36-350: Data Mining

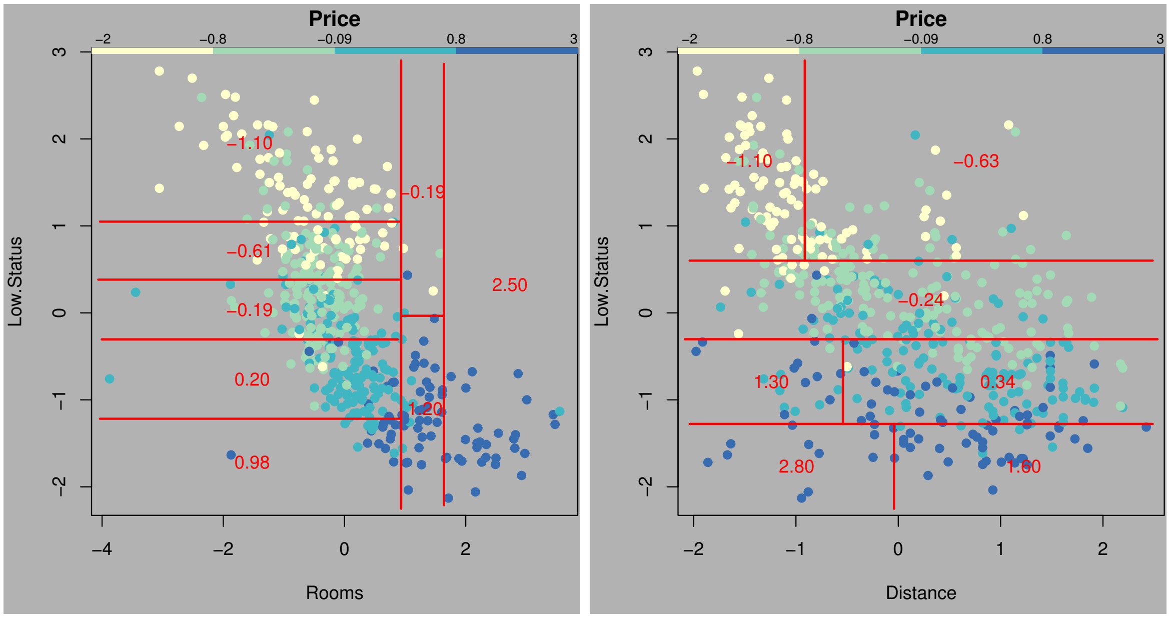
Handout 15
October 15, 2003

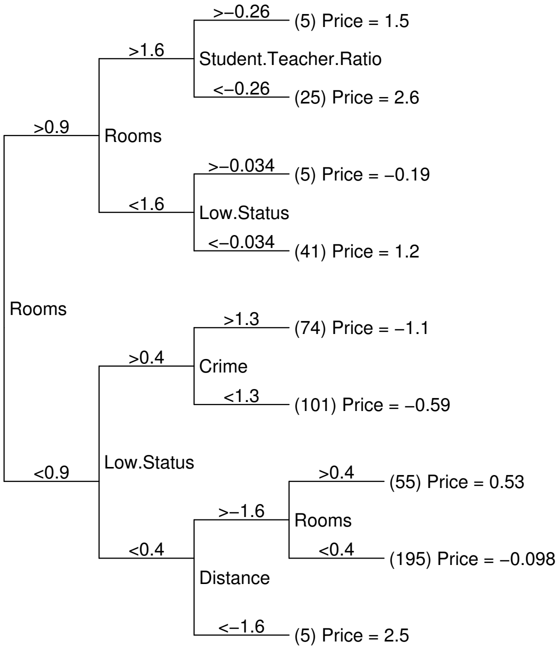
More uses of trees

What split should be at the top of the tree?

Predictor1	Predictor2	Response
0.3	1.3	49
1.4	1.6	62
0.1	0.2	46
1.2	0.4	57

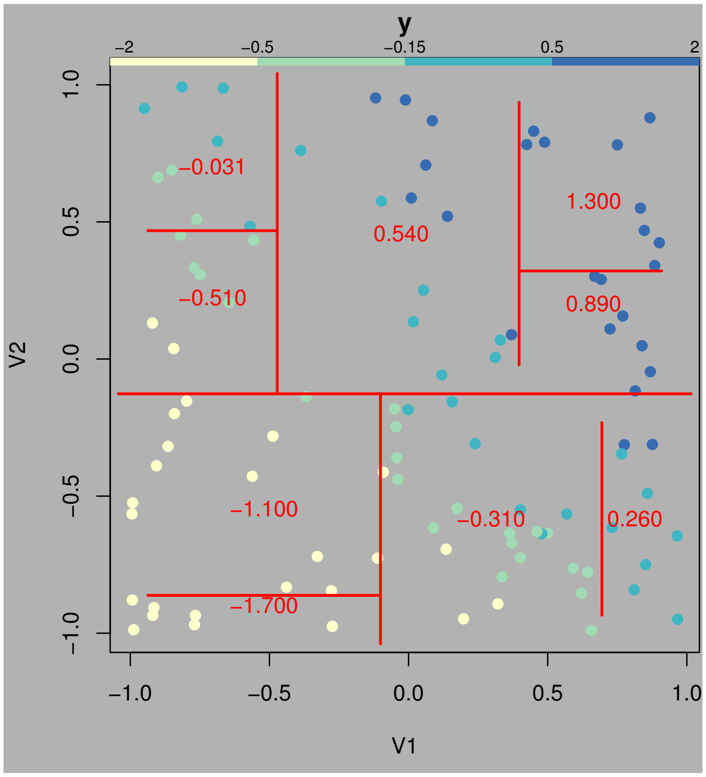
Interactions in the Housing data:

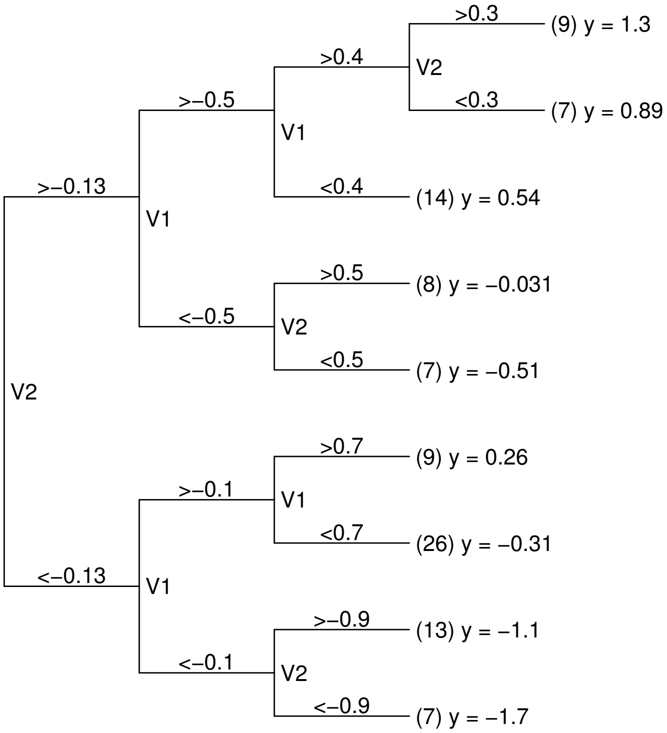




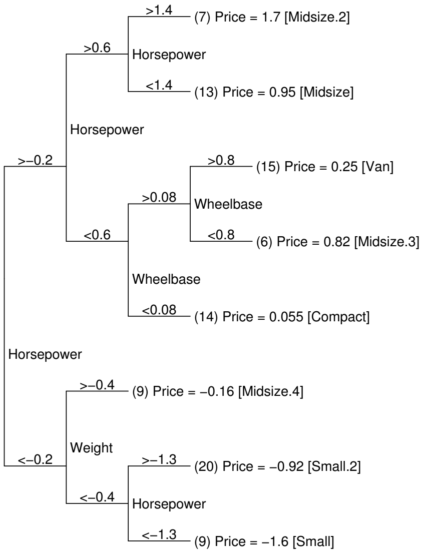
This tree has $R^2 = 0.84$, compared to 0.76 for linear regression. It correctly captures the many interactions in this data, for example the way that Low.Status changes the importance of Crime and Distance.

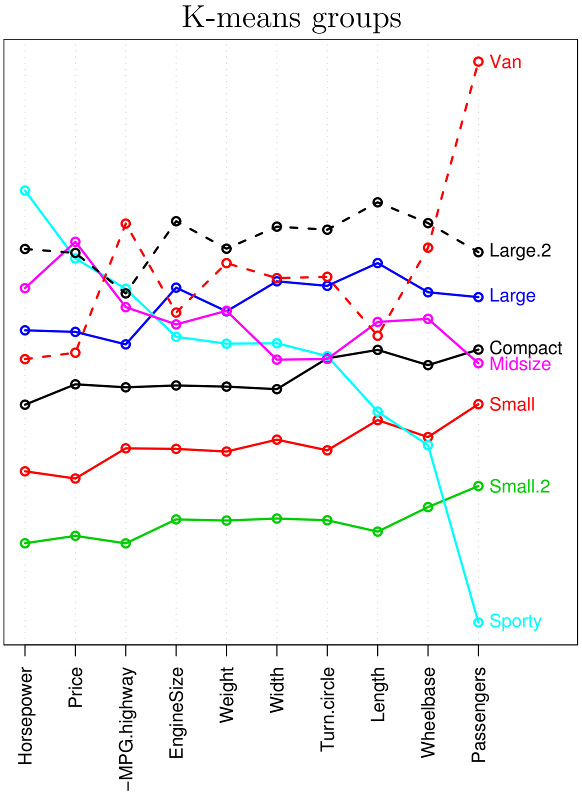
Does this data need an interaction term?

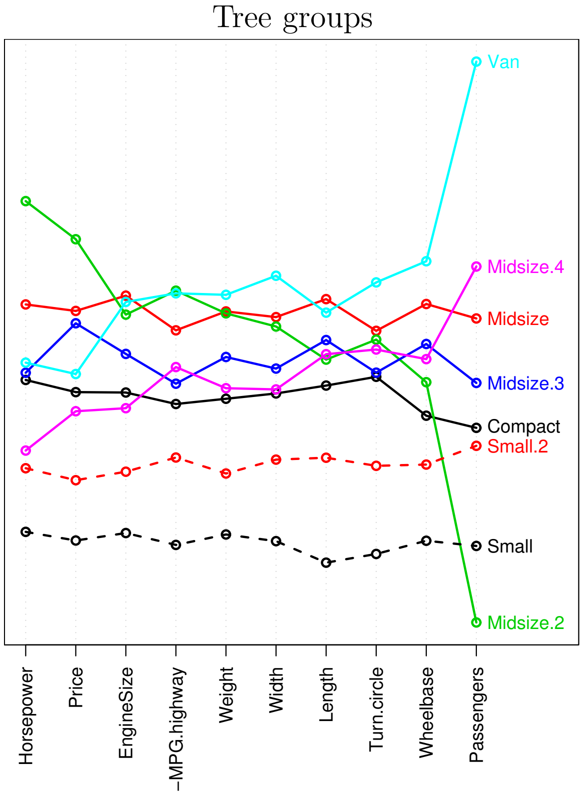




Trees are useful for constructing groups of cases whose response is similar and are easy to summarize (defined by a handful of attributes). This is better than partitioning according to the response (the groups need not be compact) or using k-means (the groups need not have different prices).



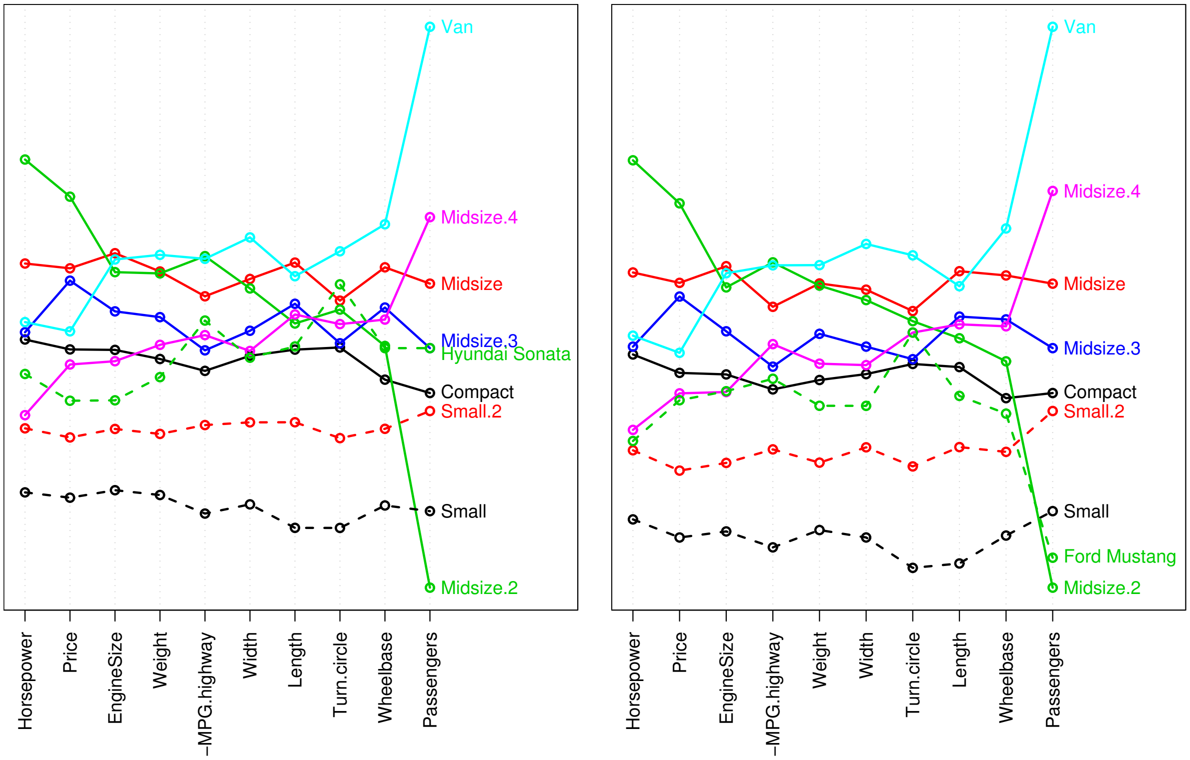




The tree groups are more distinct in price. `Large` and `Midsize` have been merged, and a new group called `Midsize.4` has been found.

Cars in `Midsize.4`:

```
                               Type Horsepower Weight Price
Ford Mustang                 Sporty        105   2850  15.9
Dodge Spirit                Compact        100   2970  13.3
Volvo 240                   Compact        114   2985  22.7
Buick Century               Midsize        110   2880  15.7
Chevrolet Lumina            Midsize        110   3195  15.9
Dodge Dynasty               Midsize        100   3080  15.6
Hyundai Sonata              Midsize        128   2885  13.9
Oldsmobile Cutlass_Ciera Midsize        110   2890  16.3
Volkswagen Eurovan              Van        109   3960  19.7
```



`Midsize.4` is an odd collection of cars—they are heavy but have low horsepower, with a price that reflects both. It is surprising to see the Ford Mustang here. It turns out that in 1993, Ford offered the "Mustang LX", a sporty-looking but low-powered car which capitalized on the Mustang image. It was discontinued the following year.

References

[1] "Fleet Uses CART Data Mining Technology to Understand Customer Characteristics and Habits." `http://www.salford-systems.com/appstories.html#fleet`