CARNEGIE MELLON UNIVERSITY
DATA, INFERENCE & APPLIED MACHINE LEARNING (COURSE 18-785)
ASSIGNMENT 3

INSTRUCTIONS

- Submissions should be made via canvas.
- **Single** Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb]**. In addition, each line of code should be documented by text. This demonstrates that the code is unique and owned by the student
- Assignment report(.pdf) with full evidence that the assignment was completed by the student and demonstrate a full understanding of each step in the process including textual descriptions of each result (statistics, table, graph etc) represents and insights that can be gained
- Indicate the libraries you have used in your code at the beginning of the report (After the title page)
- Data files (as given)

Submission process:

1. Put source code **file and data files** in a single folder
2. Name of the folder should be the same as your andrew ID
3. **Zip this folder and attach the zipped file on assignment submission page (CANVAS)**
4. After attaching zipped file, click on "Add Another File" from assignment submission page and **attach your report**
5. Submit your assignment

N.B. This process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID, report, and code file with andrewID_DIAML_AssignmentNo. For example, mcsharry_DIAML_Assignment1, mcsharry_DIAML_Assignment2 and mcsharry_DIAML_Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

The student is responsible for checking that their submission is complete. Students will lose 10% as for late submission even if the submission is repaired during the 24 hours after the deadline has passed, and receive 0 for the assignment if it is not repaired.

The submission deadline is **on Monday 03, October, 2022 17:59 Eastern Time (ET) /**

Monday 03, October, 2022 23:59 Rwandan Time (CAT) .

<table>
<tr><th>No.</th><th>Question</th><th>Format</th><th>Value</th></tr>
<tr><td>1</td><td>Daily energy intake in kJ was measured for 11 women (Altman, 1991): 5260, 5470, 5640, 6180, 6390, 6515, 6805, 7515, 7515, 8230, 8770. We wish to investigate whether the women's energy intake deviates systematically from a recommended value of 7725 kJ. Assuming this data comes from a normal distribution; use a t-test to test whether the distribution might have a mean of 7725 kJ. Explain whether a left-tail, right-tail or two-tailed test is appropriate. Give the sample mean, sample standard deviation, standard error of the mean (SEM), t statistic, degrees of freedom and p-value. Finally explain if the null hypothesis is rejected or not.</td><td>Six numbers
Two qualitative answers.</td><td>20%</td></tr>
<tr><td>2</td><td>A Guinness Overall Enjoyment Score (GOES) was used to test if Guinness served in an Irish pub tastes significantly better than pints served elsewhere around the globe. Pints consumed in Ireland received a mean GOES score of 74, while the average GOES score for Guinness tasted elsewhere was 57. The full results were as follows:
<table>
<tr><th>Location</th><th>Sample Size</th><th>Mean</th><th>Standard Deviation</th></tr>
<tr><td>Ireland</td><td>42</td><td>74</td><td>7.4</td></tr>
<tr><td>Elsewhere</td><td>61</td><td>57</td><td>7.1</td></tr>
</table>
Is this difference of 74 versus 57 significant, or is it simply due to natural, random variation? Use a t-test and explain whether a one-sample, two-sample or paired test is appropriate. Show the steps of calculating the t statistic and explain whether a left-tailed, right-tailed or two-tailed test is required. Give the resulting p-value.</td><td>Qualitative answers; t statistic; p-value and code.</td><td>20%</td></tr>
<tr><td>3</td><td>Use data from the World Bank Indicators for 2013 to study the relationship between Fertility rate, total (births per woman) versus GDP per capita PPP (current international $). Make a carefully labelled graph with one dot per country. Estimate the correlation coefficient and give your interpretation.</td><td>Graph, Correlation coefficient. Interpretation.</td><td>20%</td></tr>
<tr><td>4</td><td>Load in monthly average house price data in pounds sterling (£) from Jan 1991 to Dec 2016. Download the data from here (choose the file UK monthly indices (Post '91)). Graph the time series and label it carefully. Construct the autocorrelation function (ACF) of the monthly returns defined as r(t) = [p(t)/p(t-1)]-1 and show the values for lags of one up to 20 using a bar-graph. Indicate the values of the ACF using horizontal lines that would correspond to a statistically significant result at $p<0.05$. From the ACF of monthly data is there evidence of seasonality? Is there a trend in the time series? What is the annualized return over this period as a percentage?</td><td>Two graphs, Three qualitative answers.</td><td>20%</td></tr>
<tr><td>5</td><td>Load in the FTSE100 index from here (ticker = ^FTSE) over the same period (01-Jan-1991 to 31-Dec-2016). Plot the cumulative returns from the House market (using the price data from question 4) and the FTSE100 index on the same graph with the time series normalized such that each starts at 100 in Jan-1991. What is the average annualized return from the FTSE100? Would it have been better to invest in a UK house or the UK stock market over this period?</td><td>Graph. Average Annualized return. Qualitative answer.</td><td>20%</td></tr>
</table>