Exercise 3A: Data and tidyverse

Niels Richard Hansen

December 8, 2016

The *Data Wrangling Cheat Sheet* and the *Data Visualization Cheat Sheet* may be helpful.

Data set on supermarket sales

The data for this and several subsequent exercises is available from

http://nielsrhansen.github.io/Dong/Supermarket.txt

- Read the data set into R as a tibble called `supermarket`. Inspect the data set in the Data Viewer.

The data set contains sales for a major supermarket chain in Sweden from the first 12 weeks of 2007. Most variable names are self-explanatory. The four marketing variables are indicators of whether the different marketing strategies have been used. The variable `NormalSale` is the expected (seasonally adjusted) number of sold items. The variable `TotalSale` is the actual number of items sold.

Cleaning and filtering data

- How many observations have a missing `DiscountSEK`?
- Discard observations with a price less than 1 SEK and replace negative and missing values of `DiscountSEK` by 0 for the remaining observations.
- Construct a new variable `normalPrice` containing the price without discount.
- Carry on with further inspections of the data.

Summaries

- Cross-tabulate the four marketing variables.
- Find the 10 items with the largest number of observations.
- Find the 10 items with the largest total sale.

Visualization

- Plot the total sale against the normal sale for the top 10 most sold items. Use the color aesthetic to distinguish between the items. Change both axes to be on a log-scale. Add a line with slope 1 and intercept 0 using `geom_abline`. Give an interpretation.
- The relative total sale is `TotalSale / NormalSale`. Plot, using hexagonal binning, the relative total sale against the discount in percent. It's probably still a good idea to have the y-axis on a log-scale. Add a smoother (in red) to the plot.