PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

Data Cleansing for Predictive Models: The Next Level

Roosevelt C. Mosley, Jr., FCAS, MAAA

CAS Ratemaking & Product Management Seminar

Philadelphia, PA

March 19 – 21, 2012

Data Cleaning

Data cleansing – the next level

- Why simple visualization may not tell the whole story

Data homogeneity

- There are distinct groups in your underlying data

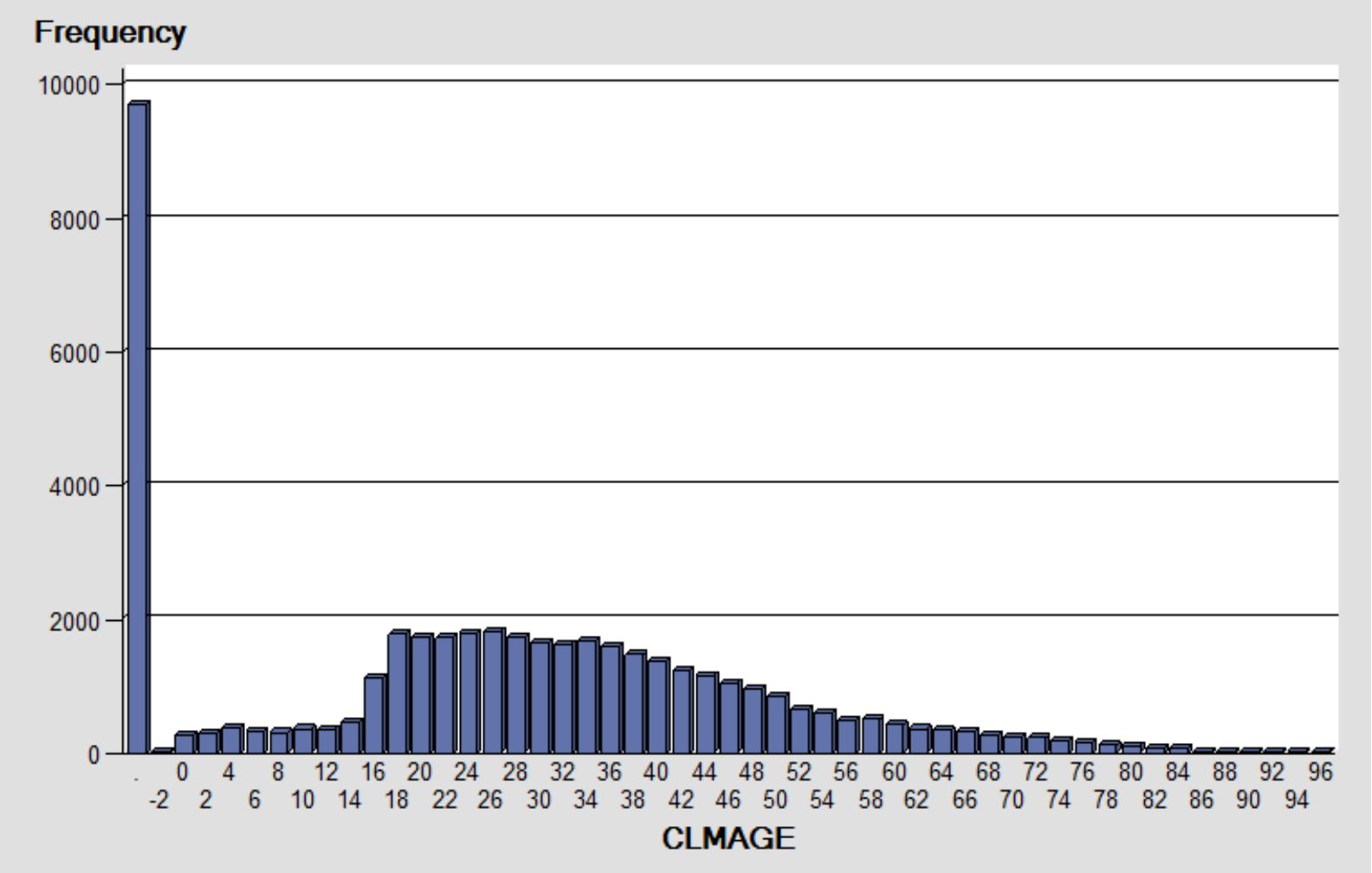
Multivariate data anomalies

- Certain combinations of variables may point to data issues

PINNACLE
ACTUARIAL RESOURCES, INC.
The Firm of Choice

Data Cleansing – The Next Level

Data Validation – One and Two Way Summaries



Data Cleansing – the Next Level

- One and two way data summarization and visualization is **absolutely key** in determining that individual factors are valid
- In building predictive models, multivariate techniques consider independent variables simultaneously to account for dependencies
- Data issues don't just exist in one and two dimensions, they can exist in n dimensions (where n is the number of individual elements)
- **Underlying causes**: heterogeneity, data anomalies
- Multivariate data exploration techniques can be used to address these issues

Data Homogeneity

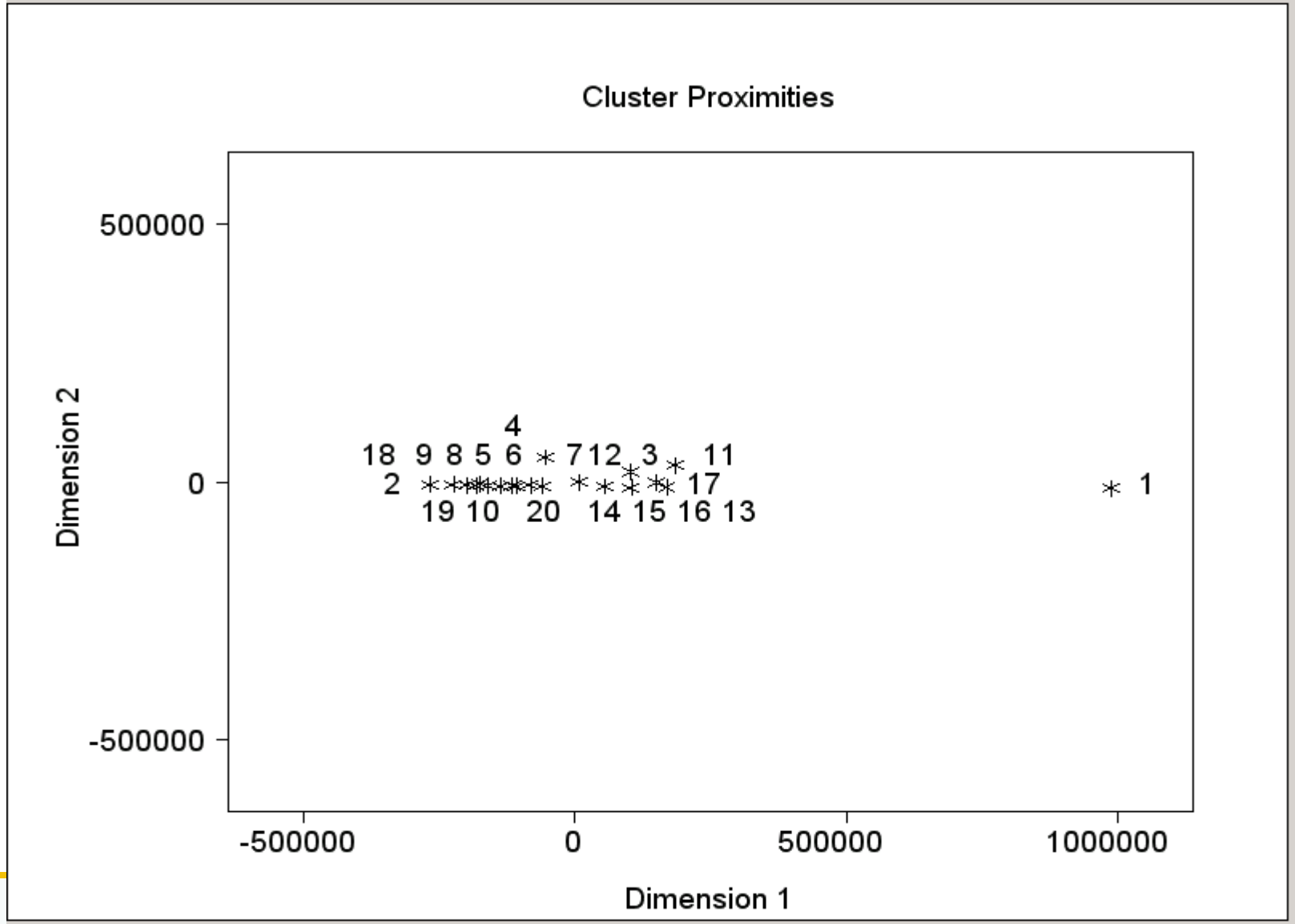
Clustering/Segmentation

- Unsupervised classification technique
- Groups data into set of discrete clusters or contiguous groups of cases
- 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
- Can be used as a dimension reduction technique

Example

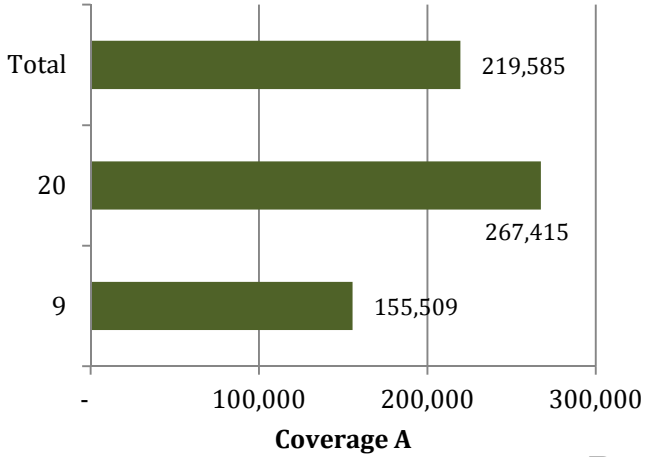
- Homeowners dataset
- Ran clustering analysis using key risk characteristics
  - Amount of insurance
  - Age of home
  - Billing option
  - Construction
  - Protection class
  - Deductible
  - Multiline
  - State/territory
- Developed predictive model on clusters independently

Cluster Distance Map

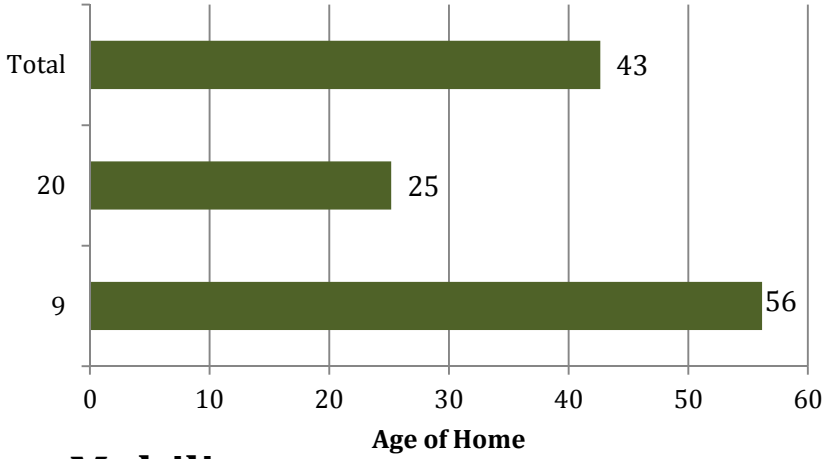


Cluster Characteristics

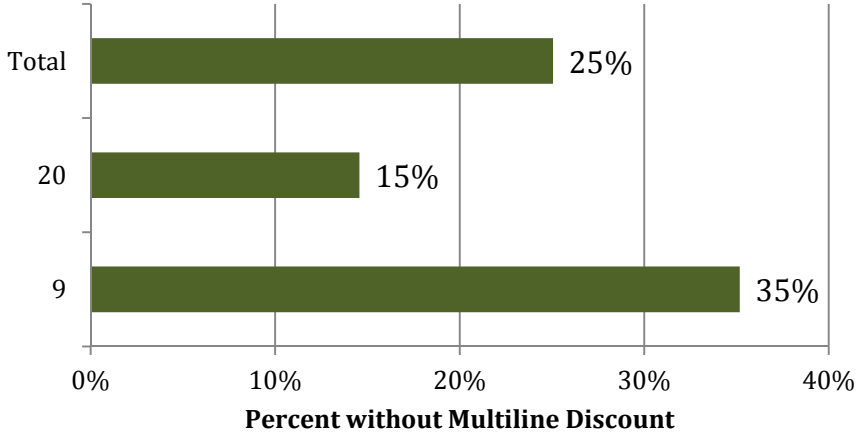
Coverage A



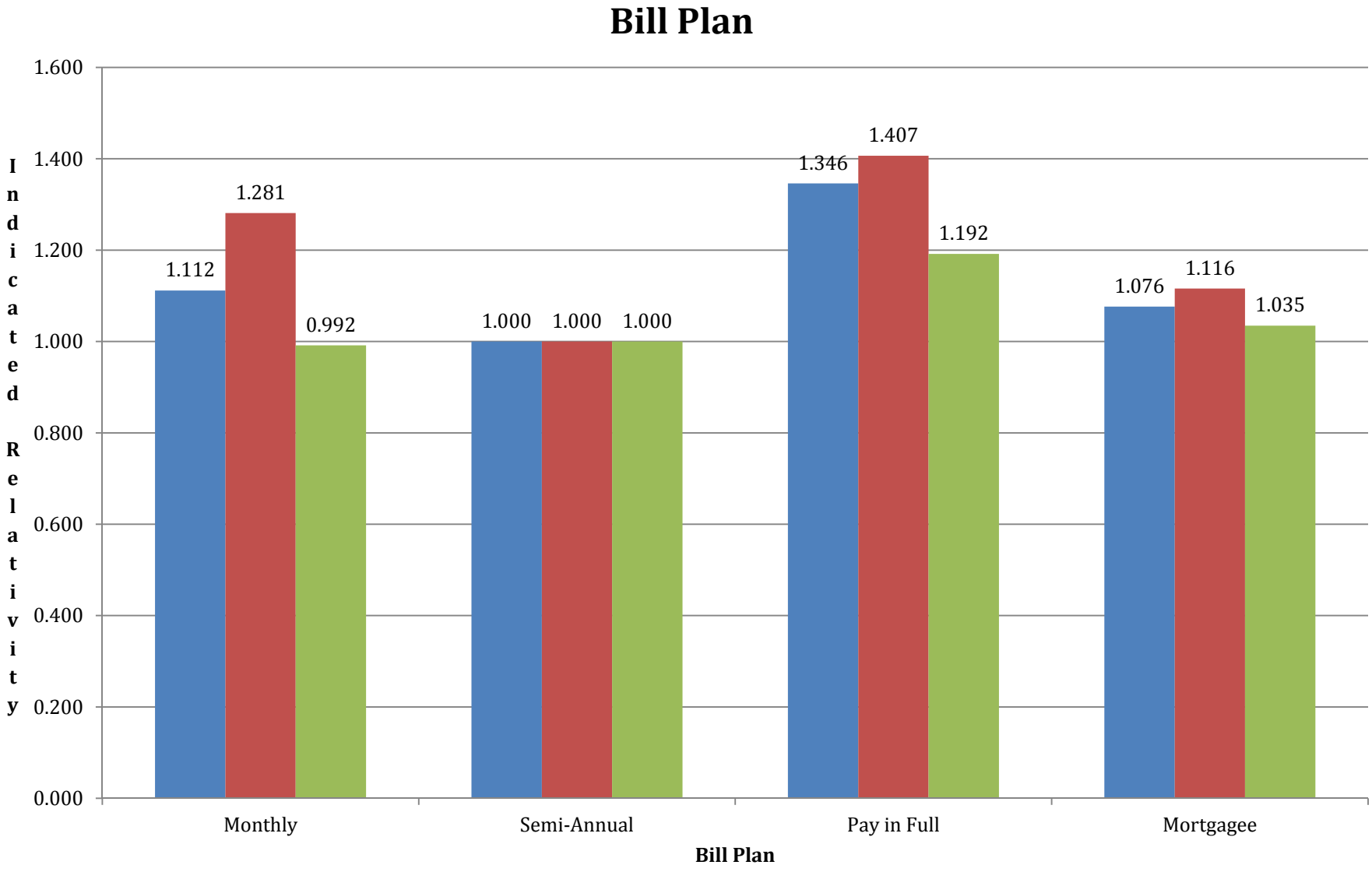
Age of Home



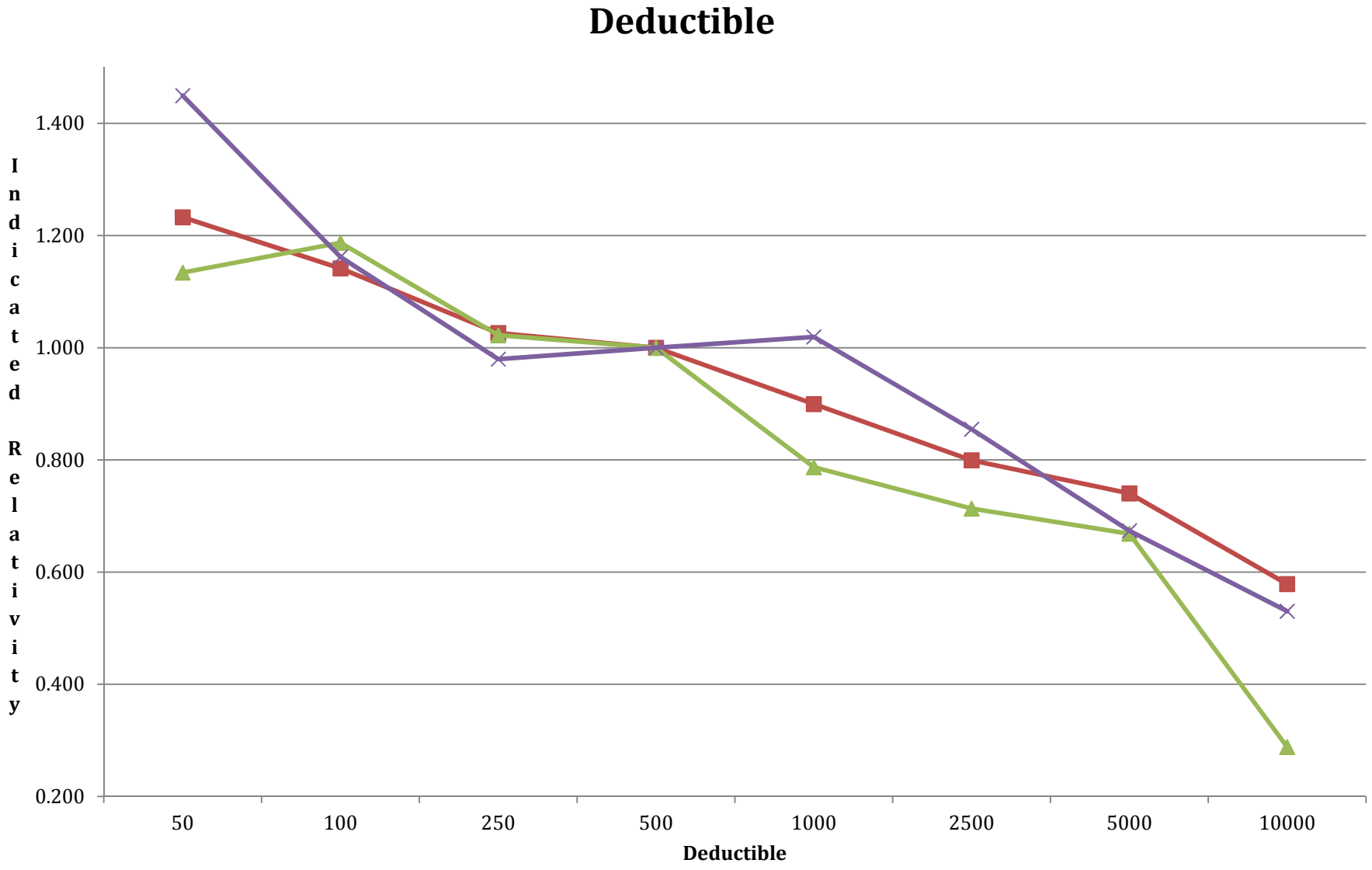
Percent without Multiline Discount



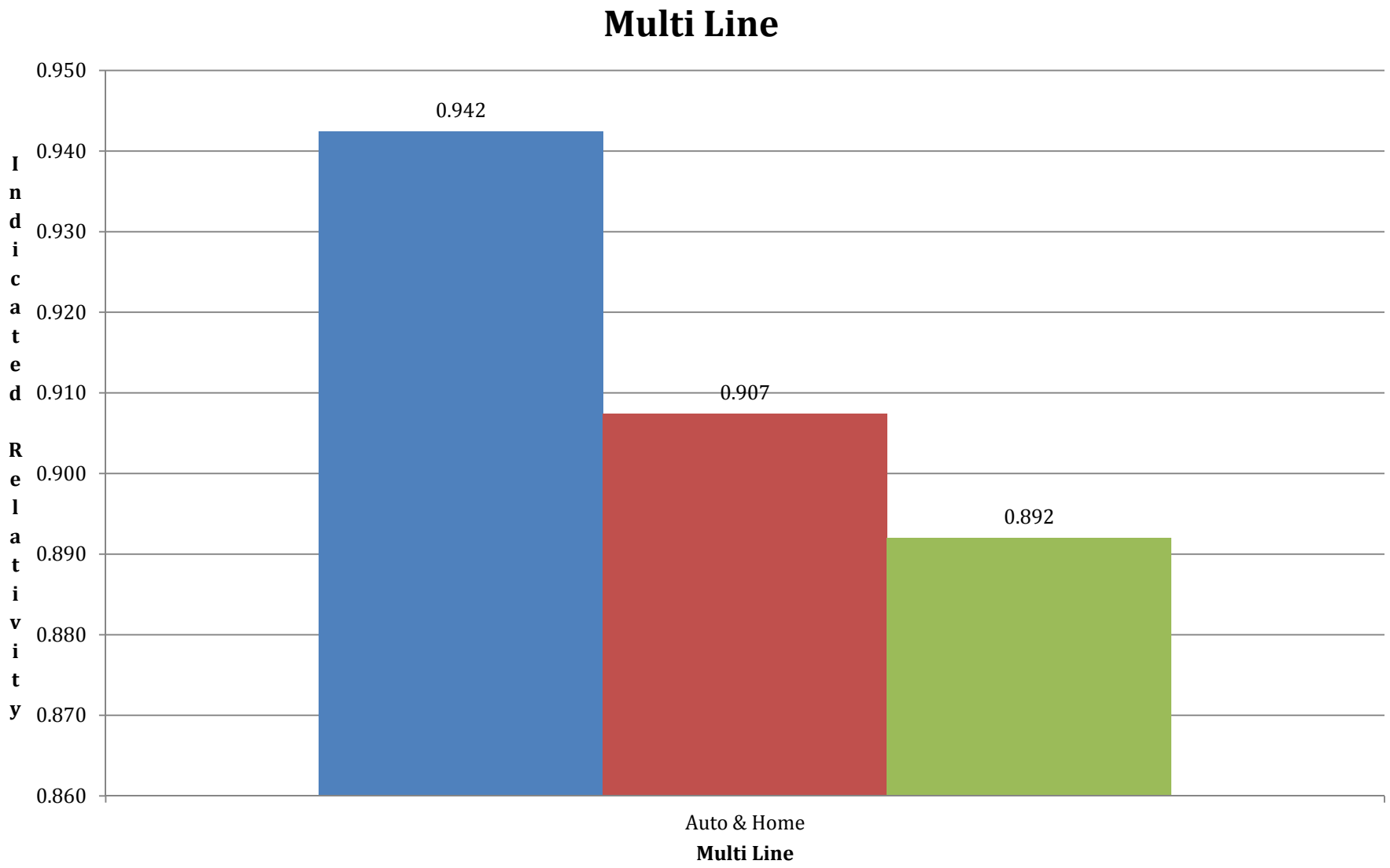
Billing Plan Indications



Deductible Indications



Multi-Line Indications

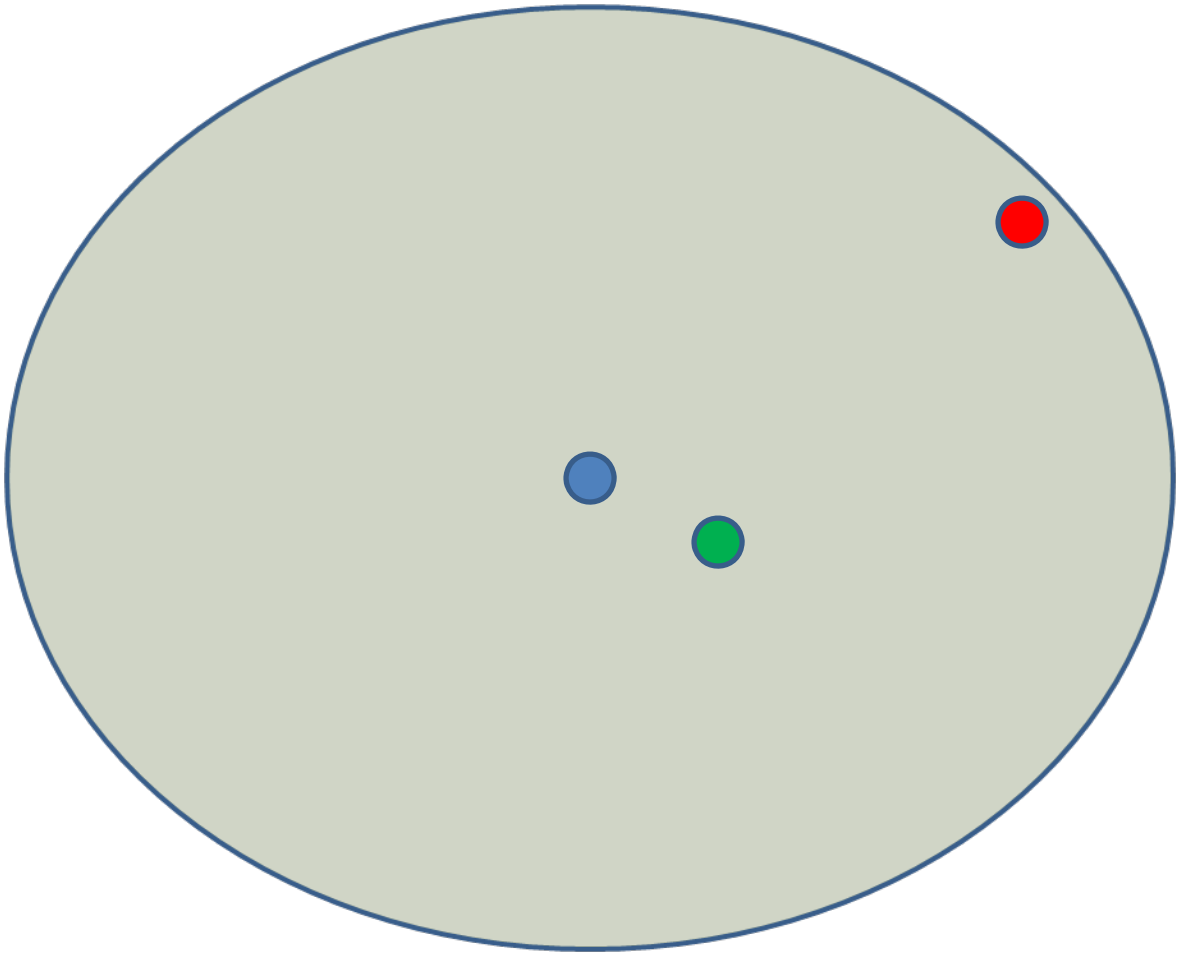


Multivariate Data Anomalies – Back to Cluster 1

	Cluster 1	Total
Average Amount of Insurance	$1,109,048	$219,585
Average Age of Home	19.6 years	42.7 years
Percentage of Deductibles > $2500	19.9%	1.9%

- Higher value homes
- Segment of the business that is certainly heterogeneous – will behave differently that overall population
- Represents 0.2% of the overall exposures
- Should we exclude data points such as these?

Outlier Data Points



Midpoint of the cluster, represents an average risk for that cluster

Risk that is slightly different than average, but still fits well with that cluster

Potential anomaly – data point fits best within this cluster but is actually an outlier for the cluster. This generally means it doesn't fit well anywhere.

Data "Cleanup"

- Reflect heterogeneity in final product (rating plan adjustments, underwriting, tiering)
- Data verification
- Modify data
- Exclude data