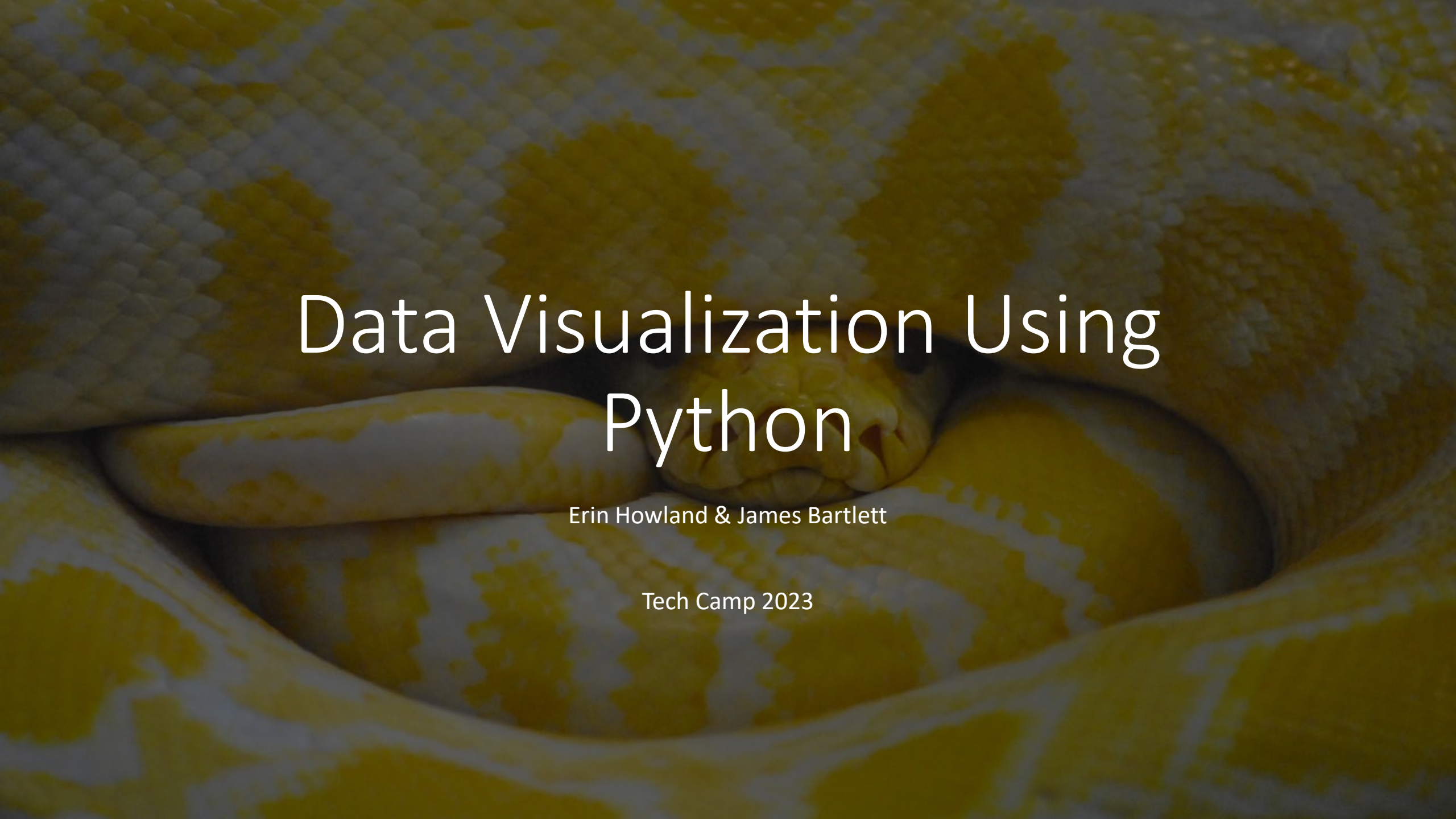


Data Visualization Using Python

Erin Howland & James Bartlett

Tech Camp 2023

Intro Questions

What brings you to Tech Camp?

What are you hoping to learn in this class?

What do you know, or think you know, already about Python and/or data visualization?

Data Jobs

What data-related jobs have you heard of?

What do you think data jobs have in common?

Data Jobs

To The Code!

Jupyter Lab Setup

- https://GitHub.com/LikeAWednesday/TechCamp_DataViz
- Follow instructions at the bottom of the page

TechCamp_DataViz Public

Watch 1

main · 2 branches · 0 tags · Go to file · Add file · Code

JamesDBartlett3 updated readme.md fd6cac5 7 hours ago 154 commits

data	removed unused data files	14 hours ago
lab	Update install.sh	10 hours ago
resources	renamed	14 hours ago
solutions	fixed "unsafe" type casting in regplot section	8 hours ago
.gitattributes	Initial commit	4 months ago
.gitignore	added temporary MS Office files to .gitignore	last week
cat_survey.ipynb	synchronized changes	12 hours ago
movie_ratings.ipynb	synchronized changes between exercise and solution	11 hours ago
readme.md	updated readme.md	7 hours ago
streaming_music.ipynb	formatting	11 hours ago

readme.md

Data Analysis & Visualization with Python

Setup

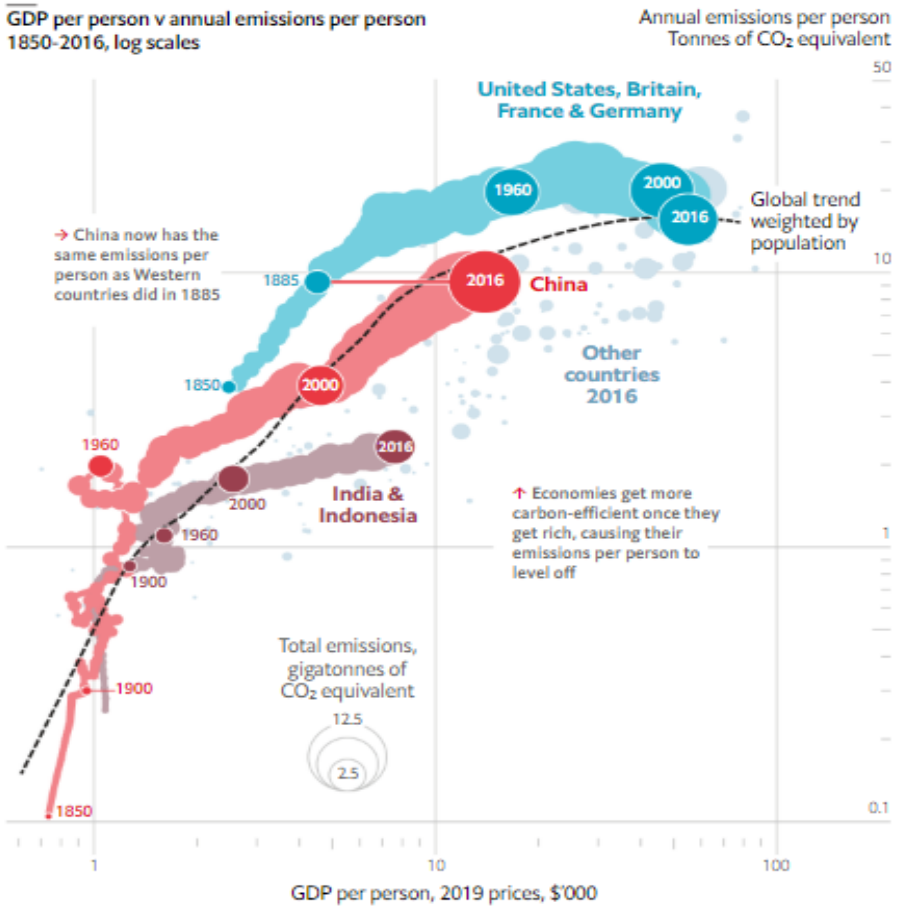
1. Install Linux on your Chromebook
2. Enter the following command in the Terminal to install Python, Jupyter Lab, and all libraries necessary for the included exercises on your Chromebook.
   Hint: Hover over the text below, click the "copy" button that appears to the right, then paste into your Terminal window by either pressing Ctrl+Shift+V or tapping the touchpad with two or three fingers, and then press Enter.

```
bash <(curl -s https://raw.githubusercontent.com/likeawednesday/TechCamp_DataViz/main/lab/install.sh)
```

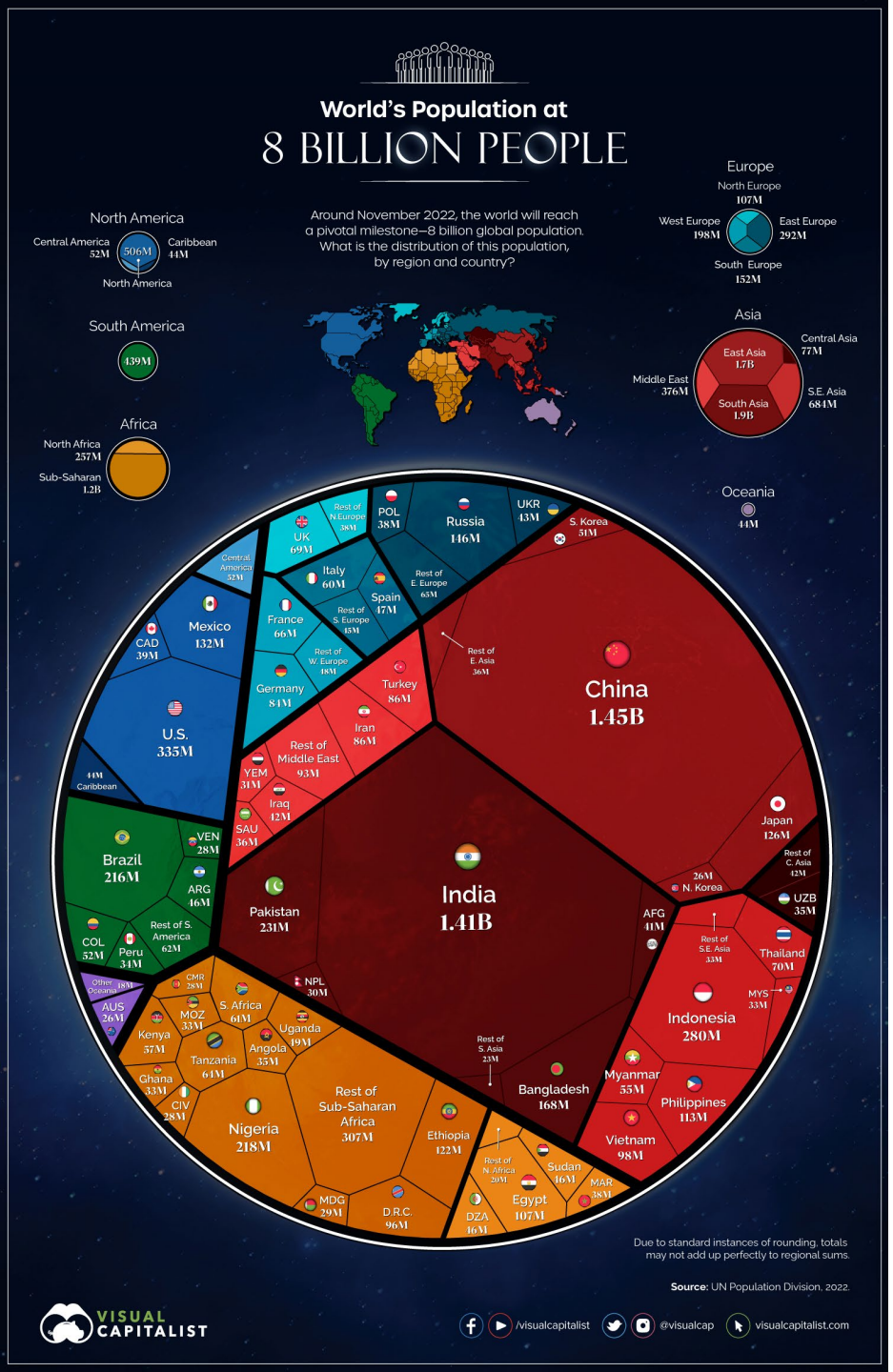
Visualizations

- What kinds of visuals can you think of?
- How do you know what visual to choose?
- What makes a visual good or bad?
- Visual Vocab

Good Visuals in the Wild, pt I



https://www.juiceanalytics.com/writing/20-best-examples-of-charts-and-graphs



Good Visuals in the Wild pt II

- Want some more examples of good visualizations?
  - Guidelines for Good Visualizations
  - MIT: How to Make Better Infographics

https://www.visualcapitalist.com/visualized-the-worlds-population-at-8-billion/

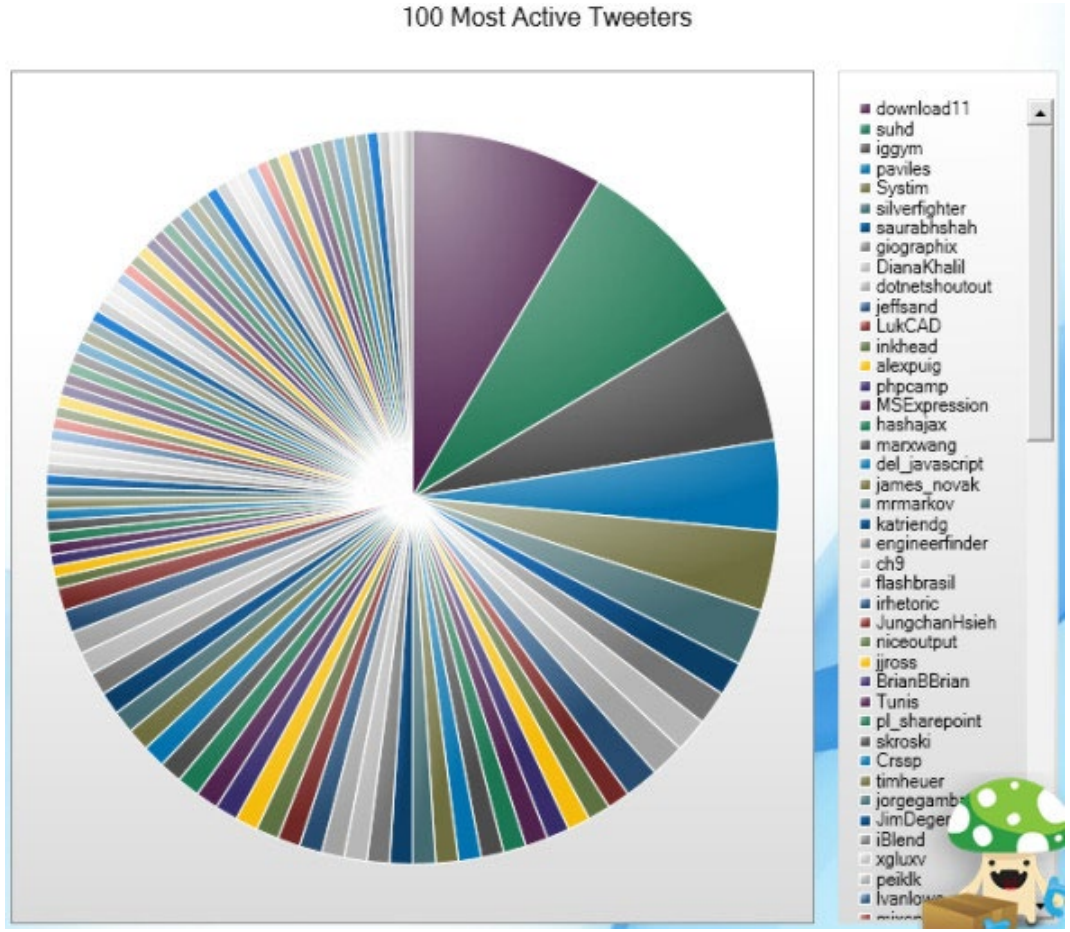
What Makes a Good Visual?

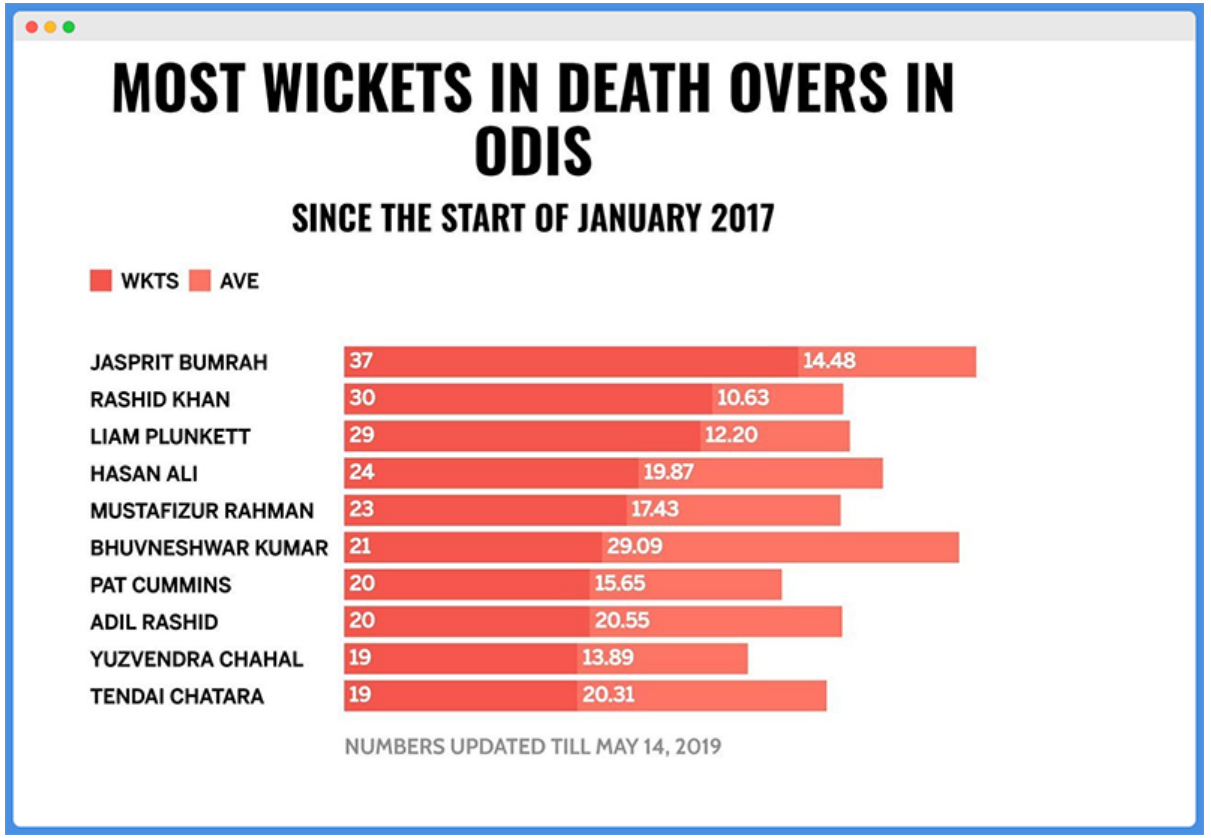
- Know what you are trying to communicate
  - This will inform chart type, whether ordering is important, if guidelines would be useful, etc.
- Understand your audience
  - Can inform labelling – technical vs. not technical, for example
- Think about use of blank space vs. visual clutter
- Use of color / accessibility
- Resilience to changes in data

Common Errors in Visualizations

- Wrong type of chart/graph
- Too many variables
- Faulty/inconsistent scale
  - Scale breaks (cropped axis) make small differences seem bigger
  - Unclear linear vs. logarithmic scale distorts understanding/perception of data
  - Inconsistent time intervals
- Poor color choices
- Poor labeling

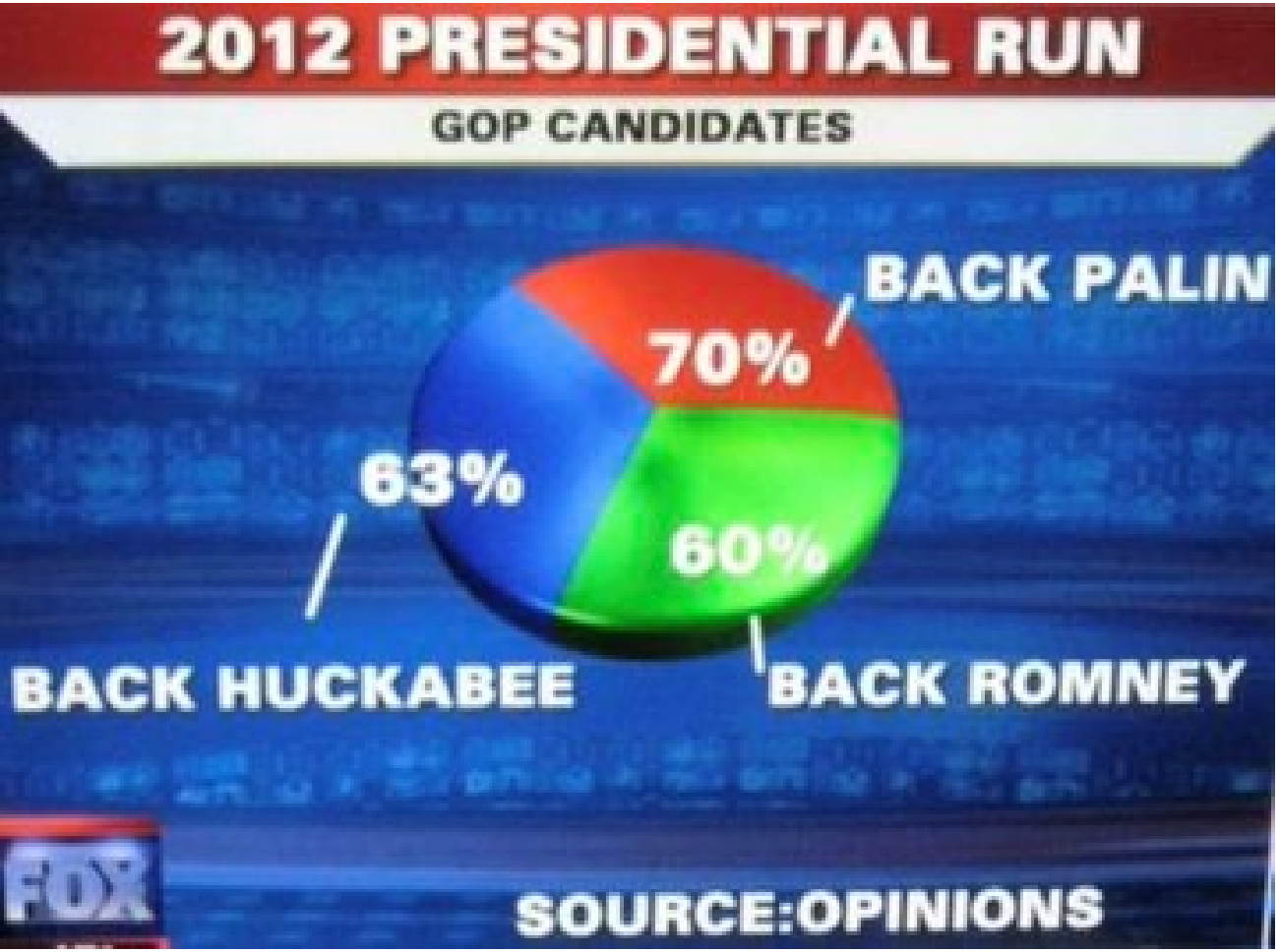
Bad Visuals in the Wild, pt I





https://chandoo.org/wp/nightmarish-pie-charts/
https://rigorousthemes.com/blog/bad-data-visualization-examples/

Bad Visuals in the Wild, pt II



- Numbers don't lie (unless you're as careless as whoever put together this chart), but the methods by which the numbers are gathered and how those numbers are presented can be deceptive.
- Want some more bad charts?
  - Business Insider - 27 Worst Charts
  - Bad Infographics

https://www.rawstory.com/fox-news-misleading-graphs/

Python Libraries

- What is a library?
  - A collection of books, journals, and other resources are stored for use.
- What is a Python library?
  - A collection of code – think of each Python library like a physical library from which you can check out functions rather than books.
  - Allows us to code more efficiently in Python by not having to rewrite common code blocks that someone else has already written.
  - Using a Python library is totally OK! It is not considered plagiarism, and is, in fact, encouraged.

What Kinds of Libraries are There?

Just as book libraries tend to focus on specific audiences (general, children, lawyers, etc.) Python libraries tend to specialize in different kinds of tasks:

- Data manipulation: Pandas, datatable, PandaPy, NumPy
- Web scraping: BeautifulSoup, Scrapy
- Visualization: Matplotlib, Seaborn, Plotly, GGplot
- AI/ML/complex analysis: Scikit-Learn, NumPy, TensorFlow, PyTorch

https://www.geeksforgeeks.org/libraries-in-python/ https://www.geeksforgeeks.org/top-8-python-libraries-for-data-visualization/

How to use Python to Make Visuals?

- Our focus is making visuals, but we'll use some libraries that do other things. This is because we will also need to be able to view and manipulate our data in order to create visuals.
- Libraries we'll import and use today:
  - Pandas
  - NumPy
  - Matplotlib
  - Seaborn
  - Squarify

To The Code!

File: cat_survey.ipynb

Dataset: data/cats.csv

Let's Make Some Visuals!

- We'll look at a sample data set with the following goals:
  - Learn how to open a flat file in Python and load to a dataframe
  - Explore data
  - Consider what story you can tell
  - Determine what visual(s) might be appropriate to communicate that story
  - Make some visuals
    - Different kinds of visuals
    - Labels and legends
    - Change colors

Reference:
Loading Data & ‘What Now?’

- Pandas can read numerous file types
  - In all instances you will need a file path, so the program knows where to look
  - We will read files into a dataframe (df), which is basically Pandas’ version of a spreadsheet, but way cooler
- I’ve loaded the data. Now what?
  - View dataframe
  - What data types do you have?
  - Is any information missing?
  - In short, *think about the data!*

Clean vs. Messy Data

Clean Data

- No empty/missing values
- Consistent data type in each column (text, integer, date, etc.)
- Column headers are names, not values
- Data is ready to use!

Messy Data

- Messy data can include:
  - Empty/missing values
  - Mixed data types in each column
  - Multiples variables stored in one column
- Data requires cleaning
  - This could be a course in and of itself!

To The Code!

File: movie_ratings.ipynb

Dataset: data/imdb.csv

Focus: Ratings

Let's Make Some More Visuals!

- How would you describe this dataset? Messy? Clean? Somewhere in the middle?
- Three primary steps:
  1. Load dataframe
  2. Summary/descriptive info
  3. Visualizations
- Think about these questions at each primary step:
  - Can you see any potential problems?
  - Why are these problems / What is the risk if these aren't ultimately addressed (i.e., how might the story change)?

Making Visuals With Less Than Perfect Data

- As you may have noticed, the IMDB data was pretty good, but it wasn't perfect – there were some missing values.
- Depending on what you're investigating, missing values may or may not be problematic. In the last example, because we looked at ratings, our visuals were not impacted since the missing data was not from the ratings column. *But what if we were missing data from an area we wanted to investigate?*

Handling Missing Data

- If you are curious and want to play with this, here are two ways to address missing data:
  - Impute (fill in) missing data using `fillna()`
    - Impute with mean:
      - `df.fillna(mean(), inplace = True)`
      - `df['column'].fillna(df['column'].mean(), inplace = True`
    - Impute with median:
      - `df.fillna(median(), inplace = True)`
    - Impute with mode:
      - `df.fillna(mode(), inplace = True)`
  - Drop rows with missing data using `dropna()`
- Any handling of missing data can influence data analysis, so choose wisely!

To The Code!

File: movie_ratings.ipynb

Dataset: data/imdb.csv

How would you handle missing data from the Certificate and/or Metascore column(s)?

To The Code!

- Visual-Ready
  - data/cats.csv
  - data/imfgdp.csv
- Challenge
  - data/imdb.csv
  - data/music.csv

Visuals: Choose Your Own Adventure!

- If you want to work with clean, perfect data and focus only on visuals and customization of visuals, see Visual-Ready List
- If you want a little more of a challenge with data quality and Python coding beyond just what's needed for visuals, see Challenge List
- Keep in mind: you can use data from either set! Just be aware of your goals and what question(s) you are looking to answer with your visuals. There are no macho points ☺