Statistical Analysis on the Boston Housing Dataset using R

First draft prepared on: October 14th, 2023
Updated on: October 18th, 2023
Prepared by: Raza Mehar
Version: 2.0

Components of the assignment

1. Installing the MASS package and loading Boston dataset to check the number of rows and columns and what they represent.
2. Scale of measurement of each variable, summary, frequency distribution, visualization and findings.
3. Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each variable.
4. How many of the census tracts in this data set bound the Charles river?
5. What is the median pupil-teacher ratio among the towns in this data set?
6. Which census tract of Boston has lowest median value of owner- occupied homes? What are the values of the other variables for that census tract, and how do those values compare to the overall ranges for those variables? Comment on your findings.
7. Pairwise scatterplots of the variables in this data set and findings
8. Are any of the predictors to explain “per capita crime rate”? If so, explain the relationship.
9. Are any of the predictors to explain the price of the houses as measured by the “Median value of owner-occupied homes in $1000”? If so, explain the relationship.
10. Which variable can play the role of the target variable and which predictors can explain it? Provide a cross-classification of the units considering the target variable and one predictor. Is there independence in distribution? Is there independence in mean? Provide measures of suitable statistical indexes to explain the eventual association, the eventual correlation, and the eventual linear correlation. Describe your findings.

Q1. Loading and Understanding the Dataset

Q1. Loading the package and the dataset. How many rows are in this data set? How many columns? What do the rows and columns represent?

Since I could not find the Boston dataset in the ISLR2 packages, I installed the MASS package from the R Studio console and loaded the library and dataset using the following commands:

```
library(MASS)
data(Boston)
```

Rows	506
Columns	14

Columns:

- The columns represent the variables or characteristics, or features being observed or measured.

Rows:

- Whereas, the rows represent the individual data points.

Note:

- Since there are 14 distinct variables, I have performed analysis on select variables: *crim, tax, ptratio* and *medv* to show my basic understanding of the concept and techniques. If required, I will provide the analysis on the remaining variables.

2a. Scale of Measurement

Q2. Which is the scale of measurement of each variable? ? Provide frequency distribution, summarization, and visualization of each variable. Describe your findings.

Variable	Description	Typology	Sub Typology
crim	The per capita crime rate by town	Numerical	Ratio scale
zn	The proportion of residential land zoned for large lots (over 25,000 sq. ft)	Numerical	Ratio scale
indus	The proportion of non-retail business acres per town	Numerical	Ratio scale
chas	Whether the town is on the Charles River (1 if it is, 0 if it is not)	Boolean	N/A
nox	Nitrogen-oxide concentration (parts per 10 million)	Numerical	Continuous
rm	Average number of rooms per dwelling	Numerical	Ratio scale
age	Proportion of owner-occupied units built before 1940	Numerical	Ratio scale
dis	The weighted distance to employment centers	Numerical	Continuous
rad	Accessibility to radial highways	Numerical	Discreet
tax	Property tax rate	Numerical	Ratio scale
ptratio	Pupil-teacher ratio	Numerical	Ratio scale
black	A measure of the proportion of residents of African American descent by town	Numerical	Ratio scale
lstat	Percentage of lower status population	Numerical	Ratio scale
medv	Median value of owner-occupied homes in $1000s	Numerical	Continuous

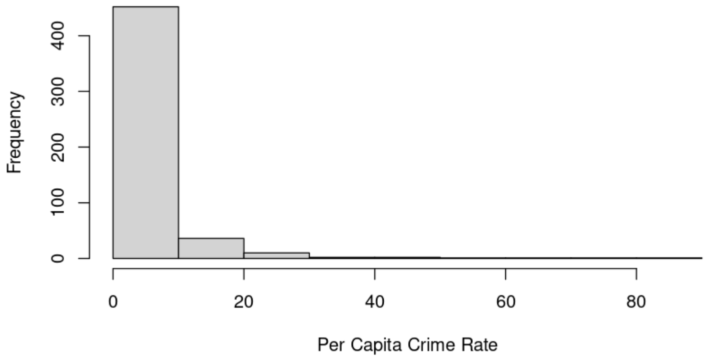
2b. Summaries, Frequency Distributions and Visualizations of Select Variables

crim variable

Min	0.00632	Max	88.9762
Mean	3.61352	Median	0.25651
Q1	0.08205	Q2	3.67708
IQR	3.595038	σ	8.601545

Intervals are created around the frequencies of the per capita crime rate

Label	Ranges	Absolute Frequency	Relative Frequency
Low	0 – 10	452	0.89
Medium	10 – 20	36	0.07
High	20 – 89	18	0.04
Total		506	1.00

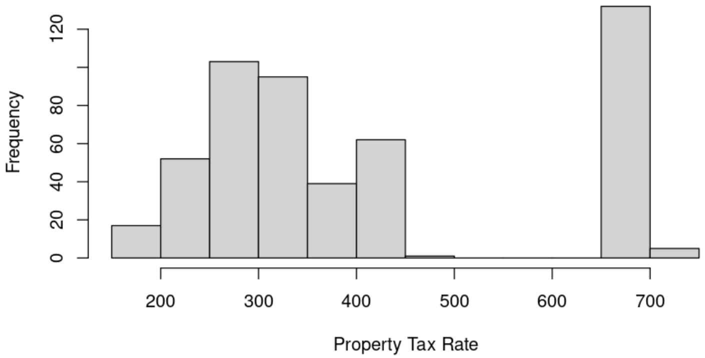


tax variable

Min	187	Max	711
Mean	408.2	Median	330
Q1	279	Q2	666
IQR	387	σ	168.5371

Intervals are created around 1/3 quantiles

Label	Ranges	Absolute Frequency	Relative Frequency
Low	187 – 279	128	0.25
Moderate	279 – 330	128	0.25
High	330 – 666	245	0.48
Very High	666 – 711	5	0.01
Total		506	1.00

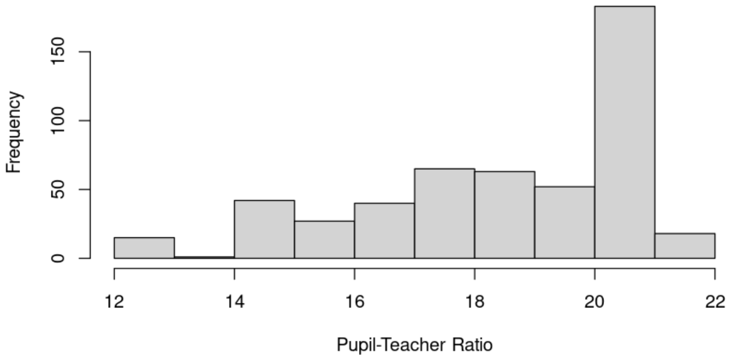


ptratio variable

Min	12.6	Max	22
Mean	18.46	Median	19.05
Q1	17.4	Q2	20.2
IQR	2.8	σ	2.164946

Intervals are created around median

Label	Ranges	Absolute Frequency	Relative Frequency
Low *(Better)*	12.6 – 19.1	253	0.5
High *(Worse)*	19.1 – 22.0	253	0.5
Total		506	1.00



2c. Findings

crim variable

- Since the median is significantly lower than the mean, this suggests that the distribution of the crime rate is positively skewed.
- The high interquartile range indicates a greater degree of variability in the data.
- The high standard deviation suggests a high degree of variability in the data.
- From the frequency distribution, it can be inferred that the majority of neighborhoods are relatively safe for living.

tax variable

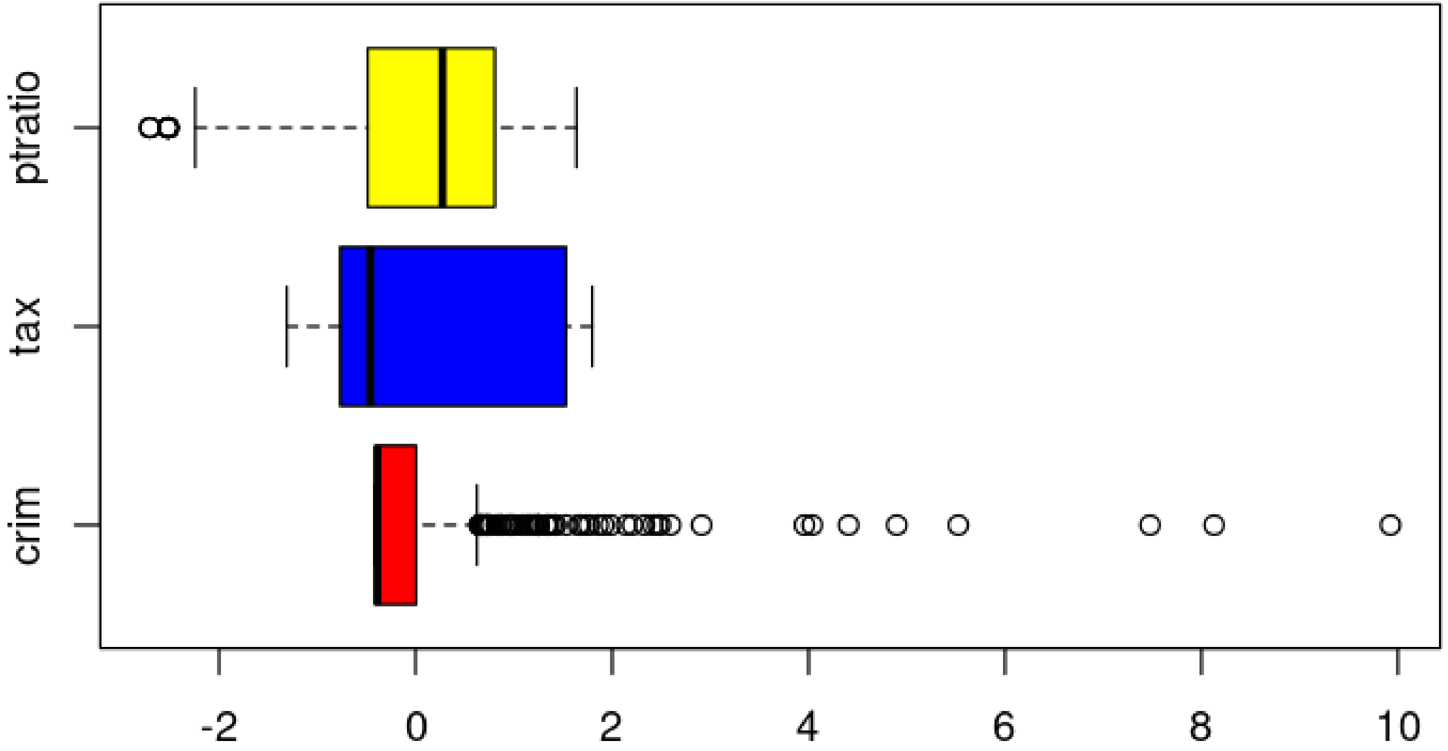
- Since the median is significantly lower than the mean, this suggests that the distribution of the tax rate is positively skewed.
- Since the interquartile range is high, this indicates there is more variability in the data.
- The high standard deviation suggests a high degree of variability in the data.
- The frequency distribution indicates that the majority (48%) of the data falls within the high tax interval.

ptratio variable

- Since the median and mean are relatively close to each other, this indicates that the distribution is somewhat symmetrical with slight negative skewness, as the median is slightly higher than the mean.
- Since the interquartile range is low, this indicates there is less variability in the data.
- The high standard deviation suggests a high degree of variability in the data.
- The frequency distribution suggests that there is a balanced distribution.

3. Identifying Outliers in the *crim*, *tax* and *ptratio* variables

Q3. Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each variable.



The variables were standardized before creating the box – whisker plot.

Observations:

Negative outliers in the *ptratio* variable are indicative of the low pupil to teacher ratio in certain areas.

There are no outliers in the *tax* variable suggesting that the data points are within range with no extreme values.

There are multiple outliers observed in the *crim* variable. This indicates that some areas have high crime rate compared to majority of the areas.

4 – 6. Miscellaneous

Q4. How many of the census tracts in this data set bound the Charles river?

- There are **35** towns on the Charles River.

Q5. What is the median pupil-teacher ratio among the towns in this data set?

- The median pupil-teacher ratio among the towns in this data set is **19.05**.

Q6. Which census tract of Boston has lowest median value of owner- occupied homes? What are the values of the other variables for that census tract, and how do those values compare to the overall ranges for those variables? Comment on your findings.

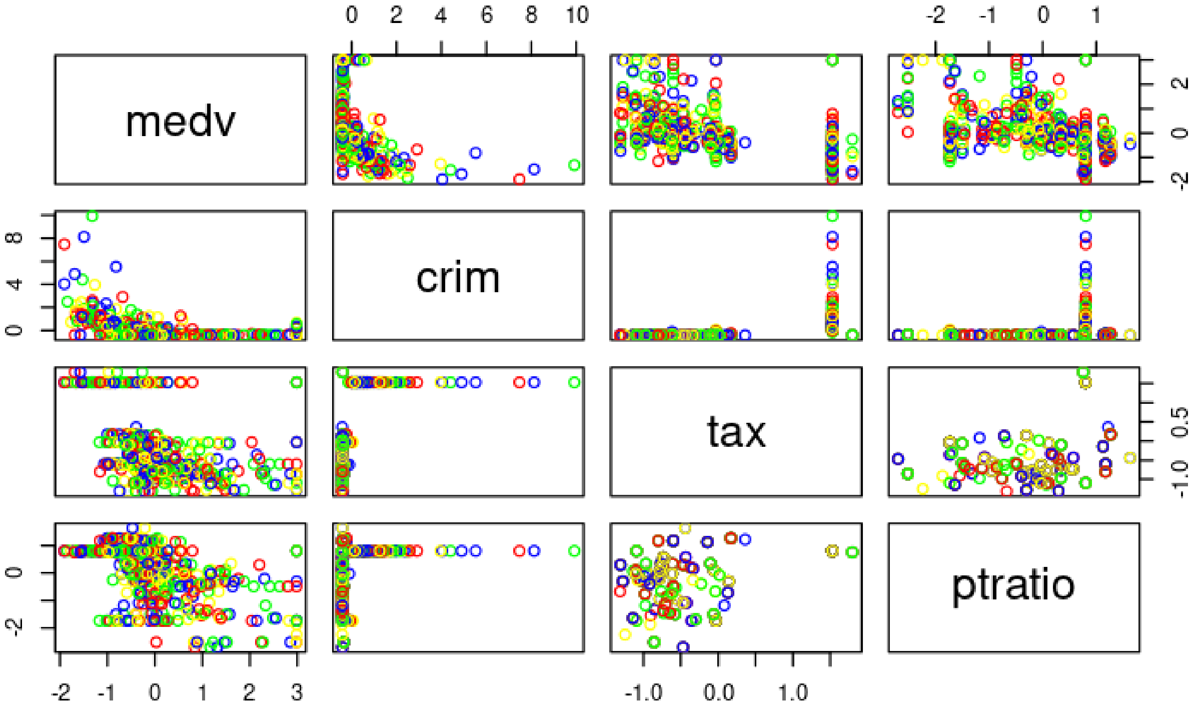
- The census tract with the lowest *medv* is at the indices **399** and **406** with the value of **5**.
- The values of other variables for these indices are as follows:

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat
38.3518	0	18.1	0	0.693	5.453	100	1.4896	24	666	20.2	396.9	30.59
67.9208	0	18.1	0	0.693	5.683	100	1.4254	24	666	20.2	384.97	22.98

- At the lowest *medv* of **5**:
    - The crime rate is significantly high.
    - The neighborhoods are not at the Charles River.
    - The property tax is surprisingly very high
    - The pupil-tutor ratio is high which might indicate less quality of schooling

7. Scatter Plot Matrix of Select Variables

Q7. Make some pairwise scatterplots of the variables in this data set. Describe your findings.



The variables were standardized before creating the scatter plot matrix.

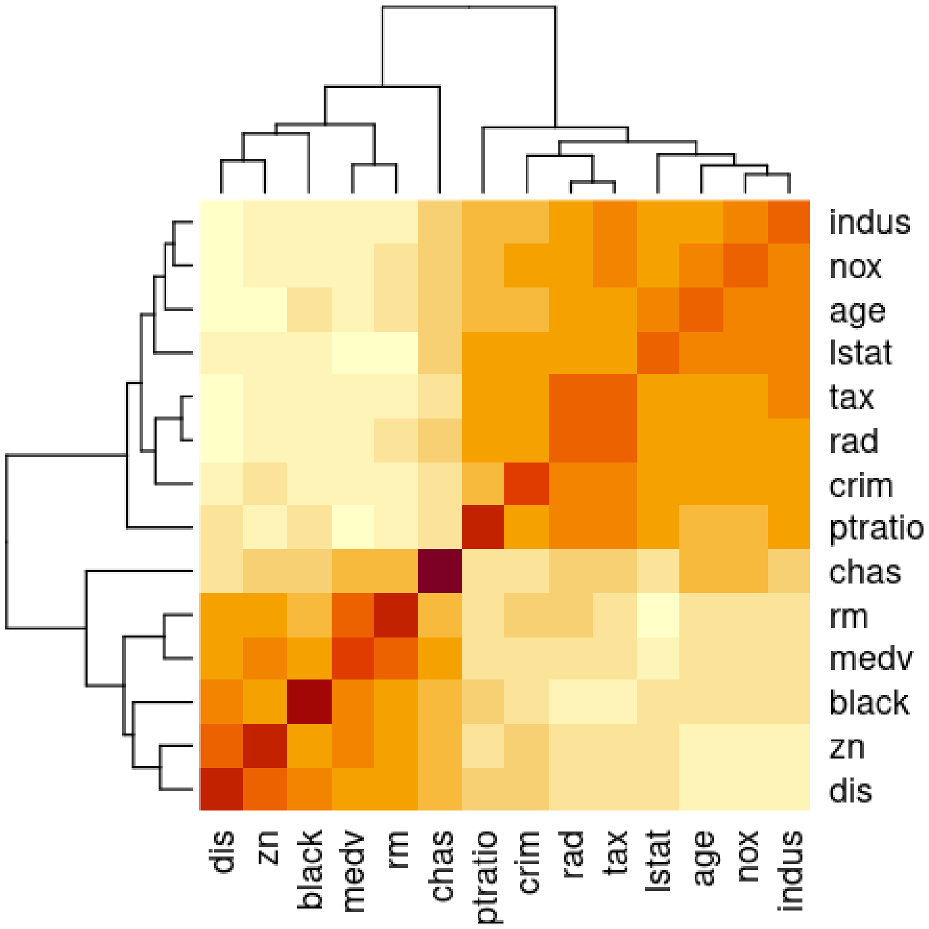
There appears to be no clear linear relationship between the variables which may indicate that there is:

- A non-linear relationship, and/or
- A week linear relationship, and/or
- A presence of outliers

Note: Additional slides' section has details on the correlation coefficients of all the variables and box plots to identify outliers..

8 – 9. Heatmap based on Correlation Coeeficients

Q8. Are any of the predictors to explain “per capita crime rate”? If so, explain the relationship.
Q9. Are any of the predictors to explain the price of the houses as measured by the “Median value of owner-occupied homes in $1000”? If so, explain the relationship.



Heatmap based on the correlation coefficients

Observations:

There appears to be a somewhat strong positive correlation between *per capita crime rate* and *accessibility radial highways*.

- Correlation: 0.62551
- Assumption: Neighborhoods with higher crime rates are situated in areas with better access to transportation.

There appears to be a moderate positive correlation between per capita crime rate and property tax rate.

- Correlation: 0.58276
- Assumption: Neighborhoods with higher crime rates have higher tax rates due to the provision of security.

There appears to be strong positive correlation between median value of owner-occupied homes and average number of rooms per dwelling.

- Correlation: 0.69536
- Assumption: Number of rooms is an important factor in deciding the property value

10. Cross-Classification and Correlation

Q10. Which variable can play the role of the target variable and which predictors can explain it? Provide a cross-classification of the units considering the target variable and one predictor. Is there independence in distribution? Is there independence in mean? Provide measures of suitable statistical indexes to explain the eventual association, the eventual correlation, and the eventual linear correlation. Describe your findings.

The variable *medv*, representing the median value of owner-occupied homes, can serve as the target variable. Among the other variables, it appears that *rm*, denoting the average number of rooms per dwelling, is the better predictor, as it exhibits the highest correlation coefficient.

Correlation	Pearson’s correlation
0.6953599	Coefficient: 0.6953599 p-Value: 2.2e-16

Observations:

There appears to be strong positive correlation between *median value of owner-occupied homes* and *average number of rooms per dwelling*.

The extremely low p-value indicates that correlation is statically significant and there is dependence between the variables.

	rm				
medv	Label	Low 3.56 – 5.99	Medium 5.99 – 6.44	High 6.44 – 8.78	Total
	Low **(5 – 18.8)**	93	52	24	**169**
	Medium **(18.8 – 23.7)**	67	83	21	**171**
	High **(23.7 – 50)**	9	33	124	**166**
	Total	**169**	**168**	**169**	**506**

Since the Boston dataset comprises numerical variables, both medv and rm have been categorized into quantiles to facilitate cross-classification.

Additional Slides

Correlation Coefficients

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
crim	1.00000	-0.20047	0.40658	-0.05589	0.42097	-0.21925	0.35273	-0.37967	0.62551	0.58276	0.28995	-0.38506	0.45562	-0.38830
zn	-0.20047	1.00000	-0.53383	-0.04270	-0.51660	0.31199	-0.56954	0.66441	-0.31195	-0.31456	-0.39168	0.17552	-0.41299	0.36045
indus	0.40658	-0.53383	1.00000	0.06294	0.76365	-0.39168	0.64478	-0.70803	0.59513	0.72076	0.38325	-0.35698	0.60380	-0.48373
chas	-0.05589	-0.04270	0.06294	1.00000	0.09120	0.09125	0.08652	-0.09918	-0.00737	-0.03559	-0.12152	0.04879	-0.05393	0.17526
nox	0.42097	-0.51660	0.76365	0.09120	1.00000	-0.30219	0.73147	-0.76923	0.61144	0.66802	0.18893	-0.38005	0.59088	-0.42732
rm	-0.21925	0.31199	-0.39168	0.09125	-0.30219	1.00000	-0.24026	0.20525	-0.20985	-0.29205	-0.35550	0.12807	-0.61381	0.69536
age	0.35273	-0.56954	0.64478	0.08652	0.73147	-0.24026	1.00000	-0.74788	0.45602	0.50646	0.26152	-0.27353	0.60234	-0.37695
dis	-0.37967	0.66441	-0.70803	-0.09918	-0.76923	0.20525	-0.74788	1.00000	-0.49459	-0.53443	-0.23247	0.29151	-0.49700	0.24993
rad	0.62551	-0.31195	0.59513	-0.00737	0.61144	-0.20985	0.45602	-0.49459	1.00000	0.91023	0.46474	-0.44441	0.48868	-0.38163
tax	0.58276	-0.31456	0.72076	-0.03559	0.66802	-0.29205	0.50646	-0.53443	0.91023	1.00000	0.46085	-0.44181	0.54399	-0.46854
ptratio	0.28995	-0.39168	0.38325	-0.12152	0.18893	-0.35550	0.26152	-0.23247	0.46474	0.46085	1.00000	-0.17738	0.37404	-0.50779
black	-0.38506	0.17552	-0.35698	0.04879	-0.38005	0.12807	-0.27353	0.29151	-0.44441	-0.44181	-0.17738	1.00000	-0.36609	0.33346
lstat	0.45562	-0.41299	0.60380	-0.05393	0.59088	-0.61381	0.60234	-0.49700	0.48868	0.54399	0.37404	-0.36609	1.00000	-0.73766
medv	-0.38830	0.36045	-0.48373	0.17526	-0.42732	0.69536	-0.37695	0.24993	-0.38163	-0.46854	-0.50779	0.33346	-0.73766	1.00000