2022학년도 1학기 기말시험(과제물)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **교과목명** | : | 데이터시각화 |
|  | **학번** | : | 202135-368864 |
|  | **성명** | : | 홍원표 |
|  | **연락처** | : | 010-5343-4341 |

1. 런던 콜레라지도(교재 p.4)를 조사하고 데이터 시각화의 필요성을 설명하시오.(7점)

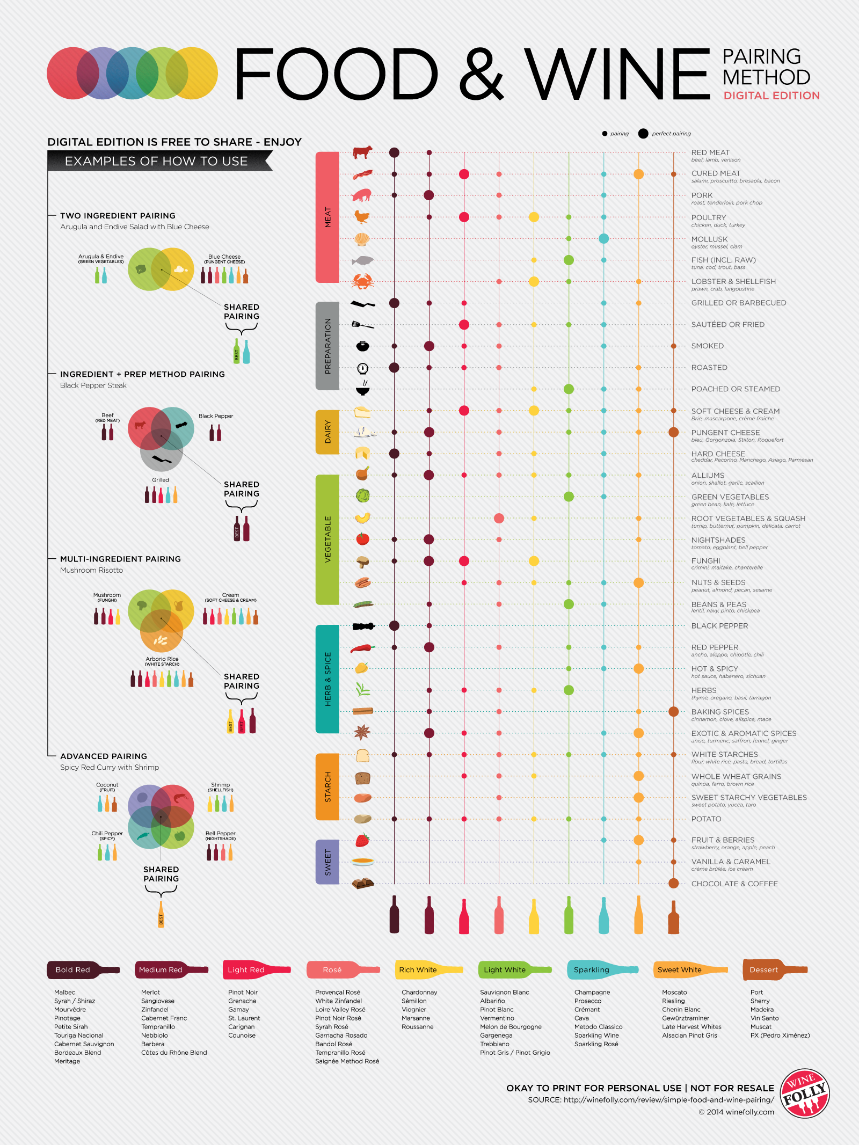
1854년 영국 런던 소호지구에서 콜레라에 걸린 환자들이 급증했다. 그때까지 콜레라는 공기로 옮겨진다고 알려져 있었다. 의사 존 스노(John Snow)는 찾아오는 환자 증상을 보면 `공기를 통한 전염`이라고 보기는 어려웠다. 공기를 통한 전염은 기관지 쪽에 증상이 있어야 하는데 내장 쪽에 증상을 보이는 환자들이 많았고 대부분 심한 설사를 동반하고 있었다. 영국 전염병 학자인 윌리엄 파(William Farr)가 사망자의 규모와 상황을 인터뷰한 기사를 보고 윌리엄 파가 사용한 감염에 따른 사망자 수와 주소 데이터를 확보하고 런던 지도를 꺼내 감염자의 주소를 찾고 검은 사각형을 표시했다. 스노는 사각형이 표시된 지도에서 검은 사각형이 브로드 가(Broad Street) 지역을 중심으로 분포해 있었고, 그 중앙에 공용펌프가 있었다. 하지만 일부 사망자의 부근에는 다른 공용펌프가 더 가깝기 때문에 스노는 사망자가 표시된 주소를 찾아가서 어느 공용펌프에서 물을 길어서 먹는지 확인하였고 사망자 모두 브로드 가에 있는 공용펌프에서 물을 떠다 마시는 것을 확인하였다.

당시 공용펌프는 시에서 관리하고 있어서 스노는 사망자가 표시된 지도와 윌리엄 파의 사망자와 주소 데이터를 들고 시청 담당자를 찾아가 설명하였고 담당 공무원은 스노의 의견대로 펌프 손잡이를 제거하였다. 이후 콜레라 환자도 급격히 줄어들었다.

윌리엄 파와 존 스노 둘 다 같은 데이터를 사용하여 분석하였으나 윌리엄 파는 공기를 통한 전염이라고 추측하였기 때문에 온도와 관련된 그래프를 보여주었고, 존 스노는 환자의 상태를 보고 음식물과 관련이 있음을 추측해 사망자의 분포를 지도에 표시하였다.

이처럼 동일한 데이터지만 어떤 방식으로 적용하는가에 따라 다른 결과를 보여주고 그것은 개인이 데이터를 바라보는 경험과 관념의 차이 달라지기 때문에 데이터를 바라보는 안목이 중요함을 알 수 있다.

2. 아름답고 의미 있는 데이터 시각화 사례를 발굴하고 그 이유를 정리하시오.(7점)



<https://winefolly.com/review/simple-food-and-wine-pairing/>

2022년 최고의 테이블 시각화로 뽑힌 사례입니다.

Rutger's University의 감각 생물학자인 Dr. Paul Breslin박사는 최근 연구에서 기름기와 떫은 맛이 어떻게 상호작용하는 지 초점을 맞추고 기름진 음식을 먹고 물로 입을 헹구면 기름진 느낌이 사라지지 않지만 와인과 같은 음료에 들어있는 타닌과 산도는 우리 혀에서 단백질을 분비해서 이러한 느끼함을 상쇄시킨다고 합니다. 하지만 실제로 타닌과 지방은 서로 잘 상쇄되지 않기 때문에 연어와 같이 기름진 생성은 레드 와인과 잘 어울리는 것 같이 생각되기도 하지만 와인의 타닌과 생선의 기름기가 서로 상쇄되어 비린 맛이 남는다고 합니다. 생선은 높은 산도의 와인과 잘 어울린다고 합니다. 샴페인과 같은 강한 풍미의 와인이 다양한 종률의 음식과 잘 어울릴 수도 있는 이유일 수도 있다고 합니다.

이 테이블은 음식에 따라 어떠한 와인이 가장 잘 어울리는지 나타내는 테이블로 음식에 따른 와인을 선택할 때 쉽게 참조할 수 있도록 만들었습니다.

이 테이블을 통해서 단맛, 신맛, 매운맛, 쓴맛, 느끼한 맛과 같은 요소가 어떻게 조화를 이루는지에 대한 기초를 배울 수 있도록 배치하였고 음식에 따라 와인의 산도, 탄닌, 알코올 도수 및 단맛의 특성에 맞춰 사용하기 쉽게 배치하였다고 합니다.

과학적인 연구의 내용을 일반인들이 쉽게 받아들일 수 있도록 만든 사례라고 생각합니다.

3. R 패키지 “vcd”에 내장된 “Arthritis” 데이터셋은 류마티스 관절염 환자를 대상으로 한 임상시험 결과 데이터이다. 각 행은 각 환자를 나타내며, 변수 Treatment는 그룹 (Treated = 새로운 치료제를 투약한 그룹, Placebo = 위약을 받은 그룹)을 나타낸다. 변수 Sex는 성별을, Improved는 치료 결과(None = 차도 없음, Some = 약간 좋아짐, Marked = 매우 좋아짐)를 나타낸다. 새로운 치료제 투약 여부가 치료 결과와 연관이 있는지, 성별과 치료 결과 간에 연관이 있는지를 데이터 시각화를 통해서 탐구하시오. (18점)

> library(vcd)

> # 기술통계 확인

> summary(Arthritis)

ID          Treatment      Sex          Age          Improved

Min.   : 1.00   Placebo:43   Female:59   Min.   :23.00   None  :42

1st Qu.:21.75   Treated:41   Male  :25   1st Qu.:46.00   Some  :14

Median :42.50                            Median :57.00   Marked:28

Mean   :42.50                            Mean   :53.36

3rd Qu.:63.25                            3rd Qu.:63.00

Max.   :84.00                            Max.   :74.00

임상시험 참여자 정보는 위약을 치료받은 그룹은 43명이고 새로운 치료제를 투약한 그룹은 41명이다. 성별로 보면 여자는 59명이고 남자는 25명이다. 치료 후 차도가 거의 없었던 참여자수는 42명이고 약간 좋아진 참여자수는 14명 매우 좋아진 참여자는 28명이다.

> # 분할표 확인

> (xtabs(~ Treatment + Improved, data = Arthritis))

Improved

Treatment None Some Marked

Placebo   29    7      7

Treated   13    7     21

분할표를 확인해 보면 위약을 투약한 그룹에서는 차도가 거의 없었던 참여자는 29명이고 약간 좋아졌거나 매우 좋아진 참여자는 14명인 반면에 새로운 치료제를 투약한 그룹에서는 차도가 거의 없었던 참여자는 13명 약간 좋아졌거나 매우 좋아진 참여자는 36명이나 됐다.

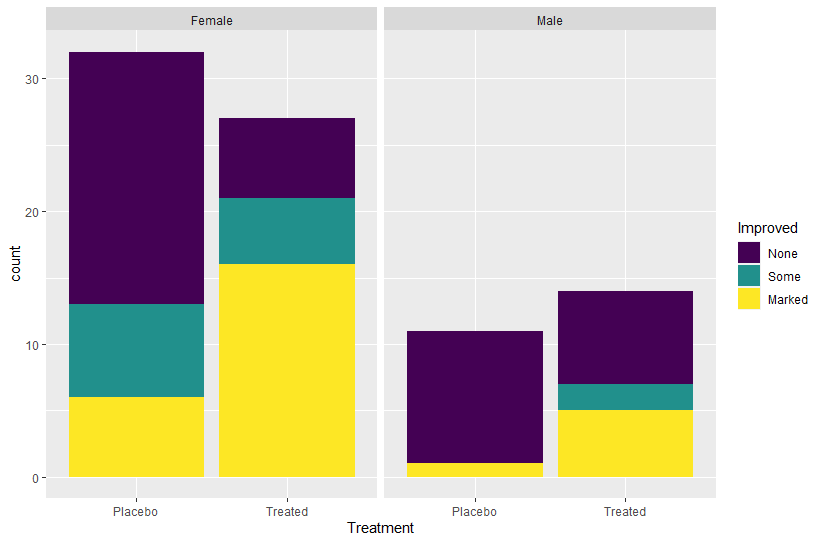
ibrary(ggplot2)

# 치료에 따른 향상을 성별로 분할하여 막대그래프로 그린다.

ggplot(data=Arthritis) +

  geom\_bar(mapping=aes(x = Treatment, fill = Improved))+

facet\_wrap(~ Sex)

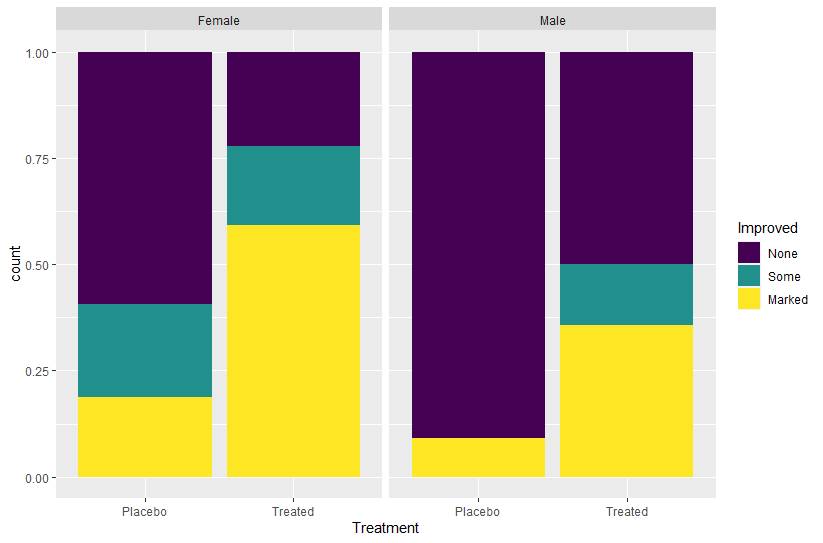


# 남성과 여성의 참여자 수가 다르기 때문에 그래프를 화면에 채워서서 표시하면

ggplot(data=Arthritis) +

  geom\_bar(mapping=aes(x = Treatment, fill = Improved), position="fill")+

  facet\_wrap(~ Sex)



막대그래프를 보면 여성이 남성보다 새로운 치료제를 투약하든 위약을 투약하든 남성에 비해 증상이 개선되는 경향을 볼 수 있다. 그러나 이 차이는 유의미한 확실한 차이는 아니라고 판단된다.

4. R에 내장된 “airquality” 데이터셋은 1973년 5월부터 9월까지 뉴욕의 대기질에 관한 데이터셋이다. 변수 Ozone은 대기 중 오존의 양, Solar.R은 태양방사선의 양, Wind는 풍속, Temp는 기온을 나타낸다. 이 네가지 변수(Ozone, Solar.R, Wind, Temp)에 대한 산점도 행렬을 그리고, 이 산점도 행렬에서 알 수 있는 변수들 간의 관계에 대하여 서술하시오. (18점)

# 먼저 NA 값들을 빼준다.

air<-na.omit(airquality)

attach(air)

## put histograms on the diagonal

panel.hist <- function(x, ...)

{

  usr <- par("usr"); on.exit(par(usr))

  par(usr = c(usr[1:2], 0, 1.5) )

  h <- hist(x, plot = FALSE)

  breaks <- h$breaks; nB <- length(breaks)

  y <- h$counts; y <- y/max(y)

  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)

}

## put (absolute) correlations on the upper panels,

## with size proportional to the correlations.

panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)

{

  usr <- par("usr"); on.exit(par(usr))

  par(usr = c(0, 1, 0, 1))

  r <- abs(cor(x, y))

  txt <- format(c(r, 0.123456789), digits = digits)[1]

  txt <- paste0(prefix, txt)

  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)

  text(0.5, 0.5, txt, cex = cex.cor \* r)

}

pairs(~Ozone+Solar.R+Wind+Temp

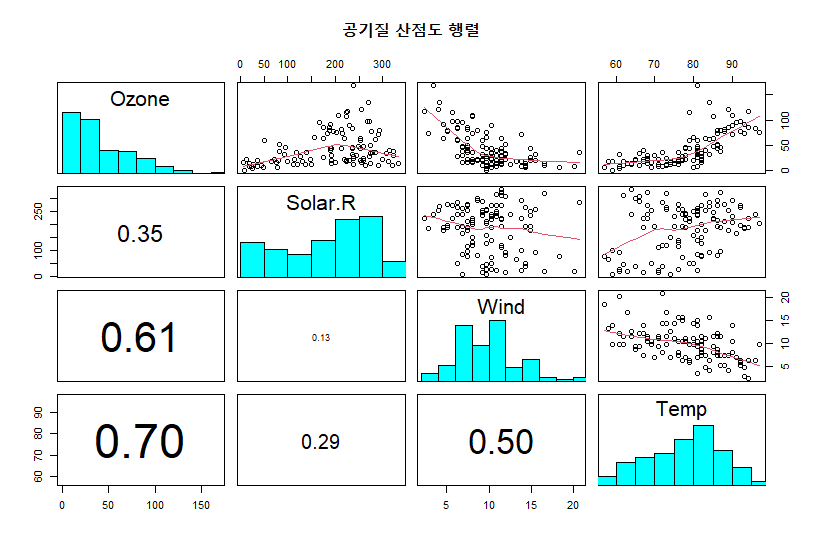
      , data=air # NA를 제거한 자료를 사용한다.

      , panel = panel.smooth # 패널에 추세선을 추가

      , diag.panel = panel.hist # 대각선에 히스토그램을 그리는 옵션

      , lower.panel = panel.cor # 하단패널에 상관계수를 출력

      , main = "공기질 산점도 행렬")



그래프의 상관계수 및 histogram 출력 함수들은 도움말(?pairs)의 샘플에 나오는 함수를 그대로 사용하였다.

공기질 산점도 행렬을 출력할 때, 행렬의 대각선 아래쪽에는 각 변수들간의 상관계수를 출력하여 쉽게 확인할 수 있도록 하였다. 또한 데이터에는 NA값이 포함되어 있어 상관계수를 츨력되지 않아서 na.omit() 함수를 이용해 NA값을 제거하였다.

변수들 간의 관계를 살펴보면 오존과 온도가 가장 높은 상관관계가 있다고 볼 수 있고 바람과 온도, 바람과 오존과의 상관계수도 0.5보다 같거나 높아서 약한 상관관계가 있는 것으로 보인다. 하지만 태양방사선의 양과 다른 변수들 간의 상관계수는 0.5보다 작은 값이기 때문에 상관관계는 없는 것으로 볼 수 있다.

-- 데이터시각화 기말과제물 끝 --