MATH 140 **Name** ______________________ **Date:**

R Activity: Descriptive Statistics

Example: Determine the median of these data: 3.2, 3.7, 5.3, 0.7, 4.6, 6.2, 7.3, 1.2, 2.4, 5.2

Answer. We want to use the `median()` function, and we also need to input the data with the `c()` function. So here's one solution:

```
median(c(3.2, 3.7, 5.3, 0.7, 4.6, 6.2, 7.3, 1.2, 2.4, 5.2))
```

```
## [1] 4.15
```

Alternatively, we could first give the data set a (descriptive) name, then use the name in the `median()` function. This approach is generally better because it makes it a lot easier to determine other statistics associated with the data. Now, truth be told, these data represent the number of miles I've walked in 10 consecutive days. I might then call the data set `dist`:

```
dist = c(3.2, 3.7, 5.3, 0.7, 4.6, 6.2, 7.3, 1.2, 2.4, 5.2)
median(dist)
```

```
## [1] 4.15
```

1. Basic Descriptive Statistics

In each case, use RStudio as in the example above to answer the question. Record your answers on this worksheet.

a. Find the standard deviation of the `dist` data set. Use `sd()` for standard deviation.

b. Find the five number summary of the `dist` data set. Use `fivenum()`.

c. Find the mean and standard deviation of the data below. **As a rule, don't type things from scratch into RStudio if it's possible to copy and paste!**
6.3, 5.7, 4.8, 9.9, 6, 10.3, 7.5, 8.1, 6.7, 8.7, 7.2, 10

2. Hank Aaron

a. How many seasons did Hank Aaron play? Record the value as well as which RStudio command you ran to find it.

b. How many home runs did Aaron hit in his career? Record the value as well as which RStudio command you ran to find it.

c. What is the maximum number of home runs Aaron hit in a single season? Record the value as well as which RStudio command you ran to find it.

d. Determine the five number summary for this distribution, and plot the corresponding box plot in RStudio. Based on this box plot, would you consider Aaron's distributions of home runs to be skewed right, skewed left, or symmetric? Explain briefly.

3. Cars Data

a. How many observations are in this data frame? How many variables?

b. What does the `am` variable tells us about a car? Is this variable categorical or numerical?

c. What is the average mpg for the 32 cars in this data frame?

d. What is the median horsepower (hp) of the cars in this data frame?

e. Note that the `cyl` variable records how many cylinders a car has. Run the code `table(df$cyl)`. What information does this provide about our data frame?

f. Describe the association between how many cylinders a car has and its full efficiency (as measured by mpg).

g. Describe the association between the weight (wt) of a car and its mpg.