Basic R

CITS4009 Computational Data Analysis

Unit Coordinator: Dr Du Huynh

Department of Computer Science and Software Engineering
The University of Western Australia

Semester 2, 2022

Subsetting

先引入库

library(ggplot2) otherwise variable mpg is undefined;
library(crayon) otherwise function chr() is undefined.

Individual elements of a vector, matrix, array or data frame are accessed with "[ ]" by specifying their index, or their name

chr函数从1遍历到n

```
a <- data.frame(row.names = "mpg" %+% chr(1:nrow(mpg)),
                manufacturer = mpg$manufacturer,
                model = mpg$model, displ = mpg$displ,
                cyl = mpg$cyl)
head(a)
```

列名自己改

```
##      manufacturer model displ cyl
## mpg1         audi    a4   1.8   4
## mpg2         audi    a4   1.8   4
## mpg3         audi    a4   2.0   4
## mpg4         audi    a4   2.0   4
## mpg5         audi    a4   2.8   6
## mpg6         audi    a4   2.8   6
```

Subsetting

- By index, by row names and column names

用[]可以和C++一样定位

```
a[3,3]
## [1] 2

a["mpg3", "displ"]
## [1] 2
```

```
a["mpg3",]
```

```
##      manufacturer model displ cyl
## mpg3         audi    a4     2   4
```

Subsetting

- Subset rows by a vector of indices

与其他语言不同，R的数组计数是从1开始的

```
a[c(1:2),]
```

```
##      manufacturer model displ cyl
## mpg1         audi    a4   1.8   4
## mpg2         audi    a4   1.8   4
```

```
a[-c(2:nrow(mpg)), ]
```

负号取补集

```
##      manufacturer model displ cyl
## mpg1         audi    a4   1.8   4
```

Subsetting

- Subset rows by a logical vector

```
a[c(T,F,T),]
```

T = TRUE F = FALSE

T和F是关键字，此处取出1,3 4,6 7,9...行

```
##        manufacturer                 model displ cyl
## mpg1           audi                    a4   1.8   4
## mpg3           audi                    a4   2.0   4
## mpg4           audi                    a4   2.0   4
## mpg6           audi                    a4   2.8   6
## mpg7           audi                    a4   3.1   6
## mpg9           audi           a4 quattro   1.8   4
## mpg10          audi           a4 quattro   2.0   4
## mpg12          audi           a4 quattro   2.8   6
## mpg13          audi           a4 quattro   2.8   6
## mpg15          audi           a4 quattro   3.1   6
## mpg16          audi           a6 quattro   2.8   6
```

Subsetting

- Subset columns

```
a$manufacturer
```

左边的数字表示这一行的
第一个在表格是第几行

```
##   [1] audi       audi       audi       audi       audi
##   [7] audi       audi       audi       audi       audi
##  [13] audi       audi       audi       audi       audi
##  [19] chevrolet  chevrolet  chevrolet  chevrolet  chevrolet
##  [25] chevrolet  chevrolet  chevrolet  chevrolet  chevrolet
##  [31] chevrolet  chevrolet  chevrolet  chevrolet  chevrolet
##  [37] chevrolet  dodge      dodge      dodge      dodge
##  [43] dodge      dodge      dodge      dodge      dodge
##  [49] dodge      dodge      dodge      dodge      dodge
##  [55] dodge      dodge      dodge      dodge      dodge
##  [61] dodge      dodge      dodge      dodge      dodge
##  [67] dodge      dodge      dodge      dodge      dodge
##  [73] dodge      dodge      ford       ford       ford
```

Subsetting

- Comparison resulting in a logical vector

```
a$manufacturer == "audi"
```

```
##   [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRU
##  [13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Subsetting

- Subset the selected rows

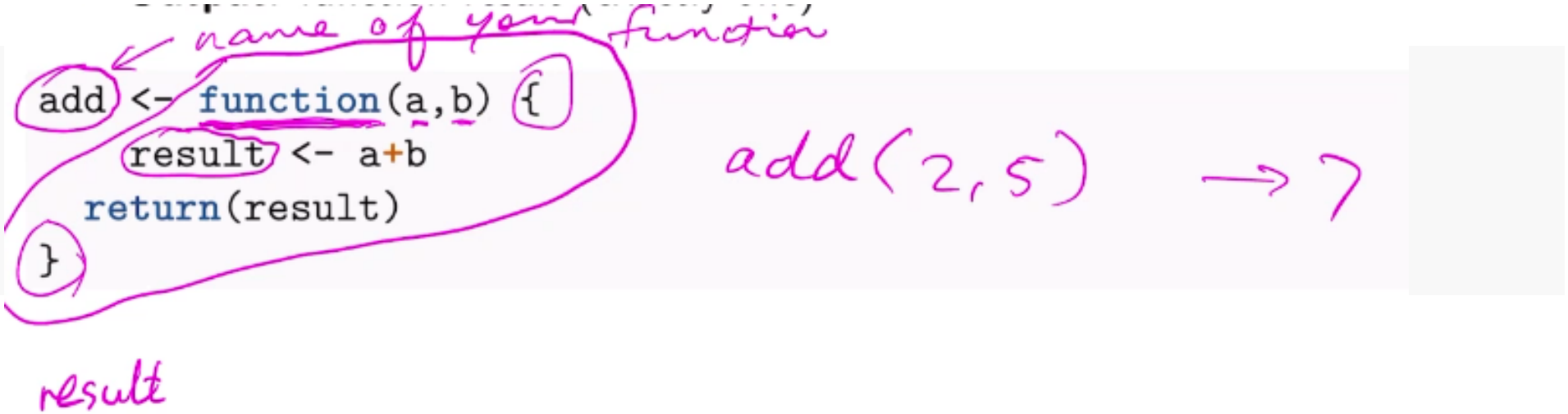
```
a[a$manufacturer == "audi" & a$model == "a4 quattro",]
```

```
##       manufacturer      model displ cyl
## mpg8          audi a4 quattro   1.8   4
## mpg9          audi a4 quattro   1.8   4
## mpg10         audi a4 quattro   2.0   4
## mpg11         audi a4 quattro   2.0   4
## mpg12         audi a4 quattro   2.8   6
## mpg13         audi a4 quattro   2.8   6
## mpg14         audi a4 quattro   3.1   6
## mpg15         audi a4 quattro   3.1   6
```

Functions

Functions take data as *input*, process it into *output*

- **Input**: function arguments (0, 1, 2, . . .)
- **Output**: function result (exactly one)



Operators

Operators: Short-cut writing for frequently used functions of one or two arguments.

- Assignment

<-	assignment operator

- Arithmetic

+	addition	-	subtraction
*	multiplication	/	division
^	exponent	%%	mod
%/%	integer division	%*%	dot product or matrix multiplication

此处A和B都是3*4，会提示不合法，需A的列数等于B的行数，如D

Operators (Why <-? Why not =?)

```
x <- rnorm(100)
y <- 2*x + rnorm(100)
lm(formula=y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     0.09097      1.99684
```

rnorm(n, mean = 0, sd = 1)

n 为产生随机值个数（长度），mean 是平均数， sd 是标准差 。

使用该函数的时候后，一般要赋予它 3个值.

rnorm （） 函数会随机正态分布，然后随机抽样 或者取值 n 次，

>rnorm （5，0,1） 以N （0,1）的正态分布，分别列出5个值。

r 这列代表随机，可以替换成dnorm, pnorm, qnorm 作不同计算

r = random = 随机， d= density = 密度， p= probability = 概率 ， q =quantile = 分位

- <- in the first two lines is used as an assignment operator;
- = in the third line does not serve as an assignment operator; instead, it is an operator that specifies a named parameter formula for the `lm` function.

Operators

- Set

`%in%` subset

- Logical

&	and
|	or
!	not

- Comparison

<	less than	>	greater than
<=	less or equal to	>=	greater or equal to
==	is equal to	!=	not equal to

Frequently used functions

- Basic stats (max, min, summary)
- Rounding (round, floor)
- Concatenate vectors (c, cbind, rbind)
- Size (length, dim, nrow, ncol)
- Vector sorting (sort, rank, order)
- Display or concatenate into a string (print, cat, paste, format)
- Others (apply, table, which)

a ← c(1, 10, 15, 5)
max(a)
min(a)

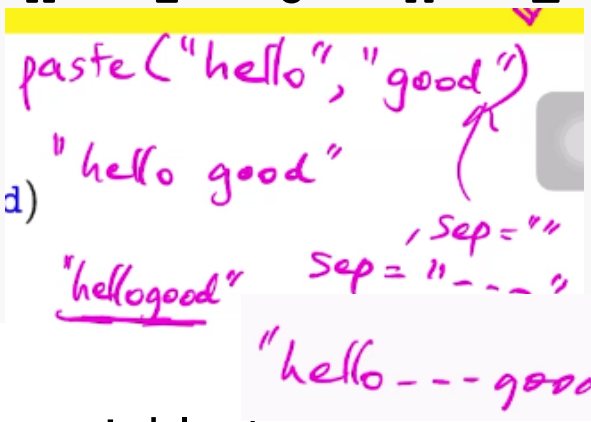
round是四舍五入，floor是向下取整

cbind可以将3*4和3*5的矩阵合并成3*9，rbind反之

length用在数组，dim,nrow,ncol用在矩阵

```
LETTERS
##  [1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N
## [20] "T" "U" "V" "W" "X" "Y" "Z"

which( LETTERS == "R" )
## [1] 18
```



(Note that LETTERS and letters are built-in variables)

Examples

a ["a", "b", "c", "d", "e"]

sample(a, 5, replace = FALSE)

```
a <- letters[1:5]
b <- table(a, sample(a))
b
```

```
## a   a b c d e
##   a 0 0 1 0 0
##   b 1 0 0 0 0
##   c 0 0 0 0 1
##   d 0 0 0 1 0
##   e 0 1 0 0 0
```

a [c a e d b]

```
apply(b, 1, mean)
```

```
##   a   b   c   d   e
## 0.2 0.2 0.2 0.2 0.2
```

Branching

分支结构

```
if (logical_expression) {
  statements
} else {
  alternative_statements
}
```

只有一句时不一定要大括号

The `else` part is optional. The braces `{ }` is optional if only one statement for the logical expression.

Branching Example

```
x <- -4
if (x >= 0) {
  print(sqrt(x))
} else {
  print(NA)
}
```

```
## [1] NA
```

More Branching

```
ifelse (logical_expression, yes_statement, no_statement)
```

```
x <- c(4:-4)
sqrt(ifelse(x >= 0, x, NA))
```

```
## [1] 2.000000 1.732051 1.414214 1.000000 0.000000       NA
## [9]       NA
```

Looping

When the same or similar tasks need to be performed multiple times; for all elements of a list; for all columns of an array; etc.

```
for (i in 1:5) {
   print(i*i)
}
```

```
## [1] 1
## [1] 4
## [1] 9
## [1] 16
## [1] 25
```

Looping

```
i <- 1
while (i <= 5) {
    print(i*i)
    i <- i + sqrt(i)
}
```

```
## [1] 1
## [1] 4
## [1] 11.65685
```

Transfer control within loop: `repeat`, `break`, `next`.

References

- **Practical Data Science with R**, *Nina Zumel, John Mount*, Manning, 2nd Ed., 2020
- **R for Data Science**, *Hadley Wickham, Garrett Grolemund*, O'Reilly, 2017 (Chapter 3)
- **Introduction to the R language:** https://users.soe.ucsc.edu/~lshiue/bioc/Rintro.ppt
- **An Introduction to R:** http://csg.sph.umich.edu/abecasis/class/815.04.pdf
- **Differences between assignment operators in R:** https://renkun.me/2014/01/28/difference-between-assignment-operators-in-r/

Data at a Glance

CITS4009 Computational Data Analysis

Unit Coordinator: Dr Du Huynh

Department of Computer Science and Software Engineering
The University of Western Australia

Semester 2, 2022

The Customer Dataset

Synthetic example data derived from Census PUMS data to predict the probability of health insurance coverage.

Data can be obtained from:
https://github.com/WinVector/zmPDSwR/tree/master/Custdata

```
custdata <- read.table('custdata.tsv', header=T, sep='\t')
```

Customer Data Structure

```
str(custdata)
```

```
## 'data.frame':    1000 obs. of  11 variables:
##  $ custid      : int  2068 2073 2848 5641 6369 8322 8521 12
##  $ sex         : Factor w/ 2 levels "F","M": 1 1 2 2 1 1 2
##  $ is.employed : logi  NA NA TRUE TRUE TRUE TRUE ...
##  $ income      : int  11300 0 4500 20000 12000 180000 12000
##  $ marital.stat: Factor w/ 4 levels "Divorced/Separated",..
##  $ health.ins  : logi  TRUE TRUE FALSE FALSE TRUE TRUE ...
##  $ housing.type: Factor w/ 4 levels "Homeowner free and cle
##  $ recent.move : logi  FALSE TRUE TRUE FALSE TRUE FALSE ...
##  $ num.vehicles: int  2 3 3 0 1 1 1 3 2 1 ...
##  $ age         : num  49 40 22 22 31 40 39 48 44 70 ...
##  $ state.of.res: Factor w/ 50 levels "Alabama","Alaska",..
```

Customer Data Summary

```
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   Homeowner free ar
##  Married            :516   FALSE:159       Homeowner with mo
##  Never Married      :233   TRUE :841       Occupied with no
##  Widowed            : 96                   Rented
##                                            NA's
```

Using Summary Statistics to spot problems

In R, you'll typically use the `summary()` command to take your first look at the data.

The goal is to understand whether you have the kind of customer information that

- can potentially help you predict health insurance coverage, and
- whether the data is of good enough quality to be informative.

Looking for several common issues:

- Missing values
- Invalid values and outliers
- Data ranges that are too wide or too narrow
- The units of the data

Read the summary

```
is.employed         income
Mode :logical   Min.   : -8700
FALSE:73        1st Qu.: 14600
TRUE :599       Median : 35000
NA's :328       Mean   : 53505
                3rd Qu.: 67000
                Max.   :615000

marital.stat
Divorced/Separated:155
Married           :516
Never Married     :233
Widowed           : 96

health.ins
Mode :logical
FALSE:159
TRUE :841
NA's :0

housing.type
Homeowner free and clear    :157
Homeowner with mortgage/loan:412
Occupied with no rent       : 11
Rented                      :364
NA's                        : 56

recent.move      num.vehicles
Mode :logical   Min.   :0.000
FALSE:820       1st Qu.:1.000
TRUE :124       Median :2.000
NA's :56        Mean   :1.916
                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
```

The variable is.employed is missing for about a third of the data. The variable income has negative values, which are potentially invalid.

About 84% of the customers have health insurance.

The variables housing.type, recent.move, and num.vehicles are each missing 56 values.

The average value of the variable age seems plausible, but the minimum and maximum values seem unlikely. The variable state.of.res is a categorical variable; summary() reports how many customers are in each state (for the first few states).

Introduction to ggplot

CITS4009 Computational Data Analysis

Unit Coordinator: Dr Du Huynh

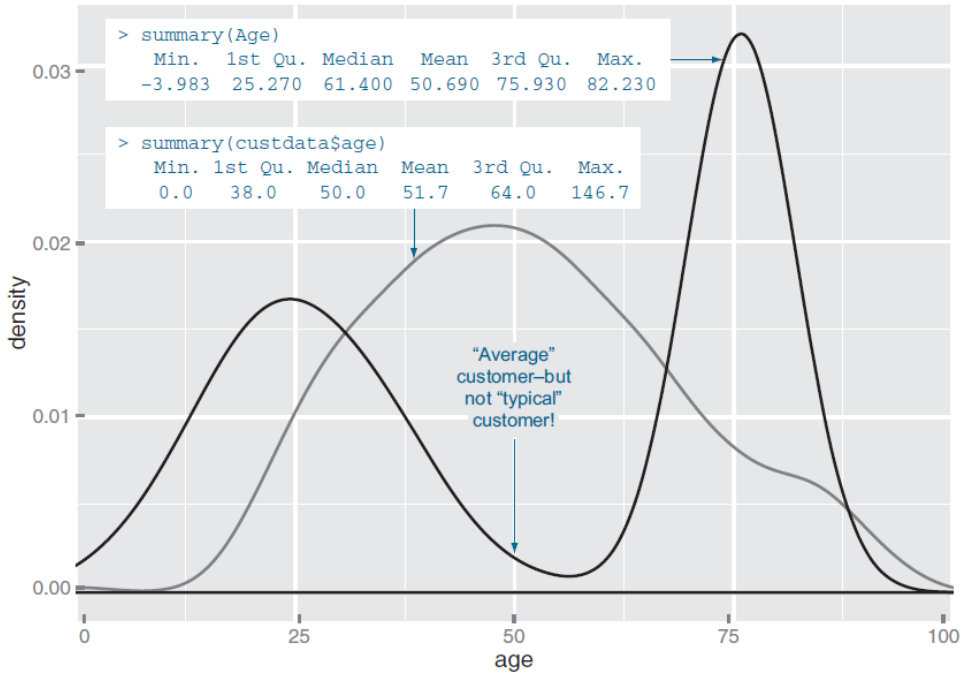
Department of Computer Science and Software Engineering
The University of Western Australia

Semester 2, 2022

Single Variable Plots

Distribution of a single variable

- What is the peak value of the distribution?
- How many peaks are there in the distribution (unimodality versus bimodality)?
- How normal (or lognormal) is the data?
- How much does the data vary? Is it concentrated in a certain interval or in a certain category?



Plots for single variable distribution

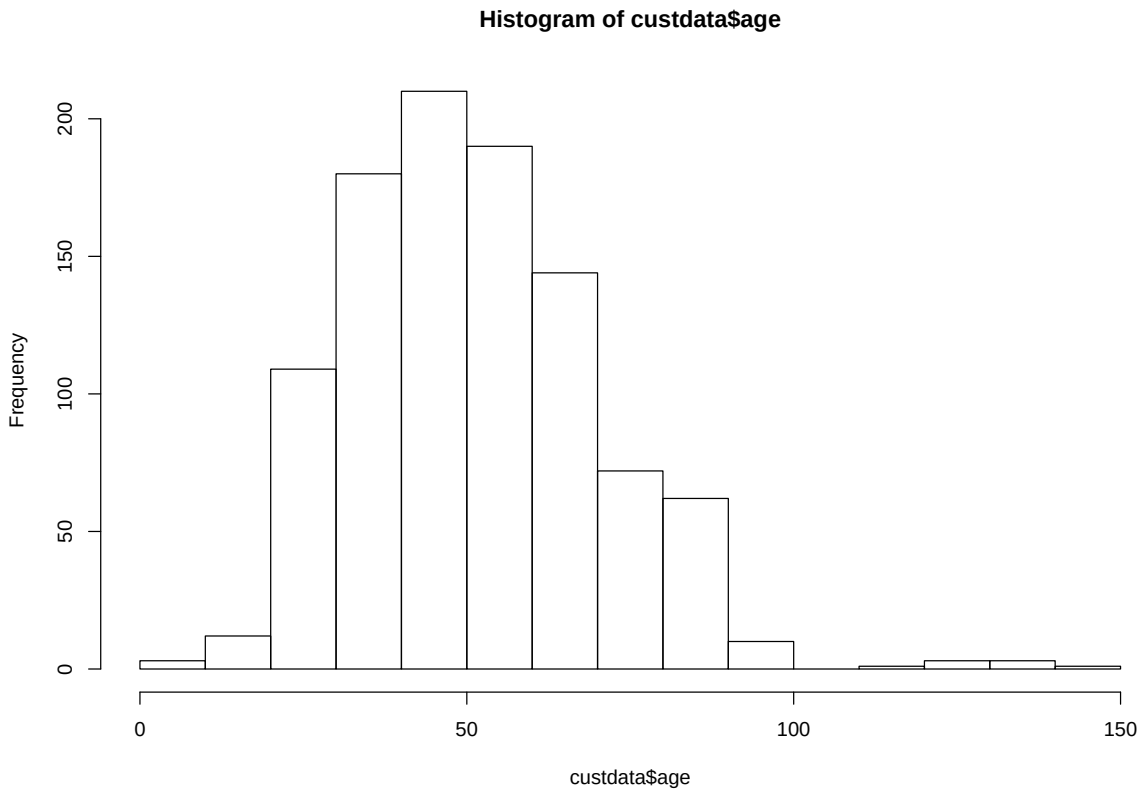
Graph Type	Uses
Histogram	Examines data range
Density Plot	Checks number of modes Checks if distribution is normal/lognormal/etc
Boxplot	Checks for anomalies and outliers
Bar Chart	Compares relative or absolute frequencies of the values of a categorical variable

Histograms

Histograms - the `hist` function in R

A basic histogram bins a variable into fixed-width buckets and returns the number of data points that falls into each bucket.

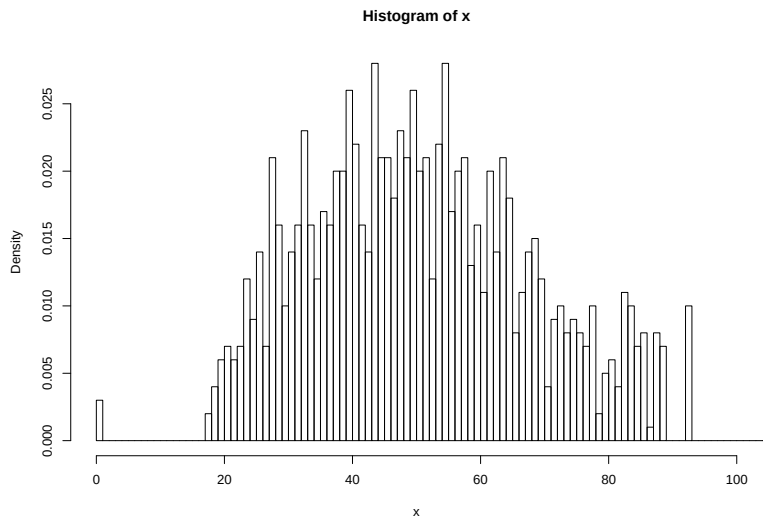
```
custdata <- read.table('custdata.tsv',header=T, sep='\t')
hist(custdata$age)
```



Histogram: Other useful options

- `breaks`: takes a sequence to specify where the breaks are
- `xlim`: takes the start and end point of x axis
- `freq`: `TRUE` for raw counts; `FALSE` for density (normalized by the total count), and the areas of the bars add to 1. This is called "density plot" in ggplot, except it is not a continuous line plot.

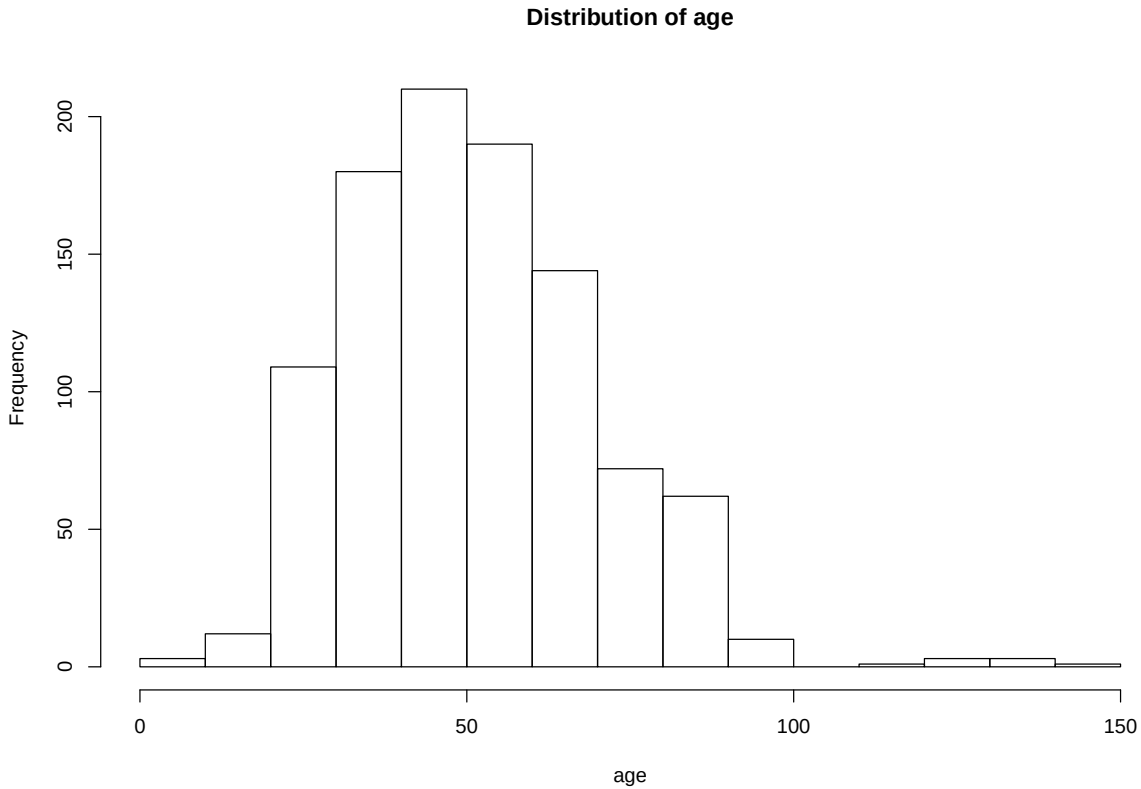
```
x <- custdata$age
hist(x, breaks=seq(0,150,1), xlim=c(0,100), freq = FALSE)
```



Adding titles

- Using Attributes of the function

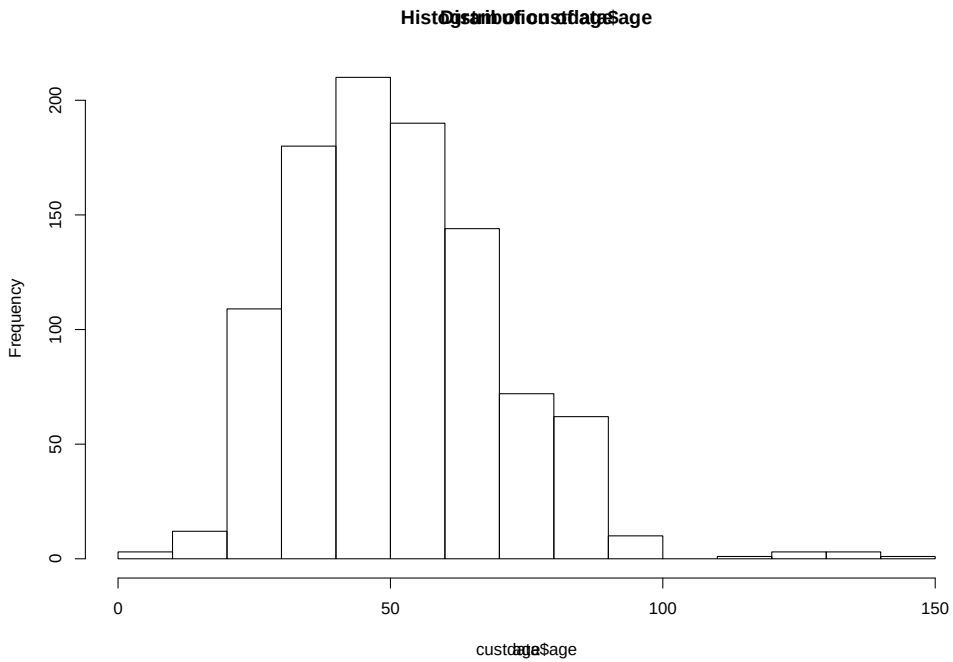
```
hist(custdata$age,main="Distribution of age",xlab="age")
```



Adding titles

- Using the `title()` function

```
hist(custdata$age)
title('Distribution of age',xlab='age')
```



To remove the default title from `hist`, do: `hist(custdata$age, main="")`

A layered grammar of graphs - ggplot

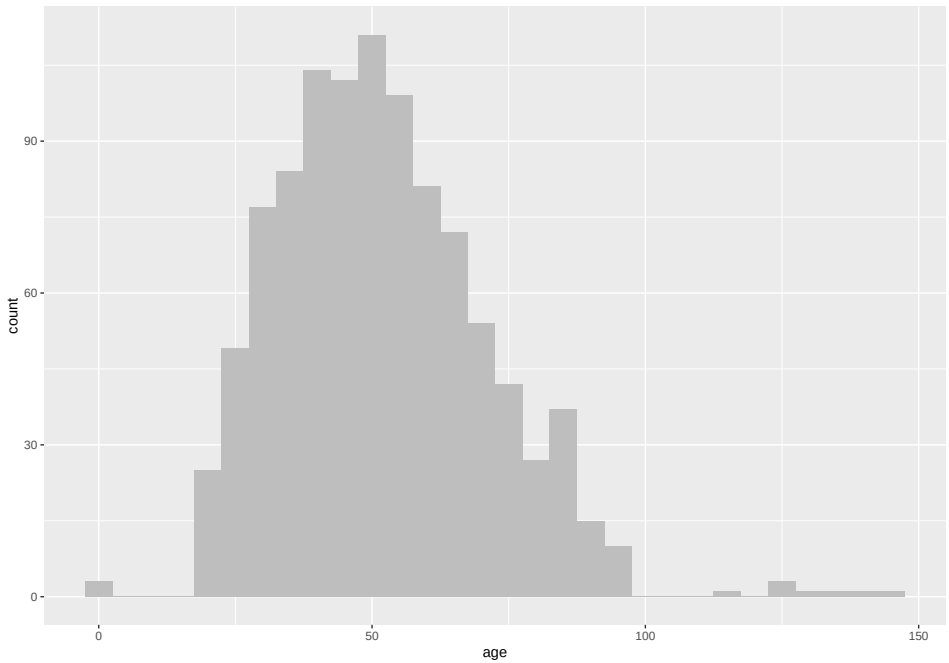
ggplot

- R has several systems for making graphs, but `ggplot2` is one of the most elegant and most versatile libraries.
- `ggplot2` implements the **grammar of graphics**, a coherent system for describing and building graphs.
- Begin a plot with the function `ggplot()`, which takes the data and create a coordinate system that you can add **layers** to.
- A reusable template for making graphs with `ggplot2` is given below. To make a graph, replace the bracketed parts in the code below with
  - a dataset,
  - a geom function (chart type), or
  - a collection of mappings (data selection for each coordinate).

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

Histograms (ggplot2)

```
library(ggplot2)
ggplot(data = custdata) +
  geom_histogram(mapping = aes(x=age),
                 binwidth=5, fill="gray")
```

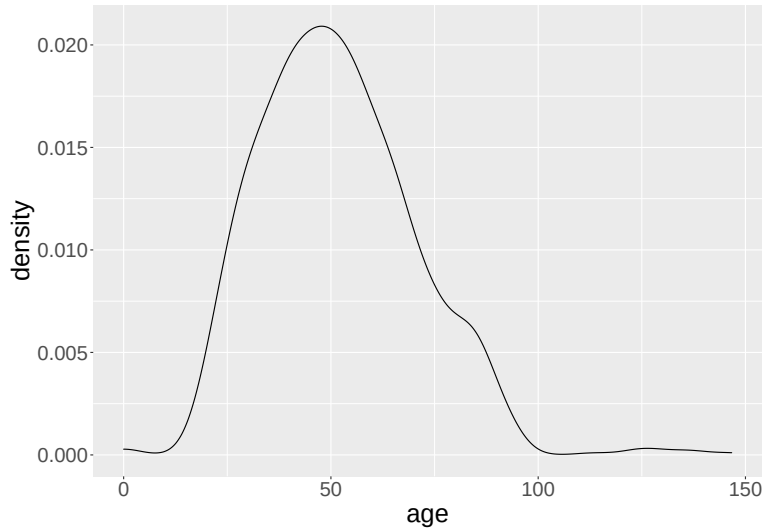


Density plot (ggplot2)

In ggplot, a **density plot** is a "continuous histogram" of a variable, except the area under the density plot is equal to 1.

- A point on a density plot corresponds to the fraction of data (or the percentage of data, divided by 100) that takes on a particular value.

```
library(ggplot2)
ggplot(custdata) + geom_density(aes(x=age)) +
    theme(text = element_text(size = 24))
```



References

- Practical Data Science with R. By Nina Zumel and John Mount, Manning, 2014. (Chapter 3)
- R for Data Science. By Garret Grokemund and Hardley Wickham, O'Relly, 2017. (Chapter 3)
- Introduction to the R language: https://users.soe.ucsc.edu/~lshiue/bioc/Rintro.ppt
- An Introduction to R: http://csg.sph.umich.edu/abecasis/class/815.04.pdf
- Differences between assignment operators in R: https://renkun.me/2014/01/28/difference-between-assignment-operators-in-r/