MONASH University

Office Use Only

Semester Two 2022
Examination Period

Faculty of Information Technology

Sample Exam

EXAM CODES: **FIT3003**

TITLE OF PAPER: Business Intelligence and Data Warehousing - SAMPLE 3

EXAM DURATION: 2 hours 10 minutes or 130 minutes

THIS PAPER IS FOR STUDENTS STUDYING AT: (tick where applicable)

□ Caulfield ☒ Clayton □ Parkville □ Peninsula
□ Monash Extension □ Off Campus Learning ☒ Malaysia □ Sth Africa
□ Other (specify)

During an exam, you must not have in your possession any item/material that has not been authorised for your exam. This includes books, notes, paper, electronic device/s, mobile phone, smart watch/device, calculator, pencil case, or writing on any part of your body. Any authorised items are listed below. Items/materials on your desk, chair, in your clothing or otherwise on your person will be deemed to be in your possession.

No examination materials are to be removed from the room. This includes retaining, copying, memorising or noting down content of exam material for personal use or to share with any other person by any means following your exam.
Failure to comply with the above instructions, or attempting to cheat or cheating in an exam is a discipline offence under Part 7 of the Monash University (Council) Regulations, or a breach of instructions under Part 3 of the Monash University (Academic Board) Regulations.

AUTHORISED MATERIALS

OPEN BOOK	□ **YES**	☒ **NO**
CALCULATORS	□ **YES**	☒ **NO**
SPECIFICALLY PERMITTED ITEMS **if yes, items permitted are:**	□ **YES**	☒ **NO**

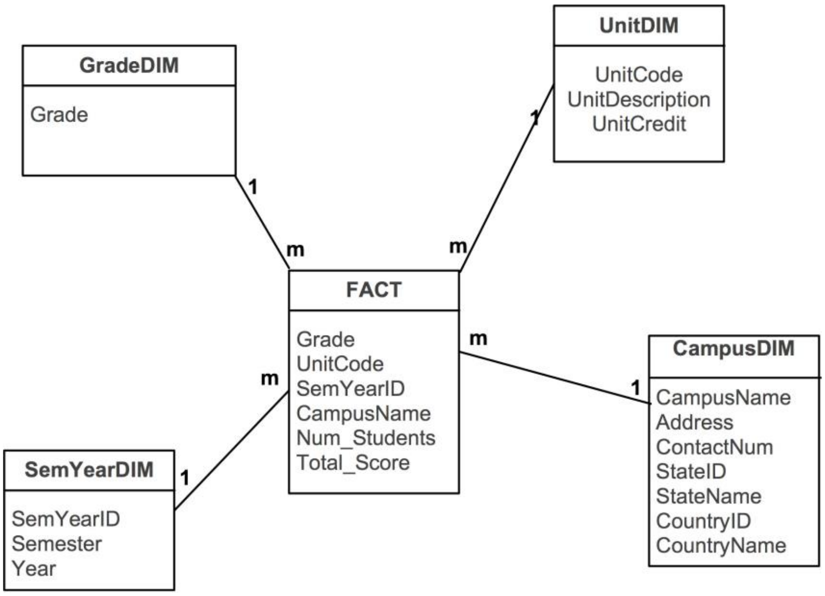
Candidates must complete this section if required to write answers within this paper

STUDENT ID: __ __ __ __ __ __ __ __ DESK NUMBER: __ __ __ __ __

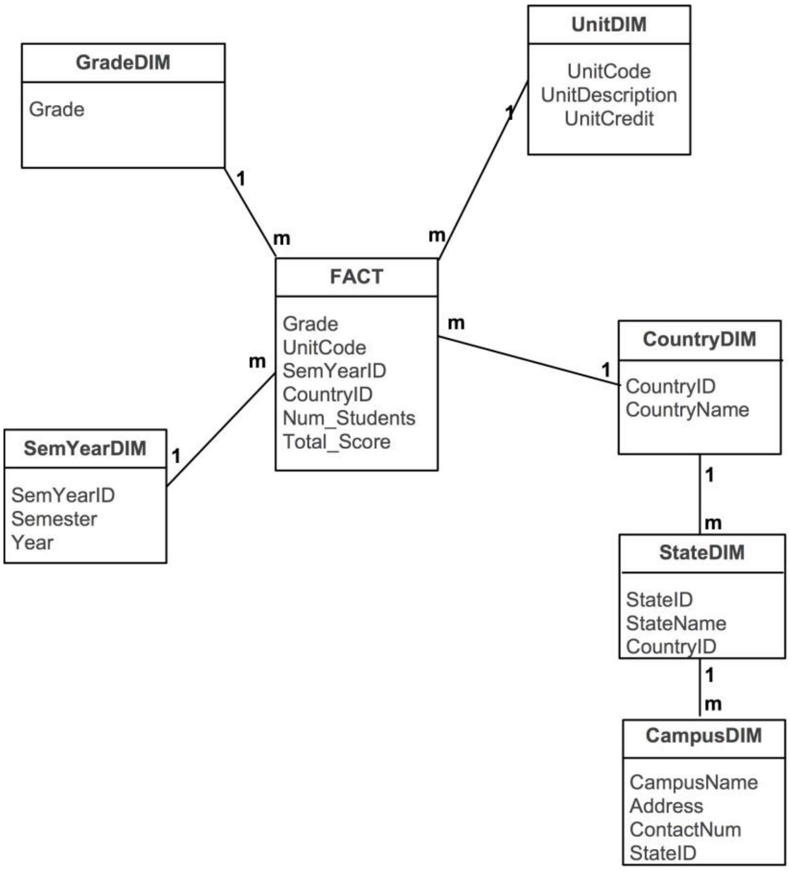
Question 1:

Consider the following Student Enrolment star schemas: Star Schema Version-1 does not have a dimension hierarchy, whereas Star Schema Version-2 has a dimension hierarchy: from country to state, and to campus.

Star Schema Version-1



Star Schema Version-2



Questions:

a. In contrasting both star schemas, is there any mistake in any of the two star schemas (Note that Star Schema Version-1 does not have a hierarchy, and Star Schema Version-2 does have)?
   - If yes, state which star schema, and explain your reason.
   - If no, also explain your reason.

b. Compare both star schemas.
   - If there are mistakes in any (or both) star schemas, you need to draw the correct schema(s) first before comparing between each other.
   - If there are no mistakes in both star schemas, you can immediately compare the two star schemas.

   Also, when you compare the two star schemas, you need to use some sample data (in the fact and in certain dimensions) to support your arguments

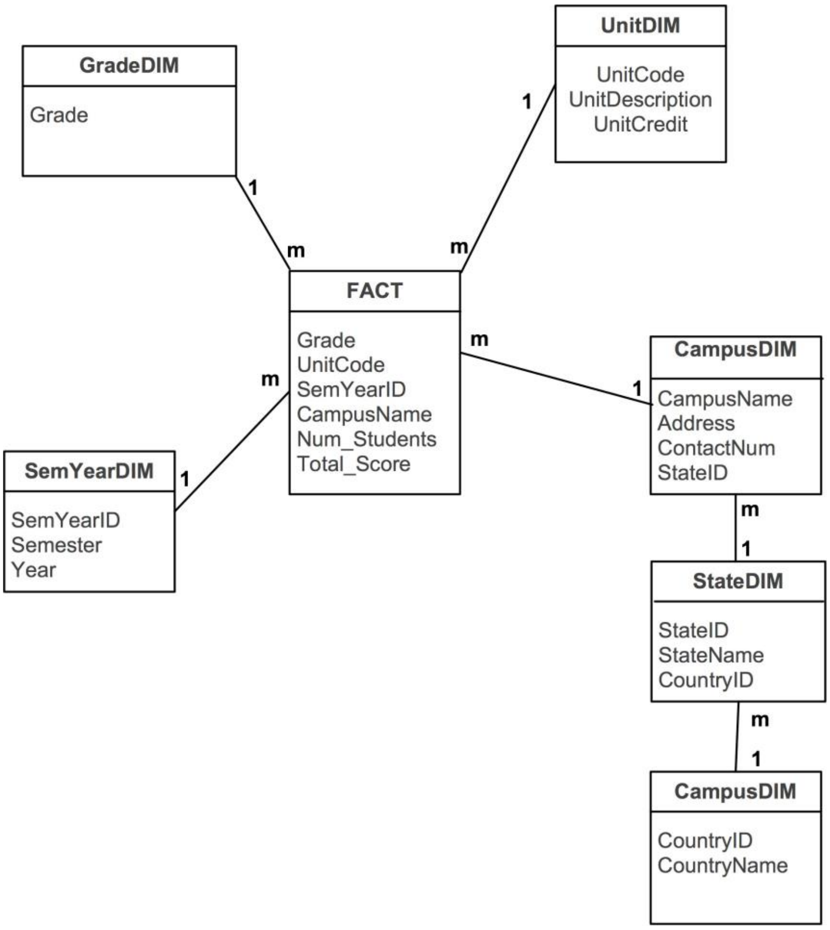
Write your answers here:

(a) There is a mistake in Star Schema Version-2; the mistake is in the hierarchy. The hierarchy should start from the most detail (e.g. Campus) to the most general (e.g. Country). Hence, the correct hierarchy should be CampusDIM➔StateDIM➔CountryDIM, and not in the opposite direction. Consequently, the fact should have CampusName, instead of CountryID.

There is no mistake in Star Schema Version-1.

Continue your answers here:

(b) The correct star schema for version-2 is as follows:
Note: the FK must also be correct.



Continue your answers here:

Data duplication or Normalization

Star Schema-1: unnormalized, has data duplication

The corrected (new) Star Schema-2: normalized, minimized data duplication

Minimise Join

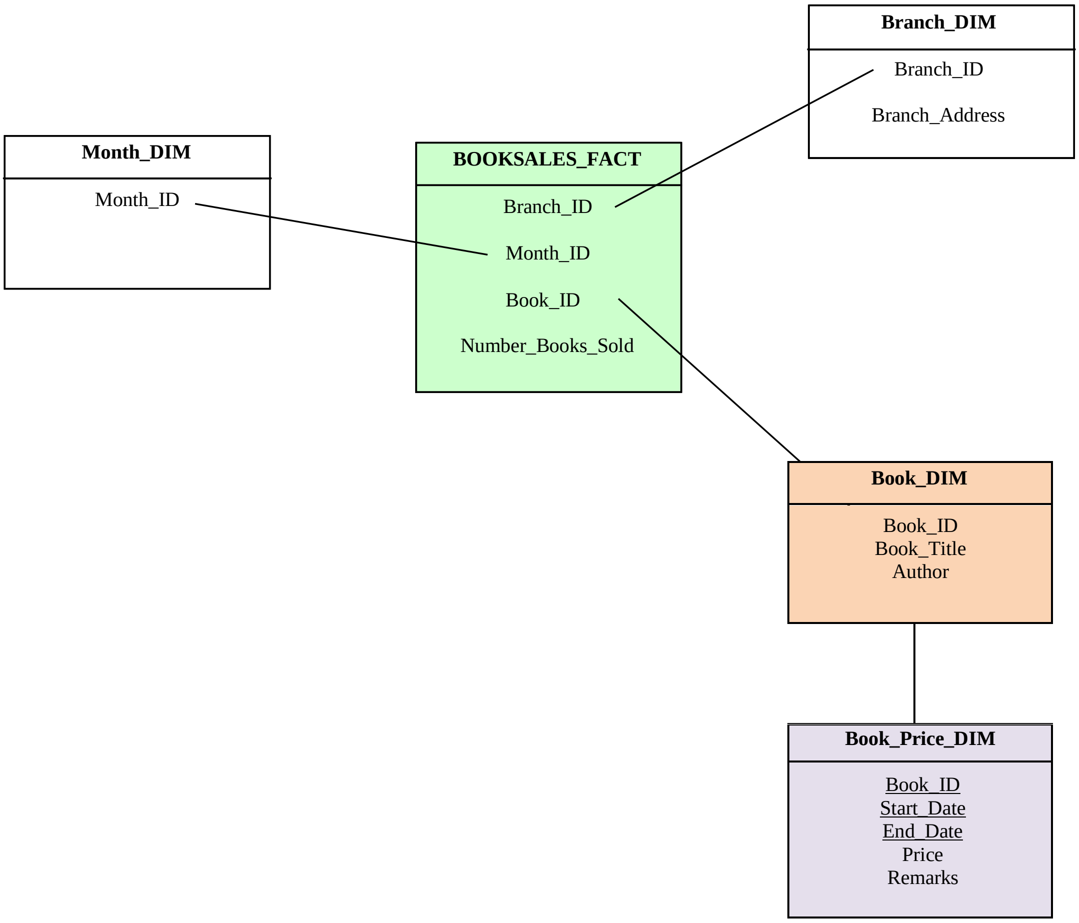
Star Schema-1: need only one join between Fact and CampusDIM

The corrected (new) Star Schema-2: need three join operations between Fact, CampusDIM, StateDIM, and CountryDIM

For example: when we answer a query “how many students from campus in Australia”, Star Schema-1 needs to join Fact with CampusDIM only, whereas Star Schema-2 needs to join tables Fact, CampusDIM, StateDIM, and CountryDIM.

Question 2

This question is taken from the *Bookshop* Case Study on Temporal Data Warehousing. The following shows a star schema shows a fact table (number of books sold) and three dimensions (e.g. Month, Branch, and Book). The Book dimension is temporal dimension, which contains a temporal attribute, called Price, which is book price.



The tables for this star schema have been created and populated from the operational database. The sample data is as follows:

Month_DIM Table

Month_ID
201503
201502
201501
201412
etc

Branch_DIM Table

Branch_ID	Branch_Address
City	Melbourne Central Shopping Centre, Melbourne
Chadstone	285 Dandenong Road, Chadstone
Camberwell	199 Burke Road, Camberwell
etc	

Book_DIM Table

Book_ID	Book_Title	Author
C1	CSIRO Diet	CSIRO Team
H6	Harry Potter 6	Rowling
DV	Da Vinci Code	Dan Brown
…	…	…

Book_Price_DIM Table

Book_ID	Start_Date	End_Date	Price	Remarks
C1	201401	201407	$45.95	Full Price
C1	201408	201410	$36.75	20% Discount
C1	201411	201501	$23.00	Half Price
C1	201502	201512	$45.95	Full Price
H6	201401	201403	$21.95	Launching
H6	201404	201501	$30.95	Full Price
H6	201502	201512	$10.00	End of Product Sale
DV	201401	201512	$27.95	Full Price
…	…	…	…	

BookSales_Fact Table

Month_ID	Branch_ID	Book ID	Number_Books_Sold
201503	City	C1	5
201503	City	H6	15
201503	City	DV	23
201503	City	…	
201503	Chadstone	C1	15
201503	Chadstone	H6	3
201503	Chadstone	DV	2
201503	Chadstone	…	
201503	Camberwell	C1	1
201503	Camberwell	H6	1
201503	Camberwell	DV	2
201503	Camberwell	…	
201503	…	…	
…	…	…	
201412	City	C1	15
201412	City	H6	6
201412	City	DV	6
201412	City	…	
201412	Chadstone	C1	10
201412	Chadstone	H6	8
201412	Chadstone	DV	1
201412	Chadstone	…	
201412	Camberwell	C1	18
201412	Camberwell	H6	3
201412	Camberwell	DV	2
201412	Camberwell	…	
201412	…	…	
…	…	…	

Question:
Write the SQL command to produce the following report (**10 marks**):

Month_ID	Branch_ID	Book_ID	Book_Title	Author	Price	Number_Books _Sold
201503	City	C1	CSIRO Diet	CSIRO Team	$45.95	5
201503	City	H6	Harry Potter 6	Rowling	$10.00	15
201503	City	DV	Da Vinci Code	Dan Brown	$27.95	23
201503	City		…			
201503	Chadstone	C1	CSIRO Diet	CSIRO Team	$45.95	15
201503	Chadstone	H6	Harry Potter 6	Rowling	$10.00	3
201503	Chadstone	DV	Da Vinci Code	Dan Brown	$27.95	2
201503	Chadstone		…			
201503	Camberwell	C1	CSIRO Diet	CSIRO Team	$45.95	1
201503	Camberwell	H6	Harry Potter 6	Rowling	$10.00	1
201503	Camberwell	DV	Da Vinci Code	Dan Brown	$27.95	2
201503	Camberwell		…			
201503	…		…			
…	…		…			
…	…		…			
201412	City	C1	CSIRO Diet	CSIRO Team	$23.00	15
201412	City	H6	Harry Potter 6	Rowling	$30.95	6
201412	City	DV	Da Vinci Code	Dan Brown	$27.95	6
201412	City		…			
201412	Chadstone	C1	CSIRO Diet	CSIRO Team	$23.00	10
201412	Chadstone	H6	Harry Potter 6	Rowling	$30.95	8
201412	Chadstone	DV	Da Vinci Code	Dan Brown	$27.95	1
201412	Chadstone		…			
201412	Camberwell	C1	CSIRO Diet	CSIRO Team	$23.00	18
201412	Camberwell	H6	Harry Potter 6	Rowling	$30.95	3
201412	Camberwell	DV	Da Vinci Code	Dan Brown	$27.95	2
201412	Camberwell		…			
201412	…		…			
…	…		…			

The structures of the above tables are as follows:

```
SQL> desc Month_DIM;
 Name                                          Null?    Type
--------------------------------------------- -------- ------------------------
 MONTH_ID                                               VARCHAR2(6)

SQL> desc Branch_DIM;

 Name                                          Null?    Type
 -------------------------------------------- -------- ------------------------
 BRANCH_ID                                              VARCHAR2(15)
 BRANCH_ADDRESS                                         VARCHAR2(50)
```

```
SQL> desc Book_DIM;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 BOOK_ID                                            VARCHAR2(5)
 BOOK_TITLE                                         VARCHAR2(20)
 AUTHOR                                             VARCHAR2(20)

SQL> desc Book_Price_DIM;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 BOOK_ID                                            VARCHAR2(5)
 START_DATE                                         VARCHAR2(6)
 END_DATE                                           VARCHAR2(6)
 PRICE                                              NUMBER(6,2)
 REMARKS                                            VARCHAR2(20)

SQL> desc BookSales_Fact;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 MONTH_ID                                           VARCHAR2(6)
 BRANCH_ID                                          VARCHAR2(15)
 BOOK_ID                                            VARCHAR2(5)
 NUMBER BOOKS SOLD                                  NUMBER
```

<u>Write your answer here:</u>

```
Select
   F.Month_ID,
   F.Branch_ID,
   F.Book_ID,
   B.Book_Title,
   B.Author,
   P.Price,
   F.Number_Books_Sold
From BookSales_Fact F, Book_DIM B, Book_Price_DIM P
Where F.Book_ID = B.Book_ID
And B.Book_ID = P.Book_ID
And F.Month_ID >= P.Start_Date
And F.Month_ID <= P.End_Date;
```

```
MONTH_ID  BRANCH_ID       BOOK_ AUTHOR                    PRICE NUMBER_BOOKS_SOLD
--------- --------------- ----- -------------------- ---------- -----------------
201503    City            C1    CSIRO Team                45.95                 5
201503    City            H6    Rowling                      10                15
201503    City            DV    Dan Brown                 27.95                23
201503    Chadstone       C1    CSIRO Team                45.95                15
201503    Chadstone       H6    Rowling                      10                 3
201503    Chadstone       DV    Dan Brown                 27.95                 2
201503    Camberwell      C1    CSIRO Team                45.95                 1
201503    Camberwell      H6    Rowling                      10                 1
201503    Camberwell      DV    Dan Brown                 27.95                 2
201412    City            C1    CSIRO Team                   23                15
201412    City            H6    Rowling                   30.95                 6

MONTH_ID  BRANCH_ID       BOOK_ AUTHOR                    PRICE NUMBER_BOOKS_SOLD
--------- --------------- ----- -------------------- ---------- -----------------
201412    City            DV    Dan Brown                 27.95                 6
201412    Chadstone       C1    CSIRO Team                   23                10
201412    Chadstone       H6    Rowling                   30.95                 8
201412    Chadstone       DV    Dan Brown                 27.95                 1
201412    Camberwell      C1    CSIRO Team                   23                18
201412    Camberwell      H6    Rowling                   30.95                 3
201412    Camberwell      DV    Dan Brown                 27.95                 2

18 rows selected.
```

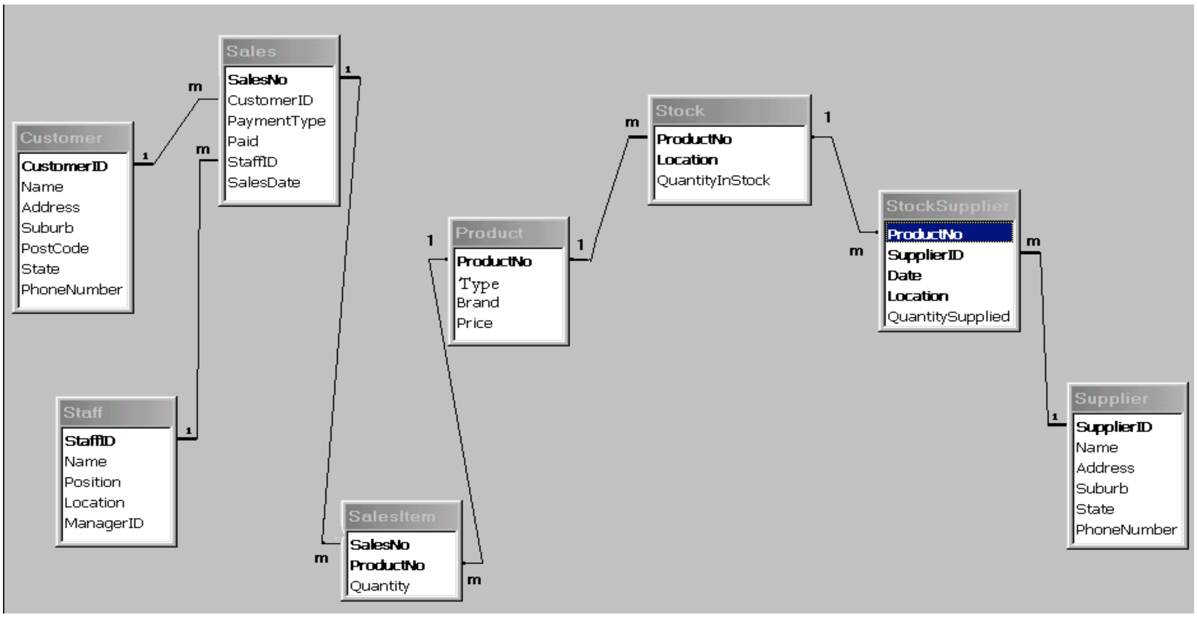
Question 3

This question is taken from the *Product-Sales-Supplier* Case Study.

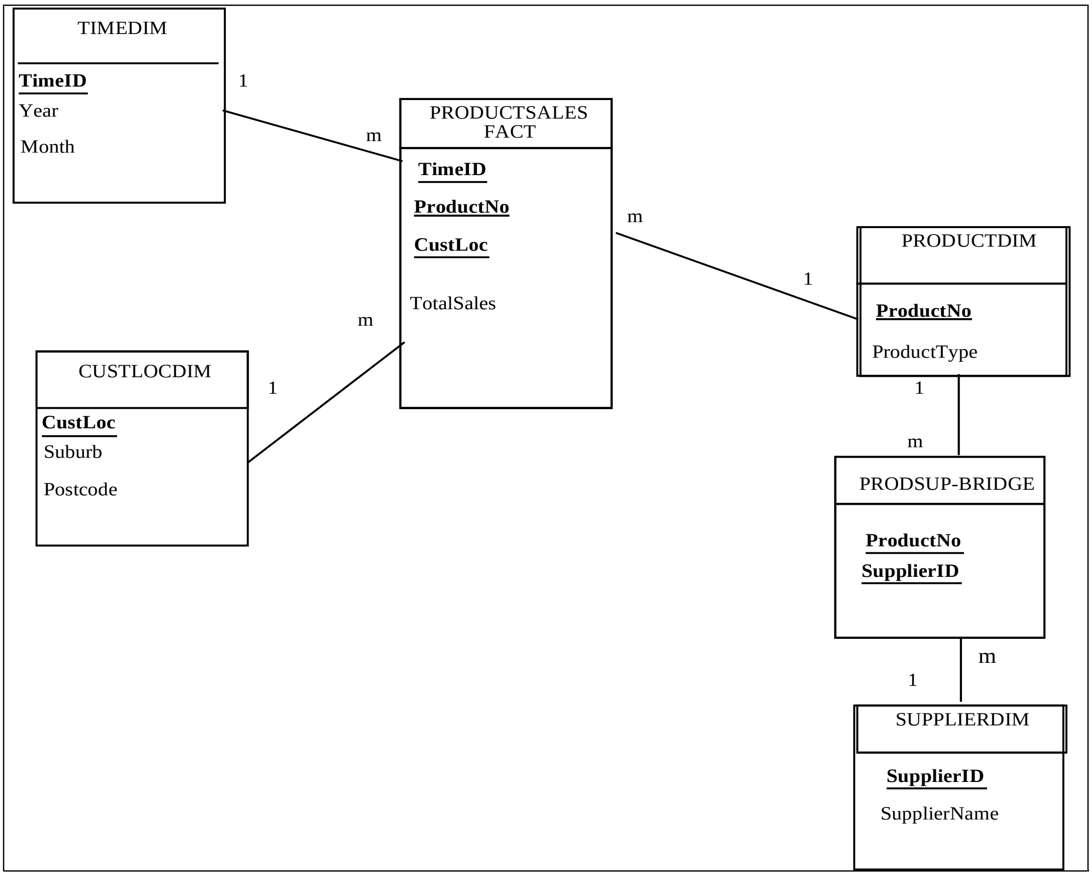
The director of a company is interested in analyzing the statistics of its product sales history. The analysis is needed for identifying which products are popular, which suppliers supply those products, when is the best time to purchase more stock, etc. You are required to design a small Data Warehouse to keep track of the statistics.

The director is particularly interested in analyzing the *total sales* (Quantity * Price) by *product*, *customer locations (suburbs and postcodes)*, *sales time periods* (monthly and yearly), and *supplier*.

The operational database currently has the following tables:



Your snowflake schema will have a Bridge Table connecting Product Dimension and Supplier Dimension. A snowflake schema with a Bridge Table as shown below:



The above snowflake schema is missing two attributes: **WeightFactor** attribute, and **ListAGG** attribute.

Questions:

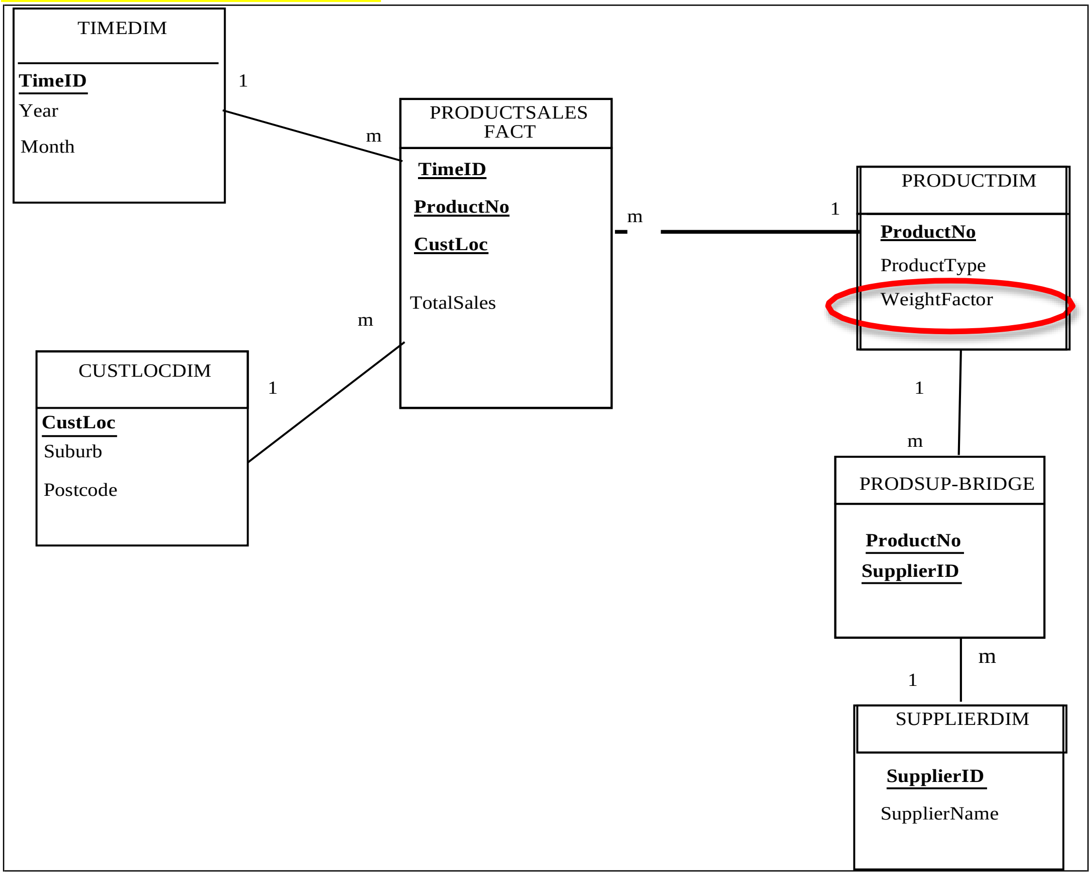
a. **Draw a new snowflake schema** (call it Snowflake Schema version 2) for the above case study, but this new snowflake schema must **use a WeightFactor attribute** (**without ListAGG attribute**). You also need to **show sample records** in the Product Dimension, the Bridge Table, and the Supplier Dimension. The sample data must show the correct values for the Weight attribute. Make sure that in your snowflake schema, the attributes are clearly shown.

b. **Draw another snowflake schema** (call it Star Schema version 3), which also has a Bridge Table and a WeightFactor attribute. But version-3 snowflake schema has the **ListAGG** attribute. You also need to **show sample records** in the Product Dimension, the Bridge Table, and the Supplier Dimension. The sample data must show the correct values for the Weight and ListAGG attributes.

c. Write the **SQL query** to create the ProductDim table for the Star Schema version 3.

Write your answer here:

Snowflake Schema Version 2



ProductDIM Table

ProductNo	ProductType	WeightFactor
P1	Shoes	0.5
P2	Jeans	0.33
etc		

ProdSup_Bridge Table

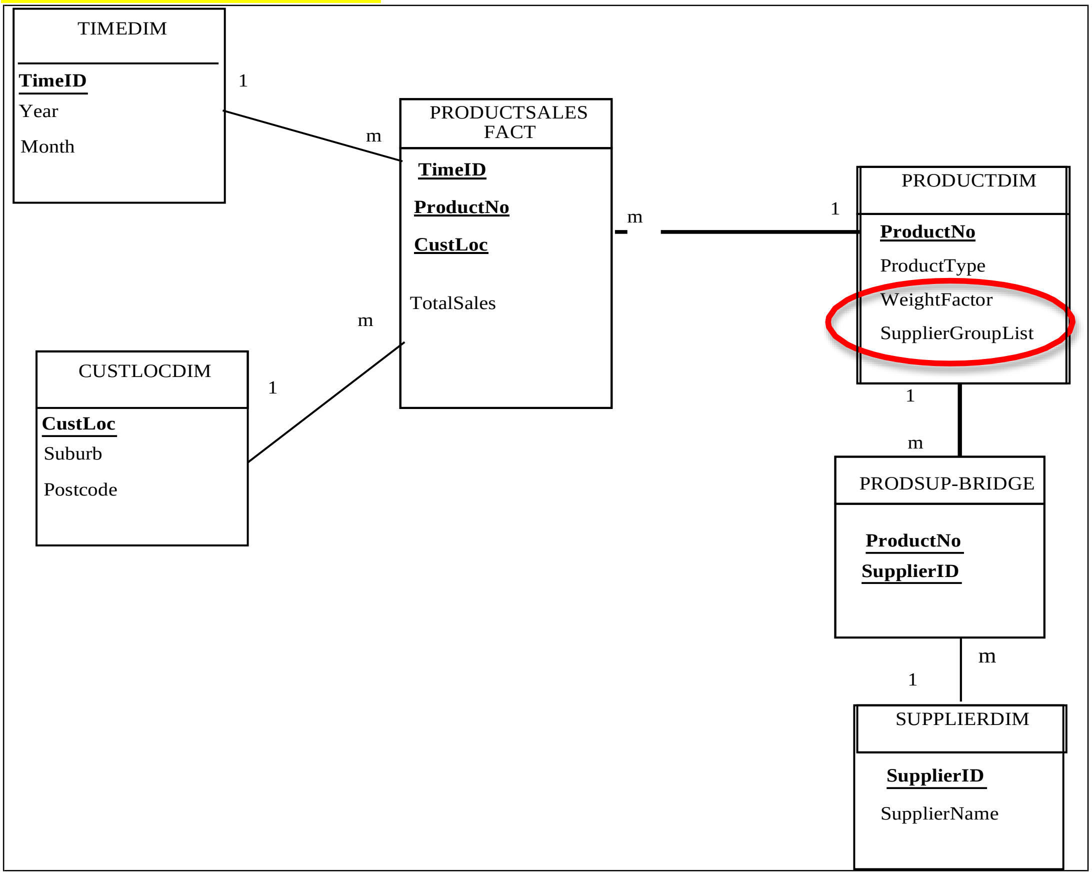
ProductNo	SupplierID
P1	S1
P1	S2
P2	S2
P2	S3
P2	S4
etc	

SupplierDIM Table

SupplierID	SupplierName
S1	Supplier-1
S2	Supplier-2
S3	Supplier-3
S4	Supplier-4
etc	

Continue your answer here:

Snowflake Schema Version 3



ProductDIM Table

ProductNo	ProductType	WeightFactor	SupplierGroupList
P1	Shoes	0.5	S1_S2
P2	Jeans	0.33	S2_S3_S4
etc			

ProdSup_Bridge Table

ProductNo	SupplierID
P1	S1
P1	S2
P2	S2
P2	S3
P2	S4
etc	

SupplierDIM Table

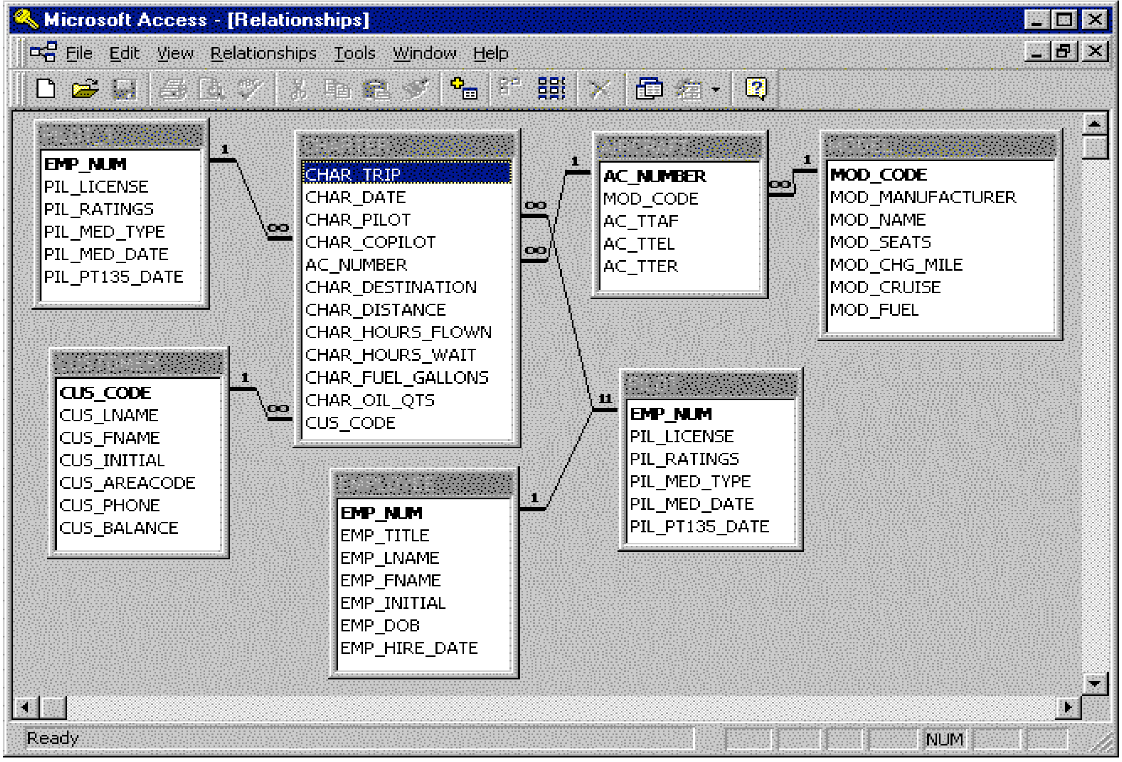
SupplierID	SupplierName
S1	Supplier-1
S2	Supplier-2
S3	Supplier-3
S4	Supplier-4
etc	

Continue your answer here:

```
Create Table ProductDim As
Select
   P.ProductNo,
   P.ProductType,
   1.0/count(SS.SupplierID) as WeightFactor,
   LISTAGG (SS.SupplierID, '_') Within Group
           (Order By SS.SupplierID) As SupplierGroupList
From Product P, Stock S, StockSupplier SS
Where P.ProductNo = S.ProductNo
And S.ProductNo = SS.ProductNo
And S.Location = SS.Location
Group By P.ProductNo, P.ProductType;
```

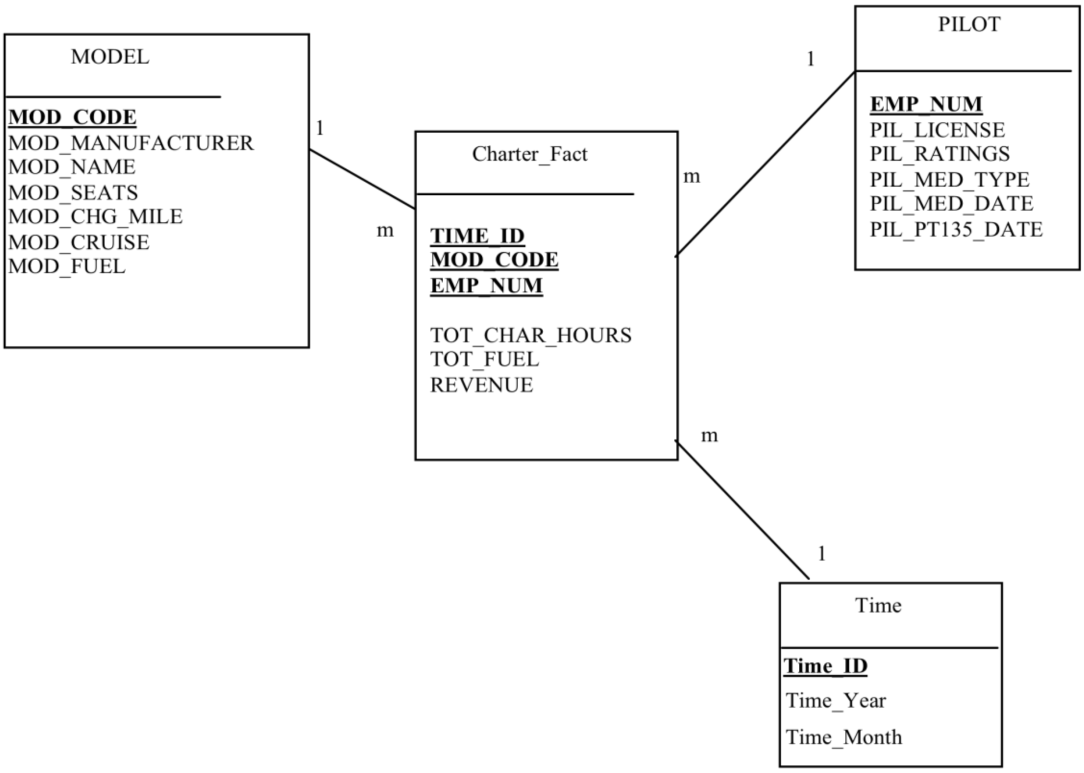
Question 4

This question is based on the Robcor case study. The following is the E/R diagram of the operational database in the Robcor case study:



A star schema for the above operational database is shown as follows:

Robcor Star Schema-1

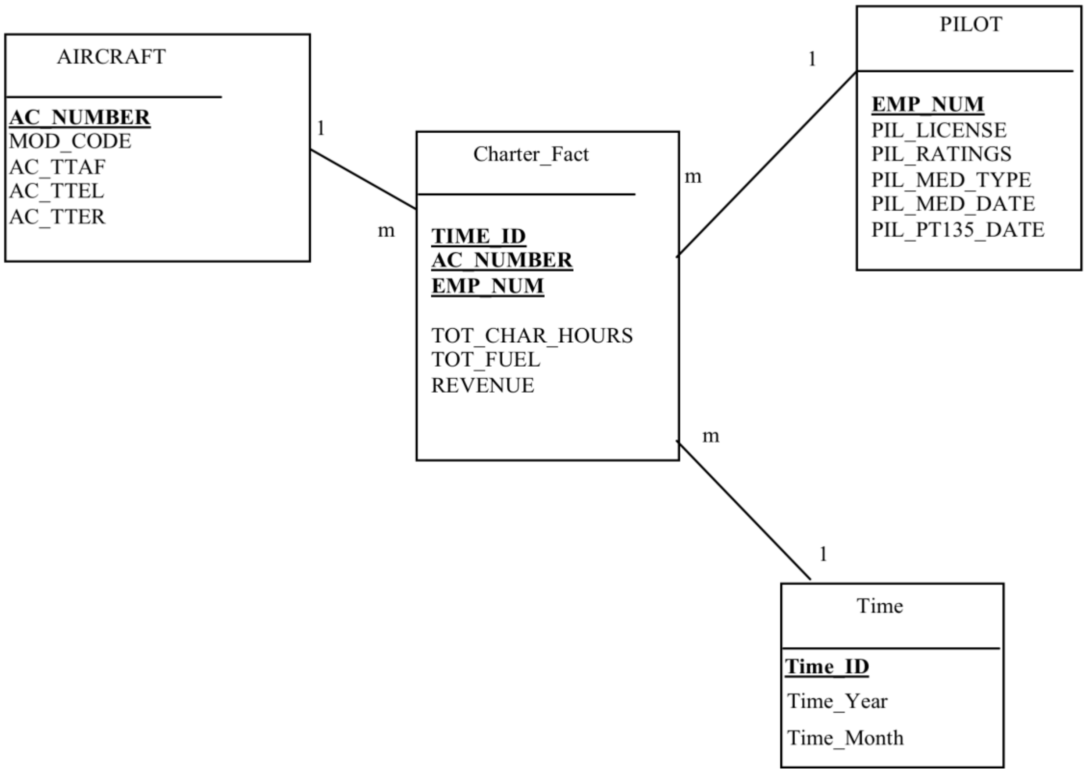


Questions:

a. Is it possible to determine which level Robcor Star Schema-1 is? If it is possible, state the level and also give the reason. If it is not possible to state the level, then give the reason.

b. Let's have a look at the following star schema (Robcor Star Schema-2). Between the two star schemas (Robcor Star Schema-1 and Robcor Star Schema-2), which one has a higher level of aggregation? State the name of the star schema, whether it is Robcor Star Schema-1 or Robcor Star Schema-2, and explain the reason.

Robcor Star Schema-2



Write your answers here:

(a)

It is not possible to determine whether this star schema is on level-2 or on level-3 or on a higher level. What we know is that Robcor Star Schema-1 is not on level-1.

Reason: Robcor Star Schema-1 is not on the lowest level, because some of the dimensions on a higher level of aggregation (e.g. time id which is based on month, instead of the actual charter date).

However, it is not possible to name whether this is level 2 or level 3, because there can be any schemas in between Robcor star-schema-1 and level-1.

If there is no star schema in between level-1 and Robcor star schema-1, then Robcor star schema-1 becomes level 2.
If there is a star schema in between level-1 and Robcor star schema-1, then Robcor star schema-1 is level 3.
If there are two star schemas in between level-1 and Robcor star schema-1, then obviously Robcor star schema-1 becomes level 4.

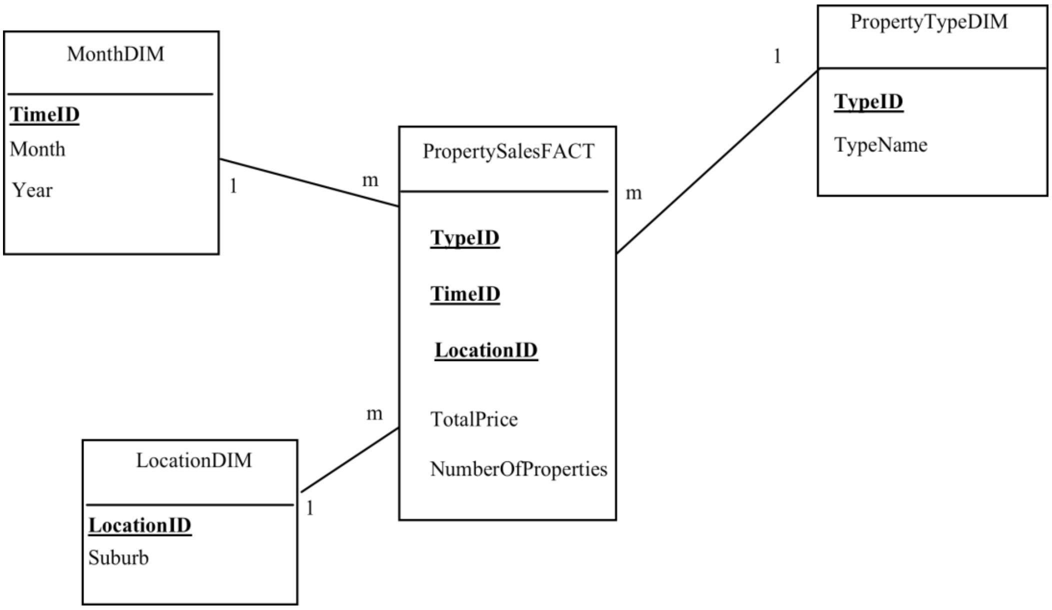
(b)

Robcor star schema-1 has a higher level of aggregation than Robcor star schema-2.

Reason: one Model can have multiple Aircrafts. Hence a star schema using Model as a dimension has a higher level of aggregation than a star schema using Aircraft as a dimension

Question 5

Given the following star schema:



The tables (e.g. Fact and three dimensions) have been created and have also been populated with an adequate number of records. The table names and attributes are shown in the star schema above.

Write the SQL for the following OLAP queries:

a. Display the top 10 average prices by suburb of property

b. Display the average price of properties by property type description and suburb. It is not required to show the subtotals or group totals or grand total

Write your answer here:

Solution a

```
Select *
From
      (Select P.LocationID, L.Suburb,
           Sum(F.TotalPrice)/Sum(F.NumberOfProperties)  as AveragePrice,
           RANK() OVER
             (ORDER BY Sum(F.TotalPrice)/Sum(F.NumberOfProperties) DESC)
             as PROPERTY_RANK
       From      PropertyFact P, LocationDim L
       Where     P.LocationID = L.LocationID
       Group by P.LocationID, L.Suburb)
Where PROPERTY_RANK <= 10;
```

Solution b :

```
SELECT T.TypeName, L.Suburb,
       Sum(F.TotalPrice)/Sum(F.NumberOfProperties)  as AveragePrice
FROM   PropertyFACT F, PropertyTypeDIM T, LocationDIM L
WHERE  F.TypeID = T.TypeID
AND    F.LocationID = L.LocationID
GROUP BY T.TypeName, L.Suburb;
```

Question 6

This question is about Top *n*% and Top *k* (such as Top 10% and Top 3) in OLAP. The tables are based on the ROBCOR data warehouse case study, which consists of one fact and three dimension tables: `charter_fact`, `time`, `pilot`, and `model`.

```
SQL> desc charter_fact;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 TIME_ID                                            VARCHAR2(6)
 MOD_CODE                                           CHAR(10)
 EMP_NUM                                            NUMBER(10)
 TOT_CHAR_HOURS                                     NUMBER
 TOT_FUEL                                           NUMBER
 REVENUE                                            NUMBER

SQL> desc time;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 TIME_ID                                            CHAR(6)
 TIME_YEAR                                          CHAR(4)
 TIME_MONTH                                         CHAR(2)

SQL> desc pilot;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 EMP_NUM                                            NUMBER(10)
 PIL_LICENSE                                        CHAR(25)
 PIL_RATINGS                                        CHAR(25)
 PIL_MED_TYPE                                       CHAR(1)
 PIL_MED_DATE                                       DATE
 PIL_PT135_DATE                                     DATE

SQL> desc model;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 MOD_CODE                                           CHAR(10)
 MOD_MANUFACTURER                                   CHAR(15)
 MOD_NAME                                           CHAR(20)
 MOD_SEATS                                          FLOAT(126)
 MOD_CHG_MILE                                       NUMBER(19,4)
 MOD_CRUISE                                         FLOAT(126)
 MOD_FUEL                                           FLOAT(126)
```

Questions:

a. Write the SQL command to display the time periods which had the revenue in the top 10% of the months.

The result should be like this:

```
TIME_ID          TOTAL  PERCENT_RANK
----------- ---------------- -------------------
199503       51144.16             1
199408       49775.51    .975609756
199510       48538.01    .951219512
199409       47647.75    .926829268
199703       45872.32    .902439024
```

b. Write the SQL command to display the `mod_code` and `mod_name` of the two airplanes that have the largest total fuel used.

The result should look like this:

```
MOD_CODE        MOD_NAME                                TOTAL           MYRANK
---------------- ------------------------------- ---------------- ----------------
PA31-350        Navajo Chieftain                      83790.5                1
C-90A           KingAir                               61708.4                2
```

Write your answer here:

a:

```
SELECT dw.time.time_id, Total, percent_rank
FROM (
  SELECT
    time_id,
    SUM(revenue) AS Total,
    PERCENT_RANK () OVER (ORDER BY SUM(revenue)) AS percent_rank
  FROM dw.charter_fact
  GROUP BY time_id
) t, dw.time
WHERE t.time_id = dw.time.time_id
AND percent_rank >= 0.9
ORDER BY percent_rank DESC;
```

b:

```
SELECT *
FROM (
      SELECT m.mod_code, m.mod_name,
         SUM(f. tot_fuel) AS total,
         RANK() OVER (ORDER BY SUM(f. tot_fuel) DESC) AS myrank
      FROM dw.charter_fact f, dw.model m
      WHERE f.mod_code = m.mod_code
      GROUP BY m.mod_code, m.mod_name
)
WHERE myrank <=2;
```

THE END