CSI4142 Midterm 2017

Memorandum

Part A: 20 marks

1. Ordinal data values have a meaningful order (ranking) but magnitude between successive values is not known. In this case, we have size ={compact, intermediate, full-size}. (4)

2. Here, you need to calculate the mean, median and mode. If they are the same, then the distribution is symmetric. Otherwise, it is skewed. The mean, median and mode are calculated as follows:

mean	median	mode
20750	22500	21000

   This implies that the distribution is **skewed.** (6)

3. This was only mentioned in class, but it is useful to be able to do (2018).
   The values are as follows: (4)

Min: 1994, Max: 2016, Q1: The mean of 2001 and 2004, i.e. 2002.5, Q2 (Median): 2010, Q3: The mean of 2011 and 2013, i.e. 2012

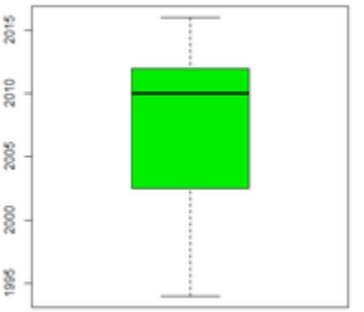

Fig. 1 Boxplot for Year

4. Missing values such as these provide us with many “headaches” when we wish to conduct analytics. The first reason is that the results of aggregate functions such as `count()`, `sum()` and so on are now uncertain. Second, for machine learning, there is the possibility to generalize over missing values. However, if the number is large, then a machine learning algorithm will have difficulty to construct accurate models.

   Let’s assume that the database is larger than these sample values. One will notice that two of the cars with missing values are older. This may imply that there was a data capturing error and that these car were actually already sold. For the third car, this may be a new car, to be sold, that has not been captured in the sales system.

   The first step would be to go to the source systems and explore them in order to trace the history of these cars. In practice, a car dealer cannot sell a car without the necessary documentation. If these values cannot be located, then the rows could either be “flagged” or removed from the data mart. This choice would be the decision of the domain expert.

   In this specific case, the most common sense thing would actually be to go and look at the three cars in the parking lot, in order to determine the Make and other details. This could be easily done through inspection, especially since the cars are i) from 1994 (very old), ii) from 1999 and intermediate, and iii) full-sized and quite expensive. (4)

5. ~~Pearson’s product moment coefficient was covered in class~~ (not yet in 2018). Students may also mention any other test that finds the correlation between numerical attributes. (2)

Part B: 20 marks

1. The grain of the data mart refers to the level of detail that we store data. Specifically, it refers to the level of details of a single row in the Fact table. In this case, the grain is a Sale of a single Car to a Customer on a particular Day, as sold by a Salesperson. (4)

2. A fact/measure is an additive (and often numeric) value or measurement that is stored in the Fact table. It is mainly used for aggregation and subsequent analytics. In this case study, the "dollar-amount" is an example of a fact/measure. Others could be the "cost-price" or the "profit-made". (4)

3. In this case, examples are the CAR(Car-key, Make, Model, Color, Year, KmReading, ManufacturingPlant, ManufacturingCity, …) or SALES-PERSON(Person-key, LastName, FirstName, Date-Employed, …) dimensions. (2)

4. Aggregates (also know as Cubes) are mainly used in order to improve query performance. They are pre-calculated and often implemented as materialized views from queries. Specifically, they involve SUM() and GROUP BY operations.
   In this case, any one of the following aggregates would potentially speed up performance (based on the frequent queries):
   a. Monthly Sales
   b. Seasonal Sales
   c. Sales per Car Maker (e.g. BMW versus FIAT) (4)

5. Snowflaking means that we normalize a dimension into two tables, and then link them using foreign keys. For example, we could split the CAR dimension and store details about the MANUFACTURER separately. It is not such a great idea, because it involves the joining of tables and (as shown in class) this is a very expensive database operation. The one place where we would use snowflaking is to avoid "monster" (i.e. very huge) dimensions. (2)

6. A surrogate key is an auto-number that is used as a primary key in the data mart. The reason for using this is that we wish to avoid using production keys that has a "meaning" in the operational system. Production keys may be changed by the source system and also even be reused. If this happen, it may cause inconsistencies in our data mart. For example, consider we have a Meet-number that is re-used. This may lead to data from two totally separate Meets being considered as one. Clearly, this will lead to inconsistencies in the data. (2)

7. The bitmap index is created during data staging and would look something like this:

Record-id	1 2 3 4 5 6 7 8 9 10 …
Male	1 1 1 0 0 0 1 0 1 0
Female	0 0 0 1 1 0 0 1 0 0
Undisclosed	0 0 0 0 0 1 0 0 0 1

In this case, we know that we only have to consider the records with "1"s in our query. Suppose we have a query that calculates the total dollar-amount (`SUM()`) of cars that were purchased by Customer where *Gender = "Undisclosed"*.

In this toy example, we know that we only have to consider records the CustId6 and CustId10. These records are selected and a semi-join is performed to link it to the FACT TABLE. Here, the FACT TABLE rows with CustId6 and CustId10 are retrieved and the query proceeds to calculate the `SUM()`.

This can speed up our queries, especially if we use hardware to calculate the AND operations.