AE 02 - Bechdel Test

Max Asmar

2026-02-17

Contents

Introduction

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".

The Bechdel Test is a measure of the representation of women in fiction. A movie passes the test if it meets three criteria: 1. It has at least two women in it 2. Who talk to each other 3. About something other than a man

In this application exercise, you will:

- Practice data filtering and summarizing
- Calculate return on investment (ROI)
- Create and interpret boxplots
- Use grouping to compare categories
- Practice using inline R code
- Interpret relationships between variables

Estimated time: 30-45 minutes

Getting Started

Fork and Clone Workflow

IMPORTANT: Just like AE 01, you need to work in YOUR forked repository.

Quick reminder: 1. Make sure you forked `application-exercises` repository (you only do this once!) 2. You should be working in YOUR fork: `YourUsername/application-exercises` 3. Navigate to the `ae-02-bechdel` folder 4. Open `ae-02-bechdel.Rmd` 5. Update the YAML with your name 6. Click **Knit** to verify it works

If you haven't forked yet: See detailed instructions in AE 01.

Data and Packages

We start with loading the packages we'll use.

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

What these packages do: - **fivethirtyeight:** Contains datasets from FiveThirtyEight articles - **tidyverse:** Data wrangling and visualization

About the Data

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))
```

What this code does: - `filter()`: Keeps only rows meeting a condition - `between()`: Checks if a value is between two numbers (inclusive) - Result: Movies from 1990-2013 only

---

Exercise 1: Count the Movies

How many movies are in our filtered dataset?

Fill in the blank using inline code:

There are 1615 movies in our dataset.

Hint: Use inline R code with `nrow(bechdel90_13)` in your narrative text.

Understanding the Variables

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 `clean_test` variables for **grouping**:

- `binary`: Whether the movie passed (PASS) or failed (FAIL) the Bechdel test
- `clean_test`: More detailed result (ok, dubious, men, notalk, nowomen)

Explore the data:

```
# View the structure
glimpse(bechdel90_13)
```

```
## Rows: 1,615
## Columns: 15
## $ year          <int> 2013, 2012, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 20~
## $ imdb          <chr> "tt1711425", "tt1343727", "tt2024544", "tt1272878", "tt0~
## $ title         <chr> "21 & Over", "Dredd 3D", "12 Years a Slave", "2 Guns", "~
## $ test          <chr> "notalk", "ok-disagree", "notalk-disagree", "notalk", "m~
## $ clean_test    <ord> notalk, ok, notalk, notalk, men, men, notalk, ok, ok, no~
## $ binary        <chr> "FAIL", "PASS", "FAIL", "FAIL", "FAIL", "FAIL", "FAIL", ~
## $ budget        <int> 13000000, 45000000, 20000000, 61000000, 40000000, 225000~
## $ domgross      <dbl> 25682380, 13414714, 53107035, 75612460, 95020213, 383624~
## $ intgross      <dbl> 42195766, 40868994, 158607035, 132493015, 95020213, 1458~
## $ code          <chr> "2013FAIL", "2012PASS", "2013FAIL", "2013FAIL", "2013FAI~
## $ budget_2013   <int> 13000000, 45658735, 20000000, 61000000, 40000000, 225000~
## $ domgross_2013 <dbl> 25682380, 13611086, 53107035, 75612460, 95020213, 383624~
## $ intgross_2013 <dbl> 42195766, 41467257, 158607035, 132493015, 95020213, 1458~
## $ period_code   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ decade_code   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
```

```
# See how many movies passed vs failed
bechdel90_13 %>%
  count(binary)
```

```
## # A tibble: 2 x 2
##   binary     n
##   <chr>  <int>
## 1 FAIL     862
## 2 PASS     753
```

How many movies passed the test? 753 movies

How many failed? 862 movies

Analysis

Binary Bechdel Test Results

Let's take a look at how median budget and gross vary by whether the movie passed the Bechdel test, which is stored in the `binary` variable.

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

```
## # A tibble: 2 x 4
##   binary med_budget med_domgross med_intgross
##   <chr>       <dbl>        <dbl>        <dbl>
## 1 FAIL    48385984.    57318606.    104475669
## 2 PASS    31070724     45330446.     80124349
```

What this code does: - `group_by(binary)`: Creates groups for PASS and FAIL - `summarise()`: Calculates summary statistics for each group - `median()`: Finds the middle value - `na.rm = TRUE`: Removes missing values before calculating

---

Exercise 2: Interpret Binary Results

Task 2.1: Which movies have higher median budgets: those that pass or fail the Bechdel test?

Your answer:

Movies that fail the test have higher median budgets.

Task 2.2: Which movies have higher median domestic gross: those that pass or fail?

Your answer:

Movies that fail the test have higher median domestic gross.

Task 2.3: What might explain these patterns?

Your answer:

I think that movies that fail the Bechdel test tend to have higher median budgets and higher median domestic gross because many movies that become popular follow tropes about men and only have women as supporting characters to those men. This could cause new directors to follow these same patterns.

Detailed Bechdel Test Results

Next, let's take a look at how median budget and gross vary by a more detailed indicator of the Bechdel test result. This information is stored in the `clean_test` variable, which takes on the following values:

- `ok` = passes test
- `dubious` = unclear whether it passes
- `men` = women only talk about men
- `notalk` = women don't talk to each other
- `nowomen` = fewer than two women

Exercise 3: Group by Detailed Test Results

Fill in the blank to group by `clean_test`:

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

```
## # A tibble: 5 x 4
##   clean_test med_budget med_domgross med_intgross
```

```
##   <ord>           <dbl>        <dbl>        <dbl>
## 1 nowomen     43373066     44891296.    89509349
## 2 notalk      56570084.    63890455    123102194
## 3 men         39737690.    56392786     99578022.
## 4 dubious     35790994     49173429     89883201
## 5 ok          31070724     45330446.    80124349
```

Task 3.1: Which category has the highest median budget?

Your answer: The "notalk" category has the highest median budget.

Task 3.2: Which category has the highest median international gross?

Your answer: The "notalk" category has the highest median international gross.

Return on Investment (ROI)

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.

ROI formula: (Total Gross) / Budget

Higher ROI = More profitable relative to what was spent

```
bechdel90_13 <- bechdel90_13 %>%
  mutate(roi = (intgross_2013 + domgross_2013) / budget_2013)
```

What this code does: - `mutate()`: Creates a new variable - `(intgross_2013 + domgross_2013)`: Total gross (domestic + international) - `/ budget_2013`: Divide by budget - Result: ROI of 2 means the movie made twice its budget

---

Exercise 4: Highest ROI Movies

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

```
bechdel90_13 %>%
  arrange(desc(roi)) %>%
  select(title, roi, year)
```

```
## # A tibble: 1,615 x 3
##    title                    roi  year
##    <chr>                  <dbl> <int>
##  1 Paranormal Activity     671.  2007
##  2 The Blair Witch Project  648.  1999
##  3 El Mariachi              583.  1992
##  4 Clerks.                  258.  1994
##  5 In the Company of Men    231.  1997
##  6 Napoleon Dynamite        227.  2004
##  7 Once                     190.  2006
##  8 The Devil Inside         155.  2012
##  9 Primer                   142.  2004
## 10 Fireproof                134.  2008
## # i 1,605 more rows
```

What this code does: - `arrange(desc(roi))`: Sorts by ROI in descending order (highest first) - `select()`: Shows only the columns we want

Task 4.1: What is the movie with the highest ROI?

Your answer: Paranormal Activity is the movie with the highest ROI in the dataset.

Task 4.2: What is its ROI value?

Your answer: It's ROI value is 671.

Task 4.3: Does this movie surprise you? Why or why not?

Your answer:

Yes, Paranormal Activity being the movie with the highest ROI in the dataset surprises me because it passes the Bechdel test and I have not heard of it before.

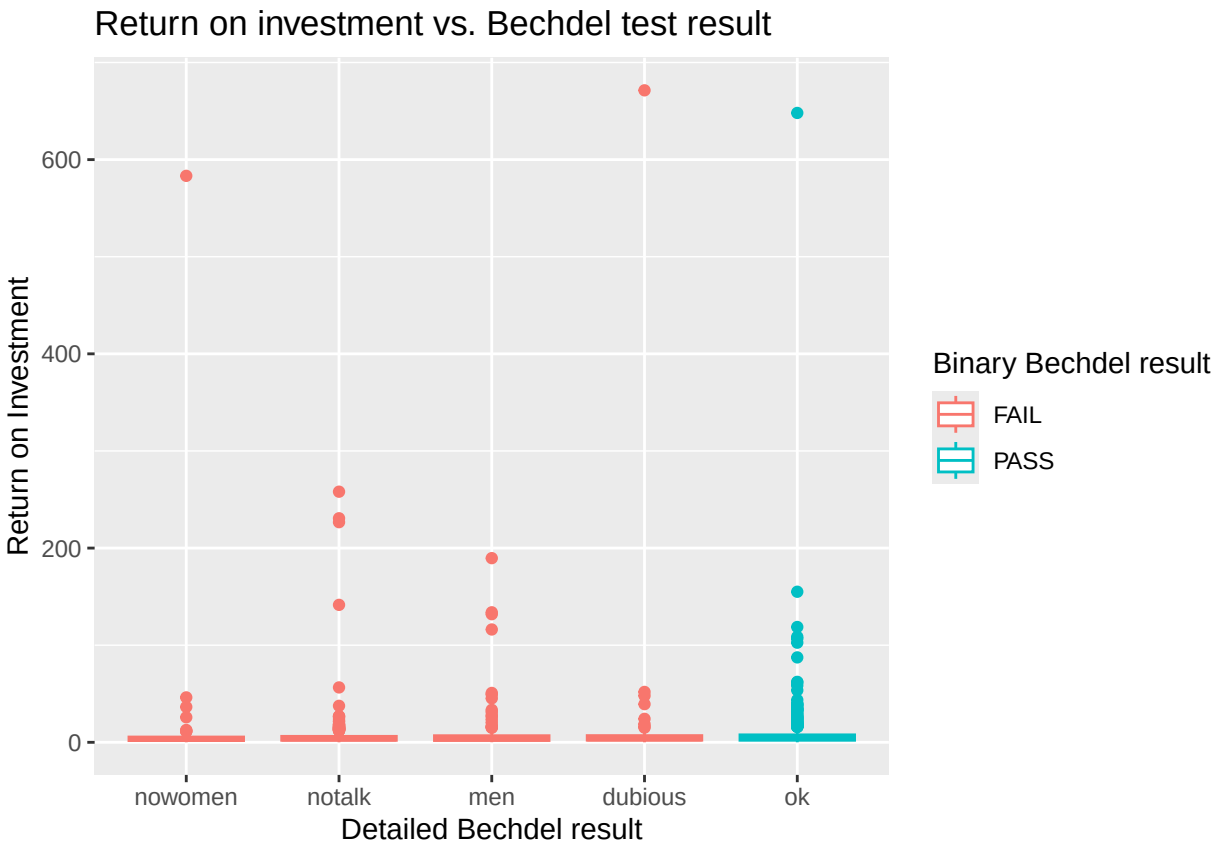
Visualizing ROI

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.

Exercise 5: Create Initial Boxplot

Fill in the blank for the y-axis label:

```
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 = "Return on Investment",
    color = "Binary Bechdel result"
  )
```



Hint: The y-axis shows the ROI values.

Task 5.1: Why is this plot difficult to interpret?

Your answer:

The plot is hard to interpret because there are movies with extremely high returns on investment that force the y axis scale to be too high.

Extreme Observations

What are those movies with *very* high returns on investment?

```
bechdel90_13 %>%
  filter(roi > 400) %>%
  select(title, budget_2013, domgross_2013, year)
```

```
## # A tibble: 3 x 4
##   title                   budget_2013 domgross_2013  year
##   <chr>                         <int>         <dbl> <int>
## 1 Paranormal Activity          505595     121251476  2007
## 2 The Blair Witch Project      839077     196538593  1999
## 3 El Mariachi                   11622       3388636  1992
```

Task 5.2: How many movies have ROI greater than 400?

Your answer: 3 movies have returns on investment greater than 400.

Task 5.3: What do you notice about the budgets of these high-ROI movies?

Your answer:

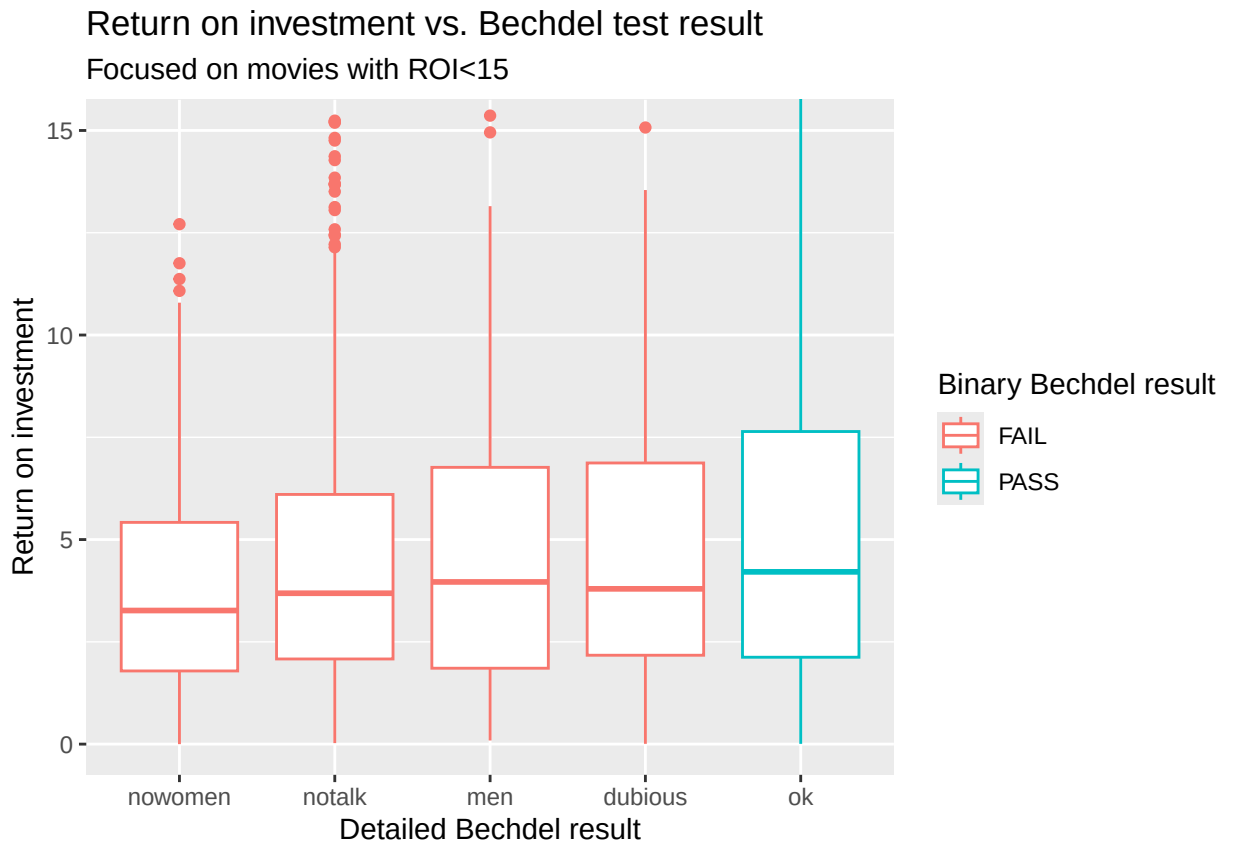
I notice that the budgets for these high return on investment movies are extremely low compared to the median budgets that we determined in Exercise 1. More specifically, the high-ROI movies have budgets less than 3% of the median budget.

Exercise 6: Zoom In for Better View

Zooming in on the movies with `roi < 15` provides a better view of how the medians across the categories compare.

Fill in the blanks:

```
ggplot(data = bechdel90_13,
       mapping = aes(x = clean_test, y = roi, color = binary)) +
  geom_boxplot() +
  labs(
    title = "Return on investment vs. Bechdel test result",
    subtitle = "Focused on movies with ROI<15 ", # Something about zooming in to a certain level
    x = "Detailed Bechdel result",
    y = "Return on investment",
    color = "Binary Bechdel result"
  ) +
  coord_cartesian(ylim = c(0, 15))
```



Hints for subtitle: - Mention that you're focusing on ROI < 15 - Something like "Focusing on movies with ROI less than 15" or "Zoomed in to ROI < 15"

What `coord_cartesian()` does: - Zooms in on the plot without removing data - `ylim = c(0, 15)`: Shows only y-values from 0 to 15 - Unlike `filter()`, this doesn't remove data, just changes the view

---

Exercise 7: Interpret the Zoomed Plot

Task 7.1: Looking at the zoomed plot, which category has the highest median ROI?

Your answer: According to the box plot, movies that have an "ok" clean_test score have the highest median return on investment.

Task 7.2: Do movies that pass the Bechdel test (binary = PASS) appear to have higher or lower ROI than those that fail?

Your answer:

Movies that pass the Bechdel test appear to have higher returns on investment than those that fail the Bechdel test.

Task 7.3: What does this suggest about the financial argument for including women in films?

Your answer:

This data suggests that including women in films has an average higher return on investment, meaning that it is a smarter financial choice to include women in films that talk to each other about subjects other than men.

Exercise 8: Create Your Own Summary

Task 8.1: Calculate the mean ROI (not median) for movies that pass vs. fail the Bechdel test.

Hint: Use `group_by()` and `summarise()` like we did earlier, but with `mean()` instead of `median()`.

```
bechdel90_13 %>%
  group_by(binary) %>%
  summarise(
    mean_roi = mean(roi, na.rm = TRUE)
  )
```

```
## # A tibble: 2 x 2
##   binary mean_roi
##   <chr>     <dbl>
## 1 FAIL       8.36
## 2 PASS       7.99
```

Mean ROI for PASS: 7.99

Mean ROI for FAIL: 8.36

Task 8.2: How do these means compare to the medians we saw earlier? Why might they be different?

Your answer:

The mean ROIs of the movies in the dataset were higher than the medians that we saw earlier. I think that they are different because the mean is more impacted by the very large outliers that we saw earlier in Exercise 5 that had ROIs greater than 15.

Exercise 9: Challenge (Optional)

Create a visualization comparing ROI across years for movies that pass vs. fail the test.

Hints: - Use `geom_point()` or `geom_line()` - Put `year` on the x-axis and `roi` on the y-axis - Color by `binary` - You might want to filter out extreme ROI values first - Consider using `geom_smooth()` to show trends

```
# Your code here
```

What trend do you notice?

Knit and Submit

Before you finish:

1. Make sure all code chunks run without errors
2. Fill in all answer spaces
3. Knit to HTML first to verify everything works
4. Review the HTML output

Git workflow:

1. Click the **Git** pane
2. Check boxes next to all changed files
3. Click **Commit**
4. Write a commit message: "Completed AE 02 - Bechdel"
5. Click **Commit**
6. Click **Push**
7. Verify your work is on GitHub in **YOUR fork**

Generate PDF for Canvas Submission

Once everything works in HTML:

1. Click the **Knit** dropdown arrow
2. Select **Knit to PDF**
3. Wait for PDF generation
4. Find `ae-02-bechdel.pdf` in the Files pane
5. Click the checkbox next to the PDF
6. Click **More → Export...**
7. Save to your computer
8. Submit the PDF to Canvas

Remember: Submit BOTH to GitHub (your .Rmd) AND Canvas (the PDF)!

Reflection Questions

Task: After completing this exercise, reflect on what you learned.

What was the most interesting finding about the Bechdel test and movie finances?

What was the most challenging part of this exercise?

How did zooming in on the plot change your interpretation of the data?

References

1. Hickey, W. (2014). "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women". FiveThirtyEight.
2. The Bechdel Test was popularized by Alison Bechdel's comic strip "Dykes to Watch Out For."

Troubleshooting

Common issues:

- **Error: "could not find function":** Make sure you ran the package loading chunk
- **ROI values look wrong:** Check that you're adding domestic and international gross before dividing by budget
- **Plot doesn't show:** Make sure you ran all previous chunks that create variables
- **NA values in results:** Use `na.rm = TRUE` in your summary functions

PDF issues:

- **LaTeX not found:** Install tinytex: `tinytex::install_tinytex()`
- **PDF looks different:** This is normal - focus on content, not formatting

If you're stuck:

- Review the walkthrough video
- Check your code against the examples
- Ask on the discussion board
- Attend office hours