import pandas as pd

import matplotlib.pylab as plt

import numpy as np

df = pd.read\_csv('autodata.csv')

df.head(5)

df.tail(5)

df.info()

df.shape

df.columns

df.size

df.dtypes

df.describe()

df.isnull()

df.isnull().sum()

df.notnull()

df.notnull().sum()

#from the above we can tell that stroke has 4 missing data

# calculate the mean vaule for "stroke" column

avg\_stroke = df["stroke"].astype("float").mean(axis = 0)

print("Average of stroke:", avg\_stroke)

# replace NaN by mean value in "stroke" column

df["stroke"].replace(np.nan, avg\_stroke, inplace = True)

df.isnull().sum()

#data standardization

#transforminng dataa into a common format

#a particular type of data normalization where

#we substrat the mean and divide standard deviation

df['city-L/100km'] = 235/df["city-mpg"]

df.head()

df['highway-L/100km'] = 235/df["highway-mpg"]

df.head()

#data normalization

# process of transforming values of several variables

# into a similar range

df['length'] = df['length']/df['length'].max()

df['width'] = df['width']/df['width'].max()

df['height'] = df['height']/df['height'].max()

df[["length","width","height"]].head()

#categorical variables to Quantitive variables

!pip install scikit-learn

# Perform one-hot encoding

df\_encoded = pd.get\_dummies(df, columns=["make", "diesel", "body-style", "wheel-base"])

# Perform label encoding

from sklearn.preprocessing import LabelEncoder

label\_encoder = LabelEncoder()

df["make\_encoded"] = label\_encoder.fit\_transform(df["make"])

df["diesel\_encoded"] = label\_encoder.fit\_transform(df["diesel"])

df["body-style\_encoded"] = label\_encoder.fit\_transform(df["body-style"])

df["wheel-base\_encoded"] = label\_encoder.fit\_transform(df["wheel-base"])

# Print the modified DataFrame

print(df\_encoded)