

```

# 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 statistica
# 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 un


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

Downloading...
From: https://d2beiqkhq929f0.cloudfront.net/public\_assets/assets/000/001/681/original/scaler\_apollo\_hospitals.csv
To: /content/scaler_apollo_hospitals.csv
100% 53.0k/53.0k [00:00<00:00, 276kB/s]


import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns


from scipy.stats import ttest_ind,ttest_rel,f_oneway
from scipy.stats import chi2_contingency,pearsonr,spearmanr,kendalltau


df = pd.read_csv('/content/scaler_apollo_hospitals.csv')

df.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

```
df.shape
(1338, 8)
```

Dataset consists of 1338 rows and 8 axes.

```
df.describe()
```

	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

```
df[df['hospitalization charges']==159426]
```

	Unnamed: 0	age	sex	smoker	region	viral load	severity level	hospitalization charges
543	543	54	female	yes	southeast	15.8	0	159426

We can see that:

- 1. Un-named column is just an index column.
- 2. From Age column, we can see that mean age is 39 years, Min age is near to 18 Years and Maximum is 64 years. So dataset consists of almost all ages person.
- 3. Viral Load mean in patients are near to 10, which seems to be greater.
- 4. Severity Level as the dataset consists of upto lvl 5 severity, so mean is around 1 which seems that most patients are in good condition/minor healt affected.
- 5. Hospitalization charges, if a person is getting admitted, the minimum he will pay is around 30-33k. Maximum hospital charge is paid around 160k, might be that person could have maximum severity level, viral load as well. And after checking record, the age is 54 and he is smoker viral load is near 16 (critical condition).

```
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
```

No null values, so no need to change data.

```
df.isnull().sum()

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

df.columns

Index(['Unnamed: 0', 'age', 'sex', 'smoker', 'region', 'viral load',
      'severity level', 'hospitalization charges'],
      dtype='object')

df.drop(['Unnamed: 0'],axis=1,inplace=True)
# Dropping Unwanted columns: Unnamed column

cols = ['age', 'sex', 'smoker', 'region', 'viral load',
        'severity level', 'hospitalization charges']
for i in cols:
    print(i,":",df[i].nunique())

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

From above columns we can see that there are 4 Categorical column excluding age, but age can be made categorical. Categorical Columns:

1. smoker
2. sex
3. region
4. severity level

```
df['age'].unique()

array([19, 18, 28, 33, 32, 31, 46, 37, 60, 25, 62, 23, 56, 27, 52, 30, 34,
       59, 63, 55, 22, 26, 35, 24, 41, 38, 36, 21, 48, 40, 58, 53, 43, 64,
       20, 61, 44, 57, 29, 45, 54, 49, 47, 51, 42, 50, 39])

print("Min Age:",min(df['age']),"Max Age: ",max(df['age']))

Min Age: 18 Max Age:  64

df['age'].value_counts()

18    69
19    68
50    29
51    29
47    29
46    29
45    29
20    29
48    29
52    29
22    28
49    28
54    28
53    28
21    28
26    28
24    28
```

```

25    28
28    28
27    28
23    28
43    27
29    27
30    27
41    27
42    27
44    27
31    27
40    27
32    26
33    26
56    26
34    26
55    26
57    26
37    25
59    25
58    25
36    25
38    25
35    25
39    25
61    23
60    23
63    23
62    23
64    22
Name: age, dtype: int64

```

Each age has near to 20-25 participants.

```

bins = [0, 18, 41, 64]
labels = ['Teen', 'Mid', 'Aged']
df['Age_Group'] = pd.cut(x = df['age'], bins = bins, labels = labels, include_lowest = True)

```

```
df['Age_Group'].value_counts()
```

```

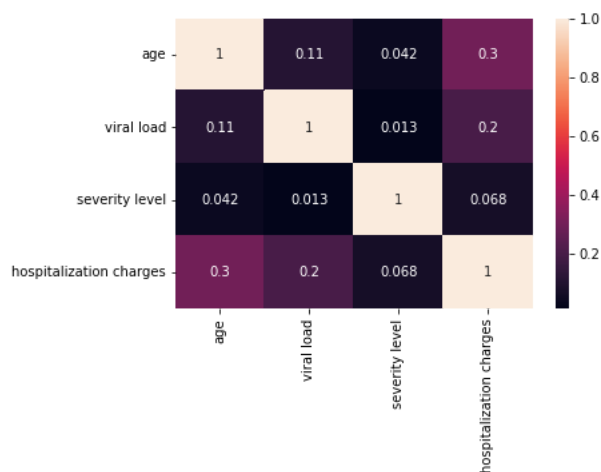
Mid      659
Aged     610
Teen      69
Name: Age_Group, dtype: int64

```

```

sns.heatmap(df.corr(),annot=True)
plt.show()

```

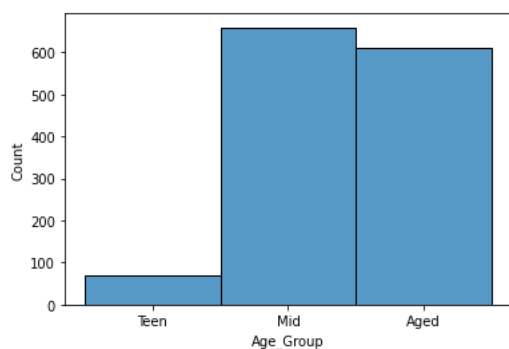


From heat map we can't conclude any direct relationship in columns, but we can conclude some columns that might have some relationship in them.

1. age vs hosp charges
2. viral load vs hosp charges
3. severity level vs hosp charges

Univariate Analysis:

```
sns.histplot(data = df['Age_Group'])
plt.show()
```



```
df.columns
```

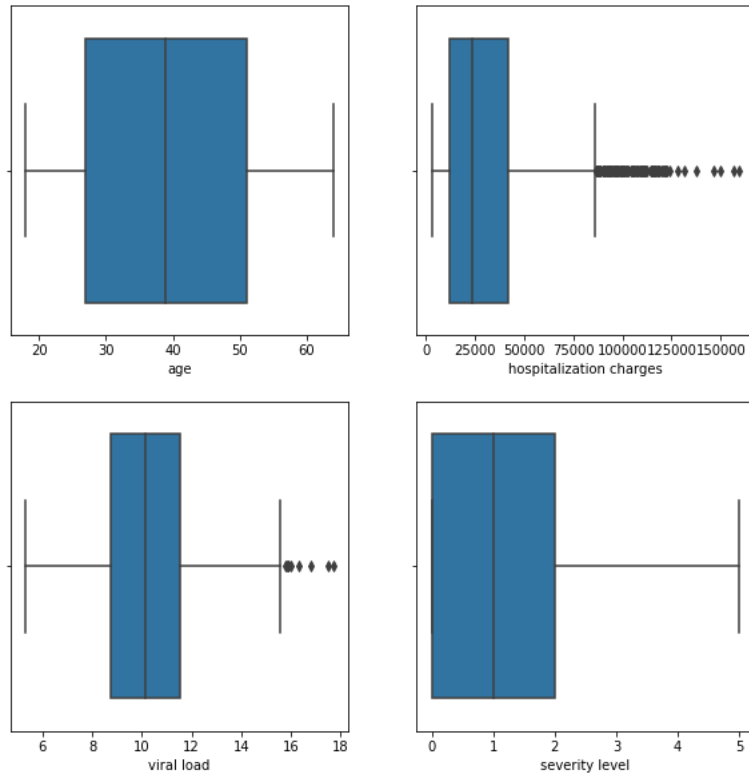
```
Index(['age', 'sex', 'smoker', 'region', 'viral load', 'severity level',
       'hospitalization charges', 'Age_Group'],
      dtype='object')
```

```
fig, axes = plt.subplots(nrows = 2, ncols = 3, figsize=(15,15))
sns.histplot(data = df, x='age', ax = axes[0,0], kde = True)
sns.histplot(data = df, x='sex', ax = axes[0,1], kde = True)
sns.histplot(data = df, x='region', ax = axes[0,2], kde = True)
sns.histplot(data = df, x='viral load', ax = axes[1,0], kde = True)
sns.histplot(data = df, x='severity level', ax = axes[1,1], kde = True)
sns.histplot(data = df, x='hospitalization charges', ax = axes[1,2], kde = True)
plt.show()
```



We can see, Ages are almost equal so no more deflection. In regions we can see that South-east region is more. We can see a Binomial distribution in Viral load and Left-Tailed skew distribution in Hospitalization Charges.

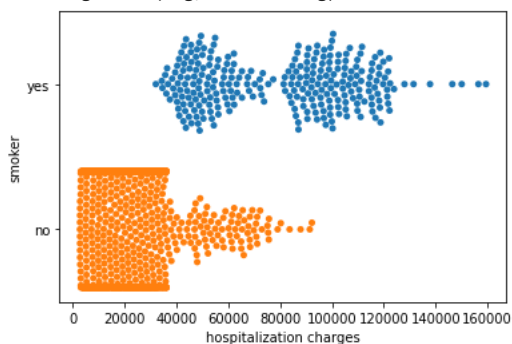
```
fig, axes = plt.subplots(nrows = 2, ncols = 2, figsize=(10,10))
sns.boxplot(data = df, x='age', ax = axes[0,0])
sns.boxplot(data = df, x='viral load', ax = axes[1,0])
sns.boxplot(data = df, x='severity level', ax = axes[1,1])
sns.boxplot(data = df, x='hospitalization charges', ax = axes[0,1])
plt.show()
```



We can see that there are almost none Outliers in Severity Level, Viral Load and Age. But we can see outliers in Hospitalization Charges, as mentioned earlier the mean is around 30-33k, but we can see that on severity levels the hospitalization charges can be more, so outliers are possible in Hospitalization Charges.

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

/usr/local/lib/python3.8/dist-packages/seaborn/categorical.py:1296: UserWarning: 68.1% of the points cannot be placed; you may want to c
warnings.warn(msg, UserWarning)



As there are around 300 Smokers in dataset and almost all smokers Hospitalization Charges are above Mean level, and we can clearly see that the ratio of non-smokers are more in dataset but still there Hospitalization charges are near 40k, leaving some cases.

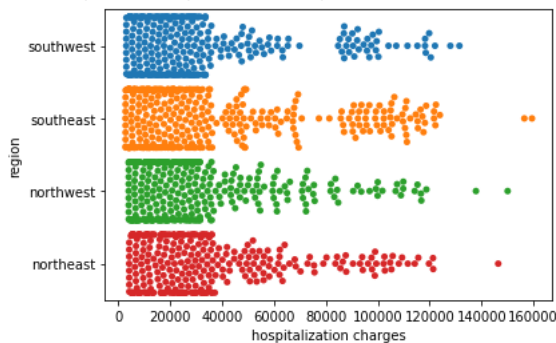
```
pd.crosstab(df['hospitalization charges'],df['region'], margins=True)
```

	region	northeast	northwest	southeast	southwest	All
hospitalization charges						
2805		0	0	1	0	1
2829		0	0	1	0	1
2840		0	0	1	0	1
2841		0	0	1	0	1
2843		0	0	1	0	1
...	
146428		1	0	0	0	1
150053		0	1	0	0	1
156482		0	0	1	0	1
159426		0	0	1	0	1
All		324	325	364	325	1338

1321 rows × 5 columns

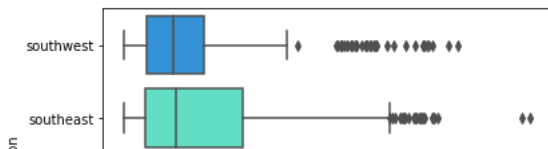
```
sns.swarmplot(x="hospitalization charges",y="region", data=df)
plt.show()
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/categorical.py:1296: UserWarning: 42.5% of the points cannot be placed; you may want to c
warnings.warn(msg, UserWarning)
/usr/local/lib/python3.8/dist-packages/seaborn/categorical.py:1296: UserWarning: 39.6% of the points cannot be placed; you may want to c
warnings.warn(msg, UserWarning)
/usr/local/lib/python3.8/dist-packages/seaborn/categorical.py:1296: UserWarning: 41.5% of the points cannot be placed; you may want to c
warnings.warn(msg, UserWarning)
/usr/local/lib/python3.8/dist-packages/seaborn/categorical.py:1296: UserWarning: 39.2% of the points cannot be placed; you may want to c
warnings.warn(msg, UserWarning)
```



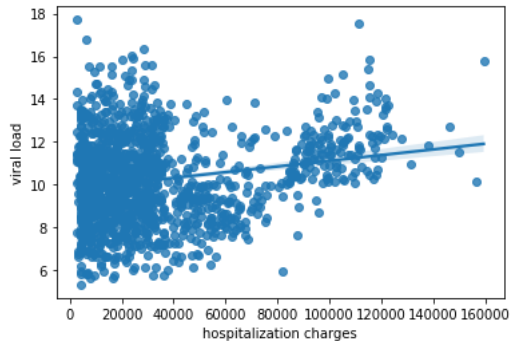
We can see that most of patients from all region are near to **Mean Levels**, but if we compare more deeply Southeast and Northeast regions patients are more likely to pay high Hospitalization Charges.

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

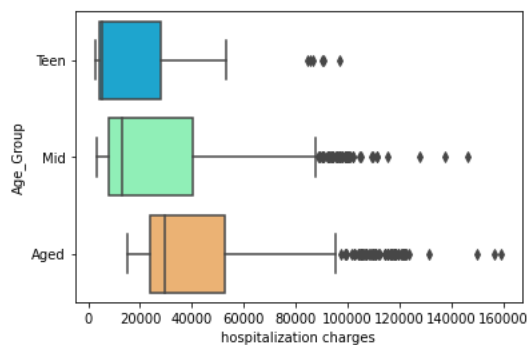


Bivariate Analysis (Relationships between important variables such as hospitalization charges with region, smoker, viral load etc)

```
sns.regplot(x='hospitalization charges',y='viral load',data=df)
plt.show()
```



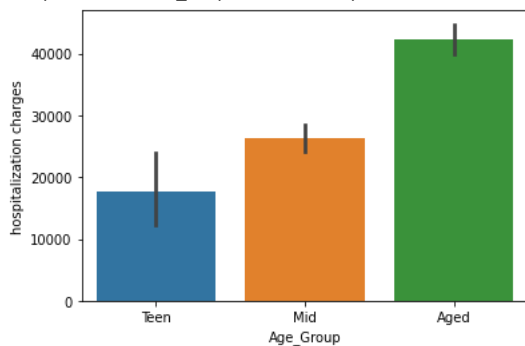
```
sns.boxplot(x='hospitalization charges',y='Age_Group',data=df,palette='rainbow')
plt.show()
```



From the above boxplots, we can conclude that Age groups matters in Hospitalization Charges as Age increases the mean of Hospitalization Charges also increases. Let us conclude this statement by performing Hypothesis Test.

```
sns.barplot(df['Age_Group'],df['hospitalization charges'])
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. Frc
warnings.warn(
<matplotlib.axes._subplots.AxesSubplot at 0x7f142ae8a310>
```



H0: Age and Hospitalization are Dependent on Each Other.

H1: Age and Hospitalization are Independent on Each Other.

Significance Level : 0.05

Test Statistics:

1. Pearson Correlation Coefficient Test
2. Spearman Rank Test

```
from scipy.stats import pearsonr
stat, p = pearsonr(df['age'], df['hospitalization charges'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably independent')
else:
    print('Probably dependent')
```

```
stat=0.299, p=0.000
Probably dependent
```

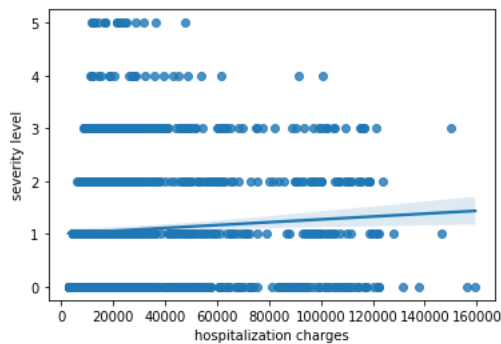
```
stat, p = spearmanr(df['age'], df['hospitalization charges'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably independent')
else:
    print('Probably dependent')
```

```
stat=0.534, p=0.000
Probably dependent
```

We fail to reject Null Hypothesis.

We can conclude from Pearson's as well as Spearman's Correlation test, that both are dependent features.

```
sns.regplot(x='hospitalization charges',y='severity level',data=df)
plt.show()
```



```
pd.crosstab(df['Age_Group'],df['smoker'], margins=True)
```

	smoker		
	no	yes	All
Age_Group			
Teen	57	12	69
Mid	520	139	659
Aged	487	123	610
All	1064	274	1338

We can see that Mid-Aged group. followed by Aged Group people are more likely to consume Smoking.

```
sns.scatterplot(x='age',y='region',data=df,hue = 'smoker',palette='rainbow')
plt.show()
```

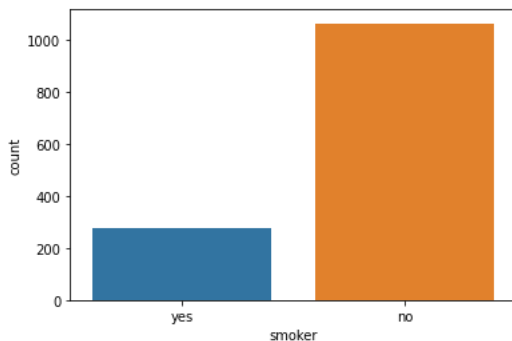


```
df['sex'].value_counts()
```

```
male      676
female    662
Name: sex, dtype: int64
```

```
# 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
```

```
sns.countplot(x=df['smoker'])
plt.show()
```



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

H0: Smoking doesnot affects Hospitalization

H1: Smoking affects Hospitalization

Significance Level: 0.05

Tests:

1. T-Test
2. ANOVA

```
smoker = df[df['smoker']=='yes']['hospitalization charges'].values
nonsmoker = df[df['smoker']=='no']['hospitalization charges'].values
stat1, p1 = ttest_ind(smoker,nonsmoker)
print('stat=%.3f, p=%.3f' % (stat1, p1))
if p1 > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')

stat=46.665, p=0.000
Probably different distributions
```

```
#ANOVA TEST
stat3, p3 = f_oneway(smoker,nonsmoker)
print('stat=%.3f, p=%.3f' % (stat3, p3))
if p3 > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')
```

```
stat=2177.612, p=0.000
Probably different distributions
```

By conducting tests , we can reject Null Hypothesis and from Plots and By tests we can conclude that by smoking the chances of Hospitalization increases and increases the Charges.

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

H0: The viral load of Females are similar to that of Males.

H1: The viral load of Females are different from Males.

Significance Level: 0.05

Test Statistics:

1. T-Test
2. Anova

```
viralload_male = df[df['sex']=='male']['viral load'].values
viralload_female = df[df['sex']=='female']['viral load'].values
```

```
stat1, p1 = ttest_ind(viralload_male,viralload_female)
print('stat=%.3f, p=%.3f' % (stat1, p1))
if p1 > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')
```

```
stat=1.696, p=0.090
Probably the same distribution
```

```
#ANOVA TEST
stat3, p3 = f_oneway(viralload_male,viralload_female)
print('stat=%.3f, p=%.3f' % (stat3, p3))
if p3 > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')
```

```
stat=2.875, p=0.090
Probably the same distribution
```

From Both tests we got same p-value, so from that we fail to reject the Null Hypothesis.

Is the proportion of smoking significantly different across different regions?

H0: The proportion of smoking is not significantly different across different regions.

H1: The proportion of smoking is significantly different across different regions.


Significance Level: 0.05

Tests:

1. Chi2 contingency test

```
smoke_effect = pd.crosstab(df['smoker'],df['region'])
```

```
smoke_effect
```

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

```
stat, p, dof, expected = chi2_contingency(smoke_effect)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably independent')
else:
    print('Probably dependent')

stat=7.343, p=0.062
Probably independent
```

As p-value greater than significance value, we reject Null Hypothesis and after more evidence we can say that the smoking is different in different regions.

```
smokers = df[df['smoker']=='yes']['region']
non_smokers = df[df['smoker']=='no']['region']
```

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

H0: All the severity level are not same.

H1: All the severity level are similar.

Significance Level: 0.05

Test Statistics:

1. Anova

```
women_df = df[df['sex']=='female']
```

```
zero_sl = women_df[women_df['severity level']==0]['viral load']
one_sl = women_df[women_df['severity level']==1]['viral load']
two_sl = women_df[women_df['severity level']==2]['viral load']
three_sl = women_df[women_df['severity level']==3]['viral load']
```

```
stat, p = f_oneway(zero_sl, one_sl, two_sl)
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')
```

```
stat=0.336, p=0.715
Probably the same distribution
```

We can reject Null Hypothesis and can say that in case of Women the mean of severity are same.

Insights:

1. From Age, we can see that mean age is 39 years, Min age is near to 18 Years and Maximum is 64 years. So dataset consists of almost all ages person.
2. Viral Load mean in patients are near to 10, which seems to be greater. Severity Level as the dataset consists of upto lvl 5 severity, so mean is around 1 which seems that most patients are in good condition/minor health affected.
3. Hospitalization charges, if a person is getting admitted, the minimum he will pay is around 30-33k. Maximum hospital charge is paid around 160k, might be that person could have maximum severity level, viral load as well. And after checking record, the age is 54 and he is smoker viral load is near 16 (critical condition).
4. Number of Patients are more from region of South-East and North-west.

5. South-east region also has most number of Active Smokers.

Recommendations: For Hospital:

1. There are less number of teens who are hospitalized so for cost cutting purpose, the doctors who are treating teens can be lay off.
2. Hospital has more number of Patients from South West region and in that region the smoking habit is also more, so more campaigns and workshops can be hosted.
3. As there are more numbers of Aged group patients are present, then more number of Doctors who treats this age group persons needs to be hired.
4. As the number of Genders are almost equal so similarly the gender ratio of Doctors should also be maintained same.

For Patients:

1. As the number of Patients increases according to increase in age, Patients/peoples of all the regions should starts excersing and starts to maintain there health by eating healthy foods and maintaining consistancy in excersice.
2. From data we can also observe that more smokers are from Mid and Aged persons, so we can also conclude that when the teenagers will grow older most of them will start smoking, so in order to prevent this situation campaigns and self awareness in people should also be increased.