CADS Workshop, R session, Fall 2022

Additional practice:

Question 1: Please take a look at the "salary" data set. In the wake of a huge public outcry on the topic (check, for instance, https://www.chronicle.com/article/Gender-Pay-Gap-Persists-Across/239553), XYZ University, located in Boston, is trying to monitor salary differences between male and female faculty members. Available info on each of 397 individuals: Gender (Male or Female). Rank (Assistant, Associate, (Full) Professor). 2008-09 nine-month academic salary (in US dollars). Years since Ph.D. Years of service. Discipline (A: "theoretical" departments or B: "applied" departments). Recall: Academic tenure-track: Assistant Prof. → Associate Prof. → Prof. Data available in file titled "salary.txt".

a. Please create a bar-plot showing the rank distribution separated by gender and discipline (with numbers attached at the top of the bars).

b. Please create histograms (raw and smoothed) and boxplots showing the salary distribution, separated by gender and discipline.

c. What proportion of faculty earn more than $75000?

d. What proportion of Assistant Professors earn more than $75000?

e. What proportion of those that make more than $75000 are Assistant Professors?

#---Hint: use dim(subset(...)) or equivalently, length(which(???))---#

#---Answers: 365/397=0.919, 46/67=0.687, 46/365=0.126---#

f. Do these fractions aid our discussion around potential discrimination? If not, find out similar proportions that do.

Question 2: Please open the "electricity" data file. It contains information on energy consumption by all-electric houses (roughly 2700 sq-feet) in the US MidWest. Available information: the bill amount (in dollars, including sales tax). the average temperature (averaged over thirty years). heating degree days (averaged over thirty years) cooling degree days (averaged over thirty years) the size of the family. total electricity consumption (in KW units). Do you notice any problems with the data set? If yes, fix those, and carry on with usual trend and seasonality analyses. Do these graphs confirm your intuition around outside temperature and electricity consumption?

Question 3: The file “diamondrings” has prices (in Singaporean dollars) of certain rings, along with information on their clarity and certification status. Create a scattercloud that plots the relationship between the size of these rings and their prices, detailing also, their clarity and certification. Thoughts on what you found?

Question 4: Does economic stress force entrepreneurs to switch jobs? Related data in “Estress”, 262 participants interviewed. column 2 (X): “estress” an index of economic stress (high scores ⇒ high stress). column 4 (Y): “withdraw” index quantifying how strongly do you want to switch (high scores ⇒ greater withdrawal intentions). column 3 (M): “depression” index quantifying how depressed you are (high scores ⇒ higher levels of depression). Conjecture: Economic stress (X) leads to a desire to disengage from entrepreneurial activities (Y). Also available: gender, tenure, etc.

Crete a scattercloud that could confirm of refute this suspicion.