```
import pandas as pd
rv=pd.read_csv('/content/drive/MyDrive/Dataset/new_insurance_data.csv')
```

```
rv.head(3)
```

	age	sex	bmi	children	smoker	Claim_Amount	past_consultations	num_of_steps	Hospital_expenditure	NUmber_of_past_hospitaliz
0	18.0	male	23.21	0.0	no	29087.54313	17.0	715428.0	4720920.992	
1	18.0	male	30.14	0.0	no	39053.67437	7.0	699157.0	4329831.676	
2	18.0	male	33.33	0.0	no	39023.62759	19.0	702341.0	6884860.774	

```
rv.tail()
```

	age	sex	bmi	children	smoker	Claim_Amount	past_consultations	num_of_steps	Hospital_expenditure	NUmber_of_past_ho
1333	33.0	female	35.530	0.0	yes	63142.25346	32.0	1091267.0	170380500.5	
1334	31.0	female	38.095	1.0	yes	43419.95227	31.0	1107872.0	201515184.8	
1335	52.0	male	34.485	3.0	yes	52458.92353	25.0	1092005.0	223644981.3	
1336	45.0	male	30.360	0.0	yes	69927.51664	34.0	1106821.0	252892382.6	
1337	54.0	female	47.410	0.0	yes	63982.80926	31.0	1100328.0	261631699.3	

```
rv.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 13 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   age                             1329 non-null   float64
 1   sex                             1338 non-null   object
 2   bmi                             1335 non-null   float64
 3   children                        1333 non-null   float64
 4   smoker                          1338 non-null   object
 5   Claim_Amount                    1324 non-null   float64
 6   past_consultations              1332 non-null   float64
 7   num_of_steps                    1335 non-null   float64
 8   Hospital_expenditure            1334 non-null   float64
 9   NUmber_of_past_hospitalizations  1336 non-null   float64
 10  Anual_Salary                    1332 non-null   float64
 11  region                          1338 non-null   object
 12  charges                         1338 non-null   float64
dtypes: float64(10), object(3)
memory usage: 136.0+ KB
```

```
rv.describe()
```

	age	bmi	children	Claim_Amount	past_consultations	num_of_steps	Hospital_expenditure	NUmber_of_past_h
count	1329.000000	1335.000000	1333.000000	1324.000000	1332.000000	1.335000e+03	1.334000e+03	
mean	39.310008	30.665112	1.090773	33361.327180	15.216216	9.100047e+05	1.584179e+07	
std	14.034818	6.101690	1.201856	15617.288337	7.467723	9.188612e+04	2.669305e+07	
min	18.000000	15.960000	0.000000	1920.136268	1.000000	6.954300e+05	2.945253e+04	
25%	27.000000	26.302500	0.000000	20768.860390	9.000000	8.471995e+05	4.077633e+06	
50%	39.000000	30.400000	1.000000	33700.310675	15.000000	9.143000e+05	7.490337e+06	
75%	51.000000	34.687500	2.000000	45052.331957	20.000000	9.716840e+05	1.084082e+07	
max	64.000000	53.130000	5.000000	77277.988480	40.000000	1.107872e+06	2.616317e+08	

```
rv.shape
```

```
(1338, 13)
```

```
rv.isnull()
```

	age	sex	bmi	children	smoker	Claim_Amount	past_consultations	num_of_steps	Hospital_expenditure	NUmber_of_past_hos
0	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	
1333	False	False	False	False	False	False	False	False	False	
1334	False	False	False	False	False	False	False	False	False	
1335	False	False	False	False	False	False	False	False	False	
1336	False	False	False	False	False	False	False	False	False	
1337	False	False	False	False	False	False	False	False	False	

1338 rows × 13 columns

Double-click (or enter) to edit

```
col_list=['bmi','past_consultations','Hospital_expenditure','NUmber_of_past_hospitalizations','Anual_Salary']
col_list
```

```
['bmi',
 'past_consultations',
 'Hospital_expenditure',
 'NUmber_of_past_hospitalizations',
 'Anual_Salary']
```

```
rv.columns.value_counts()
```

	count
age	1
sex	1
bmi	1
children	1
smoker	1
Claim_Amount	1
past_consultations	1
num_of_steps	1
Hospital_expenditure	1
NUmber_of_past_hospitalizations	1
Anual_Salary	1
region	1
charges	1

dtype: int64

```
rv.columns
```

```
Index(['age', 'sex', 'bmi', 'children', 'smoker', 'Claim_Amount',
       'past_consultations', 'num_of_steps', 'Hospital_expenditure',
       'NUmber_of_past_hospitalizations', 'Anual_Salary', 'region', 'charges'],
      dtype='object')
```

```
for i in rv.columns:
  if rv[i].dtype=='object':
    rv[i]=rv[i].fillna(rv[i].mode()[0])
```

```
  else:
    rv[i]=rv[i].fillna(rv[i].mean())
```

```
rv.info()
```

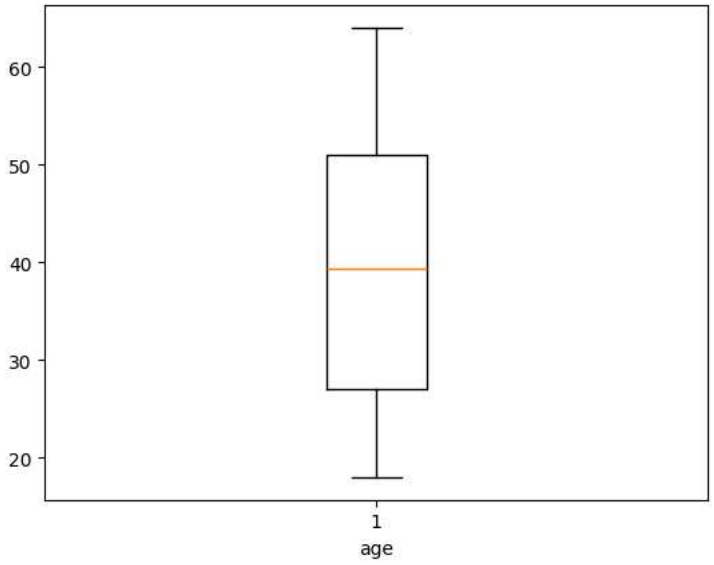
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 13 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   age                             1338 non-null   float64
 1   sex                             1338 non-null   object 
 2   bmi                             1338 non-null   float64
 3   children                        1338 non-null   float64
 4   smoker                          1338 non-null   object 
 5   Claim_Amount                    1338 non-null   float64
 6   past_consultations              1338 non-null   float64
 7   num_of_steps                    1338 non-null   float64
 8   Hospital_expenditure            1338 non-null   float64
 9   NUmber_of_past_hospitalizations 1338 non-null   float64
 10  Anual_Salary                    1338 non-null   float64
 11  region                          1338 non-null   object 
 12  charges                         1338 non-null   float64
dtypes: float64(10), object(3)
memory usage: 136.0+ KB
```

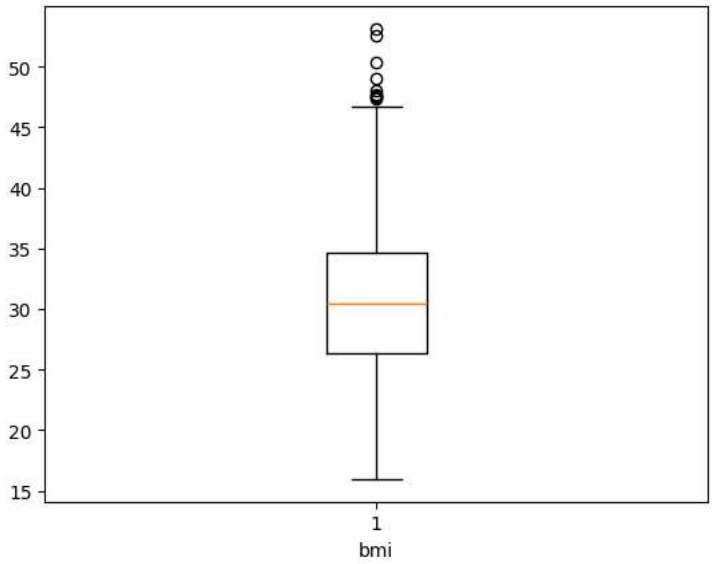
```
rv.isnull().sum()
```

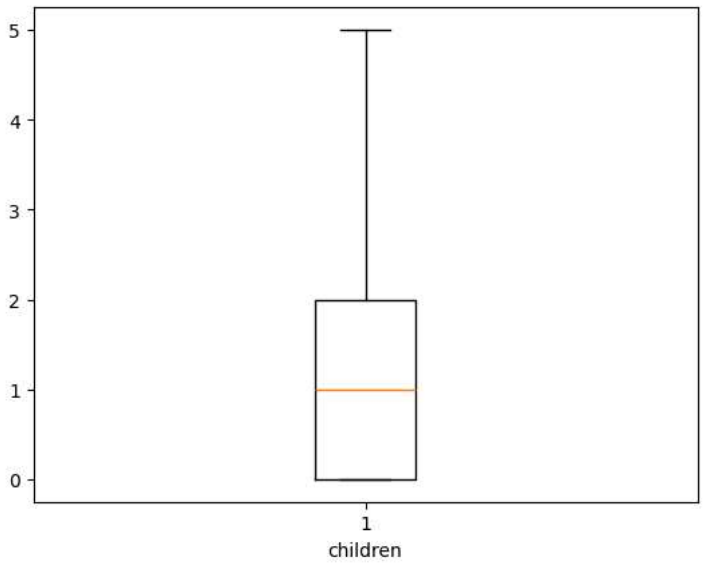
	0
age	0
sex	0
bmi	0
children	0
smoker	0
Claim_Amount	0
past_consultations	0
num_of_steps	0
Hospital_expenditure	0
NUmber_of_past_hospitalizations	0
Anual_Salary	0
region	0
charges	0

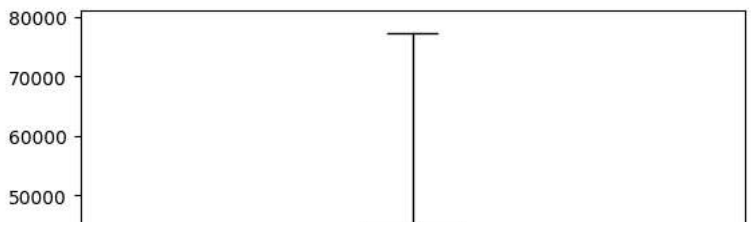
dtype: int64

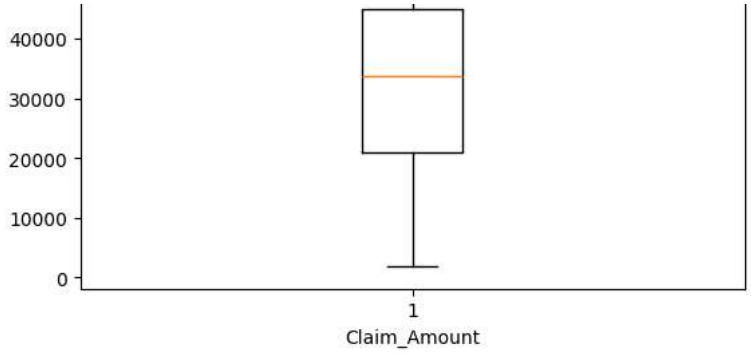
```
import matplotlib.pyplot as plt
for i in rv.columns:
  if(rv[i].dtype !='object')and (i !='charges'):
    plt.boxplot(rv[i])
    plt.xlabel(i)
    plt.show()
```

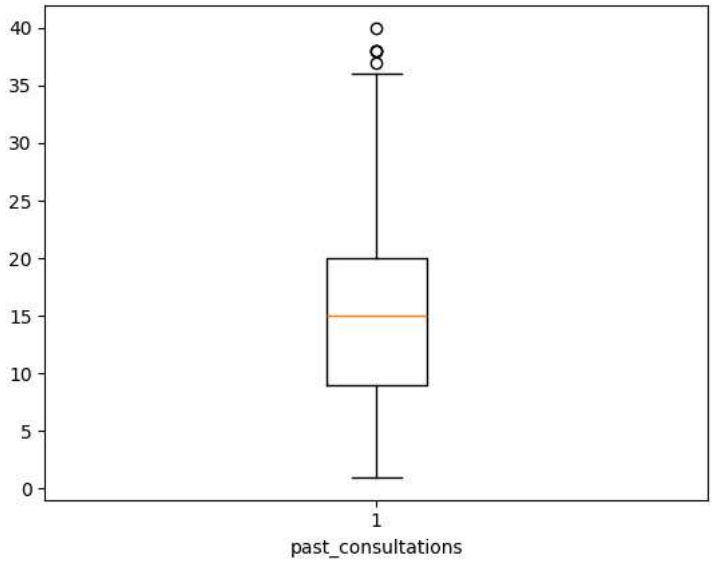

60
50
40
30
20
1
age

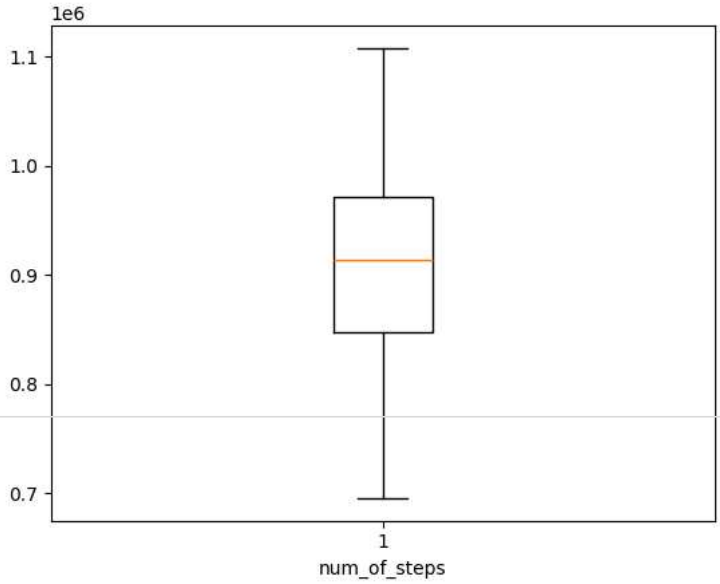

50
45
40
35
30
25
20
15
1
bmi

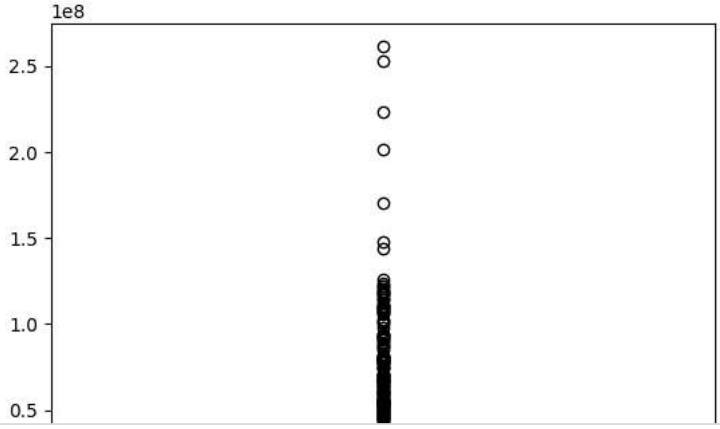

5
4
3
2
1
0
1
children


80000
70000
60000
50000


40000
30000
20000
10000
0
1
Claim_Amount


40
35
30
25
20
15
10
5
0
1
past_consultations


1e6
1.1
1.0
0.9
0.8
0.7
1
num_of_steps


1e8
2.5
2.0
1.5
1.0
0.5

```
for i in col_list:
  Q1= rv[i].quantile(0.25)
  Q3=rv[i].quantile(0.75)
  IQR=Q3-Q1
  lower_limit=Q1-1.5*IQR
  upper_limit=Q3+1.5*IQR
  rv1= rv[(rv[i]> lower_limit) & (rv[i]< upper_limit)]
```

```
rv1.head()
```

	age	sex	bmi	children	smoker	Claim_Amount	past_consultations	num_of_steps	Hospital_expenditure	NUmber_of_past_hospitaliz
0	18.0	male	23.21	0.0	no	29087.54313	17.0	715428.0	4720920.992	
1	18.0	male	30.14	0.0	no	39053.67437	7.0	699157.0	4329831.676	
2	18.0	male	33.33	0.0	no	39023.62759	19.0	702341.0	6884860.774	
3	18.0	male	33.66	0.0	no	28185.39332	11.0	700250.0	4274773.550	
4	18.0	male	34.10	0.0	no	14697.85941	16.0	711584.0	3787293.921	

```
rv1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1146 entries, 0 to 1312
Data columns (total 13 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   age                             1146 non-null   float64
 1   sex                             1146 non-null   object
 2   bmi                             1146 non-null   float64
 3   children                        1146 non-null   float64
 4   smoker                          1146 non-null   object
 5   Claim_Amount                    1146 non-null   float64
 6   past_consultations              1146 non-null   float64
 7   num_of_steps                    1146 non-null   float64
 8   Hospital_expenditure            1146 non-null   float64
 9   NUmber_of_past_hospitalizations 1146 non-null   float64
 10  Anual_Salary                    1146 non-null   float64
 11  region                          1146 non-null   object
 12  charges                         1146 non-null   float64
dtypes: float64(10), object(3)
memory usage: 125.3+ KB
```

```
rv1.isnull()
```

	age	sex	bmi	children	smoker	Claim_Amount	past_consultations	num_of_steps	Hospital_expenditure	NUmber_of_past_hos
0	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	
1147	False	False	False	False	False	False	False	False	False	
1148	False	False	False	False	False	False	False	False	False	
1150	False	False	False	False	False	False	False	False	False	
1300	False	False	False	False	False	False	False	False	False	
1312	False	False	False	False	False	False	False	False	False	

1146 rows × 13 columns

```
col_list=[]
for i in rv1.columns:
  if ((rv1[i].dtype !='object')& (i !='charges')):
          col_list.append(i)
```

```
col_list
```

```
['age',
 'bmi',
 'children',
 'Claim_Amount',
 'past_consultations',
 'num_of_steps',
 'Hospital_expenditure',
 'NUmber_of_past_hospitalizations',
 'Anual_Salary']
```

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

x = rv1[col_list]
vif_data = pd.DataFrame()
vif_data['feature'] = x.columns
vif_data['VIF'] = [variance_inflation_factor(x.values, i) for i in range(len(x.columns))]

display(vif_data)
```

	feature	VIF
0	age	13.484870
1	bmi	25.344800
2	children	2.027611
3	Claim_Amount	5.986548
4	past_consultations	6.769301
5	num_of_steps	56.669854
6	Hospital_expenditure	3.597009
7	NUmber_of_past_hospitalizations	12.055058
8	Anual_Salary	4.504657

```
vif_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9 entries, 0 to 8
Data columns (total 2 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   feature  9 non-null      object 
 1   VIF      9 non-null      float64
dtypes: float64(1), object(1)
memory usage: 276.0+ bytes
```

```
x = rv1[col_list]
vif_data = pd.DataFrame()
vif_data['feature'] = x.columns
vif_data['VIF'] = [variance_inflation_factor(x.values, i) for i in range(len(x.columns))]

display(vif_data)
```

	feature	VIF
0	age	13.484870
1	bmi	25.344800
2	children	2.027611
3	Claim_Amount	5.986548
4	past_consultations	6.769301
5	num_of_steps	56.669854
6	Hospital_expenditure	3.597009
7	NUmber_of_past_hospitalizations	12.055058
8	Anual_Salary	4.504657

```
rv1=rv1.drop('bmi',axis=1)
```

```
rv1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1146 entries, 0 to 1312
Data columns (total 12 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   age                              1146 non-null   float64
 1   sex                              1146 non-null   object 
 2   children                         1146 non-null   float64
 3   smoker                           1146 non-null   object 
 4   Claim_Amount                     1146 non-null   float64
 5   past_consultations               1146 non-null   float64
 6   num_of_steps                     1146 non-null   float64
 7   Hospital_expenditure             1146 non-null   float64
 8   NUmber_of_past_hospitalizations  1146 non-null   float64
 9   Anual_Salary                     1146 non-null   float64
 10  region                           1146 non-null   object 
 11  charges                          1146 non-null   float64
dtypes: float64(9), object(3)
memory usage: 116.4+ KB
```

```
col_list=[]
for i in rv1.columns:
  if ((rv1[i].dtype !='object')& (i !='charges')):
          col_list.append(i)
```

```
col_list
```

```
['age',
 'children',
 'Claim_Amount',
 'past_consultations',
 'num_of_steps',
 'Hospital_expenditure',
 'NUmber_of_past_hospitalizations',
 'Anual_Salary']
```

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

x = rv1[col_list]
vif_data = pd.DataFrame()
vif_data['feature'] = x.columns
vif_data['VIF'] = [variance_inflation_factor(x.values, i) for i in range(len(x.columns))]

display(vif_data)
```

	feature	VIF
0	age	13.455041
1	children	2.026930
2	Claim_Amount	5.986349
3	past_consultations	6.767323
4	num_of_steps	25.382004
5	Hospital_expenditure	3.592189
6	NUmber_of_past_hospitalizations	11.814499
7	Anual_Salary	4.146998

```
rv1=rv1.drop('num_of_steps',axis=1)
```

```
rv1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1146 entries, 0 to 1312
Data columns (total 11 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   age                              1146 non-null   float64
 1   sex                              1146 non-null   object 
```

```
 2   children                         1146 non-null   float64
 3   smoker                           1146 non-null   object
 4   Claim_Amount                     1146 non-null   float64
 5   past_consultations               1146 non-null   float64
 6   Hospital_expenditure             1146 non-null   float64
 7   NUmber_of_past_hospitalizations  1146 non-null   float64
 8   Anual_Salary                     1146 non-null   float64
 9   region                           1146 non-null   object
 10  charges                          1146 non-null   float64
dtypes: float64(8), object(3)
memory usage: 107.4+ KB
```

```
col_list=[]
for i in rv1.columns:
  if ((rv1[i].dtype !='object')& (i !='charges')):
          col_list.append(i)
```

```
col_list
```

```
['age',
 'children',
 'Claim_Amount',
 'past_consultations',
 'Hospital_expenditure',
 'NUmber_of_past_hospitalizations',
 'Anual_Salary']
```

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

x = rv1[col_list]
vif_data = pd.DataFrame()
vif_data['feature'] = x.columns
vif_data['VIF'] = [variance_inflation_factor(x.values, i) for i in range(len(x.columns))]

display(vif_data)
```

	feature	VIF
0	age	9.229855
1	children	1.978922
2	Claim_Amount	4.928278
3	past_consultations	5.646516
4	Hospital_expenditure	3.548328
5	NUmber_of_past_hospitalizations	11.088112
6	Anual_Salary	4.110859

```
rv1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1146 entries, 0 to 1312
Data columns (total 11 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   age                              1146 non-null   float64
 1   sex                              1146 non-null   object
 2   children                         1146 non-null   float64
 3   smoker                           1146 non-null   object
 4   Claim_Amount                     1146 non-null   float64
 5   past_consultations               1146 non-null   float64
 6   Hospital_expenditure             1146 non-null   float64
 7   NUmber_of_past_hospitalizations  1146 non-null   float64
 8   Anual_Salary                     1146 non-null   float64
 9   region                           1146 non-null   object
 10  charges                          1146 non-null   float64
dtypes: float64(8), object(3)
memory usage: 107.4+ KB
```

```
rv1=rv1.drop('NUmber_of_past_hospitalizations',axis=1)
```

```
rv1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1146 entries, 0 to 1312
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   age                   1146 non-null   float64
 1   sex                   1146 non-null   object 
 2   children              1146 non-null   float64
 3   smoker                1146 non-null   object 
 4   Claim_Amount          1146 non-null   float64
 5   past_consultations    1146 non-null   float64
 6   Hospital_expenditure  1146 non-null   float64
 7   Anual_Salary          1146 non-null   float64
 8   region                1146 non-null   object 
 9   charges               1146 non-null   float64
dtypes: float64(7), object(3)
memory usage: 98.5+ KB
```

```
col_list=[]
for i in rv1.columns:
  if ((rv1[i].dtype !='object')& (i !='charges')):
          col_list.append(i)
```

```
from statsmodels.stats.outliers_influence import variance_inflation_factor

x = rv1[col_list]
vif_data = pd.DataFrame()
vif_data['feature'] = x.columns
vif_data['VIF'] = [variance_inflation_factor(x.values, i) for i in range(len(x.columns))]

display(vif_data)
```

	feature	VIF
0	age	6.367623
1	children	1.754581
2	Claim_Amount	4.824929
3	past_consultations	5.623349
4	Hospital_expenditure	3.472507
5	Anual_Salary	3.740875

```
x=rv1[['children','Claim_Amount','past_consultations','Hospital_expenditure','Anual_Salary']]
y=rv1['charges']
```

```
from sklearn.model_selection import train_test_split
```

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.7,shuffle=True)
```

```
from sklearn.linear_model import LinearRegression
```

```
clf=LinearRegression()
clf.fit(x_train,y_train)
y_pred=clf.predict(x_test)
```

```
(x,y)
```

```
(      children  Claim_Amount  past_consultations  Hospital_expenditure  \
 0          0.0   29087.54313                17.0          4.720921e+06   
 1          0.0   39053.67437                 7.0          4.329832e+06   
 2          0.0   39023.62759                19.0          6.884861e+06   
 3          0.0   28185.39332                11.0          4.274774e+06   
 4          0.0   14697.85941                16.0          3.787294e+06   
 ...        ...           ...                 ...                   ...   
 1147       0.0   49372.24572                21.0          2.180519e+07   
 1148       1.0   63328.19543                14.0          2.313545e+07   
 1150       1.0   54149.85460                14.0          2.180737e+07   
 1300       0.0   57588.33715                29.0          9.365456e+07   
 1312       0.0   54661.70946                37.0          1.117967e+08   
```

```
      Anual_Salary
0     5.578497e+07
1     1.370089e+07
2     7.352311e+07
3     7.581968e+07
4     2.301232e+07
...            ...
1147  7.000519e+08
1148  6.953394e+08
1150  7.044854e+08
1300  3.696849e+08
1312  3.696849e+08
```