ERSP
UC San Diego
Early Research Scholars Program

CSE 193: Data Visualization

Sit with your research group

Last Week

Methods in CS Research

- Variables
- Dependent vs. Independent vs. Control Variables
- Experimental and Non-experimental Research Methods
- Internal and External Validity
- Mixed Methods Research
- Descriptive and Inferential Statistics

Proposal Writing Stages

- Research Context and Problem Statement
- Proposal Solution
- Evaluation and Implementation Plan

Data Collection and Visualization

- Data Science Pipeline
  - Data Collection
  - Getting ready (Cleaning and Normalizing datasets)
  - Descriptive Statistics
  - Inferential Statistics

- Data Visualization
  - Tables??
  - Figures

Why do we perform data visualization?

Why do we perform data visualization?

Anscombe's Quartet

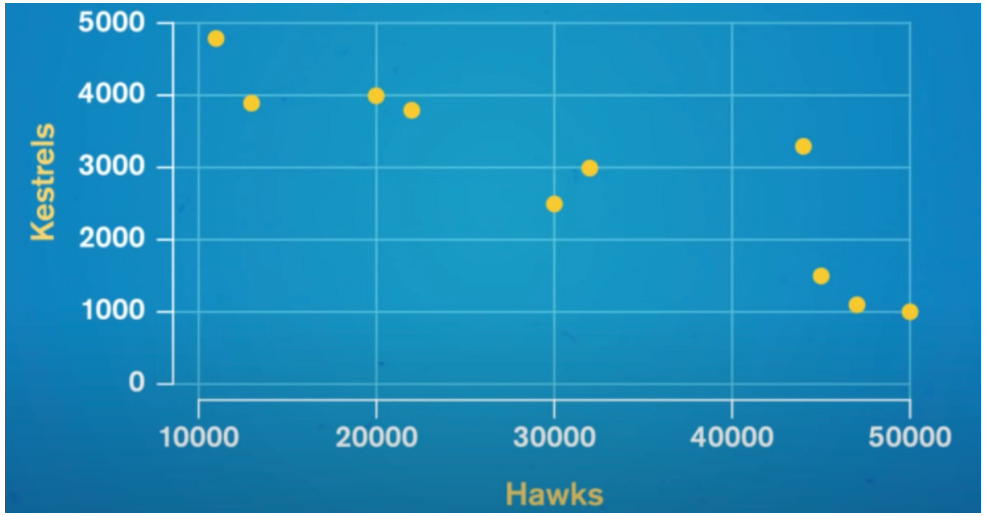
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.75
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

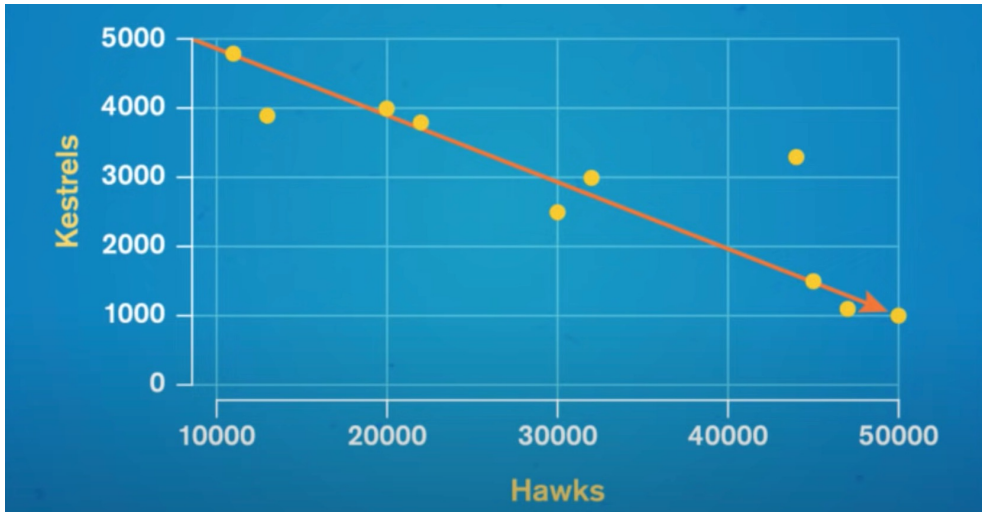
If there is a story in this table, it is completely hidden.

Scatter Plot

Tells a story about the relationships between variables

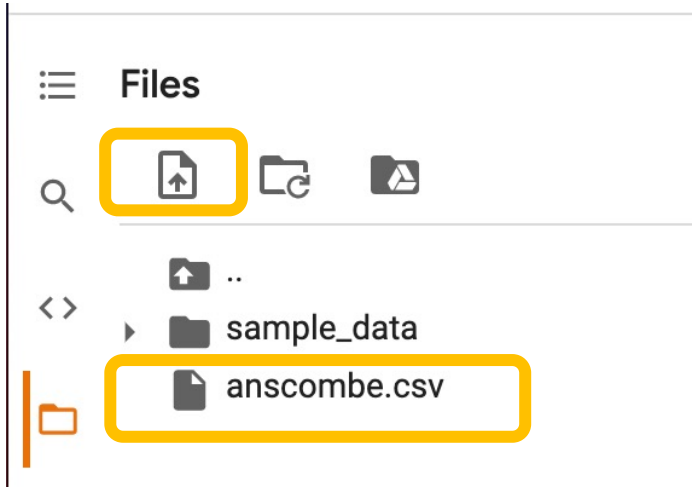
- Correlation





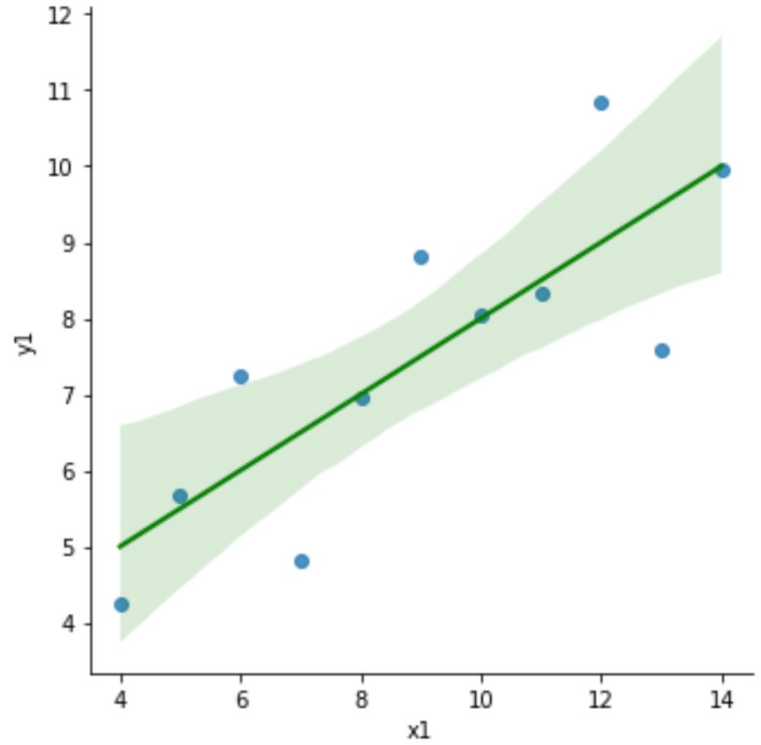
Google Colab

- **Colab** notebooks allow anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education
- File > Save a Copy in Drive

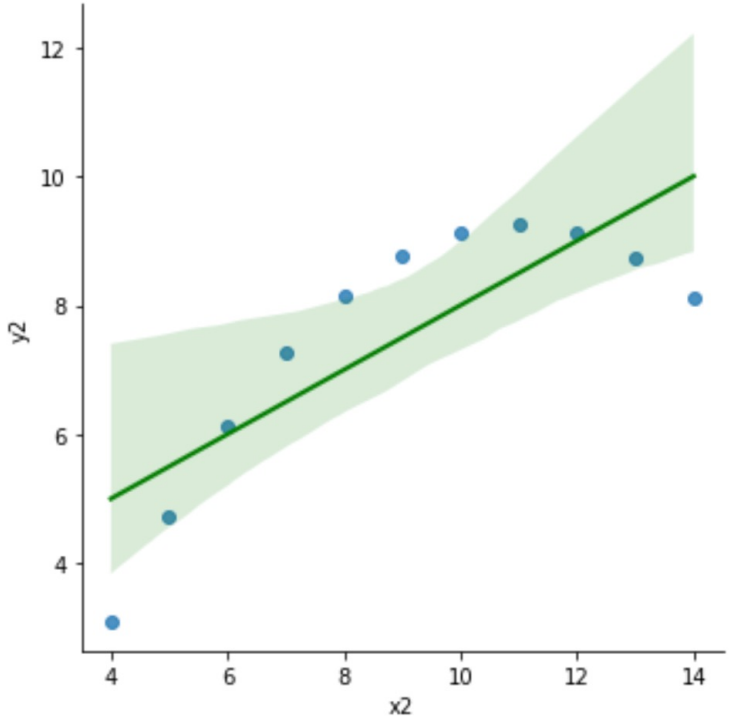


Descriptive Stats and Visualizations

Anscombe's Quartet



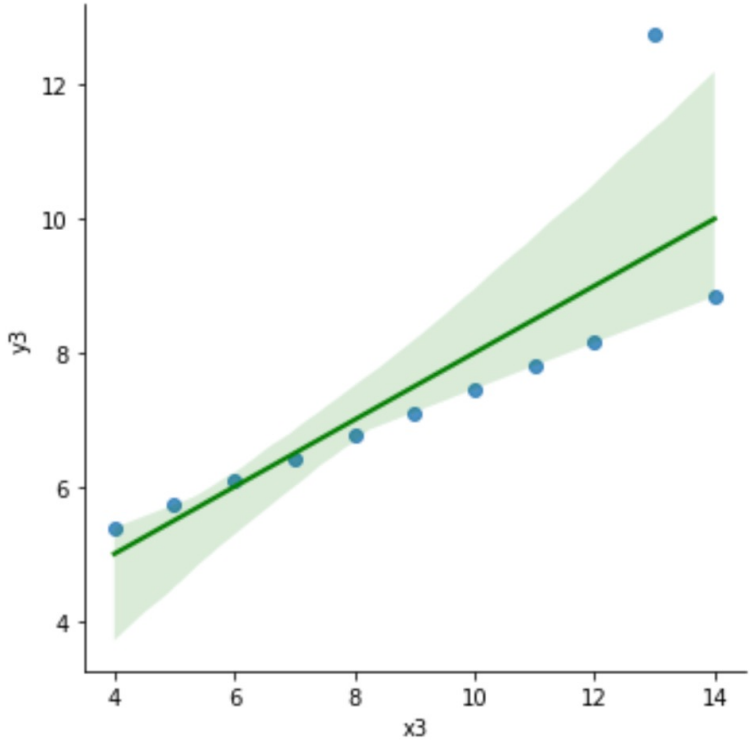
a linear relationship between x and y



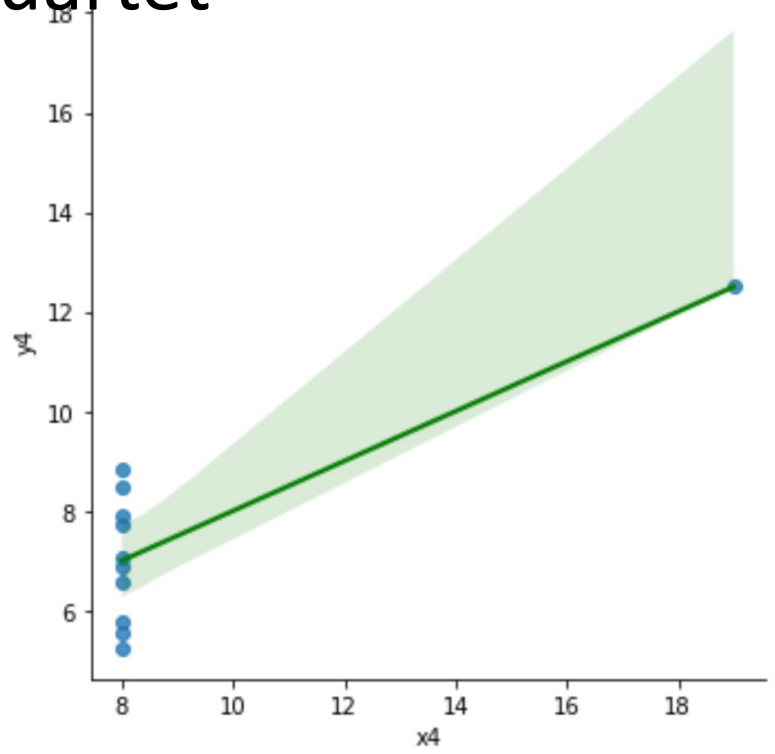
a non-linear relationship between x and y

Descriptive Stats and Visualizations

Anscombe's Quartet



Perfect linear relationship between x and y + Outlier



One high-leverage point is enough to produce a high correlation coefficient.

We observe….

Identical simple statistical properties yet appear very different when graphed

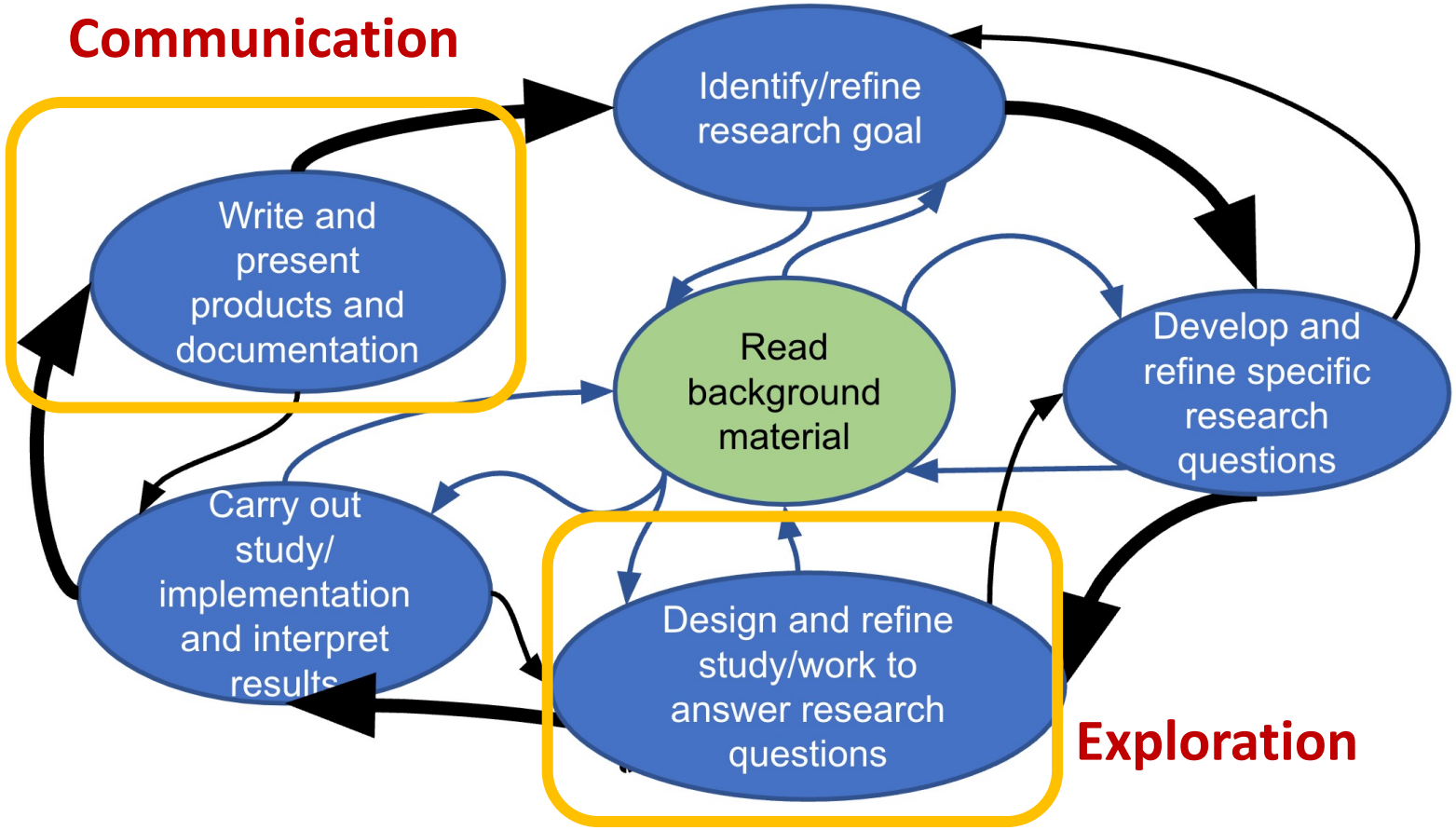
Descriptive Statistics are not enough!!

Why Visualize?

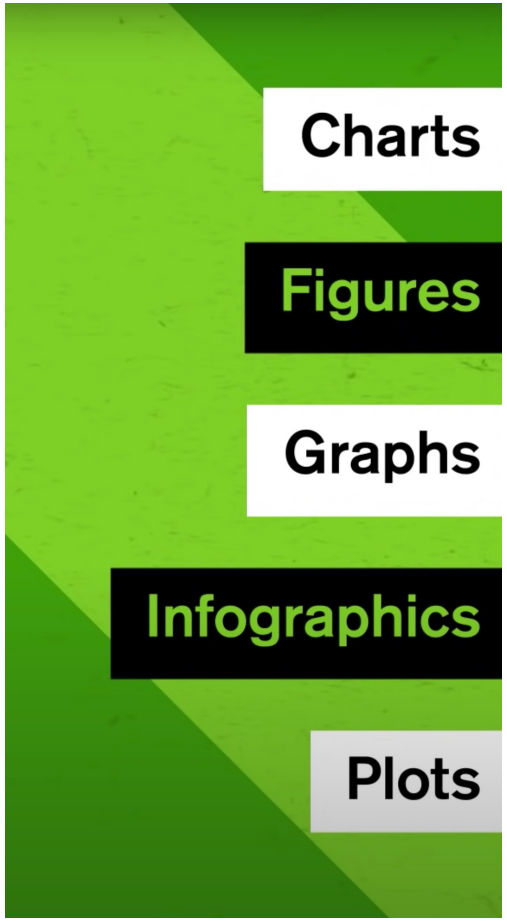
- Calculations can tell us a lot BUT visualizations can help us communicate or spot
  - Connections
  - Patterns
  - Trends
  - Outliers

- To Inform Humans: Communication
- When questions are not well defined: Exploration

Overview of the Research Process



Data visualizations go by many names…



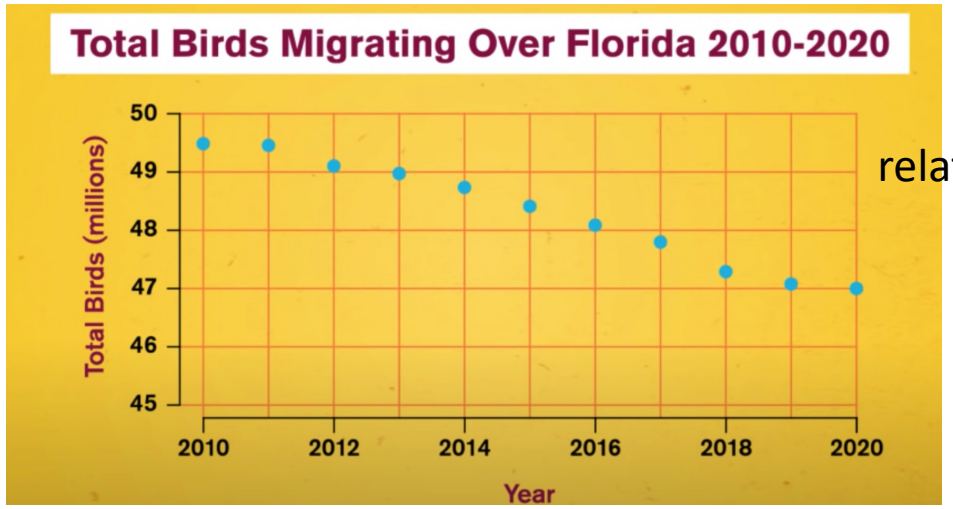
A WAY TO VISUALIZE A STORY THE DATA IS TELLING

Types of Visualizations

Used to emphasize different parts of the story

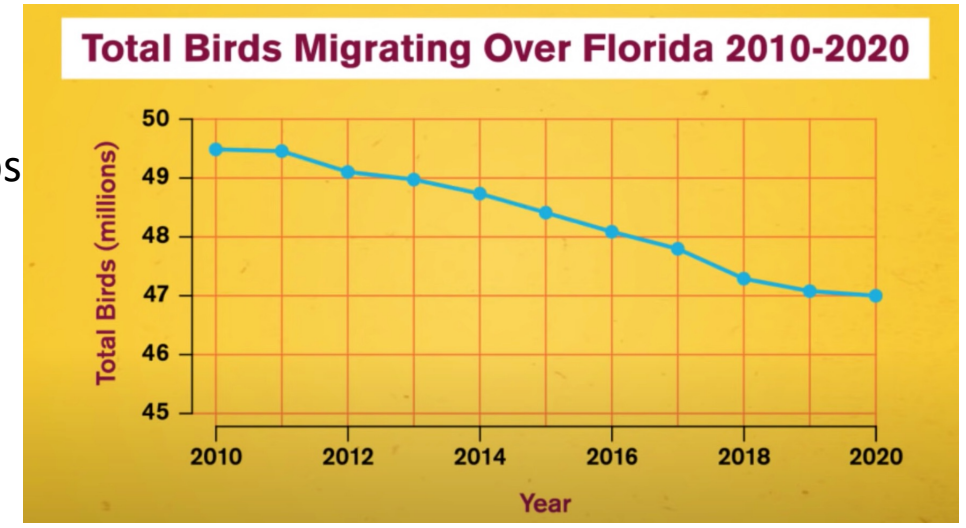
Line Plot

Track how things change over time



Scatter Plot

relationships

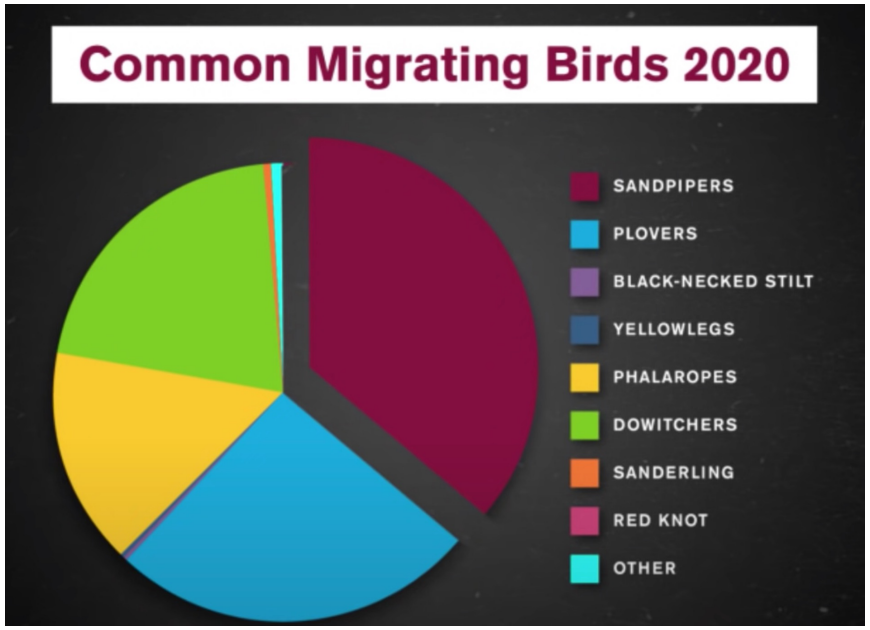


Line Plot

Jessica Pucci, Visualizing Data: Study Hall Data Literacy

Pie Chart

Stories about proportion



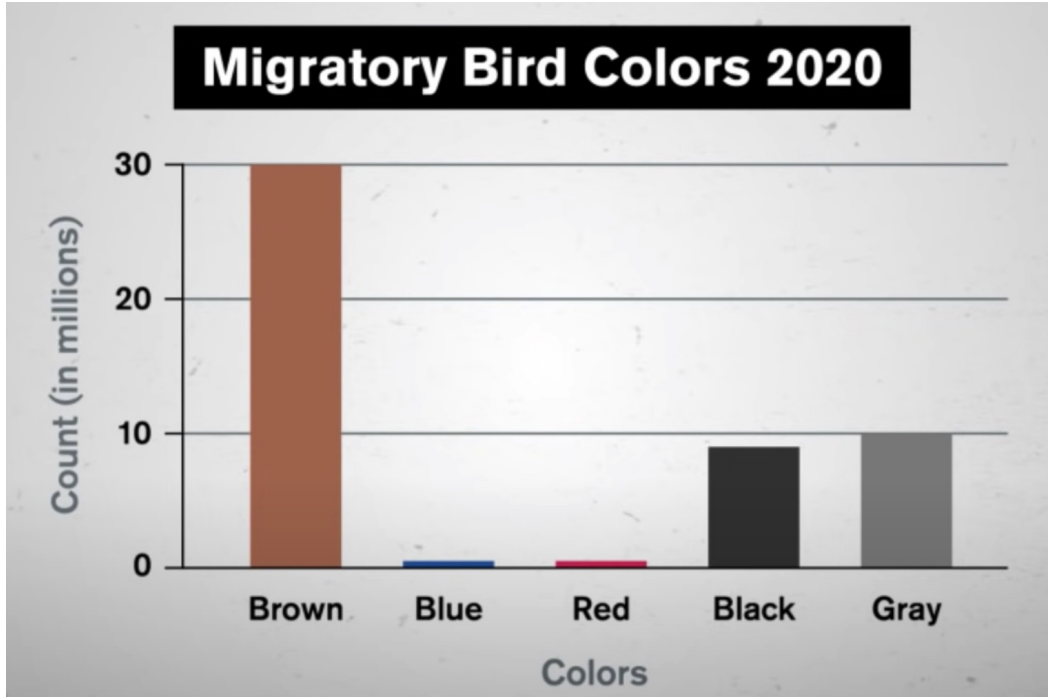
Jessica Pucci, Visualizing Data: Study Hall Data Literacy

A story about all the types of migrating birds.

Each slice shows how much each species contributes.

Bar Plot

Stories about
relationship
between a categoric
and numeric value

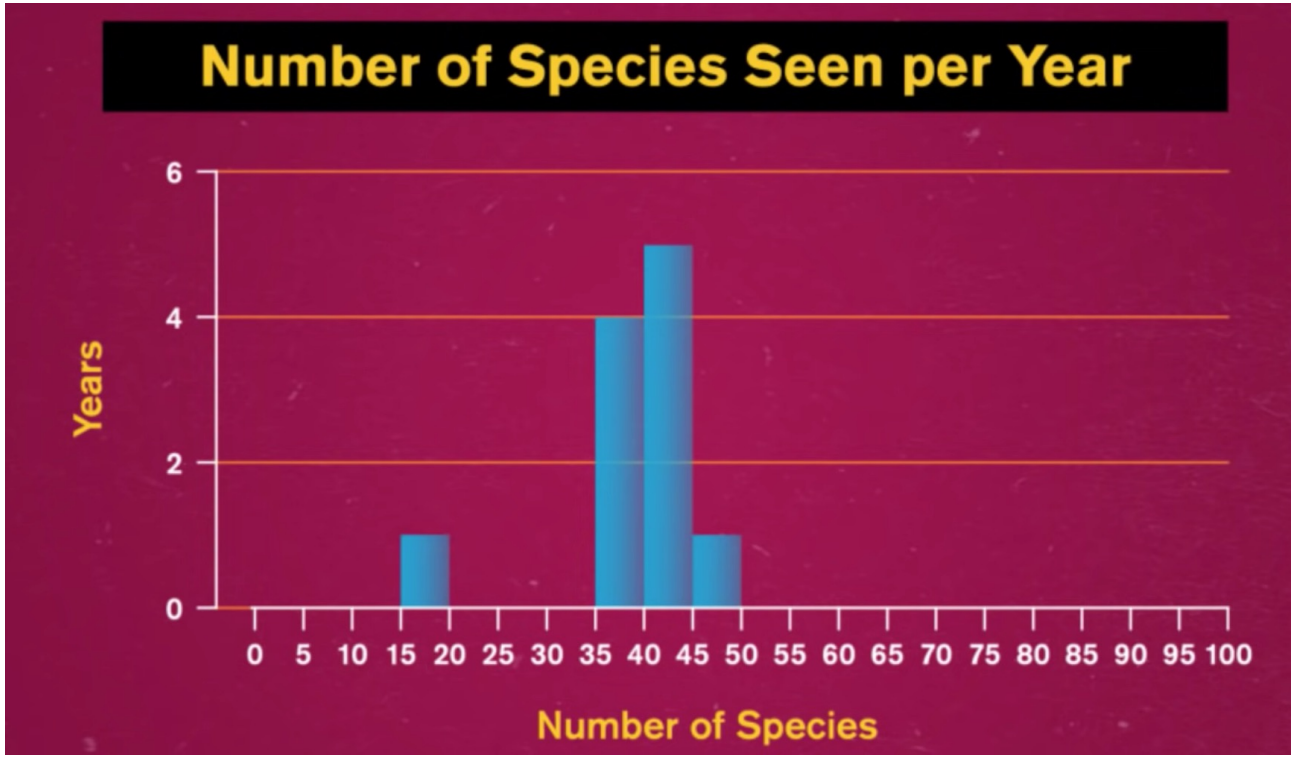


Marker for each data type

Jessica Pucci, Visualizing Data: Study Hall Data Literacy

Histogram

- Bins go across the horizontal axis
- Vertical axis tracks the frequency



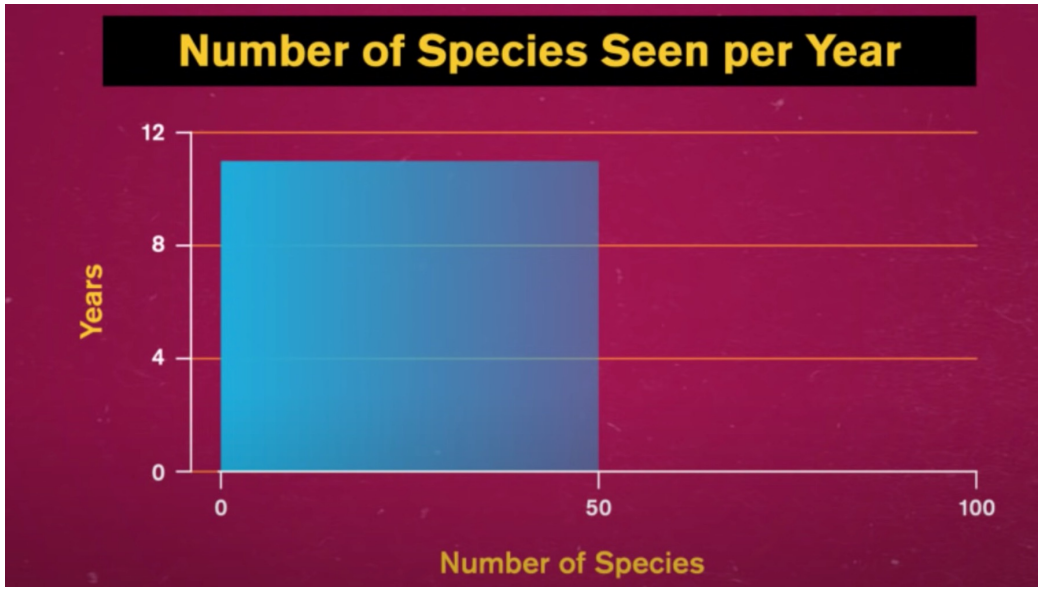
Some years, we see 35-40 or 40-45

- Pro-tip: How big we make our bins?

Jessica Pucci, Visualizing Data: Study Hall Data Literacy

Histogram

- Bins go across the horizontal axis
- Vertical axis tracks the frequency
- Pro-tip: How big we make our bins?



Bin size matters!

Jessica Pucci, Visualizing Data: Study Hall Data Literacy

Dataset: Fatal Police Shootings in the US

- The 2014 killing of Michael Brown in Ferguson, Missouri, began the protest movement culminating in Black Lives Matter and an increased focus on police accountability nationwide.

- Since Jan. 1, 2015, The Washington Post has been compiling a database of every fatal shooting in the US by a police officer in the line of duty.

https://github.com/washingtonpost/data-police-shootings

Dataset: Fatal Police Shootings in the US

The Washington Post is tracking more than a dozen details about each killing - including the race, age and gender of the deceased, whether the person was armed, and whether the victim was experiencing a mental-health issue.

https://github.com/washingtonpost/data-police-shootings

Worksheet

- Practice bar charts, histograms, and axes labeling

Resources

- **Top 6 Python Libraries for Visualization: Which one to Use?**
  - https://towardsdatascience.com/top-6-python-libraries-for-visualization-which-one-to-use-fe43381cd658

- Other Data Visualizations
  - https://rklopotek.blog.uksw.edu.pl/files/2017/09/data-visualization-2.1.pdf

Next Time

The process of peer reviews

- For another group's proposal