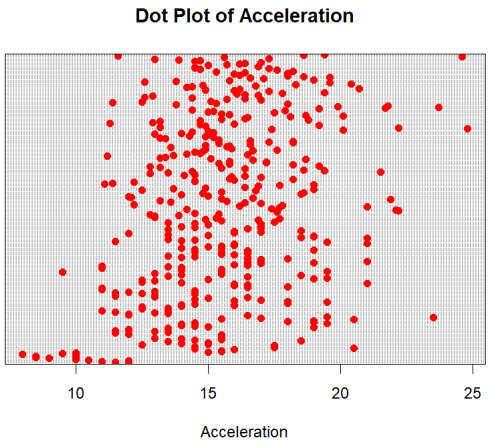
Lecture 15: Visualizations

For each question below, write the R code you would use and record the result when appropriate. Answer in complete sentences when asked to explain. We will use the `Auto` dataset from the `ISLR` package.

```
library(ISLR)
data(Auto)
```

Part 1: Dot Plot

1. Re-create the following dot plot for `acceleration`. Use `pch=16` to alter the point symbol and `col="red"` to alter the color:



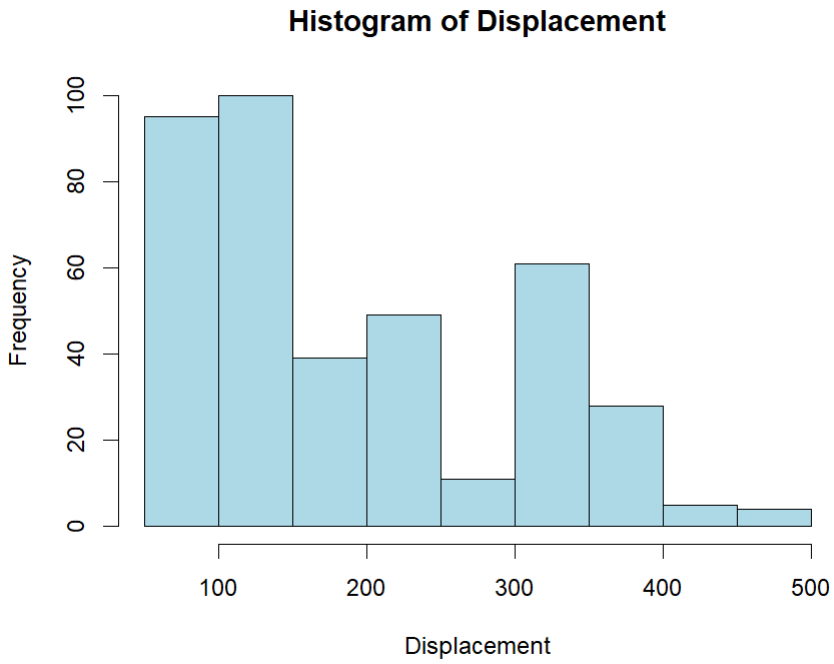
   (a) Where do most of the acceleration values seem to cluster?

   (b) Is this graph easy to read for this dataset? Explain briefly.

   (c) What is one advantage and one disadvantage of using a dot plot here?

Part 2: Histogram

2. Re-create the following histogram for `displacement`. Make sure that you include a title, an x-axis label, and color.



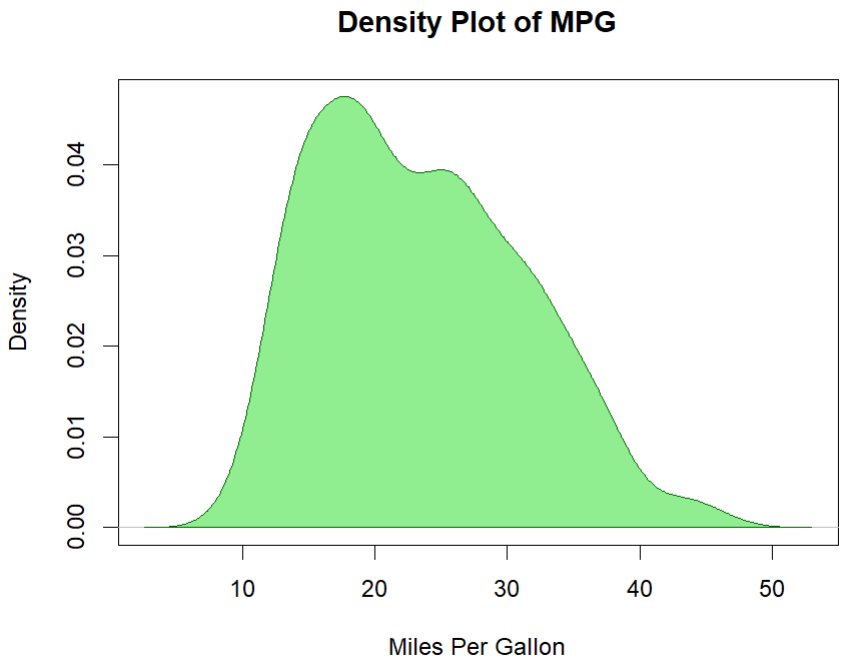
(a) Describe the shape of the distribution.

(b) Does the distribution look symmetric, skewed right, or skewed left? Explain briefly.

(c) What does the height of each bar represent? Would it look any different if we used density on the y-axis instead of frequency?

Part 3: Density Plot

3. Re-create the following density plot for `mpg`. Fill in the area under the curve using the `polygon()` function.



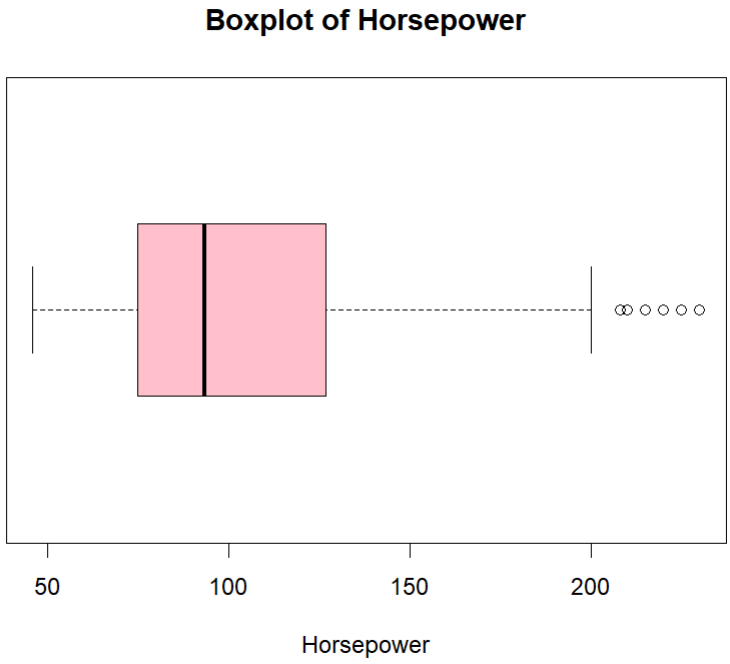
(a) What does the density curve help us see about `mpg`?

(b) Where do the `mpg` values seem to be most concentrated?

(c) How is a density plot similar to a histogram? How is it different?

Part 4: Boxplot

4. Re-create the following horizontal boxplot for `horsepower`.



(a) What does the line inside the box represent?

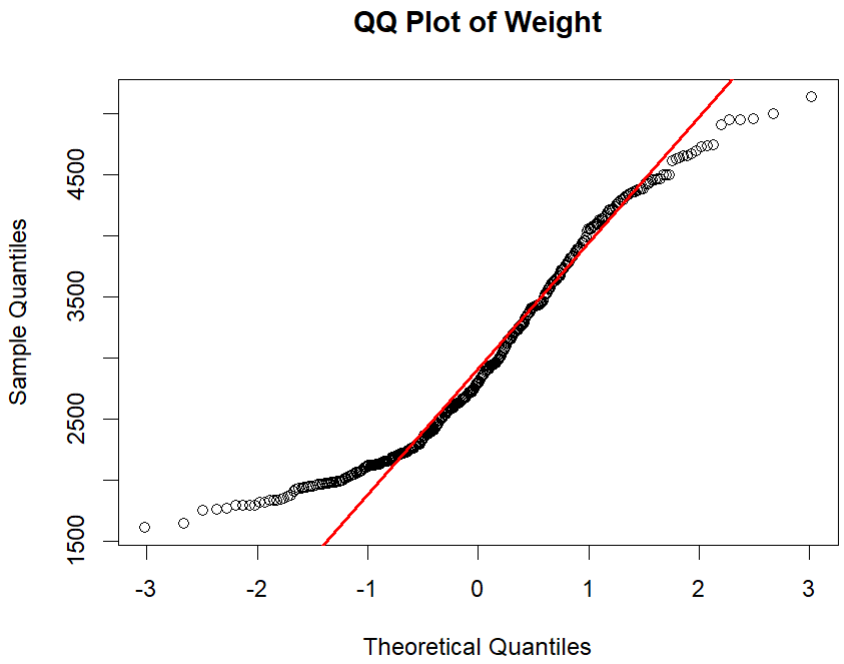
(b) Estimate the IQR

(c) Does the boxplot show any possible outliers?

(d) Based on the boxplot, does `horsepower` look fairly symmetric or skewed? Explain briefly.

Part 5: QQ Plot

5. Create a QQ plot for `weight`. Remember that you will need to use both the `qqnorm()` and `qqline()` functions.



(a) Do the points follow the line closely?

(b) What does this suggest about whether `weight` is approximately normal?

(c) What type of pattern would make you think the data are not normal?

Part 6: Tables and Pie Charts

6. Treat `cylinders` as a categorical variable. Re-create the following count table and a proportion table. To calculate the sum, use the `addmargins()` function. When making the proportion table, round all values to 3 digits by using the `round()` function.

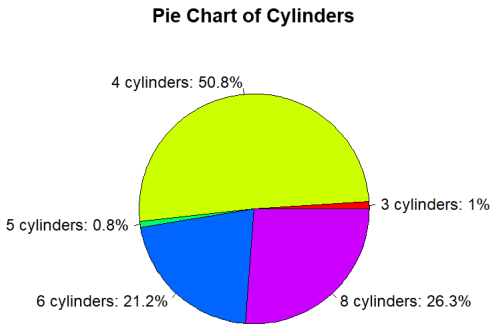
```
COUNT TABLE
3   4   5   6   8 Sum
4 199   3  83 103 392
```

```
PROPORTION TABLE
    3     4     5     6     8   Sum
0.010 0.508 0.008 0.212 0.263 1.001
```

   (a) What proportion of cars are in the most common cylinder category?

   (b) Why does it make sense to use counts and proportions instead of a mean for `cylinders` when treating it as categorical?

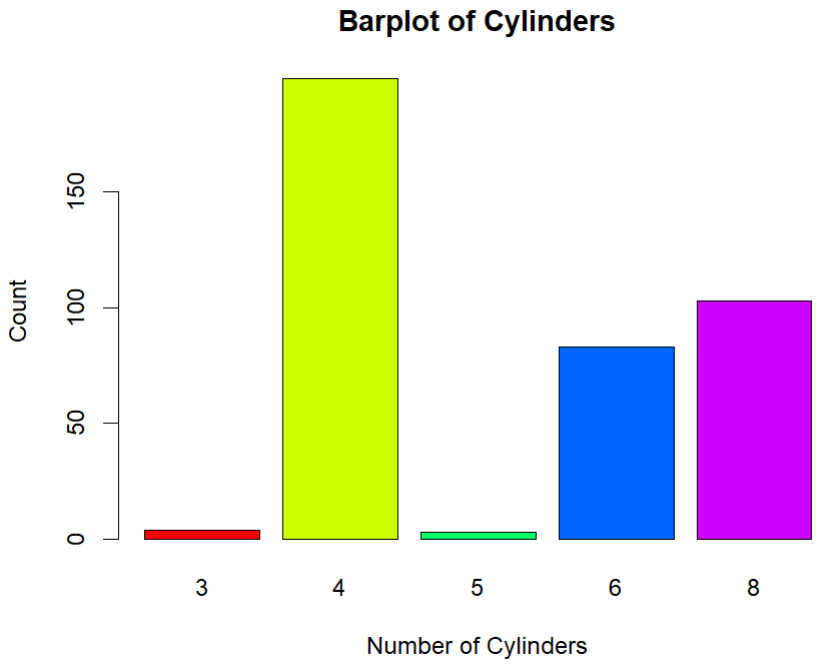
7. Re-create the following pie chart for the number of cylinders. Use the `paste0()` function to create the necessary labels and the `rainbow()` function to indicate the color.



   (a) Is the pie chart easy to interpret for this variable? Explain briefly.

Part 7: Barplot

8. Re-create the following barplot for the number of cylinders. Make sure that the label on the y-axis is correctly included.



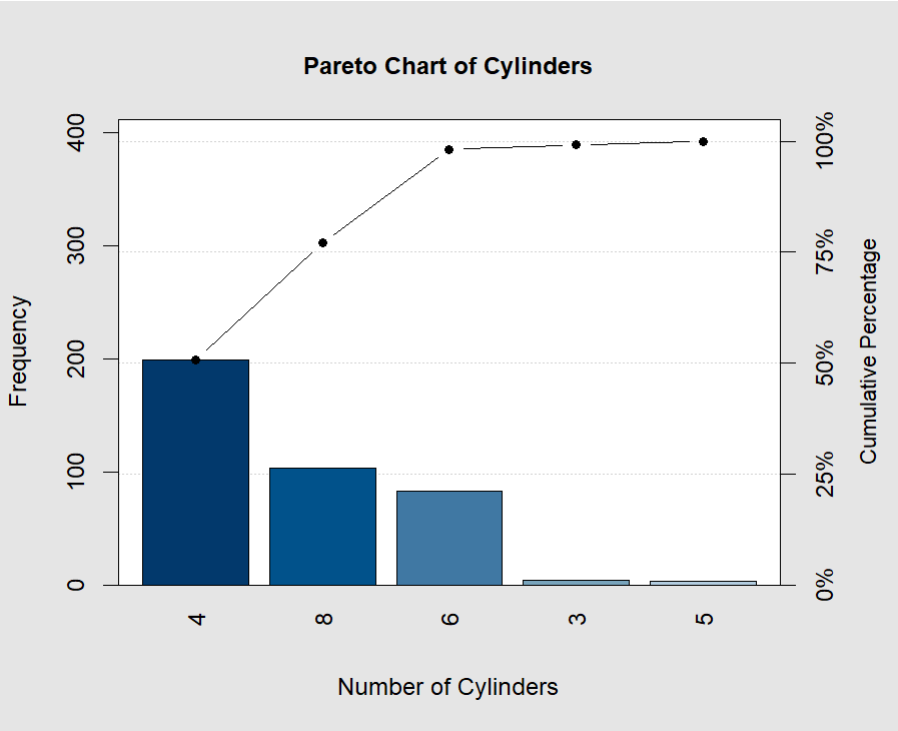
(a) What does the height of each bar represent?

(b) Which category has the tallest bar? How does this relate to the pie-chart that was previously made?

(c) Do you think the barplot is easier or harder to read than the pie chart? Explain briefly.

Part 8: Pareto Chart

9. Re-create the following Pareto chart for the number of cylinders. The `pareto.chart()` function is found within the `qcc` library.



(a) Which category appears first in the Pareto chart?

(b) How is the Pareto chart different from the regular barplot?

(c) Why might sorting the bars from largest to smallest be useful?