36-350: Data Mining

Handout 10
September 29, 2003

Visualizing subgroups

To look deeper into a dataset, it helps to divide it into **subgroups**. Some useful questions to ask:

- What is the shape of each subgroup? Does it have trends that differ from the dataset as a whole?
- How are the subgroups arranged? What distiniguishes the objects in one subgroup from another?
- Which objects are typical of a subgroup, which are on the border, and which don't belong?

Informative projection

Recall that, given a group variable c, an **informative dimension** is one which separates the groups. More generally, an **informative projection** is a projection which separates the groups. PCA is a special case, where each object is in its own group (the projection tries to separate all points).

Mathematically:

$$\begin{aligned}\mathcal{I}(c,h) &= \mathcal{H}(h) - \mathcal{H}(h \mid c) \\ &= \mathcal{H}(h) - \sum_{\mathsf{c}} p(c = \mathsf{c})\mathcal{H}(h \mid c = \mathsf{c})\end{aligned}$$

The **mv-projection** is the most informative projection under the assumption of **normality**. It is a function only of the **m**ean and **v**ariance of each group.

Entropy of a normal distribution:

$$\mathcal{H}(x) = \frac{1}{2}\log \operatorname{var}(x) + \frac{1}{2}(1 + \log(2\pi))$$

Thus $h = \mathbf{X}\mathbf{w}$ maximizes

$$2\mathcal{I}(c,h) = \log \operatorname{var}(h) - \sum_{\mathsf{c}} p(c = \mathsf{c}) \log \operatorname{var}(h \mid c = \mathsf{c})$$

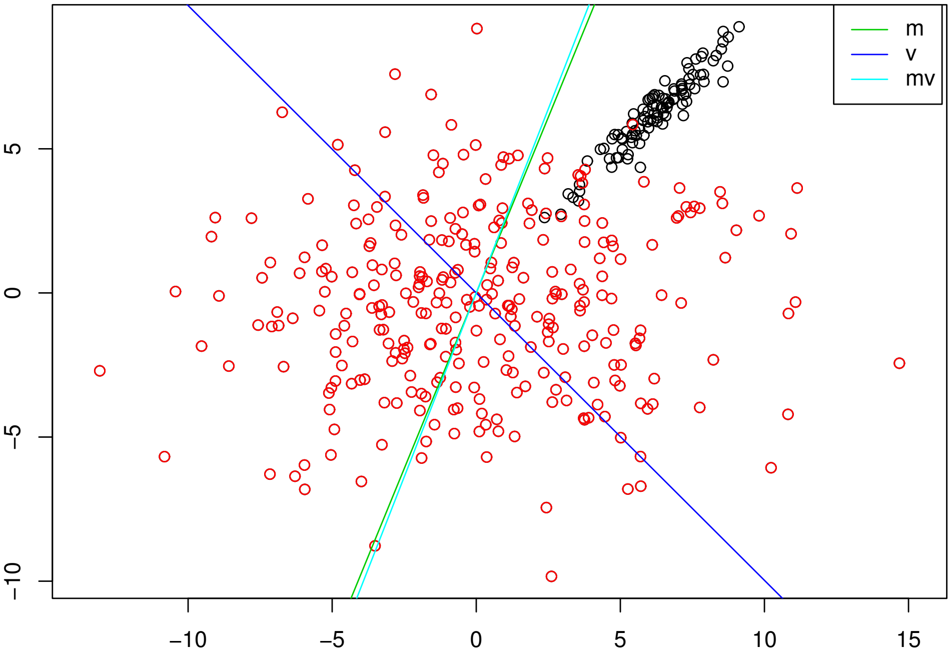
The first term wants the projected data to be spread out, the second wants the data within each group to be compact. The result is to separate the groups.

Recall that the variance of h is determined by the covariance matrix of X. If there are d dimensions, then it takes about $10d$ points to get an accurate estimate of $\mathrm{cov}(X)$. If the groups are smaller than this, the noisy covariance estimates can lead to a bad choice of projection.

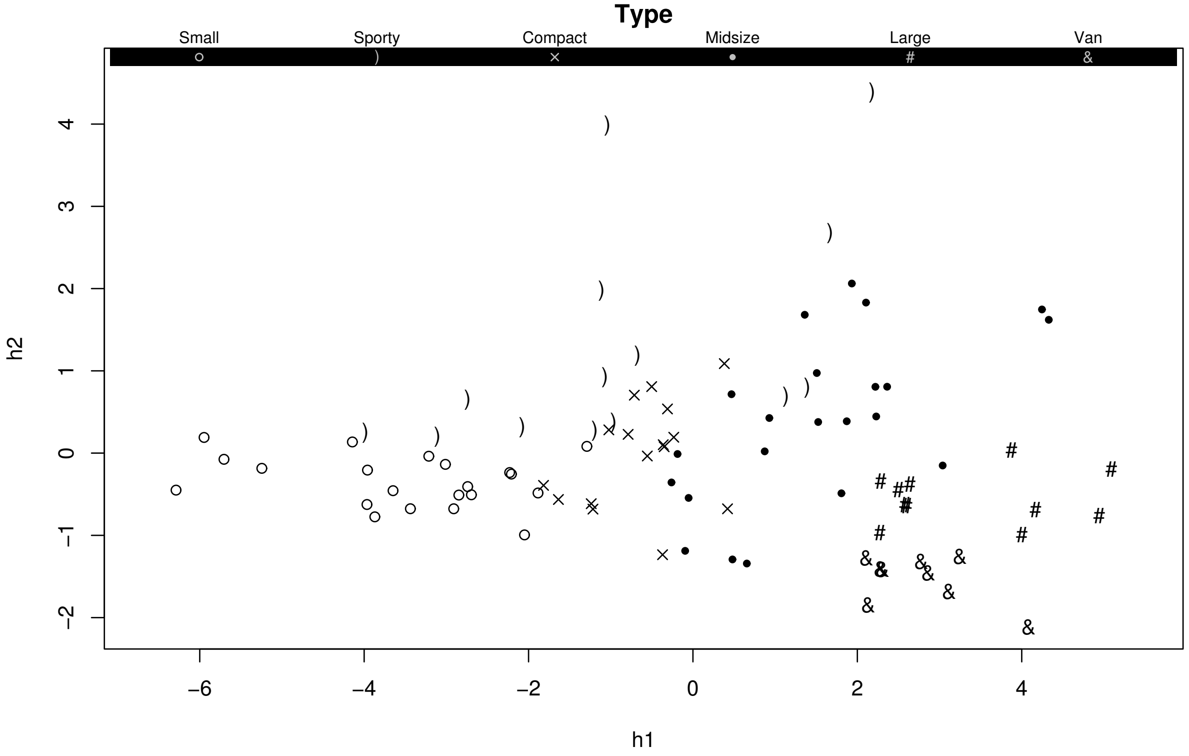
The **m-projection** is the most informative projection assuming that all groups have the *same covariance matrix*. The common covariance matrix is estimated using all points, so it can handle small groups.

The **v-projection** is the most informative projection assuming that all groups have *the same mean*. In other words, it tries to distinguish the groups based on **shape**. Differences in shape can show you attributes that vary more in one class than in others, or trends that exist in one group but not others.

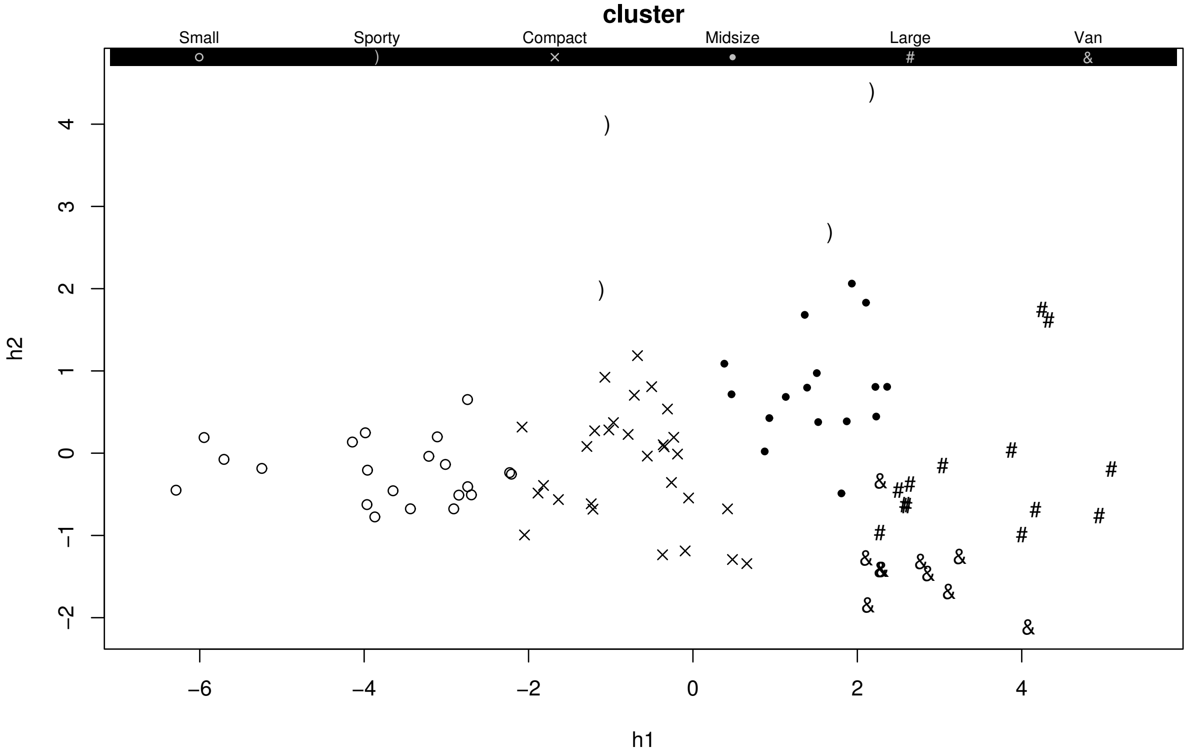
An example where v-projection is different from m-projection:



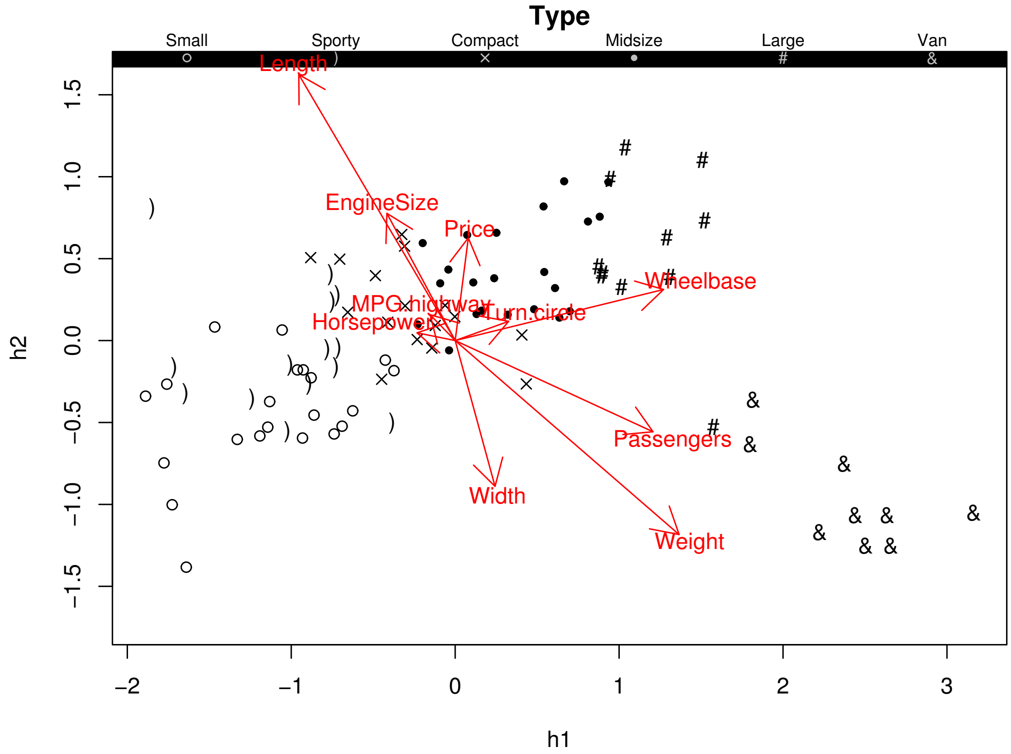
PCA projection of cars, coded by Type: (SS = 313)



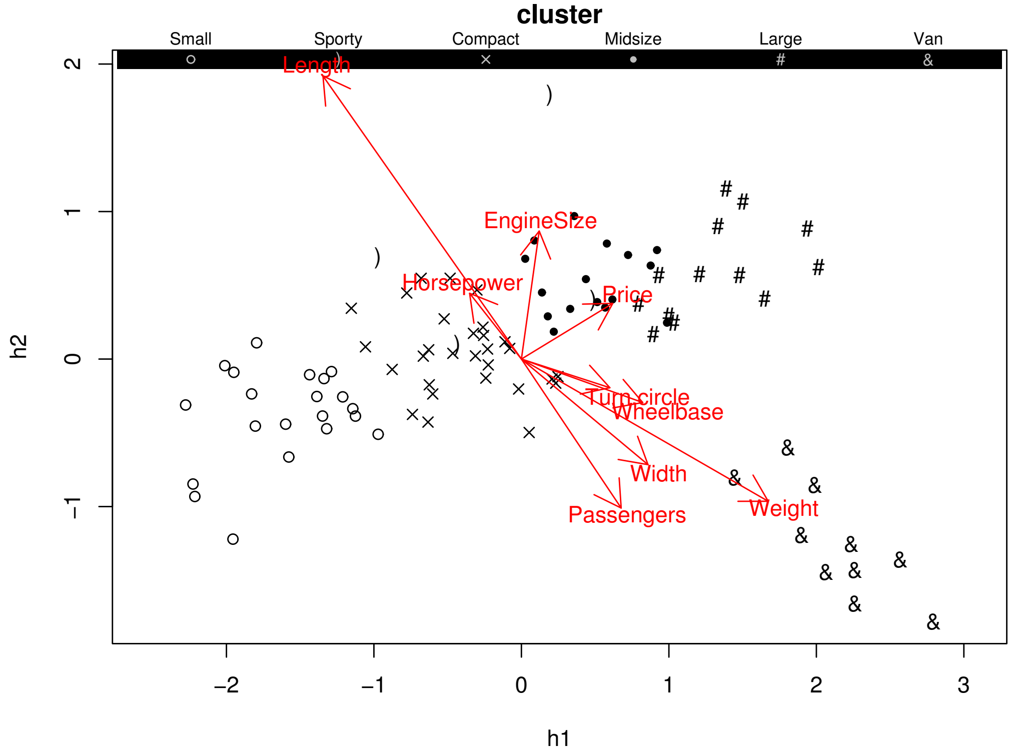
PCA projection, coded by K-means clustering: (SS = 233)



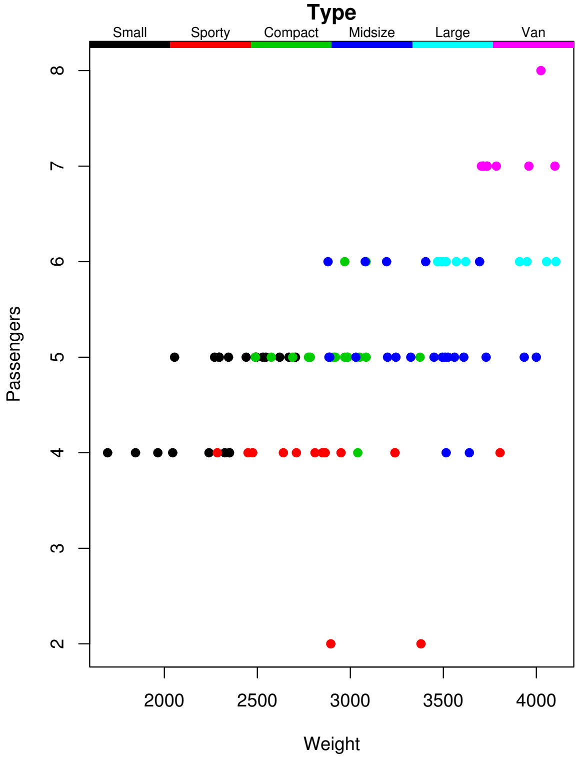
m-projection of Types:



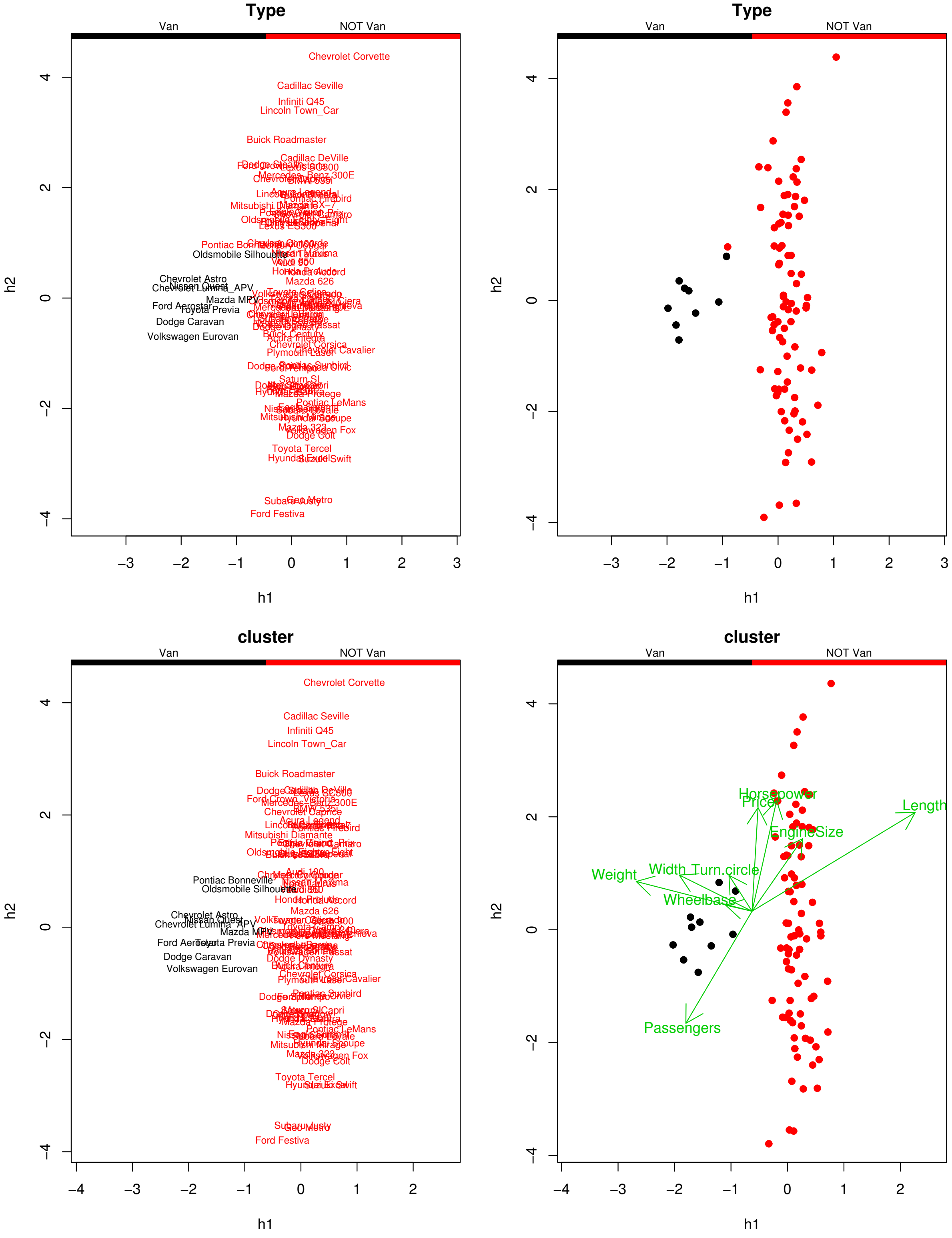
m-projection of clusters:



The two most informative attributes (about Type or cluster):

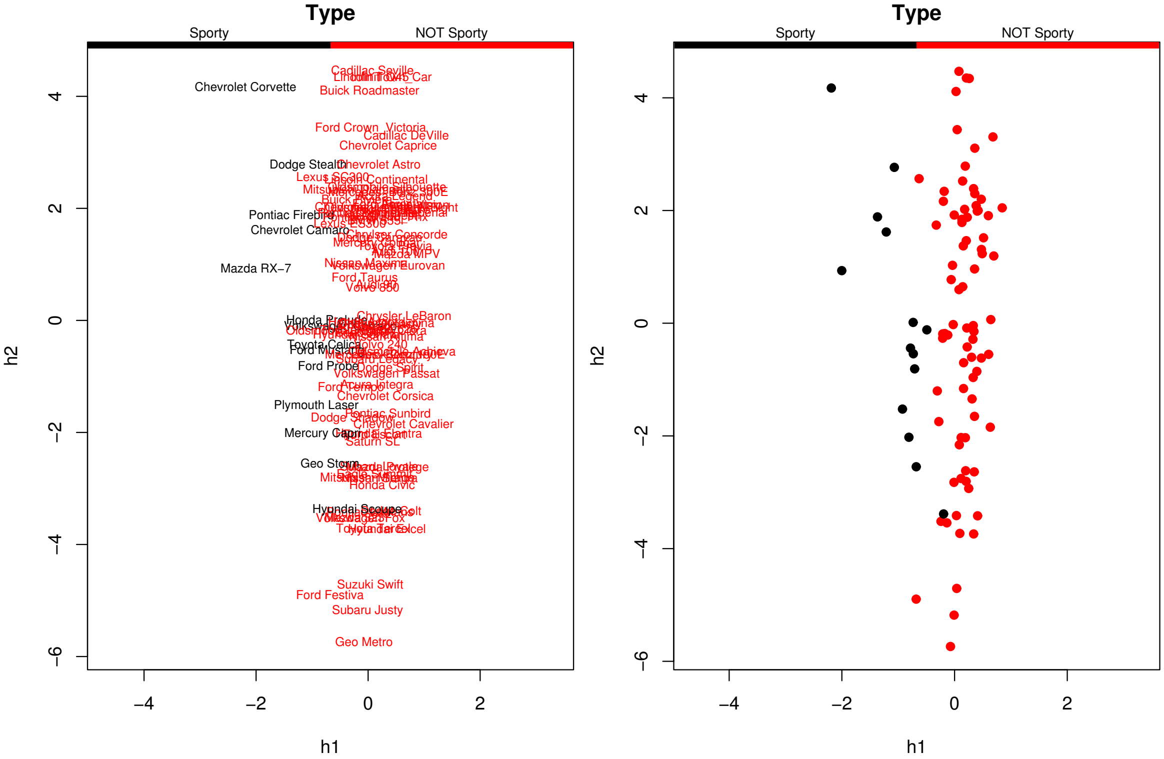


Separating Vans from non-Vans:



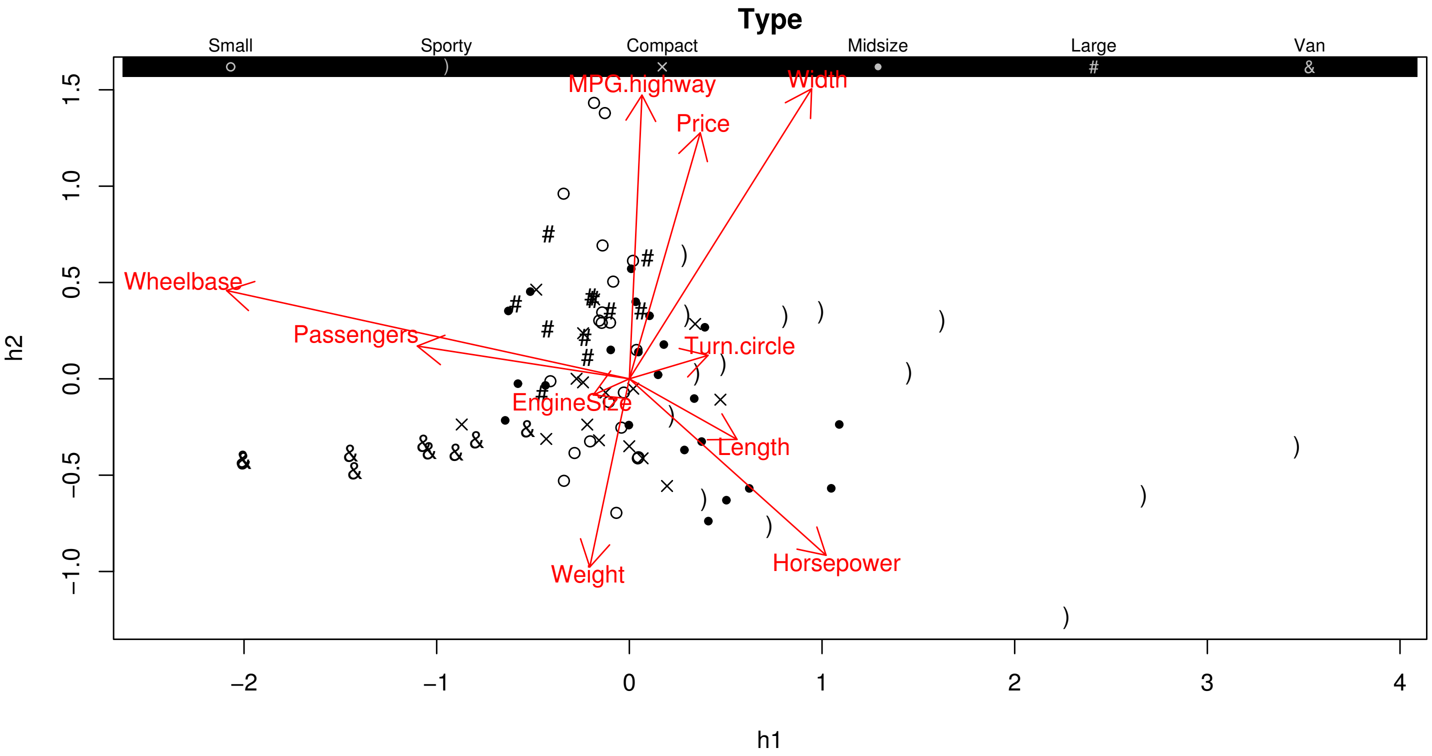
Is the Bonneville actually a Van?

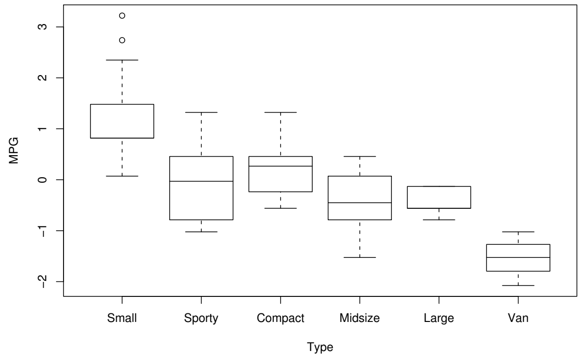
Some cars are more “Sporty” than others:

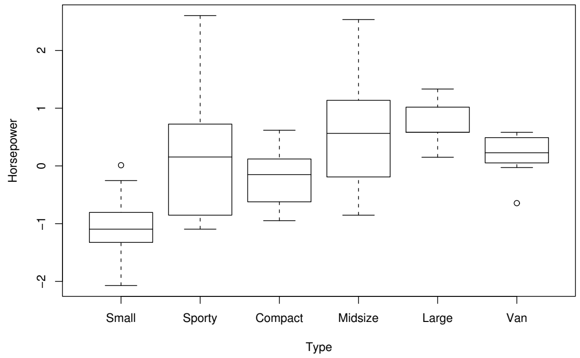


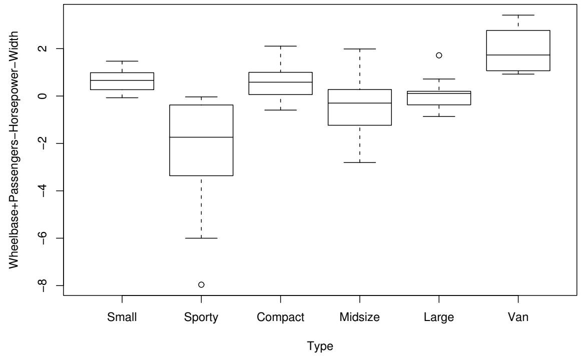
Hyundai Scoupe is not very “Sporty.”

v-projection of Types:









Shows you how the groups differ in terms of variance.