uOttawa

Assignment 1

Made by

Khadija Hesham

E-mail

khesh072@uottawa.com

Supervised by

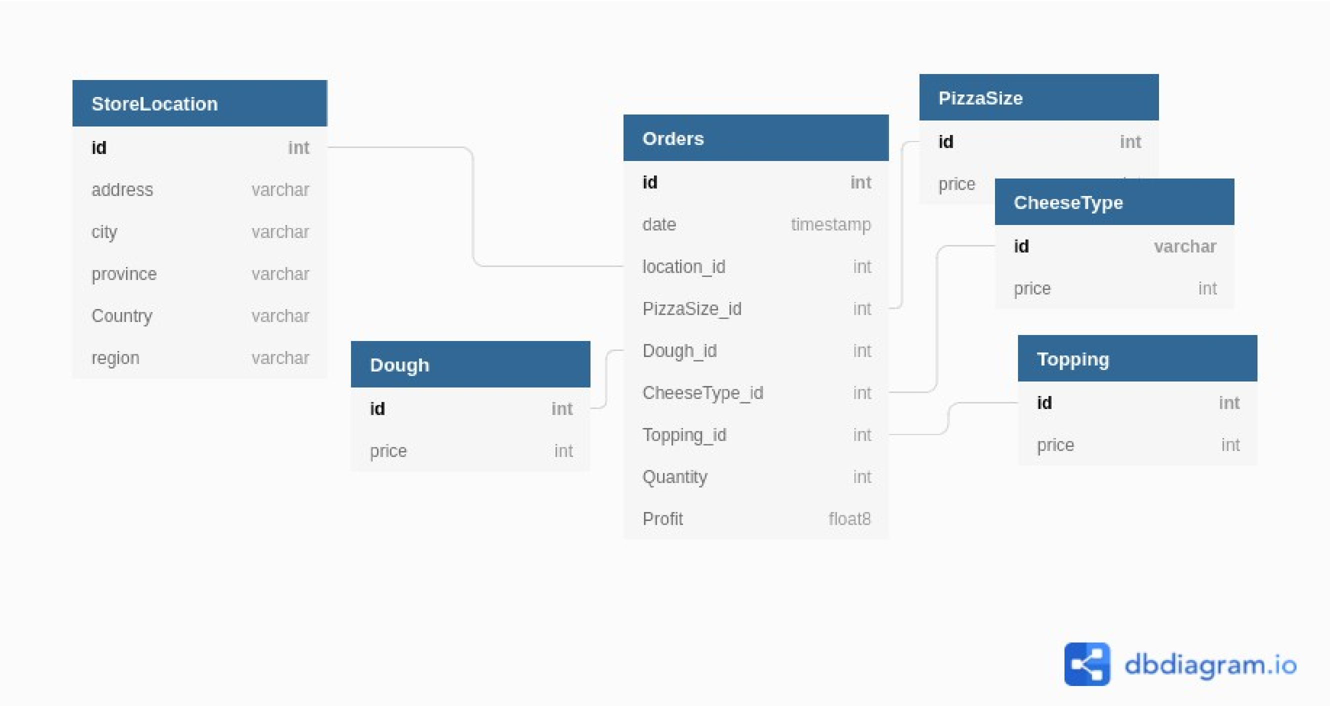
Dr, Bisi Runsewe

Part A: Data Warehousing & OLAP:

<u>Problem overview:</u>

A database containing information for pizza seller.
The information is regarding orders,store location, and pizza components.

<u>The star schema for this problem:</u>



StoreLocation	
id	int
address	varchar
city	varchar
province	varchar
Country	varchar
region	varchar

Orders	
id	int
date	timestamp
location_id	int
PizzaSize_id	int
Dough_id	int
CheeseType_id	int
Topping_id	int
Quantity	int
Profit	float8

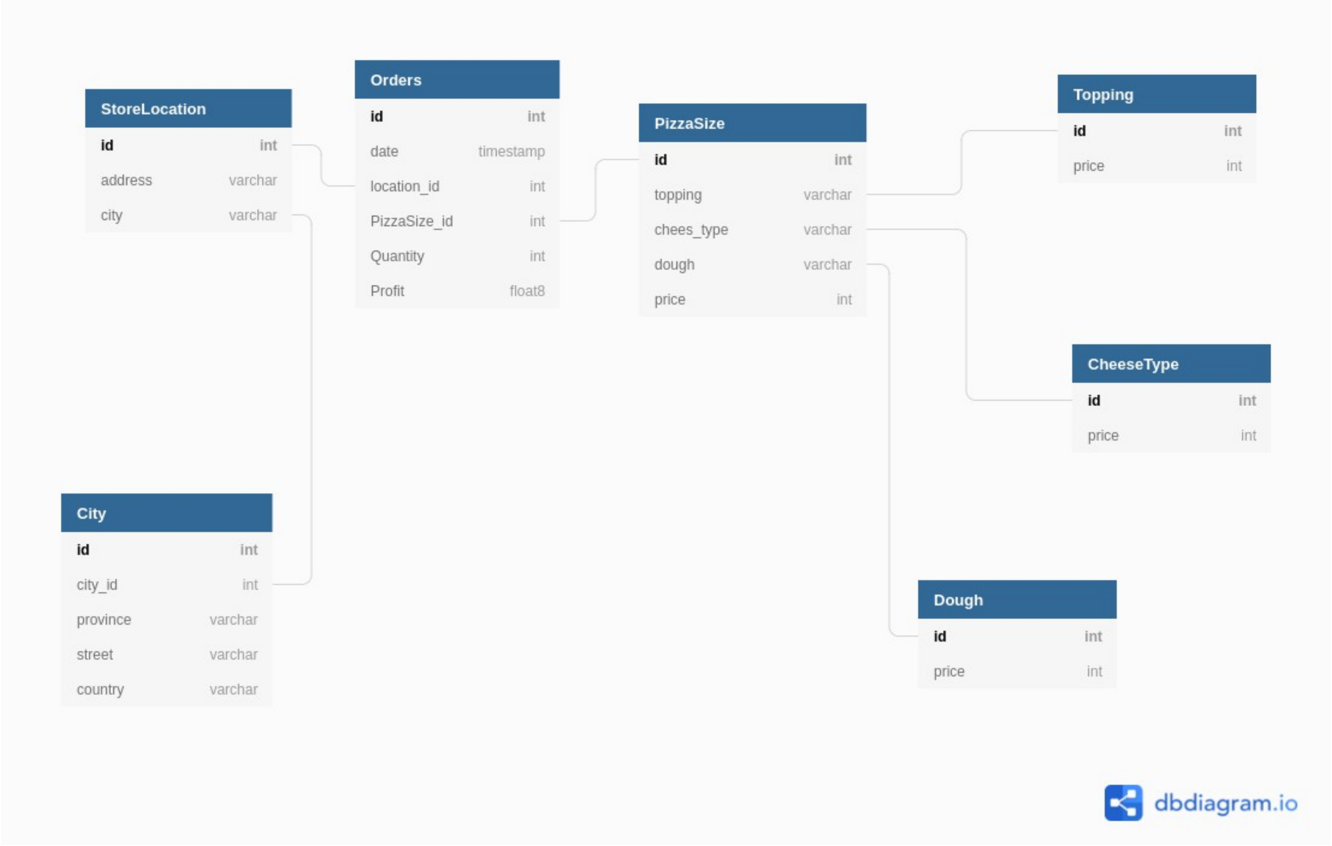
PizzaSize	
id	int
price	

CheeseType	
id	varchar
price	int

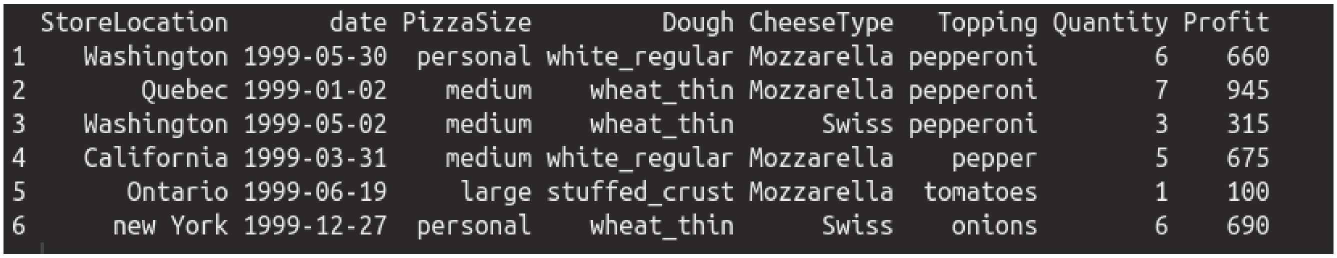
Dough	
id	int
price	int

Topping	
id	int
price	int

The snowflake schema for this problem:

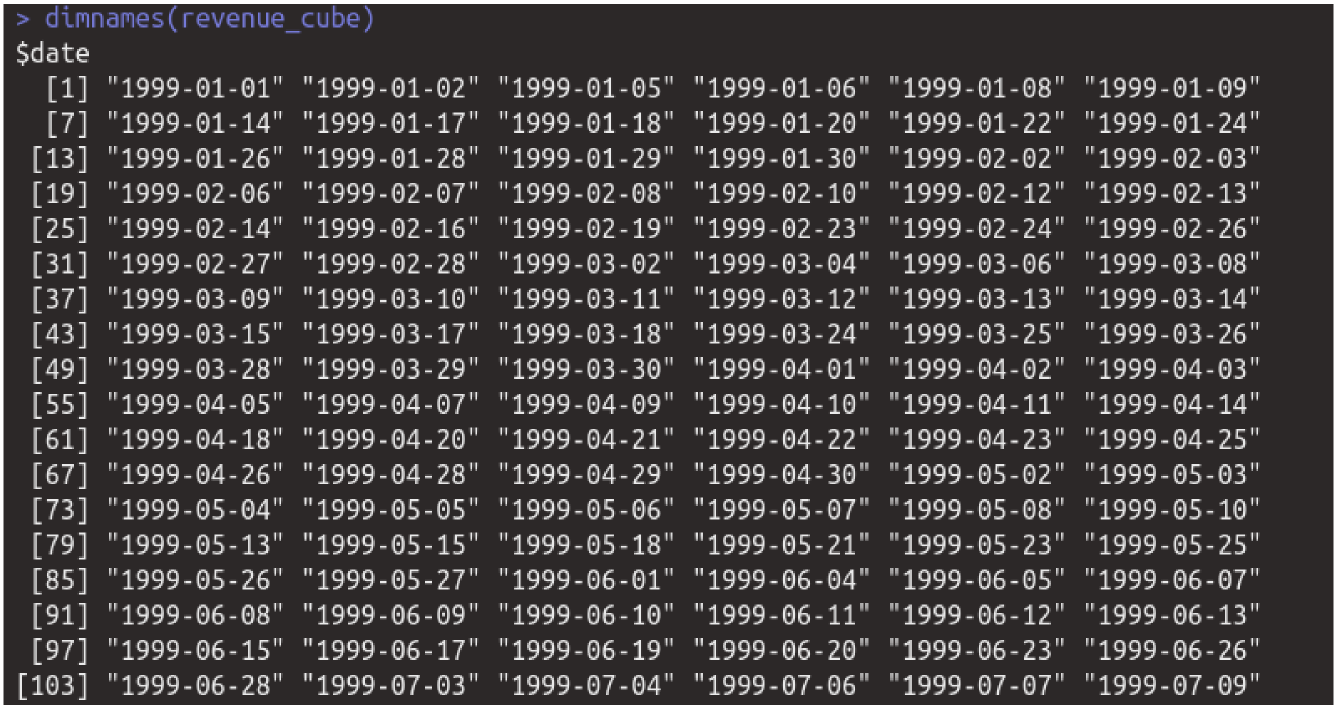


Sample data generation using R:



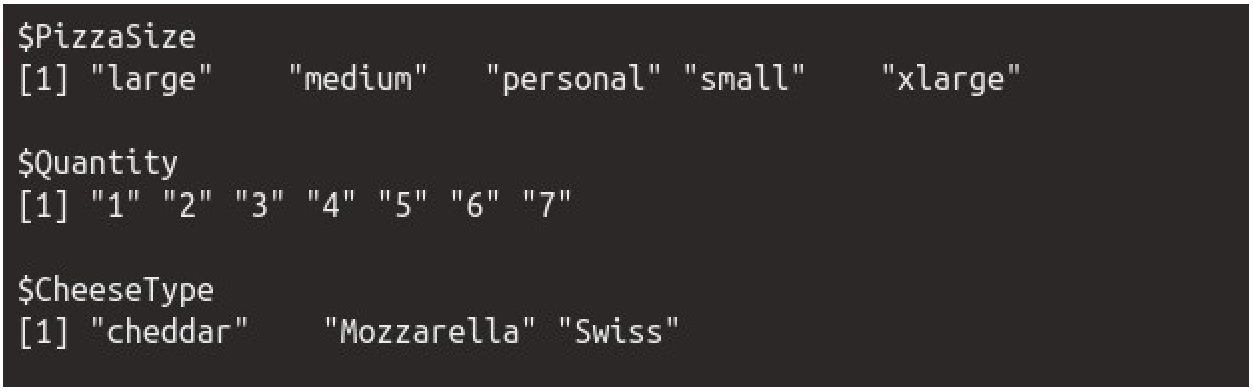
```
  StoreLocation       date PizzaSize         Dough CheeseType   Topping Quantity Profit
1    Washington 1999-05-30  personal white_regular Mozzarella pepperoni        6    660
2        Quebec 1999-01-02    medium    wheat_thin Mozzarella pepperoni        7    945
3    Washington 1999-05-02    medium    wheat_thin      Swiss pepperoni        3    315
4    California 1999-03-31    medium white_regular Mozzarella    pepper        5    675
5       Ontario 1999-06-19     large stuffed_crust Mozzarella  tomatoes        1    100
6      new York 1999-12-27  personal    wheat_thin      Swiss    onions        6    690
```

Building a cube for our generated data set:



```
> dimnames(revenue_cube)
$date
  [1] "1999-01-01" "1999-01-02" "1999-01-05" "1999-01-06" "1999-01-08" "1999-01-09"
  [7] "1999-01-14" "1999-01-17" "1999-01-18" "1999-01-20" "1999-01-22" "1999-01-24"
 [13] "1999-01-26" "1999-01-28" "1999-01-29" "1999-01-30" "1999-02-02" "1999-02-03"
 [19] "1999-02-06" "1999-02-07" "1999-02-08" "1999-02-10" "1999-02-12" "1999-02-13"
 [25] "1999-02-14" "1999-02-16" "1999-02-19" "1999-02-23" "1999-02-24" "1999-02-26"
 [31] "1999-02-27" "1999-02-28" "1999-03-02" "1999-03-04" "1999-03-06" "1999-03-08"
 [37] "1999-03-09" "1999-03-10" "1999-03-11" "1999-03-12" "1999-03-13" "1999-03-14"
 [43] "1999-03-15" "1999-03-17" "1999-03-18" "1999-03-24" "1999-03-25" "1999-03-26"
 [49] "1999-03-28" "1999-03-29" "1999-03-30" "1999-04-01" "1999-04-02" "1999-04-03"
 [55] "1999-04-05" "1999-04-07" "1999-04-09" "1999-04-10" "1999-04-11" "1999-04-14"
 [61] "1999-04-18" "1999-04-20" "1999-04-21" "1999-04-22" "1999-04-23" "1999-04-25"
 [67] "1999-04-26" "1999-04-28" "1999-04-29" "1999-04-30" "1999-05-02" "1999-05-03"
 [73] "1999-05-04" "1999-05-05" "1999-05-06" "1999-05-07" "1999-05-08" "1999-05-10"
 [79] "1999-05-13" "1999-05-15" "1999-05-18" "1999-05-21" "1999-05-23" "1999-05-25"
 [85] "1999-05-26" "1999-05-27" "1999-06-01" "1999-06-04" "1999-06-05" "1999-06-07"
 [91] "1999-06-08" "1999-06-09" "1999-06-10" "1999-06-11" "1999-06-12" "1999-06-13"
 [97] "1999-06-15" "1999-06-17" "1999-06-19" "1999-06-20" "1999-06-23" "1999-06-26"
[103] "1999-06-28" "1999-07-03" "1999-07-04" "1999-07-06" "1999-07-07" "1999-07-09"
```

Cube dimension:

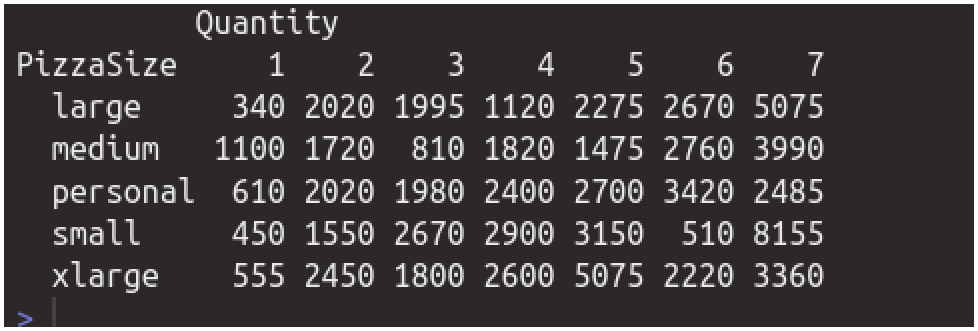


```
$PizzaSize
[1] "large"    "medium"   "personal" "small"    "xlarge"

$Quantity
[1] "1" "2" "3" "4" "5" "6" "7"

$CheeseType
[1] "cheddar"    "Mozzarella" "Swiss"
```

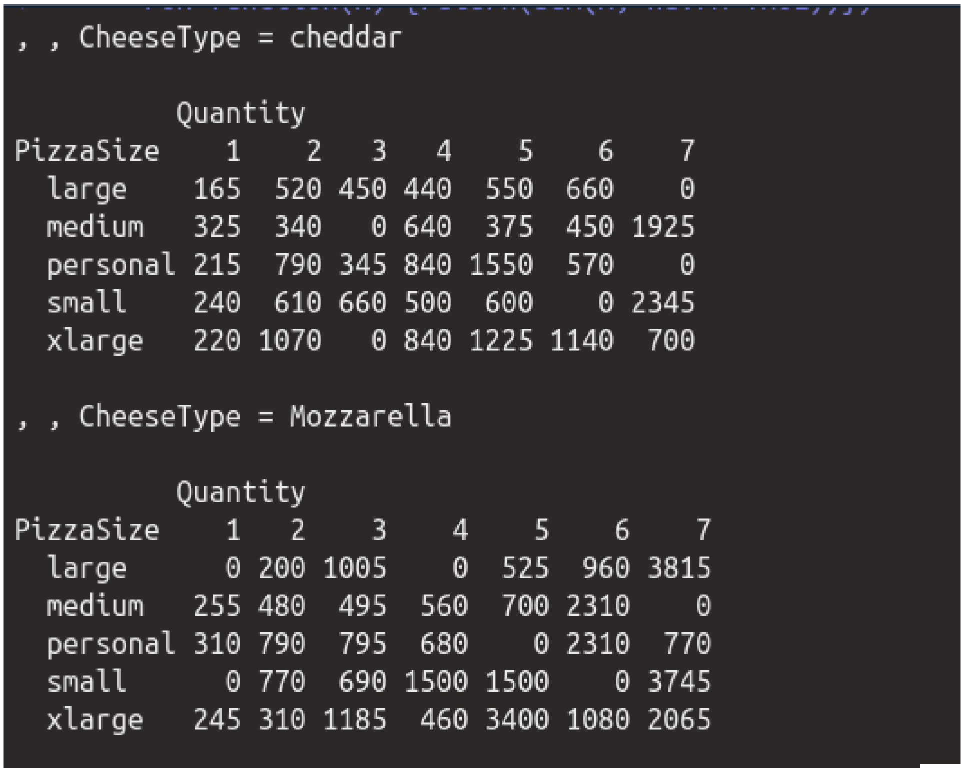
Operating rolling up:



```
          Quantity
PizzaSize     1    2    3    4    5    6    7
  large     340 2020 1995 1120 2275 2670 5075
  medium   1100 1720  810 1820 1475 2760 3990
  personal  610 2020 1980 2400 2700 3420 2485
  small     450 1550 2670 2900 3150  510 8155
  xlarge    555 2450 1800 2600 5075 2220 3360
>
```

It’s shown that the highest quantity values are for large and xlarge pizza size.

Operating drill:

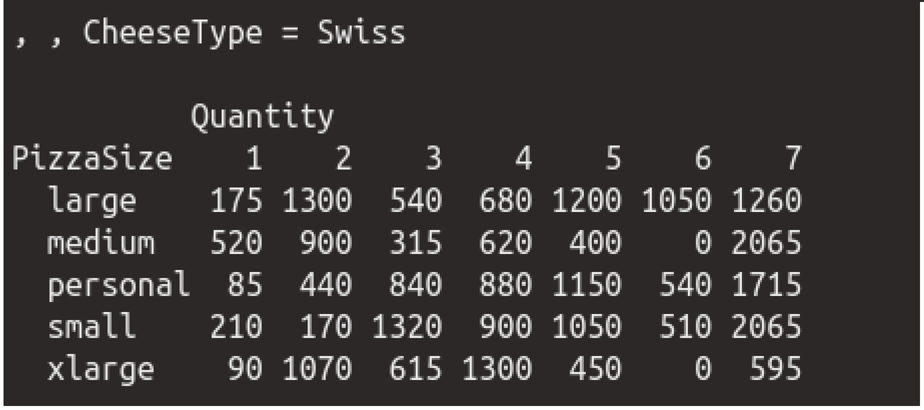


```
, , CheeseType = cheddar

          Quantity
PizzaSize    1    2   3   4    5    6    7
  large    165  520 450 440  550  660    0
  medium   325  340   0 640  375  450 1925
  personal 215  790 345 840 1550  570    0
  small    240  610 660 500  600    0 2345
  xlarge   220 1070   0 840 1225 1140  700

, , CheeseType = Mozzarella

          Quantity
PizzaSize    1   2    3    4    5    6    7
  large      0 200 1005    0  525  960 3815
  medium   255 480  495  560  700 2310    0
  personal 310 790  795  680    0 2310  770
  small      0 770  690 1500 1500    0 3745
  xlarge   245 310 1185  460 3400 1080 2065
```



```
, , CheeseType = Swiss

          Quantity
PizzaSize    1    2    3    4    5    6    7
  large    175 1300  540  680 1200 1050 1260
  medium   520  900  315  620  400    0 2065
  personal  85  440  840  880 1150  540 1715
  small    210  170 1320  900 1050  510 2065
  xlarge    90 1070  615 1300  450    0  595
```

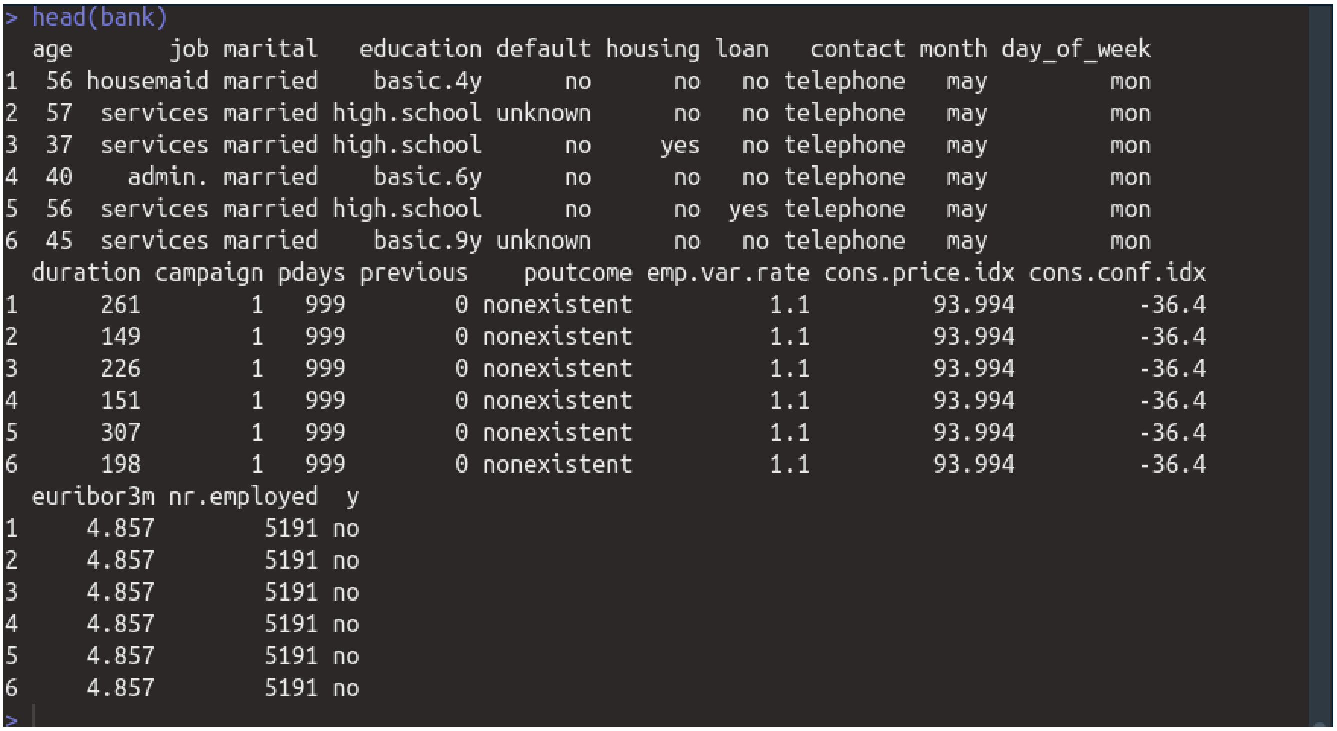
After filtering upon cheese type, we found that the highest quantity values for larger pizza size is for pizza with mozzarella.

Part B: Data Preparation:

Problem overview:

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.

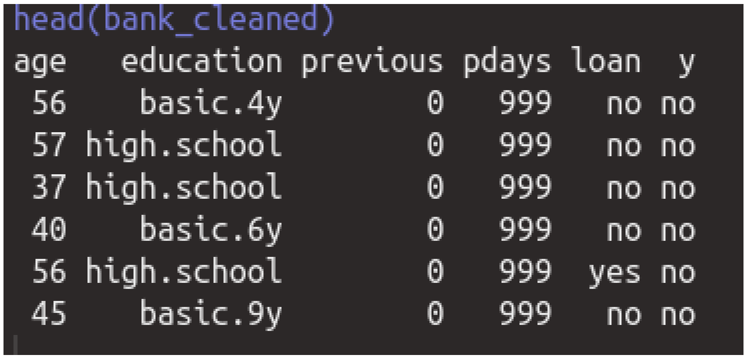
Reading the data set:



```
> head(bank)
  age       job marital   education default housing loan   contact month day_of_week
1  56 housemaid married    basic.4y      no      no   no telephone   may         mon
2  57  services married high.school unknown      no   no telephone   may         mon
3  37  services married high.school      no     yes   no telephone   may         mon
4  40    admin. married    basic.6y      no      no   no telephone   may         mon
5  56  services married high.school      no      no  yes telephone   may         mon
6  45  services married    basic.9y unknown      no   no telephone   may         mon
  duration campaign pdays previous    poutcome emp.var.rate cons.price.idx cons.conf.idx
1      261        1   999        0 nonexistent          1.1         93.994         -36.4
2      149        1   999        0 nonexistent          1.1         93.994         -36.4
3      226        1   999        0 nonexistent          1.1         93.994         -36.4
4      151        1   999        0 nonexistent          1.1         93.994         -36.4
5      307        1   999        0 nonexistent          1.1         93.994         -36.4
6      198        1   999        0 nonexistent          1.1         93.994         -36.4
  euribor3m nr.employed  y
1     4.857        5191 no
2     4.857        5191 no
3     4.857        5191 no
4     4.857        5191 no
5     4.857        5191 no
6     4.857        5191 no
>
```

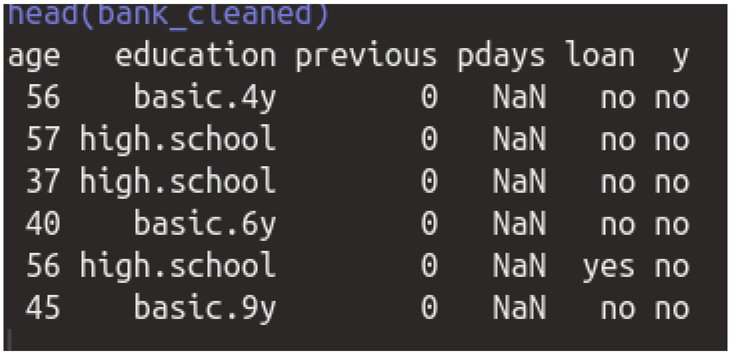
We have investigated data dimension and it has 41188 rows and 12 columns.
We need to select columns of interest.

Data after cleaning:



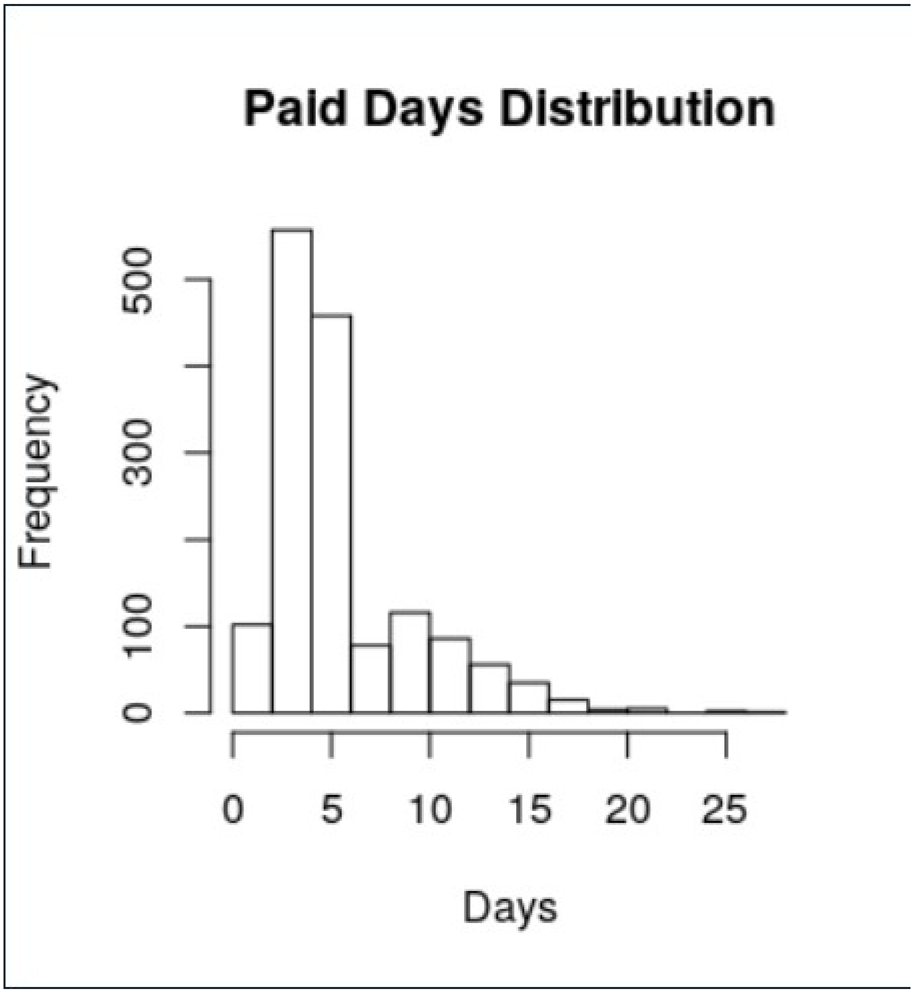
```
head(bank_cleaned)
age    education previous pdays loan  y
 56     basic.4y        0   999   no no
 57 high.school         0   999   no no
 37 high.school         0   999   no no
 40     basic.6y        0   999   no no
 56 high.school         0   999  yes no
 45     basic.9y        0   999   no no
```

We have noticed that the “999” in “pdays” column refers to clients who was last contacted from previous campaign, so we need to set these values to NAN.

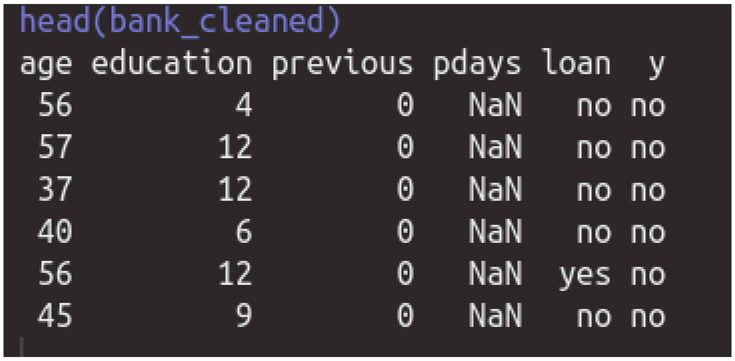


```
head(bank_cleaned)
age   education previous pdays loan   y
 56    basic.4y        0   NaN   no no
 57 high.school        0   NaN   no no
 37 high.school        0   NaN   no no
 40    basic.6y        0   NaN   no no
 56 high.school        0   NaN  yes no
 45    basic.9y        0   NaN   no no
```

we have calculated number of nans after this transformation and it was 39673 records, so this column is useless.
We have build a histogram of paid days, and we have noticed that the most frequent number of days is 5.

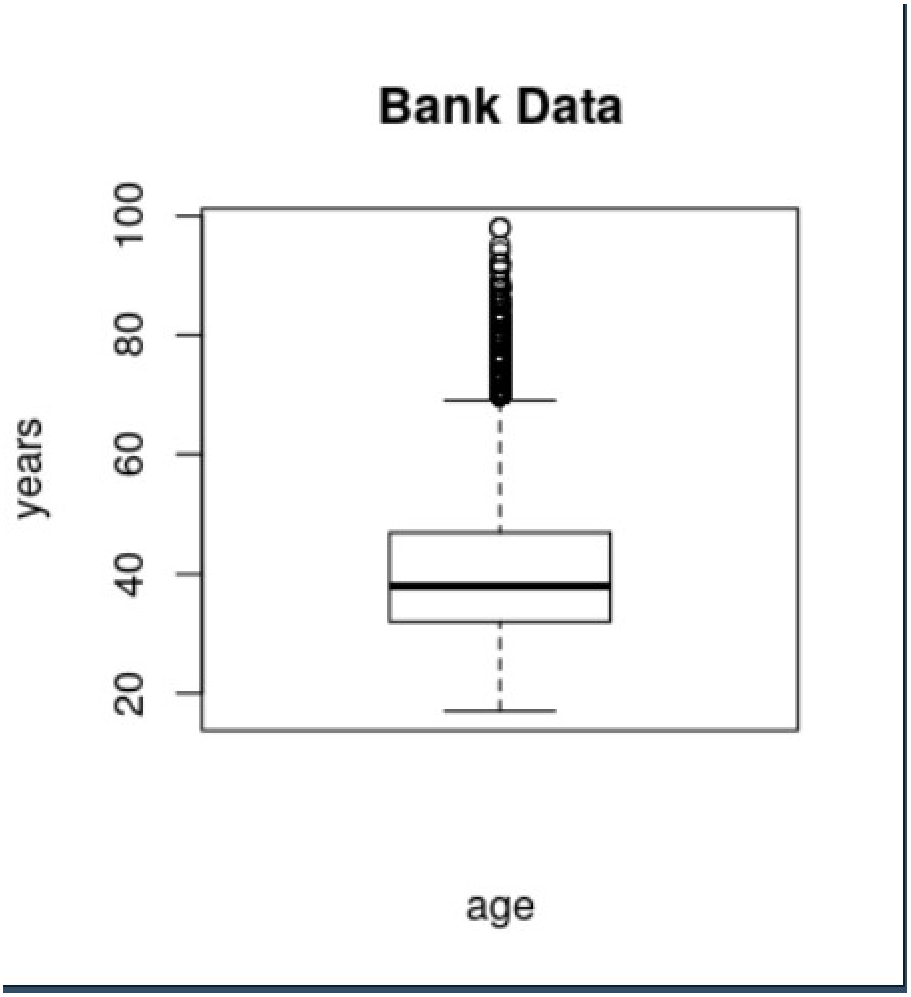


That seems to make no sense, but we lack data in this column.
We need to convert education column’s values to numeric values.

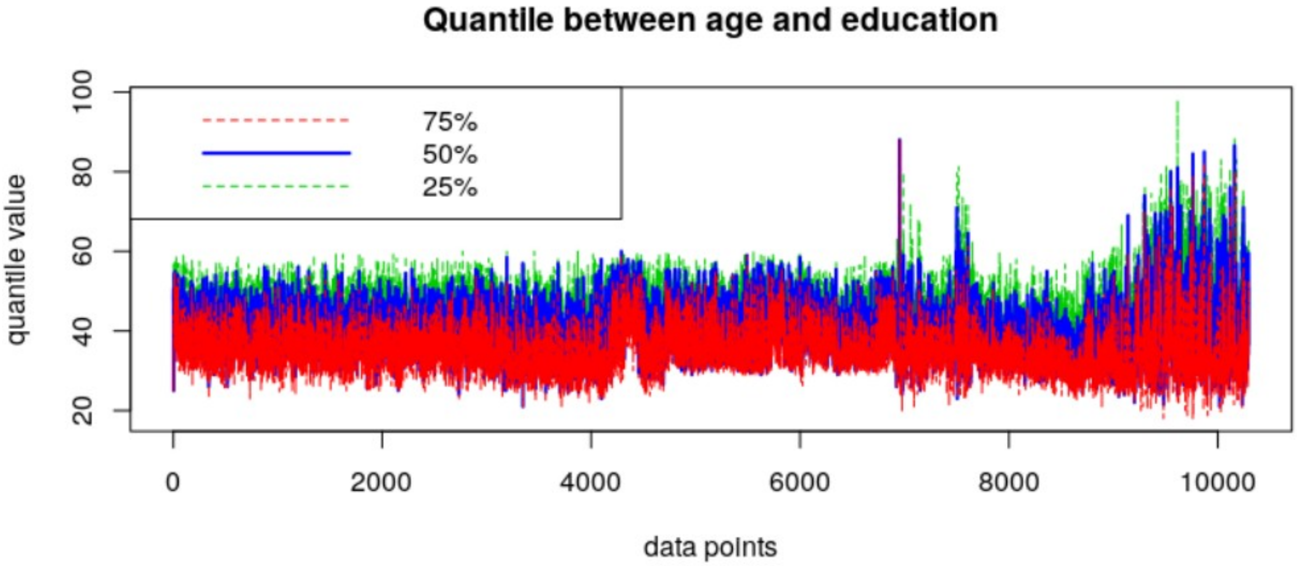


```
head(bank_cleaned)
age education previous pdays loan   y
 56         4        0   NaN   no no
 57        12        0   NaN   no no
 37        12        0   NaN   no no
 40         6        0   NaN   no no
 56        12        0   NaN  yes no
 45         9        0   NaN   no no
```

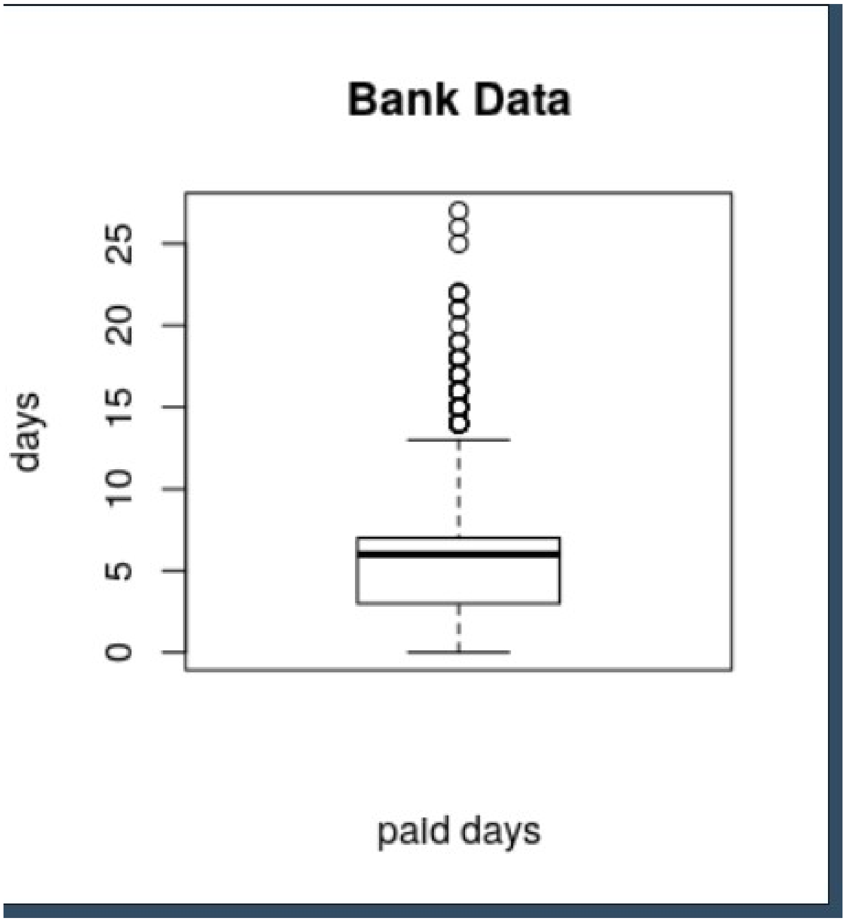
We need to investigate the age column as well.
The mean value of age is 40, the median is 38 and the mode is 31.
We need to perform some visuals for our data for better understanding.
A box plot for age column:
we see the the maximum is approx. 70, the minimum is approx. 18, the median is 38, the first quantile is approx. 30 and the third quantile is approx. 50.



Here is also another plot of quantile of the data set and to include numeric and valid data, we have plotted the age and the education.

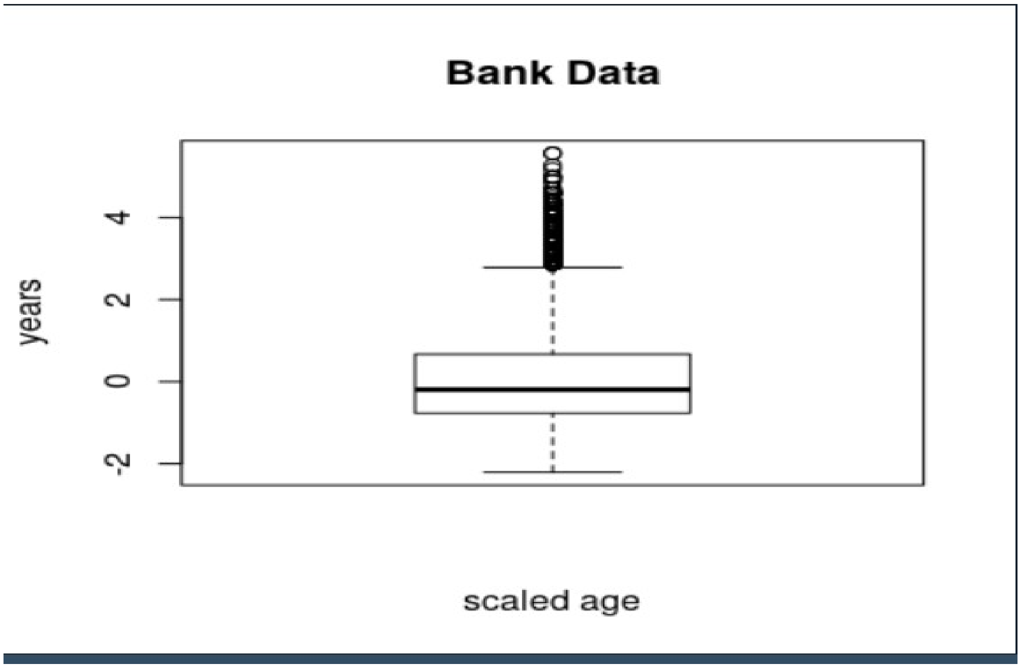


A box plot of paid days:

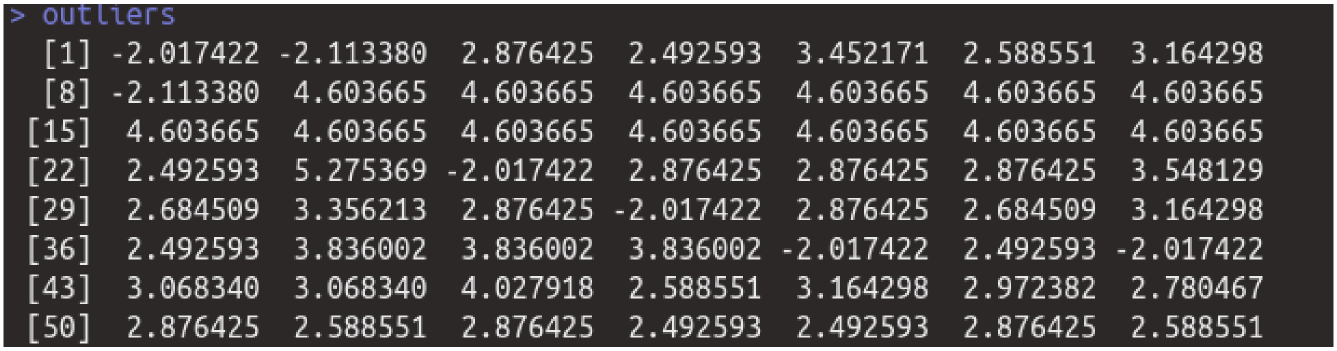


the maximum is approx. 13, the minimum is approx. 0, the median is 6, the first quantile is approx. 3 and the third quantile is approx. 7.

At last we need to scale the age column, here is a box plot again for it but scaled:



Extracting the outliers for external vector, and here is a sample of it:



```
> outliers
 [1] -2.017422 -2.113380  2.876425  2.492593  3.452171  2.588551  3.164298
 [8] -2.113380  4.603665  4.603665  4.603665  4.603665  4.603665  4.603665
[15]  4.603665  4.603665  4.603665  4.603665  4.603665  4.603665  4.603665
[22]  2.492593  5.275369 -2.017422  2.876425  2.876425  2.876425  3.548129
[29]  2.684509  3.356213  2.876425 -2.017422  2.876425  2.684509  3.164298
[36]  2.492593  3.836002  3.836002  3.836002 -2.017422  2.492593 -2.017422
[43]  3.068340  3.068340  4.027918  2.588551  3.164298  2.972382  2.780467
[50]  2.876425  2.588551  2.876425  2.492593  2.492593  2.876425  2.588551
```

References:
1) Course labs and lecture notes.
2) https://www.tutorialspoint.com/r/r_mean_median_mode.htm
3) https://stackoverflow.com/questions/19754764/plot-quantiles-in-r
4) https://stackoverflow.com/questions/16819956/warning-message-in-invalid-factor-level-na-generated