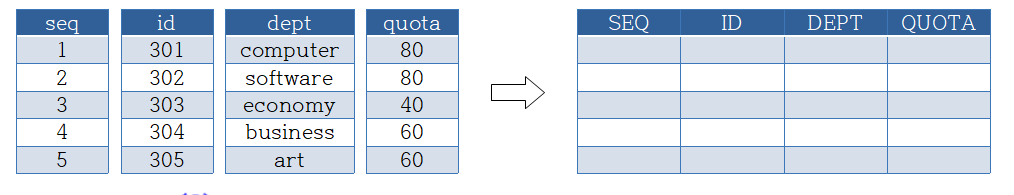
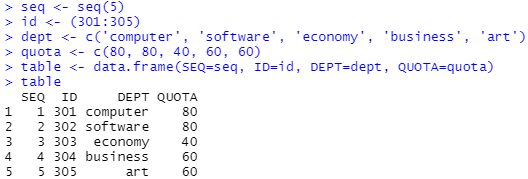
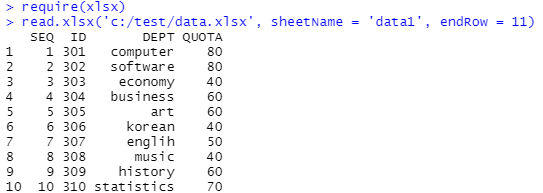
**[빅데이터 소프트웨어] - 2018301034 서 용 주**

**[1장. 연습과제]**

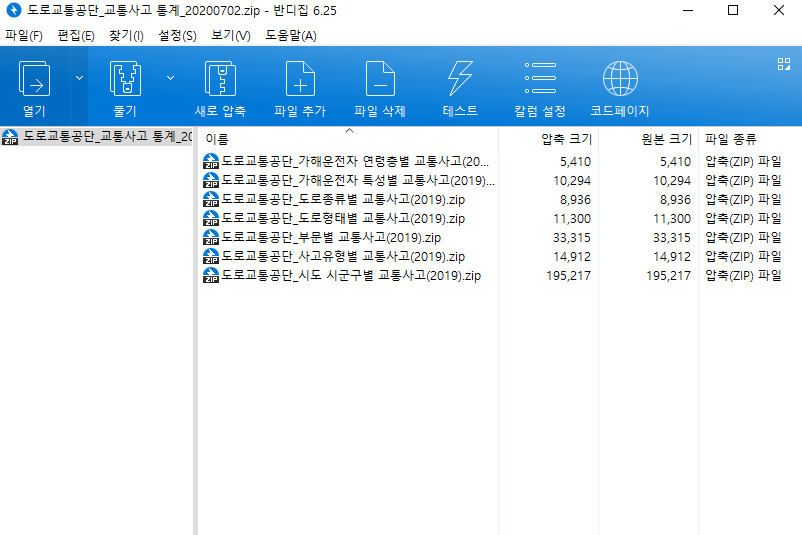
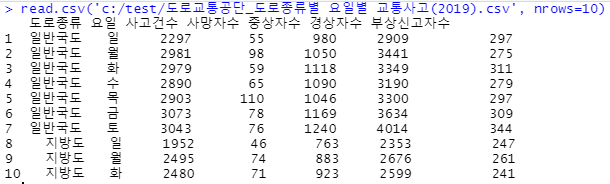
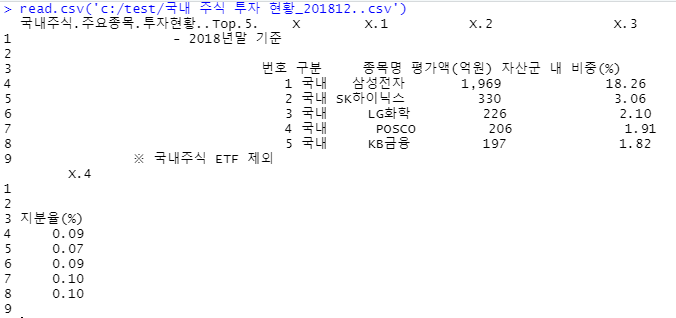
**5번.** 다음의 벡터를 만들고 데이터프레임으로 변환하여 결과를 프린트하여라.



**11번.** 저장되어 있는 엑셀 파일을 읽고 10개의 행만 화면에 프린트하여라.

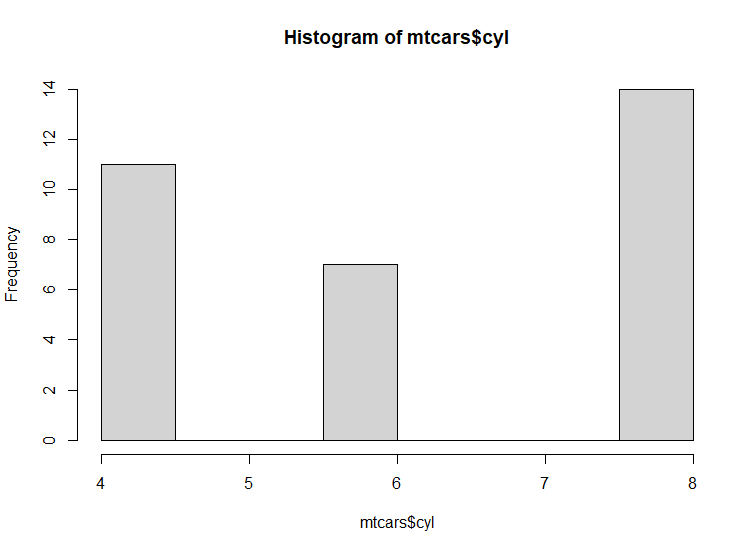


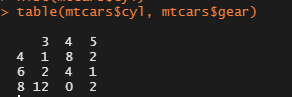
**[1장. 실습과제 1번]**

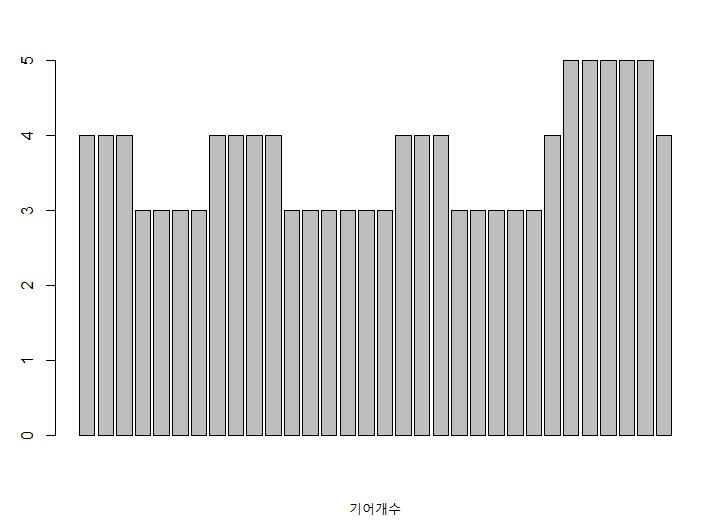


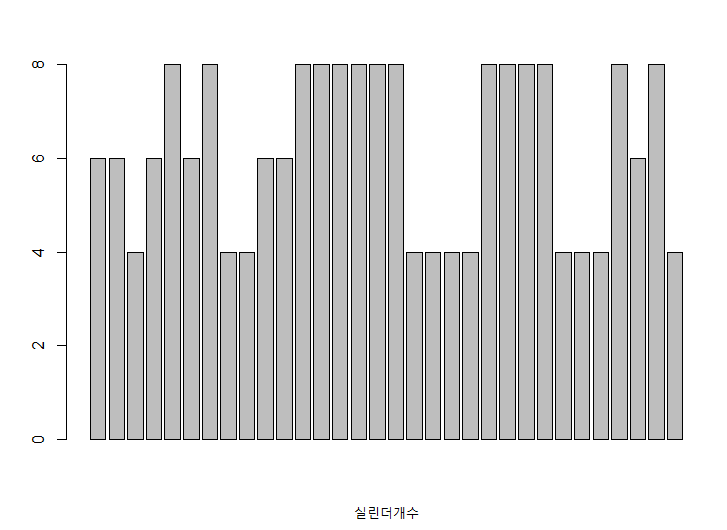
[2장. 실습과제]

1번. 실린더 개수별 히스토그램을 작성하여라.

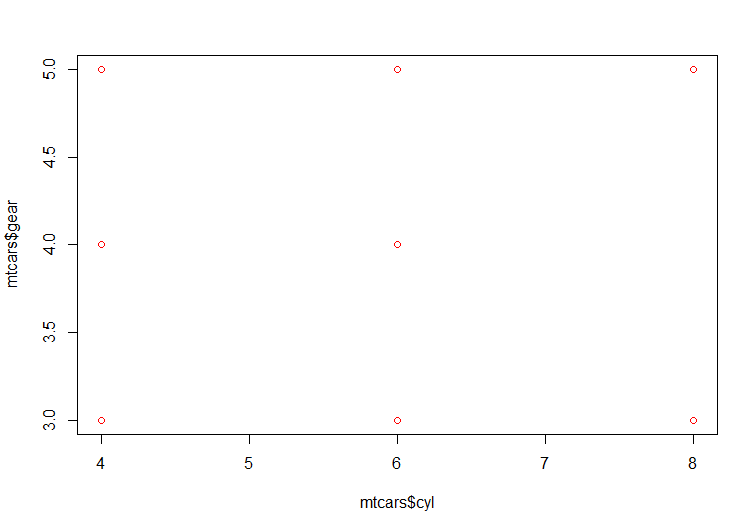
2번. 실린더 개수와 기어 개수의 교차표를 작성하여라.



3번. 실린더 개수와 기어 개수의 도수분포표를 작성하여라



4번. 실린더 개수와 기어 개수의 산포도를 작성하여라

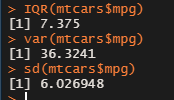


5번. 연비의 중심위치는 얼마인가? 평균과 중앙값을 비교하여 설명하여라.

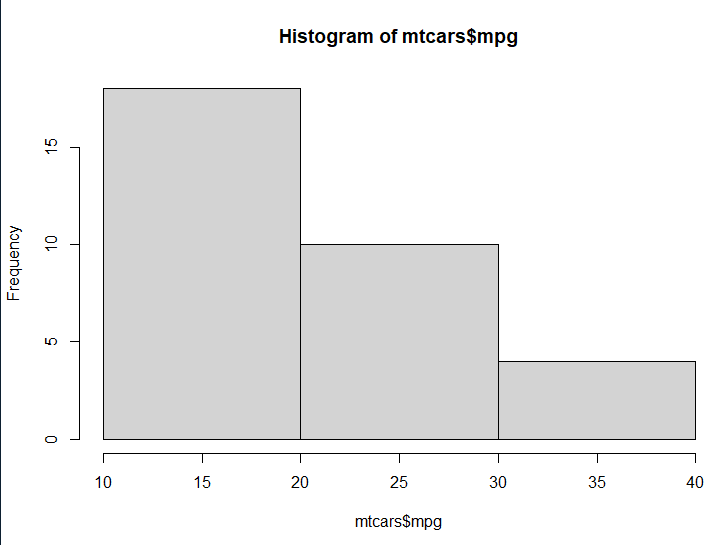
고르게 분포되어있어 비슷하다



6번. 연비의 사분위수 범위와 분산, 표준편차를 구하여라.



7번. 연비를 10구간으로 나누어 히스토그램으로 분포모양을 그려라. 그리고 왜도도 구하라.



8번. 연비를 기준으로 체비세프 정리를 만족하는지 확률로 설명해라

-체비 세프 정리는 모든 자료에 적용된다. 평균 20.09와 2표준편차(2\*6.02) 범위 내에 적어도 75%가 존재한다.

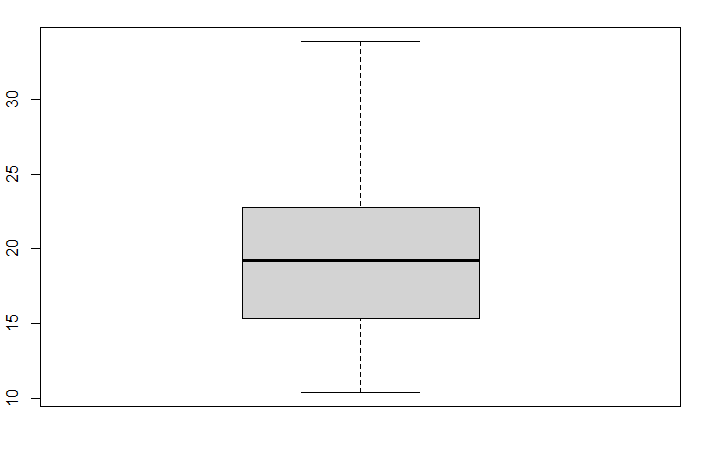
9번. 연비를 기준으로 경험법칙이 만족하는지 확률로 설명해라.

해당 자료는 정규분포가 아니기 때문에 적용하기 힘들다.

10번. 연비에 이상값을 검출하라. 어느 회사의 자동차가 이상값인가?

이상값이 따로 추출되지 않았다.

11번. 연비에 대한 5개 숫자 요약값은 ?

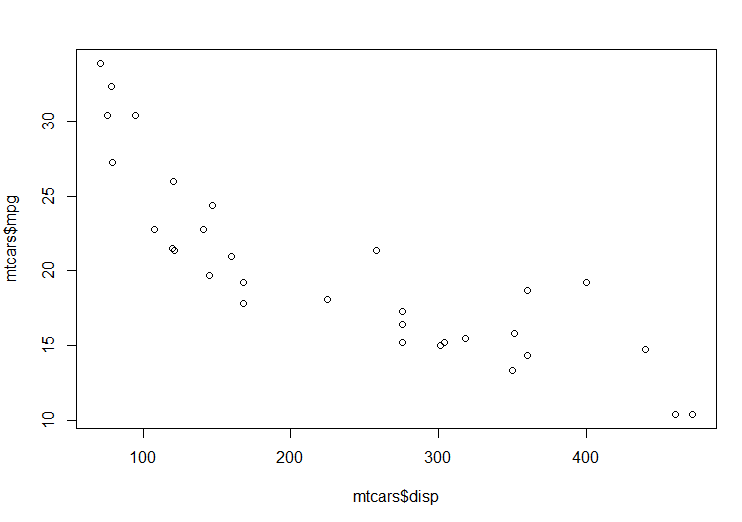
12번. 연비에 대해 박스플롯을 작성하라

13번. 연비에 대해 분포모양과 박스플롯을 비교 설명하라

박스 플롯이 더 직관적으로 볼 수 있다는 점에서 장점이 존재한다.

14번. 연비와 배기량의 공분산과 상관계수를 구하라. 상관계수 값으로 무엇을 알 수 있나?

 Cov 가 (-)값으로 크게 갖고 있다. 강한 음의 상관관계를 가질 것으로 보인다.



**[3장 연습문제]**

**6번.** mtcars 데이터셋에서 기어의 개수에 따르는 박스플롯을 작성하여라.

그리고 각 분포를 비교 설명하여라.

# gear를 명목변수로 변환

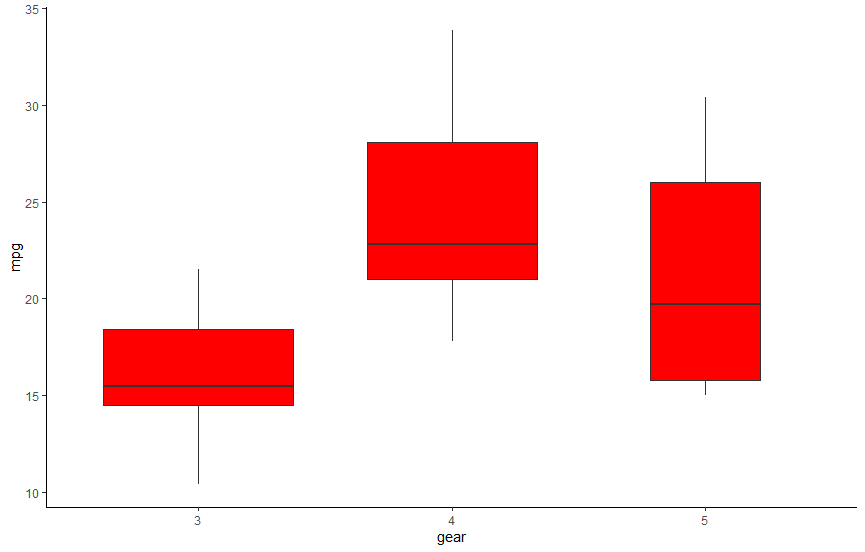
mtcars$gear <- factor(as.numeric(mtcars$gear))

# 박스플롯 그래프 그리기

ggplot(mtcars, aes(x = gear, y = mpg)) +

geom\_boxplot(varwidth = T, fill = "red")

기어개수 별 박스플롯 그래프



기어 수가 3개인 자동차는 연비(1갤런당 마일 수, mpg)가 매우 낮다.

4개와 5개인 경우를 비교할 때엔 4개의 연비가 5개인 경우보다 훨씬 높다.

따라서 연비효율은 기어 수가 4개인 자동차라 할 수 있다.

**8번.** airmiles 데이터셋을 text 파일로 저장하고 읽어서,

date형 변수를 추가하여 시계열 그래프를 그려라.

단 x축 변수 값들을 Dec. 1937 이런 방식으로 표현하여라.

# lubridate 패키지 불러오기

library(lubridate)

# text파일 생성

airmiles

write.csv(airmiles, "C:/test/airmiles.txt")

airmiles2 <- read.csv("C:/test/airmiles.txt")

airmiles2

# 필드명 재설정

names(airmiles2) <- c("year", "airmiles")

airmiles2

# year, date 변수 생성

airmiles2$year <- airmiles2$year + 1936

airmiles2$year <- as.character(airmiles2$year)

airmiles2$date <- as.Date(airmiles2$year, format = "%Y")

airmiles2

# 날짜객체 생성

year2 <- paste(month.abb[month(airmiles2$date)], sep = " ", lubridate::year(airmiles2$date))

year2

# 그래프 그리기

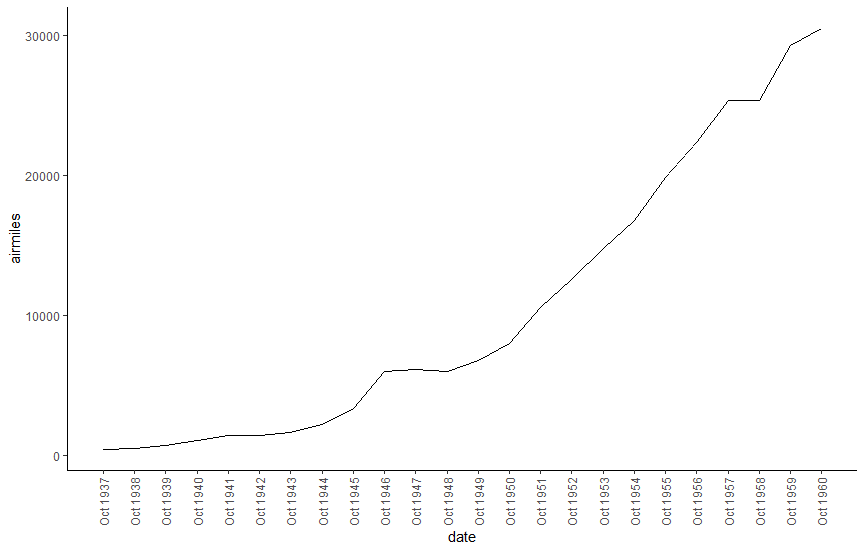
ggplot(airmiles2, aes(x = date)) +

geom\_line(aes(y = airmiles)) +

scale\_x\_date(labels = year2,

breaks = airmiles2$date) +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5))



**[4장 연습문제]**

**3번.** 로또에 대한 다음의 질문들을 R로 계산하여라.

**3 – 1)** 각 순위에 당첨될 확률을 구하여라.

# 1등

num1 <- 1 / (factorial(45) / factorial(6) \* factorial(39)) # 45 C 6 분의 1

2.950771e-100

# 2등

num2 <- 1 / (factorial(45) / factorial(5) \* factorial(40)) \* 39

4.795003e-101

# 3등

num3 <- 1 / (factorial(45) / factorial(5) \* factorial(40))

1.229488e-102

# 4등

num4 <-1 / (factorial(45) / factorial(4) \* factorial(41))

5.997503e-105

# 5등

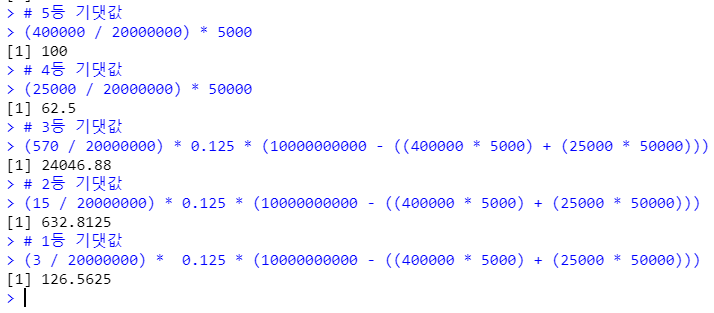
num5 <- 1 / (factorial(45) / factorial(3) \* factorial(42))

3.569942e-107

**3 – 2)** 복권 한 장당 기대 당첨금은 얼마인가?

500원

**3 – 3)** 각 순위의 기댓값은 얼마인가?



**3 – 4)** 이러한 구조에서 로또를 구매하는 것이 합리적인 행동인가?

전혀 바람직하지 않다.

1. 각 순위마다의 기댓값이 들쑥날쑥하기 때문이다.

2. 전체적으로 5000원을 투자해 500원을 기대하는 게임이기 때문에 손해보는 게임이나 마찬가지이다.

**6번.** 다음 질문들을 R로 계산하여라.

|  |  |  |  |
| --- | --- | --- | --- |
| 가입 서비스 | 남성 | 여성 | 계 |
| 웹 서비스 | 1,000,000 | 500,000 | 1,500,000 |
| 앱 서비스 | 800,000 | 1,200,000 | 2,000,000 |
|  | 2,300,000 | 2,200,000 | 4,500,000 |

**# 6-1.)** 가입자가 남성일 확률은?

2300000 / 4500000

0.5111111

**# 6-2.)** 여성가입자일 때, 동영상 서비스 가입자일 확률은?

500000 / 2200000

0.2272727

**# 6-3.)** 앱 서비스 가입자 중 남성일 확률은?

800000 / 2000000

0.4

**# 6-4.)** 여성이고 동영상 서비스 가입자일 확률은?

(2200000 / 4500000) \* (500000 / 1000000)

0.2444444

**10번**

**# 10-1.)**

0.5

**# 10-2.)**

0.8

**# 10-3.)**

(0.5 \* 0.8) / ((0.5 \* 0.8) + (0.5 \* 0.5))

0.6153846