MAT259 Project1
Bo Yan

Question:

People borrow books for different reasons, some of them wish to read the best-selling novels, others wish to learn programming skills. In other words, people have their personal preference for certain kinds of books. The specific reading interests may influence the checkout duration. It would be interesting to see if there is a pattern between checkout duration and reading interests. In this database, the reading interests could be represented by the particular dewey classes to which the borrowed books belong.

Because not all of the books in the library catalog have dewey number, the reading interests represented here are limited to those dewey classes and are only a subset of real situation. My exploration also only considers the item type book, since for other media such as DVD, people could just make a copy of it and return it very quickly.

Data Exploration:

I tried to use the whole dataset and use the standard deviation and average of the duration of each of the 10 dewey classes, but the query results does not make sense. Below is my SQL:

```
SELECT
    FLOOR(deweyClass / 100) * 100 AS Dewey,
    AVG(TIMESTAMPDIFF(DAY, cout, cin)) AS TimeAvg,
    STDDEV(TIMESTAMPDIFF(DAY, cout, cin)) AS TimeStdd
FROM
```

```
    spl2.inraw
WHERE
    itemtype = 'acbk'
GROUP BY FLOOR(deweyClass / 100) * 100
```

The results I got are:

Dewey	TimeAvg	TimeStdd
0	330.8147	2020.0455
100	270.0712	1778.5759
200	293.799	1819.7638
300	375.7069	2114.5903
400	277.9219	1761.3172
500	296.5959	1826.958
600	301.7749	1898.5511
700	321.4	1984.3802
800	272.0688	1696.4141
900	296.5808	1839.302

So it looks like the average duration is around 300 days! It’s not possible. There must be something wrong with the data. So I queried the individual duration using the SQL below:

```
SELECT
    TIMESTAMPDIFF(DAY, cout, cin) AS Duration,
    cout,
    cin,
```

```
    deweyClass,
    title
FROM
    spl2.inraw
WHERE
    itemtype = 'acbk'
```

Then I found that there are some anomalies in the database:

Duration	cout	cin	deweyClass
13150	1970-01-01 00:00:00	2006-01-02 09:46:00	973.93109

some of the check out time in this database date way back to 1970, making the average of the duration time much larger. So I decided to limit the checkout year within the range of 2006 and 2014 (inclusive).

SQL Queries:

```
SELECT
    FLOOR(deweyClass / 100) * 100 AS Dewey,
    AVG(TIMESTAMPDIFF(DAY, cout, cin)) AS TimeAvg,
    STDDEV(TIMESTAMPDIFF(DAY, cout, cin)) AS TimeStdd
FROM
    spl2.inraw
WHERE
    itemtype = 'acbk' AND YEAR(cout) >= 2006
        AND YEAR(cout) <= 2014
        AND deweyClass <> ''
GROUP BY FLOOR(deweyClass / 100) * 100
```

Processing time: 83.211 seconds.

Query Explanation:

I group the dataset by 10 dewey classes, for each class, average and standard deviation of duration is calculated. Because sometimes the average is misleading, the standard deviation will give us an idea of how spread out the duration will be across the dataset.

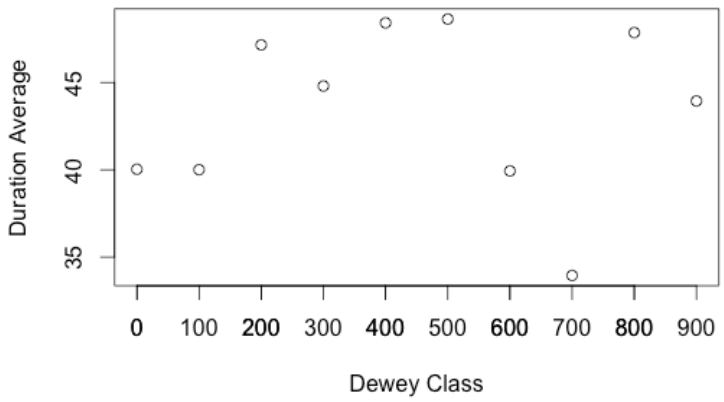
Results:

Dewey	TimeAvg	TimeStdd
0	40.0354	62.1272
100	40.0061	62.6159
200	47.156	86.0574
300	44.8015	83.4172
400	48.418	83.7641
500	48.6377	90.6569
600	39.9371	63.2997
700	33.9483	66.7933
800	47.8595	102.2354
900	43.9421	86.9274

Analysis:

We can find some interesting results. The dewey class 700 (Arts and Recreation) has the lowest duration time while classes such as 200 (Religion), 400 (Language), 500 (Science) and 800 (Literature) have relatively high duration time. One explanation could be that Arts and

Recreation books are less intensive and people tend to spend less effort and time to read those book. On the other hand, books in Science for example, are more intensive thus people are really spending more time to read those books. Actually Science books have the highest duration time.



Another interesting finding from the duration standard deviation plot is that the duration standard deviation for 800 (Literature) is very high (actually the highest), which means that the duration times within the 800 group differ a lot. There might be very short duration times and very long durations times within the 800 group. This would imply that different people spend different times to read the Literature books and it would be very interesting to find which are the books that take more duration times.