Categorical Descriptive Statistics

Grinnell College

February 2, 2025

Review

Suppose I have:

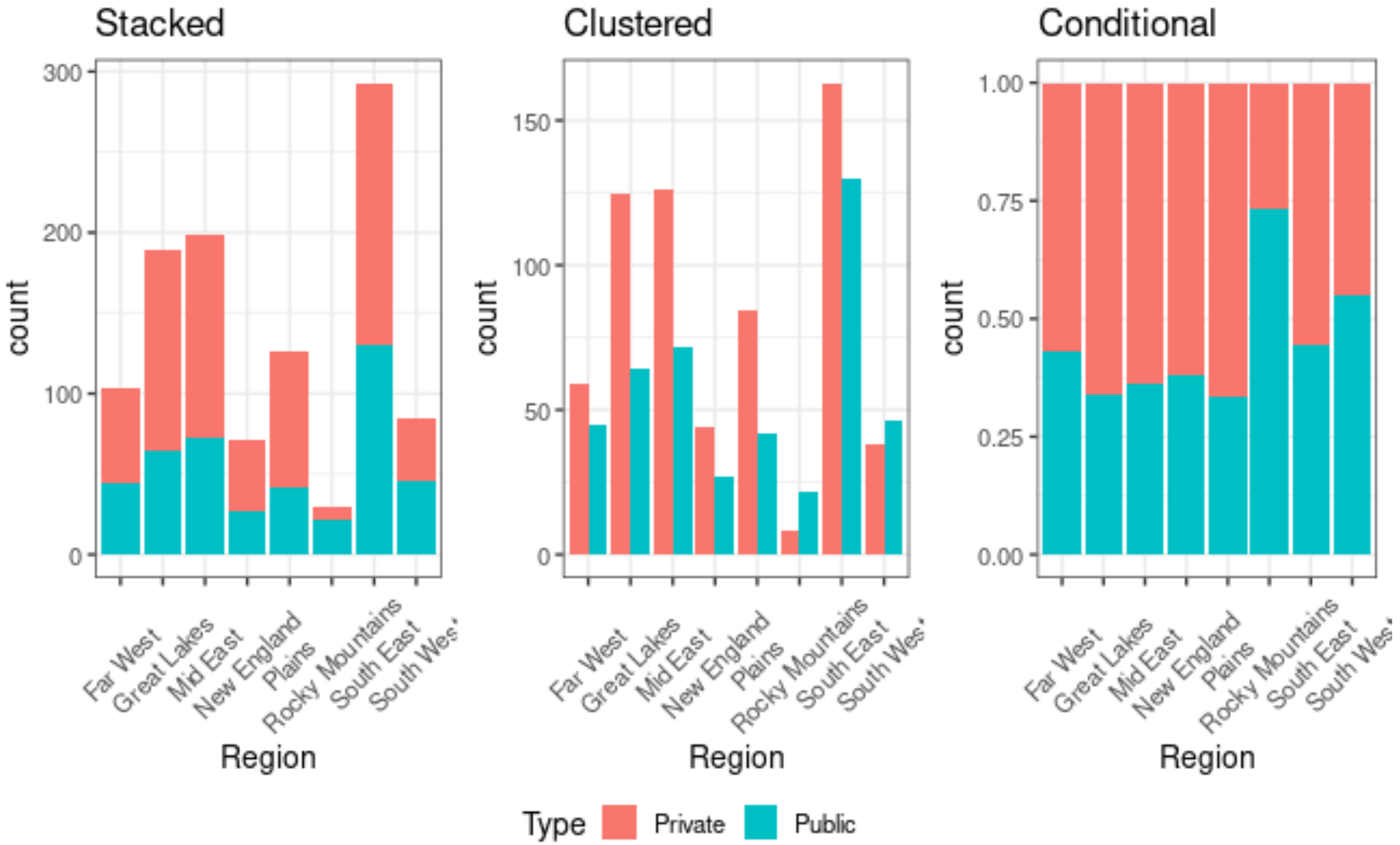
- 750 observations
- Median value of 27
- IQR of 9

How many observations would fall between the 35^{th} and 65^{th} percentiles?

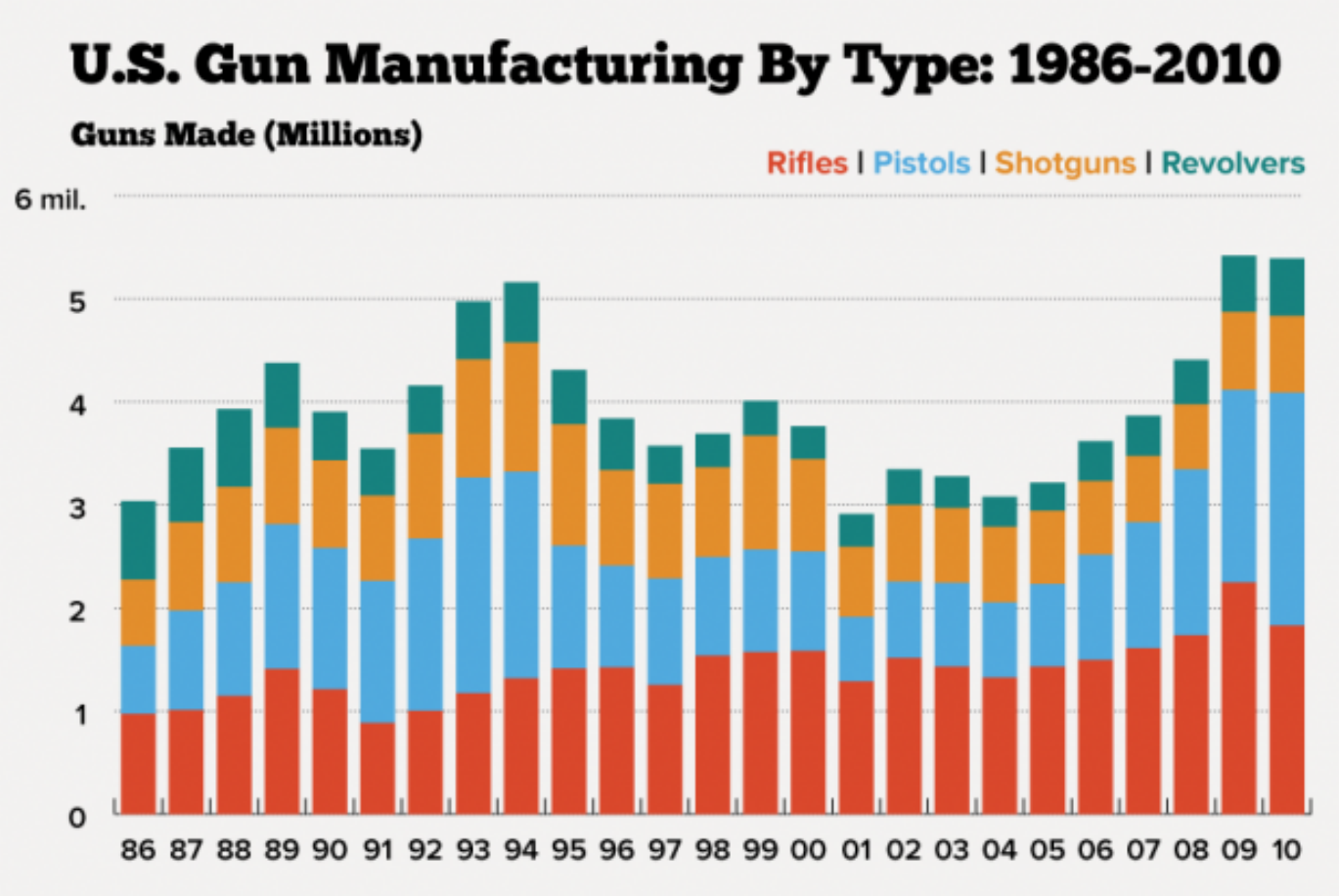
Today

- Summary of categorical variables
  - Tables
  - Bar Charts
- Types of Tables
  - Frequency
  - Proportions
  - Conditional
- Rough measures of association

Bar Charts

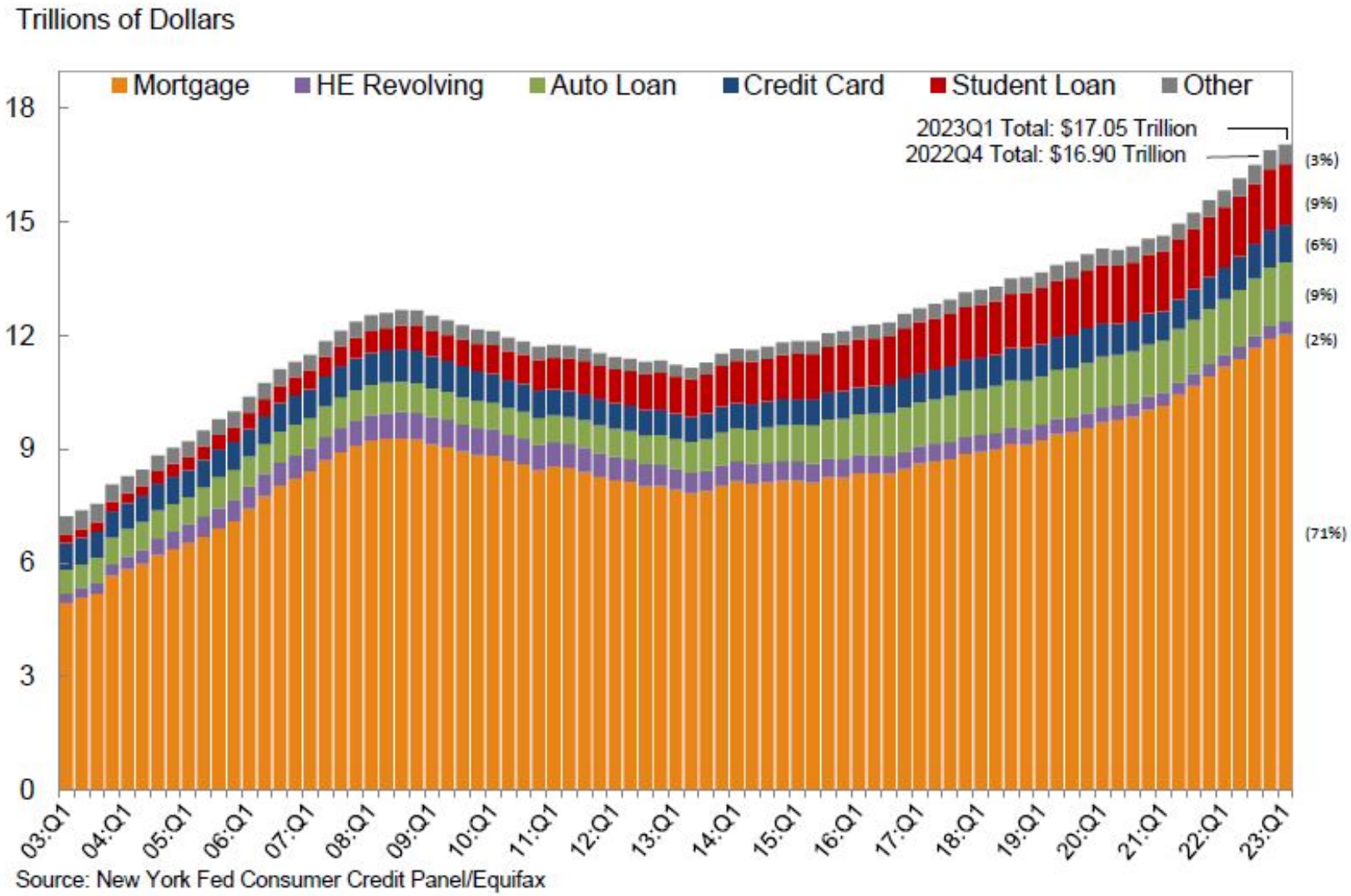

Stacked
Clustered
Conditional
count
300
200
100
0
150
50
1.00
0.75
0.50
0.25
0.00
Far West
Great Lakes
Mid East
New England
Plains
Rocky Mountains
South East
South West
Region
Type
Private
Public

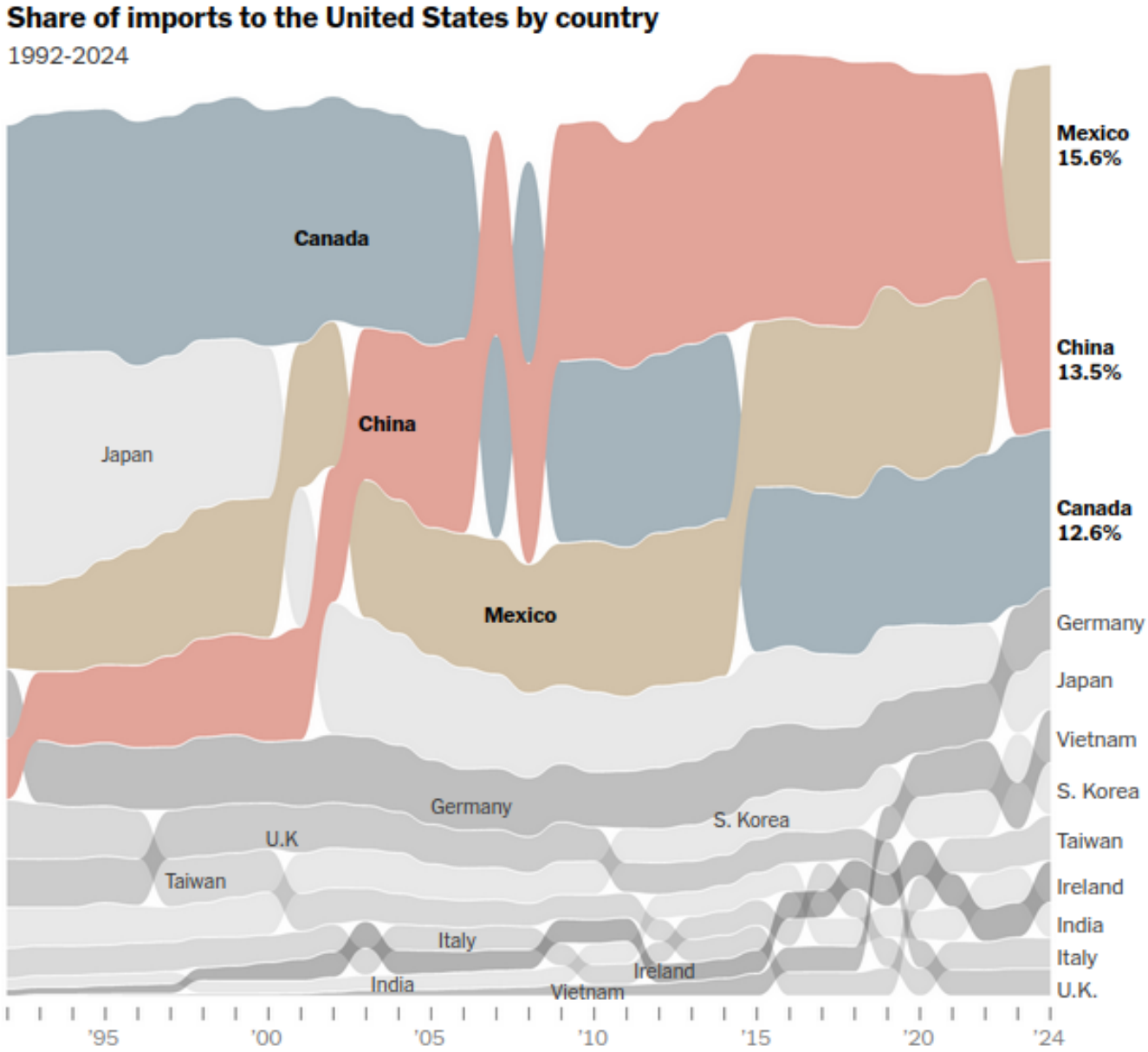
Stacked Bar Example



https://stackoverflow.com/questions/64267754/plotting-a-time-series-stacked-bar-chart

Total Debt Balance and its Composition




Share of imports to the United States by country
1992-2024
Canada
Japan
China
Mexico
Germany
U.K
Taiwan
Italy
India
Vietnam
S. Korea
Ireland
Mexico
15.6%
China
13.5%
Canada
12.6%
Germany
Japan
Vietnam
S. Korea
Taiwan
Ireland
India
Italy
U.K.
'95
'00
'05
'10
'15
'20
'24
Notes: Countries with at least a 2 percent share in 2024, through November, are shown, accounting for about three-quarters of imports. • Source: Census Bureau • By The New York Times

Descriptive Statistics – Categorical Variables

Univariate categorical variables are often presented in *tables*

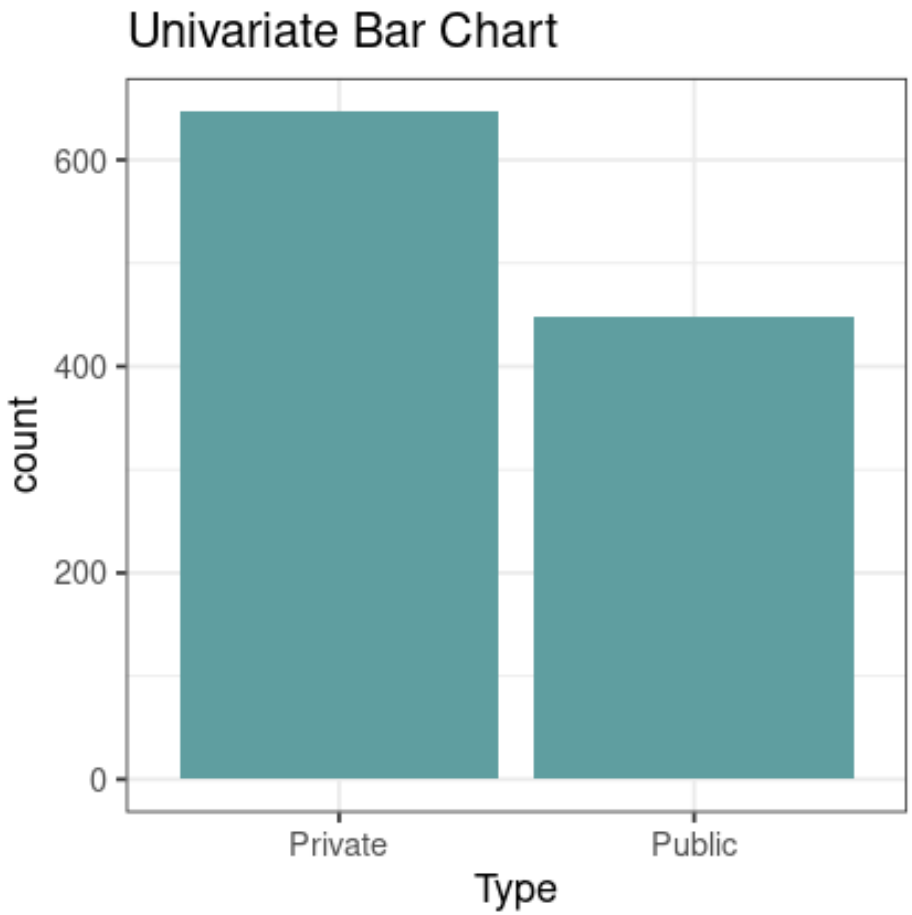
- **Frequencies:** counts how many of each case belongs to a particular category
- **Proportions:** fractions based upon frequencies, also called *relative frequencies*. Proportions will *always* add up to 1

Frequency table:

Type	Frequency
Private	647
Public	448

Table of proportions:

Type	Proportion
Private	0.591
Public	0.409



Descriptive Statistics – Categorical Variables

Univariate categorical variables are often presented in *tables*

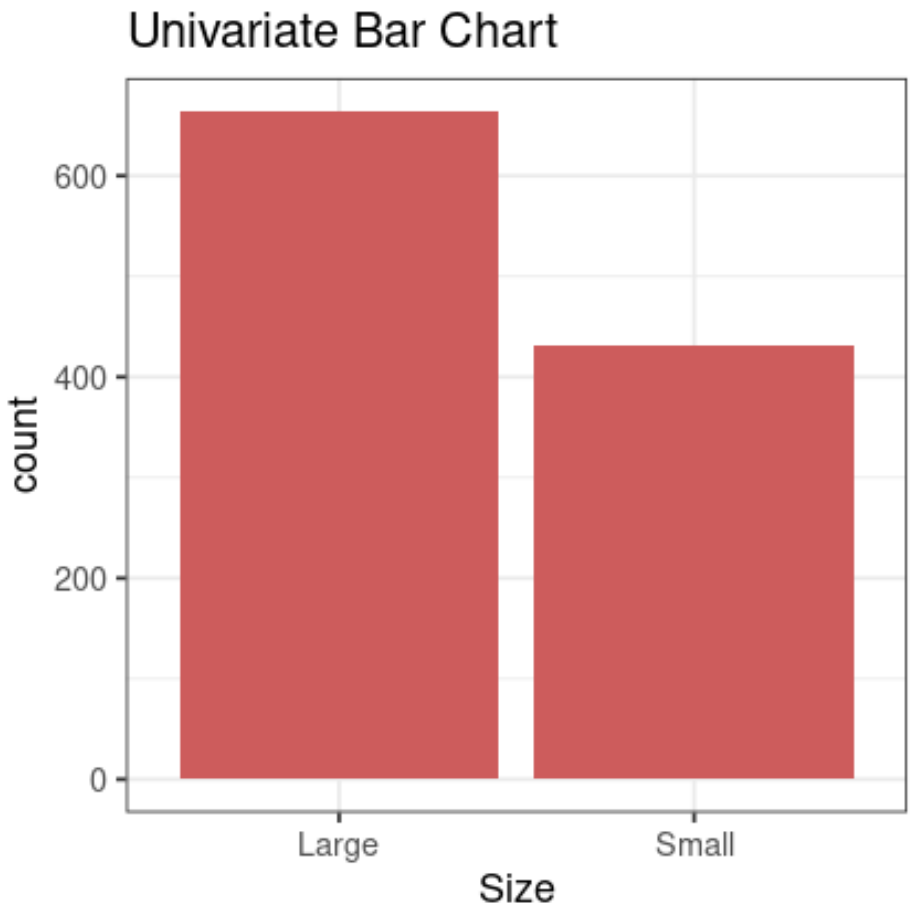
- **Frequencies:** counts how many of each case belongs to a particular category
- **Proportions:** fractions based upon frequencies, also called *relative frequencies*. Proportions will *always* add up to 1

Frequency table:

Size	Frequency
Large	664
Small	431

Table of proportions:

Size	Proportion
Large	0.606
Small	0.394



Bivariate Bar Charts

When considering two categorical variables, we typically cross-classify an observation according to its variable's values

Just as we did when looking at graphical summaries, we tend to designate variables as being either *explanatory* or *response* variables

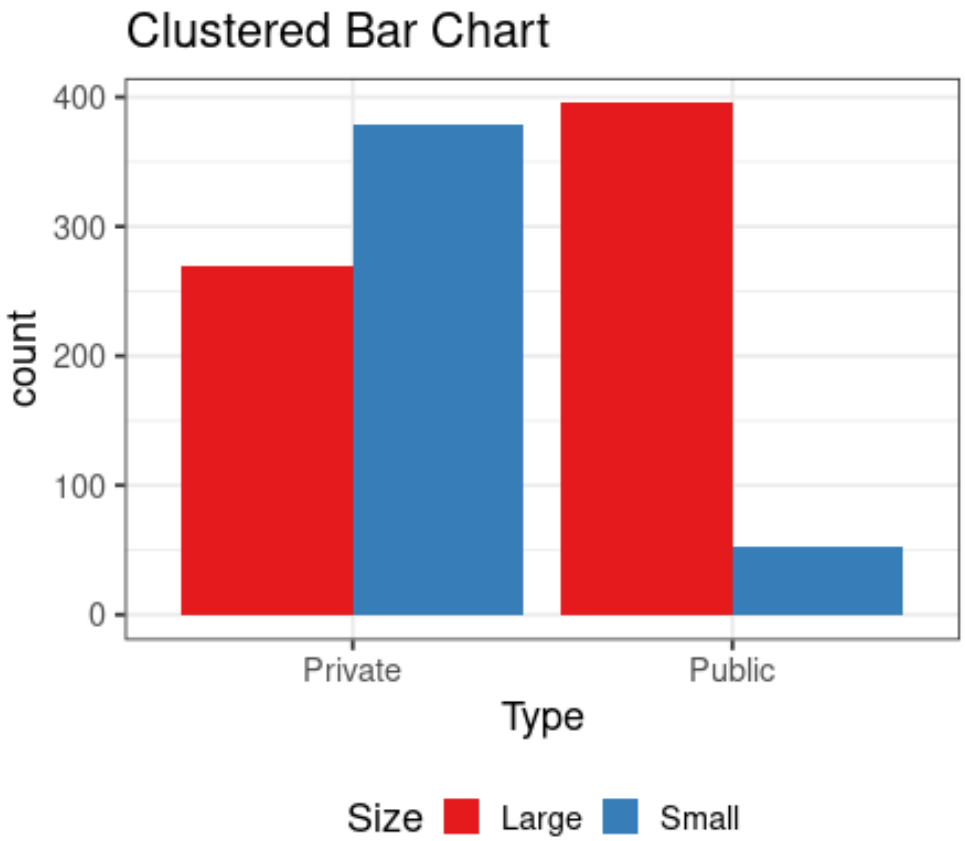
Again, this is **not** causal

Descriptive Statistics – Categorical Variables

The **joint distribution** shows us the collection and frequency of values that two variables take together

Two-way frequency table:

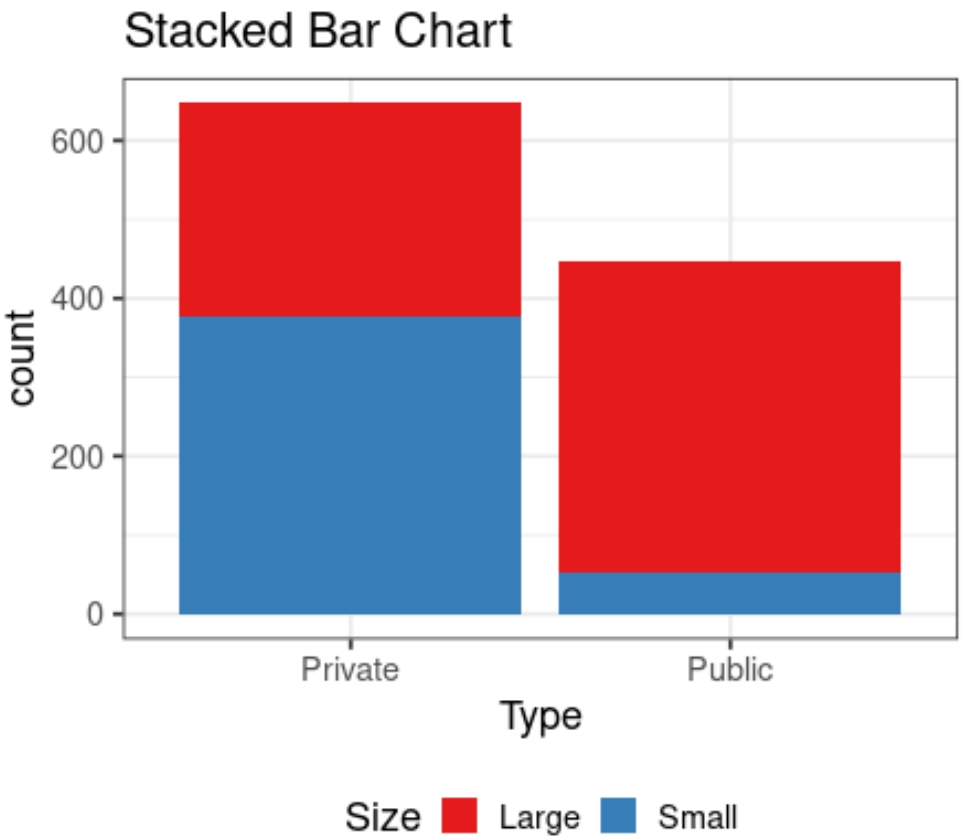
	Small	Large
Private	378	269
Public	53	395



Descriptive Statistics – Categorical Variables

Often these tables include margin sums as well, giving us **marginal distributions** of variables

	Small	Large	Sum
Private	378	269	647
Public	53	395	448
Sum	431	664	1095



Descriptive Statistics – Categorical Variables

The proportions of a joint distribution tells us the makeup of each combination, relative to the whole

	Small	Large
Private	$\frac{378}{1095}$	$\frac{269}{1095}$
Public	$\frac{53}{1095}$	$\frac{395}{1095}$

	Small	Large
Private	0.3452	0.2457
Public	0.0484	0.3607

"36% of all schools are large public schools"

Conditional Statistics

A **conditional statistic** is a statistic derived from one or more variables for all observations sharing a value of another variable

- "What is the relationship between admission rate and median ACT *given* that the school is private"
- "What is the predicted weight of an individual *given* that they are 6 ft tall?"
- "What is the proportion of public schools *given* that we are looking at the Plains region?"

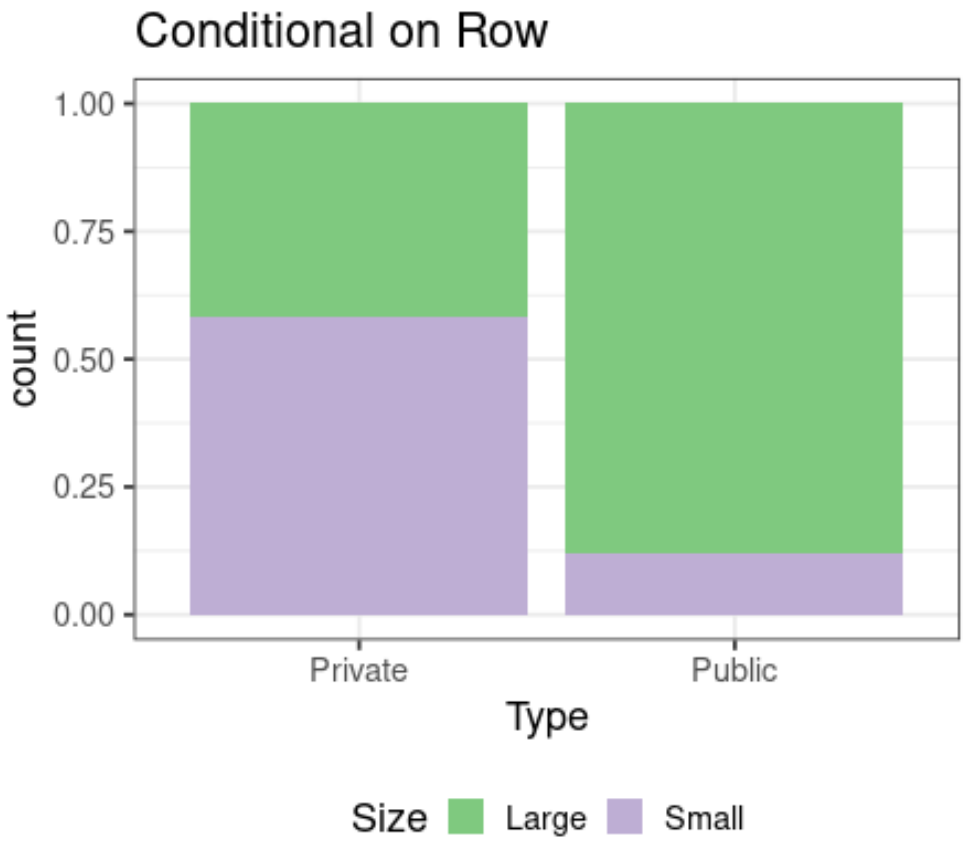
Note that we typically condition on the *explanatory* variable

Descriptive Statistics – Row Proportions

"88% of public schools are considered large"

"Given that a school is a public school, 88% of them are considered large"

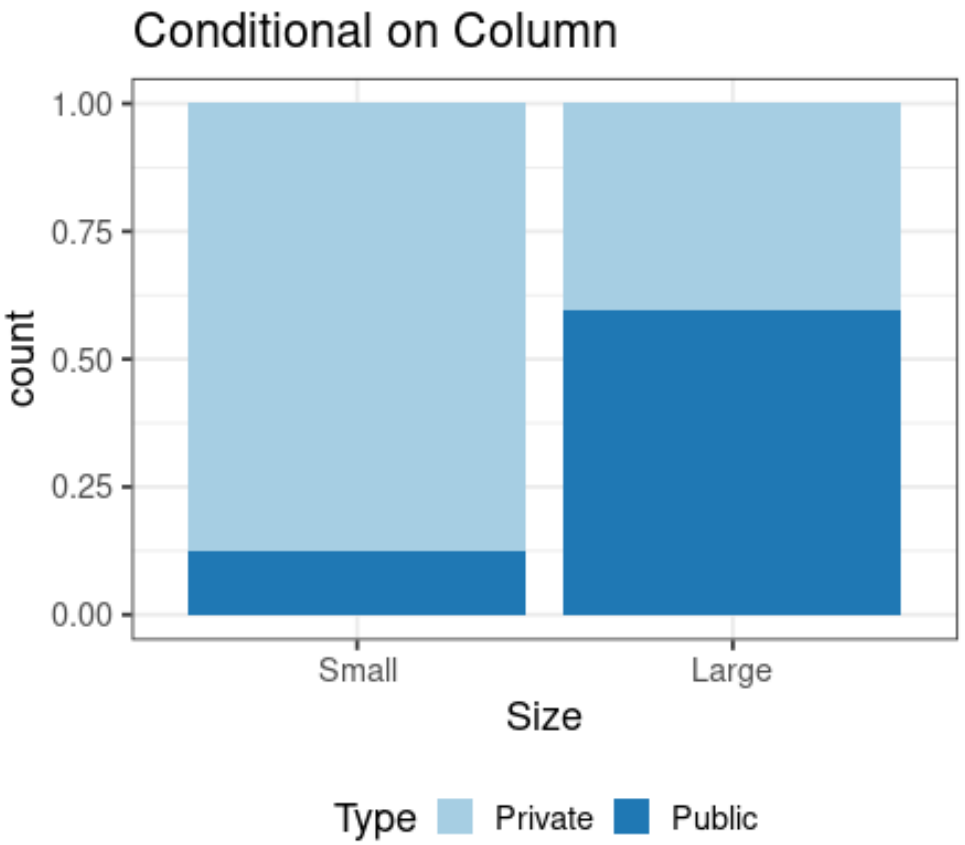
	Small	Large
Private	0.5842	0.4158
Public	0.1183	0.8817



Descriptive Statistics – Column Proportions

"12% of small colleges are public"

	Small	Large
Private	0.8770	0.4051
Public	0.1230	0.5949



Example

The two-way table below describes the survival of crew members and first class passengers aboard the Titanic

	Survived	Died
Crew	212	673
First Class	203	122

1. Given that an individual survived, is it more likely that they were a crew member or a passenger in first class?
2. Given that an individual was a crew member, is it more likely that they survived or died?
3. Which group was more likely to survive the shipwreck?

Summary

- Types of charts
  - Stacked
  - Clustered
  - Conditional
- Types of Tables
  - One and two-way tables
  - Frequency and proportions
  - Which associated with which plots?
- Association for categorical variables