

# Bechdel

YOU

2021-01-24

In this mini analysis we work with the data used in the FiveThirtyEight story titled "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women" (<https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/>).

## Data and packages

We start with loading the packages we'll use.

```
library(fivethirtyeight)
library(tidyverse)
```

The dataset contains information on 1794 movies released between 1970 and 2013. However we'll focus our analysis on movies released between 1990 and 2013.

```
bechdel90_13 <- bechdel %>%
  filter(between(year, 1990, 2013))
```

There are — such movies.

The financial variables we'll focus on are the following:

- `budget_2013`: Budget in 2013 inflation adjusted dollars
- `domgross_2013`: Domestic gross (US) in 2013 inflation adjusted dollars
- `intgross_2013`: Total International (i.e. worldwide) gross in 2013 inflation adjusted dollars

And we'll also use the `binary` and `test_clean` variables for grouping.

## Analysis

Let's take a look at how median budget and gross vary by whether the movie passed the Bechdel test.

```
bechdel90_13 %>%
  group_by(binary) %>%
  summarise(med_budget = median(budget_2013),
            med_dmgross = median(domgross_2013, na.rm = TRUE),
            med_intgross = median(intgross_2013, na.rm = TRUE))
```

```
## # A tibble: 2 x 4
##   binary med_budget med_dmgross med_intgross
##   <chr>      <dbl>      <dbl>      <dbl>
## 1 FAIL    48385984.    57318406.    104475469
## 2 PASS    31070724.    45330446.    80124149
```

Next, let's take a look at how median budget and gross vary by a more detailed indicator of the Bechdel test result (`ok` = passes test, `dubious`, `men` = women only talk about men, `notalk` = women don't talk to each other, `nowomen` = fewer than two women).

```
bechdel90_13 %>%
  #
  #>%
  summarise(med_budget = median(budget_2013),
            med_dmgross = median(domgross_2013, na.rm = TRUE),
            med_intgross = median(intgross_2013, na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##   med_budget med_dmgross med_intgross
##   <int>      <dbl>      <dbl>
## 1   37878971    52270207    95523386
```

In order to evaluate how return on investment varies among movies that pass and fail the Bechdel test, we'll first create a new variable called `roi` as the ratio of the gross to budget.

```
bechdel90_13 <- bechdel90_13 %>%
  mutate(roi = intgross_2013 / domgross_2013)
```

Let's see which movies have the highest return on investment.

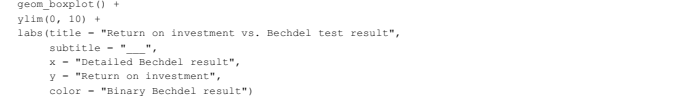
```
bechdel90_13 %>%
  arrange(desc(roi)) %>%
  select(title, clean_test, binary, roi, budget_2013, intgross_2013)
```

```
## # A tibble: 1,615 x 6
##   title                                clean_test binary    roi budget_2013 intgross_2013
##   <chr>                                <ord>    <chr>      <dbl>    <int>      <dbl>
## 1 Tropa de Elite                       ok      PASS    1638.    7345604    16086238
## 2 St. Trinian's                         ok      PASS    1496.    12808396    25219695
## 3 Jin ling shi san chai                ok      PASS    301.    103569079    9706426
## 4 Chingjeolhan geumjassi              ok      PASS    111.    5368649    28002720
## 5 Che: Part One                       notalk   FAIL    103.    62770866    32595998
## 6 Shallow Dancer                      nowomen  FAIL    87.5    13158460    52286669
## 7 Mononoke-hime                       ok      PASS    63.3    29024763    218193652
## 8 Agora                               notalk   FAIL    62.9    76001212    42335163
## 9 Perfume: The Story of a Murderer     ok      PASS    60.1    73624227    154416394
## 10 Centurion                          notalk   FAIL    53.6    16023478    9015508
## # ... with 1,605 more rows
```

Below is a visualization of the return on investment by test result, however it's difficult to see the distributions due to a few extreme observations.

```
ggplot(data = bechdel90_13, mapping = aes(x = clean_test, y = roi, color = binary)) +
  geom_boxplot() +
  labs(title = "Return on investment vs. Bechdel test result",
       x = "Detailed Bechdel result",
       y = "ROI",
       color = "Binary Bechdel result")
```

```
## Warning: Removed 15 rows containing non-finite values (stat_boxplot).
```



Zooming in on the movies with `roi < 10` provides a better view of how the medians across the categories compare:

```
ggplot(data = bechdel90_13, mapping = aes(x = clean_test, y = roi, color = binary)) +
  geom_boxplot() +
  ylim(0, 10) +
  labs(title = "Return on investment vs. Bechdel test result",
       subtitle = "____",
       x = "Detailed Bechdel result",
       y = "Return on investment",
       color = "Binary Bechdel result")
```

```
## Warning: Removed 49 rows containing non-finite values (stat_boxplot).
```



## References

1. Assignment Adapted from Mine Cetinkaya-Rundel's Data Science in a Box (<https://github.com/rstudio-education/datascience-box>)