Graphing Multivariate Categorical Data

The how, what and why of mosaic plots and alluvial diagrams

Joyce Robbins and Ludmila Janda

July 7, 2021

Agenda

Minute 0-5 Welcome

Minute 5-45 Mosaic plots with codealong (Joyce)

Minute 45-85 Alluvial diagrams with codealong (Ludmila)

Minute 85-90 Break

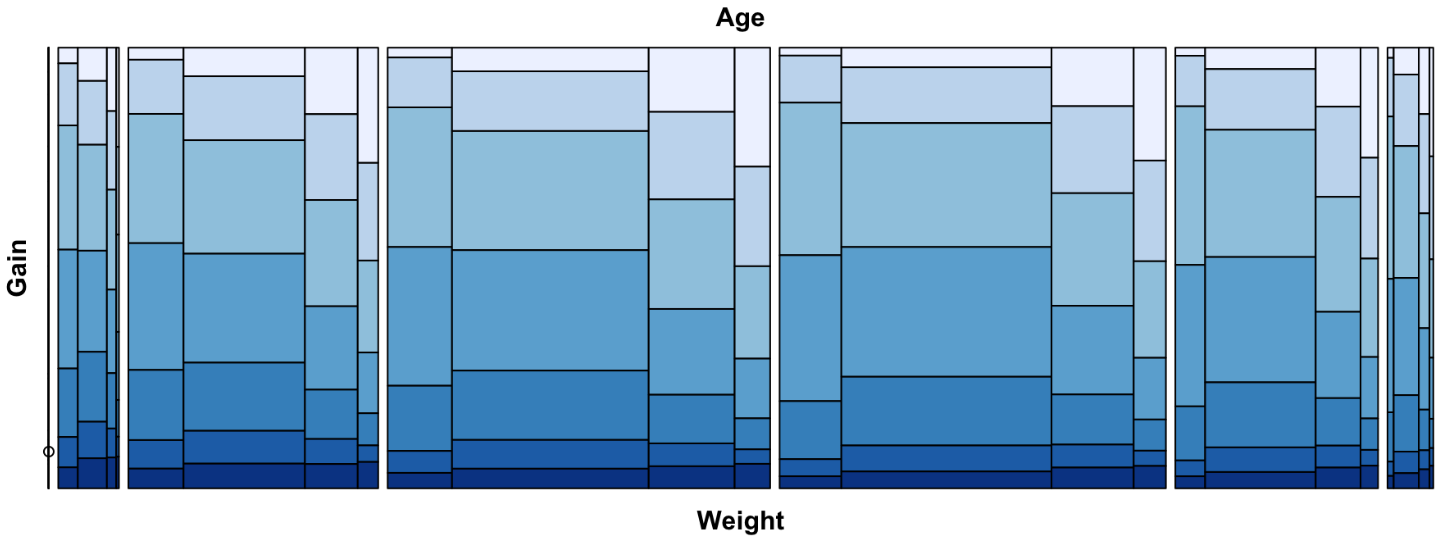
Minute 90-120 Lab (breakout rooms)

Minute 120-150 Discuss lab “results” / wrap up

Materials: https://github.com/jtr13/graphcat

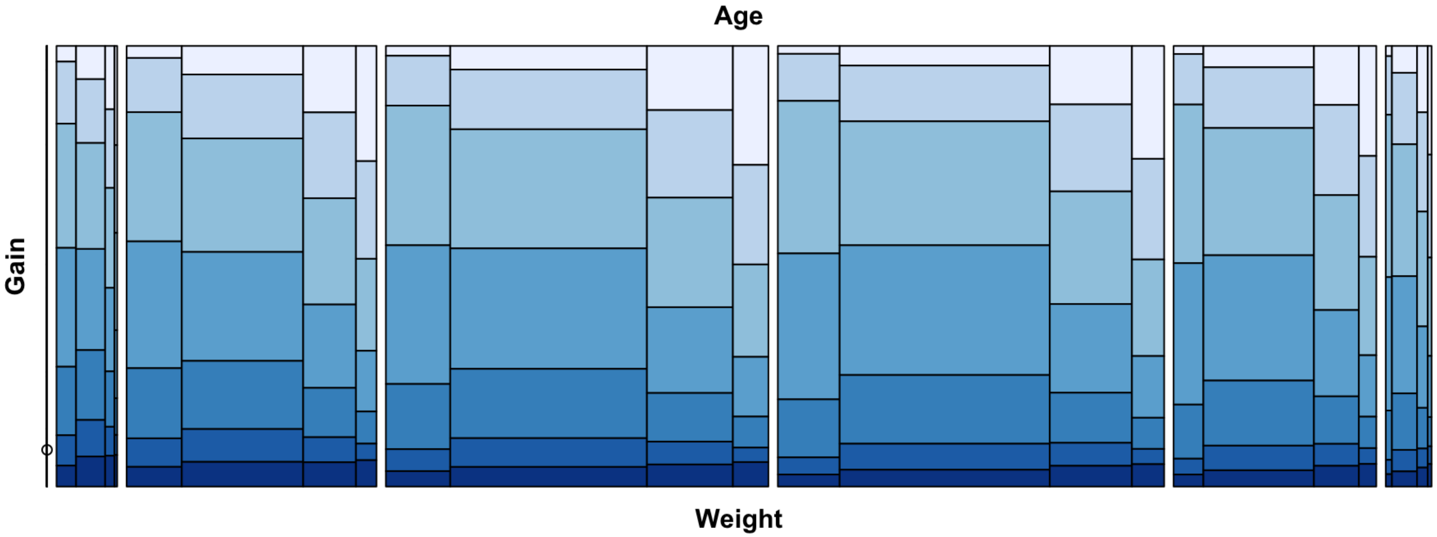
Graphing multivariate categorical data

mosaic plot space-filling visualization in which the area of each small rectangle is proportional to the frequency count for a unique combination of levels of the categorical variables displayed



Graphing multivariate categorical data

mosaic plot space-filling visualization in which the area of each small rectangle is proportional to the frequency count for a unique combination of levels of the categorical variables displayed

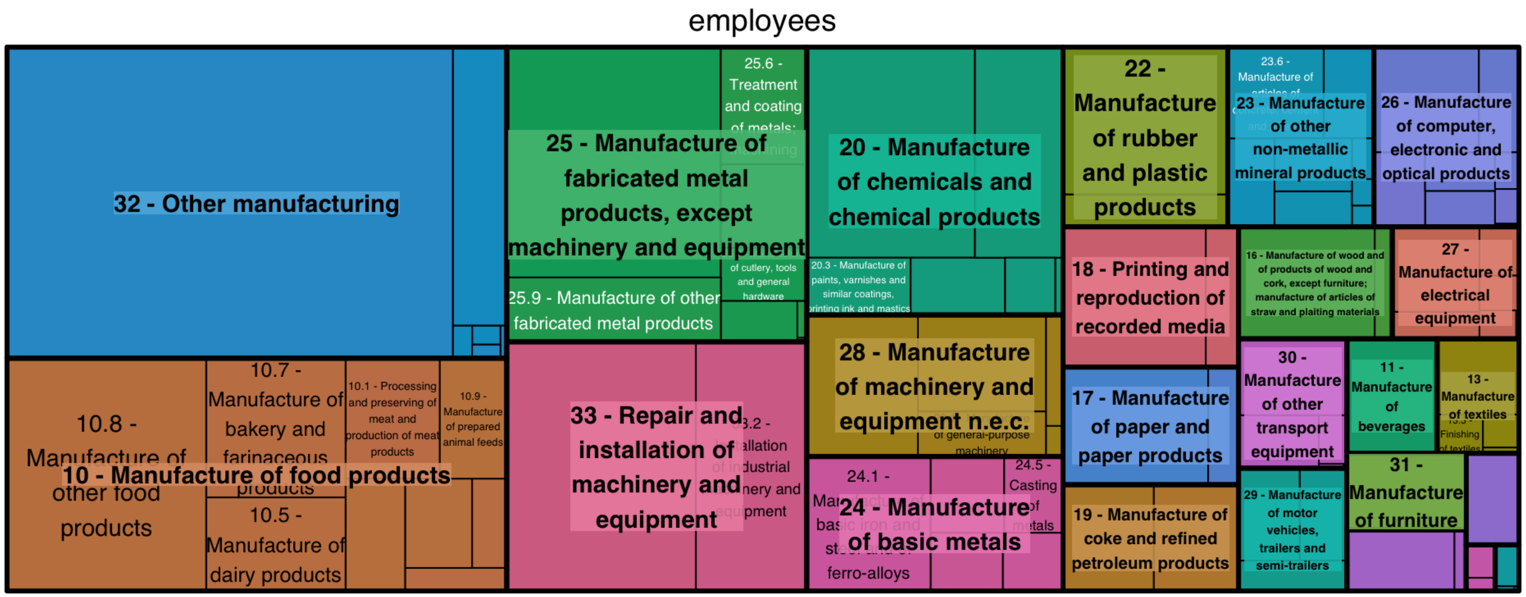


`Age`: 7 levels, `Weight`: 4 levels, `Gain`: 7 levels

–> 7 x 4 x 7 = 196 rectangles

Graphing multivariate categorical data

treemap “a space-filling visualization of hierarchical structures” (`?treemap::treemap`)



Numeric data

```
## 'data.frame':    15 obs. of  20 variables:
##  $ a1 : num  18.6 37.6 71.6 94.2 100.2 106.5 152.2 105.7 154.5 230.5 ...
##  $ a2 : num  17 38.2 67.8 106.8 64.2 134.7 133.4 108.4 108.5 264.1 ...
##  $ a3 : num  19 36.2 90.4 110.9 83.4 121.1 178.1 112.3 176.8 249.8 ...
##  $ a4 : num  6 48.6 77 115.5 94.1 208.3 199.1 79.4 95 157 ...
##  $ a5 : num  15.8 43.6 81.6 133 87.6 166.6 184.5 110.4 185.4 192.6 ...
##  $ a6 : num  0 22.8 36.6 111.2 54.8 116.5 167.1 59 150.7 144.2 ...
##  $ a7 : num  6.2 31 62 101.5 66.8 128.5 151.6 94 177.5 280.6 ...
##  $ a8 : num  5 30.2 31.1 89.7 53.5 104.6 151.5 54.2 190.1 212 ...
##  $ a9 : num  7.2 27 65 124.1 104.9 128.4 196.7 50.4 173.2 140.5 ...
##  $ a10: num  0 25.8 60.8 69.5 81.9 98.9 138.8 82 160.2 271.8 ...
##  $ a11: num  8 19.4 60.2 102.7 56.5 104.8 116.3 87.3 145.8 226.1 ...
##  $ a12: num  15 38 71.4 106.9 67.4 137.5 193.1 116.3 222.3 245.5 ...
##  $ a13: num  2.8 35.8 66.6 121.5 67.7 116.4 144.8 107.1 178.9 130.9 ...
##  $ a14: num  4.4 35.4 48 120.7 41 114.5 155.6 127.8 188.5 264.1 ...
##  $ a15: num  6.6 34.8 52 100.6 78 109.7 126.7 86.1 156.6 230.9 ...
##  $ a16: num  4 28.6 34.1 101.5 40.1 113.4 114.1 80.7 169 249.6 ...
##  $ a17: num  2.4 41.2 30 116.4 11.2 181.4 41.2 151.4 33.6 261.2 ...
##  $ a18: num  9.6 24.4 54 103.9 67.4 112.5 139.2 82.3 183.6 196 ...
##  $ a19: num  0 33.8 47.6 111.7 79.7 169.9 8 116.8 191.7 271.2 ...
##  $ a20: num  2.2 31.2 57.6 127.7 65.5 134.2 120.7 97.9 203 237.3 ...
```

Data: `clementines` from **ade4** package

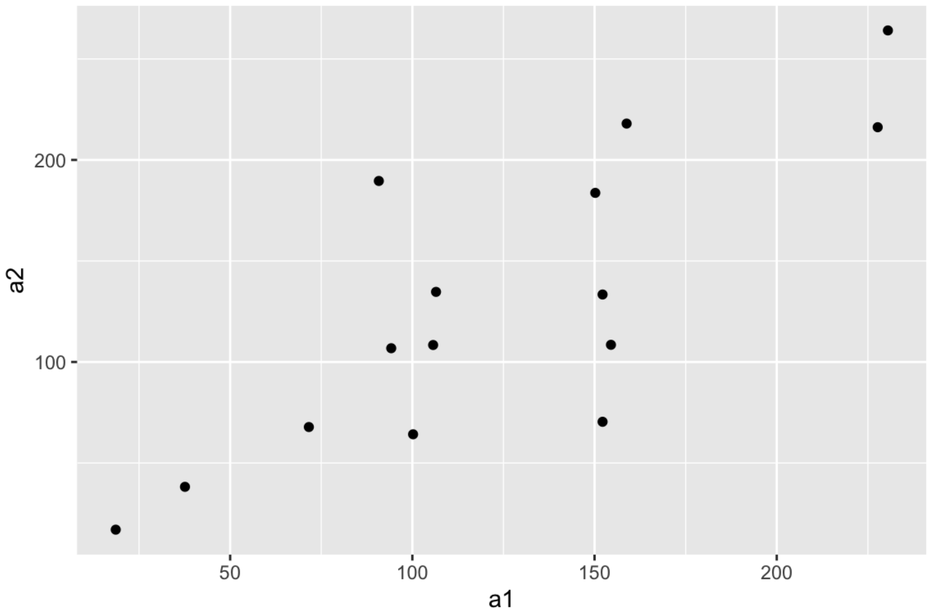
Categorical data

```
## tibble [1,373 × 12] (S3: tbl_df/tbl/data.frame)
##  $ respondent_id   : num [1:1373] 3308895255 3308891308 3308891135 3308879091 3308871671 ...
##  $ knowledge       : Ord.factor w/ 4 levels "Novice"<"Intermediate"<..: 2 1 2 1 1 3 1 3 1 1 ...
##  $ interest        : Ord.factor w/ 4 levels "Not at all"<"Not much"<..: 3 3 4 2 2 4 3 4 2 3 ...
##  $ gender          : chr [1:1373] "Male" "Male" "Male" "Male" ...
##  $ age             : Factor w/ 4 levels "18-29","30-44",..: 1 1 2 3 2 2 3 3 2 NA ...
##  $ household_income: Factor w/ 5 levels "$0 - $24,999",..: 4 4 3 1 2 3 NA 1 3 NA ...
##  $ education       : Ord.factor w/ 5 levels "Less than high school degree"<..: 1 3 5 1 2 5 2 3 3 NA ...
##  $ location        : chr [1:1373] "West South Central" "West South Central" "Pacific" "New England" ...
##  $ algeria         : chr [1:1373] "N/A" "N/A" "3" "N/A" ...
##  $ argentina       : chr [1:1373] "3" "N/A" "4" "3" ...
##  $ australia       : chr [1:1373] "5" "3" "N/A" "N/A" ...
##  $ belgium         : chr [1:1373] "4" "3" "3" "3" ...
```

Data: `food_world_cup` from **fivethirtyeight** package

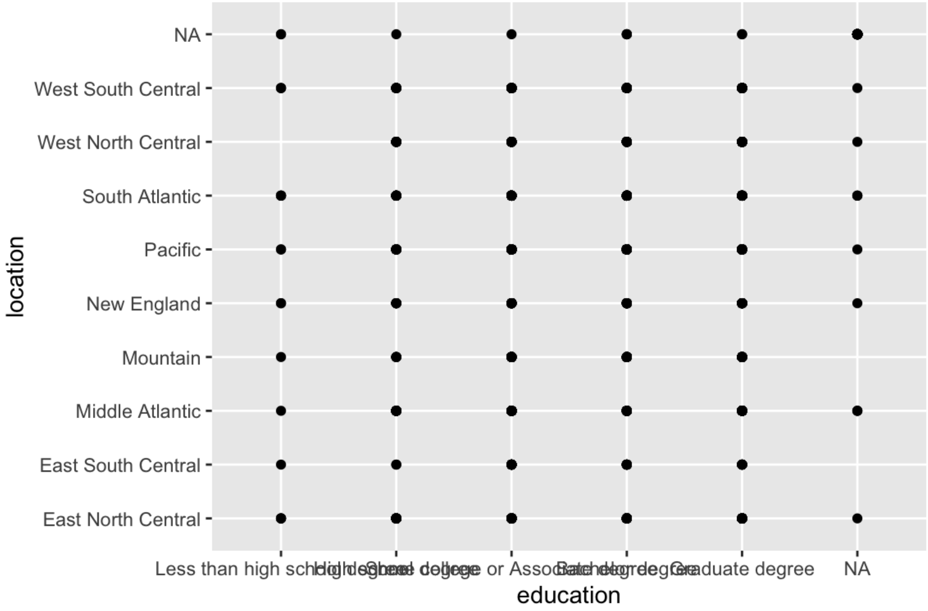
Graphing numerical data

```r
library(tidyverse)
ggplot(clementines, aes(a1, a2)) + geom_point()
```



Graphing categorical data

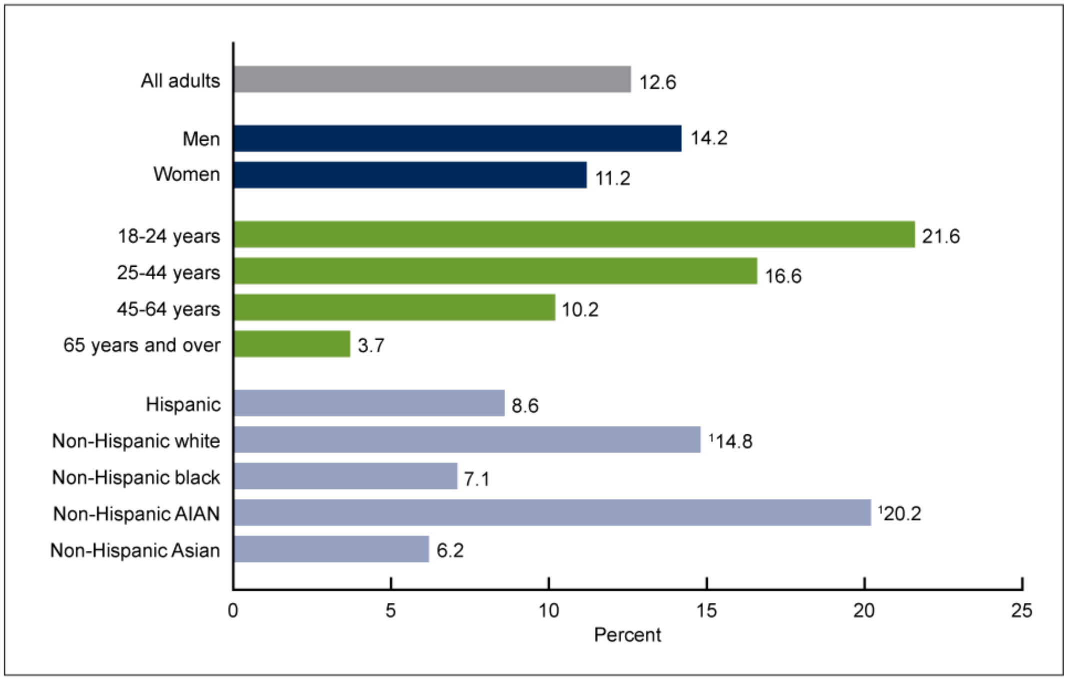
```r
ggplot(food_world_cup, aes(education, location)) + geom_point()
```



What does multivariate data look like?

Multiple variables but not a multivariate plot:

Figure 1. Percentage of adults who had ever tried an e-cigarette in their lifetime, by sex, age, and race and Hispanic or Latino origin: United States, 2014



[1]Significantly different from Hispanic, non-Hispanic black, and non-Hispanic Asian subgroups.
NOTES: AIAN is American Indian or Alaska Native. Within sex and age groups, all subgroups are significantly different from each other. There is a significant linear trend by age group.
SOURCE: CDC/NCHS, National Health Interview Survey, 2014.

https://www.cdc.gov/nchs/images/databriefs/201-250/db217_fig1.png

Multivariate categorical data

Frequency

- Bar charts
- Cleveland dot plots

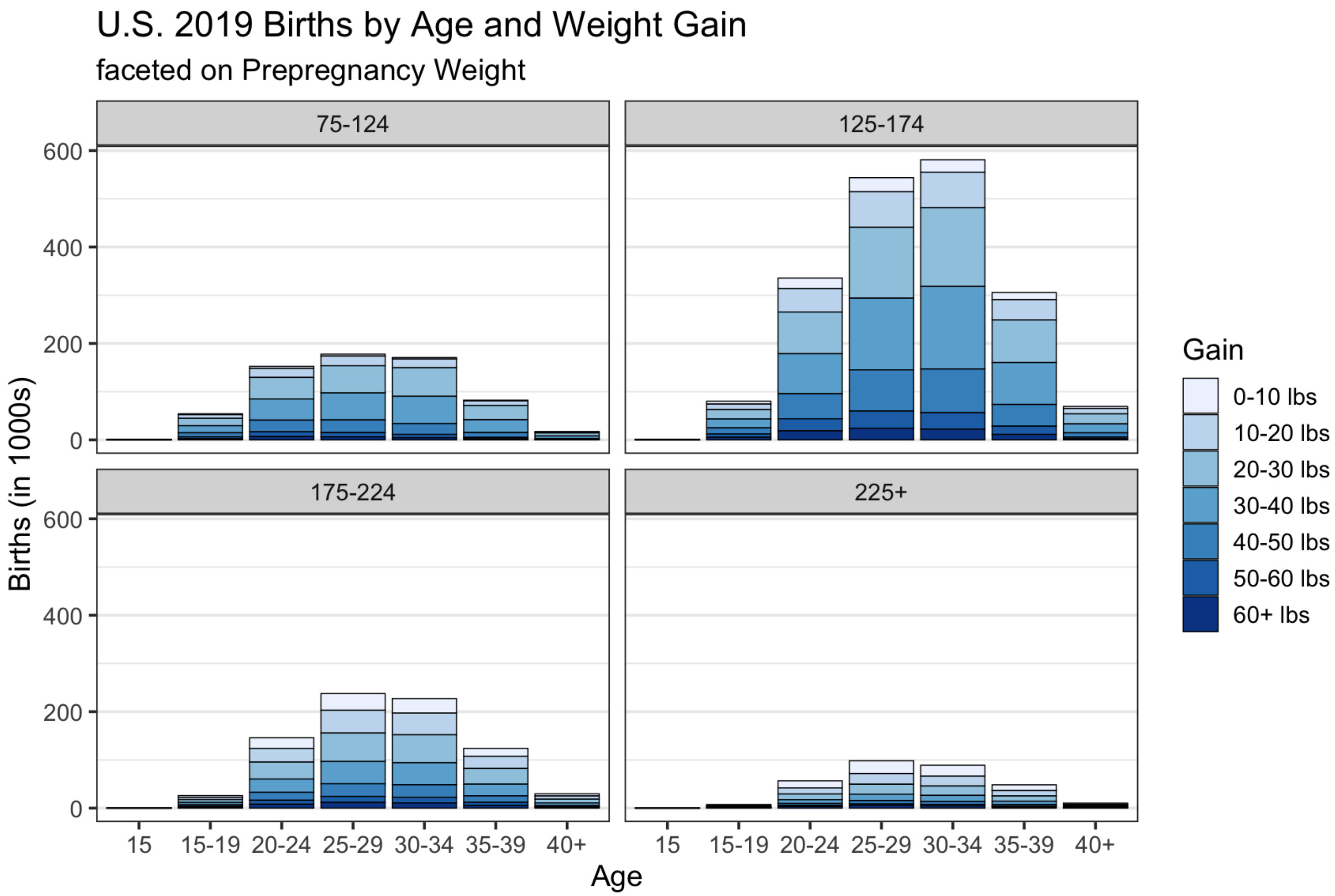
Proportion / Association

- Mosaic plots

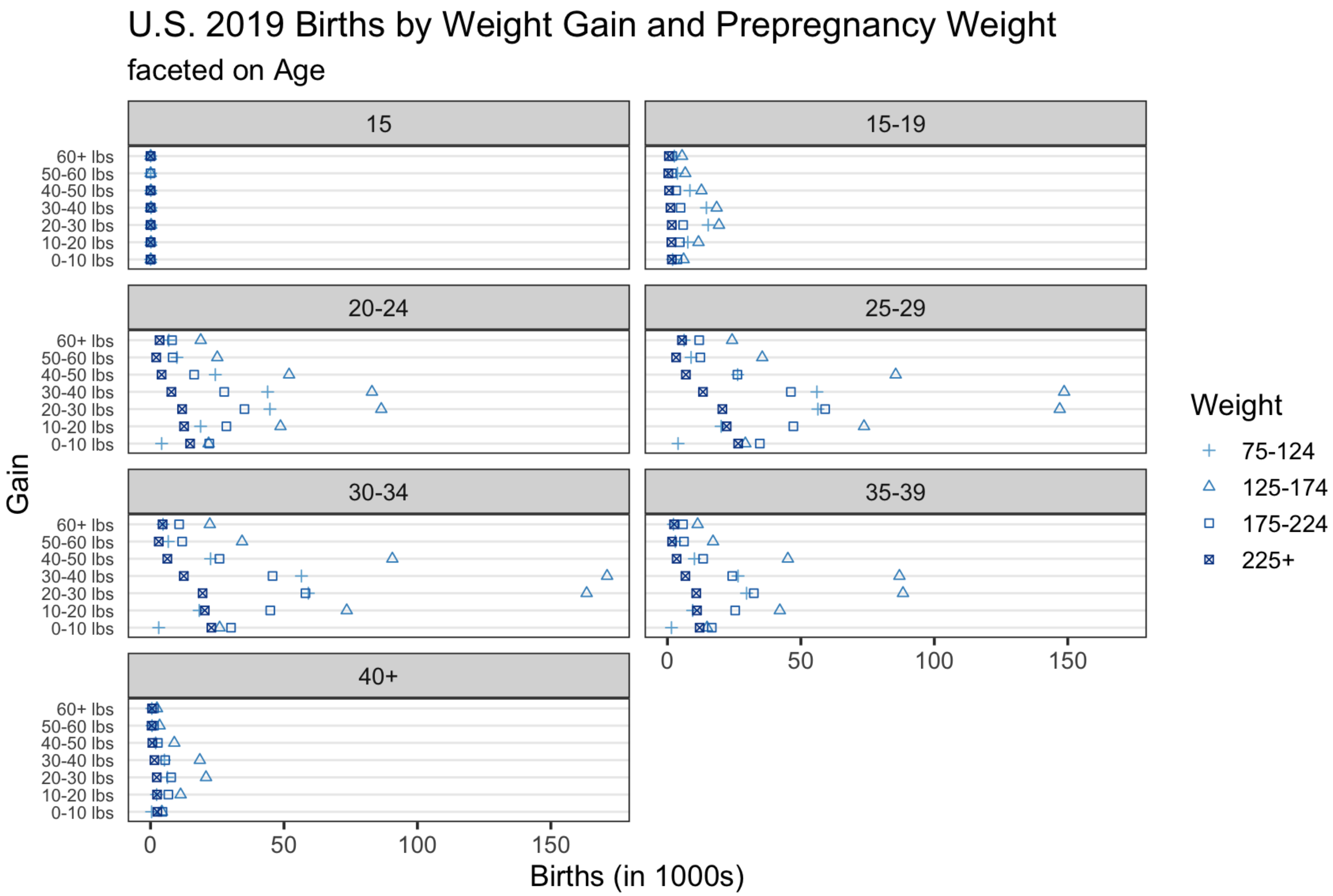
Change of state

- Alluvial diagrams

Stacked and faceted bar charts



Cleveland dot plot



Proportion / Association

Are older Americans more interested in local news than younger Americans?

34892 U.S. adults were asked whether or not they follow local news “very closely”. 34.5% said yes.

Group sizes are:

```
##     Age  Freq
## 1 18-29  2851
## 2 30-49  9967
## 3 50-64 11163
## 4   65+ 10911
```

Source: https://www.journalism.org/2019/08/14/methodology-local-news-demographics/

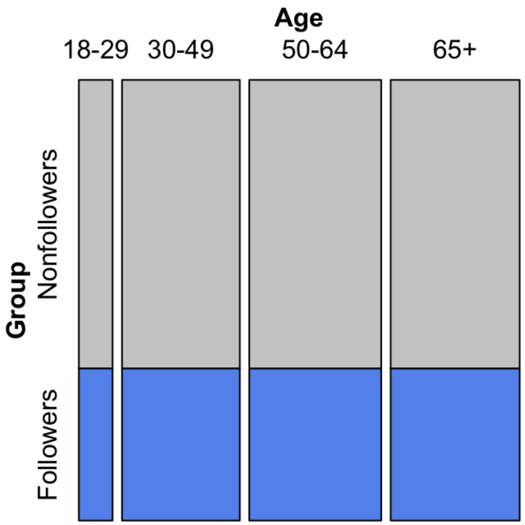
If older Americans are **NOT** more interested in local news, what would the breakdowns look like?

Assumption of no association between age and group

```
##     Age  Freq Followers Nonfollowers
## 1 18-29  2851       984         1867
## 2 30-49  9967      3439         6528
## 3 50-64 11163      3851         7312
## 4   65+ 10911      3764         7147
```

Assumption of no association between age and group

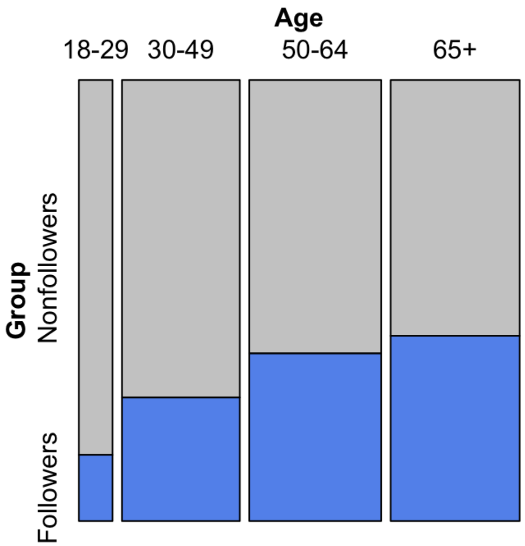
```
##     Age  Freq Followers Nonfollowers
## 1 18-29  2851       984         1867
## 2 30-49  9967      3439         6528
## 3 50-64 11163      3851         7312
## 4   65+ 10911      3764         7147
```



34.5% follow local news regardless of age

Mosaic plot of actual data

```
##     Age  Freq Followers Nonfollowers
## 1 18-29  2851       428         2423
## 2 30-49  9967      2791         7176
## 3 50-64 11163      4242         6921
## 4   65+ 10911      4583         6328
```



Chi Square Test of Independence

Null hypothesis: Age and tendency to follow local news are independent

Alternative hypothesis: Age and tendency to follow local news are NOT independent

We compare OBSERVED to EXPECTED:

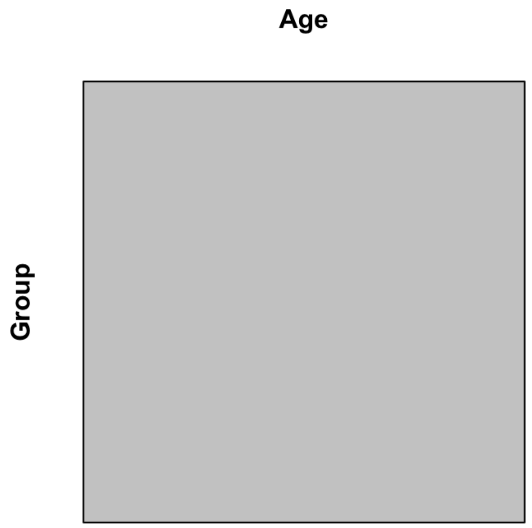
```
##       Followers Nonfollowers
## 18-29       428         2423
## 30-49      2791         7176
## 50-64      4242         6921
## 65+        4583         6328
```

```
##       Followers Nonfollowers
## 18-29       984         1867
## 30-49      3440         6527
## 50-64      3853         7310
## 65+        3766         7145
```

```
##
##  Pearson's Chi-squared test
##
## data:  localmat
## X-squared = 997, df = 3, p-value <0.0000000000000002
```

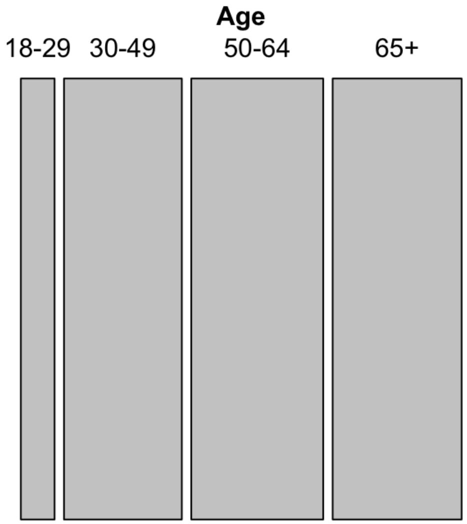
Creating mosaic plots

start with a rectangle



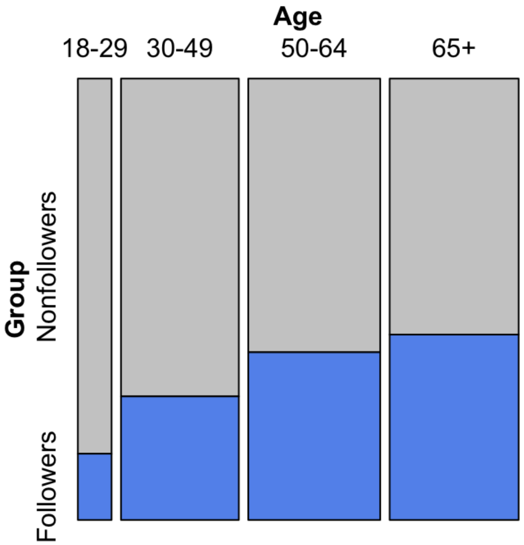
1st cut: vertical

independent variable (Age)



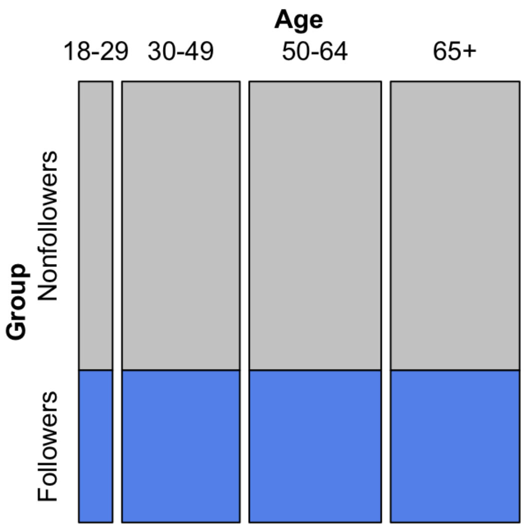
2nd cut: horizontal

dependent variable (Group)



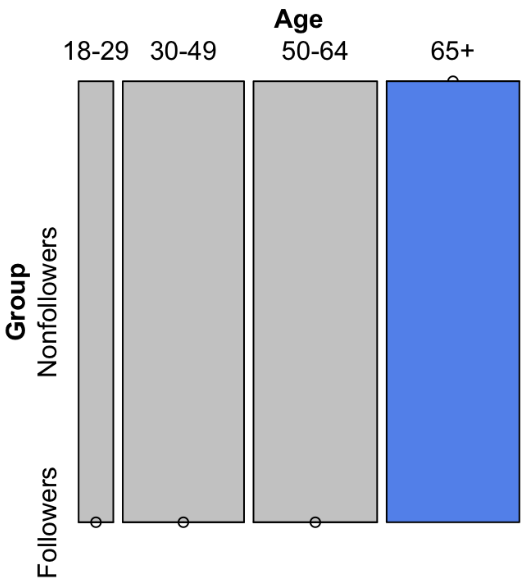
Mosaic plot

no association

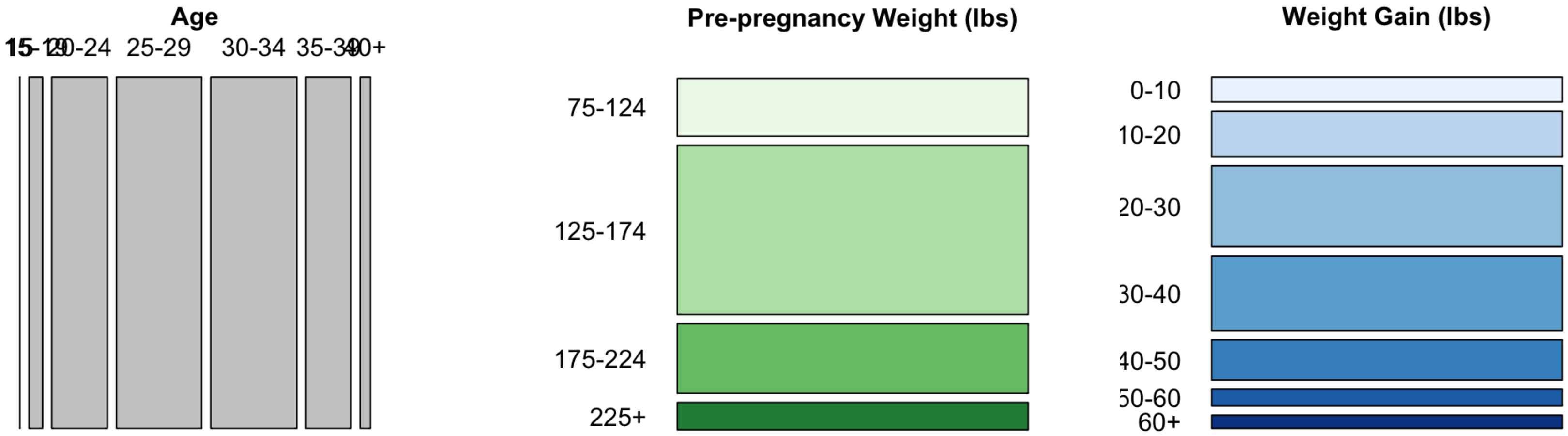


Mosaic plot

deterministic relationship



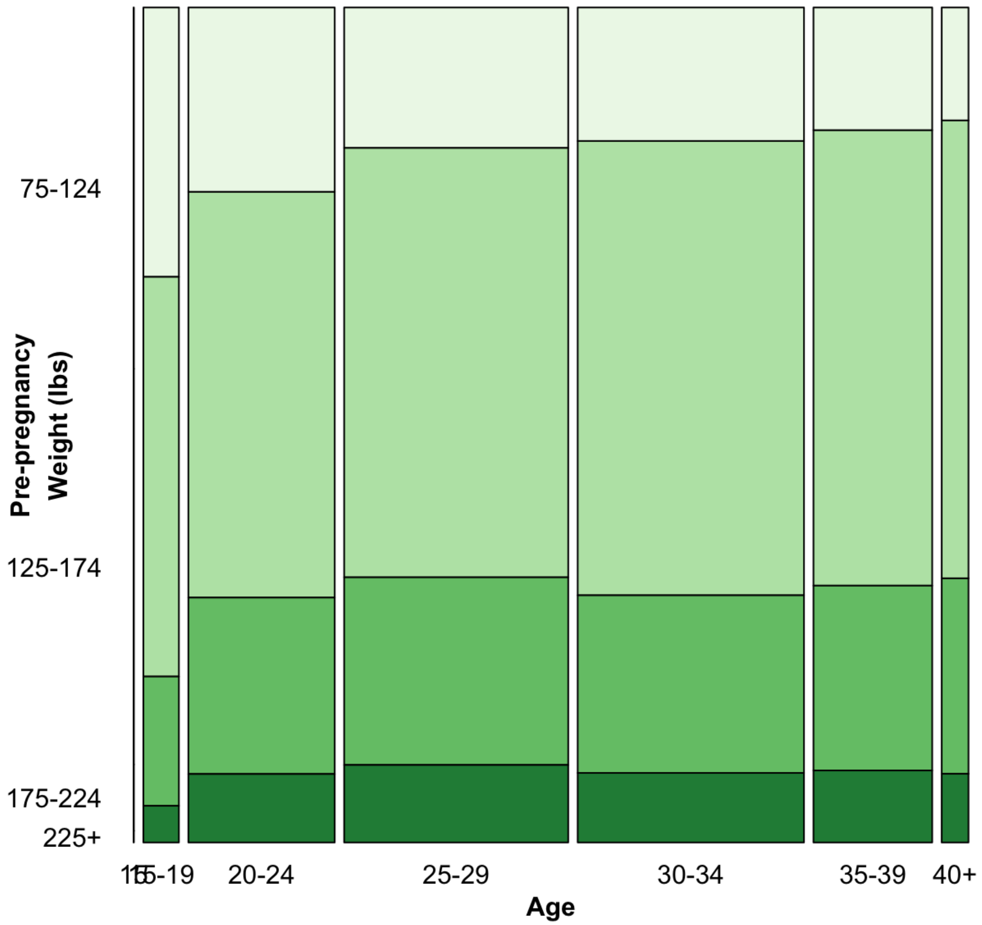
Birth data, U.S. 2019



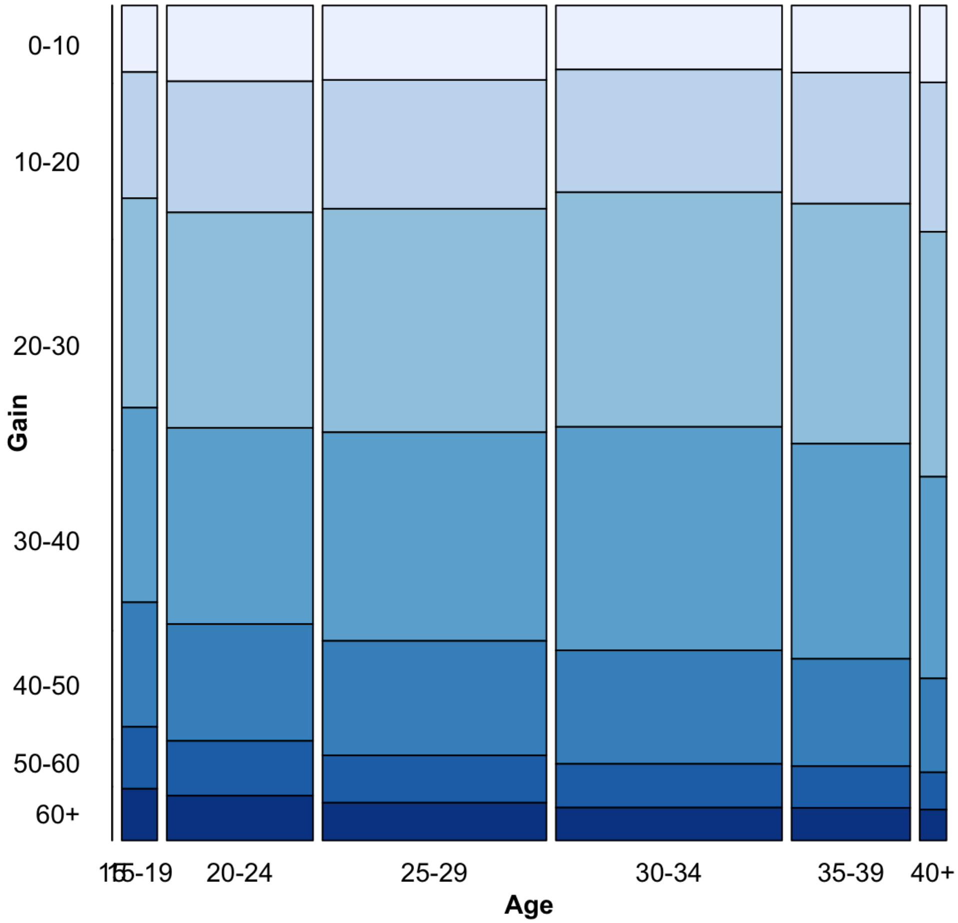
Source: https://wonder.cdc.gov/natality-current.html

https://github.com/jtr13/graphcat/blob/main/data/age_preweight_gain.txt

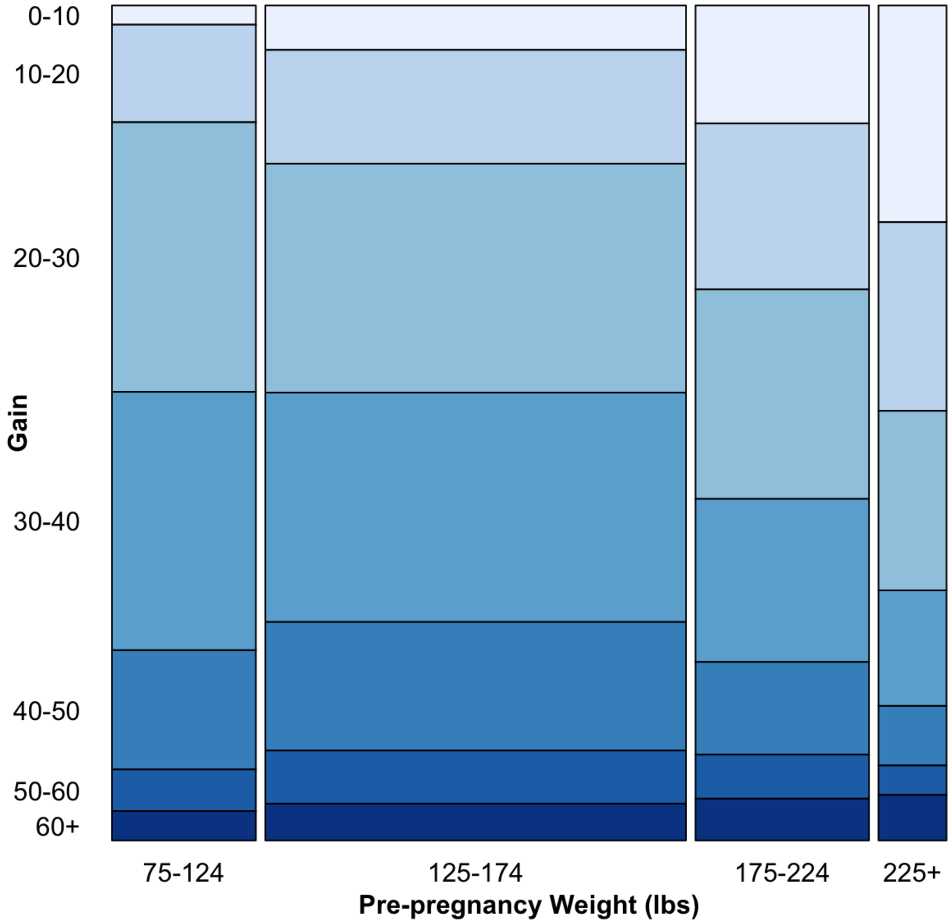
Weight vs. Age

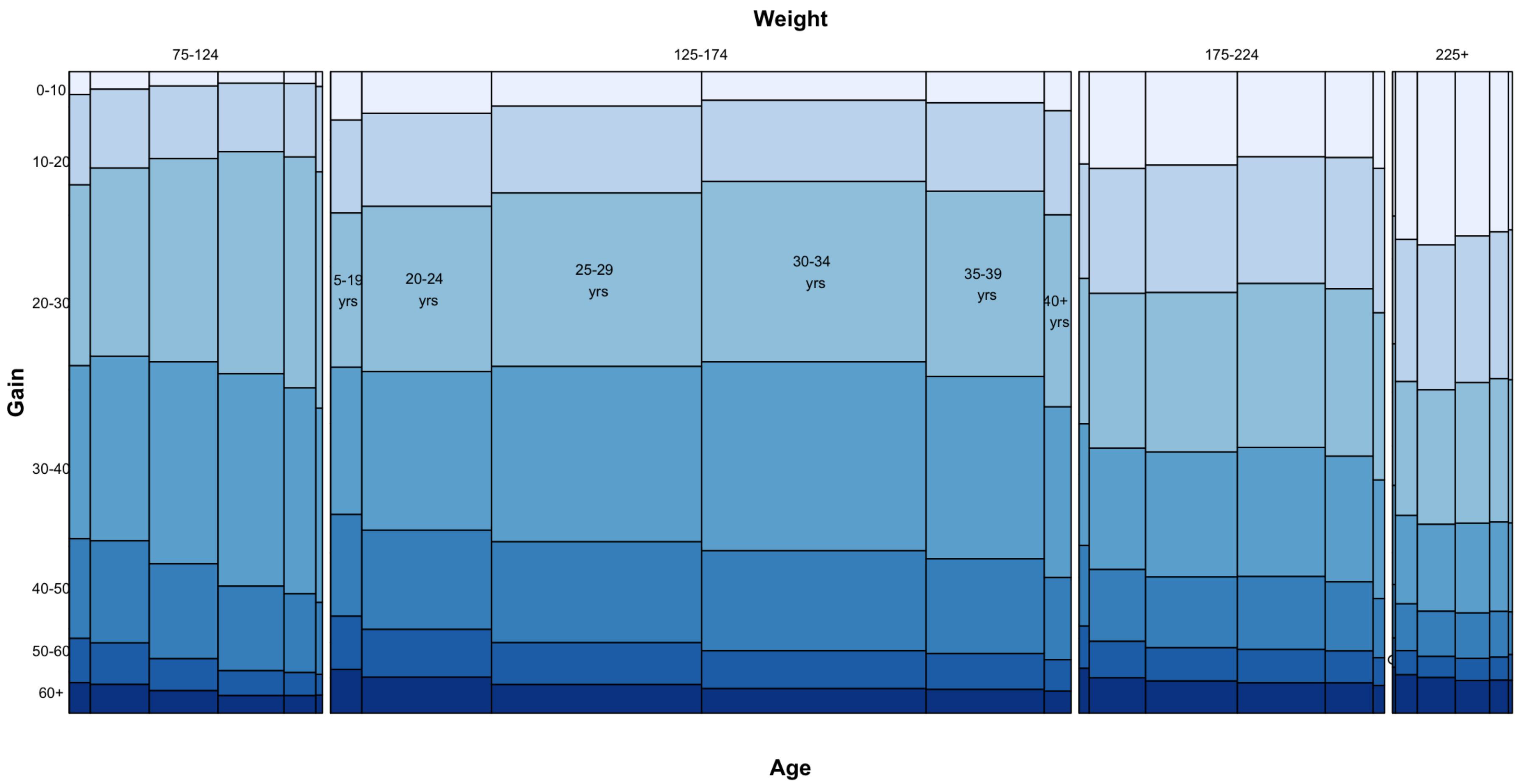


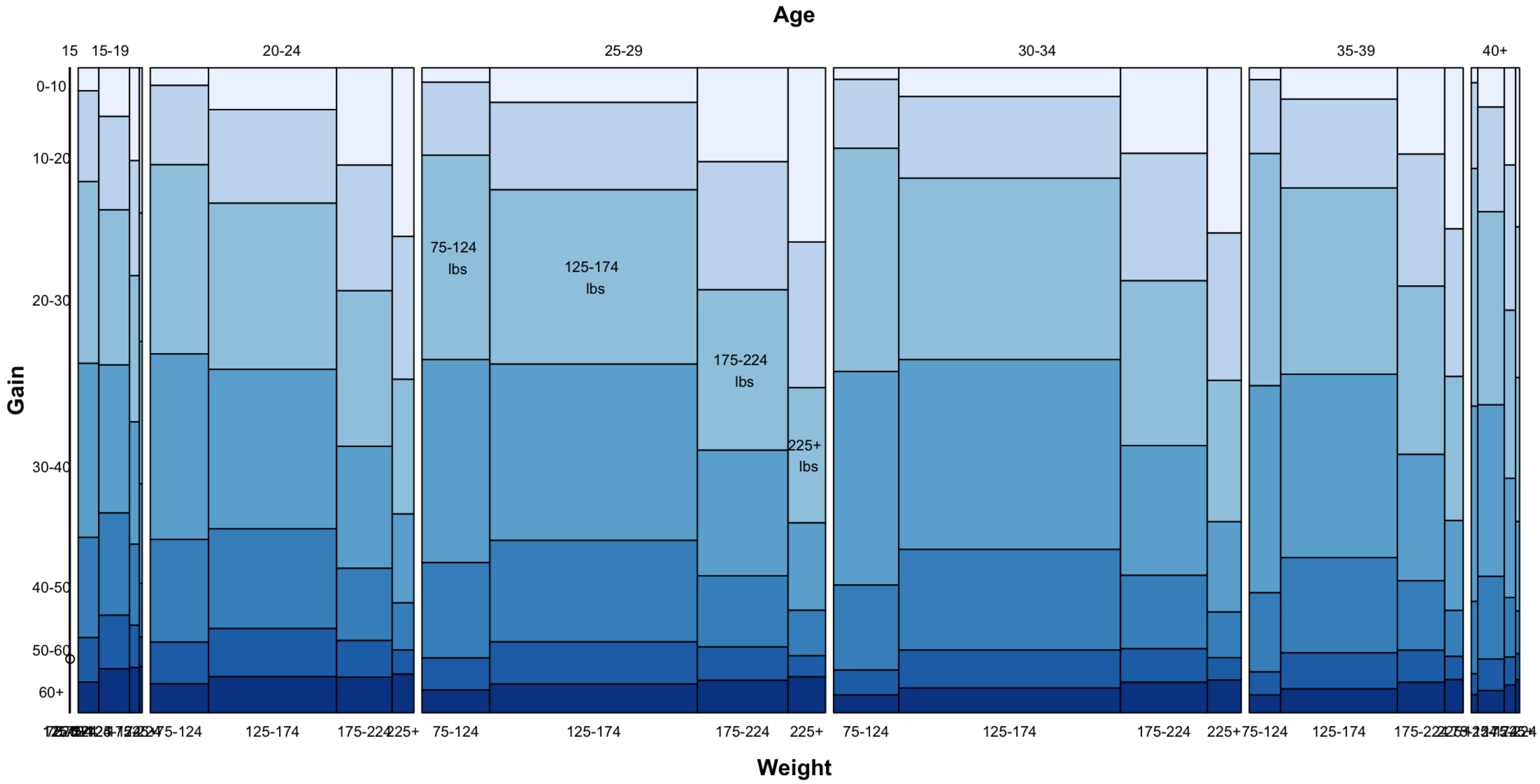
Gain vs. Age


0-10
10-20
20-30
30-40
40-50
50-60
60+
Gain
15
15-19
20-24
25-29
30-34
35-39
40+
Age

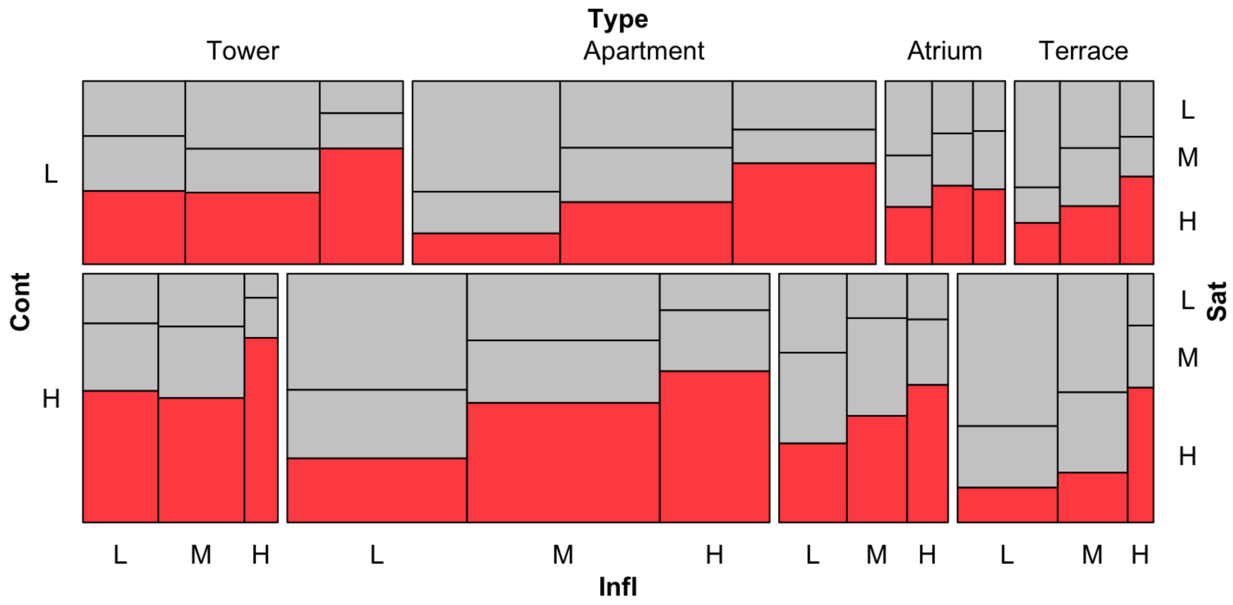
Gain vs. Weight




Weight
75-124
125-174
175-224
225+
Gain
0-10
10-20
20-30
30-40
40-50
50-60
60+
5-19
yrs
20-24
yrs
25-29
yrs
30-34
yrs
35-39
yrs
40+
yrs
Age


Age
15
15-19
20-24
25-29
30-34
35-39
40+
Gain
0-10
10-20
20-30
30-40
40-50
50-60
60+
75-124
lbs
125-174
lbs
175-224
lbs
225+
lbs
75-124
125-174
175-224
225+
Weight

Mosaic plot with four variables



`MASS::housing`

`Sat`: Satisfaction of householders with their present housing circumstances, (High, Medium or Low, ordered factor).

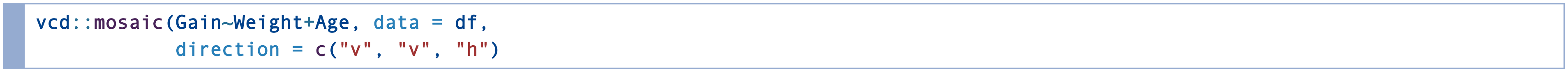
`Infl`: Perceived degree of influence householders have on the management of the property (High, Medium, Low).

`Type`: Type of rental accommodation, (Tower, Atrium, Apartment, Terrace).

`Cont`: Contact residents are afforded with other residents, (Low, High).

Mosaic plot best practices

- Dependent variables is split last and split *horizontally*
- `hightlighting_fill` only affects *dependent* variable
- Other variables are generally split vertically



```
vcd::mosaic(Gain~Weight+Age, data = df,
            direction = c("v", "v", "h")
```

- Most important *level* of dependent variable is closest to the x-axis and darkest (or most noticable shade)

See: Antony Unwin, *Graphical Data Analysis with R*, CRC Press, 2015.

next: https://github.com/jtr13/graphcat/blob/main/labs/Mosaic_codealong.R