Advanced graphics: Exercises

Erasmus Q-Intelligence B.V.

Exercise 1. Load the `patents` data from the provided `patents.Rds` file. The data set contains information on patents granted in 2012 in each of the 50 US federal states and the District of Columbia. See the appendix for a description of the variables.

(a) Produce a density plot of `logtotal` with separate density estimates for observations corresponding to different population density categories (variable `densitycat`). Use different colors for the different population density categories.

(b) Continue with the plot from (a). Fill the area under the densities with the color black and play with different values of transparancy (`alpha`) to obtain a visually appealing graph.

(c) Produce a scatterplot of `logtotal` vs `logdensity`. Use plot symbol 21 and color the points colored according to party affiliation of the governor (variable `governor`). Also play with the size of the points until you obtain a visually appealing graph.

(d) Produce the same scatterplot as in (c), but with plot symbol 16.

(e) Produce conditional boxplots of `logtotal` with observations grouped by population density category (variable `densitycat`). Color the boxes with different colors. Do you think that the displayed legend provides useful information in this plot?

(f) *Bonus*: Continue with the plot from (e). Let the box widths reflect the different group sizes. Use the help system to find out how to make this change.

Exercise 2. Load the `msleep` data from package `ggplot2`. Dataset on mammalian sleep.

(a) Produce a scatterplot of `sleep_total` vs `sleep_rem`. Use plot symbol 15 and color the observations according to `vore` with the default color scale. Set the size of the points to a suitable value.

(b) Produce the same scatterplot as in (a), but using a manual color scale with four colors of your choice. What do you observe when you compare the two plots? Which plot is more informative?

(c) Make a printer-friendly version of the plot from (b).

(d) Categorize the variable `sleep_total` using the following code:
library(`dplyr`)
breaks $<-$ seq(0, 20, by $=$ 5)
msleep $<-$ mutate(msleep, sleep_cut $=$ cut(sleep_total, breaks))

(e) Use the new variable from (d) to produce a stacked barplot by putting `vore` on the x-axis and specifying `sleep_cut` as fill color. Make the bars horizontal rather than vertical.

(f) Continue with the plot from (e). Change the axis labels to "Type" and "Frequency" and manually change the colors.

(g) *Bonus*: Produce a similar barplot as in (f), but put the bars for different `vore` side-by-side instead of stacking them. Use the help system to find out how to do this (*hint*: check the examples at the bottom of the help file of `geom_bar()`).

Exercise 3. Consider again the `patents` data.

(a) Produce a boxplot of `logdensity` to verify that there are two potential outliers. Remove the x-axis label and change the y-axis label to "Logarithm of population density".

(b) The cut-off values for outliers according to the boxplot from (a) are 0.5242581 and 6.8220640. Produce a scatterplot of `logtotal` vs `logdensity`, and add grey reference lines for the cut-off values.

(c) Refine the plot from (b). Change the axis labels to "Logarithm of total number of patents" and "Logarithm of population density", respectively. Furthermore, expand the axis limits (use suitable values of your choice).

(d) Produce a scatterplot of `logtotal` vs `logdensity` and add a linear scatterplot smoother.

(e) Change the color of the scatterplot smoother to a color of your choice.

Exercise 4. Consider the `vwgolf.Rds`-dataset from Canvas. It contains information about occasions for sale. Use `ggplot2`-functions where possible.

(a) Load the data.

(b) Produce a scatterplot with original price on the x-axis and asking price on the y-axis.

(c) Refine the plot from (a) by letting the color of the points depend on mileage and the size of the points on top speed.

(d) Change the colors in the plot from (b) such that the continuous color scale ranges from white for cars with a low mileage to red for cars with a high mileage.

(e) Further refine the plot from (c) by changing the axis labels to "New price (euro)" and "Asking price (euro)", respectively.

(f) Add a plot title and subtitle to the plot from (d). Check the help file of function `labs()` to find out how to do this. Use "Asking Price versus New Price" as plot title, and "56 second-hand VW Golfs from Marktplaats.nl" as subtitle.

(g) Look at slide 18 from Introduction. We told you you could do this after the second lecture!