Overview of “Auto” Dataset from “ISLR” Package

Julian Hatwell

February 29, 2016

This document provides a brief overview of the Auto dataset in the ISLR R package.

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##
##      weight     acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##                   name
##  amc matador       :  5
##  ford pinto        :  5
##  toyota corolla    :  5
##  amc gremlin       :  4
##  amc hornet        :  4
##  chevrolet chevette:  4
##  (Other)           :365
```

From the summary, and the associated help (not shown), the following observations can be made:

The dataframe contains 392 rows and 9 columns.

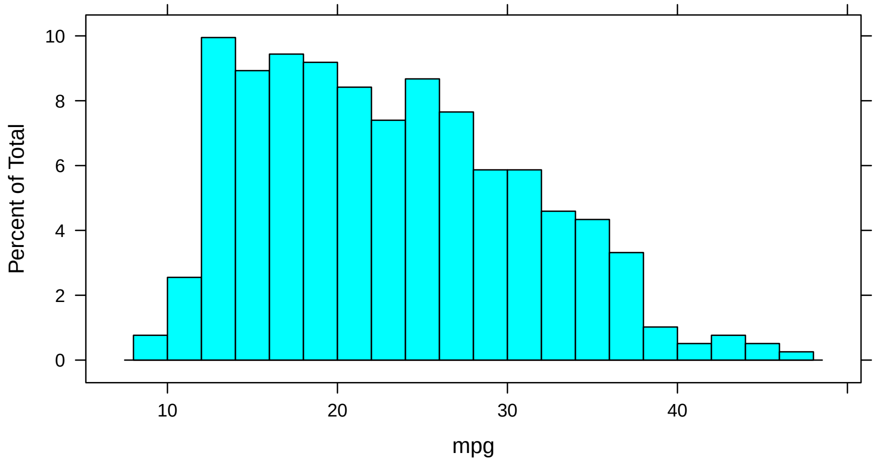


Figure 1: Histogram of the mpg variable

```
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2413  -3.1832  -0.6332   2.5491  17.9168
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.9155     0.8349   51.40   <2e-16 ***
## cylinders    -3.5581     0.1457  -24.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.914 on 390 degrees of freedom
## Multiple R-squared:  0.6047,Adjusted R-squared:  0.6037
## F-statistic: 596.6 on 1 and 390 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

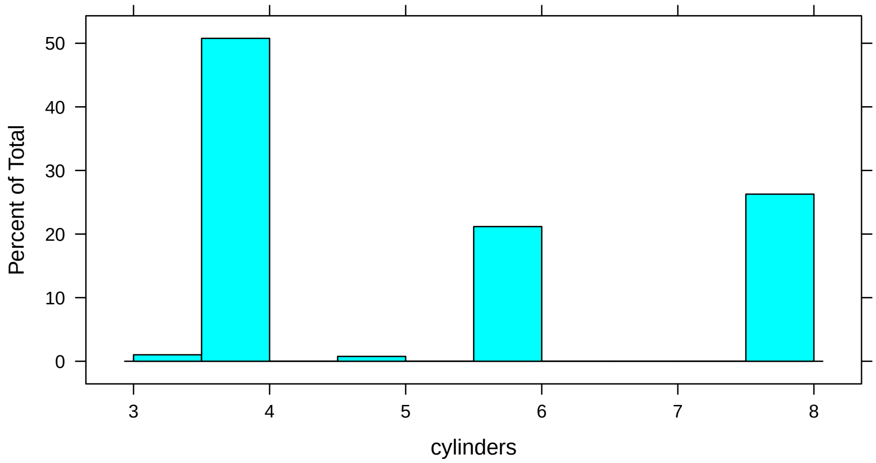


Figure 2: Histogram of the cylinders variable

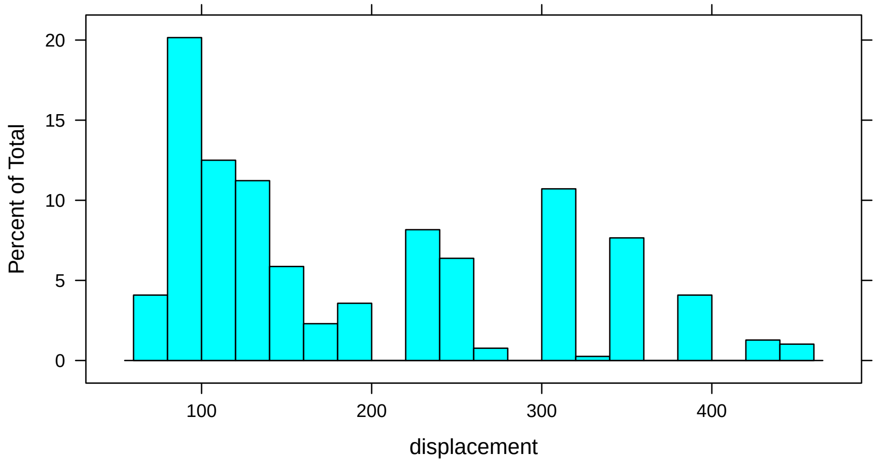


Figure 3: Histogram of the displacement variable

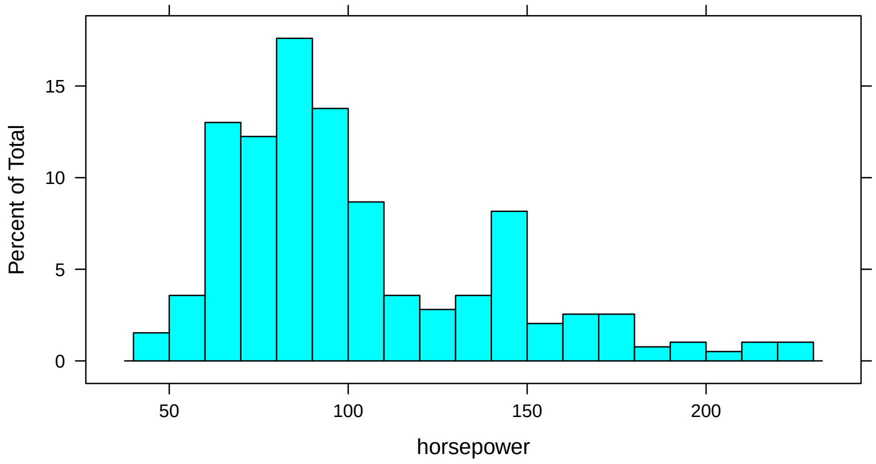


Figure 4: Histogram of the horsepower variable

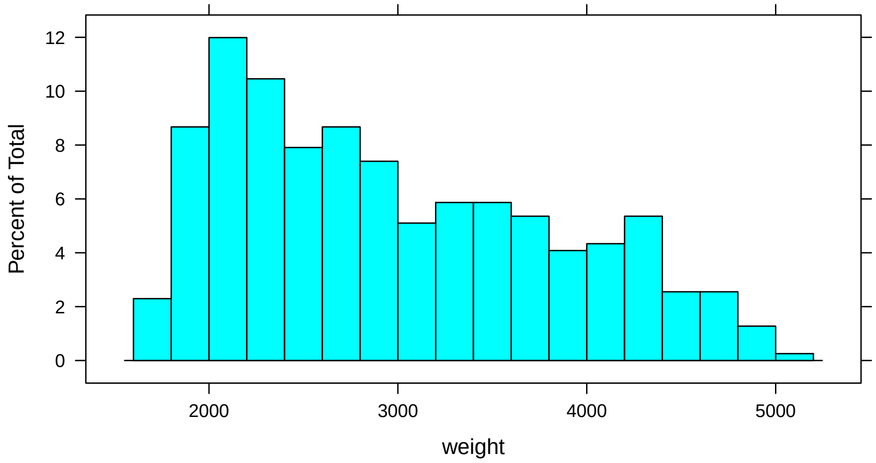


Figure 5: Histogram of the weight variable

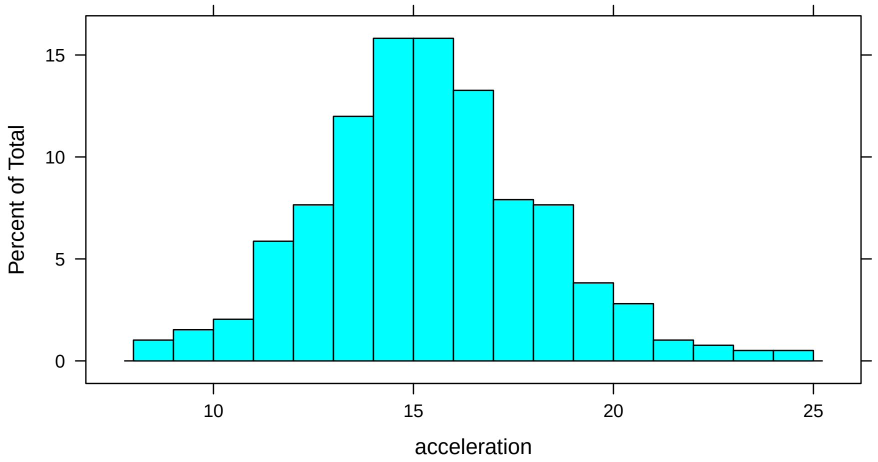


Figure 6: Histogram of the acceleration variable

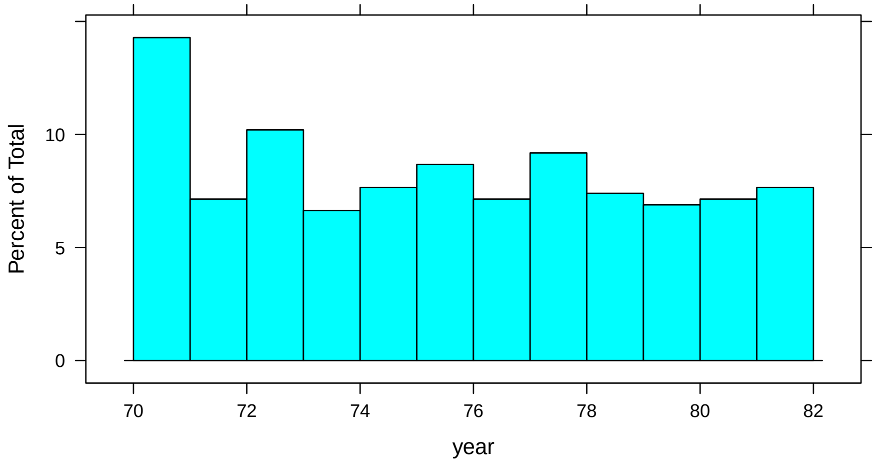


Figure 7: Histogram of the year variable

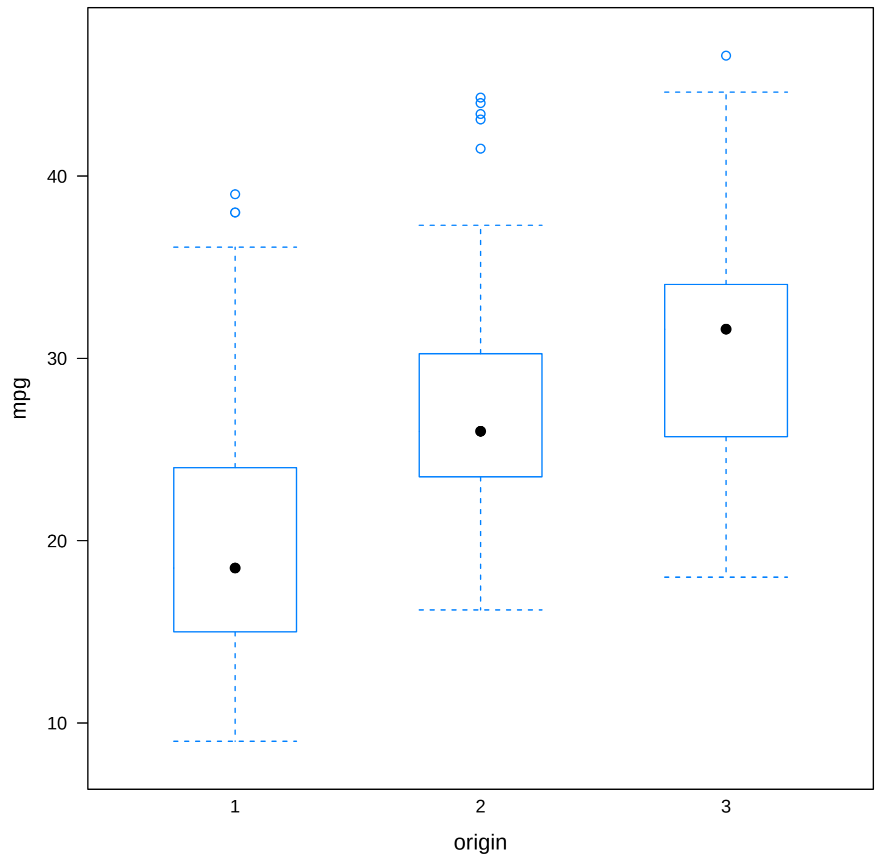


Figure 8: Boxplot of the dependent variable mpg by each factor variable

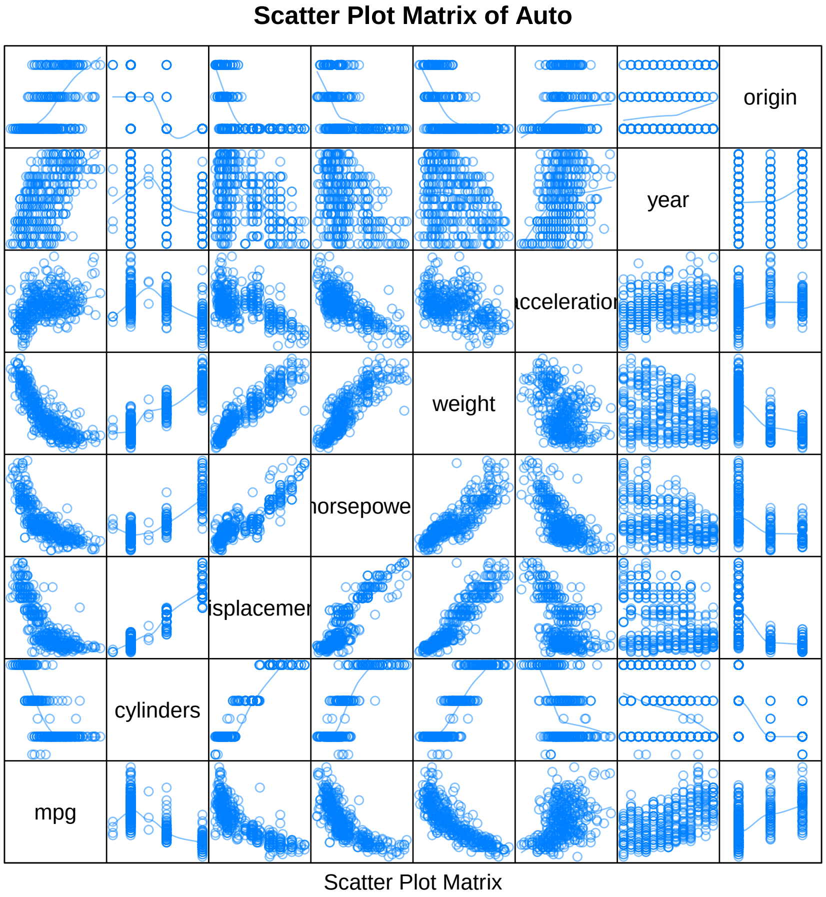


Figure 9: multi-variate comparisons

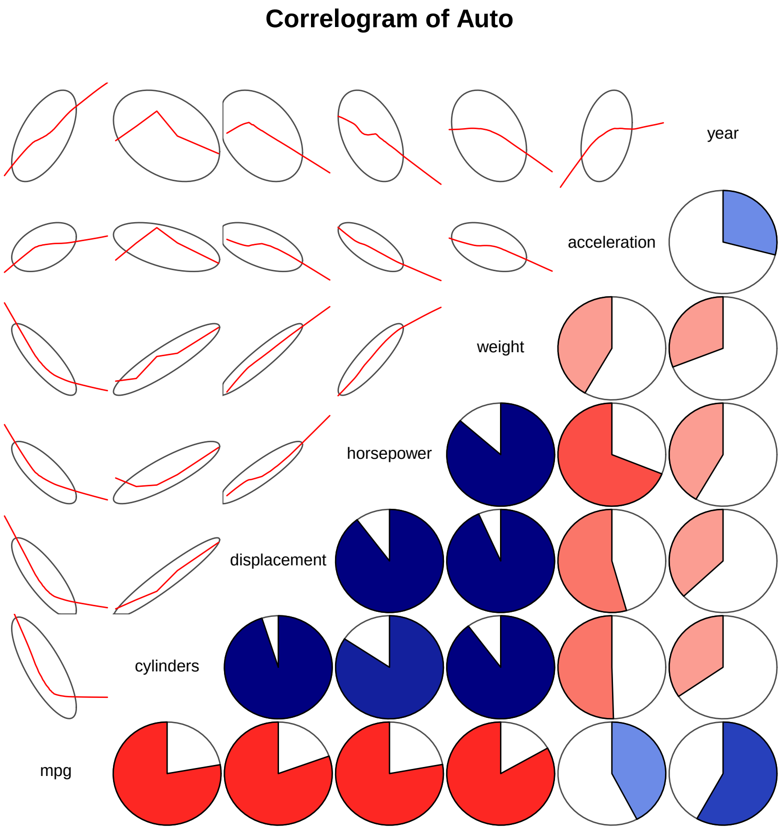
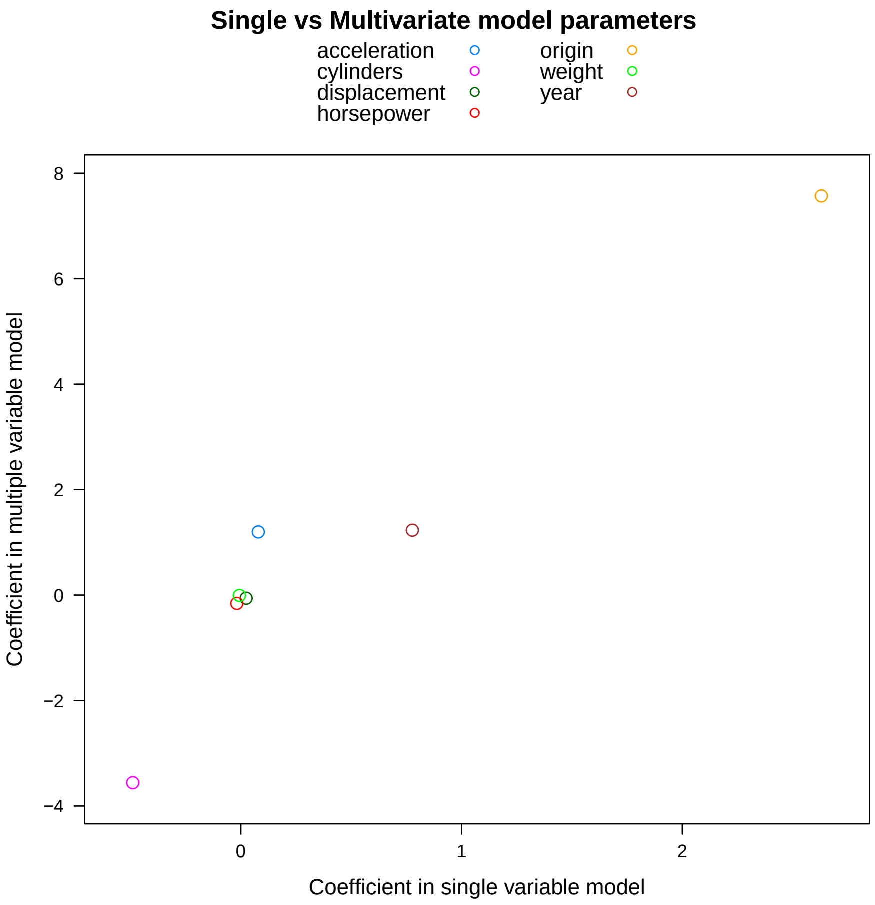


Figure 10: Correlogram

```
## -12.9170  -3.0243  -0.5021   2.3512  18.6128
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.12064    0.49443   71.03   <2e-16 ***
## displacement -0.06005    0.00224  -26.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.635 on 390 degrees of freedom
## Multiple R-squared:  0.6482,Adjusted R-squared:  0.6473
## F-statistic: 718.7 on 1 and 390 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059,Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9736  -2.7556  -0.3358   2.1379  16.5194
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.216524   0.798673   57.87   <2e-16 ***
## weight      -0.007647   0.000258  -29.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.333 on 390 degrees of freedom
## Multiple R-squared:  0.6926,Adjusted R-squared:  0.6918
## F-statistic: 878.8 on 1 and 390 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.989  -5.616  -1.199   4.801  23.239
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.8332     2.0485   2.359   0.0188 *
## acceleration   1.1976     0.1298   9.228   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.08 on 390 degrees of freedom
## Multiple R-squared:  0.1792,Adjusted R-squared:  0.1771
## F-statistic: 85.15 on 1 and 390 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0212  -5.4411  -0.4412   4.9739  18.2088
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -70.01167    6.64516  -10.54   <2e-16 ***
## year          1.23004    0.08736   14.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.363 on 390 degrees of freedom
## Multiple R-squared:  0.337,Adjusted R-squared:  0.3353
## F-statistic: 198.3 on 1 and 390 DF,  p-value: < 2.2e-16
##
##
```

```
## Call:
## lm(formula = fmla1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.451  -5.034  -1.034   3.649  18.966
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.0335     0.4086  49.025   <2e-16 ***
## origin2       7.5695     0.8767   8.634   <2e-16 ***
## origin3      10.4172     0.8276  12.588   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.396 on 389 degrees of freedom
## Multiple R-squared:  0.3318,Adjusted R-squared:  0.3284
## F-statistic:  96.6 on 2 and 389 DF,  p-value: < 2.2e-16
##
## Call:
## lm(formula = fmla, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders    -4.897e-01  3.212e-01  -1.524 0.128215
## displacement  2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower   -1.818e-02  1.371e-02  -1.326 0.185488
## weight       -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration  7.910e-02  9.822e-02   0.805 0.421101
## year          7.770e-01  5.178e-02  15.005  < 2e-16 ***
## origin2       2.630e+00  5.664e-01   4.643 4.72e-06 ***
## origin3       2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
## Multiple R-squared:  0.8242,Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```



```
df <- Auto %>% mutate(origin = factor(origin)) %>%
    select(-name)
```