Exploratory Data Analysis for custdata

<https://github.com/WinVector/zmPDSwR/tree/master/Custdata> 에서 데이터 다운

* 분석목표: 의료보험 패키지를 팔기 위해서 의료보험이 없는 고객을 예측하는 모형 생성
* 나이, 고용상태, 수입, 거주지/차량 등에 대한 정보 등

setwd("D:/Dropbox/PDSwR/Custdata")  
custdata <- read.table('custdata.tsv', header=TRUE, sep='\t')  
  
summary(custdata)

custid sex is.employed income   
 Min. : 2068 F:440 Mode :logical Min. : -8700   
 1st Qu.: 345667 M:560 FALSE:73 1st Qu.: 14600   
 Median : 693403 TRUE :599 Median : 35000   
 Mean : 698500 NA's :328 Mean : 53505   
 3rd Qu.:1044606 3rd Qu.: 67000   
 Max. :1414286 Max. :615000   
   
 marital.stat health.ins   
 Divorced/Separated:155 Mode :logical   
 Married :516 FALSE:159   
 Never Married :233 TRUE :841   
 Widowed : 96 NA's :0   
   
   
   
 housing.type recent.move num.vehicles   
 Homeowner free and clear :157 Mode :logical Min. :0.000   
 Homeowner with mortgage/loan:412 FALSE:820 1st Qu.:1.000   
 Occupied with no rent : 11 TRUE :124 Median :2.000   
 Rented :364 NA's :56 Mean :1.916   
 NA's : 56 3rd Qu.:2.000   
 Max. :6.000   
 NA's :56   
 age state.of.res  
 Min. : 0.0 California :100   
 1st Qu.: 38.0 New York : 71   
 Median : 50.0 Pennsylvania: 70   
 Mean : 51.7 Texas : 56   
 3rd Qu.: 64.0 Michigan : 52   
 Max. :146.7 Ohio : 51   
 (Other) :600

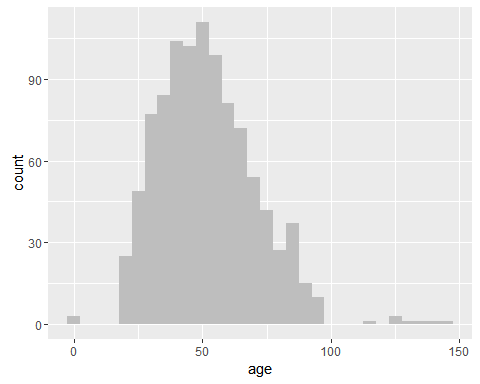
# income을 1000불 단위로 변환  
Income = custdata$income/1000  
summary(Income)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 -8.7 14.6 35.0 53.5 67.0 615.0

* is.employed 변수는 결측치가 많음. 고용상태가 알려지지 않은 것인지, 학생이나 은퇴한 사람처럼 현재 일을 하고 있지 않은지?
* income의 음수값? 빚을 진것을 나타내는지? => 어떻게 처리할 것인가?
* health.ins를 보면 대략 84%의 고객이 보험을 가짐
* housing.type, recent.move, num.vehicles 변수는 각각 56개의 결측값을 가짐 => 많지 않으므로 관측치 제거
* age의 최대값과 최소값이 이상함
* state.of.res는 고객의 주를 나타내는 범주형 변수

# 일변량 탐색

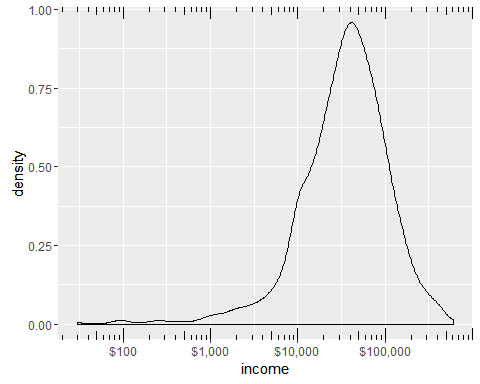
library(ggplot2)   
  
ggplot(custdata) +  
 geom\_histogram(aes(x=age), binwidth=5, fill="gray")

 age에서 0와 100이상

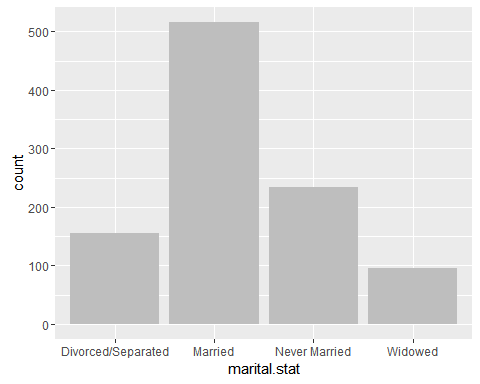
library(scales)  
  
# income에 대한 밀도 함수  
ggplot(custdata) + geom\_density(aes(x=income)) +  
 scale\_x\_continuous(labels=dollar)



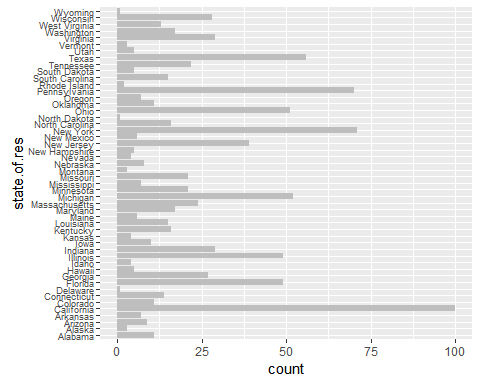
# income의 로그변환에 대한 밀도 함수  
ggplot(custdata) + geom\_density(aes(x=income)) +  
 scale\_x\_log10(breaks=c(100,1000,10000,100000), labels=dollar) +   
 annotation\_logticks(sides="bt")



# marital.stat에 대한 막대그래프  
ggplot(custdata) + geom\_bar(aes(x=marital.stat), fill="gray")



# 주별 도수  
ggplot(custdata) +  
 geom\_bar(aes(x=state.of.res), fill="gray") +   
 coord\_flip() +   
 theme(axis.text.y=element\_text(size=rel(0.8)))



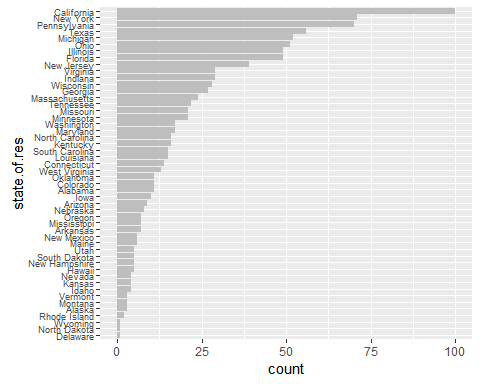
# 보기 편하게 도수가 많은 순서로 정렬해 보자  
statesums <- table(custdata$state.of.res)   
statef <- as.data.frame(statesums)   
colnames(statef)<-c("state.of.res", "count")   
summary(statef)

## state.of.res count   
## Alabama : 1 Min. : 1.00   
## Alaska : 1 1st Qu.: 5.00   
## Arizona : 1 Median : 12.00   
## Arkansas : 1 Mean : 20.00   
## California: 1 3rd Qu.: 26.25   
## Colorado : 1 Max. :100.00   
## (Other) :44

statef <- transform(statef,  
 state.of.res=reorder(state.of.res, count))   
summary(statef)

## state.of.res count   
## Delaware : 1 Min. : 1.00   
## North Dakota: 1 1st Qu.: 5.00   
## Wyoming : 1 Median : 12.00   
## Rhode Island: 1 Mean : 20.00   
## Alaska : 1 3rd Qu.: 26.25   
## Montana : 1 Max. :100.00   
## (Other) :44

ggplot(statef)+ geom\_bar(aes(x=state.of.res,y=count),  
 stat="identity",   
 fill="gray") +  
 coord\_flip() +   
 theme(axis.text.y=element\_text(size=rel(0.8)))

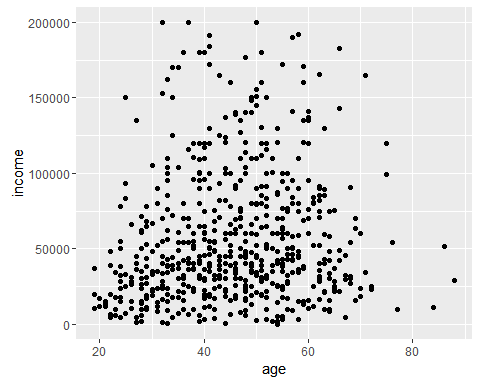


# 이변량 탐색

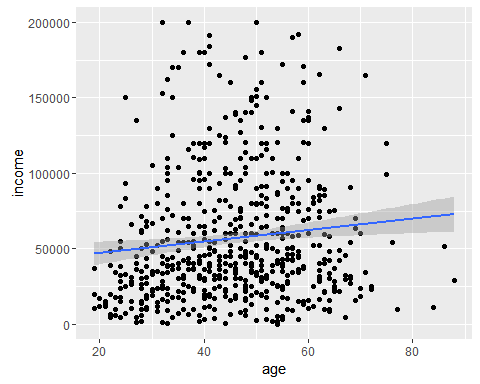
# 이상치 제거  
custdata = na.omit(custdata)  
custdata2 <- subset(custdata,  
 (custdata$age > 0 & custdata$age < 100  
 & custdata$income > 0))   
  
# age와 income간의 상관계수: 0에 가까움  
cor(custdata2$age, custdata2$income)

## [1] 0.1423229

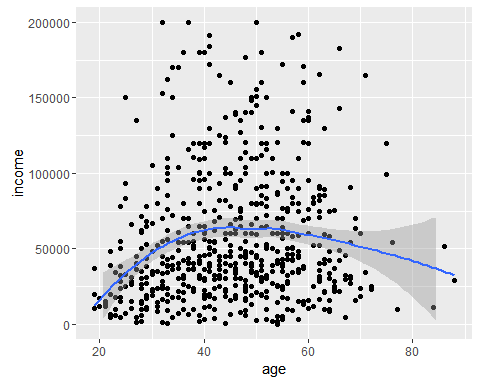
# 산점도  
ggplot(custdata2, aes(x=age, y=income)) +  
 geom\_point() + ylim(0, 200000)



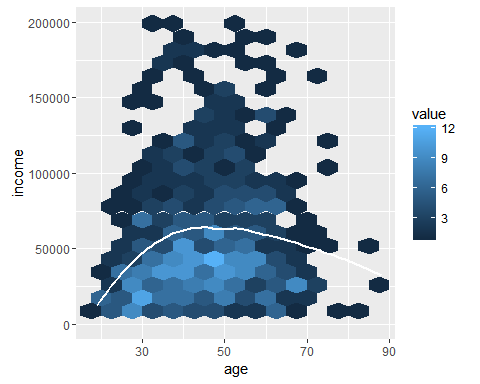
# 선형 추세선 추가  
ggplot(custdata2, aes(x=age, y=income)) + geom\_point() +  
 stat\_smooth(method="lm") +  
 ylim(0, 200000)



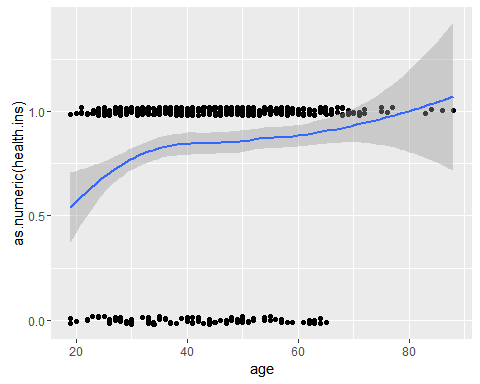
# 비선형 추세선 추가  
ggplot(custdata2, aes(x=age, y=income)) +  
 geom\_point() + geom\_smooth() +  
 ylim(0, 200000)



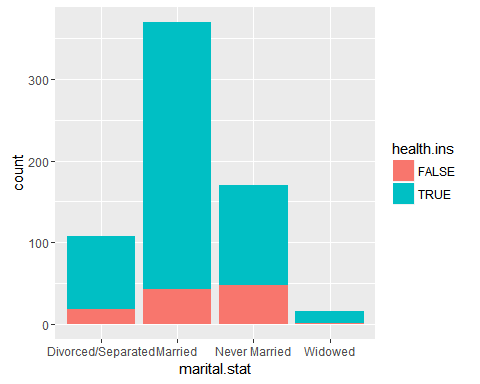
# 40세 전후   
  
# hexbin: 2차원 히스토그램과 유사(색으로 빈도 표현)  
library(hexbin)   
  
ggplot(custdata2, aes(x=age, y=income)) +  
 geom\_hex(binwidth=c(5, 10000)) +   
 geom\_smooth(color="white", se=F) +   
 ylim(0,200000)



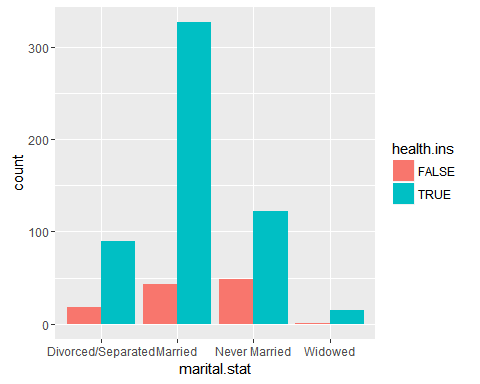
# age와 health.ins(보험 가입 여부)  
ggplot(custdata2, aes(x=age, y=as.numeric(health.ins))) +   
 geom\_point(position=position\_jitter(w=0.05, h=0.05)) +   
 geom\_smooth()



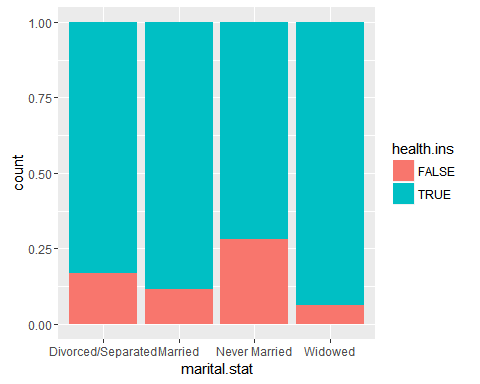
# marital.stat과 health.ins에 대한 여러가지 막대그래프  
ggplot(custdata) + geom\_bar(aes(x=marital.stat, # 디폴트: stacked  
 fill=health.ins))



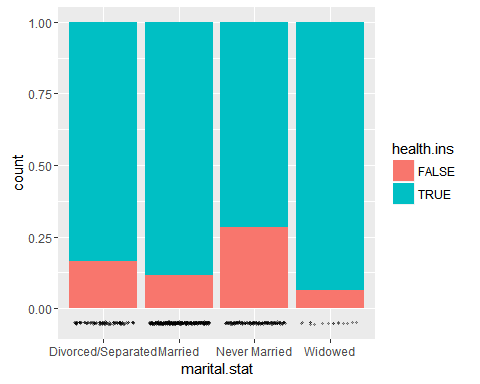
ggplot(custdata) + geom\_bar(aes(x=marital.stat, # side-by-side  
 fill=health.ins),  
 position="dodge")



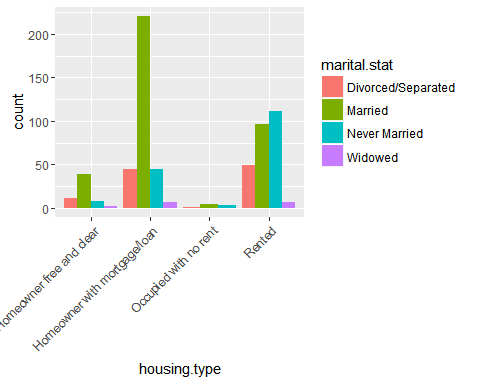
ggplot(custdata) + geom\_bar(aes(x=marital.stat, # filled  
 fill=health.ins),  
 position="fill")



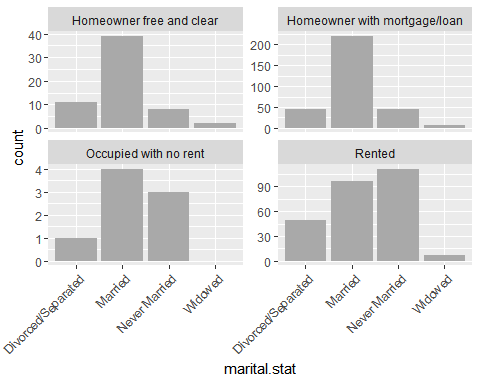
# rug 추가  
ggplot(custdata, aes(x=marital.stat)) +  
 geom\_bar(aes(fill=health.ins), position="fill") +  
 geom\_point(aes(y=-0.05), size=0.75, alpha=0.3,   
 position=position\_jitter(h=0.01)) # 점들 잘보이게 jitering



# 기혼자가 일반적이고 사별한 경우는 상대적으로 적음  
  
# housing.type과 marital.stat에 대한 side-by-side 막대그래프  
# Occupied with no rent가 잘 안보이네  
ggplot(custdata2) +   
 geom\_bar(aes(x=housing.type, fill=marital.stat ),  
 position="dodge") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# housing.type마다 다른 스케일로 막대그래프  
ggplot(custdata2) +   
 geom\_bar(aes(x=marital.stat), position="dodge",  
 fill="darkgray") +  
 facet\_wrap(~housing.type, scales="free\_y") +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# 연습문제

데이터를 더 탐색해보고 정제하시오.