Olaf Menzer,
MAT 259 Project 1 – Data Query

Empirical hours of a library

Query description

The comprehensive data base of check-outs from the Seattle Public Library between 2005 and 2014 covers a wide range of media types (e.g., books, DVDs, records) and subjects (e.g., Philosophy, Computer Science, Language). Based on more than 67 million entries one can study behavioral questions related to its users and variability of apparent patterns over time. While some patterns are expected to be found on a global data base level, other patterns may only be apparent when specifically selecting a subset of media types, subjects or a combination of both.

Here, I am posing the question of how long the library is *awake* every day in the study period in terms of the data base (empirical approach). The term *awake* will first be defined as the difference of the first check-out of the day and the last check-out of each day (in minutes) on a global data base level covering all media types and subjects. One would expect the time of the first and last check out of a day to correlate well with the actual opening hours of the library, which is the first question of my study. In the following are the current opening hours of the library that I pulled from the website:

```
Mon - Thu:      10am – 8pm
Fri - Sat:      10am – 6pm
Sun:            12pm – 6pm
```

One can easily calculate the time difference between first and last check-out of a day, to test if the opening hours are undermined well by actual observed check-outs. Specific trends in the data that are related to the day of the week can be identified, for example, some days of the week may be less *awake* than other days of the week when averaging *awake* hours over the whole study period by weekday. A study from McKenzie (2014) that uses the same data set, suggests that there is a trend in number of check-outs and their variability by media subject over the course of a typical week, with Monday having the highest number of check-outs and a steady decline throughout the course of the week. The question of when the library is more or less awake could also be further specialized to media types or media subjects to study behavioral patterns of groups of people that favor e.g., “Phineas and Ferb” comic books or records from “The Beatles”. Based on my query, one can in turn also ask the question of when the library *sleeps*, and search for media types or subjects that have not been rented out over the longest time in the data base. This could potentially lead to conclusions of which selections in the library should be reduced or extended in the future.

SQL Queries

Query 1:

```
SELECT title, ItemNumber, barcode, itemtype,
    DATE_FORMAT(cout, '%Y-%m-%d') as day_cout,
     DAYOFWEEK(cout), DAYOFMONTH(cout),
     MIN(cout) as earliest_cout,
     MAX(cout) as latest_cout,
     DATE_FORMAT(MIN(cout), '%H') as earliest_cout_hour,
     DATE_FORMAT(MAX(cout), '%H') as latest_cout_hour,
     TIMESTAMPDIFF(MINUTE, MIN(cout), MAX(cout)) as awake
FROM spl2.inraw
GROUP BY day_cout
```

The runtime for this query was 209 seconds, it returns 7499 entries in total when run without any further WHERE clauses. However, the first 4344 entries did not contain hourly information in their timestamp, so I discarded them and refined my query accordingly to start on 2005-02-23, which returned 3154 this time. I noticed a lot of check-out dates were at midnight, probably due to missing information on the actual checkout date. Therefore, I changed my query to ignore those entries (WHERE TIME_TO_SEC(cout) >0) when searching for earliest and latest check-outs of a day:

Query 2:

```
SELECT  barcode, itemtype,deweyClass, cout,
    DATE_FORMAT(cout, '%Y-%m-%d') as day_cout,
    DAYNAME(cout) as weekday,
    DAYOFWEEK(cout) as weekday_num,
    DAYOFYEAR(cout) as doy,
    MIN(cout) as earliest_cout,
    MAX(cout) as latest_cout,
    DATE_FORMAT(MIN(cout), '%H') as earliest_cout_hour,
    DATE_FORMAT(MAX(cout), '%H') as latest_cout_hour,
    TIMESTAMPDIFF(MINUTE, MIN(cout),MAX(cout)) as awake
FROM
    (SELECT barcode, itemtype, cout, cin, deweyClass
    FROM spl2.inraw
    WHERE TIME_TO_SEC(cout) >  0 and deweyClass <> '') as db1
WHERE
    date(cout) >= '2005-02-23' and date(cout) <= '2014-01-05'
GROUP BY day_cout
ORDER BY weekday_num
LIMIT 10000;
```

While Query 1 is a relatively simple query that is based on grouping the data by the day and returning the minimum and maximum of `cout` for each day, Query 2 relies on a nested (hierarchical) query design. In the inner call of the query, I selected all records that have not been made at hour 0 and minute 0 of a given day and also only targets records for which a `deweyClass` exists.
This query took about 191 seconds.

With the following query 2.1:

```
SELECT weekday, AVG(awake) as awake_mean
FROM (...(QUERY 2)*...) as db2
GROUP BY weekday_num
```

I computed the average *awake* time for each day of the week. Note that the block of Query 2 was omitted here for brevity. This query took about 272 seconds.

Another interesting question is to look at days of the year when the library was least awake throughout the year (Query 2.2), which can be done by ORDERing the results BY `awake` instead of `weekday_num` in Query 2. This query took about 241 seconds.

Query 3:
In query 3 I will extend the query above to filter out check-outs that were actually online renewals and not corresponding to check-outs associated with persons being present at the library. Currently, I did not find the check out times after opening hours to match with any check in times at the same day and will continue to search for another filtering method or corresponding flag in the data base. This can likely also provide some insight into how online renewals are captured in the data base and what type of information they represent compared to loans where a person is physically present.

Results and Discussion
The library is more awake during weekdays than during the weekend, with Wednesday being the busiest day with an active period of 12 hours and 59 minutes on average (Table 1). On Sundays, the library is only awake for 08 hours and 42 minutes on average.

weekday	awake [minutes]	awake [hh:mm]
Sunday	522	08:42
Monday	737	12:17
Tuesday	770	12:40
Wednesday	779	12:59
Thursday	767	12:47
Friday	711	11:51
Saturday	602	10:02

Table 1: Query 2.1 lead to the following average awake times shown for each day of the week.

These values generally correlate well with the actual opening hours of the library (Pearson correlation

coefficient r = 0.93), but are likely off by a few hours because they are not accounting for media renewals that can be made online. This might also be the reason why they are not in agreement with the general weekly trend that McKenzie (2014) found in his analysis. Computing another SQL query (Query 3) that filters out such check-out dates will allow for re-testing of that hypothesis and also give a more accurate description of awake hours at smaller scales of the data set.

Another interesting question was to look at the 50 days of the year when the library was least awake throughout the study period (Table 2).

	day_cout	weekday	weekday_num	earliest_cout_hour	latest_cout_hour	awake
1	2009-04-12	Sunday	1	16	16	0
2	2010-01-18	Monday	2	11	11	0
3	2009-05-25	Monday	2	12	12	0
4	2008-05-26	Monday	2	14	14	0
5	2006-09-04	Monday	2	10	10	0
6	2007-05-28	Monday	2	10	10	0
7	2008-09-01	Monday	2	17	17	0
8	2011-02-21	Monday	2	22	22	0
9	2008-01-21	Monday	2	17	17	0
10	2006-07-04	Tuesday	3	4	4	0
11	2011-09-01	Thursday	5	18	18	0
12	2010-01-01	Friday	6	15	15	0
13	2008-07-04	Friday	6	15	15	0
14	2005-12-24	Saturday	7	11	11	0
15	2009-07-04	Saturday	7	17	17	0
16	2007-07-04	Wednesday	4	12	12	1
17	2006-02-20	Monday	2	16	16	7
18	2007-02-19	Monday	2	15	17	95
19	2006-01-16	Monday	2	14	16	130
20	2007-11-22	Thursday	5	10	12	133
21	2013-01-21	Monday	2	12	14	139
22	2009-02-16	Monday	2	12	15	190
23	2006-11-11	Saturday	7	14	17	209
24	2005-03-20	Sunday	1	13	17	220
25	2005-07-24	Sunday	1	13	16	220
26	2005-05-01	Sunday	1	13	16	221
27	2007-09-03	Monday	2	12	16	221
28	2005-06-19	Sunday	1	13	16	224
29	2005-04-03	Sunday	1	13	17	225
30	2007-04-08	Sunday	1	12	16	229
31	2005-06-05	Sunday	1	13	17	229
32	2005-03-06	Sunday	1	13	17	230
33	2005-09-11	Sunday	1	13	16	231
34	2005-08-28	Sunday	1	13	16	233
35	2005-04-24	Sunday	1	13	16	233
36	2005-06-12	Sunday	1	13	17	233
37	2005-10-23	Sunday	1	13	17	233
38	2005-08-07	Sunday	1	13	16	234
39	2005-07-03	Sunday	1	13	16	234
40	2005-09-25	Sunday	1	13	16	234
41	2005-11-20	Sunday	1	13	17	234
42	2005-04-10	Sunday	1	13	16	234
43	2005-09-18	Sunday	1	13	16	235
44	2005-07-10	Sunday	1	13	16	235
45	2005-02-27	Sunday	1	13	17	236
46	2005-06-26	Sunday	1	13	16	236
47	2005-05-22	Sunday	1	13	16	236
48	2005-07-17	Sunday	1	13	17	237
49	2005-08-21	Sunday	1	13	16	237
50	2005-04-17	Sunday	1	13	17	239

Table 2: First 50 results of query 2.2 showing days where the library is least awake.

There are 15 entries where the library was awake for 0 hours which all corresponded to holidays including Memorial Day, Martin Luther King Day and President's day. A manual search showed that the first 21 entries were all holidays. In contrast, on Independence Day in the year 2011 (2011-07-04) the library was awake for 291 hours and one can only speculate if this was caused through online renewals again or if a librarian was actually present. Table 2 also shows that Sunday is consistently the least awake day of the week which indicates that the data support common patterns of human behavior.