CeDoSIA SS2020 - Exercise Sheet 4: Simple Data Manipulation & Visualization II

Vangelis Theodorakis, Xueqi Cao, Daniela Andrade Salazar, Julien Gagneur

29 June, 2020

Package

BiocStyle 2.14.4

Contents

0.1 Setup

```
library(ggplot2)
library(data.table)
library(magrittr)   # Needed for %>% operator
library(tidyr)
library(ggthemes)
```

1 Enchancing plots

Below is a graph taken from a published paper. Read the figure legend.

1) Discuss good and bad graphical properties of the plot. Make suggestions on how to improve it.
2) Implement a better visualization. As the original data is not available, we use the data simulated with the code below (also uploaded to Moodle).

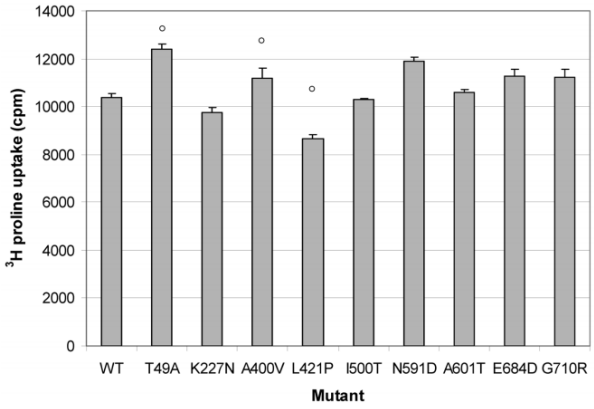


Figure 2. Maximal ^{3}H proline uptake of wildtype (WT) and all tested mutants. The maximum in uptake was measured in the presence of 3 µM cold L-proline. Data are expressed as means ± standard deviation (SD) obtained from triplicate samples. Mutants with a circle were tested in a second independent experiment.
doi:10.1371/journal.pone.0068645.g002

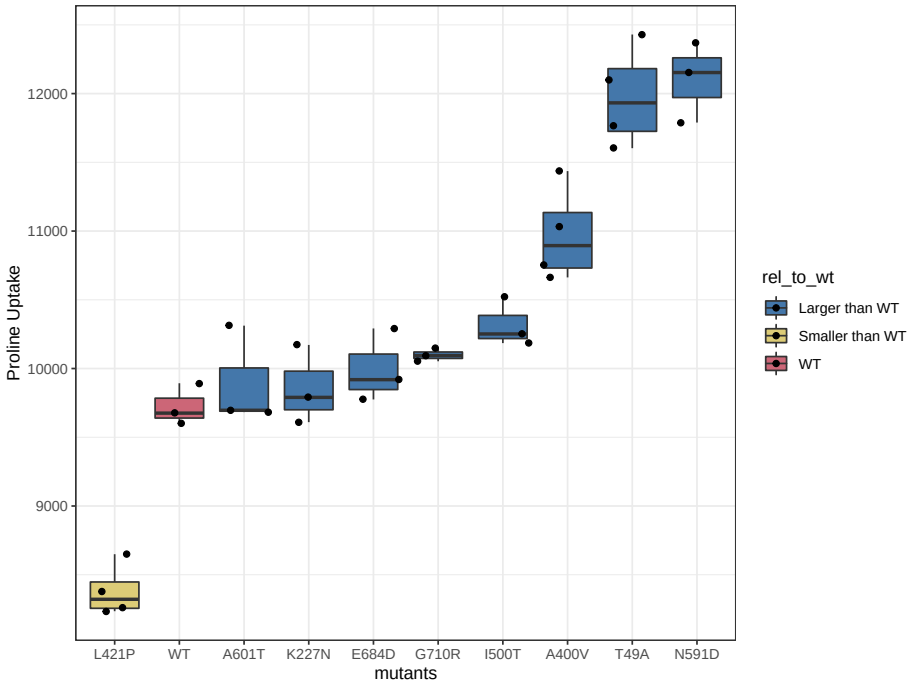
```
# GOOD
# - simple design
# - not too many colors
# - clear labels
# - no chart junk
# - horizontal grid
#
# BAD
# - no highlight, e.g. by color
# - x-axis not sorted
# - summary by mean+sd hides the data, which is at most four points per bar
#
# Suggestion
# - plot single points instead of bars, with small median line (too few points for boxplot)
```

```
# - sort Mutants by median
# - give color for above and below WT
```

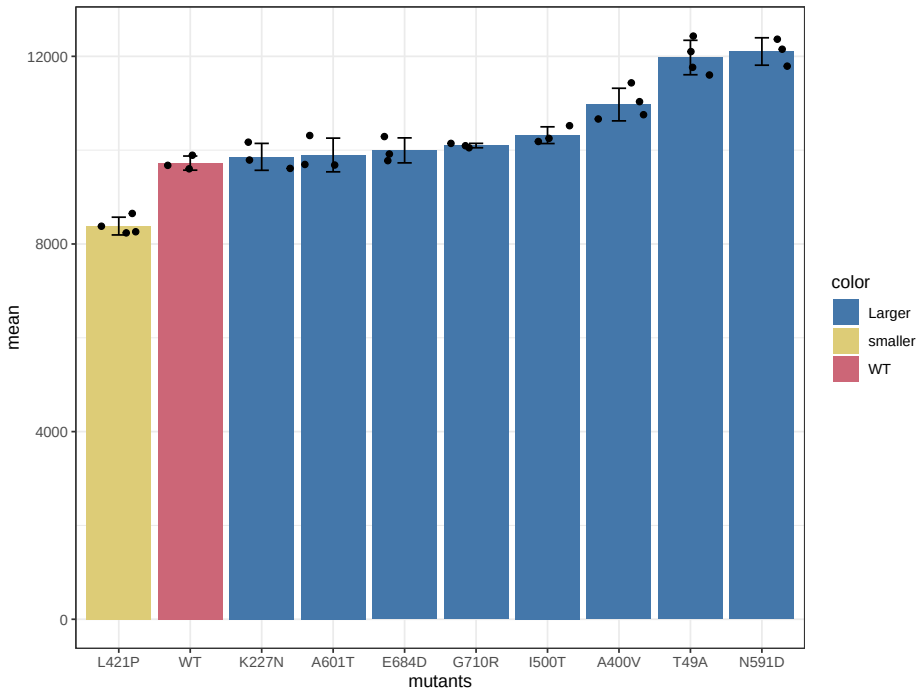
```
# simulate data
dt <- data.table(pro_uptake = c(
  rnorm(3, 10100, 300), rnorm(4, 12100, 300), rnorm(3, 9850, 300),
  rnorm(4, 11100, 300), rnorm(4,8300, 300), rnorm(3,10050, 300),
  rnorm(3, 12000, 300), rnorm(3, 10020, 300), rnorm(3, 10080, 300),
  rnorm(3, 10070, 300)),
  mutants = c(rep('WT',3), rep('T49A',4), rep('K227N',3), rep('A400V',4),
              rep('L421P',4), rep('I500T',3), rep('N591D',3), rep('A601T',3),
              rep('E684D',3), rep('G710R',3) )
)
```

```
# sort by median
dt[, median_per_mut := median(pro_uptake), by = mutants]
wt_med = unique(dt[mutants == 'WT', median_per_mut])
dt[, mutants:= factor(mutants, levels=unique(dt[order(median_per_mut), mutants]))]

# assign class by relation to WT, useful to give color
dt[, rel_to_wt := ifelse(median_per_mut < wt_med, 'Smaller than WT', 'Larger than WT'),
   by = mutants]
dt[mutants == 'WT', rel_to_wt := 'WT']
```

```
ggplot(dt, aes(mutants, pro_uptake, fill = rel_to_wt)) +
  geom_boxplot() +
  geom_jitter(width = 0.4) +
  labs(y = "Proline Uptake") + theme_bw() + scale_fill_ptol()
```



```
## Another solution with bar plot:

summary_dt <- dt[, .(mean = mean(pro_uptake),
                     sd = sd(pro_uptake)),
                 by = "mutants"]
x_order <- summary_dt[order(mean), mutants]
summary_dt[, mutants := factor(mutants, levels = x_order)]
dt[, mutants := factor(mutants, levels = x_order)]
# get wt mean
wt <- summary_dt[mutants == "WT", mean]
# group mutants to larger and smaller than wt
summary_dt[, color := ifelse(mean > wt, "Larger",
                             ifelse(mean == wt, "WT", "smaller"))]

ggplot(summary_dt) +
  geom_bar(aes(mutants, mean, fill = color), stat='identity') +
  geom_errorbar(aes(mutants, ymax=mean+sd, ymin=mean-sd), width = 0.2) +
  geom_jitter(data = dt, aes(mutants, pro_uptake)) + theme_bw() + scale_fill_ptol()
```



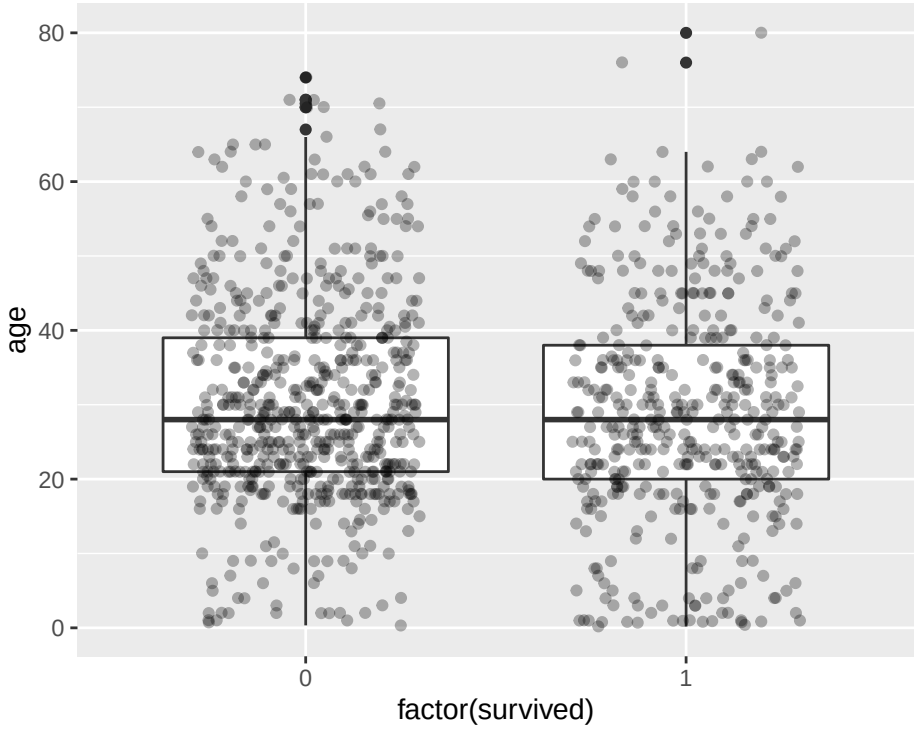
1.1 Data analysis

Read the `titanic.csv` file. You can read description of the dataset from kaggle. Did age play a role in determining survival? Visualize this with a boxplot.

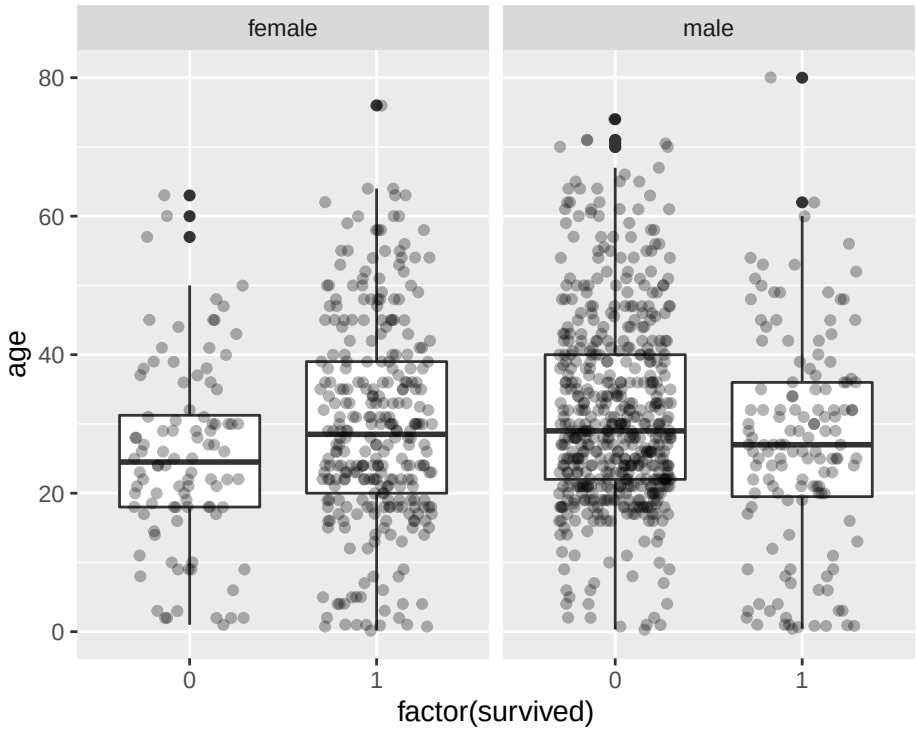
Now use facets to visualize whether age and gender combined were factors in survival. Do the same for age and passenger class.

Finally, visualize the interaction of age, gender and passenger class in determining survival.

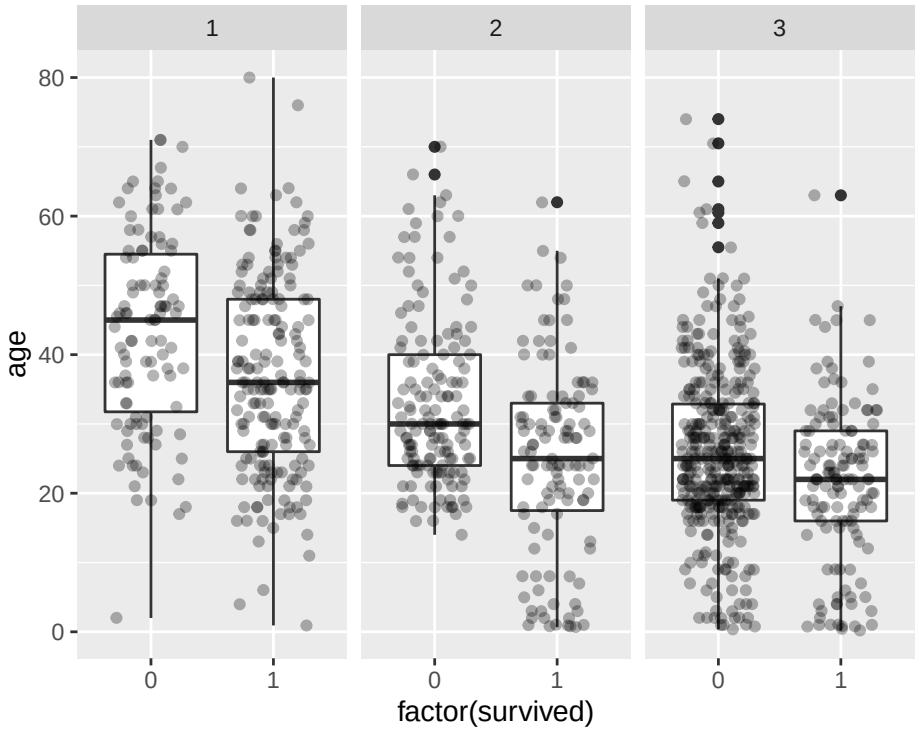
```
## Load data
titanic <- fread("extdata/titanic.csv")
titanic
##       pclass survived                                             name    sex
##    1:      1        1                    Allen, Miss. Elisabeth Walton female
##    2:      1        1                   Allison, Master. Hudson Trevor   male
##    3:      1        0                     Allison, Miss. Helen Loraine female
##    4:      1        0             Allison, Mr. Hudson Joshua Creighton   male
##    5:      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
##   ---
## 1305:      3        0                            Zabour, Miss. Hileni female
## 1306:      3        0                           Zabour, Miss. Thamine female
## 1307:      3        0                       Zakarian, Mr. Mapriededer   male
## 1308:      3        0                             Zakarian, Mr. Ortin   male
## 1309:      3        0                              Zimmerman, Mr. Leo   male
##         age sibsp parch ticket     fare   cabin embarked boat body
##    1: 29.00     0     0  24160 211.3375      B5        S    2   NA
##    2:  0.92     1     2 113781 151.5500 C22 C26        S   11   NA
##    3:  2.00     1     2 113781 151.5500 C22 C26        S        NA
##    4: 30.00     1     2 113781 151.5500 C22 C26        S       135
##    5: 25.00     1     2 113781 151.5500 C22 C26        S        NA
##   ---
## 1305: 14.50     1     0   2665  14.4542               C       328
## 1306:    NA     1     0   2665  14.4542               C        NA
## 1307: 26.50     0     0   2656   7.2250               C       304
## 1308: 27.00     0     0   2670   7.2250               C        NA
## 1309: 29.00     0     0 315082   7.8750               S        NA
##                             home.dest
##    1:                     St Louis, MO
##    2: Montreal, PQ / Chesterville, ON
##    3: Montreal, PQ / Chesterville, ON
##    4: Montreal, PQ / Chesterville, ON
##    5: Montreal, PQ / Chesterville, ON
##   ---
## 1305:
## 1306:
## 1307:
## 1308:
## 1309:

## Did age play a role?
ggplot(titanic, aes(factor(survived), age)) +
  geom_boxplot() +
  geom_jitter(width = 0.3, alpha = .3)
```



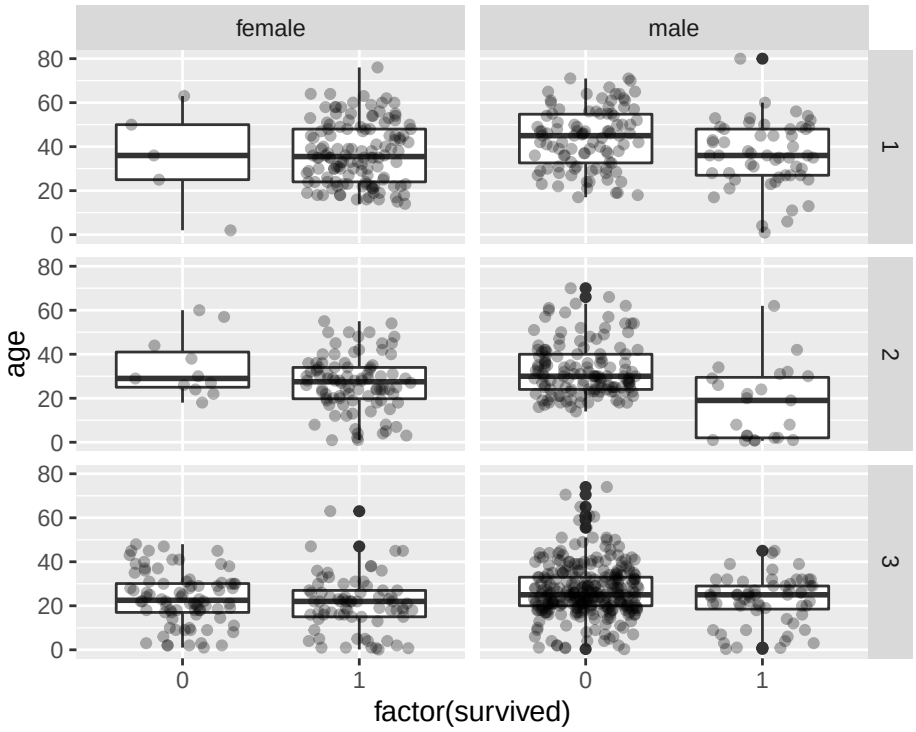

```
## Interaction between age and gender
ggplot(titanic, aes(factor(survived), age)) +
  geom_boxplot() +
  geom_jitter(width = 0.3, alpha = .3) +
  facet_wrap(~ sex)
```




```
## Interaction between age and class
ggplot(titanic, aes(factor(survived), age)) +
  geom_boxplot() +
  geom_jitter(width = 0.3, alpha = .3) +
  facet_wrap(~ pclass)
```




```
## Interaction between age, gender and class
ggplot(titanic, aes(factor(survived), age)) +
  geom_boxplot() +
  geom_jitter(width = 0.3, alpha = .3) +
  facet_grid(pclass~ sex)
```


female
male
1
2
3
age
0
20
40
60
80
0
1
factor(survived)