MOD550 – Assignment 1

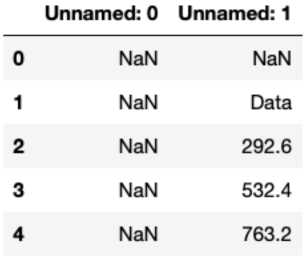
Learning Objectives:

- Understand use of various libraries of Python as Scipy, Matplotlib, Pandas
- Be able to read and do some simple cleaning of data files
- Understand Histograms and Relative Frequency Plots
- Calculate Means and Standard Deviations from first principles
- Understand Cumulative Density Plots
- Understand Percentiles and how to calculate them from a CDF
- Understand construction of Q-Q and P-P plots to compare distributions
- Improve your general Python skills

Problems

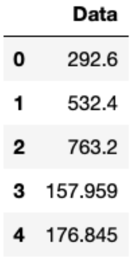
1. Data cleaning (Python)

a) Read the file "a1_data1.xlsx" into a Pandas Dataframe. Do not modify the Excel file before reading it. If you take a look at the top 5 rows of the resulting DataFrame, it will look like this



	Unnamed: 0	Unnamed: 1
0	NaN	NaN
1	NaN	Data
2	NaN	292.6
3	NaN	532.4
4	NaN	763.2

because the first row and column in the Excel file are empty. Use Python to modify (clean) the DataFrame to look like this



	Data
0	292.6
1	532.4
2	763.2
3	157.959
4	176.845

b) The data set contains several empty cells. Use Python to count how many empty cells there are.

c) In general, we want to remove empty cells before doing any calculation. Delete (drop) the rows with empty cells.

2. Mean, Variance, Median, Mode and IQR (Python)

Answer the following by performing calculations in Python (import the data from excel to python)

a) The number of data points:

b) The sum of the data points:

c) The arithmetic average calculated from the 2 previous values.

d) The arithmetic average using built in Python function:

e) How does the built in Python function treat empty cells – as zeros or as null data? (Use data from sheet2 only for this question)

For each data point, calculate the square of the difference between it and the mean.
Calculate the following quantities

f) The sum of the squared difference (hint: use for loop)
g) The variance - the average squared difference and compare with python function
h) The standard deviation and compare with python function
i) Is there any difference between the variances and standard deviations?
j) Median, mode and IQR

3. Histogram (Python)

Calculate the data range (min and max) using the Python functions.

- What are they? Plot a histogram with 20 bins

4. Data analysis (Python)

Try to change the number of bins to 10, 40, 60, 100 and plot the histogram. Try plotting all 4 of these in a 2x2 subplot.

Get the basic statistical parameters (count, min, max, mean, variance, skewness, kurtosis, Quartile1, Median, Quartile3).

5. PDF and CDF (Python)

Create a PDF and CDF using the frequency values obtained in the earlier step (bin size = 20). Plot both of them in Python. Both the plots and data points for PDF and CDF are required as an answer

a) What is the relative frequency of the 400 – 500 bin?
b) What is the PDF like where the CDF is steepest? (by visual inspection)
c) Visually estimate the 10th, 50th and 90th percentiles (P10, P50 (median) and P90) from the CDF plot.
Provide your answers in the Notebook. Is there a Python function you could use to calculate these percentiles?

6. Box Plot (Python)

Create a box plot of the data using either Matplotlib or Pandas.

7. Q-Q and P-P plots (data in “Q-Q & P-P” Sheet) (Python)

Read the Excel file “a1_qqpp_data.xlsx” into Python.

a) Use Python to generate a Q-Q plot to compare the distributions of

i. Medium Sand Porosity in Well 21-P and Medium Sand Porosity Well 32-P
ii. Medium Sand Porosity in Well 21-P and Fine Sand Porosity Well 21-P

b) Use Python to generate a P-P plot, at 1% porosity units, to compare the distributions of

i. Medium Sand Porosity in Well 21-P and Medium Sand Porosity Well 32-P
ii. Medium Sand Porosity in Well 21-P and Fine Sand Porosity Well 21-P

8. Bayes – Students (Python or any other way)

Experience has shown that students are unable to submit their homework on time (NH) for one of two reasons: computer crash (CC) or dog eating the homework (DH). The probability of CC is known to be 0.20, with the probability of no-homework submission because of CC being 0.50. The probability of DH is known to be 0.01, with the probability of no-homework submission because of DH being 0.99. If a student was unable to submit the homework on time, what is the probability that a dog ate the homework?

9. Exploration Drilling (Python or any other way)

You own an oil company planning to drill in a play where the probability of finding oil is believed to be high. You have identified two prospective locations, A and B. The probability of finding oil at location A is estimated to be 70%. The probability of finding oil at location B is estimated to be 60%. If oil is found at location A, that will be a positive sign for the play, and the probability of finding oil at location B increases to 80%.

a. Draw the probability tree that best represents this situation.

b. What is the probability of finding oil at location B given that we do not find oil at A?