역사 데이터 시각화 분석

Correlation Analysis with R

What is Correlation Analysis?

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice

- 개념(Concept)

- 두 변수나 두 데이터 세트 사이에 존재하는 선형관계의 정도를 파악하는 분석이다.
- R(상관계수)값을 이용한다.
- 상관계수의 값이 -1이면 완전한 음의 상관이고, +1이면 완전한 양의 상관이다.
- 음의 상관관계가 있는 두 변수를 산점도로 나타내면 점의 분포는 우 하향의 모습을 띄고 양의 상관관계가 있는 두 변수를 산점도로 나타내면 점의 분포는 우 상향의 모습을 띈다.
- 상관계수는 -1<R<1의 범위 값을 지닌다.

- 수식(Formula)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

What is Correlation Analysis?

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice

- 상관관계가 유의하지 않은 경우

  - 이상치(outlier)가 존재하는 경우, 이 값은 상관계수 값에 큰 영향을 미치므로 이상치의 존재 여부를 확인한 후 상관 분석을 시행하여야 한다.
  - 두 변수의 관계가 비선형인 경우 상관계수는 유의하지 않다.
  - 상관계수의 값을 통해 얻어진 결과는 상호간에 상관관계가 있는 것이지, 절대적인 인과관계가 있다고 해석하는 것은 오류이다.

- 예제(Example)

  - mtcars 데이터 셋은 데이터는 1974 년 모터 트렌드 미국 잡지에서 추출하였으며 1973년-1974년도 모델의 32종의 자동차들의 연비등 자동차의 11가지 중요 정보를 나타내고 있다.

- 변수 설명

  - mpg = 마일 / (US) 갤런
  - cyl = 실린더의 수
  - disp = 변위 (cu.in.)
  - hp = 총 마력
  - drat = 리어 액슬 비율
  - wt = 무게 (파운드 / 1000)
  - qsec = 1/4 마일 시간
  - vs = V / S
  - am = 변속기 (0 = 자동, 1 = 수동)
  - gear = 기어의 수
  - carb = 기화기의 수

```
> str(mtcars)
'data.frame':   32 obs. of  12 variables:
 $ X   : Factor w/ 32 levels "AMC Javelin",..: 18 19 5 13 14 31 7 21 20 22 ...
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : int  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : int  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  : int  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : int  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: int  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: int  4 4 1 1 2 1 4 2 2 4 ...
```

- 데이터 확인

```
> str(mtcars)
'data.frame':   32 obs. of  12 variables:
 $ X   : Factor w/ 32 levels "AMC Javelin",..: 18 19 5 13 14 31 7 21 20 22 ...
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : int  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : int  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  : int  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : int  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: int  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: int  4 4 1 1 2 1 4 2 2 4 ...
> head(mtcars)
                  X  mpg cyl disp  hp drat    wt  qsec vs am gear carb
1         Mazda RX4 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
2     Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
3        Datsun 710 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
4    Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
5 Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
6           Valiant 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

- 전체 데이터를 요약함으로써 각 변수의 데이터 특징을 한눈에 파악할 수 있다.

Correlation Analysis with R

- 상관분석
  - Pearson상관계수: 데이터가 연속형 변수(등간척도, 비율척도)일때 사용하며, 확률분포로 정규분포를 가정한다.
  - Kendall상관계수: 데이터가 질적 변수(순위척도)일때 사용하며, 확률분포에 대한 가정이 없고 비모수적 방법의 상관분석이다.
- 변수 mpg와 cyl간의 상관분석

```
> attach(mtcars)
> cor(mpg,cyl,method="pearson")
[1] -0.852162
> cor(mpg,cyl,method="kendall")
[1] -0.7953134
> cor.test(mpg,cyl)

        Pearson's product-moment correlation

data:  mpg and cyl
t = -8.9197, df = 30, p-value = 6.113e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9257694 -0.7163171
sample estimates:
      cor
-0.852162

> detach(mtcars)
```

  - mpg와 cyl에 대해 상관분석을 한 결과 p값이 0.05이하 이므로 두 변수간 연관성이 있다는 결과가 나왔다.
- 상관 계수 값 계산하기

```
> sum((mpg-mean(mpg))*(cyl-mean(cyl)))/sqrt(sum((mpg-mean(mpg))*(mpg-mean(mpg)))*sum((cyl-mean(cyl))*(cyl-mean(cyl))))
[1] -0.852162
```

Correlation Analysis with R

- 상관행렬 생성하기

```
> mtcars_2<-mtcars[,2:12]
> head(mtcars_2)
   mpg cyl disp  hp drat    wt  qsec vs am gear carb
1 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
2 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
3 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
4 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
5 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
6 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
> mcor<-cor(mtcars_2)
> round(mcor,2)
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

- 각각의 변수에 대하여 Pearson 상관계수 행렬을 생성하면 변수들간의 상관관계를 한눈에 파악 할 수 있다.

- 상관계수 행렬 값을 활용한 히트 맵(Hitmap)

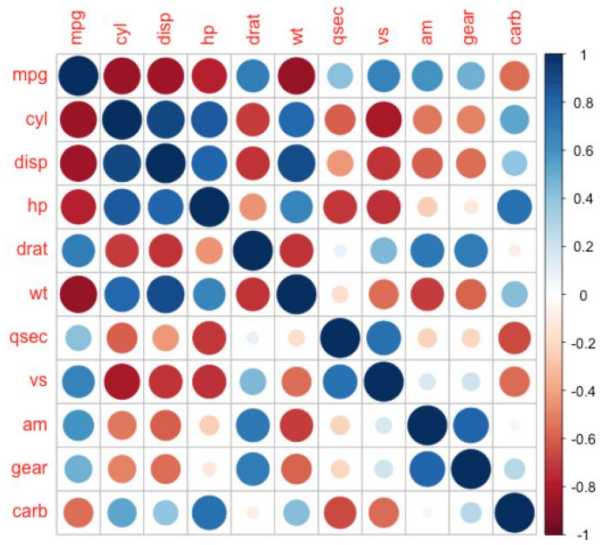
```
> install.packages("corrplot")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/corrplot_0.73.tgz'을 시도합니다
Content type 'application/x-gzip' length 2679598 bytes (2.6 Mb)
URL을 열었습니다
==================================================
downloaded 2.6 Mb


The downloaded binary packages are in
        /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//RtmpKrDACp/downloaded_packages
```

- 상관 계수 행렬을 시각화 하기에 앞서 사용할 패키지를 설치하여 준다.
- Install.packages("corrplot")

Correlation Analysis with R

- 상관계수 행렬 값을 활용한 히트 맵(Hitmap)

```
> library(corrplot)
> mcor<-cor(mtcars_2)
> round(mcor,2)
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
> corrplot(mcor)
```

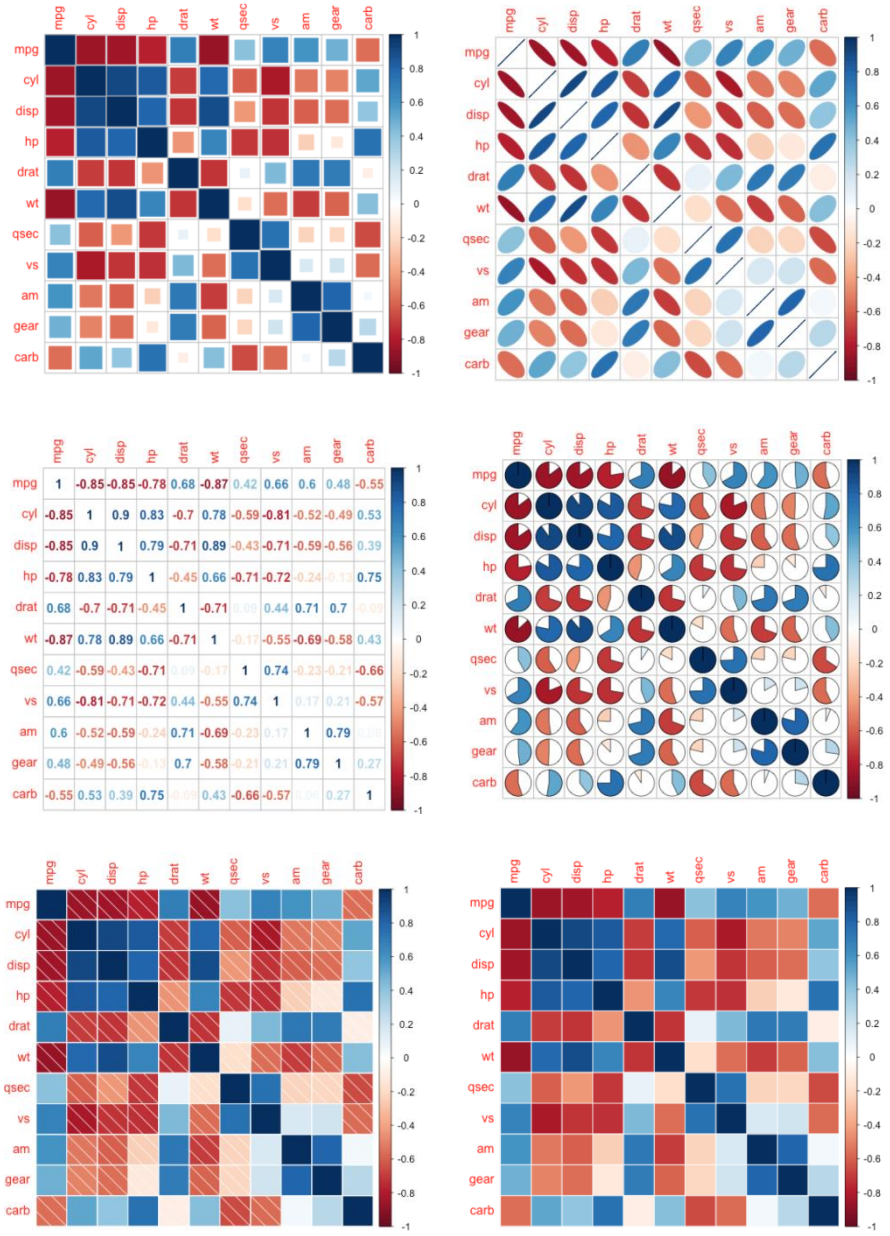


- 디폴트 값으로 설정되어 있는 히트맵은 위의 상관 행렬 표를 원으로 표시하고 계수의 크기를 원의 크기로 표현하여 각 변수의 관계를 손쉽게 확인 할 수 있다.

Correlation Analysis with R

- Method에 따른 히트맵 시각화

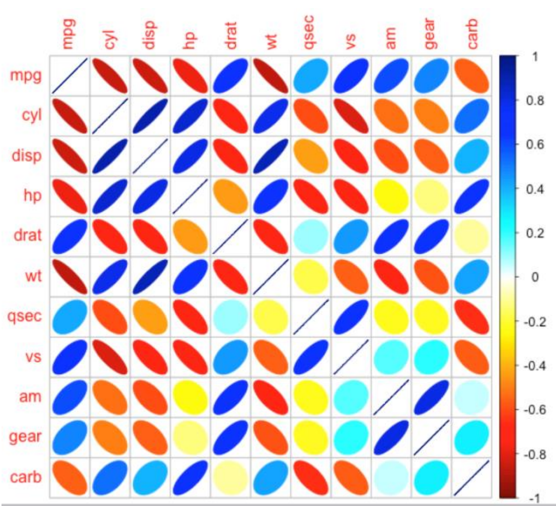
```
> corrplot(mcor, method="square")
>
> corrplot(mcor, method="ellipse")
>
> corrplot(mcor, method="number")
>
> corrplot(mcor, method="pie")
>
> corrplot(mcor, method="shade")
>
> corrplot(mcor, method="color")
```

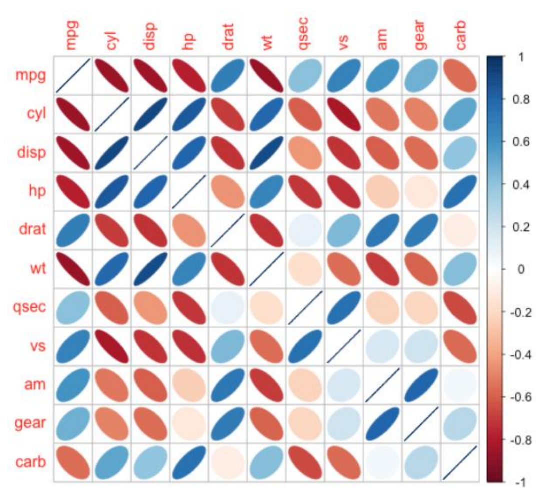


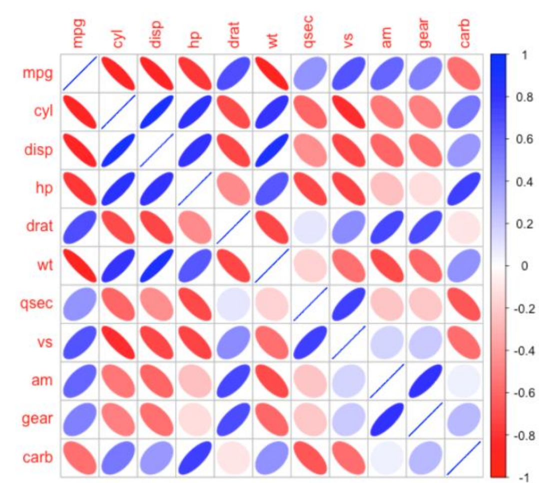
Correlation Analysis with R

- 색상에 따른 히트맵 시각화

```
> col1 <- colorRampPalette(c("#7F0000","red","#FF7F00","yellow","white",
+                            "cyan", "#007FFF", "blue","#00007F"))
>
> col2 <- colorRampPalette(c("#67001F", "#B2182B", "#D6604D", "#F4A582", "#FDDBC7",
+                            "#FFFFFF", "#D1E5F0", "#92C5DE", "#4393C3", "#2166AC", "#053061"))
>
> col3 <- colorRampPalette(c("red", "white", "blue"))
>
> wb <- c("white","black")
> corrplot(mcor, method="ellipse", col = col1(200))
> corrplot(mcor, method="ellipse", col = col2(200))
> corrplot(mcor, method="ellipse", col = col3(200))
> corrplot(mcor, method="ellipse", col = wb)
```



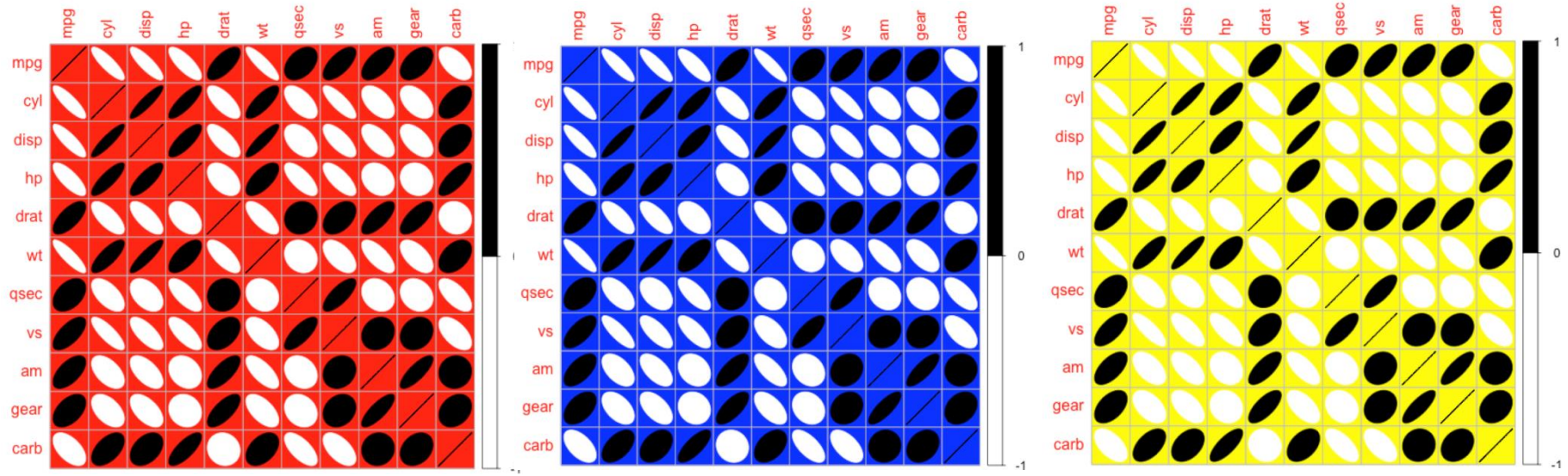




Correlation Analysis with R

- 배경에 따른 히트맵 시각화

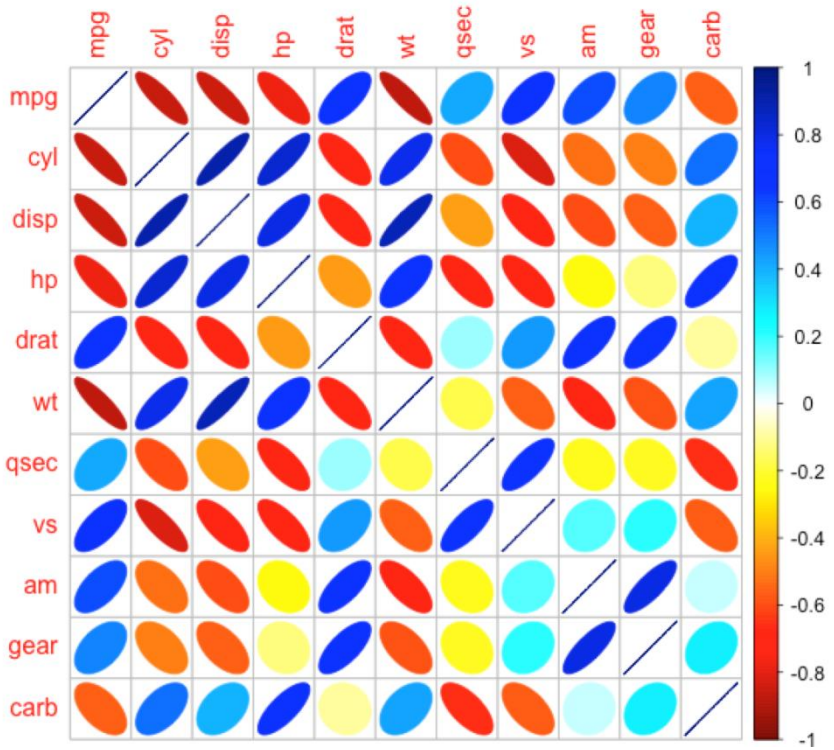
```
> corrplot(mcor, method="ellipse", col = wb, bg = "red")
> corrplot(mcor, method="ellipse", col = wb, bg = "blue")
> corrplot(mcor, method="ellipse", col = wb, bg = "yellow")
```

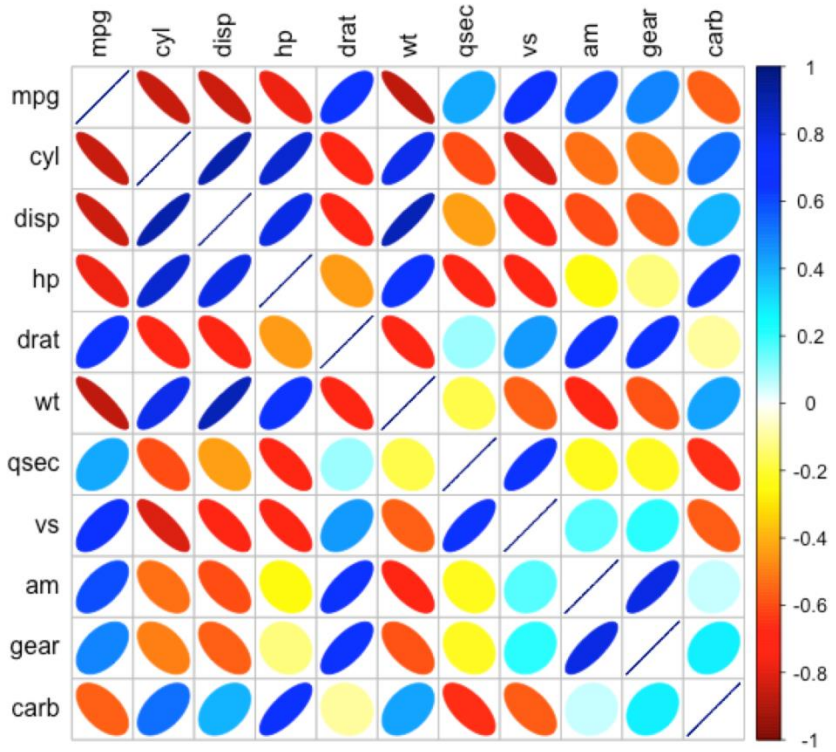


Correlation Analysis with R

- 변수명 색상에 따른 히트맵 시각화

```
> corrplot(mcor, method="ellipse", col = col1(200))
> corrplot(mcor, method="ellipse", col = col1(200), tl.col="black")
```

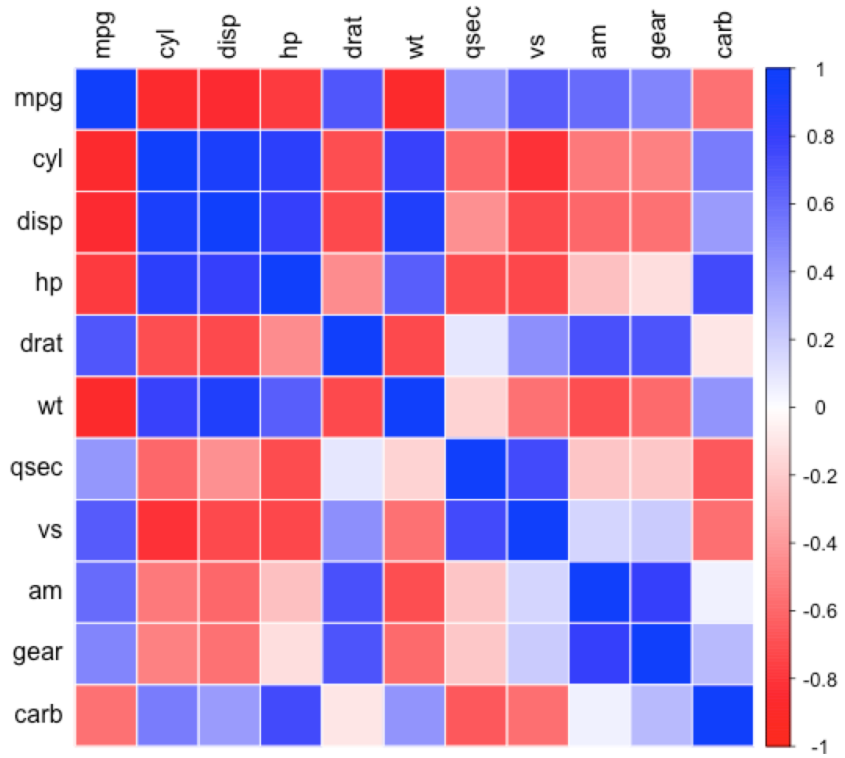


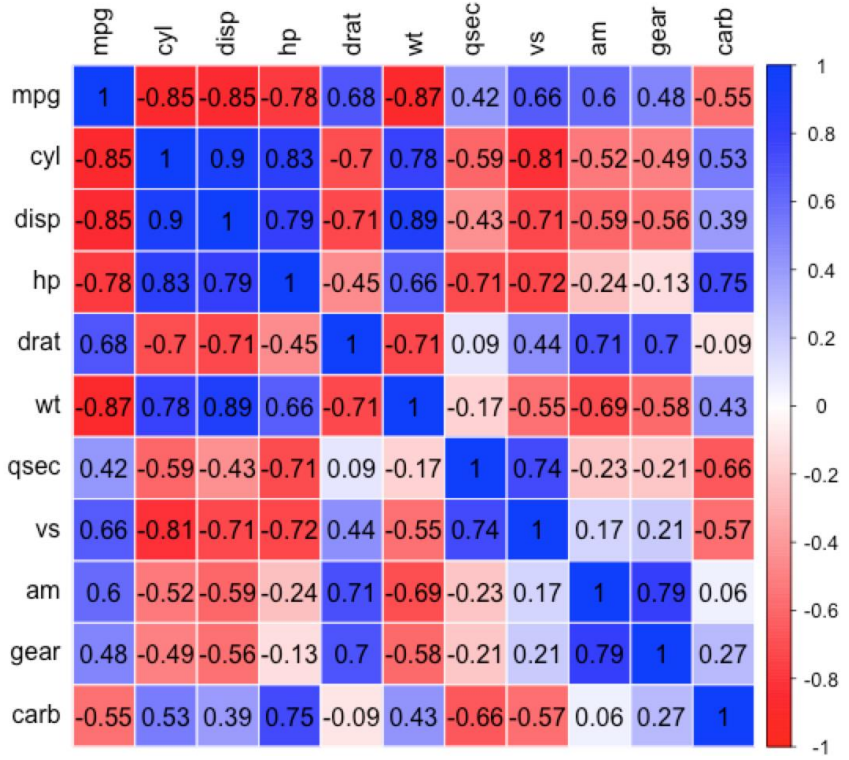


Correlation Analysis with R

- 상관계수 값 표현에 따른 히트맵 시각화

```
> corrplot(mcor, method="color", col = col3(200), tl.col="black")
> corrplot(mcor, method="color", col = col3(200), tl.col="black", addCoef.col="black")
```

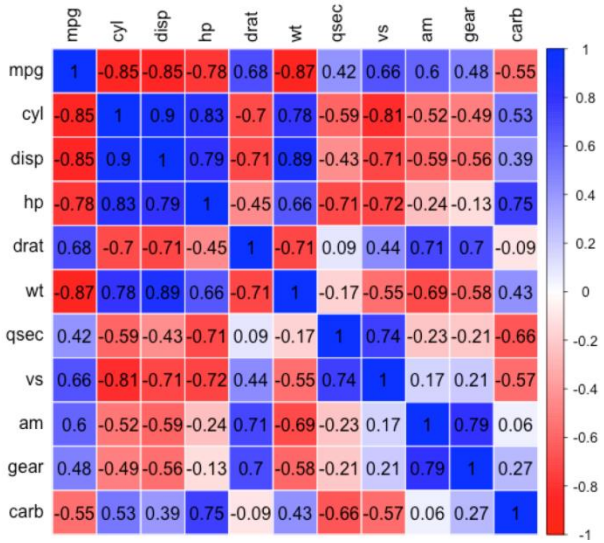


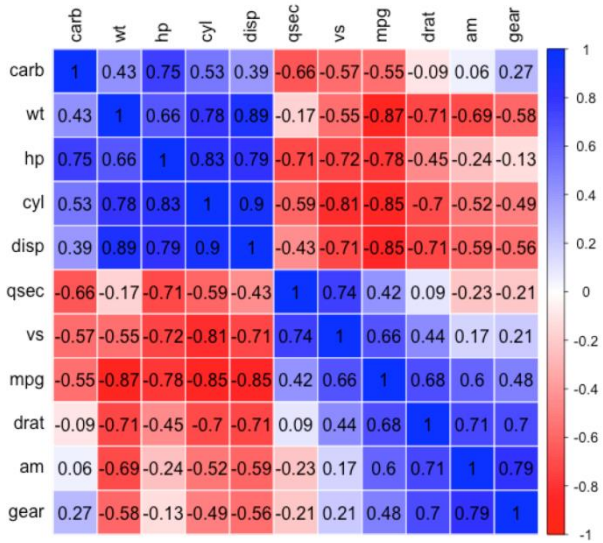


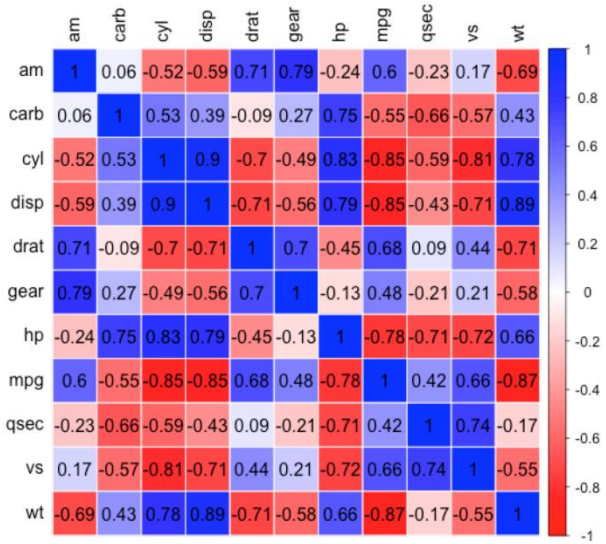
Correlation Analysis with R

- 변수간 순서 표현에 따른 히트맵 시각화

```
> corrplot(mcor, method="color", col = col3(200), tl.col="black", addCoef.col="black")
> corrplot(mcor, method="color", col = col3(200), tl.col="black", addCoef.col="black", order="hc
lust")
> corrplot(mcor, method="color", col = col3(200), tl.col="black", addCoef.col="black", order="al
phabet")
>
```







Correlation Analysis with R

- 제목 표현에 따른 히트맵 시각화

```
> corrplot(mcor, method="color", col = col3(200), tl.col="black", addCoef.col="black", order="alphabet", title="corrplot_alphabet")
```