# Introduction to Machine Learning – Exploratory Data Analysis

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

# Plan for Today

- Exploratory Data Analysis

# Exploratory Data Analysis

***Exploratory data analysis (EDA)*** refers to the practice of reducing and summarizing data in ways that:

- Help us make sense of the information that we have
- Help to inform a research question

You can think of EDA as the data version of tl;dr.

# Exploratory Data Analysis

***Exploratory data analysis (EDA)*** refers to the practice of reducing and summarizing data in ways that:

- Help us make sense of the information that we have
- Help to inform a research question

You can think of EDA as the data version of tl;dr.

**Our usual goal:**
model some phenomenon using a dataset

**Goal of EDA:**
develop an understanding of a dataset

# Exploratory Data Analysis

***Exploratory data analysis (EDA)*** refers to the practice of reducing and summarizing data in ways that:

- Help us make sense of the information that we have
- Help to inform a research question

You can think of EDA as the data version of tl;dr.

### Graphical Summaries (Data Visualizations)

A visual representation of how our data are *distributed* across the observations in our sample

### Numeric Summaries (Summary Statistics)

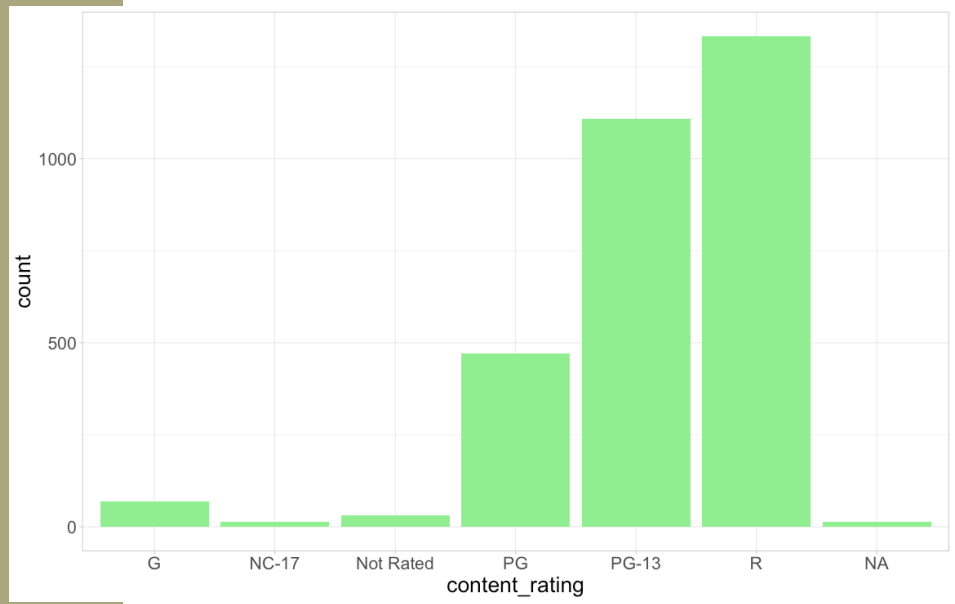
A single number or set of numbers that captures important features of that distribution, such as its *center* and *spread*

# EDA for Categorical Variables: **Bar Plots**

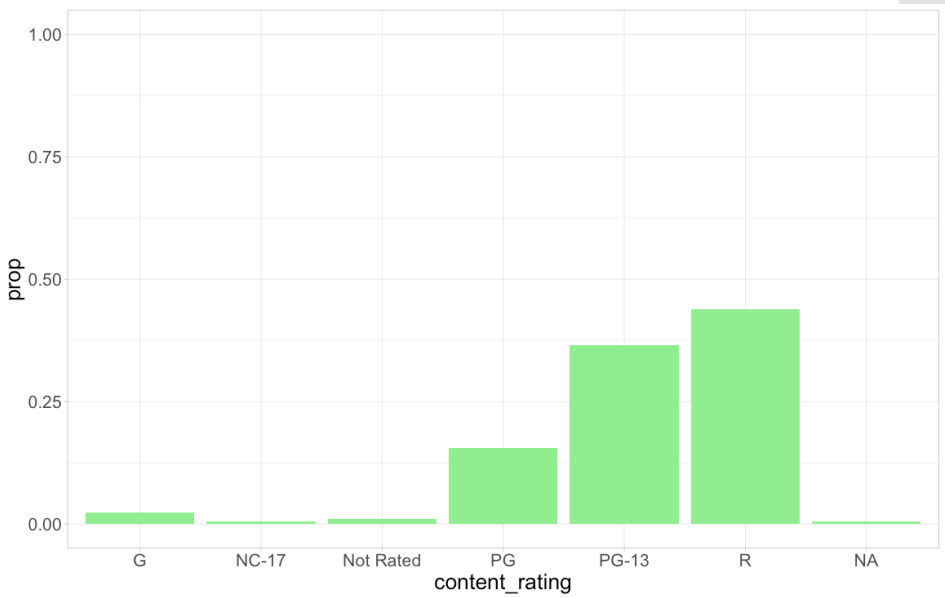
The empirical distribution of a categorical variable is comprised of:

- The possible levels or values of the categorical variable
- The (relative) frequency of those levels in the observed data

One method of visualizing this distribution is through a ***bar plot***:




*Bar plot showing frequency of MPAA ratings.*




*Bar plot showing relative frequency of MPAA ratings.*

# EDA for Categorical Variables: **Summary Statistics**

We can present this same information numerically using a ***frequency table***, which displays both:

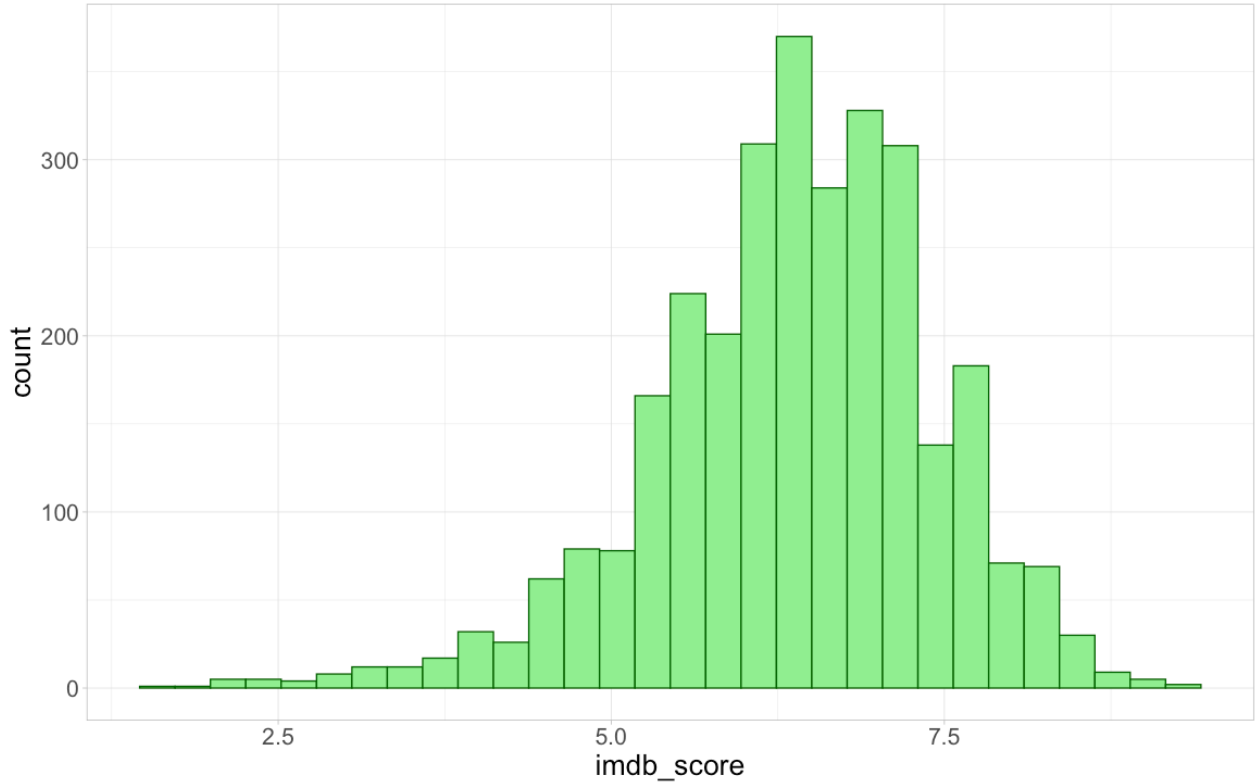
- the number of movies ($n$) that obtained each rating
- the relative frequency of (prop) those ratings

| content_rating | n | prop |
|---|---|---|
| Not Rated | 21 | 0.0069767 |
| G | 66 | 0.0219269 |
| PG | 471 | 0.1564784 |
| PG-13 | 1108 | 0.3681063 |
| R | 1331 | 0.4421927 |
| NC-17 | 13 | 0.0043189 |

# EDA for Numerical Variables: **Histograms**

When the variable that we're summarizing is numerical, we can instead visualize its distribution using either a ***histogram*** or ***density plot***

Histogram: numerical analog of the frequency bar plot




Created by:

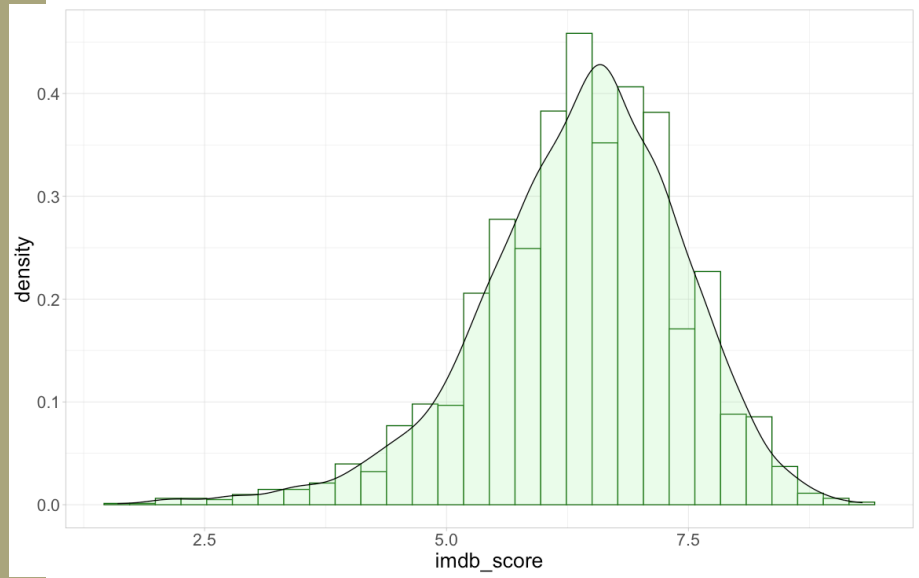
Dividing the range of IMDB ratings (here from 1.6 to 9.3) into intervals (also called "bins") of equal width

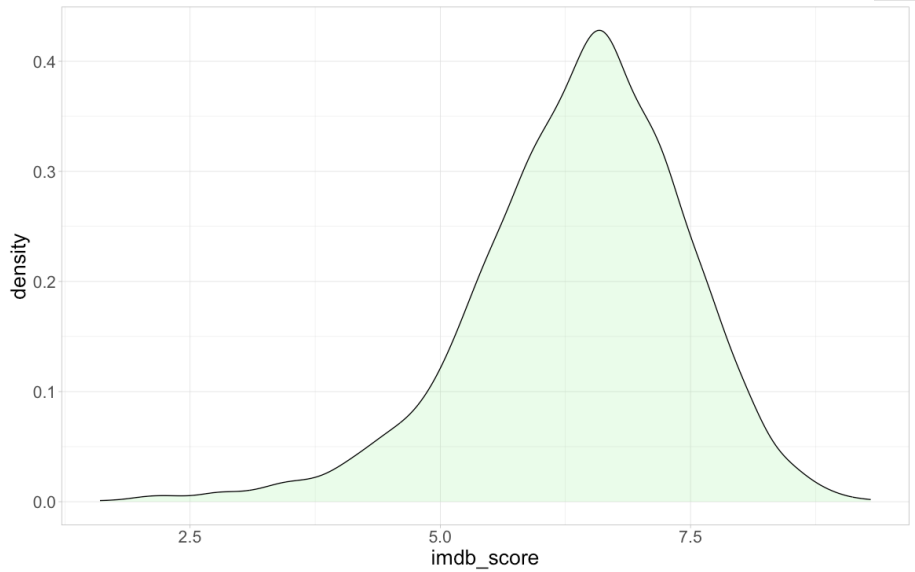
Counting the number of movies whose IMDB rating falls into each bin

# EDA for Numerical Variables: **Density Plots**

When the variable that we're summarizing is numerical, we can instead visualize its distribution using either a ***histogram*** or ***density plot***

Density plot: numerical analog of relative frequency bar plot




Created by standardizing and smoothing over the corresponding histogram

# EDA for Numerical Variables: **Describing Distributions**

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

- Skewness
- Modality
- Center
- Spread

# EDA for Numerical Variables: **Describing Distributions**

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

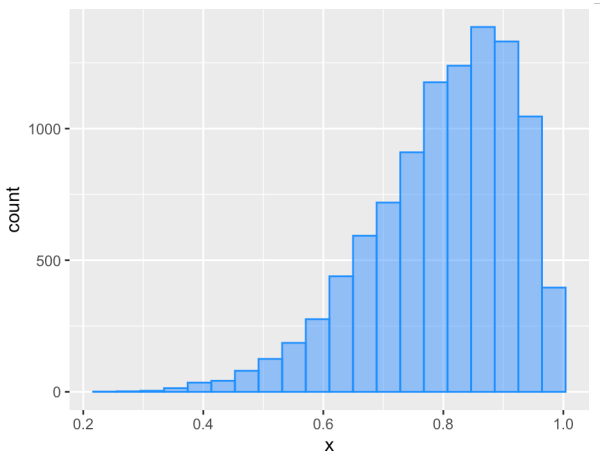
# EDA for Numerical Variables: **Describing Distributions**

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

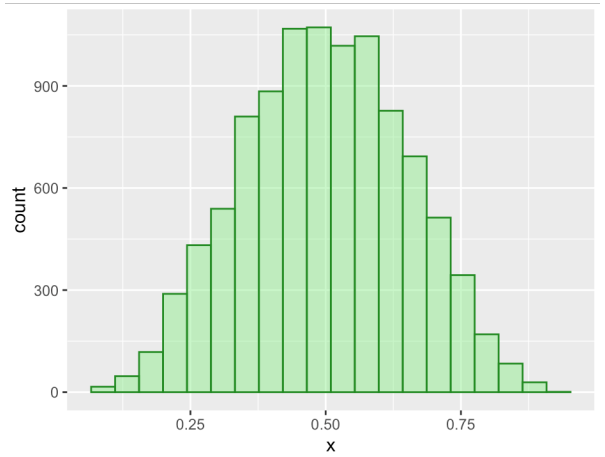
- Graphical
  - **Skewness**
  - Modality
- Summary Statistics
  - Center
  - Spread

***Skewness*** is a measure of (a)symmetry!

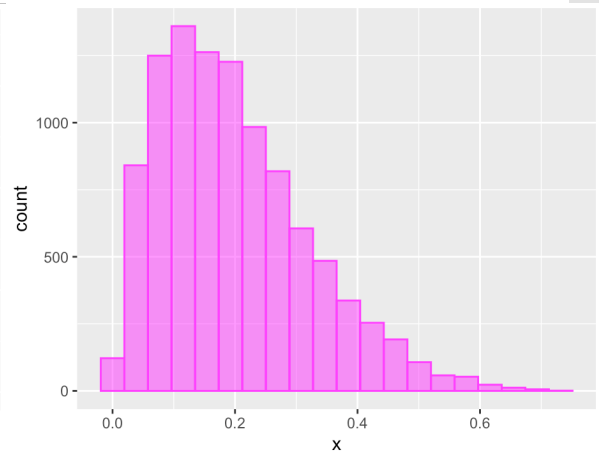
→ Why pay attention to skew? Later in this course we'll see statistical tools that assume our data are (close to) symmetric, and we need to be able to assess whether this assumption is reasonable.




Left ("negative") skewed distribution.




Symmetric distribution.




Right ("positive") skewed distribution.

**Tip**: Whatever side the long tail is on is the side of skew

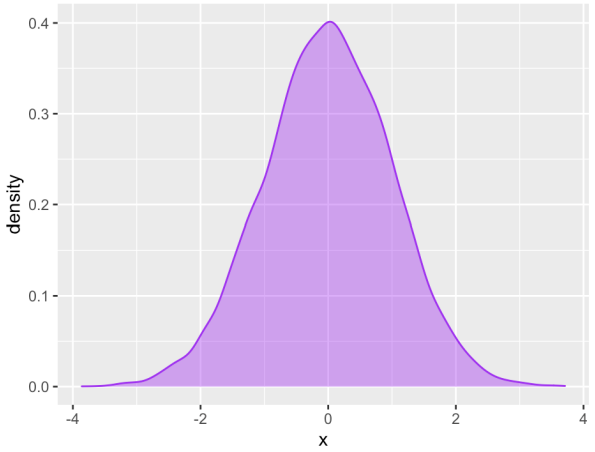
# EDA for Numerical Variables: **Describing Distributions**

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

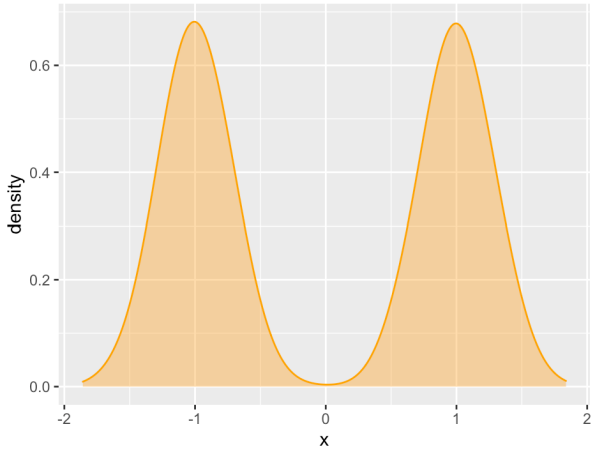
- Graphical
  - Skewness
  - **Modality**
- Summary Statistics
  - Center
  - Spread

***Modality*** is a measure of how many peaks ("modes") the distribution has

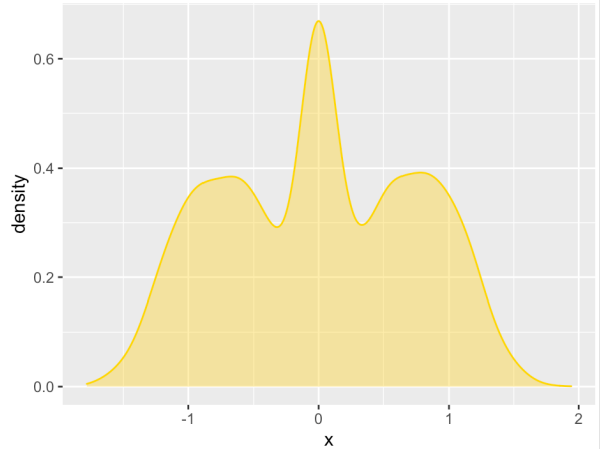
→ Why pay attention to modality? A mode is a value that occurs with high frequency in our data, and it can help to inform our understanding of what values our variable tends to take on




Unimodal distribution.




Bimodal distribution.




Multimodal distribution

# EDA for Numerical Variables: **Describing Distributions**

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

Let $x_1, x_2, x_3, \ldots x_n$ be the observed values of our variable of interest across the $n$ observational units in our dataset.

Measures of central tendency give us a sense of what the typical value of this variable might look like.

***Mean:*** the average value of the variable,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots . + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

***Median:*** suppose we order the observations from smallest to largest. the median is the value of $x_i$ that falls in the middle (or, if $n$ is even, the average of the two middle values).

⇒ At least half of our data are less than or equal to the median and at least half are greater than or equal to the median

# EDA for Numerical Variables: **Describing Distributions**

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**

Measures of dispersion give us a sense of how much observation to observation ***variability*** there is in a variable.

***Range:*** the difference between the maximum and minimum values in the dataset

***Interquartile Range:*** the difference between the 75th and 25th percentiles of the data

***Variance:*** (almost) the average squared distance between the observed data for the $i$th observational unit, $x_i$, and the sample mean, $\bar{x}$

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1}\sum_{i-1}^{n}(x_i - \bar{x})^2$$

***Standard Deviation:*** the square root of the variance, $s = \sqrt{s^2}$

# EDA for Numerical Variables: **Describing Distributions**

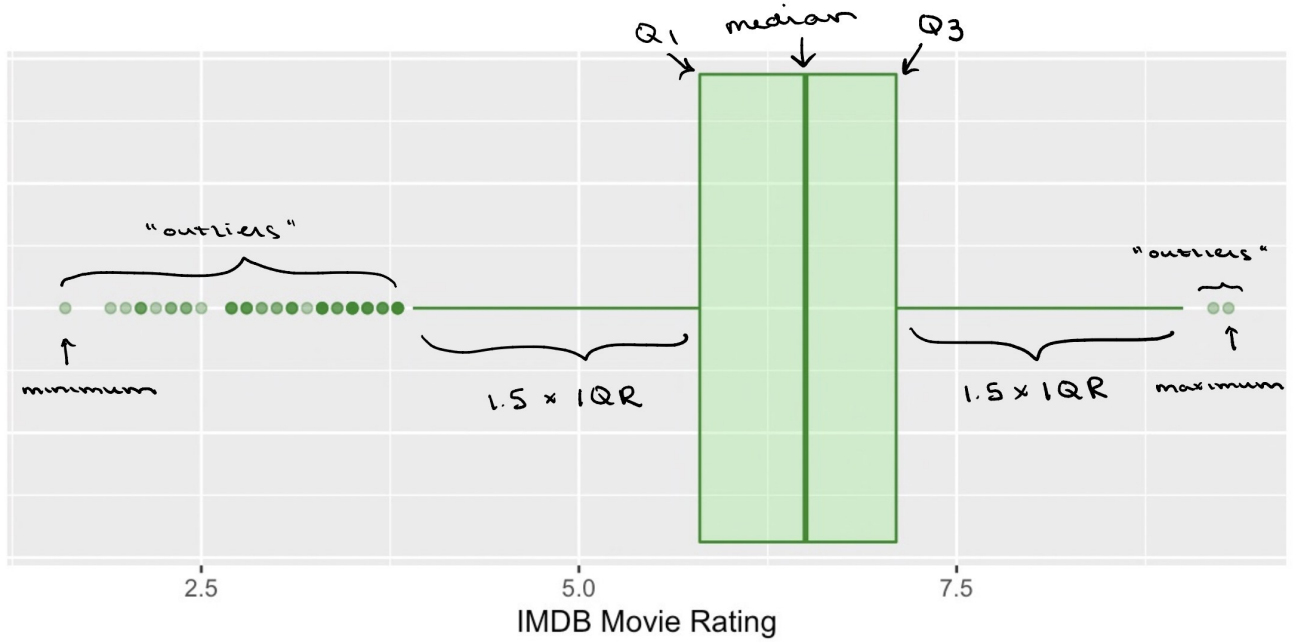
When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

- 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 | Median | 75th percentile | Maximum

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



# Multiple Variables

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

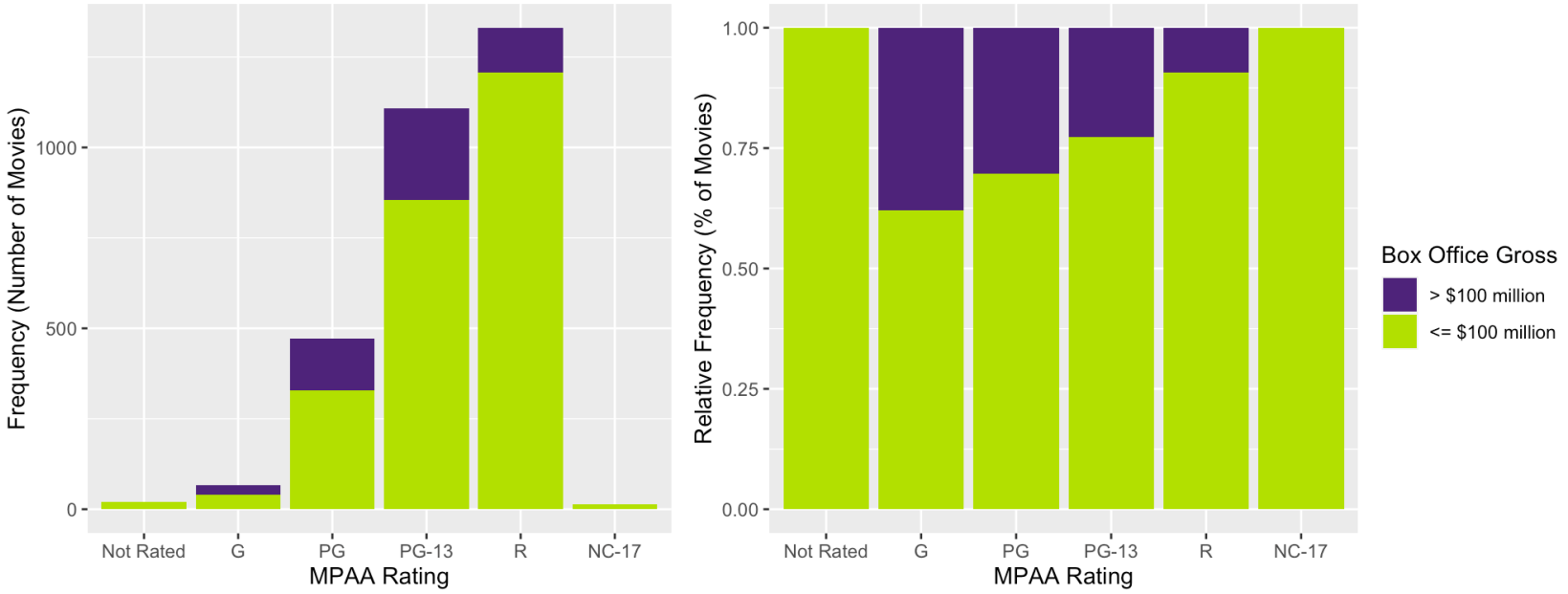
# EDA for Relationships Between Two Categorical Variables: **Stacked Bar Plots**

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



# EDA for Relationships Between Two Categorical Variables: **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!

# EDA for Relationships Between Two Categorical Variables: **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:

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

# EDA for Relationships Between Two Categorical Variables: **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:

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

# EDA for Relationships Between Two Categorical Variables: **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:

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

# EDA for Relationships Between Two Categorical Variables: **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 |

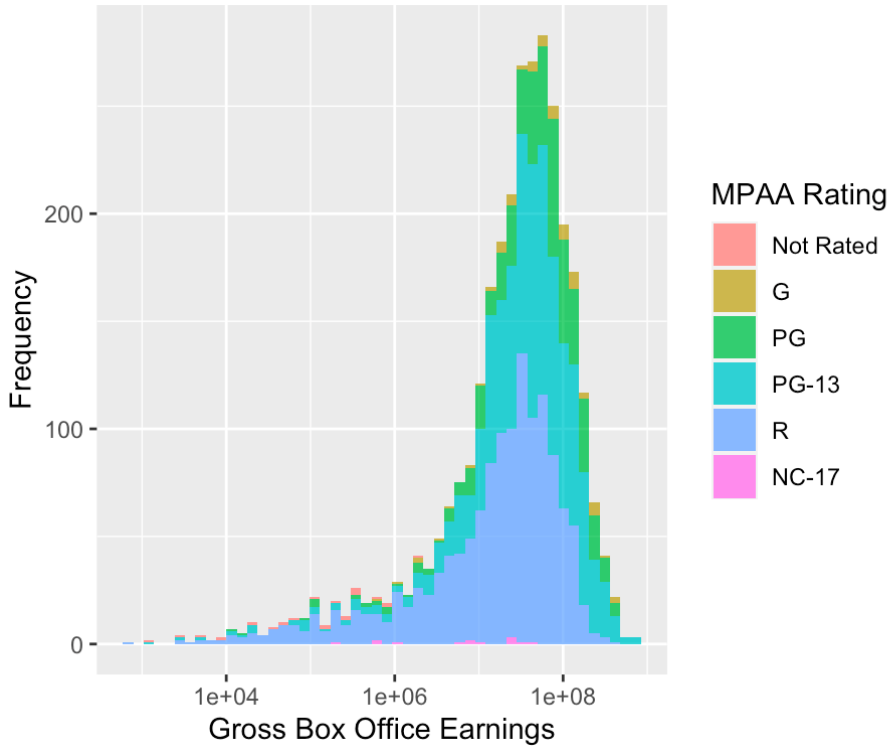
Conditional Distribution

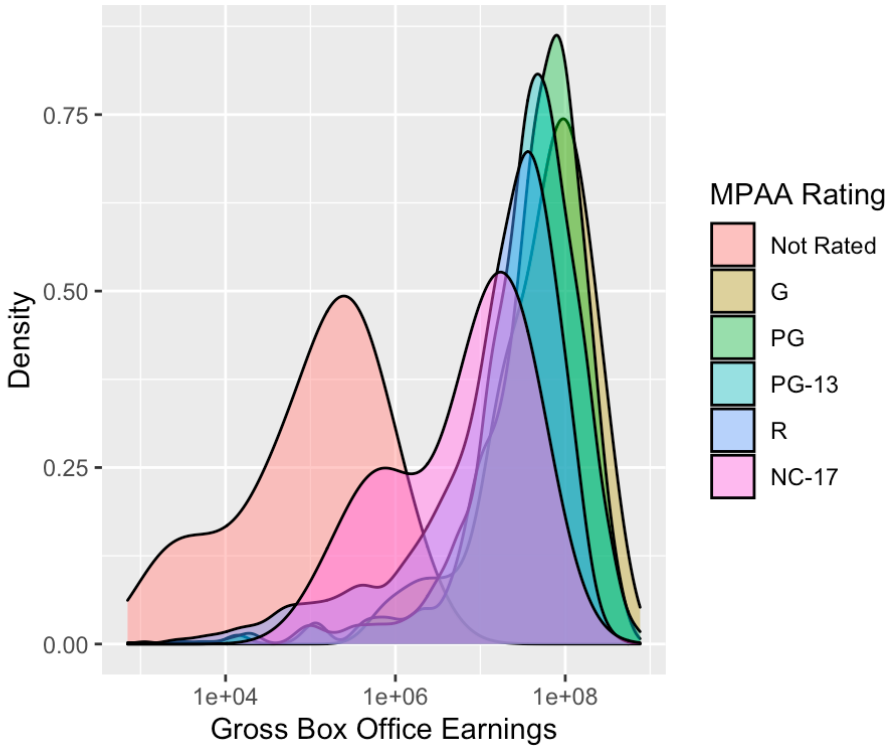
Among the movies that are rated G, 25 were high box office earners:

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

# EDA for Relationships Between One Categorical and One Numerical Variable: **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***:



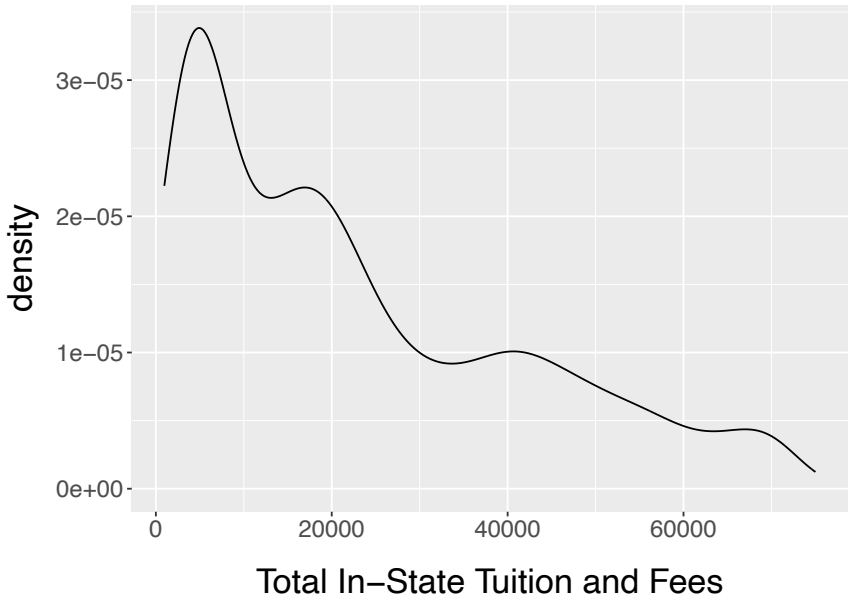




Gross box office earnings is being shown on a $\log 10$ scale

# EDA for Relationships Between One Categorical and One Numerical Variable: **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

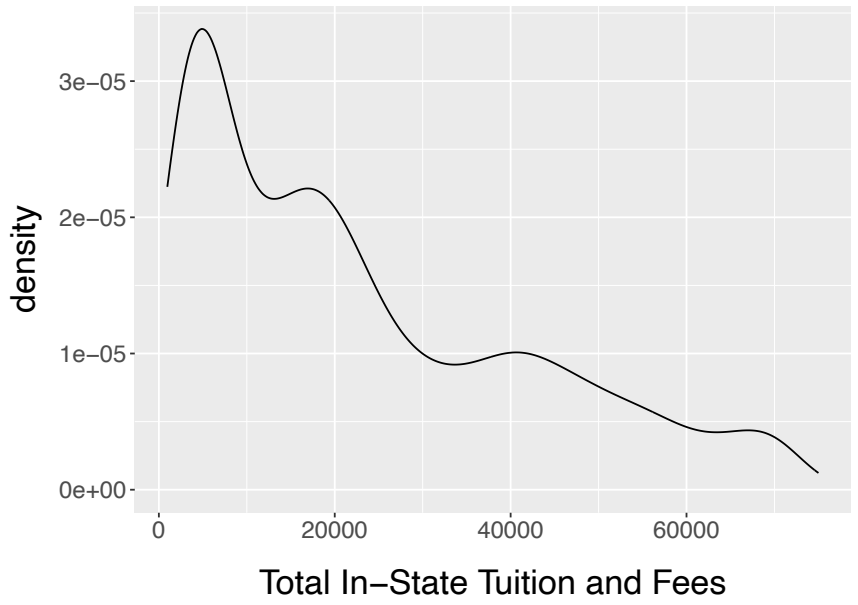
For example, consider the following density plot showing the distribution of in-state college tuition costs during the 2018–2019 academic year:

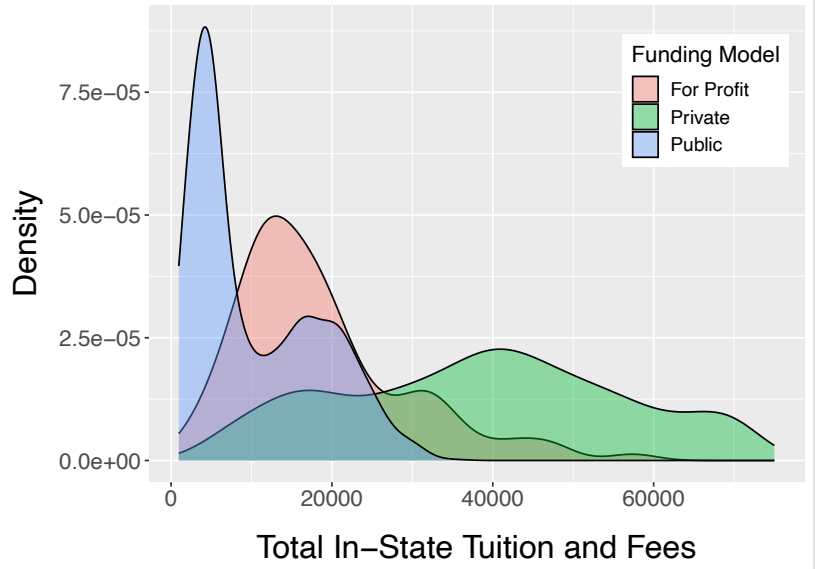


# EDA for Relationships Between One Categorical and One Numerical Variable: **Overlaid Histograms / Density Plots**

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:

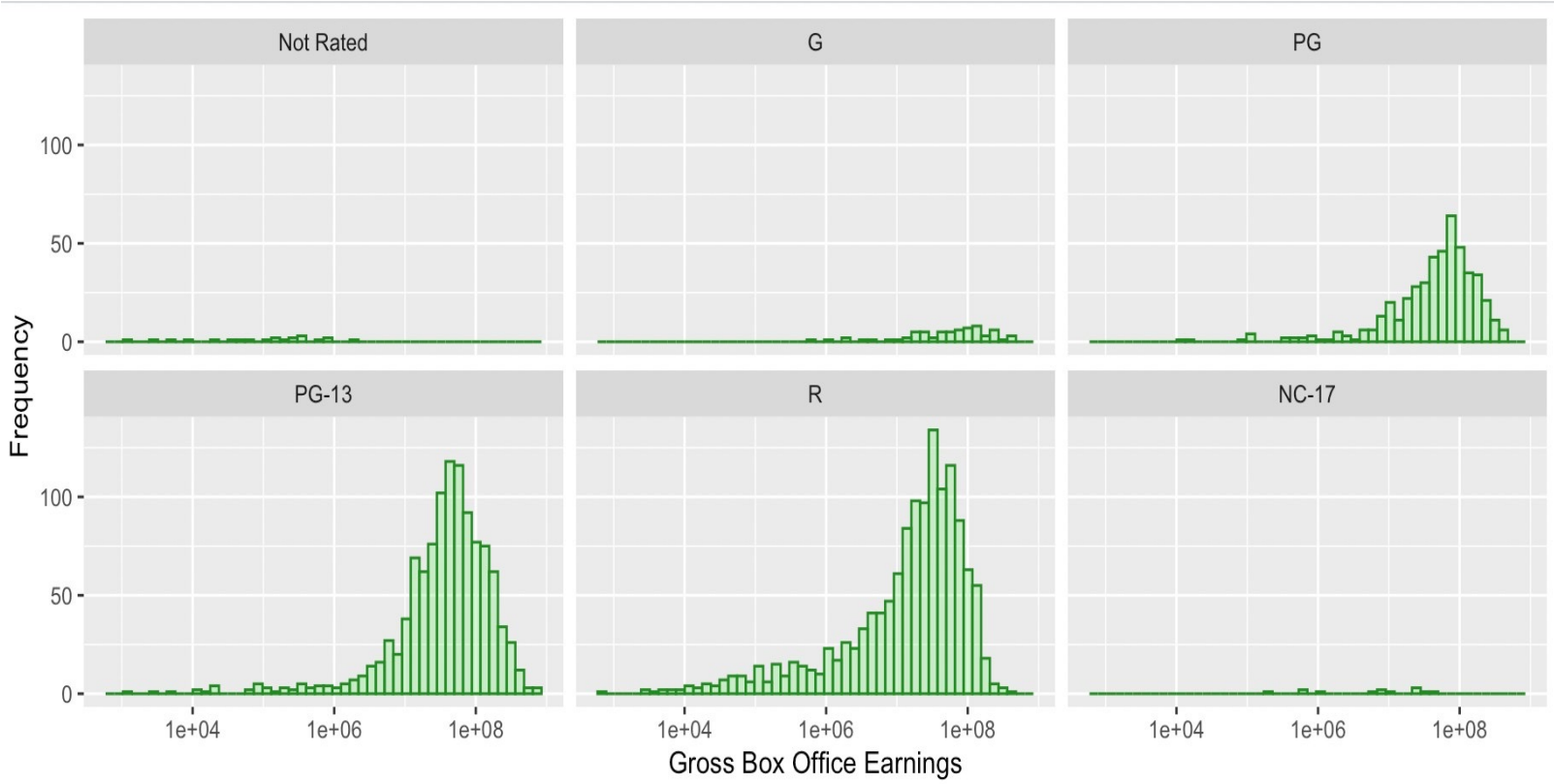


# EDA for Relationships Between One Categorical and One Numerical Variable: **Overlaid Histograms / Density Plots**

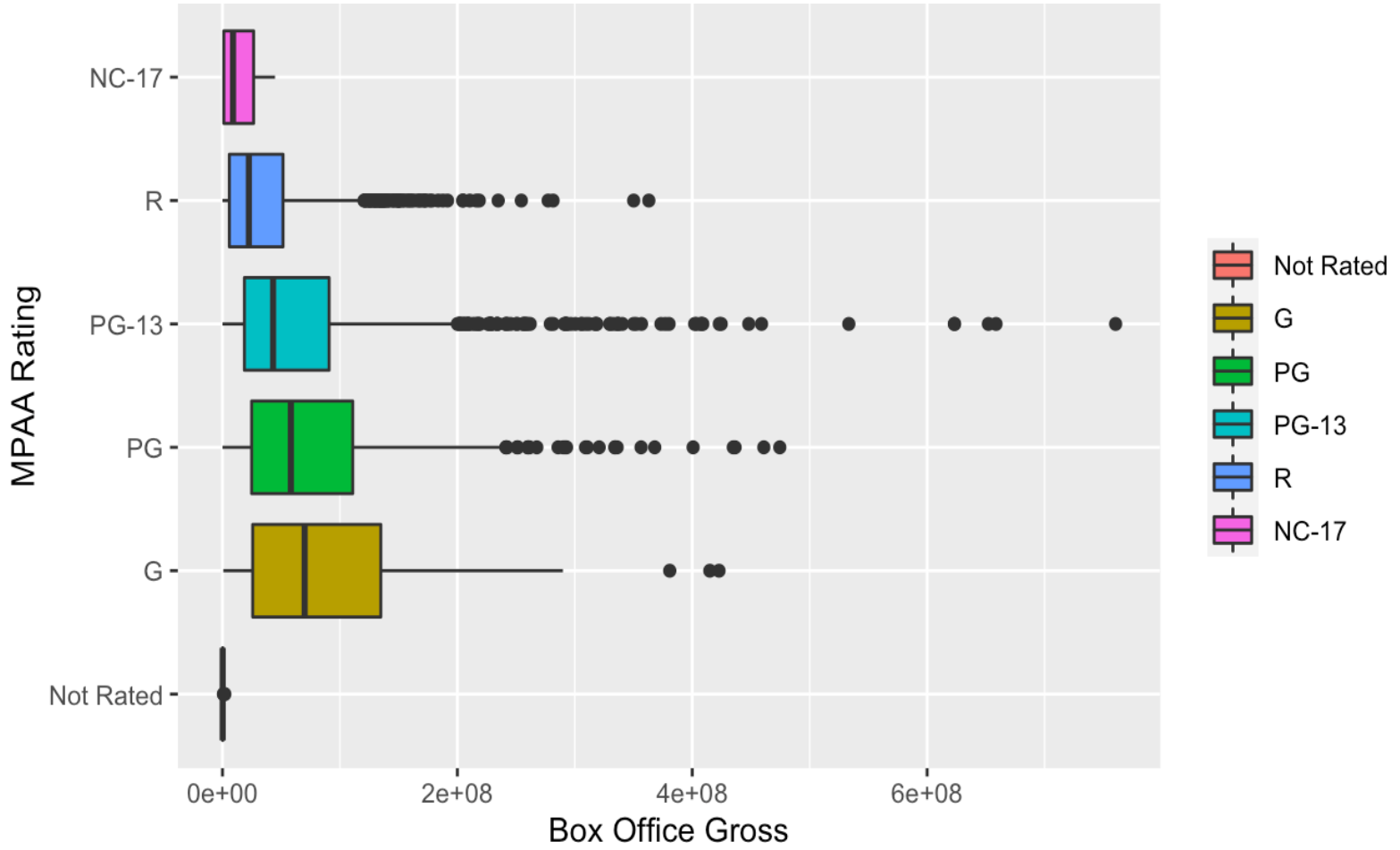
. . . Of course, overlaying all of our histograms or density plots on top of one another can sometimes be a mess, particularly if the categorical variable we're looking at has a lot of possible levels

→ 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!



# EDA for Relationships Between One Categorical and One Numerical Variable: **Side-by-Side Boxplots**

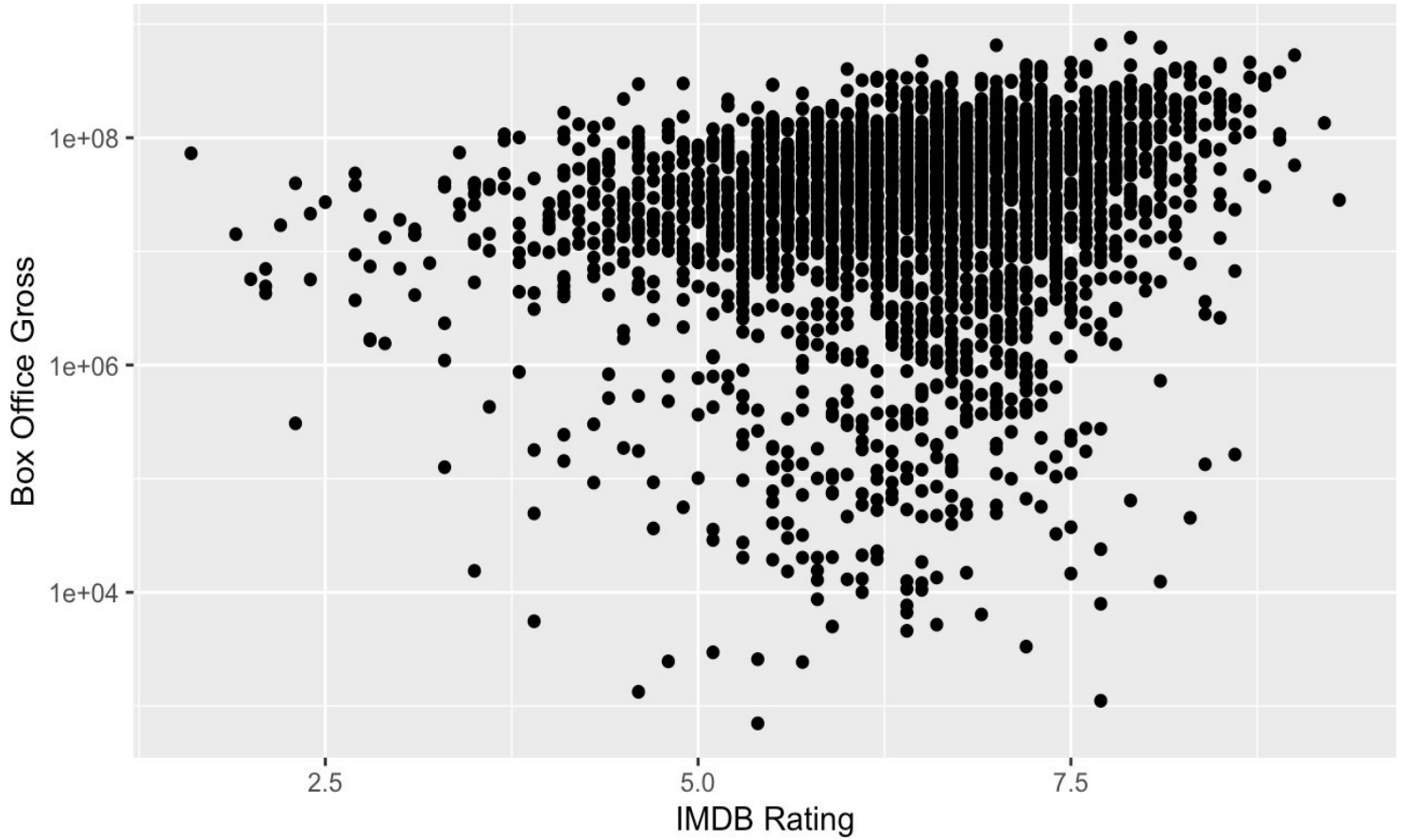
We can also create ***side-by-side boxplots*** to visually compare measures of center and spread for the numerical variable across levels of the categorical variable



# EDA for Relationships Between Two Numerical Variables: **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.



# EDA for Relationships Between Two Numerical Variables: **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.