Data Science in Medicine: Tutorial 1

Summarising and visualising data – Part 1

Semester 1, 2020-2021

- Please attempt all questions on this worksheet in advance of the tutorial, and bring with you all work, including printouts of code and other results. Tutorials cannot function properly unless you do the work in advance.
- You are welcome to bring along any questions you may have from the lectures, textbook, etc.
- Assessment is formative, meaning that tutorials do not contribute to your final grade.
- Attendance is compulsory. If you have good reasons to miss a session, you should contact your year coordinator in advance to arrange to attend a different session.

Introduction

In this tutorial you will discuss effective data visualisation and how one can use summary statistics and visualisations to explore a dataset. You will also get the chance to reflect on the importance of the use of data in health.

Part 1: What's wrong with this picture?

The aim of good data graphics is to *display data accurately and clearly*. They also help us *tell a story*. However, it is not uncommon to see difficult to read, confusing or even misleading data visualisations used. In some cases, the choice of graph is not appropriate for the story that the researcher or company is trying to tell. In this part of the tutorial we'll try to "debug" data visualisations.

(1) What are the main issues in the pie chart provided below?

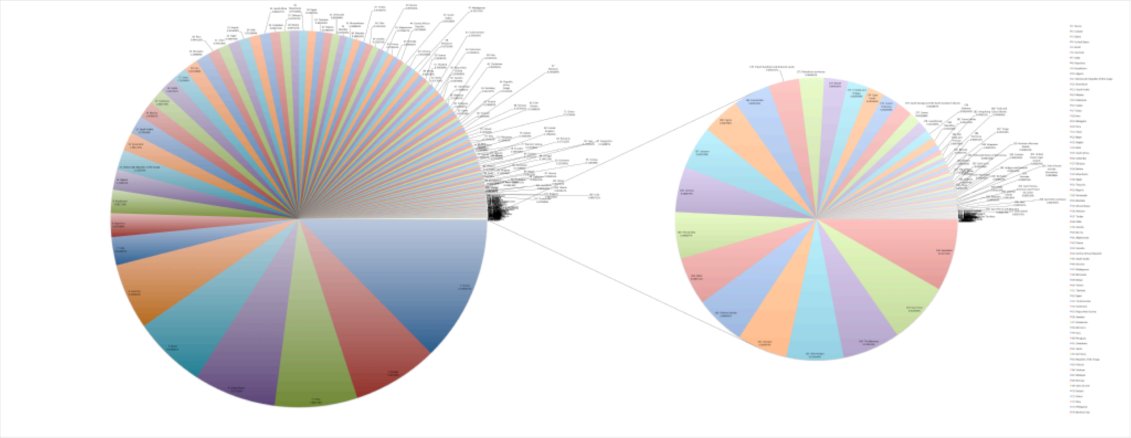

Figure 1: Pie chart of countries by area. (Attribution: Vartak.sourabh1985 at English Wikipedia)

(2) The following bar charts visualise the average time it took men and women to run 5K in a run for charity. Which of the following bar charts is more effective and why?

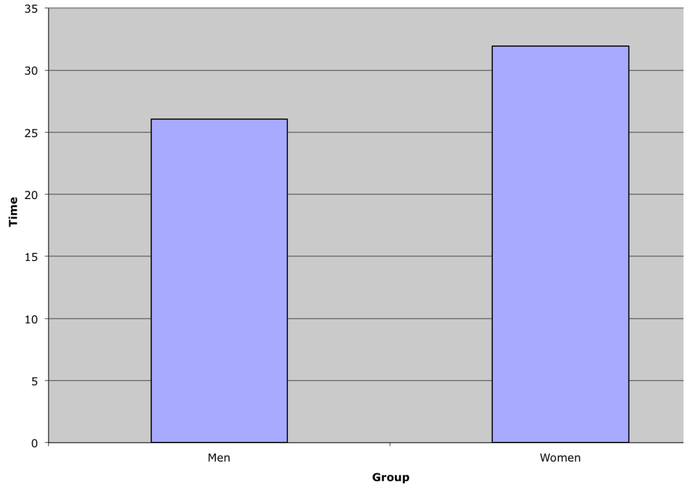

Figure 2a

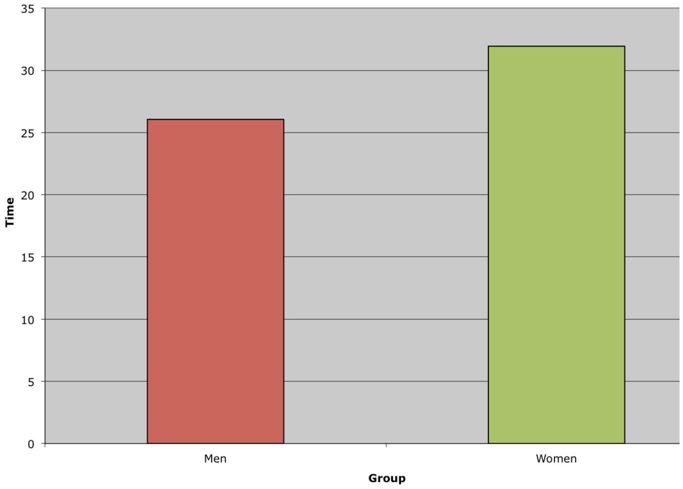

Figure 2b

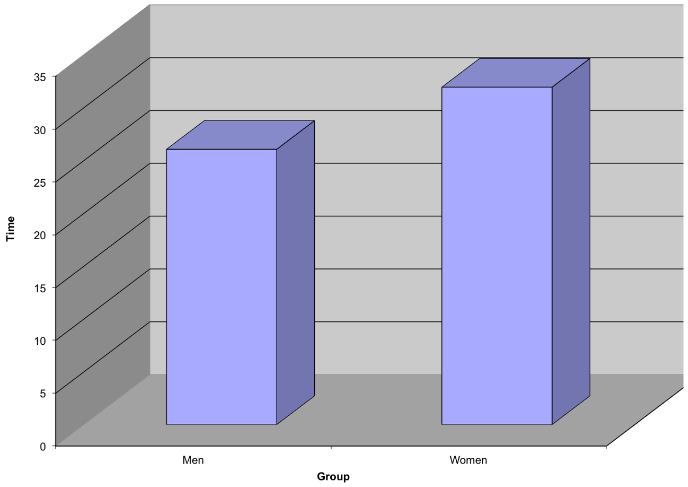

Figure 2c

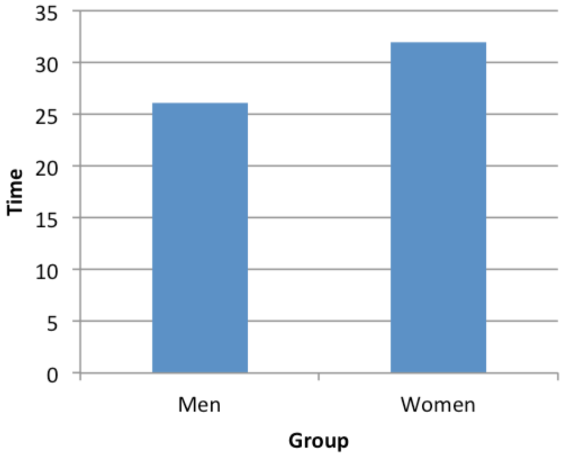

Figure 2d

(3) The following bar chart recreates a visualisation used in an ad by Chevy in the '90s to represent the percentage of trucks sold over the previous 10 years that were still on the road. What is wrong with this graph?

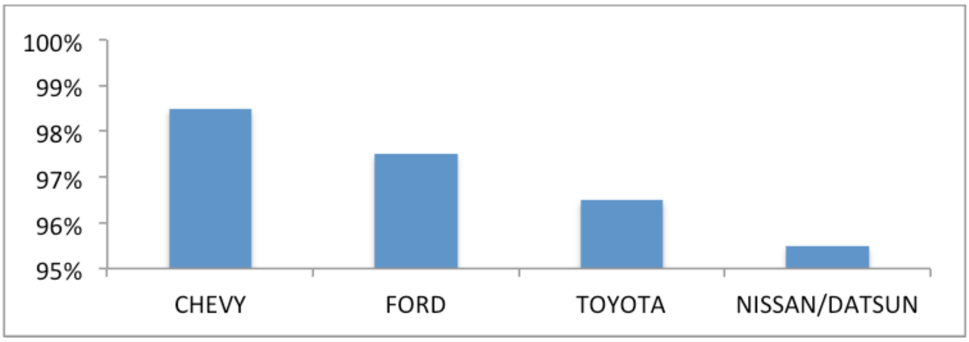

Figure 3: Recreation of a bar chart used in an ad by Chevy

(4) The following grouped bloxplots help us compare the distribution of scores per gender and group. Which of the two is more appropriate for i) focussing on the comparison between males and females for the different groups and which for ii) focussing on the comparison between the different groups per gender? What can we conclude about easing comparisons, i.e. should things to be compared be closer to each other or not?

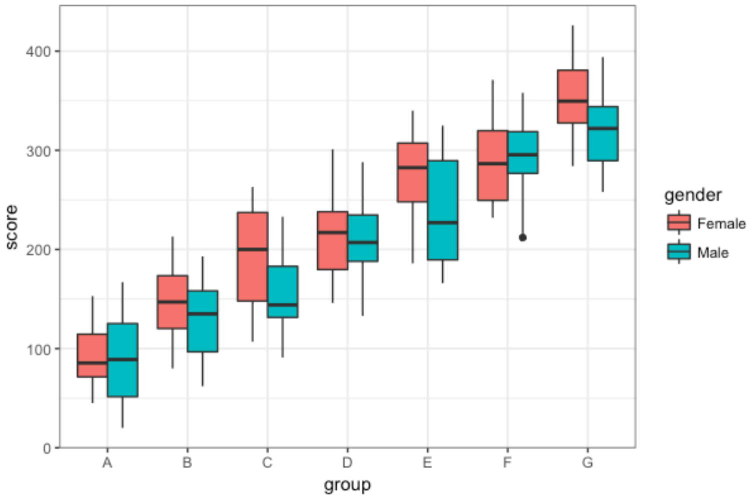

Figure 4a (Created by Holtz Yan and available at https://github.com/holtzy/R-graph-gallery)

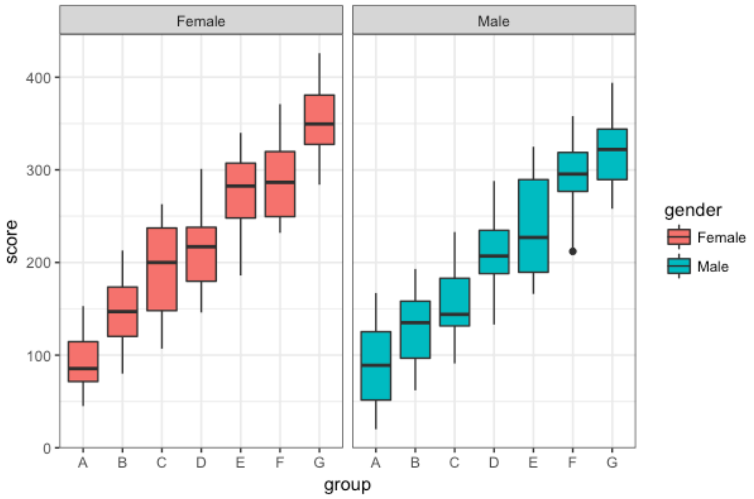

Figure 4b (Created by Holtz Yan and available at https://github.com/holtzy/R-graph-gallery)

Part 2: Exploring data through summary statistics and visualisations

Summarising and visualising data is very useful when we are given access to a dataset and we want to get a sense of what the data looks like. Data exploration is a key part of any data science project. In this part of the tutorial we'll reflect on data summarisation and visualisation in the context of data exploration, and we'll discuss outlier detection.

The (entirely fictitious) Dark Moon Hospital is recording waiting times for paediatric A&E admissions over the course of a year for audit purposes. Waiting times are calculated in minutes and manually entered in the hospital information system. The following table provides an extract of the dataset for a particular day. Note that you can obtain the same extract of the dataset through the csv file `tut1_waitTimes.csv` from the course website.

Patient case	Waiting time (min)
Case 1	34
Case 2	45
Case 3	41
Case 4	38
Case 5	47
Case 6	47
Case 7	42
Case 8	46
Case 9	35
Case 10	38
Case 11	45
Case 12	41
Case 13	49
Case 14	445
Case 15	36
Case 16	39
Case 17	42
Case 18	46

(1) Calculate the mean and median of A&E waiting time. Are the two statistics relatively close to each other or not, and why? Note that you can do the calculations by hand, using a spreadsheet application like Excel or (if you've already had your first lab) using R/RStudio.

(2) Draw a dot plot similar to Figure 6.2 of your "Learning Statistics with R" textbook to visualise the data in the table above, where the x-axis refers to the patient case and the y-axis to the waiting time. What can we learn from this plot? Is there any other type of visualisation that could convey a similar message?

Part 3: Discussion – Using data in health sciences and practice

(1) The Data Saves Lives initiative has published a number of case studies that showcase health data research worldwide. Have a look at https://datasaveslives.eu/casesummaries, choose a case study of interest, read it and keep some notes to present it in the tutorial. When presenting it, make sure to discuss what kinds of data were used, the results and the benefits to patients and healthcare systems.

(2) Health data can be used not only in research (as discussed in the previous question), but also in health practice. Can you think of different ways you could use data in your career (e.g. when working in a hospital or as a GP) to improve the health of your patients and the way care is provided? What data would you use, what data science techniques would you use and who would you need to work with to ensure the success of your project?