**Project Data Warehouse**

Fabien Tang, M1 BI2

The main goal of this DW is to better structure the data to optimize and make future querries easier. We’ll also use later the K-Mean and K-NN algorithm to determine different profile of blood donor and which kind of profile was giving on March 2007.

1. **Transfusion.data**

Structuration

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Recency | Frequency | Monetary | Time | A |
| 2 | 50 | 12500 | 98 | 1 |

Recency : Months since last donation

Frequency : Total number of donation

Monetary : Total blood donated in c.c

Time : Months since first donation

A : Whether he/she donated blood in March 2007

1. **Logical design of the database**

|  |
| --- |
| Period |
| Period\_ID  Months\_since\_first  Months\_since\_last |

|  |
| --- |
| Donation |
| Donation\_ID  Number\_of\_donation  Quantity\_of\_blood |

|  |
| --- |
| Donor |
| Period\_ID  Donation\_ID  March\_2007 |

1. **Creation of the Data Warehouse (create\_table.sql)**

**CREATE TABLE** Period

(

Period\_ID **number(7)** **primary key**,

Months\_since\_first **number(7)**,

Months\_since\_last **number(7)**

) ;

**CREATE TABLE** Donation

(

Donation\_ID **number (7) primary key**,

Number\_of\_donation **number(7)**,

Quantity\_of\_blood **number(7)**

) ;

**CREATE TABLE** Donor

(

Period\_ID **number(7)**,

Donation\_ID **number(7)**,

March\_2007 **number(1)**,

**foreign key** (Period\_ID) **references** Period(Period\_ID),

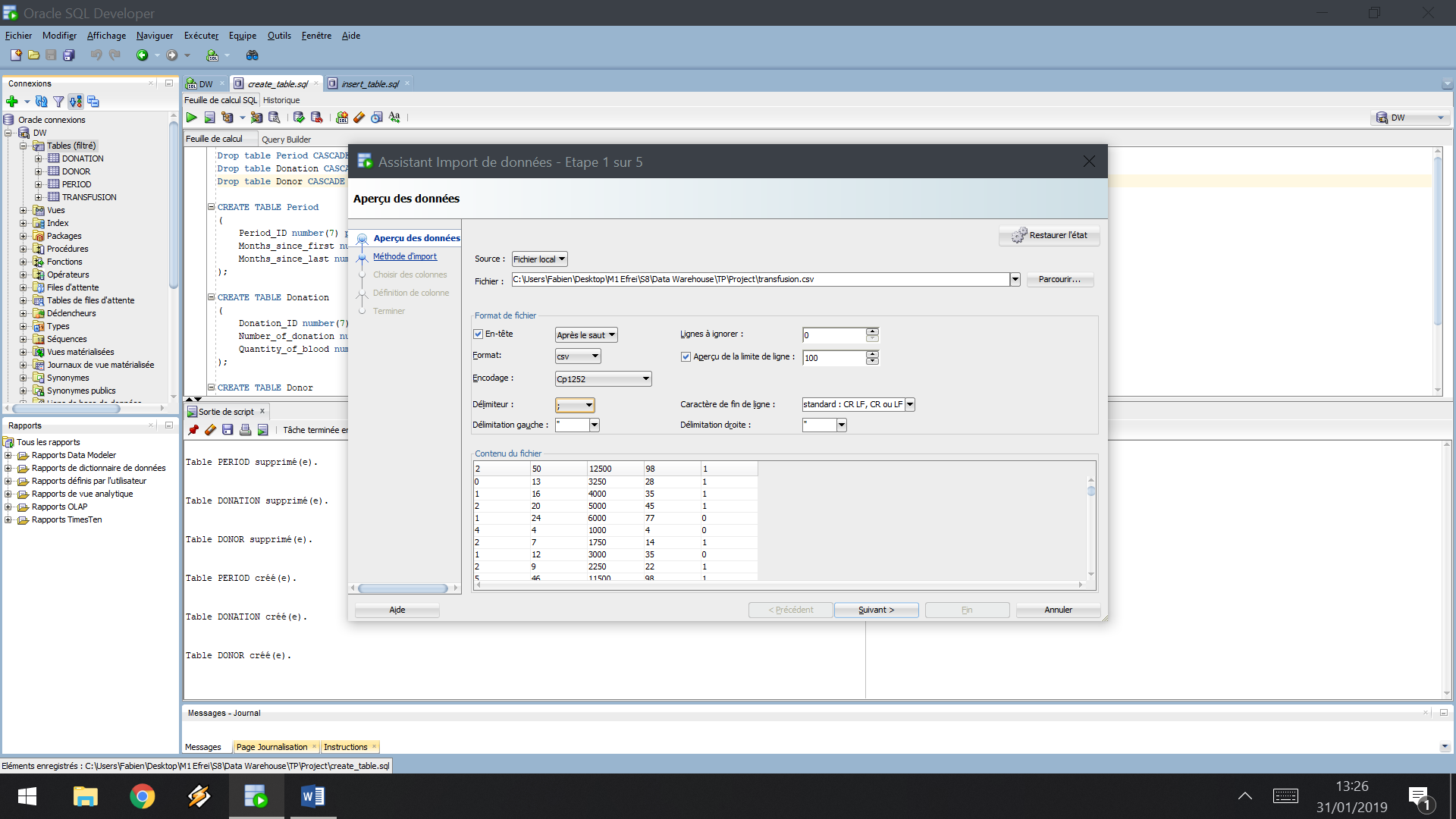
**foreign key** (Donation\_ID) **references** Donation(Donation\_ID)

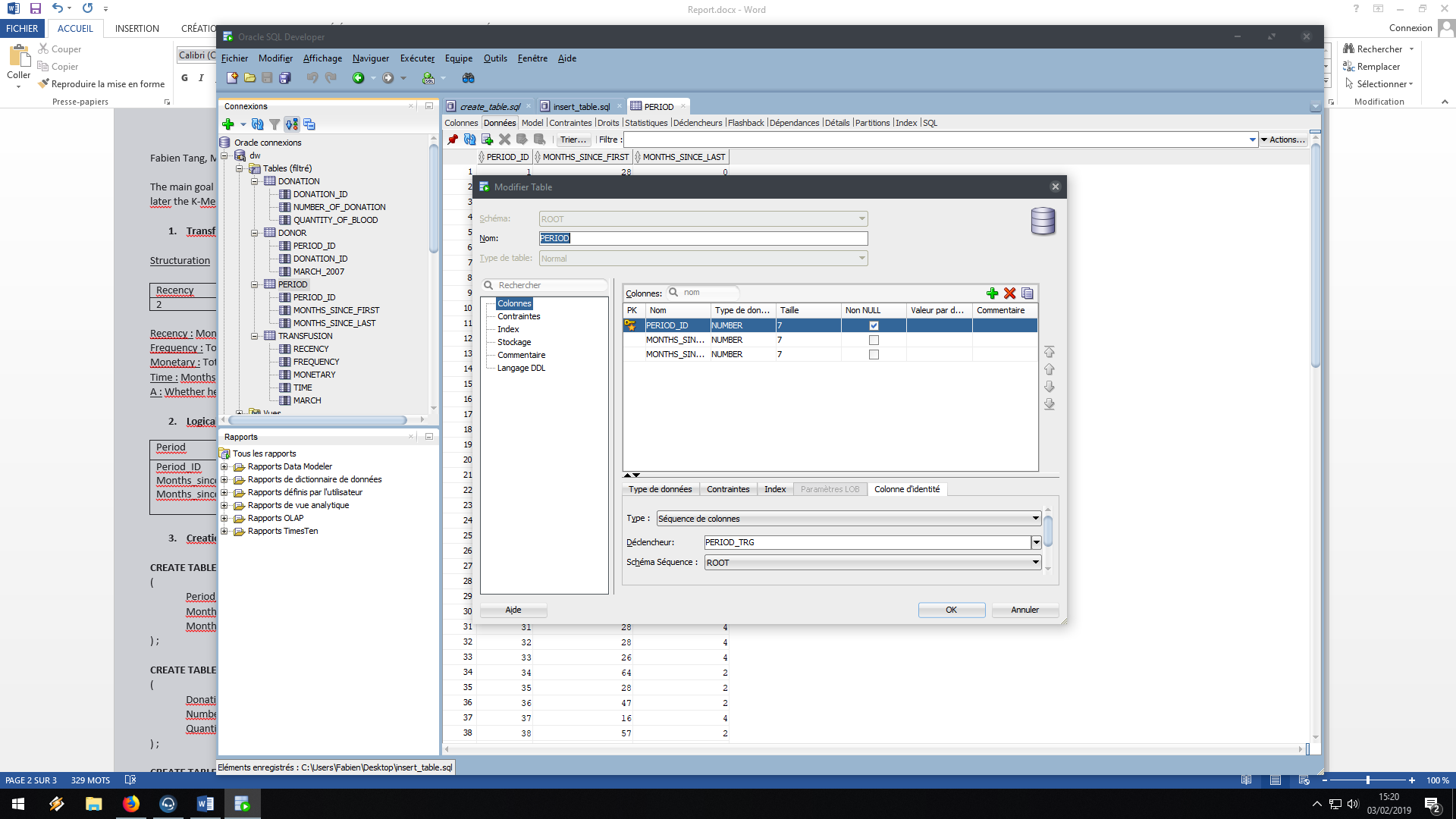
) ;

1. **Insert data into the Data Warehouse (insert\_table.sql)**

To insert the data into the DW :

* Convert Transfusion.data into Transfusion.txt file (just change the file extension)
* Open the Transfusion.txt with Excel then save it as a Transfusion.csv
* From SQLDevelopper import Transfusion.csv into a tempory table to store all the data (Photo 1)
* Create new attributes Period\_ID/Donation\_ID auto-incremented (Photo 2)
* Insert into the others tables the data from the tempory table
* Export the tables to have a backup (**export.sql**)





**Insert** into Period (months\_since\_first, months\_since\_last)

**Select** time,recency

**From** transfusion;

**Insert** into Donation (number\_of\_donation, quantity\_of\_blood)

**Select** frequency, monetary

**From** transfusion;

**Create** sequence seq\_fact start with 1;

**Create** sequence seq\_fact2 start with 1;

**Insert** into Donor (period\_id, donation\_id, march\_2007)

**Select** seq\_fact.nextval, seq\_fact2.nextval, t.march

**From** transfusion t;

1. **Query the Data Warehouse using OLAP (CUBE,…) queries in order to obtain in output a rectangular matrix X with n lines and p columns containing in lines the objects and in columns theirs characteristics. These queries depends on the data problem and application field. You can obtain several matrices. Explain the queries and the results.**

The aim with the first matrix will be with the K-Means method to determine different profil of donator based on the months since last donation and the total amount of blood given.

**Select** months\_since\_last, sum(quantity\_of\_blood)

**From** period, donor, donation

**Where** donor.period\_id=period.period\_id and donation.donation\_id=donor.donation\_id

**Group by** cube(months\_since\_last)

**Order by** 2;

The second matrix, will be usefull to determine with K-NN Method which kind of profile based on months since last donation and total number of donation is giving on March\_2007.

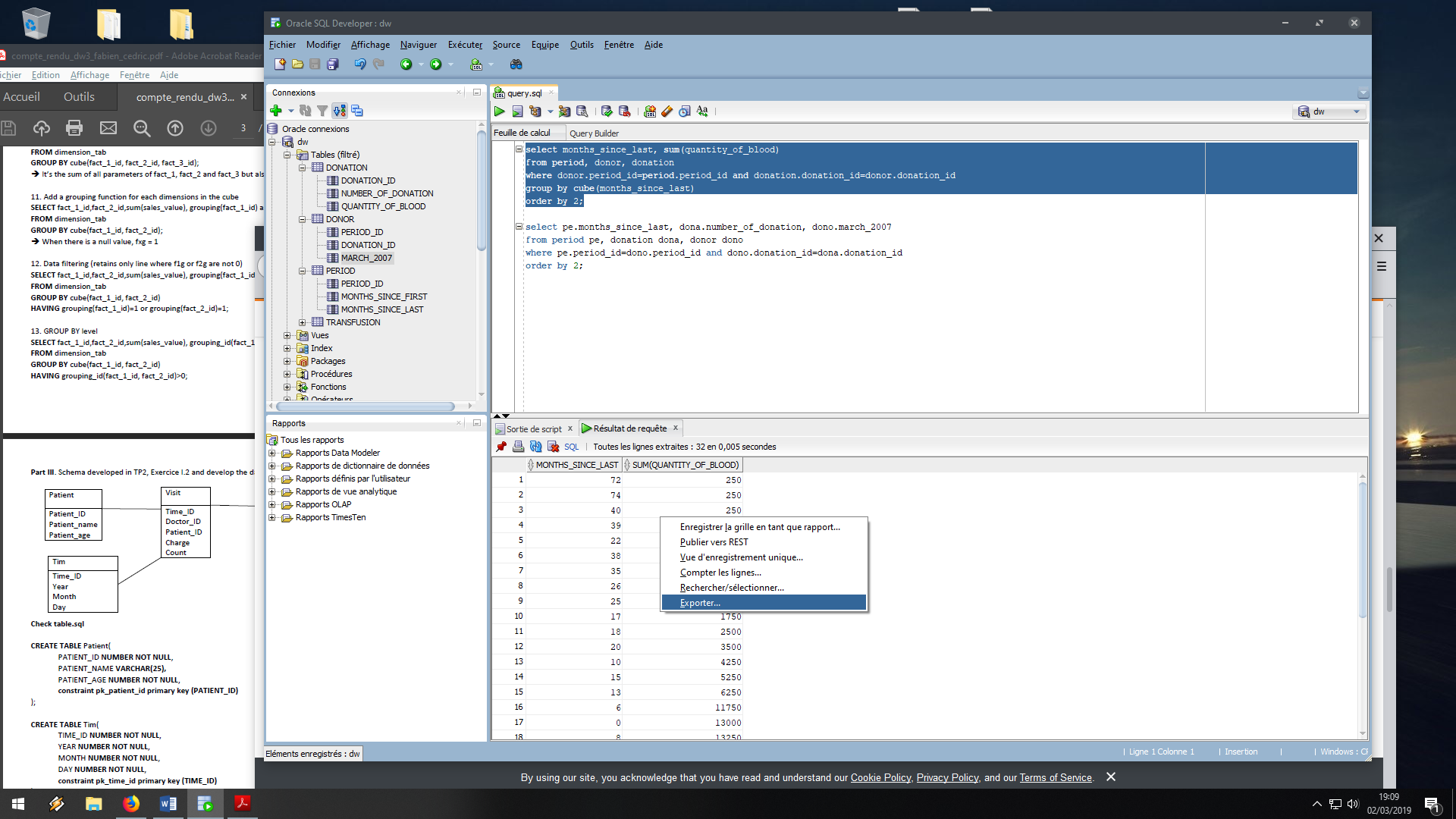
**Select** pe.months\_since\_last, dona.number\_of\_donation, dono.march\_2007

**From** period pe, donation dona, donor dono

**Where** pe.period\_id=dono.period\_id and dono.donation\_id=dona.donation\_id

**Order** by 2;

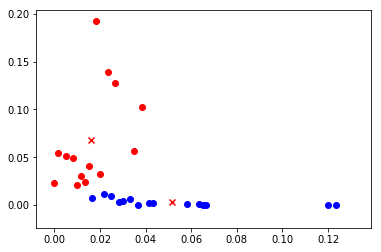
Export through sql developper : (matrix 1 : analyse\_1.csv, matrix 2 : analyse\_2.csv)



1. **Use Phyton in order to load the obtained matrix (matrices) and visualize the data** **using scatter plot. Analyse the results. Apply a clustering model using k-means method and** **visualize the results. Explain the results.**

**Analysis 1 – K-MEAN (Python : Analyse\_1.ipynb)**

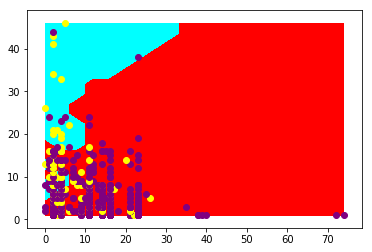
In this first analysis the K-Mean method gave us those 2 profiles of populations (2 clusters)



X = Months since last donation (x600) Y = Sum (quantity of blood) (x550250) (multiply for denormalized form)

**Analysis 2 – K-NN (Python : Analyse\_2.ipynb)**

In the second analysis with the K-NN Method we can easily say that if the number of donation is high and month since last donation low, the people would gave their blood on March 2007.



X = Months since last donation / Y = number of donation

Yellow dot = gave his blood on March 2007 / Purple dot = did not gave his blood on March 2007

Blue part : prediction the person will give his blood // Red part : prediction the person will not give his blood

**Conclusion**

In conclusion, through this project we sucessfully structured the transfusion database with a Star Schema modelization. We then used OLAP Query through group by and Cube to produce csv that were used to do some relevant analysis. Finally, the analysis shown us 2 major profile of population and help us to easily see which profile gave their blood on March 2007