https://github.com/carolinadatascience

Approaching a Data Science Project

CDC Edition

Attendance:

https://forms.gle/iPYyxizZBFET4zh
h7

https://github.com/carolinadatascience

The **Carolina Data Challenge** (CDC) is an annual data science event held at the University of North Carolina at Chapel Hill. Each year, we have the great privilege of welcoming hundreds of curious students from around the globe to our event.

This year our data challenge will start on **September 30th - October 1st** to give participants the opportunity to create unique projects they would be proud to add to their portfolio.

The theme for the 2023 Carolina Data Challenge is ***Nostalgia***! Students will work in teams to create data science projects using datasets surrounding this theme. Teams will get the change to work on projects focusing on one of five tracks: Business, Health Science, Social Science, Natural Science, and Pop Culture! Each track will have their own hand-picked dataset to explore, all with varying difficulties, unique challenges, and interesting insights.

REGISTER NOW: https://cdc.cs.unc.edu/

https://github.com/carolinadatascience

CDC Project Format

Theme: Nostalgia

Form a Team

1 - 4 people

Pick 1 track

5 options, each w/ a dataset

Project w/ dataset

Culminates in a presentation

https://github.com/carolinadatascience

CDC 2022 Example

Theme: Nature's Fury

Track	Business	Natural Sciences	Social Sciences	Health Sciences	Pop Culture
Dataset	Greenhouse Gas Reporting	Greenhouse Gas Air Emissions	San Francisco Health Vulnerability	Storm Events & Injuries	Climate Related Tweets

https://github.com/carolinadatascience

CDC 2022 Example

Theme: Nature's Fury

Track	Business	Natural Sciences	Social Sciences	Health Sciences	Pop Culture
Dataset	Greenhouse Gas Reporting	Greenhouse Gas Air Emissions	San Francisco Health Vulnerability	Storm Events & Injuries	Climate Related Tweets

We'll focus on this dataset as we go through our next steps

Forming a Problem Statement

Approach

- Start with a problem that you want to address
- Would answering this question create a meaningful impact?
  - Who would it impact?
  - How can this impact be measured?
- What does a solution to this problem look like?
- Do we need additional data to solve this problem?
- Write out a clear problem statement to guide your project

https://github.com/carolinadatascience

Forming a Problem Statement

Our Example

- **Question:** How energy efficient is transportation in different regions around the world?
- **Impact:** The International Council on Clean Transportation could use this information to decide where to focus their efforts.
- **Solution:** an index that grades each region on transportation efficiency.
- **Additional Data:** we may need information on the population of each country
- **Problem Statement:** We want to create a numerical index that quantifies the efficiency of transportation by region so the ICCT can decide where to focus their effort.

https://github.com/carolinadatascience

Cleaning/Wrangling Data

Remove Duplicates

Deal with Missing Data

Fix Structural Errors in Data

Handle Outliers

and more...

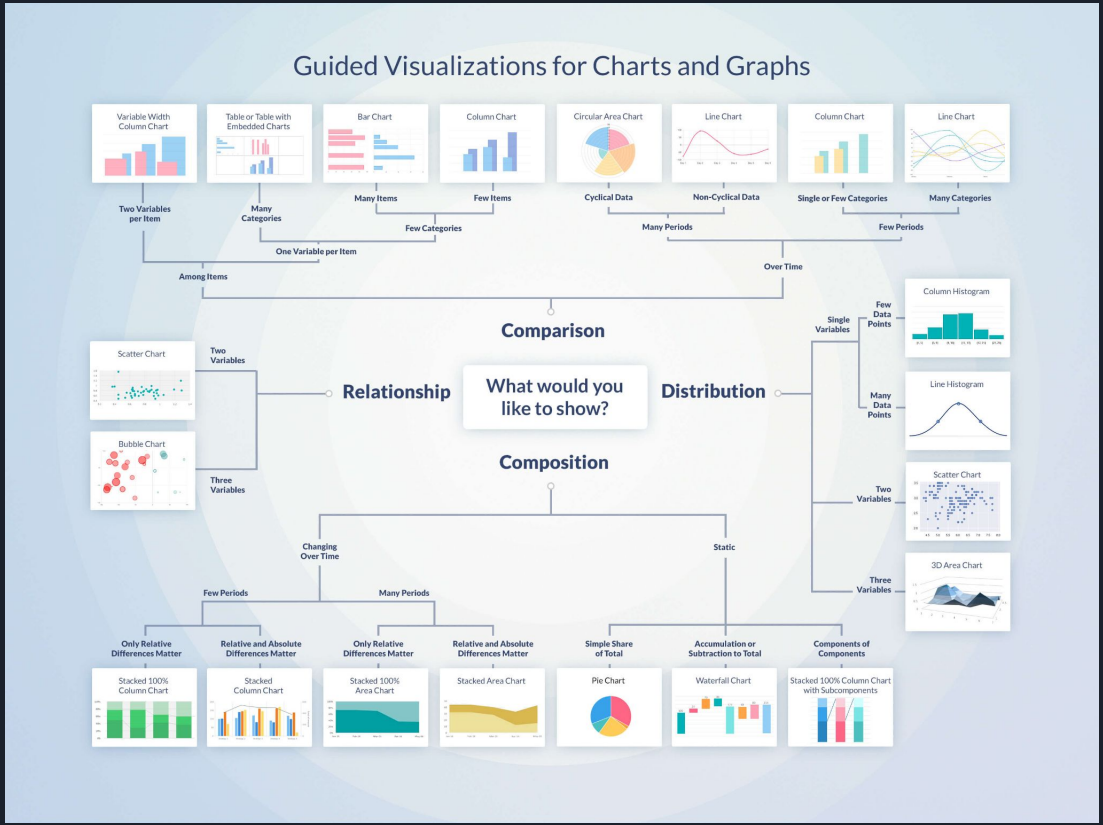
https://github.com/carolinadatascience

Analyzing Data

Descriptive Analysis	Diagnostic Analysis	Predictive Analysis	Prescriptive Analysis
Identify what has already happened. Ex. Which country has had the most CO_2 emissions in 2021?	Understand why something has happened. Ex. Which industry is making X region emit so many greenhouse gasses?	Identify future trends based on past data. Ex. How many units of methane will Africa produce in 2024?	Make recommendations for the future. Ex. What strategies could Africa take to reduce their methane production in 2024?

https://github.com/carolinadatascience

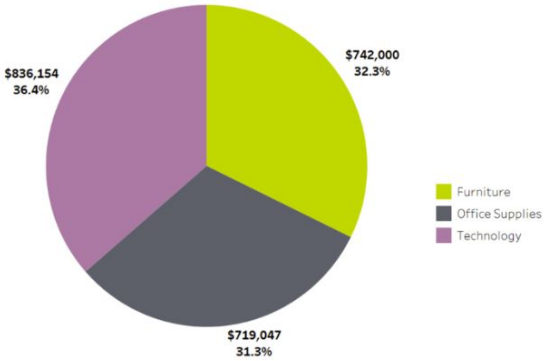
Creating Supporting Visualizations



https://github.com/carolinadatascience

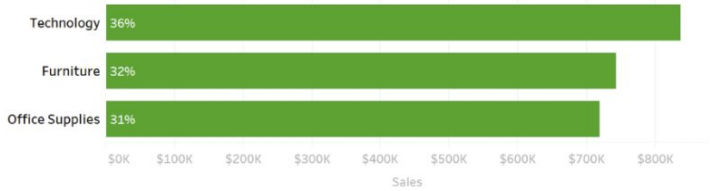
Effective Visualization Example

sales by product category



ineffective

sales by product category

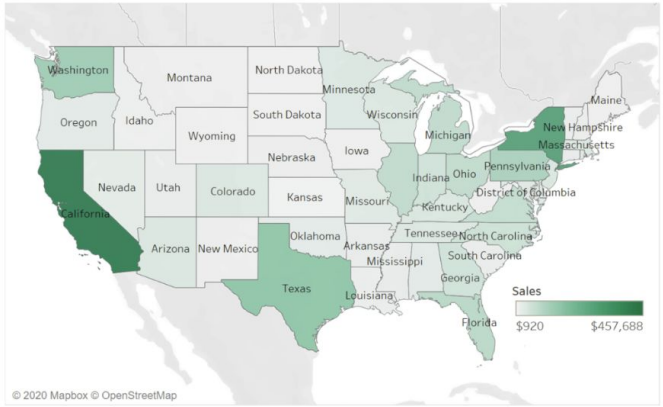


effective

https://github.com/carolinadatascience

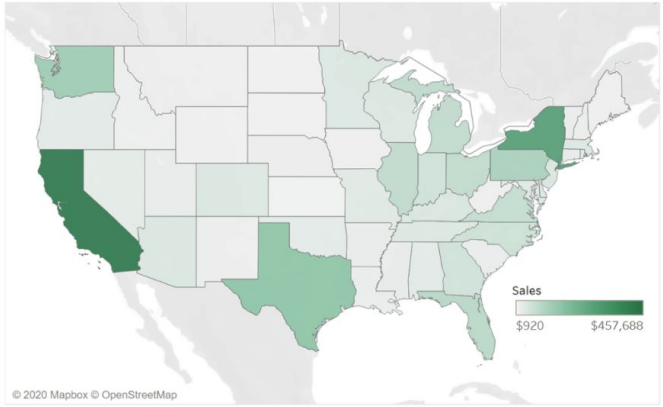
Effective Visualization Example

total sales map



ineffective

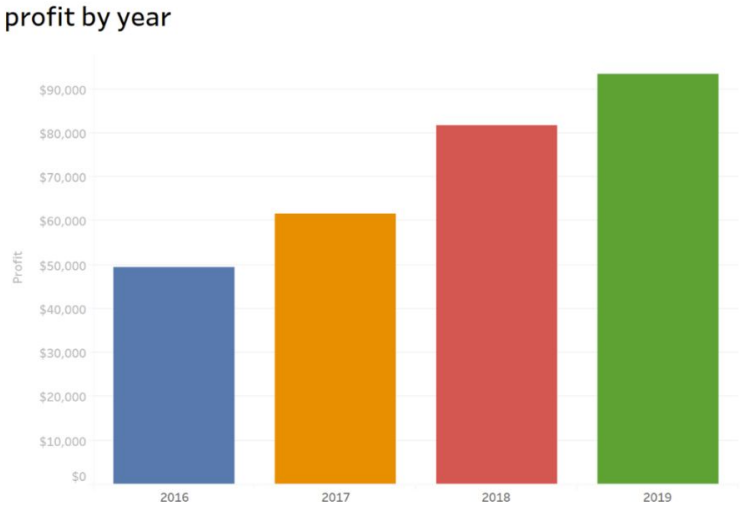
total sales map



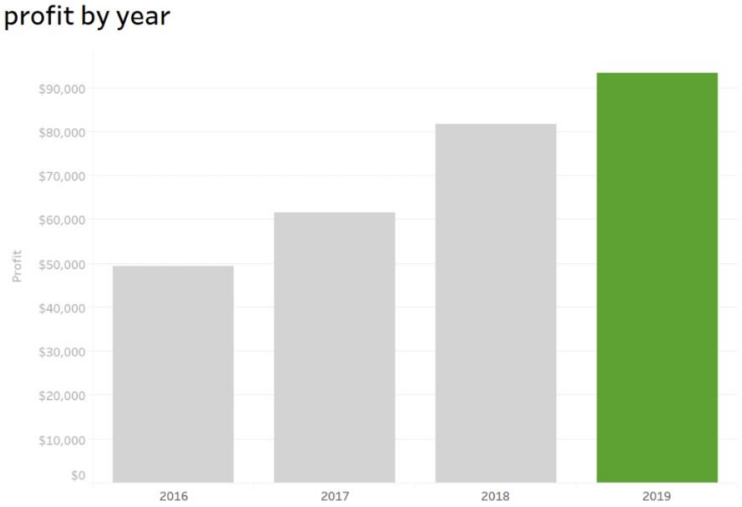
effective

https://github.com/carolinadatascience

Effective Visualization Example



ineffective



effective

Presenting to Stakeholders

- Keep it concise
- Focus on the impact
- Support your findings with visualizations
- Include limitations
- Explain your technical decisions

https://github.com/carolinadatascience

Past Projects from CDC 2022

No matter your past experience, you'll do great! Let's look at one of last year's winners that was a group of first time hackers.

Project Link

https://github.com/carolinadatascience

Stay Connected!

Join our Discord

Join as a member on HeelLife

Sign up on our Listserv

Follow us on LinkedIn

Follow us on Instagram

https://github.com/carolinadatascience

Thank you!