Retail Sports Dataset

Retail sport store dataset has Transaction details data and the Customer details data. We need to analyse the data and give some insights for business growth.
Retail Sports dataset has 2 dataset files. Transaction File and Customers File.

Understanding Data

The data format is comma separated values. There are 50000 observations.

- Transaction File contains 9 columns made up of following:
  txnno ,txndate,custno,amount,category,product,city,state,spendby
- Customer file contains 5 columns made up of following:
  custno,firstname,lastname,age, profession

UseCase = Analyse the amount spend by different age group people

```
hive> create database retail;

hive> use retail;

hive> create table Transaction(

   > txnno int,txndate string,custno string,amount double,category string,product string,city
   string,state string,spendby string)

   > row format delimited fields terminated by ','

   > stored as textfile;

hive> load data local inpath 'file:///home/cloudera/khasimbabu/HIVE_Excercise/Dataset_txns1.txt'
into table Transaction;

hive> select * from Transaction limit 5;

0       06-26-2011      4007024         40.33   Exercise & Fitness      Cardio Machine Accessories
        Clarksville     Tennessee       credit
1       05-26-2011      4006742         198.44  Exercise & Fitness      Weightlifting Gloves
        Long Beach      California      credit
2       06-01-2011      4009775         5.58    Exercise & Fitness      Weightlifting Machine
Accessories     Anaheim         California      credit
3       06-05-2011      4002199         198.19  Gymnastics      Gymnastics Rings        Milwaukee
        Wisconsin       credit
4       12-17-2011      4002613         98.81   Team Sports     Field Hockey    Nashville
        Tennessee       credit

hive> select count(*) from Transaction;
→ Hive will convert the query and launch a Map-Reduce job and give the results.

hive> create table Customer(

   > custno string,firstname string,lastname string,age int, profession string)
```

```
   > row format delimited fields terminated by ',';

hive> load data local inpath 'file:///home/cloudera/khasimbabu/HIVE_Excercise/Dataset_custs.txt'
into table Customer;

hive> select * from Customer limit 10;

OK
4000001         Kristina Chung  55      Pilot
4000002         Paige   Chen    74      Teacher
4000003         Sherri  Melton  34      Firefighter
4000004         Gretchen        Hill    66      Computer hardware engineer
4000005         Karen   Puckett 74      Lawyer
4000006         Patrick Song    42      Veterinarian
4000007         Elsie   Hamilton        43      Pilot
4000008         Hazel   Bender  63      Carpenter
4000009         Malcolm         Wagner  39      Artist
4000010         Dolores McLaughlin      60      Writer
Time taken: 0.079 seconds, Fetched: 10 row(s)

hive> create table out1(
   > custno string,firstname string,age int,profession string,amount double,product string)
   > row format delimited fields terminated by ',';
hive> insert overwrite table out1
   > select a.custno,a.firstname,a.age,a.profession,b.amount,b.product
   > from Customer a JOIN Transaction b ON a.custno=b.custno;
hive> set hive.cli.print.header=true;
hive> select * from out1 limit 10;
OK
out1.custno     out1.firstname  out1.age        out1.profession out1.amount     out1.product
4007024         Cameron         59      Actor   40.33   Cardio Machine Accessories
4006742         Gregory         36      Accountant      198.44  Weightlifting Gloves
4009775         Ruby    44      Designer        5.58    Weightlifting Machine Accessories
4002199         Keith   44      Police officer  198.19  Gymnastics Rings
4002613         Hugh    43      Engineering technician  98.81   Field Hockey
4007591         Jennifer54      Electrician     193.63  Camping & Backpacking & Hiking
4002190         Sheryl  62      Designer        27.89   Jigsaw Puzzles
4002964         Ken     67      Recreation and fitness worker   96.01   Sandboxes
4007361         Terri   52      Loan officer    10.44   Snowmobiling
4004798         Geoffrey        65      Chemist         152.46  Bungee Jumping
Time taken: 0.069 seconds, Fetched: 10 row(s)

hive> create table out2(
   > custno string,firstname string,age int, profession string, amount double, product string, level
string)
   > row format delimited fields terminated by ',';
hive> insert overwrite table out2
   > select *, case
   > when age<30 then 'low'
```

```
   > when age>=30 and age<50 then 'middle'
   > when age>=50 then 'old'
   > else 'others'
   > end
   > from out1;
hive> select * from out2 limit 5;
OK
out2.custno     out2.firstname  out2.age  out2.profession out2.amount  out2.product     out2.level
4007024         Cameron         59        Actor    40.33   Cardio Machine Accessories      old
4006742         Gregory         36        Accountant       198.44  Weightlifting Gloves     middle
4009775         Ruby    44      Designer  5.58    Weightlifting Machine Accessories       middle
4002199         Keith   44      Police officer   198.19  Gymnastics Rings        middle
4002613         Hugh    43      Engineering technician   98.81   Field Hockey    middle
Time taken: 0.066 seconds, Fetched: 5 row(s)

hive> create table out3(
   > level string, amount double)
   > row format delimited fields terminated by ',';
hive> insert overwrite table out3
   > select level, SUM(amount) from out2 group by level;
hive> select * from out3;
OK
out3.level      out3.amount
low     725221.3399999988
middle  1855861.669999996
old     2529100.310000011
Time taken: 0.079 seconds, Fetched: 3 row(s)
```