Project Id:08 (Data Anlytics Lab Mini Project) (Roll No: 29,30,31,32)

Analyze health information to make decisions for insurance business

Context: This project uses Hypothesis Testing and Visualization to leverage customer's health information like smoking habits, BMI, age, and gender for checking statistical evidence to make valuable decisions of insurance business like charges for health insurance.

DataSet: The data at hand contains the medical costs of people characterized by certain attributes.

In [1]:

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

In [12]:

```
sns.set() #setting the default seaborn style for our plots
```

1)Reading The Dataset

In [3]:

```
insurance_df = pd.read_csv("insurance.csv")
```

In [9]:

```
insurance_df.describe()
```

Out[9]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

In [10]:

```
insurance_df.shape  # shape of dataset
```

Out[10]:

```
(1338, 7)
```

In [8]:

```
#Display the first five dataset

insurance_df.head(10)
```

Out[8]:

	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
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

In [11]:

```
insurance_df.dtypes
```

Out[11]:

```
age           int64
sex          object
bmi         float64
children      int64
smoker       object
region       object
charges     float64
dtype: object
```

Basic EDA

1.Find the shape of the data,data type of individual columns

2.Check the presence of missing values

3.Descriptive stats of numerical columns

4.Find the distribution of numerical columns and the asssociated skeweness and presence of outliers

5.Distribution of categorical columns

In [12]:

```
#Info about the data

insurance_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       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
```

The data has 1338 instances with 7 attributes. 2 integer type, 2 float type and 3 object type(Strings in the column)

In [13]:

```
#Check for the null values

insurance_df.isna().apply(pd.value_counts)
```

Out[13]:

	age	sex	bmi	children	smoker	region	charges
False	1338	1338	1338	1338	1338	1338	1338

No null values in the dataset

In [14]:

```
#Five point summary for the dataset

insurance_df.describe().T
```

Out[14]:

	count	mean	std	min	25%	50%	75%	
age	1338.0	39.207025	14.049960	18.0000	27.00000	39.000	51.000000	
bmi	1338.0	30.663397	6.098187	15.9600	26.29625	30.400	34.693750	
children	1338.0	1.094918	1.205493	0.0000	0.00000	1.000	2.000000	
charges	1338.0	13270.422265	12110.011237	1121.8739	4740.28715	9382.033	16639.912515	63

What We Found????

1)Data looks legit as all the statistics seem reasonable

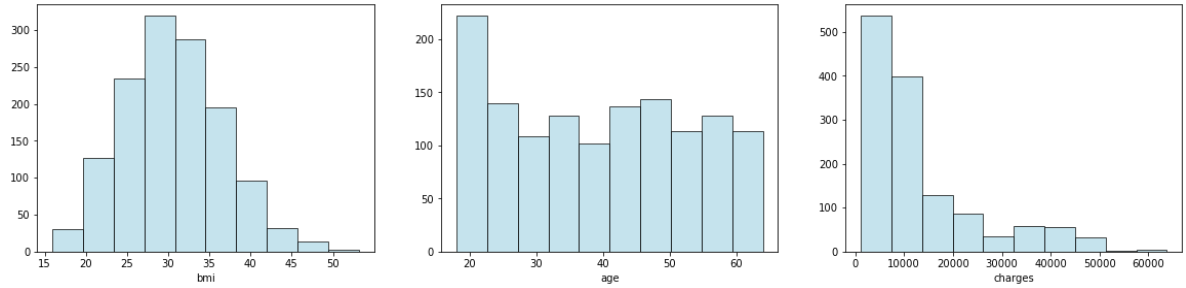
2)Looking at the age column, data looks representative of the true age distribution of the adult population.

3)Very few people have more than 2 children. 75% of the people have 2 or less children.

4)The claimed amount is higly skewed as most people would require basic medi-care and only few suffer from diseases which cost more to get rid of.

In [17]:

```
#Plots to see the distribution of the continuous features individually

plt.figure(figsize= (20,15))
plt.subplot(3,3,1)
plt.hist(insurance_df.bmi, color='lightblue', edgecolor = 'black', alpha = 0.7)
plt.xlabel('bmi')

plt.subplot(3,3,2)
plt.hist(insurance_df.age, color='lightblue', edgecolor = 'black', alpha = 0.7)
plt.xlabel('age')


plt.subplot(3,3,3)
plt.hist(insurance_df.charges, color='lightblue', edgecolor = 'black', alpha = 0.7)
plt.xlabel('charges')

plt.show()
```



- bmi looks quiet normally distributed
- Age seems be be distributed quiet uniformly¶
- As seen in the previous step, charges are highly skewed

In [18]:

```
Skewness = pd.DataFrame({'Skewness' : [stats.skew(insurance_df.bmi),stats.skew(insurance_df
                        index=['bmi','age','charges'])  # Measure the skeweness of the requ
Skewness
```

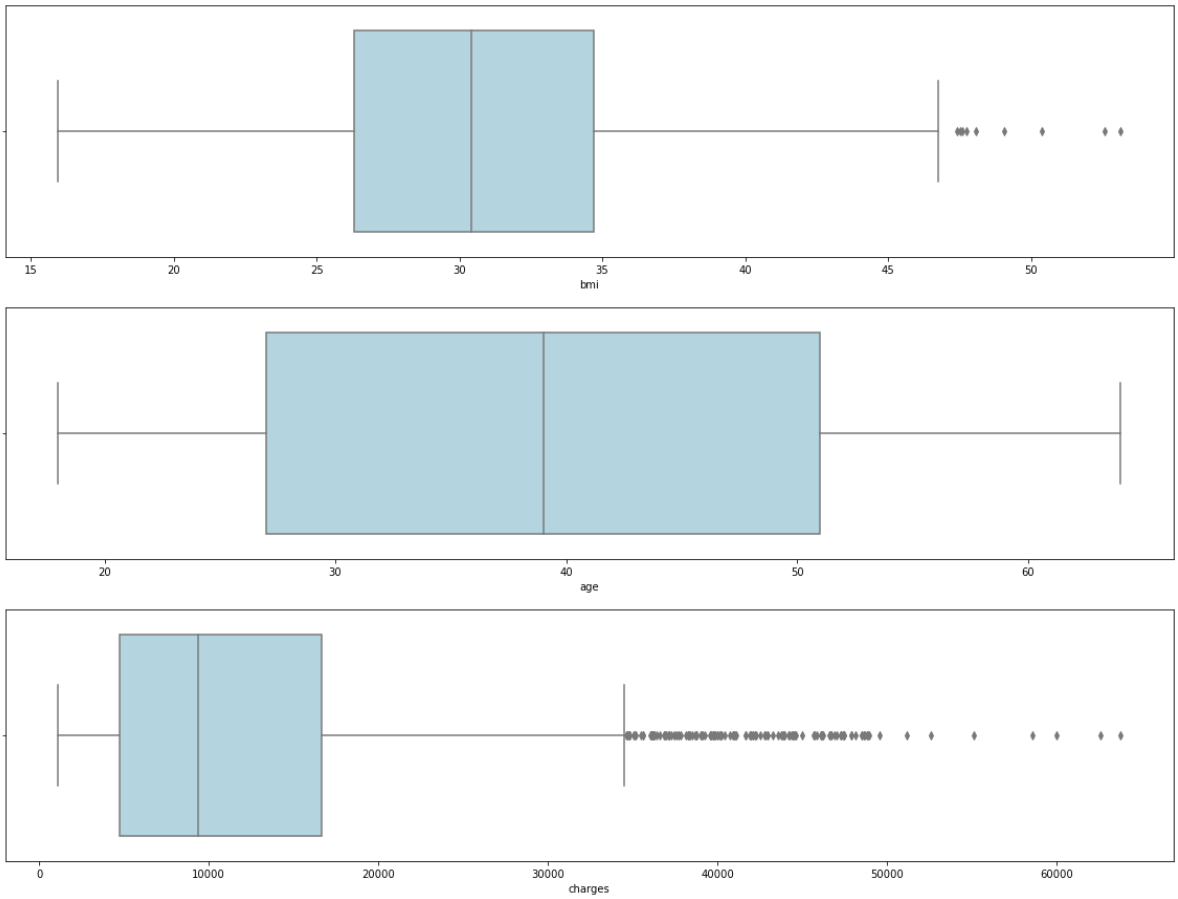
Out[18]:

	Skewness
bmi	0.283729
age	0.055610
charges	1.514180

- Skew of bmi is very less as seen in the previous step
- age is uniformly distributed and there's hardly any skew
- charges are highly skewed

In [19]:

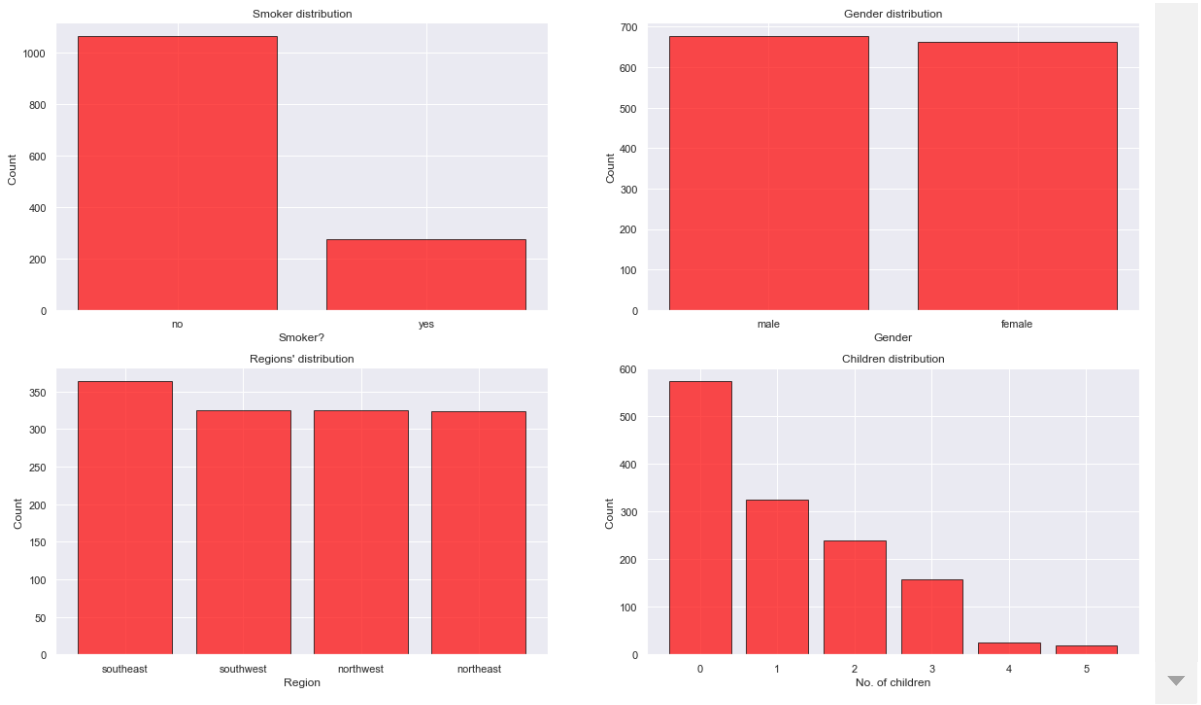
```
#Checking for the outliers
plt.figure(figsize= (20,15))
plt.subplot(3,1,1)
sns.boxplot(x= insurance_df.bmi, color='lightblue')

plt.subplot(3,1,2)
sns.boxplot(x= insurance_df.age, color='lightblue')

plt.subplot(3,1,3)
sns.boxplot(x= insurance_df.charges, color='lightblue')

plt.show()
```



- bmi has a few extreme values
- charges as it is highly skewed, there are quiet a lot of extreme values

In [19]:

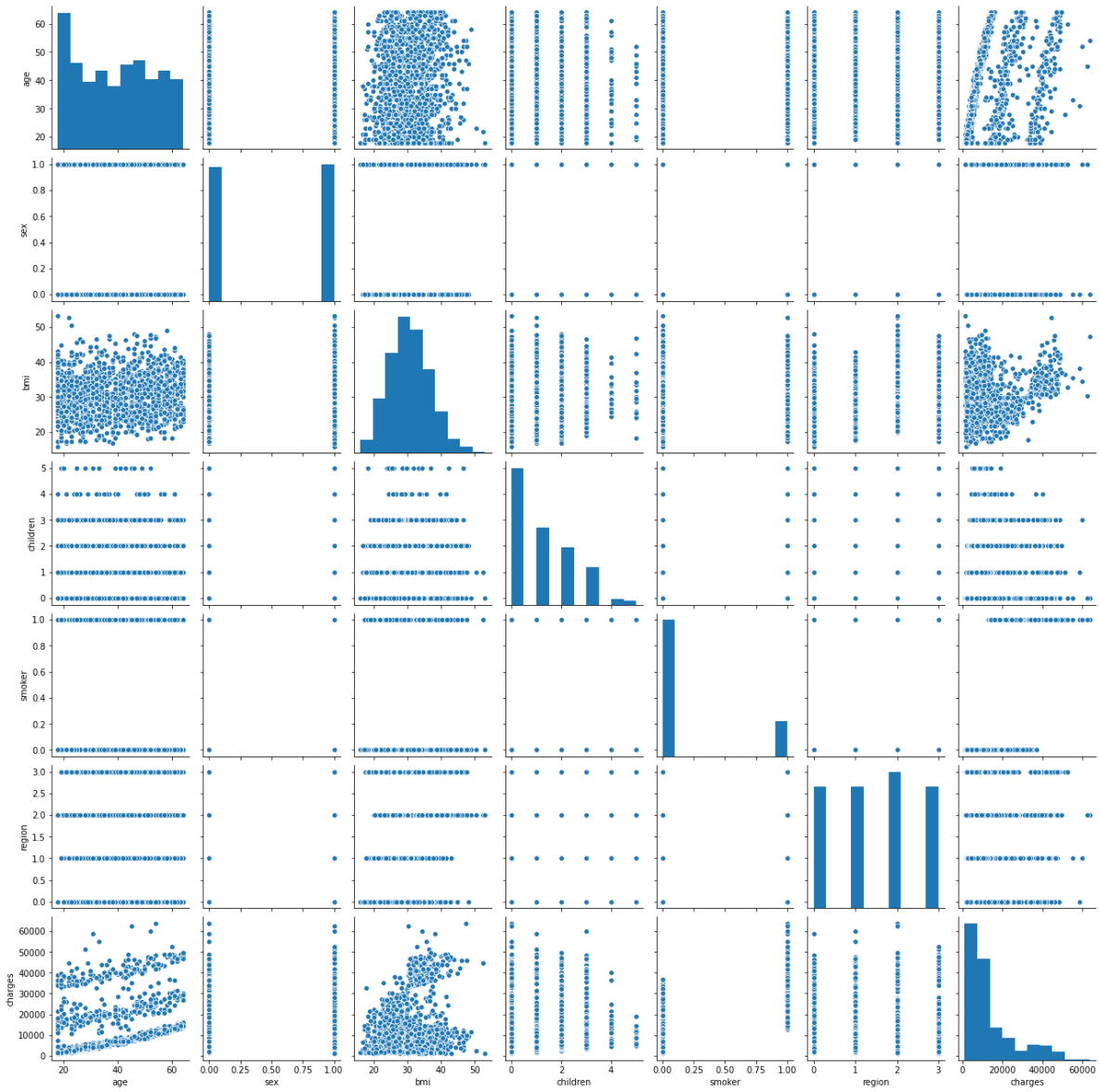
```
plt.figure(figsize=(20,25))


x = insurance_df.smoker.value_counts().index    #Values for x-axis
y = [insurance_df['smoker'].value_counts()[i] for i in x]   # Count of each class on y-axis

plt.subplot(4,2,1)
plt.bar(x,y, align='center',color = 'red',edgecolor = 'black',alpha = 0.7)  #plot a bar cha
plt.xlabel('Smoker?')
plt.ylabel('Count ')
plt.title('Smoker distribution')

x1 = insurance_df.sex.value_counts().index    #Values for x-axis
y1 = [insurance_df['sex'].value_counts()[j] for j in x1]   # Count of each class on y-axis

plt.subplot(4,2,2)
plt.bar(x1,y1, align='center',color = 'red',edgecolor = 'black',alpha = 0.7)  #plot a bar c
plt.xlabel('Gender')
plt.ylabel('Count')
plt.title('Gender distribution')

x2 = insurance_df.region.value_counts().index    #Values for x-axis
y2 = [insurance_df['region'].value_counts()[k] for k in x2]   # Count of each class on y-ax

plt.subplot(4,2,3)
plt.bar(x2,y2, align='center',color = 'red',edgecolor = 'black',alpha = 0.7)  #plot a bar c
plt.xlabel('Region')
plt.ylabel('Count ')
plt.title("Regions' distribution")

x3 = insurance_df.children.value_counts().index    #Values for x-axis
y3 = [insurance_df['children'].value_counts()[l] for l in x3]   # Count of each class on y-

plt.subplot(4,2,4)
plt.bar(x3,y3, align='center',color = 'red',edgecolor = 'black',alpha = 0.7)  #plot a bar c
plt.xlabel('No. of children')
plt.ylabel('Count ')
plt.title("Children distribution")

plt.show()
```



- There are a lot more non-smokers than there are smokers in the data
- Instances are distributed evenly accross all regions
- Gender is also distributed evenly
- Most instances have less than 2 children and very few have 4 or 5 children Bi-variate distribution of every possible attribute pair

In [20]:

```
#Label encoding the variables before doing a pairplot because pairplot ignores strings
insurance_df_encoded = copy.deepcopy(insurance_df)
insurance_df_encoded.loc[:,['sex', 'smoker', 'region']] = insurance_df_encoded.loc[:,['sex'

sns.pairplot(insurance_df_encoded)  #pairplot
plt.show()
```



- The only obvious correlation of 'charges' is with 'smoker'
- Looks like smokers claimed more money than non-smokers
- There's an interesting pattern between 'age' and 'charges. Could be because for the same ailment, older people are charged more than the younger ones

Do charges of people who smoke differ significantly from the people who don't?

In [21]:

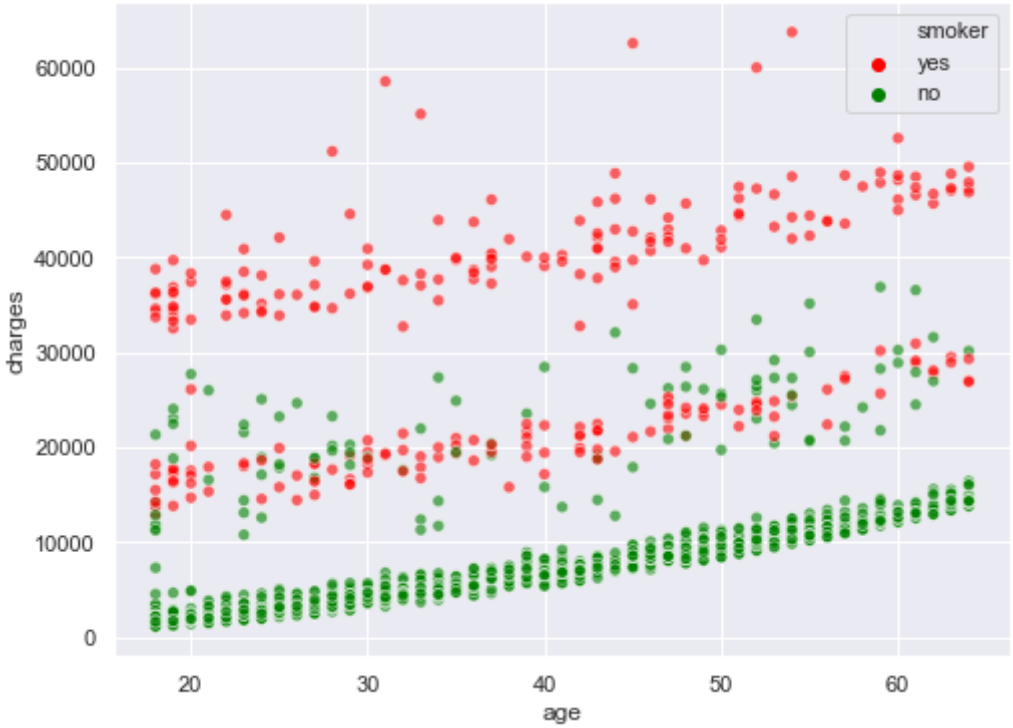
```
insurance_df.smoker.value_counts()
```

Out[21]:

```
no     1064
yes     274
Name: smoker, dtype: int64
```

In [23]:

```
#Scatter plot to look for visual evidence of dependency between attributes smoker and charg
plt.figure(figsize=(8,6))
sns.scatterplot(insurance_df.age, insurance_df.charges,hue=insurance_df.smoker,palette= ['r
plt.show()
```



Visually the difference between charges of smokers and charges of non-smokers is apparent

In [24]:

```
# T-test to check dependency of smoking on charges
Ho = "Charges of smoker and non-smoker are same"   # Stating the Null Hypothesis
Ha = "Charges of smoker and non-smoker are not the same"   # Stating the Alternate Hypothes

x = np.array(insurance_df[insurance_df.smoker == 'yes'].charges)  # Selecting charges corre
y = np.array(insurance_df[insurance_df.smoker == 'no'].charges) # Selecting charges corresp

t, p_value  = stats.ttest_ind(x,y, axis = 0)  #Performing an Independent t-test

if p_value < 0.05:  # Setting our significance level at 5%
    print(f'{Ha} as the p_value ({p_value}) < 0.05')
else:
    print(f'{Ho} as the p_value ({p_value}) > 0.05')
```

```
Charges of smoker and non-smoker are not the same as the p_value (8.27143584
2177219e-283) < 0.05
```

Smokers seem to claim significantly more money than non-smokers

Does bmi of males differ significantly from that of females?

In [25]:

```
insurance_df.sex.value_counts()   #Checking the distribution of males and females
```

Out[25]:

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

In [26]:

```
plt.figure(figsize=(8,6))
sns.scatterplot(insurance_df.age, insurance_df.charges,hue=insurance_df.sex,palette= ['pink
plt.show()
```

- Visually, there is no apparent relation between gender and charges

In [27]:

```
# T-test to check dependency of bmi on gender
Ho = "Gender has no effect on bmi"   # Stating the Null Hypothesis
Ha = "Gender has an effect on bmi"   # Stating the Alternate Hypothesis

x = np.array(insurance_df[insurance_df.sex == 'male'].bmi)  # Selecting bmi values correspo
y = np.array(insurance_df[insurance_df.sex == 'female'].bmi) # Selecting bmi values corresp

t, p_value  = stats.ttest_ind(x,y, axis = 0)  #Performing an Independent t-test

if p_value < 0.05:  # Setting our significance level at 5%
    print(f'{Ha} as the p_value ({p_value.round()}) < 0.05')
else:
    print(f'{Ho} as the p_value ({p_value.round(3)}) > 0.05')
```

```
Gender has no effect on bmi as the p_value (0.09) > 0.05
```

bmi of both the genders are identical

Is the proportion of smokers significantly different in different genders?

In [28]:

```
# Chi_square test to check if smoking habits are different for different genders
Ho = "Gender has no effect on smoking habits"   # Stating the Null Hypothesis
Ha = "Gender has an effect on smoking habits"   # Stating the Alternate Hypothesis

crosstab = pd.crosstab(insurance_df['sex'],insurance_df['smoker'])  # Contingency table of

chi, p_value, dof, expected =  stats.chi2_contingency(crosstab)

if p_value < 0.05:  # Setting our significance level at 5%
    print(f'{Ha} as the p_value ({p_value.round(3)}) < 0.05')
else:
    print(f'{Ho} as the p_value ({p_value.round(3)}) > 0.05')
crosstab
```

```
Gender has an effect on smoking habits as the p_value (0.007) < 0.05
```

Out[28]:

smoker	no	yes
sex		
female	547	115
male	517	159

** Proportion of smokers in males is significantly different from that of the females**

In [29]:

```
# Chi_square test to check if smoking habits are different for people of different regions
Ho = "Region has no effect on smoking habits"   # Stating the Null Hypothesis
Ha = "Region has an effect on smoking habits"   # Stating the Alternate Hypothesis

crosstab = pd.crosstab(insurance_df['smoker'], insurance_df['region'])  # Contingency table

chi, p_value, dof, expected =  stats.chi2_contingency(crosstab)

if p_value < 0.05:  # Setting our significance level at 5%
    print(f'{Ha} as the p_value ({p_value.round(3)}) < 0.05')
else:
    print(f'{Ho} as the p_value ({p_value.round(3)}) > 0.05')
crosstab
```

```
Region has no effect on smoking habits as the p_value (0.062) > 0.05
```

Out[29]:

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

**** Smoking habbits of people of different regions are similar (http:)***

Is the distribution of bmi across women with no children, one child and two children, the same ?

In [30]:

```
# Test to see if the distributions of bmi values for females having different number of chi

Ho = "No. of children has no effect on bmi"   # Stating the Null Hypothesis
Ha = "No. of children has an effect on bmi"   # Stating the Alternate Hypothesis


female_df = copy.deepcopy(insurance_df[insurance_df['sex'] == 'female'])

zero = female_df[female_df.children == 0]['bmi']
one = female_df[female_df.children == 1]['bmi']
two = female_df[female_df.children == 2]['bmi']


f_stat, p_value = stats.f_oneway(zero,one,two)


if p_value < 0.05:  # Setting our significance level at 5%
    print(f'{Ha} as the p_value ({p_value.round(3)}) < 0.05')
else:
    print(f'{Ho} as the p_value ({p_value.round(3)}) > 0.05')
```

```
No. of children has no effect on bmi as the p_value (0.716) > 0.05
```

BMI is not changed by the number of children a women has