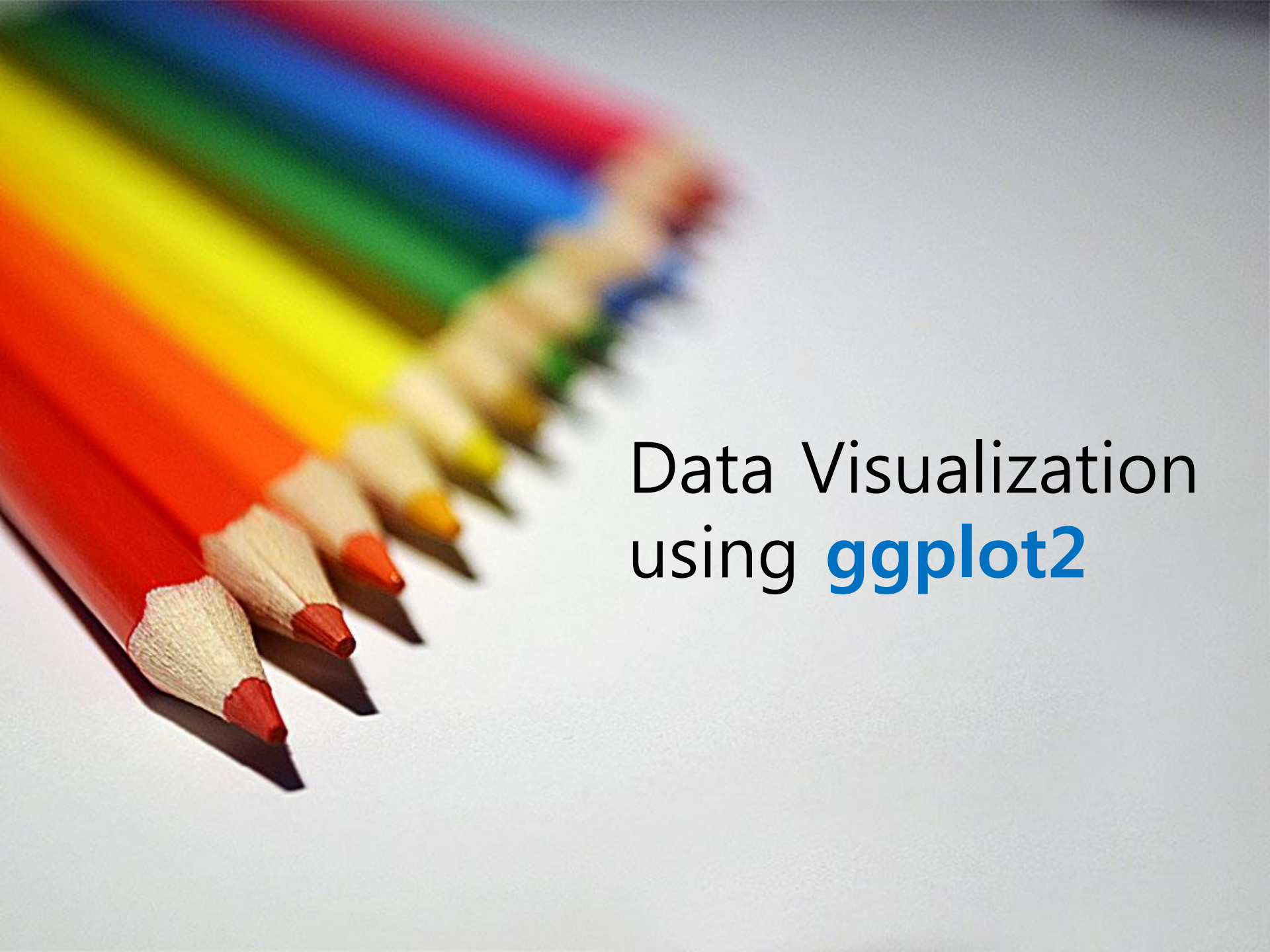

Data Visualization
using ggplot2

About the ggplot2 Package

- Grammar of Graphics의 구현체
  - 미적 매핑(aesthetic mapping)
  - 통계적 변환(stat)
  - 기하객체 적용(geom)
  - 위치 조정(position adjustment)

- ggplot2 설치
  - install.packages("ggplot2")
  - library(ggplot2)

- ggplot2 사용 문법
  - **ggplot(data= , aes(x= , y=)) + geom_*() + . . .**

- ggplot2 참조 사이트
  - http://docs.ggplot2.org/current/

ggplot2 사용 형식

ggplot(dataframe,aes(x=x축 데이터 , y=y축 데이터)) + geom 함수 .....

첫 번째 인자는 처리할 데이터 프레임 이름.

두 번째 aes 부분은 aesthetic mapping(미적 매핑)이라는 의미.
ggplot 함수로 데이터를 표현할 때 좀 더 아름답게 표현하겠다라는 그런 의미가 있으며 aes 부분이 처리.
이 부분에 올 수 있는 항목은 x 축 데이터 , y 축 데이터 , 점의 모양 , 점의 크기 , 점의 색깔과 같은 값들이 올 수 있음.
이렇게 미적 매핑하는 것을 다른 말로 스케일링(scaling) 작업이라고도 함.

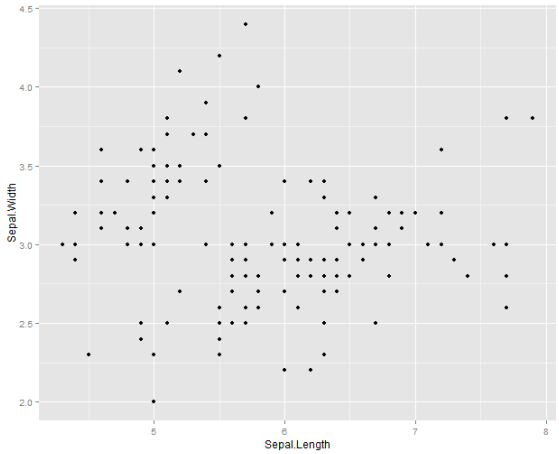
세 번째 aes 뒷부분에 + 로 geom 함수가 나올 수 있는데 이 부분은 geometric object의 약자로 앞에서 만들어진 데이터를 실제 렌더링으로 표현하는 부분을 의미.
다양한 geom 관련 함수가 있고 설정값들이 있음.

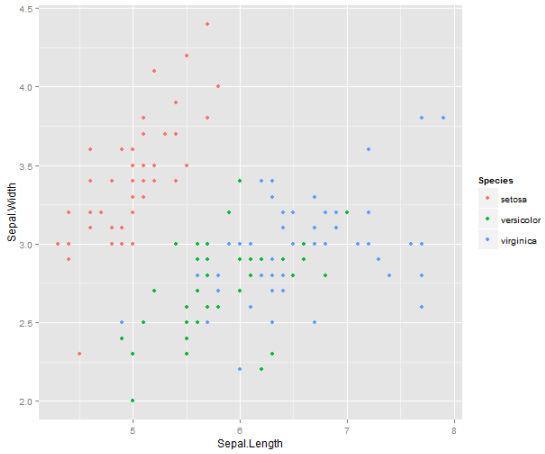
Aesthetics

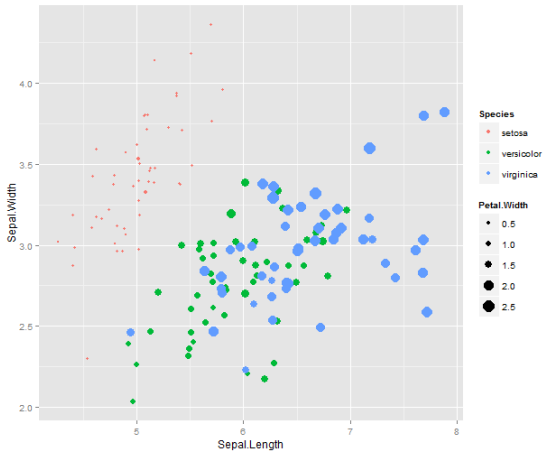
- x, y
  : x축, y축
- size
  : geom의 크기
- shape
  : geom의 모양
- linetype
  : geom의 라인 종류
- colour
  : 표면 색상
- fill
  : 채움 색상
- alpha
  : geom의 투명도(0=투명, 1=불투명)

Scatter Plots

- x <- ggplot(iris, aes(Sepal.Length, Sepal.Width))
- x + geom_point() # left
- x + geom_point(aes(colour=Species)) # middle
- x + geom_point(aes(colour=Species, size=Petal.Width)) # right

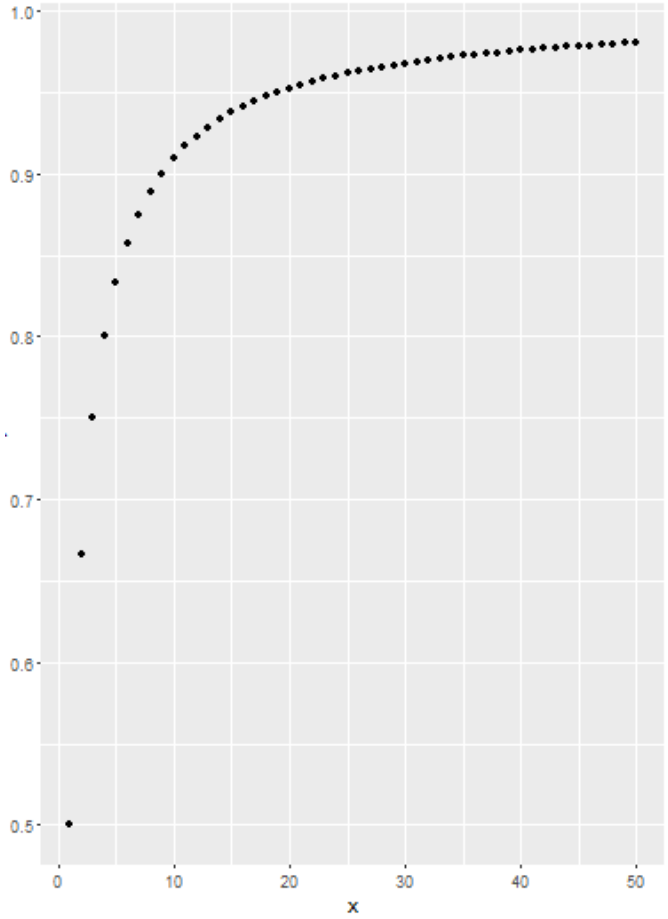






Exercises #1

〔문제〕 y = x/(x+1)를 아래와 같이 Scatter Plot으로 도식하시오. 단, x는 1에서 50까지의 정수임.



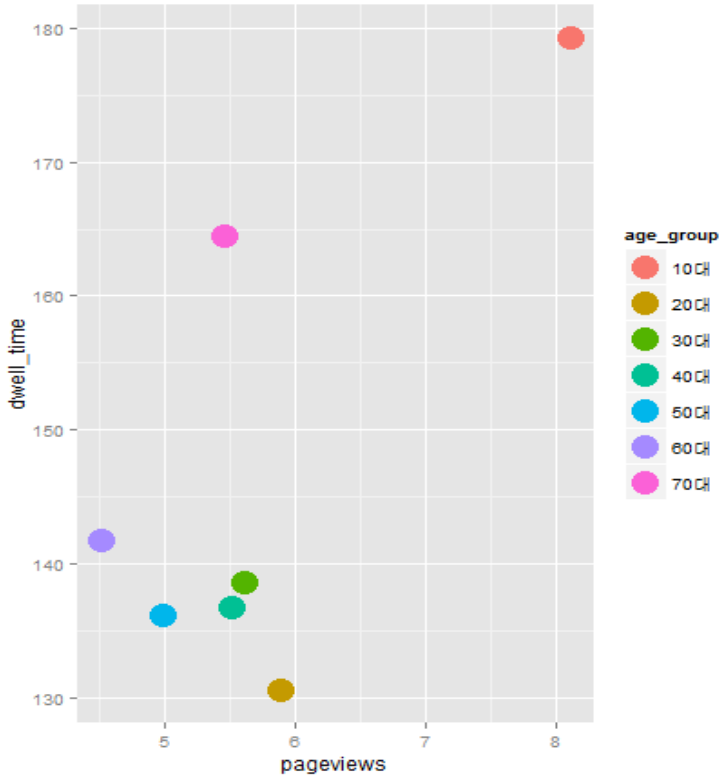
< Hint >

```
x <- 1:50
y <- y <- sapply(x, function(x) x/(x+1))
df <- data.frame(x=x, y=y)
```

Exercises #2

userDemoInfo.csv와 userLogs.csv 데이터를 이용하여 아래 문제를 해결하시오.

〔문제〕 사용자의 연령대별로 페이지뷰(pageviews)와 체류시간(dwell_time)의 평균을 구한 후 우측의 그림과 같은 Scatter Plot을 출력하시오. 단, 연령대별로 점의 색을 다르게 하고 점의 크기를 7로 설정.

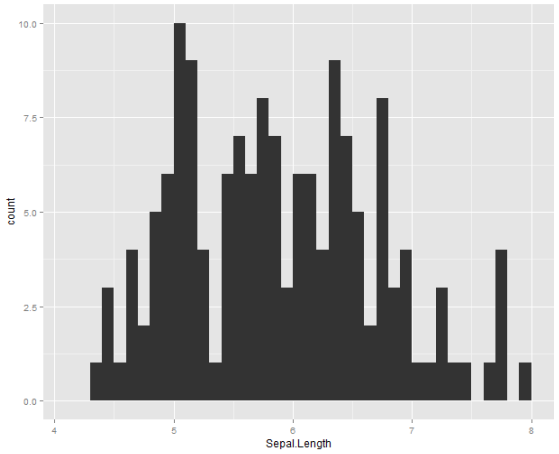


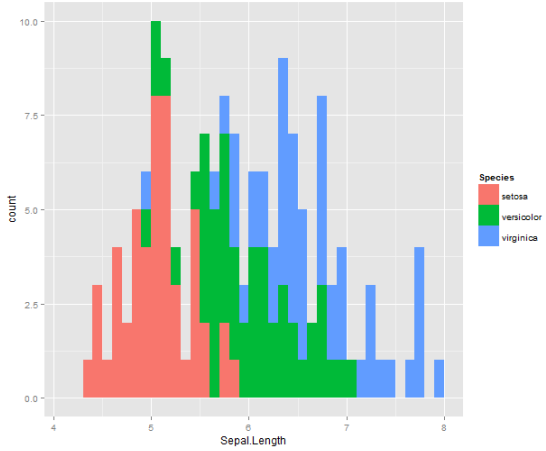
< Hint >

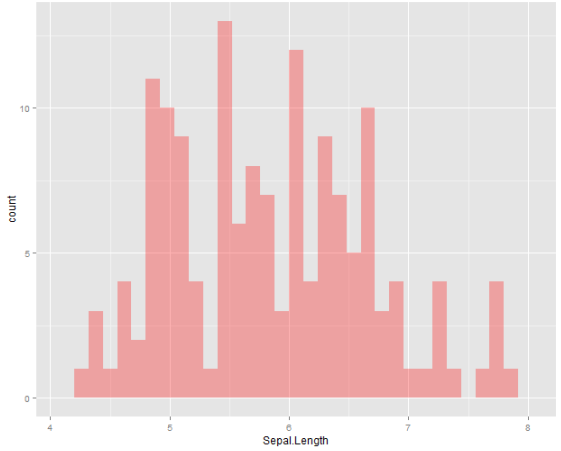
```
demo <- read.csv("userDemoInfo.csv", stringsAsFactors=FALSE)
logs <- read.csv("userLogs.csv", stringsAsFactors=FALSE)
md <- merge(demo,logs, by.x="cus_id", by.y="cus_id")
md$ageGr <- paste(md$age %/% 10, "0대", sep="")
ad <- aggregate(md[8:9], by=list(age_group=md$ageGr), mean)
```

Histograms

- x <- ggplot(iris, aes(Sepal.Length))
- x + geom_histogram(binwidth=0.1) # left
- x + geom_histogram(binwidth=0.1, aes(fill=Species)) # middle
- x + geom_histogram(fill="red", alpha=0.3) # right



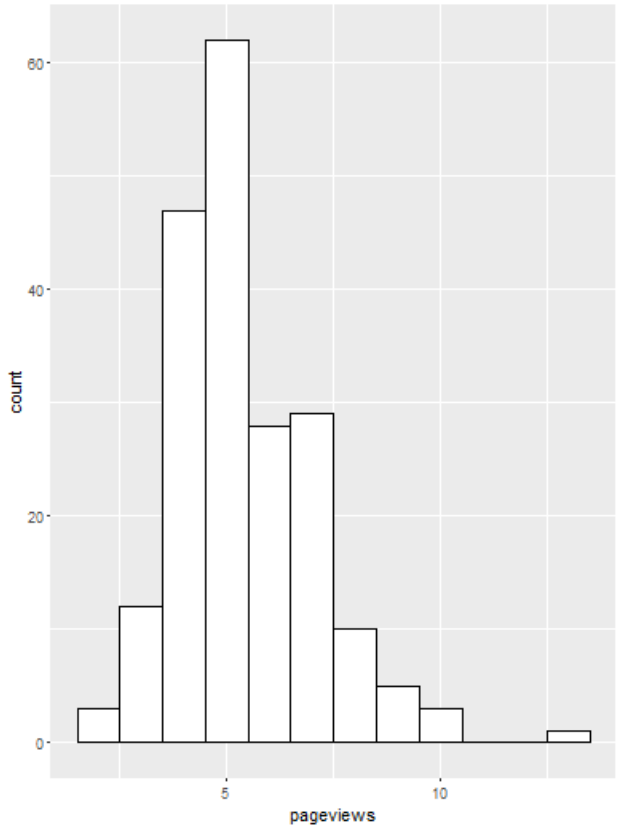




Exercises #3

userDemoInfo.csv와 userLogs.csv 데이터를 이용하여 아래 문제를 해결하시오.

〔문제〕 각 사용자의 평균 pageview를 구한 후 아래와 같은 Histogram을 출력하시오.
단, binwidth를 1로 설정.

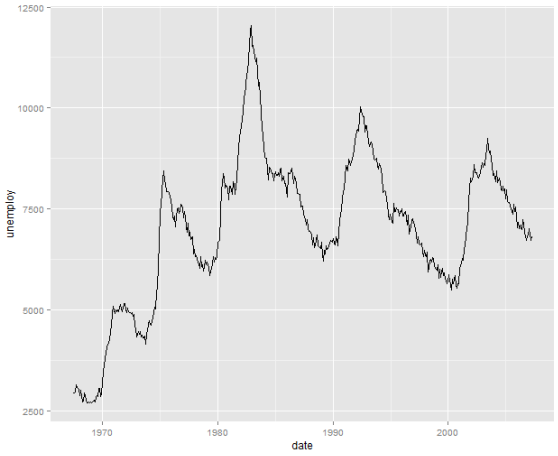


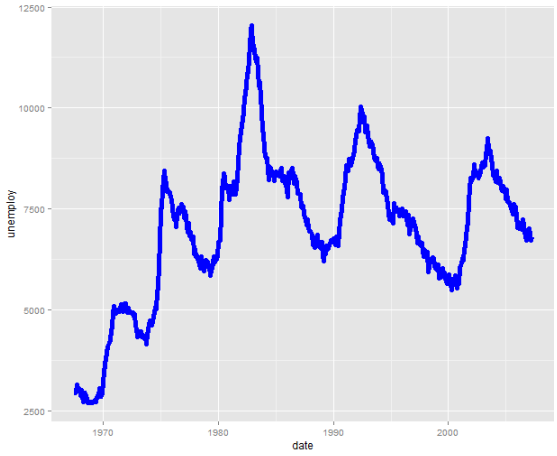
< Hint >

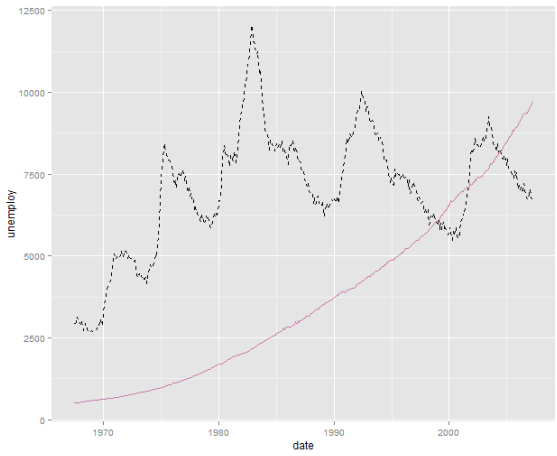
```
demo <- read.csv("userDemoInfo.csv", stringsAsFactors=F)
logs <- read.csv("userLogs.csv", stringsAsFactors=F)
md <- merge(demo,logs, by.x="cus_id", by.y="cus_id")
ag <- aggregate(md[8], by=list(cus_id=md$cus_id), mean)
```

Line Charts

- x <- ggplot(economics)
- x + geom_line(aes(x=date, y=unemploy)) # left
- x + geom_line(aes(x=date, y=unemploy), colour="blue", size=2) # middle
- x + geom_line(aes(x=date, y=unemploy), linetype=2) +
  geom_line(aes(x=date, y=pce), colour="#CC79A7") # right

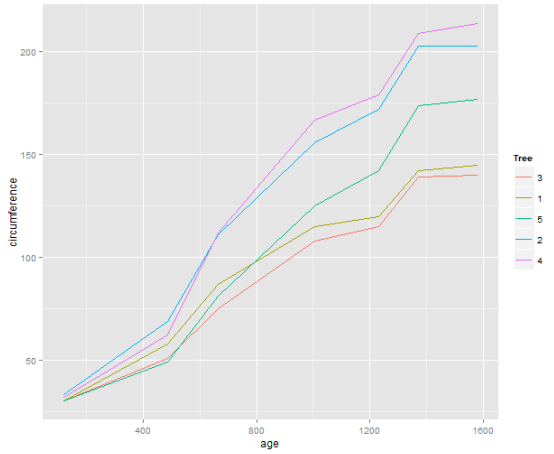


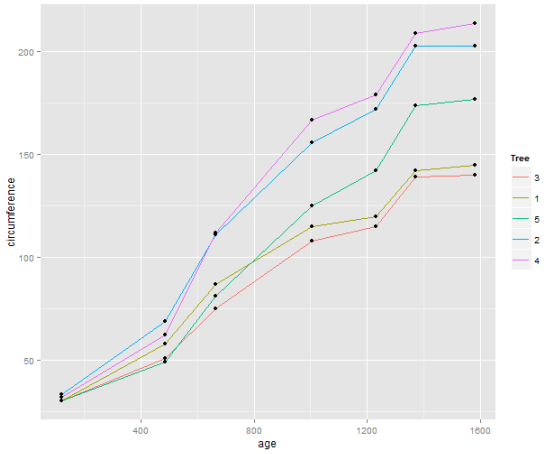


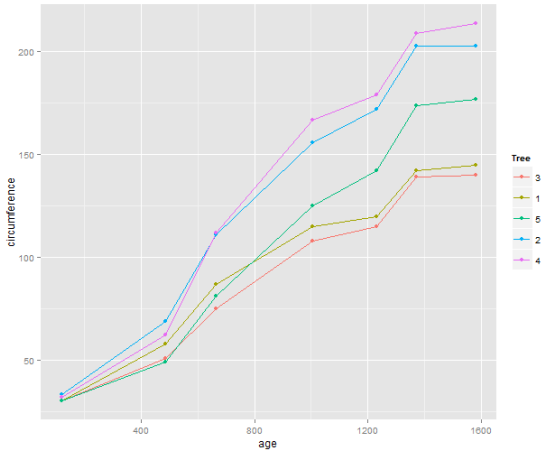


Inheritances

- x <- ggplot(Orange, aes(age, circumference))
- x + geom_line(aes(colour=Tree)) # left
- x + geom_line(aes(colour=Tree)) + geom_point() # middle
- ggplot(Orange, aes(age, circumference, colour=Tree)) + geom_line() + geom_point() # right

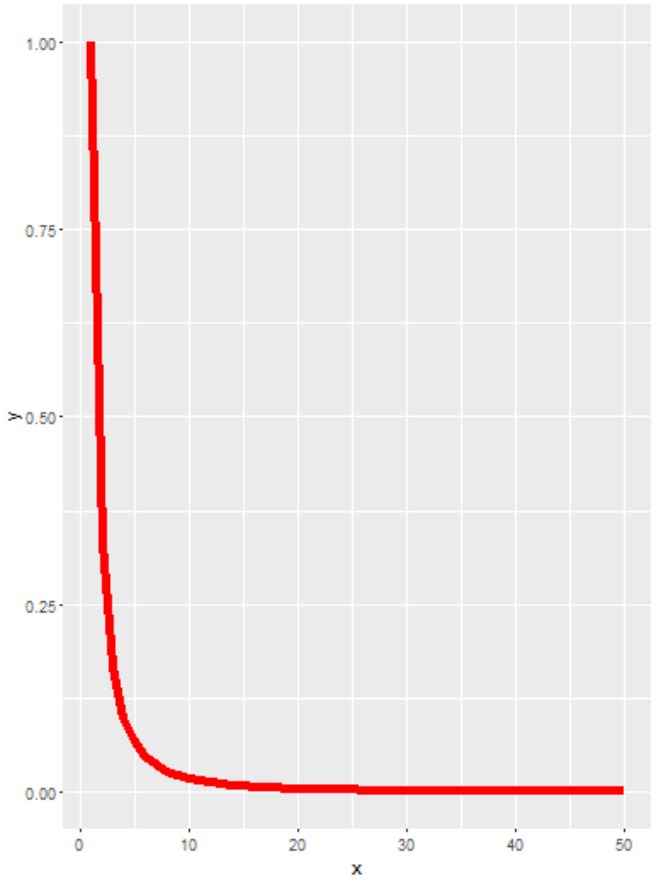






Exercises #4

〔문제〕 x값은 1 ~ 50, y값은 1, 1/(1+2), 1/(1+2+3), … , 1/(1+2+3+…50)인 Line Chart를 도식하시오. 단, color는 붉은색, size는 2임.



< Hint >

```
d <- data.frame(x=1:50, y=sapply(1:50, function(x) 1/sum(1:x)))
```

Exercises #5

userDemoInfo.csv와 userLogs.csv 데이터를 이용하여 아래 문제를 해결하시오.

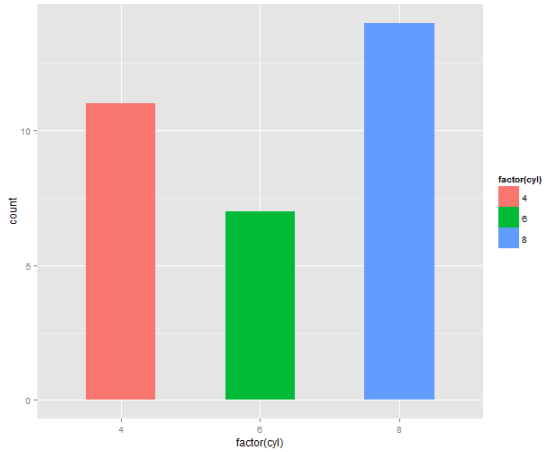
〔문제〕 시간에 따라 남녀의 평균 pageview가 어떻게 변화하는지 도식하시오. 단, 시간은 일단위임.

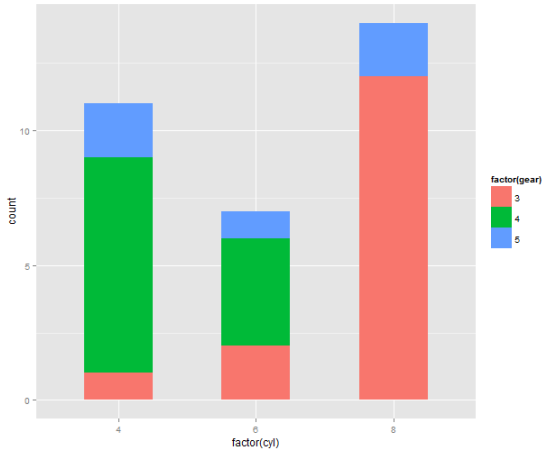
< Hint >

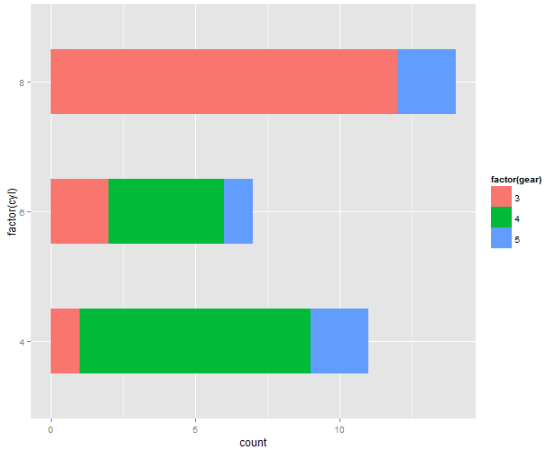
```
demo <- read.csv("userDemoInfo.csv", stringsAsFactors=F)
logs <- read.csv("userLogs.csv", stringsAsFactors=F)
md <- merge(demo, logs, by.x="cus_id")
md$date <- as.Date(as.character(md$time_id), "%Y%m%d")
ag <- aggregate(md[8], by=list(date=md$date, gender=md$gender), mean)
```

Bar Charts

- x <- ggplot(mtcars, aes(factor(cyl)))
- x + geom_bar(aes(fill=factor(cyl)), width=.5) # left
- x + geom_bar(aes(fill=factor(gear)), width=.5) # middle
- x + geom_bar(aes(fill=factor(gear)), width=.5) + coord_flip() # right



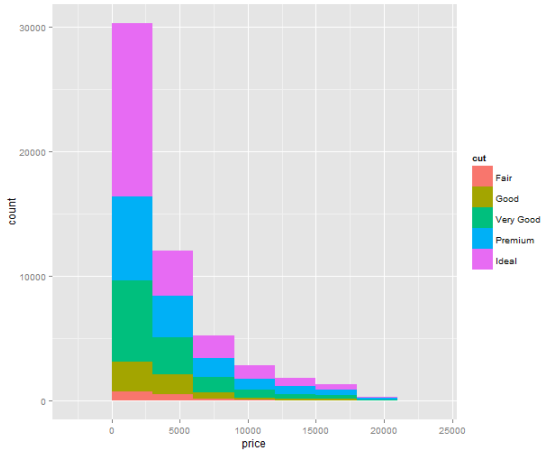


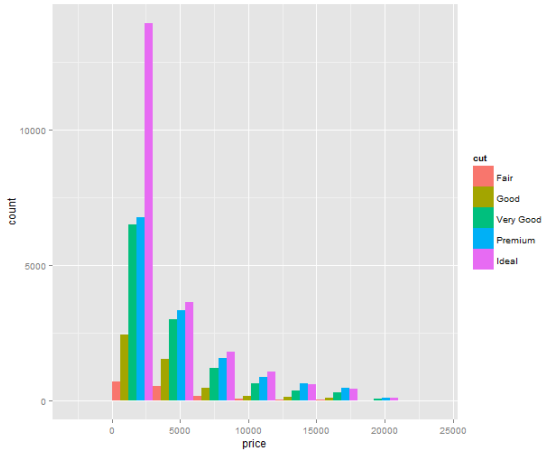


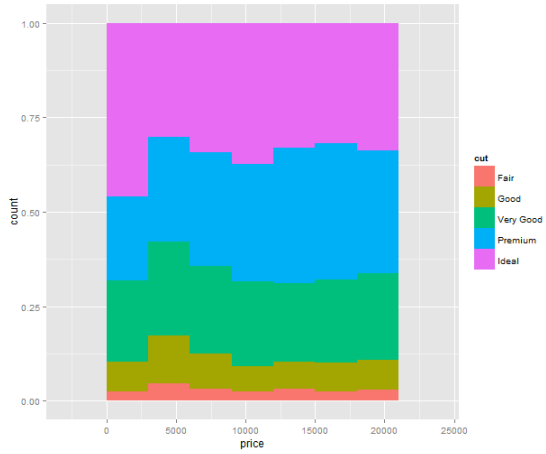
- x <- ggplot(mtcars, aes(factor(cyl), mpg))
- x + geom_bar(aes(fill=factor(cyl)), width=.5, stat="identity") # results?

Position Adjustments

- x <- ggplot(diamonds, aes(x=price))
- x + geom_bar(aes(fill=cut), binwidth=3000) # left
- x + geom_bar(aes(fill=cut), binwidth=3000, position="dodge") # middle
- x + geom_bar(aes(fill=cut), binwidth=3000, position="fill") # right



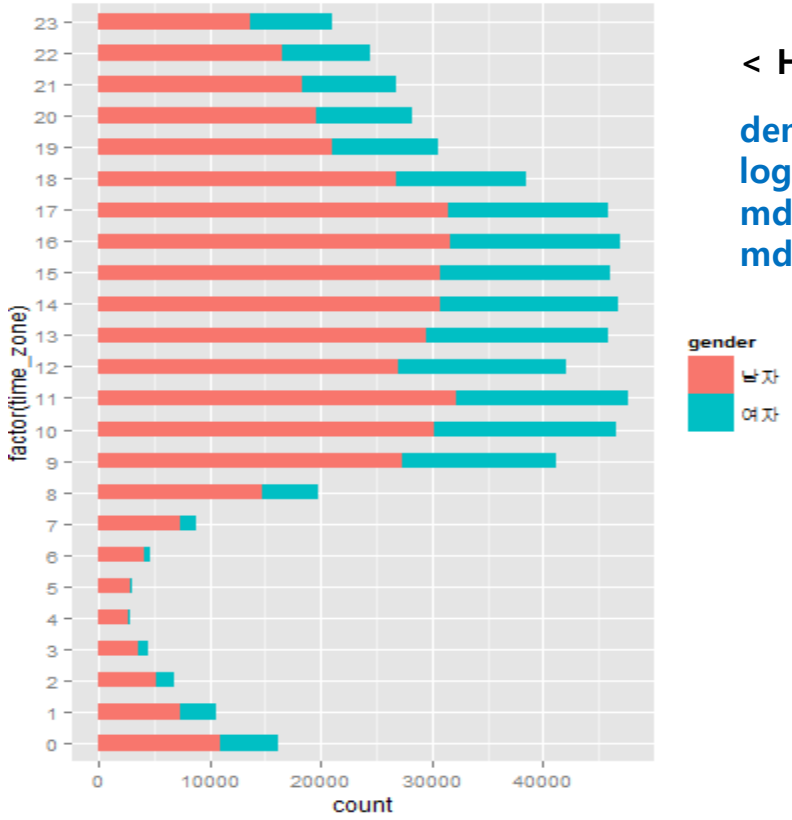




Exercises #6

userDemoInfo.csv와 userLogs.csv 데이터를 이용하여 아래 문제를 해결하시오.

〔문제〕 남녀별 각 시간대 접속 빈도를 계산하여 아래 그림과 같은 Bar Chart를 출력하시오. 단, bar의 width를 0.7로 설정.

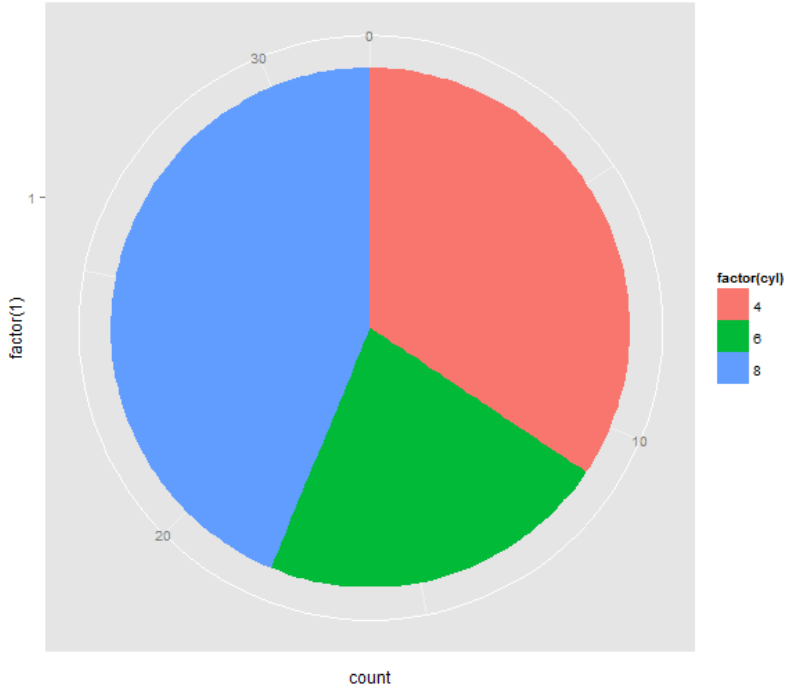


< Hint >

```
demo <- read.csv("userDemoInfo.csv", stringsAsFactors=F)
logs <- read.csv("userLogs.csv", stringsAsFactors=F)
md <- merge(demo, logs, by.x="cus_id")
md$time_zone <- md$time_id %% 100
```

Circle Charts

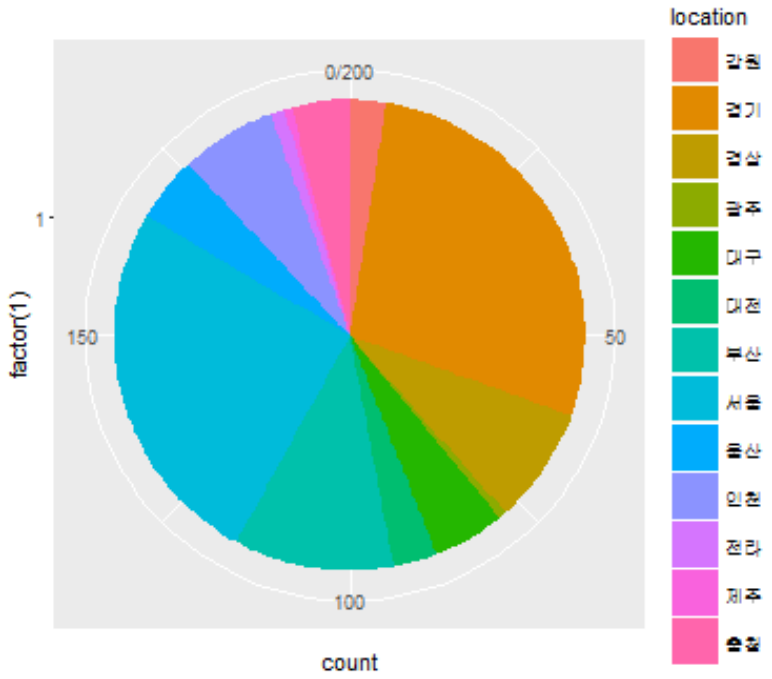
- x <- ggplot(mtcars, aes(x=factor(1), fill=factor(cyl)))
- x + geom_bar(width=1) + coord_polar(theta="y")



Exercises #7

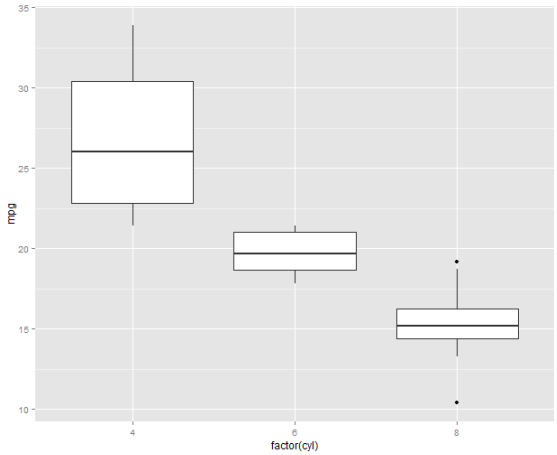
userDemoInfo.csv와 userLogs.csv 데이터를 이용하여 아래 문제를 해결하시오.

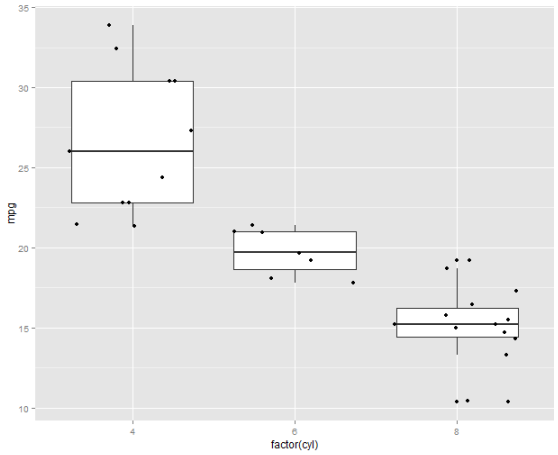
〔문제〕 200명 사용자의 지역별 분포를 아래 그림과 같은 Circle Chart로 시각화하시오.

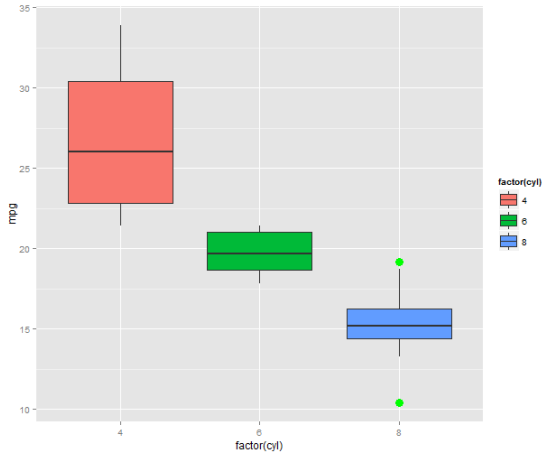


Box Plots

- x <- ggplot(mtcars, aes(factor(cyl), mpg))
- x + geom_boxplot() # left
- x + geom_boxplot() + geom_jitter() # middle
- x + geom_boxplot(aes(fill = factor(cyl)), outlier.colour = "green", outlier.size = 4) # right



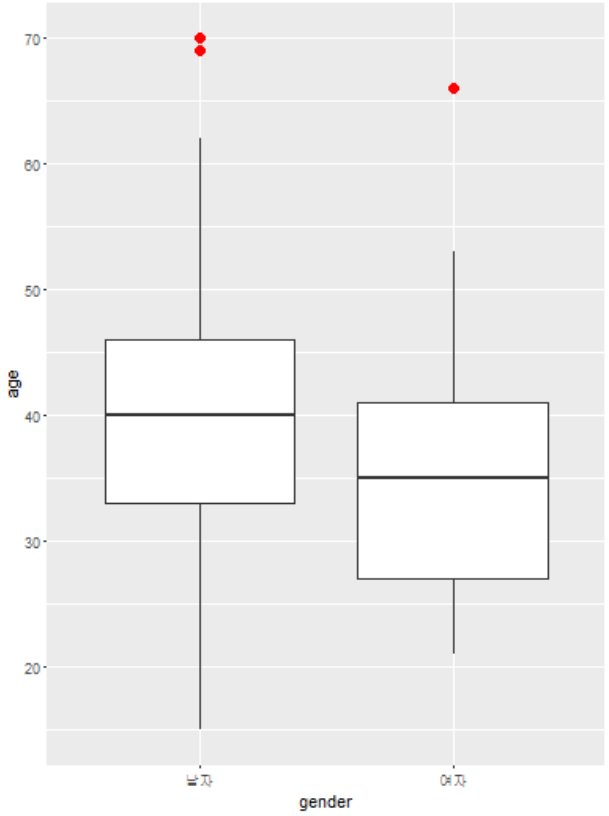




Exercises #8

userDemoInfo.csv와 userLogs.csv 데이터를 이용하여 아래 문제를 해결하시오.

〔문제〕 남녀별 나이에 대한 Box Plot을 출력하시오. 단, 크기 3의 붉은색으로 outlier를 표현할 것.

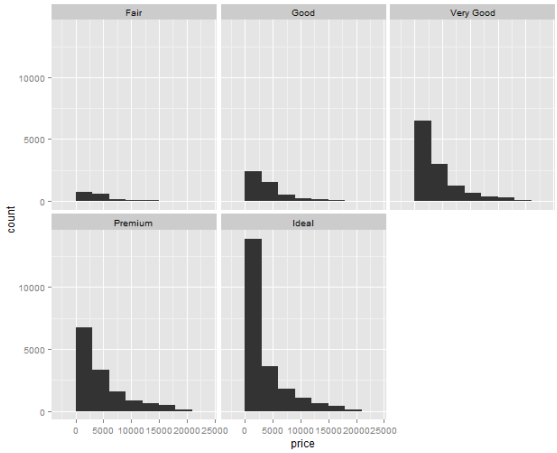


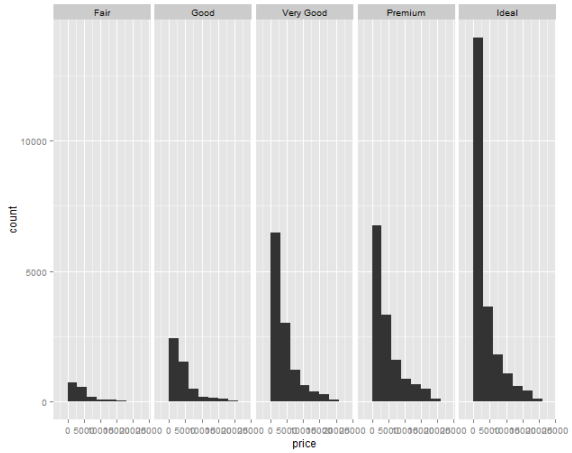
< Hint >

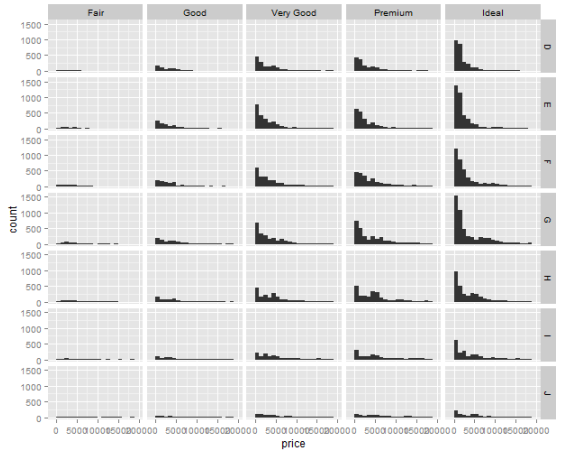
```
demo <- read.csv("userDemoInfo.csv", stringsAsFactors=F)
```

Facets

- x <- ggplot(diamonds, aes(x=price))
- x + geom_bar(binwidth=3000) + facet_wrap(~ cut) # left
- x + geom_bar(binwidth=3000) + facet_grid(. ~ cut) # middle
- x + geom_bar(binwidth=1000) + facet_grid(color ~ cut) # right



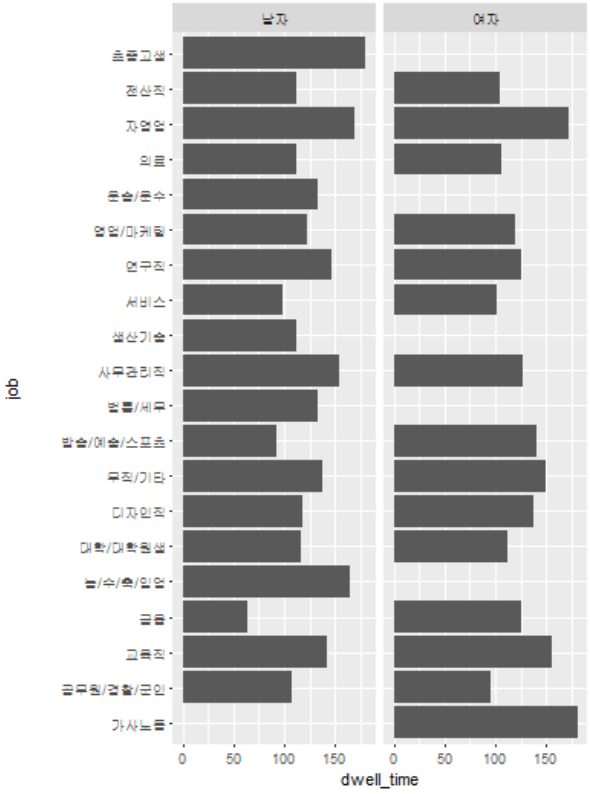




Exercises #9

userDemoInfo.csv와 userLogs.csv 데이터를 이용하여 아래 문제를 해결하시오.

〔문제〕 직업별 성별 평균 체류시간(dwell_time)를 계산하여, 우측의 그림과 같은 Facet 이 적용된 Chart를 출력하시오.

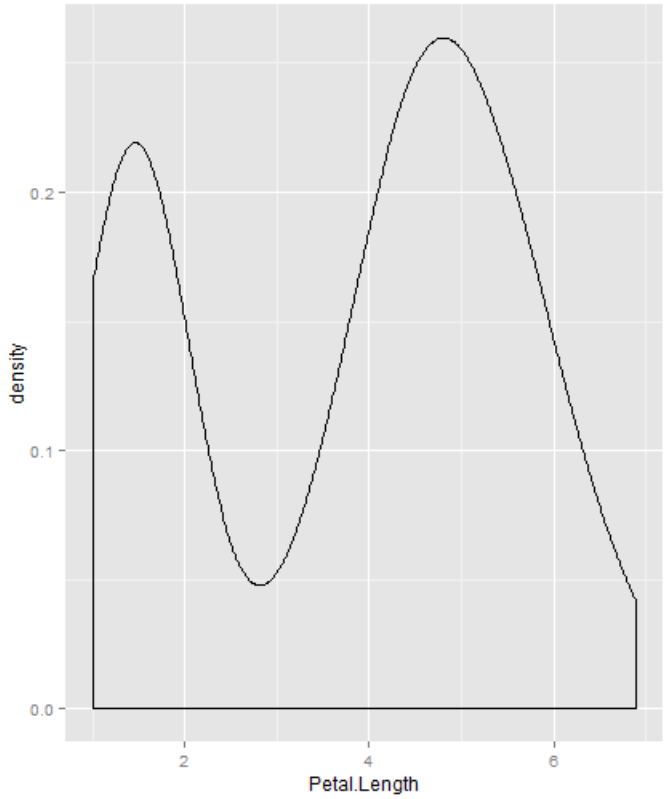


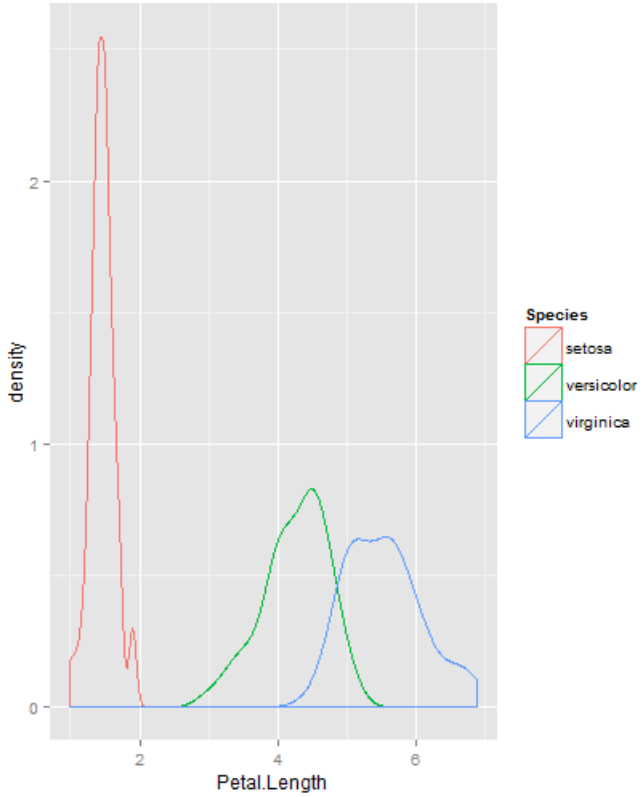
< Hint >

```
demo <- read.csv("userDemoInfo.csv", stringsAsFactors=F)
logs <- read.csv("userLogs.csv", stringsAsFactors=F)
md <- merge(demo, logs, by.x="cus_id")
ag <- aggregate(md[9], by=list(job=md$job, gender=md$gender), mean)
```

Density

- x <- ggplot(iris, aes(Petal.Length))
- x + geom_density() # left
- x + geom_density(aes(color=Species)) # right





Exercises #10

userDemoInfo.csv와 userLogs.csv 데이터를 이용하여 아래 문제를 해결하시오.

〔문제〕 200명 사용자의 남녀별 나이 분포를 아래 그림과 같이 시각화하시오.

< Hint >

```
demo <- read.csv("userDemoInfo.csv", stringsAsFactors=F)
```