

Handling High Dimensional Variables: A Practical Example

Discussion Topic

- The problem
- Techniques for handling high dimensions
- Comparisons of different techniques
- Conclusions

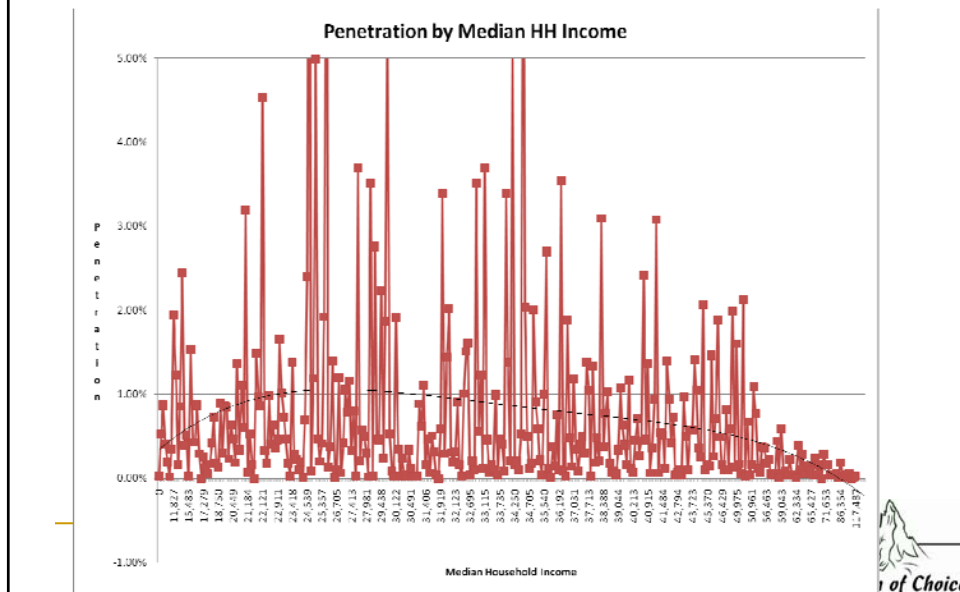
The Problem

- Data Analysis
 - Where am I penetrated in a state?
 - What tend to the be the characteristics of places where I am more highly penetrated?
- Action
 - Where am I under-penetrated?
 - What are my most likely scenarios for successful expansion?

ZIP Code Level Data

- Vehicle registration data
- Company vehicle counts
- ZIP Code Level Demographics
 - Age
 - Population density
 - Persons per household
 - Marital status
 - Urban vs. rural
 - Education

High Dimensional Explanatory Variables



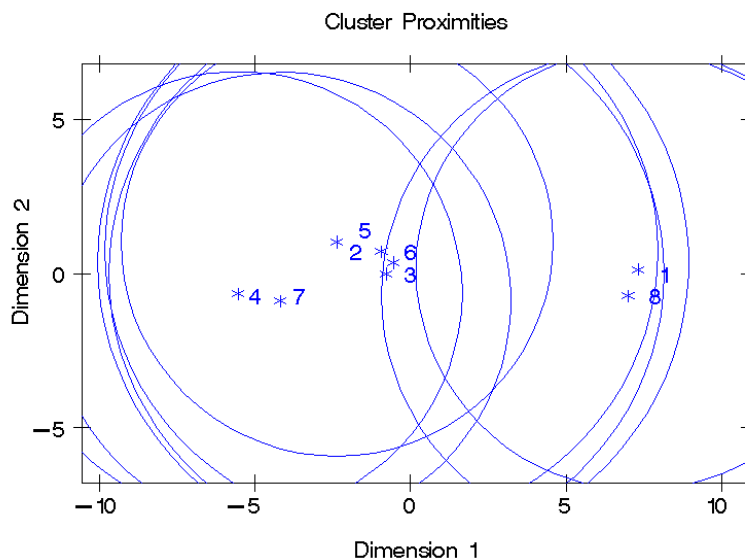
Techniques for Handling High Dimensional Variables

- Unsupervised
 - Clustering
 - Variable Clustering
 - Principal Components
- Supervised
 - Variable Selection
 - Traditional model development

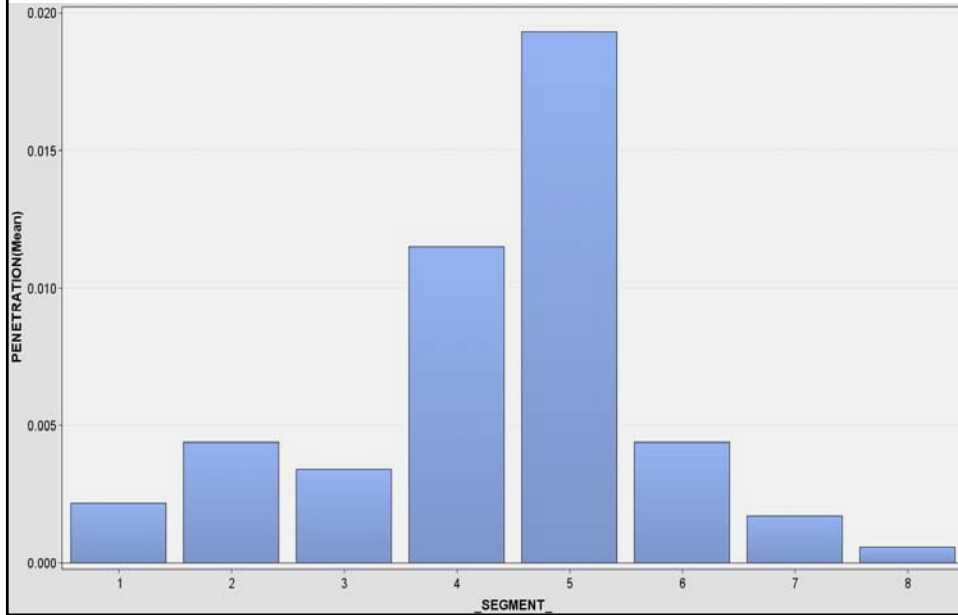
Clustering

- Observations placed into groups based on the input data
- Performs disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative input variables and cluster seeds
- Objects in each cluster tend to be similar, objects in different clusters tend to be dissimilar
- Doesn't look at the target variable

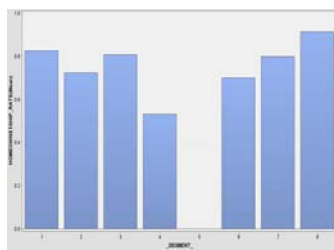
Cluster Distance Map



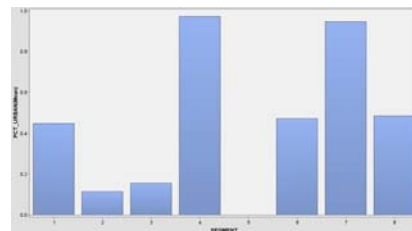
Penetration by Cluster Segment



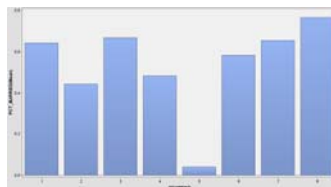
Correlation of Cluster with Independent Variables



Homeownership Rate

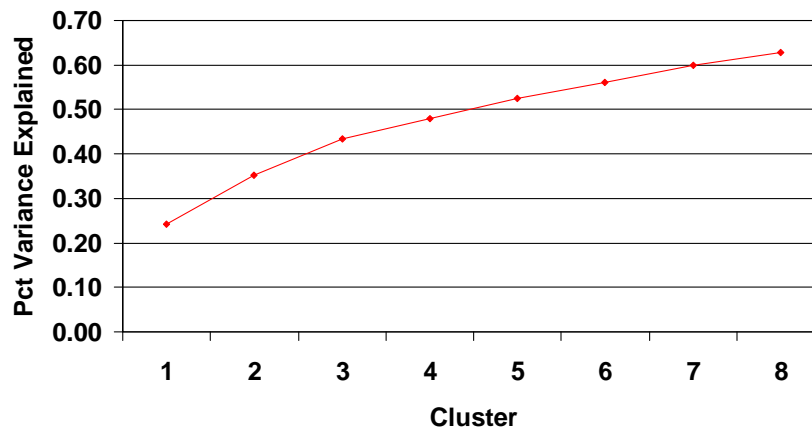


Percent Urban

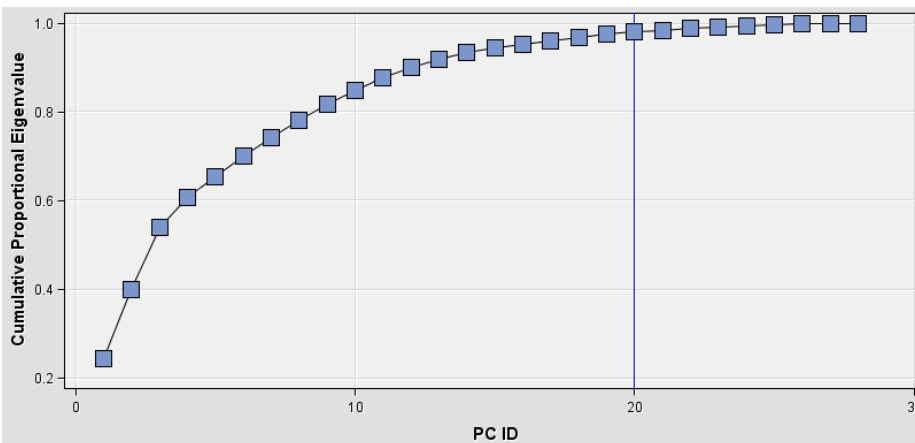


Percent Married

Variable Clustering – Variance Explained



Principal Components Result

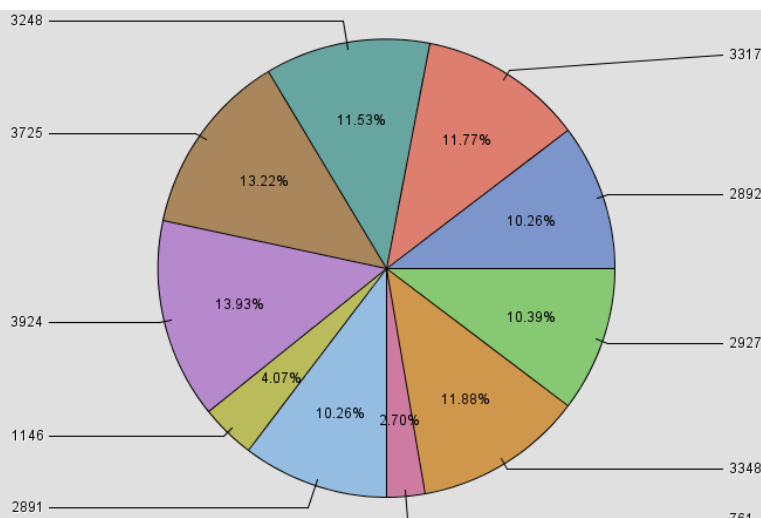


Unsupervised Learning Methods

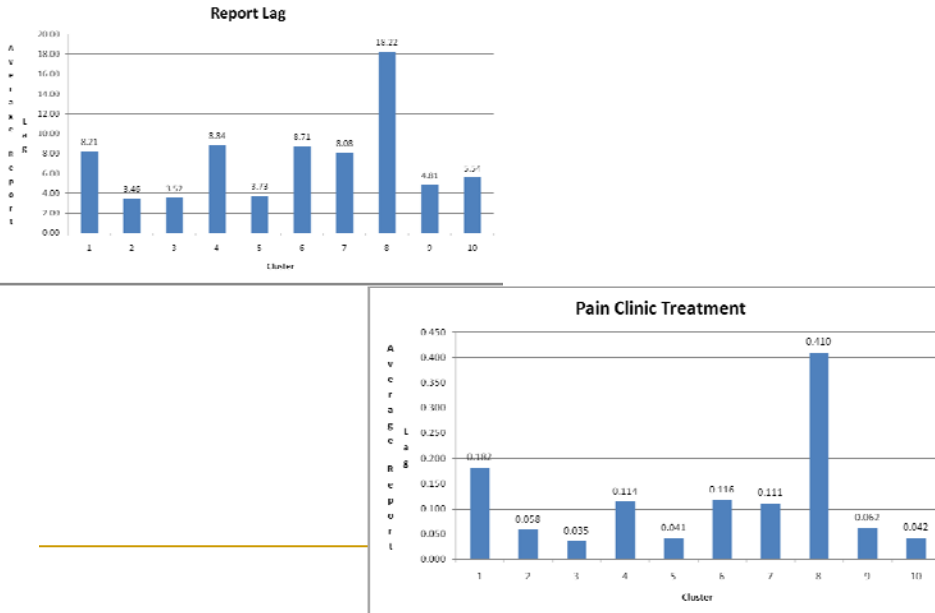
- Focus
 - Do not focus directly on the target (the dependent variables)
 - Focus is on putting observations of like independent variables together
 - If the independent variables are truly related to the dependent variable, then the clusters be related to the dependent variables
- Potential Applications
 - Claim fraud
 - Marketing targets
 - Underwriting selections



Identifying Anomalies - Segmentation



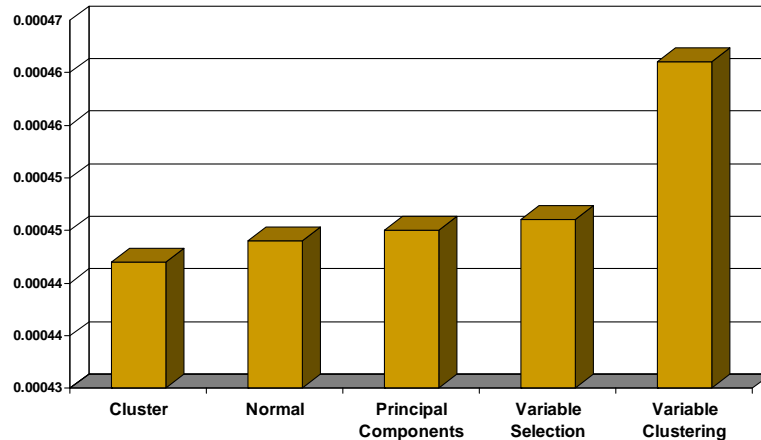
Differences in Clusters



Variable Selection

- Calculate the correlation coefficient
 - Exclude variables that do not meet criteria
- Forward stepwise regression sequentially adds variables that produce the largest incremental increase in explanatory power
- Process ends when no more variables can be added to produce a significant improvement

Comparison of Final Models



PINNACLE
ACTUARIAL RESOURCES, INC. 
The Firm of Choice

Conclusions

- Using input variable information directly is generally preferable when building predictive models
- There are many cases when this is not feasible
 - Unknown target
 - Input variable with too many levels
 - Too many input variables
- Techniques for handling high dimensional variables still result in models that produce predictive results

PINNACLE
ACTUARIAL RESOURCES, INC. 
The Firm of Choice