

Movie Data

Constance H. McLaren
Concetta A. DePaolo
Indiana State University

Journal of Statistics Education Volume 17, Number 1 (2009), www.amstat.org/publications/jse/v17n1/datasets.mclaren.html

Copyright © 2009 by Constance H. McLaren and Concetta A. DePaolo, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Time Series; Movie Box Office; Forecasting; Graphical Display of Data; Curve Fitting; Rate of Change

Abstract

The Movie dataset contains weekend and daily per theater box office receipt data as well as total U.S. gross receipts for a set of 49 movies. Dates are provided for all time series values. The diverse list of movies was selected, not at random, but to spark student interest and to provide a range of box office values. The values provide a rich dataset to use for applications such as simple graphical analysis, a variety of time series and causal forecasting models, curve-fitting, and rate of change analysis. A series of assignment questions is included and the accompanying Instructor's Manual provides representative solutions.

1. Introduction

Because time series forecasting is such a universal topic in business statistics classes, we have been intrigued with finding data sets that are both current and meaningful for our students. Although there is certainly a huge amount of financial time series data available, we have found that the movie box office data sets provide excellent examples of those forecasting features typically emphasized in business statistics textbooks: trend, seasonality, cycles, and randomness. Most students in our required business statistics classes are sophomores who have not yet studied finance. Using data that is familiar to them—they understand that receipts are higher on weekends, they know how blockbusters are released—ties statistical concepts from their classes to experiences in their lives. The accompanying data provides information on a wide variety of movies. Instructors who wish to track other movies or future releases are encouraged to visit the site from which these time series were obtained.

The dataset contains both weekend and daily per theater box office receipts and total US gross receipts for the 49

movies shown in [Table 1](#). To increase student interest, movies were chosen from lists of recent Academy Award Best Picture winners, highest grossing movies, series movies (e.g. the Harry Potter series, the Spiderman series), and from the Sundance Film Festival. Values have been retrieved from <http://www.the-numbers.com>. Movies selected include big budget as well as smaller, independent films. Receipts vary widely as well. In some cases, only weekend data is available.

Table 1: Movies in the Dataset

<i>Index</i>	<i>Movie</i>	<i>Year</i>	<i>Characteristic</i>
1	A Beautiful Mind	2001	Best Picture
2	American Beauty	1999	Best Picture
3	Batman	1989	Top 20 Gross
4	Beverly Hills Cop	1984	Top 20 Gross
5	Chicago	2002	Best Picture
6	Crash	2005	Best Picture
7	Departed, The	2006	Best Picture
8	Empire Strikes Back, The	1980	Top 20 Gross
9	ET	1982	Top 20 Gross
10	Forrest Gump	1994	Top 20 Gross
11	Ghost Busters	1984	Top 20 Gross
12	Gladiator	2000	Best Picture
13	Gods and Monsters	1998	Sundance
14	Good Girl, The	2002	Sundance
15	Harry Potter 1: Sorcerer's Stone	2001	Series
16	Harry Potter 2: Chamber of Secrets	2002	Series
17	Harry Potter 3: Prisoner of Azkeban	2005	Series
18	Harry Potter 4: Goblet of Fire	2004	Series

19	Harry Potter 5: Order of the Phoenix	2007	Series
20	Home Alone	1990	Top 20 Gross
21	In the Company of Men	1997	Sundance
22	Independence Day	1996	Top 20 Gross
23	Jurassic Park	1993	Top 20 Gross
24	Last Mimzy, The	2007	Sundance
25	Lion King, The	1994	Top 20 Gross
26	Lord of the Rings: The Return of the King	2003	Best Picture
27	Million Dollar Baby	2004	Best Picture
28	Pirates 1: Curse of the Black Pearl	2003	Series
29	Pirates 2: Dead Man's Chest	2006	Series, Top 20 Gross
30	Pirates 3: At World's End	2007	Series
31	Quinceanera	2006	Sundance
32	Raiders of the Lost Ark	1981	Top 20 Gross
33	Return of the Jedi	1983	Top 20 Gross
34	Road Home, The	2001	Sundance
35	Run Lola Run	1999	Sundance
36	Shakespeare in Love	1998	Best Picture
37	Shrek	2001	Series
38	Shrek 2	2004	Series, Top 20 Gross
39	Shrek the Third	2007	Series
40	Spider-Man	2002	Series, Top 20 Gross

41	Spider-Man 2	2004	Series
42	Spider-Man 3	2007	Series
43	Star Wars	1977	Top 20 Gross
44	Star Wars: Phantom Menace	1999	Top 20 Gross
45	Super Size Me	2004	Sundance
46	Thirteen	2003	Sundance
47	Titanic	1997	Best Picture, Top 20 Gross
48	Upside of Anger, The	2005	Sundance
49	You Can Count on Me	2000	Sundance

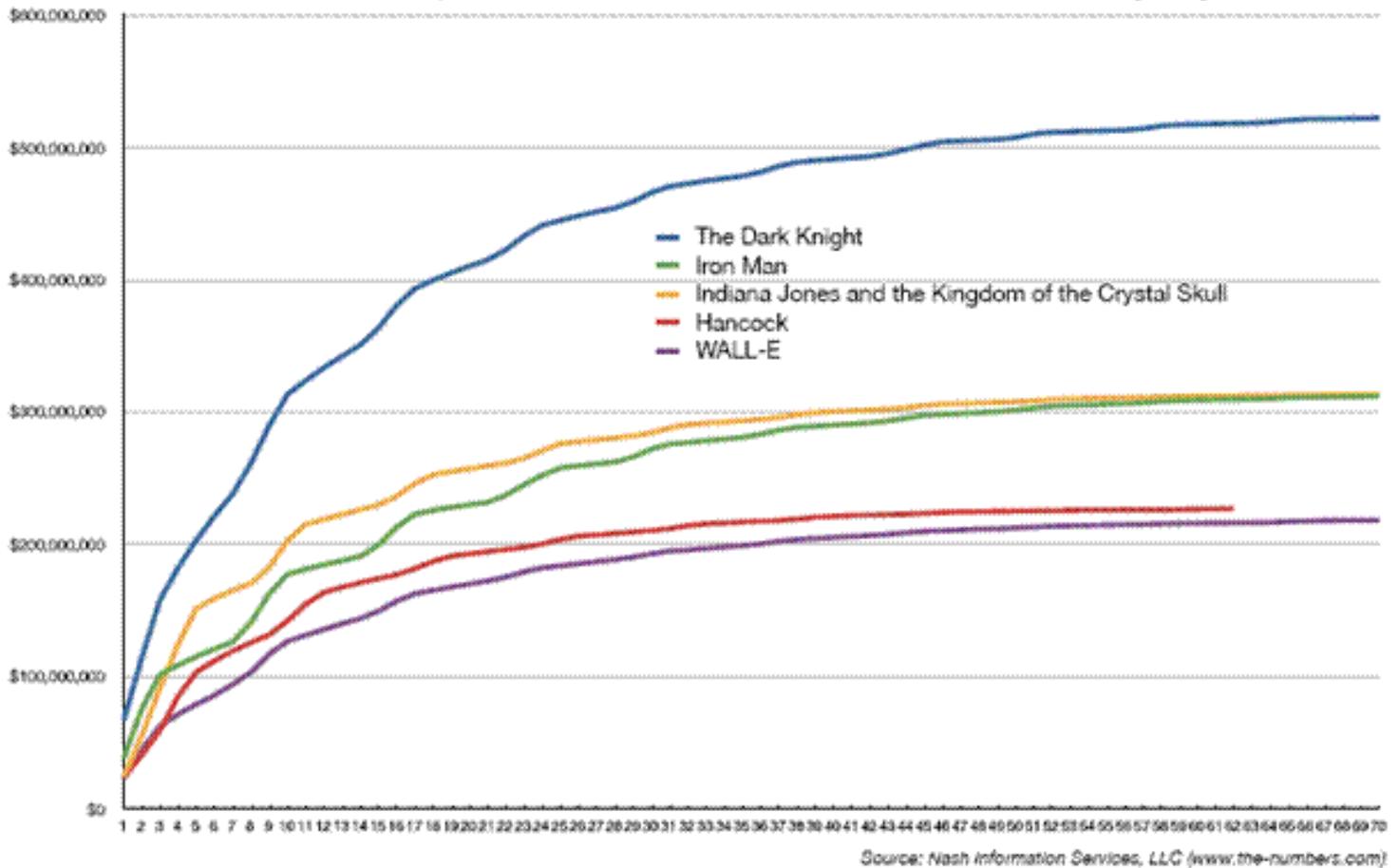
At our university, all business majors are required to complete a two-course introductory (non-calculus based) business statistics sequence, typically in their sophomore year. The first course covers data presentation, random variables and probability distributions, and inference. The second course covers tests of independence, ANOVA, regression, forecasting, and decision analysis as well as a brief unit on business applications of calculus. Typical business statistics texts include coverage of regression analysis and time series forecasting (see, for example, [Anderson, Sweeney, & Williams, 2008](#); [Bowerman, O'Connell, & Murphree, 2009](#); [Groebner, Shannon, Fry, & Smith, 2008](#); and [Levine, Stephan, Krehbiel, & Berenson, 2008](#)). We have found that the use of real data increases student interest in the topics we teach in business statistics courses and in an upper level forecasting elective, and we anticipate that this would be the case in other statistics courses. Students seem to enjoy data tied to the entertainment industry, and they are quick to connect the time series patterns they find to their own social activities.

In addition to the specific analytical questions provided in the assignments below, the data can support classroom discussions about analytical decision making. Even without additional research into the entertainment industry, students can use the data to make comparisons of similar movies, evaluate timing decisions for DVD releases, and look at the impact of holidays and award nominations on box office receipts.

A useful classroom discussion can center on "new product" forecasting. In this area, analysts usually look at analogies to learn how similar products performed in the past ([Makridakis, Wheelwright & Hyndman, 1998, page 466](#)). Students can brainstorm about whether similar movies (genre, actors, release timing, etc.) have similar patterns of receipts. Validation for this comparison process is supported by the charts created for industry watchers at The Numbers site. A typical chart, comparing major summer releases for 2008, is shown in [Figure 1](#) below.

Figure 1: Comparison Chart

Summer Blockbusters, 2008 - First 10 Weeks Cumulative Box Office by Days in Release



2. Data Sources

The data in the Movie data set were retrieved from <http://www.the-numbers.com>, a site that presents box office receipt data for hundreds of movies. For each movie, the site provides information on the number of theaters, the movie's rank, and total receipts as well as the per theater information. We have chosen to concentrate on the per theater information as it is more useful for classroom assignments, but instructors who want more detailed information or want to collect data on future releases are encouraged to visit the Movie Archive section of this site. Information on movie characteristics, such as a list of Academy Award winners, was found through various sites (www.oscars.org/awardsdatabase, www.afi.com/tvevents/100years/100yearslist.aspx, http://www.imdb.com/Sections/Awards/Sundance_Film_Festival).

3. Description of the Data

Three files contain the raw data: movietotal.dat, moviedaily.dat, and movieweekend.dat. The accompanying files movietotal.txt, moviedaily.txt, and movieweekend.txt are documentation files containing brief descriptions of the datasets. The total receipts file (movietotal.dat) has four variables: the movie's number in the alphabetical list, its title, its characteristic (type), and the gross US receipts (in \$ millions). There are two time series files (moviedaily.dat and movieweekend.dat), one showing daily per theater box office receipts in dollars, and the other showing weekend per theater box office receipts, for these movies.

The daily and weekend time series files have five variables. The first variable is the movie's number in the alphabetical list, the second is the movie title, the third is an index for the observation number, the fourth is the per theater box office receipt amount in dollars, and the fifth is the date (mm/dd/yyyy). For weekend data, the date is for the Friday of the Friday, Saturday, and Sunday that comprise the weekend total. If daily data is missing for a title, the third, fourth, and fifth variables are coded as NA. Movie titles are arranged alphabetically. The day of the week is not provided in the daily chart; if you have your students take this data to Excel, they can use the "=Weekday" function to determine the day of the week.

Some movies opened to a limited audience and so on those occasions we waited to record values until the movie was in general release. For some titles, the site does not report receipts every day and/or weekend near the end of the movie's run. It is a good exercise for students to look for missing entries in the time series and determine what to do about those instances. Alternatively, instructors might decide to cleanse the data in advance.

More detailed information appears in [Appendix A](#).

4. Pedagogical Uses

This dataset can support exercises relating to visual display of data, descriptive statistics, trend analysis, and the forecasting concepts commonly found in an introductory business statistics class. It is also appropriate for a class in operations management or a class dedicated to forecasting. If more than just a few of the observations are used, students should have access to software. Basic analyses such as graphing and descriptive statistics can be done with Excel, although use of Minitab, SPSS, or another statistical software package is preferred for many of the exercises.

Our approach to statistics follows typical business statistics books such as the widely used texts referenced above. These books commonly include at least one chapter on forecasting in addition to several chapters on regression analysis. In our approach, we first present the mathematical and statistical foundations for topics such as least squares calculations with normal equations, the relationships among entries in ANOVA tables, trend analysis, seasonal decomposition steps, and smoothing methods, so students understand the theoretical underpinnings of statistical methods before using software tools to perform calculations. When software output is presented, we focus on interpretation and analysis so that students are required to think critically about their results rather than simply reporting output without understanding.

We offer the following successive assignments for use in the classroom. Instructors would certainly have to choose those assignments that fit the educational objectives of the class and the abilities of the students. A detailed set of assignment questions and solutions is found in the accompanying Instructor's Manual.

Exercise 1: Data Retrieval and Graphing

Students will locate data for a specific movie, bring the data to the software package, format it, and create a time series plot. We use this in the first days of the introductory business statistics class; it would also be suitable for an information literacy class.

Exercise 2: Descriptive Statistics & Analysis

Students will compute descriptive statistics for several different types of movies using software, and examine these statistics to draw conclusions about the movie types. We use this exercise in the early part of the introductory business statistics class. It could also be used to illustrate the difficulty of using descriptive statistics to draw conclusions about time series data.

Exercise 3: Examination of Time Series Data

Students will create time series plots using daily and weekend movie box office data. Using visual analysis and software tools, they will prepare a discussion of the features of the plots. We use this exercise at the beginning of the forecasting unit to help students recognize trend and seasonality in time series data.

Exercise 4: Nonlinear Trend Forecasting

Using software, students will fit several nonlinear trend equations to the weekend per theater box office receipts and determine their suitability as forecasting models. We have used this exercise to illustrate nonlinear regression, trend fitting, and concepts of rate of change. It also provides the basis for a discussion of overfitting models when we ask students to consider whether their models are reasonable and appropriate.

Exercise 5: Time Series Project

This project duplicates the activities of previous exercises, combining them into one project, and adds a calculus-based activity for rate of change. We have had good results using this exercise as an out-of-class group project in the second required statistics course.

Exercise 6: Seasonal Forecasting

Students will examine the seasonal patterns in the daily per theater box office receipts. Using software tools available, they will create seasonal forecasting models and evaluate them. We have used this exercise in both the second required business statistics class, where we generally rely on seasonal decomposition, and in the specialized forecasting class, where we ask students to develop and compare results from several more advanced seasonal forecasting procedures.

Exercise 7: Comparing Several Movies

This is a more advanced exercise and could be used in our second course or a business strategy class. Students will play the role of a movie industry analyst who must predict box office revenue for a new movie. In order to find similar movies to use for comparison, they will need to determine which factors are appropriate. Data from the comparison group will be used to develop a model for the new release. We recommend this as a group exercise for upper level students.

5. Conclusion

The Movie data sets provide interesting data for use in a wide variety of statistics classes. In our business statistics classes we have found that using data from familiar products piques student interest. They are quick to see the relationship between their analysis and business decision making. By choosing those assignments that fit the

learning objectives of their classes, instructors can provide examples and exercises that augment material included with text books. The data can be used for activities as simple as plotting and finding descriptive statistics, but it also supports more advanced analysis.

Acknowledgments

The authors wish to thank Bruce Nash, The-Numbers.com, for supplying [Figure 1](#). Similar charts are posted at the site.

Appendix A - Key to Variables in Movie Data Files

For the file [movietotal.dat \(saved as tab delimited text\)](#)

Variable	Description	Label
1	Movie number	INDEX
2	Movie title	MOVIE
3	Category type	TYPE
4	Total US Gross Receipts (millions \$)	TOTAL

For the file [moviedaily.dat \(saved as tab delimited text\)](#)

Variable	Description	Label
1	Movie number	INDEX
2	Movie title	MOVIE
3	Observation number	DAY_NUM
4	Daily per theater receipts (\$)	DAILY_PER_THEATER
5	Date (mm/dd/yyyy)	DATE

Movies with missing daily data show NA for DAY_NUM, DAILY_PER_THEATER, and DATE.

For the file [movieweekend.dat \(saved as tab delimited text\)](#)

Variable	Description	Label
1	Movie number	INDEX
2	Movie title	MOVIE
3	Observation number	WEEK_NUM
4	Weekend per theater receipts (\$)	WEEKEND_PER_THEATER
5	Date (mm/dd/yyyy)	WEEKEND_DATE

Appendix B

The Movie Data Instructor's Manual, containing all exercise assignments and solutions, is available at [Appendix B Instructors Manual Assignments and Solutions.doc](#)

Data Sources

For movie box office data: <http://www.the-numbers.com/>

For a list of Academy Award winners: www.oscars.org/awardsdatabase

For a list of categorical films: www.afi.com/tvevents/100years/100yearslist.aspx

For a list of Sundance Film Festival winners: http://www.imdb.com/Sections/Awards/Sundance_Film_Festival/.

References

Anderson, D., D. Sweeney, & T. Williams (2008). *Statistics for Business and Economics*, 10th edition. Thomson South-Western, Mason, OH.

Bowerman, B., R. O'Connell, & E. Murphree (2009). *Business Statistics in Practice*, McGraw Hill/Irwin, New

York.

Groebner, D. P. Shannon, P. Fry, & K. Smith (2008). Business Statistics, 7th edition. Pearson Education, Upper Saddle River, NJ.

Levine, D., D. Stephan, T. Krehbiel, & M. Berenson (2008). Statistics for Managers, 5th edition. Pearson Education, Upper Saddle River, NJ.

Makridakis, S., Wheelwright, S., & R. Hyndman (1998). Forecasting: Methods and Applications, 3rd edition. John Wiley and Sons, New York.

Constance H. McLaren
Analytical Department
Indiana State University
Terre Haute, IN 47809
c-mclaren@indstate.edu

Concetta A. DePaolo
Analytical Department
Indiana State University
Terre Haute, IN 47809

[Volume 17 \(2009\)](#) | [Archive](#) | [Index](#) | [Data Archive](#) | [Resources](#) | [Editorial Board](#) | [Guidelines for Authors](#) | [Guidelines for Data Contributors](#) | [Home Page](#) | [Contact JSE](#) | [ASA Publications](#)