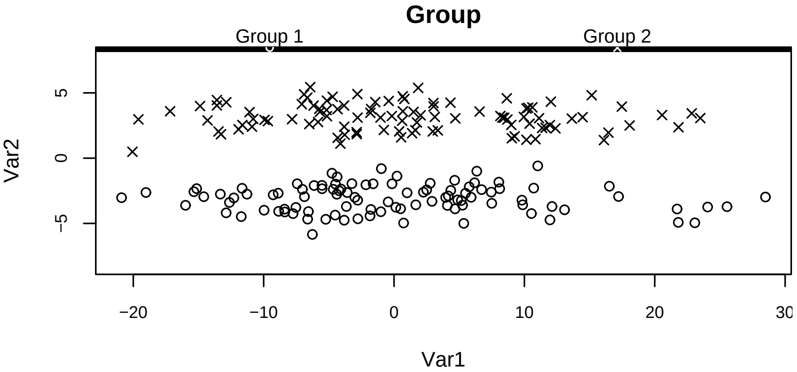
36-350: Data Mining

Homework 6
Date: October 2, 2001 **Due: start of class October 7, 2001**

1. Below is a plot of two-dimensional data on two variables, `Var1` and `Var2`.



   (a) Which variable has the most variance?

   (b) Which variable has the least overlap between the groups?

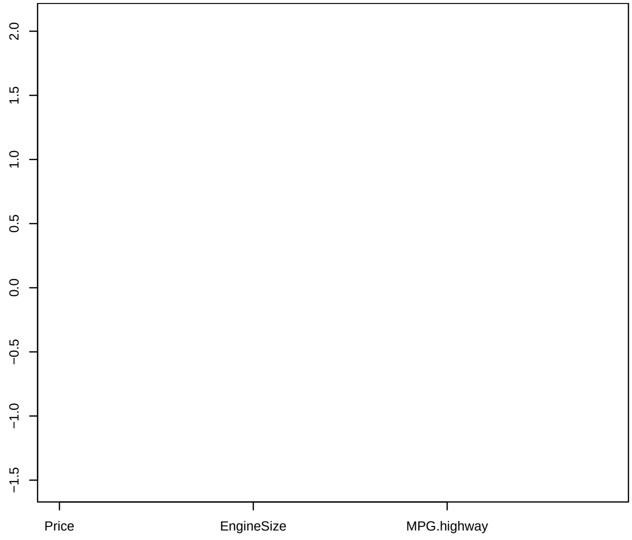
   (c) Suppose you wanted to project the data onto one dimension. What variable would PCA projection emphasize?

   (d) What variable would m-projection emphasize?

2. The following table describes a subset of the (standardized) car data:

	Price	MPG.highway	EngineSize
Toyota Previa	0.6	-1.5	-0.1
Buick Century	-0.2	0.5	-0.3
Pontiac LeMans	-1.5	2.1	-1.2

On the axes below, draw a parallel-coordinate plot of this data.



3. For each task below, state whether a projection scatterplot or parallel-coordinate plot would be more appropriate in general.

   (a) Describing why an individual is closer to the prototype for one subgroup than the prototype for another.

   (b) Describing how one individual differs from the rest of the population.

   (c) Finding individuals in a subgroup which do not belong.

   (d) Describing how several cluster prototypes differ.

4. In the computer lab, you made a parallel-coordinate plot of the State cluster prototypes.

   (a) In the plot, the variables have been scaled and shifted so that the prototypes approximately form lines, with two sloping up and two sloping down. One of the variables does not agree very well with this pattern. Which is it?

   (b) The variables `Homocide` and `Illiteracy` have been flipped, so that "up" means "low homocide/illiteracy" and "down" means "high homocide/illiteracy". Why do you suppose this was necessary? (Hint: look at the variables next to them.)

5. In the clustering from Ward's method, Texas is placed in the "Northeast" cluster, even though Texas is usually considered a Southern state.

   (a) Using your parallel-coordinate plot showing Texas, on which variables is Texas like a "South" state?

   (b) Using your parallel-coordinate plot showing Texas, on which variables is Texas like a "Northeast" state?

   (c) Describe how the plots from lab, i.e. the parallel-coordinate plot and your projection scatterplots, show that Texas should be in the "South" cluster instead of "Northeast".

6. In the clustering from Ward's method, Virgina is placed in the "Northeast" cluster, even though Virigina is usually considered a Southern state. Using your plots from lab, argue that Virigina is properly placed in "Northeast".

7. In the clustering from Ward's method, New Mexico is placed in the "South" cluster. Using your plots from lab, argue that New Mexico does not fit well in "South", or in any of the other clusters.