Cardiff School of Computer Science and Informatics

Coursework Assessment Pro-forma

Module Code: CM2105
Module Title: Data Processing and Visualisation
Lecturer: Dr Hantao Liu
Assessment Title: CM2105 Coursework
Assessment Number: 1
Date Set: 9th November 2020
Submission Date and Time: 14th December 2020 at 9:30am
Return Date: 22nd January 2021

This assignment is worth **100%** of the total marks available for this module. The penalty for late or non-submission is an award of zero marks.

Your submission must include the official Coursework Submission Cover sheet, which can be found here:

https://docs.cs.cf.ac.uk/downloads/coursework/Coversheet.pdf

Submission Instructions

Your coursework – your **code and results** should be contained within **an executed Jupyter Notebook named "CW_ your student number.ipynb"** – should be submitted via Learning Central by 9:30am on the submission date.

Description		Type	Name
Cover sheet	**Compulsory**	One PDF (.pdf) file	student number.pdf
Q1	**Compulsory**	One Jupyter Notebook (.ipynb) file	CW_student number.ipynb

Any deviation from the submission instructions above (including the number and types of files submitted) may result in a mark of zero for the assessment or question part.

Assignment

Q1. Part1:
The QS World University Rankings are a ranking of the world's top universities published annually since 2004. Along with Academic Ranking of World Universities and THE World University Rankings, the QS World University Rankings is widely recognised and cited as one of the three main world university rankings. Universities are evaluated according to the following six metrics:

– **Academic Reputation**
– **Employer Reputation**
– **Faculty/Student Ratio**
– **Citations per Faculty**
– **International Faculty**
– **International Student Ratio**

The Microsoft Excel file named "**2018-QS-World-University-Rankings-Top200.xlsx**" (available on Learning Central) contains the data of the top 200 universities in the world.

Institution Name	Location	Rank	Academic Reputation	Employer Reputation	Faculty Student	Citations per Faculty	International Faculty	International Students	Overall Score
MASSACHUSETTS INSTITUTE OF TECHNOLOGY (MIT)	United States	1	100	100	100	99.9	100	96.1	100
STANFORD UNIVERSITY	United States	2	100	100	100	99.4	99.6	72.7	98.7
HARVARD UNIVERSITY	United States	3	100	100	98.3	99.9	96.5	75.2	98.4
CALIFORNIA INSTITUTE OF TECHNOLOGY (CALTECH)	United States	4	99.5	85.4	100	100	93.4	89.2	97.7
UNIVERSITY OF CAMBRIDGE	United Kingdom	5	100	100	100	78.3	97.4	97.7	95.6
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…

1) [**cell1 – 2 marks**] Download the file "CW_your student number.ipynb" from Learning Central, and upload it to your Jupyter Notebook. Change the title of the file using your student number. Write code to read the given data (i.e., "2018-QS-World-University-Rankings-Top200.xlsx") into required tabular data structure (i.e., a DataFrame): make the "rank" (i.e., 1, 2, 3…200) be the index of the returned data structure; the first column represents the "Institution Name"; the second column represents the "Location"; the third to eighth columns represent the six QS metrics; and the last column represents the "Overall Score".
   – **Display** the returned tabular data structure in your programme (see Sample output 1). [2 marks]

2) [**cell2 – 2 marks**] Write code to create a **QS-UK-rankings** (i.e., a DataFrame) that lists all UK universities in the top 200 in the QS-World-University-Rankings. In the returned tabular data structure (i.e., a DataFrame): make the "national rank" (i.e., 1, 2, 3…) be the index; the first column represents the "Institution Name"; the second to seventh columns represent the six QS metrics; and the last column represents the "Overall Score".
   – **Display** the returned tabular data structure in your programme (see Sample output 2). [2 marks]

3) [**cell3 – 5 marks**] Write code to analyse the data contained in the variable called "**Academic Reputation**" for the **UK universities in the world's top 200**. Quantitative results should be shown as rounded values with TWO decimal places.
   – **Print** the "mean of Academic Reputation". [1 mark]
   – **Print** the "minimum of Academic Reputation". [1 mark]
   – **Print** the "maximum of Academic Reputation". [1 mark]
   – **Print** the "standard deviation of Academic Reputation". [1 mark]
   – **Print** the "95% confidence interval of Academic Reputation". [1 mark]

4) [**cell4 – 9 marks**] Write code to plot a bar graph that uses bars to compare the "**Citations per Faculty**" of universities in the United States, United Kingdom, Canada, Australia, Netherlands and Germany.
   – **Visualise** a single plot: the horizonal axis shows the data categories being compared (i.e., United States, United Kingdom, Canada, Australia, Netherlands and Germany); and the vertical axis represents the mean measure of the "Citations per Faculty". [3 marks]
   – Add **error bars** to the bar graph, showing the 95% confidence interval. [3 mark]

– Add appropriate **title**, **horizontal axis label** and **vertical axis label** to the bar graph. [3 marks]

5) [**cell5 – 7 marks**] Write code to perform appropriate statistical data analysis to reveal the difference in the "**Citations per Faculty**" of universities between the United States and Netherlands.
– **Print** the **name(s)** of chosen test(s) and **results** of test(s). [5 marks]
– **Print** ONE sentence, stating your **conclusion** and **justification** on the observed difference between two locations. [2 marks]

Q1. Part2:

You are given two sets of grayscale images, i.e., "m1.png" – "m10.png" in data file "model.zip" and "t1.png" – "t10.png" in data file "test.zip". Shown below are examples of images, i.e., "m1.png" and "m5.png". Each image is an array of integers; and the value (i.e., in the range [0, 255]) of each integer is the intensity at a pixel location. [Note: use code e.g., "***img=imread('m1.png')*255***" to read an image]
For each image, a true image quality score is given (see data file "Q_scores.xlsx").

m1.png

m5.png

6) [**cell6 – 8 marks**] Based on the **"model" dataset** (i.e., "model.zip"), write code to derive the correlation between the **"average pixel intensity"** and **"image quality"**. The "average pixel intensity" is quantified as the intensity value averaged over all pixel locations in an image.
– **Visualise** a scatter plot: the horizonal axis represents the "average pixel intensity"; and the vertical axis represents the "image quality". Add appropriate title, horizontal axis label and vertical axis label to the scatter plot. [5 marks]
– **Print** the Pearson linear correlation coefficient. **Print** one sentence to state your interpretation on the strength of the correlation. [3 marks]

7) [**cell7 – 14 marks**] Based on the **"model" dataset** (i.e., "model.zip"), write code to build a linear regression model to predict the **"image quality (note: use IQ as the name of target variable)"** from the **"average pixel intensity (note: use API as the name of predictor variable)"**.
– **Visualise** a single plot: the horizonal axis represents the "average pixel intensity"; and the vertical axis represents the "image quality". Add **data points** (representing individual images in the dataset) to the graph, using scatter plot. Add the **resulting linear regression model** (i.e., straight line). Add appropriate title, horizontal axis label and vertical axis label to the scatter plot. [6 marks]
– **Print** the resulting linear **equation** (i.e., regression model). Note, print the **equation ONLY**; use rounded values with TWO decimal places. [2 mark]
– **Print** the mean squared error (**MSE**) – the average of the squares of the errors – of the model on the **"model" dataset**. [1 mark]
– Write code to find the point on the regression line where the "image quality" score is **5.5** and rotate the straight line around this point by **two degrees** in clockwise direction. **Print** the "rotated" linear equation. Note, print the **equation ONLY**; use rounded values with TWO decimal places. **Print** the mean squared error (**MSE**) of the model on the **"model" dataset**. [5 marks]

8) [**cell8 – 13 marks**] After building the linear regression model based on the **"model" dataset** (i.e., "model.zip"), write code to evaluate its effectiveness on the **"test" dataset** (i.e., "test.zip"). This is to directly apply the above model/equation (i.e., the resulting model of **Q1.7), NOT the "rotated" version**) to predict the "image quality" of the **"test" dataset** (i.e., "test.zip").
– **Construct and display** a tabular data structure (i.e., a **DataFrame**), where the first column represents the **test "image"** (i.e., "t1.png" – "t10.png"); the second column represents the **"true quality"** (given by data file "Q_scores.xlsx"); and the last column represents the **"predicted quality"**. [5 marks]
– **Print** the mean squared error (**MSE**) – the average of the squares of the errors – of the model on the **"test" dataset**. [1 mark]

– **Print** one sentence to state the effectiveness of the model on the "test" dataset, by comparing the MSE on the "model" and "test" dataset. [2 marks]
– **Suggest and print** an improved model/equation (by adjusting the slope and/or intercept) for the "test" dataset. Note, print the **equation ONLY**; use rounded values with TWO decimal places. **Print** the MSE of the **improved model** on the **"test" dataset**. [5 marks]

Learning Outcomes Assessed

This assignment assesses the Learning outcomes 1-4 as stated in the module description.

Criteria for assessment

Credit will be awarded against the following criteria.

Your CODE and RESULTS should be contained within a Jupyter Notebook that analyses and visualises given data (should be obtained via Learning Central: CM2105 Data Processing and Visualisation). This coursework assesses the intended learning outcomes of 1, 2, 3, 4:

1. **Use Python to extract, manipulate, store and analyse information from a range of sources;**
2. **Understand statistical methods to apply to data;**
3. **Understand static visualisations of data;**
4. **Create static visualisations of data.**

Before you submit your Jupyter Notebook file, MAKE SURE you perform the following steps:
(1) Go to "Kernel", and perform "**Restart & Clear Output**";
(2) Go to "Cell", and perform "**Run All**";
(3) Carefully check the results/outputs of each cell, as they are the contents that will be marked.
Note: When marking your Jupyter Notebook submission, the module assessors will first perform steps (1) and (2), then start marking the results/outputs of all cells. It is your responsibility to make sure the **code** is **error free**. **THE PENALTY FOR UNEXECUTED CODE IS AN AWARD OF ZERO MARKS**.

The maximum mark for the coursework is **60** and the **mark obtainable** for a sub-question or part of a sub-question is **shown in brackets alongside the question**. A mark breakdown in terms of the 60 mark scale (rounded to 0.5 marks) is shown below.

Notebook cell	Maximum mark	1st	2.1	2.2	3rd	Fail
cell 1	2	>=1.4	>=1.2	>=1	>=0.8	<0.8
cell 2	2	>=1.4	>=1.2	>=1	>=0.8	<0.8
cell 3	5	>=3.5	>=3	>=2.5	>=2	<2
cell 4	9	>=6.3	>=5.4	>=4.5	>=3.6	<3.6
cell 5	7	>=4.9	>=4.2	>=3.5	>=2.8	<2.8
cell 6	8	>=5.6	>=4.8	>=4	>=3.2	<3.2
cell 7	14	>=9.8	>=8.4	>=7	>=5.6	<5.6
cell 8	13	>=9.1	>=7.8	>=6.5	>=5.2	<5.2

- **Fail**: the output of the code cell does not adequately address the stated requirement.
- **3rd**: the output of the code cell minimally addresses the stated requirement; for example, where multiple instances are required, at least one appropriate instance is provided.
- **2.2**: the output of the code cell partially addresses the stated requirement; for example, where multiple instances are required, a majority of instances are provided.
- **2.1**: the output of the code cell fully addresses the stated requirement, but has weaknesses in terms of the weakness indicators below.
- **1st**: the output of the code cell fully addresses the stated requirement, as well as meeting the excellence indicators below.

Weakness indicator: results are not presented in a professional and structured manner.
Excellent indicator: results are presented in a professional and structured manner. For example, when multiple instances are provided in the output, clear and concise descriptions are provided to help readers understand the meaning of each instance.

An indication of the level of attainment against the appropriate award is given below.

Undergraduate

1st (70-100%)
2.1 (60-69%)
2.2 (50-59%)
3rd (40-49)
Fail (0-39%)

Sample output of cell1

	Institution Name	Location	Academic Reputation	Employer Reputation	Faculty Student	Citations per Faculty	International Faculty	International Students	Overall Score
1	MASSACHUSETTS INSTITUTE OF TECHNOLOGY (MIT)	United States	100.0	100.0	100.0	99.9	100.0	96.1	100.0
2	STANFORD UNIVERSITY	United States	100.0	100.0	100.0	99.4	99.6	72.7	98.7
3	HARVARD UNIVERSITY	United States	100.0	100.0	98.3	99.9	96.5	75.2	98.4
4	CALIFORNIA INSTITUTE OF TECHNOLOGY (CALTECH)	United States	99.5	85.4	100.0	100.0	93.4	89.2	97.7
5	UNIVERSITY OF CAMBRIDGE	United Kingdom	100.0	100.0	100.0	78.3	97.4	97.7	95.6
6	UNIVERSITY OF OXFORD	United Kingdom	100.0	100.0	100.0	76.3	98.6	98.5	95.3

Sample output of cell2

	Institution Name	Academic Reputation	Employer Reputation	Faculty Student	Citations per Faculty	International Faculty	International Students	Overall Score
1	UNIVERSITY OF CAMBRIDGE	100.0	100.0	100.0	78.3	97.4	97.7	95.6
2	UNIVERSITY OF OXFORD	100.0	100.0	100.0	76.3	98.6	98.5	95.3
3	UCL (UNIVERSITY COLLEGE LONDON)	99.7	99.5	99.1	74.7	96.6	100.0	94.6
4	IMPERIAL COLLEGE LONDON	99.4	100.0	100.0	68.7	100.0	100.0	93.7
5	KING'S COLLEGE LONDON (KCL)	92.8	92.4	87.6	64.8	97.4	99.2	86.9
6	UNIVERSITY OF EDINBURGH	99.1	96.6	83.2	55.5	94.9	98.6	86.9

Feedback and suggestion for future learning

Feedback on your coursework will address the above criteria. Feedback and marks will be returned on 22nd January 2021 via Learning Central using a feedback form. If you have any questions relating to your individual solutions talk to the lecturer.