Introduction

CS215 is a class that can be tedious at times. The word ‘data’ is repeated over and over again until it begins to lose its meaning: “What is data? How can you use data? When should you not use data?” It can be easy to lose sight of the meaning of the work and start doing it mechanically.

Here are a couple of the things that I learned while taking this course, including what helped me keep the data interesting and real to me as a human in the world.

Approach your dataset with a healthy dose of skepticism

Hans Rosling talked about in his article “Factfulness: Introduction & The Gap Instinct” that misrepresentation of data has quite a powerful effect on people. Rosling distributed a quiz to a number of students, doctors, and professors that included questions such as the ones below. He found that the vast majority of people he quizzed did astonishingly badly on it, scoring more poorly than if they had guessed at random. This means something else is at play…

1. In all low-income countries across the world today, how many girls finish primary school?
   - ☐ A: 20 percent
   - ☐ B: 40 percent
   - ☐ C: 60 percent

2. Where does the majority of the world population live?
   - ☐ A: Low-income countries
   - ☐ B: Middle-income countries
   - ☐ C: High-income countries

3. In the last 20 years, the proportion of the world population living in extreme poverty has …
   - ☐ A: almost doubled
   - ☐ B: remained more or less the same
   - ☐ C: almost halved

4. What is the life expectancy of the world today?
   - ☐ A: 50 years
   - ☐ B: 60 years
   - ☐ C: 70 years

5. There are 2 billion children in the world today, aged 0 to 15 years old. How many children will there be in the year 2100, according to the United Nations?
   - ☐ A: 4 billion
   - ☐ B: 3 billion
   - ☐ C: 2 billion

6. The UN predicts that by 2100 the world population will have increased by another 4 billion people. What is the main reason?
   - ☐ A: There will be more children (age below 15)
   - ☐ B: There will be more adults (age 15 to 74)
   - ☐ C: There will be more very old people (age 75 and older)

7. How did the number of deaths per year from natural disasters change over the last hundred years?
   - ☐ A: More than doubled
   - ☐ B: Remained about the same
   - ☐ C: Decreased to less than half

We tend to offer data points a lot of weight to tell us about the world. Something about logic is so compelling to us, but remember that logic is a man made invention. It’s a method of reasoning that humans came up with to navigate the world. Try to keep this in mind when

approaching a dataset, acknowledge how the data was collected and what the limitations are of what it can tell you.

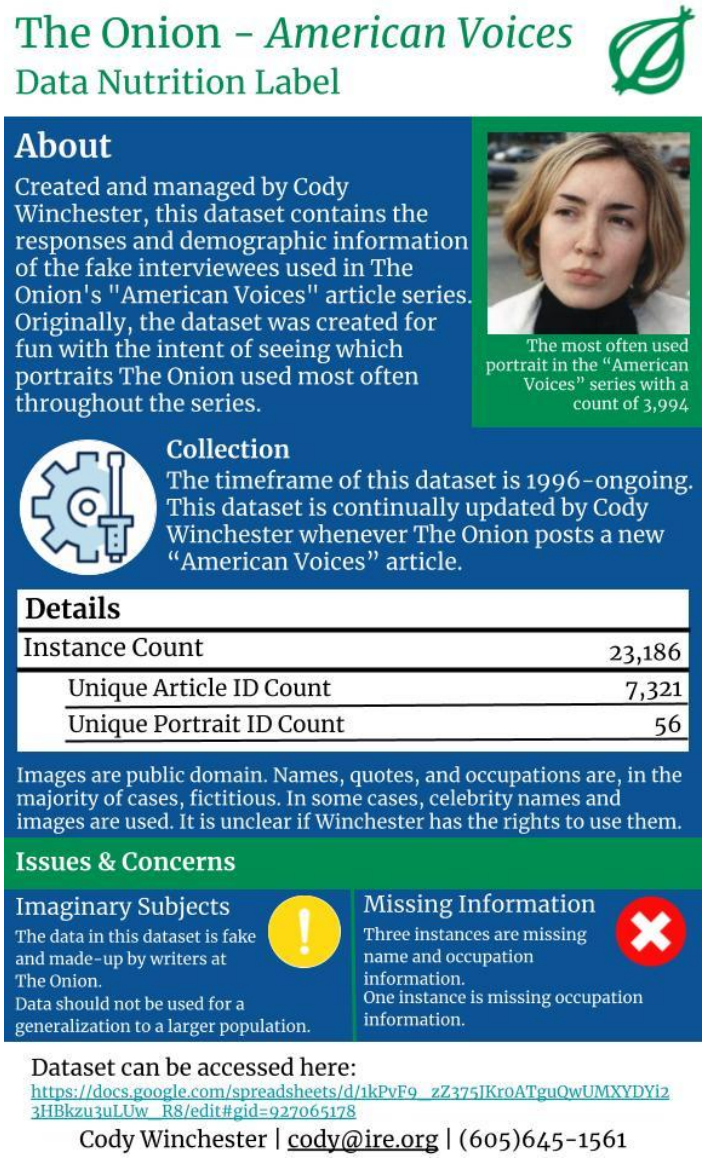
For example, for Project 2 in this class we measured the height of Stanton Hall using our phones, measuring acceleration on the elevator from the basement to the top floor. If I told you I knew that the height of Stanton Hall (at least to the bottom of the top floor from the basement) was 22 feet tall, you ought to approach that number with some skepticism. How was that data collected? What metrics were used? What human error could have occurred? That's somewhat of a simplistic and transparent example, but you get the idea.

"Healthy" is the key word here, however. I think this semester I've tended to approach things with quite a lot of skepticism, which isn't always helpful either. Flexibility is key, approaching with too much skepticism can close you off and keep you from taking advantage of opportunities to discover new things.

Be creative, have fun

Always, always choose the fun path. The website Data is Plural was my best friend this semester because they have so many fun and interesting datasets. No matter what, language is a human construction, a human structure for communication, so little pieces of you as a human are going to shine through in your data communications. If you are having fun, that is going to show through.

For my Data Nutrition Label, I chose a dataset that had compiled quotes and demographic data from all of The Onion's "American Voices" article series. This was not a traditional dataset; there weren't really trends or comparisons that it would be easy for anybody to make. For a lot of the questions that asked "What could you do with this dataset?" my answer was "Well, not a whole lot."



But I had a lot of fun working with the dataset and looking at all the silly quotes, and that pushed me to explore the dataset a little more deeply than I might have with one I wasn't interested in. Always take the fun path, even if it's harder!

Pandas.org is your friend!

Every function has its own parameters and attributes, and it would take a lot of experience to memorize them all. Pandas.org has a library that lists available functions to use in your program, as well as its parameters and attributes.

A parameter is an input that you give a function for it to manipulate. An attribute is a component of an object created by a function that can be returned in isolation. Both are super super handy when manipulating data frames in Pandas!

Here, for example, is part of the list of attributes of the plot function on Pandas.org:

pandas.DataFrame.plot

DataFrame.plot(*args, **kwargs) [source]

Make plots of Series or DataFrame.

Uses the backend specified by the option `plotting.backend`. By default, matplotlib is used.

Parameters:

data : *Series or DataFrame*

The object for which the method is called.

x : *label or position, default None*

Only used if data is a DataFrame.

y : *label, position or list of label, positions, default None*

Allows plotting of one column versus another. Only used if data is a DataFrame.

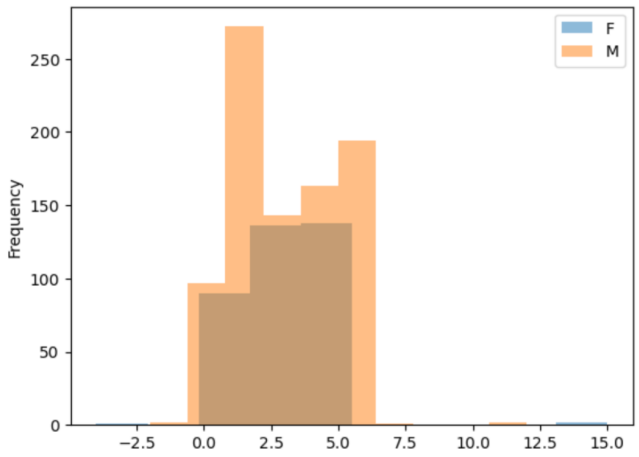
kind : *str*

The kind of plot to produce:

- 'line' : line plot (default)
- 'bar' : vertical bar plot

All of these are parameters that you can give the plot function, as well as descriptions of the type of information the parameter can take. I relied on this entry for the plot function a lot when I was doing Project 9 to visualize my Taskmaster dataset.

```
In [14]: points_gendered.groupby('gender')['points'].plot(kind='hist',alpha = 0.5)
         plot.legend()
Out[14]: <matplotlib.legend.Legend at 0x7fc6d952e310>
```



As you can see, I used the ‘kind’ parameter to choose what kind of plot I wanted to make, as well as the ‘alpha’ parameter to manipulate the transparency of the buckets.

Pandas.org also has lists of attributes for applicable functions. I relied on this entry on timestamps a lot for Project 8 when we were manipulating timestamp data.

Attributes

`asm8`	Return numpy datetime64 format in nanoseconds.
`day_of_week`	Return day of the week.
`day_of_year`	Return the day of the year.
`dayofweek`	Return day of the week.
`dayofyear`	Return the day of the year.
`days_in_month`	Return the number of days in the month.
`daysinmonth`	Return the number of days in the month.
`is_leap_year`	Return True if year is a leap year.
`is_month_end`	Check if the date is the last day of the month.
`is_month_start`	Check if the date is the first day of the month.

Attributes made it a lot easier to access the day, month, and year from the timestamp data, as opposed to having to convert the information to a string and manually pull that data using indexing. There are a lot of shortcuts like this!

```
In [10]: timestamps['day']=''
         timestamps['month']=''
         timestamps['year']=''
         timestamps['date']=''
         for i in range(len(timestamps)):
             temp = pd.to_datetime(timestamps['timestamp'][i])
             timestamps['day'][i]=temp.day
             timestamps['month'][i]=temp.month
             timestamps['year'][i]=temp.year
             timestamps['date'][i]=str(temp.month)+'-'+str(temp.day)+'-'+str(temp.year)
```

Remember the big picture

There are people and experiences behind each data point. That can be easy to lose track of. When I consider this point, I think about a visualization I saw about gun deaths from a different class, which you can see here.

With so much data, it’s sometimes hard to keep track of what you are actually working with. The meaning of each datapoint, the life behind it, fades into the background of the

visualization. It becomes meaningless in some ways, the data becomes a means to an end, numbers to crunch to try and come up with some sort of conclusion.

It's important to keep in mind the consequences of the data you are working with. What does this data actually mean? Let the gravity of it wash over you for a second. It can be helpful to get another pair of eyes on your data when you reach this point. Someone who isn't quite so deep into the nitty gritty of it might be able to see it with fresh eyes, and give you a refreshing reminder of what you're doing and why you're doing it.

Conclusion

Data can be meaningful and misleading. It's both real and fake. It's collected about real things that happen using a metric that we've created to communicate with each other. When done properly, a lot of powerful conclusions can be made. But you gotta stay safe out there.