BC2406 Analytics I
Visual and Predictive Techniques

Unit 3

Data Exploration & Summaries

NANYANG TECHNOLOGICAL UNIVERSITY
Nanyang Business School

Based on Chew C. H. (2019) textbook: Analytics, Data Science and AI. Vol 1., Chap 2.

Seminar Objectives

- Learn some techniques for Data Exploration.
  - Basic Statistics
  - Basic Charts
  - No Data Cleaning yet (see unit 5 or textbook chap 5)
- How to use R to do Data Exploration better and much faster than spreadsheet.
- Introduce a good Rpackage for Data Exploration & Summaries: data.table

Purpose of Data Exploration

- Gain some understanding of the Dataset(s)
- Compare Data to Business Problem/Opportunity:
  - Sufficient?
  - Necessary?
    - No predictive value
    - Identification value
    - Redundant
- Detect Data problems/issues
  - Data Quality
  - Anomalies (something that deviates from what is standard, normal, or expected).

Data Exploration Techniques

- Statistics
- Visualization
- Models

Example: Health Insurance Coverage

- Business Problem:
  - People may not have sufficient Health insurance.
- Analytics Problem:
  - Develop a model to predict whether someone has health insurance or not, based on demographic information.
- Potential Application:
  - Identify correct target market much more easily, faster and accurately so that they can be educated and have opportunity to be covered by health insurance.
- Data:
  - Sample from Census of customer information and status of health insurance coverage – Y/N, in USA.
- Run: ADA1.3.1 health_ins_cust.R

summary(cust.df)

```
     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
```

R treats sex as categorical (aka Factors)

About 33% of employment status data is missing!

Negative income? Income has quite a big range. Units? US$?, monthly/yearly?

summary(cust.df)

```
             marital.stat health.ins
Divorced/Separated:155   Mode :logical
Married            :516   FALSE:159
Never Married      :233   TRUE :841
Widowed            : 96   NA's :0
```

Shows the categories used in marital status column

This is the Y variable that you want model to predict. About 84% has health insurance in this sample.

summary(cust.df)

```
     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
```

Min age = 0, or just a code for something?
Max age = 146.7?

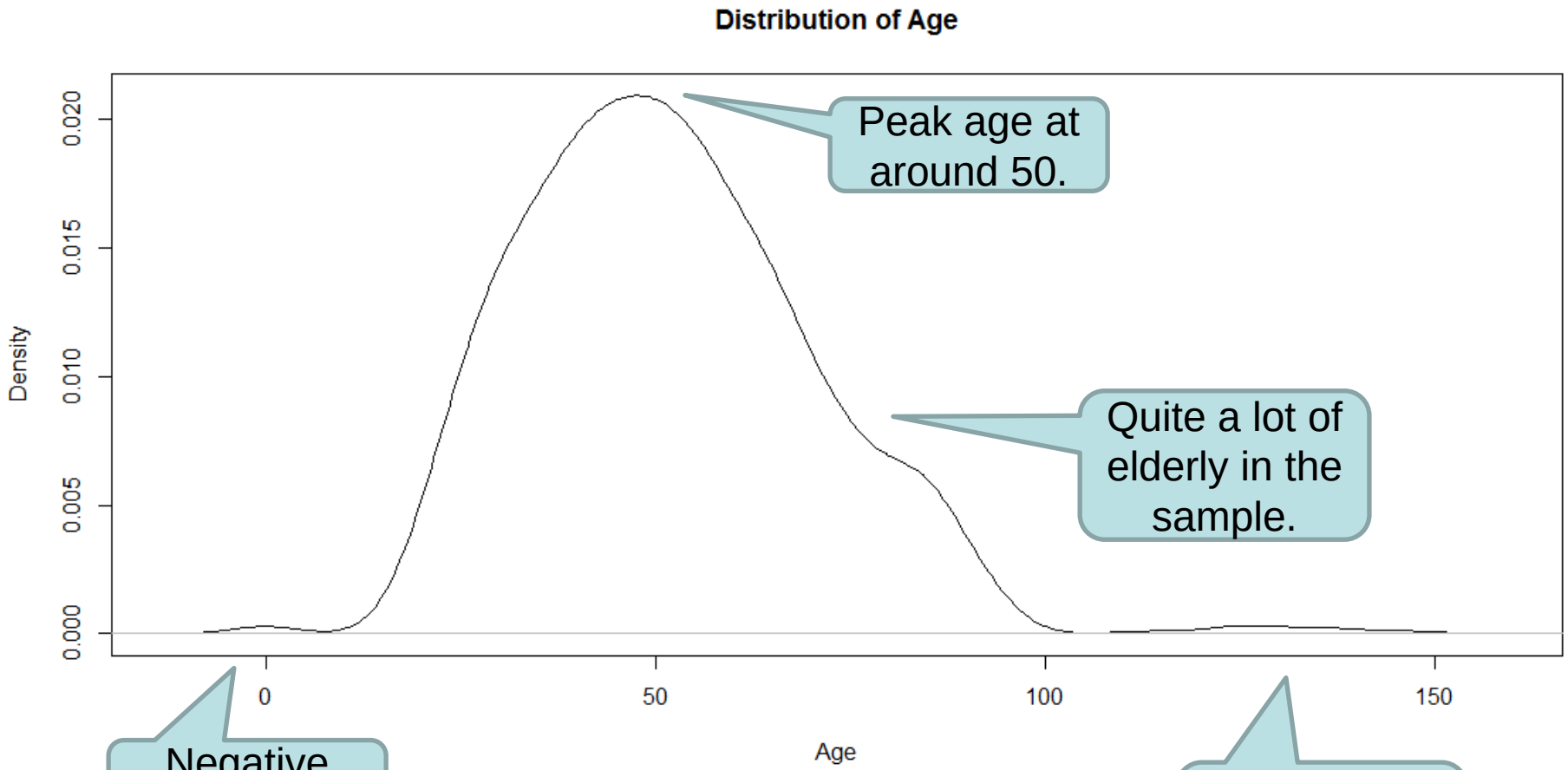
There are many other states of residence. For summary purpose, R used “(other)” category to group the less frequent states

summary(cust.df)

```
                        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
```

56 missing values in 3 columns. Are these the same 56 customers?

plot(density(cust.df$age)) for continuous variable age



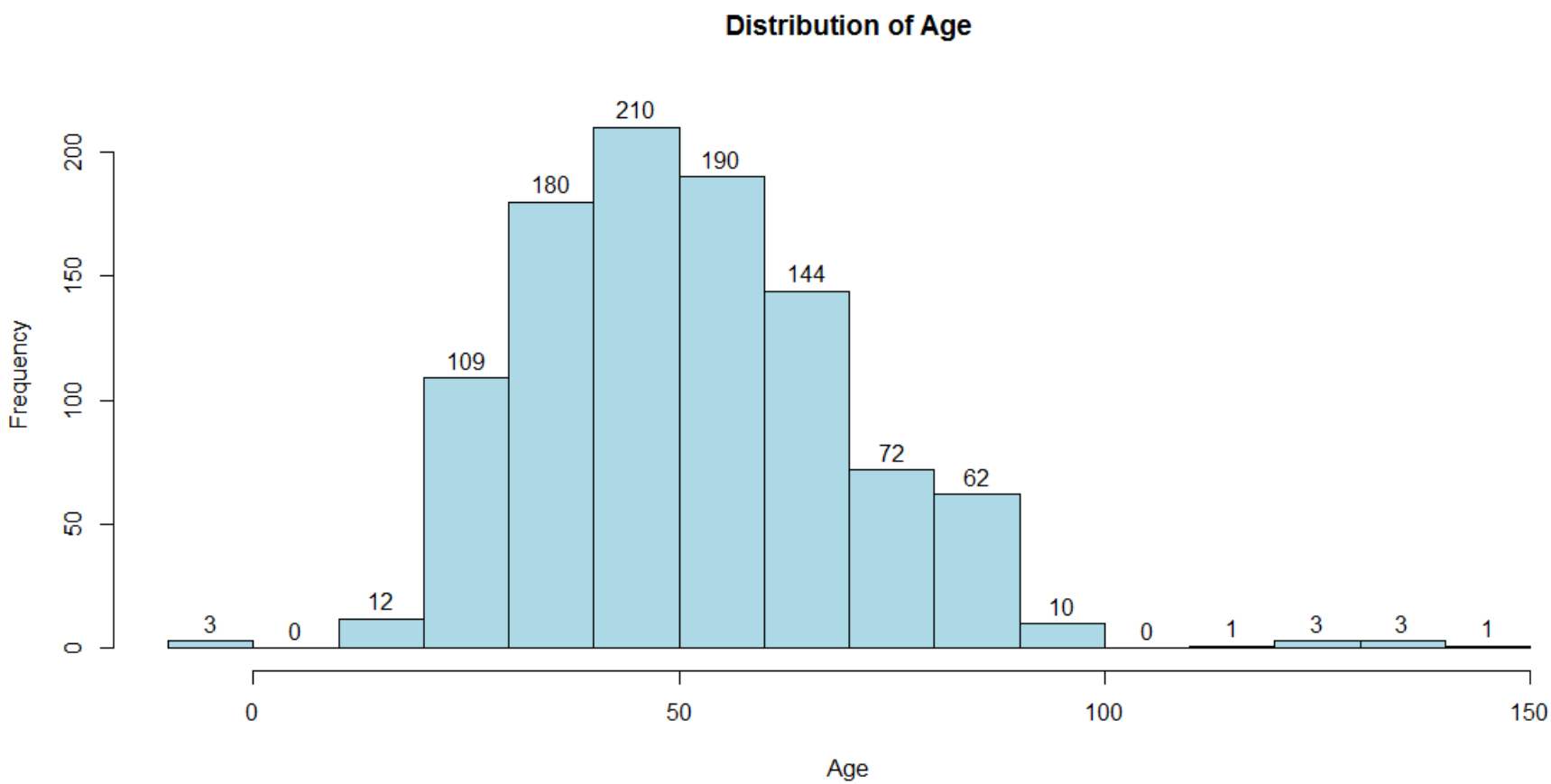
Peak age at around 50.

Quite a lot of elderly in the sample.

Negative age? 0 age?

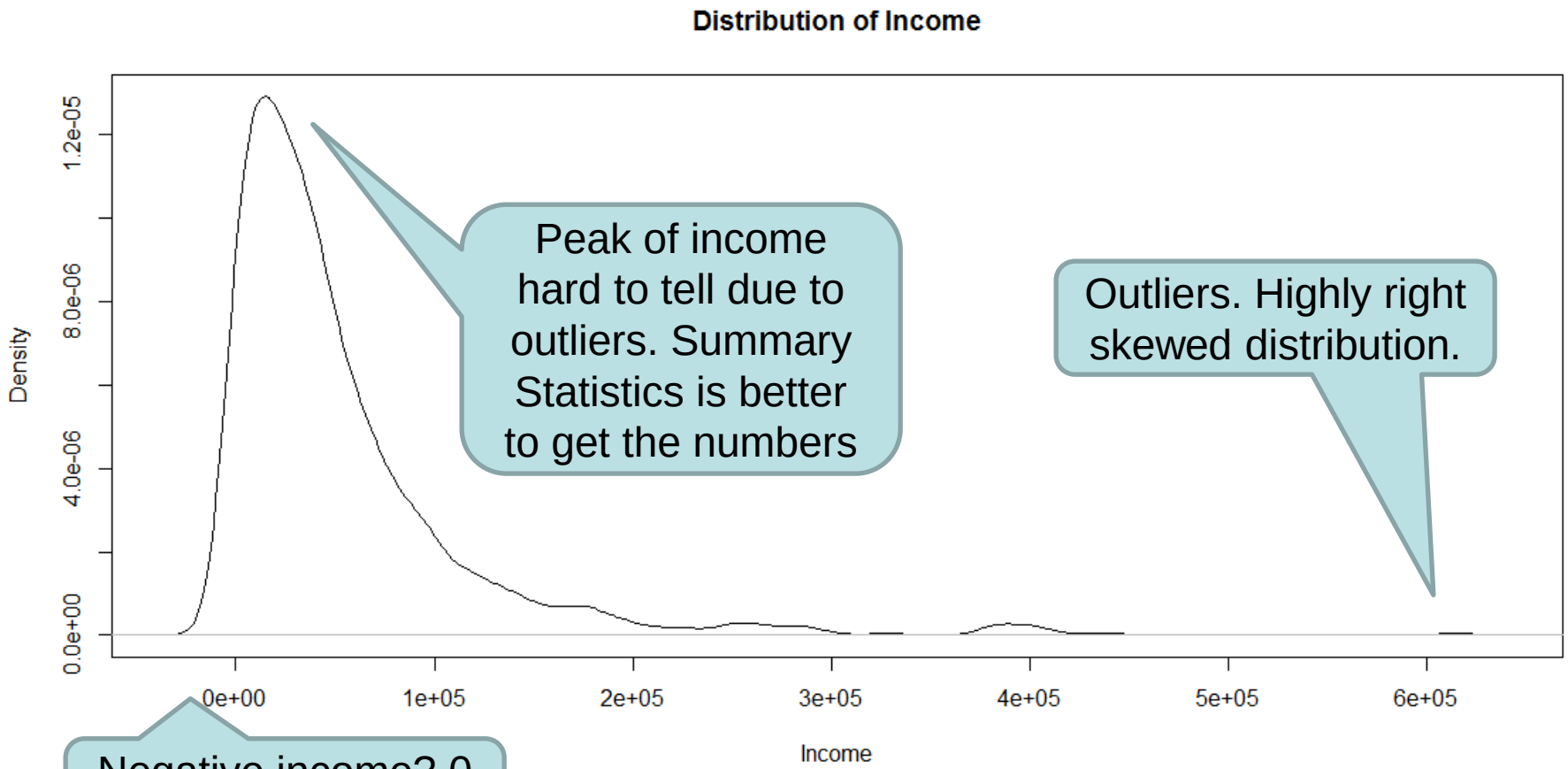
Way above 100 yrs old. Really?

hist(cust.df$age, ylim=c(0,220), breaks = c(-10, 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150), labels = T, col ="light blue")



Intervals can be controlled by the breaks argument.
Note: Left open, right closed interval by default.

plot(density(cust.df$income)) for continuous variable income

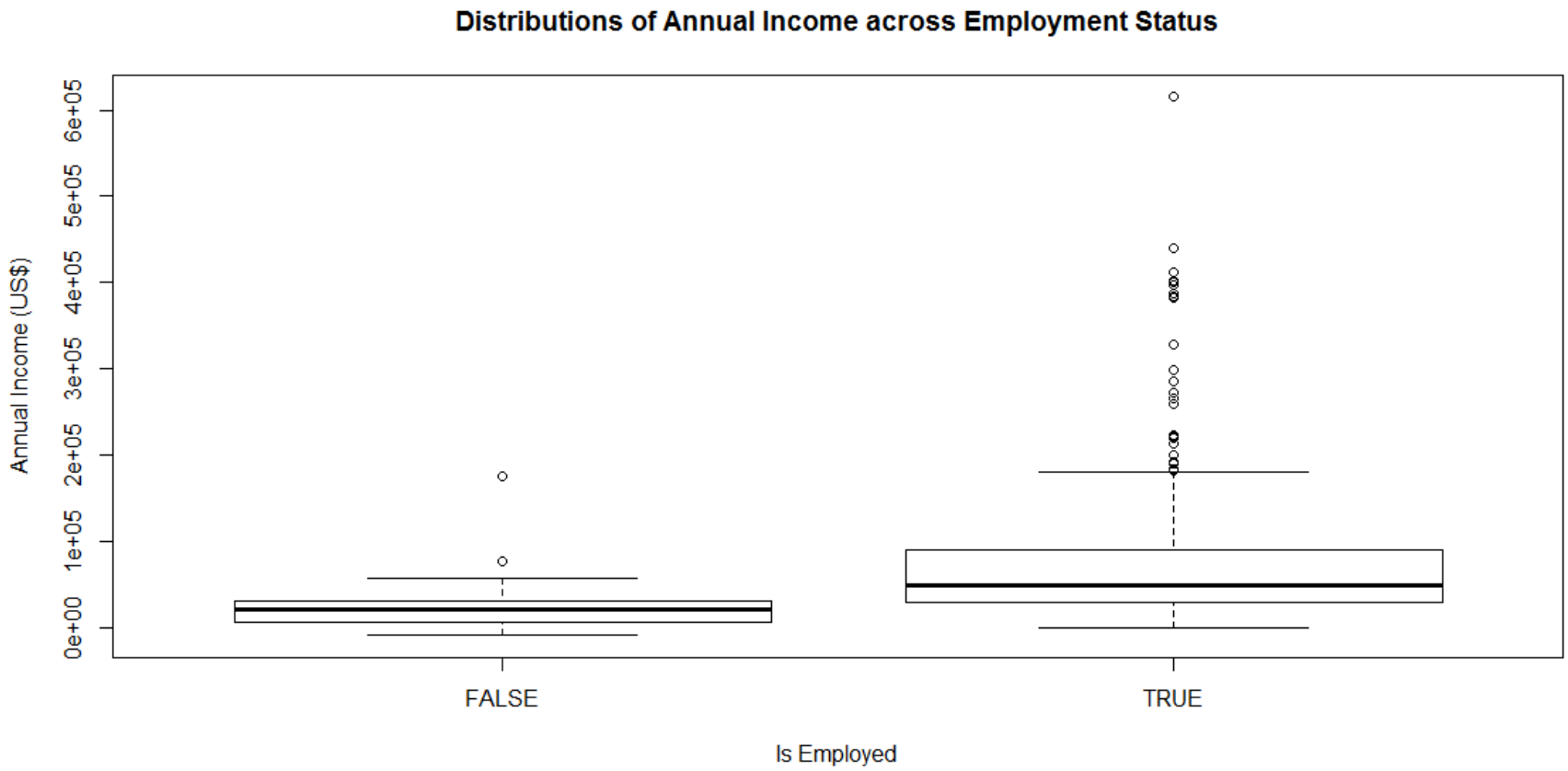


Peak of income hard to tell due to outliers. Summary Statistics is better to get the numbers

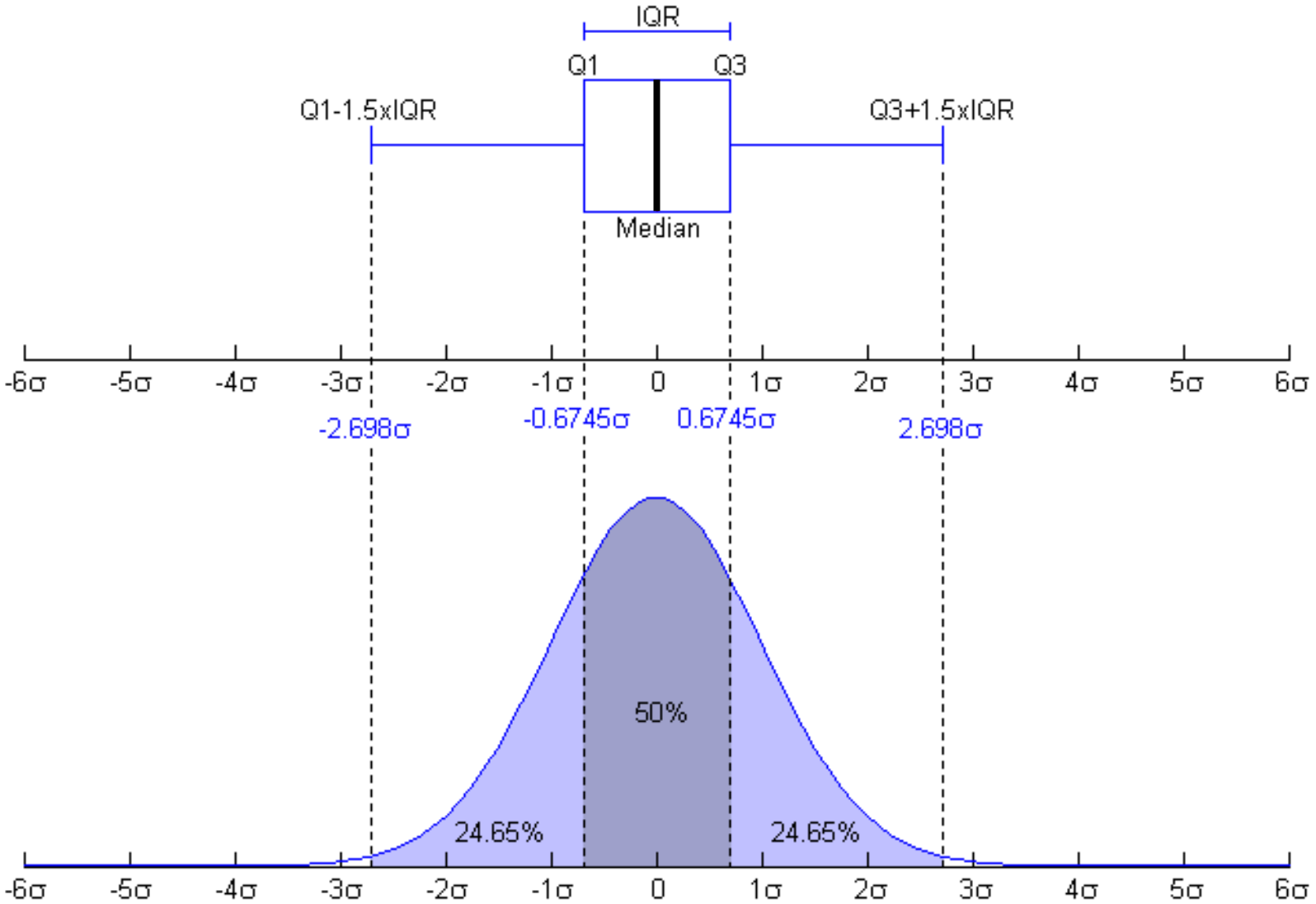
Outliers. Highly right skewed distribution.

Negative income? 0 income or coded?

Boxplot of a continuous variable (Income) across Employment Status



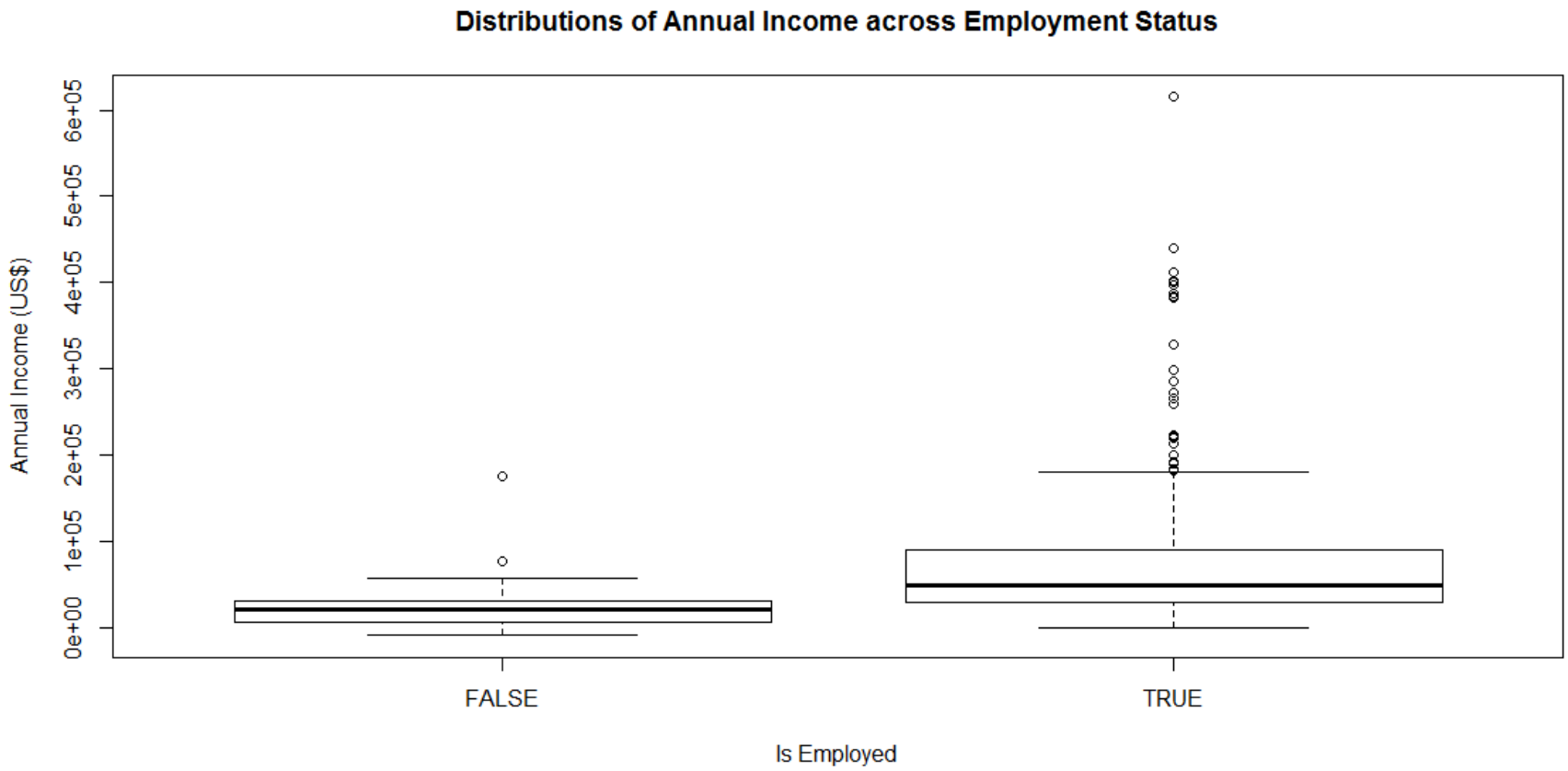
Who is contributing most to the Income outliers?
Whose income is more variable?


IQR
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Median
-6σ
-5σ
-4σ
-3σ
-2σ
-1σ
0
1σ
2σ
3σ
4σ
5σ
6σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
50%
24.65%
24.65%

Outliers

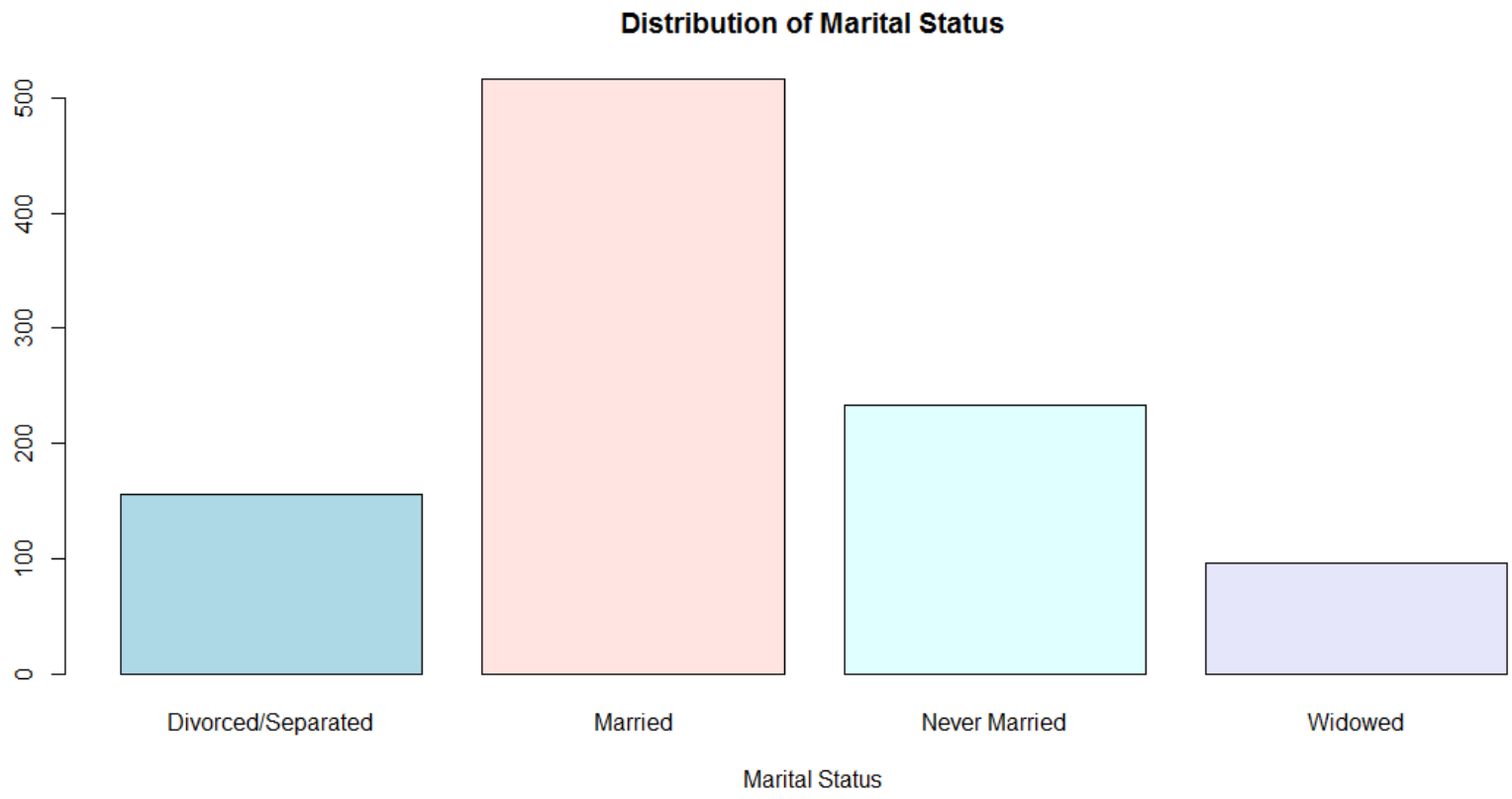
- There is a vast body of literature on outlier detection, and several definitions of outlier exist. For example, Tukey's box-and-whisker method for outlier detection is often appropriate.
- In this method, an observation is an outlier when it is larger than the so-called "whiskers"' of the set of observations. The upper whisker is computed by adding 1.5 times the interquartile range to the third quartile and rounding to the nearest lower observation.

Boxplot of a continuous variable (Income) across Employment Status



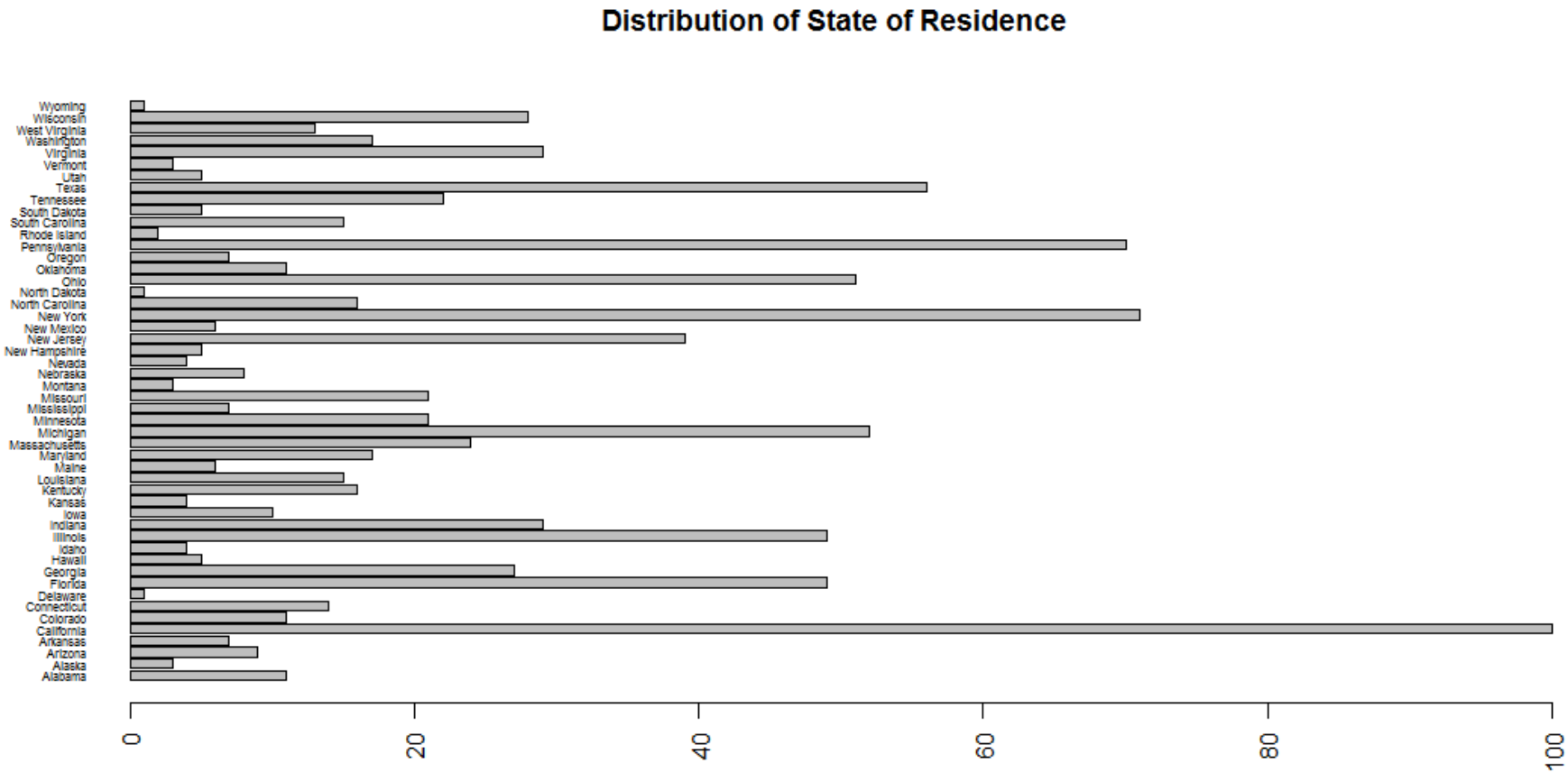
Who is contributing most to the Income outliers?
Whose income is more variable?

barplot(table(cust.df$marital.stat)) to see distribution of categorical variables



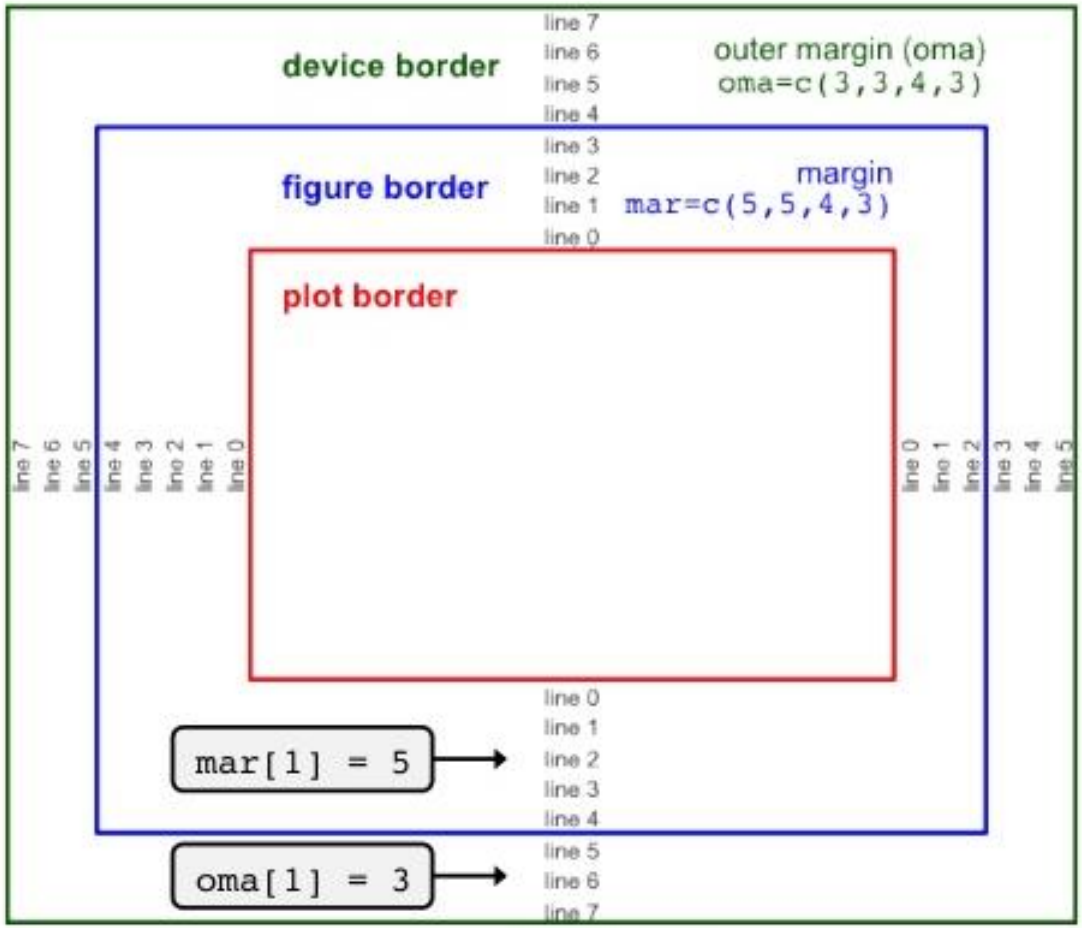
```
par(las=2)      # Default is las = 0
par(mar=c(5,8,4,2))      # Default is mar = c(5,4,4,2)

barplot(table(cust.df$state.of.res),  horiz = T, cex.names=0.5)
```



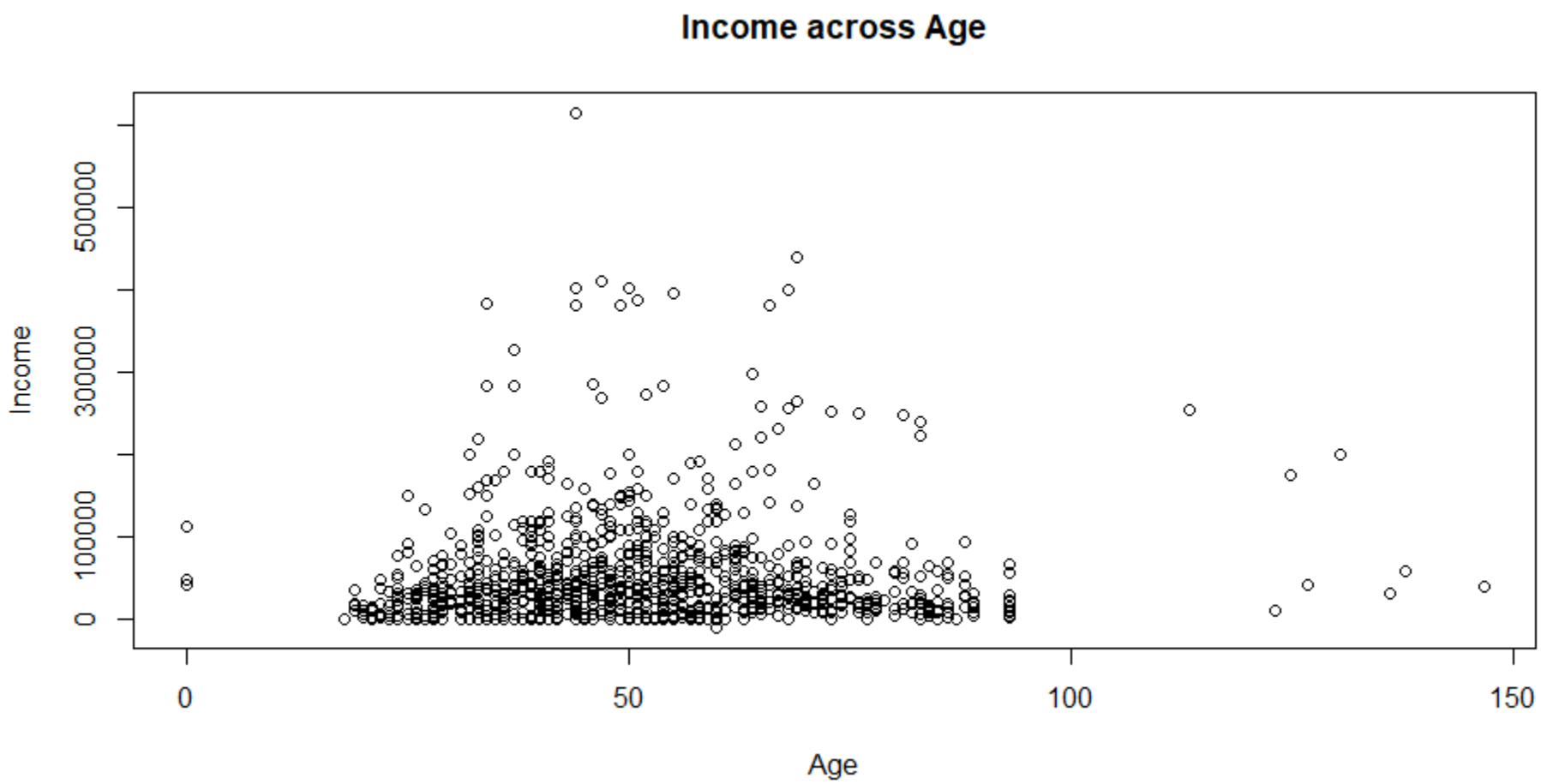
```
par(las=2)      # Default is las = 0
par(mar=c(5,8,4,2))      # Default is mar = c(5,4,4,2)

barplot(table(cust.df$state.of.res),  horiz = T, cex.names=0.5)
```

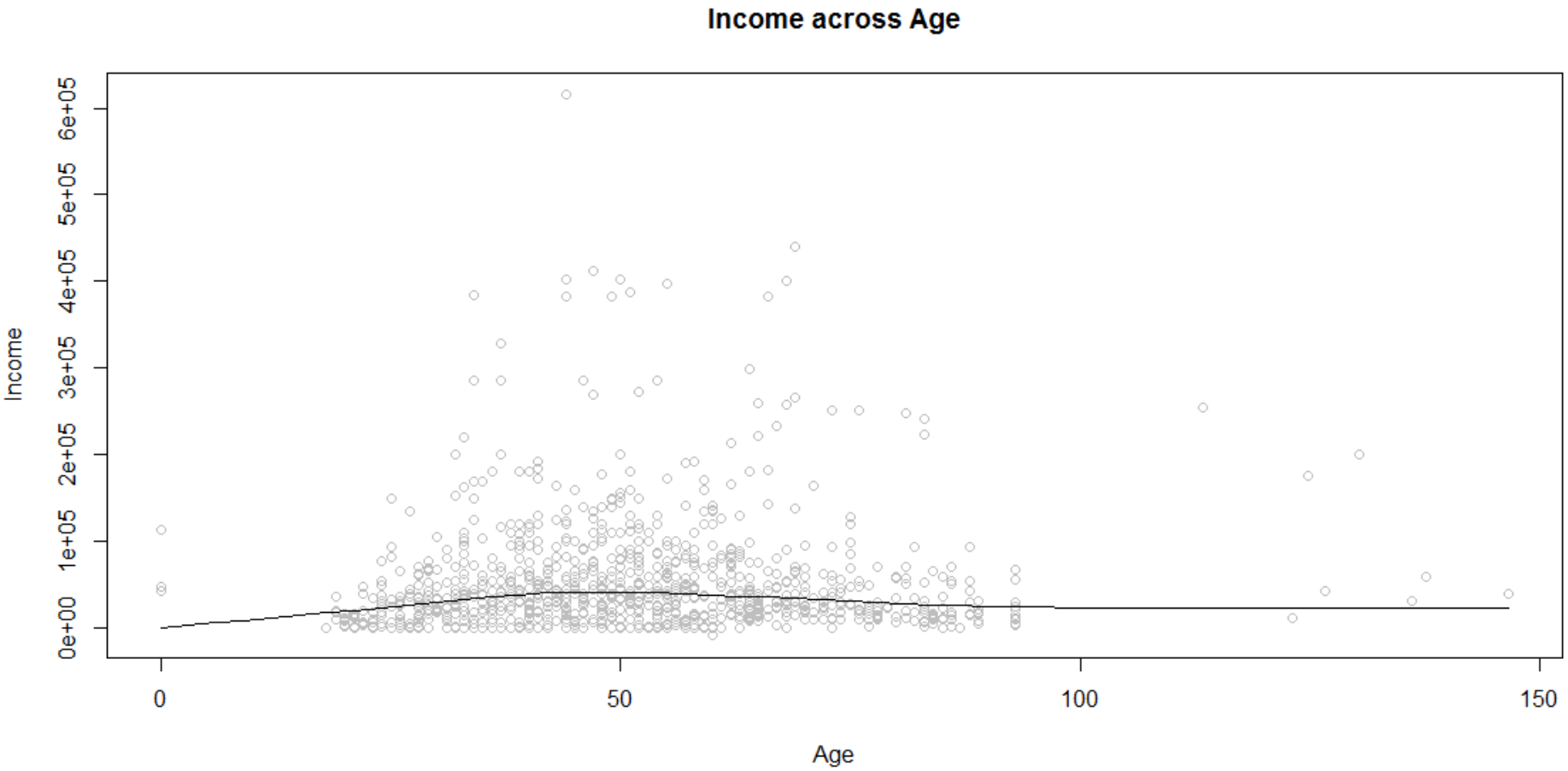


Siehe auch: Clarkson, 2010, chr

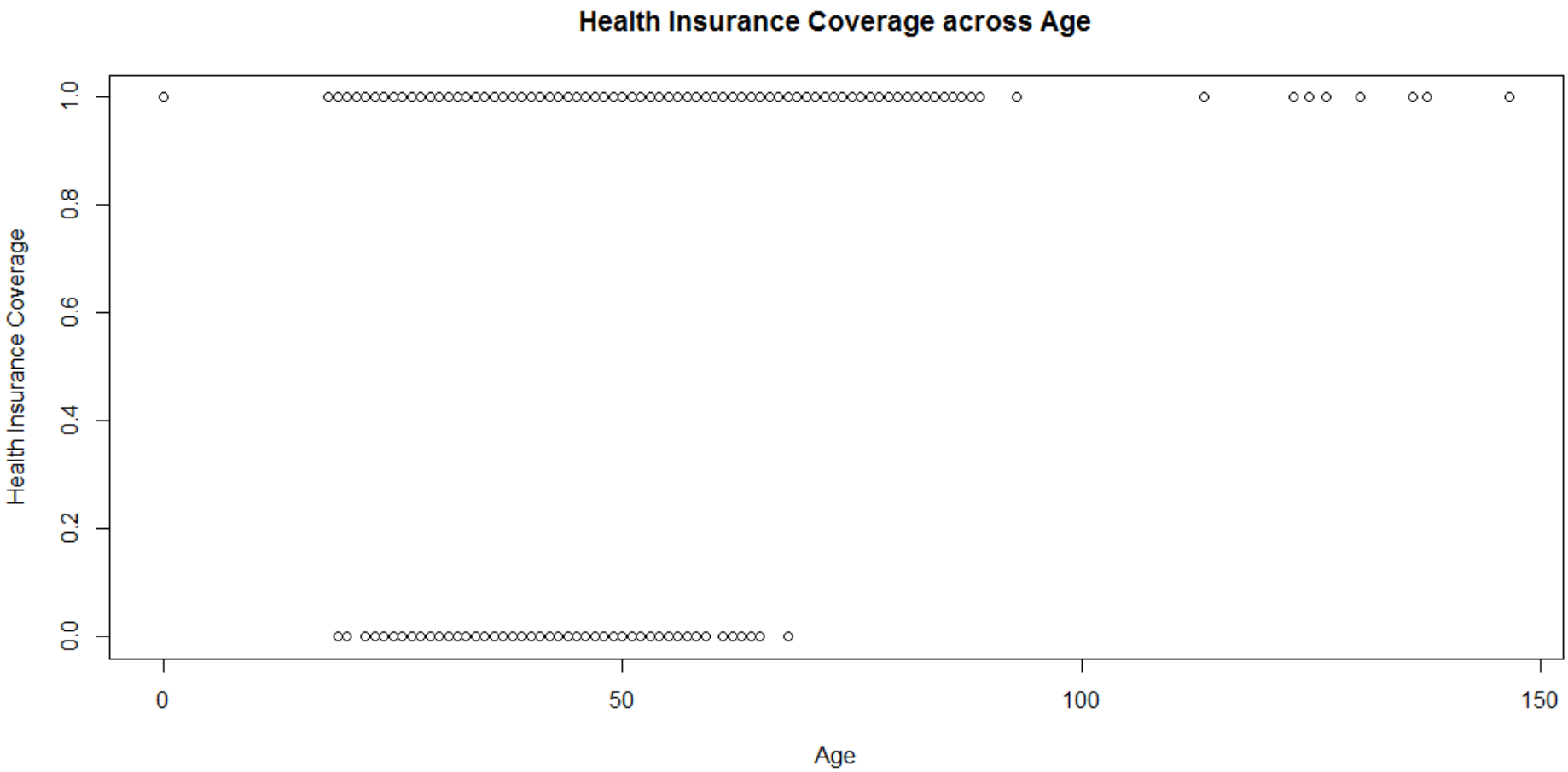
Scatterplot of two continuous variables (Age and Income)



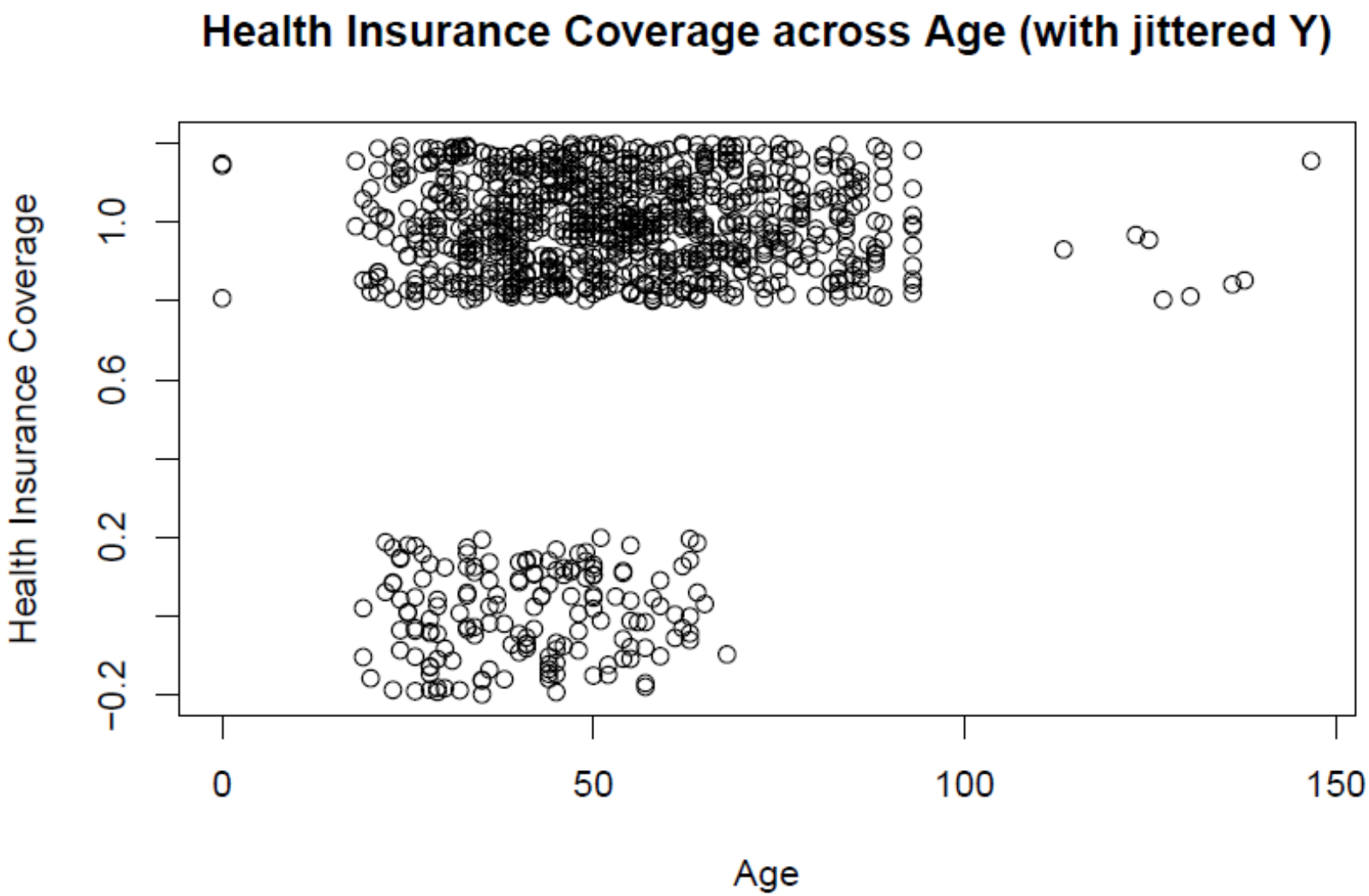
Scatterplot of Income across Age, with smooth curve



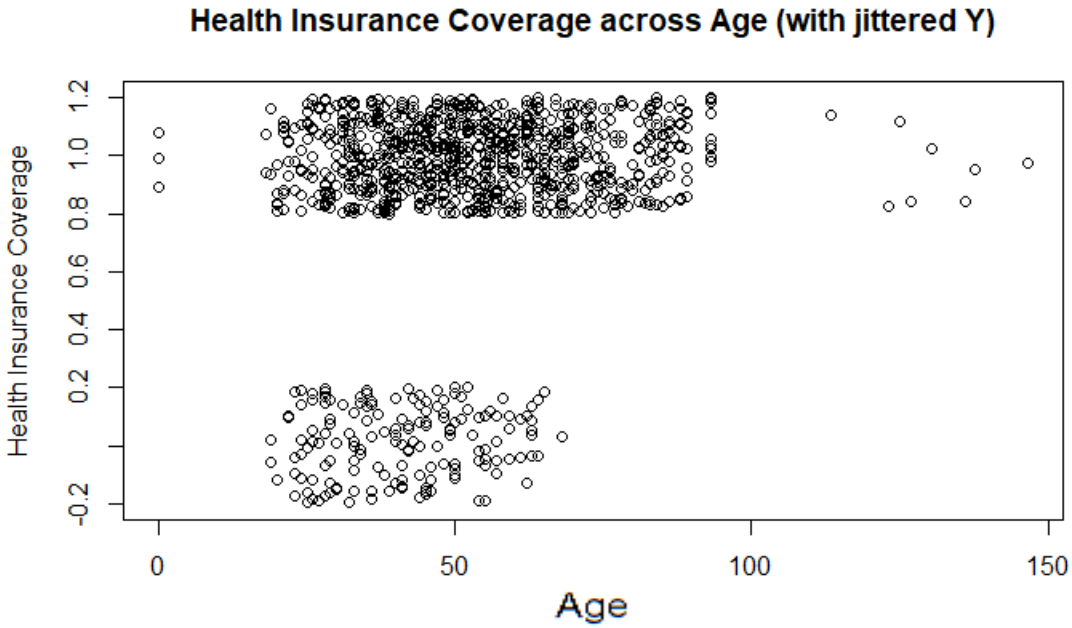
Scatterplot of Health Insurance Coverage and Age

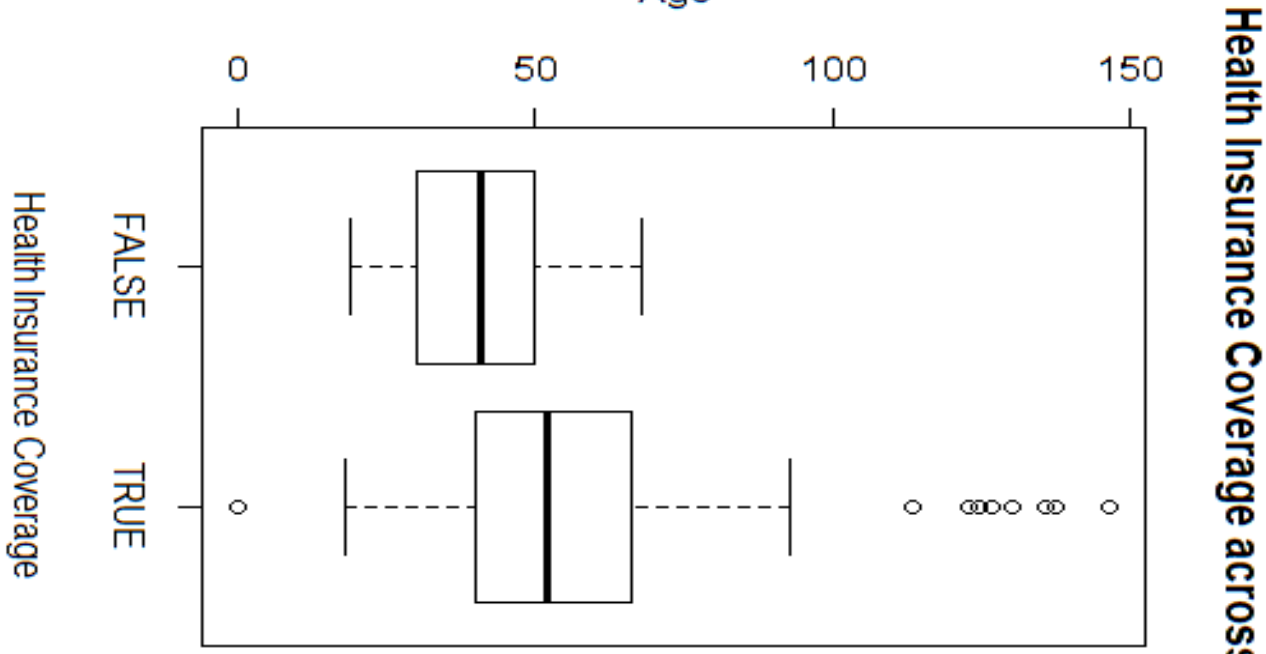


What can you see?

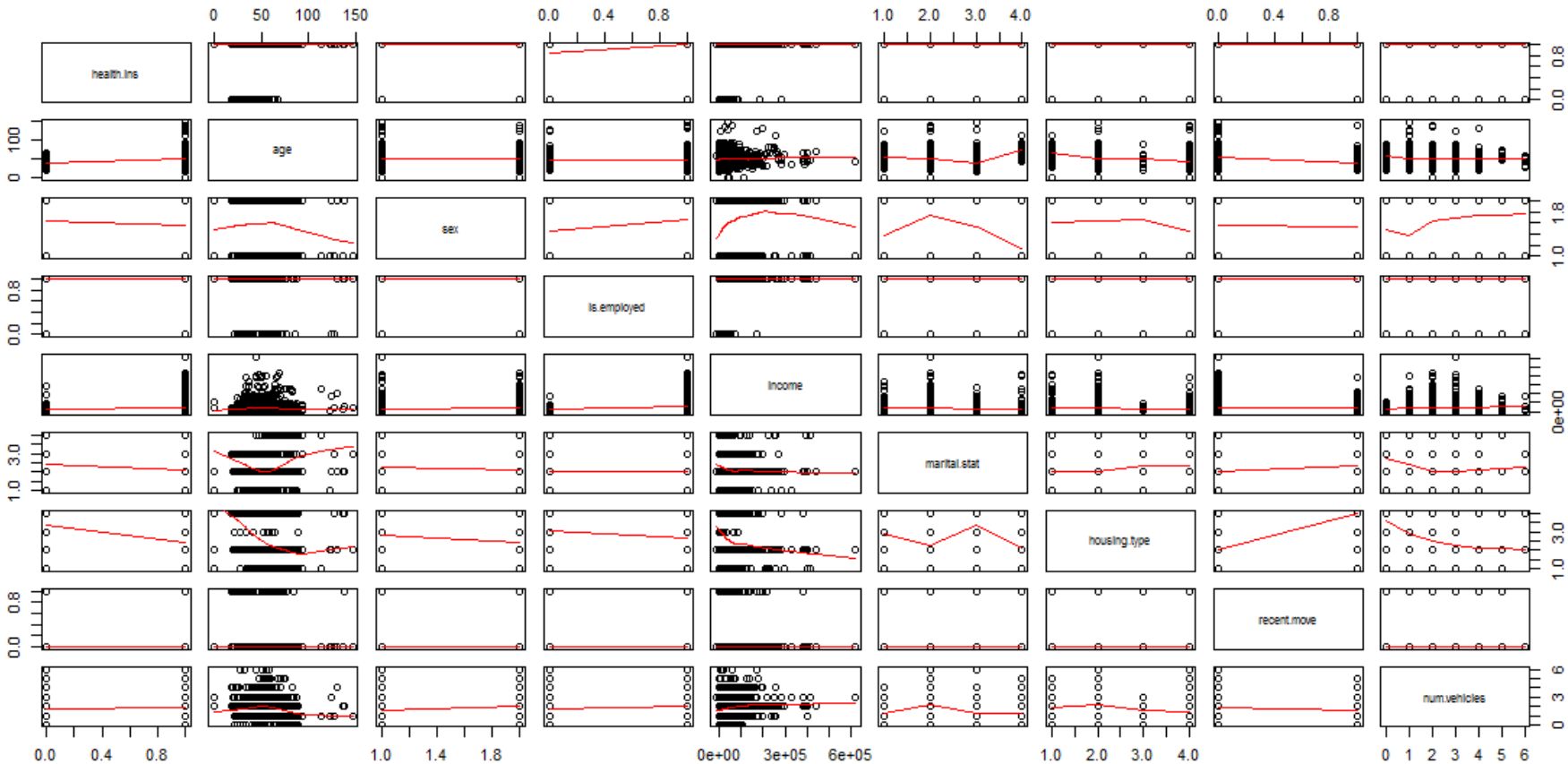


What can you see now?


Health Insurance Coverage across Age (with jittered Y)
Health Insurance Coverage
-0.2
0.2
0.4
0.6
0.8
1.0
1.2
0
50
100
150
Age

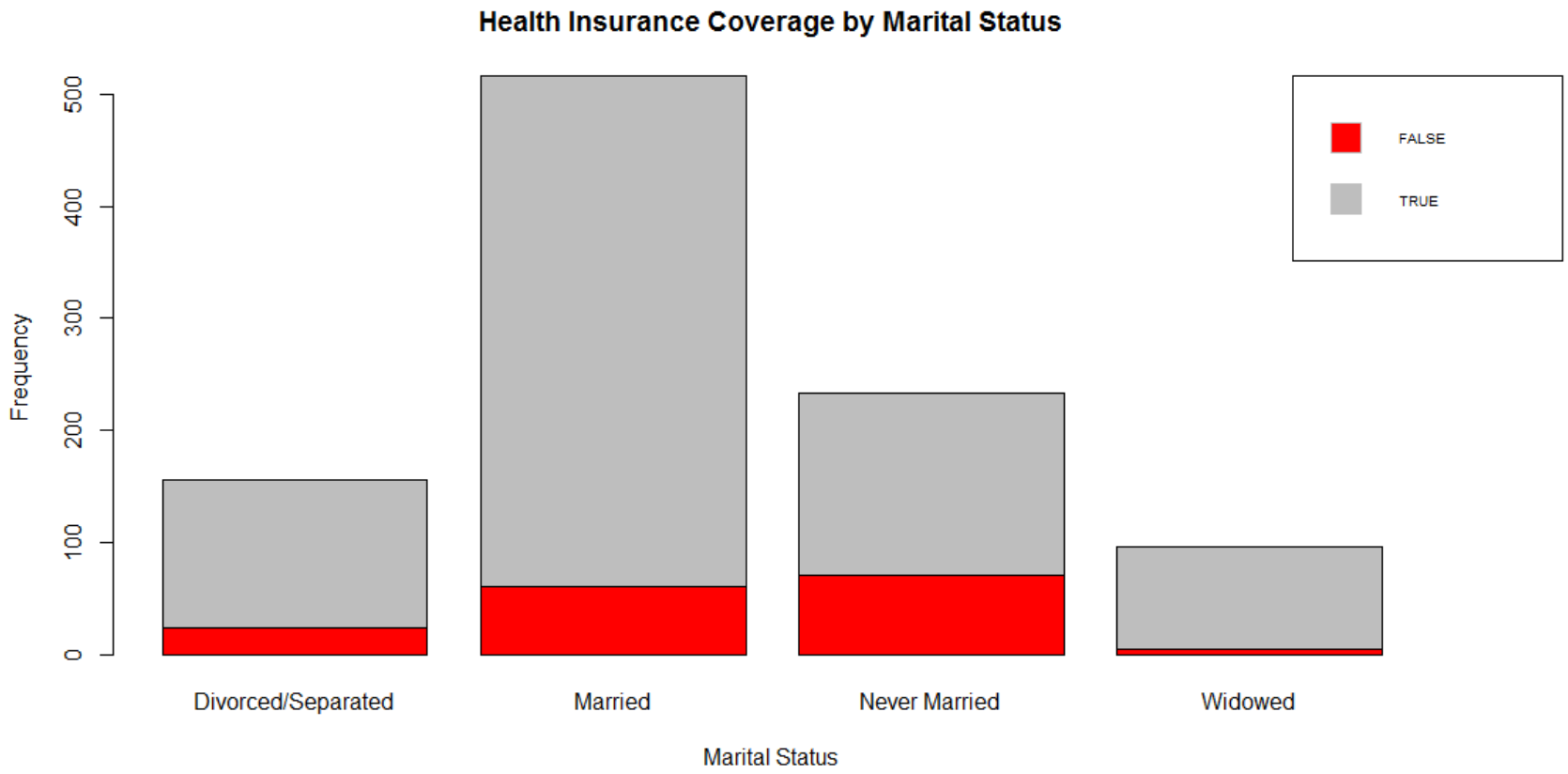

Health Insurance Coverage across Ag
0
50
100
150
FALSE
TRUE
Health Insurance Coverage

Scatterplot Matrix of Selected Variables with smooth curves



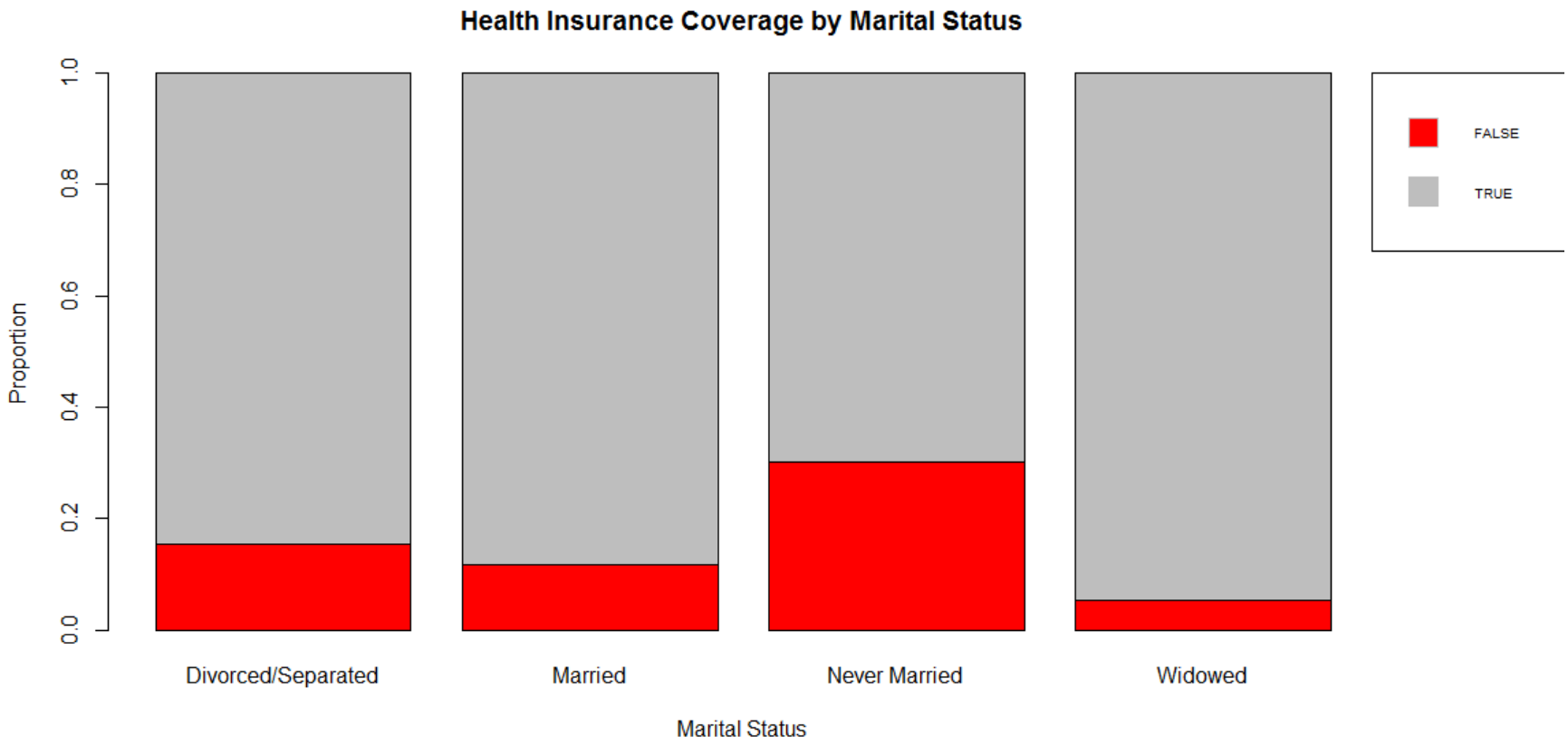
What's the usefulness of such a chart?

Stacked Bar Chart (Frequency)



But “widowed” is the smallest sub-population.

Stacked Bar Chart (Proportion)



"Never Married" has the highest proportion of no health insurance, while "Widowed" has the lowest proportion of no health insurance.

Others

- There are many other statistics and visualizations available that may help Data Exploration.
- Guideline:
  1. What is the business problem/Opportunity?
  2. Ask specific questions about the Data. i.e. What do I need to know about the Data that could help answer/address the business problem/opportunity?
  3. How do I answer those data questions using R (or any other software)?
- It's not rushing in to try all possible statistics or charts from the Data in hope of finding something useful – waste of effort.
- It begins with the business problem/Opportunity. Understand this first. Data exploration should be purpose-driven.

data.table package

FAST DATA EXPLORATION

Package data.table

- Fast
  - Read-in Data
  - Rscript Development
- Simple, consistent Syntax

$$DT[i, j, by]$$

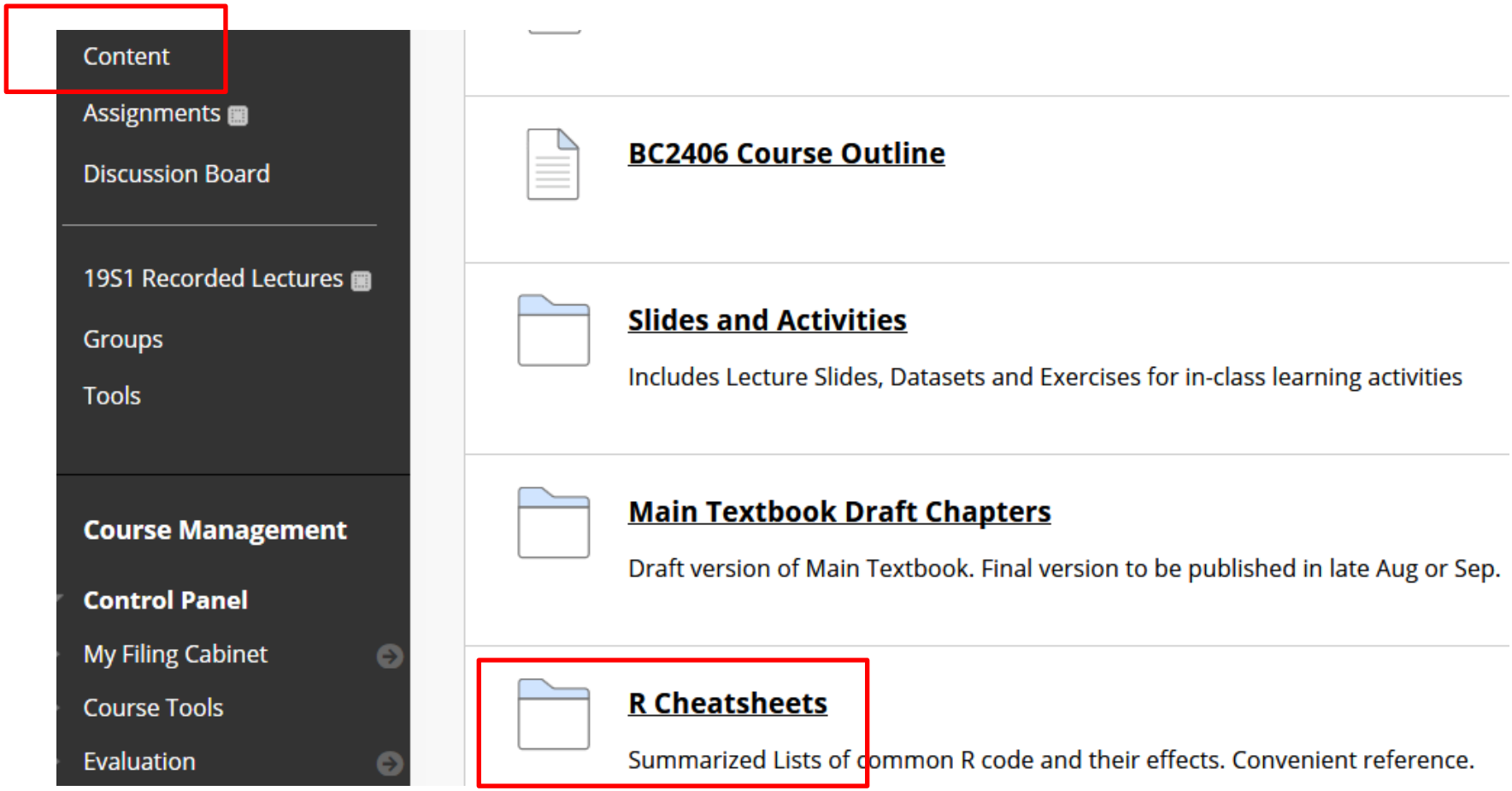
- DT: Name of the Data Table
- i: Criteria for selecting rows
- j: Actions on the selected rows in terms of column variable(s)
- by: Grouping variable(s)

Cheatsheet: datatable

- Refer to datatable Cheatsheet posted in NTULearn main site for summarized list of common procedures and their effects.

data.table documentation

- Reference document with detailed explanation from package creator
- See Main Site > Content > Slides and Activities > Unit 3 sub-folder > data.table documentation.pdf


Content
Assignments
Discussion Board
19S1 Recorded Lectures
Groups
Tools
Course Management
Control Panel
My Filing Cabinet
Course Tools
Evaluation
BC2406 Course Outline
Slides and Activities
Includes Lecture Slides, Datasets and Exercises for in-class learning activities
Main Textbook Draft Chapters
Draft version of Main Textbook. Final version to be published in late Aug or Sep.
R Cheatsheets
Summarized Lists of common R code and their effects. Convenient reference.

Import data with read.csv() from Base R vs fread() from data.table

- Base R: data1 <- read.csv('health_ins_cust.csv')
- data.table: data2 <- fread('health_ins_cust.csv')

- Data values in data1 is the same as data2, but their structure is different.

Summary

- Use of simple Summaries to explore data
- Use of simple Visualizations to explore data
- Some problems/issues may be detected now, others may be discovered with more sophisticated techniques or more subject-matter knowledge later.
- Start from understanding the business problem/opportunity/challenge.
  - Don’t be too quick to jump in to explore data.
- Package data.table
  - Good for Big Data
  - Good for Small Data