Introduction:

This project aims to classify if a person smoker or not smoker using their health feature and personal details. We will explore at the data and make some visualization , and test different machine learning and deep learning algorithms to see which one works best for telling if someone is a smoker or not.

Executive Summary

Exploratory Data Analysis (EDA )

In this project, I visualized data before using machine learning. So it helps us understand the data better. Here's what I did:

Reading the Data : I used the ReaderData.py scode to load the data and see what's inside its .  
this is a first 5 rows from the data set :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age** | **Gender** | **BMI** | **Region** | **No. Children** | **Insurance Charges** | **Smoker** |
| **27** | **male** | **30.5** | **north** | **0** | **2494.022** | **no** |
| **41** | **female** | **31.02** | **north** | **0** | **6185.3208** | **no** |
| **28.47004946** | **male** | **26.66785034** | **north** | **4** | **5549.324781** | **no** |
| **50.30353686** | **female** | **30.8816603** | **south** | **4** | **11366.35084** | **no** |
| **50.77674293** | **male** | **26.84404232** | **south** | **3** | **25729.18463** | **yes** |

These columns represent the following :

Age : The age of a person

Gender : the gender of a person

BMI : body mass index  
Region : north or south  
No. children : number of children  
insurance Charges : the amount paid for the insurance company by the person

Smoker : yes for positive (smoker) and no otherwise .  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

we notice that there are two categorical columns, which means they contain non-numeric, string values . These columns are:

**Gender**: This column classifies into categories based on their gender such that male or female.

**Region**: This column classifies into categories based on their gender such that north or south.

**Smoker** : This column classifies into categories based on their smoker such that yes or no.

So the algorithm that will be used in this project its need the numerical data to work successfully .   
so must be convert these categories values to numerical values .   
there is a many algorithm that using to convert to numeric values but since the data is a binary categories I am used a label encoding approach to handle a non numeric values .

Therefore, I adopted the following for non-numeric columns as this :   
for Gender Column : male = 1 and female = 0 .   
for Region Column : south = 1 and north = 0 .

For Smoker : yes = 1 and no = 0 ;

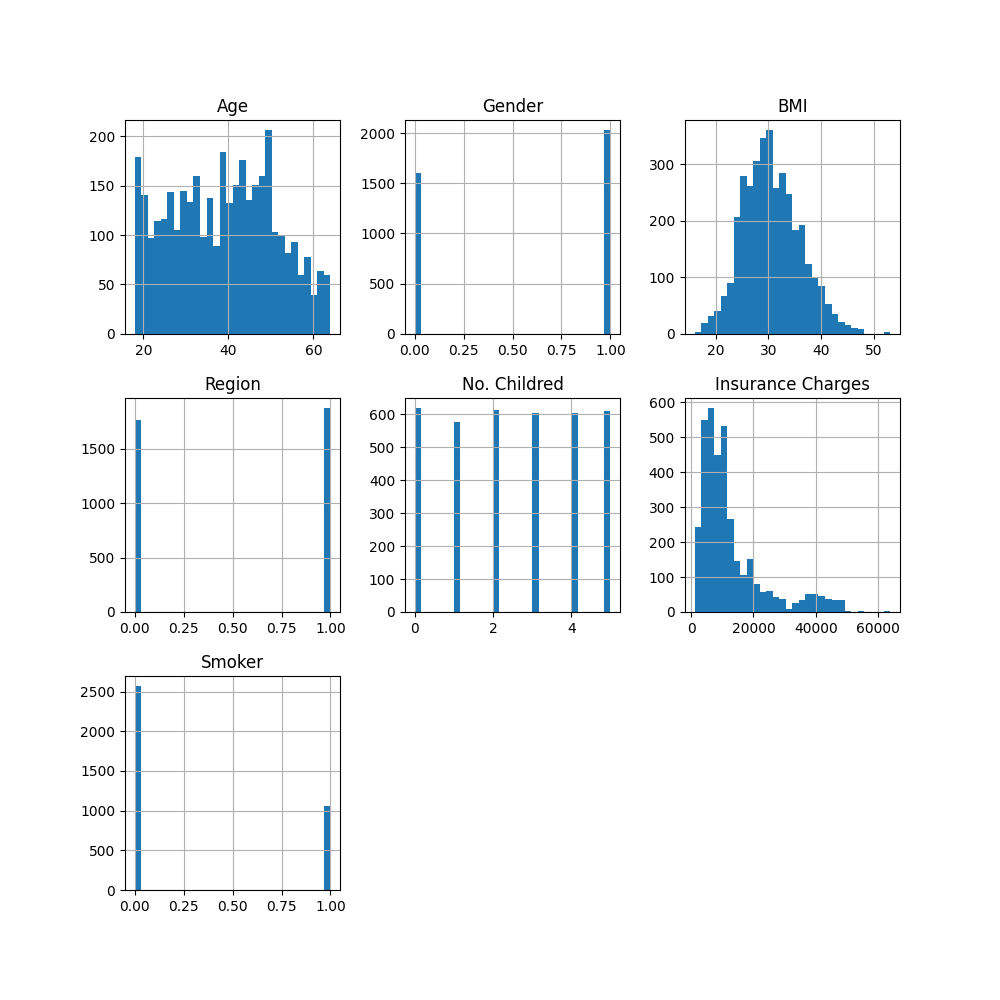
Now after encoding the non numeric values to numeric values . the data become as this :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age** | **Gender** | **BMI** | **Region** | **No. Children** | **Insurance Charges** | **Smoker** |
| **27.000000** | **1** | **30.50000** | **1** | **0** | **2494.022000** | **0** |
| **41.000000** | **0** | **31.02000** | **1** | **0** | **6185.320800** | **0** |
| **28.470049** | **1** | **26.66785** | **1** | **4** | **5549.324781** | **0** |
| **50.303537** | **0** | **30.88166** | **0** | **4** | **11366.350840** | **0** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column** | **Count** | **Mean** | **Std Dev** | **Min** | **25%** | **50%** | **75%** | **Max** |
| **Age** | **3630.000** | **38.887** | **12.151** | **18.000** | **29.000** | **39.171** | **48.343** | **64.000** |
| **Gender** | **3630.000** | **0.559** | **0.497** | **0.000** | **0.000** | **1.000** | **1.000** | **1.000** |
| **BMI** | **3630.000** | **30.630** | **5.441** | **15.960** | **26.695** | **30.200** | **34.100** | **53.130** |
| **Region** | **3630.000** | **0.515** | **0.500** | **0.000** | **0.000** | **1.000** | **1.000** | **1.000** |
| **No. Children** | **3630.000** | **2.504** | **1.713** | **0.000** | **1.000** | **3.000** | **4.000** | **5.000** |
| **Insurance Charges** | **3630.000** | **12784.809** | **10746.167** | **1121.874** | **5654.818** | **9443.807** | **14680.408** | **63770.428** |
| **Smoker** | **3630.000** | **0.293** | **0.455** | **0.000** | **0.000** | **0.000** | **1.000** | **1.000** |

Here is a summery statistics about data set :

In the below image represent the visualization for dataset :



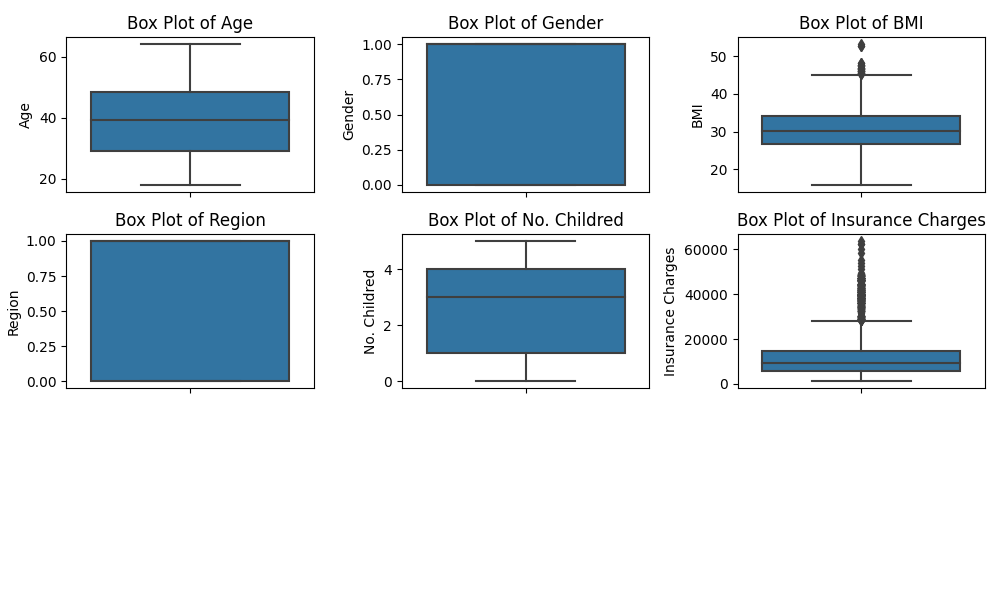
Data Cleaning :

1 - For my smoker dataset, we did not find any missing but there is a 668 Raws are completely duplicated .   
Therefore, I removed these duplicate rows from the data set. the size of data set after that = 2963 raw .

2- Conversion of Age Data from Double to Integer Values

Next, I notice that the "Age" column in the data set contain a decimal values i.e. float values , such as 48.44, but this is not correct . To fix this problem, I rounded all age values and converted them to integers .

3- Check the Outliers :

In the this steps I am visualize the data using box plot to determine if there is outliers or not .   


We conclude from the image above that there are outliers in the BMI and insurence charge columns .

Therefore This data must be cleaned by replacing it with appropriate values to increase the strength of the models later on .

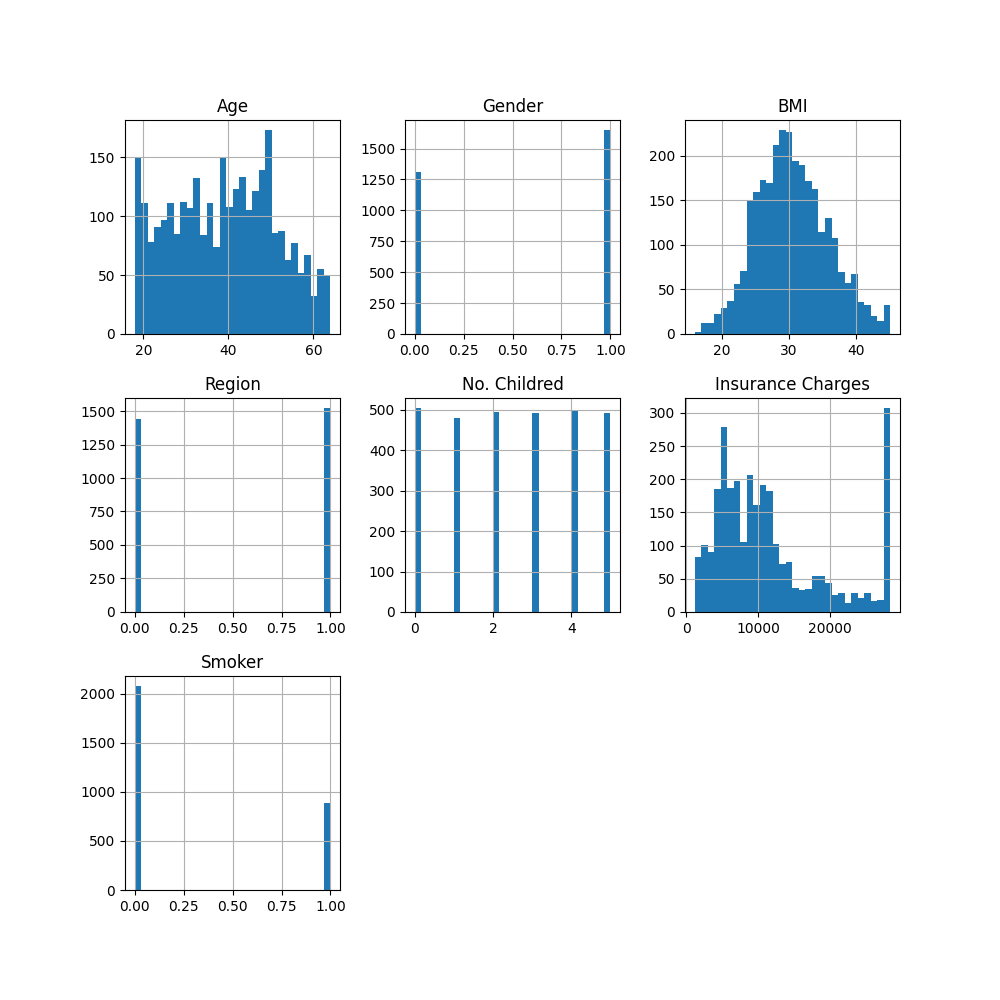
**Handling Outliers :**

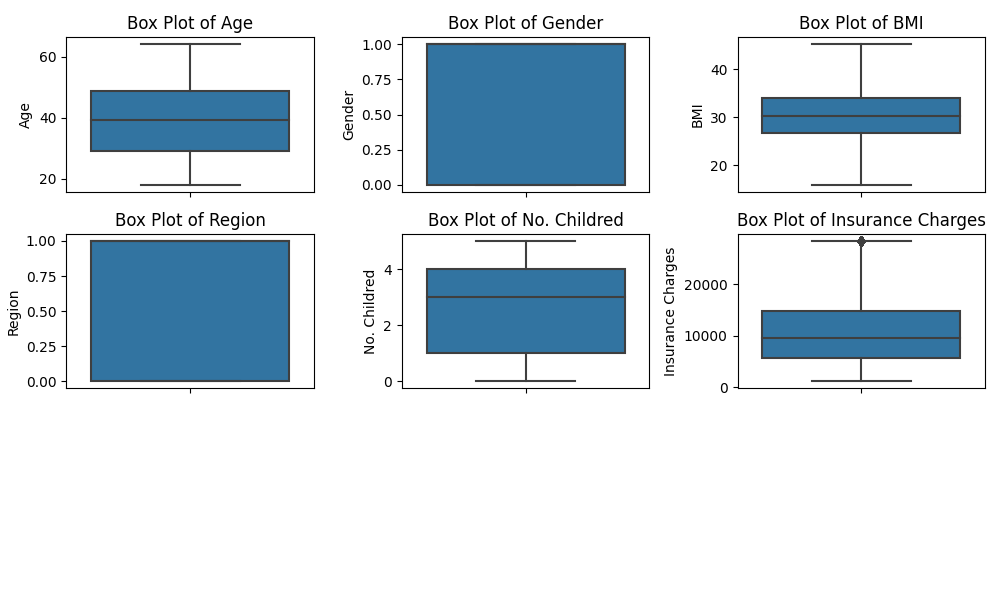
cap\_outliers, designed to reduce outliers. It calculates the acceptable data range using the interquartile range (IQR) and sets extreme values accordingly. Values below Q1 - 1.5IQR or above Q3 + 1.5IQR are set to these limits. This method maintains the integrity of the data while minimizing the impact of outliers

Data After Cleaning :   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The description of dataset after cleaning noise :

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column** | **Count** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| **Age** | **2962.000** | **39.023** | **12.20** | **18.0** | **29.000** | **39.350208** | **48.659834** | **64.000** |
| **Gender** | **2962.000** | **0.556** | **0.49** | **0.000** | **0.000** | **1.000** | **1.000** | **1.000** |
| **BMI** | **2962.000** | **30.608** | **5.36** | **15.0** | **26.730** | **30.210** | **34.099324** | **45.153310** |
| **Region** | **2962.000** | **0.51** | **0.499902** | **0.000** | **0.000** | **1.000** | **1.000** | **1.000** |
| **No. Children** | **2962.000** | **2.5** | **1.7118** | **0.000** | **1.000** | **3.000** | **4.000** | **5.000** |
| **Insurance Charges** | **2962.000** | **11705.78** | **7838.632999** | **1121.873900** | **5709.962155** | **9563.616073** | **14791.278335** | **28413.252605** |
| **Smoker** | **2962.000** | **0.298** | **0.457802** | **0.000** | **0.000** | **0.000** | **1.000** | **1.000** |

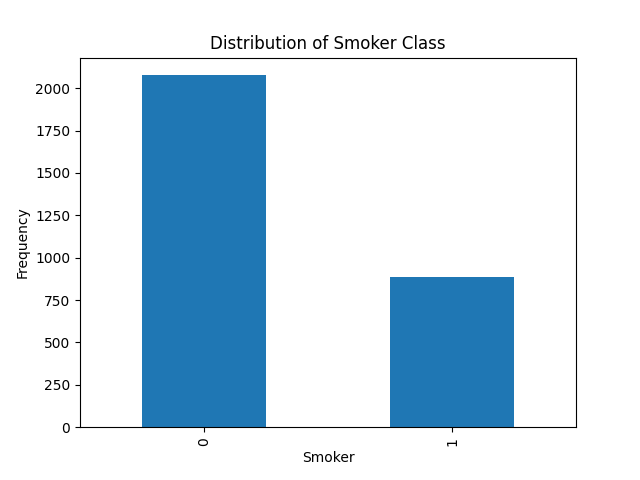




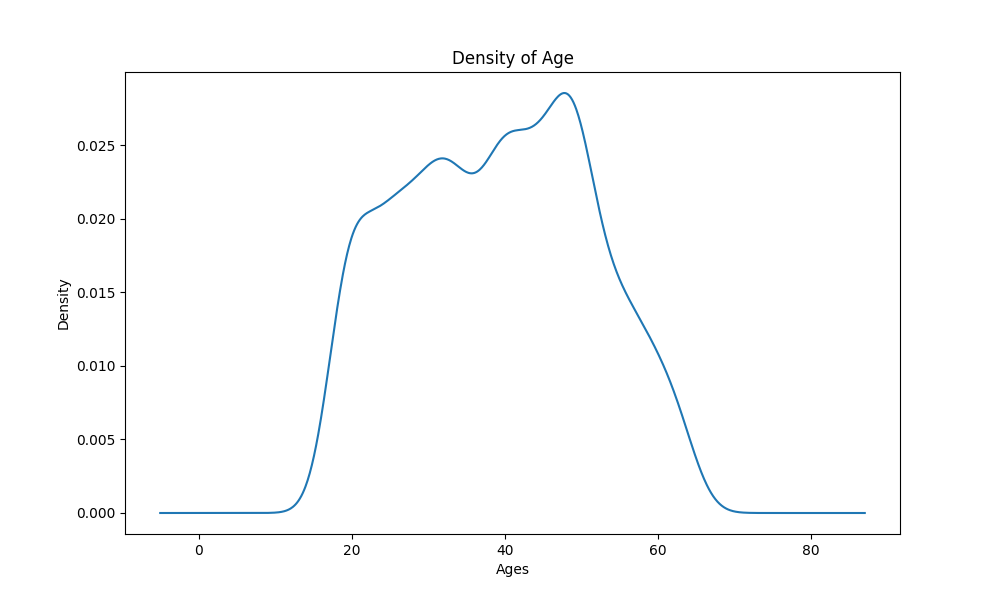
Data And Analysis :  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Histogram for Age Group :

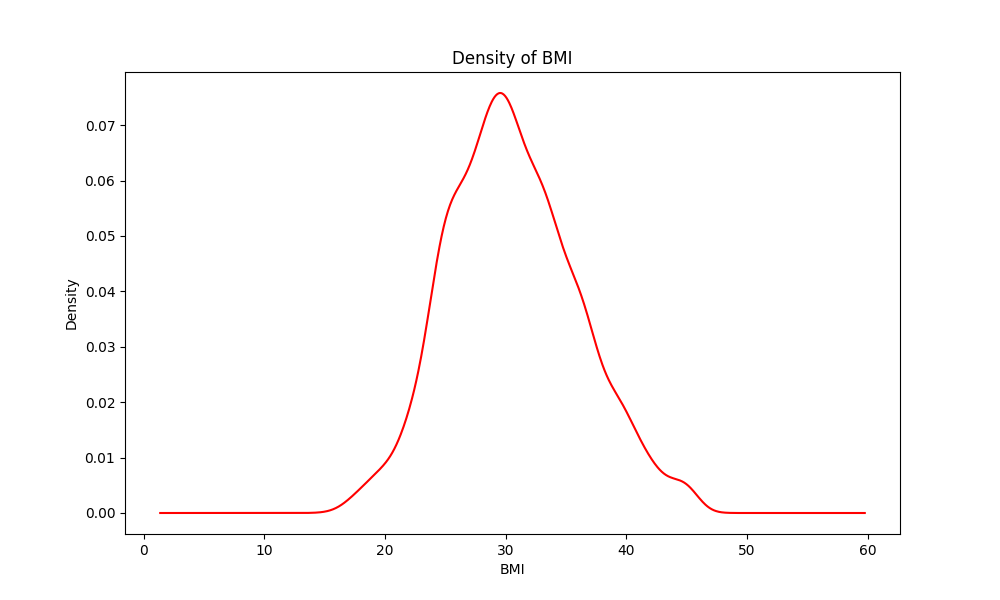
The chart shows the number of smoker and not smoker people.



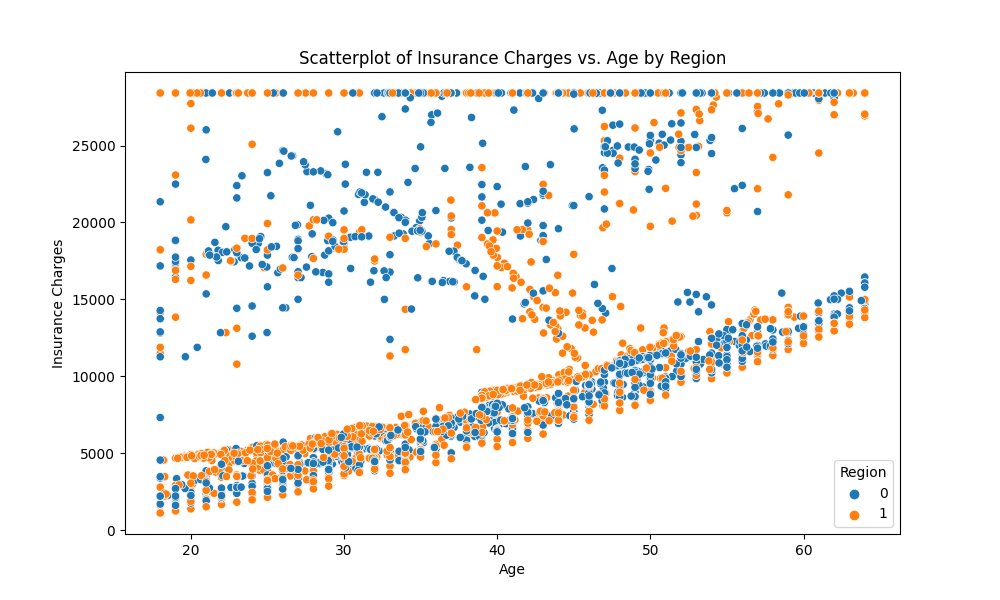
The density chart for age shows an increase in the age group between approximately 20 and 55 years of age and then a gradual decrease with increasing age.



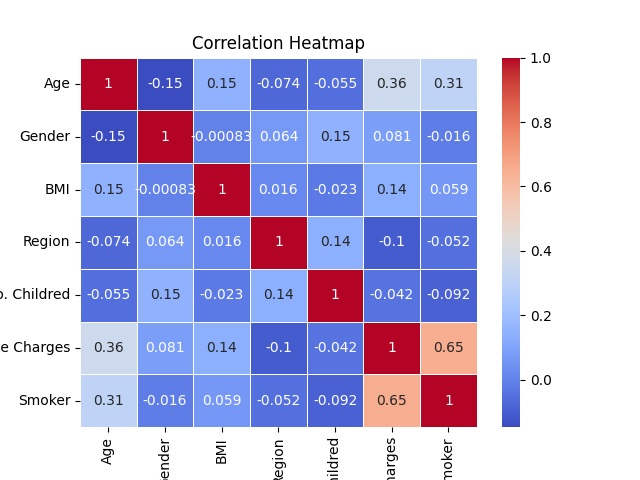
density plot for BMI, with a peak around the middle range at 30 age and tails off towards the lower and higher BMI values



Visualise the scatterplot of data and split based on Region attribute



Visualize the correlation between all features



**Insurance Charges:** has 0.65, which is a strong positive with Smoker coulmn . This means that people who are smokers tend to have higher insurance charges.

**Age** has 0.31, which is a middle positive link. This suggests that older people might be more likely to be smokers.

**Region** : has -0.052, which is a very weak negative link. This means there isn't much of a pattern to where people live and if they smoke.

**No . Children** : this means is that there is a slight tendency for individuals with more children to be is not smokers

**BMI** : This indicates that there is a very slight increase in BMI among smokers compared to not smokers

**Gender** : This indicates that there is hardly any relationship between gender and smoking status in this dataset

Normalization of Data (Feature Scalling ) :

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. [1]

In this project I applied the Min-Max scale, which is to normalize the data. This method transforms features by scaling each one to a specific range, from 0 to 1. The formula used is (value - min) / (max - min).

Split the Data Set :   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In this project, I divided the data into two parts: 80% for training, and 20% for testing. This means training the models and ensuring their validity and generalizability. So how to detect smokers and non-smokers.

I am use these feature as an inputs : Age , Insurance Charges , No.children .  
  
note : I am select a NO.children after trying several testing on model . without use it’s the result was not good . but when use it’s the result it become very good . so I am select its .

Machine Learning Algorithms

KNN Algorithm:  
in this algorithms chose to focus on four main features