Dale Deans

Dr. Block

IST 652

December 16, 2023

Project Report

**Data and Preprocessing:**

For my project, I decided to go with the Most Profitable Hollywood Stories CSV file I found on GitHub. The dataset contains the information of the best performing films through the years 2007 to 2011. The dataset contains 643 lines along with 15 different attributes. To prepare the data, I first imported the proper packages then used a created function to upload the csv into a pandas dataframe which would allow me to manipulate the incoming information. Next, I used the dropna() function to drop rows with null values in any of the columns. Lastly, I felt like the “Story” and “Genre” columns were a little redundant, so I combined the two values into the “Genre” column and dropped the “Story” column.

**Questions:**

* For the top 4 grossing lead studios in 2010, what is their average gross from movies released that year?
* For the three top grossing and bottom grossing movies of 2008, what are the Rotten Tomatoes scores and profitability of those films? Is there a correlation between the score and the profit margin?
* For movies with a budget ranging between $50 to $60 million, what is the average profitability by year?
* For the top 5 studios in total domestic gross for 2011, What is the average number of theaters showing the movie on opening weekend? What is their average opening weekend sales?

**Methods:**

To answer the questions above, I will be using methods of transformation as well as group by functions and calculation functions like summarization and mean. I will also be using conditional functions to manipulate the data.

**Program:**

In the program below, you can see the packages being imported, the preprocessing steps, and the series of data transformations and group by calculations done in order to answer the questions. The notes at the top of the accompanying blocks of code in figure 1 specify what is being done/the purpose.

**Figure 1: Program**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Results:**

* For the top 4 grossing lead studios in 2010, what is their average gross from movies released that year?

**output: Figure 2: Average Worldwide Gross of top 4 Studios**

A screenshot of a computer

Description automatically generated

The results in figure 2 show us that Disney by far has the highest gross average followed by Paramount, Warner Bros., and then independent studios last. However, when we look at films released for the year, Disney has much less than everyone else. To see if this average was the result of a skew, I created a dataframe that showed only Disney movies for that year. Disney did indeed have two of those movies gross over 1 billion dollars with Toy Story 3 and Alice in Wonderland while the other films did no more than 355 million dollars in earnings. Nevertheless, at least four of their films that year were higher than the worldwide gross averages of the other three studios, who on the other hand put out at least double the films Disney did. I think this speaks to Disney’s popularity and reputation as an attribute to their success. Looking at the total worldwide gross for the year, they still beat out independent leading studios and finished behind Warner Bros. and Paramount by about $500 million and $630 million respectively even with the big difference in movies released.

* For the three top grossing and bottom grossing movies of 2008, what are the Rotten Tomatoes scores and profitability of those films? Is there a correlation between the score and the profit margin?

**output:**

**Figure 3a: Top 3**

A screenshot of a computer

Description automatically generated

**Figure 3b: Bottom 3**

A screenshot of a computer

Description automatically generated

To answer the question, by looking at the outputs in figure 3a and 3b it is evident there isn't much of a correlation between a movie's Rotten Tomatoes score and it's profitability. If we were to compare the top 3 and bottom 3 we can make more of a case towards there being a correlation, but when you compare the data with their respective dataframes, Kung Fu Panda's score is 11 points higher than the Indiana Jones movie despite having a more than 100% lower profitability. Also, The Express has the lowest profitabilty out of the bottom 3 but has a much higher Rotten Tomatoes score when compared to the other two.

* For movies with a budget ranging between $50 to $60 million, what is the average profitability by year?

**output: Figure 4: Average Profitability**

A screenshot of a computer

Description automatically generated

I decided to focus on a certain range because it gives the results a little more meaning and allows us to measure the change more accurately in profitability. Looking at the output in figure 4, we can see how the profitability percentage fluctuates throughout the years 2007 to 2011. It starts with a decent 20% increase from 2007 to 2008 and then the percentage skyrockets and peaks in 2009. The profitability changes in 2009 is followed by an almost 100% decrease in 2010 and another decrease in 2011. A New York Times article in 2009 supports the findings, “theater owners have sometimes found fortunes in times of distress, and distress in good times”, referring to the 2008-2009 great recession and the anomaly of the theater business thriving during this period.[[1]](#footnote-1)

* For the top 5 studios in total domestic gross for 2011, What is the average number of theaters showing the movie on opening weekend? What is their average opening weekend sales?

**output:**

**Figure 5: Average Number of U.S. Theaters and Opening Weekend Sales**

A screenshot of a computer

Description automatically generated

For this question I wanted to look at the correlation between the number of U.S. theaters showing a film in opening weekend and domestic opening weekend sales. I focused on the year 2011 and the top 5 grossing leading studios of that year to narrow down the data and make the results more readable. As we can see by the results in figure 5, there does seem to be a correlation between the average number of U.S. theaters in opening weekend and average domestic opening weekend sales. Comparing the two rankings, the order is almost identical with the exception to Dreamworks Pictures and Disney. Interestingly, Dreamworks Pictures leads in average domestic opening weekend sales for this year despite only releasing three films so the order of the average weekend sales may very well just be a result of a skew.

**Conclusion:**

The Pandas and NumPy packages allowed me to take the Most Profitable Hollywood Stories CSV file and put it into a pandas dataframe where I can make transformations and calculations in order to answer my questions. The questions presented touched on several aspects of the movie business. The first and fourth question touches on how the popularity and reputation of a lead studio may affect overall sales. The second question shows how critic ratings little if any relationship with the profitability of movies. Lastly, the third question reflects how outside factors can affect the movie business on a year-by-year basis.

1. Cieply, Michael, and Brooks Barnes. “Despite Downturn, Americans Flock to Movie Theaters.” The New York Times, March 1, 2009. https://www.nytimes.com/2009/03/01/movies/01films.html. [↑](#footnote-ref-1)