UNIVERSITY OF MALAYA

Midterm Test FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2022/2023 : SEMESTER 1

WQD7005 : Data Mining

Name: Sadman Chowdhury ID: S2199546

Duration: From 3/12/2022, 8.00 PM to 3/12/2022, 11.00 PM

INSTRUCTIONS TO CANDIDATES :

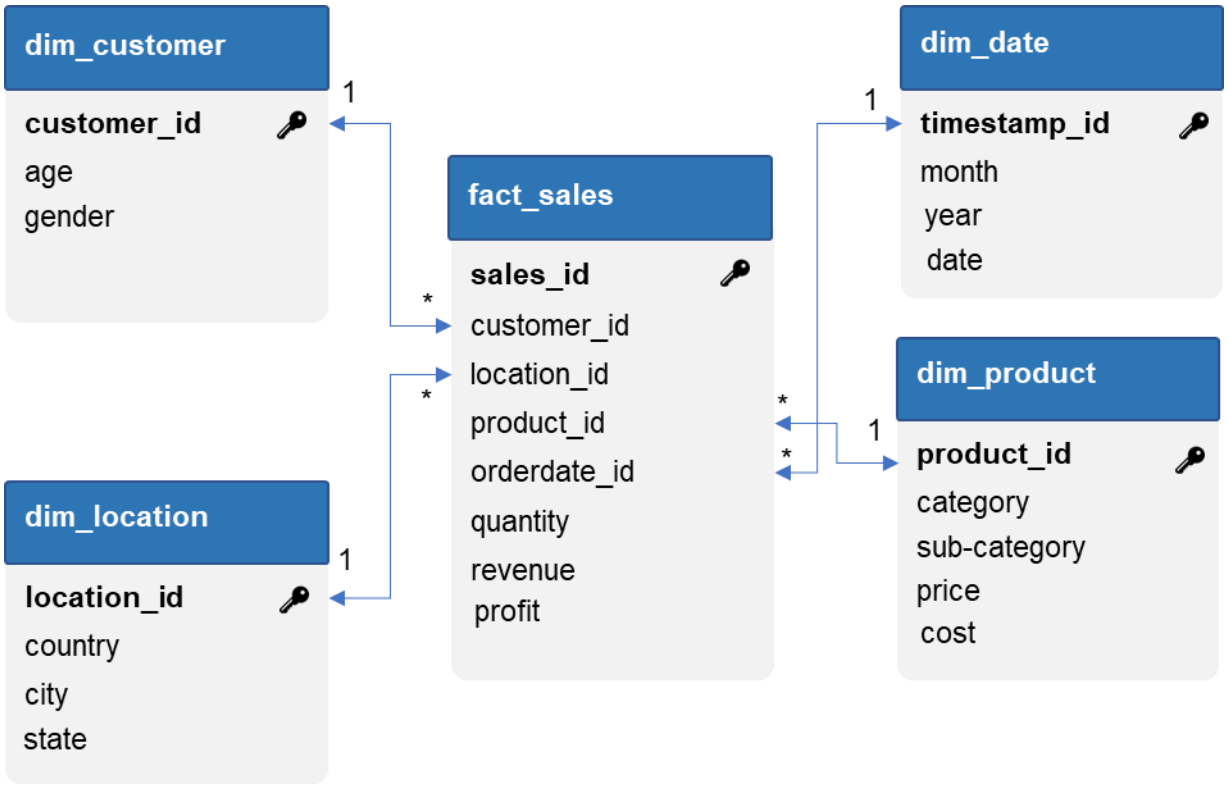
Answer **ALL** questions (10 marks).

(This question paper consists of 4 questions on 2 printed pages)

1. Use your group assignment dataset and draw a schema diagram for the data warehouse using Star schema.

Ans:

Dataset Link: https://github.com/KAFSALAH/WQD7005-DATA-MINING-PROJECT



(4 marks)

2. Define your words of Data Mining.

Ans: Data mining is a process used to extract knowledge from the raw collection of massive datasets by using pattern recognition techniques, clustering, and mathematical and statistical analysis to identify different variations, anomalies, correlations and hidden patterns and trends in the data to improve decisions and make a prediction.

(2 marks)

3. Discuss the differences between database and data warehouse with own three examples which are based on group assignment dataset.

(2 marks)

Ans: Database: Databases are special repositories used to store data that are generally recorded for operational purposes (OLTP). In our assignment, we used a data set of Bike Sales information. Typically, in the database, we store customer information, product information and store purchase data. We also store transactions related to all the purchases in here. These pieces of information are stored in a normal form to minimize data redundancy following the ACID principle in the database. Generally, a database table looks long and narrow, and the size varies from 1GB-1TB.

Datawarehouse: These are data repositories that generally take some incoming data and store them comprehensively, allowing further analysis and providing insights from those data (OLAP). A data warehouse is the knowledge base that helps a company's top managers make decisions and generate reports. For example, we can generate helpful insight into our bike sales dataset by including valuable columns in the fact table like order date, revenue, profit and other critical information. We can answer questions like in which month we have the most sales, filter sales by location, gender, and product type and easily visualize profit and loss. Such data repositories are generally wide and tall as lots of data are aggregated together, which results in a size greater than 1TB.

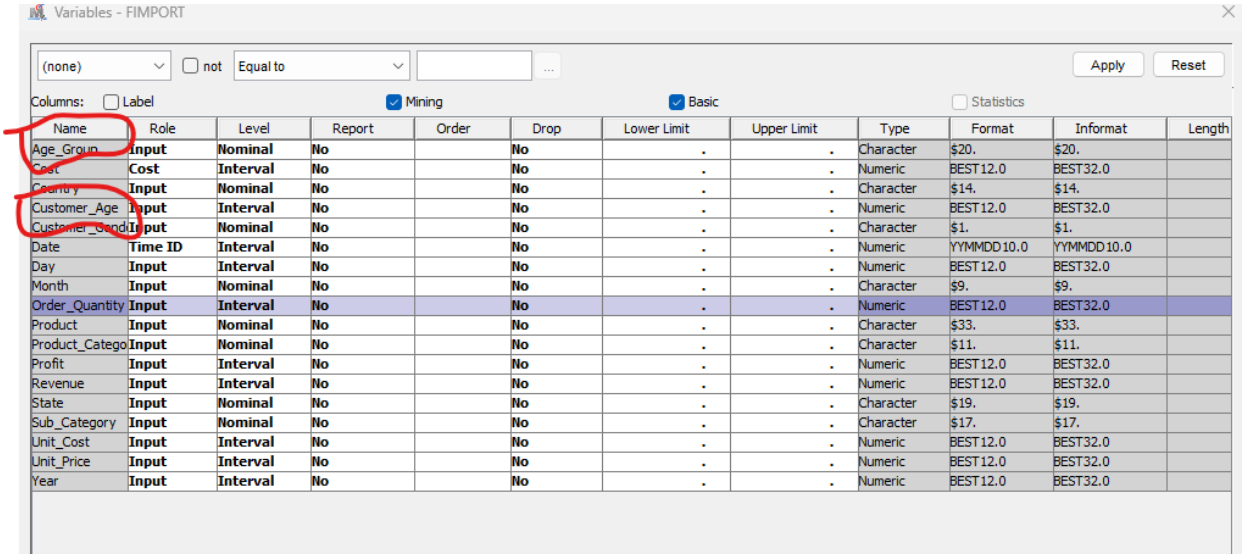
4. Data quality can be assessed in terms of several issues, including incomplete, noisy, inconsistent and intentional. Show any two issues in your group assignment dataset.

(2 marks)

Ans: In our dataset, we have divided the data in 5 samples among our groupmates. Here is the data quality issues that I found in my dataset:

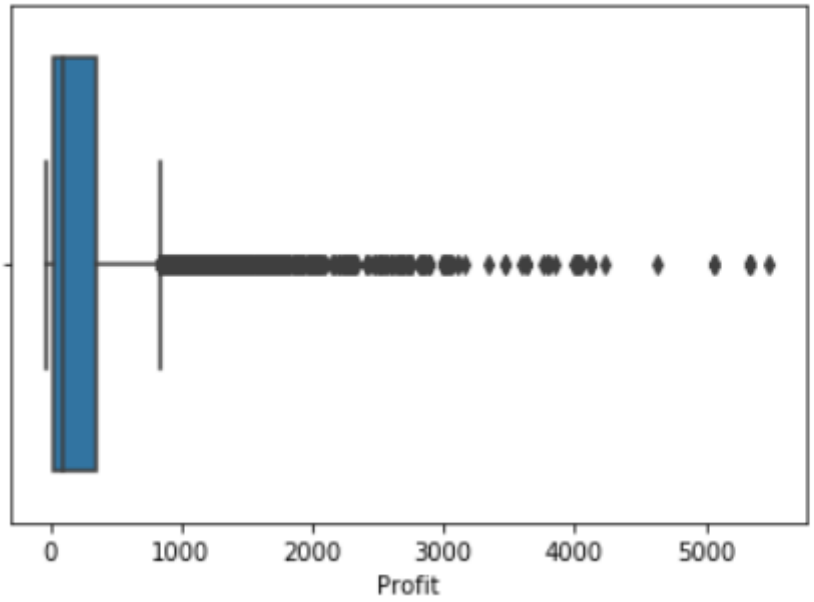
Dataset Description (of Sample):

Total number of records: 22607
No of columns: 18

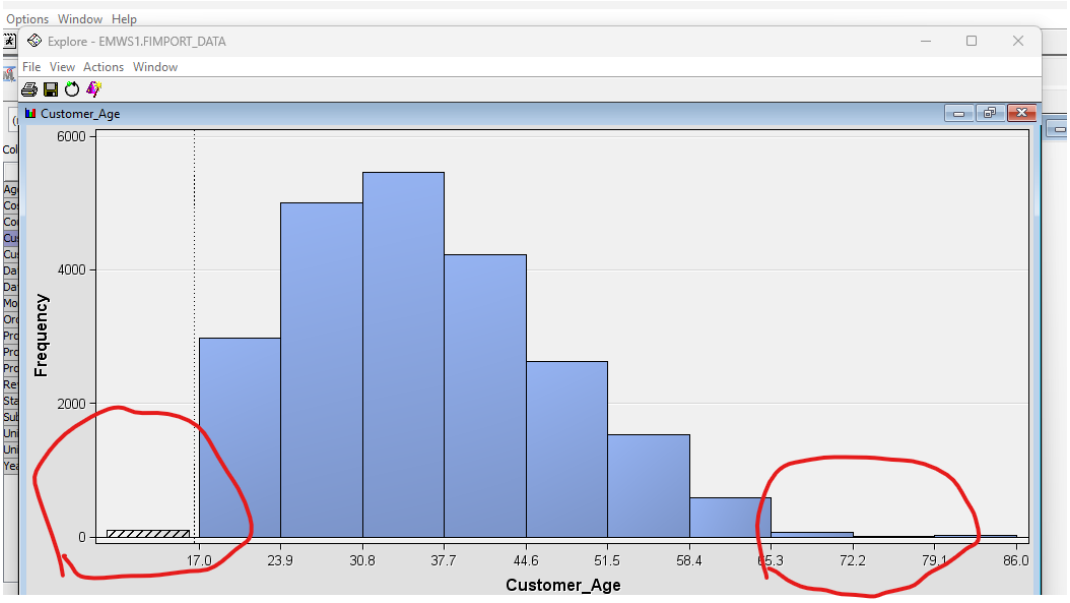


Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Type	Format	Informat	Length
Age_Group	Input	Nominal	No		No	.	.	Character	$20.	$20.	
Cost	Cost	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	
Country	Input	Nominal	No		No	.	.	Character	$14.	$14.	
Customer_Age	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	
Customer_Gend	Input	Nominal	No		No	.	.	Character	$1.	$1.	
Date	Time ID	Interval	No		No	.	.	Numeric	YYMMDD10.0	YYMMDD10.0	
Day	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	
Month	Input	Nominal	No		No	.	.	Character	$9.	$9.	
Order_Quantity	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	
Product	Input	Nominal	No		No	.	.	Character	$33.	$33.	
Product_Catego	Input	Nominal	No		No	.	.	Character	$11.	$11.	
Profit	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	
Revenue	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	
State	Input	Nominal	No		No	.	.	Character	$19.	$19.	
Sub_Category	Input	Nominal	No		No	.	.	Character	$17.	$17.	
Unit_Cost	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	
Unit_Price	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	
Year	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	

Here we can see there are multiple columns that represents the same value.



i) In the profit columns we can see the inter quartile range is from 0$ to 1000$ but there are lots of outliers in the example which indicates some extreme cases of very high profit.



ii) In the figure we can see there is 101 records with missing values in the customer age column.

END