Apollo - Hypothesis Testing

March 1, 2023

1 Problem Statement

Apollo Hospitals,established in 1983, renowned as the architect of modern healthcare in India.

As Data scientist our aim is to extract actionable and meanigful insights

such as comparing viral loads of male in comparison to female, checking hospitalization charge for smoker and non smoker patient, checking smoking and viral loads impacts in the various regions, comparing viral load with their severity level , comparing viral load respect to age .

on the basis of insights generated from the partient level data suggesting more efficient way to influence diagnostic and treatment processes, create awarness among the society regarding the ongoing health condition/pandemic

[1]:
```
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
import warnings
from scipy import stats
from scipy.stats import chi2_contingency
from scipy.stats import chi2
import numpy as np
sns.set_style(style='darkgrid')
plt.rcParams['figure.dpi'] = 200
sns.set(rc={'figure.figsize':(15.27,8)})
warnings.filterwarnings("ignore")
```

[2]:
```
data=pd.read_csv(r'/Users/surbhi/Desktop/scaler_apollo_hospitals.
 ↪csv',index_col='Unnamed: 0')
```

[3]:
```
data.head()
```

[3]:
```
   age     sex smoker     region  viral load  severity level  \
0   19  female    yes  southwest        9.30               0
1   18    male     no  southeast       11.26               1
2   28    male     no  southeast       11.00               3
3   33    male     no  northwest        7.57               0
4   32    male     no  northwest        9.63               0
```

```
   hospitalization charges
0                    42212
1                     4314
2                    11124
3                    54961
4                     9667
```

```
[4]: print(f"No of rows : {data.shape[0]}\nNo of columns : {data.shape[1]}")
```

```
No of rows : 1338
No of columns : 7
```

```
[5]: data.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
```

```
[6]: data.isna().sum()
```

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

No Missing values present in the data

```
[7]: data.shape
```

```
[7]: (1338, 7)
```

```
[8]: print("No. of duplicate rows: ",data.duplicated().sum())
```

```
No. of duplicate rows:  1
```

[9]:

```
data=data.drop_duplicates()
```

[10]:

```
print("Number of duplicates: {}".format(data.duplicated().sum()))
```

```
Number of duplicates: 0
```

[11]:

```
data.nunique()
```

[11]:

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

we have 6 unique values in severity level so lets convert it to categorical

[12]:

```
data['severity level']=data['severity level'].astype('object')
```

[13]:

```
data.describe()
```

[13]:

```
               age   viral load  hospitalization charges
count  1337.000000  1337.000000              1337.000000
mean     39.222139    10.221249             33197.806283
std      14.044333     2.033556             30275.900411
min      18.000000     5.320000              2805.000000
25%      27.000000     8.760000             11866.000000
50%      39.000000    10.130000             23465.000000
75%      51.000000    11.570000             41644.000000
max      64.000000    17.710000            159426.000000
```

[14]:

```
data.describe(include='object')
```

[14]:

```
         sex smoker     region  severity level
count   1337   1337       1337            1337
unique     2      2          4               6
top     male     no  southeast               0
freq     675   1063        364             573
```

[15]:

```
categorical_column=['sex','smoker','region','severity level']
```

[16]:

```
for i in categorical_column:
    print("***********************************")
    print(i)
    print(data[i].value_counts(normalize=True)*100)
```

```
***********************************
sex
male      50.486163
female    49.513837
Name: sex, dtype: float64
***********************************
smoker
no     79.506358
yes    20.493642
Name: smoker, dtype: float64
***********************************
region
southeast    27.225131
southwest    24.308153
northwest    24.233358
northeast    24.233358
Name: region, dtype: float64
***********************************
severity level
0    42.857143
1    24.233358
2    17.950636
3    11.742708
4     1.869858
5     1.346298
Name: severity level, dtype: float64
```

Observations:

1. we have almost equal amount of male and female customer , 49.51% female customer and 50.48% male customer respectively.
2. 79% non smoker and 20.49% smoker patient
3. from southeast we have 27.22% patient and almost 24% from all other region i.e southwest,northwest,norteast
4. 42% of data having severity level 0 followed by 24% having severity level 1,17.95% having severity level 2, 11.74% having severity level 3, 1.87% having severity level 4 and 1.35% having severity level 5

[17]:
```
data.columns
```

[17]:
```
Index(['age', 'sex', 'smoker', 'region', 'viral load', 'severity level',
       'hospitalization charges'],
      dtype='object')
```

[18]:
```
five_point_summary=data.describe().T
five_point_summary['IQR']=np.
 ↪round(five_point_summary['75%']-five_point_summary['25%'],2)
five_point_summary['Upper Whisker']=np.round(five_point_summary['75%']+1.
 ↪5*five_point_summary['IQR'],2)
five_point_summary['Lower Whisker']=np.round(five_point_summary['25%']-1.
 ↪5*five_point_summary['IQR'],2)
five_point_summary
```

[18]:
```
                          count          mean           std      min  \
age                      1337.0     39.222139     14.044333    18.00
viral load               1337.0     10.221249      2.033556     5.32
hospitalization charges  1337.0  33197.806283  30275.900411  2805.00

                              25%       50%       75%        max       IQR  \
age                         27.00     39.00     51.00      64.00     24.00
viral load                   8.76     10.13     11.57      17.71      2.81
hospitalization charges  11866.00  23465.00  41644.00  159426.00  29778.00

                         Upper Whisker  Lower Whisker
age                              87.00          -9.00
viral load                       15.78           4.54
hospitalization charges       86311.00      -32801.00
```

[19]:
```
numerical_column = ['age', 'viral load','hospitalization charges']
```

[20]:
```
outlier_count=data[(data[numerical_column]>five_point_summary.T.loc['Upper␣
 ↪Whisker',numerical_column])|(data[numerical_column]<five_point_summary.T.
 ↪loc['Lower Whisker',numerical_column])].count()
print(outlier_count)
outlier_count_percent=(outlier_count/len(data))*100
print("outlier count in %")
outlier_count_percent[numerical_column]
```

```
age                          0
sex                          0
smoker                       0
region                       0
viral load                   9
severity level               0
hospitalization charges    139
dtype: int64
outlier count in %
```

[20]:
```
age                         0.000000
viral load                  0.673149
hospitalization charges    10.396410
```

```
dtype: float64
```

Observation

1. median age of patient is 39 with minimum age 18 and maximum 64.
2. median viral load is 10.13 having 0.67% outlier.
3. minimum viral load is 5.32 and maximum 17.71
4. median hospitalization charge is 23465.0 having 10.39 % of outlier
5. minimum hospitalization charge is 2805 and maximum 159426

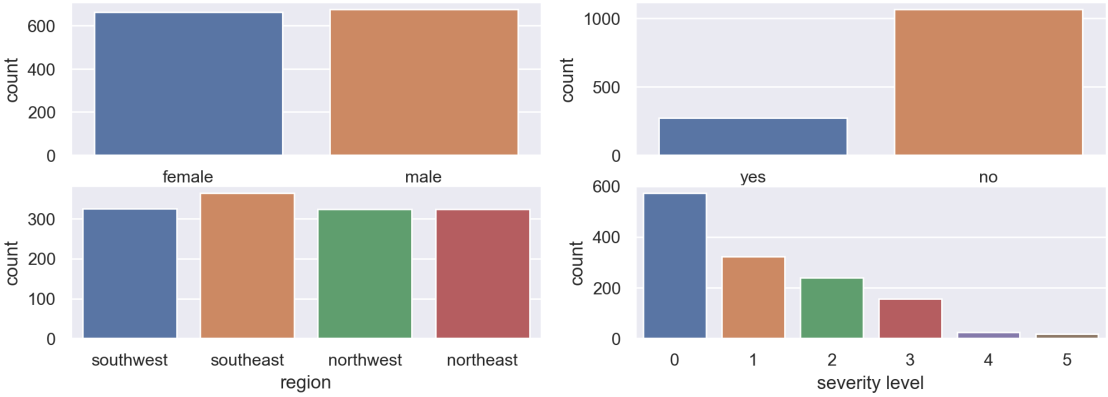
2 univariate Analysis

[21]:

```
categorical_column=['sex','smoker','region','severity level']
numerical_column = ['age', 'viral load','hospitalization charges']
```

[22]:

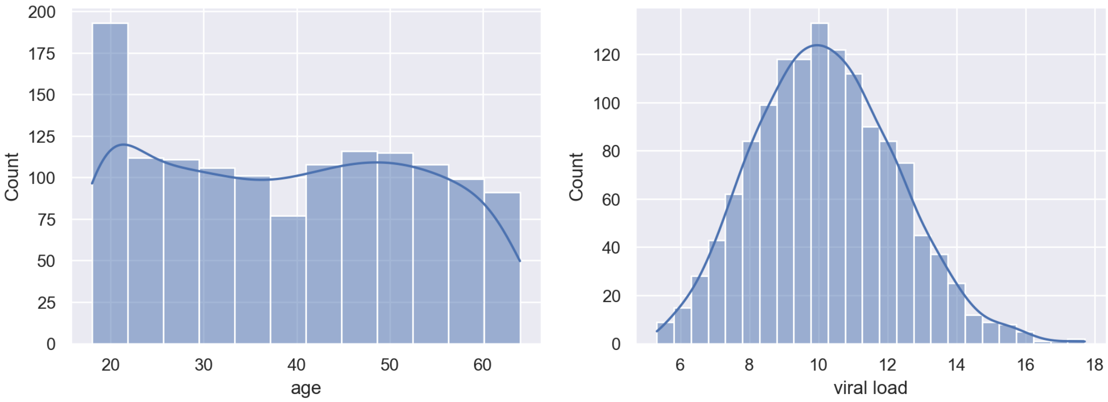
```
fig, axs = plt.subplots(nrows=2, ncols=2, figsize=(12, 4))
ind=0
for i in range(2):
    for j in range(2):
        sns.countplot(x=categorical_column[ind], data=data,ax=axs[i,j])
        ind+=1

plt.show()
```

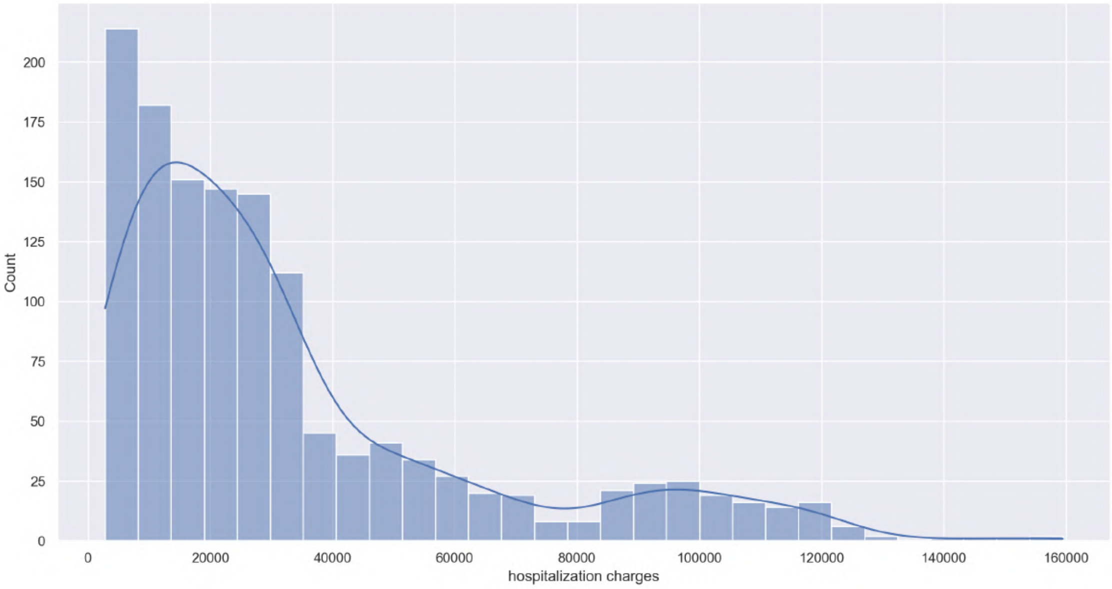


1. sex & region have almost equal number of values in each category.
2. Most of the patients are non-smoker.
3. Number of patients decreses as severity level increses.

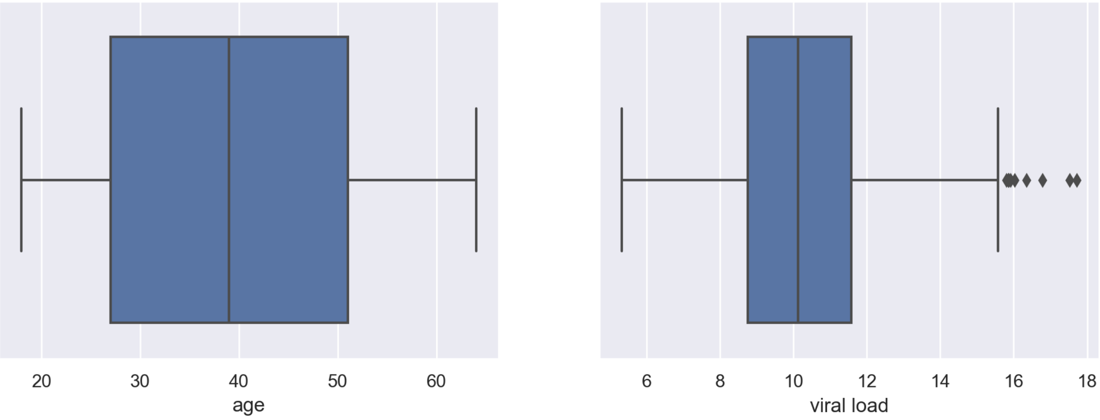
[23]:

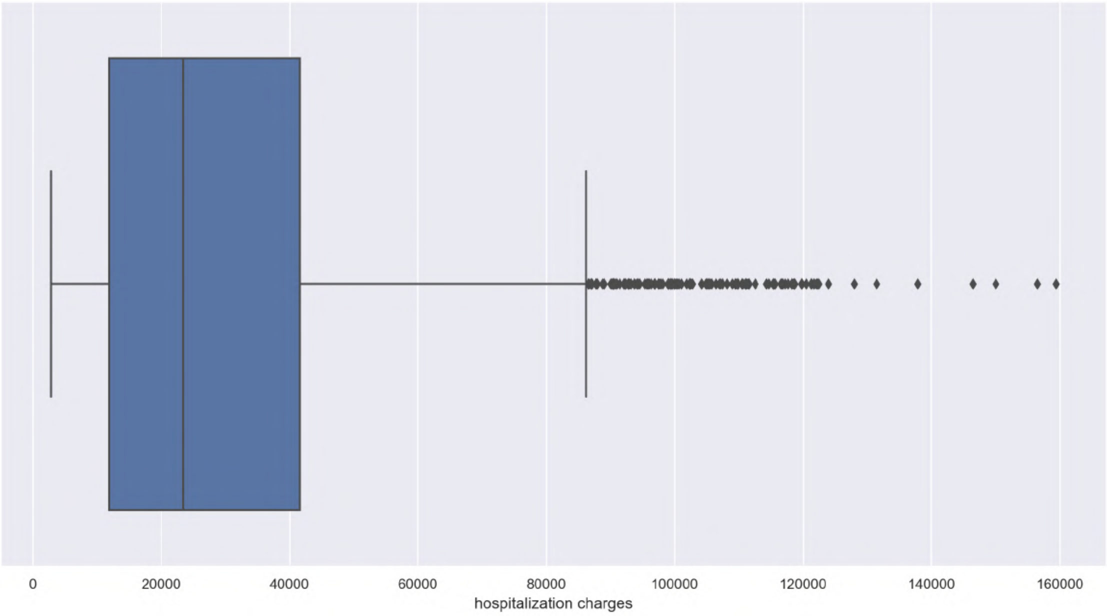
```
fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(12, 4))
ind=0
numerical_column = ['age', 'viral load']
for i in range(1):
    for j in range(2):

        sns.histplot(x=data[numerical_column[ind]],kde=True,ax=axs[j])
        ind+=1

plt.show()
sns.histplot(x=data['hospitalization charges'],kde=True)
plt.show()
```





[24]:
```
fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(12, 4))
ind=0
numerical_column = ['age', 'viral load']
for i in range(1):
    for j in range(2):

        sns.boxplot(x=data[numerical_column[ind]],ax=axs[j])
        ind+=1

plt.show()
sns.boxplot(x=data['hospitalization charges'])
plt.show()
```





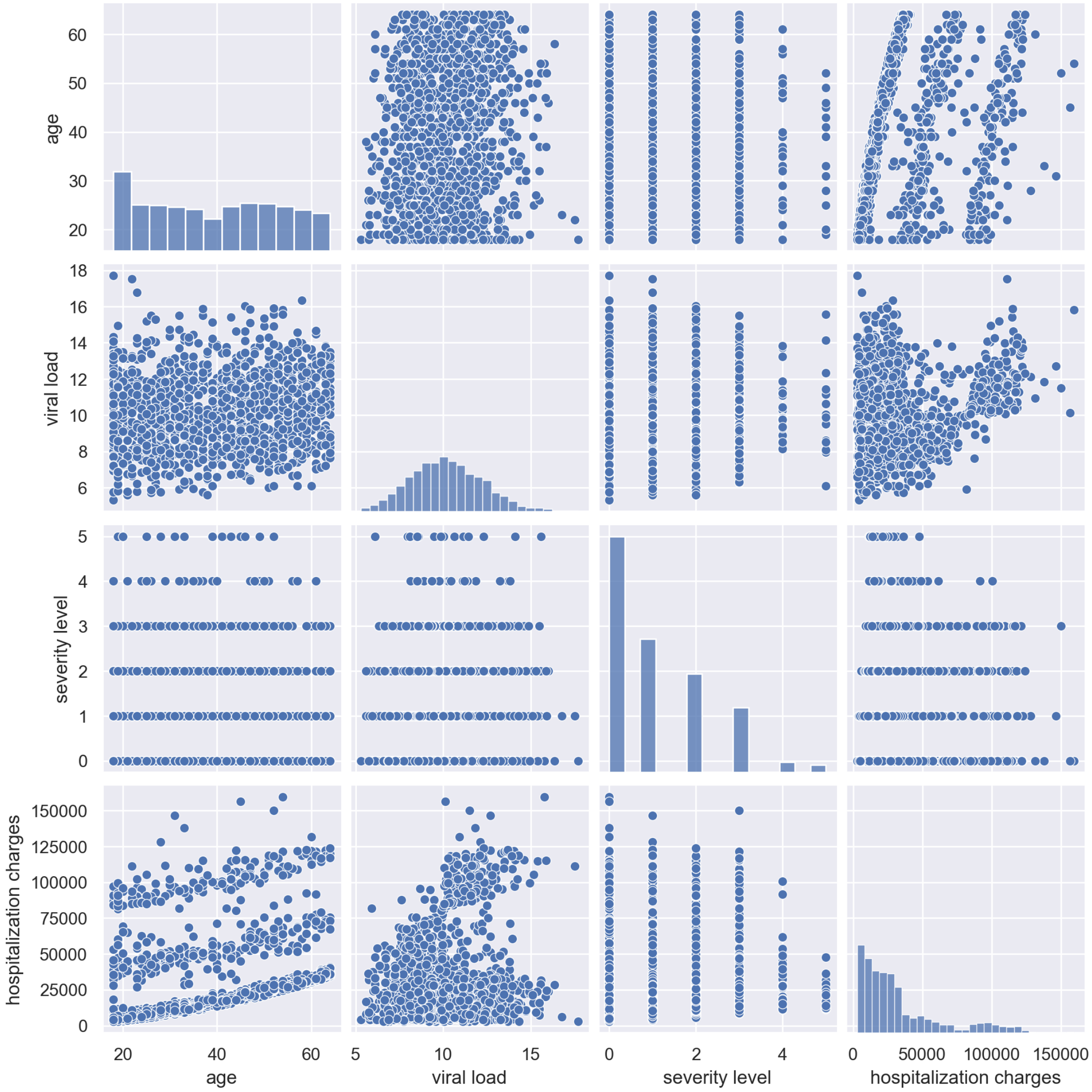
1. age Follows somewhat uniform distribution
2. viral load Looks like the normal distribution
3. hospitalization charges is right skewed

4. age doesn't have any outliers.
5. viral load with few outliers
6. hospitalization charges with lot of outliers

3 Bivariate Analysis

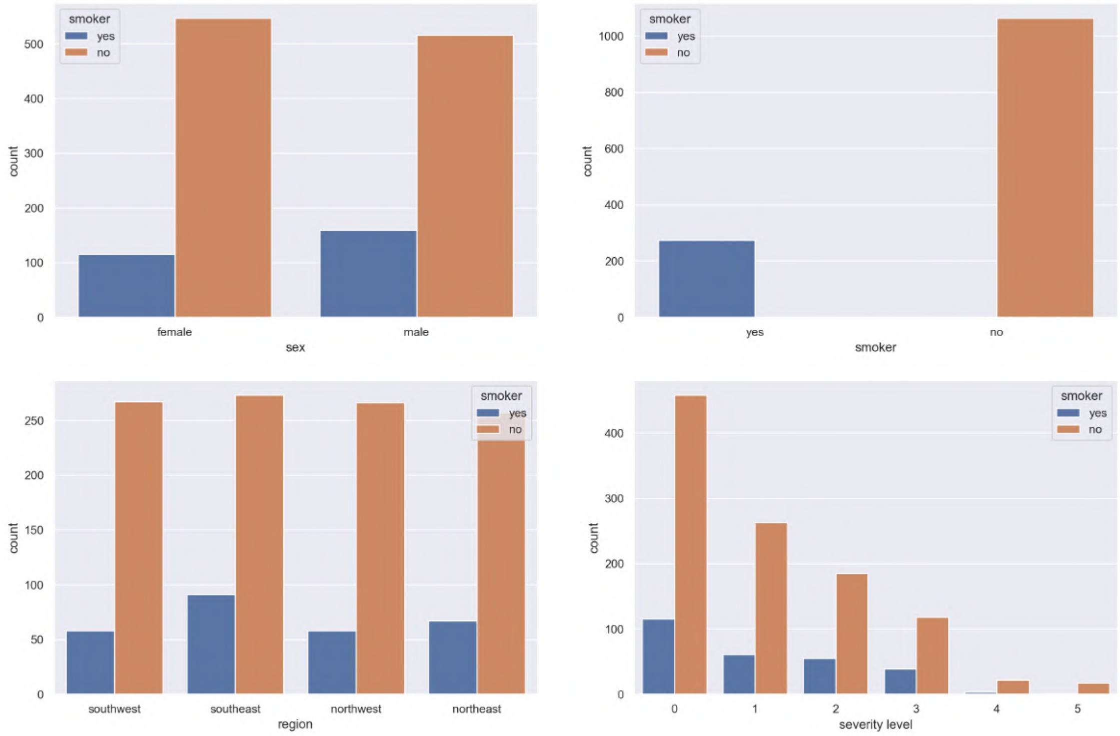
[25]:
```
sns.pairplot(data)
```

[25]:
```
<seaborn.axisgrid.PairGrid at 0x13628ce20>
```



[26]:
```
fig, axs = plt.subplots(nrows=2, ncols=2, figsize=(18, 12))
index = 0
```

```
for i in range(2):
    for j in range(2):
        sns.countplot(hue='smoker', x=categorical_column[index], data=data, ax=axs[i, j])
        index += 1
plt.show()
```



[27]:
```
data.value_counts(['smoker','sex'],normalize=True)
```

[27]:
```
smoker  sex
no      female    0.409125
        male      0.385939
yes     male      0.118923
        female    0.086013
dtype: float64
```

[28]:
```
data.value_counts(['smoker','region'],normalize=True)
```

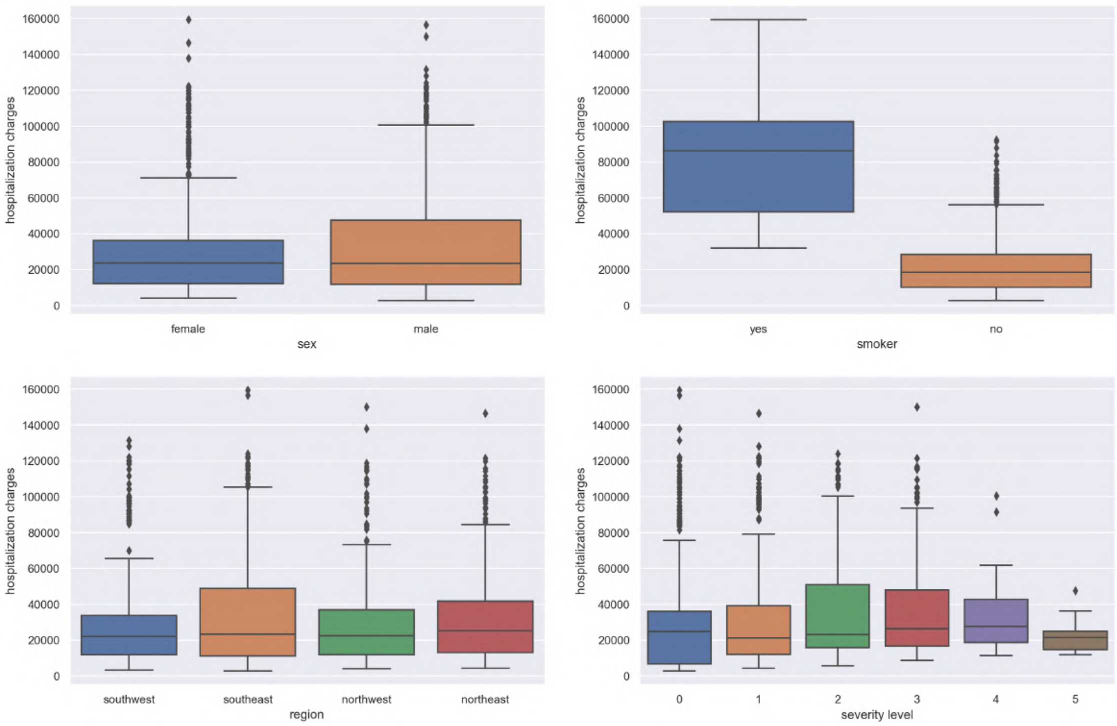
[28]:
```
smoker  region
no      southeast    0.204188
        southwest    0.199701
        northwest    0.198953
        northeast    0.192221
```

```
yes     southeast    0.068063
        northeast    0.050112
        northwest    0.043381
        southwest    0.043381
dtype: float64
```

```
[29]: data.value_counts(['smoker','severity level'],normalize=True)
```

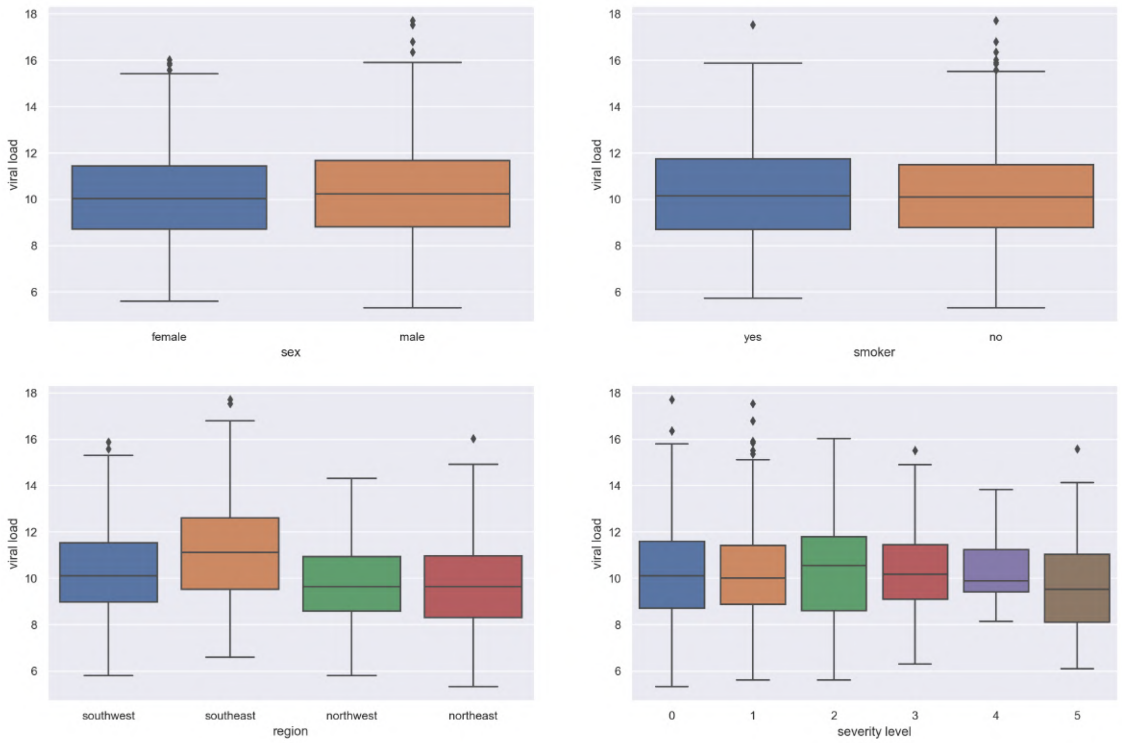
```
[29]: smoker  severity level
      no      0                 0.342558
              1                 0.196709
              2                 0.138369
              3                 0.088257
      yes     0                 0.086013
              1                 0.045625
              2                 0.041137
              3                 0.029170
      no      4                 0.016455
              5                 0.012715
      yes     4                 0.002244
              5                 0.000748
      dtype: float64
```

```
[30]: fig, axs = plt.subplots(nrows=2, ncols=2, figsize=(18, 12))
      index = 0
      for i in range(2):
          for j in range(2):
              sns.boxplot(y='hospitalization charges', x=categorical_column[index],␣
       ↪data=data, ax=axs[i, j])
              index += 1
      plt.show()
```



1. smoker patients will have high hospitalization charges.
2. medians for sex seems like both genders have similar hospitalization charges. statistical test need to get true picture.
3. Patients living in southeast and northeast , have slightly higher hospitalization charges as compared to southweat and northwest respectively.
4. with increase in severity level from 1 to 5 hospitalization charges also increses.

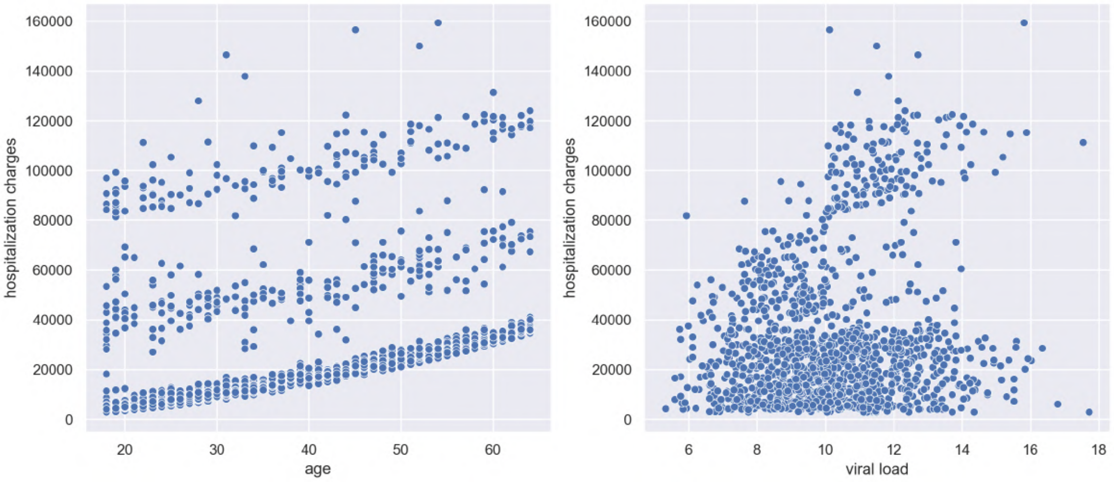
[31]:
```
numerical_column = ['age', 'viral load']
fig, axs = plt.subplots(nrows=2, ncols=2, figsize=(18, 12))
index = 0
for i in range(2):
    for j in range(2):
        sns.boxplot(y='viral load', x=categorical_column[index], data=data,␣
 ↪ax=axs[i, j])
        index += 1
plt.show()
```

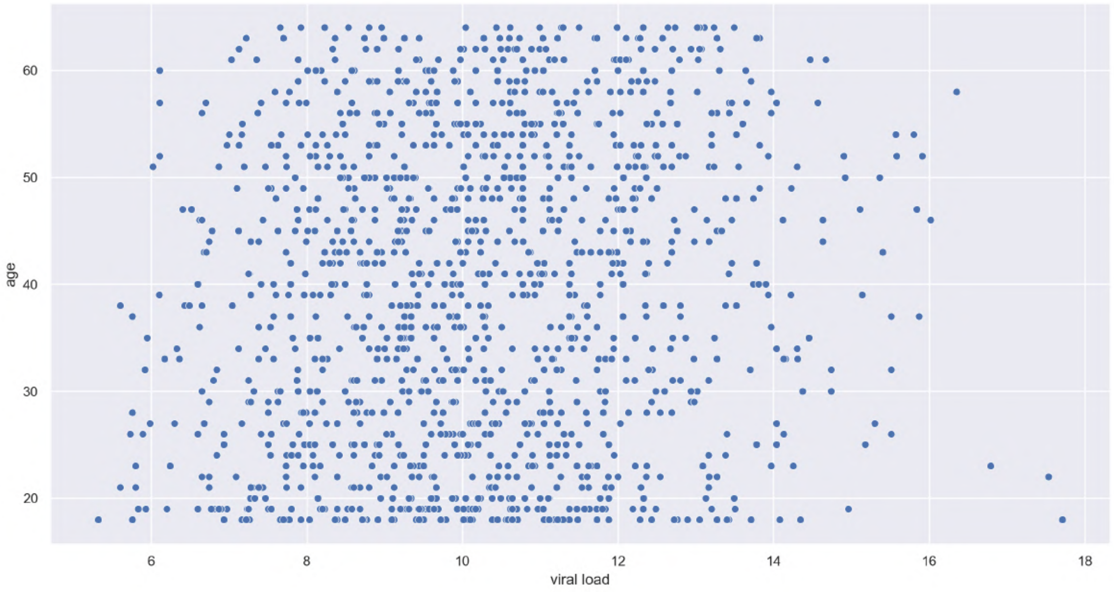


1. viral load not related to either sex or smoker
2. Patients from south east have higher viral load as compared to other regions.

[32]:

```
fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(14, 6))
index = 0
for i in range(2):
    sns.scatterplot(y='hospitalization charges', x=numerical_column[index],␣
 ↪data=data, ax=axs[i])
    index += 1
plt.show()
sns.scatterplot(x='viral load', y='age', data=data)
plt.show()
```





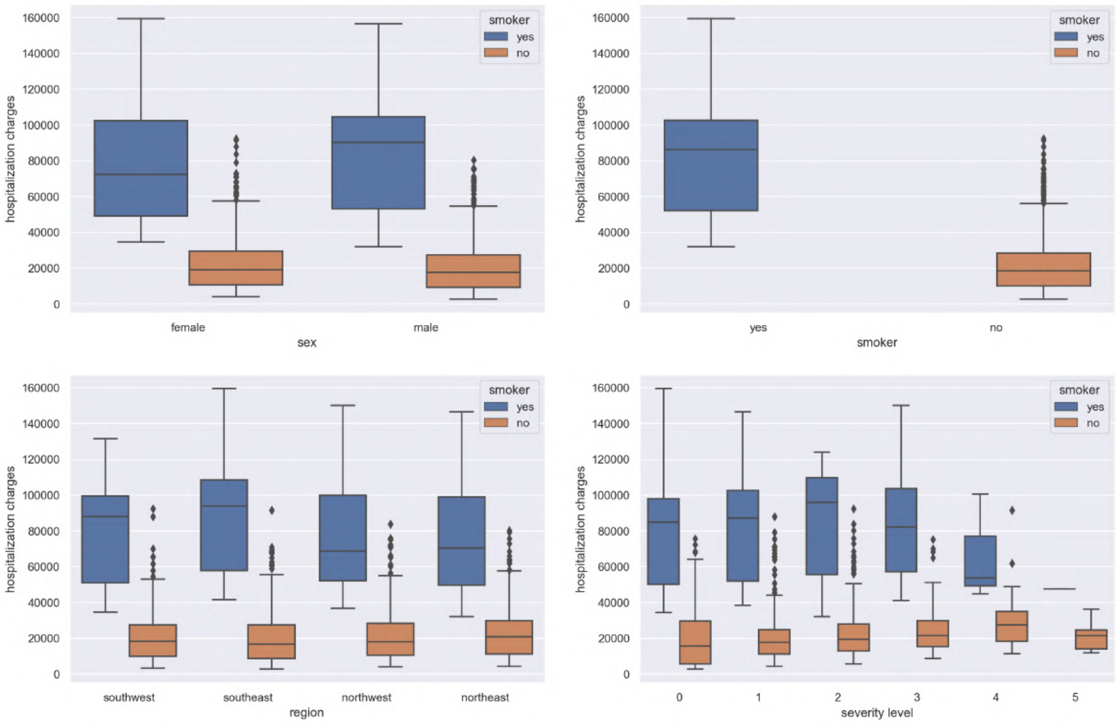
```
[33]: corr = data.corr()
      corr
```

```
[33]:                          age  viral load  hospitalization charges
      age                      1.000000    0.109373                 0.298308
      viral load               0.109373    1.000000                 0.198449
      hospitalization charges  0.298308    0.198449                 1.000000
```

1. slight correlation between hospitalization charges and age
2. alomost No correlation between hospitalization charges and viral load
3. No corelation between viralload and age as well

4 Multivariate Analysis

[34]:
```
fig, axs = plt.subplots(nrows=2, ncols=2, figsize=(18, 12))
index = 0
for i in range(2):
    for j in range(2):
        sns.boxplot(y='hospitalization charges',␣
 ↪x=categorical_column[index],hue='smoker',data=data, ax=axs[i, j])
        index += 1
plt.show()
```



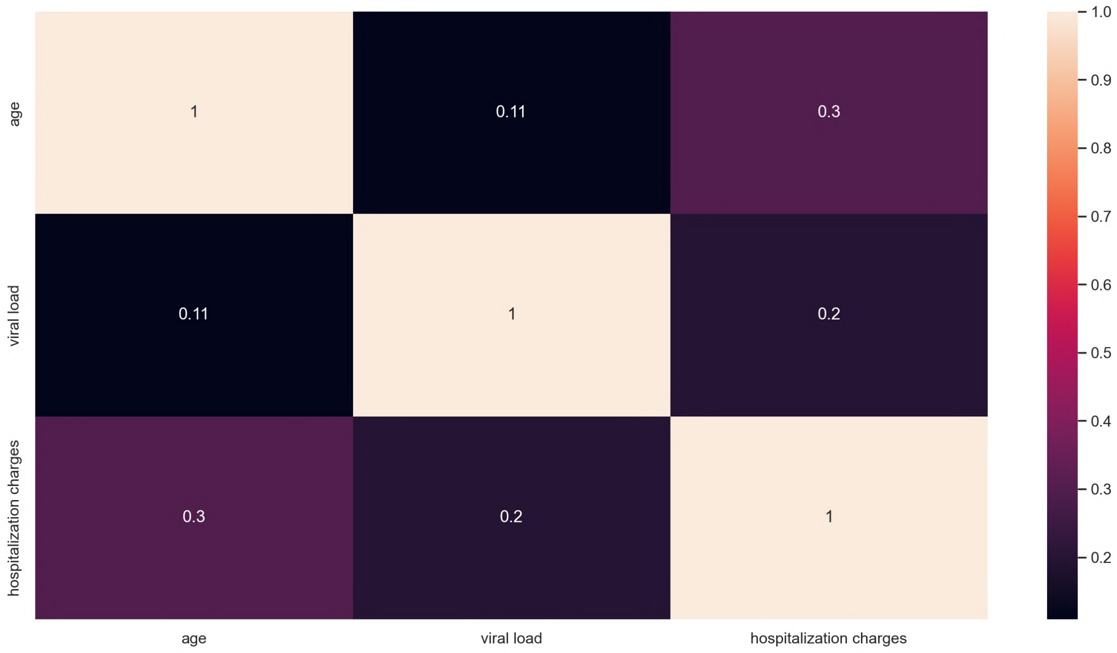
Observations

1. High hospitalization charges for Male smoker patients as compared to Female smoker patients.
2. Hospitalization charges for smoker patient is comparatively larger than non smoker.
3. southeast region have Highest Hospitalization charge for smoker patient in comparision to other region followed by southwest region
4. Hospitalization charges for smoker patient is comparatively greater than non smoker patient having same save severity level.
5. severity level 2 has highest hospitalization mark followed by 1 then 0 for smoker category patient

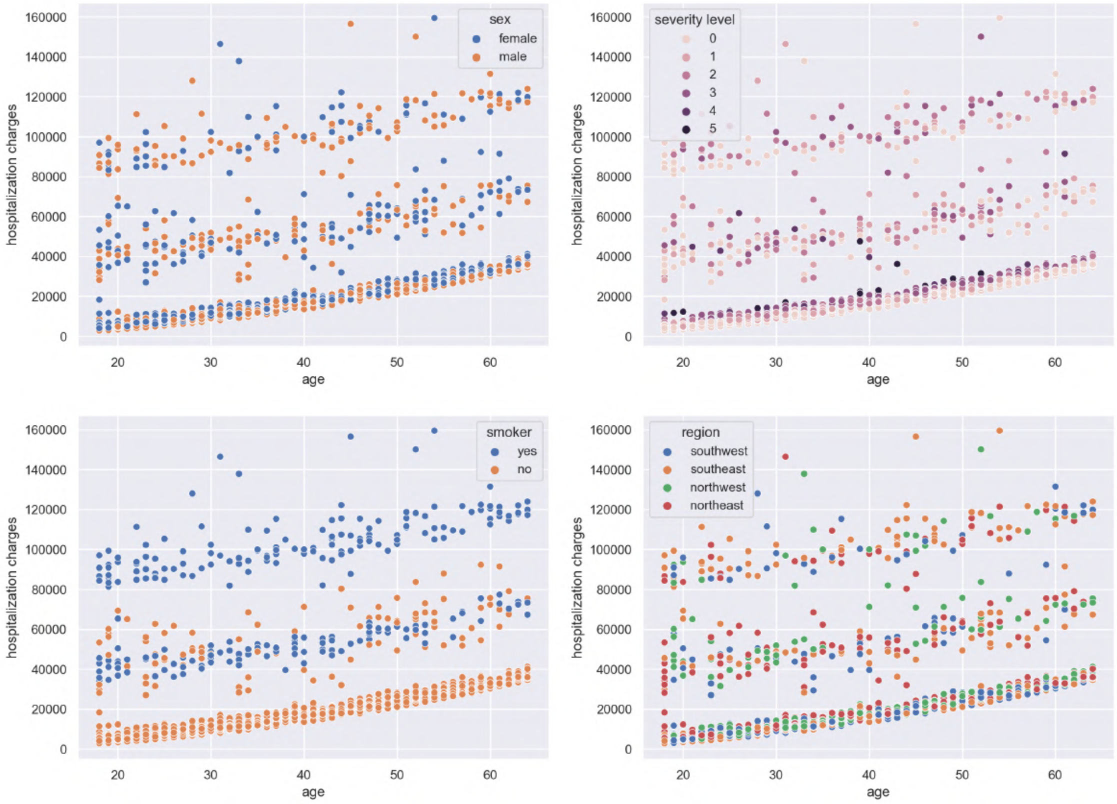
[35]:
```
corr = data.corr()
corr
```

[35]:
```
                              age  viral load  hospitalization charges
age                      1.000000    0.109373                 0.298308
viral load               0.109373    1.000000                 0.198449
hospitalization charges  0.298308    0.198449                 1.000000
```

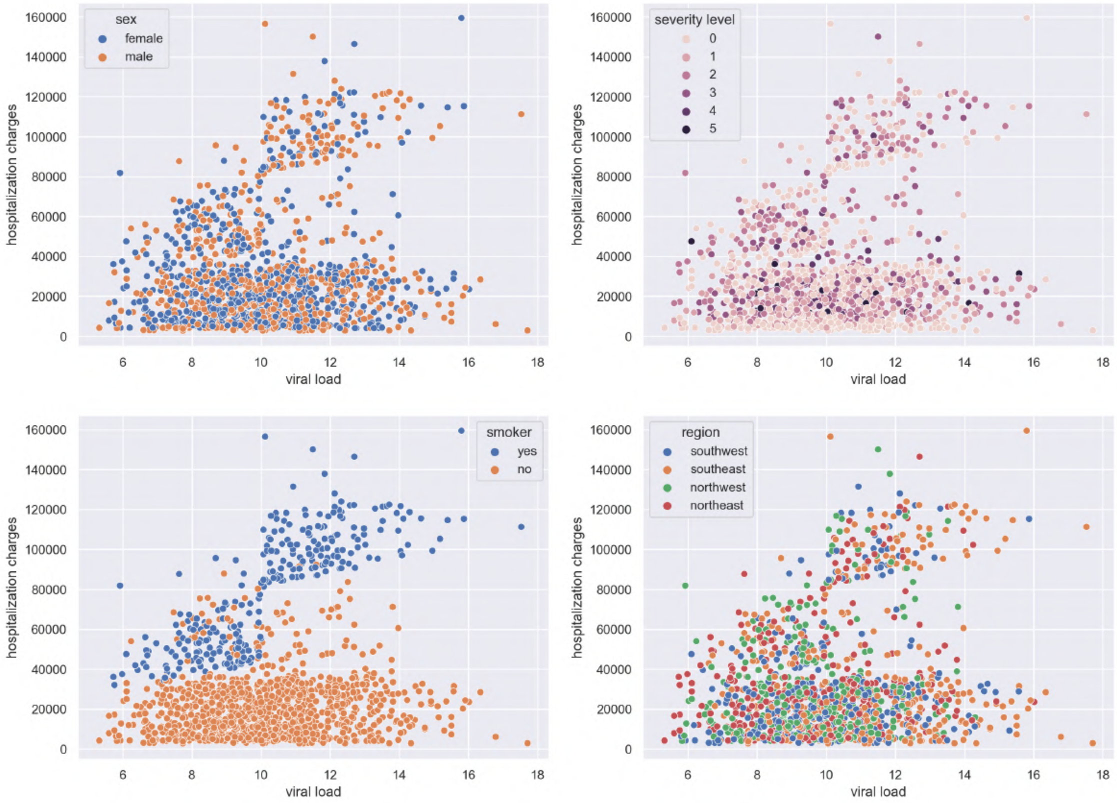
[36]:
```python
sns.heatmap(corr, annot=True)
plt.show()
```



[37]:
```python
fig, axs = plt.subplots(nrows=2, ncols=2, figsize=(16, 12))

sns.scatterplot(y='hospitalization charges', x='age', data=data, hue='sex',␣
 ↪ax=axs[0, 0])
sns.scatterplot(y='hospitalization charges', x='age', data=data, hue='severity␣
 ↪level', ax=axs[0,1])
sns.scatterplot(y='hospitalization charges', x='age', data=data, hue='smoker',␣
 ↪ax=axs[1,0])
sns.scatterplot(y='hospitalization charges', x='age', data=data, hue='region',␣
 ↪ax=axs[1,1])
plt.show()
```

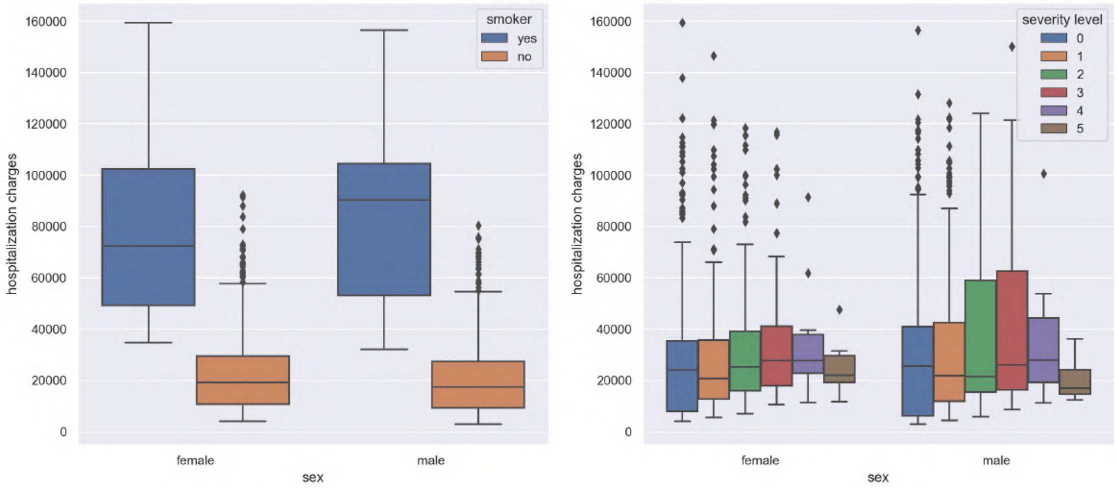


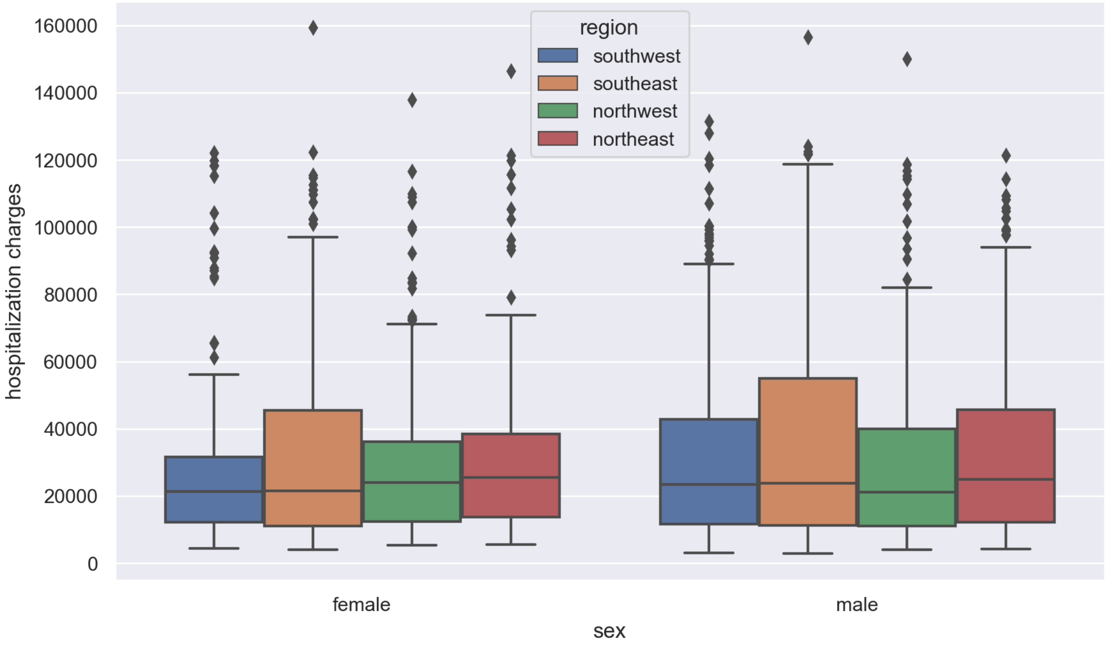
[38]:
```
fig, axs = plt.subplots(nrows=2, ncols=2, figsize=(16, 12))

sns.scatterplot(y='hospitalization charges', x='viral load', data=data, hue='sex', ax=axs[0, 0])
sns.scatterplot(y='hospitalization charges', x='viral load', data=data, hue='severity level', ax=axs[0,1])
sns.scatterplot(y='hospitalization charges', x='viral load', data=data, hue='smoker', ax=axs[1,0])
sns.scatterplot(y='hospitalization charges', x='viral load', data=data, hue='region', ax=axs[1,1])
plt.show()
```



from above charts we can see smoker patients will have high hospitalization charges. other attributes doesn't have any corelation/pattern with hospitalisation charge

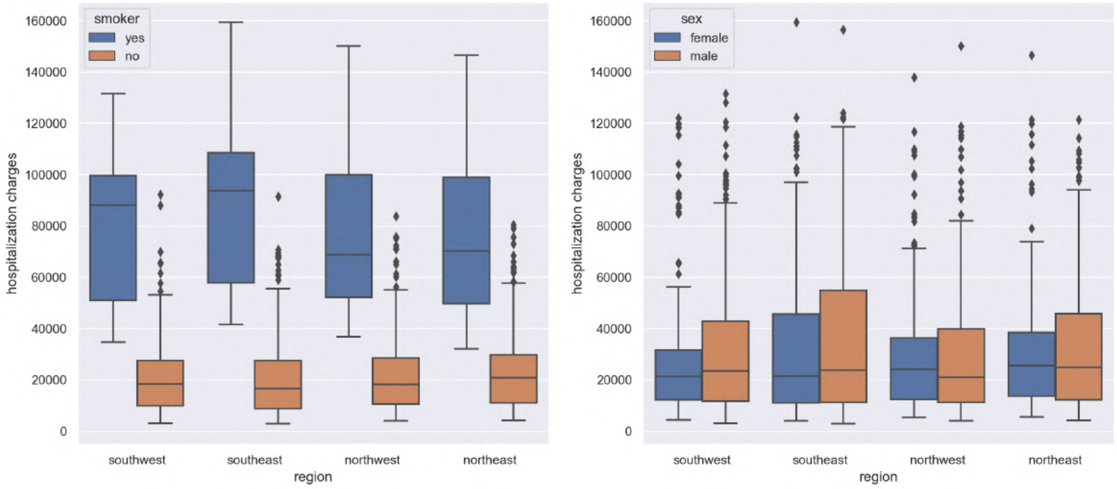
[39]:
```
fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(16, 7))

sns.boxplot(y='hospitalization charges', x='sex', hue='smoker', data=data,␣
 ↪ax=axs[0])
sns.boxplot(y='hospitalization charges', x='sex', hue='severity level',␣
 ↪data=data, ax=axs[1])
plt.show()
plt.figure(figsize=(10,6))
sns.boxplot(y='hospitalization charges', x='sex', hue='region', data=data)
plt.show()
```

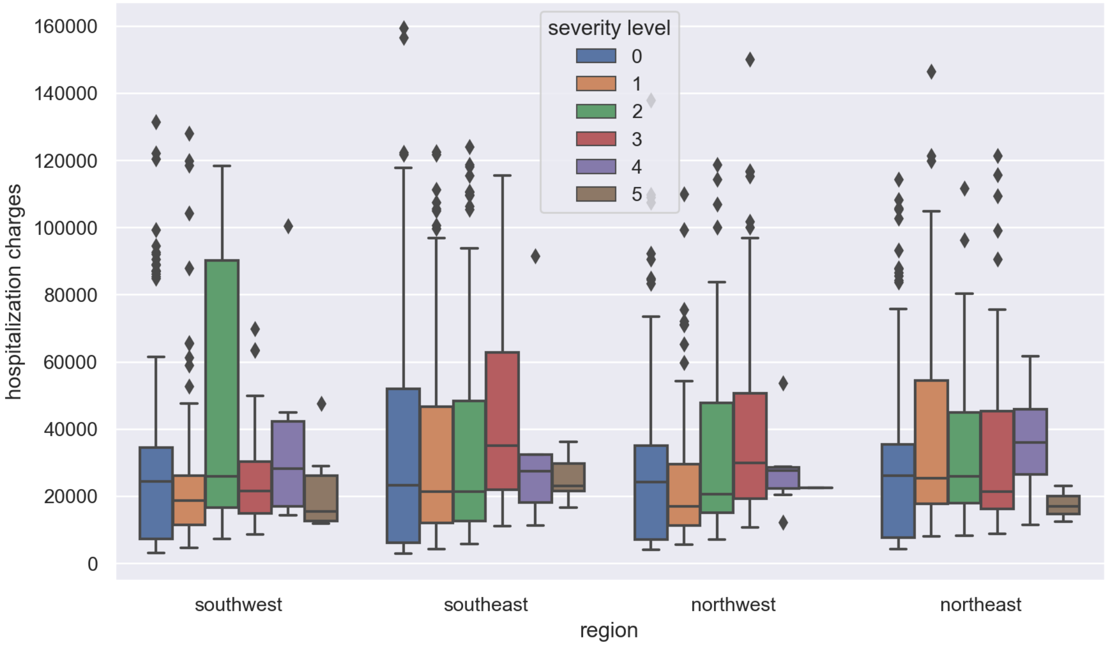




[40]:
```
fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(16, 7))

sns.boxplot(y='hospitalization charges', x='region', hue='smoker', data=data,␣
 ↪ax=axs[0])
sns.boxplot(y='hospitalization charges', x='region', hue='sex', data=data,␣
 ↪ax=axs[1])
plt.show()
plt.figure(figsize=(10,6))
```

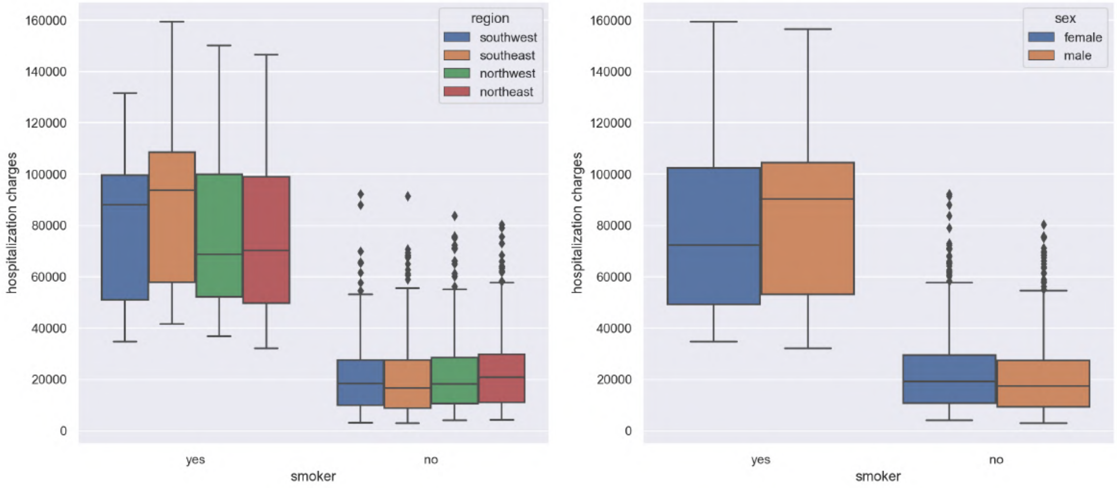
```
sns.boxplot(y='hospitalization charges', x='region', hue='severity level',␣
 ↪data=data)
plt.show()
```

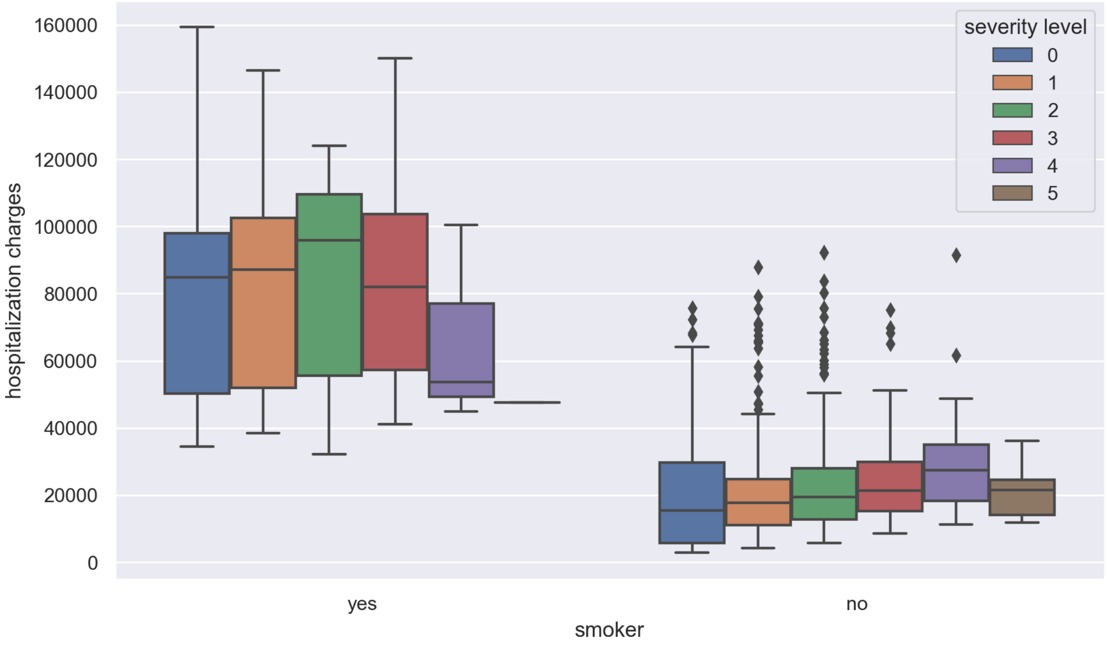




```
[41]: fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(16, 7))

      sns.boxplot(y='hospitalization charges', x='smoker', hue='region', data=data,␣
       ↪ax=axs[0])
```

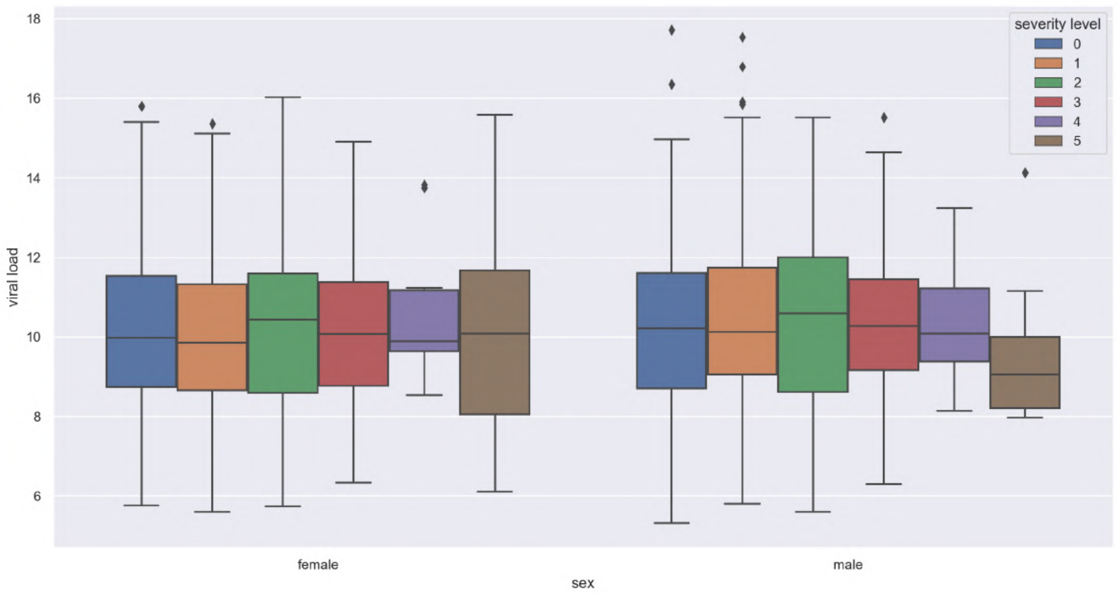
```
sns.boxplot(y='hospitalization charges', x='smoker', hue='sex', data=data, ax=axs[1])
plt.show()
plt.figure(figsize=(10,6))
sns.boxplot(y='hospitalization charges', x='smoker', hue='severity level', data=data)
plt.show()
```





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

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



Observations

1. Male smoker patients have high hospitalization charges as compared to Female smoker patients .
2. female with severity level 3 & 4 will have high hospitalization charges while male with severity level 0 & 4 will have almost similary hospitalization charges .
3. In smoker patients , patients with severity level - 2 will have higher hospitalization charges as comapred to other severity levels.
4. In non-smoker patients , patients with severity level - 4 will have higher hospitalization charges as compared to other severity lavels.
5. In smoker patients , patients living in southwest region will have higher hospitalization charges as compared to other regions.
6. male patients in southweat & southeast regions will have slightly higher hospitalization charges as compare females while female patients in northweat & northeast will have slightly higher hospitalization charges as compared to male patients.
7. smoker patients living in southwest & southeast regions will have high hospitalization charges as compared to smoker patients living in north regions.
8. Patients living in southwest & northeast and have severity level-4 will have higher hospitalization charges as compared to other severity level patients
9. Patients living in southeast & northwest and have severity level-3 will have higher hospitalization charges as compared to other severity level patients

4.0.1 Outlier treatment

[43]:

```
from scipy import stats
five_point_summary=data.describe().T
five_point_summary['IQR']=np.
 ↪round(five_point_summary['75%']-five_point_summary['25%'],2)
five_point_summary['Upper Whisker']=np.round(five_point_summary['75%']+1.
 ↪5*five_point_summary['IQR'],2)
five_point_summary['Lower Whisker']=np.round(five_point_summary['25%']-1.
 ↪5*five_point_summary['IQR'],2)
five_point_summary
```

[43]:

```
                         count          mean           std      min  \
age                     1337.0     39.222139     14.044333    18.00
viral load              1337.0     10.221249      2.033556     5.32
hospitalization charges 1337.0  33197.806283  30275.900411  2805.00

                              25%       50%       75%        max       IQR  \
age                         27.00     39.00     51.00      64.00     24.00
viral load                   8.76     10.13     11.57      17.71      2.81
hospitalization charges  11866.00  23465.00  41644.00  159426.00  29778.00

                         Upper Whisker  Lower Whisker
age                              87.00          -9.00
viral load                       15.78           4.54
hospitalization charges       86311.00      -32801.00
```

[44]:

```
numerical_column=['age', 'viral load','hospitalization charges']
```

[45]:

```
outlier_count=data[(data[numerical_column]>five_point_summary.T.loc['Upper␣
 ↪Whisker',numerical_column])|(data[numerical_column]<five_point_summary.T.
 ↪loc['Lower Whisker',numerical_column])].count()
print(outlier_count)
outlier_count_percent=(outlier_count/len(data))*100
print("outlier count in %")
outlier_count_percent[numerical_column]
```

```
age                          0
sex                          0
smoker                       0
region                       0
viral load                   9
severity level               0
hospitalization charges    139
dtype: int64
outlier count in %
```

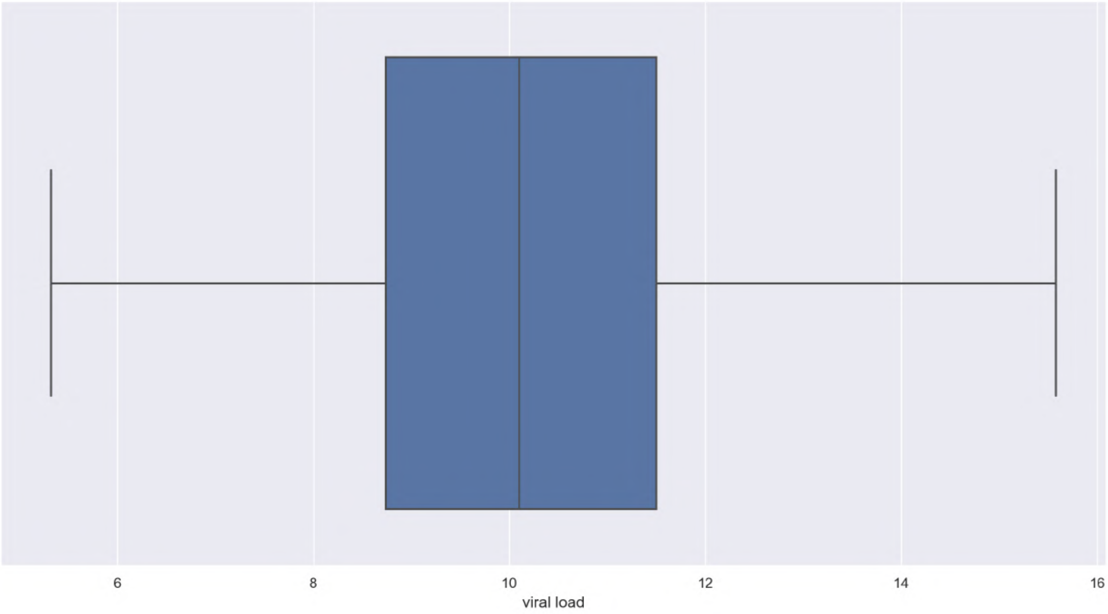
[45]:
```
age                        0.000000
viral load                 0.673149
hospitalization charges   10.396410
dtype: float64
```

[46]:
```
#since viral load has just 9 outlier which 0.67 so we can remove that
new_data= data
q1 = np.percentile(new_data['viral load'], 25)
q3 = np.percentile(new_data['viral load'], 75)
iqr = q3-q1
new_data= new_data[(new_data['viral load'] > (q1-1.5*iqr)) & (new_data['viral␣
 ↪load'] < (q3+1.5*iqr))]
```

[47]:
```
# In order to remove outlier from hospitalization lets take log and check

data['hospitalization charges']=np.log(data['hospitalization charges'])
```

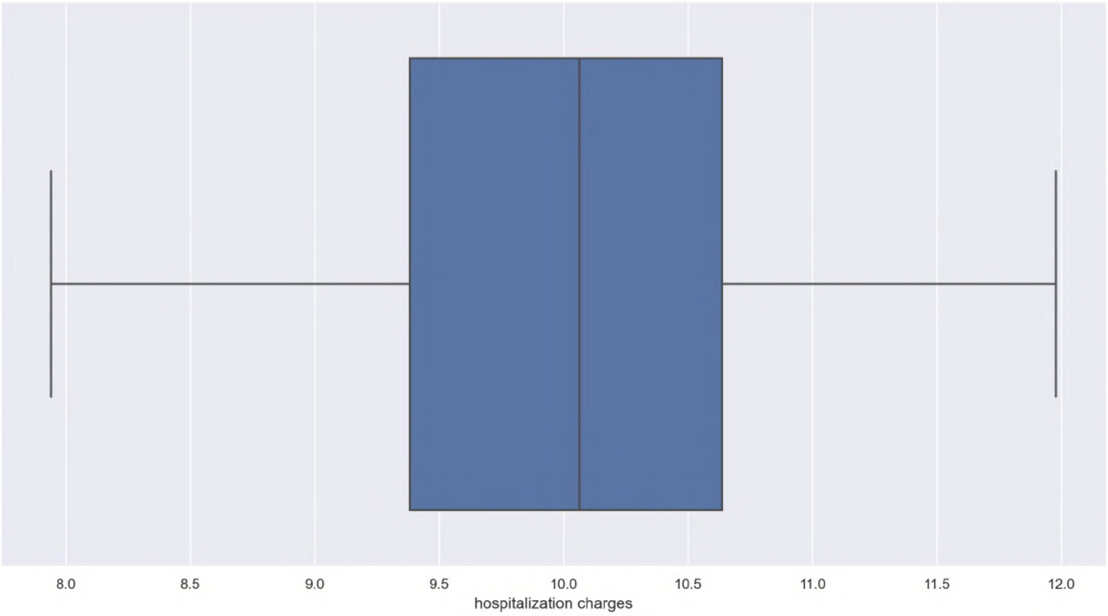
[48]:
```
sns.boxplot(x='viral load',data=new_data)
```

[48]:
```
<AxesSubplot:xlabel='viral load'>
```



[49]:
```
sns.boxplot(x='hospitalization charges',data=data)
```

[49]:
```
<AxesSubplot:xlabel='hospitalization charges'>
```



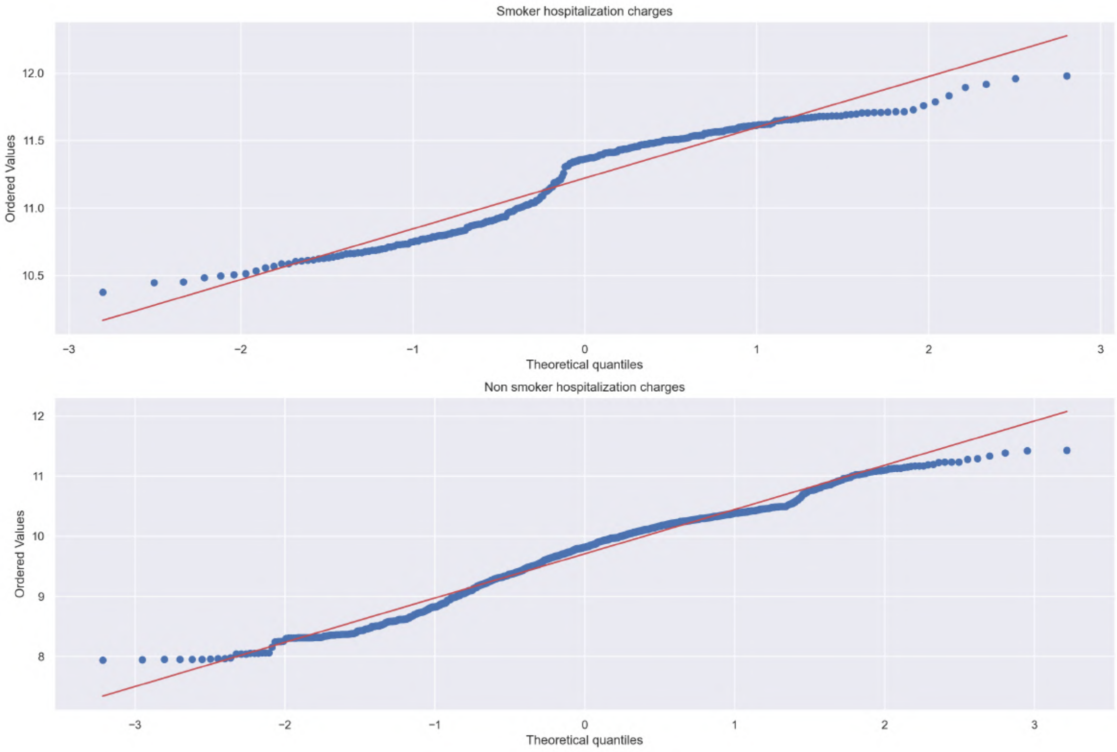
Hypothesis Testing - 1

1. **Null Hypothesis:** Mean hospitalization charges for smoker and non smoker are same.
2. **Alternate Hypothesis:** Mean hospitalization charges are greater for smoker than those who don't smoke.
3. **Significance Level:** 0.05

using right tailed 2-sample T-test.

[50]:
```
smoker = data[data['smoker']=='yes']['hospitalization charges']
non_smoker = data[data['smoker']=='no']['hospitalization charges']
```

[51]:
```
# normality test
fig, axs = plt.subplots(nrows=2, ncols=1, figsize=(18,12))

stats.probplot(smoker, plot=axs[0])
stats.probplot(non_smoker, plot=axs[1])

axs[0].set_title("Smoker hospitalization charges")
axs[1].set_title("Non smoker hospitalization charges")
plt.show()
```



values looks much closer to normal distribution, so we can proceed with the test as our assumption of normality holds

[52]:
```
res = stats.ttest_ind(smoker, non_smoker, alternative='greater')
res
```

[52]:
```
Ttest_indResult(statistic=32.59670864311422, pvalue=3.0113983195275365e-172)
```

p-value is less than 0.05 we reject the null hypothesis i.e failed to accept null hypothesis. Meaning hospitalization charges for smoker are greater than non smoker.

Hypothesis Testing - 2

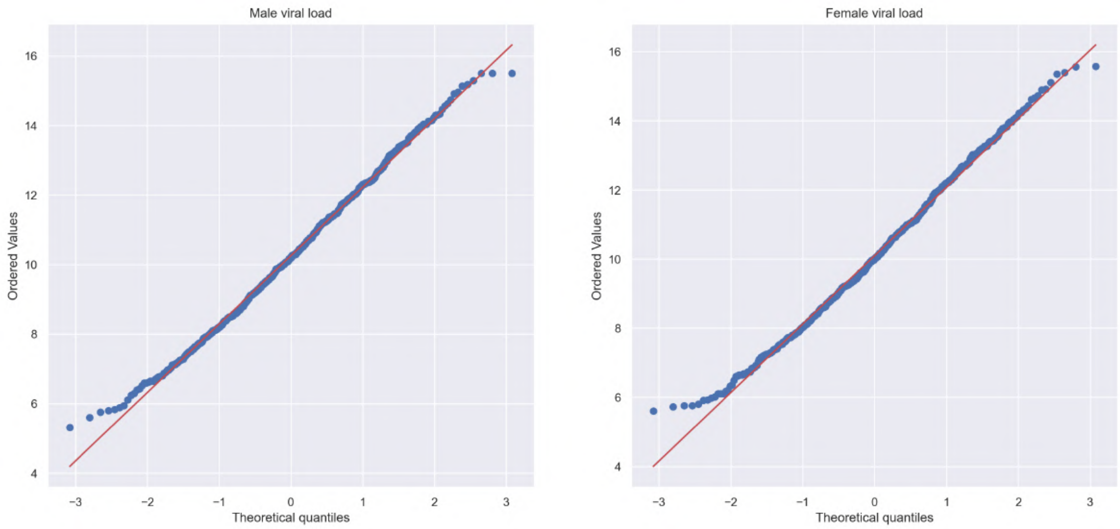
1. **Null Hypothesis:** Viral load is same for both male and females.
2. **Alternate Hypothesis:** Viral load is different for male and females.
3. **Significance Level:** 0.05

Here we will be using two tailed 2-sample T-test.

[53]:
```
male = new_data[new_data['sex']=='male']['viral load']
female = new_data[new_data['sex']=='female']['viral load']
```

[54]:
```
# normality check before and after log transformations
fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(18,8))
```

```
stats.probplot(male, plot=axs[0])
stats.probplot(female, plot=axs[1])

axs[0].set_title("Male viral load")
axs[1].set_title("Female viral load")
plt.show()
```



above values very much closer to normal distribution, so we can proceed with the test as our assumption of normality holds

[55]:
```
res = stats.ttest_ind(male, female, alternative='two-sided')
res
```

[55]:
```
Ttest_indResult(statistic=1.4554178775791846, pvalue=0.14579030641597715)
```

p-value is greater than 0.05, hence not enough evidence to reject null hypothesis so viral load of females is no differnet from that of males.

Hypothesis Testing - 3

1. **Null Hypothesis:** proportion of smoking is same across different regions.
2. **Alternate Hypothesis:** proportion of smoking is different across different regions.
3. **Significance Level:** 0.05

Here we will be using chisquare test.

[56]:
```
data_table = pd.crosstab(data['smoker'], data['region'])
print("Oberved values:")
data_table
```

```
Oberved values:
```

[56]:

```
region  northeast  northwest  southeast  southwest
smoker
no            257        266        273        267
yes            67         58         91         58
```

[57]:

```
# contingency table
stat, p, dof, expected = chi2_contingency(data_table)
print('dof=%d' % dof)
print(expected)
# interpret test-statistic
prob = 0.95
critical = chi2.ppf(prob, dof)
print('probability=%.3f, critical=%.3f, stat=%.3f' % (prob, critical, stat))
if abs(stat) >= critical:
    print('Dependent (reject H0)')
else:
    print('Independent (fail to reject H0)')
# interpret p-value
alpha = 1.0 - prob
print('significance=%.3f, p=%.3f' % (alpha, p))
if p <= alpha:
     print('Dependent (reject H0)')
else:
     print('Independent (fail to reject H0)')
```

```
dof=3
[[257.60059835 257.60059835 289.40314136 258.39566193]
 [ 66.39940165  66.39940165  74.59685864  66.60433807]]
probability=0.950, critical=7.815, stat=7.278
Independent (fail to reject H0)
significance=0.050, p=0.064
Independent (fail to reject H0)
```

p-value is greater than 0.05, hence not enough evidence to reject null hypothesis

so proportion of smoking is same across different regions.

Hypothesis Testing - 4

1. **Null Hypothesis:** mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level the equal.
2. **Alternate Hypothesis:** mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level the are unequal.
3. **Significance Level:** 0.05

Here we will be using Anova

[58]:

```
sample_data=new_data[new_data['sex']=='female']
data1 = sample_data[sample_data['severity level']==0]['viral load']
```

```
data2 = sample_data[sample_data['severity level']==1]['viral load']
data3 = sample_data[sample_data['severity level']==2]['viral load']
```

[59]:

```
# Example of the Analysis of Variance Test
from scipy.stats import f_oneway
stat, p = f_oneway(data1, data2, data3)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')
```

```
stat=0.117, p=0.889
Probably the same distribution
```

p-value is greater than 0.05, hence not enough evidence to reject null hypothesis

so mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level the equal.

4.0.2 Business Insights

1. we have almost equal amount of male and female customer , 49.51% female customer and 50.48% male customer respectively.
2. 79% non smoker and 20.49% smoker patient
3. from southeast we have 27.22% patient and almost 24% from all other region i.e southwest,northwest,norteast
4. smoker patients have high hospitalization charges.
5. Mean hospitalization charges are greater for smoker than those who don't smoke.
6. viral load of females is no differnet from that of males.
7. proportion of smoking is same across different regions.
8. mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level the equal.
9. Patients living in southeast and northeast , have slightly higher hospitalization charges as compared to southweat and northwest respectively.
10. with increase in severity level from 1 to 5 hospitalization charges also increses.
11. High hospitalization charges for Male smoker patients as compared to Female smoker patients. 12.female with severity level 3 & 4 will have high hospitalization charges while male with severity level 0 & 4 will have almost similary hospitalization charges .
12. In smoker patients , patients with severity level - 2 will have higher hospitalization charges as comapred to other severity levels.
13. In non-smoker patients , patients with severity level - 4 will have higher hospitalization charges as compared to other severity lavels.

4.0.3 Recommendations

1. More Awarness drive can be created among the people describing consequences of smoking/any viral pandemic regarding infection ratio among various categories .

2. since Patients from south east have higher viral load as compared to other regions so number of hospitals can be increased in that reason for providing easy and smooth treatment facility.

3. smokers are more prone to chronic bronchitis cancerous diseases and we have good amount of patients belong to smoker categories , hence hospital can make sure to have good quality of doctor and treatment facilities for diseases cause by smoking along with other diseases.

[ ]:

```
```