Data Modelling and Analysis

Mock Lab submission: 2020/2021

Dr Mercedes Torres Torres

Contents

Submission Information

Please complete the following information here:

1. First Name:
2. Last Name(s):
3. Student ID:

Information

This lab submission is composed of three parts, designed to assess your knowledge of data modelling, pre-processing, mining, and visualisation. It covers concepts seen in the the first seven labs of COMP4030.

This submission is open book and amounts to 25% of your final mark for COMP4030.

You have 2 hours to complete this submission and upload it to Moodle, from 11:00 to 13:00 on the 21st of April 2021.

Submission instructions

1. The submission deadline is on the *11th of May (Monday) at 13:00*.
2. Fill this .Rmd file with the answers to the questions.
3. All exercises must be done in R.
4. You may create as many auxiliary functions as you need.
5. You may use annotations to your code as you see fit.
6. Rename this .Rmd file as *COMP4030-XXXXXXXXXX.Rmd*, where *XXXXXXXXXX* should be replaced with your student ID number (e.g. COMP4030-40781818.Rmd), and submit the single .Rmd file via Moodle (see website for the dedicated link).

6 *Make sure your full name and student ID are shown in the first page of your report.*

Assessment Criteria

The main assessment criteria for the lab submission are:

- *Correctness*: that is, do you apply techniques correctly; do you make correct assumptions; do you interpret the results in an appropriate manner; etc.?
- *Completeness*: that is, do you apply a technique only to small subsets of the data; do you apply only one technique, when there are multiple alternatives; do you consider all options; etc.?
- *Originality*: that is, do you combine techniques in new and interesting ways; do you make any new and/or interesting findings with the data?
- *Argumentation*: that is, do you explain and justify all of your choices?

Plagiarism and Collusion

Plagiarism and collusion are completely unacceptable and will be dealt with according to the University's standard policies. Do NOT, under any circumstances, share code, figures, graphs, charts, results, etc.

1 Modelling

```
knitr::opts_chunk$set(echo=FALSE, eval=FALSE)
```

For your first assignment as a wildlife conservationist in Brazil, you have been sent to the Amazonian rainforest to study the behaviour of a population of jaguars and anacondas with particularly violent tendencies. Natural enemies, these two species compete for survival and control of resources in the rainforest.

Suppose that in the absence of an encounter against each other, both species exhibit unconstrained growth. In this case, the change in population during one month is proportional to the population size at the beginning of the month.

In this scenario, you are aware of the following information:

- an_growth: the growth rate for anacondas
- jag_growth: the growth rate for jaguars
- You are looking at monthly changes.

1. Following the steps seen in the labs, build a complete mathematical model to show how both populations change when there are no violent encounters between them. Use R to solve this model. What would happen in the next two years if there were 141 jaguars and 132 anacondas, when observable growth rates are 2.1% for the jaguars and 2.8% for the anacondas?

1.a Answer the following questions:

- Is the model deterministic or stochastic? Why?
- Is the model natural of artificial? Why?
- Is the model distributed or lumped? Why?
- Is the model linear or non-linear? Why?

2. Following the steps seen in the labs, build a complete mathematical model to show how both populations change when there are violent encounters between. Use R to solve this model and interpret the solution for the next two years when there are 300 jaguars and 300 anacondas. Growth rates are maintained from the previous scenarios, and you have calculated the following rates:

- jag_kill: the proportion of jaguars killed in one-on-one fights with anacondas (0.1%)
- an_kill: proportion of anacondas killed in one-on-one fights with jaguars (0.2%)

2.a. Answer the following questions:

- Is the model discrete or continuous? Why?
- Is the model natural of artificial? Why?
- Is the model dynamic or static? Why?
- Is the model linear or non-linear? Why?
- Is the model robust? Why?

2 Data Analysis, Visualisation, and Mining

In this exercise, we will use the *rock* dataset in R. Load it, save it into a dataframe called *r*. Then, create the following functions:

1. Write a function, called *analysis(r)* returns a dataframe with the number of elements, mean, median, variance and interquartile range of the attribute *r*. Test this function with *rock*.
2. Use Principal Component Analysis (PCA) to create a reduced version of the dataset *r*, called *r_redux*. Show a summary of the results of the PCA. What is the minimum number of dimensions that you retain to ensure 95% of the variance?
3. Write a function, called *visualisation(r)* that prints boxplot of all relevant fiels in *r*.
4. Write a function called *shape_rate(r)*, which modifies *r* by creating a new attribute called *rating*. *rating* is equal to *normal* if the shape is between 0.1 and 0.3 and *abnormal* otherwise. This function should also prints the ratio of *normal* to *abonormal* observations.
5. Write a function *class_analysis(r)* that calculates the median *area* and *shape* of *normal* and *abnormal* samples in *r*
6. Write a function, called *create_scatterplot(r)*, which takes a dataframe *r* as input and prints a scatterplot of *area* and *shape* of those observations in *r* with a permeability *perm* over 500. Include the *perm* information in the scatterplot. Test this function.