Bechdel Tsst

[YOUR NAME]

[DATE]

In this mini analysis we work with the data used in the 2014 FiveThirtyEight story titled "The Dollar-And-Cents Case Against Hollywood's 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 (US) gross revenue in 2013 inflation adjusted dollars
- `intgross_2013`: Total interational (i.e., worldwide including US) gross revenue in 2013 inflation adjusted dollars

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

Analysis

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

```
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
```

Next, let's take a look at how median budget and gross revenue 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_domgross = median(domgross_2013, na.rm = TRUE),
            med_intgross = median(intgross_2013, na.rm = TRUE))
```

```
## # A tibble: 1 x 3
##   med_budget med_domgross med_intgross
##        <int>        <dbl>        <dbl>
## 1   37878971     52270207     93523336
```

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 total international gross revenue to the budget.

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

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

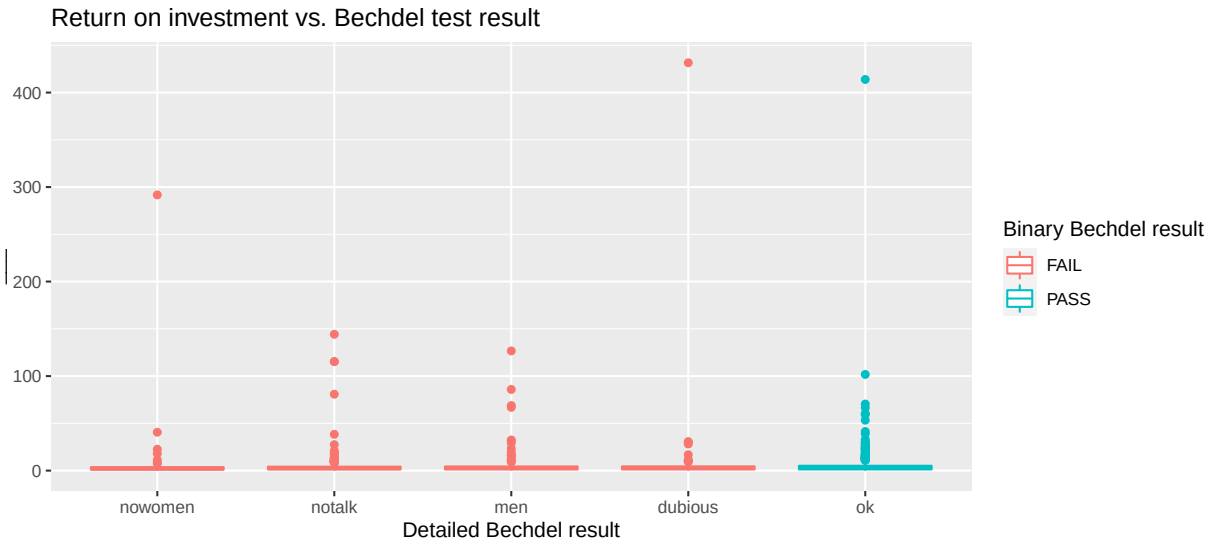
```
bechdel90_13 %>%
  arrange(desc(roi)) %>%
  select(title, clean_test, binary, roi, budget_2013, intgross_2013)
```

```
## # A tibble: 1,615 x 6
##    title                   clean_test binary   roi budget_2013 intgross_2013
##    <chr>                   <ord>      <chr>  <dbl>       <int>         <dbl>
##  1 Paranormal Activity     dubious    FAIL   432.       505595     218173082
##  2 The Blair Witch Project ok         PASS   414.       839077     347238122
##  3 El Mariachi             nowomen    FAIL   292.        11622       3390310
##  4 Clerks.                 notalk     FAIL   144.        42435       6120440
##  5 Once                    men        FAIL   127.       173369      21956864
##  6 In the Company of Men   notalk     FAIL   115.        36281       4184879
##  7 Napoleon Dynamite       notalk     FAIL   115.       493277      56878201
##  8 The Devil Inside        ok         PASS   102.      1014639     103248087
##  9 Saw                     men        FAIL    85.9     1479831     127137678
## 10 Primer                  notalk     FAIL    80.8        8632        697797
## # ... 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 = "___",
       color = "Binary Bechdel result")
```

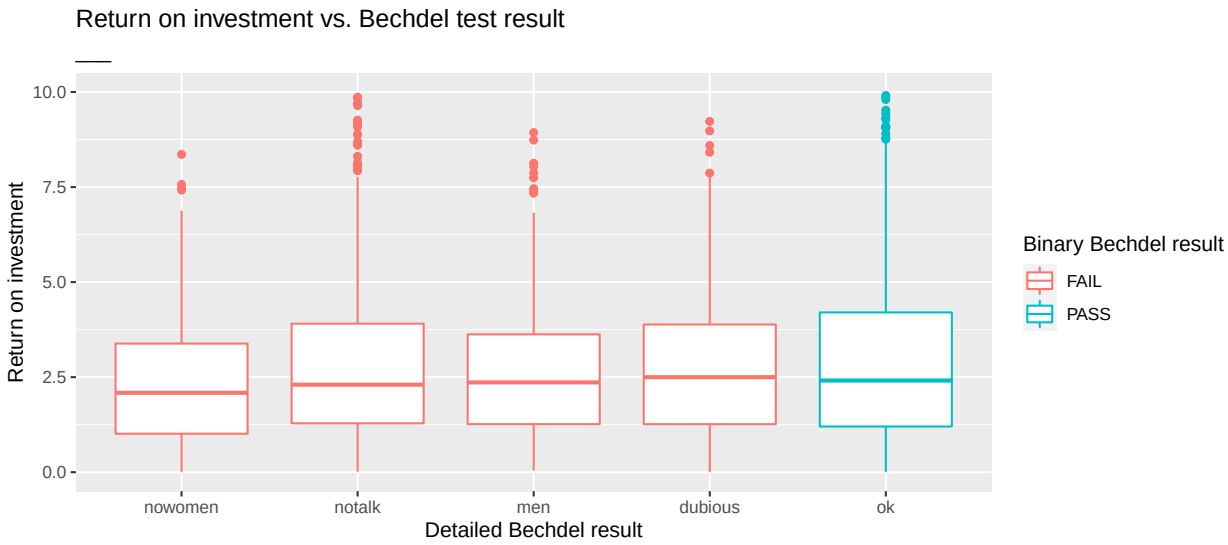
```
## Warning: Removed 8 rows containing non-finite values (stat_boxplot).
```



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 121 rows containing non-finite values (stat_boxplot).
```



- What are the advantages to each plot? What are the disadvantages to each plot?