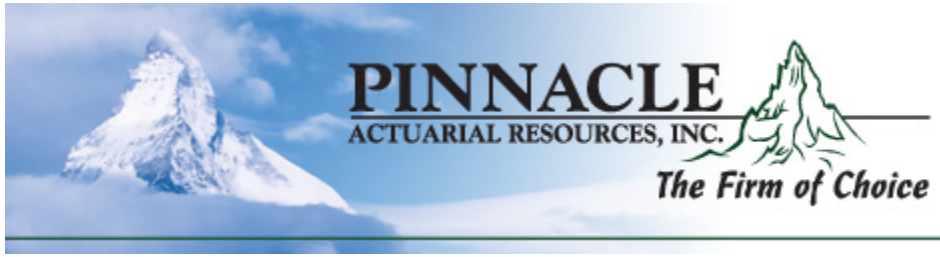
Handling High Dimensional Variables

Casualty Actuaries of the Northwest
Shawna Ackerman, Pinnacle Actuarial Resources
September 25, 2009



Discussion Topic

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

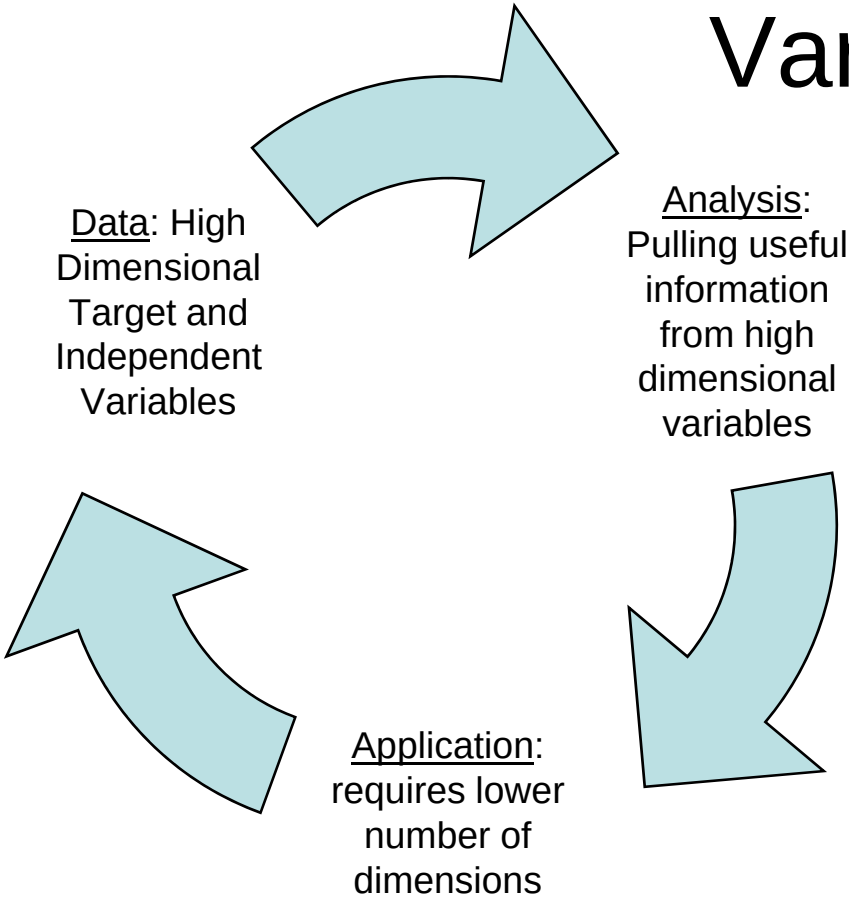
The Problem

- 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
- Demographics at zip code level
  - Age
  - Population density
  - Persons per household
  - Marital status
  - Urban vs. rural
  - Education

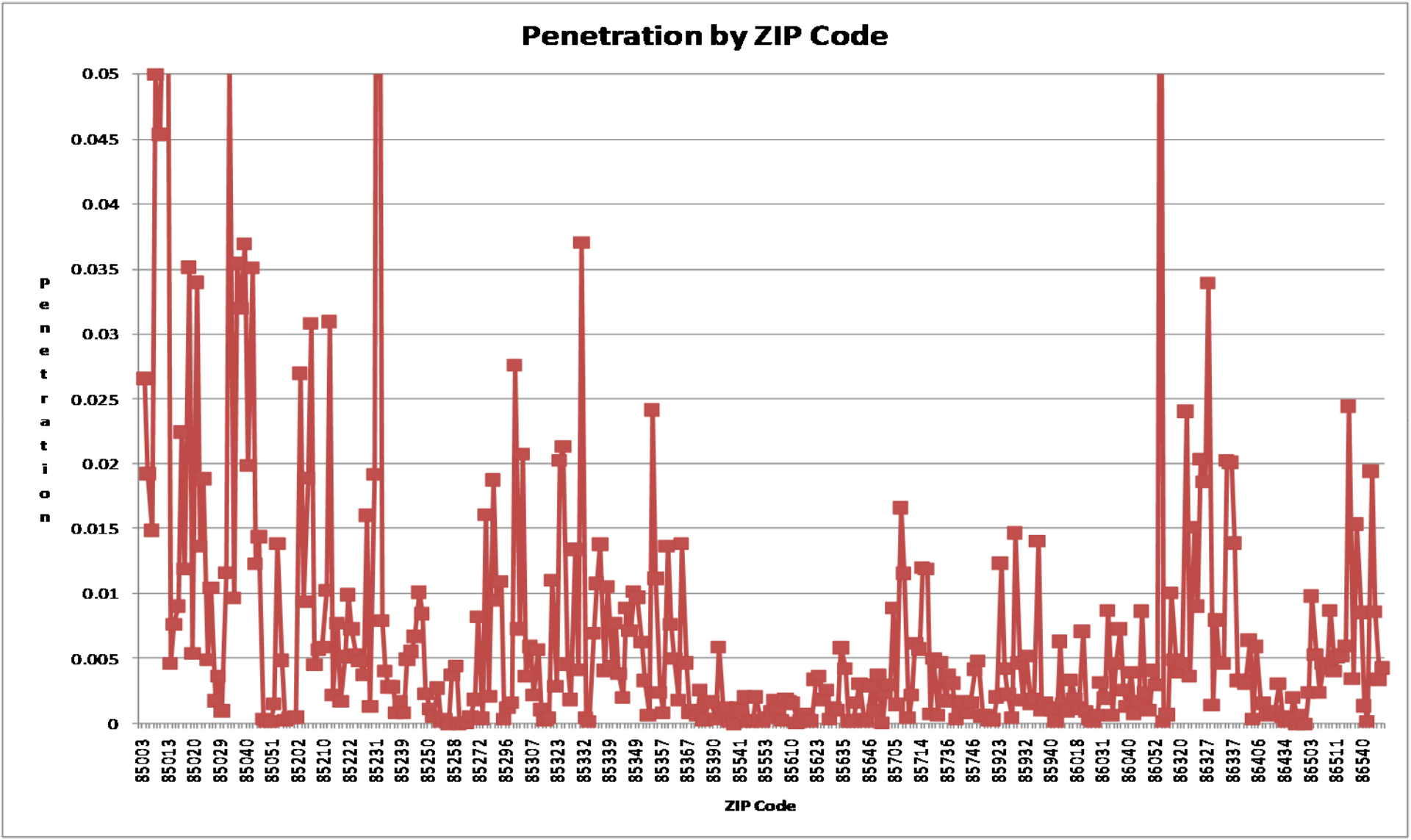
The Problem With High Dimensional Variables



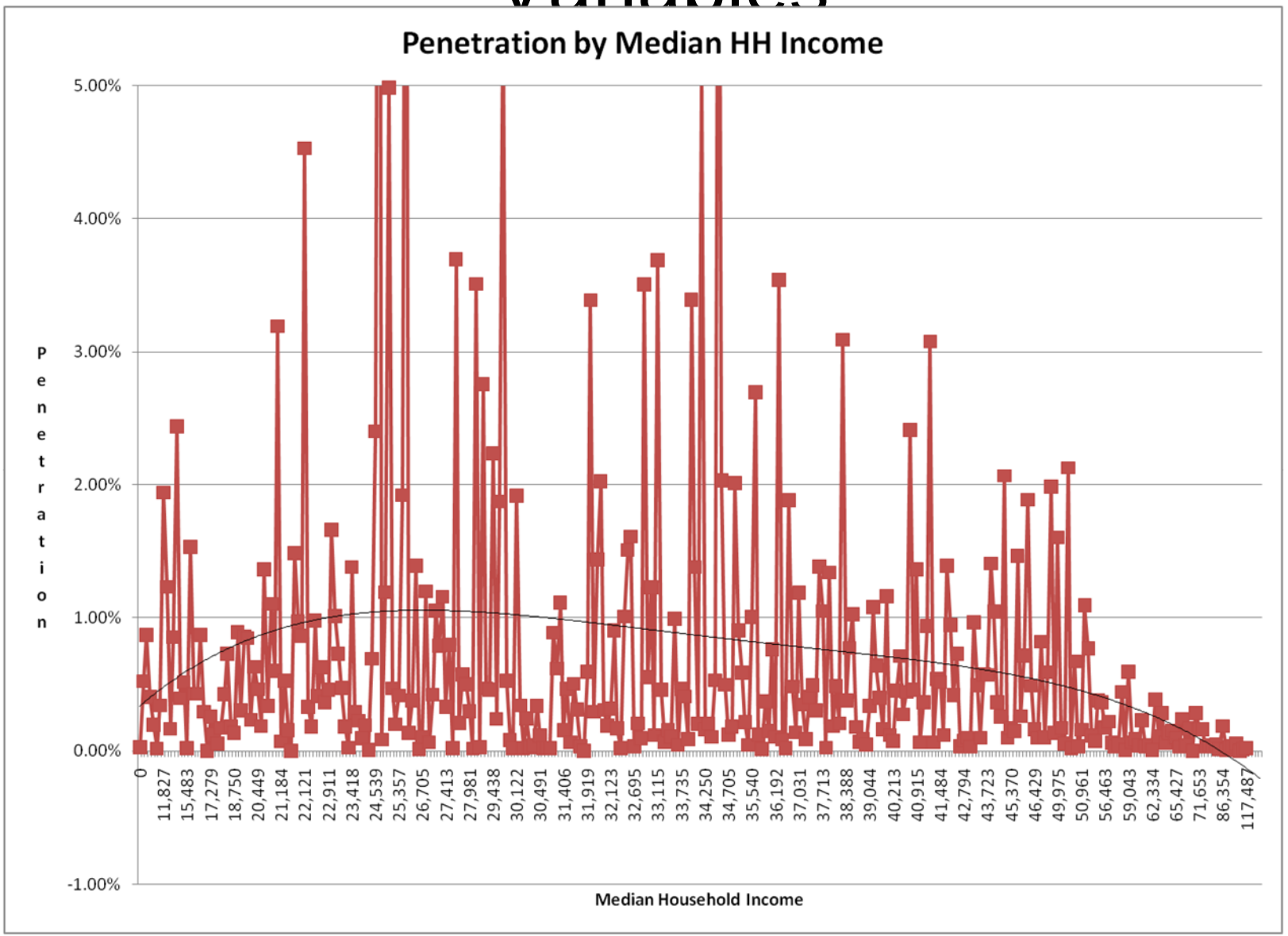
- How do I understand trends in the data?
- How do I make this information actionable for the business units?

"We are drowning in information and starving for knowledge."

High Dimensional Target



High Dimensional Explanatory Variables



Techniques for Handling High Dimension Variables

- Unsupervised
  - Clustering
  - Variable Clustering
  - Principal Component Analysis
- Supervised
  - Variable Selection
  - Traditional model development

Techniques for Handling High Dimension Variables

Supervised

- Model defined by target / outcome
- Clear measure of success

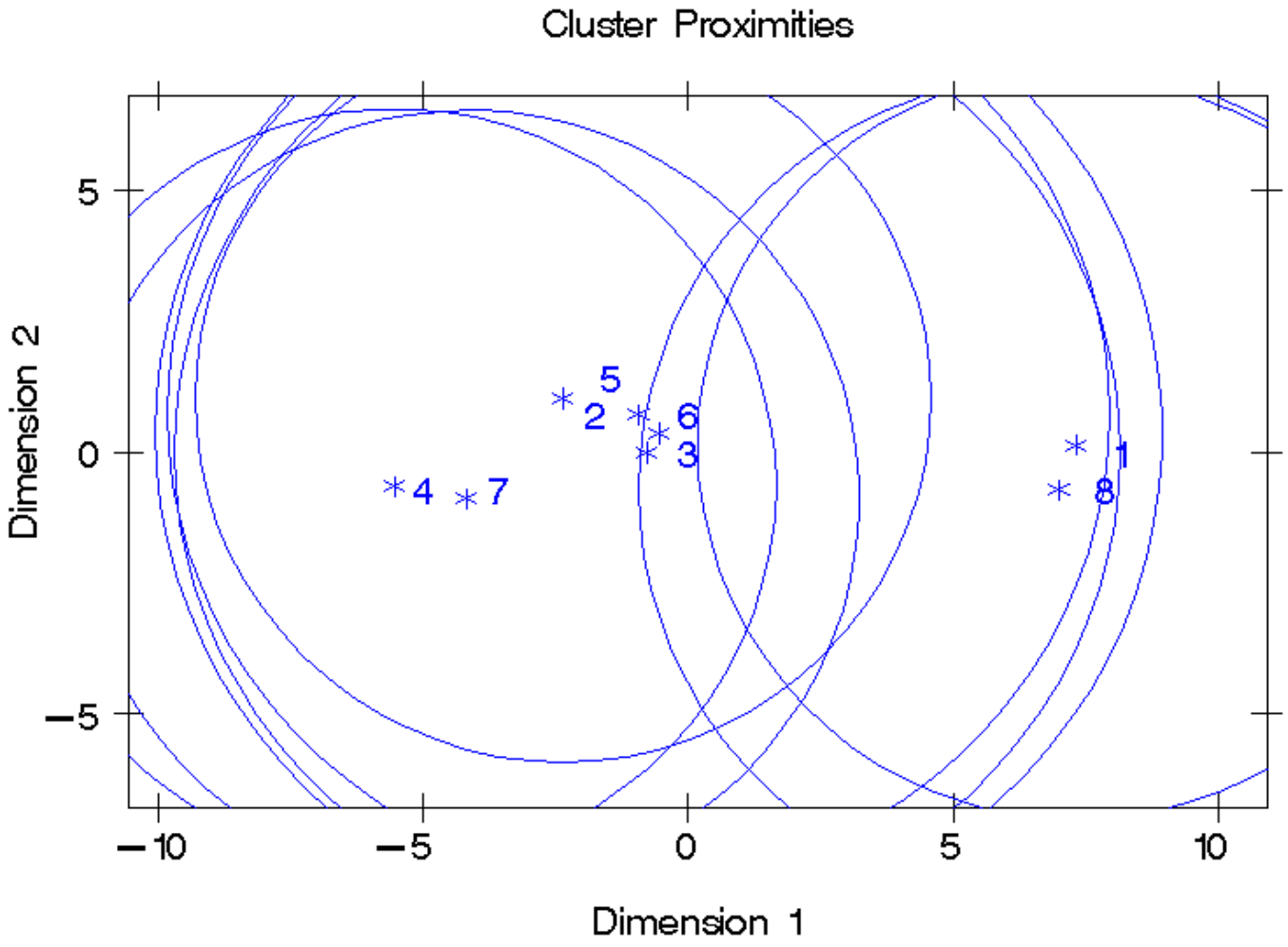
Unsupervised

- Describe data independent of outcome
- No direct measure of success

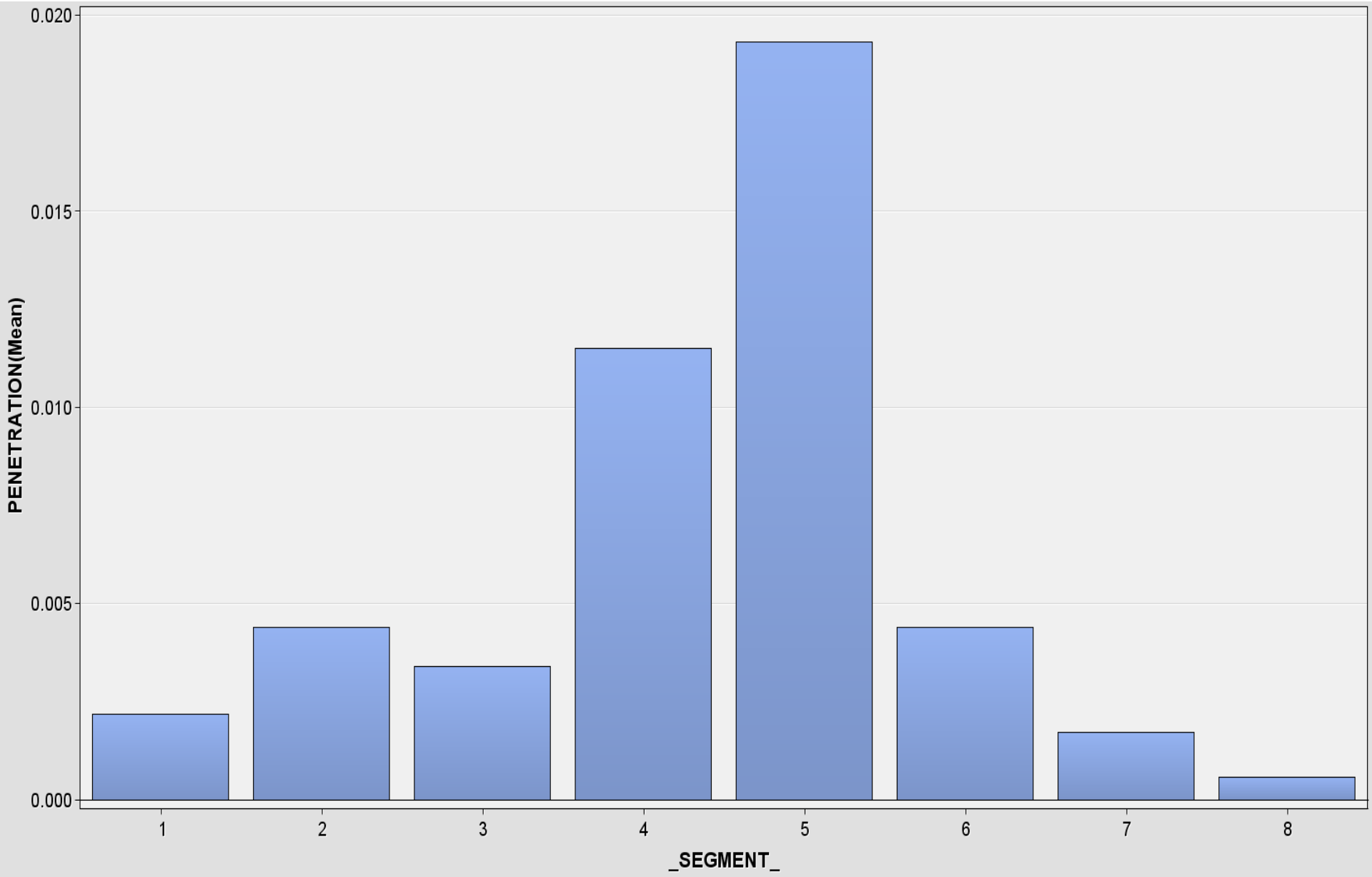
Clustering

- Divides a data set into groups of similar characteristics without regard to the target variable
- Groups created such that objects within each cluster are more closely related to one another than objects assigned to different clusters
- Key choice – distance or dissimilarity measure
  - e.g., squared distance, absolute difference
- Goal is to reduce the number of levels

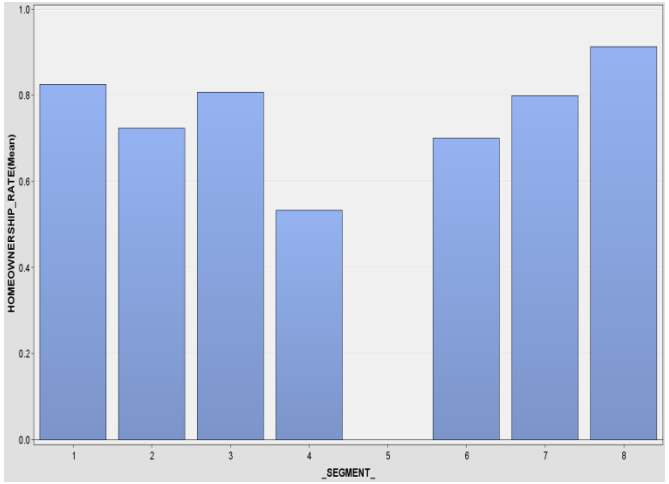
Cluster Distance Map



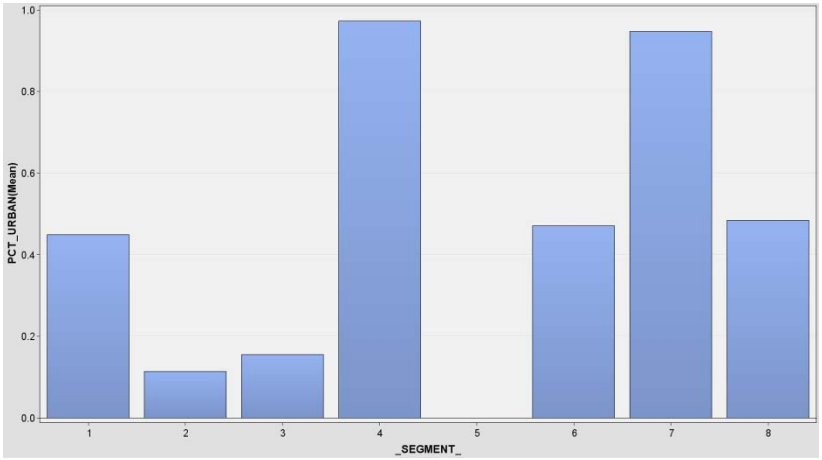
Penetration by Cluster Segment



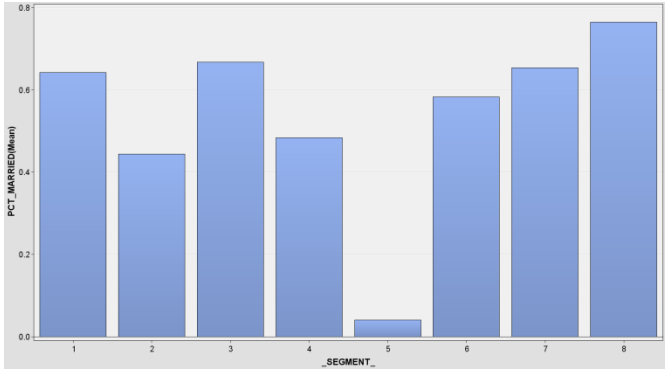
Correlation of Cluster with Independent Variables



Homeownership Rate



Percent Urban



Percent Married

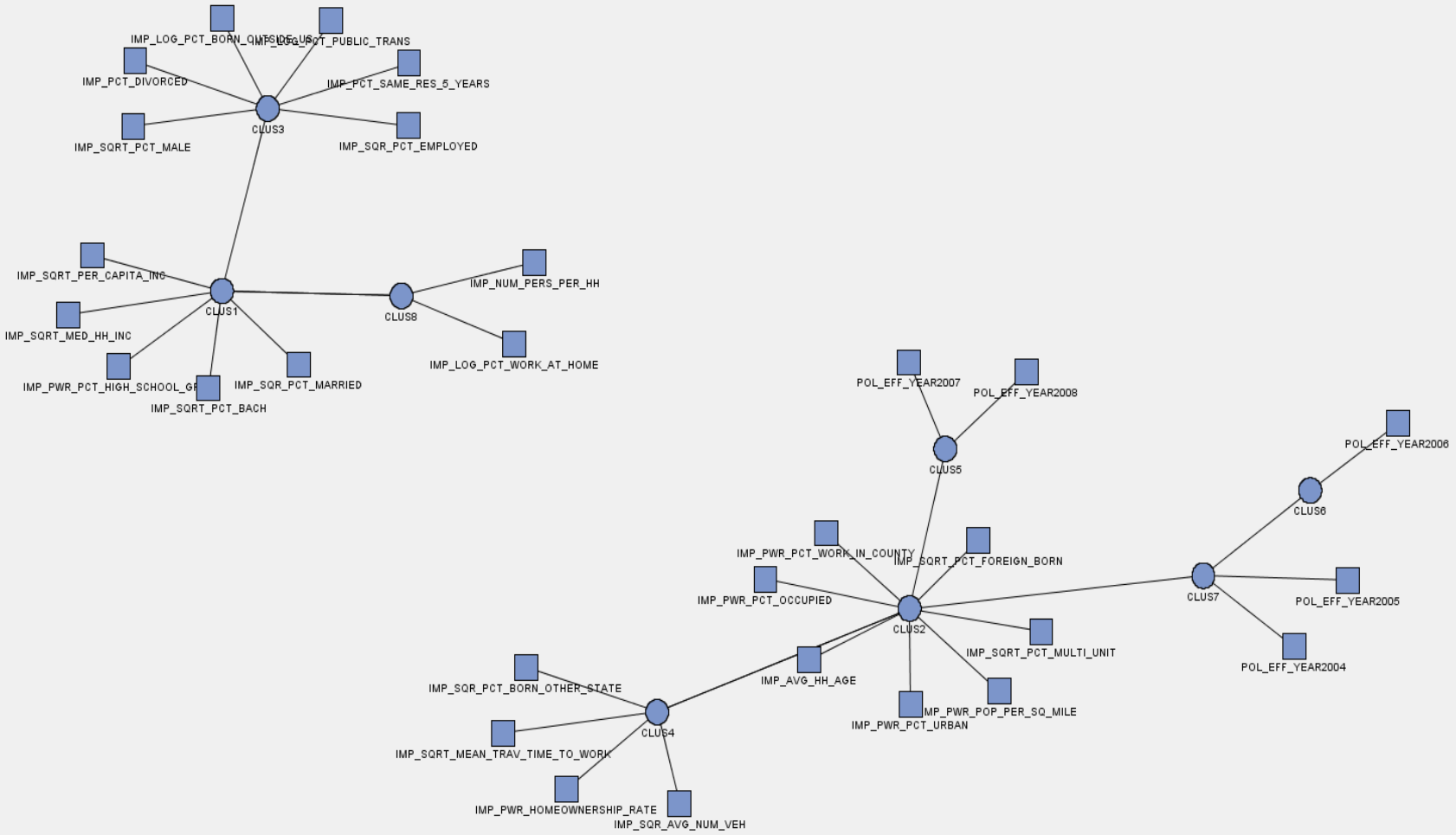
Variable Clustering

- 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
- Goal is to remove redundant variables

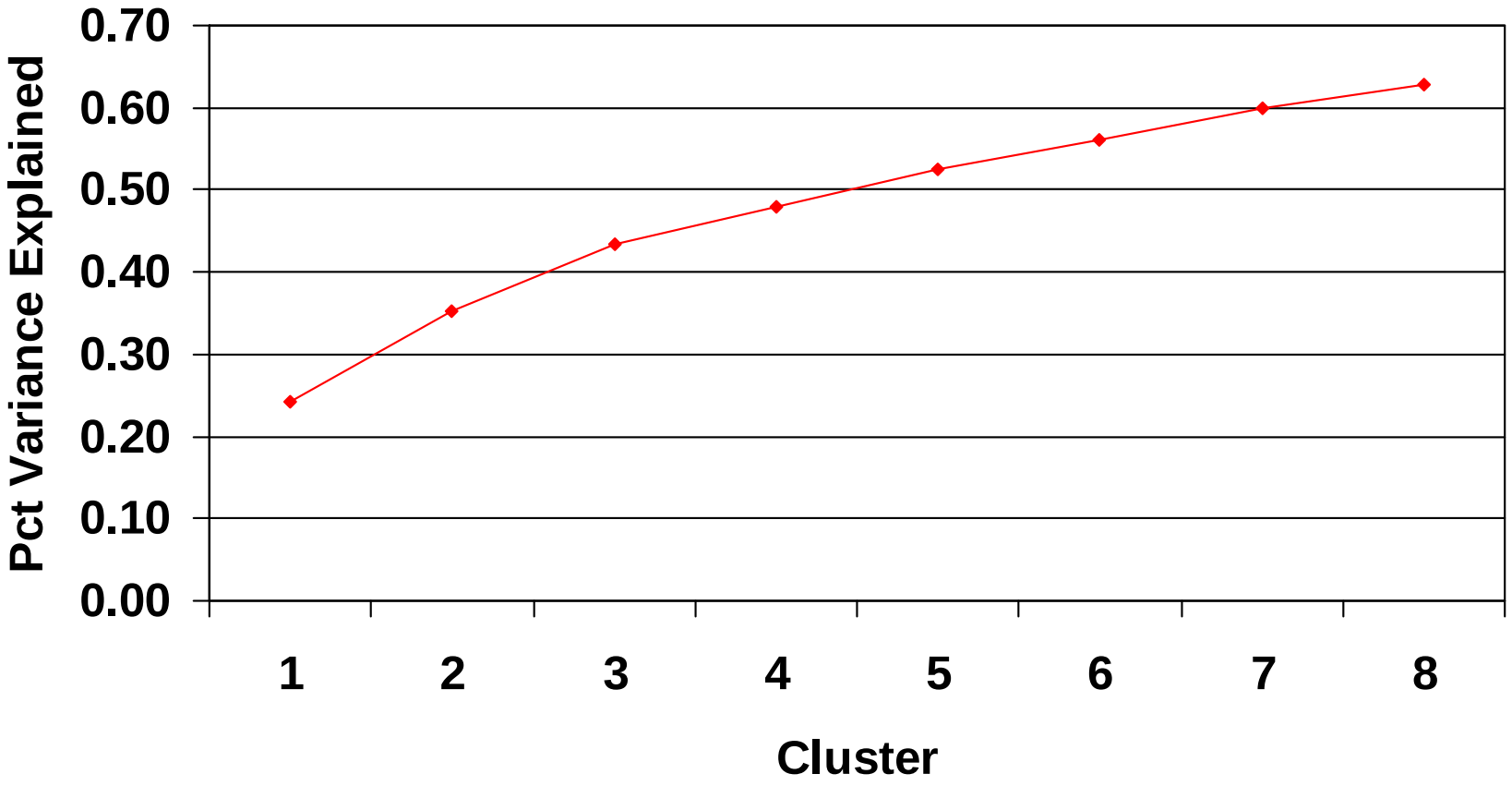
Variable Clustering

- Clustering rule – select variable with minimum $1\text{-}R^2_{\text{ratio}}$ as cluster representative
- $1\text{-}R^2_{\text{ratio}} = 1\text{-}R^2_{\text{own}} / 1\text{-}R^2_{\text{nearest}}$

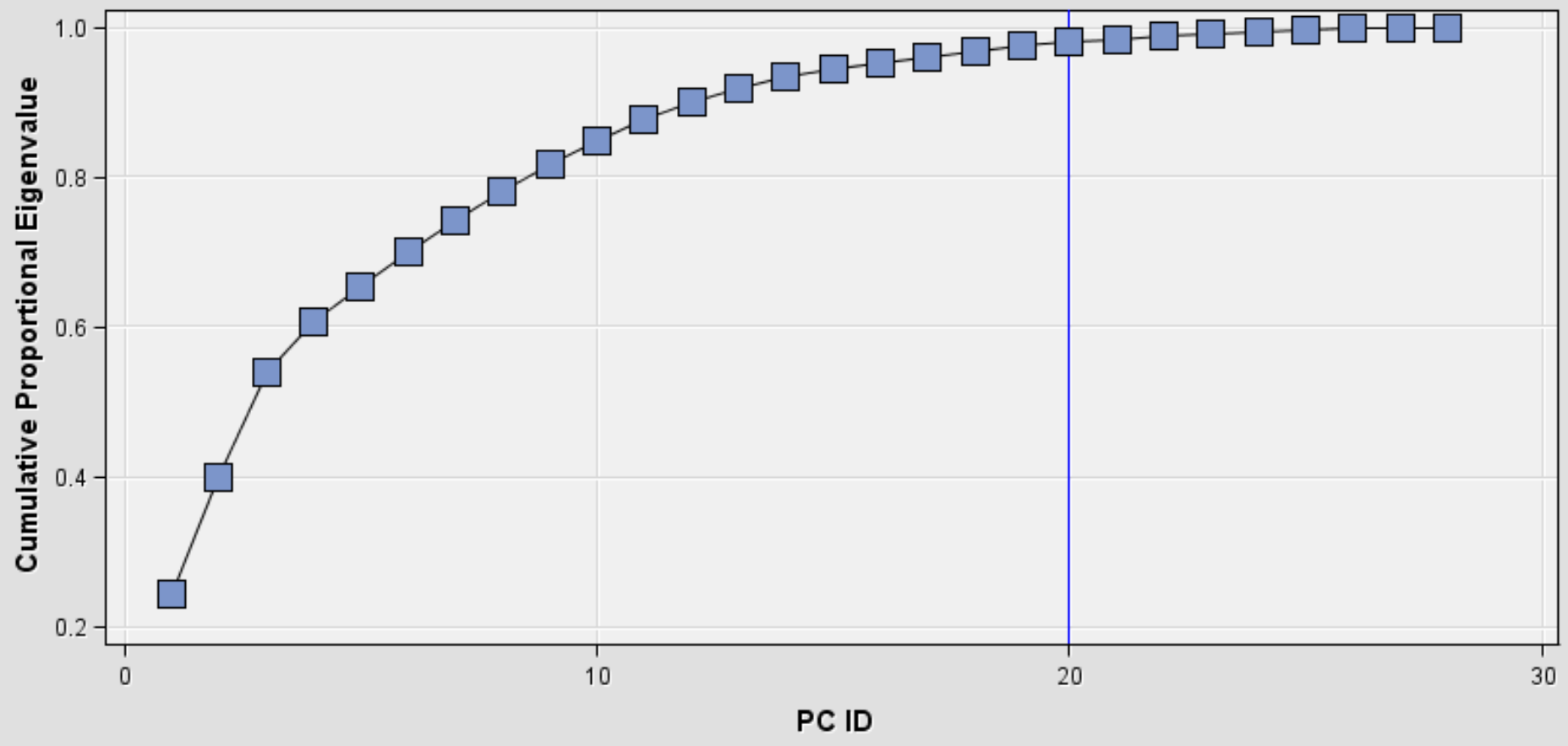
Variable Clustering



Variable Clustering – Variance Explained



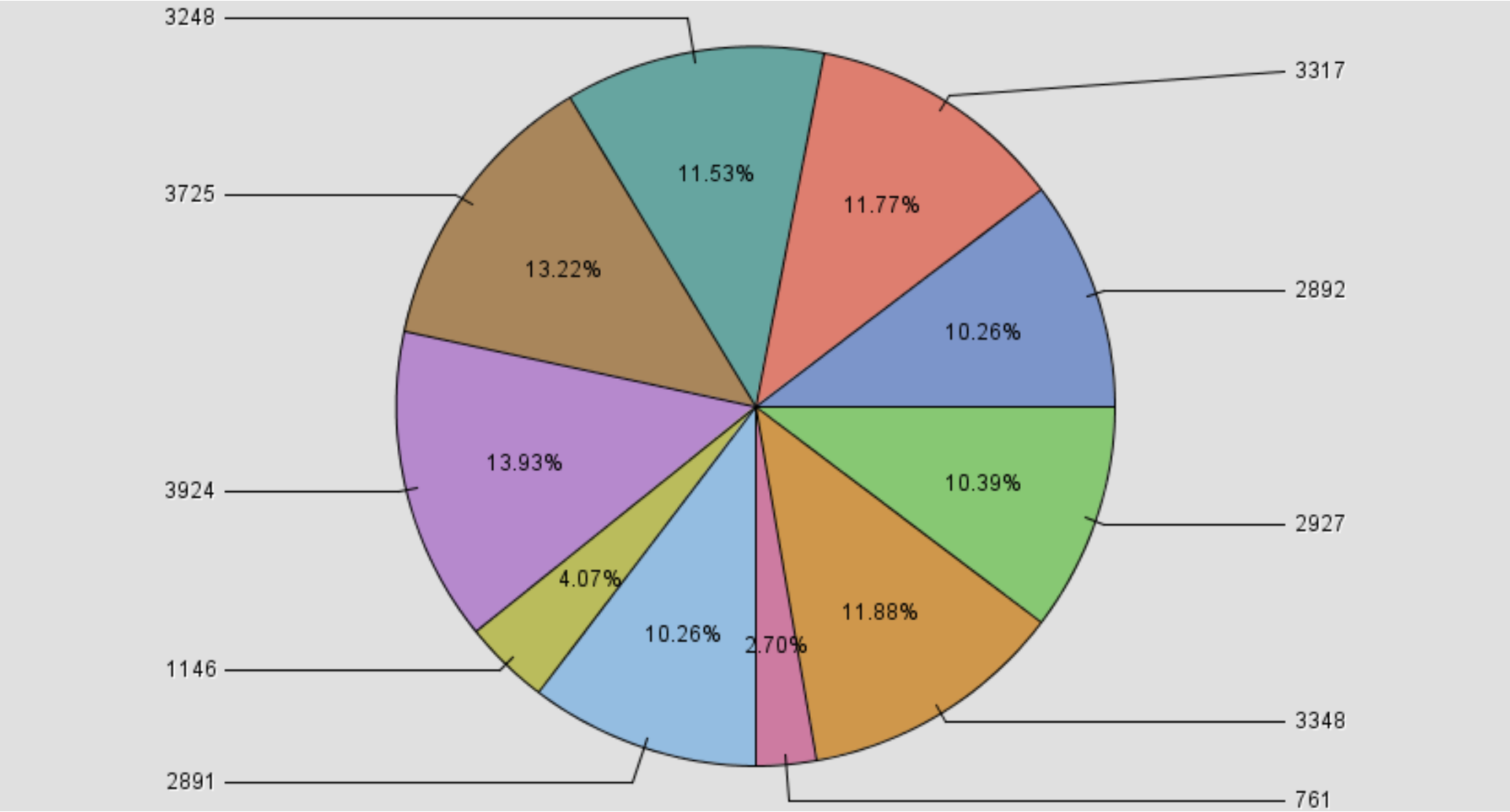
Principal Components Result



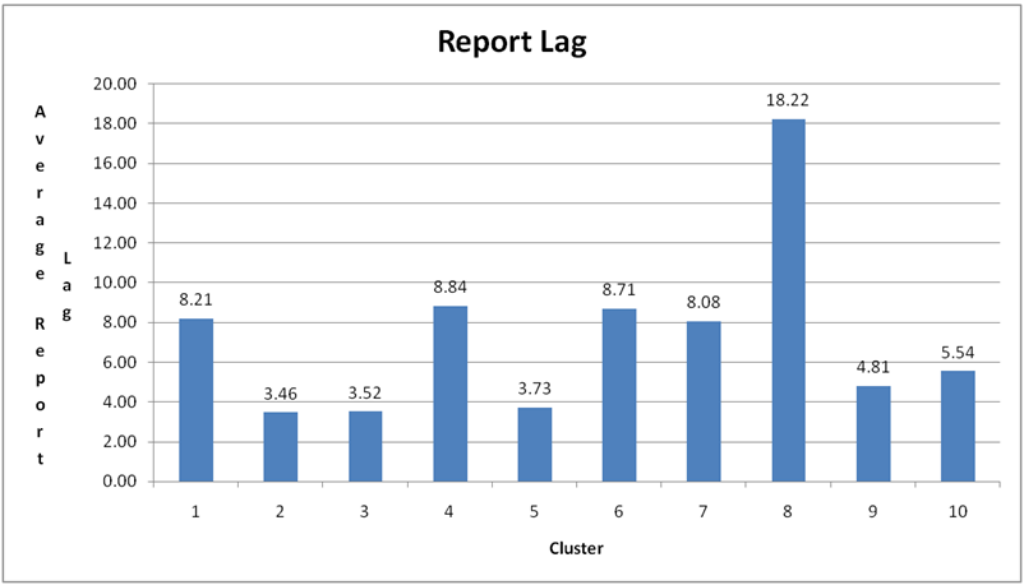
Unsupervised Learning Methods

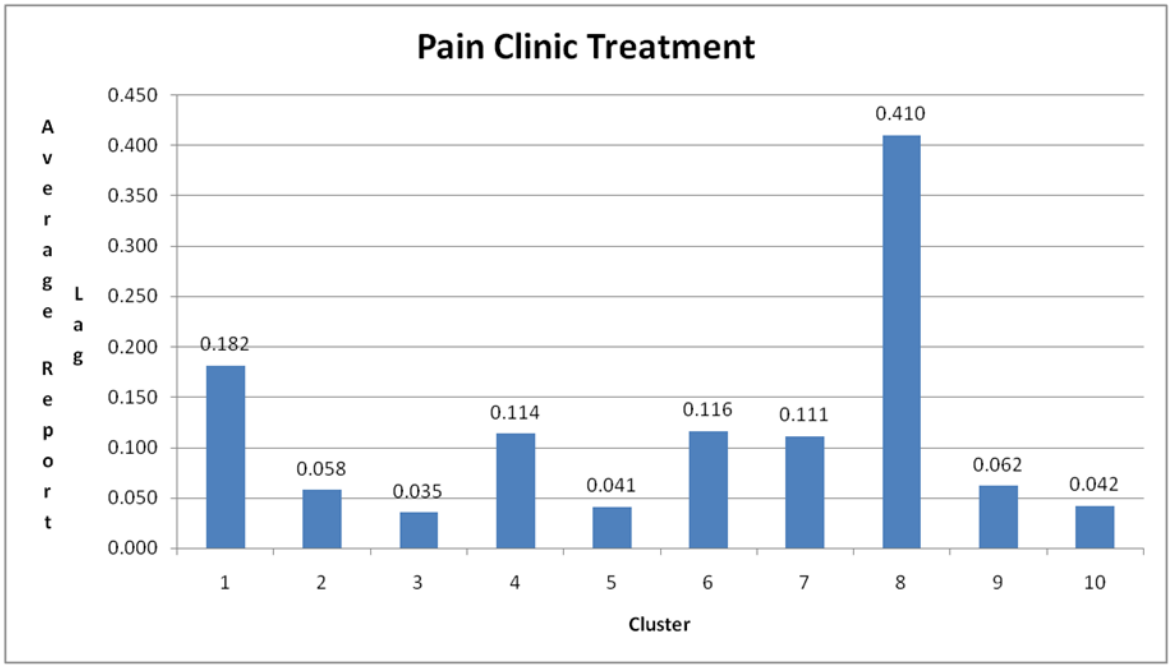
- Focus
  - Do not focus directly on the target (the dependent variables)
  - Focus is on putting observations of like independent variables together
  - If the independent variables are truly related to the dependent variable, then the clusters will be related to the dependent variables
- Potential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

Identifying Anomalies - Segmentation



Differences in Clusters

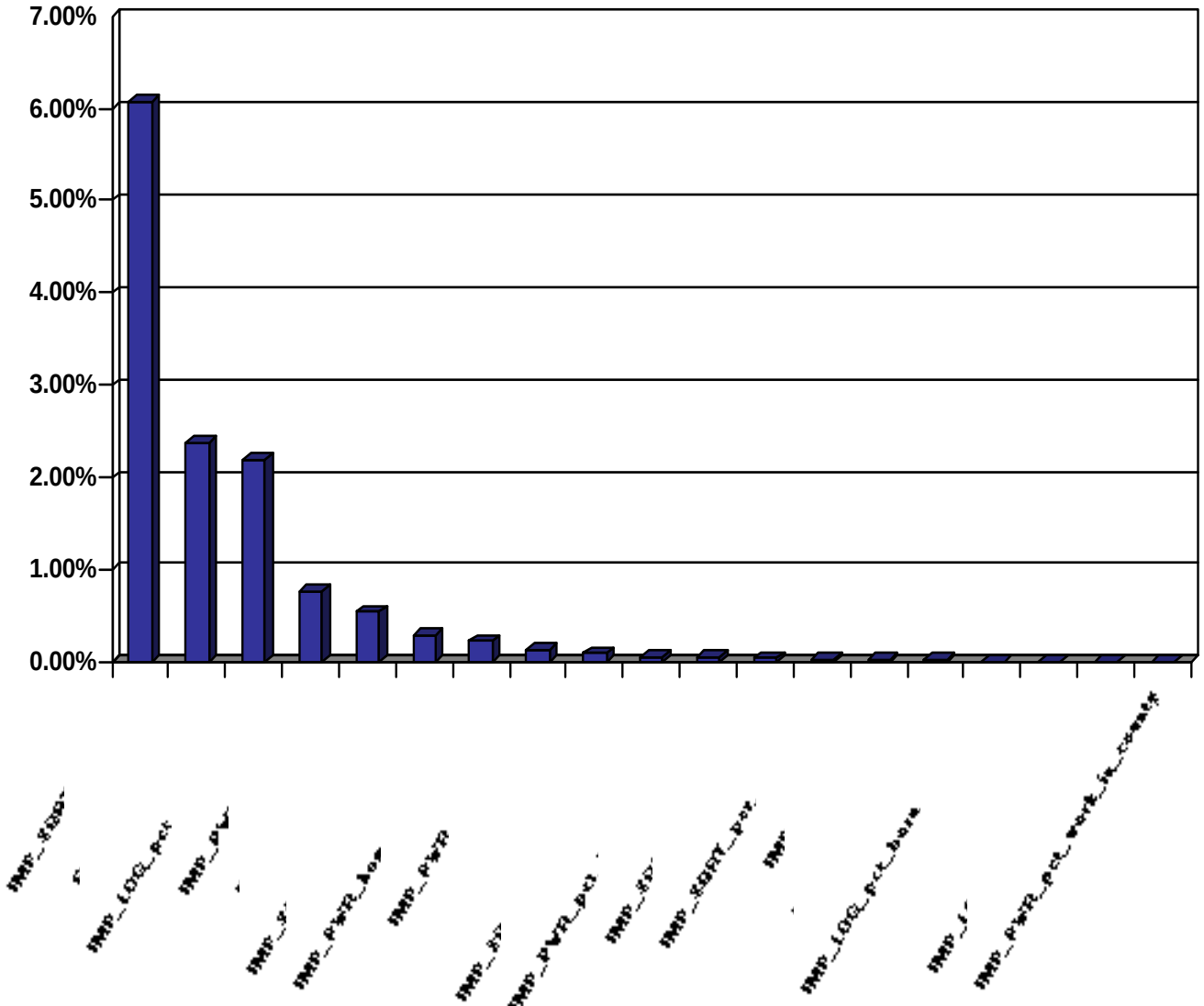




Variable Selection

- Calculate the correlation coefficient
  - Exclude variables that do not meet specified 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

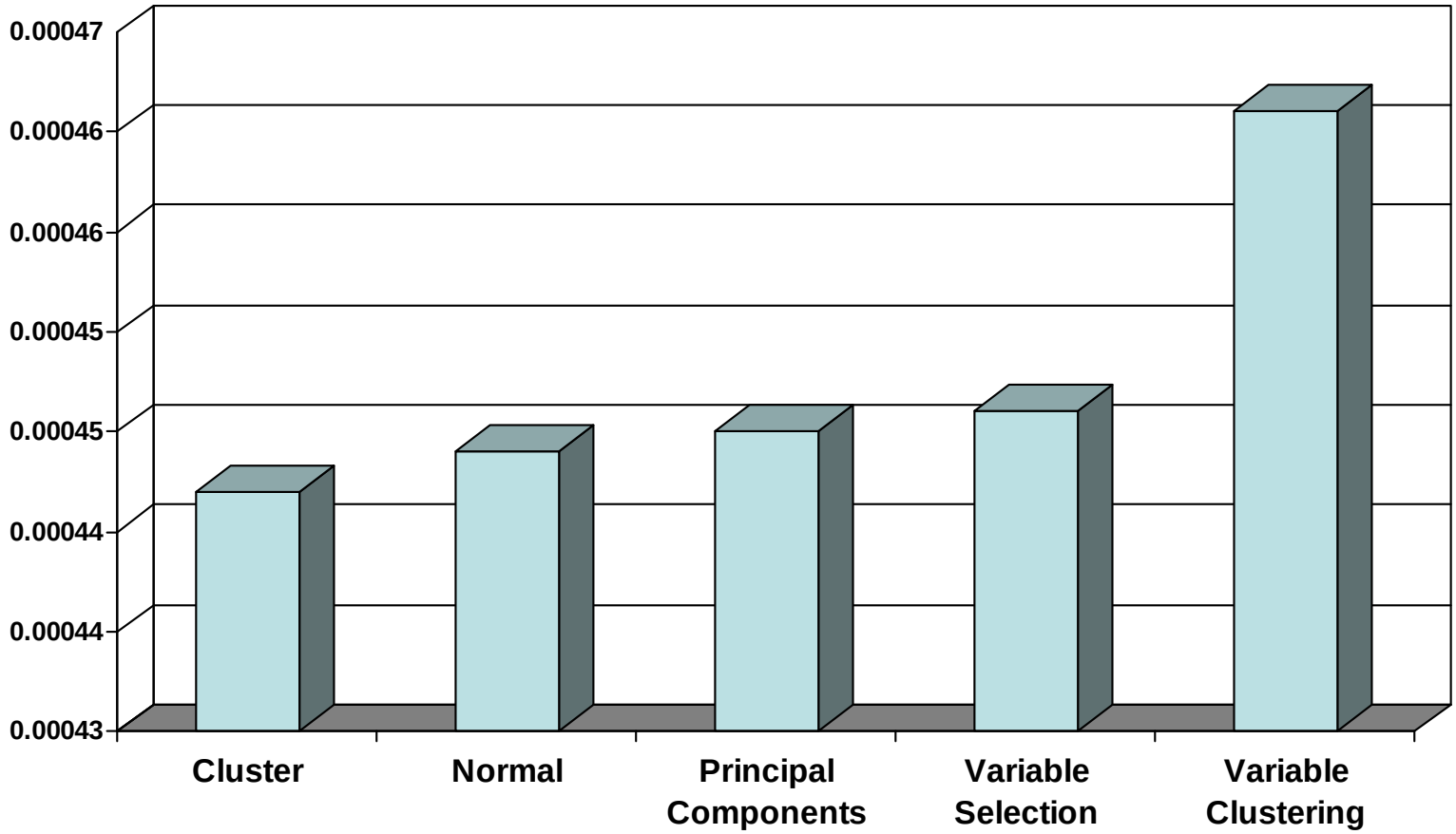
Variable Selection - Sequential R-squared



Final Comparison

- Comparison of models based on five sets of inputs
  - Clustering
  - Variable clustering
  - Variable selection
  - Principal components
  - Raw inputs

Comparison of Final Models



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

References

- *The Elements of Statistical Learning*, Hastie, Tibshirani, Friedman
- *Variable Reduction for Predictive Modeling with Clustering*, Robert Sanche and Kevin Lonergan, CAS Winter Forum 2006
- SAS User Guide 9.2

Contact Info

Shawna Ackerman

Pinnacle Actuarial Resources

shawnaa@pinnacleactuaries.com

(415) 692-0937