Shanqing Gu

MSDS 7330 HW4 Due Date: 06/11/2018

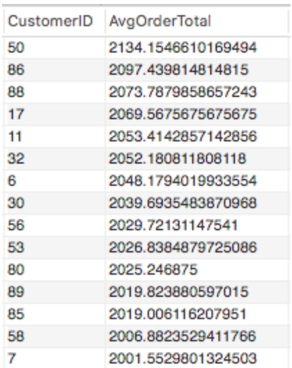
Unit 5 HomeWork

1. Write an SQL to list the customers in terms of the Avg Order Totals sorted in a descending order

SQL scripts:

```
select CustomerID, avg(Quantity*PricePerQty) as AvgOrderTotal
from (select CustomerID, OrderNumber, PricePerQty, Quantity
         from CustomerTable ct, OrderHeader oh, Orderline ol
         where ct.CustomerID=oh.CustomerTable_CustomerID
         and oh.orderNumber=ol.OrderHeader_orderNumber) as c
group by CustomerID
order by AvgOrderTotal desc;
```

SQL results:



CustomerID	AvgOrderTotal
50	2134.1546610169494
86	2097.439814814815
88	2073.7879858657243
17	2069.5675675675675
11	2053.4142857142856
32	2052.180811808118
6	2048.1794019933554
30	2039.6935483870968
56	2029.72131147541
53	2026.8384879725086
80	2025.246875
89	2019.823880597015
85	2019.006116207951
58	2006.8823529411766
7	2001.5529801324503

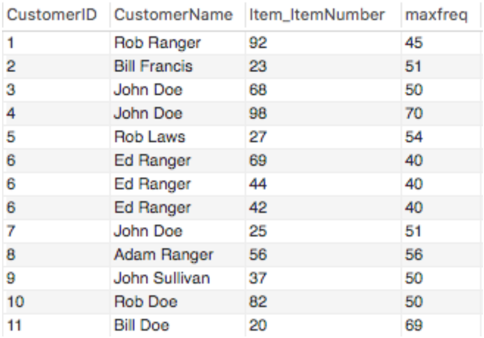
2. Write an SQL query to list the most frequently order item for each customer

SQL scripts:

```
With c as (select
ct.CustomerID, ct.CustomerName, sum(Quantity) as TotalQty, ol.Item_ItemNumber
from
OrderHeader as oh,
OrderLine as ol,
CustomerTable as ct
where
oh.OrderNumber = ol.OrderHeader_OrderNumber and
oh.CustomerTable_CustomerID = ct.CustomerID
group by CustomerID, Item_ItemNumber)

select c.CustomerID, c.CustomerName, Item_ItemNumber, maxfreq
from c, (select CustomerID,max(TotalQty) as maxfreq from c group by CustomerID) as b
where c.CustomerID=b.CustomerID and c.TotalQty=maxfreq;
```

SQL results:

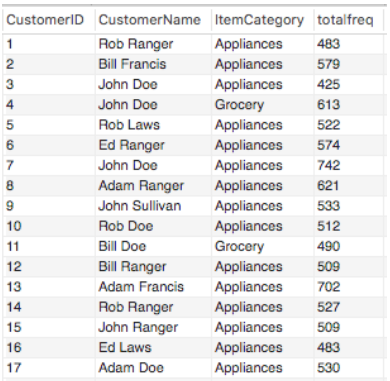


CustomerID	CustomerName	Item_ItemNumber	maxfreq
1	Rob Ranger	92	45
2	Bill Francis	23	51
3	John Doe	68	50
4	John Doe	98	70
5	Rob Laws	27	54
6	Ed Ranger	69	40
6	Ed Ranger	44	40
6	Ed Ranger	42	40
7	John Doe	25	51
8	Adam Ranger	56	56
9	John Sullivan	37	50
10	Rob Doe	82	50
11	Bill Doe	20	69

3. Write an SQL query to list the most frequently ordered item category for each customer

SQL scripts:

```
with c as (SELECT CustomerID, CustomerName,ItemCategory, sum(Quantity) as totalfreq
                 FROM CustomerTable ct,OrderHeader oh, OrderLine ol,Item it
                 where ct.CustomerID=oh.CustomerTable_CustomerID and
oh.OrderNumber=ol.OrderHeader_OrderNumber
                 and ol.Item_ItemNumber=it.ItemNumber
                 group by CustomerID, ItemCategory)
select c.*
from c, (select distinct CustomerID, max(totalfreq) as maxcategory from c group by
CustomerID) d
where c.CustomerID=d.CustomerID and c.totalfreq=d.maxcategory;
```



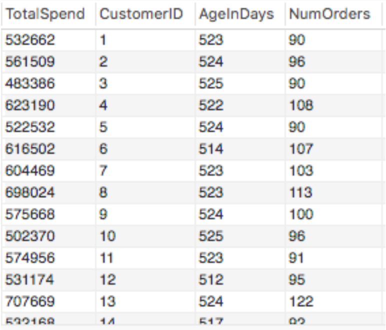
CustomerID	CustomerName	ItemCategory	totalfreq
1	Rob Ranger	Appliances	483
2	Bill Francis	Appliances	579
3	John Doe	Appliances	425
4	John Doe	Grocery	613
5	Rob Laws	Appliances	522
6	Ed Ranger	Appliances	574
7	John Doe	Appliances	742
8	Adam Ranger	Appliances	621
9	John Sullivan	Appliances	533
10	Rob Doe	Appliances	512
11	Bill Doe	Grocery	490
12	Bill Ranger	Appliances	509
13	Adam Francis	Appliances	702
14	Rob Ranger	Appliances	527
15	John Ranger	Appliances	509
16	Ed Laws	Appliances	483
17	Adam Doe	Appliances	530

4. The following query we executed in the class calcs various parameters for each customer. Using statistical procedures test and report the columns that have the highest correlations. Are there any conclusions you can draw from this test?

```
select
sum(Quantity*PricePerQty) TotalSpend ,
oh.CustomerTable_CustomerID CustomerID  ,
datediff(curdate(),min(oh.OrderDate) ) AgeInDays ,
count(distinct oh.OrderNumber) NumOrders
from
OrderHeader oh ,
OrderLine ol
```

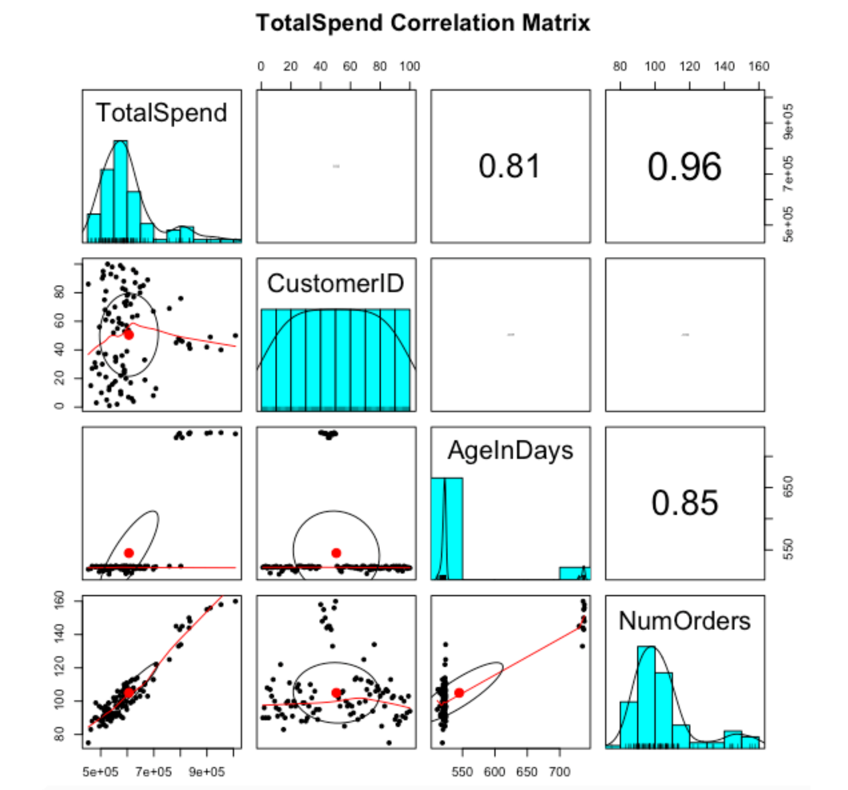
```
where
oh.OrderNumber=ol.OrderHeader_OrderNumber
group by
CustomerID;
```

Export the csv data after running the SQL query.



TotalSpend	CustomerID	AgeInDays	NumOrders
532662	1	523	90
561509	2	524	96
483386	3	525	90
623190	4	522	108
522532	5	524	90
616502	6	514	107
604469	7	523	103
698024	8	523	113
575668	9	524	100
502370	10	525	96
574956	11	523	91
531174	12	512	95
707669	13	524	122
532168	14	517	92

As shown below, the highest correlation columns are TotalSpend and NumOrders with correlation efficient at 0.96, indicating their strong positive linear relationship.



```
R Codes:

hw5_4 <- read.csv (file='/Users/shanqinggu/Desktop/hw5_4.csv')
str(hw5_4)
library(psych)
pairs.panels(hw5_4, scale=TRUE, main="TotalSpend Correlation Matrix")
```