# Women's Institute of Technology & Innovation (WITI)
# Introduction to Data Science & Machine Learning
# Practice Questions

Course code: CSD 114
Course duration: 45 hours
Credit Units: 3 CUs

Cohort II 2022

In answering these questions, full marks are given for explanations, not just right answers. Good luck!

1. If we start with a principal sum, $P$, and earn compounded interest, the total accumulated value, $V$, at the end of time $t$ is:

$$V = P\left(1 + \frac{r}{n}\right)^{nt} \tag{1}$$

where $r$ is the annual interest rate and $n$ is the compounding frequency. For example, if you deposit \$2,100 in a bank paying an annual interest rate of 3.4% compounded quarterly, compute the balance after 7 years using the above variables.

   (a) Continuing the previous example, suppose you start with the same principle and the same interest rate, but interest is compounded twice per year, so n = 2. What would the total value be after 7 years? Hint: The expected answer to be a bit less than the previous answer.

   (b) If interest is compounded continuously, the value after time is given by the formula:

$$V = P\ e^{rt} \tag{2}$$

   Translate this equation into Python and use it compute the value of the investment in the previous example with continuous compounding. Hint: The expected answer to be a bit more than the previous answers.

2. In September 2019, The Economist published an article comparing sandwich prices in Boston and London: "Why Americans pay more for lunch than Britons do". It includes a graph in Figure 1 showing prices of several sandwiches in the two cities: The Economist data used data on sandwich prices in Boston and London to create the graph below. Use the following data to replicate the image in python.

   - sandwich prices in Boston: boston price list = $[9.99, 7.99, 7.49, 7.00, 6.29, 4.99]$
   - sandwich prices in London: london price list =$[7.5, 5, 4.4, 5, 3.75, 2.25]$

3. Consider the polynomial expression

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots a_nx^n = \sum_{i=0}^{n} a_ix^i \tag{3}$$

Write a function $p$ such that p(x, coeff) that computes the value in above expression given a point $x$ and a list of coefficients coeff () $(a_1, a_2, \cdots a_n)$. HINT: Try to use enumerate() in your loop.

$$p(x) = a_0 + a_1x + a_2x^2 + \cdots a_Nx^N = \sum_{n=0}^{N} a_nx^n \tag{4}$$

Earlier, you wrote a simple function p(x, coeff) to evaluate the above expression without considering efficiency. Now write a new function that does the same job, but uses NumPy arrays and array operations for its computations, rather than any form of Python loop. HINT: Use np.cumprod().

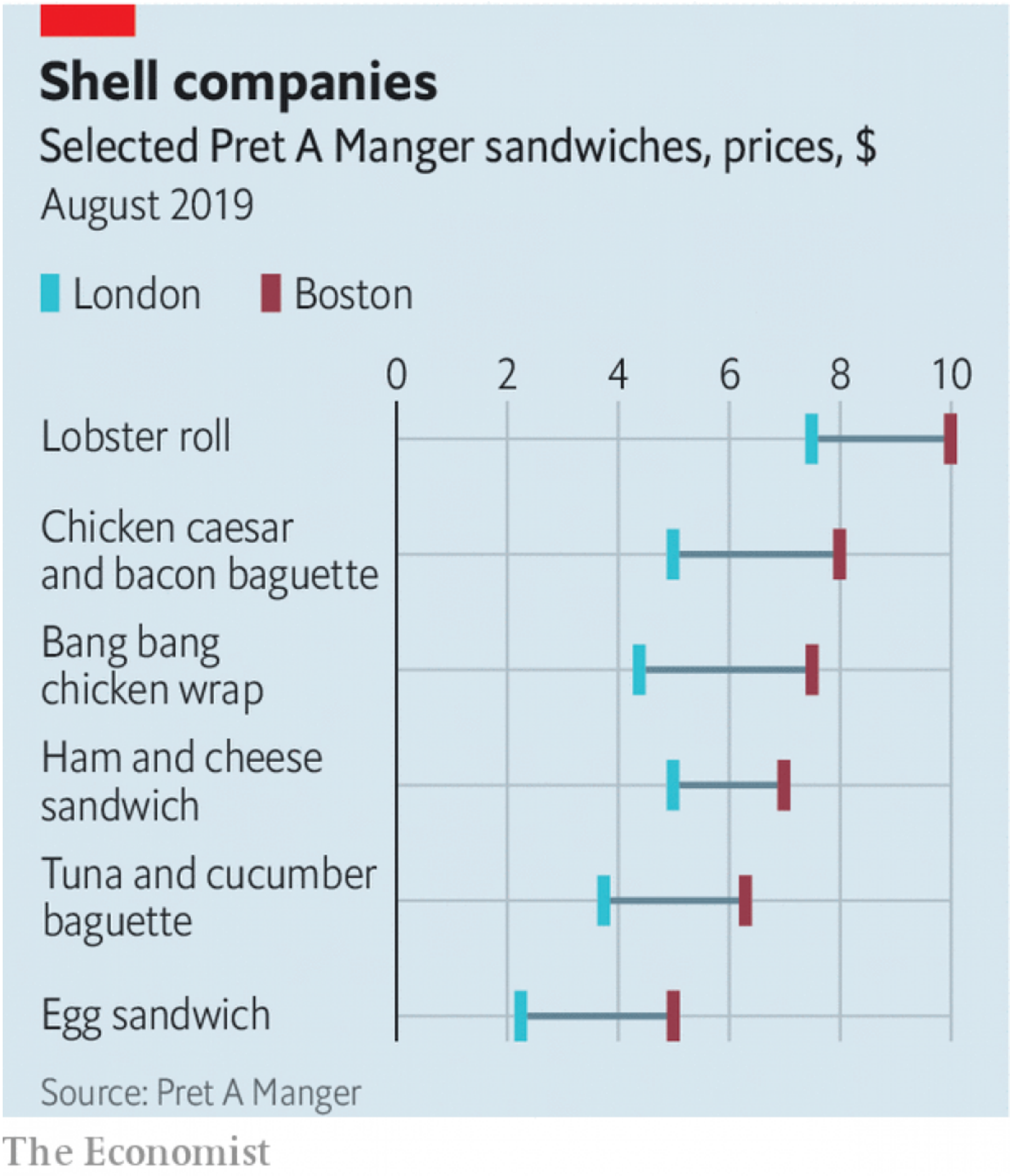

Figure 1: The Economist sandwich

4. Is the normal distribution a good model for the distribution of ages in Uganda's. population as at last census data? To answer this question. Plot the CDF of the ages in the dataset.

   (a) Compute the mean and standard deviation of ages in the dataset.

   (b) Use 'linspace' to create an array of equally spaced values between 18 and 89.

   (c) Use 'norm' to create a normal distribution with the same mean and standard deviation as the data, then use it to compute the normal CDF for each value in the array.

   (d) Plot the normal CDF with a gray line.

5. In general it is a good idea to visualize the relationship between variables before you compute a correlation. Using the covid-19 dataset on muhame _ alon github.com under the capstone repo. Generate a visualization of the relationship between any two continouse variables of your choice. How would you describe the relationship, if any?

   (a) Compute the descriptive statistics.

   (b) Explain and visualise the regression analysis results using the dataset. HINT: We visualise the regression results after fitting the models and then visualize the results.

6. The underlying objective function for K-Means Cluster algorithm tries to find cluster centers such that, if the data are partitioned into the corresponding clusters, distances between data points and their closest cluster centers become as small as possible.

   Given a set of datapoints $x_1, ..., x_n$ and a positive number $k$ , find the clusters $C_1, ..., C_k$ that minimize

   $$J = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \, ||x_i - \mu_j||_2 \qquad (5)$$

   where:
   $z_{ij} \in \{0, 1\}$ defines whether of not datapoint $x_i$ belongs to cluster $C_j$
   $\mu_j$ denotes the cluster center of cluster $C_j$
   $|| \, ||_2$ denotes the Euclidean distance

   (a) Implement a very simple K-Means clustering algorithm (clustering belongs to unsupervised learning) to cluster flowers in the IRIS dataset.

7. Recall that the linear regression problem can be solved using the Least Squares method by optimizing the normal equation below:

   $$\hat{\Theta} = (X^T X)^{-1} X^T y \qquad (6)$$

   Implement the simple linear regression and multiple linear regression by following the steps using the covid-19 dataset:

   (a) Import the Libraries

   (b) Import the Dataset

   (c) Check for any missing data points

   (d) Split the Dataset

   (e) Additionally: One may need to carryout feature scaling for some cases under multiple linear regression but not simple linear regression

MERRY X-MAS AND HAPPY SIMPLE JOYS IN THE HOLIDAYS.