ACSC/STAT 3740, Predictive Analytics

WINTER 2023

Toby Kenney

Homework Sheet 1

Due: Thursday 26th January: 11:30

Note: This homework assignment is only valid for WINTER 2023. If you find this homework in a different term, please contact me to find the correct homework sheet.

[Note: all data in this homework are simulated.]

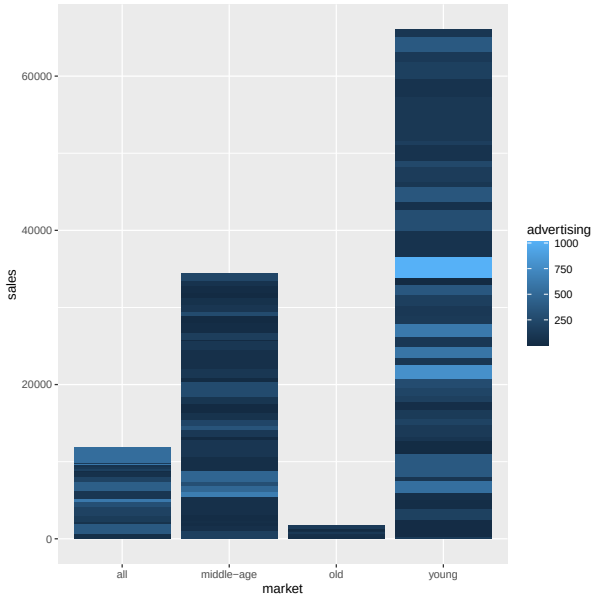
Basic Questions

1. The file `HW1Q1.txt` is from an experiment about the effect of acid rain on forestry. It includes the following variables:

Variable name	Meaning
land.area	The area of the plot of land in square km
yearly.rain	The total annual rainfall in mm
soil.type	The predominant type of soil
winter.temp	The average daytime maximum temperature during winter months
summer.temp	The average daytime maximum temperature during summer months
water.pH	The pH of the groundwater in the region
production	Annual lumber production in tonnes

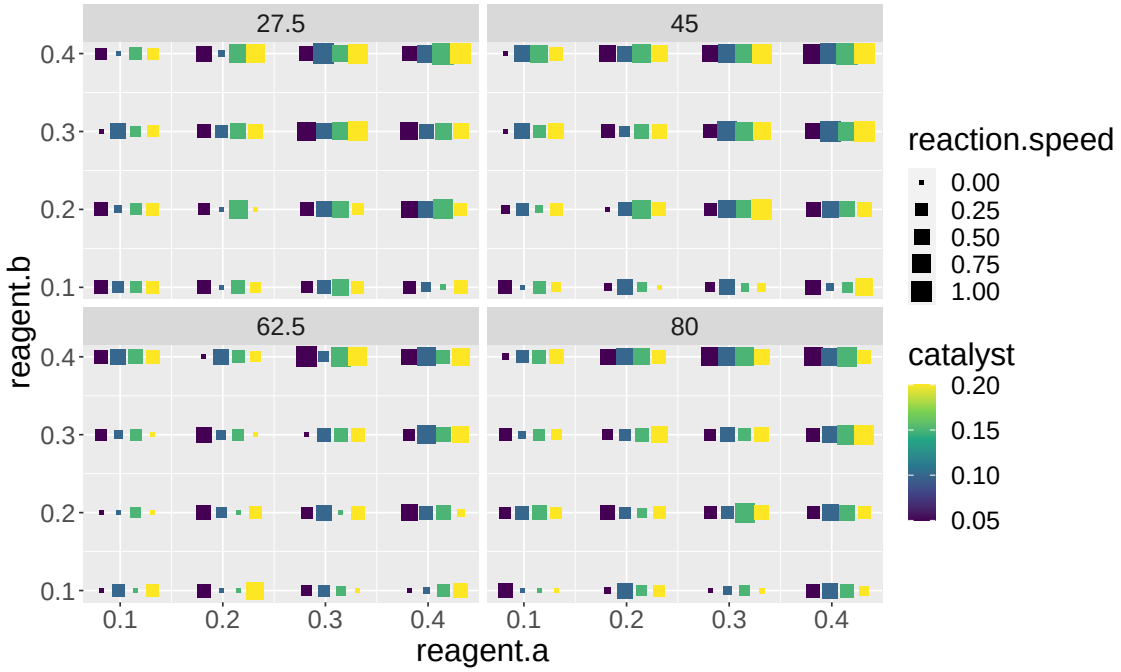
   Display this data set in a plot.

2. A retail company is studying the effect of advertising on sales, and has produced the plot below. Identify at least three issues with the plot and produce a new plot that better displays the data.



Provide `R` code for the new plot. [The data used to produce the figure is in the file `HW1Q2.txt`. You should include more information from that file in the plot as appropriate.]

3. Use `ggplot` to produce the following plot from the data in file `HW1Q3.txt`. [Make sure to reproduce all aspects of the plot — axis scales, labels, etc.]



4. The file `HW1Q4.txt` contains the following data from an insurance company about worker's compensation insurance claims.

Variable	Meaning
company.size	The number of workers at the company
industry	The industry of the company
years.history	The number of years of history with this company
previous.claims	The per-worker average previous annual claims from the company
safety.protocols	An index indicating how many safety protocols the company adheres to
aggregate.claims	The company's aggregate average claims per worker.

Construct a plot or plots to show these data for the purpose of data exploration.

Standard Questions

5. A bank collects the following data on loan repayments by customers. The data are contained in the file `HW1Q5.txt` and include the following variables:

Variable	Meaning
loan.amount	The size of the loan
yearly.income	The borrower's annual income
credit.score	The borrower's credit score
age	The borrower's age
employment.status	The borrower's employment status
default	Whether the loan is repaid

Make a plot to show these data.

6. The file `HW1Q6.txt` contains data on the effect of pollution on cancer incidence.

Variable name	Meaning
size	The population of the town or city
over.60	The percentage of the population that are over 60.
poverty	The percentage of the population with income below the poverty level
temperature	The average temperature ($^\circ C$) in the settlement.
rainfall	The average annual rainfall (mm).
pollutant.a	The average levels of pollutant a in the air (ppm)
pollutant.b	The average levels of pollutant b in the air (ppm)
cancer.incidence	The average annual number of new cancer diagnoses per 1000 residents.

(a) Produce a figure to show these data for the purpose of data exploration.

(b) After analysing the data, you conclude that for fixed values of the other parameters, `pollutant.b` is positively associated with `cancer.incidence`, while for fixed values of the other parameters, `pollutant.a` is positively associated with `cancer.incidence` if `rainfall` is small. Make a plot which makes these conclusions more obvious.