```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import statsmodels.api as sm
```

```
!wget 'https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/681/original/scaler_apollo_hospitals.csv'
```

```
--2023-03-03 05:09:14--  https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/681/original/scaler_apollo_hospitals.cs
Resolving d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)... 108.157.172.10, 108.157.172.183, 108.157.172.173, ...
Connecting to d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)|108.157.172.10|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 53047 (52K) [text/plain]
Saving to: ‘scaler_apollo_hospitals.csv’

scaler_apollo_hospi 100%[===================>]  51.80K  --.-KB/s    in 0.02s

2023-03-03 05:09:14 (2.41 MB/s) - ‘scaler_apollo_hospitals.csv’ saved [53047/53047]
```

```
data=pd.read_csv('scaler_apollo_hospitals.csv')
```

```
data.head()
```

	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

```
data.shape
```

```
(1338, 8)
```

```
data.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
```

```
data.isnull().sum()
```

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

```
data['Unnamed: 0'].value_counts()
```

```
0       1
898     1
896     1
895     1
894     1
       ..
445     1
```

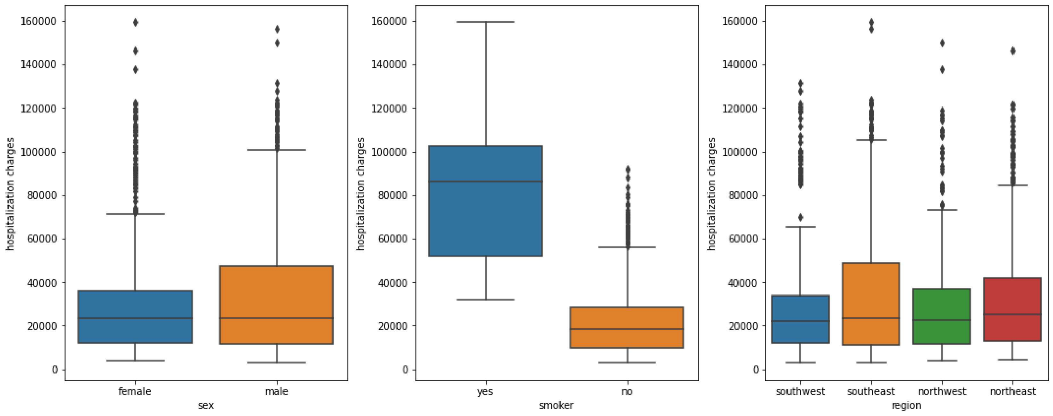
```
    444     1
    443     1
    442     1
    1337    1
    Name: Unnamed: 0, Length: 1338, dtype: int64
```

```
data.drop(columns=['Unnamed: 0'],inplace=True)
```

```
data.select_dtypes('object').columns
```

```
    Index(['sex', 'smoker', 'region'], dtype='object')
```

```
plt.figure(figsize=(15,6))
for i,j in enumerate(list(data.select_dtypes('object').columns)):
  plt.subplot(1,3,i+1)
  plt.subplots_adjust(hspace=0.8)
  sns.boxplot(x=j,y='hospitalization charges',data=data)
  plt.tight_layout(pad=1)
```



```
# median of male and female is same so by visualization we can see that sex doesnt have any effect on hospitalization charges, but we hav
# Median of yes is more than no in smoker on hospitalization charges
# There can be no effect of region on hospitalization charges
# The above conclusions were made based on visual only but we have prove them mathematically through hypo testing
```

```
data.describe(include='object').T
```

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

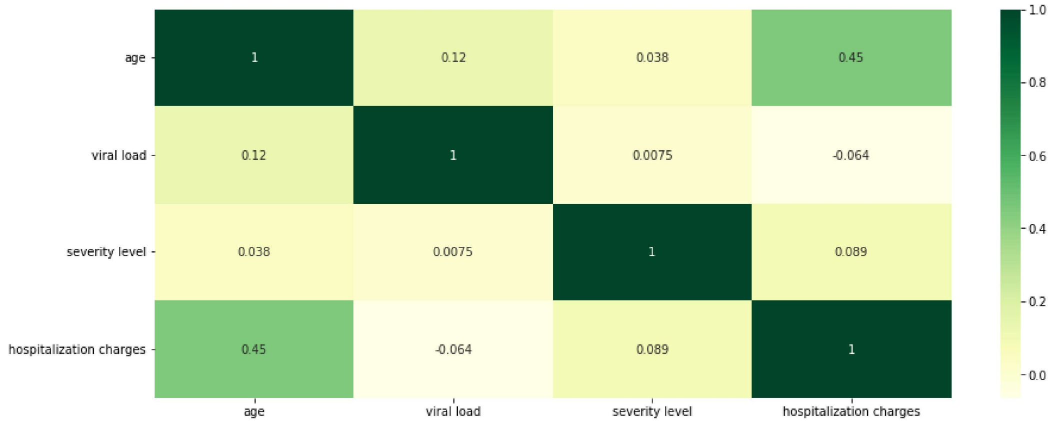
```
data.describe(include=np.number).T
```

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

```
for i,j in enumerate(list(data.select_dtypes(np.number).columns)):
  q1=data[j].quantile(0.25)
  q3=data[j].quantile(0.75)
  iqr=q3-q1
  data=data[(data[j]>=q1-1.5*iqr)&(data[j]<=q3+1.5*iqr)]
```

```
plt.figure(figsize=(15,6))
sns.heatmap(data.corr(),annot=True,cmap='YlGn')
```

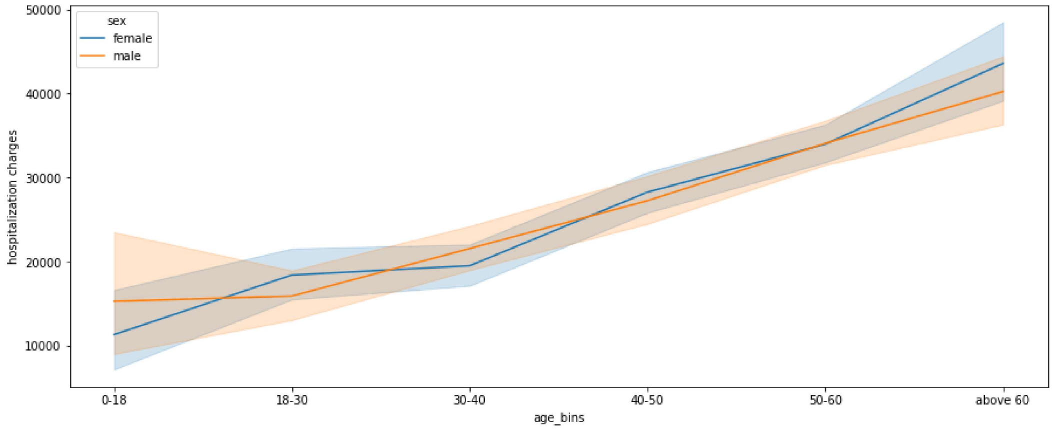
```
plt.show()
```



```
# here we have correlation between age and hospitalization charges
#The best way to explore their relation is to check the age in terms of bins
data['age_bins']=pd.cut(data['age'],bins=[0,18,30,40,50,60,100],labels=['0-18','18-30','30-40','40-50','50-60','above 60'])
```

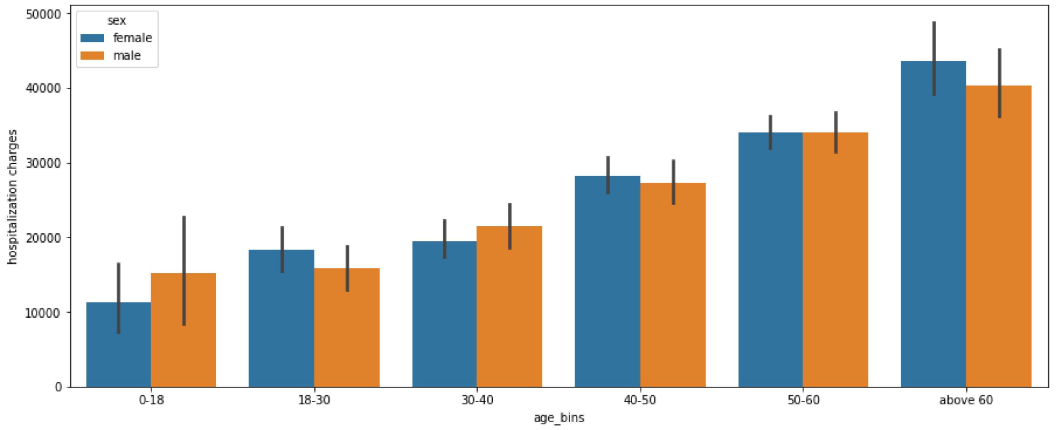
```
plt.figure(figsize=(15,6))
sns.lineplot(x='age_bins',y='hospitalization charges',data=data,hue='sex')
```

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



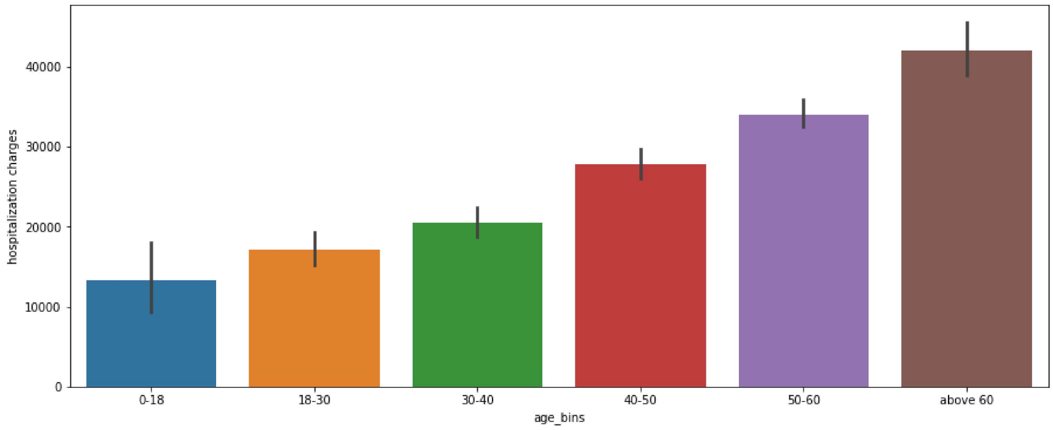
```
plt.figure(figsize=(15,6))
sns.barplot(x='age_bins',y='hospitalization charges',hue='sex',data=data)
```

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



```
plt.figure(figsize=(15,6))
sns.barplot(x='age_bins',y='hospitalization charges',data=data)
```

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



```
data
```

	age	sex	smoker	region	viral load	severity level	hospitalization charges	age_bins
0	19	female	yes	southwest	9.30	0	42212	18-30
1	18	male	no	southeast	11.26	1	4314	0-18
2	28	male	no	southeast	11.00	3	11124	18-30
3	33	male	no	northwest	7.57	0	54961	30-40
4	32	male	no	northwest	9.63	0	9667	30-40
...	...	...	...	...	...	...	...	...
1333	50	male	no	northwest	10.32	3	26501	40-50
1334	18	female	no	northeast	10.64	0	5515	0-18
1335	18	female	no	southeast	12.28	0	4075	0-18
1336	21	female	no	southwest	8.60	0	5020	18-30
1337	61	female	yes	northwest	9.69	0	72853	above 60

1191 rows × 8 columns

```
data.groupby(['region','sex','smoker']).mean()['hospitalization charges'].unstack()
```

	smoker	no	yes
region	sex		
northeast	**female**	24105.053435	48756.263158
	male	21660.096000	56480.500000
northwest	**female**	21967.518519	58942.350000
	male	20801.734848	56219.444444
southeast	**female**	20590.739130	56240.470588
	male	19123.868217	53855.058824
southwest	**female**	19585.122302	54129.909091
	male	19447.293651	50628.571429

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

Here we have to do t test right tailed and t test independent

H0: mu1 <= mu2 --> the average hospitalization charges of smoking people is less than or equal to non smoking people

Ha: mu1 > mu2 --> The average hospitalization charges of smoking people is greater than non smoking people

alpha=0.05

```
data['smoker'].value_counts()
```

```
no     1055
yes     136
Name: smoker, dtype: int64
```

```
aplha=0.05
```

```
# Here we have to take equal samples
# Let the sample size be 110
smokers=data[data['smoker']=='yes']['hospitalization charges'].sample(110)
non_smokers= data[data['smoker']=='no']['hospitalization charges'].sample(110)
```

```
t_statistic,p=stats.ttest_ind(smokers,non_smokers,equal_var=False)
## Here we took equal _variance as fasle as we can see below that their variances are not equal
```

```
data.groupby('smoker')['hospitalization charges'].describe()
```

	count	mean	std	min	25%	50%	75%	max
smoker								
no	1055.0	20907.971564	14563.067125	2805.0	9962.5	18313.0	28387.5	83680.0
yes	136.0	54578.154412	13360.849267	32074.0	44663.5	51899.5	61421.5	85758.0

```
onetailed_p=p/2
```

```
print('Test statistic = {} ,P-value = {} ,onetailed_p = {} '.format(t_statistic,p,onetailed_p))
```

```
Test statistic = 17.075044899859243 ,P-value = 2.513391454551179e-41 ,onetailed_p = 1.2566957272755896e-41
```

```
# Here we can see that onetailed_p<alpha so we cannot accept null hypothesis.
# Therefore The average hospitalization cahrges of smokers is greater than non smokers
```

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

```
data.groupby('sex')['viral load'].describe()
```

	count	mean	std	min	25%	50%	75%	max
sex								
female	610.0	9.966541	1.96940	5.60	8.585	9.855	11.1725	15.58
male	581.0	10.030947	1.94495	5.32	8.600	9.940	11.2900	15.51

Here we have to perform two tailed test

H0: The average viral load of females and males are same ----> mu_1=mu_2

Ha: The average viral load of females and males are not same ----> mu_1!=mu_2

aplha=0.05

```
males=data[data['sex']=='female']['viral load'].sample(581)
females=data[data['sex']=='male']['viral load'].sample(581)
```

```
alpha=0.05
t_statistic,p=stats.ttest_ind(males,females,alternative='two-sided')
print('test_statistic = {},p-value ={}'.format(t_statistic,p))
```

```
test_statistic = -0.627233980931372,p-value =0.5306293066829824
```

```
if p>alpha:
  print("We fail to reject null hypothesis")
else:
  print('We reject null hypothesis')
```

```
We fail to reject null hypothesis
```

3. Is the proportion of smoking significantly different across different regions? (chi-square)

Here why we are using chi-square test because there are two categorical variables which are to be compared

```
x=pd.crosstab(data['region'],data['smoker'])
x
```

smoker	no	yes
region		
northeast	256	39
northwest	267	38
southeast	267	34
southwest	265	25

```
chi2, pval, dof, exp_freq = stats.chi2_contingency(x, correction = False)
print('chi-square statistic: {} , Pvalue: {} , Degree of freedom: {} ,expected frequencies: {} '.format(chi2, pval, dof, exp_freq))
```

```
chi-square statistic: 3.5220357595425758 , Pvalue: 0.31791538258247426 , Degree of freedom: 3 ,expected frequencies: [[261.31402183
 [270.17212427  34.82787573]
 [266.62888329  34.37111671]
 [256.88497061  33.11502939]]
```

```
if pval>0.05:
  print('Failed to reject null hypothesis')
else:
  print('Reject Null Hypothesis')
```

```
Failed to reject null hypothesis
```

4. 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.

H0: μ1 = μ2 = μ3 ---> The mean viral load of women with no severity level , one severity level,two severity level is same

Ha: Atleast one of mean viral load of women is not same

```
data[data['sex']=='female'].groupby('severity level')['viral load'].describe()
```

	count	mean	std	min	25%	50%	75%	max
severity level								
0	268.0	9.963209	1.936862	5.76	8.6075	9.695	11.1300	14.92
1	147.0	9.908844	1.918987	5.60	8.6000	9.670	11.1300	15.36
2	106.0	9.945000	2.092305	5.73	8.3450	10.060	11.2025	15.57
3	71.0	10.014366	1.950361	6.33	8.5950	10.030	11.1750	14.90
4	10.0	10.601000	1.815063	8.53	9.6200	9.825	11.0350	13.82
5	8.0	10.206250	2.975480	6.10	8.0500	10.080	11.6625	15.58

```
# But here we need only 0,1,2 severity levels
data_female=data[data['sex']=='female'].loc[data[data['sex']=='female']['severity level']<=2]
```

```
data_female
```

	age	sex	smoker	region	viral load	severity level	hospitalization charges	age_bins
0	19	female	yes	southwest	9.30	0	42212	18-30
5	31	female	no	southeast	8.58	0	9392	30-40
6	46	female	no	southeast	11.15	1	20601	40-50
9	60	female	no	northwest	8.61	0	72308	50-60
11	62	female	yes	southeast	8.76	0	69522	above 60
...	...	...	...	...	...	...	...	...
1331	23	female	no	southwest	11.13	0	26990	18-30
1334	18	female	no	northeast	10.64	0	5515	0-18
1337	61	female	ves	northwest	9.69	0	72853	above 60

- Here we have to perform ANova test
- for Anova there are mainly two type of assumptions: 1) Normal distribution assumptions 2) Variance equal

For Normal distribution assumption we use Shapiro-wilk test

H0: Data is Noramlly distributed

Ha: Data is not normally distributed

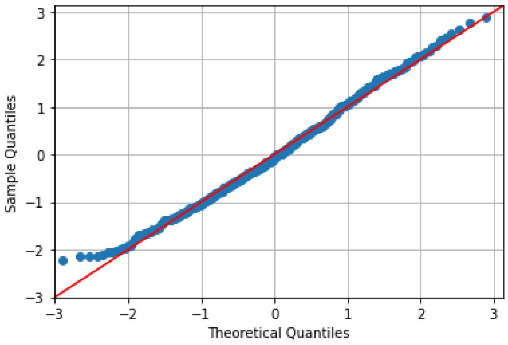
```
w,p=stats.shapiro(data_female['viral load'])
p
```

```
    0.0068216221407055855
```

```
wlog,plog=stats.shapiro(np.log(data_female['viral load']))
plog
```

```
    0.009918625466525555
```

```
# Here we can see both p,plog values are less than alpha
# Lets check this with qqplot
f=sm.qqplot(data_female['viral load'],line='45',fit=True)
plt.grid()
```



```
# Through visual analysis we can see that the data is normal distributed
```

2) Levene test

H0: All viral load have same variance

Ha: All viral load doesnt have same variance

```
statistic, p_value = stats.levene( data_female[data_female['severity level']==0]['viral load'].sample(106),
                             data_female[data_female['severity level']==1]['viral load'].sample(106),
                             data_female[data_female['severity level']==2]['viral load'].sample(106))
```

```
p_value
```

```
    0.13916641311791475
```

```
if p_value>0.05:
  print("All have same variance")
```

```
else:
  print('All data doesnt have same variance')
```

```
    All have same variance
```

ANOVA

```
test_statistic,p_value=stats.f_oneway(data_female[data_female['severity level']==0]['viral load'].sample(106),
                                      data_female[data_female['severity level']==1]['viral load'].sample(106),
                                      data_female[data_female['severity level']==2]['viral load'].sample(106))
p_value
```

```
    0.3800708037771745
```

```
if p_value>0.05:
  print("The mean viral load of women with no severity level , one severity level,two severity level is same")
else:
  print("Atleast one of mean viral load of women is not same")
```

```
    The mean viral load of women with no severity level , one severity level,two severity level is same
```

RECOMMENDATIONS

- People who are smoking having greater chance of hospitalization so apollo should consider making a seperate department for treating them.
- Apollo should make some campaigns on the effects of smoking.
- Apollo should consider that the viral load is independent of gender.
- They should consider of establishing the departments to treat smokers irrespective of regions.

✓ 0s completed at 1:05 PM