Homework **Week 4** - Assignment

Data preparation

This assignment will use the `mpg` data from `ggplot2`. We will add two columns for use later in the analysis; script for adding these columns has already been added for you in Code Chunk 3. You should not need to modify this script. Note that `echo=FALSE` so you should not see this script in the Word file, but `results='verbatim'` so we should see the results of the call to `str()`.

Continuous data

Distribution

Make a graph

In Code Chunk 4, insert script from class this week to plot a histogram and density estimate for city mileage (`mpg$cty`). *Remember you will likely need to adjust* `binwidth=` *for the histogram.*

Interpret the graph

Below the code chunk, give answers to the following questions:

- what does the histogram represent, and how do you interpret it?
- What does the density estimate represent, and how do you interpret it?
- Make a couple comments on the distribution of these data.

Probability Density Function

Add PDF curve to the graph

In Code Chunk 5, insert script from class to overlay a PDF for a normal distribution of the `mpg` city mileage data. This graph should include three components: histogram, density estimate, and PDF.

Interpret the graph

Below the code chunk, give answers to the following questions:

- What does the new curve represent, and how do you interpret it?
- How well do you think the normal distribution models these data?

Graph a different continuous distribution

Recall that a proper normal distribution should be symetrical around the mean, with two nearly-equal tails on either side. What might be done to improve the symmetry of the distribution?

Transformation graph

In Code Chunk 6, modify script as we did in class this week to plot a histogram, density estimate, and PDF for transformated data that are more normally-distributed.

Interpretation

Describe how the transformation affected the distribution of the data. Why might this be of value?

Discrete data

Distribution

Discrete data can be handled in ways similar to continuous data (although we disscused exceptions in class). Use Code Chunk 7 to convert the variable of interest to a proportion of total counts per group.

Make a bar graph

In Code Chunk 8, insert script from class this week to plot a bar graph of counts per each category in `class`.

Add Probability Mass Function

In Code Chunk 9, insert script from class this week to overlay a PMF based on the negative binomial distribution using `prob=` as one of the arguments.

Interpret the graph

Below the code chunk, give answers to the following questions:

- How is the PMF similiar to the PDF? How is is different?
- In a simulated dataset based on this PMF, which class of vehicle do you expect would be most represented? Which would be the least? How do you figure?