MOD550 – Problems

1. Data cleaning

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

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

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

- What are they? Plot a histogram with 20 bins

4. Data analysis

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

Create a PMF and CDF (bin size = 20). Plot both of them in Python.

a) What is the relative frequency of the 500 – 540 bin?
b) What is the PDF like where the CDF is steepest? (by visual inspection)
c) Visually estimate the 10^{th}, 50^{th} and 90^{th} 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

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

7. Scatterplots and Bivariate Distributions ("scatterplots_DATA.xlsx")

The spreadsheet contains porosity and permeability data from the medium sand in Well 21-P. kH refer to horizontal permeability and kV to vertical permeability. The units for porosity is % and for permeability mD.

a) Generate a scatterplot of
    i. Horizontal Permeability (kH) versus Porosity in Well 21-P
    ii. Log10 of kH versus Porosity in Well 21-P
    iii. Vertical Permeability (kV) versus Horizontal Permeability (kH) in Well 21-P

b) Generate a bivariate
    i. histogram of Porosity (5% increments) and log10 kH (0.5 increments)
        - Your poro_bin edges should be from 0 to 50% in 5% steps and
        - Your perm_bin edges should be from 0 to 4 in 0.5 mD steps
    ii. cumulative frequency distribution of porosity and log10 kH

- The counts in each bin must be normalized so that the cumulative distribution goes from 0 (at poro = < 5% and log10_kH < 0.5) to 1 (at poro < 50% and log10_kH < 4)
- Matplotlib's imshow function can be used for this.

c) Generate plots of the marginal frequency distributions for both porosity and log10 kH

d) Generate plots of the conditional frequency distributions for
   i. Porosity when log10 kH is in the 2.0 to 2.5 range
   ii. log10 kH when porosity is in the 30%-35% range

8. Correlation ("scatterplots_DATA.xlsx")

a) Use *Pandas and Numpy* libraries to calculate both the covariance and correlation coefficient between
   i. Porosity and Permeability (kH)
   ii. Porosity and Log10 kH
   iii. Horizontal Permeability (hK) and Vertical Permeability (kV) - using the first 100 data points from the "Scatterplots" sheet
   iv. Log10 kH and Log10 kV - using the first 100 data points from the "Scatterplots" sheet

b) Use *Numpy* library functions to calculate the rank correlation coefficient between
   i. Porosity and Permeability (kH)
   ii. Horizontal Permeability and Vertical Permeability - using the first 100 data points from the "Scatterplots" sheet

c) Why is there no point in calculating the Rank Correlation between Porosity and Log10 kH?

9. Regression ("regress_DATA.xlsx")

Use the 100 porosity and permeability data to generate a linear predictor for permeability and test its unbiasedness. (The poro-perm data are the same as for the correlation exercise)

a) Calculate the "m" and "b" coefficients of a linear predictor of permeability from porosity (think carefully about exactly what variables you will use!). Which is the independent variable and which is the dependent variable?

b) Use the regression line to predict the permeability variable for each porosity measurement. Create a scatter plot of the data pairs as points and also plot the predicted line

c) Calculate the residuals and plot them against the porosity (independent variable). Calculate the correlation coefficient between them. Comment on how good the predictor is.

d) Plot a histogram of both the predicted and actual permeability data. Also calculate the variance of each. Observe and explain the difference between them.

10. Spatial Correlations ("covar_correl_semivario_DATA.xlsx")

There are 6 sets of equally spaced Log Data in the "Well_Data" sheet. The first column shows the depth.

a) Calculate number of data points, mean and variance for each data set

b) For each of the log data sets, create a plot of the data versus depth (a log trace) – with depth on the y- axis, shallowest depth at the TOP.

c) For lags from 0 to 150 in steps of 1, calculate

i. The Covariance function
ii. The Correlogram
iii. The Semi-variogram

d) Plot all three functions on a single graph (Hint: functions will display best if you put the Semi-Variogram and Covariance on the primary Y axis and the Correlogram on the Secondary Y axis)

Visually examine the results for each set of data. Answer the following questions (include your answers in your Notebook).

f) What do you notice about the relationship between the Correlogram, Covariance Function and Semi-variogram

g) What do you observe about the relationship between the data and the semi-variograms

h) Estimate the range and sill for each semi-variogram

i) What do you observe about the relationship between the sills and the variances

11. EW- and NS Directional semi-variograms ("2d_directional_semivariograms_DATA.xlsx")

a) Using the gridded porosity data in the "grid_data" worksheet, calculate the semi-variograms in the NS and EW directions for lags of 1 to 15 in steps of 1. Plot both on a single chart.

b) Calculate the variance of the data. How does it compare with the sill?

c) What are the ranges of the two variograms?

12. Omni-Directional semi-variogram ("2d_directional_semivariograms_DATA.xlsx")

a) Using the well coordinates and their porosities in the "well_data" worksheet, calculate the omni-directional semi-variogram for lags of 1 to 12 in steps of 1. Make the tolerance an input variable and start with a value of 0.5. Also compute the variance of the data.
(Hint: First develop a matrix of distances between well pairs)

b) How does the range compare with the ranges of the EW and NS directional semi-variograms?

c) Vary the tolerance (say 0.1, 0.25, 0.75, 1, 2, 5, 10, 25) and observe the semi-variogram and number of pairs.