# TASK REPORT
# STATISTICS

**By:**
**Akbar Maulana**
**Data Science Student**

**Data Fellowship**
**2020**

# Chapter 1
# Introduction

Statistics is a term used to summarize a process that an analyst uses to characterize a data set. If the data set depends on a sample of a larger population, then the analyst can develop interpretations about the population primarily based on the statistical outcomes from the sample. Statistical analysis involves the process of gathering and evaluating data and then summarizing the data into a mathematical form. Statistics is used in various disciplines such as psychology, business, physical and social sciences, humanities, government, and manufacturing. Statistical data is gathered using a sample procedure or other method. Two types of statistical methods are used in analyzing data: descriptive statistics and inferential statistics. Descriptive statistics are used to synopsize data from a sample exercising the mean or standard deviation. Inferential statistics are used when data is viewed as a subclass of a specific population. In addition, statistics are the core of data science and therefore a data scientist must master statistics well.

A CEO of the Mallianzs insurance company, wants to know its customer profile in a detailed way. He has several questions about this. The questions are:

1. Perform basic exploratory data analysis which should include the following and print out your insights at every step:
   a. The shape of the data
   b. The data type of each attribute
   c. Checking the presence of missing values
   d. 5 points summary of numerical attributes
   e. Distribution of 'bmi', 'age' and 'charges' columns
   f. The measure of skewness of 'bmi', 'age', and 'charges' columns
   g. Checking the presence of outliers in 'bmi', 'age' and 'charges' columns
2. Do charges of people who smoke differ significantly from the people who don't? (HypothesisTesting)

# Chapter 2
# Progress Report

In this chapter you will have to fill in the table below according to the progress of the project that you have made along the way. We need to know how long it takes for you and how big the effort that you have done in order to complete this task. We appreciate detailed information.

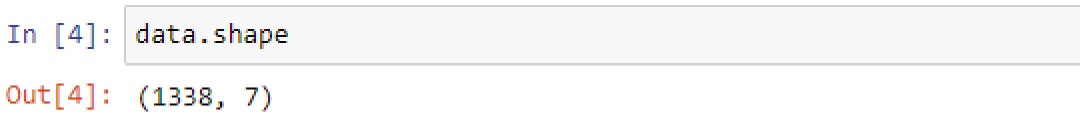
| **Day/Date** | **Task** | **Level**<br>(easy/medium/hard) | **Comments** |
|---|---|---|---|
| 24/04/2020 | Analyzing Data using Jupyter Notebook | Easy | |

# Chapter 3
# Task Report

## Customer Profile of Mallianzs Insurance Company

In this 4th practice case, I was asked to explore customer profile data from the insurance company Mallianz. The following are the results of exploration carried out:

### 1. Shape Of Data

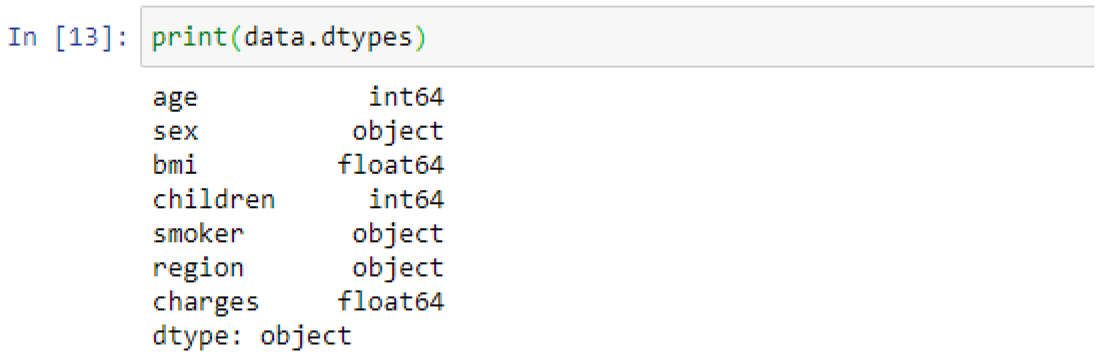


```
In [4]: data.shape

Out[4]: (1338, 7)
```

Customer profile data Mallianzs insurance company has dimensions (1338,7) which means that there are 1338 data rows and 7 data columns. 1338 rows represent the number of customers while 7 columns namely age, sex, bmi, children, smoker, region and charges represent the profile of each customer. Following is an explanation of each variable:

| Coloumn name | Definition |
|---|---|
| **Age** | The age of the insurance account holder |
| **Sex** | The gender of the insurance account holder |
| **Bmi** | The Body Mass Index of the insurance account holder |
| **Children** | The number of children of the insurance account holder |
| **Smoker** | The smoking status of the insurance account holder |
| **Region** | The region of the insurance account holder |
| **Charges** | The charges/insurance fee  paid by the insurance account holder |

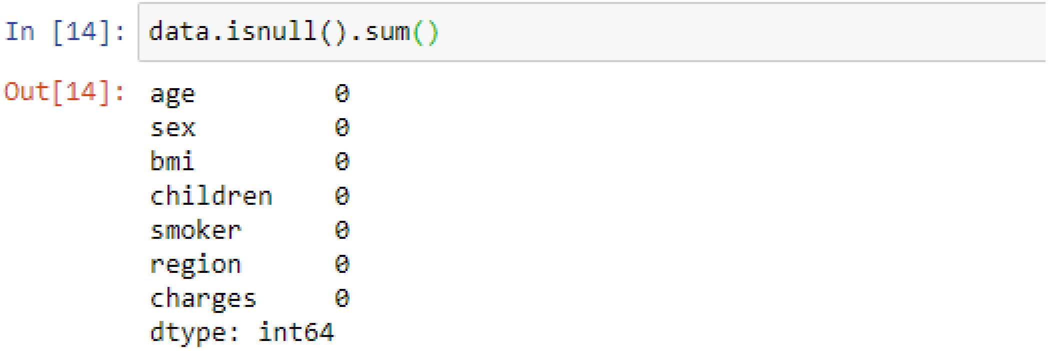
### 2. Type of Data



```
In [13]: print(data.dtypes)

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

There are seven variables in the customer profile data, namely age, sex, bmi, children, smoker, region and charges. The age and children variables are variables with integer data types. The bmi and charges variables are variables with the float data type. While sex, smoker and region variables are variables with object data types.

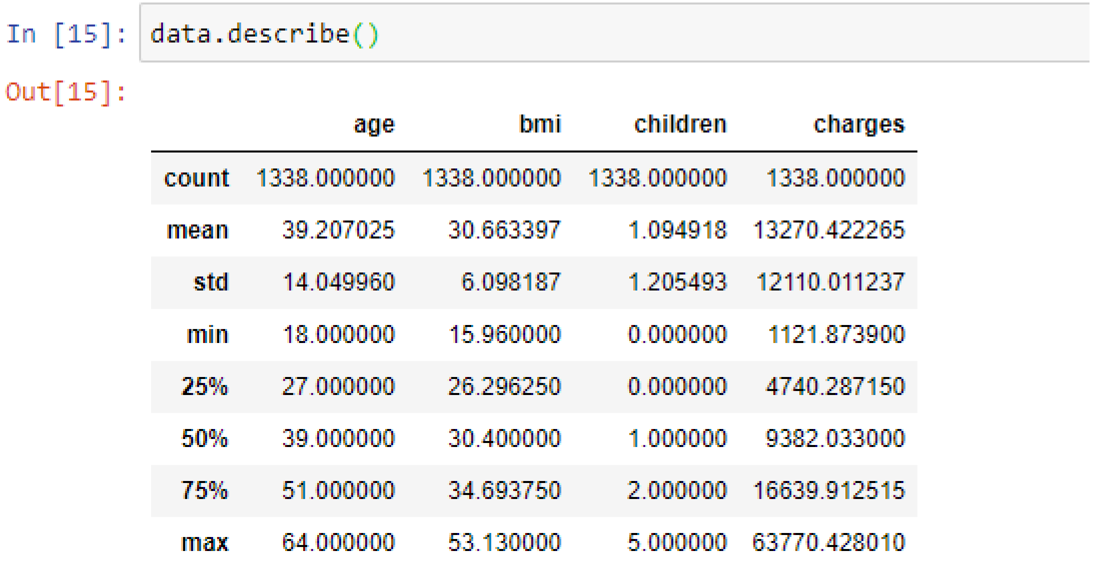
### 3. Check Missing Value



```
In [14]: data.isnull().sum()

Out[14]: age         0
         sex         0
         bmi         0
         children    0
         smoker      0
         region      0
         charges     0
         dtype: int64
```

Based on the results above, it can be concluded that there is no missing value in the data.

### 4. Summary of Numerical Variables



```
In [15]: data.describe()

Out[15]:
```

| | 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 |

Based on the results above, it can be seen that the number of insurance account holders at the Mallianz company is 1338 people. The following is a profile description of the insurance account holder:

(1) The mean and median age of insurance account holders is 39 years. The youngest insurance account holder is 18 years old, while the oldest insurance account holder is 64 years old. The age of insurance account holders varies considerably from those who are still teenagers to those who are elderly, it can be seen from the high standard deviation which is quite high at 14.05.

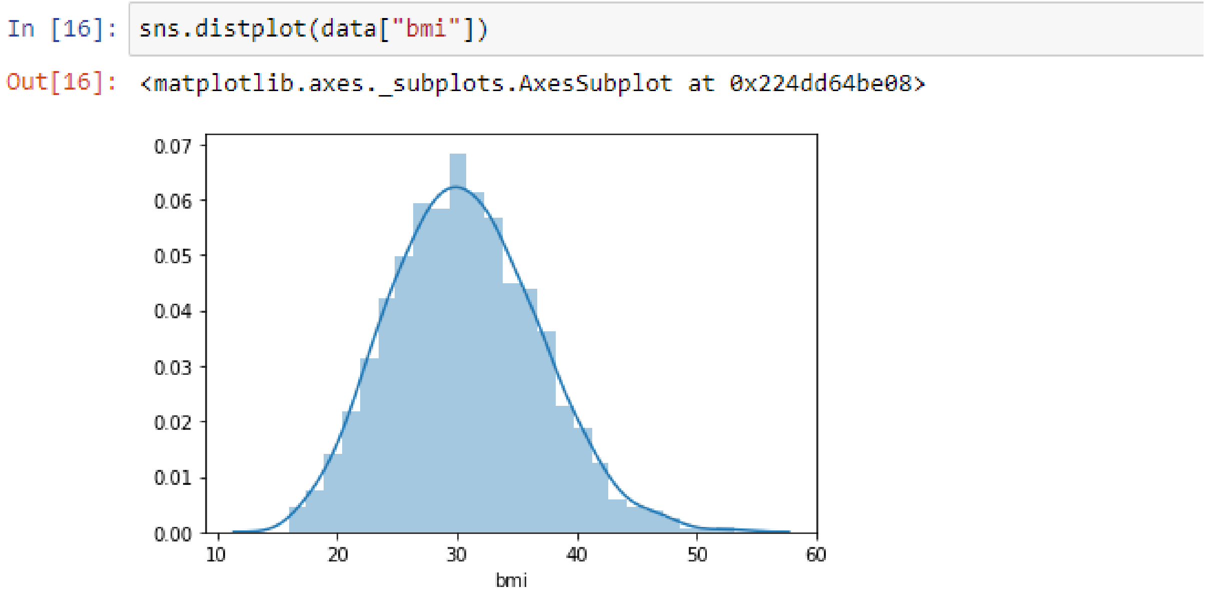
(2) The mean body mass index of the insurance account holder is 30.66 while the median value is 30.4. The lowest body mass index of insurance account holders is 15.96 while the highest is 53.13.

(3) Insurance account holders have an mean of 1 children. But there are also some who do not have children and some who have 5 children.

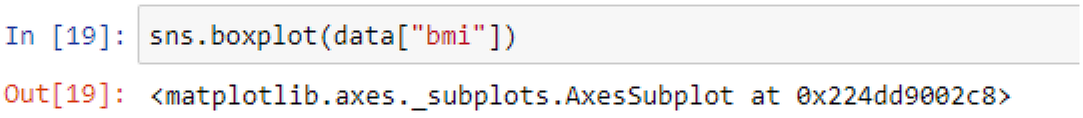
(4) The mean of insurance charges by insurance account holders is 13270.42 US $, while the median is 9382.03 US $. The median value is more appropriate than the mean value, it is because there are many insurance account holders whose insurance charges are much higher than the normal price. The lowest insurance charges is 1121.87 US $ while the highest insurance charges is 63770.43 US $. The insurance charges by insurance account holders vary greatly in value.

## 5. Distribution of Numerical Variables

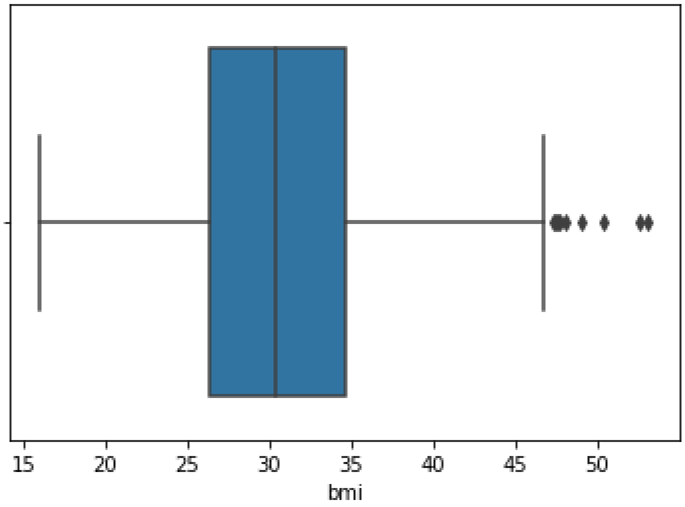
**BMI:**



```
In [16]: sns.distplot(data["bmi"])

Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x224dd64be08>
```

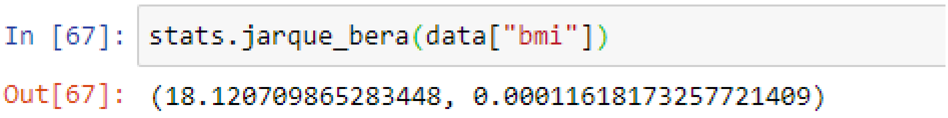


```
In [19]: sns.boxplot(data["bmi"])
Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x224dd9002c8>
```



When viewed based on plot density and histogram, the body mass index (bmi) variable is a variable that approaches the normal distribution. However, if checked based on the box plot, the variable value is not symmetrical and protrudes upward because there are several observations of the outlier above. But do these outlier observations affect the data distribution? When we check using visualization, sometimes it can cause differences in perceptions between people. Therefore, in this case we use the fallow jarque test to prove this, here are the results of the test:

a. Hypothesis
   $H_0$: The bmi variable is normally distributed
   $H_1$: The bmi variable is not normally distributed
b. Level of significance
   $\alpha = 0.05$
c. Critical area:
   $H_0$ is rejected if test statistics > 5.991 or p-value<0.05
d. Test Statistics

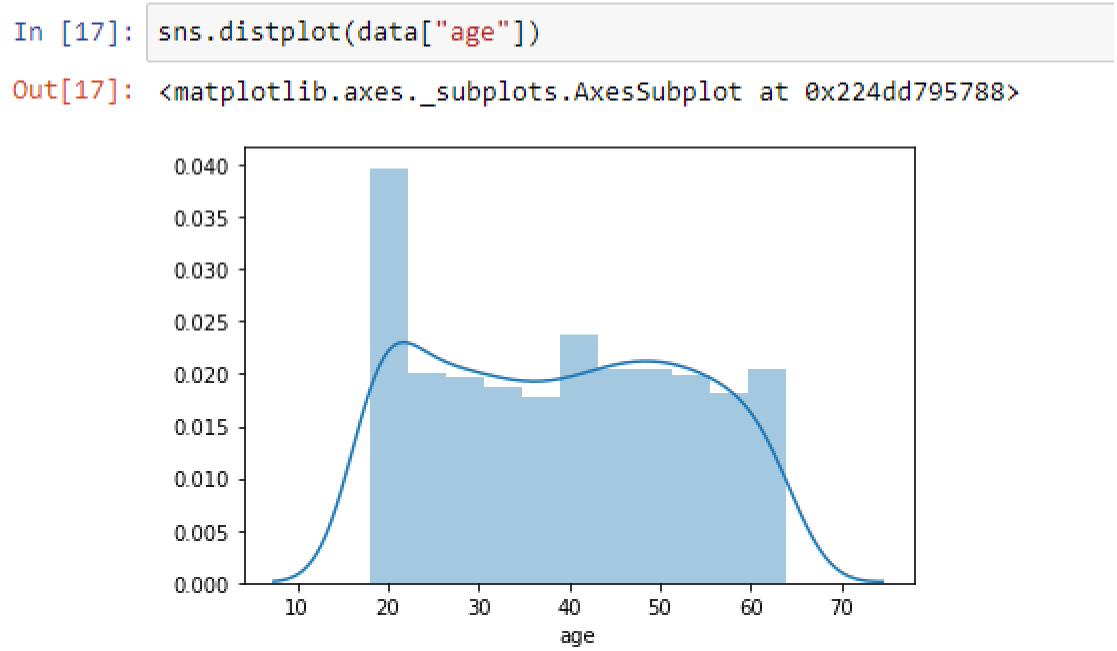


```
In [67]: stats.jarque_bera(data["bmi"])
Out[67]: (18.120709865283448, 0.00011618173257721409)
```

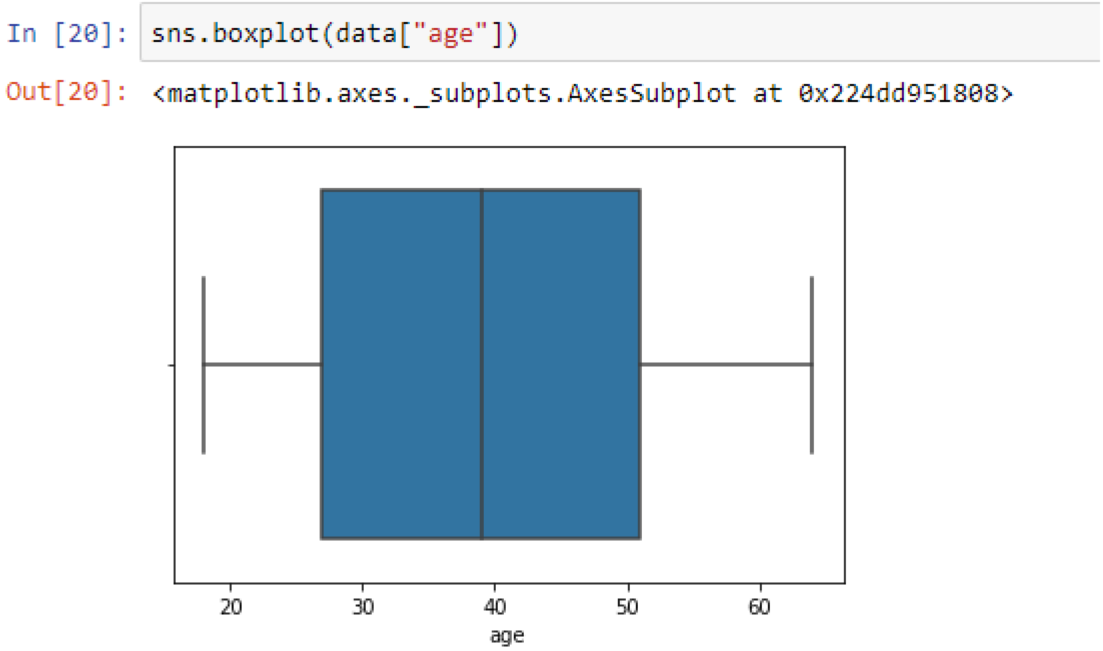
e. Conclusion
   $H_0$ is rejected, which means that the bmi variable is not normally distributed.

Even though based on the density plot and histogram the bmi variable looks close to the normal distribution, the results are quite deceptive because there are outlier observations that greatly affect the data distribution. Normal distribution is one distribution that is sensitive to the presence of outliers.

**AGE:**



```
In [17]: sns.distplot(data["age"])

Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x224dd795788>
```



```
In [20]: sns.boxplot(data["age"])

Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x224dd951808>
```

Based on plot density and histogram, age variable is data with bimodal distribution. This can be seen from the presence of two maximum peak or local points in the data. Furthermore, when viewed from the boxplot, the boxplot of the age variable is asymmetrical and protrudes upward even though there are no outliers. Bimodal distribution shows that the data contained a mixture of two distributions or commonly called a mixture distribution. Furthermore, to ensure that the age variable is not normally distributed, we can use the following fallow test:

a. Hypothesis

   $H_0$: The age variable is normally distributed

   $H_1$: The age variable is not normally distributed

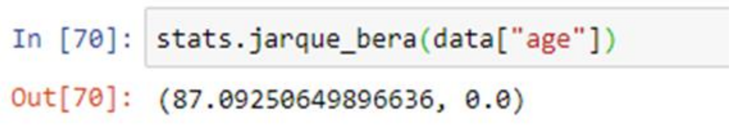
b. Level of significance

   $\alpha = 0.05$

c. Critical area:

   $H_0$ is rejected if test statistics > 5.991 or p-value<0.05

d. Test Statistics

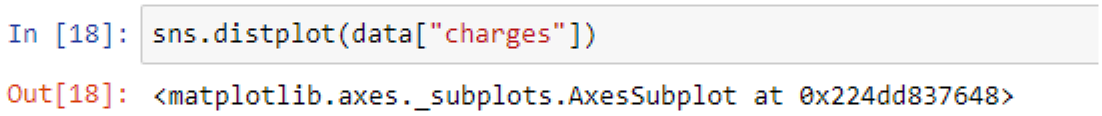


```
In [70]: stats.jarque_bera(data["age"])

Out[70]: (87.09250649896636, 0.0)
```

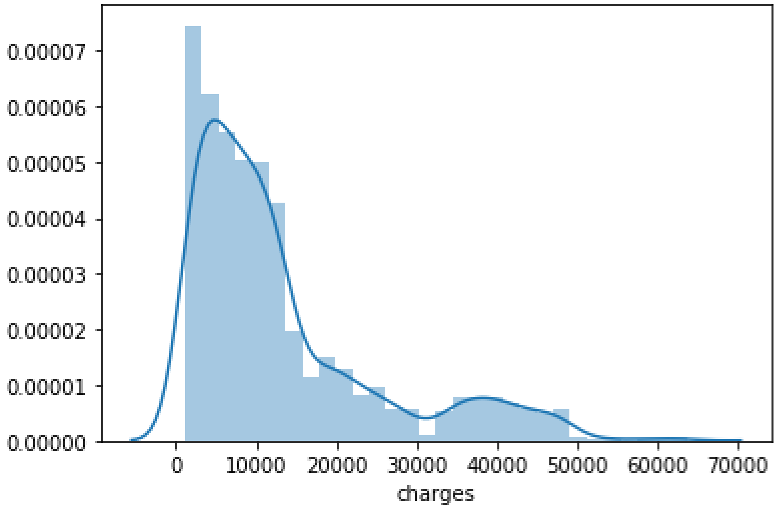
e. Conclusion

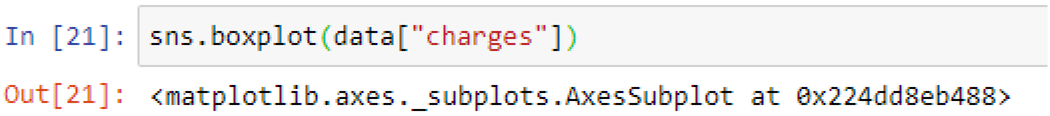
   $H_0$ is rejected, which means that the age variable is not normally distributed.

**CHARGES:**

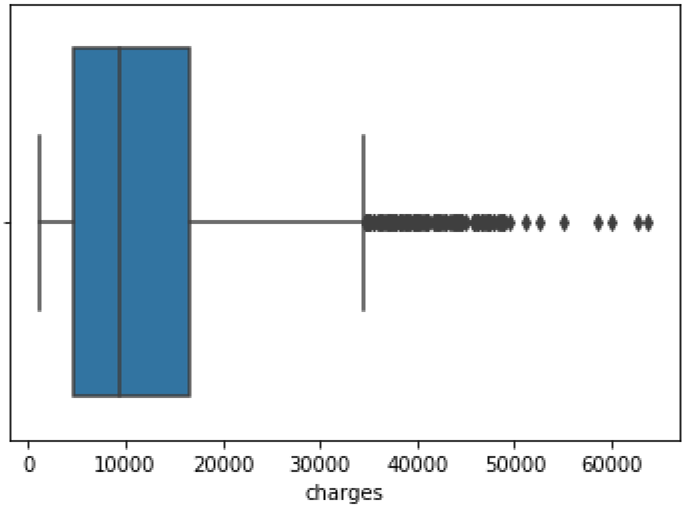


```
In [18]: sns.distplot(data["charges"])

Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x224dd837648>
```



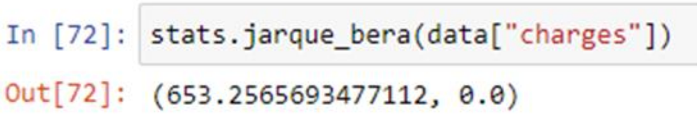


```
In [21]: sns.boxplot(data["charges"])

Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x224dd8eb488>
```



Based on plot density and histogram, variable charges are data with positive skewness. This is caused because there are many top outliers in the data. So the mode value <median <mean. Furthermore, when viewed from the boxplot, the boxplot of the variable charges is not symmetrical and very protruding upwards. Positive skewness shows that the data is not normally distributed. Furthermore, to ensure that the variable charges are not normally distributed, we can use the following fallow test:

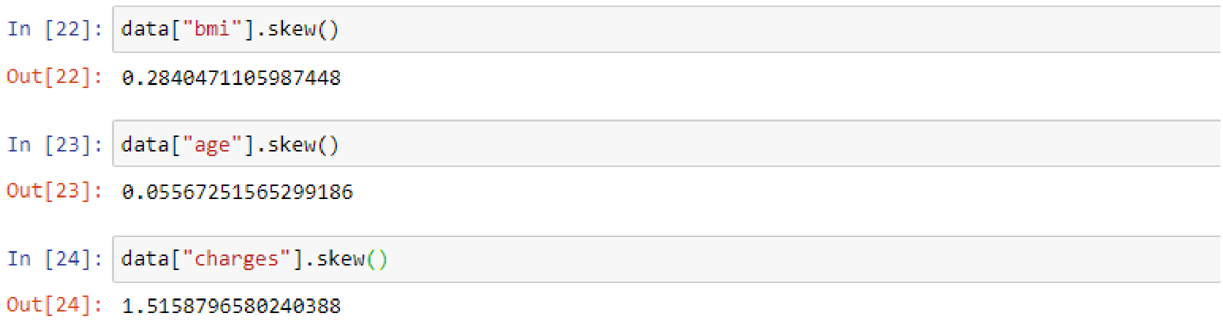
a. Hypothesis
   $H_0$: The charges variable is normally distributed
   $H_1$: The charges variable is not normally distributed
b. Level of significance
   $\alpha = 0.05$
c. Critical area:
   $H_0$ is rejected if test statistics > 5.991 or p-value<0.05
d. Test Statistics



```
In [72]: stats.jarque_bera(data["charges"])

Out[72]: (653.2565693477112, 0.0)
```

e. Conclusion
   $H_0$ is rejected, which means that the charges variable is not normally distributed.

## 6. The measure of skewness of numerical variables



```
In [22]: data["bmi"].skew()

Out[22]: 0.2840471105987448

In [23]: data["age"].skew()

Out[23]: 0.05567251565299186

In [24]: data["charges"].skew()

Out[24]: 1.5158796580240388
```

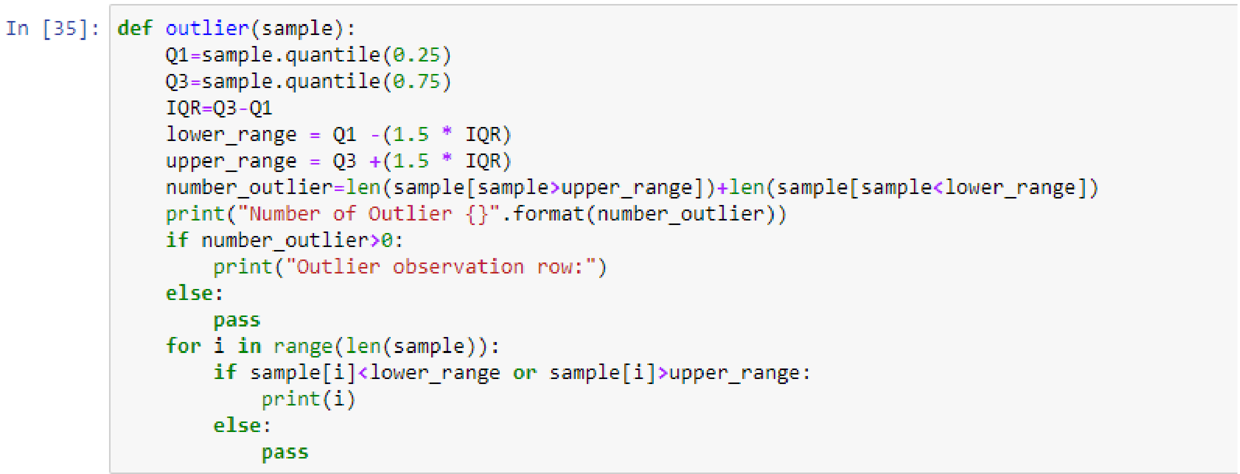
Skewness can be used to see the distribution of data, but it is not enough to just use skewness. Skewness only measures the degree of slope of the distribution. The following is an explanation of the skewness value of each numeric variable:

(1) The bmi variable has a skewness value of 0.28, which is positive skewness, which means the mode value <median <mean. Although the skewness value is quite small, however, outlier observations on the bmi variable greatly affect the data distribution so that the bmi variable is not normally distributed.

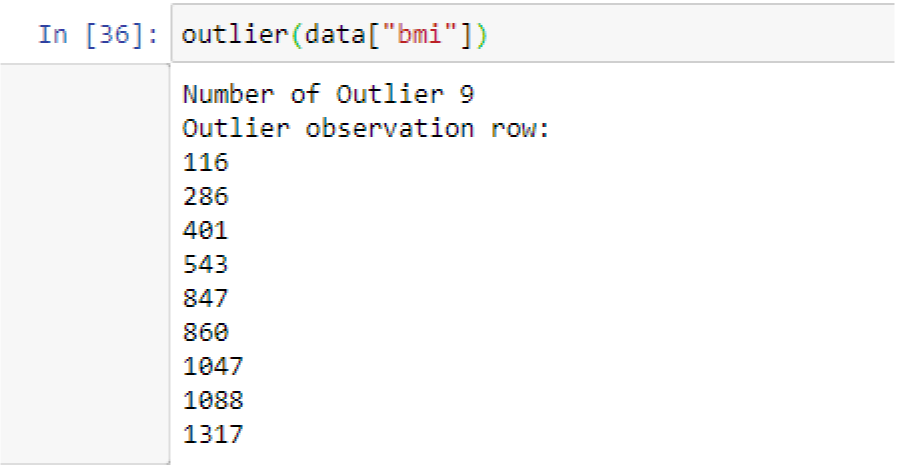
(2) The age variable has a skewness value of 0.055. This value is close to zero, but skewness alone is not enough to conclude that the data is normally distributed and it needs to be checked as well as kurtois values. When checking the value of the kurtois, there are two peaks in the density plot or commonly called a bimodal distribution. This shows that the age variable is not normally distributed.

(3) Variable charges are variables that have a large enough skewness value of 1.515. This value shows that the charges variable has positive skewness with the mode value <median <mean.

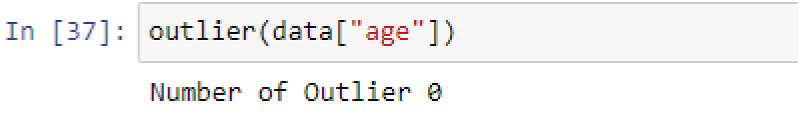
## 7. Checking the presence of outliers



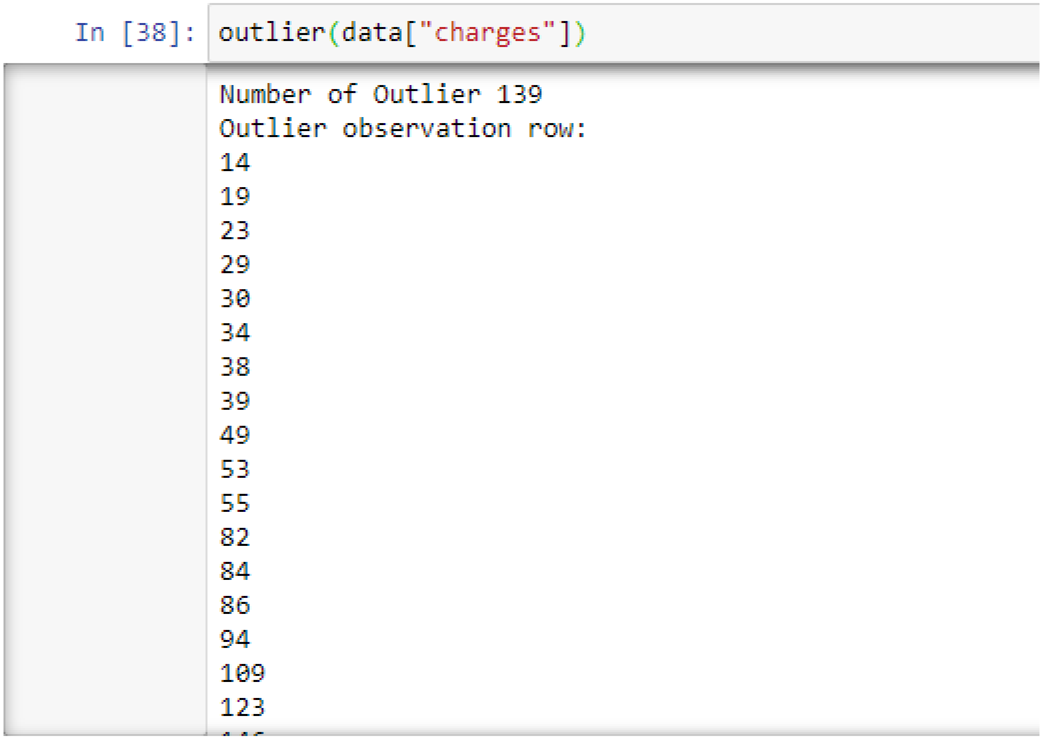
```
In [35]: def outlier(sample):
             Q1=sample.quantile(0.25)
             Q3=sample.quantile(0.75)
             IQR=Q3-Q1
             lower_range = Q1 -(1.5 * IQR)
             upper_range = Q3 +(1.5 * IQR)
             number_outlier=len(sample[sample>upper_range])+len(sample[sample<lower_range])
             print("Number of Outlier {}".format(number_outlier))
             if number_outlier>0:
                 print("Outlier observation row:")
             else:
                 pass
             for i in range(len(sample)):
                 if sample[i]<lower_range or sample[i]>upper_range:
                     print(i)
                 else:
                     pass
```



```
In [36]: outlier(data["bmi"])

Number of Outlier 9
Outlier observation row:
116
286
401
543
847
860
1047
1088
1317
```



```
In [37]: outlier(data["age"])

Number of Outlier 0
```

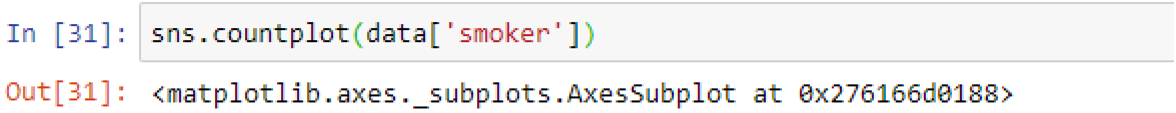


```
In [38]: outlier(data["charges"])

Number of Outlier 139
Outlier observation row:
14
19
23
29
30
34
38
39
49
53
55
82
84
86
94
109
123
```

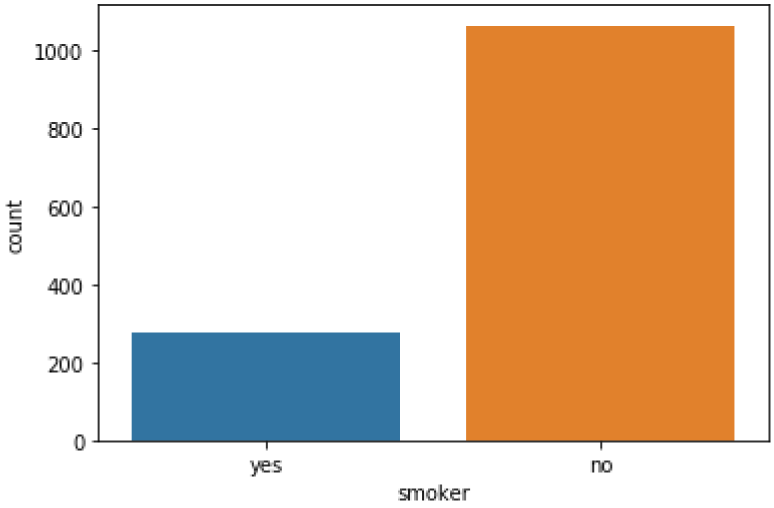
The bmi variable has 9 outlier observations, the age variable has no outlier observations while the charges variable has 139 outlier observations.

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

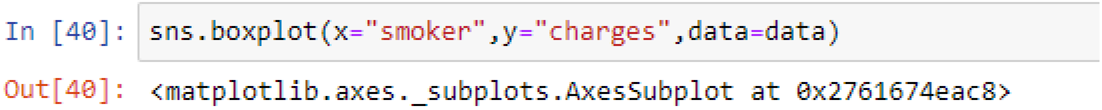
The first step taken is to look at the comparison between customers who smoke and not smoke. Based on the picture below, it can be seen that there are 275 customers who smoke and 1064 customers who don't smoke.



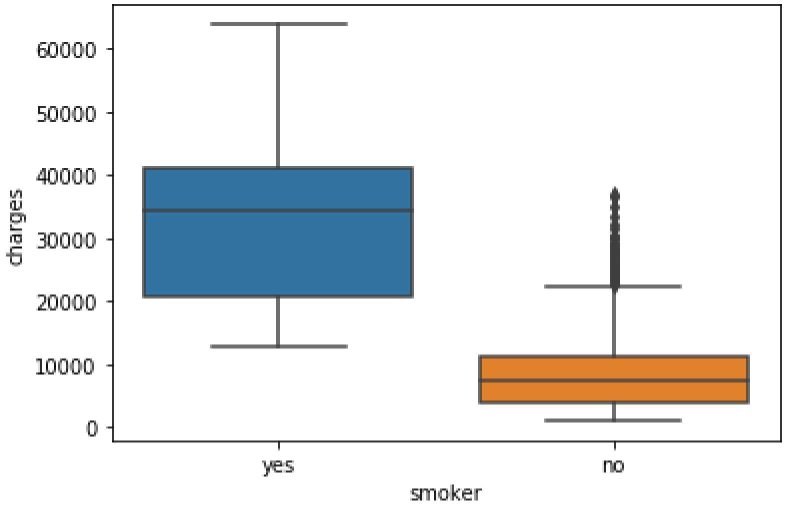
```
In [31]: sns.countplot(data['smoker'])

Out[31]: <matplotlib.axes._subplots.AxesSubplot at 0x276166d0188>
```



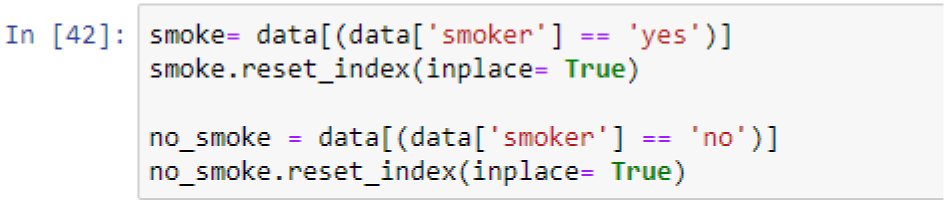
The next step is to look at the distribution of each class using the boxplot. If seen from the boxplots below, between customers who smoke and don't smoke, there really is a difference in charges. Customers who smoke on average have greater charges than customers who don't smoke.



```
In [40]: sns.boxplot(x="smoker",y="charges",data=data)

Out[40]: <matplotlib.axes._subplots.AxesSubplot at 0x2761674eac8>
```



Furthermore, the data is partitioned into two parts namely smokers and nonsmokers customer data.

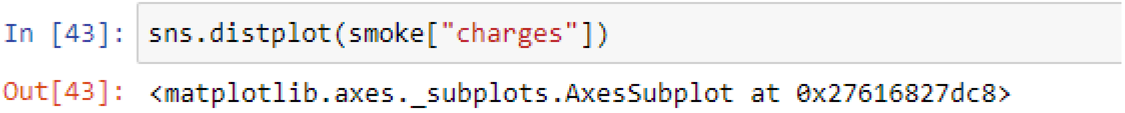


```
In [42]: smoke= data[(data['smoker'] == 'yes')]
         smoke.reset_index(inplace= True)

         no_smoke = data[(data['smoker'] == 'no')]
         no_smoke.reset_index(inplace= True)
```

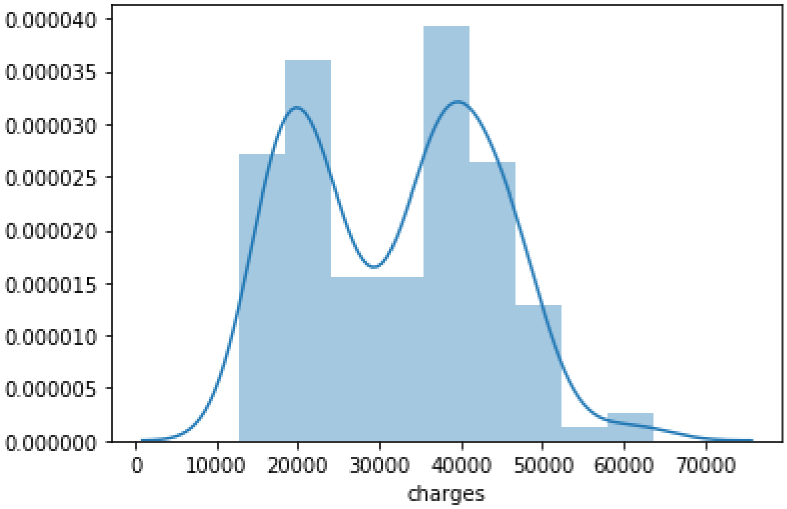
In testing the mean difference in two populations, it can be used parametric and non-parametric tests. Parametric test is two independent sample t tests while non-parametric test is Mann Whitney. Parametric tests must meet the assumptions of normality and homogeneity of variance, if these two assumptions are not met then a non-parametric test can be used.

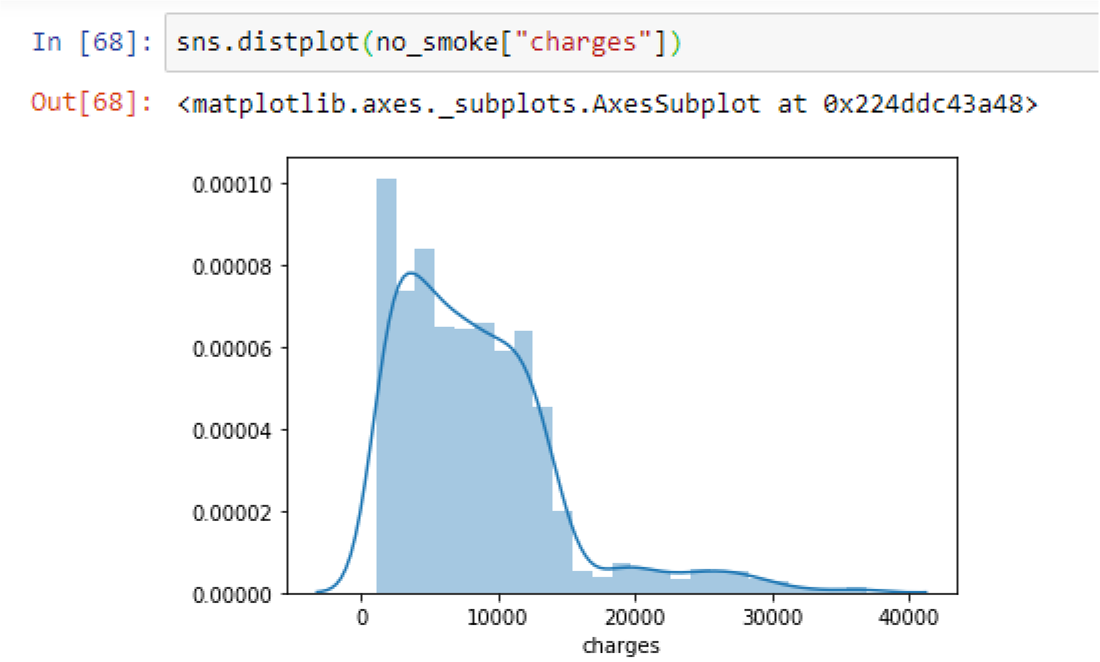
**Normality Test:**

When viewed from the density plot and histogram, the two data are not normally distributed. Data on insurance charges for smoker customers is a bimodal distribution, while insurance charges for nonsmokers have positive skewness.



```
In [43]: sns.distplot(smoke["charges"])

Out[43]: <matplotlib.axes._subplots.AxesSubplot at 0x27616827dc8>
```

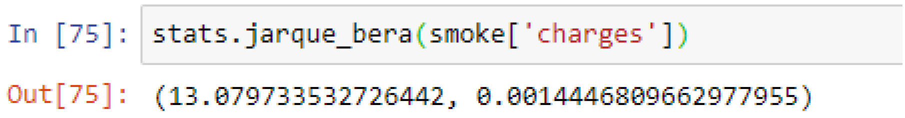




```
In [68]: sns.distplot(no_smoke["charges"])

Out[68]: <matplotlib.axes._subplots.AxesSubplot at 0x224ddc43a48>
```

Next, to emphasize the results, Jarque's normality test was carried out on both data:
Smoker Customers:

a. Hipotesis
   $H_0$: Insurance charges for smoker customers data is normally distributed
   $H_1$: Insurance charges for smoker customers data is not normally distributed
b. Level of significance
   $\alpha = 0.05$
c. Critical area:
   $H_0$ is rejected if test statistics > 5.728 or p-value<0.05
d. Test statistics



```
In [75]: stats.jarque_bera(smoke['charges'])

Out[75]: (13.079733532726442, 0.0014446809662977955)
```

e. Conclusion
   $H_0$ is rejected, which means insurance cost data for smoker customers are not normally distributed.

Non Smoker Customer:

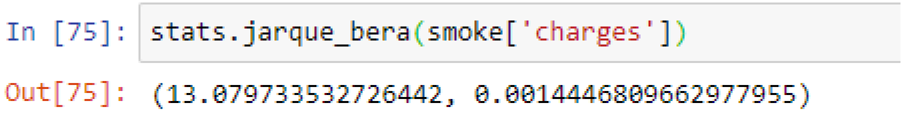
a. Hypothesis
   $H_0$: Insurance charges for non-smoker customers data is normally distributed
   $H_1$: insurance charges for non-smoker customers data is not normally distributed
b. Level of significance

$\alpha = 0.05$

c. Critical region:

$H_0$ is rejected if test statistics> 5.991 atau p-value<0.05

d. Test statistics



```
In [75]: stats.jarque_bera(smoke['charges'])

Out[75]: (13.079733532726442, 0.0014446809662977955)
```

e. Conclusion

$H_0$ is rejected, which means that insurance cost data for non-smoker customers is not normally distributed.

**Homogeneity Variance Test:**

To test whether the variance between the two homogeneous data can be used Leven's test. Following are the results of Levene's test:

a. Hypothesis

$H_0$: Both data have homogeneous variances

$H_1$: Both data have heterogeneous variance

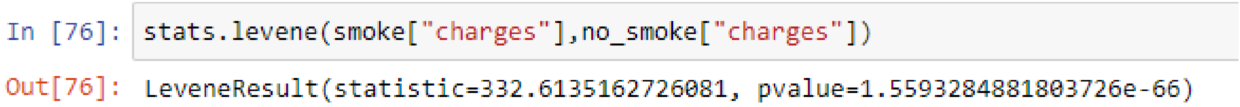
b. Level of significance

$\alpha = 0.05$

c. Critical region:

$H_0$ is rejected if p-value<0.05

d. Test statistics:



```
In [76]: stats.levene(smoke["charges"],no_smoke["charges"])

Out[76]: LeveneResult(statistic=332.6135162726081, pvalue=1.5593284881803726e-66)
```

e. Conclusion

$H_0$ is rejected, which means that both data have heterogeneous variance.

Because the two assumptions, namely the normality and homogeneity of variance are not fulfilled, we cannot use the two independent sample t test. We can use the non-parametric test, the Mann Whitney test. Here are the results of the Mann Whitney test:

a. Hypothesis

$H_0: \mu_1 = \mu_2$ (Smokers and nonsmokers pay the same insurance charges)

$H_1: \mu_1 \neq \mu_2$ (Smokers and nonsmokers pay different insurance charges)

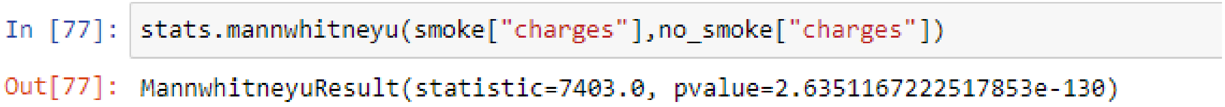
b. Level of significance

$\alpha = 0.05$

c. Critical region:

$H_0$ is rejected if p-value<0.05

d. Test statistics:



```
In [77]: stats.mannwhitneyu(smoke["charges"],no_smoke["charges"])

Out[77]: MannwhitneyuResult(statistic=7403.0, pvalue=2.6351167222517853e-130)
```

e. Conclusion:
$H_0$ is rejected, which means smokers and nonsmokers pay different insurance charges.