Julius Niyonzima

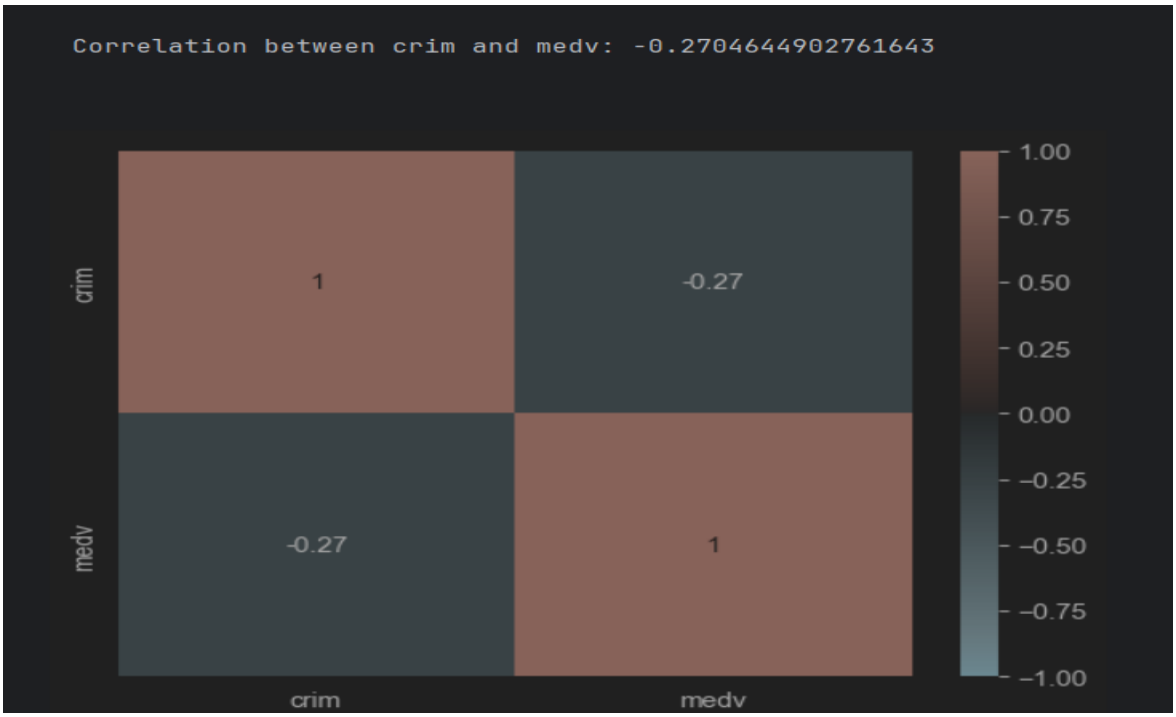
Professor Adu Baffour

Class: CS456R

HW#2 Classification

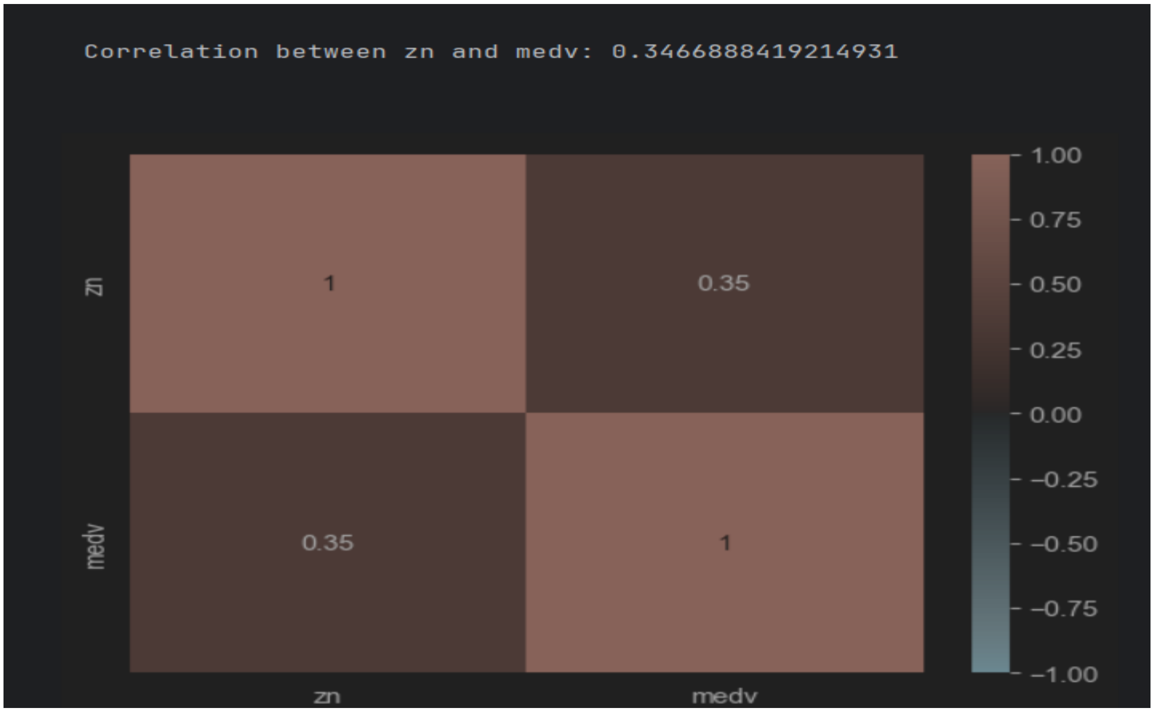
#2. Explanation regarding the graphs generated for the Boston dataset.

Median value of Owner-occupied homes in $1000s(medv) vs Per Capita Crime Rate by Town(crim)



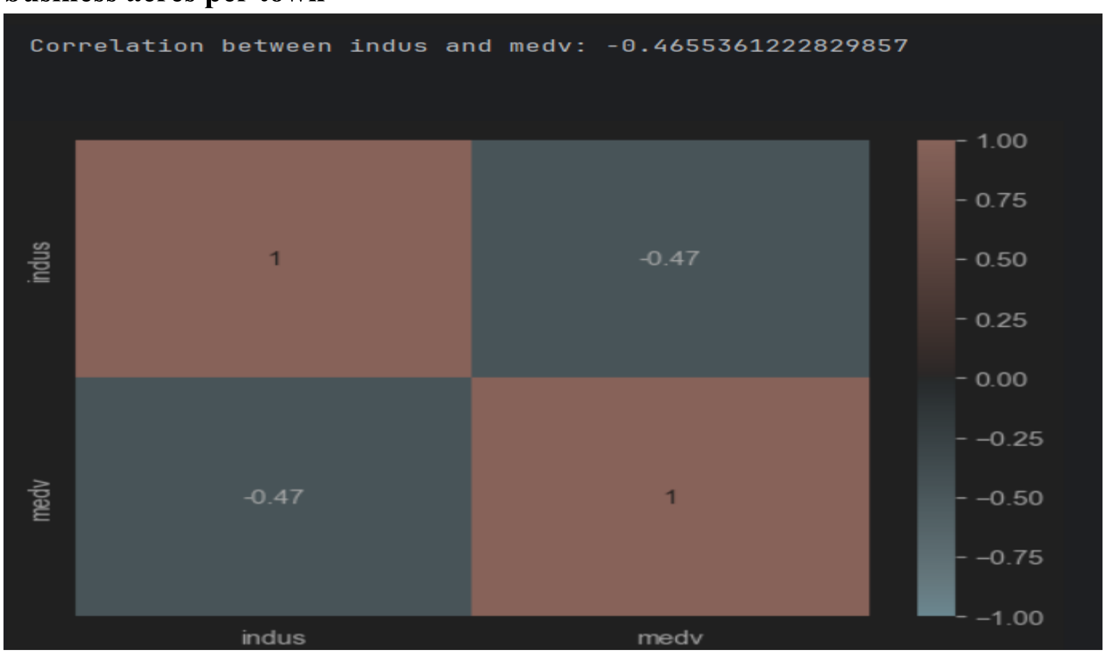
The correlation coefficient between the crime rate and median house value is -0.27, which shows that there's tendance for crime rate to increases as median house value decreases. However, the correlation is weak, so we cannot ascertain a strong relationship between the crime rate and median house value.

Median value of Owner-occupied homes in $1000s(MEDV) vs Proportion of residential Land Zoned for lots over 25,000 square ft (ZN)



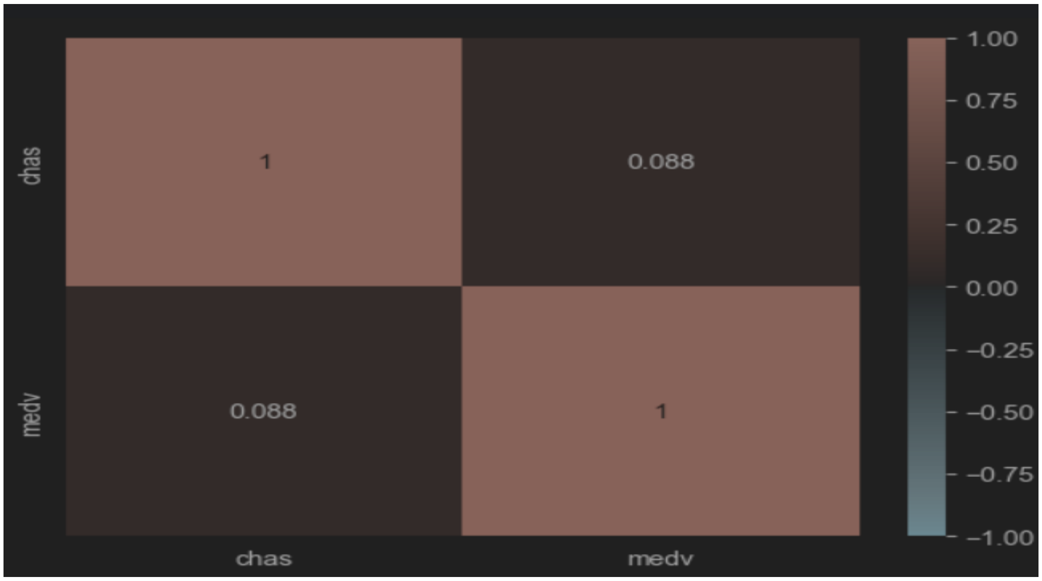
The correlation between ZN and MEDV shows a moderate positive relationship, indicating that as the median house value increases, the zoned area (ZN) also increases, and vice versa.

Median value of Owner-occupied homes in $1000s(MEDV) Vs Proportion of non_retail business acres per town



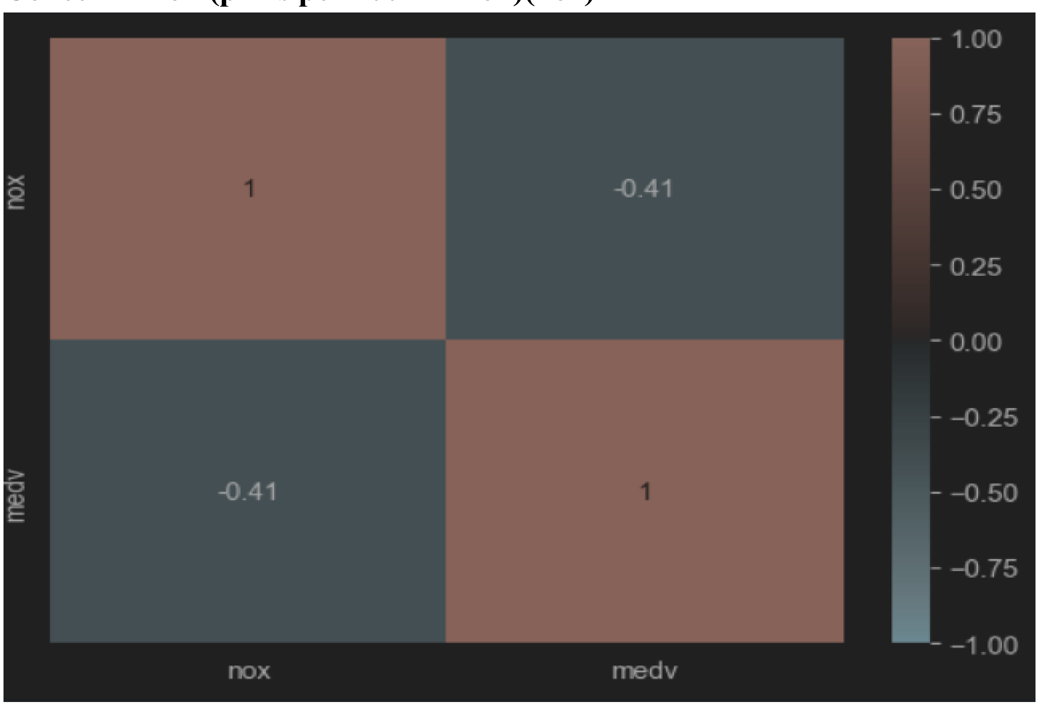
The correlation between INDUS and MEDV shows a moderate negative correlation. This means that as the proportion of non-retail business acres increases per town, the housing value decreases, although it is not a perfect prediction.

Median value of Owner-occupied homes in $1000s(MEDV) vs (CHAS) Charles River dummy variable(1 = tract bounds river, 0 otherwise)



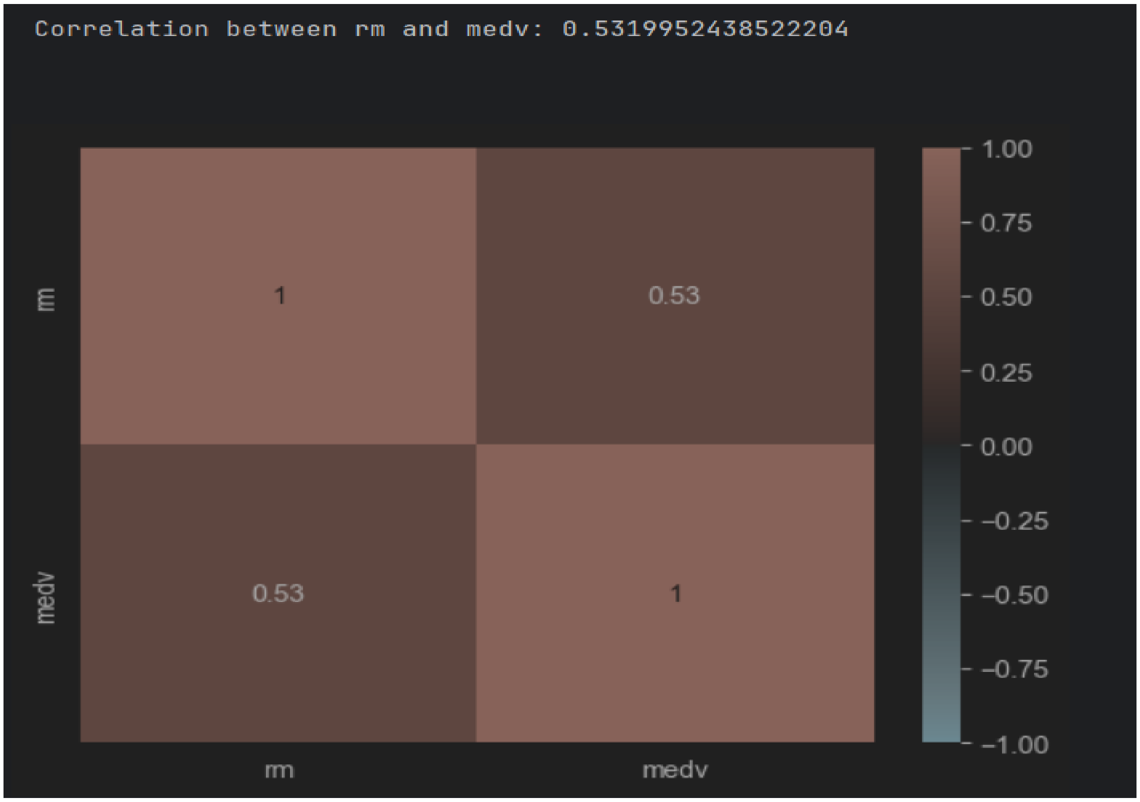
The correlation between MEDV and CHAS is 0.088, suggesting that houses located near the Charles River slightly increase in value. However, the correlation is negligible

Median value of Owner-occupied homes in $1000s(MEDV) vs Nitrogen Oxides Concentration (parts per 100 milllion)(nox)



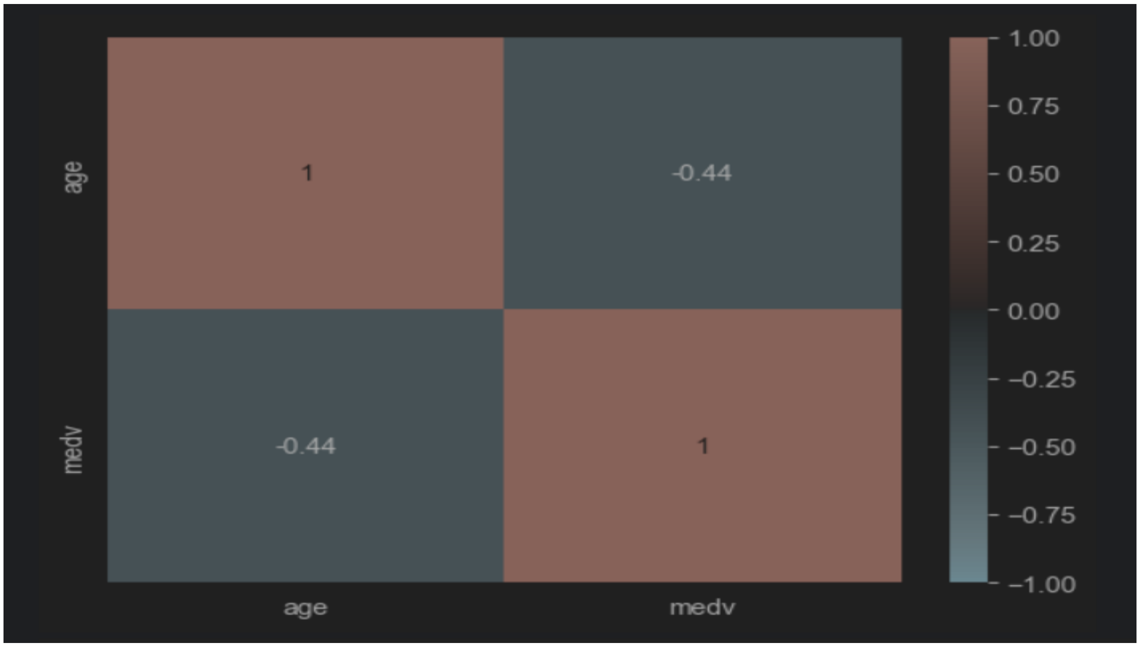
The correlation coefficient between the median house value and the concentration of nitrogen oxide is -0.41, indicating a moderate negative relationship between them. As the concentration of nitrogen oxide increases, the median house value tends to decrease**.**

Median value Owner-occupied homes in $1000s(MEDV) vs Average Number of rooms per dwelling(RM).



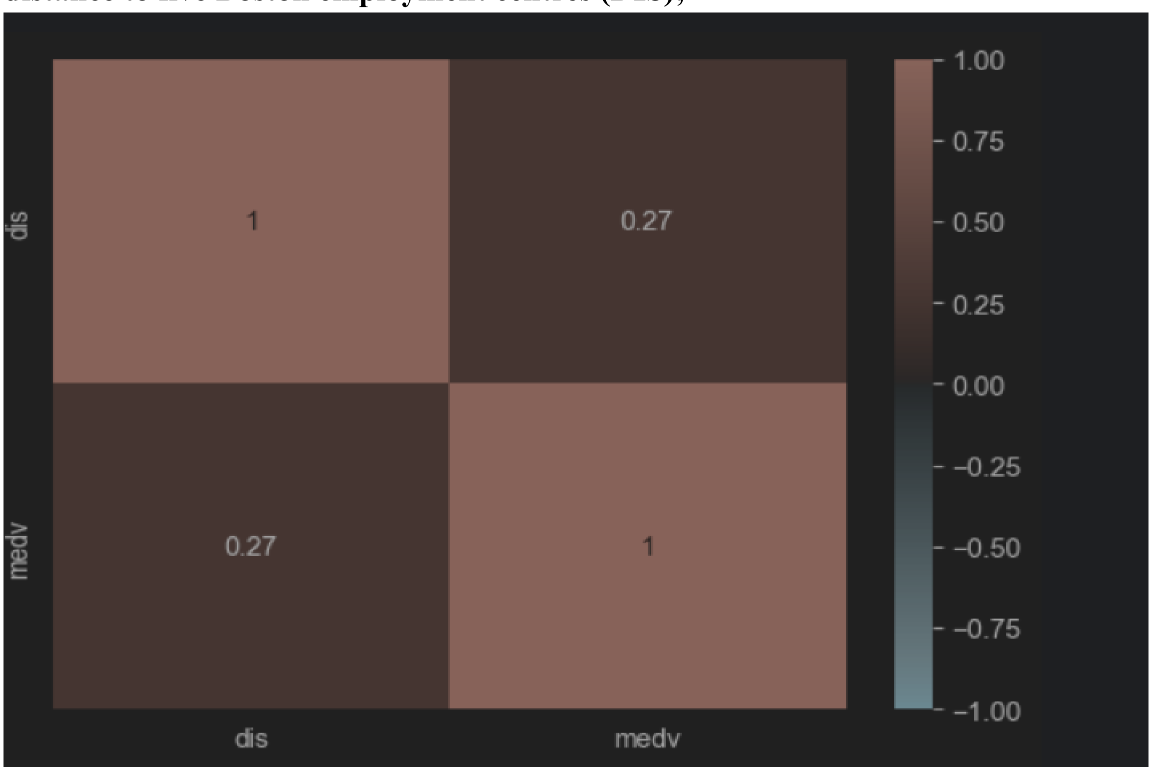
The coefficient correlation between MEDV and RM is o.53. This indicates a positive relationship between house value and number of rooms. As the number of rooms increases, the house value tends to increase**.**

Median value of Owner-occupied homes in $1000s(MEDV) vs proportion of owner-occupied units built prior to 1940(age).



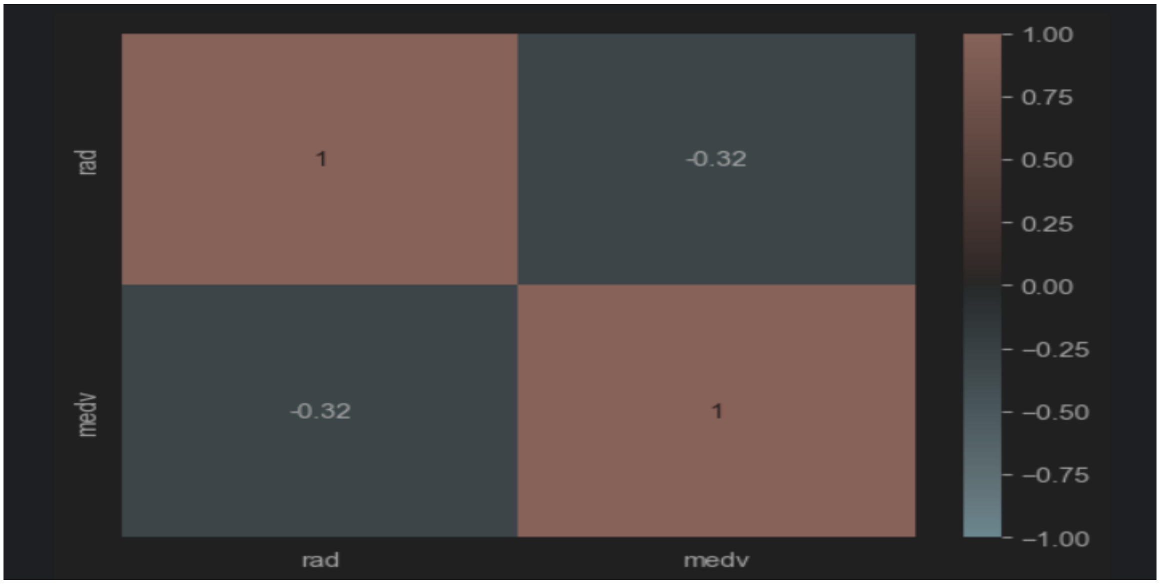
The correlation coefficient between the age of the house and the median value is -0.44. This shows a moderate negative correlation between house value and age. As the house increases in age, its value tends to decrease.

Median value of Owner-occupied homes in $1000s(MEDV) vs (Weighted mean of distance to five Boston employment centres (DIS);



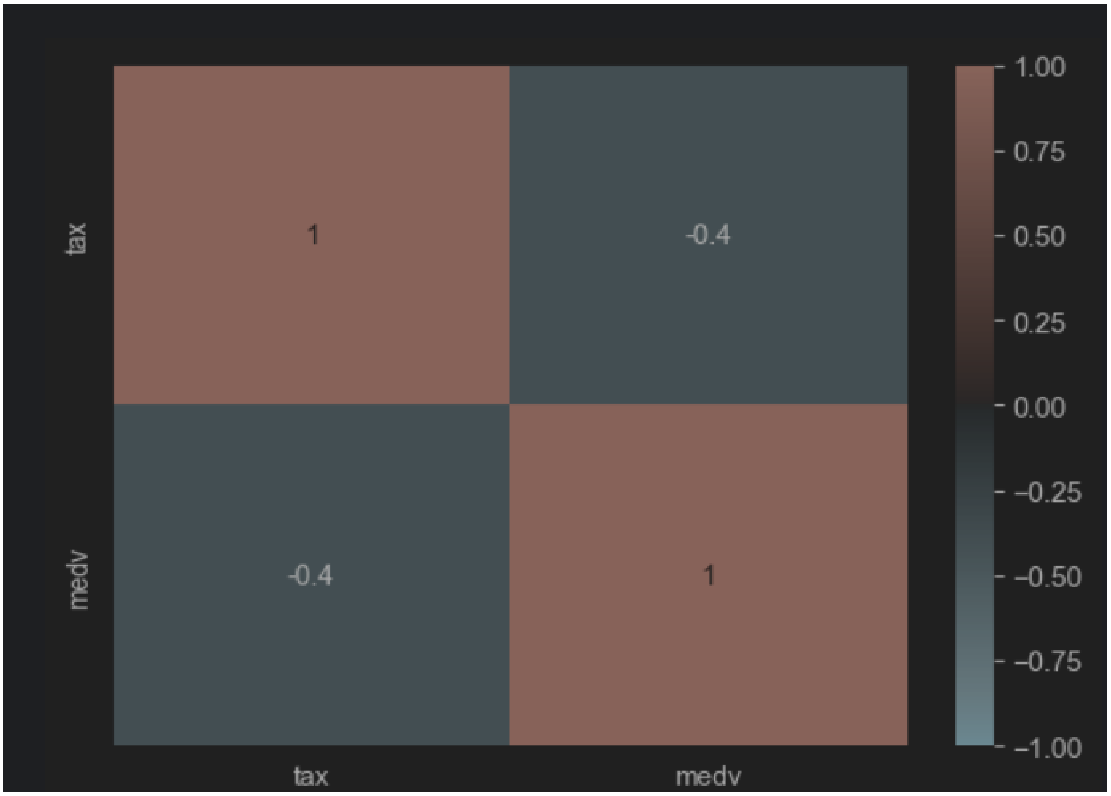
The correlation coefficient between house value and distance from the Boston employment center is 0.27, which shows a weak positive relationship between the two. As the distance from the Boston employment center increases, the value of the house tends to increase slightly. However, the coefficient does not strongly support accurate predictions

Median value of Owner-occupied homes in $1000s(MEDV) vs Index of accessibility to radial highways(RAD)



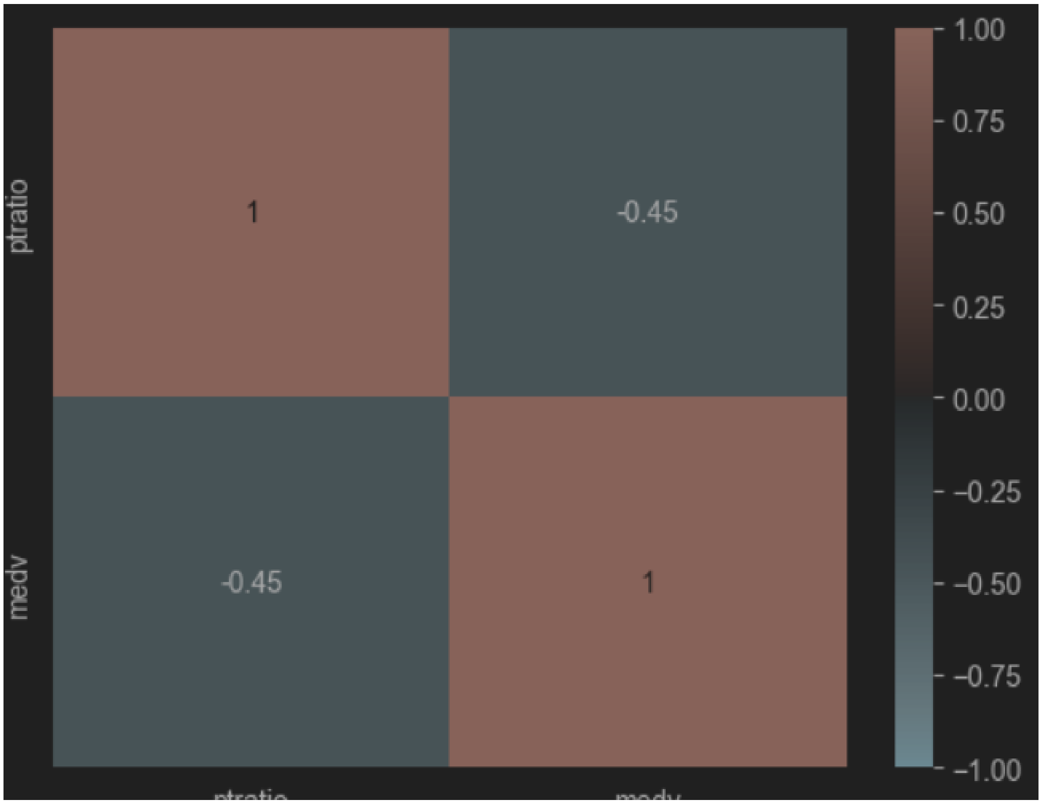
The correlation coefficient between MEDV and RAD is -0.32. This shows a moderate relationship between them. As the accessibility to radial high increases, the house price tends to decrease**.**

Median value of Owner-occupied homes in $1000s(MEDV) vs full-value property-tax rate per $10,000(TAX).



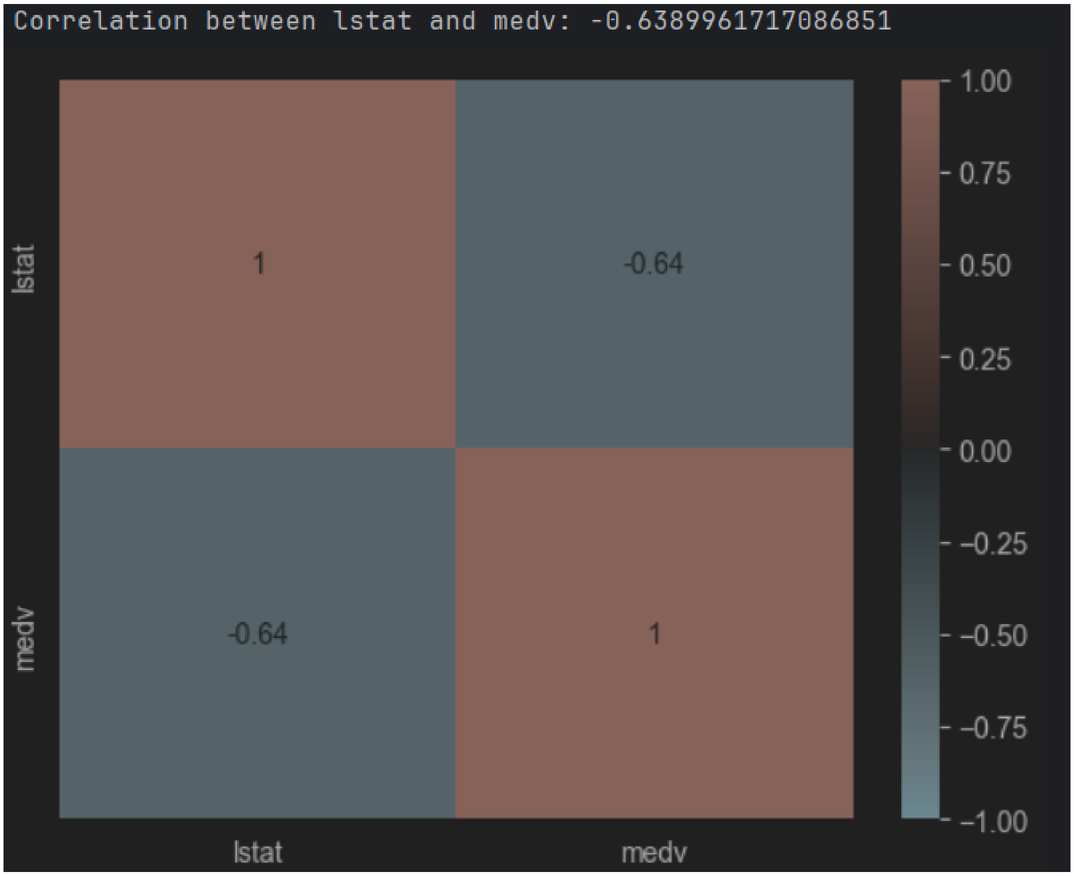
The correlation between the full-value property tax rate and the median house value is -0.4. This shows that if the tax increases, the house value tends to decrease.

Median value of Owner-occupied homes in $1000s(MEDV) vs Pupil-teacher ratio by town(PTRATIO)



The correlation coefficient between ptratio and the median value of homes is -0.45. This shows that as the number of pupils increases relative to the number of teachers in the area, housing prices tend to decrease due to lower-quality education.

Median value of Owner-occupied homes in $1000s(MEDV) vs lower status of the population (percent)(LSTAT).



The correlation coefficient between the lower status of the population and the median value of homes is -0.64. This indicates a strong negative correlation. It predicts that as the proportion of lower-status individuals increases, the median house value decreases**.**

#4. Logistic Regression summary result.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                   medv   No. Observations:                  404
Model:                          Logit   Df Residuals:                      391
Method:                           MLE   Df Model:                           12
Date:                Sun, 09 Mar 2025   Pseudo R-squ.:                  0.5534
Time:                        16:22:07   Log-Likelihood:                -123.06
converged:                       True   LL-Null:                       -275.56
Covariance Type:            nonrobust   LLR p-value:                 4.182e-58
==============================================================================
```

The likelihood ratio p-value is 4.12 * 10^-58. This shows that the independent variables used to predict the dependent variable (MEDV) are significantly associated with predicting the outcome of housing values according to the Boston dataset.

	coef	std err	z	P>|z|	[0.025	0.975]
const	10.4497	4.086	2.558	0.011	2.442	18.457
crim	-0.0372	0.046	-0.803	0.422	-0.128	0.054
zn	0.0064	0.012	0.549	0.583	-0.016	0.029
indus	-0.0332	0.050	-0.660	0.509	-0.132	0.065
chas	0.2231	0.686	0.325	0.745	-1.122	1.568
nox	-2.8181	2.987	-0.944	0.345	-8.672	3.036
rm	1.3135	0.444	2.960	0.003	0.444	2.183
age	-0.0243	0.010	-2.320	0.020	-0.045	-0.004
dis	-0.5435	0.164	-3.321	0.001	-0.864	-0.223
rad	0.2226	0.066	3.370	0.001	0.093	0.352
tax	-0.0076	0.003	-2.312	0.021	-0.014	-0.001
ptratio	-0.4917	0.115	-4.290	0.000	-0.716	-0.267
lstat	-0.3056	0.057	-5.353	0.000	-0.418	-0.194

The p-value of the number of rooms per dwelling, the proportion of units built prior to 1940, distance from five Boston employment centers, accessibility to radial highways, full-value property tax, pupil-teacher ratio, and the lower status of the population (percent) is less than 0.05. This shows that all these independent variables are statistically significant in predicting the median house value.

#5 COMMENTS ON CONFUSION MATRIX

```
6   print()
    ✓ [7] 15ms

    Accuracy: 92.16
    ConfusionMatrix is [[64  1]
     [ 7 30]]
```

This shows that the model had an accuracy of 92.2% in making correct predictions regarding how the independent variables predicted the median house value.

It was

The confusion matrix shows that the model was able to correctly predict 64 negative and 30 positive outcomes. The model made 1 mistake by predicting a positive when it should have been negative, and 7 errors by predicting negative when it should have been positive.

NO 9 a. LDA results

```
Accuracy:  92.15686274509804
Confusion Matrix:
 [[64  1]
 [ 7 30]]
```

KNN RESULTS

```
The accuracy score is 86.27 %
```

```
The confusion matrix:
 [[60  5]
 [ 9 28]]
```

NAÏVE BAYES RESULTS

```
The accuracy score is 80.39 %
```

```
the confusion matrix is [[53 12]
 [ 8 29]]
```

I will select Linear Discriminant Analysis and Logistic Regression because they had an accuracy score of 92% in making correct predictions on how independent variables like the age of a house, distance from Boston employment centers, accessibility to radial highways, full-value property tax, pupil-teacher ratio, and the lower status of the population can affect the median value of homes in the city.

9b.

The Naïve bayes had an accuracy score of 80% which is poor. The model predicted 53 negative outcomes correctly (True Negatives) and 29 positive outcomes, but made 12 false positive predictions and 8 false negative predictions. The reason could have been due to imbalanced datasets, errors in cleaning the data, and the fact that most of the features used did not follow normal distribution. Additionally, the features used in the Boston dataset, such as crime rate and lower status, are highly correlated, which could have hindered the Naive Bayes classifier from making correct predictions.

No 10. Conclusion.

The Logistic Regression(LR) and Linear Discriminant Analysis (LDA)models achieved a 92% accuracy score. They both correctly predicted 64 negative outcomes and 30 positive outcomes, with 1 false positive and 7 false negatives. On the other hand, (K-nearest neighbor) KNN had an 86% accuracy score. It correctly predicted 60 negative and 28 positive outcomes, with 5 false positives and 9 false negatives. Naïve Bayes classifier had an 80% accuracy score. It correctly predicted 53 negatives and 29 positives, with 12 false positives and 8 false negatives. The LR, LDA, and KNN models agree on correctly predicting 60+ negative outcomes. LDA and LR agree in making predictions because they produced similar results. However, the Naïve Bayes model completely disagrees with the results, as it produced different outcomes and performed poorly in terms of real estate. The models show that factors like crime, house age, poor quality education, and poverty affect housing prices. According to Martin Maximino, criminal behavior imposes both direct costs on victims and indirect costs on society, including economic effects that can spread beyond the immediate crime scene. While residents may avoid dangerous neighborhoods or move elsewhere, research suggests that criminal activity can shift between areas, much like infectious diseases. Additionally, according to Habitat for Humanity, the location and condition of a child's house play a significant role in their physical, cognitive, and emotional development and well-being, which impacts their education through improved attendance, better cognitive and behavioral health, and improved academic achievement. Making direct connections between housing and its impact on children's education is challenging.

Work cited:

The impact of crime on property values: Research roundup - The Journalist's Resource

21-81776_RD_EvidenceBrief-6-Education_FASH-lores_1.pdf

Evaluation Metrics in Machine Learning - GeeksforGeeks

Naive Bayes Classifiers - GeeksforGeeks

Understanding the Confusion Matrix in Machine Learning - GeeksforGeeks