Hand in your completed assignment to the Student Resource Centre by the above date.

The data set for this assignment is in the file house.csv which is available on the course web page. The data have originally been used in a study on influences on housing prices in the greater Boston area in 1978[1]. The data set contains 506 observations which concern the housing values in suburbs of Boston.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRIM</td>
<td>per capita crime rate by town.</td>
</tr>
<tr>
<td>ZN</td>
<td>proportion of residential land zoned for lots over 25,000 sq.ft.</td>
</tr>
<tr>
<td>INDUS</td>
<td>proportion of non-retail business acres per town.</td>
</tr>
<tr>
<td>CHAS</td>
<td>Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).</td>
</tr>
<tr>
<td>NOX</td>
<td>nitric oxides concentration (parts per 10 million).</td>
</tr>
<tr>
<td>RM</td>
<td>average number of rooms per dwelling.</td>
</tr>
<tr>
<td>AGE</td>
<td>proportion of owner-occupied units built prior to 1940.</td>
</tr>
<tr>
<td>DIS</td>
<td>weighted distances to five Boston employment centres.</td>
</tr>
<tr>
<td>RAD</td>
<td>index of accessibility to radial highways.</td>
</tr>
<tr>
<td>TAX</td>
<td>full-value property-tax rate per $10,000.</td>
</tr>
<tr>
<td>PTRATIO</td>
<td>pupil-teacher ratio by town.</td>
</tr>
<tr>
<td>B</td>
<td>1000(Bk − 0.63)^2 where Bk is the proportion of blacks by town.</td>
</tr>
<tr>
<td>LSTAT</td>
<td>% lower status of the population.</td>
</tr>
<tr>
<td>MEDV</td>
<td>Median value of owner-occupied homes in $1000’s (our response).</td>
</tr>
</tbody>
</table>

The purpose of the original study was to identify the impact of clean air on housing prices in the greater Boston area. In this assignment, however, I want you to build a model for prediction.

Please provide your R-code and any necessary output with the answers of the respective section.

1. Read the data into a data frame. Use visualisation and summary statistics for an exploratory analysis. Are there any potential outliers? Are the variable ranges appropriate? Do all variables have the appropriate type for the analysis? (5)

2. Fit a regression model to these data, using MEDV as the response. Test your model for collinearity. Use the stepwise approach to build a submodel, and test its appropriateness. (10)

3. Test your model for linearity, and transform the variables if necessary. Can you get your $R^2$ over 85%? (10)

4. Test your model for influential points. Does your model have high leverage points or large studentised residuals? If so, are these points having an undue influence on any aspect of the fitted model? Justify your reasoning. Remove if necessary. (turn page) (10)

---

5. Test your model for independence of residuals. Explain your findings. Adapt if necessary. 

6. Test your model for normality using QQ plots and the Weisberg-Bingham test. Can we rely on our model for normality? 

7. Use your model to predict the median household in a Boston neighbourhood with the following specifications:

\[
\begin{align*}
\text{CRIM} & = 4, \\
\text{ZN} & = 12, \\
\text{INDUS} & = 10, \\
\text{CHAS} & = 1, \\
\text{NOX} & = 0.5, \\
\text{RM} & = 4, \\
\text{AGE} & = 50, \\
\text{DIS} & = 10, \\
\text{RAD} & = 2, \\
\text{TAX} & = 500, \\
\text{PTRATIO} & = 16, \\
\text{B} & = 380, \\
\text{LSTAT} & = 15.
\end{align*}
\]

In the light of all previous findings, how much do you trust your prediction? Would you use your model to make statements about causal relationships? 

---

Total: (50)

Hints:

- Exploratory analysis can consist of a pairs plot, but may also contain extra plots for more detailed views. Be aware that the pairs plot for the full data will be of size \(14 \times 14\).

- Recall, that good prediction models have a high \(R^2\) and satisfy the normality of residuals condition for reliable prediction models.

- Recall, that the \texttt{newdata} argument for the \texttt{predict} function only needs values for the variables used in the model.

- If you prefer alternative tests to Weisberg-Bingham or Durbin-Watson, feel free to use them.