UNIVERSITY OF LEEDS

School of Computing
Faculty of Engineering

Assessed Coursework Header/Feedback Sheet

This form should be completed and submitted with each piece of assessed coursework.

My signature here confirms acceptance of the declaration of academic integrity printed overleaf.	
Student Name:	Shreyas Honnalli Abhishek Paradkar Chemseddine-Wassim Benimoussa Amazigh Bouldjenet
Signature:	SHREYAS HONNALLI ABHISHEK PARADKAR CHEMSEDDINE-WASSIM BENIMOUSSA AMAZIGH BOULDJENET
SID Number:	201392806, 201469616, 201429246, 201398339

You must make sure that you have read the declaration of academic integrity.

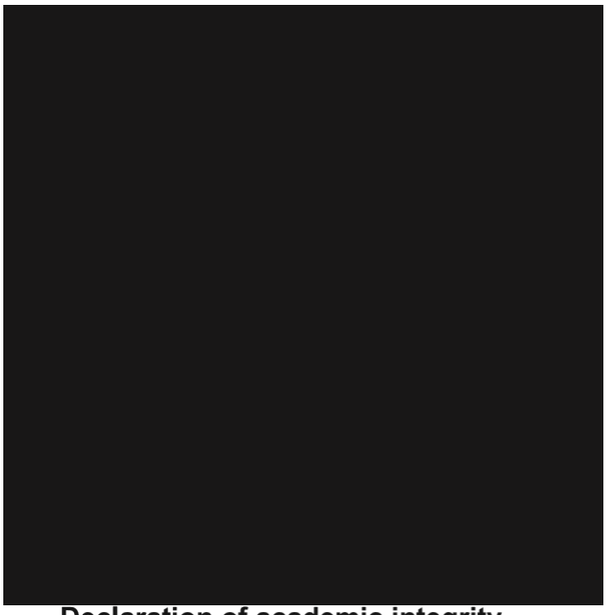
To be completed by the student submitting work
(please make sure that you complete ALL fields)

Module Code	C O M P 3736
Module Title	Information Visualisation
Coursework Number	2
Name of Lecturer Marking Work	Roy Ruddle ****You must fill this in****
Deadline Date	15/12/22
Date Handed In	14/12/2022

Feedback section to be completed by the marker

Section	Marks Available	Marks Awarded	Comments on sections
Total			
Late Penalty (marks to be deducted)			

General Comments on Work



Declaration of academic integrity

Further information about plagiarism, fraudulent or fabricated coursework and malpractice in University assessments is available at:
http://www.leeds.ac.uk/aaandr/cpff.htm

I am aware that the University defines plagiarism as **presenting someone else's work, in whole or in part, as your own.** Work means any intellectual output, and typically includes text, data, images, sound or performance.

*(On the understanding that other members of the group have made contributions to the attached submission,)** I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

**For the submission of group work*

Penalties for late submission

University rules on penalties for late submission of coursework require 5% of the total marks available to be deducted for each calendar day that passes after the date of required submission. The deduction will be applied to the grade/mark for the coursework component concerned before any conflation with other grades/marks to give the overall result for the module. If coursework is not submitted by the end of 14 calendar days following the prescribed deadline, a grade/mark of zero should be returned for that component.

Student IDs: 201392806, 201469616, 201429246, 201398339

Coursework 2: Comparison of Map and Bar charts in expressing plastic pollution of different countries

I. Introduction:

Effective and interactive visualisations help users to explore large datasets, identify patterns and outliers, and answer questions about the dataset. It plays a pivotal role in analysis of data and investigations of hypotheses. The aim of this report is to design a laboratory experiment that compares the effectiveness of bar graphs vs map charts for displaying data related to plastic pollution from different countries by allowing the user to read and interpret randomised data visualised by these graphs. The domain of countries chosen for this experiment is from the Western Europe region: United Kingdom, Ireland, France, Spain, Portugal, Italy, and Germany. Participants that take part in this experiment would be presented with a random series of Multiple Choice Questions(MCQs), each question having attached a randomised bar chart or a map describing an aspect of the chosen topic (plastic pollution). Each MCQ is a question that can be answered only by reading and interpreting the data given in the visualisation associated with it. There are 10 questions in total, each of which are asked twice to the user in a randomised order – once for the bar chart and once for the map. This allows us to measure the time taken by the user to answer the same question using different forms of visualisation, each with randomised data, essentially giving an insight into the effectiveness of bar charts over maps or vice versa in conveying certain forms of data. The software designed to conduct this experiment implements a command line interface using python along with several different libraries such as Matplotlib and Plotly used for plotting and displaying the visualisations.

This report outlines the experiment design, i.e., the software implementation, the background of participants that took part in the experiment, the method and conditions in which the experiment took place, the results of the experiment, i.e., the average time taken as well as the average score (right vs. wrong) to answer each of the questions and for each chart type shown.

II. Experiment:

A. Participants:

The experiment will be tested on 10 participants in total, each of whom will undertake the quiz once. The 10 participants will be shown a consent form (see Appendix) which indicates that the participants are consenting on taking part in the experiment which assesses the effectiveness of different visualisations and that each one of them can withdraw from the experiment at any given time. In addition to the informed consent the participants are given the date of the experiment, duration of the experiment (depends on how long they take to answer the 20 questions, but in practice does not take longer than 15-30 minutes), and the structure of the experiment (20 randomised multiple choice questions, being shown 20 graphs via a python script). All the participants of this experiment are volunteers which means that no money has been spent on finding the volunteers or to get them to take part in the experiment. This is done to reduce bias. We conclude in more accurate results this way as there is no bias due to the influence of money.

This experiment uses a “within-participants” design: each participant is exposed to data represented in both conditions - bar and map visualisations (each participant answers all questions). The reason for this is to eliminate as many uncontrollable variables as possible. If using a “between-subjects” design i.e. different participants being exposed to different visualisations (for example, in a single

experiment, 1 participant being exposed to bar graphs only and another participant being exposed to maps only), external variables such as participant IQ, experience with the specific visualisation, and visualisation perspicacity would have varied for any single experiment. Using a "within-participants" design we have controlled the variability because of the following-

- The participant answering questions represented by bars will have the same IQ as the participant answering questions represented by maps (as it is the same person answering both question types).
- In order to mitigate the possibility that a participant has little experience with one of the visualisations and a familiarity with other visualisations (for example, user may have experience with bar graphs but may not know what a tree map is, leading to longer response times and higher incorrect answer frequencies for questions with data represented by tree maps. Longer response times in this scenario are not because tree maps are a poor way of representing the data, but simply because the user has no prior experience with tree maps, leading to skewed results), we ensured prior to the experiment that the participant knows each of the visualisation chart types in the experiment. If they do not know, multiple examples of the unknown chart type is shown with example data, in order to ensure they are familiar with the graph structure and how to infer data from each graph.

B. Design:

We have used a python language command line script as a framework for our experiment. In order to display our bar graph and the different kinds of maps (including heat maps, tree maps, choropleth maps, heated tree maps, etc) we have used four python libraries. We have used plotly.express library in order to display the heated tree maps and the tree maps. In order to compute and display our choropleth maps, we have used plotly.graph_objs library. Plotly is a company based in Montreal that has built the Plotly Open-Source Graphing Library for Python. This library allows us to make interactive graphs which are displayed on the web browser. The graphs made from this library display on a web browser. A strong advantage of using plotly graphs is interactivity with visualisations. Hovering over elements of the visualisation allow us to see additional information about specific labels such as actual number, what it represents, etc. We took full advantage of this feature in plotly by asking the user questions which can only be answered by using interactive data. The additional information can also be used to help the user answer questions indirectly. In order to display our heat maps we have used seaborn module. Seaborn is a python data visualisation library based on matplotlib. Seaborn displays its graphs on an external interactive window. Finally in order to display all of our bar graphs we have used matplotlib. Matplotlib does not use a web browser in order to display its graphs but also opens up a new interactive window, allowing users to zoom in and zoom out on graphs. Matplotlib is also a python data visualisation library allowing us to visualise data easily.

In order to help us record the amount of time a user has taken in order to answer a question, we have imported the in-built python time module, which has a large amount of functions related to computational time. To generate the randomised data for each visualisation, we have also used another python library – random. The in-built random library for python has a huge number of functions related to pseudo-randomness. The numbers generated by module are considered pseudo random as they are not truly random (this is because computers are inherently deterministic. Pseudo random number generators generate the same sequence of numbers for a given seed), but for the purposes of our experiment the module will suffice.

The software works as a python script which is run on the command line. We first go to the file location where all the software files are stored in the command line and then type in “python ui.py”. This script starts the program and the questionnaire, where the participant will be prompted to press enter to start the questionnaire. This is illustrated by figure 1, which shows the user interface of the software after starting the program.

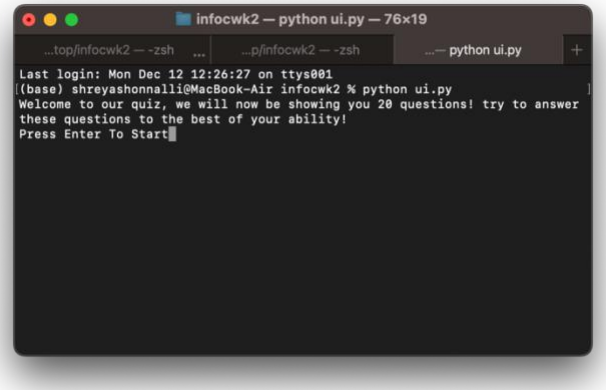


Figure 1: User Interface of the software upon starting the questionnaire program.

Upon pressing enter, the participant is prompted with a question, and the corresponding graph is displayed on an external window or the web browser installed on the computer depending on the library used to display the visualisation. After answering the question, the graph (which is open in an external window) is automatically closed. A one second time delay will come into effect after which the user is asked the next question, and the corresponding graph of the next question is opened on an external window. An example of this is shown in figures 2 and 3. Figure 2 shows an example of the user interface of the software after the user has pressed enter and has answered several questions. Figure 3 shows us the corresponding graph for the current question. In the scenario of question 2, the question at hand involved a visualisation from the library plotly, hence it was displayed on the web browser of the computer.

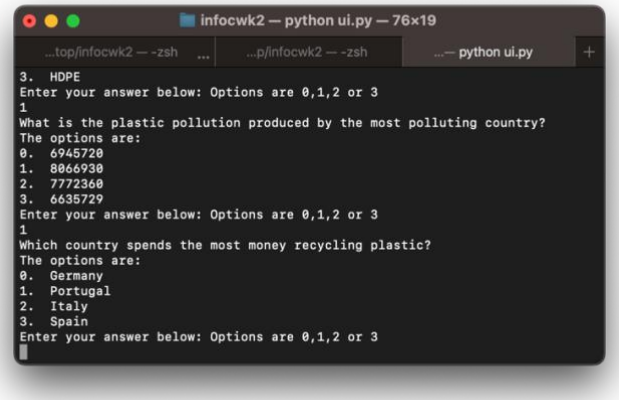


Figure 2: User Interface of the software after pressing enter and after user has answered several questions.

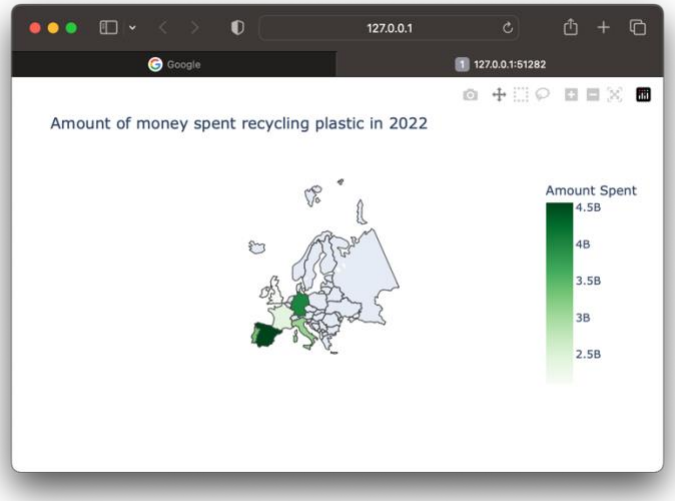


Figure 3: Graph displayed in external window corresponding to question from figure 2.

C. Procedure:

The procedure of each experiment is simple: The experiment commences once the participant has been shown the consent form and has given the approval. Each participant will run the above python script containing the questionnaire once and their answers are recorded after completing the questionnaire. The script contains 10 unique multiple choice style questions. Within the script, each participant will be asked each unique question twice in a randomised order, the difference between these two attempts being the way the data is represented (one attempt for bar, other for map). The data within these graphs are randomised and generated using bespoke implemented random number generators (as described in method) and will not be the same for any two given graphs. This way the participant must visually identify and infer data required to answer the questions and cannot answer correctly from memory of previous questions. They will repeat this process of answering the same question twice(for two different chart types) 10 times, summing up to 20 questions in total. The order of the 20 questions is randomised each time the python script is ran. After they have answered all the questions and pressed enter, the user's results – incorrectly/correctly answered questions as well as time taken on each question - are stored in a csv file and they are exited from the application.

In context of the overall procedure, we have run each experiment ten times in total - once for each participant. Each experiment records the time taken to answer each question, the type of map each question was representing and whether the participant got the answer to the question correct. Upon obtaining this data in a csv file, we can look at the average score of each question, the average amount of time taken to answer that particular question, and the corresponding chart type. We can use this data for all questions to perform empirical analysis on the effectiveness of the chart type by looking at the average times and average score of each chart type.

III. Results:

The questions asked in the experiment are ordered on the result tables as follow:

1. What is the plastic pollution produced by the most polluting country?
2. Which country pollutes the most plastic
3. What is the plastic pollution produced by the least polluting country?

4. Which country pollutes the least plastic?
5. Which kind of plastic waste is most found in United Kingdom?
6. Which country has the best plastic pollution to emission ratio?
7. Across all years, how much pollution is produced by the most polluting country?
8. Which country has the best ratio of recyclable plastics to non-recyclable plastics?
9. Which country generates the most plastic waste per person?
10. Which country spends the most money recycling plastic?

Results for average time taken for each of these questions and for each chart type shown (seconds):

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Mean	SD
Bar	21.89	11.54	24.65	11.95	26.90	34.05	28.28	31.46	15.21	12.50	21.84	8.07
Map	18.85	16.62	17.20	18.21	47.44	58.29	60.43	41.97	16.65	18.60	31.42	17.51

Results for average score for each of these questions and for each chart type shown (between 0 and 1):

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Mean	SD
Bar	0.6	1	0.6	1	0.6	0.8	0.7	0.9	0.8	1	0.8	0.17
Map	1	1	1	0.9	0.6	0.6	0.2	0.7	0.9	0.9	0.78	0.257

Observations:

- From the average score table for each question, we can notice that scores for question 1 and 3 are higher in maps than bars. Question 1 and 3 are both about reporting the numerical value of a certain polluting country. We can see that the time lapse of answering these two questions is higher in bars than maps. The reason for which map charts give better scores and time lapse than bar charts is related to the interactive element of plotly. The hovering element of plotly enabled participants to obtain exact numbers.
- For questions 2 and 5 we can see that the scores are the same for both type charts even though they are low scores for question 5. Question 5 is about finding what type of plastic is found in a specific country. When considering the time lapse table and the result for question 5 we can see the time lapse for bars is almost half the time that it took for maps which means that when handling a visualization that compares different types of variables it is faster to get the idea by using bar chart rather than maps.
- Apart from question 1, 3, and 5 all the scores are higher in bars than maps for the rest of the questions especially in questions that are related to finding the best ratio (for example see question 6 and 8). This is because spotting the ratio between two variables from stacked bars is much easier than doing from map charts as it can be done just by looking at the size of each of the stacked bar while comparing them, but for maps it would be harder to do the same.
- Overall, the time lapse to answer all the questions apart from question 1 and 3 were all lower when visualizing bar charts as the mean for the time laps considering bars is 21.84 seconds comparing to considering maps is 31.42 seconds.

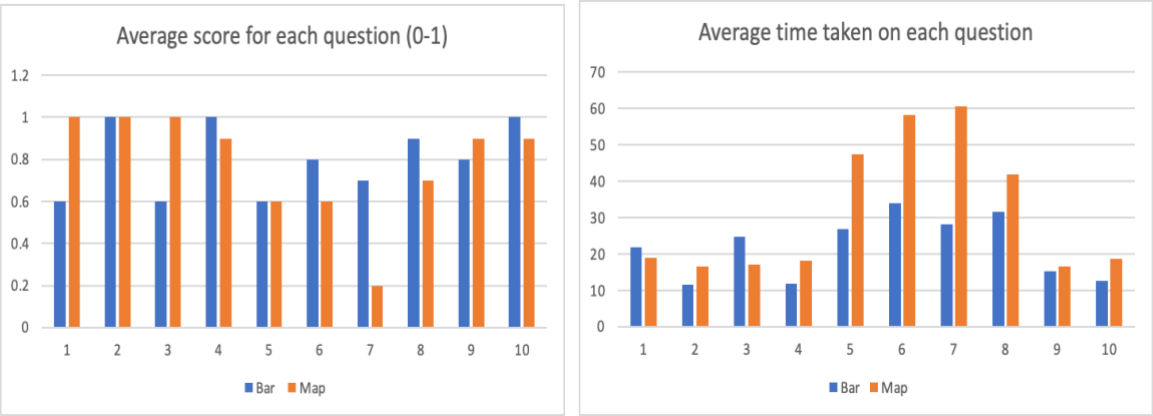


Figure 4.1 and 4.2: Bar graph of average score of 10 participants for each question and Bar graph of average time taken of all 10 participants for each question

T-tests Results:

Correctness T-test:

After analysing the difference between Bar charts and Map charts in term of correctness of the results by using T-tests, there was not enough evidence to suggest that the mean score of each visualization is truly statistically significant :
(t = 0.2308 (degrees of freedom = 9); p-value < 0.8227; standard-error of difference = 0.087)
As shown from the results the standard error and t value are quiet low which indicates that the two chart types are not that different in term of correctness.

Time Taken T-test:

In the other hand, T-tests results related to the difference between bar charts and Map charts in term of time lapse to answer each question gave statistically enough evidence to suggest the mean response times for the visualizations are significant :
(t(10) = 2.6820(degrees of freedom = 10); p-value < 0.0230; standard-error of difference = 3.573)
As shown from the results the standard error and t value are high which indicates that the two chart types are different in term of time taken to answer the questions.

IV. Conclusion:

- The level of understanding of the data gained from viewing bar charts and map charts is statistically the same when the data is related to plastic pollution, however it takes the viewer a statistically longer amount of time in order to grasp the information when presented in the form of tree maps, choropleth maps or heat maps.
- This is most likely because audiences can read and understand a bar graph and grasp the information it conveys. Audiences are not familiar with various different map forms.
- Reading the exact value of each bar in a bar graph without labelling value of bars with the exact value is difficult. Questions that required the volunteers to retrieve the precise value such as Q1 and 5, mean correctness was low.
- Some types of maps such as tree-maps and heat-maps are not only more complex to understand without additional explanation, but they are also area-based data visualization that can be hard to interpret precisely. Viewers understand data quicker when represented by height over area.

V. Appendix:

Information Sheet: Visualization Experiment

Students: Shreyas Honnalli, Abhishek Paradkar, Chemseddine-Wassim Benimoussa, Amazigh Bouldjenet

Address: School of Computing, University of Leeds

We are students at the University of Leeds, interested in how to design visualizations of descriptive statistics. If you volunteer for the experiment you will perform a series of 20 trials. In each trial you will be prompted to answer a multiple choice question where the data for the answer of the question will be inside a graph corresponding to each question. The experiment is a questionnaire which will be conducted on a computer platform through a command line script. The order in which you perform the trials is randomised. Each trial will show a visualisation and a corresponding multiple choice question with 4 options, and your task is to answer the question with the available information from the visualisation to the best of your ability. If you are not able to provide an exact numerical answer to the question, please calculate an estimate and choose the option which is closest to that estimate. You are also able to obtain additional information about each graph by hovering your mouse over the window in which the graph is being displayed in.

The experiment will take approximately 15 to 30 minutes. Your responses and time you take to answer each question will be recorded, but you will remain anonymous. The research may be reported in student work, but no-one should be able to identify you and at no point will your identity be divulged.

By taking part in the experiment you are indicating informed consent for us to use your data and . You are free to withdraw from the experiment at any time.

Finally, please let us know if you have any questions or would like to discuss anything with us regarding data collection and/or usage.