The Rest of the Story: Applications and Practical Considerations of GLM & Predictive Modeling

Fun with Factors

CAS Ratemaking Seminar 2005
New Orleans Marriott

Raw Variable Data

Types of Variables

- Continuous
  - Numerical Values or Statistics (Age, Miles, etc)
- Categorical
  - Discrete Classes or Groups (Symbols, Gender, etc)
- Ordinal
  - Ranks or Scores
- Spatial/Temporal
  - Boundary or Point Spatial Data (Lat/Lon, Zip, etc)

Why create factors?

- Challenges
  - Low credibility in distinct classifications, either univariately or multivariately
  - Sparse values across range of continuous or ordinal values
  - Unmanageable number of distinct classifications

- Solutions
  - Split or combine variable(s) to create new groupings which are better suited to the GLM modelling process

Creating Predictive Factors

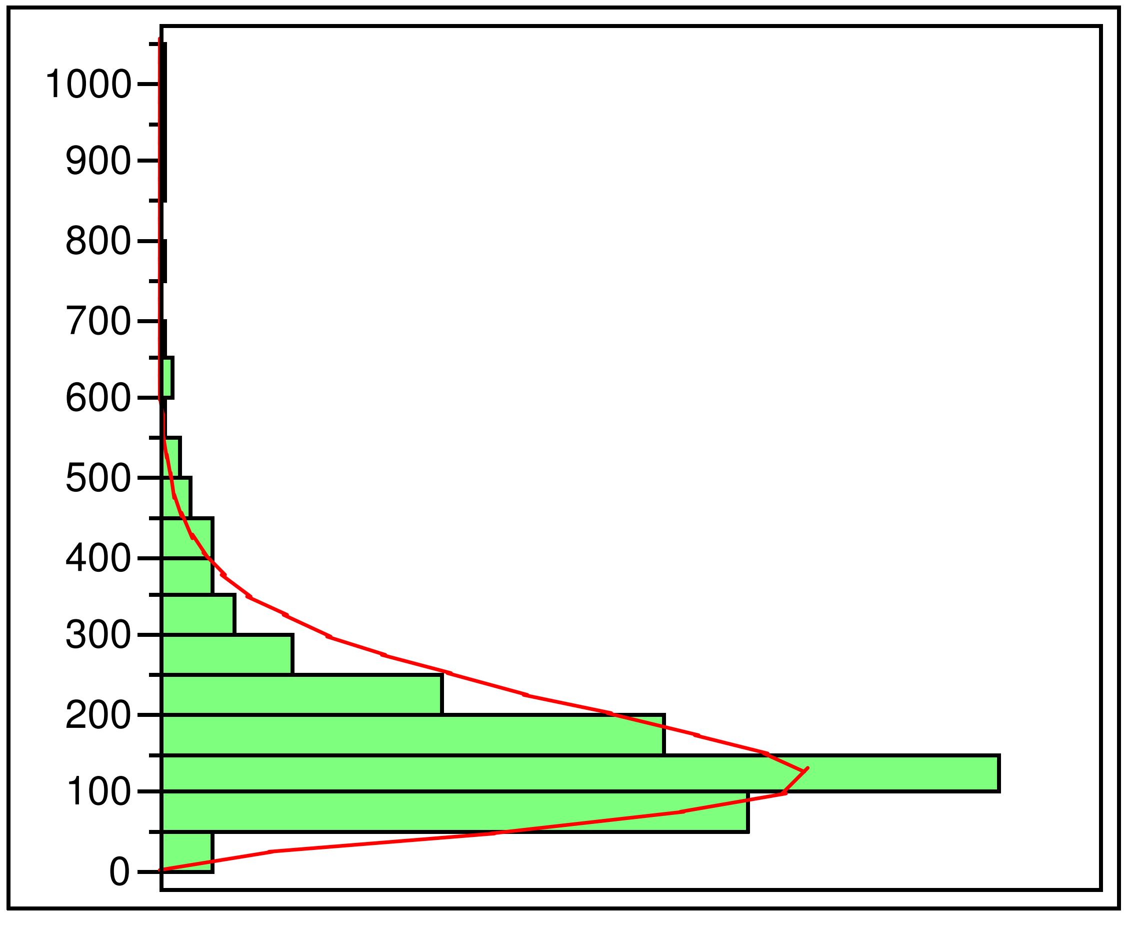
- Methods
  - Ad Hoc
    - EDA to manually combine factor levels
  - Decision Trees/Partitioning
    - Tree based methods to split or combine variables
    - Use either leaf as final classification, or splits and combines from tree
  - Cluster Analysis
    - Combine level(s) of factors(s) to form new groupings
  - Genetic Algorithms
    - Genetic rules select “best” combination of splits or combines

Examples

- Fundamentals
  - Any type of dependent variable (loss, frequency, severity, response) can be used
  - Can accommodate measures of credibility appropriate to method
- Data
  - Personal Auto Loss Data
  - Modeling Losses
  - Multiple potential predictive variables
  - Low multivariate credibility and predictive power

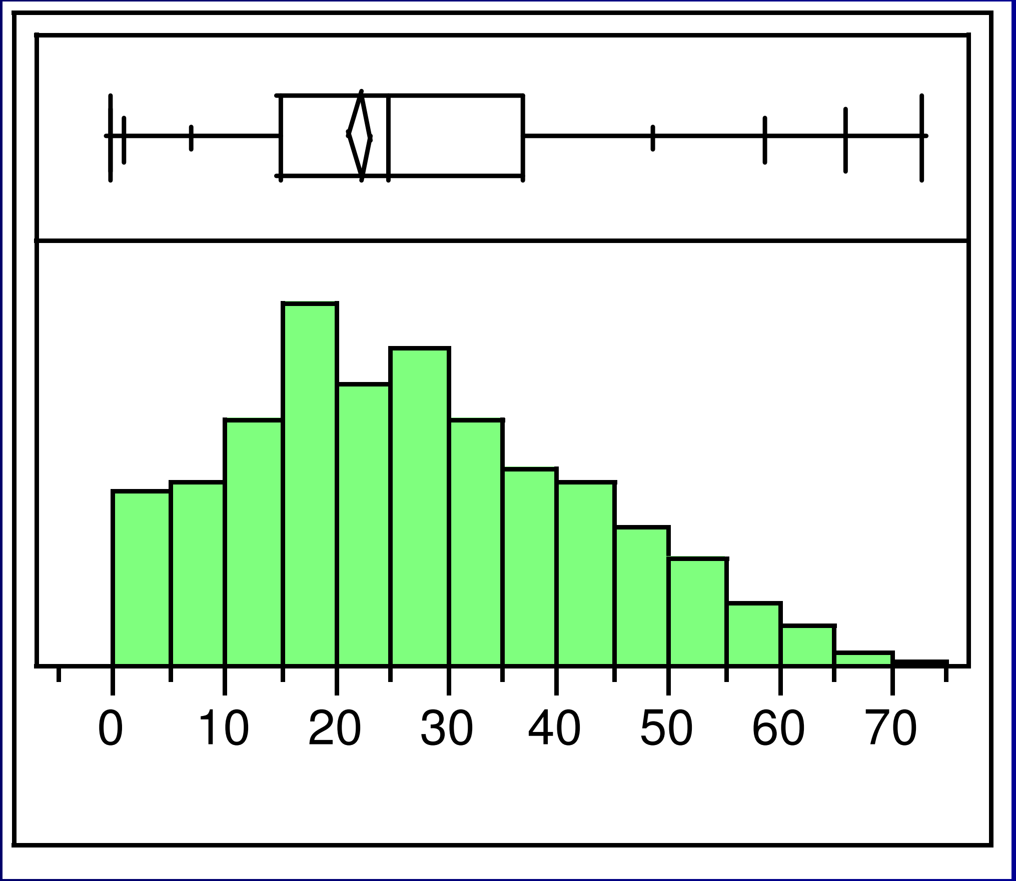
Loss Distribution

- Gamma distributed Auto losses

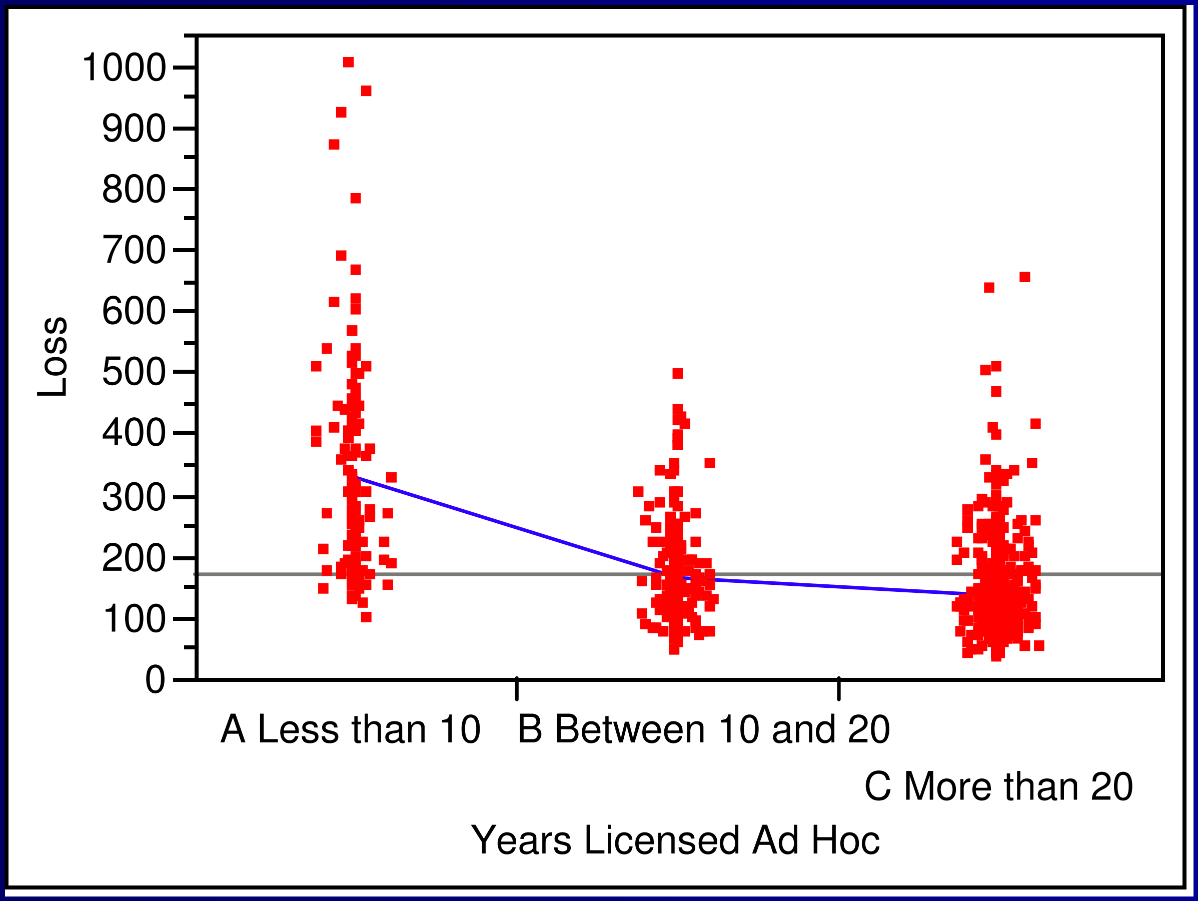


Continuous Variable: Years Licensed

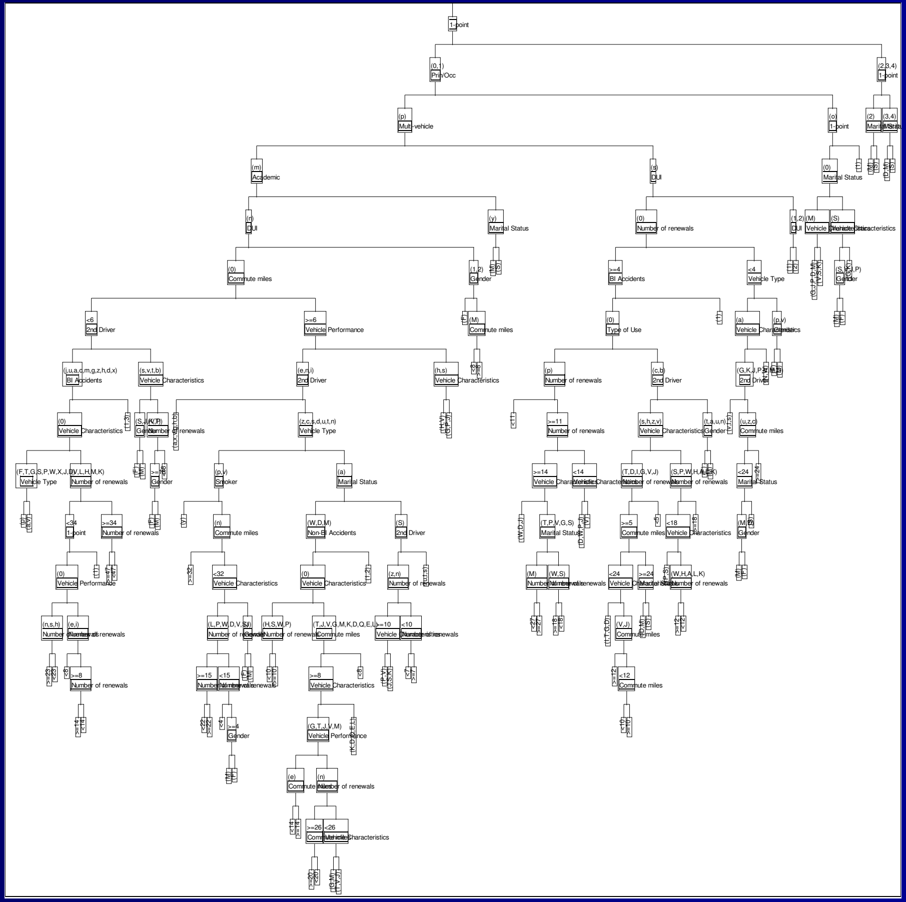
Min=0, Max=73, Mean=22.4, Median=25



Ad Hoc Categorization

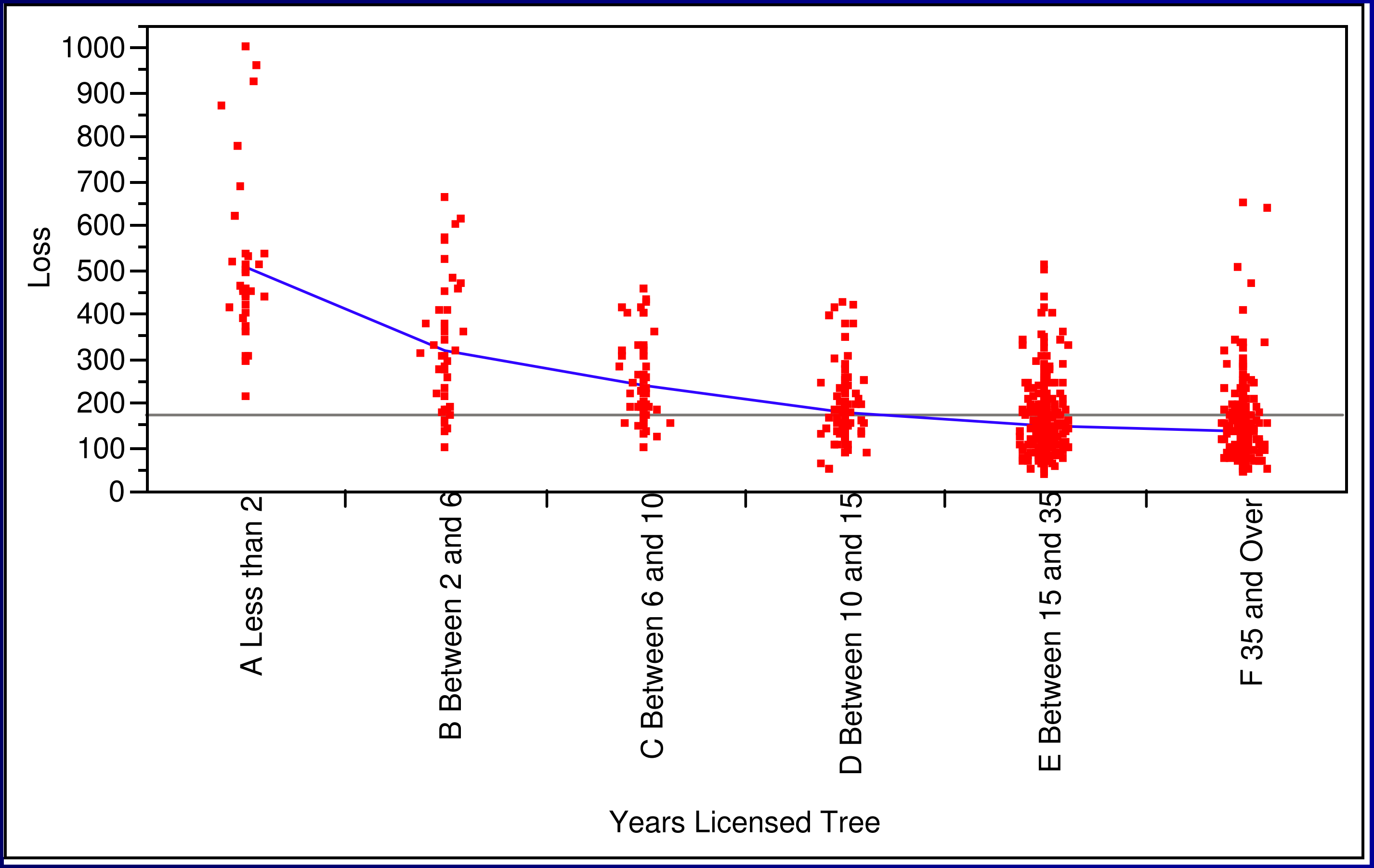


Partition – Multivariate with GDD & Miles

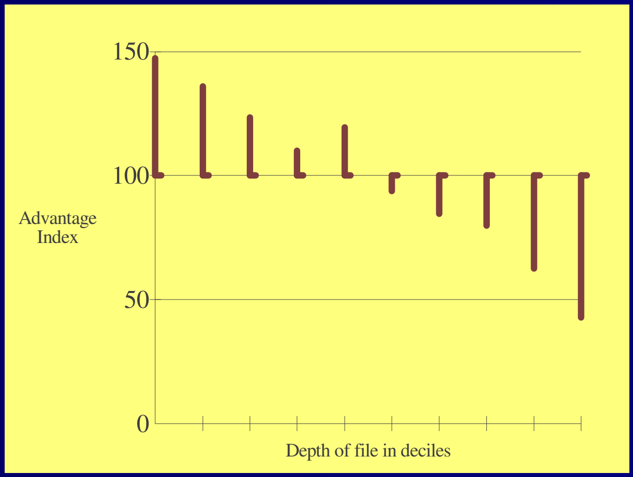


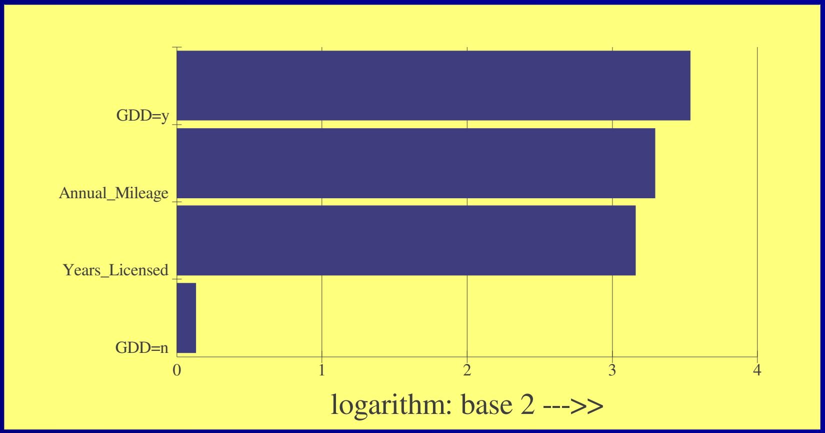
ChoicePoint

Partition Categorization



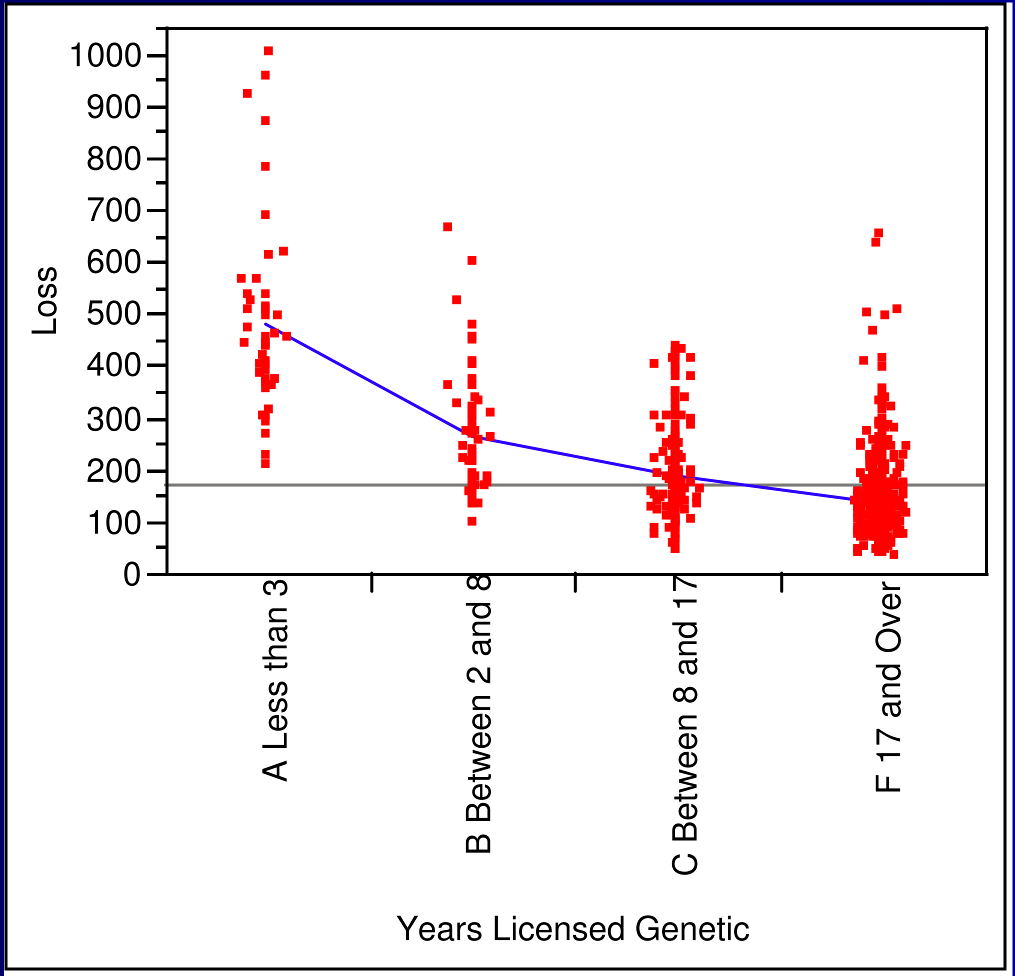
Genetic – Multivariate with GDD & Miles



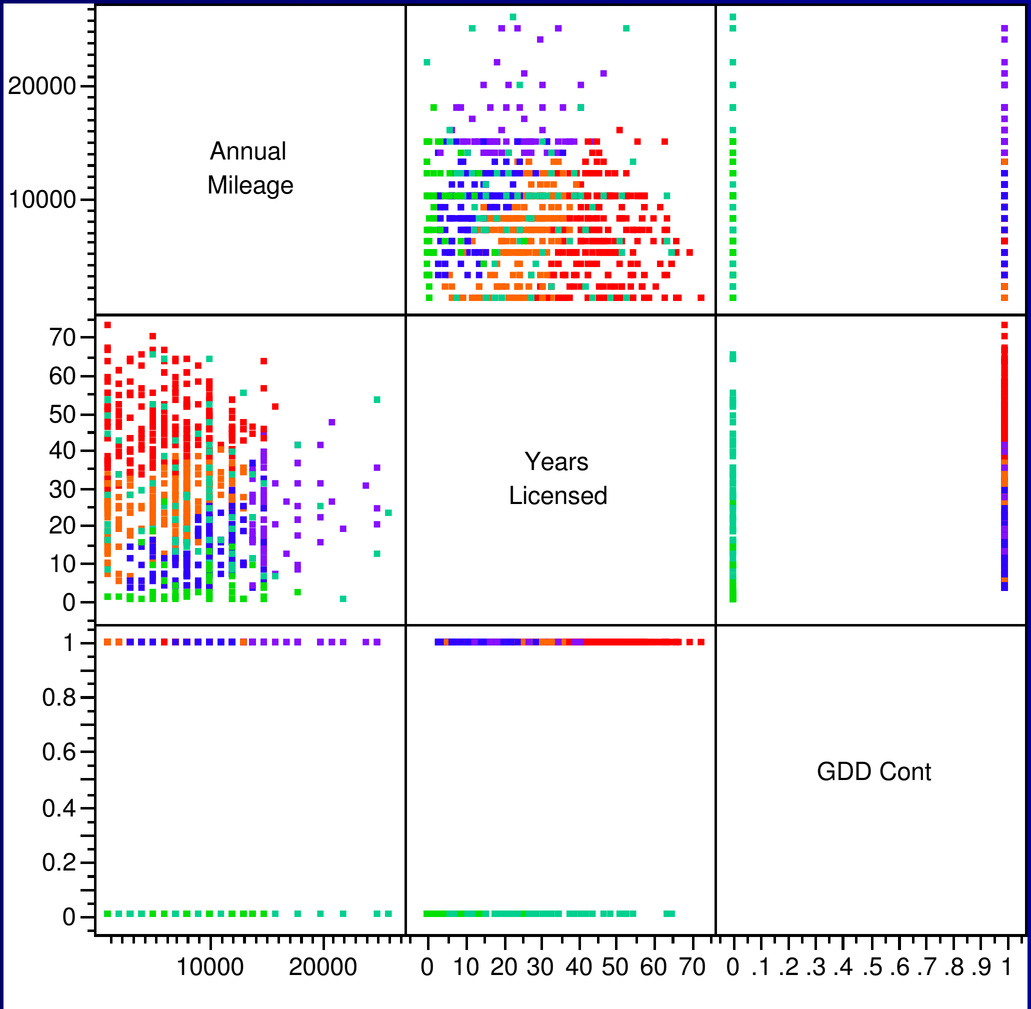


	Predicted Profit		Avrg Profit		Cum. Avrg	
	Cum. Gain		Min score		Max score	
Top	1	0.01	0.01	141	1.006	1.067
2nd	1	0.01	0.01	128	1.003	1.006
3rd	1	0.01	0.01	114	1.001	1.003
Bottom	1	0.00	0.01	100	0.000	1.001

Genetic Categorization

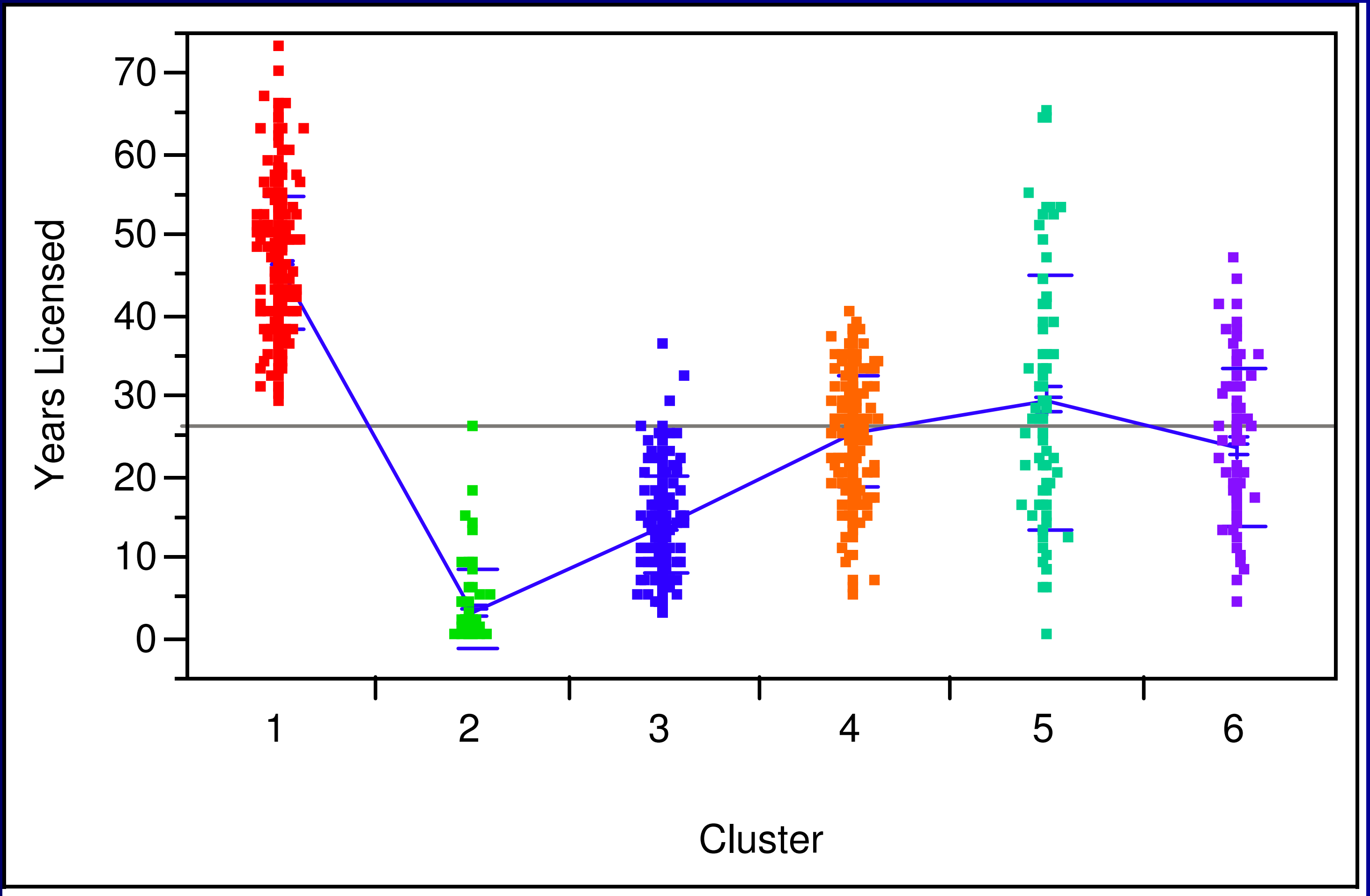


Cluster – Multivariate with GDD & Miles



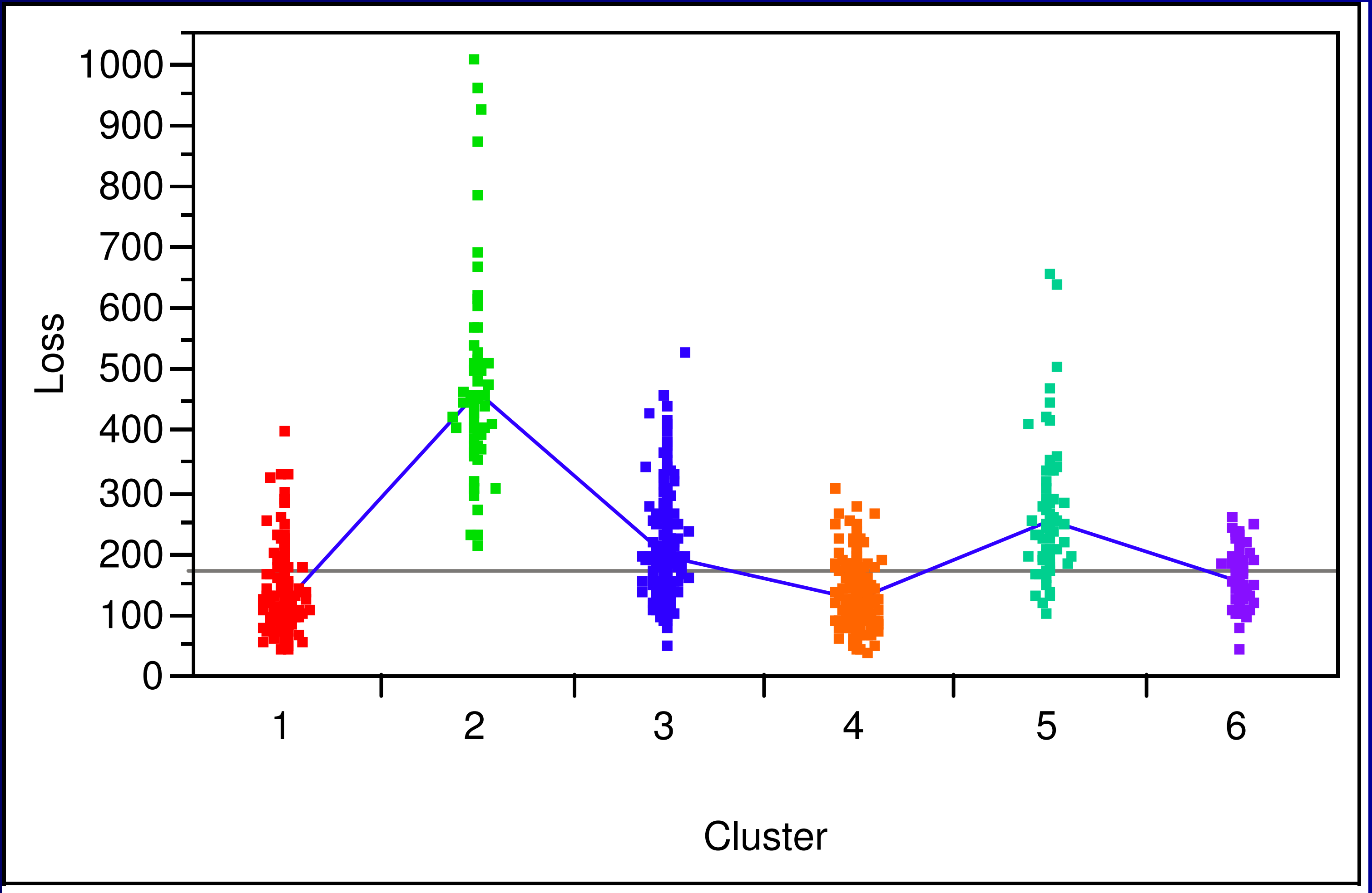
Cluster View

- Wards



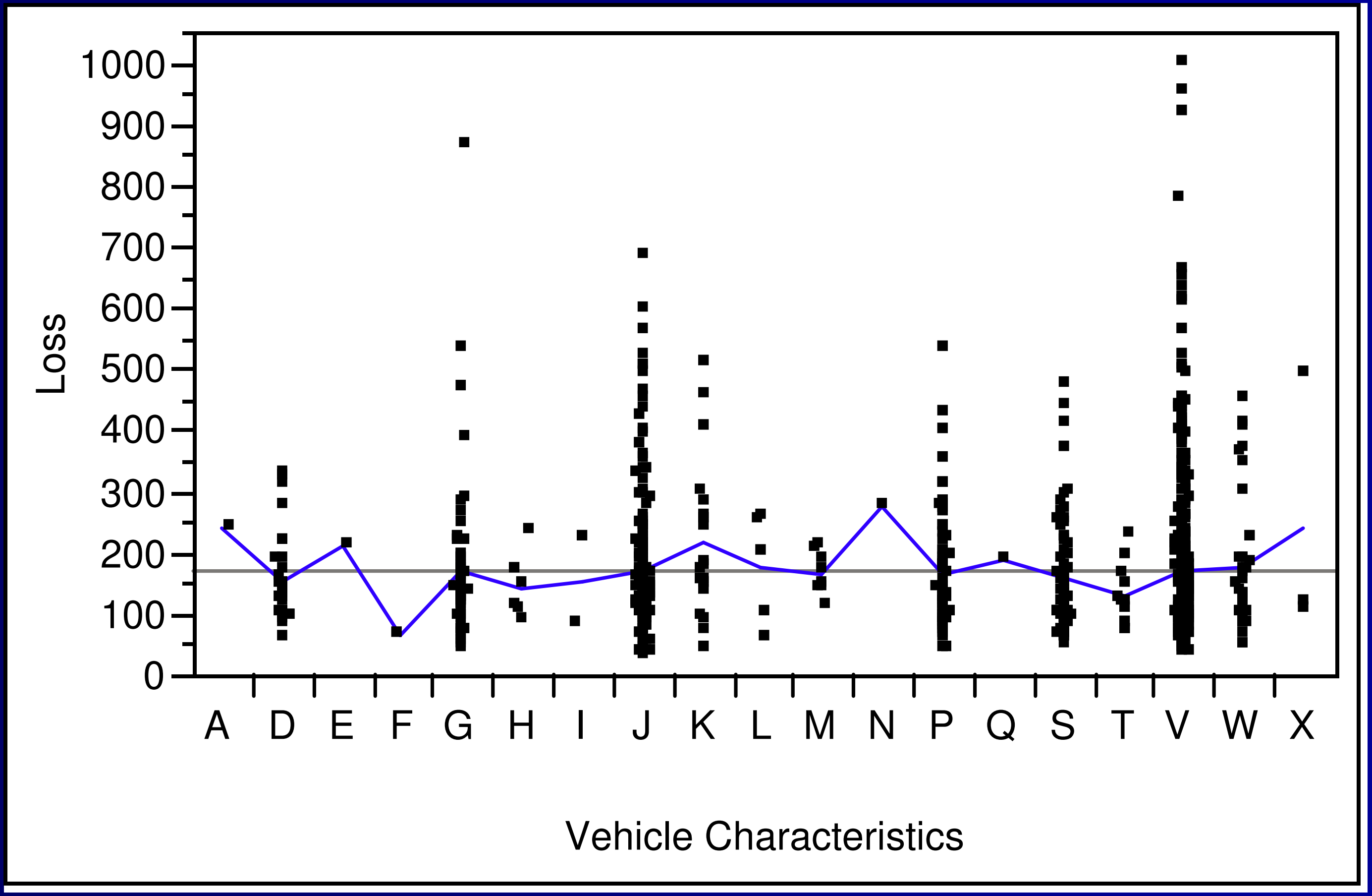
ChoicePoint

Cluster Categorization



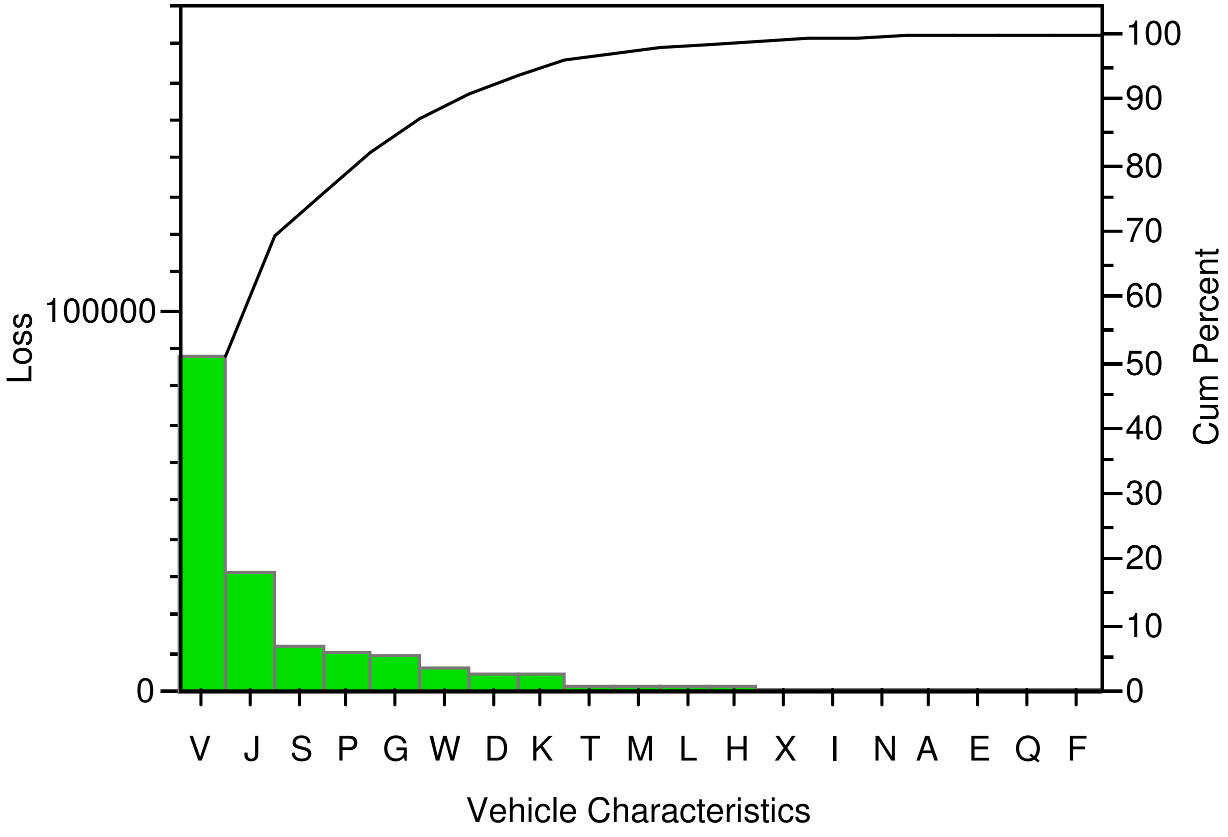
ChoicePoint

Categorical Variable: Vehicle Characteristics

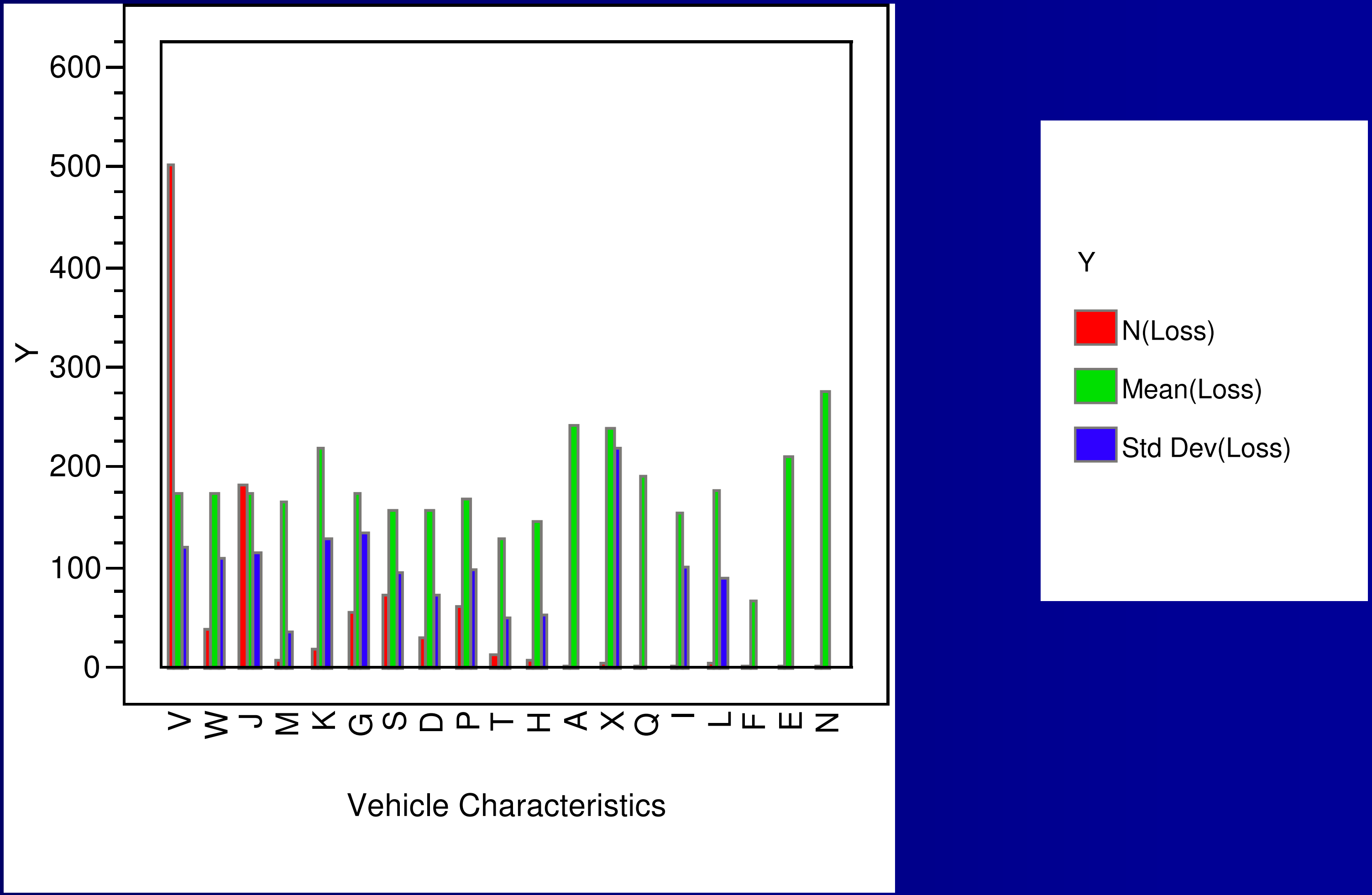


ChoicePoint

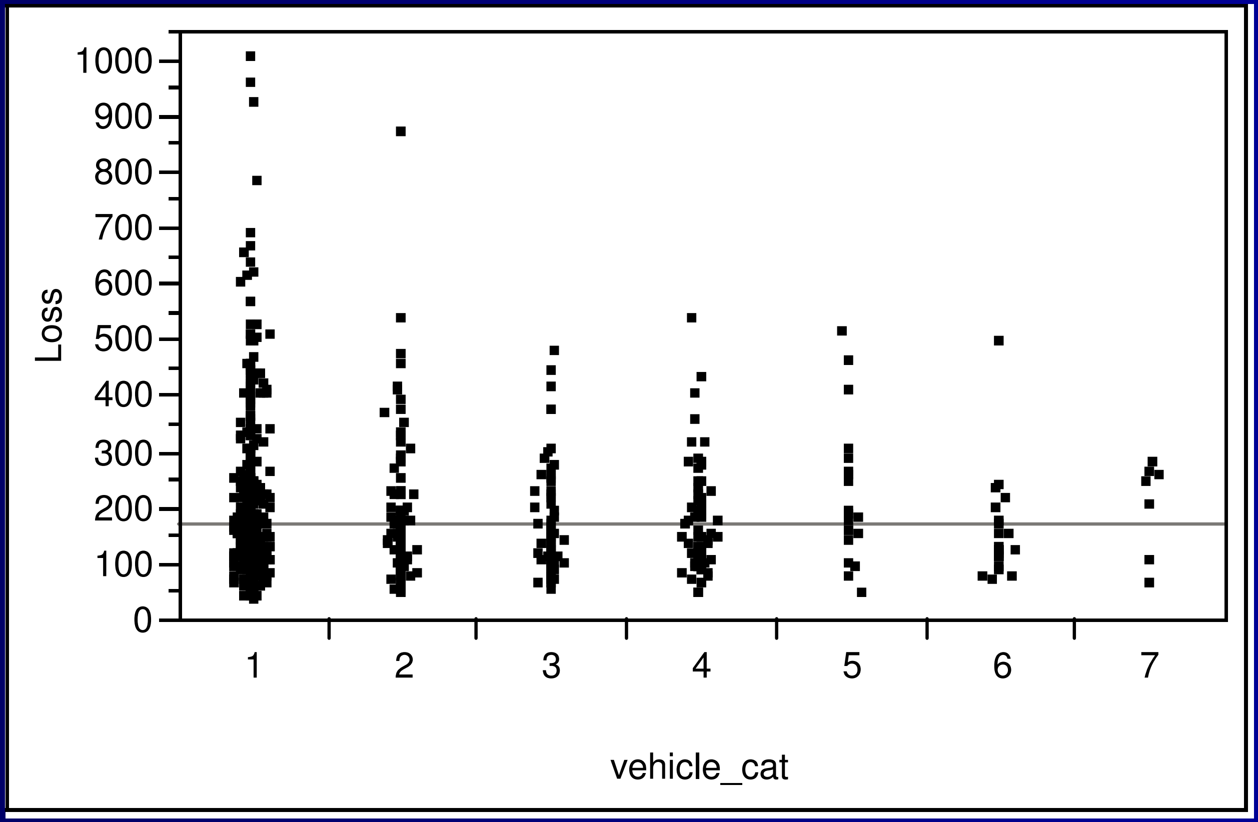
Ad Hoc



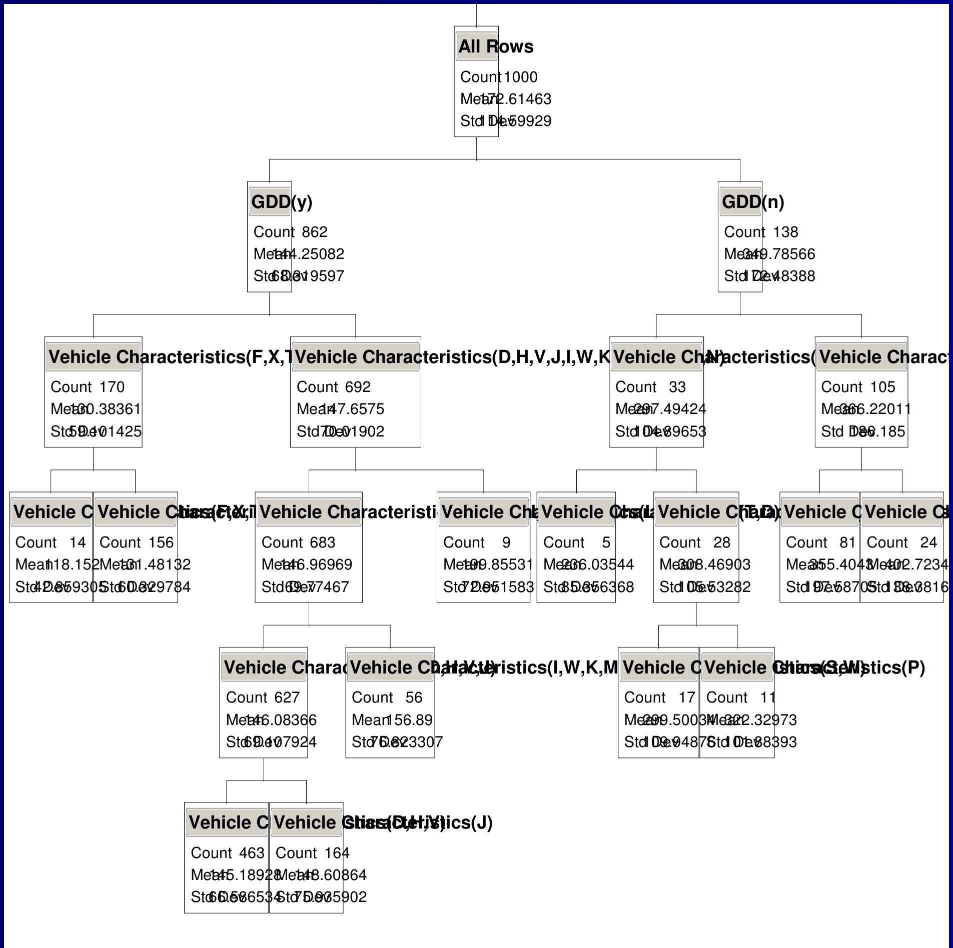
Empirical View



Ad Hoc Classification



Partition – Multivariate with GDD



All Rows
Count 1000
Mean 172.61463
Std Dev 104.59929

GDD(y)
Count 862
Mean 144.25082
Std Dev 68.819597

GDD(n)
Count 138
Mean 349.78566
Std Dev 172.48388

Vehicle Characteristics(F,X,T
Count 170
Mean 130.38361
Std Dev 59.101425

Vehicle Characteristics(D,H,V,J,I,W,K
Count 692
Mean 147.6575
Std Dev 70.01902

Vehicle Characteristics(N)
Count 33
Mean 297.49424
Std Dev 104.89653

Vehicle Charact
Count 105
Mean 366.22011
Std Dev 186.185

Vehicle C
Count 14
Mean 118.152
Std Dev 42.859305

Vehicle Characteristics(F,X,T
Count 156
Mean 131.48132
Std Dev 60.829784

Vehicle Characteristi
Count 683
Mean 146.96969
Std Dev 69.77467

Vehicle Ch
Count 9
Mean 199.85531
Std Dev 72.951583

Vehicle Characteristics(L
Count 5
Mean 236.03544
Std Dev 85.356368

Vehicle Characteristics(T,D)
Count 28
Mean 308.46903
Std Dev 105.53282

Vehicle C
Count 81
Mean 355.4043
Std Dev 197.5870

Vehicle C
Count 24
Mean 402.7234
Std Dev 138.3816

Vehicle Characteristics(D,H,V,J)
Count 627
Mean 146.08366
Std Dev 69.107924

Vehicle Characteristics(I,W,K,M
Count 56
Mean 156.89
Std Dev 75.823307

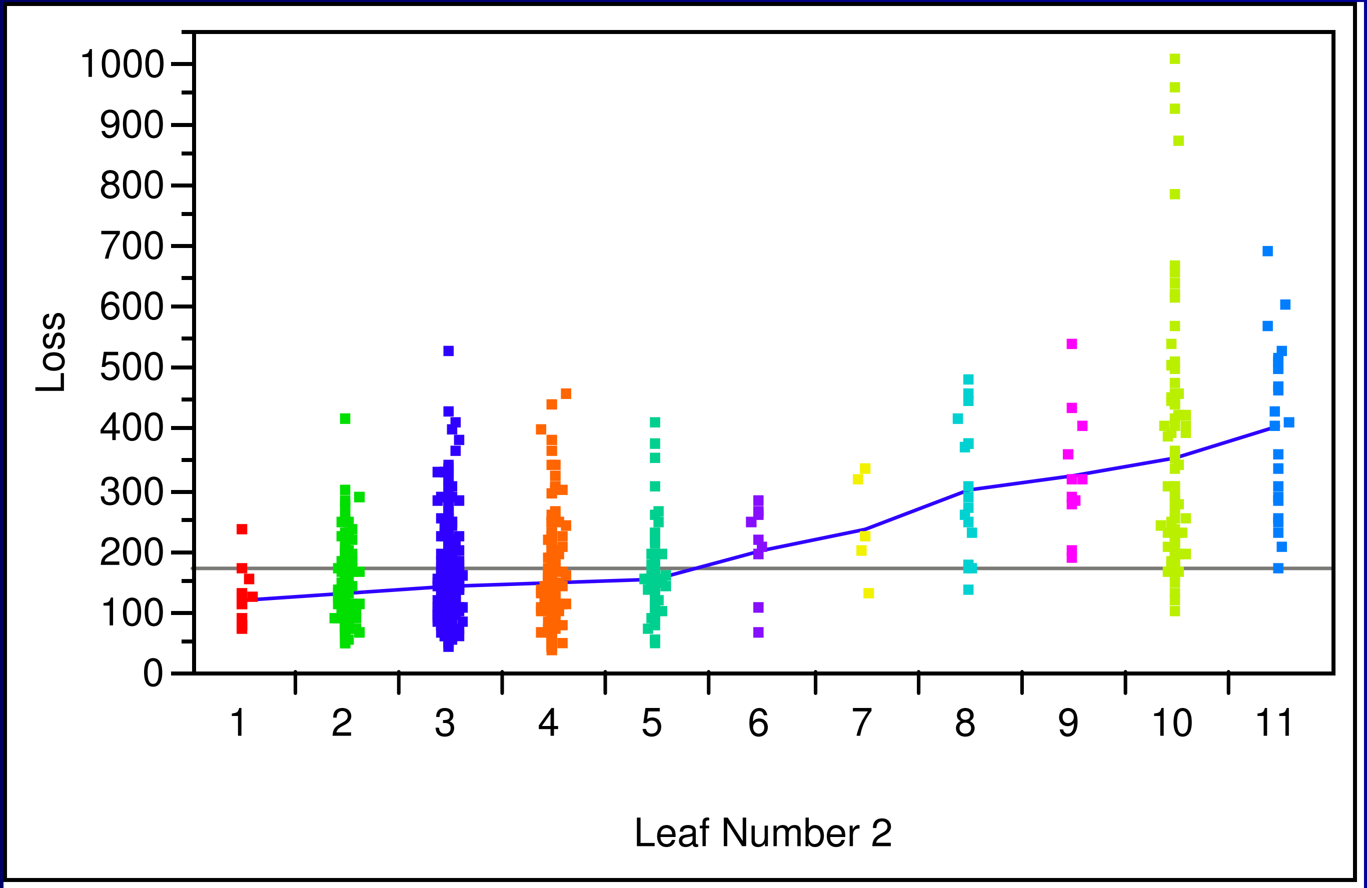
Vehicle Characteristics(S,W)
Count 17
Mean 299.50034
Std Dev 109.9487

Vehicle Characteristics(P)
Count 11
Mean 322.32973
Std Dev 101.88393

Vehicle Characteristics(D,H,V)
Count 463
Mean 145.18928
Std Dev 66.586534

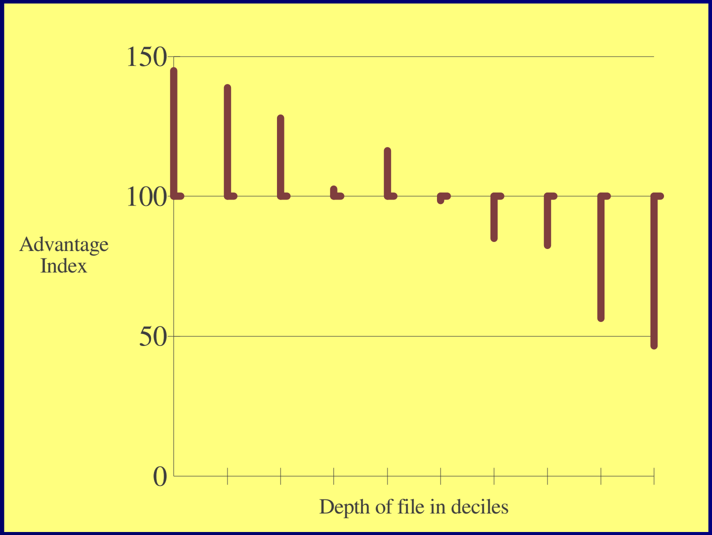
Vehicle Characteristics(J)
Count 164
Mean 148.60864
Std Dev 75.935902

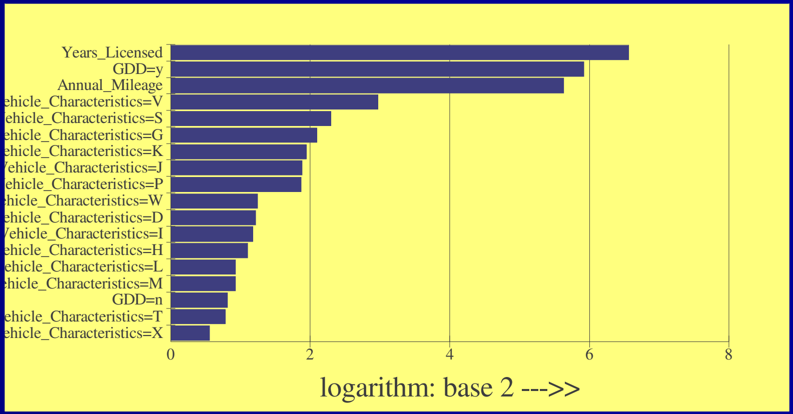
Partition Classification



Genetic – Multivariate

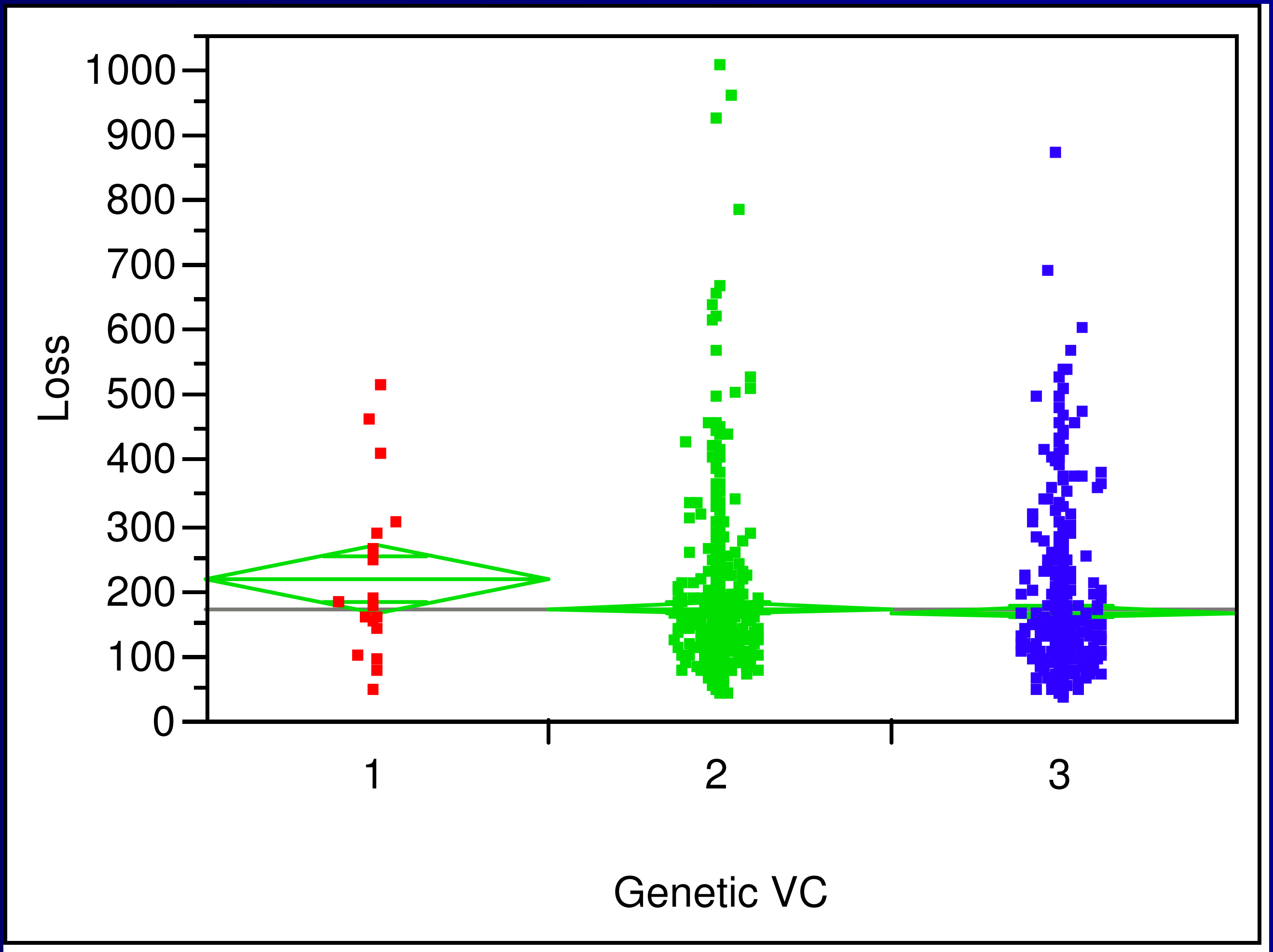
- GDD, Annual Miles, Years Licensed





	Predicted Profit		Avrg Profit		Cum. Avrg	
	Cum. Gain		Min score		Max score	
Top	1	0.01	0.01	140	0.0098	0.1258
2nd	1	0.01	0.01	126	0.0052	0.0098
3rd	1	0.01	0.01	114	0.0023	0.0052
Bottom	1	0.00	0.01	100	0.0000	0.0023

Genetic Classification



Methods for Other Factor Types

- Spatio-Temporal
  - Correlated in one or two dimensions
  - Spatial smoothing
  - Geostatistical methods (Variogram modeling, kriging)
- Ordinal
  - Specialized Methods
  - Marketing Research Preference Studies

Final Notes

- Model continuous variables if sufficient range and credibility
- Check at least bivariate with best predictor prior to modeling
- Best modeling practices and full model evaluation

Software Used

- SAS/JMP (Graphics, Partition, Cluster)
  - http://www.jmp.com
- ANGOSS Knowledgeseeker (Partition)
  - http://www.angoss.com
- SAS/STAT (Data manipulation, GLM, Cluster)
  - http://www.sas.com
- Minetech GMax (Genetic Selection)
  - http://www.minetech.com

Contact Information

Kate Phinney
Insurity
170 Huyshope Avenue
Hartford, CT 06106

860-616-7413
Kate.Phinney@Insurity.com

Thank you!