Importing Libraries

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import math as m
import numpy as np
```

Reading dataset

In [3]:
```
df=pd.read_csv("C:/Users/Hp/Downloads/insurance.csv")
```

In [183]:
```
pd.options.mode.chained_assignment = None
```

Understanding Dataset

The insurance dataset typically refers to a dataset containing information about individuals and their health insurance-related attributes. It's a common dataset used for analysis

In [199]:
```
df
```

Out[199]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	NaN	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 7 columns

Data types of coloumn

In [56]:
```
df.dtypes
```

Out[56]:
```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

In [57]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1332 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Number of rows and column

In [58]:
```
df.shape
```

Out[58]:
```
(1338, 7)
```

In [59]:
```
df.head()
```

Out[59]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	NaN	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

In [60]:
```
df.tail()
```

Out[60]:

	age	sex	bmi	children	smoker	region	charges
1333	50	male	30.97	3	no	northwest	10600.5483
1334	18	female	31.92	0	no	northeast	2205.9808
1335	18	female	36.85	0	no	southeast	1629.8335
1336	21	female	25.80	0	no	southwest	2007.9450
1337	61	female	29.07	0	yes	northwest	29141.3603

In [61]:
```
df.sample()
```

Out[61]:

	age	sex	bmi	children	smoker	region	charges
1115	55	male	32.67	1	no	southeast	10807.4863

Cleaning Dataset

In [62]:
```
df.dtypes
```

Out[62]:
```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

In [63]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1332 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

In [64]:
```
df.isnull()
```

Out[64]:

	age	sex	bmi	children	smoker	region	charges
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	True	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...
1333	False	False	False	False	False	False	False
1334	False	False	False	False	False	False	False
1335	False	False	False	False	False	False	False
1336	False	False	False	False	False	False	False
1337	False	False	False	False	False	False	False

1338 rows × 7 columns

lets check the null values in dataset

In [70]:
```
df.isnull().sum()
```

Out[70]:
```
age         0
sex         0
bmi         6
children    0
smoker      0
region      0
charges     0
dtype: int64
```

checking the count of values in children column

In [71]:
```
df.bmi.value_counts()
```

Out[71]:
```
32.300    13
28.310     9
30.875     8
28.880     8
30.800     8
          ..
46.700     1
46.200     1
23.800     1
44.770     1
30.970     1
Name: bmi, Length: 548, dtype: int64
```

In [72]:
```
len(df.bmi.unique())
```

Out[72]:
```
549
```

In [73]:
```
value = df.bmi.mean()

print(value)
```

```
30.665195195195196
```

In [74]:
```
df.fillna(value,inplace=True)
```

here we remove the null value of column 'bmi' with mean value of columns

In [75]:
```
df.isnull().sum()
```

Out[75]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

lets check the bmi column

In [86]:
```
df.iloc[::,2:3:]
```

Out[86]:

	bmi
0	27.900000
1	33.770000
2	30.665195
3	22.705000
4	28.880000
...	...
1333	30.970000
1334	31.920000
1335	36.850000
1336	25.800000
1337	29.070000

1338 rows × 1 columns

understanding columns

Age: It contains distribution of age.

Sex: Gender (male or female).

BMI (Body Mass Index): A measure of body fat based on height and weight.

Children: Number of children covered by the insurance.

Smoker: Number of individual is a smoker or non-smoker.

Region: The geographical region of individual (southeast, southwest, northeast, northwest).

Charges: The insurance cost by the individual.

Categorical Data :

sex

smoker

region

Quantitative Data :

age

bmi

children

charges

(1) sex

In [91]:
```
df.sex.unique()
```

Out[91]:
```
array(['female', 'male'], dtype=object)
```

In [92]:
```
df.sex.value_counts()
```

Out[92]:
```
male      676
female    662
Name: sex, dtype: int64
```

(2) age

In [99]:
```
df.age.unique()
```

Out[99]:
```
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], dtype=int64)
```

In [100]:
```
df.age.value_counts()
```

Out[100]:
```
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
```

In [101]:
```
df.age.describe()
```

Out[101]:
```
count    1338.000000
mean       39.207025
std        14.049960
min        18.000000
25%        27.000000
50%        39.000000
75%        51.000000
max        64.000000
Name: age, dtype: float64
```

In [277]:
```
df['age'].aggregate(['max'])
```

Out[277]:
```
max    64
Name: age, dtype: int64
```

In [279]:
```
df['age'].aggregate(['min'])
```

Out[279]:
```
min    18
Name: age, dtype: int64
```

(3) bmi

In [106]:
```
df.bmi.unique()
```

Out[106]:
```
array([27.9      , 33.77     , 30.6651952, 22.705    , 28.88     ,
       25.74     , 33.44     , 27.74     , 29.83     , 25.84     ,
       26.22     , 26.29     , 34.4      , 39.82     , 42.13     ,
       24.6      , 30.78     , 23.845    , 40.3      , 36.005    ,
       32.4      , 34.1      , 31.92     , 28.025    , 27.72     ,
       23.085    , 32.775    , 17.385    , 36.3      , 35.6      ,
       26.315    , 28.6      , 28.31     , 36.4      , 20.425    ,
       32.965    , 20.8      , 36.67     , 39.9      , 26.6      ,
       36.63     , 21.78     , 30.8      , 37.05     , 37.3      ,
       38.665    , 34.77     , 24.53     , 35.2      , 35.625    ,
       33.63     , 28.       , 34.43     , 28.69     , 36.955    ,
       31.825    , 31.68     , 22.88     , 37.335    , 27.36     ,
       33.66     , 24.7      , 25.935    , 22.42     , 28.9      ,
       39.1      , 36.19     , 23.98     , 24.75     , 28.5      ,
       28.1      , 32.01     , 27.4      , 34.01     , 29.59     ,
       35.53     , 39.805    , 26.885    , 38.285    , 37.62     ,
       41.23     , 34.8      , 22.895    , 31.16     , 27.2      ,
       26.98     , 39.49     , 24.795    , 31.3      , 38.28     ,
       19.95     , 19.3      , 31.6      , 25.46     , 30.115    ,
```

In [107]:
```
df.bmi.value_counts()
```

Out[107]:
```
32.300    13
28.310     9
34.100     8
28.880     8
30.875     8
          ..
46.200     1
23.800     1
44.770     1
32.120     1
30.970     1
Name: bmi, Length: 549, dtype: int64
```

In [109]:
```
df['bmi'].aggregate(['max'])
```

Out[109]:
```
max    53.13
Name: bmi, dtype: float64
```

In [111]:
```
df['bmi'].aggregate(['min'])
```

Out[111]:
```
min    15.96
Name: bmi, dtype: float64
```

In [112]:
```
df['bmi'].aggregate(['mean'])
```

Out[112]:
```
mean    30.665195
Name: bmi, dtype: float64
```

(4) smoker

In [102]:
```
df.smoker.unique()
```

Out[102]:
```
array(['yes', 'no'], dtype=object)
```

In [103]:
```
df.smoker.value_counts()
```

Out[103]:
```
no     1064
yes     274
Name: smoker, dtype: int64
```

(4) region

In [104]:
```
df.region.unique()
```

Out[104]:
```
array(['southwest', 'southeast', 'northwest', 'northeast'], dtype=object)
```

In [105]:
```
df.region.value_counts()
```

Out[105]:
```
southeast    364
southwest    325
northwest    325
northeast    324
Name: region, dtype: int64
```

(5) charges

In [113]:
```
df.charges.unique()
```

Out[113]:
```
array([16884.924 ,  1725.5523,  4449.462 , ...,  1629.8335,  2007.945 ,
       29141.3603])
```

In [114]:
```
df.charges.value_counts()
```

Out[114]:
```
1639.56310     2
16884.92400    1
29330.98315    1
2221.56445     1
19798.05455    1
              ..
7345.08400     1
26109.32905    1
28287.89766    1
1149.39590     1
29141.36030    1
Name: charges, Length: 1337, dtype: int64
```

In [115]:
```
df['charges'].aggregate(['max'])
```

Out[115]:
```
max    63770.42801
Name: charges, dtype: float64
```

In [116]:
```
df['charges'].aggregate(['min'])
```

Out[116]:
```
min    1121.8739
Name: charges, dtype: float64
```

In [117]:
```
df['charges'].aggregate(['mean'])
```

Out[117]:
```
mean    13271.008266
Name: charges, dtype: float64
```

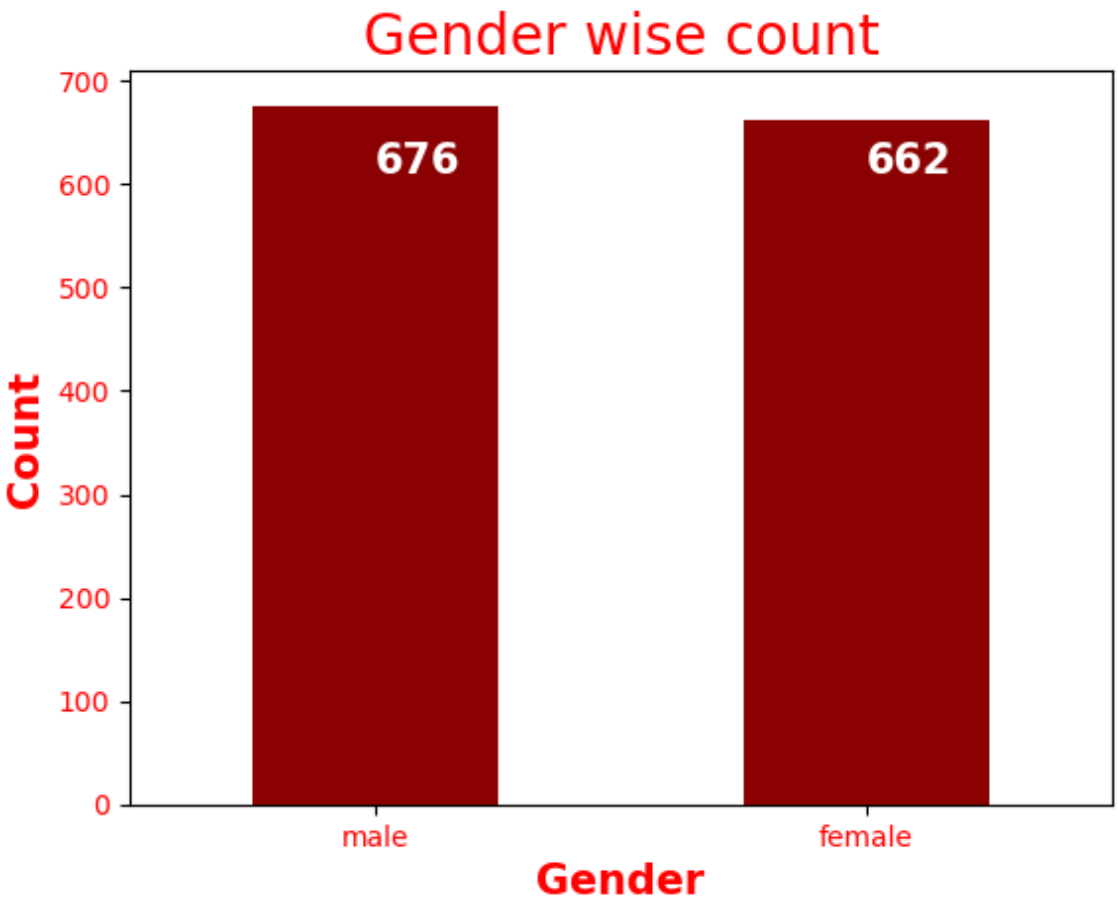
UNIVARIATE ANALYSIS

categorical data

In [4]:
```
df.info()
```

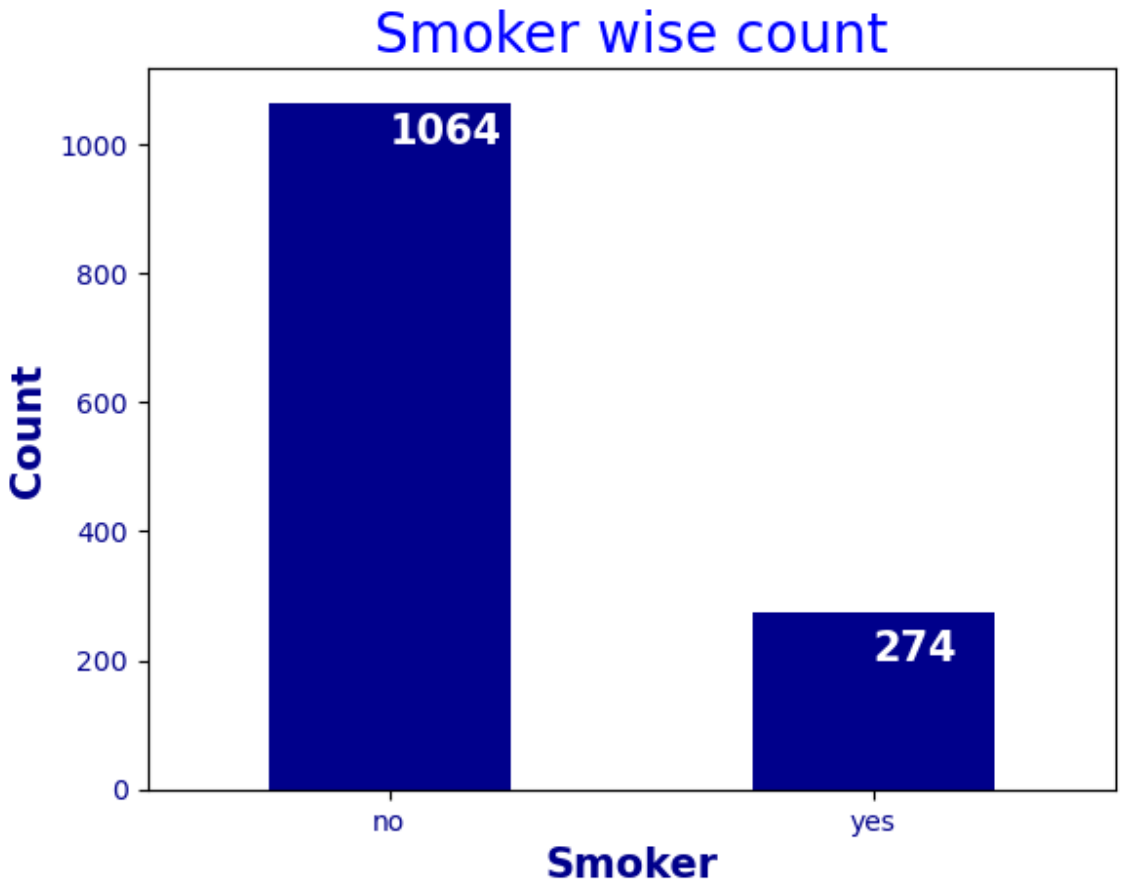
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1332 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

In [210]:
```
df['sex'].value_counts().plot(kind='bar',color=(['darkred','darkred']))

plt.title('Gender wise count',size=20,c='red')

plt.xlabel('Gender',c='red',size=15,fontweight='bold')
plt.ylabel('Count',c='red',size=15,fontweight='bold')

plt.text(0,610,'676',color='white',size=15,fontweight='bold')
plt.text(1,610,'662',color='white',size=15,fontweight='bold')

plt.xticks(rotation='horizontal',color='red',fontsize=10)
plt.yticks(color='red',fontsize=10)

plt.show()
```



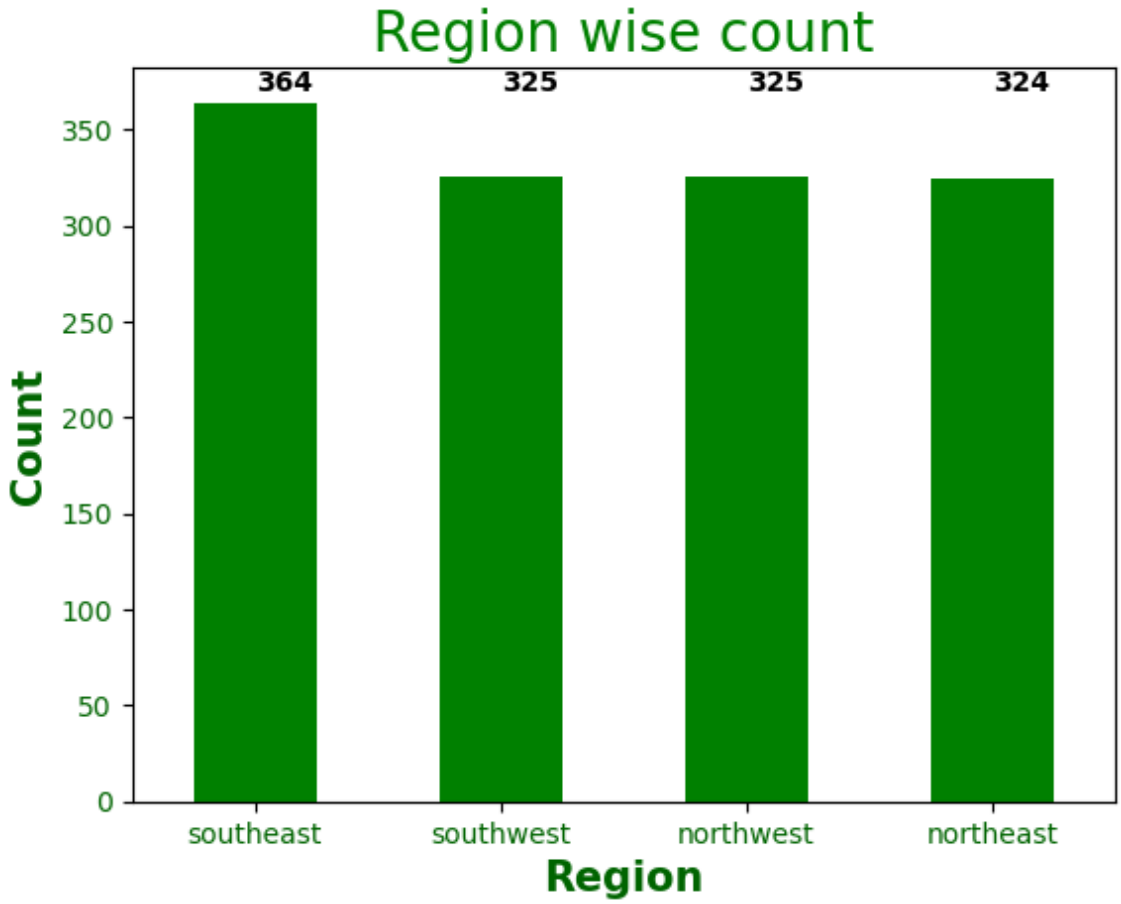
Here we can observe number of male are more

In [204]:
```
df['smoker'].value_counts().plot(kind='bar',color='darkblue')

plt.title('Smoker wise count',size=20,color='blue')

plt.xticks(rotation='horizontal',color='darkblue')
plt.yticks(color='darkblue')

plt.xlabel('Smoker',c='darkblue',size=15,fontweight='bold')
plt.ylabel('Count',c='darkblue',size=15,fontweight='bold')

plt.text(0,1000,'1064',color='white',size=15,fontweight='bold')
plt.text(1,200,'274',color='white',size=15,fontweight='bold')

plt.show()
```

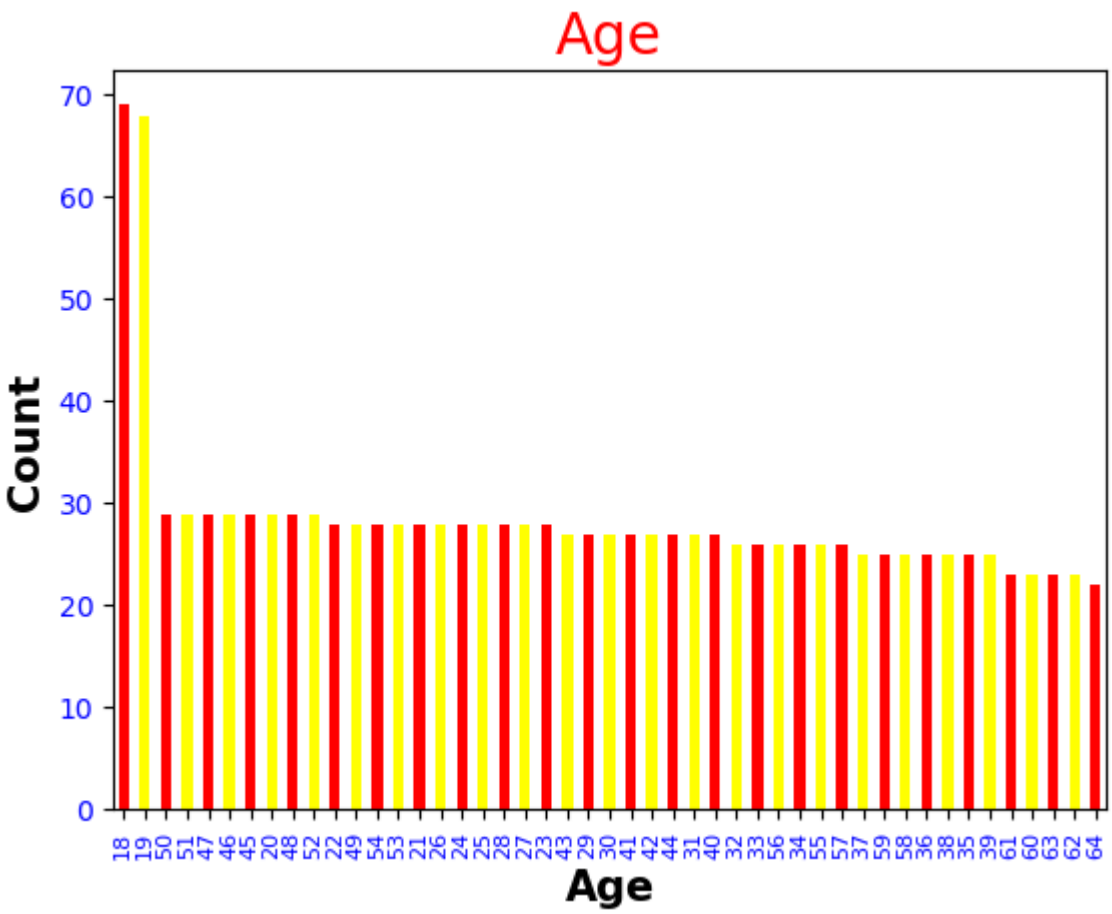


Here we can observe number of no's are more

In [208]:
```
df['region'].value_counts().plot(kind='bar',color='green')

plt.title('Region wise count',size=20,color='green')

plt.xticks(rotation='horizontal',color='darkgreen')
plt.yticks(color='darkgreen')

plt.xlabel('Region',c='darkgreen',size=15,fontweight='bold')
plt.ylabel('Count',c='darkgreen',size=15,fontweight='bold')

plt.text(0,370,'364',color='black',size=10,fontweight='bold')
plt.text(1,370,'325',color='black',size=10,fontweight='bold')
plt.text(2,370,'325',color='black',size=10,fontweight='bold')
plt.text(3,370,'324',color='black',size=10,fontweight='bold')

plt.show()
```



In [244]:
```
df['age'].value_counts().plot(kind='bar',color=['red','yellow'])

plt.xticks(color='blue',size=8)
plt.yticks(color='blue')

plt.title('Age',size=20,c='red')

plt.xlabel('Age',size=15,fontweight='bold')
plt.ylabel('Count',size=15,fontweight='bold')

plt.show()
```



Observations of univariate Analysis

1.In gender column number of male are more.

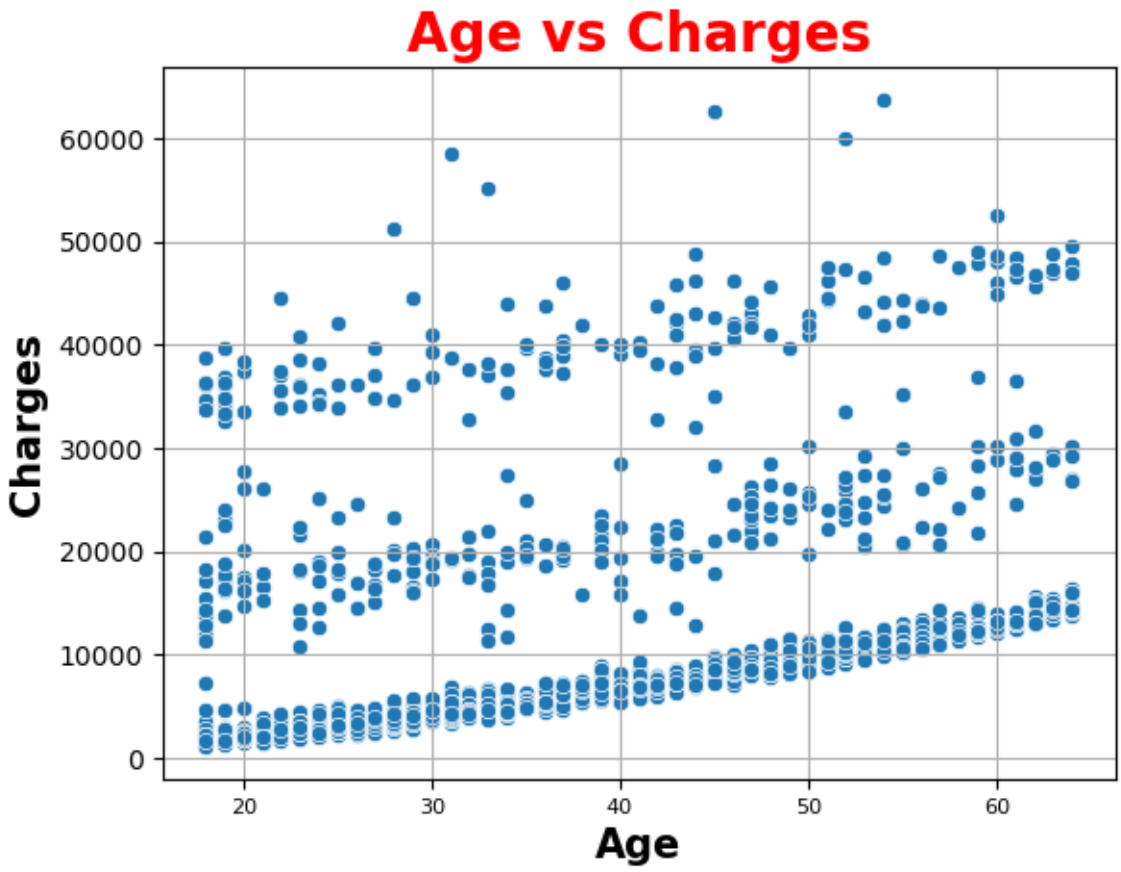
2.number of non-smokers is more .

3.people are more from southeast.

Bivariate Analysis

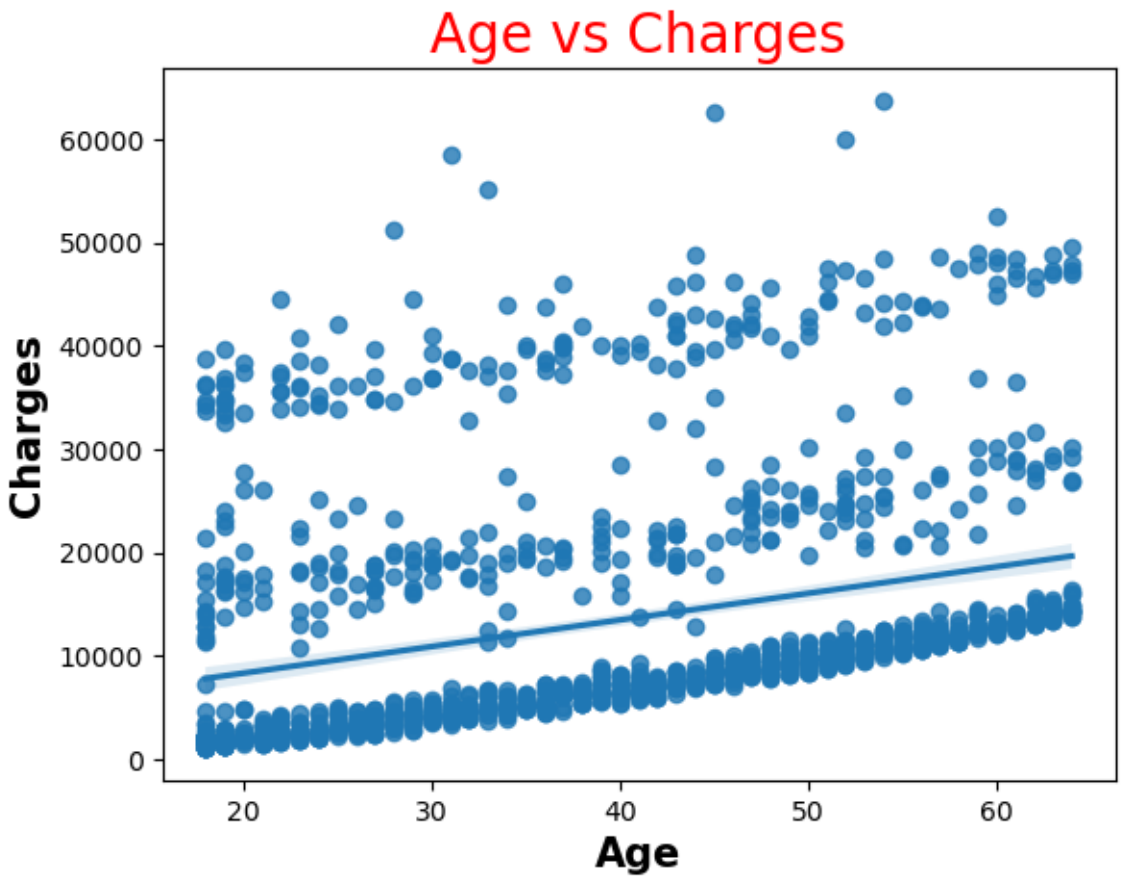
Reletion between the age and the charges

In [246]:
```
sns.scatterplot(data=df,y='charges',x='age')

plt.xticks(color='k',size=8)
plt.yticks(color='k')

plt.title('Age vs Charges',size=20,c='red',fontweight='bold')

plt.xlabel('Age',size=15,fontweight='bold')
plt.ylabel('Charges',size=15,fontweight='bold')

plt.grid()

plt.show()
```



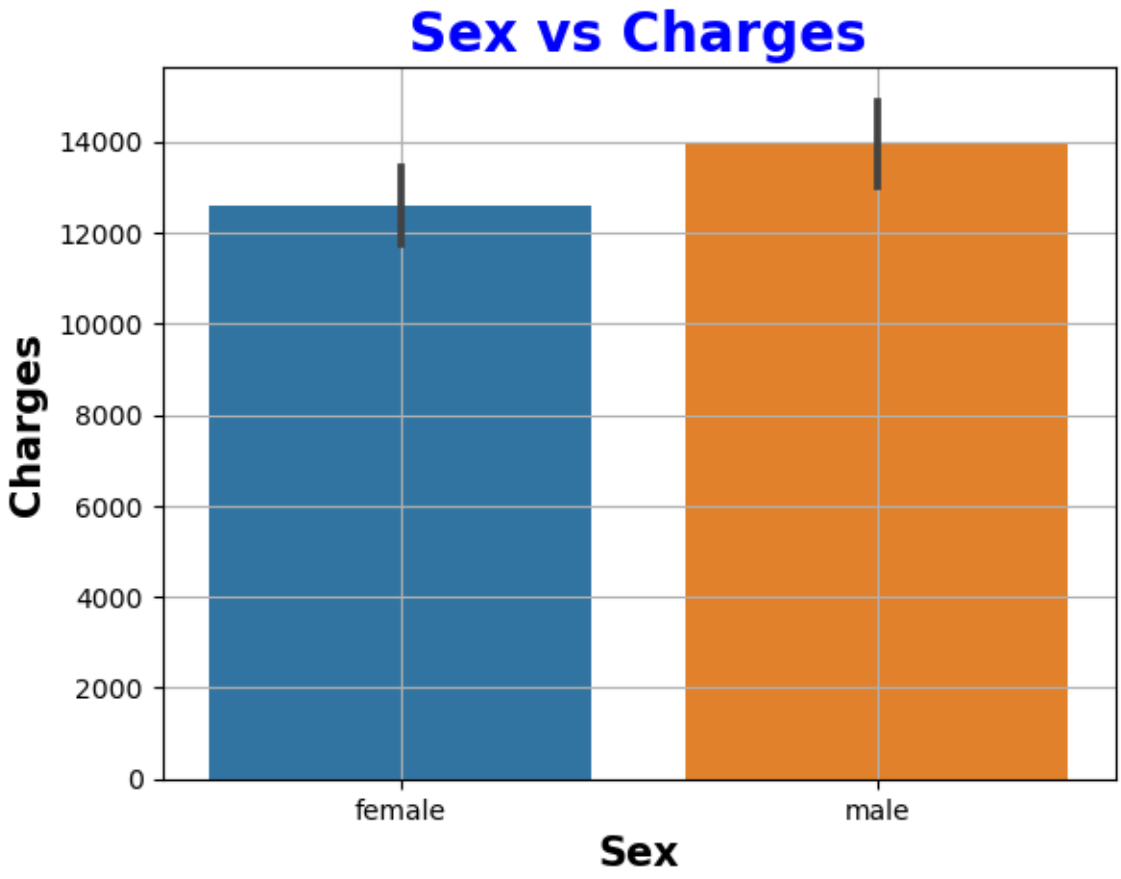
Based on the data we observe that there is a positive correlation between age and charges as age increases the charges also increase.

In [253]:
```
sns.regplot(data=df,y='charges',x='age')

plt.title('Age vs Charges',size=20,c='red')

plt.xlabel('Age',size=15,fontweight='bold')
plt.ylabel('Charges',size=15,fontweight='bold')

plt.show()
```



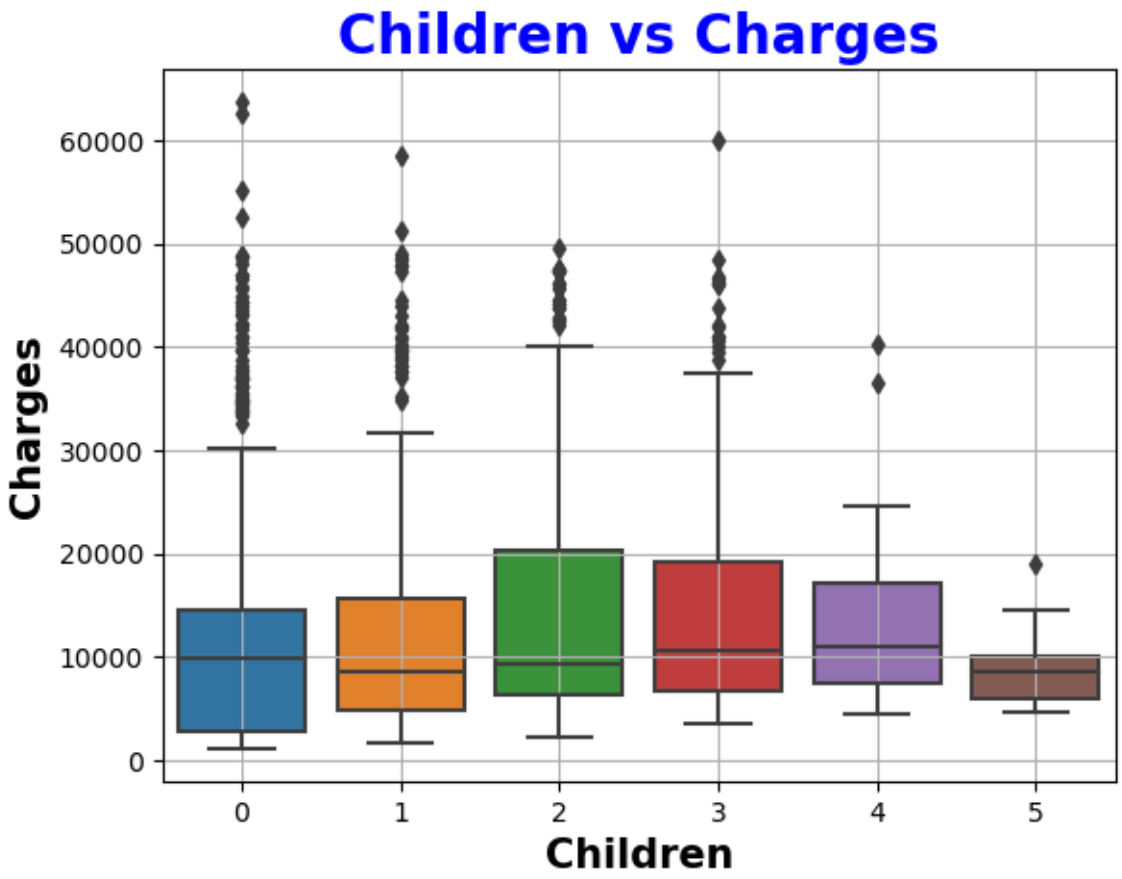
Sex vs Charges

In [257]:
```
sns.barplot(data=df,y='charges',x='sex')

plt.xticks(color='k')
plt.yticks(color='k')

plt.title('Sex vs Charges',size=20,c='Blue',fontweight='bold')

plt.xlabel('Sex',size=15,fontweight='bold')
plt.ylabel('Charges',size=15,fontweight='bold')

plt.grid()

plt.show()
```



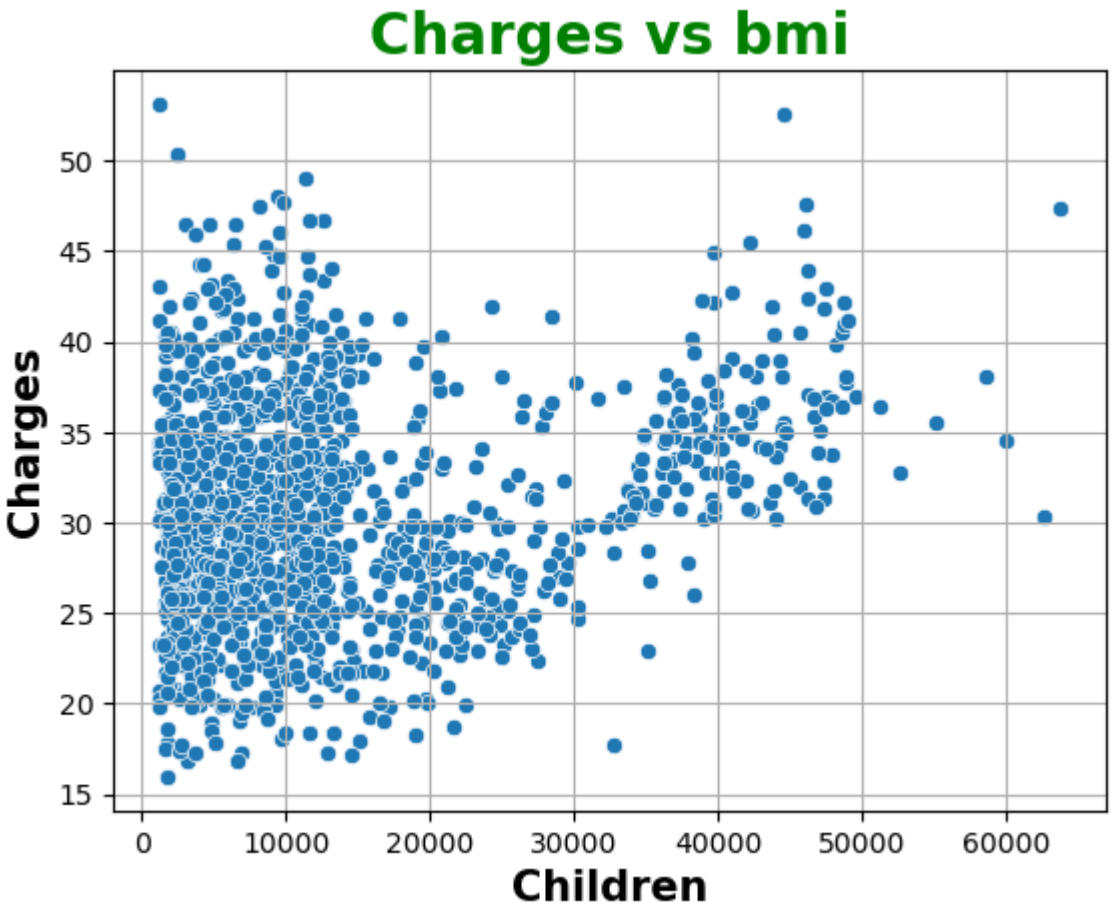
Based on the data we observe that there is correlation between sex and charges as count of male is more

In [258]:
```
sns.boxplot(data=df,y='charges',x='children')

plt.xticks(color='k')
plt.yticks(color='k')

plt.title('Children vs Charges',size=20,c='Blue',fontweight='bold')

plt.xlabel('Children',size=15,fontweight='bold')
plt.ylabel('Charges',size=15,fontweight='bold')

plt.grid()

plt.show()
```



Chrages vs Bmi

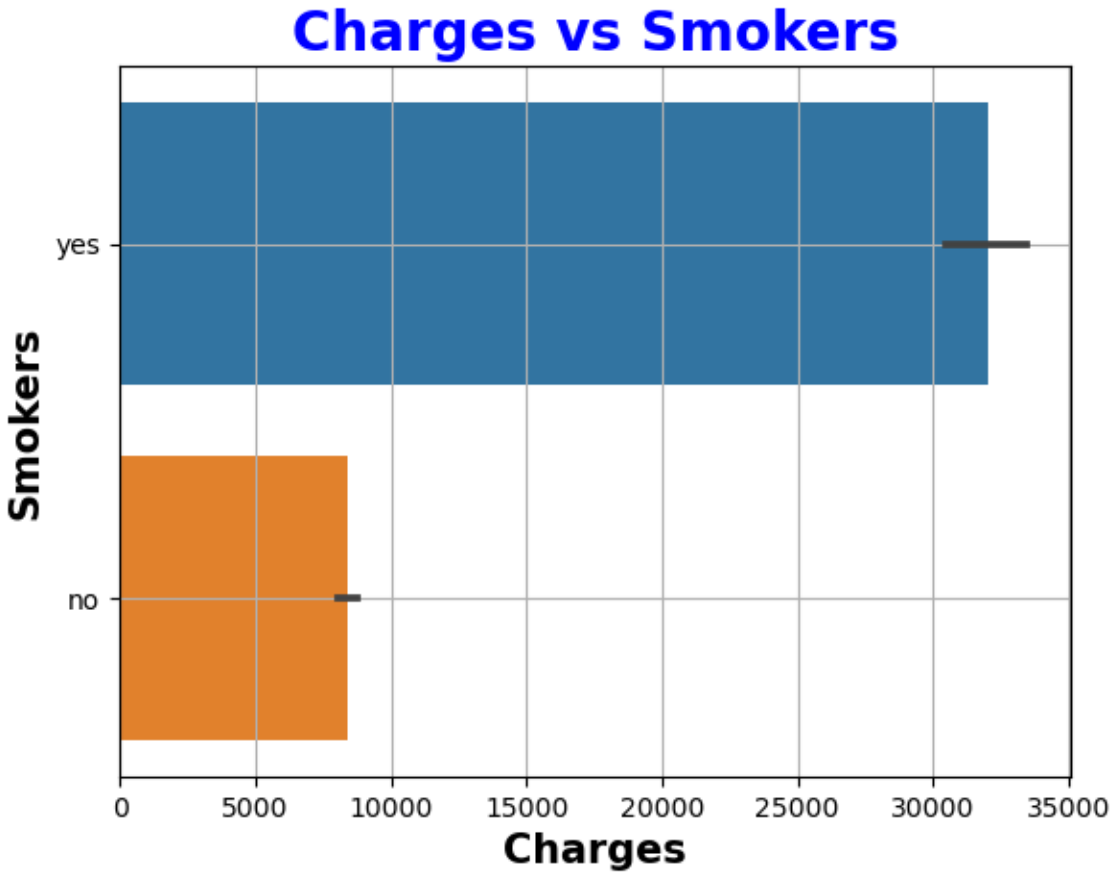
In [259]:
```
sns.scatterplot(data=df,x='charges',y='bmi')

plt.xticks(color='k')
plt.yticks(color='k')

plt.title('Charges vs bmi',size=20,c='Green',fontweight='bold')

plt.xlabel('Children',size=15,fontweight='bold')
plt.ylabel('Charges',size=15,fontweight='bold')

plt.grid()

plt.show()
```



there is no relation betwwen the bmi , children with charges

Charges vs Smokers

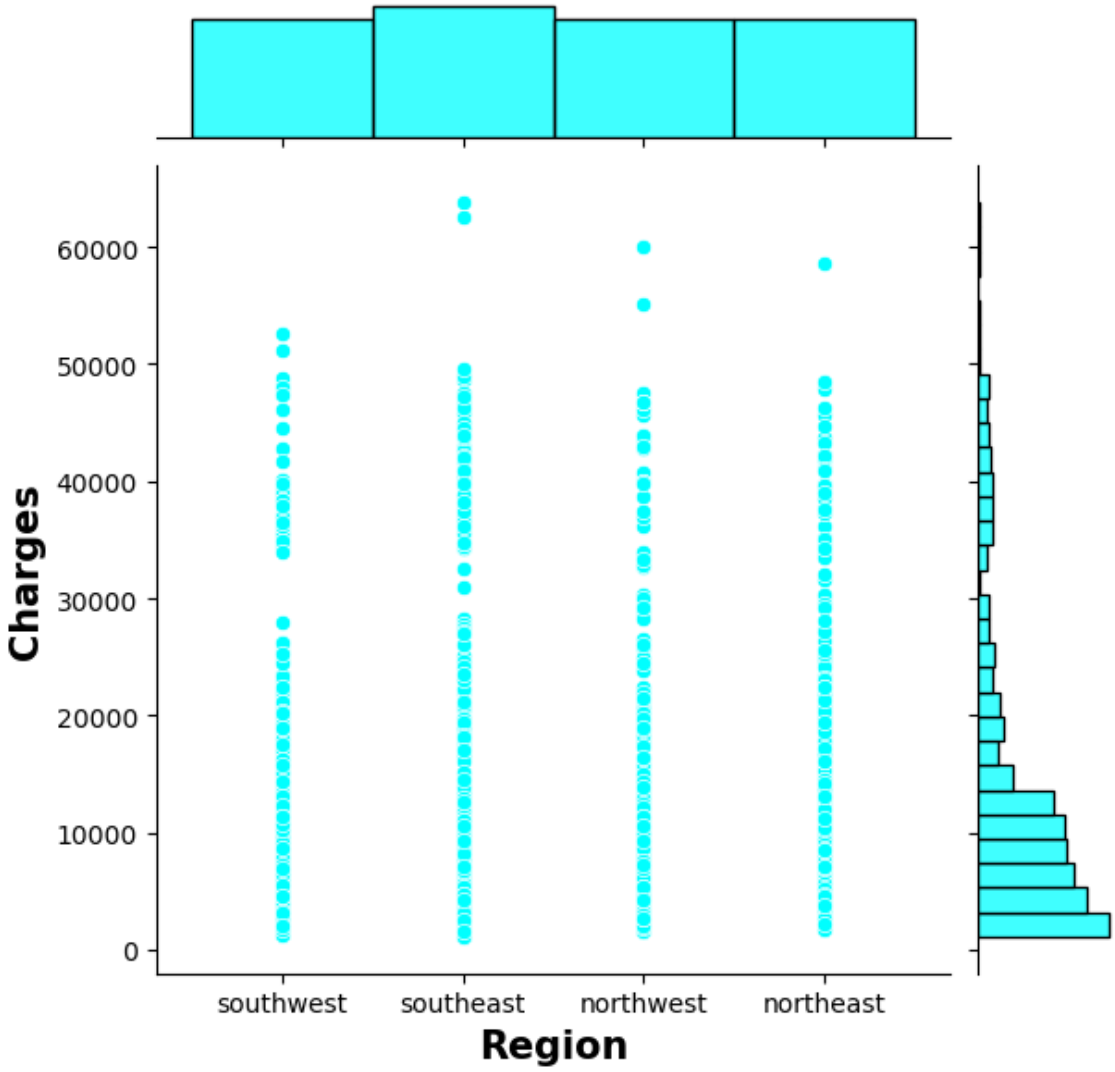
In [261]:
```
sns.barplot(data=df,x='charges',y='smoker')

plt.xticks(color='k')
plt.yticks(color='k')

plt.title('Charges vs Smokers',size=20,c='Blue',fontweight='bold')

plt.xlabel('Charges',size=15,fontweight='bold')
plt.ylabel('Smokers',size=15,fontweight='bold')

plt.grid()

plt.show()
```



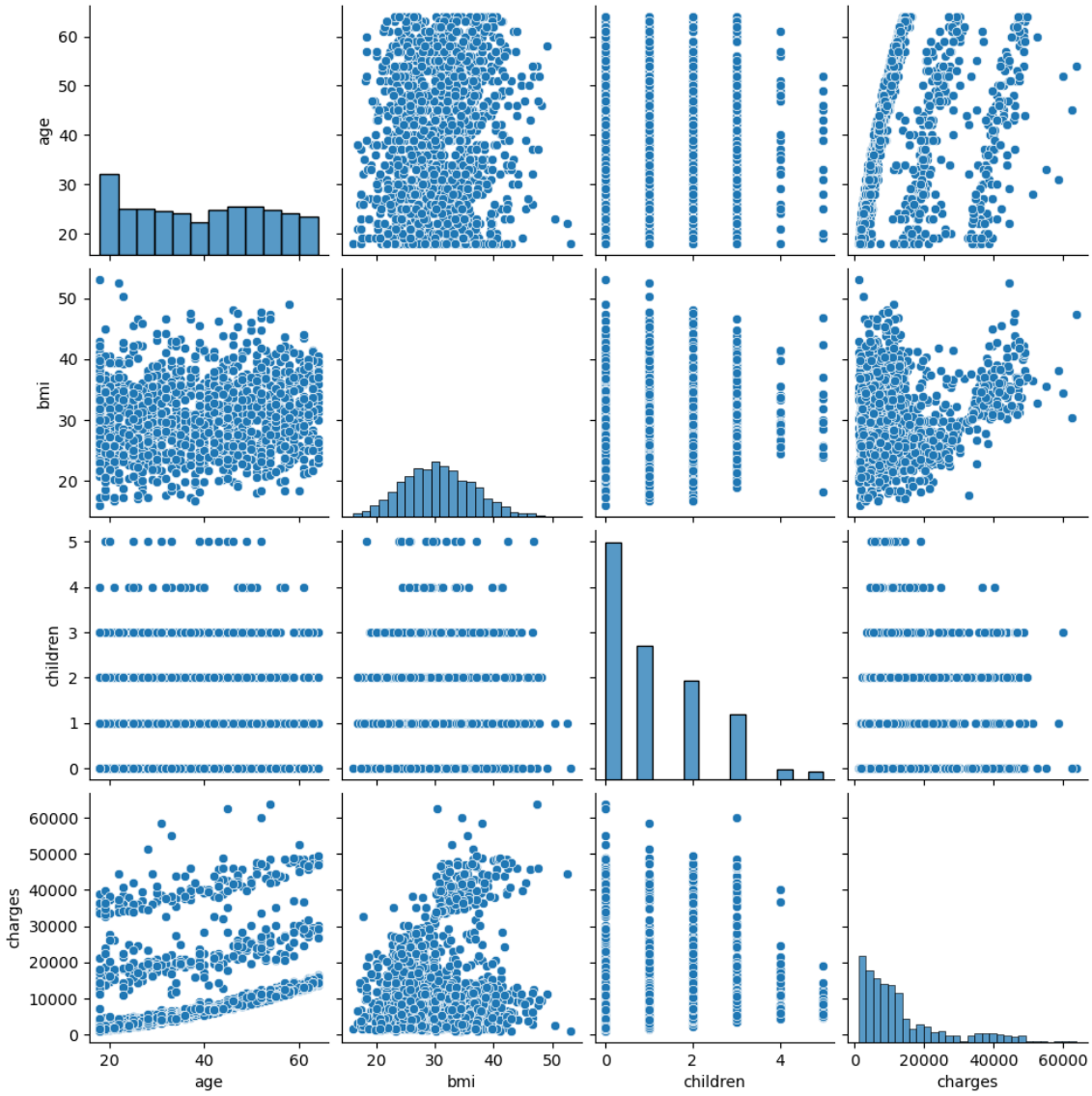
Based on the data we observe that there is correlation between smokers and charges. smoker have to pay the more charges as compare to non-smokers

Region vs Charges

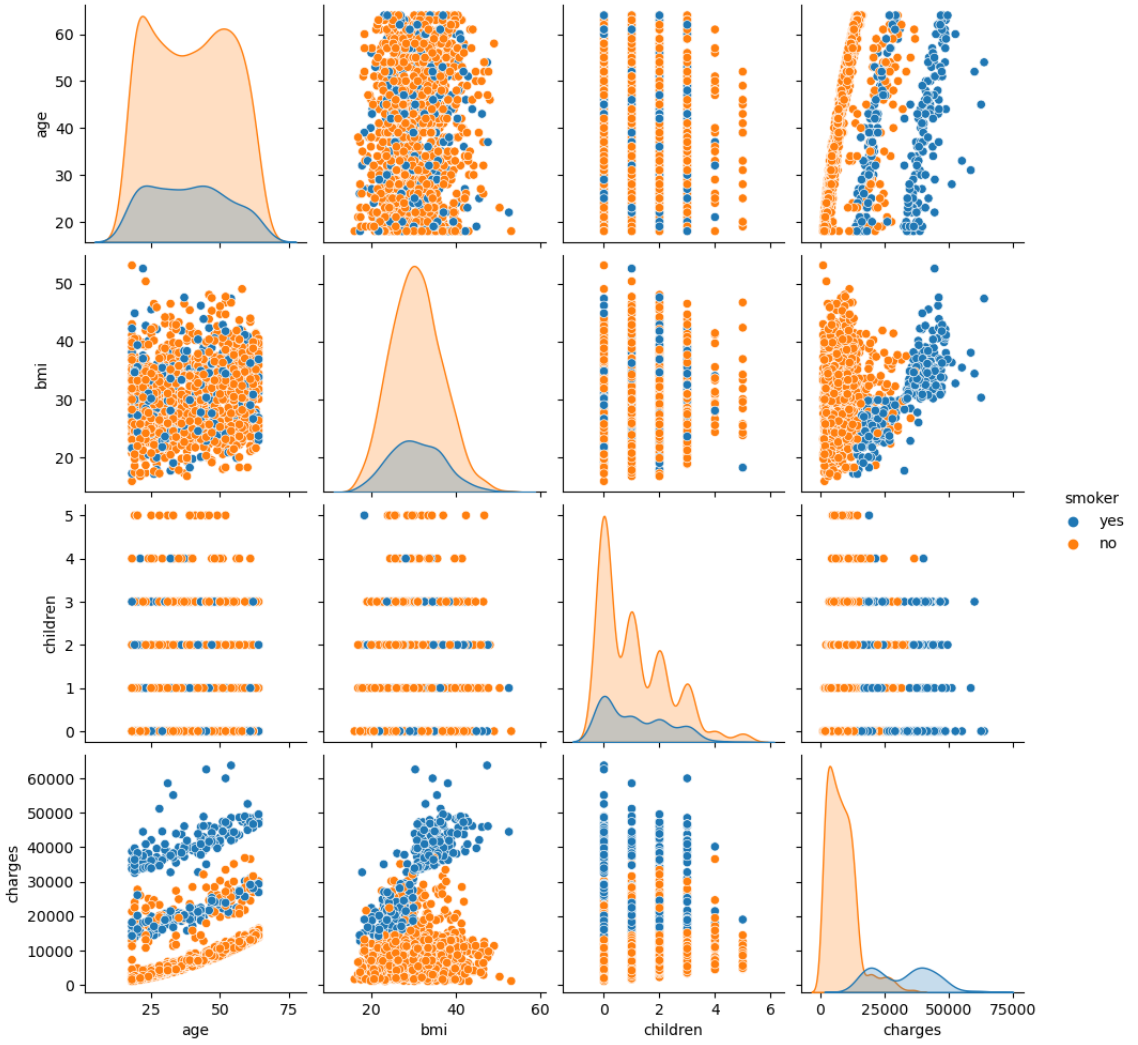
In [264]:
```
sns.jointplot(data=df,x='region',y='charges',color='cyan')

plt.xticks(color='k')
plt.yticks(color='k')

plt.xlabel('Region',size=15,fontweight='bold')
plt.ylabel('Charges',size=15,fontweight='bold')

plt.show()
```



In [269]:
```
sns.pairplot(data=df)

plt.show()
```



In [195]:
```
sns.pairplot(data=df,hue='smoker')

plt.show()
```



Observation of Bivariate Analysis

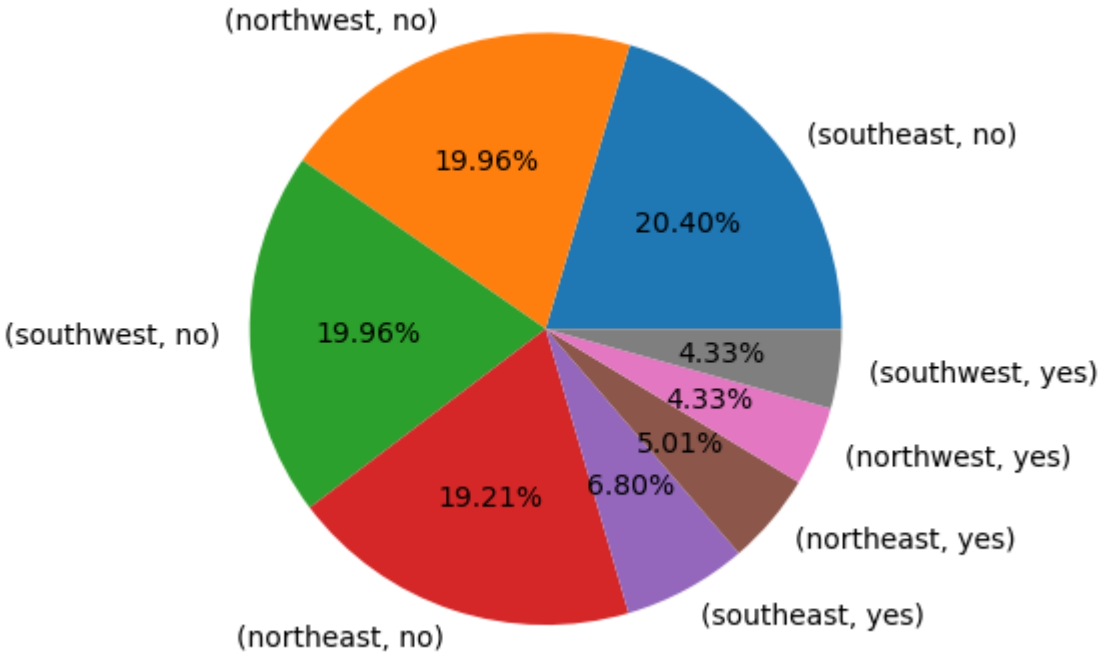
1.There is a positive correlation between age and charges as age increases the charges also increase.

2.Count of male is more as per insurance charges

Multivariate Analysis

smokers count by region

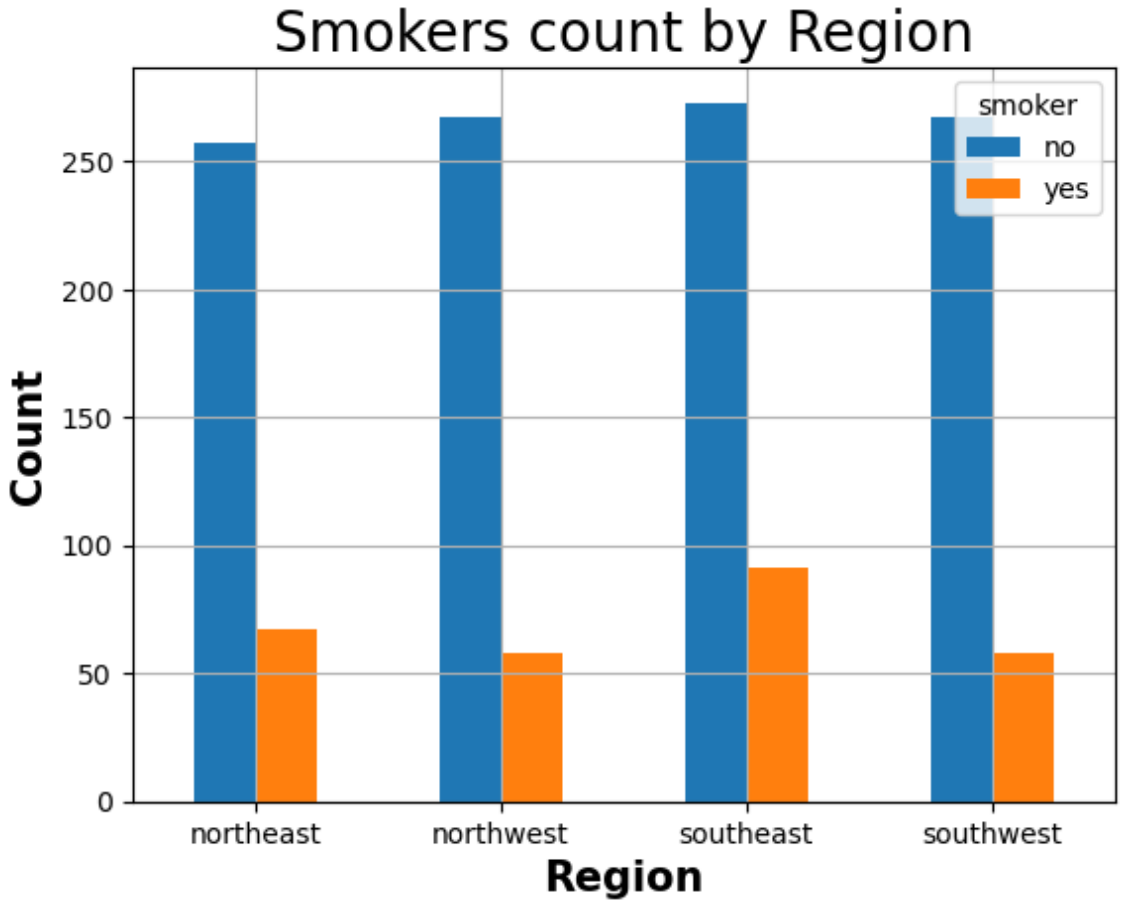
In [170]:
```
df[['region','smoker']].value_counts().plot(kind='pie',autopct='%0.2f%%')

plt.grid()

plt.show()
```



In [120]:
```
data=df.groupby(['region','smoker']).size().unstack()

print(data)
```

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

In [275]:
```
data.plot(kind='bar')

plt.title('Smokers count by Region',size='20')

plt.xticks(color='k',rotation='horizontal')
plt.yticks(color='k')

plt.xlabel('Region',size=15,fontweight='bold')
plt.ylabel('Count',size=15,fontweight='bold')

plt.grid()

plt.show()
```



this is the distribution.

Distribution between the charges and region

In [173]:
```
data5=df[['region','charges']]

print(data5)
```

```
         region      charges
0     southwest  16884.92400
1     southeast   1725.55230
2     southeast   4449.46200
3     northwest  21984.47061
4     northwest   3866.85520
...         ...          ...
1333  northwest  10600.54830
1334  northeast   2205.98080
1335  southeast   1629.83350
1336  southwest   2007.94500
1337  northwest  29141.36030

[1338 rows x 2 columns]
```

In [175]:
```
data5.aggregate(['max'])
```

Out[175]:

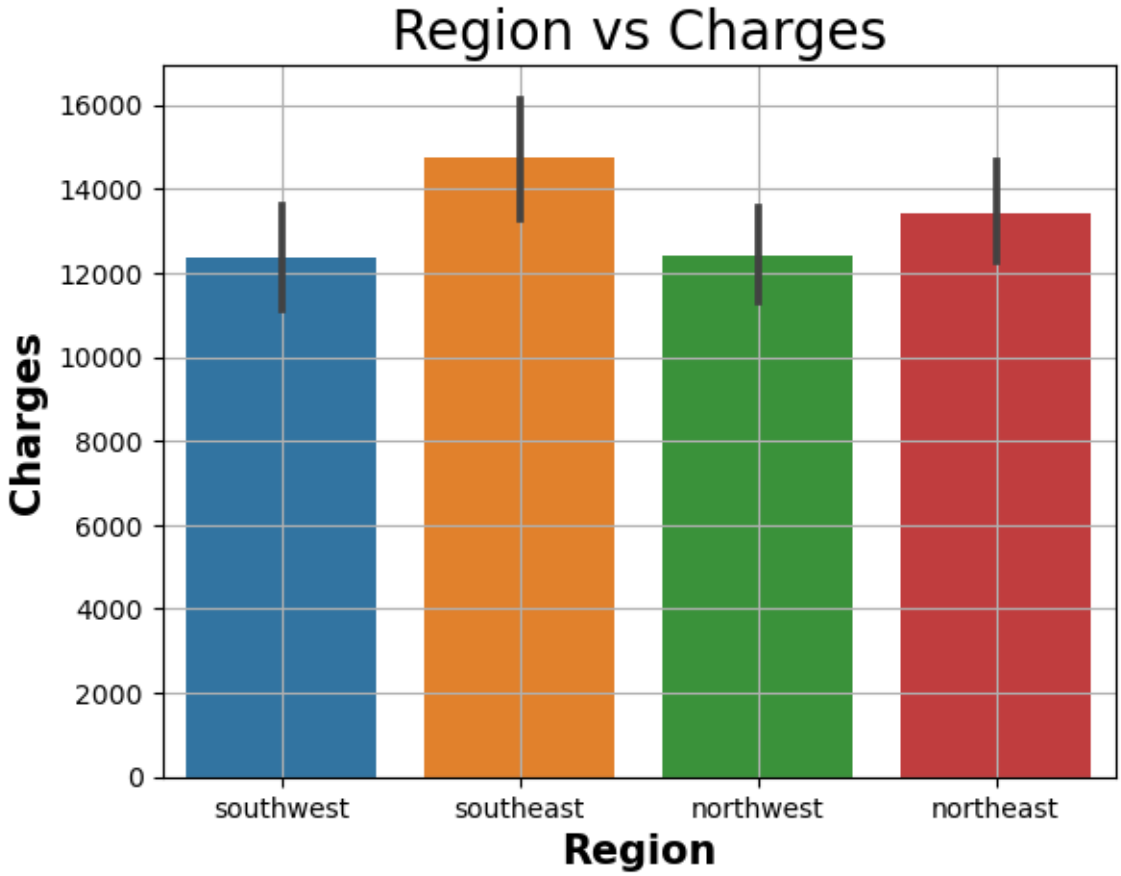
	region	charges
max	southwest	63770.42801

In [177]:
```
data5.aggregate(['min'])
```

Out[177]:

	region	charges
min	northeast	1121.8739

In [276]:
```
sns.barplot(data=df,x='region',y='charges')

plt.title("Region vs Charges",size=20)

plt.xticks(color='k',rotation='horizontal')
plt.yticks(color='k')

plt.xlabel('Region',size=15,fontweight='bold')
plt.ylabel('Charges',size=15,fontweight='bold')

plt.grid()

plt.show()
```



number of people from southeast's pay more charges and you can observe distribution .

southwest - maximum insurance charge

northeast - minimum insurance charges

Age wise smokers count

In [146]:
```
data1=df.groupby(['age','smoker']).size().unstack()

print(data1)
```

```
smoker  no  yes
age
18      57   12
19      50   18
20      20    9
21      26    2
22      22    6
23      21    7
24      22    6
25      23    5
26      25    3
27      19    9
28      25    3
29      21    6
30      18    9
31      22    5
32      21    5
33      20    6
34      21    5
35      20    5
36      19    6
37      16    9
38      23    2
39      19    6
40      22    5
41      25    2
42      19    8
43      15   12
44      21    6
45      24    5
46      24    5
47      19   10
48      24    5
49      24    4
50      25    4
51      23    6
52      23    6
53      23    5
54      23    5
55      24    2
56      22    4
57      22    4
58      24    1
59      21    4
60      18    5
61      17    6
62      19    4
63      18    5
64      15    7
```

In [184]:
```
df.corr()
```

```
C:\Users\Hp\AppData\Local\Temp\ipykernel_14280\1134722465.py:1: FutureWarn
ing: The default value of numeric_only in DataFrame.corr is deprecated. In
a future version, it will default to False. Select only valid columns or s
pecify the value of numeric_only to silence this warning.
  df.corr()
```

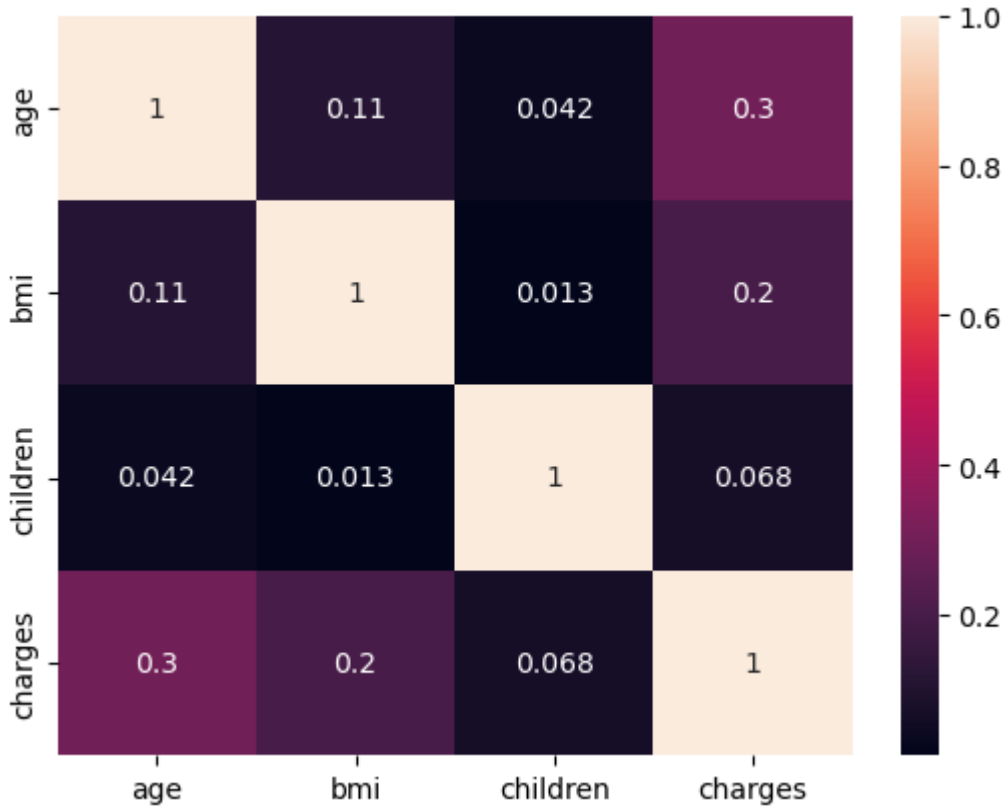
Out[184]:

	age	bmi	children	charges
age	1.000000	0.109516	0.042469	0.298970
bmi	0.109516	1.000000	0.012513	0.198433
children	0.042469	0.012513	1.000000	0.068073
charges	0.298970	0.198433	0.068073	1.000000

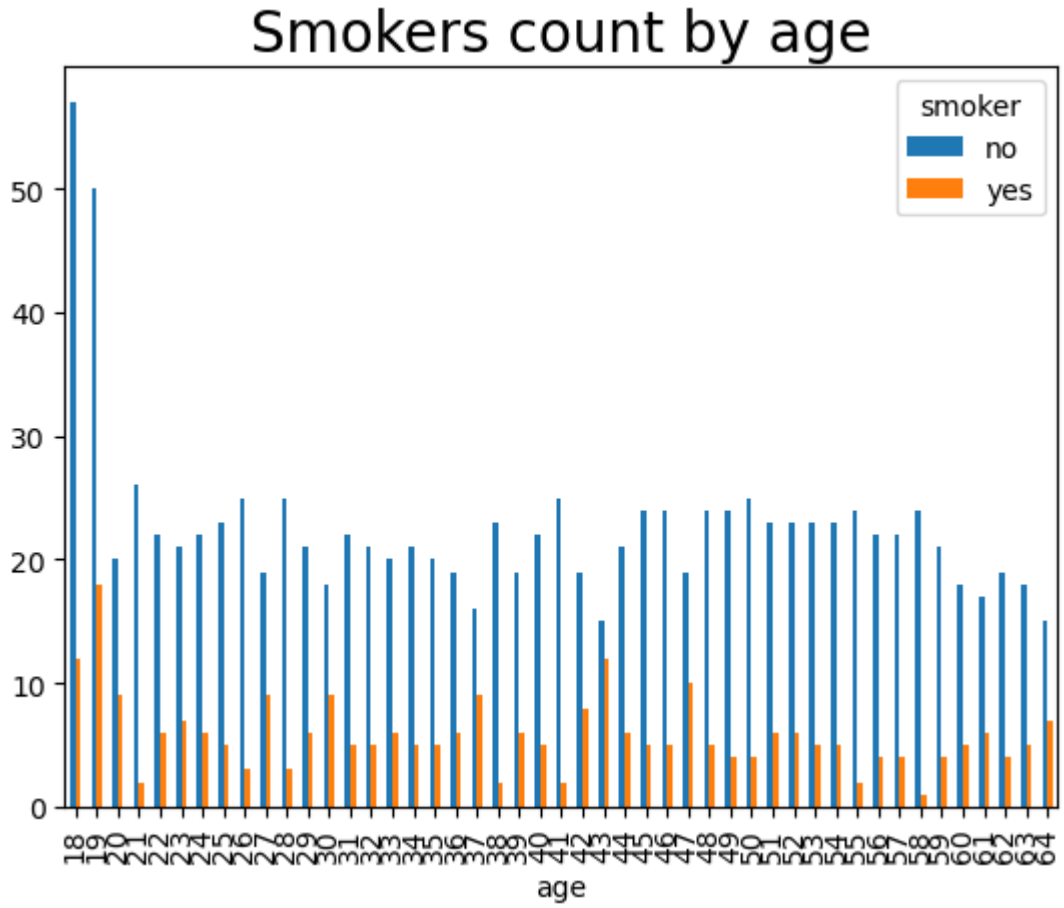
In [185]:
```
sns.heatmap(df.corr(),annot=True)
```

```
C:\Users\Hp\AppData\Local\Temp\ipykernel_14280\4277794465.py:1: FutureWarn
ing: The default value of numeric_only in DataFrame.corr is deprecated. In
a future version, it will default to False. Select only valid columns or s
pecify the value of numeric_only to silence this warning.
  sns.heatmap(df.corr(),annot=True)
```

Out[185]: <Axes: >



In [151]:
```
data1.plot(kind='bar')

plt.title('Smokers count by age',size='20')

plt.show()
```



Obesrvations on Mutivariate Analysis

1.It show that southeast people smoke more .

2.number of people from southeast's pay more charges and you can observe distribution .

3.southwest - maximum insurance charge

4.northeast - minimum insurance charges

5.teenagers smoke more and large number of non-smokers also belong to teenagers.