Sandy Schoettler Project 2 Report

February 5, 2019

```
In [2]: %%html
        <style>
        table td, table th, table tr {text-align:left !important;}
        table {float:left}
        </style>
```

```
<IPython.core.display.HTML object>
```

1 Motivating Question

What can we learn about the distribution of book checkouts in time over the course of a week?

1.1 Problems with data

1.1.1 First Problem

We notice that there are a very large number of checkouts in 1970 as compared to nearby years. This is the UNIX ephoc time and reflects a date stamp of 0 in the database. We should discard these because the checkout time does not reflect reality. The most data activity is past 2005 so we'll start most queries here. But it's important to exclude 1970.

Total checkouts by Year:

```
SELECT Year(checkout) AS year,
       Count(*)
FROM   transactions
GROUP  BY year;
```

Year	Checkouts
1970	936295
1987	4
1988	84
1989	119
1990	172
1991	204
1992	292

Year	Checkouts
1993	519
1994	738
1995	880
1996	1053
1997	1426
1998	2001
1999	2703
2000	4023
2001	4470
2002	7970
2003	15758
2004	38982
2005	543324
2006	13029176
2007	13769669
2008	18427707
2009	19813291
2010	18535749
2011	16027444
2012	12969344
2013	13939772
2014	13080441
2015	12285802
2016	11418873
2017	10200509
2018	7442800
2019	803889

1.1.2 Second Problem

We notice that when looking at the time of day, many checkout times are at 12:00 AM. This does not match intuition because checkouts should be most common while the library is open. So we should discard these records as well.

Checkouts by Hour from 1990 to present

```
SELECT Hour(sample.checkout) AS hr,
       Count(*)
FROM   (SELECT checkout
        FROM   transactions
        WHERE  checkout >= '1990-01-01') sample
GROUP  BY hr;
```

hour	Checkouts
0	102863
1	436

hour	Checkouts
2	405
3	81
4	492
5	852
6	2248
7	26475
8	384043
9	617853
10	9196210
11	14524855
12	15952591
13	22242891
14	23524377
15	24859556
16	26992550
17	23490490
18	10684293
19	9649194
20	102756
21	2201
22	3187
23	8082

Notice that the entry for 0 is relatively large. There seem to be a lot of midnight checkouts. We can examine how the number of midnight checkouts has compares to a more recent sample:

Checkouts by Hour from 2015 to present:

```
SELECT Hour(sample.checkout) AS hr,
       Count(*)
FROM   (SELECT checkout
        FROM   transactions
        WHERE  checkout >= '2005-01-01') sample
GROUP  BY hr;
```

hour	Checkouts
0	21673
1	436
2	405
3	81
4	492
5	852
6	2248
7	26475
8	384043
9	617852

hour	Checkouts
10	9196210
11	14524855
12	15952591
13	22242891
14	23524377
15	24859556
16	26992550
17	23490490
18	10684293
19	9649194
20	102756
21	2201
22	3187
23	8082

1.1.3 Third Problem

These tables are almost exactly the same!

Let's take a closer look at how many yearly checkouts are logged at midnight.

Analysis of Midnight Checkouts from 1990 to present

```
SELECT
    YEAR(checkout) AS yr,
    COUNT(checkout),
    SUM(CASE WHEN HOUR(checkout) = 0 THEN 1 ELSE 0 END)
FROM transactions
WHERE checkout > '1990-01-01'
GROUP BY YEAR(checkout);
```

Results

Year	Total Checkouts	Midnight Checkouts
1990	172	172
1991	204	204
1992	292	292
1993	519	519
1994	738	738
1995	880	880
1996	1053	1053
1997	1426	1426
1998	2001	2001
1999	2703	2703
2000	4023	4023
2001	4470	4469
2002	7970	7970
2003	15758	15758

Year	Total Checkouts	Midnight Checkouts
2004	38982	38982
2005	543324	19164
2006	13029176	1635
2007	13769669	754
2008	18427707	80
2009	19813291	4
2010	18535749	25
2011	16027444	6
2012	12969344	2
2013	13939772	0
2014	13080441	3
2015	12285802	0
2016	11418873	0
2017	10200509	0
2018	7442800	0
2019	803889	0

1.1.4 Conclusion

We should only consider data from 2007 to present in analyzing any timestamp information, or alternatively exclude timestamps which are at midnight.

1.2 Components

Checkouts by Weekday

```
SELECT Weekday(sample.checkout) AS day,
       Count(*)
FROM   (SELECT checkout
        FROM   transactions
        WHERE  checkout >= '1990-01-01'
        LIMIT 10000) sample
GROUP  BY day;
```

Results

Weekday	Checkouts (sample)	Checkouts (whole table)
0	1802	28016941
1	2059	29150448
2	1847	29848679
3	1823	27792077
4	1387	22149467
5	799	30114174
6	283	15297195

(Monday is 0, weekends are days 5 & 6)

Checkouts by weekday & hour

```
USE spl_2016;

SELECT Count(*) AS qty, day, hr
FROM   (SELECT checkout,
               Day(checkout)  AS day,
               Hour(checkout) AS hr
        FROM   transactions
        WHERE  checkout > '1990-01-01'
        LIMIT  1000000) sample
WHERE  hr != 0
GROUP  BY day, hr;
```

The query returned the following table showing the number of book checkouts grouped by day of the week, and the hour of the day the checkout happened.

Checkouts	Day of week	Hour of day
2	1	6
12	1	7
54	1	8
252	1	9
802	1	10
1083	1	11
912	1	12
...	...	...

Checkouts by Hour & weekday in 2D

```
SELECT
   hr,
   SUM(CASE WHEN day=0 THEN qty ELSE 0 end) Monday,
   SUM(CASE WHEN day=1 THEN qty ELSE 0 end) Tuesday,
   SUM(CASE WHEN day=2 THEN qty ELSE 0 end) Wednesday,
   SUM(CASE WHEN day=3 THEN qty ELSE 0 end) Thursday,
   SUM(CASE WHEN day=4 THEN qty ELSE 0 end) Friday,
   SUM(CASE WHEN day=5 THEN qty ELSE 0 end) Saturday,
   SUM(CASE WHEN day=6 THEN qty ELSE 0 end) Sunday
FROM   (SELECT Count(*) AS qty,
               day,
               hr
        FROM   (SELECT checkout,
                       Weekday(checkout) AS day,
                       Hour(checkout)    AS hr
                FROM   transactions
                WHERE  checkout > '1990-01-01'
                LIMIT  1000000) sample
```

```
        WHERE  hr != 0
        GROUP  BY day, hr)
        AS hourly
GROUP  BY hr;
```

hr	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
2	0	0	4	0	2	4	0
3	0	0	0	7	1	0	0
4	0	0	1	0	0	0	0
5	0	0	0	6	2	0	0
6	0	0	0	2	0	0	0
7	109	227	83	99	49	0	2
8	937	1383	991	867	790	4	5
9	998	1500	1294	1504	755	92	16
10	2507	3232	9793	9660	10833	8616	85
11	3399	3117	14898	14454	15298	12487	21
12	3663	4549	16697	14316	17190	13467	18
13	16456	18632	17025	15990	16862	14612	6583
14	19996	19491	20247	17805	15976	15842	8335
15	20884	22581	22106	19530	18395	17071	7781
16	21980	23537	22225	19706	19207	14369	10309
17	19156	23397	21559	25296	21740	15428	227
18	14156	17648	17777	322	499	135	0
19	13831	17901	16210	13	17	1	0
20	172	320	211	0	4	0	0
21	0	0	4	0	0	0	0
22	3	6	0	0	0	0	0
23	5	0	0	0	0	0	0

Book Checkouts in 2006

```
SELECT
   hr,
   SUM(CASE WHEN day=0 THEN qty ELSE 0 end) Monday,
   SUM(CASE WHEN day=1 THEN qty ELSE 0 end) Tuesday,
   SUM(CASE WHEN day=2 THEN qty ELSE 0 end) Wednesday,
   SUM(CASE WHEN day=3 THEN qty ELSE 0 end) Thursday,
   SUM(CASE WHEN day=4 THEN qty ELSE 0 end) Friday,
   SUM(CASE WHEN day=5 THEN qty ELSE 0 end) Saturday,
   SUM(CASE WHEN day=6 THEN qty ELSE 0 end) Sunday
FROM   (SELECT Count(*) AS qty,
               day,
               hr
        FROM   (SELECT checkout,
                       itemtype,
                       Weekday(checkout) AS day,
```

```
                Hour(checkout)    AS hr
         FROM   transactions,
                itemType
         WHERE  checkout BETWEEN '2006-01-01' AND '2007-01-01'
                AND transactions.itemnumber = itemType.itemnumber) sample
  WHERE  hr != 0
         AND ( itemtype = 'acbk'
                OR itemtype = 'arbk'
                OR itemtype = 'bcbk'
                OR itemtype = 'drbk'
                OR itemtype = 'jcbk'
                OR itemtype = 'jrbk' )
  GROUP  BY day, hr) hourly
GROUP  BY hr;
```

hr	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1	3	2	1	6	1	0	0
2	5	3	3	4	2	13	2
3	0	0	0	0	0	0	1
4	0	21	6	18	1	0	0
5	0	7	0	5	3	0	0
6	0	11	15	1	2	4	0
7	77	212	214	56	440	13	7
8	8590	7651	6922	7498	4930	83	59
9	9228	8908	6911	7105	7194	1022	232
10	15331	18580	74400	71409	74432	77946	343
11	22973	22365	111724	114829	104054	119854	1046
12	25436	26998	108180	94497	107172	129096	46520
13	107671	125027	112343	99118	111399	131384	87096
14	118640	133812	122332	109672	123475	140085	95457
15	123489	139373	136664	119841	137642	142563	95164
16	136972	150134	150010	135653	148246	139704	110763
17	131536	144595	146790	140494	170291	130034	17064
18	98925	112922	112647	81759	3700	1960	295
19	96666	110952	109427	75498	65	3	14
20	1478	1660	1669	1080	4	0	0
21	1	2	20	31	28	0	0
22	14	10	0	4	0	0	1
23	17	20	116	108	99	0	7

2 Final Result

Expanding on these results, this SQL code collects the checkout totals in 15-minute time intervals, rather than hours. Here, it gathers information about DVDs in 2016:

This code takes about 3-5 minutes to run.

```
USE spl_2016;

SELECT
   tslice,
   SUM(CASE WHEN day=0 THEN qty ELSE 0 end) Monday,
   SUM(CASE WHEN day=1 THEN qty ELSE 0 end) Tuesday,
   SUM(CASE WHEN day=2 THEN qty ELSE 0 end) Wednesday,
   SUM(CASE WHEN day=3 THEN qty ELSE 0 end) Thursday,
   SUM(CASE WHEN day=4 THEN qty ELSE 0 end) Friday,
   SUM(CASE WHEN day=5 THEN qty ELSE 0 end) Saturday,
   SUM(CASE WHEN day=6 THEN qty ELSE 0 end) Sunday
FROM   (SELECT Count(*) AS qty,
               day,
               tslice
        FROM   (SELECT checkout,
                       itemtype,
                       Weekday(checkout) AS day,
                       floor(Hour(checkout) * 4 + (Minute(checkout)/15)) AS tslice
                FROM   transactions,
                       itemType
                WHERE  checkout BETWEEN '2016-01-01' AND '2017-01-01'
                       AND transactions.itemnumber = itemType.itemnumber) sample
        WHERE  tslice != 0
               AND ( itemtype = 'acdvd'
                      OR itemtype = 'ardvd'
                      OR itemtype = 'bcbk'
                      OR itemtype = 'bcdvd'
                      OR itemtype = 'jcdvd'
                      OR itemtype = 'jrdvd'
                      OR itemtype = 'scmed')
        GROUP  BY day, tslice) daytime
GROUP BY tslice;
```

This grid layout compares the checkout activity of books and dvds in 2006 and 2016. Each graph shows a heatmap of the checkout activity over the course of the year, for one item type (books or dvds).

** See Picture

2.1 Analysis

In comparing these data, we notice a few trends: 1. Weekends became much more active over the 2006 - 2016 decade 2. Books overall are much more popular than DVDs 3. In 2006 the most active time / item was book checkouts on friday nights 4. Friday activity decreased dramatically for both books and DVDs 5. It seems the library hours may have changed over time. In 2006 there is almost no activity on Mon/Tue mornings or Fri / Sat evenings, but this is not true in 2016.