About Apollo Hospitals

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.

Purpose / Problem Statement

The ultimate goal is to tease out meaningful and actionable insights from Patient-level collected data at Apollo 24/7

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

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

In [1]:

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import shapiro
from scipy.stats import levene
from scipy.stats import mannwhitneyu
from scipy.stats import chi2_contingency
from scipy.stats import f_oneway
```

In [2]:

```
df = pd.read_csv('scaler_apollo_hospitals.csv')
df.head()
```

Out[2]:

	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 [3]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 8 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Unnamed: 0               1338 non-null   int64
 1   age                      1338 non-null   int64
 2   sex                      1338 non-null   object
 3   smoker                   1338 non-null   object
 4   region                   1338 non-null   object
 5   viral load               1338 non-null   float64
 6   severity level           1338 non-null   int64
 7   hospitalization charges  1338 non-null   int64
dtypes: float64(1), int64(4), object(3)
memory usage: 83.8+ KB
```

In [4]:

```
# There are no Missing Values
df.isnull().sum(axis = 0)
```

Out[4]:

```
Unnamed: 0                 0
age                        0
sex                        0
smoker                     0
region                     0
viral load                 0
```

```
severity level            0
hospitalization charges    0
dtype: int64
```

Statistical Summary of Data

In [5]:

```
# Numerical Data
df.describe()
```

Out[5]:

	Unnamed: 0	age	viral load	severity level	hospitalization charges
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	668.500000	39.207025	10.221233	1.094918	33176.058296
std	386.391641	14.049960	2.032796	1.205493	30275.029296
min	0.000000	18.000000	5.320000	0.000000	2805.000000
25%	334.250000	27.000000	8.762500	0.000000	11851.000000
50%	668.500000	39.000000	10.130000	1.000000	23455.000000
75%	1002.750000	51.000000	11.567500	2.000000	41599.500000
max	1337.000000	64.000000	17.710000	5.000000	159426.000000

In [6]:

```
# Categorical Data
df.describe(include = 'object')
```

Out[6]:

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

Univariate Analysis

In [81]:

```
def numerical_data_univariate(c):
    x = df[c]
    sns.kdeplot(x)
    plt.title(f'Distribution of {c.title()}')
    plt.show()
    print('****** Insights ******')
    m = df[c].mean()
    me = df[c].median()
    ma = df[c].max()
    mi = df[c].min()
    print(f'Mean {c} = {m}\nMedian {c} = {me}\nMaximum {c} = {ma}\nMinimum {c} = {mi}')
```

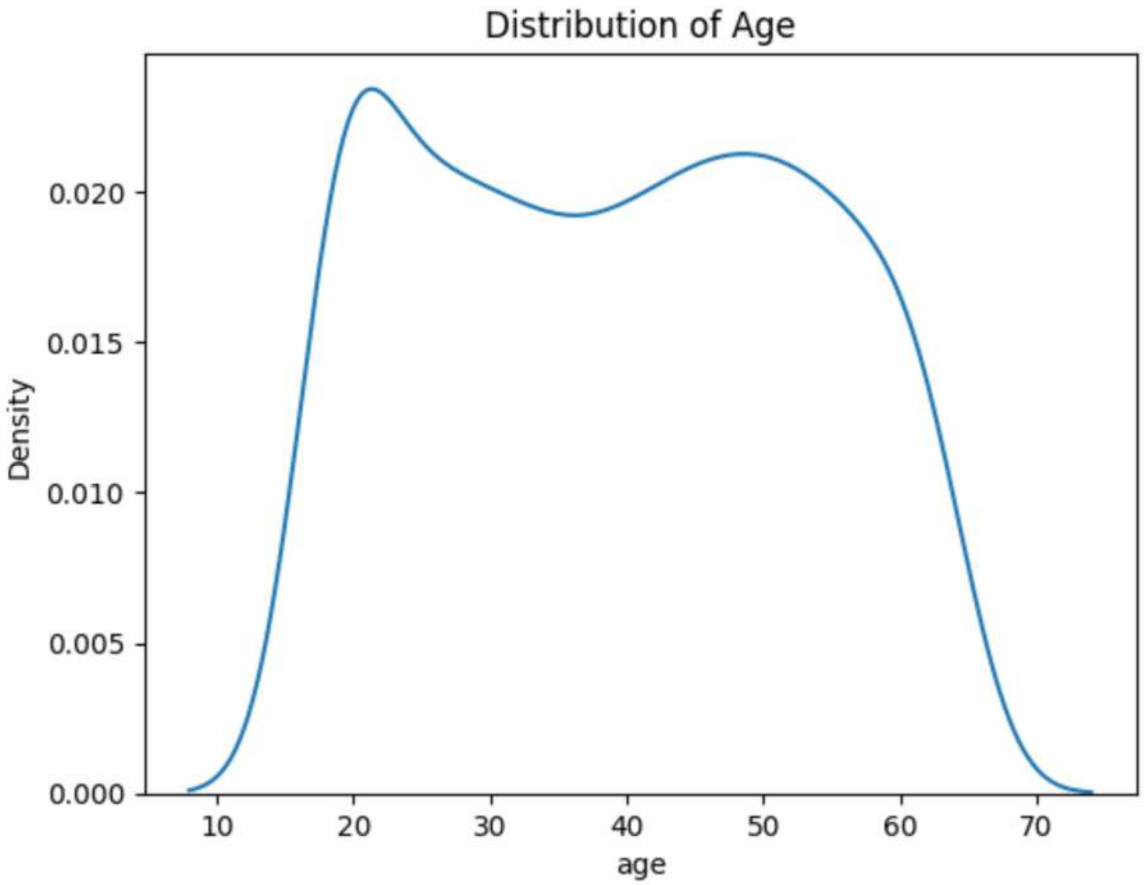
In [82]:

```
def categorical_data_univariate(c):
    sns.countplot(data = df, x = c)
    plt.title('Count of '+c.title())
    plt.show()
    print('****** Insights ******')
    zipped = zip(df[c].value_counts().index, df[c].value_counts().values)
    for i in zipped:
        print(f'Count of {i[0].title()} = {i[1]}')
```

Age

In [83]:

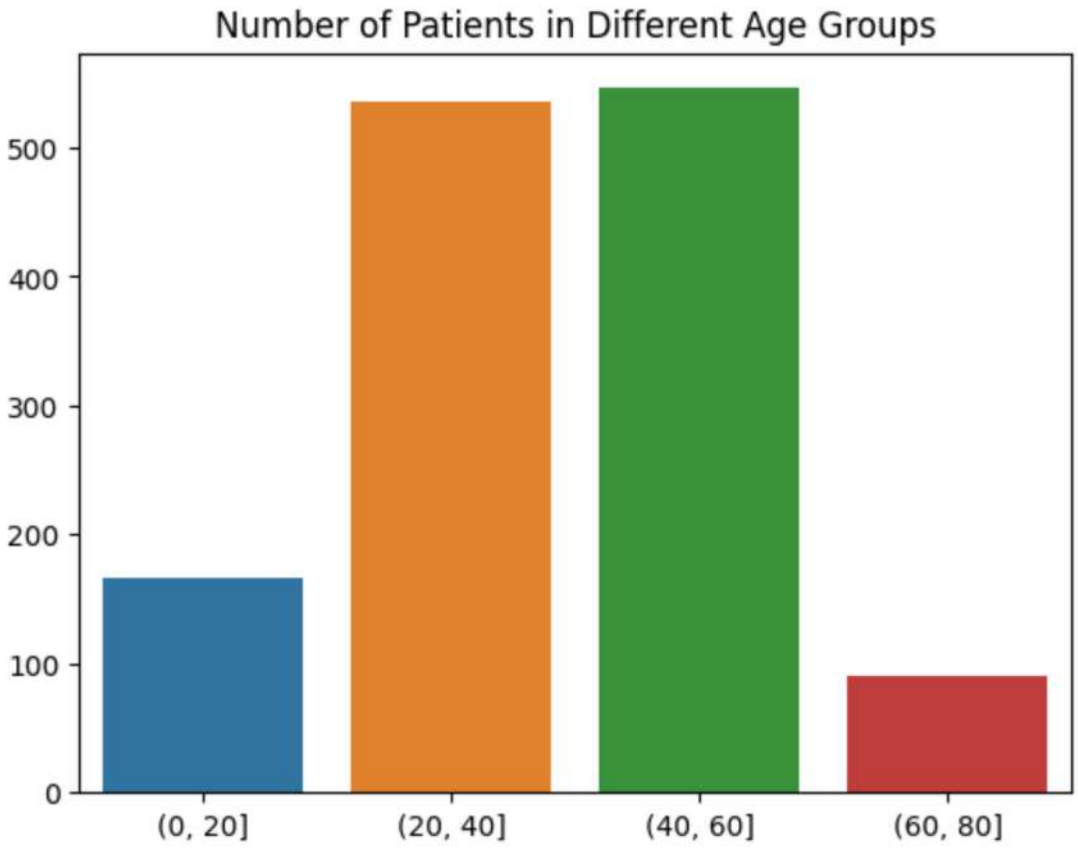
```
numerical_data_univariate('age')
```



```
****** Insights ******
Mean age = 39.20702541106129
Median age = 39.0
Maximum age = 64
Minimum age = 18
```

In [145…

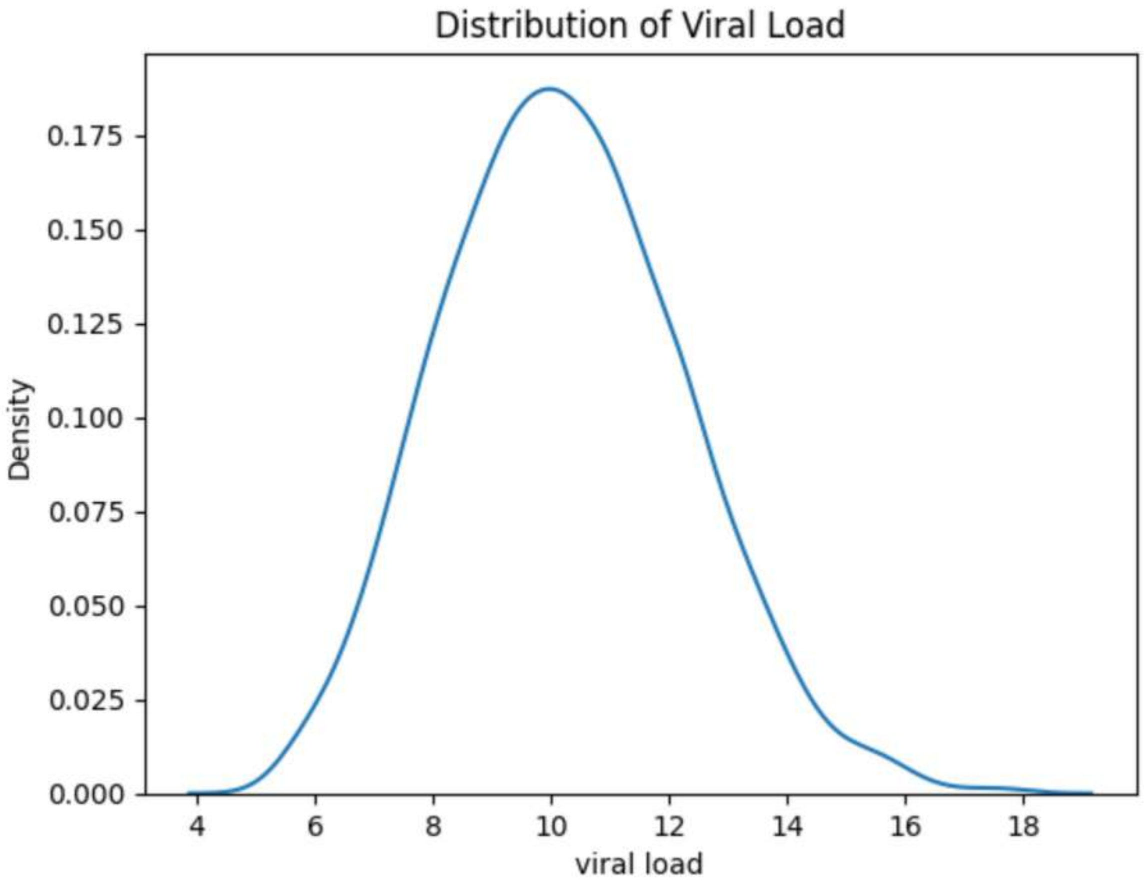
```
df['age_cat'] = pd.cut(bins = [0, 20, 40, 60, 80], x = df['age'])
x = df.groupby('age_cat').count()['age'].index.values
y = df.groupby('age_cat').count()['age'].values
sns.barplot(x = x, y = y)
plt.title('Number of Patients in Different Age Groups')
plt.show()
```



Viral Load

In [84]:

```
numerical_data_univariate('viral load')
```

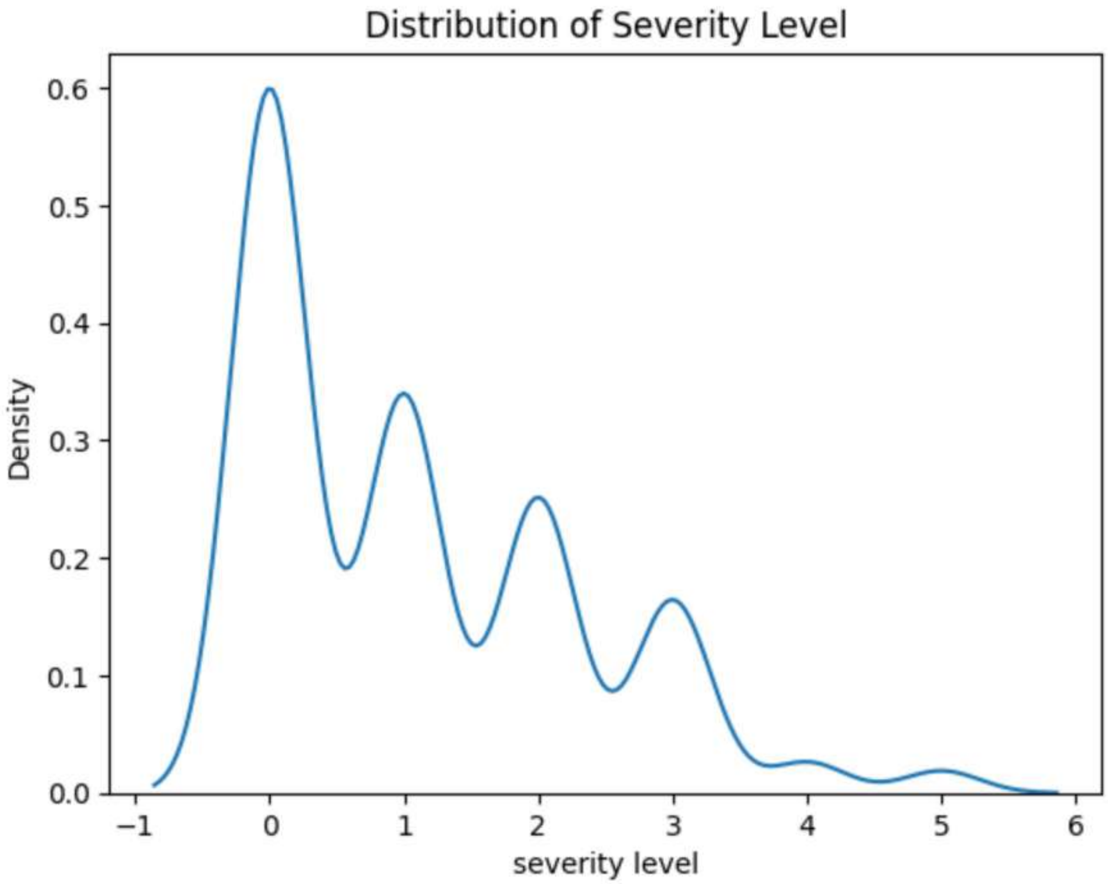


```
****** Insights ******
Mean viral load = 10.221233183856526
Median viral load = 10.13
Maximum viral load = 17.71
Minimum viral load = 5.32
```

Severity Level

In [85]:

```
numerical_data_univariate('severity level')
```

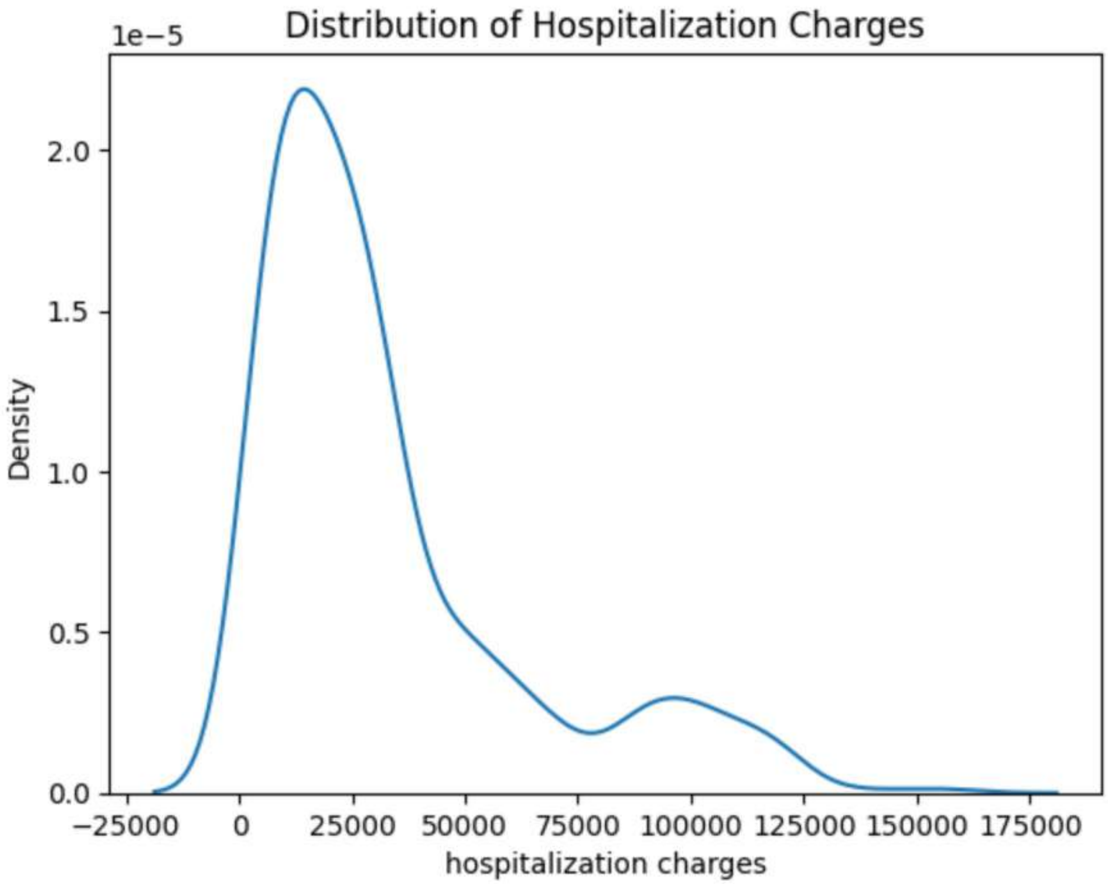


```
****** Insights ******
Mean severity level = 1.0949177877429
Median severity level = 1.0
Maximum severity level = 5
Minimum severity level = 0
```

Hospitalization Charges

In [86]:

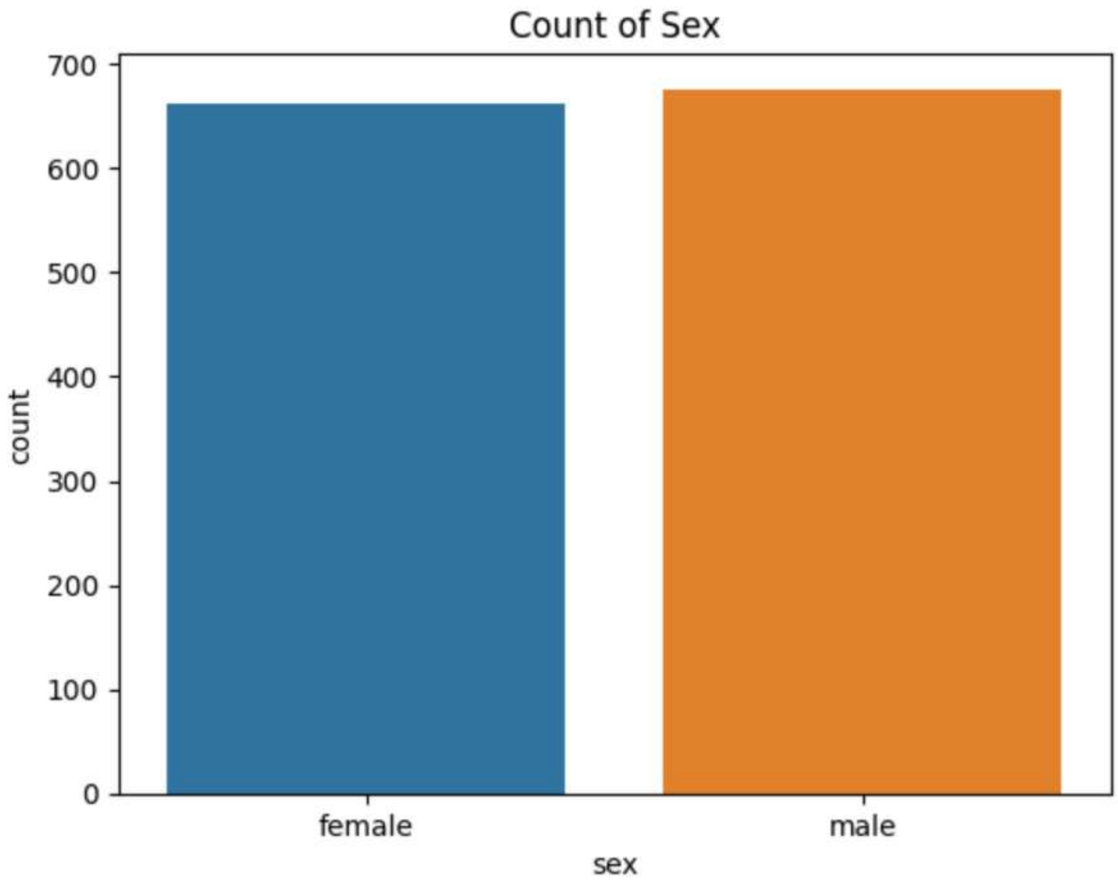
```
numerical_data_univariate('hospitalization charges')
```



```
****** Insights ******
Mean hospitalization charges = 33176.058295964125
Median hospitalization charges = 23455.0
Maximum hospitalization charges = 159426
Minimum hospitalization charges = 2805
```

In [87]:

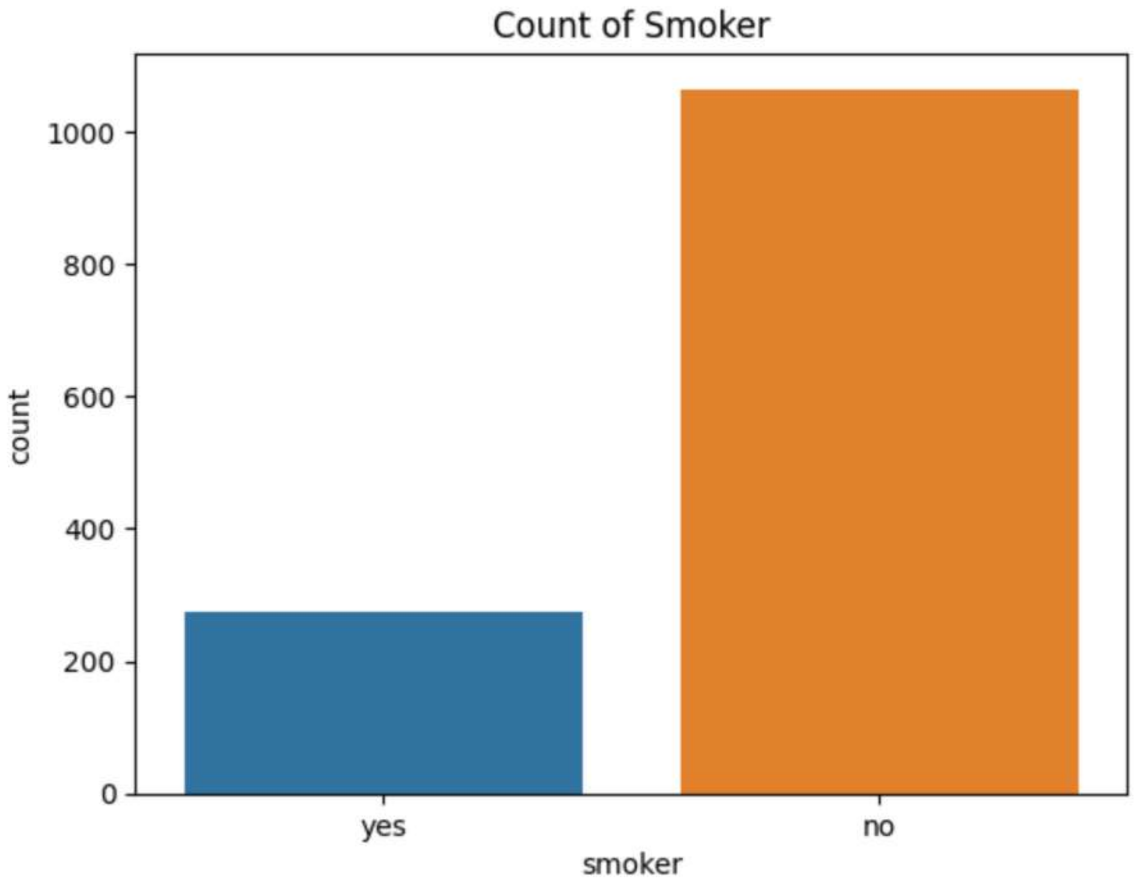
```
categorical_data_univariate('sex')
```



```
****** Insights ******
Count of Male = 676
Count of Female = 662
```

In [88]:

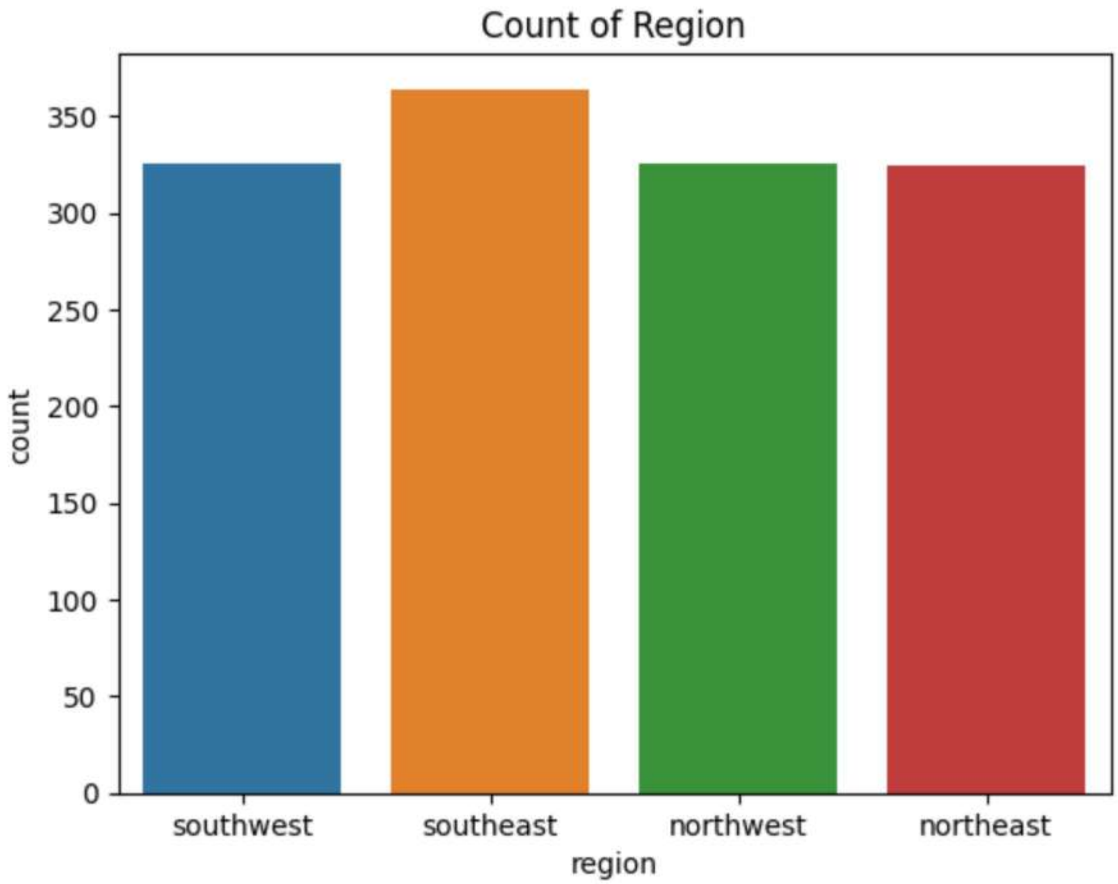
```
categorical_data_univariate('smoker')
```



```
****** Insights ******
Count of No = 1064
Count of Yes = 274
```

In [89]:

```
categorical_data_univariate('region')
```

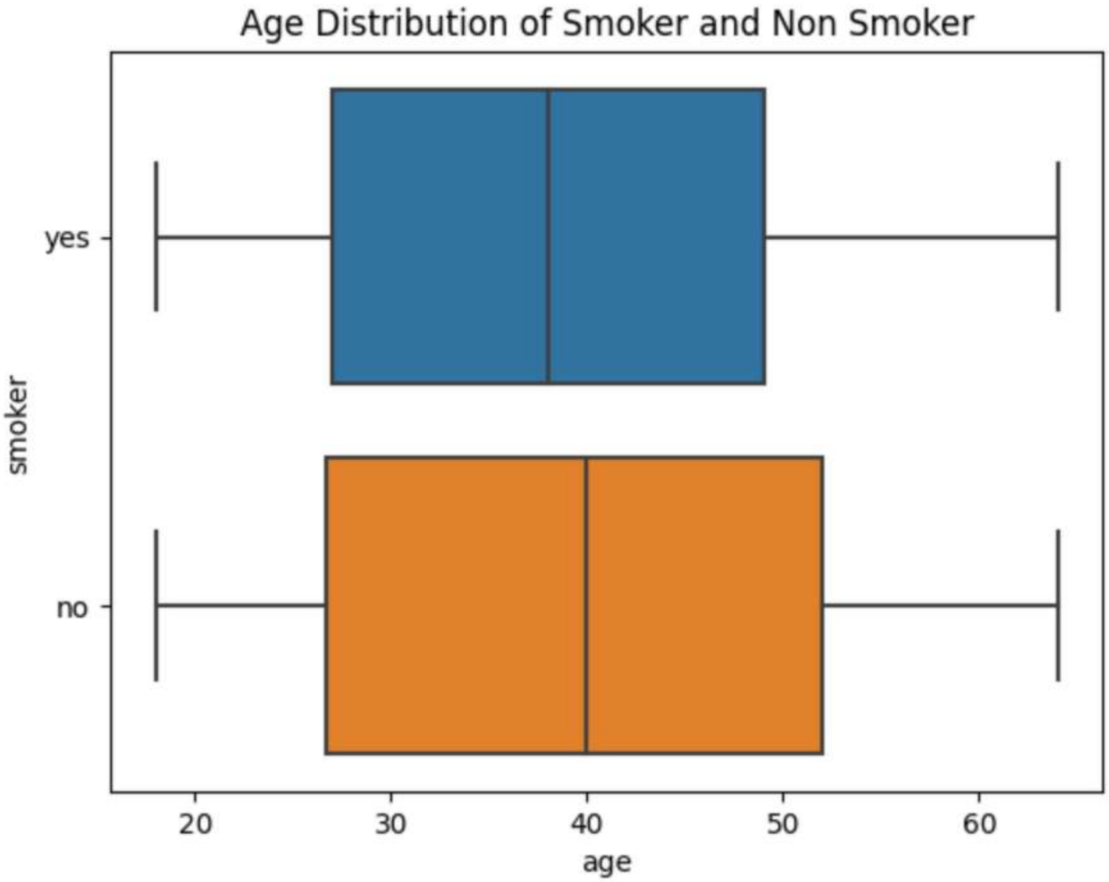


```
****** Insights ******
Count of Southeast = 364
Count of Southwest = 325
Count of Northwest = 325
Count of Northeast = 324
```

Bivariate Analysis

In [92]:

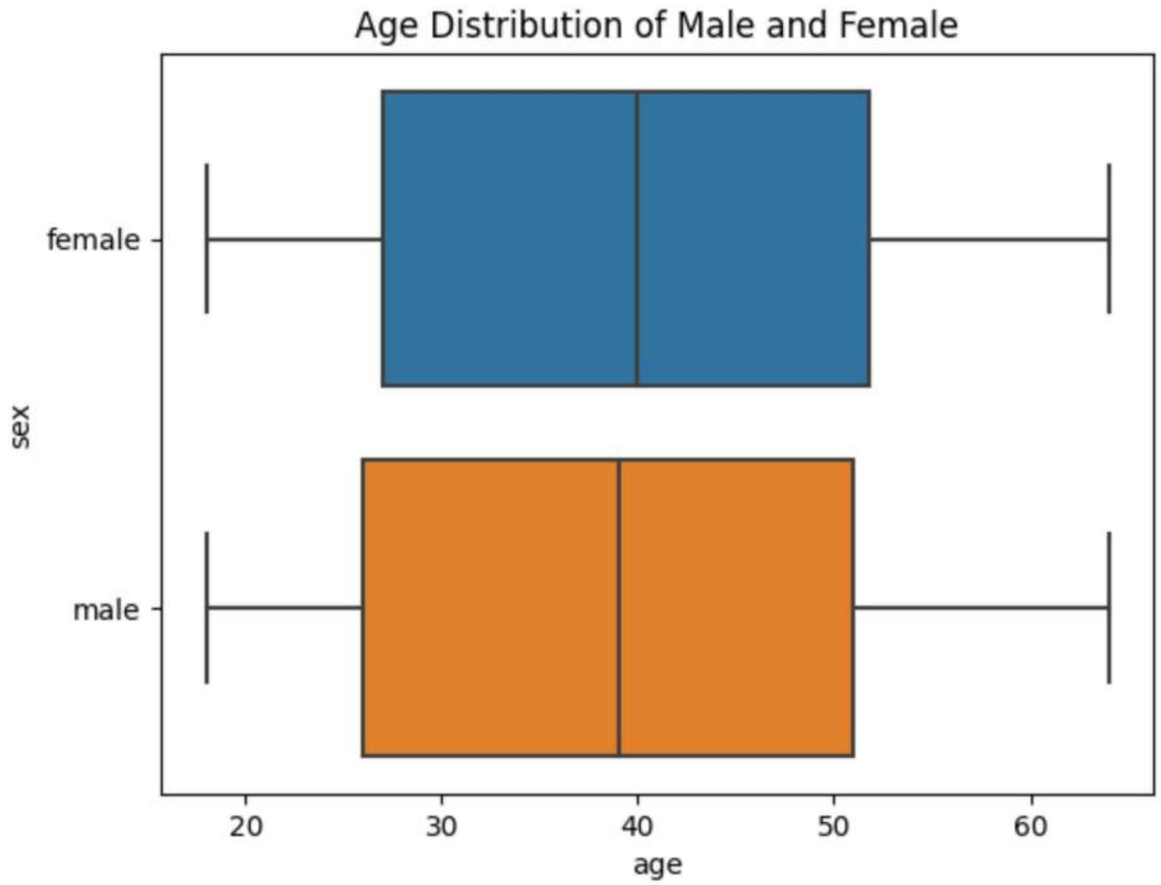
```
sns.boxplot(data = df, x = 'age', y = 'smoker')
plt.title('Age Distribution of Smoker and Non Smoker')
plt.show()
```



Insights - The People who do not smoke have slightly Larger distribution of Age

In [74]:

```
sns.boxplot(data = df, x = 'age', y = 'sex')
plt.title('Age Distribution of Male and Female')
plt.show()
```



In [111…

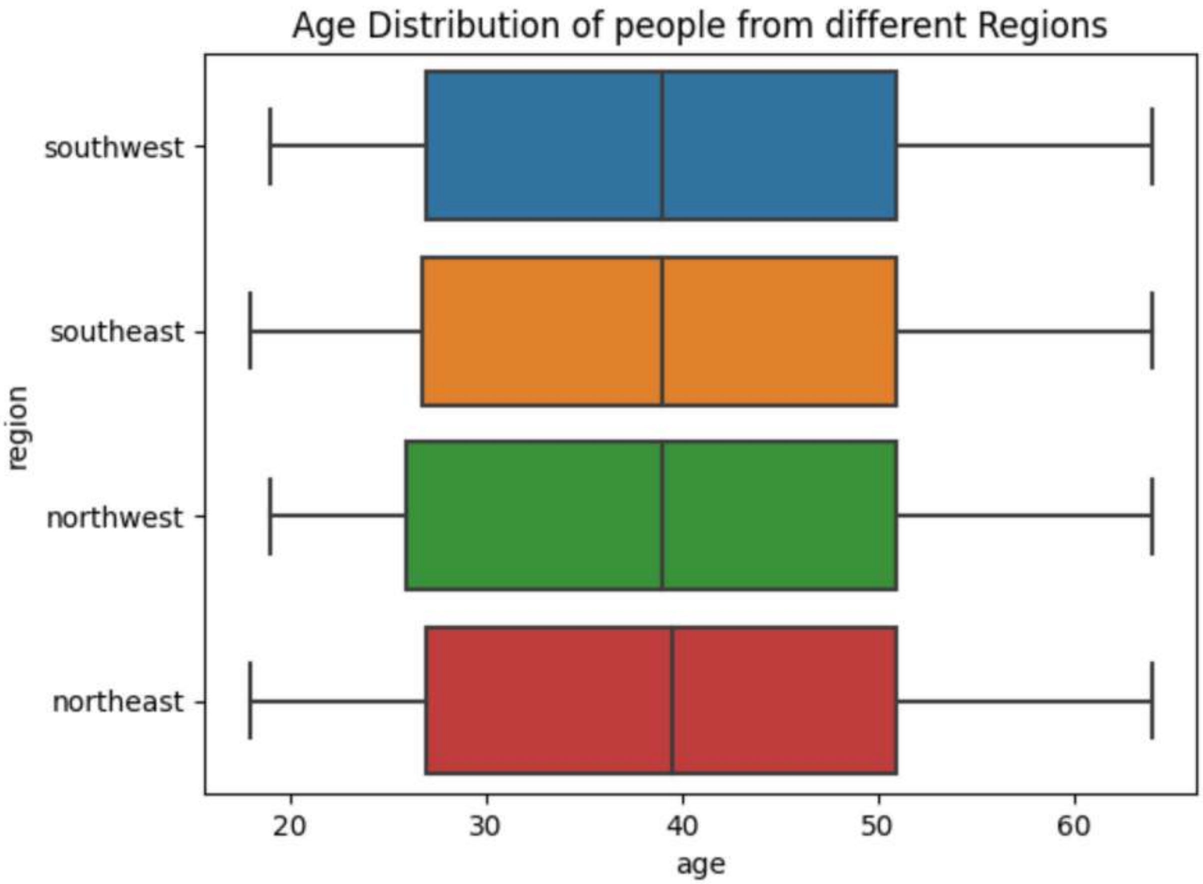
```
m = df[df['sex']=='male']['age'].median()
f = df[df['sex']=='female']['age'].median()
print(f'Median Age of Males = {m} and Median age of Females = {f}')
```

```
Median Age of Males = 39.0 and Median age of Females = 40.0
```

Insight - Median age of Female (40 years) is larger then Male (39 years)

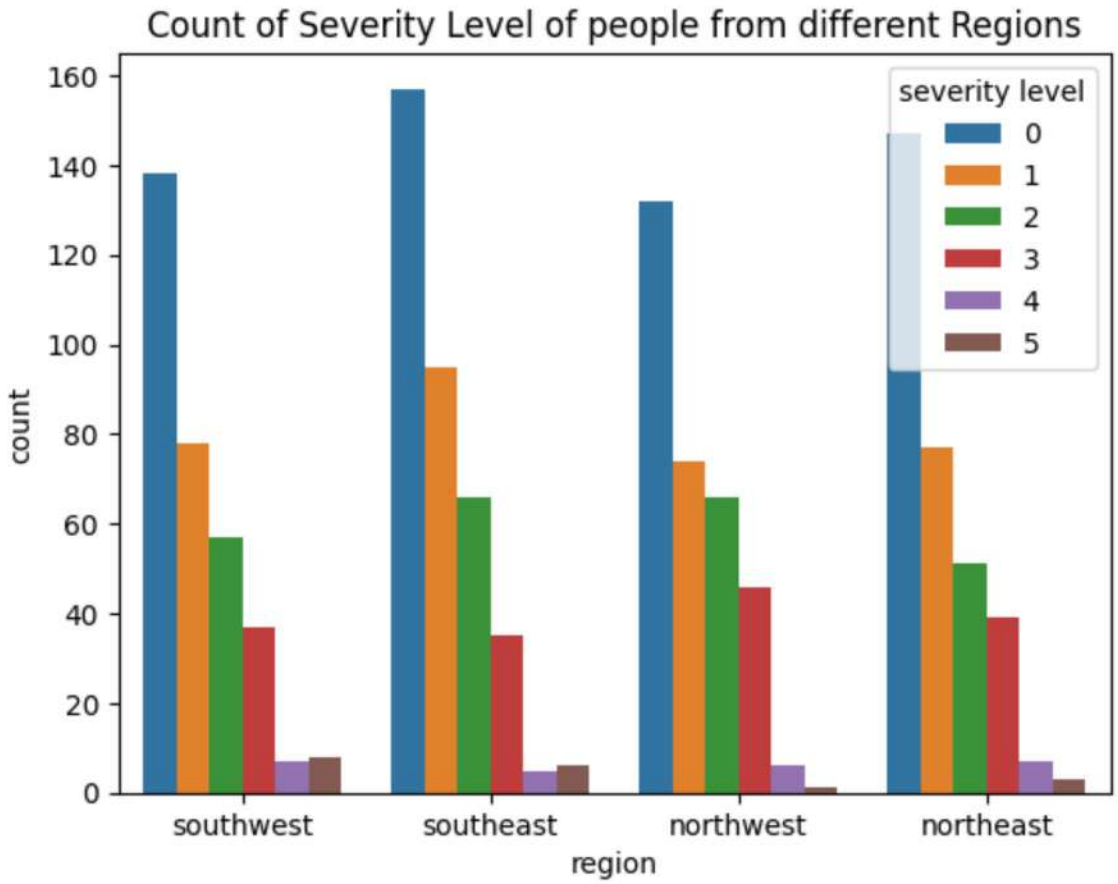
In [125…

```
sns.boxplot(data = df, x = 'age', y = 'region')
plt.title('Age Distribution of people from different Regions')
plt.show()
```



In [96]:

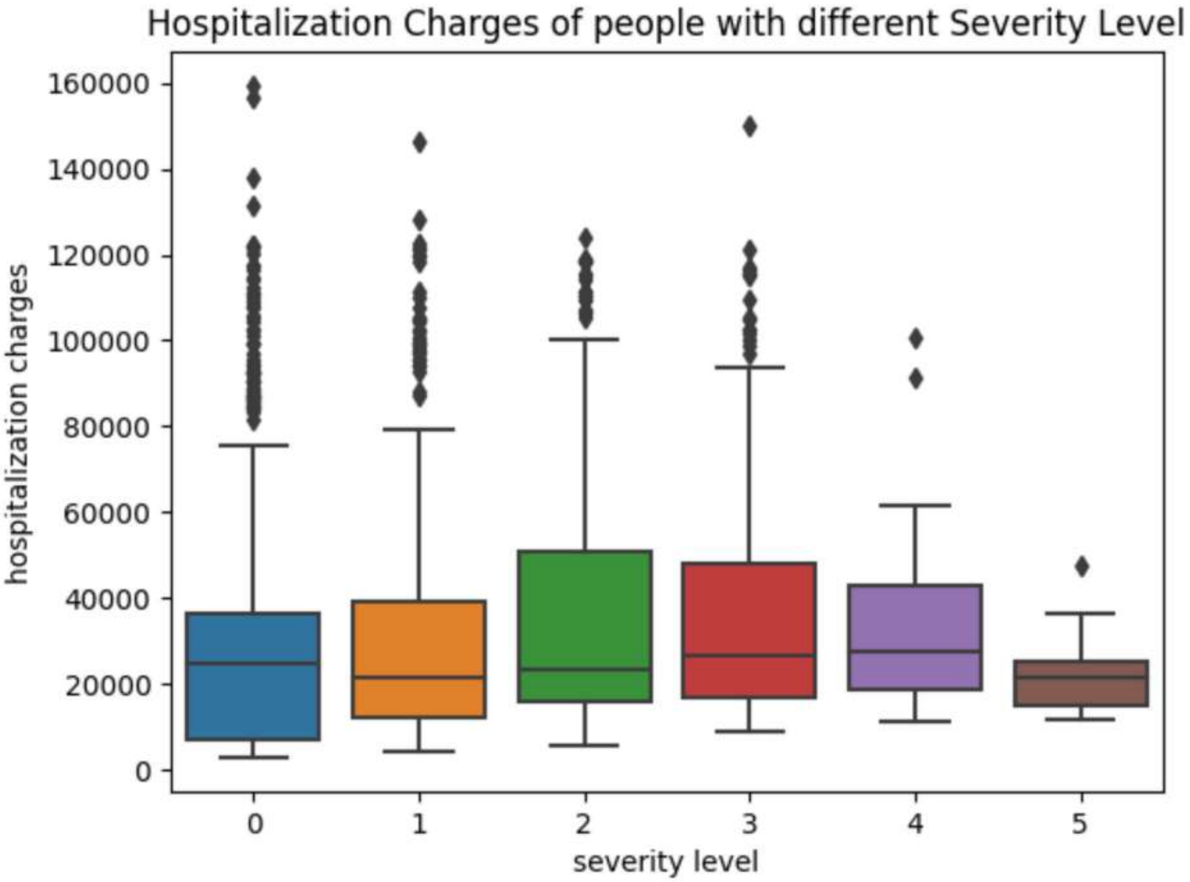
```
sns.countplot(data = df, x = 'region', hue = 'severity level')
plt.title('Count of Severity Level of people from different Regions')
plt.show()
```



Insight - From all the regions, most of the people have zero Severity level, followed by 1, 2 and 3

In [70]:

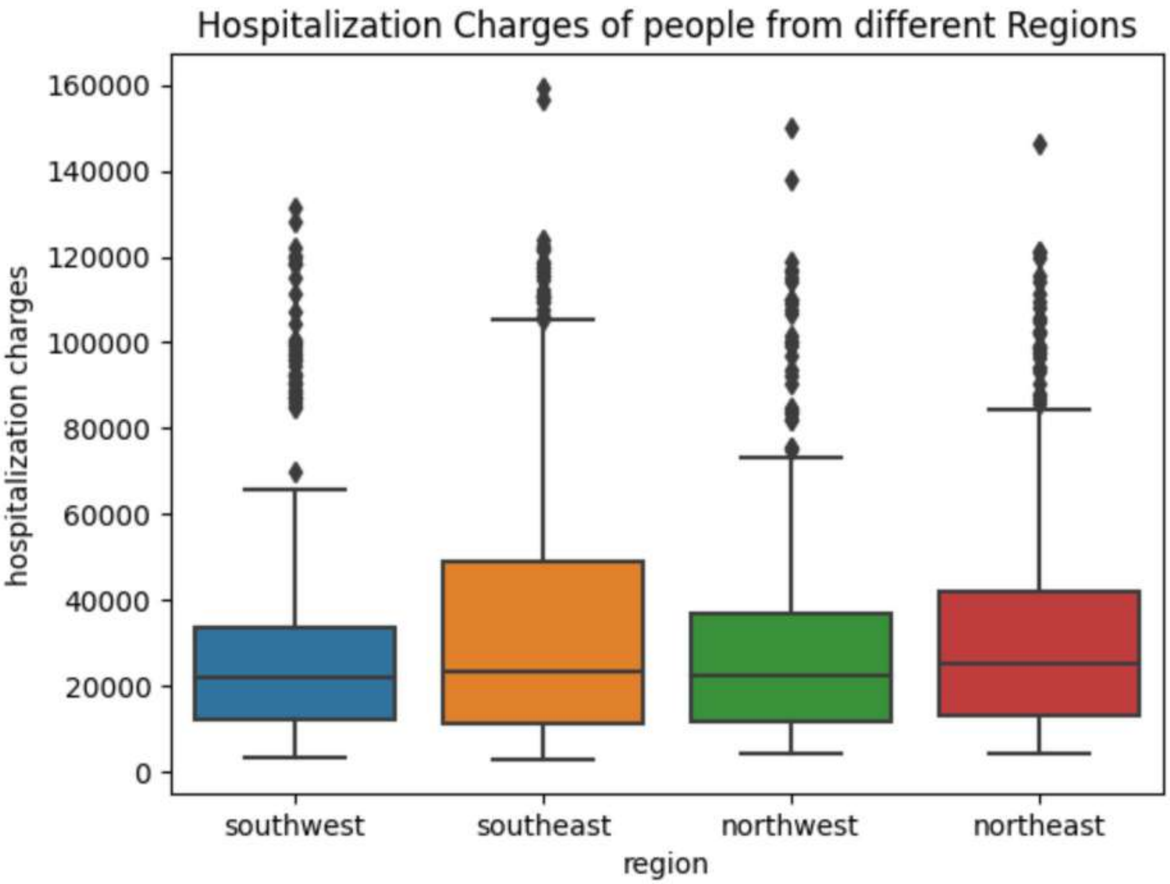
```
sns.boxplot(data = df, x = 'severity level', y = 'hospitalization charges')
plt.title('Hospitalization Charges of people with different Severity Level')
plt.show()
```



Insight - If we consider the minimum Hospitalization charges, Severity Level 5 has the Maximum. If we consider the maximum Charges, its maximum for Severity level 2

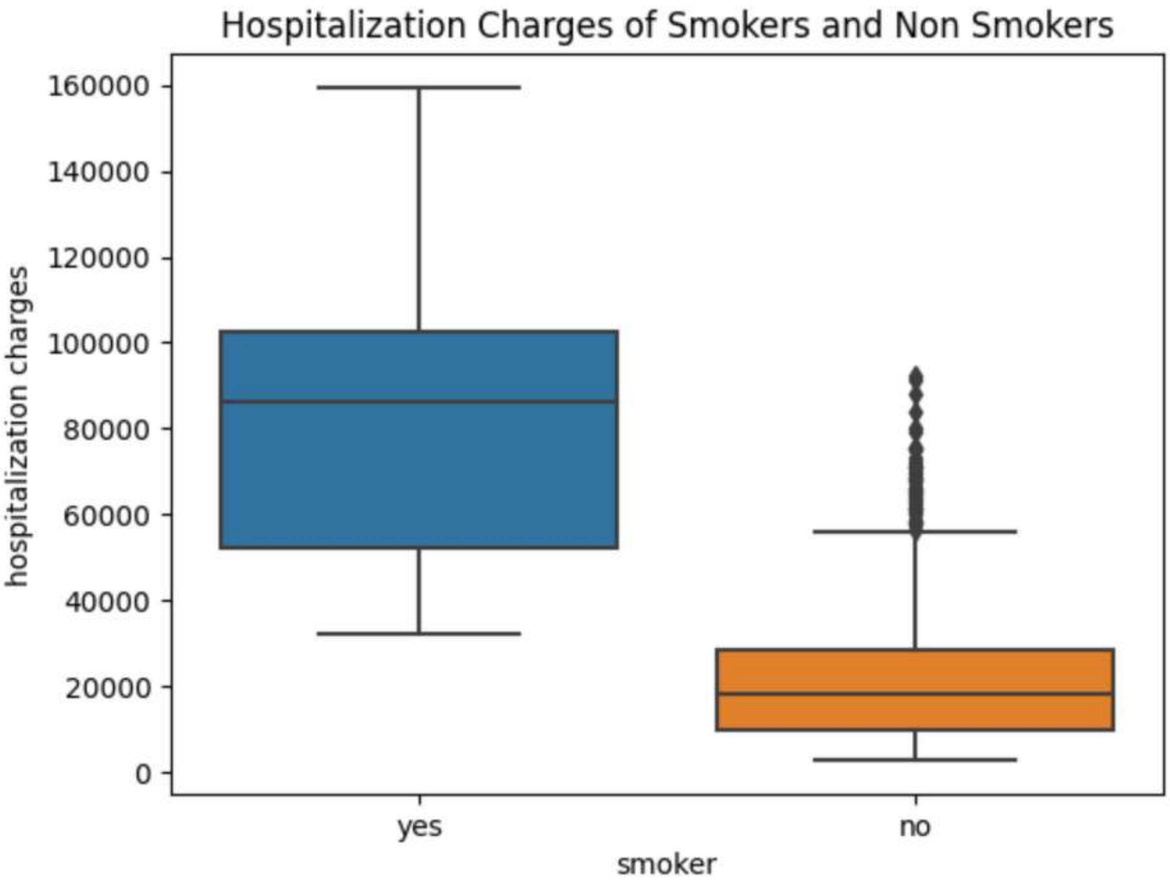
In [115…

```
sns.boxplot(data = df, x = 'region', y = 'hospitalization charges')
plt.title('Hospitalization Charges of people from different Regions')
plt.show()
```



In [76]:

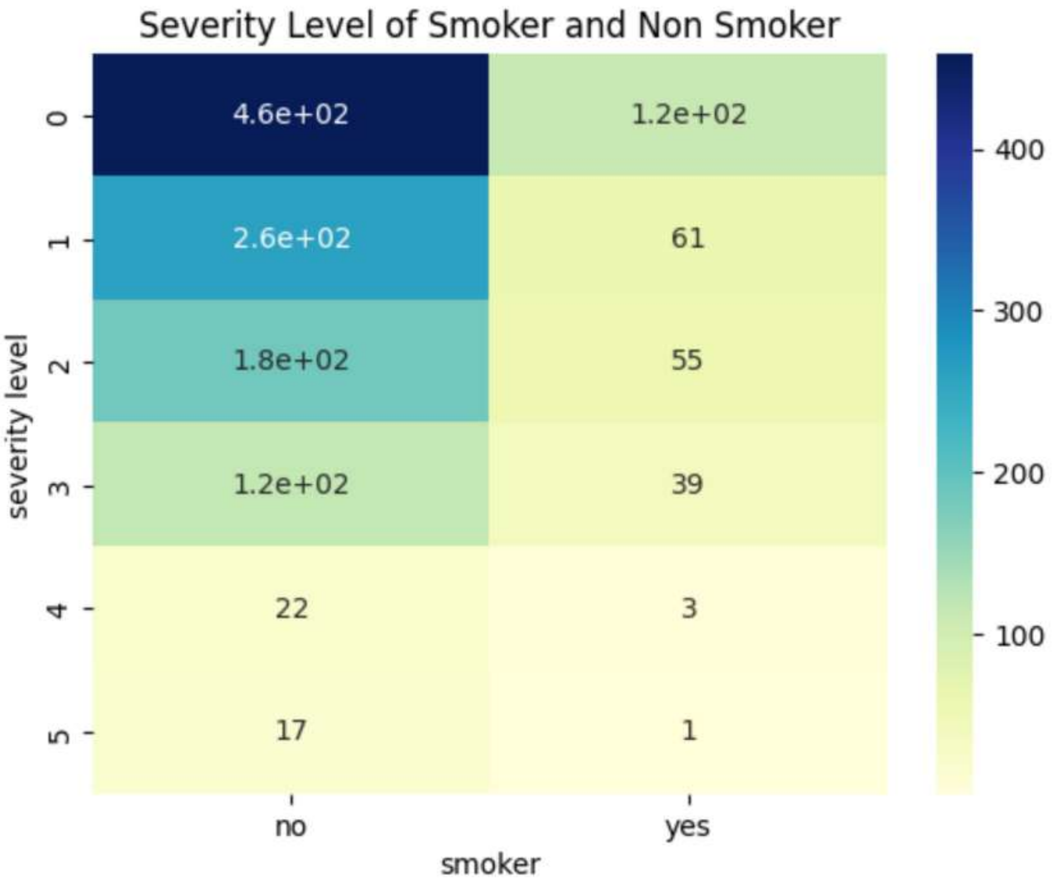
```
sns.boxplot(data = df, x = 'smoker', y = 'hospitalization charges')
plt.title('Hospitalization Charges of Smokers and Non Smokers')
plt.show()
```



Insights - There seems to be a very significant difference between the Hospitalization charges of Smokers and Non Smokers

In [77]:

```
sns.heatmap(pd.crosstab(df['severity level'], df['smoker']), cmap = "YlGnBu", annot=True)
plt.title('Severity Level of Smoker and Non Smoker')
plt.show()
```



In [78]:

```
def check_normality(data):

    if shapiro(data)[1]<0.05:
        print(f'The Dataset is not Normal based on Shapiro-Wilk test with P Value = {shapiro(data)[1]} and Statistic =
    else:
        print(f'The Dataset is Normal based on Shapiro-Wilk test with P Value = {shapiro(data)[1]} and Statistic = {sh
```

In [79]:

```
def check_variance(data1, data2):

    if levene(data1, data2)[1]<0.05:
        print(f'The difference in variance of Dataset is significant based on Levene Test with P Value = {levene(data1
    else:
        print(f'The difference in variance of Dataset is not significant based on Levene Test with P Value = {levene(d
```

Hypothesis Testing

Is hospitalization Charges of people who do smoking is greater than those who don't?

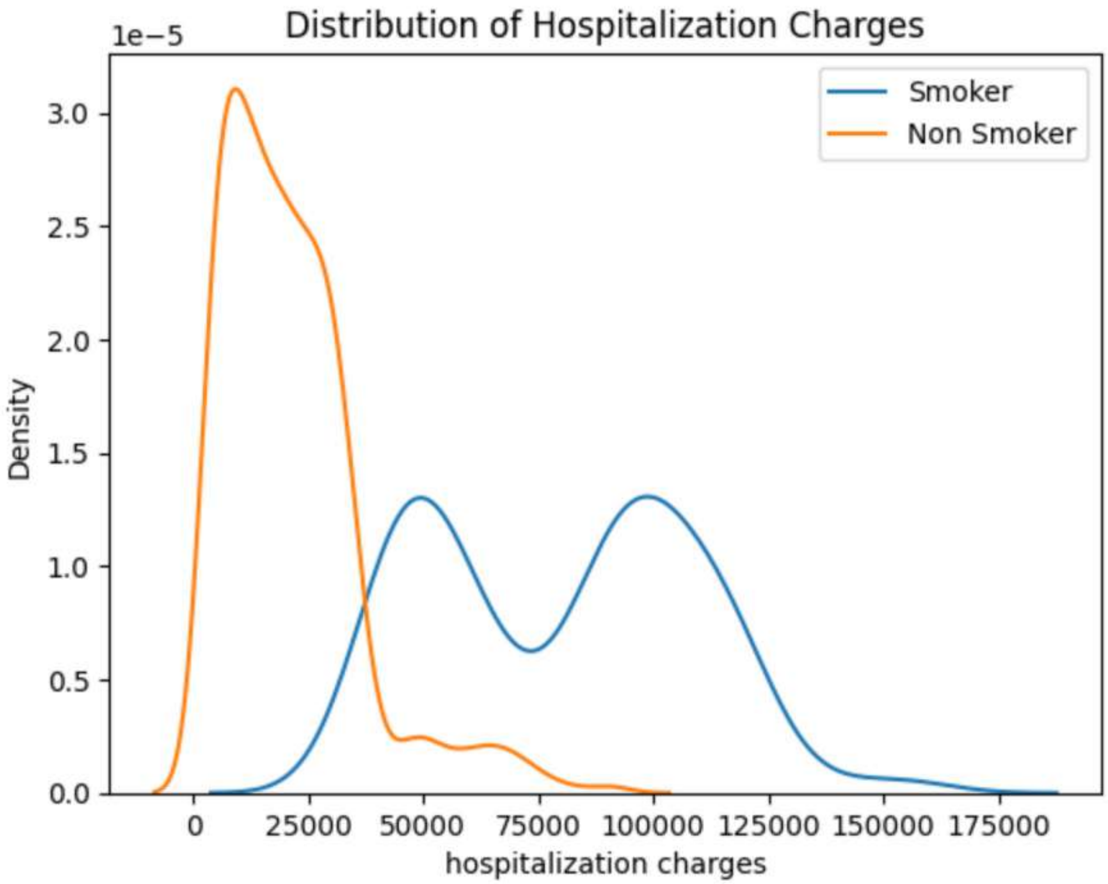
- Null Hypothesis (Ho) = The Hospitalization Charges of People who smoke and don't smoke is Same.
- Alternative Hypothesis (Ha) = The Hospitalization Charges of People who smoke and don't smoke is Different.

In [36]:

```
hospitalization_charges_smokers = df[df['smoker'] == 'yes']['hospitalization charges']
hospitalization_charges_non_smokers = df[df['smoker'] == 'no']['hospitalization charges']
```

In [37]:

```
sns.kdeplot(hospitalization_charges_smokers, label = 'Smoker')
sns.kdeplot(hospitalization_charges_non_smokers, label = 'Non Smoker')
plt.title('Distribution of Hospitalization Charges')
plt.legend()
plt.show()
```



Assumptions of T Test -

1. The data is continuous.
2. The sample data have been randomly sampled from a population.
3. There is homogeneity of variance (i.e., the variability of the data in each group is similar).
4. The distribution is approximately normal.

In [38]:

```
check_variance(hospitalization_charges_smokers, hospitalization_charges_non_smokers)
check_normality(df['hospitalization charges'])
```

```
The difference in variance of Dataset is significant based on Levene Test with P Value = 1.5595259401311176e-66 and St
atistic = 332.6132009308764
The Dataset is not Normal based on Shapiro-Wilk test with P Value = 1.1505333015369624e-36 and Statistic = 0.814688205
```

The assumptions of T Test is not satisfied, because the data is not Normally ditributed and the Difference in Variance is Significant.

We use Mann–Whitney U test, which is a Non Paramteric Test

In [39]:

```
mannwhitneyu(hospitalization_charges_smokers, hospitalization_charges_non_smokers, alternative='greater')
```

Out[39]:

```
MannwhitneyuResult(statistic=284132.5, pvalue=2.6407031043303346e-130)
```

In [97]:

```
stat, p = mannwhitneyu(hospitalization_charges_smokers, hospitalization_charges_non_smokers, alternative='greater')
print(f'P Value = {p} and Statistic = {stat}')
if p <0.05:
    print('The Hospitalization Charges of smoker is Greater than Hospitalization Charges of non - smoker\nHence we rej
else:
    print('The Hospitalization Charges of smoker is not Greater than Hospitalization Charges of non - smoker\nHence we
```

```
P Value = 2.6407031043303346e-130 and Statistic = 284132.5
The Hospitalization Charges of smoker is Greater than Hospitalization Charges of non - smoker
Hence we reject the Null Hypothesis
```

Is the viral load of females different from that of males ?

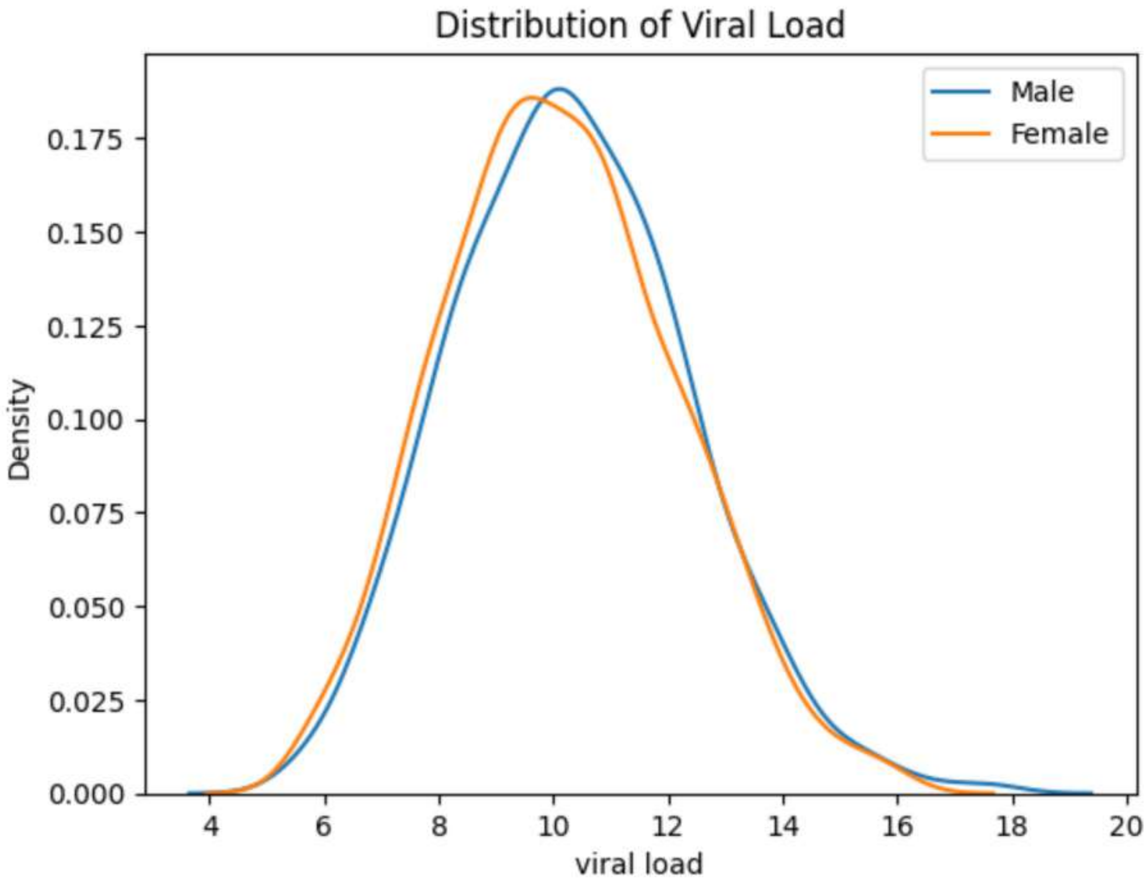
- Null Hypothesis (Ho) = The Viral load of Females is same as Males.
- Alternative Hypothesis (Ha) = The Viral Load of Males is Different from Females.

In [100…

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

In [101…

```
sns.kdeplot(male_viral_load, label = 'Male')
sns.kdeplot(female_viral_load, label = 'Female')
plt.title('Distribution of Viral Load')
plt.legend()
plt.show()
```



Assumptions of T Test -

1. The data is continuous.
2. The sample data have been randomly sampled from a population.
3. There is homogeneity of variance (i.e., the variability of the data in each group is similar).
4. The distribution is approximately normal.

In [102…

```
check_normality(df['viral load'])
check_variance(male_viral_load, female_viral_load)
```

```
The Dataset is not Normal based on Shapiro-Wilk test with P Value = 2.6902040190179832e-05 and Statistic = 0.993904888
6299133
The difference in variance of Dataset is not significant based on Levene Test with P Value = 0.9503708012456551  and S
```

The assumptions of T Test is not satisfied, because the data is not Normally ditributed

We use Mann–Whitney U test, which is a Non Paramteric Test

In [103…

```
stat, p = mannwhitneyu(male_viral_load, female_viral_load, alternative='two-sided')
print(f'P Value = {p}, and Statistic = {stat}')
if p <0.05:
    print('The difference of Viral load between Male and Female is Significant\nHence we reject the Null Hypothesis')
else:
    print('The difference of Viral load between Male and Female is not Significant\nHence we fail to reject the Null H
```

```
P Value = 0.10178463776495861, and Statistic = 235319.0
The difference of Viral load between Male and Female is not Significant
Hence we fail to reject the Null Hypothesis
```

Is the proportion of smoking significantly different across different regions?

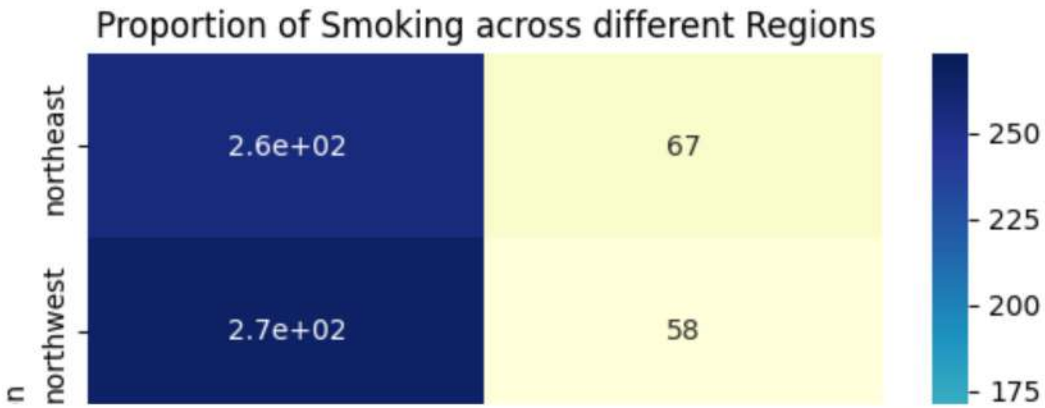
- Null Hypothesis (Ho) = The Proportion of Smoking is not significantly different across different regions.
- Alternative Hypothesis (Ha) = The Proportion of Smoking is not significantly different across different regions.

In [48]:

```
sns.heatmap(pd.crosstab(df['region'], df['smoker']),cmap = "YlGnBu", annot=True)
plt.title('Proportion of Smoking across different Regions')
```

Out[48]:

```
Text(0.5, 1.0, 'Proportion of Smoking across different Regions')
```



Assumptions of Chi-Squared Test

```
1. Observations used in the calculation of the contingency table are independent.
2. 25 or more examples in each cell of the contingency table.
```

In [56]:

```
stats, p, dof, expected_values = chi2_contingency(pd.crosstab(df['region'], df['smoker']))

print(f'P Value = {p} and Statistic = {stats}')
if p <0.05:
    print('The proportion of smoking is significantly different across different regions\nHence we reject the Null Hyp
else:
    print('The proportion of smoking is not significantly different across different regions\nHence we fail to reject
```

```
P Value = 0.06171954839170541 and Statistic = 7.343477761407071
The proportion of smoking is not significantly different across different regions
Hence we fail to reject the Null Hypothesis
```

Is the mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level the same?

- Null Hypothesis (Ho) = The mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level is same.
- Alternative Hypothesis (Ha) = The mean viral load of women with 0 Severity level , 1 Severity level, and 2 Severity level is not same.

In [57]:

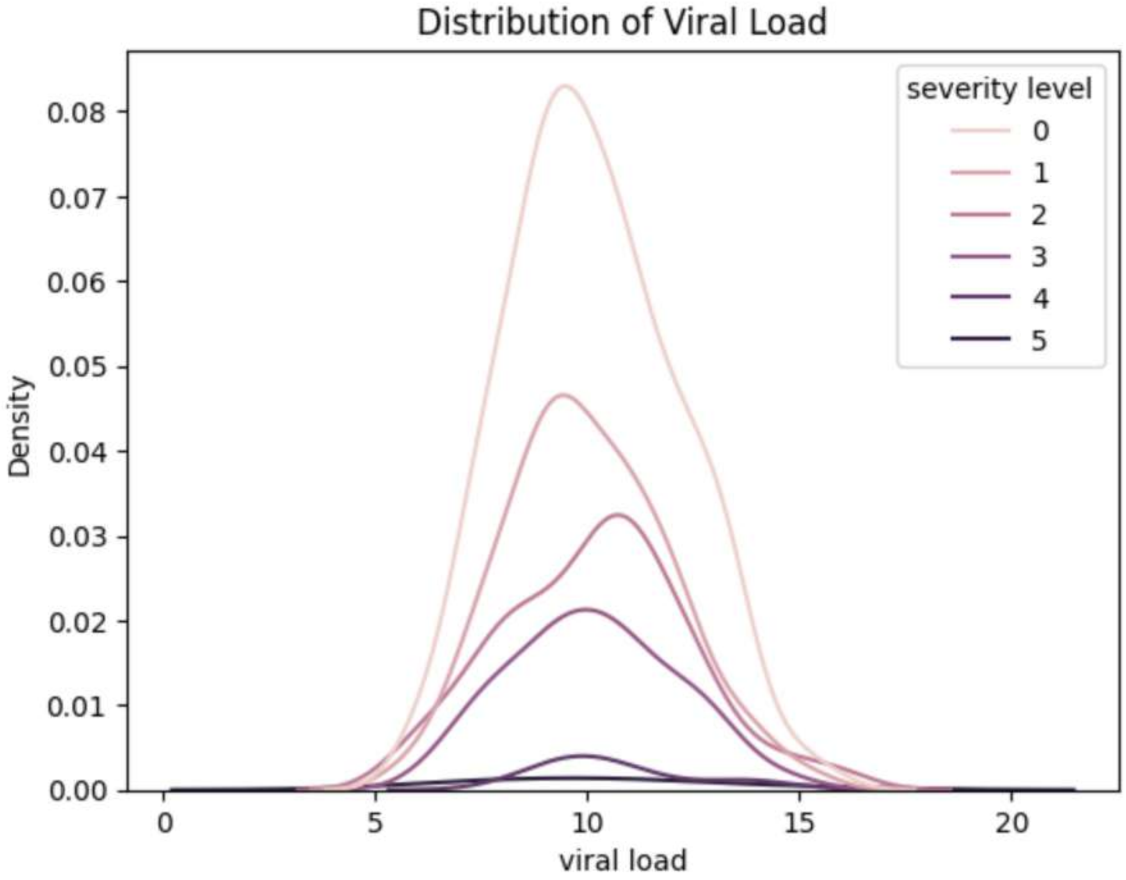
```
df_f = df[df['sex']=='female'] # Data of Sex = Female
x1 = df_f[df_f['severity level']==0]['viral load']
x2 = df_f[df_f['severity level']==1]['viral load']
x3 = df_f[df_f['severity level']==2]['viral load']
```

In [63]:

```
sns.kdeplot(data = df_f, x = 'viral load', hue = 'severity level')
plt.title('Distribution of Viral Load')
```

Out[63]:

```
Text(0.5, 1.0, 'Distribution of Viral Load')
```



Assumptions of ANOVA -

1. Each group sample is drawn from a normally distributed population
2. All populations have a common variance
3. All samples are drawn independently of each other

In [64]:

```
p = levene(x1, x2, x3)[1]
stat = levene(x1, x2, x3)[0]
if p>0.05:
    print(f'The difference in Variance is not Significant, with P Value = {p} and Statistic = {stat}')
else:
    print(f'The difference in Variance is Significant, with P Value = {p} and Statistic = {stat}')
```

```
The difference in Variance is not Significant, with P Value = 0.38987253596513605 and Statistic = 0.9435131022565071
```

We assume that the Data Sample comes from Normal Population

In [65]:

```
p = f_oneway(x1, x2, x3)[1]
stat = f_oneway(x1, x2, x3)[0]
if p>0.05:
    print(f'The difference in Means of Viral Load is not Significant, with P Value ={p} and Statistic = {stat}\nHence
else:
    print(f'The difference in Means of Viral Load is Significant, with P Value = {p} and Statistic = {stat}\nHence we
```

```
The difference in Means of Viral Load is not Significant, with P Value =0.7151189650367746 and Statistic = 0.335506143
4584082
Hence we fail to reject the Null Hypothesis
```

Business Insights

1. 79.52% of The patients are Non Smoker.
2. The People who do not smoke have slightly Larger distribution of Age.
3. Median age of Female (40 years) is slightly larger then Male (39 years)
4. From all the regions, most of the people have zero Severity level, followed by 1, 2 and 3
5. If we consider the minimum Hospitalization charges, Severity Level 5 has the Maximum.
6. If we consider the maximum Hospitalization Charges, its maximum for Severity level 2
7. The Hospitalization Charges of smoker is Greater than Hospitalization Charges of non - smoker
8. The difference of Viral load between Male and Female is not Significant
9. The proportion of smoking is not significantly different across different regions
10. The difference in Means of Viral Load of Women with Different severity level is not Significant
11. Most of the people are coming from South East Region

12. Smoker Patient are more likely to be younger than Non Smoker Patients (Based on the 75th Percentile of age, Smokers have lesser 75th percentile than Non Smokers).
13. Most the Patients are from (20-40] and (40-60] age Group

Recommendations

1. More number of patients are from the Soth East Region, hence its recommended to do appropriate research if these patients have any similar traits.
2. Its also recommended to increase the Staff in these regions, so that they are able to serve the larger patients.
3. North West Region seems to have Most number of Severity level 3 cases and South West seems to have most number Severity level 5 cases, hence its important to have the doctors and staff ready to treat them accordingly.
4. South East Region seems to have more hospitalization charges, hence its recomnmended to provide more suitable financial services like instant loans espcially to these regions.
5. Since the Hospitalization charges of Smokers is more than Non Smokers, the Hospital can easily convince the soker to buy a health insurance or Term life Insurance.
6. Since most the patients are Middle aged, its good to have respective Specialist doctors to treat them.

```
In [ ]:
```