Recipes 데이터 분석

강사 : 문성민

Text Mining about Novel

Text Mining about Novel

- 데이터 가져오기

```
> setwd("/Users/Seongmin_M/Desktop/Class")
>
> God_No_1 <- file("/Users/Seongmin_M/Desktop/Class/GOD_1.txt", blocking=F)
>
> txtLines <- readLines(God_No_1)
>
> close(God_No_1)
```

- 라이브러리 불러오기

```
> library(tm)
필요한 패키지를 로딩중입니다: NLP
> library(KoNLP)
```

- 불필요한 요소 삭제

```
> txtLines <- gsub("()", "", txtLines)
> txtLines <- gsub("<", "", txtLines)
> txtLines <- gsub(">", "", txtLines)
> txtLines <- gsub("[ |\t]{2,}", "", txtLines)
```

- 세종사전을 활용한 형태소 분석

```
> useSejongDic()
Backup was just finished!
87007 words were added to dic_user.txt.
>
> txtLines_Nouns <- sapply(txtLines, function(x) {paste(extractNoun(x), collapse = " ")})
```

- 분석 결과 확인(1)

```
> head(unlist(txtLines_Nouns))
```

"지은이 소개"
베르베르는 일곱 살 때부터 단편소설을 쓰기 시작한 타고난 글쟁이이다. 1961년 툴루스에서 태어나 법학을 전공하고 국립 언론 학교에서 저널리즘을 공부했다. 별들의 전쟁 세대에 속하기도 하는 그는 고등학교 때 만화와 시나리오에 탐닉하면서 만화 신문 유포리를 발행하였고, 이후 올더스 헉슬리와 H.G. 웰스를 사숙하면서 소설과 과학을 익혔다. 대학 졸업 후에는 르 누벨 옵세르바퇴르에서 저널리스트로 활동하면서 과학 잡지에 개미에 관한 평론을 발표해 오다가, 드디어 1991년 120여 회의 개작을 거친 개미를 발표, 전 세계 독자들을 사로잡으며 단숨에 주목받는 파랑스의 천재 작가로 떠올랐다.

"베르베르는 일곱 살 때 단편소설 시작 한 글쟁이 1961 년 툴루스에서 법학 전공 국립 언론 학교 저널리즘 공부 별들의 전쟁 세대 그 고등학교 때 만화 시나리오 탐닉 만화 신문 유포 리 발행 이후 올더스 헉슬리와 H G 웰스를 사숙 소설 과학 대학 졸업 후 르 누벨 옵세르바퇴르에서 저널리스트 활동 과학 잡지 개미 평론 발표 해 1991 년 120 회의 개작 개미 발표 전 세계 독자 들 주목 파랑 스 천재 작가"

Text Mining about Novel

- 분석 결과 확인(2)

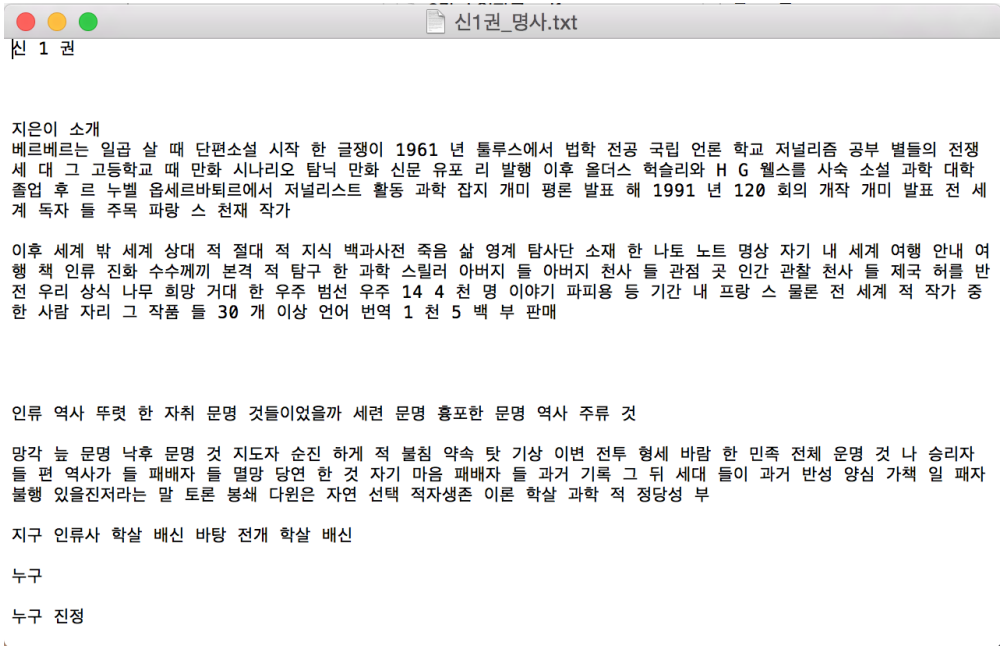
```
> head(unlist(txtLines_Nouns), 20)
```

이렇듯 지구의 인류사는 학살과 배신을 바탕으로 전개되었고, 그 학살과 배신은 잊혔다.

"지구 인류사 학살 배신 바탕 전개 학살 배신"

- 데이터 내보내기

```
> write(unlist(txtLines_Nouns),"/Users/Seongmin_M/Desktop/Class/신1권_명사.txt")
```



Text Mining about Novel

- 명사 합산 하기

```
> Nouns_wordcount <- table(unlist(txtLines_Nouns))
>
> length(Nouns_wordcount)
[1] 1338
```

- 분석 결과 내림차순 정렬

```
> head(sort(Nouns_wordcount, decreasing=T))
```

```
1425
에드몽 웰즈 상대 적 절대 적 지식 백과사전 제5권 (헤시오도스 신통 기 프랑시스 라조르박 글 근거 한 것
5
라울
3
나 무엇
2
누구
2
로디테
2
```

아프

Text Mining about Novel

- 분석 결과 오름차순 정렬

```
> head(sort(Nouns_wordcount,decreasing=F))

"미카엘 팽송 종아 142,857 호 빌라

1

"미카엘 팽송입니다

1

"아에덴 도성 일세 올림 피 자네 이름 뭐 내 말 자네 인간 시절 이름 뭐냐는

1

"어쨌거나 말 수 거 자네 그 눈길 낮춘다벌거숭이라는

1
"흔히들 나 디오니소스라 자 포도 나무 포도주 축제 음주 가무 방탕 따위 나 연결 그것 잘못 일세 것들은 나 진면목 거리 나 자유
 일세 통속 적 상상 체계 자유 의심 방탕 연결 되기 나 일세 저 자기 안 것 자유 나 방탕 한 신 그것 나 수

1

“ 작업 저 최선 몇 세 자멸 인류 창조 해 ” 헤르메스는 미소 우리 공중

1
```

Text Mining about Novel

- Corpus형태로 형변환

```
> GOD.text.Corpus <- Corpus(VectorSource(txtLines_Nouns))
>
```

- stopwords제거

```
> GOD.text.Corpus <- tm_map(GOD.text.Corpus, function(x)removeWords(x,stopwords()))
>
```

- TermDocumentMatrix를 사용하여 수치형으로 데이터 변환

```
> God_TDM_1 <- TermDocumentMatrix(GOD.text.Corpus, control = list(wordLengths = c(2, Inf)))
>
```

- 빈도수가 10 이상인 명사들 출력

```
> findFreqTerms(God_TDM_1, lowfreq = 10)
  [1] "“그들은"      "“나는"       "“내가"       "“이"
  [5] "“이제"        "142857"      "17"          "18"
  [9] "가슴"         "가운데"      "가정"        "가족"
 [13] "가지"         "각자"        "강둑"        "강력"
 [17] "강물"         "강의"        "개념"        "개미"
 [21] "거기"         "거대"        "거리"        "거울"
 [25] "거인"         "거지"        "건너편"      "건물"
 [29] "건설"         "게임"        "결과"        "결합"
 [33] "경우"         "경험"        "계속"        "고개"
 [37] "고대"         "고안"        "고통"        "곡선"
 [41] "공격"         "공기"        "공동체"      "공룡"
 [45] "공포"         "과거"        "과학"        "관심"
 [49] "관찰"         "광물"        "괴물"        "구름"
 [53] "구멍"         "구성"        "구체"        "궁전"
 [57] "귀스타브"     "규칙"        "그것"        "그녀"
 [61] "그다음"       "그들"        "그때"        "그리스"
```

- 가정과 관련된 명사 출력

```
> findAssocs(God_TDM_1, "가정", 0.25)
$가정
      불안케 파괴한다면(이제       핵무기       외계인         가증 
        0.53         0.53         0.53         0.48         0.38 
        과오         도리       메시지         아찔         오버 
        0.38         0.38         0.38         0.38         0.38 
        오염         외계       책임감         실패       생명체 
        0.38         0.38         0.38         0.34         0.28 
    “수사는     “피해자들         1830           33         가난 
        0.27         0.27         0.27         0.27         0.27 
        가스         건강         결핵         고뇌         공로 
        0.27         0.27         0.27         0.27         0.27 
        궁녀         권세         꿈속         농부       덩어리 
        0.27         0.27         0.27         0.27         0.27 
  데메테르는         도래     드루이드교         막연         만약 
        0.27         0.27         0.27         0.27         0.27 
        목숨       무용수         무희 백파이프를         벤치 
        0.27         0.27         0.27         0.27         0.27 
```

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)

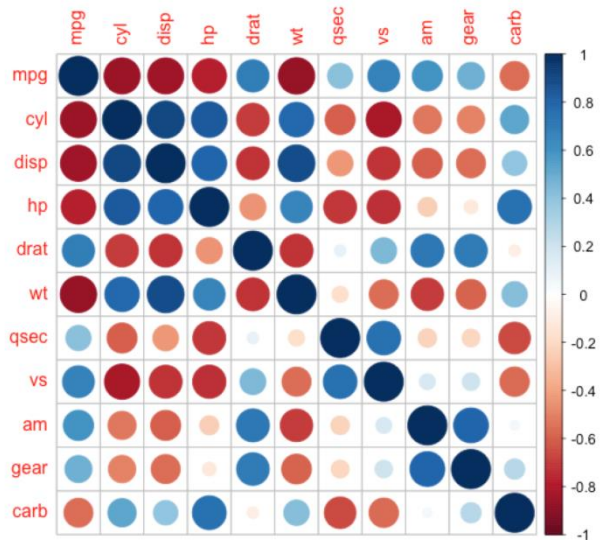
```
> install.packages("corrplot")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/corrplot_0.73.tgz'을 시도합니다
Content type 'application/x-gzip' length 2679598 bytes (2.6 Mb)
URL을 열었습니다
==================================================
downloaded 2.6 Mb


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)
```

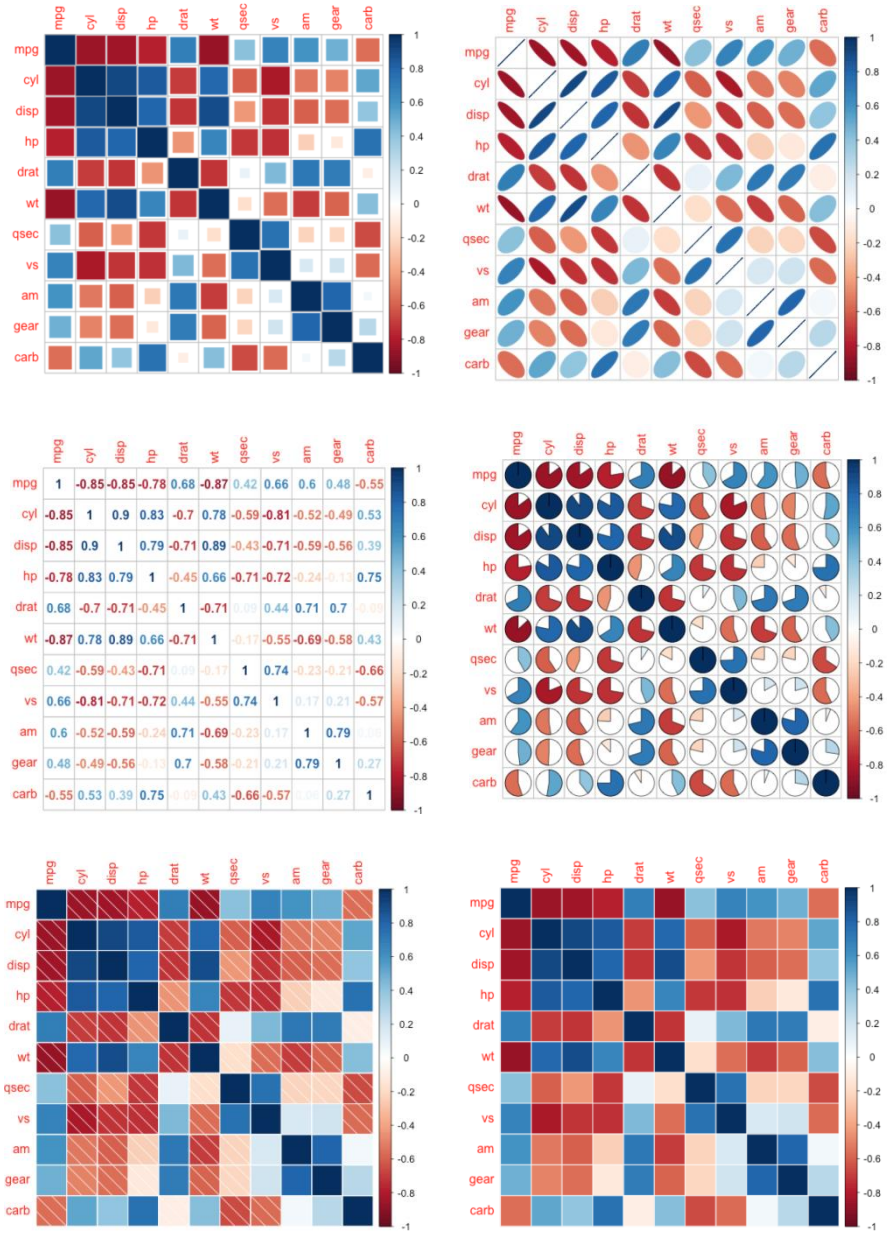


- 디폴트 값으로 설정되어 있는 히트맵은 위의 상관 행렬 표를 원으로 표시하고 계수의 크기를 원의 크기로 표현하여 각 변수의 관계를 손쉽게 확인 할 수 있다.

Correlation Analysis with R

- Method에 따른 히트맵 시각화

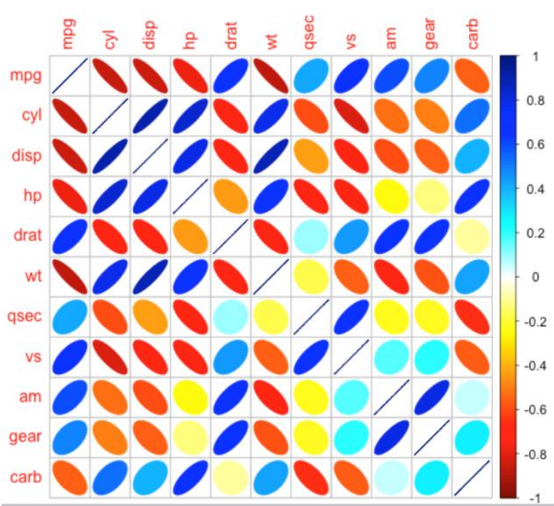
```
> corrplot(mcor, method="square")
>
> corrplot(mcor, method="ellipse")
>
> corrplot(mcor, method="number")
>
> corrplot(mcor, method="pie")
>
> corrplot(mcor, method="shade")
>
> corrplot(mcor, method="color")
```

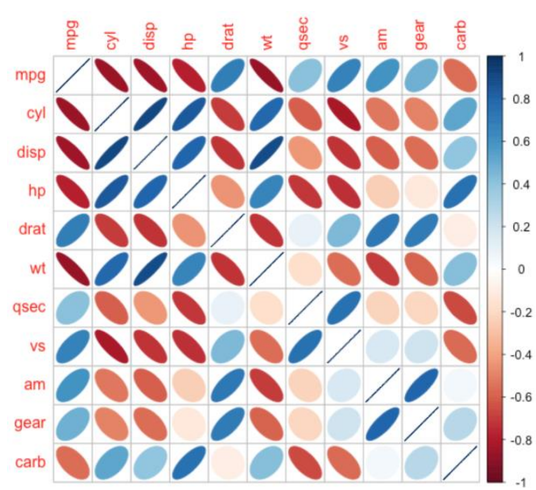


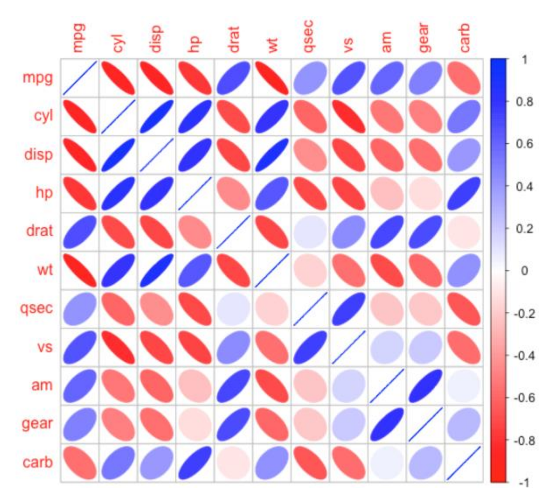
Correlation Analysis with R

- 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)
```



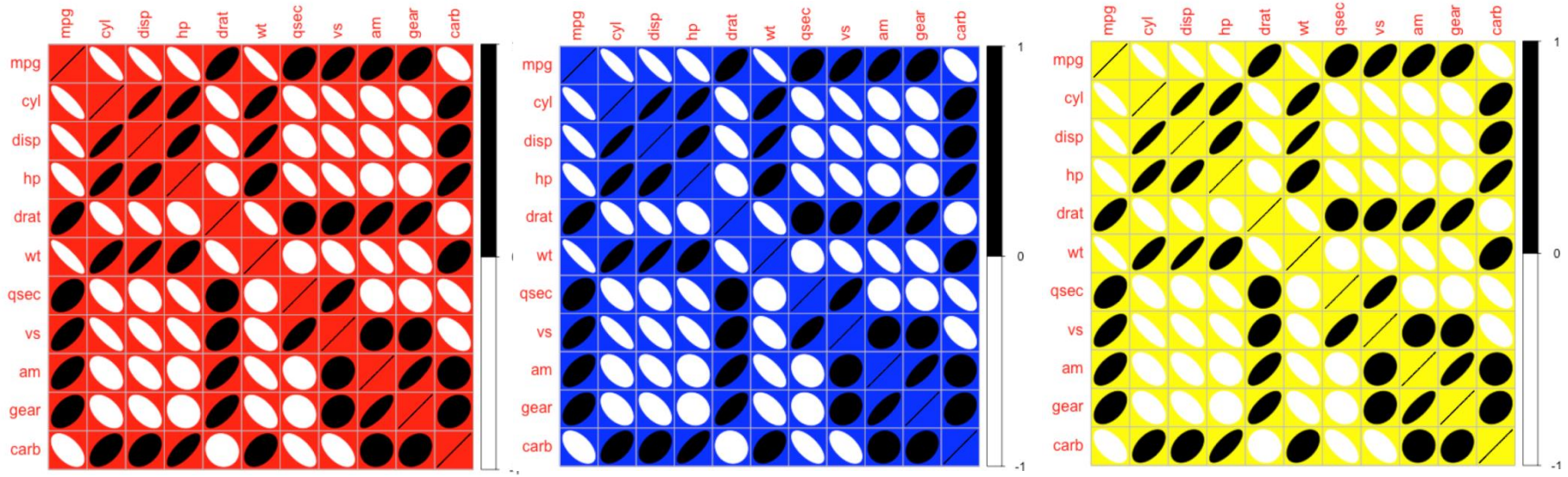




Correlation Analysis with R

- 배경에 따른 히트맵 시각화

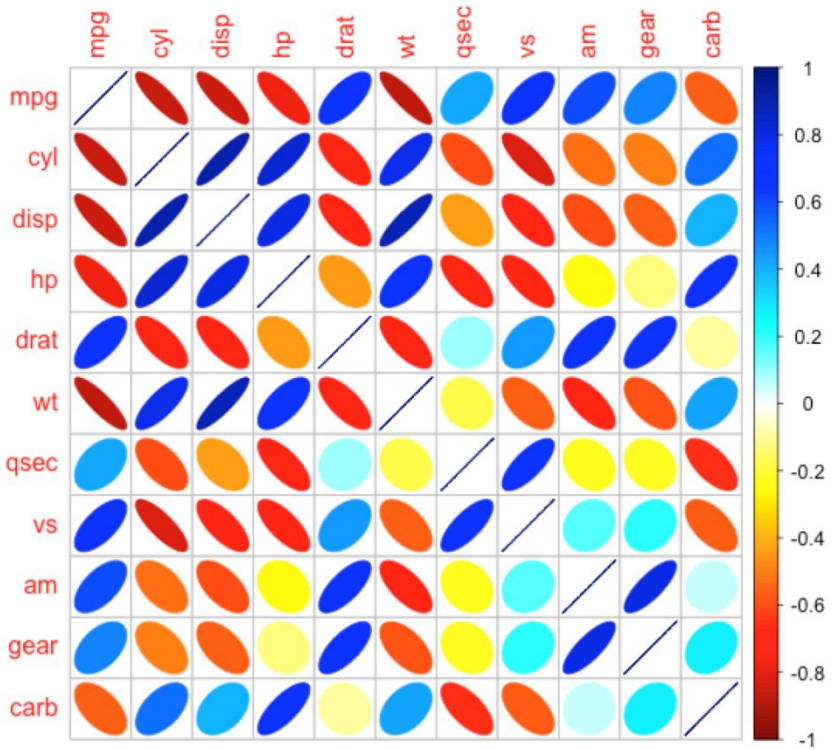
```
> corrplot(mcor, method="ellipse", col = wb, bg = "red")
> corrplot(mcor, method="ellipse", col = wb, bg = "blue")
> corrplot(mcor, method="ellipse", col = wb, bg = "yellow")
```

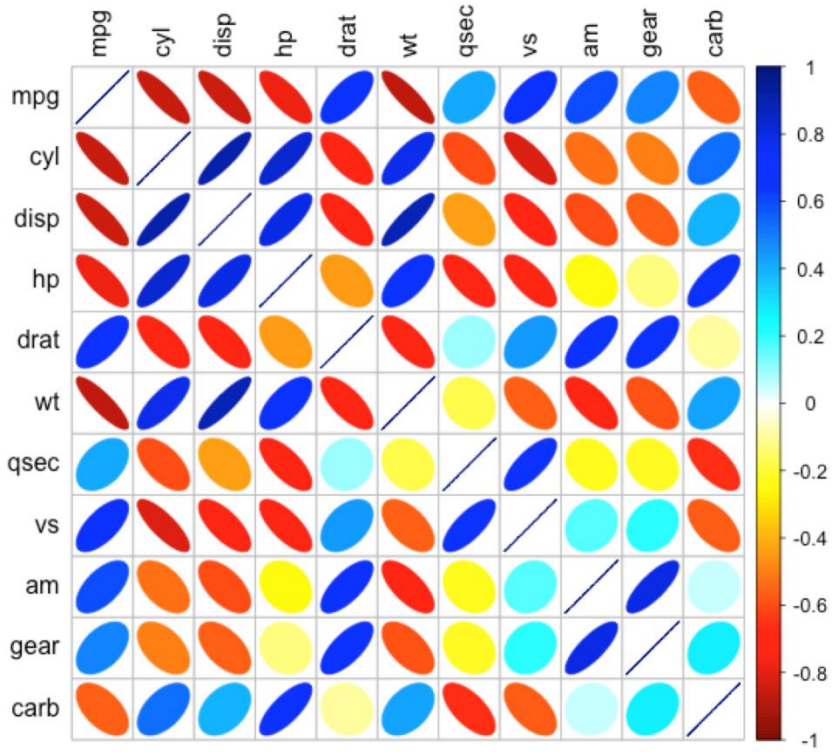


Correlation Analysis with R

- 변수명 색상에 따른 히트맵 시각화

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

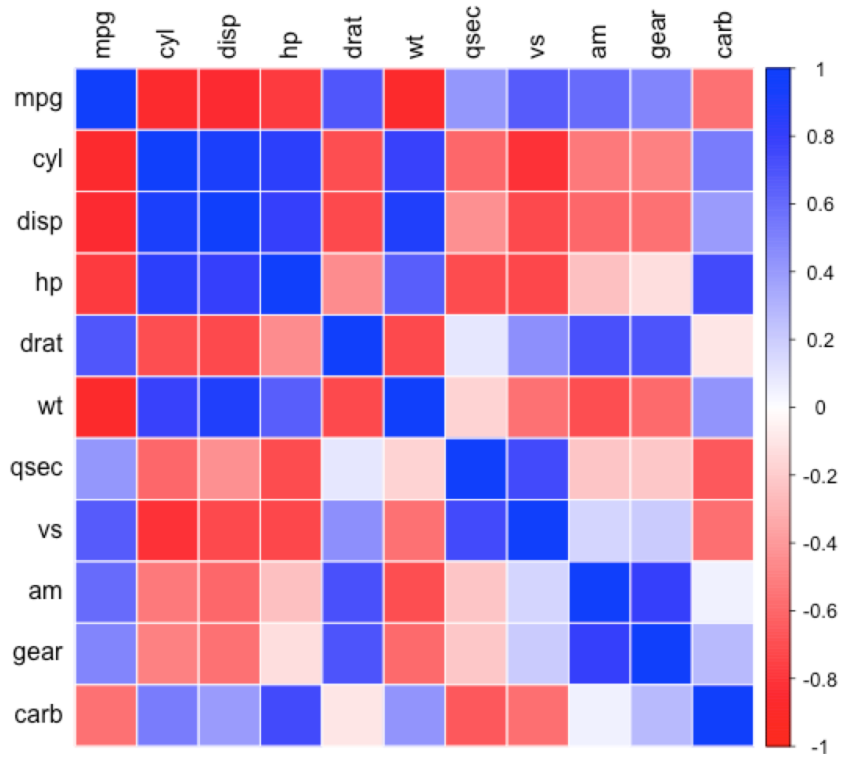


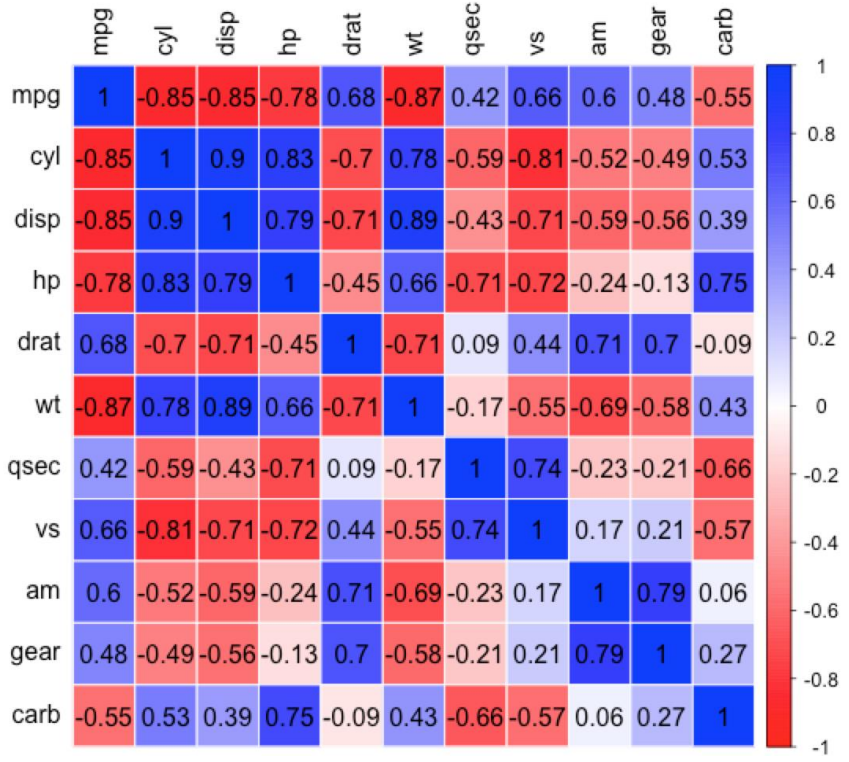


Correlation Analysis with R

- 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")
```

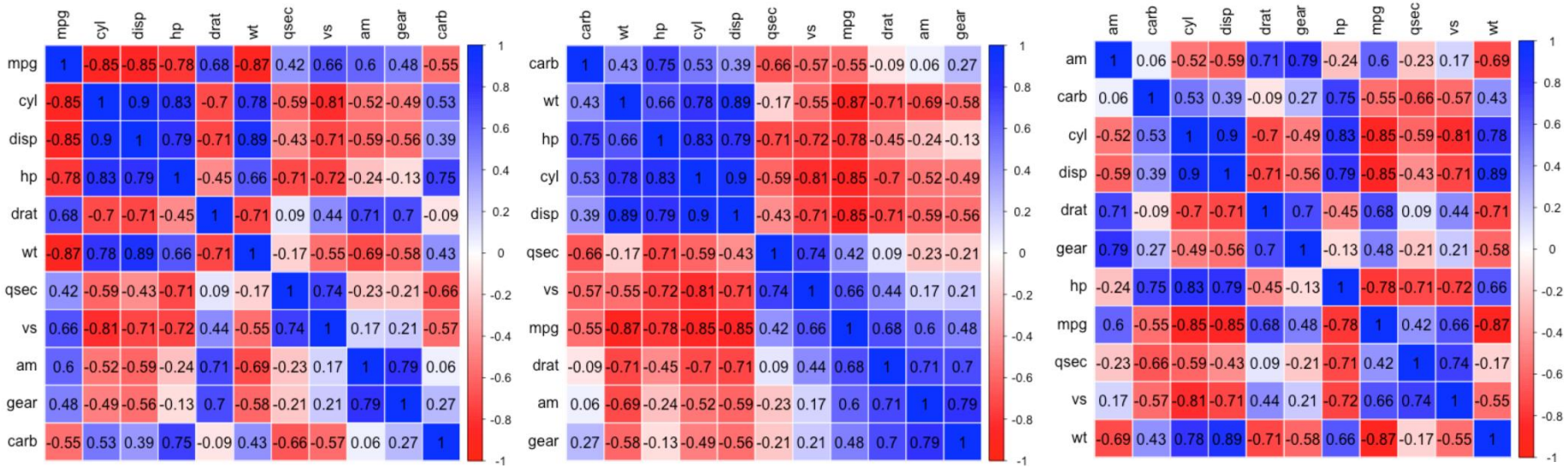




Correlation Analysis with R

- 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")
```