# Data Analysis and Visualization in R (IN2339)

Exercise Session 4 - Low dimensional visualization

Daniela Klaproth-Andrade, Jun Cheng, Daniel Bader, Julien Gagneur

## Quizzes (from the lecture)

The following quizzes will be solved orally by the students and the professor during the lecture.

1. When do we use a line plot for visualizing data?

   a. To show a connection between a series of individual data points
   b. To show a correlation between two quantitative variables
   c. To highlight individual quantitative values per category
   d. To compare distributions of quantitative values across categories

2. What's the result of the following command?

```
ggplot(data = mpg)
```

   a. Nothing happens
   b. A blank figure will be produced
   c. A blank figure with axes will be produced
   d. All data in `mpg` will be visualized

3. What's the result of the following command?

```
ggplot(data = mpg, aes(x = hwy, y = cty))
```

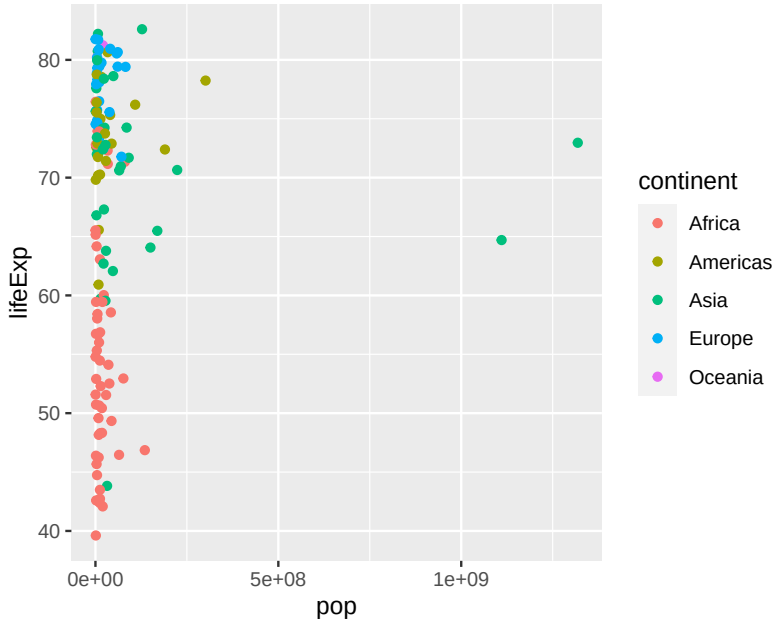
   a. Nothing happens
   b. A blank figure will be produced
   c. A blank figure with axes will be produced
   d. All data in `mpg` will be visualized

4. For which type of data will boxplots produce meaningful visualizations? (2 possible answers)

   a. For discrete data.
   b. For bi-modal distributions.
   c. For non-Gaussian, symmetric data.
   d. For exponentially distributed data.

5. Observe the following plot and select the correct answer.

```
library(gapminder)
library(ggplot2)
library(data.table)
gm_dt <- data.table(gapminder)[year == 2007]
ggplot(gm_dt, aes(pop, lifeExp, color=continent)) +
  geom_point()
```




a. The coloration of the points reduces legibility.
b. The scaling of the x-axis makes the plot difficult to interpret.
c. Life expectancy is lower in Europe than in Asia.
d. Larger populations always cause lower life expectancy.

# Tutorial

The following exercises will be solved during the tutorial sessions.

## Section 00 - Getting ready

1. Make sure you have already installed and loaded the following libraries:

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

## Section 01 - Plot types

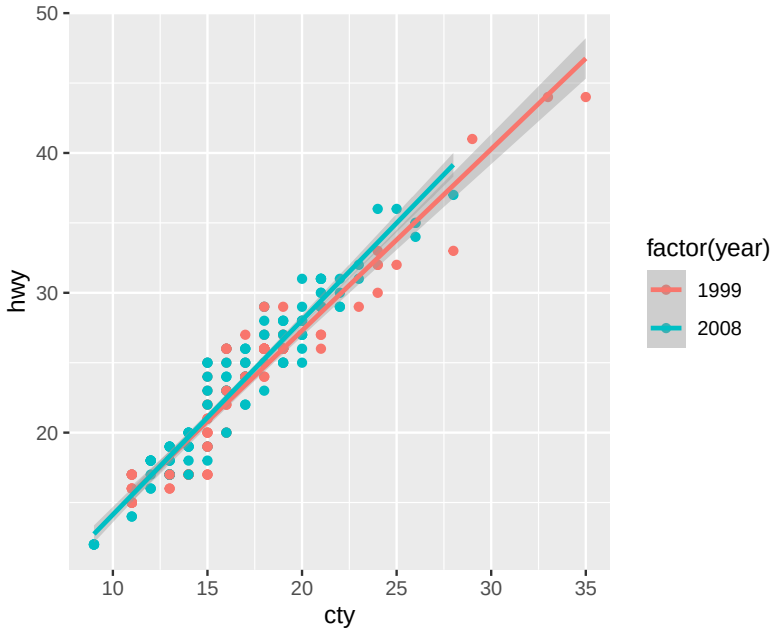
Match each chart type with the relationship it shows best.

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

Options: bar chart, line chart, scatterplot, boxplot

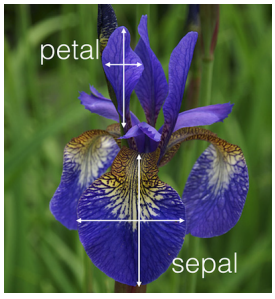
## Section 02 - Reproducing a plot with ggplot2

1. Reproduce the following visualization of the association between the variables `cty` and `hwy` for the years 1999 and 2008 from the dataset `mpg` using the library `ggplot2`:




## Section 03 - Visualizing distributions

`Iris` is a classical dataset in machine learning literature. It was first introduced by R.A. Fisher in his 1936 paper. The dataset gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris.



1. Load the *iris* data and transform it to a `data.table`. Have a look at its first and last rows.

2. How are the lengths and widths of sepals and petals distributed? Make one plot of the distributions with multiple facets. *Hint:* You will need to reshape your data so that the different measurements (petal length, sepal length, etc.) are in one column and the values in another. Remember which is the best plot for visualizing distributions.

3. Vary the number of bins in the created histogram. Describe what you see.

4. Visualize the lengths and widths of the sepals and petals from the iris data with boxplots.

5. Add individual data points as dots on the boxplots to visualize all points. Discuss: in this case, why is it not good to visualize the data with boxplots? *Hint:* `geom_jitter()`

6. Alternatives to boxplot are violin plots (`geom_violin()`). Try combining a boxplot with a violinplot to show the the lengths and widths of the sepals and petals from the iris data.

7. Which pattern shows up when moving from boxplot to a violin plot? Investigate the dataset to explain this kind of pattern, provide with visualization.

### Section 04 - Visualizing relationships between continuous variables

1. Are there any relationships/correlations between petal length and width? How would you visually show it?

2. Do petal lengths and widths correlate in every species? Show this with a plot.

## Homework

Please solve the exercises below at home. The solutions will be discussed in the central exercise.

### Section 05 - Axes scaling and text labeling

1. Load the medals dataset stored in the file `medals.csv`. Plot the total number of medals won against population size in the 2016 Rio Olympics with a scatter plot. You can load the dataset with the following code:

```
medals_dt <- fread('extdata/medals.csv')
```

2. What are the problems with the previous plot? Solve these issues with an adapted version of the plot.

3. Add the country labels to the points in the scatter plot. Compare the differences of using the library `ggplot2` and the library `ggrepel` for this task

### Section 06 - The importance of data visualization

Anscombe's quartet was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it, and the effect of outliers on statistical properties. `anscombe` is directly built in R. You do not need to download it.

1. We reshaped the original `anscombe` data to `anscombe_reshaped`. Which one is tidier?

```r
anscombe_reshaped <- anscombe %>%
  as.data.table %>%
  .[, ID := seq(nrow(.))] %>%
  melt(id.var=c("ID")) %>%
  separate(variable, c('xy', "group"), sep=1) %>%
  as.data.table %>%
  dcast(... ~ xy) %>%
  .[, group := paste0("dataset_", group)]
```

2. Compute the mean and standard deviation of each variable for each group. What do you see?

3. For each dataset, what is the Pearson correlation between x and y? *Hint:* `cor()` and Wikipedia[1] for Pearson correlation.

4. Only by computing statistics, we could conclude that all 4 datasets have the same data. Now, plot x and y for each dataset and discuss.

5. Consider now the datasets given in the file `boxplots.csv`. Load the data and visualize the different datasets with a boxplot. What do you observe? What can you conclude?

6. Exchange the boxplots by violin plots in the previous exercise. What do you observe? Do you conclude the same as you did when visualizing the datasets with the boxplots?

## Section 07 - Understanding and recreating boxplots

1. Using the `mtcars` dataset, make a boxplot of the miles per gallon (mpg) per cylinder (cyl).

2. Now, recreate the same plot without using `geom_boxplot`. You have to add all the layers manually: IQR box, median line, whiskers and outlier points. *Hint*: Remember how a boxplot is constructed[2]. You may find these functions useful: `IQR`, `geom_crossbar`, `geom_segment`, `geom_point`. Use `data.table` commands.

[1] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
[2] http://docs.ggplot2.org/current/geom_boxplot.html