**BUS5DWR – Data Wrangling**

**Student Name: Phu Qui Dang**

**Student ID: 22095834**

**Assignment 01: Database design and SQL**

**Marks: 60 marks (30% of the final grade)**

**Assignment Type: Individual**

|  |  |
| --- | --- |
| **SILOS** | **SILO 1:** Design, formulate and compose database schema and query structures for analytics initiatives |
|  | **SILO 3:** Construct SQL and computer programs to wrangle data. |

# Overview

Over the past few weeks, you have gained a general understanding of designing databases and how to use SQL to retrieve information from business data. This assignment allows you to demonstrate what you have learned.

Academic Integrity

**Plagiarism** occurs when you use words, ideas, or work products attributable to another identifiable person or source:

* without attributing the work to the source from which it was obtained
* in a situation in which there is a legitimate expectation of original authorship
* in order to obtain some benefit, credit, or gain which need not be monetary

**Collusion** is a form of cheating which occurs when people work together in a deceitful way to develop a submission for an assessment which has been restricted to individual effort.

**By submitting\* this piece of work and signing this document, I declare that:**

* 1. The work is my own individual work.
  2. I have not previously submitted all or part of this work for assessment in any subject, unless the subject coordinator for the current subject (or my research supervisor, if applicable) has given me written permission to reuse specific material and I have correctly referenced the material taken from my own earlier work.
  3. I have read and agree to be bound by the Statutes, Regulations and Policies of the University relating to Academic Integrity available at [http://www.latrobe.edu.au/students/academic-integrity;](http://www.latrobe.edu.au/students/academic-integrity) and
  4. I may be subject to student discipline processes in the event of an act of academic misconduct by me including an act of plagiarism or cheating.

I further grant to the University or any third party authorised by the University (www.latrobe.edu.au/text-match) the right to reproduce and/or communicate (make available online or electronically transmit) the work I have submitted for the purpose of detecting plagiarism.

# Assignment Description

**Part 1: Database design**

**Question 1 [30 marks]**

**AUT Motors’ Database System**

AUT Motors is a large automobile dealership that sells a wide range of local and international vehicles from the world's leading manufacturers at competitive prices. Rather than being situated in the heart of a large metropolis, AUT Motors' branches are located in beautiful regional areas.

The current record-keeping system at the dealership uses on manual processes, involving hardcopies for each transaction and separate files for every car sale. This outdated method has several drawbacks including being time consuming, insecure, slow in retrieving information, and prone to human errors, all of which reflect poorly on the company's reputation. By incorporating digital transformation approaches into the company processes, AUT Motors decided to address these challenges by implementing a software system with a database to store this information. This initiative aims to improve data management efficiency and alleviate the significant costs associated with manual transactions. The database will encompass customer details, branch information, sales agent records, and sales transactions for comprehensive tracking and streamlined operations.

The current process involves a customer filling out a form that contains their personal information and the features that they are seeking in a vehicle (Appendix). It is the policy of AUT Motors to list cars based on the interest of the customer. However, the features section of the form may be left blank by some customers. These customers speak directly to the sales agent to determine which car they are interested in. Besides car feature, AUT Motors records car model, type (e.g., 4WD, sedan), manufacturer (name, country, etc), year of manufacturer. A manufacturer may have production line in different countries.

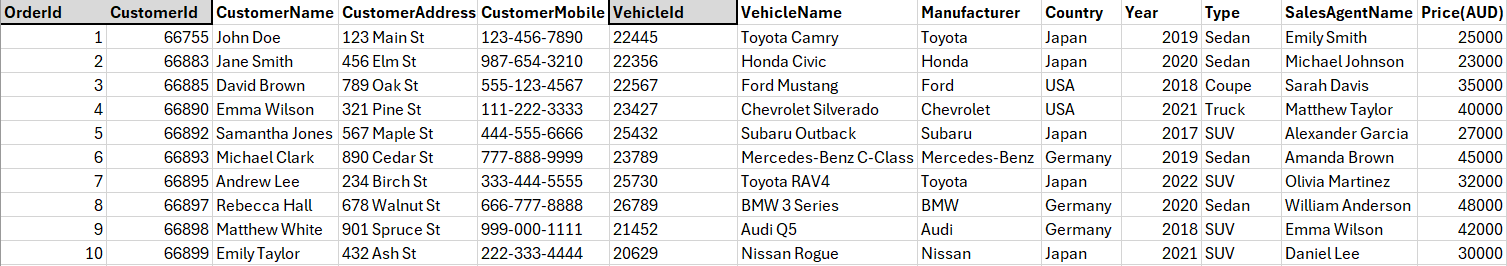
Once a customer purchases a car, AUT Motors records the details of customers, details of car, details of sales agent, sales date, sales price, etc. In addition, the dealership has records of the car, including the price paid by the dealership and the date the car entered the dealership.

**Appendix**

|  |  |
| --- | --- |
| AUT Motors  Car with solid fill | |
| **Customer:**  First Name --------  Last Name ---------  Address: Street …………. Suburb ……… State……. Postal Code ………… | Gender --------  Age ---------  Phone Number ----------  E-mail Address ---------- |
| **Interested Features**  Electric  Automatic Transition  Keyless  Airbag  ABS  Cruise control  MP4 Player | Leader Seats  Remote Start  Bluetooth System  Auto Drive  Parking Sensors  Rear Camera  Lane Keeping Assist |
| **Description** (fill in by staff) | |

Prior to establishing a software development division with database experts, existing members of AUT Motor’s did a quick proof of concept (PoC) design for the database. As they are not experts in database design, they have come up with below design which only captures the most important details during a vehicle purchase.

**Question 1.1 (10 marks)**



1. Will this table satisfy 1NF, 2NF, and 3NF? Explain why.

**Answer:**

|  |  |
| --- | --- |
| **Normal Form** | **Explanation** |
| **1 NF** | 1NF is met because each column seems to contain only atomic values, and there are no repeating groups or arrays within the rows. |
| **2 NF** | The table does not satisfy the 2NF because there are partial dependencies of non-key attributes on the primary key. For example, CustomerName, CustomerAddress, and CustomerMobile depend on CustomerId, not on the entire primary key (OrderId). |
| **3 NF** | Since the table doesn’t satisfy 2NF, it can’t satisfy 3NF. There are also transitive dependencies present; for instance, Manufacturer and Country depend on VehicleName. |

1. Give examples for insertion anomaly, deletion anomaly and update anomaly that may happen when using this table to record data.

**Answer:**

* Insertion Anomaly: Suppose a new customer wants to register in the system but they haven’t placed an order yet. In the current table structure, you can’t insert a new customer without having an OrderId, VehicleId, and other order-related details.
* Deletion Anomaly: If an order is canceled or deleted from the table, the information about the customer who placed the order and the vehicle ordered will also be lost if that was their only order.
* Update Anomaly: If a customer changes their address or mobile number, you must update every row in the table where that customer’s details appear. If you miss any row, the database will have inconsistent data for that customer.

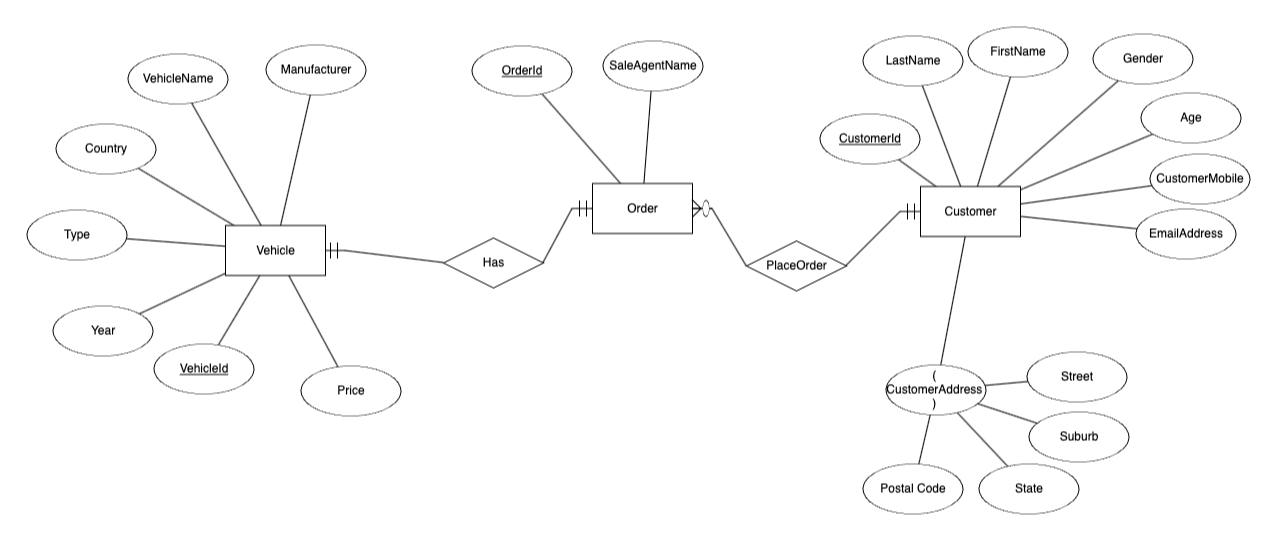
**Question 1.2 (20 marks)**

Using [https://erdplus.com](https://erdplus.comL), design a database in the normal form. You are asked to:

1. Present your entity-relationship diagram (ERD). Any assumption should be stated clearly.
2. Present the generated relational schema. The primary key and possible foreign keys of each table should be shown.
3. Present the generated SQL statements to compose the tables with the appropriate data type set for each attribute.

**NOTE**: The initial table design (Question 1.1) doesn’t capture the details of the order request from a customer which shown in the Appendix. Hence, when you design the ER diagram, make sure to capture the details of initial vehicle order request as well.

**Answer:**

1. 

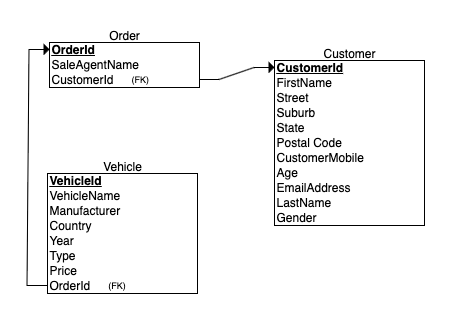
A customer can place no or many orders.

An order needs to belong to one and only one customer.

An order needs to belong to one and only one vehicle.

A vehicle may be ordered by one or only one order.

2)



3)

CREATE TABLE Customer

(

FirstName INT NOT NULL,

Street INT NOT NULL,

Suburb INT NOT NULL,

State INT NOT NULL,

Postal\_Code INT NOT NULL,

CustomerMobile INT NOT NULL,

CustomerId INT NOT NULL,

Age INT NOT NULL,

EmailAddress INT NOT NULL,

LastName INT NOT NULL,

Gender INT NOT NULL,

PRIMARY KEY (CustomerId)

);

CREATE TABLE Order

(

SaleAgentName INT NOT NULL,

OrderId INT NOT NULL,

CustomerId INT NOT NULL,

PRIMARY KEY (OrderId),

FOREIGN KEY (CustomerId) REFERENCES Customer(CustomerId)

);

CREATE TABLE Vehicle

(

VehicleName INT NOT NULL,

Manufacturer INT NOT NULL,

Country INT NOT NULL,

Year INT NOT NULL,

Type INT NOT NULL,

VehicleId INT NOT NULL,

Price INT NOT NULL,

OrderId INT NOT NULL,

PRIMARY KEY (VehicleId),

FOREIGN KEY (OrderId) REFERENCES Order(OrderId)

);

**Part 2: Database retrieval using SQL**

We will make use of the data contained in bus5dwr.Purchasing schema within the lmban-teaching.database.windows.net. Please refer to Workshop 1 for instructions on how to access the server via DBeaver if you have not accessed it already.

A screenshot of a computer

Description automatically generated

Start by exploring the contents of the tables to understand the meaning of the attributes of each table.

**Question 2.1 [30 marks]:** The manager wants a better insight into the company business to support his future decisions. He requests the IT department to provide answers to the following questions. You are asked to construct SQL code to find the answer for each question.

Fill in below the code ***in text*** and screenshot the result table highlighting the columns (limit to the first 8 rows and must show the number of fetched rows and time). Result of one question must be shown in only one return table.

***Example (No marks)***

|  |  |
| --- | --- |
| ***Question*** | Display the orders that have an order quantity greater than 3. |
| ***Code*** | **SELECT** pod.PurchaseOrderID, pod.ProductID , pod.OrderQty  **FROM** Purchasing.PurchaseOrderDetail pod  **WHERE** pod.OrderQty > 3 |
| ***Returned table*** | (*Adjust your screen to include the yellow line in the screenshot.*) |

**2.1  *(2 Marks)***

|  |  |
| --- | --- |
| ***Question*** | Display product standard price and vendor name of products with a standard price higher than the average standard price. |
| ***Code*** | **SELECT** *pv*.StandardPrice **AS** *ProductStandardPrice* , *v*.Name **AS** *VendorName*  **FROM** Purchasing.ProductVendor *pv*  **JOIN** Purchasing.Vendor *v*  **ON** *pv*.BusinessEntityID = *v*.BusinessEntityID  **WHERE** *pv*.StandardPrice >  (**SELECT** **AVG**(StandardPrice)  **FROM** Purchasing.ProductVendor *pv*) |
| ***Returned table*** |  |

**2.2 *(2 Marks)***

|  |  |
| --- | --- |
| ***Question*** | How many distinct orders made in February 2012 that have been either shipped by “TRUCK” or “CARGO”? Display also the order quantity. (Note: TRUCK or CARGO should be in the name of shipment method) |
| ***Code*** | **SELECT** **COUNT**(**DISTINCT** *poh*.PurchaseOrderID) **AS** *NumberOfOrders*, *pod*.OrderQty **AS** *OrderQuantity*  **FROM** Purchasing.PurchaseOrderHeader *poh*  **JOIN** Purchasing.ShipMethod *sm*  **ON** *poh*.ShipMethodID = *sm*.ShipMethodID  **JOIN** Purchasing.PurchaseOrderDetail *pod*  **ON** *pod*.PurchaseOrderID = *poh*.PurchaseOrderID  **WHERE** (*sm*.Name **LIKE** **'%TRUCK%'** **OR** *sm*.Name **LIKE** **'%CARGO%'**)  **AND** **MONTH**(*poh*.OrderDate) = 2  **AND** **YEAR**(*poh*.OrderDate) = 2012  **GROUP** **BY** *pod*.OrderQty |
| ***Returned table*** |  |

**2.3 *(2 Marks)***

|  |  |
| --- | --- |
| ***Question*** | Display vendors name, state and its latest receipt cost in descending order of both standard price and average lead times. |
| ***Code*** | **SELECT** *v*.Name **AS** *VendorName*, *vvwa*.StateProvinceName **AS** *State*, *pv*.LastReceiptCost **as** *LatestReceiptCost*  **FROM** Purchasing.Vendor *v*  **JOIN** Purchasing.ProductVendor *pv*  **ON** *v*.BusinessEntityID = *pv*.BusinessEntityID  **JOIN** Purchasing.vVendorWithAddresses *vvwa*  **ON** *vvwa*.BusinessEntityID = *v*.BusinessEntityID  **ORDER** **BY** *pv*.StandardPrice **DESC**, *pv*.AverageLeadTime **DESC** |
| ***Returned table*** |  |

**2.4 *(2 Marks)***

|  |  |
| --- | --- |
| ***Question*** | Select Top 4 vendors based on total sales (calculated from order quantity and unit price). |
| ***Code*** | **SELECT** **TOP** 4 v.Name **AS** VendorName, **SUM**(*pod*.OrderQty \* *pod*.UnitPrice) **AS** *TotalSales*  **FROM** Purchasing.Vendor *v*  **JOIN** Purchasing.ProductVendor *pv*  **ON** *v*.BusinessEntityID = *pv*.BusinessEntityID  **JOIN** Purchasing.PurchaseOrderDetail *pod*  **ON** *pod*.ProductID = *pv*.ProductID  **JOIN** Purchasing.PurchaseOrderHeader *poh*  **ON** *poh*.PurchaseOrderID = *pod*.PurchaseOrderID  **GROUP** **BY** *v*.Name  **ORDER** **BY** **SUM**(*pod*.OrderQty \* *pod*.UnitPrice) **DESC** |
| ***Returned table*** |  |

**2.5 *(3 Marks)***

|  |  |
| --- | --- |
| ***Question*** | Calculate the sales of vendors in Washington province. Results must have columns that named as vendor name, year, total quantity, and total sales (calculated from order quantity and unit price) and displayed in time order. |
| ***Code*** | **SELECT** *v*.Name **AS** *VendorName*, **YEAR**(*poh*.OrderDate) **AS** **Year**, **SUM**(*pod*.OrderQty) **AS** *TotalQuantity*, **SUM**(*pod*.OrderQty \* *pod*.UnitPrice) **AS** *TotalSales*  **FROM** Purchasing.Vendor *v*  **JOIN** Purchasing.vVendorWithAddresses *vvwa*  **ON** *vvwa*.BusinessEntityID = *v*.BusinessEntityID  **JOIN** Purchasing.ProductVendor *pv*  **ON** *v*.BusinessEntityID = *pv*.BusinessEntityID  **JOIN** Purchasing.PurchaseOrderDetail *pod*  **ON** *pod*.ProductID = *pv*.ProductID  **JOIN** Purchasing.PurchaseOrderHeader *poh*  **ON** *poh*.PurchaseOrderID = *pod*.PurchaseOrderID  **WHERE** *vvwa*.StateProvinceName = **'Washington'**  **GROUP** **BY** *v*.Name, **YEAR**(*poh*.OrderDate)  **ORDER** **BY** **YEAR**(*poh*.OrderDate) |
| ***Returned table*** |  |

**2.6 *(3 Marks)***

|  |  |
| --- | --- |
| ***Question*** | Calculate the sum of freight charges for each shipping method used for the orders placed in the year 2011. The result should include the shipping method and the total freight charges for that method. |
| ***Code*** | **SELECT** *sm*.Name **AS** *ShippingMethod*, **SUM**(*poh*.Freight) **AS** *TotalFreightCharge*  **FROM** Purchasing.PurchaseOrderHeader *poh*  **JOIN** Purchasing.ShipMethod *sm*  **ON** *sm*.ShipMethodID = *poh*.ShipMethodID  **WHERE** **YEAR**(*poh*.OrderDate) = 2011  **GROUP** **BY** *sm*.Name |
| ***Returned table*** |  |

**2.7 *(3 Marks)***

|  |  |
| --- | --- |
| ***Question*** | Retrieve the details of vendors from whom the company has not purchased any products. The details should include the vendor's full name with title, phone number, and email address. |
| ***Code*** | **SELECT** *vvwc*.Title **AS** *Title* ,  *vvwc*.FirstName + **' '** +*vvwc*.LastName **AS** *FullName*,  *vvwc*.PhoneNumber **AS** *PhoneNumber*,  *vvwc*.EmailAddress **AS** *EmailAddress*  **FROM** Purchasing.vVendorWithContacts *vvwc*  **LEFT** **JOIN** Purchasing.Vendor *v*  **ON** *v*.BusinessEntityID = *vvwc*.BusinessEntityID  **WHERE** *v*.BusinessEntityID **NOT** **IN**  (**SELECT** *poh*.VendorID  **FROM** Purchasing.PurchaseOrderHeader *poh*) |
| ***Returned table*** |  |

**2.8 *(3 Marks)***

|  |  |
| --- | --- |
| ***Question*** | The company has decided to cease purchasing items from vendors who are from Arizona and Washington. You need to write a query to identify products that will have no available vendors to supply them after ceasing purchases from vendors located in Arizona and Washington. The query should return product IDs if any exist. |
| ***Code*** | **SELECT** *pv*.ProductID  **FROM** Purchasing.vVendorWithAddresses *vvwa*  **JOIN** Purchasing.ProductVendor *pv*  **ON** *vvwa*.BusinessEntityID = *pv*.BusinessEntityID  **WHERE** *vvwa*.StateProvinceName = **'Arizona'OR** *vvwa*.StateProvinceName = **'Washington'**  **GROUP** **BY** *pv*.ProductID |
| ***Returned table*** |  |

**2.9 *(5 Marks)***

|  |  |
| --- | --- |
| ***Question*** | The company wants to choose the best vendor for each shipping type to process orders faster that can improve the company sales and reputation. To help the manager suggest these vendors, you are required to propose a criterion/criteria with justification, and write code to find the answer. |
| ***Criteria*** | Shortest Average Lead Time: The lead time is the period between initiating and completing a process. In this case, it refers to the time it takes for a vendor to fulfill an order. Vendors with shorter average lead times can process orders faster, leading to quicker deliveries.  Cheapest Freight Cost: The freight cost is the price at which a specific cargo is delivered from one point to another. The company can save on shipping expenses by choosing vendors with the cheapest freight costs. These savings could be passed on to customers through lower prices, making the company’s products more attractive. |
| ***Justification*** | Average lead time directly impacts the speed of order processing and delivery.  Choosing vendors with lower lead times ensures faster order fulfillment, which is crucial for customer satisfaction and retention.  Faster delivery times can lead to higher customer satisfaction, repeat purchases, and positive word-of-mouth, ultimately enhancing the company's reputation. |
| ***Code*** | **SELECT** *TT2*.ShipMethodID, **MIN**(*TT2*.Name) **AS** *Name*, **MIN**(*TT2*.*AverageLeadTime*) **AS** *ShortestDeliveryTime*, **MIN**(*TT2*.*CheapestFreight*) **AS** *CheapestShippingCharge*  **FROM** (  **SELECT** *v*.Name, *sm*.ShipMethodID, **AVG**(*pv*.AverageLeadTime) **AS** *AverageLeadTime*, **MIN**(*poh*.Freight) **AS** *CheapestFreight*  **FROM** Purchasing.Vendor *v*  **JOIN** Purchasing.ProductVendor *pv*  **ON** *v*.BusinessEntityID = *pv*.BusinessEntityID  **JOIN** Purchasing.PurchaseOrderHeader *poh*  **ON** *v*.BusinessEntityID = *poh*.VendorID  **JOIN** Purchasing.ShipMethod *sm*  **ON** *sm*.ShipMethodID = *poh*.ShipMethodID  **GROUP** **BY** *v*.Name, *sm*.ShipMethodID ) **AS** *TT2*  **GROUP** **BY** *TT2*.ShipMethodID |
| ***Returned table*** |  |

**2.10 *(5 Marks)***

|  |  |
| --- | --- |
| ***Question*** | The company wants to choose vendors based on their reputation and speed of processing orders. You are required to propose a criterion/criteria with justification, and write code to find the answer. Based on the results suggest two vendors and justify. |
| ***Criteria*** | Speed of Processing Orders: The speed at which a vendor can process orders is crucial for maintaining a fast and efficient supply chain. Faster order processing can lead to quicker deliveries, improving customer satisfaction and leading to repeat business. This can measure the average lead time from order placement to delivery.  Lowest Shipping Cost: The cost of shipping can significantly impact the overall cost of products. The company can save on shipping expenses by choosing vendors with the cheapest shipping. These savings could be passed on to customers through lower prices, making the company’s products more attractive. |
| ***Justification*** | These criteria aim to enhance the company’s operational efficiency and reduce costs. The company can ensure timely deliveries and minimize shipping expenses by choosing vendors with quick order processing times and the lowest shipping cost. This can lead to increased customer satisfaction, repeat business, and a positive reputation for the company while also improving the company’s bottom line. |
| ***Code*** | **SELECT** v.Name, **AVG**(pv.AverageLeadTime), **MIN**(poh.Freight)  **INTO** #TempTable5  **FROM** Purchasing.Vendor v  **JOIN** Purchasing.ProductVendor pv  **ON** v.BusinessEntityID = pv.BusinessEntityID  **JOIN** Purchasing.PurchaseOrderDetail pod  **ON** pod.ProductID = pv.ProductID  **JOIN** Purchasing.PurchaseOrderHeader poh  **ON** poh.PurchaseOrderID = pod.PurchaseOrderID  **JOIN** Purchasing.ShipMethod sm  **ON** sm.ShipMethodID = poh.ShipMethodID  **WHERE** v.PreferredVendorStatus = 1  **GROUP** **BY** v.Name  **SELECT** Name, **MIN**(**AVG**(pv.AverageLeadTime)) **AS** FastestLeadTime, **MIN**(**MIN**(poh.Freight)) **AS** CheapestShipping  **INTO** #TempTable6  **FROM** #TempTable5  **GROUP** **BY** Name  **SELECT** **TOP** 2 Name, FastestLeadTime, CheapestShipping  **FROM** #TempTable6  **ORDER** **BY** FastestLeadTime, CheapestShipping **ASC** |
| ***Returned table*** |  |

# Submission Guide

Your submission is this WORD file with all the questions answered. The SQL code has to be copied as text (not an image or screenshot) so that the markers can immediately re-run and check your answers. Marks will not be given if the code is not provided in plain text.

Submissions with high similarity in the SQL code with another submission, especially having the same wrong code, will be considered plagiarism/collusion and will be reported to the Academic Integrity Advisors (AIA) for further investigation.

# Marking rubric

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Criterion | High Distinction  (> 80%) | Distinction  (70-79%) | Credit  (60-69%) | Pass  (50-59%) |
| ERD, Relational Schema, and SQL Statements (20) | Exceptional effort to address questions and present information and insights.  In depth  knowledge of ERD and database schema.  All entities and attributes are identified. | Excellent effort to address questions and present information and insights.  Comprehensive  knowledge and understanding of ERD and database schema.  Majority of entities and attributes are identified. | Fair effort to address questions and present information and insights.  Developed knowledge and understanding of ERD and database schema.  Almost majority of entities and attributes are identified. | Limited effort to address questions.  Acceptable knowledge and understanding of ERD and database schema.  Acceptable number of entities and attributes are identified. |
| Anomaly (10) | Demonstrates in depth  knowledge of anomalies. | Demonstrates comprehensive  knowledge and understanding  of anomalies. | Demonstrates a  developed knowledge  and understanding of anomalies. | Demonstrates  acceptable knowledge  and understanding of anomalies. |
| SQL code (30) | Demonstrates in depth  knowledge of constructing SQL. | Demonstrates comprehensive  knowledge and understanding  of constructing SQL. | Demonstrates a  developed knowledge  and understanding of constructing SQL. | Demonstrates  acceptable knowledge  and understanding of constructing SQL. |