MSc Data Science Bootcamp Skills Test

Harper Adams University

Ed Harris

2022

Instructions

1) This assessment covers material in the Data Science Bootcamp, found here: **Bootcamp webpage** https://dsgarage.netlify.app/bootcamp/
2) The assessment will require the use of R and RStudio, and is “open book” in that you are permitted and encouraged to use your own notes, the web, books, articles and the Bootcamp pages themselves to help you answer questions. However, you must complete the test on your own and no collaboration with others is permitted. There is an ‘honour system’ for this, which you must agree to by submitting your assessment.
3) Your deliverable for this quiz is an R script named in the format “**YOUR-NAME.R**” (for me, my script would be named “**ED-HARRIS.R**”)
4) Your script must be returned to Ed by email at **eharris@harper-adams.ac.uk** within 24 hours of receiving the questions by email
5) Questions or issues can be raised in the Bootcamp Slack channels

Assessment questions

1) Start a reproducible script for the assessment

Start, save and name your assessment script file. Set the script up according to best practice as instructed in the Data Science Bootcamp (e.g. with header, contents, clickable sections and comments as appropriate throughout.). This is the file you will return for this assessment and it should fully run with no faults to reproduce your results for the assessment questions, as appropriate. The clickable sections in your code should correspond to the questions in this assessment.

2) Show the R code to produce the following output:

```
> mymat
       dog cat
female  32  34
male    28  42
```

3) Show the R code to make a *good* histogram of the following data, describing the height in centimeters of domestic cats at the shoulder:

```
cat <- c(20.7, 21.7, 23.7, 27.1, 20.0, 27.0, 27.4,
24.6, 24.3, 18.6, 20.1, 19.8, 24.9, 21.8, 25.7,
23.0, 25.2, 27.9, 21.8, 25.8, 27.3, 20.1, 24.5,
19.3, 20.7, 21.9, 18.1, 21.8, 26.7, 21.4, 22.8,
24.0, 22.9, 19.9, 26.3, 24.7, 25.9, 19.1, 25.2,
22.1)
```

4) Using the `cat` data object from the previous question, perform the Shapiro test to decide whether the data adhere to a Gaussian distribution. Show your R code to do this and also, in comments, report and briefly interpret your results in the technical style.

5) Using the following code, examine the 'CO2' data frame and help file

```
data(CO2)
help(CO2)
head(CO2)
```

Show the code to select all columns of the data frame for the rows where `conc` is equivalent to 350, then to make a *good* boxplot showing the `uptake` variable as a function of the `Treatment` variable (i.e. just for the rows where `conc` is equivalent to 350).

6) Given the data:

```
y <- c(1.36, -0.10,  0.39, -0.05, -1.38, -0.41, -0.39, -0.06,  1.10,  0.76)
x <- c(-0.16, -0.25,  0.70,  0.56, -0.69, -0.71,  0.36,  0.77, -0.11,  0.88)
```

Show the code to make a figure identical to **Figure 1** below (figures appear at the end of this document). Note the line on the figure is the regression line of prediction for the given data (`y` as a function of `x`).

7) Show the code to calculate the residuals for a simple linear regression of `y` as a function of `x` for the data in the previous question, and calculate the mean and standard deviation of the residuals. In comments, report your findings appropriately and briefly interpret them.

8) Based on on Figure 2 below, relate in comments the degree to which you believe the Excel dataset shown adheres to Tidy Data standards. Be brief, but be as specific as you can.

9) Perform 1-way Analysis of Variance to evaluate the overall effect of insect spray type (`spray`) on the number of insects counted (`count`) for the `InsectSprays` dataset. Show your code and briefly summarize and interpret your results in comments. Also, show the code to make a *good* and appropriate graph for the data and statistical test.

Use `data(InsectSprays)` to load the data frame.

10) Create a github repository called `bootcamp`, and create an html R Markdown document with your complete answer to question #9 in it, showing both code and graph output. Push your R markdown file, your html file and other related files to your new repository. Your answer to this question should simply be a commented html link to your new github repository.

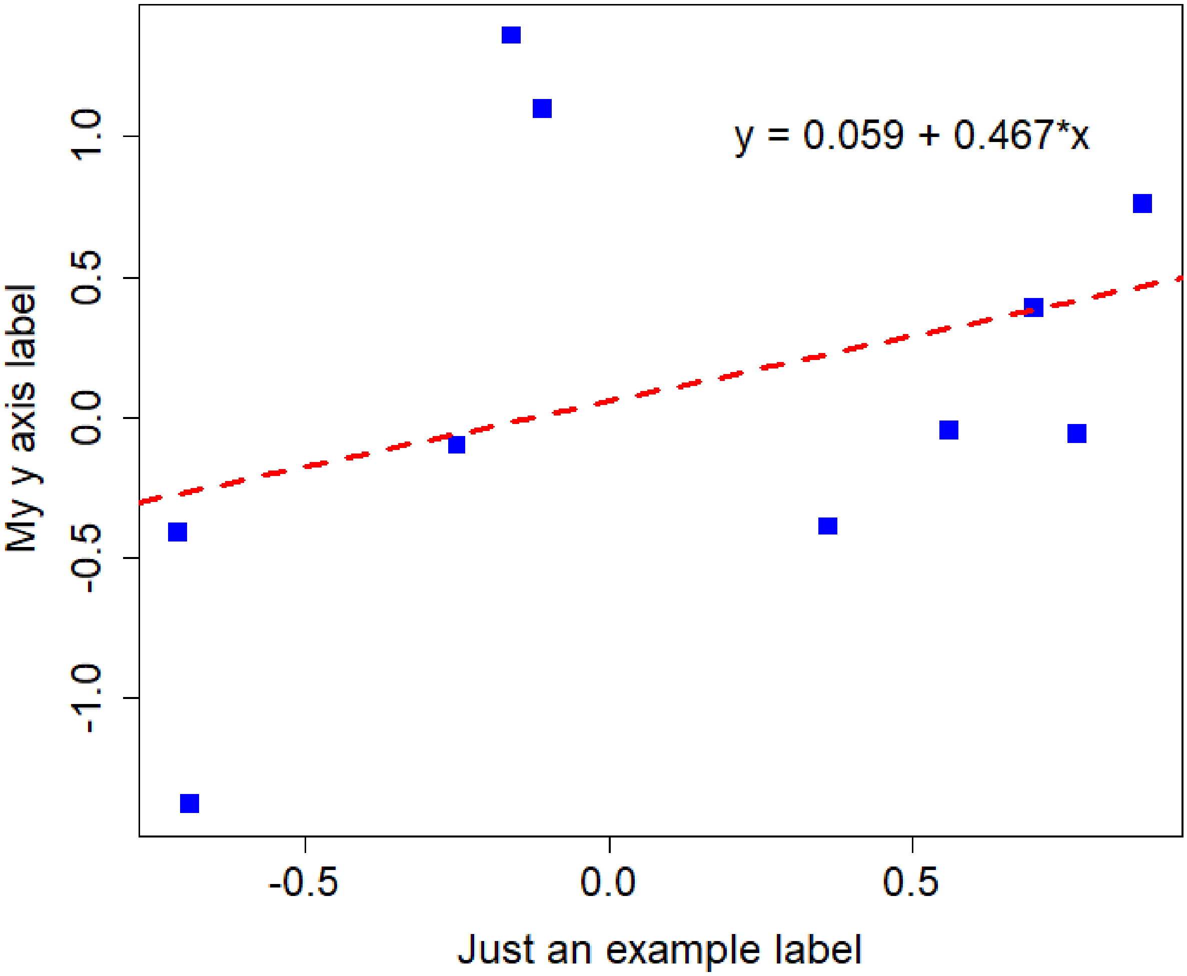
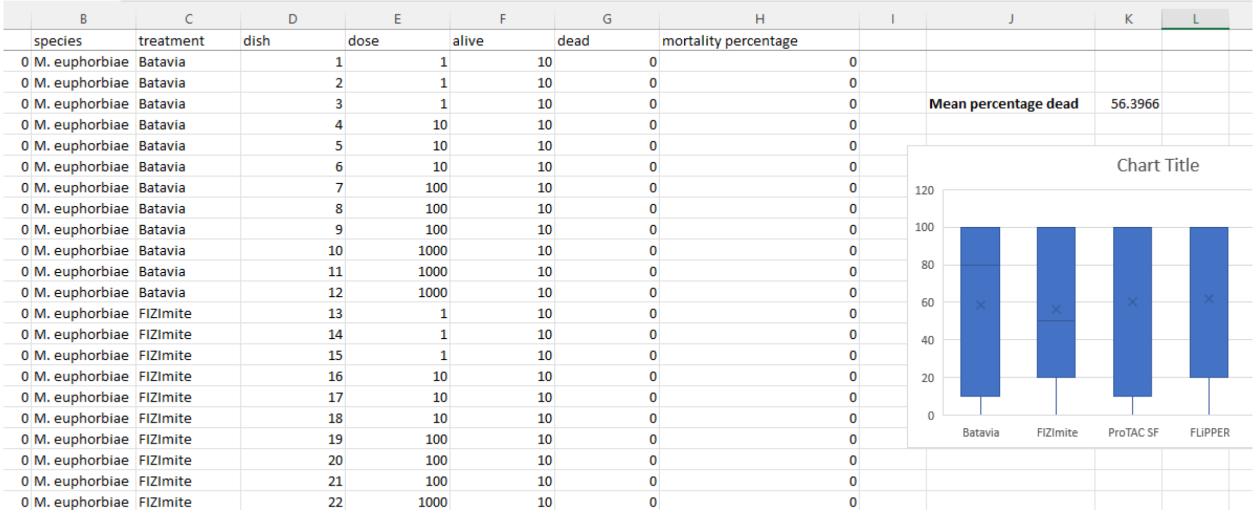


Figure 1: Show the code to reproduce this image



	B	C	D	E	F	G	H	I	J	K	L
	species	treatment	dish	dose	alive	dead	mortality percentage				
0	M. euphorbiae	Batavia	1	1	10	0	0				
0	M. euphorbiae	Batavia	2	1	10	0	0				
0	M. euphorbiae	Batavia	3	1	10	0	0		**Mean percentage dead**	56.3966	
0	M. euphorbiae	Batavia	4	10	10	0	0				
0	M. euphorbiae	Batavia	5	10	10	0	0				
0	M. euphorbiae	Batavia	6	10	10	0	0				
0	M. euphorbiae	Batavia	7	100	10	0	0				
0	M. euphorbiae	Batavia	8	100	10	0	0				
0	M. euphorbiae	Batavia	9	100	10	0	0				
0	M. euphorbiae	Batavia	10	1000	10	0	0				
0	M. euphorbiae	Batavia	11	1000	10	0	0				
0	M. euphorbiae	Batavia	12	1000	10	0	0				
0	M. euphorbiae	FIZImite	13	1	10	0	0				
0	M. euphorbiae	FIZImite	14	1	10	0	0				
0	M. euphorbiae	FIZImite	15	1	10	0	0				
0	M. euphorbiae	FIZImite	16	10	10	0	0				
0	M. euphorbiae	FIZImite	17	10	10	0	0				
0	M. euphorbiae	FIZImite	18	10	10	0	0				
0	M. euphorbiae	FIZImite	19	100	10	0	0				
0	M. euphorbiae	FIZImite	20	100	10	0	0				
0	M. euphorbiae	FIZImite	21	100	10	0	0				
0	M. euphorbiae	FIZImite	22	1000	10	0	0				

Figure 2: Tidy data in Excel - how tidy is it?