Part 2.1 | Data Cleaning

Outline

we've developed a systematic method for summarizing data

and we've developed the skills to do this summarization in software

we've done this with nice looking data that i've prepared for you

but often you'll not start with such nice data

remember that survey I gave you at the beginning of the semester? take a look at some of the dates

they are in a consistent format, but when we put them into a visualization package, it gives us an unusable figure

here we'd actually want years instead of birthdays

we can do this conversion very easily in both Excel and python, just extracting the year from this date variable

this is a common problem with dates: we have a consistent format but it's not exactly the format we want

and remember when I asked you how far away from Pittsburgh is your hometown?

that's in even worse shape

i just gave you a box to type into, so there is no consisten format

so we get a very messy and unusable figure

there are a couple of tricks here

the first is to recognize that many answers look numerical but are actually strings

this means the computer doesn't recognize them as numbers without telling it to convert to numbers

all we have to do is tell the computer to convert this variable to numbers

and if it's not easily recognizable as a number we'll tell it to leave it blank

so we get some nice numbers and the more complex ones are left blank

so what do we do with the blank ones?

well we have a few options: either we go through them and try to fill them in manually, fill them in as zero or the average or something, or we drop them from the dataset

each option has their pros and cons, but we generally stay away from filling them in

these are some of the basic data cleaning steps you might encounter

but there are many other kinds

you'll see a couple more on your homework

but lets do some work with this survey data in our exercise

Excel Exercise 2.1

We can extract year with the `=LEFT()` function, which takes the date cell `A2` and gets the leftmost 4 characters.

```
=LEFT(A2,4)
```

We can also convert the distances column into a numerical column using the `=VALUE()` function.

```
=VALUE(D2)
```

This will give us numbers where possible. But we'll get a bunch of errors. This is where we have to deal with the missing numerical values. One approach is to simply enter the values in. We want to be careful to keep a record of what we're doing. So instead of replacing the original values, we'll simply override the function that's returning an error. Or better yet, create a new column with the final distances.