# CS4250/5250 Data Visualisation and Exploratory Analysis
# Assessed Coursework Assignment 2

This coursework will count for **10%** of the total marks for the course, and half of the total assessed coursework marks for the course.

## Learning outcome assessed

- Be able to perform open-ended exploratory analysis of data, and master the analytical presentation and critical evaluation of the results of statistical analyses.
- Be able to demonstrate practical experience of using standard graph visualisation methods and evaluation of results.
- Be able to critically assess and evaluate a visualisation.

## Instructions

This coursework should be submitted on Moodle: there will a submission place for the assignment. **Please submit all your work in one file (in PDF) and do not compress your file before submitting it**.

Note: All the work you submit should be solely your own work. Coursework submissions are routinely checked for this.

**The submission deadline is Wednesday 18th March 2020, 17:00. An extension can only be given by the academic advisor (and in some cases by the office, but not by the course lecturer).** Feedback will be given by 21st April 2020.

## Coursework

There are **FOUR** parts in the coursework and answer all parts: 10 marks in total - 2 marks for Part 1, 1 mark for Part 2, 3 marks for Part 3, and 4 marks for Part 4.

## **Part 1**. Diamonds dataset [2 marks].

In assignment 1, you have investigated the diamonds dataset. How could you find out using appropriate visualisations, from the diamonds dataset, whether diamonds that weigh slightly more than one carat are significantly more valuable than diamonds that weigh just less than one carat? Show your visualisations to support your answers.

## **Part 2**. 'Wide' and 'long' (also known as 'tall') data formats [1 mark].

Look at the lecture slides on Data Wrangling.

MSc students should also look at the article "Tidy Data" by Hadley Wickham, Journal of Statistical Software. Preprint available from
http://vita.had.co.nz/papers/tidy-data.pdf

This paper is not technically complicated, and it is particularly nice in that it covers case-by-case several common ways in which data may be strangely formatted, and how to convert these cases to 'long' (aka 'tall') format.

The following is a 'wide' data table: please re-write it in a 'tall' format suitable for use with ggplot2 (*no R programming is required*).

| Country | Population.1960 | Population.1980 | Population.2000 |
|---|---|---|---|
| Ruritania | 1200 | 2400 | 600 |
| Atlantis | 17 | 67 | 68 |
| Oceania | 15000 | 20000 | 53000 |
| Eurasia | 39000 | 120000 | 230000 |

## **Part 3 [3 marks].**

This is a conceptual question. Roughly **one page of A4** please, definitely not more than 2 pages, but write as carefully and precisely as you can. (This is to give you practice in answering exam-style questions.) The aim of the question is to get you thinking about the difficulties of analysing observational data (the database described below is observational), and also some of the difficulties of carrying out experiments in practice.

Read the following scenario carefully.

Imagine that you own a chain of coffee shops (such as Pret-a-Manger or Caffe Nero).

You have collected a lot of sales data over several years. You are very pleased with the amount of data you have collected: it is in the form of records with the following fields (that is, a data frame with rows with the following fields) which records the following information for every item sold:

- shop : the shop where the item was sold
- date : the date the time of the transaction
- time : the time of the transaction (to within a second )
- server : the person who sold the item
- customer : customer number (known if they have a loyalty card, otherwise missing)
- item : the item that was sold
- price : the price

If a customer buys several items at once – for example, an egg and salmon breakfast roll, a fruit-pot, and a strong white Americano – they appear as separate rows in the data frame, but with the same times. Some coffee shops are large, with up to 10 servers, so that sales can be made at the same times.
There are ten coffee shops, each with its own manager.

You – as the central owner – set a 'menu' of 30 different items that managers can sell: on each day, each manager chooses which ingredients to order and which items the staff in that shop will make.

At any time, some of the staff in the shop are selling to customers, others are working the coffee machines, and others are making the food or clearing up.

Each manager selects which items from the 'menu' to sell in his or her shop on any particular day (the staff in the shop make the sandwiches, and they can only make a limited number of different sandwiches on each day).

The shops are in different types of location (stations, streets, shopping malls) and each manager selects the items that they think will sell best in their location, and on the particular day.

a) Summarising the total sales by server shows that some servers sell far more than others. Should you get rid of the servers who don't sell much, or do you need more information? Explain carefully.
b) You suspect that some servers persuade customers to buy extra items, and so generate you more profits. Describe how you might investigate this from the data.
c) You believe that some managers are better than others, as some of the shops are much more profitable than others, and some managers change their menus more often than others. Should you sack your less profitable managers, or do you need more information? Explain carefully.
d) Could you devise an experiment to assess which managers are more competent? Describe how you might do this. Do you think the experiment is feasible?
e) Some types of sandwich have higher sales than others. Do you think it is reasonable to eliminate the sandwiches that sell least in total? Can you devise an experiment to find out whether different sandwiches sell at different rates? Would the results of this experiment directly tell you whether to stop selling some types of sandwich? (Hint: this requires careful thought!)

## Part 4 [4 marks].

Find an example of a visualisation on the web which is intended to persuade you of something, or which is intended to be part of an argument. (That is, don't choose visualisations that are really works of art with no particular purpose.)

As before, please include the following in your submission:

- an image of the visualisation;
- a link to where you found the visualisation;
- a short (less than half a side of A4) discussion of what the visualisation is intended to persuade people of, whether you think it succeeds, and why.

Marks will be given for:

- choice of substantial or unusual visualisation and correct description of the visualisation (30%);
- insightful critique of the visualisations (70%).