MAT259 Assignment 1 - Interesting Query

Donghao Ren
01/14/2015

One interesting question about this dataset is to see the difference between the check-out and check-in time (lending time), which can reveal how long people read the books to some extent. Although people may not read the book during the entire lending time, but this number is the only available measure of reading time in the database. In this assignment, I explored the statistics of the lending times of top-level Dewey classes using SQL queries.

Query

```
SELECT
  SUBSTRING(deweyClass, 1, 2) AS class,
  (SELECT COUNT(*) FROM spl2.dewey WHERE SUBSTRING(spl2.dewey.dewey, 1, 2) = class) AS class_count,
  COUNT(*) AS activity_count,
  AVG(TIMESTAMPDIFF(DAY, cout, cin)) AS average_lending_time,
  SQRT(VARIANCE(TIMESTAMPDIFF(DAY, cout, cin))) AS std_lending_time,
  MAX(TIMESTAMPDIFF(DAY, cout, cin)) AS maximum_lending_time
FROM
  spl2.inraw
WHERE
  cout != "1970-01-01 00:00:00"
  AND itemtype LIKE "%bk"
GROUP BY
  class
```

Query Time

248.811 seconds (note that the query results may be cached, so if you run it twice, the second time might return instantly).

Explanation

The granularity of the results is very important. In this assignment, I chose to group the records by the first two digits of the Dewey class, yielding around 100 rows in the query result, which is not too general and not too specific.

For the columns, I chose to return the dewey class, the number of books in that class, the number of check-in/check-out activities, and the average, standard deviation, and maximum of the lending times. The result set contains 6 columns.

Result

See the dewey-count-average_lending_time.csv file attached.

Basic Analysis

First, it's interesting that the average lending time of all classes are greater than the 21 days loan period stated in http://www.spl.org/using-the-library/get-started/check-out-and-return. The policy says most items can be renewed twice, so I believe that the total loan time must be smaller than 63 days, however, it's not the case in the database. Some Dewey classes' average lending time is even more than 100 days.

MAT259 Interesting Query

I explored the Dewey classes with the minimum/maximum average lending times. The classes with smallest lending times are:

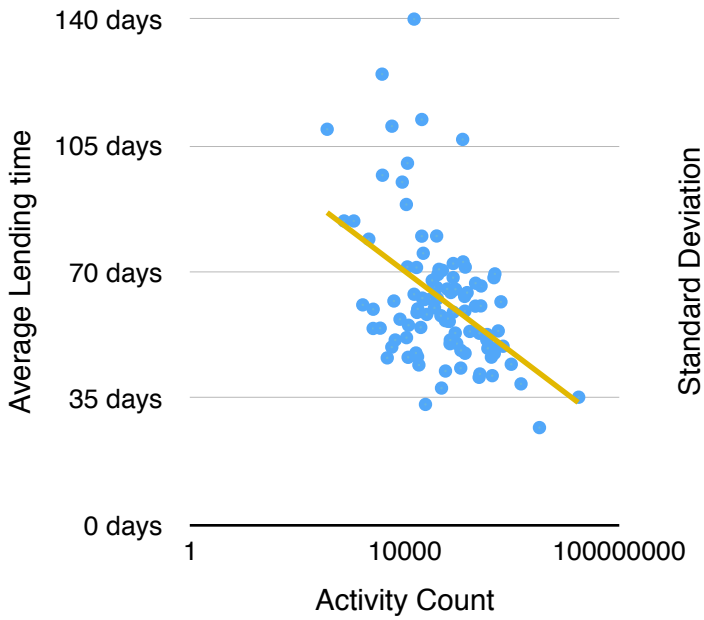
- 740 Graphic arts & decorative arts
- 030 Encyclopedias & books of facts
- 560 Fossils & prehistoric life
- 640 Home & family management
- 000 Computer science, knowledge & systems

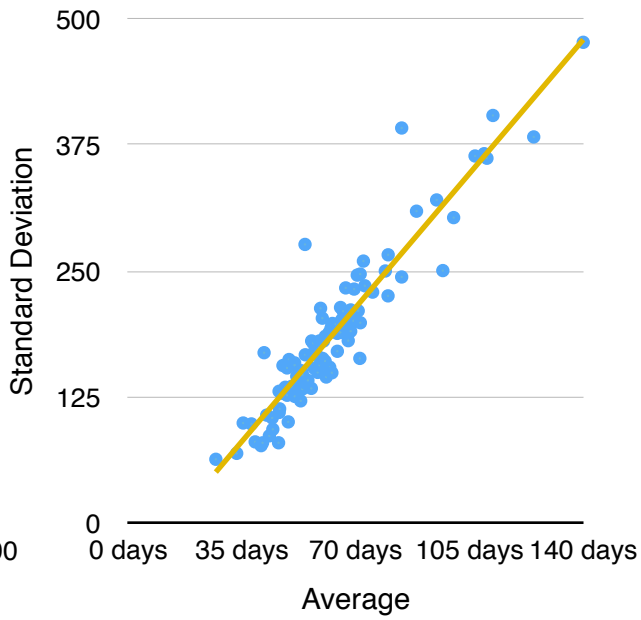
The classes with largest lending times are:

- 830 German & related literatures
- 870 Latin & Italic literatures
- 840 French & related literatures
- 850 Italian, Romanian, & related literatures
- 090 Manuscripts & rare books

They are mostly literatures, which I believe needs more time to read through. Among the literatures, German books takes more time to read, then Latin and Italic, French and Italian.

I also tried to visualize the correlation between several variables in the resulting table.





The left graph shows the relationship between activity count (number of check-in/check-outs) and the average lending time. The overall trend is the more activities there are, the less the average lending time.

The right graph shows the relationship between average lending time and its standard deviation, these two columns has a very strong linear relationship. The two outliers are 050 Magazines, journals & serials and 310 Statistics, which have a higher variance than the general trend.

As a note for further expansion this idea, we can relate the average lending time with a measure of the length of the book, say number of pages (which currently not available in the dataset), in this way we can measure “reading speed”.

Update (for Assignment 2): Query 2 - Yearly Results

```
SELECT class, GROUP_CONCAT(CAST(year AS CHAR(40)) ORDER BY year) AS years,
GROUP_CONCAT(CAST(average_lending_time AS CHAR(40)) ORDER BY year) AS average_lending_times
FROM (
  SELECT
    SUBSTRING(deweyClass, 1, 2) AS class,
    YEAR(cout) AS year,
    AVG(TIMESTAMPDIFF(DAY, cout, cin)) AS average_lending_time
  FROM
    spl2.inraw
  WHERE
    cout >= "2006" AND cout < "2014"
        AND itemtype LIKE "%bk"
    AND deweyClass != ""
  GROUP BY
    class, year
) AS innerTable
GROUP BY
  class
ORDER BY
  class ASC;
```

Result

See the dewey-count-average_lending_time-yearly.csv file attached.

Query Time

129.254 seconds.

Analysis

In this query, we group the average lending time by year and dewey class. Please refer to the next assignment for my visualization of this result.