Lecture 5: Data Visualization
Modeling Social Data, Spring 2017
Columbia University

February 17, 2017

Notes from ea2711

1 Guest Lecture

1.1 What is data visualization?

- "The use of computer-generated interactive visual representations to amplify cognition."
- The goal of visualizations is to develop and support hypothesis, and to inspire and convince others

1.2 Exploratory data analysis

There are multiple methods of showing data

- Plain data – very difficult to perceive patterns and make sense of data
- Summary-statistics, e.g. mean, median, etc. – can get an intuitive understanding
- Plot Data – can identify patterns and trends

To create a visualization, we must record data, analyze data, and communicate the findings to the community.

1.3 What makes "good" visualizations?

Identify which dimensions of data should be matched to which areas

1.3.1 Design Principles

- Expressiveness
  - express all the facts/info in the set of data, nothing less and nothing more
  - choose correct type of chart – ex. a scatter plot cannot express a one to many relation
- Effectiveness
  - More effective if information can be conveyed more quickly
  - Easy to understand

Some things to remember about visualizations:

- Tell the truth and nothing but the truth – i.e. Don't lie!
- Use encodings that people decode better
- Not all visual encodings are equal
  - Some may contain biases for example

1.4 Visualization Effectiveness

Steven's Power law

$$S = I^p$$

where S is the perceived sensation, I is the physical intensity and p is the exponential relation.

1.4.1 Perception Biases

- Area - we underestimate large areas over small areas
- Perception of shock increases quicker than the actual level of shock
- We perceive length and position well, much better than color saturation or pie charts (Cleveland and McGill Experiment)
- We can actually rank human perception biases

1.4.2 Data Types

- **Normal** – non intrinsic ordering – eye color, gender, etc.
- **Ordinal** – contains a natural ordering – socioeconomic class, month, etc.
- **Quantitative** – is described numerically

1.4.3 Other Decisions

Color

- How should I color my plot? – not all colors are equal
- Choose colors that maintain distinguishability
- Small changes in color should correlate with proportional changes in value

Tools

- There is a tradeoff between speed and expressiveness
- Declarative Encoding Languages
  - program by describing *what* not *how*
  - separate specification from execution
  - examples are: HTML/CSS, SQL, D3
  - Advantages
    - faster iteration
    - performance
    - reuse-ability
    - portability
  - Disadvantages
    - debugging is difficult
- The Grammar of Graphics
  - Set of principles for graphical APIs
  - "Don't give a pie, give primitives to make a pie and more"
  - Provide small tools that provide more flexibility and customization

2 ggplot

Visit the Jupyter Notebook for exact source code, some observations are listed here.

- Purpose of a plot is to communicate a 10 word point
- Use geoms to represent data points, use aes() function to add aesthetics, variables, axes, etc.
- Pipe commands using '+' not '%>%' – the data frame will default to first argument
- When using aes(), order of arguments doesn't matter – they are added just as descriptors
- **geom_histogram()** – implicit stat counting happening, then maps the count for you
  - Be sure to specify the number of bins – If you don't, some random number will be assumed and a warning will be thrown
  - Identifies categorical variabels and maps them to bins if needed
- **geom_smooth()** – fits a model to the data
  - specify method="lm" to force linear model
- **geom_density()** – fill in plot
- Be careful what to include in aes() – if it is a constant, best to keep it out of the aesthetic mappings
- Be careful when using xlim – may either zoom into the plot, or eliminate points from computation
- R will default categorical axes to alphabetical ordering – often it is better to change this to something that conveys pattern/point better

Notes from hn2284

1 Introduction

In Lecture 5 of Modeling Social Data, we studied Data Visualization. The first half of the class was a lecture by Cagatay Demiralp on the history and theory of data visualization, and the second half was an introduction to using $ggplot2$, a library in R that is frequently used to perform practical data vizualizations. These notes correspondingly are divided into two parts, first for the lecture by Demirlap and the second for the $ggplot2$ demonstration.

2 Demirlap's Lecture on Data Visualization

2.1 What is Data Visualization?

- Data is everywhere, young field of study—representing it is a challenge
- Lots of different definitions, but recommended one by Demirlap is "The use of computer-generated, interactive, visual representations of data to amplify cognition", with emphasis on *amplyifying cognition*
- Literature has been defining data visualization, with the first example in class going back to 1967

2.2 Why Create Visualizations?

Can be used to...

- record data, for example E.J Marey's sphymograph (Braun 83)
- form a hypothesis, such as John Snow's map in 1854 of cholera across London. One of the first experimental hypotheses
- make a point, inspire, amplify findings (i.e Florence Nightingale's Crimean War Deaths graphs)
- explore data in ways that raw numbers with the naked eye don't allow us to. Figure 1 shows four different datasets that have the same summary statistics, but where the X and Y variables have very different relationships that you need a graph to identify. Another example in class is popularity of John Tukey which spiked on January 12, 2011. By using a visualization of search data on Wikipedia it is possible to identify the anomaly in data for traffic to his wikipedia page, and find the date where he was mentioned on Jeapordy.

In summary, to *Record*, *Analyze*, and *Communicate* information!

2.3 What makes a good visualization?

In the making of any good visualization, there are several decisions that need to be made. These decisions can have a large effect on the utility of the visualization. A good way to understand what makes a good visualization are two specific design principles, outlined in visualization literature

- Expressiveness, can be defined as the extent to which the visualization expresses all the facts in the dataset, and only the facts. Certain visualizations cannot express all the facts of certain kinds of datasets, for example a horizontal dot plot cannot fully express the information of a one-to-many relations dataset. Additionally, expressiveness means not expressing facts that aren't true. *Tell the truth and nothing but the truth.*
- Effectiveness, measured by how readily the information conveyed is perceived, relative to other visualizations. *If the visualization is an encoding, how quickly and accurately can it be decoded by viewers.*

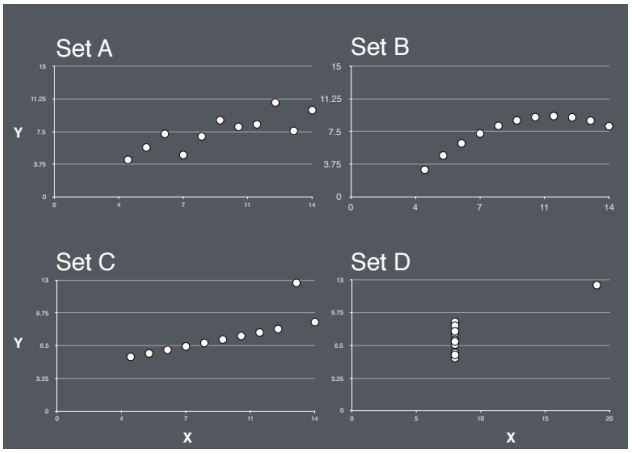

Figure 1: Anscombe's Quartet, a dataset constructed in (Anscombe 1973) to show why graphing data is important

2.4 Steven's Power Law

- $S = I^P$, where S is the perceived sensation, I is physical intensity, and P is an empirally determined exponent that varies based on the way the data is conveyed
- What this means in terms of data visualization is that different ways to convey the same numerical values of data result in different forms of bias when perceived by humans. The closer the exponent P is to 1 for a way to convey data, the more linear the relationship is, making it a better. For example length is preferable to area when showing the relative size of different quantities.
- Effectiveness of different encodings to visualize data have been extensively researched, such as by Cleveland and McGill which compared different porportionality estimates across spatial encodings (position, length, and angle) and the bias associated with them when perceived graphically by users.
- Crowdsourced and in-person studies generally show the same results for effectiveness of data visualizations, making crowdsourcing a cheap and effective alternative to traditional studies.

2.5 Additional Data Stuff

- Three main data types: Quantiative (i.e Position, Length, Area), Ordinal (i.e Texture, Slope, Shape–anything that is categorical but where the data has a natural and intuitive ordering scheme), and Nominal (i.e Color)–which is categorical but with no natural ordering.
- Color is an important consideration, important to take into account color blindess, distinguishability, and also black and white printing (when the colors are transferred to grayscale)

2.6 Tools

- There is a tradeoff between speed and expressiveness, where the most fast tools for data visualization are Excel, Google Charts, etc. that are very user-friendly.
- For developers, data scientists, etc., more expressive tools are libraries like D3, ggplot, etc. Most extreme end are graphical libraries like OpenGL, DirectX which can be used to make never-before-seen visualizations
- Main categories of tools are Chart Typologies, Declarative Encoding Languages, Component Model Architectures, and Graphics APIs (in order from least to most expressive)

- *Declarative Langauges* (as opposed to imperative) focus on what you want from execution, rather than the logical steps (how to execute it).
- *Wilkinson's Grammar of Graphics* which is a theoretical foundation for producting pretty much all visualizations of data, which has inspired libraries such as D3 and ggplot. Outlines a pipeline for the production of graphics, and also discusses Algebraic foundation (Sets, Operators and Rules) for producing such graphics

3 ggplot2

In this section of the lecture we went over how to practically use ggplot for data visualizations. The Jupyter notebook that corresponds to the lecture material can be found HERE (click for link). These notes serve as a companion to the Jupyter notebook with important points that were mentioned in class regarding using ggplot during the demonstration.

- Visualization is *always* about communication something. If you are producting visualizations for research for example, you might be trying to communicate a point to readers with each graph. It is important to keep that point in mind when producing these graphics, as way to shape the decisions you make. Similarly, exploratory data analysis is about communicating a point to yourself.
- Some helpful libraries are $lubridate$, which helps with dates, $scales$ which helps make good scales, and $theme\_set$, which helps change color schemes.
- ggplot is an implementation of the ideas developed by Wilkinson's *Grammar of Graphics*
- It is important to pay attention *warnings* in ggplot. ggplot may do things for you, such as automatically determining the numbers of bins, but this can in some cases be very unreliable, so it is a best practice to always address the warnings.
- can use pipe to pass things into ggplot, but when you're using ggplot you need to use a plus instead of a pipe
- Tip: When using a log scale, it's a good idea to use intervals of 1 and 3
- Important to know the difference between $coord\_cartesian$ and x_lim to know which you actually want to use
- Always be sure to $ungroup()$
- Facets can be helpful at times, and not helpful at other times. They allow you to view the same type of plot for different data.

Notes from rs3505

1 Introduction

Data visualization is an important step in data science, with which we can present the abstract data to people in a more direct way. It can help us explain, study and find the information we want from the original raw data.

2 What is data visualization

1. There can be many different definitions which emphasize on different perspective to describe data visualization
2. One of the definition is the use of computer generated entire representation of data to amplify commission

3 Why create visualization

1. Record information
   Example: Medical devices in 19th century to measure blood pressure
2. Verify and support hypothesis
   Example:John Snow uses scatter plot to map of London to study the relationship between cholera and water use on Broad Road, which is the first natural experiment.Convinced other people about the hypothesis
3. Inspire others
   Visulization as the artificial memory, extract data and make sense of it
   Example: Four sets of data in different patterns,but they can have the same mean value and same standard derivation

4 Expository data analysis

Example: John Tukey, the wikipedia popularity of him, and we can find that at some time points, the popularity becomes very high, so there can be some incident happen at this time point
On January 13th,2011, there is a very high peak–A question:What is software

5 What is good visualization

1. Should be able to express whatever in your data nothing much, nothing more
2. Different visualization tools can express different kind of data
   Example: 1-D data cannot show correlation
3. One visualization can be more effective than another if you can express faster and more accurately
   Activity: Guess the area/length ratio
   Steven Power Law:
   $$\S = I^P \tag{1}$$
   Different shapes have different p, for example, length is the easiest to estimate and p 1
   Experiments:
   1:5 charts, ask people to estimate the position on y axis in different bars
   2:Judge the ratio with two charts:bar and pie chart
4. Ranking of effectiveness: Position is the most effective¿norminal/ordinal/quantitative

6 Color

1. Color of charts can have some effects on the expression of data
2. Color mapping should also be somewhat linear and have a constant using measurement
3. Small changes in the colors should correspond to the linear perceptual changes in per section
4. Take care of color blind people

7 Tools

1. There is a balance between speed to use and expressiveness
2. Declarative languages
3. Advantages and disadvantages of declarative languages
   difficult to debug,etc
4. Example: Population for some cities at different time
5. Do not give users a pie, give them how to make a pie

Notes from ts2837

1 Part 1: Guest Lecture

In the first part of today's class, we had Çağatay Demiralp, reseacher from IBM T. J. Watson Research Center, present on data visualization. You can view his slides here. The following is a summary, although much is borrowed from his slides.

1.1 What is visualization?

Why create visualizations? Some examples:

- E.J Marey's sphygmograph (Braun 83)
- John Snow's mapping of cholera cases on Broad St. (Tufte 83)
- Florence Nightingale's Crimean War Deaths (1856)

Jacques Bertin, cartographer and theorist, considered *"visualization as the artificial memory"*:

- Consider time taken for multiplication by mental calculation vs pen and paper.
- Anscombe's quartet: Four datasets have identical summary statistics and linear regression lines, but appear very different when graphed. This is shown in Figure **??**.

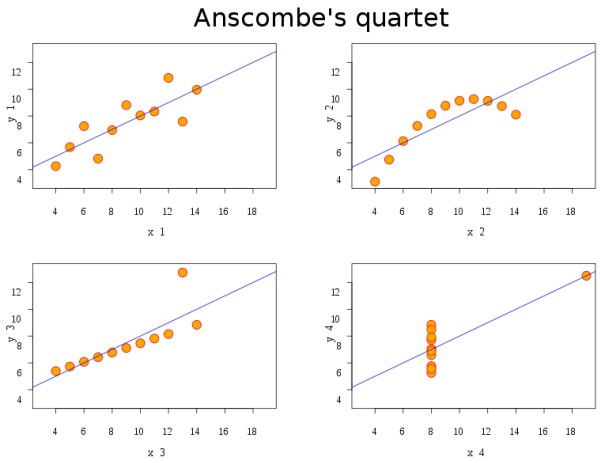

Figure 2: Anscombe's quartet - With proper encoding of the data, we are able to make more sense of it.

Figure 3: *
Source: Wikipedia.org

- Popularity of John W. Tukey (1915-2000) on Wikipedia: After observing the number of page views across time, we notice a spike in January 2011 (Figure **??**). Why?
  - Jeopardy Final Round #6063 on Wednesday, January 12, 2011: "John Tukey coined this compound word in 1958 saying it was as important as tubes, transistors, wires, tapes..."
  - Answer: *What is software?*

Bottom line: We create visualizations to

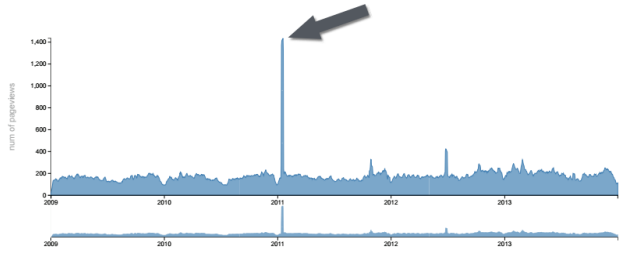

Figure 4: Popularity of John W. Tukey on Wikipedia across time.

Figure 5: *
Source: Çağatay Demiralp

- Record information
- Analyze data to support reasoning
- Communicate information to others

1.2 What is a "good" visualization?

Visualization specification involves several choices, i.e. graph type, use of color, etc... How much do these choices matter?

Design Principles (Mackinlay 86)

- Expressiveness - A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data. Watch out for visualizations that:
  - Cannot express the facts, e.g. a one-to-many relation cannot be expressed in a single horizontal dot plot
  - Express facts not in the data, e.g. length of bar in graph says something untrue about the data
- Effectiveness - A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Principles - animation (B. Tversky 02)

- Congruence - The structure and content of the external representation should correspond to the desired structure and content of the internal representation.
- Apprehension - The structure and content of the external representation should be readily and accurately perceived and comprehended.

Design Principles translated:

- Tell the truth and nothing but the truth.
- Use encodings that people decode better.

1.3 Not all visual encoding variables are created equal...

Stevens Power Law (Figure **??**) describes the relationship between the magnitude of a physical variable and its perceived sensation:

$$S = I^p \tag{2}$$

where S is perceived sensation, I is physical intensity, and p is an empirically determined exponent. For example, length (such as comparing the length of bars) is perceived linearly, whereas area (such as comparing the area of circles) is underestimated. Stevens' Power Law predicts bias, not necessarily accuracy.

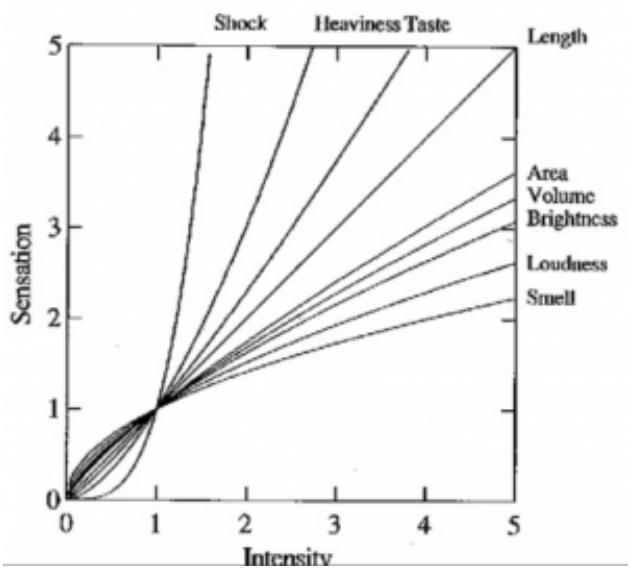

Figure 6: Stevens' Power Law

Figure 7: *
Source: Stevens, S. The psychophysics of sensory function. Am. Sci. 48, 226–253 (1960).

Graphical Perception (Cleveland & McGill 84)

- Task in Experiment I (position - length), Figure **??**: "For the two marked bars or divisions, what percent the smaller is of the larger?"
  Type 1, 2, and 3 compare "position" along a common scale, while Type 4 and 5 compare "length".
- Task in Experiment II (position - angle), Figure **??**: "What percent each of the other segments or bars is of the largest?"
  The pie chart on the left compares "angle" while the bar chart on the right compares "position".
- Figure **??** shows the results of Cleveland and McGill's Experiment I (top) and II (bottom). Bias was measured by the log absolute estimation error.

1.4 Not all data types are created equal...

Data variable types:

- *Nominal* - has two or more categories, but no instrinsic ordering
- *Ordinal* - has two or more categories with natural ordering
- *Quantitative* - numerical variables
- Mackinlay (86) published effectiveness rankings on the different data variable types.

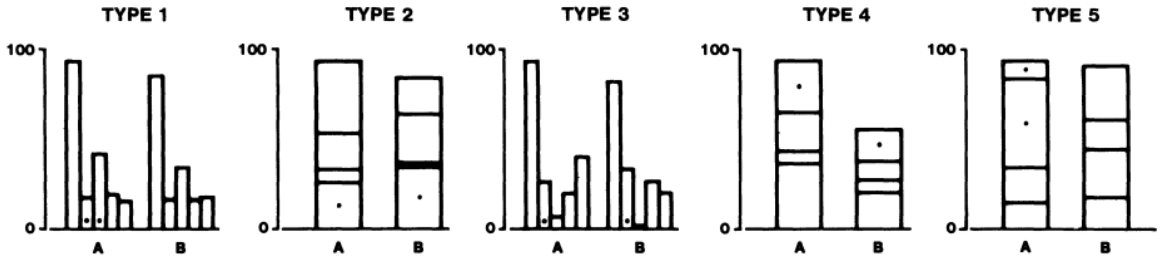

Figure 8: Experiment I (position - length)

Figure 9: *
Source: Cleveland, W. & McGill, R. Graphical perception: Theory, experimentation, and application to the development of graphical methods. J. Am. Stat. 79, 531–554 (1984).

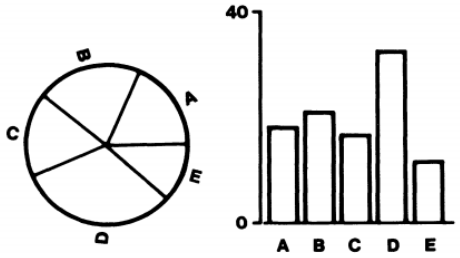

Figure 10: Experiment II (position - angle)

Figure 11: *
Source: Cleveland, W. & McGill, R.

1.5 Color

How should I color my bar chart?

Color Design Guidelines

- Maintain perceptual distinguishability
- Use only a few (¡7) & named, when possible
- Avoid unintended "pop out" of colors
- Get it right in black & white
- Respect cultural norms & the color blind
- Beware of segmentation (rainbows!)
- Consider spatial frequency effects

1.6 Tools

Which tool should I use to create my bar chart?

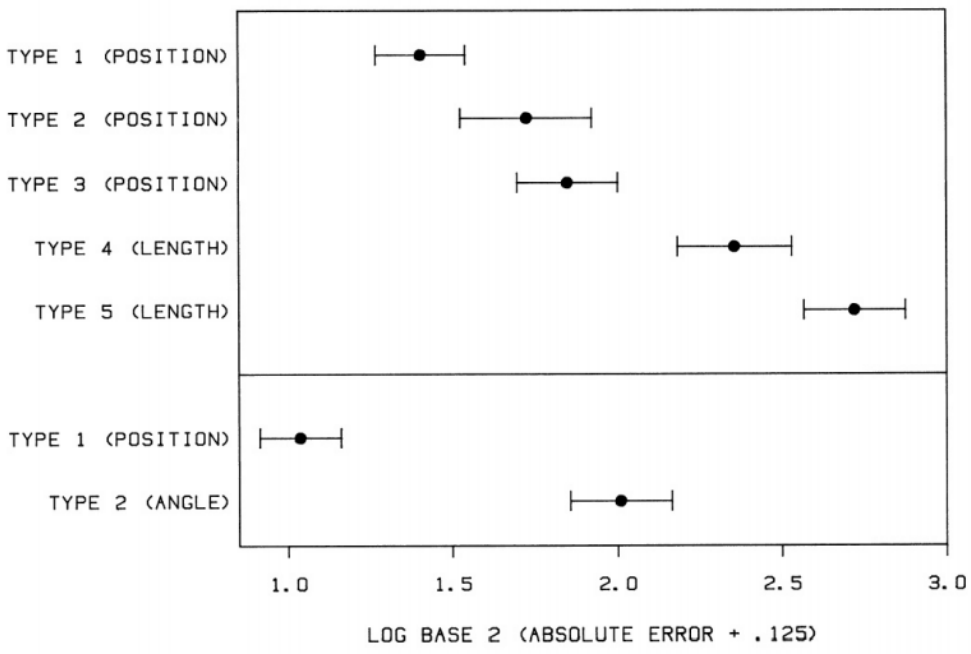

Figure 12: Results

Figure 13: *
Source: Cleveland, W. & McGill, R.

- Tools come and go, but underlying ideas are important
- Consider the expressiveness vs the speed of tools

Declarative languages:

- Programming by describing *what*, not *how*
  - In contrast to imperative programming, where you must give explicit steps
- Separate specification (what you want) from execution (how it should be computed)
- Examples include HTML/CSS, SQL, and D3
- Advantages:
  - Faster iteration
  - Less code
  - Larger user base than imperative languages
  - Potential to create better visualization - Smart defaults
  - Reusable - Write-once, then re-apply
  - Easier programmatic generation - Write programs which output visualizations e.g., automated search & recommendation

The Grammar of Graphics (Wilkinson)

- Algebra is useful! (Sets, operators, rules)
  - Operators: + (blend), * (cross), / (nest)

- Geometric primitives (marks)
  - "Don't give a pie, give primitives to make a pie and more!"

1.7 The Upshot

Always consider what the one point you are trying to make with this visualization is. Then, how can you make that point the most obvious thing when your visualization is seen?

2 Part 2: Intro to ggplot2

In the second part of today's class, Professor Hofman went over some basic ideas in ggplot2 and how to effectively use visualization tools like ggplot2 to convey a point. To do this, we revisited the Movielens data from Lecture 2. You can find his code in this Jupyter notebook. The following are some pointers and best practices that Professor Hofman mentioned throughout the demo:

- Convert timestamps to datetime objects using package *lubridate* for easy manipulation
  - For example, `year(datetime_object)` can quickly extract the year
- General idea of ggplot2:
  `ggplot(<dataframe>, 'aes' aesthetic-mapping('x' variable = column_name)) + geom_type`
- Use piping operator %>% (similar to command line) so that code is easier to follow
  - But in ggplot2, remember to use + to precede `geom_histogram/geom_point/geom_line` etc. instead of %>% because geoms implicitly compute y variable counts.
- If x is categorical, i.e. `as.factor(rating)`, ggplot will automatically bin by category instead of breaking up by numerical intervals
- `scale_y_continuous(label = comma)` adds commas to give you easily readable numbers on your y axis
- Good practice for making a plot with discrete data: Plot data as points, your model as the line, and optional confidence band as an underlying shade
- Changing the window of your data:
  - Using `coord_cartesian(xlim = c(starting_value, ending_value))` will only change the visual window of your graph, not the data your graph is using
  - Use `xlim(c(starting_value, ending_value)` instead, which limits your data before any transformation occurs
- Be careful when ignoring warnings!
- Playing around with log scales may give the data a completely different story
- `geom_density(fill = <color>)` can be used to smooth and fill the area under a graph
- Consider using a vertical bar to depict mean: `geom_vline(aes(intercept = mean(num_ratings)), linetype = 'dashed')`
- Keep in mind when you are specifying something constant (which you would want to keep outside the aesthetic mapping) vs transforming something IN the data (keep inside the aesthetic mapping)
- `cumsum` gives a running sum of values down the column, whereas `sum` gives one value
- `coord_flip` is good for seeing long x-axis labels like movie titles
- Be aware of the ordering of factors (`mutate(title = reorder(title, mean_rating))`)

- Every design choice depends on the point you want to make!!
- If you want to split up data and visualize individual plots together, you can use `title` to break out facets: `facet_wrap(~ title)`
- Add standard errors to a line graph with `geom_ribbon`
- `extract` can pull out specific texts (regex)