Exercise 1 - The Data Journal - Part 1

Data, CDAD-UH 1001Q, Spring 2022

Assigned: January 24, 2022 Due: January 31, 2022

Preliminaries

For this exercise, you will work ***individually***. You need to work with Microsoft Excel, Google Sheets or any spreadsheet software of your choice.

We prefer you begin getting familiar with Microsoft Excel. We will use Excel extensively for in-class demonstrations and future exercises.

If you haven't used spreadsheet software make sure you go through the following resources:

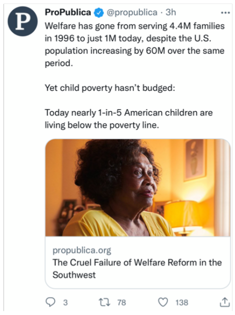
1. Chapter 1 of Data Smart
2. https://education.microsoft.com/en-us/resource/86835256

You can get a free student version by getting in touch with the center for academic technology at the NYUAD library.

1 Personal "Data" Exposure Diary

In our daily lives, we are exposed to a variety of ***data stories*** on the social media and news platforms we subscribe to. You may be part of a WhatsApp group or mailing list where family and friends send you articles with some form of data-backed claim.

For example, on January 16, this is a sample of some of the data stories I was exposed to:



ProPublica @propublica · 3h

Welfare has gone from serving 4.4M families in 1996 to just 1M today, despite the U.S. population increasing by 60M over the same period.

Yet child poverty hasn't budged:

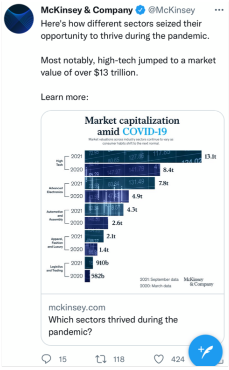
Today nearly 1-in-5 American children are living below the poverty line.

propublica.org

The Cruel Failure of Welfare Reform in the Southwest

3 78 138

(a) A story I got on my twitter feed in the morning



McKinsey & Company @McKinsey

Here's how different sectors seized their opportunity to thrive during the pandemic.

Most notably, high-tech jumped to a market value of over $13 trillion.

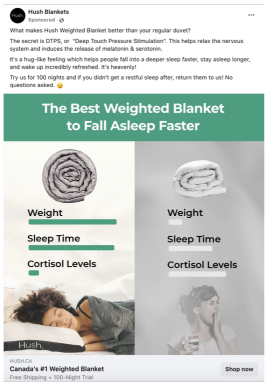
Learn more:

mckinsey.com

Which sectors thrived during the pandemic?

15 118 424

(b) Good to know tech is doing well according to McKinsey & co



Hush Blankets

Sponsored ·

What makes Hush Weighted Blanket better than your regular duvet?

The secret is DTPS, or "Deep Touch Pressure Stimulation". This helps relax the nervous system and induces the release of melatonin & serotonin.

It's a hug-like feeling which helps people fall into a deeper sleep faster, stay asleep longer, and wake up incredibly refreshed. It's heavenly!

Try us for 100 nights and if you didn't get a restful sleep after, return them to us! No questions asked. 😏

HUSH.CA

Canada's #1 Weighted Blanket

Free Shipping + 100-Night Trial

Shop now

(c) Sleepless scrolling lands me on this data gem!

For this exercise, you will track for **one week** your exposure to such data articles and stories to better understand how you personally are subjected to, influenced or manipulated by data.

In part 2 of this exercise, you will create a visualization of this personal data-set that illustrates a personal finding, or lesson learned from this exercise.

You need to approach this journaling exercise with a bit of a scientific, critical-thinking mindset. You are building a data-set rather than simply a collection or folder of data stories. So, for each story, you will also record the different features of the story and you will thoroughly explain your process.

1.1 The Process

Turn in: **Writeup.pdf/.txt/.doc**: A 1-2 page explanation of how you approached this journaling exercise.

As you begin your collection process, consider the following questions. Your write-up should provide answers to these questions.

1. [1 point] What defines a data story for you? This definition should help you determine whether or not you will include the data story in your journal or not: so aim for a non-ambiguous definition.
2. [1 point] What features will you track for each data story? Here are some examples of features to inspire you that are in no way exhaustive or required:
   - Timestamp, *when* did you read the story?
   - Source, *where* did you read the story?
   - Summary, *what* was the main claim?
   - Fake?, how *plausible* is the claim?
   - Reference(s), what are the scientific papers/data sets, if any, does it refer to ?
   - Impact, how did you *feel* reading the claim?
   - Link/Screenshot

   You may wish to track other features as well such as social behaviors around the story that you witnessed.
3. [2 points] Why are you tracking these features?
   - What are some of the questions you hope to answer through this data set?
   - Which features will help you answer these questions?
   - Which features are not immediately pertinent to this self-exploration exercise but can be useful and why?
4. [2 points] Revisions and Extensions
   - Did you decide to change your process at any point and why? How did you handle the data collected retrospectively?
   - Would/Did you re-encode a specific feature? e.g. re-encode free form text to a categorical representation, or represent a binary (yes/no) value as some numerical score (1-5) or vice-versa? Why did you do so?

     An example of re-encoding a free-form text title feature into a categorical one. For data story (c) “The Best Weighted Blanket to Fall Asleep Faster” I might assign it the categorical value health for the topic feature.
5. [2 points] What are some of the limitations, flaw and challenges you faced during the journaling process? Here are some issues that we often run into during data collection but again they are not exhaustive and you may wish to elaborate on other challenges that you faced or other issues with your sample.
   - How representative this data set is with respect to your typical data exposure?

- How difficult was it to categorize or track the different features?
- Are there missing values in your data set? Why?

[2 points] Quality of your write up in terms of clarity, writing style and organization.

1.2 The Spreadsheet

Turn in: A spreadsheet where you record information on each individual story in a table.
Each data story will occupy a row in this table and the different features will form the columns of this spreadsheet table.

1. [5 points] The spreadsheet and data set quality will be judged on at least the following criteria:
   - completeness: are all your features populated? are there missing values for good reason?
   - consistency: do you use a consistent representation, scoring, labeling scheme, etc. when assigning values to your data
   - clarity: is it easy to understand what the *schema* (feature name and range of possible values) is for your data and do you provide comments to describe what the different features are? If you hand over your data set to a peer, can they easily understand and use your data set?
   - style: is your spreadsheet well formatted/styled/organized and organized.

Submission

1. You will package your solution into a folder with the title: 'Ex1-Part1-netID-firstname'. (e.g. Ex1-Part1-aa175-azza is the name of the folder I would submit).

2. This folder will contain exactly **two** files: (a) a Spreadsheet file (.csv, .xls or .xlsx), (b) a write-up file (.txt or .doc or .pdf).

3. **Zip this folder** and submit a .zip folder via DropBox at the following link: http://bit.ly/Data-F22-Ex1-Part1

We will not grade any submission that does not strictly follow the submission rules.