STAT40830 Advanced Data Programming with R - Homework 1

Aoife Doyle Student No: 15419052

Dataset Description

For this assignment we have chosen to use the built in **Boston Housing** dataset from the MASS package which contains housing data for various suburbs of Boston. Utliising the `?` and `str()` functions we are able to call the R help file and looking at the structure of our dataset we are able to determine that our dataset has 14 variables and 506 observations. It should be noted that almost all of our variables are numeric in nature aside from our chas (river dummy variable) and rad (an index variable for access to urban highways) which are both classified as being of type interger.

```
library(MASS) # loading in the MASS package
?Boston # calls the R help file to see a brief descriptions of the datasets
# variables
str(Boston) # function that allows us to see the structure of our dataset and
```

```
'data.frame':   506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm     : num  6.58 6.42 7.18 7 7.15 ...
 $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```
# its variables
head(Boston, 10) # prints the first 10 rows of our dataset
```

```
      crim   zn indus chas   nox    rm   age    dis rad tax ptratio  black
1  0.00632 18.0  2.31    0 0.538 6.575  65.2 4.0900   1 296    15.3 396.90
2  0.02731  0.0  7.07    0 0.469 6.421  78.9 4.9671   2 242    17.8 396.90
3  0.02729  0.0  7.07    0 0.469 7.185  61.1 4.9671   2 242    17.8 392.83
4  0.03237  0.0  2.18    0 0.458 6.998  45.8 6.0622   3 222    18.7 394.63
5  0.06905  0.0  2.18    0 0.458 7.147  54.2 6.0622   3 222    18.7 396.90
6  0.02985  0.0  2.18    0 0.458 6.430  58.7 6.0622   3 222    18.7 394.12
7  0.08829 12.5  7.87    0 0.524 6.012  66.6 5.5605   5 311    15.2 395.60
```

```
8  0.14455 12.5  7.87    0 0.524 6.172  96.1 5.9505   5 311    15.2 396.90
9  0.21124 12.5  7.87    0 0.524 5.631 100.0 6.0821   5 311    15.2 386.63
10 0.17004 12.5  7.87    0 0.524 6.004  85.9 6.5921   5 311    15.2 386.71
   lstat medv
1   4.98 24.0
2   9.14 21.6
3   4.03 34.7
4   2.94 33.4
5   5.33 36.2
6   5.21 28.7
7  12.43 22.9
8  19.15 27.1
9  29.93 16.5
10 17.10 18.9
```

Descriptive Statistics Analysis

Examing the summary statistics below reveals a wide range of values across key housing attributes for our Boston dataset. For instance, examining our variable for the crime rate per capita (crim) shows that crime rates in Boston can vary drastically from as low as 0.006 to nearly 89, indicating stark contrasts in different neighborhoods safety levels. The median value of homes (medv) ranges from **$5,000** to **$50,000**, with a mean of **$22,530**, suggesting some skewness and possible value capping at the upper end of our distributions tails.

There is also variation in socio-economic indicators such as the percentage of lower-status residents (lstat), which ranges from *1.73%* to *37.97%*, and property tax rates (tax), which range from *187* to *711*. Additionally, the average number of rooms per dwelling (rm) is around *6.28*, with most homes having between *5* and just under *7* rooms. These variables provide valuable insights into the structural and environmental characteristics influencing housing prices in the Boston area.

```r
summary(Boston) # prints summary statistics for each variable in our dataset
```

```
      crim                zn              indus            chas        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
      nox               rm             age              dis        
 Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
 1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
 Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
 Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
 3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
 Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
      rad              tax           ptratio          black       
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
 Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
     lstat            medv      
 Min.   : 1.73   Min.   : 5.00  
```

```
 1st Qu.: 6.95   1st Qu.:17.02
 Median :11.36   Median :21.20
 Mean   :12.65   Mean   :22.53
 3rd Qu.:16.95   3rd Qu.:25.00
 Max.   :37.97   Max.   :50.00
```

In line with our summary statistics Figure 1 below presents boxplots of all variables in our Boston dataset. For several variables such as crime rate (crim), property tax (tax), and the percentage of lower-income residents (lstat), we find that they exhibit strong right-skewed distributions with significant outliers which suggests the presence of extreme values in certain Boston neighborhoods. In contrast, variables such as number of rooms (rm) and the pupil-teacher ratio (ptratio) appear to be more symmetrically distributed.

Overall we find that our plot also highlights the wide variability across features, which is line with previous findings around substantial variation in socio-economic and environmental disparities among different neighbourhoods. However,to further understand how these variables are interrelated, we now turn to the correlation matrix in Figure 2, which provides deeper insights into the strength and direction of linear relationships between our datasets key variables.

```r
library(ggplot2)  # Functions used for visualization
library(reshape2) # Functions used for reshaping our data

# Utilising the melt function to reshape the data into a long format for data
# visualisation
boston_long <- melt(Boston)
```

```
No id variables; using all as measure variables
```

```r
# Creating a boxplots for our variables
ggplot(boston_long, aes(x = variable, y = value)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", outlier.color = "darkred") +
  theme_minimal() +
  labs(title = "Figure 1: Boxplot of our Boston Housing Variables",
       x = "Variable", y = "Value") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

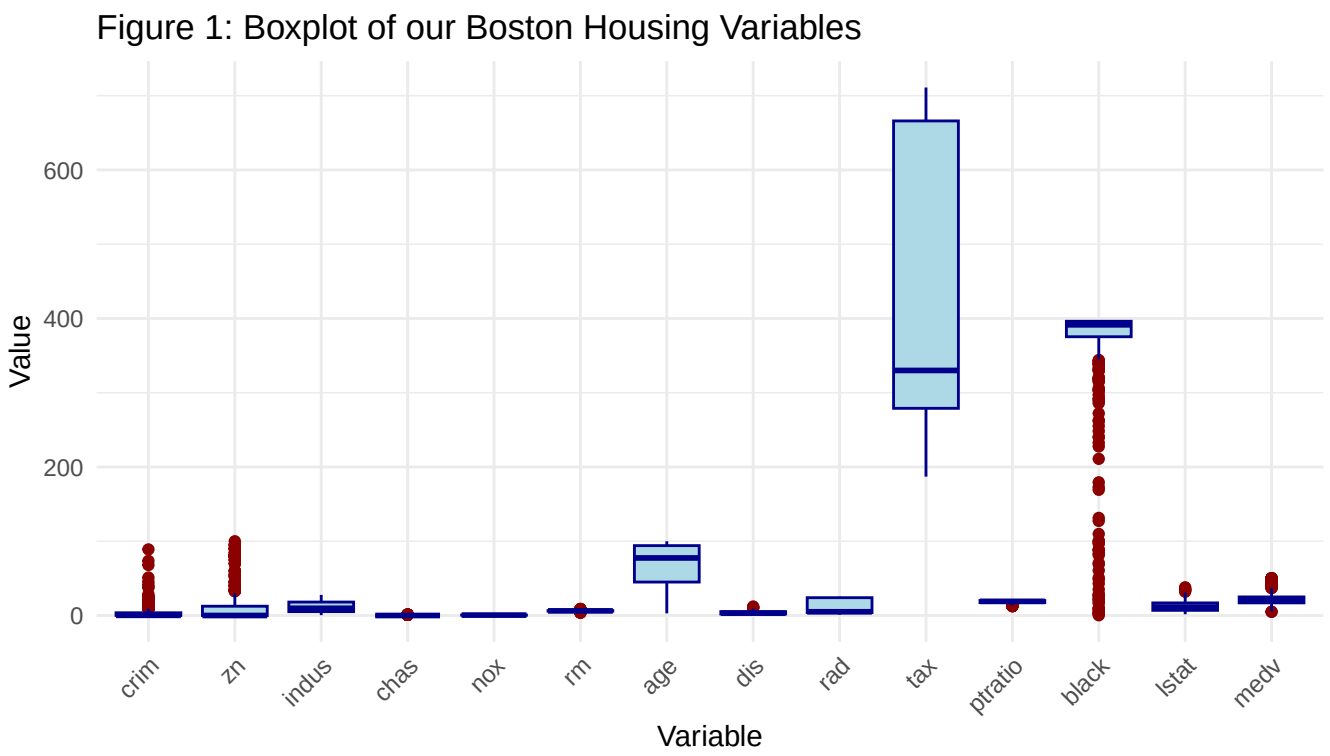

Figure 1: Boxplot of our Boston Housing datasets Variables

Figure 2 below presents the correlation matrix for our Boston Housing dataset in order to highlight the strength and direction of relationships between our numeric variables. Overall we find that the strongest positive correlation is between rad (index of accessibility to radial highways) and tax (property tax rate), with a coefficient of *+0.91*, suggesting that areas with greater access to highways are significantly more likely to have higher property taxes.

In contrast, we observed that our strongest negative correlation was between lstat (percentage of lower-status population) and medv (median home value), with a correlation coefficient of *–0.74*, thus reinforcing the fact that areas with higher concentration of economically disadvantaged residents will tend to have lower housing prices.

Lastly we find that houses in proximity to the Charles River (chas) shows the weakest correlations among all of our variables, particularly with zn (the proportion of land zoned for lots over 25,000 sq.ft.) with *-0.04*, indicating that being close to a river has little to no linear relationship with zoning and other housing characteristics in our dataset overall.

In sum our correlation analysis highlights strong impact that key socioeconomic and structural factors such as access to highways, number of rooms, and income level can have on influencing housing values in Boston, while other variables like proximity to the river appear to have minimal to no impact.

```
# Creating our correlation matrix for our heat map
numeric_boston <- Boston[, sapply(Boston, is.numeric)]
corr_matrix <- cor(numeric_boston)

library(ggcorrplot) # functions for reordering the correlation matrix and
# displays the significance level on the plot.

# Creating a Heatmap using ggcorrplot to see the relationship between variables
ggcorrplot(corr_matrix,
           method = "square", # squares instead of circles for data visuals
           type = "lower", # show only lower triangle
           lab = TRUE, # add in correlation variables
           lab_size = 3,
```

```
          colors = c("darkred", "white", "darkblue"),
          title =
            "Figure 2: Correlation Matrix for our Boston Dataset") +
theme_minimal()
```

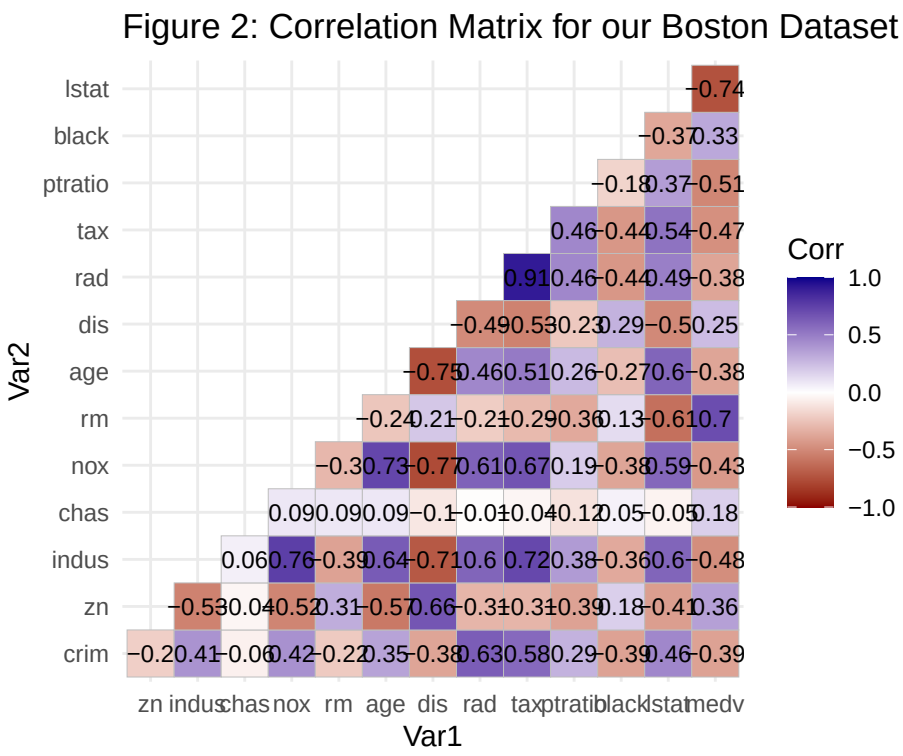

Figure 2: Correlation Matrix for our Boston Housing Dataset