Harry Potter: Movies vs Reading - Checkouts Over Time
by Nataly Moreno

Do movies with sequels lose popularity in the sequels? I've often heard that the sequels are not better than the first so I wanted to explore this question, and I have also heard that movies that come from books tend to be better than movies that do not. I was not able to find all the movie titles I had in mind, so I decided to take a closer look at the Harry Potter series. I wanted to know which of the 7 movies was checked out most frequently and how the movie checkouts compared to the book checkouts. I also wanted to see when the books and movies were checked out most often over time.

Initially I was just looking at movies, but it turned out to be inefficient as I was getting multiple results. However, it allowed me to explore the database. Here is the first attempt:

```
#Harry Potter and the Sorcerer's Stone         November 16, 2001
SELECT
    COUNT(*)
FROM
    spl2.inraw
WHERE
    (title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title LIKE '%stone%'
        AND title LIKE '%sorcerers%'
        AND title NOT LIKE '%soundtrack%')
;

#Harry Potter and the Chamber of Secrets       November 15, 2002
SELECT
    COUNT(*)
FROM
    spl2.inraw
WHERE
    (title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title LIKE '%chamber%'
        AND title LIKE '%secrets%'
        AND title NOT LIKE '%soundtrack%')
        AND itemtype = 'acdvd'
        AND DATE(cout) >= '2002-11-15'
;
```

```
#Harry Potter and the Prisoner of Azkaban      June 4, 2004
SELECT
    COUNT(*)
FROM
    spl2.inraw
WHERE
    (title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title LIKE '%prisoner%'
        AND title LIKE '%azkaban%'
        AND title NOT LIKE '%soundtrack%')
        AND itemtype = 'acdvd'
        AND DATE(cout) >= '2004-06-04'
;
#Harry Potter and the Goblet of Fire           November 18, 2005
SELECT
    COUNT(*)
FROM
    spl2.inraw
WHERE
    (title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title LIKE '%goblet%'
        AND title LIKE '%fire%'
        AND title NOT LIKE '%soundtrack%')
        AND itemtype = 'acdvd'
        AND DATE(cout) >= '2005-11-18'
;
#Harry Potter and the Order of the Phoenix     July 11, 2007
SELECT
    COUNT(*)
FROM
    spl2.inraw
WHERE
    (title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title LIKE '%order%'
        AND title LIKE '%phoenix%'
        AND title NOT LIKE '%soundtrack%')
        AND itemtype = 'acdvd'
        AND DATE(cout) >= '2007-07-11'
;
```

```
#Harry Potter and the Half-Blood Prince        July 15, 2009
SELECT
    COUNT(*)
FROM
    spl2.inraw
WHERE
    (title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title LIKE '%half%'
        AND title LIKE '%blood%'
        AND title LIKE '%prince%'
        AND title NOT LIKE '%soundtrack%')
        AND itemtype = 'acdvd'
        AND DATE(cout) >= '2009-07-15'
;

#Harry Potter and the Deathly Hallows P1       November 19, 2010
SELECT
    COUNT(*)
FROM
    spl2.inraw
WHERE
    (title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title LIKE '%deathly%'
        AND title LIKE '%hallows%'
        AND title LIKE '%part 1%'
        AND title NOT LIKE '%soundtrack%')
        AND itemtype = 'acdvd'
        AND DATE(cout) >= '2010-11-19'
;

#Harry Potter and the Deathly Hallows P2       July 15, 2011
SELECT
    COUNT(*)
FROM
    spl2.inraw
WHERE
    (title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title LIKE '%deathly%'
        AND title LIKE '%hallows%'
        AND title LIKE '%part 2%'
        AND title NOT LIKE '%soundtrack%')
        AND itemtype = 'acdvd'
        AND DATE(cout) >= '2011-07-15'
;
```

I fine tuned the query further to display all the data I wanted:

```
SELECT
    title,
    IF(itemtype = 'acdvd' OR 'acvhs' OR 'jcvhs',
        'movie',
        'book') AS itemtype,
    DATE(cout) AS 'checkout date'
FROM
    spl2.inraw
WHERE
    title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title NOT LIKE '%soundtrack%'
        AND ((title LIKE '%stone%'  AND title LIKE '%sorcerers%')
        OR (title LIKE '%chamber%'  AND title LIKE '%secrets%')
        OR (title LIKE '%prisoner%' AND title LIKE '%azkaban%')
        OR (title LIKE '%goblet%'   AND title LIKE '%fire%')
        OR (title LIKE '%order%'    AND title LIKE '%phoenix%')
        OR (title LIKE '%half%'     AND title LIKE '%prince%')
        OR (title LIKE '%deathly%'  AND title LIKE '%hallows%'))
        AND (itemtype = 'acdvd' OR itemtype = 'acvhs'
        OR itemtype = 'jcvhs'
        OR itemtype = 'acbk'
        OR itemtype = 'jcbk')
ORDER BY title
```

The query took 127.244 seconds.

Query Explanation

This query will print out 3 columns: title, itemtype, checkout date
Since I was looking at movie checkouts, it does not matter whether someone watched harry potter as a dvd or a vhs. Querying the database showed that there were two types of vhs and two types of book categories in which Harry Potter was found so I checked for those and labeled them “movie” or “book” because it doesn’t matter as long as it is one of those two media forms. The “where” part of the query narrows down the search so I only get the Harry Potter series titles, otherwise I would get many other titles that are not necessarily the book and movie titles. I also excluded soundtracks and media types that were not what I was looking for. Lastly, to make it easy to read I ordered the list by the title. I was surprised to find that Harry Potter is in different languages and also checked out, but not as often.

However, this does not say how many checkouts there are per title in an easy to view format for me, so I did one additional query to do the math for me. The number of lines from the first query is equal to the sum of all the totals from the second query.

This query says how many checkouts there are per title.

```
SELECT
    title,
    IF(itemtype = 'acdvd' OR 'acvhs' OR 'jcvhs',
        'movie',
        'book') AS itemtype,
        COUNT(*) AS 'total'
FROM
    spl2.inraw
WHERE
    title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title NOT LIKE '%soundtrack%'
        AND ((title LIKE '%stone%'  AND title LIKE '%sorcerers%')
        OR (title LIKE '%chamber%'  AND title LIKE '%secrets%')
        OR (title LIKE '%prisoner%' AND title LIKE '%azkaban%')
        OR (title LIKE '%goblet%'   AND title LIKE '%fire%')
        OR (title LIKE '%order%'    AND title LIKE '%phoenix%')
        OR (title LIKE '%half%'     AND title LIKE '%prince%')
        OR (title LIKE '%deathly%'  AND title LIKE '%hallows%'))
        AND (itemtype = 'acdvd' OR itemtype = 'acvhs'
        OR itemtype = 'jcvhs'
        OR itemtype = 'acbk'
        OR itemtype = 'jcbk')
        AND DATE(cout) >= '1800-01-01'
group by itemtype, title
ORDER BY title
```

This query took 102.335 seconds.

Query Explanation

The differences are minor. Instead of the “date” column, it has a “total” column which refers to the total number of checkouts. Furthermore, I only wanted the checkouts so I added a check for a date that is very far back to make sure I got all the checkouts. Lastly, I grouped by itemtype and title to get the results in fewer lines.

Analysis

It’s difficult to do a full analysis without being able to see the data in 3D. I think it has potential to be viewed in 3D with the x-axis as the number of book checkouts, the y-axis as the number of movie checkouts, and the z-axis as the date. From a glance, it appears as if people prefer to watch movies over reading, although for some the number of book checkouts were greater than the number of movie checkouts.

January 16, 2015

Modification to Query to give Numerical Results Only

```
SELECT
    title,
    IF(itemtype = 'acdvd' OR 'acvhs' OR 'jcvhs',
        '1',
        '0') AS itemtype,
    month(cout) AS 'month',
    day(cout) as 'day',
    year(cout) as 'year'
FROM
    spl2.inraw
WHERE
    title LIKE '%harry%'
        AND title LIKE '%potter%'
        AND title NOT LIKE '%soundtrack%'
        AND ((title LIKE '%stone%'  AND title LIKE '%sorcerers%')
        OR (title LIKE '%chamber%'  AND title LIKE '%secrets%')
        OR (title LIKE '%prisoner%' AND title LIKE '%azkaban%')
        OR (title LIKE '%goblet%'   AND title LIKE '%fire%')
        OR (title LIKE '%order%'    AND title LIKE '%phoenix%')
        OR (title LIKE '%half%'     AND title LIKE '%prince%')
        OR (title LIKE '%deathly%'  AND title LIKE '%hallows%'))
        AND (itemtype = 'acdvd' OR itemtype = 'acvhs'
        OR itemtype = 'jcvhs'
        OR itemtype = 'acbk'
        OR itemtype = 'jcbk')
ORDER BY title
```

Python Code to Separate Results into Smaller CSV Files

```
sorcerersStone  = []
chamberSecrets  = []
prisonerAzkaban = []
gobletFire      = []
orderPhoenix    = []
halfPrince      = []
deathlyHallows  = []
deathlyHallows1 = []
deathlyHallows2 = []

substrings = ["sorcerers stone", "chamber of secrets",
                      "prisoner of azkaban", "goblet of fire",
                      "order of the phoenix", "prince", "part 1",
                      "part 2", "deathly hallows"];

titles = ["sorcerersStone", "chamberSecrets", "prisonerAzkaban",
                "gobletFire", "orderPhoenix", "halfPrince",
                "deathlyhallowsp1", "deathlyhallowsp2",
                "deathlyHallows"];
```

```
def parseFile():
	inputFile = "HarryPotterBookMovieNumbericResults.csv"

	with open(inputFile, "r") as openfileobject:
		for line in openfileobject:
			line = line.lower()
			line = line.split(',')

			assignToList(line)

	ii = 0
	for t in titles:
		file = open(t + ".csv", 'w')
		file.write("itemtype,month,day,year\n")

		if(ii == 0):
			writeToFile(file, sorcerersStone)
		if(ii == 1):
			writeToFile(file, chamberSecrets)
		if(ii == 2):
			writeToFile(file, prisonerAzkaban)
		if(ii == 3):
			writeToFile(file, gobletFire)
		if(ii == 4):
			writeToFile(file, orderPhoenix)
		if(ii == 5):
			writeToFile(file, halfPrince)
		if(ii == 6):
			writeToFile(file, deathlyHallows1)
		if(ii == 7):
			writeToFile(file, deathlyHallows2)
		if(ii == 8):
			writeToFile(file, deathlyHallows)

		ii += 1

	ss = len(sorcerersStone)
	cs = len(chamberSecrets)
	pa = len(prisonerAzkaban)
	gf = len(gobletFire)
	op = len(orderPhoenix)
	hp = len(halfPrince)
	dh = len(deathlyHallows)
	p1 = len(deathlyHallows1)
	p2 = len(deathlyHallows2)

	print ss, cs, pa, gf, op, hp, dh, p1, p2
	print ss + cs + pa + gf + op + hp + dh + p1 + p2
```

```
def writeToFile(file, listy):
	for l in listy:
		i = 0
		for e in l:
			file.write(e)
			if(i < 3):
				file.write(",")
			i += 1
	file.close()

def assignToList(row):
	i = 0
	for s in substrings:
		found = row[0].find(s)
		if(found != -1):
			if(i == 0):
				sorcerersStone.append(row[1:])
				break
			if(i == 1):
				chamberSecrets.append(row[1:])
				break
			if(i == 2):
				prisonerAzkaban.append(row[1:])
				break
			if(i == 3):
				gobletFire.append(row[1:])
				break
			if(i == 4):
				orderPhoenix.append(row[1:])
				break
			if(i == 5):
				halfPrince.append(row[1:])
				break
			if(i == 6):
				deathlyHallows.append(row[1:])
				break
			if(i == 7):
				deathlyHallows1.append(row[1:])
				break
			if(i == 8):
				deathlyHallows2.append(row[1:])
				break
		i += 1

def main():
	parseFile()

if __name__ == "__main__":
	main()
```