The Hong Kong Polytechnic University
Department of Computing
COMP 5121 Data Mining and Data Warehousing
Semester 1 2011-12
Assignment 3
Due Date: November 22, 2011

Part A (Individual)

Question 1

The *Consumer Magazine* measured the gas mileage of 15 model cars. Each car is characterized by seven attributes which are: (a) Car name (Make and model); (b) weight; (c) Miles per gallon, a measure of gas mileage; (d) Drive ratio of the automobile; (e) Horsepower; (f) Displacement of the car (in cubic inches); (g) Number of cylinders. The data set is given below:

(a)	(b)	(c)	(d)	(e)	(f)	(g)
Buick Estate Wagon	16.9	4.36	2.73	155	350	8
Ford Country Squire Wagon	15.5	4.054	2.26	142	351	8
Chrysler LeBaron Wagon	18.5	3.94	2.45	150	360	8
Chevette	30	2.155	3.7	68	98	4
Dodge Omni	30.9	2.23	3.37	75	105	4
Buick Century Special	20.6	3.38	2.73	105	231	6
Mercury Zephyr	20.8	3.07	3.08	85	200	6
AMC Concord D/L	18.1	3.41	2.73	120	258	6
Mercury Grand Marquis	16.5	3.955	2.26	138	351	8
Dodge Colt	35.1	1.915	2.97	80	98	4
AMC Spirit	27.4	2.67	3.08	80	121	4
Honda Accord LX	29.5	2.135	3.05	68	98	4
Buick Skylark	28.4	2.67	2.53	90	151	4
Olds Omega	26.8	2.7	2.84	115	173	6
Plymouth Horizon	34.2	2.2	3.37	70	105	4

a) Find a clustering arrangement of records using the *k*-means algorithm by setting *k*=2 and using the **first two records** as the initial cluster centers.
b) Repeat a) by setting *k*=3 and using the **last three records** as the initial cluster centers.
c) Instead of clustering the data manually as in a) and b), use PASW's *K*-Means algorithm to get models of 2-cluster and 3-cluster.
d) For the four different clustering arrangements obtained from a) to c) above, which one do you think gives the best results? Why? Is there any way that the results can be further improved? If so, how can they be proved?
e) Repeat a) and b) using the Hierarchical Agglomerative Single-Linkage clustering algorithm. In your submission, please show the first distance matrix, the partitions produced at each stage and the final dendrogram. In what ways are your results different from that when *k*-means is used?

Part B (Groups of 2 or 3)

Question 1

For the last two years, ABC Telecom has not been performing up to the expectation of their investors. The management of ABC Telecom would like to find ways to improve their competitiveness in a very mature market in Hong Kong. They decided that if they could understand the data they collected better, they may be able to become more profitable. You are hired as their data mining consultant and are given access to a number of databases, including: (i) a database containing a number of call-detailed records which are generated whenever a phone call is made, (ii) a database containing the customers' demographic data and (iii) databases containing information about payment and marketing. Using the data mining techniques you know, mine these databases for patterns that may have ABC Telecom to see if there are any interesting patterns that can help the management of ABC Telecom better make business decisions.

You are to submit a report to the management of ABC Telecom detailing every step you take to mine the databases. You are to explain why you do what you do and highlight interesting patterns discovered in the databases that may help them better serve their customers, increase their revenue and reduce costs.

Data Characteristics

The databases include the followings: PhoneCallDetails, CDemographics, PaymentRecord and MarketingData. PhoneCallDetails contains data collected when every phone call is made. CDemographics contains demographic information of each customer. PaymentRecord contains payment details like payment method and date of settlement and MarketingData contains information about special sales package offered to selected customers. The details of these databases are given below.

PhoneCallDetails (No. of records:)

	Attribute Name	Data Type	Description
1	TID	int	Transaction identifier
2	Origin	int	Origin phone number
3	Destination	int	Destination phone number
4	StartTime	Timestamp	Time of making the call
5	EndTime	Timestamp	Time of ending the call
6	Cell	char	The region of making the phone call
7	Cell2	char	The region of ending the phone call
8	Disconnect	char	With disconnection -Y; No disconnection - N

CDemographics (No. of records: 256)

	Attribute Name	Data Type	Description
1	CustomerID	int	Customer identifier
2	Sex	char	M/F
3	DateOfBirth	Date	Date of Birth
4	EducationLevel	char	Education Level. "P": Primary, "S": Secondary, "T": Tertiary, "D": Degree, "A": Degree or above
5	PhoneNo	int	Phone number
6	Joining Date	Date	Date of joining ABC Telecom
7	JoinFrom	String	New or join from another telecom company

PaymentRecord (No. of records: 2180)

	Attribute Name	Data Type	Description
1	PID	int	Payment identifier
2	CustomerID	int	Customer identifier
3	Method	int	Payment method (1- credit card, 2-PPS, 3-Cash, 4-online)
4	Charge	double	Current charges
5	Balance	double	Current balance (current + previous charges)
6	PaymentDate	Date	Date of payment
7	DueDate	Date	Payment Due Date

MarketingData (No. of records: 645)

	Attribute Name	Data Type	Description
1	CustomerID	int	Transaction identifier
2	SpecialOffer	char	Special offer (A - free MTR, Tunnel & License Fee, B – Additional min. , C – Discount purchase of phone)

Data File

The tables are stored in an Access file which contains the dataset described above. See attached file, *telecomData.mdb*.

What to do

You are to use the modeling techniques you learned in this course to mine the dataset. You are to submit a report as if you are submitting to the management of ABC Telecom.

The report should detail every step of the process you took to mine the data and what you have discovered (Hint: KDD process, possible preprocessing steps, different goal**s** and modeling techniques). You are to convince the management the effectiveness of data mining, both objectively and subjectively.

In the report, you need define your objective in each mining process. You may need to use various functions in the tools or write your own programs or modules to help you prepare your mining data into suitable format. You may need to read the user manual to discover further functions.

Please submit the report and state, at the end of it, the role each team member played in the assignment.

Question 2

The credit card transaction data in this assignment are stored in five tables: CATEGORY, CUSTOMERS, ORDERDETAILS, ORDERS, and PRODUCTS. You are required to import the data from attached zipped file (*dw_work.zip*) into a data warehouse for further analysis.

You are preparing a sales report that shows the sales figures of shops in different countries for the last few years. It may be: "Daily Sales", "Monthly Sales", "Quarterly Sales" or "Yearly Sales" of shops in "London", "UK", for example or other places. In the meantime, the sales of each product or product category can be displayed if required.

a) Sketch a data warehouse schema that can implement the above requirements. Explain how the design can fulfill the requirements and what the advantages are.
b) Use Oracle Warehouse Builder to construct a data warehouse which you have just designed.
c) Generate a pie chart of "Quarterly Sales" of a product in Germany for the year 1997.
d) Show how you can use the application to retrieve other sales figures (monthly or quarterly or individual product) in different cities/countries.

Category (No. of records: 8)

	Attribute Name	Data Type	Description
1	CategoryID	int	Category identifier
2	CategoryName	String	Name of category
3	Description	String	Category description

Customers (No. of records: 91)

	Attribute Name	Data Type	Description
1	CustomerID	String	Customer identifier
2	CompanyName	String	Name of company
3	ContactName	String	Contact person
4	ContactTitle	String	Title of the contact person
5	Address	String	Address
6	City	String	City
7	Region	String	Region
8	PostalCode	String	Postal code of the address
9	Country	String	Country
10	Phone	String	Contact number
11	Fax	String	Fax number

OrderDetails (No. of records: 2,155)

	Attribute Name	Data Type	Description
1	OrderID	int	Order identifier
2	ProductID	int	Product identifier
3	UnitPrice	double	Price of each product
4	Quantity	int	Order quantity
5	Discount	double	Discount rate

Orders (No. of records: 830)

	Attribute Name	Data Type	Description
1	OrderID	int	Order identifier
2	CustomerID	String	Customer identifier
3	EmployerID	int	Employer identifier
4	OrderDate	Date	Date of placing order
5	RequiredDate	Date	Date of requiring the order
6	ShippedDate	Date	Date of shipping
7	ShipVia	int	Shipping method

8	Freight	double	Price of freight
9	ShipName	String	Ship name
10	ShipAddress	String	Shipping address
11	ShipCity	String	Shipping city
12	ShipRegion	String	Shipping region
13	ShipPostalCode	String	Shipping postal code
14	ShipCountry	String	Shipping country

Products (No. of records: 77)

	Attribute Name	Data Type	Description
1	ProductID	int	Product identifier
2	ProductName	String	Product name
3	SupplierID	int	Supplier identifier
4	CategoryID	int	Category identifier
5	QuantityPerUnit	String	Quantity per unit
6	UnitPrice	double	Price of each unit
7	UnitsInStock	int	Remaining stock units
8	UnitsOnOrder	int	Ordering units
9	RecorderLevel	int	Level of reorder
10	Discontinued	boolean	Status of discontinued

**Make assumptions where appropriate.*