Project 1 **Math 222**
Exploratory Data Analysis **Due Friday, February 1**

1. The file rainfall.csv contains monthly and annual precipitation totals for Farmville, VA from 1931 to 2011. Save a copy of this file in the same folder where you keep your R-markdown documents. Then use the command

```
rain = read.csv("rainfall.csv")
```

   to load the file into `R` as a data frame called `rain`.

   (a) Make a graph that displays the distribution of annual rainfall totals. Make sure to clearly label the axes of the graph. How would you describe the shape of the distribution?

   (b) Find the five number summary of the annual rainfall totals. Are there any years that would be considered outliers by the $1.5 \times$ IQR rule? Explain your findings.

   (c) One important question is whether the rainfall amounts are independent from year to year. One way to check is to make a scatterplot of rainfall totals each year against the rainfall totals the following year. Do this, and describe what you see. Do you think the graph is consistent with the assumption that precipitation amounts in consecutive years are independent? (Two helpful functions: the command `head(x,-1)` removes the last entry from a vector `x` and `tail(x,-1)` removes the first entry.)

   (d) Using the mean and standard deviation of the annual rainfall totals as parameters, we could make a normal distribution model to evaluate how unlikely extremely wet and dry years might be here in Farmville. 2018 was an extremely wet year. At Richmond International Airport, the total precipitation for 2018 was 63.7 inches. Assuming the total for Farmville was the same, what percent of years are as extreme as 2018 according to your normal distribution model? Do you think this is a reasonable estimate? Why or why not?

2. Over the last decade there has been an explosion in the number of confirmed planets orbiting other stars outside our solar system. The NASA Exoplanet Archive keeps a record of confirmed exoplanet data available online. The file exoplanets.csv contains data on all of the confirmed exoplanets that have been discovered to date. The variables in the file are:

   - `pl_hostname`: Host star name
   - `pl_name`: Planet name
   - `pl_discmethod`: Discovery method
   - `pl_pnum`: Number of planets in system
   - `pl_orbper`: Orbital period [days]
   - `pl_orbsmax`: Orbit semi-major axis radius [astronomical units (AU)])
   - `pl_orbeccen`: Orbit eccentricity

- `pl_bmassj`: Planet mass [Jupiter masses]
- `st_dist`: Distance from our solar system [parsecs]
- `pl_facility`: Discovery facility

Load this file into `R` as a data frame and use it to help answer the following questions.

(a) How many different methods have been used to discover exoplanets? What are they? Which methods have been used to discover the most exoplanents?

(b) How far away are most of the exoplanets that have been discovered? Make a histogram showing the distribution of distances.

(c) Which method has been the most effective for discovering far away exoplanets?

(d) Do some research on your own. How big is a parsec? How big is the Milky Way galaxy? How far away are the exoplanents that have been discovered relative to the size of the Milky Way galaxy?

(e) The estimated mass of each exoplanet is given by the `pl_bmassj` variable, which has units equal to Jupiter masses. What is the mass Earth measured in Jupiter masses? How many of the confirmed exoplanets are more massive than Earth? How many are less massive?

(f) If you plot a histogram of exoplanet masses, it isn't very interesting because the data is extremely right skewed. One common approach to making nicer graphs with right-skewed data is to graph the logarithm of the data. In `R`, the `log()` function computes the natural log of a numeric variable. Make a histogram of the logarithms of the masses of the exoplanets. Describe the distribution. Where would the Earth and Jupiter fall in this histogram?