# Elementary Statistics – Exploratory Data Analysis (EDA) Pt 2

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

# Reminder!

- First quiz is out today! (On PLATO)
- Quizzes and homeworks are *week long* assignments; expect to spend 5-7 hours on them (this is standard for a college class)

# Plan for Today

- EDA for Two Variables
  - Categorical & Categorical
  - Categorical and Numeric
  - Numeric and Numeric

# Warm Up / Recap: Describing Distributions for Continuous Variables

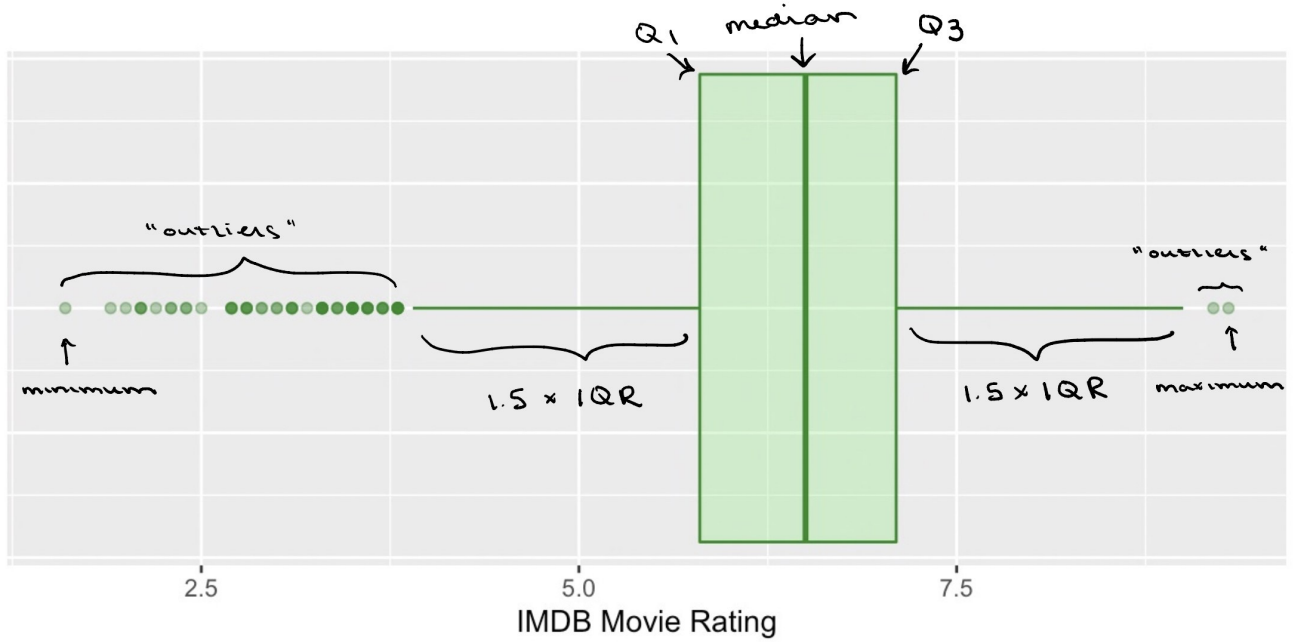
When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

- Graphical
  - Skewness
  - Modality
- Summary Statistics
  - Center
  - **Spread**

The following five statistics make up the five-number summary, which captures information about both the center *and* spread of the data:

Minimum / 25th percentile (Q1) / Median / 75th percentile (Q3) / Maximum

We can use a box plot to visualize all of these statistics in one go:



## Practice

Generate a box plot to summarize `Hours`:

| Hours |
|---|
| 4.5 |
| 4 |
| 6.25 |
| 9 |
| 1.75 |
| 8 |
| 3 |
| 8.33 |
| 9.67 |
| 0 |

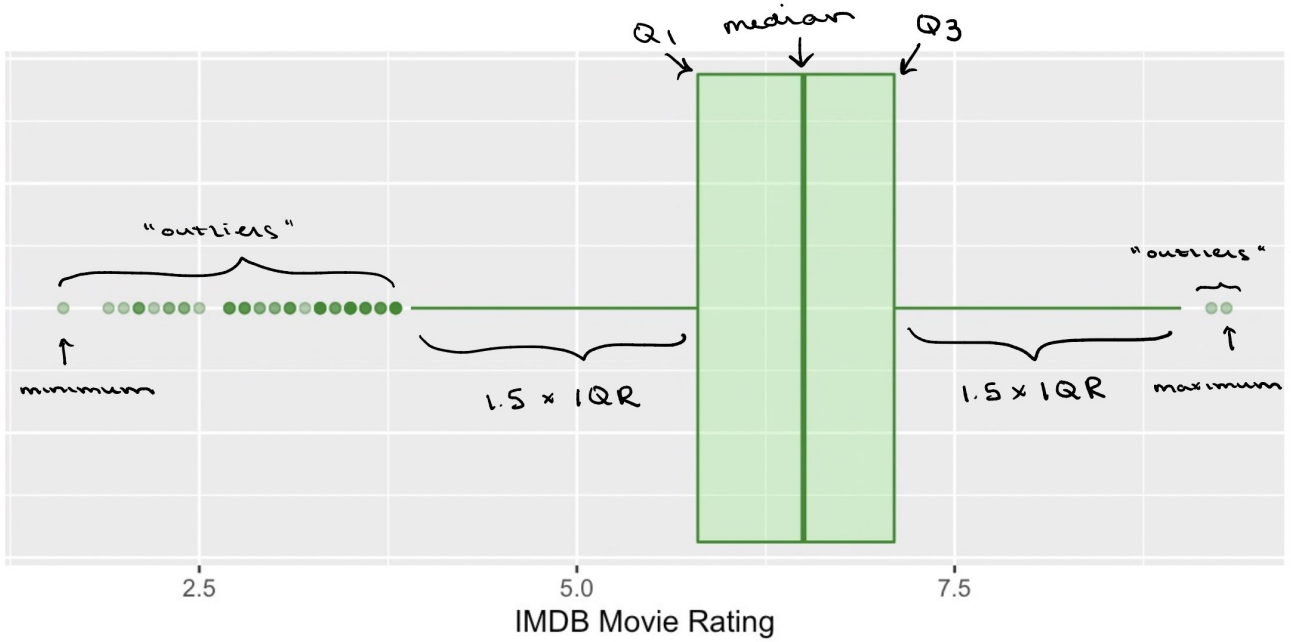
When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

- Graphical
  - Skewness
  - Modality
- Summary Statistics
  - Center
  - **Spread**

The following five statistics make up the five-number summary, which captures information about both the center *and* spread of the data:

Minimum / 25$^{th}$ percentile (Q1) / Median / 75$^{th}$ percentile (Q3) / Maximum

We can use a box plot to visualize all of these statistics in one go:



# 20 Minute Activity: EDA Practice

Open movies.csv (under Demos on the course website) in excel or google sheets.

Work with 1-2 other people.

Choose 1 categorical and 1 numerical variable. For each variable, generate the appropriate summary visualizations and summary statistics.

You in some cases, you will need to manipulate the raw data and use formulas. Helpful tips can be found here:

- Excel
  - https://www.princeton.edu/~otorres/Excel/excelstata.htm
  - https://statisticsbyjim.com/basics/descriptive-statistics-excel/
- Google Sheets
  - http://www.comfsm.fm/~dleeling/statistics/text6.html#page-031
  - https://www.groovypost.com/howto/quickly-get-column-statistics-in-google-sheets/

# Big Picture

Last time, we discussed how we might use both numbers and visuals to summarize individual variables in our dataset:

## Categorical variables

- Bar plots and frequency tables

## Numerical variables

- Histograms and density plots ⇒ the *distribution* of the variable
- Statistics like the mean and standard deviation ⇒ *center* and *spread*

# Big Picture

Last time, we discussed how we might use both numbers and visuals to summarize individual variables in our dataset:

## Categorical variables

- Bar plots and frequency tables

## Numerical variables

- Histograms and density plots ⇒ the *distribution* of the variable
- Statistics like the mean and standard deviation ⇒ *center* and *spread*

What if we want to use EDA understand relationships between variables?

# Relationships Between Two Categorical Variables

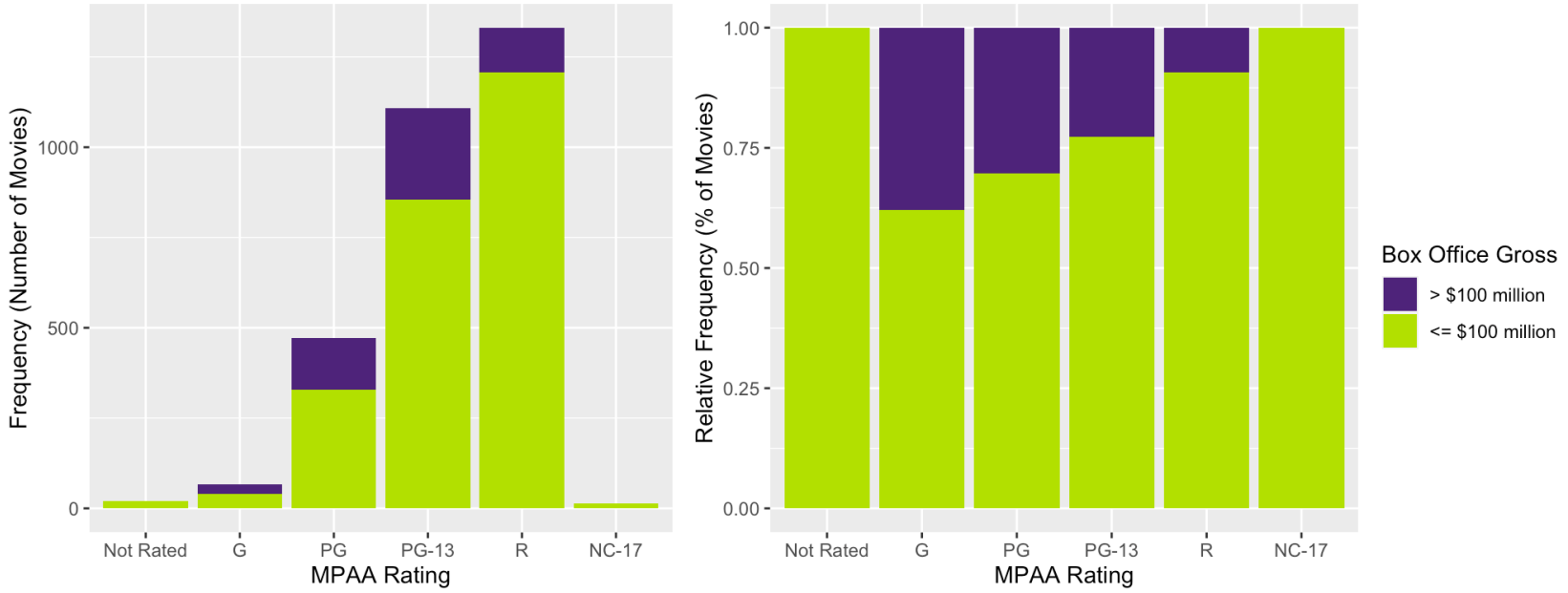
# Data Visualizations: Stacked Bar Plots

Suppose we want to understand the relationship between a movie's MPAA rating and whether it grosses more than $100 million at the box office

→ How does the distribution of movies with large versus small to moderate box office earnings differ based on MPAA rating?

We can use a stacked barplot!

- Each bar in a standard barplot is divided into stacked sub-bars, each corresponding to the level of the second categorical variable



## Practice

Show how the distribution of high versus low hours differs based on activity.

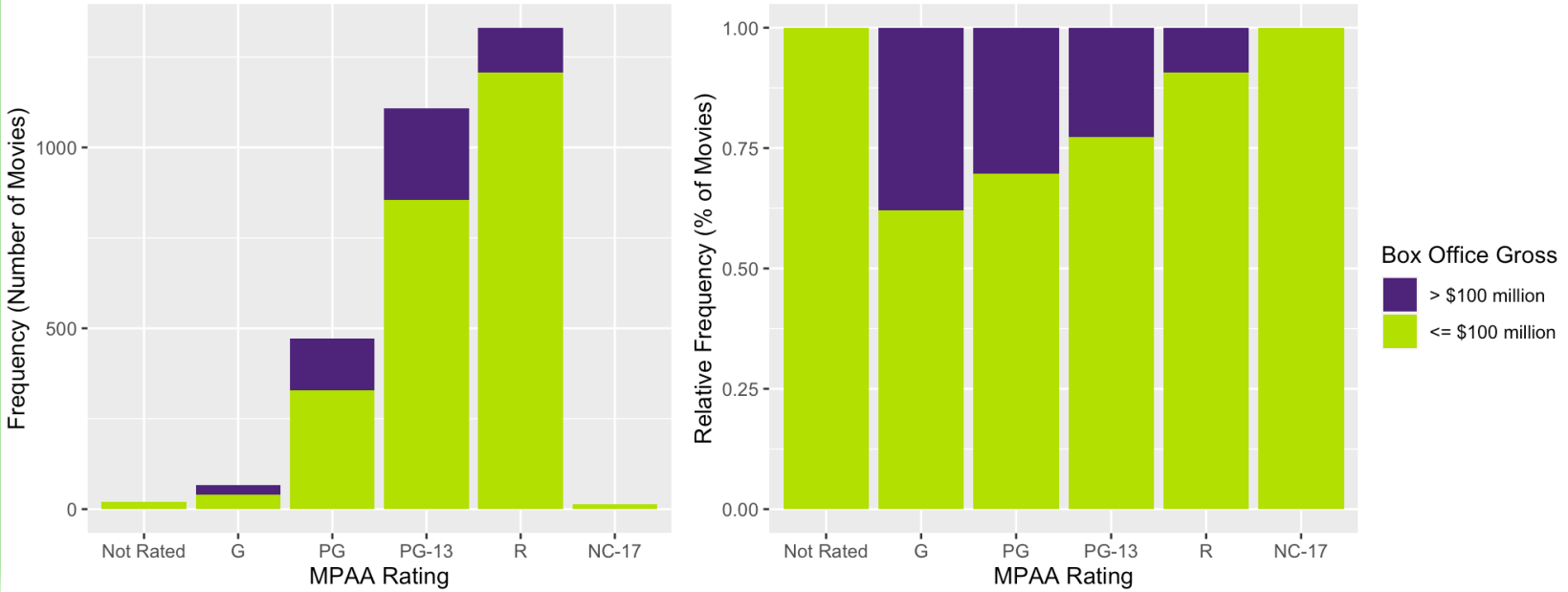
| Hours | Activity |
|---|---|
| low | Exercise |
| low | Exercise |
| high | Work |
| high | Chores |
| low | Work |
| high | Work |
| low | Driving |
| high | Exercise |
| high | Work |
| low | Driving |

Suppose we want to understand the relationship between a movie's MPAA rating and whether it grosses more than $100 million at the box office

→ How does the distribution of movies with large versus small to moderate box office earnings differ based on MPAA rating?

We can use a stacked barplot!

- Each bar in a standard barplot is divided into stacked sub-bars, each corresponding to the level of the second categorical variable



# Summary Statistics: Contingency Tables

Just as a frequency table contains the same information as a univariate barplot, we can use a contingency table to numerically summarize the distribution of two categorical variables!

- Displays the number of observations falling into each unique combination of levels for the two variables:

| | Box Office Gross | | |
|---|---|---|---|
| **MPAA Rating** | Low | High | Total |
| Not Rated | 21 | 0 | 21 |
| G | 41 | 25 | 66 |
| PG | 328 | 143 | 471 |
| PG-13 | 856 | 252 | 1108 |
| R | 1207 | 124 | 1331 |
| NC-17 | 13 | 0 | 13 |
| Total | 2466 | 544 | 3010 |

We can use these tables to glean a lot of information about our two variables!

# Summary Statistics: Contingency Tables

Just as a frequency table contains the same information as a univariate barplot, we can use a contingency table to numerically summarize the distribution of two categorical variables!

- Displays the number of observations falling into each unique combination of levels for the two variables:

| | Box Office Gross | | |
|---|---|---|---|
| **MPAA Rating** | Low | High | Total |
| Not Rated | 21 | 0 | 21 |
| G | 41 | 25 | 66 |
| PG | 328 | 143 | 471 |
| PG-13 | 856 | 252 | 1108 |
| R | 1207 | 124 | 1331 |
| NC-17 | 13 | 0 | 13 |
| Total | 2466 | 544 | 3010 |

**Marginal Distribution**

66 of the movies in our dataset are rated G:

$$p_G = \frac{66}{3010} = 2.2\%$$

# Summary Statistics: Contingency Tables

Just as a frequency table contains the same information as a univariate barplot, we can use a contingency table to numerically summarize the distribution of two categorical variables!

- Displays the number of observations falling into each unique combination of levels for the two variables:

| | Box Office Gross | | |
|---|---|---|---|
| **MPAA Rating** | Low | High | Total |
| Not Rated | 21 | 0 | 21 |
| G | 41 | 25 | 66 |
| PG | 328 | 143 | 471 |
| PG-13 | 856 | 252 | 1108 |
| R | 1207 | 124 | 1331 |
| NC-17 | 13 | 0 | 13 |
| Total | 2466 | 544 | 3010 |

Marginal Distribution

544 of the movies in our dataset were high box office earners:

$$p_{high} = \frac{544}{3010} = 18\%$$

# Summary Statistics: Contingency Tables

Just as a frequency table contains the same information as a univariate barplot, we can use a contingency table to numerically summarize the distribution of two categorical variables!

- Displays the number of observations falling into each unique combination of levels for the two variables:

| | Box Office Gross | | |
|---|---|---|---|
| **MPAA Rating** | Low | High | Total |
| Not Rated | 21 | 0 | 21 |
| G | 41 | 25 | 66 |
| PG | 328 | 143 | 471 |
| PG-13 | 856 | 252 | 1108 |
| R | 1207 | 124 | 1331 |
| NC-17 | 13 | 0 | 13 |
| Total | 2466 | 544 | 3010 |

Joint Distribution

25 of the movies in our dataset are rated G and were high box office earners:

$$p_{GandHigh} = \frac{25}{3010} = 0.08\%$$

# Summary Statistics: Contingency Tables

Just as a frequency table contains the same information as a univariate barplot, we can use a contingency table to numerically summarize the distribution of two categorical variables!

- Displays the number of observations falling into each unique combination of levels for the two variables:

| | Box Office Gross | | |
|---|---|---|---|
| **MPAA Rating** | Low | High | Total |
| Not Rated | 21 | 0 | 21 |
| G | 41 | 25 | 66 |
| PG | 328 | 143 | 471 |
| PG-13 | 856 | 252 | 1108 |
| R | 1207 | 124 | 1331 |
| NC-17 | 13 | 0 | 13 |
| Total | 2466 | 544 | 3010 |

**Conditional Distribution**

Among the movies that are rated G, 25 were high box office earners:

$$p_{high|G} = \frac{25}{66} = 37.9\%$$

# Start here Tuesday

# Summary Statistics: Contingency Tables

***Practice:*** Generate the contingency table that shows the distribution of hours across activities. Then, calculate the

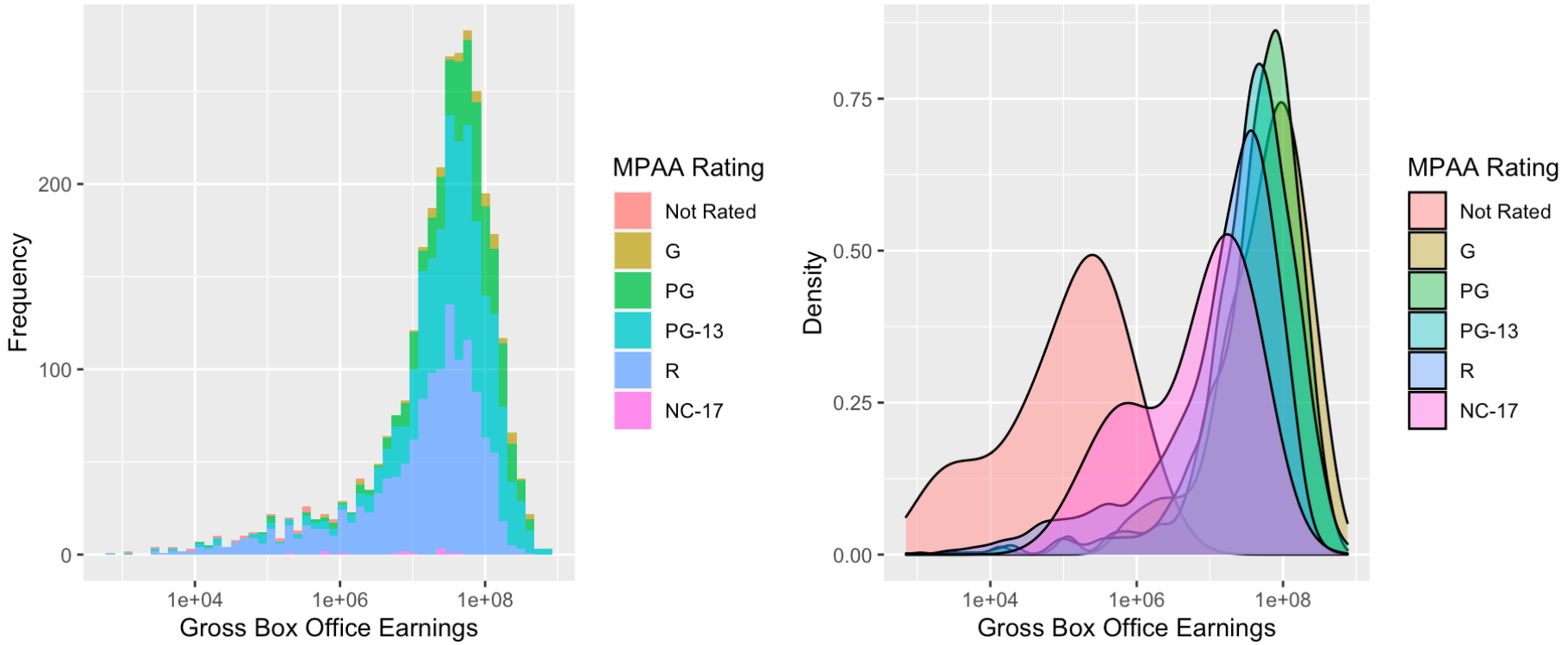
1. The ***marginal distribution*** of Work in the dataset
2. The ***marginal distribution*** of high hours in the dataset
3. The ***joint distribution*** of Work and high hours
4. The ***conditional distribution*** of low hours among Work

| Hours | Activity |
|---|---|
| low | Exercise |
| low | Exercise |
| high | Work |
| high | Chores |
| low | Work |
| high | Work |
| low | Driving |
| high | Exercise |
| high | Work |
| low | Driving |

# Relationships Between One Categorical Variable and One Numerical Variable

# Data Visualizations: Overlaid Histograms / Density Plots

We can visualize the distribution of gross box office earnings within each level of MPAA ratings—and compare these distributions with one another—using overlaid histograms and density plots:

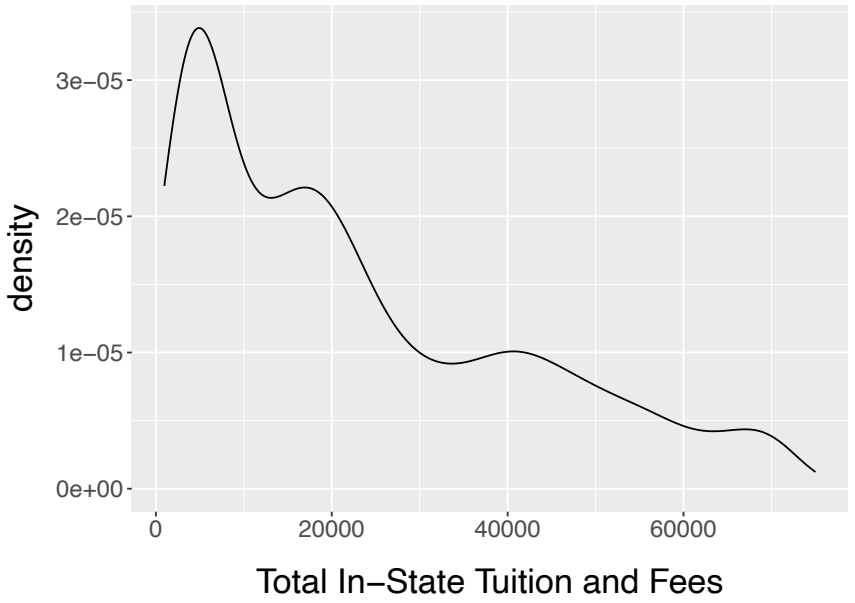



Gross box office earnings is being shown on a $\log 10$ scale

# Data Visualizations: Overlaid Histograms / Density Plots

These visualizations can be particularly informative if your data appear to be multi-modal, as there may be (and often is) something more going on in the story

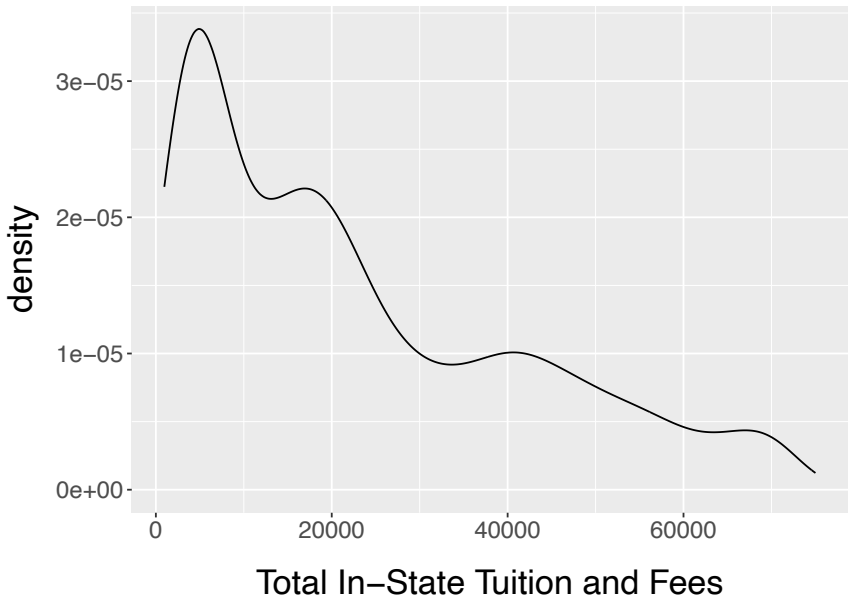
For example, consider the following density plot showing the distribution of in-state college tuition costs during the 2018–2019 academic year:

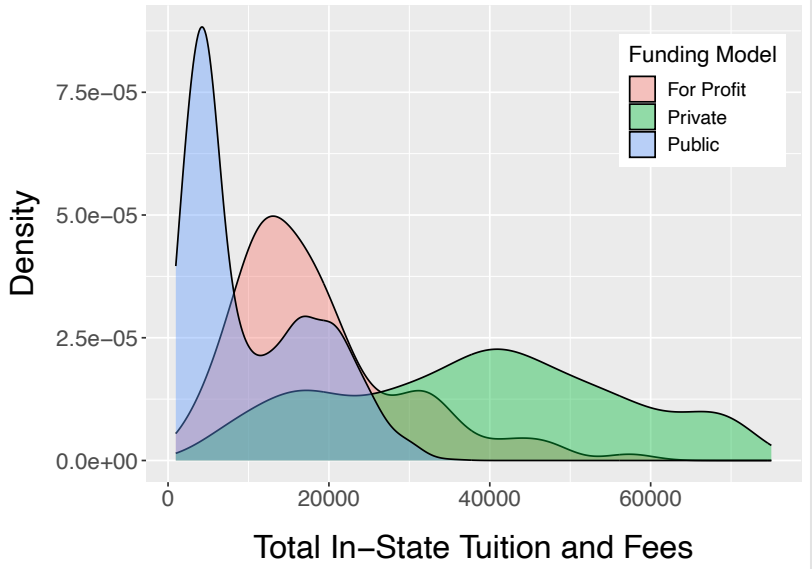


# Data Visualizations: Overlaid Histograms / Density Plots

These visualizations can be particularly informative if your data appear to be multi-modal, as there may be (and often is) something more going on in the story

For example, consider the following density plot showing the distribution of in-state college tuition costs during the 2018–2019 academic year:



# Data Visualizations: Faceted Histograms/ Density Plots

. . . Of course, overlaying all of our histograms or density plots on top of one another can sometimes be a mess, particularly if the categorical variable we're looking at has a lot of possible levels

→ We can instead display the histograms in side-by-side plots, each with the same $x$ and $y$ axis limits, for easier comparison across levels!

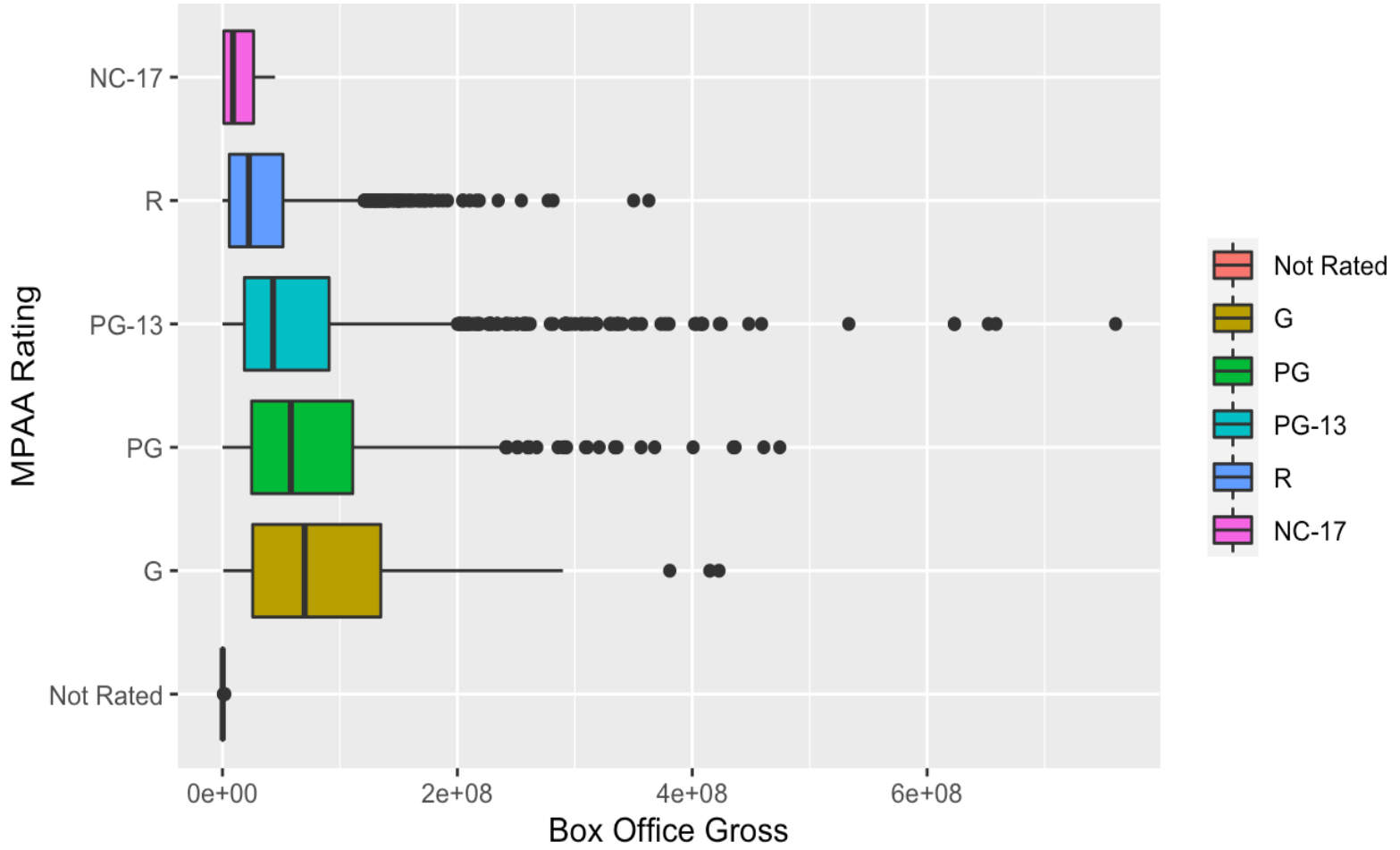
# Data Visualization Faceted Histograms/ Density Plots

***Practice:*** Generate an overlaid histogram / density plot or faceted histogram / density plot to show the distribution of minutes across each activity.

| Minutes | Activity |
|---|---|
| 100 | Exercise |
| 45 | Exercise |
| 60 | Work |
| 35 | Chores |
| 20 | Work |
| 90 | Work |
| 15 | Driving |
| 120 | Exercise |
| 300 | Work |
| 12 | Driving |

# Data Visualizations: Side-by-Side Boxplots

We can also create side-by-side boxplots to visually compare measures of center and spread for the numerical variable across levels of the categorical variable



# Data Visualization Side-by-Side Boxplots

***Practice:*** Generate side-by-side boxplots to show the distribution of minutes across each activity.

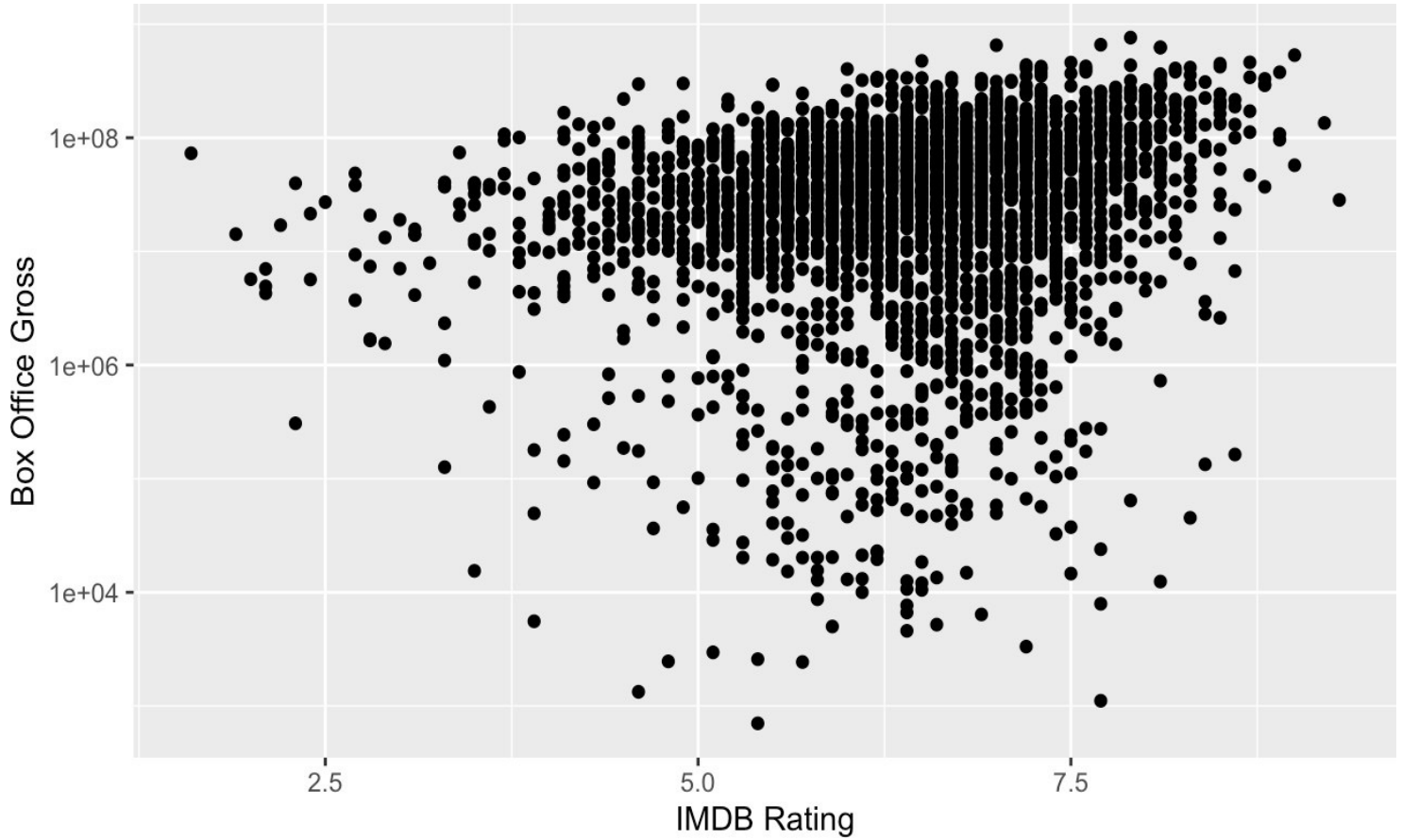
| Minutes | Activity |
|---|---|
| 100 | Exercise |
| 45 | Exercise |
| 60 | Work |
| 35 | Chores |
| 20 | Work |
| 90 | Work |
| 15 | Driving |
| 120 | Exercise |
| 300 | Work |
| 12 | Driving |

# Relationships Between Two Numerical Variables

# Data Visualizations: Scatterplots

Scatterplots are one of the most common ways of visualizing the relationship between two numerical variables.

- For the $i$th observational unit, let $x_i$ be the value of the explanatory variable and $y_i$ the value of the response variable.
- We plot each $(x_i, y_i)$ pair for all $n$ observations in our sample.



# Data Visualization Scatterplots

***Practice:*** Generate a scatterplot show the relationship between minutes and cost.

| Minutes | Cost |
|---|---|
| 100 | $45 |
| 45 | $27 |
| 60 | $56 |
| 35 | $15 |
| 20 | $21 |
| 90 | $62 |
| 15 | $5 |
| 120 | $55 |
| 300 | $100 |
| 12 | $7 |

# Summary Statistics: Pearson Correlation Coefficient

The Pearson correlation coefficient quantifies the strength of the (linear) relationship between our explanatory and response variables:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right),$$

where $\bar{x}$ and $\bar{y}$ are the sample means and $s_x$ and $s_y$ are the sample standard deviations.

# Summary Statistics: Pearson Correlation Coefficient

The Pearson correlation coefficient quantifies the strength of the (linear) relationship between our explanatory and response variables:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right),$$

where $\bar{x}$ and $\bar{y}$ are the sample means and $s_x$ and $s_y$ are the sample standard deviations.

What aspects of the distribution does $r$ capture?

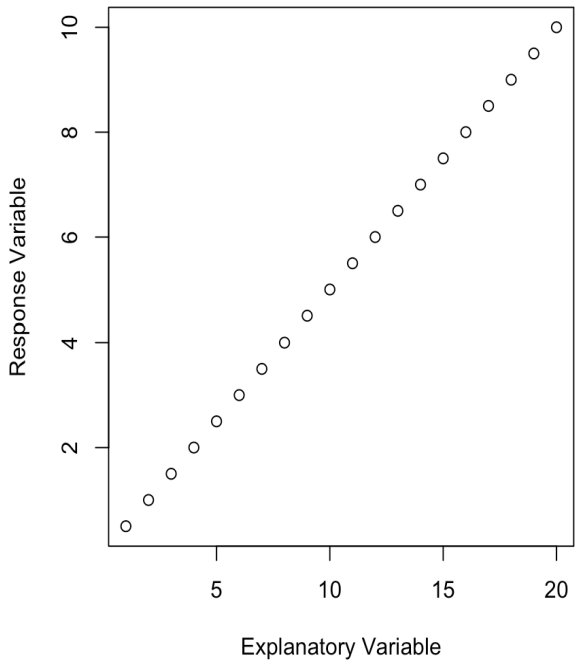
1. Direction of the association
   - Ex: Do higher-rated movies tend to make more or less money at the box office?
2. Degree of noisiness ⇒ think two-dimensional spread!
   - Ex: If a movie receives a 7.4 IMDB rating, how certain are we (and how much uncertainty remains) in the box office totals?

# Summary Statistics: Pearson Correlation Coefficient

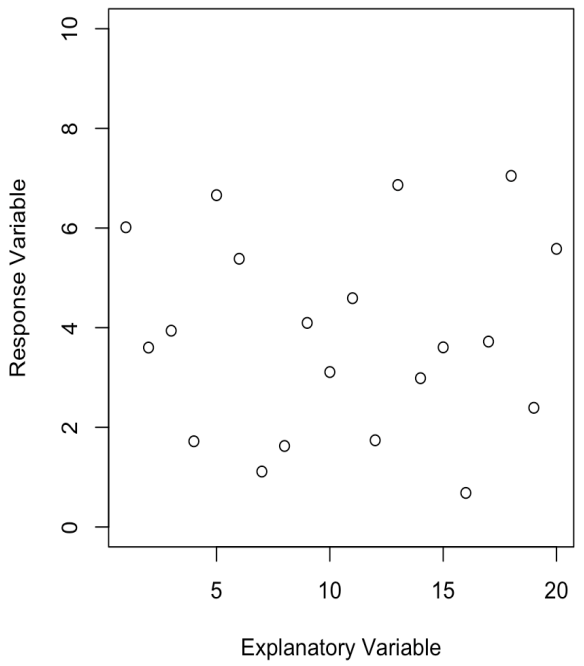
The Pearson correlation coefficient quantifies the strength of the (linear) relationship between our explanatory and response variables:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right),$$

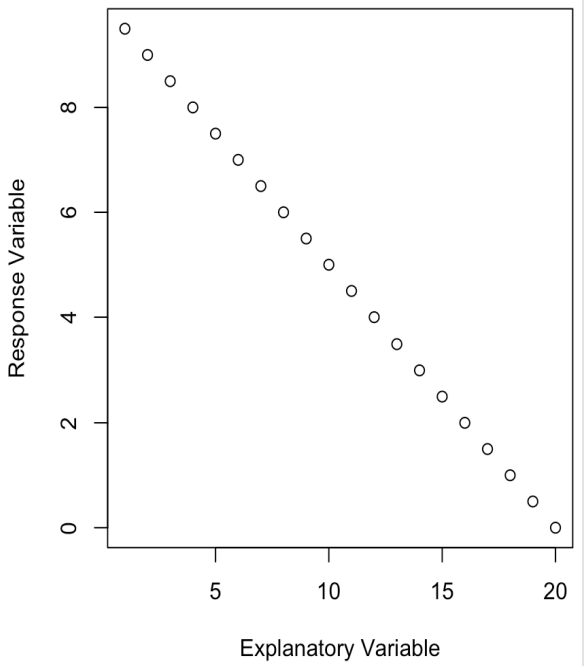
The correlation takes on values between $-1 \leq r \leq 1$, where:




Perfect positive correlation, $r = 1$




No correlation, $r = 0$




Perfect negative correlation, $r = -1$

# Summary Statistics: Pearson Correlation Coefficient

The Pearson correlation coefficient quantifies the strength of the (linear) relationship between our explanatory and response variables:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right),$$

***Practice:*** Based on your scatterplot what correlation value would you expect between Minutes and Cost?

Calculate it to check your answer.

| Minutes | Cost |
|---|---|
| 100 | $45 |
| 45 | $27 |
| 60 | $56 |
| 35 | $15 |
| 20 | $21 |
| 90 | $62 |
| 15 | $5 |
| 120 | $55 |
| 300 | $100 |
| 12 | $7 |


tive

$r = -1$

# 20-30 Minute Activity: EDA Practice

Open movies.csv (under Demos on the course website) in excel or google sheets.

For (a) two categorical, (b) a categorical and numerical, and (c) two numerical variables, generate the appropriate summary visualizations and summary statistics.

You in some cases, you will need to manipulate the raw data and use formulas. Helpful tips can be found here:

- Excel
  - https://www.princeton.edu/~otorres/Excel/excelstata.htm
  - https://statisticsbyjim.com/basics/descriptive-statistics-excel/
- Google Sheets
  - http://www.comfsm.fm/~dleeling/statistics/text6.html#page-031
  - https://www.groovypost.com/howto/quickly-get-column-statistics-in-google-sheets/