In [1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as plx
```

In [2]:
```
from plotly.offline import init_notebook_mode
init_notebook_mode(connected = True)
```

In [3]:
```
from scipy.stats import ttest_ind
```

In [4]:
```
df=pd.read_csv(r"C:\Vidhya\MyWork\scaler\probability\Apollo hospital case study\sca
```

In [5]:
```
df.sample(10)
```

Out[5]:

	Unnamed: 0	age	sex	smoker	region	viral load	severity level	hospitalization charges
168	168	19	female	no	northwest	10.61	1	6798
912	912	59	female	no	northwest	8.90	3	35957
97	97	55	male	no	southeast	12.76	0	25566
62	62	64	male	no	northwest	8.23	1	75417
1245	1245	28	male	no	southwest	8.10	5	14038
1164	1164	41	female	no	northwest	9.44	1	17884
404	404	31	male	no	southwest	6.80	0	8150
776	776	40	male	no	northwest	10.77	2	17467
1240	1240	52	male	yes	southeast	13.93	2	118175
1183	1183	48	female	no	northeast	9.12	1	23618

Objective : Extract meaningful and actionable insights

Datatype and shape of data

In [6]:
```
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
```

Statistical Summary

In [7]:
```
df.describe()
```

Out[7]:

	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 [8]:
```
print(df["sex"].value_counts())
print(df["smoker"].value_counts())
print(df["region"].value_counts())
print(df["severity level"].value_counts())
```

```
male      676
female    662
Name: sex, dtype: int64
no     1064
yes     274
Name: smoker, dtype: int64
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
0    574
1    324
2    240
3    157
4     25
5     18
Name: severity level, dtype: int64
```

In [9]:
```
df["sex"]=df["sex"].astype("category")
df["smoker"]=df["smoker"].astype("category")
```

```
df["region"]=df["region"].astype("category")
df["severity level"]=df["severity level"].astype("category")
```

In [10]:
```
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   category
 3   smoker                  1338 non-null   category
 4   region                  1338 non-null   category
 5   viral load              1338 non-null   float64 
 6   severity level          1338 non-null   category
 7   hospitalization charges  1338 non-null   int64   
dtypes: category(4), float64(1), int64(3)
memory usage: 47.8 KB
```

Missing Value Detection

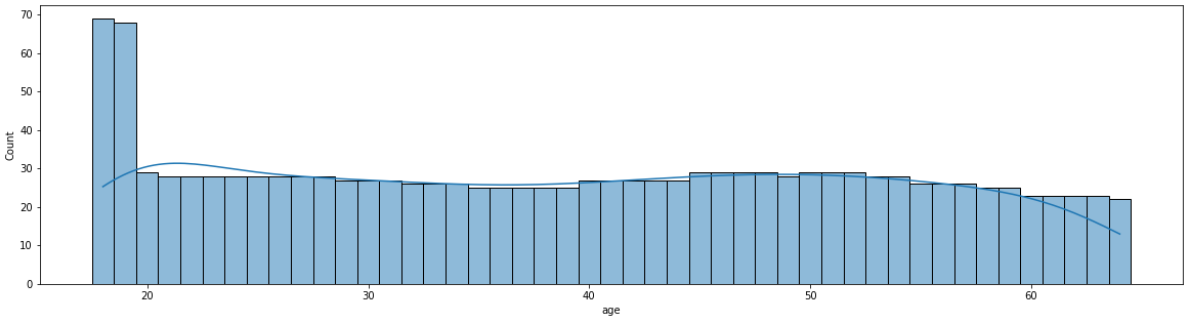
In [11]:
```
df.isna().sum()
```

Out[11]:
```
Unnamed: 0                0
age                       0
sex                       0
smoker                    0
region                    0
viral load                0
severity level            0
hospitalization charges    0
dtype: int64
```

There are no missing values

In [12]:
```
fig=plt.figure(figsize=(20,5))
sns.histplot(x='age',data=df,discrete=True,kde=True)
```

Out[12]:
```
<AxesSubplot:xlabel='age', ylabel='Count'>
```



In [13]:
```
bins=[1,20,30,40,50,60,70]
df["age_bin"]=pd.cut(x=df["age"],bins=bins,labels=['18-20','21-30','31-40','41-50'
df.sample(10)
```

Out[13]:

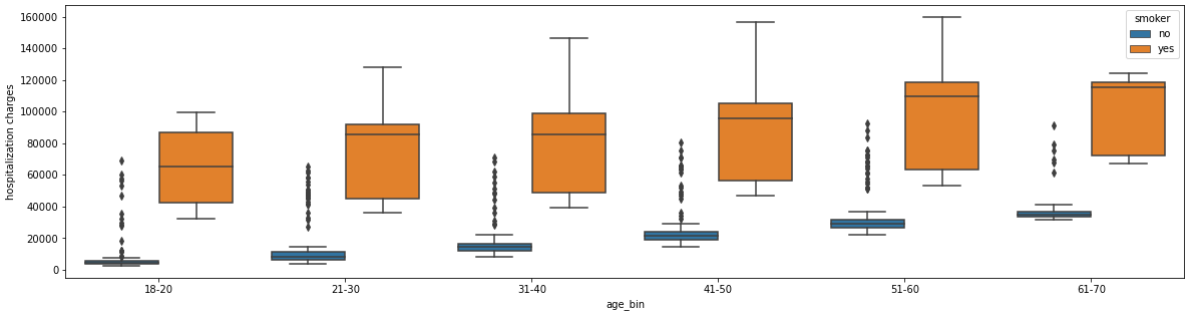
	Unnamed: 0	age	sex	smoker	region	viral load	severity level	hospitalization charges	age_bin
1280	1280	48	female	no	southeast	11.11	0	20709	41-50
841	841	59	male	no	northeast	8.23	0	30810	51-60
1259	1259	52	female	no	northeast	7.73	0	25494	51-60
38	38	35	male	yes	northeast	12.22	1	99436	31-40
1325	1325	61	male	no	northeast	11.18	0	32858	61-70
386	386	58	female	no	southeast	13.02	0	29641	51-60
463	463	56	male	no	northeast	8.64	0	27914	51-60
201	201	48	female	no	southeast	10.74	1	22178	41-50
1047	1047	22	male	yes	southeast	17.53	1	111253	21-30
681	681	19	male	no	southwest	6.77	0	3106	18-20

In [14]:
```
df['age_bin'].value_counts()
```

Out[14]:
```
41-50    281
21-30    278
51-60    265
31-40    257
18-20    166
61-70     91
Name: age_bin, dtype: int64
```

In [15]:
```
fig=plt.figure(figsize=(20,5))
sns.boxplot(data=df,x='age_bin',y='hospitalization charges',hue="smoker")
```

Out[15]:
```
<AxesSubplot:xlabel='age_bin', ylabel='hospitalization charges'>
```



In [16]:
```
pd.crosstab(index=df['smoker'],columns=df['age_bin'])
```

Out[16]:

age_bin	18-20	21-30	31-40	41-50	51-60	61-70
smoker						
no	127	222	203	220	223	69
yes	39	56	54	61	42	22

In [17]:
```
pd.crosstab(index=df['smoker'],columns=df['age_bin'],normalize='index')
```

Out[17]:

age_bin	18-20	21-30	31-40	41-50	51-60	61-70
smoker						
no	0.119361	0.208647	0.190789	0.206767	0.209586	0.064850
yes	0.142336	0.204380	0.197080	0.222628	0.153285	0.080292

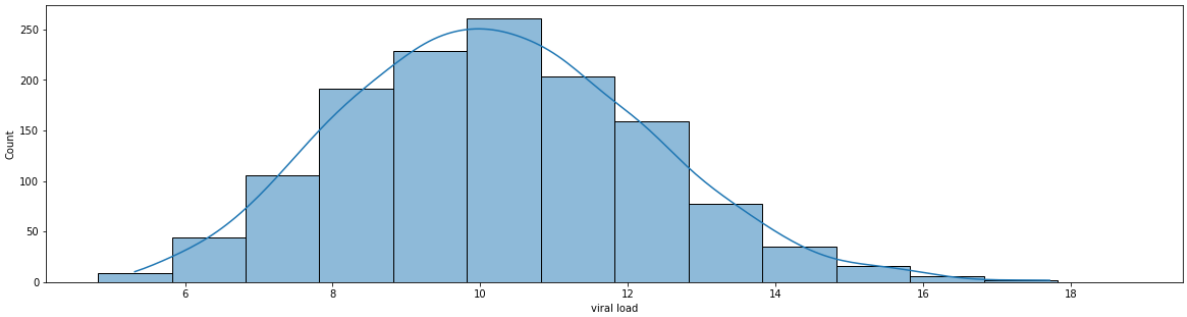
In [18]:
```
pd.crosstab(index=df['smoker'],columns=df['region'],normalize='index')
```

Out[18]:

region	northeast	northwest	southeast	southwest
smoker				
no	0.241541	0.250940	0.256579	0.250940
yes	0.244526	0.211679	0.332117	0.211679

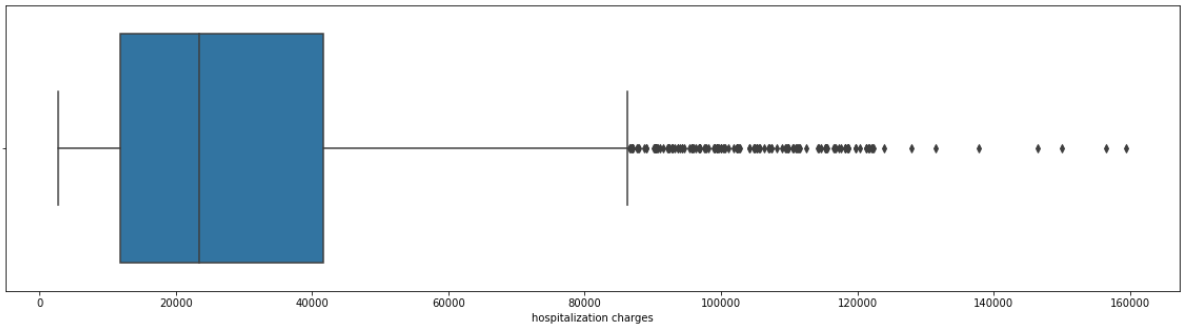
In [19]:
```
fig=plt.figure(figsize=(20,5))
sns.histplot(x='viral load',data=df,discrete=True,kde=True)
```

Out[19]:
```
<AxesSubplot:xlabel='viral load', ylabel='Count'>
```

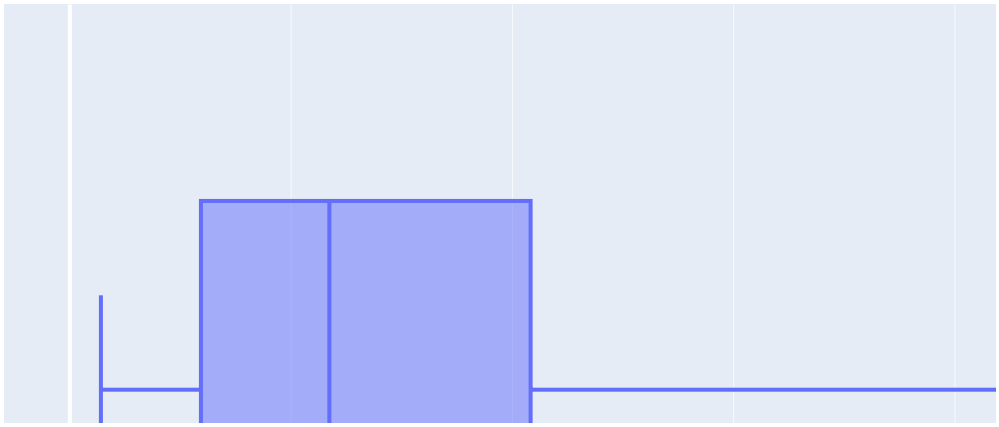


In [20]:
```
fig=plt.figure(figsize=(20,5))
sns.boxplot(x='hospitalization charges',data=df)
```

Out[20]:
```
<AxesSubplot:xlabel='hospitalization charges'>
```



In [21]:
```
fig=plt.figure(figsize=(20,5))
plx.box(x='hospitalization charges',data_frame=df)
```



```
<Figure size 1440x360 with 0 Axes>
```

Outlier data with hospitalization charges

In [22]:
```
df["hospitalization charges"].quantile(0.9)
```

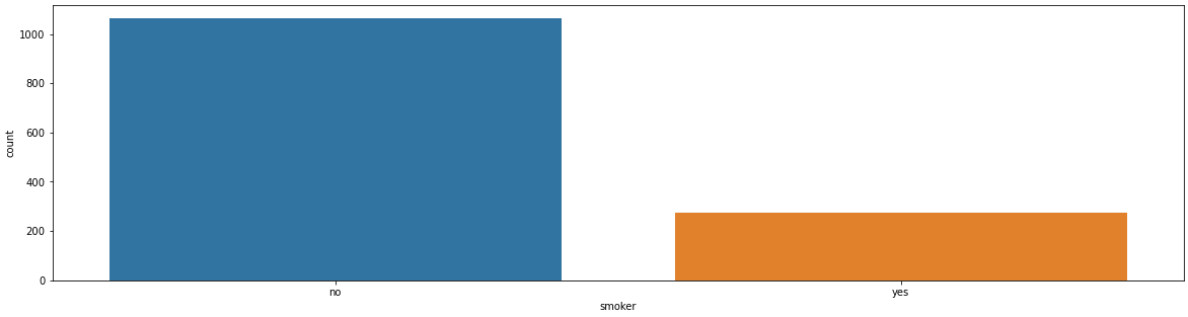
Out[22]:
```
87079.5
```

In [23]:
```
hospital_charges_outlier=df[df["hospitalization charges"] > df["hospitalization cha
print(hospital_charges_outlier.count())
display(hospital_charges_outlier.sort_values())
```

```
134
314      87097
917      87673
476      87869
242      87900
322      88729
         ...
819     137839
577     146428
1230    150053
1300    156482
543     159426
Name: hospitalization charges, Length: 134, dtype: int64
```

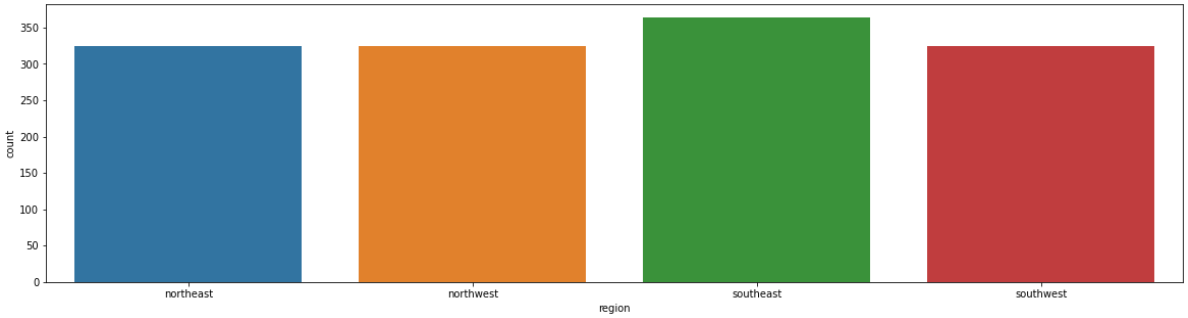
In [24]:
```
fig=plt.figure(figsize=(20,5))
sns.countplot(x='smoker',data=df)
```

Out[24]:
```
<AxesSubplot:xlabel='smoker', ylabel='count'>
```



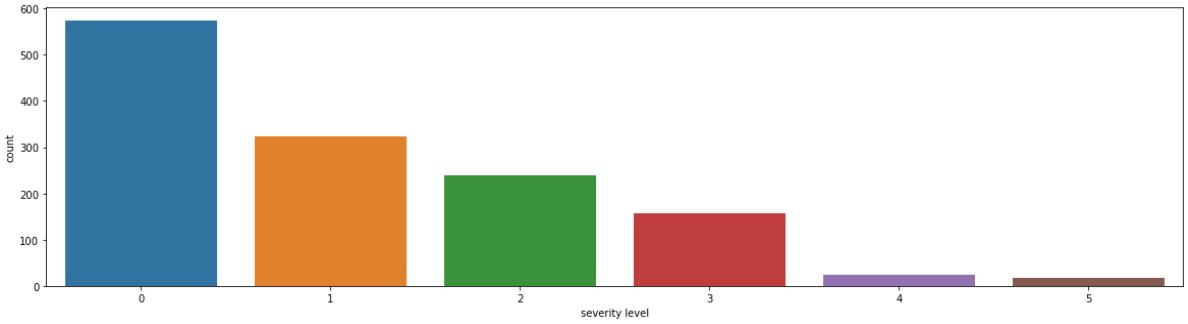
In [25]:
```
fig=plt.figure(figsize=(20,5))
sns.countplot(x='region',data=df)
```

Out[25]:
```
<AxesSubplot:xlabel='region', ylabel='count'>
```

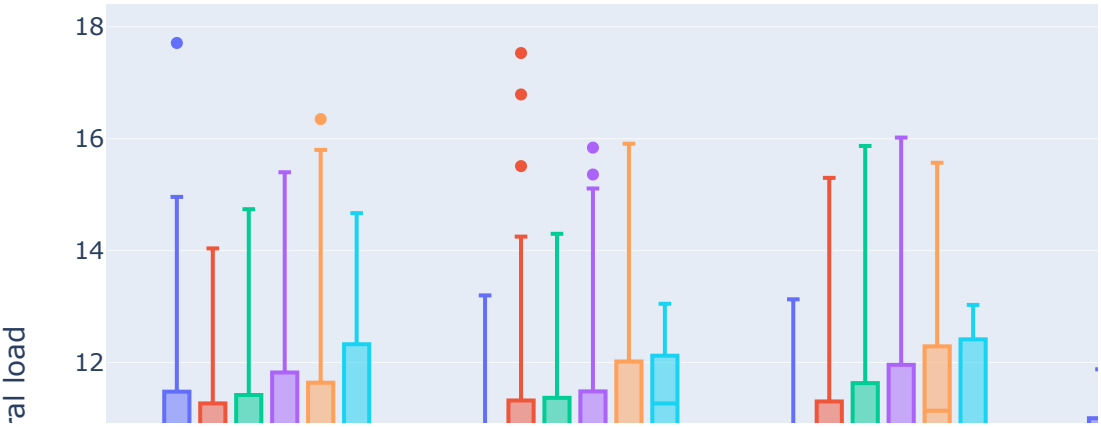


In [26]:
```
fig=plt.figure(figsize=(20,5))
sns.countplot(x='severity level',data=df)
```

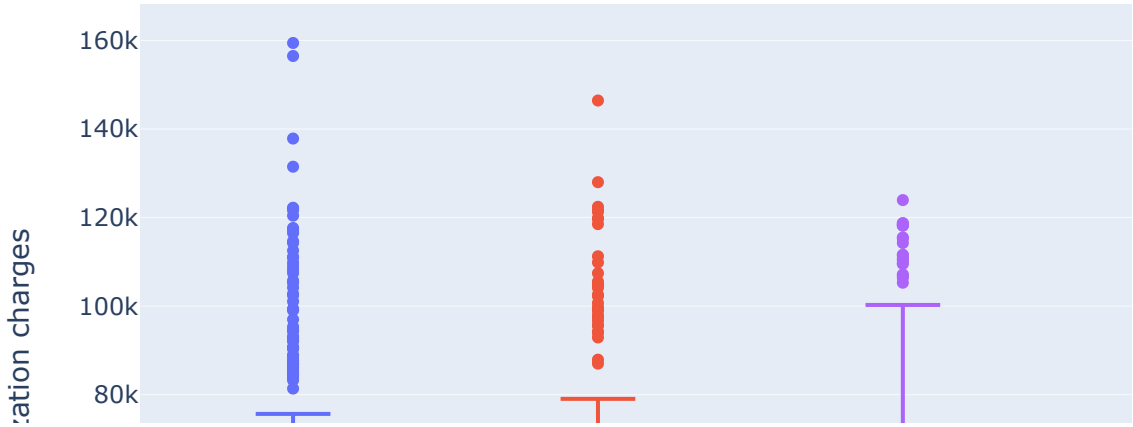
Out[26]:
```
<AxesSubplot:xlabel='severity level', ylabel='count'>
```



In [27]:
```
plx.box(data_frame=df,x='severity level',y='viral load',color='age_bin')
```



In [28]:
```
plx.box(data_frame=df,x='severity level',y='hospitalization charges',color='severi
```

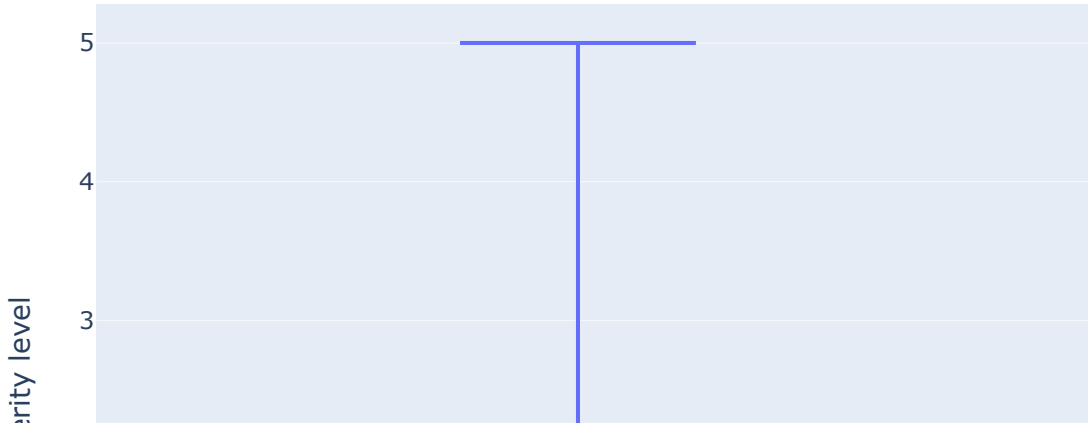


In [29]:
```
pd.crosstab(index=df['severity level'],columns=df['age_bin'],normalize='index')
```

Out[29]:

severity level	18-20	21-30	31-40	41-50	51-60	61-70
0	0.205575	0.226481	0.106272	0.123693	0.231707	0.106272
1	0.080247	0.182099	0.250000	0.283951	0.169753	0.033951
2	0.058333	0.200000	0.254167	0.300000	0.150000	0.037500
3	0.025478	0.210191	0.248408	0.222930	0.235669	0.057325
4	0.040000	0.240000	0.320000	0.240000	0.120000	0.040000
5	0.166667	0.111111	0.388889	0.277778	0.055556	0.000000

In [30]:
```
plx.box(data_frame=df,y='severity level',x='smoker')
```



In [31]:
```
fig=plt.figure(figsize=(20,5))
plx.box(y='hospitalization charges',data_frame=df,x='region',color='region')
```


```
<Figure size 1440x360 with 0 Axes>
```

southeast region appears more skewed

In [32]:
```
fig=plt.figure(figsize=(20,5))
plx.box(y='hospitalization charges',data_frame=df,x='smoker',color='smoker')
```


```
<Figure size 1440x360 with 0 Axes>
```

In [33]:
```
pd.crosstab(index=df['sex'],columns=df['age_bin'],normalize='index')
```

Out[33]:

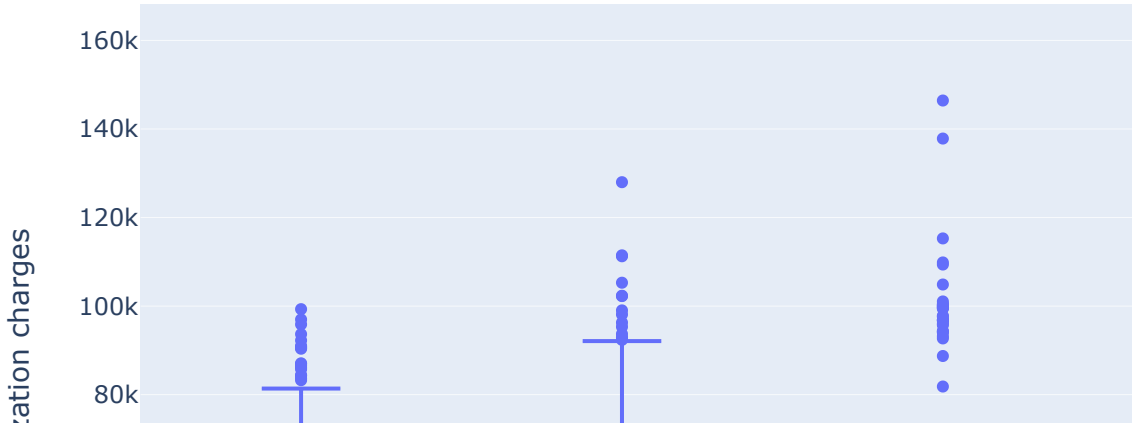
sex	18-20	21-30	31-40	41-50	51-60	61-70
female	0.120846	0.202417	0.191843	0.21148	0.202417	0.070997
male	0.127219	0.213018	0.192308	0.20858	0.193787	0.065089

In [34]:
```
pd.crosstab(index=df['sex'],columns=df['age_bin'],normalize='columns')
```

Out[34]:

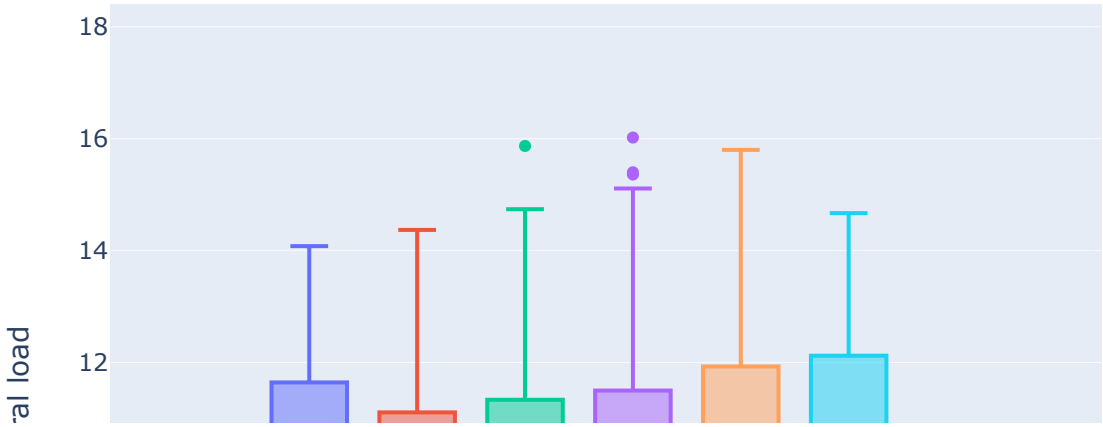
sex	18-20	21-30	31-40	41-50	51-60	61-70
female	0.481928	0.482014	0.494163	0.498221	0.50566	0.516484
male	0.518072	0.517986	0.505837	0.501779	0.49434	0.483516

In [35]:
```
fig=plt.figure(figsize=(20,5))
plx.box(y='hospitalization charges',data_frame=df,x='age_bin')
```



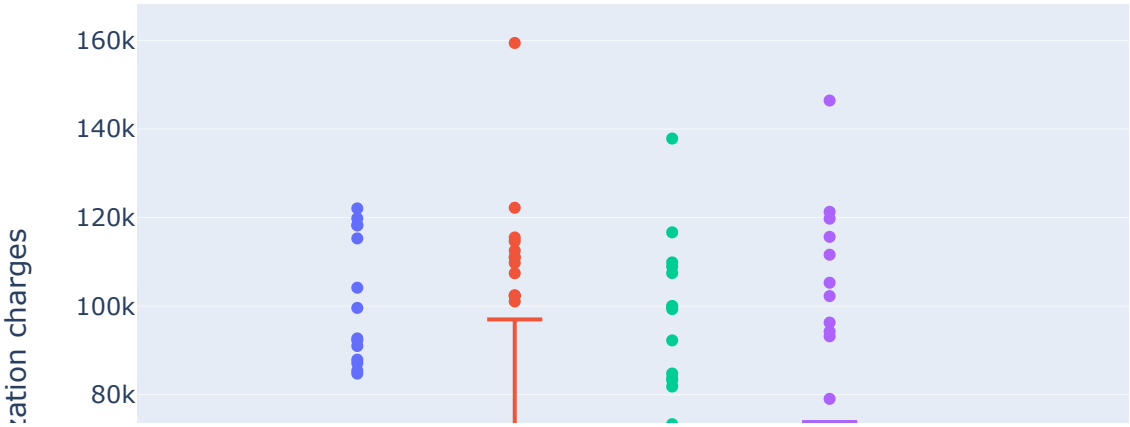

```
<Figure size 1440x360 with 0 Axes>
```

In [36]:
```
fig=plt.figure(figsize=(20,5))
plx.box(y='viral load',data_frame=df,x='sex',color='age_bin')
```




```
<Figure size 1440x360 with 0 Axes>
```

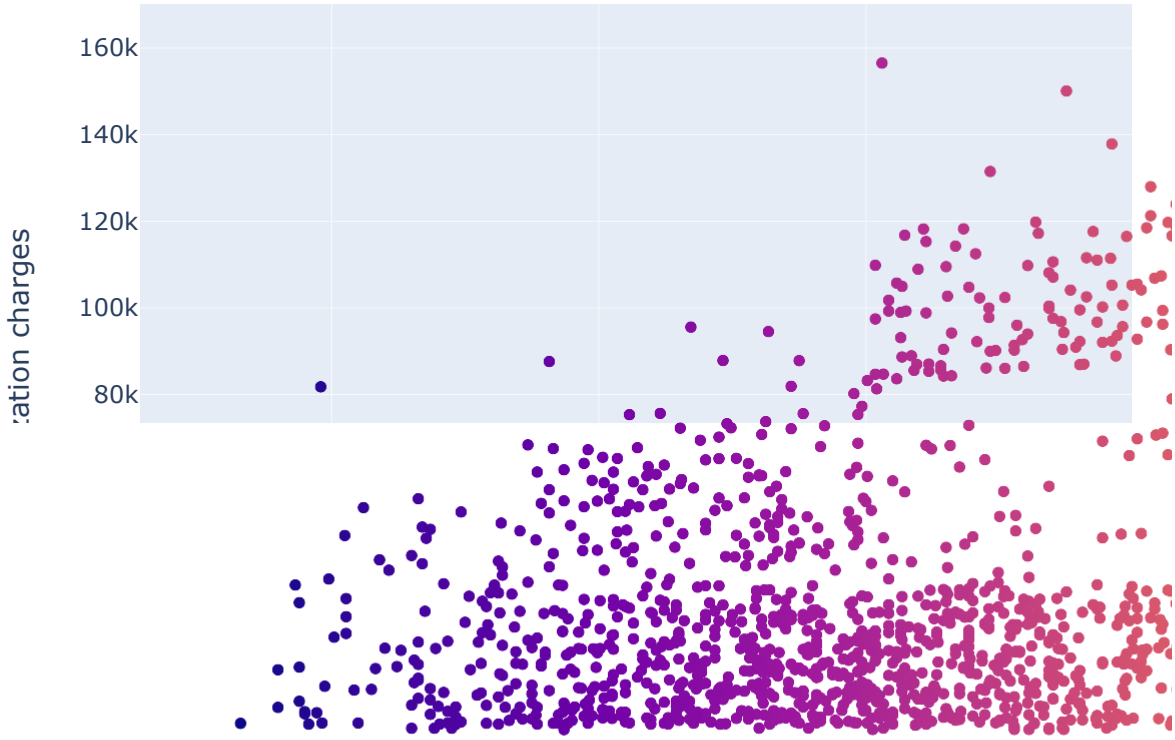
In [37]:
```
fig=plt.figure(figsize=(20,5))
plx.box(y='hospitalization charges',data_frame=df,x='sex',color='region')
```




```
<Figure size 1440x360 with 0 Axes>
```

Clearly smokers had higher hospitalization charge

In [38]:
```
fig=plt.figure(figsize=(20,5))
plx.scatter(y='hospitalization charges',data_frame=df,x='viral load',color='viral
```




```
<Figure size 1440x360 with 0 Axes>
```

Hypothesis testing

Hospitalization charges relation to smoking

H0 : Data is gaussian

H1 : Data is not gaussian

In [39]:
```
from scipy.stats import shapiro
from scipy.stats import normaltest
from scipy.stats import mannwhitneyu
```

In [40]:
```
smokers=df[df["smoker"]=="yes"]["hospitalization charges"]
non_smokers=df[df["smoker"]=="no"]["hospitalization charges"]
print(smokers.count(),non_smokers.count())
```

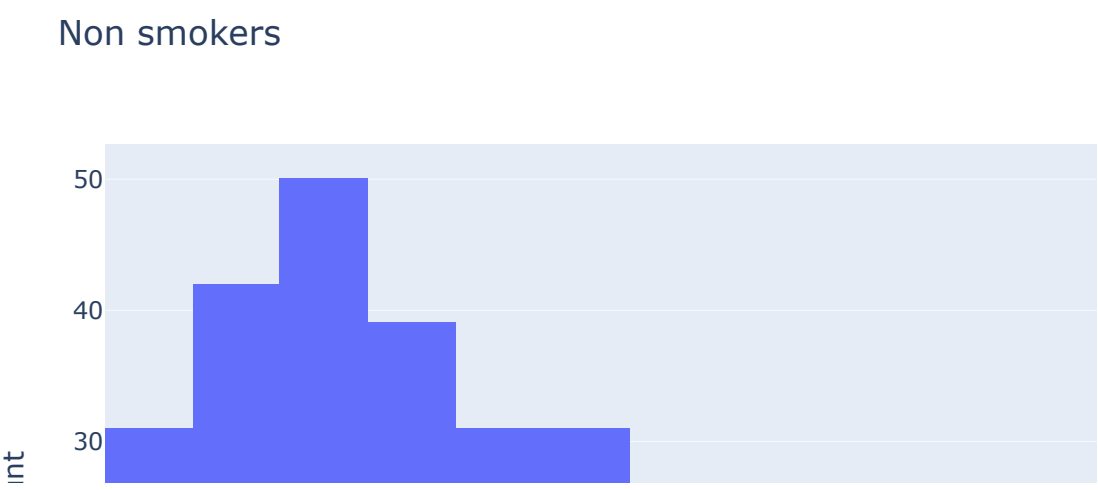
```
274 1064
```

In [41]:
```
non_smokers_final=non_smokers.sample(274)
```

In [42]:
```
display(smokers.count())
display(non_smokers_final.count())
```

```
274
274
```

In [43]:
```
plx.histogram(data_frame=non_smokers_final,title="Non smokers")
```



Non smokers data looks bimodal gaussian

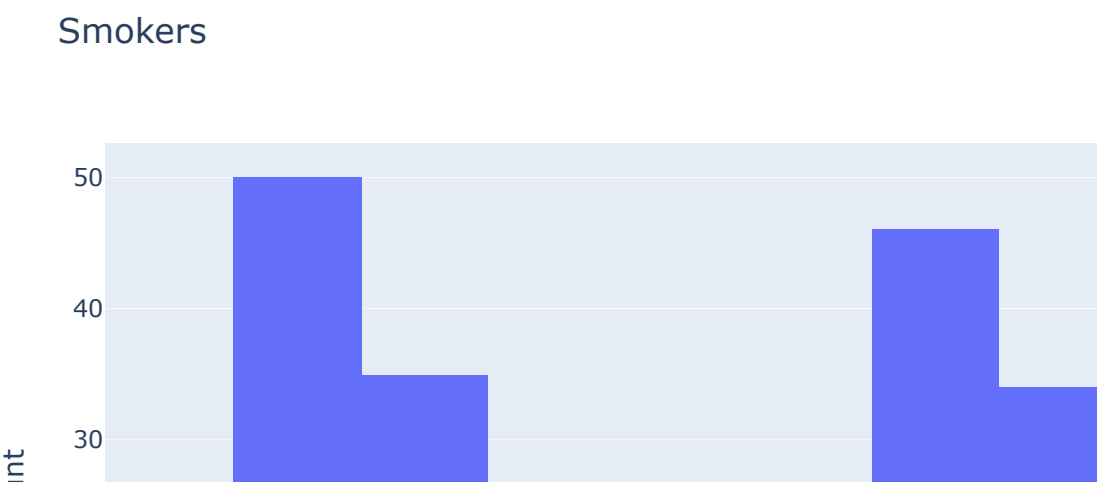
In [44]:
```
def test_for_gaussian(ip_df):
    stat1,p_value1=shapiro(ip_df)
    stat2,p_value2=normaltest(ip_df)
    type_of_data="Gaussian"
    print(p_value1,p_value2)

    if p_value1 > 0.05 and p_value2 > 0.05:
        print("Data looks guassian")
    elif ( p_value1 < 0.05 and p_value2 > 0.05 ) or ( p_value1 > 0.05 and p_value2
        print("Data is gaussian like but not exactly gaussian")
    elif p_value1 < 0.05 and p_value2 < 0.05:
        print("Data is not guassian")
        type_of_data="Gaussian"
    return type_of_data
```

In [45]:
```
non_smokers_type=test_for_gaussian(non_smokers_final)
```

```
2.4437178460972383e-16 1.2311967519842144e-24
Data is not guassian
```

Non smokers data --> NOT gaussian

In [47]:
```
plx.histogram(data_frame=smokers,title="Smokers")
```



In [48]:
```
smokers_type=test_for_gaussian(smokers)
```

```
3.6248792856241607e-09 5.560432703059049e-14
Data is not guassian
```

Smokers data --> NOT Gaussian

Manwhitneyu non parametric test

H0: Smokers hospitalization charge mean = non smokers hospitalization charge mean

H1: Smokers > Non smokers

In [49]:
```
stat,p_value=mannwhitneyu(smokers,non_smokers,alternative="greater")
print(stat,p_value)
```

```
284132.5 2.6407031043303346e-130
```

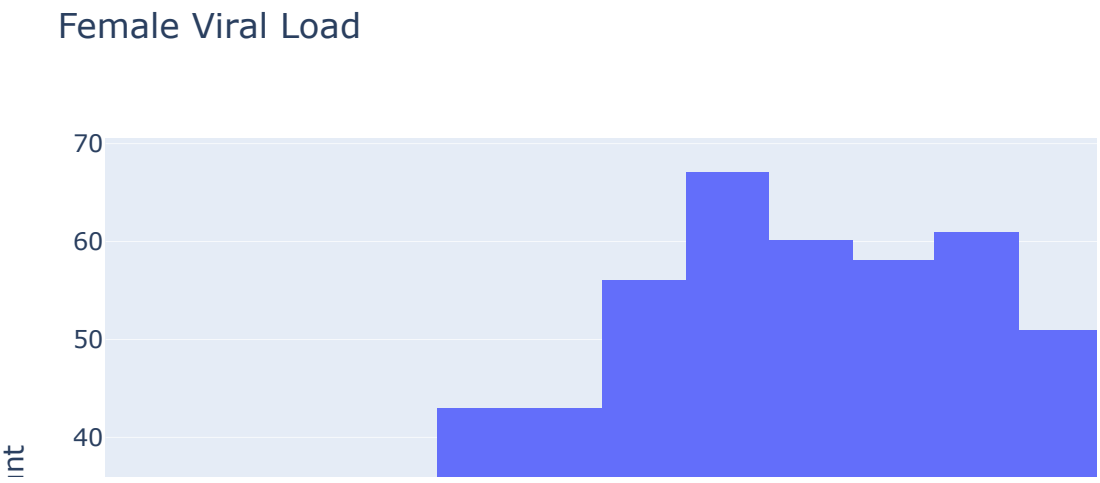
Reject Null Hypothesis

Result : Smokers hospitalization charge > Non smokers

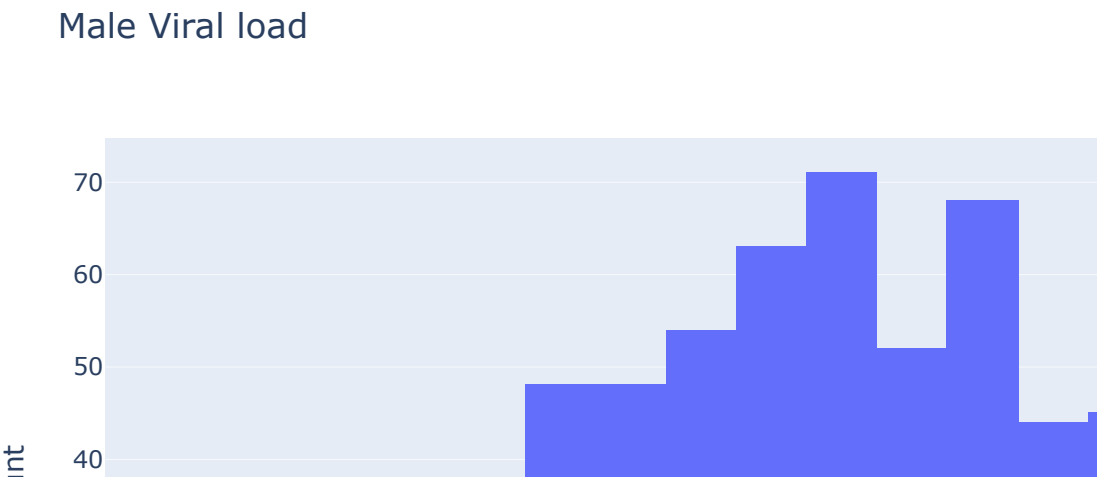
In [67]:
```
female=df[df["sex"]=="female"]
male=df[df["sex"]=="male"].sample(662)
```

In [52]:
```
female_vload=female["viral load"]
male_vload=male["viral load"]
```

In [53]:
```
plx.histogram(data_frame=female_vload,title="Female Viral Load")
```



In [54]:
```
plx.histogram(data_frame=male_vload,title="Male Viral load")
```



In [55]:
```
female_vload_type=test_for_gaussian(female_vload)
male_vload_type=test_for_gaussian(male_vload)
```

```
0.003624602919444442 0.013092448927449781
Data is not guassian
0.016298236325383186 0.01659416547079123
Data is not guassian
```

female viral load and male viral load data is not gaussian

H0: two distributions are equal

H1: two distributions are not equal

In [56]:
```
stat,p_value=mannwhitneyu(female_vload,male_vload,alternative="two-sided")
print(stat,p_value)
```

```
208141.5 0.11446268970467466
```

Fail to Reject Null Hypothesis

Result : Female viral load and male viral load are equal

In [57]:
```
from scipy.stats import chi2_contingency
```

H0: Smoker and region is independent

H1: Smoker and region are dependent

In [58]:
```
smoker_region_contigency_table=pd.crosstab(index=df["smoker"],columns=df["region"]
smoker_region_contigency_table
```

Out[58]:

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

In [59]:
```
chi2_contingency(smoker_region_contigency_table)
```

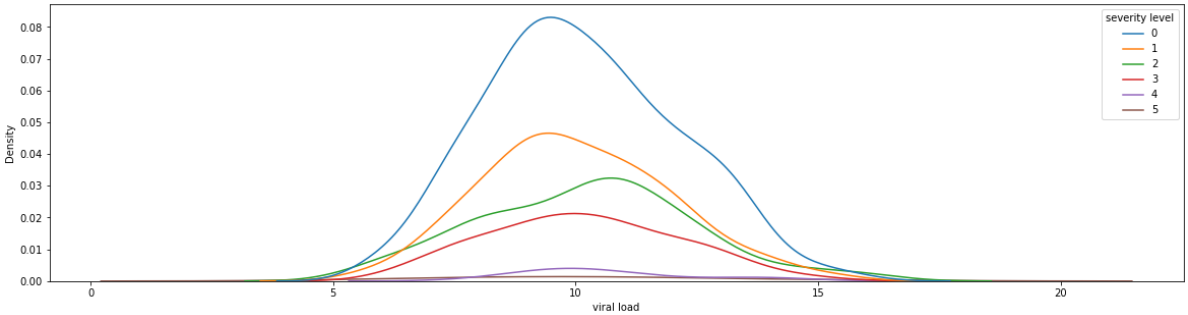
Out[59]:
```
(7.34347776140707,
 0.06171954839170547,
 3,
 array([[257.65022422, 258.44544096, 289.45889387, 258.44544096],
        [ 66.34977578,  66.55455904,  74.54110613,  66.55455904]]))
```

Fail to Reject Null Hypothesis

Result : Smoker and region are independent. Propotion of smoking is not different across different regions

In [60]:
```
viral_load_female=df[df["sex"]=="female"]
fig=plt.figure(figsize=(20,5))
sns.kdeplot(data=viral_load_female,x="viral load",hue="severity level")
```

Out[60]:
```
<AxesSubplot:xlabel='viral load', ylabel='Density'>
```



The severity levels 0,1,2 are looking approximately gaussian. we are good to go ahead with ANOVA

In [61]:
```
from scipy.stats import levene
from scipy.stats import f_oneway
```

In [66]:
```
sev0=viral_load_female[viral_load_female["severity level"]==0]["viral load"]
sev1=viral_load_female[viral_load_female["severity level"]==1]["viral load"]
sev2=viral_load_female[viral_load_female["severity level"]==2]["viral load"]
```

In [63]:
```
levene(sev0,sev1,sev2)
```

Out[63]:
```
LeveneResult(statistic=0.9435131022565071, pvalue=0.38987253596513605)
```

Difference in variance across the 3 groups is not significant. Fail to reject null hypothesis

H0: All 3 severity levels have same mean of viral load

H1: 3 Severity levels have different mean of viral load

In [64]:
```
f_oneway(sev0,sev1,sev2)
```

Out[64]:
```
F_onewayResult(statistic=0.3355061434584082, pvalue=0.7151189650367746)
```

Fail to reject Null hypothesis. All 3 severity levels have same mean of viral load

Business Insights

1. Hospitalization charges is clearly dependent on Age and smoking factor

2. South east has highest percentage of smokers which also echoes in the hospitalization charges

3. Smokers are more prevalent in the age group of 21-50

4. Non smokers vs smokers propotion is skewed. 75% of the data is from non smokers

Recommendations

1. Smokers can be urged to buy health insurance and the cost of policy can be increased citing the risk factor

2. Severity level 2 & 3 has the highest hospitalization expense and these 2 levels are more prominent in 31-50 age group. People in this age require more attention.

3. South east people might need special packages from government to meet the higher hospitalization expense