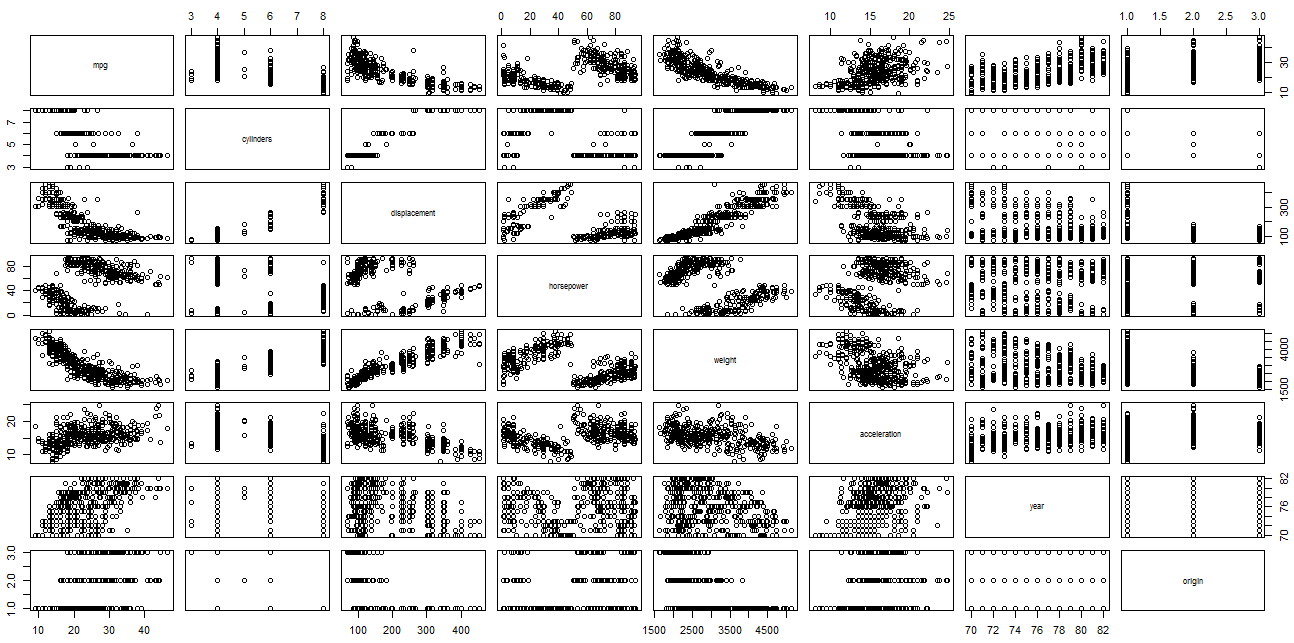
# Exploratory data analysis

plot(auto1)



Mpg decreases with increase in number of cylinders, displacement, weight, horsepower and increases with acceleration

ggplot(auto1, aes(x = displacement, y =cylinders)) + geom\_point()



ggplot(auto, aes(x = mpg, y = weight)) + geom\_point(aes(color = cylinders))



More weight more cylinder 🡪 weight ~ cylinder

More weight less mpg

ggplot(auto, aes(x = mpg, y = weight)) + geom\_point(aes(color = horsepower))



More weight less horsepower 🡪 weight ~1/ horsepower

ggplot(auto, aes(x = mpg, y = acceleration)) + geom\_point(aes(color = cylinders))

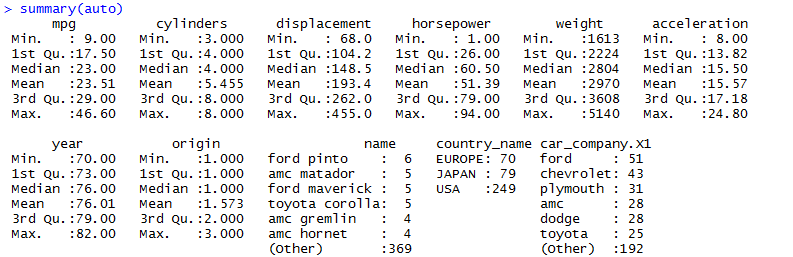


Liner relation between acceleration and mpg

More number of cylinder still low acceleration and low mpg

Mostly cars with cylinder 3 have more acceleration and fairly high mpg and average acceration ----

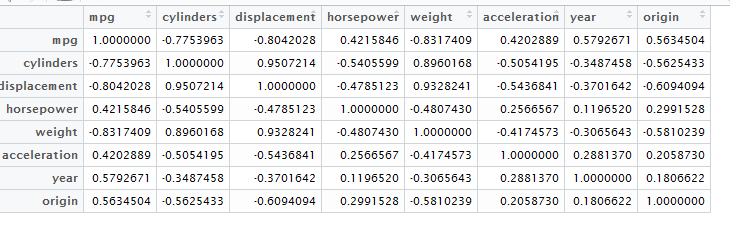
# Statistical Analysis



Average Mpg - 23.51

With number of cylinder ~5.4 with max cars belonginging to year -73

Correlation between all fields: correlations = cor(auto1)



Displacement 🡪 cylinder

Weight🡪 cylinder, displacement

# Find correlating dimensions

autoModel1 = lm(mpg ~ cylinders + horsepower + weight + displacement + year + acceleration, data = auto)

summary(autoModel1)

Residuals:

Min 1Q Median 3Q Max

-8.7735 -2.4087 -0.0915 1.9604 14.3342

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.615e+01 4.271e+00 -3.783 0.00018 \*\*\*

cylinders -1.018e-01 3.451e-01 -0.295 0.76821

horsepower 1.020e-02 6.957e-03 1.467 0.14331

weight -6.848e-03 5.978e-04 -11.454 < 2e-16 \*\*\*

displacement 5.640e-03 7.221e-03 0.781 0.43531

year 7.601e-01 5.081e-02 14.961 < 2e-16 \*\*\*

acceleration 7.548e-02 7.835e-02 0.963 0.33595

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.435 on 391 degrees of freedom

Multiple R-squared: 0.8097, Adjusted R-squared: 0.8068

F-statistic: 277.3 on 6 and 391 DF, p-value: < 2.2e-16

R-Squared ~80.68 highly predictive …

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. **0%**indicates that the model explains none of the variability of the response data around its mean

low p-value – **statistically significant** as P < 0.05 and **statistically highly significant** as P < 0.001 (less than one in a thousand chance of being wrong).

specially(weight and year)

Removing higher p value to notice the difference

autoModel2 = lm(mpg ~ cylinders + weight + displacement + year + acceleration, data = auto)

summary(autoModel2)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.457e+01 4.138e+00 -3.521 0.00048 \*\*\*

cylinders -2.586e-01 3.286e-01 -0.787 0.43177

weight -6.926e-03 5.963e-04 -11.614 < 2e-16 \*\*\*

displacement 7.268e-03 7.146e-03 1.017 0.30977

year 7.553e-01 5.078e-02 14.875 < 2e-16 \*\*\*

acceleration 8.035e-02 7.839e-02 1.025 0.30604

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.44 on 392 degrees of freedom

Multiple R-squared: 0.8087, Adjusted R-squared: 0.8062

F-statistic: 331.4 on 5 and 392 DF, p-value: < 2.2e-16

R-square reduce still weight and year p-value not effected

Removing all variable except weight and year

autoModel3 = lm(mpg ~ weight + year , data = auto)

summary(autoModel3)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.420e+01 3.968e+00 -3.578 0.000389 \*\*\*

weight -6.664e-03 2.139e-04 -31.161 < 2e-16 \*\*\*

year 7.566e-01 4.898e-02 15.447 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.435 on 395 degrees of freedom

Multiple R-squared: 0.8079, Adjusted R-squared: 0.8069

F-statistic: 830.4 on 2 and 395 DF, p-value: < 2.2e-16

Higher predictivity , less variable

**Final MPG:**

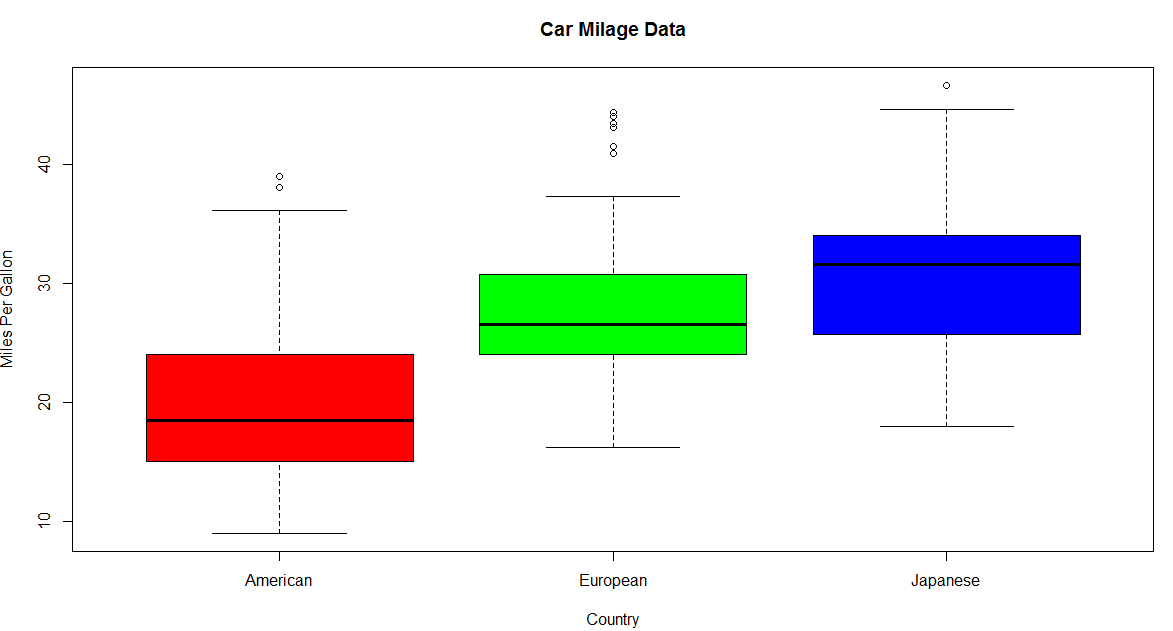
MPG=-14.2 - .006664\*weight + 0.7566\* year

# Compare miles per gallon for different countries

From summary … max cars belong to Americans – 249 out of 398 ~ 62 %

boxplot(mpg~country\_name,data=auto,col=(c("red","green","blue")), main="Car Milage Data",

xlab="Country", ylab="Miles Per Gallon")



23.5

(avg)

62 % cars from America have low MPG and Japanese car give high mileage.

Milege per company ---need to figure out way to represent for each country.

