Inf 2 - Foundations of Data Science
Task: Preparation for S1 Week 5 Workshop – Visualisation

In the S1 Week 5 workshop, we will explore the design and evaluation of effective data presentations. The goal of this workshop is twofold:

1. To familiarise yourself with the principles and guidance used for data visualisation in the course, and which we will use when marking Coursework 1.
2. To learn how to evaluate and critique (in a constructive way) visualisations prepared by others.

To this end, we will give you a chance to peek at some of the students' submissions of CW1 from 2020/21. (Don't worry, this year's CW1 is different, so you won't be biasing yourself towards a particular solution).

Do this individually in Week 4

- Read chapter 1 ("Data Visualizations: A primer") from the Big Book of Dashboards by Wexler, Shaffer and Colgreave (2017). An online version of the book is available in the library. You can find it from the FDS course pages as follows:
    - Click on the course page Resource List
    - Click on Informatics 2 - Foundations of Data Science Resource List
    - Click on Essential
    - Click on Book chapter: Data Visualization: Primer
    - Select either the Wiley or Proquest Ebook option
    - If you've selected Wiley, you should find a direct link to the PDF of chapter 1 in the table of contents; if you've selected Proquest Ebook, you'll need to click on Part I: A Strong Foundation to find it.
- Read the principles and guidance for creating visualisations in the lecture notes, which are summarised in the Visualisation principles and guidance handout sheet available from the task page.
- Read the description of Coursework 1 from 2020/21 (below).
- Go to **Learn →Assessment →Examples of Previous Assessment → Coursework 1 - Data wrangling and visualisation** and find sample submissions of three students from 2020/21 (we've received their permission to share their work).
- Try assessing Question 6, Question 3 and Question 4 in each submission using the marking spreadsheet `FDS-S1-05-visualisation-2-mark-sheet-individual.xlsx` available from the task page to record your marks. For each question, there are visualisation criteria corresponding to the visualisation principles. You should assess how well the visualisation in each question meets the principles on a scale of 0 (absent) to 4 (excellent). We've also given you space to assess the quality of the explanation and the readability of the code. You can write comments about why you awarded that number of points too. We will *not* be assessing your marking – the point is for you to learn about what makes a visualisation good.
- In the workshop, we will ask you to compare your marks with the marks given by other students at your table.

CW1 questions (FDS 2020-21 – not this year's 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 the data in the hpg_reserve and air_reserve datasets. We wish to compare customer visits in those restaurants that appear in both hpg and air datasets (see store_id_relation.csv file). Which dataset generated more visits to these restaurants in 2016? Please visualize your answer. To determine whether a reservation was made for 2016 use the "visit_datetime" field in the respective reservation file hpg_reserve and air_reserve.
3. (25 marks) The following question relates to the data in the air_store_info 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_visit 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_reserve data set. Create a visualization for describing 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_reserve and air_store_info 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.
   a. Compute the time difference between reservation time and visit time.
   b. 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.

Some Clarifications:

- You can assume people show up for their reservations.
- If a reservation is made for n people on a given day, you can assume that n visits occurred on that day. In Q2, we care about the visit dates (when people actually came to the restaurant, not when they made the reservations), so use the "visit_datetime" field