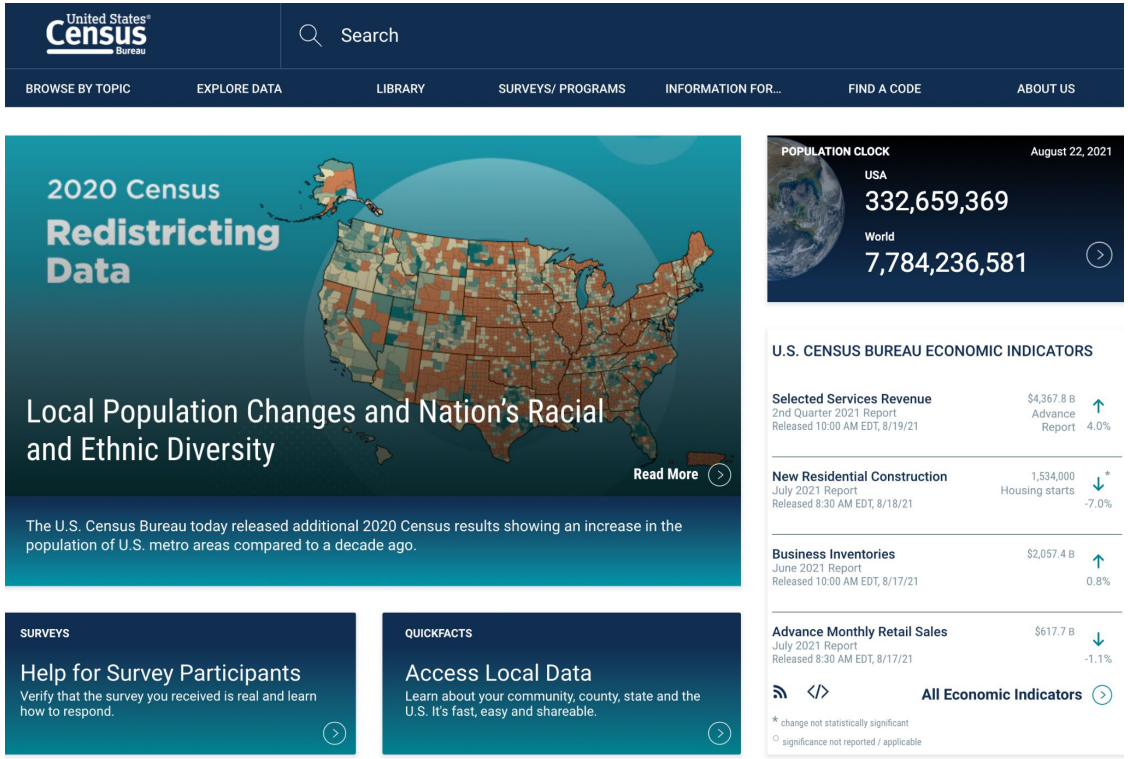
EDA with categorical data

Comparing numerical data across groups

Prof. Dr. Jan Kirenz
HdM Stuttgart

Data for 3142 counties in the United States

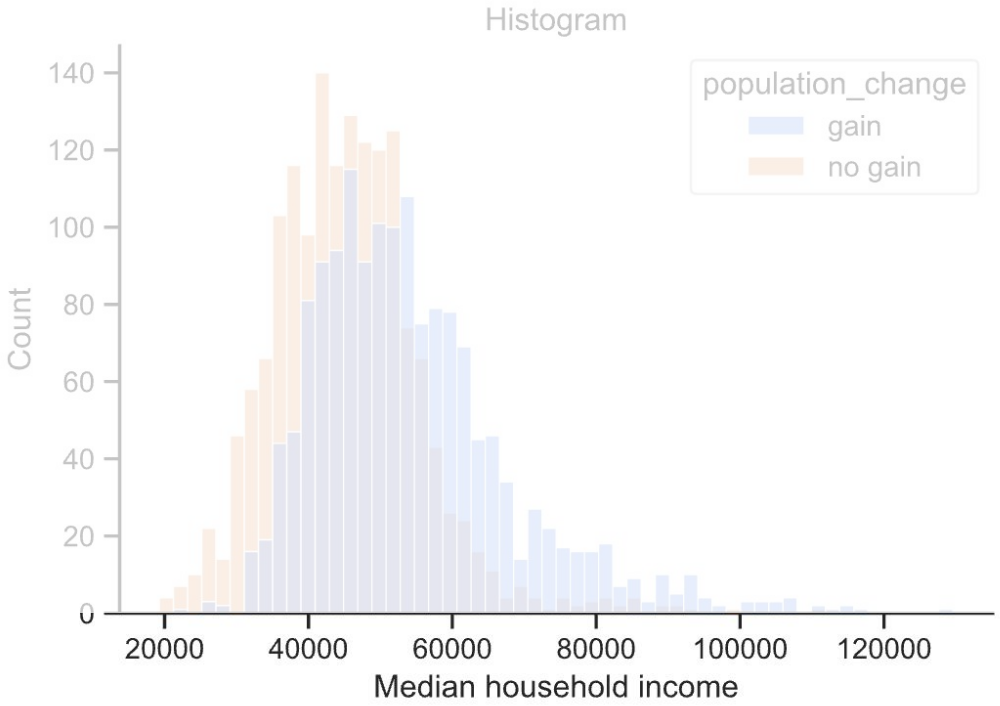


County dataset: https://www.openintro.org/data/index.php?data=county

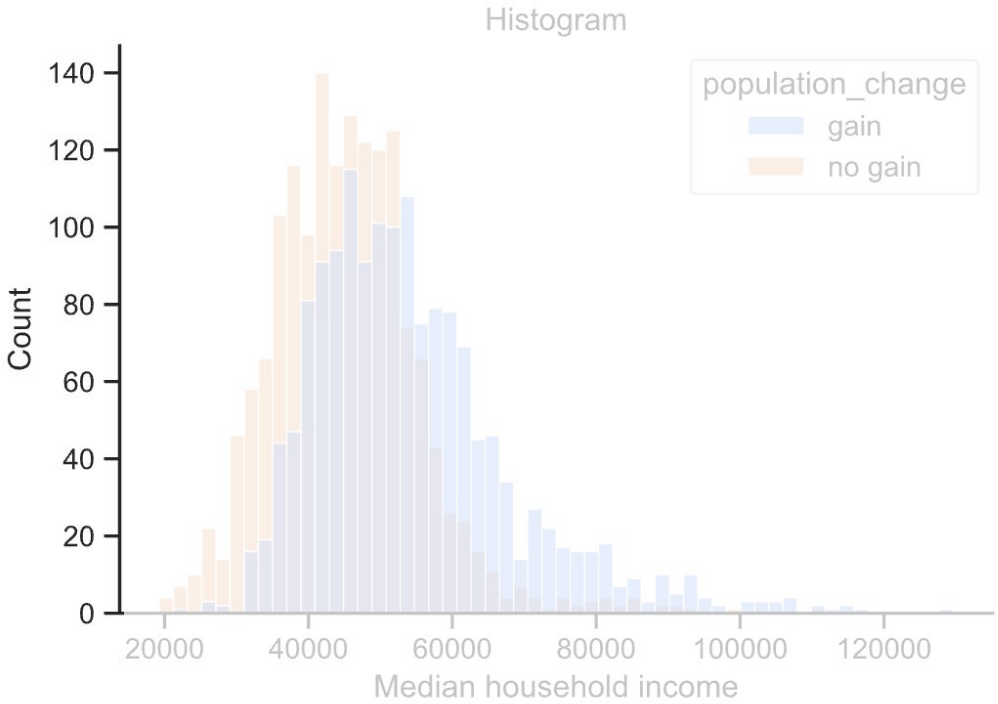
Overview about some counties with two groups: with and without population change

	state	name	pop_change	population_change	median_hh_income
0	Alabama	Autauga County	1.48	gain	55317.0
1	Alabama	Baldwin County	9.19	gain	52562.0
2	Alabama	Barbour County	-6.22	no gain	33368.0
3	Alabama	Bibb County	0.73	gain	43404.0
4	Alabama	Blount County	0.68	gain	47412.0

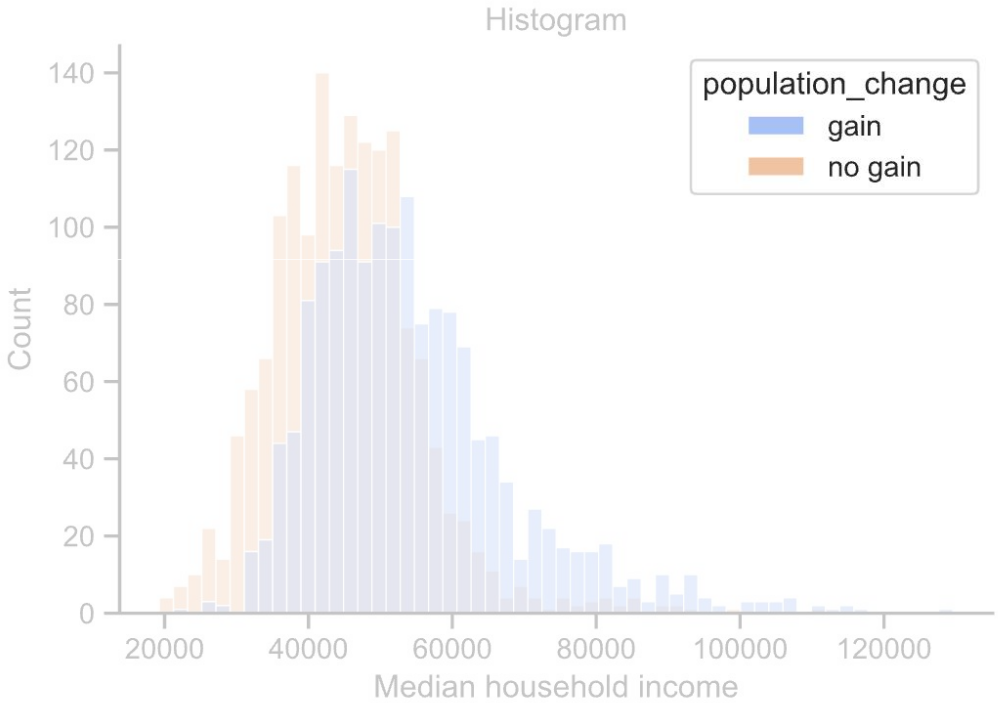
Median household income is on x-axis



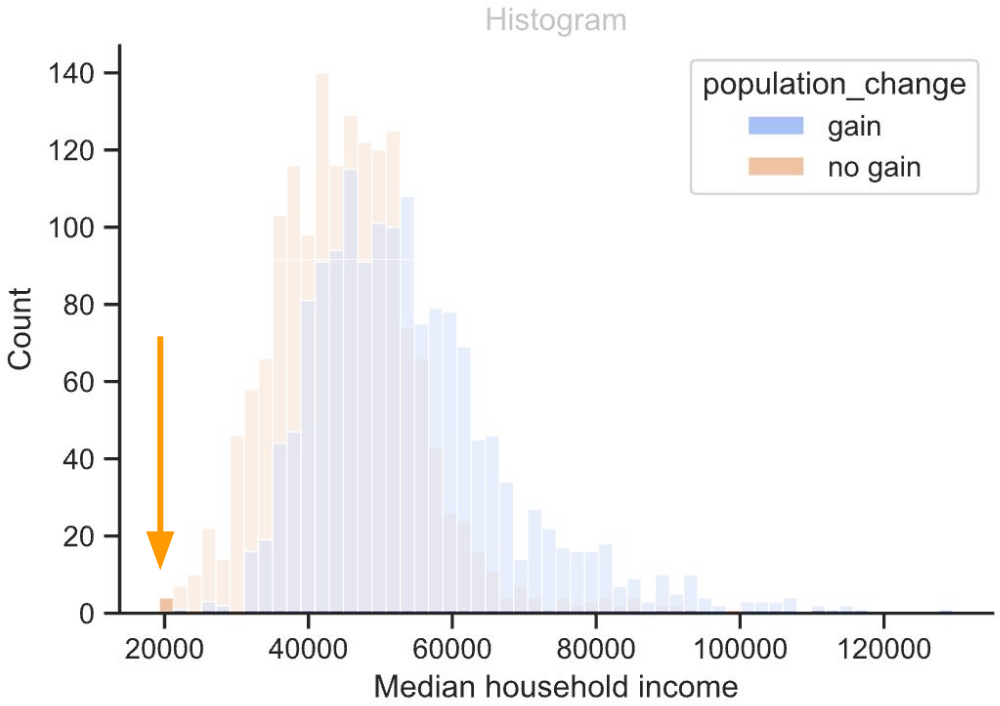
Count is on y-axis



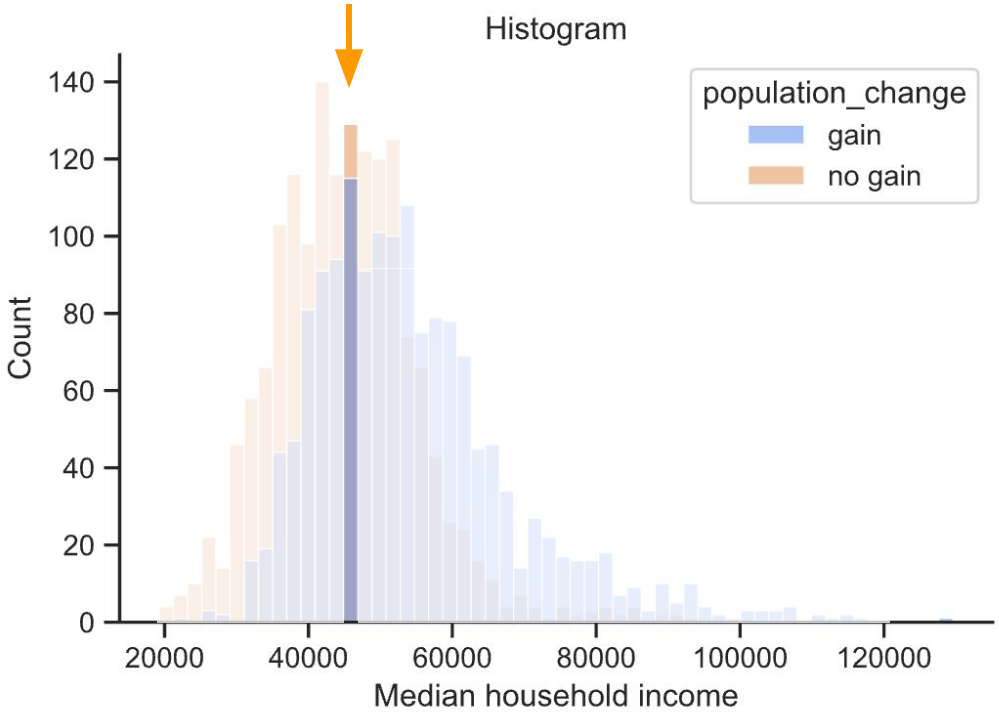
We compare two groups: **change in population**



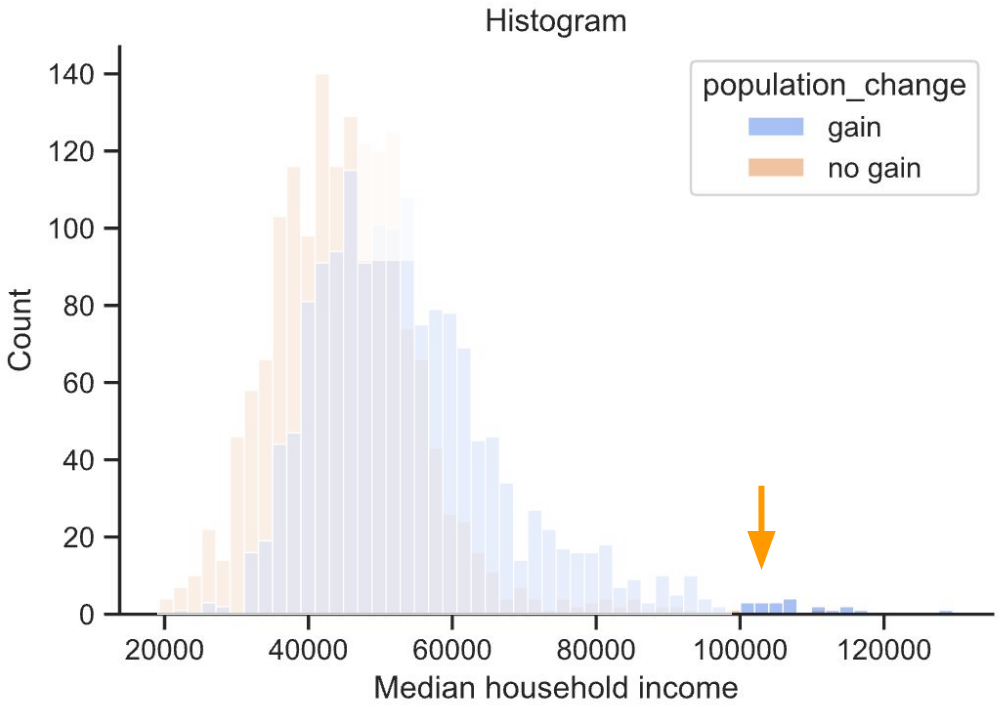
Only “no gain” in the first bin



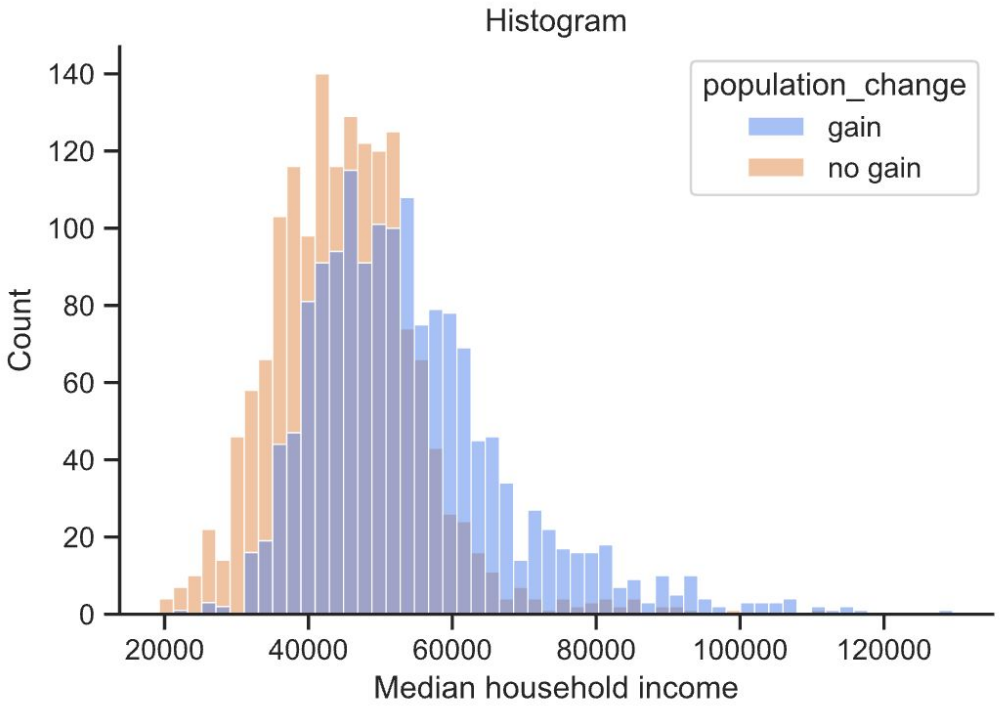
Now let's consider this bin



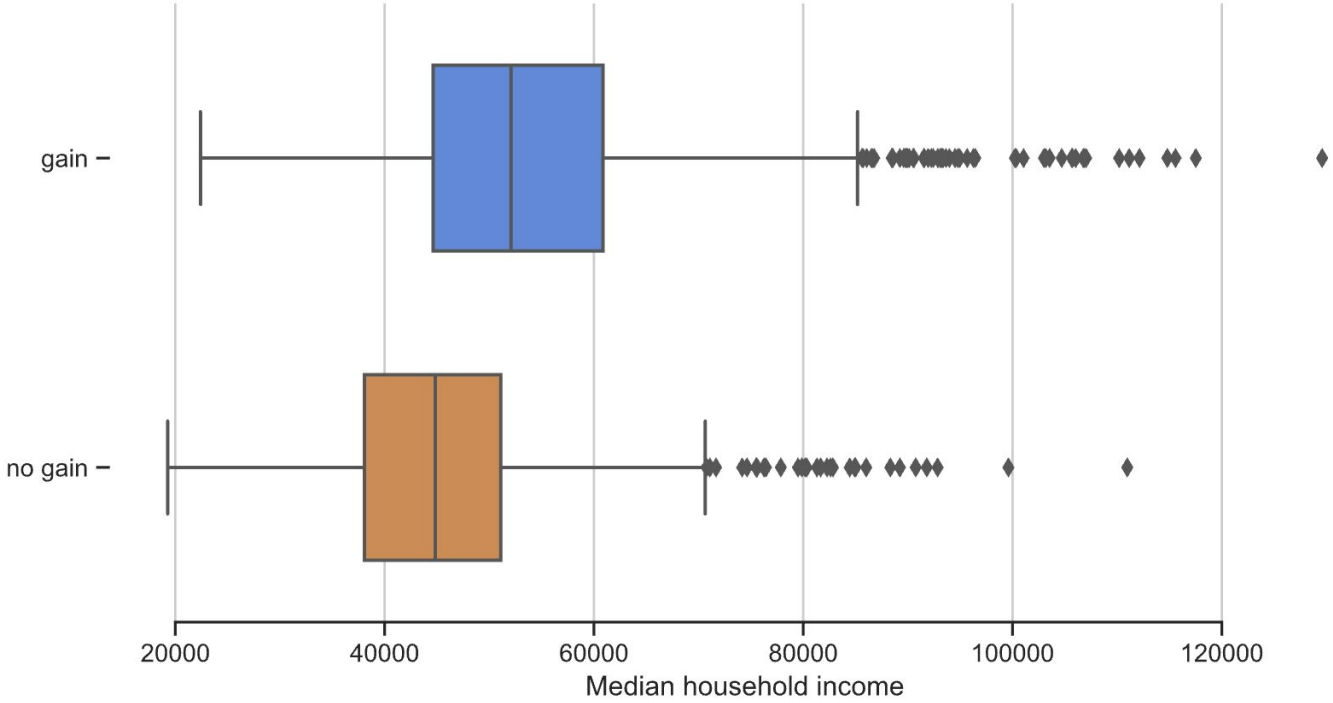
Now let’s consider these bins



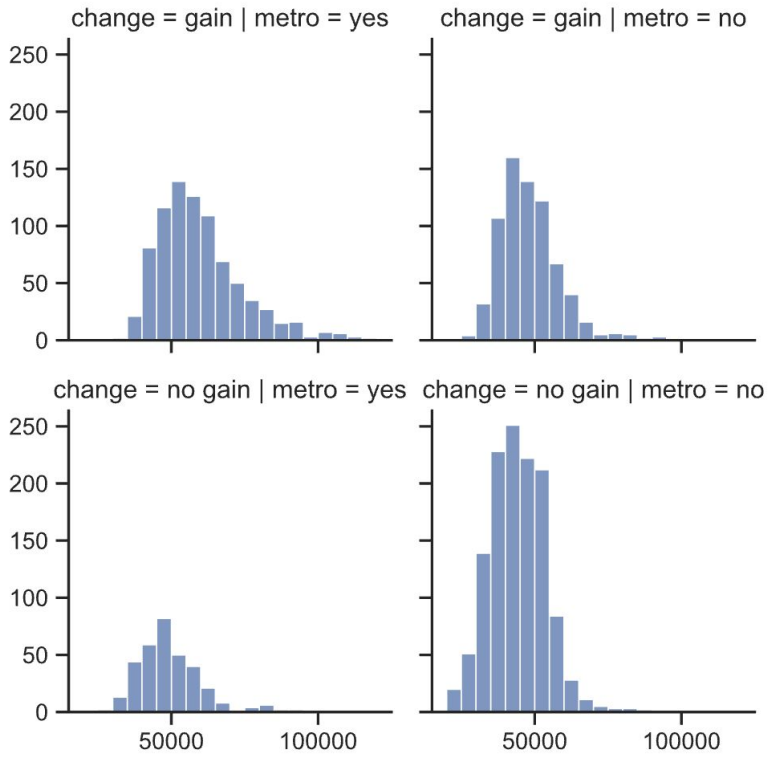
Histogram with two groups



Boxplot with two groups

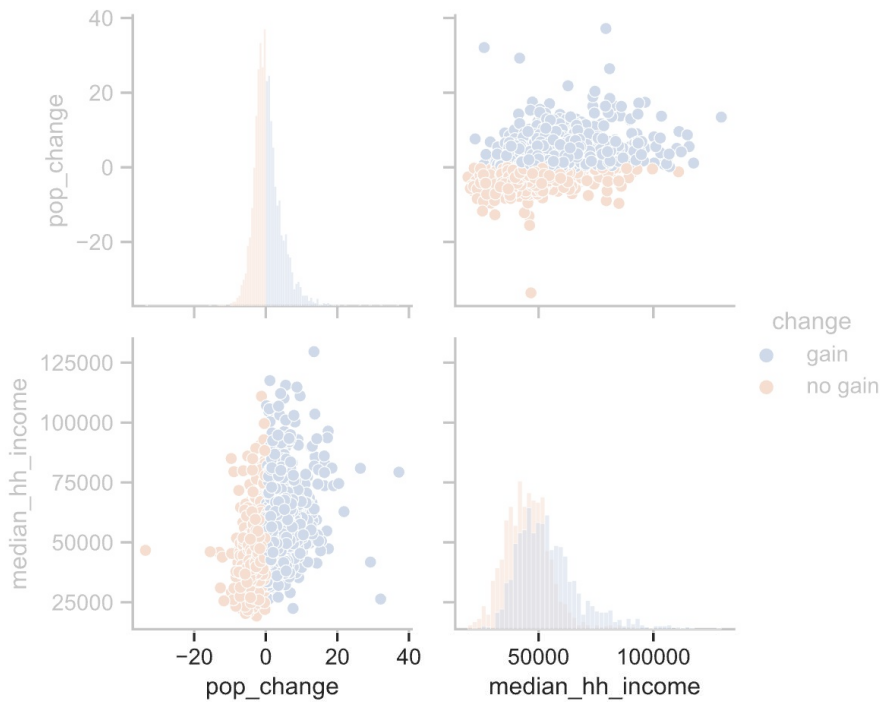


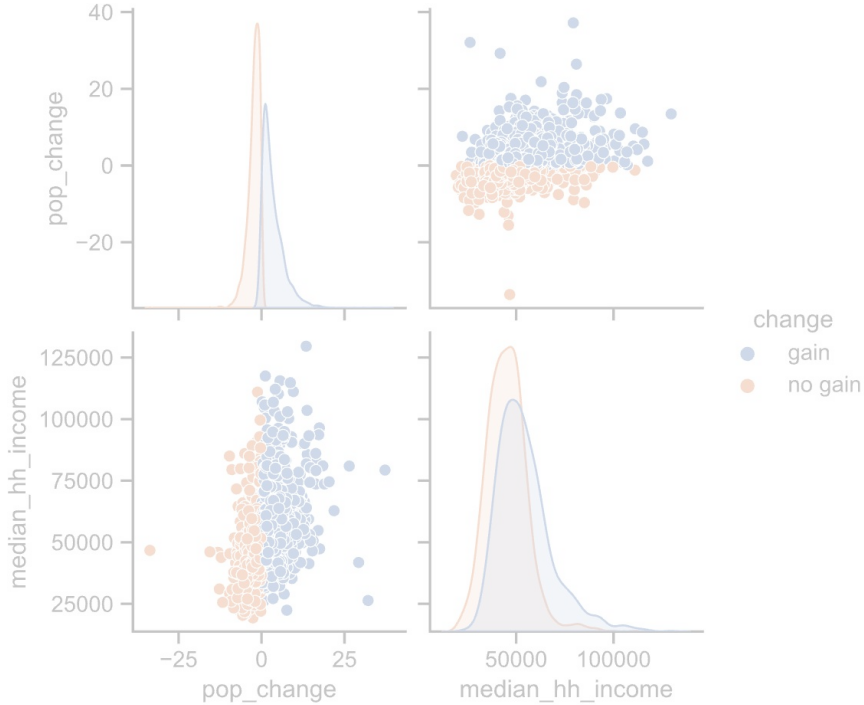
Faceting


change = gain | metro = yes
change = gain | metro = no
change = no gain | metro = yes
change = no gain | metro = no
250
200
150
100
50
0
50000
100000

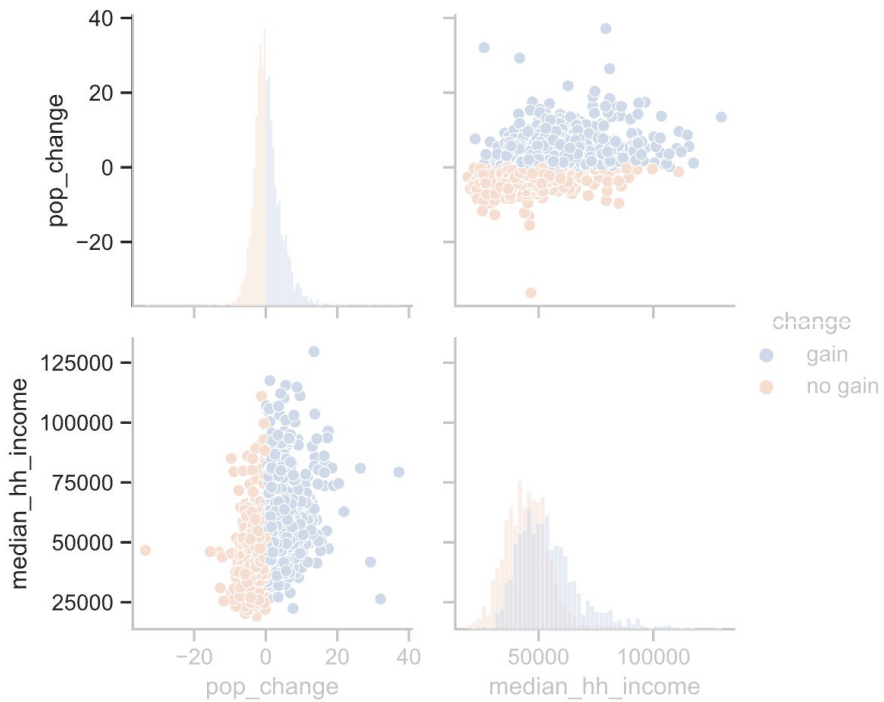
Additional graphs

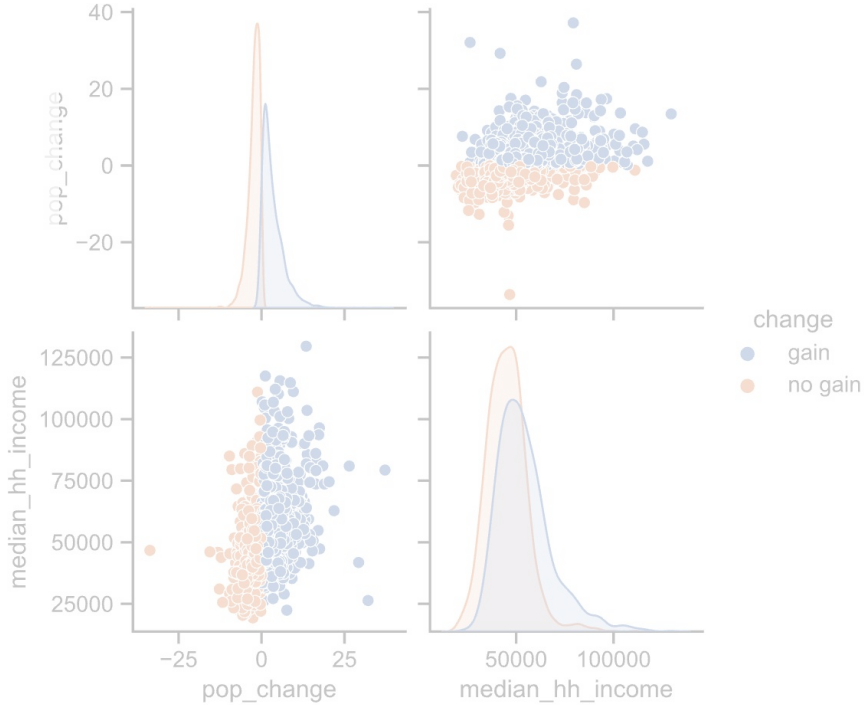
Plotting pairwise data relationships



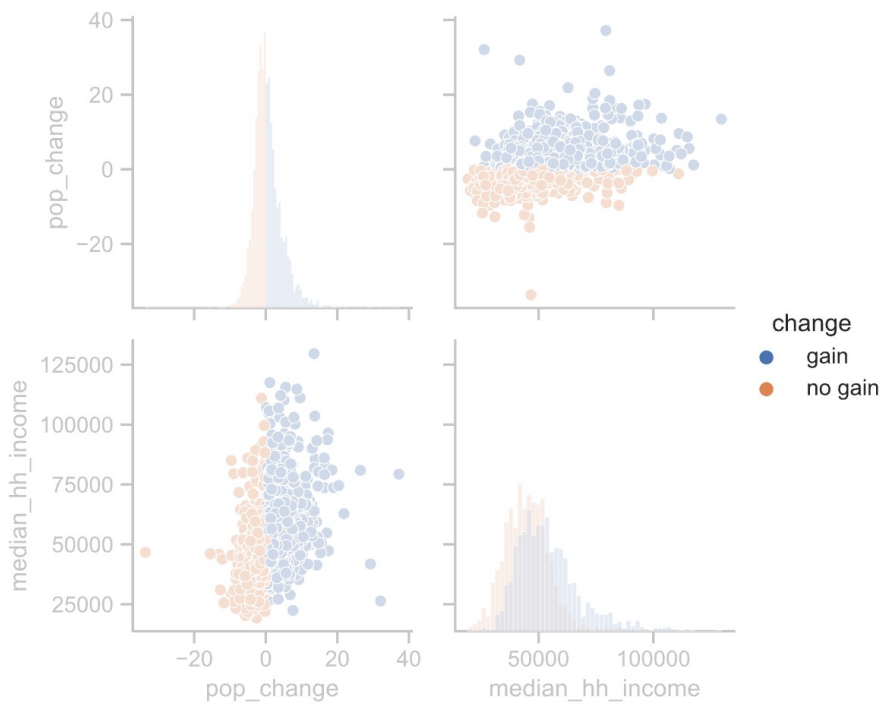


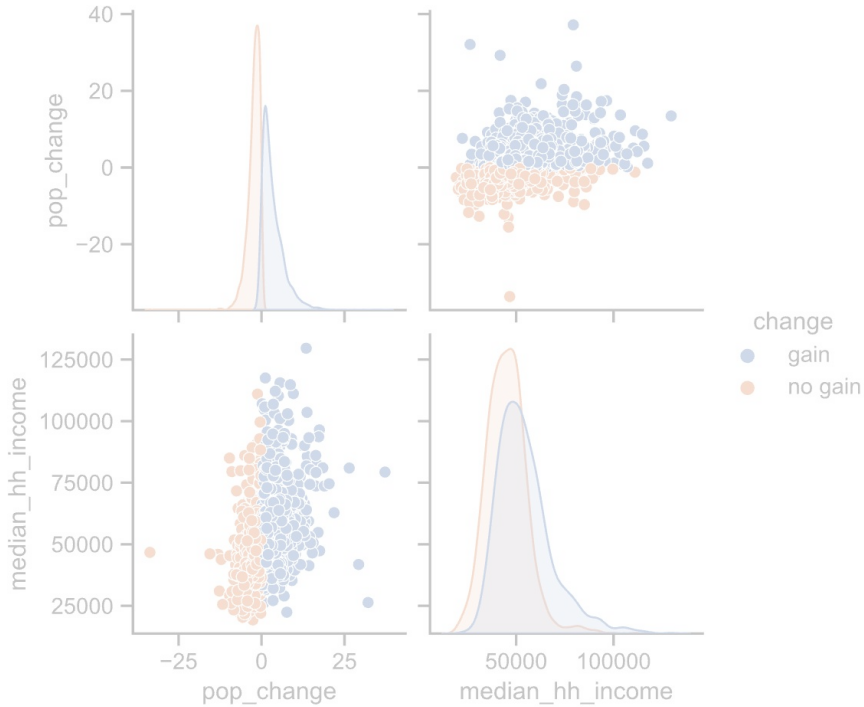
Plotting pairwise data relationships



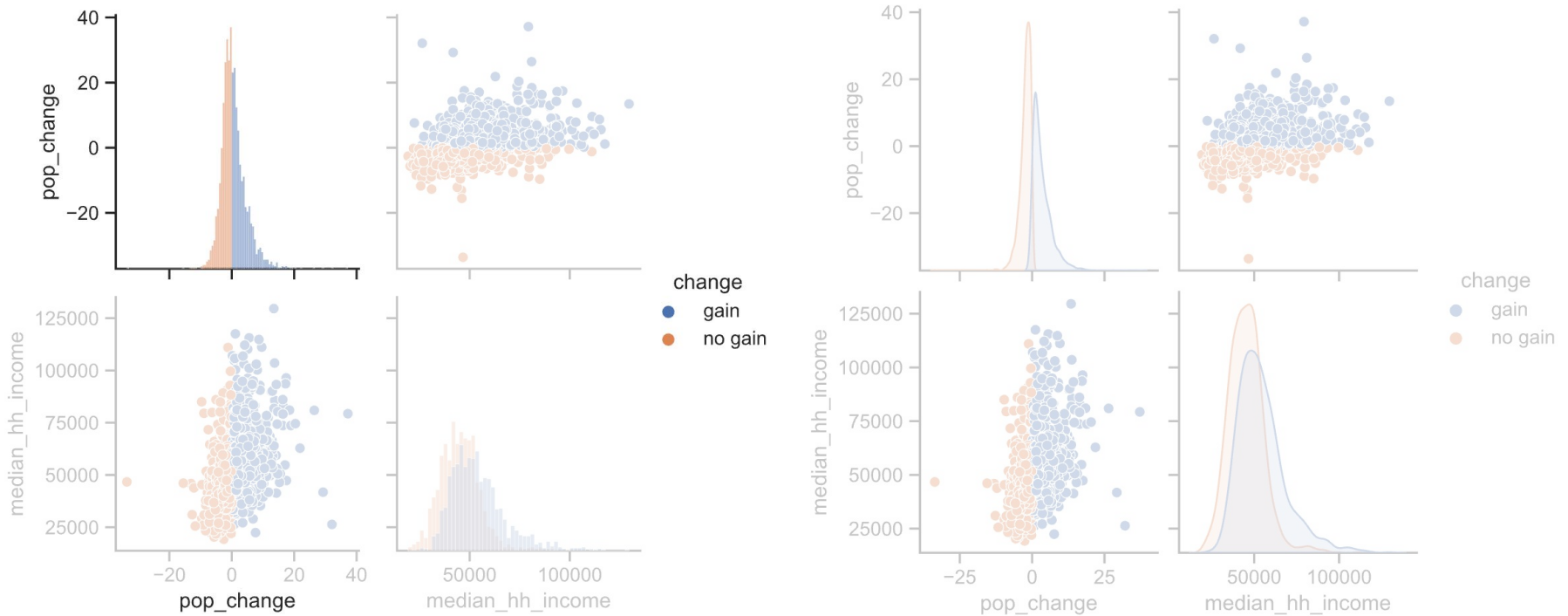


Plotting pairwise data relationships

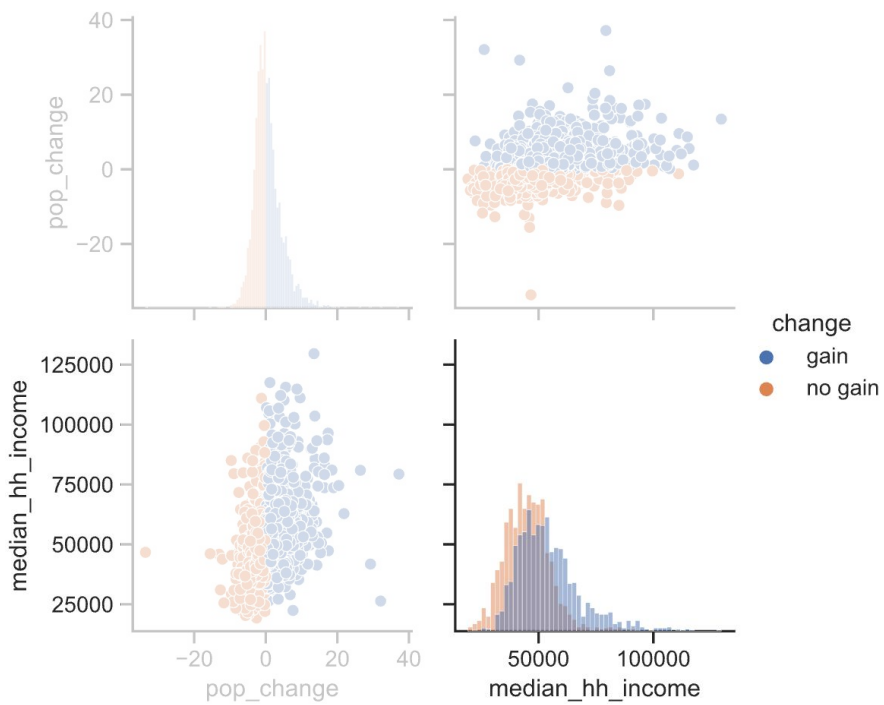


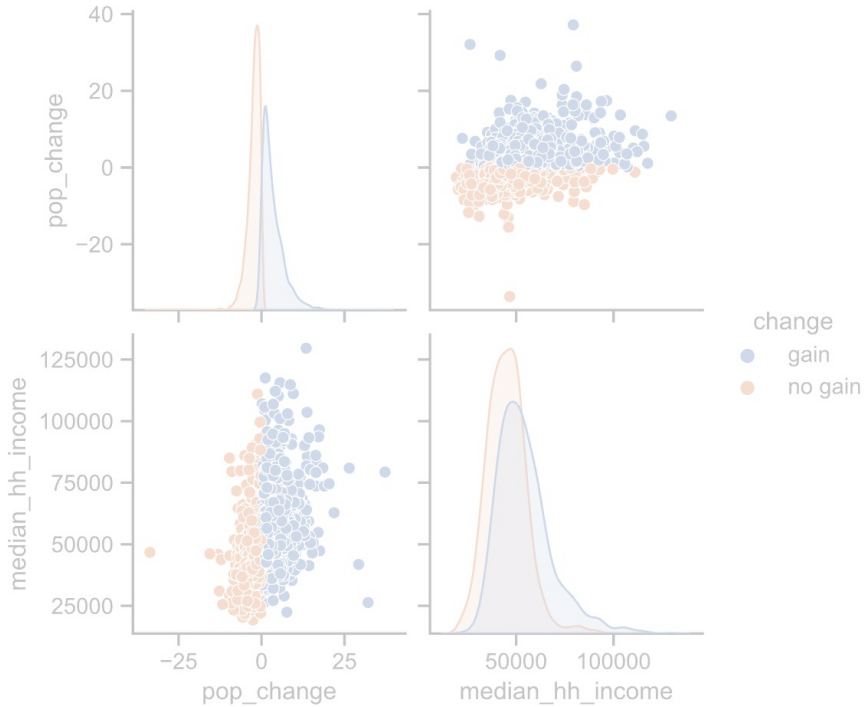


Plotting pairwise data relationships

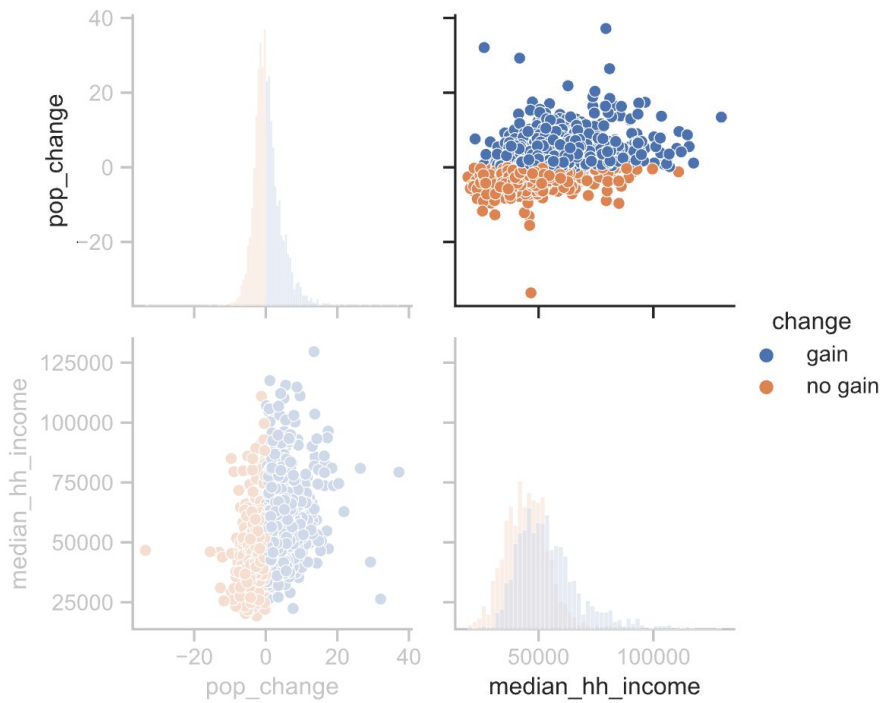


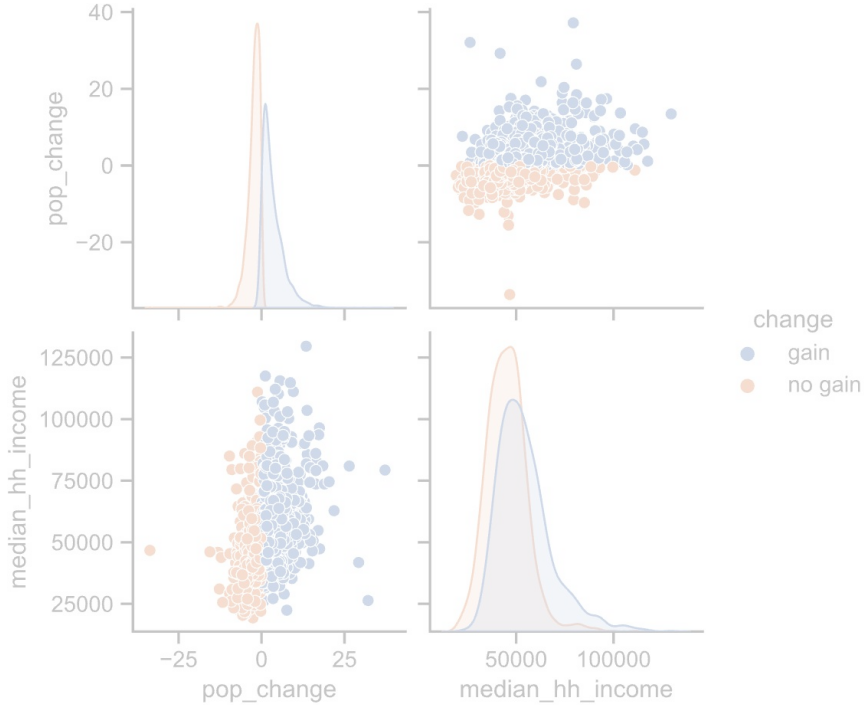
Plotting pairwise data relationships



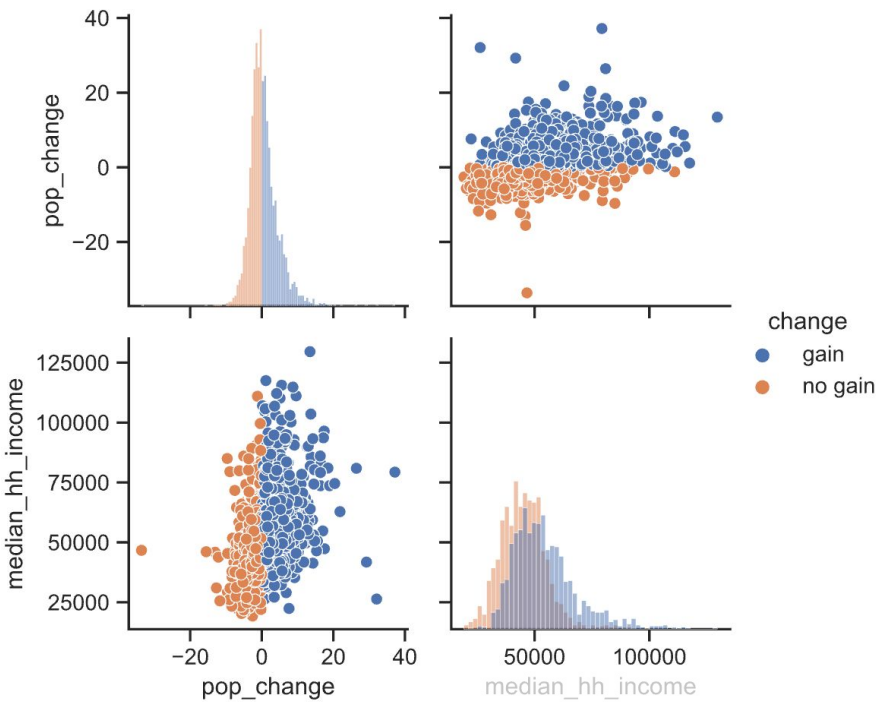


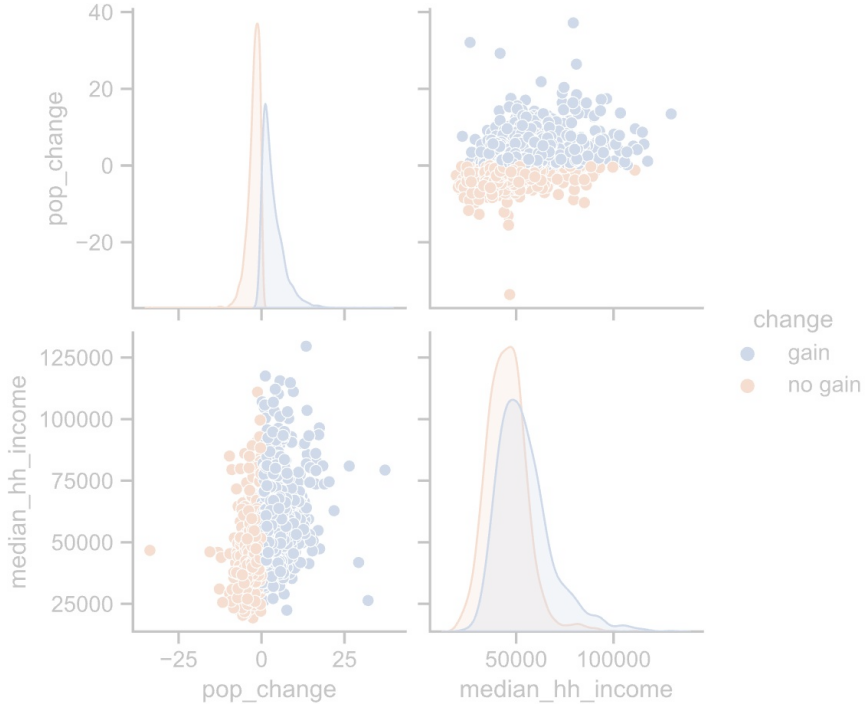
Plotting pairwise data relationships



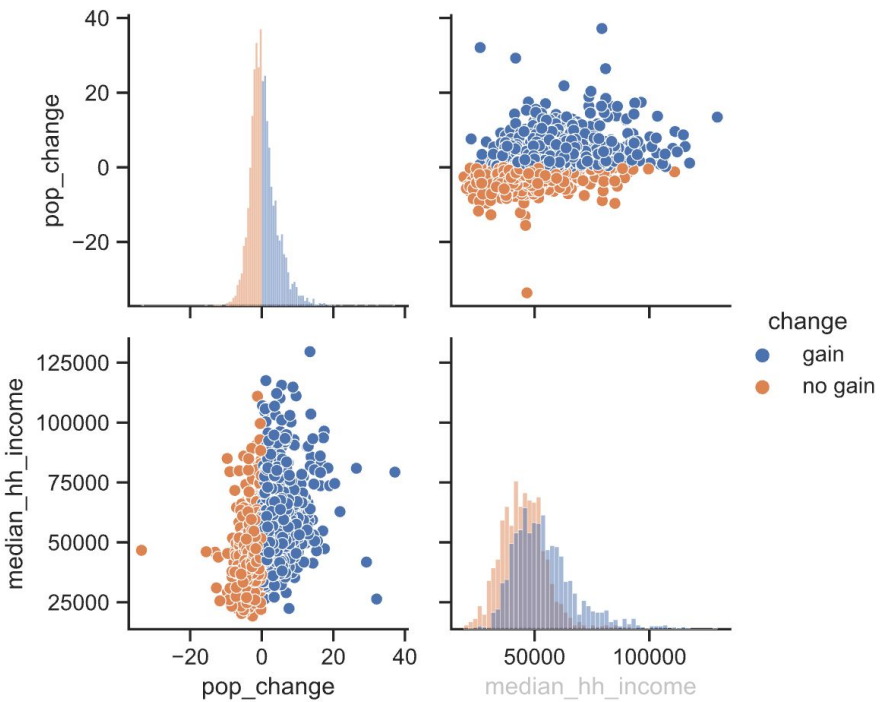


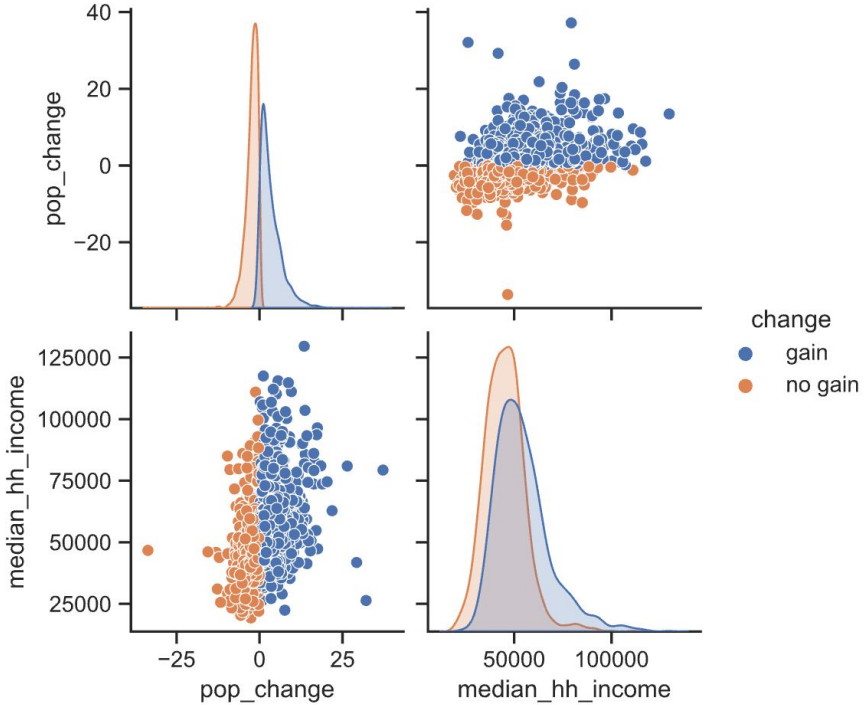
Plotting pairwise data relationships





Plotting pairwise data relationships





Terms you should know

column proportions

column totals

conditional proportions

contingency table

dodged bar plot

faceted plot

filled bar plot

mosaic plot

(ridge plot)

row proportions

row totals

side-by-side box plot

stacked bar plot

standardized bar plot

Resources

The slides are based on the excellent book “Introduction to Modern Statistics” by Mine Çetinkaya-Rundel and Johanna Hardin.

The online version can be **accessed for free**:

https://openintro-ims.netlify.app/explore-categorical.html

