MTH 3240 Lab 1

Due Thu., Jan. 30

1 Part A: Graphing and Summarizing Data

1.1 Lightning Deaths Data Set

The following data represent numbers of deaths by lightning strikes in the U.S. for each of the years 1959 - 2005 (in time order), as compiled by the National Climatic Data Center from reports by the National Weather Service.

75,	48,	61,	48,	150,	49,	57,	39,	27,	51,	46,	50,	62,	51,	50,	58,
38,	34,	59,	44,	24,	39,	40,	33,	49,	33,	34,	32,	35,	30,	23,	39,
36,	25,	20,	32,	43,	52,	42,	44,	46,	51,	47,	51,	44,	33,	38	

1. Use the `c()` function and the assignment operator `<-` to create a vector containing these data.

2. Use `length()` to determine how many observations there are in the data set.

3. Use `sum()` to count the **total** number of deaths over the years 1959 - 2005.

4. Use the functions `min()` and `max()` to find the minimum and maximum values in the data set.

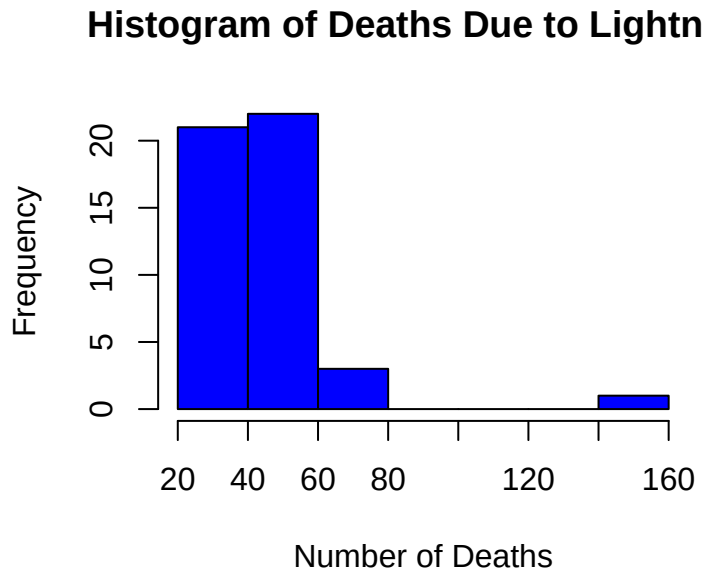
5. The function `hist()` will produce a histogram a data set. The main argument passed to `hist()` is a data vector `x`, but it accepts other optional arguments too. Among its other arguments are:

`x`	a data vector.
`col`	a color used to fill the histogram bars.
`xlab`	a label for the x-axis.
`ylab`	a label for the y-axis.
`main`	a main title.

   Make a histogram of the lightning deaths data, for example by typing something like this (assuming you named your data vector `deaths`):

```
hist(x = deaths, col = "blue", xlab = "Number of Deaths", ylab = "Frequency",
     main = "Histogram of Deaths Due to Lightning")
```

   Your histogram should look similar to the one below.



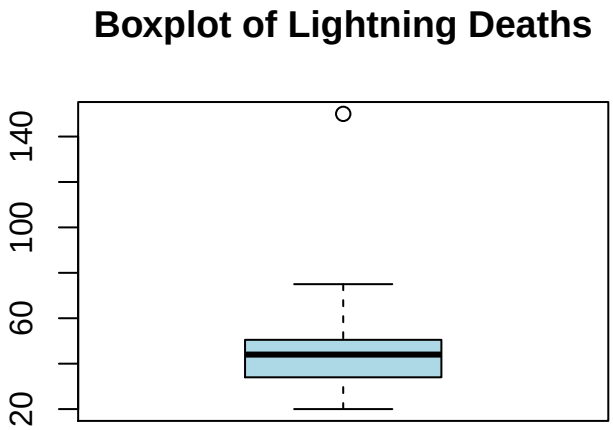
6. The function `boxplot()` will produce a boxplot of a data set. The main argument passed to `boxplot()` is a data vector `x`, but it accepts other optional arguments too. Among its arguments are:

`x`	a data vector.
`col`	a color used to fill the body of the box.
`main`	a main title.

Make a boxplot of the lightning deaths data, for example by typing something like this:

```
boxplot(x = deaths,  col = "lightblue", main = "Boxplot of Lightning Deaths")
```

Your plot should look something like the one below.



7. Notice the outlier in the histogram and boxplot. Use `mean()` and `median()` to compute the mean and median of the lightning deaths data set and compare their values.

8. Use `sort()` to print the data in sorted order, and compare the median to the sorted data set.

9. Use `sd()` to compute the standard deviation s of the data. How does the outlier affect the value of s?

10. R has a function called `plot()` that takes two main arguments, `x` and `y`, both vectors, and produces a scatterplot of the data. But `plot()` accepts other optional arguments too. Among its arguments are:

`x`	a data vector.
`y`	a data vector.
`type`	what type of plot should be drawn. Use `"l"` for "lines".
`xlab`	a label for the x-axis.
`ylab`	a label for the y-axis.
`main`	a main title.

We want to plot the **lightning deaths** over **time**. Either of the commands below will create a vector called `year` containing the years 1959 - 2005:

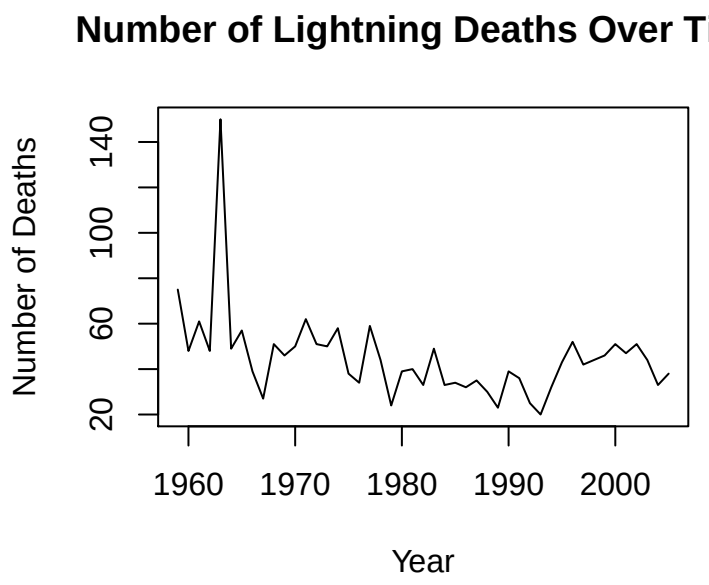
```
year <- seq(from = 1959, to = 2005, by = 1)
year <- 1959:2005
```

Create the vector `year` using one of the commands above.

Now plot the lightning deaths over time, connecting the points by lines, for example, by typing something like this:

```
plot(x = year, y = deaths, type = "l", xlab = "Year", ylab = "Number of Deaths",
     main = "Number of Lightning Deaths Over Time")
```

You should get something like this:



1.2 The Normal Distribution

A study suggests that **blood glucose levels** in *johnny darter* fish follow a **normal** distribution with mean **37.5** mg/100 ml and standard deviation **15.3** mg/100 ml.

1. We use the function `pnorm()` to compute the **probability** $\boldsymbol{P(X \leq x)}$ from a $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ distribution.

   The `pnorm()` function takes arguments:

`q`	the value x in $P(X \leq x)$.
`mean`	the mean μ
`sd`	the standard deviations σ

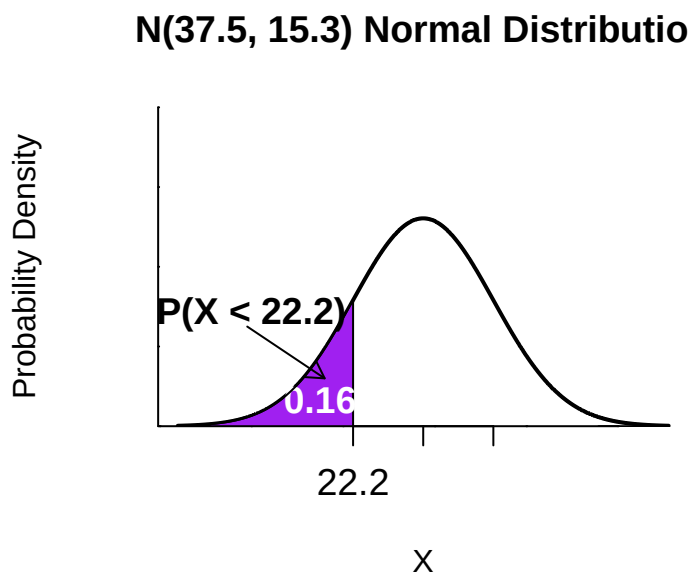
   It returns the value of $P(X \leq x)$.

   For example, to find the **probability** $P(X \leq 27.0)$ that the **glucose** level in a *johnny darter* fish will be **below 27.0**, type:

```
pnorm(q = 27.0, mean = 37.5, sd = 15.3)
```

```
## [1] 0.24627
```

Use `pnorm()` to compute $P(X \leq 22.2)$, the **probability** that the **glucose** level in a fish will be **below 22.2**.



2. To find the **probability** $P(22.2 \leq X \leq 52.8)$ that the **glucose** level in fish will be **between 22.2** and **52.8**, compute

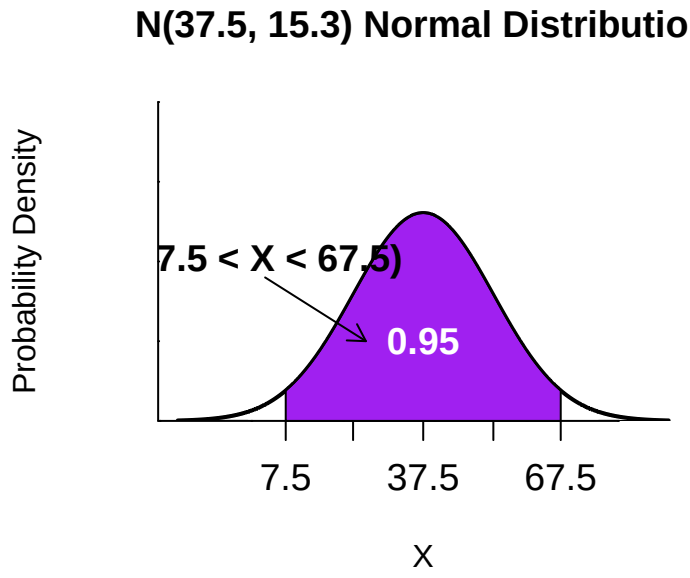
$$P(22.2 \leq X \leq 52.8) \;=\; P(X \leq 52.8) - P(X \leq 22.2)$$

by typing:

```
pnorm(q = 52.8, mean = 37.5, sd = 15.3) -  pnorm(q = 22.2, mean = 37.5, sd = 15.3)
```

```
## [1] 0.6826895
```

Now use `pnorm()` to compute $P(7.5 \leq X \leq 67.5)$.



2 Part C: Log Transformations for Right Skewed Data

2.1 Exhaust Hydrocarbons Data Set

The table below shows data on the amounts (in g/mi) of **hydrocarbon (HC) emissions** in the exhaust of $\boldsymbol{n = 46}$ randomly selected vehicles of the same type, measured under standard conditions prescribed by the U.S. Environmental Protection Agency.

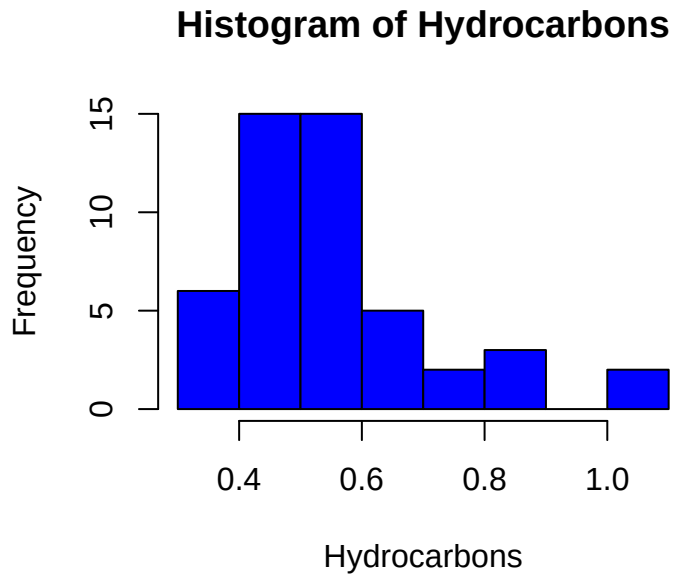
0.50,	0.65,	0.46,	0.41,	0.41,	0.39,	0.44,	0.55,	0.72,	0.64,	0.83,	0.38,	0.38,	0.50,
0.60,	0.73,	0.83,	0.57,	0.34,	0.41,	0.37,	1.02,	0.87,	1.10,	0.65,	0.43,	0.48,	0.41,
0.51,	0.41,	0.47,	0.52,	0.56,	0.70,	0.51,	0.52,	0.51,	0.52,	0.57,	0.51,	0.36,	0.48,
0.52,	0.61,	0.58,	0.46,	0.47,	0.55								

1. Use `c()` to create a vector containing the **HC emissions** data.

2. Recall that many statistical procedures rest on an assumption that either the sample was drawn from a **normal population** or $\boldsymbol{n}$ **is large**.

   If neither condition is met, the procedures may be invalid.

   A histogram a of the data (below) indicates that the sample is from a ***right skewed*** population.

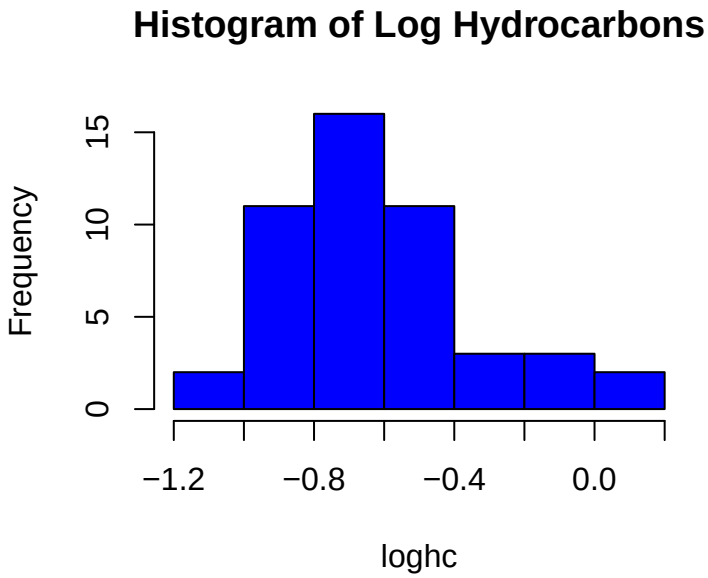


3. Create a new vector containing the *logs* of the HC emissions data, for example by typing:

```
loghc <- log(hc)
```

4. Use `hist()` to make a histogram of the *logs* of the HC emissions data:

```
hist(loghc, col = "blue", main = "Histogram of Log Hydrocarbons")
```


Histogram of Log Hydrocarbons
Frequency
0
5
10
15
−1.2
−0.8
−0.4
0.0
loghc