İST438 - HW - Week 1

Uğur DAR

01 03 2021

Contents

Figure 1: (https://rajivsworklife.files.wordpress.com/2018/02/boston.jpg?w=675&h=448)

Boston Housing Data

Housing data for 506 census tracts of Boston from the 1970 census. The dataframe BostonHousing contains the original data by Harrison and Rubinfeld (1979), the dataframe BostonHousing2 the corrected version with additional spatial information.

Packages Importing

```
# install.packages("mlbench") # Installing the package for the data.
library(mlbench) # Importing the package.
library(dplyr) # For glimpse function.
library(ggplot2)
library(purrr)
library(tidyr)
library(e1071) # For skewness and kurtosis functions
#knitr::opts_chunk$set(echo = FALSE)
library(knitr)  # For tables
library(kableExtra) # For tables
library(ggpubr)
library(corrplot)
library(RColorBrewer)
```

```
data(BostonHousing) # Calling the data from mlbench
```

```
?BostonHousing # For data description
```

```
## starting httpd help server ... done
```

Features

The original data are 506 observations on 14 variables, medv being the target variable:

- 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)
- nox nitric 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 distances to five Boston employment centres
- rad index of accessibility to radial highways
- tax full-value property-tax rate per USD 10,000
- ptratio pupil-teacher ratio by town
- b 1000(B - 0.63)^2 where B is the proportion of blacks by town
- lstat percentage of lower status of the population
- medv median value of owner-occupied homes in USD 1000's

The corrected data set has the following additional columns: - cmedv corrected median value of owner-occupied homes in USD 1000's - town name of town - tract census tract - lon longitude of census tract - lat latitude of census tract

Structure of The Boston Housing Data

```
glimpse(BostonHousing)
```

```
## Rows: 506
## Columns: 14
## $ crim    <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, 0.088...
## $ zn      <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5, 12.5...
## $ indus   <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7.87, 7.87,...
## $ chas    <fct> 0, 0, 0, 0, 0, 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, 0.524, 0.524, 0.5...
## $ rm      <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6.172, 5.6...
## $ age     <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 100.0, 85.9...
## $ dis     <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5.5605, 5.9...
## $ rad     <dbl> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4,...
## $ tax     <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311,...
## $ ptratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2,...
## $ b       <dbl> 396.90, 396.90, 392.83, 394.63, 396.90, 394.12, 395.60, 396...
## $ lstat   <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.93, 17...
## $ medv    <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16.5, 18.9,...
```

Data has 13 features, 1 target which is **medv**, 506 instances. Target variable is continues, 12 features are numeric, 1 feature (**chas**) is categorical. Also **rad** feature is discrete, other numeric features are continues.

Exploratory Data Analysis

Missing Values

```
anyNA(BostonHousing)
```

```
## [1] FALSE
```

Boston data hasn't missing value.

Summary Statistics

```
sumStats <- function(x){   # An alternative summary function,
  sum <- sum(x)            # I added sum, mean, variance, skewness, kurtosis statistics.
  min <- min(x)
  max <- max(x)
  mean <- mean(x)
  q1 <- quantile(x,0.25);median <- median(x);q3 <- quantile(x,0.75)
  var <- var(x)
  skew <- skewness(x)
  kurt <- kurtosis(x)
  df <- data.frame(Sum = sum, Min = min, Max = max, Mean=mean, Q1=q1, Median = median,
                   Q3=q3, Variance = var, Skewness = skew, Kurtosis = kurt)
  df <- round(df,4)
}
```

Table 1: Summary Statistics of Boston Housing Data

	crim	zn	indus	nox	rm	age
Sum	1828.443	5750	5635.21	280.6757	3180.025	34698.9
Min	0.0063	0	0.46	0.385	3.561	2.9
Max	88.9762	100	27.74	0.871	8.78	100
Mean	3.6135	11.3636	11.1368	0.5547	6.2846	68.5749
Q1	0.082	0	5.19	0.449	5.8855	45.025
Median	0.2565	0	9.69	0.538	6.2085	77.5
Q3	3.6771	12.5	18.1	0.624	6.6235	94.075
Variance	73.9866	543.9368	47.0644	0.0134	0.4937	792.3584
Skewness	5.1922	2.2125	0.2933	0.725	0.4012	-0.5954
Kurtosis	36.5958	3.9524	-1.2402	-0.0874	1.8418	-0.978

Table 2: Summary Statistics of Boston Housing Data

	dis	rad	tax	ptratio	b	lstat	medv
Sum	1920.292	4832	206568	9338.5	180477.1	6402.45	11401.6
Min	1.1296	1	187	12.6	0.32	1.73	5
Max	12.1265	24	711	22	396.9	37.97	50
Mean	3.795	9.5494	408.2372	18.4555	356.674	12.6531	22.5328
Q1	2.1002	4	279	17.4	375.3775	6.95	17.025
Median	3.2074	5	330	19.05	391.44	11.36	21.2
Q3	5.1884	24	666	20.2	396.225	16.955	25
Variance	4.434	75.8164	28404.76	4.687	8334.752	50.9948	84.5867
Skewness	1.0058	0.9989	0.666	-0.7976	-2.8733	0.9011	1.1015
Kurtosis	0.4576	-0.8789	-1.1503	-0.3048	7.1037	0.4628	1.451

```
summaries <- sapply(BostonHousing %>% select(where(is.numeric)), sumStats)
summaries <- as.data.frame(summaries)
```

```
kable(summaries[,1:6], format="latex", booktabs=TRUE,
      caption = "Summary Statistics of Boston Housing Data") # Table 1
```

```
kable(summaries[,7:13], format="latex", booktabs=TRUE,
      caption = "Summary Statistics of Boston Housing Data")  # Table 2
```

In Tables 1 and 2, it is seen that the variance of **tax** and **b** features is very high. **age**, **ptratio**,**b** are negative(left) skewed. The means of the **rm** and **medv** variables are in the middle of their minimum and maximum values, and their medians are close to their mean, these features may have a normally distributed. Graphs can used to better examine the distribution of features.

Histograms of Boston Housing

```
BostonHousing %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram() +
    theme_minimal()
```

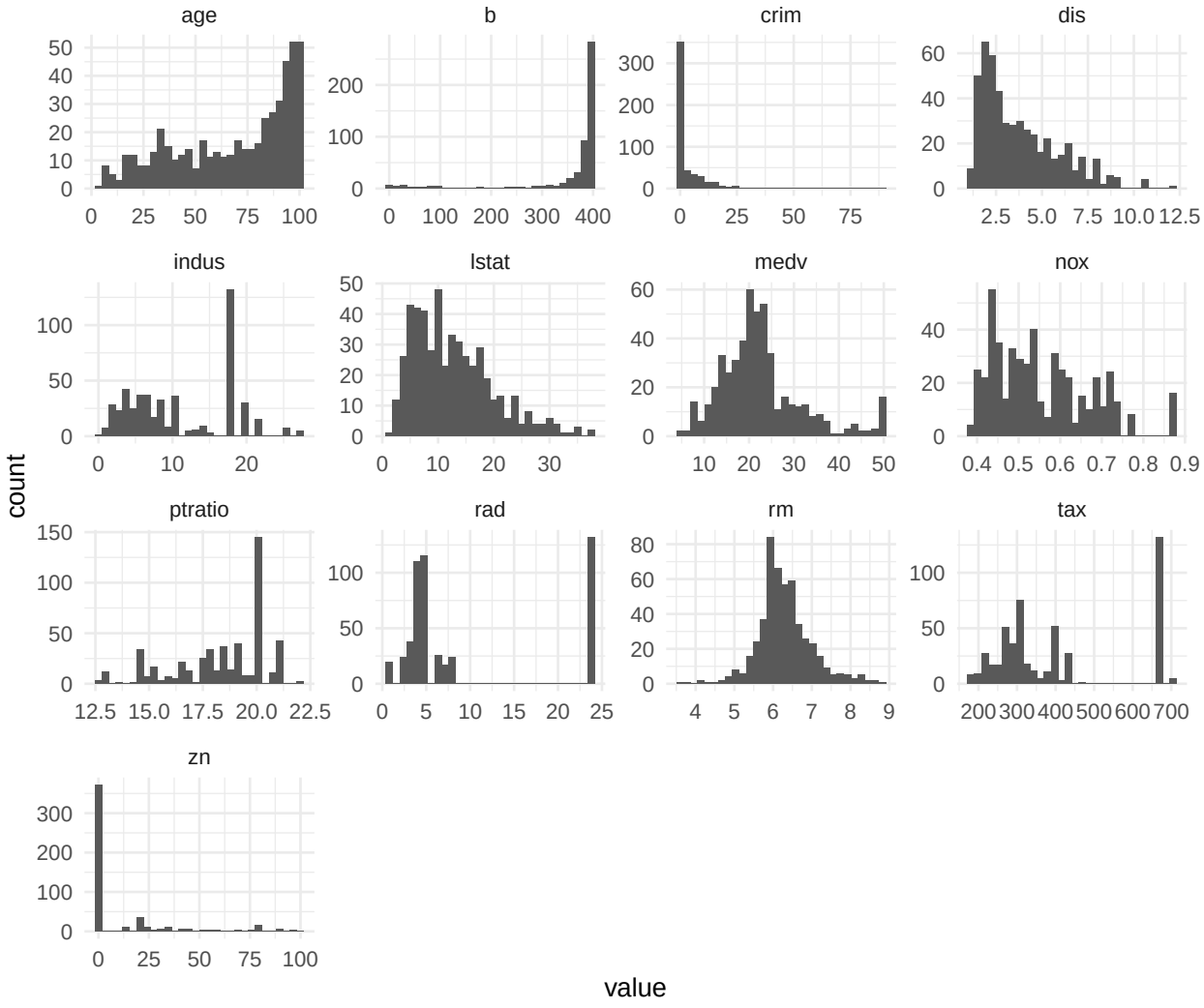


Figure 2: Histograms of Boston Housing - Numeric Features

Box Plots of Boston Housing

```
BostonHousing %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(y=value)) +
    facet_wrap(~ key,scales="free") +
    geom_boxplot() +
    theme_minimal()
```

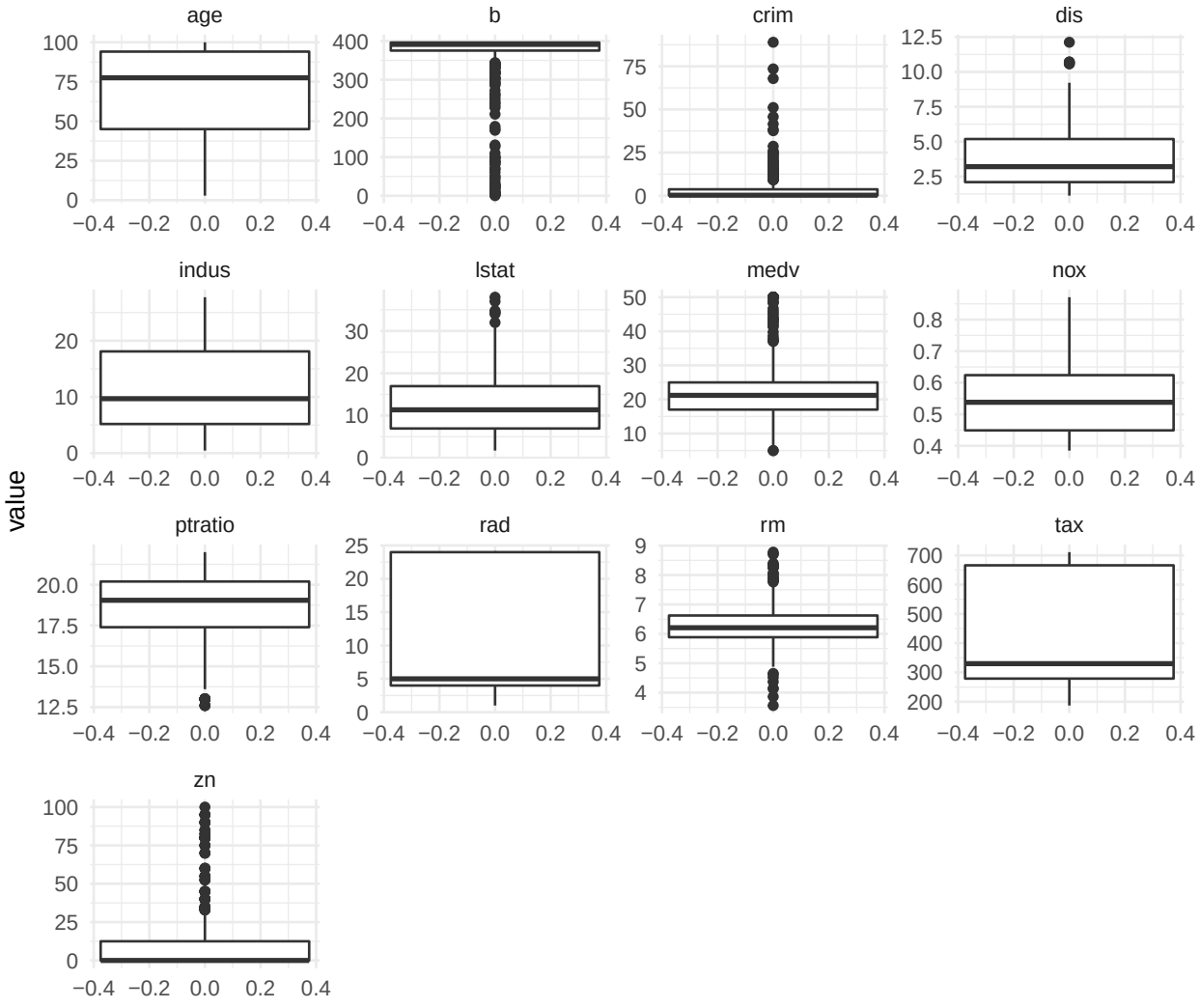


Figure 3: Box Plots of Boston Housing - Numeric Features

b,**crim**, **medv**,**rm**, **lstat** have too many outliers, **zn**,**b**,**crim** highly skewed.

Barplot of Boston Housing

```
ggplot(data=BostonHousing, aes(x=chas)) +
    geom_bar() +
    geom_text(aes(label = scales::percent(..prop..), group = 1),
              fontface = "bold",colour = "#CE2929", size = 5,stat= "count") +
    theme_minimal()
```

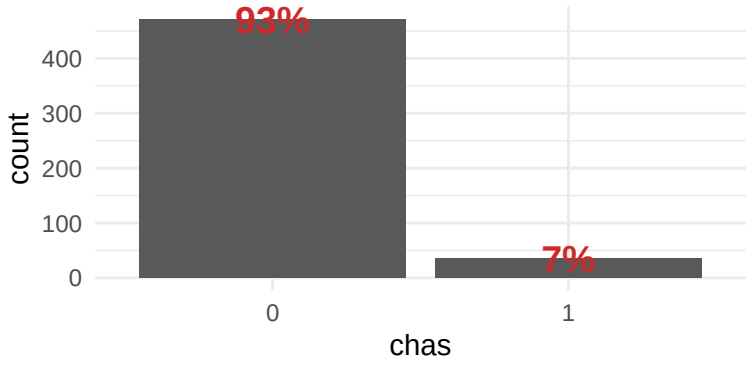


Figure 4: Barplot of Boston Housing - chas

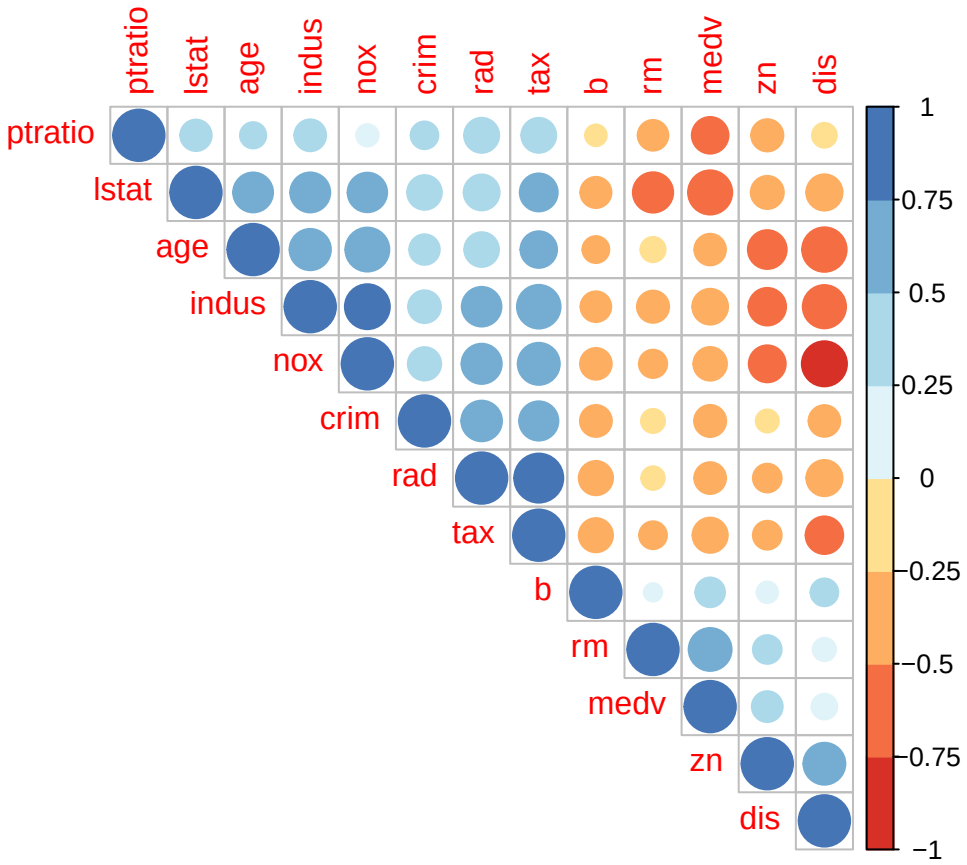
In Boston Housing data set, only 7% house close to Charles River.

Examining Relationship Between Target and Features

Correalation Plot

Correlation plot can use for examining linear relationship in the data.

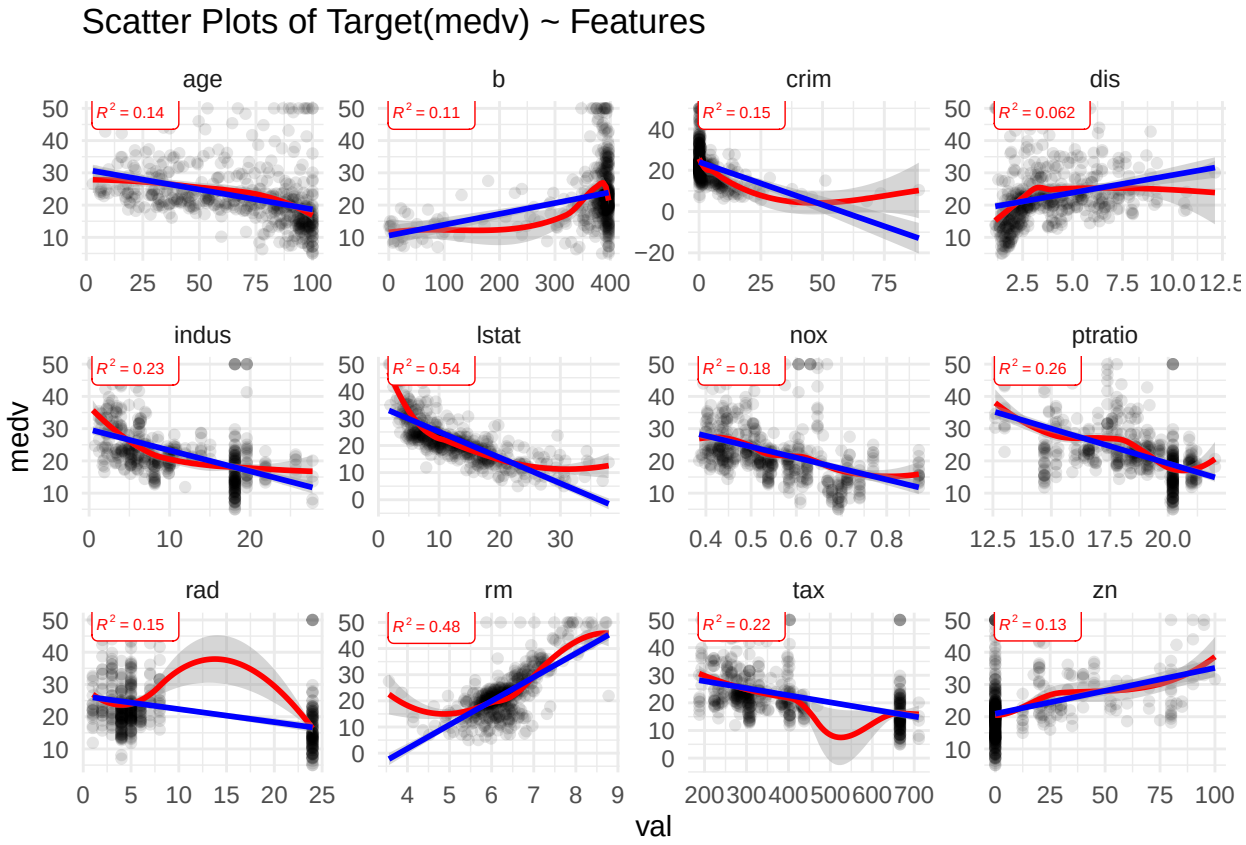
```
M <-cor(BostonHousing %>% select(-chas))
corrplot(M, type="upper", order="hclust",
         col=brewer.pal(n=8, name="RdYlBu"))
```



Some features have linear relationship like **nox** and **dis**,**rad** and **tax**, multicollinearity problem can be seen when modelling.

```
BostonHousing %>%
  select(-chas) %>%
  gather(key, val, -medv) %>%
  ggplot(aes(x = val, y = medv)) +
  geom_point(alpha=0.1) +
  stat_smooth(formula = y ~ x,method = "loess",size = 1, se = TRUE, col = "red") +
  stat_smooth(formula = y ~ x,method = "lm",size = 1, se = TRUE, col = "blue") +
  facet_wrap(~key, scales = "free") +
  theme_minimal() +
  ggtitle("Scatter Plots of Target(medv) ~ Features") +
  stat_cor(aes(label = ..rr.label..), color = "red",
           geom = "label",size=2,hjust =0.01, vjust = 0.6)
```

Scatter Plots



Red lines shows non-linear smooth ,blue lines shows linear smooth between features and target. The feature **lstat** seems to be the most contributing feature, as expected, where income is low, house prices are cheap, the relationship between **lstat** and **medv** variables is non-linear. There seems to be a decrease in house prices as the crime rate increases. While the distance to employment centers is below 2.5, a rapid increase is seen in house prices as the distance increases. There appears to be a weak negative relationship between the **indus** variable and the **medv**. There is a positive relationship between **rm** and **medv**, it seems that as the number of rooms in the house increases, the price of the house increases. **pratio** and **medv** has negative relationship, pupil-teacher income may be considered low. There is a weak negative relationship between the **age** of the building and its price **medv**. It can be thought that the old buildings were restored and used.