CSC4008 Assignment 1

121020163 沈驰皓

1 Sales Analysis

Here we retain two decimal places.

(a)

Type	Total Sales
A	4331014722.75
B	2000700736.82
C	405503527.54

(b)

Is Holiday	Average Sales
Holidays	17035.82
Non-Holidays	15901.45

From the data above, we can conclude that sales are generally higher during holidays.

2 People You Might Know

(a) Pipeline: Firstly, I create an rdd containing all edges and output a dictionary containing all friends corresponding to each user by using this rdd. Then, I do join to this rdd and use subtraction to eliminate those who have already been friends to find all potential mutual friends pairs. By using the dictionary above, we can find the mutual friend number for each pair by doing intersection. Finally, we sort the number and give the output pairs for each user.

(b) The recommendations for the users with following user IDs: 10, 152, 288, 603, 714, 1525, 2434, 2681.

10 [2, 3, 4, 5, 6, 7, 8, 9, 11, 12]

152 [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]

288 [71, 1525, 69, 90, 217, 2348, 2351, 2352, 2354, 2356]

603 [1, 289, 290, 291, 292, 293, 294, 295, 296, 297]

714 [1, 712, 713, 715, 717, 718, 1525, 90, 217, 247]

1525 [288, 1, 710, 714, 603]

2434 [71, 288, 711, 716, 719, 720, 2348, 2351, 2352, 2354]

2681 [71, 288, 710, 711, 716, 719, 720, 721, 722, 2348]