MOVIES IN THE U.S - PROFIT AND REVENUE ANALYSIS

Minh K. Chau

2023-02-04

About this project

In this project, I used the Movies data, provided by a course in Google Data Analytics certificate on Coursera. This data table includes movies titles, director, five cast members, released years, budgets, and revenues released from 2012 - 2016. I address the following questions in my analysis:

- What are the top 10 directors and cast members for movies that generated the most revenue, from 2012 - 2016?
- What is the most profitable genre, by year and of all time?
- Can we predict revenue just from budget for a particular movie?

LIBRARY & Read-in Files

```
library(tidyverse)
library(skimr)
library(ggpubr)        # customization ggplot2 & ggarrange
setwd("C:/Users/minhk/OneDrive/Data Science stuffs/PORTFOLIO/Movie Analysis")
options(digits = 3)
movies <- read.csv("movies.csv")
```

Data At first glance

The following code chunks will perform the following:

- Convert all revenues and budgets into millions of U.S. Dollars
- Create a column for Profit (in millions of U.S. Dollars), which is calculated by: (revenue - profit)/1,000,000
- Check for unique values in:
  - year: 2012 - 2016
  - genres: multiple

```
movies <- movies %>%
  mutate(revenue.mils = revenue/1000000,
         budget.mils = budget/1000000,
         profit.mils = (revenue - budget)/1000000)
skim(movies)
```

Data summary

Name	movies

Number of rows	508
Number of columns	15
_____	
Column type frequency:	
character	9
numeric	6
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
title	0	1	1	48	0	508	0
release.date	0	1	8	10	0	330	0
genre	0	1	5	11	0	17	0
director	0	1	3	27	0	414	0
cast1	0	1	5	38	0	345	0
cast2	0	1	0	40	5	403	0
cast3	0	1	0	25	23	413	0
cast4	0	1	0	24	56	408	0
cast5	0	1	0	25	119	354	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
year	0	1	2.01e+03	1.35e+00	2012.0	2.01e+03	2.01e+03	2.02e+03
budget	0	1	4.89e+07	4.92e+07	1000000.0	1.40e+07	3.00e+07	6.50e+07
revenue	0	1	1.52e+08	1.83e+08	1000000.0	3.11e+07	7.94e+07	2.04e+08
revenue.mils	0	1	1.52e+02	1.83e+02	1.0	3.11e+01	7.93e+01	2.04e+02
budget.mils	0	1	4.89e+01	4.92e+01	1.0	1.40e+01	3.00e+01	6.50e+01
profit.mils	0	1	1.03e+02	1.49e+02	-58.6	5.65e+00	4.88e+01	1.38e+02

At first glance, we have 508 rows with 15 columns. There are no missing values. No Data Cleaning process is necessary.

/ Top 10 Director and First Cast Member

```
movies %>%
  select(director,
         revenue.mils) %>%
  group_by(director) %>%
  summarize(sum = sum(revenue.mils)) %>%
  arrange(desc(sum)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    director           sum
##    <chr>            <dbl>
##  1 Chris Renaud     2044.
##  2 Zack Snyder      1541.
##  3 Francis Lawrence 1409.
##  4 Bryan Singer     1292.
##  5 Steve Martino    1123.
##  6 Ridley Scott     1034.
##  7 Justin Lin       1032.
##  8 Phil Lord        1002
##  9 Peter Jackson     956
## 10 David Ayer        914.
```

The top 10 directors whose movies received the highest revenues returned are shown by the tibbit above. How about Cast Member?

```
movies %>%
  select(cast1,
         revenue.mils) %>%
  group_by(cast1) %>%
  summarize(sum = sum(revenue.mils)) %>%
  arrange(desc(sum)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    cast1               sum
##    <chr>             <dbl>
##  1 Jennifer Lawrence 2204.
##  2 Tom Cruise        1914.
##  3 Hugh Jackman      1733.
##  4 Ben Affleck       1504.
##  5 Adam Sandler      1451.
##  6 Will Smith        1311
##  7 Matt Damon        1272.
##  8 Ray Romano        1245.
##  9 Bradley Cooper    1209.
## 10 Ryan Reynolds     1174
```

The table above showed the top 10 actors/actresses whose movies that were casting on generated the highest sum of revenues.

What genre generates the most profit for each year in 2012 - 2016?

```
movies %>%
  select(year,
         genre,
         profit.mils) %>%
  group_by(year,genre) %>%
  summarize(sum = sum(profit.mils)) %>%
  arrange(desc(sum)) %>%
  head(5)
```

```
## # A tibble: 5 x 3
## # Groups:   year [4]
##    year genre    sum
##   <int> <chr>  <dbl>
## 1  2014 Action 4225.
## 2  2016 Action 4215.
## 3  2012 Action 3947.
## 4  2013 Action 3897.
## 5  2012 Comedy 2829.
```

Action movies generated the most profit for: 2012, 2013, 2014, 2016. So the next question is: ***Does that mean Action moves generate the most profit per movie released?***

```
movies %>%
  select(year,
         genre,
         profit.mils) %>%
  group_by(year,genre) %>%
  summarize(average = mean(profit.mils)) %>%
  arrange(desc(average)) %>%
  head(5)
```

```
## # A tibble: 5 x 3
## # Groups:   year [3]
##    year genre     average
##   <int> <chr>       <dbl>
## 1  2014 Adventure    352.
## 2  2012 Adventure    331.
## 3  2013 Thriller     298
## 4  2012 Fantasy      282.
## 5  2013 Adventure    246.
```

According to the output, **Action movies do NOT generate the most profit per movie released in any year**.

Even though action movies generated the most profits for each year from 2012 - 2016 (except 2015), each action movie does not generate the most profit on average, in comparing to other genres.

To explain this contradiction, let's look at the number of movies released by year for each genre.

```
movies %>%
  select(year,
         genre,
         title) %>%
  group_by(year,genre) %>%
  summarize(count = n()) %>%
```

```
  arrange(desc(count)) %>%
  head(5)
```

```
## # A tibble: 5 x 3
## # Groups:   year [4]
##    year genre  count
##   <int> <chr>  <int>
## 1  2012 Comedy    28
## 2  2015 Drama     28
## 3  2012 Action    27
## 4  2013 Action    26
## 5  2014 Action    26
```

Based on this number, we can see that **Action movies stand in the top 5 in terms of number released (i.e., 2012, 2013, 2014)**. How about In terms of budget and Revenue?

```
movies %>%
  select(year,
         genre,
         budget.mils) %>%
  group_by(year,genre) %>%
  summarize(sum = sum(budget.mils)) %>%
  arrange(desc(sum)) %>%
  head(5)
```

```
## # A tibble: 5 x 3
## # Groups:   year [5]
##    year genre    sum
##   <int> <chr>  <dbl>
## 1  2013 Action 2212
## 2  2016 Action 2202
## 3  2014 Action 2116.
## 4  2012 Action 1958.
## 5  2015 Action 1449
```

From 2012 - 2016, action movies received the most budget.

```
movies %>%
  select(year,
         genre,
         revenue.mils) %>%
  group_by(year,genre) %>%
  summarize(sum = sum(revenue.mils)) %>%
  arrange(desc(sum)) %>%
  head(5)
```

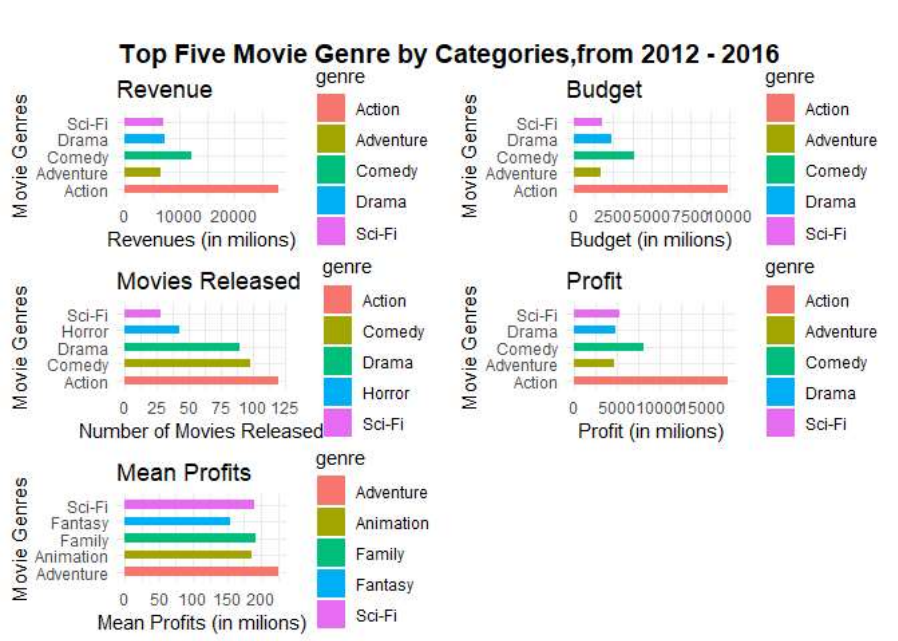
```
## # A tibble: 5 x 3
## # Groups:   year [4]
##    year genre    sum
##   <int> <chr>  <dbl>
## 1  2016 Action 6417.
## 2  2014 Action 6341
## 3  2013 Action 6109.
```

```
## 4  2012 Action 5905.
## 5  2012 Comedy 4151.
```

Except for 2015, Action movies made the most revenues.

so let's look at this trend as a whole.

From 2012 - 2016, what genres generate the most revenues, receive the most budget, release the most movies, generate the most net profit, and average profit/movie?



Results

Action movies do not generate the most profit per movies. But they do generate the most Revenues and tend to be the most popular genre (e.g., with more budgets and revenues).

Let's confirm this finding: ***what are the top 10 movies that generate the most revenues?***

```
movies %>%
  select(title,
         genre,
         year,
         revenue.mils) %>%
  group_by(genre) %>%
  arrange(desc(revenue.mils)) %>%
  head(10)
```

```
## # A tibble: 10 x 4
## # Groups:   genre [5]
##    title                                    genre     year revenue.mils
##    <chr>                                    <chr>    <int>        <dbl>
##  1 Despicable Me 2                          Comedy    2013         971.
##  2 The Hobbit: The Battle of the Five Armies Adventure 2014         956
##  3 Ice Age: Continental Drift               Adventure 2012         877
##  4 Batman v Superman: Dawn of Justice       Action    2016         873.
##  5 The Twilight Saga: Breaking Dawn â€“ Part 2 Fantasy   2012         830.
##  6 Fast & Furious 6                         Action    2013         789.
##  7 Deadpool                                 Action    2016         783.
##  8 The Amazing Spider-Man                   Action    2012         758.
```

```
##  9 The Hunger Games: Mockingjay â€“ Part 1     Sci-Fi     2014         755.
## 10 X-Men: Days of Future Past                 Action     2014         748.
```

We can see that 5/10 movies with the highest revenues belong to Action movies. This confirm our hypothesis that Action Movies are the most profitable genre.

Can we predict Revenue based on Budget?

For this question, we can use regression to predict Revenue by Budget for each movies, regardless of genre.

Linear Regression of the entire data table (N = 508)

```
summary(lm(data = movies, revenue.mils~budget.mils))
```

```
##
## Call:
## lm(formula = revenue.mils ~ budget.mils, data = movies)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -332.9  -57.6  -16.3   32.2  742.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.774      7.470    1.84    0.066 .
## budget.mils    2.828      0.108   26.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 119 on 506 degrees of freedom
## Multiple R-squared:  0.576,  Adjusted R-squared:  0.576
## F-statistic:  689 on 1 and 506 DF,  p-value: <2e-16
```

Results: In general, budget significantly predicts revenues, both in million of Dollars, b = 2.83, F(1,506) = 689, p<.05.

Interpretation: for every $1 million increase in budget, revenue is expected to increase by $2.83 million.

Now, let's analyzed this by group. We will first create a column that convert budget (in millions) into four subgroup:

- ">100": greater than $100 millions (>100)
- "<100": Between $100 millions and $50 millions (<100)
- "<50": between $50 millions and $25 millions
- "<25": $25 millions or less

```
movies <- movies %>%
  mutate(bud_group = ifelse(budget.mils >100,">100",ifelse(
    budget.mils <= 100 & budget.mils > 50,"<100",ifelse(
      budget.mils <=50 & budget.mils >25, "<50","<25"))))

movies %>%
  select(bud_group,
```

```
        budget.mils) %>%
  group_by(bud_group) %>%
  count()
```

```
## # A tibble: 4 x 2
## # Groups:   bud_group [4]
##   bud_group     n
##   <chr>     <int>
## 1 <100         97
## 2 <25         230
## 3 <50         107
## 4 >100         74
```

Because each group has a sample size of greater than 50, I decided that this sample size is sufficient to conduct linear regression analysis. Proceed to analysis.

Linear Regression in movies with budget greater than $100 million (N = 74)

```
budget100more <- movies %>%
  filter(budget.mils >100) %>%
  select(budget.mils,
         genre,
         revenue.mils) %>%
  data.frame()
summary(lm(data = budget100more, revenue.mils~budget.mils))
```

```
##
## Call:
## lm(formula = revenue.mils ~ budget.mils, data = budget100more)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -339.4 -141.0   -6.2  127.2  478.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.130     87.847    0.01     0.99
## budget.mils    2.915      0.577    5.05  3.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 178 on 72 degrees of freedom
## Multiple R-squared:  0.262,  Adjusted R-squared:  0.251
## F-statistic: 25.5 on 1 and 72 DF,  p-value: 3.2e-06
```

Results: For movies with budget of greater than $100 million, budget significantly predicts revenues, both in million of Dollars, b = 2.92, F (1,72) = 25.5, p<.05.

Interpretation: for every $1 million increase in budget, revenue is expected to increase by $2.92 millions, for movies with a budget of greater than $100 millions.

Linear Regression in movies with budget between $50 - $100 millions (N = 97)

```
budget100less <- movies %>%
  filter(budget.mils <=100 & budget.mils >50) %>%
```

```
  select(budget.mils,
         genre,
         revenue.mils) %>%
  arrange(desc(budget.mils)) %>%
  data.frame()
summary(lm(data = budget100less, revenue.mils~budget.mils))
```

```
##
## Call:
## lm(formula = revenue.mils ~ budget.mils, data = budget100less)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -244.7 -126.6  -56.7   55.7  734.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    77.85      97.41    0.80     0.43
## budget.mils     2.08       1.33    1.57     0.12
##
## Residual standard error: 187 on 95 degrees of freedom
## Multiple R-squared:  0.0253, Adjusted R-squared:  0.015
## F-statistic: 2.46 on 1 and 95 DF,  p-value: 0.12
```

Results: For movies with budget of between $50 - $100 million, budget does not significantly predicts revenues, both in million of Dollars, p >.05.

Interpretation: There exists no relationship between budgets and revenue returned (in million of dollars) in movies with budget between $50 - $100 millions Dollar.

//// Linear Regression in movies with budget between $25 - $50 millions (N = 107)

```
budget50less <- movies %>%
  filter(budget.mils <= 50 & budget.mils >25) %>%
  select(budget.mils,
         genre,
         revenue.mils) %>%
  arrange(desc(budget.mils)) %>%
  data.frame()
summary(lm(data = budget50less, revenue.mils~budget.mils))
```

```
##
## Call:
## lm(formula = revenue.mils ~ budget.mils, data = budget50less)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -128.1  -56.3  -12.5   28.0  354.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -17.55      41.75   -0.42   0.6750
## budget.mils     3.17       1.10    2.90   0.0046 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.7 on 105 degrees of freedom
```

```
## Multiple R-squared:  0.074,  Adjusted R-squared:  0.0652
## F-statistic: 8.39 on 1 and 105 DF,  p-value: 0.00458
```

Results: For movies with budget of between $25 - $50 million, budget significantly predicts revenues, both in million of Dollars, b = 3.17, F(1,105) = 8.39, p >.05.

Interpretation: for every $1 million increase in budget, revenue is expected to increase by $3.17 millions, for movies with a budget between $25 - $50 millions of Dollars.

//// Linear Regression in movies with budget less than $25 millions (N = 230)

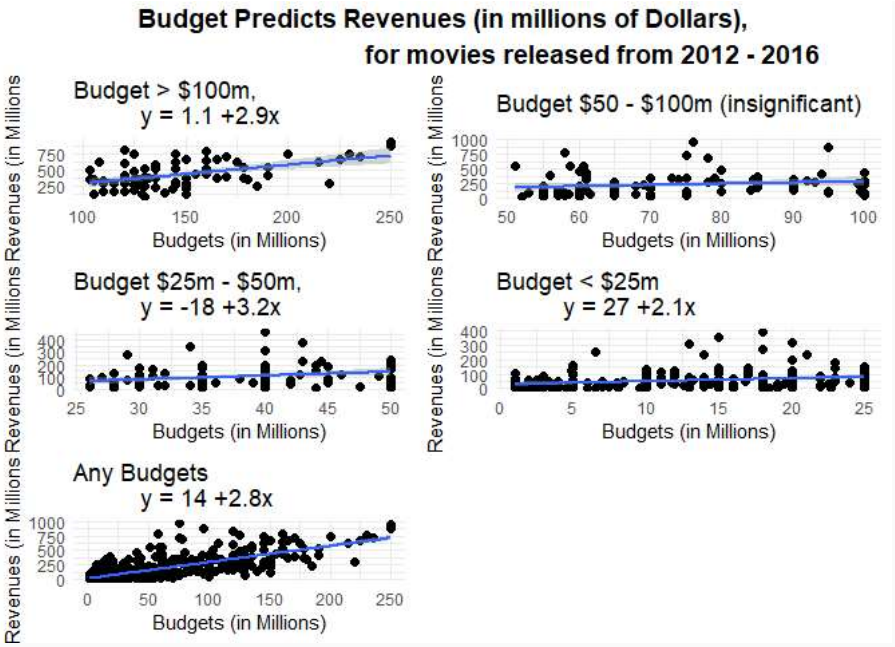
```
budget25less <- movies %>%
  filter(budget.mils <= 25) %>%
  select(budget.mils,
         genre,
         revenue.mils) %>%
  arrange(desc(budget.mils)) %>%
  data.frame()
summary(lm(data = budget25less, revenue.mils~budget.mils))
```

```
##
## Call:
## lm(formula = revenue.mils ~ budget.mils, data = budget25less)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -76.6  -34.3  -18.4   16.0  325.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   26.954      8.090    3.33  0.00101 **
## budget.mils    2.107      0.561    3.76  0.00022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.9 on 228 degrees of freedom
## Multiple R-squared:  0.0583, Adjusted R-squared:  0.0541
## F-statistic: 14.1 on 1 and 228 DF,  p-value: 0.00022
```

Results: For movies with budget of less than $25 million, budget significantly predicts revenues, both in million of Dollars, b = 2.12, F(1, 228) = 14.1, p >.05.

Interpretation: for every $1 million increase in budget, revenue is expected to increase by $2.11 millions, for movies with a budget less than $25 millions of Dollars.

//// Put it together



When analyzed by different budget groups, budget significantly predict revenues returned by each movies, across genre, from 2012 - 2016. This regression coefficient does not apply to movies with budget between $50 - $100 however.

/ TAKE-HOME MESSAGE

- Action movies generated the most profits and revenues. With lots of budgets, it is the most popular genre.
- It is best to have a movie with a budget around $25 - $50 millions, or $100 millions or more, for those movies to generate profits.