시각화 핵심 요약

Library(ggplot2)

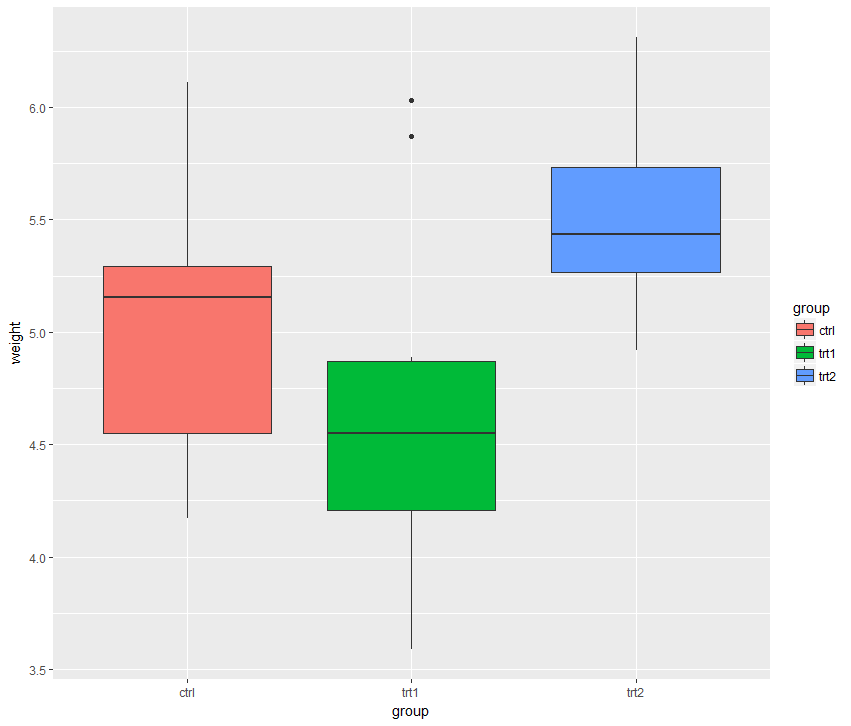
**데이터 분포 요약 : 1. 히스토그램,2. 밀도곡선,3.박스플롯,4. 이차원 데이터의 밀도 그래프**

|  |  |
| --- | --- |
| 옵션 | 설명 |
| **X,y** | X축, y축에 해당하는 변수를 지정함 |
| **data** | Dataframe을 지정하여 축과 색, 크기 등을 지정 |
| **color** | 변수의 레벨을 이용하여 색상을 지정 |
| **shape** | 변수 값을 이용하여 심볼 모양을 사각형, 원 등 다양하게 차별화 |
| **size** | 심볼의 크기를 변수를 이용하여 지정 |
| **alpha** | 투명도를 표시하는 값으로 0~1, 1이 가장 진한상태 |
| **geom** | 그래프 유형을 선정, 라인, 바 등 |
| **facet** | 조건에 따른 격자형으로 그래프 작성 |
| **Xlim, ylim** | 각각 2개의 값으로 최소, 최대값을 각 축에 설정 |
| **xlab,ylab** | 각 축에 대한 라벨 |
| **main, sub** | 메인과 서브타이틀로 그래프에 대한 설명 추가 |

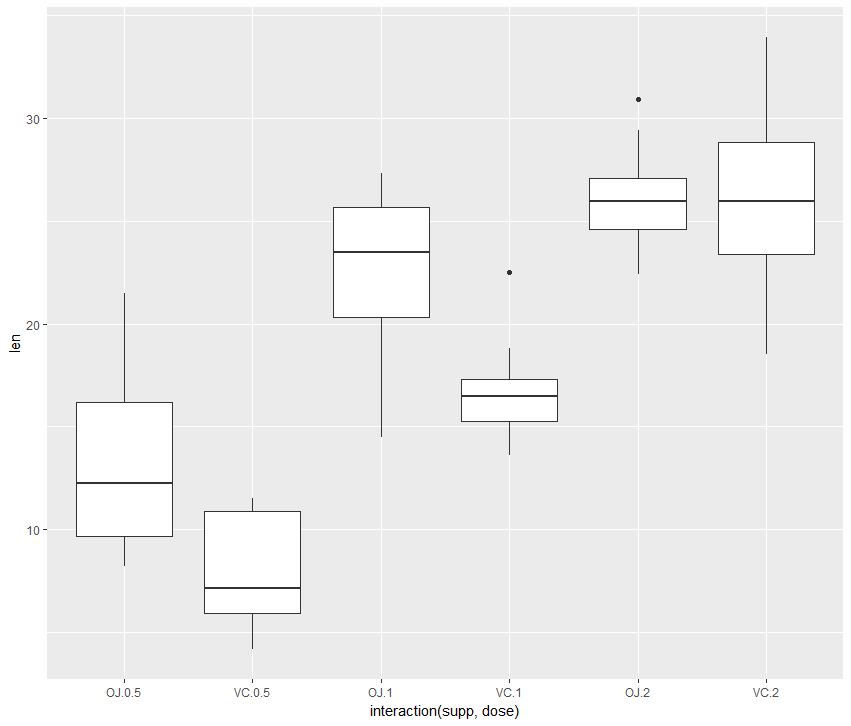
#박스플롯

data("PlantGrowth")

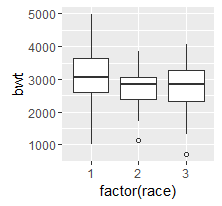
ggplot(PlantGrowth, aes(x=group, y=weight, fill=group))+geom\_boxplot()



ggplot(ToothGrowth, aes(x=interaction(supp,dose), y=len))+geom\_boxplot()



ggplot(birthwt, aes(x=factor(race), y=bwt))+geom\_boxplot(outlier.size=1.5, **outlier.shape=21)**

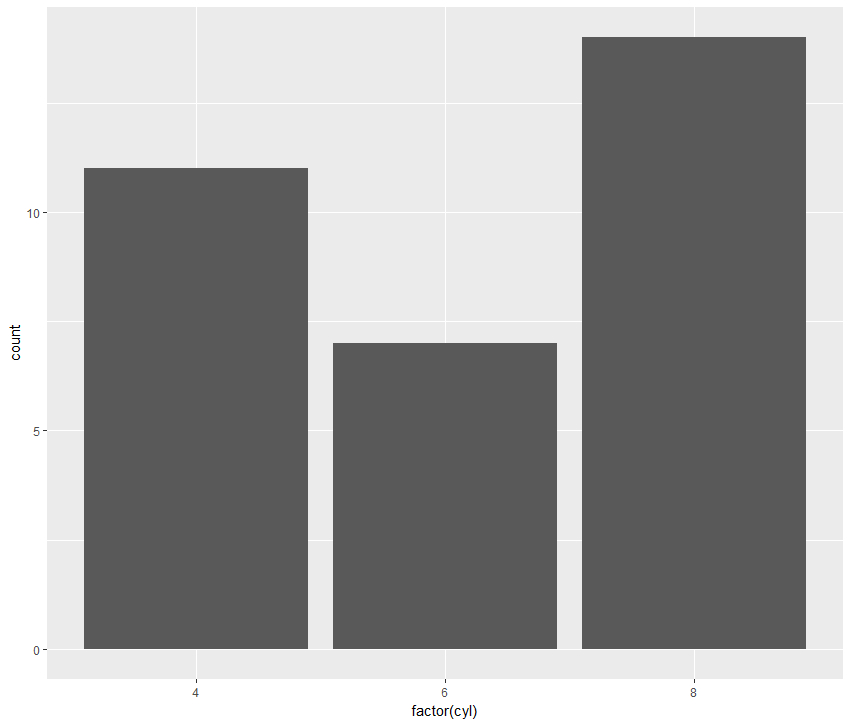


#아웃라이어 확인, Outlier.size,shape, 기본크기 2, 기본모양 16, 아웃라이어 겹치는 것

#막대그래프

data(mtcars)

ggplot(mtcars, aes(x=factor(cyl)))+geom\_bar()



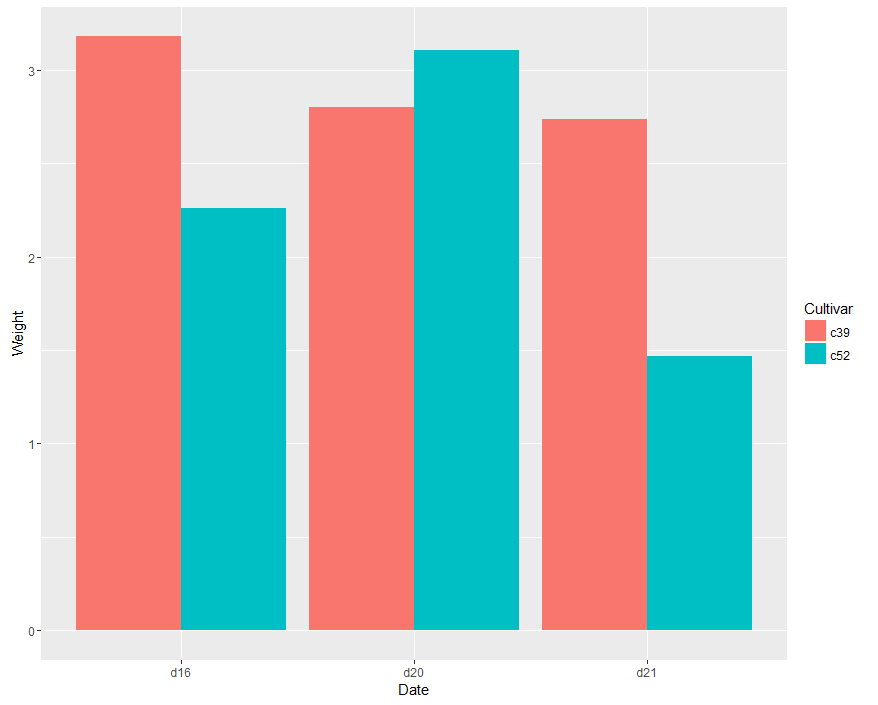
> **ggplot(mtcars, aes(x=mpg))+geom\_histogram(binwidth=10)**

# Binwidth=디폴트 30개, 디폴트는 검정색으로



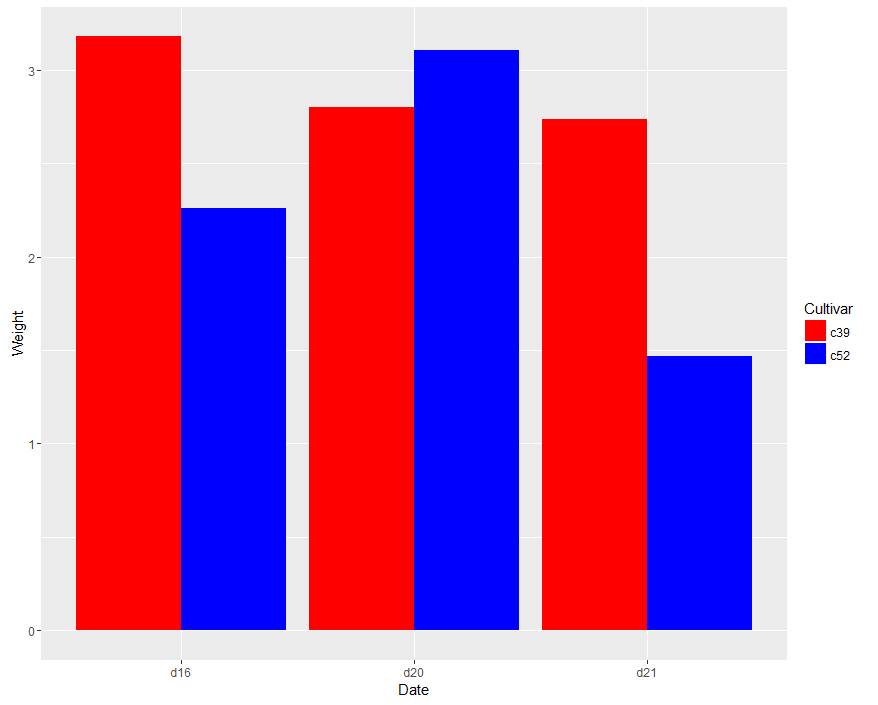
ggplot(cabbage\_exp, aes(x=Date, y=Weight, fill=Cultivar))+geom\_bar(stat="identity", **position="dodge")**

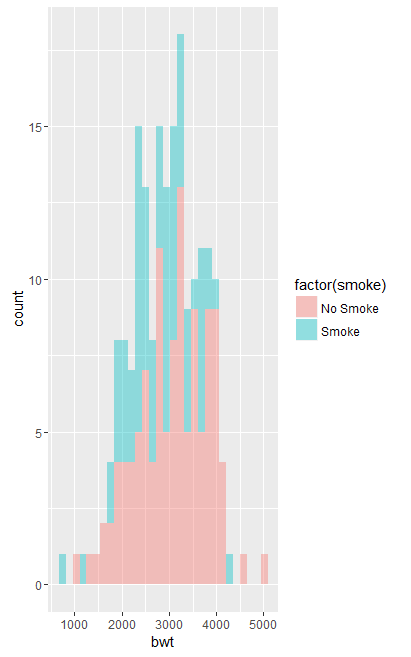
#Geom\_bar(stat=“identity”)해줘야함



> ggplot(cabbage\_exp, aes(x=Date, y=Weight, fill=Cultivar))+ geom\_bar(stat="identity", position="dodge")+**scale\_fill\_manual(values=c("red","blue"))**

#색깔을 각각 넣고 싶을때 scale\_fill\_manual





ggplot(birthwt,aes(x=bwt,**fill=factor(smoke)**))+**geom\_histogram(alpha=.4)**

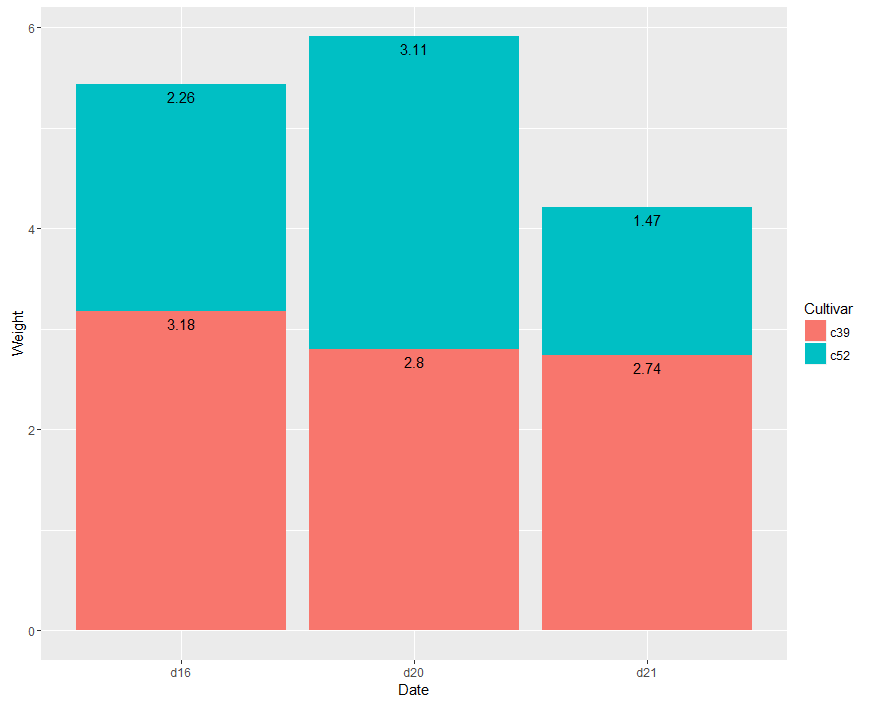
Fill=그룹정보, 집단들의 개별정인분포를 알아보기힘듬, Identity: 겹쳐저서 그려짐

ce=arrange(cabbage\_exp, Date, Cultivar)

ce=ddply(ce,"Date", transform, label\_y=**cumsum**(Weight))

PLYR패키지의 ARRANGE함수를 이용해서 정렬미리 해놔야함. CUMSUM위해서…

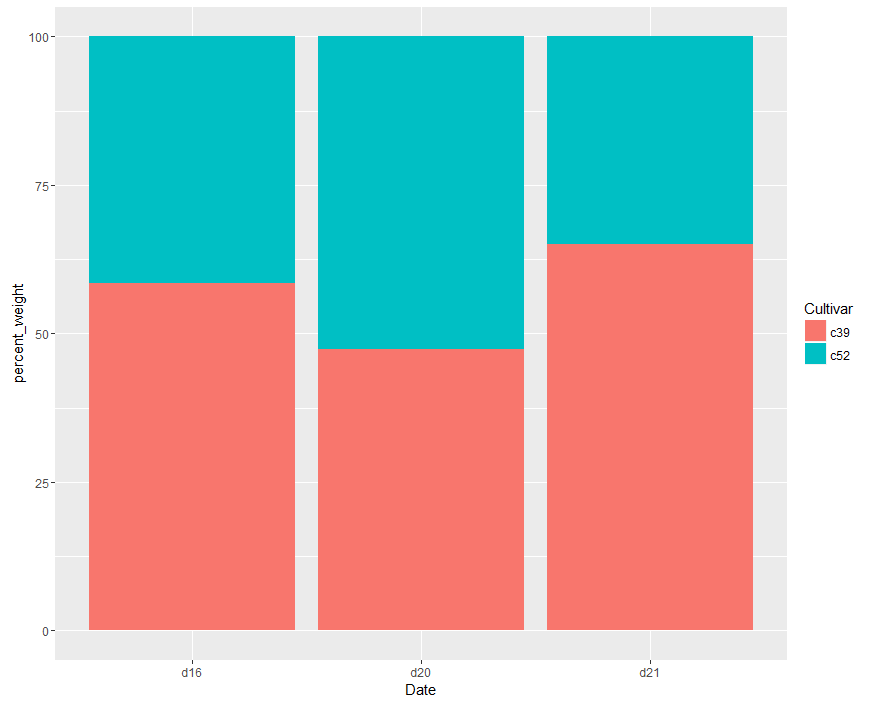
ggplot(ce,aes(x=Date,y=Weight,fill=Cultivar))+geom\_bar(stat="identity")+geom\_text(aes(label=Weight, y=label\_y), **vjust=1.5)**



**비율 누적 막대그래프**

ce=ddply(cabbage\_exp,"Date", transform, percent\_weight=**Weight/sum(Weight)\*100)**

ggplot(ce,aes(x=Date,y=percent\_weight,fill=Cultivar))+geom\_bar(stat="identity")

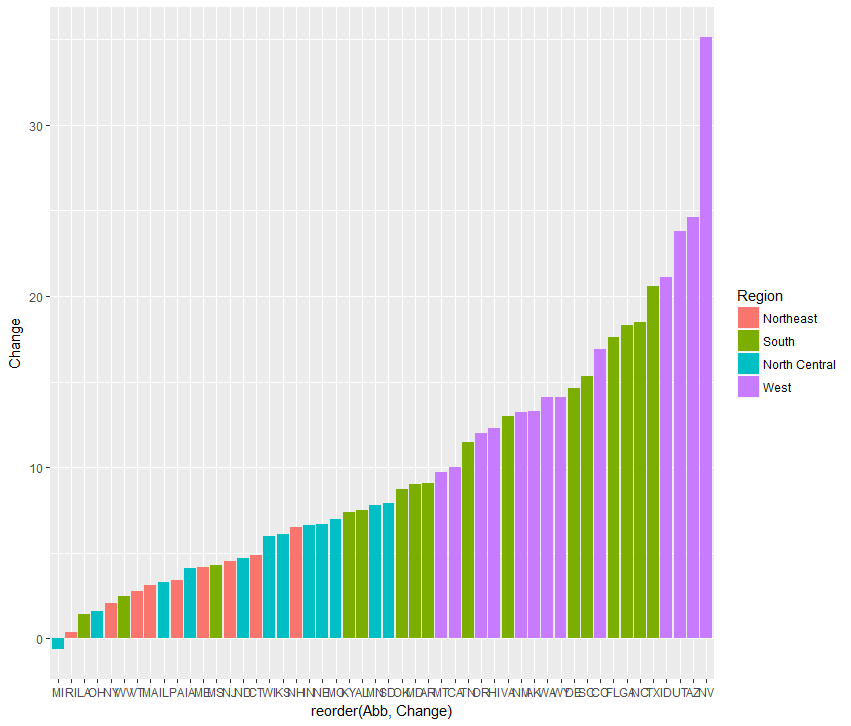


TRANSFORM 새로운 열도 추가해줌, 데이터프레임형태를 그래도 유지한채로

library(gcookbook)

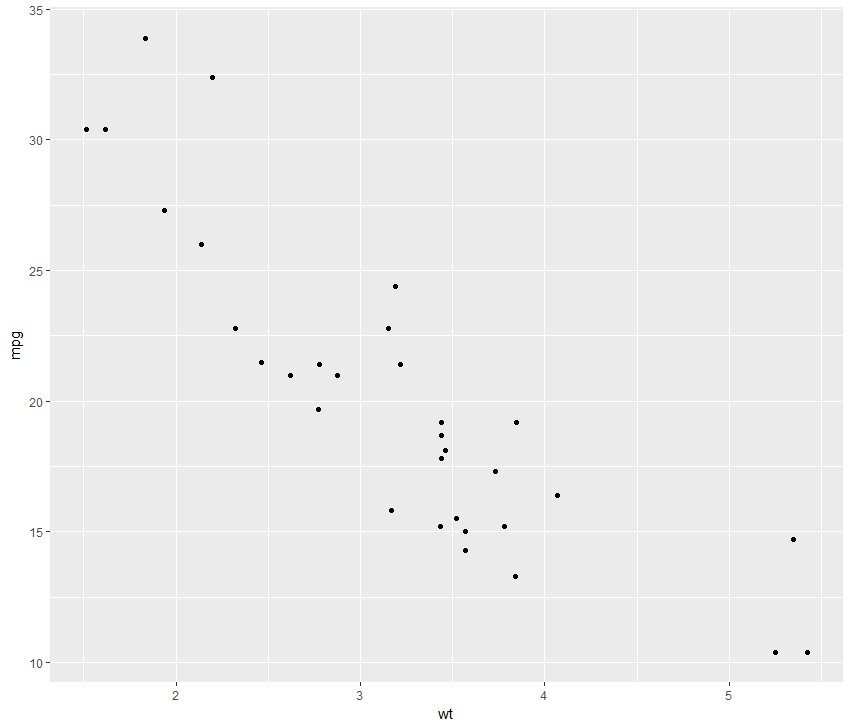
data(uspopchange)

ggplot(uspopchange, aes(x=reorder(Abb, Change), y=Change, fill=Region))+geom\_bar(stat="identity")



#산점도

boxplot(mpg~factor(cyl), data=mtcars)

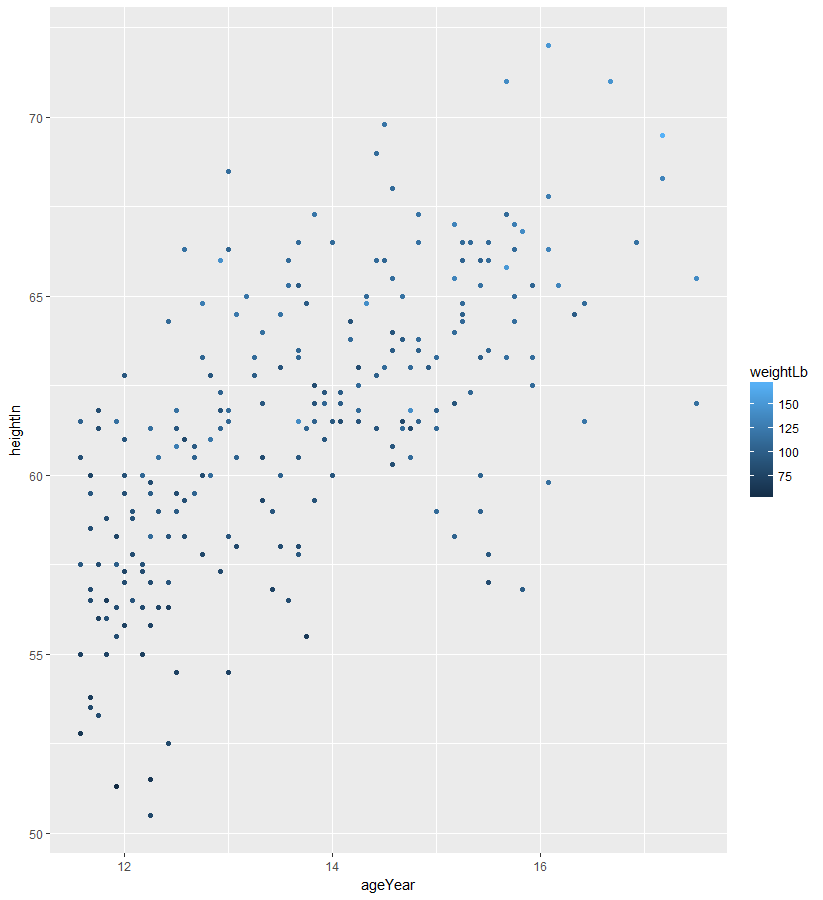


> library(ggplot2)

> library(gcookbook)

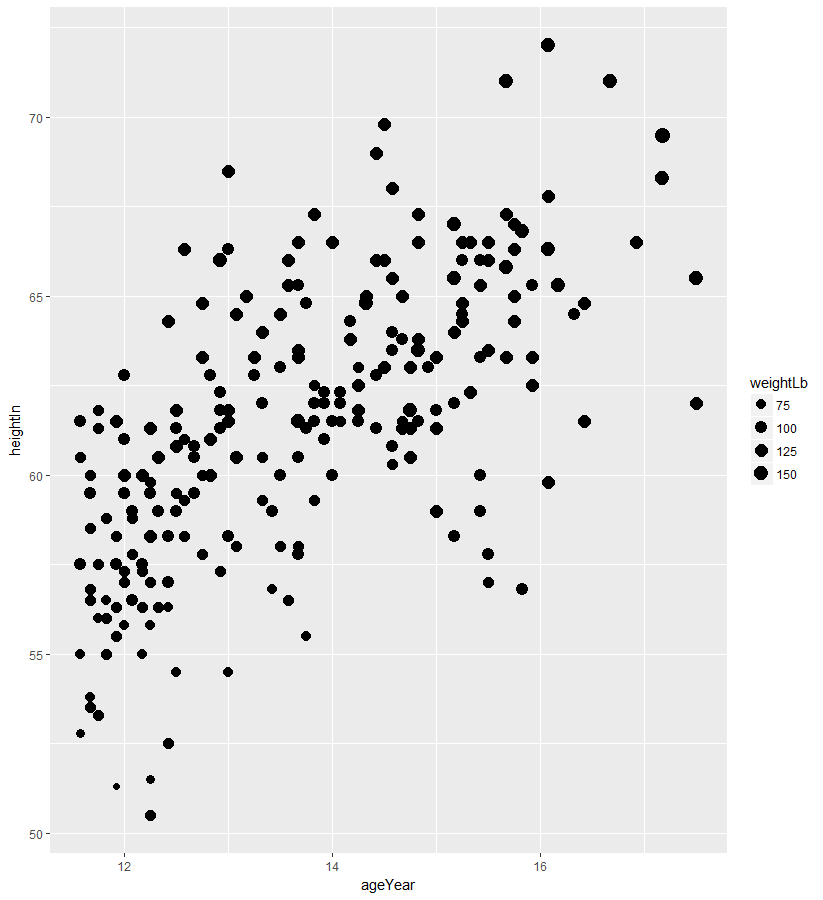
> ggplot(heightweight,aes(x=ageYear,y=heightIn,colour=weightLb))+geom\_point()

몸무게가 낮을수록 어두운색깔로 표현됨, gradient



ggplot(heightweight,aes(x=ageYear,y=heightIn,size=weightLb))+geom\_point()+scale\_size\_continuous(range=c(2,5))

연속형 사이즈 지정 size\_continuoius(range=c(2,5)), Scale\_size\_area() 점의 크기가 데이터에 대입되어 왜곡되지 않음

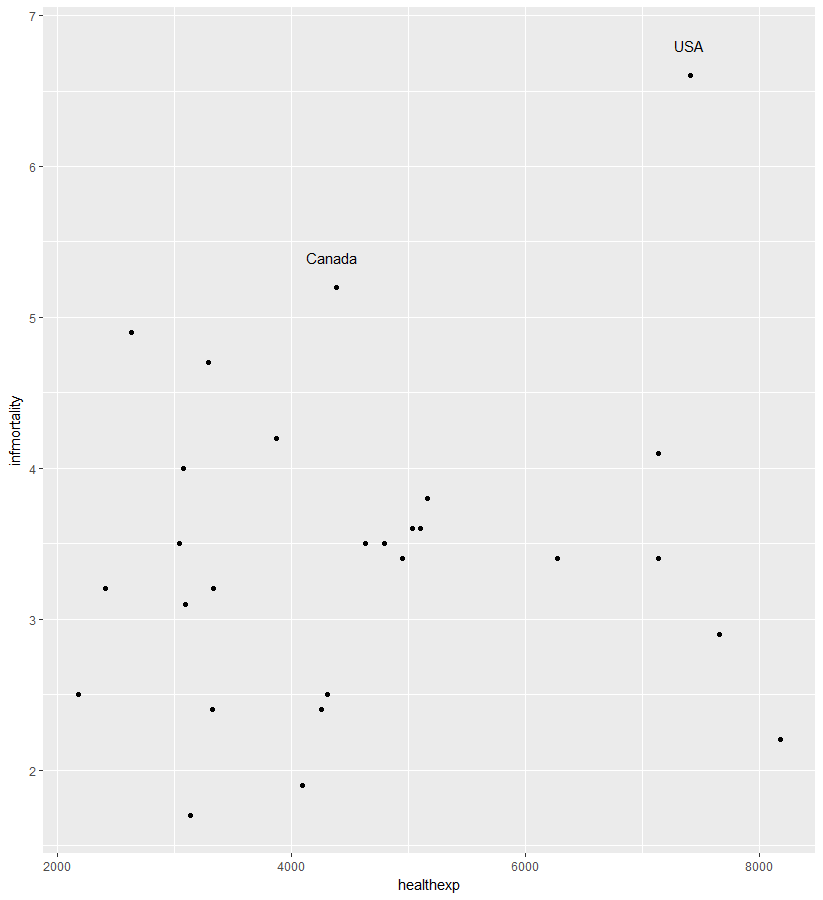


> df<-subset(countries, Year==2009 & healthexp>2000)

> View(df)

> sp=ggplot(df, aes(x=healthexp, y=infmortality))+geom\_point()

> sp+annotate("text",x=4350, y=5.4, label="Canada")+annotate("text", x=7400, y=6.8, label="USA")



# label에 열을 따로 생성하여 내가 찾고싶은 나라의 이름이 포함되어 있는지 검색하고 TRUE, FALSE로 반환시켜준다.

> cdat<-subset(countries, Year==2009 & healthexp>2000)

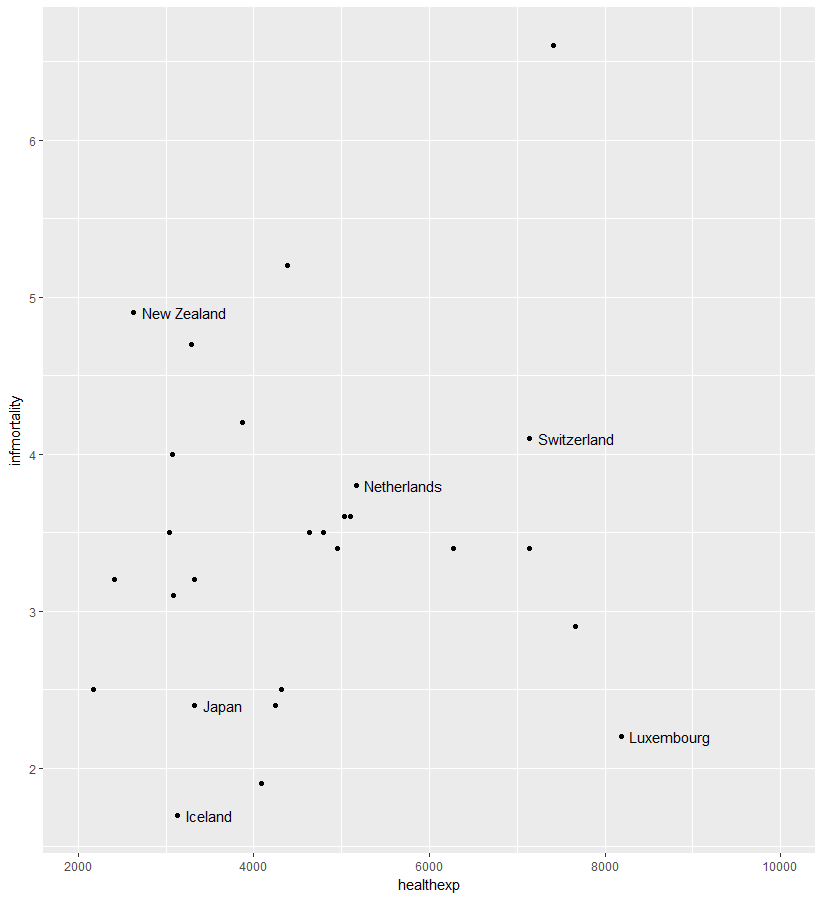
> cdat$Name1<-cdat$Name

> idx<-cdat$Name1 %in% c("New Zealand", "Iceland", "Japan", "Luxembourg", "Netherlands", "Switzerland")

> cdat$Name1[!idx]<-NA

> View(cdat)

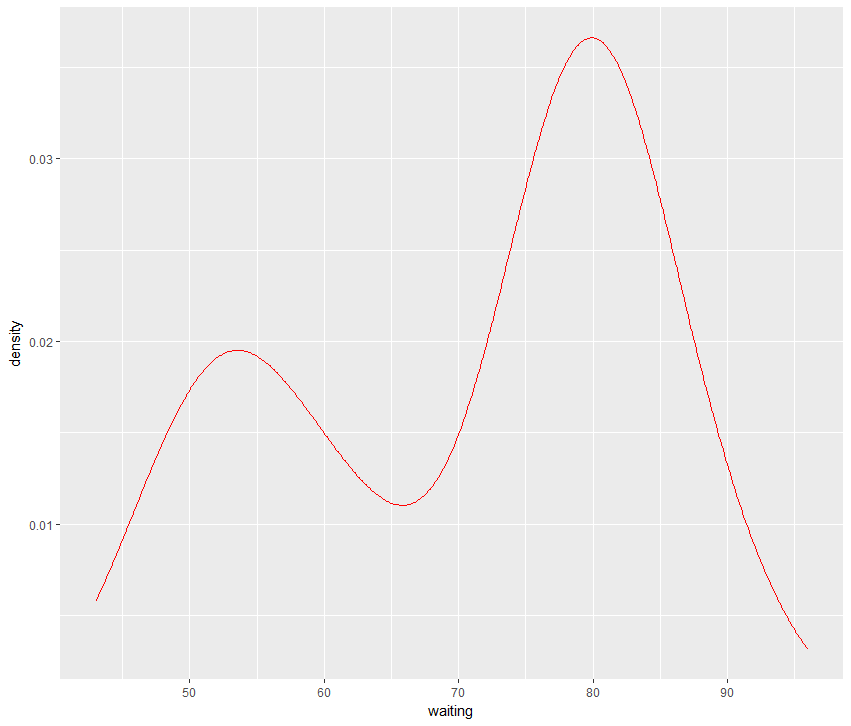
> ggplot(cdat, aes(x=healthexp, y=infmortality))+geom\_point()+geom\_text(aes(x=healthexp+100, label=Name1), size=4, hjust=0)+xlim(2000,10000)



#라인 그래프

data("faithful")

ggplot(faithful, aes(x=waiting))+geom\_line(stat="density", colour="red")

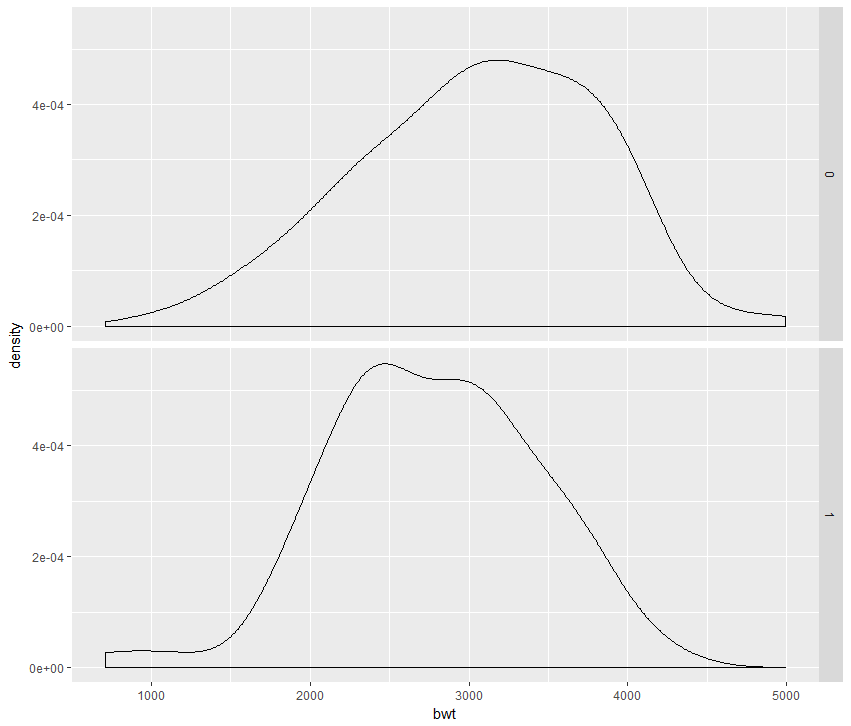


> library(MASS)

> data(birthwt)

ggplot(birthwt, aes(x=bwt))+geom\_density()+facet\_grid(smoke~.)

#Facet분할됨



> library(plyr)

> tg<-ddply(ToothGrowth,c("supp","dose"), summarise, length=mean(len))

> str(tg)

'data.frame': 6 obs. of 3 variables:

$ supp : Factor w/ 2 levels "OJ","VC": 1 1 1 2 2 2

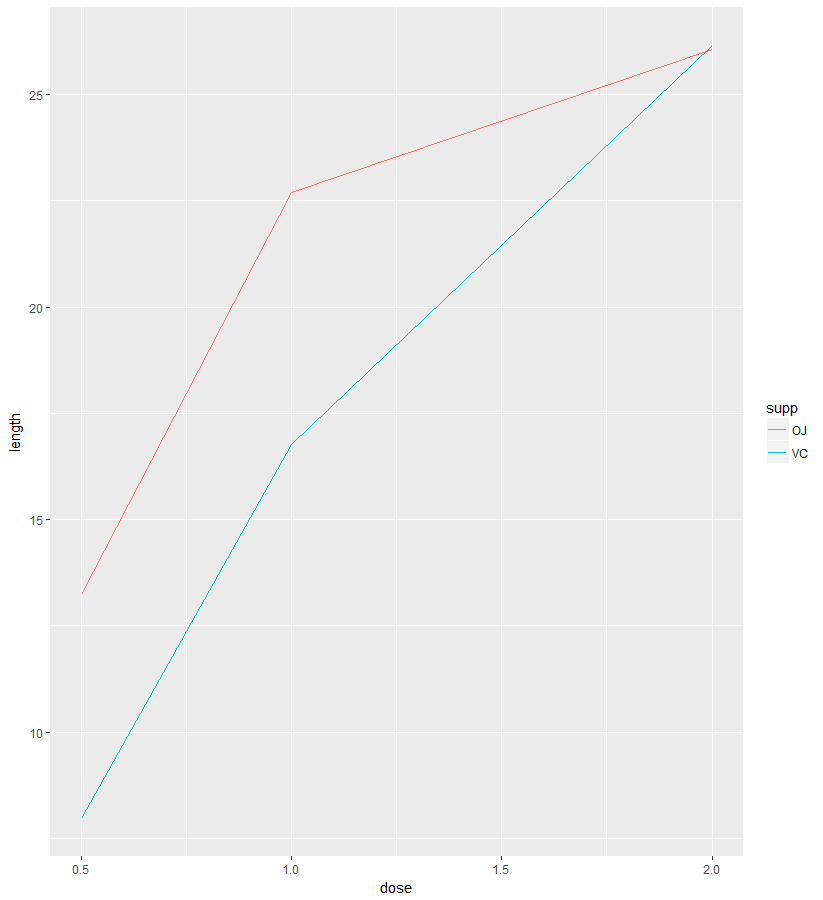
$ dose : num 0.5 1 2 0.5 1 2

$ length: num 13.23 22.7 26.06 7.98 16.77 ...

> View(tg)

그룹별 group=supp

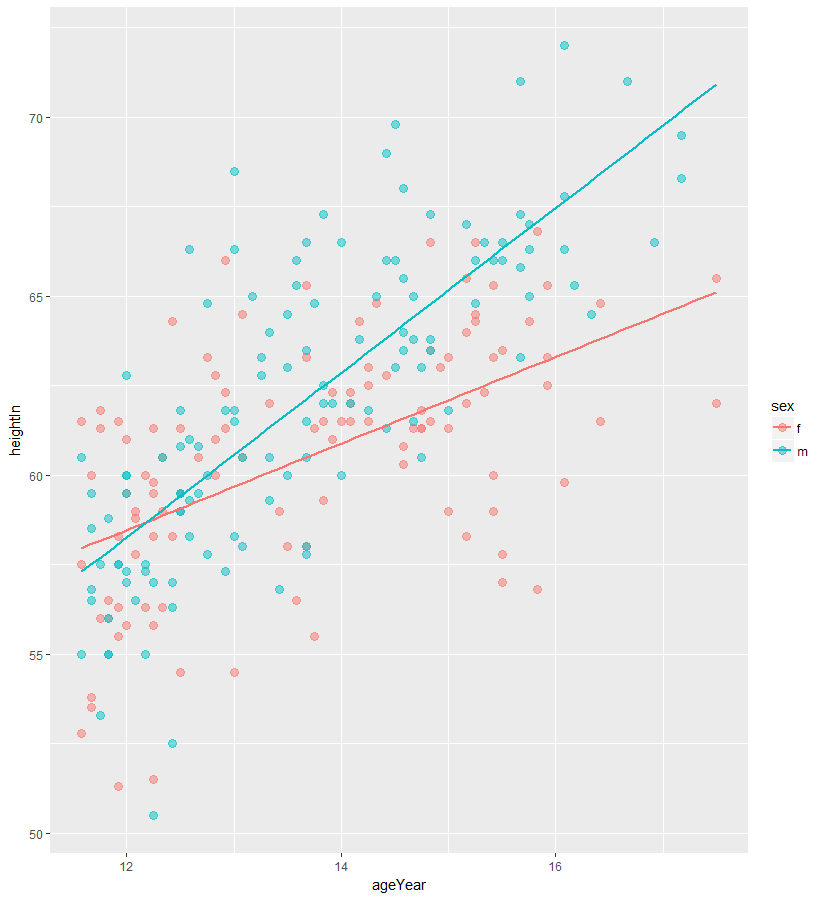
> ggplot(tg, aes(x=dose, y=length, colour=supp))+geom\_line()



p+stat\_smooth()#비선형 모델로 적합

**p+stat\_smooth(method=loess)#**loess(locally weighted polynomial)

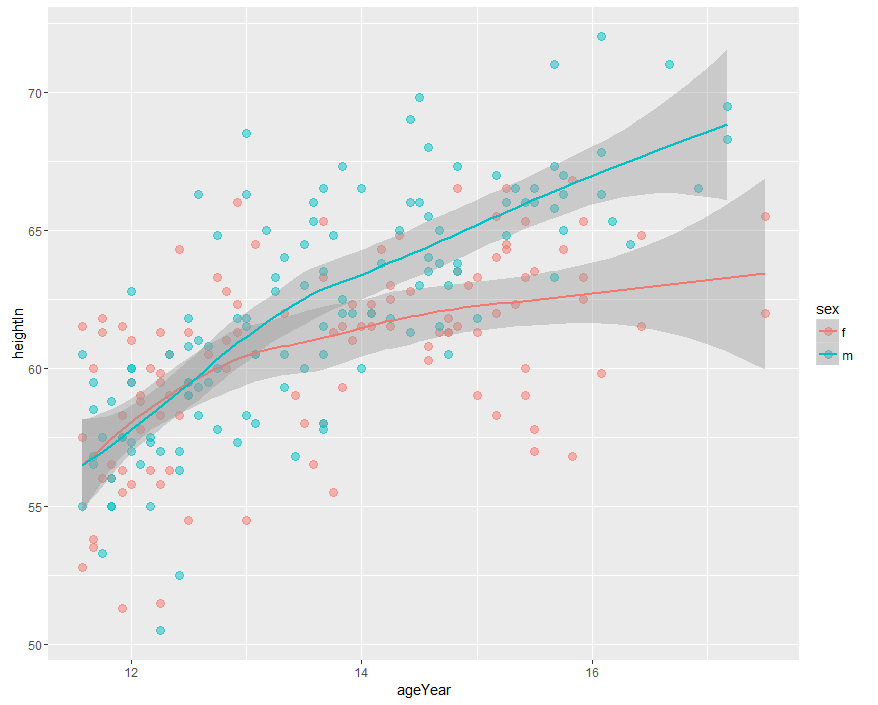
ggplot(heightweight,aes(x=ageYear,y=heightIn,**colour=sex))**+geom\_point(size=3,alpha=.5)+**geom\_smooth(method=loess)**



집단별로 적합선이 하나씩그려짐, stat=\_smooth기능이 예측변수 데이터 범위로 x축 상 범위로만 사용됨, 외사법을 사용하더라도 회귀선이 오른쪽 끝까지 표시되지 않음.,Fullrange=T로 하면 데이터의 끝까지 표시가됨

ggplot(heightweight,aes(x=ageYear,y=heightIn,colour=sex))+geom\_point(size=3, alpha=.5)+geom\_smooth(method=lm, **fullrange=T, se=F)**

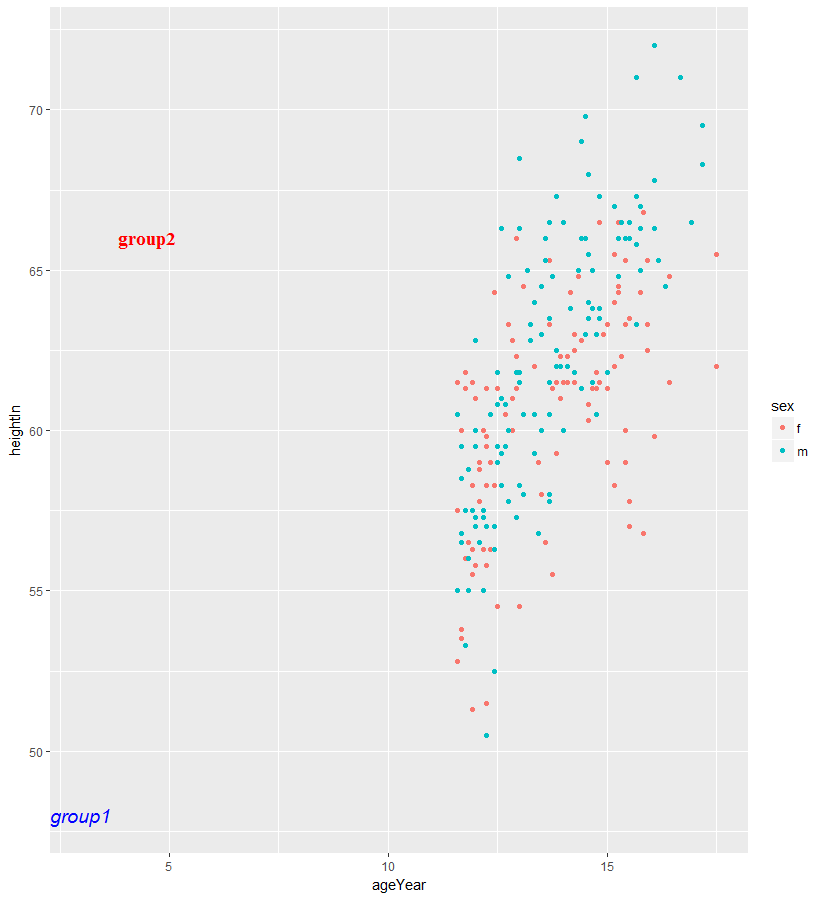
ggplot(heightweight,aes(x=ageYear,y=heightIn,**colour=sex))**+geom\_point(size=3,alpha=.5)+**geom\_smooth(method=loess)**



집단별로 적합선이 하나씩그려짐, stat=\_smooth기능이 예측변수 데이터 범위로 x축 상 범위로만 사용됨, 외사법을 사용하더라도 회귀선이 오른쪽 끝까지 표시되지 않음.,Fullrange=T로 하면 데이터의 끝까지 표시가됨

**1.텍스트 주석넣기(수식사용/분할면에 주석넣기)**

p=ggplot(heightweight, aes(x=ageYear, y=heightIn, colour=sex))+geom\_point()



p+annotate("text", x=3, y=48, label="group1",fontface="italic", colour="blue",size=5)+annotate("text",x=4.5,y=66,label="group2",**family="serif",fontface="bold",**colour="red",size=5)

**x, y축 지정, family=글씨체, fondface=굵음,이탤릭체**