Hypothesis Testing

Prepared for

Shanti Sekhar

Prepared by

Tan Dai Jun

Table of Contents

1.0 Introduction

Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called null hypothesis and is denoted by $H_{0.}$ An alternative hypothesis, denoted by $H_{a,}$ which is the opposite of what is stated in the null hypothesis is then defined. The hypothesis testing procedure involves using sample data to determine whether or not H_0 can be rejected. If H_0 is rejected, the statistical conclusion is that the alternative hypothesis H_a is true.

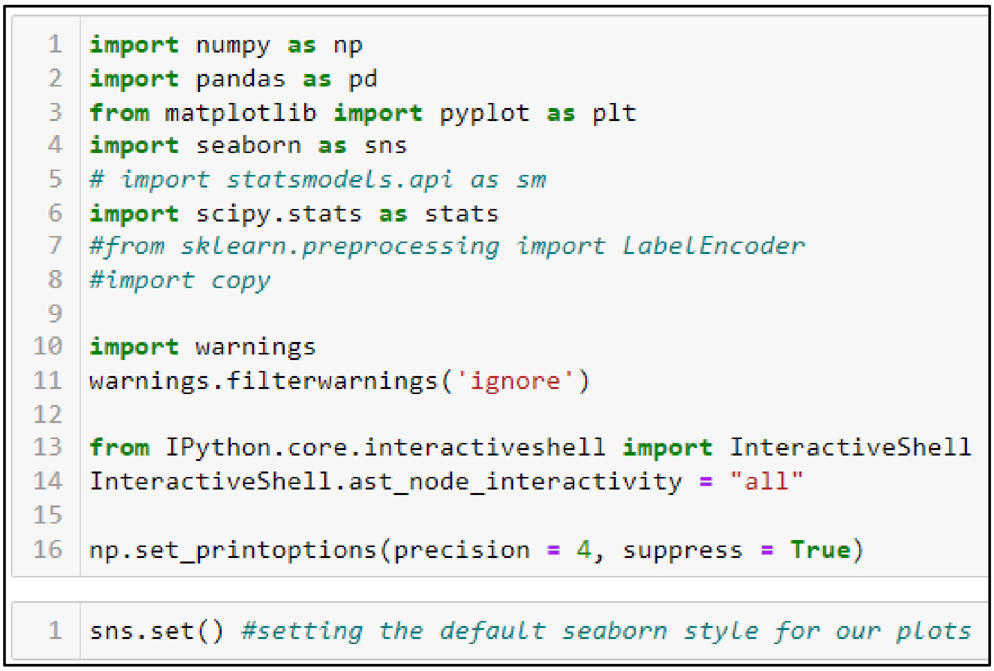
In this project, we will try to understand an insurance dataset that describes different characteristics of insurers and their respective medical claim amount. After that, a hypothesis test will be conducted to study if smoking habit will affect the medical claim amount of the insurer.

2.0 Methodology

In this project, we will use Exploratory Data Analysis to explore the dataset. We will first make sure the data is pre-processed before understand all the features in the dataset as well as the relationship between each feature. After the dataset exploration, a hypothesis test will be conducted to prove that if smoking will increase the amount of medical claim of the insurer.

The entire project will be executed using python in Jupyter Notebook. Several necessary libraries to support the execution of this project are listed in table below.

Library	Descriptions
numpy	House with large collection of mathematical functions to operate arrays
pandas	Built on top on numpy, offers data structure and operations for manipulating numerical tables and time series
matplotlib	A cross-platform, data visualization and graphical plotting library
seaborn	Built on top of matplotlib, provides high level interface for drawing attractive and informative statistical graphics
scipy	House with large collection of scientific and technical computing functions
warnings	Enable user to hide unnecessary warnings



```python
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

import warnings
warnings.filterwarnings('ignore')

from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

np.set_printoptions(precision = 4, suppress = True)
```

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

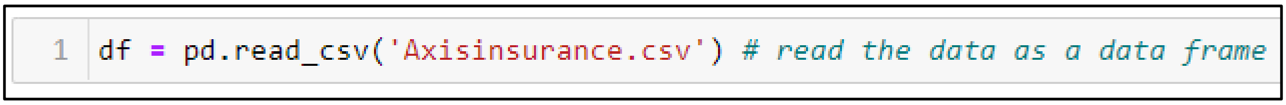
Figure 2.0.0: Import necessary libraries

3.0 Exploratory Data Analysis

The objective of exploratory data analysis is to first ensure the dataset is cleaned and transformed, then use dedicated plots to visualize the dataset. This will enable us to have an overview about the dataset and gain some useful insights by understanding the relationship between each feature in our dataset.

3.1 Data pre-processing

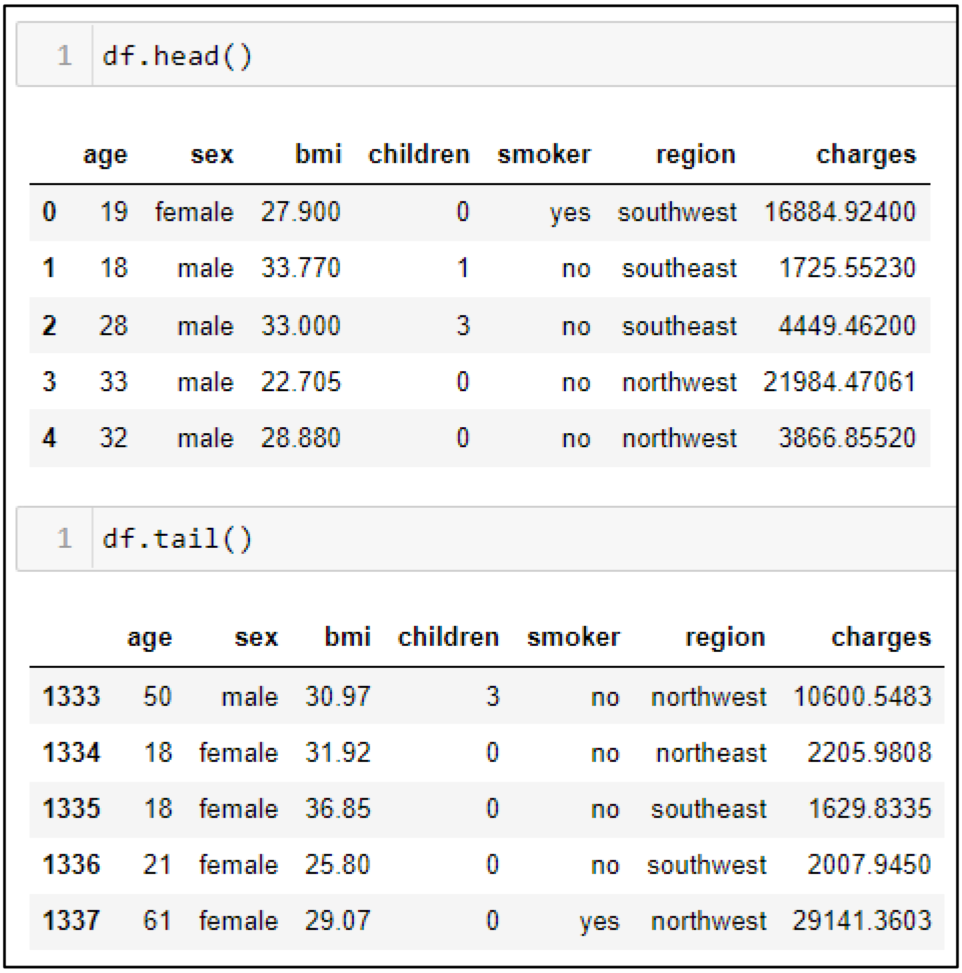
We first read the dataset, “Axisinsurance.csv” by using read csv function in pandas library.



```python
df = pd.read_csv('Axisinsurance.csv') # read the data as a data frame
```

Figure 3.1.0: Read dataset “Axisinsurance.csv”

To get an overall idea of how our dataset looks like, we extract the first and last 5 rows of data from the dataset.



```python
df.head()
```

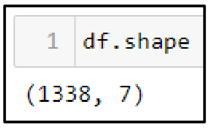
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

```python
df.tail()
```

	age	sex	bmi	children	smoker	region	charges
1333	50	male	30.97	3	no	northwest	10600.5483
1334	18	female	31.92	0	no	northeast	2205.9808
1335	18	female	36.85	0	no	southeast	1629.8335
1336	21	female	25.80	0	no	southwest	2007.9450
1337	61	female	29.07	0	yes	northwest	29141.3603

Figure 3.1.1: First and last 5 rows of data

We can observe that the dataset has **1338 observations** and **7 features**.

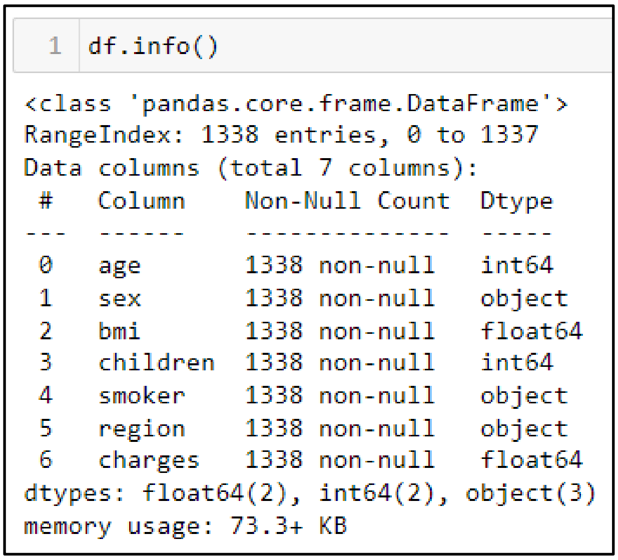


```
df.shape
```

```
(1338, 7)
```

Figure 3.1.2: Dimension of dataset

Besides, we can understand the data types of the columns of the dataset.

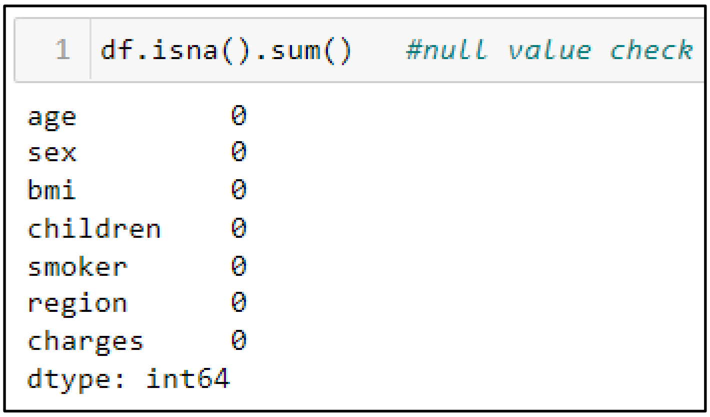


```
df.info()
```

```
<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
```

Figure 3.1.3: Data types of dataset’s columns

Most importantly, we will now check if there are any missing or null values in the dataset. We can see that there are no missing values.

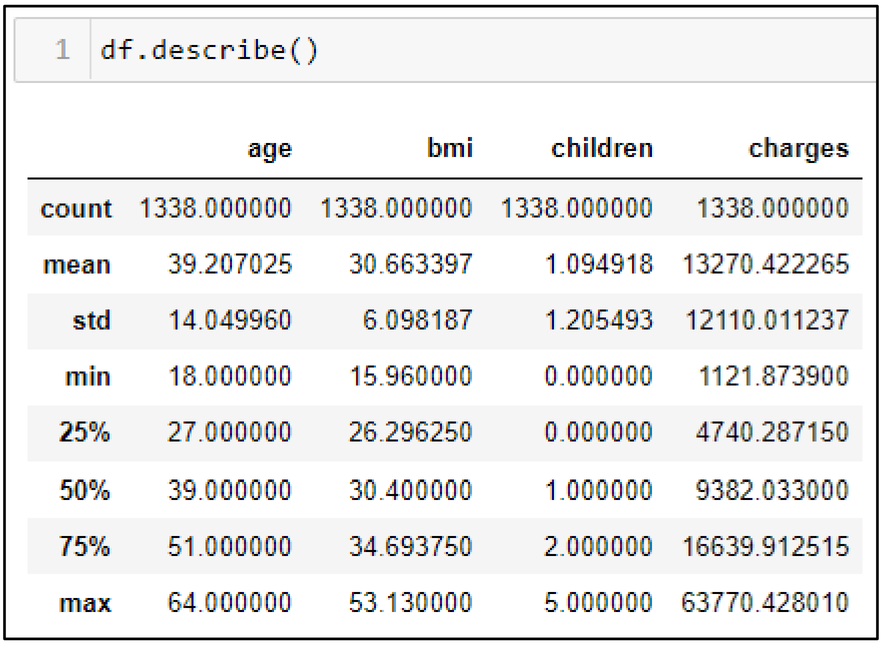


```
df.isna().sum()   #null value check
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Figure 3.1.4: Check missing or null values in dataset

Lastly, we will observe the descriptive statistics of the numerical columns in the dataset. This will enable us to understand the distribution of each column.



```
df.describe()
```

	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

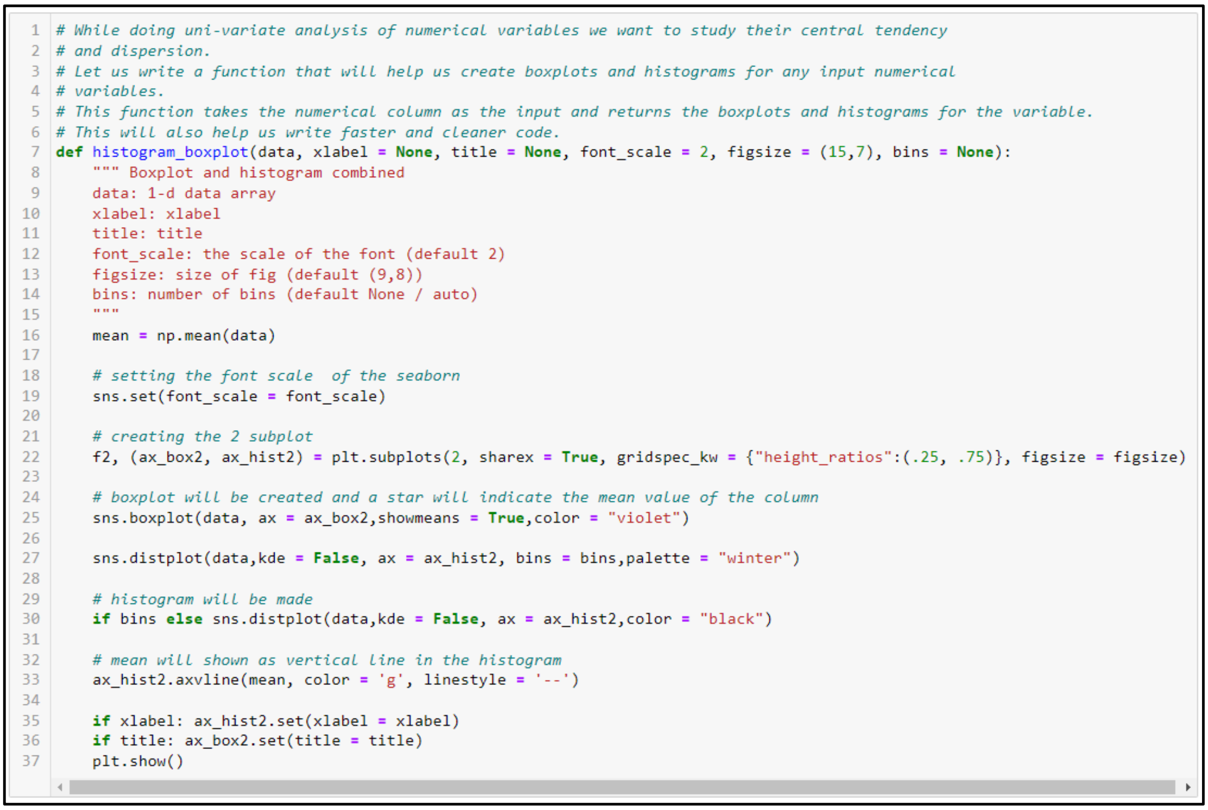
Figure 3.1.5: Descriptive analysis of numerical columns in dataset

Based on the descriptive analysis, we can presume the followings:

- Data looks legitimate as all the statistics seem reasonable
- In age column, data looks representative of the true age distribution of the adult population
- Very few people have more than 2 children. 75% of the people have 2 or less children
- The claimed amount, charges column has rightly skewed distribution as majority of the people would only require basis medical care while minority of people suffer from disease which cost more

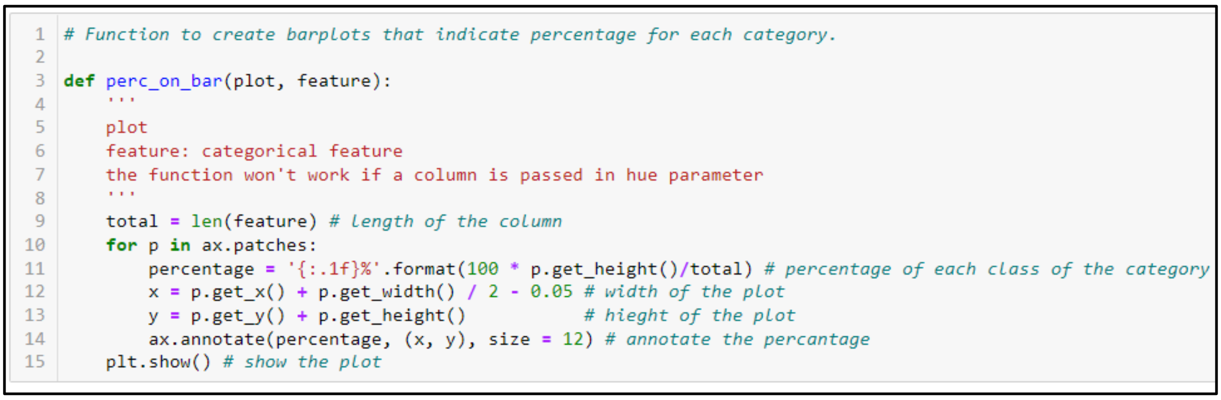
3.2 Univariate analysis

Univariate analysis refers to the analysis of one variable. We can get the idea of overall distribution of that variable for instance its central tendency and dispersion.



```python
# While doing uni-variate analysis of numerical variables we want to study their central tendency
# and dispersion.
# Let us write a function that will help us create boxplots and histograms for any input numerical
# variables.
# This function takes the numerical column as the input and returns the boxplots and histograms for the variable.
# This will also help us write faster and cleaner code.
def histogram_boxplot(data, xlabel = None, title = None, font_scale = 2, figsize = (15,7), bins = None):
    """ Boxplot and histogram combined
    data: 1-d data array
    xlabel: xlabel
    title: title
    font_scale: the scale of the font (default 2)
    figsize: size of fig (default (9,8))
    bins: number of bins (default None / auto)
    """
    mean = np.mean(data)

    # setting the font scale  of the seaborn
    sns.set(font_scale = font_scale)

    # creating the 2 subplot
    f2, (ax_box2, ax_hist2) = plt.subplots(2, sharex = True, gridspec_kw = {"height_ratios":(.25, .75)}, figsize = figsize)

    # boxplot will be created and a star will indicate the mean value of the column
    sns.boxplot(data, ax = ax_box2,showmeans = True,color = "violet")

    sns.distplot(data,kde = False, ax = ax_hist2, bins = bins,palette = "winter")

    # histogram will be made
    if bins else sns.distplot(data,kde = False, ax = ax_hist2,color = "black")

    # mean will shown as vertical line in the histogram
    ax_hist2.axvline(mean, color = 'g', linestyle = '--')

    if xlabel: ax_hist2.set(xlabel = xlabel)
    if title: ax_box2.set(title = title)
    plt.show()
```

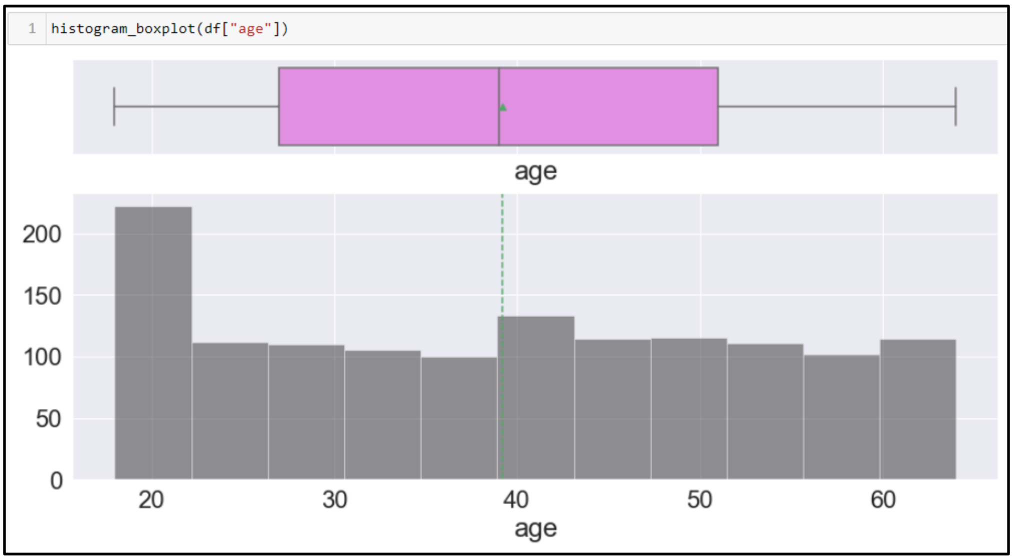
Figure 3.2.0: Source code for visualization in univariate analysis of numerical variable



```python
# Function to create barplots that indicate percentage for each category.

def perc_on_bar(plot, feature):
    '''
    plot
    feature: categorical feature
    the function won't work if a column is passed in hue parameter
    '''
    total = len(feature) # length of the column
    for p in ax.patches:
        percentage = '{:.1f}%'.format(100 * p.get_height()/total) # percentage of each class of the category
        x = p.get_x() + p.get_width() / 2 - 0.05 # width of the plot
        y = p.get_y() + p.get_height()           # hieght of the plot
        ax.annotate(percentage, (x, y), size = 12) # annotate the percantage
    plt.show() # show the plot
```

Figure 3.2.1: Source code for visualization in univariate analysis of categorical variable

3.2.1 Age column

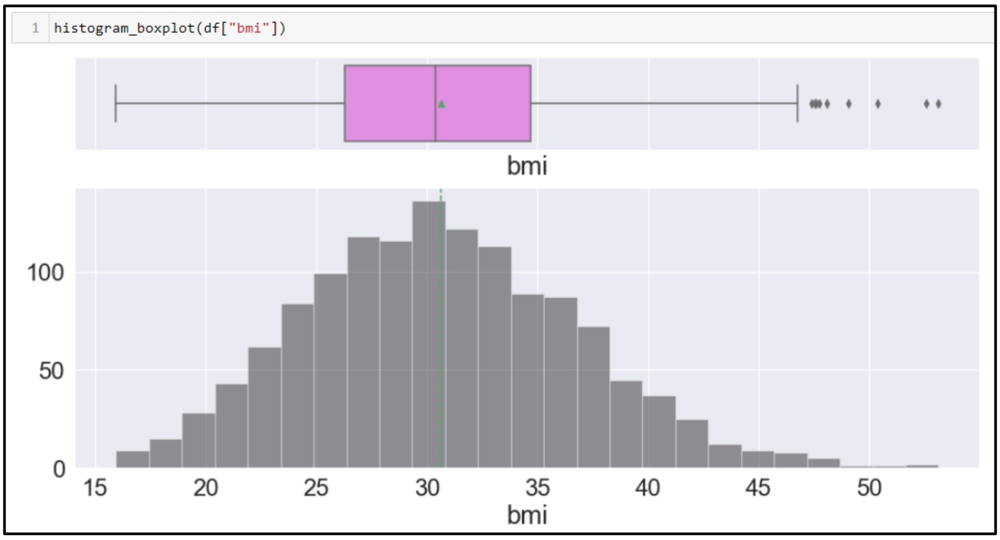


```
histogram_boxplot(df["age"])
```

Figure 3.2.1.0: Box plot and histogram of age column

Based on the plot, we can infer that data in age column is uniformly distributed as mean and median are around 40 years.

3.2.2 BMI column

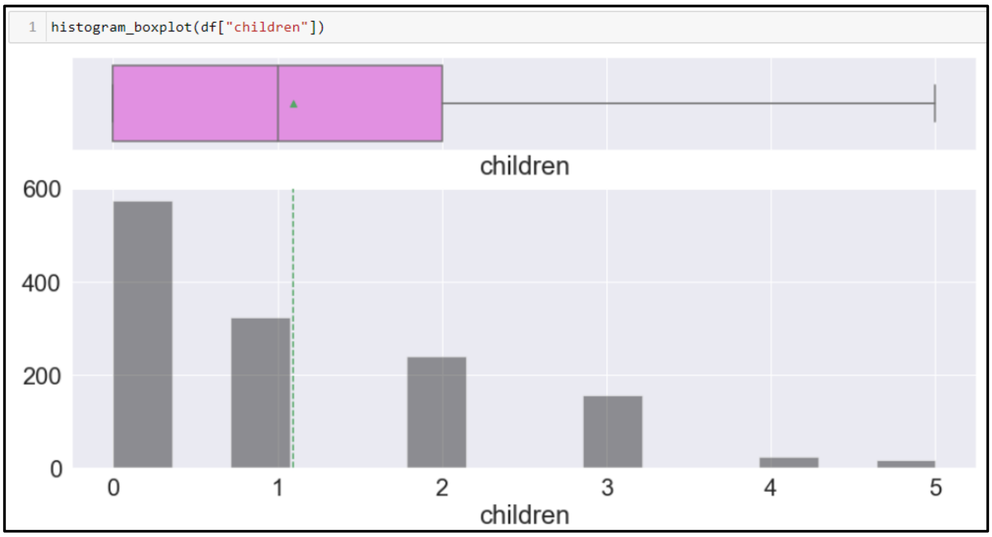


```
histogram_boxplot(df["bmi"])
```

Figure 3.2.2.0: Box plot and histogram of BMI column

Based on the plot, we can infer that data in BMI column has a fairly normal distribution as mean and median are around 30 kg/m^2.

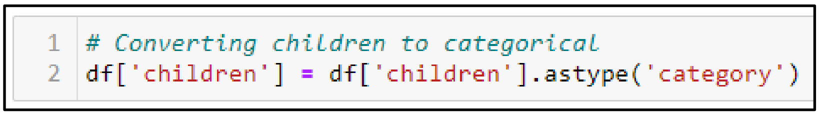
3.2.3 Children column



```
histogram_boxplot(df["children"])
```

Figure 3.2.3.0: Box plot and histogram of children column

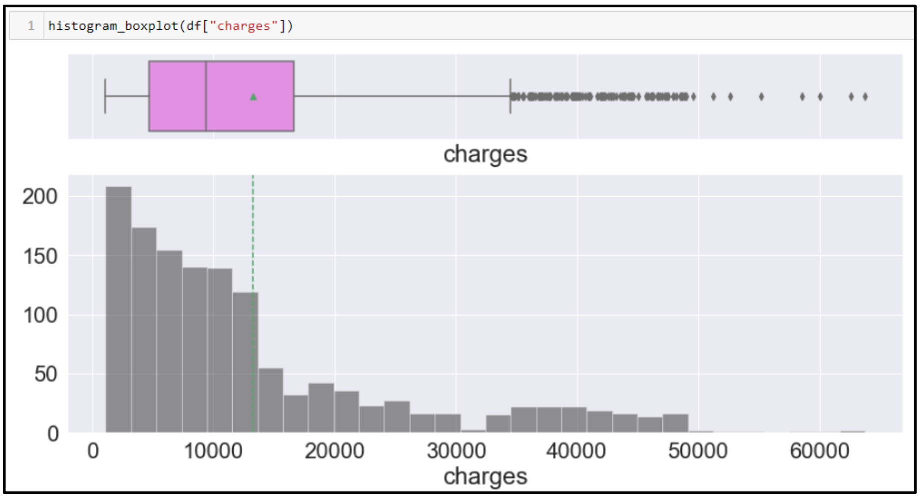
Based on the plot, we can infer that data in children column has a right skewed distribution. Since children column has only 6 different values as its data, we can make it into categorical for gain a more meaningful insight.



```
# Converting children to categorical
df['children'] = df['children'].astype('category')
```

Figure 3.2.3.1: Convert children column from numerical to categorical

3.2.4 Charges column

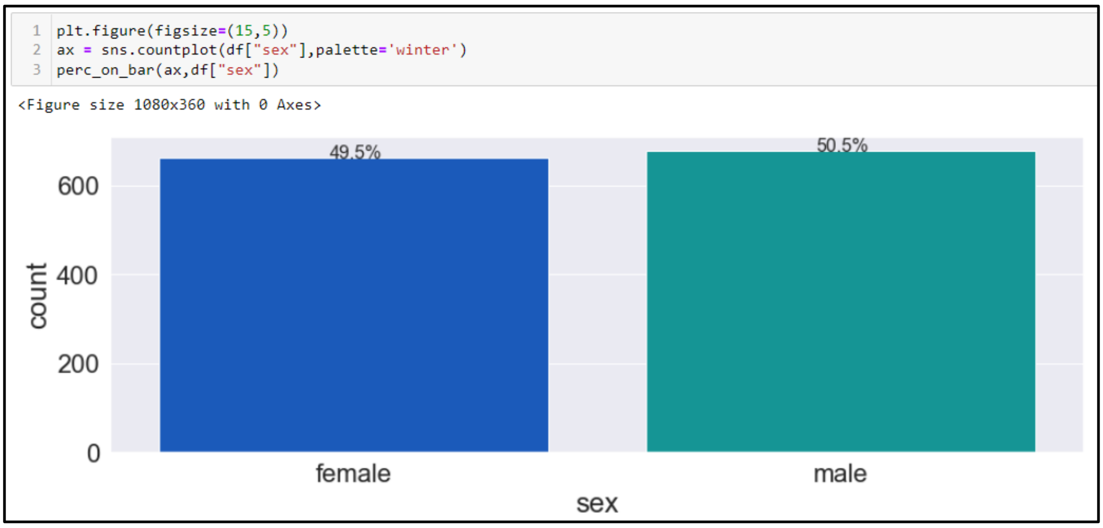


```
histogram_boxplot(df["charges"])
```

Figure 3.2.4.0: Box plot and histogram of charges column

Based on the plot, we can infer that data in charges column has a right skewed distribution as the mean is greater than the median (*13270.4223 > 9382.0330*). This is because of the outliers towards the higher end indicating that some people spend very high on their medicals.

3.2.5 Sex column



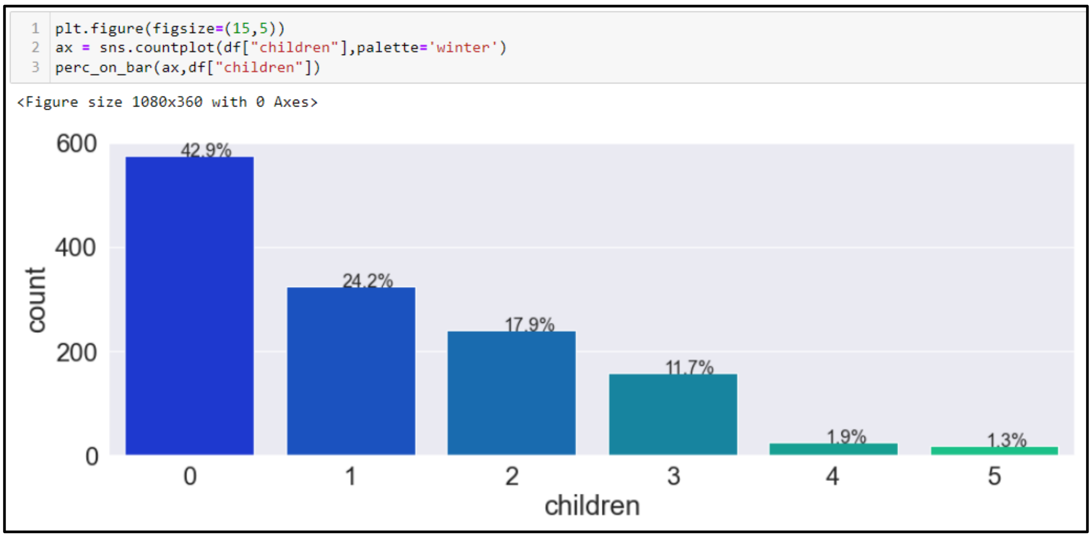
```
plt.figure(figsize=(15,5))
ax = sns.countplot(df["sex"],palette='winter')
perc_on_bar(ax,df["sex"])
```

```
<Figure size 1080x360 with 0 Axes>
```

Figure 3.2.5.0: Bar plot with percentage for sex column

Based on the plot, we can infer that the observations across genders is fairly distributed.

3.2.6 Children column



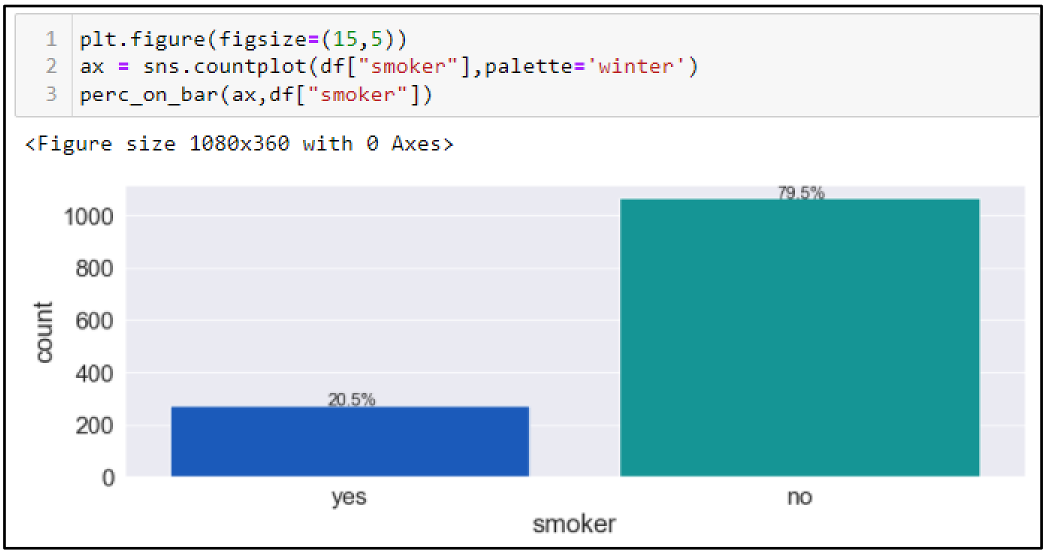
```
plt.figure(figsize=(15,5))
ax = sns.countplot(df["children"],palette='winter')
perc_on_bar(ax,df["children"])
```

```
<Figure size 1080x360 with 0 Axes>
```

Figure 3.2.6.0: Bar plot with percentage for children column

Based on the plot, we can infer that the 42% insurers do not have a child and nearly 42% of insurers have either 1 or 2 children.

3.2.7 Smoker column



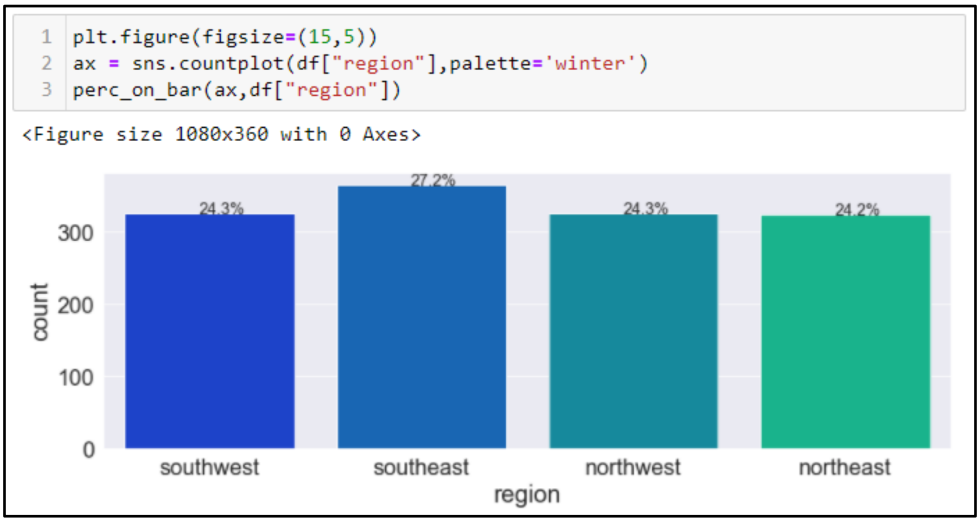
```
plt.figure(figsize=(15,5))
ax = sns.countplot(df["smoker"],palette='winter')
perc_on_bar(ax,df["smoker"])
```

```
<Figure size 1080x360 with 0 Axes>
```

Figure 3.2.7.0: Bar plot with percentage for smoker column

Based on the plot, we can infer that out of all the insurers in the dataset, around 20% of them are smokers.

3.2.8 Region column



```
plt.figure(figsize=(15,5))
ax = sns.countplot(df["region"],palette='winter')
perc_on_bar(ax,df["region"])
```

```
<Figure size 1080x360 with 0 Axes>
```

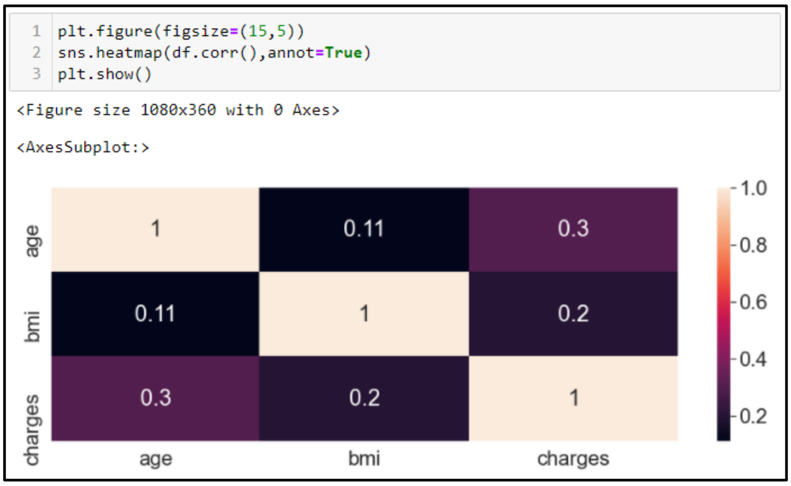
Figure 3.2.8.0: Bar plot with percentage for region column

Based on the plot, we can infer that the distribution of insurers across various regions of United States of America is fairly uniform. South east region does have around 3% more observations as compared to others.

3.3 Bivariate analysis

Bivariate analysis can be understood as exploring the relationship between 2 variables from the dataset. In the last part of this section, we will investigate the relationship between smoking and amount of medical claim using dedicated plots.

3.3.1 Explore correlation between numerical variables



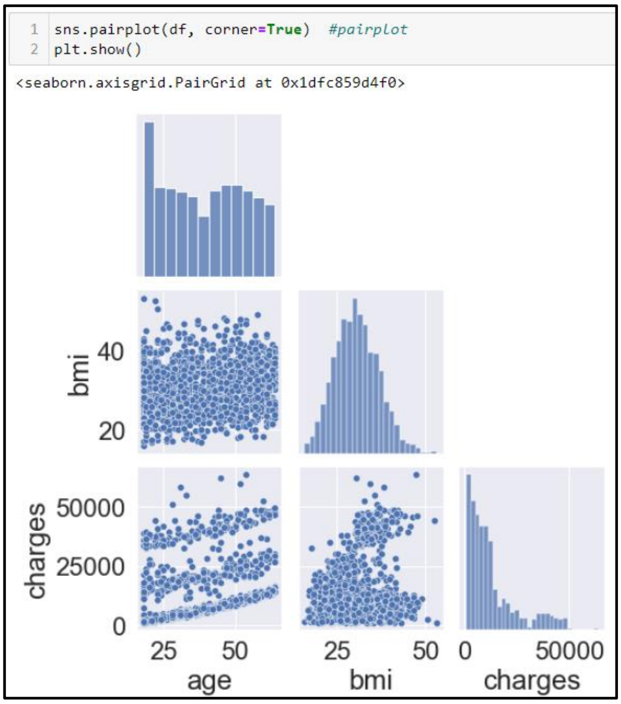
```
plt.figure(figsize=(15,5))
sns.heatmap(df.corr(),annot=True)
plt.show()
```

```
<Figure size 1080x360 with 0 Axes>

<AxesSubplot:>
```

	age	bmi	charges
age	1	0.11	0.3
bmi	0.11	1	0.2
charges	0.3	0.2	1

Figure 3.3.1.0: Heatmap to explore correlation relationship between numerical variables

Based on the plot, we can deduce that there’s no correlation between all the continuous variable.



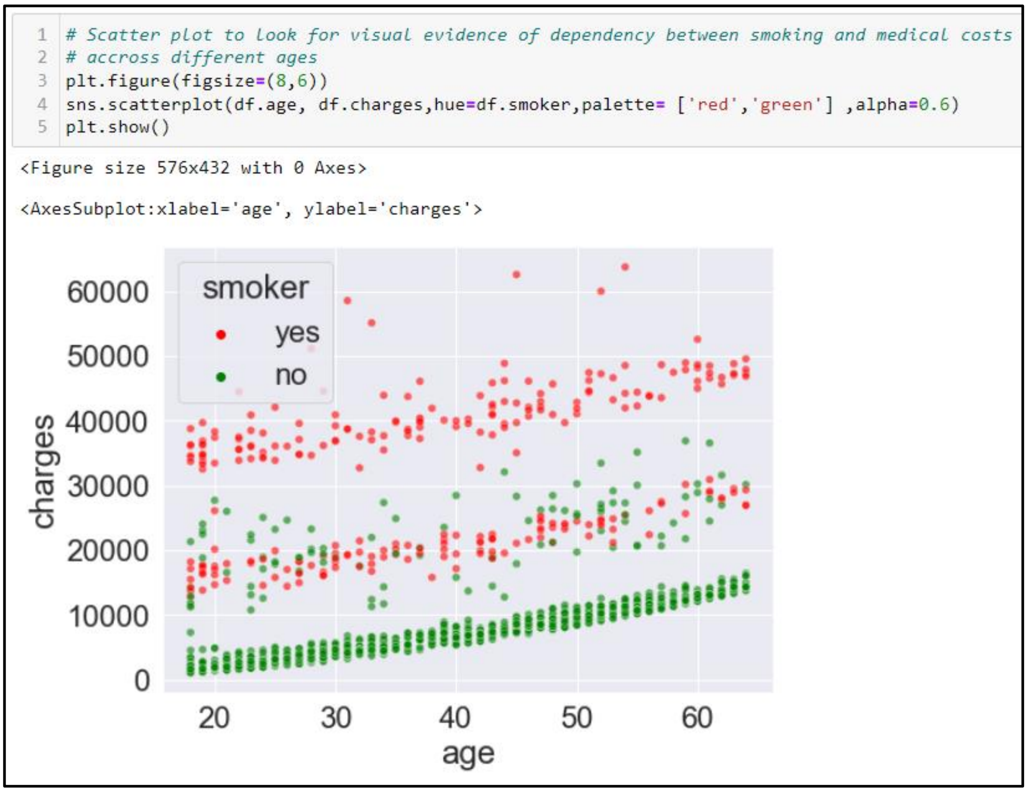
```
sns.pairplot(df, corner=True)  #pairplot
plt.show()
```

```
<seaborn.axisgrid.PairGrid at 0x1dfc859d4f0>
```

Figure 3.3.1.1: Pair plot to explore relationship and distribution of all continuous variables

Based on the plot, we can deduce that there’s an interesting pattern between “age” and “charges” column. It is possible that for the same ailment, other people are charged more than the younger ones.

3.3.2 Explore relationship between age and amount of medical claim explained with smoking



```
# Scatter plot to look for visual evidence of dependency between smoking and medical costs
# accross different ages
plt.figure(figsize=(8,6))
sns.scatterplot(df.age, df.charges,hue=df.smoker,palette= ['red','green'] ,alpha=0.6)
plt.show()
```

```
<Figure size 576x432 with 0 Axes>

<AxesSubplot:xlabel='age', ylabel='charges'>
```

Figure 3.3.2.0: Scatter plot of charges against age explained with smoking

Based on the scatter plot, we can infer that there is difference between charges for smokers and non-smokers as the amount of medical claim for non-smokers are much lower compared to the smokers.

4.0 Hypothesis Testing for Smoking Affects Amount of Medical Claim

As mentioned earlier, hypothesis testing is used to test the validity of a claim (*null hypothesis*) that is made about a population using sample data. The *alternative hypothesis* is the one to believe of the null hypothesis is concluded to be false.

In order to examine if the null hypothesis is true, we will use p value to check if our claim is statistically significant based on the significance level. Significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis. For instance, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

On the other hand, we can understand p-value in a continuous probability context, or in a distribution, we are only interested in adding more extreme values to the p-value rather than the rarer values. Therefore, if it's a one-tailed test, we will only be interested in finding the area of upper or lower bound region of the distribution from the point of our interest. On the other hand, if it's a two-tailed test, we are interested in finding the area of upper or lower bound region of the distribution from the point of our interest, as well as the symmetrical point from our point of interest.

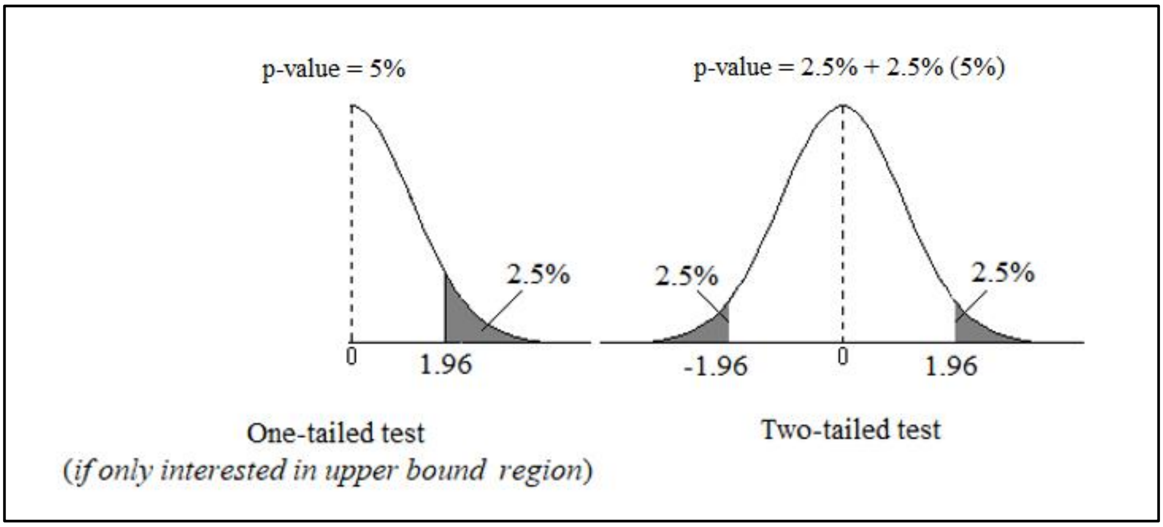


Figure 4.0.0: Conceptual diagram of p-value for one-tailed and two-tailed test

In this hypothesis testing, we would like to prove that insurer who is smoker has the medical claim amount greater than insurer who is non-smoker. Having understood our problem statement, we can construct our null hypothesis and alternative hypothesis statements as follow:

Null hypothesis, H_0: Mean amount of medical claim of smokers is less than or equal to non-smokers

Alternative hypothesis, H_a: Mean amount of medical claim of smokers is greater than non-smokers

In this project, we are only interested in whether the amount of claim amount is greater than average, we will conduct a one-tailed t test. The next step is to decide which inferential test method to conduct our hypothesis experiment.

Since both of our interested group, smokers and non-smokers are come from same population, and we would like to find out if there are any differences between the mean amount of medical claim of smokers and non-smoker, we can use independent t test.

In order to conduct an independent t test, there are several assumptions to be fulfilled.

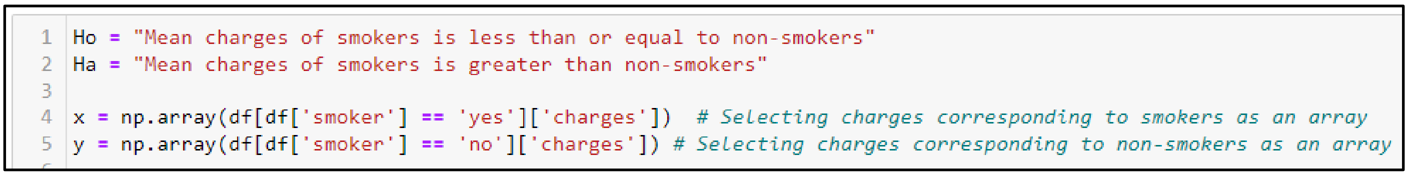
Assumption	Validity
Dependent variable is continuous	Since dependent variable in this project is average amount of medical claim and it's continuous, it's valid.
Independent variable that is categorical	Since the independent variable for the test is a categorical variable, smoker or non-smoker; it's valid.
Independent observations	Since whether subject 1 is smoker does not affect if subject 2 smoke, it's valid.
Homogeneity of variances	We need to ensure that variances of two interested subject group (smoker and non-smoker) are same. **Yet to be proven**.

To prove the homogeneity of our subject groups, we can use Levene's Test. In order to carry out Leneve's test, we first need to extract the total count for both smokers and non-smokers. The hypotheses of Levene's test are as follow:

Null hypothesis, H_0: $\sigma_1^2 - \sigma_2^2 = 0$ ("the population variances of group 1 and 2 are equal")

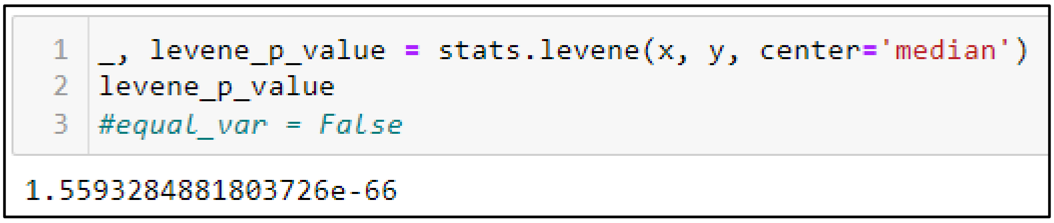
Alternative hypothesis, H_1: $\sigma_1^2 - \sigma_2^2 \neq 0$ ("the population variances of group 1 and 2 are not equal")

Levene's test indicates that the variances are equal across the two groups if the p-value is large and vice versa.



```python
Ho = "Mean charges of smokers is less than or equal to non-smokers"
Ha = "Mean charges of smokers is greater than non-smokers"

x = np.array(df[df['smoker'] == 'yes']['charges'])  # Selecting charges corresponding to smokers as an array
y = np.array(df[df['smoker'] == 'no']['charges']) # Selecting charges corresponding to non-smokers as an array
```

Figure 4.0.1: Extract number of counts of both smoker and non-smoker from dataset

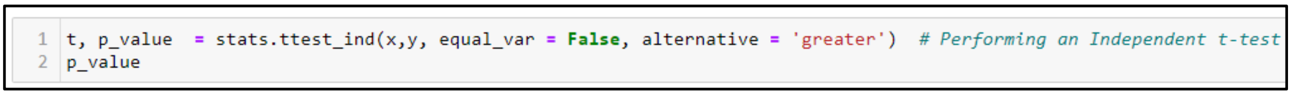


```python
_, levene_p_value = stats.levene(x, y, center='median')
levene_p_value
#equal_var = False
```

```
1.5593284881803726e-66
```

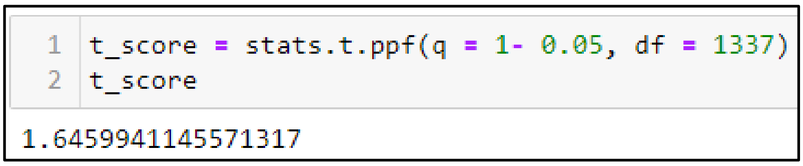
Figure 4.0.2: Leven's test and computed p-value

Based on figure 4.0.2, we can notice that p-value obtained from Levene's test is very low. It indicates that the assumption of homogeneity of variances for both of our smoker and non-smoker data values are not valid. Therefore, when we conduct independent t test, we need to indicate that it's the independent t test with equal variances not assumed. Since the test is a one-tailed test which interested in upper bound region, we pass the parameter "greater" in the alternative argument in the t test function.



```
t, p_value  = stats.ttest_ind(x,y, equal_var = False, alternative = 'greater')  # Performing an Independent t-test
p_value
```

Figure 4.0.3: Independent t test with equal variance not assumed and computed value

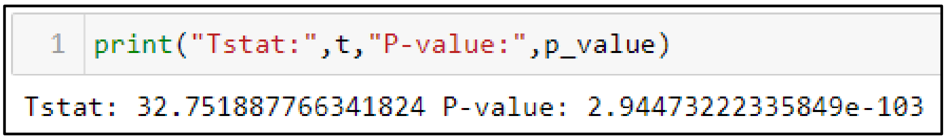


```
t_score = stats.t.ppf(q = 1- 0.05, df = 1337)
t_score
```

```
1.6459941145571317
```

Figure 4.0.4: T score with alpha value of 0.05

Based on the theory explained in figure 4.0.0, we do not need to divide the obtained p-value from t test by 2 because we are only interested in proving if mean amount of medical claim of smokers is higher than non-smoker, a one-tailed test.



```
print("Tstat:",t,"P-value:",p_value)
```

```
Tstat: 32.751887766341824 P-value: 2.94473222335849e-103
```

Figure 4.0.5: P-value for one-tailed test

Based on the p-value of almost 0, we can reject the null hypothesis, which is "Mean amount of medical claim of smokers is less than or equal to non-smokers". There is statistically significant evidence mean amount of medical claim of smokers is greater than non-smokers at a significance level of 0.05. The obtain p-value shows there is almost 0% chance that our result, the alternative hypothesis occurred because of random noise.

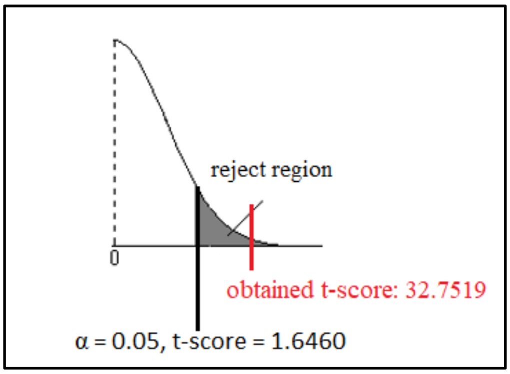


Figure 4.0.6: Conceptual diagram to illustrate rejection of null hypothesis for the project

5.0 Conclusion

Throughout this project, we have learnt the importance of data exploratory analysis to understand the relationships of features within data and its distribution as well as ensure the data is transformed and cleansed.

On the other hand, in order to prove our claim regarding on a population with sample, we can conduct a hypothesis testing. Levene’s test is one of the tools to identify if our subject groups violated the assumptions of homogeneity of variances in order for us to use independent t test in our hypothesis testing.

Lastly, p-value is sued to compared against the significant level of our hypothesis test. If the p-value if less than the significance level of the test, we can reject the null hypothesis and that the occurrence of statement in alternative hypothesis by random noise is very low.