Rodger (Jieliang) Luo
MAT 259
15 January 2015

Project 1: Forgotten Books

Project Description:

Every library needs to face an awkward situation that many books have been kept by audience way later than they are supposed to be. Can we just simply assume that's because of people's bad memory? Is it possible that there is a deeper connection between the types of books and their delayed returned date? Can we find a reasonable explanation for this phenomenon? In this project, the books which have been checked out for more than 3 years will be examined to see if any interesting patterns will be found.

Queries/Analysis:

1. How many books have been checked out for more than three years?

Query: SELECT * FROM spl2.inraw WHERE TIMESTAMPDIFF(YEAR, cout, cin) >= 3
Result: 1,008,577
Duration: 71.653 secs
Analysis: There are 1,008,577 books have been kept by for at least 3 years by audiences, which is obviously beyond the due date. Since the number is incredibly huge, a simple explanation that people accidentally forgot to return is not convincing.

2. Regarding of years, what is the distribution of these 'forgotten books'?

Query:
```
SELECT
        TIMESTAMPDIFF(YEAR, cout, cin) AS Years,
        COUNT(*) AS 'Books'
FROM
        spl2.inraw
WHERE
        TIMESTAMPDIFF(YEAR, cout, cin) >= 3
GROUP BY TIMESTAMPDIFF(YEAR, cout, cin)
```

Result:

Years	Books
3	32324
4	15751
5	9000
6	5560
7	3831
8	2740
9	2122
10	1610
11	1271
12	997
13	729
14	540
15	400
16	335
17	211
18	166
19	89
20	54
21	27
22	8
23	5
36	117235
37	129490
38	207238
39	197387
40	172382
41	105903
42	448
43	716
44	8

Duration: 58.018 secs
Analysis: The statistics look reasonable from 3 years to 23 years that the numbers of books are descending due to the increase of years. However, after 23 years, there is a 13 years gap, and then the numbers of books increased tremendously. What happened behind the anomalous numbers?

3. In the chart above, what are the reasons to cause the gap and unreasonable increased number of books?

Query: SELECT cout, cin FROM spl2.inraw WHERE TIMESTAMPDIFF(YEAR, cout, cin) >= 36
Result (Partial):

cout	cin
1970-01-01 00:00:00	2006-01-02 08:53:00
1970-01-01 00:00:00	2006-01-02 09:46:00
1970-01-01 00:00:00	2006-01-02 10:16:00
1970-01-01 00:00:00	2006-01-02 10:20:00
1970-01-01 00:00:00	2006-01-02 10:25:00
1970-01-01 00:00:00	2006-01-02 10:32:00
1970-01-01 00:00:00	2006-01-02 10:35:00
1970-01-01 00:00:00	2006-01-02 10:36:00
1970-01-01 00:00:00	2006-01-02 10:36:00
1970-01-01 00:00:00	2006-01-02 10:36:00
1970-01-01 00:00:00	2006-01-02 10:46:00
1970-01-01 00:00:00	2006-01-02 10:52:00

Duration: 67.794 secs
Analysis: For all the books having been checked out for more 36 years, they share the same checking out time: 1970-01-01. It might be a system error. Or it might be all those books were checked out before the library setup the digital database, so once the digital database was setup, they were all considered as the same checking out time. In that case, it looks that people were more likely to keep the books before the digital database implemented into the library. Interestingly, they somehow returned the books after 36 or more years.

4. What are the most popular type of books that people like to keep?

Query: SELECT
TIMESTAMPDIFF(YEAR, cout, cin) AS Years,
FLOOR(deweyClass / 100) * 100 AS Dewey,
COUNT(*) AS 'Books'
FROM
spl2.inraw
WHERE
TIMESTAMPDIFF(YEAR, cout, cin) >= 3
GROUP BY FLOOR(deweyClass / 100) * 100 , TIMESTAMPDIFF(YEAR, cout, cin)
ORDER BY Books DESC
Result (Partial):

Years	Dewey	Books
38	0	113044
39	0	105096
40	0	93836
37	0	65603
36	0	59496
41	0	57187
38	700	47667
39	700	45387
40	700	36302
37	700	31696
36	700	27046
41	700	22756
39	600	15723
38	600	15287
40	600	14621

Duration: 59.590 secs
Analysis: Without considering books having no dewey numbers, it looks like arts (700) and technology (600) are the two most popular types that people want to keep at their home before 1970s.

5. If we consider the years longer than 36 as noise, what are the most popular type of books that people like to keep without the noise?

Query: SELECT
TIMESTAMPDIFF(YEAR, cout, cin) AS Years,
FLOOR(deweyClass / 100) * 100 AS Dewey,
COUNT(*) AS 'Books'

```
FROM
        spl2.inraw
WHERE
        DATE(cout) > '1970-01-01'
        AND TIMESTAMPDIFF(YEAR, cout, cin) >= 3
GROUP BY FLOOR(deweyClass / 100) * 100 , TIMESTAMPDIFF(YEAR, cout, cin)
ORDER BY Books DESC
```

Result:

Years	Dewey	Books
3	0	11484
4	0	5278
3	700	4717
3	300	3759
3	900	3599
3	800	3201
5	0	2917
3	600	2574
4	700	2289
4	900	1955
4	300	1856
4	800	1854
6	0	1750
5	700	1372
3	500	1364

Duration: 74.185 secs
Analysis: Without the noise, the result seems more random rather than a clear consequence. However, many books on the top of the lists, both in 4 and 5, don't a dewey number. So a further research needs to check the bibNumber to see if any pattern can be revealed.