Gender Bias in the Movie Biz

Put your name here

Put due date here

### Overview

In 1985, cartoonist [Alison Bechdel](http://dykestowatchoutfor.com/) proposed "The Rule." To pass, a movie has to sastisfy three basic requirements:

1. It has to have at least two women in it,
2. The two women have to talk to each other, and
3. They have to talk to each other about something besides a man.

"The Rule" (see a copy of the original comic strip [here](http://www.npr.org/templates/story/story.php?storyId=94202522)) is commonly refered to as the Bechdel test. It's a seemingly low bar, and it's surprising how many films **fail** the test.

In 2014, [FiveThirtyEight](http://fivethirtyeight.com/) analyzed 1,615 films between 1990 and 2013 to explore the financial effect of a these films' portrayal of women in an article titled [The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women](http://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/). Their analysis relied on data sets from [BechdelTest.com](http://bechdeltest.com/) and [The-Numbers.com](http://www.the-numbers.com/). They concluded that the "the median budget of movies that passed the test - those that featured a conversation between two women about something other than a man — was substantially lower than the median budget of all films in the sample." Overall, movies that passed the test may have a better financial return than those that don't.

### The data

The data set is available for download from FiveThirtyEight's GitHub repository: <https://raw.githubusercontent.com/fivethirtyeight/data/master/bechdel/movies.csv>. (To take a look at the raw data, copy and paste this address into your web browser.) We've read in the data and saved it as the R object movies. Here are the first three rows of the dataframe:

## # A tibble: 3 x 15  
## year imdb title test clean\_test binary  
## <int> <chr> <chr> <chr> <chr> <chr>  
## 1 2013 tt1711425 21 &amp; Over notalk notalk FAIL  
## 2 2012 tt1343727 Dredd 3D ok-disagree ok PASS  
## 3 2013 tt2024544 12 Years a Slave notalk-disagree notalk FAIL  
## # ... with 9 more variables: budget <int>, domgross <chr>, intgross <chr>,  
## # code <chr>, `budget\_2013$` <int>, `domgross\_2013$` <chr>,  
## # `intgross\_2013$` <chr>, `period code` <int>, `decade code` <int>

The original data set includes movies from 1970 to 2013. We want to make sure we only include movies that came out *after* 1990. Once you clean the data for limit to movies from later than 1990, the last few rows in the year and title columns will look like this:

## # A tibble: 3 x 2  
## year title  
## <int> <chr>  
## 1 1991 The Naked Gun 2 1/2: The Smell of Fear  
## 2 1991 The Silence of the Lambs  
## 3 1991 White Fang

You should also clean up the data a bit, including renaming a few of the column names to make them easier to work with. Here are the column names of the original dataset:

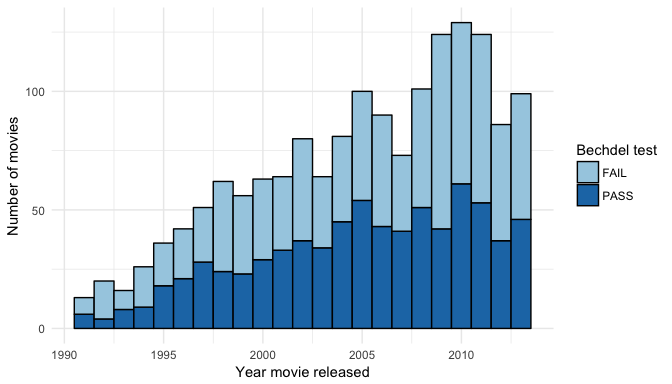
## [1] "year" "imdb" "title" "test"   
## [5] "clean\_test" "binary" "budget" "domgross"   
## [9] "intgross" "code" "budget\_2013$" "domgross\_2013$"  
## [13] "intgross\_2013$" "period code" "decade code"

To clean the data, (1) rename "binary" to "bechdel", "budget\_2013$" to "budget\_2013", "domgross\_2013$" to "domgross\_2013", and "intgross\_2013$" to "intgross\_2013", (2) limit the dataframe to only include the variables "bechdel", "year", "clean\_test", "budget\_2013", "domgross\_2013", "intgross\_2013", and "title", and (3) change "intgross\_2013" and "domgross\_2013" to have a numeric class (use as.numeric). After that change, the first three rows of the dataframe will look like this:

## # A tibble: 3 x 7  
## bechdel year clean\_test budget\_2013 intgross\_2013 domgross\_2013  
## <chr> <int> <chr> <int> <dbl> <dbl>  
## 1 FAIL 2013 notalk 13000000 42195766 25682380  
## 2 PASS 2012 ok 45658735 41467257 13611086  
## 3 FAIL 2013 notalk 20000000 158607035 53107035  
## # ... with 1 more variables: title <chr>

### "The Rule"

The bechdel variable specifies whether a movie passed or failed the Bechdel rule. Here is a plot of how the total and relative number of movies in this dataset that have failed and passed the test changed over time:



The clean\_test variable offers a finer detail of the Bechdel Test than Fail vs. Pass. Here are the different levels of this variable:

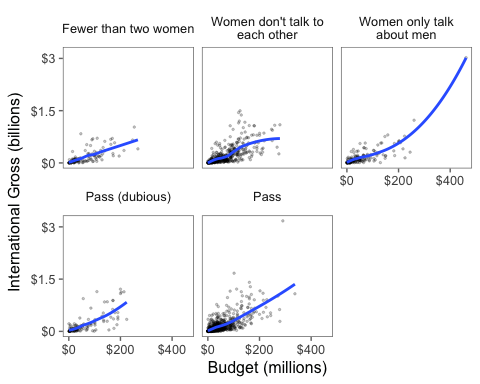
## [1] "notalk" "ok" "men" "nowomen" "dubious"

Based on FiveThirtyEight's [article](http://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/), these levels correspond to the following designations:

|  |  |  |
| --- | --- | --- |
| clean\_test | Category\_Description | Bechdel\_Test |
| nowomen | Fewer than two women | Fail |
| notalk | Women don't talk to each other | Fail |
| men | Women only talk about men | Fail |
| dubious | Dubious | Pass |
| ok | Passes Bechdel Test | Pass |

Next, for the purposes of later plotting, reorder the levels of clean\_test so that they are in the order given in the above table, and rename them to more verbose descriptions: "Fewer than two women", "Women don't talk to each other", "Women only talk about men", "Pass (dubious)", and "Pass", for "nowomen", "notalk", "men", "dubious", and "ok", respectively. You may want to use "\n" to make cleaner label breaks in plots for the second and third categories.

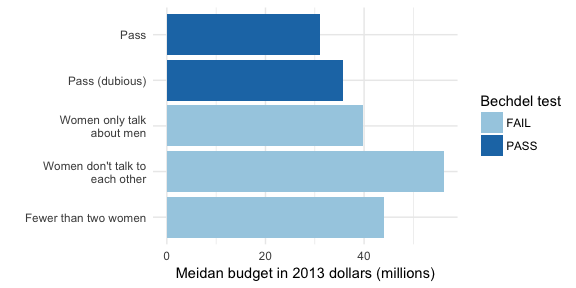
Here is a plot of the association between movie budget and international gross separately for each of these categories:



We can check out the median budget in 2013 dollars for movies fitting into each of the categories of the Bechdel test. First, create a dataframe called med\_budget with the median value of budget\_2013 for each movie using dplyr's group\_by() and summarize() functions. Change the dollar amount so that its units are in millions of dollars, and add a column called bechdel with "PASS" or "FAIL" values for each category ("FAIL" for the first three, "PASS" for the other two).

## # A tibble: 5 x 3  
## clean\_test median\_budget\_2013 bechdel  
## <fctr> <dbl> <chr>  
## 1 Fewer than two women 44.01686 FAIL  
## 2 Women don't talk to  
## each other 56.17717 FAIL  
## 3 Women only talk  
## about men 39.73769 FAIL  
## 4 Pass (dubious) 35.79099 PASS  
## 5 Pass 31.11689 PASS

Now, following [FiveThirtyEight's](http://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/) lead, plot this dataframe as a bar graph:

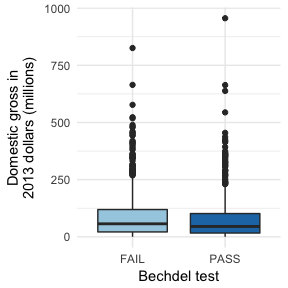


We can see that past movies that have failed the Bechdel test had higher median budgets.

### Extra challenge-- not graded

What about how much money these movies made? In FiveThirtyEight's analysis, they focused on "Return on investment", which involved dividing movie profits (domgross\_2013, for example) by movie budgets. For simplicity's sake, let's compare domestic gross for movies that passed and failed the test. Again, we should mutate domgross\_2013 so that it is in units of millions of dollars.

Here's a boxplot comparing domestic gross for movies that failed and passed the Bechdel test:



With a lot of outliers, it looks like the mean domestic gross for movies that failed is *slightly* higher than that for movies that passed.

Let's find out what the highest grossing movies in "Pass" and "Fail" are. For failed, it looks like the higest grossing movie in this data set that failed the Bechdel test earned $825.7 million (in 2013 dollars):

## # A tibble: 1 x 2  
## title domgross\_2013  
## <chr> <dbl>  
## 1 Avatar 825.7072

And its title is "Avatar"!

For passed, it looks like "Titanic" made $955.9 million (in 2013 dollars).

## # A tibble: 1 x 2  
## title domgross\_2013  
## <chr> <dbl>  
## 1 Titanic 955.8904

To wrap up, add both of these titles as interesting labels to the plot. (Note: to match the lables shown here, use size = 4 and hjust = 1.2 in your geom\_text() addition.)

