CeDoSIA SS2020 - Exercise Sheet 3: Simple Data Manipulation & Visualization I

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

29 June, 2020

Package

BiocStyle 2.14.4

Contents

1 Setup

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

2 Choosing the appropriate visualization method

Match each chart type with the relationship it shows best.

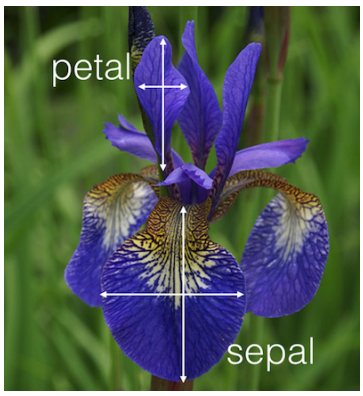
1. shows distribution and quantiles, especially useful when comparing distributions.
2. highlights individual values, supports comparison, can show rankings or deviations categories and totals
3. shows overall changes and patterns, usually over time
4. shows relationship between two quantitative variables.

Options: bar chart, line chart, scatterplot, boxplot

```
# 1. boxplot
# 2. bar chart
# 3. line chart
# 4. scatterplot
```

3 Getting to know your dataset

`Iris` is a classical and widely used dataset in machine learning literature. It was first introduced by R.A. Fisher in his 1936 paper. Load the *iris* data into your R environment. What is the dimension of the dataset? What kind of data type does each column has? How many Species does it contain?




```
# Solution
dim(iris)
## [1] 150   5
```

```
head(iris)
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
sapply(iris, class)
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
##    "numeric"    "numeric"    "numeric"    "numeric"     "factor"
## iris %>% as.data.table %>% .[, .N, by=Species]
table(iris$Species)
##
##     setosa versicolor  virginica
##         50         50         50
```

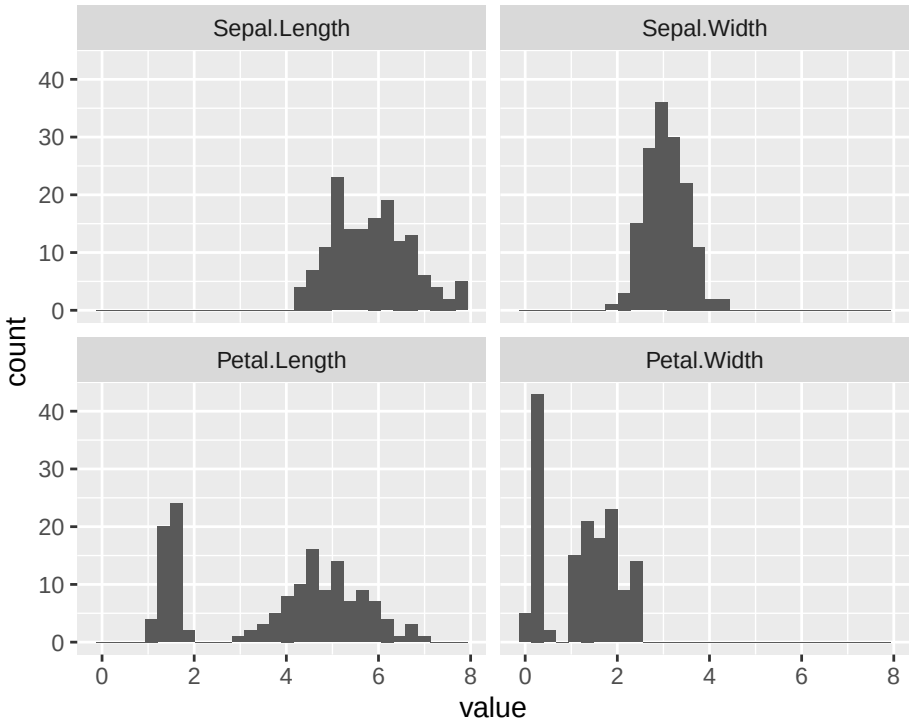
4 Data exploration

How are the lengths and widths of sepals and petals distributed? How would you visualize them? `Hint: tidy the data set and facet_wrap()`.

```
# Solution
iris_melt <- melt(iris, id.var=c("Species"))

head(iris_melt)
##   Species     variable value
## 1  setosa Sepal.Length   5.1
## 2  setosa Sepal.Length   4.9
## 3  setosa Sepal.Length   4.7
## 4  setosa Sepal.Length   4.6
## 5  setosa Sepal.Length   5.0
## 6  setosa Sepal.Length   5.4

ggplot(data = iris_melt, aes(x = value)) +
  geom_histogram() +
  facet_wrap( ~ variable)
```



5 Misleading plots

1) Visualize the lengths and widths of the sepals and petals from the iris data with boxplots.
2) Add jitter (`geom_jitter()`) to visualize all points. Discuss: in this case, why is it not good to visualize the data with boxplots?
3) Alternatives to boxplot are violin plots (`geom_violin()`) and beanplots (`geom_beeswarm()` from library `ggbeeswarm`. Install it with `install.packages("ggbeeswarm")`). Apply both options to the same data.
4) Which pattern shows up when moving from boxplot to violin/bean plot? Give possible explanations for it. How could you prove your theories graphically?

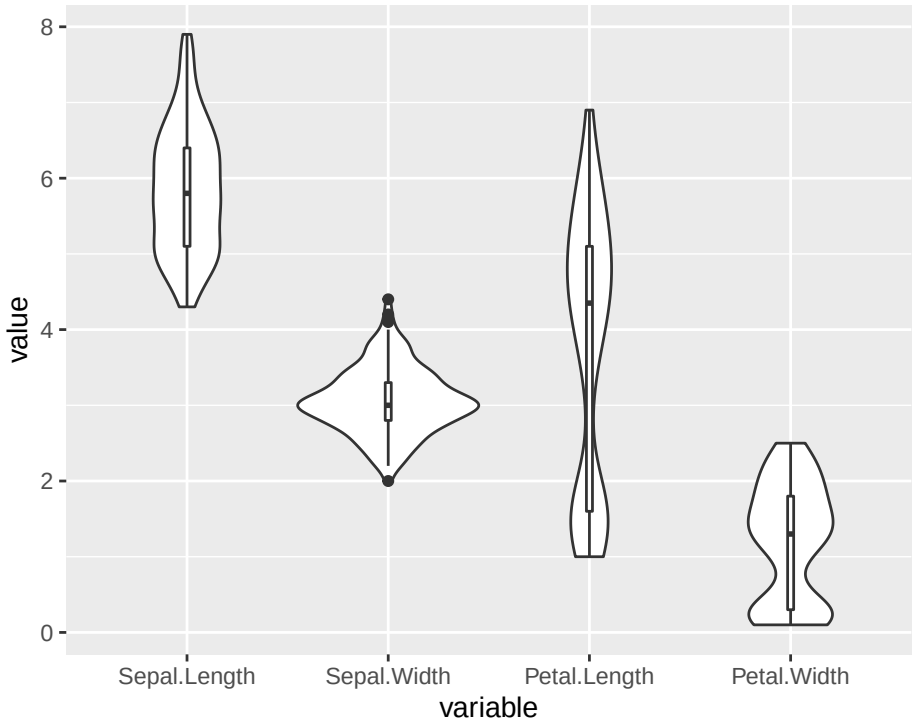
```
# Solution
library(ggbeeswarm)

# 1)
ggplot(iris_melt, aes(variable, value)) +
  geom_boxplot()
```

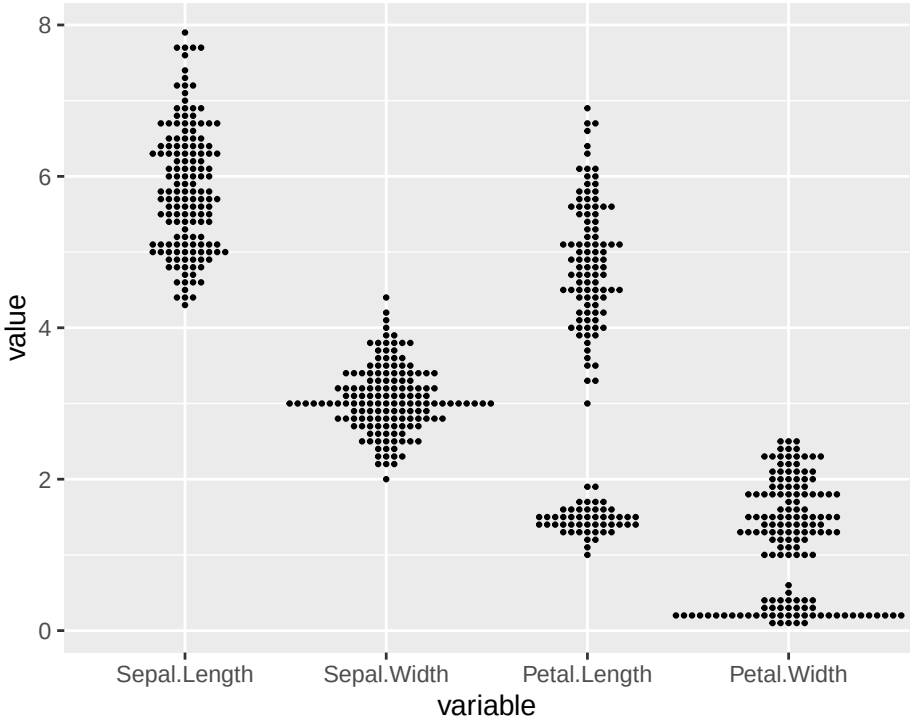

```
# 2)
ggplot(iris_melt, aes(variable, value)) +
  geom_boxplot() +
  geom_jitter(width = 0.3, size = .5)
```

```
# petal distributions are bimodal, boxplot cannot visualize this property.

# 3)
ggplot(iris_melt, aes(variable, value)) +
  geom_violin() +
  geom_boxplot(width=0.03) # Overlay boxplot to visualize median can interquartile range.
```




```
ggplot(iris_melt, aes(variable, value)) +
  geom_beeswarm(size=0.5)
```




```
# 4) The difference in the measurements might be due to the Species
ggplot(iris_melt, aes(variable, value, color = Species)) +
  geom_beeswarm(size=0.5)
```