**University of Westminster**

School of Computer Science & Engineering

|  |  |
| --- | --- |
| **7BUIS010W Data Warehousing and Business Intelligence – Coursework (2023/4)** | |
| Module leader | Dr Panagiotis Chountas |
| Unit | Coursework-CWK2 |
| Weighting: | 50% |
| Qualifying mark | 40% |
| Description | The in-module assessment will consist of a single coursework that will assess students’ ability to utilise conceptual modelling in Data Warehouses for the needs of subject oriented analysis; it will also assess students in depth and systematic understanding of key issues, advantages and problems related to data integration and warehousing. Finally, it will assess students’ ability to conceive and implement OLAP applications, to devise effective multi-dimensional databases and to use appropriate querying languages for effective decision making. |
| Learning Outcomes Covered in this Assignment: | This assignment contributes towards the following Learning Outcomes (LOs):  LO 4. define customer relationship management, change management problems, select and apply appropriate Business  Intelligence (BI) methodologies and evaluate BI solutions to these problems;  LO 5. demonstrate competence in using BI Technologies and Tools on business data for the purposes of CRM and CM;  LO 6. apply CRM knowledge and CM to support change and improve operational processes of service organizations. |
| Handed Out: | 28TH March 2024 |
| Due Date | 30TH April 2024, Submission by 13:00 |
| Expected deliverables | Submit on Blackboard a single file containing the required documentation (either in docx or pdf format). All implemented codes should be included in your documentation together with the results/analysis. |
| Method of Submission: | Electronic submission on BB via a provided link close to the submission time. |
| Type of Feedback and Due Date: | Feedback will be provided on BB, on 20th May 2024 (appx.15 working days) |
| BCS CRITERIA MEETING IN THIS ASSIGNMENT | * 7.1.1 Critical review of literature * 7.1.2 Development of the self-directed learner * 7.1.3 Respond to opportunities for innovation * 7.1.6 Use appropriate processes * 7.1.7 Investigate and define a problem * 7.1.8 Apply principles of supporting disciplines * 8.1.1 Systematic understanding of knowledge of the domain with depth in particular areas * 8.1.2 Comprehensive understanding of essential principles and practices * 8.2.1 Produce work informed by research at the forefront * 9.1.1 Systematic understanding of knowledge at the forefront in development and implementation * of systems * 9.1.2 Comprehensive understanding of the state of the art techniques * 10.2.1 Critical awareness of current research issues, problems and/or insights |

Assessment regulations

Refer to section 4 of the “How you study” guide for undergraduate students for a clarification of how you are assessed, penalties and late submissions, what constitutes plagiarism etc.

Penalty for Late Submission

If you submit your coursework late but within 24 hours or one working day of the specified deadline, 10 marks will be deducted from the final mark, as a penalty for late submission, except for work which obtains a mark in the range 50 – 59%, in which case the mark will be capped at the pass mark (50%). If you submit your coursework more than 24 hours or more than one working day after the specified deadline you will be given a mark of zero for the work in question unless a claim of Mitigating Circumstances has been submitted and accepted as valid.

It is recognised that on occasion, illness or a personal crisis can mean that you fail to submit a piece of work on time. In such cases you must inform the Campus Office in writing on a mitigating circumstances form, giving the reason for your late or non-submission. You must provide relevant documentary evidence with the form. This information will be reported to the relevant Assessment Board that will decide whether the mark of zero shall stand. For more detailed information regarding University Assessment Regulations, please refer to the following website:<http://www.westminster.ac.uk/study/current-students/resources/academic-regulations>

**Data Set Information:** The dataset has 38765 rows comprising purchase orders of customers from grocery stores. These orders can be analysed using Market Basket Analysis or clustering for performing RFM analysis.

Attribute Information:

Member Number: Nominal, a 4-digit integral number uniquely assigned to each member.

Item Description: Product (item) name. Nominal.  
Date: Numeric, the day and time when each transaction was generated.  
**The dataset is available** on BB

**Guidelines:**

You are required to deliver a report (max 15 pages including figures) describing the methods adopted and the discussion of achieved results with reference to the tasks listed below. Assume that the report is targeted to a marketing strategist, who is interested in learning the business insights inferred in your analysis and receiving suggestions on how to take appropriate actions.

**Tasks**

1. **Data Understanding**: useful as a preliminary step to capture basic data properties and schema. To understand the basic data properties, you need to perform distribution analysis, statistical exploration, suitable transformation of variables and elimination of missing values.

**Identify as part of a UML class diagram the main classes and associations by assigning the correct cardinality constraints**.

**[12 Marks]**

1. **Perform** **RFM Segmentation in SQL**: The first step is to build an RFM model to assign Recency, Frequency and Monetary values to each member via an SQL query.

**[11 Marks]**

1. **Customer segmentation with DBSCAN:** The second step is to divide the customer list into tiered groups using DBSCAN and discuss the profile of each found cluster (in terms of the properties that describe the customers of each cluster). The report should illustrate the adopted clustering methodology and the cluster interpretation. In particular, it is necessary to discuss the identification of the best value of **ε and minPts**.

**[10 Marks]**

1. **Review of Results:** Discuss briefly the business value for marketers of the specific clusters and segments of customers and their behaviour – in terms of increased customer loyalty and customer lifetime value.

**[8 Marks]**

1. **Data Mart Design:** Based on your findings (Tasks (2,3)) and conclusions (Task 4), suggest the main dimensions and metrics for designing a data mart for the analysis needs of the marketing department.

[9 Marks]

Total [50 Marks]

Marking Scheme

Due to the nature of the assessment candidates may come up with more than one equally, good solutions. Thus marks will be allocated as follows

**Tasks**

1. **Data Understanding**: useful as a preliminary step to capture basic data properties and schema. To understand the basic data properties, you need to perform distribution analysis, statistical exploration, suitable transformation of variables and elimination of missing values.

**Identify as part of a UML class diagram the main classes and associations by assigning the correct cardinality constraints**

* Distribution analysis **[1 Mark]**
* Statistical exploration **[1 Mark]**
* Suitable transformation of variables **[1.5 Marks]**
* Elimination of redundant variables **[1.5 Marks]**
* Data visualisation **[1.5 Marks]**
* Management of missing values **[1 Mark]**
* Class identification **[1.5 Marks**]
* Association identification **[1.5 Marks]**
* Cardinality constraints  **[1.5 Marks]**

**[12 Marks]**

1. **Perform** **RFM Segmentation in SQL**: The first step is to build an RFM model to assign Recency, Frequency and Monetary values to each member via an SQL query.
   * + - Definition of RFM metrics [6 Marks]
       - Implementation in SQL/Python of the RFM metrics **[5 Marks]**

**[11 Marks]**

1. **Customer segmentation with DBSCAN:** The second step is to divide the customer list into tiered groups using DBSCAN and discuss the profile of each found cluster (in terms of the properties that describe the customers of each cluster). The report should illustrate the adopted clustering methodology and the cluster interpretation. In particular, it is necessary to discuss the identification of the best value of ε and minPts.
   * + - Build DBSCAN Model in Python [3 Marks]
       - Correct Justification of ε and minPts [5 Marks]
       - Testing of **DBSCAN Model**  in Python **[2 Marks]**

**[10 Marks]**

1. **Review of Results:** Discuss briefly the business value for marketers of the specific clusters and segments of customers and their behaviour – in terms of increased customer loyalty and customer lifetime value.
   * + - Identification of business value customer segments [4 Marks]
       - Correct Justification of their business value [4 Marks]

**[8 Marks]**

1. **Data Mart Design:** Based on your findings (Tasks (2,3)) and conclusions (Task 4), suggest the main dimensions and metrics for designing a data mart for the analysis needs of the marketing department.
   * + - Identification of Dimensions [2 Marks]
       - Justification of Selected Dimensions [2 Marks]
       - Identification of Measures [2.5 Marks]
       - Justification of Selected Dimensions [2.5 Marks]

[9 Marks]

Total [50 Marks]