Categorical Data

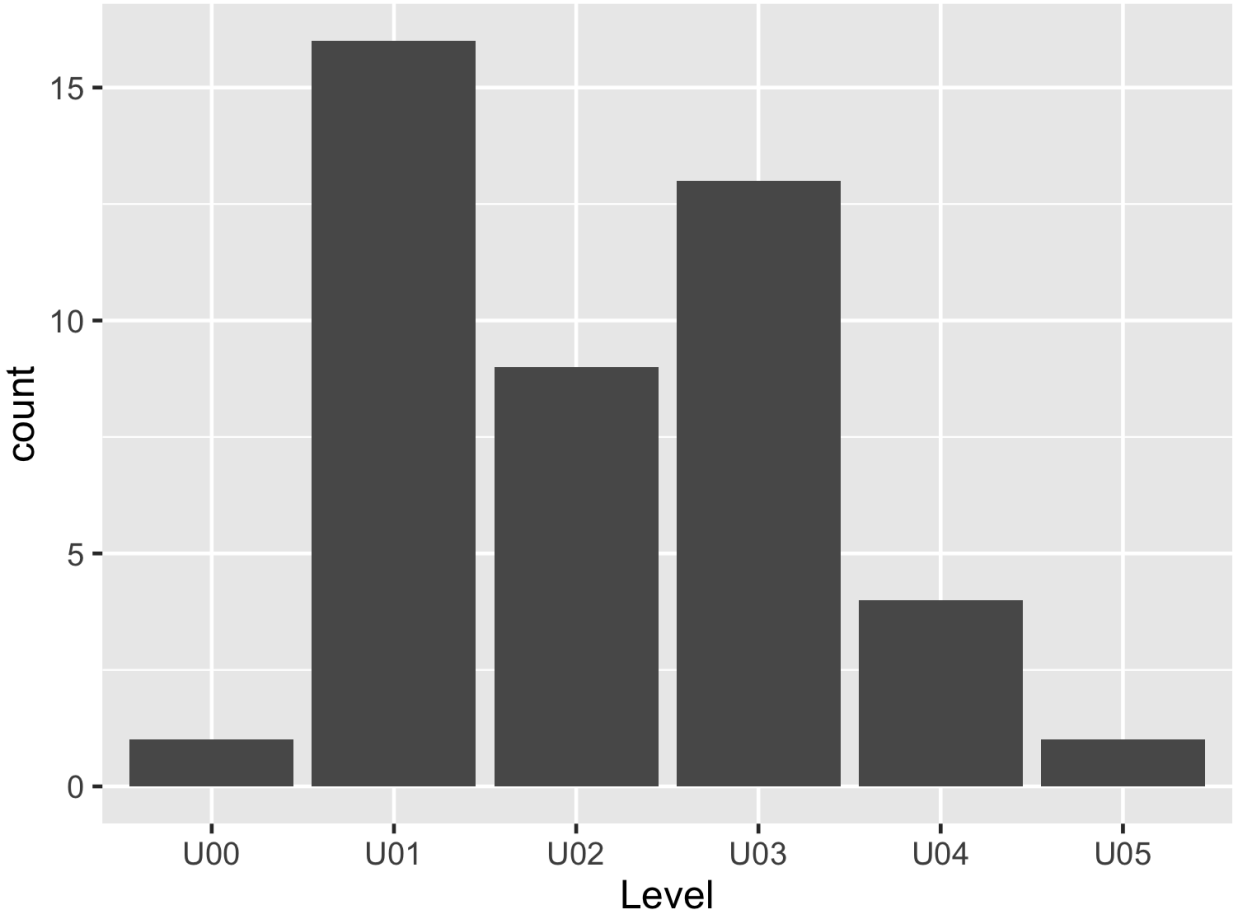
- hard to work with
- not a lot of options (esp. for 1 dimension)
- choice about which categories to display
- choice of the order of categories
- data cleaning takes more time

Types of data

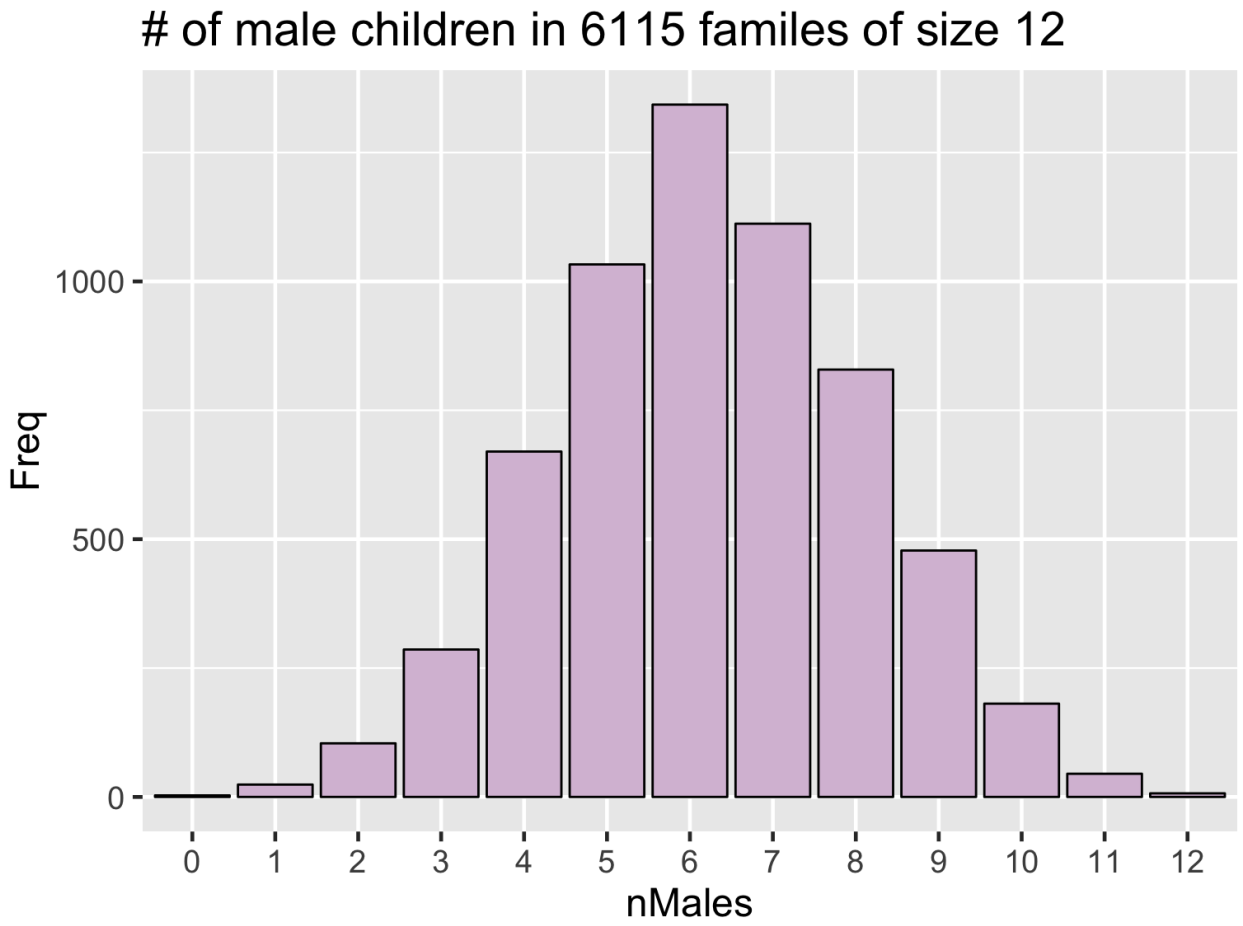
- nominal – no fixed category order
- ordinal – fixed category order
- (“real”) discrete, small # of possibilities
- Not always clearcut: nominal vs. ordinal, ordinal vs. discrete, and…
- Sometimes numbers = nominal, not discrete

Level (ordinal)

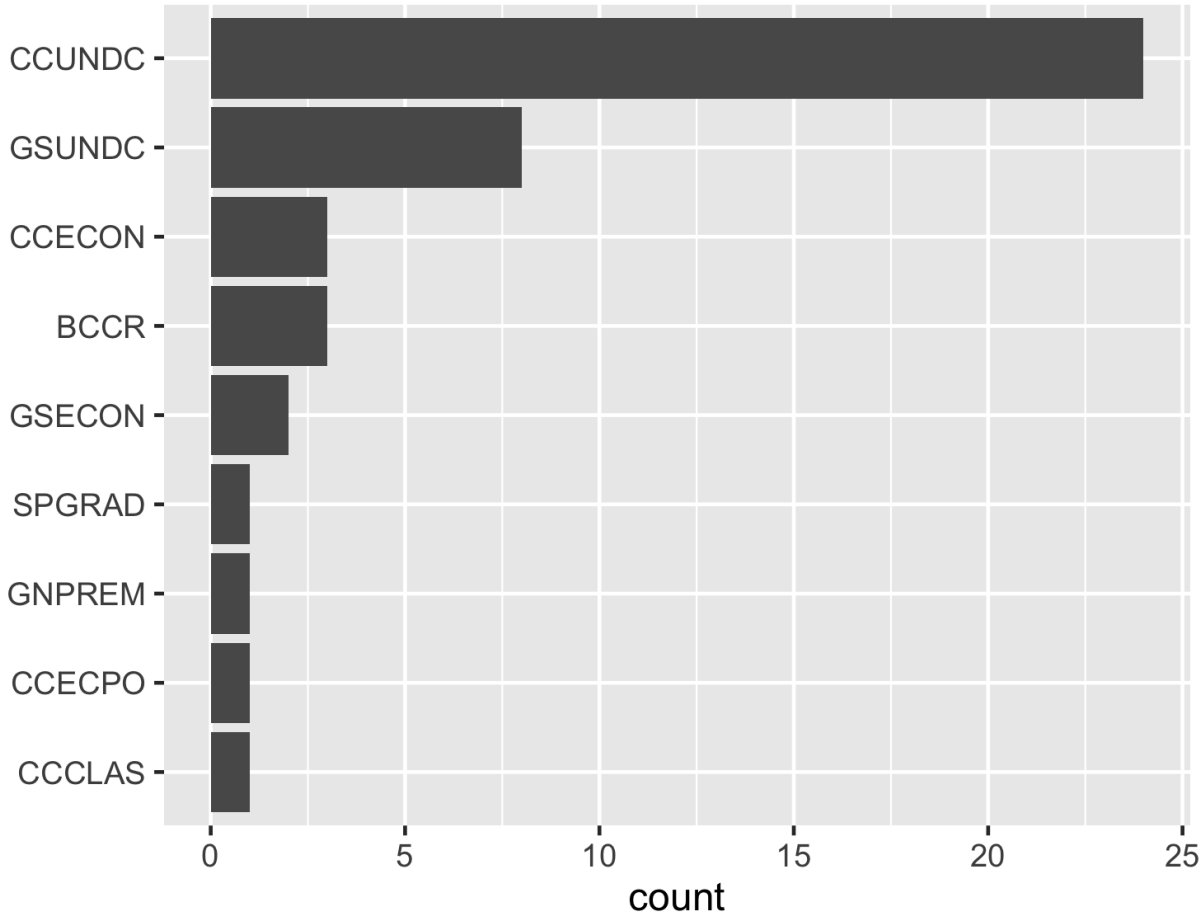
Sort in logical order of the categories



Number of children (19th century Saxony) (real discrete)

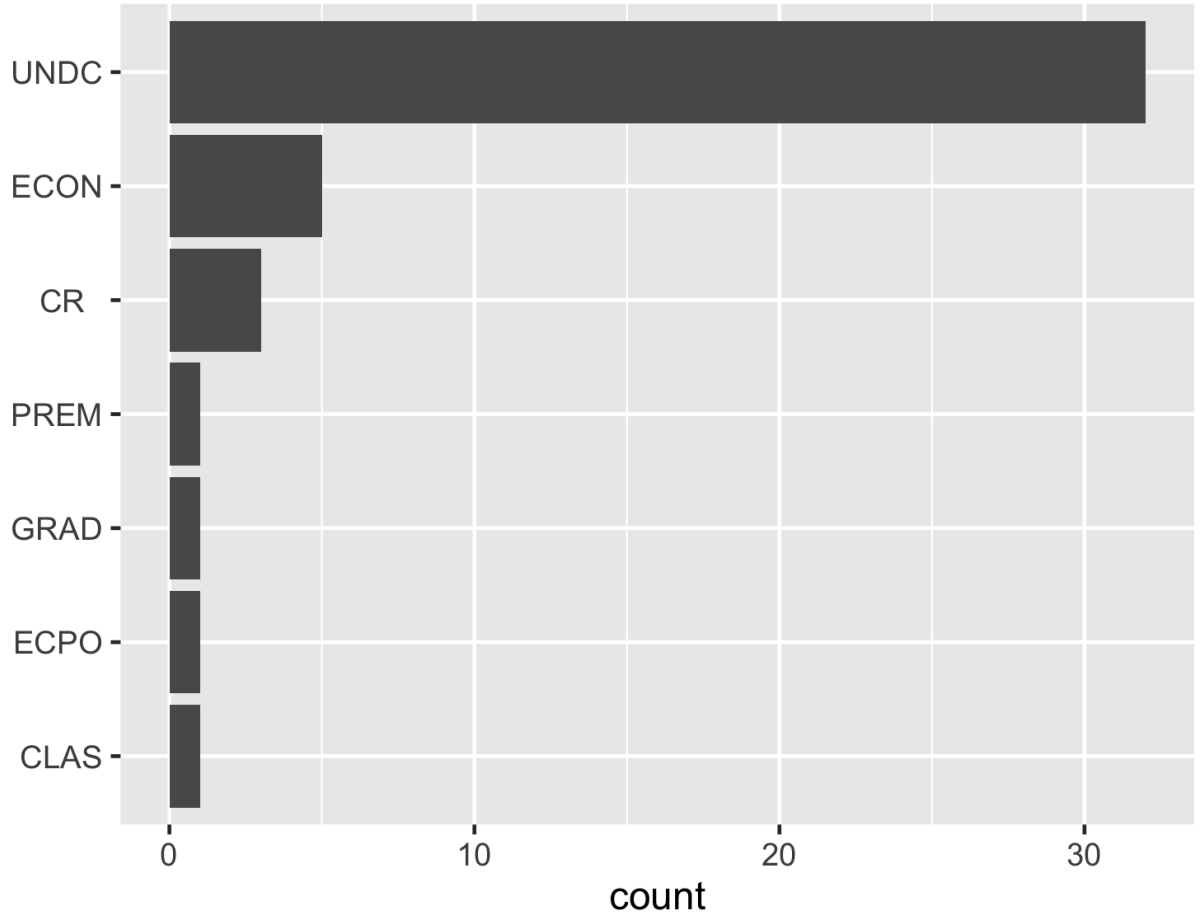


Affiliation (nominal)

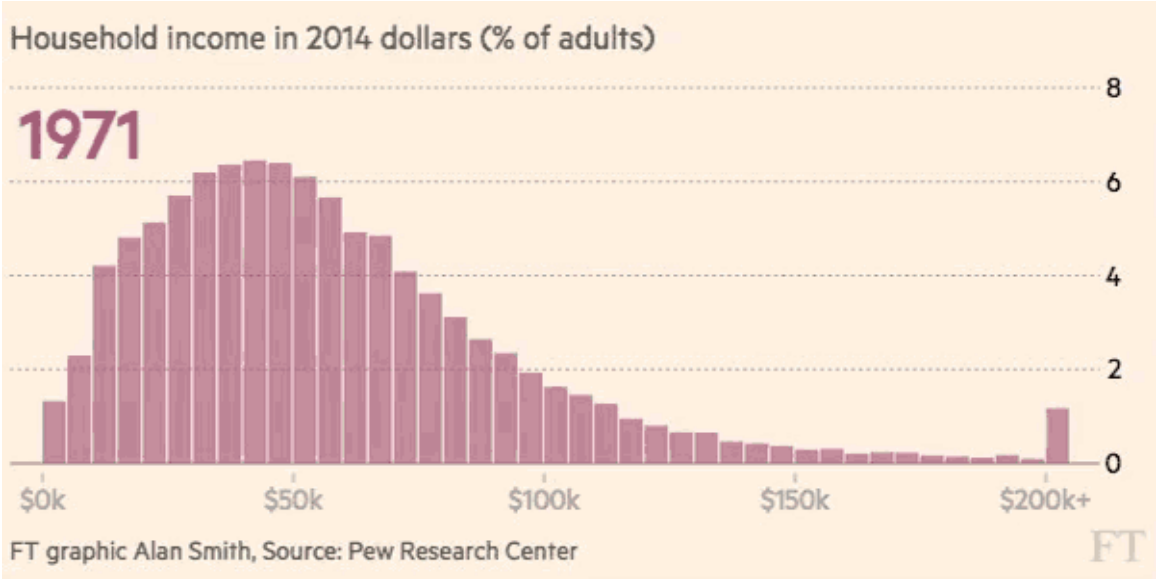


What's the problem?

Affiliation (last 4 characters only = major) (nominal)



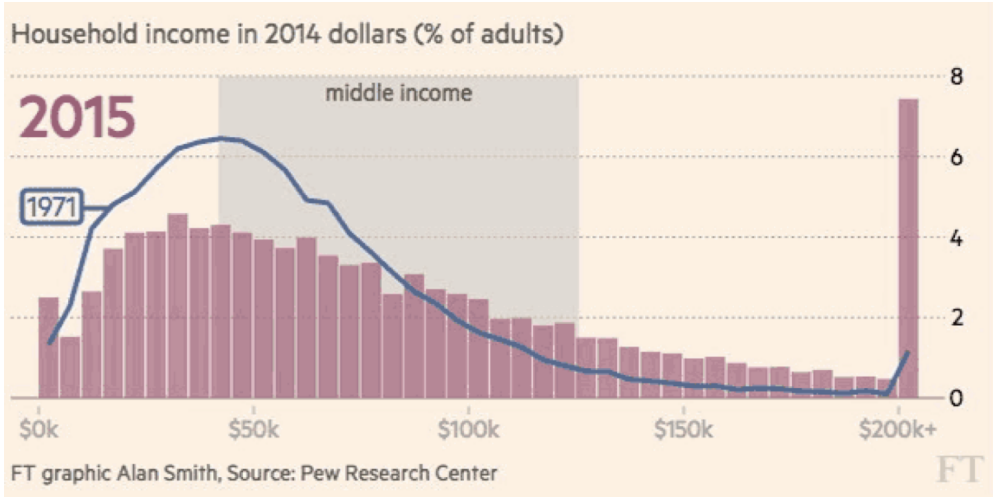
“Other” “or more” categories: “topcoding”



Source: “America’s explosion of income inequality, in one amazing animated chart”

http://www.latimes.com/business/hiltzik/la-fi-hiltzik-ft-graphic-20160320-snap-htmlstory.html

“Other” “or more” categories: “topcoding”



Source: “America’s explosion of income inequality, in one amazing animated chart”

http://www.latimes.com/business/hiltzik/la-fi-hiltzik-ft-graphic-20160320-snap-htmlstory.html

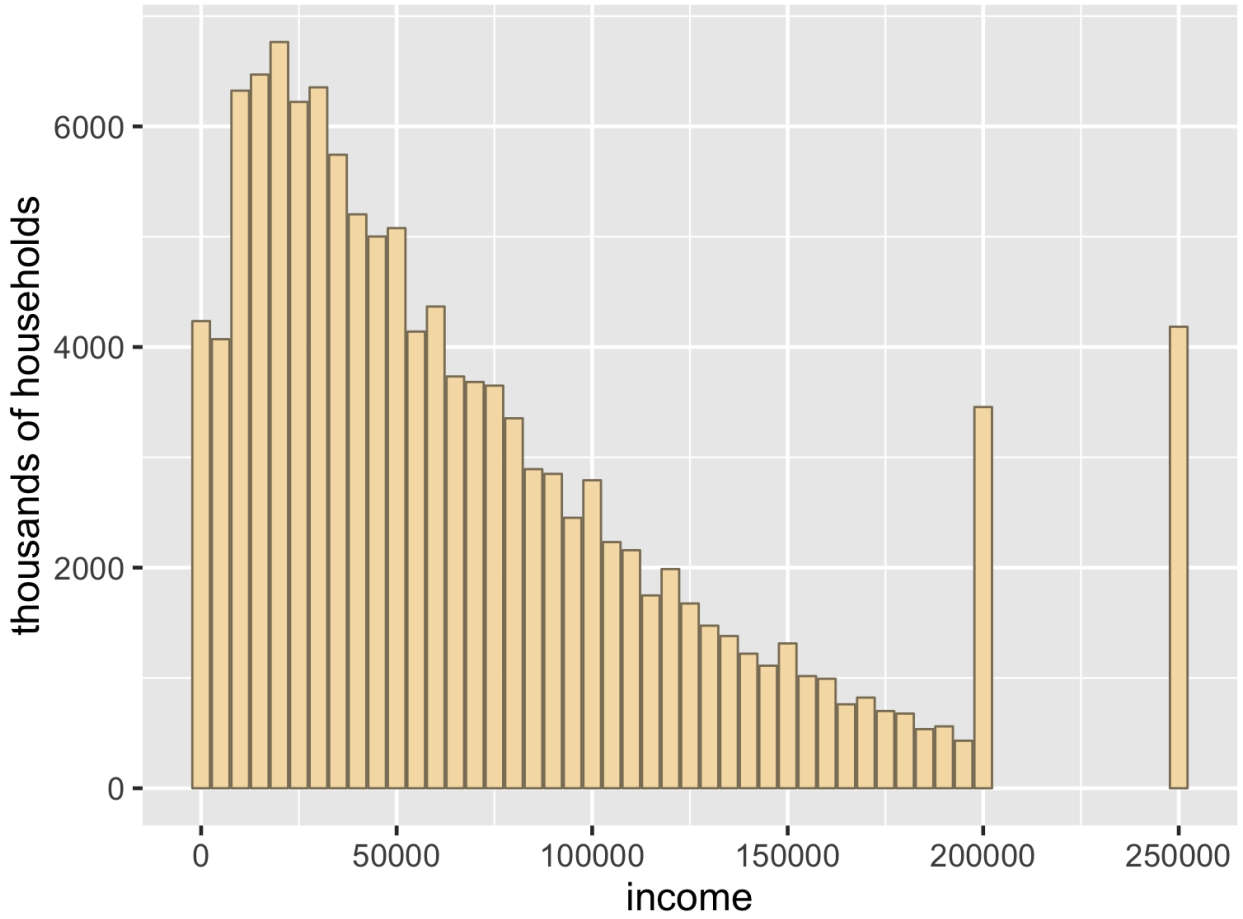
“Other” “or more” categories

$82,500 to $84,999	$85,000 to $87,499	$87,500 to $89,999	$90,000 to $92,499	$92,500 to $94,999	$95,000 to $97,499	$97,500 to $99,999	$100,000 and over	Va (D
1,102	1,683	892	2,065	894	1,306	770	22,426	

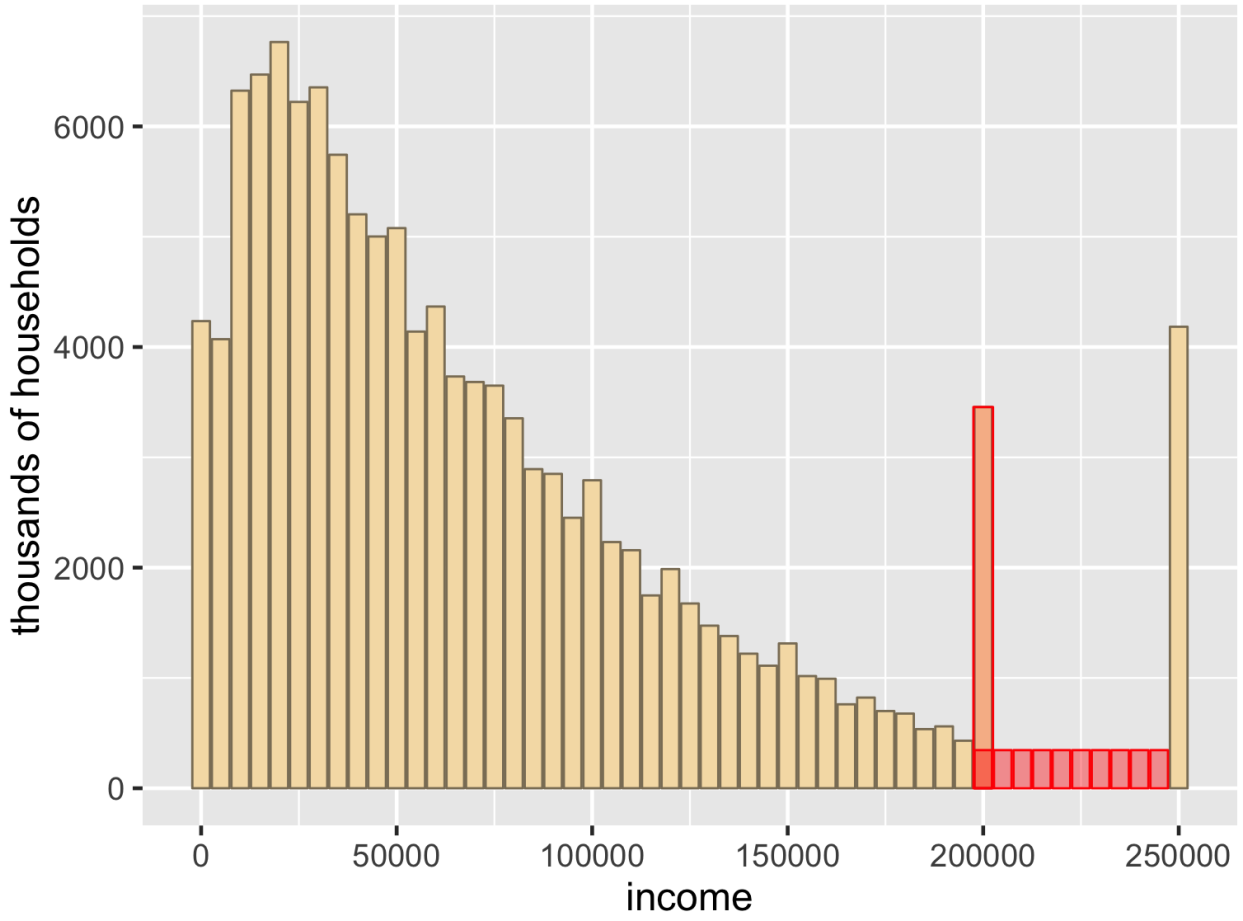
$82,500 to $84,999	$85,000 to $87,499	$87,500 to $89,999	$90,000 to $92,499	$92,500 to $94,999	$95,000 to $97,499	$97,500 to $99,999	$100,000 and over	Va (D
973	1,520	775	1,880	824	1,172	711	20,773	
323	550	265	634	309	381	238	7,479	
650	970	509	1,246	515	790	473	13,295	
129	163	117	185	70	134	59	1,653	

Source: https://www2.census.gov/programs-surveys/cps/tables/pinc-01/2017/pinc01_1_1_1.xls

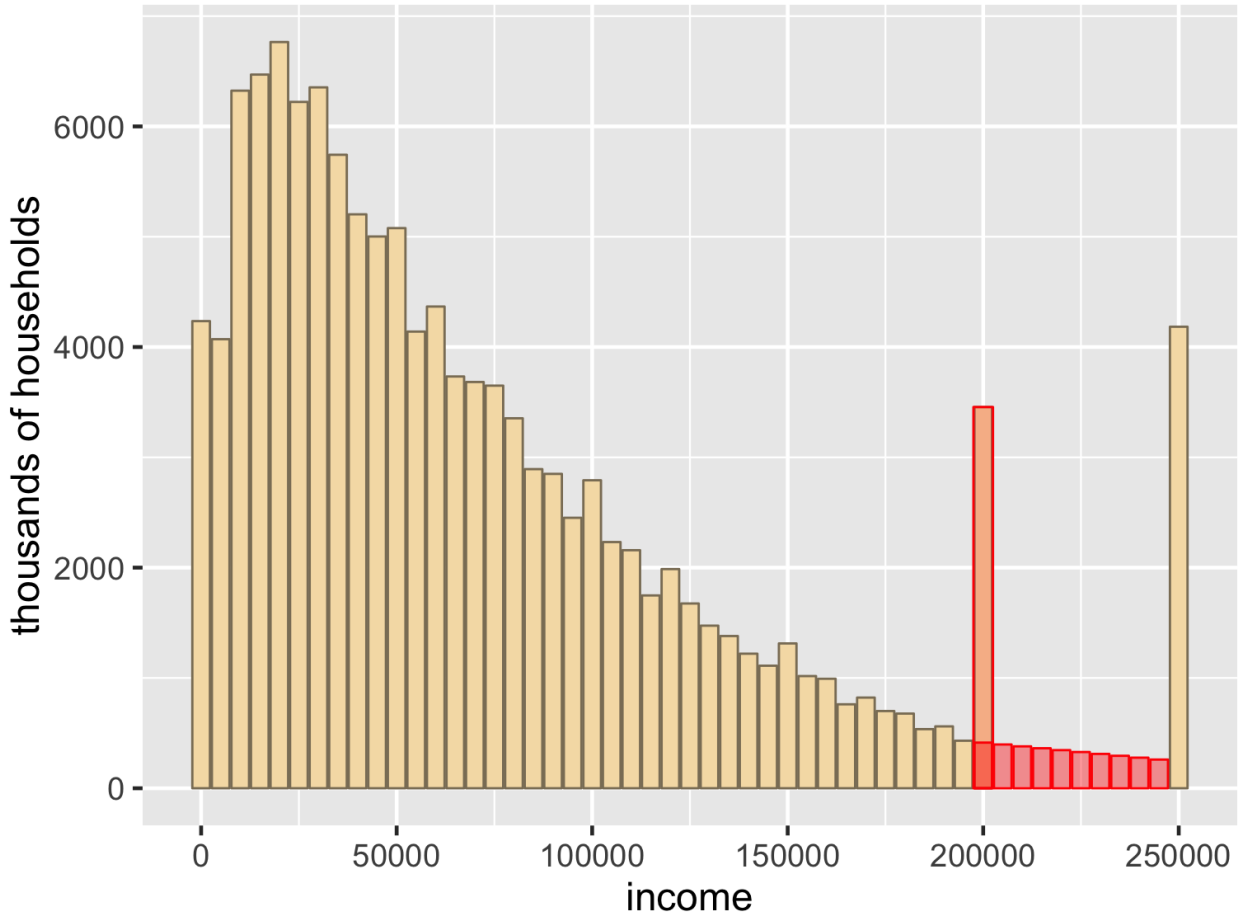
Household Income in 2015



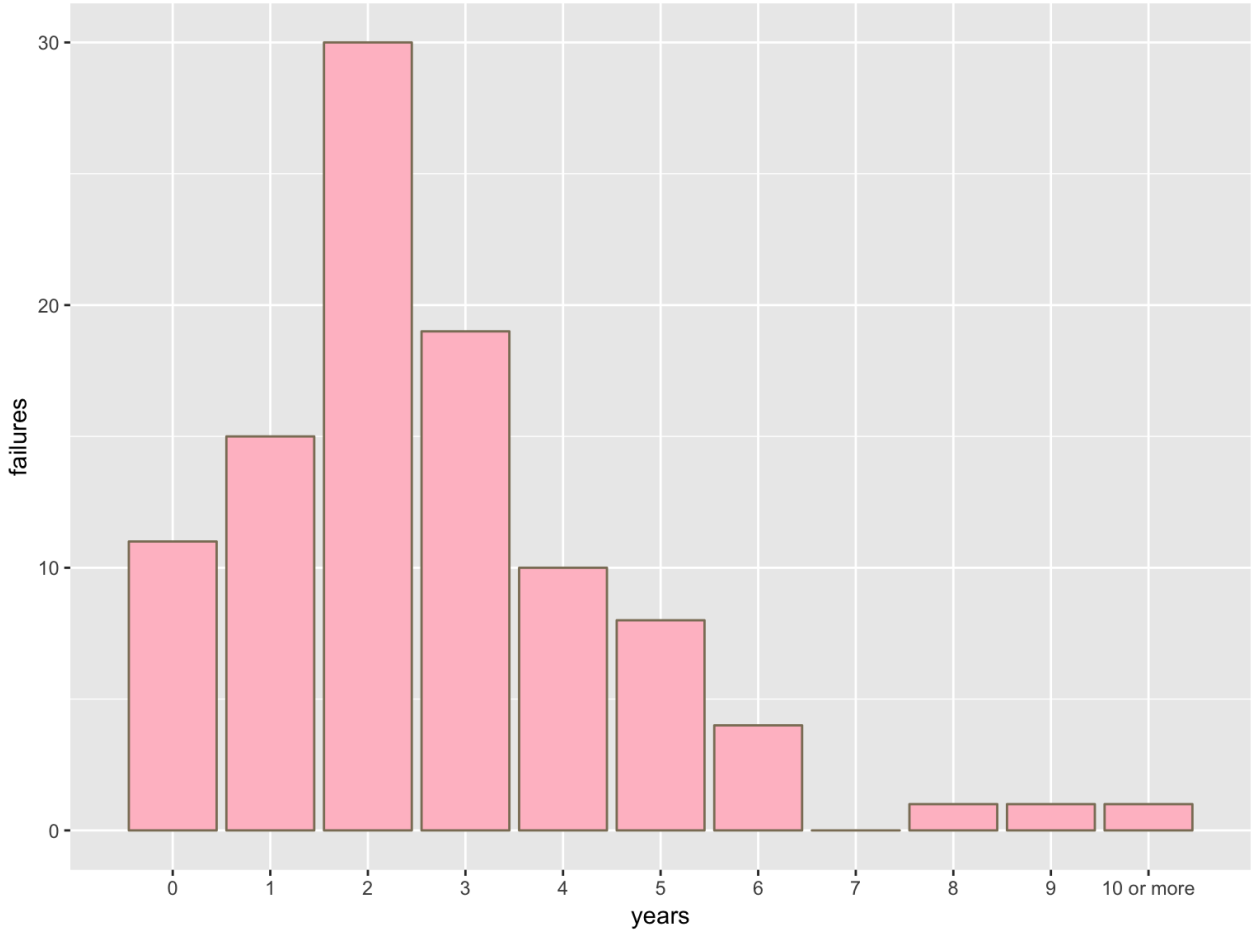
Household Income in 2015



Household Income in 2015



Reasonable use of “or more”



Data cleaning / transforming

http://toddwschneider.com/posts/the-simpsons-by-the-data/