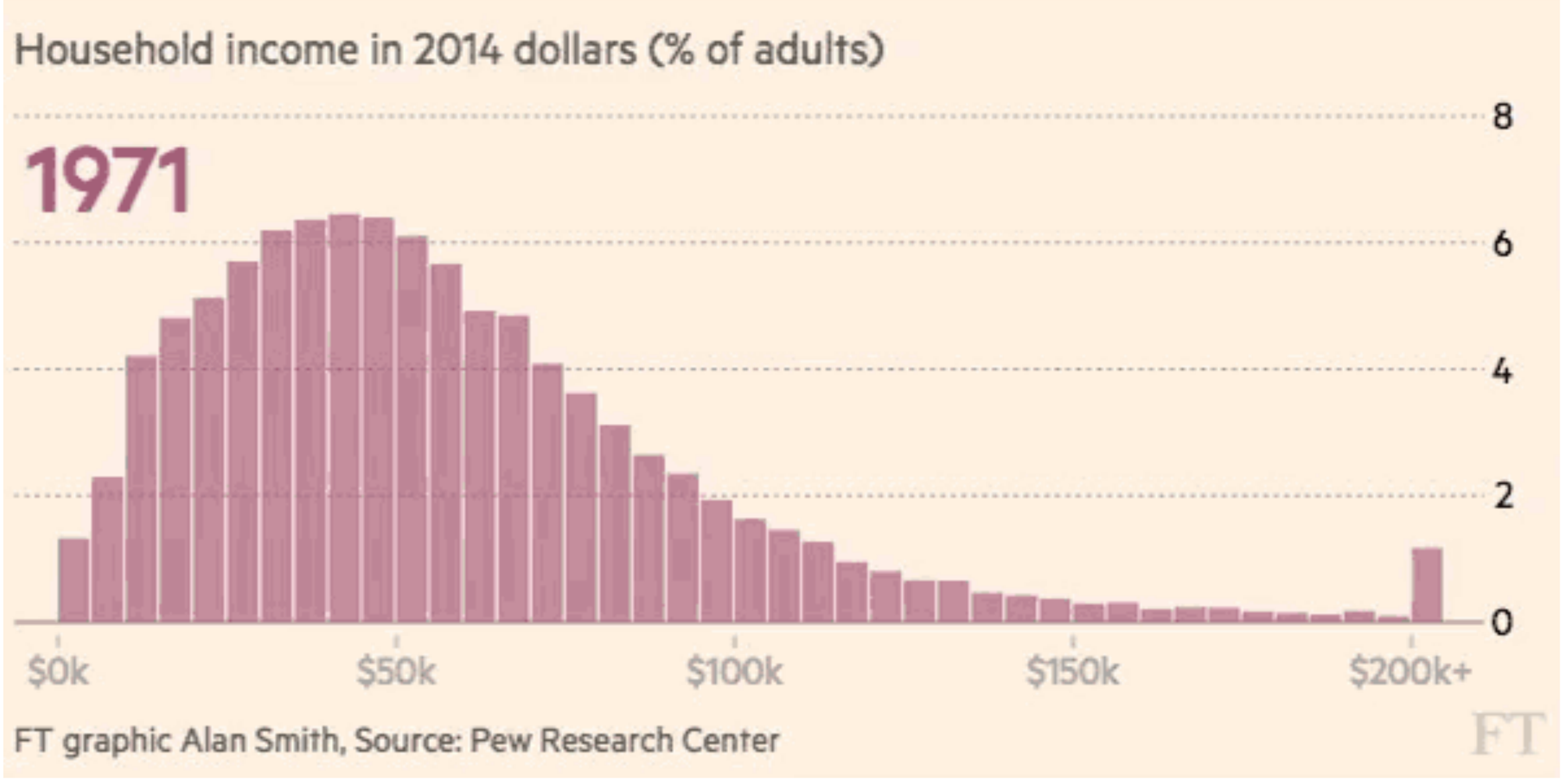
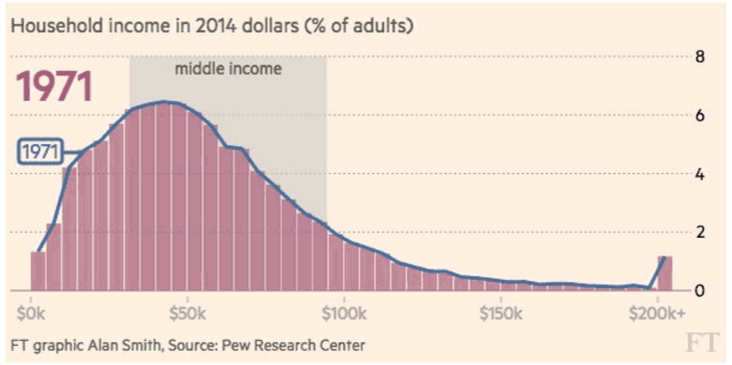
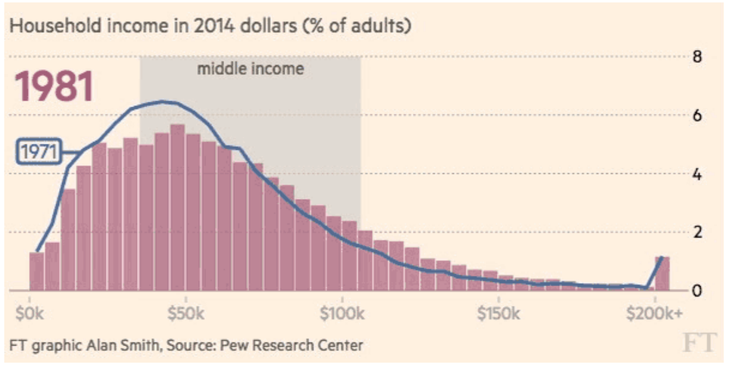
GR5702
Exploratory Data Analysis and Visualization

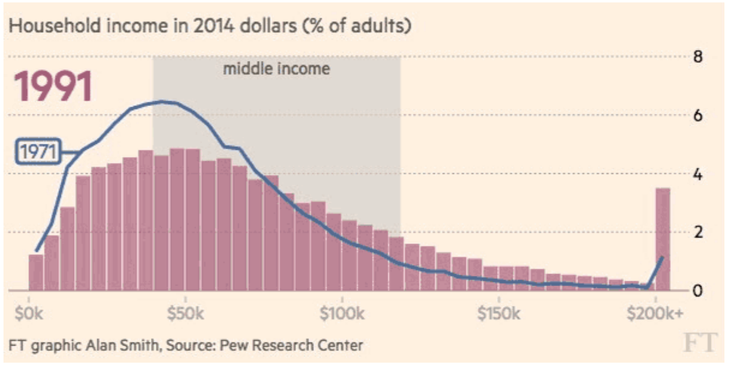
Prof. Joyce Robbins

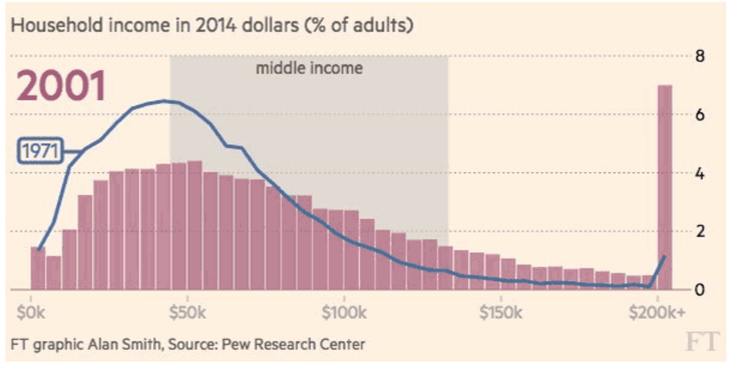


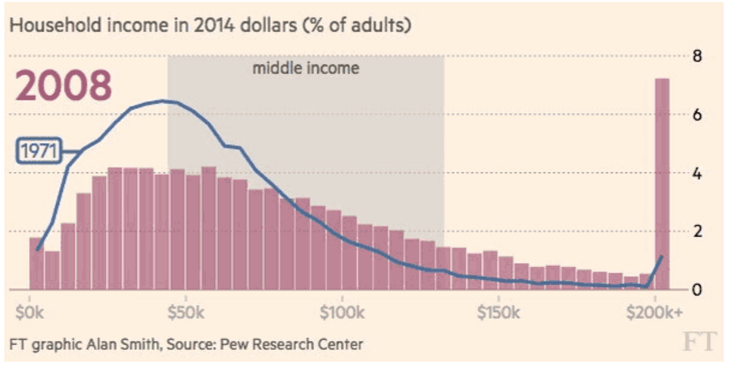
http://www.latimes.com/business/hiltzik/la-fi-hiltzik-ft-graphic-20160320-snap-htmlstory.html

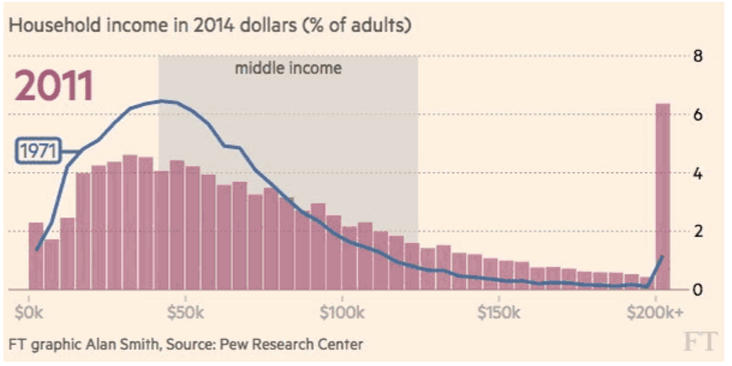

Household income in 2014 dollars (% of adults)
1971
middle income
1971
8
6
4
2
0
$0k
$50k
$100k
$150k
$200k+
FT graphic Alan Smith, Source: Pew Research Center
FT

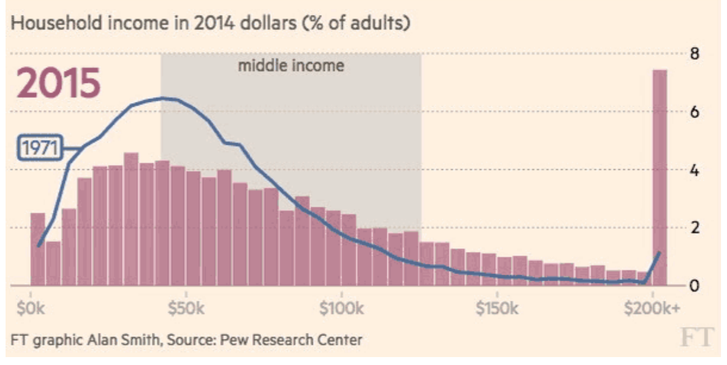

Household income in 2014 dollars (% of adults)
1981
middle income
1971
8
6
4
2
0
$0k
$50k
$100k
$150k
$200k+
FT graphic Alan Smith, Source: Pew Research Center
FT

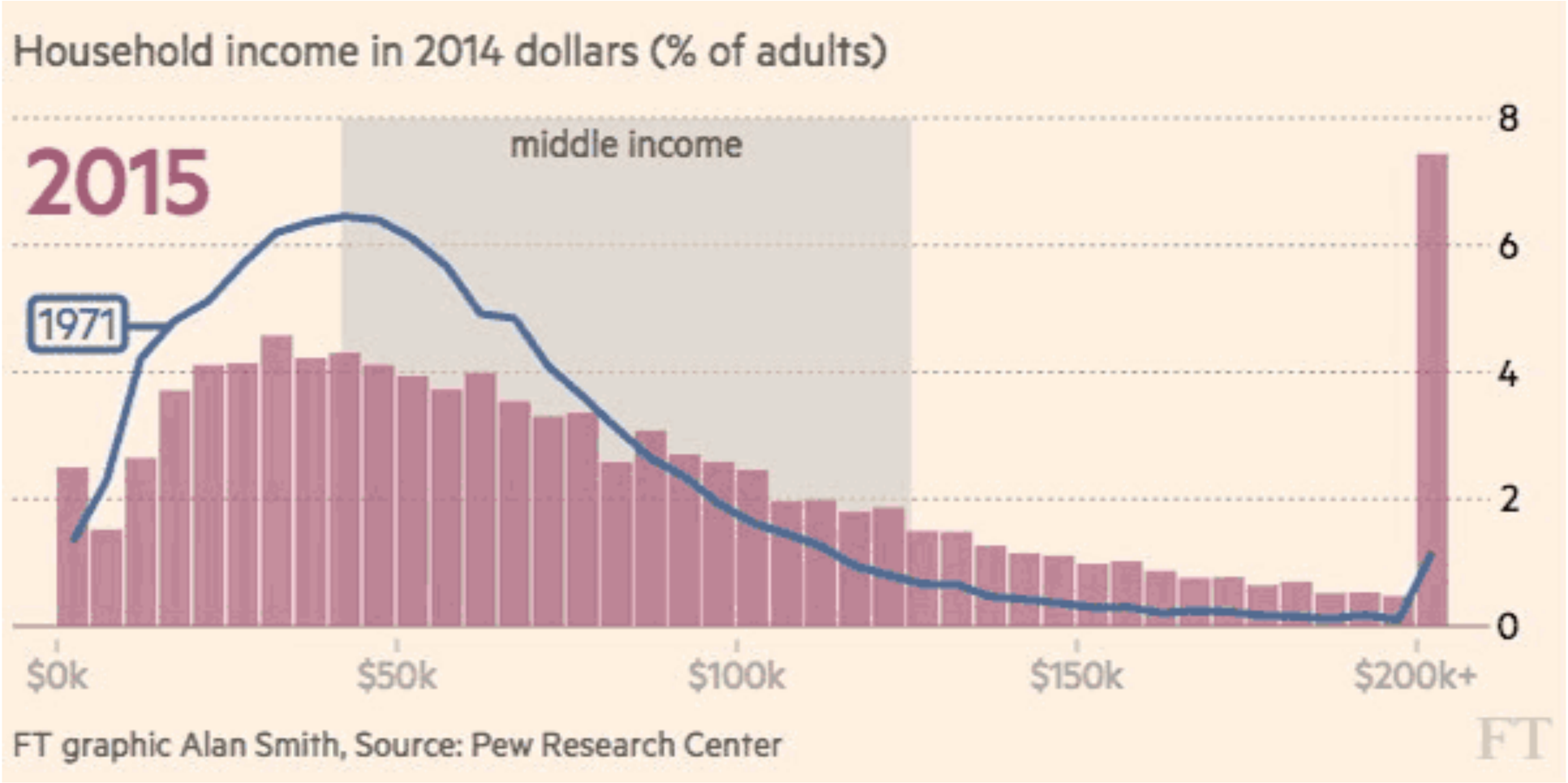

Household income in 2014 dollars (% of adults)
1991
middle income
1971
8
6
4
2
0
$0k
$50k
$100k
$150k
$200k+
FT graphic Alan Smith, Source: Pew Research Center
FT


Household income in 2014 dollars (% of adults)
2001
middle income
1971
8
6
4
2
0
$0k
$50k
$100k
$150k
$200k+
FT graphic Alan Smith, Source: Pew Research Center
FT


Household income in 2014 dollars (% of adults)
2008
middle income
1971
8
6
4
2
0
$0k
$50k
$100k
$150k
$200k+
FT graphic Alan Smith, Source: Pew Research Center
FT


Household income in 2014 dollars (% of adults)
2011
middle income
1971
8
6
4
2
0
$0k
$50k
$100k
$150k
$200k+
FT graphic Alan Smith, Source: Pew Research Center
FT


Household income in 2014 dollars (% of adults)
2015
middle income
1971
8
6
4
2
0
$0k
$50k
$100k
$150k
$200k+
FT graphic Alan Smith, Source: Pew Research Center
FT


Household income in 2014 dollars (% of adults)
2015
middle income
1971
8
6
4
2
0
$0k
$50k
$100k
$150k
$200k+
FT graphic Alan Smith, Source: Pew Research Center
FT

Tracking down data sources

- Pew Research Center

  http://www.pewsocialtrends.org/2015/12/09/the-american-middle-class-is-losing-ground/

  http://www.pewsocialtrends.org/2015/12/09/methodology/

- *Income and Poverty in the United States*, 2014 http://www.census.gov/content/dam/Census/library/publications/2015/demo/p60-252.pdf

  80 pages, no $5000 bin breakdowns

Current Population Survey (CPS), Annual Social and Economic Supplements (ASEC)

- http://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-06.html#.html

(354 pages)

A3 | For information on confidentiality protection, sampling error, nonsampling error, and definitions, see www2.census.gov/programs-surveys/cps/techdocs/cpsmar16.pdf

Table HINC-06. Income Distribution to $250,000 or More for Households: 2015

For information on confidentiality protection, sampling error, nonsampling error, and definitions, see www2.census.gov/programs-surveys/cps/techdocs/cpsmar16.pdf

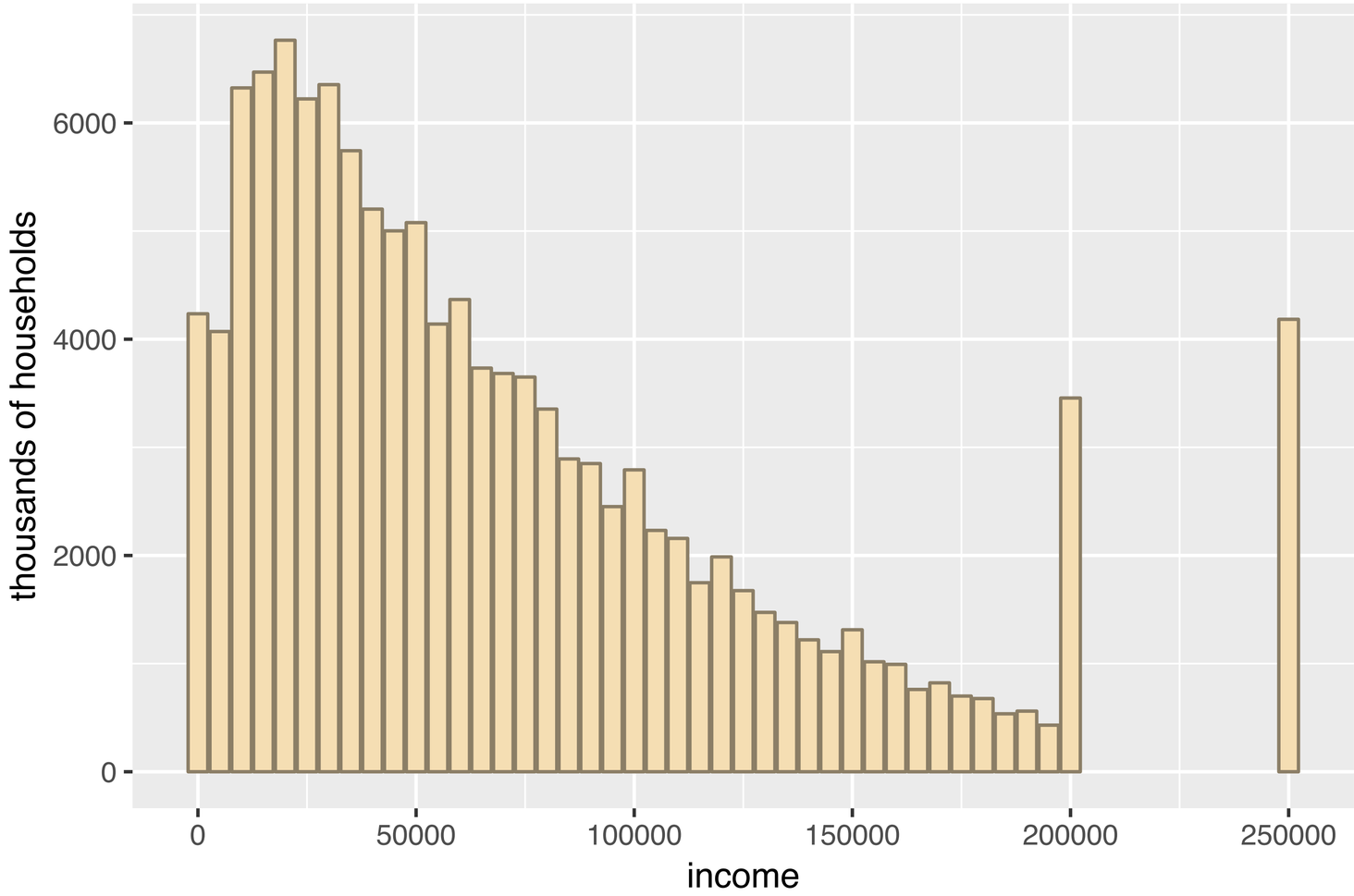
Source: U.S. Census Bureau, Current Population Survey, 2016 Annual Social and Economic Supplement.
(Numbers in thousands. Households as of March of the following year. A.O.I.C. stands for alone or in combination. Standard errors calculated using replicate weights)

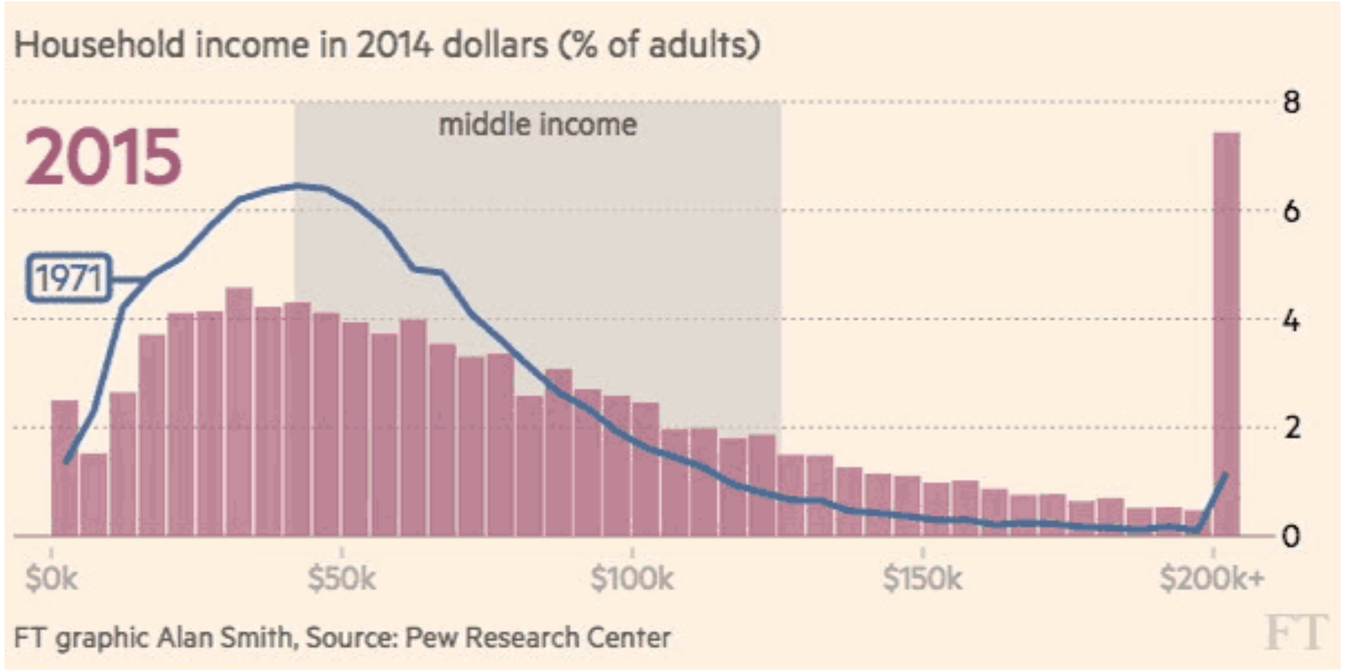
Income of Household	All Races			White A.O.I.C.			White alone (1)		
		Mean Income			Mean Income			Mean Inc	
	Number	Dollars	Standard Error	Number	Dollars	Standard Error	Number	Dollars	St
Total	125,819	79,263	403	100,990	82,052	457	99,313	82,226	
Under $5,000	4,235	1,080	40	2,796	1,094	50	2,709	1,097	
$5,000 to $9,999	4,071	8,018	31	2,647	8,033	40	2,569	8,035	
$10,000 to $14,999	6,324	12,397	30	4,646	12,461	34	4,561	12,467	
$15,000 to $19,999	6,470	17,297	29	4,922	17,323	36	4,813	17,327	
$20,000 to $24,999	6,765	22,199	28	5,290	22,229	33	5,202	22,225	
$25,000 to $29,999	6,222	27,116	30	4,958	27,111	33	4,877	27,112	
$30,000 to $34,999	6,354	32,027	31	5,024	32,069	35	4,949	32,070	
$35,000 to $39,999	5,743	37,115	28	4,620	37,142	32	4,531	37,139	

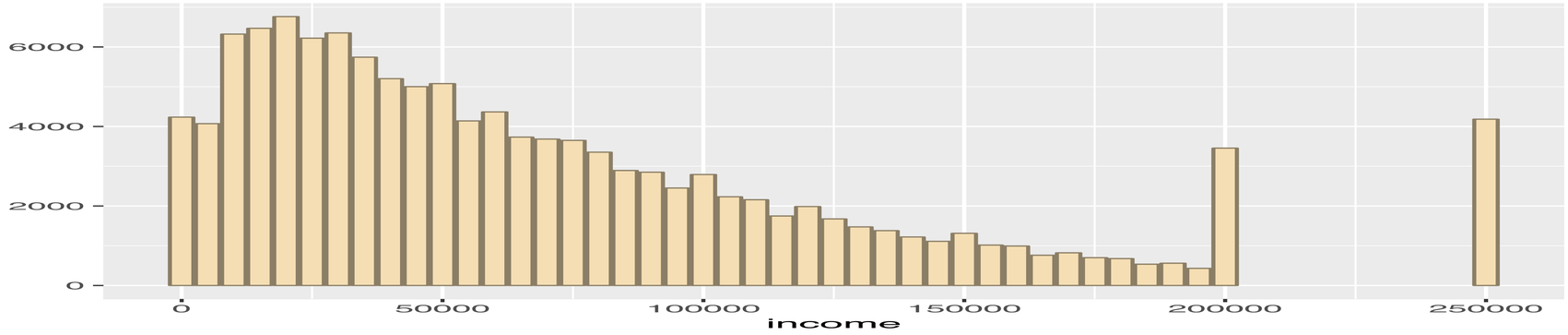
Current Population Survey (CPS), Annual Social and Economic Supplements (ASEC)

	A	B	C	D	E	F
34	$120,000 to $124,999	1,987	121,965	54	1,685	121,999
35	$125,000 to $129,999	1,675	126,961	57	1,462	126,974
36	$130,000 to $134,999	1,474	131,990	68	1,260	131,955
37	$135,000 to $139,999	1,380	137,101	66	1,168	137,161
38	$140,000 to $144,999	1,220	141,979	70	1,032	141,976
39	$145,000 to $149,999	1,111	147,145	73	910	147,143
40	$150,000 to $154,999	1,313	151,768	61	1,072	151,759
41	$155,000 to $159,999	1,017	157,137	76	854	157,101
42	$160,000 to $164,999	993	162,080	88	825	162,123
43	$165,000 to $169,999	761	167,171	80	652	167,144
44	$170,000 to $174,999	822	172,199	88	728	172,228
45	$175,000 to $179,999	700	177,157	99	608	177,146
46	$180,000 to $184,999	677	182,009	78	570	182,056
47	$185,000 to $189,999	536	187,224	94	446	187,219
48	$190,000 to $194,999	561	191,951	97	455	191,926
49	$195,000 to $199,999	431	197,339	104	381	197,364
50	$200,000 to $249,999	3,456	220,874	410	2,906	220,991
51	$250,000 and over	4,184	411,551	5,980	3,512	410,189

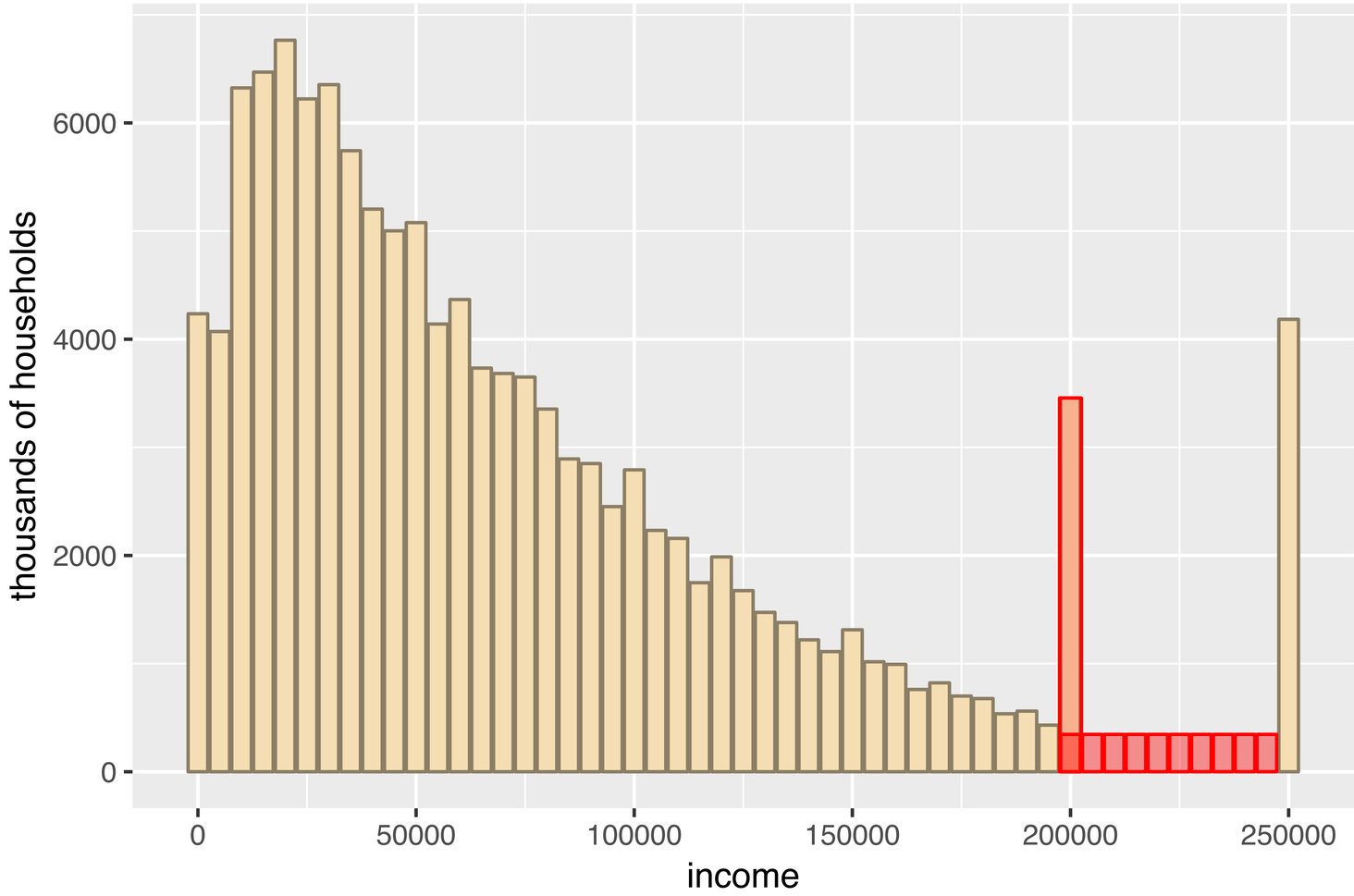
Household Income in 2015



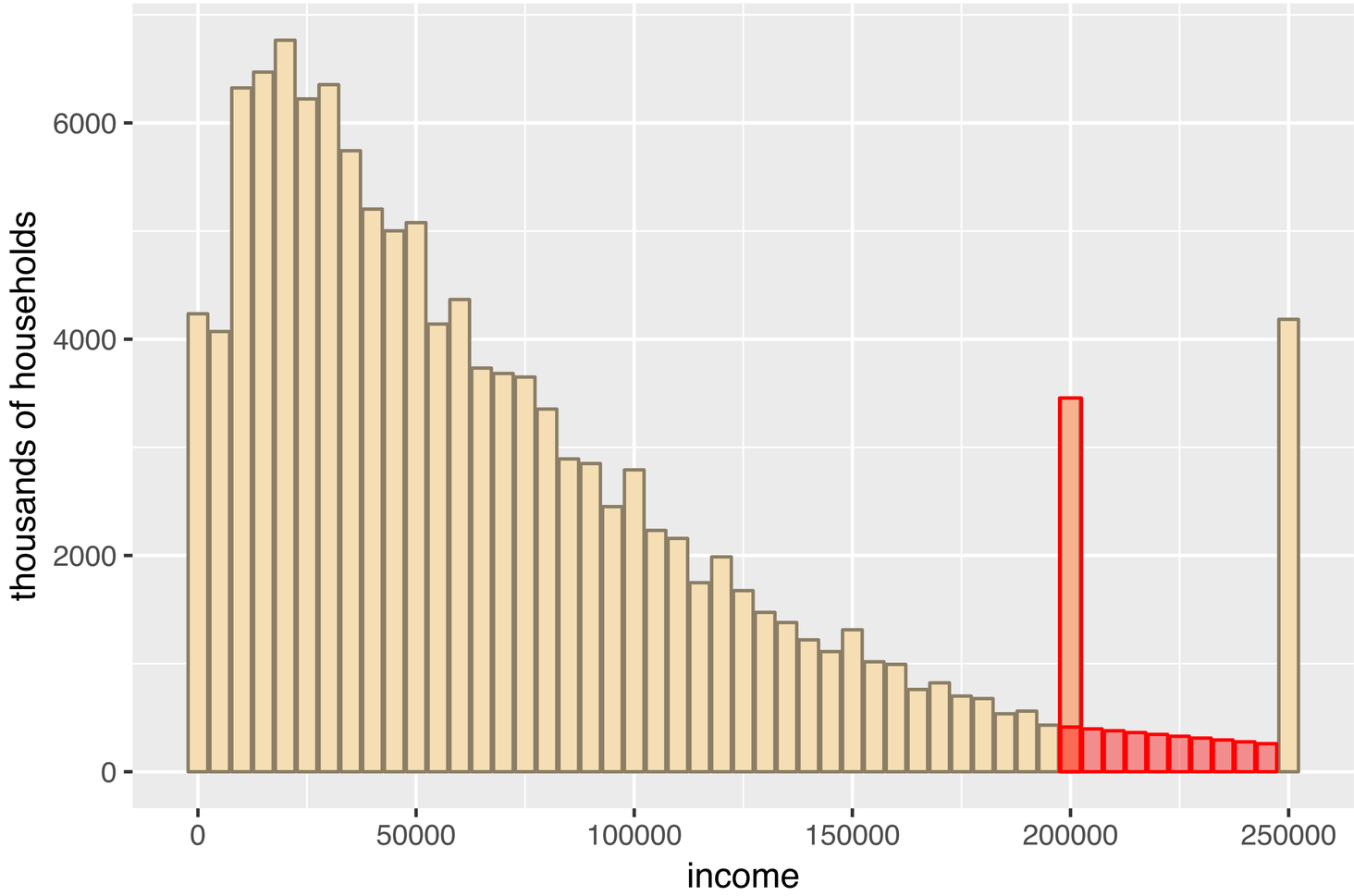

Household income in 2014 dollars (% of adults)
2015
middle income
1971
8
6
4
2
0
$0k
$50k
$100k
$150k
$200k+
FT graphic Alan Smith, Source: Pew Research Center
FT


6000
4000
2000
0
0
50000
100000
150000
200000
250000
income

Household Income in 2015



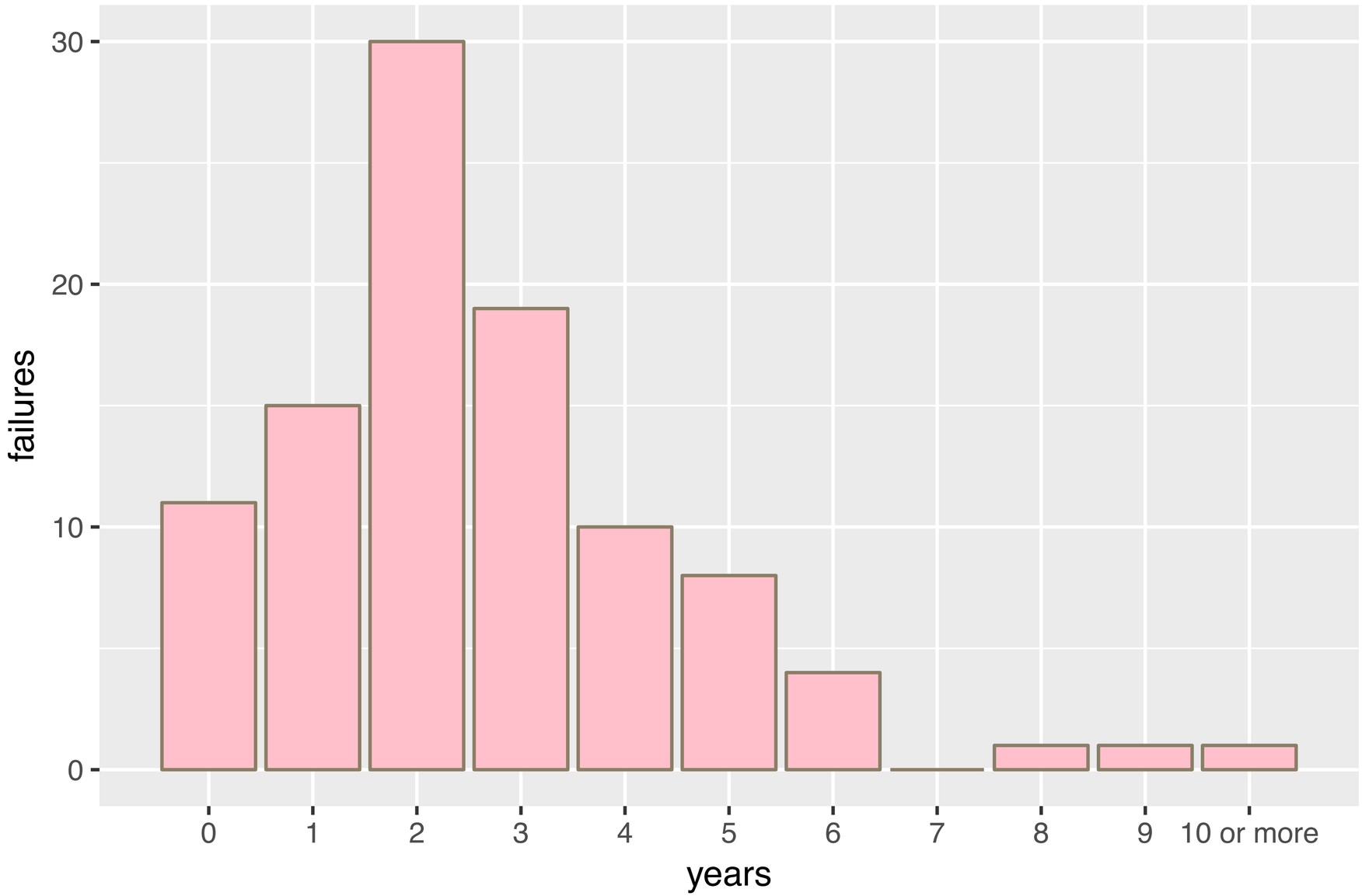
Household Income in 2015



Sources of Data on Income Distribution in the U.S.

- U.S. Census *Current Population Survey*
- U.S. Department of Labor, Bureau of Labor Statistics
- Internal Revenue Service
- Federal Reserve

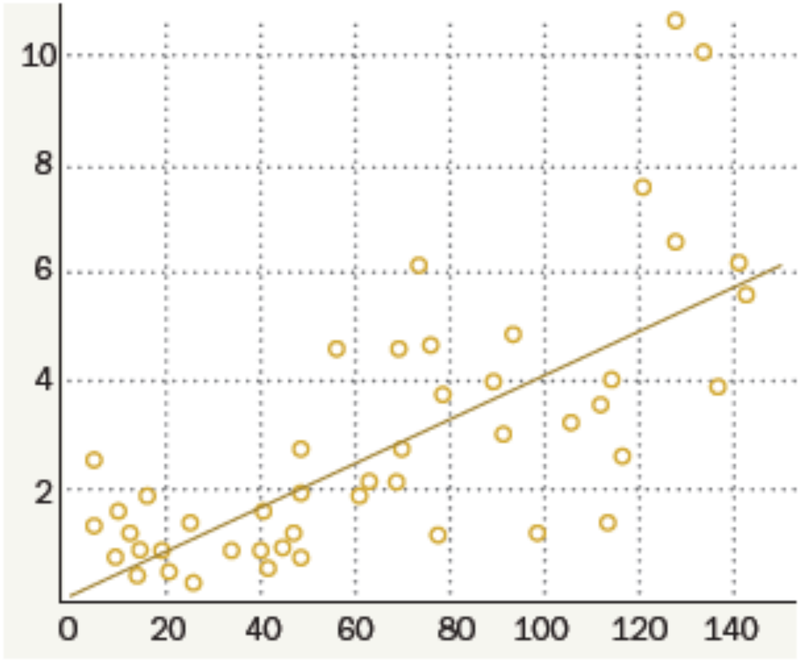
  https://www.federalreserve.gov/pubs/bulletin/2014/pdf/scf14.pdf


30
20
10
0
failures
0
1
2
3
4
5
6
7
8
9
10 or more
years

63% of American Adults Can Correctly Read This Chart

Which of the following statements best describes the data in the graph below?

Average number of decayed teeth per person in different countries



Average sugar consumption (grams per person per day)

A. In recent years, the rate of cavities has increased in many countries

B. In some countries, people brush their teeth more frequently than in other countries

C. The more sugar people eat, the more likely they are to get cavities

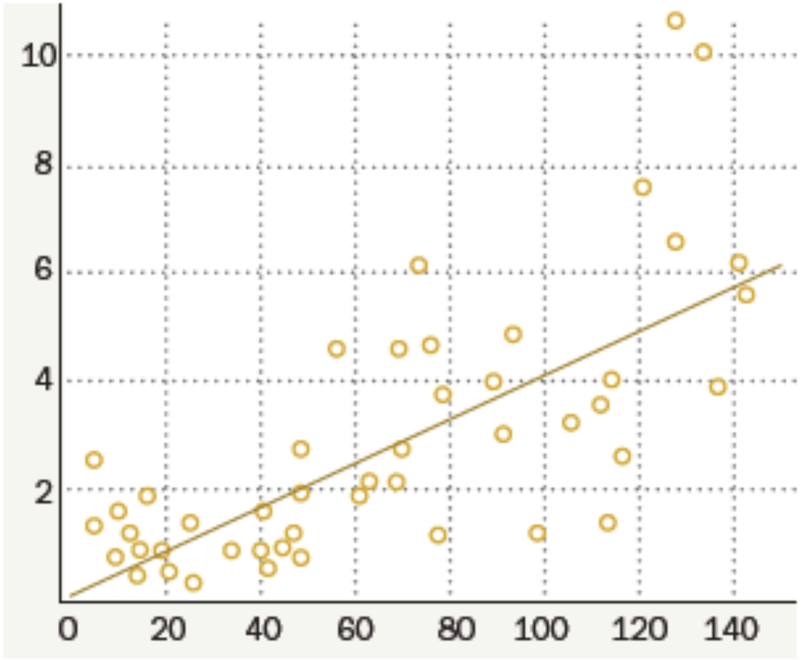
D. In recent years, the consumption of sugar has increased in many countries.

http://www.pewresearch.org/fact-tank/2015/09/16/the-art-and-science-of-the-scatterplot/ft_15-09-11_scatterplot/

63% of American Adults Can Correctly Read This Chart

Which of the following statements best describes the data in the graph below?

Average number of decayed teeth per person in different countries



A. In recent years, the rate of cavities has increased in many countries

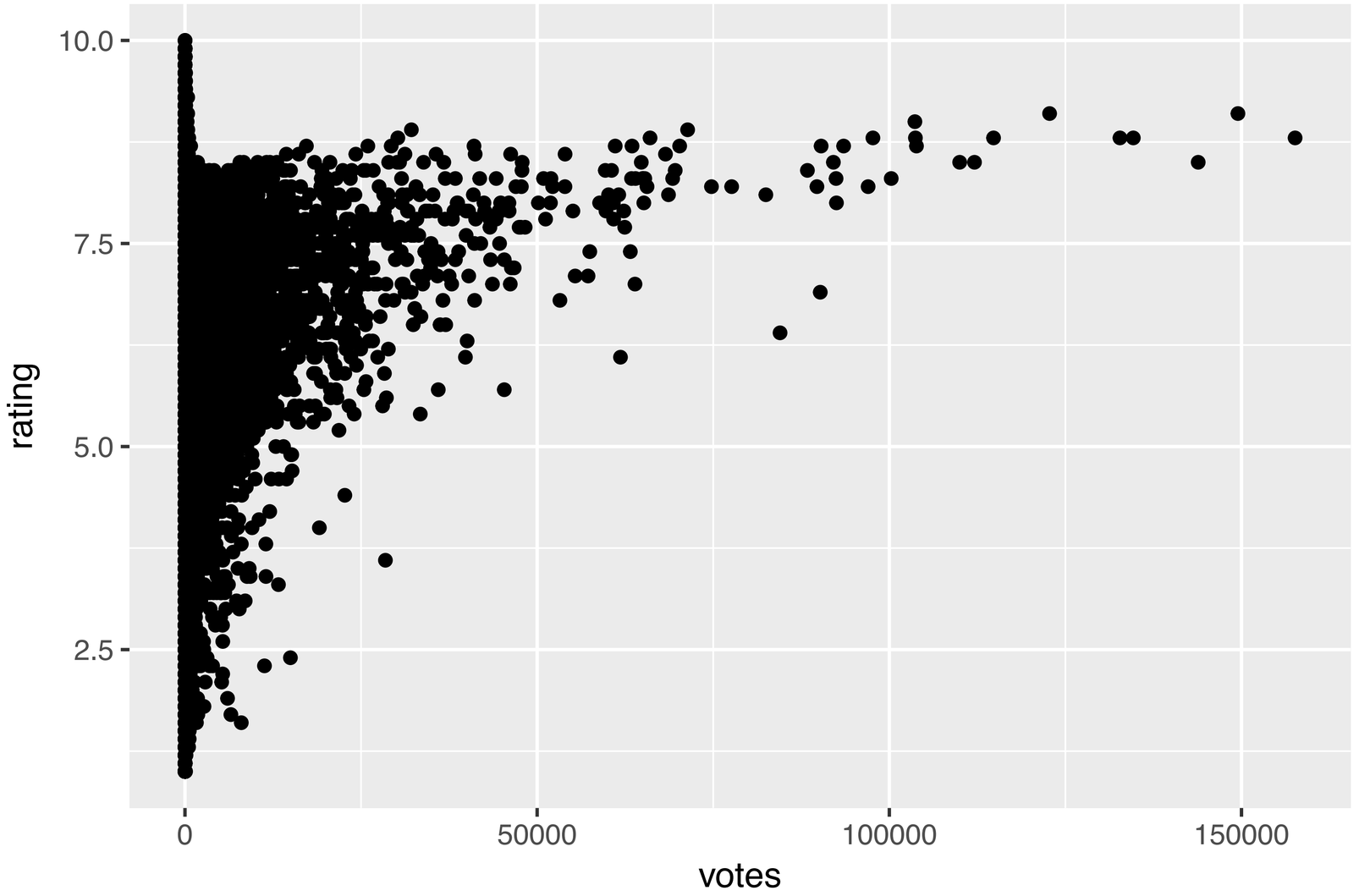
B. In some countries, people brush their teeth more frequently than in other countries

C. The more sugar people eat, the more likely they are to get cavities

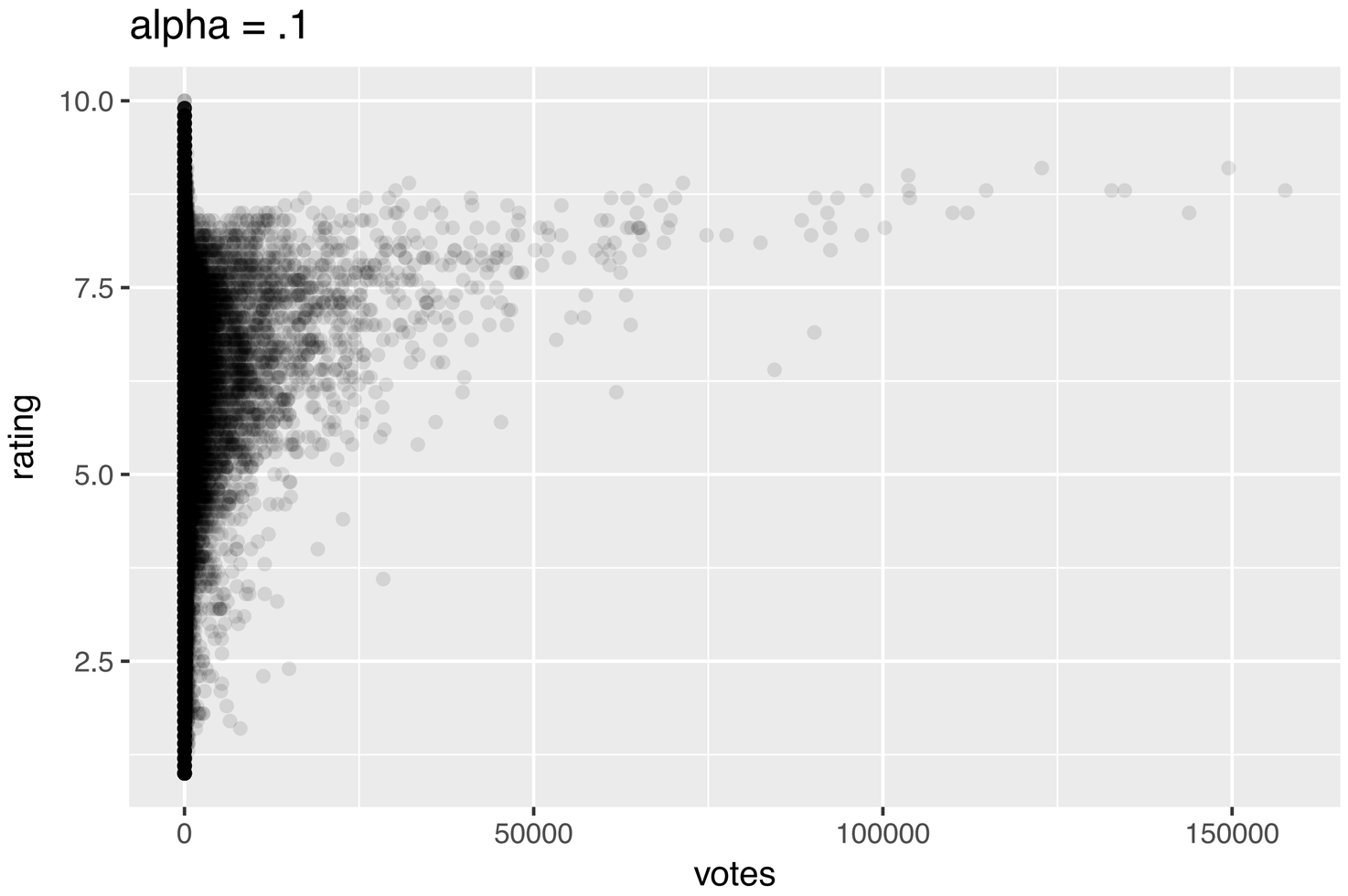
D. In recent years, the consumption of sugar has increased in many countries.

http://www.pewresearch.org/fact-tank/2015/09/16/the-art-and-science-of-the-scatterplot/ft_15-09-11_scatterplot/

```
library(ggplot2)
library(ggplot2movies)
ggplot(movies, aes(votes, rating)) + geom_point()
```

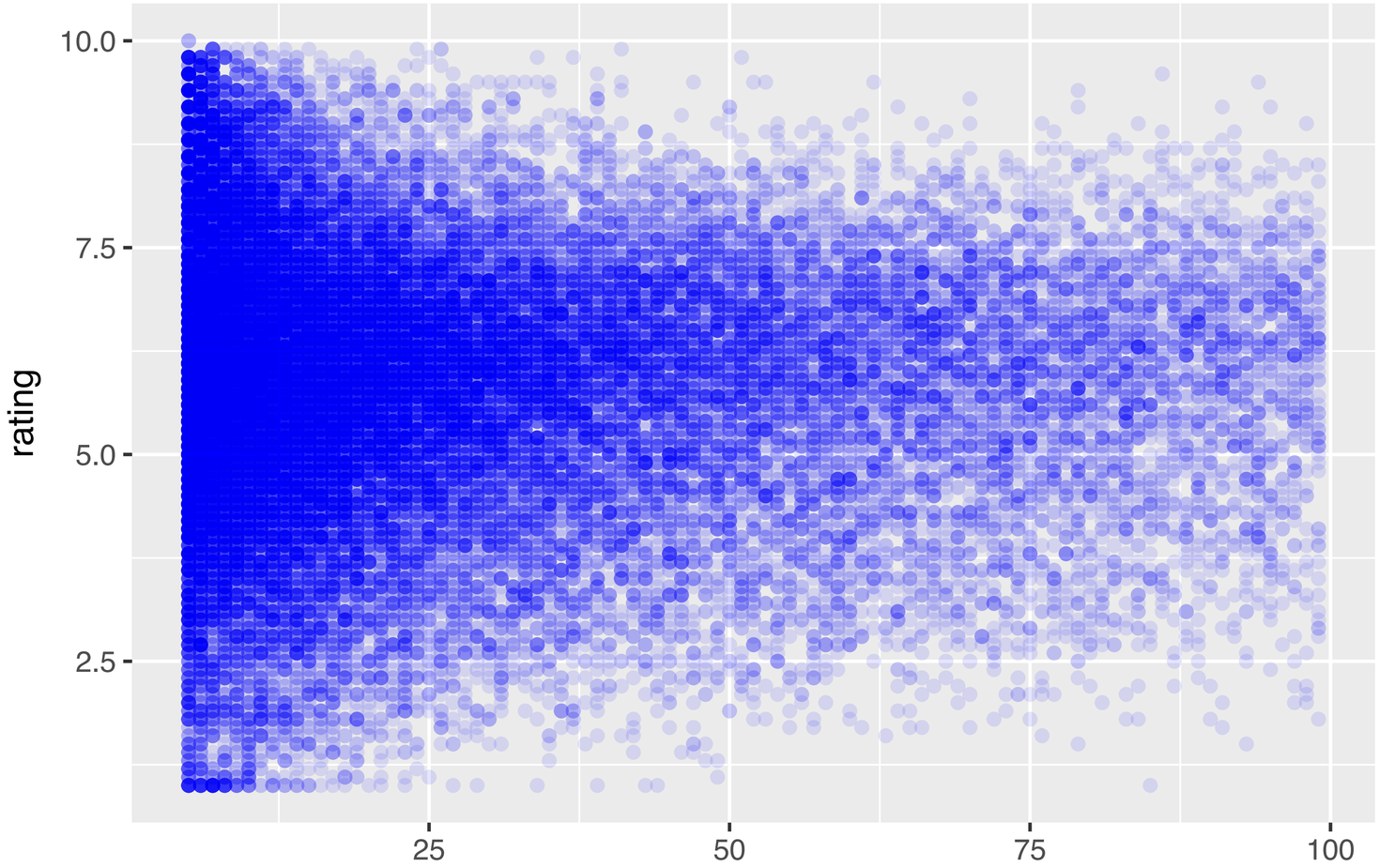


```
ggplot(movies, aes(votes, rating)) + geom_point(alpha = .1) +
    ggtitle("alpha = .1")
```



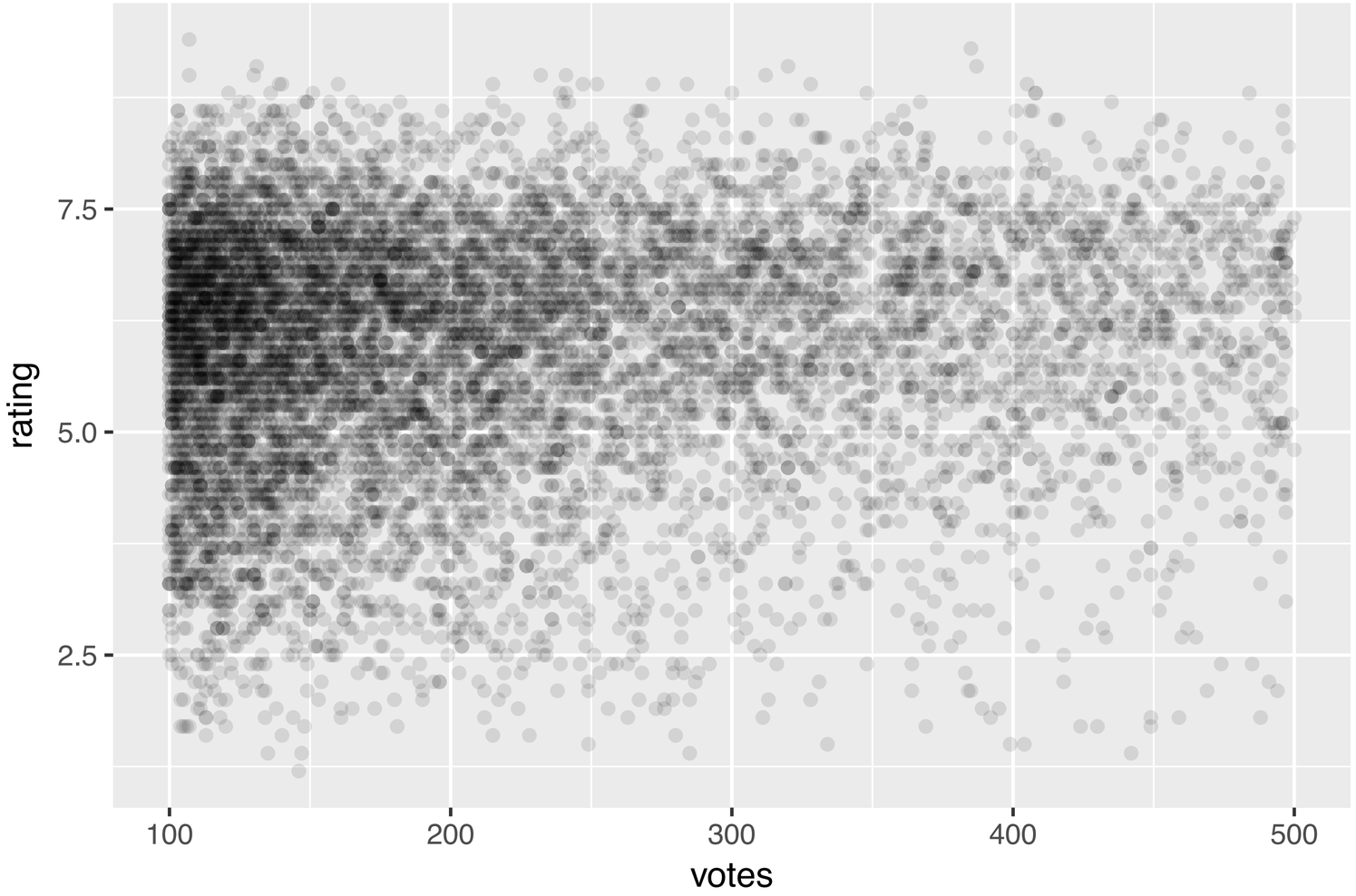
Movies with fewer than 100 votes

```
movies100 <- movies %>% filter(votes < 100)
ggplot(movies100, aes(votes, rating)) +
    geom_point(alpha = .1, color = "blue")
```



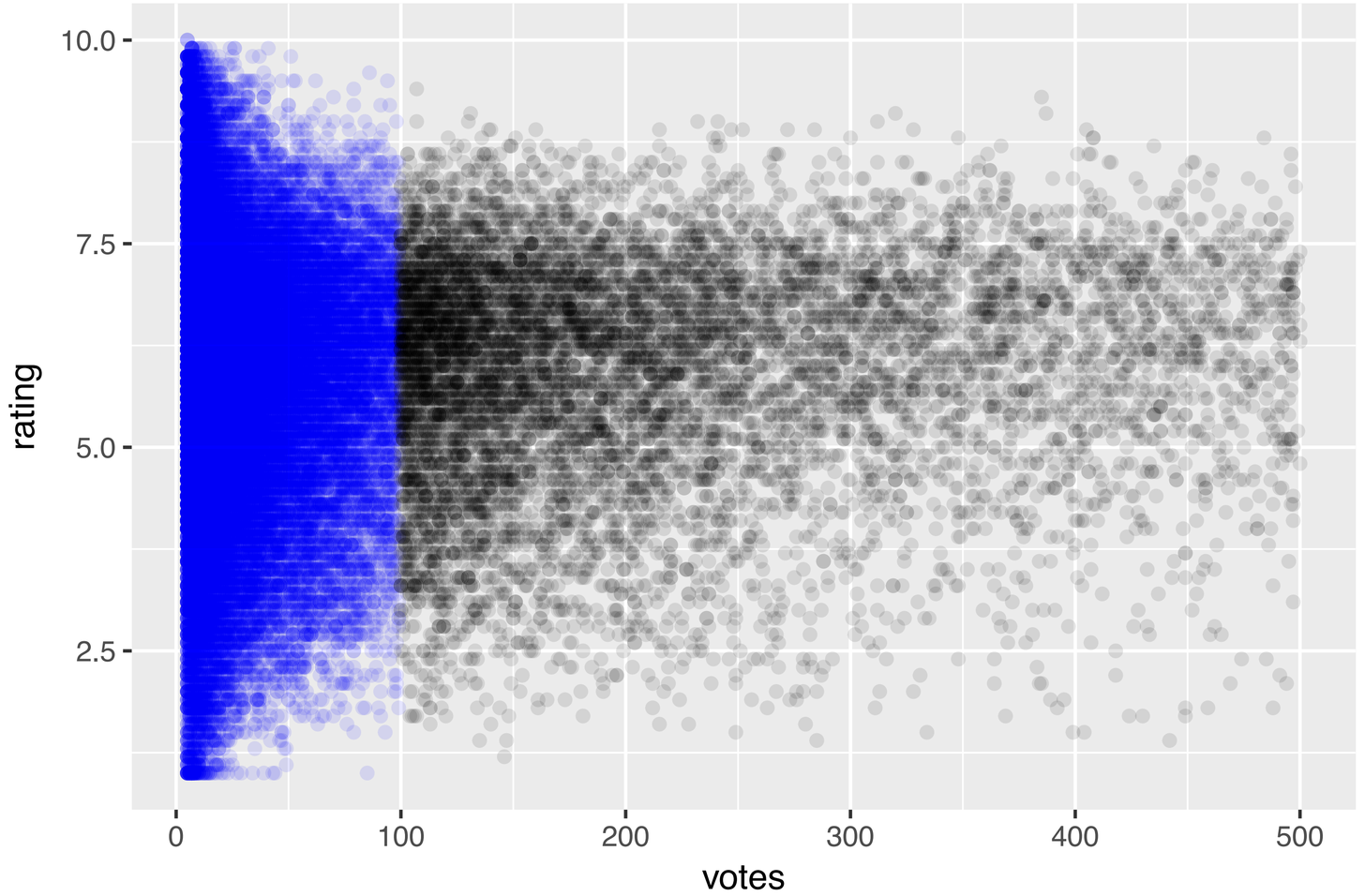
Movies with 100 to 500 votes

```
movies100to500 <- movies %>% filter(votes >= 100 & votes <=500 )
ggplot(movies100to500, aes(votes, rating)) + geom_point(alpha =
```

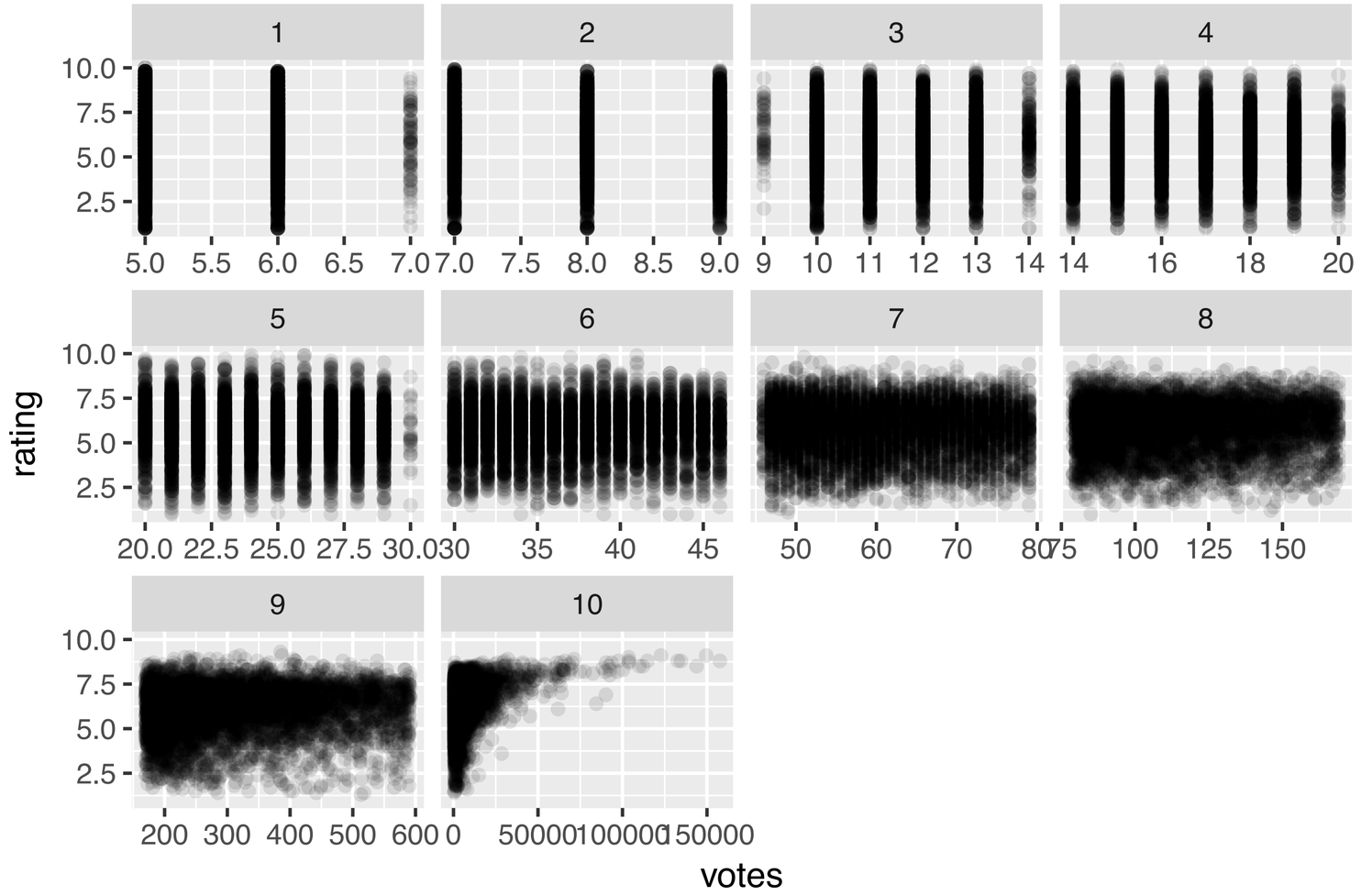


```
ggplot(movies100, aes(votes, rating)) +
    geom_point(alpha = .1, color = "blue") +
    geom_point(data = movies100to500, aes(votes, rating),
               alpha = .1)
```

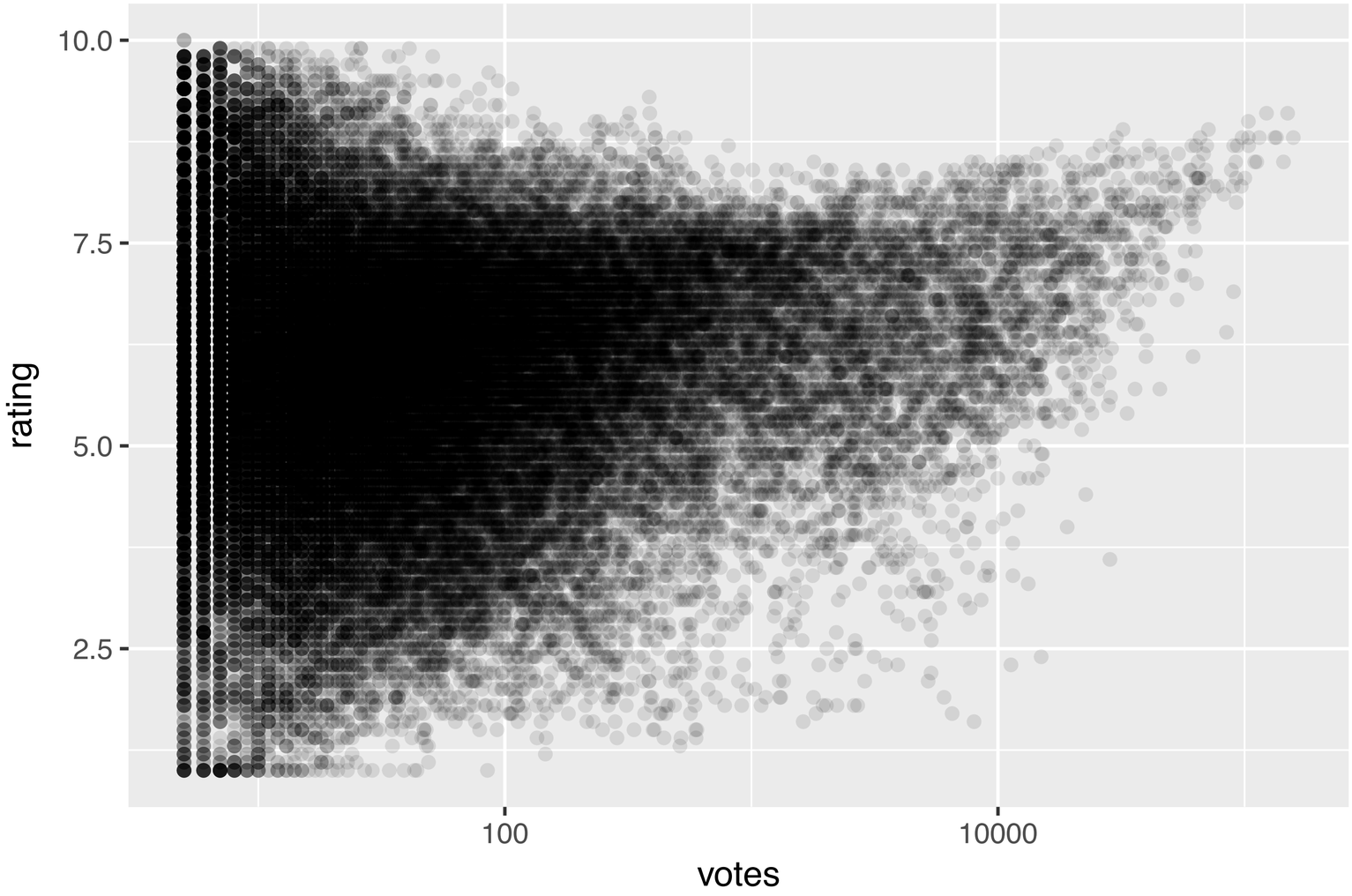
Movies with 500 votes or fewer



```
binnedmovies <- movies %>% mutate(mybin = ntile(votes, 10))
ggplot(binnedmovies, aes(votes, rating)) + geom_point(alpha = .1
    facet_wrap(~mybin, scales = "free_x")
```



```
ggplot(movies, aes(votes, rating)) + geom_point(alpha = .1) +
    scale_x_log10()
```



str(movies)

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    58788 obs. of  2
##  $ title      : chr  "$" "$1000 a Touchdown" "$21 a Day Once
##  $ year       : int  1971 1939 1941 1996 1975 2000 2002 2002
##  $ length     : int  121 71 7 70 71 91 93 25 97 61 ...
##  $ budget     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ rating     : num  6.4 6 8.2 8.2 3.4 4.3 5.3 6.7 6.6 6 ...
##  $ votes      : int  348 20 5 6 17 45 200 24 18 51 ...
##  $ r1         : num  4.5 0 0 14.5 24.5 4.5 4.5 4.5 4.5 4.5 ..
##  $ r2         : num  4.5 14.5 0 0 4.5 4.5 0 4.5 4.5 0 ...
##  $ r3         : num  4.5 4.5 0 0 0 4.5 4.5 4.5 4.5 4.5 ...
##  $ r4         : num  4.5 24.5 0 0 14.5 14.5 4.5 4.5 0 4.5 ...
##  $ r5         : num  14.5 14.5 0 0 14.5 14.5 24.5 4.5 0 4.5 .
##  $ r6         : num  24.5 14.5 24.5 0 4.5 14.5 24.5 14.5 0 44
##  $ r7         : num  24.5 14.5 0 0 0 4.5 14.5 14.5 34.5 14.5
##  $ r8         : num  14.5 4.5 44.5 0 0 4.5 4.5 14.5 14.5 4.5
##  $ r9         : num  4.5 4.5 24.5 34.5 0 14.5 4.5 4.5 4.5 4.5
##  $ r10        : num  4.5 14.5 24.5 45.5 24.5 14.5 14.5 14.5 2
##  $ mpaa       : chr  "" "" "" "" ...
##  $ Action     : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Animation  : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ Comedy     : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ Drama      : int  1 0 0 0 0 1 1 0 1 0 ...
```

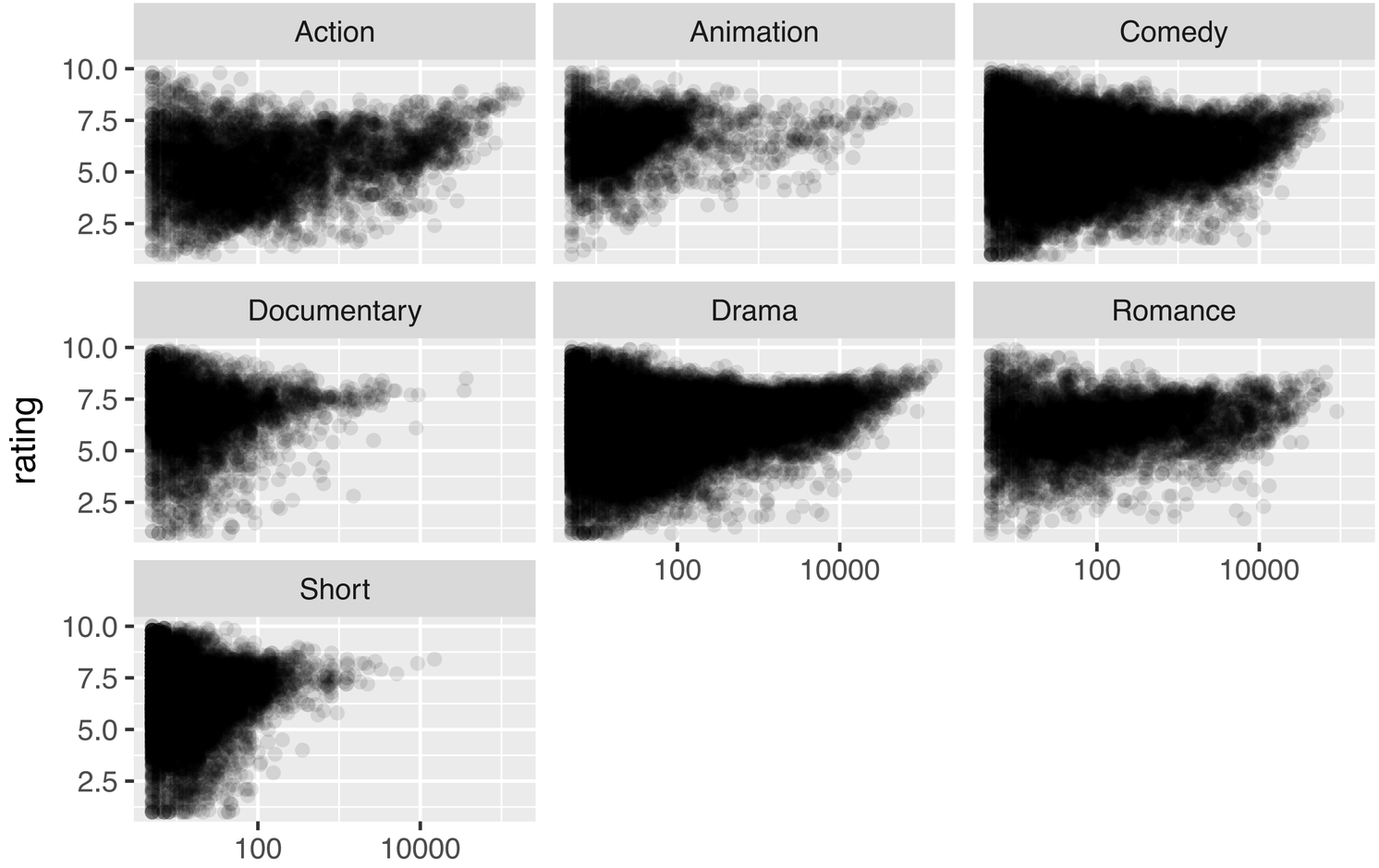
```r
cb <- movies %>% filter(title == "Casablanca") %>%
    select(-(r1:r10))
str(cb)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    2 obs. of
##  $ title       : chr  "Casablanca" "Casablanca"
##  $ year        : int  1942 2002
##  $ length      : int  102 14
##  $ budget      : int  950000 NA
##  $ rating      : num  8.8 4
##  $ votes       : int  66030 31
##  $ mpaa        : chr  "" ""
##  $ Action      : int  0 0
##  $ Animation   : int  0 0
##  $ Comedy      : int  0 0
##  $ Drama       : int  1 0
##  $ Documentary: int  0 0
##  $ Romance     : int  1 0
##  $ Short       : int  0 1
```

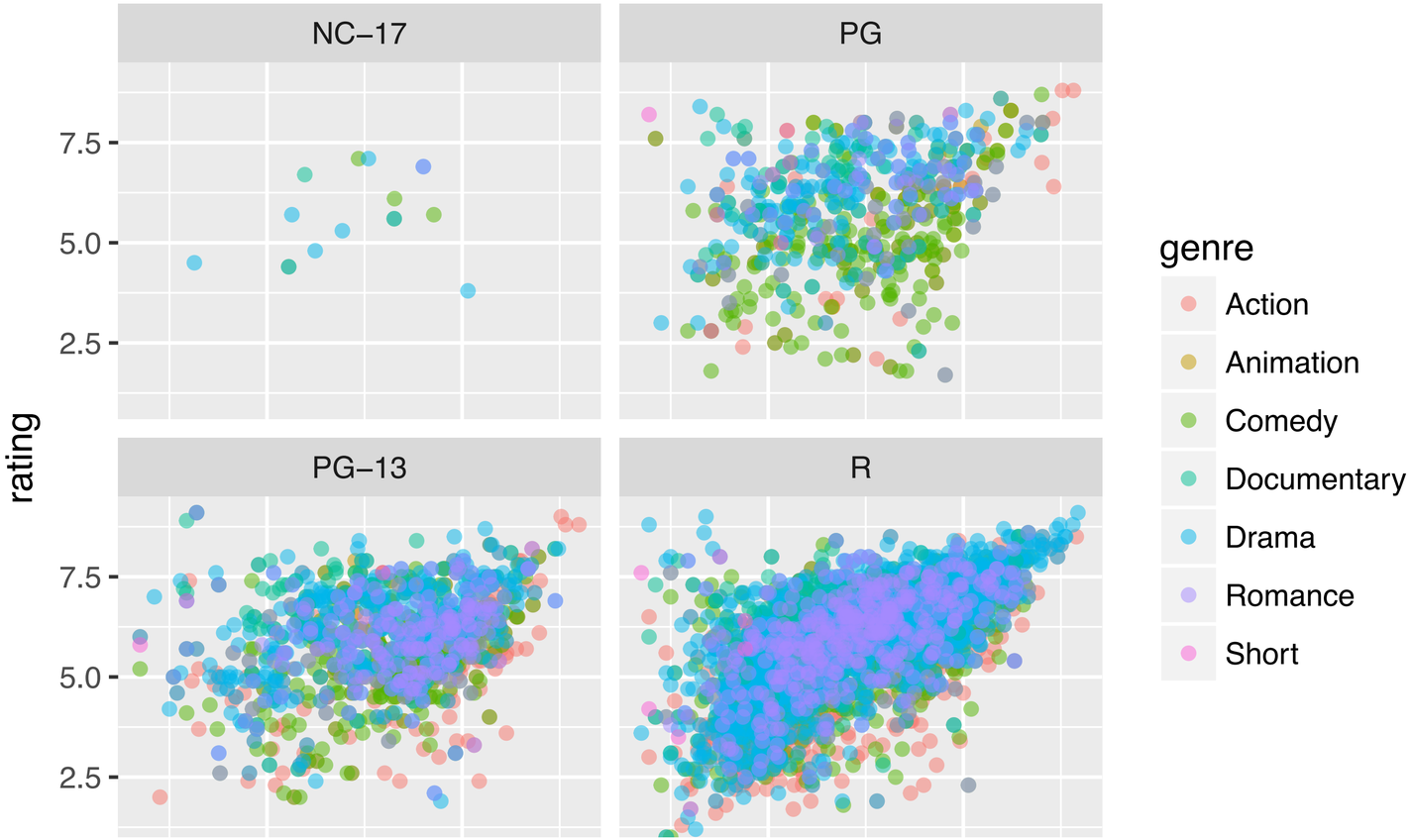
```
tidymovies %>% filter(title == "Casablanca") %>%
    select(title, year, genre) %>% kable()
```

title	year	genre
Casablanca	1942	Drama
Casablanca	1942	Romance
Casablanca	2002	Short

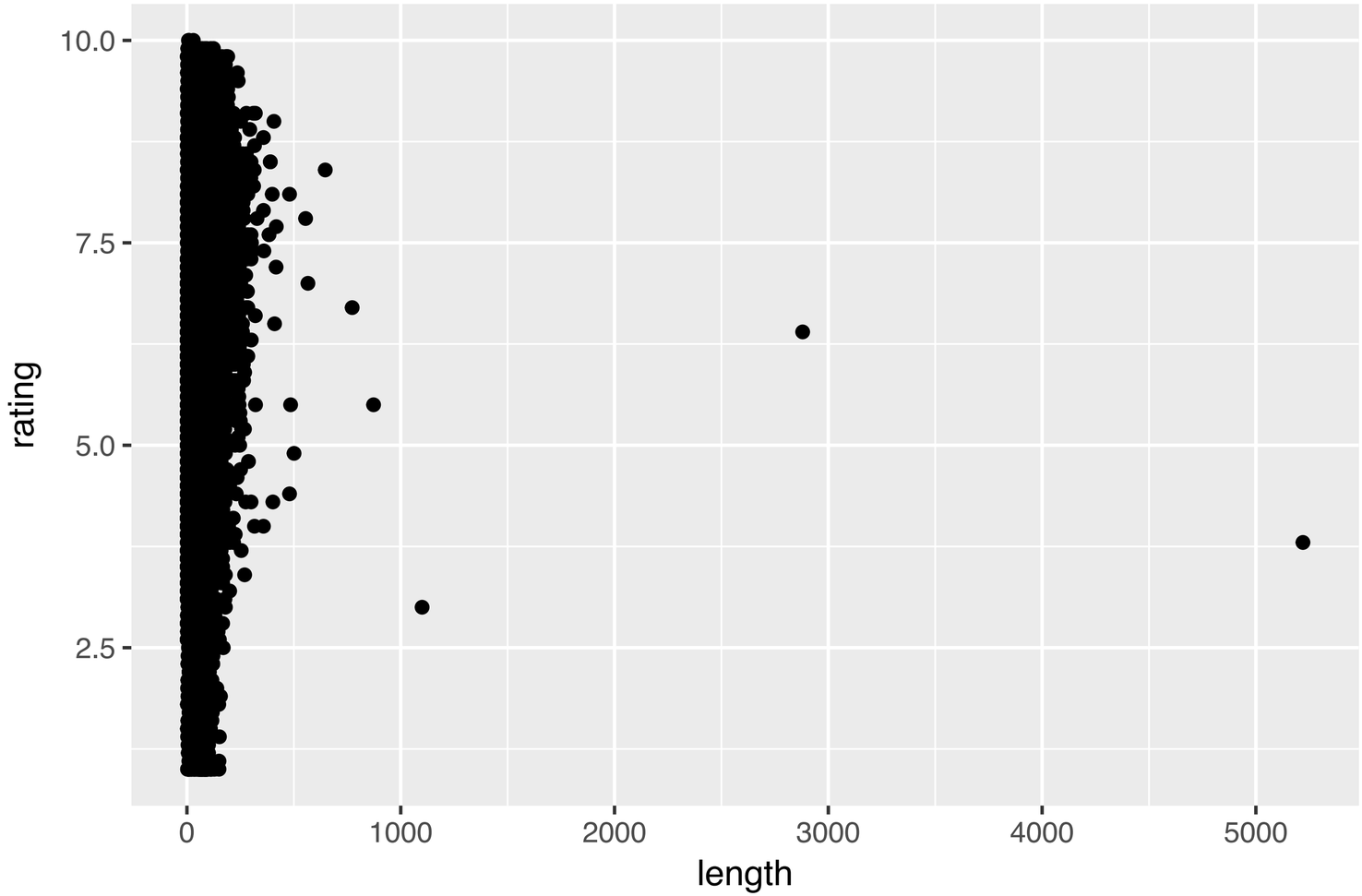
```
library(tidyr)
tidymovies <- gather(movies, key = genre, value, Action:Short) %
    filter(value == 1) %>% select(-value)
ggplot(tidymovies, aes(votes, rating)) + geom_point(alpha = .1)
    facet_wrap(~genre) + scale_x_log10()
```



```
ratings <- c("R", "PG-13", "PG", "NC-17")
tidyrated <- tidymovies %>% filter(mpaa %in% ratings)
ggplot(tidyrated, aes(votes, rating, color = genre,
                      fill = genre)) +
    geom_point(alpha = .5) +
    facet_wrap(~mpaa) + scale_x_log10()
```



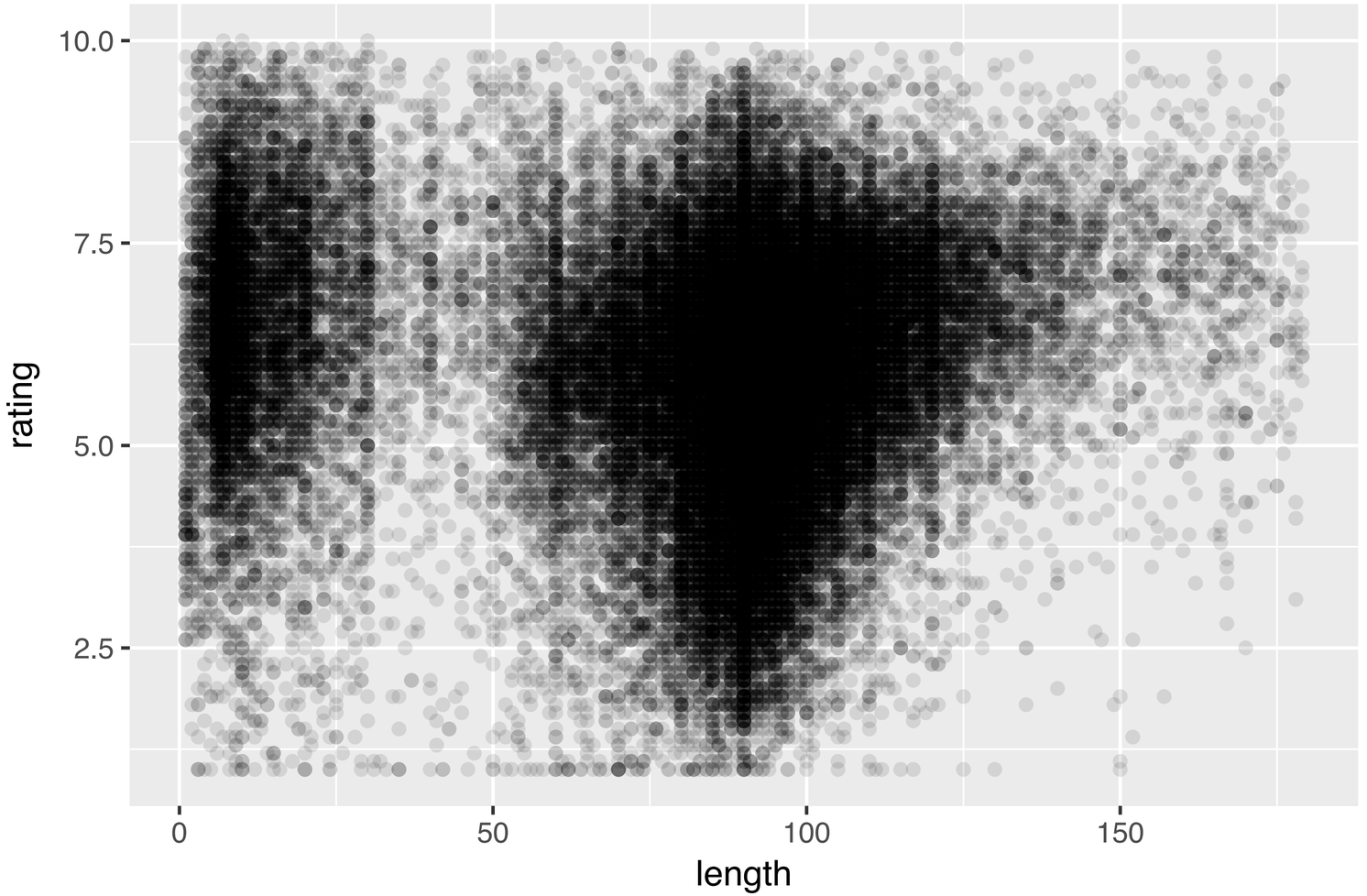
```
ggplot(movies, aes(length, rating)) + geom_point()
```



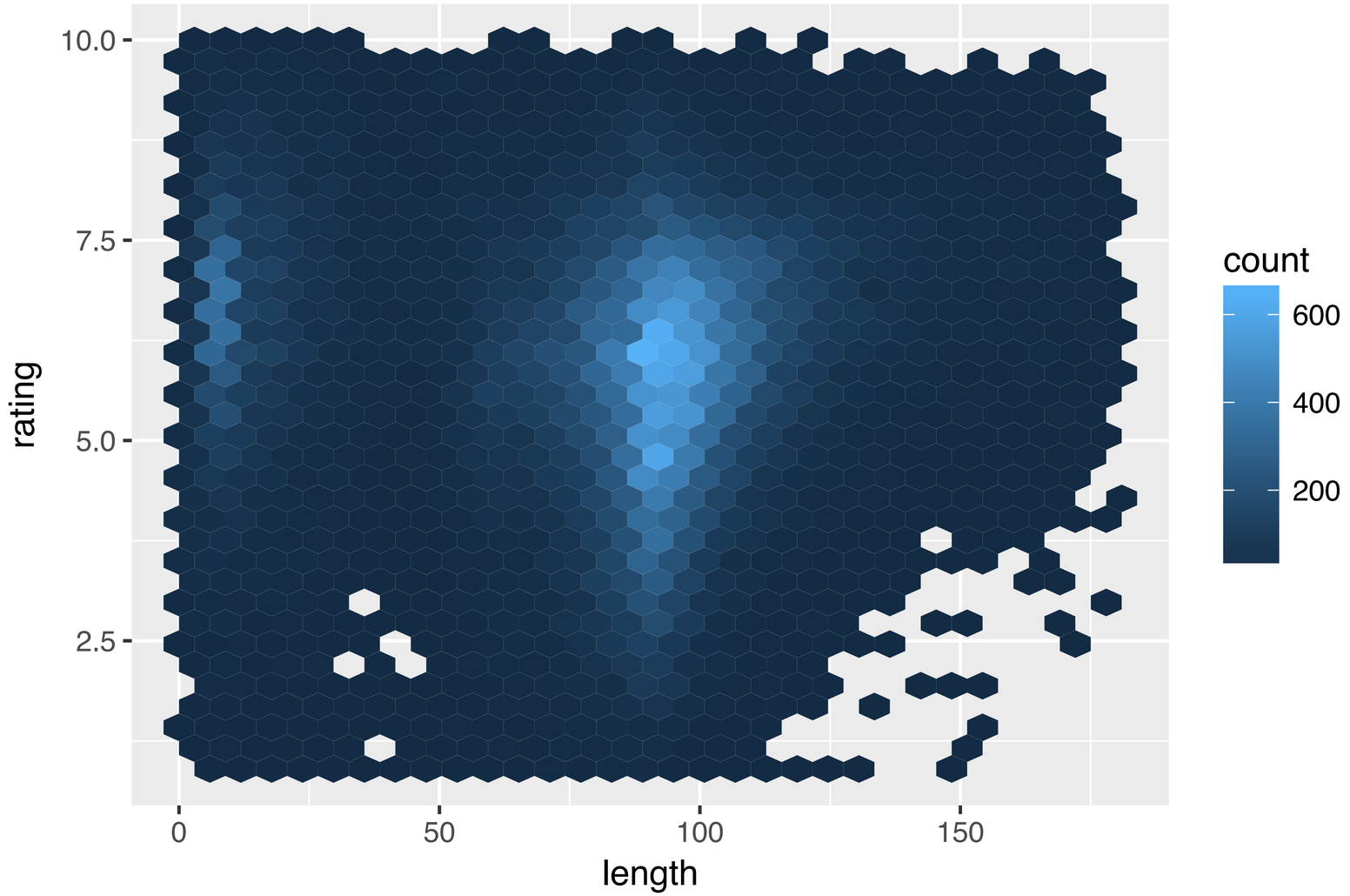
```
movies %>% filter(length > 5000)
```

```
## # A tibble: 1 × 24
##                     title  year length budget rating votes
##                     <chr> <int>  <int>  <int>  <dbl> <int> <db
## 1 Cure for Insomnia, The  1987   5220     NA    3.8    59  44
## # ... with 16 more variables: r3 <dbl>, r4 <dbl>, r5 <dbl>, r
## #   r7 <dbl>, r8 <dbl>, r9 <dbl>, r10 <dbl>, mpaa <chr>, Acti
## #   Animation <int>, Comedy <int>, Drama <int>, Documentary <
## #   Romance <int>, Short <int>
```

```
normalmovies <- movies %>% filter (length < 180)
ggplot(normalmovies, aes(length, rating)) + geom_point(alpha = .
```



```
ggplot(normalmovies, aes(length, rating)) + geom_hex()
```



```
ggplot(normalmovies, aes(length, rating)) + geom_point(alpha = .
```