Data Science Toolbox Question Sheet

01.2 Exploratory Data Analysis

Daniel Lawson

Block 1

Short questions

1. Describe the most common way to standardize a variable, both by formula and by its meaning. How would it be used with a test dataset?
2. Describe the difference between a quantitative variable and a categorical variable. Give an example of each. How might they be provided to data science algorithms?
3. Describe a two-way table. What is it useful for?
4. Describe and sketch a "segmented bar chart" and "heatmap". When would either be appropriate to use?
5. What are the advantages and disadvantages of a histogram vs a density plot?
6. Consider the following dataset:

attack	service
DOS	TCP
DOS	TCP
DOS	UDP
remoteAccess	UDP
remoteAccess	UDP
remoteAccess	TCP
portDetection	ICMP
portDetection	ICMP
portDetection	TCP

   a. Make a contingency table from this data.
   b. State and give a high-level description of the simplest **non-parametric** statistical test that can be applied to test whether there is a difference in the types of attack seen on different services. What limitations would this test have?
   c. Name and briefly describe, with a sketch, visualization for **categorical** data, for each of **one** dimension, **two** dimensions, and **very large** dimensions.