Christopher Broderick

Data Mining Homework 2

1)

Make

Make	sum(Price)
*	670500
Honda	248500
Mazda	85000
Ford	170000
GM	110000
Toyota	57000

Make, Type

Make	Type	sum(Price)
*	*	670500
Honda	*	248500
Honda	Car	36500
Honda	SUV	32000
Honda	Van	180000
Mazda	Car	85000
Ford	*	170000
Ford	Truck	79000
Ford	Van	91000

GM	*	110000
GM	Car	88000
GM	Truck	22000
Toyota	*	57000
Toyota	SUV	29000
Toyota	Car	28000
*	Car	237500
*	SUV	61000
*	Truck	101000
*	Van	271000

Make, Type, Year

Make	Type	Year	sum(price)
*	*	*	670500
Honda	*	*	248500
Honda	Car	*	36500
Honda	Car	2008	12000
Honda	Car	2014	24500
Honda	SUV	2009	32000
Honda	Van	*	180000
Honda	Van	2010	97000
Honda	Van	2011	42000
Honda	Van	2015	41000
Mazda	Car	*	85000
Mazda	Car	2011	32000
Mazda	Car	2012	25000

Mazda	Car	2015	28000
Ford	*	*	170000
Ford	Truck	*	79000
Ford	Truck	2008	50000
Ford	Truck	2009	29000
Ford	Van	*	91000
Ford	Van	2008	43000
Ford	Van	2009	24000
Ford	Van	2011	24000
Ford	*	2008	93000
Ford	*	2009	53000
GM	*	*	110000
GM	Car	2012	88000
GM	Truck	2010	22000
Toyota	*	*	57000
Toyota	Car	2011	29000
Toyota	SUV	2009	28000
*	Car	*	237500
*	Car	2011	60000
*	Car	2012	113000
*	SUV	2009	61000
*	Truck	*	101000
*	Van	*	271000
*	Van	2011	66000
*	*	2008	105000
*	*	2009	114000

*	*	2010	119000
*	*	2011	126000
*	*	2015	69000

Make, Type, Year, Color

Make	Type	Year	Color	sum(Price)
*	*	*	*	670500
Honda	*	*	*	248500
Honda	Car	*	Red	36500
Honda	Car	2008	Red	12000
Honda	Car	2014	Red	24500
Honda	SUV	2009	Red	32000
Honda	Van	*	*	180000
Honda	Van	2010	*	97000
Honda	Van	2010	Green	68000
Honda	Van	2010	White	29000
Honda	Van	2011	Red	42000
Honda	Van	2011	Black	41000
Honda	*	*	Red	110500
Mazda	Car	*	*	85000
Mazda	Car	*	Black	60000
Mazda	Car	2011	Black	32000
Mazda	Car	2015	Black	28000
Mazda	Car	2012	Blue	25000
Ford	*	*	*	170000

Ford	Truck	*	*	79000
Ford	Truck	2008	Blue	50000
Ford	Truck	2009	Black	29000
Ford	Van	*	*	91000
Ford	Van	2008	Red	43000
Ford	Van	*	Black	48000
Ford	Van	2009	Black	24000
Ford	Van	2011	Black	24000
Ford	*	2008	*	93000
Ford	*	2009	Black	53000
Ford	*	*	Black	77000
GM	*	*	*	110000
GM	Truck	2010	Red	22000
GM	Car	2012	*	88000
GM	Car	2012	Blue	42000
GM	Car	2012	Red	22000
GM	Car	2012	White	24000
GM	*	*	Red	44000
Toyota	*	*	Black	57000
Toyota	SUV	2009	Black	29000
Toyota	Car	2011	Black	28000
*	Car	*	*	237500
*	Car	2012	*	113000
*	Car	2012	Blue	67000
*	Car	*	Black	88000
*	Car	2011	Black	60000

*	Car	*	Red	58500
*	SUV	2009	*	61000
*	Truck	*	*	101000
*	Van	*	*	271000
*	Van	2011	*	66000
*	Van	*	Black	89000
*	Van	*	Red	85000
*	*	2008	*	105000
*	*	2008	Red	55000
*	*	2009	*	114000
*	*	2009	Black	82000
*	*	2010	*	119000
*	*	2011	*	126000
*	*	2011	Black	84000
*	*	2015	Black	69000
*	*	*	Black	235000
*	*	*	Blue	117000
*	*	*	Red	197500
*	*	*	White	53000

2)

Code and results are attached in the submission along with this document.

a)

Node ID	Outflow
1	93
2	31
3	97
4	106
5	176

b)

Node ID	[indegree, outdegree]
1	[3,4]
2	[2,3]
3	[2,4]
4	[2,6]
5	[2,5]
6	[2,5]
8	[1,0]
17	[1,0]
18	[3,0]
19	[1,0]
20	[1,0]

26	[1,0]
29	[2,0]
30	[1,0]
31	[1,0]
32	[2,0]

3)

Input:

003	Alicia	student
001	John	student
001	Tim	student
002	Gary	student
002	Jason	student
003	Ohio	State
001	Indiana	State
003	Phil	student
002	Kentucky	State
001	Mary	student
002	Carla	student
003	Pete	student

After Mapper:

The mapper takes the school Id and makes it the key. Then for the value it makes it [0,name] if it is a state and [1,name] if it is a student. So [003 Alicia Student] goes to 003 [1, Alicia] and [003 Ohio State] goes to 003 [0, Ohio]

Key	Value
003	[1, Alicia]
001	[1, John]
001	[1, Tim]
002	[1, Gary]
002	[1, Jason]
003	[0, Ohio]
001	[0, Indiana]
003	[1, Phil]
002	[0, Kentucky]
001	[1, Mary]
002	[1, Carla]
003	[1, Pete]

After Shuffle Sort:

The shuffle sort brings all the values together under each of their keys. In this instance a secondary sort is also added to sort within these keys by the first value in the pair. That is the states (first value 0) will come first and the students will be the rest of the entries

Key	Value
001	[[0,Indiana], [1, John], [1, Tim], [1, Mary]]
002	[[0, Kentucky], [1, Gary], [1, Jason], [1, Carla]]
003	[[0, Ohio], [1, Alicia], [1, Phil], [1, Pete]]

After Reduce:

The reducer takes the state for each school id and then adds it to each student. Each of the student state pairs are output along with the school ID.

001	John	Indiana
001	Tim	Indiana
001	Mary	Indiana
002	Gary	Kentucky
002	Jason	Kentucky
002	Carla	Kentucky
003	Alicia	Ohio
003	Phil	Ohio
003	Pete	Ohio