HUDM5123_Lab01_LinearRegressionInR

Zhuqian Karen Zhou

February 6, 2020

Task 1: Examine Data

```
dim(mtcars)
```

```
## [1] 32 11
```

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

According to the functions applied above, there are 32 rows and 11 columns in the dataset. The 11 variables are 1) mpq (Mile/US gallon), 2) cyl (Number of cylinders), 3) disp (Displacement (cu.in.)), 4) hp (Gross horsepower), 5) drat (Rear axle ratio), 6) wt (Weight (1000 lbs)), 7) qsec (1/4 mile time), 8) vs (Engine: 0=V-shaped,1=straight), 9) am (Transmission: 0=automatic, 1=manual), 10) gear (Number of forward gears), and 11) carb (Number of carburetors).

```
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
tail(mtcars)
```

```
##                 mpg cyl  disp  hp drat    wt qsec vs am gear carb
## Porsche 914-2  26.0   4 120.3  91 4.43 2.140 16.7  0  1    5    2
## Lotus Europa   30.4   4  95.1 113 3.77 1.513 16.9  1  1    5    2
## Ford Pantera L 15.8   8 351.0 264 4.22 3.170 14.5  0  1    5    4
## Ferrari Dino   19.7   6 145.0 175 3.62 2.770 15.5  0  1    5    6
## Maserati Bora  15.0   8 301.0 335 3.54 3.570 14.6  0  1    5    8
## Volvo 142E     21.4   4 121.0 109 4.11 2.780 18.6  1  1    4    2
```

The performance and other indices of the first six brands (i.e. Mazda RX4, Mazda RX4 Wag, Datsun 710, Hornet 4 Drive, Hornet Sportabout, and Valinant) and the last six brands (i.e. Porsche 914-2, Lotus Europa, Ford Pantera L, Ferrari Dino, Maserati Bora, and Volvo 142E) in the dataset are shown above.

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

According to the structure of the dataset presented above, all variables are numeric.

```
round(var(mtcars), digits = 3)
```

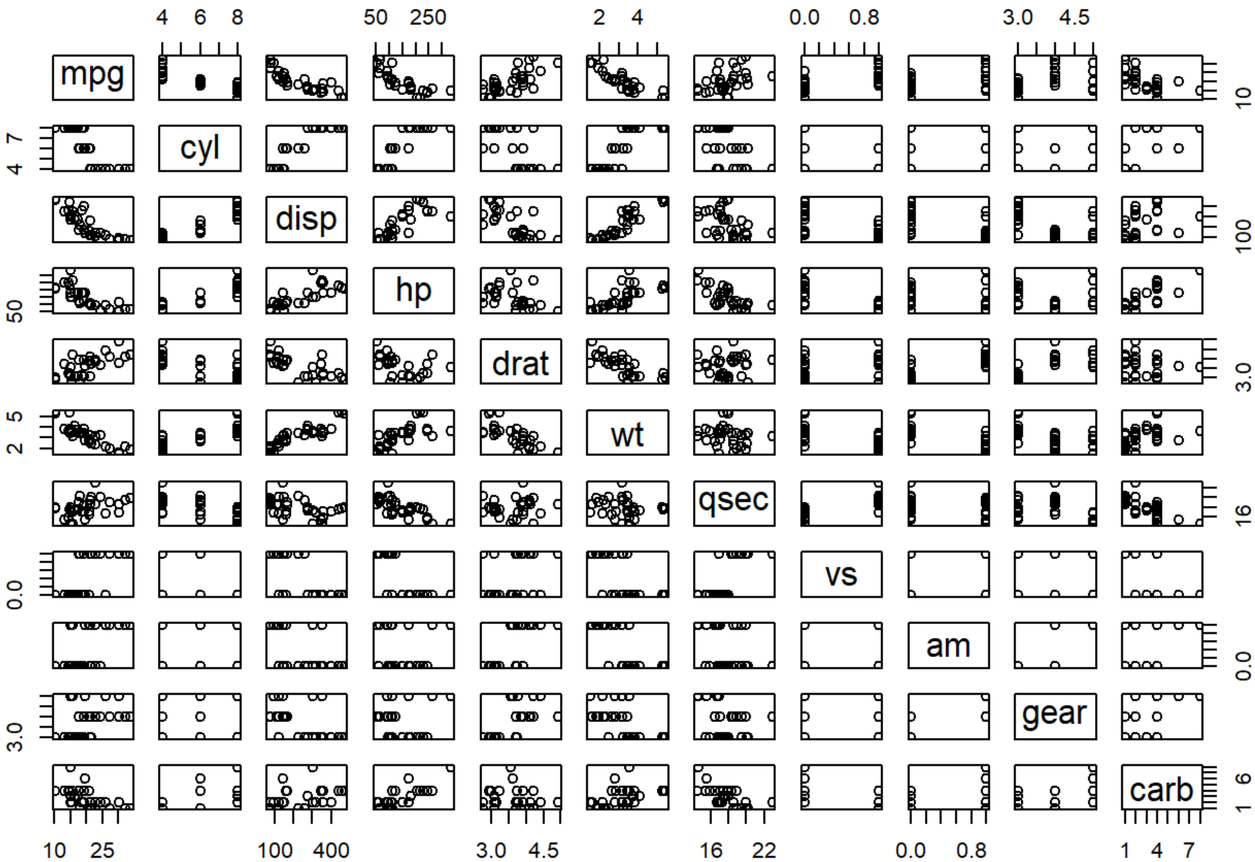
```
##           mpg     cyl      disp       hp    drat      wt    qsec      vs
## mpg    36.324  -9.172  -633.097 -320.732   2.195  -5.117   4.509   2.017
## cyl    -9.172   3.190   199.660  101.931  -0.668   1.367  -1.887  -0.730
## disp -633.097 199.660 15360.800 6721.159 -47.064 107.684 -96.052 -44.378
## hp   -320.732 101.931  6721.159 4700.867 -16.451  44.193 -86.770 -24.988
## drat    2.195  -0.668   -47.064  -16.451   0.286  -0.373   0.087   0.119
## wt     -5.117   1.367   107.684   44.193  -0.373   0.957  -0.305  -0.274
## qsec    4.509  -1.887   -96.052  -86.770   0.087  -0.305   3.193   0.671
## vs      2.017  -0.730   -44.378  -24.988   0.119  -0.274   0.671   0.254
## am      1.804  -0.466   -36.564   -8.321   0.190  -0.338  -0.205   0.042
## gear    2.136  -0.649   -50.803   -6.359   0.276  -0.421  -0.280   0.077
## carb   -5.363   1.520    79.069   83.036  -0.078   0.676  -1.894  -0.464
##           am    gear   carb
## mpg    1.804   2.136 -5.363
## cyl   -0.466  -0.649  1.520
## disp -36.564 -50.803 79.069
## hp    -8.321  -6.359 83.036
## drat   0.190   0.276 -0.078
## wt    -0.338  -0.421  0.676
## qsec  -0.205  -0.280 -1.894
## vs     0.042   0.077 -0.464
## am     0.249   0.292  0.046
## gear   0.292   0.544  0.327
## carb   0.046   0.327  2.609
```

```
round(cor(mtcars), digits = 3)
```

```
##         mpg    cyl   disp     hp   drat     wt   qsec     vs     am   gear
## mpg   1.000 -0.852 -0.848 -0.776  0.681 -0.868  0.419  0.664  0.600  0.480
## cyl  -0.852  1.000  0.902  0.832 -0.700  0.782 -0.591 -0.811 -0.523 -0.493
## disp -0.848  0.902  1.000  0.791 -0.710  0.888 -0.434 -0.710 -0.591 -0.556
## hp   -0.776  0.832  0.791  1.000 -0.449  0.659 -0.708 -0.723 -0.243 -0.126
## drat  0.681 -0.700 -0.710 -0.449  1.000 -0.712  0.091  0.440  0.713  0.700
## wt   -0.868  0.782  0.888  0.659 -0.712  1.000 -0.175 -0.555 -0.692 -0.583
## qsec  0.419 -0.591 -0.434 -0.708  0.091 -0.175  1.000  0.745 -0.230 -0.213
## vs    0.664 -0.811 -0.710 -0.723  0.440 -0.555  0.745  1.000  0.168  0.206
## am    0.600 -0.523 -0.591 -0.243  0.713 -0.692 -0.230  0.168  1.000  0.794
## gear  0.480 -0.493 -0.556 -0.126  0.700 -0.583 -0.213  0.206  0.794  1.000
## carb -0.551  0.527  0.395  0.750 -0.091  0.428 -0.656 -0.570  0.058  0.274
##        carb
## mpg  -0.551
## cyl   0.527
## disp  0.395
## hp    0.750
## drat -0.091
## wt    0.428
## qsec -0.656
## vs   -0.570
## am    0.058
## gear  0.274
## carb  1.000
```

The variance/covariance matrix and the correlation matrix are also displayed above.

```
plot(mtcars)
```



We can see bivariate scatterplots of all combinations of two variables above.

At a glance, there seems to be *positive* linear relationships 1) between number of cylinders and displacement, 2) between displacement and gross hoursepower, 3) between miles per gallon and 1/4 mile time, 4) between displacement and weight, 5) between hoursepower and weight, and 6) between hoursepower and number of carburetors.

Also, there seems to be *negative* linear relationships 1) between miles per gallon and numbers of cylinders, 2) between miles per gallon and displacement, 3) between miles per gallon and gross hoursepower, 4) between miles per gallon and weight, 5) between number of cylinders and number of rear axle ratio, 6) between displacement and rear axle ratio, 7) between gross hoursepower and 1/4 mile time, and 8) between rear axle ratio and weight.

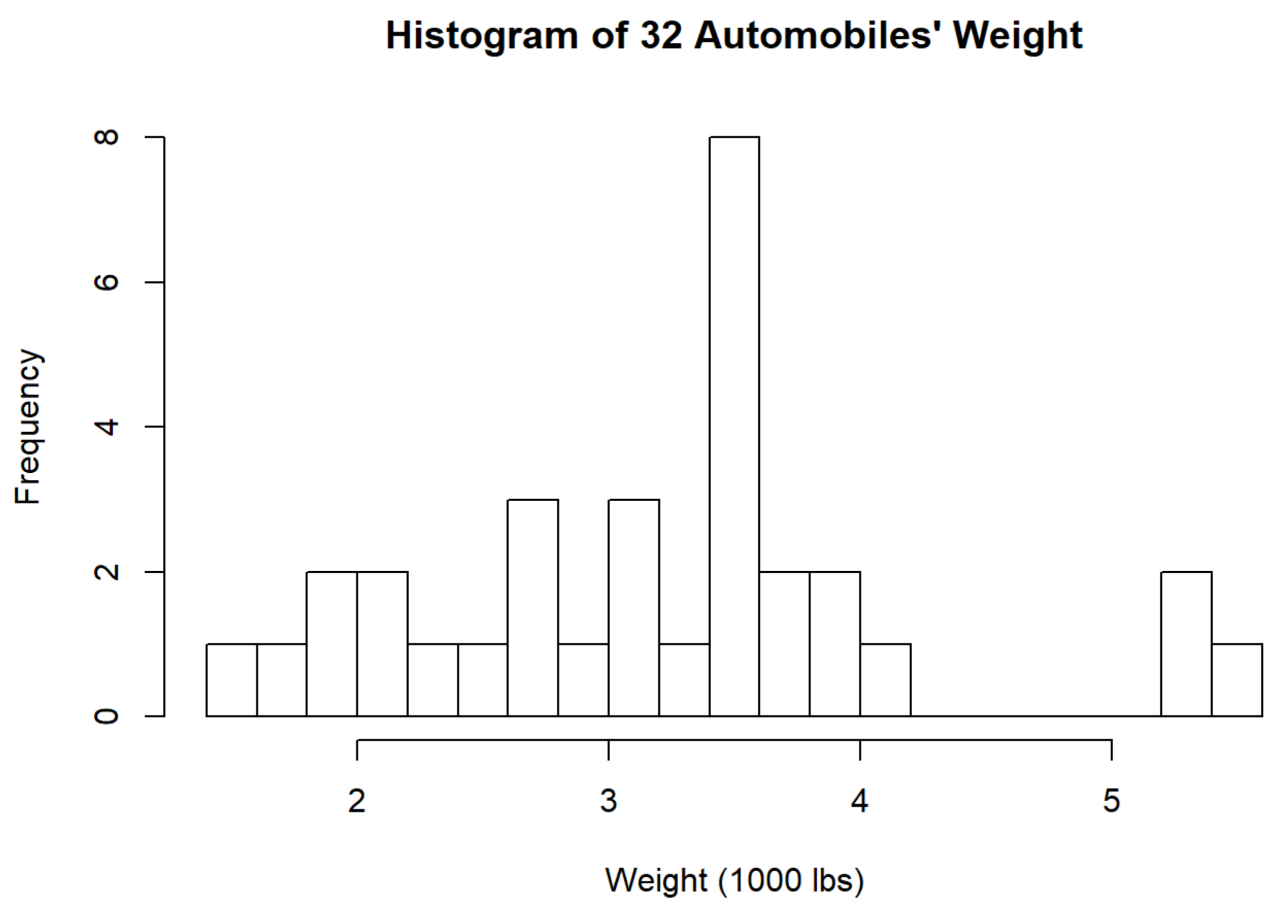
Task 2: Create a new variable *am_f*

```
df <- mtcars
df$am_f <- factor(x=df$am, levels = c(0,1), labels = c("automatic", "manual"))
df[,c(9,12)] # print out am and am_f
```

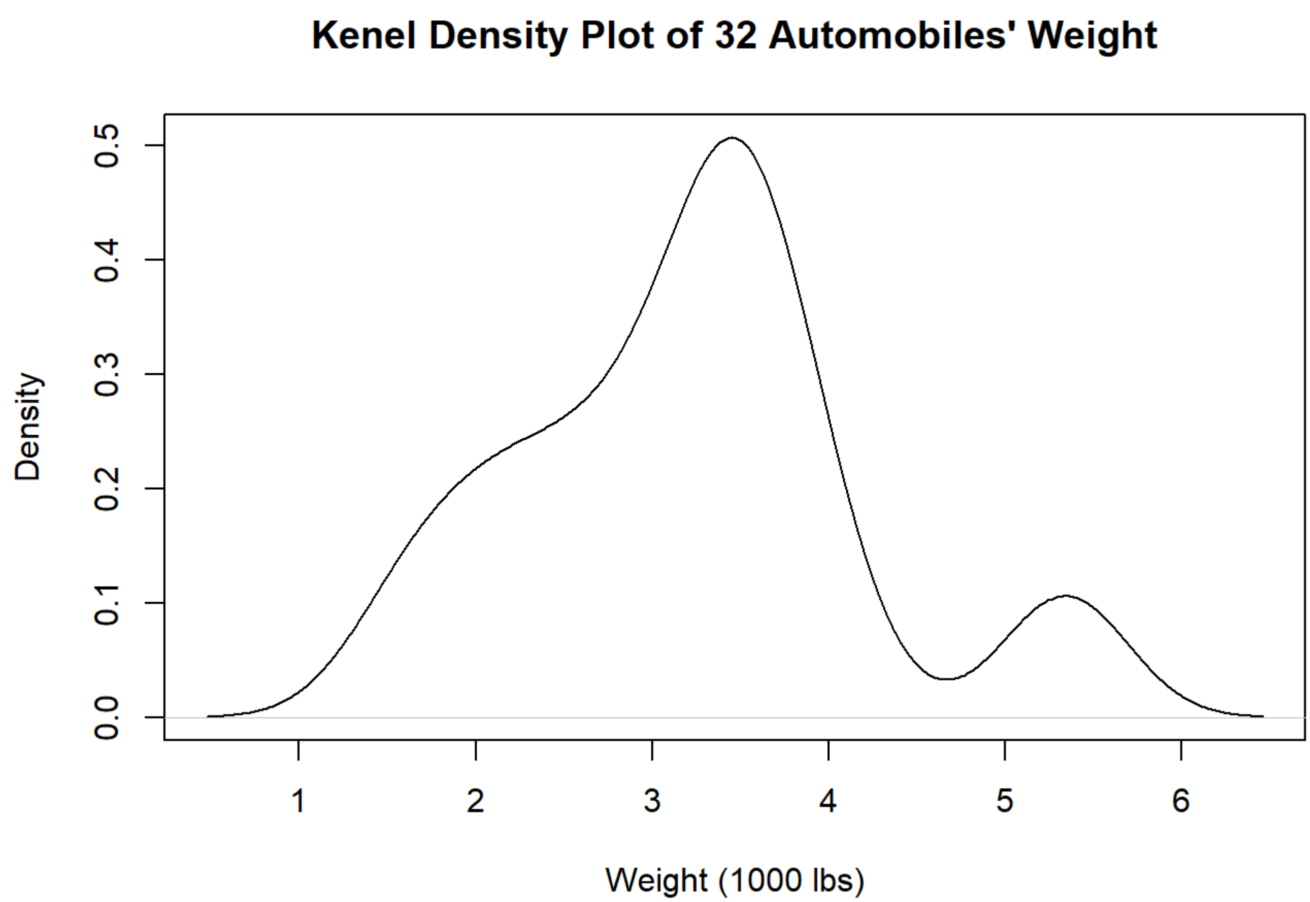
```
##                      am      am_f
## Mazda RX4             1    manual
## Mazda RX4 Wag         1    manual
## Datsun 710            1    manual
## Hornet 4 Drive        0 automatic
## Hornet Sportabout     0 automatic
## Valiant               0 automatic
## Duster 360            0 automatic
## Merc 240D             0 automatic
## Merc 230              0 automatic
## Merc 280              0 automatic
## Merc 280C             0 automatic
## Merc 450SE            0 automatic
## Merc 450SL            0 automatic
## Merc 450SLC           0 automatic
## Cadillac Fleetwood    0 automatic
## Lincoln Continental   0 automatic
## Chrysler Imperial     0 automatic
## Fiat 128              1    manual
## Honda Civic           1    manual
## Toyota Corolla        1    manual
## Toyota Corona         0 automatic
## Dodge Challenger      0 automatic
## AMC Javelin           0 automatic
## Camaro Z28            0 automatic
## Pontiac Firebird      0 automatic
## Fiat X1-9             1    manual
## Porsche 914-2         1    manual
## Lotus Europa          1    manual
## Ford Pantera L        1    manual
## Ferrari Dino          1    manual
## Maserati Bora         1    manual
## Volvo 142E            1    manual
```

Task 3: Graphical Exploration

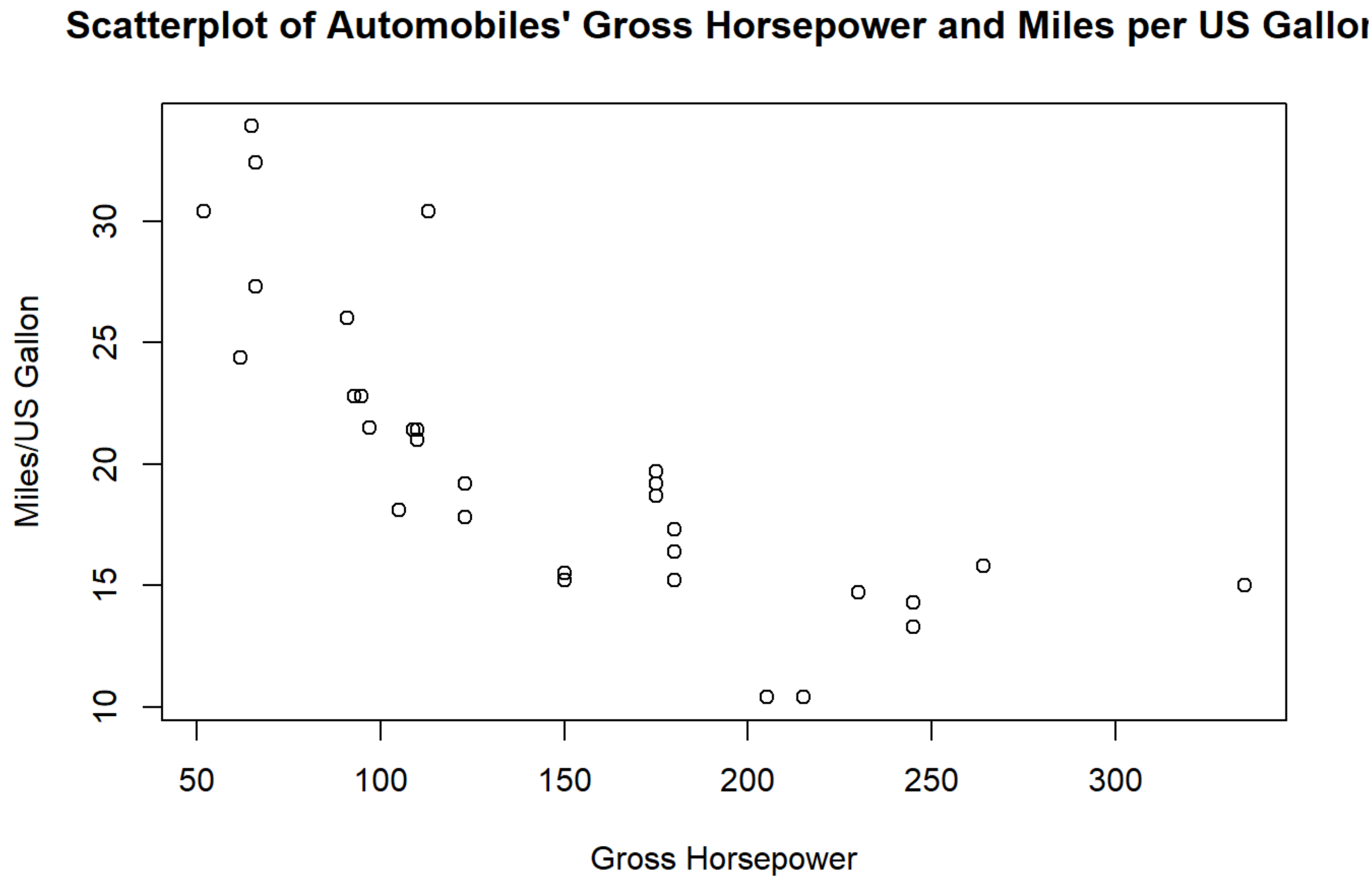
```
hist(df$wt, breaks = 20, xlab = c("Weight (1000 lbs)"), ylab = c("Frequency"), main = c("Hist
ogram of 32 Automobiles' Weight"))
```




```
plot(density(df$wt), xlab = c("Weight (1000 lbs)"), ylab = c("Density"), main = c("Kenel Dens
ity Plot of 32 Automobiles' Weight"))
```




```
plot(x = df$hp, y = df$mpg, xlab = c("Gross Horsepower"), ylab = c("Miles/US Gallon"), main =
c("Scatterplot of Automobiles' Gross Horsepower and Miles per US Gallon"))
```

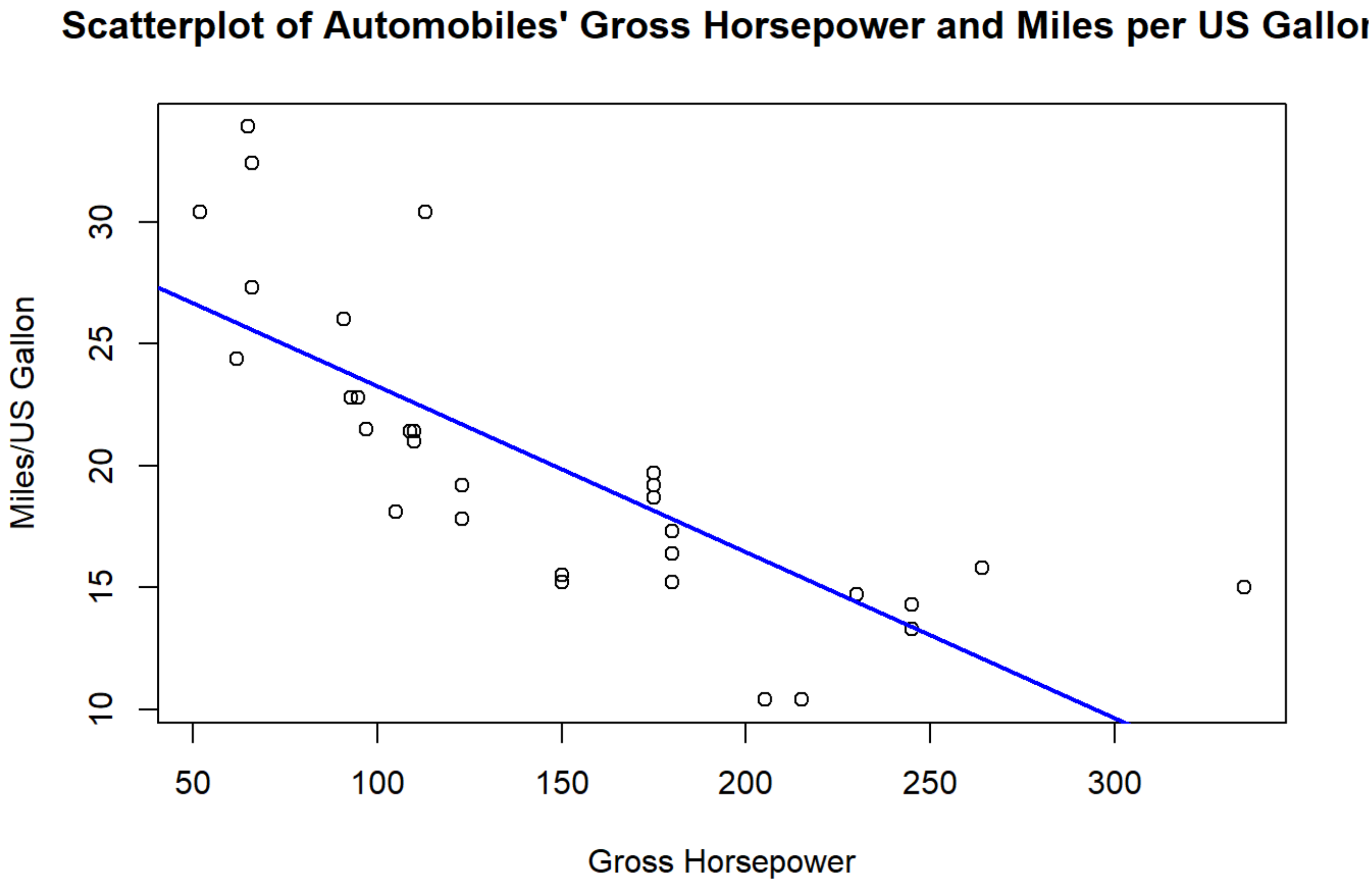


Task 4: Simple Linear Regression (1 IV)

```
lm1 <- lm(formula = mpg ~ hp, data = df)
summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ hp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

```
#Plot the Regression Line
plot(x = df$hp, y = df$mpg, xlab = c("Gross Horsepower"), ylab = c("Miles/US Gallon"), main =
c("Scatterplot of Automobiles' Gross Horsepower and Miles per US Gallon"))
abline(a = 30.09886,
       b = -0.06823,
       lwd = 2,
       col = "blue")
```



The estimated intercept, 30.09886, means that when an automobile's gross horsepower approaches zero, the milage that this automobile can run per US gallon is estimated to be around 30. However, it is not genuinely meaningful since a car can hardly has zero horsepower.

The estimated slope, -0.06823, means that in average, when gross horsepower of one automobile is one unit more than the other automobile, the distance it can run per US gallon tends to be around 0.068 miles less than the other one.

The R-square, 0.6024, means that the linear model, mpg=30.09886-0.06823*hp, can explain around 60.24% of the related variability between miles per US gallon and gross horsepower of the 32 automobiles in the dataset.

Task 5: Multiple Linear Regression (2 IVs)

```
lm2 <- lm(formula = mpg ~ hp + wt, data = df)
summary(lm2)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

```
#Plot the 3D scatterplot
library(rgl)
```

```
## Warning: package 'rgl' was built under R version 3.5.3
```

```
open3d()
```

```
## wgl
##   1
```

```
plot3d(df[,c("mpg", "hp", "wt")], col = "red", size = 2)

# Plot the regression plane
planes3d(a=-1, b=coef(lm2)[2], c=coef(lm2)[3],
         d=coef(lm2)[1], alpha=.5, col = "pink")
```

The estimated intercept, 37.22727, means that when both an automobile’s gross horsepower and its weight approaches zero, the milage that this automobile can run per US gallon is estimated to be around 37.2. However, it is not genuinely meaningful since a car can hardly weigh zero and has zero horsepower.

The estimated slope of hp, -0.03177, means that holding the weight of cars constant, when gross horsepower of one automobile is one unit more than the other automobile, the distance it can run per US gallon tends to be around 0.032 miles less than the other one.

The estimated slope of wt, -3.87783, means that holding the gross horsepower of cars constant, when the weight of one automobile is one lbs more than the other automobile, the distance it can run per US gallon tends to be around 3.878 miles less than the other one.

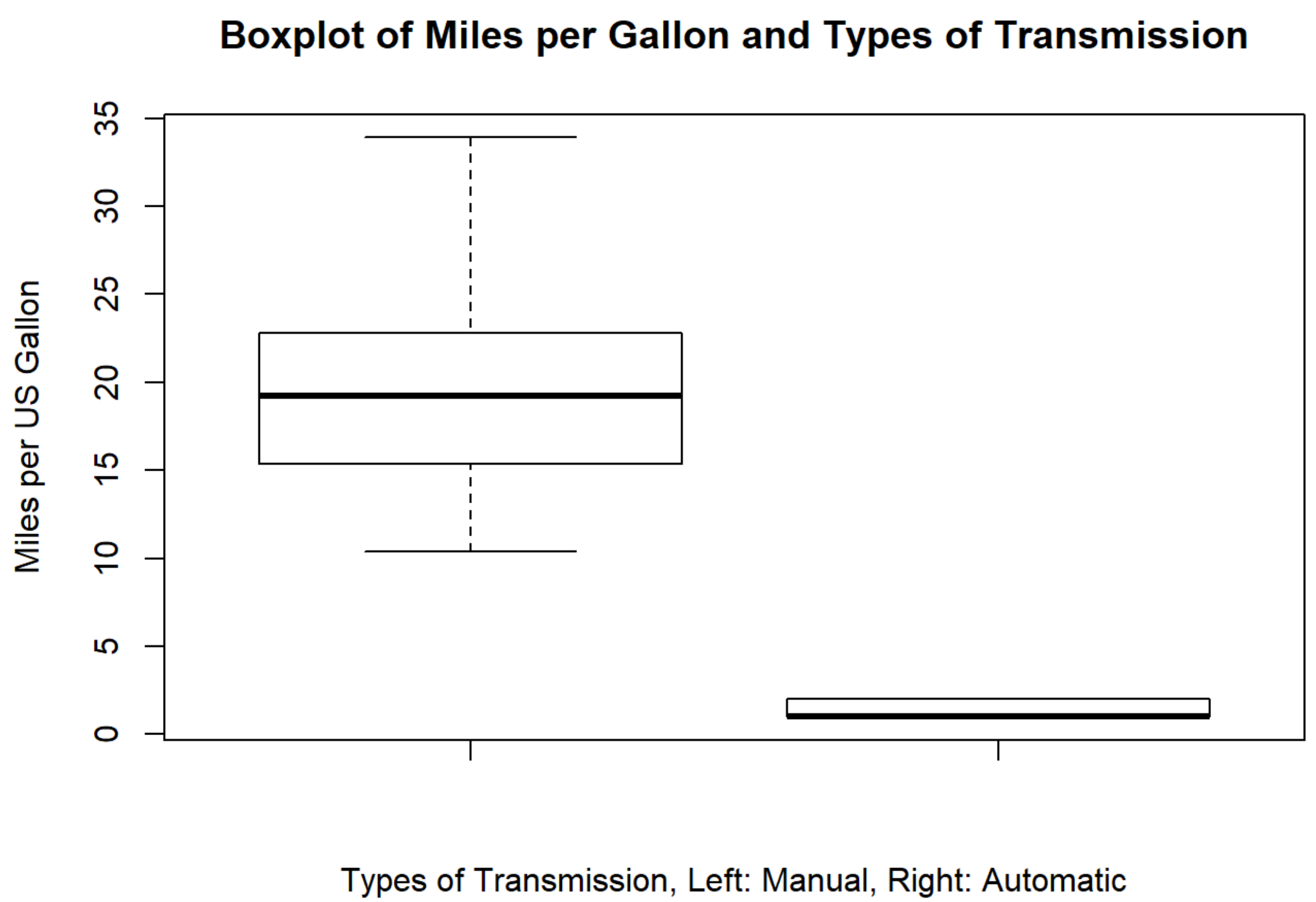
The R-square, 0.8268, means that the linear model, mpg=37.22727-0.03177*hp-3.87783wt, can explain around 82.68% of the related variability among miles per US gallon, gross horsepower, and weight of the 32 automobiles in the dataset.

Task 6: Linear Regression with a factor variable

```
lm3 <- lm(formula = mpg ~ am_f, data = df)
summary(lm3)
```

```
##
## Call:
## lm(formula = mpg ~ am_f, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am_fmanual     7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
boxplot(df$mpg, df$am_f, xlab = c("Types of Transmission, Left: Manual, Right: Automatic"), y
lab=c("Miles per US Gallon"), main = c("Boxplot of Miles per Gallon and Types of Transmissio
n"))
```

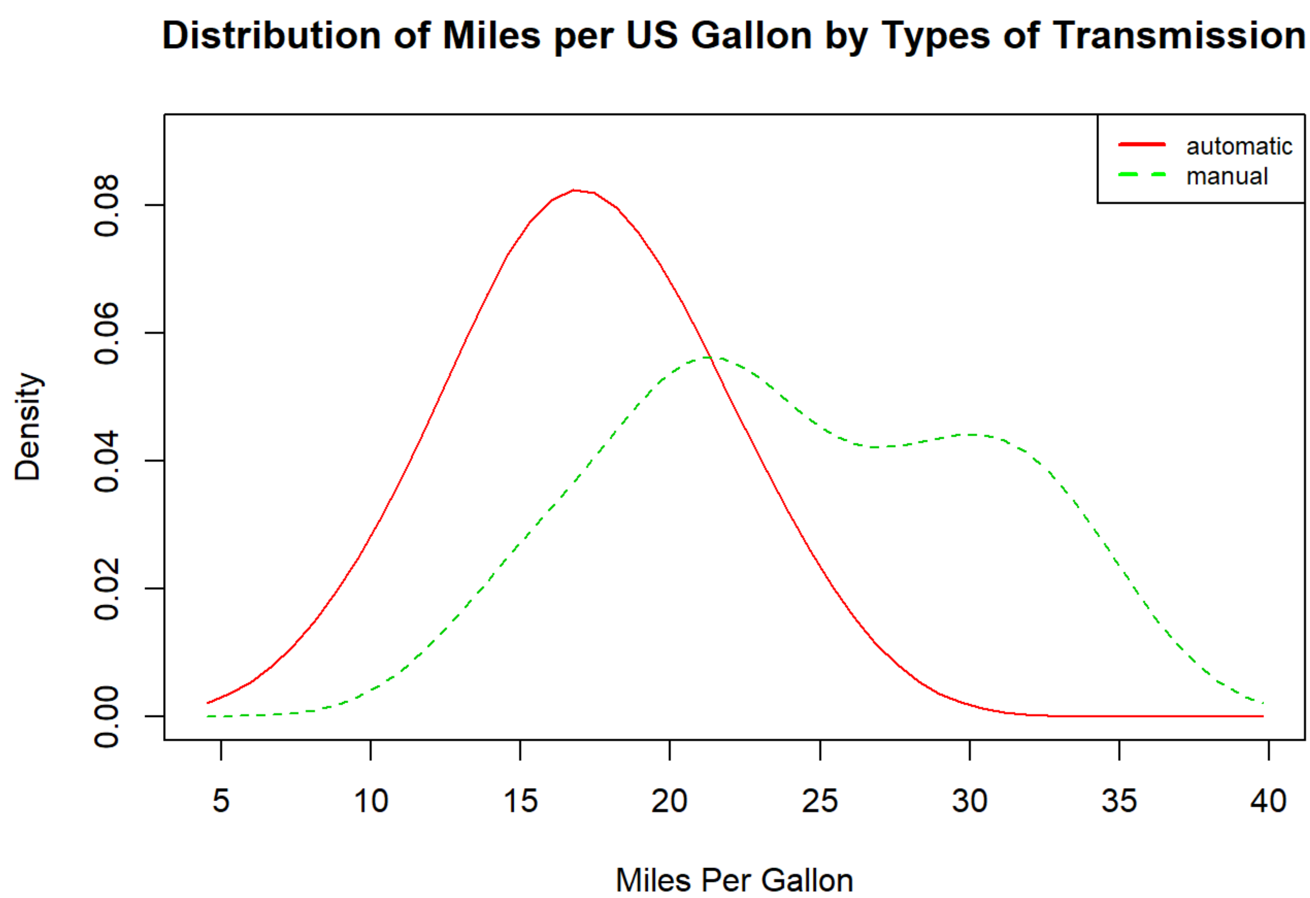



```r
# Density plots
library(sm)
```

```
## Warning: package 'sm' was built under R version 3.5.3
```

```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

```r
sm.density.compare(df$mpg, df$am_f, xlab="Miles Per Gallon")
title(main="Distribution of Miles per US Gallon by Types of Transmission")
legend(x = "topright",
       col = c("red", "green"),
       lty=c("solid", "dashed"),
       legend = c("automatic", "manual"),
       lwd = 2,
       cex = 0.75)
```



The estimated intercept, 17.147, means that the average distance that automatic transmissions can run per US gallon is around 17.147 miles.

The estimated slope, 7.245, means that the average distance that manual transmissions can run is around 7.245 miles more than that of automatic transmissions.

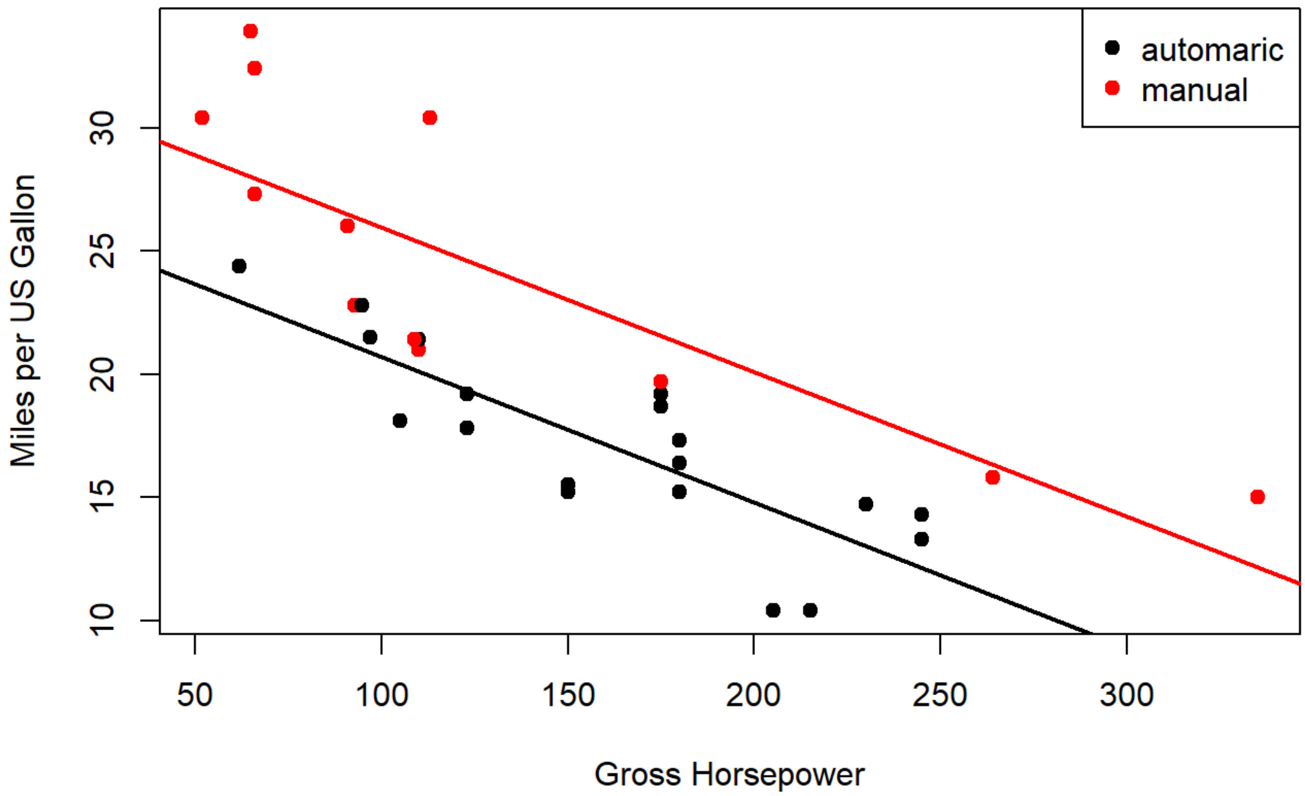
The R-square, 0.3598, means that the linear model, mpg=17.147+7.245*am, can explain around 35.98% of the related variability between miles per US gallon and types of transmission of the 32 automobiles in the dataset.

Task 7: Multiple Linear Regression with interaction

```
lm4<-lm(formula = mpg ~ hp*am_f, data = df)
summary(lm4)
```

```
##
## Call:
## lm(formula = mpg ~ hp * am_f, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3818 -2.2696  0.1344  1.7058  5.8752
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   26.6248479  2.1829432  12.197 1.01e-12 ***
## hp            -0.0591370  0.0129449  -4.568 9.02e-05 ***
## am_fmanual     5.2176534  2.6650931   1.958   0.0603 .
## hp:am_fmanual  0.0004029  0.0164602   0.024   0.9806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 28 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.7587
## F-statistic: 33.49 on 3 and 28 DF,  p-value: 2.112e-09
```

```
plot(x = df$hp,
     y = df$mpg,
     col = df$am + 1,
     pch = 19,
     xlab = "Gross Horsepower",
     ylab = "Miles per US Gallon")
legend(x = "topright",
       pch = 19, col = 1:2,
       legend = c("automaric", "manual"))
abline(a = 26.6248479, b = -0.0591370,
       col = 1, lwd = 2)
abline(a = (26.6248479 + 5.2176534),
       b = (-0.0591370 + 0.0004029),
       col = 2, lwd = 2)
```



The estimated intercepts of the regression line of miles per US gallon on gross horsepower for automatic transmissions, 26.6248479, is smaller than that for manual transmissions, 26.6248479 + 5.2176534.

The estimated slope of the regression line of miles per US gallon on gross horsepower for automatic transmissions, -0.0591370, is slightly smaller than that for manual transmissions, -0.0591370 + 0.0004029, and both are negative.

The estimated intercepts and slopes indicate that for automatic transmissions the higher the gross horsepower the lower the milage per gallan a car can run, which is also true for the manual transmissions, but they have overall higher milage per gallon compared to automatic transmissions.

The R-square, 0.782, means that the linear model, mpg=26.62-0.059hp when cars are automatic transmissions & mpg=31.84-0.059hp when cars are manual transmissions, can explain around 78.2% of the related variability among miles per US gallon, gross horsepower, and types of transmission of the 32 automobiles in the dataset.