개요

측정형 변수의 함수 관계(직선 관계만 분석하는 것이 선형회귀모형)를 시각적으로 표현 - 두 측정형 변수의 직선관계에 대한 척도(상관계수) [상관분석강의노트 참고]

종속변수(Y), 설명변수(X)의 함수 관계 표현 - 굳이 직선일 필요는 없음

미국 MLB 타자 연봉 데이터 (http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/baseball.csv)

타자 연봉 및 능력 데이터

```
baseball<-read.csv('http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/baseball.csv')
dim(baseball); names(baseball)
```

```
> dim(baseball)
[1] 322  23
> dim(baseball); names(baseball)
[1] 322  23
 [1] "Player_Name"      "Team"             "TimesatBat"       "Hits"
 [5] "HomeRuns"         "Runs"             "RBIs"             "Walks"
 [9] "YearsinMLB"       "CareerTimesatBat" "CareerHits"       "CareerHome"
[13] "CareerRuns"       "CareerRBIs"       "CareerWalks"      "League"
[17] "Division"         "Position"         "PutOuts"          "Assists"
[21] "Errors"           "Salary"           "League2"
```

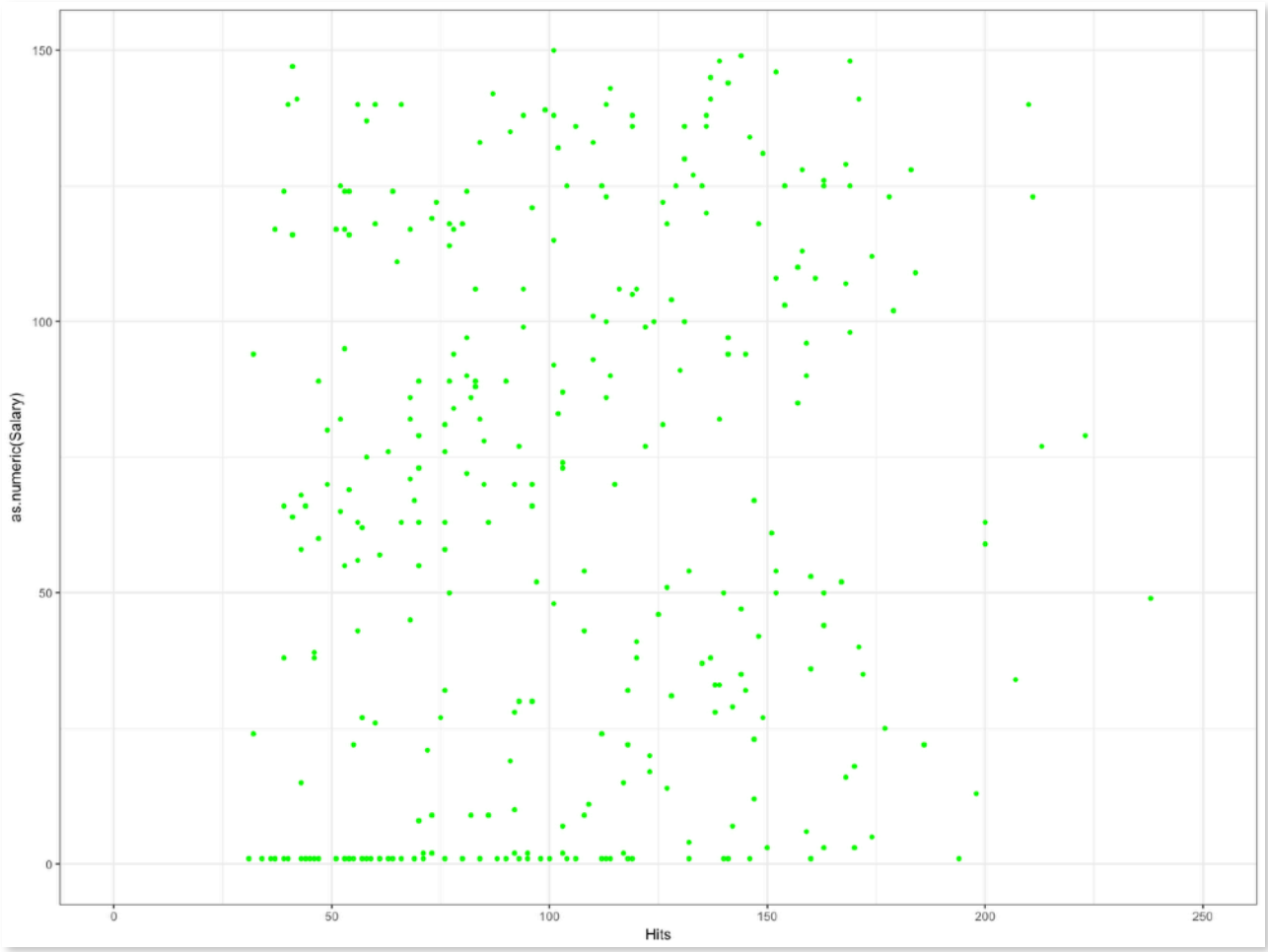
타자 안타수와 연봉의 관계

연봉(salary) 결측치(na)가 있어 as.numeric() 함수를 이용하여 숫자형 변수로 만들어야 한다. 결측치가 없는 경우는 굳이 as.numeric() 함수를 사용할 필요가 없음

```
summary(as.numeric(baseball$Salary))
summary(baseball$Hits)
library(ggplot2)
ggplot(baseball, aes(x=Hits,y=as.numeric(Salary))) +
  geom_point(shape=16,col='green') +
  xlim(0,250) + ylim(0,150) +
  theme_bw()
```

```
> summary(as.numeric(baseball$Salary))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   12.25   63.50   64.17  108.75  150.00
> summary(baseball$Hits)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   31.0    68.0    98.5   103.4   137.8   238.0
```

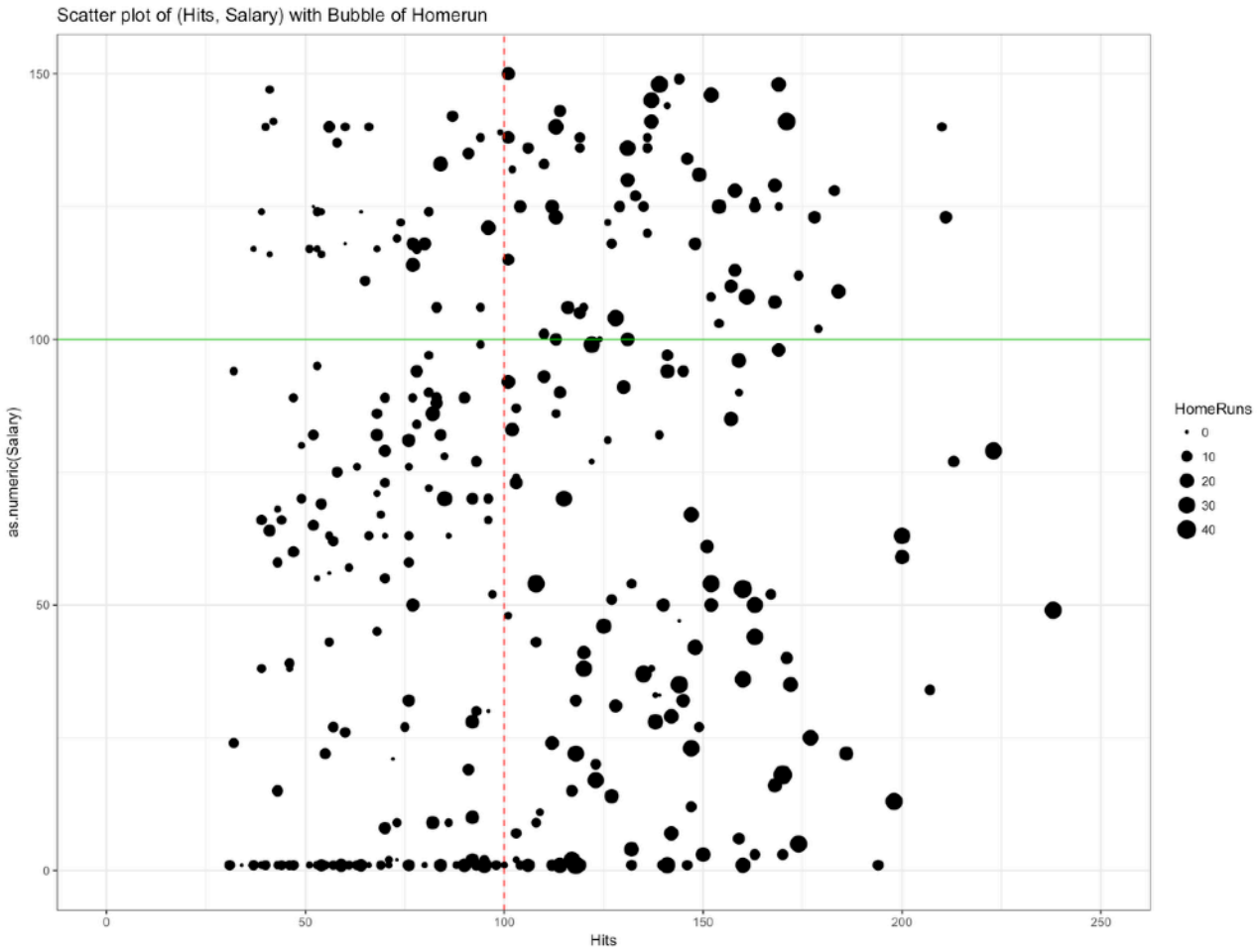
축의 범위를 결정하기 위하여 미리 최소, 최대값을 summary() 함수 이용하여 출력한다.


150
100
50
0
as.numeric(Salary)
0
50
100
150
200
250
Hits

Bubble plot : 타자 (안타수와 연봉의 관계) - 점의 크기를 홈런 개수

```
library(ggplot2)
ggplot(baseball, aes(x=Hits,y=as.numeric(Salary))) +
  geom_point(shape=16,aes(size=HomeRuns)) +
  geom_vline(xintercept=100,col=2,lty='dashed') +
  geom_hline(yintercept=100,col=3,lty='solid') +
  xlim(0,250) + ylim(0,150) +
  ggtitle('Scatter plot of (Hits, Salary) with Bubble of Homerun') +
  theme_bw()
```

- geom_vline <- 수직선 vertical line 그리기, lty=line type약어

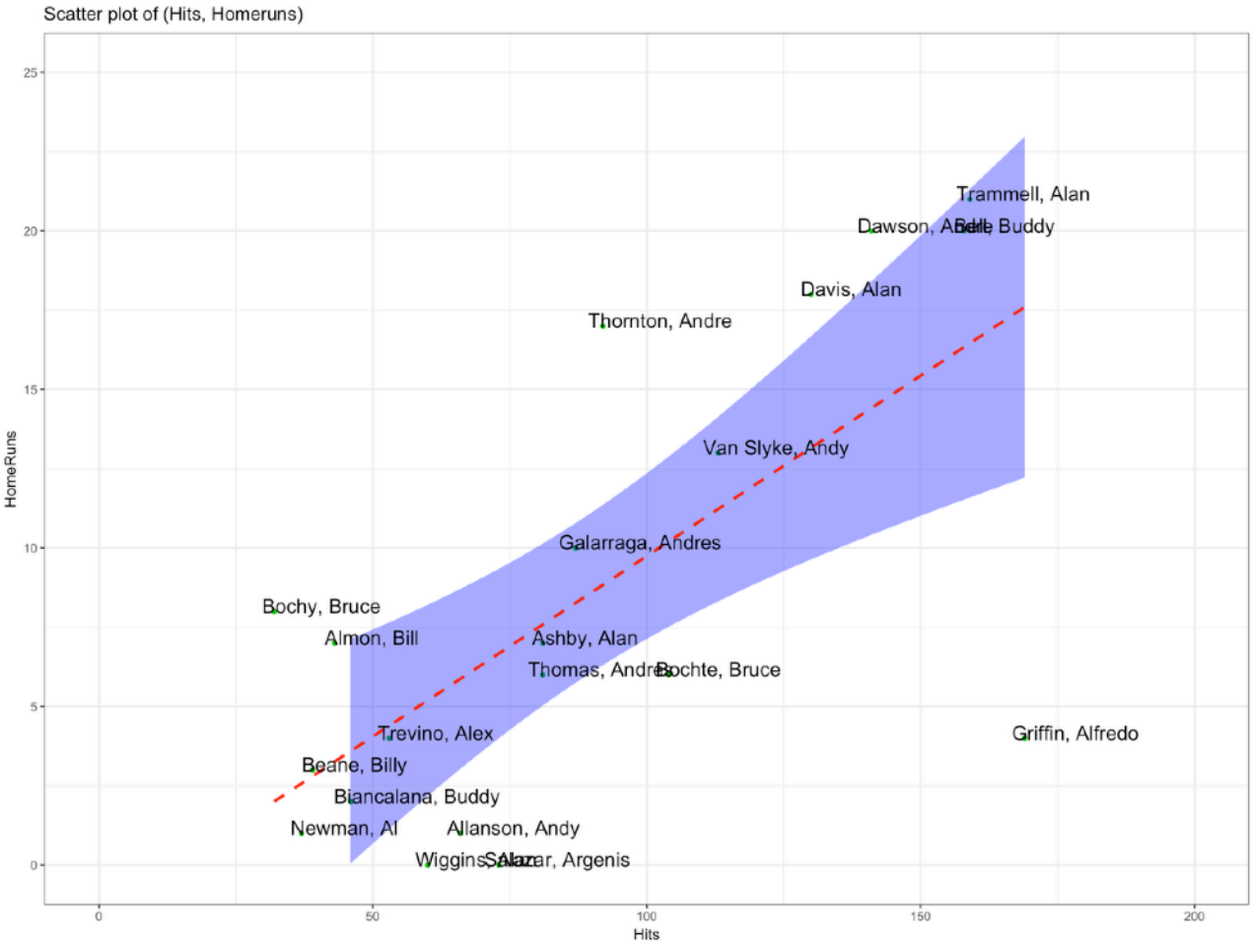


선형 회귀식(method=lm) 및 신뢰구간(fill=) 표현 (개체 이름 표현 label=개체식별변수)

```
baseball0<-baseball[c(1:20), ] #타자 20명만 : 그래프 간편성
summary(baseball0$HomeRuns); summary(baseball0$Hits)
library(ggplot2)
ggplot(baseball0, aes(x=Hits,y=HomeRuns,label=Player_Name)) +
  geom_point(shape=16,col=3) +
  geom_smooth(method=lm,linetype="dashed",color="red", fill="blue") +
  geom_text(size=5,hjust=0.1,vjust=0.1) +
  xlim(0,200) + ylim(0,25) +
  ggtitle('Scatter plot of (Hits, Homeruns)') +
  theme_bw()
```

```
> summary(baseball0$HomeRuns)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    2.75    6.50    8.40   14.00   21.00
> summary(baseball0$Hits)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  32.00   51.25   81.00   88.20  117.25  169.00
```

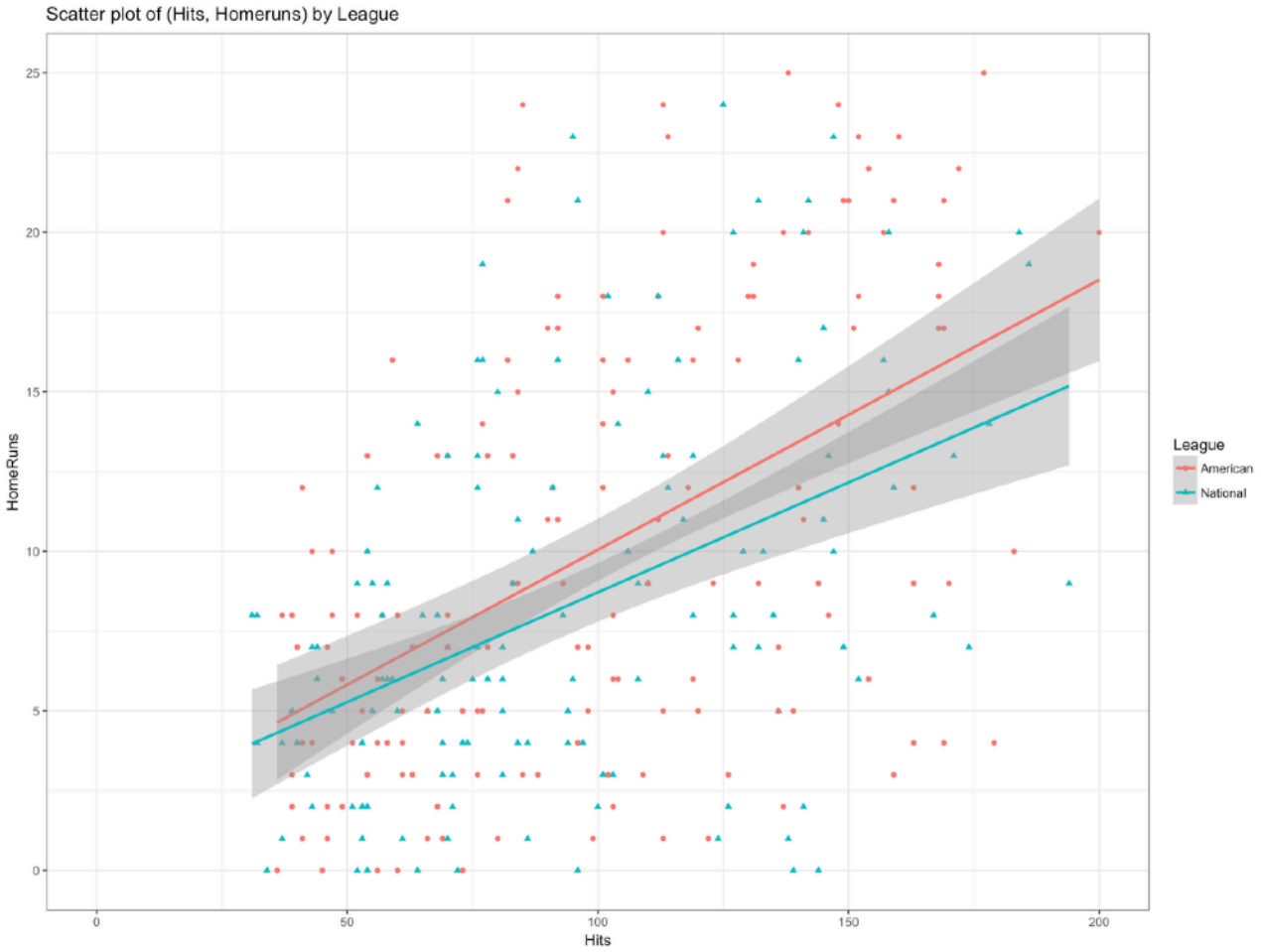
hjust, vjust는 수평(horizon), 수직(vertical) 개체 식별 이름 위치를 지정하게 된다. (0_1 사이값)



집단변수(리그)에 의한 선형회귀모형 산점도(안타수*홈런)

se=T :신뢰구간을 표현하게 됨, F를 사용하면 표현되지 않음

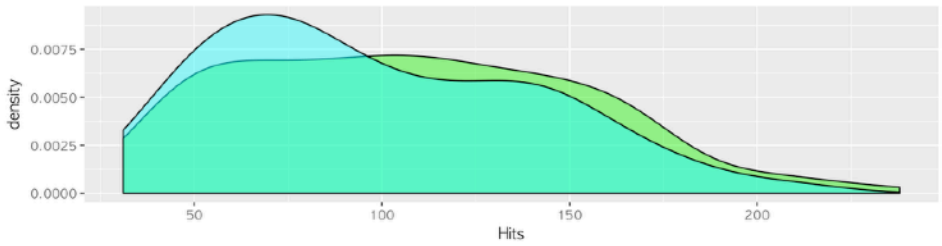
```
library(ggplot2)
ggplot(baseball,aes(x=Hits,y=HomeRuns,color=League,shape=League)) +
  geom_point() +
  geom_smooth(method=lm,se=T) +
  xlim(0,200) + ylim(0,25) +
  ggtitle('Scatter plot of (Hits, Homeruns) by League') +
  theme_bw()
```

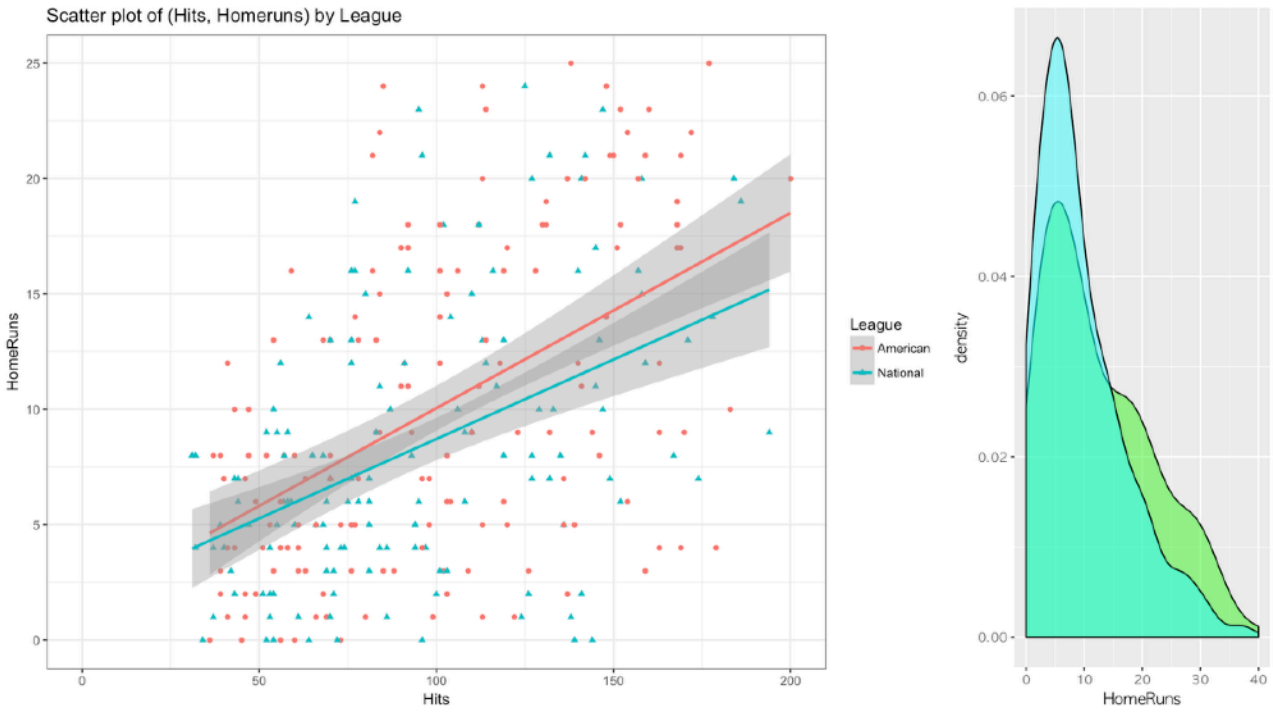


(집단별) 주변밀도함수 표현하기

```
library(ggplot2)
#main scatter plot(left_down)
scatterPlot<-
ggplot(baseball,aes(x=Hits,y=HomeRuns,color=League,shape=League)) +
  geom_point() +
  geom_smooth(method=lm,se=T) +
  xlim(0,200) + ylim(0,25) +
  ggtitle('Scatter plot of (Hits, Homeruns) by League') +
  theme_bw()

# Marginal density plot of x (Hits)
xdensity<-ggplot(baseball,aes(Hits,fill=League)) +
  geom_density(alpha=.5) +
  scale_fill_manual(values = c('green','cyan')) +
  theme(legend.position = "none")
xdensity
# Marginal density plot of y (Homeruns)
ydensity<-ggplot(baseball,aes(HomeRuns,fill=League)) +
  geom_density(alpha=.5) +
  scale_fill_manual(values = c('green','cyan')) +
  theme(legend.position = "none")
ydensity

#no plot : right-up
blankPlot<-ggplot() + theme_bw()

#install.packages('gridExtra')
library("gridExtra")
grid.arrange(xdensity, blankPlot, scatterPlot, ydensity,
             ncol=2, nrow=2, widths=c(4, 1.4), heights=c(1.4, 4))
```

4개의 ggplot() 을 한 화면에 그리게 됨. 오른쪽 위의 빈 플롯에 다른 ggplot 넣어도 됨


density
0.0075
0.0050
0.0025
0.0000
50
100
150
200
Hits


Scatter plot of (Hits, Homeruns) by League
HomeRuns
25
20
15
10
5
0
0
50
100
150
200
Hits
League
American
National

density
0.06
0.04
0.02
0.00
0
10
20
30
40
HomeRuns