**Ans 1.**

import numpy as np

import pandas as pd

from matplotlib import pyplot as plt

import seaborn as sns

from sklearn.preprocessing import LabelEncoder

import statsmodels.api as sm

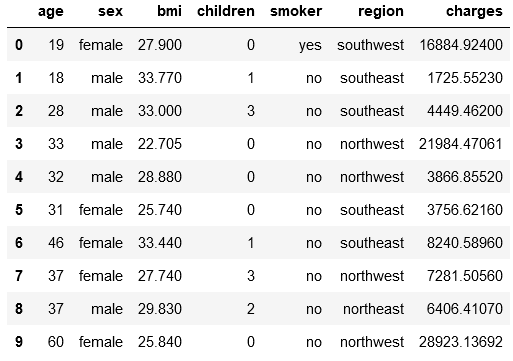
import scipy.stats as stats

import copy

**Ans 2.**

Insu\_df=pd.read\_csv('Insurance.csv')

Insu\_df.head(10)



**Ans 3.**

**a.**

#a. Shape of the data

Insu\_df.shape

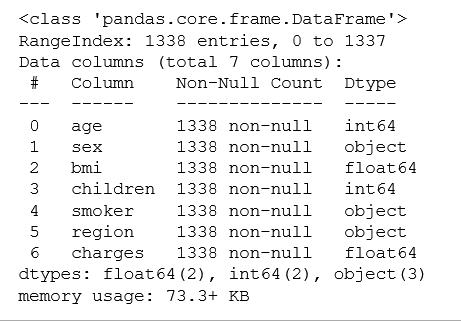


There are 1338 rows and 7 columns.

**b.**

#b. Data type of each attribute

Insu\_df.info()

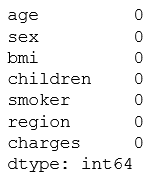


* The columns 'sex', 'smoker' and 'region' are of type object. Rest all other attributes are of int and float type. No column has any null or missing values.

**c.**

#c. Checking the presence of missing values

Insu\_df.isnull().sum()

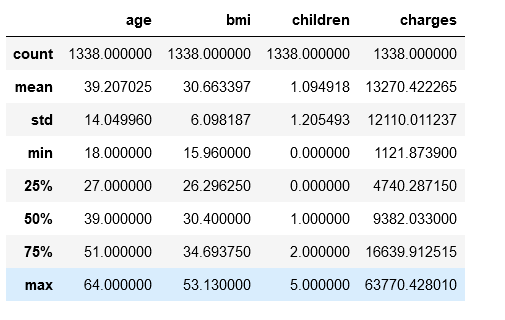


* No missing values are present in the dataframe.

**d.**

#d 5 point summary of numerical attributes

Insu\_df.describe()

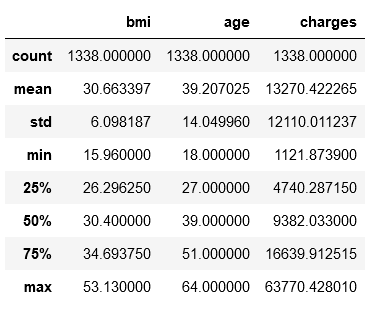


* 75% of the population has two or less than two children and the number of outliers in charges is a huge number.

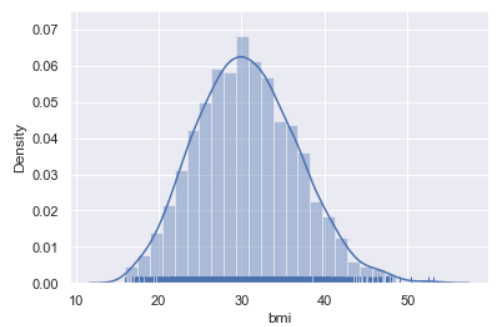
**e.**

#e. Distribution of ‘bmi’, ‘age’ and ‘charges’ columns

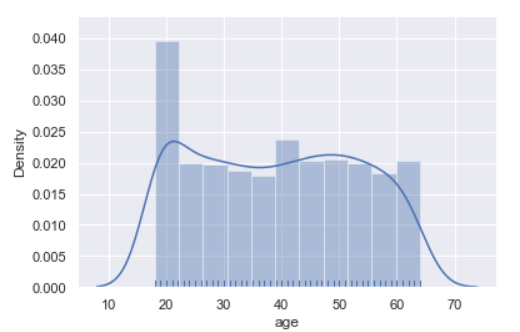
Insu\_df[['bmi','age','charges']].describe()



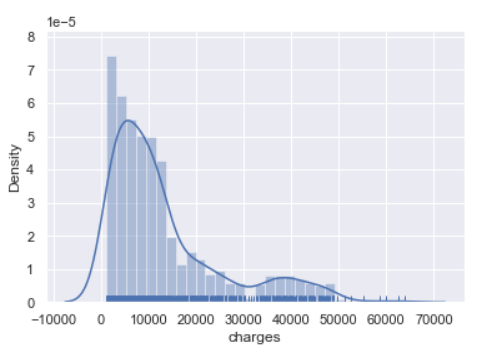
sns.distplot(Insu\_df['bmi'], kde=True, rug=True)



sns.distplot(Insu\_df['age'], kde=True, rug=True)



sns.distplot(Insu\_df['charges'], kde=True, rug=True)

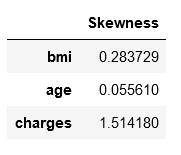


* bmi is normally distributed(approx).
* Age is quiet uniformly distributed.
* Charges are highly skewed towards the left.

**f.**

#f. Measure of skewness of ‘bmi’, ‘age’ and ‘charges’ columns

Skewness = pd.DataFrame({'Skewness' : [stats.skew(Insu\_df.bmi), stats.skew(Insu\_df.age),stats.skew(Insu\_df.charges)]}, index=['bmi','age','charges'])



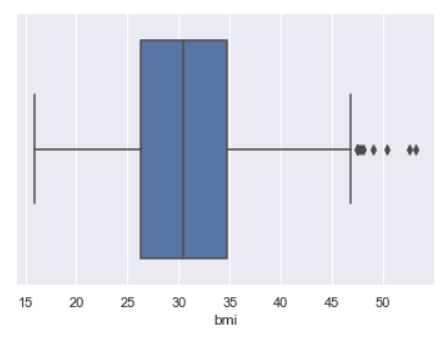
* Skew of bmi is very less thus it is more like normally distributed.
* Age is uniformly distributed and there's hardly any skew.
* Charges are highly skewed.

**g.**

#g.Checking the presence of outliers in ‘bmi’, ‘age’ and ‘charges columns

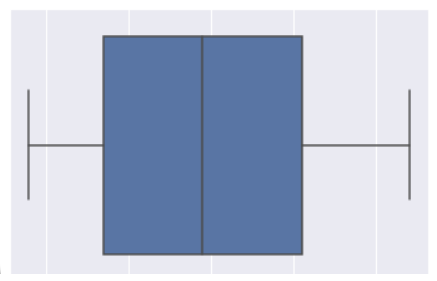
sns.set(style="darkgrid")

sns.boxplot(Insu\_df["bmi"])



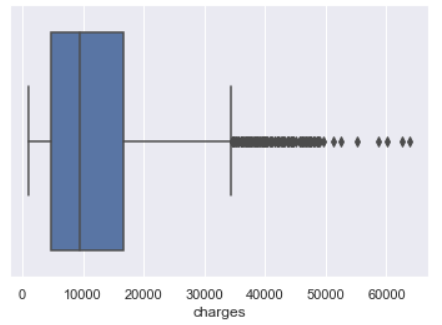
* Bmi has outliers.

sns.boxplot(Insu\_df["age"])



* Age doesn’t have any outliers.

sns.boxplot(Insu\_df["charges"])



* Charges has a large number of outliers.

iqr\_bmi = np.subtract(\*np.percentile(Insu\_df['bmi'], [75, 25]))

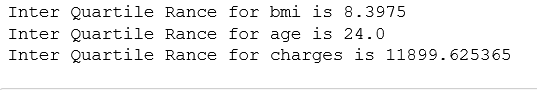
print("Inter Quartile Rance for bmi is",iqr\_bmi)

iqr\_age = np.subtract(\*np.percentile(Insu\_df['age'], [75, 25]))

print("Inter Quartile Rance for age is",iqr\_age)

iqr\_charges = np.subtract(\*np.percentile(Insu\_df['charges'], [75, 25]))

print("Inter Quartile Rance for charges is",iqr\_charges)



* Thus we can see that the inter quartile range for charges is very large.

# identify outliers for bmi

q25 = np.percentile(Insu\_df['bmi'], 25)

q75 = np.percentile(Insu\_df['bmi'], 75)

iqr = q75 - q25

cut\_off = iqr \* 1.5

lower = q25 - cut\_off

upper = q75 + cut\_off

outliers = [x for x in Insu\_df['bmi'] if x < lower or x > upper]

print('The total outliers for bmi out of 1338 records are %d' % len(outliers))



# identify outliers for age

q25 = np.percentile(Insu\_df['age'], 25)

q75 = np.percentile(Insu\_df['age'], 75)

iqr = q75 - q25

cut\_off = iqr \* 1.5

lower = q25 - cut\_off

upper = q75 + cut\_off

outliers = [x for x in Insu\_df['age'] if x < lower or x > upper]

print('The total outliers for age out of 1338 records are %d' % len(outliers))



# identify outliers for charges

q25 = np.percentile(Insu\_df['charges'], 25)

q75 = np.percentile(Insu\_df['charges'], 75)

iqr = q75 - q25

cut\_off = iqr \* 1.5

lower = q25 - cut\_off

upper = q75 + cut\_off

outliers = [x for x in Insu\_df['charges'] if x < lower or x > upper]

print('The total outliers for charges out of 1338 records are %d' % len(outliers))

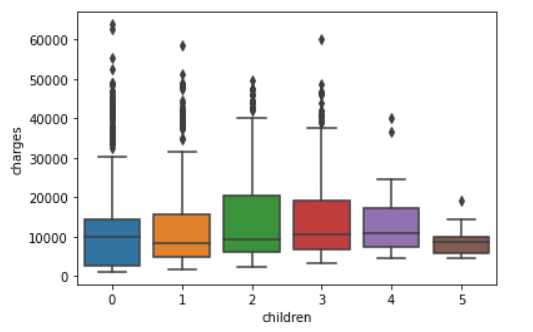


**h.**

#h. Distribution of categorical columns (include children)

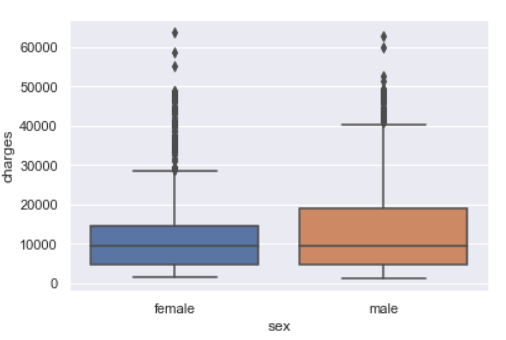
#sns.histplot(x='children', y='charges', data= Insu\_df,kde=True)

sns.boxplot(x='children', y='charges', data= Insu\_df)



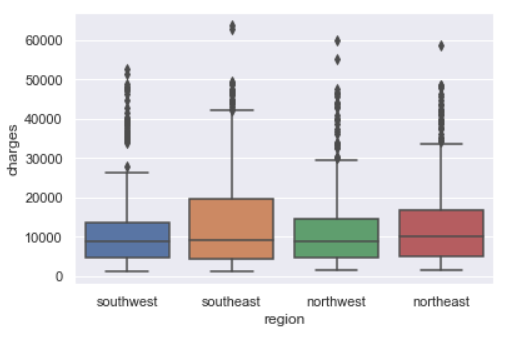
* Thus we can see that the charges paid by customers having no children is the highest and customers having four or more children is very less.

sns.boxplot(x='sex', y='charges', data= Insu\_df)



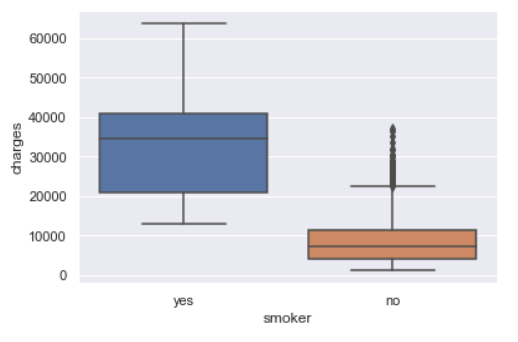
* Female has more outliers while males have a right skew telling more of them pay higher charges.

sns.boxplot(x='region', y='charges', data= Insu\_df)



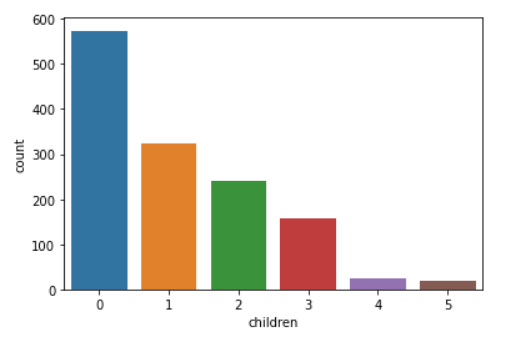
* Charges paid by the customers from southeast are the highest and those by from the southwest is the lowest. Also, every region has some outliers.

sns.boxplot(x='smoker', y='charges', data= Insu\_df)



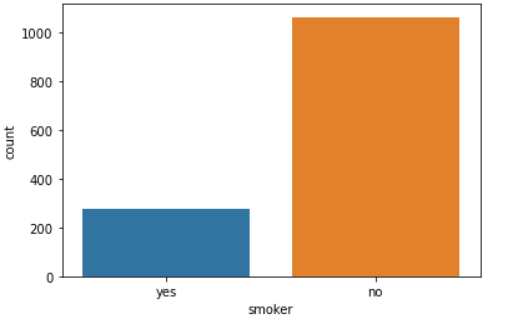
* Smokers pay higher medical costs billed by health insurance than the non-smokers. However, there are some outliers exists in the nonsmoker who pay higher charges.

sns.countplot(Insu\_df['children'])



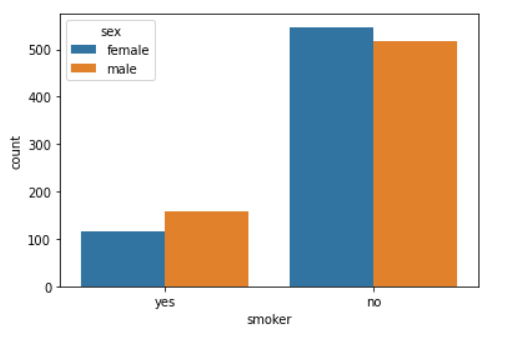
* The number of customers having no children is higher.

sns.countplot(Insu\_df['smoker'])



* Larger population of customers don’t smoke.

sns.countplot(Insu\_df['smoker'],hue = Insu\_df['sex'])



**i.**

#i. Pair plot that includes all the columns of the data frame

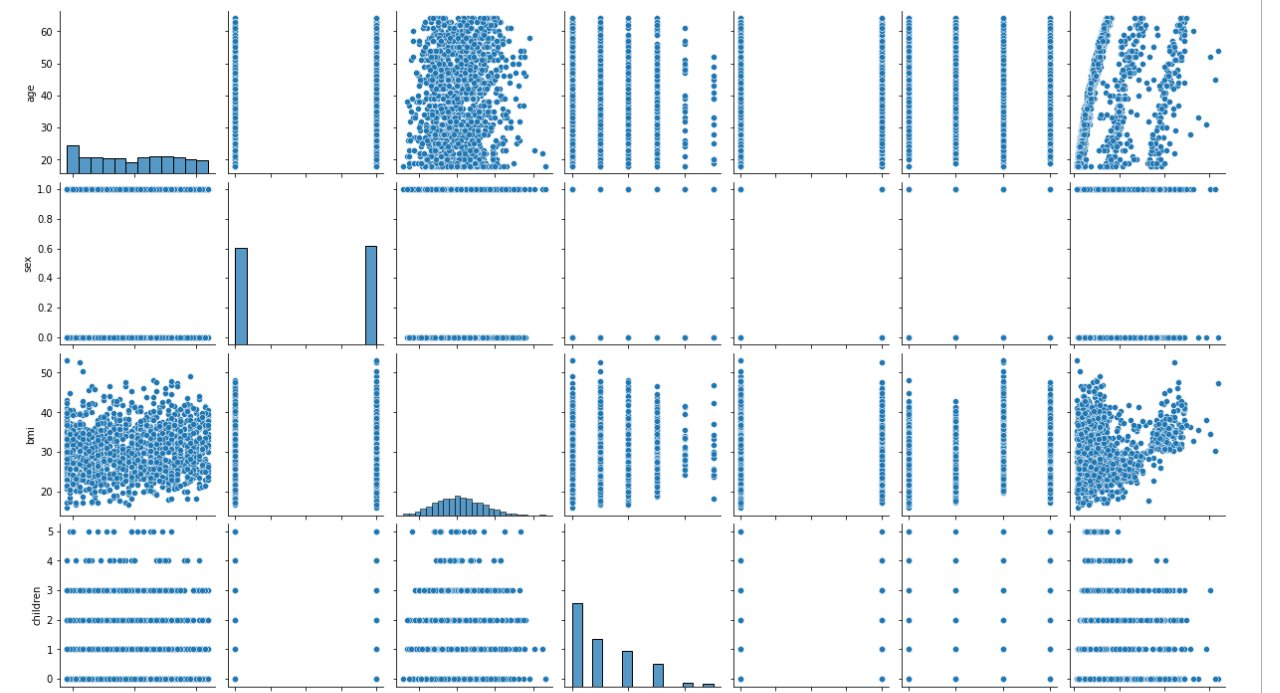
#Encoding the variables before doing a pairplot because pairplot ignores string datatype

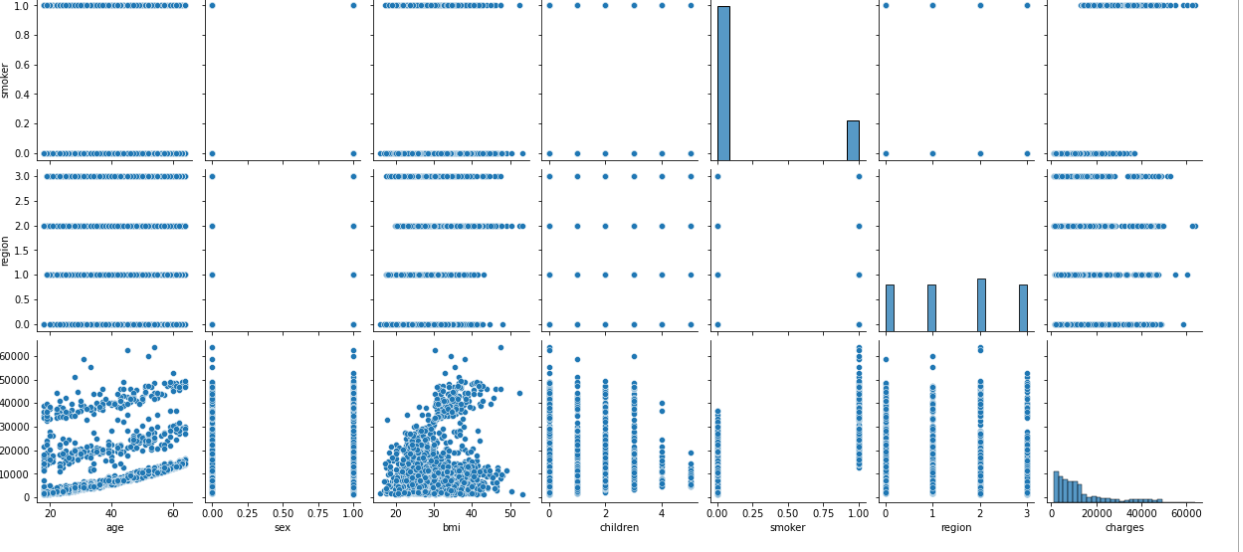
insu\_df\_encoded = copy.deepcopy(Insu\_df)

insu\_df\_encoded.loc[:,['sex', 'smoker', 'region']] = Insu\_df.loc[:,['sex', 'smoker', 'region']].apply(LabelEncoder().fit\_transform)

sns.pairplot(insu\_df\_encoded)

plt.show()





* From the plot of age vs charges older people tend to pay higher than younger people.

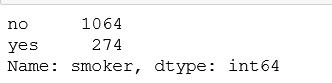
**Ques 4.**

#Ques-4

**a.**

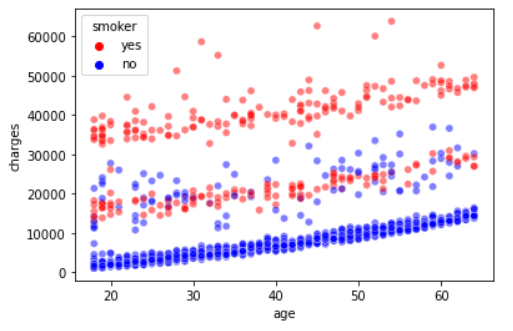
#a) Do charges of people who smoke differ significantly from the people who don't?

Insu\_df.smoker.value\_counts()



sns.scatterplot(Insu\_df.age,Insu\_df.charges,hue=Insu\_df.smoker,palette= ['red','blue'] ,alpha=0.5)

plt.show()



* It can be seen that smokers are charged more than non-smokers.

#Applying T-test to determine the impact of smoking on the charges.

Ho = "Charges of smokers and non-smokers are the same"

Ha = "Charges of smokers and non-smokers are not the same"

# Selecting charges corresponding to smokers as an array

x = np.array(Insu\_df[Insu\_df.smoker == 'yes'].charges)

# Selecting charges corresponding to non-smokers as an array

y = np.array(Insu\_df[Insu\_df.smoker == 'no'].charges)

#Performing an Independent t-test

t, p\_value = stats.ttest\_ind(x,y, axis = 0)

print(p\_value)

#Rejecting the null hypothesis as the p\_value is lesser than 0.05.

#This tells us that the paid charges by the smokers and non-smokers is different.

#Smokers pay higher charges in comparison to the non-smokers.

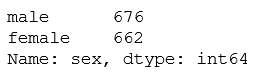


* It can be seen that we can reject the null hypothesis as p value is lesser than 0.05.
* And smokers pay a higher charge than non-smokers.

**b.**

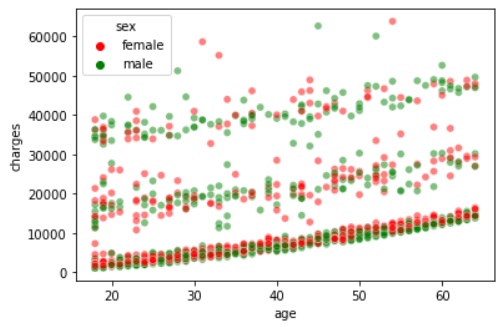
#b Does bmi of males differ significantly from that of females?

Insu\_df.sex.value\_counts()



sns.scatterplot(Insu\_df.age,Insu\_df.charges,hue=Insu\_df.sex,palette= ['red','green'] ,alpha=0.5 )

plt.show()



* Nothing can be inferred from the scatter plot.

#Applying T-test to determine the impact of Gender on the bmi.

Ho = "Gender has no impact on bmi"

Ha = "Gender has an impact on bmi"

x = np.array(Insu\_df[Insu\_df.sex == 'male'].bmi)

y = np.array(Insu\_df[Insu\_df.sex == 'female'].bmi)

t, p\_value = stats.ttest\_ind(x,y, axis = 0)

print(p\_value)

#Accepting the null hypothesis as pvalue is greater than 0.05. Hence we can say that Gender has no impact on the bmi.



* We fail to reject the null hypothesis this could mean that the bmi of males and females doesn’t differ significantly.

**c.**

#c Is the proportion of smokers significantly different in different genders?

# We will perform Chi\_square test to check if the proportion of smokers differs as per gender.

Ho = "Gender has no effect on smoking habits"

Ha = "Gender has an effect on smoking habits"

crosstab = pd.crosstab(Insu\_df['sex'],Insu\_df['smoker'])

chi, p\_value, dof,

expected = stats.chi2\_contingency(crosstab)

print(p\_value)



* We can reject the null hypothesis and say that smoking habits differ with genders.

**d.**

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

# Applying anova test to check the proportion as there are more than two cases involved

Ho = "No. of children has no effect on the bmi"

Ha = "No. of children has an effect on the bmi"

newdf = copy.deepcopy(Insu\_df[Insu\_df['sex'] == 'female'])

zero = newdf[newdf.children == 0]['bmi']

one = newdf[newdf.children == 1]['bmi']

two = newdf[newdf.children == 2]['bmi']

f\_stat, p\_value = stats.f\_oneway(zero,one,two)

print(p\_value)



* We fail to reject the null hypothesis thus concluding that no. of children has no effect on bmi.

**By- Shantanu Singh**