MATH 70096 - Exploratory Data Analysis & Visualisation

Spring 2022 - Assessed Coursework 1

Deadline: 01 February 2022, 23:59 (UK time)

You should submit an R Markdown notebook, containing your answers to these questions, via the Imperial College Blackboard VLE, by the deadline stated above. Your submission should include both the .Rmd source file for your notebook, as well as the corresponding, compiled PDF document.

This coursework should involve approximately **2 hours** of effort. The available marks are indicated in square brackets for each question.

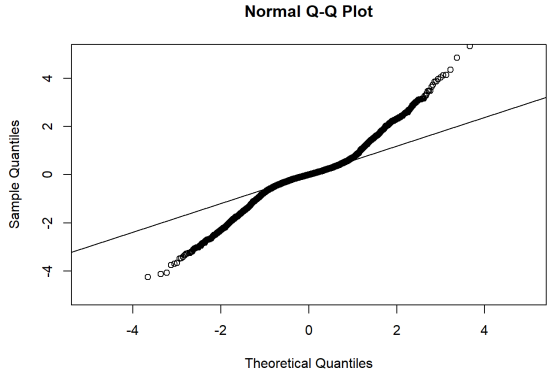
This coursework counts for 5% of your total mark for EDAV.

Plagiarism: Your submission should be *your own work*. Note that software tools are used for plagiarism detection.

Q1) Quantile-quantile (Q-Q) plots are used to compare the empirical distributions of two samples. Recall that they can also be used to test whether a set of observed data, $\underline{x} = \{x_1, \ldots, x_m\}$, is distributed according to a proposed 'theoretical' distribution, by *simulating* n variates $\underline{y} = \{y_1, \ldots, y_n\}$ from the proposed distribution and using this as the second sample for constructing a Q-Q plot. **[3]**

a) Clearly explain the role of n, the size of the sample simulated from the proposed 'theoretical' distribution, in our ability to judge whether $\underline{x}$ is distributed according to the proposed distribution.

b) Suppose a call to `qqnorm(x)` returns the following plot.



Describe how the empirical skewness and kurtosis of the observed data in `x`, compare with the skewness and kurtosis of the Gaussian distribution.

Q2) Consider the dataset provided in the accompanying file `travel-times.csv`. Load this into R, using the `readr` package. **[7]**

a) State the data type for each variable in this dataset, using the NOIR classification.

b) The variable `AvgSpeed` contains missing data, which you are told is ***not*** MNAR. Diagnose the mechanism of missingness for this data. Justify your answer using appropriate visualisations.

c) Propose a method for imputing the missing `AvgSpeed` values, such that the mean of the complete data (i.e. including both observed and imputed values) is an unbiased estimate of the underlying population mean.

d) By calculating the modified Z-scores, establish whether any observations of the variable `TotalTime` are outliers. If any exist, use the `dplyr` package to create a tibble that contains only these observations (and which includes all columns/variables from the original dataset).