

# Inf2 - Foundations of Data Science

## Coursework 1 (CW1)

### Data Wrangling and visualization

Released: Monday 19 October 2020

Deadline for submission: **Friday 30 Oct 4pm** on Learn.

As per the Informatics [Late Coursework & Extensions](#) policy, late coursework without an authorised extension will be recorded as late and the following penalties will apply:

- 5 percentage points will be deducted for every calendar day or part thereof it is late, up to a maximum of 7 calendar days.
- If you have not submitted coursework within these 7 days, a mark of zero will be recorded.

**It is very important that you read and follow the instructions below to the letter: you will be deducted marks for not adhering to the advice below.**

### Good Scholarly Practice

Please remember the University requirement as regards all assessed work for credit. Details about this can be found at: <http://web.inf.ed.ac.uk/infweb/admin/policies/academic-misconduct>

Specifically, this coursework **must be your own work**. We want you to be able to discuss the class material with each other, but the coursework you submit must be your own work. You are free to form study groups and discuss the concepts related to, and the high-level approach to the coursework. **You may never share code or share write-ups**. It is also not permitted to discuss this coursework on Piazza. The only exception is that if you believe there is an error in the coursework, you may ask a private question to the instructors, and if we feel that the issue is justified, we will send out an announcement.

### General Instructions

- This assignment accounts for 20% of the mark for this course and is graded based on a Jupyter notebook, which you are to submit via Learn (see below). The actual assignment is marked out of 100 and the number of marks is indicated by each question below.
- The criteria on which you will be judged include the functional code, and the quality of the textual answers and plots asked for.
- Read the instructions carefully, answering what is required and only that. Keep your answers brief and concise.

- For answers involving figures, make sure to clearly label your plots and provide legends where necessary. You will gain marks for clear visualizations.
- For answers involving numerical values, use correct units where appropriate and format floating point values to a reasonable number of decimal places.

## Submission

Your submission should be in the form of a Jupyter notebook including your code, the visualizations you created for each question, and your analysis. The notebook will read the given data and generate the requested visualizations - it is not necessary to show the exploration of the data. Please submit the notebook in three formats: ipynb, html, and pdf.

## Coursework questions

In this coursework, we will ask you to analyse and visualize two datasets. The coursework includes data from Japanese restaurants that was collected from multiple review sites.

The Japanese Restaurant Review Dataset includes reviews from the following websites:

- [Hot Pepper Gourmet](#) (**hpg**): similar to [Yelp](#), here users can search restaurants and also make a reservation online.
- [AirREGI](#) / Restaurant Board (**air**): similar to [Square](#), a reservation control and cash register system.

The data includes the ID of the restaurant, the date and time of visits and reservation-making, restaurant location in Japan, and restaurant type. You will find explanations about the files and variables in them in the “Information\_On\_Files.txt” file in the files you download in step 1 below.

In this assignment you will need to obtain datasets from different sources, preprocess the data, and generate visualizations that show interesting patterns.

Please proceed as instructed below:

1. Download the data sets from <https://github.com/Inf2-FDS/fds-coursework1>  
If you are using Noteable, you can clone this repository into Noteable as you have done with the Lab exercises. If you are not using Noteable, please do not fork the repository, leave the forked repository public, and thereby allow others to see your solutions.
2. (20 marks) The following question relates to customer visits in 2016 to restaurants that appear in both **hpg** and **air** datasets (see `store_id_relation.csv` file). Which dataset generated more visitors to these restaurants? Please visualize your answer. Note we are referring to restaurant visits and not reservations.
3. (25 marks) The following question relates to restaurants in the **air** dataset. Use a scatterplot to visualize restaurants by location and type of restaurant (you might want to add a small random noise to the restaurant location to reduce overlap in the plot). You can choose how to represent the restaurant location (e.g., longitude/latitude, distance from city centre, or other). Note there are many restaurant categories. You should

collapse the different categories to 4 or 5 categories based on your best judgement (e.g., Asian, International, Bar and party, Cafe and sweets). What can you infer from the plot you created about the relationship of the restaurant categories and their location?

4. (15 marks) The following question relates to the **air** data set. Plot the trend of the mean number of visitors to restaurants vs. a time unit of your choice (e.g, daily, weekly, etc...). What is the trend that you see?
5. (15 marks) The following question relates to the **hpg** data set. Create a histogram of the number of visitors per day of the week (Sunday, Monday, ..., Saturday). What can you infer from this graph?
6. (25 marks) The following question relates to the **air** data set. We wish to determine which restaurant genre Japanese people are most passionate about. To this end we will analyse how much time people plan ahead before visiting a restaurant.
  - Compute the time difference between reservation time and visit time.
  - Compare the time differences among restaurant genres using a visualization of your choice. Tip: In order to avoid outliers, it might be best to choose an upper threshold for values of preparation time. You may want to use the genre categories used in question 3 to reduce clutter.