# EXAM 1

JOEY GEBIN
SEC 01

## DATA SCIENCE WORKFLOW

1) Data collection
2) Data Preperation
3) Data Vizualization
4) Data Analysis
5) Data Storytelling

## UNIVARIATE VIZ.

Benefits of viz
- understanding what we are working w/
  - → scales, typp. outcomes
  - → outliers
  - → patterns + relationships
- refine research questions
- communicate findings + tell a story

★ GGplot: "grammar of graphics" (tidyverse)

Comp. of graphics
→ frame (coordinate system, layer, scales, facetting, theme

## BIVARIATE VIZ.

Bivariate viz explores:
- relationship trends
- relationship strength
- outliers in relationship

Variable Roles
- response variable — variable whose variability we want to explain
- predictors — variables that might explain variability

Building Bivariate Plots
- ea. quantatative variable needs a new axis
- ea. categ. variable needs new "group"
- viz. that overlaps needs facetting + transparency

## MULTIVARIATE VIZ.

Components of a plot
- setting up a frame → adding layers → splitting plot into facets for dif. groups → change the theme → scales
  - ↳ changes color, fill, size, shape, or other prop. of new variable

## SPATIAL VIZ.

TYPES:
- Point maps — plotting locations of indiv. observations
- contour maps — plotting density or distrib. of observations
- choropleth maps — plotting outcomes in dif. regions

→ can be static or dynamic/interactive

## EFFECTIVE VIZ

COMPONENTS:
- professionalism (axis labels + caption)
- accessibility (alt. text + color palette)
- Design details
- Ethics (visibility, privacy, power, emotion, embodiement etc.)

## WRANGLING

ARRANGE, FILTER, SELECT, MUTATE, SUMMARIZE, GROUP_BY

## RESHAPING

TYPES:
- aggregate data — group_by + summarize (gains info but loses data on indiv. observations)
- raw data — retains all info but reshaped to perform task

EX: Pivot_longer — collapses several columns into two (lengthen)
Pivot_wider — spread out values accross new variables (widen)

## JOINING

Mutating Joins
- left_join — keeps all L disc. all R
- inner_join — keeps all L w/ match on R
- full_join — keeps all L + R

Filtering Joins
- semi_join — disc. ob. L table if no match in R table
- anti_join — disc. ob in L table if they have match in R table