Chapter 2

Summarizing Data[1]

Department of Mathematics & Statistics
North Carolina A&T State University

[1] These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

Considering categorical data

Contingency tables

A table that summarizes data for two categorical variable is called a **contingency table**.

Contingency tables

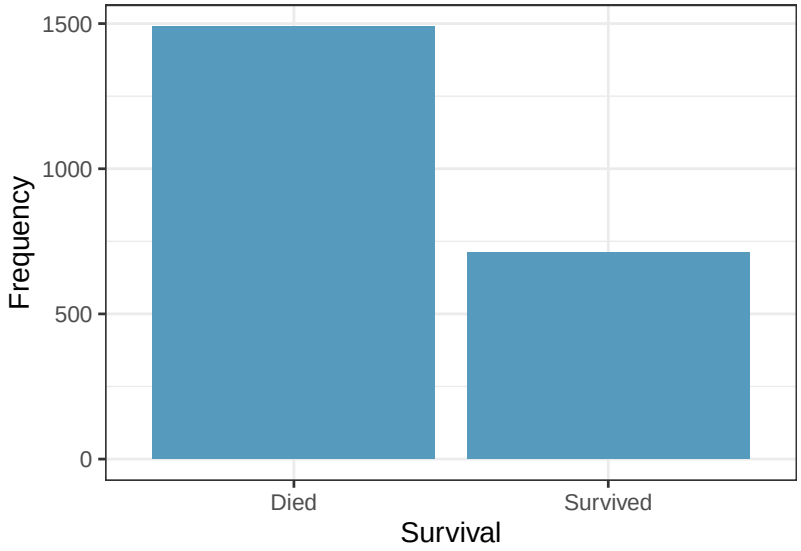
A table that summarizes data for two categorical variable is called a **contingency table**.

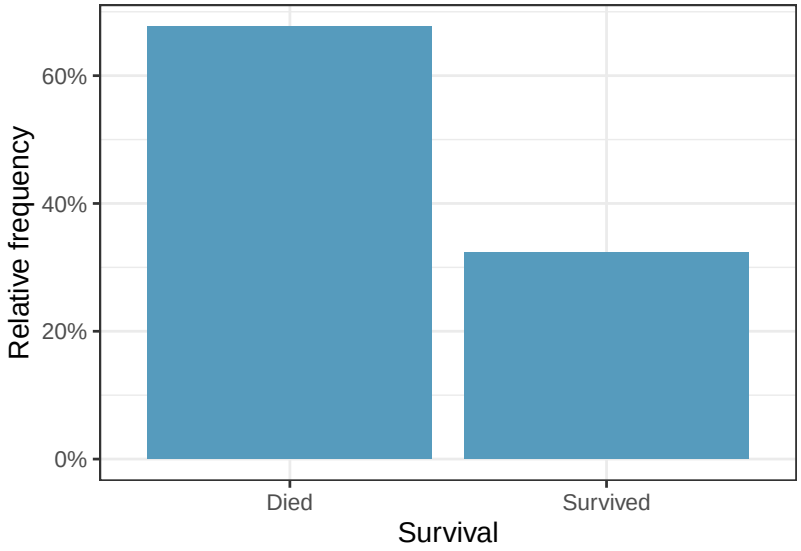
The contingency table below shows the distribution of survival and ages of passengers on the Titanic.

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

Bar Plots

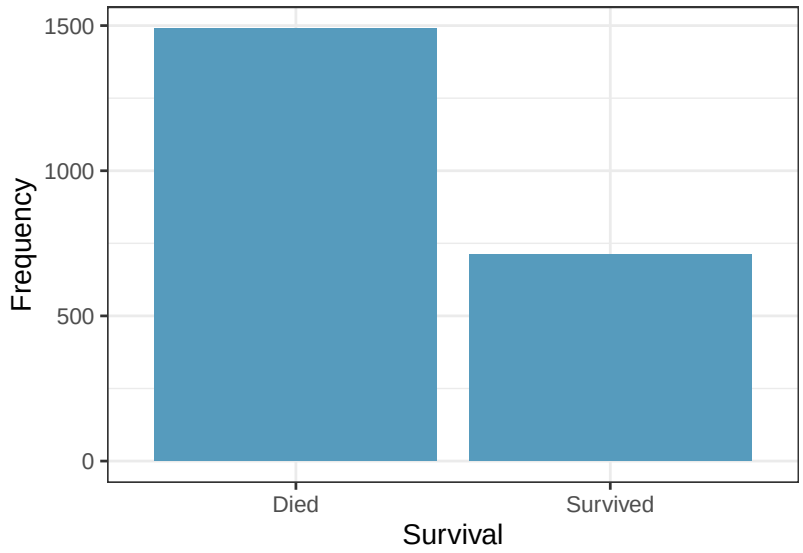
A **bar plot** is a common way to display a single categorical variable. A bar plot where proportions instead of the frequencies are shown is called a **relative frequency bar plot**.

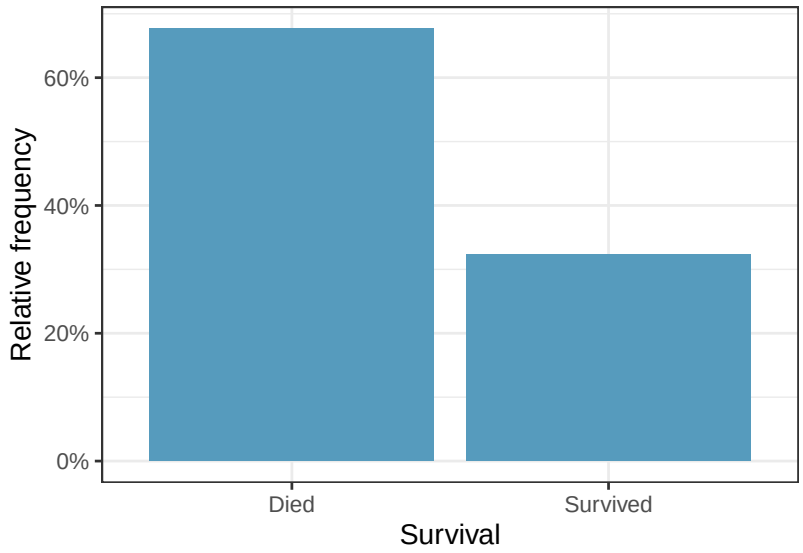




Bar Plots

A **bar plot** is a common way to display a single categorical variable. A bar plot where proportions instead of the frequencies are shown is called a **relative frequency bar plot**.

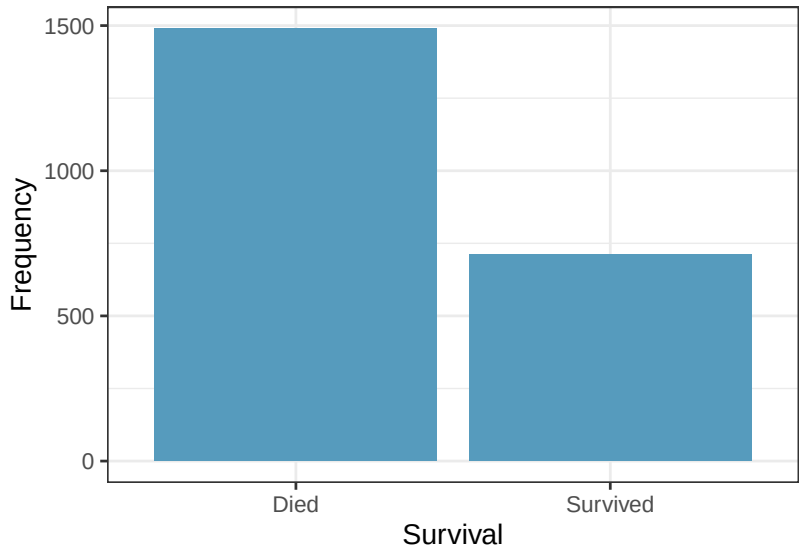


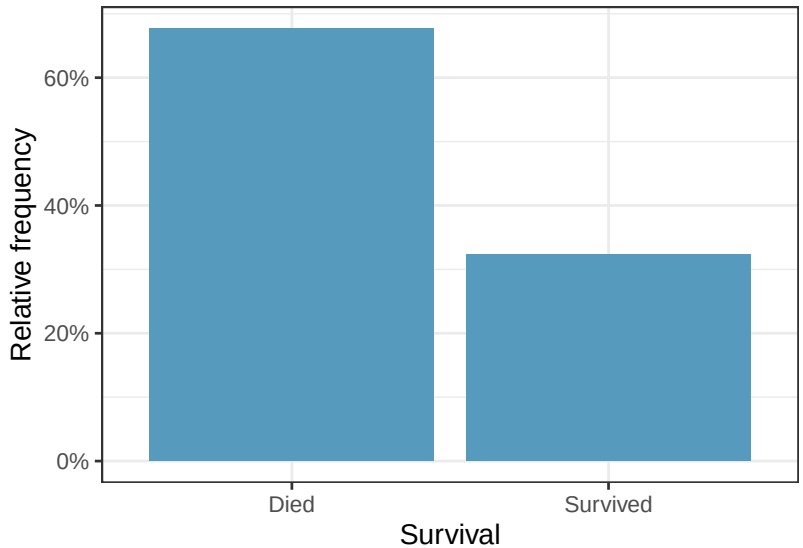


How are bar plots different than histograms?

Bar Plots

A **bar plot** is a common way to display a single categorical variable. A bar plot where proportions instead of the frequencies are shown is called a **relative frequency bar plot**.





How are bar plots different than histograms?

Bar plots are used for displaying distributions of categorical variables, histograms are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bards cannot be changed. In a bar plot, the categories can be listed in any order (though some ordering make more sense than others, especially for ordinal variables.)

Choosing the appropriate proportion

Does there appear to be a relationship between age and survival for passengers on the Titanic?

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

Choosing the appropriate proportion

Does there appear to be a relationship between age and survival for passengers on the Titanic?

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

To answer this question we examine the row proportions:

Choosing the appropriate proportion

Does there appear to be a relationship between age and survival for passengers on the Titanic?

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

To answer this question we examine the row proportions:

- % Adults who survived: $654 / 2091 \approx 0.31$

Choosing the appropriate proportion

Does there appear to be a relationship between age and survival for passengers on the Titanic?

		Survival		
		Died	Survived	Total
Age	Adult	1438	654	2092
	Child	52	57	109
	Total	1490	711	2201

To answer this question we examine the row proportions:

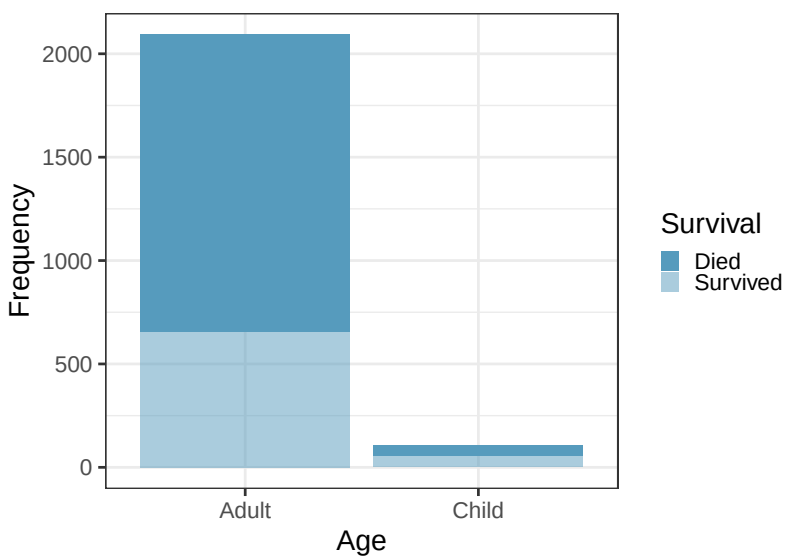
- % Adults who survived: $654 / 2091 \approx 0.31$
- % Children who survived: $57 / 109 \approx 0.52$

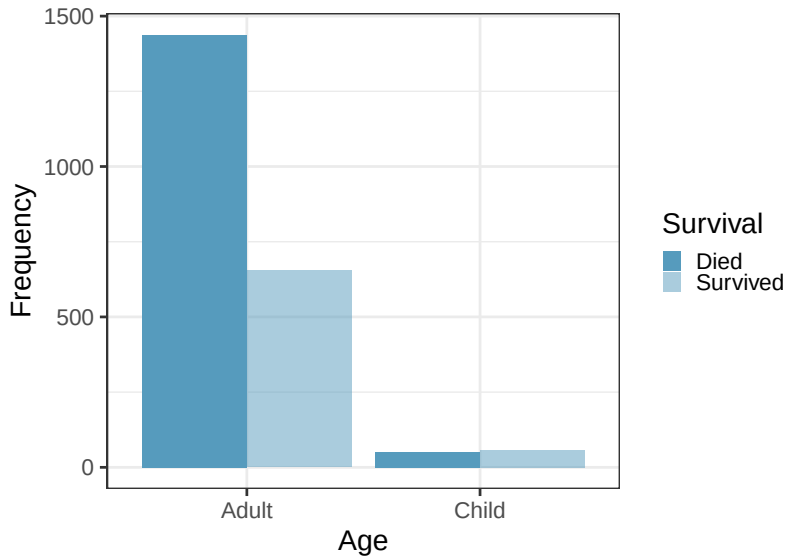
Bar plots with two variables

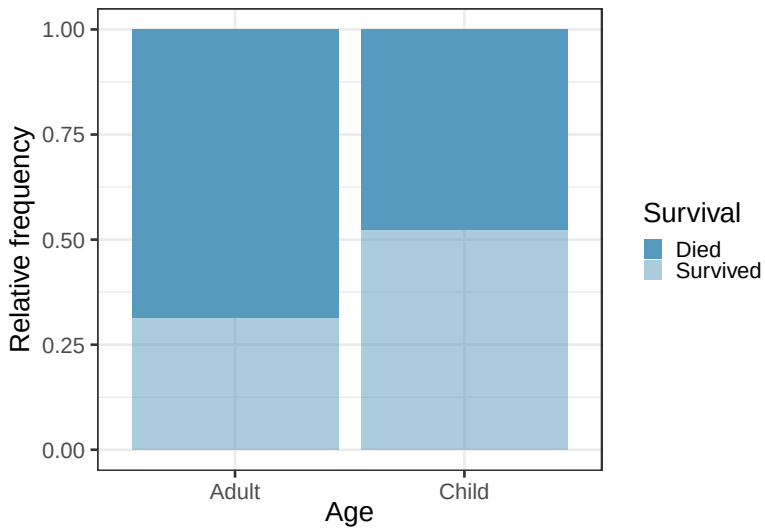
- **Stacked bar plot:** Graphical display of contingency table information, for counts.
- **Side-by-side bar plot:** Displays the same information by placing bars next to, instead of on top of, each other.
- **Standardized stacked bar plot:** Graphical display of contingency table information, for proportions.

Bar plots with two variables

What are the difference between the three visualizations shown below?

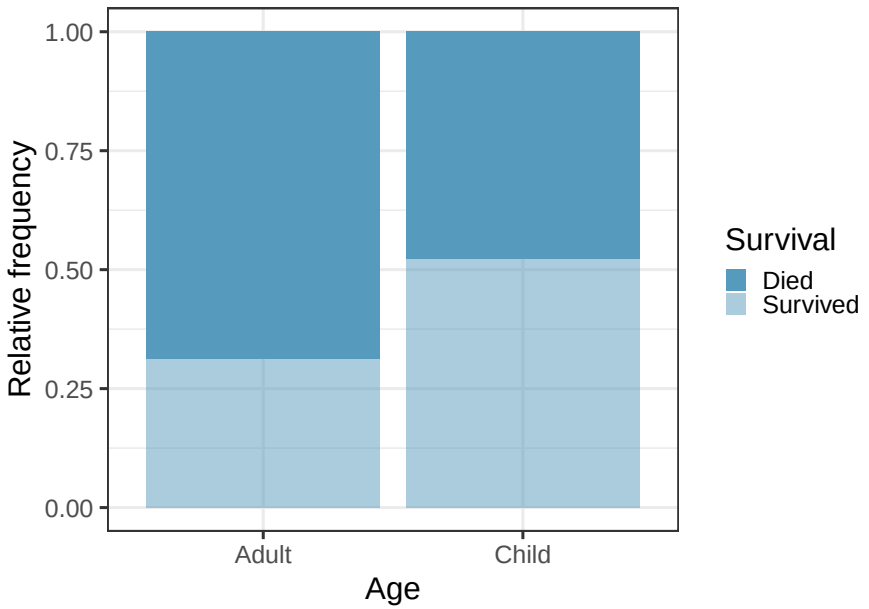


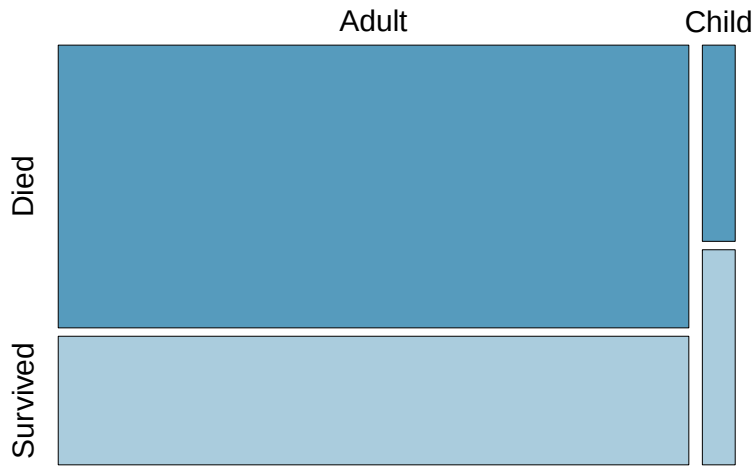




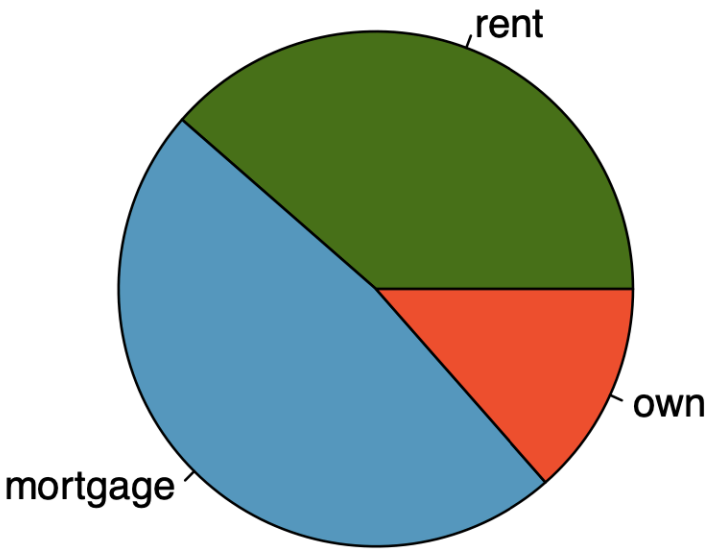
Mosaic plots

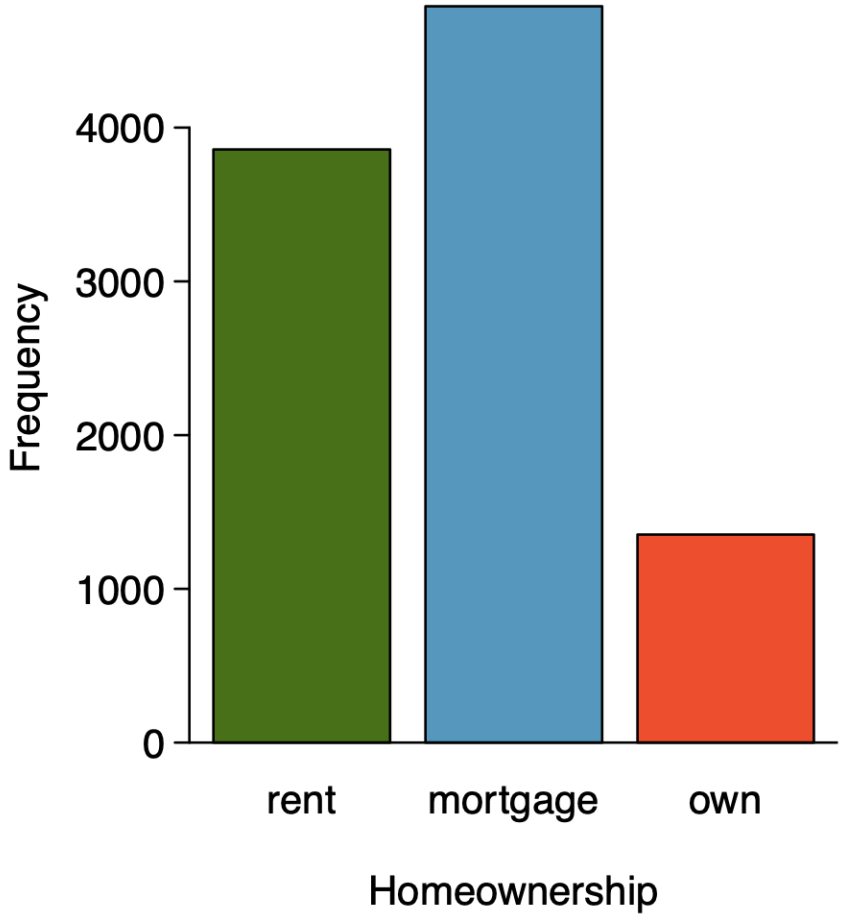
What is the difference between the two visualizations shown below?





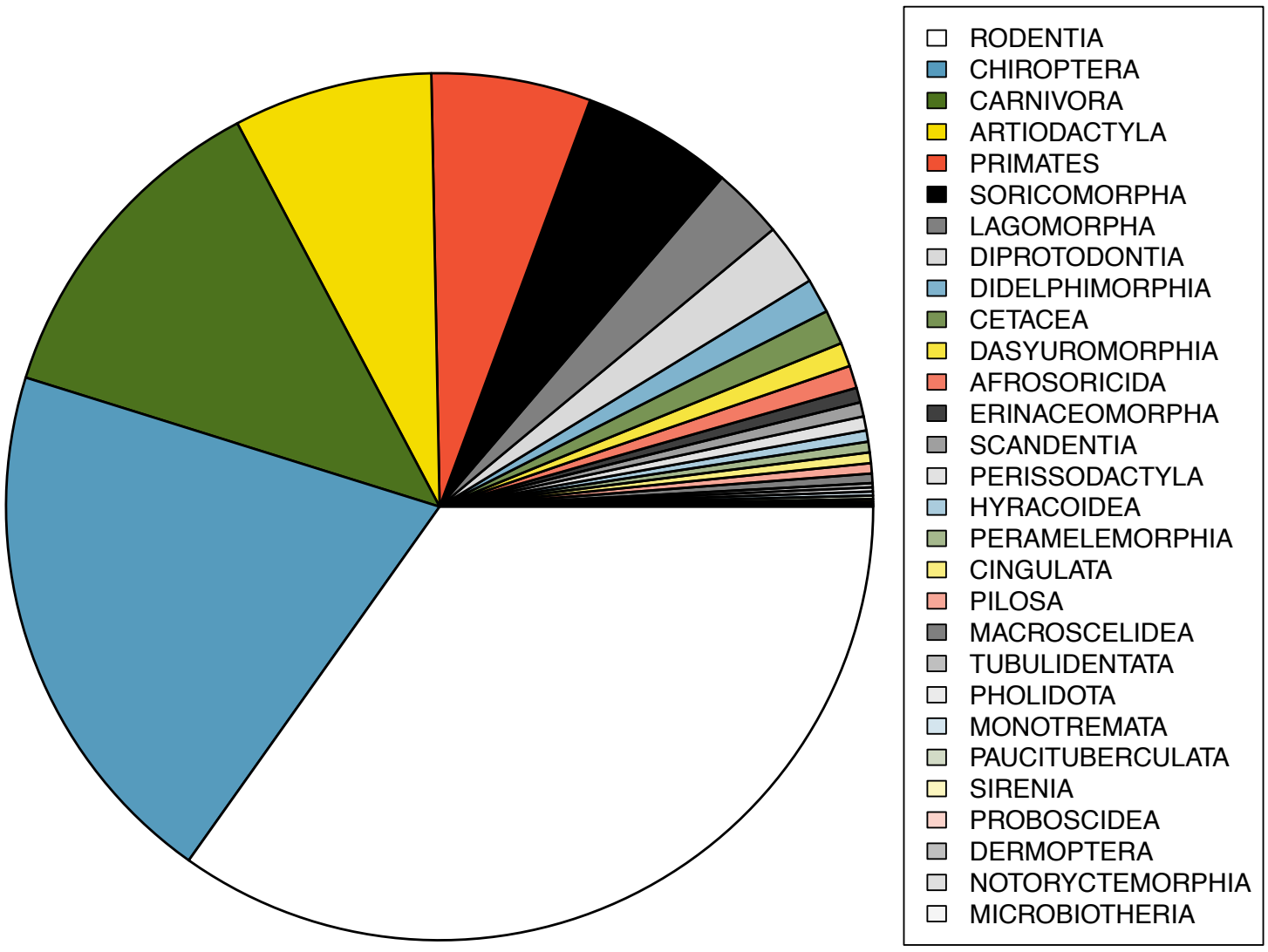
Pie charts


rent
own
mortgage


4000
3000
2000
1000
0
Frequency
rent
mortgage
own
Homeownership

Pie charts

Can you tell which order excompasses the lowest percentage of mammal species?



Side-by-side box plots

Does there appear to be a relationship between class year and number of clubs studetns are in?