Project - Retail Store Data

----------------------------------

The data set contains the data of the following 4 months:

D11: Transaction data collected in November, 2000

D12: Transaction data collected in December, 2000

D01: Transaction data collected in January, 2001

D02: Transaction data collected in February, 2001

**Format of Transaction Data:**

data columns separated by “;”

**Column definition:**

1)Transaction date and time (Time is invalid)

2)Customer I.D

3)Age: 10 possible values

A <25,B 25-29,C 30-34,D 35-39,E 40-44,F 45-49,G 50-54,H 55-59,I 60-64,J >65

4)Residence Area: 8 possible values, A-F: zip code area: 105,106,110,114,115,221,G:others, H: Unknown Distance to store, from the closest: 115,221,114,105,106,110

5)Product subclass (category)

6)Product ID

7)Qty or Number of units

8)Total Cost

9)Total Sales

Txn\_dt : String

Custno : String

Age : String

Zipcode : String

Category : String

Product : String

Qty : Int

Cost : bigint

Sales : bigint

Row format delimited

Fields terminated by ‘\;’

1. Count of unique customers and total sales for each age group and for a given month = Jan

**Hint : where month(txn\_dt) = 1**

A 5000 600000

B 4500 540000

2) Count of unique customers and total sales for one age group(A) for all products - [ sort data on totalsales desc- find top 10]

ProdA count of unique cust total sales

ProdB ‘’ ‘’

**Hint : where trim(age) = 'A'**

3) Area wise sales

4) find top 10 viable products (prod which give highest profit)

5) find all loss making products – Display all the loss making products from highest loss to the least

Total number records : 817741 records