**Question 1. [15 marks]**

Consider the matrix given below.

1. Write Python code to enter the matrix as a numpy array, print out , and calculate thesum of each row and column using numpy.sum().*Include both Python code and output in your answer.*

[5 marks]

|  |
| --- |
| *.* |

1. Consider the diagonal matrices and given in the Python code above. Calculate the matrix product and print out . What property do the rows and columns of have? Hint: is called a “doubly stochastic” matrix.*Include both Python code and output in your answer.*

[5 marks]

|  |
| --- |
|  |

1. Write Python code to build the matricesand . Use numpy.allclose() to demonstrate that one of and is asymmetricmatrix, the other one is anantisymmetric matrix, and that gives.*Include both Python code and output in your answer.*

[5 marks]

|  |
| --- |
|  |

**Question 2. [10 marks]**

Consider the *triangular* distribution which has three parameters: (left), (middle), and (right). The Python code given below plots a particular triangular distribution with the given values of , and .

1. The *mean* of a triangular distribution is given by. Calculate the mean, the median, and the upper quartileof the particular triangular distribution above and add them to the plot using vertical dashed lines.*Include both Python code and output in your answer.*

[4 marks]

|  |
| --- |
|  |

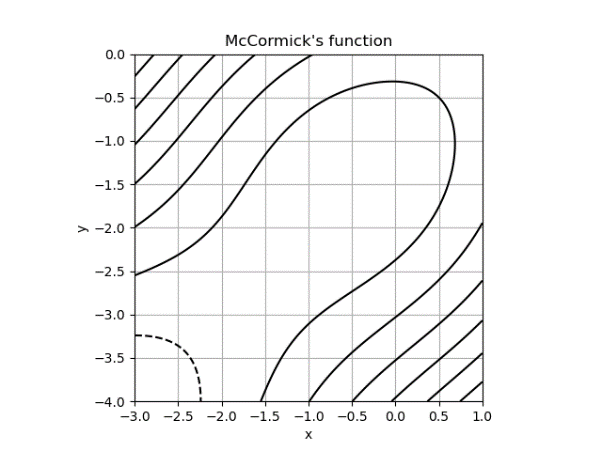
1. Generate a random sample of 10000 values from the particular triangular given above. Plot a histogram of your sample using 50 bins. Calculate the mean, median, and upper quartile from your sample and compare to the equivalent values from part (a).*Include both Python code, histogram and output in your answer.*

[6 marks]

|  |
| --- |
|  |

**Question 3. [10 marks]**

Consider *McCormick’s* function and the corresponding contour plot given below.



1. Briefly explain what is mean by a “contour” on a contour plot, i.e., what does each individual contour represent?

[2 marks]

|  |
| --- |
|  |

1. Write Python code to reproduce the contour plot of McCormick’s function given above (this plot uses only the default levels). You may find the function numpy.sin() helpful. *Include only your Python codein your answer.*

[4 marks]

|  |
| --- |
|  |

1. Add further Python code to label the contours and add further contours that clearly highlight the valley bottom near the point and investigate the behaviour of the function near the point.

[4 marks]

|  |
| --- |
|  |

**Question 4. [20 marks]**

Consider the dataset in the file “animals.csv” (provided along with these questions on Moodle) is extract from the Living Planet Index database <https://www.livingplanetindex.org/>.Each row of the dataset corresponds to a published survey of one particular species of animal.

*In RStudio, make sure you go to the Session menu, select Set Working Directory and then Source File Location. Save the “animals.csv” file to the same folder as where your R code is saved.*

1. The variable (column) *System* gives the ecosystem that each species of animal was found in. Write R code to produce a summary table that gives a count of the number of rows of the dataset corresponding each ecosystem. *Include both R code and output in your answer.*

[3 marks]

|  |
| --- |
|  |

1. Write R code using a dplyr pipe to find each *Country* in which a Hippopotamus has been found. Make sure that each country appears at most once. *Include both R code and list (or table) of countries in your answer.*

[5 marks]

|  |
| --- |
|  |

1. Write R code to reproduce the graphical plot below as accurately as possible.*Include both R code and the plot that your code produces in your answer.*

[6 marks]

|  |
| --- |
|  |

1. Consider the Cheloniamydas (known by its common name of “Green turtle”). Use the geographical coordinates *Latitude* (degrees North-South) and *Longitude* (degrees East-West) to plot the locations where Green turtles have been found. Use the categorical variable *Region* to colour each point (together with a legend). What is the name of the island that is the southern-most *Location* where Green turtles have been found? *Include both R code and the output in your answer.*

[6 marks]

|  |
| --- |
|  |

**Question 5. [25 marks]**

Consider the dataset “bull\_sales.csv” (provided along with these questions on Moodle)which gives the sale price (the variable *sale\_price*) of bulls sold at auction and the *breed* of bull: Angus (a), Hereford (h), or Simmental (s). Several further measurements are given for each bull, including height at one year old (*yearling\_height* in inches), percentage of fat free body weight (*fat\_free\_body*), amount of back fat (*back\_fat* in inches), height when sold (*sale\_height* in inches) and weight when sold (*sale\_weight* in pounds).



Angus bull – image from <https://www.aberdeen-angus.co.uk/sire-verification/>

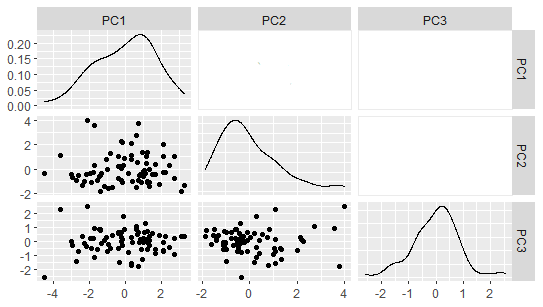
*In RStudio, make sure you go to the Session menu, select Set Working Directory and then Source File Location. Save the “bull\_sales.csv” file to the same folder as where your R code is saved.*

1. The R code above produces a scatter matrix. Comment on the relationship between *sale\_price* and *yearling\_height* and the relationship between *sale\_height* and*back\_fat*.

[4 marks]

|  |
| --- |
|  |

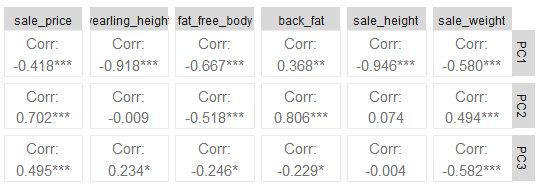
1. Suppose the dataset is scaled and Principal Component Analysis (PCA) is applied. Explain why the following scatter matrix is (mostly) reassuring and fill in the missing parts of the scatter matrix.



[4 marks]

|  |
| --- |
|  |

1. Scaling the dataset and applying PCA gives the partial scatter matrix showing the correlations between the first three principal components (PC1, PC2, and PC3) and the sixoriginal quantitative variables.

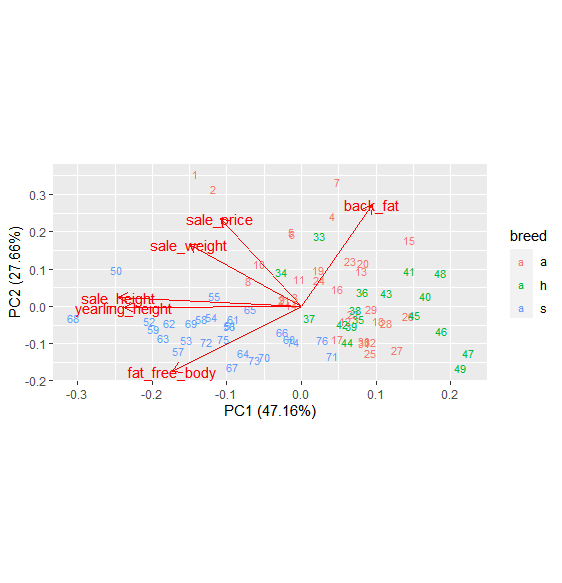


Use R to produce a PCA loadings plot showingPC1, PC2 and PC3. By considering the loadings plot, the variance explained by each principal component, and the partial scatter matrix above, give one possibleinterpretation of each of PC1, PC2 and PC3.Justify whether {PC1, PC2, PC3} are sufficient*Include your R code and output in your answer.*

[9 marks]

|  |
| --- |
|  |

1. Consider the PCA biplotgiven below. Use R to produce an appropriatedendrogram giving a hierarchical clustering of individual bulls. Compare the clusters formed in the dendrogram(using five clusters) with the clusters of individual bullsvisible in the PCA biplot(consider both location of clusters and mixing of breeds).*Include any R code and dendrogram you produce in your answer.*



[8 marks]

|  |
| --- |
|  |

**Question 6. [20 marks]**

This question uses the same “bull\_sales.csv” dataset as Question 5.

1. Write R code to fit the linear model “sale\_price~yearling\_height” to the dataset and construct an appropriate scatterplot including the line of best fit. Write down the equation of the fitted model. *Include your R code and scatterplot in your answer.*

[6 marks]

|  |
| --- |
|  |

1. Assess whether the linear model from part (a) is a good fit to the dataset by looking carefully at the diagnostic plots. *Include your comments only. You may find the R code below useful.*

[6 marks]

|  |
| --- |
|  |

1. Write R code to fit the linear model “sale\_price~yearling\_height+breed” to the dataset and clearly interpret the fitted model.*Include your R code in your answer.*

[5 marks]

|  |
| --- |
|  |

1. Would you recommend adding *sale\_weight* to the model from part (c) as an additional predictor? *Justify your answer.Include any R code in your answer.*

[3 marks]

|  |
| --- |
|  |