Descriptive Statistics II

Juan Pablo Lewinger

6/21/2024

Descriptive Statistics

- Univariate methods:
  - quantitative variables: histograms, boxplots, mean/sd (for symmetric vars), median/IQR (for skewed vars)
  - Categorical variables: barplots, table, counts, percentages
- Typical goal in data analysis is to understand the relationship (associations) between pairs of variables
- Today we'll focus on bivariate descriptive statistics
- Bivariate descriptive statistics can provide initial clues about associations

CHS data

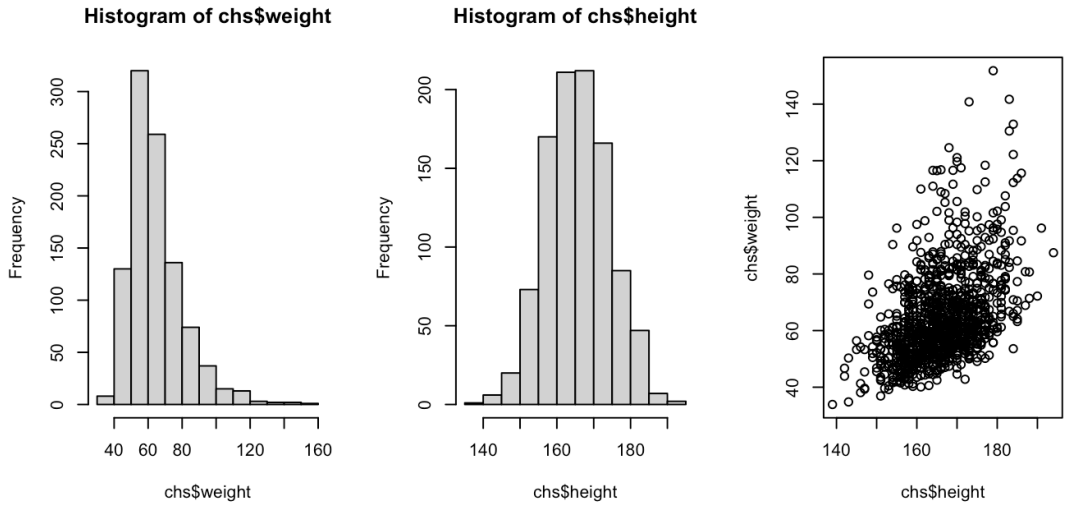
```
setwd("/Users/JP/My Drive/Teaching/LAs BEST/2023/Lectures/3. Descriptive stats bivariate")
chs = read.csv('CHS_cohortE_final_subset.csv')
str(chs)
```

```
## 'data.frame':    1000 obs. of  26 variables:
##  $ id          : int  54577 50863 52081 53817 54683 55339 55766 51056 54919 52992 ...
##  $ townabbr    : chr  "SA" "SD" "SD" "RV" ...
##  $ age         : num  15.1 16.5 15.6 15.2 14.2 15.2 15.8 16 15.2 16.1 ...
##  $ male        : int  1 1 0 1 1 0 1 1 0 0 ...
##  $ race        : chr  "Others" "Mixed" "Caucasian" "Unknown or Missing" ...
##  $ hisp        : chr  "Hispanic" "Hispanic" "Non-Hispanic" "Hispanic" ...
##  $ asthma      : int  0 NA 0 0 1 0 0 1 0 0 ...
##  $ height      : int  168 168 167 160 169 161 185 183 163 165 ...
##  $ weight      : num  52 50.2 55.6 60.9 62.1 ...
##  $ bmi         : num  18.4 17.8 19.9 23.8 21.8 20.6 33.2 39 28.6 28.4 ...
##  $ educ        : int  1 1 3 2 5 5 2 3 1 2 ...
##  $ HomeBuilt   : chr  "1980 or later" "Unknown or Missing" "1960s to 1970s" "Unknown or Missing" ...
##  $ BaseGasstove: int  1 0 1 1 1 0 1 1 1 1 ...
##  $ BasePets    : int  1 0 1 0 1 1 1 1 1 0 ...
##  $ ETS_base    : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ wheeze      : int  0 NA 0 0 0 0 0 0 0 0 ...
##  $ fev1        : int  4090 3790 3240 3890 3730 3530 5420 4480 3290 3390 ...
##  $ fvc         : int  4950 4810 3370 4190 4930 4010 6360 5590 3450 3930 ...
##  $ pm25        : num  8.84 14.28 15 15.76 14.18 ...
##  $ sulfate     : num  0.93 1.38 1.46 1.57 1.32 ...
##  $ nitrate     : num  1.87 2.28 2.48 2.45 2.18 ...
##  $ ec          : num  0.702 0.873 0.884 0.762 0.893 ...
##  $ dust        : num  0.449 1.302 1.246 1.29 1.34 ...
##  $ longitude   : num  -120 -118 -118 -117 -118 ...
##  $ latitude    : num  34.5 34.1 34.1 34 34.1 ...
##  $ obesity     : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

Quantitative vs. quantitative variables

Graphical summary: scatter plots

```
par(mfrow = c(1,3))
hist(chs$weight)
hist(chs$height)
plot(chs$height, chs$weight)
```



Many R packages for generating plots. ggplot2 is among the most popular

Quantitative vs. quantitative

Numerical summary: Pearson correlation coefficient

$$r = corr(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{sd(x)sd(y)}$$

$$-1 \leq r \leq 1$$

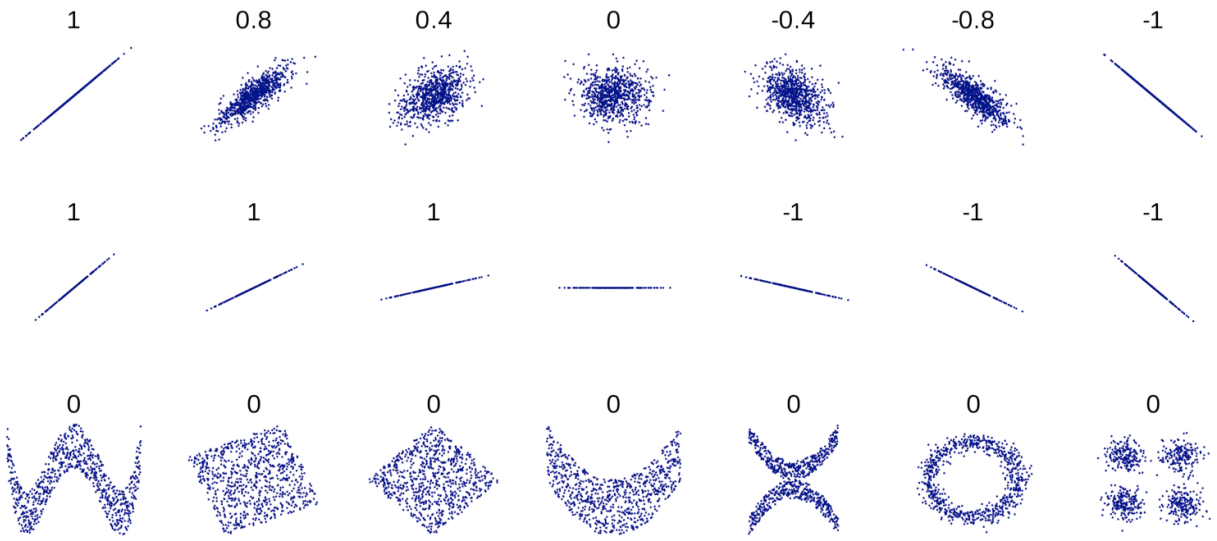
Captures strength of linear relationship between x and y

```
cor(chs$height, chs$weight)
```

```
## [1] 0.450752
```

Quantitative vs. quantitative

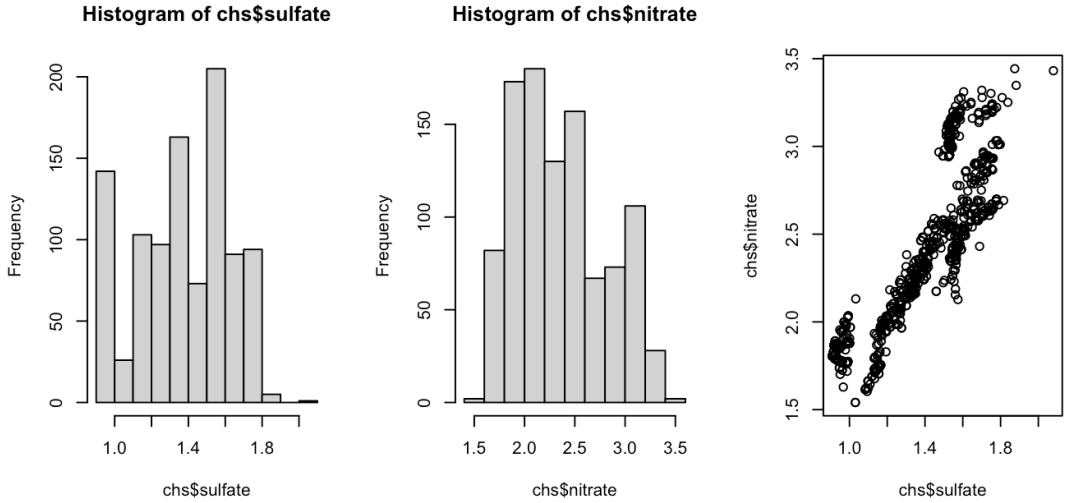
Correlation examples



Source: Wikipedia

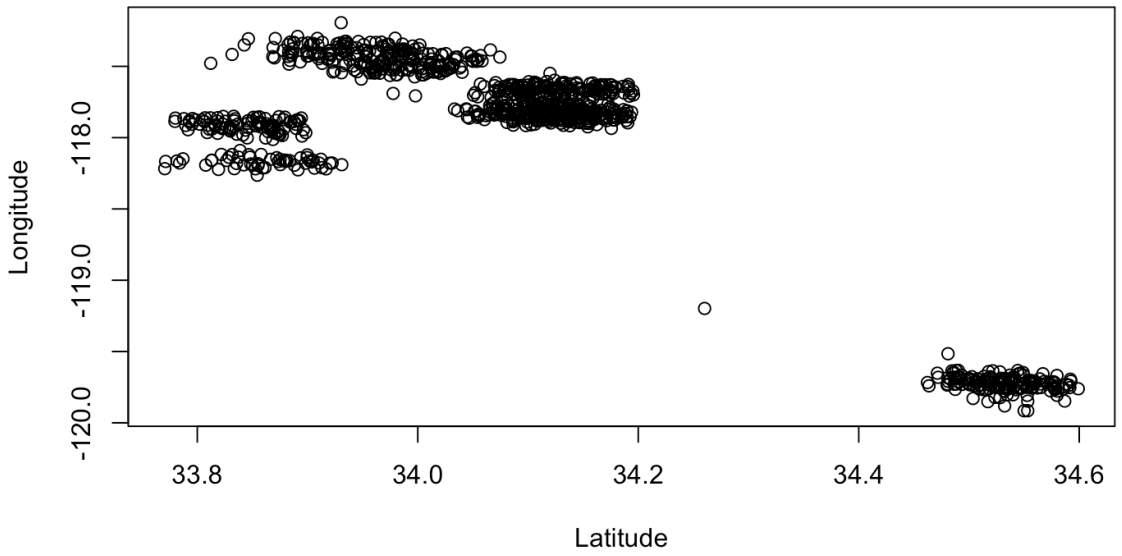
Quantitative vs. quantitative

```
par(mfrow = c(1,3))
hist(chs$sulfate)
hist(chs$nitrate)
plot(chs$sulfate, chs$nitrate)
```



Quantitative vs. quantitative

```
plot(chs$latitude, chs$longitude, xlab = 'Latitude', ylab = 'Longitude')
```

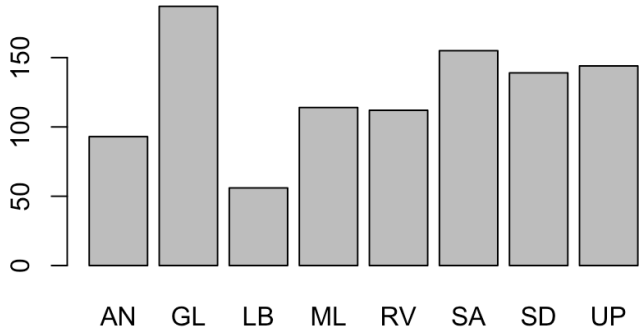


Quantitative vs. categorical

```
table(chs$townabbr)
```

```
##
##  AN  GL  LB  ML  RV  SA  SD  UP
##  93 187  56 114 112 155 139 144
```

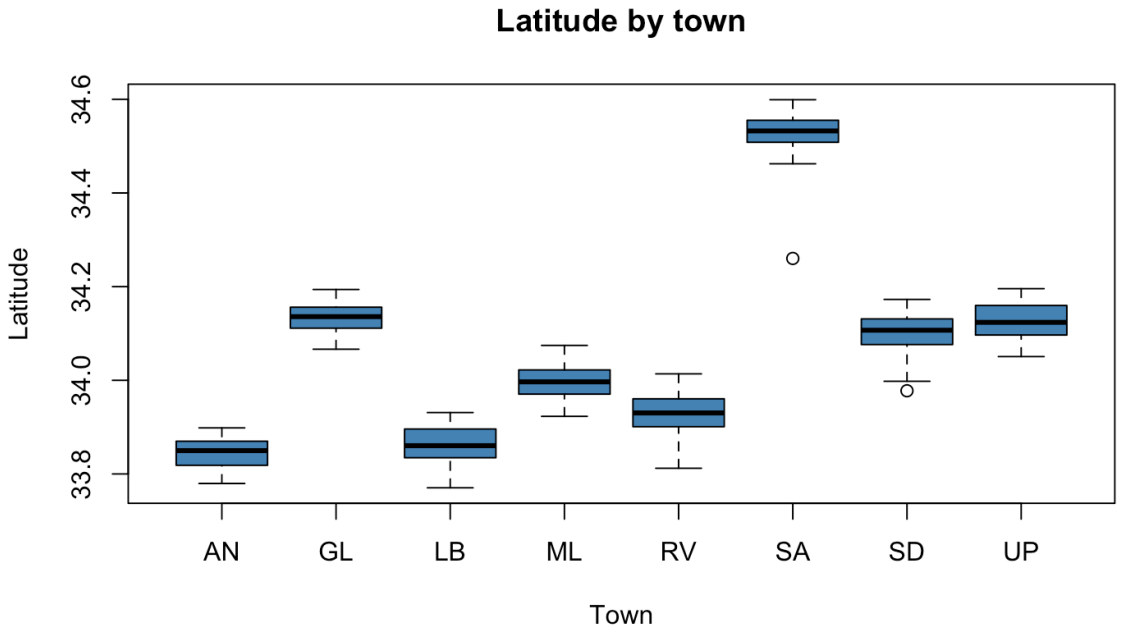
```
barplot(table(chs$townabbr))
```



Quantitative vs. Categorical

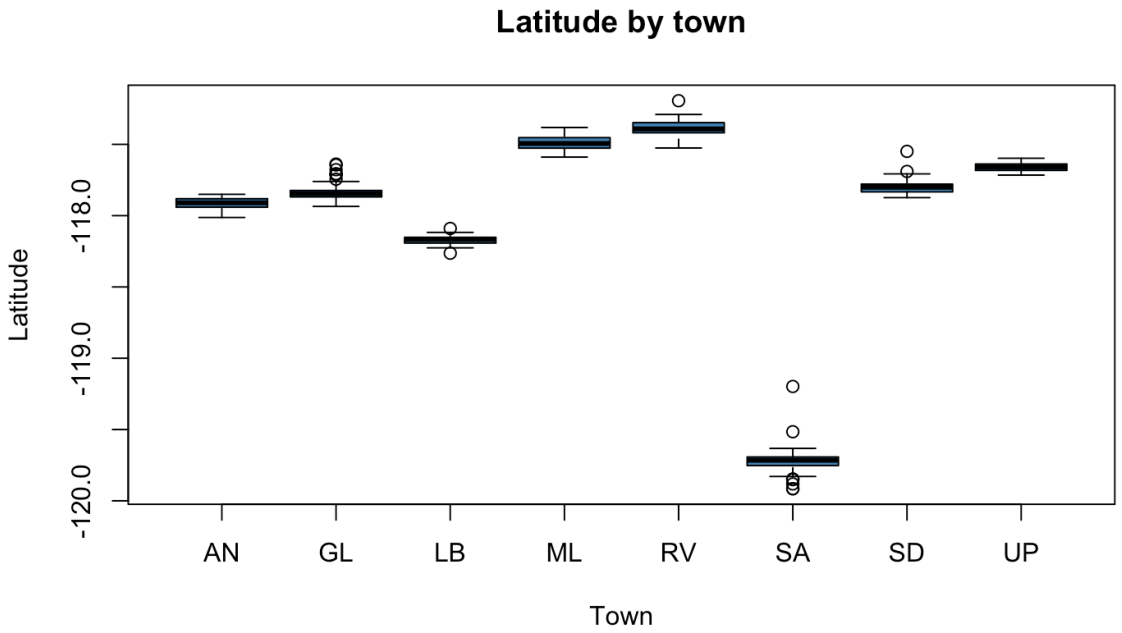
Graphical summary:Side by side Boxplots

```
boxplot(chs$latitude ~ chs$townabbr, main = 'Latitude by town',
        xlab = 'Town', ylab='Latitude', col = 'steelblue')
```



Quantitative vs. categorical

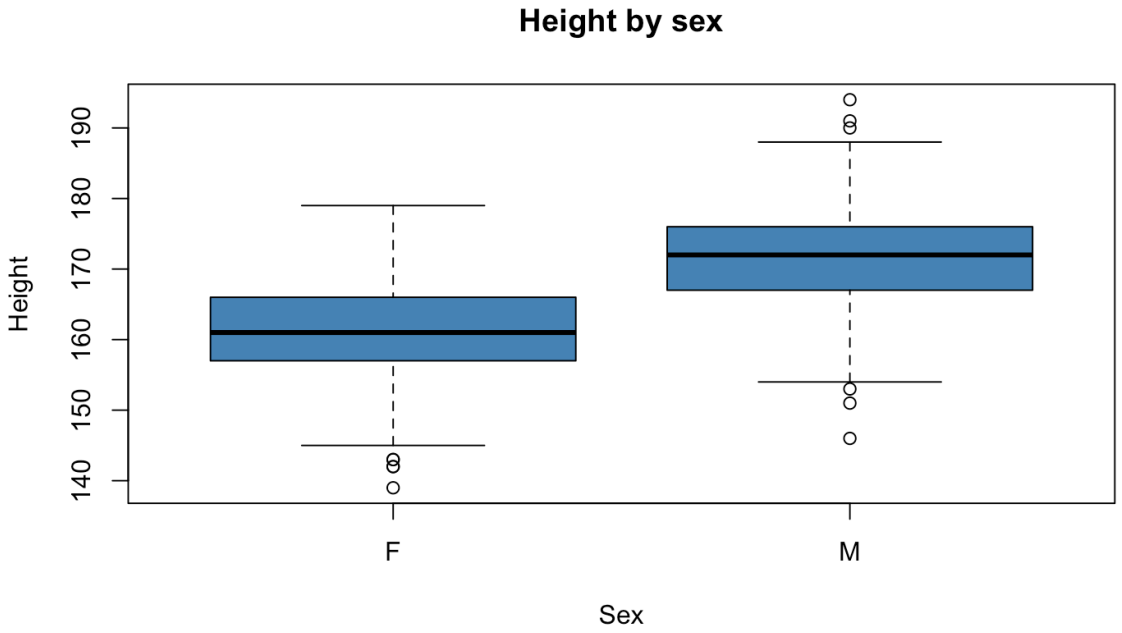
```
boxplot(chs$longitude ~ chs$townabbr, main = 'Latitude by town',
        xlab = 'Town', ylab='Latitude', col = 'steelblue')
```



Quantitative vs. categorical

Numerical summary:

```
chs$sex = factor(chs$male, levels = 0:1, labels = c('F', 'M'))
boxplot(chs$height ~ chs$sex, main = 'Height by sex',
        xlab = 'Sex', ylab='Height', col = 'steelblue')
```



Quantitative vs. categorical

Numerical summary: mean/sd or median/IQR **by** levels of the categorical variable

```
aggregate(chs$height, by = list(chs$sex), FUN=mean)
```

```
##   Group.1        x
## 1       F 161.2863
## 2       M 171.4741
```

```
aggregate(chs$height, by = list(chs$sex), FUN=sd)
```

```
##   Group.1        x
## 1       F 6.657984
## 2       M 7.192727
```

Many nice alternatives using R packages like dplyr for general data manipulation

Categorical vs. categorical

```
chs$educ <- factor(chs$educ, levels = 1:5, labels = c('< grade 12', 'grade 12','some post high-school', 'college',

table(chs$educ)
```

```
##
##            < grade 12              grade 12 some post high-school
##                   171                     153                     323
##               college        Some post-grad
##                   151                     138
```

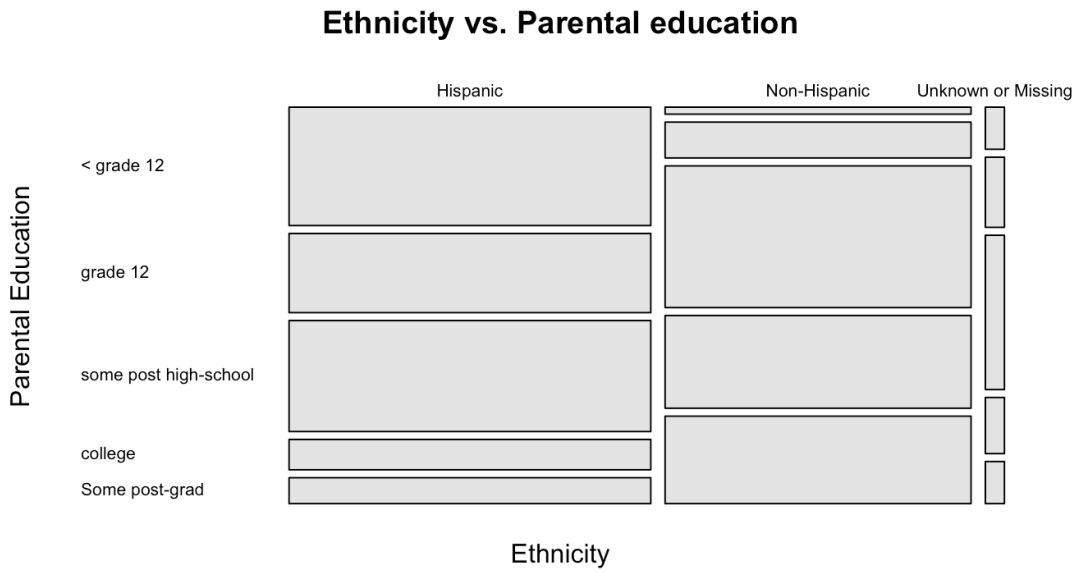
```
table(chs$hisp)
```

```
##
##           Hispanic       Non-Hispanic Unknown or Missing
##                522                 422                  56
```

Categorical vs. categorical

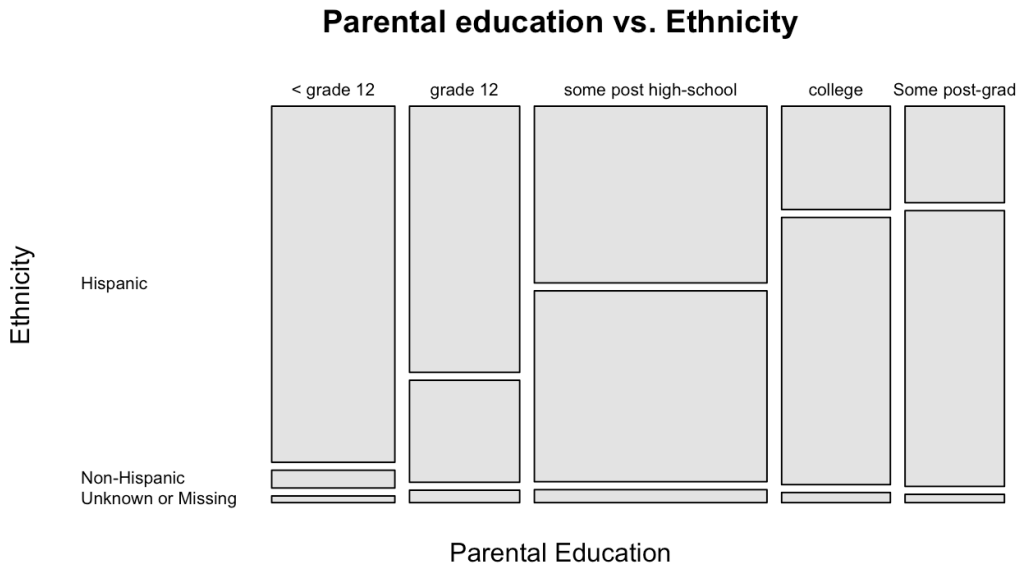
Graphical summary: Mosaic plots

```
mosaicplot(chs$hisp ~ chs$educ, ylab ='Parental Education',
           xlab = 'Ethnicity', col='gray90', las=1, main = 'Ethnicity vs. Parental education')
```



Categorical vs. categorical

```
mosaicplot(chs$educ ~ chs$hisp, xlab ='Parental Education',
           ylab = 'Ethnicity', col='gray90',las=1,  main = 'Parental education vs. Ethnicity')
```



Categorical vs. Categorical

Numerical summary: cross tabulation / contingency table

```
table(chs$educ, chs$hisp)
```

```
##
##                          Hispanic Non-Hispanic Unknown or Missing
##   < grade 12                  160            8                   3
##   grade 12                    107           41                   5
##   some post high-school       150          162                  11
##   college                      41          106                   4
##   Some post-grad               35          100                   3
```

- the base R `table` is not great for generating richly-featured crosstabs
- Many packages: crosstable, ctabs, xtable, ftable, function `CrossTable` in gmodels, and many more

Categorical vs. Categorical

Cross tabulation

```
library(crosstable)

c1 = crosstable(chs, c(educ), by = 'hisp', total="both", percent_pattern="{n} ({p_row})", percent_digits=1)
c1
```

```
## # A tibble: 7 × 7
##   .id   label variable        Hispanic `Non-Hispanic` `Unknown or Missing` Total
##   <chr> <chr> <chr>           <chr>    <chr>          <chr>                 <chr>
## 1 educ  educ  < grade 12      160 (93… 8 (4.7%)       3 (1.8%)              171 …
## 2 educ  educ  grade 12        107 (69… 41 (26.8%)     5 (3.3%)              153 …
## 3 educ  educ  some post high… 150 (46… 162 (50.2%)    11 (3.4%)             323 …
## 4 educ  educ  college         41 (27.… 106 (70.2%)    4 (2.6%)              151 …
## 5 educ  educ  Some post-grad  35 (25.… 100 (72.5%)    3 (2.2%)              138 …
## 6 educ  educ  NA              29       5              30                    64
## 7 educ  educ  Total           522 (52… 422 (42.2%)    56 (5.6%)             1000…
```

Categorical vs. Categorical

```
as_flextable(c1)
```

label	variable	hisp			Total
		Hispanic	Non-Hispanic	Unknown or Missing	
	< grade 12	160 (93.6%)	8 (4.7%)	3 (1.8%)	171 (18.3%)
	grade 12	107 (69.9%)	41 (26.8%)	5 (3.3%)	153 (16.3%)
	some post high-school	150 (46.4%)	162 (50.2%)	11 (3.4%)	323 (34.5%)
educ	college	41 (27.2%)	106 (70.2%)	4 (2.6%)	151 (16.1%)
	Some post-grad	35 (25.4%)	100 (72.5%)	3 (2.2%)	138 (14.7%)
	NA	29	5	30	64
	Total	522 (52.2%)	422 (42.2%)	56 (5.6%)	1000 (100.0%)

Categorical vs. Categorical

Cross tabulation

```
as_flextable(crosstable(chs, c(hisp), by = 'educ', total="both", percent_pattern="{n} ({p_row})", percent_digits=1)
```

label	variable	educ						Total
		< grade 12	grade 12	some post high-school	college	Some post-grad	NA	
hisp	Hispanic	160 (32.5%)	107 (21.7%)	150 (30.4%)	41 (8.3%)	35 (7.1%)	29	522 (52.2%)
	Non-Hispanic	8 (1.9%)	41 (9.8%)	162 (38.8%)	106 (25.4%)	100 (24.0%)	5	422 (42.2%)
	Unknown or Missing	3 (11.5%)	5 (19.2%)	11 (42.3%)	4 (15.4%)	3 (11.5%)	30	56 (5.6%)
	Total	171 (18.3%)	153 (16.3%)	323 (34.5%)	151 (16.1%)	138 (14.7%)	64	1000 (100.0%)