MAT 265
Ilia Nikiforov
Assignment 2

Exploratory analysis of SPL database

This report analyzes several disconnected questions using Seattle Public Library database. Specifically, the report covers: (1) the aftermath of J.K. Rowling's Twitter scandal; (2) items that were not returned to the library and associated check-in issues; and (3) general information regarding the busiest day in the SPL to date.

1. JK Rowling scandal continued

Following last week's report, I am interested in the aftermath of J.K. Rowling's Twitter scandal. Specifically, I want to know whether public outrage regarding her controversial tweet on 6/6/2020 has resulted in the boycott of her books. This time, I look into a short, granular query, and build a timeline of Harry Potter book orders between May and October of 2020.

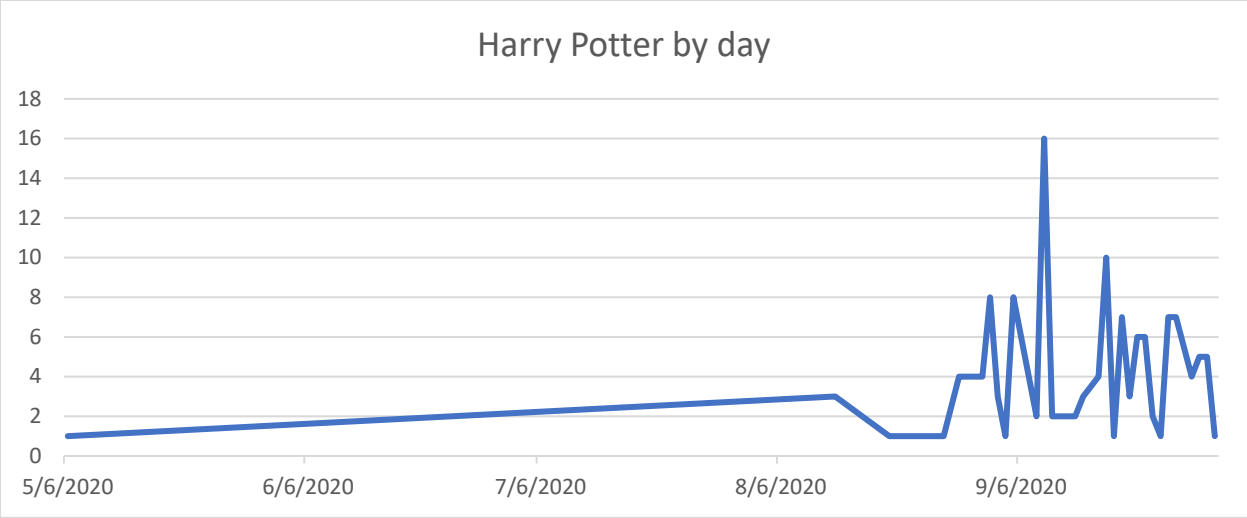
Query:

```
select
        date_format(cout, '%Y-%m-%d') as days,
        count(*) as Harry_Potter_counts
from spl_2016.inraw
where title like 'Harry Potter%' and date_format(cout, '%Y-%m-%d') between '2020-05-01' and '2020-10-01'
group by 1
order by 1
```

Result:

The results are stored in the jkrowling_by_date.csv.

The plot below visualizes the results of the query by date. As we can see, between May and September 2020, there was not a single instance of a library patron checking out any Harry Potter book. I find it unlikely, as the scandal itself occurred in June. Therefore, this might be indicative of two possibilities: (1) missing data; (2) all Harry Potter books were checked out and missing from the library at that time.



2. Items that were not returned.

In this section I attempt to identify instances where a book was checked out but not returned to the library. I begin with an exploratory query that identifies instances where check-in date ('cin') is null:

Query 1:

```
select * from spl_2016.inraw where cin is null limit 10
```

Result: No results. Every entry has a check-in date. As the next step, I inspect for any instances when the check-in date ('cin') was coded as a technical date of '1970-01-01', which is not an actual date but rather a placeholder for a date, which might be used to denote instances where no check-in date has occurred yet.

Query 2:

```
select * from spl_2016.inraw where cin like '1970%' limit 10
```

Result: No results again. As the next step, I look for instances where check-in date occurred before check-out date ('cout') and limit the search to 10 entries.

Query 3:

```
select * from spl_2016.inraw where cin < cout limit 10
```

Result:

itemNumber	cout	cin
273228	1/2/2006 15:01	1/2/2006 15:00
1052922	1/2/2006 15:50	1/2/2006 15:49
636624	1/3/2006 10:44	1/3/2006 10:43
342959	1/3/2006 12:22	1/3/2006 12:21

291265	1/3/2006 16:28	1/3/2006 16:27
2378585	1/3/2006 16:39	1/3/2006 16:38
2028357	1/3/2006 17:29	1/3/2006 17:28
905125	1/3/2006 18:04	1/3/2006 18:03
789964	1/4/2006 13:45	1/4/2006 13:44
406561	1/4/2006 16:02	1/4/2006 16:01

The result is stored in 'checkin_before_checkout_10.csv'. Some items have check-in date exactly one minute before the check-out date, like the table shows. Is it possible that in this case check-in date is another person returning the book, and check-out is the next person taking it? Probably not – I would assume it takes more than 1 minute to process the returned book. It appears, then, that this 1-minute anomaly is a technical error in the database.

At the next step, let's see how many instances of this we have per month in 2018, for example.

Query 4:

```
select

date_format(cout, '%Y-%m') as months,

count(*) as in_out_issue_counts from spl_2016.inraw

where cin < cout and cout like '2018%'

group by 1

limit 10
```

Result:

months	in_out_issue_counts
2018-03	18
2018-04	29
2018-05	35
2018-06	16
2018-07	18
2018-08	22
2018-09	5
2018-10	78
2018-11	34
2018-12	60

Result: The results are stored in 'checkin_before_checkout_monthly_2018'. It appears that every month there is a small number of those occurrences.

Let's investigate further: are all of these instances only 1 minute, or is there more of a difference? In the following query I will utilize the command "timestampdiff", with a parameter set to "minute", which will produce an absolute difference between check-out and check-in times, in minutes. I will further limit this query to the cases when check-in times occurred prior to check-out times. I will limit the number of the results by 5 to optimize runtime and sort them in descending order.

Query 5:

```
select TIMESTAMPDIFF(minute, cin,cout) from spl_2016.inraw

where cin<cout

order by TIMESTAMPDIFF(minute, cin,cout) desc

limit 5
```

Result: the result is stored in "checkin_before_checkout_top5_difference.csv". The table below illustrates the results. It appears that there are entries where the check-in times was logged in 550+ minutes earlier than the check-out time. Perhaps we can attribute that to human error when logging things in the database. If this process is automated, however, the possible explanation could be that check-in time from the previous patron, was somehow merged into the row associated with the check-out time of the next patron.

TIMESTAMPDIFF(minute, cin,cout)
572
568
566
554
552

Conclusion: I was not able to identify entries where the library item was not returned at all. I was, however, able to detect a number of rows where the check-in time occurred prior to check-out time. This anomaly can be attributed to a database issue, or a human error (in case librarians are able to log check-in times into the database manually).

3. The busiest day at the library

In this part I will examine the busiest day in the library. I will use check-out counts ('cout') per day as a measure of library activities, sort in descending order and display the top-10 busiest days.

Query 1:

```
select

        date_format(cout, '%Y-%m-%d') as days,

        count(*) as counts

from spl_2016.inraw
```

group by 1

order by 2 desc

limit 10

days	counts
1/1/1970	941006
3/13/2020	100724
11/12/2009	45434
2/22/2011	45163
2/16/2010	45121
1/18/2011	44712
11/26/2008	44610
1/2/2009	44264
8/26/2009	43808
7/12/2010	43477

Result:

The result is stored in the 'busiest_days.csv' and displayed in the table above. The "technical date" of 1/1/1970 has the most counts – I will ignore that. The second busiest date is March 13, 2020. This day also appears to be an anomaly, as everything afterwards is somewhere around ~40-45k range. Incidentally, March 13, 2020, is the day when the President of the US declared the state of emergency concerning the COVID outbreak. Could it be that people panicked and started researching this topic at the SPL of all places?

In the following query I will attempt a more detailed look into the types of items that were checked out on that day. Specifically, I will group them by item type (which includes books, video materials, etc.) and Dewey Class.

Query 2:

select

itemtype, deweyClass, count(*) as counts

from spl_2016.inraw

where cout like '2020-03-13%'

group by 1,2

order by 3 desc

limit 10

Result: the result is stored in the "busiest_day_details" and displayed in the table below. Evidently, kid's books, adult books, and DVD materials were checked out the most on March 13, 2020. Most of those

items do not have a Dewey Class attached to them in the database, however. From the information available on those media that do have Dewey Class, we can see that people were checking out some comic books (741.59..) and rock music (782.42..).

itemtype	deweyClass	Counts
jcbk		33665
acbk		17886
acdvd		11766
jcdvd		2242
acbk	741.5973	1725
pkbknh		1368
accd	782.42166	1002
jccd		985
accd		816
acbk	741.5952	524

In the following query I will try to identify exactly the most popular items that were checked out on March 13, 2020.

Query 3:

```
select

title, count(*) as counts

from spl_2016.inraw

where cout like '2020-03-13%'

group by 1

order by 2 desc

limit 10
```

Result: the result is stored in “busiest_day_top10_items.csv”. Given the total count of items above 100 thousand, the top performing entries are not exactly very informative. We can see a novel that was released 3 days prior (“My Dark Vanessa”), and a few other items.

title	counts
My dark Vanessa a novel	101
Something that may shock and discredit you	61
Nickel boys a novel	60
Bombshell	59
Untamed	58
testaments	55
In the dream house a memoir	54

Red at the bone	44
Uncut gems	43
warning	42

In the following query I will try to find out whether there were any items related to the COVID-19 pandemic that were checked out that day. For that, I will look up items that contain several keywords, such as “emergency”, or “pandemic”

Query 4:

```
select *

from spl_2016.inraw

where cout like '2020-03-13%'

and title like 'COVID%' or '%COVID%' or '%coronavirus%' or '%emergency%' or '%pandemic%'

limit 10
```

Result: No results.

Conclusion: It does not seem that people were rushing to the library to research the situation with COVID-19, given the state of emergency that was declared on March 13, 2020. Given the composition of popular demand by items, and some of the most checked out items on that day, it appears that people were generally trying to stock up on reading materials and media, anticipating a forthcoming lockdown at the onset of the COVID-19 pandemic.