Exploratory Data Analysis and Visualisation 11374/11517

Week 4 Lab Exercises

1. The “Cars93” dataset available in the *MASS* package contains characteristics of 93 cars that were on sale in the US in 1993.

   a. Read the help file about the “Cars93” dataset (?Cars93)
   b. Suppose you are curious about the relationship between fuel efficiency and horsepower. You suspect that as the horsepower of the car increases, the fuel efficiency decreases. You want to verify this with a plot. Create such a plot with informative axis labels. Interpret the plot. Does this relationship change by vehicle type (perhaps use the filter() function)?
   c. Add a smoothed trend line as well as a smoothed trend line. Which one suits the visualisation better and why?
   d. What is the most common car manufacturer in this dataset? Create a plot which shows the frequency of the top 10 manufacturers. Hint: wrangling the dataset to create a tibble of the top 10 does not require reordering in ggplot.
   e. Create a bivariate barplot which shows the median midrange price of each type of vehicle. What does this plot reveal? Should we be cautious of anything? How could we resolve these cautions (how could we create a more detailed plot)?
   f. Which manufacturer is the most fuel efficient? Take fuel efficiency to be the average of the “MPG.city” and “MPG.highway” variables. Create a bivariate boxplot for the five most fuel-efficient manufacturers.
   g. Determine the relationship between the number of passenger seats and a variable of interest, such as price. Is there anything misleading or strange about the output? What function could we use to improve it?

2. Consider the “crabs” dataset found in the *MASS* package. It contains information on crabs. Wow!

   a. Read the help file on the “crabs” dataset (?crabs)
   b. Create an extended scatter plot between FL and BD. Creating meaningful titles and axis labels!
   c. Now add a geom_jitter() layer, setting the width to 2. Does this help the interpretability of the plot heavily? Is it required to view the actual relationship? Why or why not?
   d. Add a linear trend line. Are the two variables linearly correlated? Positive or negative? Does the variation change in BD depending on FL?
   e. There is a function which can also help overplotting on a scatter plot when so many points are on a graph you cannot tell which areas have more points and which have less. Which function could help you solve such an issue?
   f. Apply the function to a plot with the two variables CL and CW. Explain any notable observations and based off the help file, try to explain the trends.