Handling High Dimensional Variables: A Practical Example

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

Discussion Topic

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

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

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?

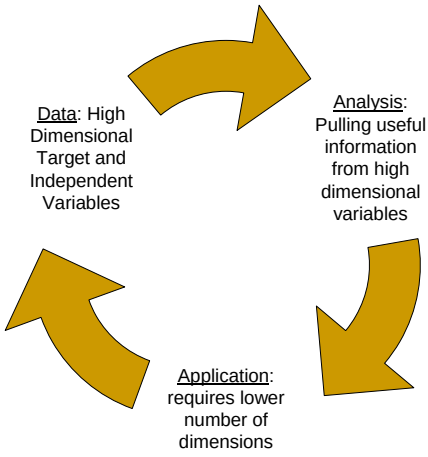
PINNACLE ACTUARIAL RESOURCES, INC. *The Firm of Choice*

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

PINNACLE ACTUARIAL RESOURCES, INC. *The Firm of Choice*

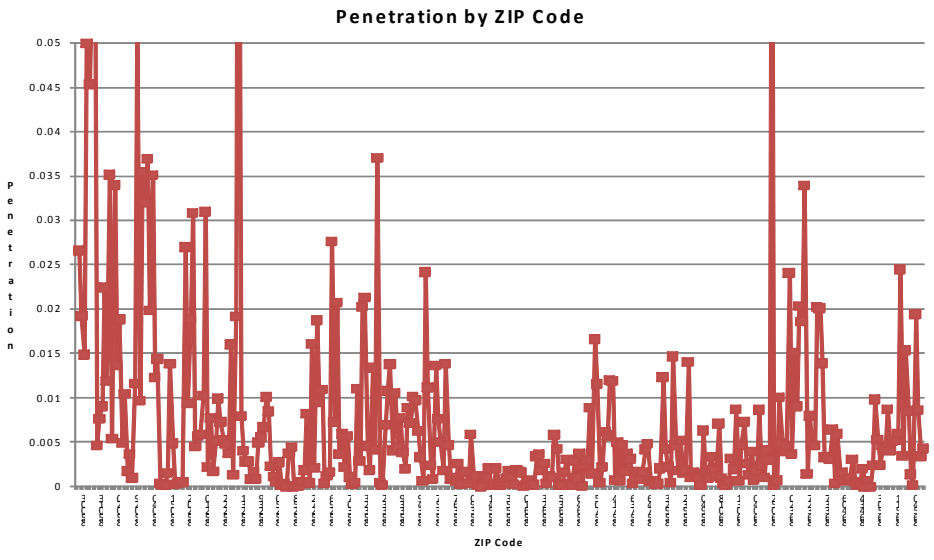
The Problem With High Dimensional Variables



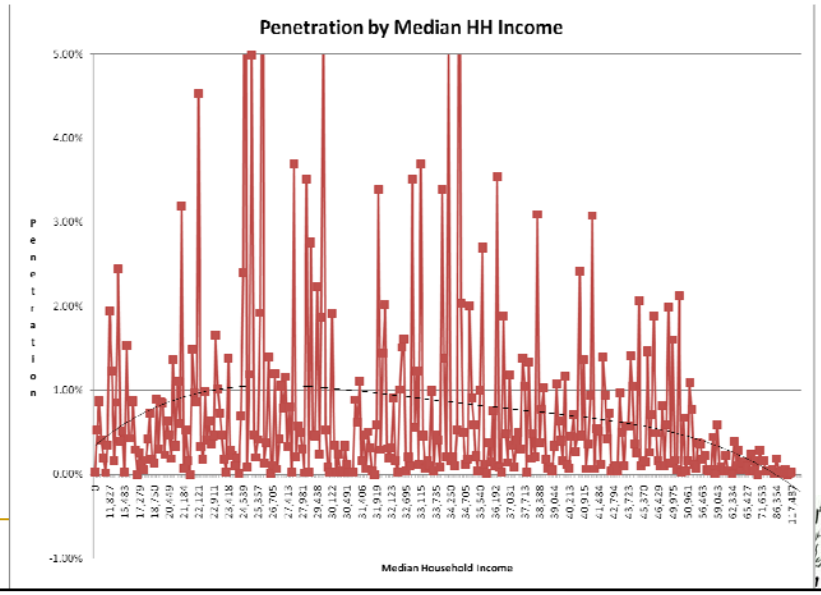
- How do I understand trends in the data?
- How do I make this information actionable for the business units?

PINNACLE ACTUARIAL RESOURCES, INC.
The Firm of Choice

High Dimensional Target



High Dimensional Explanatory Variables



Techniques for Handling High Dimensional Variables

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

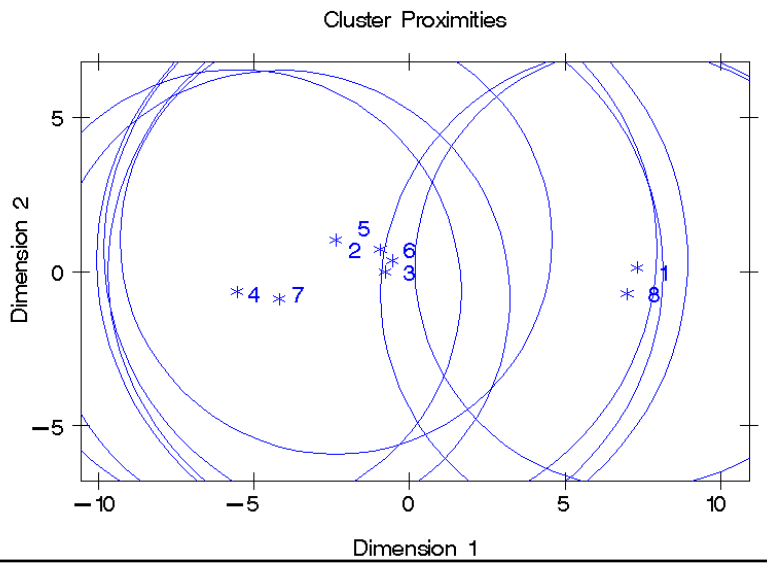
PINNACLE ACTUARIAL RESOURCES, INC.
The Firm of Choice

Clustering

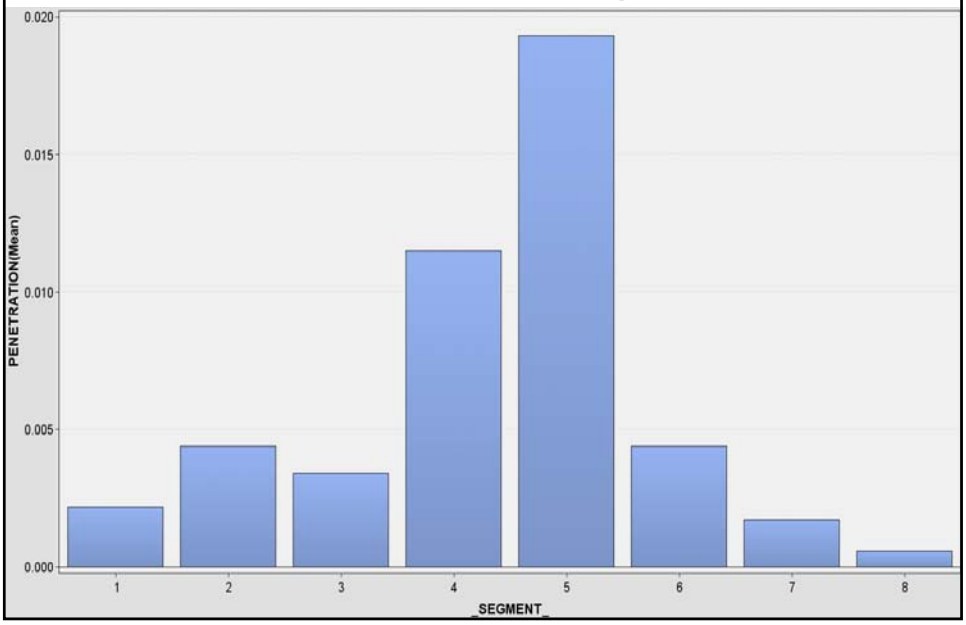
- Observations placed into groups based on the input data
- Performs disjoint cluster analysis on the basis of Euclidean distances computed from one or more quantitative input variables and cluster seeds
- Objects in each cluster tend to be similar, objects in different clusters tend to be dissimilar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

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

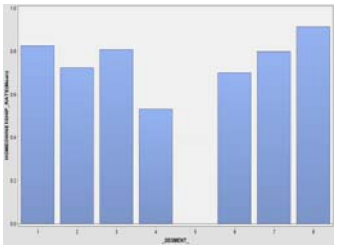
Cluster Distance Map



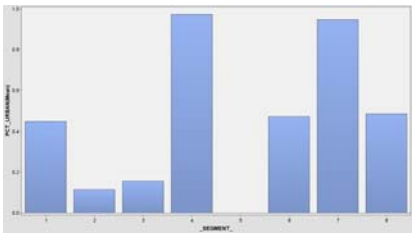
Penetration by Cluster Segment



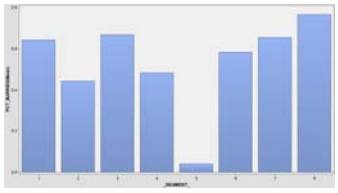
Correlation of Cluster with Independent Variables



Homeownership Rate



Percent Urban



Percent Married

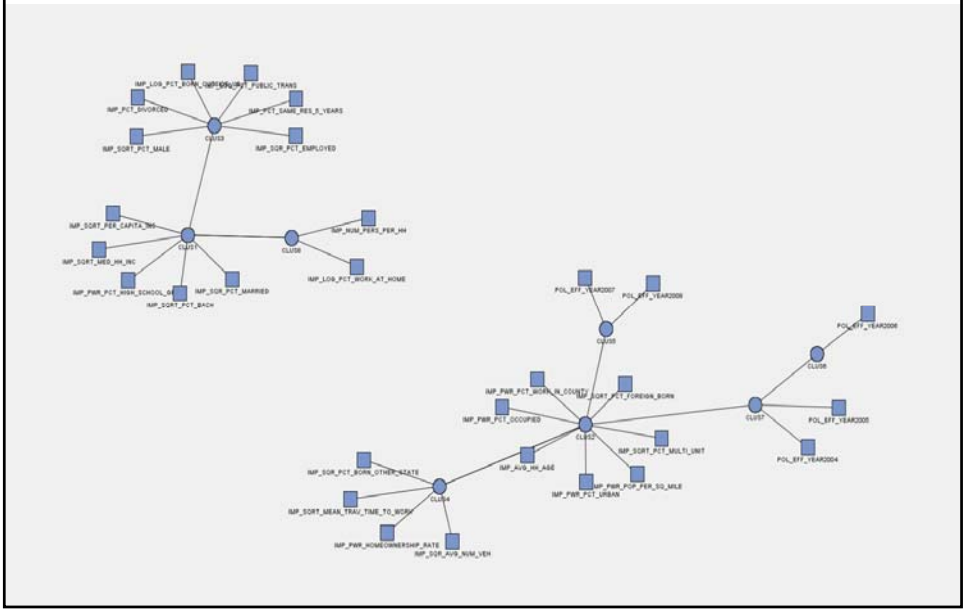
PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

Variable Clustering

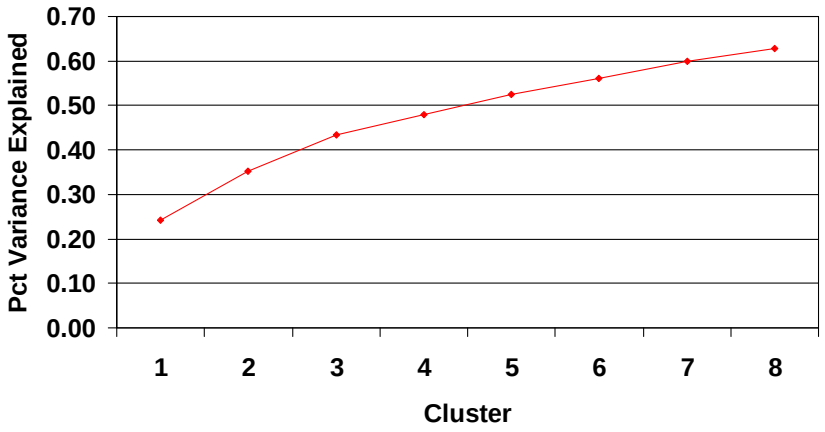
- Divides variables into clusters
- Resulting cluster is a linear combination of variables in cluster
  - First principal component
- Attempts to explain the maximum variance in the inputs

PINNACLE ACTUARIAL RESOURCES, INC.
The Firm of Choice

Variable Clustering

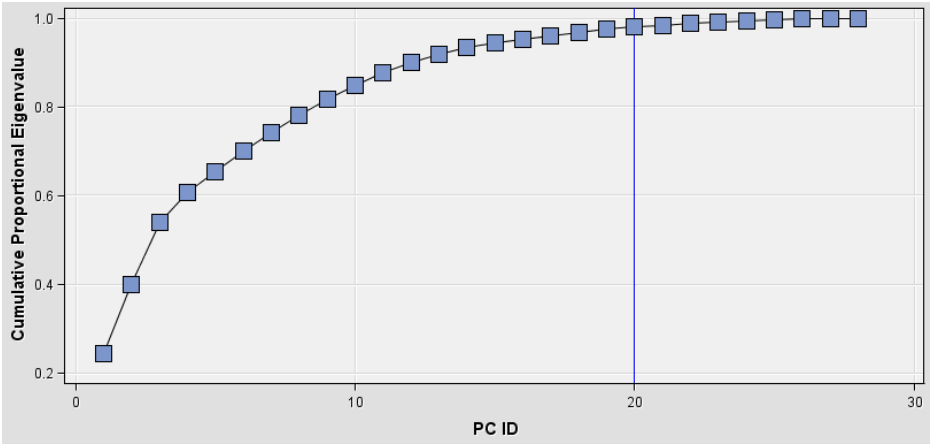


Variable Clustering – Variance Explained



PINNACLE ACTUARIAL RESOURCES, INC. *The Firm of Choice*

Principal Components Result



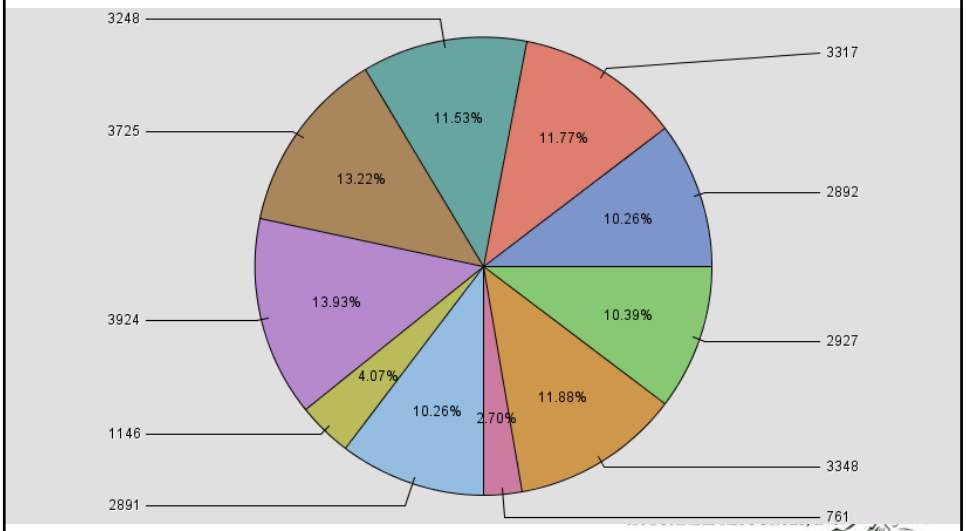
PINNACLE ACTUARIAL RESOURCES, INC. *The Firm of Choice*

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

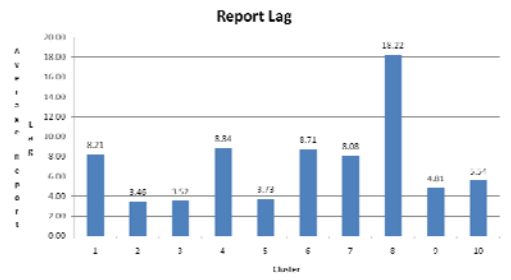
PINNACLE ACTUARIAL RESOURCES, INC. The Firm of Choice

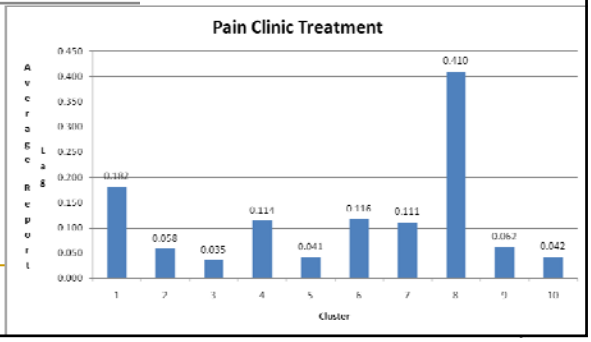
Identifying Anomalies - Segmentation



The Firm of Choice

Differences in Clusters



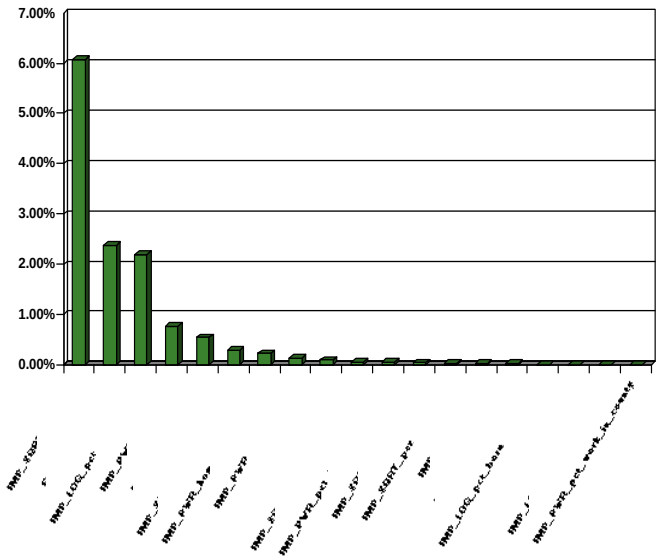


Variable Selection

- 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

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

Variable Selection - Sequential R-squared

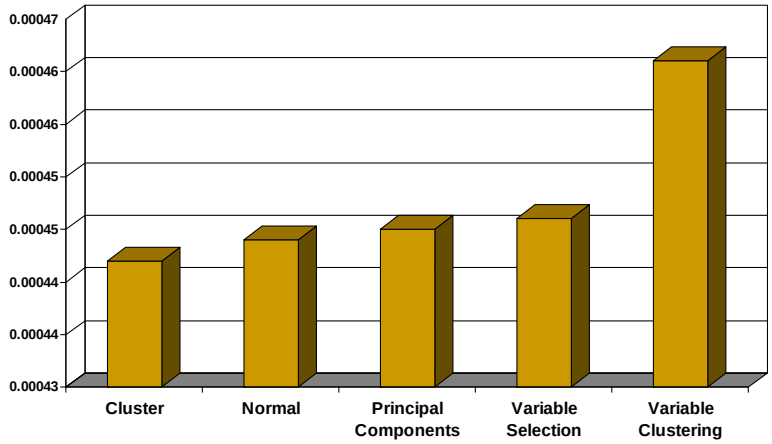


Final Comparison

- Comparison of models based on five sets of inputs
  - Clustering
  - Variable clustering
  - Variable selection
  - Principal components
  - Raw inputs
- Review numerous model comparison techniques

PINNACLE ACTUARIAL RESOURCES, INC.
The Firm of Choice

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*