Regression in R

```
housing <- read.csv("housing.csv")
head(housing)
```

```
##      CRIM ZN INDUS CHAS   NOX    RM  AGE    DIS RAD TAX PTRATIO      B LSTAT
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   MDEV
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

Examine the descriptive statistics (e.g., mean, median, sd) of the median value of owner-occupied homes, no. of rooms, and % of lower status of the population.

```
mean(housing$MDEV)
```

```
## [1] 22.53281
```

```
median(housing$MDEV)
```

```
## [1] 21.2
```

```
sd(housing$MDEV)
```

```
## [1] 9.197104
```

```
mean(housing$LSTAT)
```

```
## [1] 12.65306
```

```
median(housing$LSTAT)
```

```
## [1] 11.36
```

```r
sd(housing$LSTAT)
```

```
## [1] 7.141062
```

```r
mean(housing$RM)
```

```
## [1] 6.284634
```

```r
median(housing$RM)
```

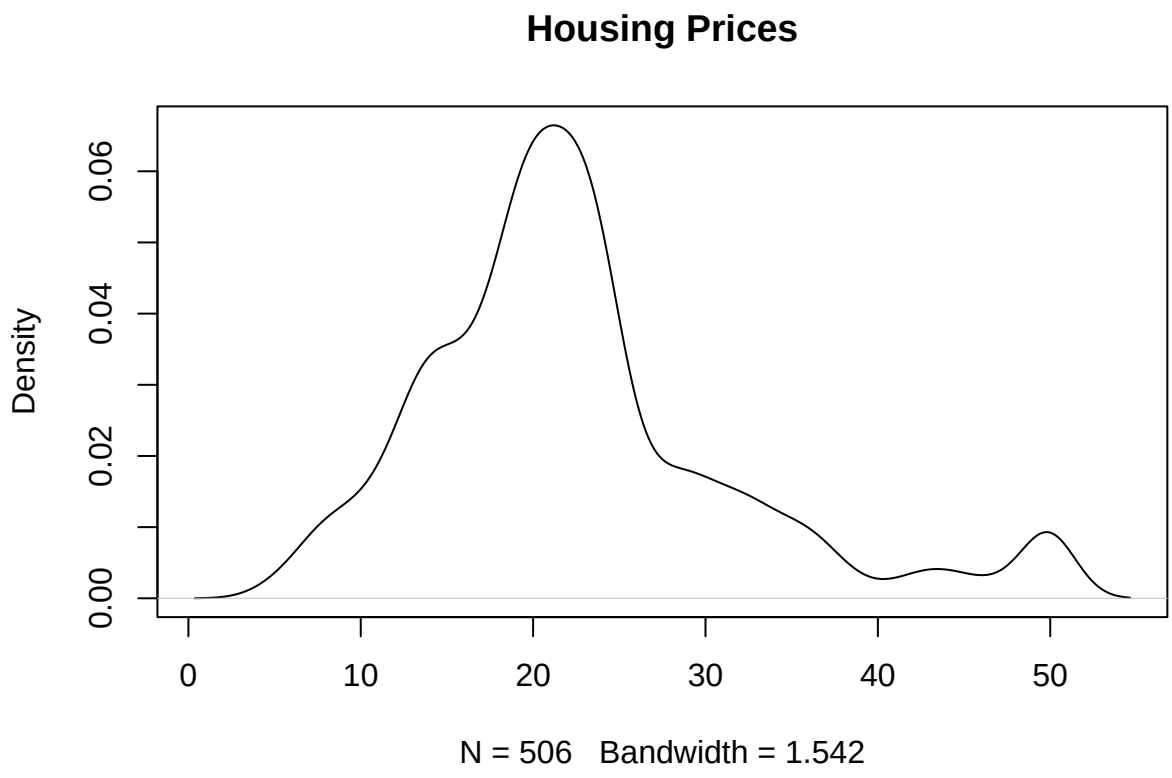
```
## [1] 6.2085
```

```r
sd(housing$RM)
```

```
## [1] 0.7026171
```

Create a plot of the median home value variable you used in (a) above. Can you create a density plot? (i.e., plot(density(...))).

```r
plot(density(housing$MDEV), main = "Housing Prices")
```



We want to compare home values in areas with greater (vs lesser) % of lower status of the population. Can you create a new variable/column in the dataset called low_status, which takes the value 1 if the value of

LSTAT > median(LSTAT)? Check how many rows belong to the "low_status" (1 vs 0) using table. What do you notice?

```
housing$low_status <- 1*(housing$LSTAT > median(housing$LSTAT))
table(housing$low_status)
```

```
##
##   0   1
## 253 253
```

Now conduct a ttest to compare the median home value for homes with low_status == 1 & homes with low_status == 0. Interpret the results.

```
t.test(housing$MDEV[which(housing$low_status==1)], housing$MDEV[which(housing$low_status==0)])
```

```
##
##  Welch Two Sample t-test
##
## data:  housing$MDEV[which(housing$low_status == 1)] and housing$MDEV[which(housing$low_status == 0)]
## t = -18.565, df = 391.19, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.95083 -10.47051
## sample estimates:
## mean of x mean of y
##  16.67747  28.38814
```

The results show that there is a significant difference between home prices in high vs low status areas.

Can you examine the correlations of median home value with other variables: average no. of rooms and % of lower status of the population.

```
cor(housing$MDEV, housing$RM)
```

```
## [1] 0.6953599
```

```
cor(housing$MDEV, housing$LSTAT)
```

```
## [1] -0.7376627
```

Can you create a list of these two variable names (i.e., average no. of rooms and % of lower status of the population) then write a loop over these two variables that prints the correlation of median home value with each of the two? HINT: Use the exact variable name in strings, e.g., c("RM") so you can use that column name inside your loop.

```
l <- c("RM", "LSTAT")
for (item in l){
  print(cor(housing$MDEV, housing[,(item)]))
}
```

```
## [1] 0.6953599
## [1] -0.7376627
```

Estimate a regression model for median home price as the Y variable using different sets of X variables.

a. Only LSTAT
b. Only RM (no. of rooms)
c. Both LSTAT and RM (no. of rooms)

```
summary(lm(MDEV ~ LSTAT, data = housing))
```

```
##
## Call:
## lm(formula = MDEV ~ LSTAT, data = housing)
##
## 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
```

```
summary(lm(MDEV ~ RM, data = housing))
```

```
##
## Call:
## lm(formula = MDEV ~ RM, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## RM             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
summary(lm(MDEV ~ LSTAT + RM, data = housing))
```

```
##
## Call:
## lm(formula = MDEV ~ LSTAT + RM, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.076  -3.516  -1.010   1.909  28.131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.35827    3.17283  -0.428    0.669
## LSTAT       -0.64236    0.04373 -14.689   <2e-16 ***
## RM           5.09479    0.44447  11.463   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.54 on 503 degrees of freedom
## Multiple R-squared:  0.6386, Adjusted R-squared:  0.6371
## F-statistic: 444.3 on 2 and 503 DF,  p-value: < 2.2e-16
```

Model with both variables has the highest R-squared. Both LSTAT and RM are significantly related with median prices, but LSTAT has a negative and RM has a positive relationship.

BONUS

Write a function that takes a dataframe name and two column ids as inputs and prints the results of the regression of the first column (Y) on the second column (X). Invoke this function with (your_data_name, 14, 13) as inputs. What do you get? Check if the results are equivalent to the results in 3a(a).

```r
reg_func <- function(data, y, x){
  print(summary(lm(data[,y] ~ data[,x])))
}

reg_func(housing, 14, 13)
```

```
##
## Call:
## lm(formula = data[, y] ~ data[, x])
##
## 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 ***
## data[, x]   -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
```