WELCOME TO THE 2023 NDACAN SUMMER TRAINING SERIES!

- The session will begin at 12pm EST.
- Please submit questions to the Q&A box.
- This session is being recorded.

NDACAN SUMMER TRAINING SERIES

National Data Archive on Child Abuse and Neglect

Cornell University & Duke University

NATIONAL DATA ARCHIVE ON CHILD ABUSE AND NEGLECT

NDACAN SUMMER TRAINING SERIES SCHEDULE 2023

July 5 — Introduction to NDACAN and the Administrative Data Series

July 12 — New Data Acquisition: CCOULD Data

July 19 — Causal Inference Using Administrative Data

July 26 — Evaluating and Dealing with Missing Data in R

August 2 — Time Series Analysis in Stata

August 9 — Data Visualization in R

SESSION AGENDA

- Topics in data visualization
- Making visualizations in R with ggplot2

TOPICS IN DATA VISUALIZATION

WHY VISUALIZE DATA

- Visualize raw data to uncover patterns and gain better understanding, such as trends or outliers
- Display results from modeling or estimation
- Figures can help assess model fit
- More palatable, memorable, and usually easier to compare trends than tables

EFFECTIVE VISUALS

- Who is the intended audience?
- What are you trying to convey?
- What do you want to highlight in your visualizations?
- Don't overcomplicate or make too 'busy'
- Add informative titles and labels (axes, legends, variables) so the figure can stand on its own
- Concise and clear legends
- Use the appropriate figure for your data, e.g. bar charts for percentages or categorical variables, density plots or scatterplots for continuous data

CONSIDERATIONS

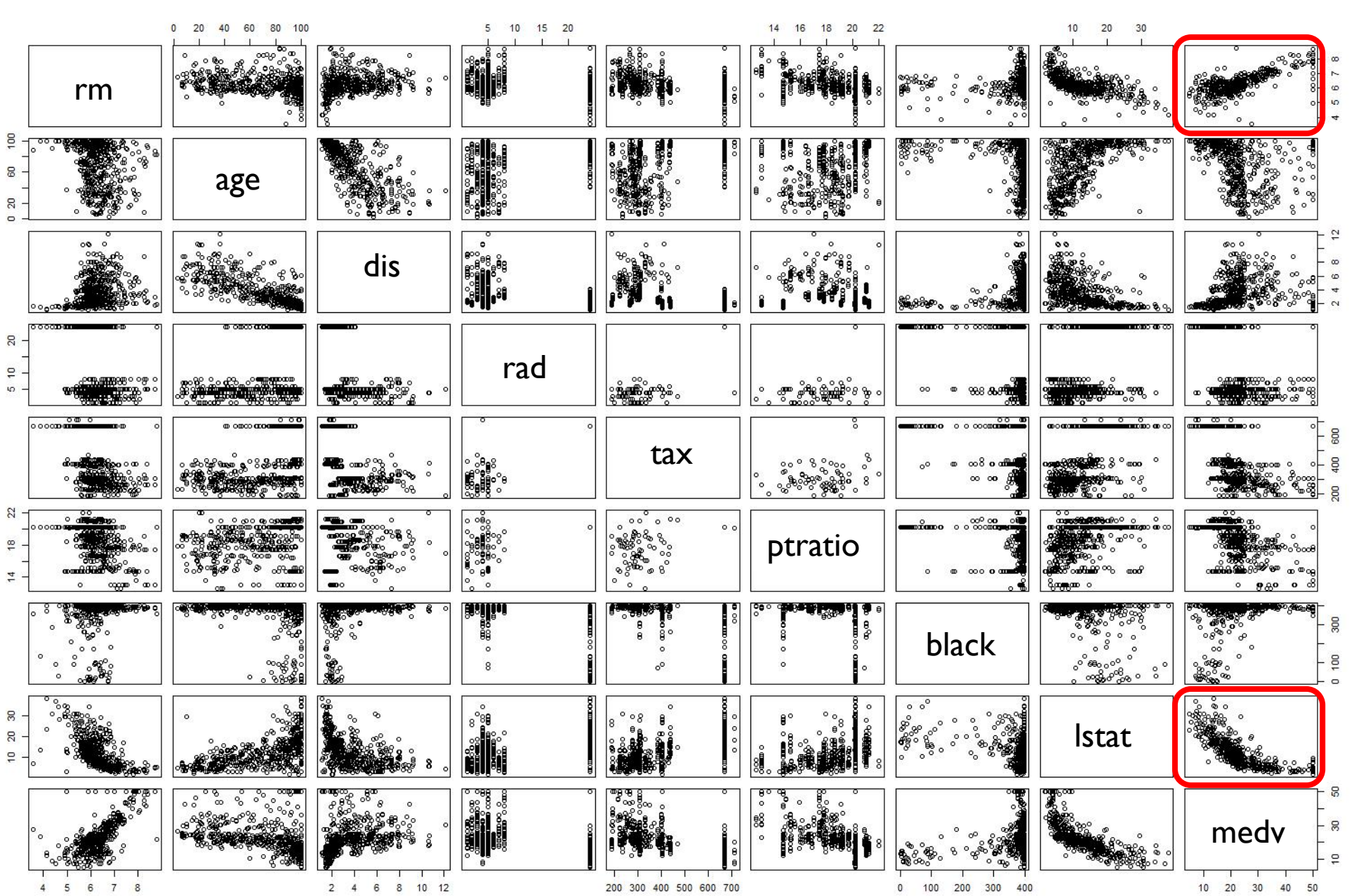
- Color blind accessibility
- Colors may look different on different computer screens
- Aesthetics is important but subjective
- What type of graph is appropriate for your data

DIFFERENT GRAPHS FOR DIFFERENT DATA

- Density plots – continuous data
- Scatter plots – continuous data best (could use ordinal)
- Bar plots – ordinal data
  - Stacked bar plots
- Heat maps – ordinal and continuous, 3 dimensions

CAUTIONS

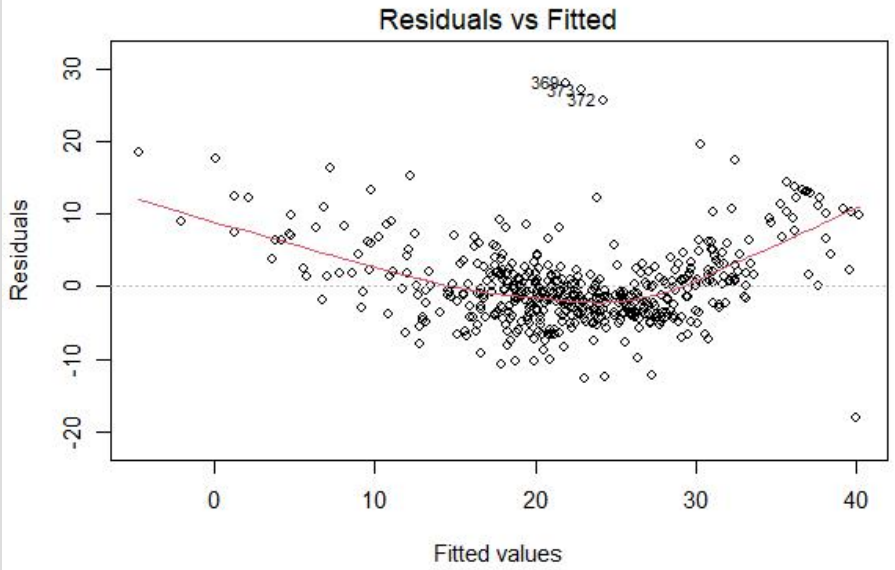
- Axes scales
- Aspect ratio, e.g. width and height of final figure
- What estimate is being shown, e.g. rates vs counts
  - Is it misleading?
  - Is it appropriate?
  - Is it the most effective at telling your story?
- Misleading information or presentation

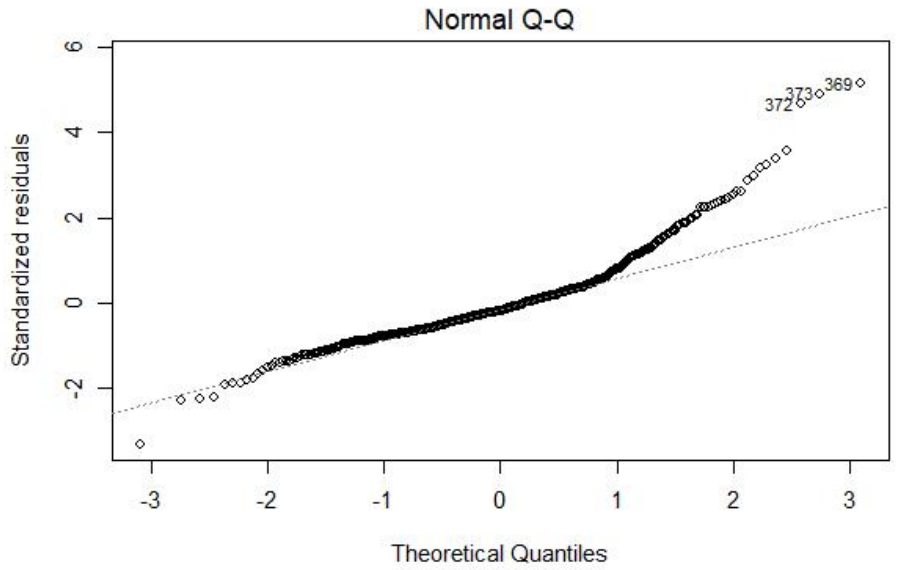


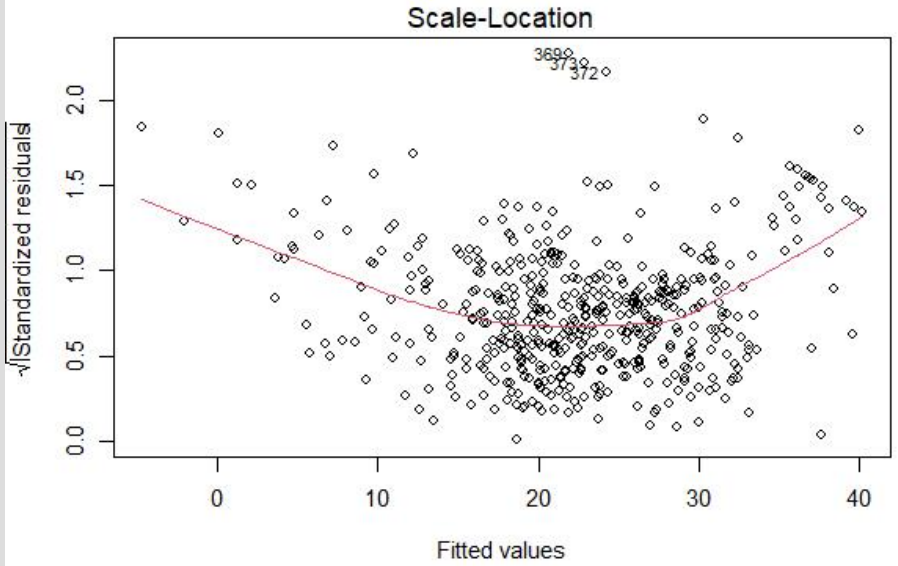
Example of scatter plots of the data

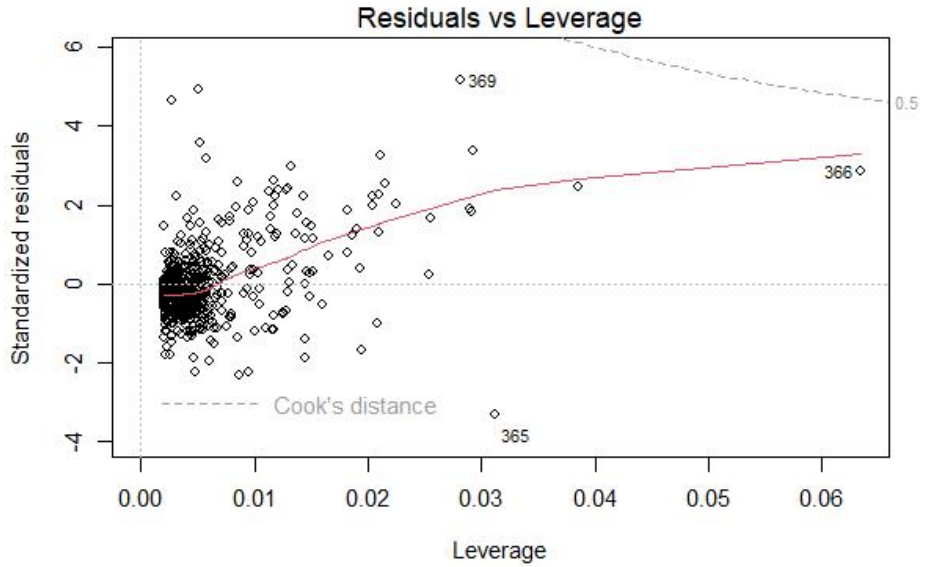
```
# load data from
# MASS package
library(MASS)
data("Boston")

# see data codebook
help(Boston)

# make figure
pairs(Boston[,-(1:5)])
```

Example of residuals and diagnostics plots










```
# fit linear model
# medv is the response
# lstat and rm are
# significant covariates
m1 = lm(medv ~ lstat + rm,
        data = Boston)

# make 2x2 figure
par(mfrow = c(2,2))

# plot model results
plot(m1)
```

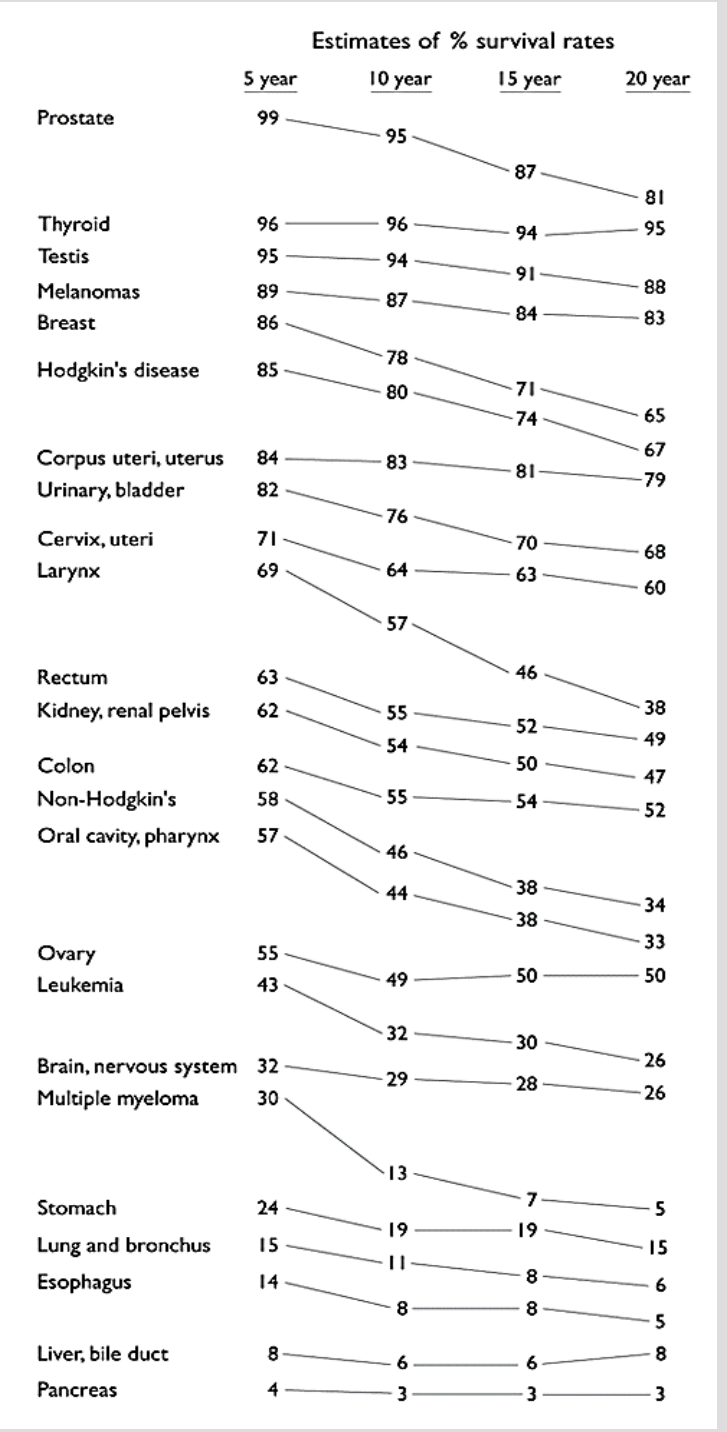
Estimates of relative survival rates, by cancer site

	% survival rates and their standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8

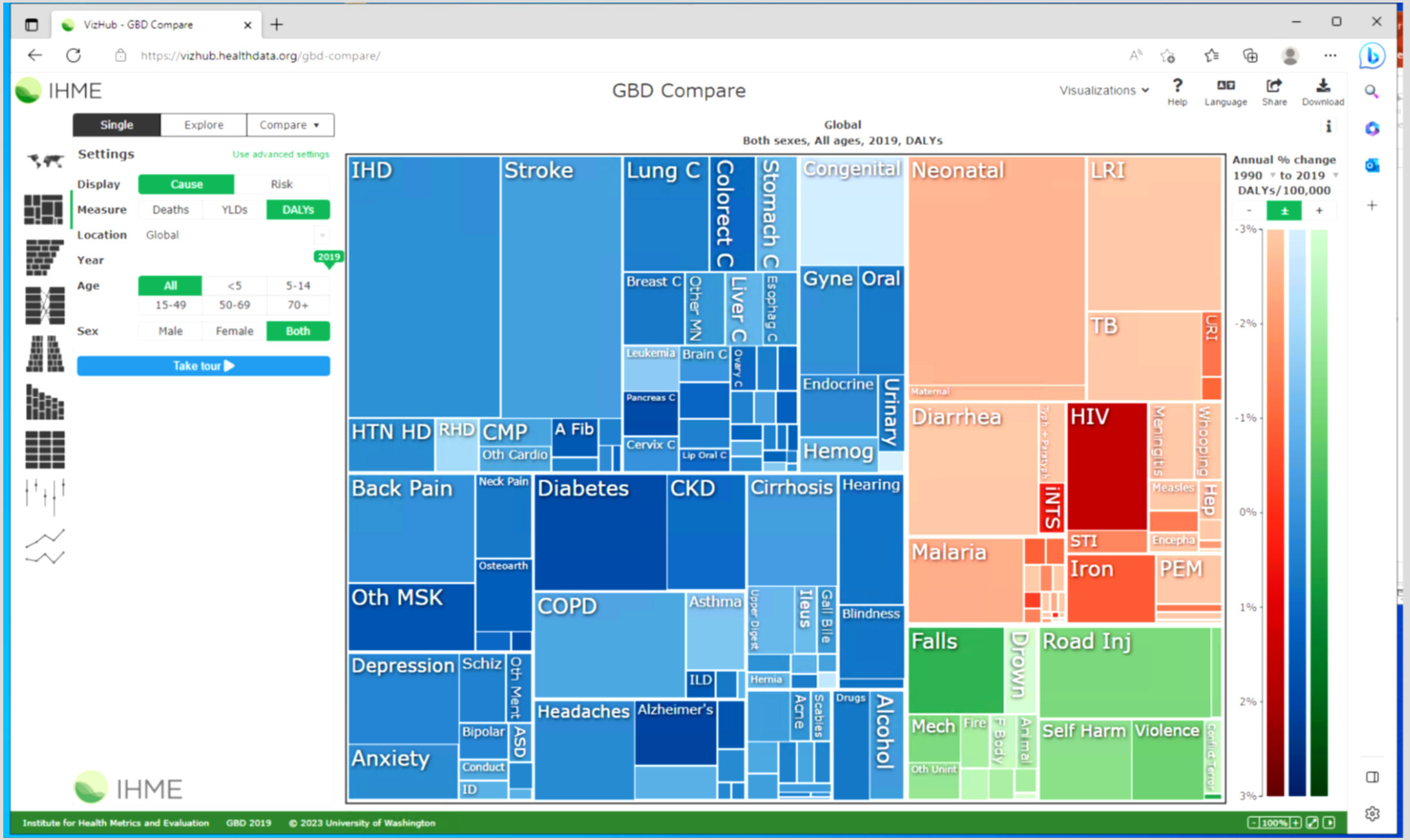
Source: Edward Tufte, *Beautiful Evidence,* pp. 174, 176.
Obtained from https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0003nk

Example of an effective and simple visual

Table → Slope Graph



Example of confusing and overwhelming visual

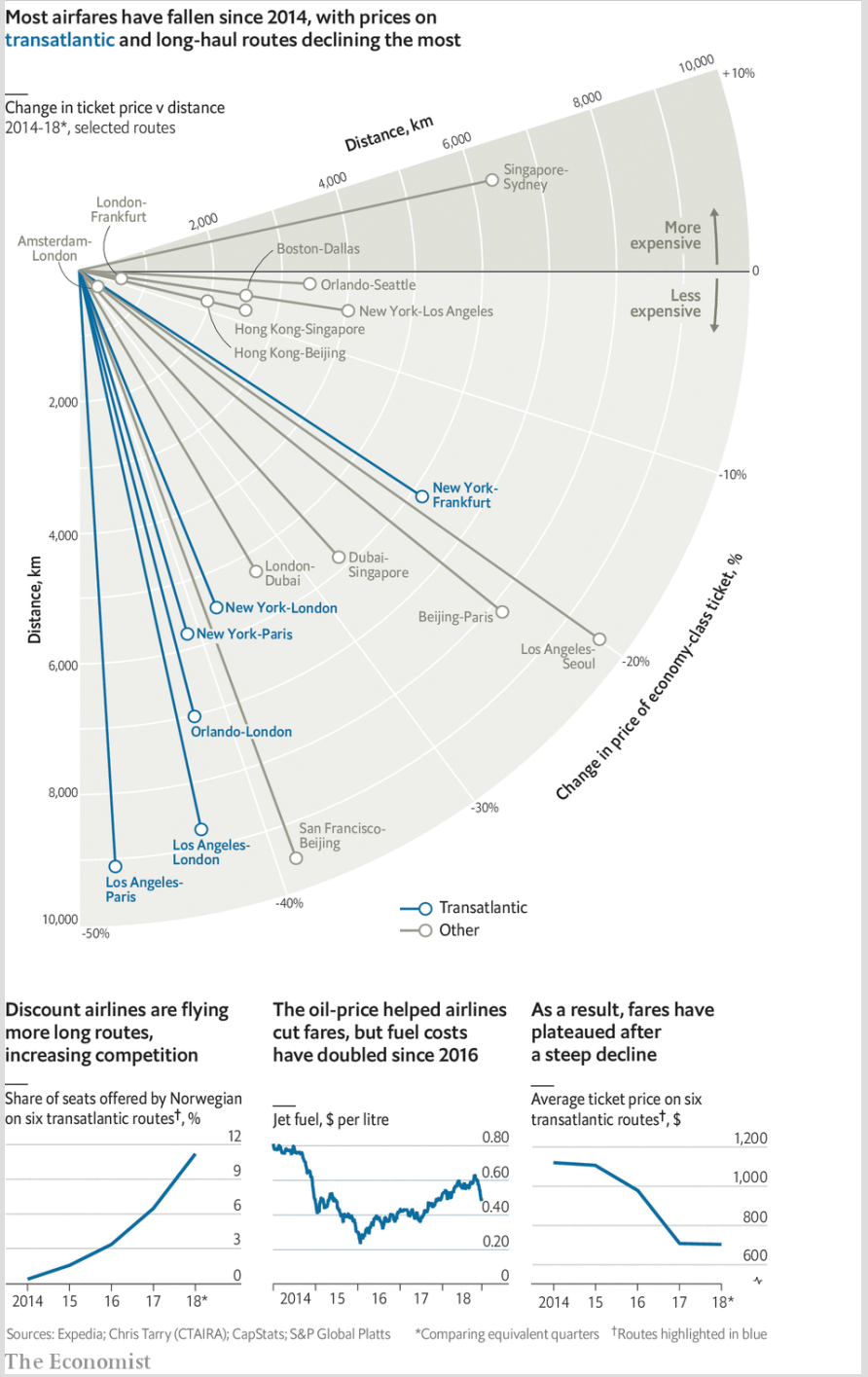


Source: https://vizhub.healthdata.org/gbd-compare/

Example of confusing and misleading visual

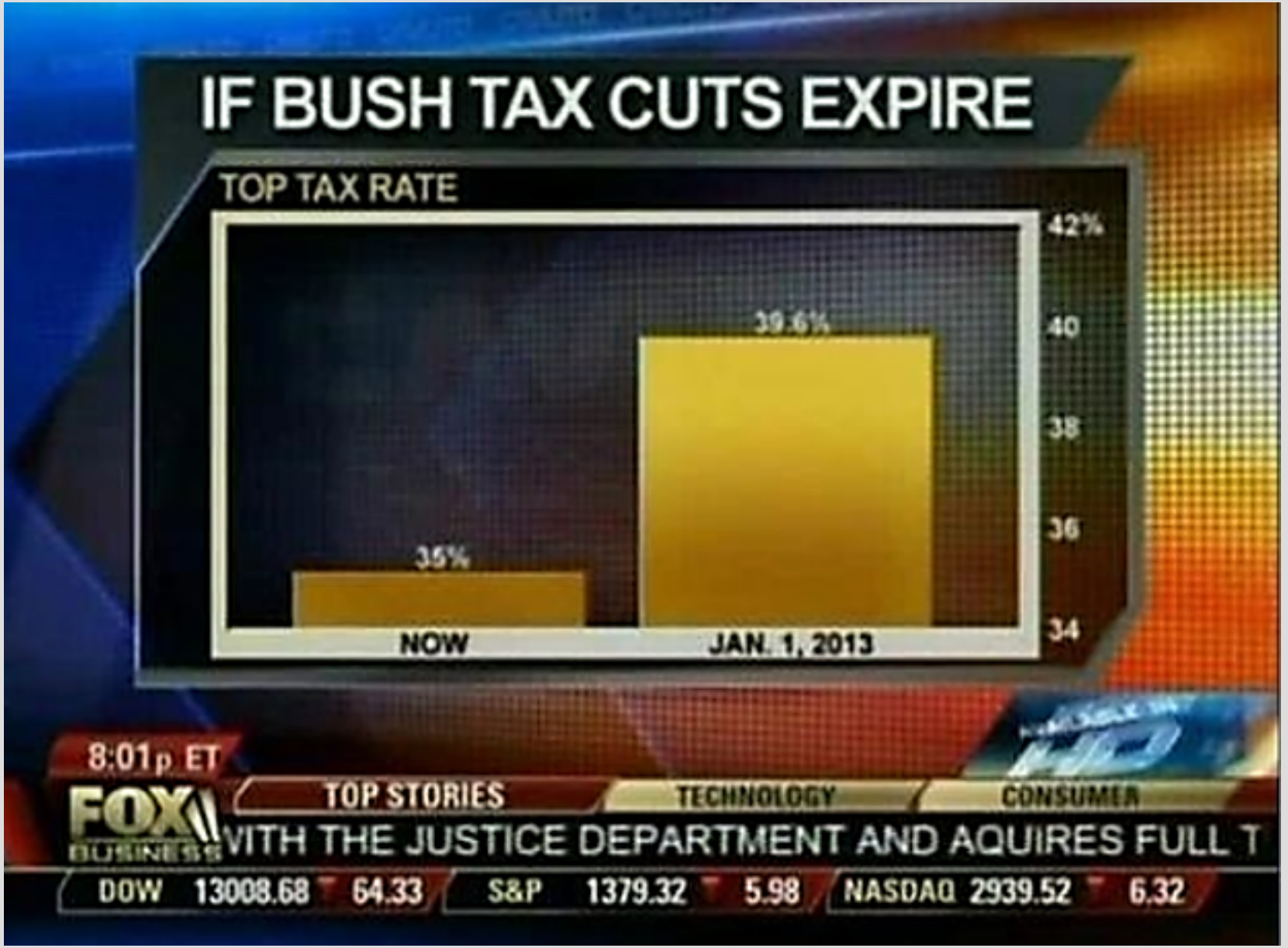
Source: https://www.cnbc.com/2019/06/12/tesla-looks-like-netflix-did-in-2011and-it-may-see-a-similar-recovery.html

Example of busy and confusing visual



Source: https://www.economist.com/graphic-detail/2018/12/08/why-ticket-prices-on-long-haul-flights-have-plummeted?__twitter_impression=true

Example of misleading axes



Obtained from: https://www.kdnuggets.com/2012/12/taking-misleading-statistics-to-a-new-level.html

MAKING VISUALIZATIONS IN R

MAKING FIGURES IN R

- The package ggplot2 (grammar of graphics) within the tidyverse universe produces beautiful figures
- Lots of customization possible
- Able to construct complex plots
- Works by constructing figures with ‘layers’

USING GGPLOT

- Figures made with ggplots are built by layering functions

```
ggplot(DATA, aes(x = XVAR, y = YVAR) )
```

- This would not plot anything! Need a layer for points, and/or lines

USING GGPLOT

- Figures made with ggplots are built by layering functions

```
ggplot(DATA, aes(x = XVAR, y = YVAR) ) +
    geom_point()
```

- This would make a basic scatter plot of XVAR by YVAR from dataset DATA with just points

USING GGPLOT

- Figures made with ggplots are built by layering functions

```
ggplot(DATA, aes(x = XVAR, y = YVAR) ) +
    geom_point() +
    geom_line()
```

- This would make a basic scatter plot of XVAR by YVAR from dataset DATA with lines connecting the points

EXAMPLE IN R

- Using NCANDS data (National Child Abuse National Data System, e.g. data on child maltreatment reports)
  - Number of unsubstantiated and substantiated reports by year (2010-2020), state, race/ethnicity, sex
- Linked with Census population data to get respective populations, e.g. for using population as the denominator for rates
  - Population by year, state, race/ethnicity, sex

FOLLOW THE EXAMPLE

- NCANDS data are available to individuals with an IRB, view the list of years from the NDACAN website: www.ndacan.acf.hhs.gov/datasets/datasets-list-ncands-child-file.cfm and follow the "Order dataset" link on the right side of the details page of the dataset of interest
- Similar state level-data are publicly available, though, in excel format from the Child Maltreatment reports in a few years: 2010, 2012, 2013. A list of all Child Maltreatment reports can be found here: www.acf.hhs.gov/cb/data-research/child-maltreatment
- Census population data can be found here: www.census.gov/programs-surveys/popest/data/data-sets.All.List_1725564412.html#list-tab-List_1725564412
- Alternative source for population data: https://seer.cancer.gov/popdata/download.html

A SMALL HANDFUL OF USEFUL GRAPHING FUNCTIONS

- Plot multiple graphs into one figure
  - `cowplot::plot_grid()`
  - `ggpubr::ggarrange()`
- Add facets to ggplot, add in as layer
  - `… + facet_grid()`
  - `… + facet_wrap()`
- Grpahs by state, add in as layer
  - `… + geofacet::facet_geo()`
  - `… + usmap::plot_usmap()`
  - `… + coord_map()`
- Interaction plot
  - `interaction.plot()`
- Check model fit
  - `plot(m1)` where `m1` would be the fitted model output, from `lm()` for example

REFERENCES AND RESOURCES

Practical & coding

- Kieran Healy. *Data Visualization: A Practical Introduction*
- Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*
- Winston Chang. *R Graphics Cookbook*

Theoretical

- Edward Tufte. *The Visual Display of Quantitative Information*
- Edward Tufte. *Envisioning Information*

QUESTIONS?

SARAH SERNAKER
STATISTICIAN

SARAH.SERNAKER@DUKE.EDU

R CODE PAGE 1 OF 8

```
library(data.table)    # package for reading the data
library(tidyverse)
library(ggplot2)
library(scales)        # package for label formats
library(geofacet)      # package for state graphs

# set directory to folder where data are
setwd("C:/Users/ss1216/Box/NDACAN/Presentations/Summer Series 2023/S6 - Data Viz")

#### LOAD DATA ####
# load NCANDS data of number of substantiated/unsubstantiated reports
ncands = fread("CF_summerseries.csv")
head(ncands)

ncands2 = ncands  %>% # filter out PR (b/c not in census data) and counts less than 10 (for data protection)
                  filter(staterr != "PR",
                         unsubst > 10,
                         subst > 10)

# load census data
census = fread("census_pop.csv")
head(census)

## join census and ncands
# left join because some states not reported in ncands from 2010-2012
dat  = ncands2 %>% # rename variables to link with census
                   rename(year = subyr,
                          st = staterr,
                          sex = chsex) %>%
                # left join because ncands may not have all states all years like census
                left_join(census) %>%
                # reorder variables
                dplyr::select(year, st, state, stfips, everything()) %>%
                # sort by year, state, and race
                arrange(year, stfips, raceEthn)
head(dat)
```

R CODE PAGE 2 OF 8

```
## Data cleaning
dat2 = dat %>% # add informative labels to race and sex
            mutate(raceEthn2 = case_when(raceEthn == 1 ~ "White NH",
                                         raceEthn == 2 ~ "Black NH",
                                         raceEthn == 3 ~ "Native Am NH",
                                         raceEthn %in% 4:5 ~ "AAPI NH",
                                         raceEthn == 6 ~ "Multiracial NH",
                                         raceEthn == 7 ~ "Hispanic"),
                        sex2 = ifelse (sex == 1, "Male", "Female"))
head(dat2)

#### Summarize data to national level
# totals in each year - grouped by race and sex
natdat = dat2 %>% group_by(year, raceEthn2, sex2) %>%
              summarise(unsubst = sum(unsubst,na.rm = TRUE),
                        subst = sum(subst, na.rm = TRUE),
                        pop = sum(pop,na.rm = TRUE))
head(natdat)

# total in each year - total over everyone
natdat_tot = dat2 %>% group_by(year) %>%
                summarise(unsubst = sum(unsubst,na.rm = TRUE),
                          subst = sum(subst, na.rm = TRUE),
                          pop = sum(pop,na.rm = TRUE))
head(natdat_tot)

# put natdat_tot data in long format
natdat_tot_long = natdat_tot %>% pivot_longer(cols = c(unsubst, subst),
                                              names_to = "rptoutcome ",
                                              values_to = "rpts ")

head(natdat_tot_long)

#### FIGURES #######
# basic scatter plot of substantiated reports, at national level
p = ggplot(natdat_tot_long %>% filter(rptoutcome == "subst"),
           aes(x = year, y = rpts)) +
           geom_point()
p
```

R CODE PAGE 3 OF 8

```
# add unsubstantiated data, at national level
# make the lines different color based on substantiation/outcome
p2 = ggplot(natdat_tot_long,
            aes(x = year, y = rpts, color = rptoutcome)) +
  geom_point() +
  geom_line()
p2

# take previous figure but fix labels and some reformatting
p2 +
  # change x axes lines to 2010-2020, incremented by 1 yr
  scale_x_continuous(breaks = 2010:2020) +

  # change y axes to start at 0 and go to 3,000,000, incremented by 500,000 - formatted with commas
  scale_y_continuous(limits = c(0,3e6),
                     breaks = seq(0,3e6, by = 5e5),
                     label = scales::comma) +

  # relabel x and y axes, and title
  xlab("Year") +
  ylab("Number children") +
  ggtitle("Number of children on reports of maltreatment, substantiated or unsubstantiated") +

  # remove the color legend title name
  labs(color = "") +

  # relabel the values of "subst" and "unsubst" respectively, need to specify 'values'/colors for each one too
  scale_color_manual(values = c("red","blue"),
                     breaks =  c("subst", "unsubst"),
                     labels = c("Substantiated", "Unsubstantiated")) +

  # put the legend horizontally on the bottom
  theme(legend.position = "bottom")
```

R CODE PAGE 4 OF 8

```
### just look at substantiated cases
p +
  geom_line() +
  scale_x_continuous(breaks = 2010:2020) +
  scale_y_continuous(label = scales::comma,
                     #limits = c(0,650000)
                     ) +
  xlab("Year") +
  ylab("Number substantiated") +
  ggtitle("Number of children on reports of substantiated maltreatment")

## look at national trends of race

# make national level data - totals by race/ethnicity
natdat_race = natdat %>% group_by(year, raceEthn2) %>%
                         summarise(unsubst = sum(unsubst, na.rm = TRUE),
                                   subst = sum(subst, na.rm = TRUE),
                                   pop = sum(pop, na.rm = TRUE))

# Plot number substantiated by race
ggplot(natdat_race, aes(x = year, y = subst, color = raceEthn2)) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = 2010:2020) +
  scale_y_continuous(label = scales::comma,
                     breaks = seq(0,300000, by = 50000)) +
  guides(color = guide_legend("Race"))+
  xlab("Year") +
  ylab("Number of substantiated cases") +
  ggtitle("Number of children on reports of substantiated reports of maltreatment")
```

R CODE PAGE 5 OF 8

```
## Create rates to standardize comparison
# national level rates of substantiated reports per 100k children - by race
natdat_race3 = natdat_race %>% mutate(subst_rate = 100000*subst/pop)

# plot substantiated rate
ggplot(natdat_race3,
       aes(x = year, y = subst_rate, color = raceEthn2)) +
  geom_point() +
  geom_line() +
  scale_x_continuous(breaks = 2010:2020) +
  scale_y_continuous(label = scales::comma,
                     limits = c(0,1700),
                     breaks = seq(0,1600, by = 400)) +
  guides(color = guide_legend("Race")) +
  xlab("Year") +
  ylab("Rate of substantiated cases (per 100k children)") +
  ggtitle("Rate of substantiated reports of maltreatment (per 100,000 children)") +
  theme(legend.position = "bottom")

### look at national trends of race and sex ######
# grouping by sex too now
natdat_race_sex = natdat %>% group_by(year, raceEthn2, sex2) %>%
                             summarise(unsubst = sum(unsubst, na.rm = TRUE),
                                       subst = sum(subst, na.rm = TRUE),
                                       pop = sum(pop, na.rm = TRUE)) %>%
                             mutate(subst_rate = 100000*subst/pop)
```

R CODE PAGE 6 OF 8

```
# facet by sex
ggplot(natdat_race_sex, aes(x = year, y = subst_rate, color = raceEthn2)) +
  geom_point() +
  geom_line() +
  facet_grid(~sex2) +
  scale_x_continuous(breaks = 2010:2020) +
  scale_y_continuous(label = scales::comma,
                     limits = c(0,1700),
                     breaks = seq(0,1600, by = 400)) +
  guides(color = guide_legend("Race"))+
  xlab("Year") +
  ylab("Rate of substantiated cases (per 100k children)") +
  ggtitle("Rate of substantiated reports of maltreatment (per 100,000 children)") +
  theme(legend.position = "bottom")

# facet by race instead
ggplot(natdat_race_sex, aes(x = year, y = subst_rate, color = sex2)) +
  geom_point() +
  geom_line() +
  # using facet_wrap now, can easily specify 2 rows and free scales between figures
  facet_wrap(~raceEthn2,nrow = 2, scales = "free") +
  scale_x_continuous(breaks = 2010:2020) +
  scale_y_continuous(label = scales::comma) +
  guides(color = guide_legend(""))+
  xlab("Year") +
  ylab("Rate of substantiated cases (per 100k children)") +
  ggtitle("Rate of substantiated reports of maltreatment (per 100,000 children)") +
  theme(legend.position = "bottom")
```

R CODE PAGE 7 OF 8

```
##### make figures by state ########
# collapse data over state
statedat = dat %>% group_by(year, st, state, stfips) %>%
                summarise(unsubst = sum(unsubst, na.rm = TRUE),
                          subst = sum(subst, na.rm = TRUE),
                          pop = sum(pop, na.rm = TRUE)) %>%
              arrange(year,stfips)


# plot substantiated by state
ggplot(statedat, aes(x = year, y = subst)) +
  geom_point() +
  geom_line() +
  facet_geo(~st, grid = "us_state_grid1"#, scales = "free_y"
            ) +
  scale_x_continuous(breaks = 2010:2020)+
  scale_y_continuous(label = scales::comma) +
  xlab("Year") +
  ylab("Number children") +
  ggtitle("Number of children on reports of maltreatment, substantiated or unsubstantiated") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

## make rates instead
statedat2 = statedat %>% mutate(subst_rate = 10000*subst/pop,
                                unsubst_rate = 10000*unsubst/pop)

# plot substantiated rate by state
ggplot(statedat2, aes(x = year, y = subst_rate)) +
  geom_point() +
  facet_geo(~st, grid = "us_state_grid1") +
  scale_x_continuous(breaks = 2010:2020) +
  xlab("Year") +
  ylab("Rate per 10k Children ") +
  ggtitle("Rate of children on reports of maltreatment, substantiated or unsubstantiated") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

R CODE PAGE 8 OF 8

```
## plot subst and unsubt by color - make long first
statedat_long = statedat2 %>% dplyr::select(year,st,state,stfips,ends_with("rate")) %>%
                              pivot_longer(cols = subst_rate:unsubst_rate,
                                           names_to = "rptoutcome",
                                           values_to = "rate")

# plot substantiated and unsubstantiated rate by state
ggplot(statedat_long, aes(x = year, y = rate, color = rptoutcome)) +
  geom_point() +
  geom_line() +
  facet_geo(~st, grid = "us_state_grid1") +
  scale_x_continuous(breaks = 2010:2020) +
  xlab("Year") +
  ylab("Rate per 10k Children ") +
  ggtitle("Rate of children on reports of maltreatment, substantiated or unsubstantiated") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(color = "") +
  scale_color_manual(values = c("red","blue"),
                     breaks =  c("subst_rate", "unsubst_rate"),
                     labels = c("Substantiated", "Unsubstantiated")) +
  theme(legend.position = "bottom",
        # edit axis text to be a little smaller and vertical
        axis.text.x = element_text(angle = 90, vjust = 0,
                                   size = 8))
```