Est. 1841 | YORK ST JOHN UNIVERSITY

Assessment Brief

Module Code	COM7036M	**Module Lecturers**	Dr Gayathri Karthick (Lecturer)		
Module Title	Big Data				
Level	7	**Credit Value**	15		
Assessment Title	Portfolio – A data collection, structuring and processing project, with accompanying analytical report.				
Workload	N/A				
Assessment Number	1	**of**	1	**Weighting**	100%
Submission Type	Portfolio (Anonymous)				
Submission Method	Turnitin within Moodle				
Publication Date					
Due Date	26 May 2023, 12:00 Noon				
Expected Feedback Date	16 June 2023				
Format of Feedback	Written feedback within Turnitin/Moodle				

Programme Learning Outcomes (PLO)

This coursework is designed to achieve the following PLOs:
PLOs **7.1-7.7**

7.1Evaluate computer science concepts and principles and their application to the effective design, implementation, and usability of computer-based systems.
7.2Apply the findings of advanced scholarship and/or contemporary research and practice to the solution of computer science problems
7.3Critically evaluate computer science problems, including those at the forefront of the field.
7.4Demonstrate operation within applicable professional, legal, social, and ethical frameworks.
7.5Demonstrate originality and creativity in the solution of computer science problems.
7.6Recommend, with detailed justification, the appropriate computer science principles and practices to apply to the significant domain-specific activity.
7.7Apply standards, quality processes, and engineering principles to the solution of computer science problems.

Assignment Description

For Big Data assessment, you need to select **ONE** dataset from the dataset details available in this document, and perform the required task listed below:

1. **Acquisition and understanding the dataset:** Although we are providing the location of the dataset, you need to research and comment on how the dataset was collected. **(10 marks)**
2. **Storage:** You are probably storing the dataset in a local resource. In your report present a brief **What-if? Scenario** and what you will need to do to store your data in a cloud-based system or warehouse. **(5 marks)**
3. **Analysis:** This is the core of your assessment. You need to perform a full data

analyticsframework (based on the lectures and tutorials):

3.1. **Data Wrangling:** Applying your learning on missing value handling and outlier detection from the lectures and lab sessions to clean your data. Some cleaning could be to remove any feature/column with 60% missing values or holding NULL values, constants, NaN values, or to remove duplicate and highly correlated information. You can also perform outlier detection at this stage if this seems appropriate. You must save and submit a clean version of the dataset. **(20 marks)**

3.2. **Descriptive analytics:** You are expected to provide summary statistics and basic descriptive visualizations of the dataset you are describing in the report. Remember that you are answering what happened question. You must include, as a minimum, a histogram and a scatter plot or a Density-KDE plot. **(15 marks)**

3.3. **Diagnostic analytics:** Here you should try to answer the why question. As before, you must support your analysis with adequate plots such as wordclouds for text, correlation plots, and scatterplots with trendlines for numerical data as we have done in the labs. **(15 marks)**

3.4. **Predictive analytics:** Split the data into a training set and a test set once cleansing is done. Use suitable toolkit and libraries (Python, Orange, or whichever platform you are comfortable with) to train models (e.g., a Decision Tree or a Random Forest) from the training set to build a suitable classifier. You will need to test the performance of your model on your test set**. (20 marks)**

4. **Recommendation/Conclusion:** As part of your final report, you also need to provide a sensible recommendation based on your analysis and findings. **(10 marks)**

5. **Overall presentation and references. (5 marks)**

General considerations:
Please be aware that each plot should be fully described in your assessment. But the plot itself, must be clear enough to tell a story by itself.

For each dataset you must include a proper citation. This is usually provided at the source website. Remember that when dealing with text data, you must trim leading and trailing blanks before conducting any analysis.

In the following list you will find further details for each dataset: Name of file, source of file, suggested activities etc.

Data Sets

1. **Loan Default Analysis and Prediction**
   a. Files: Loan_status_2007-2020Q3.gzip(1.77 GB), LCDataDictionary.xlsx
   b. Source: https://www.kaggle.com/code/jkashish18/lending-club-python/data
   c. Task: Present an analysis of the features used in this dataset by looking into the source website. Do appropriate descriptive and diagnostic analysis on the data. If you are using this dataset, you also need to build a Loan Default Analysis and Prediction classifier, based on the 'loan status' (Current, Fully Paid, charged off, late etc. column provided with the dataset.
   d. Action: Create a feature importance plot and create some recommendations from the

top features to help the lending club predict applications that may default/be late in paying or can be identified as potential bad loans.

2. **Malware Memory Analysis for Intrusion Detection**
    a. Files: Obfuscated-MalMem2022.csv
    b. Source: http://insideairbnb.com/get-the-data.html
    c. Task: Present an analysis of the features used in this dataset by looking into the source website. Do appropriate descriptive and diagnostic analysis on the data. If you are using this dataset, you also need to Build a benign vs malware classifier, based on the 'class' column provided with the dataset.
    d. Action: Create a feature importance plot and discuss on the top five features why you think those would are effective for identifying the malware.
3. **Text Analytics: Airbnb reviews data for London**
    a. Files: reviews.csv.gz, locations.csv, positive.txt, negative.txt
    b. Source: http://insideairbnb.com/get-the-data.html
    c. Task: Find where in London hosts have the most positive and negative reviews. If you are using this dataset, you also need to build a sentiment classifier, creating the labels from the files positive.txt and negative.txt. Do appropriate descriptive and diagnostic analysis of the data. You can take as a starting point the labs covered in week 8 for doing the sentiment analysis.
    d. Action: What would you recommend for reducing the number of negative comments? For example, properties with a high number of negative comments are in a particular area, you can recommend renting somewhere else. You can also recommend the landlord to improve the description of the property or change the location of the property.

After you have completed the challenge, you need to submit a report no more than 15 pages. You also need to submit your findings and the implemented code in a Jupyter notebook file (ie ".ipynb" file)

Est. 1841 | YORK ST JOHN UNIVERSITY

Additional Information

The work you present should be your own work, and not just copied from others. You can quote from others, but you must say who the author is and use quotation marks or paraphrase. If you do not do so, we will investigate your work for academic misconduct. This is particularly likely if your Turnitin similarity score is above 25% and/or individual matches are above 6%.

If you require support with your study skills, please visit https://www.yorksj.ac.uk/students/study-skills/

Assessment Regulations

Please refer to the York St John University Code of Practice for Assessment and Academic Related Matters 2022-23.

We ask that you pay particular attention to the academic misconduct policy. Penalties will be applied where a student is found guilty of academic and/or ethical misconduct, including termination of programme (**Policy Link**).

You are required to keep to the word limit set for an assessment and to note that you may be subject to penalty if you exceed that limit. You are required to provide an accurate word count on the cover sheet for each piece of work you submit (**Policy Link**).

For late or non-submission of work by the published deadline or an approved extended deadline, a mark of 0NS will be recorded. Where a re-assessment opportunity exists, a student will normally be permitted only one attempt to be re-assessed for a capped mark (**Policy Link**).

An extension to the published deadline may be granted to an individual student if they meet the eligibility criteria of the (**Policy Link**).

Please see the assessment criteria below.

Est. 1841 | YORK ST JOHN UNIVERSITY

York St John University Level 7 Assessment Marking Descriptor- Big Data

Criteria	Deliverables	Marks
1. Subject knowledge & understanding	Learner should demonstrate good knowledge/understanding of Big Data and practice for this level through the identification and critical analysis of Big Data acquisition techniques. Learners should provide a rationale for the choices they have made.	10
2.Higher cognitive skills & originality	Learners should provide detailed criticality and evidence of storing the Big data set in a local infrastructure and critical analysis of storing Big Data on a cloud-based system or warehouse. These techniques should follow up with logical and sustained conclusions.	5
3. Advanced technical, professional and/or research expertise	Learners should be able to demonstrate technical and professional analysis of Data Wrangling techniques.	20
	The Learners are required to explain description of the data set using statistical methods and its visualisation using graphs. The evidence should include in the report.	15
	The Learners should explain the Diagnostics analysis. The explanation should contain a description of the methodology.	15
	Learner should demonstrate a suitable predictive method to analysis the data. Learners are required to evaluate the outcomes based on the preparation and processing of the Big Data analytic. The evaluation should show that the learner has made judgements taking into account different factors and using available knowledge, experience and evidence .	20

	Learners are required to make recommendations for how the analytics have been carried out with the selected Big Data sources. The recommendations should focus on how the data should be managed from selection through to processing. The evidence should include in the report.	10
4. References and Presentation	Citations are provided in Harvard's style and the report is well organized and presented.	5