import pandas as pd

import numpy as np

from sklearn.datasets import load\_boston

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

from sklearn.linear\_model import LinearRegression, Ridge

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

import matplotlib.pyplot as plt

import seaborn as sns

data=load\_boston()

df = pd.DataFrame(data.data, columns=data.feature\_names)

df['MEDV'] = data.target

#1

print("Dataset Features\n", df.head())

print()

print("Checking for missing values:")

print()

print(df.isnull().sum())

print("No of features in given dataset is")

print(df.shape[1])

df.fillna(df.mean(), inplace=True)

print("\nMissing values after replacement:")

print(df.isnull().sum())

#2

print("\nStatistical details of the dataset:")

print(df.describe())

#3

plt.figure(figsize=(10, 6))

plt.scatter(df['RM'], df['MEDV'], color='red')

plt.title("Relationship between RM and MEDV ")

plt.xlabel("RM (Average Number of Rooms)")

plt.ylabel("MEDV (Median Value of Homes)")

plt.show()

#4

x = df.drop('MEDV',axis=1)

y = df['MEDV']

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.3,random\_state=50)

lr = LinearRegression()

lr.fit(x\_train, y\_train)

y\_predlinear = lr.predict(x\_test)

rg=Ridge()

rg.fit(x\_train,y\_train)

y\_predridge=rg.predict(x\_test)

#5

plt.subplot(1, 2, 1)

plt.scatter(y\_test, y\_predlinear, color='green')

plt.plot([y\_test.min(), y\_test.max()], [y\_test.min(), y\_test.max()])

plt.title("Linear Regression")

plt.xlabel("Actual Values")

plt.ylabel("Predicted Values")

plt.subplot(1, 2, 2)

plt.scatter(y\_test, y\_predridge, color='purple')

plt.plot([y\_test.min(), y\_test.max()], [y\_test.min(), y\_test.max()])

plt.title("Ridge Regression")

plt.xlabel("Actual Values")

plt.ylabel("Predicted Values")

plt.show()

mlr = mean\_absolute\_error(y\_test, y\_predlinear)

rlr = r2\_score(y\_test, y\_predlinear)

mridge = mean\_absolute\_error(y\_test, y\_predridge)

rridge = r2\_score(y\_test, y\_predridge)

print("Linear Regression:")

print("mean absolute error")

print(mlr)

print("r-squared")

print(rlr)

print("Ridge Regression:")

print("mean absolute error")

print(mridge)

print("r-squared")

print(rridge)

if mlr>mridge:

print ("mean in linear")

else:

print("mean in ridge")

if rlr>rridge:

print ("r in linear")

else:

print("r in ridge")

Dataset Features

CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX \

0 0.00632 18.0 2.31 0.0 0.538 6.575 65.2 4.0900 1.0 296.0

1 0.02731 0.0 7.07 0.0 0.469 6.421 78.9 4.9671 2.0 242.0

2 0.02729 0.0 7.07 0.0 0.469 7.185 61.1 4.9671 2.0 242.0

3 0.03237 0.0 2.18 0.0 0.458 6.998 45.8 6.0622 3.0 222.0

4 0.06905 0.0 2.18 0.0 0.458 7.147 54.2 6.0622 3.0 222.0

PTRATIO B LSTAT MEDV

0 15.3 396.90 4.98 24.0

1 17.8 396.90 9.14 21.6

2 17.8 392.83 4.03 34.7

3 18.7 394.63 2.94 33.4

4 18.7 396.90 5.33 36.2

Checking for missing values:

CRIM 0

ZN 0

INDUS 0

CHAS 0

NOX 0

RM 0

AGE 0

DIS 0

RAD 0

TAX 0

PTRATIO 0

B 0

LSTAT 0

MEDV 0

dtype: int64

No of features in given dataset is

14

Missing values after replacement:

CRIM 0

ZN 0

INDUS 0

CHAS 0

NOX 0

RM 0

AGE 0

DIS 0

RAD 0

TAX 0

PTRATIO 0

B 0

LSTAT 0

MEDV 0

dtype: int64

Statistical details of the dataset:

CRIM ZN INDUS CHAS NOX RM \

count 506.000000 506.000000 506.000000 506.000000 506.000000 506.000000

mean 3.613524 11.363636 11.136779 0.069170 0.554695 6.284634

std 8.601545 23.322453 6.860353 0.253994 0.115878 0.702617

min 0.006320 0.000000 0.460000 0.000000 0.385000 3.561000

25% 0.082045 0.000000 5.190000 0.000000 0.449000 5.885500

50% 0.256510 0.000000 9.690000 0.000000 0.538000 6.208500

75% 3.677083 12.500000 18.100000 0.000000 0.624000 6.623500

max 88.976200 100.000000 27.740000 1.000000 0.871000 8.780000

AGE DIS RAD TAX PTRATIO B \

count 506.000000 506.000000 506.000000 506.000000 506.000000 506.000000

mean 68.574901 3.795043 9.549407 408.237154 18.455534 356.674032

std 28.148861 2.105710 8.707259 168.537116 2.164946 91.294864

min 2.900000 1.129600 1.000000 187.000000 12.600000 0.320000

25% 45.025000 2.100175 4.000000 279.000000 17.400000 375.377500

50% 77.500000 3.207450 5.000000 330.000000 19.050000 391.440000

75% 94.075000 5.188425 24.000000 666.000000 20.200000 396.225000

max 100.000000 12.126500 24.000000 711.000000 22.000000 396.900000

LSTAT MEDV

count 506.000000 506.000000

mean 12.653063 22.532806

std 7.141062 9.197104

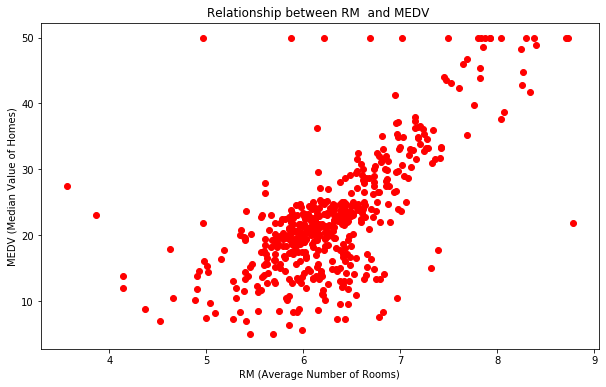
min 1.730000 5.000000

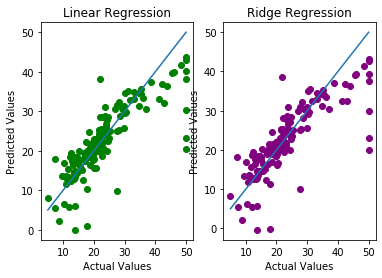
25% 6.950000 17.025000

50% 11.360000 21.200000

75% 16.955000 25.000000

max 37.970000 50.000000





Linear Regression:

mean absolute error

3.678977534499423

r-squared

0.6685538790447987

Ridge Regression:

mean absolute error

3.6691280566480144

r-squared

0.6641307057316166

mean in linear

r in linear