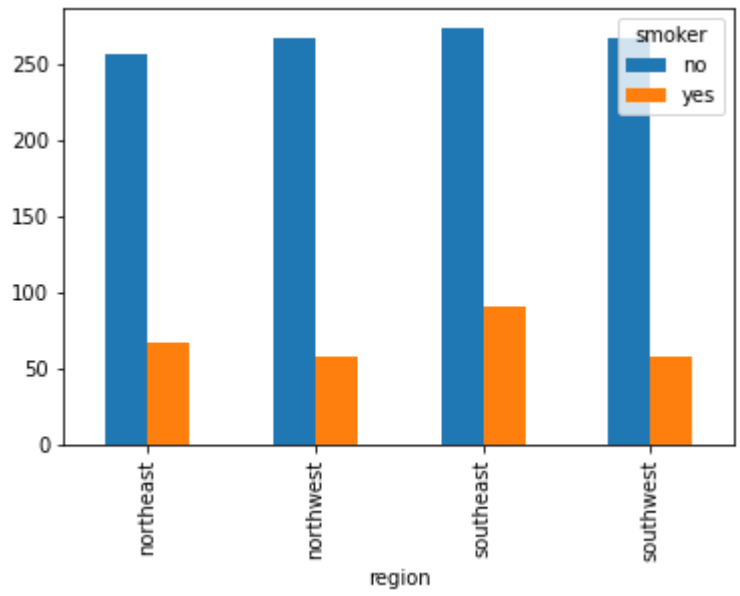
we are going to do here The chi-square test of independence which is used to test whether two categorical variables are related to each other.

In [177]:

```
1 pd.crosstab(index = df["region"],
2              columns= df["smoker"]).plot.bar()
```

Out[177]:

```
<AxesSubplot:xlabel='region'>
```



In [178]:

```
observed = pd.crosstab(columns = df["region"],
            index= df["smoker"])
observed
```

Out[178]:

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

In [179]:

```
row_sum = np.array(np.sum(observed,axis = 1))

col_sum = np.array(np.sum(observed,axis = 0))
```

In [180]:

```
row_sum,col_sum
```

Out[180]:

```
(array([1064,  274], dtype=int64), array([324, 325, 364, 325], dtype=int64))
```

In [181]:

```
total_sum = np.sum(np.sum(observed))
total_sum
```

Out[181]:

```
1338
```

In [182]:

```
expected = []
for i in row_sum:
    expected.append((i*col_sum)/total_sum)
expected
```

Out[182]:

```
[array([257.65022422, 258.44544096, 289.45889387, 258.44544096]),
 array([66.34977578, 66.55455904, 74.54110613, 66.55455904])]
```

In [183]:

```
expected = pd.DataFrame(expected, columns= observed.columns)
```

In [184]:

```
expected.index = observed.index
```

In [186]:

```
1 expected
```

Out[186]:

region	northeast	northwest	southeast	southwest
smoker				
no	257.650224	258.445441	289.458894	258.445441
yes	66.349776	66.554559	74.541106	66.554559

In [187]:

```
1 o_e_2_by_e = ((observed-expected)**2)/expected
```

In [188]:

```
1 np.sum(np.sum(o_e_2_by_e))   # test statistic
```

Out[188]:

```
7.343477761407071
```

In [189]:

```
1 stats.chi2.ppf(0.95,df=3)  # chi-sq critical value
```

Out[189]:

```
7.814727903251179
```

In [190]:

```
1 1-stats.chi2.cdf(7.343477761407071,3)
```

Out[190]:

```
0.06171954839170546
```

In [191]:

```
1 stats.chi2_contingency(observed)
```

Out[191]:

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

Observation -

from above chi-square test of independence we failed to reject null hypothesis , hence we conclude that proportion of smoking across different region is same.

so , smoker is independent of region !

In [ ]:

```
1
```

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)

One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. One-Way ANOVA is a parametric test. This test is also known as: One-Factor ANOVA.

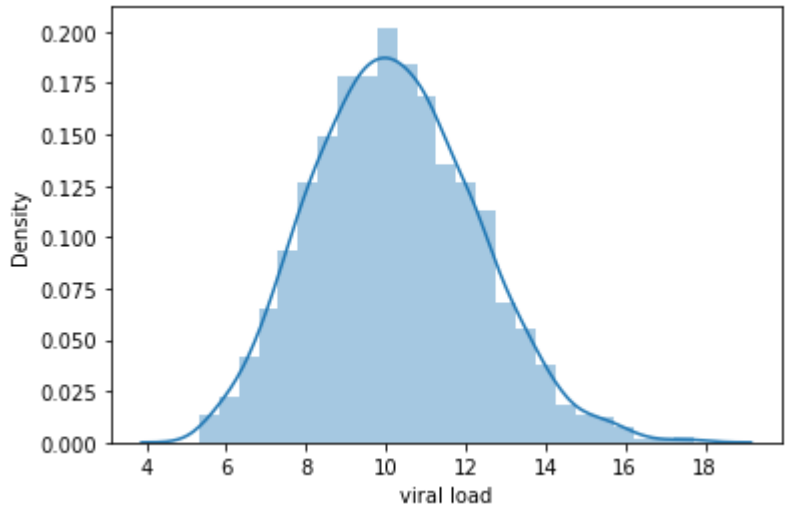
Statistical Evidence and test

In [194]:

```
1 sns.distplot(df["viral load"])
```

Out[194]:

```
<AxesSubplot:xlabel='viral load', ylabel='Density'>
```

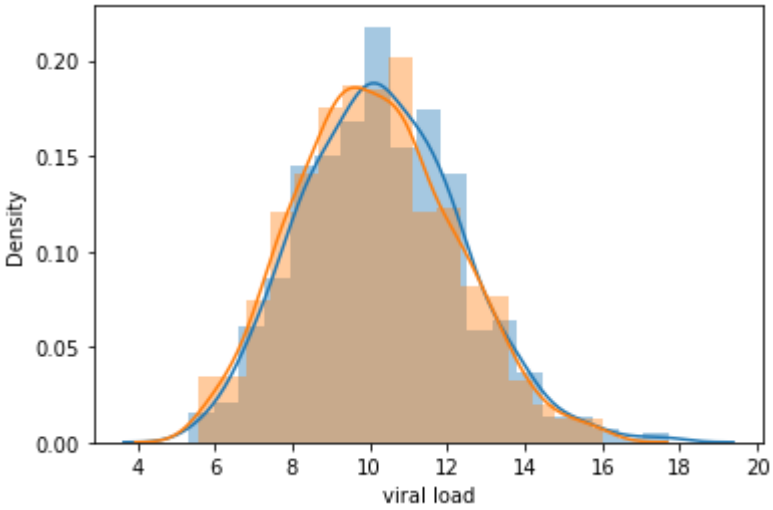


In [196]:

```
sns.distplot(df.loc[df["sex"]=="male"]["viral load"])
sns.distplot(df.loc[df["sex"]=="female"]["viral load"])
```

Out[196]:

```
<AxesSubplot:xlabel='viral load', ylabel='Density'>
```

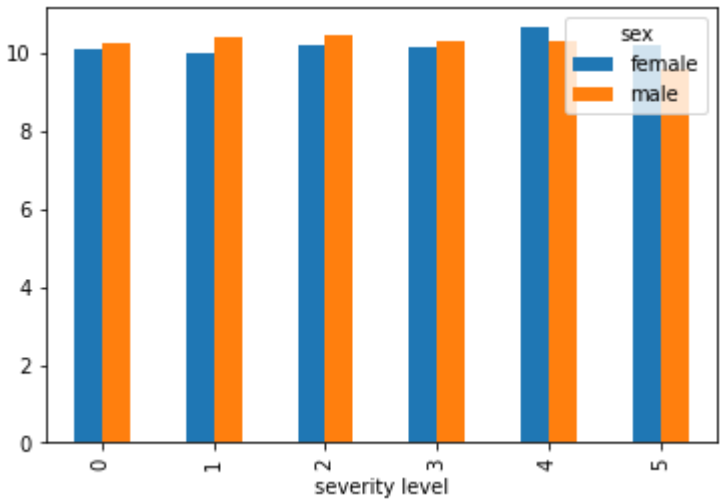


In [197]:

```
pd.crosstab(columns = df["sex"],
            index  = df["severity level"],
            values = df["viral load"],
            aggfunc = np.mean).plot.bar()
```

Out[197]:

```
<AxesSubplot:xlabel='severity level'>
```



from above bar plot , we can observe the mean viral load as per different severirty level is similar for male and female.

In [15]:

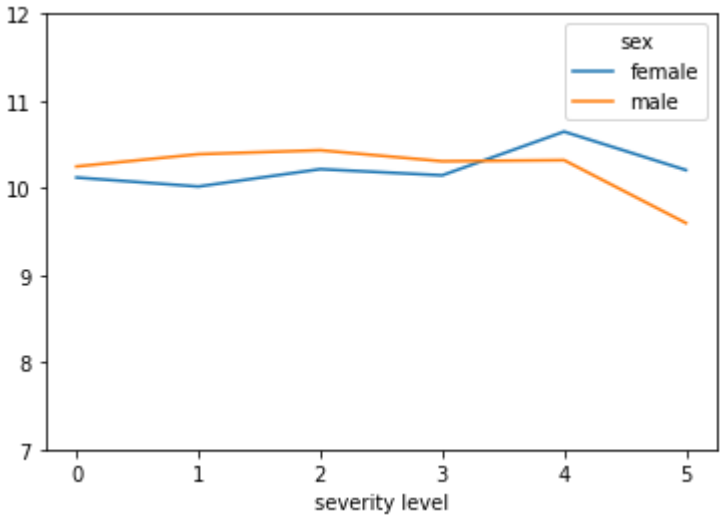
```
pd.crosstab(columns = df["sex"],
            index  = df["severity level"],
            values = df["viral load"],
            aggfunc = np.mean)
```

Out[15]:

sex	female	male
severity level		
0	10.120727	10.247544
1	10.017468	10.388494
2	10.216807	10.433554
3	10.145974	10.307375
4	10.647273	10.320000
5	10.206250	9.598000

In [16]:

```
pd.crosstab(columns = df["sex"],
            index  = df["severity level"],
            values = df["viral load"],
            aggfunc = np.mean).plot()
plt.ylim(7,12)
plt.show()
```

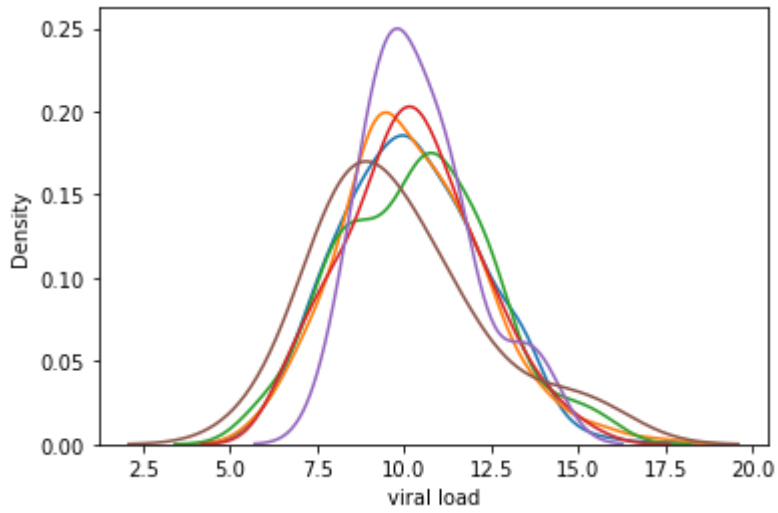


In [17]:

```
sns.kdeplot(df[df["severity level"]==0]["viral load"])
sns.kdeplot(df[df["severity level"]==1]["viral load"])
sns.kdeplot(df[df["severity level"]==2]["viral load"])
sns.kdeplot(df[df["severity level"]==3]["viral load"])
sns.kdeplot(df[df["severity level"]==4]["viral load"])
sns.kdeplot(df[df["severity level"]==5]["viral load"])
```

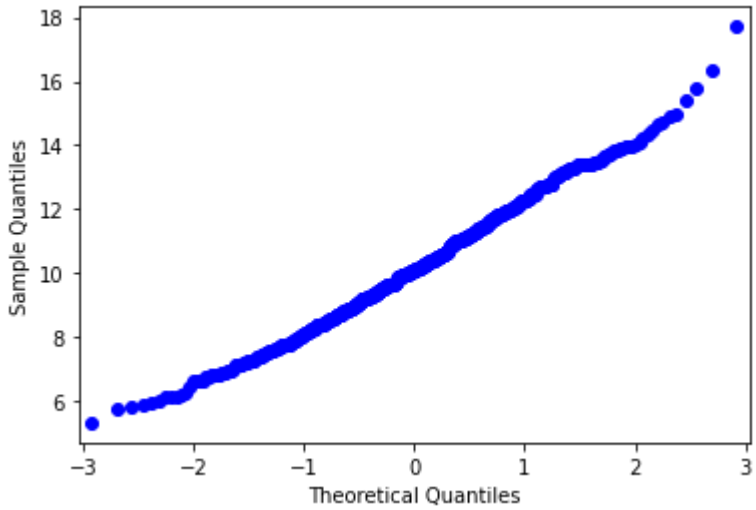
Out[17]:

```
<AxesSubplot:xlabel='viral load', ylabel='Density'>
```



In [18]:

```
sm.qqplot((df[df["severity level"]==0]["viral load"]))
plt.show()
```



In [19]:

```
df[df["severity level"]==0]["viral load"].mean(),df[df["severity level"]==1]["viral loa
```

Out[19]:

```
(10.183693379790936,
 10.207561728395063,
 10.326083333333333,
 10.22821656050955,
 10.464,
 9.868333333333332)
```

In [21]:

```
from scipy import stats
stats.f_oneway(df[df["severity level"]==0]["viral load"],
df[df["severity level"]==1]["viral load"],
df[df["severity level"]==2]["viral load"],
df[df["severity level"]==3]["viral load"],
df[df["severity level"]==4]["viral load"],
df[df["severity level"]==5]["viral load"])
```

Out[21]:

```
F_onewayResult(statistic=0.3491094504719582, pvalue=0.883007195713889)
```

In [22]:

```
womendata = df[df["sex"]=="female"]
```

In [23]:

```
stats.f_oneway(womendata[womendata["severity level"]==0]["viral load"],
womendata[womendata["severity level"]==1]["viral load"],
womendata[womendata["severity level"]==2]["viral load"],
womendata[womendata["severity level"]==3]["viral load"],
womendata[womendata["severity level"]==4]["viral load"],
womendata[womendata["severity level"]==5]["viral load"])
```

Out[23]:

```
F_onewayResult(statistic=0.2900065466233716, pvalue=0.9185708092374022)
```

Observation -

from the anova test , we can observe the viral load across different severity level is similar.

In [ ]:

```

```

Normality Assumption Check

Shapiro - Wilk's test

H0 : viral load follows normal distribution

Ha : viral load doesn't follow normal distribution

In [24]:

```
# Assumption 1 : Normality
# Import required functions
from scipy.stats import shapiro

# find the p-value
w,p_value = shapiro(womendata["viral load"])
print("The p-value is" , p_value)
```

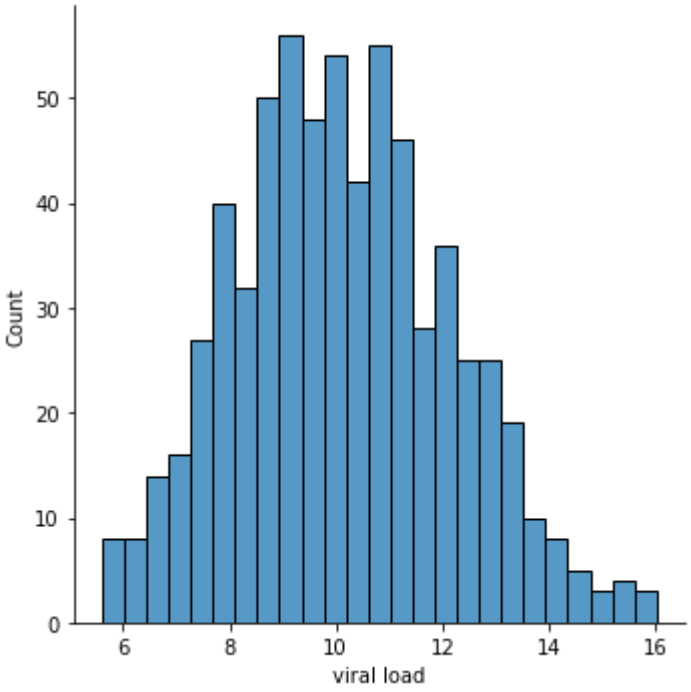
```
The p-value is 0.003624602919444442
```

In [27]:

```
import seaborn as sbn
sbn.displot(womendata["viral load"],bins = 25)# this looks normal
```

Out[27]:

```
<seaborn.axisgrid.FacetGrid at 0x2a15282ea00>
```



In [28]:

```
# Assumption 1 : Normality
from scipy.stats import shapiro
# find the p-value
w,p_value = shapiro(np.log(womendata["viral load"]))
print("the p-value",p_value)
```

```
the p-value 0.0028097345493733883
```

In [31]:

```
p_value > 0.05
```

Out[31]:

```
True
```

fail to reject null hypothesis

viral load follows normal distribution

Homogeneity of variance Asssumption Check

Levene's Test

we will Test the null Hypothesis

H0 : All the viral load variance are equal

Ha : At least one variance is Different from the rest

In [29]:

```
# Assumption 2: Homogeneity of variance
from scipy.stats import levene
statistics , p_value = levene(womendata[womendata["severity level"]==0]["viral load"].s
                              womendata[womendata["severity level"]==1]["viral load"].s
                              womendata[womendata["severity level"]==2]["viral load"].s
print("The p-value is ", p_value)
```

```
The p-value is  0.51447870116085
```

In [30]:

```
p_value >0.05
```

Out[30]:

```
True
```

fail to reject null hypothesis

All the viral load variance are equal

In [ ]:

```

```

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

In [ ]:

```
1
```

In [ ]:

```
1
```

Insights -

1) From the (T-Tset Right Tailed) Hypothesis Test we obsereved Hospitalization Charges for smokers than Non-smokers is very high.

2) By the (T-Test Two tailed) Hypothesis Test ,It is clear that probability of having the viral load for male and female is almost same .

3) We did (Chi-Square Test)Hypothesis Test and the point got that Proportion of smoking across the different region is same so smoker is independent of region .

4) From the (One Way Anova) Hypothesis Test we obsereved that the viral load across different severity level is similer .

5) There are maximum people are of Age 64 ,who has the virus and very less people of age of 18 That means People who are younger have less chances to got infected ,and Hospitalization Charges also would be low for them .

6)high age ,viral load and severity level are the major reason for the high hospitalization charges.

7)There are most probably equal count of male and female in dataset So this means that the Hospitalization Charges are not effected by the Sex.

8) Region wise also not much effect on the Hospitalization charges.

9) Age has the positive correlation with the Hospitalization charges

10)In every region for znon-smoker female having more Hospitalization charges as compare to male.

11) The males who smoke have the most claims and have higher bills.

12)there is one important point which helps us to make recommandation that is viral load is not affected by the severity .

13) Viral load has the normal distribution.

14)and the variance are also equal.

15)we can observe the mean viral load as per different severirty level is similar for male and female

Recommendations -

> We can encourage customers to Quit smoking by providing them incentive points fro talking life coach , get help for improving lifestyle habits, Quit Tabacco-28day program. Give gift cards when customers accumulates specific number of points .

> Based on the average or meadian charges for people and than come up with the premium charges for male and female different cases.

> High viral load is primarily because of less immunity in the body . We can provide patients vitamins and other Info to boost Impunity power, and rict diet plans and health coaches which can help them to make the right choices.

> Based on the insights that we genereted we can ask to take premium for thoseare the smokers . Because their Hospitalization charges are more than Non-smokers.

> In the Northeast area probabily charge higher for Non-smoker female .So may be there is Pollution kind of atmosphare for them so we can start a program to provide some info how can we prevent ourselves from this by medical.

> We can charge more on medical for smokers and female Non-smokers in Northeast areas.

> Don't need to change the treatment stratagy for male and female because as we can see that the viral load effect on male and female are same .

> viral load is not effected by the severity.So Hospital can be proactive steps in identifing if the patient have other existing conditions because of which they are serious or severity is more.

> so premium charge should not be less for these people.