6203hw1

2025-01-08

R Markdown

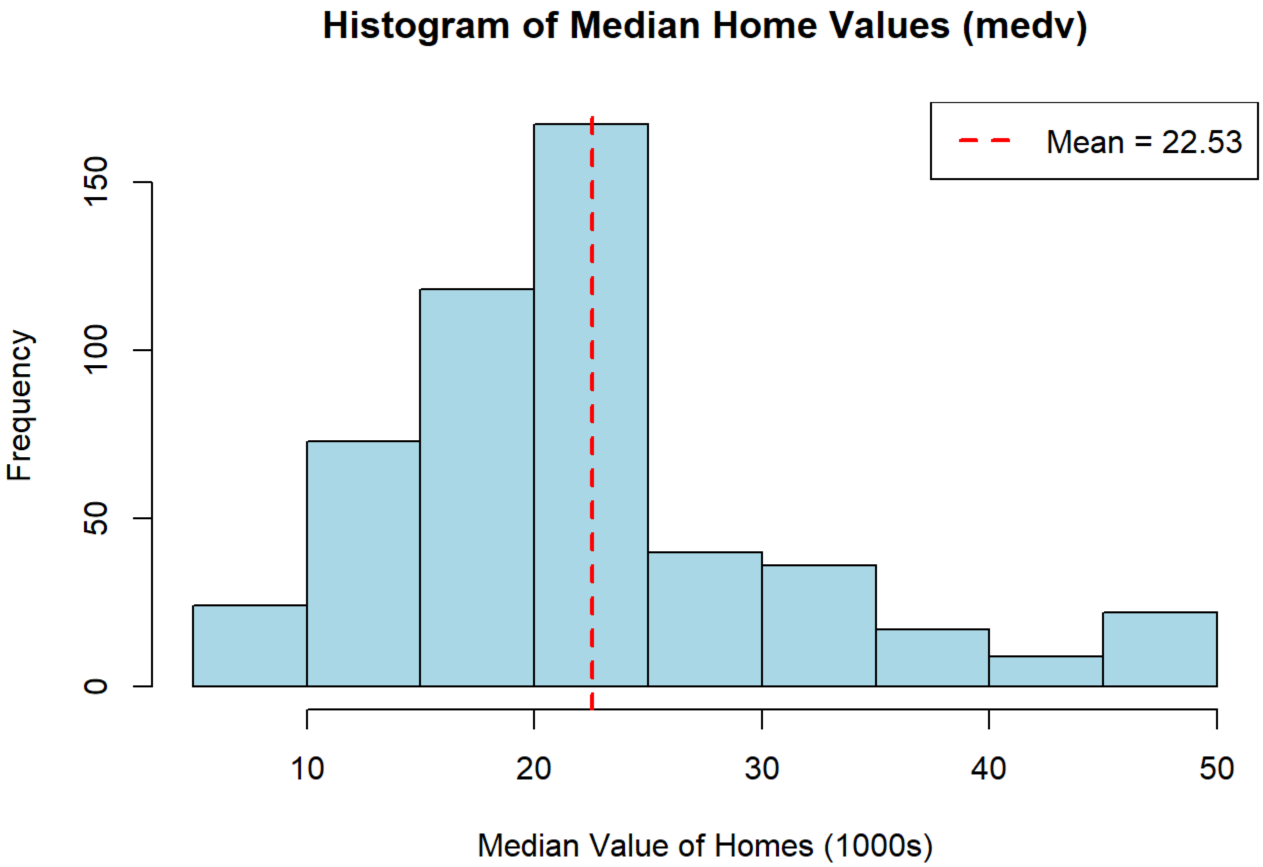
```
# 3. Show a summary of all variables
summary(data)
```

```
##       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          lstat      
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95  
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97  
##       medv      
##  Min.   : 5.00  
##  1st Qu.:17.02  
##  Median :21.20  
##  Mean   :22.53  
##  3rd Qu.:25.00  
##  Max.   :50.00
```

```
# 4. Calculate the mean value of the 'medv' variable

mean_medv <- mean(data$medv, na.rm = TRUE)
# Include na.rm = TRUE to handle missing values if any
print(paste("Mean of medv:", mean_medv))
```

```
## [1] "Mean of medv: 22.5328063241107"
```

```
# 5. Plot a histogram of the 'medv' variable and mark the mean value
hist(data$medv,
     main = "Histogram of Median Home Values (medv)",
     xlab = "Median Value of Homes (1000s)",
     col = "lightblue",
     border = "black")
abline(v = mean_medv, col = "red", lwd = 2, lty = 2) # Add vertical line for mean
legend("topright", legend = paste("Mean =", round(mean_medv, 2)),
       col = "red", lty = 2, lwd = 2)
```



```
# 6. Create a variable called cat.medv
data$cat.medv <- ifelse(data$medv > 30, 1, 0)

# 7. Calculate the mean of cat.medv
mean_cat_medv <- mean(data$cat.medv)
print(paste("Mean of cat.medv:", mean_cat_medv))
```

```
## [1] "Mean of cat.medv: 0.166007905138340"
```

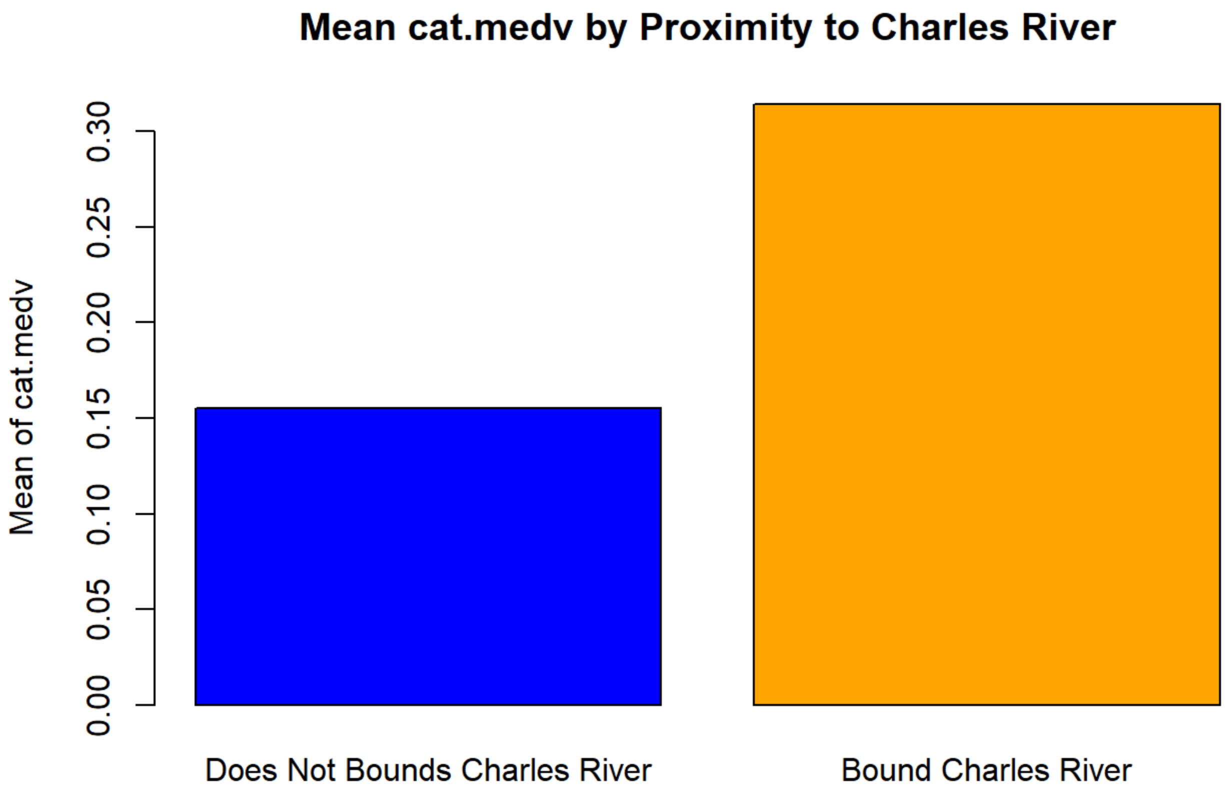
```
# 8. Calculate the mean of cat.medv for tracts that bound the Charles River (chas == 1)
mean_cat_medv_chas1 <- mean(data$cat.medv[data$chas == 1])
print(paste("Mean of cat.medv for tracts that bound Charles River:", mean_cat_medv_chas1))
```

```
## [1] "Mean of cat.medv for tracts that bound Charles River: 0.314285714285714"
```

```
# 9. Calculate the mean of cat.medv for tracts that don’t bound Charles River (chas == 0)
mean_cat_medv_chas0 <- mean(data$cat.medv[data$chas == 0])
print(paste("Mean of cat.medv for tracts that don't bound Charles River:", mean_cat_medv_chas0))
```

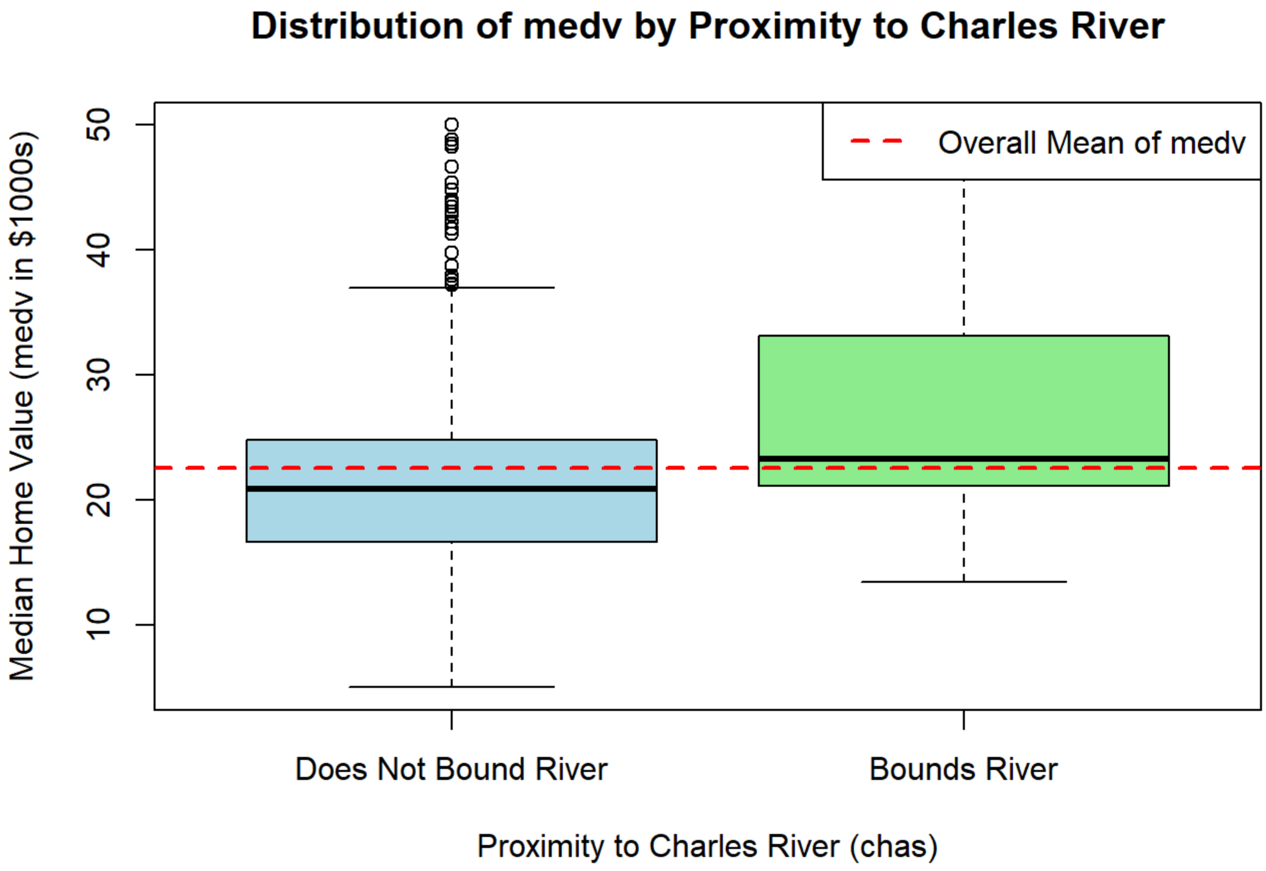
```
## [1] "Mean of cat.medv for tracts that don't bound Charles River: 0.154989384288747"
```

```
# 10. Create a vector of the two mean values
means_vector <- c( mean_cat_medv_chas0,mean_cat_medv_chas1)
names(means_vector) <- c("Does Not Bounds Charles River", "Bound Charles River")

# 11. Plot a bar chart
barplot(means_vector,
        main = "Mean cat.medv by Proximity to Charles River",
        ylab = "Mean of cat.medv",
        col = c("blue", "orange"),
        names.arg = c("Does Not Bounds Charles River", "Bound Charles River"))
```



Comparison: The height of the bars represents the proportion of tracts where the median home value exceeds $30,000 (cat.medv = 1) for tracts near the Charles River vs. those not near the river. Observation: If the bar for tracts that bound the Charles River is significantly taller, it suggests that homes near the river are generally more expensive. Conversely, a smaller or equal height indicates little to no premium for being near the river. Possible Implications: Proximity to the Charles River might influence housing prices due to factors such as aesthetics, desirability, or environmental benefits.:

```
# 13. Create a side-by-side boxplot of medv over chas
boxplot(medv ~ chas,
        data = data,
        main = "Distribution of medv by Proximity to Charles River",
        xlab = "Proximity to Charles River (chas)",
        ylab = "Median Home Value (medv in $1000s)",
        col = c("lightblue", "lightgreen"),
        names = c("Does Not Bound River", "Bounds River"))

# Add a horizontal line for the overall mean of medv for reference
abline(h = mean(data$medv, na.rm = TRUE), col = "red", lwd = 2, lty = 2)
legend("topright", legend = "Overall Mean of medv", col = "red", lty = 2, lwd = 2)
```



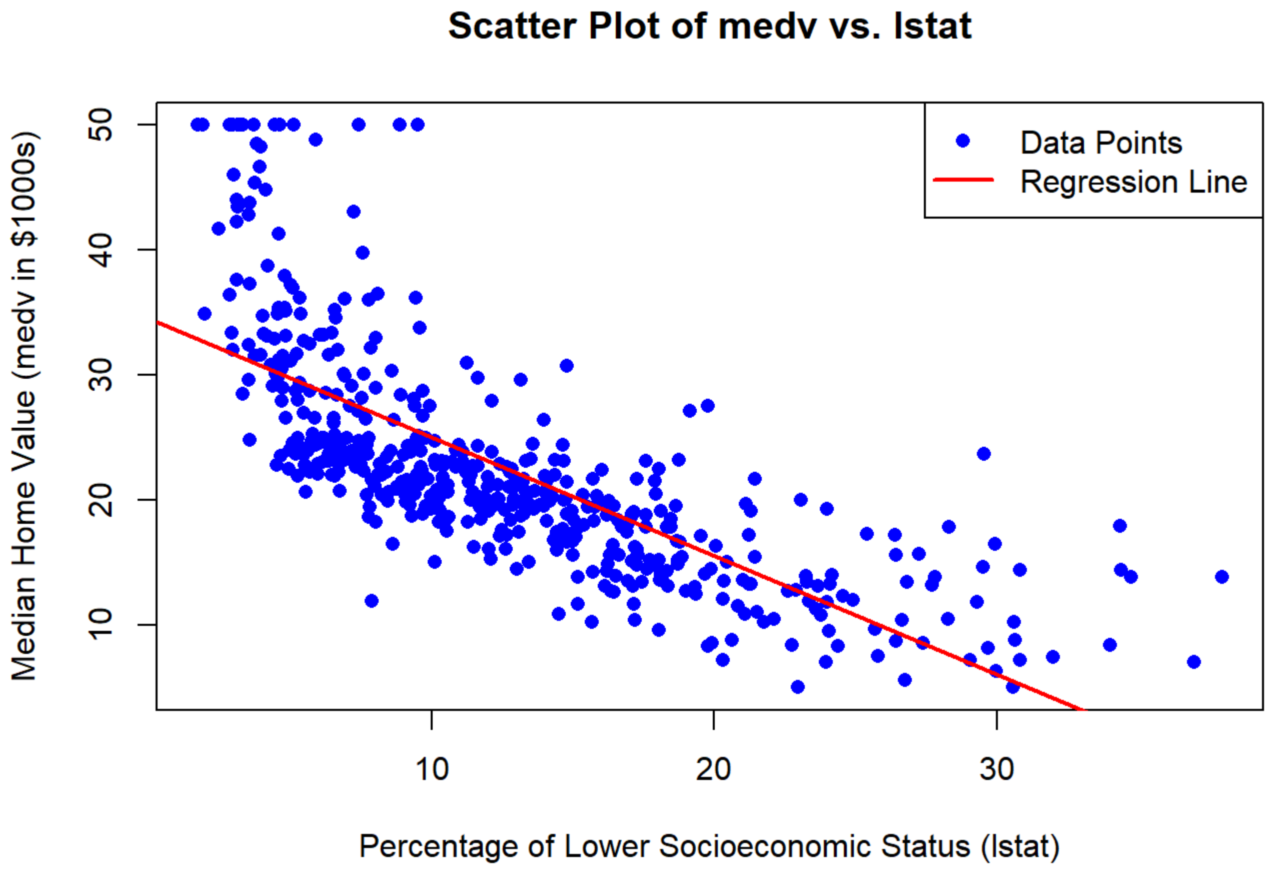
Including Plots

You can also embed plots, for example:

```
# 14. Create a scatter plot of medv (y-axis) versus lstat (x-axis)
plot(data$lstat, data$medv,
     main = "Scatter Plot of medv vs. lstat",
     xlab = "Percentage of Lower Socioeconomic Status (lstat)",
     ylab = "Median Home Value (medv in $1000s)",
     col = "blue",
     pch = 16)

# 15. Run a simple linear regression of medv on lstat
reg_model <- lm(medv ~ lstat, data = data)
summary(reg_model)
```

```
## 
## Call:
## lm(formula = medv ~ lstat, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -15.168  -3.990  -1.318   2.034  24.500 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432 
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
# 16. Add the regression line onto the scatter plot
abline(reg_model, col = "red", lwd = 2)

# Add a legend
legend("topright",
       legend = c("Data Points", "Regression Line"),
       col = c("blue", "red"),
       pch = c(16, NA),
       lty = c(NA, 1),
       lwd = c(NA, 2))
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.