Introduction to Data Visualization

October 28, 2020

How are you all feeling?

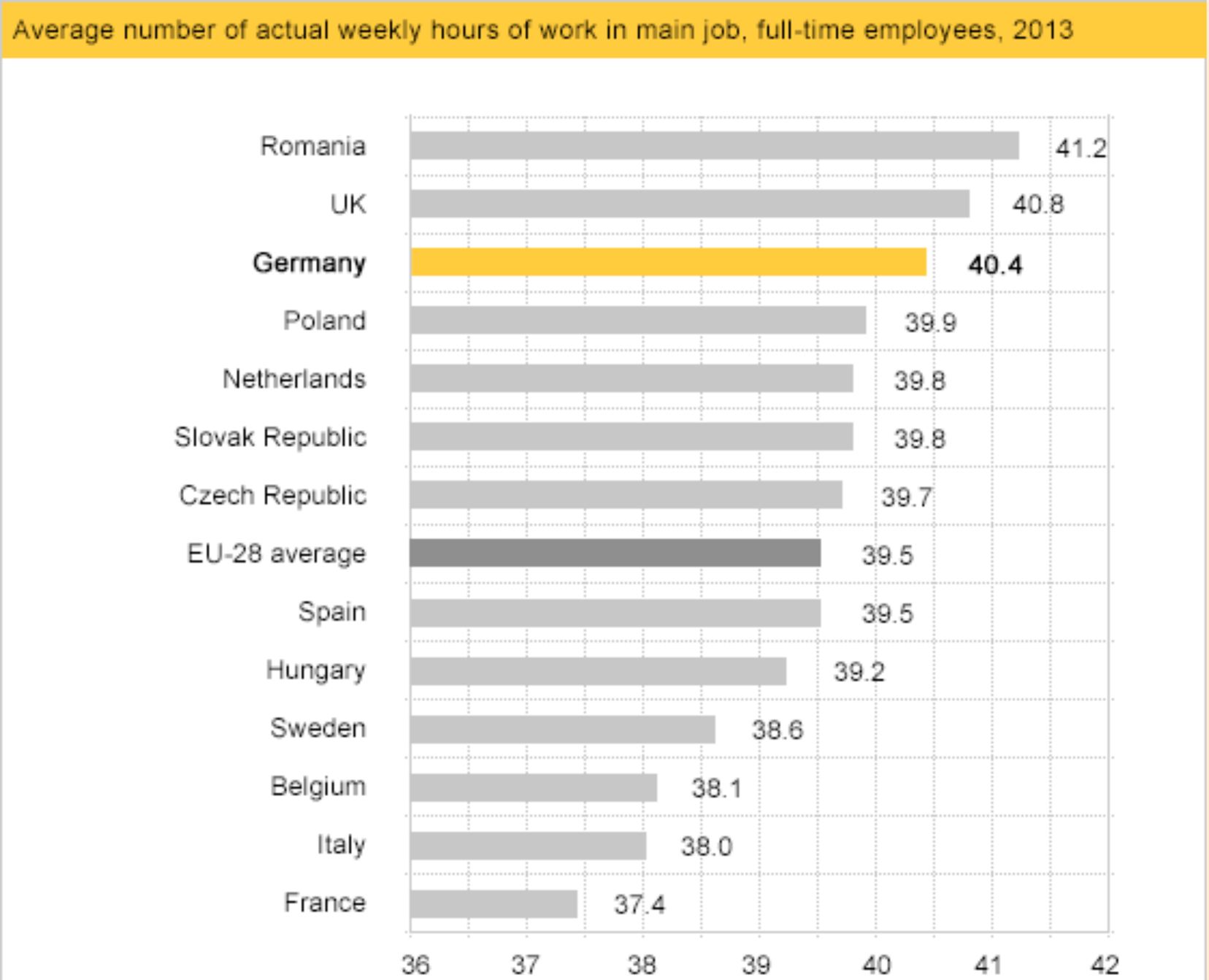
10 minutes

Announcements

- Everyone should have their Project Topics approved by now!
- Lab 4 was released earlier this week
- Milestone 2 will be released on Monday, it will be “similar” to Lab 4 in terms of what you will be expected to do, but slightly “deeper”.
- Couple of git things:
    - `git pull`
    - .gitignore

Part 1: Importance of data visualization

Credit to: callingbull.org



Weekly hours for full-time employees

Claim: "German workers are more motivated and work more hours than workers in other EU nations."

Q1: How confident are you in the claim based on the visualization?

A. Very confident

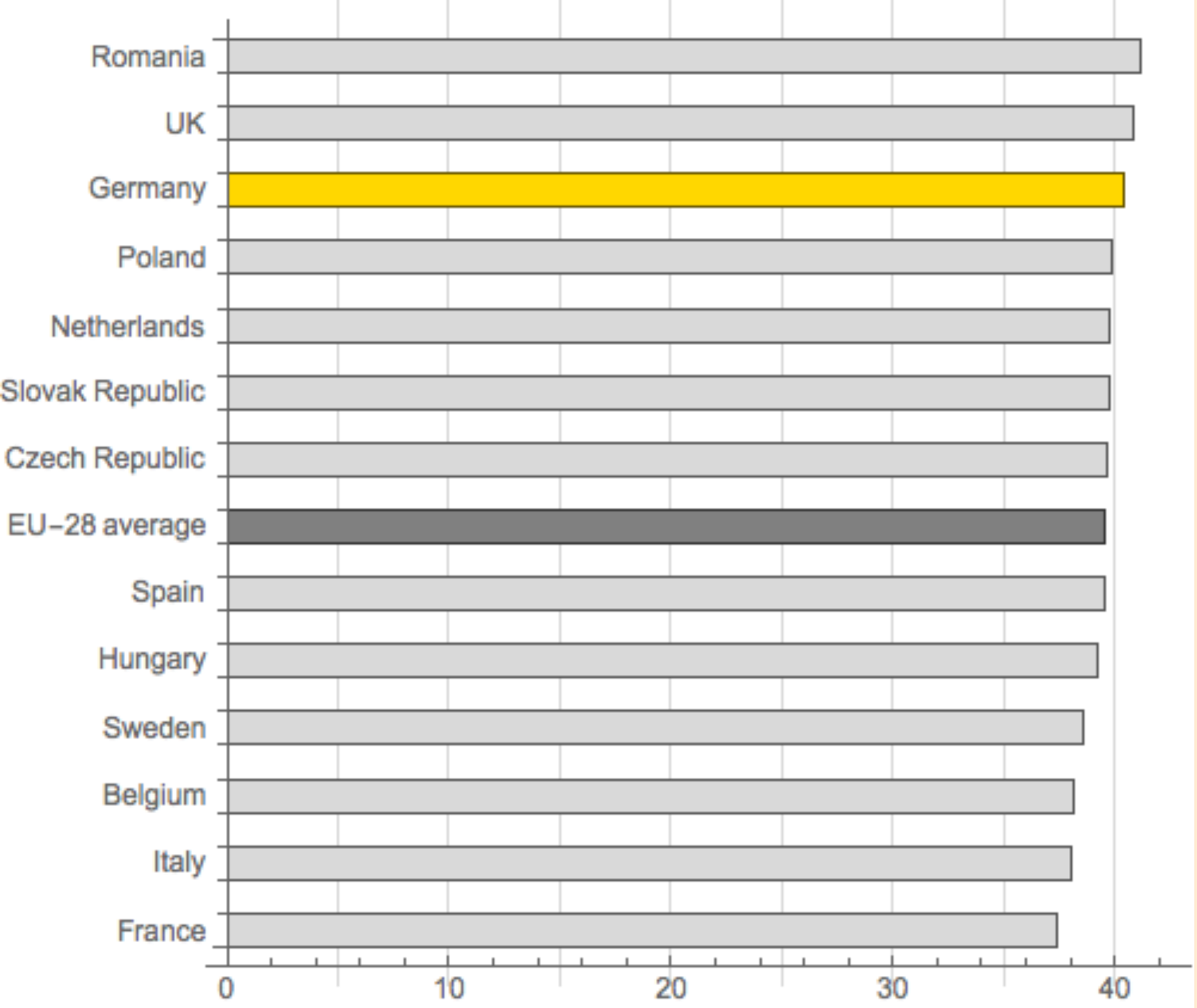
B. Confident

C. Not Confident

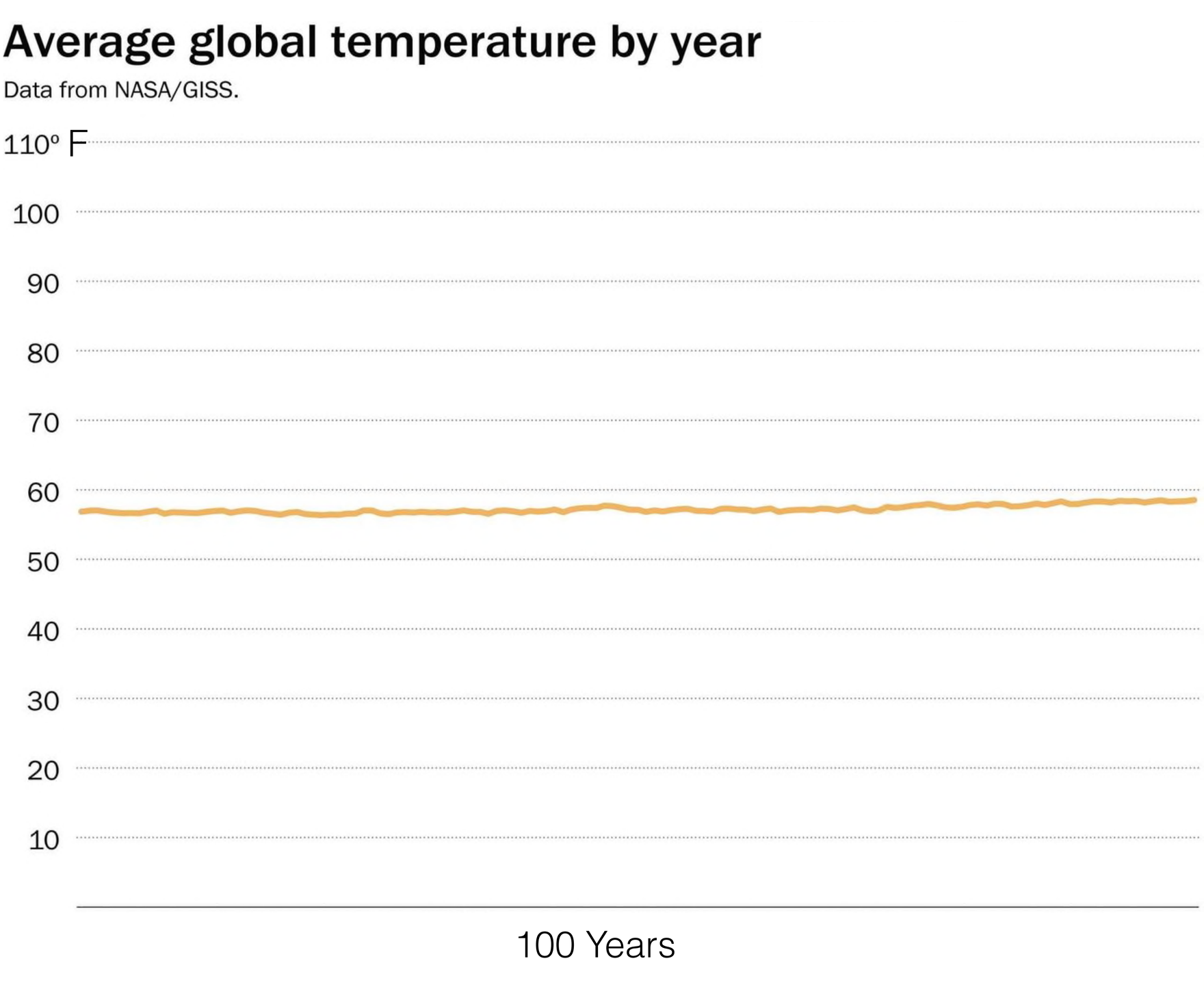
D. Claim is wrong

Data Source: Eurofound 2014
Example Source: Callingbull.org

Weekly hours for full-time employees



Example Source: Callingbull.org



Average Global Temperatures by Year

Claim: “Over **these 100 years**, there is a negligible change in global temperature.”

Q2: How confident are you in the claim based on the visualization?

A. Very confident

B. Confident

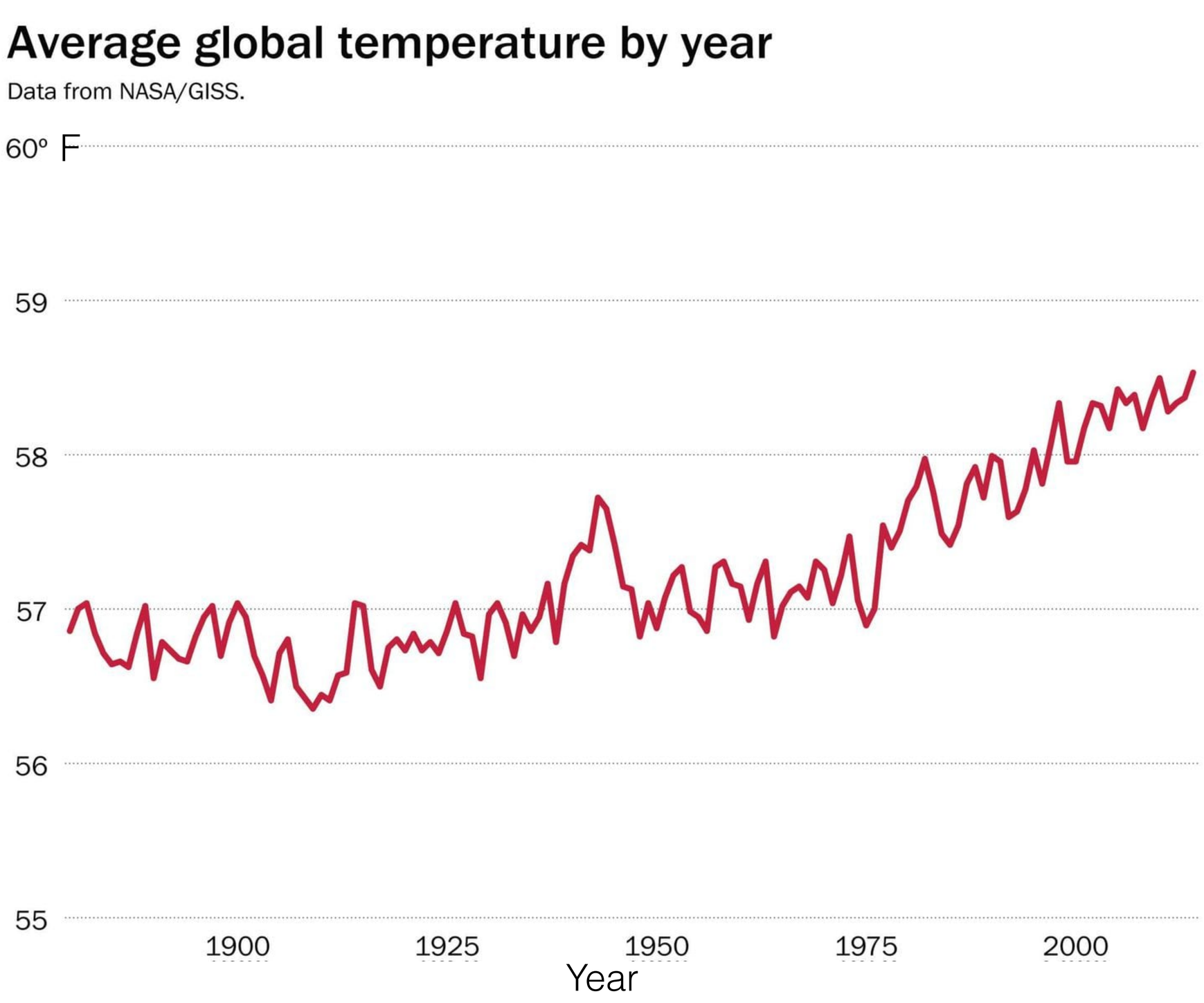
C. Not Confident

D. Claim is wrong

Data Source

Example Source

Average Global Temperatures by Year



Data Source

Example Source

nature
International journal of science

Brief Communication | Published: 29 September 2004

Athletics

Momentous sprint at the 2156 Olympics?

Andrew J. Tatem ✉, Carlos A. Guerra, Peter M. Atkinson & Simon I. Hay

Nature **431**, 525 (2004) | Download Citation ↓

1743 Accesses | **46** Citations | **78** Altmetric | Metrics »

Women sprinters are closing the gap on men and may one day overtake them.

Abstract

The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.

Gender parity in the Olympics (100m race)

Claim: "Women sprinters are closing the gap on men and may one day overtake them."

Q3: How confident are you in the claim based on the visualization?

A. Very confident

B. Confident

C. Not Confident

D. Claim is wrong

Gender parity in the Olympics (100m race)

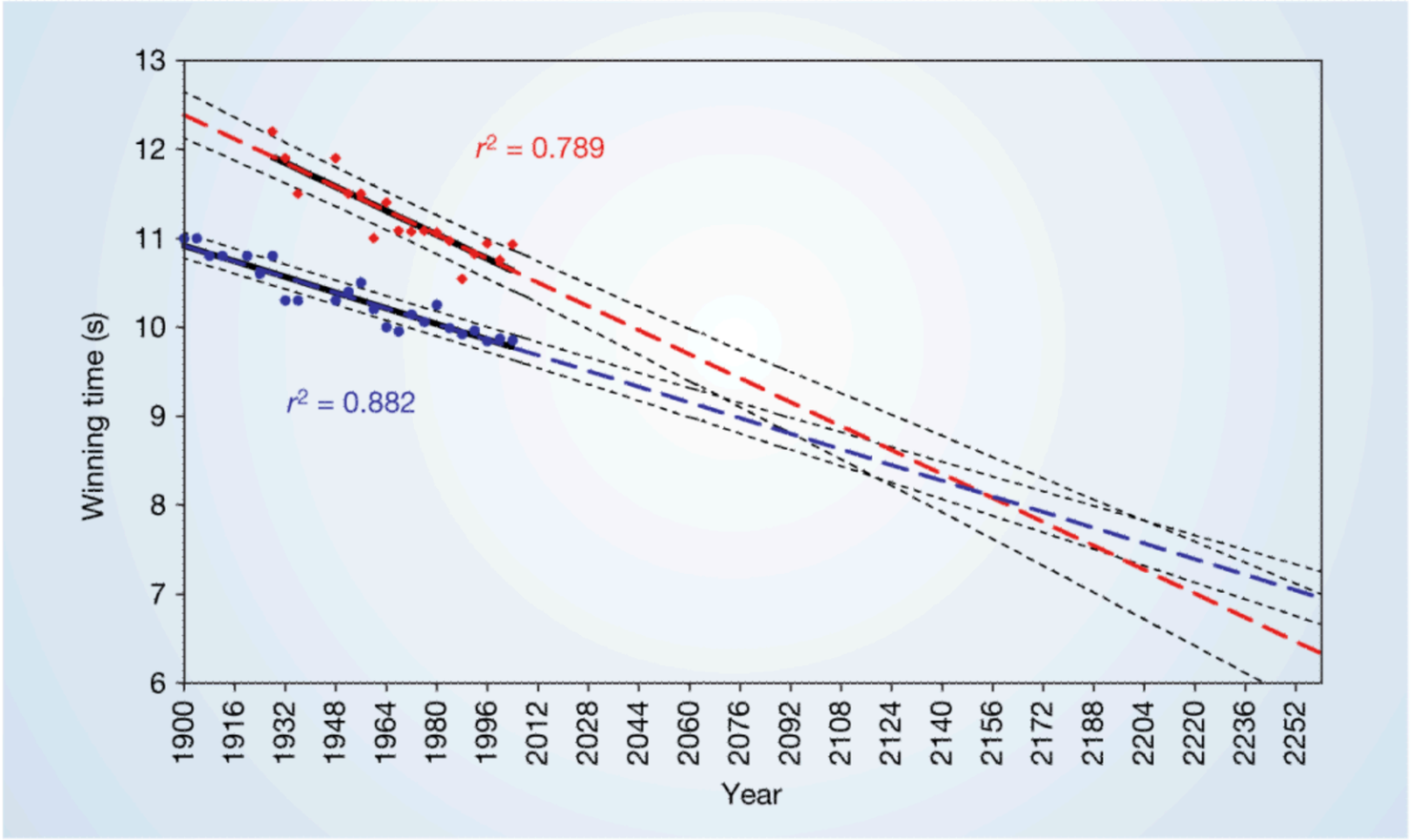
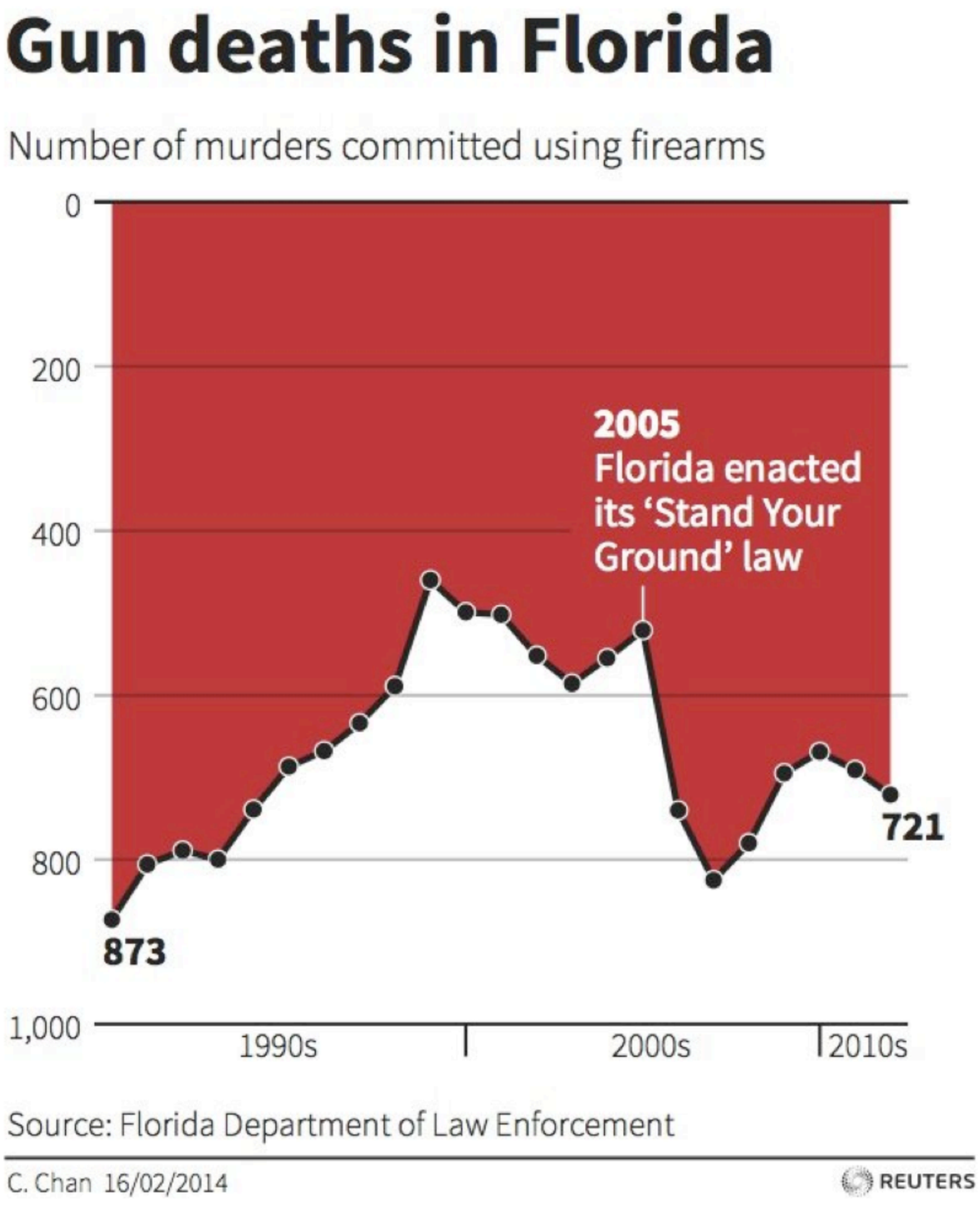


Figure 1.
The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.



Gun deaths in Florida after legislation

Claim: "After enacting new gun legislation in Florida, gun deaths sharply declined."

Q4: How confident are you in the claim based on the visualization?

A. Very confident

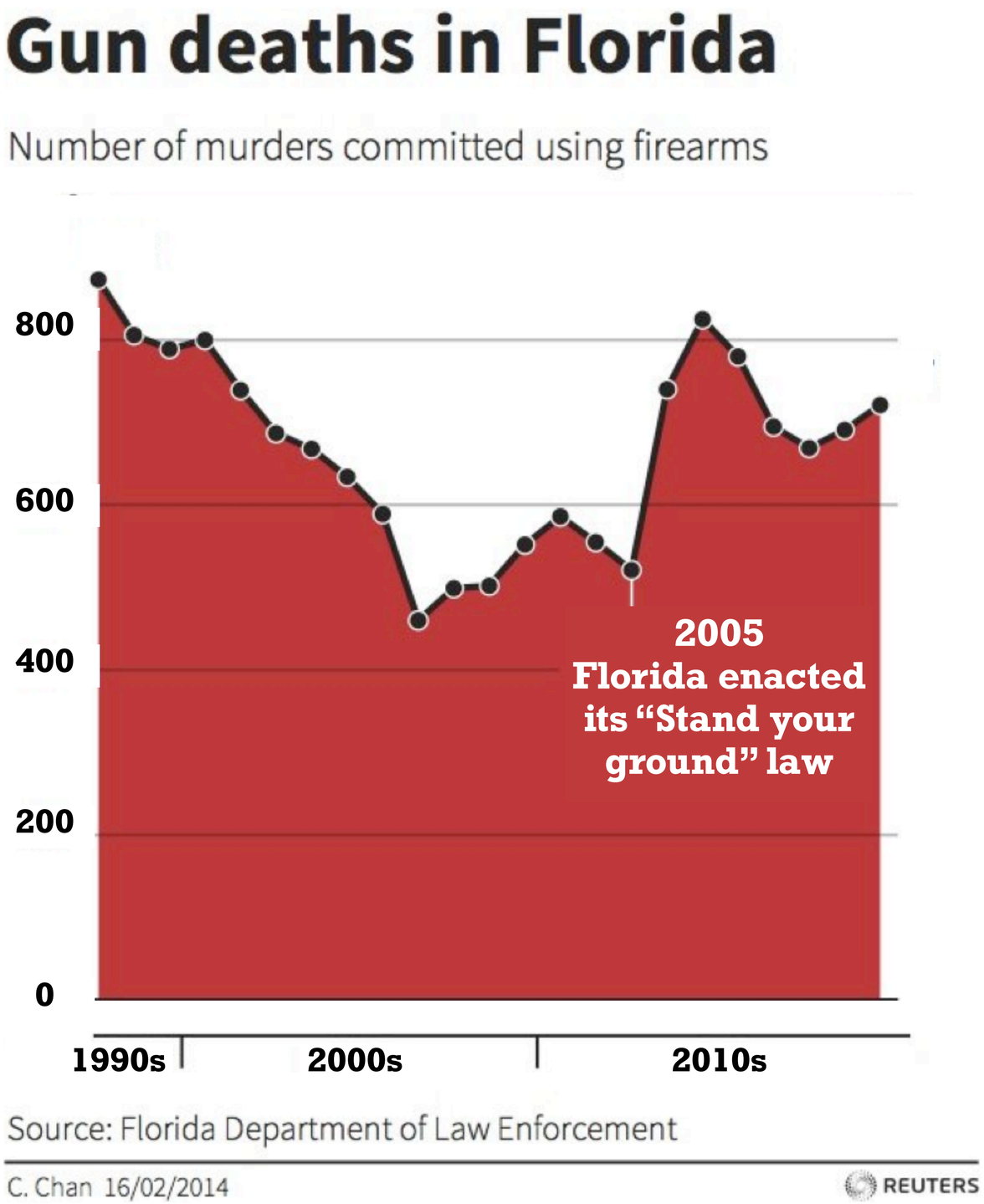
B. Confident

C. Not Confident

D. Claim is wrong

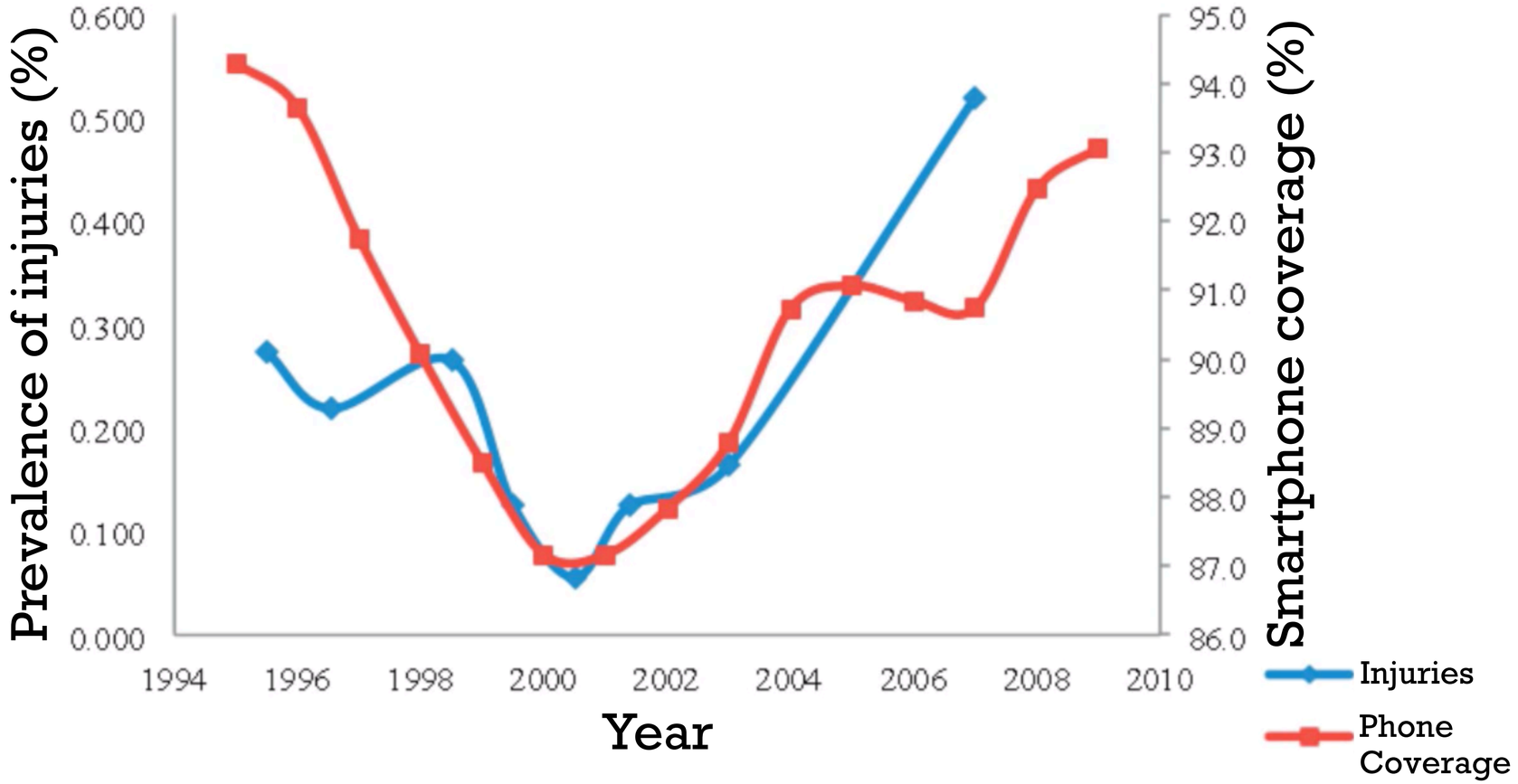
Data Source: Florida Department of Law Enforcement
Example Source: Callingbull.org & Reuters

Gun deaths in Florida after legislation



Data Source: Florida Department of Law Enforcement
Example Source: Callingbull.org & Reuters

Prevalence of wrist/thumb injuries in population and smartphone coverage in Bristol, UK



Injuries after smartphone coverage

Claim: "The rise of smartphones in the population have dramatically increased prevalence of wrist/thumb injuries."

Q5: How confident are you in the claim based on the visualization?

A. Very confident

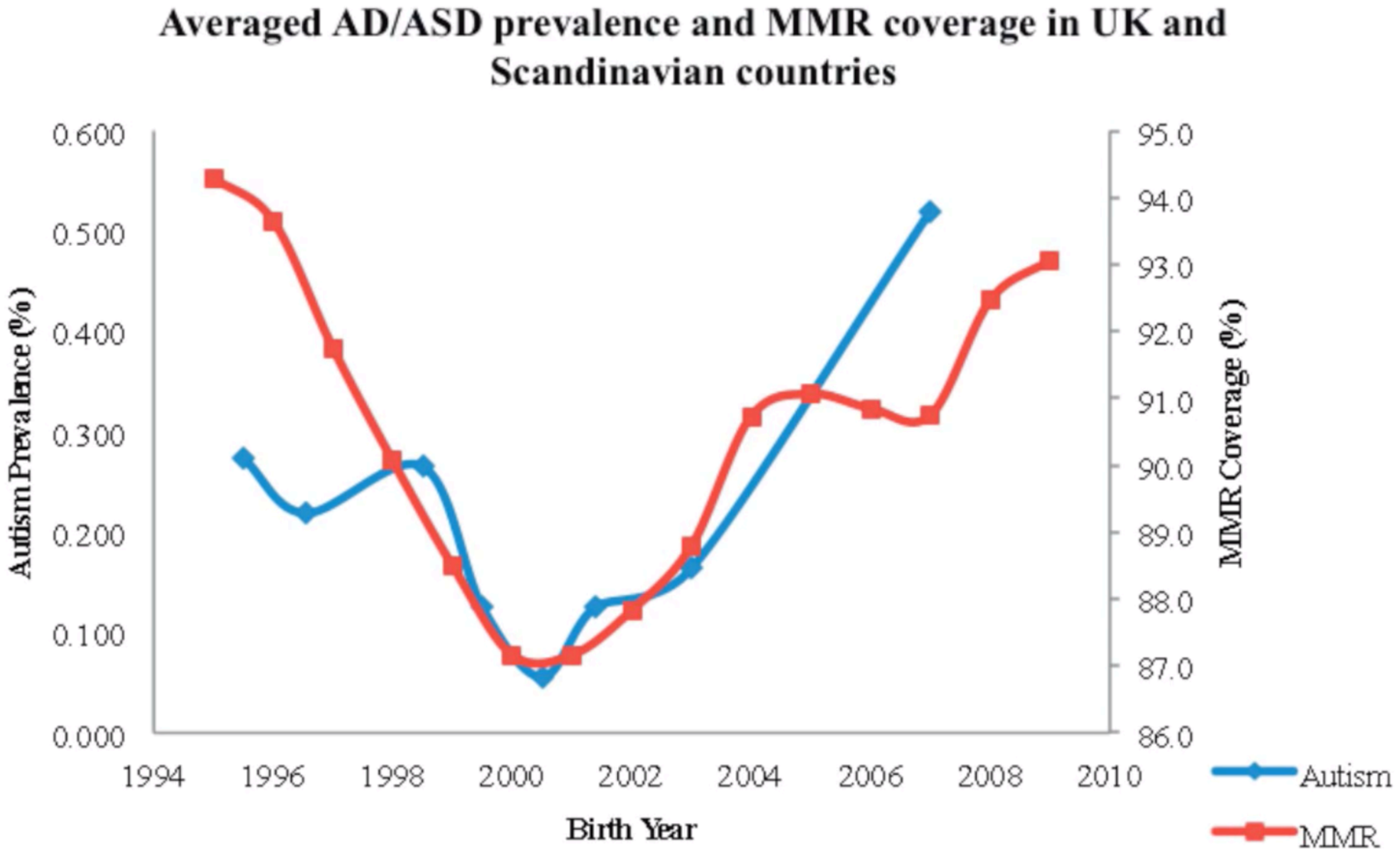
B. Confident

C. Not Confident

D. Claim is wrong

Adapted from: callingbull.org

Autism and MMR coverage



Claim: "The rise of smartphones in the population have dramatically increased prevalence of wrist/thumb injuries."

Q5: How confident are you in the claim based on the visualization?

A. Very confident

B. Confident

C. Not Confident

D. Claim is wrong

Adapted from: callingbull.org

Autism and MMR coverage

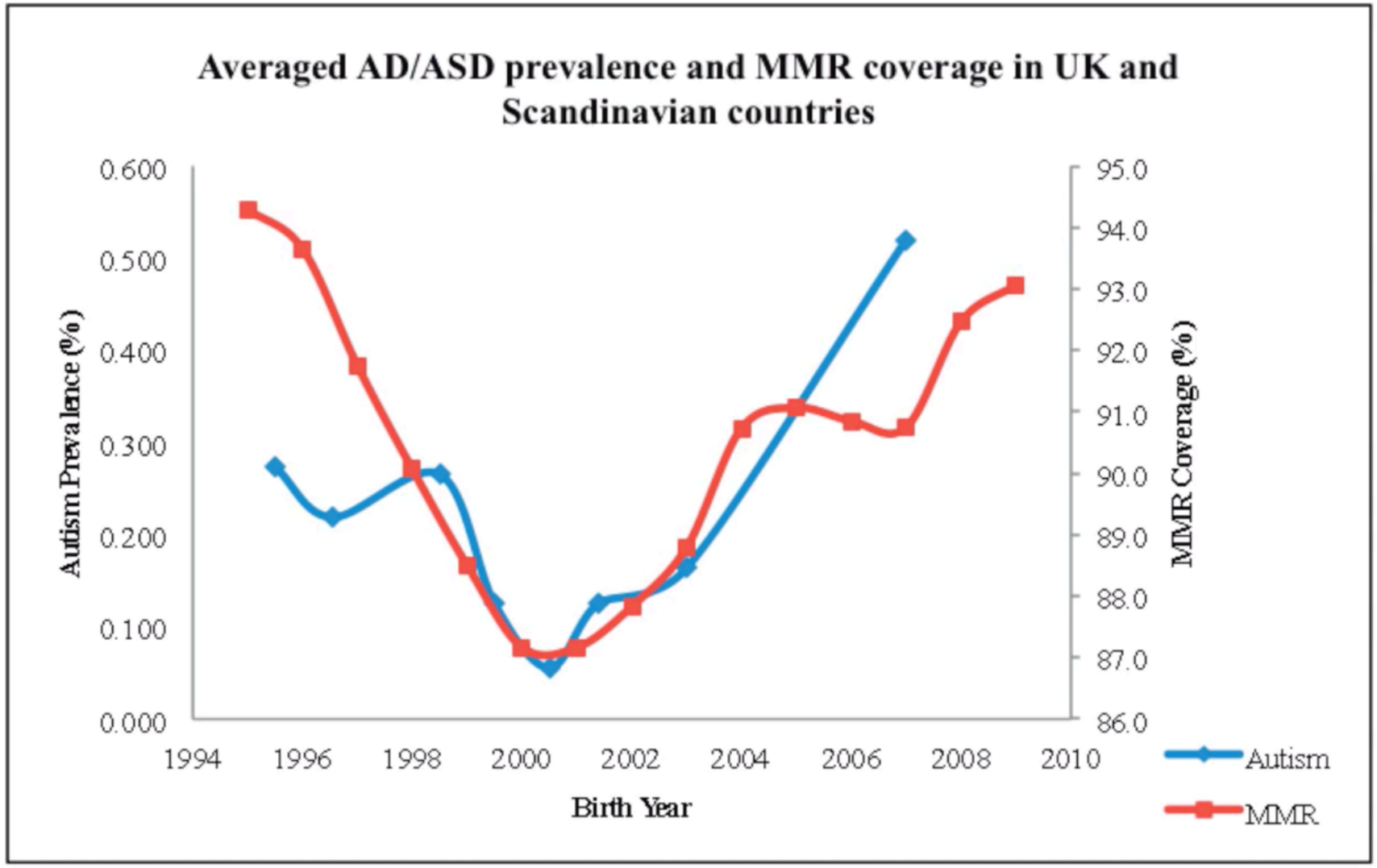


Figure 1-Averaged AD/ASD prevalence and MMR coverage in UK, Norway and Sweden. Both MMR and AD/ASD data are normalized to the maximum coverage/prevalence during the time period of this analysis.

Claim: "The rise of MMR coverage in the population have dramatically increased prevalence of Autism Spectrum Disorder."

Q5: How confident are you in the claim based on the visualization?

A. Very confident

B. Confident

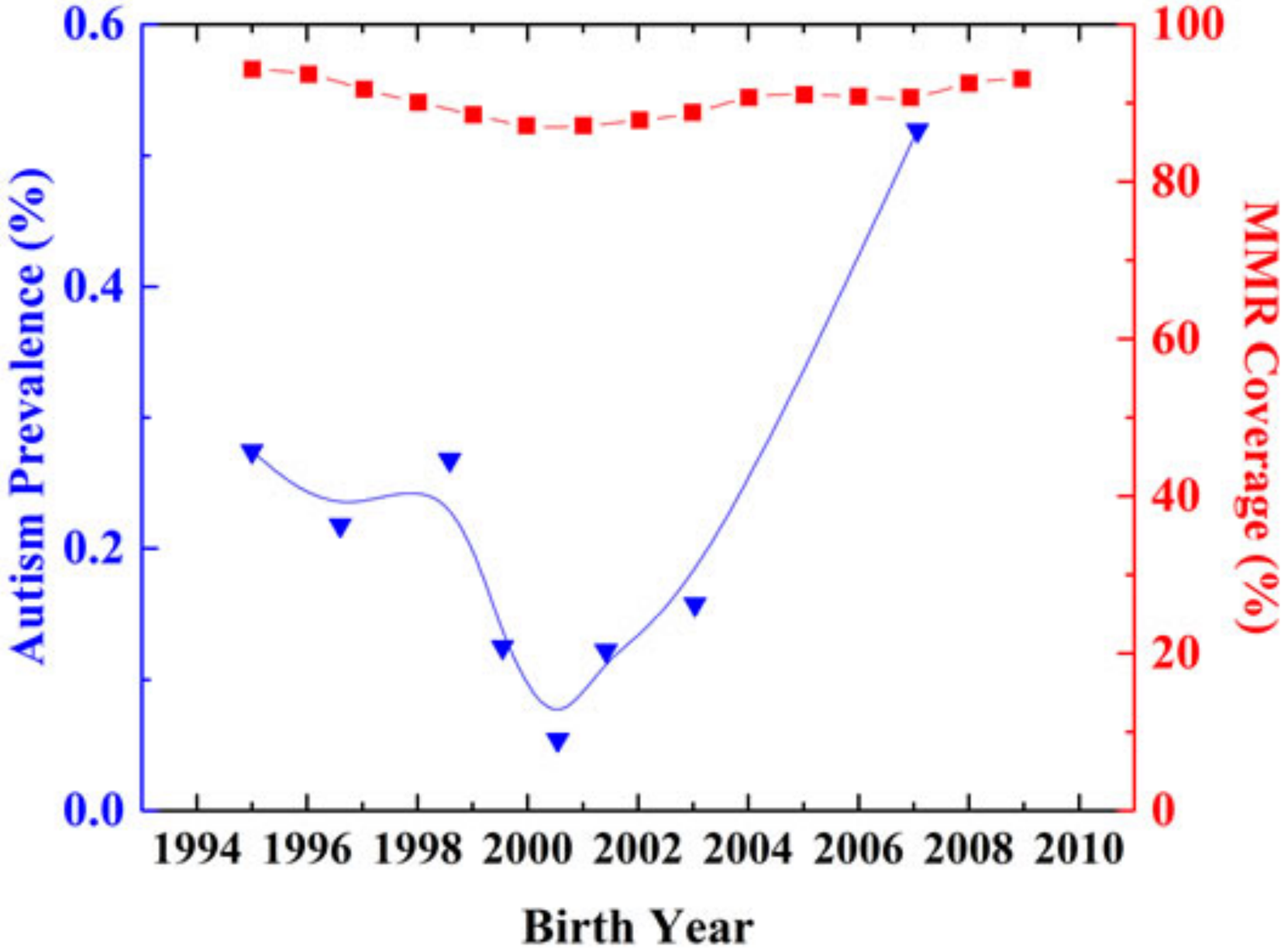
C. Not Confident

D. Claim is wrong

Diesher *et al.* 2015 *Issues in Law and Medicine*

Adapted from: callingbull.org

Autism and MMR coverage



Revised graph courtesy of Matt Carey

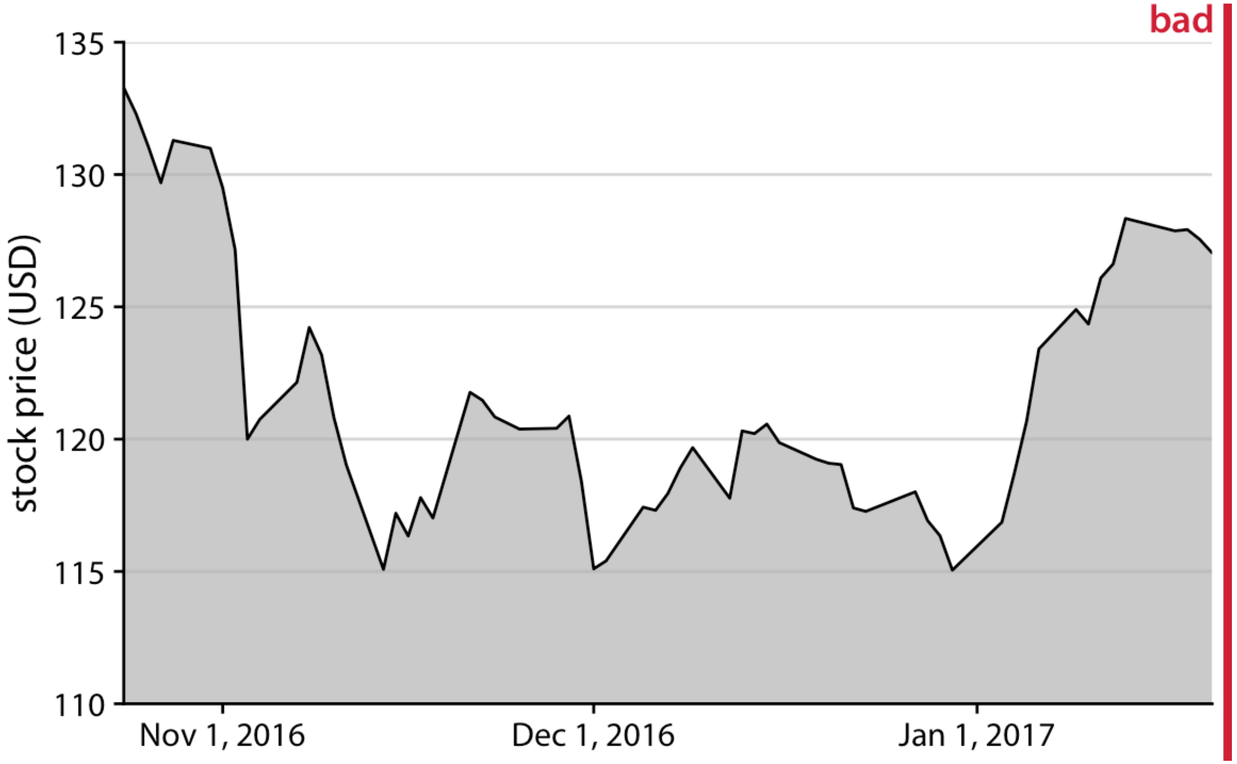
Full analysis/critique of paper

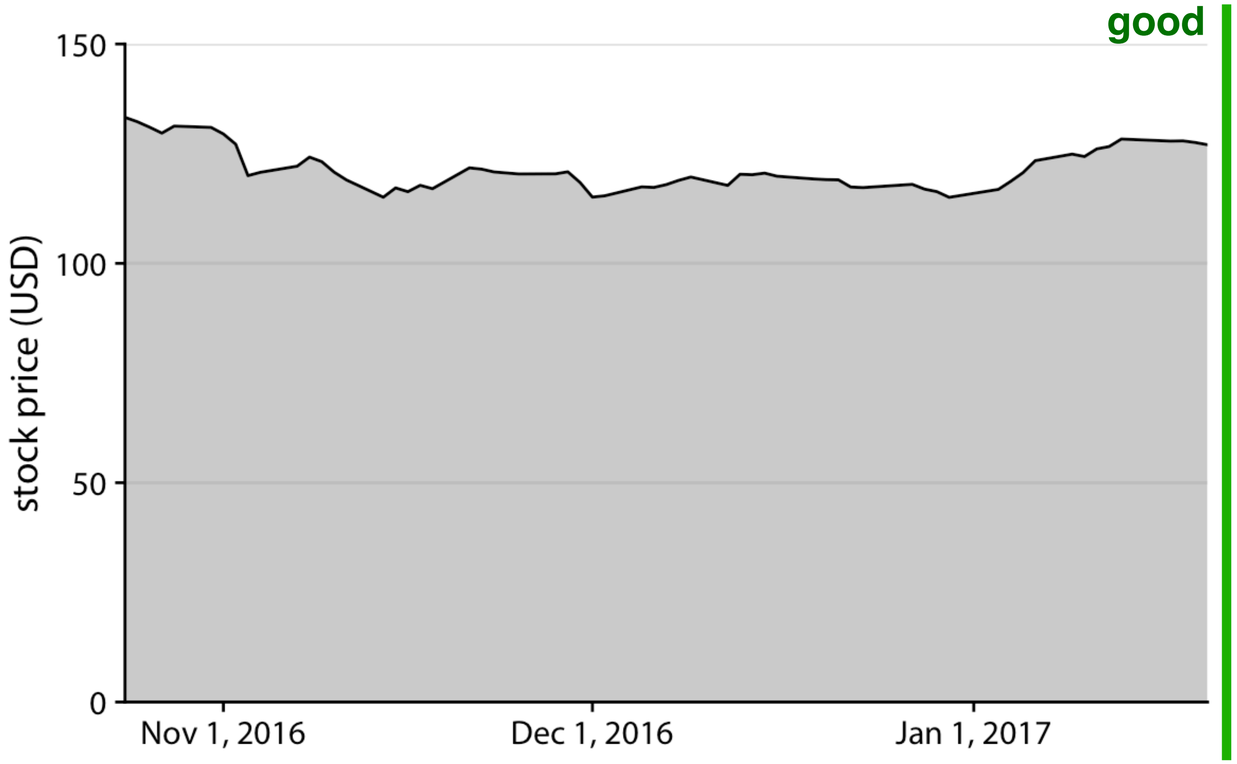
Adapted from: callingbull.org

20 minutes

Part 2: Principles of Effective Visualizations

Principle	Definition	Examples
• Proportional Ink	The amount of ink used to indicate a value should be proportional to the value itself.	Truncating the y-axis on a bar chart to exaggerate the difference between bars violates the principle of proportional ink

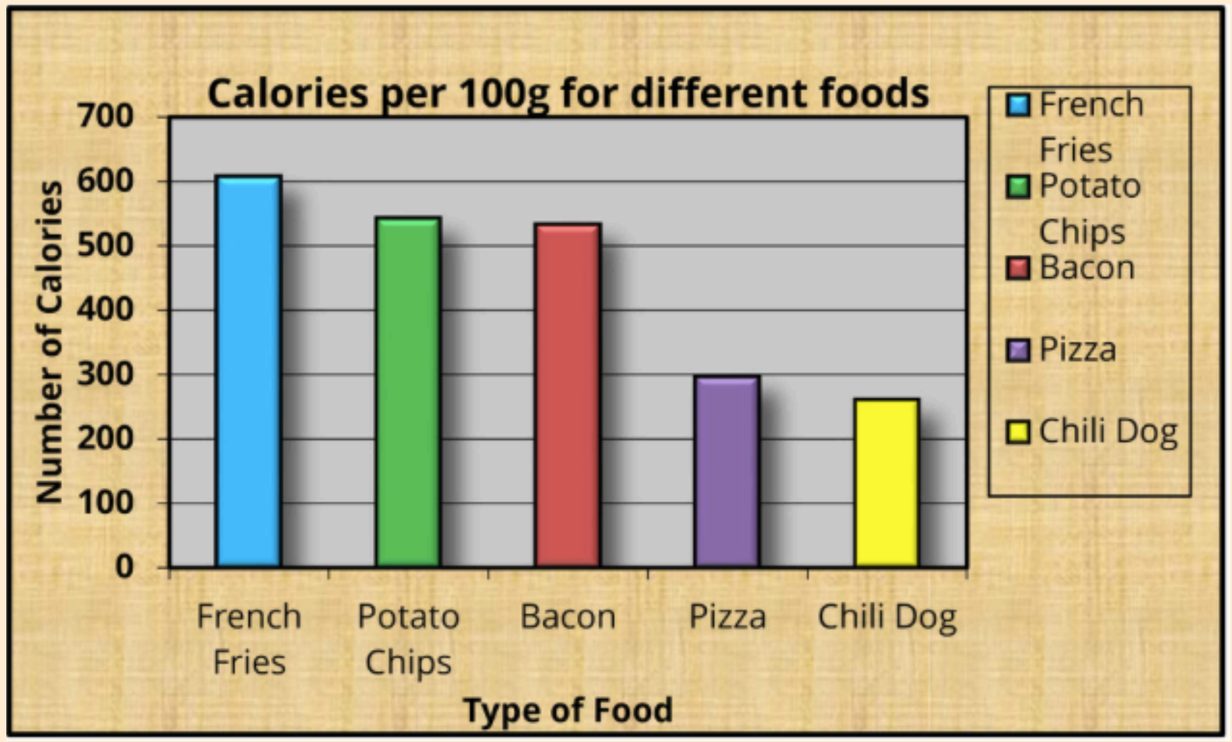




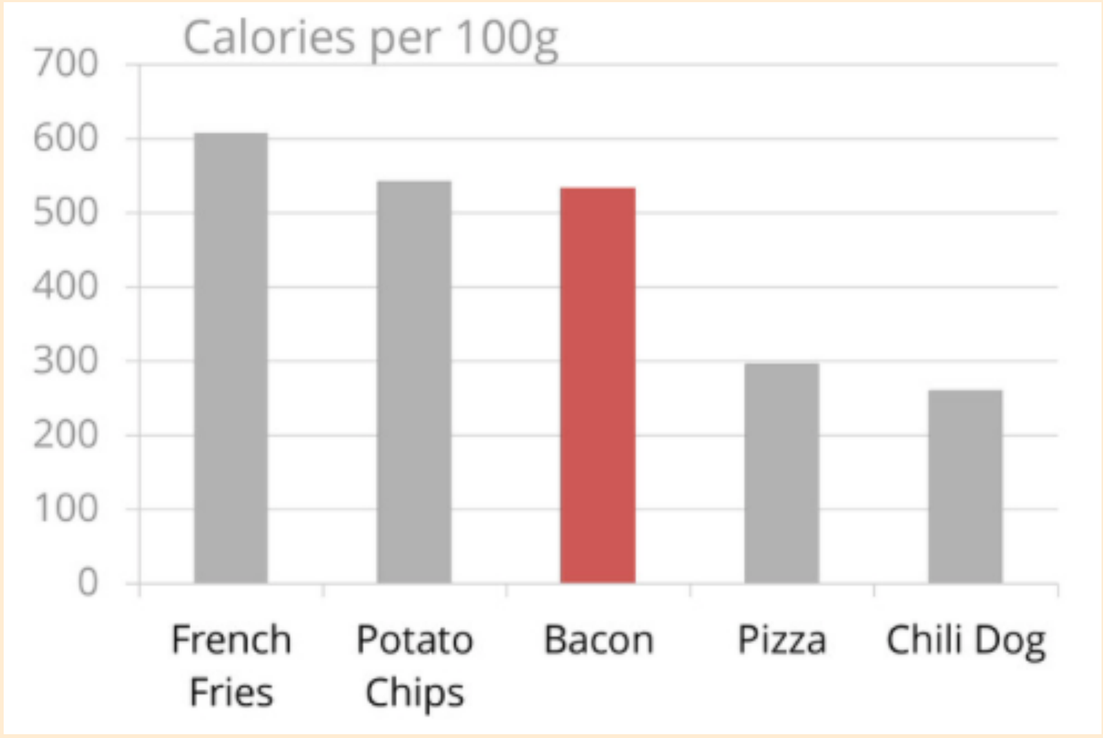
Source: Fig 17.3 and 17.4 of Fundamentals of Data Visualization

Principle	Definition	Examples
• Data:ink ratio	Remove distracting visual elements to focus attention on the data	Lighten line weights, remove backgrounds, never use 3D or special effects, remove unnecessary/redundant labels, etc

bad



good



Source: Dark Horse Analytics

Principles of Effective Visualizations

Principle	Definition	Examples
• Proportional Ink	The amount of ink used to indicate a value should be proportional to the value itself.	Truncating the y-axis on a bar chart to exaggerate the difference between bars violates the principle of proportional ink.
• Data:ink ratio	Remove distracting visual elements to focus attention on the data	Lighten line weights, remove backgrounds, never use 3D or special effects, ~~remove~~ avoid unnecessary/redundant labels.
• Labels & legends	Use axes labels and titles to highlight/ communicate data	Never leave your data column names as axes labels! Generally good to add a title.
• Overplotting	With large datasets, points overlap, resulting in large clouds of data	To fix overplotting, could plot just a sample subset of the data, use alpha, and use smaller points. Or, jitter - but check if appropriate!
• Visualization choice	Must be informed by the **data** you have, the **research question** being asked and the **audience** that cares.	Pick the simplest plot that best shows most/all of the data needed to answer the research question. If you only have summary statistics, cannot show distributions. Tailor the visualization to your audience (within reason) but don't dumb it down.
• Colour & Accessibility	Colour can be used to encode information or for aesthetics/style/ design. However, colour can also be distracting if used inappropriately or poorly.	Choose a perceptually uniform colour palette; can be sequential or diverging for quantitative data. Opt for colour-blind friendly palettes. Categorical data can use qualitative colour schemes.

20 minutes

Part 3:
Exploratory Data Analysis

7 Exploratory Data Analysis

7.1 Introduction

This chapter will show you how to use visualisation and transformation to explore your data in a systematic way, a task that statisticians call exploratory data analysis, or EDA for short. EDA is an iterative cycle. You:

1. Generate questions about your data.

2. Search for answers by visualising, transforming, and modelling your data.

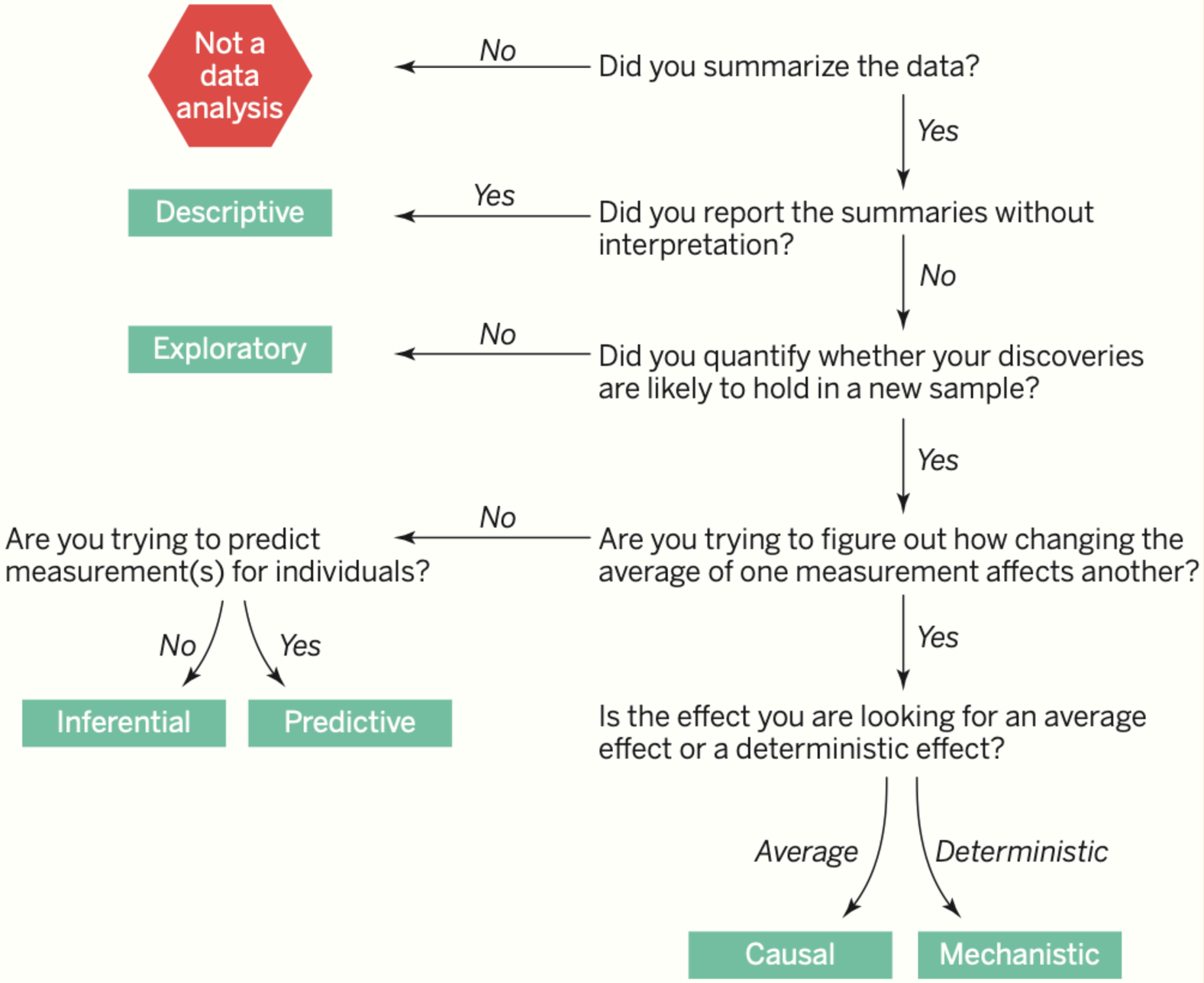
3. Use what you learn to refine your questions and/or generate new questions.

EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will home in on a few particularly productive areas that you'll eventually write up and communicate to others.

Types of Research Questions

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Data analysis flowchart



Source: What is the Question?
By Jeffery Leek and Roger Peng

Research Questions

1. Descriptive

- one that seeks to summarize a characteristic of a set of data
- no interpretation of the result itself as the result is a fact, an attribute of the data set you are working with

- e.g., What is the frequency of viral illnesses in a set of data collected from a group of individuals?
- e.g., How many people live in each US state?

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

2. Exploratory

- one in which you analyze the data to see if there are patterns, trends, or relationships between variables
- looking for patterns that would support proposing a hypothesis to test in a future study

- e.g., Do diets rich in certain foods have differing frequencies of viral illnesses in a set of data collected from a group of individuals?
- e.g., Does air pollution correlate with life expectancy in a set of data collected from groups of individuals from several regions in the United States?

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

3. Inferential

- **one in which you analyze the data to see if there are patterns, trends, or relationships between variables in a representative sample**
- **want to quantify how much the patterns, trends, or relationships between variables is applicable to all individuals units in the population**

- **e.g., Is eating at least 5 servings a day of fresh fruit and vegetables is associated with fewer viral illnesses per year?**
- **e.g., Does air pollution correlate with life expectancy in the United States?**

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

4. Predictive

- **one where you are trying to predict measurements or labels for individuals (people or things)**
- **less interested in what causes the predicted outcome, just what predicts it**

- **e.g., How many viral illnesses will someone have next year?**
- **e.g., What political party will someone vote for in the next US election?**

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

5. Causal

- asks about whether changing one factor will change another factor, on average, in a population.
- Sometimes the underlying design of the data collection, by default, allows for the question that you ask to be causal (e.g., randomized experiment or trial)

- e.g., Does eating at least 5 servings a day of fresh fruit and vegetables cause fewer viral illnesses per year?
- e.g., Does smoking cause cancer?

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Research Questions

6. Mechanistic

- **one that tries to explain the underlying mechanism of the observed patterns, trends, or relationship (how does it happen?)**

- **e.g., How do changes in diet lead to a reduction in the number of viral illnesses?**
- **e.g., How does airplane wing design changes air flow over a wing, leading to decreased drag?**

Source: Art of Data Science, by Roger Peng and Elizabeth Matsui

Types of Research Questions

Exploratory Data Analysis

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

Steps of EDA

1. Describe your dataset

rubric:{reasoning:4}

Task: Describe your dataset. Consider the following questions to guide you in your exploration

- Who: Which company/agency/organization provided this data?
- What: What is in your data?
- When: When was your data collected (for example, for which years)?
- Why: What is the purpose of your dataset? Is it for transparency/accountability, public interest, fun, learning, etc...
- How: How was your data collected? Was it a human collecting the data? Historical records digitized? Server logs?

Steps of EDA

2. Load the dataset

rubric:{correctness:1}

Task: Load your dataset from a file, or URL. This needs to be a pandas dataframe so you can use it with Seaborn. Remember that others may be running your jupyter notebook so it's important that the data is accessible to them. If your dataset isn't accessible as a URL, make sure to commit it into your repo.

Steps of EDA

3. Explore your dataset

rubric:{correctness:5}

Task: Explore the columns in your dataset. Which ones are interesting/relevant? You can use df.profile_report()

To install pandas-profiling:

conda install -c conda-forge pandas-profiling

Source: EDA with Pandas Profiling

Steps of EDA

4. Initial thoughts

rubric={correctness:1}

Task: Use this a place to record any observations you come up with, anything jump out at you as surprising or particularly interesting? Where do you think you'll go with exploring this dataset? Feel free to take notes in this section and use it as a scratch pad. Any content in this area will not be marked.

Steps of EDA

5. Wrangling

rubric={correctness:1}

Task: You can do any wrangling you need to do here. Describe what you're doing (or did) using comments within your code.

Steps of EDA

6. Research questions

rubric={reasoning:5}

Task: come up with at least two research questions about your dataset that will require data visualizations to help answer. Recall that for this purpose, you should only aim for "Descriptive" or "Exploratory" research questions.

Steps of EDA

7. Data Analysis & Visualizations

rubric={viz:40, reasoning:10}

Task: Create data visualizations (and justify your choices) using Seaborn that will help you answer your research questions.

Steps of EDA

8. Summary and conclusions

rubric={reasoning:10}

Task: Summarize your findings and describe any conclusions and insight you were able to draw from your visualizations.