Week 3 Practical: Data Cleaning

Overview

This week's practical will introduce students to data cleaning and basic imputation methods in python using pandas and numpy.

Practical Learning Outcomes

After completing this practical students will be able to:

PLO1 – identify and replace missing data.

PLO2 – transform data.

Class Activity

You will work on this activity either individually, in a small group or with the whole class as directed. This part of the practical does not need to be submitted but will provide the formative learning experience that will help you to do the Individual Problem solving task, which is required to be submitted.

Task 1: Data Cleaning

In this task you will use pandas to automatically identify missing values and replace these with the values you decide to add. You will also transform data using functions or mapping.

Students should open first open "0. Data Cleaning.ipynb" in Jupyter and work through and run each cell. Ensure you understand what each code cell is doing and how it works. Try alternatives to ensure you understand numpy and matplotlib.

Individual Problem Solving Task

Your individually worked solution to this section should be submitted as part of your Assignment 1 Problem Solving Task Part A (Weeks 1 -3).

Task 2: Problem Solving in Python

Your task this week is to create a new notebook called "[yourstudentID]_Week_3_Problem Solving Tasks.ipynb". In this you need to create cells to load the abalone.data data and identify any missing data. Where possible replace values appropriately or remove rows or columns that should be replaced. It is your choice what to do with the values found to be incorrect. Include a written statement explaining what errors you found and what you did to correct those errors.

Task 3: Problem Solving on Paper

Given the following table of data use z-score to calculate and identify any outliers where their z value is greater than 3 or less than -3.

4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
54	3.7	1.5	0.2
4.8	3.4	1.6	0.2
4.8	3	1.4	0.1
4.3	3	1.1	0.1
5.8	4	1.2	0.2

5.7	4.4	1.5	0.4
5.4	3.9	1.3	0.4
5.1	3.5	1.4	0.3
5.7	3.8	1.7	0.3
5.1	3.8	1.5	0.3
5.4	0.0	1.7	0.2
5.1	3.7	1.5	0.4
4.6	3.6	1	0.2
5.1	3.3	1.7	0.5

Task 4: Problem Solving on Paper

Now use linear regression to impute a value for the values you identified as missing in Task 3 above.