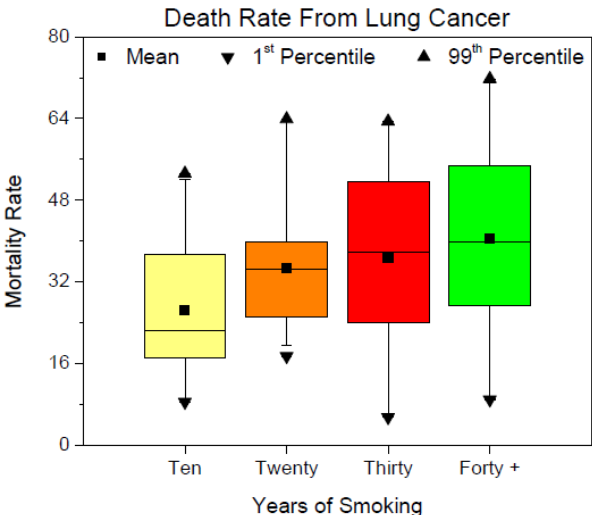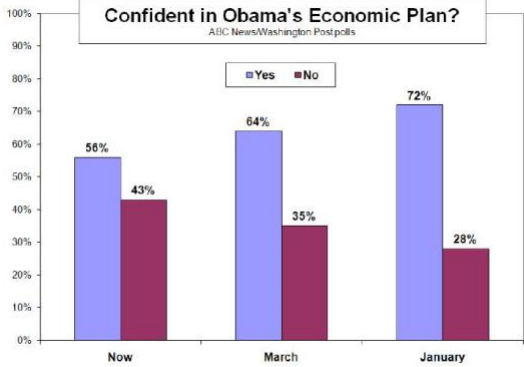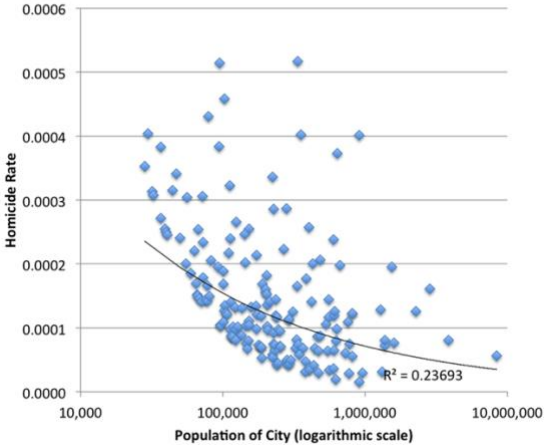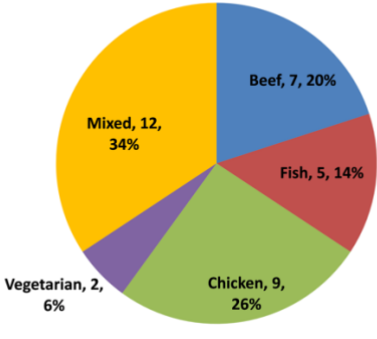# MATH 345: Homework 2

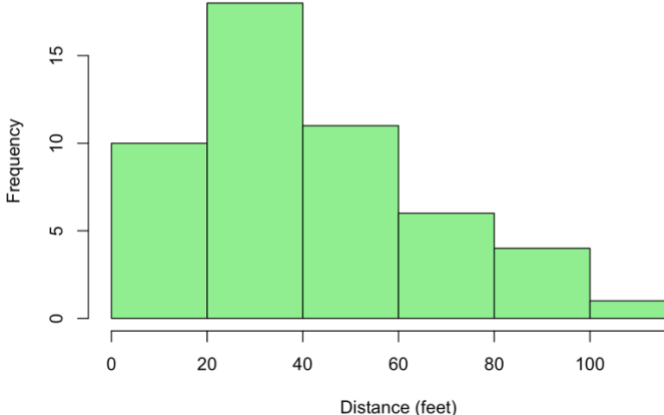UMass Boston – Spring 2023

Exercise 1

For each of the following graphs:
- Indicate the type of the graph (for example, scatter plot, histogram).
- Name the variable(s) displayed – if this information is available.
- State the nature of the variable(s): numerical or categorical. If a variable is numerical, indicate whether it's continuous or discrete.

Answers for the first graph are given as an example.

Graph type: Box plot.	Graph type:
Variable(s): Years of smoking, mortality rate.	Variable(s):
Nature of variable(s): Years of smoking is numerical and continuous (although presented here as categorical). Mortality rate: numerical and continuous.	Nature of variable(s):

Population vs. 2009 Homicide Rate
(for the 174 US cities with at least 10 murders)

R² = 0.23693

Homicide Rate

Population of City (logarithmic scale)

Graph type:

Variable(s):

Nature of variable(s):

Number of dishes (35 in total)

Beef, 7, 20%
Mixed, 12, 34%
Fish, 5, 14%
Vegetarian, 2, 6%
Chicken, 9, 26%

Graph type:

Variable(s):

Nature of variable(s):

Distribution of Stopping Distancs

Frequency

Distance (feet)

Graph type:

Variable(s):

Nature of variable(s):

NUMBER OF BOXES BOUGHT

0	5 8 9
1	3 5 6 7 8 8 9 9
2	1 1 2 2 3 4 5 8
3	0 1 2 3 3 5 6 7
4	5

KEY

1 | 5 = 15

Graph type:

Variable(s):

Nature of variable(s):

Exercise 2

One of these two figures is misleading. Which one and why?
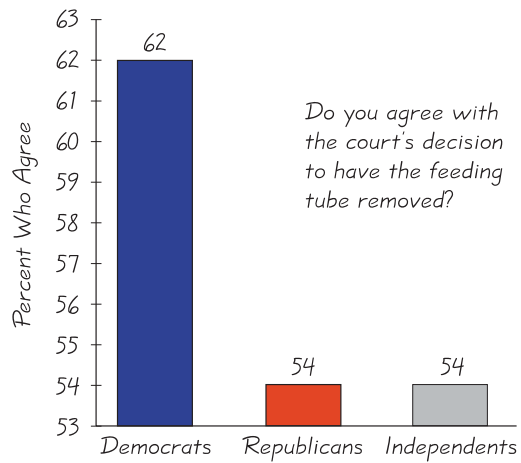
Figure 2-1 **Survey Results by Party**
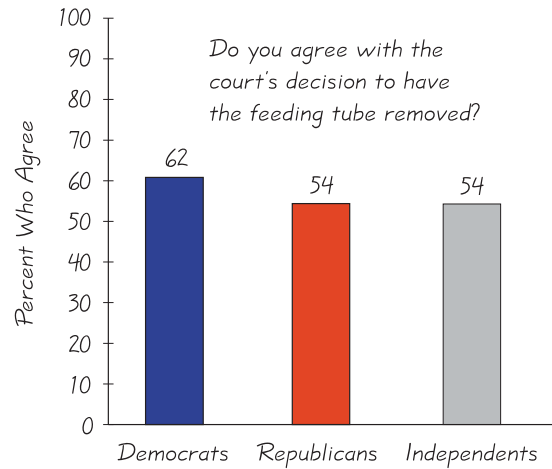
Figure 2-9 **Survey Results by Party**

Exercise 3

This R exercise utilizes the file `world-covid-2023-02-03.txt` available under the same item as HW2 on Blackboard. The source of the data is
https://www.worldometers.info/coronavirus/

1) Import the data in R or RStudio and give the dataset a shorter name for easier manipulation. For example, `covid`. To do this, open and run the lines of the attached file `HW2-Ex3.R`.
   Note 1: make sure that the data file is in your current working directory OR specify the entire file path in `read.table` so that R knows where to find the file.
   Note 2: if you use instead the button "Import Dataset" in RStudio, you may run into difficulties because it won't let you indicate missing values as "N/A ".
2) How many rows does the dataset have? How many columns, ii.e., variables? Hint: you can use the functions `str` or `dim`.
3) What may be a cause for the missing values in the dataset?
4) Using the function `plot`, create a scatter plot of the population size versus the total number of cases. After viewing the two variables on a linear scale, plot them again on a logarithmic scale using the argument `log = "xy"`. Comment on the relationship between these two variables.
5) Create a histogram of the number of tests per million people. Comment on its shape. Then plot the histogram of the logarithm in base 10 of the same variable. How do the two histograms compare?
6) Which country has the highest ratio of active cases over population size?
   Hint: divide the first corresponding column by the second and use `which.max`.

Exercise 4

Exercise 1.30 of the textbook (HDL, "good cholesterol").