Welcome to DATA 151

I'm so glad you're here!

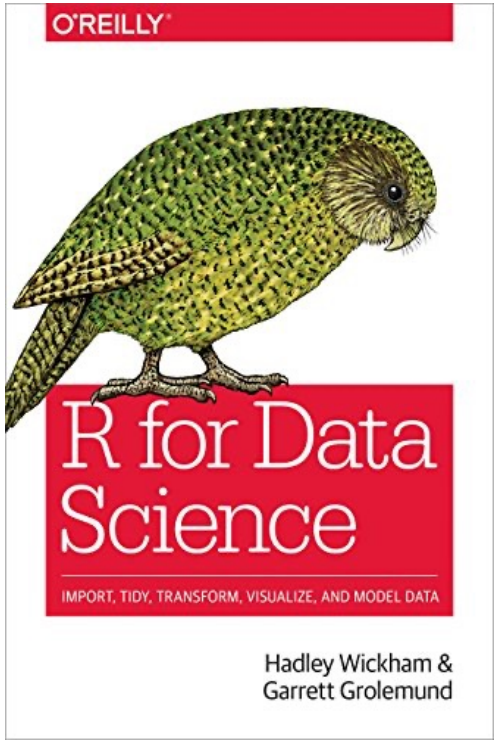
DATA 151: CLASS 7B
INTRODUCTION TO DATA SCIENCE (WITH R)

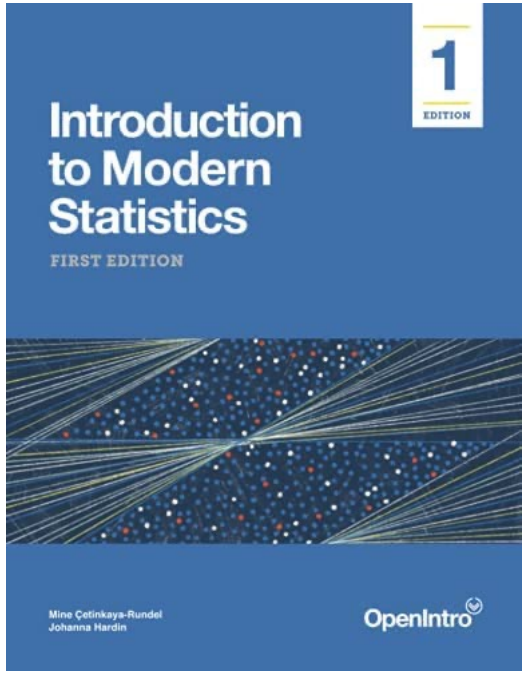
CATEGORICAL DATA ANALYSIS: TABLES AND BARS

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING





Introduction to Data Science:

- Tuesday:
  - R for Data Science
  - Ch 7: Exploratory Data Analysis
- Thursday:
  - Introduction to Modern Statistics
  - Ch 4: Exploring Categorical Data

HOMEWORK REMINDER

Due this/next week: (EXTENSION DUE 10/18)

- *HW #6: DC Introduction to Data Visualization in ggplot2*
  - ***No submission on WISE necessary, do on DataCamp***
- *Project Milestone #3: EDA Step 1*
  - Ask questions and form hypotheses

HOMEWORK REMINDER

Due next week: (DUE 10/20)

- *HW #7: DC Exploratory Data Analysis with Categorical Data*
  - *Just one chapter*
  - ***No submission on WISE necessary, do on DataCamp***
- *Project Milestone #4: EDA Step 2*
  - Create Tables and Bar Graphs

EXTRA CREDIT OPPORTUNITY

Data & Computing Tea

On the Research Experience for Undergraduates,

Thursday the 20th, 11:30 AM, Ford 201

Meelad Doroodchi, a major in the department who completed a REU (Research Experience for Undergraduates) this summer. Meelad will share some thoughts on the program and his experience, and then we will have time for open discussion.

Pizza will be provided.

EXTRA CREDIT OPPORTUNITY

If you go to the presentation and do a 1-page write up about your take-aways and how this work relates to how we are learning data science in this class, I will give you **4 extra credit points** toward your

Midterm #1 grade!

BOARD OF TRUSTEES

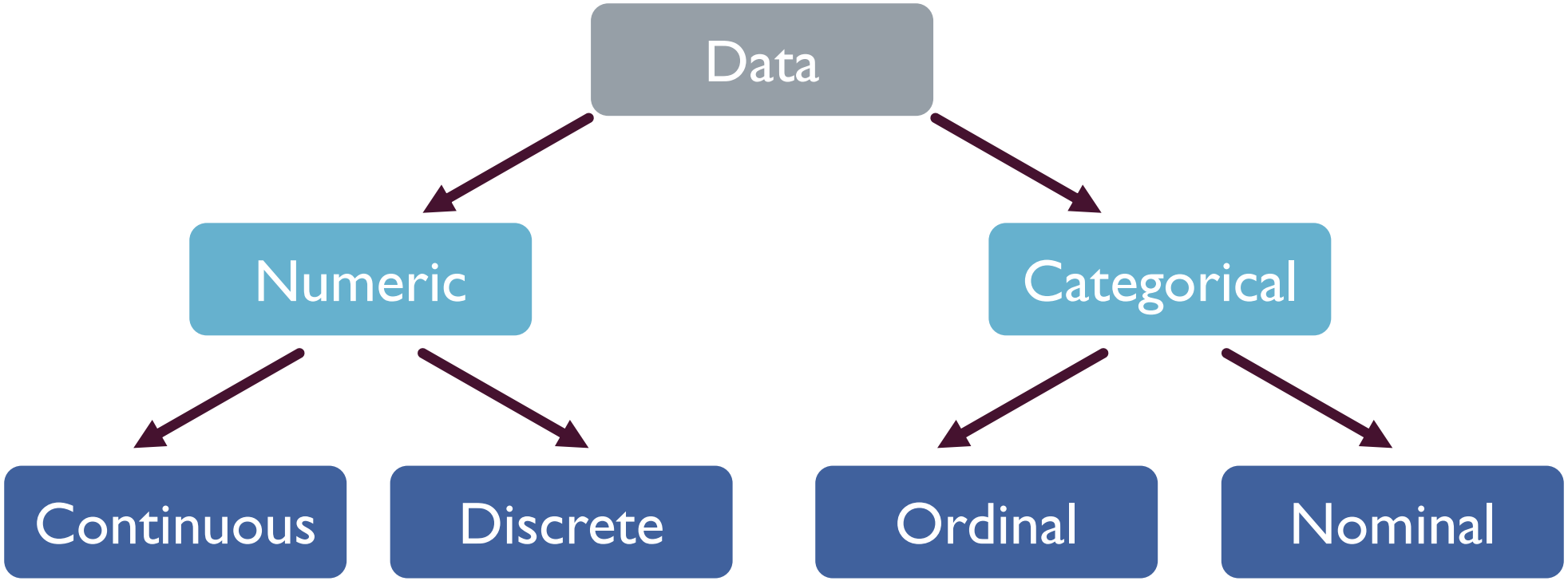
Next week on Thursday Oct 20

The Willamette Board of Trustees will be meeting in FORD 102 (our classroom space)

For this one day only we have been asked to relocate our class to Eaton 209

LET'S GET STARTED!

TYPES OF VARIABLES



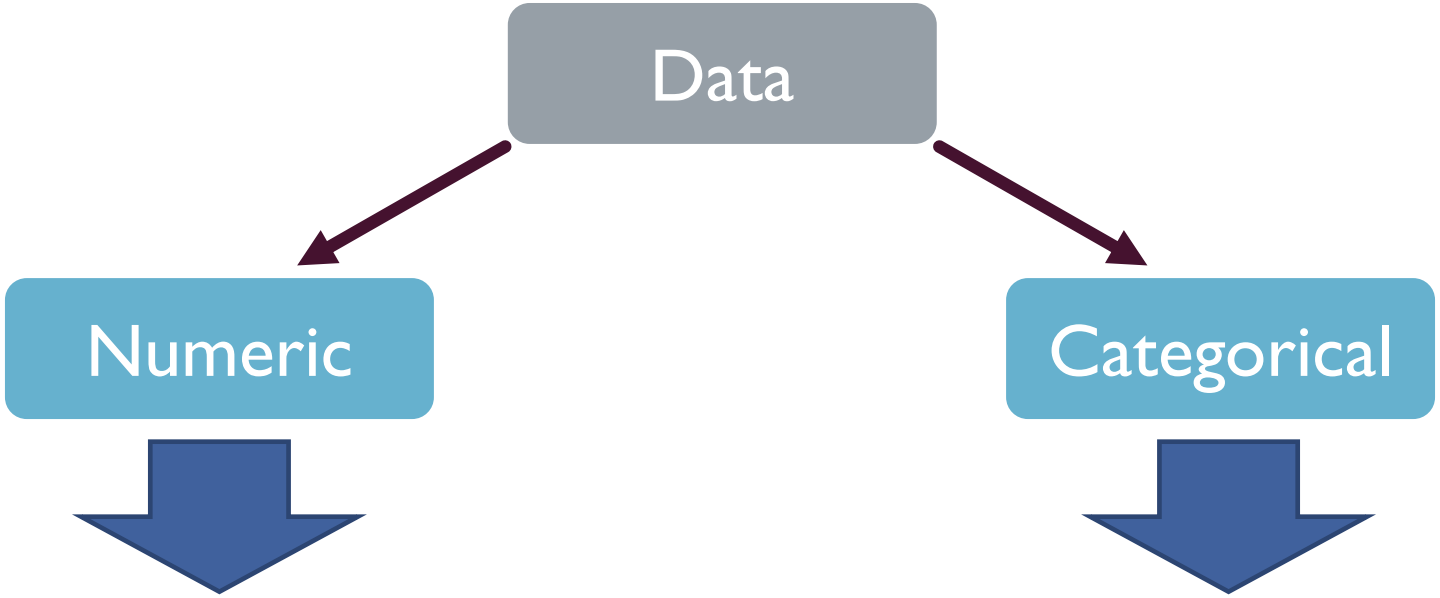
EXPLORING YOUR DATA

DISTRIBUTIONS

- **What is a distribution?**
  - The distribution of a variable tells us what values the variable takes and how often the variable assume those values.

- **How can you see a distribution?**
  - Graphical tools/ Data Visualization
  - Exploratory Data Analysis

UNIVARIATE GRAPHICAL TOOLS



- Stem-and-Leaf plots
- Histograms
- Density (approximation) plot
- Boxplot / Box-and-whisker plot

- Bar graphs
- Pie charts

WHAT TO DO WHEN YOU HAVE CATEGORICAL DATA

The distribution of a categorical variable lists the categories and provides the count or percent of individuals who fall into each category.

- **Tables**
  - Shows the counts/proportion of observations that fall within a category or categories
  - Ex: Joint, marginal, conditional distributions
- **Bar Graphs**
  - Represent each category as a bar whose heights show the category counts or percent.
- **Pie Charts**
  - Show the distribution of a categorical variable as a "pie" whose slices are sized by the counts or percent for the categories relative to the whole
  - ***NOTE: Statisticians avoid using pie charts because differences in angles are difficult for the view to perceive. We want to elucidate the story that the data is telling, not muddle it.***

EXAMPLE 1: TITANIC

MOTIVATING EXAMPLE #1: THE TITANIC

- At 11:40 pm on April 14, 1912 the *Titanic* hit an iceberg on its maiden voyage from Southampton to New York City.
- Our data represent 2201 passengers
  - This data is very popular and is included in an R package

PassengerId	Pclass	Name	Sex	Age
892	3	Kelly, Mr. James	male	34.5
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0
894	2	Myles, Mr. Thomas Francis	male	62.0
895	3	Wirz, Mr. Albert	male	27.0
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0
897	3	Svensson, Mr. Johan Cervin	male	14.0

MOTIVATING EXAMPLE: THE TITANIC

Knowledge check:

Are these data "tidy"?

Recall: The three principles of tidy data from class

PassengerId	Pclass	Name	Sex	Age
892	3	Kelly, Mr. James	male	34.5
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0
894	2	Myles, Mr. Thomas Francis	male	62.0
895	3	Wirz, Mr. Albert	male	27.0
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0
897	3	Svensson, Mr. Johan Cervin	male	14.0

MOTIVATING EXAMPLE: THE TITANIC

Knowledge check:

1) What are the observations?

2) What the variables? What types of variables are they?

PassengerId	Pclass	Name	Sex	Age
892	3	Kelly, Mr. James	male	34.5
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0
894	2	Myles, Mr. Thomas Francis	male	62.0
895	3	Wirz, Mr. Albert	male	27.0
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0
897	3	Svensson, Mr. Johan Cervin	male	14.0

Titanic Passenger Survival Data Set

Documentation for package ‘titanic’ version 0.1.0

- DESCRIPTION file.

Help Pages

titanic-package	titanic: Titanic Passenger Survival Data Set
titanic	titanic: Titanic Passenger Survival Data Set
titanic_gender_class_model	Titanic gender class model data.
titanic_gender_model	Titanic gender model data.
titanic_test	Titanic test data.
titanic_train	Titanic train data.

DIFFERENT FORMATS OF DATA

In this package there are a couple different forms of data:

- Cross-tabulated data
- Individual level raw data

We will work on examples of each so that you can get practice with different types of data structures.

MOTIVATING EXAMPLE: THE TITANIC

Since 2201 passengers are represented in these data, its hard to comprehend it all at the same time and discern important patters, so lets make **tables and graphics!**

FREQUENCY TABLES

Frequency Table: Records the totals and the category names

In this case, we might be interested in how many people were are in each class.

	Class	n
1	1st	325
2	2nd	285
3	3rd	706
4	Crew	885

FREQUENCY TABLES

Create a one-way frequency table for the distribution of class.

```
## frequency table for class
## can you think of how you would do this in dplyr?
titanClass<-Titanic%>%
  group_by(Class)%>%
  summarise(n_class=sum(Freq))

titanClass
```

```
## # A tibble: 4 × 2
##   Class n_class
##   <fct>   <dbl>
## 1 1st       325
## 2 2nd       285
## 3 3rd       706
## 4 Crew      885
```

FREQUENCY TABLES

Relative Frequency Table: Displays percentages in each category instead of counts

Note: This must add up to 100%

	Class	n	freq
1	1st	325	0.1476602
2	2nd	285	0.1294866
3	3rd	706	0.3207633
4	Crew	885	0.4020900

FREQUENCY TABLES

We might also want to display proportions.

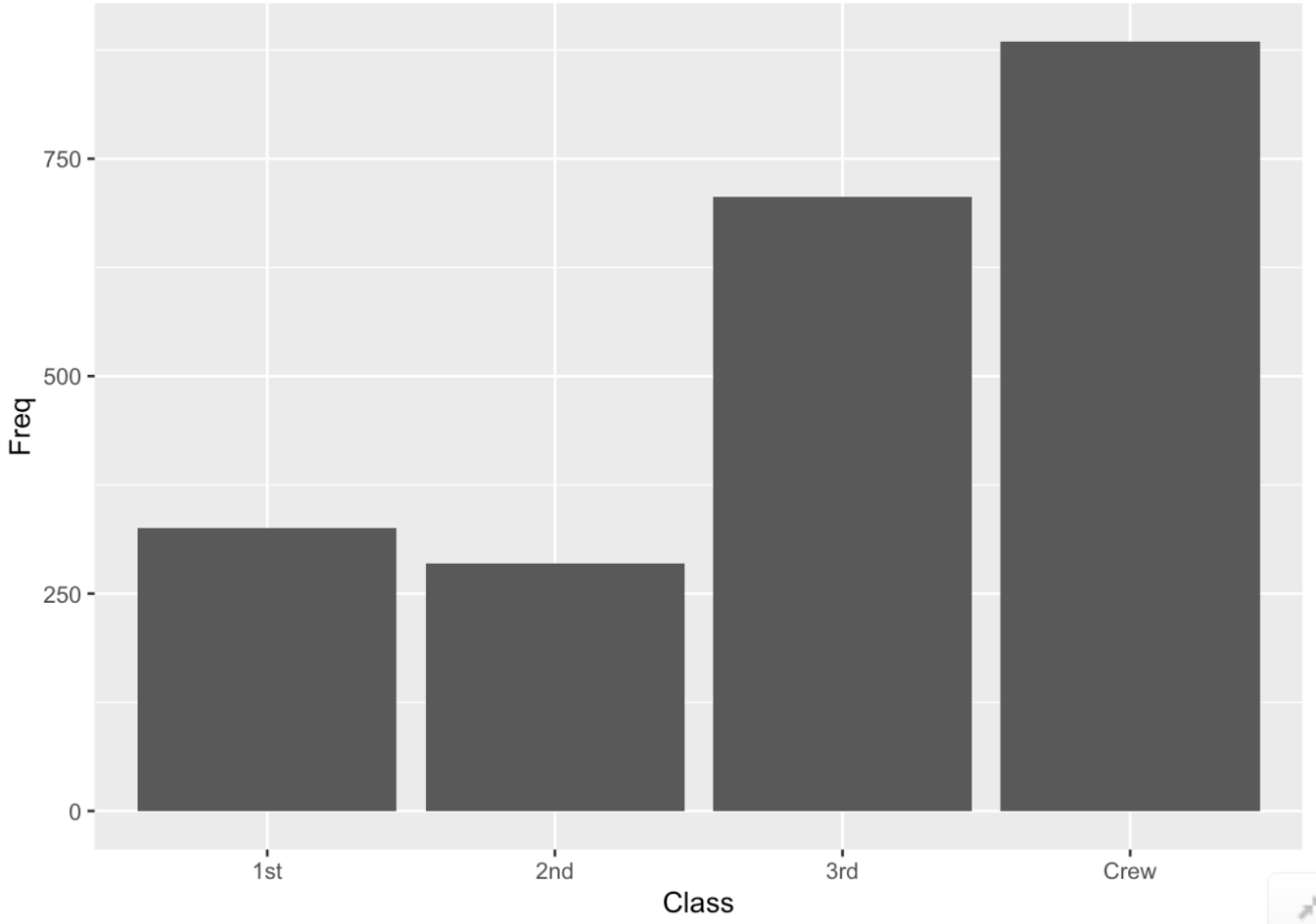
```
## What could we do if we want proportions?

titanClassProp<-titanClass%>%
  mutate(prop=n_class/sum(n_class))

titanClassProp
```

```
## # A tibble: 4 × 3
##   Class n_class  prop
##   <fct>   <dbl> <dbl>
## 1 1st       325 0.148
## 2 2nd       285 0.129
## 3 3rd       706 0.321
## 4 Crew      885 0.402
```

BAR CHARTS

Bar Charts:
Illustrates counts or percentage for each category

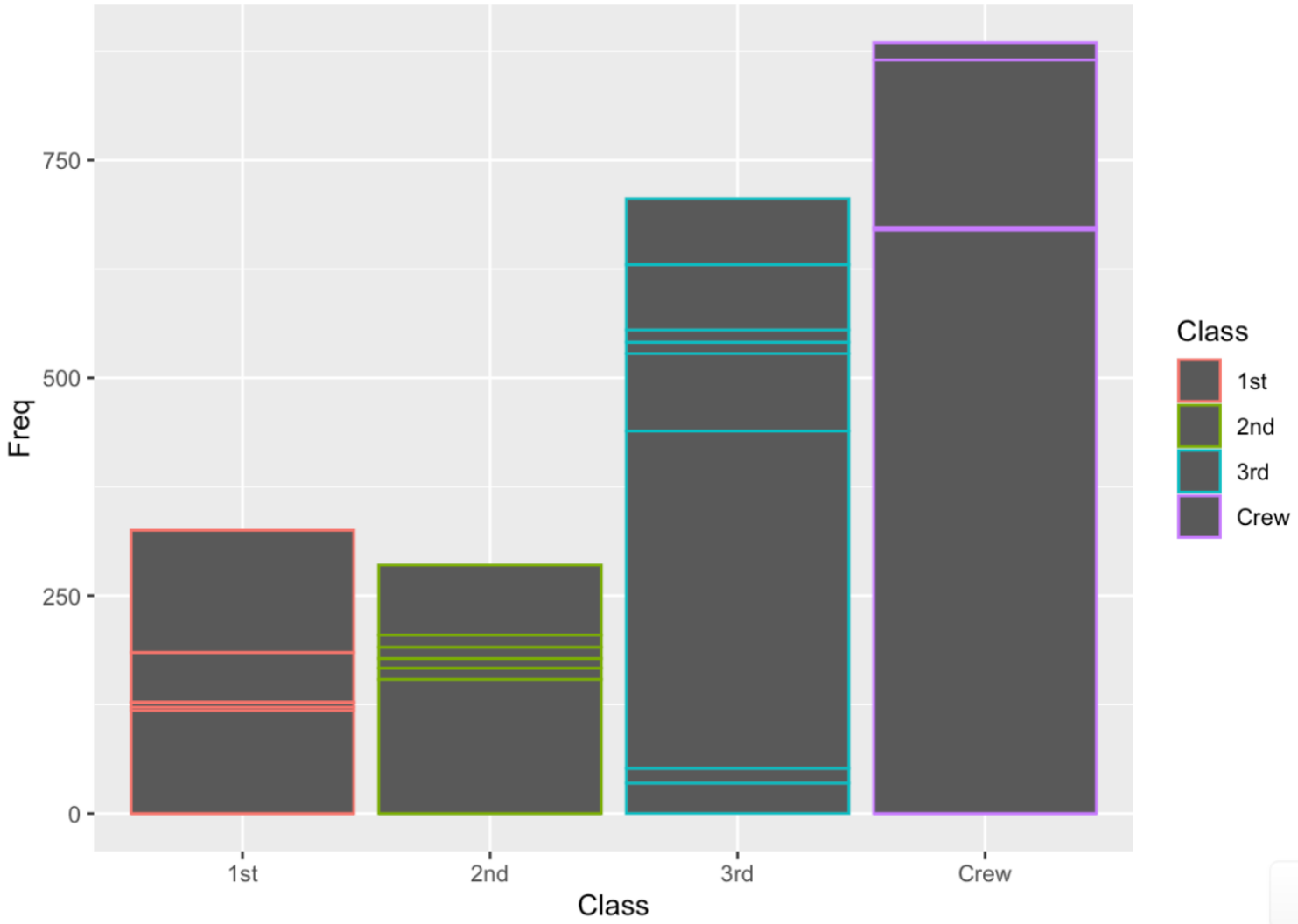
```
## bar graphs
ggplot(Titanic, aes(x=Class, y=Freq))+
  geom_bar(stat = "identity")
```



What is going on in this graph?

ADD COLOR!

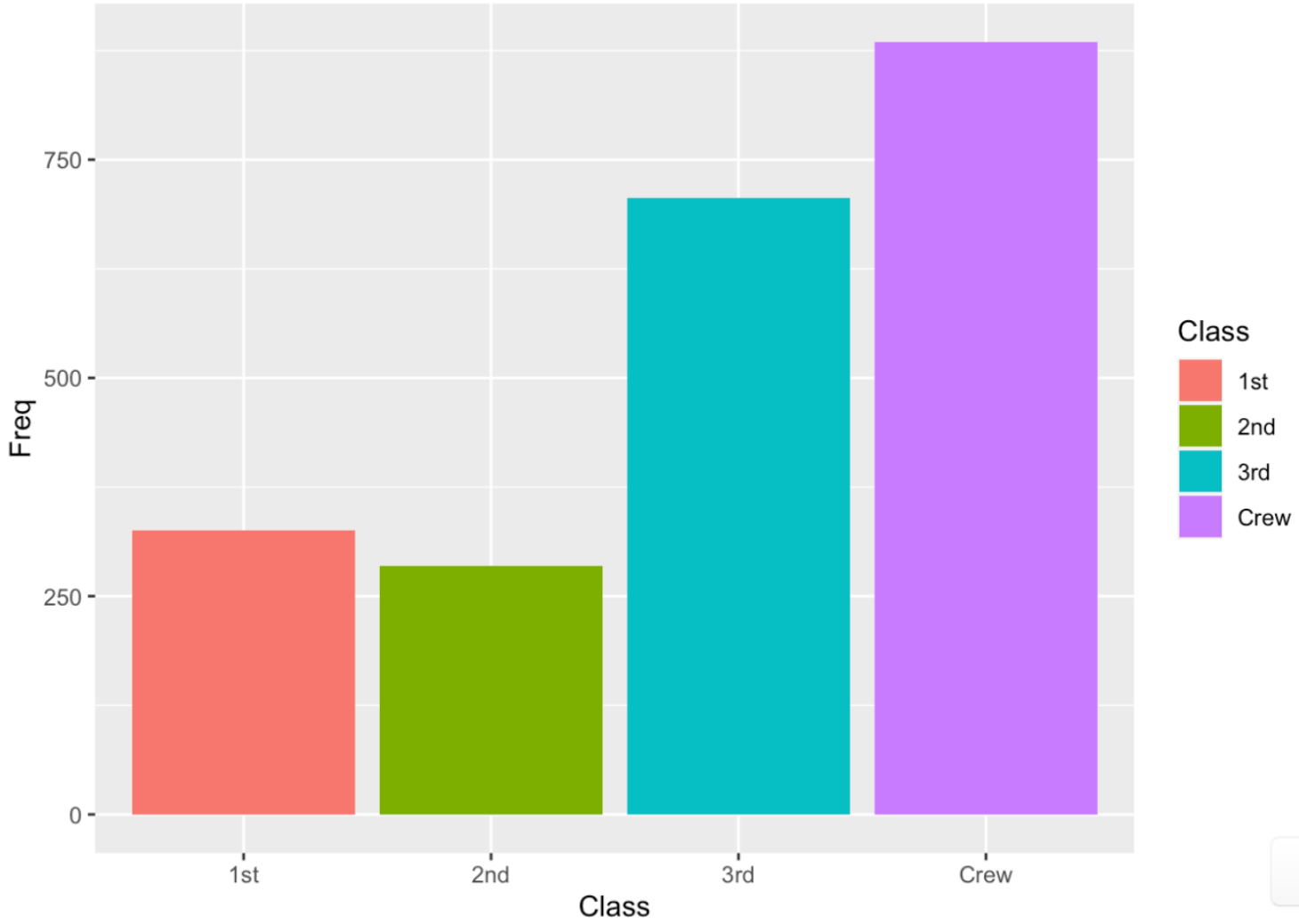
```
## add color
ggplot(Titanic, aes(x=Class, y=Freq, color=Class))+
  geom_bar(stat = "identity")
```



ADD COLOR!

C. Fill

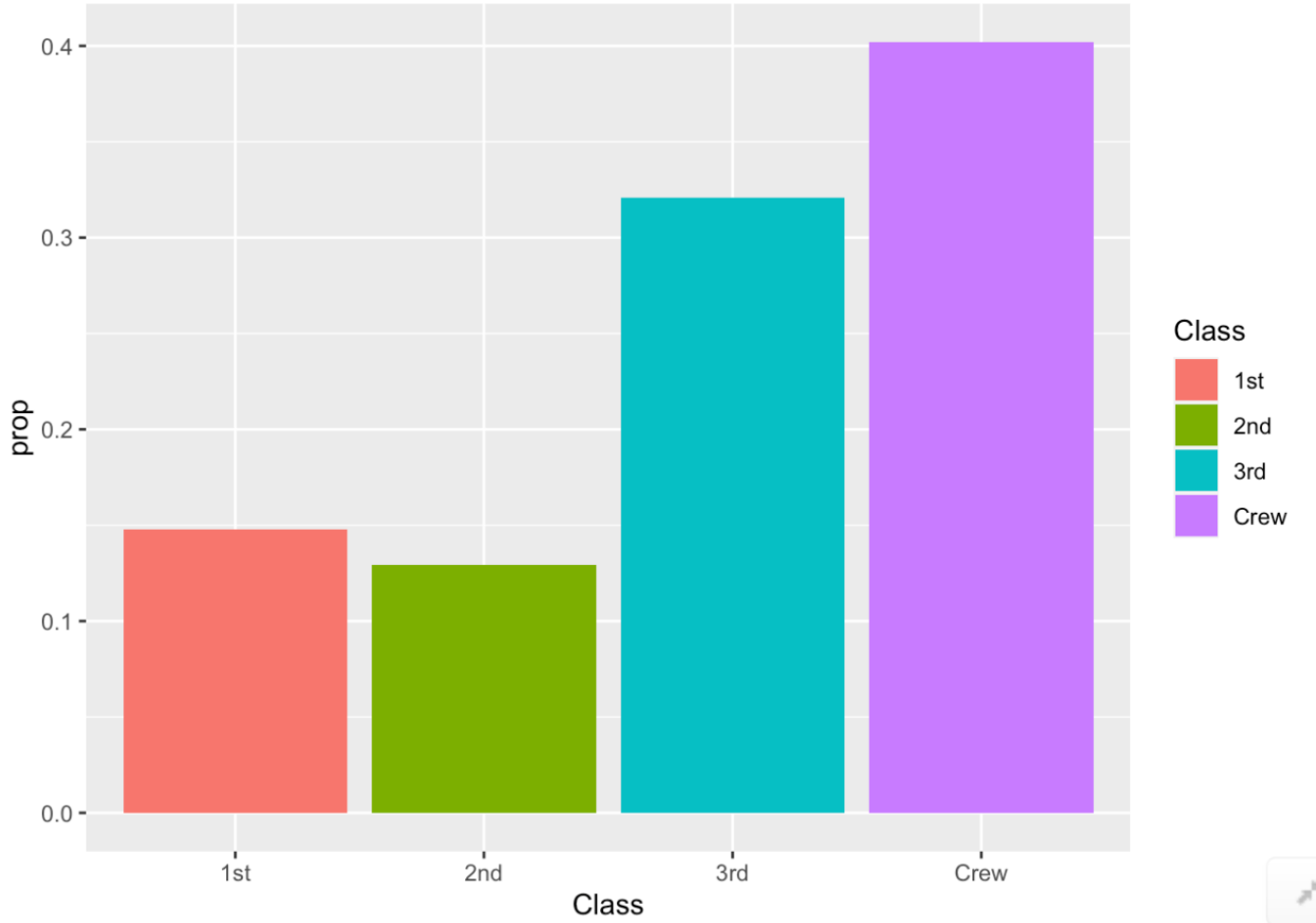
```
## oops! let's use fill
ggplot(Titanic, aes(x=Class, y=Freq, fill=Class))+
  geom_bar(stat = "identity")
```



Y-AXIS

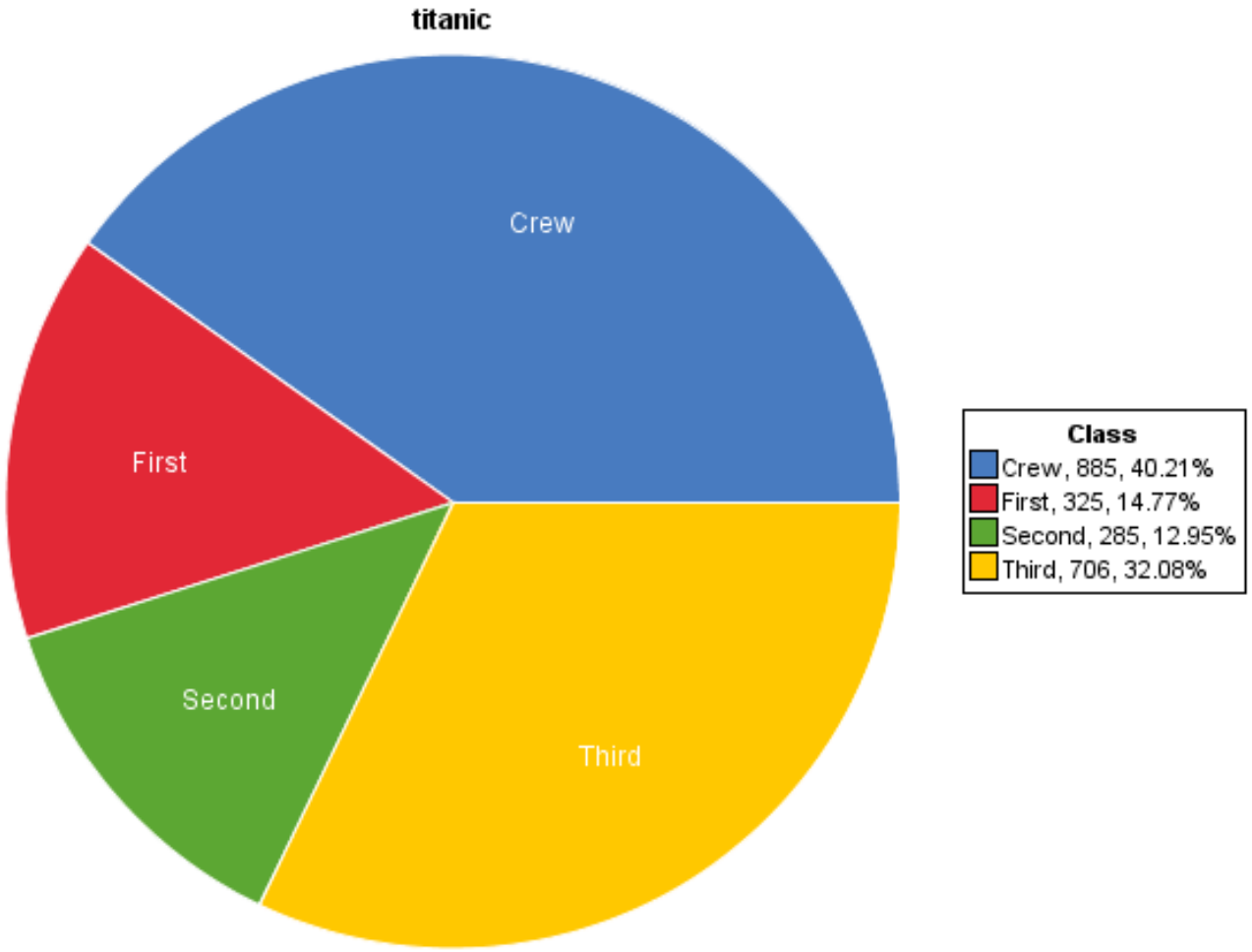
If we want the height of the bar to be a non-integer value (such as proportions) we can use `geom_col`.

```
## change y-axis
ggplot(titanClassProp,
       aes(x=Class, y=prop, fill=Class))+
  geom_col()
```



PIE CHARTS

Pie Charts show the distribution of a categorical variable as a "pie" whose slices are sized by the counts or percent for the categories.

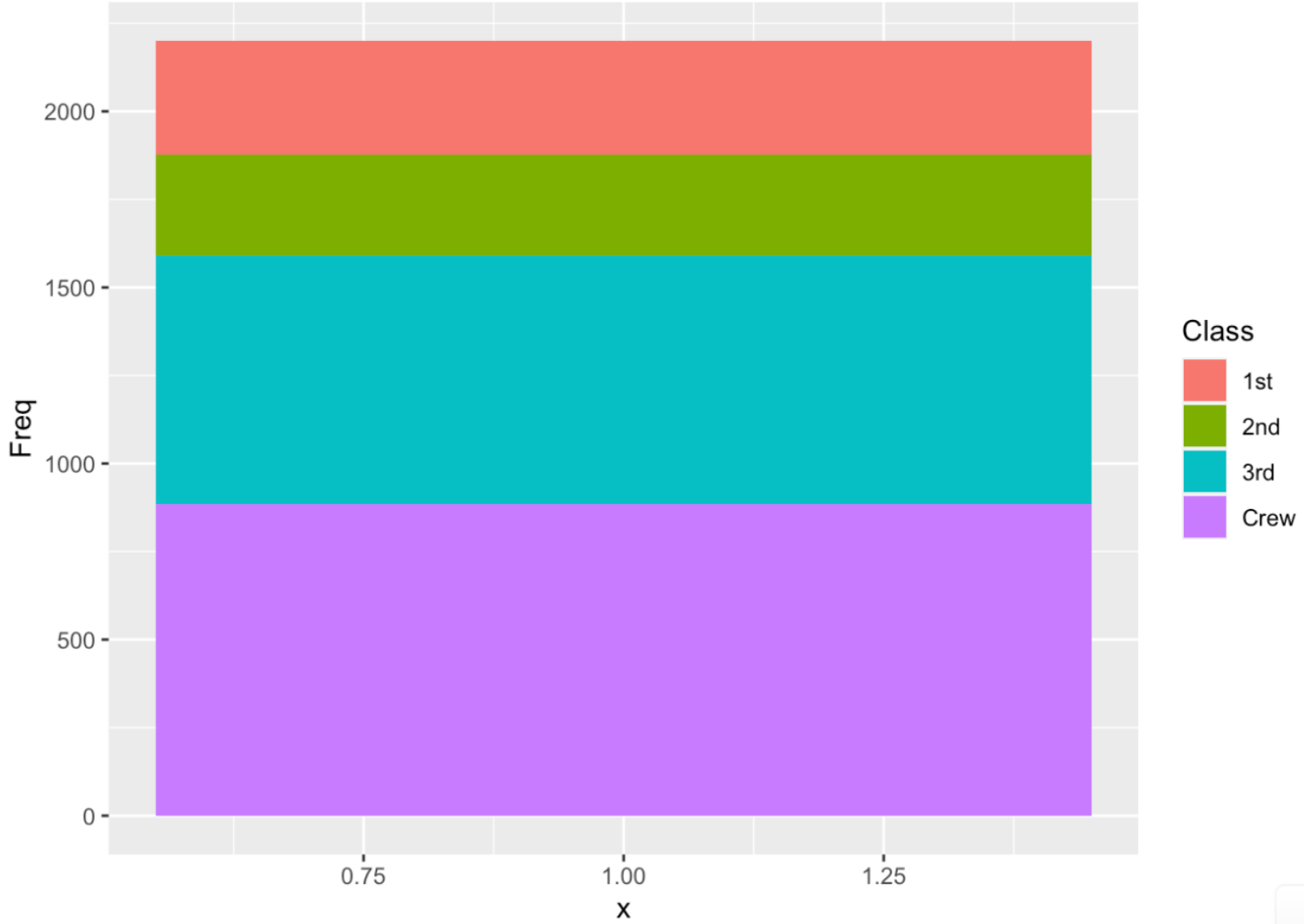


PIE CHARTS

Recipe for a Pie chart in R

RECIPE FOR A PIE CHART

```
## Pie chart
## 1) Start with a stacked bar
ggplot(Titanic, aes(x=1, y=Freq, fill=Class))+
  geom_bar(stat = "identity")
```

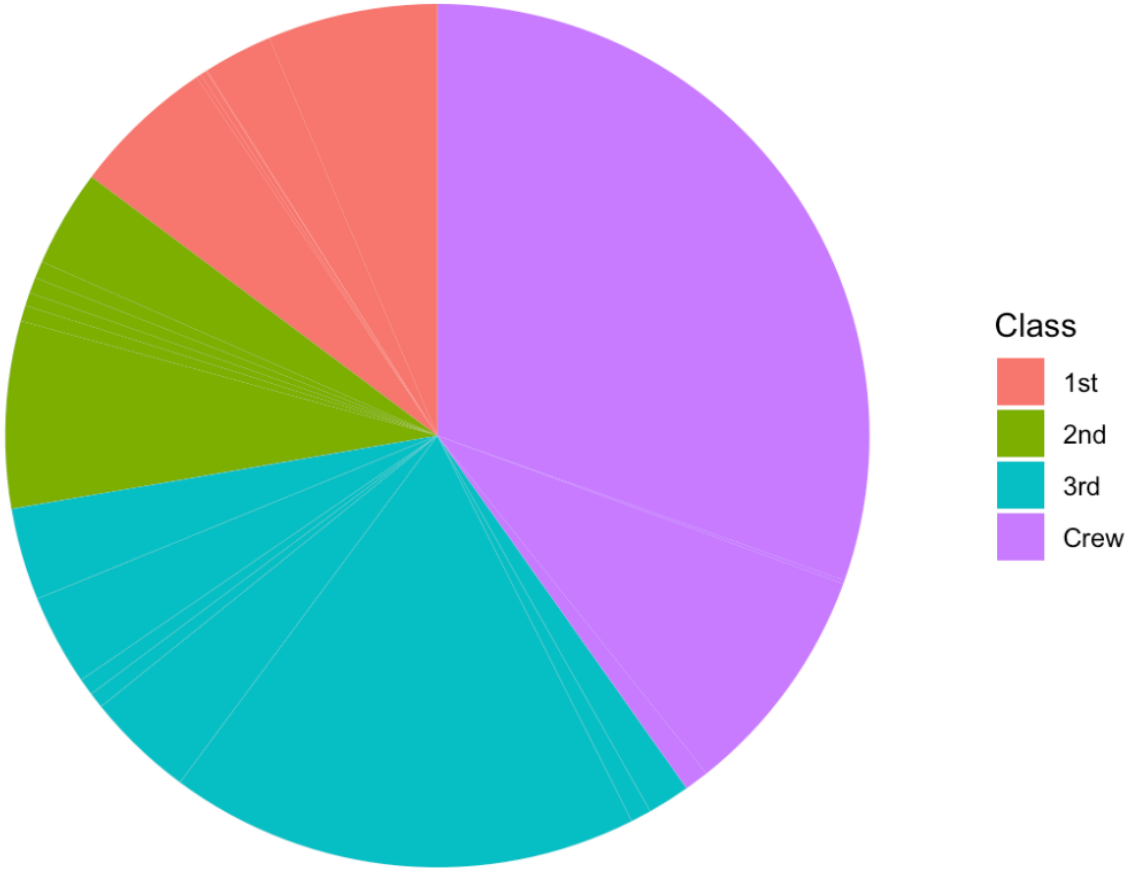


Step 1: Make a stacked bar graph

RECIPE FOR A PIE CHART

```
## 2) plot it in a circle
ggplot(Titanic, aes(x=1, y=Freq, fill=Class))+
  geom_bar(stat = "identity")+
  coord_polar("y", start=0)+
  theme_void()
```

Step 2: Use polar coordinates



I'LL NEVER LET GO, JACK!

We have only displayed one dimension of the data, but we have other variables in our data.

Let's explore the survival rates across classes…

CONTINGENCY TABLES

Contingency Tables (aka two-way tables): The counts of how individuals are distributed across two characteristics simultaneously.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

DISTRIBUTIONS

Joint Distribution

The joint probability distribution of X, Y gives the probability that each X, Y fall in a particular range or discrete set of values specified for the variables.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	5.5%	7.6%	24.0%	30.6%	67.7%
	Total	14.8%	12.9%	32.1%	40.2%	100%

DISTRIBUTIONS

Marginal Distribution

The marginal probability distribution is where we are only interested in one of the random variables. The marginal gives the probabilities of the subject of interest without reference to the other variables.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	5.5%	7.6%	24.0%	30.6%	67.7%
	Total	14.8%	12.9%	32.1%	40.2%	100%

Marginal Distribution

Where can I find the marginal distribution for survival?

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	5.5%	7.6%	24.0%	30.6%	67.7%
	Total	14.8%	12.9%	32.1%	40.2%	100%

DISTRIBUTIONS

Conditional Distribution

A measure of the probability of an event given that another event has occurred.

Conditional distribution of survival, given ticket class.

			Class				
			First	Second	Third	Crew	Total
Survival	Alive	Count	203	118	178	212	711
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
	Dead	Count	122	167	528	673	1490
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
	Total	Count	325	285	706	885	2201
			100%	100%	100%	100%	100%

DISTRIBUTIONS

Try this at home!

Conditional distribution of class, given survival status.

Survival	Class: First	Second	Third	Crew	Total
Alive	**203**	**118**	**178**	**212**	**711**
	28.6%	16.6%	25.0%	29.8%	**100%**
Dead	**122**	**167**	**528**	**673**	**1490**
	8.2%	11.2%	35.4%	45.2%	**100%**

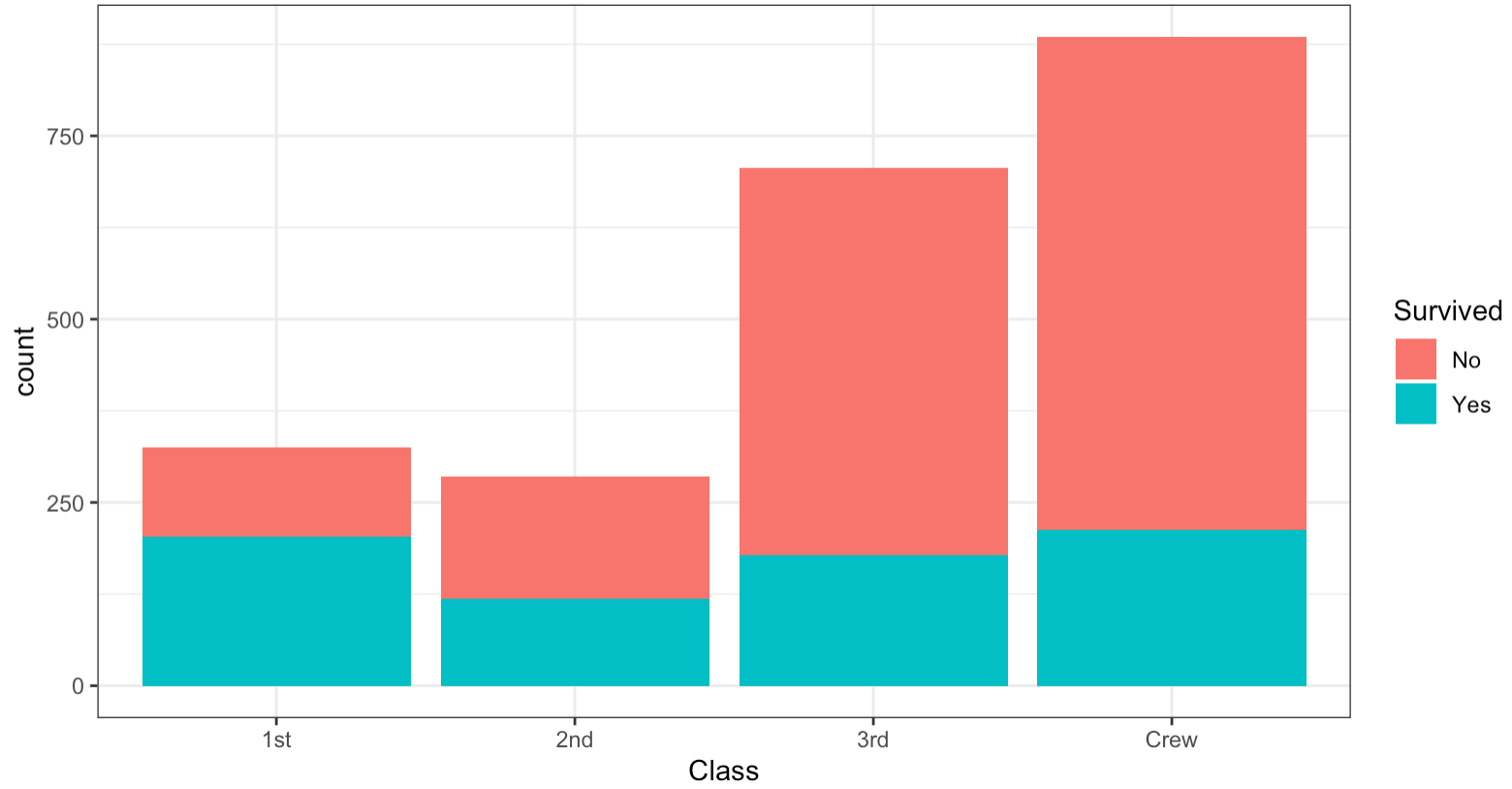
CAUTION!!!

Be careful not to confuse similar sounding probabilities:

- The percentage of the passengers who were both in first class and survived: 203/2201 = 9.2%
- The percentage of the first-class passengers who survived: 203/325 = 62.5%
- The percentage of the survivors who were in first class: 203/711 = 28.6%

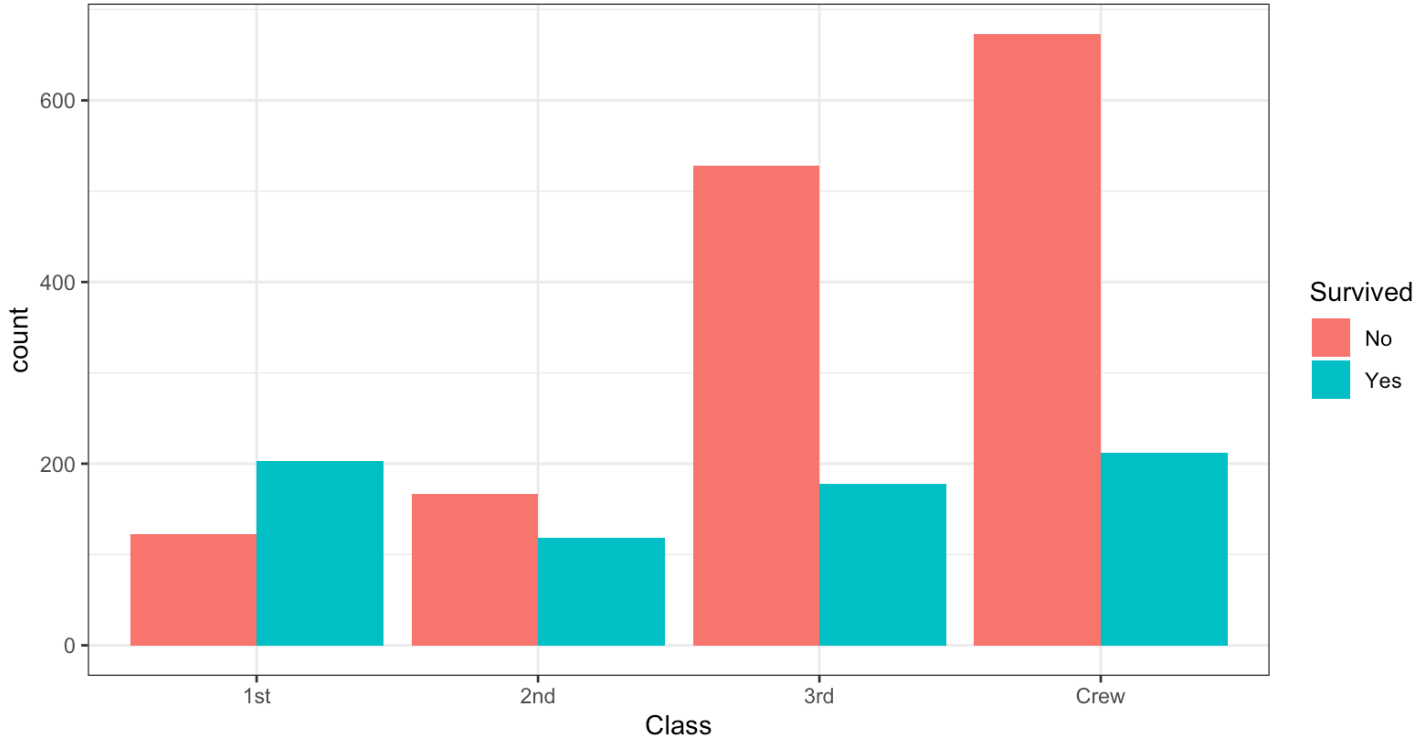
STACKED BAR CHARTS

Stacked bar charts (aka segmented bar charts): Illustrates category components within bars

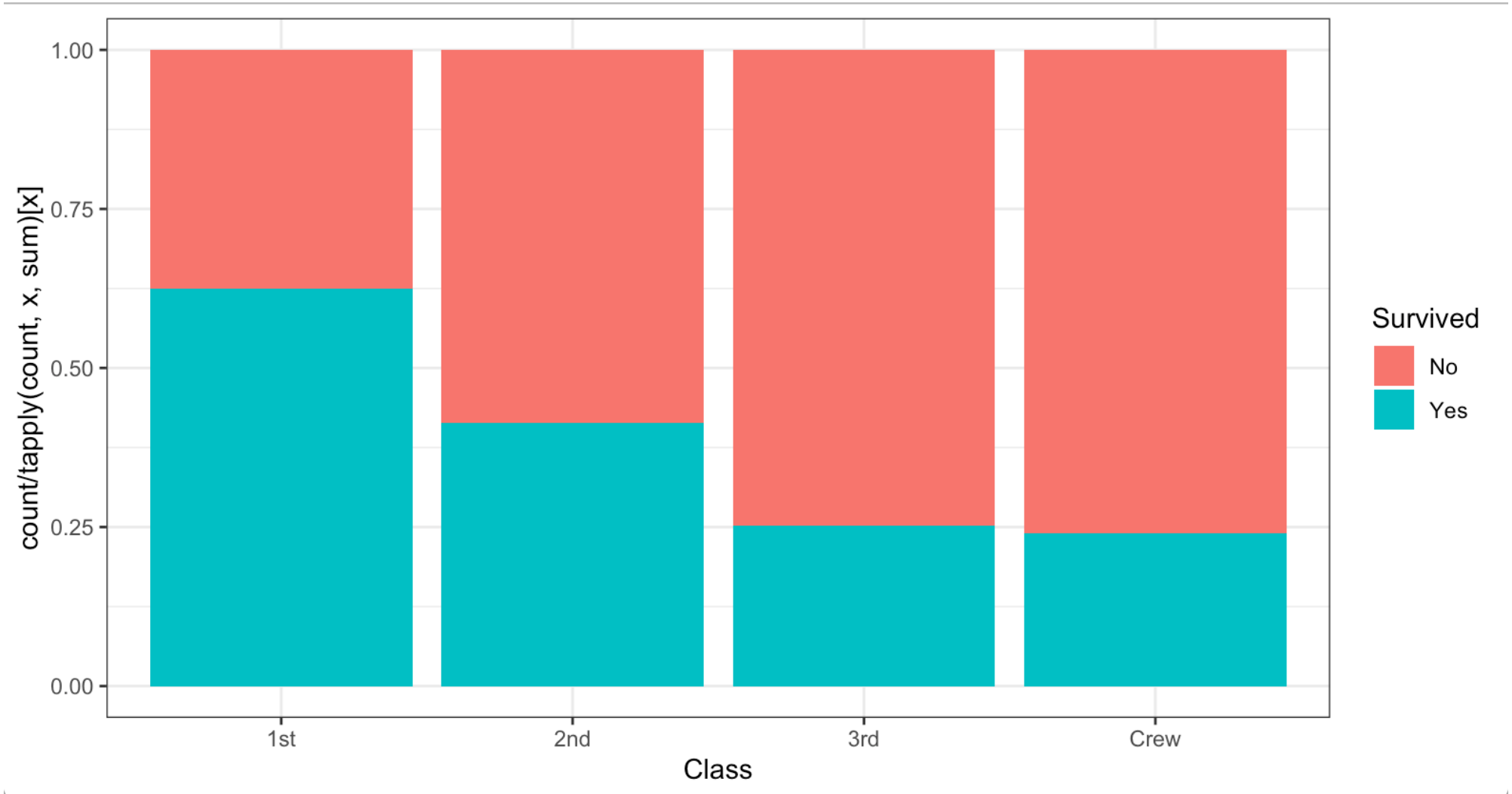


SIDE-BY-SIDE BAR CHARTS

Side-by-side bar charts: Show two dimensions of categorical data at the same time! If percentages are used, they can illustrate the joint or conditional distributions.

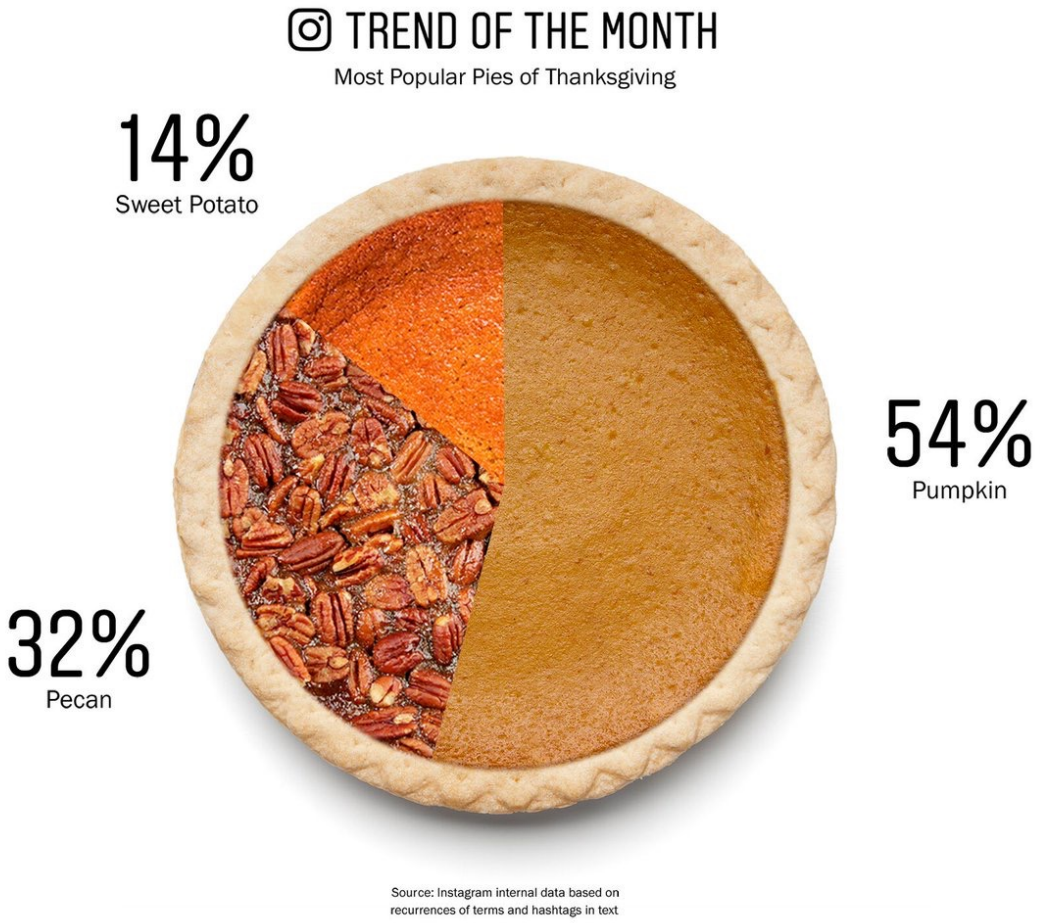


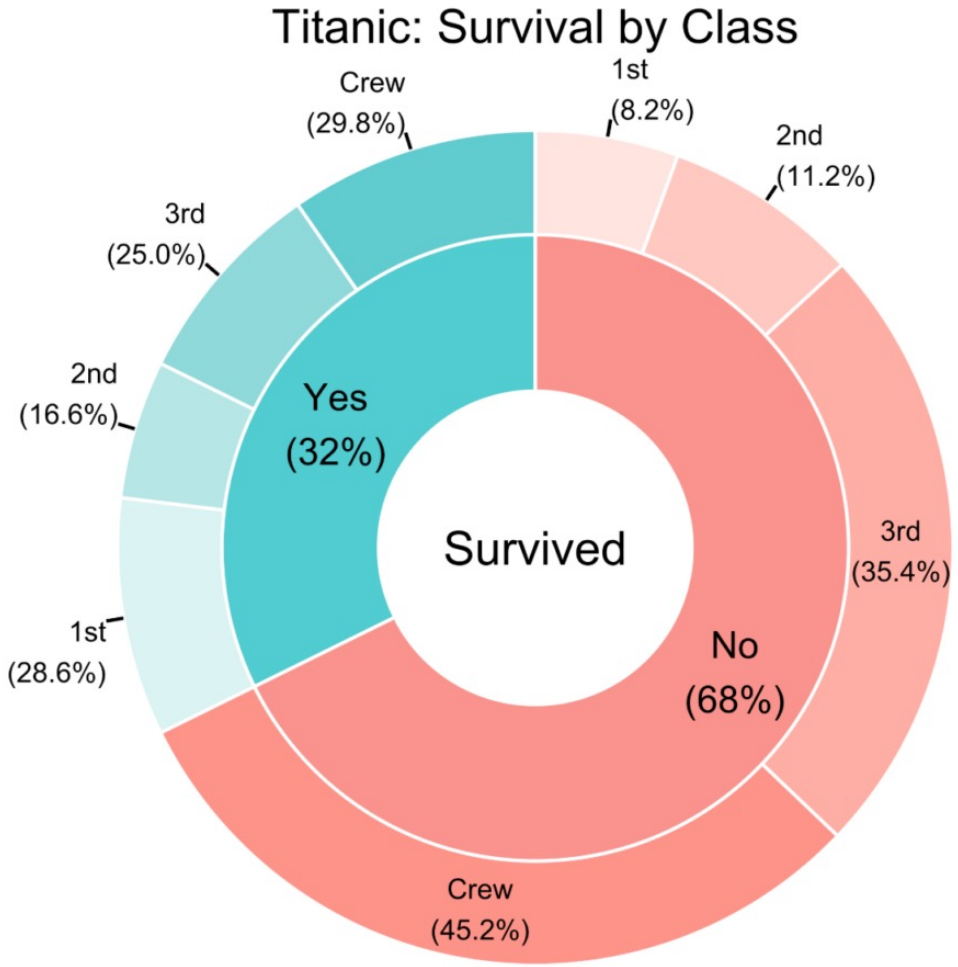
STACKED BAR CHARTS



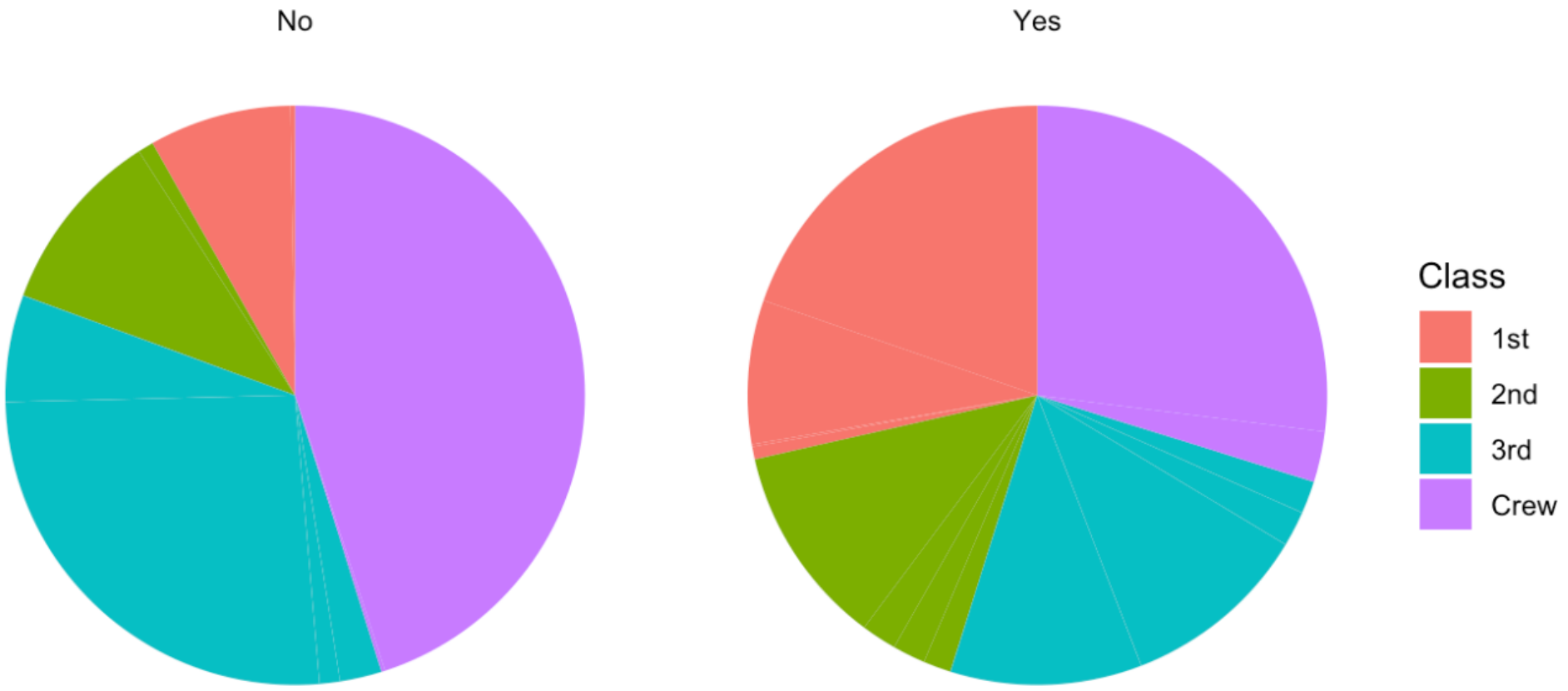
BEWARE OF CHARTS NAMED AFTER DESSERTS

AVOID DESSERT THEMED CHARTS! PIE AND DONUT CHARTS

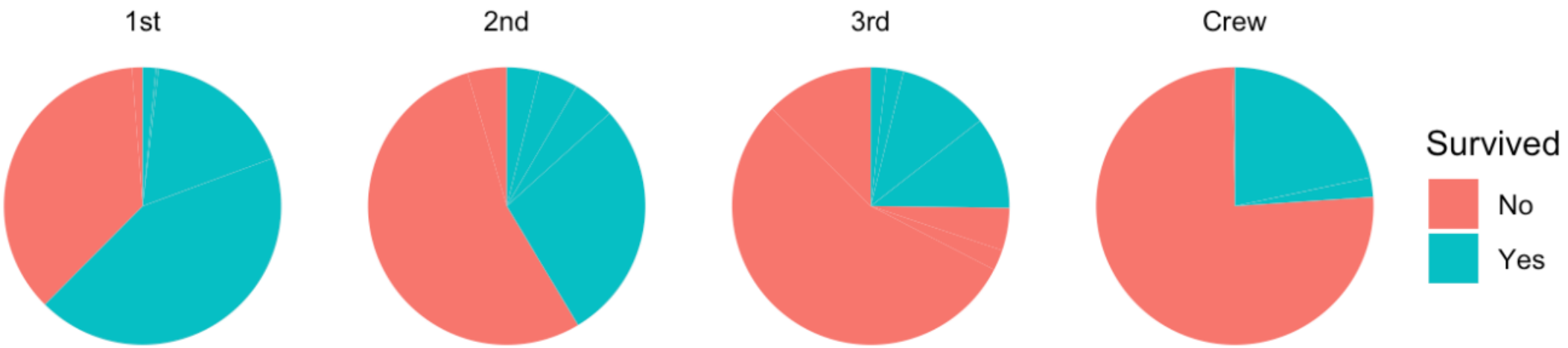




COMPARING ACROSS SURVIVAL



COMPARING ACROSS CLASS



EXAMPLE 2: IMMIGRATION POLICY

MOTIVATING EXAMPLE #2: IMMIGRATION POLICY

Views on immigration. Nine-hundred and ten (910) randomly sampled registered voters from Tampa, FL were asked if they thought workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.[48]

QUESTIONS OF INTEREST

a. What percent of these Tampa, FL voters identify themselves as conservatives?

b. What percent of these Tampa, FL voters are in favor of the citizenship option?

c. What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

d. What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

e. Do political ideology and views on immigration appear to be associated? Explain your reasoning.

TRANSITION TO R STUDIO FOR OUR HANDS-ON ACTIVITY

GETTING STARTED

Step 0: Install the package

```
#install.packages("openintro")
library(openintro)
```

Step 1: Load the Data

```
data("immigration")
str(immigration)
```

```
## tibble [910 × 2] (S3: tbl_df/tbl/data.frame)
##  $ response : Factor w/ 4 levels "Apply for citizenship",..:
1 1 1 1 1 1 1 1 1 1 ...
##  $ political: Factor w/ 3 levels "conservative",..: 1 1 1 1 1
1 1 1 1 1 ...
```

STEP 2

Step 2: Re-level categories

By default R will order a variable alphabetically, but we might not want that.

```
immigration$political<-as.character(immigration$political)
immigration$political<-factor(immigration$political,
                              levels = c("conservative","moderat
e", "liberal"))
```

QUESTION 1

What percent of these Tampa, FL voters identify themselves as conservatives?

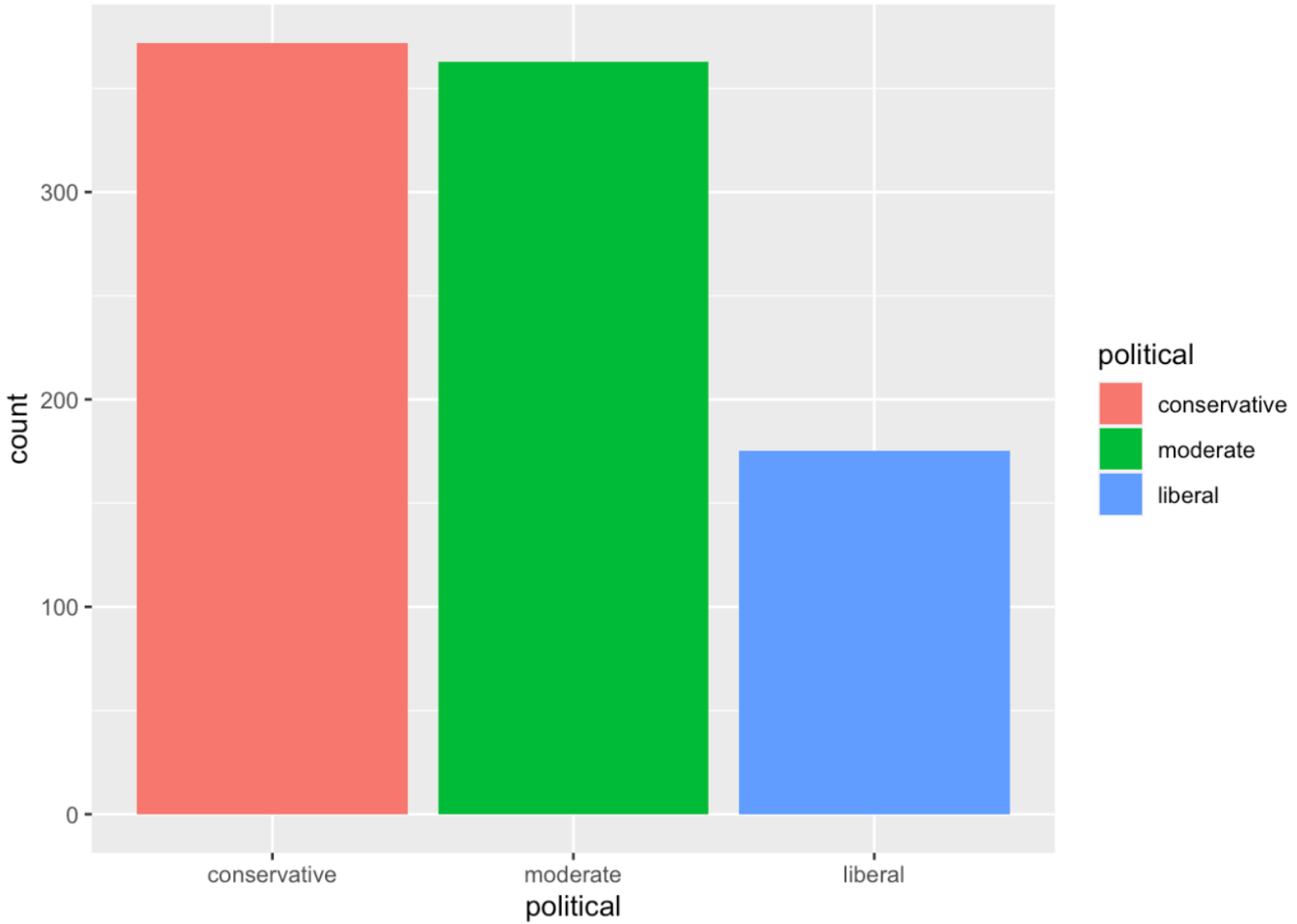
We will learn two new functions to work with individual level data:

- `table()`
- `prop.table()`

```r
# Table for Political affiliation
# use table() function
tabPol<-table(immigration$political)

# the prop.table() function must take a table object
prop.table(tabPol)
```

```
##
## conservative     moderate      liberal
##    0.4087912    0.3989011    0.1923077
```

```
# create a graph to display the distribution
ggplot(immigration, aes(x=political, fill=political))+
  geom_bar()
```



QUESTION 2

What percent of Tampa, FL voters are in favor of the citizenship option?

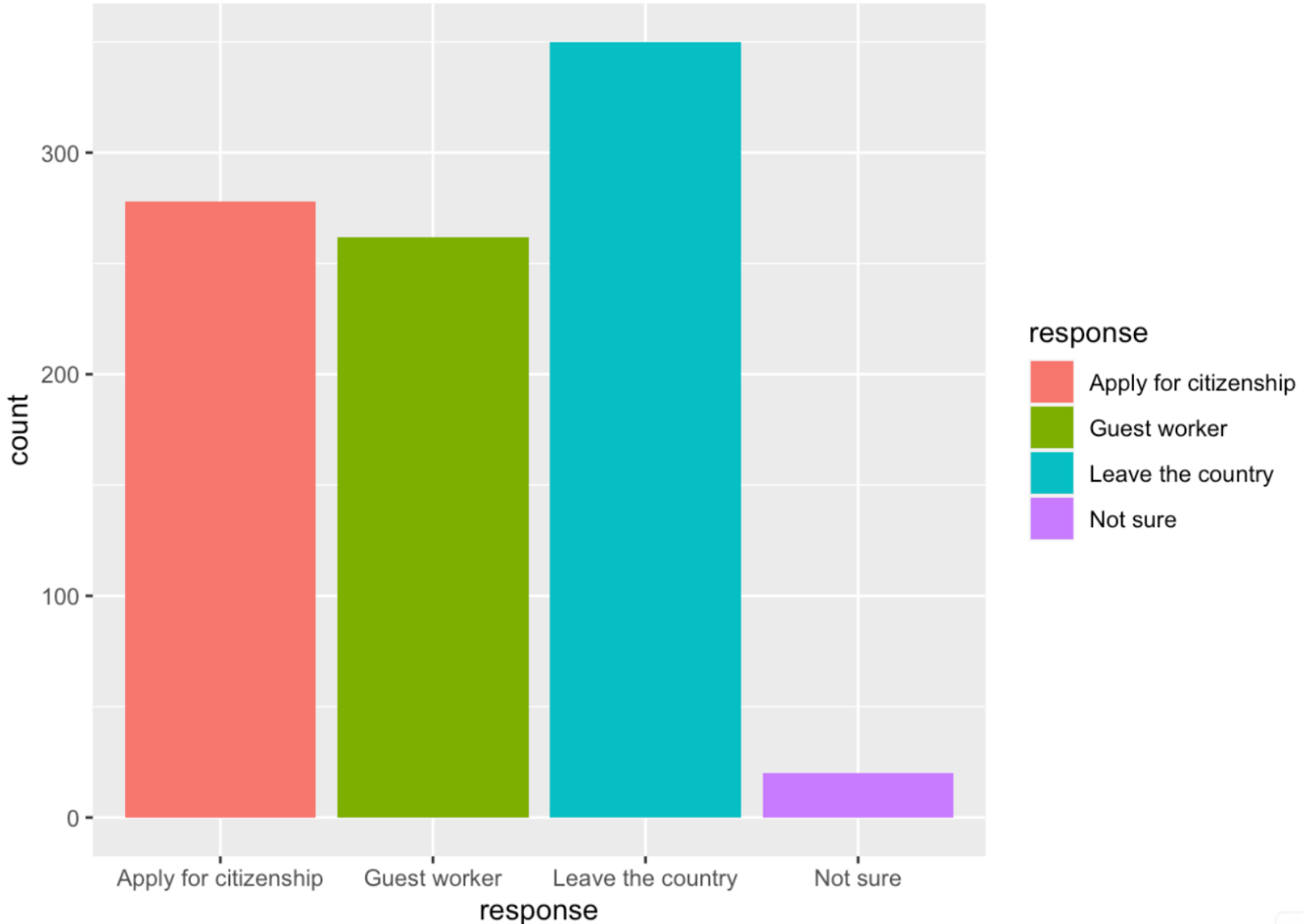
Use the functions

- `table()`
- `prop.table()`

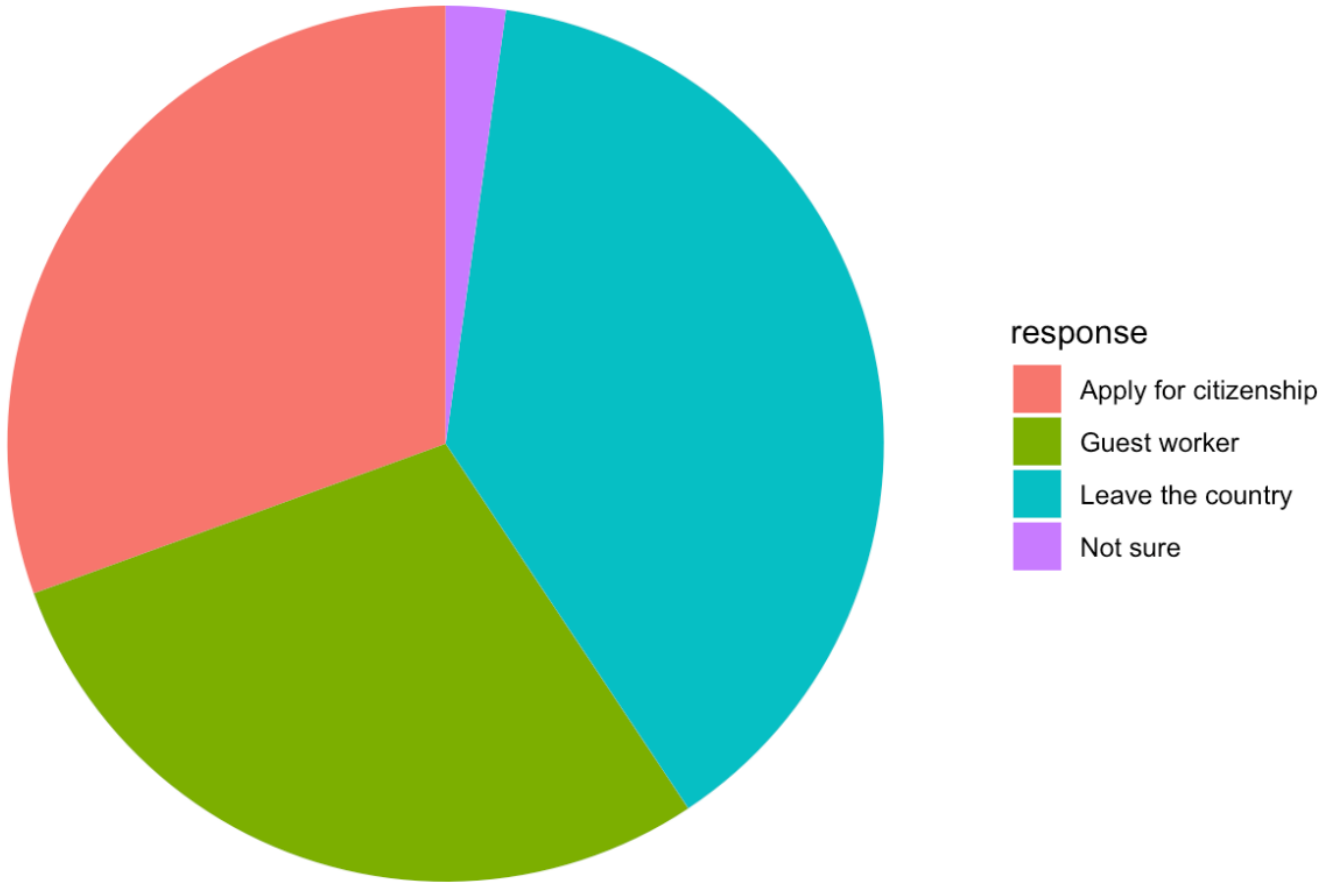
```r
# Table for citizenship response
# use table() function
tabResp<-table(immigration$response)

# use prop.table()
prop.table(tabResp)
```

```
##
## Apply for citizenship          Guest worker     Leave the cou
ntry
##             0.30549451            0.28791209             0.3846
1538
##               Not sure
##             0.02197802
```

```
# create a graph to display the distribution
ggplot(immigration, aes(x=response, fill=response))+
  geom_bar()
```



```
# pie graph
ggplot(immigration, aes(x=1, fill=response))+
  geom_bar()+
  coord_polar("y", start=0)+
  theme_void()
```



QUESTION 3

What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?

Use the functions

- `table()`
- `prop.table()`

```r
## conservative and citizen
# Row then col
tabPolResp<-table(immigration$political, immigration$response)
tabPolResp
```

```
##
##              Apply for citizenship Guest worker Leave the country Not sure
##   conservative                    57           121                179       15
##   moderate                       120           113                126        4
##   liberal                        101            28                 45        1
```

```
## joint
prop.table(tabPolResp)
```

```
##
##                Apply for citizenship Guest worker Leave the country    Not sure
##   conservative            0.062637363  0.132967033       0.196703297 0.016483516
##   moderate                0.131868132  0.124175824       0.138461538 0.004395604
##   liberal                 0.110989011  0.030769231       0.049450549 0.001098901
```

```
sum(prop.table(tabPolResp))
```

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

QUESTION 4

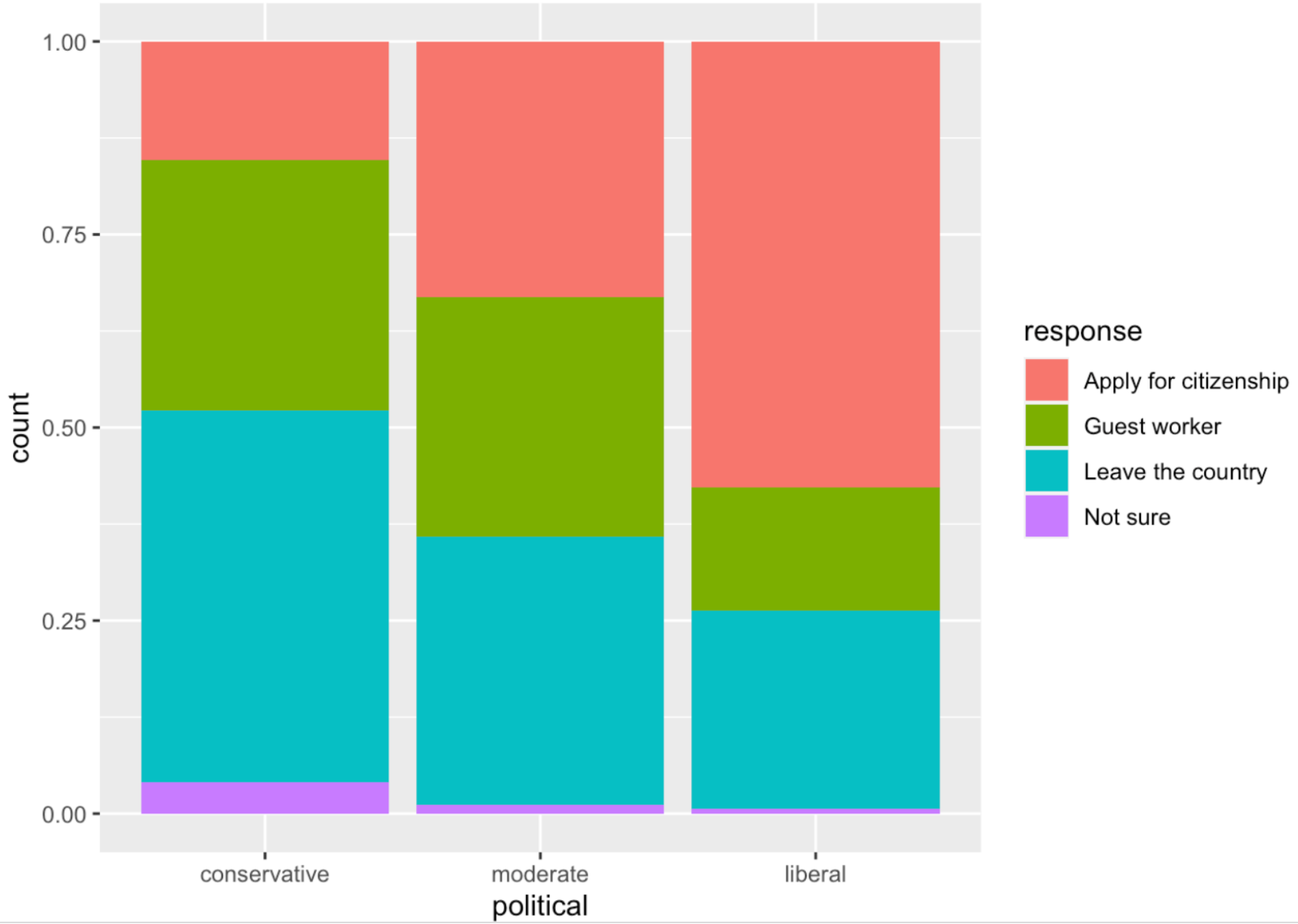
What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates share this view? What percent of liberals share this view?

```
## marginal prop
prop.table(tabPolResp, 1)
```

1 = marginal on the row dimension

```
##
##                Apply for citizenship Guest worker Leave the country     Not sure
##   conservative             0.153225806  0.325268817        0.481182796 0.040322581
##   moderate                 0.330578512  0.311294766        0.347107438 0.011019284
##   liberal                  0.577142857  0.160000000        0.257142857 0.005714286
```

```
ggplot(immigration, aes(x=political, fill=response))+
  geom_bar(position="fill")
```



EXAMPLE 3: SIMPSON'S PARADOX

SIMPSON'S PARADOX

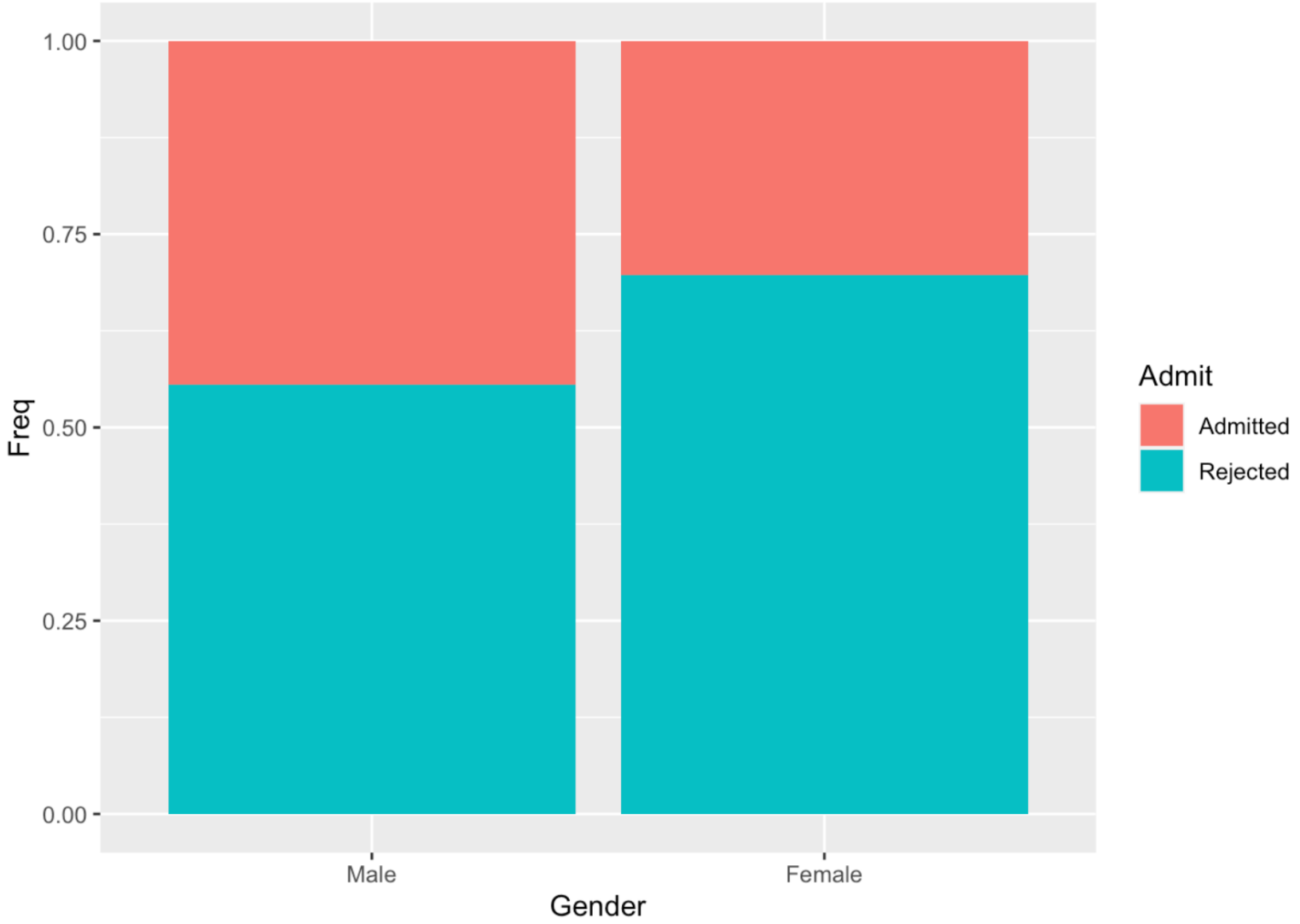
Simpson's Paradox (aka The Ecological Fallacy):
A phenomenon in which a trend appears in several different groups but disappears or reverses when the groups are combined.

EXAMPLE #3: SIMPSON'S PARADOX

- 1973 UC Berkeley Gender Bias in Admissions
- "One of the first universities to be sued for sexual discrimination" (with a statistically significant difference)

	Applicants Admitted	
Men	8442	44%
Women	4321	35%

EXAMPLE #3: SIMPSON'S PARADOX

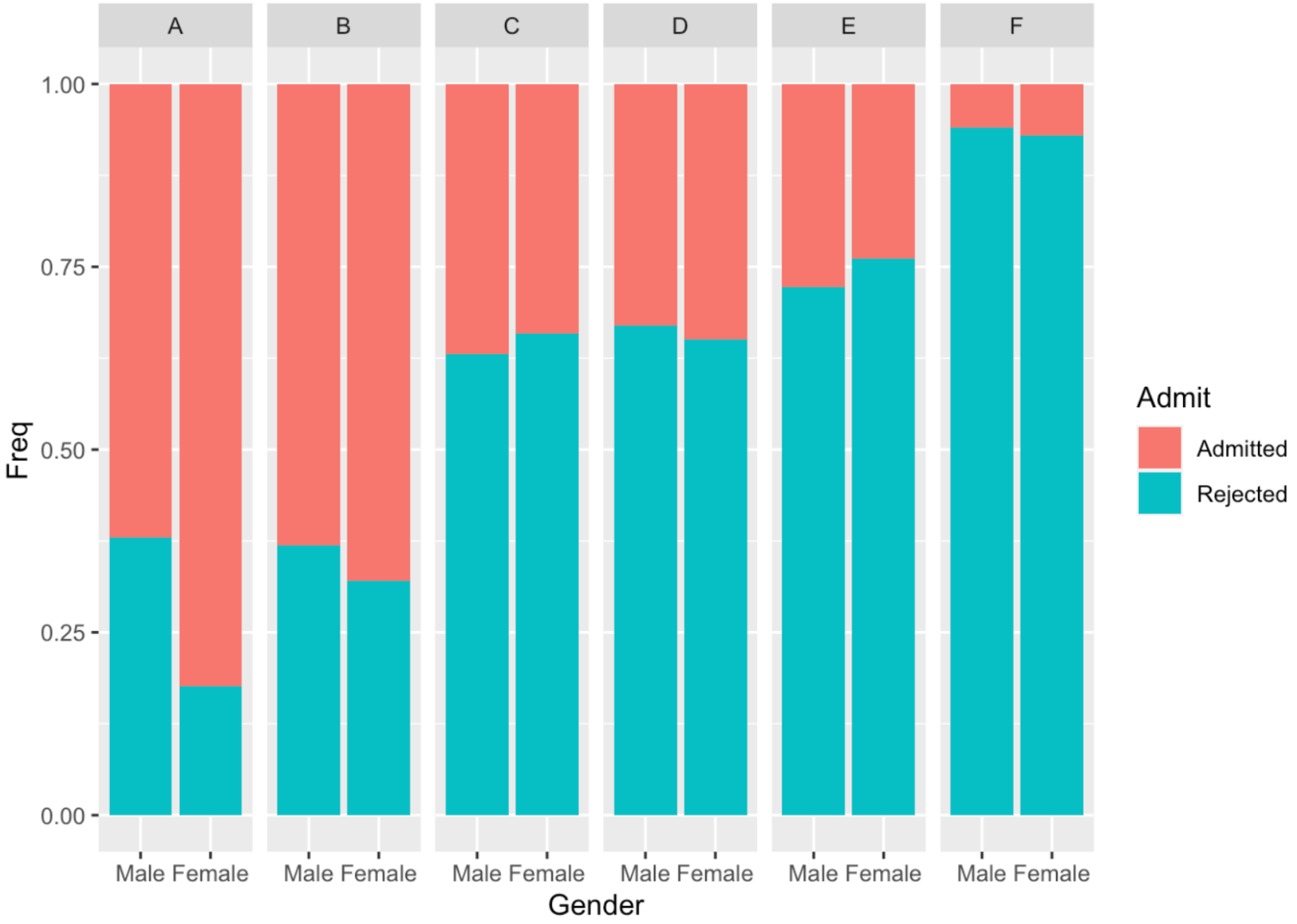


SIMPSON'S PARADOX

- But, when you dig into the data…

Department	# of Men	# of Women	Men Accepted	Women Accepted
A	**825**	**108**	62%	**82%**
B	560	25	63%	**68%**
C	325	593	37%	34%
D	417	375	33%	**35%**
E	191	393	28%	24%
F	**373**	**341**	6%	**7%**
Total	**8442**	**4321**		

EXAMPLE #3: SIMPSON'S PARADOX



SIMPSON'S PARADOX

How does this happen?

"The simple explanation is that women tended to apply to the departments that are the hardest to get into, and men tended to apply to departments that were easier to get into. (Humanities departments tended to have less research funding to support graduate students, while science and engineer departments were awash with money.) So women were rejected more than men. Presumably, the bias wasn't at Berkeley but earlier in women's education, when other biases led them to different fields of study than men."

BONUS EXAMPLE

THE DONNER PARTY

Now its your turn to make tables!

To do this we will be importing data from an online source that contains data about the Donner Party.

The Donner Party was a group of pioneers that departed Missouri on the Oregon Trail in the Spring of 1846. On their journey the group experienced delays and rugged terrain that caused them to travel in extreme winter weather with low food supplies. This group is well known for the fact that they resorted to cannibalism.

MOTIVATING EXAMPLE #2: THE DONNER PARTY

Step 1: Import your data

```
# Import Data
donner<-read.table("https://raw.githubusercontent.com/kitadasmalley/MATH138/main/FALL_2021/Data/donner.txt",
                   header=TRUE)
```

LOOK AT YOUR DATA

Step 2: Look at your data

```
# Look at the first 6 rows
head(donner)
```

```
##   Age    Sex Survived
## 1  40 Female Survived
## 2  40   Male Survived
## 3  30   Male     Died
## 4  28   Male     Died
## 5  40   Male     Died
## 6  45 Female     Died
```

```
# Look at the last 6 rows
tail(donner)
```

```
##    Age    Sex Survived
## 39  25   Male     Died
## 40  30   Male     Died
```

ONE-DIMENSIONAL TABLE

Step 3: Look at the one-dimensional distribution for survival

A) Create a table

```
# One dim table
table_surv<-table(donner$Survived)
table_surv
```

```
##
##      Died Survived
##        24       20
```

B) Let's look at the relative frequency

```
prop.table(table_surv)
```

```
##
##       Died   Survived
## 0.5454545 0.4545455
```

SIMPLE BAR GRAPH

SIMPLE PIE CHART

TWO WAY TABLE

Step 4: Look at the two-dimensional distribution for sex and survival

A) Create a two-way table

```
# Create a frequency table
# Row = Sex
# Col = Survived
table_survFM<-table(donner$Sex, donner$Survived)
table_survFM
```

```
##
##           Died Survived
##   Female     5       10
##   Male      19       10
```

B) Look at distributions using this table

I) These are marginal tables

```
# Sex frequencies (summed over Survival)
# Use 1, to sum over columns
margin.table(table_survFM, 1)
```

```
##
## Female   Male
##     15     29
```

```
# Survival frequencies (summed over Sex)
# Use 2, to sum over rows
margin.table(table_survFM, 2)
```

```
##
##     Died Survived
##       24       20
```

II) Find the relative proportions

Table 1: Joint distribution

```
# cell percentages (joint distribution)
prop.table(table_survFM)
```

```
##
##                Died  Survived
##   Female 0.1136364 0.2272727
##   Male   0.4318182 0.2272727
```

Table 2: Conditional distribution for survival by sex

```
# row percentages (conditional distribution for survival by sex)
prop.table(table_survFM, 1)
```

```
##
##                Died  Survived
##   Female 0.3333333 0.6666667
##   Male   0.6551724 0.3448276
```

Table 3: Conditional distribution for sex by surival

```
# column percentages (conditional distribution for sex by survival)
prop.table(table_survFM, 2)
```

```
##
##                Died  Survived
##   Female 0.2083333 0.5000000
##   Male   0.7916667 0.5000000
```

Which one of these tables tells the most compelling story?