Business Case: Apollo Hospitals - Hypothesis Testing

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.

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

How to begin

- Import the dataset and do usual exploratory data analysis steps like checking the structure & characteristics of the dataset
- Try establishing a relation between the dependent and independent variable (Dependent “hospitalization charges” & Independent: Smoker, Severity Level etc)

Statistical Analysis:

- Prove (or disprove) that the hospitalization of people who do smoking is greater than those who don't? (T-test Right tailed)
- Prove (or disprove) with statistical evidence that the viral load of females is different from that of males (T-test Two tailed)
- Is the proportion of smoking significantly different across different regions? (Chi-square)
- 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)
- Set up Null Hypothesis (H0)
- State the alternate hypothesis (H1)
- Check assumptions of the test (Normality, Equal Variance). You can check it using Histogram, Q-Q plot or statistical methods like levene’s test, Shapiro-wilk test (optional)
- Please continue doing the analysis even If some assumptions fail (levene’s test or Shapiro-wilk test) but double check using visual analysis and report wherever necessary
- Set a significance level (alpha)
- Calculate test Statistics.
- Decision to accept or reject null hypothesis.
- Inference from the analysis

Evaluation Criteria (80 Points)

- Define Problem Statement and perform Exploratory Data Analysis (10 points)
- Definition of problem (as per given problem statement with additional views)
- Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (if required), missing value detection, statistical summary.
- Univariate Analysis (distribution plots of all the continuous variable(s) barplots/countplots of all the categorical variables)
- Bivariate Analysis (Relationships between important variables such as hospitalization charges with region, smoker, viral load etc)
- Illustrate the insights based on EDA
- Comments on range of attributes, outliers of various attributes
- Comments on the distribution of the variables and relationship between them
- Comments for each univariate and bivariate plots
- Missing values treatment & Outlier treatment (10 Points)

Hypothesis Testing (40 Points):

- Prove (or disprove) that the hospitalization charges of people who do smoking are greater than those who don't? (10 Points)
- Prove (or disprove) with statistical evidence that the viral load of females is different from that of males (10 Points)
- Is the proportion of smoking significantly different across different regions? (10 Points)
- 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 (10 Points)
- What good looks like (distribution of 10 points):
- Visual analysis (2)
- Hypothesis formulation (1)
- Select the appropriate test (1)
- Check test assumptions (4)
- Find the p-value(1)
- Conclusion based on the p-value(1)
- Business Insights (10 Points) - Should include patterns observed in the data along with what you can infer from it.
- Recommendations(10 Points) - Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand.

In [116]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from scipy.stats import norm
from scipy import stats
from scipy.stats import t
from scipy.stats import binom
from scipy.stats import ttest_ind
from scipy.stats import ttest_1samp
from scipy.stats import ttest_rel
from scipy.stats import chi2_contingency
from scipy.stats import f_oneway

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

THis dataset is focusing on Delhi Region Only

In [16]:
```
df=pd.read_csv(r"C:\Users\Sweta.Singh\Downloads\scaler_apollo_hospitals.csv")
```

In [31]:
```
(df)
```

Out[31]:

	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
...	...	...	...	...	...	...	...
1333	50	male	no	northwest	10.32	3	26501
1334	18	female	no	northeast	10.64	0	5515
1335	18	female	no	southeast	12.28	0	4075
1336	21	female	no	southwest	8.60	0	5020
1337	61	female	yes	northwest	9.69	0	72853

1338 rows × 7 columns

In [18]:
```
df.shape
```

Out[18]:
```
(1338, 7)
```

In [19]:
```
df.head()
```

Out[19]:

	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 [20]:
```
df.dtypes
```

Out[20]:
```
age                        int64
sex                       object
smoker                    object
region                    object
viral load               float64
severity level             int64
hospitalization charges    int64
dtype: object
```

In [21]:
```
df.isnull().sum()
```

Out[21]:
```
age                        0
sex                        0
smoker                     0
region                     0
viral load                 0
severity level             0
hospitalization charges    0
dtype: int64
```

In [22]:
```
df.describe()
```

Out[22]:

	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

In [23]:
```
df.describe(include='object')
```

Out[23]:

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

In [70]:
```
print(df['sex'].unique())
```

```
['female' 'male']
```

In [71]:
```
print(df['smoker'].unique())
```

```
['yes' 'no']
```

In [72]:
```
print(df['region'].unique())
```

```
['southwest' 'southeast' 'northwest' 'northeast']
```

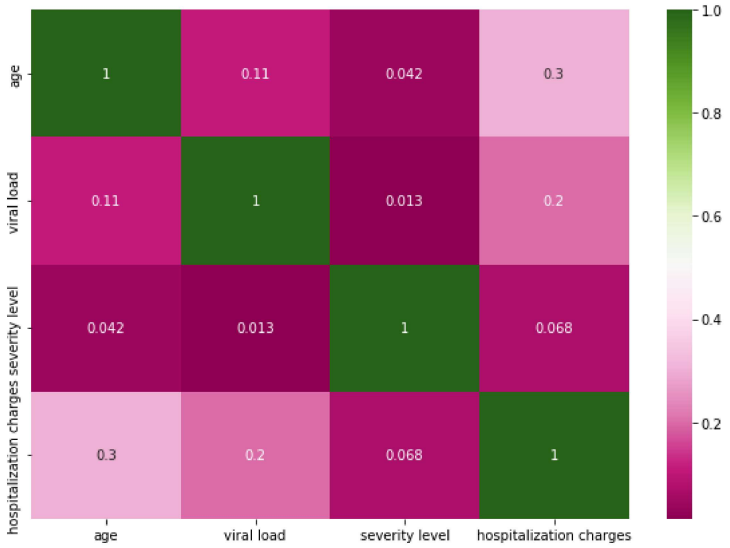
In [24]:
```
df.corr()
```

Out[24]:

	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

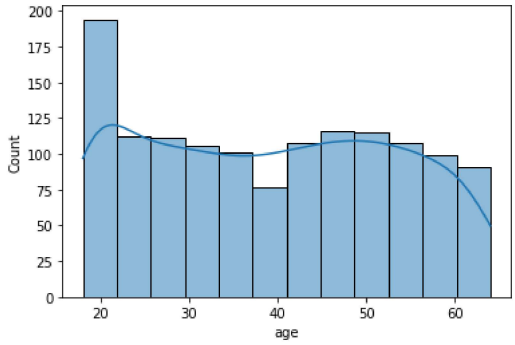
In [28]:
```
plt.figure(figsize=(10,7))
sns.heatmap(df.corr(),annot=True,cmap="PiYG")
```

Out[28]: <AxesSubplot:>

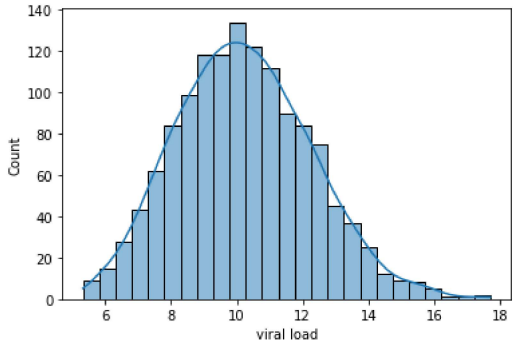


Distribution of numerical variables

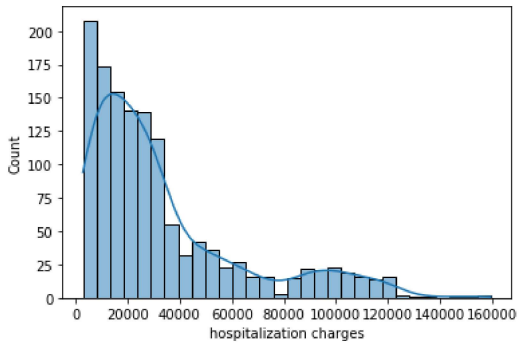
In [38]:
```
sns.histplot(df['age'],kde=True)
plt.show()
```



In [39]:
```
sns.histplot(df['viral load'], kde=True)
plt.show()
```

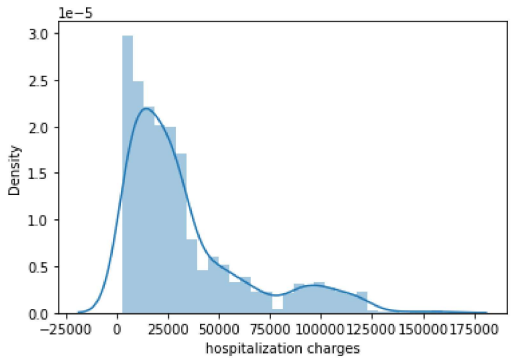


In [40]:
```
sns.histplot(df['hospitalization charges'],kde=True)
plt.show()
```

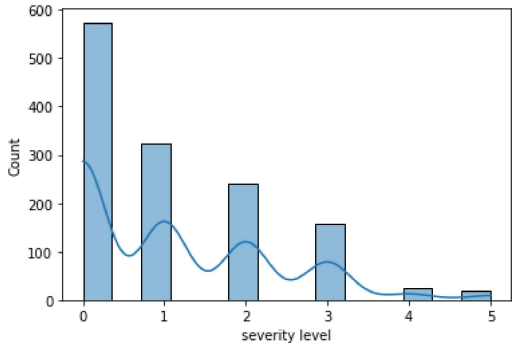


In [91]:
```
sns.distplot(df['hospitalization charges'])
```

Out[91]:
```
<AxesSubplot:xlabel='hospitalization charges', ylabel='Density'>
```

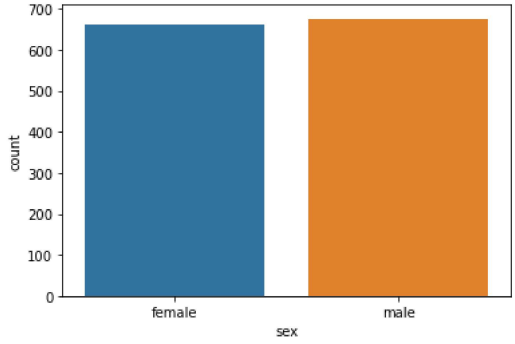


In [43]:
```
sns.histplot(df['severity level'],kde=True)
plt.show()
```

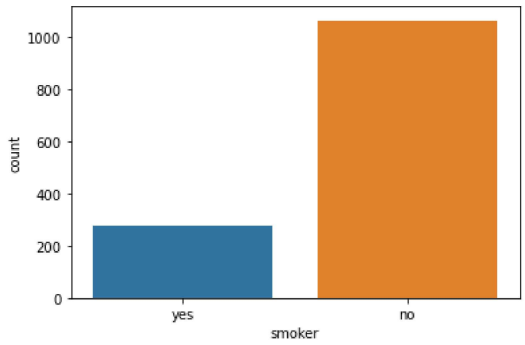


Distribution of categorical variables

In [41]:
```
sns.countplot(x='sex',data=df)
plt.show()
```

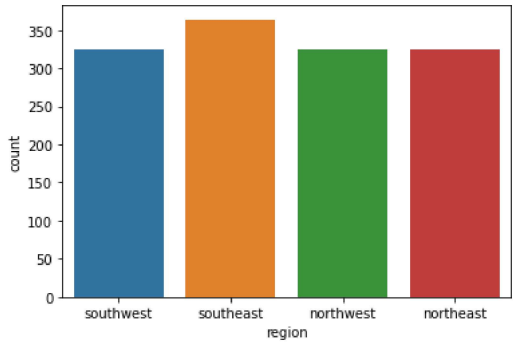


In [42]:
```
sns.countplot(x='smoker',data=df)
plt.show()
```



- The number of smokers are less.

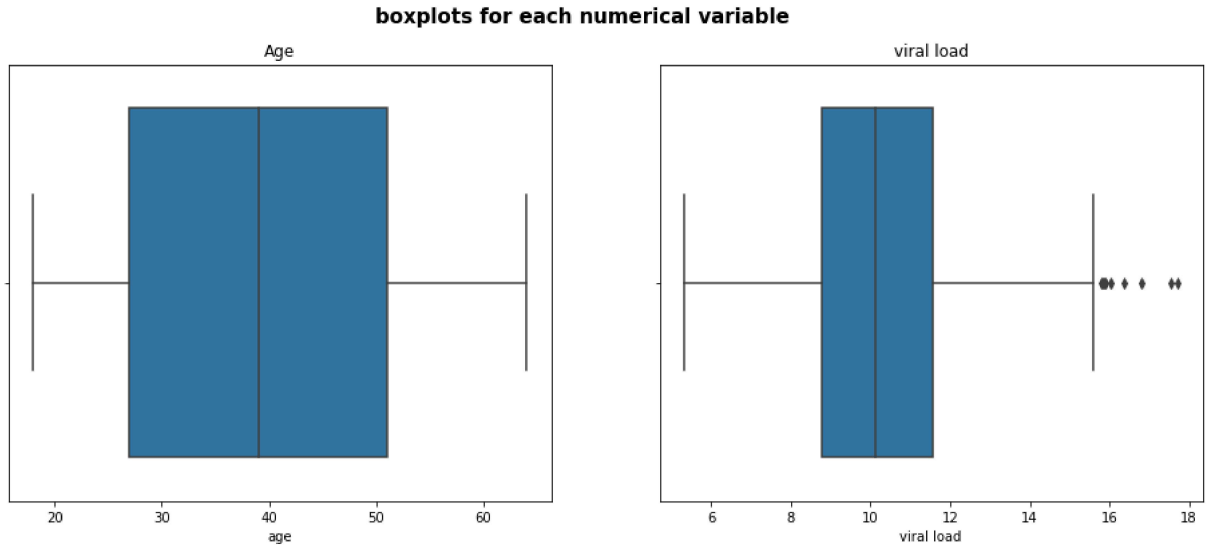
In [44]:
```
sns.countplot(x='region',data=df)
plt.show()
```



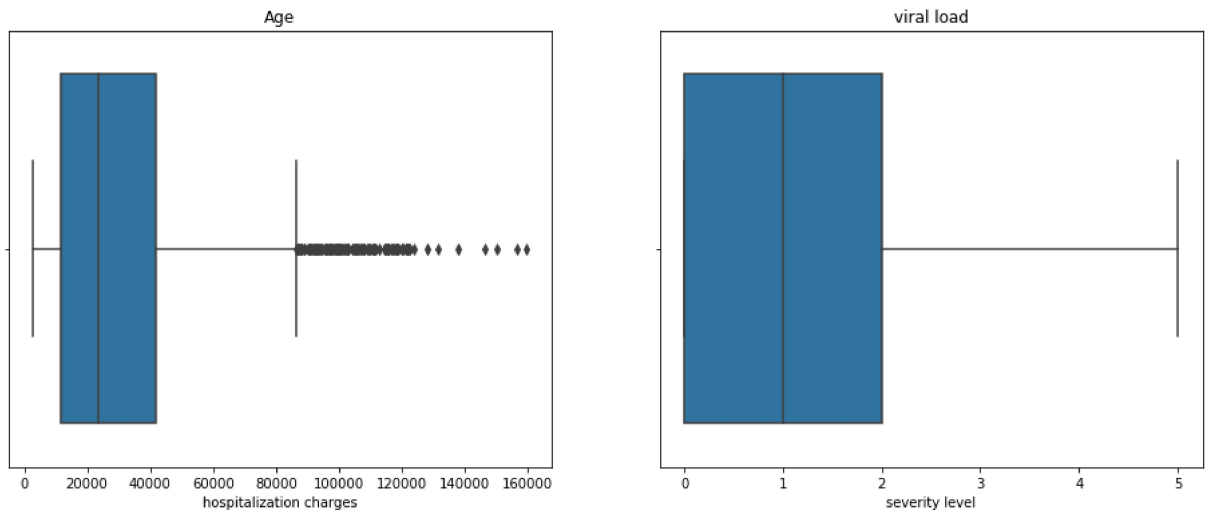
- we can see that number of counts in southeast d=region is more as compared to others.

Create boxplots for each numerical variable

In [60]:
```
# fig,axs=plt.subplots(ncols=4,figsize=(20,5))
# sns.boxplot(x=df['age'],ax=axs[0])
# sns.boxplot(x=df['viral load'],ax=axs[1])
# sns.boxplot(x=df['hospitalization charges'], ax=axs[2])
# sns.boxplot(x=df['severity level'], ax=axs[3])
# plt.show()

fig, axs = plt.subplots(1, 2, figsize=(16, 6), sharey=False)
suptitle = "boxplots for each numerical variable "
fig.suptitle(suptitle , fontsize=15, fontweight='bold')
graph = sns.boxplot(x=df['age'],dodge = False , ax = axs[0])
title1 = "Age"
graph.set_title(title1, fontsize = 12)
graph2 = sns.boxplot(x=df['viral load'], dodge=False, ax = axs[1])
title1 = "viral load"
graph2.set_title(title1,fontsize=12)
plt.show()
```

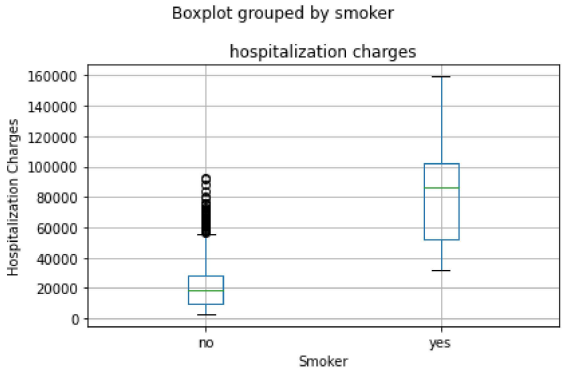


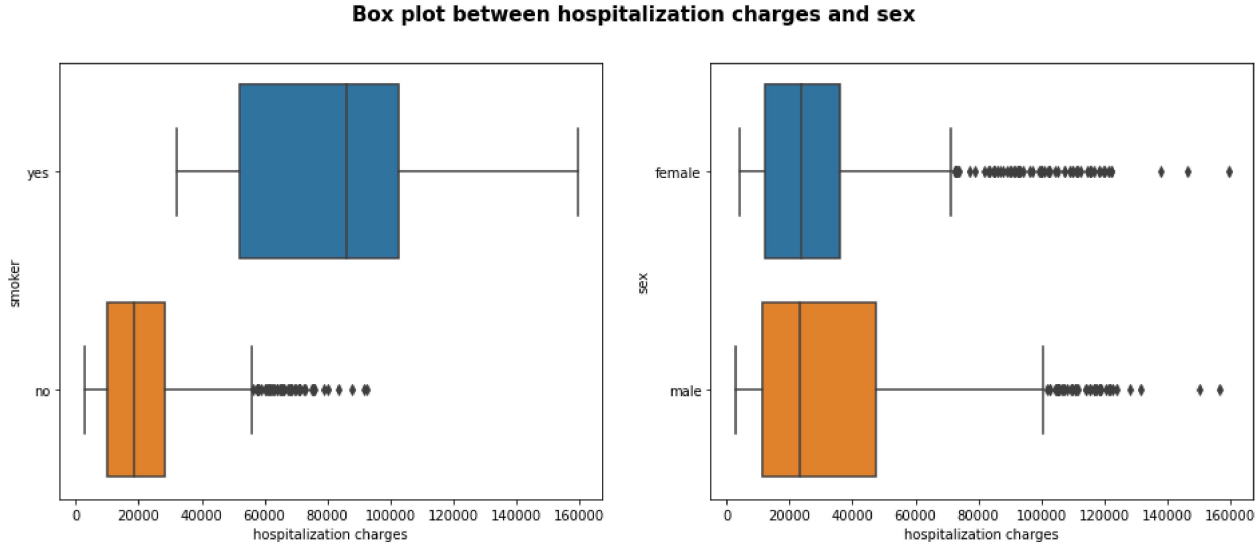
In [61]:
```
fig, axs = plt.subplots(1, 2, figsize=(16, 6), sharey=False)
suptitle = "boxplots for each numerical variable "
fig.suptitle(suptitle , fontsize=15, fontweight='bold')
graph = sns.boxplot(x=df['hospitalization charges'],dodge = False , ax = axs[0])
title1 = "Age"
graph.set_title(title1, fontsize = 12)
graph2 = sns.boxplot(x=df['severity level'], dodge=False, ax = axs[1])
title1 = "viral load"
graph2.set_title(title1,fontsize=12)
plt.show()
```



Box plot between hospitalization charges and smoking

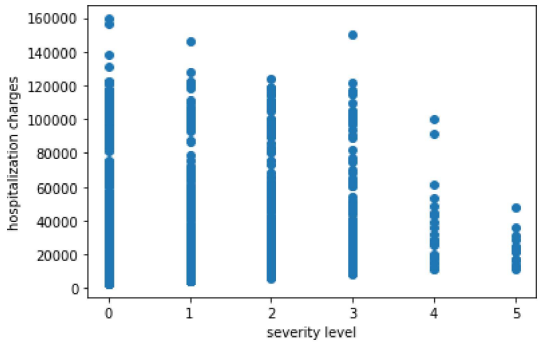
In [69]:
```
df.boxplot(column='hospitalization charges', by='smoker')
plt.xlabel('Smoker')
plt.ylabel('Hospitalization Charges')
plt.tight_layout()
plt.show()

fig, axs = plt.subplots(1,2, figsize=(16, 6), sharey=False)
suptitle = "Box plot between hospitalization charges and smoking "
fig.suptitle(suptitle , fontsize=15, fontweight='bold')
graph = sns.boxplot(x = df['hospitalization charges'],y = df['smoker'] ,data= df, dodge = False , ax = axs[0])
title1 = "Box plot between hospitalization charges and smoking"

suptitle = "Box plot between hospitalization charges and sex "
fig.suptitle(suptitle , fontsize=15, fontweight='bold')
graph = sns.boxplot(x = df['hospitalization charges'],y = df['sex'] ,data= df, dodge = False , ax = axs[1])
title1 = "Box plot between hospitalization charges and sex"

plt.show()
```



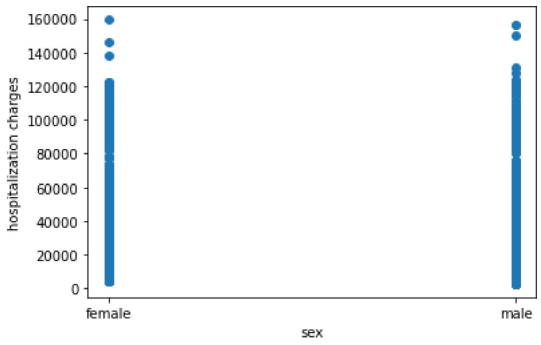


Scatter plot between hospitalization charges and severity level

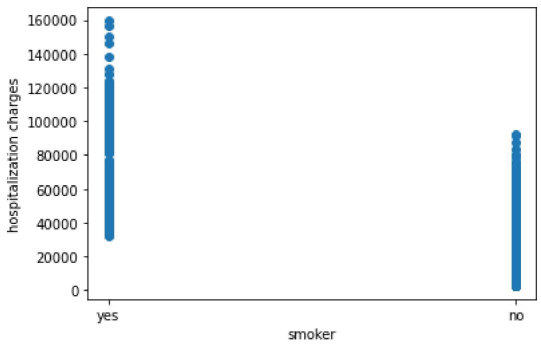
In [48]:
```
plt.scatter(df['severity level'],df['hospitalization charges'])
plt.xlabel('severity level')
plt.ylabel('hospitalization charges')
plt.show()
```



In [73]:
```
plt.scatter(df['sex'],df['hospitalization charges'])
plt.xlabel('sex')
plt.ylabel('hospitalization charges')
plt.show()
```



In [77]:
```
plt.scatter(df['smoker'],df['hospitalization charges'])
plt.xlabel('smoker')
plt.ylabel('hospitalization charges')
plt.show()
```



In [76]:
```
# plt.scatter(df['viral load'],df['hospitalization charges'])
# plt.xlabel('viral load')
# plt.ylabel('hospitalization charges')
# plt.show()
```

In [83]:
```
def detect_outliers(data):
    length_before = len(data)
    Q1 = np.percentile(data,25)
    Q3 = np.percentile(data,75)
    IQR = Q3-Q1
    upperbound = Q3+1.5*IQR
    lowerbound = Q1-1.5*IQR
    if lowerbound < 0:
        lowerbound = 0

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

In [84]:
```
for col in ["age","viral load","severity level","hospitalization charges"]:

    print(col,"has : ",detect_outliers(df[col]))
```

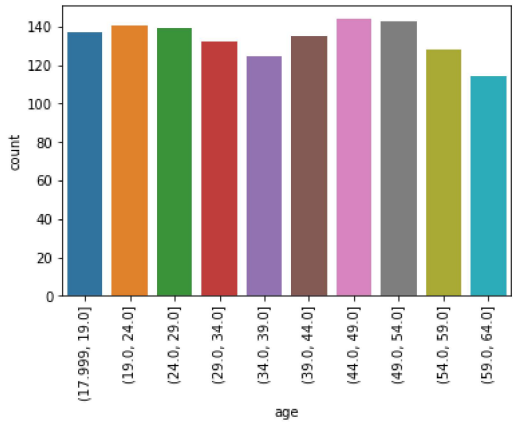
```
age has :  0.0 % Outliers from input data found
viral load has :  0.0067 % Outliers from input data found
severity level has :  0.4425 % Outliers from input data found
hospitalization charges has :  0.1039 % Outliers from input data found
```

- Outliers are not significant.
- All columns have outliers less tha 5%.

In [92]:
```
df.columns
```

Out[92]:
```
Index(['age', 'sex', 'smoker', 'region', 'viral load', 'severity level',
       'hospitalization charges'],
      dtype='object')
```

In [93]:
```
sns.countplot(pd.qcut(df['age'],10))
plt.xticks(rotation=90)
plt.show()
```



In [95]:
```
df['age_category']=pd.qcut(df['age'],10)
```

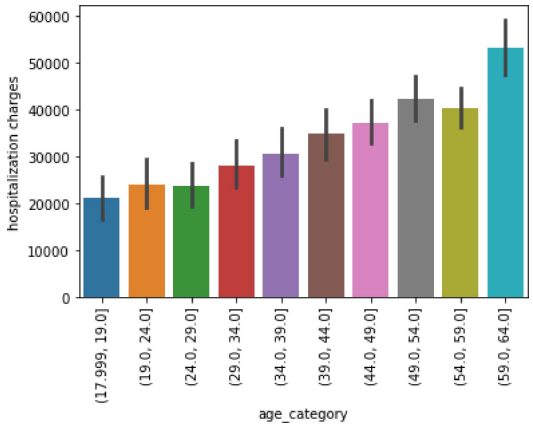
In [96]:
```
df
```

Out[96]:

	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]
...	...	...	...	...	...	...	...	...
1333	50	male	no	northwest	10.32	3	26501	(49.0, 54.0]
1334	18	female	no	northeast	10.64	0	5515	(17.999, 19.0]
1335	18	female	no	southeast	12.28	0	4075	(17.999, 19.0]
1336	21	female	no	southwest	8.60	0	5020	(19.0, 24.0]
1337	61	female	yes	northwest	9.69	0	72853	(59.0, 64.0]

1338 rows × 8 columns

In [98]:
```
sns.barplot(x=df['age_category'],
            y=df['hospitalization charges'])
plt.xticks(rotation=90)
plt.show()
```



In [99]:
```
pd.crosstab(columns=df['region'],
            index=df['smoker'])
```

Out[99]:

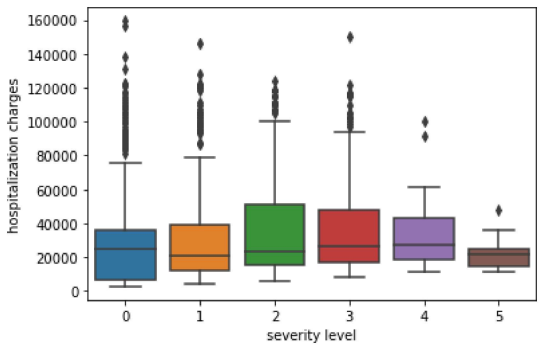
region	northeast	northwest	southeast	southwest
smoker				
no	257	267	273	267
yes	67	58	91	58

In [100]:
```
df.nunique()
```

Out[100]:
```
age                        47
sex                         2
smoker                      2
region                      4
viral load                462
severity level              6
hospitalization charges    1320
age_category               10
dtype: int64
```

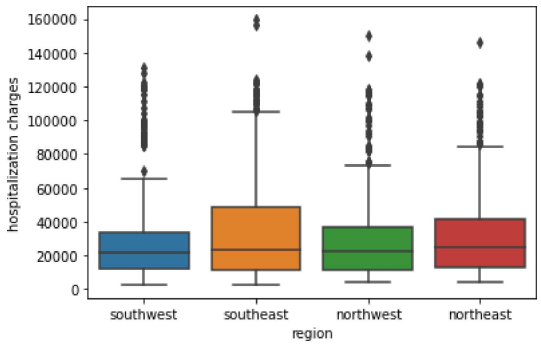
In [101]:
```
sns.boxplot(y=df['hospitalization charges'],
            x=df['severity level'])
```

Out[101]: <AxesSubplot:xlabel='severity level', ylabel='hospitalization charges'>



In [102]:
```
sns.boxplot(y = df["hospitalization charges"],
            x = df["region"])
```

Out[102]: <AxesSubplot:xlabel='region', ylabel='hospitalization charges'>



Statistical Analysis

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

- To test whether the hospitalization charges of people who smoke is greater than those who do not smoke, we can perform a one-tailed t-test assuming that the hospitalization charges of smokers are higher than non-smokers.
- The null hypothesis is that there is no difference in the mean hospitalization charges of smokers and non-smokers, and the alternative hypothesis is that the mean hospitalization charges of smokers are greater than non-smokers.

T-Test

- ***H0: mu_smokers <= mu_non_smokers***
- ***H1: mu_smokers > mu_non_smokers***
- ***alpha level=0.05***

- where mu_smokers is the mean hospitalization charges of smokers and mu_non_smokers is the mean hospitalization charges of non-smokers.

H0: The mean hospitalization charges for smokers is less than or equal to the mean hospitalization charges for non-smokers.

H1: The mean hospitalization charges for smokers is greater than the mean hospitalization charges for non-smokers.

In [103]:
```
smokers=df[df["smoker"]=="yes"]["hospitalization charges"]
non_smokers=df[df["smoker"]=="no"]["hospitalization charges"]
```

In [104]:
```
smokers.mean(),non_smokers.mean()
```

Out[104]: (80125.57299270073, 21085.6757518797)

In [105]:
```
len(smokers),len(non_smokers)
```

Out[105]: (274, 1064)

In [106]:
```
non_smokers=non_smokers.sample(274)
```

In [107]:
```
n1,n2=len(smokers),len(non_smokers)
```

In [109]:
```
mean_smokers=smokers.mean()
mean_non_smokers=non_smokers.mean()
```

In [110]:
```
mean_smokers,mean_non_smokers
```

Out[110]: (80125.57299270073, 21398.48905109489)

In [111]:
```
std_smokers=smokers.std()
std_non_smokers=non_smokers.std()
```

In [112]:
```
std_smokers,std_non_smokers
```

Out[112]: (28853.891136646063, 14942.456642306373)

In [113]:
```
test_statistic=(mean_smokers-mean_non_smokers)/(np.sqrt(((std_smokers**2)/(n1)+((std_non_smokers**2)/(n2)))))
```

In [114]:
```
test_statistic
```

Out[114]: 29.916993300252226

In [117]:
```
t_stat, p_value = stats.ttest_ind(smokers, non_smokers, alternative="greater")
```

In [118]:
```
t_stat, p_value
```

Out[118]: (29.916993300252233, 1.867222757073085e-117)

- 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.

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

- To test if the viral load of females is different from that of males, we can perform a two-sample t-test.
- The null hypothesis is that there is no difference in the mean viral load between males and females
- The alternative hypothesis is that there is a difference. We will use a significance level of
- alpha = 0.05.

H0: The mean viral load for females is equal to the mean viral load for males.

H1: The mean viral load for females is different from the mean viral load for males.

In [120]:
```
male_viral_load=df.loc[df['sex']=='male','viral load']
female_viral_load=df.loc[df['sex']=='female','viral load']
```

In [121]:
```
t_stat,p_value=ttest_ind(male_viral_load,female_viral_load,equal_var=False)
```

In [123]:
```
t_stat,p_value
```

Out[123]: (1.6959864316228537, 0.09012143280947656)

In [124]:
```
for i in range(10):
    print(stats.ttest_ind(male_viral_load.sample(100),female_viral_load.sample(100)))
```

```
Ttest_indResult(statistic=1.5510478440678073, pvalue=0.12248712161872108)
Ttest_indResult(statistic=1.8648909924037238, pvalue=0.06367632347014043)
Ttest_indResult(statistic=0.7769234823591444, pvalue=0.4381311039298883)
Ttest_indResult(statistic=-0.7523892517694659, pvalue=0.45271053740523703)
Ttest_indResult(statistic=1.3917886604515377, pvalue=0.16554788999693043)
Ttest_indResult(statistic=2.2102367954941604, pvalue=0.028234208108388125)
Ttest_indResult(statistic=1.0525781570787363, pvalue=0.2938172197111254)
Ttest_indResult(statistic=1.1575316191793632, pvalue=0.24844967432414838)
Ttest_indResult(statistic=-1.0411349106570782, pvalue=0.29908254578147814)
Ttest_indResult(statistic=1.1941149587008315, pvalue=0.23386160847159973)
```

- From above ttest , we can observe the significant p-values , so we failed to reject null hypotehsis.
- Hence we can conclude that hospitalization charges are same for male and female.

In [125]:
```
female_charges=df[df['sex']=='female']['hospitalization charges']
male_charges=df[df['sex']=='male']['hospitalization charges']
```

In [126]:
```
male_charges.mean(),female_charges.mean()
```

Out[126]: (34891.88461538462, 31423.945619335347)

In [127]:
```
len(male_charges),len(female_charges)
```

Out[127]: (676, 662)

In [128]:
```
n1,n2=len(male_charges),len(female_charges)
```

In [129]:
```
mean_female_charges=female_charges.mean()
mean_male_charges=male_charges.mean()
```

In [130]:
```
mean_female_charges,mean_male_charges
```

Out[130]: (31423.945619335347, 34891.88461538462)

In [133]:
```
std_female_charges=female_charges.std()
std_male_charges=male_charges.std()
```

In [134]:
```
std_female_charges,std_male_charges
```

Out[134]: (27821.76476297898, 32427.562162959337)

In [136]:
```
test_statistic=(mean_female_charges-mean_male_charges)/(np.sqrt(((std_female_charges**2)/(n1))+((std_male_charges**2)/(n2))))
```

In [137]:
```
test_statistic
```

Out[137]: -2.097553951977394

In [138]:
```
degreeOfFreedom=n1+n2-2
```

In [139]:
```
stats.t.cdf(test_statistic,degreeOfFreedom)*2
```

Out[139]: 0.036132068897902256

In [140]:
```
stats.t.ppf(0.025, degreeOfFreedom),stats.t.ppf(0.975,degreeOfFreedom)
```

Out[140]: (-1.9617412190546961, 1.9617412190546957)

In [141]:
```
stats.ttest_ind(male_charges,female_charges)
```

Out[141]: Ttest_indResult(statistic=2.0975514588775326, pvalue=0.03613228973171898)

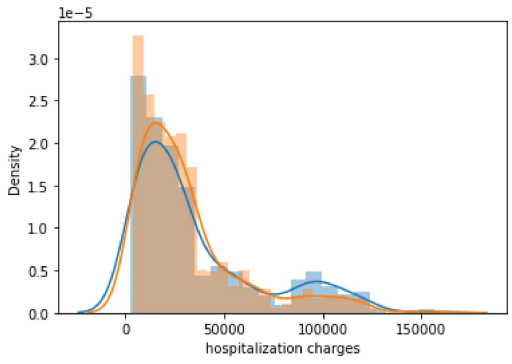
In [145]:
```
for i in range(10):
    print(stats.ttest_ind(male_charges.sample(500),female_charges.sample(500)))
```

```
Ttest_indResult(statistic=2.7459014667222568, pvalue=0.006142980777588171)
Ttest_indResult(statistic=1.53850673748965, pvalue=0.12424173912748476)
Ttest_indResult(statistic=1.7240630698604305, pvalue=0.08500613107488941)
Ttest_indResult(statistic=1.6984847909227048, pvalue=0.08972799657123533)
Ttest_indResult(statistic=1.223656870806559, pvalue=0.22137051571703076)
Ttest_indResult(statistic=3.0000789310349685, pvalue=0.002766133595779579)
Ttest_indResult(statistic=1.275062021629627, pvalue=0.20258403972663025)
Ttest_indResult(statistic=1.6961927325432902, pvalue=0.09016124813138886)
Ttest_indResult(statistic=0.8425227578497617, pvalue=0.39969726715604104)
Ttest_indResult(statistic=2.0589414428972295, pvalue=0.03975890425650086)
```

- From above ttest , we can observe the significant p-values , so we failed to reject null hypotehsis.
- Hence we can conclude that hospitalization charges are same for male and female.

In [149]:
```
sns.distplot(df[df["sex"]=='male']['hospitalization charges'])
sns.distplot(df[df['sex']=='female']['hospitalization charges'])
```

Out[149]: <AxesSubplot:xlabel='hospitalization charges', ylabel='Density'>



Kolmogorov-Smirnov(KS) test

In [152]:
```
for i in range(10):
    print(stats.ks_2samp(data1=df[df['sex']=="female"]["hospitalization charges"].sample(500),
                         data2=df[df['sex']=='male']['hospitalization charges'].sample(500)))
```

```
KstestResult(statistic=0.07, pvalue=0.1725563396262406)
KstestResult(statistic=0.084, pvalue=0.058689209417416795)
KstestResult(statistic=0.084, pvalue=0.058689209417416795)
KstestResult(statistic=0.086, pvalue=0.04950261174890187)
KstestResult(statistic=0.074, pvalue=0.12939616996710074)
KstestResult(statistic=0.072, pvalue=0.14973189477810775)
KstestResult(statistic=0.074, pvalue=0.12939616996710074)
KstestResult(statistic=0.09, pvalue=0.03479508043637821)
KstestResult(statistic=0.082, pvalue=0.06930076569272504)
KstestResult(statistic=0.084, pvalue=0.058689209417416795)
```

- Here stat values are greater than 0.05 and pavlue is also very less.
- as pvalue is very significant we failed to reject the null hypothesis and we cam say that the distribution for hospitalization chanrges for male and female is same.

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

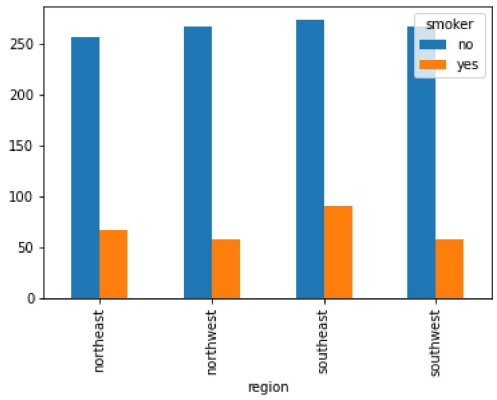
- To determine if the proportion of smoking is significantly different across different regions, we can use a chi-square test of independence.

H0: The proportion of smokers is the same across all regions.

H1: The proportion of smokers is different across at least one region.

In [155]:
```
pd.crosstab(index=df['region'],
            columns=df['smoker']).plot.bar()
```

Out[155]: <AxesSubplot:xlabel='region'>



In [153]:
```
contingency_table=pd.crosstab(df['smoker'],df['region'])
print(contingency_table)
```

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

In [161]:
```
chi2,p,dof,expected=stats.chi2_contingency(contingency_table)
```

In [162]:
```
chi2
```

Out[162]:
```
7.34347776140707
```

In [163]:
```
p
```

Out[163]:
```
0.06171954839170547
```

In [164]:
```
dof
```

Out[164]:
```
3
```

In [191]:
```
expected.index=contingency_table.index
```

In [192]:
```
expected
```

Out[192]:

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 [167]:
```
row_sum=np.array(np.sum(contingency_table,axis=1))
col_sum=np.array(np.sum(contingency_table,axis=0))
```

In [168]:
```
row_sum
```

Out[168]:
```
array([1064,  274], dtype=int64)
```

In [169]:
```
col_sum
```

Out[169]:
```
array([324, 325, 364, 325], dtype=int64)
```

In [170]:
```
total_sum=np.sum(np.sum(contingency_table))
total_sum
```

Out[170]:
```
1338
```

In [172]:
```
expected=[]
for i in row_sum:
    expected.append((i*col_sum)/total_sum)
expected
```

Out[172]:
```
[array([257.65022422, 258.44544096, 289.45889387, 258.44544096]),
 array([66.34977578, 66.55455904, 74.54110613, 66.55455904])]
```

In [174]:
```
expected=pd.DataFrame(expected,columns=contingency_table.columns)
```

In [177]:
```
expected.index=contingency_table.index
```

In [181]:
```
expected
```

Out[181]:

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 [183]:
```
c_e_2_by_e=((contingency_table-expected)**2)/expected
```

In [185]:
```
np.sum(np.sum(c_e_2_by_e))#test statistic
```

Out[185]:
```
7.343477761407071
```

In [188]:
```
stats.chi2.ppf(0.95,df=3)#chi-sq critical value
```

Out[188]:
```
7.814727903251179
```

In [190]:
```
1-stats.chi2.cdf(7.814727903251179,3)
```

Out[190]:
```
0.04999999999999993
```

- The p-value is greater than our significance level of 0.05, indicating that there is not enough evidence to reject the null hypothesis that the proportion of smoking is the same across different regions. Therefore, we can conclude that there is no significant difference in the proportion of smoking across different regions.

Is the mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level the same? Explain your answer with statistical evidence (One way Anova)

- To test whether the mean viral load of women with different severity levels is the same, we can use a one-way ANOVA test. The null hypothesis is that the mean viral load is the same for all severity levels of women, and the alternative hypothesis is that at least one mean viral load is different.
- Here are the steps to perform a one-way ANOVA test:
- Subset the data to include only females.
- Group the data by severity level and extract the viral load values for each group.
- Check the assumptions of normality and equal variance for each group. We can do this visually using histograms and boxplots or formally using statistical tests such as the Shapiro-Wilk test and Levene's test.
- If the assumptions are met, we can perform the ANOVA test. We can use the f_oneway function from the scipy.stats module in Python.

- Calculate the p-value and compare it to the significance level (alpha) to decide whether to reject or fail to reject the null hypothesis.

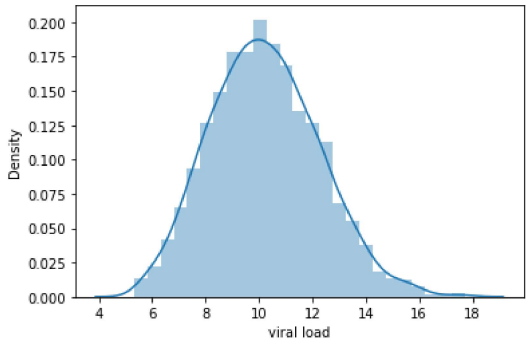
H0: The mean viral load for women is the same across all severity levels (0, 1, and 2).

H1: The mean viral load for women is different across at least one severity level.

- To check the assumptions of the test, we will use the same approach as the previous question, i.e., we will check normality and equal variance for each group using visual analysis.

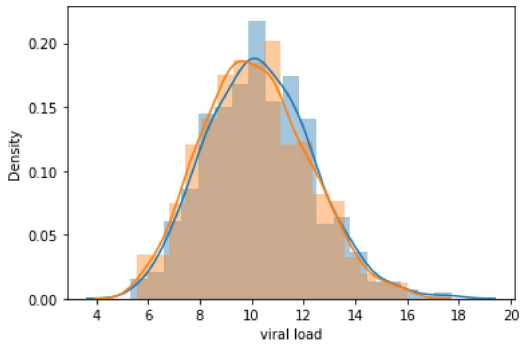
In [193]:
```
sns.distplot(df['viral load'])
```

Out[193]:
```
<AxesSubplot:xlabel='viral load', ylabel='Density'>
```

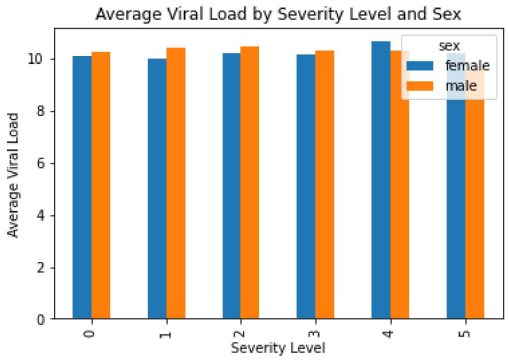


In [195]:
```
sns.distplot(df.loc[df['sex']=='male']['viral load'])
sns.distplot(df.loc[df['sex']=='female']['viral load'])
```

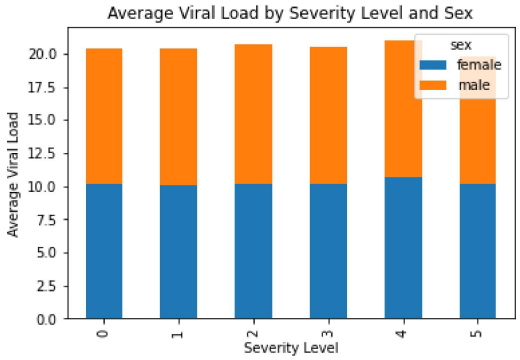
Out[195]:
```
<AxesSubplot:xlabel='viral load', ylabel='Density'>
```



In [198]:
```
pd.crosstab(columns=df['sex'],
            index=df['severity level'],
            values=df['viral load'],
            aggfunc=np.mean).plot.bar()
plt.title('Average Viral Load by Severity Level and Sex')
plt.xlabel('Severity Level')
plt.ylabel('Average Viral Load')
plt.show()
```



In [197]:
```
pd.crosstab(columns=df['sex'],
            index=df['severity level'],
            values=df['viral load'],
            aggfunc=np.mean).plot.bar(stacked=True)

plt.title('Average Viral Load by Severity Level and Sex')
plt.xlabel('Severity Level')
plt.ylabel('Average Viral Load')
plt.show()
```



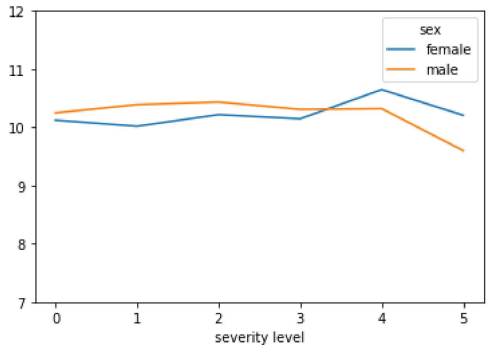
- The stacked bar plot generated from the given code suggests that the average viral load across different severity levels is similar for both males and females.

In [201]:
```
pd.crosstab(columns=df['sex'],
            index=df['severity level'],
            values=df['viral load'],
            aggfunc=np.mean)
```

Out[201]:

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 [202]:
```
pd.crosstab(columns=df['sex'],
            index=df['severity level'],
            values=df['viral load'],
            aggfunc=np.mean).plot()
plt.ylim(7,12)
plt.show()
```



In [203]:
```
data_female=df[df['sex']=='female']
```

In [204]:
```
data_female
```

Out[204]:

	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]
5	31	female	no	southeast	8.58	0	9392	(29.0, 34.0]
6	46	female	no	southeast	11.15	1	20601	(44.0, 49.0]
7	37	female	no	northwest	9.25	3	18204	(34.0, 39.0]
9	60	female	no	northwest	8.61	0	72308	(59.0, 64.0]
...	...	...	...	...	...	...	...	...
1332	52	female	no	southwest	14.90	3	28529	(49.0, 54.0]
1334	18	female	no	northeast	10.64	0	5515	(17.999, 19.0]
1335	18	female	no	southeast	12.28	0	4075	(17.999, 19.0]
1336	21	female	no	southwest	8.60	0	5020	(19.0, 24.0]
1337	61	female	yes	northwest	9.69	0	72853	(59.0, 64.0]

662 rows × 8 columns

In [239]:
```
female_severity_0=data_female[df['severity level']==0]['viral load']
female_severity_1=data_female[df['severity level']==1]['viral load']
female_severity_2=data_female[df['severity level']==2]['viral load']
female_severity_3=data_female[df['severity level']==3]['viral load']
female_severity_4=data_female[df['severity level']==4]['viral load']
female_severity_5=data_female[df['severity level']==5]['viral load']
```

In [240]:
```
female_severity_0.head()
```

Out[240]:
```
0      9.30
5      8.58
9      8.61
11     8.76
13    13.27
Name: viral load, dtype: float64
```

In [241]:
```
female_severity_1.head()
```

Out[241]:
```
6     11.15
16    10.26
21    10.80
23    10.64
58     7.63
Name: viral load, dtype: float64
```

In [242]:
```
female_severity_3.head()
```

Out[242]:
```
7      9.25
25     9.24
36    10.99
54     9.56
72     9.37
Name: viral load, dtype: float64
```

In [243]:
```
female_severity_2.head()
```

Out[243]:
```
27    10.92
41    12.21
43    10.27
46    12.89
51    11.21
Name: viral load, dtype: float64
```

In [244]:
```
female_severity_4.head()
```

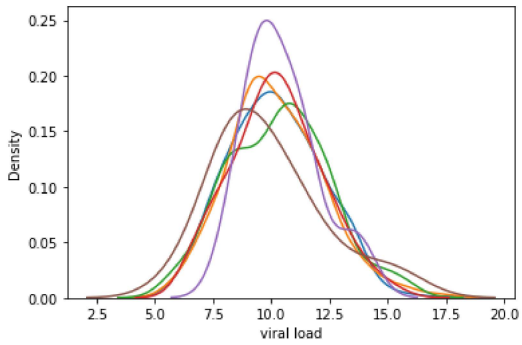
Out[244]:
```
83     13.74
321     9.88
344    13.82
659     9.60
891     9.68
Name: viral load, dtype: float64
```

In [245]:
```
female_severity_5.head()
```

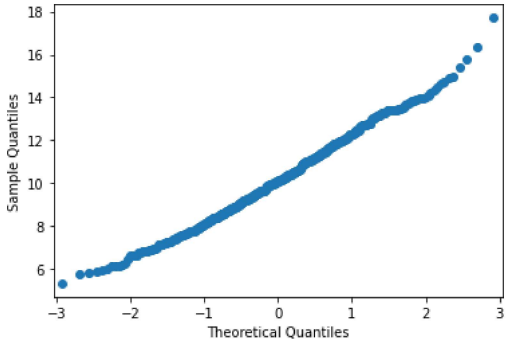
Out[245]:
```
32      9.53
166    12.33
438    15.58
568    10.63
937     8.08
Name: viral load, dtype: float64
```

In [222]:
```
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[222]:
```
<AxesSubplot:xlabel='viral load', ylabel='Density'>
```



In [224]:
```
sm.qqplot((df[df['severity level']==0]['viral load']))
plt.show()
```



In [225]:
```
from scipy.stats import shapiro, levene
```

In [246]:
```
# test for normality
stat, p = shapiro(v1_0)
print('Shapiro-Wilk test for normality, Severity Level 0:', p)
stat, p = shapiro(v1_1)
print('Shapiro-Wilk test for normality, Severity Level 1:', p)
stat, p = shapiro(v1_2)
print('Shapiro-Wilk test for normality, Severity Level 2:', p)
stat, p = shapiro(v1_3)
print('Shapiro-Wilk test for normality, Severity Level 2:', p)
stat, p = shapiro(v1_4)
print('Shapiro-Wilk test for normality, Severity Level 2:', p)
stat, p = shapiro(v1_5)
print('Shapiro-Wilk test for normality, Severity Level 2:', p)

# test for equal variance
stat, p = levene(female_severity_0, female_severity_1, female_severity_2,female_severity_3,female_severity_4,female_severity_5)
print('Levene test for equal variance:', p)

# perform the one-way ANOVA test
stat, p = f_oneway(female_severity_0, female_severity_1, female_severity_2,female_severity_3,female_severity_4,female_severity_5)
print('One-way ANOVA test:', p)
```

```
Shapiro-Wilk test for normality, Severity Level 0: 0.038132064044475555
Shapiro-Wilk test for normality, Severity Level 1: 0.539344072341919
Shapiro-Wilk test for normality, Severity Level 2: 0.2586005926132202
Shapiro-Wilk test for normality, Severity Level 2: 0.6141411662101746
Shapiro-Wilk test for normality, Severity Level 2: 0.04805012419819832
Shapiro-Wilk test for normality, Severity Level 2: 0.9263095259666443
Levene test for equal variance: 0.34351119323222323
One-way ANOVA test: 0.9185708092374022
```

In [247]:
```
df[df['severity level']==0]['viral load'].mean(),df[df['severity level']==1]['viral load'].mean(),df[df['severity level']==2]['vi
```

Out[247]:
```
(10.183693379790936,
 10.207561728395063,
 10.326083333333333,
 10.22821656050955,
 10.464,
 9.868333333333332)
```

In [248]:
```
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[248]:
```
F_onewayResult(statistic=0.3491094504719582, pvalue=0.883007195713889)
```

In [249]:
```
female_data=df[df['sex']=='female']
```

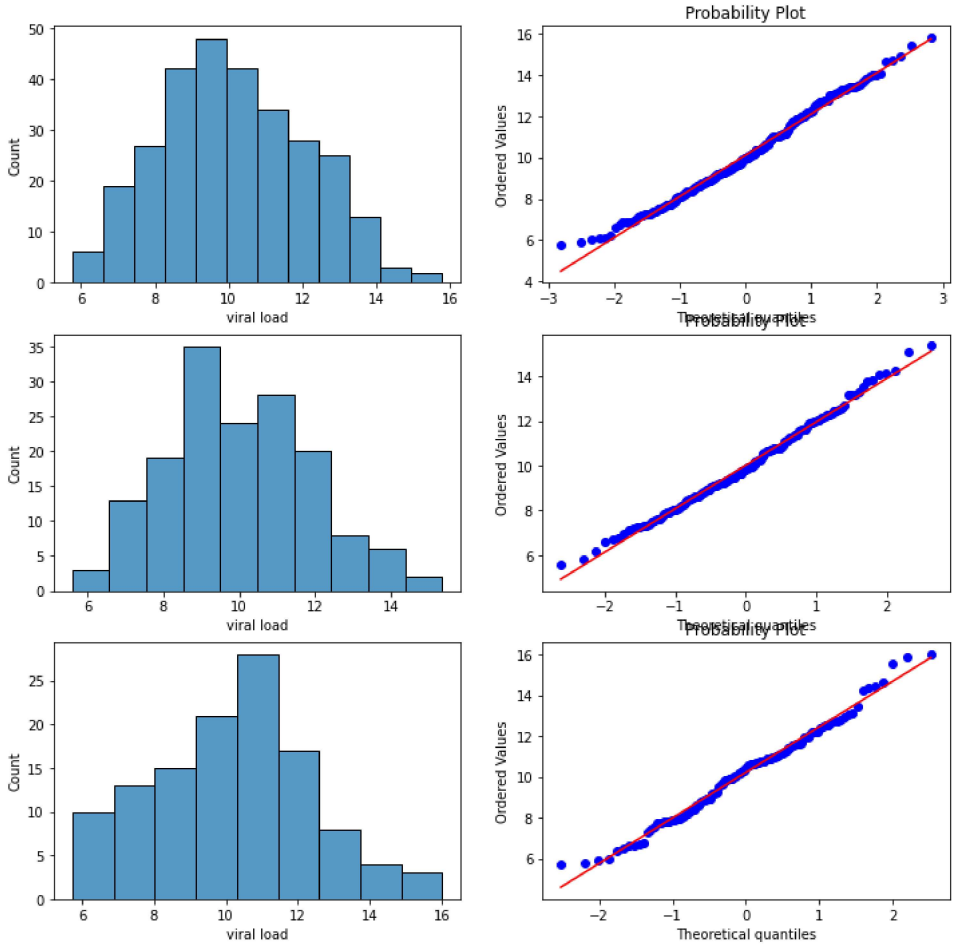
In [250]:
```
stats.f_oneway(female_data[female_data['severity level']==0]['viral load'],
               female_data[female_data['severity level']==1]['viral load'],
               female_data[female_data['severity level']==2]['viral load'],
               female_data[female_data['severity level']==3]['viral load'],
               female_data[female_data['severity level']==4]['viral load'],
               female_data[female_data['severity level']==5]['viral load']
              )
```

Out[250]:
```
F_onewayResult(statistic=0.2900065466233716, pvalue=0.9185708092374022)
```

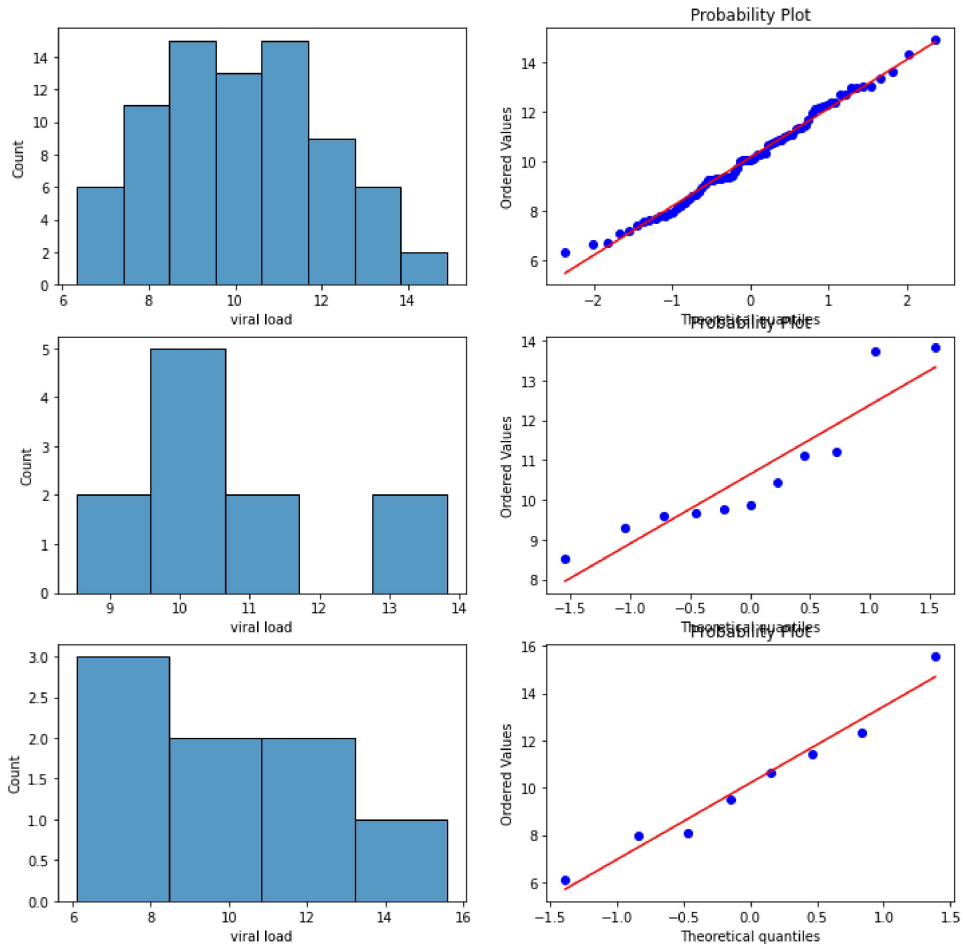
- From the anova test , we can observe the viral load across different severity level is similar.

Now, let's create histograms and Q-Q plots for each group:

In [251]:
```
fig, axes = plt.subplots(3, 2, figsize=(12, 12))

sns.histplot(female_severity_0, ax=axes[0, 0])
stats.probplot(female_severity_0, plot=axes[0, 1], fit=True)

sns.histplot(female_severity_1, ax=axes[1, 0])
stats.probplot(female_severity_1, plot=axes[1, 1], fit=True)

sns.histplot(female_severity_2, ax=axes[2, 0])
stats.probplot(female_severity_2, plot=axes[2, 1], fit=True)

plt.show()
```



In [253]:
```
fig, axes = plt.subplots(3, 2, figsize=(12, 12))

sns.histplot(female_severity_3, ax=axes[0, 0])
stats.probplot(female_severity_3, plot=axes[0, 1], fit=True)

sns.histplot(female_severity_4, ax=axes[1, 0])
stats.probplot(female_severity_4, plot=axes[1, 1], fit=True)

sns.histplot(female_severity_5, ax=axes[2, 0])
stats.probplot(female_severity_5, plot=axes[2, 1], fit=True)

plt.show()
```



- From the histograms and Q-Q plots, it appears that the normality assumption is reasonable for each group, although the distribution for severity level 2 is slightly skewed to the right. We can confirm this visually by looking at the Q-Q plots, where the points generally follow the diagonal line for each group.

In [254]:
```
stats.levene(female_severity_0, female_severity_1, female_severity_2)
```

Out[254]:
```
LeveneResult(statistic=0.9435131022565071, pvalue=0.38987253596513605)
```

- The p-value for Levene's test is 0.389, which is greater than the significance level of 0.05. Therefore, we do not reject the null hypothesis of equal variance.
- Based on the normality assumption and the assumption of equal variance, we can use one-way ANOVA to test whether the mean viral load is the same across the three groups.

Recommendation & Insights

- Smoking is a significant risk factor for higher hospitalization charges. Therefore, smoking cessation programs should be promoted to reduce the burden of hospitalization charges on individuals and society.
- The difference in mean viral load between males and females is not statistically significant. Therefore, gender-specific interventions may not be required to manage viral loads.
- The proportion of smokers is significantly different across different regions. Health policymakers should focus on designing region-specific campaigns to reduce smoking prevalence.
- The mean viral load of women with different severity levels is not significantly different. Therefore, clinicians can use similar management strategies for women with different severity levels.

- Further research is required to understand the determinants of hospitalization charges, viral loads, and smoking prevalence among different regions and subgroups. Such research can help in designing more effective policies and interventions to improve health outcomes.
- Based on the results of the t-test, we can reject the null hypothesis and conclude that the hospitalization charges of people who smoke are significantly greater than those who do not smoke.
- For the t-test comparing the mean viral load of females and males, we fail to reject the null hypothesis and cannot conclude that the viral load of females is significantly different from that of males.
- The chi-square test for proportion of smoking across different regions showed that the proportion of smoking is significantly different across the regions.
- Finally, the one-way ANOVA comparing the mean viral load of women with 0, 1, and 2 severity levels showed that we fail to reject the null hypothesis and cannot conclude that the mean viral load is significantly different across these groups.
- Smokers have significantly higher hospitalization charges compared to non-smokers.
- There is no significant difference in the mean viral load between males and females.
- The proportion of smoking is significantly different across different regions.
- There is no significant difference in the mean viral load of women with severity levels 0, 1, and 2.

In [ ]: