MAT 259 – Winter 2016 – Project 1

May ElSherif

Question: For non-fictional items, is there a correlation between the number of subject entries (labels/tags) that describe an item in the Seattle Public Library and the number of checkouts? Does it differ across different Dewey classes?

Example:

As shown in Fig.1, the item with bibNumber = 3 has only one entry that describes it while the item with bibNumber = 7 has more than one entry. Could this discrepancy lead to a difference in the number of checkouts as well?

bibNumber	subject
3	Naturalists Biography
6	Marx Brothers
7	Authors Chinese 20th century Biography
7	Authors English 20th century Biography
7	China Description and travel
7	China History 20th century

Figure 1. Different items with a different number of entries in the subject table.

Labeling and tagging items with keywords are important aspects in data organization. It is used to create search indexes that help users, especially on the web retrieve the information they need.

Methodology:

To tackle the aforementioned question, I began by investigating the deweyClass table, the x_checkOutCountBib, and the subject table. To

acquire the data in the needed form to answer the question, a triple join was needed between the previously mentioned tables.

```
SELECT
    deweyClass AS Dewey,
    spl3.deweyClass.bibNumber,
    checkOutCount,
    COUNT(spl3.subject.bibNumber) AS SubjectEntriesCount
FROM
    spl3.x_checkOutCountBib,
    spl3.deweyClass,
    spl3.subject
WHERE
    deweyClass > 0
        AND spl3.x_checkOutCountBib.bibNumber =
spl3.deweyClass.bibNumber
        AND spl3.x_checkOutCountBib.bibNumber = spl3.subject.bibNumber
        AND (spl3.subject.subject != ‘’
        OR spl3.subject.subject IS NOT NULL)
GROUP BY bibNumber
ORDER BY deweyClass;
```

Query Explanation:

The query captures for every bibNumber, the number of subject entries that describe this item and its deweyClass through a triple join. The deweyClass > 0 eliminates fictional items. The conditions spl3.x_checkOutCountBib.bibNumber = spl3.deweyClass.bibNumber

AND spl3.x_checkOutCountBib.bibNumber = spl3.subject.bibNumber

join the three tables. The important condition is that to check for empty string subjects or null subject which is captured by the last condition spl3.subject.subject != ‘’ OR spl3.subject.subject IS NOT NULL. This last condition provoked me to do a separate analysis for items with bibNumbers that has no subject entries.

Results:

Figure 2 depicts a sample of the output CSV file which shows for each bibNumber: the checkOutCount, the number of subject entries and its Dewey classification. To better understand these numbers, a scatter plot that incorporates the number of subject entries on the x-axis, the number of checkouts on the y-axis and the Dewey classification grouped into 10 classes is shown in Figure 3.

Processing time:

For the previously described query, the result contained 476936 rows, the duration was 54.448 seconds and the fetch time was 6.249 seconds.

Dewey	bibNumber	checkOutCou	SubjectEntriesCount
1	13703	2	2
1	24571	18	1
1	28596	4	1
1	37304	14	1
1	54681	8	3
1	64726	4	2
1	93270	2	1
1	93859	1	1
1	94976	4	1
1	95354	1	2
1	98459	5	1
1	99264	4	2
1	103880	9	4
1	105526	3	1

Figure 2. A sample of the output CSV file.

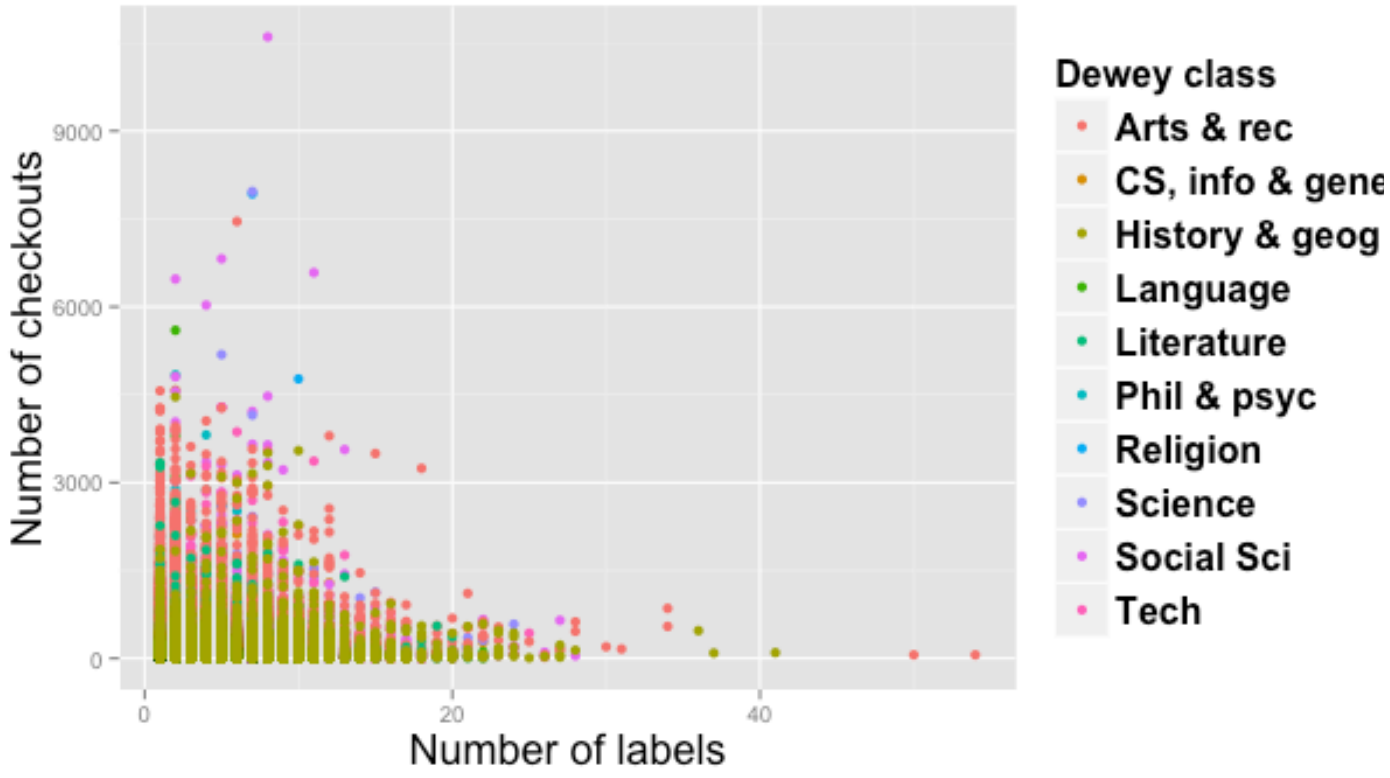

Figure 3. Number of checkouts vs. number of subject entries.

Results explanation and further analysis:

As we can see from Figure 3, there is a trend of a decreasing number of checkouts when the number of labels increases. Although many aspects affect the number of checkouts, this result can be interpreted in the sense that people in the Seattle public library tend to check out very specific books. Since a generic book incorporates multiple topics, the probability of this book having more than one subject entry increases. A specialty book (i.e. a book that focuses on fewer points) has the probability of having fewer labels. The above result suggests that people tend to check specialty books more than generic books. On the other hand, the question still remains: "***What about unlabeled items?***" (i.e. items without subject entries).

I was interested to know how many entries in the subject table contain an empty subject (select * from spl3.subject where subject = ' ') and the number of rows was 134, 966 unlabeled items. From these entries, most of them belonged to Dewey Classes from 800 to 900 (as shown in Figure. 4) which suggests that the majority of unlabeled items are Literature items.

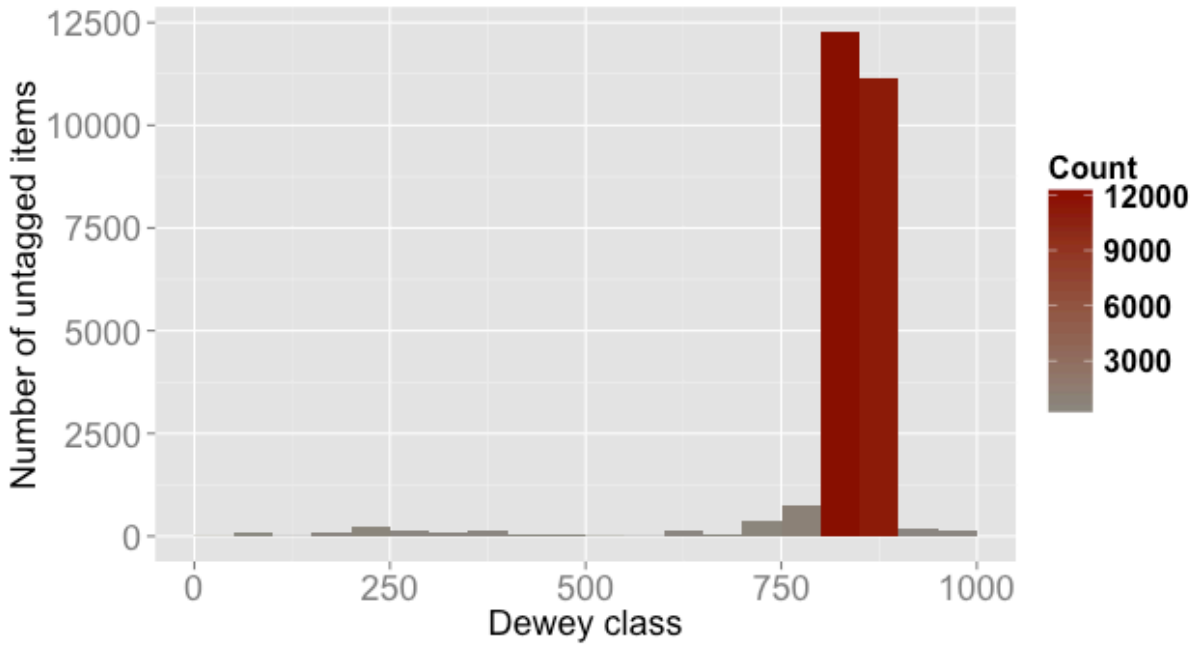

Figure 4. Histogram of Dewey class for untagged items.

Another question that seems compelling was "***Do unlabeled items tend to have lower or higher checkout rate?***"

To answer this question, I constructed another query to check for entries with empty string subject entries in the subject table as follows.

```
SELECT
    spl3.deweyClass.bibNumber,
    deweyClass AS Dewey,
    checkOutCount
FROM
    spl3.x_checkOutCountBib,
    spl3.deweyClass,
    spl3.subject
WHERE
    deweyClass > 0
        AND spl3.x_checkOutCountBib.bibNumber =
spl3.deweyClass.bibNumber
        AND spl3.x_checkOutCountBib.bibNumber = spl3.subject.bibNumber
        AND spl3.subject.subject = ''
ORDER BY deweyClass;
```

This query returned 25, 933 rows which suggests that there are 109, 027 entries with zero check out rate. The checkout rate of the rest of the entries is depicted in Figure 5.

If we compare Figure 3 and Figure 4, we can see that the rate of checkout for labeled items is on average higher than the rate of checkout for unlabeled items. This indicates that labeling an item could result in higher checkout rates for that item.

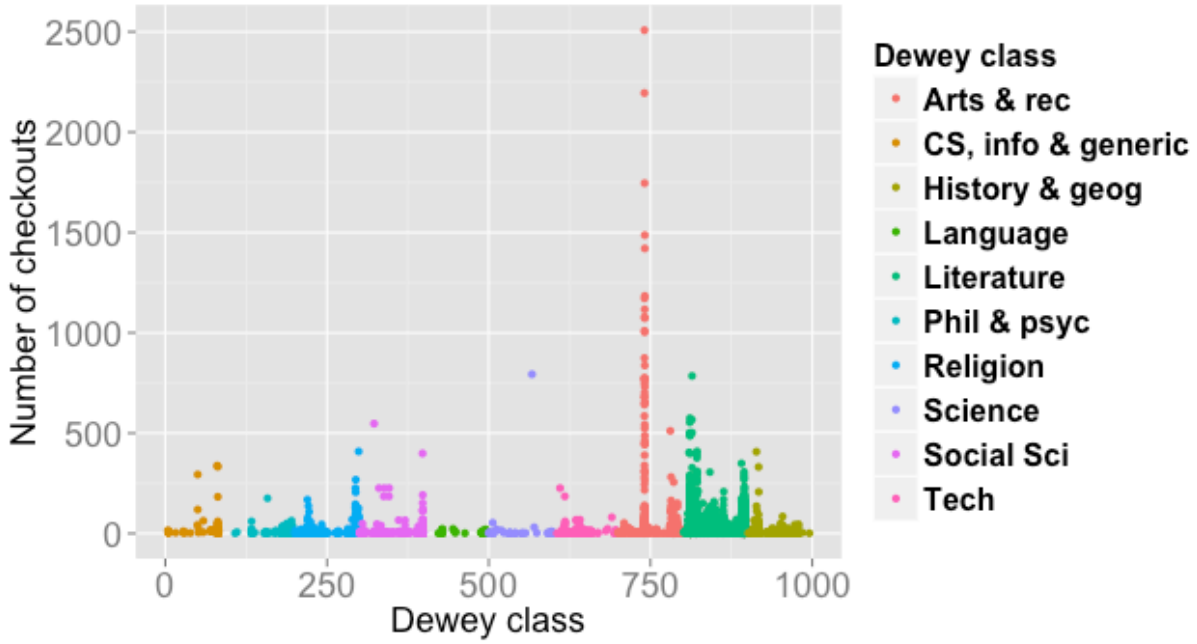

Figure 5. Checkout rate for unlabeled items.