Module Code
COM00017C

BSc, BEng, MEng and MMath Degree Examinations 2020-2021

Department Computer Science
Title Introduction to Data Science
Issued 09:00am on Friday 28th May 2021
Submission due 09:00am on Saturday 29th May 2021
Feedback & marks due Monday 28th June 2021
Time Allowed 24 Hours (NOTE: papers late by up to 30 minutes will be subject to a 5 mark penalty; papers later than 30 minutes will receive 0 marks).

Notification of errors in the paper may be made up to **one hour** after the start time. If you wish to raise a possible error it must be done through ⟨cs-exams@york.ac.uk⟩ with enough time for a response to be considered and made within the first hour.

Time Recommended THREE hours

Word Limit Question 1 has **strict** limits on the length of your answers; text exceeding the limit will not be marked.

Allocation of Marks:

The exam consists of four (4) questions and each question is worth 25 marks.

Instructions:

Candidates should answer **all** questions using Jupyter Notebooks running Python 3. Failing to do so will result in a mark of 0%. All questions are independent and can be answered in any order. For programming questions, you may use additional Jupyter Notebook cells.

Download the paper and the Jupyter Notebook exam file (DAT1-Exam.ipynb) and the required datasets from the VLE, in the "Assessment>DAT1 Exam" section. If VLE does not work, you can find the material also on this Google Drive directory.

You must save the Jupyter Notebook file named with your student number (e.g., 123456789.ipynb) and upload the Jupyter Notebook file on the DAT1 Assessment submission point on the Department's Teaching Portal. Other formats such as .pdf and .html will not be accepted.

Submit your answers to the Department's Teaching Portal as a single PDF file.

If a question is unclear, answer the question as best you can, and note the assumptions you

have made to allow you to proceed. Please inform ⟨cs-exams@york.ac.uk⟩ about any suspected errors on the paper immediately after you submit.

Do not use colour: use black-on-white only, unless otherwise instructed.

Start each top-level question on a new page.

A Note on Academic Integrity

We are treating this online examination as a time-limited open assessment, and you are therefore permitted to refer to written and online materials to aid you in your answers.

However, you must ensure that the work you submit is entirely your own, and for the whole time the assessment is live you must not:

- communicate with departmental staff on the topic of the assessment
- communicate with other students on the topic of this assessment
- seek assistance with the assignment from the academic and/or disability support services, such as the Writing and Language Skills Centre, Maths Skills Centre and/or Disability Services. (The only exception to this will be for those students who have been recommended an exam support worker in a Student Support Plan. If this applies to you, you are advised to contact Disability Services as soon as possible to discuss the necessary arrangements.)
- seek advice or contribution from any third party, including proofreaders, online fora, friends, or family members.

We expect, and trust, that all our students will seek to maintain the integrity of the assessment, and of their award, through ensuring that these instructions are strictly followed. Failure to adhere to these requirements will be considered a breach of the Academic Misconduct regulations, where the offences of plagiarism, breach/cheating, collusion and commissioning are relevant: see AM1.2.1 *(Note this supercedes Section 7.3 of the Guide to Assessment)*.

1 (25 marks) Data Science Process

EnergiseMe is a Yorkshire company which, since its establishment in 1974 in York, UK, produces and trades natural energy drinks in more than 50 countries worldwide. Compared to the typical energy drinks on the market (e.g., Monster, Red Bull), the products of EnergiseMe offer equivalent amounts of energy (hence providing mental and physical stimulation), but without the heaps of sugar, artificial sweeteners and colourings. Some of its most well-known products are LukoMade, Tiger Matcha and GojiBeri.

EnergiseMe has produced a new natural energy drink named PowerHerb, whose main ingredients are Indian gooseberry, green coffee, and green tea. Initial tests carried out using a small number of tasters have shown promising results. Based on these results, EnergiseMe is considering proceding with prodution.

During a planning meeting that involved key stakeholders of the company, including the heads of the Marketing and Customer Service departments, the lead of Product Design team, the head of the Product Line team, and, you, as the head of the Data Science team, the following opinions were raised:

- "Since we are a Yorkshire-based company, the launch of our new product PowerHerb should happen here in York. Doing this will give a boost to York's economy".
- "As a company that produces natural energy drinks, we should produce a PowerHerb Zero Sugar and a PowerHerb with stevia (a natural sweetener). Each product's variant should come in two different flavours (e.g., berries and orange) to satisfy our customer base".
- "Since Indian gooseberry is one of the main ingredients of PowerHerb and this ingredient is particularly famous in Asia, we should consider producing and selling PowerHerb in Asia too".
- "Brexit has already taken place and additional taxes will be imposed on our products imported in certain European countries. However, we should not change our pricing policy and should keep the price of our new product at a maximum 10% higher than all our other products. Although we will have a smaller profit, PowerHerb would have more chances of penetrating the European market".

Interested in investigating these ideas further, the CEO of EnergiseMe has given you £20000 (as a budget) and one month to determine their validity by implementing a data science project.

(i) [20 marks] Using no more than **three** sentences for each data science lifecycle step, describe how you would implement this project to provide conclusive answers regarding the validity of these ideas. You can use examples, fictional data or plots to explain your arguments.

(ii) [5 marks] Provide details of **two** ethics-related concerns that should be considered when carrying out this data science project. Use no more than **two** sentences per ethical concern.

2 (25 marks) Data Analysis

The Yorkshire BeeKeepers Association collects data for the honey produced by its members each year and the average price per kilo (KG) for which the honey was sold to distribution companies. This data is available in the file named "honeyProduction.csv".

You should use the dataset named "honeyProduction.csv" and answer the following data questions using Python 3, Pandas, NumPy and Matplotlib.

(i) [5 marks] Compute measures of central tendency (mean, median) and dispersion (range, standard deviation.) Also, for each variable calculate its interquartile range. You **must** show your calculations. You **must not** use the describe() function of Pandas for the calculations but you **may** use it to check your results.

(ii) [2 marks] Calculate the Pearson's and Spearman's correlation coefficients between Production and Price.

(iii) [3 marks] Visualise the dataset as a scatter plot using the Matplotlib library in Python. The plot should be complemented with suitable axis legends and title.

(iv) [5 marks] Using Python and scikit-learn, generate the simple linear regression model for the dataset. Also, report the linear regression equation. Finally, plot the corresponding regression model on the scatter plot generated in (iii).

(v) [5 marks] Using Python and scikit-learn, generate a polynomial linear regression model of degree 5 for the dataset. Also, plot the corresponding regression model on the scatter plot generated in (iii). Finally, predict the average price per KG (£) for a production of 1010 KG.

(vi) [5 marks] Using Python and scikit-learn, calculate the Mean Absolute Error, Mean Square Error, and Root Mean Square Error both for the simple linear regression and polynomial regression models for the following list of production values - average price tuples:[(1010,0.45), (1011,0.36), (1012,0.39), (1013,0.53)]. In no more than two sentences, analyse the results and comment on the quality of the developed linear regression models.

3 (25 marks) Data Processing

The Association of Tennis Professionals (ATP) is the governing body organising and managing a wide range of professional tennis tournaments including the four Grand Slams (i.e., Australian Open, French Open, Wimbledon UK, and US Open) and ATP and Masters Cups.

In singles tennis matches, two players compete against each other over a number of sets. By winning their tennis matches, a player can proceed through the rounds and win the tournament. Each set comprises a sequence of games played with the service alternating between games, ending when the count of games won meets certain criteria. Typically, a player wins a set by winning at least six games and at least two games more than the opponent. Each set consists of a sequence of points played with the same player serving. A game is won by the first player to have won at least four points in total and at least two points more than the opponent. You can read more about the tennis rules at https://en.wikipedia.org/wiki/Tennis#Manner_of_play. (The HTML file is also available on VLE for your reference.)

DataVision has won a major contract with ATP. The main objective of this contract is to analyse the dataset for the tennis matches played in 2019 and 2020 and extract actionable knowledge the will help ATP to determine whether any changes are needed to the scheduling or structure of its various tournaments.

As the head of the Data Science team at DataVision, your task is to guide your team in undertaking this data analysis task.

You should use the dataset named "tennis20192020.csv" and answer the following data questions using Python 3, Pandas and Matplotlib.

(i) [1 mark] What are the dimensions of the dataset?

(ii) [2 marks] What does each column mean and what is its datatype?

(iii) [2 marks] Remove the variables 'WPts' and 'LPts' which are irrelevant for data analysis. Print the new shape of the data frame.

(iv) [5 marks] Is there any missing data? If yes, apply the most appropriate imputation strategy to handle the missing data and print the affected entries after imputation.

- For each match in which the number of sets won by each player is missing, assume that the minimum number of sets was played for that match.
- For each match in which the odds of match winner/loser is missing, use the average odds for that particular player across the entire dataset.

(v) [5 marks] When Novak Djokovic and Rafael Nadal participated in tournaments, how many matches has each won each against the top 10 best players of the tournament? How many of

these have been in the final?

(vi) [3 marks] Which tennis players have won at least 4 titles?

(vii) [4 marks] What percentage of ATP250 and ATP500 matches were won by an outsider player (i.e., a player with higher odds than his/her opponent's)?

(viii) [3 marks] For each series in the dataset, how many matches were played in each surface type available in indoor and outdoor courts? You can answer this question using either the 'groupby' or 'pivot' commands.

4 (25 marks) Hypothesis Testing

The service team at the Golden Lion restaurant in York collects the tips given by their customers and puts them in a common safe place. At the end of every week, the total amount is split equally between members of the service team. Recently, the team started recording several information about the tips received, including the total bill, the gender of the bill payer and the number of customers served per order.

The service team is interested in analysing this dataset and identifying whether there are factors that affect the amount of tips they receive. To this end, DataVision has been tasked to help and provide their insights.

You should use the dataset named "goldenLionTips.csv" and answer the following data questions using Python 3, Pandas, Matplotlib and Scipy.

(i) [2 marks] The service team believes that customers give a higher tip when they come for dinner rather than lunch. State the Null and Alternative hypotheses that DataVision would need to test to evaluate this situation.

(ii) [4 marks] Execute the appropriate test and report the result. You must show your calculations, i.e., calculate the result analytically, but you may use the Scipy library to check your results.

(iii) [2 marks] Draw appropriate conclusions making explicit reference to the critical value for a significance level $\alpha = 0.05$.

(iv) [5 marks] The team has observed that a different amount of tip is given depending on the number of customers served per order (group size). To analyse this hypothesis, DataVision partitions the orders into 'small' orders, i.e., involving up to 2 customers, 'medium' orders, i.e., with 3 or 4 customers, and 'large' orders, with 5 or more customers.

Perform the appropriate test to evaluate the hypotheses for significance level $\alpha = 0.05$. You must show your calculations, i.e., calculate the result analytically, but you may use the Scipy library to check your results.

(v) [4 marks] Plot the distributions for the 'small', 'medium' and 'large' orders.

(vi) [3 marks] Execute a post hoc analysis using the Bonferroni correction to identify the pairs which exhibit statistically significant difference.

(vii) [3 marks] Which entries may be reported as outliers using the IQR outlier detection method?

(viii) [2 marks] Produce a boxplot to confirm the findings from (vii).

End of examination paper