Chapter 4. Managing Data

coop711

2015년 10월 4일

Cleaning Data

- Data Preparation
    - `zmPDSwR-master/Custdata/exampleData.rData` 를 필요로 함.
    - working directory 정리

```
rm(list=ls())
ls()
```

```
## character(0)
```

```
load("../../zmPDSwR-master/Custdata//exampleData.rData")
ls()
```

```
## [1] "custdata"     "hhdata"       "medianincome"
```

```
str(custdata)
```

```
## 'data.frame':    1000 obs. of  19 variables:
##  $ state.of.res    : Factor w/ 50 levels "Alabama","Alaska",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ custid          : int  1063014 1192089 16551 1079878 502705 674271 15917
467335 462569 1216026 ...
##  $ sex             : Factor w/ 2 levels "F","M": 1 2 1 1 2 2 1 2 2 2 ...
##  $ is.employed     : logi  TRUE NA NA NA TRUE FALSE ...
##  $ income          : int  82000 49000 7000 37200 70000 0 24000 42600 22000 9
600 ...
##  $ marital.stat    : Factor w/ 4 levels "Divorced/Separated",..: 2 2 2 1 2 2
1 3 4 3 ...
##  $ health.ins      : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
##  $ housing.type    : Factor w/ 4 levels "Homeowner free and clear",..: 4 1 2
2 4 4 1 4 1 4 ...
##  $ recent.move     : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ num.vehicles    : int  2 2 2 1 4 1 1 1 0 6 ...
##  $ age             : num  43 77 46 62 37 54 70 33 89 50 ...
##  $ is.employed.fix1: chr  "employed" "missing" "missing" "missing" ...
##  $ age.normalized  : num  -0.461 1.341 -0.302 0.546 -0.779 ...
##  $ Median.Income   : num  52371 52371 52371 52371 52371 ...
##  $ income.norm     : num  1.566 0.936 0.134 0.71 1.337 ...
##  $ gp              : num  0.935 0.116 0.991 0.187 0.849 ...
##  $ income.lt.30K   : logi  FALSE FALSE TRUE FALSE FALSE TRUE ...
##  $ age.range       : Factor w/ 3 levels "[0,25]","(25,65]",..: 2 3 2 2 2 2 3
2 3 2 ...
##  $ Income          : num  NA NA 4500 20000 12000 180000 120000 40000 NA 2400
0 ...
```

```
custdata.2 <- custdata
(v.to.add <- c("age.normalized", "Median.Income", "income.norm", "gp", "incom
e.lt.30K", "age.range"))
```

```
## [1] "age.normalized" "Median.Income"  "income.norm"    "gp"
## [5] "income.lt.30K"  "age.range"
```

```
#(index.to.add <- which(names(custdata) %in% v.to.add))
(v.to.retain <- setdiff(names(custdata), v.to.add))
```

```
##  [1] "state.of.res"     "custid"           "sex"
##  [4] "is.employed"      "income"           "marital.stat"
##  [7] "health.ins"       "housing.type"     "recent.move"
## [10] "num.vehicles"     "age"              "is.employed.fix1"
## [13] "Income"
```

```
#custdata <- custdata[-index.to.add]
custdata <- custdata[v.to.retain]
str(custdata)
```

```
## 'data.frame':    1000 obs. of  13 variables:
##  $ state.of.res    : Factor w/ 50 levels "Alabama","Alaska",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ custid          : int  1063014 1192089 16551 1079878 502705 674271 15917
467335 462569 1216026 ...
##  $ sex             : Factor w/ 2 levels "F","M": 1 2 1 1 2 2 1 2 2 2 ...
##  $ is.employed     : logi  TRUE NA NA NA TRUE FALSE ...
##  $ income          : int  82000 49000 7000 37200 70000 0 24000 42600 22000 9
600 ...
##  $ marital.stat    : Factor w/ 4 levels "Divorced/Separated",..: 2 2 2 1 2 2
1 3 4 3 ...
##  $ health.ins      : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
##  $ housing.type    : Factor w/ 4 levels "Homeowner free and clear",..: 4 1 2
2 4 4 1 4 1 4 ...
##  $ recent.move     : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ num.vehicles    : int  2 2 2 1 4 1 1 1 0 6 ...
##  $ age             : num  43 77 46 62 37 54 70 33 89 50 ...
##  $ is.employed.fix1: chr  "employed" "missing" "missing" "missing" ...
##  $ Income          : num  NA NA 4500 20000 12000 180000 120000 40000 NA 2400
0 ...
```

Treating missing values

- Checking locations fo missing data

```
options(width=132)
summary(custdata[is.na(custdata$housing.type), c("recent.move", "num.vehicle
s")])
```

```
##  recent.move     num.vehicles
##  Mode:logical   Min.   : NA
##  NA's:56        1st Qu.: NA
##                 Median : NA
##                 Mean   :NaN
##                 3rd Qu.: NA
##                 Max.   : NA
##                 NA's   :56
```

- `is.employed` 변수 확인

```
str(custdata)
```


```
## 'data.frame':    1000 obs. of  13 variables:
##  $ state.of.res    : Factor w/ 50 levels "Alabama","Alaska",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ custid          : int  1063014 1192089 16551 1079878 502705 674271 15917
467335 462569 1216026 ...
##  $ sex             : Factor w/ 2 levels "F","M": 1 2 1 1 2 2 1 2 2 2 ...
##  $ is.employed     : logi  TRUE NA NA NA TRUE FALSE ...
##  $ income          : int  82000 49000 7000 37200 70000 0 24000 42600 22000 9
600 ...
##  $ marital.stat    : Factor w/ 4 levels "Divorced/Separated",..: 2 2 2 1 2 2
1 3 4 3 ...
##  $ health.ins      : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
##  $ housing.type    : Factor w/ 4 levels "Homeowner free and clear",..: 4 1 2
2 4 4 1 4 1 4 ...
##  $ recent.move     : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ num.vehicles    : int  2 2 2 1 4 1 1 1 0 6 ...
##  $ age             : num  43 77 46 62 37 54 70 33 89 50 ...
##  $ is.employed.fix1: chr  "employed" "missing" "missing" "missing" ...
##  $ Income          : num  NA NA 4500 20000 12000 180000 120000 40000 NA 2400
0 ...
```


```
summary(custdata[c("housing.type", "recent.move", "num.vehicles", "is.employe
d")])
```

```
##                        housing.type recent.move      num.vehicles   is.emplo
yed
##  Homeowner free and clear    :157   Mode :logical   Min.   :0.000   Mode :lo
gical
##  Homeowner with mortgage/loan:412   FALSE:820       1st Qu.:1.000   FALSE:73
##  Occupied with no rent       : 11   TRUE :124       Median :2.000   TRUE :59
9
##  Rented                      :364   NA's :56        Mean   :1.916   NA's :32
8
##  NA's                        : 56                   3rd Qu.:2.000
##                                                     Max.   :6.000
##                                                     NA's   :56
```

- `is.employed` 의 NA를 `missing` 이라는 새로운 범주로 설정

```
custdata$is.employed.fix <- ifelse(is.na(custdata$is.employed), "missing", ifel
se(custdata$is.employed == TRUE, "employed", "not employed"))
summary(custdata$is.employed.fix)
```

```
##    Length     Class      Mode
##      1000 character character
```

```
summary(factor(custdata$is.employed.fix))
```

```
##     employed      missing not employed
##          599          328           73
```

```
summary(as.factor(custdata$is.employed.fix))
```

```
##     employed      missing not employed
##          599          328           73
```

```
summary(factor(custdata$is.employed.fix, levels=c("employed", "not employed",
"missing")))
```

```
##     employed not employed      missing
##          599           73          328
```

```
# summary(as.factor(custdata$is.employed.fix, levels=c("employed", "not employe
d", "missing")))
```

- `missing` 의 성격 파악, not in the active workforce? (from the summary of age)

```
summary(custdata[custdata$is.employed.fix=="missing", ])
```

```
##         state.of.res     custid        sex     is.employed        income
marital.stat health.ins
##  California   : 43   Min.   :   2068   F:172   Mode:logical   Min.   :     0
Divorced/Separated: 47   Mode :logical
##  New York     : 33   1st Qu.: 314974   M:156   NA's:328       1st Qu.:  1550
Married           :145   FALSE:48
##  Ohio         : 21   Median : 623182                          Median : 14450
Never Married     : 56   TRUE :280
##  Pennsylvania : 17   Mean   : 684007                          Mean   : 27524
Widowed           : 80   NA's :0
##  Michigan     : 15   3rd Qu.:1050329                          3rd Qu.: 31650
##  Massachusetts: 14   Max.   :1412971                          Max.   :269000
##  (Other)      :185
##                        housing.type recent.move      num.vehicles        age
is.employed.fix1       Income
##  Homeowner free and clear    :96    Mode :logical   Min.   :0.000   Min.   :
0.00   Length:328         Min.   :     0
##  Homeowner with mortgage/loan:89    FALSE:257       1st Qu.:1.000   1st Qu.:
49.00   Class :character   1st Qu.: 24000
##  Occupied with no rent       : 3    TRUE :23        Median :2.000   Median :
67.00   Mode  :character   Median : 45000
##  Rented                      :92    NA's :48        Mean   :1.643   Mean   :
63.22                      Mean   : 62990
##  NA's                        :48                    3rd Qu.:2.000   3rd Qu.:
78.00                      3rd Qu.: 80000
##                                                     Max.   :6.000   Max.
:123.06                      Max.   :388000
##                                                     NA's   :48
NA's   :99
##  is.employed.fix
##  Length:328
##  Class :character
##  Mode  :character
##
##
##
##
```

- Rename

```
custdata$is.employed.fix <- ifelse(is.na(custdata$is.employed), "not in active
workforce", ifelse(custdata$is.employed == TRUE, "employed", "not employed"))
summary(factor(custdata$is.employed.fix))
```

```
##               employed           not employed not in active workforce
##                    599                     73                      328
```

- Missing values is numeric data

```
summary(custdata$Income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   25000   45000   66200   82000  615000     328
```

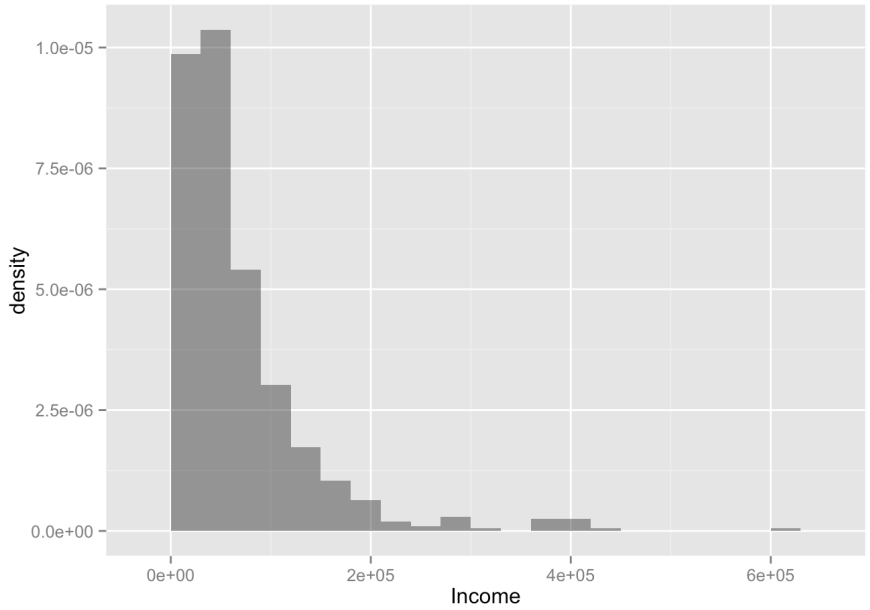
- Imputation

```
mean.income <- mean(custdata$Income, na.rm = TRUE)
Income.fix <- ifelse(is.na(custdata$Income), mean.income, custdata$Income)
summary(Income.fix)
```

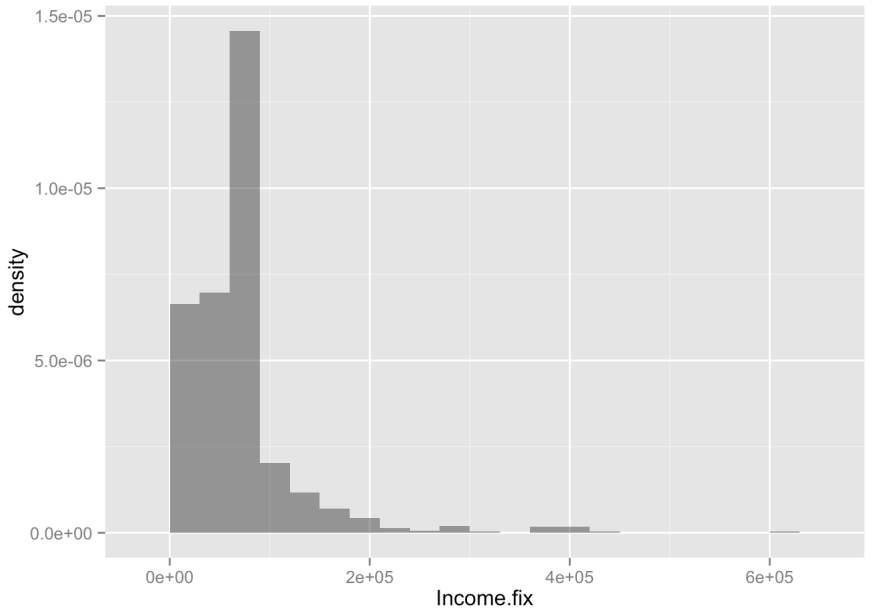
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   35000   66200   66200   66200  615000
```

- graph로 확인

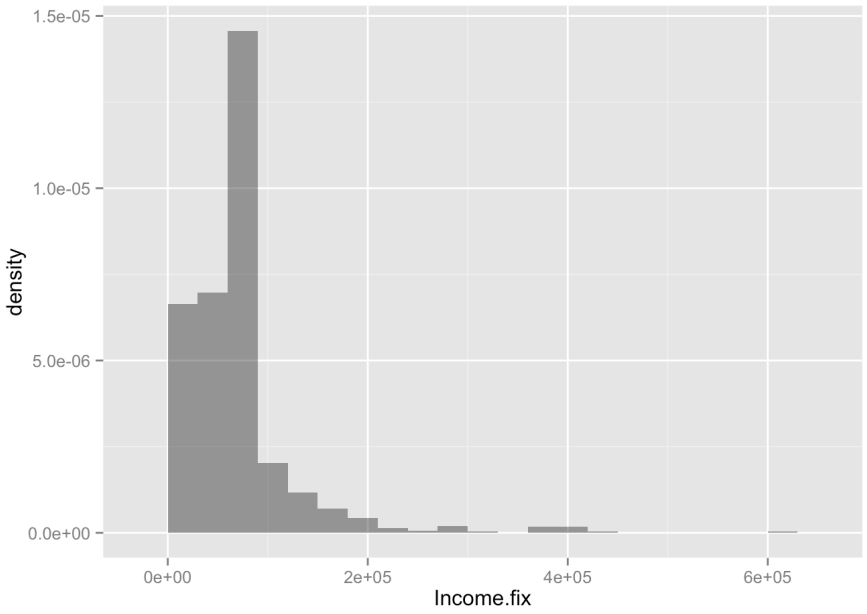
```
library(ggplot2)
ggplot(custdata, aes(x = Income)) + geom_histogram(binwidth=30000, aes(y = ..de
nsity..), alpha=0.5)
```




```
ggplot(data.frame(Income.fix), aes(x = Income.fix)) + geom_histogram(binwidth=3
0000, aes(y = ..density..), alpha=0.5)
```




```
ggplot(data.frame(Income.fix), aes(x = Income.fix)) + geom_bar(stat="bin", binw
idth=30000, aes(y = ..density..), alpha=0.5)
```



- Categorize

```
Income.breaks <- c(0, 10000, 50000, 100000, 250000, 1000000)
Income.groups <- cut(custdata$Income, breaks = Income.breaks, include.lowest =
TRUE)
summary(Income.groups)
```

```
##       [0,1e+04]   (1e+04,5e+04]   (5e+04,1e+05] (1e+05,2.5e+05] (2.5e+05,1
e+06]            NA's
##              63             312             178              98
21             328
```

```
table(Income.groups, useNA = "ifany")
```

```
## Income.groups
##       [0,1e+04]   (1e+04,5e+04]   (5e+04,1e+05] (1e+05,2.5e+05] (2.5e+05,1
e+06]            <NA>
##              63             312             178              98
21             328
```

```
str(Income.groups)
```

```
##  Factor w/ 5 levels "[0,1e+04]","(1e+04,5e+04]",..: NA NA 1 2 2 4 4 2 NA 2
...
```

```
Income.groups <- as.character(Income.groups)
Income.groups <- ifelse(is.na(Income.groups), "no income", Income.groups)
str(Income.groups)
```

```
##  chr [1:1000] "no income" "no income" "[0,1e+04]" "(1e+04,5e+04]" "(1e+04,5
e+04]" "(1e+05,2.5e+05]" ...
```

```
summary(Income.groups)
```

```
##    Length     Class      Mode
##      1000 character character
```

```
summary(factor(Income.groups))
```

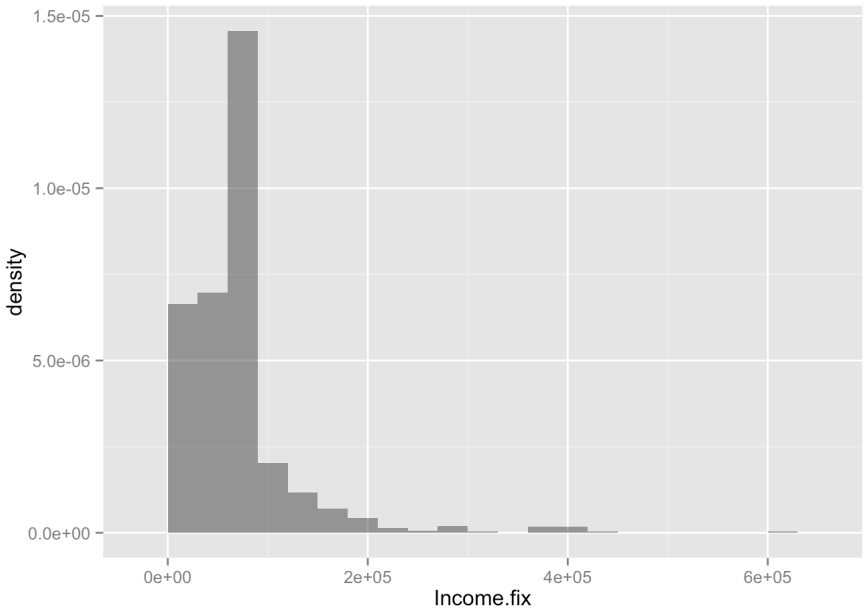
```
##   (1e+04,5e+04] (1e+05,2.5e+05] (2.5e+05,1e+06]   (5e+04,1e+05]       [0,1
e+04]       no income
##             312              98              21             178
63             328
```

```
table(Income.groups)
```

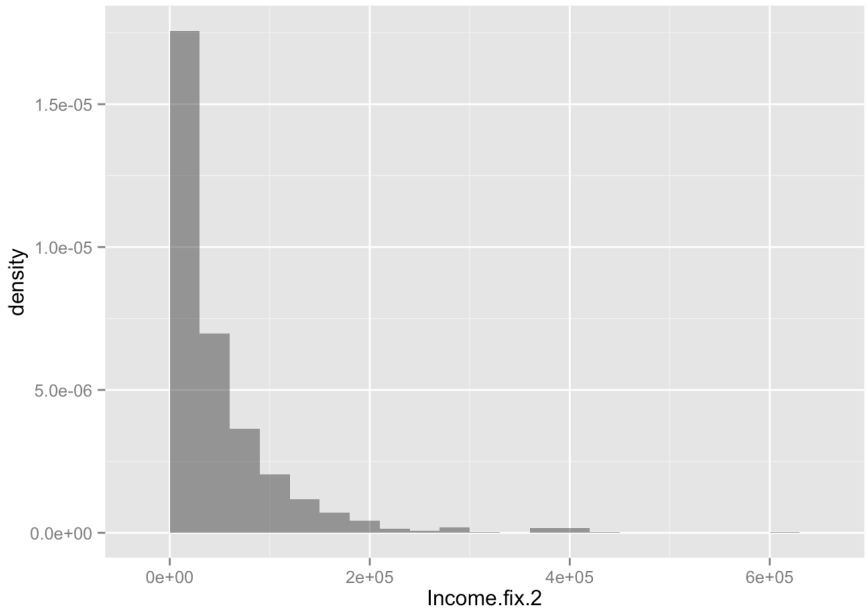
```
## Income.groups
##   (1e+04,5e+04] (1e+05,2.5e+05] (2.5e+05,1e+06]   (5e+04,1e+05]       [0,1
e+04]       no income
##             312              98              21             178
63             328
```

- zero income 구분

```
missing.Income <- is.na(custdata$Income)
Income.fix.2 <- ifelse(is.na(custdata$Income), 0, custdata$Income)
ggplot(data.frame(Income.fix), aes(x = Income.fix)) + geom_histogram(binwidth=3
0000, aes(y = ..density..), alpha=0.5)
```




```
ggplot(data.frame(Income.fix.2), aes(x = Income.fix.2)) + geom_histogram(binwid
th=30000, aes(y = ..density..), alpha=0.5)
```



Data Transformation

- Median Income

```
str(custdata)
```

```
## 'data.frame':    1000 obs. of  14 variables:
##  $ state.of.res    : Factor w/ 50 levels "Alabama","Alaska",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ custid          : int  1063014 1192089 16551 1079878 502705 674271 15917
467335 462569 1216026 ...
##  $ sex             : Factor w/ 2 levels "F","M": 1 2 1 1 2 2 1 2 2 2 ...
##  $ is.employed     : logi  TRUE NA NA NA TRUE FALSE ...
##  $ income          : int  82000 49000 7000 37200 70000 0 24000 42600 22000 9
600 ...
##  $ marital.stat    : Factor w/ 4 levels "Divorced/Separated",..: 2 2 2 1 2 2
1 3 4 3 ...
##  $ health.ins      : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
##  $ housing.type    : Factor w/ 4 levels "Homeowner free and clear",..: 4 1 2
2 4 4 1 4 1 4 ...
##  $ recent.move     : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ num.vehicles    : int  2 2 2 1 4 1 1 1 0 6 ...
##  $ age             : num  43 77 46 62 37 54 70 33 89 50 ...
##  $ is.employed.fix1: chr  "employed" "missing" "missing" "missing" ...
##  $ Income          : num  NA NA 4500 20000 12000 180000 120000 40000 NA 2400
0 ...
##  $ is.employed.fix : chr  "employed" "not in active workforce" "not in activ
e workforce" "not in active workforce" ...
```

```
str(medianincome)
```

```
## 'data.frame':    52 obs. of  2 variables:
##  $ State        : Factor w/ 52 levels "","Alabama","Alaska",..: 2 3 4 5 6 7
8 9 10 11 ...
##  $ Median.Income: num  52371 44191 65720 48484 39832 ...
```

```
summary(medianincome)
```

```
##         State    Median.Income
##            : 1   Min.   :37427
##  Alabama   : 1   1st Qu.:47483
##  Alaska    : 1   Median :52274
##  Arizona   : 1   Mean   :52655
##  Arkansas  : 1   3rd Qu.:57195
##  California: 1   Max.   :68187
##  (Other)   :46
```

```
custdata <- merge(custdata, medianincome, by.x = "state.of.res", by.y = "Stat
e")
str(custdata)
```

```
## 'data.frame':    1000 obs. of  15 variables:
##  $ state.of.res    : Factor w/ 50 levels "Alabama","Alaska",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ custid          : int  1063014 1192089 16551 1079878 502705 674271 15917
467335 462569 1216026 ...
##  $ sex             : Factor w/ 2 levels "F","M": 1 2 1 1 2 2 1 2 2 2 ...
##  $ is.employed     : logi  TRUE NA NA NA TRUE FALSE ...
##  $ income          : int  82000 49000 7000 37200 70000 0 24000 42600 22000 9
600 ...
##  $ marital.stat    : Factor w/ 4 levels "Divorced/Separated",..: 2 2 2 1 2 2
1 3 4 3 ...
##  $ health.ins      : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
##  $ housing.type    : Factor w/ 4 levels "Homeowner free and clear",..: 4 1 2
2 4 4 1 4 1 4 ...
##  $ recent.move     : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
##  $ num.vehicles    : int  2 2 2 1 4 1 1 1 0 6 ...
##  $ age             : num  43 77 46 62 37 54 70 33 89 50 ...
##  $ is.employed.fix1: chr  "employed" "missing" "missing" "missing" ...
##  $ Income          : num  NA NA 4500 20000 12000 180000 120000 40000 NA 2400
0 ...
##  $ is.employed.fix : chr  "employed" "not in active workforce" "not in activ
e workforce" "not in active workforce" ...
##  $ Median.Income   : num  52371 52371 52371 52371 52371 ...
```

```
summary(custdata[, c("state.of.res", "income", "Median.Income")])
```

```
##        state.of.res     income       Median.Income
##  California  :114   Min.   : -8700   Min.   :37427
##  New York    : 94   1st Qu.: 14600   1st Qu.:44819
##  Pennsylvania: 63   Median : 35000   Median :50118
##  Ohio        : 59   Mean   : 53505   Mean   :50919
##  Illinois    : 52   3rd Qu.: 67000   3rd Qu.:55534
##  Texas       : 51   Max.   :615000   Max.   :68187
##  (Other)     :567
```

```
custdata$income.norm <- with(custdata, income/Median.Income)
summary(custdata$income.norm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.1956  0.2812  0.6712  1.0780  1.3510 11.7900
```

- Converting continuous variables to discrete

```
custdata$income.lt.20K <- custdata$income < 20000
summary(custdata$income.lt.20K)
```

```
##    Mode   FALSE    TRUE    NA's
## logical     678     322       0
```

- Converting age into ranges

```
age.breaks <- c(0, 25, 65, Inf)
custdata$age.range <- cut(custdata$age, breaks = age.breaks, include.lowest = T
RUE)
summary(custdata$age.range)
```

```
##   [0,25]  (25,65] (65,Inf]
##       56      732      212
```

```
str(custdata$age.range)
```

```
##  Factor w/ 3 levels "[0,25]","(25,65]",..: 2 3 2 2 2 2 3 2 3 2 ...
```

- Centering on mean age

```
summary(custdata$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    38.0    50.0    51.7    64.0   146.7
```

```
mean.age <- mean(custdata$age)
custdata$age.normalized <- custdata$age/mean.age
summary(custdata$age.normalized)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.7350  0.9671  1.0000  1.2380  2.8370
```

- Summarizing age

```
summary(custdata$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    38.0    50.0    51.7    64.0   146.7
```

```
sd.age <- sd(custdata$age)
custdata$age.normalized <- (custdata$age - mean.age)/sd.age
summary(custdata$age.normalized)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.74100 -0.72630 -0.09011  0.00000  0.65210  5.03500
```

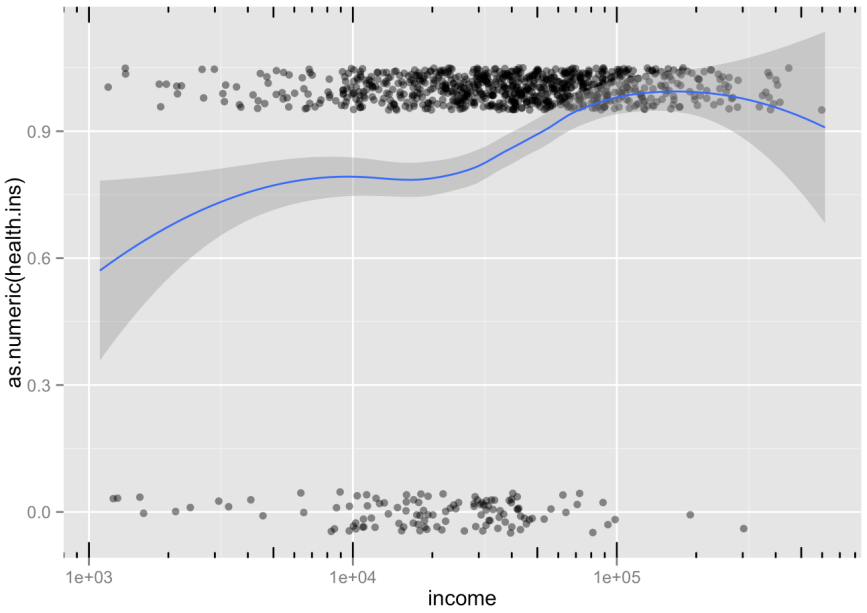
```
summary(scale(custdata$age))
```

```
##        V1
##  Min.   :-2.74074
##  1st Qu.:-0.72626
##  Median :-0.09011
##  Mean   : 0.00000
##  3rd Qu.: 0.65207
##  Max.   : 5.03516
```

- Figure 4.2
  - `y = as.numeric(health.ins)` 와 `y = health.ins` 라고 했을 때의 차이 유의.

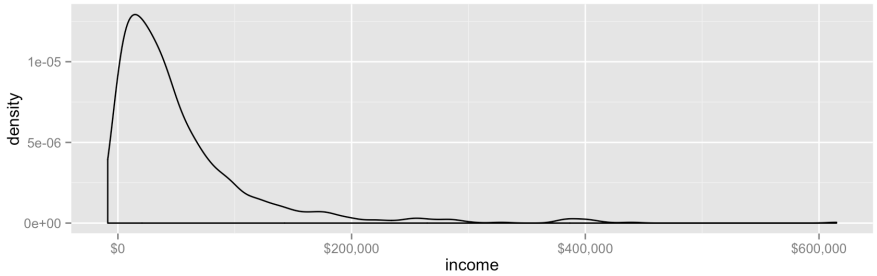
```
ggplot(subset(custdata, custdata$income > 1000), aes(x = income, y = as.numeri
c(health.ins))) +
  geom_point(alpha = 0.5, position = position_jitter(w = 0.05, h = 0.05)) +
  geom_smooth() + scale_x_log10() + annotation_logticks(sides = "bt")
```

```
## geom_smooth: method="auto" and size of largest group is <1000, so using loes
s. Use 'method = x' to change the smoothing method.
```

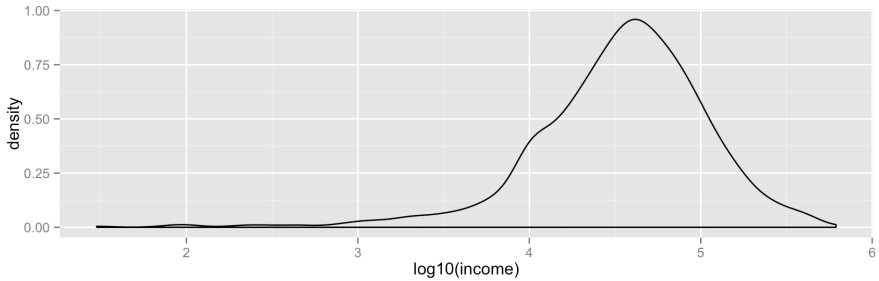


- Figure 4.4
  - `subset()` 설정을 하지 않으면 어떻게 될까?

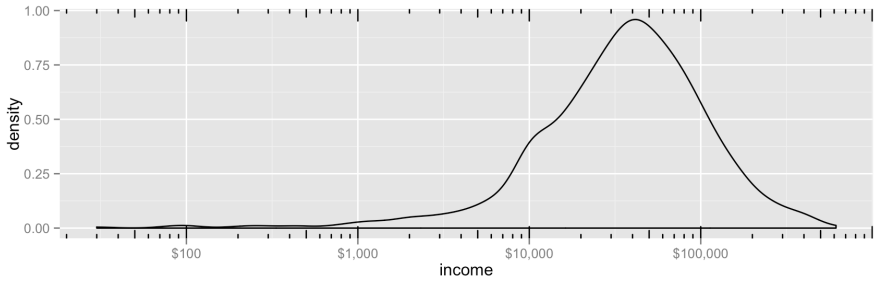
```
library(scales)
ggplot(custdata, aes(x = income)) + geom_density() + scale_x_continuous(labels
= dollar)
```




```
ggplot(subset(custdata, custdata$income > 0), aes(x = log10(income))) + geom_de
nsity()
```




```
ggplot(subset(custdata, custdata$income > 0), aes(x = income)) + geom_density()
+
  scale_x_log10(breaks = c(100, 1000, 10000, 100000), labels = dollar) +
  annotation_logticks(sides = "bt")
```



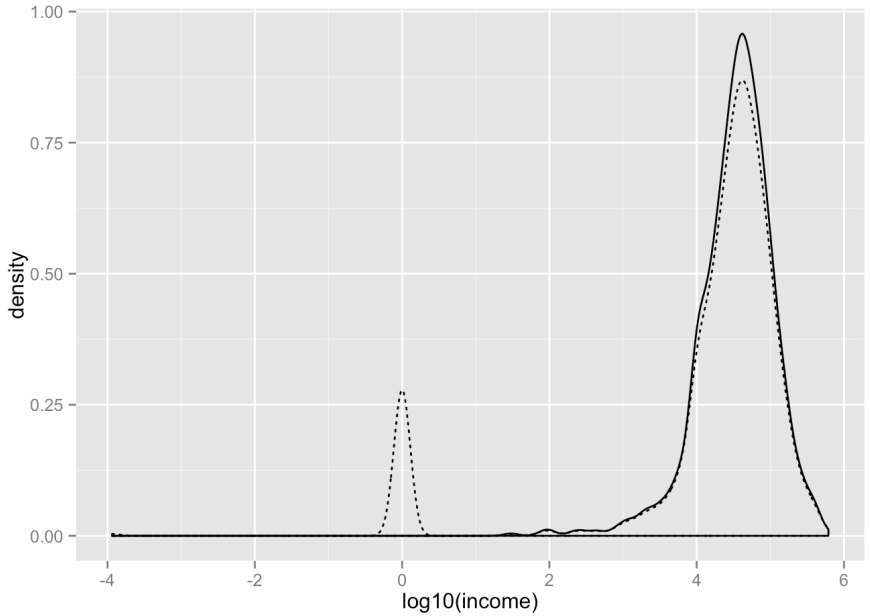
- Signed log10

```
signed.log10 <- function(x) {
  ifelse(abs(x) <= 1, 0, sign(x) * log10(abs(x)))
}
ggplot(custdata, aes(x = log10(income))) + geom_density() +
  geom_density(aes(x = signed.log10(income)), linetype="dotted")
```

```
## Warning: NaN이 생성되었습니다
```

```
## Warning: NaN이 생성되었습니다
```

```
## Warning: Removed 79 rows containing non-finite values (stat_density).
```




```
dump("signed.log10", file="signed.log10.R")
```

Sampling for modelling and validation

Creating a sample group column

- Splitting into test and training using a random group mark

```
set.seed(123456)
custdata$gp <- runif(nrow(custdata))
custdata.test <- subset(custdata, custdata$gp <= 0.1)
custdata.train <- subset(custdata, custdata$gp > 0.1)
nrow(custdata.test)
```

```
## [1] 108
```

```
nrow(custdata.train)
```

```
## [1] 892
```

```
custdata$gp.2 <- factor(ifelse(1:nrow(custdata) %in% sample(nrow(custdata), siz
e=100), "test", "train"))
summary(custdata$gp.2)
```

```
##  test train
##   100   900
```

```
table(custdata$gp.2)
```

```
##
##  test train
##   100   900
```

- Record grouping

```
set.seed(123456)
str(hhdata)
```

```
## 'data.frame':    12 obs. of  4 variables:
##  $ household_id: Factor w/ 5 levels "hh1","hh2","hh3",..: 1 1 2 3 3 3 4 4 4
5 ...
##  $ cust_id     : Factor w/ 3 levels "cust1","cust2",..: 1 2 1 1 2 3 1 2 3 1
...
##  $ income      : Factor w/ 12 levels "0","100000","110000",..: 12 1 7 2 10 4
8 11 5 9 ...
##  $ gp          : num  0.626 0.626 0.88 0.711 0.711 ...
```

```
(hhdata.2 <- hhdata[1:3])
```

```
##    household_id cust_id income
## 1           hh1   cust1  95000
## 2           hh1   cust2      0
## 3           hh2   cust1  60000
## 4           hh3   cust1 100000
## 5           hh3   cust2   8000
## 6           hh3   cust3  35020
## 7           hh4   cust1  65000
## 8           hh4   cust2  86000
## 9           hh4   cust3  36000
## 10          hh5   cust1  68000
## 11          hh5   cust2 110000
## 12          hh5   cust3  47950
```

```
(hh <- unique(hhdata$household_id))
```

```
## [1] hh1 hh2 hh3 hh4 hh5
## Levels: hh1 hh2 hh3 hh4 hh5
```

```
(households <- data.frame(household_id = hh, gp = runif(length(hh))))
```

```
##   household_id        gp
## 1          hh1 0.7977843
## 2          hh2 0.7535651
## 3          hh3 0.3912557
## 4          hh4 0.3415567
## 5          hh5 0.3612941
```

```
(hhdata.3 <- merge(hhdata.2, households, by = "household_id"))
```

```
##    household_id cust_id income        gp
## 1           hh1   cust1  95000 0.7977843
## 2           hh1   cust2      0 0.7977843
## 3           hh2   cust1  60000 0.7535651
## 4           hh3   cust1 100000 0.3912557
## 5           hh3   cust2   8000 0.3912557
## 6           hh3   cust3  35020 0.3912557
## 7           hh4   cust1  65000 0.3415567
## 8           hh4   cust2  86000 0.3415567
## 9           hh4   cust3  36000 0.3415567
## 10          hh5   cust1  68000 0.3612941
## 11          hh5   cust2 110000 0.3612941
## 12          hh5   cust3  47950 0.3612941
```