Econ 106: Data Analysis for Economics

Lecture 10

slides adapted from:
https://jhudatascience.org/tidyversecourse/model.html#descriptive-and-exploratory-analysis

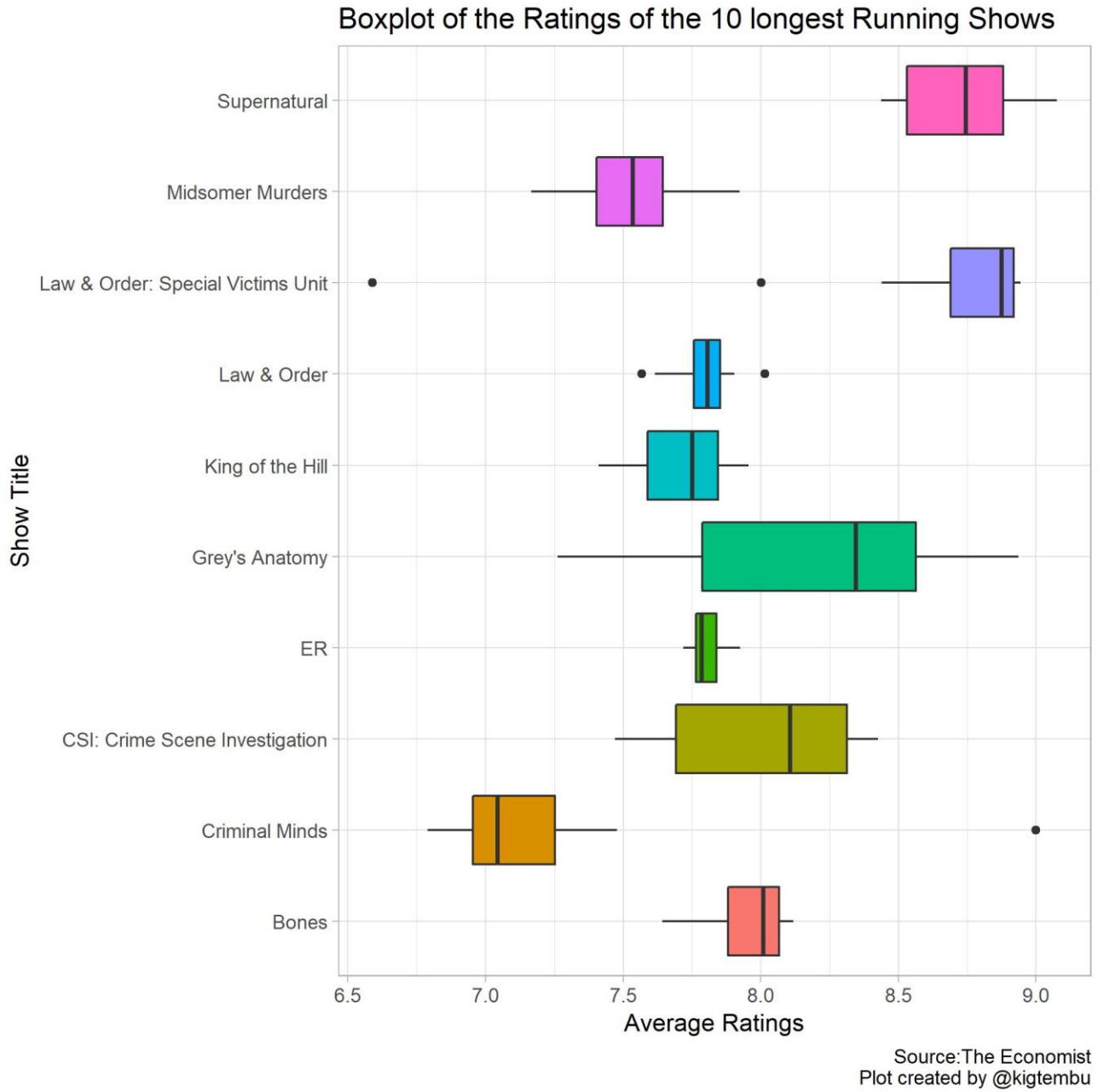
Reminder

- Research Milestone #1 is due Sunday 11:59pm

https://pollev.com/vsovero

#tidytuesday

Law and Order: SVU is the best (except for two bad seasons)



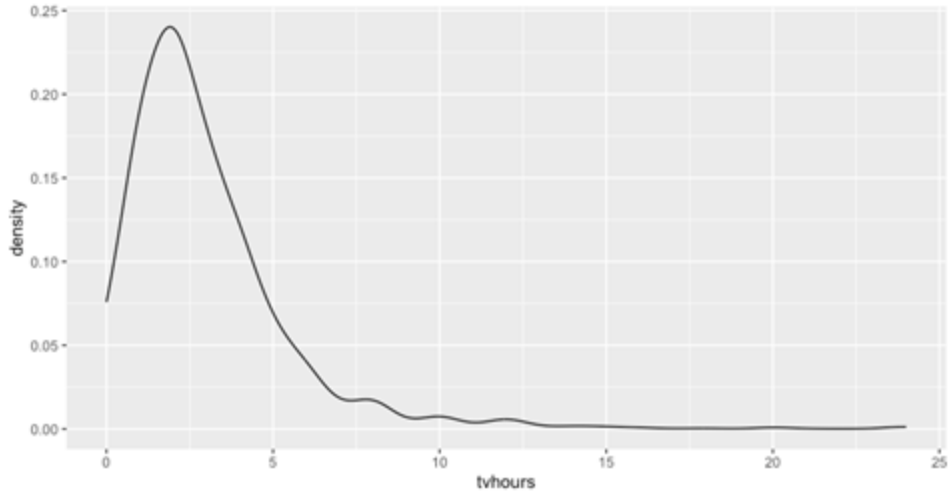
https://x.com/kigtembu/status/1082355212176814081?s=20

Descriptive and Exploratory Analysis

- The goal of a descriptive analysis is to generate simple **summaries** to **describe** the data you're working with
- The goal of an exploratory analysis is to **explore** the data and find **relationships** that weren't previously known.

Describing a Quantitative Variable

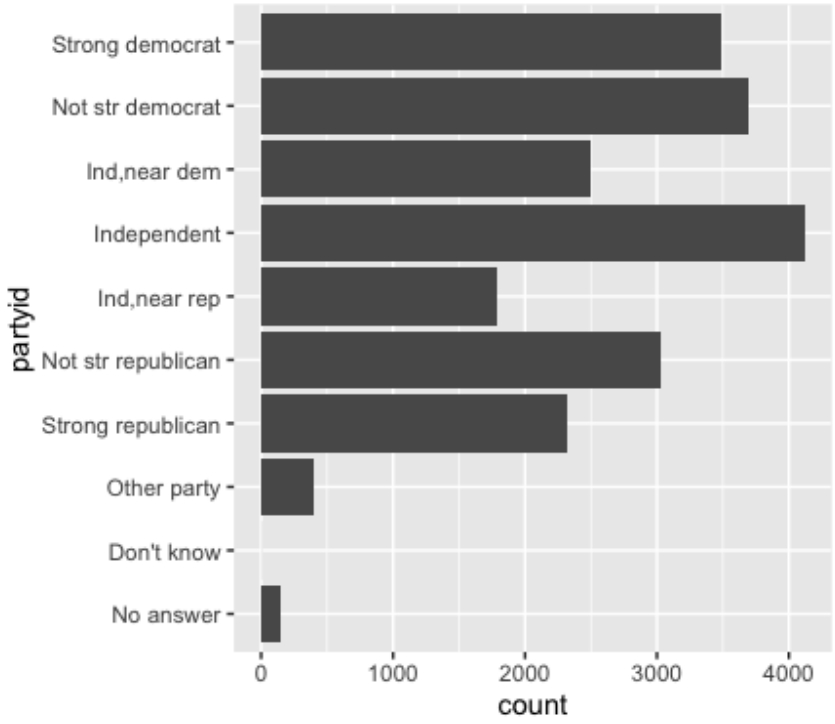
```
ggplot(data = gss_cat,
              mapping=aes(x=tvhours)) +
 geom_density(adjust=2)
```



Describing a Categorical Variable

```
ggplot(data=gss_cat,
       mapping= aes(x=partyid))+
  geom_bar() +
coord_flip()
```

- We flip the bar chart horizontally when there are a lot of levels or levels with long labels



Filter, then combine levels

Some of the partyid levels have very few cases:

- no answer
- other party
- don't know)

We're first going to filter out cases with these answers

Next, we're going to collapse the remaining levels into:

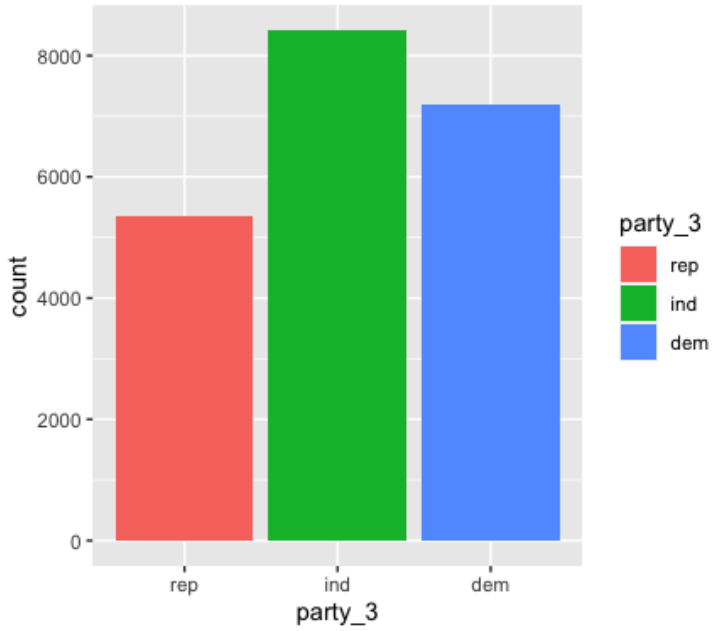
- rep
- ind
- dem

```
gss_party<-gss_cat %>%
filter(partyid!="No answer")%>%
filter(partyid!="Don't know")%>%
filter(partyid!="Other party")%>%
mutate( party_3 = fct_collapse(partyid,
                    "rep" = c("Strong republican", "Not str republican"),
                    "ind" = c("Ind,near rep", "Independent", "Ind,near dem"),
                    "dem" = c("Not str democrat", "Strong democrat")
                      )
```

That's better

- I also added color to the bar plot with the fill argument

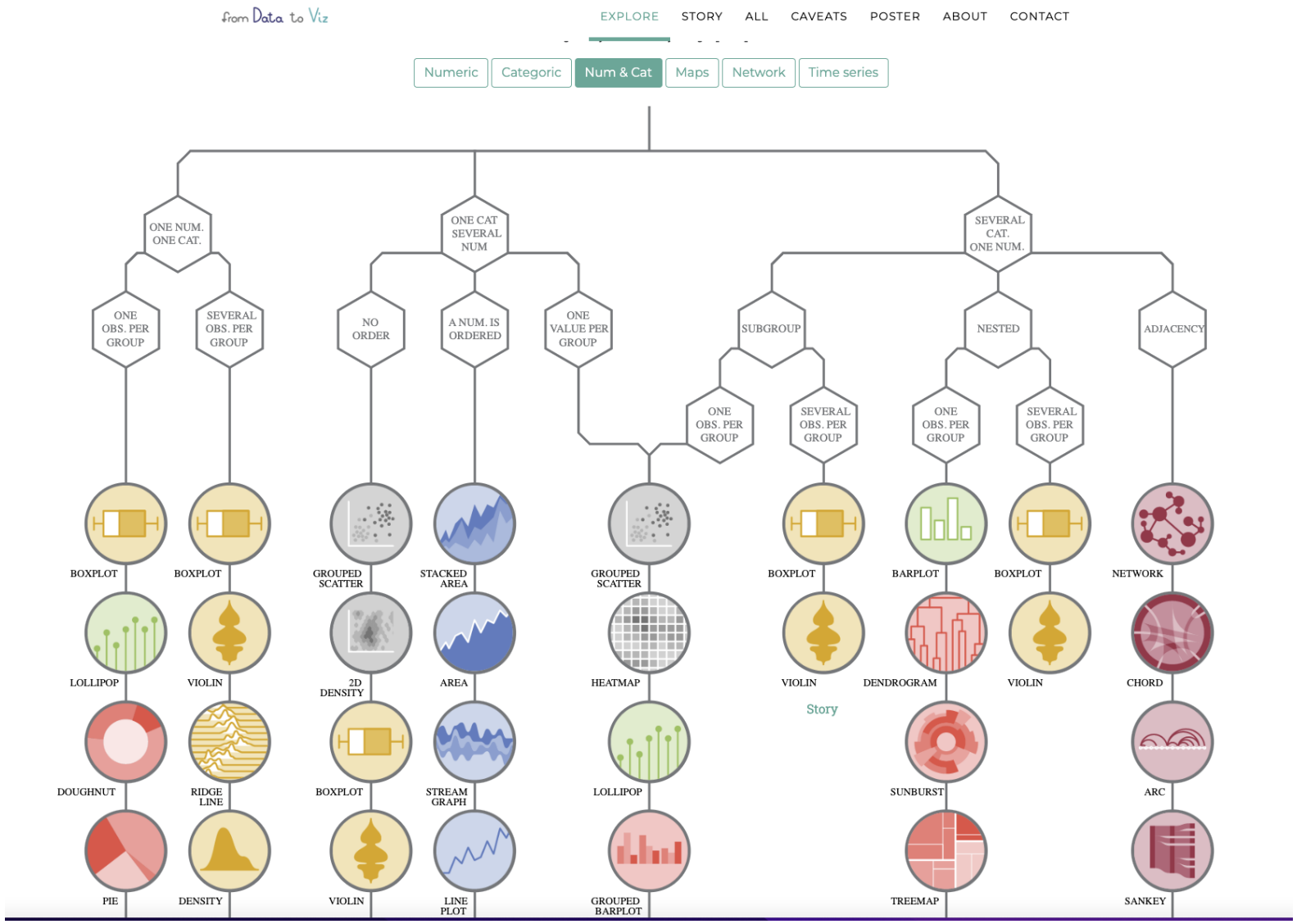
```
ggplot(data=gss_party,
       mapping= aes(x=party_3))+
 geom_bar(aes(fill=party_3))
```



Exploring Relationships Between Variables

- The best way to spot a relationship is to visualize the relationship between two or more variables.
- How you do that should depend on the type of variables involved:
  - categorical and quantitative
  - categorical and categorical
  - quantitative and quantitative

A road map to choosing the right visualization

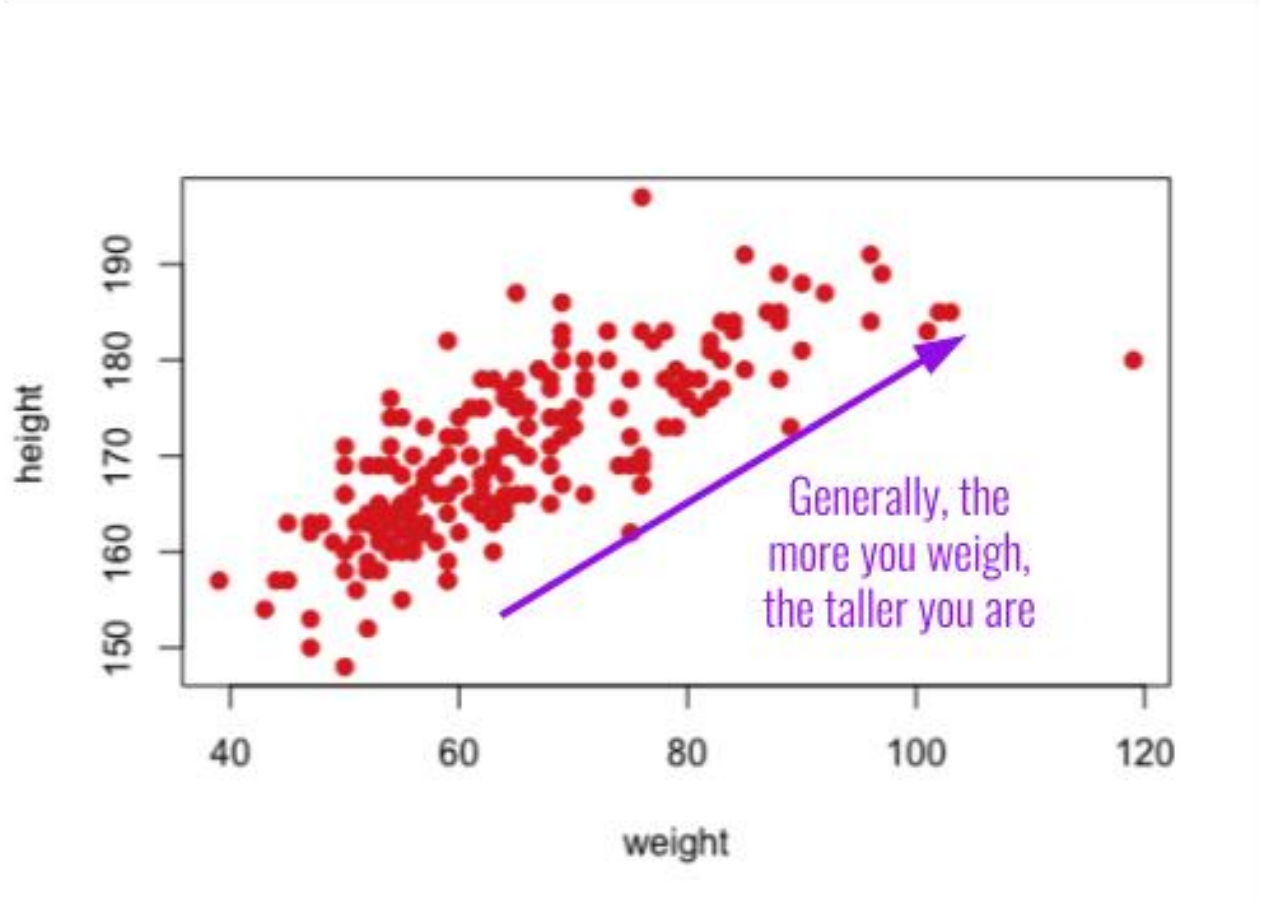


https://www.data-to-viz.com/

Visualizations for Two Quantitative Variables

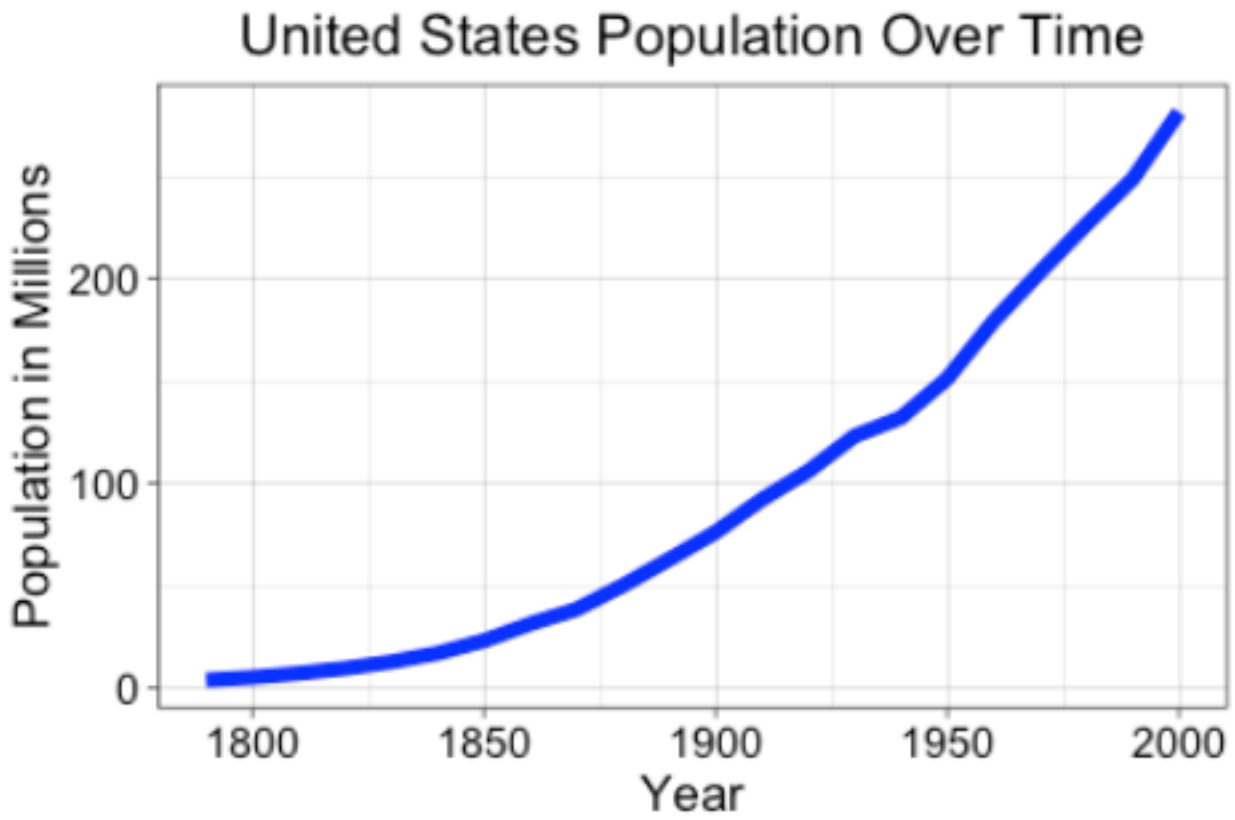
Scatterplot

Relationship between two quantitative variables



Line plot

quantitative trend over time

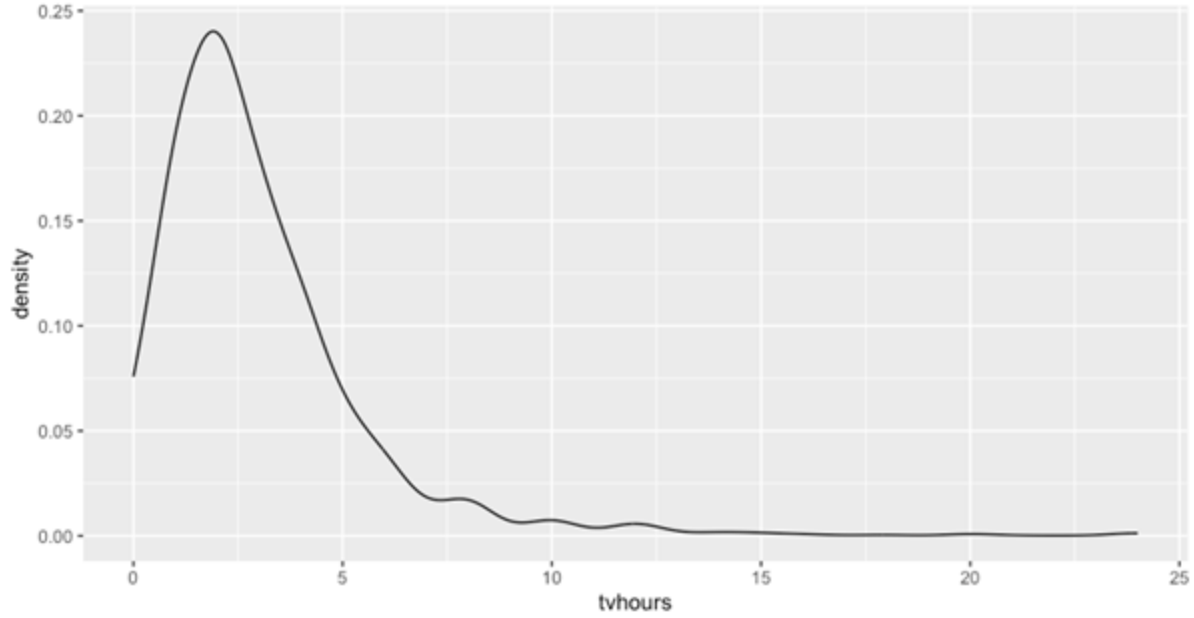


https://pollev.com/vsovero

Visualizations for a Quantitative Variable and a Categorical Variable

GSS TV Hours

- Does the distribution of TV hours look different based on political affiliation?
- Let’s find out.



Distribution of tvhours by political affiliation

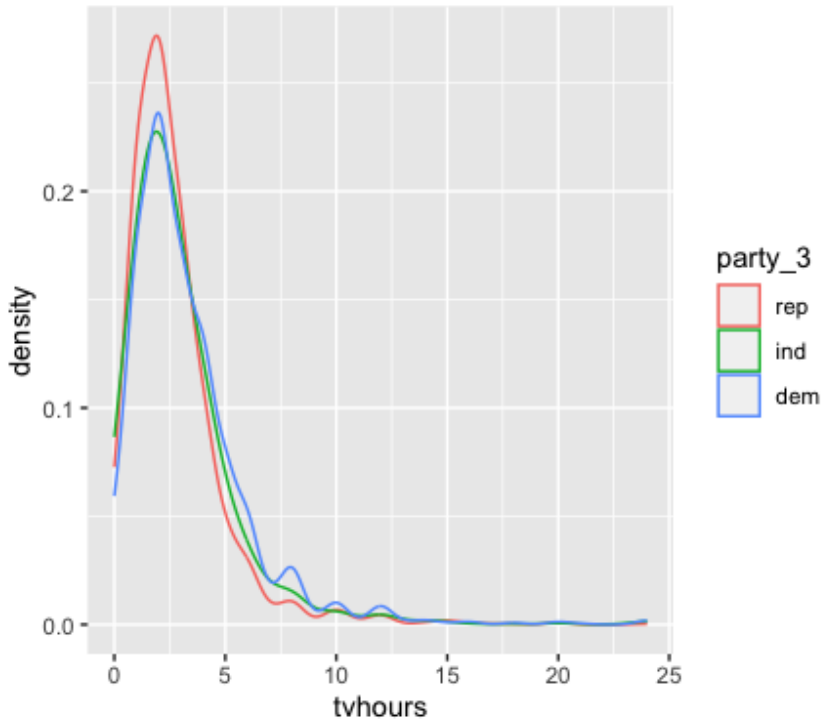
- **Color Mapping:** color will vary based on the value of party_3
- remember to wrap variable names *inside* the **aes**() function

```
ggplot(data = gss_party,
              mapping=aes(x=tvhours)) +
 geom_density(adjust=2, aes(color=party_3) )
```

Distribution of tvhours by political affiliation

- This give us a separate density plot of tvhours for each level of party_3
- Republicans watch the least amount of tv
- Similar tv watching habits between independents and democrats
- Democrats watch the most

```
ggplot(data = gss_party,
       mapping=aes(x=tvhours)) +
geom_density(adjust=2, aes(color=party_3) )
```



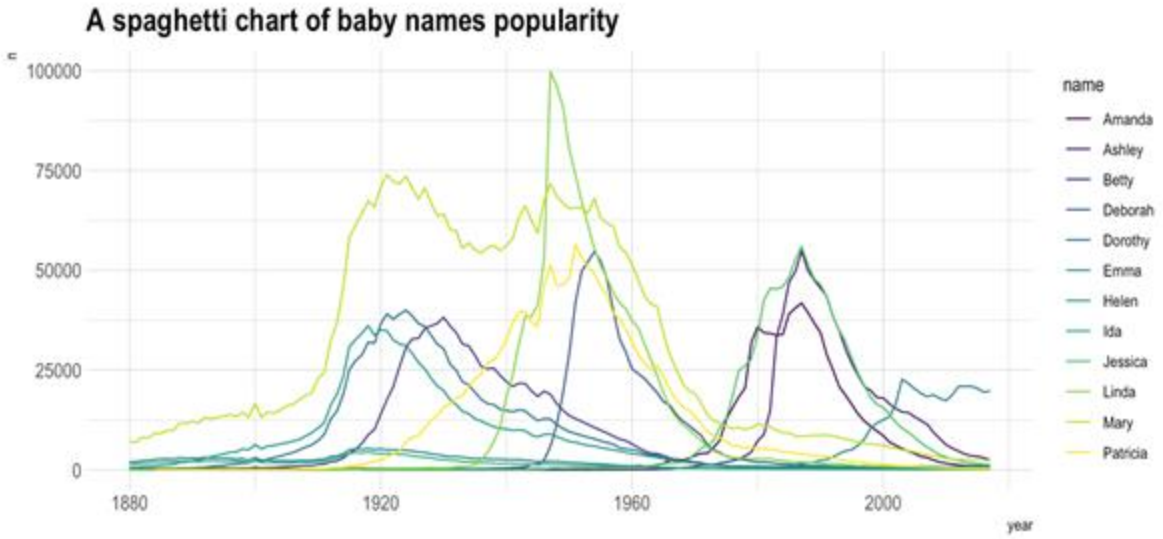
Class Exercise

- Filter out cases where marital is “No answer”
- Create a factor variable (marry_fewer) with the following levels:
  - Married
  - Previously Married
  - Never Married
- Create a density plot for age by marry_fewer

https://pollev.com/vsovero

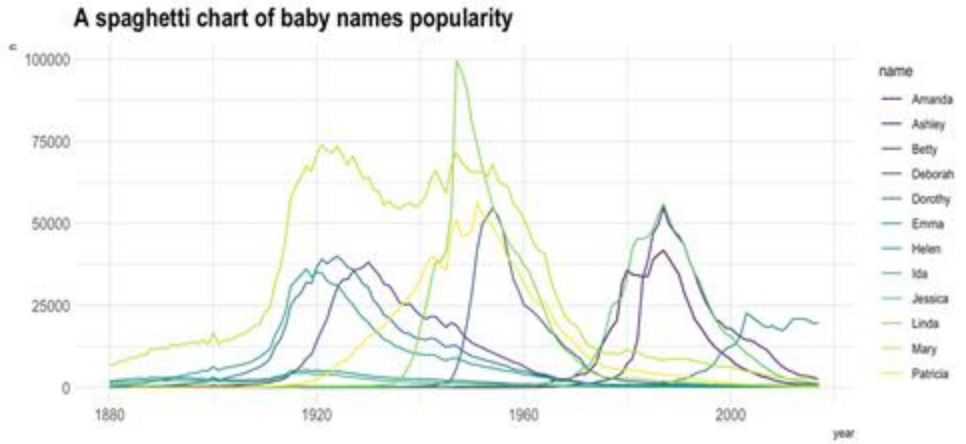
Try Multiple Graphs Instead of a Single Plot

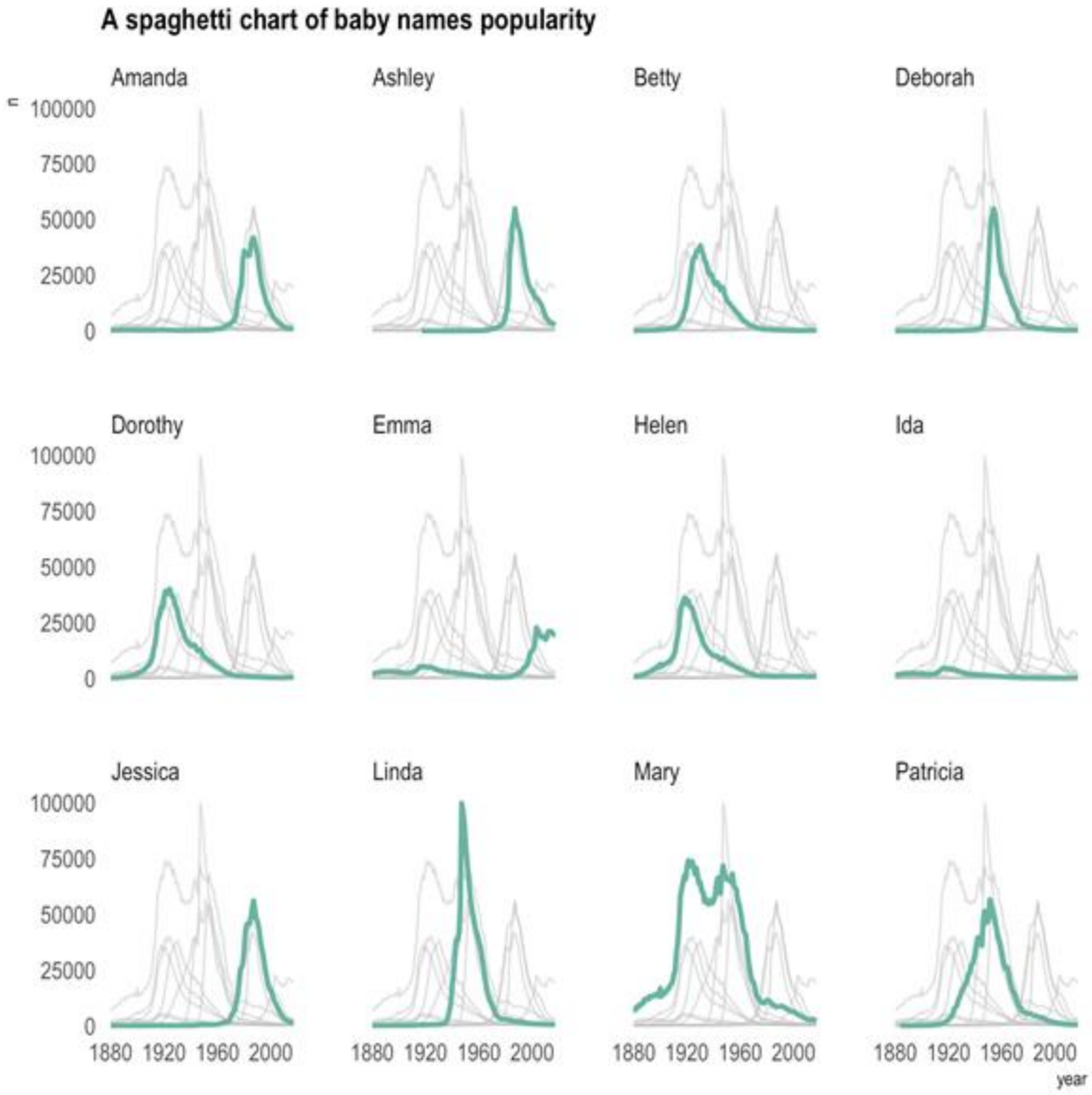
- color can help break down the information, but we can still end up with an overcluttered graph
- Instead, we can try faceting (break up the information into multiple plots)



https://www.data-to-viz.com/caveat/spaghetti.html

Faceting



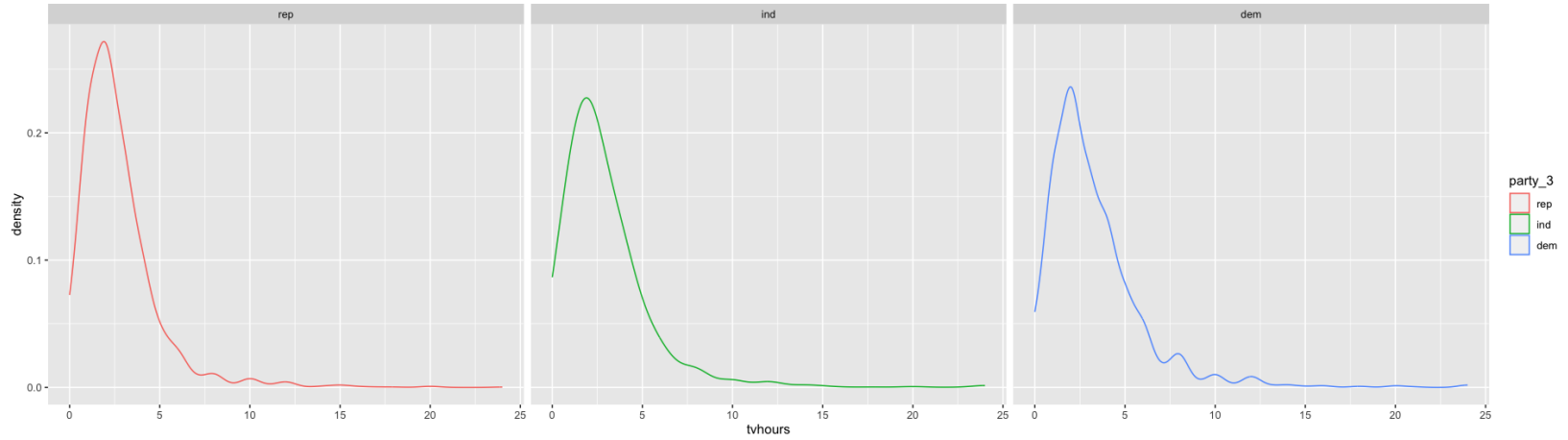


https://www.data-to-viz.com/caveat/spaghetti.html

facet_wrap()

- **Argument:** the name of the variable you want to facet on
- **Output**: separate graph for each level of your faceted variable

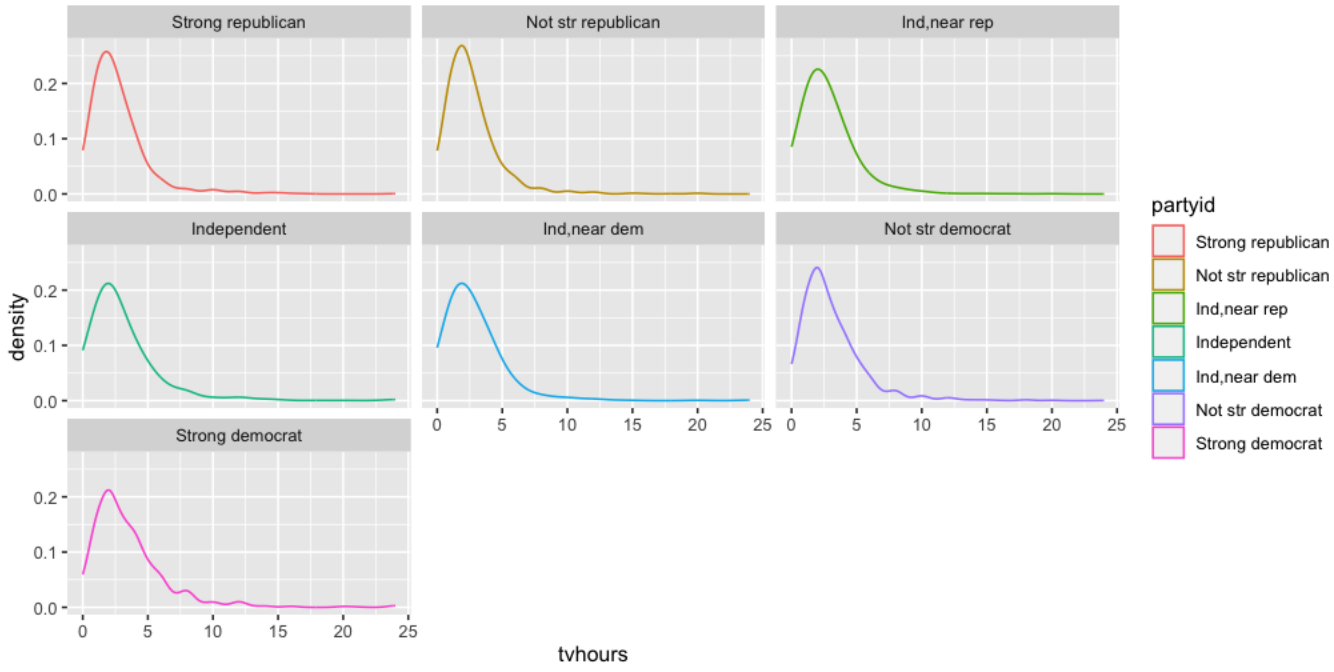
```
gss_party%>%

ggplot(data=gss_party, mapping=aes(x = tvhours)) +
   geom_density(adjust=2, aes(color=party_3)) +
   facet_wrap(~ party_3)
```



facet_wrap()

- it's called facet_wrap() because the plots will wrap to the following row as needed
- In this example, partyid has 7 levels, so the graphs will wrap over to the next row

```
ggplot(data = gss_party,
       mapping=aes(x = tvhours)) +
  geom_density(adjust=2, aes(color=partyid)) +
  facet_wrap(~ partyid)
```



facet_grid()

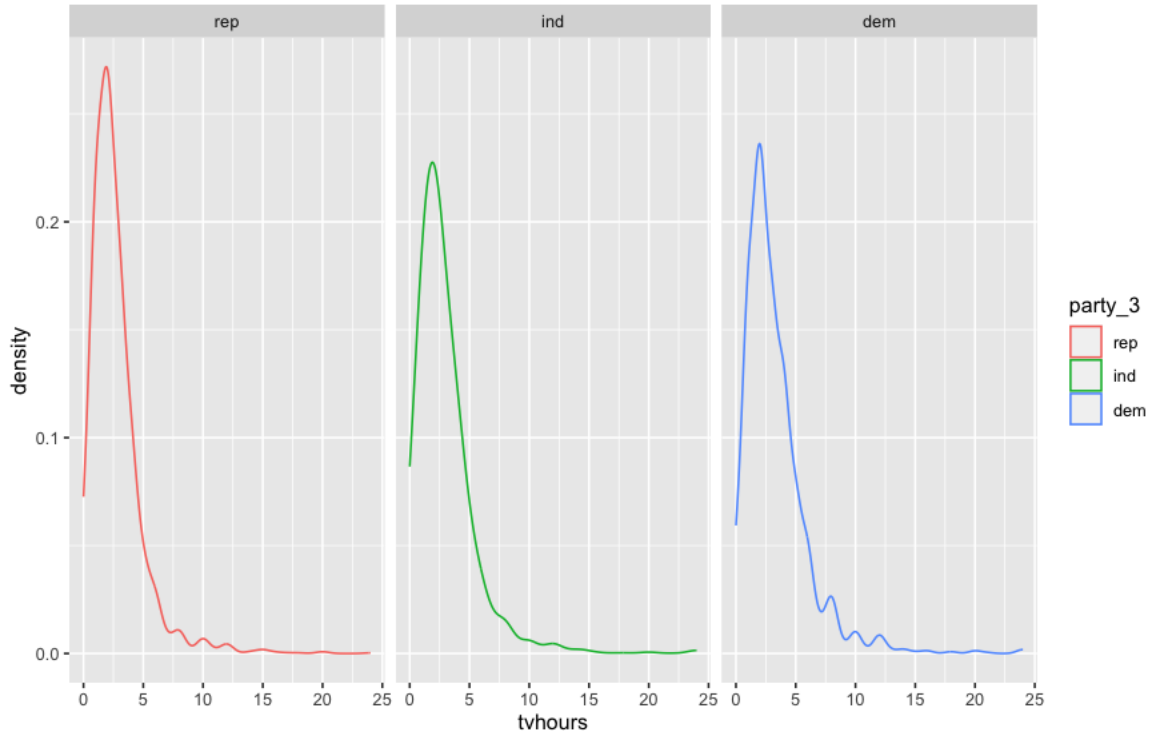
- Do you want everything in a single row? Single column?
- facet_grid() gives you this control

```
ggplot(data = gss_party,
       mapping=aes(x = tvhours)) +
  geom_density(adjust=2, aes(color=party_3)) +
  facet_grid(.~ party_3)
```

facet_grid()

- Everything in a single row

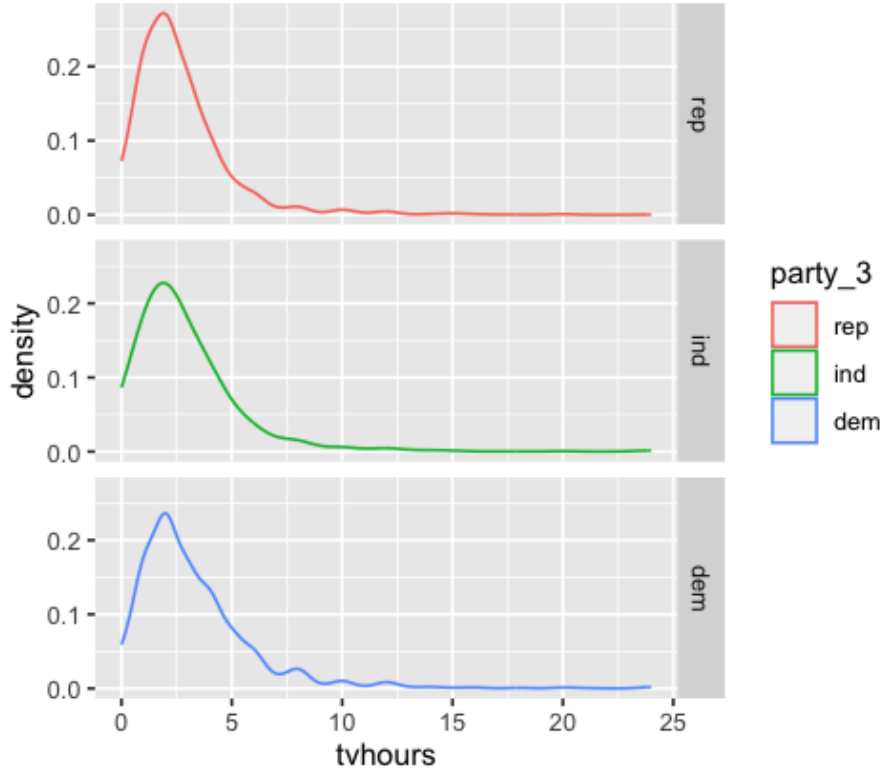
```
ggplot(data = gss_party,
       mapping=aes(x = tvhours)) +
  geom_density(adjust=2, aes(color=party_3)) +
  facet_grid(.~ party_3)
```



facet_grid()

- Everything in a single column

```
ggplot(data = gss_party,
       mapping=aes(x = tvhours)) +
  geom_density(adjust=2, aes(color=party_3)) +
  facet_grid(party_3~.)
```



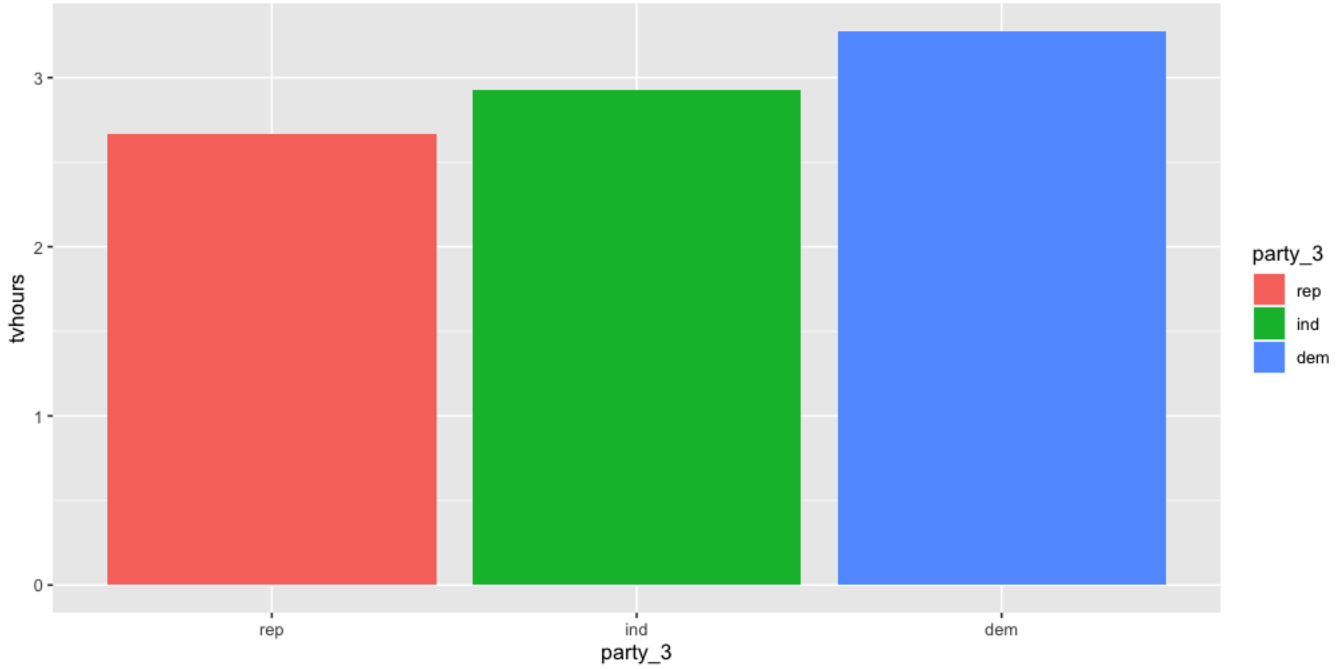
Class Exercise

Step 1. Create separate density plots for age by marry_fewer (don’t include “No answer”). Put all the plots into a single column.

https://pollev.com/vsovero

Categorical and Quantitative (Summary Stats)

- Instead of showing the entire distribution of tvhours for each level of a categorical variable, we can compare summary statistics
- In this example, we are exploring if the average amount of time spent watching tv differs by political affiliation



geom_col()

- to use geom_col(), you first need to create a summary table of mean tvhours by party_fewer
- We will also calculate the standard deviation

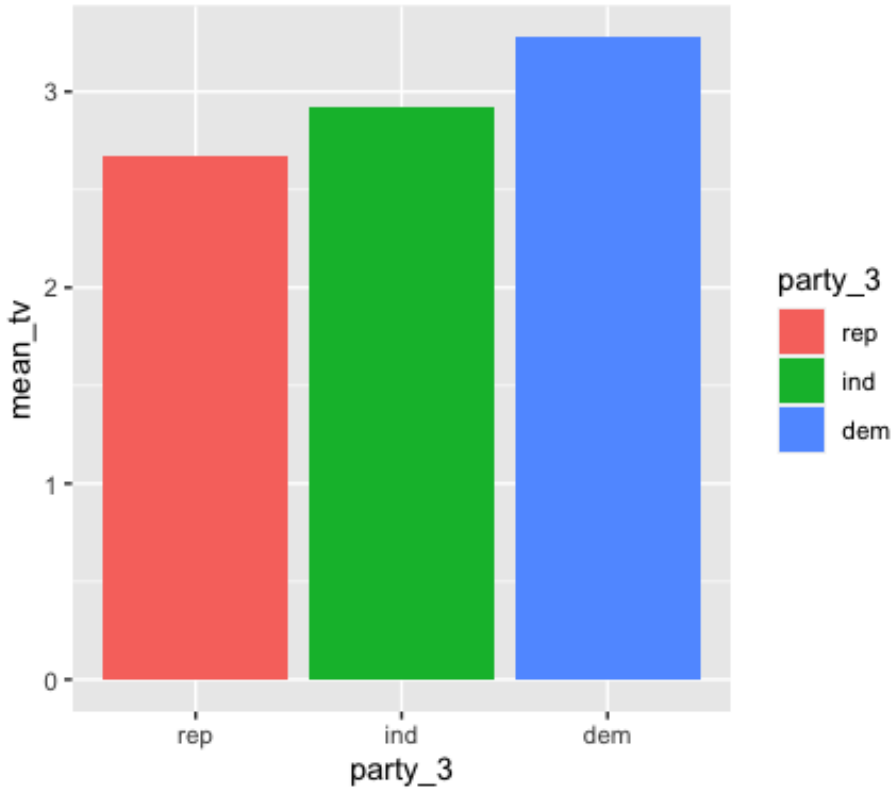
```
tv_summary<-gss_party%>%
  group_by(party_3)%>%
  summarize(mean_tv=mean(tvhours, na.rm=TRUE),
                    sd_tv=sd(tvhours, na.rm=TRUE))
```

	party_3	mean_tv	sd_tv
1	rep	2.666180	2.201986
2	ind	2.925984	2.593218
3	dem	3.275351	2.754989

geom_col()

- when using geom_col(), your y variable is mean_tv from the summary table

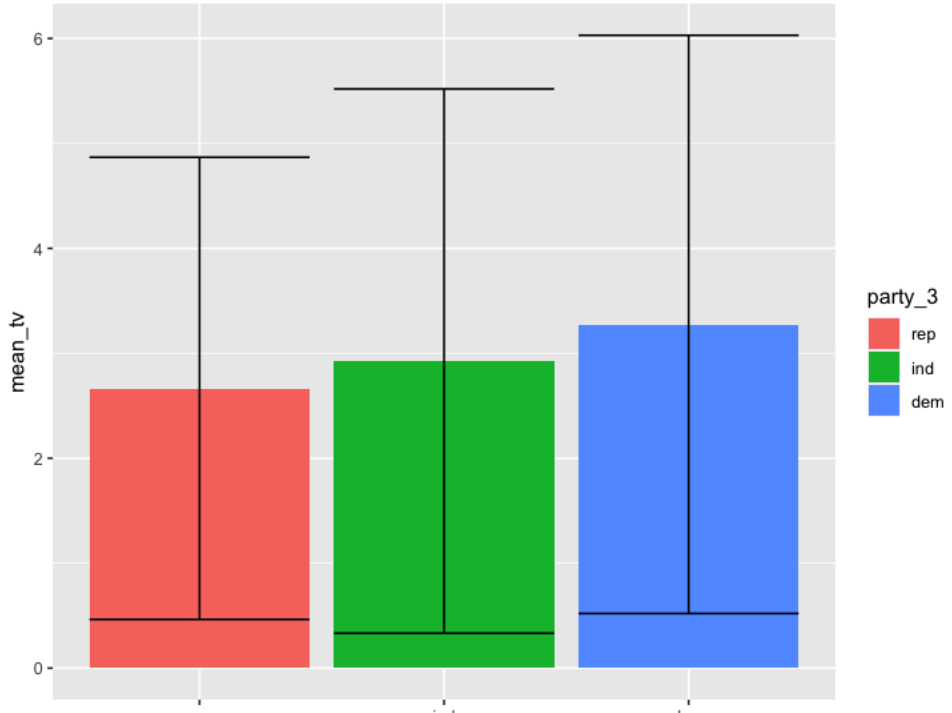
```
ggplot(data=tv_summary,
       mapping=aes(x=party_3, y=mean_tv))+
 geom_col(aes(fill=party_3))
```



Adding Standard Deviation

- We can also add error bars to represent standard deviation
- **geom_errorbar**() arguments:
  - ymin: the low point of the error bar
  - ymax: the high point of the error bar

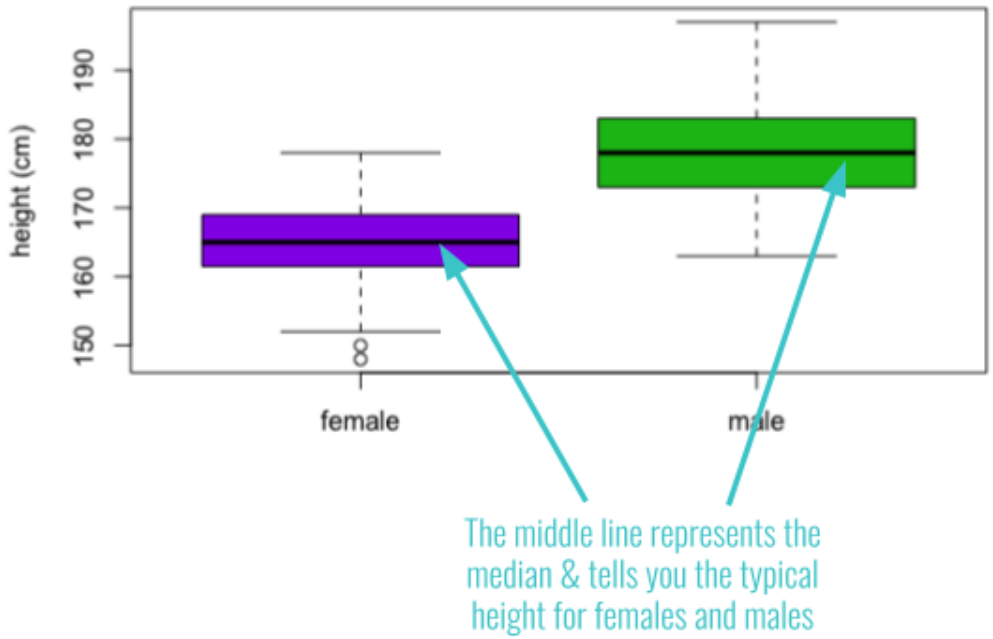
```
ggplot(data=tv_summary,
       mapping=aes(x=party_3, y=mean_tv))+
 geom_col(aes(fill=party_3)) +
geom_errorbar(aes(ymin=mean_tv-sd_tv,
ymax=mean_tv+sd_tv))
```



Box plots

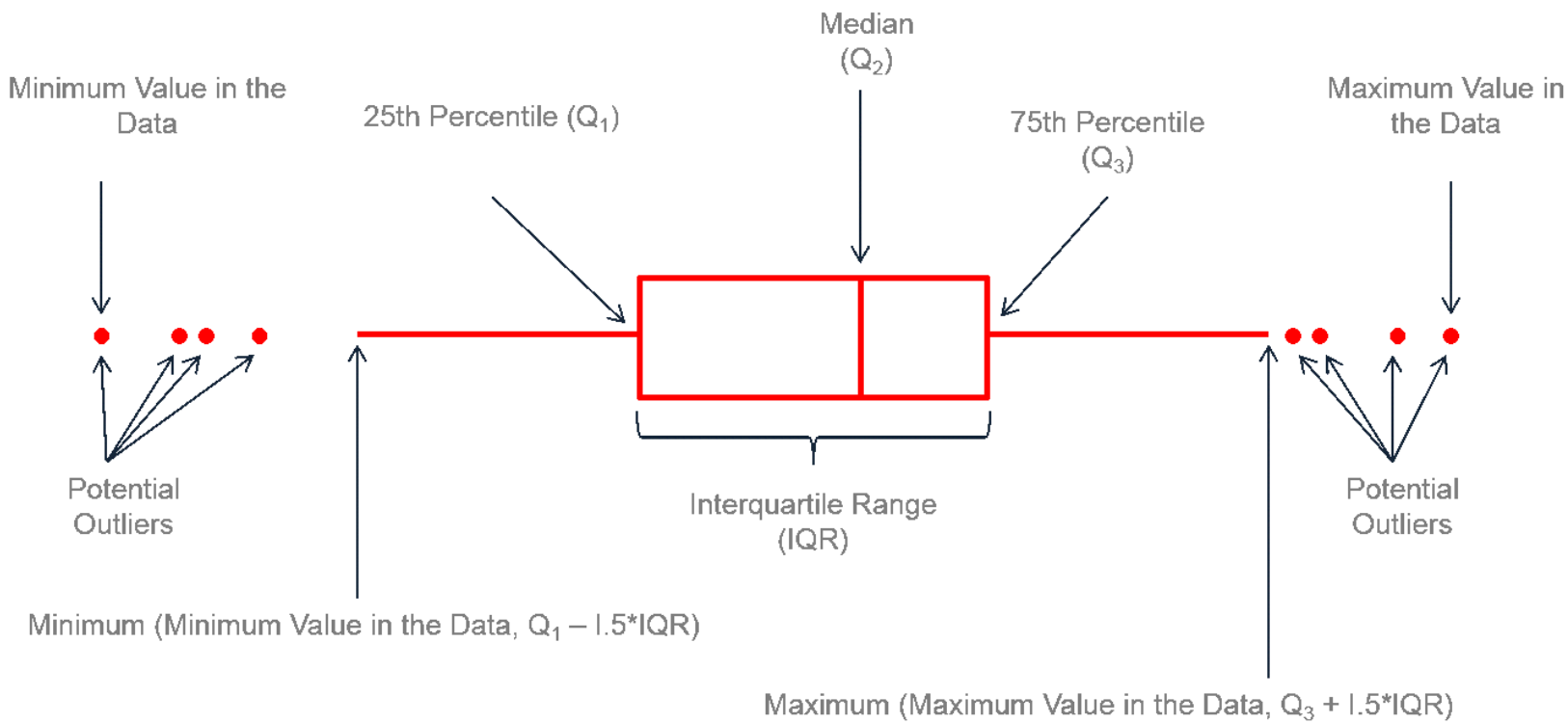
Boxplot

Summary of a quantitative variable broken down by a categorical variable



Components of a Box Plot

- a box plot gives more information about the distribution of a numeric variable (center and spread)



https://leansigmacorporation.com/box-plot-with-minitab/

Categorical and Quantitative (box plot)

Mapping Arguments:

- x= the categorical variable
- y= the quantative variable

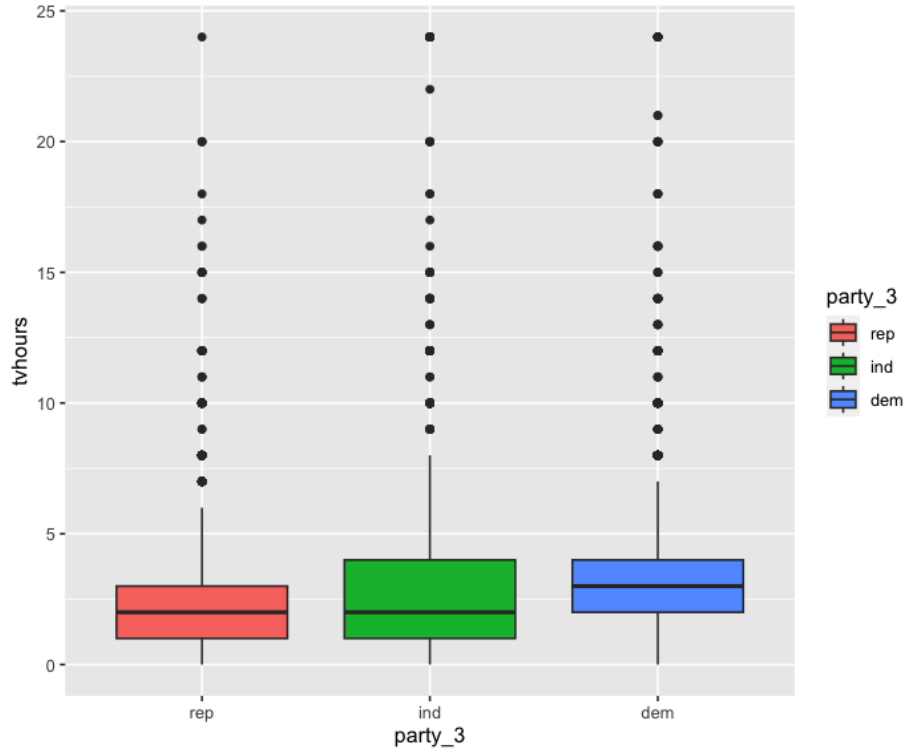
```
ggplot(data=gss_party,
       mapping=aes(x=party_3, y=tvhours))+
  geom_boxplot(aes(fill=party_3))
```

- using fill to color the box plots by party_3

Categorical and Quantitative (box plot)

- The box plot is showing that median tv watching is actually similar for republicans and independents
- The mean of independents is pulled up by some extreme values (right skewed)

```
ggplot(data=gss_party,
       mapping=aes(x=party_3, y=tvhours))+
  geom_boxplot(aes(fill=party_3))
```



Class Exercise

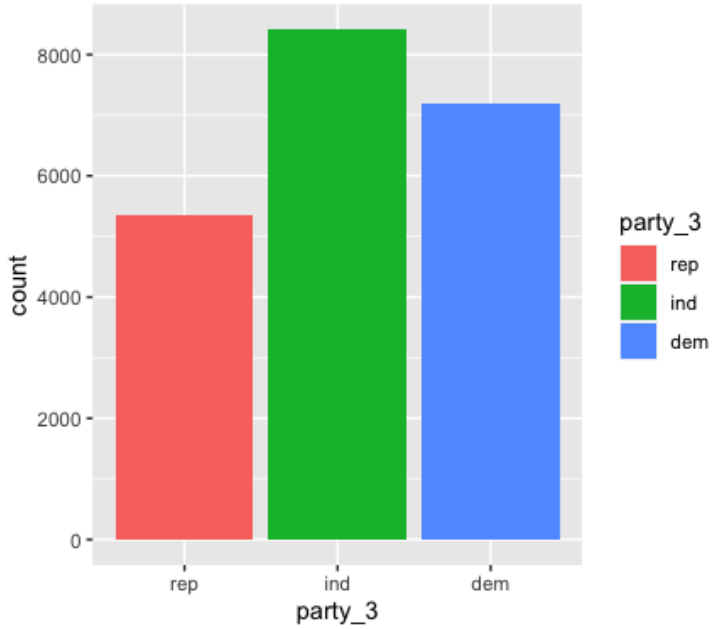
- create a box plot for age by marry_fewer

https://pollev.com/vsovero

Visualizations for Two Categorical Variables

Bar graphs

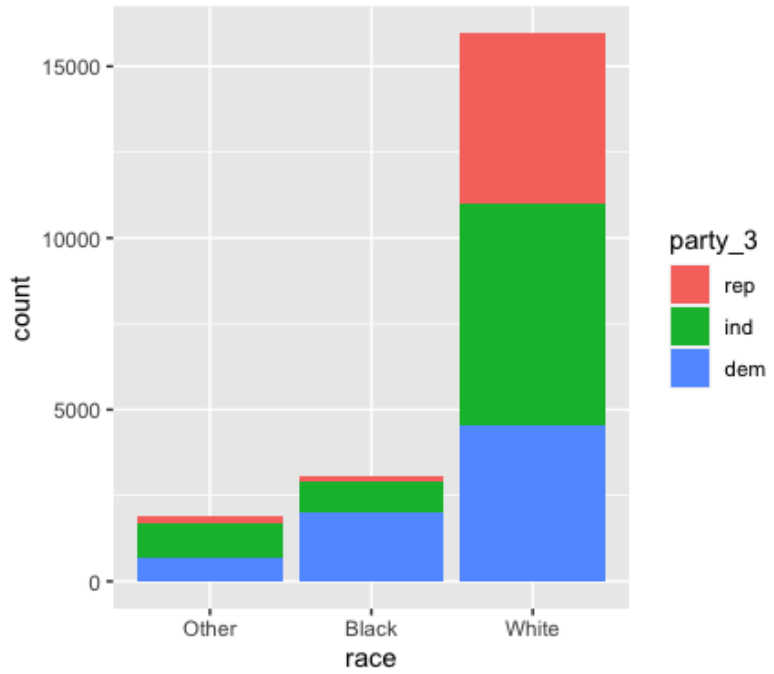
- We use bar plots to display information about categorical variables
- We can also examine the relationship between two categorical variables
- Example: are there differences in political affiliation by race?



Stacked Barplot

- Frequencies of political parties within each race

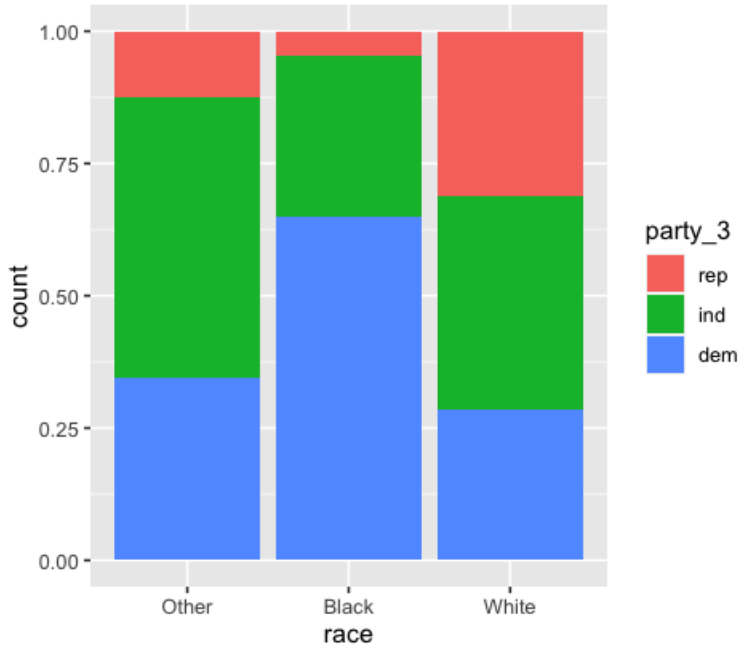
```
ggplot(data=gss_party,
       mapping= aes(x=race)) +
 geom_bar(aes(fill=party_3))
```



Stacked barplot (proportions)

- Proportion of political parties within each race

```
ggplot(data=gss_party,
       mapping= aes(x=race)) +
geom_bar(aes(fill=party_3), position ="fill")
```



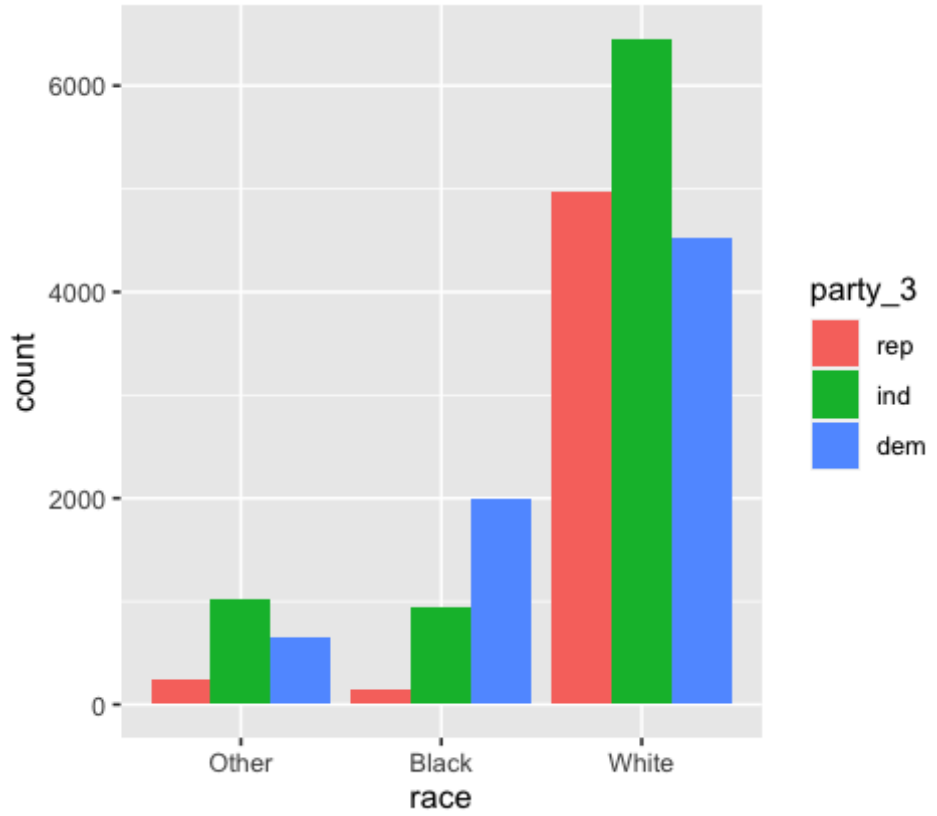
Class Exercise

- Create a stacked bar plot that shows the proportions of marital status for each political affiliation

Unstacked barplot

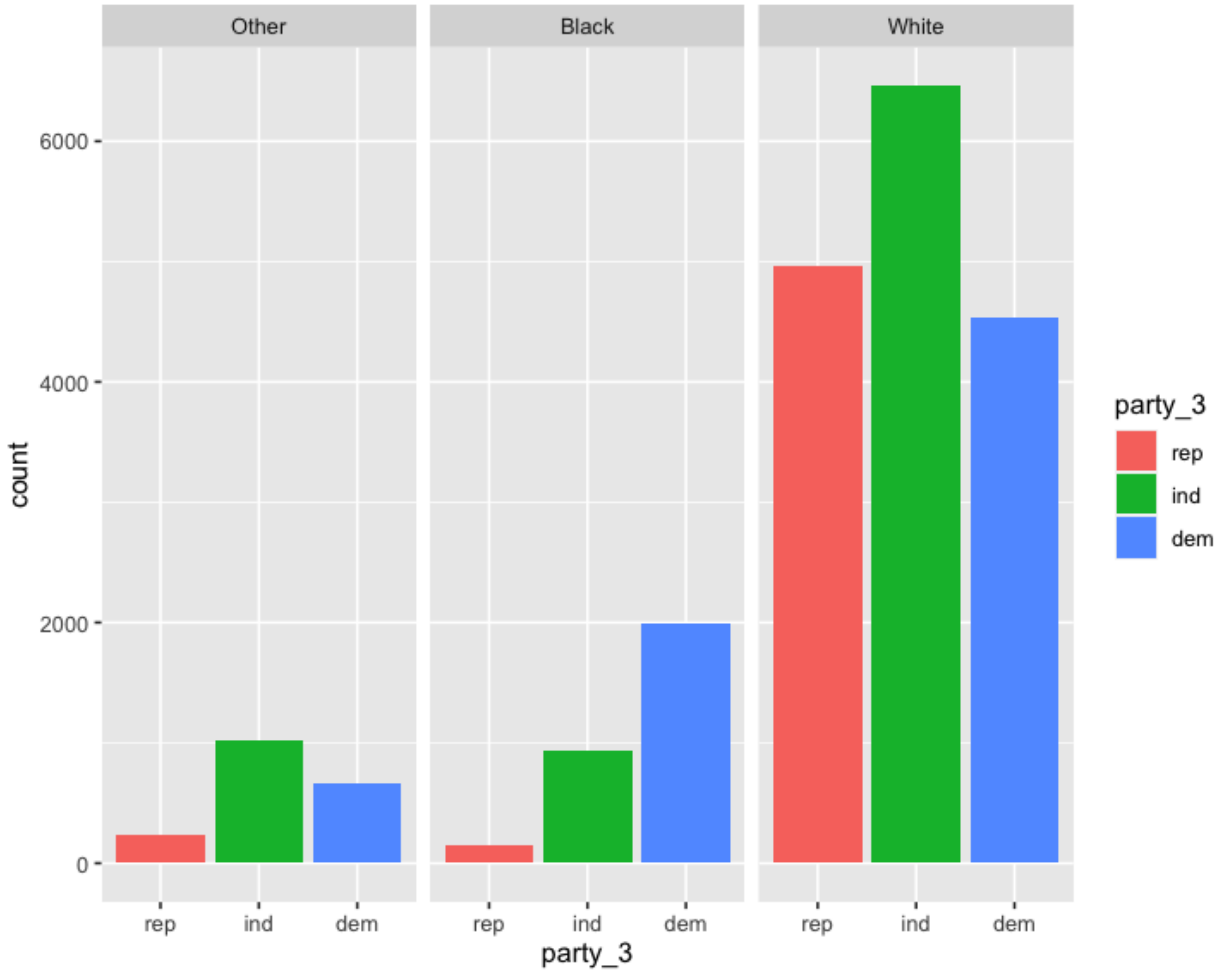
- Frequencies of political parties within each race

```
ggplot(data=gss_party,
       mapping= aes(x=race)) +
geom_bar(aes(fill=party_fewer), position ="dodge"))
```



Faceted barplot

```
ggplot(data=gss_party,
       mapping= aes(x=party_3)) +
 geom_bar(aes(fill=party_3))+
 facet_wrap(~race)
```



Class Exercise

- Create a faceted bar plot that shows the marital status for each political affiliation