Data Compilation from BoxOfficeMojo.com

VBA Final Project

4/8/2010
BYU Marriott School MBA Program
Adam Solter

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Trend data in cinema box office results can tell us much about how tastes are changing, how ratings affect box office results, and how well movies do internationally. BoxOfficeMojo.com compiles this data in a simple format that would be easy enough to download by hand (there are only five pages), but much of the important data is buried in individual pages for each movie.

This project is designed to:

- Extract the basic data automatically
- Drill down into each sub-page to get additional information based on the link from the main page.
- Compile the information from multiple-release titles which must be handled differently from single-release titles.
- Output the data into a worksheet where ad hoc analysis can be done.
- Show summary data.

The process used in this project is applicable when trying to compile data that requires a click on each link to retrieve more detailed data from sub-pages. Just as important, it is designed to handle exceptions when clicking on the link has a different result that expected and a different process must be used to extract the data. The automation of this process can save countless hours of hand-picking the data, especially if this is a process that must be done often.

IMPLEMENTATION

WEBSITE

The main page for this website is: http://boxofficemojo.com/alltime/domestic.htm

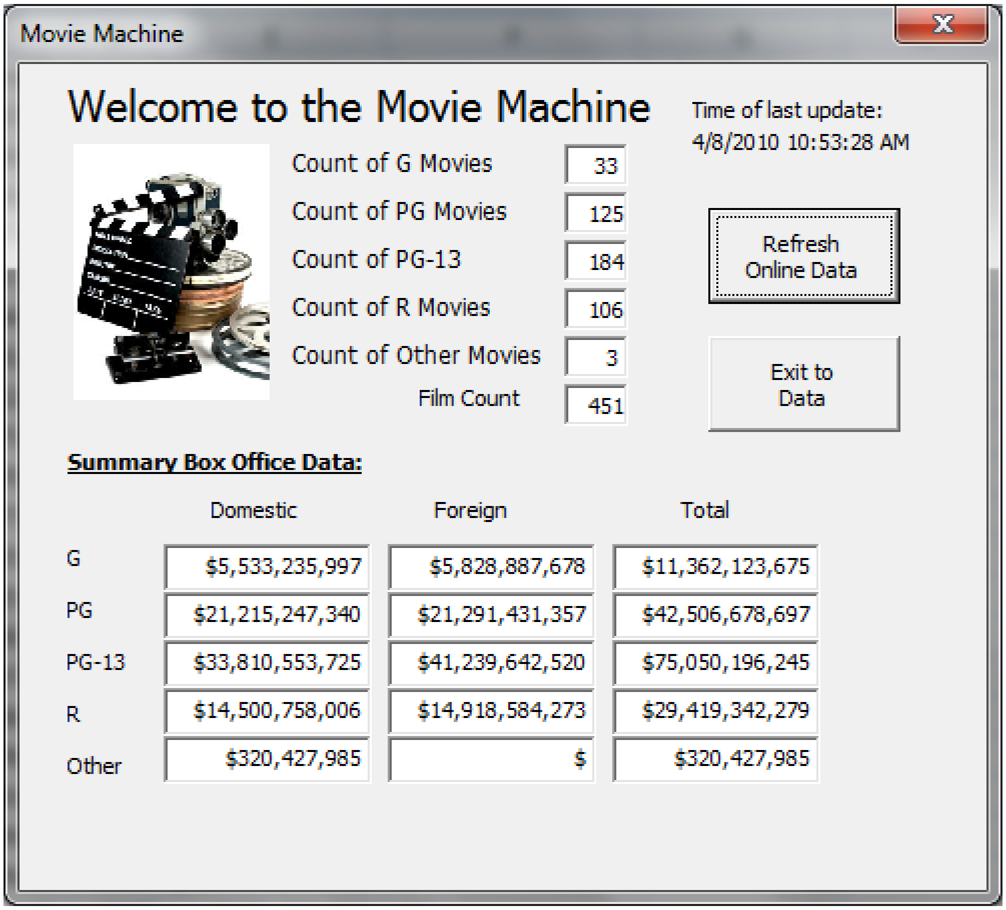
The robots.txt file does not prohibit the use of this data. Below are the contents of the file:

```
# robots.txt for http://www.boxofficemojo.com

User-agent: *
Disallow: /movies/default.movies.htm
Disallow: /showtimes/buy.php
Disallow: /forums/
Disallow: /derbygame/
Disallow: /grades/
Disallow: /moviehangman/
Disallow: /users/
```

INTERFACE

The interface is a simple form that shows summary data and has button to refresh that data or to exit so that the user can perform whatever analysis they wish.



Welcome to the Movie Machine

Time of last update:
4/8/2010 10:53:28 AM

Count of G Movies	33
Count of PG Movies	125
Count of PG-13	184
Count of R Movies	106
Count of Other Movies	3
Film Count	451

Summary Box Office Data:

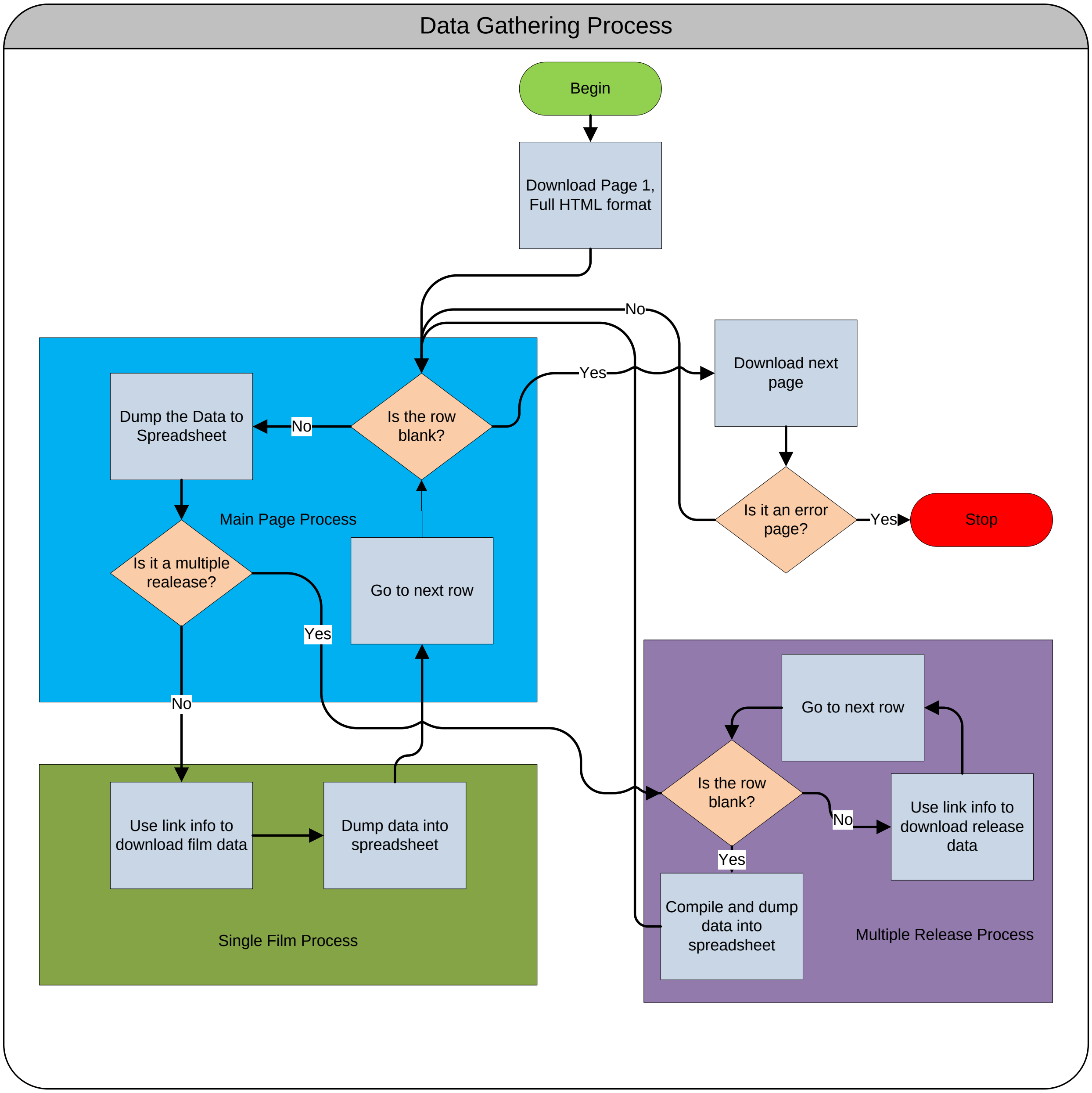
	Domestic	Foreign	Total
G	$5,533,235,997	$5,828,887,678	$11,362,123,675
PG	$21,215,247,340	$21,291,431,357	$42,506,678,697
PG-13	$33,810,553,725	$41,239,642,520	$75,050,196,245
R	$14,500,758,006	$14,918,584,273	$29,419,342,279
Other	$320,427,985	$	$320,427,985

If the "Refresh Online Data" button it pushed, a message box appears asking if the user really wants to refresh the data because it does take a while to compile the complete list. If they proceed, the form stays visible while the user can see each line of the data being entered (although the web queries that are part of the collection process are hidden). Once the data is refreshed, a message box announce the completion and the summary info on the form is refreshed.

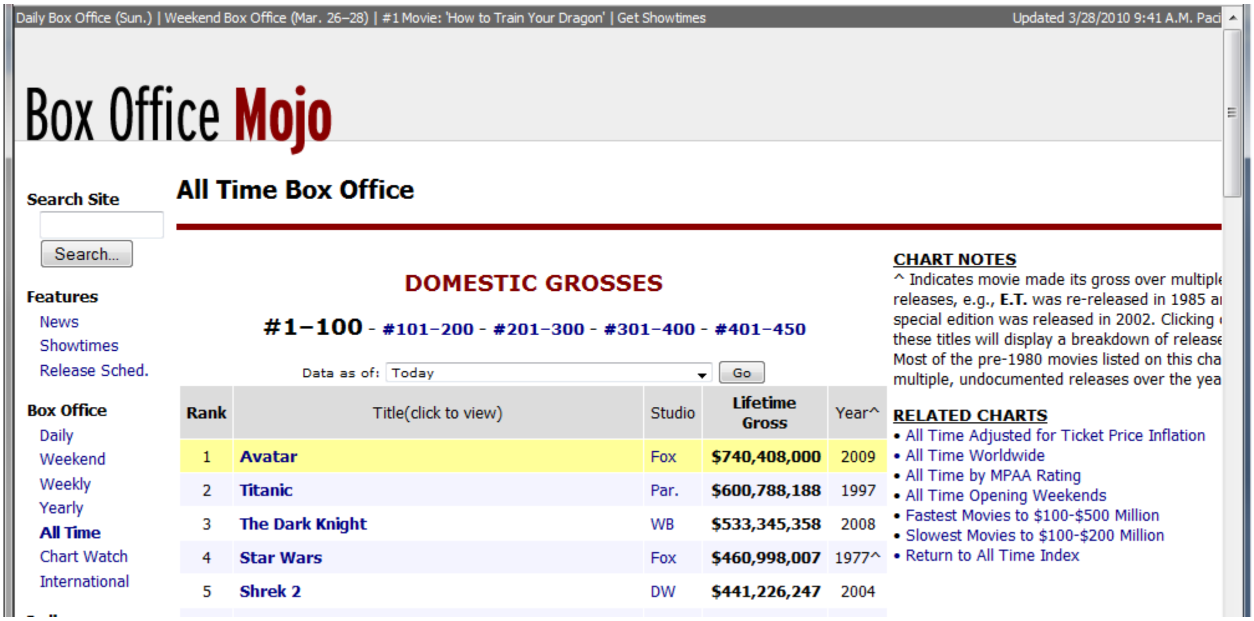
Because the types of analysis are far too many to contemplate, the option is available for the user to exit the form and perform their own measurements on the data. The data can be used in charts, pivot tables, and any other form of analysis the user wishes to perform.

PROCESS

MAIN DATA PROCESS


Data Gathering Process
Begin
Download Page 1, Full HTML format
No
Yes
Download next page
Dump the Data to Spreadsheet
No
Is the row blank?
Is it an error page?
Yes
Stop
Main Page Process
Is it a multiple realease?
Go to next row
Yes
No
Go to next row
Is the row blank?
No
Use link info to download release data
Yes
Use link info to download film data
Dump data into spreadsheet
Compile and dump data into spreadsheet
Single Film Process
Multiple Release Process

The basic data gathering process begins with the download of the first page. It is important that the option for Full HTML Formatting be checked so that the links are downloaded in addition to text of the site. Without the links it is impossible to tell from the text what the appropriate site is for each individual film. Below is a screenshot of the site followed by the code for the query. Notice the line of the query that specifies HTML formatting in bold.



Daily Box Office (Sun.) | Weekend Box Office (Mar. 26–28) | #1 Movie: 'How to Train Your Dragon' | Get Showtimes

Updated 3/28/2010 9:41 A.M. Paci

Box Office Mojo

Search Site

Search...

Features
- News
- Showtimes
- Release Sched.

Box Office
- Daily
- Weekend
- Weekly
- Yearly
- **All Time**
- Chart Watch
- International

All Time Box Office

DOMESTIC GROSSES

#1–100 - #101-200 - #201-300 - #301-400 - #401-450

Data as of: Today ▾ Go

Rank	Title(click to view)	Studio	Lifetime Gross	Year^
1	**Avatar**	Fox	**$740,408,000**	2009
2	**Titanic**	Par.	**$600,788,188**	1997
3	**The Dark Knight**	WB	**$533,345,358**	2008
4	**Star Wars**	Fox	**$460,998,007**	1977^
5	**Shrek 2**	DW	**$441,226,247**	2004

CHART NOTES
^ Indicates movie made its gross over multiple releases, e.g., **E.T.** was re-released in 1985 ar special edition was released in 2002. Clicking these titles will display a breakdown of release Most of the pre-1980 movies listed on this cha multiple, undocumented releases over the yea

RELATED CHARTS
- All Time Adjusted for Ticket Price Inflation
- All Time Worldwide
- All Time by MPAA Rating
- All Time Opening Weekends
- Fastest Movies to $100-$500 Million
- Slowest Movies to $100-$200 Million
- Return to All Time Index

```
With Sandbox.QueryTables.Add(Connection:= _
        urlstring, Destination:=Range( _
        "$A$1"))
        '.Name = "domestic_1"
        .FieldNames = True
        .RowNumbers = False
        .FillAdjacentFormulas = False
        .PreserveFormatting = False
        .RefreshOnFileOpen = False
        .BackgroundQuery = True
        .RefreshStyle = xlInsertDeleteCells
        .SavePassword = False
        .SaveData = True
        .AdjustColumnWidth = True
        .RefreshPeriod = 0
        .WebSelectionType = xlEntirePage
        .WebFormatting = xlWebFormattingAll
        .WebPreFormattedTextToColumns = True
        .WebConsecutiveDelimitersAsOne = True
        .WebSingleBlockTextImport = False
        .WebDisableDateRecognition = False
        .WebDisableRedirections = False
        .Refresh BackgroundQuery:=False
    End With
```

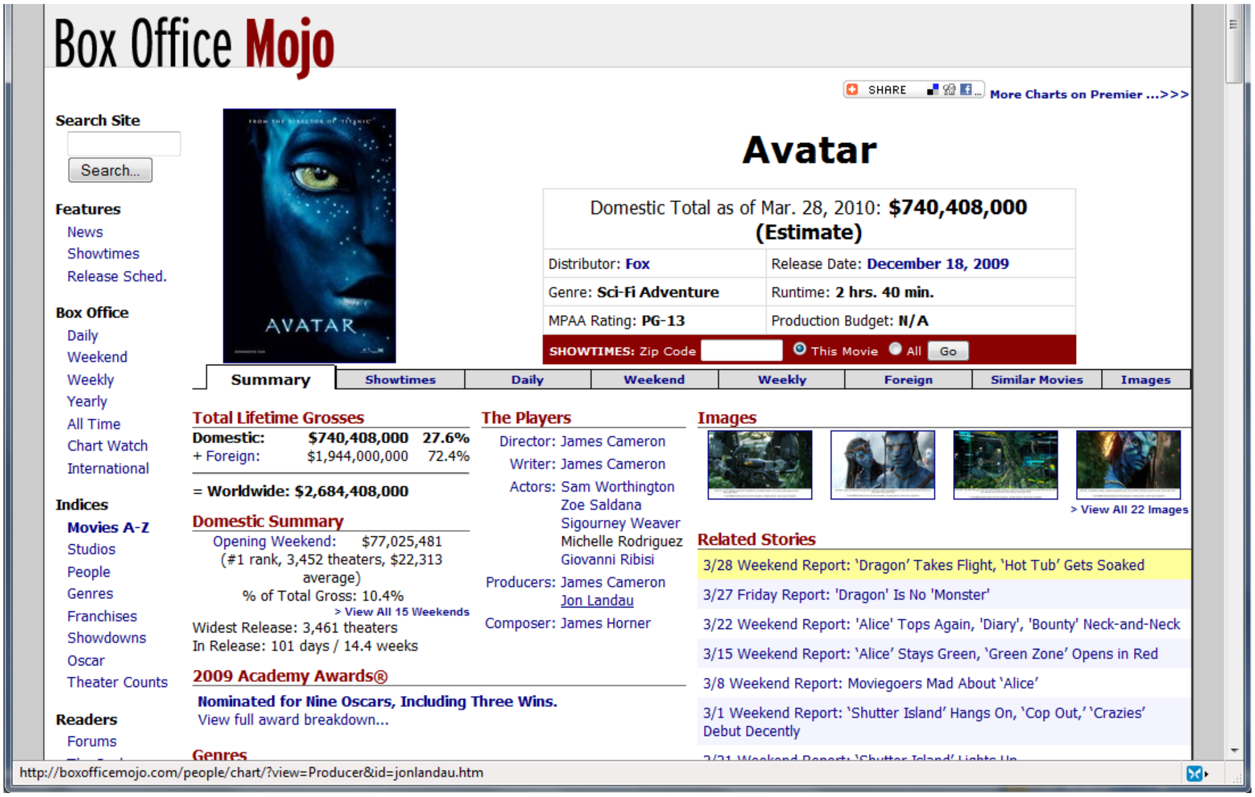
SINGLE RELEASE PROCESS

Once the main data is downloaded, it parses each line of data. First the relevant data is dumped to the spreadsheet where the data is compiled. As part of this process, it checks the date for a caret (^) symbol appearing next to it. This signifies that the movie was released multiple times, and therefore must be handled differently. If there is a caret, then it is stripped from the date and sent to the multiple release process described later. If not, then for each entry it retrieves the link to use in another web data query.

The code that retrieves the link is below. Note that there is only one link per cell.

```
Dim link as Hyperlink
Dim links as Hyperlinks
.
.
.
Set links = datarange.Cells(i, 1).Hyperlinks
For Each link In links
  SingleFilm = link.Address
Next
```

The individual movie page is retrieved, this time in plain text mode. A sample page is below:



Box Office Mojo

SHARE More Charts on Premier ...>>>

Search Site

Search...

Features
News
Showtimes
Release Sched.

Box Office
Daily
Weekend
Weekly
Yearly
All Time
Chart Watch
International

Indices
Movies A-Z
Studios
People
Genres
Franchises
Showdowns
Oscar
Theater Counts

Readers
Forums

Avatar

Domestic Total as of Mar. 28, 2010: **$740,408,000**
(Estimate)

Distributor: **Fox**	Release Date: **December 18, 2009**
Genre: **Sci-Fi Adventure**	Runtime: **2 hrs. 40 min.**
MPAA Rating: **PG-13**	Production Budget: **N/A**

SHOWTIMES: Zip Code ● This Movie ○ All Go

Summary | Showtimes | Daily | Weekend | Weekly | Foreign | Similar Movies | Images

Total Lifetime Grosses

Domestic:	**$740,408,000**	**27.6%**
+ Foreign:	$1,944,000,000	72.4%

= **Worldwide: $2,684,408,000**

Domestic Summary

Opening Weekend: $77,025,481
(#1 rank, 3,452 theaters, $22,313 average)
% of Total Gross: 10.4%
> View All 15 Weekends
Widest Release: 3,461 theaters
In Release: 101 days / 14.4 weeks

2009 Academy Awards®

Nominated for Nine Oscars, Including Three Wins.
View full award breakdown...

Genres

The Players

Director: James Cameron
Writer: James Cameron
Actors: Sam Worthington
Zoe Saldana
Sigourney Weaver
Michelle Rodriguez
Giovanni Ribisi
Producers: James Cameron
Jon Landau
Composer: James Horner

Images

> View All 22 Images

Related Stories

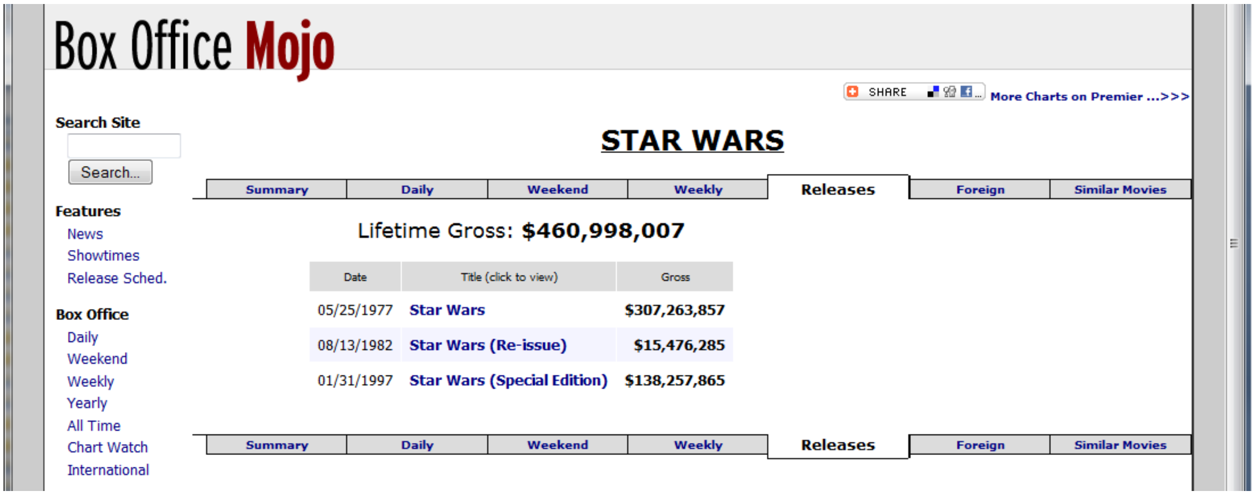
3/28 Weekend Report: 'Dragon' Takes Flight, 'Hot Tub' Gets Soaked
3/27 Friday Report: 'Dragon' Is No 'Monster'
3/22 Weekend Report: 'Alice' Tops Again, 'Diary', 'Bounty' Neck-and-Neck
3/15 Weekend Report: 'Alice' Stays Green, 'Green Zone' Opens in Red
3/8 Weekend Report: Moviegoers Mad About 'Alice'
3/1 Weekend Report: 'Shutter Island' Hangs On, 'Cop Out,' 'Crazies' Debut Decently
3/21 Weekend Report: 'Shutter Island' Lights Up

http://boxofficemojo.com/people/chart/?view=Producer&id=jonlandau.htm

Once the relevant data such as MPAA Rating and Foreign Gross are retrieved and added to the data, the next film in the list is processed.

MULTIPLE RELEASE PROCESS

As mentioned before, some films have multiple releases which create problems for the single release process. The link from the main page (notice Star Wars in the first screenshot has a caret) instead of an informational page produces a page such as the one below:



STAR WARS

Lifetime Gross: **$460,998,007**

Date	Title (click to view)	Gross
05/25/1977	**Star Wars**	**$307,263,857**
08/13/1982	**Star Wars (Re-issue)**	**$15,476,285**
01/31/1997	**Star Wars (Special Edition)**	**$138,257,865**

When this happens, the each link must be processed individually. For our purposes, we simply add together the foreign gross of each release to arrive at one unified datum for each film. As with the main data load, this page must be downloaded in HTML format so that the links remain intact.

TERMINATION OF THE PROCESS

Once the main page is finished (defined by hitting a blank spot) it goes to the next page by substituting the page number in the web data query. In order to be useful regardless of the number of films added in the future, it does this until it reaches an error page which means the page number is too high and not valid.

LEARNING POINTS

Some of the difficulties involved in this project are:

- Working with links – learning how to use the Hyperlink object and Hyperlinks collection is crucial to making this project work.
- Working with ranges – finding the appropriate data and setting up the correct ranges makes working with the data much easier.
- Reusing code – many of the process, although different, are similar and learning how to write code so that it can be easily reused simplifies the coding process and readability of the code.
- Parsing data – finding unique markers in the data that set the relevant information apart and manipulating the data is an important skill.

Although this data doesn’t have any direct business use other than my own curiosity, the process of using information from a data query to perform another query can be very useful in a business setting. When retrieving market data, detailed stock information, or any data that requires a drilling deeper into a website, as long as the drilling process is predictable and uniform, this basic process can simplify and automate the data gather process. There are 448 movies in this list, and retrieving this data by hand would take hours and because this information is constantly updated as movies are added, the automation of this type of data gathering can save many hours are monotonous work.