Université d'Ottawa
Faculté de génie

École de science informatique
et de génie électrique

uOttawa
L'Université canadienne
Canada's university

University of Ottawa
Faculty of Engineering

School of Electrical Engineering
and Computer Science

CSI4142: Introduction to Data Science

Midterm Winter 2017

Name	
Student number	

Instructions:

1. No aid allowed, **except for one letter-size page** (a so-called cheat sheet), printed or written on both sides.
2. Calculators are allowed.
3. This exam is out of 40 marks and contains 6 pages.
4. Answer all questions on this exam paper, using the space provided in the answer sheets.
5. Good luck!!!

PART A: Data summarization and Data Preprocessing (20 marks)

Answer all of the following questions, by indicating your answers on the answer sheet on page 3.

Consider the following table that you obtained from the *Bumpy Used Car Dealership*. (Some Dealers prefer to use the term "Previously Owned".) This table contains the Make of the automobiles, the Year of manufacture, the Size and Number of Doors, together with the Sales Price of all the Cars on sale.

Make	Size	Doors	Year	Price ($)
BMW	Compact	4-door	2008	34,000
Honda	Intermediate	4-door	2011	21,000
Fiat	Compact	2-door	2015	24,000
	Intermediate	4-door	1999	8,000
Toyota	Full-size	4-door	2001	9,000
	Full-size		2010	21,000
BMW	Intermediate	4-door	2004	25,000
Smart	Compact	2-door	2013	15,000
Fiat	Compact	2-door	2016	28,000
BMW	Intermediate	4-door	2010	29,000
BMW	Compact	4-door	2010	31,000
			1994	4,000

1. Explain what ordinal data are and give an example of ordinal data contained in the table. (4 marks)

2. Show the steps you would follow to determine whether the distribution of the prices is symmetric or skewed. (6 marks)

3. Draw the boxplot for the Year attribute. (4 marks)

4. The above table contains some missing values, notably for the Make attribute. Explain the approach that would be the best to use, when handling such missing values for a Car Dealer, and be sure to motivate your answer. (4 marks)

5. You are asked to determine whether the Price and Year attributes are correlated. Name the statistical test you would you use to accomplish this task. (2 marks)

Part A: Answer sheet

Please enter your answers in the table below.

Question	Answer
Question 1	
Question 2	
Question 3	
Question 4	
Question 5	

PART B: Data Marts (20 marks)

Answer all of the following questions, by indicating your answers on the answer sheets.

The *Bumpy Used Car Dealership* currently records the data about their daily car sales transactions in a PostgreSQL relational database. This database contains data about the Customers, such as their names, social security numbers, address, and so on. It also includes the data from the sales Invoices, such as the date of the car sale as well as the sales amount. In addition, the database records some detailed information about the Car, such as the make, color, year, model, and so on. Data about the car Salesperson, i.e. the person who sold a specific car, are also recorded.

Suppose that the *Bumpy Used Car Dealership* decides to create a data mart to keep track of the sales of cars over the last ten years. They are interested in exploring daily sales data, as well as studying monthly and seasonal trends. Further, they aim to use this data mart in order to identify trends in sales, such as which makes are popular, when sales are at a peak, what color people prefer, etc.

You are hired to implement this data mart.

Answer the following questions.

1. Explain what the grain of a data mart is and declare the grain of the data mart you will design. (4 marks)

2. Explain what a fact/measure is and identify one fact/measure that you will include in your data mart. (4 marks)

3. Identify and show the attributes of one dimension, *other than* Date or Customer, that you will include in your data mart. (2 marks)

4. Explain the benefit of aggregates and provide one aggregate that would aid *Bumpy Used Car Dealership* in their data analytics. This choice should be based on the above-mentioned "potential queries". (4 marks)

5. Explain, (a) what snowflaking is, and, (b) why it should be avoided. (2 marks)

(Note: Question 6 and 7 continue on page 6.)

Part B: Answer sheet (1 of 2)

Please enter your answers in the table below.

Question	Answer
Question 1	

Question 2	
Question 3	
Question 4	
Question 5	

PART B: Continued

6. Explain, (a) what a surrogate key is, and, (b) why it should be used in all data marts. (2 marks)

7. Suppose that your data mart contains a Customer dimension that includes the Gender attribute. Illustrate, by means of your own example, how you would use a bitmap index to speed up the queries against the Gender attribute. (2 marks)

Part B: Answer sheet (2 of 2)

Please enter your answers in the table below.

Question	Answer
Question 6	
Question 7	

fini