Unit 2 – Data Visualization
Solutions
R Users

Question #1

The World Health Organization (WHO) records the annual number of confirmed cases of human Avian Influenza A/(H5N1). Following are a subset of their data:

Year:	2003	2004	2005	2006	2007	2008
# cases:	4	32	43	79	59	26

a) By any means you like, construct a **frequency/relative frequency table** summarization of these data.

b) By any means you like, construct a **bar graph** summarization of these data. Include a title and label your axes.

c) State the **facts** of your **bar graph**.

d) In 1-2 sentences, **interpret** the table and bar graph you have produced.

Before you begin, from the course website, right click to download the following:
(1) https://people.umass.edu/biep540w/datasets/q1data.Rdata
(2) https://people.umass.edu/biep540w/datasets/cholesterol_smokers.Rdata
(3) https://people.umass.edu/biep540w/datasets/cholesterol_nonsmokers.Rdata

Before you begin, in R-Studio console window ONE TIME installation of packages (if you have not done so already)

```
install.packages("summarytools")
install.packages("ggplot2")
install.packages("gridExtra")
install.packages("aplpack")
```

Launch R-Studio and load data. User edits yellow highlighted

```
setwd("/Users/cbigelow/Desktop")
options(scipen=1000)                                    # Turn OFF scientific notation (handy)
```

__ 1a) Frequency/relative frequency table

Good to Know:
In R Studio, you load or import **files** into your session (these might be ".Rdata" or ".xlsx" files and so on). Once that is accomplished, you should then see in your **environment** window a listing of the **objects** that are now available for you to use. To be clear, R commands operate on **objects** not on files. Objects are all kinds of things. We'll get to these. For the moment, the objects that you will be working with are **dataframes**. You can think of a dataframe as a spreadsheet (rows define subjects, columns define variables)

Hack:
Keep track of object names by looking at what is in your environment window In this question

Name of file	**Associated name of object (dataframe)**
q1data.Rdata	q1data Yup. It happens to be the same….

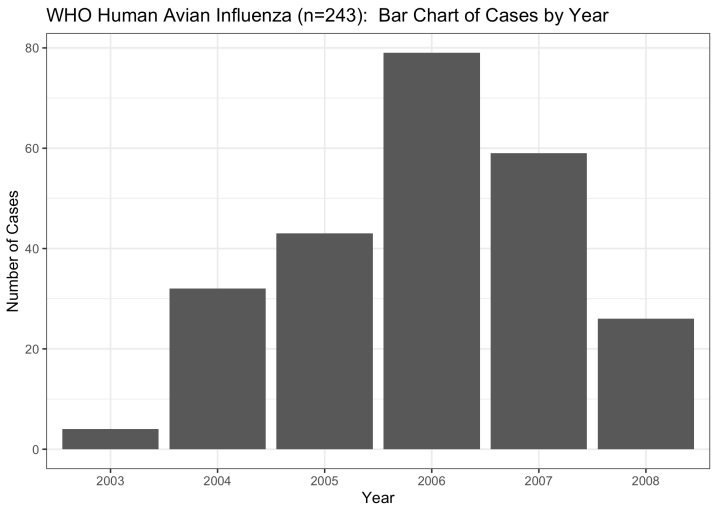
```
library(summarytools)
load(file="q1data.Rdata")
freq(q1data$year)
```

```
q1data$year
Type: Numeric

             Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
----------- ------ --------- -------------- --------- --------------
       2003      4      1.65           1.65      1.65           1.65
       2004     32     13.17          14.81     13.17          14.81
       2005     43     17.70          32.51     17.70          32.51
       2006     79     32.51          65.02     32.51          65.02
       2007     59     24.28          89.30     24.28          89.30
       2008     26     10.70         100.00     10.70         100.00
       <NA>      0                               0.00         100.00
      Total    243    100.00         100.00    100.00         100.00
```

__ 1b) Bar Graph

```
library(ggplot2)
ggplot(data=q1data,aes(x=factor(year))) +
     geom_bar(show.legend = FALSE) +
     ggtitle("WHO Human Avian Influenza (n=243):  Bar Chart of Cases by Year") +
     xlab("Year") + ylab("Number of Cases") +
     theme(legend.position = "none") +
     theme_bw()
```



__1c) Facts of Graph

This is a plot of the distribution of 243 cases of human avian influenza A/H5N1 recorded by the World Health Organization (WHO) over the six-year period 2003-2008. The plot shows that the annual number of cases ranged from a minimum of 4 in 2003 to a maximum of 79 in 2006. The annual numbers were: 4, 32, 43, 79, 59, and 26.

__1d. Interpretation

The 2003 number of confirmed cases, 4, is substantially lower than the other 5 years' records; the next higher annual number was 26 and was observed in 2008. Inspection of the annual numbers with advancing year shows that increases with year for the period 20013-2006 followed by a decline.

Question #2

A study examining the health risks of smoking measured the cholesterol (mg/dL) levels of people in two independent groups: 1) SMOKERS: those who had smoked for at least 25 years; and 2) NON-SMOKERS: persons of similar ages who had never smoked. The following are the data.

Load data.

```
load(file="cholesterol_smokers.Rdata")
load(file="cholesterol_nonsmokers.Rdata")
```

Check out the environment window …

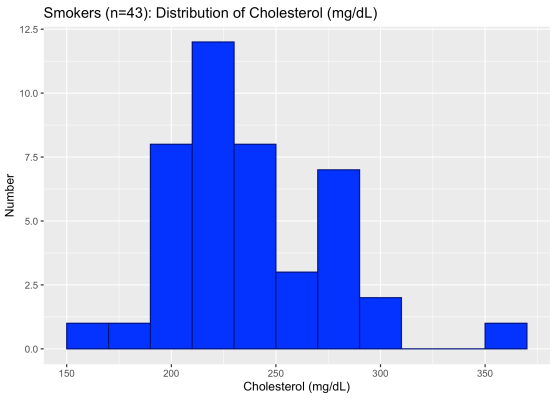
Name of file	Associated name of object (dataframe)
cholesterol_smokers.Rdata	q2smokers
cholesterol_nonsmokers.Rdata	q2nonsmokers

a) By any means you like, produce a **histogram** of cholesterol for **SMOKERS**
b) By any means you like, produce a **histogram** of cholesterol for **NON-SMOKERS**
c) Produce side-by-side histograms
d) In 1-2 sentences**, state the facts** of these histograms.
e) In 1-2 sentences, provide an **interpretation** of the comparison of these histograms

2a. Histogram for smokers

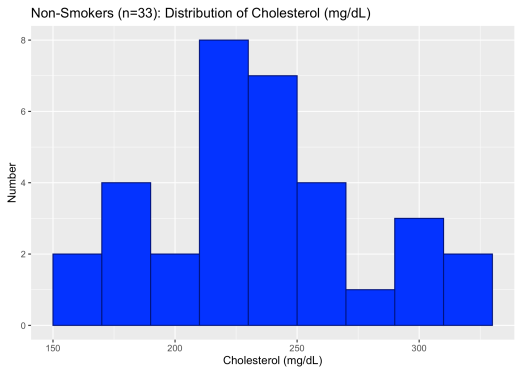
```
library(ggplot2)

ggplot(data=q2smokers, aes(x=chol)) +
  geom_histogram(binwidth=20, color="navy", fill="blue") +
  ggtitle("Smokers (n=43): Distribution of Cholesterol (mg/dL)") +
  labs(y="Number", x="Cholesterol (mg/dL)")
```



2b. Histogram for Non-smokers

```
load(file="cholesterol_nonsmokers.Rdata")
library(ggplot2)
ggplot(data=q2nonsmokers, aes(x=chol)) +
  geom_histogram(binwidth=20, color="navy", fill="blue") +
  ggtitle("Non-Smokers (n=33): Distribution of Cholesterol (mg/dL)") +
  labs(y="Number", x="Cholesterol (mg/dL)")
```



__ 2c. Side-by-side Histograms for smokers and nonsmokers

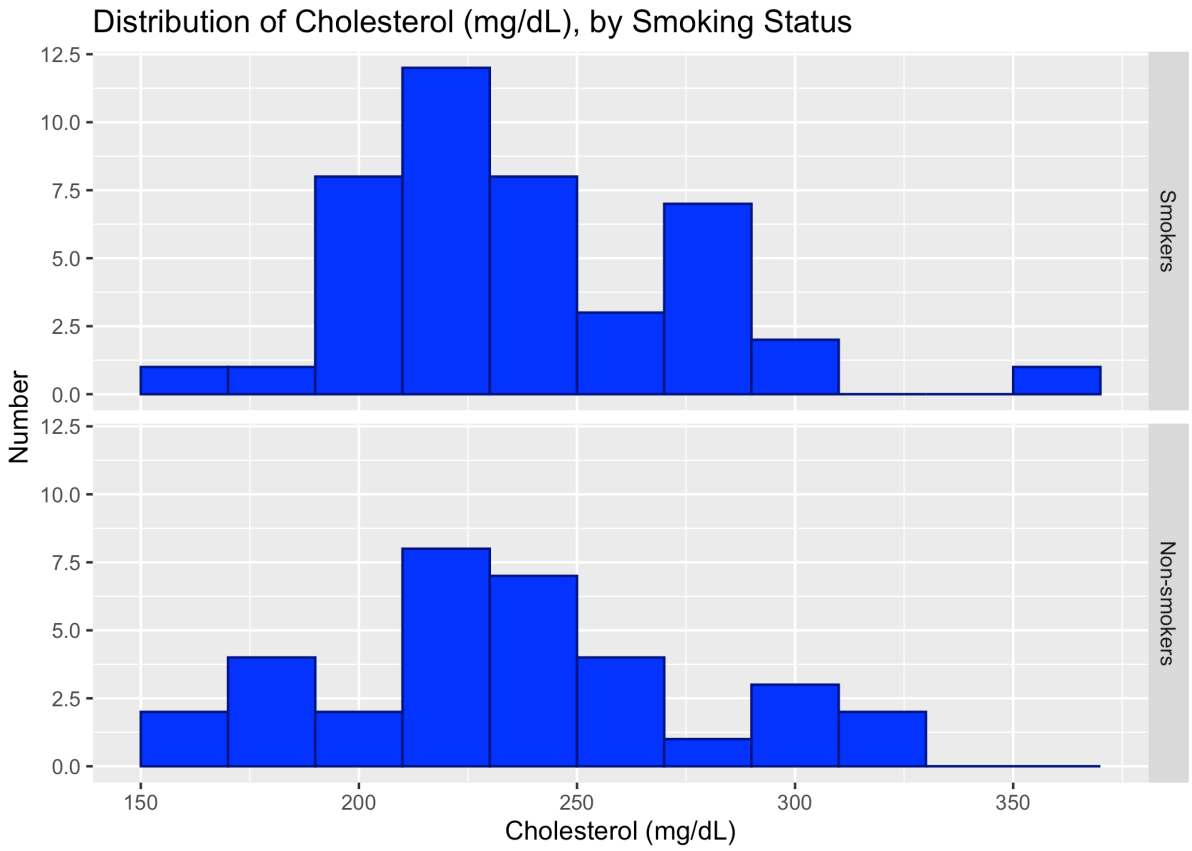
Again, check out your environment window …

Name of file	Associated name of object (dataframe)
cholesterol_540.Rdata	q2all

```
library(openxlsx)
library(ggplot2)
q2all <- read.xlsx("cholesterol_540.xlsx")
save(q2all, file="cholesterol_540.Rdata")

# Preliminary:  Label values of group so that your graph has a nice aesthetic
q2all$group <- factor(q2all$group,
                              levels = c(1,0),
                              labels = c("Smokers", "Non-smokers"))

# Side-by-side histogram
ggplot(data=q2all, aes(x=chol)) +
  geom_histogram(binwidth=20, color="navy", fill="blue") +
  facet_grid(group ~ .) +
  ggtitle("Distribution of Cholesterol (mg/dL), by Smoking Status") +
  labs(y="Number", x="Cholesterol (mg/dL)")
```



__ 2d. Facts of Graph

Cholesterol (mg/dL) was measured in n=43 smokers and n=33 non-smokers. Among smokers, cholesterol ranged from 155 mg/dL to 351 mg/dL, with a mean and standard deviation of 237.98 mg/dL and 38.54 mg/dL, respectively. Among non-smokers, cholesterol ranged from 160 mg/dL to 328 mg/dL, with a mean and standard deviation of 235.45 mg/dL and 44.43 mg/dL, respectively. In both groups, the mean and median values were similar, suggesting that the distributions are each symmetric around their central values. Inspection of the histograms confirms this.

__ 2e. Interpretation

These samples of cholesterol (mg/dL) among smokers and non-smokers, on the basis of simple descriptive statistics and side-by-side histograms, suggest that the distributions are similar and thus, not different depending on smoking status.

Question #3

Consider again the cholesterol (mg/dL) data in smokers and non-smokers introduced in question #2.
By any means you like, produce a **side-by-side stem and leaf** diagram.
a) By any means you like, complete the following table:

3a. Side-by-side Stem & Leaf Diagram for Smokers and Non-smokers

```
library(aplpack)
stem.leaf.backback(q2smokers$chol,q2nonsmokers$chol)
```

```
_____________________________________
  1 | 2: represents 12, leaf unit: 1
    q2smokers$chol      q2nonsmokers$chol
_____________________________________
   1          5| 15 |
               | 16 |03         2
               | 17 |45         4
   2           3| 18 |38         6
   3           6| 19 |2          7
  10     9990000| 20 |0          8
  18    77666631| 21 |333778    14
  21         555| 22 |78        16
  (4)       7220| 23 |8         (1)
  18        6633| 24 |22399     16
  14         860| 25 |07        11
  11           7| 26 |377        9
  10           1| 27 |1          6
   9      874000| 28 |
               | 29 |2          5
   3          95| 30 |0          4
               | 31 |0          3
               | 32 |18         2
               | 33 |
_____________________________________
HI: 351
n:            43      33
_____________________________________
```

3b. Complete the following table

```
library(summarytools)
paste("Smokers")
descr(q2smokers$chol, stats = c("n.valid","mean", "sd", "min","q1", "med", "q3", "max","iqr"))

paste("Non-Smokers")
descr(q2nonsmokers$chol, stats = c("n.valid","mean", "sd", "min","q1", "med", "q3", "max","iqr"))
(file="cholesterol_smokers.Rdata")
```

```
[1] "Smokers"
Descriptive Statistics
q2smokers$chol
N: 43

                  chol
------------- --------
      N.Valid    43.00
         Mean   237.98
      Std.Dev    38.54
          Min   155.00
           Q1   211.00
       Median   230.00
           Q3   267.00
          Max   351.00
          IQR    50.50
```

```
[1] "Non-Smokers"
Descriptive Statistics
q2nonsmokers$chol
N: 33

                  chol
------------- --------
      N.Valid    33.00
         Mean   235.45
      Std.Dev    44.43
          Min   160.00
           Q1   213.00
       Median   238.00
           Q3   263.00
          Max   328.00
          IQR    50.00
```

	Smokers	NON-Smokers
Number in group, n =	43	33
P_{25} = Q1 = Lower Quartile =	212	213
P_{50} = Q2 = Median Quartile =	230	238
P_{75} = Q3 = Upper Quartile =	262.50	263
Interquartile Range (IQR) =	50.50	50
1.5* IQR=	75.75	75
Value of Lower Fence =	155	160
Value of Upper Fence =	309	328
Outliers (if any) below lower fence (LIST) =	*none*	*none*
Outliers (if any) above upper fence (LIST) =	351	*none*

Dear Class – Some notes:
(1) Sometimes, the calculation of the lower (or upper fence) yields the actual minimum (or maximum), as occurred here.
(2) I calculated the 1.5*IQR by hand

Question #4

Consider again the cholesterol (mg/dL) data in smokers and non-smokers introduced in question #2. By any means you like, produce a **side-by-side boxplot**

4. Side-by-side Boxplot for Smokers and Non-smokers

```
library(ggplot2)

ggplot(data=q2all, aes(x=factor(group), y=chol)) +
    geom_boxplot(color="black", fill="blue") +
    labs(x=" ",y="Cholesterol (mg/dL)") +
    coord_flip() +
    ggtitle("Distribution of Cholesterol (mg/dL), by Smoking Status")
```