MAT 259 Assignment 1: Knowledge Discovery

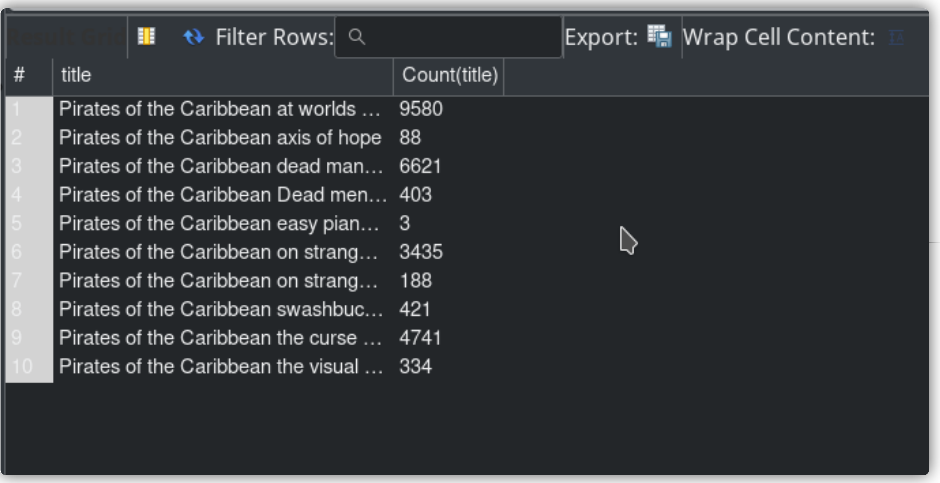
Intro

I did my assigment on the Pirates of the Caribbean movie series. I was am interested in visualizing the number of checkouts for the Pirate of the Caribbean movies to see how they fluctuate when a new movie in the series is released.

Investigation

I first ran the following code to see if there were any titles that were categorized as different items, even though we would consider them the same.(Aka check to see if the library got a new verson of the same movie). To account for unforseen capitalizations, I used SQL's `LOWER` operator.

```
SELECT
    title, Count(title)
FROM
    spl_2016.inraw
WHERE LOWER(title) LIKE  'pirates of the caribbean%'
GROUP By title
```



#	title	Count(title)
1	Pirates of the Caribbean at worlds ...	9580
2	Pirates of the Caribbean axis of hope	88
3	Pirates of the Caribbean dead man...	6621
4	Pirates of the Caribbean Dead men...	403
5	Pirates of the Caribbean easy pian...	3
6	Pirates of the Caribbean on strang...	3435
7	Pirates of the Caribbean on strang...	188
8	Pirates of the Caribbean swashbuc...	421
9	Pirates of the Caribbean the curse ...	4741
10	Pirates of the Caribbean the visual ...	334

Although it may be hard to tell from this screenshot, but the titles for the movies are all unique, as the repeats are not different versions, but supplementary material. Also, it seems that the third movie "At World's End" garnered the most number of checkouts. Keeping in mind our confirmation bias, this could be because it was the last movie in the original trilogy. Also fun fact: "At World's End" is the second most expensive movie ever made, costing 300 millon dollars.

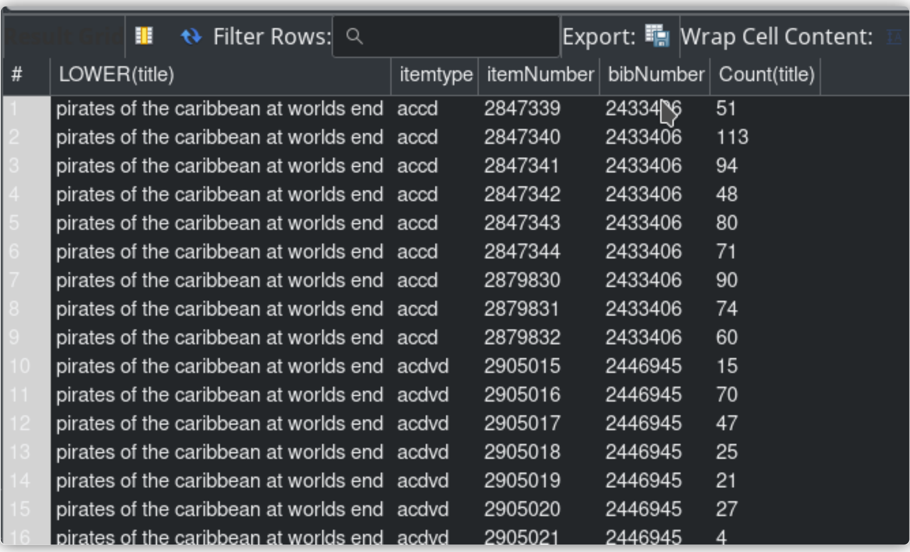
For future reference, the actual title names for the movies, as listed in this dataset, are:

1. Pirates of the Caribbean the curse of the Black Pearl
2. Pirates of the Caribbean dead mans chest
3. Pirates of the Caribbean at worlds end
4. Pirates of the Caribbean on stranger tides
5. Pirates of the Caribbean Dead men tell no tales

Next, to break the degenercy, I grouped by the title, itemtype, itemNumber, and bibNumber. This time, I used the `IN` operator to specify the exact titles.

```
SELECT
    title, itemtype, itemNumber, bibNumber, Count(title)
FROM
    spl_2016.inraw
WHERE LOWER(title)
IN
('pirates of the caribbean the curse of the black pearl',
'pirates of the caribbean dead mans chest',
'pirates of the caribbean at worlds end',
'pirates of the caribbean on stranger tides',
'pirates of the caribbean dead men tell no tales')

GROUP By title, itemtype, itemNumber, bibNumber
```

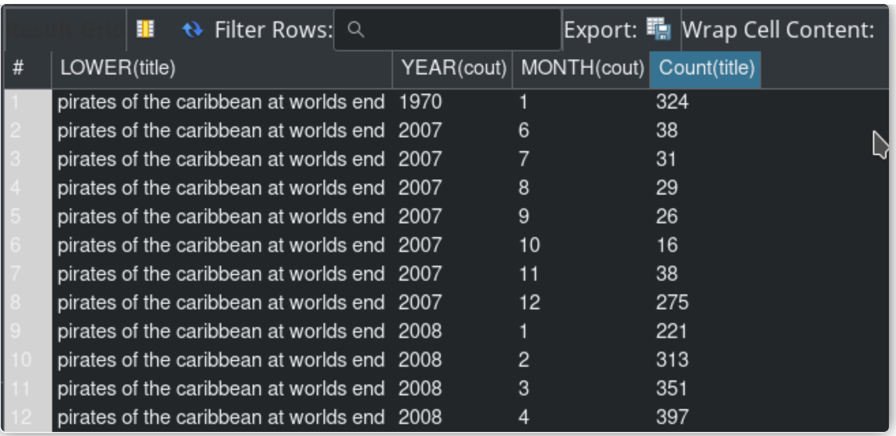


#	LOWER(title)	itemtype	itemNumber	bibNumber	Count(title)
1	pirates of the caribbean at worlds end	accd	2847339	2433406	51
2	pirates of the caribbean at worlds end	accd	2847340	2433406	113
3	pirates of the caribbean at worlds end	accd	2847341	2433406	94
4	pirates of the caribbean at worlds end	accd	2847342	2433406	48
5	pirates of the caribbean at worlds end	accd	2847343	2433406	80
6	pirates of the caribbean at worlds end	accd	2847344	2433406	71
7	pirates of the caribbean at worlds end	accd	2879830	2433406	90
8	pirates of the caribbean at worlds end	accd	2879831	2433406	74
9	pirates of the caribbean at worlds end	accd	2879832	2433406	60
10	pirates of the caribbean at worlds end	acdvd	2905015	2446945	15
11	pirates of the caribbean at worlds end	acdvd	2905016	2446945	70
12	pirates of the caribbean at worlds end	acdvd	2905017	2446945	47
13	pirates of the caribbean at worlds end	acdvd	2905018	2446945	25
14	pirates of the caribbean at worlds end	acdvd	2905019	2446945	21
15	pirates of the caribbean at worlds end	acdvd	2905020	2446945	27
16	pirates of the caribbean at worlds end	acdvd	2905021	2446945	4

I didn't see anything useful that would come from seperating into cd's or dvd's, or knowing exactly which copy was checked out, so I decided that moving forward, I would not consider the itemtype, itemNumber, or bibNumber.

Now, on to my main goal, which was to get the data for the number of checkouts of each movie for every combination of month and year.

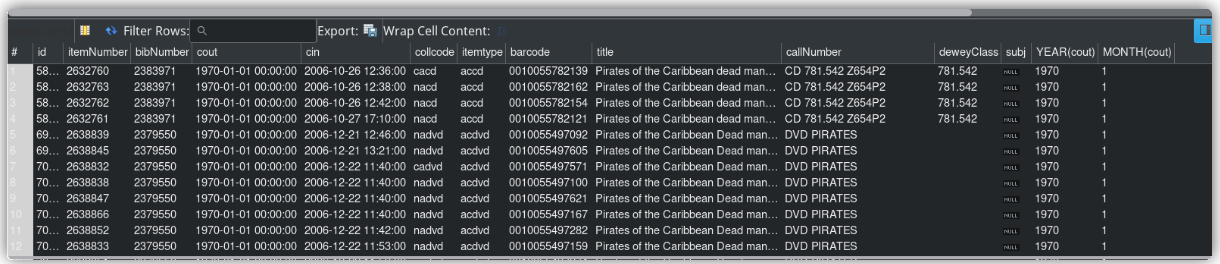
```
SELECT
    LOWER(title), YEAR(cout), MONTH(cout), COUNT(title)
FROM
    spl_2016.inraw
WHERE
    LOWER(title) IN ('pirates of the caribbean the curse of the black pearl' ,
        'pirates of the caribbean dead mans chest',
        'pirates of the caribbean at worlds end',
        'pirates of the caribbean on stranger tides',
        'pirates of the caribbean dead men tell no tales')
GROUP BY LOWER(title), YEAR(cout), MONTH(cout)
```



#	LOWER(title)	YEAR(cout)	MONTH(cout)	Count(title)
1	pirates of the caribbean at worlds end	1970	1	324
2	pirates of the caribbean at worlds end	2007	6	38
3	pirates of the caribbean at worlds end	2007	7	31
4	pirates of the caribbean at worlds end	2007	8	29
5	pirates of the caribbean at worlds end	2007	9	26
6	pirates of the caribbean at worlds end	2007	10	16
7	pirates of the caribbean at worlds end	2007	11	38
8	pirates of the caribbean at worlds end	2007	12	275
9	pirates of the caribbean at worlds end	2008	1	221
10	pirates of the caribbean at worlds end	2008	2	313
11	pirates of the caribbean at worlds end	2008	3	351
12	pirates of the caribbean at worlds end	2008	4	397

However, it seems there is a problem with this dataset, as one of the checkout years is 1970, which shouldn't be possible. To investigate further, I ran:

```
SELECT
    *, YEAR(cout), MONTH(cout)
FROM
    spl_2016.inraw
WHERE
    (LOWER(title) IN ('pirates of the caribbean the curse of the black pearl' ,
        'pirates of the caribbean dead mans chest',
        'pirates of the caribbean at worlds end',
        'pirates of the caribbean on stranger tides',
        'pirates of the caribbean dead men tell no tales'))
    AND (YEAR(cout)=1970)
```



#	id	itemNumber	bibNumber	cout	cin	collcode	itemtype	barcode	title	callNumber	deweyClass	subj	YEAR(cout)	MONTH(cout)
1	58...	2632760	2383971	1970-01-01 00:00:00	2006-10-26 12:36:00	cacd	accd	0010055782139	Pirates of the Caribbean dead man...	CD 781.542 Z654P2	781.542	NULL	1970	1
2	58...	2632763	2383971	1970-01-01 00:00:00	2006-10-26 12:38:00	nacd	accd	0010055782162	Pirates of the Caribbean dead man...	CD 781.542 Z654P2	781.542	NULL	1970	1
3	58...	2632762	2383971	1970-01-01 00:00:00	2006-10-26 12:42:00	nacd	accd	0010055782154	Pirates of the Caribbean dead man...	CD 781.542 Z654P2	781.542	NULL	1970	1
4	58...	2632761	2383971	1970-01-01 00:00:00	2006-10-27 17:10:00	nacd	accd	0010055782121	Pirates of the Caribbean dead man...	CD 781.542 Z654P2	781.542	NULL	1970	1
5	69...	2638839	2379550	1970-01-01 00:00:00	2006-12-21 12:46:00	nadvd	acdvd	0010055497092	Pirates of the Caribbean Dead man...	DVD PIRATES		NULL	1970	1
6	69...	2638845	2379550	1970-01-01 00:00:00	2006-12-21 13:21:00	nadvd	acdvd	0010055497605	Pirates of the Caribbean Dead man...	DVD PIRATES		NULL	1970	1
7	70...	2638832	2379550	1970-01-01 00:00:00	2006-12-22 11:40:00	cadvd	acdvd	0010055497571	Pirates of the Caribbean Dead man...	DVD PIRATES		NULL	1970	1
8	70...	2638838	2379550	1970-01-01 00:00:00	2006-12-22 11:40:00	nadvd	acdvd	0010055497100	Pirates of the Caribbean Dead man...	DVD PIRATES		NULL	1970	1
9	70...	2638847	2379550	1970-01-01 00:00:00	2006-12-22 11:40:00	nadvd	acdvd	0010055497621	Pirates of the Caribbean Dead man...	DVD PIRATES		NULL	1970	1
10	70...	2638866	2379550	1970-01-01 00:00:00	2006-12-22 11:40:00	nadvd	acdvd	0010055497167	Pirates of the Caribbean Dead man...	DVD PIRATES		NULL	1970	1
11	70...	2638852	2379550	1970-01-01 00:00:00	2006-12-22 11:42:00	nadvd	acdvd	0010055497282	Pirates of the Caribbean Dead man...	DVD PIRATES		NULL	1970	1
12	70...	2638833	2379550	1970-01-01 00:00:00	2006-12-22 11:53:00	nadvd	acdvd	0010055497159	Pirates of the Caribbean Dead man...	DVD PIRATES		NULL	1970	1

And indeed, there is a glitch in the database, as the check in times seem to be ok. I have yet to figure out why this is so and how to account for this in my visualization