ProjectWeek4

Executive Summary:

Explore the relationship between a set of variables and miles per gallon (MPG) (outcome). The interest is to answer if an automatic or manual transmission is better for MPG, plus, quantify the MPG difference between automatic and manual.

```
library(datasets)
data(mtcars)
```

```
names(mtcars) ##Shows names of the columns
```

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

```
dim(mtcars) ## (rown,columns)
```

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

```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.42   19.20   20.09   22.80   33.90
```

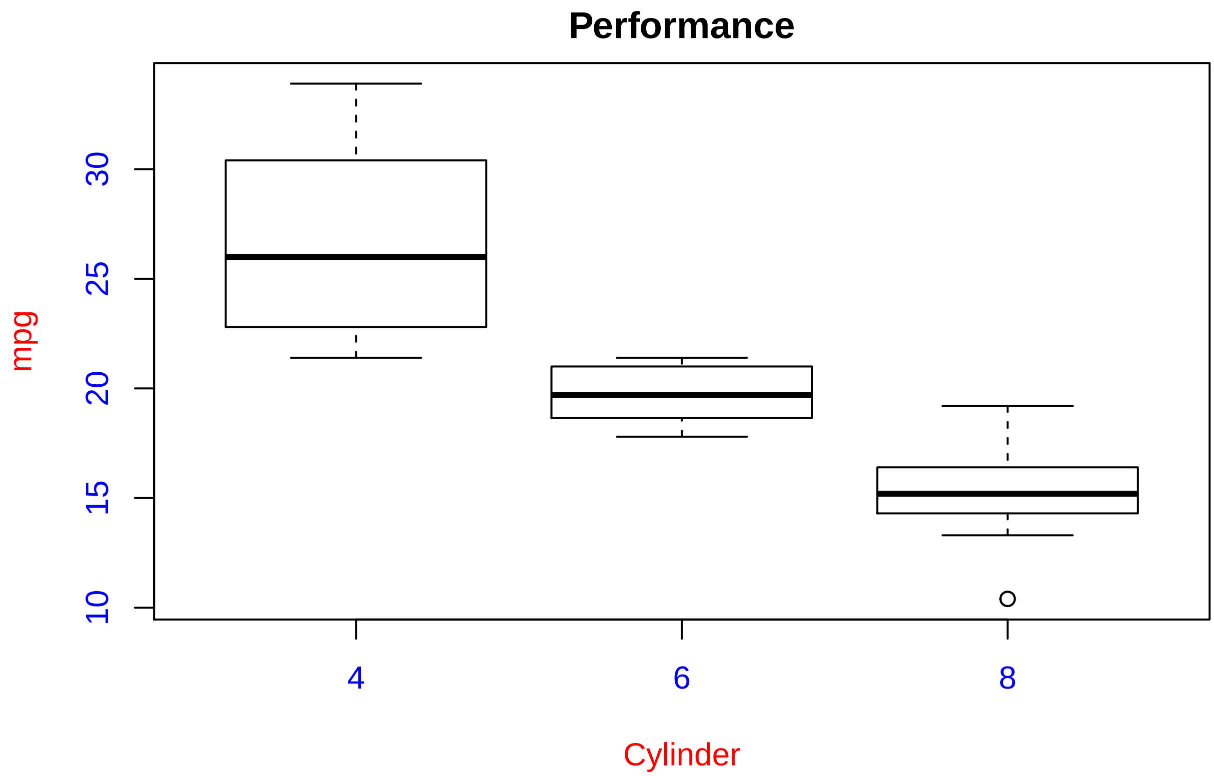
Exploratory data analysis

Make "cyl" and "am" as factors:

```
mtcars$cyl<-as.factor(mtcars$cyl)
mtcars$am<- factor(mtcars$am,labels=c("Automatic","Manual"))
```

Observe how the mean looks per cylinder:

```
par(mfrow=c(1,1),mar=c(4,4,2,1),oma = c(0, 0, 2, 0))
boxplot(mpg~cyl, mtcars,xlab = "Cylinder",ylab = "mpg", col.axis = "blue", col.lab = "red")
title(main="Performance")
```



Clustering data. The chart shows cars' performance based on number of cylinders. The first chart shows strong similarity among the cars with 4 cylinders while with 8 cylinders are two well defined groups

```
subset(mtcars, cyl== 8, select = c(hp, wt)) ## another subset
```

```
##                       hp    wt
## Hornet Sportabout   175 3.440
## Duster 360          245 3.570
## Merc 450SE          180 4.070
## Merc 450SL          180 3.730
## Merc 450SLC         180 3.780
## Cadillac Fleetwood  205 5.250
## Lincoln Continental 215 5.424
## Chrysler Imperial   230 5.345
## Dodge Challenger    150 3.520
## AMC Javelin         150 3.435
## Camaro Z28          245 3.840
## Pontiac Firebird    175 3.845
## Ford Pantera L      264 3.170
## Maserati Bora       335 3.570
```

```
#dist(subset(mtcars, cyl== 8, select = c(hp, wt))) ##calc dist between all points
Distance<-dist(subset(mtcars, cyl== 8, select = c(hp, wt)))
```

```r
subset<-subset(mtcars, cyl== 8, select = c(hp, wt))
cluster<-hclust(Distance)


subset(mtcars, cyl== 6, select = c(hp, wt)) ## another subset
```

```
##                 hp    wt
## Mazda RX4      110 2.620
## Mazda RX4 Wag  110 2.875
## Hornet 4 Drive 110 3.215
## Valiant        105 3.460
## Merc 280       123 3.440
## Merc 280C      123 3.440
## Ferrari Dino   175 2.770
```

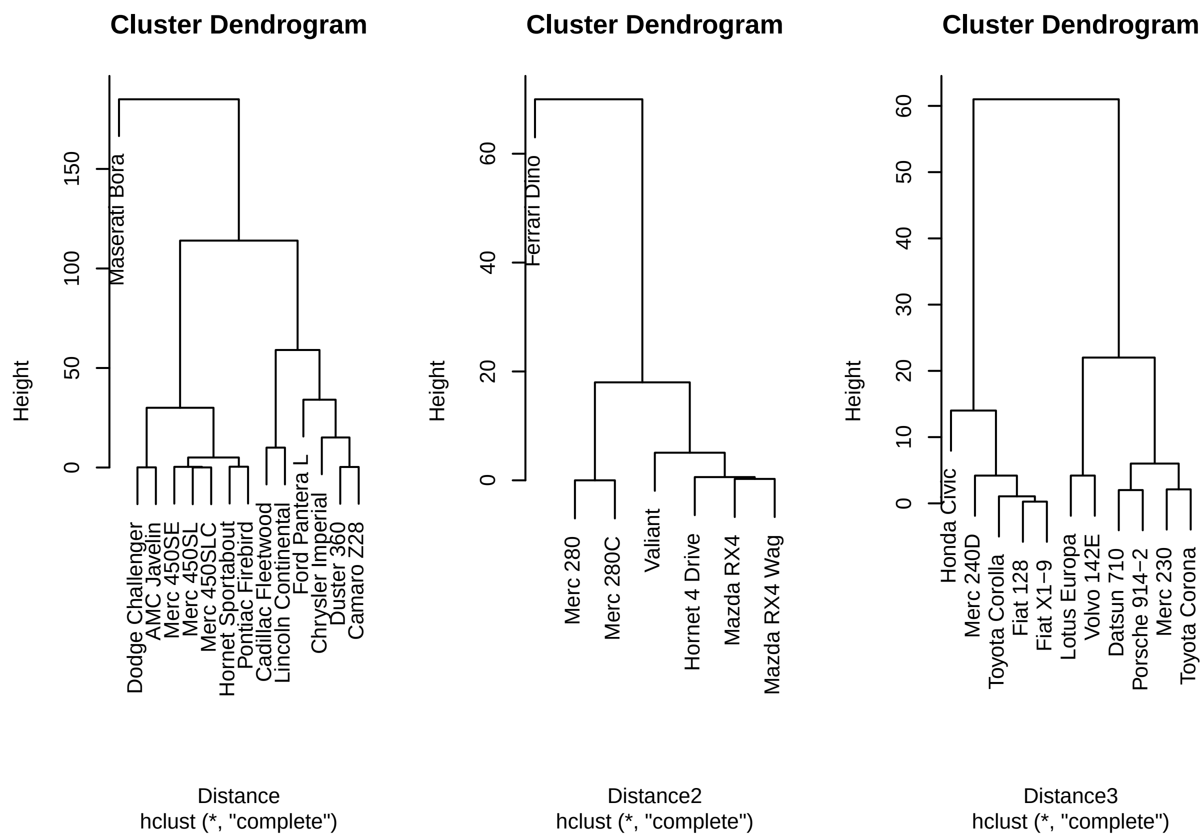
```r
#dist(subset(mtcars, cyl== 6, select = c(hp, wt))) ##calc dist between all points
Distance2<-dist(subset(mtcars, cyl== 6, select = c(hp, wt)))
subset2<-subset(mtcars, cyl== 6, select = c(hp, wt))
cluster2<-hclust(Distance2)


subset(mtcars, cyl== 4, select = c(hp, wt)) ## another subset
```

```
##                 hp    wt
## Datsun 710      93 2.320
## Merc 240D       62 3.190
## Merc 230        95 3.150
## Fiat 128        66 2.200
## Honda Civic     52 1.615
## Toyota Corolla  65 1.835
## Toyota Corona   97 2.465
## Fiat X1-9       66 1.935
## Porsche 914-2   91 2.140
## Lotus Europa   113 1.513
## Volvo 142E     109 2.780
```

```r
#dist(subset(mtcars, cyl== 4, select = c(hp, wt))) ##calc dist between all points
Distance3<-dist(subset(mtcars, cyl== 4, select = c(hp, wt)))
subset3<-subset(mtcars, cyl== 4, select = c(hp, wt))
cluster3<-hclust(Distance3)


par(mfrow=c(1, 3))
plot(cluster)
plot(cluster2)
plot(cluster3)
```



Regression analysis:

Regression with single predictor (fit1)

```
fit<-lm(mpg ~ am, data=mtcars) # regression model with "mpg" as the outcome and "am" as the predictor.
summary(lm(mpg ~ am, data=mtcars))$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

The intercept is the expected mean value of Y when all X=0, which means when the residual has mean = zero

If we concider only "am", it says that for every 1% increase in "amManual", we expect a 7.24 increase in mpg

Regression with multiple-predictors (fit2)

```r
fit2<-lm(formula = mpg ~ ., data = mtcars)
shapiro.test(fit2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit2$residuals
## W = 0.96212, p-value = 0.3135
```

It's convenient to check normality, If p-value>0.05 then fails to reject normality. That's the case here.

```r
summary(lm(formula = mpg ~ ., data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4734 -1.3794 -0.0655  1.0510  4.3906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.81984   16.30602   1.093   0.2875
## cyl6        -1.66031    2.26230  -0.734   0.4715
## cyl8         1.63744    4.31573   0.379   0.7084
## disp         0.01391    0.01740   0.799   0.4334
## hp          -0.04613    0.02712  -1.701   0.1045
## drat         0.02635    1.67649   0.016   0.9876
## wt          -3.80625    1.84664  -2.061   0.0525 .
## qsec         0.64696    0.72195   0.896   0.3808
## vs           1.74739    2.27267   0.769   0.4510
## amManual     2.61727    2.00475   1.306   0.2065
## gear         0.76403    1.45668   0.525   0.6057
## carb         0.50935    0.94244   0.540   0.5948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.582 on 20 degrees of freedom
## Multiple R-squared:  0.8816, Adjusted R-squared:  0.8165
## F-statistic: 13.54 on 11 and 20 DF,  p-value: 5.722e-07
```

For every 1% increase in "cyl6", we expect a 1.66031 decrease in mpg, holding all other variables constant

For every 1% increase in "amManual", we expect a 2.61726 increase in mpg, holding all other variables constant

Regression with multiple-predictors (fit3)

```
fit3<-lm(formula = mpg ~cyl+wt+am, data = mtcars)
shapiro.test(fit3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit3$residuals
## W = 0.95915, p-value = 0.2602
```

Residuals show normality

```
summary(lm(formula = mpg ~cyl+wt+am, data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ cyl + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4898 -1.3116 -0.5039  1.4162  5.7758
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7536     2.8135  11.997  2.5e-12 ***
## cyl6         -4.2573     1.4112  -3.017  0.00551 **
## cyl8         -6.0791     1.6837  -3.611  0.00123 **
## wt           -3.1496     0.9080  -3.469  0.00177 **
## amManual      0.1501     1.3002   0.115  0.90895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.603 on 27 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8134
## F-statistic: 34.79 on 4 and 27 DF,  p-value: 2.73e-10
```

For every 1% increase in "cyl6", we expect a 4.2573 decrease in mpg, holding all other variables constant

For every 1% increase in "amManual", we expect a 0.1501 increase in mpg, holding all other variables constant, which show less increase than considering all the other predictors.

Let's evaluate the 3 Regression models

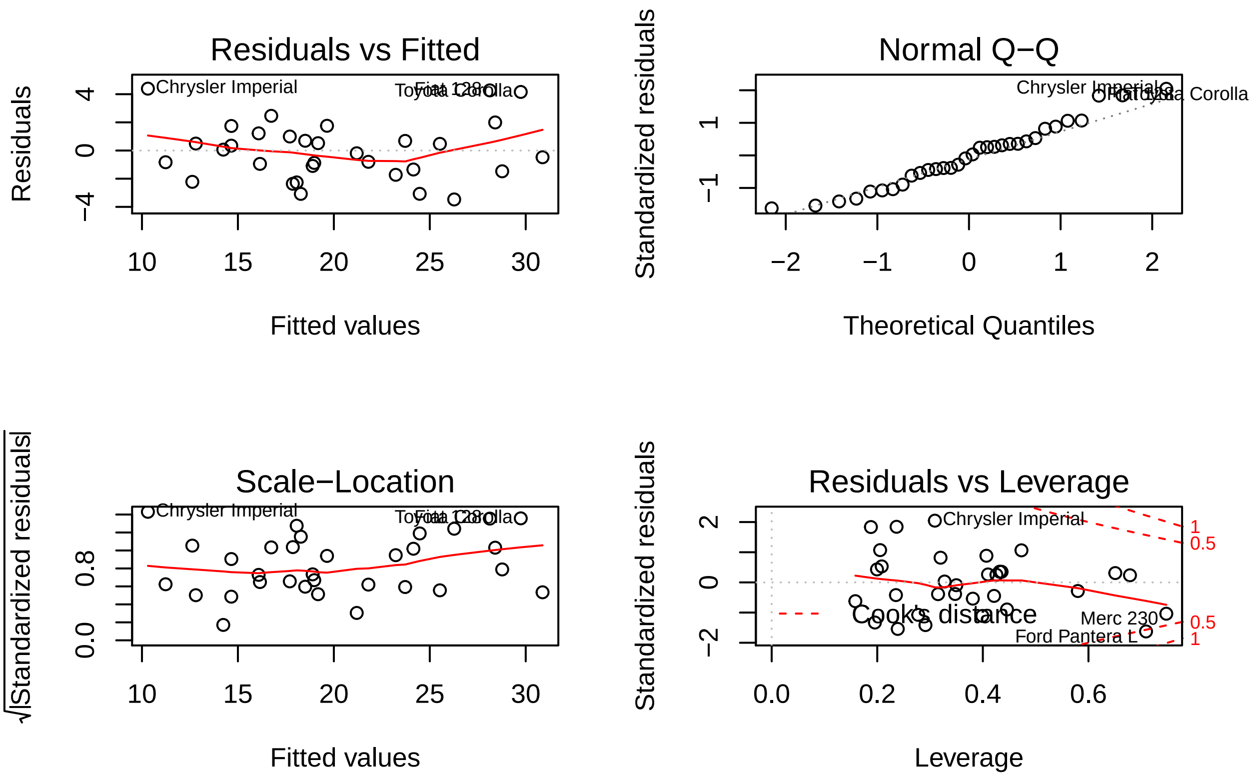
```r
anova(fit,fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 3: mpg ~ cyl + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     20 133.32 10    587.57 8.8143 2.249e-05 ***
## 3     27 182.97 -7    -49.64 1.0639    0.4211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It shows high significance when using all the predictors versus only amManual or cyl+disp+hp+wt

Residuals

```r
par(mfrow=c(2, 2))
plot(fit2)
```



Most of points fall on the line of Normal Q-Q plot indicating normality

Evaluate collinearity of the best fit (fit2), Variance Inflation Factors

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.2.5
```

```
vif(fit2)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## cyl  36.345193  2        2.455341
## disp 21.631435  1        4.650961
## hp   16.078598  1        4.009813
## drat  3.736566  1        1.933020
## wt   15.182209  1        3.896435
## qsec  7.739648  1        2.782022
## vs    6.101654  1        2.470153
## am    4.653616  1        2.157224
## gear  5.371501  1        2.317650
## carb 10.775733  1        3.282641
```

These VIFs show, for each regression coefficient, the variance inflation due to including all the others. For instance, the variance in the estimated coefficient of cyl is 36.34 times what it might have been if cyl were not correlated with the other regressors. Since "cyl" and score on an "disp" are likely to be correlated, we might guess that most of the variance inflation for cyl is due to including disp.