*Data Exploration*

**The Bechdel Test**

In 1985, cartoonist Alison Bechdel\*\* suggested a method of measuring gender bias in works of fiction.

A work passes the ‘Bechdel Test’ if:

*“…Take “American Hustle.” The film passes the test, but only barely, thanks to a single scene where a con artist’s wife, played by Jennifer Lawrence, discusses nail polish with a politician’s wife, played by Elisabeth Röhm. On the other hand, “Gravity” — which is dominated almost entirely by Sandra Bullock, in a highly praised performance — fails the test, as Bullock never speaks to another woman in the film….”*

* It has at least two women,
* Who have at least one conversation,
* That is about something other than a man.

The test is far from perfect (see the quote on the right, from an article referenced at the end of this activity), but it does give an approximate assessment, and it has gained in popularity over the years. Committed moviegoers have analyzed hundreds of films, assessing whether the criteria are satisfied (*BechdelTest.com*). The site *fivethirtyeight.com* has taken this data and linked it with financial data from *The-Numbers.com*, resulting in the data we will use here. Before we look at the data, we’d like you to think about your expectations and intuitions regarding some of the it.

**\*\* Note: When you’re asked for ‘Intuitions, it is not expected that you do calculations or research or that your answers will be exactly correct. The point is for you to explore what you know already, and to prepare you for a more thorough investigation later in the activity.**

**Intuitions**

1. What percentage of films would you expect to pass the test?

Very little.

1. We will look at films from the 70’s, 80’s, 90’s, 00’s, and 10’s. Do you think the percentage of films that pass the test has changed over the decades? If so, how?

I hope it has increased but I think only by a little.

1. Would you expect that films that pass the test have larger or smaller budgets? Explain.

Um smaller?, like indie movies

1. Would you expect that films that pass the test have larger or smaller profits? Explain.

Smaller profits (?)

\*\*Fun facts: Alison Bechdel wrote a graphic novel that was adapted for the stage (*Fun Home*), and the play won the Tony award for Best Musical of 2015. Alison lives in Bolton, Vermont.

Now, we’ll go on to the data. From our class Blackboard site, download to your laptop the file **bechdel87.csv**. There are basically three types of variables:

**General film description:** *year, decade, imdb#, title*

**Bechdel test:** *Btest* (pass or fail), *Btest5* (which criteria were satisfied)

**Money data:** Budget and Gross earnings – domestic and international in 2013 dollars, Percent profit domestic and international

**Rating Data:** *averating* is the average star rating (on a scale from 0 to 5), based on crowd-sourced data from MovieLens.org. The last variable is the number of people who rated this film; we used average ratings only for movies that had been rated by at least 5 people.

Read the file into R as a data frame called **bech**. We’ll look at two types of data: Quantitative and Categorical.

**Categorical Data:**

1. When individual values are ‘categories’ that represent a characteristic of data that is not inherently numeric, we say they are **categorical** data. For example, your major in college or your state of residence are both categorical variables. We’ll look plots of **Decade**, **BTest**, and **BTest5**.
2. Using ggplot, make a bar graph of **Decade**, and sketch the trend below. Why do you think the number increases over the decades? Why do you think the last bar is lower?
3. Using ggplot, make a bar graph of **BTest.**  What proportion of films pass the Bechdel Test. Is this what you predicted?
4. Using ggplot, make a bar graph of **BTest5**. Here, you can see the reasons for failure to pass the test:

**Nowomen:** film had no women

**Notalk:** film had women, but they did not speak with each other

**Men:** film had women who spoke, but they only spoke about men

**Dubious:** film may pass, but there is some disagreement. (for example, see <http://bechdeltest.com/view/157/mulan/> for debate on whether Mulan passes the test)

Of the four reasons above, which is most frequent?

**Quantitative Data:**

When individual values have numeric meaning – often, a ‘quantity’ of something – we say they are **quantitative** data. For example, your height, or the number of credit hours you’re taking are both quantitative variables.

1. Use ggplot to create a histogram, and then a density plot of **budget13** -- on which each value represents the cost of making of a particular film, in millions of dollars. The amounts have been adjusted to 2013 dollars, to account for inflation. (That is, since everything was cheaper in, say, 1970, they’ve calculated the amount it would have cost in 2013. This way, all of the numbers can be compared fairly).

Sketch or paste the plot of the budget data below, labelling the axes.

1. Use ggplot to create a histogram, and then a density plot of **domgross13**. This plot shows the amount of money that the film made overall, **in the U.S**., also adjusted to 2013 dollars. (It does not account for film expenses – it is “gross” earnings.) How does its shape compare to the plot of budget?
2. Use ggplot to create a histogram, and then a density plot of **intgross13**. This plot shows the amount of money that the film made **internationally**, also adjusted to 2013 dollars. How does its shape compare to the previous plots? What data transformation might be useful for these three variables?
3. Use ggplot to create a histogram, and then a density plot of **averating**. Describe the shape. Why does it have this shape?

**RELATIONSHIPS AMONG VARIABLES**

**Two Categorical Variables:**

1. For example, we’d like to know if the proportion of films passing the Bechdel test has changed over time. Create a bar graph showing **BTest** for each **Decade**. How have the proportion of films that PASS change over time? Is this what you expected? Explain
2. Create a bar graph showing **BTest5** for each **Decade**.

**One Quantitative and One Categorical Variable:**

1. Do films that pass the test tend to get higher or lower ratings?

Create boxplots or density plots of **averating** for Passing and Failing films (by **BTest)**. Using summarise and group\_by, calculate summary statistics of **averating** for Passing and Failing films (by **BTest)**.

Write down the mean, median, and standard deviation of ratings for each group below.

PASS FAIL

Mean rating: \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

Median rating: \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

Standard deviation: \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

Do you see a noticeable difference in ratings for the films that pass and those that fail?

1. Do films that pass the Bechdel test tend to have higher or lower budgets? Create boxplots or density plots, then use summarise and group\_by to calculate summary statistics of **Budget13** for Passing and Failing films (by **BTest)**.

Write down the mean, median and standard deviation of the budget (in millions) for each group below.

PASS FAIL

Mean budget: \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

Median budget: \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

Standard deviation: \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

1. Do films that pass the test appear to have higher or lower budgets? Why are the mean values so much higher than the median values? Do you think that **means** or **medians** would be better to use for this data? The standard deviations are large – would the 2/3 rule apply? Explain.
2. Following the same procedure, create boxplots or density plots, then use summarise and group\_by to find summary statistics of gross earnings (in millions), **domgross13,** by **BTest** and write means and medians and sds below. Do films that pass the test tend to have higher or lower earnings?

PASS FAIL

Mean domgross: \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

Median domgross: \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

Standard deviation: \_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_

1. But wait! If films that pass the Bechdel test have lower earnings, is this likely related to passing the test (having women in them)? These same films had lower budgets. Do films with lower budgets tend to make less money than films with higher budgets? What is your intuition?

**Two Quantitative Variables**

1. Let’s determine whether films with larger budgets tend to have larger profits. Create a scatterplot with **Budget13** on the X axis, and **domgross13 (**domestic gross income) on the Y axis. Add the layer: geom\_smooth()

Sketch or paste the plot below. Does the smoothed curve show a trend in the data? Compared to films with smaller budgets, do films with larger budgets tend to make more or less or about the same? Do the films fall on a perfect trend, or is there some variation from film to film?

**Final Question:**

Some suggest that an explanation for the low percentage of films passing the test is that most writers, directors, and producers are men. It has also been suggested that films with more women simply do not make as much money. We’ve seen that, indeed, films that pass the test have lower gross earnings.

However, this could be simply because they are lower budget films, on average, and lower budget films tend to make less money. Can we somehow ‘correct’ for budget in looking at profit? Here’s one way:

1. Using mutate, create a column in the data called **domprofit13**. This is the amount of profit, expressed as a percentage of the budget:

**100% x (domestic gross – budget) / budget**

If a film lost money, this value would be negative. If it broke even (made as much as it spent), it would be zero. If the film spent $10 million, and made $15 million, its value would be 100% x (15 – 10) / 10 = 50%.

Hand-check a few values of **domprofit13**, to make sure your computation worked.

1. We will now compare **domprofit13** for films that do and do not pass the test. Using ggplot, calculate density plots by BTest, either overlaying the density plots, or using facets. Sketch or paste the plots below.
2. Using summarise and group\_by, calculate the median domprofit13 for films that Pass and Fail, and record below. (We will use only medians, because they are more appropriate for very skewed data, such as this.)

Median % domestic profit: PASS: \_\_\_\_\_\_\_\_\_\_ FAIL: \_\_\_\_\_\_\_\_\_\_

1. Repeat 16, 17, 18 for international profits: **intprofit13**, and write down the median percentage **international** profit for each group below.

Median % international profit: PASS: \_\_\_\_\_\_\_\_\_\_ FAIL: \_\_\_\_\_\_\_\_\_\_

1. What is your conclusion? Do films that pass the Bechdel test have a higher or lower percentage profit, compared to films that do not? What do you think is the reason for the low percentage of films that pass the test?
2. To follow up, please read the article “*The Dollar-And-Cents Case Against Hollywood’s Exclusion of Women*,” on **fivethirtyeight.com**. (<https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/>)