WELCOME TO NDACAN MONTHLY OFFICE HOURS!

NATIONAL DATA ARCHIVE ON CHILD ABUSE AND NEGLECT
DUKE UNIVERSITY, CORNELL UNIVERSITY, & UNIVERSITY OF CALIFORNIA: SAN FRANCISCO

NDACAN

Children's Bureau
An Office of the Administration for Children & Families

- The session will begin at 11am EST
  - 11:00 - 11:30am – LeaRn with NDACAN (Introduction to R)
  - 11:30 - 12:00pm – Office hours breakout sessions
- Please submit LeaRn questions to the Q&A box
- This session is being recorded.
- See ZOOM Help Center for connection issues: https://support.zoom.us/hc/en-us
  - If issues persist and solutions cannot be found through Zoom, contact Andres Arroyo at aa17@cornell.edu.

LEARN WITH NDACAN

Presented by Frank Edwards

MATERIALS FOR THIS COURSE

- Course Box folder (https://cornell.box.com/v/LeaRn-with-R-NDACAN-2024-2025) contains
  - Data (will be released as used in the lessons)
    - Census state-level data, 2015-2019
    - AFCARS state-aggregate data, 2015-2019
    - AFCARS (FAKE) individual-level data, 2016-2019
    - NYTD (FAKE) individual-level data, 2017 Cohort
  - Documentation/codebooks for the provided datasets
  - Slides used in each week's lesson
  - Exercises as that correspond to each week's lesson
  - An .R file that will have example, usable R code for each lesson – will be updated and appended with code from each lesson

WEEK 6: DATA VISUALIZATION

March 21, 2025

DATA USED IN THIS WEEK'S EXAMPLE CODE

- AFCARS fake individual level data ./Data/afcars_2018_indv_fake.csv
  - Simulated foster care data following the AFCARS structure
  - Can order full data from NDACAN:
    - https://www.ndacan.acf.hhs.gov/datasets/request-dataset.cfm

BASIC ANATOMY OF A PLOT IN R

GGPLOT2

- ggplot2 uses the following basic ingredients for a plot
  - 1) Data, 2) aesthetic mappings, 3) graphics to draw

This takes the form in syntax of

```
ggplot(DATA,
       aes(x = VARIABLE)) +
       geom_histogram()
```

COMMON AESTHETIC PARAMETERS

- For univariate visuals, we will generally only use aes(x = VARIABLENAME)
- For bivaraites
  - continuous: use aes(x = VAR1, y = VAR2)
  - Continuous + categorical: aes(x = VAR1, color = VAR2)
- Can also use shape, size, color (for lines), fill (for solid fills)
- For continuous ranges, try xmin, xmax, ymin, ymax.
- Group is also useful: aes(x = VAR1, y = VAR2, group = VAR3)

COMMON GEOMS

- Here are my most commonly used geoms
- Histogram: geom_histogram()
- Density: geom_density()
- Scatterplot: geom_point()
- Line plot: geom_line()
- Bar plot: geom_col() or geom_bar()
- Maps: geom_sf()

OVER TO RSTUDIO

R CODE, PAGE 1 OF 8

```
### leaRn week 6
### data visualization with ggplot2

library(tidyverse)

### read data
afcars_ind<-read_csv("./data/afcars_2018_indv_fake.csv")
# take a look
head(afcars_ind)

## Univariate visuals
# histogram of age at first entry
ggplot(afcars_ind,
       aes(x = agefirstrem_f)) +
  geom_histogram()
# density of age at first entry
ggplot(afcars_ind,
       aes(x = agefirstrem_f)) +
  geom_density()
# distribution of race/ethnicity
ggplot(afcars_ind,
       aes(x = raceth_f)) +
  geom_bar()
# weird, oh because it is numeric
ggplot(afcars_ind,
       aes(x = factor(raceth_f))) +
  geom_bar()
```

R CODE, PAGE 2 OF 8

```
## Bivariate continuous / categorical
# age at first entry by child sex
ggplot(afcars_ind,
       aes(x = agefirstrem_f,
           color = factor(sex_f))) +
  geom_density()
## thats a bit difficult because of overlap.
# let's try small multiples with facet_
# use facet_grid when you want to fix the number of rows or columns
# facet_wrap is more generic
ggplot(afcars_ind,
       aes(x = agefirstrem_f)) +
  geom_density() +
  facet_grid(~sex_f)
# density of age at first entry by child race/ethnicity
ggplot(afcars_ind,
       aes(x = agefirstrem_f,
           color = factor(raceth_f))) +
  geom_density()
# And let's also look at sex
ggplot(afcars_ind,
       aes(x = agefirstrem_f,
           color = factor(raceth_f))) +
  geom_density() +
  facet_wrap(~sex_f)

### Two continuous measures
# let's look at the joint distribution of age at first and last rem
ggplot(afcars_ind,
       aes(x = agefirstrem_f,
           y = ageatlatrem_f)) +
  geom_point()
```

R CODE, PAGE 3 OF 8

```
##### ok those 99s are missing, let's remove them
ggplot(afcars_ind %>%
         filter(agefirstrem_f<25,
                ageatlatrem_f<25),
       aes(x = agefirstrem_f,
           y = ageatlatrem_f)) +
  geom_point()
### this doesn't do a good job of showing the density of data
## at each point because age is an integer
# Let's add some random noise to each observation with a jitter
ggplot(afcars_ind %>%
         filter(agefirstrem_f<25,
                ageatlatrem_f<25),
       aes(x = agefirstrem_f,
           y = ageatlatrem_f)) +
  geom_point(position = position_jitter())

# better! let's make the points a little transparent
# alpha does the trick here, 0 is transparent, 1 is opaque
ggplot(afcars_ind %>%
         filter(agefirstrem_f<25,
                ageatlatrem_f<25),
       aes(x = agefirstrem_f,
           y = ageatlatrem_f)) +
  geom_point(position = position_jitter(),
             alpha = 0.25)
```

R CODE, PAGE 4 OF 8

```
# not bad! Let's provide useful axis labels
ggplot(afcars_ind %>%
         filter(agefirstrem_f<25,
                ageatlatrem_f<25),
       aes(x = agefirstrem_f,
           y = ageatlatrem_f)) +
  geom_point(position = position_jitter(),
             alpha = 0.25) +
  labs(x = "Age at first removal",
       y = "Age at last removal")

# and what's with that grey background, I don't like it
# we can swap to a different theme easily
ggplot(afcars_ind %>%
         filter(agefirstrem_f<25,
                ageatlatrem_f<25),
       aes(x = agefirstrem_f,
           y = ageatlatrem_f)) +
  geom_point(position = position_jitter(),
             alpha = 0.25) +
  labs(x = "Age at first removal",
       y = "Age at last removal") +
  theme_minimal()
```

R CODE, PAGE 5 OF 8

```
## Ok cool! Is this pattern the same for all groups?
ggplot(afcars_ind %>%
         filter(agefirstrem_f<25,
                ageatlatrem_f<25),
       aes(x = agefirstrem_f,
           y = ageatlatrem_f)) +
  geom_point(position = position_jitter(),
             alpha = 0.25) +
  facet_wrap(~raceth_f) +
  labs(x = "Age at first removal",
       y = "Age at last removal")

# and are there differences by child sex?
ggplot(afcars_ind %>%
         filter(agefirstrem_f<25,
                ageatlatrem_f<25),
       aes(x = agefirstrem_f,
           y = ageatlatrem_f,
           color = sex_f)) +
  geom_point(position = position_jitter(),
             alpha = 0.25) +
  facet_wrap(~raceth_f) +
  labs(x = "Age at first removal",
       y = "Age at last removal")
```

R CODE, PAGE 6 OF 8

```
### oops sex is binary, force it to a factor
ggplot(afcars_ind %>%
         filter(agefirstrem_f<25,
                ageatlatrem_f<25),
       aes(x = agefirstrem_f,
           y = ageatlatrem_f,
           color = factor(sex_f))) +
  geom_point(position = position_jitter(),
             alpha = 0.25) +
  facet_wrap(~raceth_f) +
  labs(x = "Age at first removal",
       y = "Age at last removal")
## and clean up the legend a bit
ggplot(afcars_ind %>%
         filter(agefirstrem_f<25,
                ageatlatrem_f<25)),
       aes(x = agefirstrem_f,
           y = ageatlatrem_f,
           color = factor(sex_f))) +
  geom_point(position = position_jitter(),
             alpha = 0.25) +
  facet_wrap(~raceth_f) +
  labs(x = "Age at first removal",
       y = "Age at last removal",
       color = "Sex")
```

R CODE, PAGE 7 OF 8

```
# this doesn't really tell us what we want to know though

# one more, placement setting by sex and age
ggplot(afcars_ind,
       aes(x = factor(curplset_f))) +
  geom_bar()

# let's make sex color and keep placement setting as x
ggplot(afcars_ind,
       aes(x = factor(curplset_f),
           color = factor(sex_f))) +
  geom_bar()
## oops we want fill
ggplot(afcars_ind,
       aes(x = factor(curplset_f),
           fill = factor(sex_f))) +
  geom_bar()
## and I want to see the bars side by side, not stacked
ggplot(afcars_ind,
       aes(x = factor(curplset_f),
           fill = factor(sex_f))) +
  geom_bar(position = position_dodge())
# ok now let's add age at last removal
ggplot(afcars_ind,
       aes(x = factor(curplset_f),
           fill = factor(sex_f))) +
  geom_bar(position = position_dodge()) +
  facet_wrap(~ageatlatrem_f)
```

R CODE, PAGE 8 OF 8

```
# the y axis makes this tough - many more 1 year olds than 15 year olds
# we can let the y axis vary for each facet
ggplot(afcars_ind,
       aes(x = factor(curplset_f),
           fill = factor(sex_f))) +
  geom_bar(position = position_dodge()) +
  facet_wrap(~ageatlatrem_f, scales = "free_y")

# and get it ready for presentation
ggplot(afcars_ind,
       aes(x = factor(curplset_f),
           fill = factor(sex_f))) +
  geom_bar(position = position_dodge()) +
  facet_wrap(~ageatlatrem_f, scales = "free_y") +
  labs(y = "Number of children",
       x = "Placement setting",
       fill = "Child sex",
       title = "Foster care placement settings for 2018",
       subtitle = "by child age (panels) and sex (color)") +
  theme_bw()
```