CS24420 Scientific Python --- Semester 2 --- Practical Worksheet 3

Descriptive Statistics

The tasks this week are about using descriptive statistics (and a bit of plotting).

1. Write a Python program to calculate the centre and spread of a uniformly-distributed array

a) Generate an array of 100 **uniformly**-distributed random real numbers in the range 0 - 100.

b) Calculate the mean, median, variance and standard deviation of this array.

c) Calculate the skewness and kurtosis of this array.

2. Write a Python program to calculate the center and spread of a normally-distributed array

a) Repeat Q1, but using **normally**-distributed random numbers with a mean of 50 and a standard deviation of 10.

b) Compare the two sets of results - do they show what you would expect?

3. Write a Python program to compute some statistics for a DataFrame, and plot the results

Download the file `camden_trees.csv`, which contains data from a tree survey conducted in Camden. The columns are:

Identifier	: Tree identification string
Name	: Common name of tree species
Height	: Height in metres
Spread	: Spread (width) in metres
Diameter	: Diameter at chest height (cm)
Maturity	: Description of age of tree (text)
Condition	: Estimation of condition of tree (text)

a) Read the file into a pandas `DataFrame` object using `read_csv`

b) Calculate and print the summary statistics of the data: count, min, max, mean, median, mode etc.

c) Plot the **Height** column as a histogram. Add a title and axis labels. Remember you can use `plt.savefig` to save the plot and `plt.clf` to clear the drawing for a new plot.

d) Plot the **Spread** column as a histogram. Add a title and axis labels.

e) Group the data by tree maturity and assign the resulting object to a variable.
[Hint: see `DataFrame.groupby`]

f) Plot the grouped **Height** columns as histograms. Include a legend.
[Hint: By default, they will draw on top of each other. You could use the parameter `alpha=0.5` to make the histograms semi-transparent, or loop over the groups doing an individual plot for each, or use `plt.subplot` to combine them]

g) Using the grouped **Spread** columns, plot the *maximum* spread for each age of tree as a bar chart. You might need to use `plt.tight_layout` if the labels are too long to fit.

2023-02-17 LGT