Formative Assessment: Deceptive Visualization

Deceptive Visualization

Data visualizations are becoming a key medium for the public to understand news and information. It's crucial to recognize how the design of a visualization can affect what people understand and remember from the data. In this task, you need to pick a dataset you find interesting and create two static visualizations using the same dataset. The first should be a **truthful** representation of the data. The second should be a **deceptive** visualization, designed to deceive the viewer. However, you should avoid clear distortions or leaving out information for this deceptive visualization.

Assignment

Your objective is to create two static (single image) visualizations of a selected dataset. The first visualization should be designed to clearly and sincerely convey insights from the data. In contrast, the second should be crafted to intentionally mislead the viewer, causing them to make incorrect inferences. Additionally, you are required to **write a brief explanation**, limited to four paragraphs, outlining your design strategy for both visualizations.

In this task, an earnest visualization is defined as one that:

- Is easily understandable and can be interpreted by the general public.
- Uses visual encodings that are suitable and effective for the desired purpose.
- Clearly and openly describes any transformations made to the data.
- Transparently communicates the source of the data and any possible biases involved.

Conversely, a deceptive visualization typically displays these characteristics:

- The graphical depiction is deliberately unsuitable or deceptive.
- Headings are crafted to influence the viewer's understanding in a biased manner.
- There is intentional manipulation or selective filtering of data or axes to deceive.
- It's not transparent about possible bias present in the data.

For the earnest visualization, your goal is to be as clear and transparent as possible to help viewers answer your intended question. For the deceptive visualization, your goal is to trick the viewer (including the course staff!) into believing that the visualization is legitimate and earnest. **It should not be immediately obvious which visualization is trying to be deceptive.** Subtle ineffective choices in the design should require close and careful reading to be identified.

For the deceptive visualization, misleading strategies are fine but outright lying is not. For example, sketchy, unreliable or untrustworthy input datasets are discouraged, but misleading omission, filtering, or transformation of trustworthy data records is fine. Deliberate lies in the title, axes, labels, or annotations is discouraged, but technically true/relevant but otherwise misleading text in the visualization is fine.

For both visualization designs, start by choosing a question you would like to answer. Design your visualization to answer that question either correctly (for the earnest visualization) or incorrectly (for the deceptive visualization). You may choose to address a different question with each visualization. Be sure to document the question as part of the visualization design (e.g., title, subtitle, or caption) and in your assignment write-up.

Your write-up should contain the following information:

- The specific question each visualization aims to answer.
- A description of your design rationale and important considerations for each visualization.

Recommended Data Sources

To help get you started, this assignment, we've provided two possible datasets for you to use, although you're welcome to select any dataset you prefer. In class we demonstrated the data analysis workflow using a dataset of emissions data from Apple. We also looked at climate change data of the global mean temperature anomaly. Both of these datasets have been provided for you on Moodle and are well suited to use in the assignment. However, you may use any dataset you find interesting, whether it's one we've used in class before, the others listed below, or any dataset you can find.

You must use the same dataset for both visualizations, but you may transform the data differently, use additional data variables, or choose to address a different question for each design. These datasets are intentionally chosen to cover politically charged topics for the simple reason that these are typically the types of data where deceptive visualizations may proliferate.

Apple Emissions

You may use the Apple Emissions dataset we started to look at in class which can be downloaded from Moodle. Find the source here

A breakdown of Apple's greenhouse gas emissions from 2015 to 2022 as they aim to reach net zero emissions by 2030. This includes every source of emissions from both their corporate operations and their product life cycle, the carbon footprint of their baseline iPhone in the same period, and normalizing factors like sales, market cap, and employees.

Some Recommended Analysis

- How much has Apple reduced their emissions from 2015 to 2022?
- How does this trend compare to their revenue & market cap trend in the same period?
- Which areas have seen the most improvement? What about the least?
- Is Apple on track to meet their 2030 goal of net zero emissions?

Other Data Sources

Data on Energy by Our World in Data, 1900-2022

Our World in Data, a non-profit that gathers and analyzes data about global issues, has published data about energy usage for countries (e.g. coal consumption, hydropower consumption, etc.) around the world since 1900. You can download the data here.

Education Data

Every year, the federal government releases large amounts of data on US schools, districts, and colleges. However, this information is scattered across multiple datasets. Urban Institute's Education Data Explorer tries to fix this issue by putting together data from various sources such as the National Center for Education Statistics' Common Core of Data (CCD), the Civil Rights Data Collection (CRDC), the US Department of Education's EDFacts, and IPUMS' National Historical Geographic Information System (NHGIS) and makes it available as an API. You can download the data by making an API call using the code available on the website or alternatively clicking on the downloads button on the website.

Internet Usage Data

UNdata brings international statistical databases within easy reach of users through a single-entry point. It is maintained by the Development Data Section of the Development Data and Outreach Branch within the Statistics Division of the Department of Economic and Social Affairs (UN DESA) of the UN Secretariat. You can find the internet usage data here. Feel free to take a look at some of the other datasets made available by UNdata here.

This data has the following columns:

- `Region/country Code:` code representing the country or region.
- `Region or Country Name:` Field containing the country name.
- `Year:` Field containing the year at which the data was collected.
- `Value:` Field denoting the Percentage of individuals using the internet.
- `Source:` Field denoting the source of the data.

Here are some other possible sources to consider. You are also free to use data from a source different from those included here. If you have any questions on whether a dataset is appropriate, please ask the course staff ASAP!

- City of San Diego open data
- U.S. Government Open Datasets
- U.S. Census Bureau - Census Datasets
- IPUMS.org - Integrated Census & Survey Data from around the World
- Federal Elections Commission - Campaign Finance & Expenditures
- Federal Aviation Administration - FAA Data & Research
- NOAA Daily Weather - NOAA Daily Global Historical Climatology Network Data
- yelp.com/dataset - Yelp Open Dataset
- fivethirtyeight.com - Data and Code behind the Stories and Interactives
- Buzzfeed News - Open-source data from BuzzFeed's newsroom

Bibliography