**MEDICAL INSURANCE PRICE PREDICTION**

INTRODUTION:

*People are always confused about their medical insurance and don’t know the cost of insurance at different ages and conditions. This data is useful these people and useful to make predictions of the insurance cost they will have to pay.*

*Having a health insurance policy protects your savings from getting drained due to medical treatments. Medical emergencies are unpredictable , and with rising healthcare costs, quality management can get very expensive. Without medical insurance, there can be a rapid loss of savings.*

***Problem statement*** *: Medical Insurance Price prediction using multiple linear regression.*

***APPROACH:***

***Multiple linear regression*** *: multiple regression is like linear regression,but with more than one independent value,meaning that we try to predict a value based on two or more variables.It is basically used type for predictive analysis.*

*Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.*

*Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.*

Formula and Calculation of Multiple Linear Regression

*yi*​=*β*0​+*β*1​*xi*1​+*β*2​*xi*2​+...+*βp*​*xip*​+*ϵ*

**where, for***i*=*n***observations:**

*yi*​=dependent variable

*xi*​=explanatory variables

*β*0​=y-intercept (constant term)

*βp*=slope coefficients for each explanatory variable*ϵ*=the model’s error term (also known as the residuals)​

*INPUT :*

Age: It tells the age of the person

Sex: it tells about the person is man or women

bmi(bodymass index):it gives the information about the person’s health. If the range is is less than 18.5 ,it fall under under weight.If the range is between 18.5 to 25 ,the person is healthy.

Smoker:If the person is a smoker he may has more chances of getting health issues like lung cancer etc.

Region:insurance prices are changes region by region,under current law,you will need to buy health insurance in state in which you have permanent address.

Children:If the person has more children insurance price will be increases ,because the children health issues also comes under the parents insurance.

*OUTPUT*

*price of medical insurance*

*LIBRARIES USED:*

*1.import pandas as pd*

*(pandas allowing importing data from various files such as Excel,csv,JSON)*

*2.import numpy as np*

*(numpy can be used to perform a wide variety of mathematical operatins on array)*

*3.import matplotlib.pyplot as plt*

*(matpltlib is a cross-platform,data visualization and graphical plotting libra*

*4.import seaborn as sns*

*(It is also used for viualization and helps you explore and understand your data)*

*5. Label encoder:*

(*In machine learning, we usually deal with datasets that contain multiple labels in one or more than one columns. These labels can be in the form of words or numbers. To make the data understandable or in human-readable form, the training data is often labelled in words.*

*Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.*

*6.r\_2 score:*

*(It tells about the accuracy of the model)*

PROGRAM:

Step1:

# Importing Libraries:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

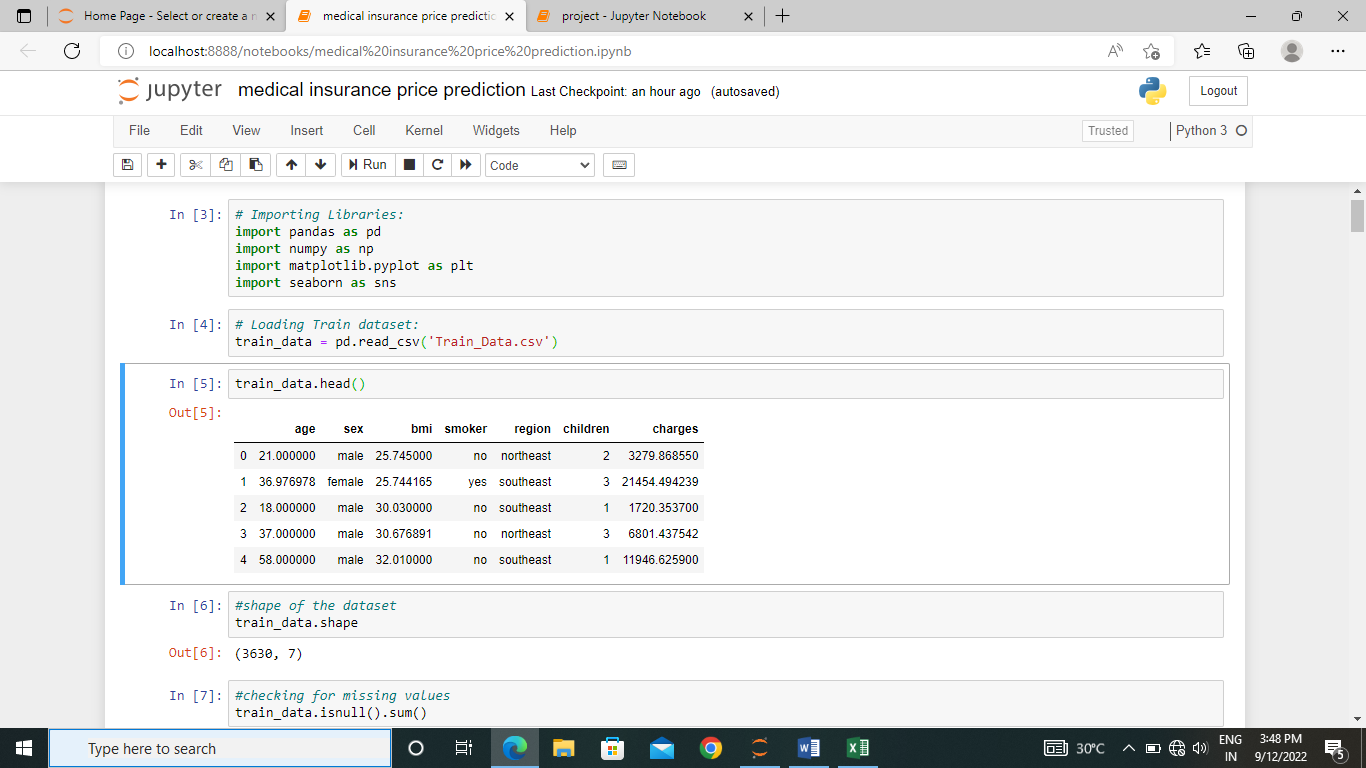
import seaborn as sns

step 2:

# Loading Train dataset:

train\_data = pd.read\_csv('Train\_Data.csv')

train\_data.head()



#shape of the dataset

train\_data.shape

(3630, 7)

Step 3:

#checking for missing values

train\_data.isnull().sum()

age 0

sex 0

bmi 0

smoker 0

region 0

children 0

charges 0

dtype: int64

#information about the dataset

train\_data.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 3630 entries, 0 to 3629

Data columns (total 7 columns):

age 3630 non-null float64

sex 3630 non-null object

bmi 3630 non-null float64

smoker 3630 non-null object

region 3630 non-null object

children 3630 non-null int64

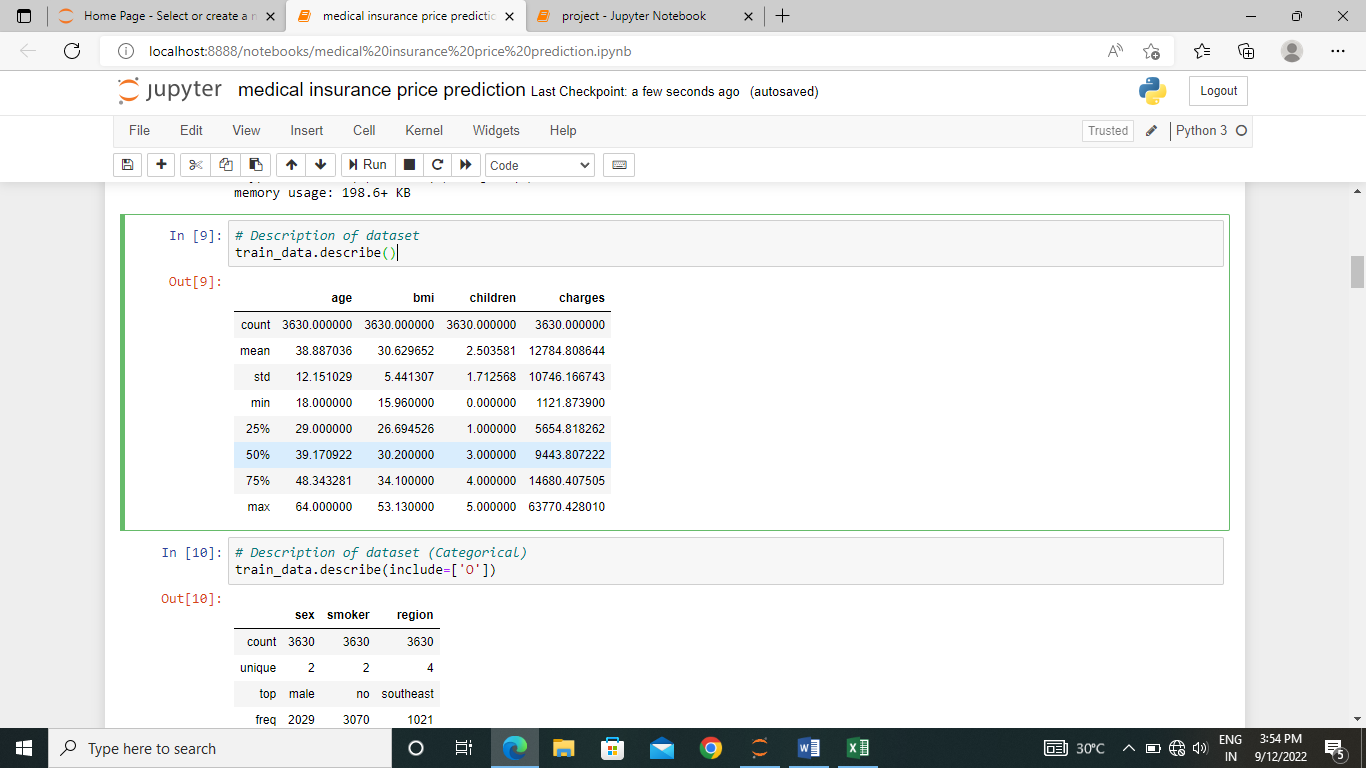
charges 3630 non-null float64

dtypes: float64(3), int64(1), object(3)

memory usage: 198.6+ KB

# Description of dataset

train\_data.describe()



Step 4 :

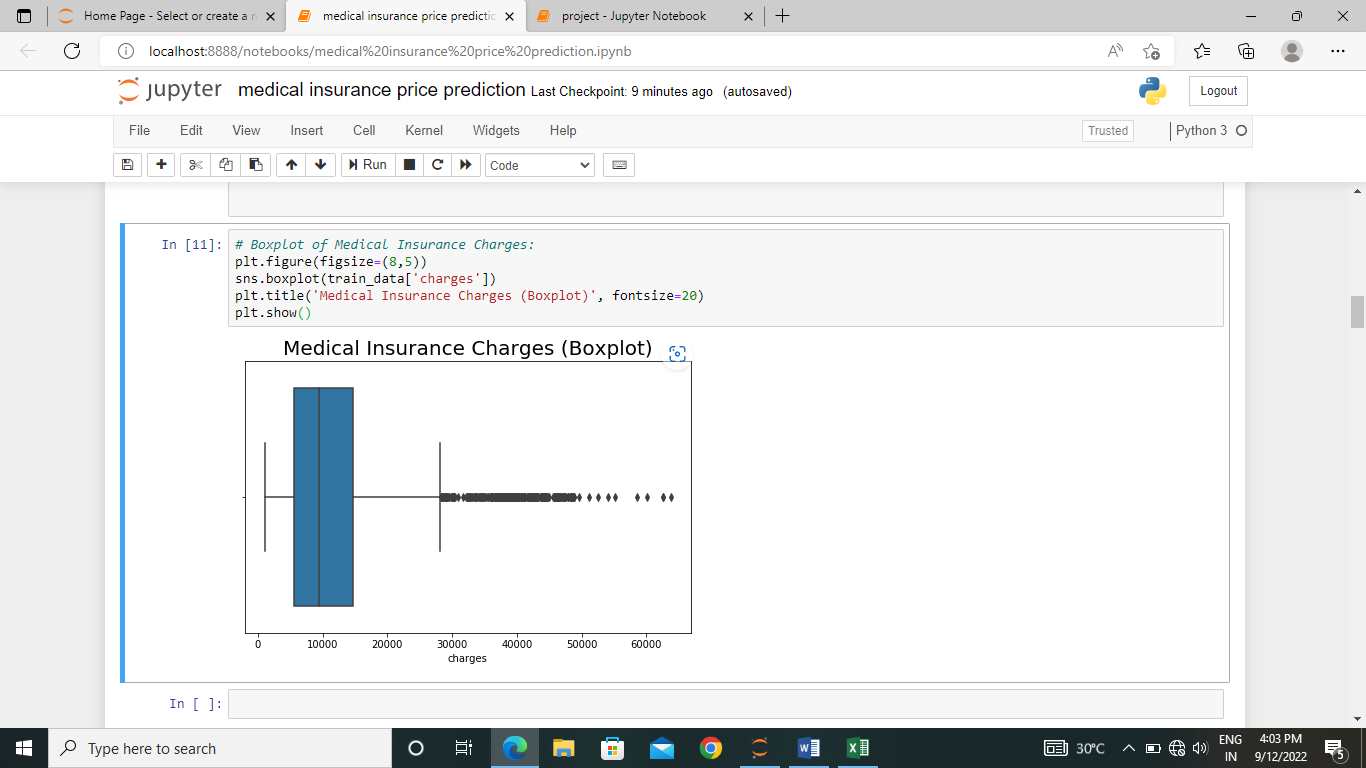
# Boxplot of Medical Insurance Charges:

plt.figure(figsize=(8,5))

sns.boxplot(train\_data['charges'])

plt.title('Medical Insurance Charges (Boxplot)', fontsize=20)

plt.show()



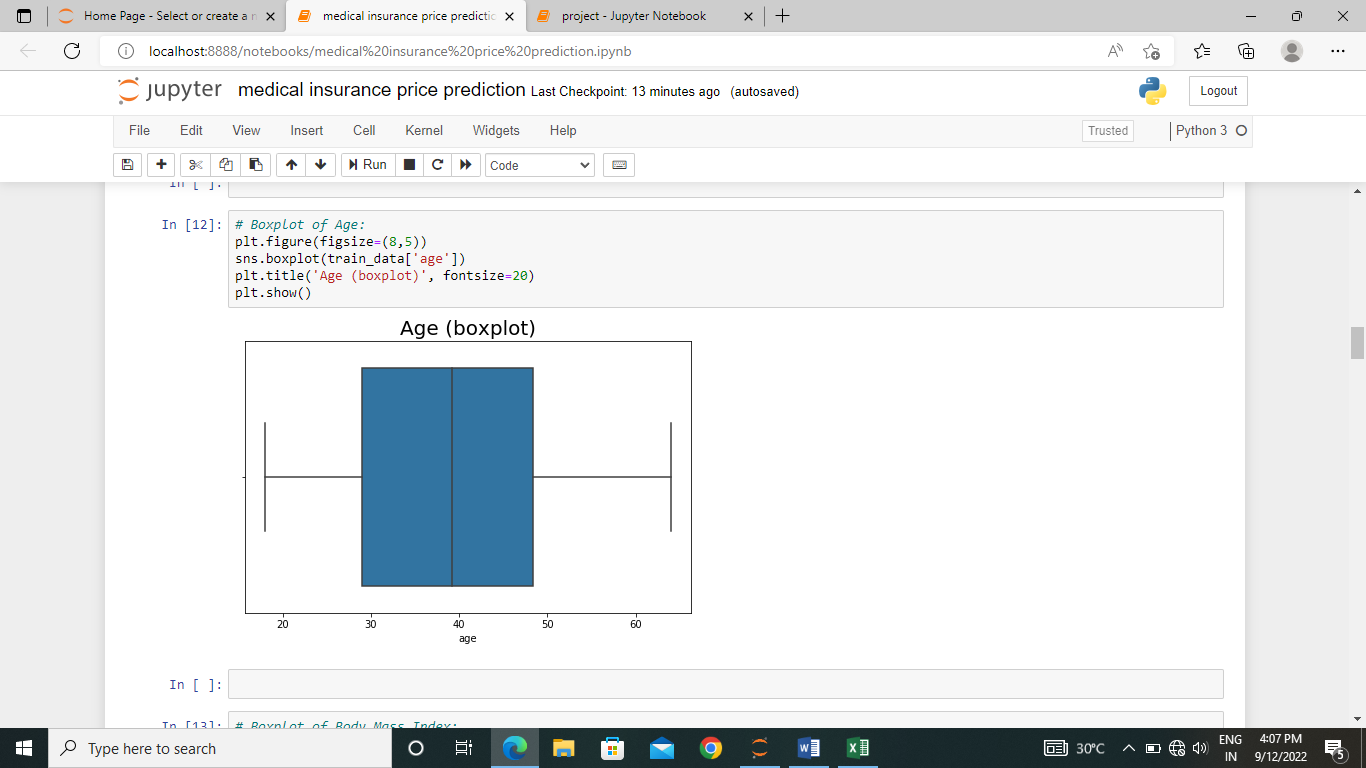
# Boxplot of Age:

plt.figure(figsize=(8,5))

sns.boxplot(train\_data['age'])

plt.title('Age (boxplot)', fontsize=20)

plt.show()



print("Male :", train\_data['sex'].value\_counts()[0])

print("Female :", train\_data['sex'].value\_counts()[1])

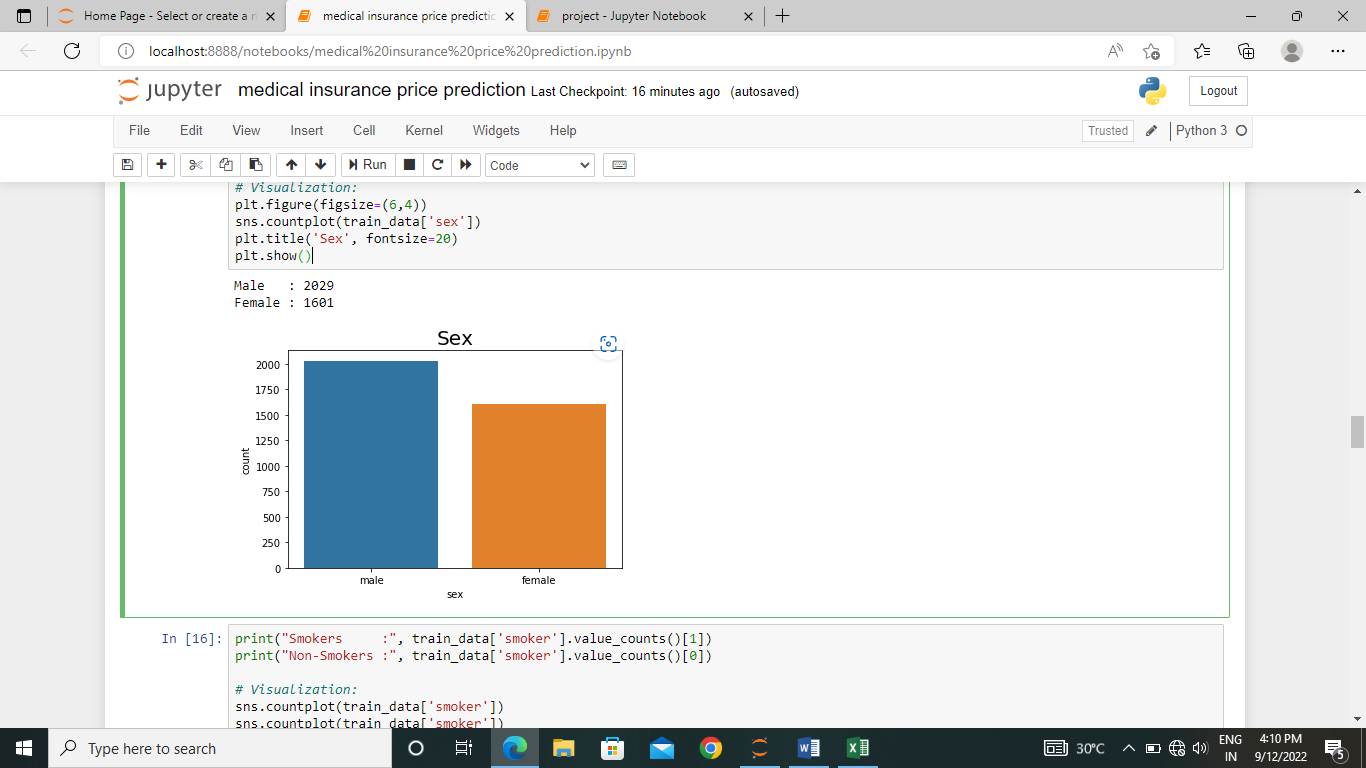
# Visualization:

plt.figure(figsize=(6,4))

sns.countplot(train\_data['sex'])

plt.title('Sex', fontsize=20)

plt.show()



print("Smokers :", train\_data['smoker'].value\_counts()[1])

print("Non-Smokers :", train\_data['smoker'].value\_counts()[0])

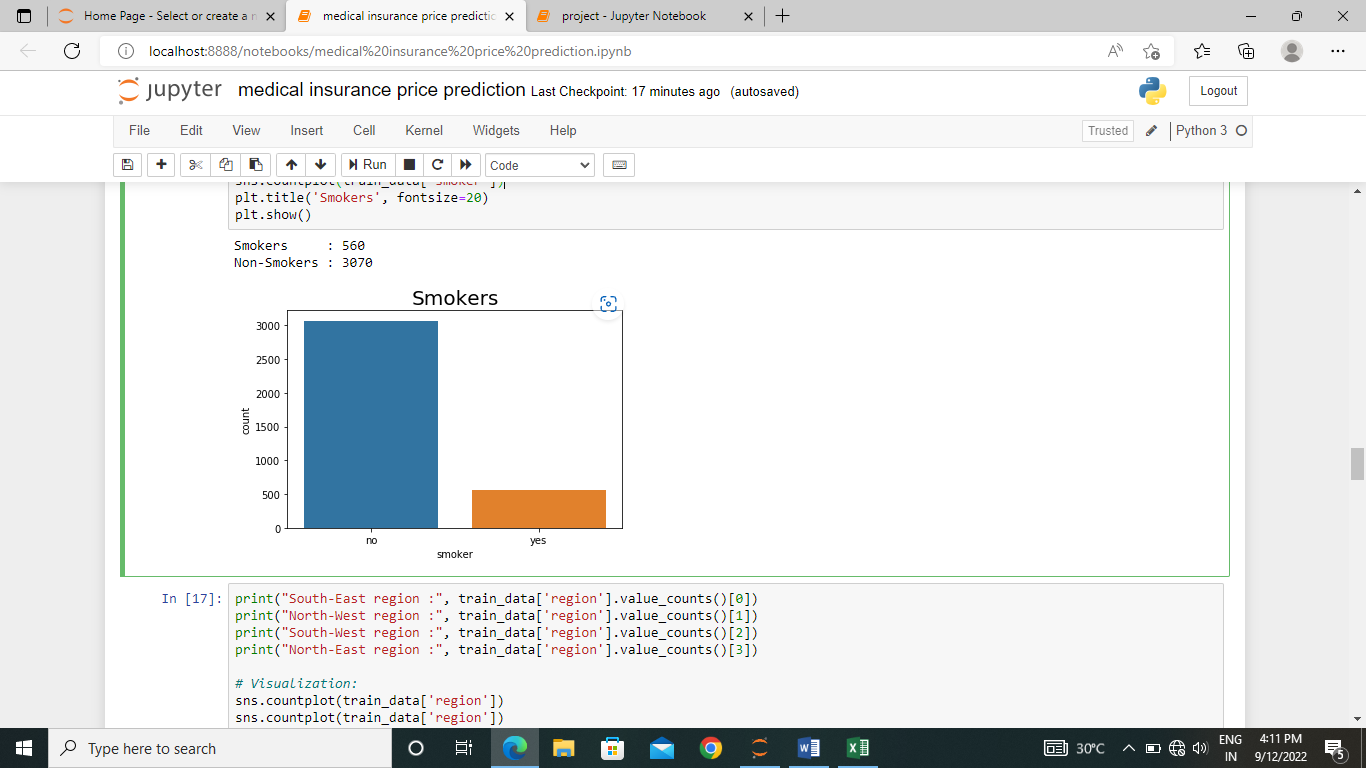
# Visualization:

sns.countplot(train\_data['smoker'])

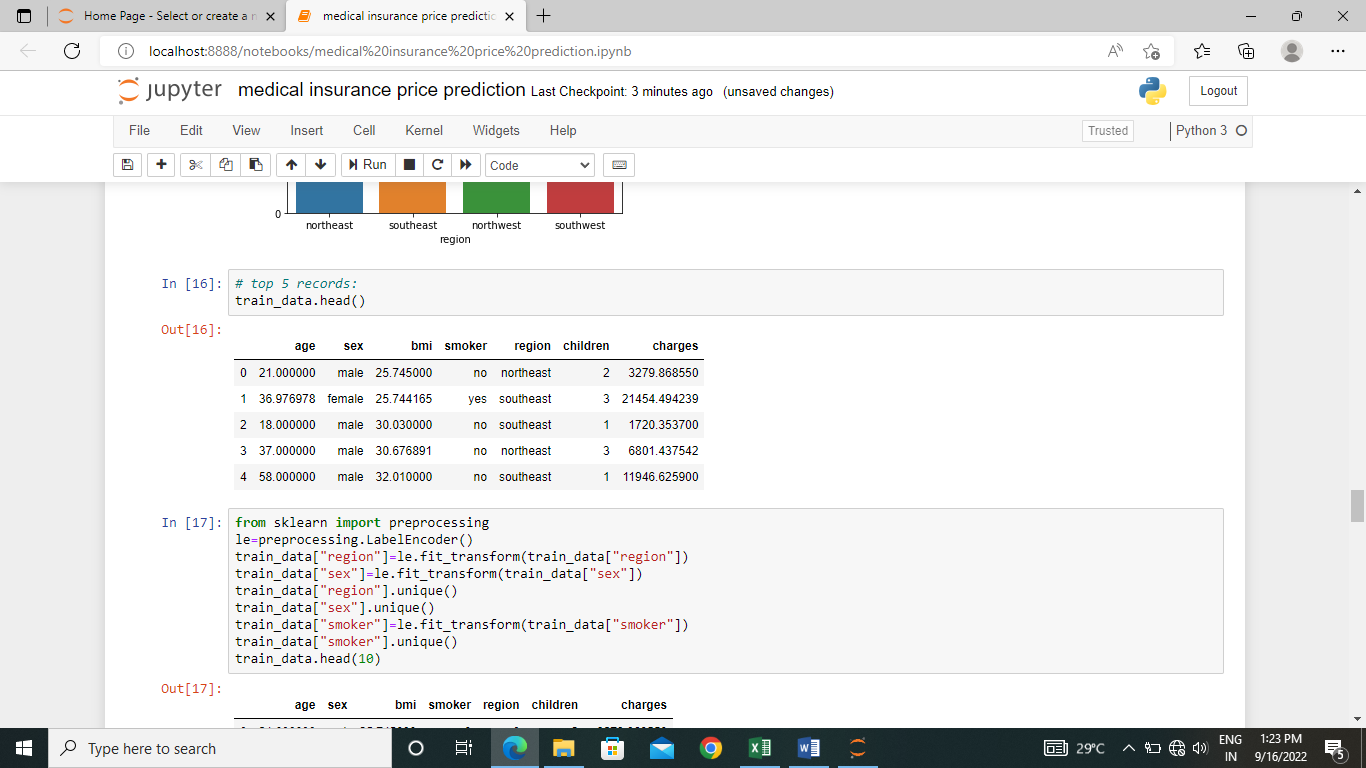
sns.countplot(train\_data['smoker'])

plt.title('Smokers', fontsize=20)

plt.show()



train\_data.head()



from sklearn import preprocessing

le=preprocessing.LabelEncoder()

train\_data["region"]=le.fit\_transform(train\_data["region"])

train\_data["sex"]=le.fit\_transform(train\_data["sex"])

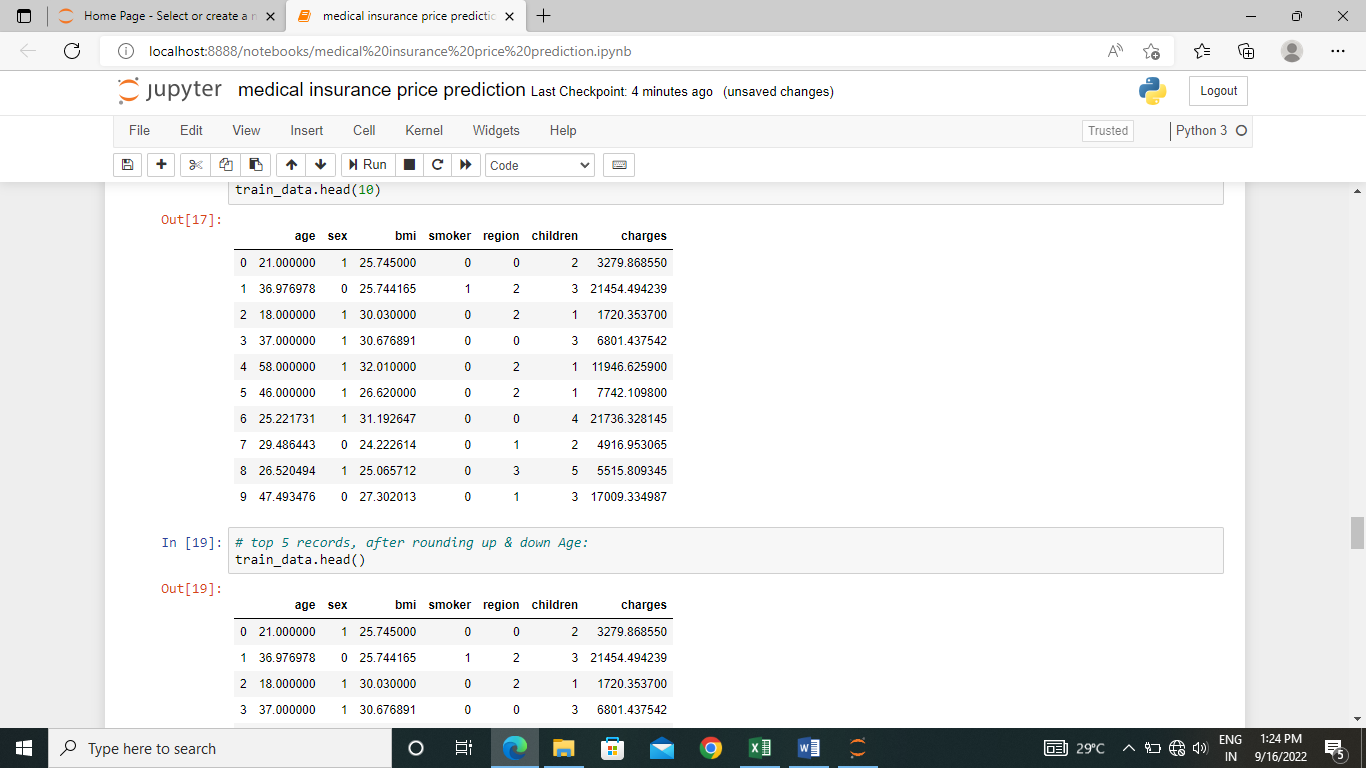
train\_data["region"].unique()

train\_data["sex"].unique()

train\_data["smoker"]=le.fit\_transform(train\_data["smoker"])

train\_data["smoker"].unique()

train\_data.head(10)



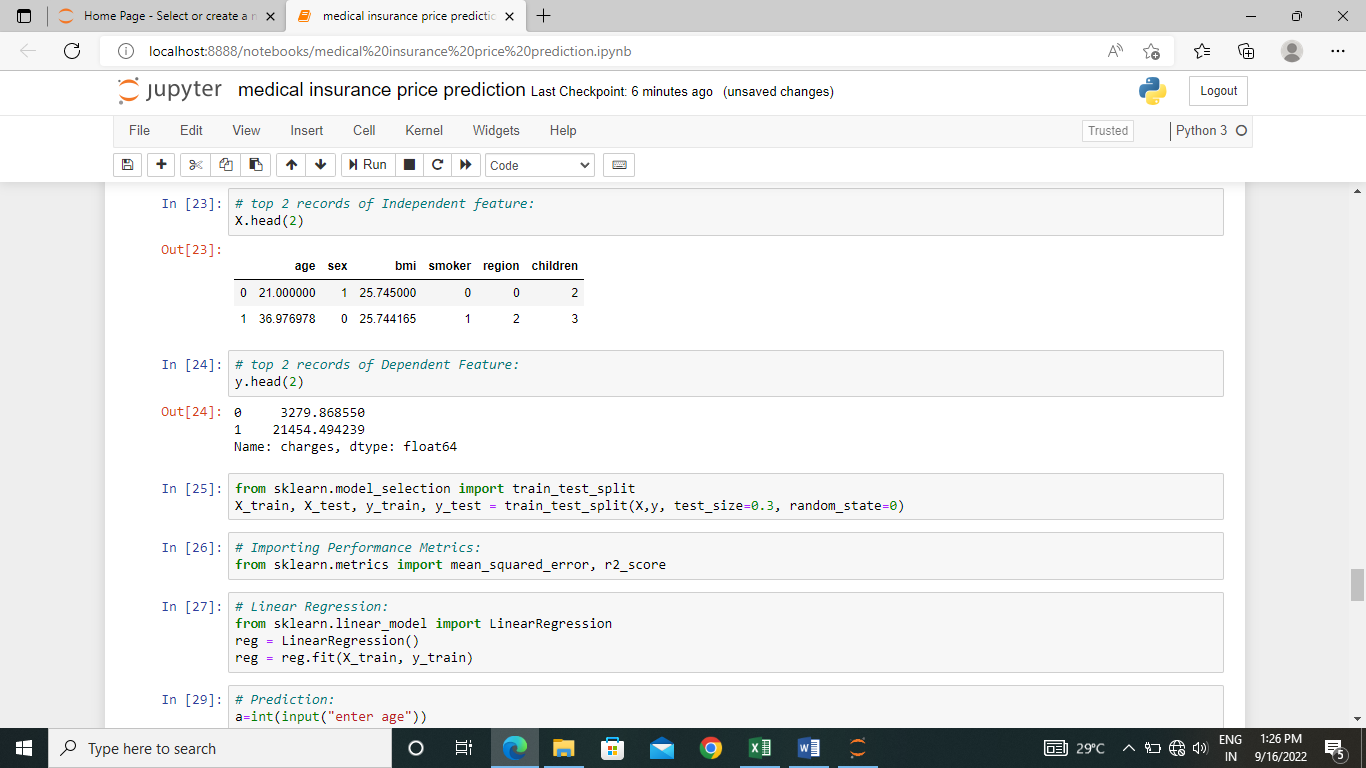
Step 5:

# Splitting Independent & Dependent variables:

X = train\_data.iloc[:, :-1]

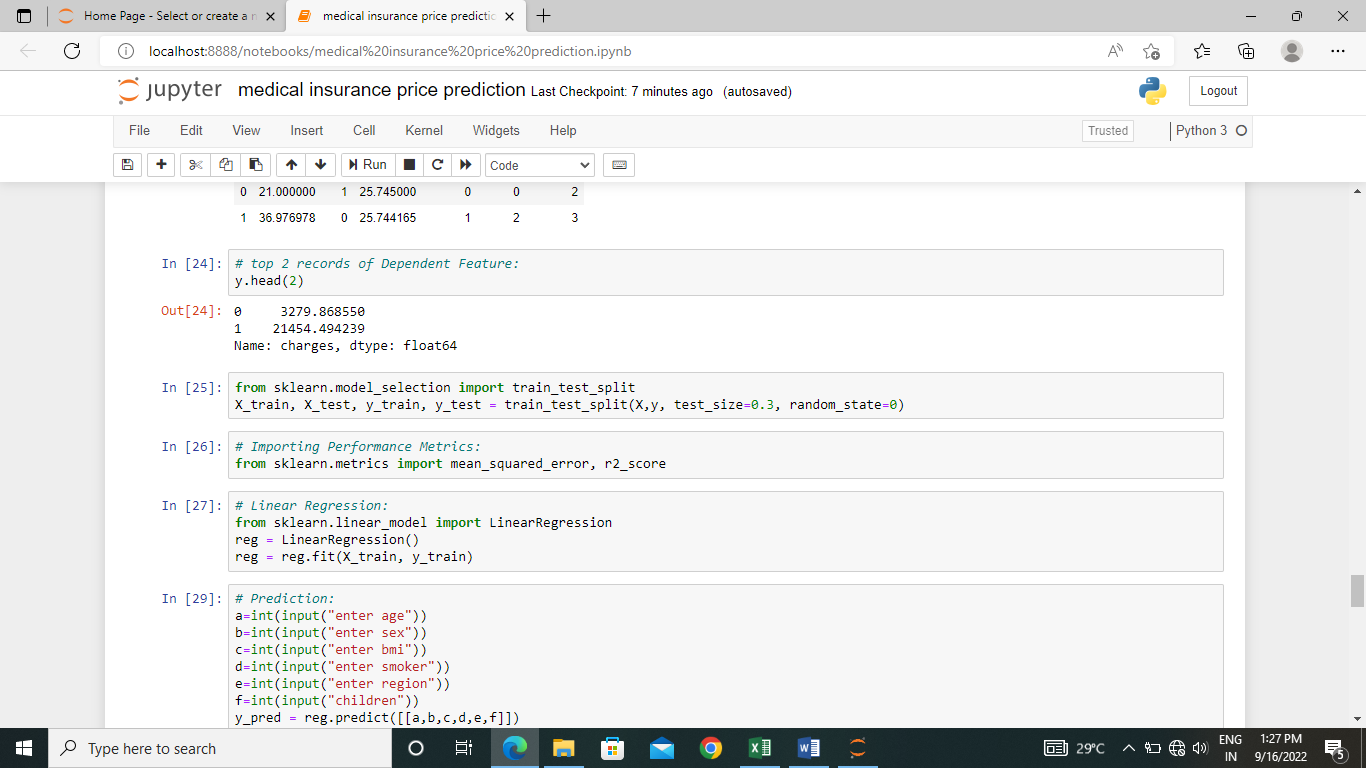
y = train\_data.iloc[:, -1]

X.head(2)



# top 2 records of Dependent Feature:

y.head(2)



Step 6:

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y, test\_size=0.3, random\_state=0)

# Importing Performance Metrics:

from sklearn.metrics import mean\_squared\_error, r2\_score

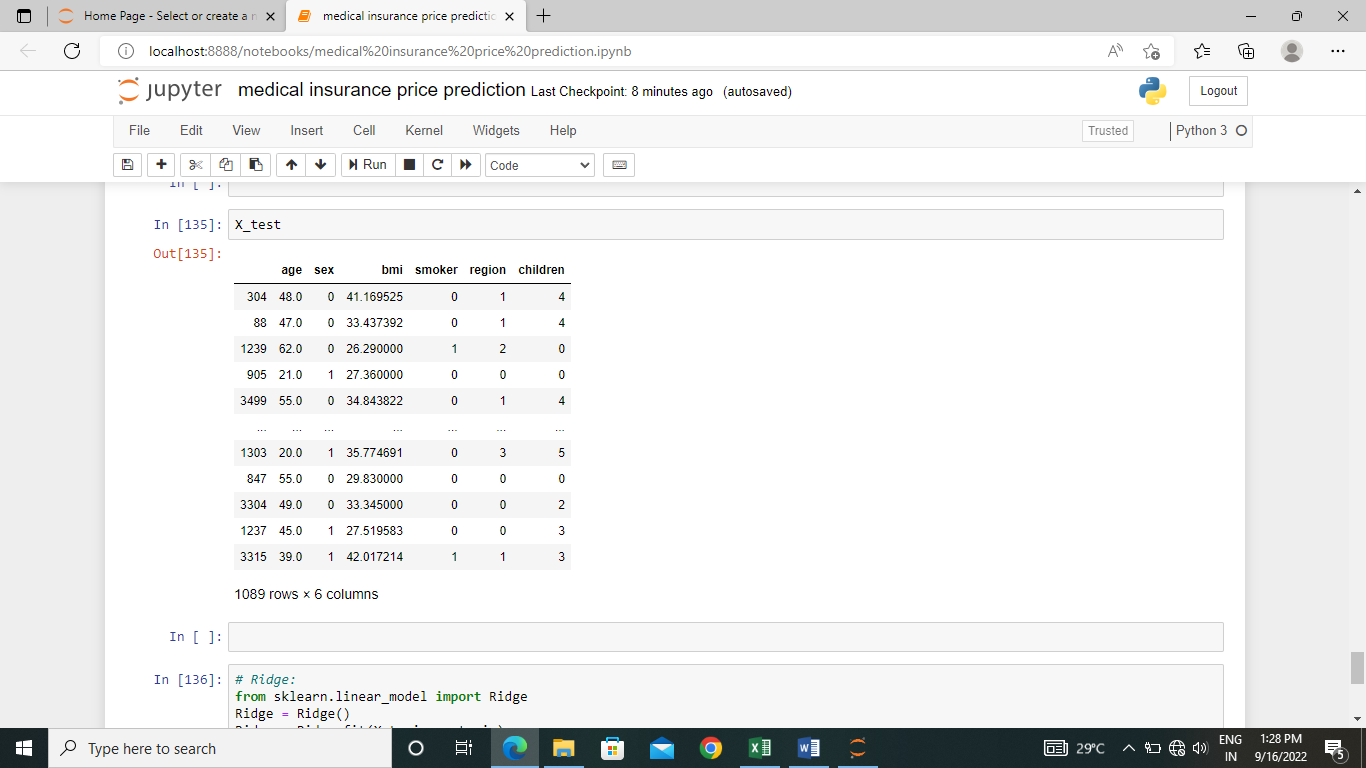
# Linear Regression:

from sklearn.linear\_model import LinearRegression

reg = LinearRegression()

reg = reg.fit(X\_train, y\_train)

X\_test



Step 7:

#predictions

a=int(input("enter age"))

b=int(input("enter sex"))

c=int(input("enter bmi"))

d=int(input("enter smoker"))

e=int(input("enter region"))

f=int(input("children"))

y\_pred = reg.predict([[a,b,c,d,e,f]])

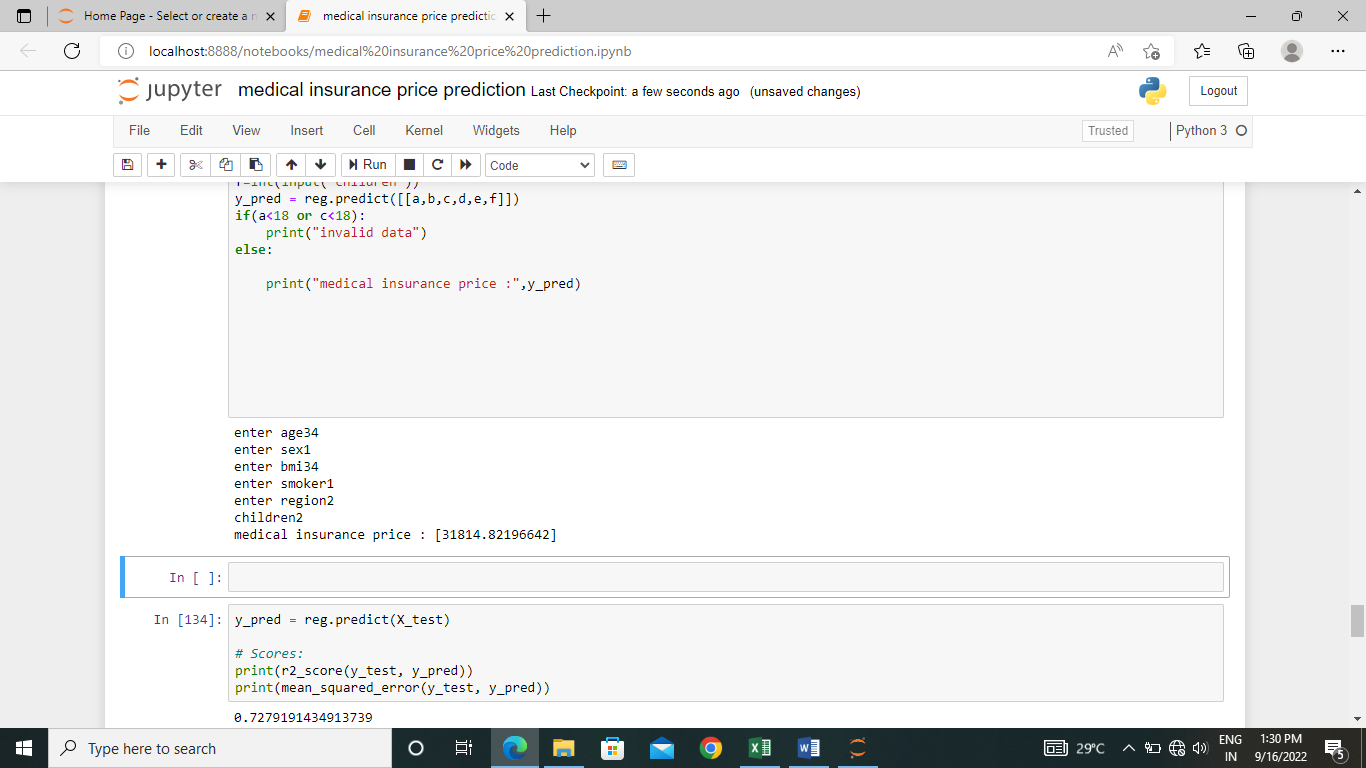
if(a<18 or c<12):

print("invalid data")

else:

print(y\_pred)

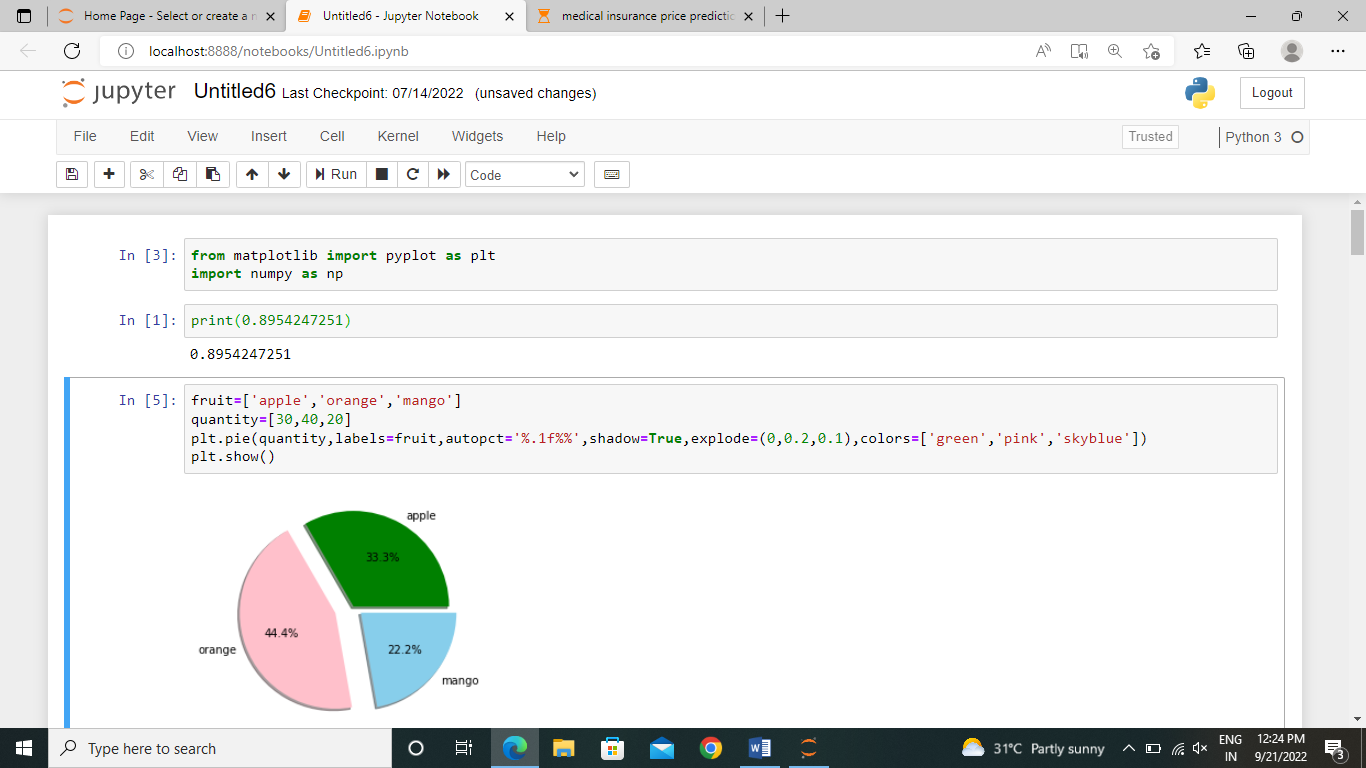
print("medical insurance price :")



y\_pred = reg.predict(X\_test)

# Scores:

print(r2\_score(y\_test, y\_pred))



Sources:

Downloaded the data from kaggale:

<https://www.kaggle.com./datasets/noordeen/insurance-premium-prediction>

git link:

<https://github.com/vinuthanakalepu/dsp-project.git>

SUMMARY:

It is *useful to make predictions of the insurance cost they will have to pay.*

*Medical emergencies are unpredictable , and with rising healthcare costs, quality management can get very expensive. Without medical insurance, there can be a rapid loss of savings.*

*This model is useful to estimate the cost of our medical insurance.*

*PRESENTING:*

*K.VINUTHANA*

*S180065@rguktsklm.ac.in*

*S180065*

*CSE-2A*