Final Project

STAE04 - Statistics: Data Visualisation

Thi Kim Hong Nguyen

2024-03-19

In this project, I am choosing Boston Dataset from the MASS package in R, which contains information collected from the U.S. Census about housing values in the Boston area from 1978. The dataframe BostonHousing contains the original data by Harrison and Rubinfeld (1979) Hedonic prices and the demand for clean air. J. Environ. Economics and Management, contains 506 observations and 14 variables, which can be used to explore various aspects of the city's housing market. In order to view the Boston data set, we must load the MASS package and take a glimpse of the data.

```
## Rows: 506
## Columns: 14
## $ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, ~
## $ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12~
## $ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.8~
## $ chas    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ nox     <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.5~
## $ rm      <dbl> 6.58, 6.42, 7.18, 7.00, 7.15, 6.43, 6.01, 6.1~
## $ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.~
## $ dis     <dbl> 4.09, 4.97, 4.97, 6.06, 6.06, 6.06, 5.56, 5.9~
## $ rad     <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, ~
## $ tax     <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, ~
## $ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.~
## $ black   <dbl> 397, 397, 393, 395, 397, 394, 396, 397, 387, ~
## $ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19~
## $ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.~
```

The variables can be described as follows:

Table 1: The variables of the data set

Attribute	Description
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft.
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

Attribute	Description
nox	nitrogen oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted mean of distances to five Boston employment
centres	
rad	index of accessibility to radial highways
tax	full-value property-tax rate per $10,000
ptratio	pupil-teacher ratio by town
black	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by
town.	
lstat	lower status of the population (percent)
medv	median value of owner-occupied homes in $1000s

1 Cleaning and Visualizing the dataset

As we can see from the output, there are no missing or duplicated values in the Boston dataset.

```
## Missing values:  0
```

```
##    crim      zn   indus    chas     nox      rm     age
##       0       0       0       0       0       0       0
##     dis     rad     tax ptratio   black   lstat    medv
##       0       0       0       0       0       0       0
```

```
## package 'ggcorrplot' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\DellVostro5590\AppData\Local\Temp\Rtmpq81W3y\downloaded_packages
```

The plot clearly illustrates the correlation between variables, with red indicating negative correlation, white representing no correlation, and green indicating positive correlation. Various shades of colors are employed to depict different levels of correlation, where darker shades of green signify a stronger positive correlation, and darker shades of red signify a stronger negative correlation. From the plot, it can be inferred that the variables DIS and NOX exhibit a robust negative correlation, while the variables TAX and RAD display a pronounced positive correlation. Now We can create a similar histogram of average number of rooms per dwelling.

Then, I create a scatter plot to visualize the relationship between age and median home value. The plot reveals that, for most houses, there is a tendency for the value of owner-occupied houses to decrease as the age of the house increases. However, there is also a small proportion of houses where the price increases with age.

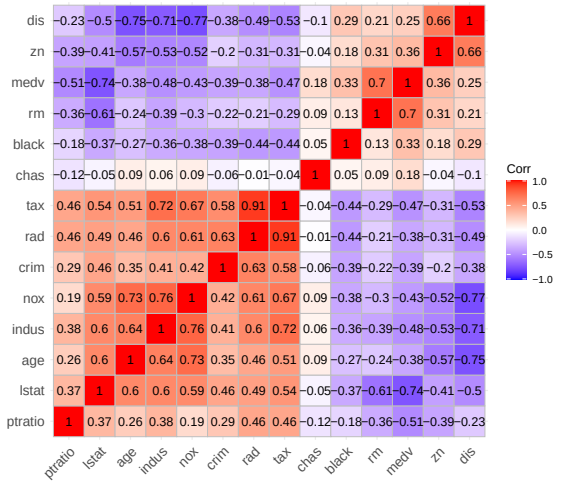

Figure 1: Correlation Heatmap Between Variables

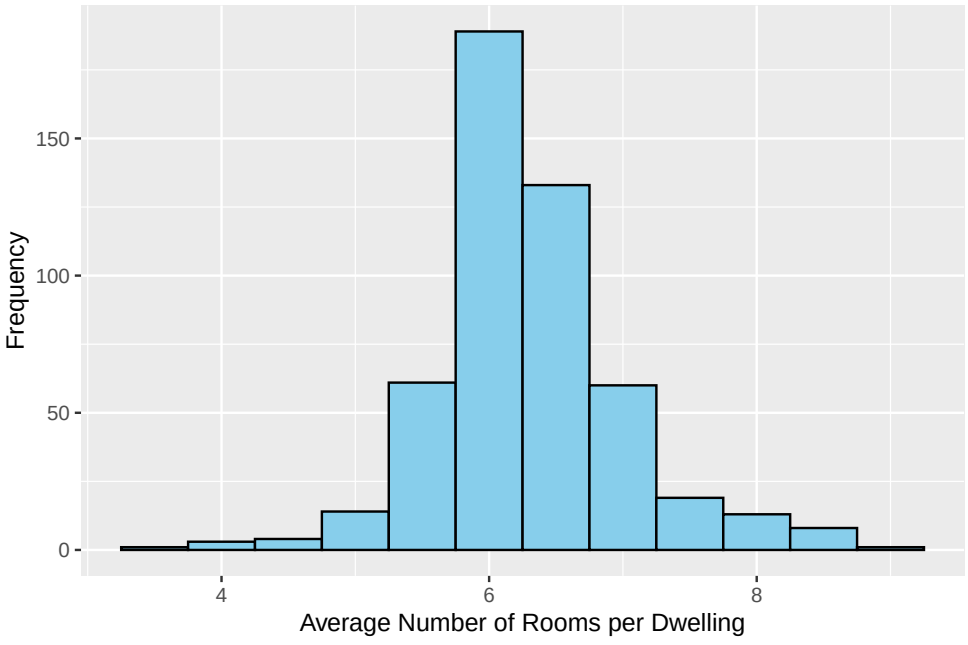

Figure 2: Histogram of Rooms per Dwelling

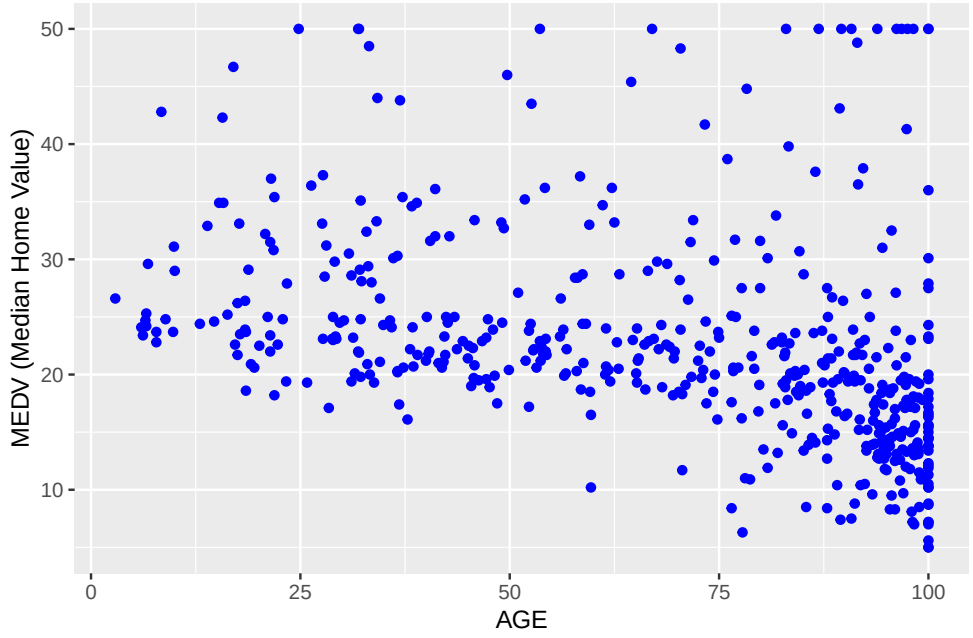

Figure 3: Relationship between Age and Median Home Value

2 Fitting a Linear Regression Model

Let's fit a model for the Boston data set, which help to predict crime rate from median home value, the average numbers of rooms per house and proximity to the Charles River.

We first try to regress lower status of population on median home value.

In conclusion, this linear regression model crafted to forecast the per capital crime rate in Boston exhibited satisfactory performance. Notably, variables like median home value, average number of rooms per dwelling, and proximity to the Charles River were identified as significant influencer of crime rates. While the model provides valuable insights, it's essential to acknowledge and address its underlying assumptions and potential limitations.

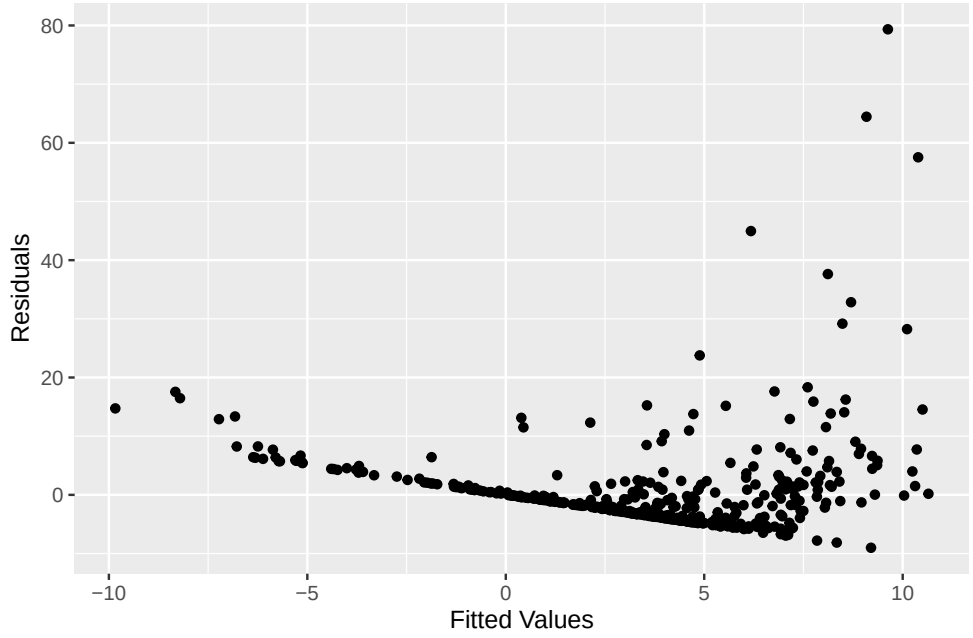

Figure 4: Residuals vs. Fitted Values Plot

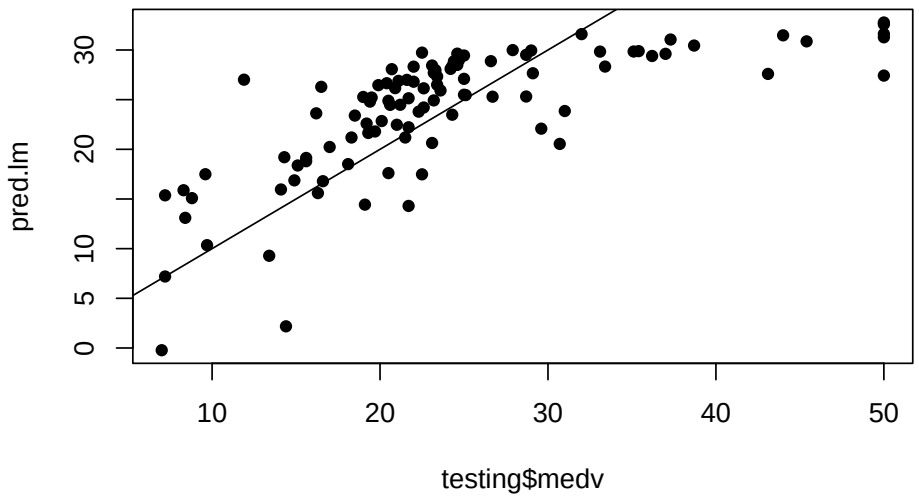

Figure 5: Regress lower status of population on median home value

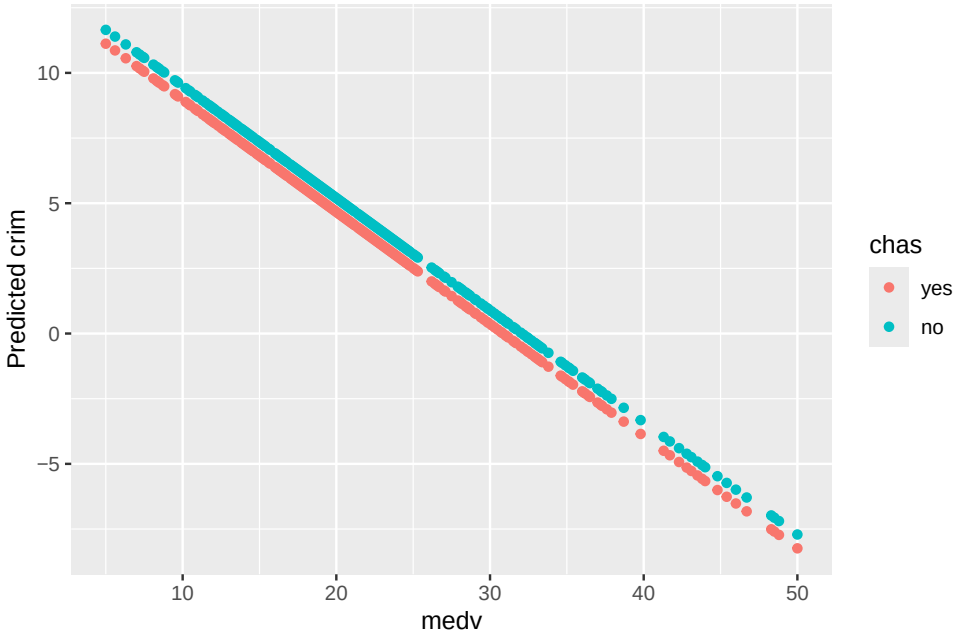

Figure 6: Prediction Plot