DTI 5126: Fundamentals for Applied Data Science

Summer 2021

Assignment 1

Submission Deadline: 31st May 2021 on Brightspace.

This assignment should be completed individually using R. The assignment is in two parts: Data Warehousing & Data Preparation. Upon completion, present your result in one submission, including the answers generated or plots. Where applicable, submit the source codes used to generate your results as a separate attachment.

Part A: Data Warehousing & OLAP (50 points)

An on-line seller of Pizza wishes to maintain data about orders. Customers can order their Pizzas with a selected size (personal, small, medium large, xlarge), a selected type of the dough (e.g., whole wheat thin, white regular, stuffed crust), a selected type of cheese (e.g., Swiss, cheddar, Mozzarella) and any one topping (e.g., tomatoes, pepper, onions, pepperoni). The fact table for such a database might be:

Orders (StoreLocation, date, PizzaSize, Dough, CheeseType, Topping, Quantity, Profit)

StoreLocation reflects which store has performed the service and points to a dimension table, and PizzaSize, Dough, CheeseType and Topping are also pointing to other dimension tables. For example, the StoreLocation might be pointing to a dimension table about location (address, city, province, Country, region). The Quantity attribute is the number of pizzas ordered of that type of pizza, and Profit is the profit made for that type of Pizza.

Deliverables:

1.
    a. Sketch a star schema that represents this problem.
    b. Sketch a snowflake schema that represents this problem.
    c. Generate a set of sample data stored in csv files for the dimensions and fact table for the snowflake schema in c.
2. Using R, read the dimensions files and the profit fact table. Build an OLAP cube for your revenue and show the cells of a subset of the cells
3. Suppose that we want to examine the data of the above store to find trends and thus to predict which Pizza components the store should order more of. Describe a series of drill-down and roll-up operations that would lead to the conclusion that customers are beginning to prefer bigger pizzas.

Part B: Data Preparation (50 points)

Using the *bank-additionalfull.csv* dataset from the UCI Machine Learning Repository given. The data relates to a phone-based direct marketing campaign conducted by a bank in Portugal. The bank was interested in whether or not the contacts would subscribe to a term deposit account with the bank. Complete the following preprocessing tasks to prepare the data for analysis:

1. Import the data set into RStudio and reduce the dataset to only four predictors (*age, education, previous,* and *pdays*), and the target, *response*.
2. The field *pdays* is a count of the number of days since the client was last contacted from a previous campaign. The code *999* in the value represents customers who had not been contacted previously. Change the field value *999* to *"NA"* to represent missing values.
3. Explain why the field *pdays* is essentially useless until you handle the *999* code
4. Create a histogram of the *pdays* variable showing the missing value excluded.
5. Transform the data values of the *education* field into numeric values using the chart in Table 1 below.

Table 1: Numerical values for education

Categorical Value	Numeric Value
illiterate	0
basic.4y	4
basic.6y	6
basic.9y	9
high.school	12
professional.course	12[a]
university.degree	16
unknown	Missing

Note: use *"NA"* where the value is "Missing".

6. Compute the mean, median & mode of the *age* variable. Using a boxplot, give the five-number summary of the data. Plot the quantile information.
7. Some machine learning algorithms perform better when the numeric fields are standardized. Standardize the *age* variable and save it as a new variable, *age_z*.
8. Obtain a listing of all records that are outliers according to the field *age_z*.