Data Analysis and Visualization Exercise 5

Jun Cheng, Daniel Bader, Julien Gagneur, Matthias Heinig, Jan Krumsiek, Vicente Yépez

20 November 2019

Setup

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

Questions

Question 1:

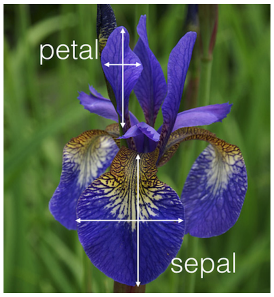
Match each chart type with the relationship it shows best.

1. shows distribution and quantiles, especially useful when comparing distributions.
2. highlights individual values, supports comparison 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 continues variables.

Options: bar chart, line chart, scatterplot, boxplot

Question 2:

`Iris` is a classical dataset in machine learning literature. It was first introduced by R.A. Fisher in his 1936 paper. Load the *iris* data and transform it to a `data.table`. How are the lengths and widths of sepals and petals distributed? Make one plot with multiple facets. 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. `Hint: remember which is the best plot for visualizing distributions; facet_wrap(~variable)`.



Question 3

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

Question 4:

1) Visualize the lengths and widths of the sepals and petals from the iris data with boxplots.
2) Add jitter (`geom_jitter()`: overlays individual data points as dots on the boxplot) 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 plot (`geom_violin()`) and beanplots (`geom_beeswarm` from the `ggbeeswarm` package. Install it with `install.packages("ggbeeswarm")`). Apply both approaches to the same data.
4) Which pattern shows up when moving from boxplot to violin/bean plot? Investigate the dataset to explain this kind of pattern, provide with visualization.

Question 5:

1) Are there any relationships/correlations between petal length and width? How would you show it?
2) Change your plot title and axis labels, for instance, to “Relationship between petal length and width”, “Petal Length” and “Petal Width”, respectively.
3) Do petal lengths and widths correlate in every species?

Question 6: Anscombe’s dataset of four x-y pairs with 11 values.

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 don’t need to load it.

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

```
anscombe_reshaped <- anscombe %>%
  as.data.table %>%
  .[, ID := seq(nrow(.))] %>%
  melt(id.var=c("ID")) %>%
  separate(variable, c('xy', 'group'), sep=1) %>%
  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 correlation between x and y?
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.

Question 7:

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

Question 8:

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 (lecture, http://docs.ggplot2.org/current/geom_boxplot.html). You may find these functions useful: `IQR`, `geom_crossbar`, `geom_segment`, `geom_point`. Use `data.table` commands.