Exploratory Data Analysis (EDA) is a short name given to what we have been doing in class: describing the shape, center, and spread of a data set. This project shows how to use R to do these things.

First we need to start up R and load some data sets. Here are the steps:

1. Start R by double-clicking on the desktop icon
2. At the R command line type the command

   ```
   library(pmg)
   ```

3. While waiting for the pmg GUI to start, minimize the main R window. It just causes confusion.
4. Under the `File` menu is an entry `Load package...`. Select this, and then load the `MASS` package.
5. Next, under the `Data` menu is an entry `Load data set...`. Open this and load the following packages (double click on the entry): `Aids2`, `Animals`, `Cars93`, `USCereal`, `galaxies`, `geyser` and `Michelson`.

These data sets should appear in the variable browser on the left.

1 Boxplots

We learned how to use the lattice explorer to make various graphs, we see now how to make boxplots too.

Open the lattice explorer which is found under the `Plots` menu. Drag the `galaxies` data set onto the explorer. By default a density plot is drawn.

Question 1.1 *Describe the data set* ***`galaxies`*** *as unimodal, bimodal or multimodal.*

Change the plot type from `densityplot` to `bwplot`. A boxplot of the data should appear.

Question 1.2 *Are there any outliers identified? If so, what is the smallest? What is the largest?*

Question 1.3 *Identify the following: the minimum, Q_1, the median, Q_3 and the maximum.*

The `michelson` data set contains measurements made regarding the speed of light. (The Michelson-Morley experiment of 1887 discounted the theory of aether.) We see that experiments produce randomness due to several factors.

Now clear the lattice explorer, and expand the values for `michelson` in the variable browser. You should see `Speed` and `Expt` (additionally `Run`. **First** drag `Speed` to the lattice explorer and then `Expt`. (You did clear it first?). Change the plot to `bwplot`. You should now have 5 different boxplots drawn.

Question 1.4 *Which experiment has the biggest median?*

Question 1.5 *Which experiment had the smallest recorded value? The largest?*

Question 1.6 *Which experiment had the smallest IQR? The largest?*

Question 1.7 *Which experiment(s) had outliers?*

Question 1.8 *Which experiment had the smallest range, the largest? Does this suggest that the experimenters got better as they progressed with the experiment?*

Now clear the lattice explorer and drag the `waiting` variable from the `geyser` data set onto the lattice explorer.

Question 1.9 *Is the data set unimodal, bimodal, or multimodal?*

Question 1.10 *Switch to a boxplot. Why can't you answer the previous question with a boxplot?*

Question 1.11 *From the boxplot, does the data set look symmetric?*

2 Numeric summaries

The `Dynamic Summaries` dialog, under the `Data` menu, allows the easy computation of the mean, median, IQR and standard deviation. These are our numeric summaries of a data set. Open the dialog, but keep open the lattice explorer.

2.1 The center

The center of a data set is usually described by the mean or the median.

Inside the `UScereal` data set is a variable `calories`.

Question 2.1 *Make a density plot of the* ***`calories`*** *variable and describe its shape. (modes? Symmetric, skewed or neither?)*

Question 2.2 *From the graphic you can* ***guess*** *the mean and median. How?*

- *The mean is the balance point for the graph*
- *The median splits the area in half.*

Guess both the mean and median for this variable.

Question 2.3 *Now drag the* ***`calories`*** *variable over to the* ***`Dynamic summaries`*** *dialog and drop it at the top in the spot labeled* ***`x:`****. Switch the "select summary:" popup to* ***`mean`****. What is the value? Was your guess close? Repeat with the median.*

Question 2.4 *For a skewed right data set the mean is usually more than the median. Is this the case with this data set?*

Now drag the `vitamins` variable onto the lattice explorer. This creates different graphs for each type of vitamin fortification.

Question 2.5 *Are the shapes basically the same, or are they different?*

Question 2.6 *Which type of vitamin seems to have the most data?*

Now drag the `vitamins` variable onto the "[group by]:" area of the Dynamic Summaries dialog. Which type has the largest mean? The largest median?

Question 2.7 *In the* ***`Animals`*** *data set is the incredibly skewed* ***`brain`*** *variable. What are the mean and median of this variable?*

3 Spread

Spread is measured by the IQR and the standard deviation, s. These are both options when you select a summary, although the standard deviation is referred to by `sd`.

Question 3.1 *For bell shaped data sets, the IQR is usually about 25% more than standard deviation. For skewed ones this can be much different. What are the values of the IQR and s for the* `brain` *variable?*

Question 3.2 *Look at the* `age` *variable in the* `Aids2` *data set. Compare the difference between the IQR and s. This should be consistent with a bell-shaped data set. Now make a densityplot of the data and describe it.*

Now drag the `sex` variable over to the "[group by]:" area.

Question 3.3 *Is there a difference in the spread between the two genders? Is there a difference in the mean? Make a boxplot and discuss how it relates to your answers.*