Assignment 2

11520G Data Capture and Preparations

Due date: 11:55pm, 7 May

- Laboratory work forms an essential part of the Unit, and as such, students are expected to complete and submit the scripts on time.
- It is a requirement to achieve **at least 25% marks** in this assignment.
- This assignment is divided in three parts, in the first part you will collect data from a website; in the second part, you will clean and wrangling the data; and in the third part, you will answer some questions using the data.
- Remember, **plagiarism is a serious offense** and it will not be tolerated.

1.1. Submission

Each team is expected to submit the code (R script) from their project. Your submission should include:

- You should submit in a .zip folder (named: "yourteamname").
- A description of how the project was divided between the members (including full name and student ID) of the group (who is responsible for what part of the code).
- Consistent indentation.
- Comments thought all the code.
- Your code should not include hard-coded directory paths (e.g., C:\Raul\MyFolder\Code\), except from any folder within the same submitted folder.
- A short description (as a comment in your code) of how each question was solved is required at the beginning of each question.
- Number your answers using the number for each question and separate each question e.g., ********** Question 1 **********.
- As University Policy, a 5% penalty per day is applied for late submissions.

Failing to do any of the previous points will result in **5% penalty deduction** for each of the missing points not included. For example, code with no consistent indentation (-5%), code with no comments (-5%), and late submission (-5%): a 15% penalty is applied to your group.

All deliverables should be submitted in Canvas **by a single member of your team**. Individual work is not allowed, work needs to be done within a team.

1.2. Peer Review

This is a group assignment, thus, the final mark for this assessment will be based on a peer assessment from each member of the team using the following method:

Each member will be asked to evaluate peers' individual contributions to group work. For each student, a peer assessment factor (PAF) is calculated based on their contribution to the group assignment. The following table shows the individual PAF values and their interpretation.

PAF	Interpretation	Comments
>1.5	Alarm! Team failure	Something has gone wrong – either there is a student who is not participating at all or this student has taken all the work home and done it by themselves. Either way, learning objectives are probably not being achieved.
1.15 - 1.5	Super Leader	The team balance probably needs to be addressed as to achieve such a high score other students must not be participating, or this student is doing far too much.
1.05 - 1.15	Leader	The student is showing definite leadership qualities and/or has been putting in significant extra effort.
1.00 - 1.05	Good teamwork	The student is working well with the group and has been recognised as pulling their weight (1.00) and perhaps a little more (>1.00).
0.95 - 1.00	Acceptable teamwork	This student has probably only been penalised because another team member has shown leadership and put in extra effort.
0.85 - 0.95	Social Loafer	Any PAF below 0.95 is unacceptable. Social loafers who lie in this band can usually be mentored with the group's help and become productive members of the group.
0.75 - 0.85	Super Social Loafer	As above and below.
< 0.75	Alarm! Individual failure!	The individual is in grave danger of failing the course. Much work is required for this student to be accepted back into the group and there will be trust issues with allocating this student any work.

Reference: https://elearning.uq.edu.au/guides/group-peer-assessment/paf-formula-and-moderation-overview

For example, if the group assignment receives a mark of 80 out of 100 and each student receive an average PAF from the members of the team of:

	Peer Reviews from			
	Student A	Student B	Student C	Average
Student A	1.1	1.0	1.1	**1.066**
Student B	0.9	1.2	0.9	**1**
Student C	1.0	0.8	1.0	**0.933**
Total available marks	3.0	3.0	3.0	3.0

Therefore, the final mark for each student will be:

- Student A: 1.066 X 80 = 85.28
- Student B: 1 X 80 = 80
- Student C: 0.933 X 80 = 74.64

If the group is formed by 4 people, the total available marks will be adjusted to be 4.0.

Each member of the team must submit their peer assessment in the canvas site for the Assignment 1 before the submission due date. If no submission is received, a 1.0 PAF mark will be used as your peer review.

Overview

In this assignment, you will be able to demonstrate your ability to integrate and synthesize the knowledge acquired throughout the unit by understanding the basic concepts of data capture from online sources (i.e., web scraping), and cleaning and wrangling data using R. This assignment is divided into three main parts:

- First, you will need to collect the information using web scraping techniques learned in this unit.
- Second, you will need to clean and wrangle the scraped data.
- Third, answer research questions using R.

You will need to collect specific data from the United Nations Educational, Scientific and Cultural Organisation (UNESCO). This organisation supports the preservation of the world's natural and cultural heritage. As of today (April 2021), there are 1121 heritage sites around the world, most of which are man-made like the Acropolis of Athens or the Taj Mahal in India, but also natural sites like the Great Barrier Reef here in Australia are listed. Unfortunately, some of the sites in the list are threatened by human intervention or natural reasons. We want to know more about the endangered sites, their location, and the reasons that put a site at risk.

For this assignment, we will use the website of the list of currently and previously endangered sites in Wikipedia. This site can be accessed at: https://en.wikipedia.org/wiki/List_of_World_Heritage_in_Danger. You will the table with the current endangered sites, which contains information about the sites and their reasons to be in the list, e.g., their name, location, type of threat that is facing the site. Now, let's collect some specific information about the sites to answer some specific questions.

You can complete this assignment during the computer labs of week 12 and week 13. During these sessions, you will be able to ask your tutor questions regarding this assignment. You can also use the drop-in session to ask questions regarding the assignment.

Part 1. Web Scraping – 30 points

In this first part, you need to scrape the website using R. You should apply the concepts learned during the lectures and tutorials/labs. It is desirable to use functions/methods employed in this unit. Used the above url for this part.

1. Retrieve and load all the data from the url into R. **(2 points)**
2. Obtain the table legend (Figure 1) and store all its elements. You can use any data structure to store the data. **(6 points)**


Table legend

Name: as listed by the World Heritage Committee[16]
Location: at city or provincial level and country name, with coordinates; column sorts by state[nb 2]
Criteria: the site was listed under
Area: in hectares and acres if provided by UNESCO
Year (WHS): the year the site was inscribed on the World Heritage List
Endangered: the year the site appeared on the List of World Heritage in Danger
Reason: threats to the site which prompted UNESCO to list it as in danger

Figure 1. Table listing the legends used in the currently listed sites.

3. Scrape the endangered list, which contains the current listed sites. You can use any data structure to store the table. **(6 points)**
4. Scrape all available hyperlinks in the url. **(4 points)**
5. Using computational methods, obtain the hyperlink that you can click on to jump to a new page that contains the selection criteria used to classify a site as cultural or natural heritage. **(8 points)**
6. Use the hyperlink obtained in the previous step and scrape the two lists (cultural and natural) and store them in two separated data structures within a list. **(4 points)**

Part 2. Data cleaning and wrangling – 40 points

In this second part, you need to clean and wrangle the data. You should apply the concepts learned during the lectures and tutorials.

1. From the table containing the endangered sites, remove the undesired variables: Image and Reds. **(2 points)**
2. Then, obtain the country from the "Location" variable. Using computational methods (e.g., Regex) fix any inconsistencies in the variable and then to extract the country only. **(14 points)**
3. Using computational methods (Regex), split the variable that contains the criteria ("Criteria") into two variables: "Type" (cultural/natural) and "Criteria" (containing roman numbers). **(6 points)**
   *** For those aiming for Credit: ***
4. Then, maintain only the data in acres and remove the hectares (ha) from the "Area" variable. Remove any extra characters from the numbers. **(8 points)**
5. Using computational methods (Regex), clean the variable Endangered and maintain only the very last year. Remove any unwanted characters. **(8 points)**
   *** For those aiming for Distinction: ***
6. Make sure that you have numeric vectors and characters vectors only. **(2 points)**

Part 3. Data Analysis – 30 points

In this last part, you need answer the questions using the scraped data in the previous sections.

1. What type of site (cultural or natural) is the most common in the endangered list and how many does each type of site have? **(3 points)**
2. What site has the largest area (in m^2) and what site has the smallest area (in m^2)? (**3 points)**
   *** For those aiming for High Distinction: ***
3. What is the frequency (in years) with which sites were put on the endangered list? For example, how many were put on the list between 2010 and 2015? Use a plot to answer this question (e.g. histogram), remember to label and title you plot correctly. **(6 points)**

4. What country has more sites in the list? and how many sites has each country in the list? **(6 points)**
5. How long took each site to be in the endangered list? **(6 points)**
6. What is the frequency with which sites were put on the endangered list after they were inscribed in the World Heritage List? For instance, how many sites were in the endangered list after 3 years in the World Heritage list. Use a plot to answer this question. **(6 points)**

If you cannot achieve the outcome using the expected computational methods, a 30% penalty will be applied to the weight of that question.