import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

dataset = pd.read\_csv('diamonds.csv')

print(dataset.info())

df1 = dataset.loc[dataset['cut'].isnull()]

df1

print(dataset.describe())

dataset.isnull().sum()

#dataset = dataset.fillna(dataset['carat'].value\_counts().index[0])

print((dataset==0).sum())

#dataset.iloc[:,6:]=dataset.iloc[:,6:].replace(0, np.NaN)

#dataset = dataset.fillna(dataset['price'].value\_counts().index[0])

sns.boxplot(x=dataset['carat'])

sns.boxplot(x=dataset['x'])

sns.boxplot(x=dataset['y'])

sns.boxplot(x=dataset['z'])

sns.boxplot(x=dataset['depth'])

sns.boxplot(x=dataset['table'])

sns.boxplot(x=dataset['price'])

dataset['cut'].unique()

dataset['color'].unique()

dataset['clarity'].unique()

print(dataset['clarity'].value\_counts())

print(dataset['color'].value\_counts())

print(dataset['cut'].value\_counts())

dataset.drop(dataset.loc[dataset['clarity']=='XXX'].index, inplace=True)

dataset.drop(dataset.loc[dataset['clarity']==' '].index, inplace=True)

dataset.drop(dataset.loc[dataset['cut']==0.3].index, inplace=True)

dataset.drop(dataset.loc[dataset['cut']=='Wonderful'].index, inplace=True)

dataset.drop(dataset.loc[dataset['color']== 'AAA'].index, inplace=True)

dataset.drop(dataset.loc[dataset['color']==0.3].index, inplace=True)

dataset['cut'].fillna('null entry', inplace = True)

dataset['cut'] = pd.Categorical(dataset['cut'], ['null entry','Fair','Good','Very Good','Ideal','Premium'], ordered = True)

dataset['cut'] = dataset['cut'].cat.codes

dataset['color'] = pd.Categorical(dataset['color'], ['D','E','F','G','H','I','J'], ordered = True)

dataset['color'] = dataset['color'].cat.codes

dataset['clarity'] = pd.Categorical(dataset['clarity'], ['I1','SI2','SI1','VS2','VS1','VVS2','VVS1','IF'], ordered = True)

dataset['clarity'] = dataset['clarity'].cat.codes

#dataset = dataset.fillna(0)

dataset['price'].fillna(0, inplace = True)

dataset.head()

dataset.info()

df = dataset.loc[dataset.price==0]

df

x = dataset.drop(dataset.loc[dataset.price==0].index,axis=0)

x

# regression

allCols = dataset.columns.tolist()

allCols.remove('depth')

allCols.remove('x')

allCols.remove('y')

allCols.remove('z')

allCols.remove('table')

allCols.remove('price')

allCols

print(allCols)

x1=x[allCols]

y1=x.price

y\_pred=df.price

ynew=df[allCols]

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

#x\_train, x\_test, y\_train, y\_test = train\_test\_split(x,y, test\_size=0.25)

from sklearn.linear\_model import LinearRegression

model = LinearRegression()

model.fit(x1,y1)

#predict

y\_pred = model.predict(ynew)

print(y\_pred)

# classification

allCols2 = dataset.columns.tolist()

allCols2.remove('depth')

allCols2.remove('x')

allCols2.remove('y')

allCols2.remove('z')

allCols2.remove('table')

allCols2.remove('cut')

print(allCols2)

df1 = dataset.loc[dataset['cut']==0]

df1

p = dataset.drop(dataset.loc[dataset.cut==0].index,axis=0)

p

xp=p[allCols2]

yp=p.cut

y\_pred=df1.cut

ynew=df1[allCols2]

#x2=dataset[allCols2]

#y2=dataset.cut

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

#x\_train2, x\_test2, y\_train2, y\_test2 = train\_test\_split(x2,y2, test\_size=0.25)

from sklearn.neighbors import KNeighborsClassifier

model2 = KNeighborsClassifier()

model2.fit(x1,y1)

#predict

y\_pred2 = model2.predict(ynew)

print(y\_pred2)

pd.crosstab(dataset.carat,dataset.cut, values=dataset.price,  aggfunc='mean').fillna('-')

# In[25]:

pd.crosstab(dataset.carat,dataset.clarity, values=dataset.price,  aggfunc='mean').fillna('-')

# In[26]:

pd.crosstab(dataset.carat,dataset.color, values=dataset.price,  aggfunc='mean').fillna('-')