Agenda

1. Center, Shape, and Spread

Warmup: Lurking Variables For each of the following pairs of variables, a statistically signficant positive relationship has been observed. Identify a potential lurking variable that might cause the spurious correlation.

1. The amount of ice cream sold in New England and the number of deaths by drowning

2. The salary of U.S. ministers and the price of vodka

3. The number of doctors in a region and the number of crimes committed in that region

4. The number of storks sighted and the population of Oldenburg, Germany, over a six-year period

5. The amount of coffee consumed and the prevalence of lung cancer

IMDB movie data Today, we'll focus on data about movies. I've chosen to use a dataset available on Kaggle.com which includes information scraped from the Internet Movie Database (IMDB). `https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset/discussion`

```
library(mosaic)
movies <- read_csv("http://www.science.smith.edu/~amcnamara/sds220/data/movies2.csv")
movies %>%
  glimpse()
```

```
## Observations: 5,043
## Variables: 28
## $ color                     <chr> "Color", "Color", "Color", "Color", ...
## $ director_name             <chr> "James Cameron", "Gore Verbinski", "...
## $ num_critic_for_reviews    <int> 723, 302, 602, 813, NA, 462, 392, 32...
## $ duration                  <int> 178, 169, 148, 164, NA, 132, 156, 10...
## $ director_facebook_likes   <int> 0, 563, 0, 22000, 131, 475, 0, 15, 0...
## $ actor_3_facebook_likes    <int> 855, 1000, 161, 23000, NA, 530, 4000...
## $ actor_2_name              <chr> "Joel David Moore", "Orlando Bloom",...
## $ actor_1_facebook_likes    <int> 1000, 40000, 11000, 27000, 131, 640,...
## $ gross                     <int> 760505847, 309404152, 200074175, 448...
## $ genres                    <chr> "Action|Adventure|Fantasy|Sci-Fi", "...
## $ actor_1_name              <chr> "CCH Pounder", "Johnny Depp", "Chris...
## $ movie_title               <chr> "Avatar", "PiratesoftheCaribbeanAtWo...
## $ num_voted_users           <int> 886204, 471220, 275868, 1144337, 8, ...
## $ cast_total_facebook_likes <int> 4834, 48350, 11700, 106759, 143, 187...
## $ actor_3_name              <chr> "Wes Studi", "Jack Davenport", "Step...
```

```
## $ facenumber_in_poster      <int> 0, 0, 1, 0, 0, 1, 0, 1, 4, 3, 0, 0, ...
## $ plot_keywords             <chr> "avatar|future|marine|native|paraple...
## $ movie_imdb_link           <chr> "http://www.imdb.com/title/tt0499549...
## $ num_user_for_reviews      <int> 3054, 1238, 994, 2701, NA, 738, 1902...
## $ language                  <chr> "English", "English", "English", "En...
## $ country                   <chr> "USA", "USA", "UK", "USA", NA, "USA"...
## $ content_rating            <chr> "PG-13", "PG-13", "PG-13", "PG-13", ...
## $ budget                    <dbl> 237000000, 300000000, 245000000, 250...
## $ title_year                <int> 2009, 2007, 2015, 2012, NA, 2012, 20...
## $ actor_2_facebook_likes    <int> 936, 5000, 393, 23000, 12, 632, 1100...
## $ imdb_score                <dbl> 7.9, 7.1, 6.8, 8.5, 7.1, 6.6, 6.2, 7...
## $ aspect_ratio              <dbl> 1.78, 2.35, 2.35, 2.35, NA, 2.35, 2....
## $ movie_facebook_likes      <int> 33000, 0, 85000, 164000, 0, 24000, 0...
```

As the output suggests, the data has 5,043 observations (movies) and 28 variables.

Thought Experiment Consider the following two variables:

- The `duration` of all the movies in the IMDB dataset.
- The `facenumber_in_poster` (number of faces detected in the movie poster).

Think about the distribution of each variable, and discuss the following questions with a neighbor.

1. Draw the shape you believe each distribution has. What features does it have? Is it symmetric? Is it normal? It is unimodal? [Make sure you label the axes on your distribution plot.] What is the range of each variable?

2. How would you summarize each distribution numerically? Which measures are most appropriate?

3. Suppose we added an additional face to each movie poster. How would the distribution of `facenumber_in_poster` change? What would happen to your measures of center and spread?

Describing Distributions We are going to hone in on the quantitative variables in this analysis, and start doing some EDA of their distributions. When describing distributions, three concepts are particularly useful: *Center*, *Shape*, and *Spread*.

- Graphical techniques for summarizing the *shape* of the distribution of one variable:
  - Histogram [`geom_histogram()`]
  - Density plot [`geom_density()`]

  - Box (and whisker) plot [`geom_boxplot()`]

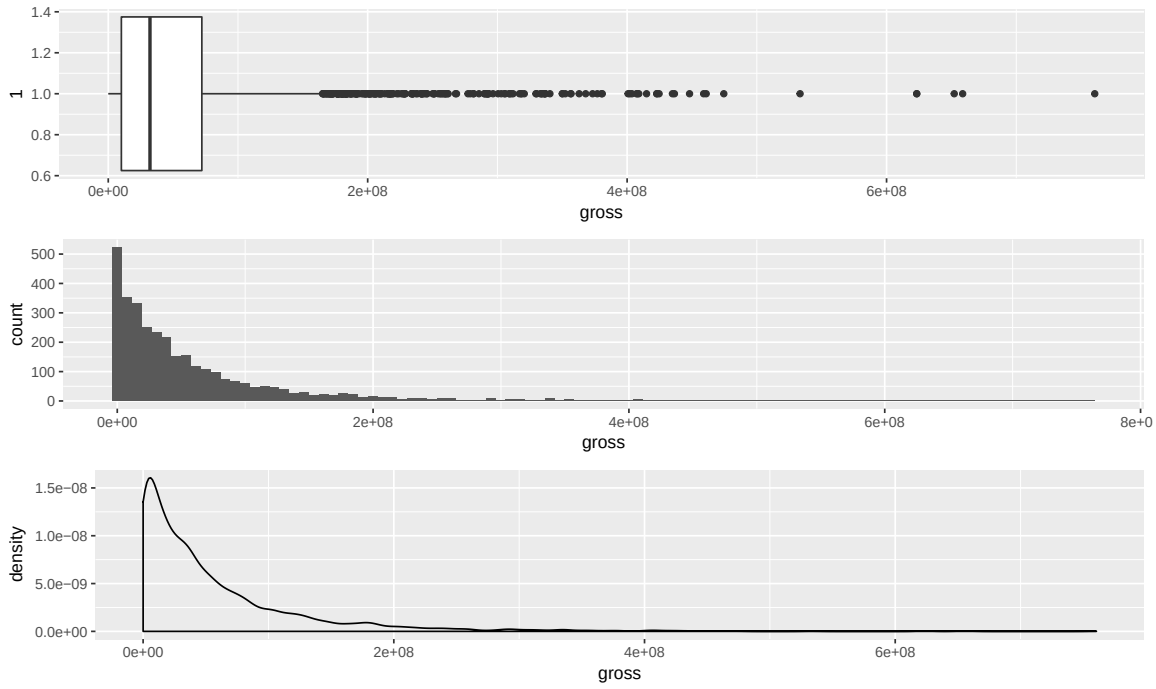
- Numerical Techniques for summarizing the *center* and *spread* of the distribution of one variable:
  - Center: mean [`mean()`], median [`median()`]
  - Spread: standard deviation [`sd()`], variance [`var()`], range [`range()`], IQR [`IQR()`]
  - (Center and spread can be seen together in `favstats()`)

US box-office gross A box plot, histogram, and density plot reveal different features of the distribution of box-office gross.

```
movies <- movies %>%
  filter(country=="USA")
favstats(~gross, data = movies)

##  min       Q1   median       Q3       max     mean       sd    n missing
##  703 10110274 32178777 72147000 760505847 55214607 71733124 3235     572
```

```
ggplot(data=movies, aes(y=gross, x=1)) + geom_boxplot()
ggplot(data=movies, aes(x=gross)) + geom_histogram()
ggplot(data=movies, aes(x=gross)) + geom_density()
```



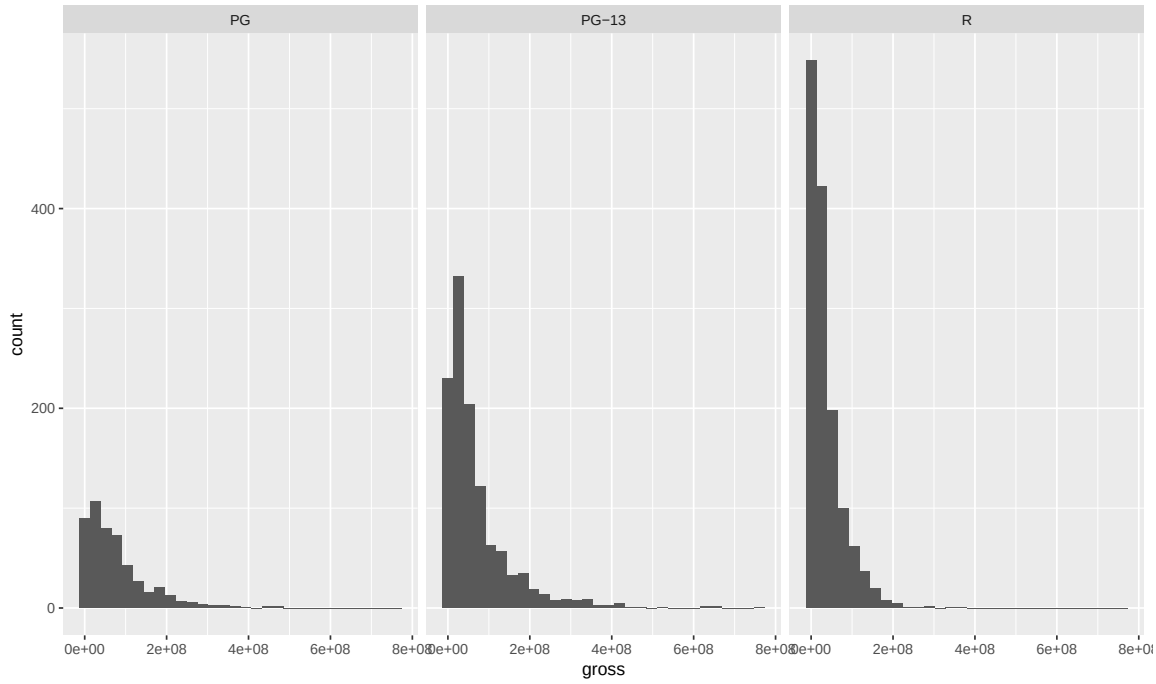
1. Describe the distribution of box office gross using words.

2. What information can you glean from the histogram or density plot that is not revealed by the numerical table or the box plot?

3. What information does the numerical table give you that is not available in the plots?

faceting One great thing about `ggplot2` is that it allows you to 'facet' by another variable. (Essentially, make the same plot several times for different values of a second variable.) For example, maybe we think the distribution of `gross` will be different depending on the `content_rating`

```
movies <- movies %>%
  filter(content_rating %in% c("PG", "PG-13", "R"))
ggplot(data=movies, aes(x=gross)) + geom_histogram() + facet_grid(~content_rating)
```



Relationships With a partner, brainstorm some relationships you think might exist in the data. What distributions of quantitative variables likely look different depending on the value of another categorical variable?