Project

Brandon Hom

11/2/2021

Introduction

Exploratory data analysis

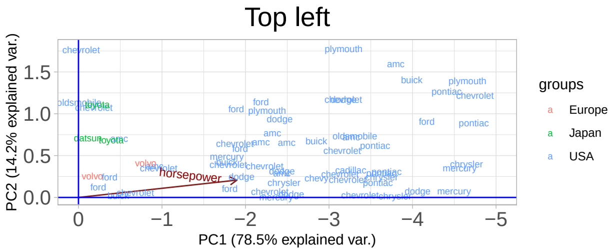
```
# data cleaning up
data <- read.csv('auto-mpg.csv')
#convert horsepower chr->dbl
data$horsepower <- as.numeric(data$horsepower)
#remove rows with missing values
data <- na.omit(data)
#translate origin numbers to country strings
data$origin <- ifelse(data$origin==1,"USA",ifelse(data$origin==2,"Europe","Japan"))
data$origin <- as.factor(data$origin)
#cylinders count for 3 and 5 low combine with 4 and 6 respectively
data$cylinders <- replace(data$cylinders,data$cylinders %in% c(3,5),c(4,6))
data$cylinders <- as.factor(data$cylinders)
#remove model.year, not interested in this feature
data <- data[-c(7)]
data$car.name <- word(data$car.name,1)
```

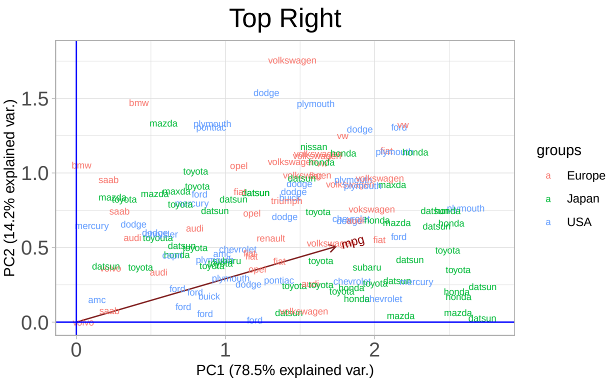
```
cor(data[-c(2,7,8)])
```

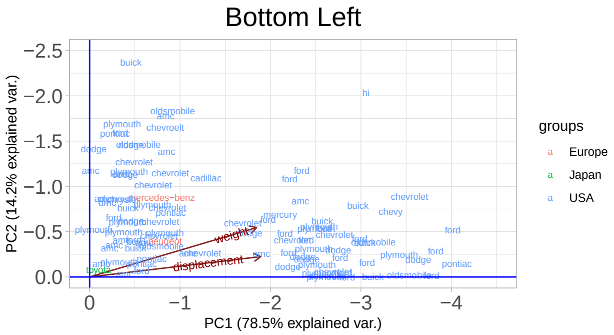
```
##                     mpg displacement horsepower     weight acceleration
## mpg           1.0000000   -0.8051269 -0.7784268 -0.8322442    0.4233285
## displacement -0.8051269    1.0000000  0.8972570  0.9329944   -0.5438005
## horsepower   -0.7784268    0.8972570  1.0000000  0.8645377   -0.6891955
## weight       -0.8322442    0.9329944  0.8645377  1.0000000   -0.4168392
## acceleration  0.4233285   -0.5438005 -0.6891955 -0.4168392    1.0000000
```

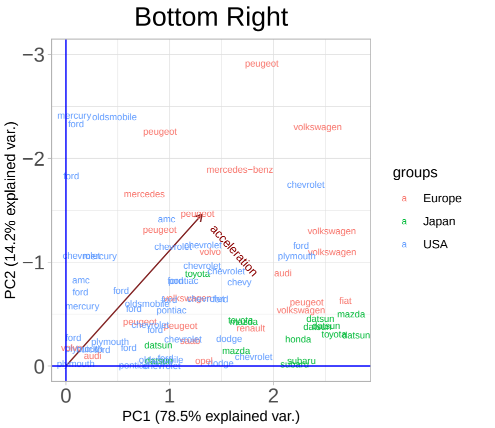
```
#h.clustering.complete <- hclust(dist(data),method="complete")
#dendro.data <- dendro_data(as.dendrogram(h.clustering.complete),type="rectangle")

#ggplot(dendro.data$segments) +
  #geom_segment(aes(x = x, y = y, xend = xend, yend = yend))+
  #geom_text(data = dendro.data$labels, aes(x, y, label = label),
           # hjust=1,angle = 90, size = 3.5)+ylim(-1,5000)+
  #theme_light()+
  #theme(panel.grid=element_blank(),
   #     plot.title=element_text(hjust=.5,size=20),
      #  axis.text = element_text(size=15)
       # )+
  #labs(title="Complete linkage Dendrogram of Cars",
   #    x="Car Name",
    #    y="height")
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5
## Standard deviation     1.9816 0.8438 0.47500 0.28788 0.22966
## Proportion of Variance 0.7853 0.1424 0.04512 0.01658 0.01055
## Cumulative Proportion  0.7853 0.9277 0.97288 0.98945 1.00000
```

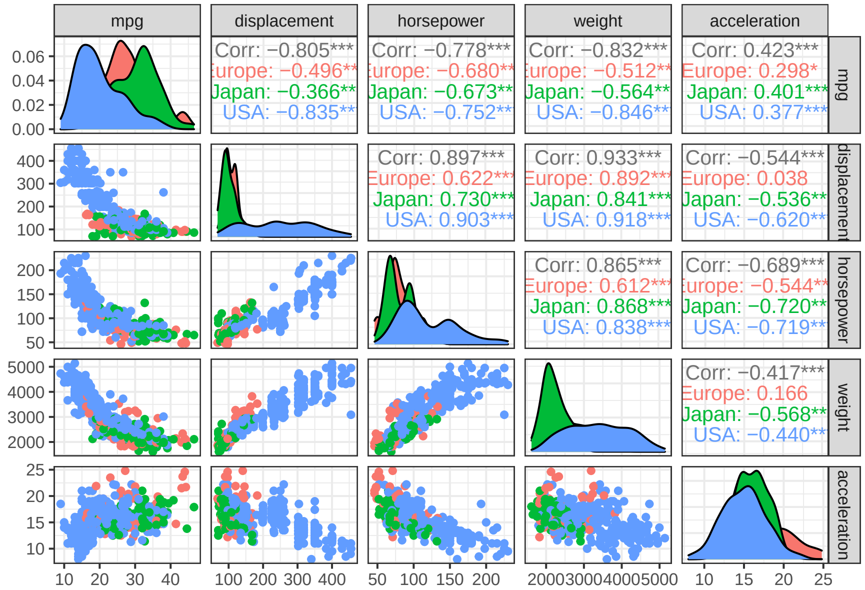








```
ggpairs(data[-c(2,7,8)],aes(color=data$origin))+theme_bw()
```



```
data.transformed <- data
data.transformed$mpg <- log(data.transformed$mpg,base=10)
data.transformed <- data.transformed[-c(8)]
```

```
fit <- lm(mpg~.,data=data.transformed)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = data.transformed)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.167155 -0.040523 -0.005129  0.040544  0.241478
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.773e+00  3.980e-02  44.563  < 2e-16 ***
## cylinders6   -6.762e-02  1.365e-02  -4.955 1.09e-06 ***
## cylinders8   -6.033e-02  2.471e-02  -2.442   0.0151 *
## displacement  1.624e-04  1.431e-04   1.135   0.2573
## horsepower   -1.662e-03  2.695e-04  -6.167 1.77e-09 ***
## weight       -7.290e-05  1.279e-05  -5.701 2.38e-08 ***
## acceleration -2.664e-03  1.932e-03  -1.379   0.1688
## originJapan   2.812e-02  1.097e-02   2.563   0.0108 *
## originUSA    -3.729e-03  1.112e-02  -0.335   0.7375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06509 on 383 degrees of freedom
## Multiple R-squared:  0.8097, Adjusted R-squared:  0.8057
## F-statistic: 203.7 on 8 and 383 DF,  p-value: < 2.2e-16
```

```
library(MASS)
```

```
##
```

```
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
step.model <- stepAIC(fit, direction = "both",
trace = FALSE)
summary(step.model)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + horsepower + weight + acceleration +
##     origin, data = data.transformed)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.174422 -0.038163 -0.005526  0.040005  0.251095
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.772e+00  3.979e-02  44.534  < 2e-16 ***
## cylinders6   -5.992e-02  1.185e-02  -5.059 6.56e-07 ***
## cylinders8   -4.330e-02  1.963e-02  -2.206   0.0280 *
## horsepower   -1.573e-03  2.579e-04  -6.099 2.60e-09 ***
## weight       -6.697e-05  1.168e-05  -5.736 1.96e-08 ***
## acceleration -2.836e-03  1.927e-03  -1.472   0.1419
## originJapan   2.821e-02  1.098e-02   2.570   0.0105 *
## originUSA     5.768e-04  1.045e-02   0.055   0.9560
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06512 on 384 degrees of freedom
## Multiple R-squared:  0.8091, Adjusted R-squared:  0.8056
## F-statistic: 232.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

Conclusion

Appendix: R code used

```
# keeps this here to remove comments from knitted output.
knitr::opts_chunk$set(comments=NA)
library(tidyverse)
library(knitr)
library(leaps)
library(GGally)
library(ggbiplot)
# data cleaning up
data <- read.csv('auto-mpg.csv')
#convert horsepower chr->dbl
data$horsepower <- as.numeric(data$horsepower)
#remove rows with missing values
data <- na.omit(data)
#translate origin numbers to country strings
```

```
data$origin <- ifelse(data$origin==1,"USA",ifelse(data$origin==2,"Europe","Japan"))
data$origin <- as.factor(data$origin)
#cylinders count for 3 and 5 low combine with 4 and 6 respectively
data$cylinders <- replace(data$cylinders,data$cylinders %in% c(3,5),c(4,6))
data$cylinders <- as.factor(data$cylinders)
#remove model.year, not interested in this feature
data <- data[-c(7)]
data$car.name <- word(data$car.name,1)
cor(data[-c(2,7,8)])
#h.clustering.complete <- hclust(dist(data),method="complete")
#dendro.data <- dendro_data(as.dendrogram(h.clustering.complete),type="rectangle")

#ggplot(dendro.data$segments) +
  #geom_segment(aes(x = x, y = y, xend = xend, yend = yend))+
  #geom_text(data = dendro.data$labels, aes(x, y, label = label),
           # hjust=1,angle = 90, size = 3.5)+ylim(-1,5000)+
  #theme_light()+
  #theme(panel.grid=element_blank(),
   #     plot.title=element_text(hjust=.5,size=20),
      #  axis.text = element_text(size=15)
       # )+
  #labs(title="Complete linkage Dendrogram of Cars",
   #    x="Car Name",
    #   y="height")
pcs.out <- prcomp(data[-c(2,7,8)],scale.=T)
summary(pcs.out)

ggbiplot(pcs.out,labels = data$car.name,groups=data$origin,obs.scale = 1,labels.size = 2.3)+
  geom_hline(yintercept = 0,col="blue")+
  geom_vline(xintercept = 0,col="blue")+
  ylim(0,1.8)+
  xlim(0,-5)+
  theme_light()+
  theme(plot.title=element_text(hjust=.5,size=20),
        axis.text = element_text(size=15)
        )+
  labs(title="Top left",
       )


ggbiplot(pcs.out,labels = data$car.name,groups=data$origin,obs.scale = 1,labels.size = 2.3)+
  geom_hline(yintercept = 0,col="blue")+
  geom_vline(xintercept = 0,col="blue")+
  ylim(0,1.8)+
  xlim(0,2.8)+
  theme_light()+
  theme(plot.title=element_text(hjust=.5,size=20),
        axis.text = element_text(size=15)
        )+
  labs(title="Top Right",
       )

ggbiplot(pcs.out,labels = data$car.name,groups=data$origin,obs.scale = 1,labels.size = 2.3)+
  geom_hline(yintercept = 0,col="blue")+
  geom_vline(xintercept = 0,col="blue")+
  ylim(0,-2.5)+
  xlim(0,-4.5)+
  theme_light()+
  theme(plot.title=element_text(hjust=.5,size=20),
        axis.text = element_text(size=15)
        )+
  labs(title="Bottom Left",
       )

ggbiplot(pcs.out,labels = data$car.name,groups=data$origin,obs.scale = 1,labels.size = 2.3)+
  geom_hline(yintercept = 0,col="blue")+
  geom_vline(xintercept = 0,col="blue")+
  ylim(0,-3)+
  xlim(0,2.8)+
  theme_light()+
  theme(plot.title=element_text(hjust=.5,size=20),
        axis.text = element_text(size=15)
        )+
  labs(title="Bottom Right",
       )
ggpairs(data[-c(2,7,8)],aes(color=data$origin))+theme_bw()
data.transformed <- data
data.transformed$mpg <- log(data.transformed$mpg,base=10)
data.transformed <- data.transformed[-c(8)]
fit <- lm(mpg~.,data=data.transformed)
summary(fit)
library(MASS)
step.model <- stepAIC(fit, direction = "both",
trace = FALSE)
summary(step.model)
```