BSc, BEng and MEng Degrees Examination 2020—21

DEPARTMENT OF COMPUTER SCIENCE

Data Analysis and Management

ANSWERS

Time Allowed: TWENTY-FOUR hours
Time Recommended: 2 hours 30 minutes

Word limit: NA

Allocation of Marks:
This assessment is out of 100 marks:

- Q1 (14 marks) | Q2 (14 marks) | Q3 (22 marks) | Q4 (9 marks)
- Q5 (11 marks) | Q6 (13 marks) | Q7 (17 marks)

Instructions:

- Answer **all** questions.
- For some of the questions you will need: Jupyter notebook for Python 3 and SQLite DB Browser (version 3.12); and some files, which are available on VLE (Assessment -- 24_Hour_Exam).
- Submit your answers to the Department's Teaching Portal as a zipped directory containing four files:
  - Data2.pdf (This file must include all answers (such as text, SQL commands, code, figures) for all questions
  - Data2_Q5.ipynb (This file must include the code as part of your answers for Q5)
  - Data2_Q6.ipynb (This file must include the code as part of your answers for Q6)
  - Data2_Q7.ipynb (This file must include the code as part of your answers for Q7)

  Note: The **'ipynb'** files will be used to test and run your solution

If a question is unclear, answer the question as best you can, and note the assumptions you have made to allow you to proceed.

A note on Academic Integrity

We are treating this online examination as a time-limited open assessment, and you are therefore permitted to refer to written and online materials to aid you in your answers.

However, you must ensure that the work you submit is entirely your own, and for the whole time the assessment is live you must not:

- communicate with departmental staff on the topic of the assessment
- communicate with other students on the topic of this assessment.
- seek assistance with the assignment from the academic and/or disability support services, such as the Writing and Language Skills Centre, Maths Skills Centre and/or Disability Services. (The only exception to this will be for those students who have been recommended an exam support worker in a Student Support Plan. If this applies to you, you are advised to contact Disability Services as soon as possible to discuss the necessary arrangements.)
- seek advice or contribution from any third party, including proofreaders, friends, or family members.

We expect, and trust, that all our students will seek to maintain the integrity of the assessment, and of their award, through ensuring that these instructions are strictly followed. Failure to adhere to these requirements will be considered a breach of the Academic Misconduct regulations, where the offences of plagiarism, breach/cheating, collusion and commissioning are relevant - see AM.1.2.1" *(Note this supersedes section 7.3 of the Guide to Assessment).*

ANSWERS

Q1. [Total: 14 Marks]

Go through the following case study, and then answer the questions that follow:

ABC Ltd. is an online supplier who requires a database system for processing sales orders. Customer accounts are created when they place their first order, with many orders can be placed. According to the company's policy, each order can only be from one customer.

Each order placed by a customer can consist of one or many items, but a given item refers to exactly one order. Each order raises an invoice, and an invoice belongs to one order. A product could be part of one or many items, but each item refers to only one product.

Each order is then processed by a single employee on the sales accounts team, who prepares the shipment of the goods to the customer. An employee can be scheduled for processing many separate orders and preparing many shipments each day, but may not have any schedules for both tasks at all.

Items in each order can be packaged together into one shipment, or packaged separately and sent by several shipments, if required. A shipment can only belong to one order.

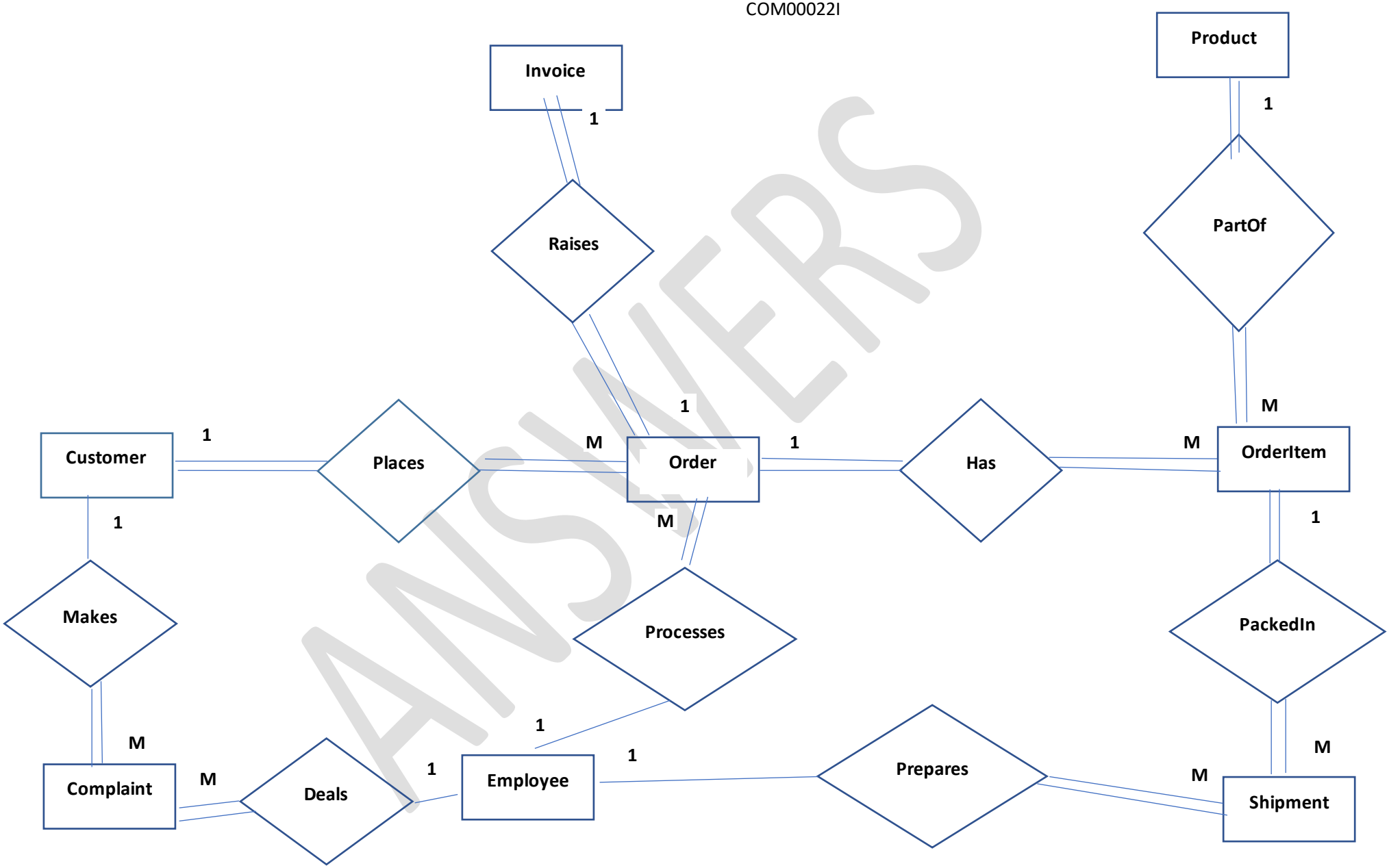
A customer may not have any complaints to make, but they can make as many they want. Each complaint comes from one customer, which is then dealt by one employee. Not all employees will handle the complaints, and sometimes an employee may have to deal with many complaints.

Task:

(i) [6 marks] You are required to draw the Entity Relationship Model (ERM) for the proposed system, clearly showing all entities, relationships among participation entities, multiplicity constraints and cardinalities.

Solution: Please refer to next page.

ANSWERS


Invoice
1
Raises
Product
1
PartOf
1
M
Customer
1
Places
M
Order
1
Has
M
OrderItem
1
M
1
Makes
Processes
PackedIn
M
1
M
Complaint
M
Deals
1
Employee
1
Prepares
M
Shipment

ANSWERS

Note: In the ERD we did not show any weak entities as they were not required in the question.

(ii) [6 marks] From (i):

(a) [2 marks] What is the entity pair involved in the relationship for raising an order?

In the question the word 'raising' has created ambiguity for some students. Most students answered 'Customer and Order' and few of them answered 'Order and Invoice'. We have accepted both answers.

(b) [1 mark] From (a) what is the cardinality of this relationship?

For those who answered 'Customer and Order': One to Many

For those who answered 'Order and Invoice': One to One

(c) [1 mark] From (a) state with explanation which entity is the parent and which entity is the child. If there is not a parent and a child, you still need to provide a justification.

For those who have answered 'Customer and Order':
The entity on the one side (which is Customer) is the parent and the entity on the many side (which is Order) is the child.

For those who have answered 'Order and Invoice':
Since we have total participation on both sides, we do not have any parent and child

(d) [2 marks] From (a) when mapping the ERM to tables, how will the relationship between these two entities be represented?

For those who answered 'Customer and Order':

We will have two tables: Customer (parent) and Order (child). A copy of the primary key of the parent will be designated to the child as foreign

ANSWERS

key.

For those who answered 'Order and Invoice':

The two entities need to be combined in a single table. Of the original primary keys specified for each entity, one is designated as the primary key for the table, the other is retained as an alternate key

(iii) [2 marks] Are there any weak entities ? If yes, list them.

Item, Complaint, Invoice, Shipment. Since we do not know any attributes for the entities, other possible weak entities could be Order and Product.

ANSWERS

Q2. [Total: 14 Marks]

The following sample table (refer to Figure 1) details students who are being enrolled on one or more modules, their grades, and the tutors who are teaching these modules. The composite key will be (sID, moduleNo).

sID	moduleNo	moduleName	grade	sName	tutorID	tutorName	tutorOffice
S101	M1001	AI	A	TC	T1	ZD	R121
S102	M1002	HCI	B	SR	T2	AS	R220
S103	M1003	DI	C	SG	T3	BN	R010
S104	M1003	DI	B	KP	T3	BN	R010
S101	M1004	SE	A	TC	T4	CF	R111
S103	M1002	HCI	A	SG	T2	AS	R220

Figure 1: Students, grades, and tutors

Task:

(i) [2 marks] Which columns in the table contain redundant data? [moduleName, sName, tutorID, tutorName, tutorOffice]

(ii) [4 marks] List all functional dependencies which exist in the table.

- sID,moduleNo - -> grade
- sID - -> sName
- moduleNo - -> moduleName
- moduleNo - -> tutorID
- moduleNo - -> tutorName
- moduleNo - -> tutorOffice
- tutorID - -> tutorName,
- tutorID - -> tutorOffice

Other functional dependencies:

- *tutorName - -> tutorOffice*
- *tutorOffice -- > tutorName*

- *tutorID - - > moduleName*
- *moduleName - -> tutorID*

- *tutorName - -> tutorID*
- *tutorOffice - -> tutorID*

- *moduleName - - > tutorName*
- *tutorName - - > moduleName*

- *moduleName - - > tutorOffice*
- *tutorOffice - - > moduleName*

ANSWERS

(iii) [3 marks] Of the functional dependencies that **actually exist** which, if any, violate the rules of the 2NF?

- sID - -> sName [partial]
- moduleNo - - > moduleName, tutorID, tutorName, tutorOffice [partial]

(iv) [2 marks] How could you resolve the issues in (iii), whilst maintaining the required relationship?

- Move moduleName, tutorID, tutorName, tutorOffice to a new table with moduleNo as the primary key
- Move sName to a new table with sID as the primary key

(v) [3 marks] Having answered (iv), show all the tables that are required, clearly showing the primary keys, foreign keys, and any composite keys.

- StudentGrade: PK/FK: sID |PK/FK: moduleNo | grade
- Module: PK: moduleNo | moduleName | tutorID | tutorName | tutorOffice
- Student: PK: sID| sName

For Question 3 you need SQLite 3.12 DB Browser

Q3. [Total: 22 Marks]

DVD XYZ is a simple database designed to support the business activities of a DVD rental library. The multiplicity constraints are as follows:

- A distribution center has one or many members of staff. Each staff member works at one distribution center.
- A staff member manages zero or one distribution center. Each distribution center is managed by staff member.
- A DVD has one or many copies. Each DVD copy belongs to one DVD.
- A supplier supplies one or many DVDs. Each DVD is supplied by one supplier.
- A distribution center has one or many DVD copies. Each DVD copy is at one distribution center.

With regards to the above information, a set of tables has been designed, consisting of the following five tables.

ANSWERS

A_DistributionCenter Table

dCenterNo	dStreet	dCity	dState	dZipCode	staffNo
B001	8 Jefferson Way	Portland	OR	97201	S1500
B002	City Center Plazza	Seatle	WA	98122	S0010
B003	14 -8th Avenue	New York	NY	10012	S0415
B004	16 -14th Avenue	Seatle	WA	98128	S2250

B_Staff Table

staffNo	name	position	salary	dCenterNo
S0003	Sally Adams	Snr Assistant	30000	B001
S0010	Mary Martinez	Manager	50000	B002
S0415	Art Peters	Manager	41000	B003
S1500	Tom Daniels	Manager	46000	B001
S2250	Sally Stern	Manager	48000	B004
S2350	Robert Chin	Supervisor	32000	B002

ANSWERS

C_Supplier Table

supplierNo	name	address	telNo	status
S01	Universal Home Videos	100 Universal City Plaza	8188666000	OK
S02	MGM Home Videos	2500 Broadway St, Santa Monica, CA,90404	8189002000	OK
S03	Buena Vista Pictures	1100 Santa Monica Bivid, CA, 90041	3208406500	OK
S04	Paramount Pictures	5555 Melrose Avenue, Hollywood, CA, 90038	3238621130	OK
S05	20th Century Fox Home Video	900 Center Piaza, Beverly Hills, CA, 90213	6007772300	OK

D_DVD Table

catalogNo	title	genre	rating	supplierNo
207132	Casino Royale	Action	PG-13	S02
330553	Lord of the Rings III	Action	PG-13	S04
445624	Mission Impossible III	Action	PG-13	S03
634817	War of the Worlds	Sci-Fi	PG-13	S05
781132	Shrek 2	Children	PG	S03
902355	Harry Potter	Children	PG	S01

E_DVDCOPY Table

videoNo	available	catalogNo	dCenterNo
178643	False	634817	B001
199004	True	207132	B001
200900	True	330553	B002
210087	True	902355	B002
243431	True	634817	B002
245456	True	207132	B002
245457	True	207132	B002
317411	True	781132	B003

The tables are in the design phase, and they need to be implemented.

Task:

Using **SQLite3.12 DB Browser**, you are required to:

(i) [3 marks] Give the SQL commands to create each table, clearly showing the primary and foreign keys.

```
BEGIN TRANSACTION;
CREATE TABLE A_DistributionCenter(
dCenterNo TEXT PRIMARY KEY NOT NULL,
dStreet TEXT,
dCity TEXT,
dState TEXT,
dZipCode INTEGER,
staffNo TEXT,
FOREIGN KEY(staffNo) REFERENCES B_Staff(staffNo)
ON DELETE CASCADE
ON UPDATE CASCADE
);
COMMIT;
**********************************************

BEGIN TRANSACTION;
CREATE TABLE B_Staff(
staffNo TEXT PRIMARY KEY NOT NULL,
name TEXT,
position TEXT,
salary INTEGER,
dCenterNo TEXT,
FOREIGN KEY(dCenterNo) REFERENCES A_DistributionCenter(dCenterNo)
ON DELETE CASCADE
ON UPDATE CASCADE
);
COMMIT;
***********************************************

BEGIN TRANSACTION;
CREATE TABLE C_Supplier(
supplierNo TEXT PRIMARY KEY NOT NULL,
name TEXT,
address TEXT,
telNo INTEGER,
status TEXT
);
COMMIT;
*****************************************************

BEGIN TRANSACTION;
CREATE TABLE D_DVD(
catalogNo INTEGER PRIMARY KEY NOT NULL,
title TEXT,
genre TEXT,
rating TEXT,
supplierNo TEXT,
FOREIGN KEY(supplierNo) REFERENCES C_Supplier(supplierNo)
```

ANSWERS

```
ON DELETE CASCADE
ON UPDATE CASCADE
);
COMMIT;
**********************************************************

BEGIN TRANSACTION;
CREATE TABLE E_DVDCOPY(
videoNo INTEGER PRIMARY KEY NOT NULL,
available TEXT,
catalogNo INTEGER,
dCenterNo TEXT,
FOREIGN KEY(catalogNo) REFERENCES D_DVD(catalogNo)
FOREIGN KEY(dCenterNo) REFERENCES A_DistributionCenter(dCenterNo)
ON DELETE CASCADE
ON UPDATE CASCADE
);
COMMIT;
```

(ii) [3 marks] Give the SQL commands to populate each table with all the records.

```
INSERT INTO A_DistributionCenter
(dCenterNo,dStreet,dCity,dState,dZipCode)
VALUES ('B001','8 Jefferson Way', 'Portland', 'OR',97201),
('B002','City Center Plaza', 'Seatle', 'WA', 98122),
('B003','14 -8th Avenue', 'New York', 'NY', 10012),
('B004','16 -14th Avenue', 'Seatle', 'WA', 98128);


INSERT INTO B_Staff
VALUES ('S003','Sally Adams','Snr Assistant',30000,'B001'),
('S0010','Mary Martinez','Manager',50000,'B002'),
('S0415','Art Peters','Manager',41000,'B003'),
('S1500','Tom Daniels','Manager',46000,'B001'),
('S2250','Sally Stern','Manager',48000,'B004'),
('S2350','Robert Chin','Supervisor',320000,'B002');


UPDATE A_DistributionCenter
SET staffNo = 'S1500'
WHERE dCenterNo = 'B001';


UPDATE A_DistributionCenter
SET staffNo = 'S0010'
WHERE dCenterNo = 'B002';
```

ANSWERS

```
UPDATE A_DistributionCenter
SET staffNo = 'S0415'
WHERE dCenterNo = 'B003';

UPDATE A_DistributionCenter
SET staffNo = 'S2250'
WHERE dCenterNo = 'B004';
```

```
INSERT INTO C_Supplier
VALUES ('S01','Universal Home Videos','100 Universal City Plaza',818-866-
6000,'OK'),
('S02','MGM Home Videos','2500 Broadway St, Santa Monica,
CA,90404',818-900-2000,'OK'),
('S03','Buena Vista Pictures','1100 Santa Monica Bivid, CA, 90041',320-840-
6500,'OK'),
('S04','Paramount Pictures','5555 Melrose Avenue, Hollywood, CA,
90038',323-862-1130,'OK'),
('S05','20th Century Fox Home Video','900 Center Piaza, Beverly Hills, CA,
90213', 600-777-2300,'OK');
```

```
INSERT INTO D_DVD
VALUES (207132,'Casino Royale','Action','PG-13','S02'),
(330553,'Lord of the Rings III','Action','PG-13','S04'),
(445624,'Mission Impossible III','Action', 'PG-13','S03'),
(634817,'War of the Worlds','Sci-Fi','PG-13','S05'),
(781132,'Shrek 2','Children','PG','S03'),
(902355,'Harry Potter','Children','PG','S01');
```

```
INSERT INTO E_DVDCOPY
VALUES (178643,'False',634817,'B001'),
(199004,'True',207132,'B001'),
(200900,'True',330553,'B002'),
(210087,'True',902355,'B002'),
(243431,'True',634817,'B002'),
(245456,'True',207132,'B002'),
(245457,'True',207132,'B002'),
(317411,'True',781132,'B003');
```

(iii) [16 marks] Having answered ((i)-(ii)), you are now required to write and show

your commands for each of the following queries:

(a) [2 marks] Write a query so that the result lists the titles of all DVDs along with their length in descending order of length.

```
SELECT title, Length(title) AS length
FROM D_DVD
ORDER BY Length(title) DESC;
```

Output:

title	length
Mission Impossible III	22
Lord of the Rings III	21
War of the Worlds	17
Casino Royale	13
Harry Potter	12
Shrek 2	7

(b) [2 marks] Write a query so that the result lists the full details of all DVD copies (discs) held in the Portland and New York distribution centers.

```
SELECT * FROM E_DVDCOPY
WHERE dCenterNo IN ('B001','B003');
```

Output:

videoNo	available	catalogNo	dCenterNo
178643	False	634817	B001
199004	True	207132	B001
317411	True	781132	B003

(c) [3 marks] Write a query so that the result lists all distribution centers and their total staff salary costs but only where such costs are greater than 50,000.

```
SELECT  dCenterNo, SUM(salary) as totalSalary
FROM B_Staff
GROUP BY dCenterNo
HAVING SUM(salary) > 50000;
```

ANSWERS

Output:

dCenterNo	totalSalary
B001	76000
B002	82000

(d) [3 marks] Write a query so that the result lists the total number of DVD copies (identified by video number) available for rental in each of the distribution centers, with rows displayed in descending order of copies available.

```
SELECT COUNT(videoNo) AS totalDVD, dCenterNo
FROM E_DVDCOPY
WHERE available = 'True' GROUP BY dCenterNo ORDER BY
COUNT(videoNo) DESC;
```

Output:

totalDVD	dCenterNo
5	B002
1	B003
1	B001

(e) [3 marks] Write a query so that the result lists all of the DVD copies supplied by 20th Century Fox Home Video along with their titles and current availability for rental.

```
SELECT title, videoNo, available
FROM D_DVD, E_DVDCOPY, C_Supplier
WHERE C_Supplier.supplierNo = D_DVD.supplierNo AND
D_DVD.catalogNo = E_DVDCOPY.catalogNo AND name = '20th
Century Fox Home Video';
```

ANSWERS

Output:

title	videoNo	available
War of the Worlds	178643	False
War of the Worlds	243431	True

(f) [1 mark] From the results of your last query, what is the video number of the copy that is available? 243431

(g) [2 marks] Sally Adams (staff number = S0003) has been given a raise of 1000 per year. She has also just got married and her surname has changed to Daniels. Write a query so that these changes are reflected in the database.

```
UPDATE B_Staff
SET name = 'Sally Daniels', salary = salary + 1000
WHERE staffNo= 'S0003';
```

Q4. [Total: 9 Marks]

Below is a dataset (see Figure 2) on students and grades. Write an XML document for this data. You should clearly show the use of internal DTD to define the structure of the XML document.

		Maths	**English**	**IT**	
ID	**Student**	**Test**	**Essay**	**Activity 1**	**Test 2**
1643022	Pace,Camden G.	54%	C+	Pass	32%
1647021	Weber, Lucy X.	32%	C-	Fail	73%
1606023	Branch, Caesar A.	73%	C-	Pass	63%

Figure 2: Students and grades

```
<?xml version='1.0' encoding='utf-8'?>
<!DOCTYPE studentgrade[
<!ELEMENT studentgrade (ID,Student,Maths,English,IT)>
<!ELEMENT ID (#PCDATA)>
<!ELEMENT Student (#PCDATA)>
<!ELEMENT Maths (Test)>
<!ELEMENT Test (#PCDATA)>
<!ELEMENT English (Essay)>
<!ELEMENT Essay (#PCDATA)>
<!ELEMENT IT (Activity1,Test2)>
<!ELEMENT Activity1 (#PCDATA)>
<!ELEMENT Test2 (#PCDATA)>
]>


<studentgrade>
        <Student>
        <ID>1643022></ID>
        <Student>Place,Camden G.</Student>
        <Maths>
                <Test>54%</Test>
        </Maths>

        <English>
                <Essay>C+</Essay>
        </English>

        <IT>
                <Activity1>Pass</Activity1>
                <Test2>32%</Test2>
        </IT>

        </Student>



        <Student>
        <ID>1647021></ID>
        <Student>Weber, Lucy X.</Student>
        <Maths>
                <Test>32%</Test>
        </Maths>
```

ANSWERS

```
<English>
        <Essay>C-</Essay>
</English>

<IT>
        <Activity1>Fail</Activity1>
        <Test2>73%</Test2>
</IT>
</Student>



<Student>
<ID>1606023></ID>
<Student>Branch, caesar A.</Student>
<Maths>
        <Test>73%</Test>
</Maths>

<English>
        <Essay>C-</Essay>
</English>

<IT>
        <Activity1>Pass</Activity1>
        <Test2>63%</Test2>
</IT>
</Student>

</studentgrade>
```

ANSWERS

For Question 5 you need Jupyter notebook for Python 3

Q5. [Total: 11 Marks]

Consider the given dataset (refer to **kmeans.csv**) on VLE (Assessment -- 24_Hour_Exam). Using **Jupyter notebook for Python 3** you are required to use the k-means algorithm to cluster this dataset into 3 groups. Specifically, you need to show:

(i) [7 marks] The coordinates of the point that belong to each cluster.

(ii) [4 marks] Graphically, the points that belong to each cluster, including the means. The means should be of a different shape and colour to those provided for the points.

Include your code in your answer.

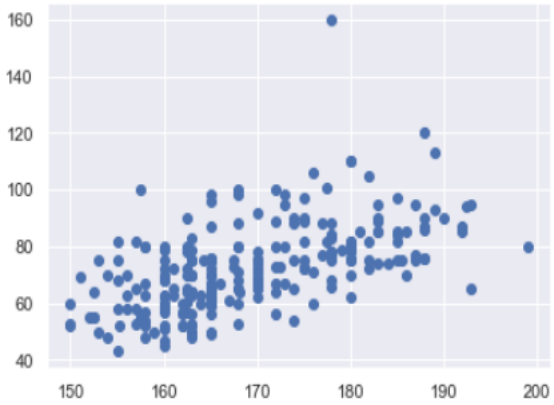
(i)

```
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
import numpy as mp
import pandas as pd
from sklearn.cluster import KMeans
from pandas import DataFrame
```

```
mydata=pd.read_csv('kmeans.csv')
```

```
plt.scatter(mydata['height'],mydata['weight'])
```

```
<matplotlib.collections.PathCollection at 0x1f7091b71d0>
```



In [4]:
```
kmeans=KMeans(n_clusters=3)
kmeans
```

Out[4]:
```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=3, n_init=10, n_jobs=None, precompute_distances='auto',
    random_state=None, tol=0.0001, verbose=0)
```

In [5]:
```
clustering= kmeans.fit_predict(mydata[['height','weight']])
```

In [6]:
```
clustering
```

Out[6]:
```
array([1, 0, 1, 0, 1, 0, 2, 1, 0, 2, 2, 2, 1, 1, 2, 0, 2, 2, 2, 2, 0, 0,
       1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 2, 1, 0, 0,
       1, 0, 0, 1, 0, 2, 2, 0, 0, 0, 0, 0, 1, 0, 2, 2, 1, 2, 0, 0, 1, 0,
       0, 2, 2, 2, 2, 0, 0, 1, 0, 2, 2, 2, 1, 1, 0, 0, 1, 2, 0, 0, 2, 2,
       2, 0, 0, 1, 1, 2, 1, 2, 1, 1, 0, 0, 2, 1, 0, 0, 2, 0, 1, 0, 1, 0,
       1, 1, 1, 0, 1, 0, 1, 2, 0, 2, 2, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0,
       2, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 1, 2,
       1, 0, 1, 2, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 2, 1, 1, 2, 1,
       0, 0, 1, 2, 1, 0, 2, 0, 2, 1, 1, 0, 2, 1, 1, 1, 0, 2, 1, 1, 1, 2,
       2, 1, 1, 2, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 2, 0, 1,
       2, 2, 0, 0, 2, 2, 0, 1, 0, 0, 1, 1, 2, 1, 1, 0, 2, 0, 0, 0, 2, 0,
       1, 1, 0, 1, 0, 1, 0])
```

In [7]:
```
mydata['cluster']=clustering
mydata.head(20)
```

Out[7]:

	id	weight	height	cluster
0	417	63.0	165.0	1
1	423	68.0	168.0	0
2	244	62.0	170.0	1
3	517	75.0	163.0	0
4	428	69.0	151.0	1
5	414	75.0	160.0	0
6	276	100.8	177.5	2
7	57	62.0	165.0	1
8	1	78.0	170.0	0
9	426	93.0	189.0	2
10	519	90.0	190.0	2
11	204	82.0	180.0	2
12	275	53.0	150.0	1
13	24	55.0	152.0	1
14	20	82.0	182.0	2
15	508	74.5	165.0	0
16	2	90.0	188.0	2
17	424	160.0	178.0	2
18	279	88.0	174.0	2
19	427	113.0	189.0	2

(ii)

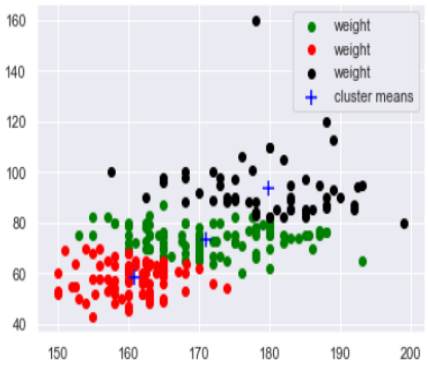
In [8]:
```
a=kmeans.cluster_centers_
a
```

Out[8]:
```
array([[170.93942308,  73.46201923],
       [160.83777778,  58.30777778],
       [179.8       ,  93.45090909]])
```

In [9]:
```
x=mydata[mydata.cluster==0]
y=mydata[mydata.cluster==1]
z=mydata[mydata.cluster==2]
plt.scatter(x.height,x.weight,color='green')
plt.scatter(y.height,y.weight,color='red')
plt.scatter(z.height,z.weight,color='black')


plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color='blue',marker='+',s=100,label='cluster means')
plt.legend()
```

Out[9]: <matplotlib.legend.Legend at 0x1f70969cc88>



ANSWER

For Question 6 you need Jupyter Notebook for Python 3

Q6. [Total: 13 Marks]

You will be using data generated from a survey conducted to explore the prevalence and impact of sleep problems on aspects of people's lives. Refer to the file (**survey_responses.csv**) on VLE (Assessment -- 24_Hour_Exam). Below (refer to Figure 3) is a description of the data source and how the collected data has been coded.

Description of variable	Variable name	Coding instructions
Gender	gender	male; female
Rate quality of sleep	qualslp	1=very poor, 2=poor, 3=fair, 4=good, 5=very good, 6=excellent
Hours sleep/week nights	hourwnit	Hours sleep on average each weeknight
Hours sleep/week ends	hourwend	Hours sleep on average each weekend night

Figure 3: Codebook

With regards to the data, suppose we want to examine the following: "Is there a significant difference in quality of sleep reported by men and women?"

Task

To answer this question:

(i) [6 marks] Show the code **(written in Jupyter notebook for Python 3)** to justify whether a parametric or non-parametric test would be needed.

ANSWERS

```
In [268]: import pandas as pd
          from scipy import stats
          df=pd.read_csv("survey_responses.csv")
          df.groupby("gender")['qualslp'].describe()
```

Out[268]:

gender	count	mean	std	min	25%	50%	75%	max
female	13.0	3.615385	1.502135	2.0	2.0	4.0	5.0	6.0
male	13.0	3.769231	1.235168	1.0	4.0	4.0	4.0	6.0

```
In [269]: df.head()
```

Out[269]:

	id	gender	hourwnit	hourwend	qualslp
0	417	female	6.0	6.0	5.0
1	423	female	7.0	8.0	2.0
2	244	male	6.0	6.0	3.0
3	517	female	4.0	4.0	2.0
4	428	female	4.5	4.5	4.0

```
In [270]: male = df[(df['gender']=='male')]
          female = df[(df['gender']=='female')]
```

```
In [274]: stats.shapiro(male['qualslp'].dropna())
Out[274]: (0.8384102582931519, 0.020229365676641464)

In [275]: stats.shapiro(female['qualslp'].dropna())
Out[275]: (0.8370558023452759, 0.01944321021437645)
```

A non-parametric test would be needed.

Justification: Using shapiro-Wilk test, for each variable, the p value is < 0.05, meaning that they do not follow a normal distribution.

(ii) [3 marks] Having answered (i), which statistical test will you apply? Justify your answer.

We will apply Mann-Whitney U Test. This is because we are examining the difference between an independent variable (two groups – male and female)

on a dependent variable (quality of sleep). % 2 marks

Another statistical test would be: Kruskal-Wallis. Unlike Mann-Whitney U, with Kruskal-Wallis more than two groups can be accommodated. % 1 mark

(iii) [4 marks] From (ii), apply the appropriate statistical test and clearly explain your result. You need to show the code **(written in Jupyter notebook for Python 3)** that you have used to reach your answer.

Using Mann-Whitney U % 2 marks

```
stats.mannwhitneyu(male['qualslp'].dropna(),female['qualslp'].dropna(),alternative='two-sided')
```

```
Output

MannwhitneyuResult(statistic=86.5, pvalue=0.9362216404545397)
```

Since the p value of 0.936 is > 0.05, we can reveal that there is no significant difference between male and female on quality of sleep.%2 marks

Using Kruskal-Wallis %2 marks

```
stats.kruskal(male['qualslp'].dropna(),female['qualslp'].dropna(),alternative='two-sided')
```

```
Output

KruskalResult(statistic=0.011383363214675927, pvalue=0.915032658918408)
```

Since the p value of 0.915 is > 0.05, we can reveal that there is no significant difference between male and female on quality of sleep. %2 marks

For Question 7 you need Jupyter Notebook for Python 3

Q7. [Total: 17 Marks]

Consider the training dataset (refer to **training_data.csv** on VLE) (Assessment -- 24_Hour_Exam) which describes a set of objects using eight attributes, A1-A8. The dataset also lists whether each object is a '0' (as negative) or '1' (as positive) example of a certain, unnamed concept (see column 'Output_Class_Label' in the csv file). We want to build a logistic regression model in Python to try to predict this class.

ANSWERS

Task

Using Jupiter Notebook for Python 3:

(i) [1 mark] Show the code to read the file 'training_data'.

```
import numpy as np
import pandas as pd

df=pd.read_csv("training_data.csv")
```

(ii) [1 mark] We do not want the ID column in our analysis. Show the code to drop this column.

```
df.drop(['ID'],axis=1,inplace=True)
df.head()
```

(iii) [5 marks] Show the code to clean the dataset and run the logistic regression. You should use 80% of the data for training and 20% for testing.

```
df.dropna(inplace=True)
df.isnull().sum() %not required. Some students will, some will not.



var_y=df['Output_Class_Label']

var_x=df[['A1','A2','A3','A4','A5','A6','A7','A8']]

*********************************************************************************
from sklearn.model_selection import train_test_split

var_xtrain,var_xtest,var_ytrain,var_ytest=train_test_split(var_x,var_y,test_size
=0.20)

*********************************************************************************
from sklearn.linear_model import LogisticRegression

lg_model = LogisticRegression()
```

ANSWERS

****

```
lg_model.fit(var_xtrain,var_ytrain)

y_prediction=lg_model.predict(var_xtest)
```

(iv) [2 marks] Show the magnitudes of the coefficients. You must show your code. Which attribute corresponds to which weight must be clearly shown when running the code?

```
coeff=pd.Series(lg_model.coef_[0],index=var_x.columns.values)
print(coeff)
```

Output:

```
A1    0.121293
A2    0.028932
A3   -0.016495
A4   -0.003196
A5    0.000365
A6    0.055168
A7    0.748815
A8    0.003518
```

(v) [1 mark] Which attribute is impacting the result the most?

A7

(vi) [2 marks] Show the code to display the confusion matrix . The confusion matrix should clearly show the following labels: 0 and 1, Actual (or True), and Predicted.

****

Code for displaying the confusion matrix *without any labels:*

```
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(var_ytest,y_prediction)
```

Output:

```
array([[85, 12],
       [27, 27]],
```

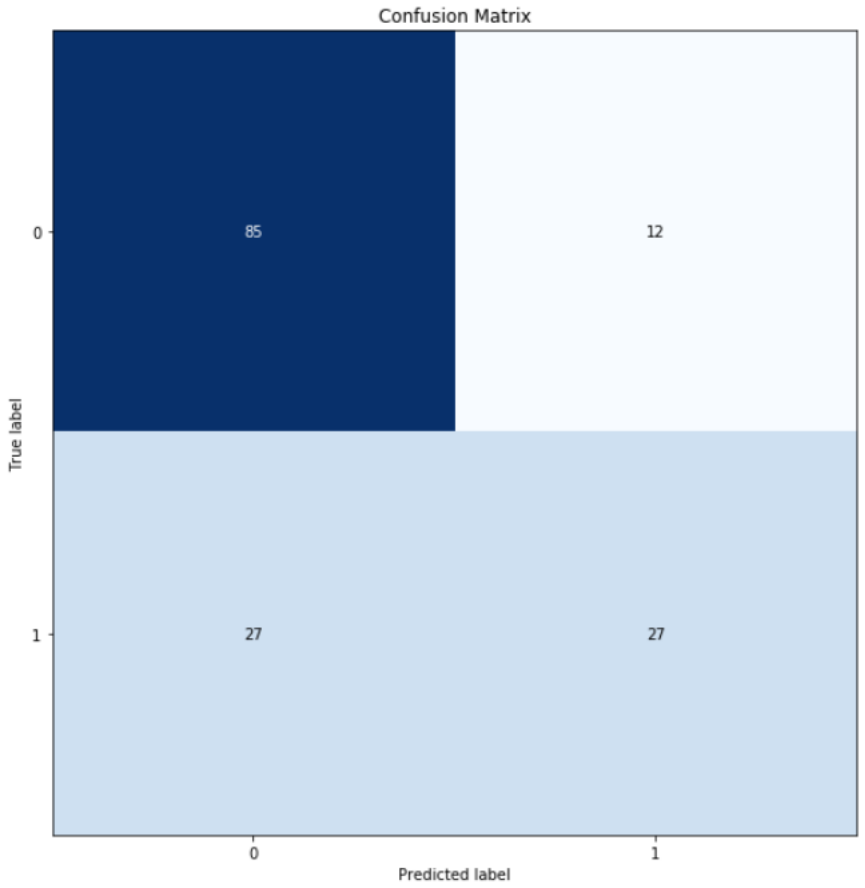
****

Code for displaying the confusion matrix and with prescribed labels:

```
! pip install -q scikit-plot

import scikitplot as skplt
skplt.metrics.plot_confusion_matrix(var_ytest,y_prediction,figsize=(12,12))
```

Output:



Note: Another word for 'True' will be 'Actual'

Note: The values in the confusion matrix will differ

(vii) [2 marks] Interpret the confusion matrix.

Interpretation.

1) We had 97 objects which were from class 0 (negative). The model has correctly predicted 85 objects as 0, but incorrectly classified 12 cases as 1 (positive). **%1 mark**

2) We had 54 objects which from class 1 (positive). The model has correctly predicted 27 cases as 1, but incorrectly classified 27 objects as 0 (negative). **%1 mark**

(viii) [3 marks] From (vi) calculate the Sensitivity, Positive Predicted Value, and F1 score. Show your workings including the formulas.

****

Sensitivity = TP/(TP + FN) = 27/(27+27) = 0.5

****

Positive Predicted Value (Precision) = TP/(TP + FP) = 27/(27+12)= 0.7

****

F1 score = (2*Precision*Recall)/(Precision + Recall)

F1 score = 2* 0.7*0.5/0.7+0.5 = 0.7/1.2 = 0.6

Note: Recall is the same as sensitivity (calculated earlier)

Note: For F1 score we combine the results of the precision and sensitivity into a single metric, by taking their harmonic mean

****

END OF PAPER

ANSWERS