Biostatistics I: Introduction to R

Data transformation, exploration and visualization

Eleni-Rosalina Andrinopoulou

Department of Biostatistics, Erasmus Medical Center

e.andrinopoulou@erasmusmc.nl

@erandrinopoulou

Erasmus MC
University Medical Center Rotterdam

In this Section

- Data transformation
- Data exploration
- Data visualization
- A lot of practice

Data Transformation

You will never receive the perfect data set!

- **Round** continuous variables
- Convert **numeric** variables to **factors**
- Compute **new variables**
  - transform variables
- **Sort** the data set
- Data sets of **wide** $\Longleftrightarrow$ **long** format

Data Transformation

- **Round** continuous variables

```
pbc[1:3, c("time", "age", "bili", "chol")]
```

```
  time      age bili chol
1  400 58.76523 14.5  261
2 4500 56.44627  1.1  302
3 1012 70.07255  1.4  176
```

```
round(pbc[1:3, c("time", "age", "bili", "chol")], digits = 2)
```

```
  time   age bili chol
1  400 58.77 14.5  261
2 4500 56.45  1.1  302
3 1012 70.07  1.4  176
```

Data Transformation

- Convert **numeric** variables to **factors**

```
DF <- pbc[,c("id", "time", "status", "trt", "age",
                    "sex", "bili", "chol")]
head(DF)
```

```
  id time status trt      age sex bili chol
1  1  400      2   1 58.76523   f 14.5  261
2  2 4500      0   1 56.44627   f  1.1  302
3  3 1012      2   1 70.07255   m  1.4  176
4  4 1925      2   1 54.74059   f  1.8  244
5  5 1504      1   2 38.10541   f  3.4  279
6  6 2503      2   2 66.25873   f  0.8  248
```

Data Transformation

- Convert **numeric** variables to **factors**

```
DF <- pbc[,c("id", "time", "status", "trt", "age",
                    "sex", "bili", "chol")]
DF$trt <- factor(x = DF$trt, levels = c(1, 2),
                 labels = c("D-penicillmain", "placebo"))
head(DF)
```

```
  id time status            trt      age sex bili chol
1  1  400      2 D-penicillmain 58.76523   f 14.5  261
2  2 4500      0 D-penicillmain 56.44627   f  1.1  302
3  3 1012      2 D-penicillmain 70.07255   m  1.4  176
4  4 1925      2 D-penicillmain 54.74059   f  1.8  244
5  5 1504      1        placebo 38.10541   f  3.4  279
6  6 2503      2        placebo 66.25873   f  0.8  248
```

Data Transformation

- Compute **new variables**
  - transform variables

```
DF <- pbc[,c("id", "time", "status", "trt", "age",
                    "sex", "bili", "chol")]
head(DF)
```

```
  id time status trt      age sex bili chol
1  1  400      2   1 58.76523   f 14.5  261
2  2 4500      0   1 56.44627   f  1.1  302
3  3 1012      2   1 70.07255   m  1.4  176
4  4 1925      2   1 54.74059   f  1.8  244
5  5 1504      1   2 38.10541   f  3.4  279
6  6 2503      2   2 66.25873   f  0.8  248
```

Data Transformation

- Compute **new variables**
  - transform variables

```
DF <- pbc[,c("id", "time", "status", "trt", "age",
                    "sex", "bili", "chol")]
DF$time <- DF$time/30
DF$time_years <- DF$time/12
head(DF)
```

```
  id      time status trt      age sex bili chol time_years
1  1  13.33333      2   1 58.76523   f 14.5  261   1.111111
2  2 150.00000      0   1 56.44627   f  1.1  302  12.500000
3  3  33.73333      2   1 70.07255   m  1.4  176   2.811111
4  4  64.16667      2   1 54.74059   f  1.8  244   5.347222
5  5  50.13333      1   2 38.10541   f  3.4  279   4.177778
6  6  83.43333      2   2 66.25873   f  0.8  248   6.952778
```

Data Transformation

- **Sort** the data set in either ascending or descending order
  - The variable by which we sort can be a numeric, string or factor

```
head(sort(pbc$bili))
```

```
[1] 0.3 0.3 0.3 0.4 0.4 0.4
```

Data Transformation

- **Sort** the data set in either ascending or descending order
  - The variable by which we sort can be a numeric, string or factor

```
head(pbc[order(pbc$bili), ])
```

```
     id time status trt      age sex ascites hepato spiders edema bili chol
8     8 2466      2   2 53.05681   f       0      0       0     0  0.3  280
36   36 3611      0   2 56.41068   f       0      0       0     0  0.3  172
163 163 2055      2   1 53.49760   f       0      0       0     0  0.3  233
84   84 4032      0   2 55.83025   f       0      0       0     0  0.4  263
108 108 2583      2   1 50.35729   f       0      0       0     0  0.4  127
135 135 3150      0   1 42.96783   f       0      0       0     0  0.4  263
    albumin copper alk.phos    ast trig platelet protime stage
8      4.00     52   4651.2  28.38  189      373    11.0     3
36     3.39     18    558.0  71.30   96      311    10.6     2
163    4.08     20    622.0  66.65   68      358     9.9     3
84     3.76     29   1345.0 137.95   74      181    11.2     3
108    3.50     14   1062.0  49.60   84      334    10.3     2
135    3.57    123    836.0  74.40  121      445    11.0     2
```

Data Transformation

- **Sort** the data set in either ascending or descending order
  - The variable by which we sort can be a numeric, string or factor

```
head(pbc[order(pbc$bili, pbc$age), ])
```

```
     id time status trt      age sex ascites hepato spiders edema bili chol
8     8 2466      2   2 53.05681   f       0      0       0   0.0  0.3  280
163 163 2055      2   1 53.49760   f       0      0       0   0.0  0.3  233
36   36 3611      0   2 56.41068   f       0      0       0   0.0  0.3  172
135 135 3150      0   1 42.96783   f       0      0       0   0.0  0.4  263
320 320 2403      0  NA 44.00000   f      NA     NA      NA   0.5  0.4   NA
168 168 2713      0   2 47.75359   f       0      1       0   0.0  0.4  257
    albumin copper alk.phos   ast trig platelet protime stage
8      4.00     52   4651.2 28.38  189      373    11.0     3
163    4.08     20    622.0 66.65   68      358     9.9     3
36     3.39     18    558.0 71.30   96      311    10.6     2
135    3.57    123    836.0 74.40  121      445    11.0     2
320    3.81     NA       NA    NA   NA      226    10.5     3
168    3.80     44    842.0 97.65  110       NA     9.2     2
```

Data Transformation

- Data sets of **wide** $\Longleftrightarrow$ long format

```
head(pbc[,c("id", "time", "status", "trt", "age",
                    "sex", "bili", "chol")])
```

```
  id time status trt      age sex bili chol
1  1  400      2   1 58.76523   f 14.5  261
2  2 4500      0   1 56.44627   f  1.1  302
3  3 1012      2   1 70.07255   m  1.4  176
4  4 1925      2   1 54.74059   f  1.8  244
5  5 1504      1   2 38.10541   f  3.4  279
6  6 2503      2   2 66.25873   f  0.8  248
```

Data Transformation

- Data sets of wide $\Longleftrightarrow$ **long** format

```
head(pbcseq[, c("id", "futime", "status", "trt", "age", "day",
                "sex", "bili", "chol")])
```

```
  id futime status trt      age day sex bili chol
1  1    400      2   1 58.76523   0   f 14.5  261
2  1    400      2   1 58.76523 192   f 21.3   NA
3  2   5169      0   1 56.44627   0   f  1.1  302
4  2   5169      0   1 56.44627 182   f  0.8   NA
5  2   5169      0   1 56.44627 365   f  1.0   NA
6  2   5169      0   1 56.44627 768   f  1.9   NA
```

Data Transformation

- Data sets of **wide** $\Longleftrightarrow$ **long** format

```
?reshape
```

Data Exploration

- Common questions for the pbc data set
  - What is the mean and standard deviation for age?
  - What is the mean and standard deviation for time?
  - What is the median and interquartile range for age?
  - What is the percentage of placebo patients?
  - What is the percentage of females?
  - What is the mean and standard deviation for age in males?
  - What is the mean and standard deviation for baseline serum bilirubin?
  - What is the percentage of missings in serum bilirubin?

All these questions can be answered using R!

Data Exploration

- **Hints**

- Check functions: **mean(...)**, **sd(...)**, **percent(...)**, **median(...)**, **IQR(...)**, **table(...)**

Data Exploration

- **Hints**

- Check functions: **mean(...)**, **sd(...)**, **percent(...)**, **median(...)**, **IQR(...)**, **table(...)**

What is the mean value for age?

```
mean(pbc$age)
```

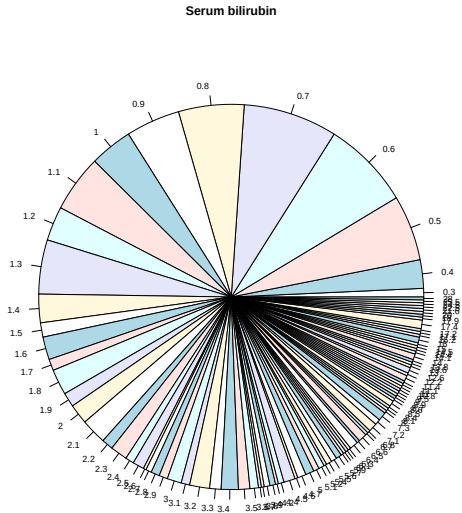
```
[1] 50.74155
```

Data Visualization

- It is important to investigate each variable in our data set using plots
  - Descriptive statistics for continuous and categorical variables
  - Distributions of variables
  - Distributions of variables per group
  - Extreme values
  - Linear/nonlinear evolutions

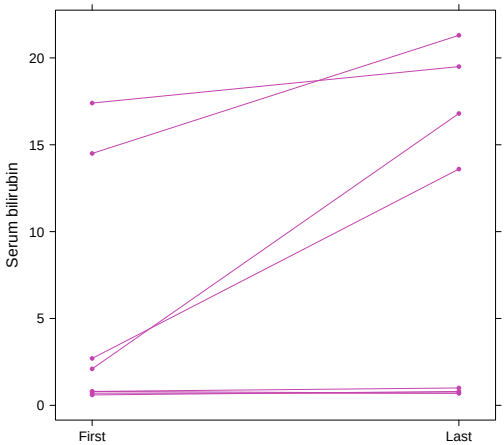
Data Visualization

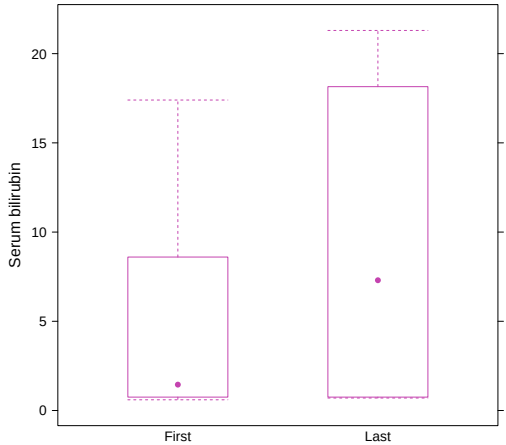
Take care!



Data Visualization

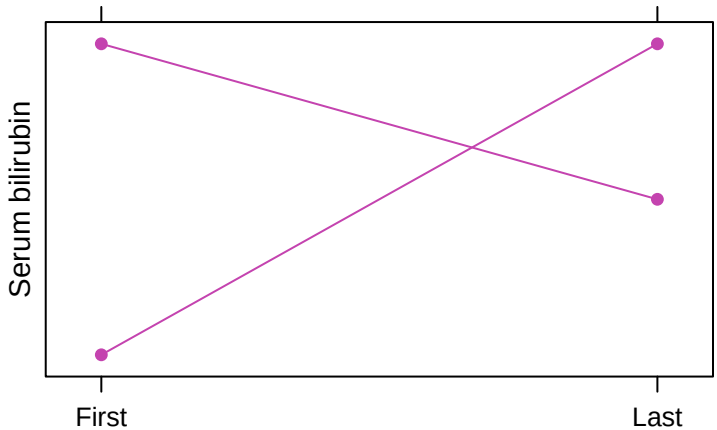
Take care!





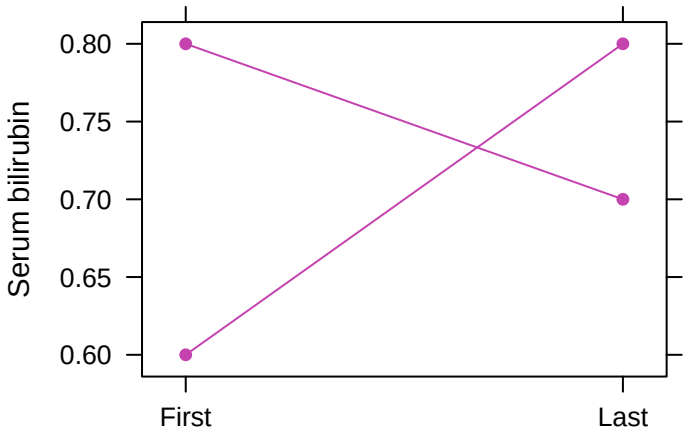
Data Visualization

Take care!



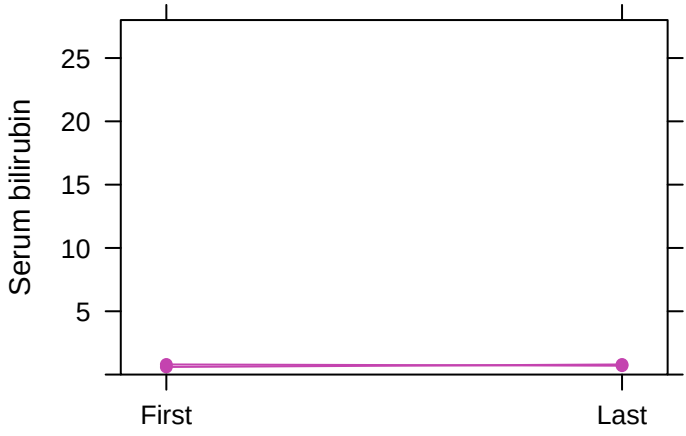
Data Visualization

Take care!



Data Visualization

Take care!



Data Visualization

- R has very powerful graphics capabilities
- Some good references are
  - Murrel, P. (2005) R Graphics. Boca Raton: Chapman & Hall/CRC.
  - Sarkar, D. (2008) Lattice Multivariate Data Visualization with R. New York: Springer-Verlag.

Data Visualization

- Traditional graphics system
  - package **graphics**
- Trellis graphics system
  - package **lattice** (which is based on package grid)
- Grammar of Graphics implementation (i.e., Wilkinson, L. (1999) The Grammar of Graphics. New York: Springer-Verlag)
  - packages **ggplot** & **ggplot2**

Data Visualization

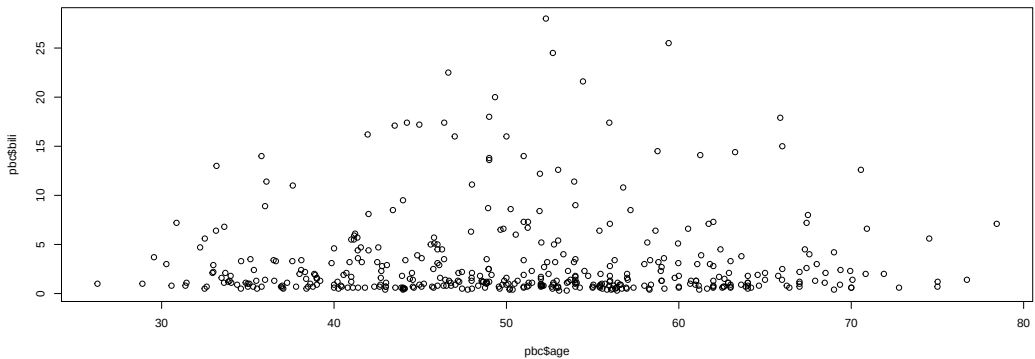
Important plotting basic functions

- **plot()**: scatter plot (and others)
- **barplot()**: bar plots
- **boxplot()**: box-and-whisker plots
- **hist()**: histograms
- **dotchart()**: dot plots
- **pie()**: pie charts
- **qqnorm()**, **qqline()**, **qqplot()**: distribution plots
- **pairs()**: for multivariate data

Data Visualization

Continuous variables

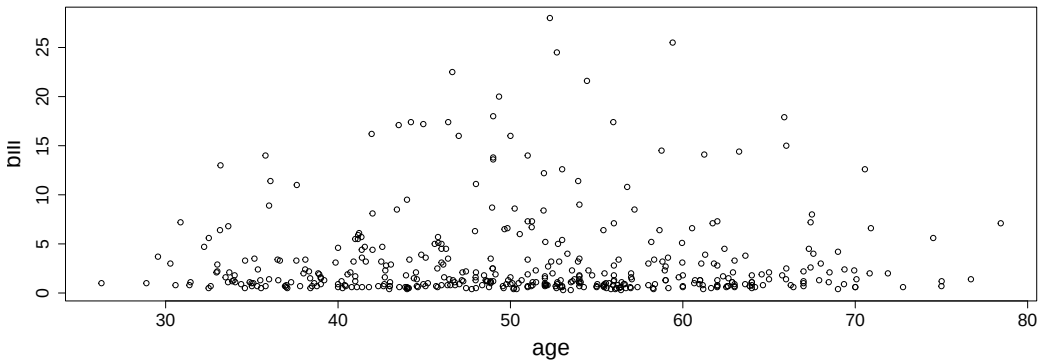
```
plot(x = pbc$age, y = pbc$bili)
```



Data Visualization

Continuous variables

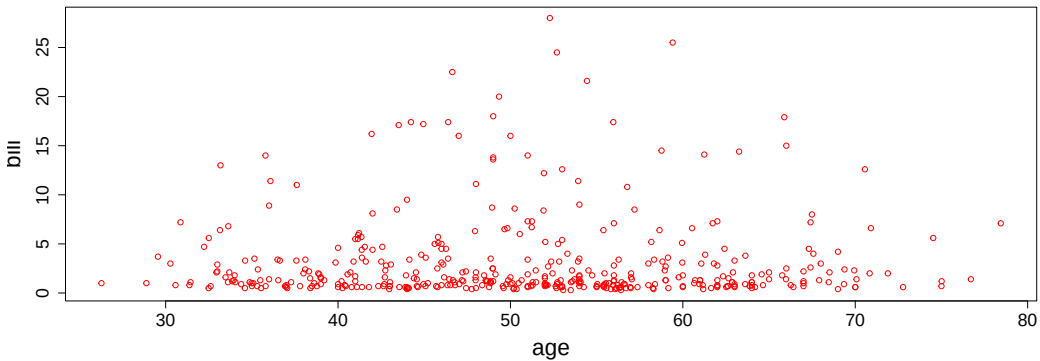
```
plot(x = pbc$age, y = pbc$bili, xlab = "age", ylab = "bili",
     cex.lab = 1.9, cex.axis = 1.5)
```



Data Visualization

Continuous variables

```
plot(x = pbc$age, y = pbc$bili, xlab = "age", ylab = "bili",
     cex.lab = 1.9, cex.axis = 1.5, col = "red")
```



Data Visualization

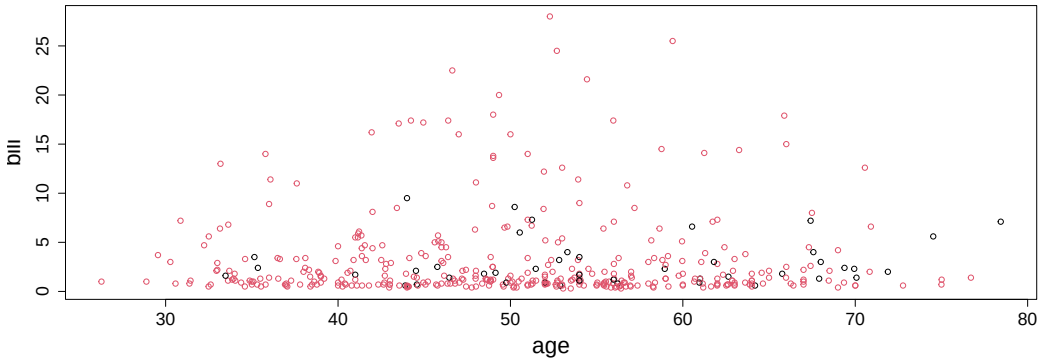
- For more options check

```
?plot
```

Data Visualization

Continuous variables per group

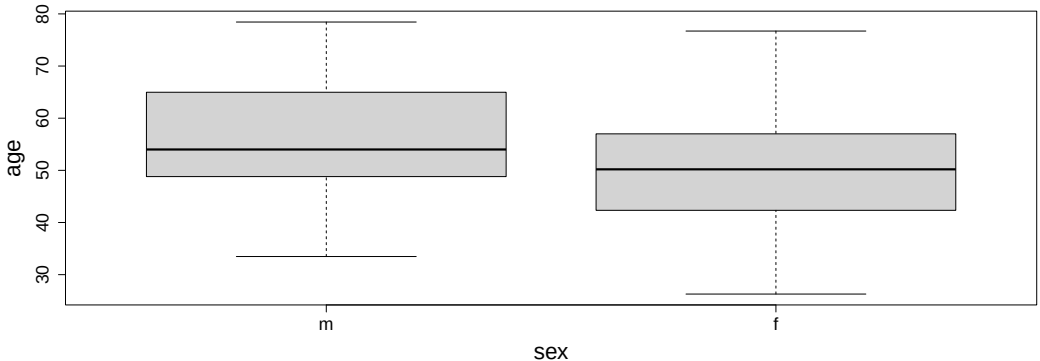
```
plot(x = pbc$age, y = pbc$bili, xlab = "age", ylab = "bili",
     cex.lab = 1.9, cex.axis = 1.5, col = pbc$sex)
```



Data Visualization

Continuous variables per group

```
boxplot(formula = pbc$age ~ pbc$sex, xlab = "sex", ylab = "age",
     cex.lab = 1.9, cex.axis = 1.5)
```



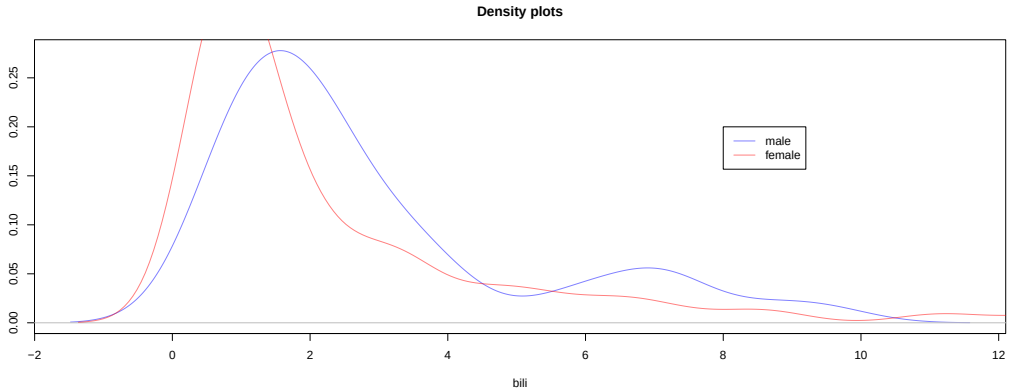
Data Visualization

Continuous variables per group

```
pbc_male_bili <- pbc$bili[pbc$sex == "m"]
pbc_female_bili <- pbc$bili[pbc$sex == "f"]
plot(density(x = pbc_male_bili), col = rgb(0,0,1,0.5),
     main = "Density plots", xlab = "bili", ylab = "")
lines(density(x = pbc_female_bili), col = rgb(1,0,0,0.5))
legend(x = 8, y = 0.2, legend = c("male", "female"),
       col = c(rgb(0,0,1,0.5), rgb(1,0,0,0.5)), lty = 1)
```

Data Visualization

Continuous variables per group



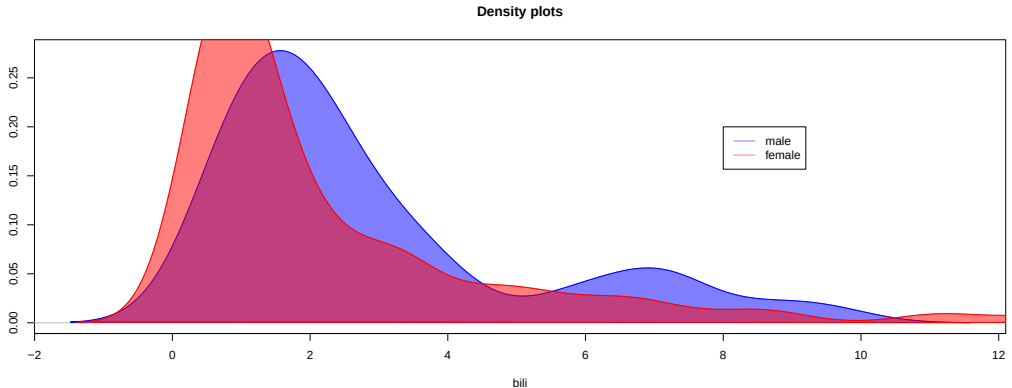
Data Visualization

Continuous variables per group

```
pbc_male_bili <- pbc$bili[pbc$sex == "m"]
pbc_female_bili <- pbc$bili[pbc$sex == "f"]
plot(density(x = pbc_male_bili), col = rgb(0,0,1,0.5),
     main = "Density plots", xlab = "bili", ylab = "")
polygon(density(x = pbc_male_bili), col = rgb(0,0,1,0.5),
        border = "blue")
lines(density(x = pbc_female_bili), col = rgb(1,0,0,0.5))
polygon(density(x = pbc_female_bili), col = rgb(1,0,0,0.5),
        border = "red")
legend(x = 8, y = 0.2, legend = c("male", "female"),
       col = c(rgb(0,0,1,0.5), rgb(1,0,0,0.5)), lty = 1)
```

Data Visualization

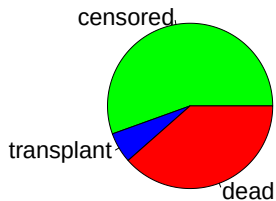
Continuous variables per group



Data Visualization

Categorical variables

```
pbc$status <- factor(x = pbc$status, levels = c(0, 1, 2),
                     labels = c("censored", "transplant", "dead"))
pie(table(pbc$status), col = c("green", "blue", "red"), cex = 2)
```



Summary

Transformation

- `round()`
- `factor()`
- `order()`
- `reshape()`

Exploration

- `mean()`, `sd()`
- `median()`, `IQR()`
- `table()`

Visualization

- `plot()`, `legend()`
- `hist()`
- `barchart()`
- `boxplot()`
- `xyplot()`, `ggplot()`
- `par()`