

# MS&E 125: Intro to Applied Statistics

## Data Munging

Professor Udell

Management Science and Engineering  
Stanford

April 10, 2023

## Announcements

- ▶ hw 1 due Friday
- ▶ quiz 1 Friday
- ▶ convert colab to pdf

# Outline

Messy data

SQL

## Data types

- ▶ continuous values (e.g., 4.2,  $\pi$ )
- ▶ discrete values (e.g., 0, 4, 994)
- ▶ nominal values (e.g., apple, banana, pear)
- ▶ ordinal values (e.g., rarely, sometimes, often)
- ▶ graphs or networks (e.g., person 1 is friends with person 2)
- ▶ text (e.g., doctor's note describing symptoms)
- ▶ sets (e.g., items purchased)

## Messy data

- ▶ heterogeneous: values of many different types
- ▶ missing: some values are missing, inconsistent, not recorded, or lost
- ▶ noise: some (or all) values suffer errors, inaccuracies, or malicious corruption
- ▶ duplicated values

## Data cleaning

- ▶ remove duplicates
- ▶ remove missing values
- ▶ remove noise
- ▶ convert to a single type (usually, numeric)

how? by taking a careful look...

## Demo

`https://colab.research.google.com/github/  
stanford-mse-125/demos/blob/main/fires.ipynb`

# Outline

Messy data

SQL



# SQL

- ▶ most data is stored in relational databases
- ▶ Structured Query Language (SQL) is a language for querying relational databases
- ▶ we will use pandas in python, not SQL
- ▶ but if you know the ideas, you can easily write SQL queries

“ChatGPT, please write an SQL query to find the average age of all users in the database.”

## SQL-style munging

- ▶ select rows
- ▶ select columns
- ▶ on condition
- ▶ sort
- ▶ group (aggregate using a function)
- ▶ join (combine tables)

## Demo

`https://colab.research.google.com/github/  
stanford-mse-125/demos/blob/main/join.ipynb`