In [90]:

Out[90]:

PROJECT ON MEDICAL INSURANCE

Context

- **From a Medical Insurance Costs dataset we'll build models using Regression to predict individual insurance costs. To help us with this task we have some informations of the persons wih insurance, like the Age, Sex, BMI (Body Mass Index), Children, Smokers, Region and their Charges. In this notebook we'll clean and organize the data, build the models and compare the results.**

Data Description

- **The information about children's medical health insurance includes details about their age, sex, smokers, regions and their insurance charges depends on health conditions, they receive to help understand how to support their healthcare needs better.**

Objectives

- **Children medical health insurance is a type of special protection that helps children without parents get the medical care they need, like going to the doctor or staying in the hospital, If they become sick or injured. It's like having a safety net to make sure**

Importing Libraries

In [1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
```

Read Dataset

In [2]:
```
A = pd.read_csv("insurance.csv")
A
```

Out[2]:

	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	33.000	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

shows number of rows and columns

In [24]:
```
A.shape
```

Out[24]:
```
(1338, 7)
```

In [25]:
```
A.size
```

Out[25]:
```
9366
```

shows all column name in dataframe

In [26]:
```
A.columns
```

Out[26]:
```
Index(['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges'], dtype
='object')
```

shows information of dataset

In [22]:
```
A.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       1338 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 [30]:
```
A.describe(include="all")
```

Out[30]:

	age	sex	bmi	children	smoker	region	charges
count	1338.000000	1338	1338.000000	1338.000000	1338	1338	1338.000000
unique	NaN	2	NaN	NaN	2	4	NaN
top	NaN	male	NaN	NaN	no	southeast	NaN
freq	NaN	676	NaN	NaN	1064	364	NaN
mean	39.207025	NaN	30.663397	1.094918	NaN	NaN	13270.422265
std	14.049960	NaN	6.098187	1.205493	NaN	NaN	12110.011237
min	18.000000	NaN	15.960000	0.000000	NaN	NaN	1121.873900
25%	27.000000	NaN	26.296250	0.000000	NaN	NaN	4740.287150
50%	39.000000	NaN	30.400000	1.000000	NaN	NaN	9382.033000
75%	51.000000	NaN	34.693750	2.000000	NaN	NaN	16639.912515
max	64.000000	NaN	53.130000	5.000000	NaN	NaN	63770.428010

Check NaN Values

In [31]:
```
A.isna()
```

Out[31]:

	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	False	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

In [36]:
```
A.isna().sum()
```

Out[36]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Duplicates Data

In [38]:
```
A[A.duplicated()]
```

Out[38]:

	age	sex	bmi	children	smoker	region	charges
581	19	male	30.59	0	no	northwest	1639.5631

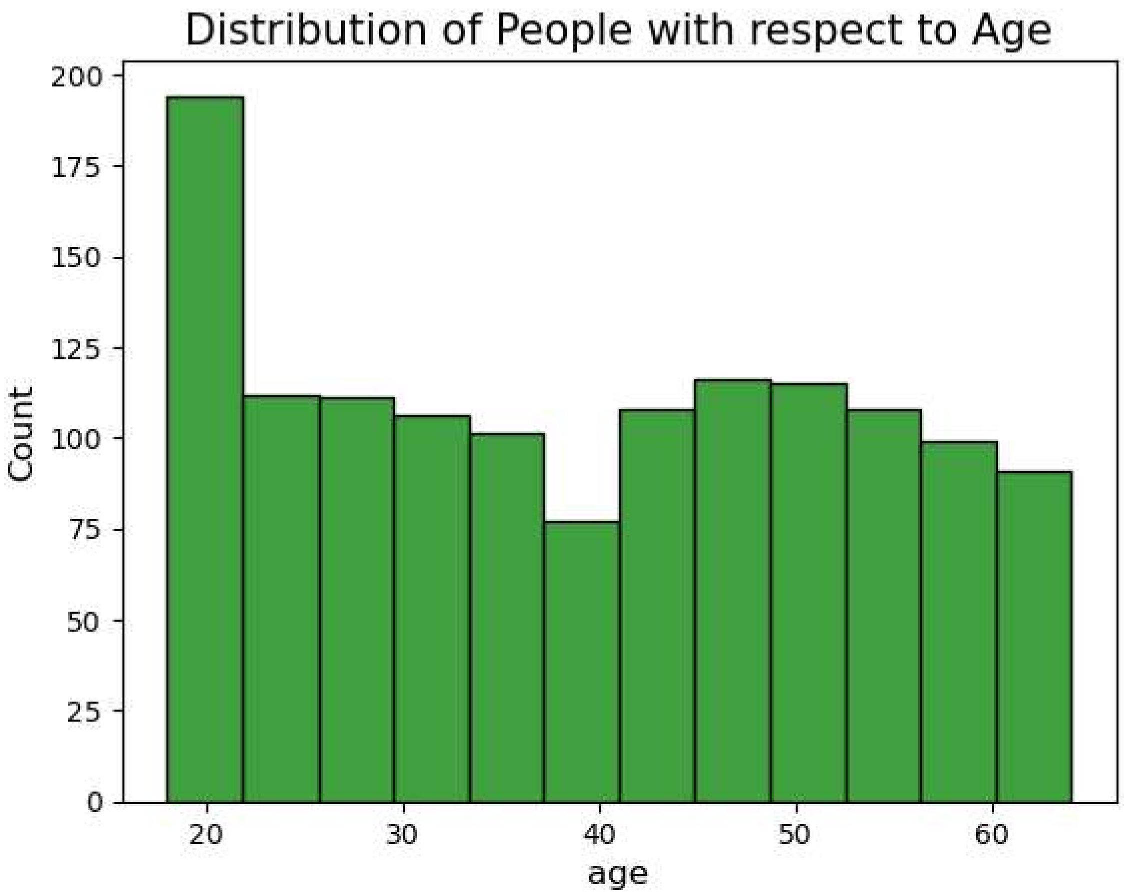
In [40]:
```
A.drop_duplicates(inplace=True)
```

In [ ]:

Data Analysis and Visualization on Dataset

Age Column

In [17]:
```
sns.histplot(data=A,x="age",color="green")
plt.title("Distribution of People with respect to Age",size=15)
plt.xlabel("age", size=12)
plt.ylabel("Count", size=12)
plt.show()
```

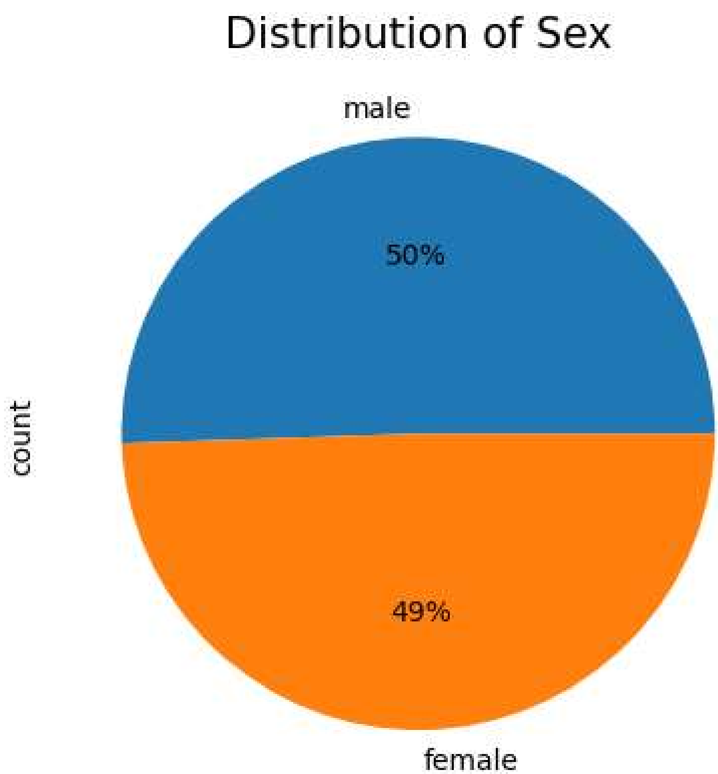


Observation

- In this histplot graph show that more number of people in the dataset near the age of range of 20 and ramaining range of age 25 to 70 distribution of age is normal.

Sex Column

In [46]:
```
A["sex"].value_counts().plot(kind="pie",autopct="%i%%")
plt.title("Distribution of Sex",size=15)
plt.show()
```



Observation-

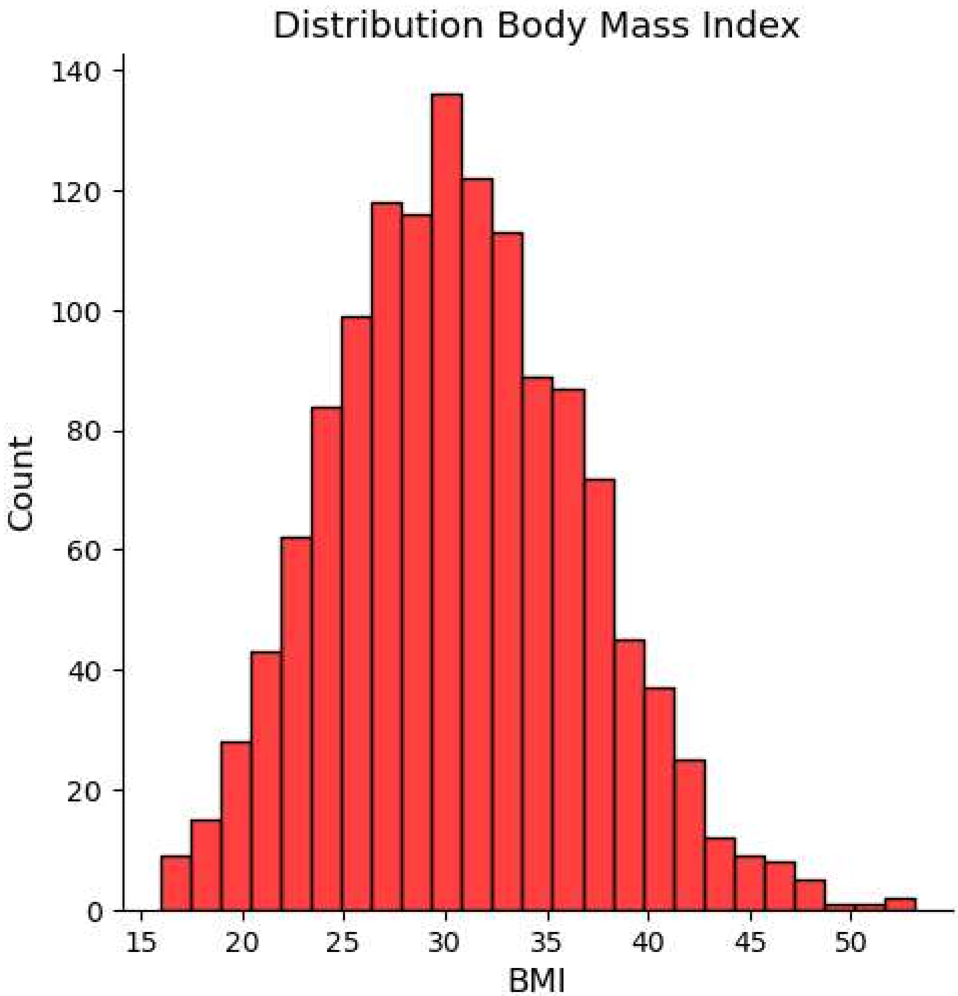
- In this piechart graph show that distribution of sex in which count of male is 50% and count of female is 49%.

In [39]:
```
A["sex"].value_counts()
```

Out[39]:
```
sex
male      676
female    662
Name: count, dtype: int64
```

BMI Column (Body Mass Index)

In [88]:
```
sns.displot(data=A, x="bmi", color="red")
plt.title("Distribution Body Mass Index",size=13)
plt.xlabel("BMI", size=12)
plt.ylabel("Count", size=12)
plt.show()
```

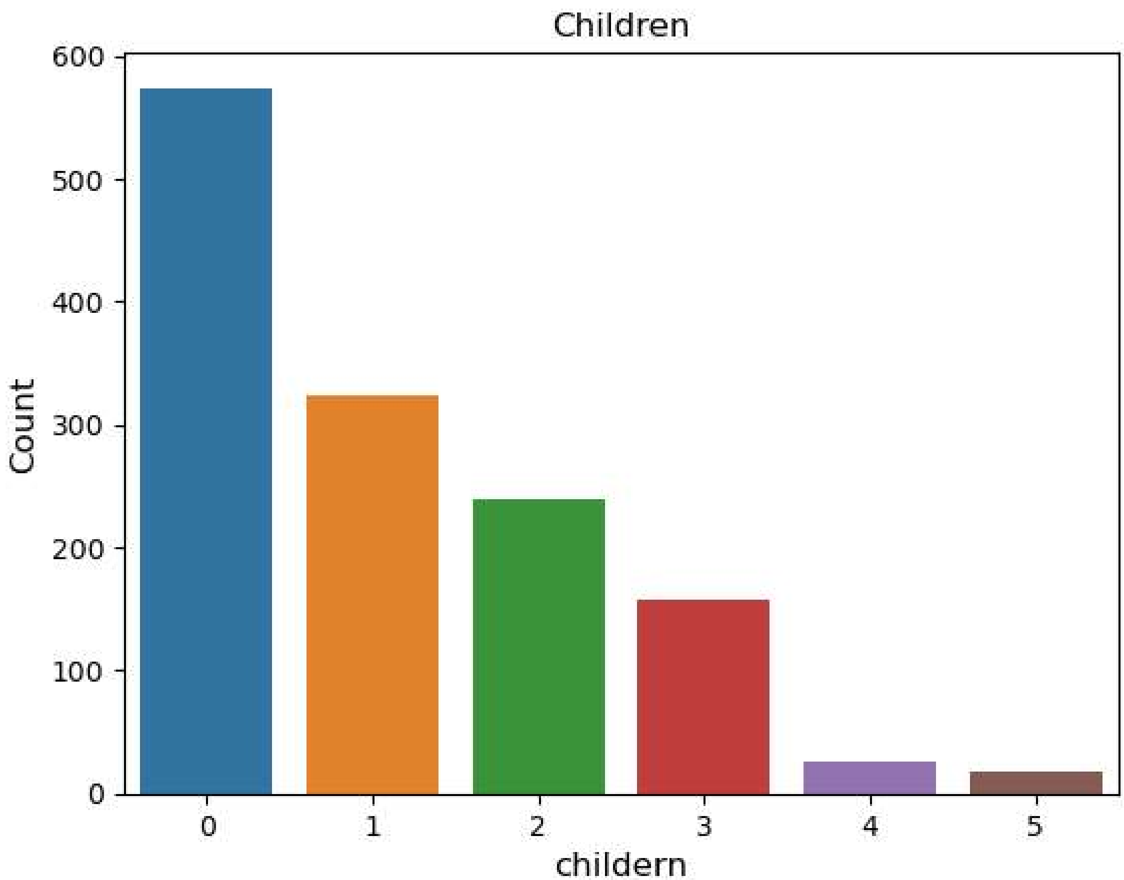


Observation:

- In this distplot graph show that **highest** BMI index is **30**. It is under **overweight**.
- Normal BMI range is 18.5 to 24.9
  - The value of BMI range **below 18.5** is **underweight**
  - The value of BMI range **18.5 to 24.9** is **mediumweight**
  - The value of BMI range **above 24.9** is **overweight**

Children Column

In [84]:
```
sns.countplot(data=A, x="children")
plt.title("Children")
plt.xlabel("childern", size=12)
plt.ylabel("Count", size=12)
plt.show()
```



Observation

- In this countplot graph we can see that number of people who doesn't have a childern and count is around 570. and other people whose have a childern is around 320. and so on in decline order.

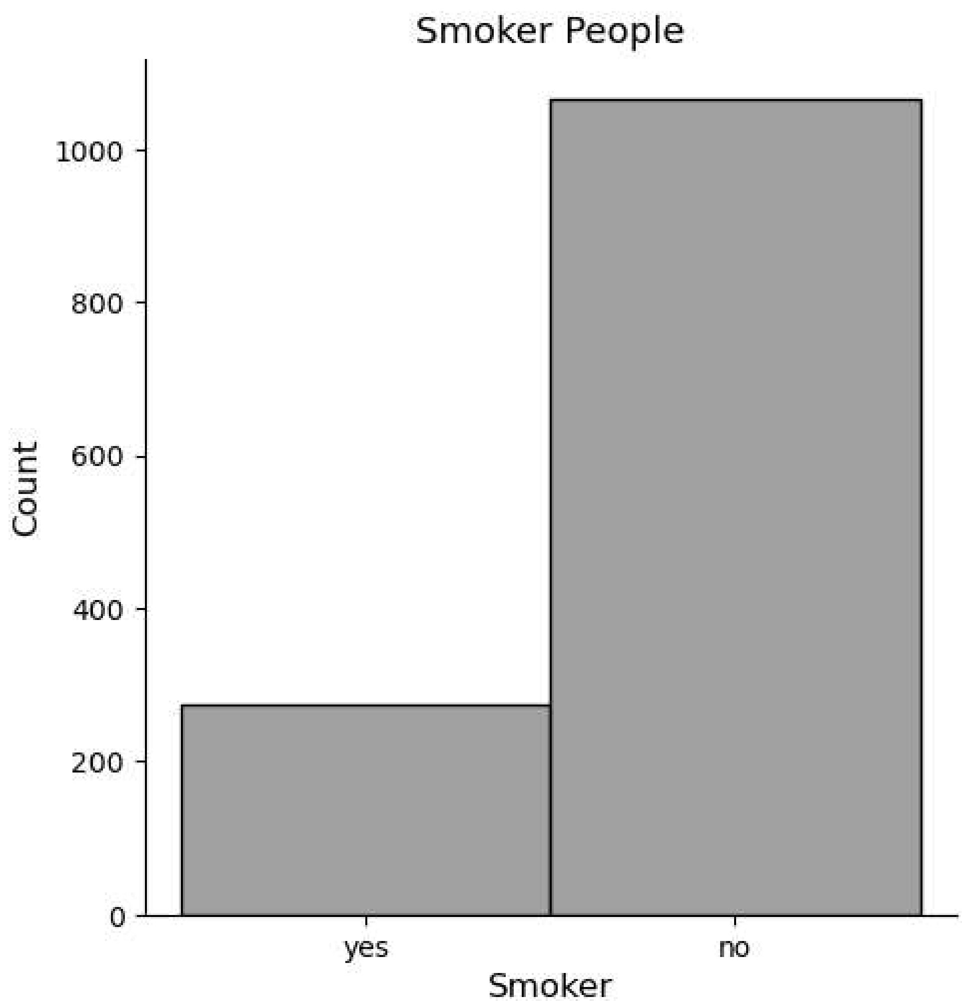
In [48]:
```
A["children"].value_counts()
```

Out[48]:
```
children
0    574
1    324
2    240
3    157
4     25
5     18
Name: count, dtype: int64
```

Smokers Column

In [91]:
```
plt.figure(figsize=(15,10))
sns.displot(data=A, x="smoker", color="grey", bins=20)
plt.title("Smoker People",size=13)
plt.xlabel("Smoker", size=12)
plt.ylabel("Count", size=12)
plt.show()
```

```
<Figure size 1500x1000 with 0 Axes>
```



Observation

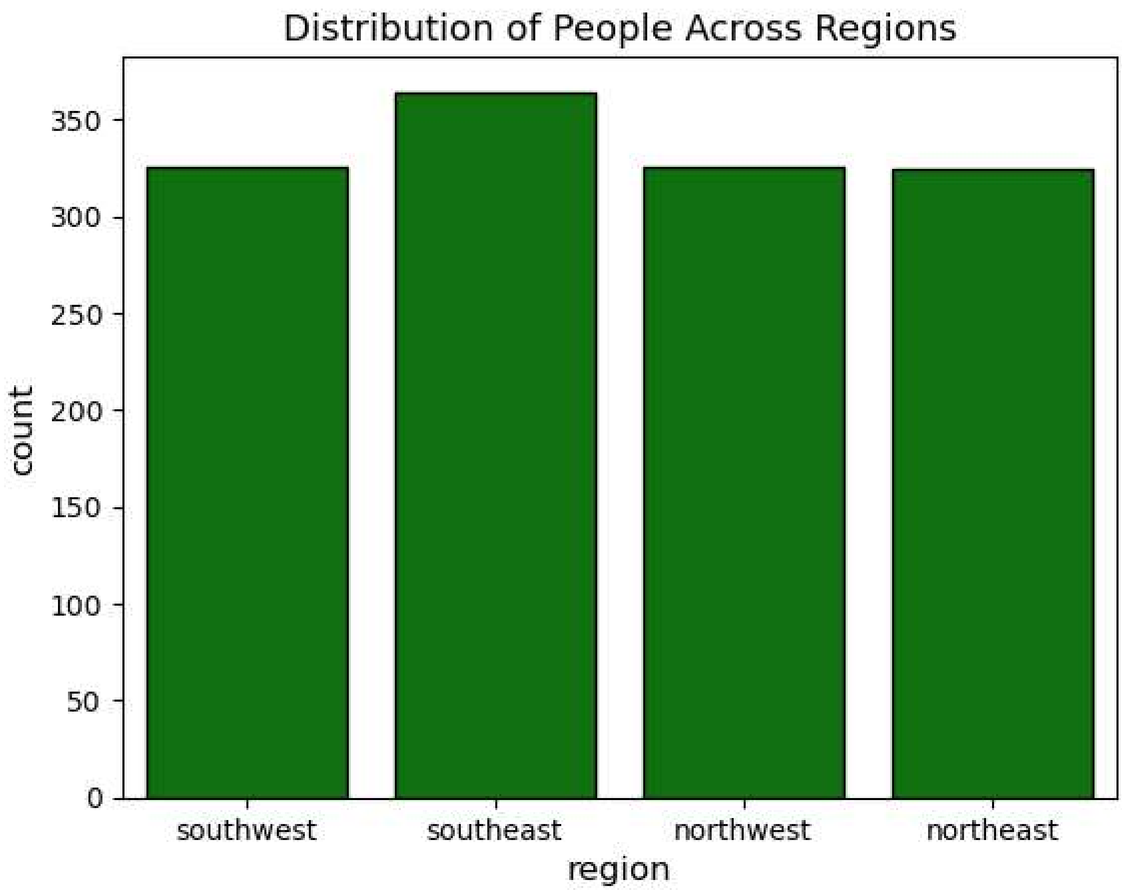
- In this displot graph we can see that number of people who does not smoke is more rather than whose does.

In [58]:
```
A["smoker"].value_counts()
```

Out[58]:
```
smoker
no     1064
yes     274
Name: count, dtype: int64
```

Regions Column

In [29]:
```
sns.countplot(data=A,x="region", color="green", edgecolor="black")
plt.title("Distribution of People Across Regions",size=13)
plt.xlabel("region", size=12)
plt.ylabel("count", size=12)
plt.show()
```



Observation

- In this countplot graph we can see that count of people in this four regions are almost same but littlebit more is southeast region.

In [74]:
```
A["region"].value_counts()
```

Out[74]:
```
region
southeast    364
southwest    325
northwest    325
northeast    324
Name: count, dtype: int64
```

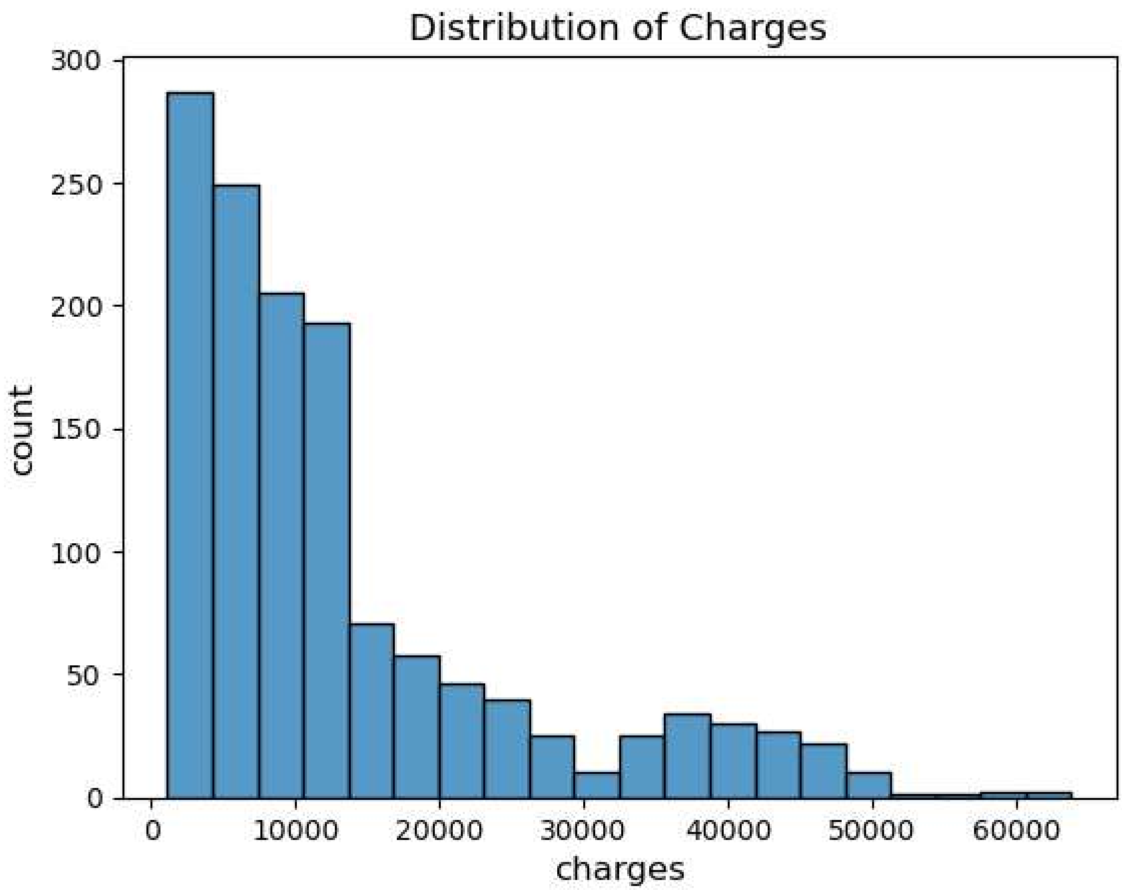
In [73]:
```
A_new=A.groupby("region")
for x,y in A_new:
    print()
    print(x)
    print(y)
```

```
northeast
      age     sex     bmi  children smoker     region      charges
8      37    male  29.830         2     no  northeast   6406.41070
10     25    male  26.220         0     no  northeast   2721.32080
16     52  female  30.780         1     no  northeast  10797.33620
17     23    male  23.845         0     no  northeast   2395.17155
20     60  female  36.005         0     no  northeast  13228.84695
...   ...     ...     ...       ...    ...        ...          ...
1321   62    male  26.695         0    yes  northeast  28101.33305
1325   61    male  33.535         0     no  northeast  13143.33665
1326   42  female  32.870         0     no  northeast   7050.02130
1328   23  female  24.225         2     no  northeast  22395.74424
1334   18  female  31.920         0     no  northeast   2205.98080

[324 rows x 7 columns]

northwest
      age     sex     bmi  children smoker     region      charges
3      33    male  22.705         0     no  northwest  21984.47061
4      32    male  28.880         0     no  northwest   3866.85520
7      37  female  27.740         3     no  northwest   7281.50560
9      60  female  25.840         0     no  northwest  28923.13692
24     37    male  28.025         2     no  northwest   6203.90175
...   ...     ...     ...       ...    ...        ...          ...
1319   39  female  26.315         2     no  northwest   7201.70085
1320   31    male  31.065         3     no  northwest   5425.02335
1324   31    male  25.935         1     no  northwest   4239.89265
1333   50    male  30.970         3     no  northwest  10600.54830
1337   61  female  29.070         0    yes  northwest  29141.36030

[325 rows x 7 columns]

southeast
      age     sex    bmi  children smoker     region     charges
1      18    male  33.77         1     no  southeast   1725.5523
2      28    male  33.00         3     no  southeast   4449.4620
5      31  female  25.74         0     no  southeast   3756.6216
6      46  female  33.44         1     no  southeast   8240.5896
11     62  female  26.29         0    yes  southeast  27808.7251
...   ...     ...    ...       ...    ...        ...         ...
1322   62    male  38.83         0     no  southeast  12981.3457
1323   42  female  40.37         2    yes  southeast  43896.3763
1327   51    male  30.03         1     no  southeast   9377.9047
1330   57  female  25.74         2     no  southeast  12629.1656
1335   18  female  36.85         0     no  southeast   1629.8335

[364 rows x 7 columns]

southwest
      age     sex   bmi  children smoker     region      charges
0      19  female  27.9         0    yes  southwest  16884.92400
12     23    male  34.4         0     no  southwest   1826.84300
15     19    male  24.6         1     no  southwest   1837.23700
18     56    male  40.3         0     no  southwest  10602.38500
19     30    male  35.3         0    yes  southwest  36837.46700
...   ...     ...   ...       ...    ...        ...          ...
```

```
1316   19  female  20.6         0     no  southwest   1731.67700
1329   52    male  38.6         2     no  southwest  10325.20600
1331   23  female  33.4         0     no  southwest  10795.93733
1332   52  female  44.7         3     no  southwest  11411.68500
1336   21  female  25.8         0     no  southwest   2007.94500

[325 rows x 7 columns]
```

Charge Column

In [4]:
```
sns.histplot(data=A, x="charges",bins=20)
plt.title("Distribution of Charges",size=13)
plt.xlabel("charges", size=12)
plt.ylabel("count", size=12)
plt.show()
```



Observation

- In this hisplot graph show that we have lots of data in distribution charges in this 10,000 values and very little bit values in 30,000 to 60,000.

Correlation

In [13]:
```
#sex column
A.replace({"sex":{"male":0, "female":1}},inplace=True)

#smoker column
A.replace({"smoker":{"yes":0, "no":1}},inplace=True)

#region column
A.replace({"region":{"southeast":0, "southwest":1, "northwest":3, "northeast":4
```

Out[13]:

	age	sex	bmi	children	smoker	region	charges
0	19	1	27.900	0	0	1	16884.92400
1	18	0	33.770	1	1	0	1725.55230
2	28	0	33.000	3	1	0	4449.46200
3	33	0	22.705	0	1	3	21984.47061
4	32	0	28.880	0	1	3	3866.85520
...	...	...	...	...	...	...	...
1333	50	0	30.970	3	1	3	10600.54830
1334	18	1	31.920	0	1	4	2205.98080
1335	18	1	36.850	0	1	0	1629.83350
1336	21	1	25.800	0	1	1	2007.94500
1337	61	1	29.070	0	0	3	29141.36030

1338 rows × 7 columns

In [14]:
```
A[["charges","sex","smoker","region"]].corr()
```

Out[14]:

	charges	sex	smoker	region
charges	1.000000	-0.057292	-0.787251	-0.037020
sex	-0.057292	1.000000	0.076185	0.012741
smoker	-0.787251	0.076185	1.000000	0.036749
region	-0.037020	0.012741	0.036749	1.000000

In [15]:
```
sns.heatmap(A[["charges","sex","smoker","region"]].corr(),annot=True)
```

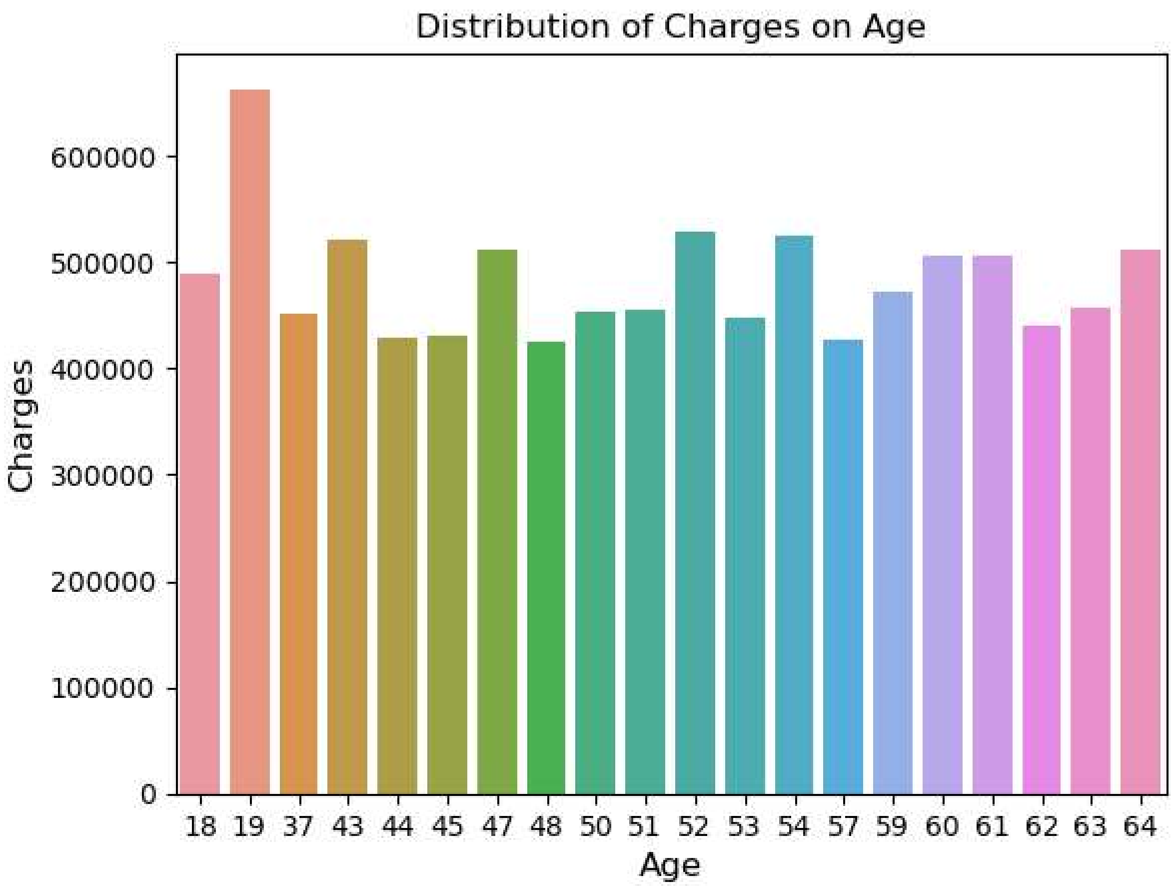
Out[15]: <Axes: >

From the Heatmap we can see that there is a **strong correlation** between being a **smoker** and the **medical charges**.

There is also small correlation between **age** and **medical charges** as well.

Insurance Charges with respect to Age

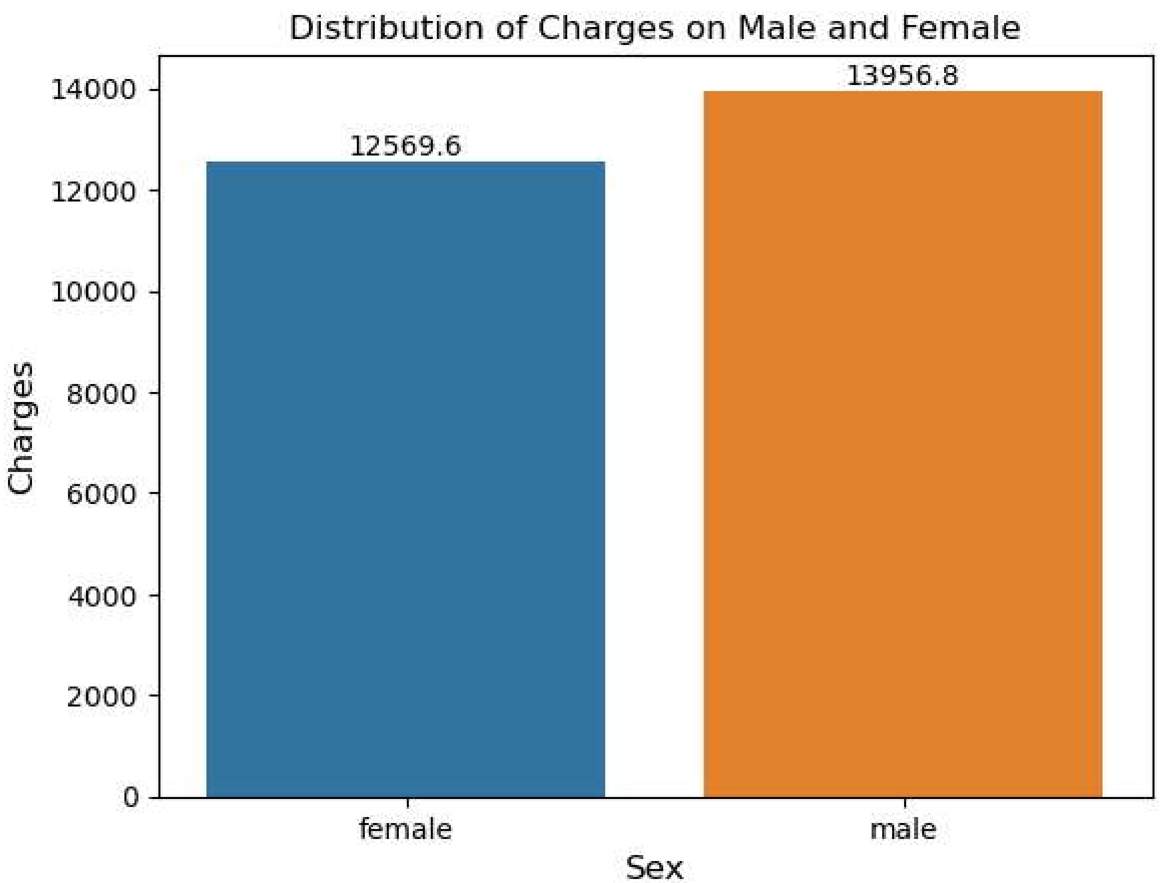
In [24]:
```
B=A.groupby(["age"], as_index=False)["charges"].sum().sort_values(by="charges".
sns.barplot(data=B, x="age", y="charges")
plt.title("Distribution of Charges on Age", size=12)
plt.xlabel("Age", size=12)
plt.ylabel("Charges", size=12)
plt.show()
```



Observation:- In this barplot graph show that whose age 19 people has more charge as compare to other age group of people.

Insurance Charges with respect to Sex

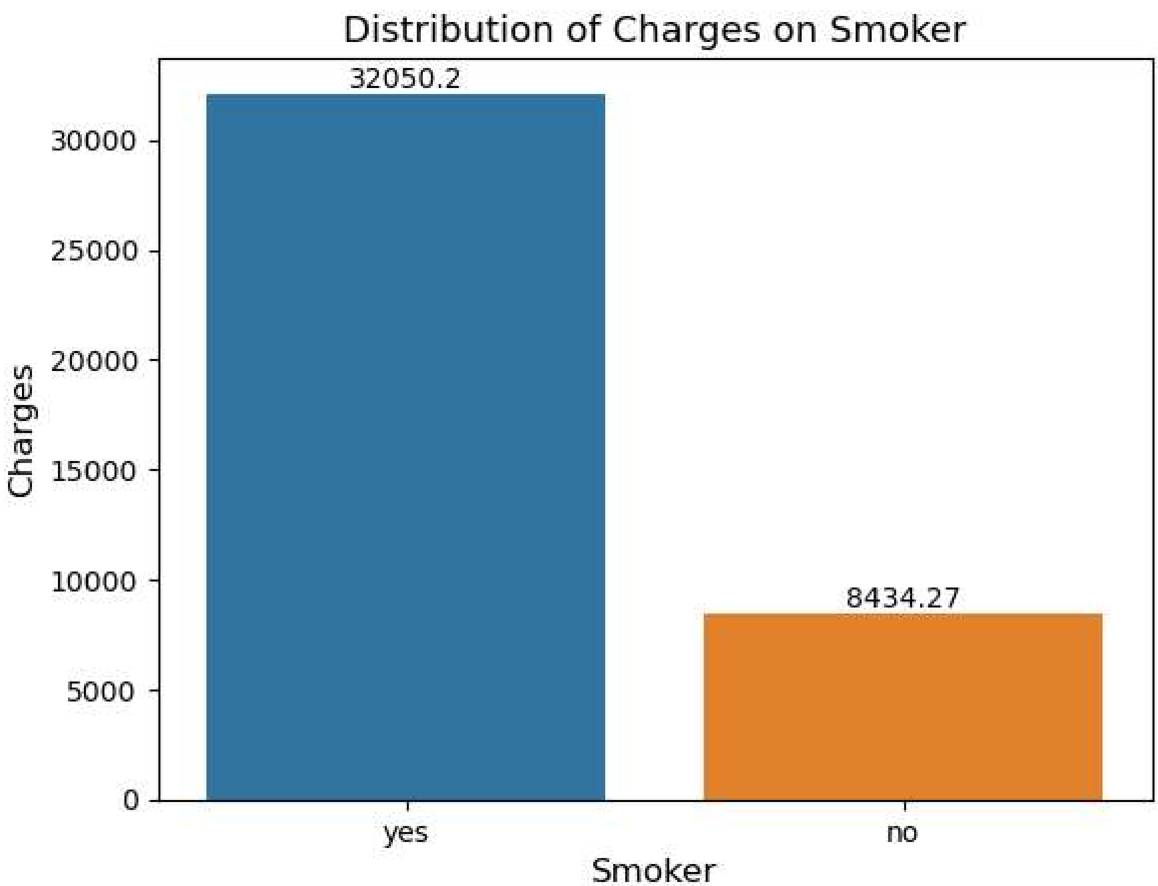
In [23]:
```
C=sns.barplot(data=A, x="sex",y="charges", ci=False)
plt.title("Distribution of Charges on Male and Female", size=12)
plt.xlabel("Sex", size=12)
plt.ylabel("Charges", size=12)
for i in C.containers:
        C.bar_label(i)
plt.show()
```



Observation: Barplot graph show that medical health insurance charges on male is high as compare to female.

Insurance Charges with respect to Smoker

In [41]:
```
D=sns.barplot(data=A, x="smoker",y="charges", ci=False)
plt.title("Distribution of Charges on Smoker", size=13)
plt.xlabel("Smoker", size=12)
plt.ylabel("Charges", size=12)
for i in D.containers:
        D.bar_label(i)
plt.show()
```

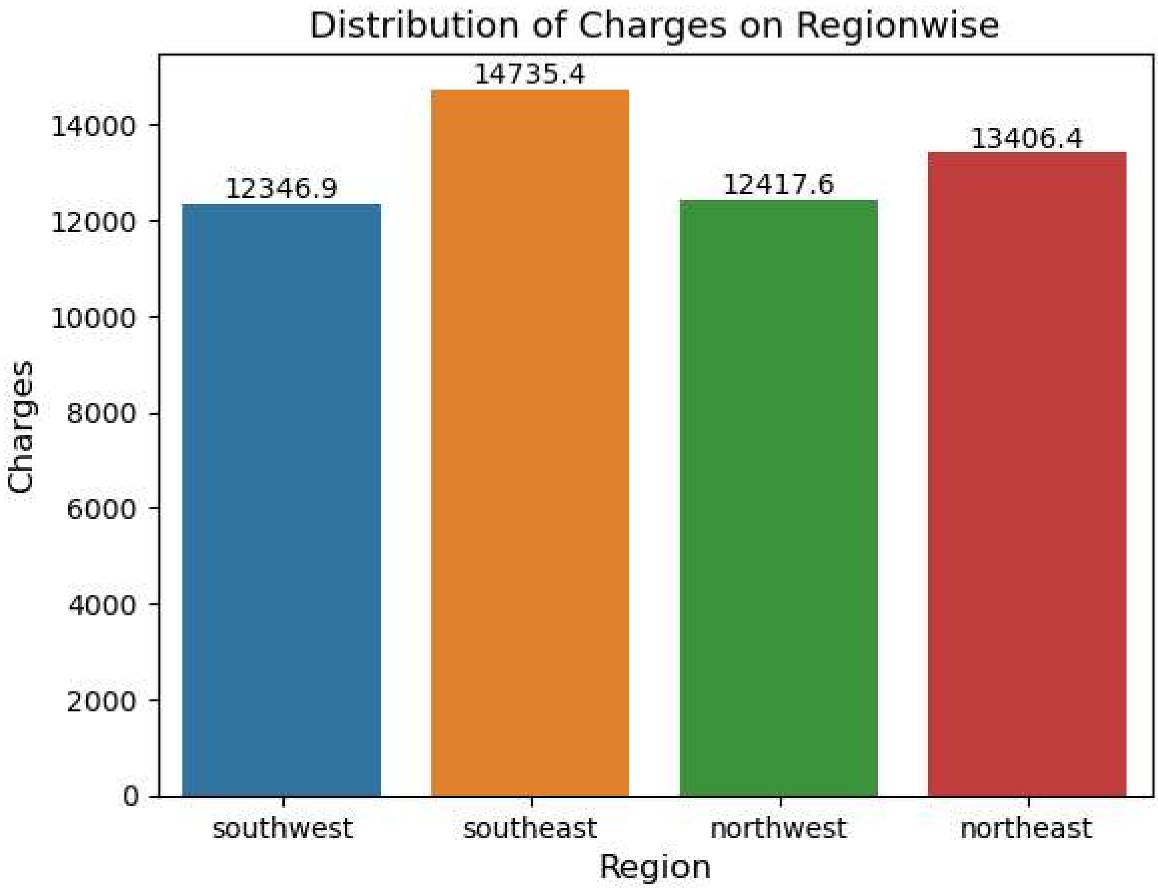


Observation

- In this barplot graph show medical health insurance charges for smoker tend to be higher compared to non-smokers. This reflects the increased health risks associated with smoking, which may lead to more frequent and costly medical treatments.

Insurance Charges with respect to Region

In [27]:
```
D=sns.barplot(data=A, x="region",y="charges", ci=False)
plt.title("Distribution of Charges on Regionwise", size=13)
plt.xlabel("Region", size=12)
plt.ylabel("Charges", size=12)
for i in D.containers:
        D.bar_label(i)
plt.show()
```

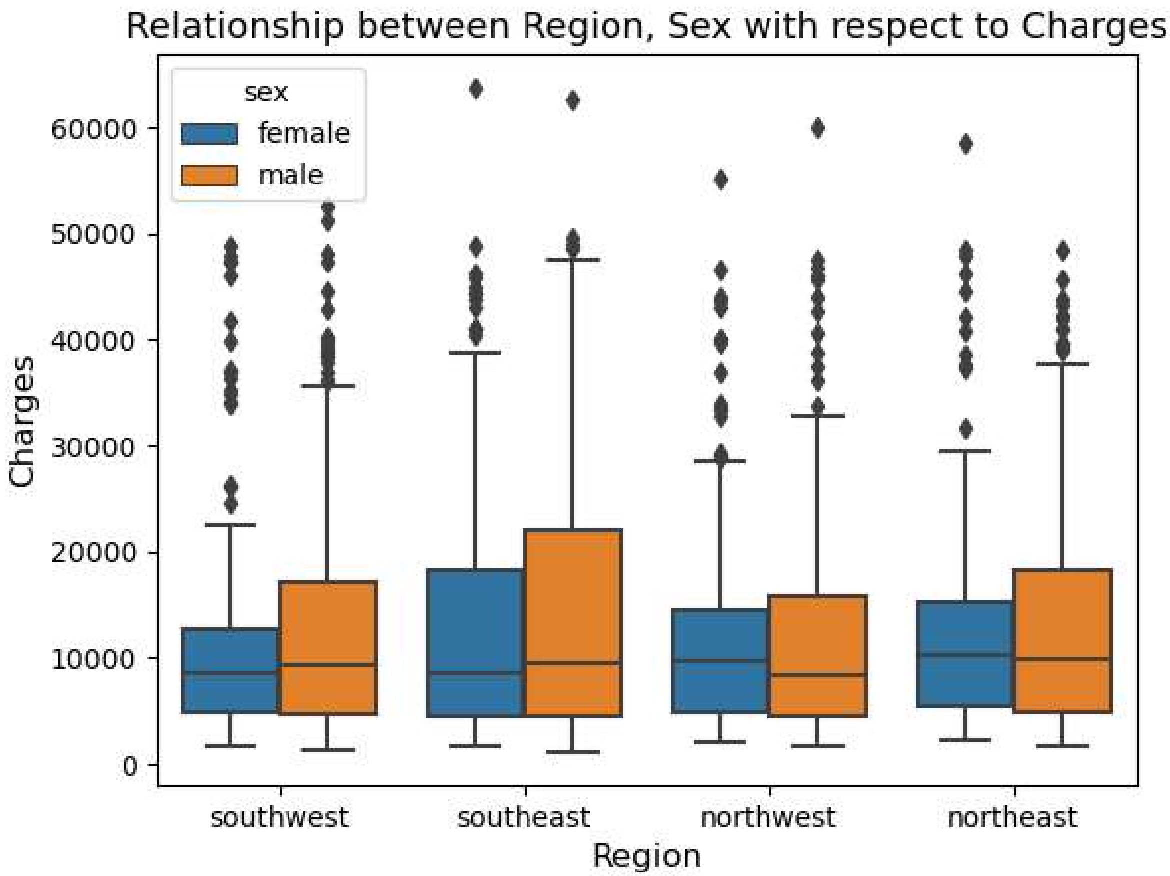


Observation:

- In this barplot graph show southeast region insurance charges is higher due to greater healthcare demand, expensive medical facilities. and other regions with lower healthcare costs and healtheir people may have more affordable insurance charges.

Relationship between Region, Sex with respect to Charges

In [4]:
```
D=sns.boxplot(data=A, x="region",y="charges",hue="sex")
plt.title("Relationship between Region, Sex with respect to Charges", size=13)
plt.xlabel("Region", size=12)
plt.ylabel("Charges", size=12)
plt.show()
```

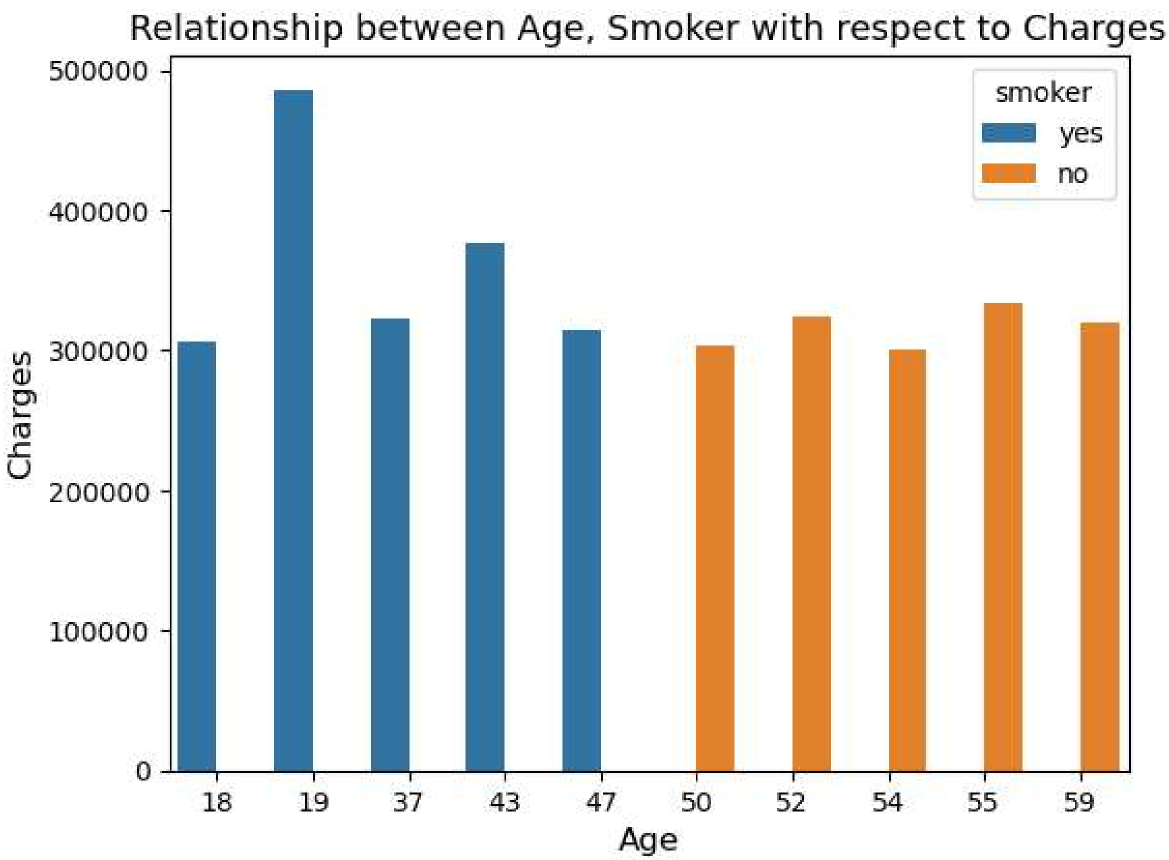


Observation

- In this swarmplot graph show Medical health insurance charges may differ due to varying healthcare costs and population health. similiarly, sex based differences in insurance charges may reflect differences in healthcare utilization patterns and risk factors between male and female. There variations aim to ensure that insurance premiums align with the specific healthcare needs and risk profiles of different regions and sex, promoting fair and equitable access to healthcare coverage.

Relationship between Age, Smokers with respect to Charges

In [38]:
```
B=A.groupby(["age","smoker"], as_index=False)["charges"].sum().sort_values(by='
sns.barplot(data=B,x="age", y="charges", hue="smoker")
plt.title("Relationship between Age, Smoker with respect to Charges", size=13)
plt.xlabel("Age", size=12)
plt.ylabel("Charges", size=12)
plt.show()
```



Obseravation

- In this barplot graph show that Medical health insurance charges are often influenced by both age and smoking status. As individuals age, insurance chargs tend to increase due to higher healthcare needs associated with aging. Additionally, smokers typically face higher insurance charges compared to non-smokers due to the increased health risks associated with smoking. Combining age and smoking status, older smokers often encounter the highest insurance charges, refecting both age related healthcare needs and the additonal risks posed by smoking.

Relationship between Smoker, Sex with respect to Charges

In [39]:
```
sns.swarmplot(data=A,x="sex", y="charges", hue="smoker")
plt.title("Relationship between Sex, Smoker with respect to Charges", size=13)
plt.xlabel("Sex", size=12)
plt.ylabel("Charges", size=12)
plt.show()
```

Observation

- Medical health insurace charges for smokers may differ based on sex. In this graph male smokers might face higher premiums compared to female smokers due to statistically higher rates of certain smoking related condition among men. However, insurance charges for smokers, regardless of sex, generally tend to be higher than those for non-smokers to cover the increased risk of health issues associated with smoking.

Conclusion

- **Age**, **BMI** and **being a smoker** affects the price of medical charges for individuals.
  - Medical charges **increase** as age and BMI **increases**.
  - Medical charge will always be **high** if you're a **smoker**.

- It helps you pay for medical expenses when you're sick or injured. With insurance, you can get treatment without worrying too much about the cost. It's like having a safety net for your health. So, having medical health insurance is a smart move to stay protected and financially secure during times of medical need.