ACSC/STAT 3740, Predictive Analytics

WINTER 2024
Toby Kenney

Homework Sheet 1

Due: Wednesday 24th January: 11:30

Note: This homework assignment is only valid for WINTER 2024. 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

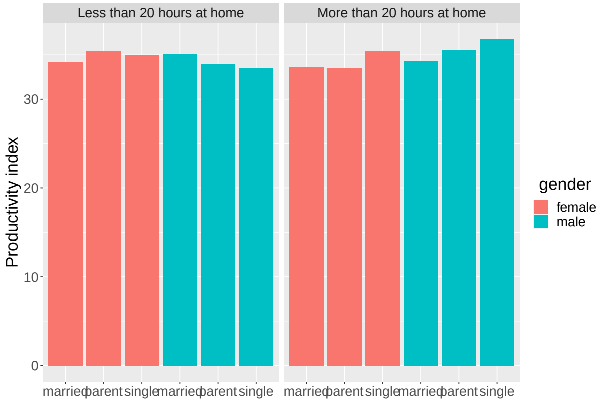
1. The file `HW1Q1.txt` contains data from a university about student research projects. The data are not formatted in a very convenient way. Read the data into `R` and reformat into a more convenient way, and use it to create a plot showing funding (y-axis) vs GPA (x-axis) with colour showing student age and size showing professor age, with a facet grid of professor subject versus student subject. Make a list of all corrections made to the data.

2. The file `HW1Q1.txt` is from an experiment about the effect of alcohol consumption on depression in young adults. It includes the following variables:

Variable name	Meaning
Age	The subject's age
Gender	The subject's gender
Employment.Status	The subject's employment status.
Weekly.work.hours	The average number of weekly hours spent working or studying.
Average.daily.sleep	The average number of hours spent asleep each day.
Weekly.alcohol.consumption	The subject's weekly alcohol consumption
Depression.score	An index indicating how many symptoms of depression the subject has.

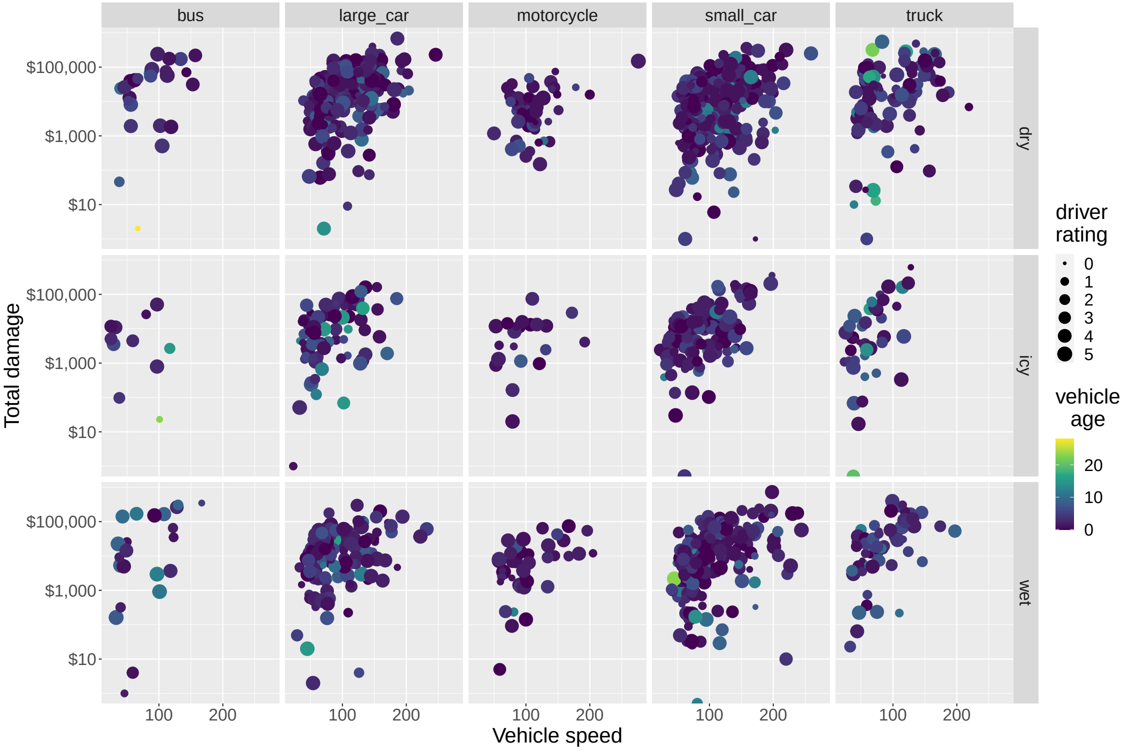
   Display this data set in a plot.

3. A company is studying the effect of working from home on productivity, 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 HW1Q3.txt. You should include more information from that file in the plot as appropriate.]

4. Use ggplot to produce the following plot from the data in file HW1Q4.txt. [Make sure to reproduce all aspects of the plot — axis scales, labels, etc.]



5. The file HW1Q5.txt contains the following data from an experiment into

the effect of pollution on bacterial growth in lakes.

Variable	Meaning
nitrate.conc	Concentration of nitrates in the lake
phosphate.conc	Concentration of phosphates in the lake
pH	pH of the lake (0=strong acid, 14=strong alkali, 7=neutral)
salt	Concentration of salt in the lake
temperature	Water temperature of the lake.
weekly.rainfall	Total rainfall in the past week (mm)
cyanobacteria	Abundance of cyanobacteria in the lake
toxin.level	Concentration of toxins in the lake

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

6. A doctor collects the following data on patients. The data are contained in the file `HW1Q6.txt` and include the following variables:

Variable	Meaning
age	The patient's age
sex	The patient's sex
weekly.exercise	The number of hours per week spent exercising
daily.calorie.intake	The patient's average estimated daily number of calories
family.history	Whether the patient has a family history of heart disease
bmi	The patient's BMI
dbp	The patient's diastolic blood pressure
year.heart.attack	Whether the patient suffers a heart attack in the year from the appointment.
year.stoke	Whether the patient suffers a stroke in the year from the appointment.

Make a plot to show these data.

7. The file `HW1Q7.txt` contains data on the effect of traffic cameras on road safety.

Variable name	Meaning
population	The population of the town or city
area	The area of the town or city.
cars	The number of cars in the town or city
speed.cameras	The total number of speed cameras in the city.
bicycle.lanes	The proportion of roads with bicycle lanes.
three.year.deaths	The number of deaths in traffic accidents during the past 3 years.

(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, `speed.cameras` is negatively associated with `three.year.deaths` when `bicycle.lanes` is relatively low for the population; but when `bicycle.lanes` is high for the population `speed.cameras` may be positively associated with `three.year.deaths`. Make a plot that emphasises these conclusions.