Report 1 Brief

MTH1004 - University of Exeter

Submission deadline: noon on Friday Dec 13, 2024 through ELE

Introduction

Your task is to summarise and analyse a data set about pet dogs. You will produce both numerical summaries and graphical visualisations, explain your results, and present your work in the form of a 1-page poster.

You may *not* collaborate with others on this coursework nor are you permitted to use any generative AI tools. You may use material presented in the module (e.g., lecture notes, R lab exercises) without attribution, but any other external resources used must be properly cited, including the references mentioned in this brief (all of which are available through the university library). For detailed information on University policies regarding academic conduct and practice, please refer to the online course on "Academic Honesty and Plagiarism" via ELE (link) and Chapter 12 of the Teaching Quality Assurance Manual (link).

This coursework brief is structured as follows: The data is discussed in the next section and is followed by instructions about how to carry out your analysis. The final two sections discuss how to structure your submission files and the marking guidelines. At the end of this document is Table 1 which describes the columns in the data set.

Data

The data in `manydogs_data.csv` is a subset of the data collected through a study in the ManyDogs Project, a collaboration across researchers from 8 countries. In this study, researchers explored "dogs' responses to human pointing" and recruited dogs through previously known contacts and social media. The citation for this data is:

ManyDogs Project, Espinosa, J., Hare, E., Alberghina, D., Valverde, B.M.P., and Stevens, J.R. (2024). Data from the ManyDogs 1. *Journal of Open Psychology Data,* 12:7, pp. 1-26. DOI: `https://doi.org/10.5334/jopd.109`.

Each row of the data set represents a dog and the columns provide information about the dog's characteristics (`sex`, `age`) along with information about where the dog lives (`environment`). The columns `cbarq_train1` through `cbarq_train8` are related to the Canine Behavioral Assessment and Research Questionnaire (C-BARQ) which is a way to measure the "temperment" along with the behaviour of dogs. All of this information was reported by the dog owner to the researchers. The citation for the C-BARQ measure is:

Hsu, Y. and Serpell, J.A. (2003). Development and validation of a questionnaire for measuring behavior and temperment traits in pet dogs. *Journal of the American Veterinary Medical Association*, 223(9), 1293-1300. DOI: `https://doi.org/10.2460/javma.2003.223.1293`.

Further information about the columns in `manydogs_data.csv` are described in Table 1, which can be found on page 5; missing data has been encoded as `NA` in the data file.

Assignment

In the following three parts of this assignment, you will analyse data about pet dogs contained in the data set, `manydogs_data.csv`. To start, download the data from the "Summative Assessments" tile on ELE and read it into R using `read_delim()` from the package `tidyverse`. Your analysis can (and should) be completed using only those functions discussed in the lecture notes and R labs.

Part 1: Introduction

1. How many dogs are in the data set?

2. Describe the study and comment on how representative the dogs in the study may be among all dogs on the planet. Provide a reference for your answer. (No coding.)

3. Describe the dogs in the study using suitable summary statistics based on the following columns: `age`, `sex`, and `owned_status`. (Drop any relevant missing data, if applicable, but report how many observations were omitted.)

Part 2: C-BARQ Scores

1. What do the letters in C-BARQ stand for? Who invented this measure and when? Provide a reference for your answer. (No coding.)

2. Create a new column in the data set called `cbarq` which is the sum of the individual C-BARQ scores: `cbarq_train1`, `cbarq_train2`, `cbarq_train3`, `cbarq_train4`, `cbarq_train5`, `cbarq_train6`, `cbarq_train7`, and `cbarq_train8`. What range of C-BARQ scores are possible? How should we interpret the C-BARQ score? For example, if dog A has a *higher* C-BARQ score than dog B, what does that mean?

3. Using `cbarq`, how many dogs in the study, if any, are missing C-BARQ scores?

4. Create a box plot of `cbarq` and save your graph as a file using `ggsave()`. Comment on the shape of the distribution of C-BARQ scores including a discussion of the outliers, if applicable.

5. Use suitable summary statistics to describe the distribution of `cbarq`, connecting your explanation with your box plot from question (4), the range of possible C-BARQ scores, and any other information you deem relevant. (Drop any relevant missing data, if applicable, but report how many observations were omitted.)

Part 3: Exploring Further

1. Is there any evidence that `cbarq` scores differ substantially by age and/or sex? Use a suitable data visualisation and sample correlation calculations to support your answer. (Drop any relevant missing data, if applicable, but report how many observations were omitted. Save your visualisation using `ggsave()`.)

Submission

Submit two files zipped together: a one-page, A4-sized pdf of your poster and your .R script through the "Report 1" ELE submission point. Do *not* include your name in the file or file names.

- **Poster Format:** Your answers to the above questions should be saved and submitted as a **one-page, A4-sized poster saved as a pdf file.** The poster can be in either landscape or portrait orientation. It must be possible to highlight and copy text on the poster when viewed in a pdf viewer (like Adobe Acrobat); that is, do **not** submit a pdf that just contains an image file of your poster. (Note: your graphs can be added to the poster as image files.) Do not submit your poster in any other file format than pdf (such as ppt, doc, png, etc.).

  Your poster should be structured so as to contain an informative title and section headings corresponding to each part of the analysis. Furthermore, each section should include a brief written summary of the results and answers to all questions and graph(s) of the corresponding part. Answers that do not appear on the poster will not be marked. Write in a clear and objective tone, avoid slang or conversational language, and be understandable to someone who has not worked with the data set but knows basic statistics. You may include further information about the data or research that you think could make the poster interesting; cite any references you have used. The text on the poster should be mostly written in full sentences. As a soft guideline, aim to write between 300 and 500 words in total.

You can find more information on good formatting and writing in the Week 8 ELE tile along with the mock coursework (link). The book, *Communicating with Data: The Art of Writing for Data Science* by Deborah Nolan and Sara Stoudt (2021), may be helpful and is available through the university library.

- **R Code Format:** Submit all R code you used to produce your answers and figures as a **single .R script file**. It should be possible to open and run the file in RStudio without errors. The R code should be formatted for good readability, using appropriate spaces, indentation, line breaks, and code comments. You can complete this coursework using the code used throughout the lectures and lab exercises. If you use a function that was not covered in the course, add a comment on how or where you found it.
- **How to Zip Files:** Since your submitted coursework is comprised of two files, you must zip them together to create a single, compressed file to upload to ELE. Here is how to create a zipped file: (a) link for Windows (e.g., a PC) or (b) Mac.

Marking Guidelines

This summative coursework will be worth 15% of your final module mark. Your poster and code will be marked against the following criteria, with some indicative questions that will be considered by the marker:

- **Accuracy, 40%:** Is the data source cited? Are all the reported answers correct? Do the visualisations display the correct data? Do the written descriptions match the visualisations? Does the summary provide a coherent description of the analyses and results? Do the conclusions logically follow from the data analysis results? Are the reported answers relevant to the questions?
- **Coding, 30%:** Does the submitted R script run without errors and produce all results reported on the poster? Is the code include suitable comments and formatted for good readability? Does the code make use of functions taught in the module?
- **Writing Style, Visualisations, and Poster Design, 30%:** Is the written description of data and figures clear and concise? Does the text contain typos and/or grammatical errors? Are the graphs appropriately labeled (e.g., titles, units of measurement) and is the text on the graph legible? Are suitable units of measurement provided with all calculated summary statistics? Does the poster look visually attractive? Were font types, sizes, and colors sensibly chosen? Is there a stuiable title and section headings?
- **Creativity Bonus, +10%:** A bonus of up to 10% can be awarded if the report contains additional interesting data analyses that were not explicitly asked in any of the questions and/or if the visualisations are particularly well-done. (Note: The total coursework mark cannot exceed 100%.)

Table 1: Description of columns in `manydogs_data.csv`; missing data has been encoded as `NA` in the data file.

	name	description	type
1	`age`	age of the dog in years	number
2	`sex`	sex of the dog (`Female`, `Male`)	text
3	`owned_status`	where the dog lives (`Private home`, `Group housing` (e.g., working dog kennel), or `Other`)	text
4	`cbarq_train1`	"When off the leash, returns immediately when called" (Options: `0`=Never, `1`=Seldom, `2`=Sometimes, `3`=Usually, and `4`=Always)	number
5	`cbarq_train2`	"Obeys the 'sit' command immediately" (Options: `0`=Never, `1`=Seldom, `2`=Sometimes, `3`=Usually, and `4`=Always)	number
6	`cbarq_train3`	"Obeys the 'stay' command immediately" (Options: `0`=Never, `1`=Seldom, `2`= Sometimes, `3`=Usually, and `4`=Always)	number
7	`cbarq_train4`	"Seems to attend/listen closely to everything you say" (Options: `0`=Never, `1`=Seldom, `2`=Sometimes, `3`=Usually, and `4`=Always)	number
8	`cbarq_train5`	"Slow to respond to correction or punishment" (Options: `0`=Never, `1`=Seldom, `2`=Sometimes, `3`=Usually, and `4`=Always)	number
9	`cbarq_train6`	"Slow to learn new tricks or tasks" (Options: `0`=Never, `1`=Seldom, `2`=Sometimes, `3`=Usually, and `4`=Always)	number
10	`cbarq_train7`	"Easily distracted by interesting sights, sounds, or smells" (Options: `0`=Never, `1`=Seldom, `2`=Sometimes, `3`=Usually, and `4`=Always)	number
11	`cbarq_train8`	"Will 'fetch,' or attempt to fetch, sticks, balls, or objects" (Options: `0`=Never, `1`=Seldom, `2`=Sometimes, `3`=Usually, and `4`=Always)	number