# Data Science with R

## Data Visualization Assignment

## Case Study 1:

Fine particulate matter (PM_{2.5}) is an ambient air pollutant for which there is strong evidence that it is harmful to human health. In the United States, the Environmental Protection Agency (EPA) is tasked with setting national ambient air quality standards for fine PM and for tracking the emissions of this pollutant into the atmosphere. Approximately every 3 years, the EPA releases its database on emissions of PM_{2.5}. This database is known as the National Emissions Inventory (NEI). You can read more information about the NEI at the [http://www.epa.gov/ttn/chief/eiinformation.html]

For each year and for each type of PM source, the NEI records how many tons of PM_{2.5} were emitted from that source over the course of the entire year. The data that we use for this assignment are for 1999, 2002, 2005, and 2008. The data is available - here

Goal of Case Study: The overall goal is to explore the National Emissions Inventory database and see what it say about fine particulate matter pollution in the United states over the 10-year period 1999-2008.

Ques1. Have total emissions from PM_{2.5} decreased in the United States from 1999 to 2008?

Ques2. Have total emissions from PM_{2.5} decreased in the Baltimore City, Maryland from 1999 to 2008?

Ques3. Of the four types of sources indicated by the =type= (point, nonpoint, onroad, nonroad) variable, which of these four sources have seen decreases in emissions from 1999-2008 for Baltimore City? Which have seen increases in emissions from 1999-2008?

Ques4. Across the United States, how have emissions from coal combustion-related sources changed from 1999-2008?

Ques5. How have emissions from motor vehicle sources changed from 1999-2008 in Baltimore City?

Ques6. Compare emissions from motor vehicle sources in Baltimore City with emissions from motor vehicle sources in Los Angeles County, California. Which city has seen greater changes over time in motor vehicle emissions?

Ques7. Answer the following questions given below:

Use the `bfs` data object from the `Beginning.RData` file

1. Look at the `warpbreaks` data that comes built into R. Create a box-whisker plot of the number of `breaks` for the different `tension`. Make the plot using horizontal bars and display the whiskers to the extreme range of the data. How can you draw this graph to display only a single type of wool?

2. The `trees` data come as part of R. The data is composed of three columns: the `Girth` of black cherry trees (in inches), the `Height` (in feet), and the `Volume` of wood (in cubic feet). How can you make a scatter plot of girth versus volume and display a line of best-fit? Modify the axes so that the intercept is shown clearly. Use an appropriate plotting symbol and colors to help make the chart more "interesting."

3. The `HairEyeColor` data are built into R. These data are in the form of a table, which has three dimensions. As well as the usual rows and columns, the table is split in two: `Male` and `Female`. Use the "males" table to create a Cleveland dot chart of the data. Use the mean values for the columns as an additional grouping summary result.

4. Look at the `HairEyeColor` data again. This time make a bar chart of the "female" data. Add a legend and make the colors reflect the different hair colors.

5. The `bfs` data object is part of the example data used in this book (you can download the entire data set from the companion website). Here you have two columns, butterfly abundance (`count`) and habitat (`site`). How can you draw a bar chart of the median butterfly abundance from the three sites?

## Ques8: Answer the following questions:

1. Use the `hogl` data object from the `Beginning.RData` file for this activity. Create a bar chart of the mean values for the two samples. Alter the aspect ratio of the plot to produce a graph 4 inches wide and 7 inches tall. Now add error bars to show the standard error using the `arrows()` command.

2. Use the `hoglouse` data object from the `Beginning.RData` file for this activity. Make a bar chart of these data. Use blocks of bars for each sample (that is, not stacked) and use a palette of rainbow colors. Include a legend; there will be room at the top left of the plot.

3. Examine the `hoglouse` data again. This time make a horizontal bar chart using stacked bars to highlight differences between *fast* and *slow*. You will find that there is no room for the category labels, so make the margin a bit wider. Add a legend to your plot separately and use a mouse-click to position it in an appropriate location.

4. Look at the `mf` data, which you've seen previously. Make a scatter plot using the `matplot()` command that shows the `Length` against both `Speed` and `Algae` variables. You should be able to use different colors and plotting symbols for each series, and add a legend (it will fit nicely at the bottom right of the plot). Add appropriate axis titles and include a subscript to indicate that `Length` was measured in mm.