**Name : Dina Abdelhady**

**DTI 5126: Fundamentals for Applied Data Science**

**Summer 2021**

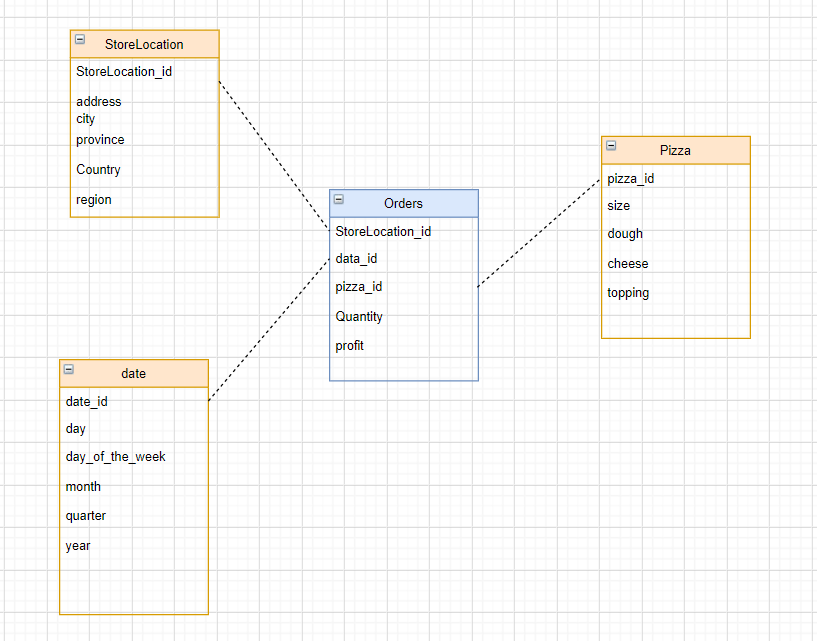
**Assignment 1**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

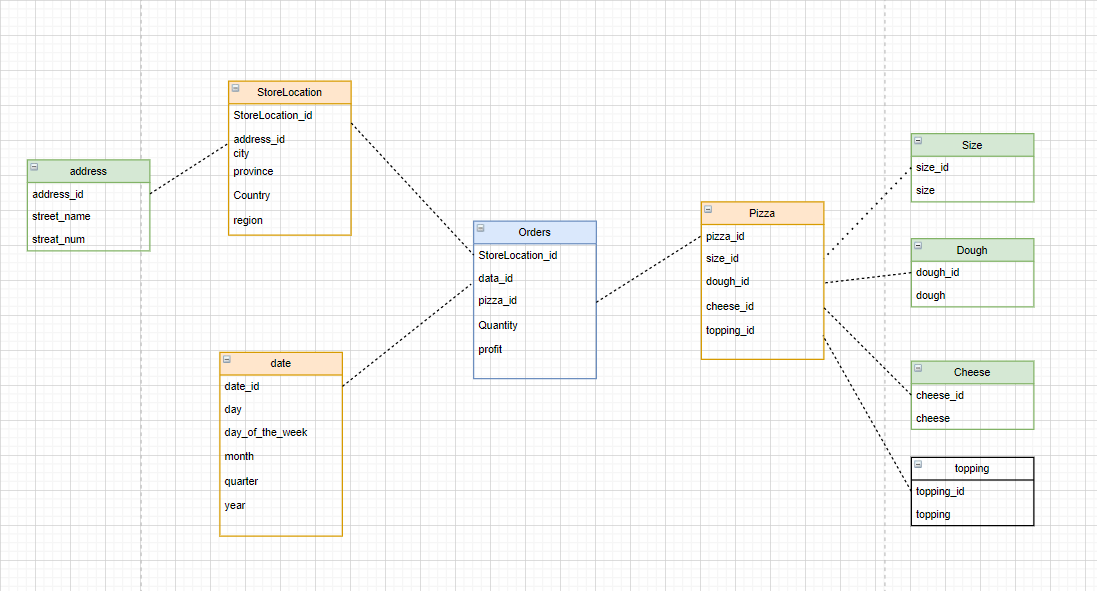
**Part A: Data Warehousing & OLAP**

**1.**

1. Sketch a star schema that represents this problem

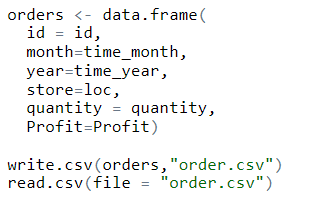


1. Sketch a snowflake schema that represents this problem

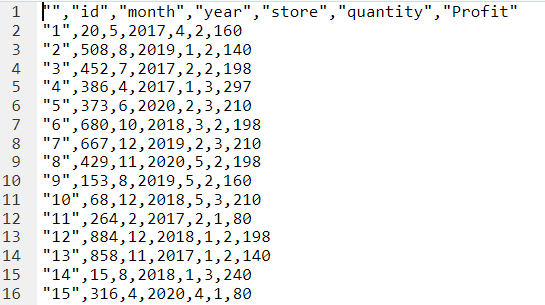


1. Generate a set of sample data stored in csv files for the dimensions and fact table for the snowflake schema in c.
2. Orders

Code :

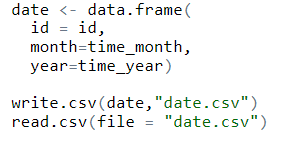


File:

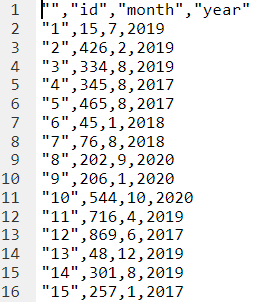


1. Date

Code :

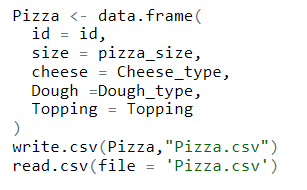


File:

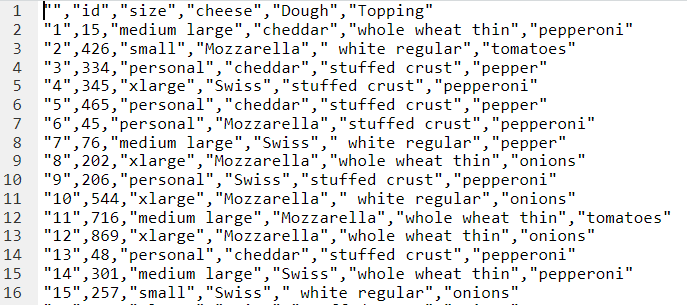


1. Pizza

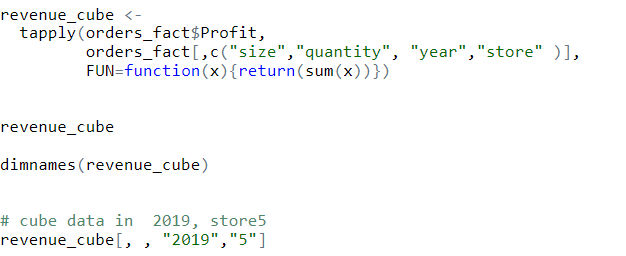
Code:



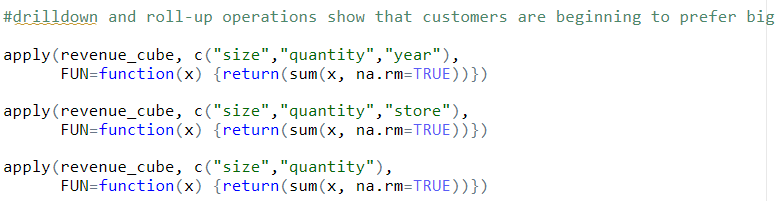
File:



**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

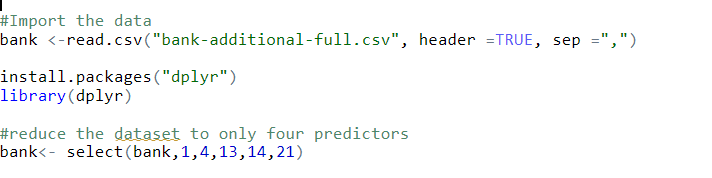


**3.** Suppose that we want to examine the data of the above store to find trends and thus to predict which Pizza components the store should order more of. Describe a series of drilldown and roll-up operations that would lead to the conclusion that customers are beginning to prefer bigger pizzas.



**Part B : Data Preparation**

1. Import the data set into RStudio and reduce the dataset to only four predictors (age, education, previous, and pdays), and the target, response

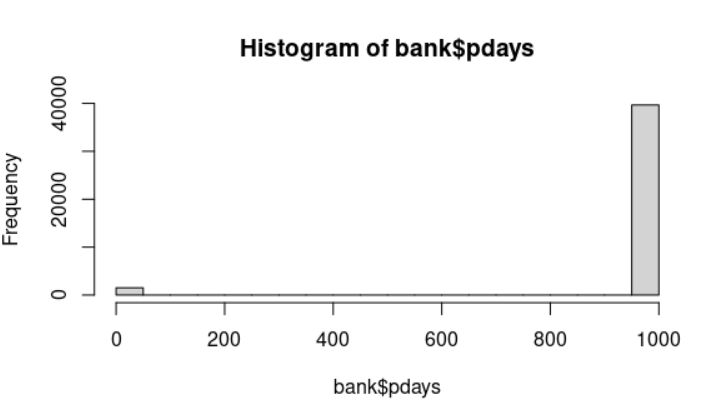


1. The field pdays is a count of the number of days since the client was last contacted from a previous campaign. The code 999 in the value represents customers who had not been contacted previously. Change the field value 999 to “NA” to represent missing values.

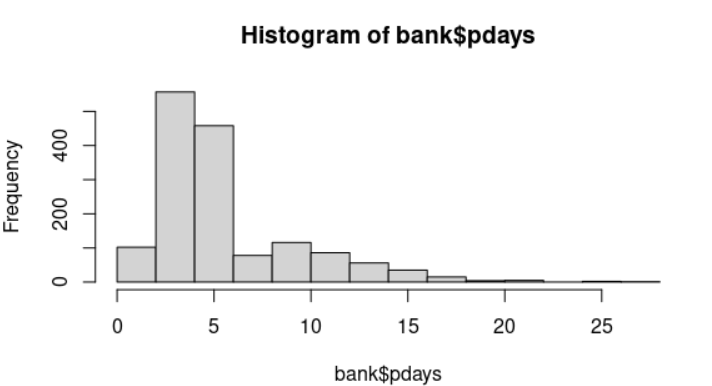


1. Explain why the field pdays is essentially useless until you handle the 999 code

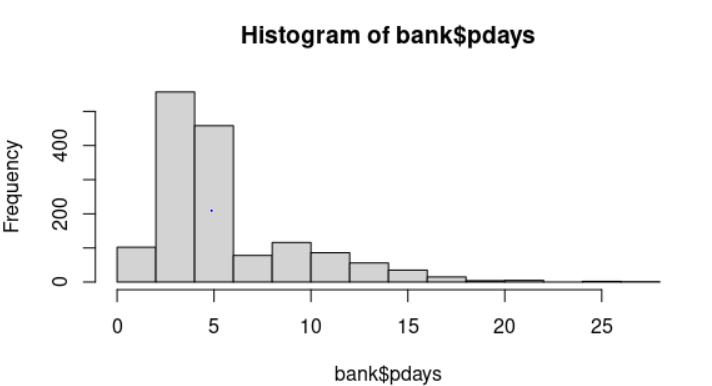
Before :



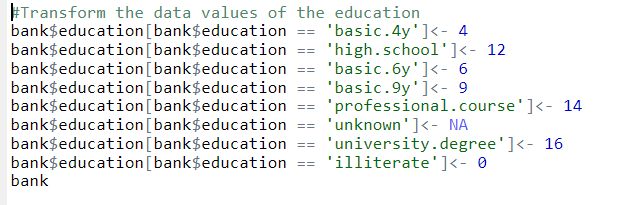
After :



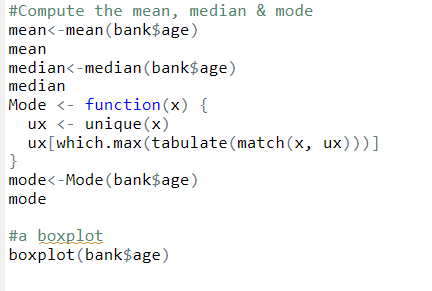
1. Create a histogram of the pdays variable showing the missing value excluded.

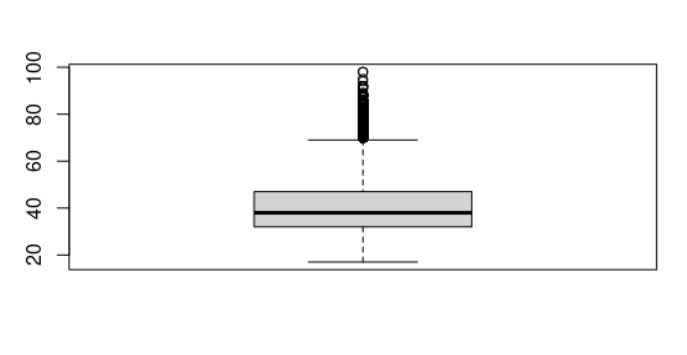


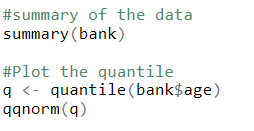
1. Transform the data values of the education field into numeric values using the chart in Table 1 below.

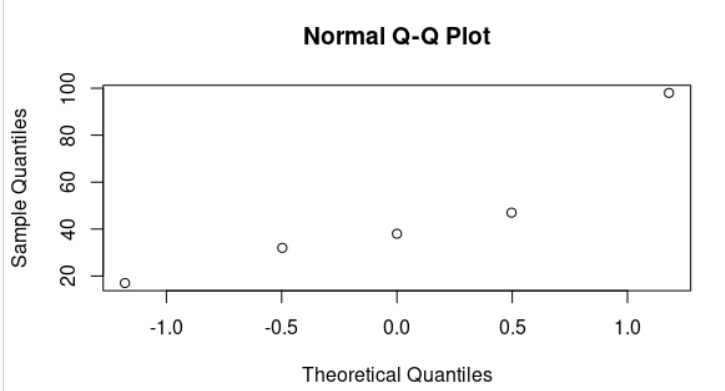


1. Compute the mean, median & mode of the age variable. Using a boxplot, give the fivenumber summary of the data. Plot the quantile information.

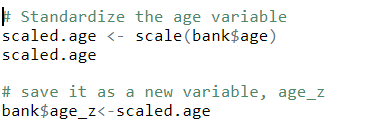








1. Some machine learning algorithms perform better when the numeric fields are standardized. Standardize the age variable and save it as a new variable, age\_z.



1. Obtain a listing of all records that are outliers according to the field age\_z.

