Categorical Variables (Chapter 4)

Prof. Joyce Robbins

Categorical Data

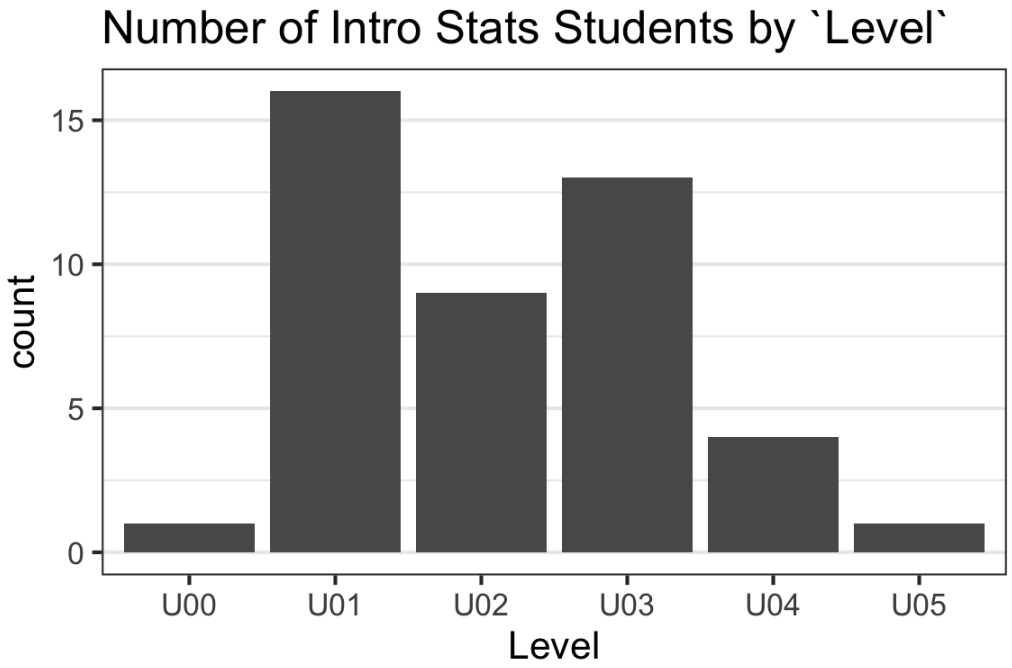
- hard to work with
- not a lot of options (esp. for 1 dimension)
- choice about which categories to display
- choice of the order of categories
- data cleaning takes more time

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

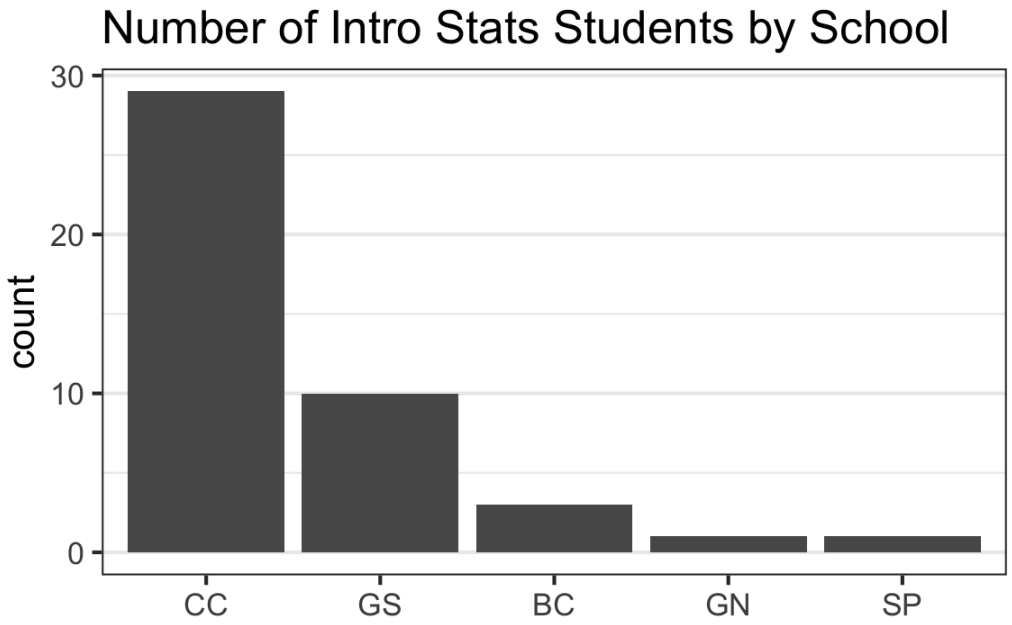
Ordinal data

Sort in logical order of the categories



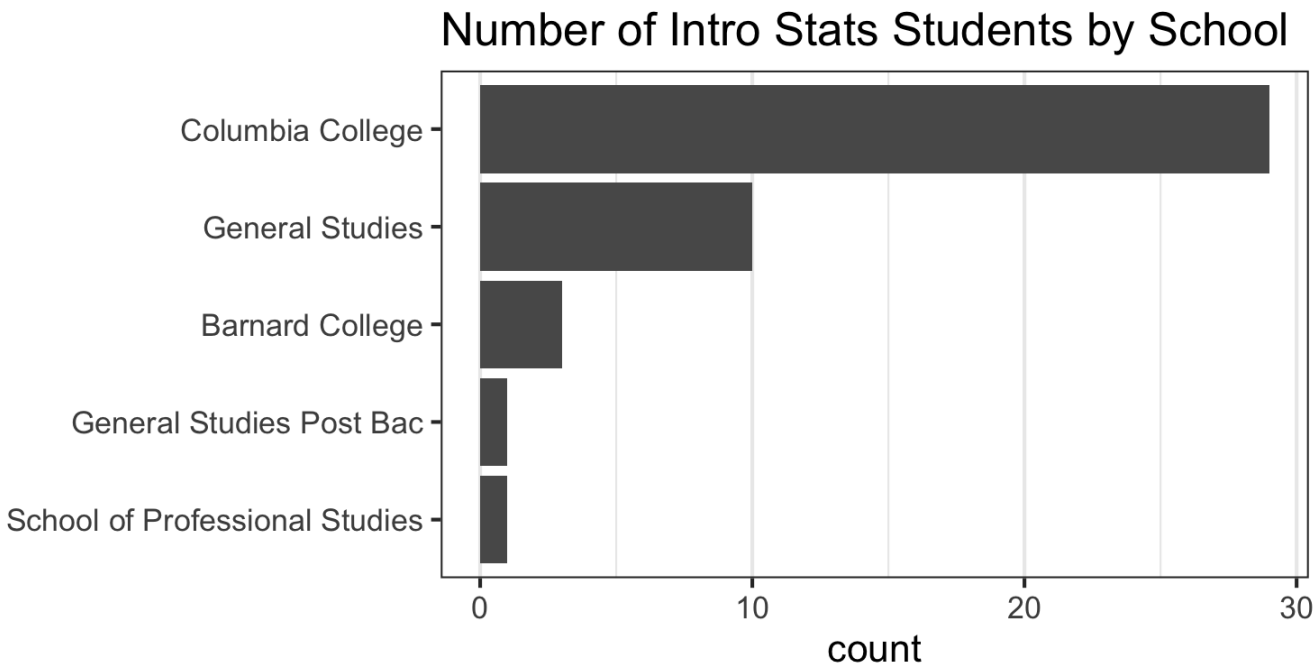
Nominal data

Sort from highest to lowest count (left to right)



Nominal data

… or top to bottom



Unbinned data

`forcats` package (part of tidyverse)

`geom_bar()` takes x only

- Vertical bars:
  ```
  ggplot(df, aes(fct_infreq(x))) + geom_bar()
  ```
- Horizontal bars:
  ```
  ggplot(df, aes(fct_rev(fct_infreq(x)))) + geom_bar() + coord_flip()
  ```

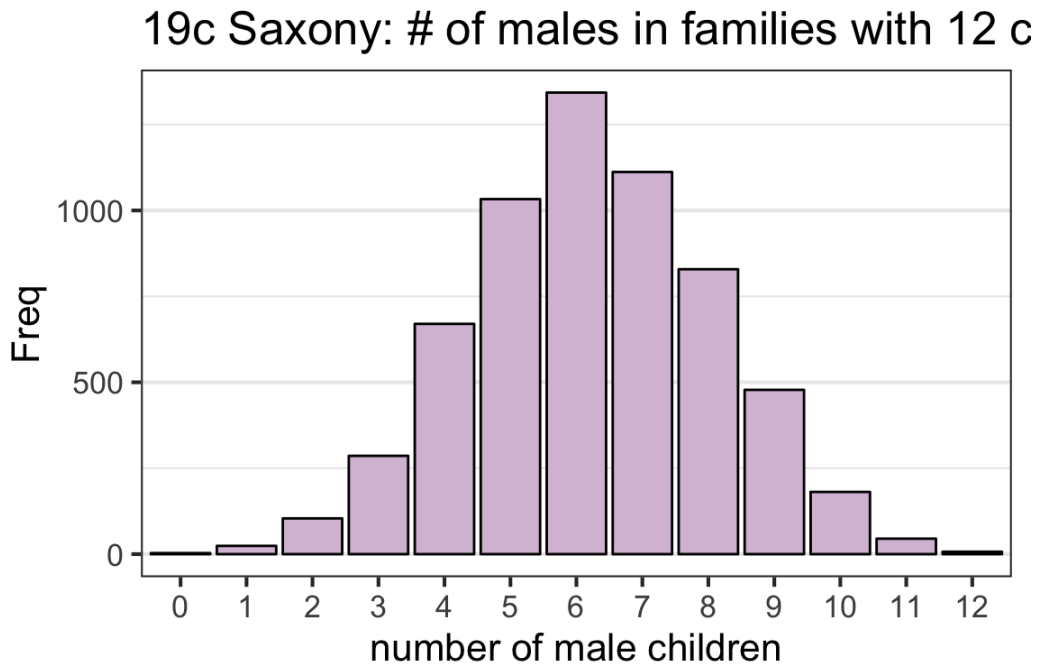
Binned data

`geom_col()` – takes x & y

`...aes(fct_reorder(x, y), y)`

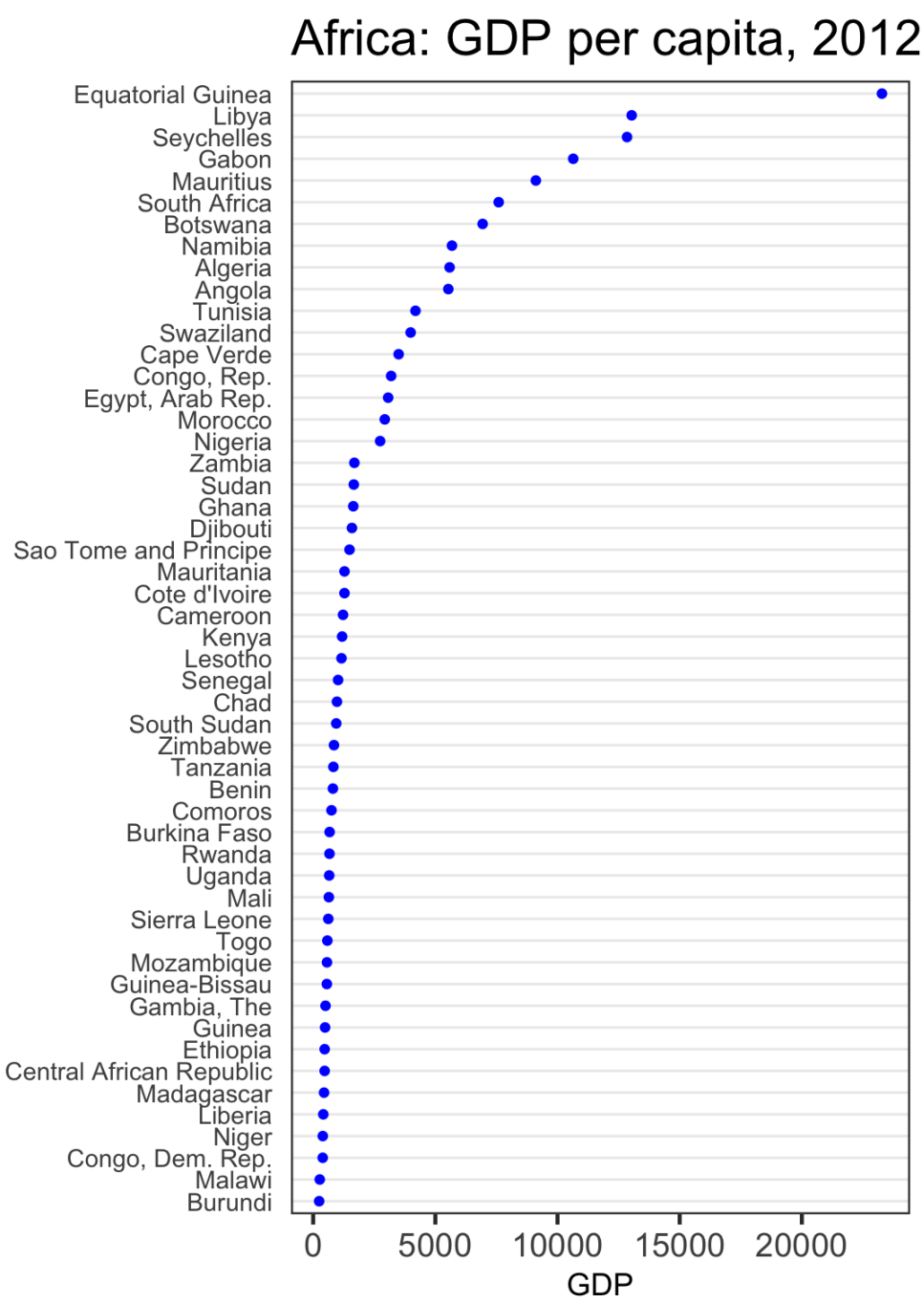
https://github.com/jtr13/codehelp/blob/master/R/reorder.md

Discrete data

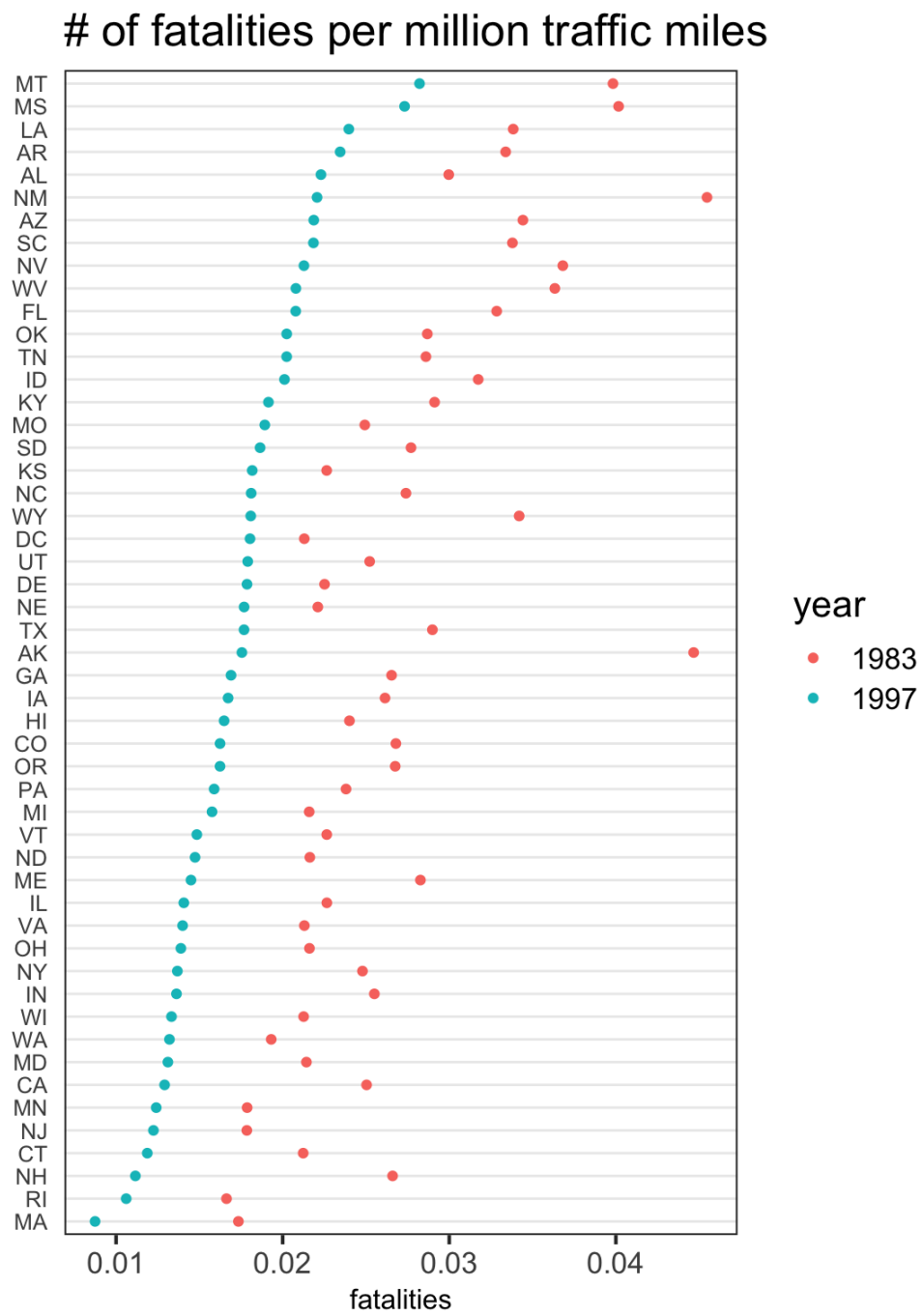


Large number of categories

Cleveland dot plot


Africa: GDP per capita, 2012
Equatorial Guinea
Libya
Seychelles
Gabon
Mauritius
South Africa
Botswana
Namibia
Algeria
Angola
Tunisia
Swaziland
Cape Verde
Congo, Rep.
Egypt, Arab Rep.
Morocco
Nigeria
Zambia
Sudan
Ghana
Djibouti
Sao Tome and Principe
Mauritania
Cote d'Ivoire
Cameroon
Kenya
Lesotho
Senegal
Chad
South Sudan
Zimbabwe
Tanzania
Benin
Comoros
Burkina Faso
Rwanda
Uganda
Mali
Sierra Leone
Togo
Mozambique
Guinea-Bissau
Gambia, The
Guinea
Ethiopia
Central African Republic
Madagascar
Liberia
Niger
Congo, Dem. Rep.
Malawi
Burundi
0
5000
10000
15000
20000
GDP

Cleveland dot plot with multiple dots

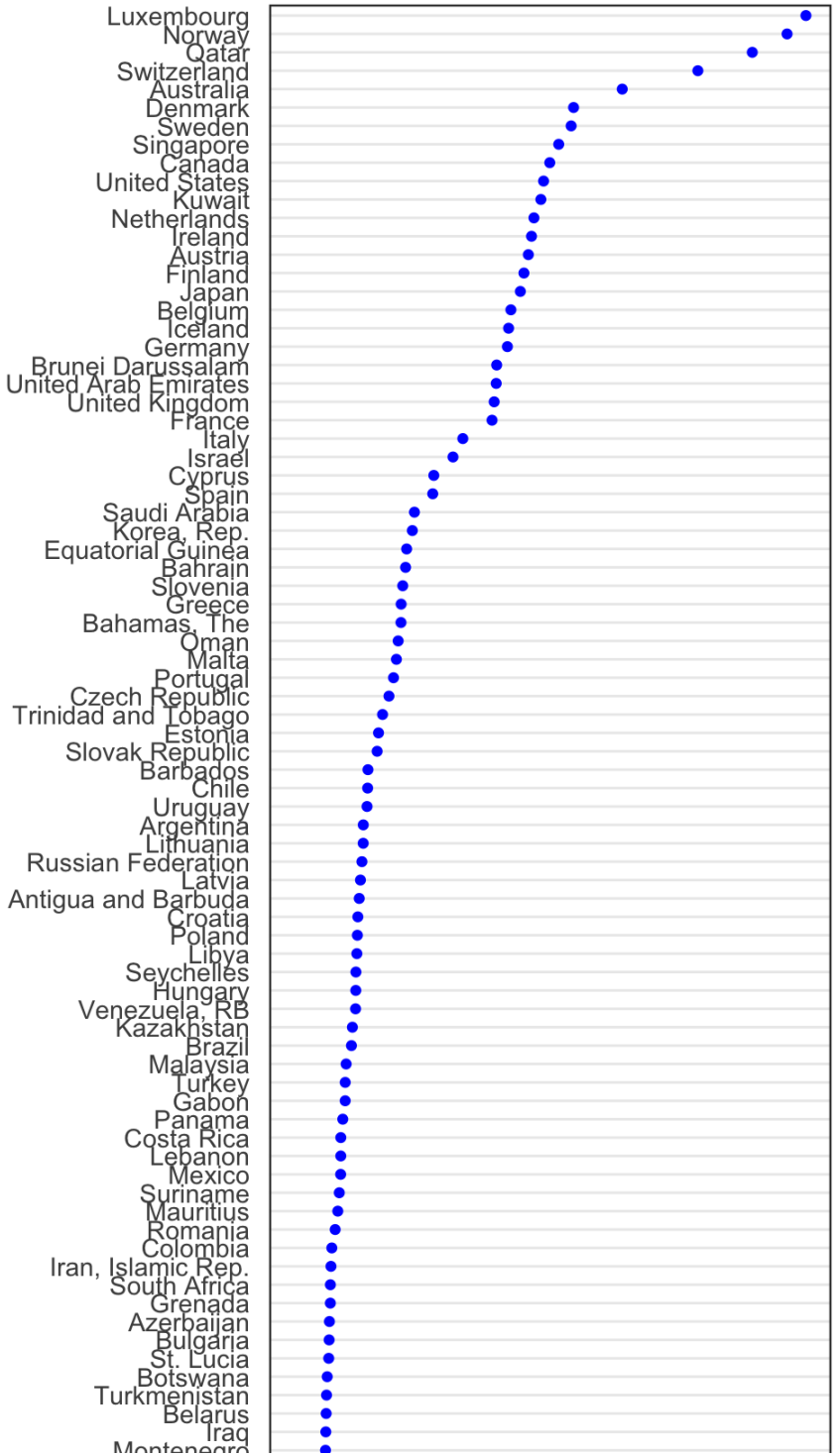


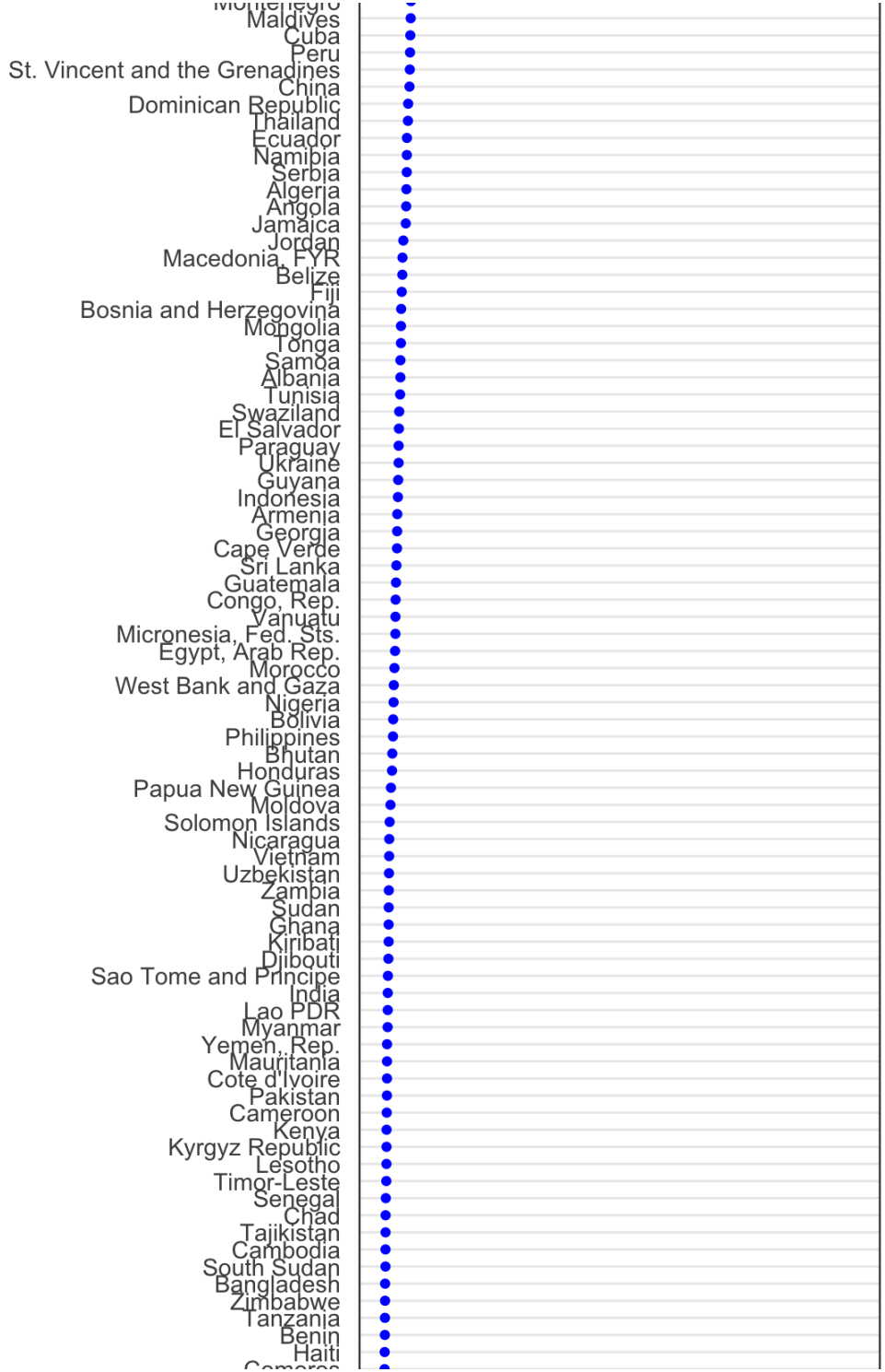
`fct_reorder2` -> double sort: year, then fatalities

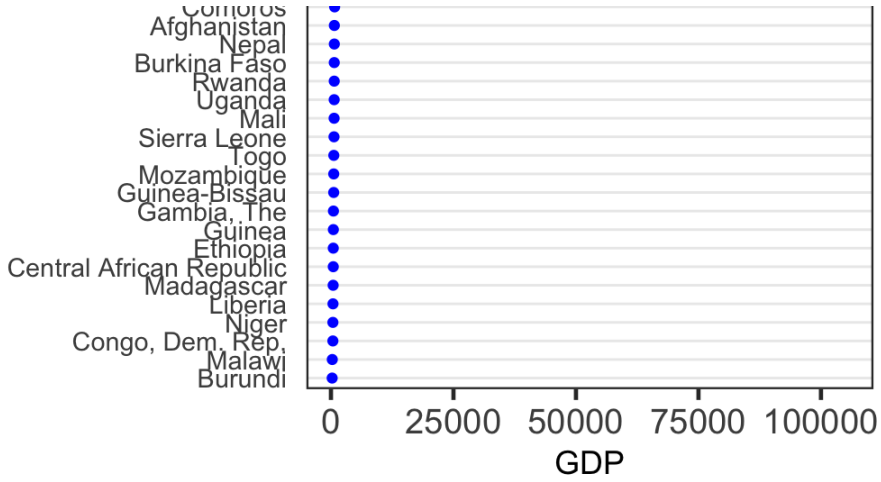
Large number of categories

Scroll

In chunk options: `{r fig.height = 20}`

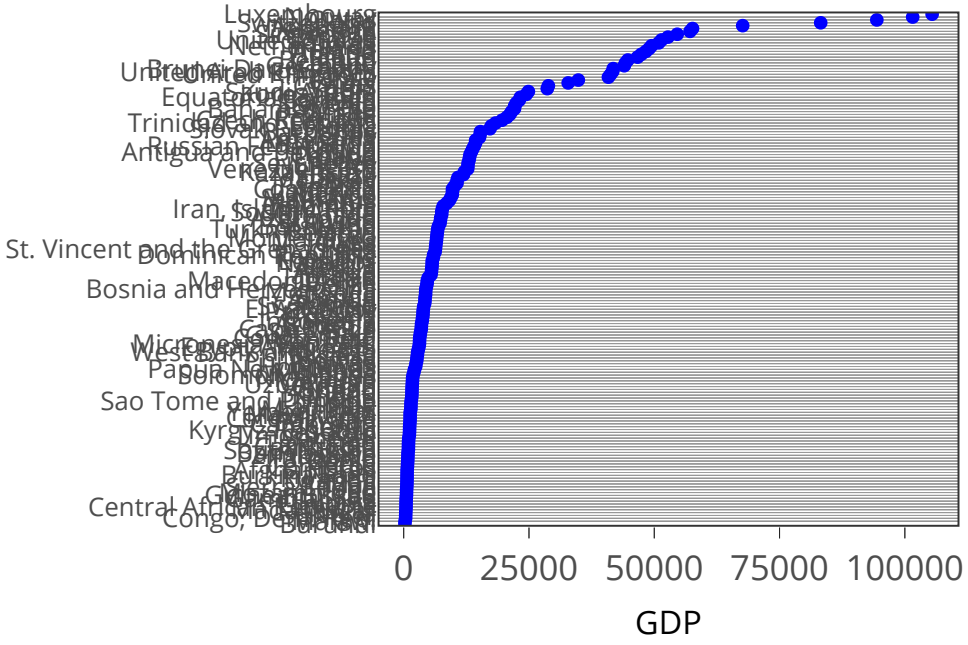

Luxembourg
Norway
Qatar
Switzerland
Australia
Denmark
Sweden
Singapore
Canada
United States
Kuwait
Netherlands
Ireland
Austria
Finland
Japan
Belgium
Iceland
Germany
Brunei Darussalam
United Arab Emirates
United Kingdom
France
Italy
Israel
Cyprus
Spain
Saudi Arabia
Korea, Rep.
Equatorial Guinea
Bahrain
Slovenia
Greece
Bahamas, The
Oman
Malta
Portugal
Czech Republic
Trinidad and Tobago
Estonia
Slovak Republic
Barbados
Chile
Uruguay
Argentina
Lithuania
Russian Federation
Latvia
Antigua and Barbuda
Croatia
Poland
Libya
Seychelles
Hungary
Venezuela, RB
Kazakhstan
Brazil
Malaysia
Turkey
Gabon
Panama
Costa Rica
Lebanon
Mexico
Suriname
Mauritius
Romania
Colombia
Iran, Islamic Rep.
South Africa
Grenada
Azerbaijan
Bulgaria
St. Lucia
Botswana
Turkmenistan
Belarus
Iraq
Montenegro


Montenegro
Maldives
Cuba
Peru
St. Vincent and the Grenadines
China
Dominican Republic
Thailand
Ecuador
Namibia
Serbia
Algeria
Angola
Jamaica
Jordan
Macedonia, FYR
Belize
Fiji
Bosnia and Herzegovina
Mongolia
Tonga
Samoa
Albania
Tunisia
Swaziland
El Salvador
Paraguay
Ukraine
Guyana
Indonesia
Armenia
Georgia
Cape Verde
Sri Lanka
Guatemala
Congo, Rep.
Vanuatu
Micronesia, Fed. Sts.
Egypt, Arab Rep.
Morocco
West Bank and Gaza
Nigeria
Bolivia
Philippines
Bhutan
Honduras
Papua New Guinea
Moldova
Solomon Islands
Nicaragua
Vietnam
Uzbekistan
Zambia
Sudan
Ghana
Kiribati
Djibouti
Sao Tome and Principe
India
Lao PDR
Myanmar
Yemen, Rep.
Mauritania
Cote d'Ivoire
Pakistan
Cameroon
Kenya
Kyrgyz Republic
Lesotho
Timor-Leste
Senegal
Chad
Tajikistan
Cambodia
South Sudan
Bangladesh
Zimbabwe
Tanzania
Benin
Haiti
Comoros


Comoros
Afghanistan
Nepal
Burkina Faso
Rwanda
Uganda
Mali
Sierra Leone
Togo
Mozambique
Guinea-Bissau
Gambia, The
Guinea
Ethiopia
Central African Republic
Madagascar
Liberia
Niger
Congo, Dem. Rep.
Malawi
Burundi
0
25000
50000
75000
100000
GDP

Interactive Cleveland Dot Plot

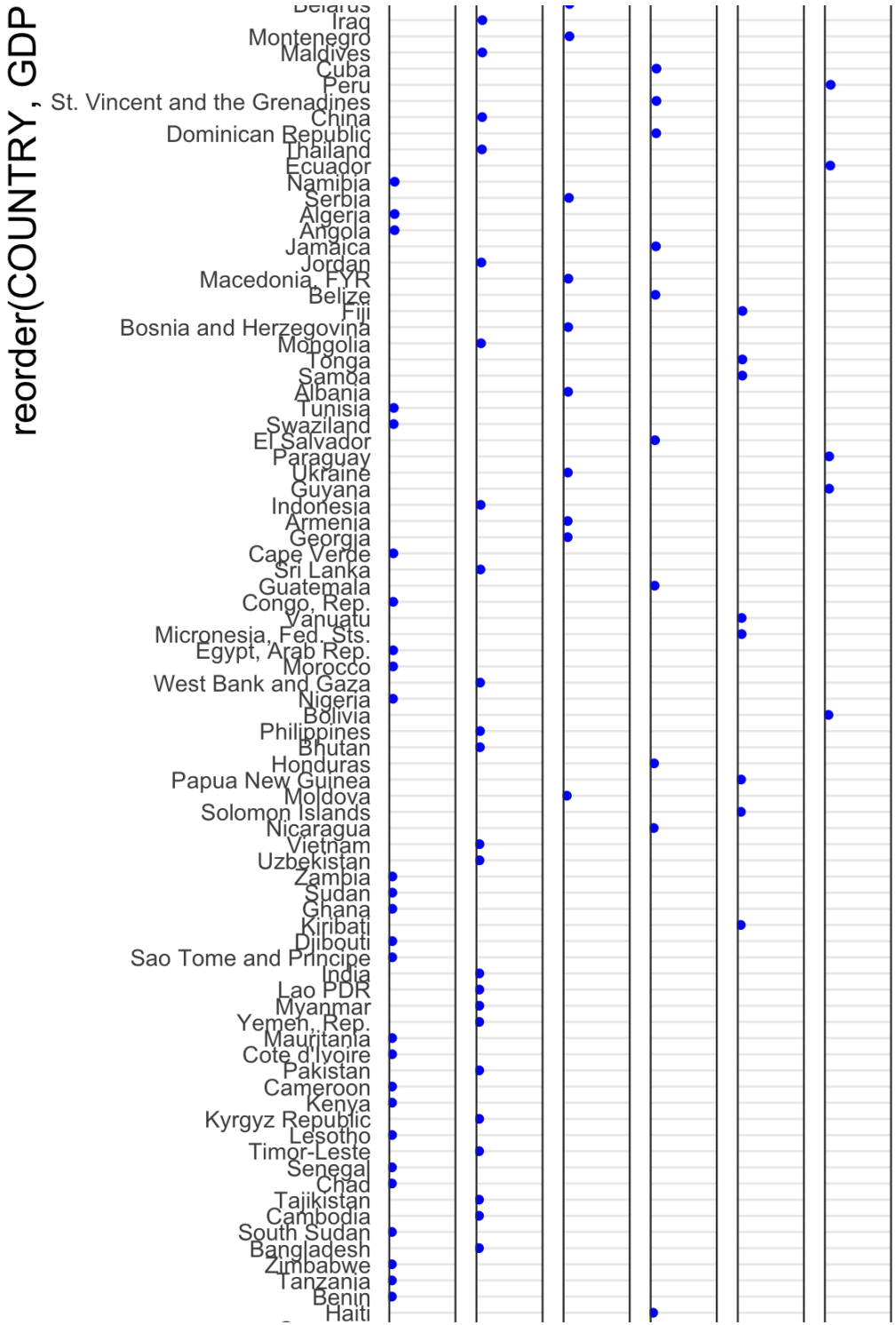
```
library(plotly)
ggplotly(g)
```

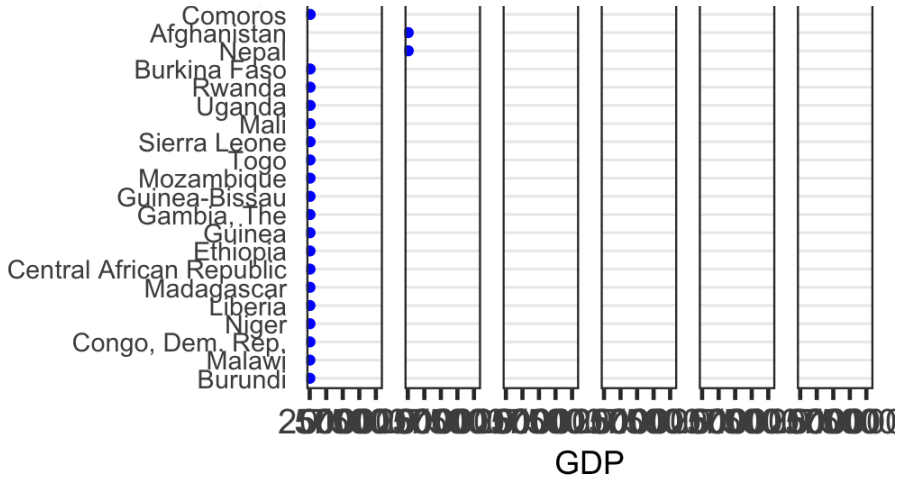


Cleveland dot plot with facets

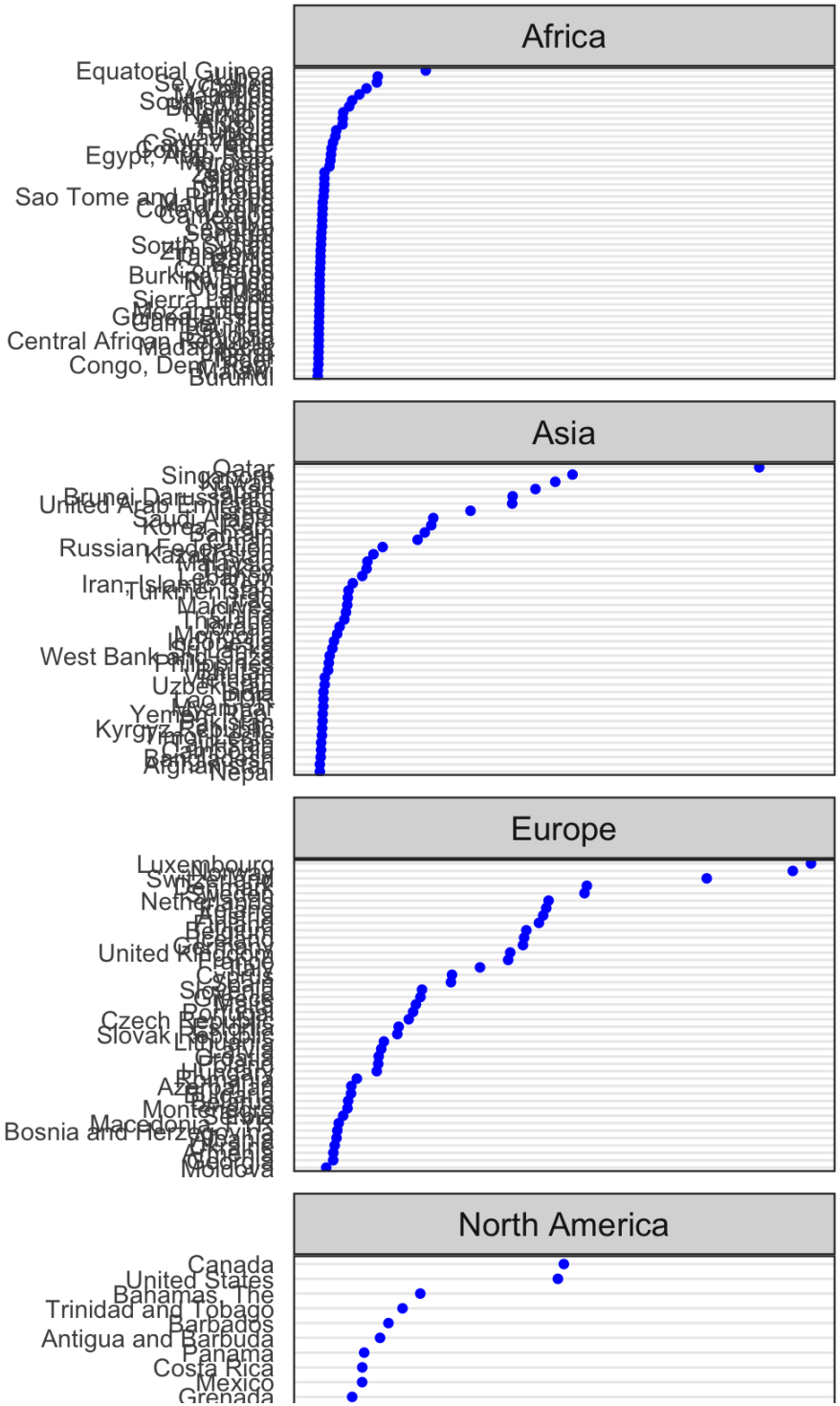
What’s wrong?

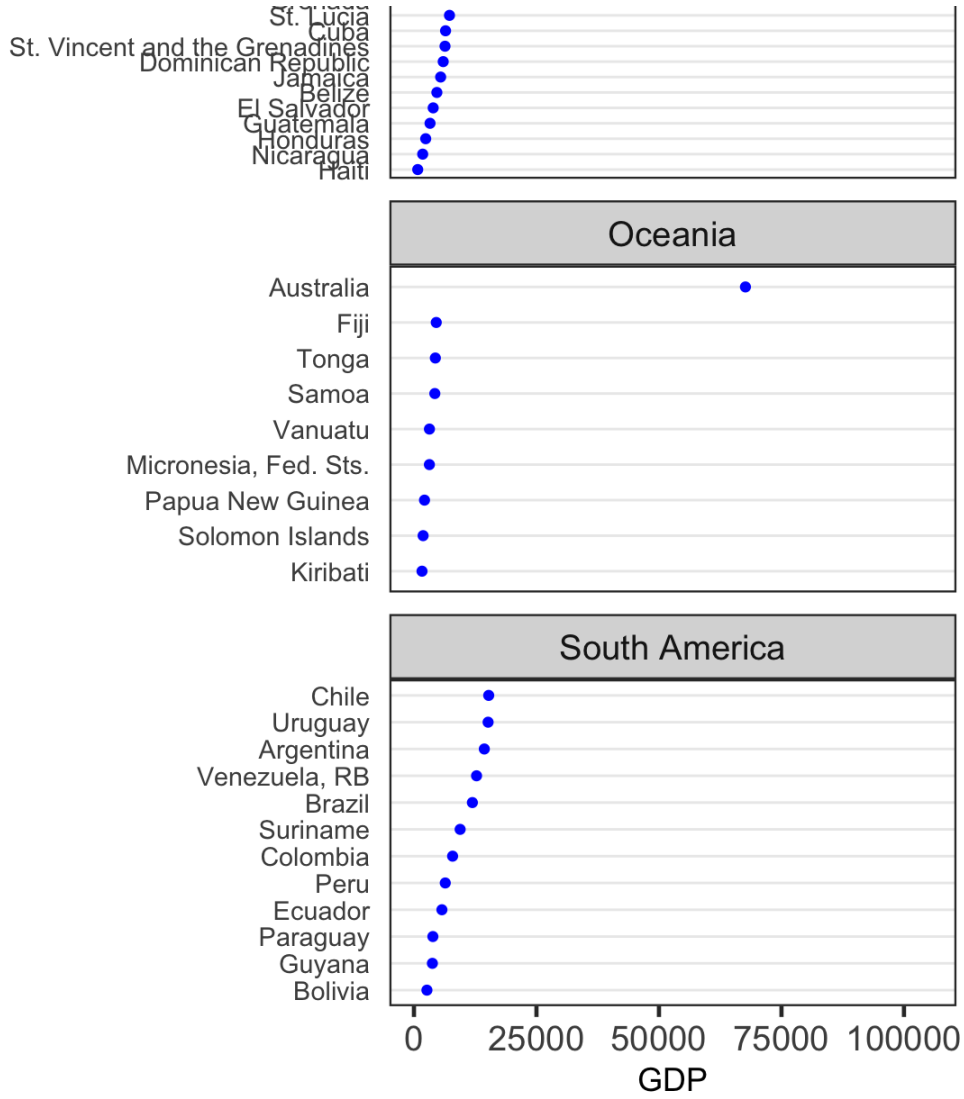

Africa
Asia
Europ
n Am
ocean
n Am
Luxembourg
Norway
Qatar
Switzerland
Australia
Denmark
Sweden
Singapore
Canada
United States
Kuwait
Netherlands
Ireland
Austria
Finland
Japan
Belgium
Iceland
Germany
Brunei Darussalam
United Arab Emirates
United Kingdom
France
Italy
Israel
Cyprus
Spain
Saudi Arabia
Korea, Rep.
Equatorial Guinea
Bahrain
Slovenia
Greece
Bahamas, The
Oman
Malta
Portugal
Czech Republic
Trinidad and Tobago
Estonia
Slovak Republic
Barbados
Chile
Uruguay
Argentina
Lithuania
Russian Federation
Latvia
Antigua and Barbuda
Croatia
Poland
Libya
Seychelles
Hungary
Venezuela, RB
Kazakhstan
Brazil
Malaysia
Turkey
Gabon
Panama
Costa Rica
Lebanon
Mexico
Suriname
Mauritius
Romania
Colombia
Iran, Islamic Rep.
South Africa
Grenada
Azerbaijan
Bulgaria
St. Lucia
Botswana
Turkmenistan
Belarus


reorder(COUNTRY, GDP
Belarus
Iraq
Montenegro
Maldives
Cuba
Peru
St. Vincent and the Grenadines
China
Dominican Republic
Thailand
Ecuador
Namibia
Serbia
Algeria
Angola
Jamaica
Jordan
Macedonia, FYR
Belize
Fiji
Bosnia and Herzegovina
Mongolia
Tonga
Samoa
Albania
Tunisia
Swaziland
El Salvador
Paraguay
Ukraine
Guyana
Indonesia
Armenia
Georgia
Cape Verde
Sri Lanka
Guatemala
Congo, Rep.
Vanuatu
Micronesia, Fed. Sts.
Egypt, Arab Rep.
Morocco
West Bank and Gaza
Nigeria
Bolivia
Philippines
Bhutan
Honduras
Papua New Guinea
Moldova
Solomon Islands
Nicaragua
Vietnam
Uzbekistan
Zambia
Sudan
Ghana
Kiribati
Djibouti
Sao Tome and Principe
India
Lao PDR
Myanmar
Yemen, Rep.
Mauritania
Cote d'Ivoire
Pakistan
Cameroon
Kenya
Kyrgyz Republic
Lesotho
Timor-Leste
Senegal
Chad
Tajikistan
Cambodia
South Sudan
Bangladesh
Zimbabwe
Tanzania
Benin
Haiti

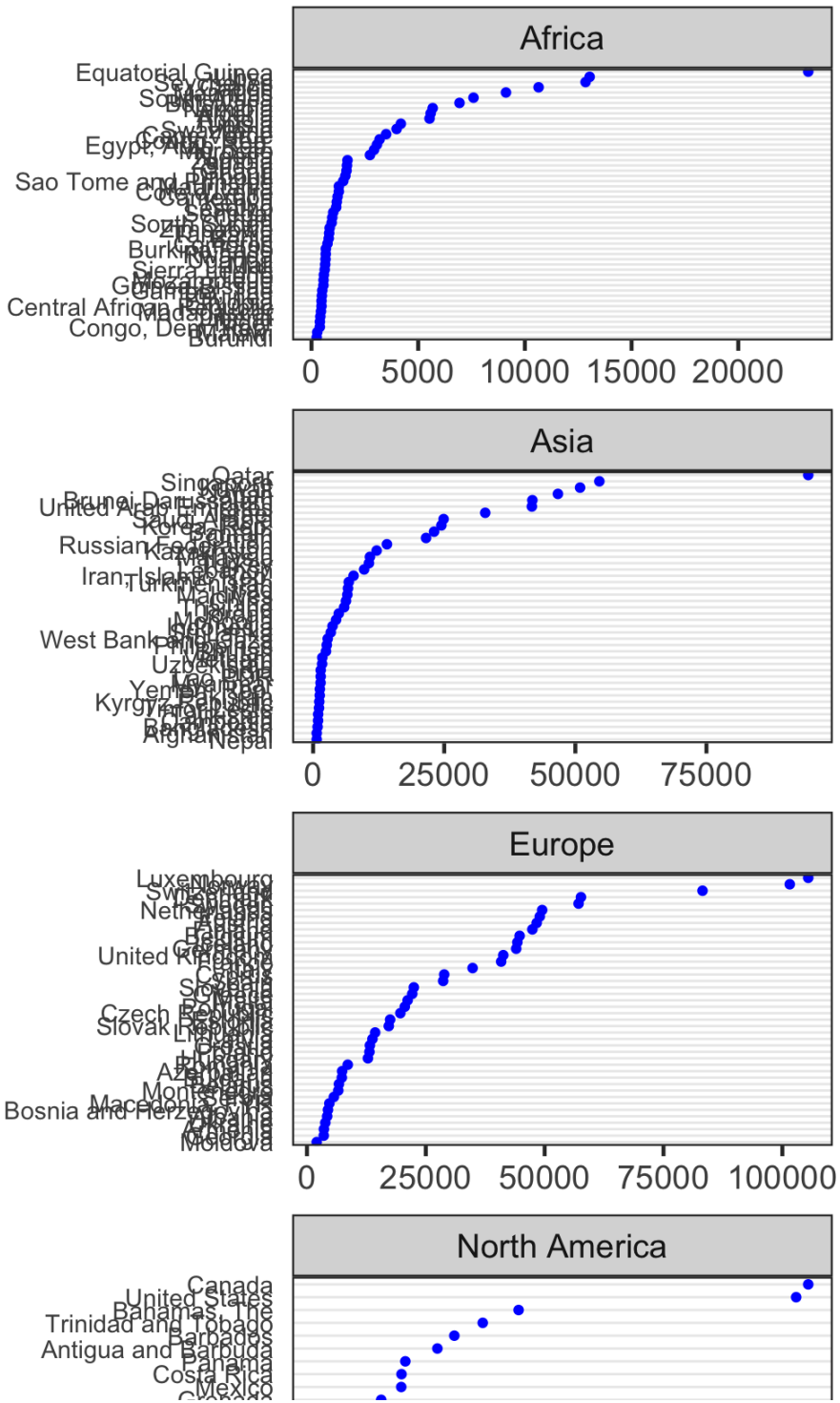

Comoros
Afghanistan
Nepal
Burkina Faso
Rwanda
Uganda
Mali
Sierra Leone
Togo
Mozambique
Guinea-Bissau
Gambia, The
Guinea
Ethiopia
Central African Republic
Madagascar
Liberia
Niger
Congo, Dem. Rep.
Malawi
Burundi
GDP

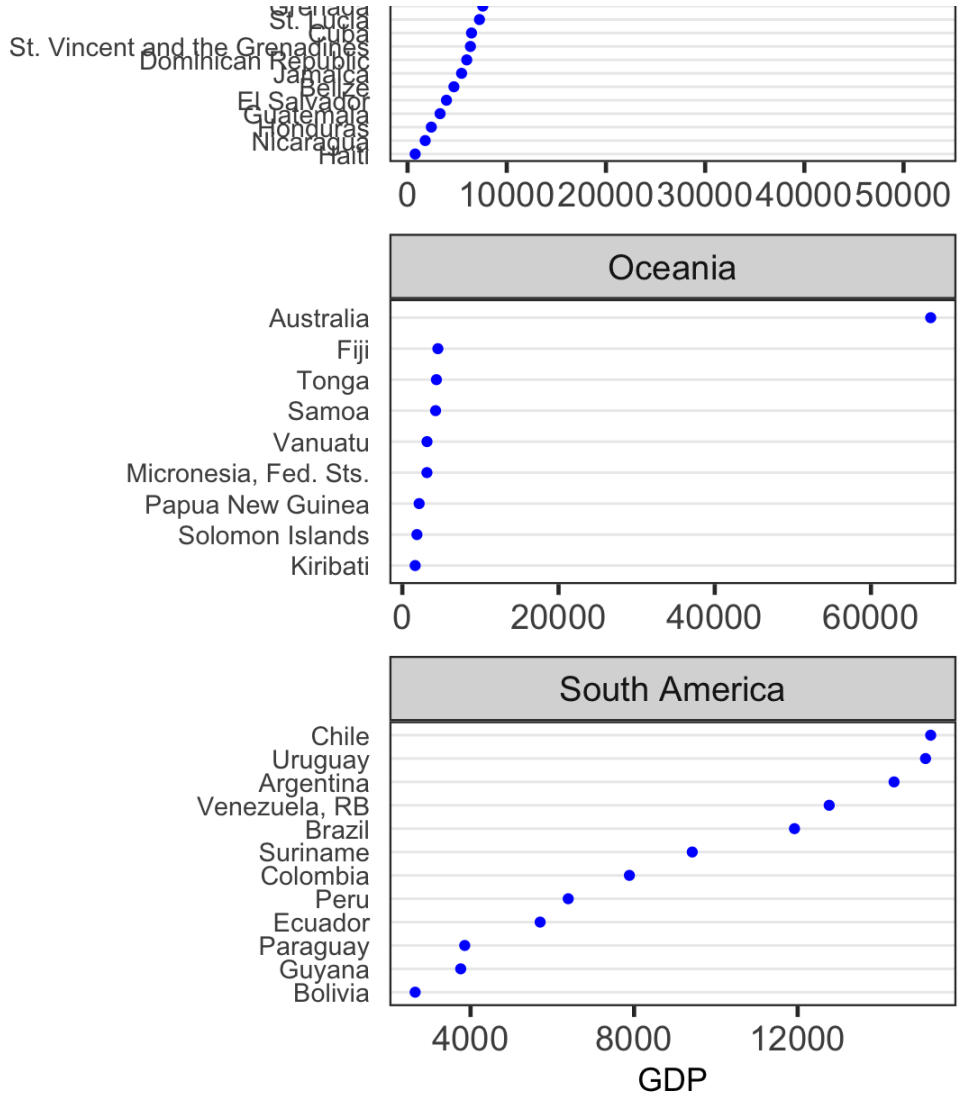
Cleveland Dot Plot with Facets `scales = "free_y"`


Africa
Equatorial Guinea
Egypt, Arab Rep.
Sao Tome and Principe
South Africa
Burkina Faso
Sierra Leone
Central African Republic
Congo, Dem. Rep.
Burundi
Asia
Qatar
Singapore
Brunei Darussalam
United Arab Emirates
Saudi Arabia
Korea, Rep.
Bahrain
Russian Federation
Kazakhstan
Malaysia
Iran, Islamic Rep.
Turkmenistan
Maldives
Thailand
Mongolia
Indonesia
West Bank and Gaza
Philippines
Vietnam
Uzbekistan
Lao PDR
Myanmar
Yemen, Rep.
Pakistan
Kyrgyz Republic
Tajikistan
Cambodia
Bangladesh
Afghanistan
Nepal
Europe
Luxembourg
Norway
Switzerland
Ireland
Denmark
Netherlands
Sweden
Belgium
Austria
Germany
United Kingdom
France
Finland
Cyprus
Slovenia
Italy
Spain
Malta
Portugal
Czech Republic
Slovak Republic
Estonia
Lithuania
Latvia
Poland
Croatia
Hungary
Greece
Azerbaijan
Romania
Bulgaria
Montenegro
Serbia
Macedonia, FYR
Bosnia and Herzegovina
Albania
Ukraine
Georgia
Moldova
North America
Canada
United States
Bahamas, The
Trinidad and Tobago
Barbados
Antigua and Barbuda
Panama
Costa Rica
Mexico
Grenada

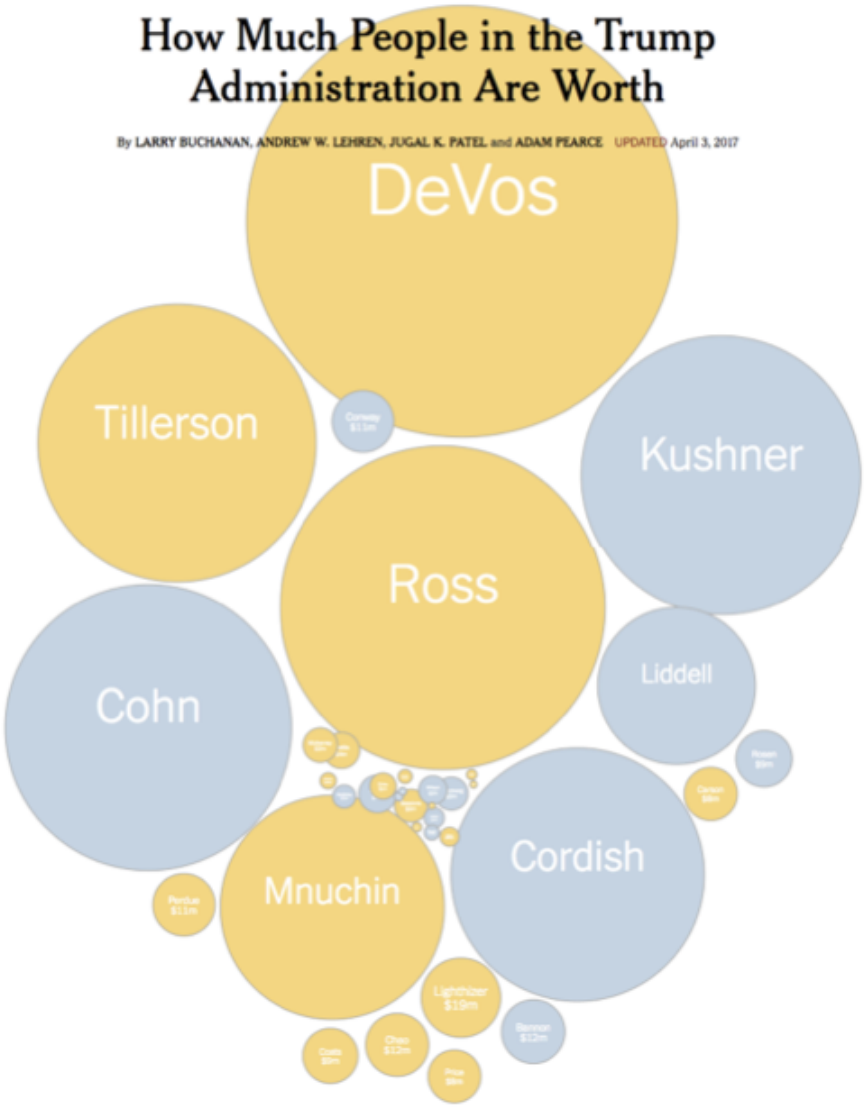

St. Lucia
Cuba
St. Vincent and the Grenadines
Dominican Republic
Jamaica
Belize
El Salvador
Guatemala
Honduras
Nicaragua
Haiti
Oceania
Australia
Fiji
Tonga
Samoa
Vanuatu
Micronesia, Fed. Sts.
Papua New Guinea
Solomon Islands
Kiribati
South America
Chile
Uruguay
Argentina
Venezuela, RB
Brazil
Suriname
Colombia
Peru
Ecuador
Paraguay
Guyana
Bolivia
0
25000
50000
75000
100000
GDP

Cleveland Dot Plot with Facets `scales = "free"`

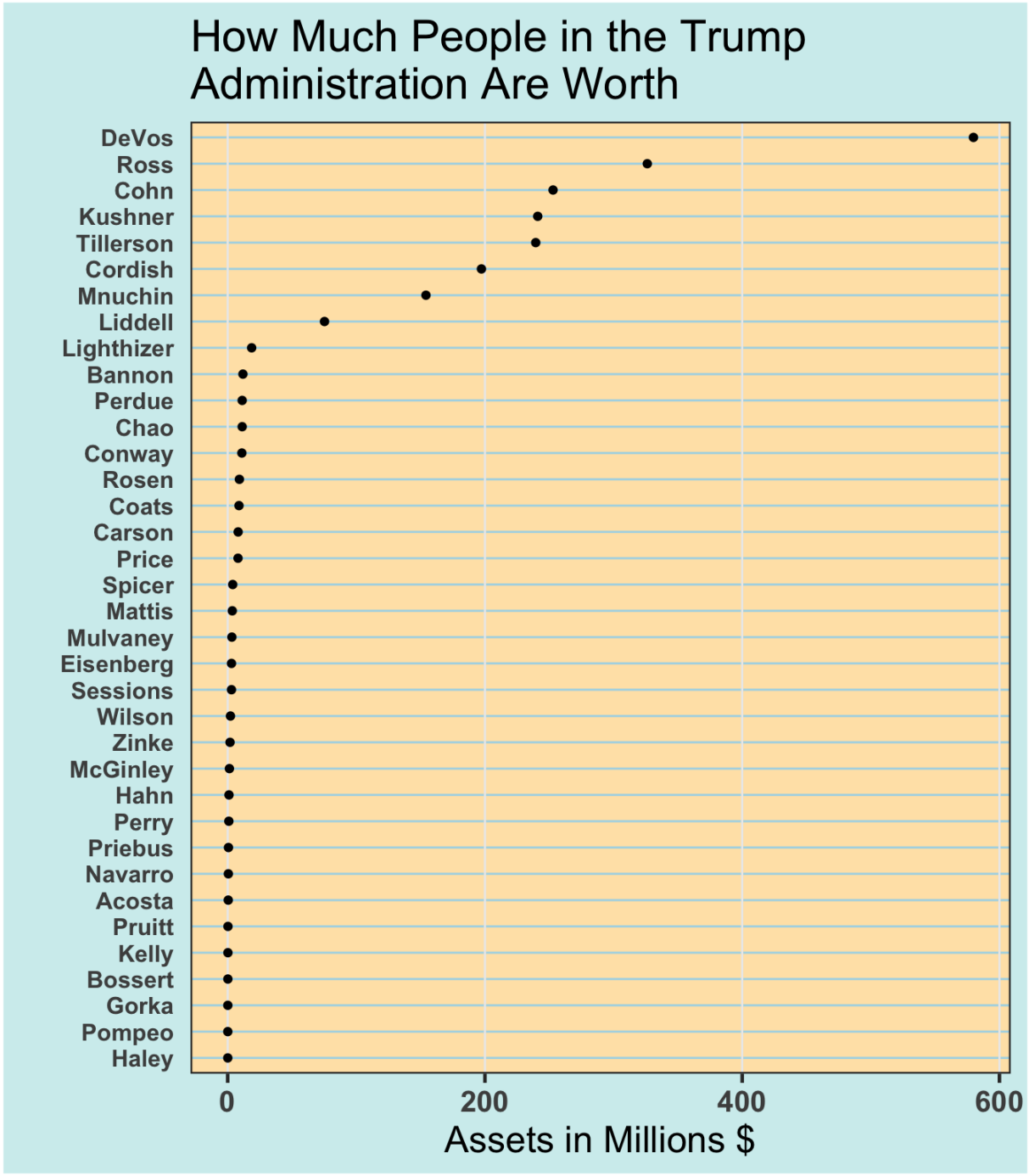

Africa
Equatorial Guinea
Egypt, Arab Rep.
Sao Tome and Principe
Central African Republic
Congo, Dem. Rep.
Burundi
0
5000
10000
15000
20000
Asia
Qatar
Singapore
Brunei Darussalam
United Arab Emirates
Russian Federation
Iran, Islamic Rep.
West Bank and Gaza
Kyrgyz Republic
Nepal
0
25000
50000
75000
Europe
Luxembourg
Switzerland
Netherlands
Ireland
United Kingdom
Czech Republic
Slovak Republic
Bosnia and Herzegovina
Moldova
0
25000
50000
75000
100000
North America
Canada
United States
Bahamas, The
Trinidad and Tobago
Barbados
Antigua and Barbuda
Panama
Costa Rica
Mexico

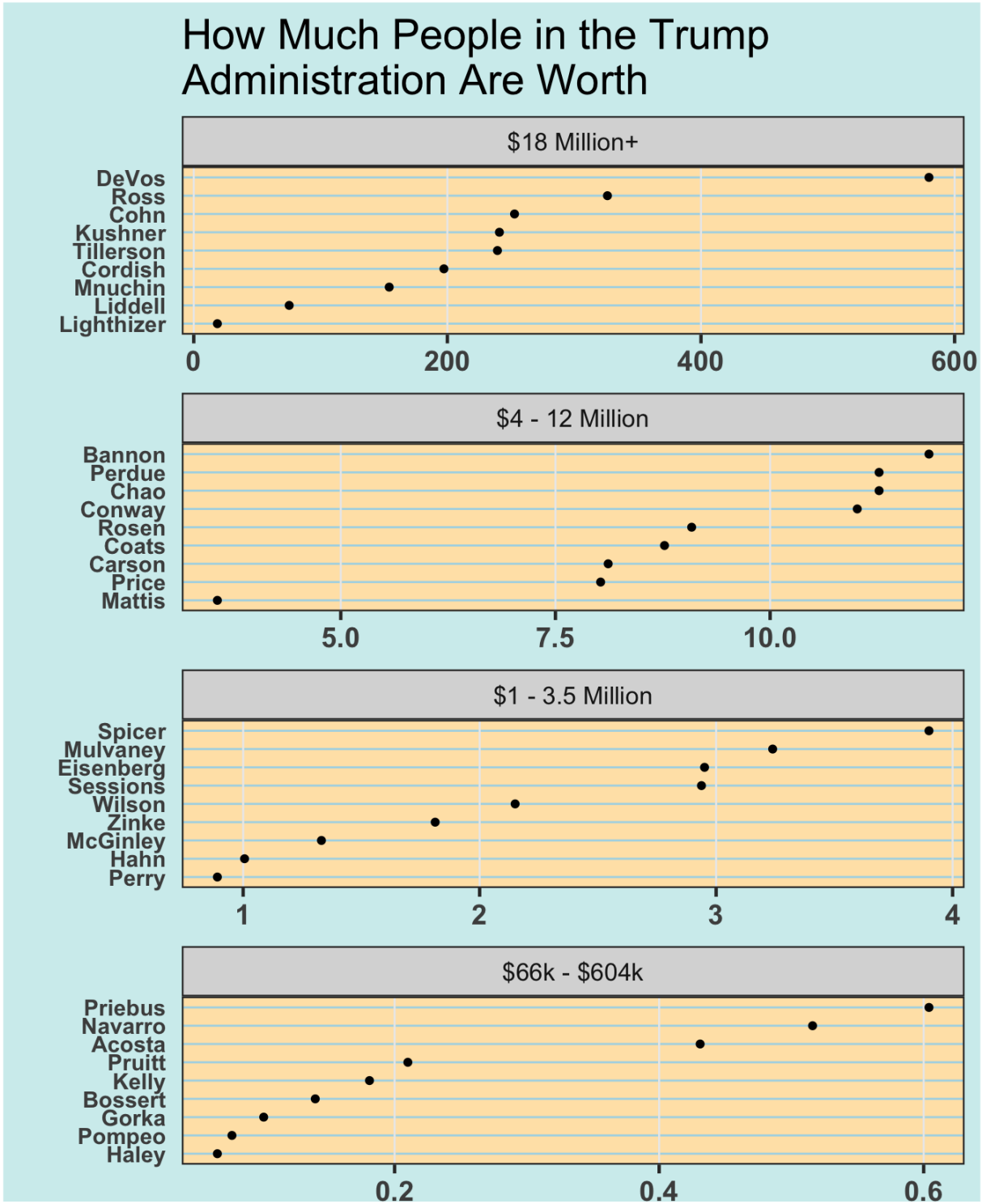

St. Lucia
Cuba
St. Vincent and the Grenadines
Dominican Republic
Jamaica
Belize
El Salvador
Guatemala
Honduras
Nicaragua
Haiti
0
10000
20000
30000
40000
50000
Oceania
Australia
Fiji
Tonga
Samoa
Vanuatu
Micronesia, Fed. Sts.
Papua New Guinea
Solomon Islands
Kiribati
0
20000
40000
60000
South America
Chile
Uruguay
Argentina
Venezuela, RB
Brazil
Suriname
Colombia
Peru
Ecuador
Paraguay
Guyana
Bolivia
4000
8000
12000
GDP

Trump Administration Assets



Redraw as Cleveland Dot Plot


How Much People in the Trump Administration Are Worth
DeVos
Ross
Cohn
Kushner
Tillerson
Cordish
Mnuchin
Liddell
Lighthizer
Bannon
Perdue
Chao
Conway
Rosen
Coats
Carson
Price
Spicer
Mattis
Mulvaney
Eisenberg
Sessions
Wilson
Zinke
McGinley
Hahn
Perry
Priebus
Navarro
Acosta
Pruitt
Kelly
Bossert
Gorka
Pompeo
Haley
0
200
400
600
Assets in Millions $


How Much People in the Trump Administration Are Worth
$18 Million+
DeVos
Ross
Cohn
Kushner
Tillerson
Cordish
Mnuchin
Liddell
Lighthizer
0
200
400
600
$4 - 12 Million
Bannon
Perdue
Chao
Conway
Rosen
Coats
Carson
Price
Mattis
5.0
7.5
10.0
$1 - 3.5 Million
Spicer
Mulvaney
Eisenberg
Sessions
Wilson
Zinke
McGinley
Hahn
Perry
1
2
3
4
$66k - $604k
Priebus
Navarro
Acosta
Pruitt
Kelly
Bossert
Gorka
Pompeo
Haley
0.2
0.4
0.6

Assets in Millions $

Recoding factor levels

Keep a trail of breadcrumbs

```
x <- factor(c("G234", "G452", "G136"))
levels(x)
```

```
## [1] "G136" "G234" "G452"
```

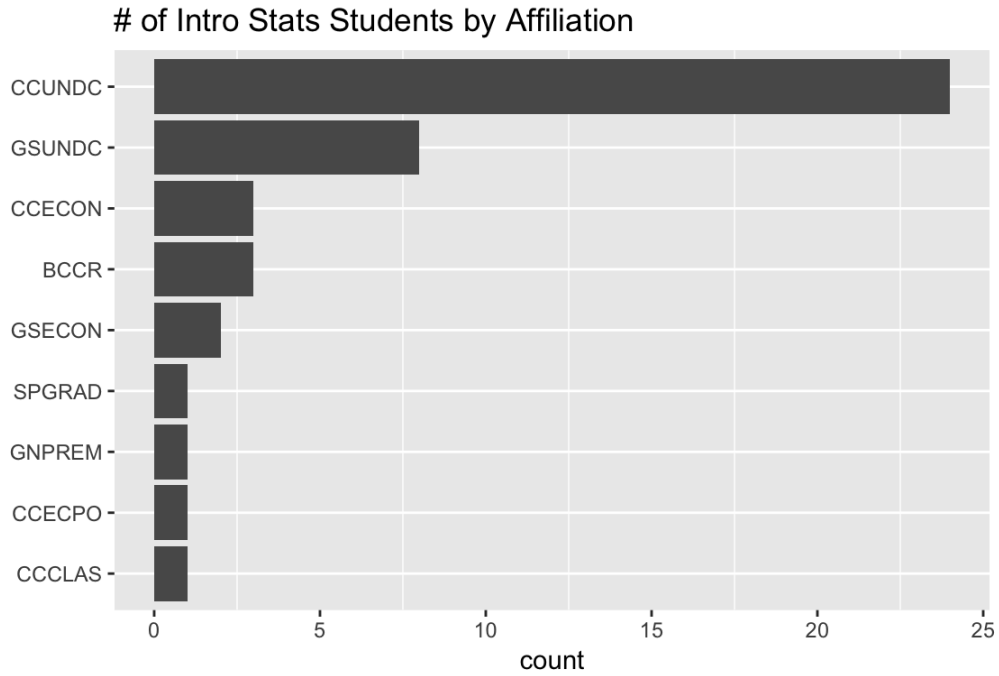
```
levels(x) <- c("Physics", "Math", "Chemistry")
x
```

```
## [1] Math      Chemistry Physics
## Levels: Physics Math Chemistry
```

```
# clear connections between old and new names, old version is preserved
x <- factor(c("G234", "G452", "G136"))
y <- fct_recode(x, Physics = "G234", Math = "G452", Chemistry = "G136")
y
```

```
## [1] Physics   Math      Chemistry
## Levels: Chemistry Physics Math
```

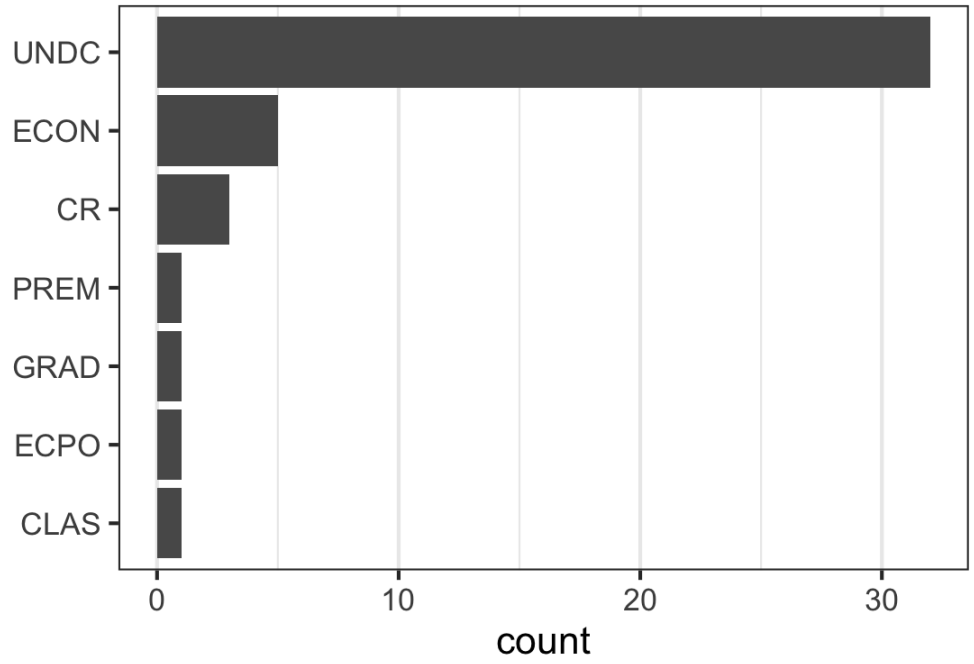
Cleaning data



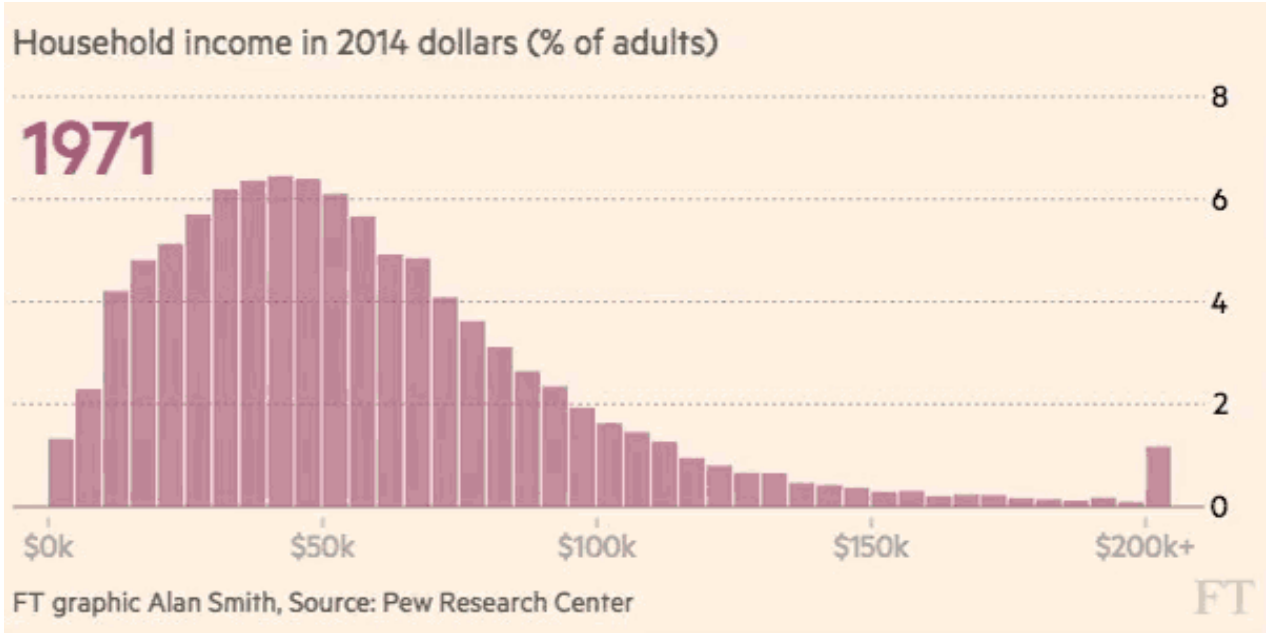
What's the problem?

Major only

Remove school from affiliation



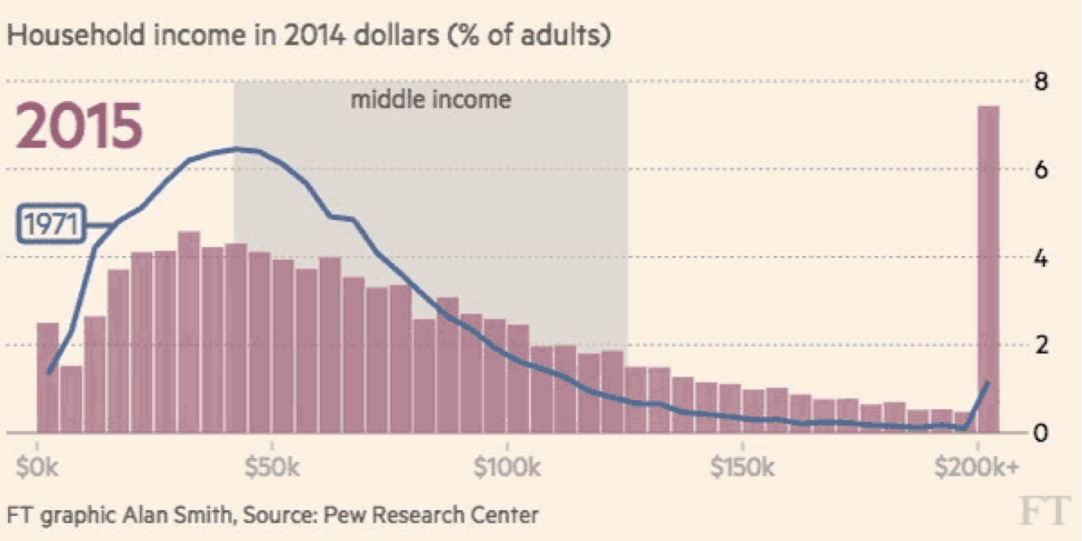
“Other” “or more” categories: “topcoding”



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

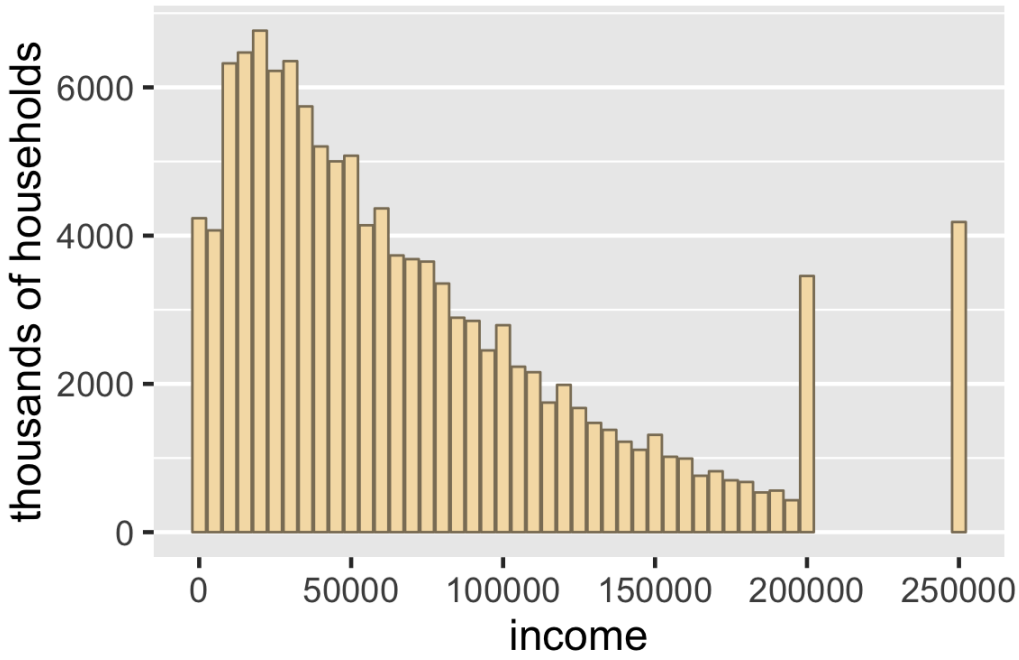
“Other” “or more” categories

$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	

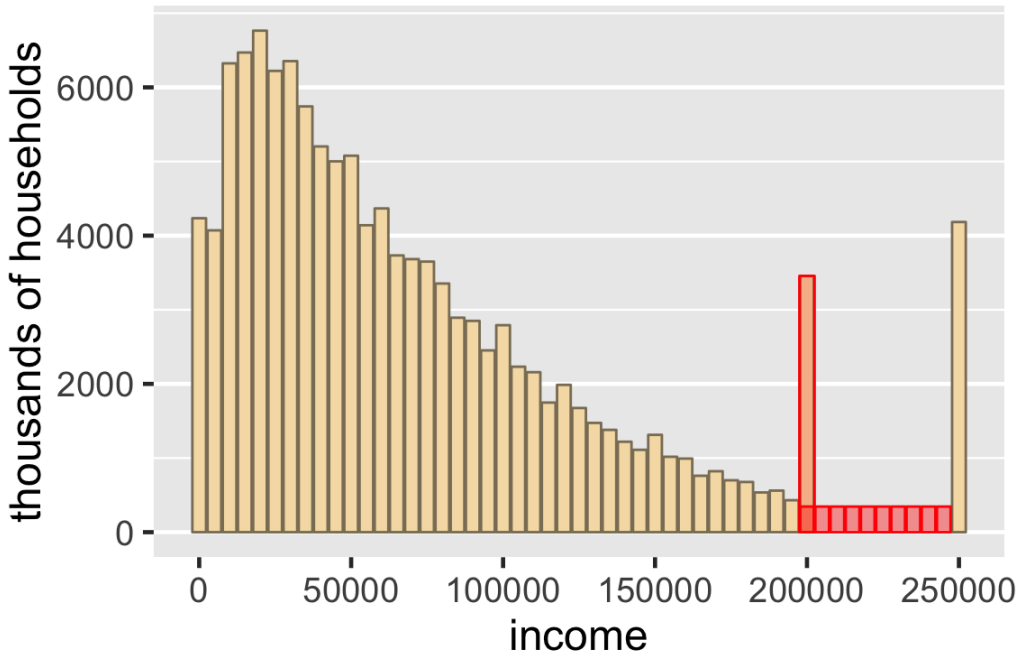
$82,500 to $84,999	$85,000 to $87,499	$87,500 to $89,999	$90,000 to $92,499	$92,500 to $94,999	$95,000 to $97,499	$97,500 to $99,999	$100,000 and over	Va (D
973	1,520	775	1,880	824	1,172	711	20,773	
323	550	265	634	309	381	238	7,479	
650	970	509	1,246	515	790	473	13,295	
129	163	117	185	70	134	59	1,653	

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

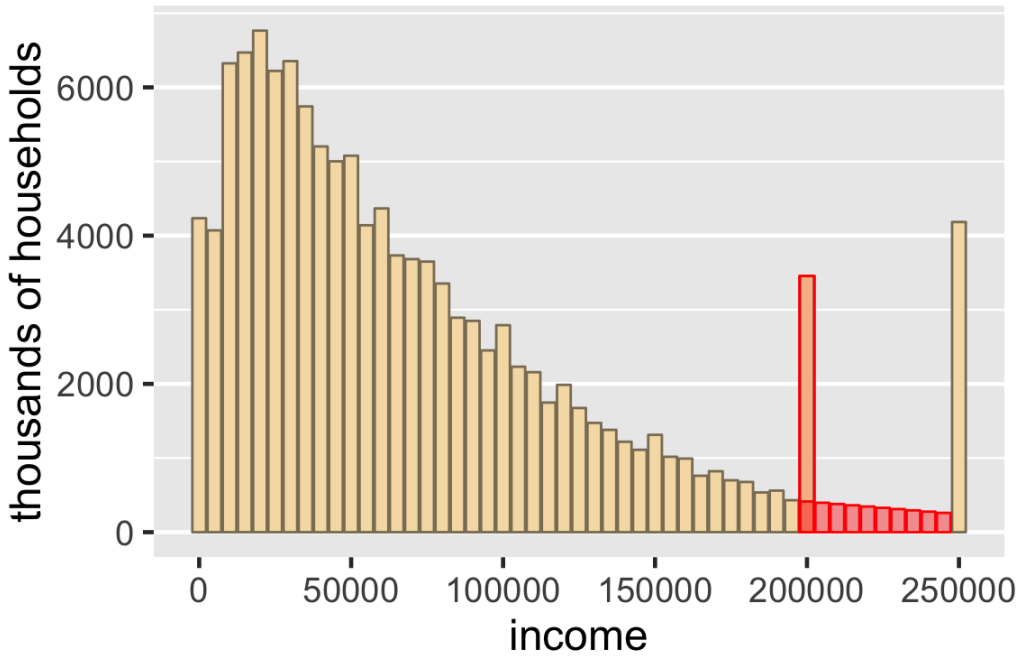
Household Income in 2015



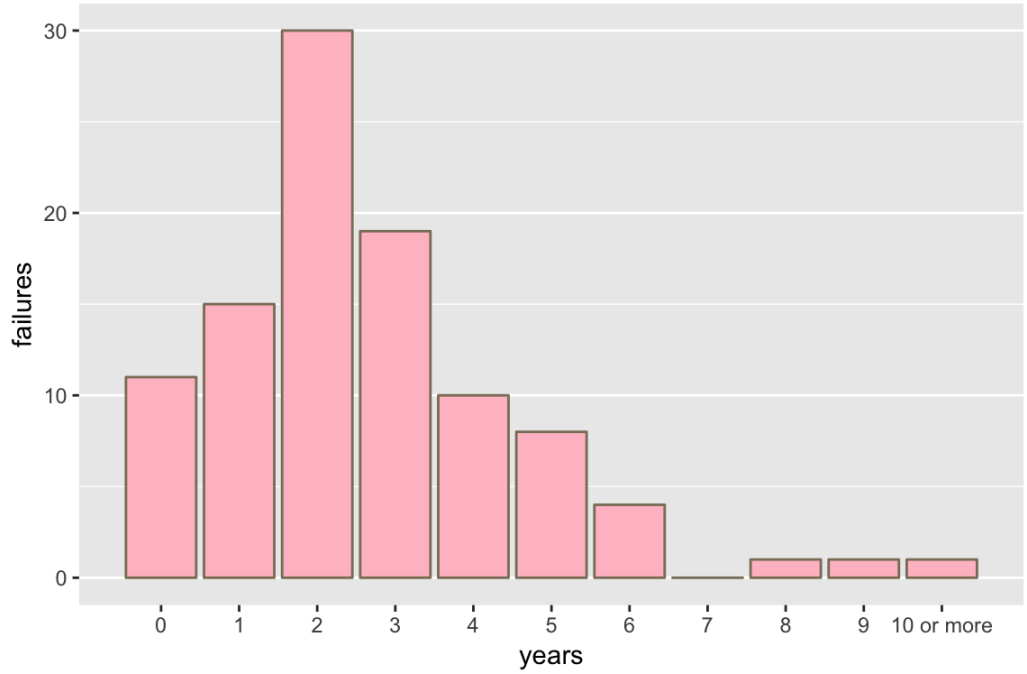
Household Income in 2015



Household Income in 2015



Reasonable use of “or more”



Data cleaning / transforming

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