Question 1 (12 pts)

Import the data set into R. Important things to note:

1. (6 pts) The data set does not contain any column headers, so we have some basic data prep to do. You will need to figure out how to add column headers to headless data in R and end up with a data frame that has column headers and is ready to be used for analytics.
2. (6 pts) Upon successful import you should see 506 total observations and 14 variables in the Environment pane. (iii) Use the View() and head() commands to verify that the data loaded properly.

A screenshot of a computer

Description automatically generated

Question 2 (14 pts)

1. (4 pts) Report summary statistics for the dependent/outcome variable, MEDV - Median value of owner-occupied homes, in the data set.

A close up of numbers

Description automatically generated

1. (6 pts) Plot a histogram and a QQ plot of the dependent variable and comment on normality.



Both the plots suggest that the data deviate from the normal distribution.

Histogram: It shows right skewed distribution. With more data concentrated on the left side and higher values or the outliers on the right side.

QQplot: The points deviate from the straight line at the tails.

1. (4 pts) Would the dependent variable benefit from a log transformation? Briefly explain.



The histogram shows that the data is approximately normally distributed. The QQplot has the points deviated at the tails with most points on the line.

Question 3 (8 pts)

1. (4 pts) Tabulate all the categorical variables in the data set.

A number with numbers on it

Description automatically generated with medium confidence

1. (4 pts) Does there seem to be enough variation in the categorical variables to include them in a model for home values in the Boston area?

The data is more skewed towards ‘0’. Approximately ‘0’ covers 94% variation and ‘1’ has only 6%.

Question 4 (20 pts)

Now that we have looked at some summary statistics and have made a decision on whether (or not) we should log transform our dependent variable, it is time to identify good candidates that could serve as explanatory variables in our model. Estimating a simple regression model for each candidate explanatory variable will allow us to identify and later on filter out the insignificant ones.

1. (13 pts) Estimate a simple linear regression model for the dependent variable using each of the variables in the data set and report the results using the stargazer library.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A white sheet with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A white sheet with black text and numbers

Description automatically generated

A screenshot of a computer

Description automatically generated

A white sheet with black text

Description automatically generated

A white sheet of paper with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. (7 pts) Which variables seem to be significant determinants of home values at the 5% level?

All the variables are significant determinants of home because all of them have’\*\*\*’, it means they all are significant at the 1% level, which is below 5 % level. Therefore, all the variables are significant determinants of home values at the 5% level.

Question 5 (40 pts)

1. (8 pts) Estimate a multiple linear regression model for home values using all the significant explanatory variables in the data set based on your results from Question 4 above and report the regression results using the stargazer library.

Effects of All Variables on Median

===============================================

Dependent variable:

---------------------------

MEDV

-----------------------------------------------

CRIM -0.108\*\*\*

(0.033)

ZN 0.046\*\*\*

(0.014)

INDUS 0.021

(0.061)

CHAS 2.687\*\*\*

(0.862)

NOX -17.767\*\*\*

(3.820)

RM 3.810\*\*\*

(0.418)

AGE 0.001

(0.013)

DIS -1.476\*\*\*

(0.199)

RAD 0.306\*\*\*

(0.066)

TAX -0.012\*\*\*

(0.004)

PTRATIO -0.953\*\*\*

(0.131)

B 0.009\*\*\*

(0.003)

LSTAT -0.525\*\*\*

(0.051)

Constant 36.459\*\*\*

(5.103)

-----------------------------------------------

Observations 506

R2 0.741

Adjusted R2 0.734

Residual Std. Error 4.745 (df = 492)

F Statistic 108.077\*\*\* (df = 13; 492)

===============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

MODEL SUMMARY WITHOUT INDUSTRY AND AGE

Effects of All Variables on Median

===============================================

Dependent variable:

---------------------------

MEDV

-----------------------------------------------

CRIM -0.108\*\*\*

(0.033)

ZN 0.046\*\*\*

(0.014)

CHAS 2.719\*\*\*

(0.854)

NOX -17.376\*\*\*

(3.535)

RM 3.802\*\*\*

(0.406)

DIS -1.493\*\*\*

(0.186)

RAD 0.300\*\*\*

(0.063)

TAX -0.012\*\*\*

(0.003)

PTRATIO -0.947\*\*\*

(0.129)

B 0.009\*\*\*

(0.003)

LSTAT -0.523\*\*\*

(0.047)

Constant 36.341\*\*\*

(5.067)

-----------------------------------------------

Observations 506

R2 0.741

Adjusted R2 0.735

Residual Std. Error 4.736 (df = 494)

F Statistic 128.206\*\*\* (df = 11; 494)

===============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

1. (6 pts) Which coefficients are not statistically significant at the 5% level?

INDUS, AGE coefficients are not statistically significant at the 5% level.

1. (14 pts) Do all the estimated coefficients have the expected sign? Briefly explain for each variable.

* INTERCEPT
  + Positive. This acts a baseline value when all variables are zero.
* Crime Rate
  + Negative. Higher crime rate decreases the house price
* Zoning
  + Positive. More land zoned for large lots increases housing prices
* Industrial areas
  + Ambiguous. Generally, more industrial areas can lower property values but there can be exceptions.
* Charles River proximity
  + Positive. Having a good view helps to increase price
* NOX
  + Negative. More pollution leads lower price
* Rooms
  + Positive. More rooms means more cost
* Age
  + Ambiguous. Older homes can be valued for their charm or lower their value if they need repairs
* Distance
  + Negative. More distant you are from the City, lower price for the house.
* RAD ( highway access)
  + Positive. More access leads to higher price
* TAX
  + Negative. Higher Property tax means Lower prices. As people don’t want to pay more in taxes.
* PT-ratio
  + Negative. High ratio indicates poorer school quality, which can lower property value
* Blacks proportion
  + Ambiguous. Generally, more black population might be associated with Lower property values, but this is influenced by other socio-economic factors.
* LSTAT
  + Negative. A higher percentage of lower status residents typically means lower housing prices.

1. (4 pts) What percentage of the total variation in Boston home values is the model able to explain?

74% of the variation is explained by the model.

1. (8 pts) Does the model appear to violate any of the OLS assumptions (normality, homoskedasticity, etc.)? Support your answer with the necessary graphs and metrics.







The histogram approximately looks like a bell- curve but has little skewness and outliers. The QQplot shows deviations from the straight line at the tails. Overall, this indicates that the residuals may not be normally distributed, which violates the assumption of normality.

The Residuals vs fitted plot does not show a clear pattern which is good but there is are some outliers and increase in spread which might indicate that there is Heteroscedasticity.

The assumptions may not fully met due to deviations from normality and potential heteroscedasticity.

Question 6 (26 pts) In multiple regression modeling that involves many explanatory variables, multicollinearity is always a prime concern and something to be checked and corrected for. One of the tools at our disposal to detect multicollinearity is the what is known as a correlation matrix. The correlation matrix for a data set is simply a symmetric k x k matrix that reports the correlation coefficient of each variable with itself and every other variable in the data set, where k denotes the number of variables.

1. (10 pts) Calculate the correlation matrix using all the explanatory variables in the multiple regression model you estimated in Question 5 above (Hint: the built-in R command to create a correlation matrix is cor()). Regardless of what programming language we use (R, Python, SAS, etc.), as data scientists, we will rarely be told what libraries to pick and which commands to use to get to desired results and perform our analytics work. Therefore, we will need to develop the essential skills to learn about new libraries and how to use them by looking at examples online. This gets easier with experience and Professor Google is our friend here.

2. (10 pts) Create a heat map for the correlation matrix using a suitable library of your choice in R (Hint: corrplot and ggplot2 are some of the common libraries that can be used for this purpose.)



1. (6 pts) Which explanatory variables display the highest degree of multicollinearity? Is there an intuitive relationship between these variables? Briefly explain.

There is high level of multicollinearity between access to radial highways and high property tax. Generally, well connected places have increased demand, potentially driving up the taxes.

Question 7 (22 pts)

1. (10 pts) Calculate the VIF values for all the explanatory variables in the multiple regression model you estimated in Question 5 above. (Hint: you will need to import the car library)



2. (4 pts) Use the barplot() function in R to create a visualization of the VIF values.



3. (4 pts) Add a horizontal reference line to your bar plot using the abline() function at the value 5.



1. (4 pts) Do any of the explanatory variables cross the VIF threshold of 5? Which ones?

Access to radial highways and property tax cross threshold limit of 5.

Question – 8







1. (12 pts) Estimate a revised multiple linear regression model for home values and report the results using the stargazer library.

After taking into consideration the above graphs, changes in adjusted r-square, coefficients and their significance. Model, where Tax variable is removed, is comparatively better than the model where RAD variable is removed.

In terms of adjusted r- squared there is very little variation. Coming to the Coefficients and their significance in Tax model(the model where tax is included) Crime variable got less significance whereas in RAD model all the variables are statistically significant at 1% level.

In the above graph in the scatterplot at the lower tail of the qqplot there is little improvement in TAX removed model.

Considering all this TAX variable is not taken into consideration in the model.

Effects of Variables on Median after removing variables with high multicollinearity

===============================================

Dependent variable:

---------------------------

MEDV

-----------------------------------------------

CRIM -0.105\*\*\*

(0.033)

ZN 0.037\*\*\*

(0.013)

CHAS 2.968\*\*\*

(0.861)

NOX -20.314\*\*\*

(3.472)

RM 3.977\*\*\*

(0.408)

DIS -1.429\*\*\*

(0.187)

RAD 0.129\*\*\*

(0.041)

PTRATIO -1.015\*\*\*

(0.129)

B 0.010\*\*\*

(0.003)

LSTAT -0.528\*\*\*

(0.048)

Constant 34.712\*\*\*

(5.103)

-----------------------------------------------

Observations 506

R2 0.734

Adjusted R2 0.729

Residual Std. Error 4.790 (df = 495)

F Statistic 136.714\*\*\* (df = 10; 495)

===============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

1. (12 pts) Using your revised multiple regression model, calculate a home value prediction for three different scenarios of your choosing. (Hint: Please do not just report three numbers here, report the scenario as well: e.g. ”This is the predicted value for a Boston home with ABCD characteristics:”)

**SCENARIO- 1 Urban Renewal Project**

**Context:** A town undergoing an urban renewal project aimed at improving living conditions and increasing property values.

* **CRIM:** 1.5 (within 1st Quartile)
* **ZN:** 25.0 (within 3rd Quartile)
* **INDUS:** 8.0 (slightly below the Median)
* **CHAS:** 1 (near the river)
* **NOX:** 0.4 (decent air quality)
* **RM:** 6.8 (larger homes)
* **AGE:** 30.0 (modernized homes)
* **DIS:** 3.5 (good access to employment)
* **RAD:** 6 (moderate access)
* **TAX:** 350 (reasonable tax rate)
* **PTRATIO:** 16 (decent pupil-teacher ratio)
* **B:** 200 (moderate diversity)
* **LSTAT:** 8 (lower status of the population)

ANSWER - 34.60333

This could indicate that the predicted median value of **Urban Renewal Project** homes is approximately $34,603

**Scenario 2: Suburban Family Community**

**Context:** A suburban community focused on attracting families and improving education.

* **CRIM:** 0.5 (very low crime rate)
* **ZN:** 50.0 (high proportion of residential land)
* **INDUS:** 4.0 (low industrial presence)
* **CHAS:** 0 (not near the river)
* **NOX:** 0.3 (good air quality)
* **RM:** 7.5 (larger homes for families)
* **AGE:** 15.0 (most homes built post-1950)
* **DIS:** 4.0 (close to employment)
* **RAD:** 3 (good highway access)
* **TAX:** 280 (affordable taxes)
* **PTRATIO:** 14 (good schools)
* **B:** 300 (diversity maintained)
* **LSTAT:** 5 (low lower status)

ANSWER - 40.95488

this value would suggest a predicted median **Suburban Family Community** home value is around $40,955

**Scenario 3: Aging Industrial Town**

**Context:** An older industrial town facing economic challenges.

* **CRIM:** 15.0 (higher crime rate)
* **ZN:** 5.0 (limited residential zones)
* **INDUS:** 20.0 (high industrial presence)
* **CHAS:** 0 (not near the river)
* **NOX:** 0.7 (poor air quality)
* **RM:** 4.5 (smaller, older homes)
* **AGE:** 70.0 (many homes pre-1940)
* **DIS:** 2.0 (far from jobs)
* **RAD:** 2 (poor access)
* **TAX:** 450 (high tax burden)
* **PTRATIO:** 19 (overcrowded schools)
* **B:** 100 (low diversity)
* **LSTAT:** 30 (high lower status)

ANSWER - 0.2407628.

it would indicate a predicted Median Aging Industrial Town home value is approximately **$240.76**