# Introduction to Data Visualization

## Choosing the Right Visualization Software




**Halil Bisgin, Ph.D.**


# What is the best visualization tool?

- People have strong emotional bonds to the specific tools they are familiar with.
- Instead of investing time in learning a new approach, people vigorously tend to defend their preferences.
- Sticking one tool is not an unreasonable choice since learning new one will require time and effort.
- Finally, if you can make the figures you want to make, without excessive effort, then that's all that matters.

# How to assess a tool?

- We can assess any new approach through some principles.
- These principles roughly break down by
  - *how reproducible the visualizations are*
  - *how easy it is to rapidly explore the data*
  - *to what extent the visual appearance of the output can be tweaked*

# Reproducibility and Repeatability

- A work is reproducible if the overarching scientific finding of the work will remain unchanged if a different research group performs the same type of study.
  - *A drug being tested by two different group and results remain the same*
- A work is repeatable if very similar or identical measurements can be obtained by the same person repeating the exact same measurement procedure on the same equipment.
  - *Me weighing my cat in two different occasions on the same scale and finding it 9lbs.*

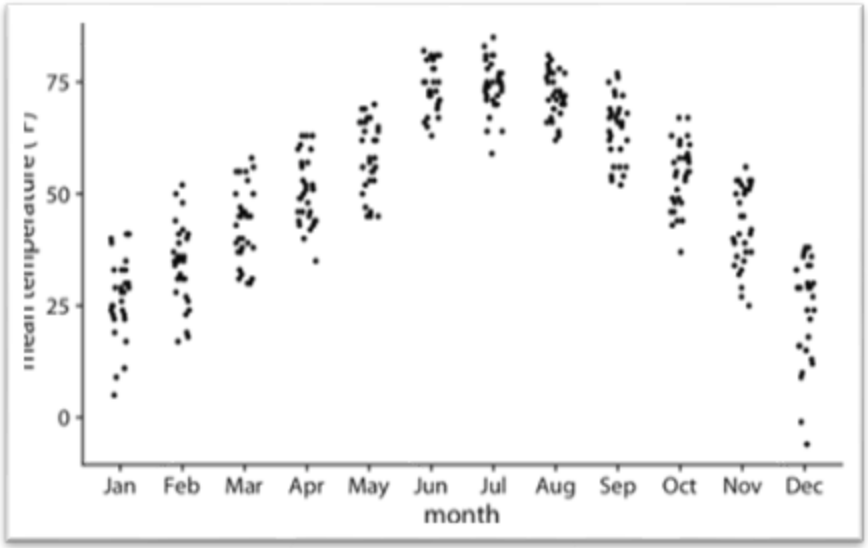
# Reproducibility in visualization

- A visualization is reproducible if the plotted data is available and any data transformations that may have been applied before plotting are exactly specified.
  - *If you make a figure and then share the exact data that you plotted, then anyone else can prepare a figure that looks substantially similar.*
  - *Fonts or colors or point sizes may differ, so the figures may not be exactly identical, but two figures convey the same message and therefore are reproductions of each other.*

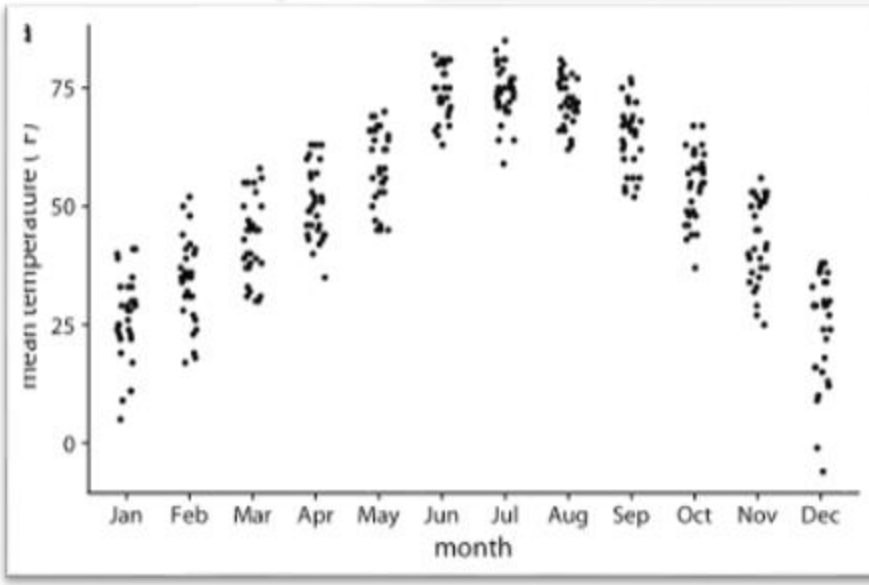
# Repeatability in visualization

- A visualization is repeatable, if it is possible to recreate the exact same visual appearance, down to the last pixel, from the raw data.
- Repeatability requires that even if there are random elements in the figure, such as jitter, those elements were specified in a repeatable way and can be regenerated at a future date.
  - *For random data, repeatability generally requires that we specify a particular random number generator for which we set and record a seed.*

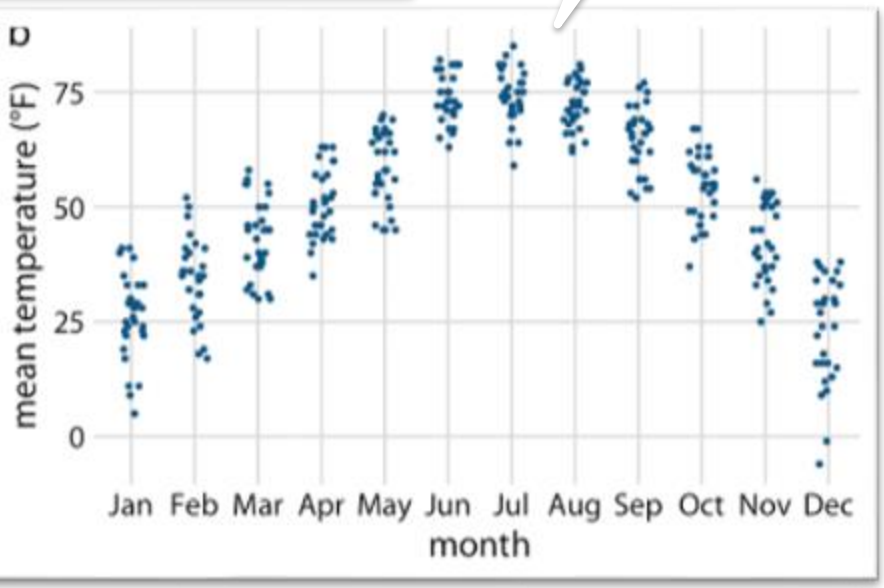
# Repeated vs. Reproduced




Repeated




Reproduced



# RR can be difficult w/ interactive tools

- Many interactive programs allow you to transform or otherwise manipulate the data, but don't keep track of every individual data transformation you perform, only of the final product.
  - *E.g., Tableau*
- If you make a figure using this kind of program, and then somebody asks you to reproduce the figure or create a similar one with a different dataset, you might have difficulty doing so.

# Data Exploration vs. Data Presentation

- Two distinct phases of data visualization, and they have very different requirements.
- Data exploration:
  - *Looking at it from different angles and trying various ways of visualizing it, just to develop an understanding of the dataset's key features*
  - *Trying different types of visualizations, different data transformations, and different subsets of the data*
- Data presentation:
  - *The key objective to prepare a high-quality, publication-ready figure.*

# Primary focus in data exploration

- Whether the figures look appealing is secondary.
  - *It's fine if the axis labels are missing, the legend is messed up, or the symbols are too small, as long as you can evaluate the various patterns in the data.*
- A well-designed data exploration tool will allow you to easily change which variables are mapped onto which aesthetics with a wide range of different visualization options.
- Check whether your software allows for rapid data exploration or whether it tends to get in the way.

# Different venues for representation

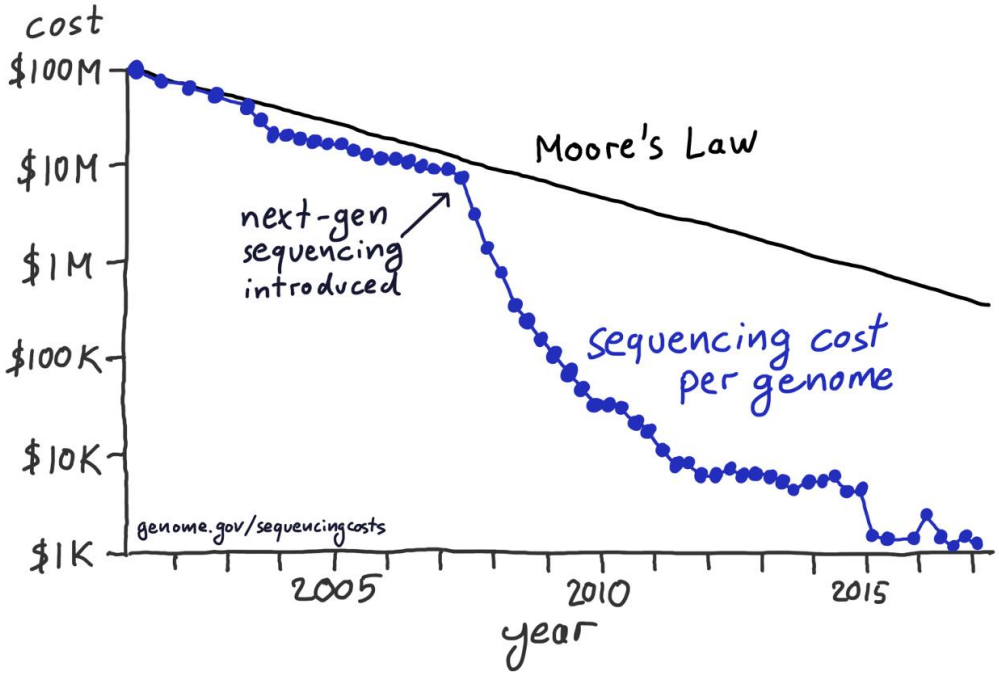
- We can finalize the figure using the same software platform we used for initial exploration. (e.g., R→R)
- We can switch to a platform that provides us finer control over the final product, even if that platform makes it harder to explore. (e.g., Python → R)
- We can produce a draft figure with visualization software and then manually post-process it with an image manipulation program. (e.g., R→Photoshop)
- We can manually redraw the entire figure from scratch, either with pen and paper or using an illustration program.

# Be cautious about manual intervention

- All these avenues are reasonable, but manually sprucing up figures in routine data analysis pipelines or for scientific publications is not recommended.
- Manual steps in the figure preparation pipeline make repeating or reproducing a figure inherently difficult and time-consuming.

# Hand-drawn and post-processed figures

- Not totally against manual processing.
  - *change axis labels, add annotations, or modify colors.*
- These approaches can yield beautiful and unique figures that couldn't easily be made in any other way.



# Separation of Content and Design

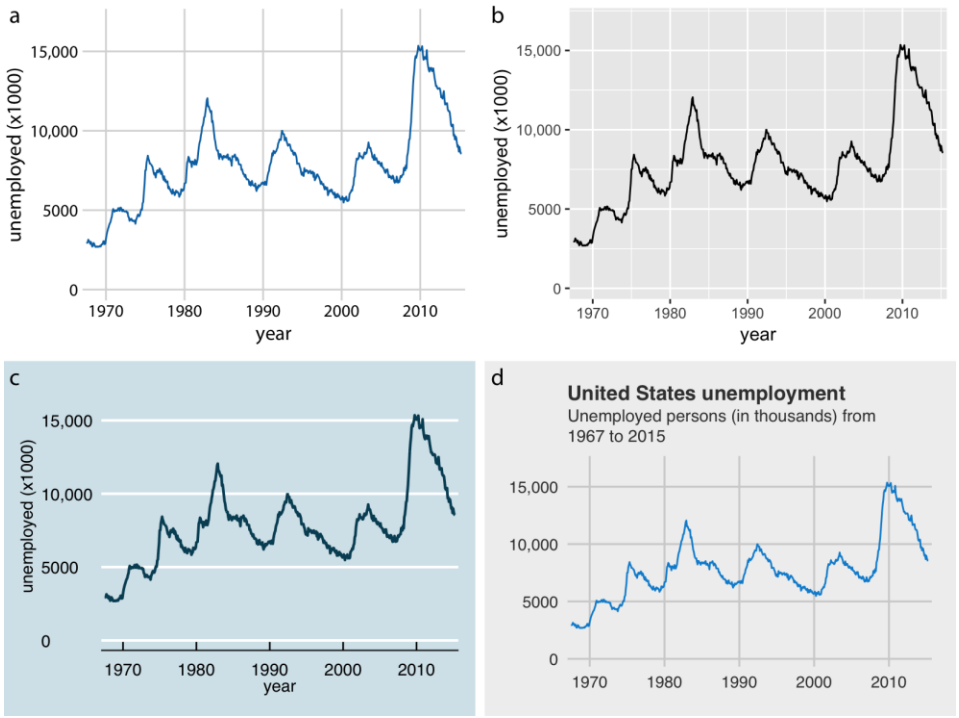
- Good visualization software should allow you to think separately about the content and the design of your figures.
- Content:
  - *the specific dataset shown, the data transformations applied (if any), the specific mappings from data onto aesthetics, the scales, the axis ranges, and the type of plot (scatterplot, line plot, etc.)*
- Design:
  - *features such as the foreground and background colors, font specifications (e.g., font size, face, and family), symbol shapes and sizes, whether or not the figure has a background grid, and the placement of legends, axis ticks, axis titles, and plot titles.*

# Content + Design

- You can first determine what the content should be, using the kind of rapid exploration.
- Once the content is set, you can tweak the design, or you can apply a predefined design that you like and/or that gives the figure a consistent look in the context of a larger body of work.

# Separation example in ggplot2

- Separation is achieved via themes.
- A theme specifies the visual appearance of a figure.
- It is easy to take a figure and apply different themes.



# Separation helps focus

- Most data scientists are not designers, and therefore their primary concern should be the data, not the design of a visualization.
- Most designers are not data scientists, and they should be able to provide a unique and appealing visual language for figures without having to worry about specific data, appropriate transformations, and so on

# Summary

- Think about how easily you can reproduce and redo.
- Consider if you can rapidly explore different visualizations of the same data, and to what extent you can tweak the visual design.
- It may be beneficial to use different tools for the data exploration and data presentation stages.
  - *You can do final visual tweaking interactively or by hand*
- If you have to make figures interactively, consider taking careful notes on how you make each figure for reproducibility.

# Introduction to Data Visualization

## Story Telling with Your Data

**Halil Bisgin, Ph.D.**

# Data visualization for communication

- We have an insight about a dataset, and we have a potential audience, and we would like to convey our insight to our audience.
- To communicate successfully, we will have to present the audience with a meaningful and exciting story.
- The need for a story may seem disturbing to scientists and engineers, who may equate it with making things up, putting a spin on things, or overselling results.

# Stories play in reasoning and memory

- We get excited when we hear a good story, and we get bored when the story is bad or there is no story.
- Any communication creates a story in audiences' minds.
- If we don't provide a clear story ourselves, then our audience will make one up.
  - *In the best-case scenario, the story they make up is reasonably close to our own view of the material presented. However, it can be and often is much worse: "this is boring," "the author is wrong," or "the author is incompetent."*

# What Is a Story?

- A story is a set of observations, facts, or events, true or invented, that are presented in a specific order such that they create an emotional reaction in the audience.
  - *Tension at the beginning - some type of resolution toward the end.*

"Stephen Hawking. was diagnosed with motor neuron disease at age 21—one year into his PhD—and was given two years to live. Hawking did not accept this predicament and started pouring all his energy into doing science. He ended up living to be 76, became one of the most influential physicists of his time, and did all of his seminal work while being severely disabled."

Opening–Challenge–Action–Resolution

# Story formats

- Lead–Development–Resolution

"The influential physicist Stephen Hawking, who revolutionized our understanding of black holes and of cosmology, outlived his doctors' prognosis by 53 years and did all of his most influential work while being severely disabled."

Lead

- Action–Background–Development–Climax–Ending

The young Stephen Hawking, facing a debilitating disability and the prospect of an early death, decided to pour all his efforts into his science, determined to make his mark while he still could

Early connection

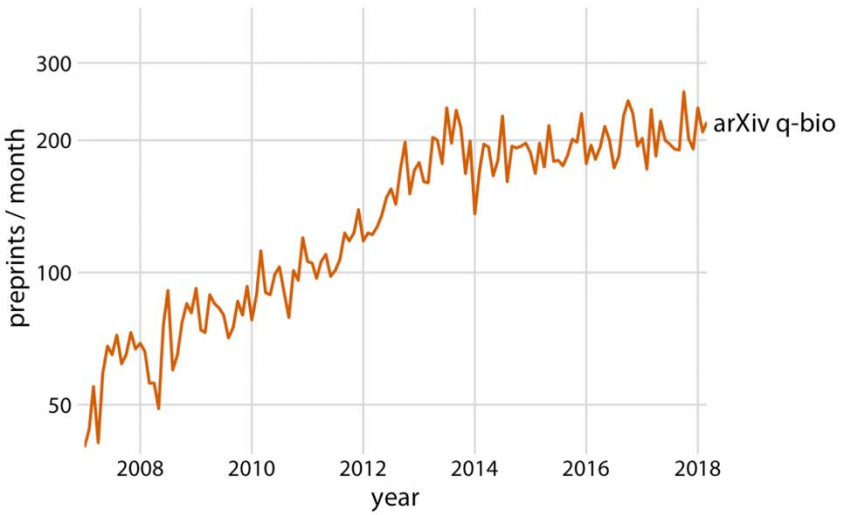
- Opening–Challenge–Action–Resolution

# Your story in visuals

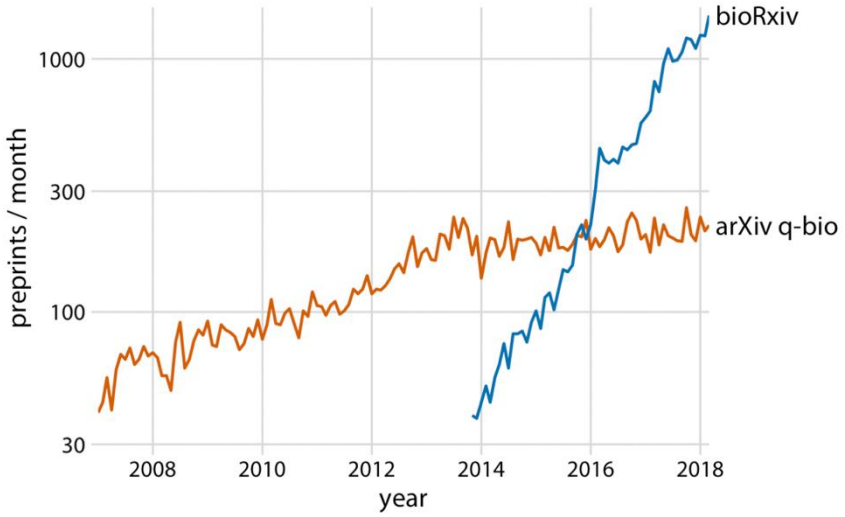
- One visual rarely enough.
- Story formats in visualizations.
- Need multiple visualizations for a complete story.

# Challenge and resolution

- Visualizing number of submissions to pre-print servers in a two-slide story.




Almost no growth after 2014 in arXiv.org



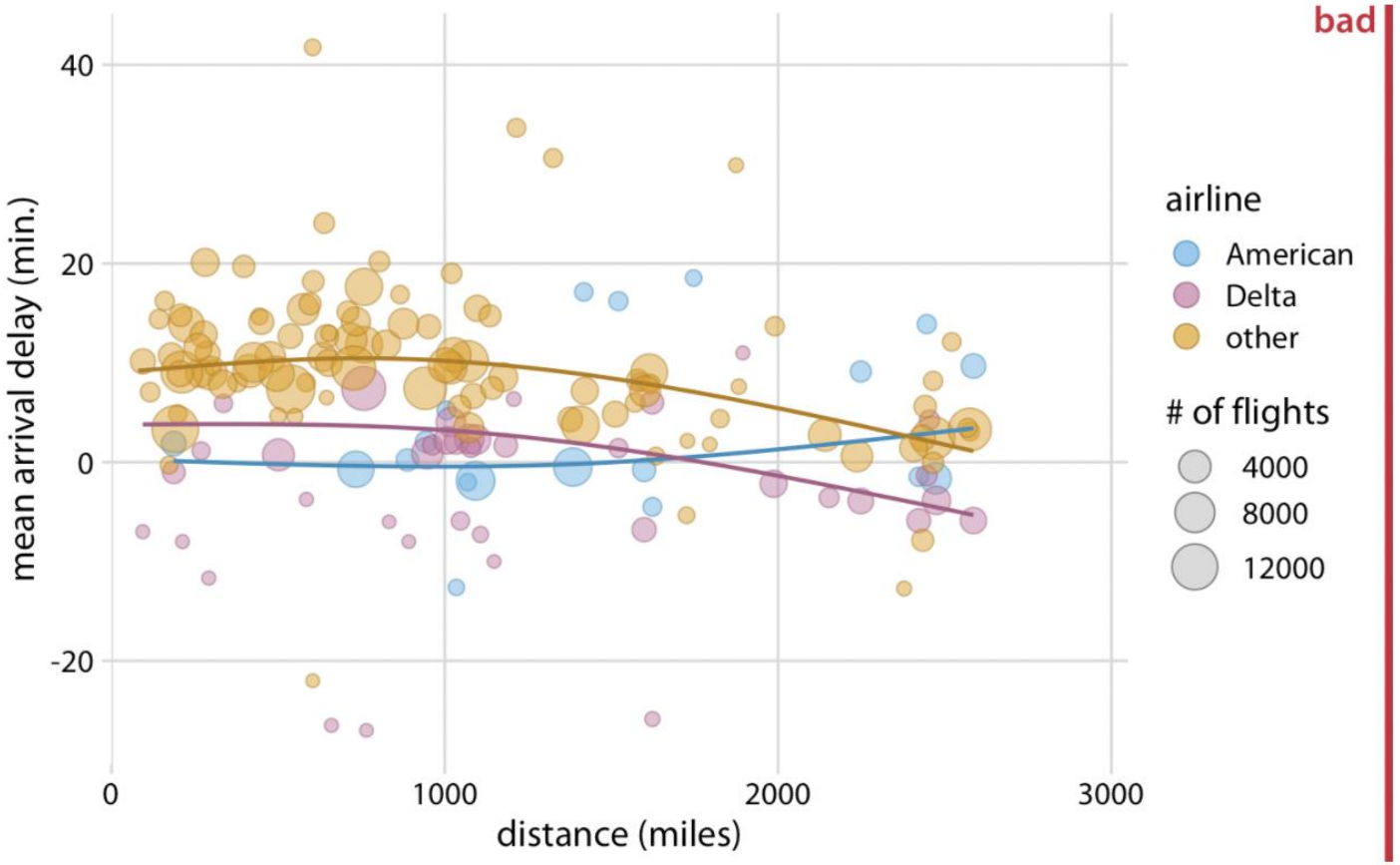

Introduction of the bioRxiv just before 2014

# How to make a figure for the Generals?

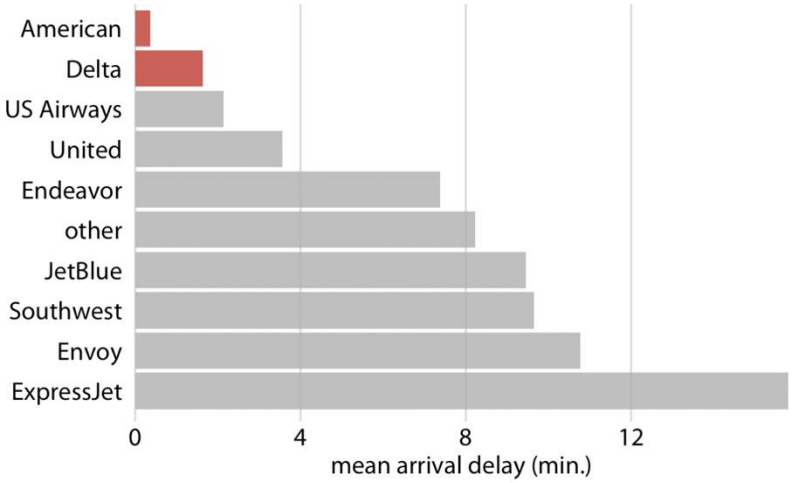
- You need to show your audience figures they can actually understand.
- Never assume your audience can rapidly process complex visual displays.
- Anyone outside your domain, especially those who don not have enough time, should be able to make inferences in a short time.

# Impressive images may still be complex

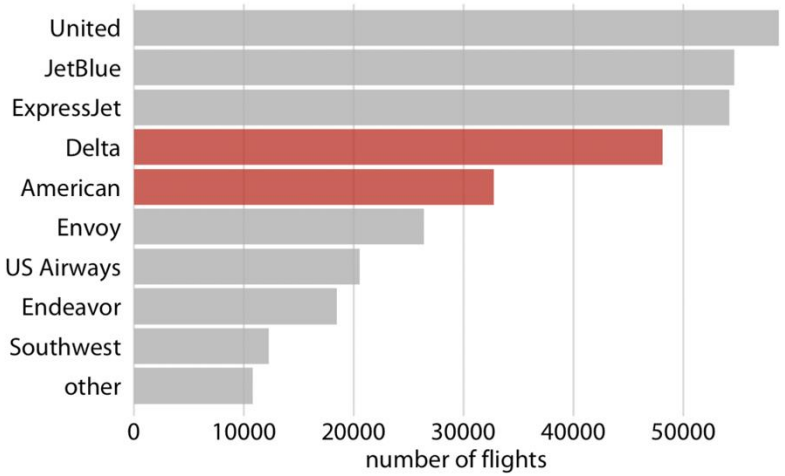
- Too much to process.



# Simplified airline report



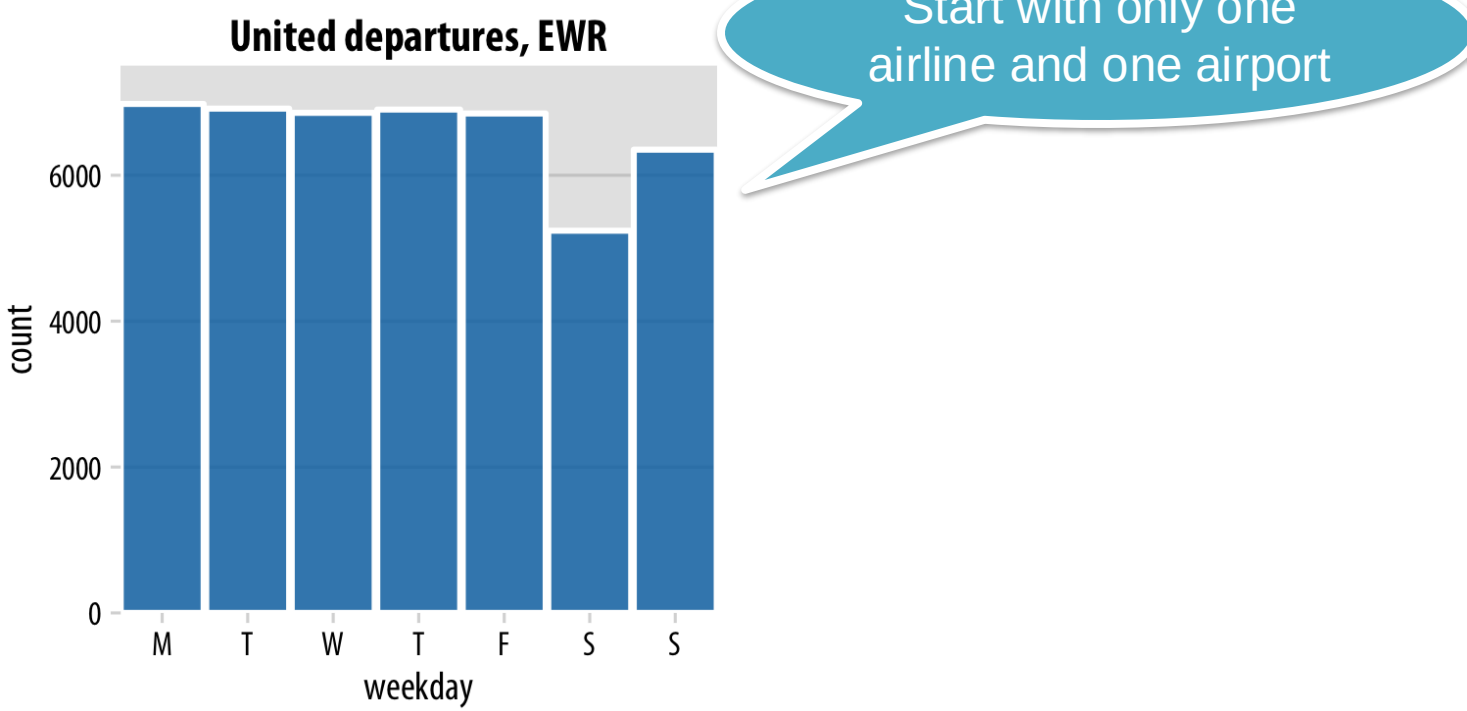

If you want to highlight the least delayed




If you wonder why these two. They fly only NYC area

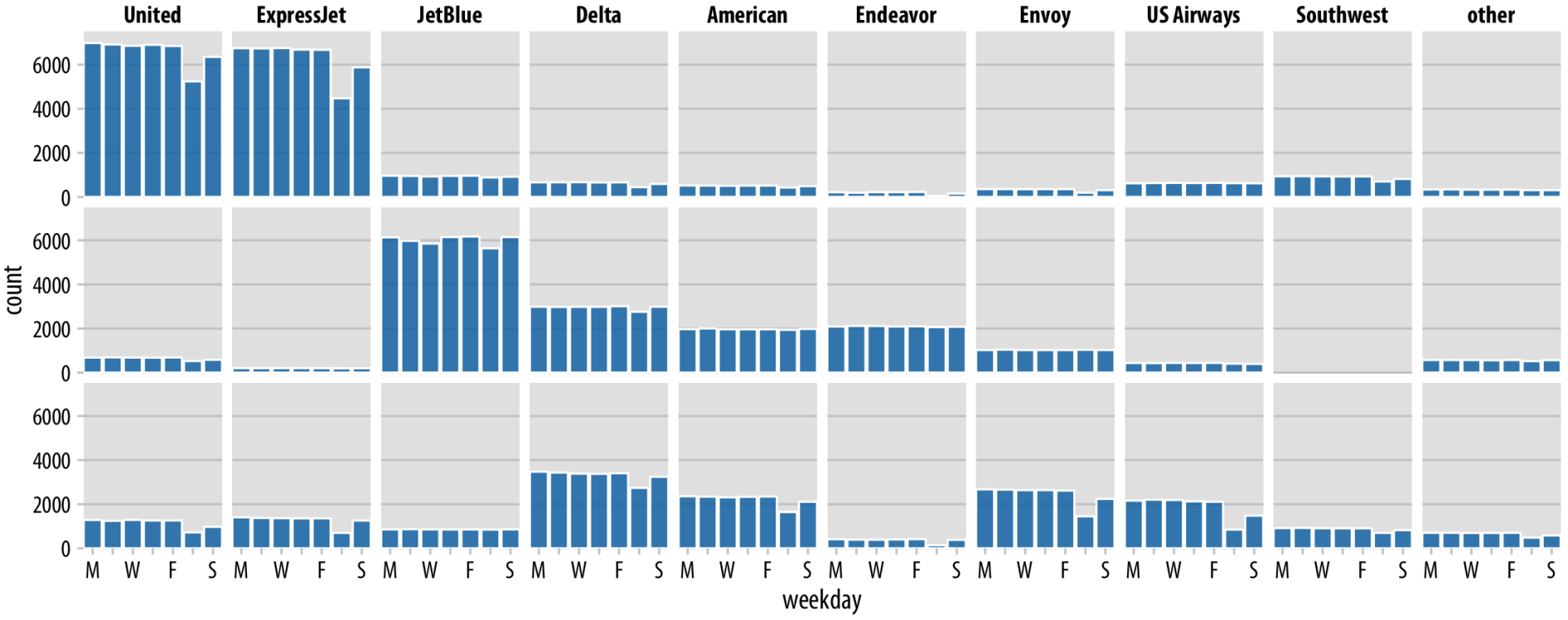
# Build Up Toward Complex Figures

- You may want to show more complex figures that contain a large amount of information at once.
- Start a simplified version of the figure.



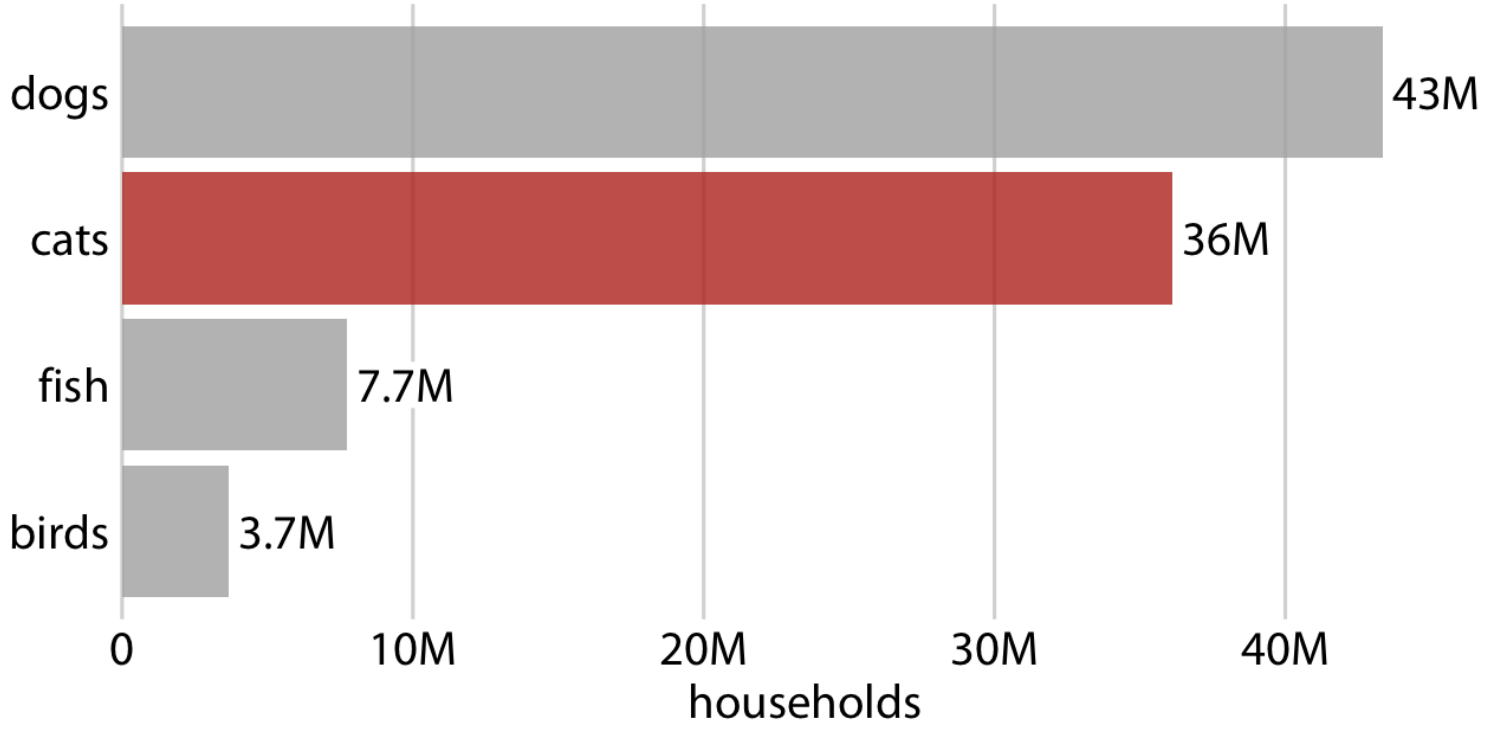
# Then 10 airlines and 3 airports at once

- Multi-panel plot for each airport and airline



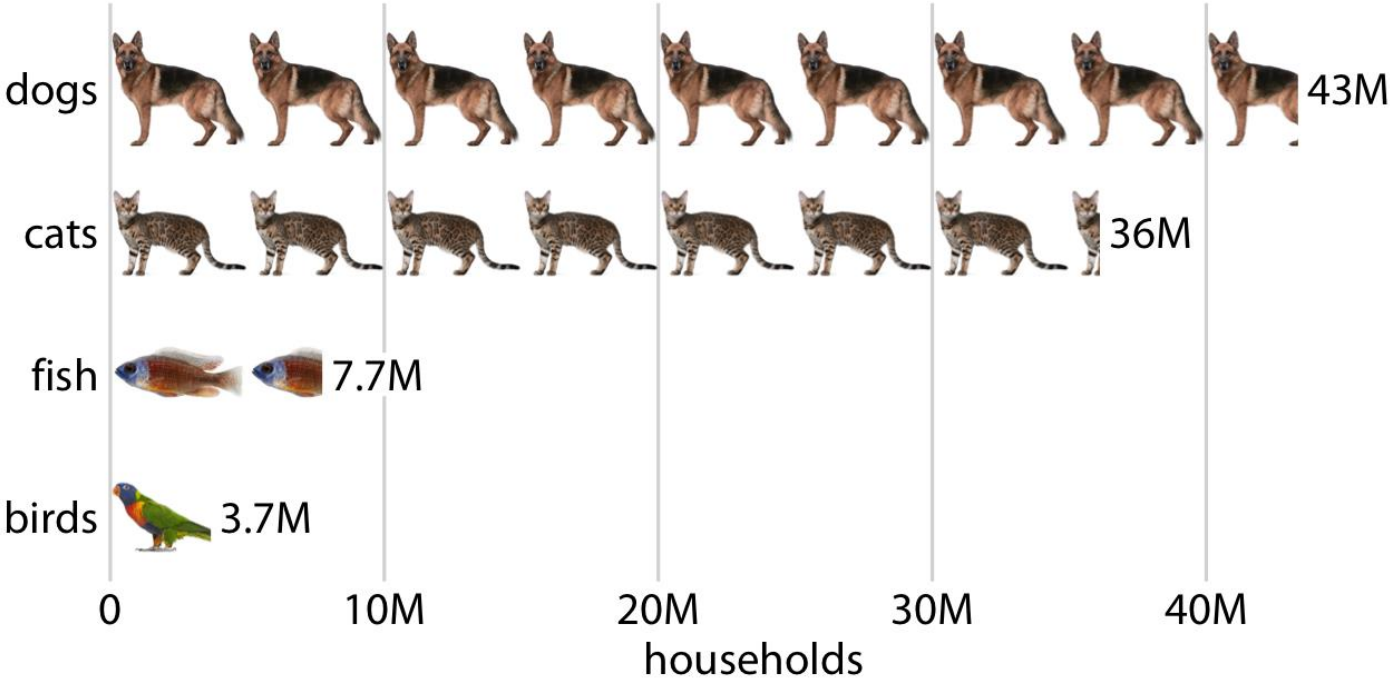
# Make Your Figures Memorable

- Simple is good, but people may easily forget something common.



# A modified "bar plot"

- Strike a balance between the two extremes and make our figures both memorable and clear.
- Memorability is not a concern for publications, but carry importance for other venues such as blogs.

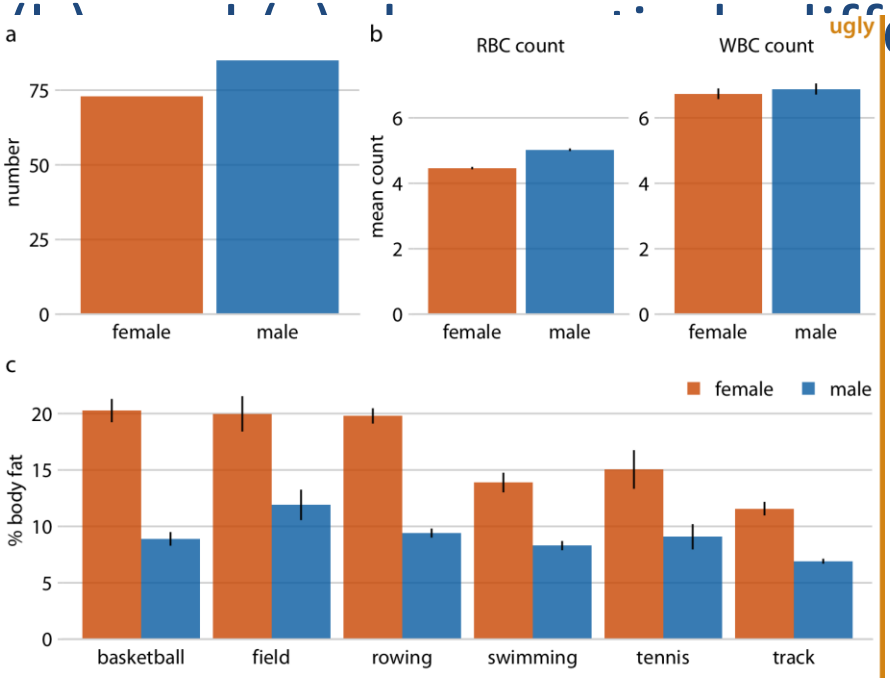


# Be Consistent but Don't Be Repetitive

- Using a consistent visual language does not mean everything should look exactly the same.
- It is important that figures describing different analyses look visually distinct, so that your audience can easily recognize where one analysis ends and another one starts.
- Use different visualization approaches for different parts of the overarching story.
  - *If you have used a bar plot already, next use a scatterplot, or a boxplot, or a line plot.*

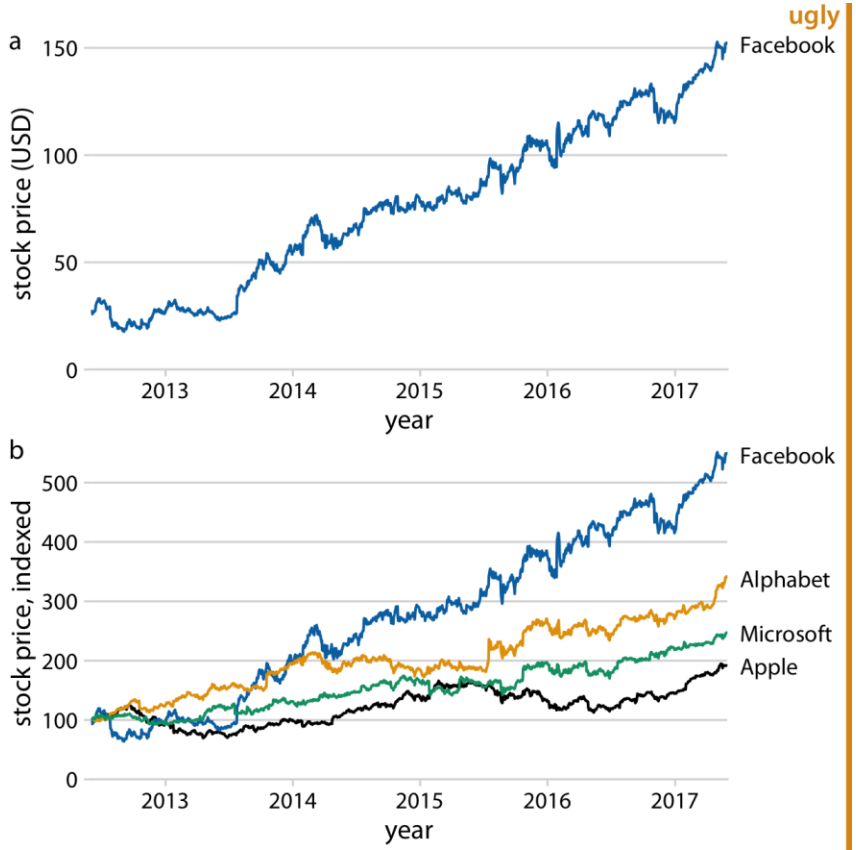
# Repetitive visualization example

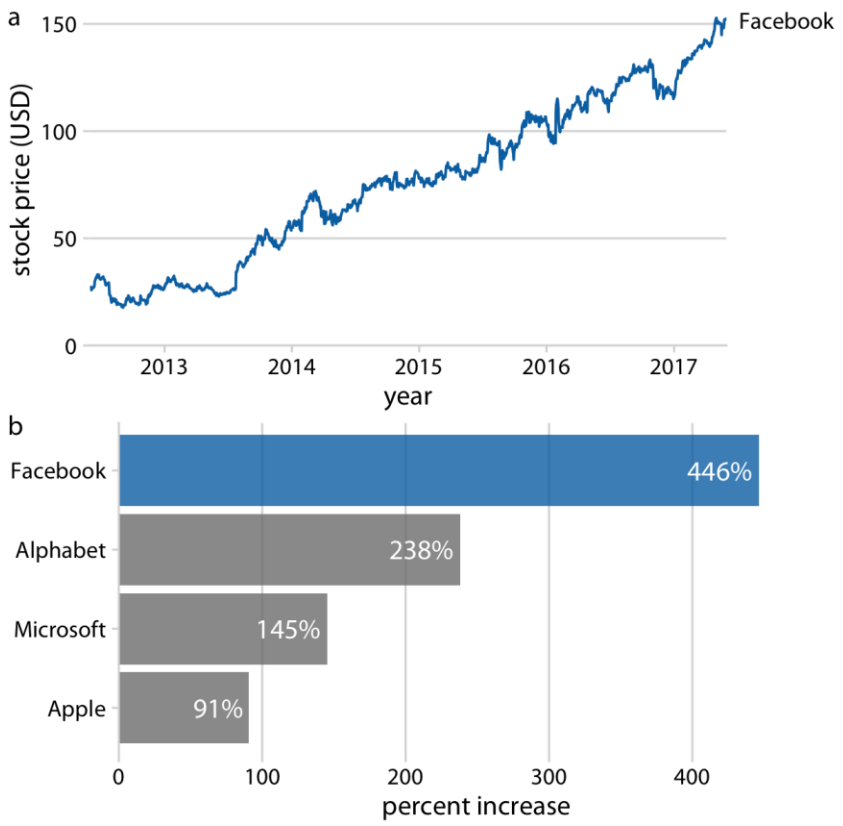
- Uses consistent visual language, but all the subfigures use the same type of visualization (bar plots).
- This makes it difficult for the reader to process that parts (a), (b), and (c) show conceptually different results.



# Diversify plot types to hold attention

- Avoid repetition even for the same raw data
  - *Facebook vs. other tech companies from the same data source*



# Thank you!