SCORECARD VARIABLE GROUPING AND SELECTION

Dr. Aric LaBarr
Institute for Advanced Analytics

Process Flow

Data Collection
- Variable Selection
- Sample Size
- Sample / Performance Window

Data Cleaning
- Eliminate Duplicates
- Examine / Remove Outliers

Variable Grouping and Selection
- Weights of Evidence (WOE)
- Information Value (IV)
- Gini Criterion

Initial Scorecard Creation
- Logistic Regression
- Accuracy
- Threshold
- Assessment

Reject Inference
- Remove bias resulting from exclusion of rejects

Final Scorecard Creation
- Final Model Assessment

VARIABLE GROUPING

Variable Grouping and Selection

- Scorecards end up with only just groups within a variable.
- Objectives:
  1. Eliminate weak characteristics (variables) or those that do not conform to good business logic.
  2. Group the strongest characteristics' attribute levels in order to produce a model in scorecard format.
- Interactive Grouping node in SAS EM.
- Function "smbinning" in R

Variable	Level
MISS	$x < 24$
MISS	$24 \leq x < 36$
MISS	$36 \leq x < 48$
MISS	$x \geq 48$
HOME	OWN
HOME	RENT

Why Grouping (Binning)?

- Goal is to help simplify analysis through grouping:
  - Useful for understanding relationships – no worries about explaining coefficients.
  - Modeling nonlinearities – similar to decision trees. (NO MORE LOGISTIC REGRESSION LINEARITY ASSUMPTION!)
  - Dealing with outliers – contained in the smallest / largest group.
  - Missing values typically in own group.

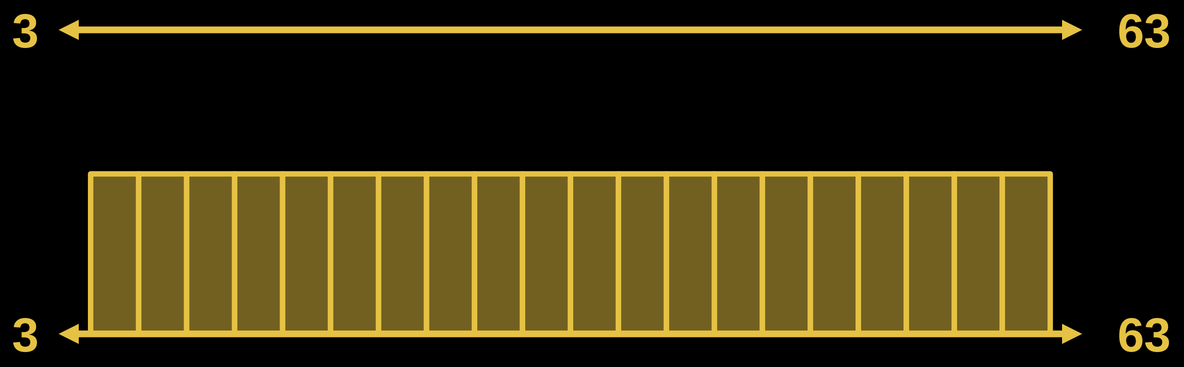
Initial Characteristic Analysis – SAS

- Need a starting point for the grouping / binning.
  - Quantiles are most popular technique.
- Pre-bin the interval variables into a number of user-specified quantiles / buckets for fine detailed groupings.
- Aggregate the fine detailed groupings into a smaller number to produce coarse groupings.
  - Chi-squared tests to combine groups.

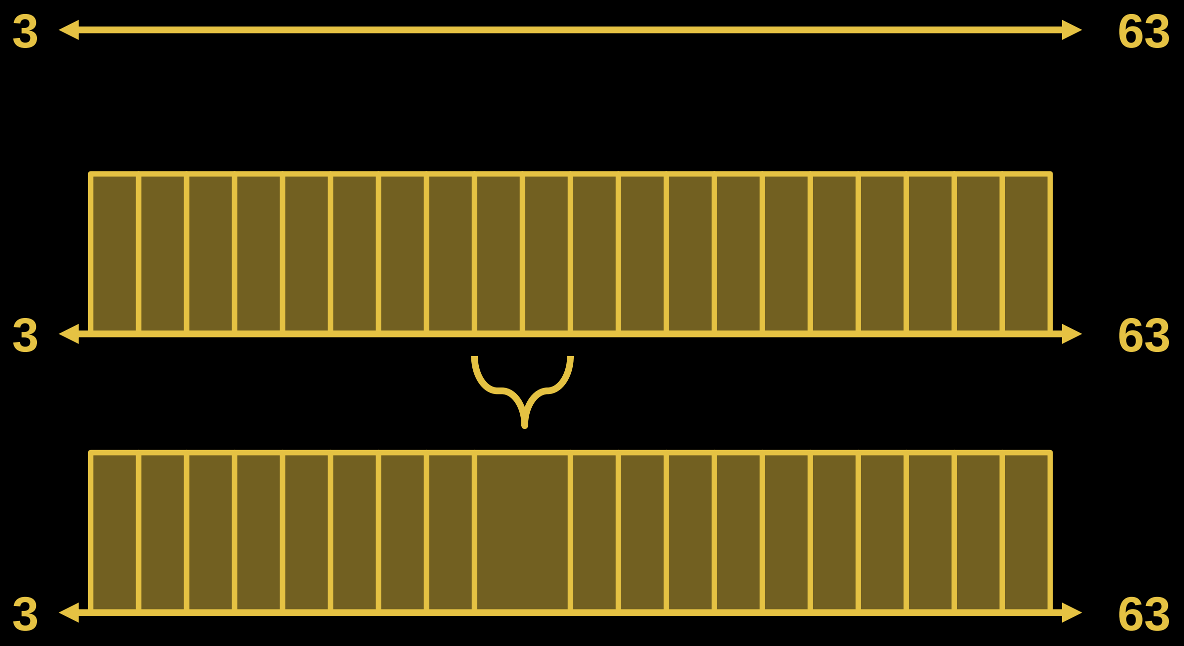
Initial Characteristic Analysis – SAS

3 ⟷ **63**

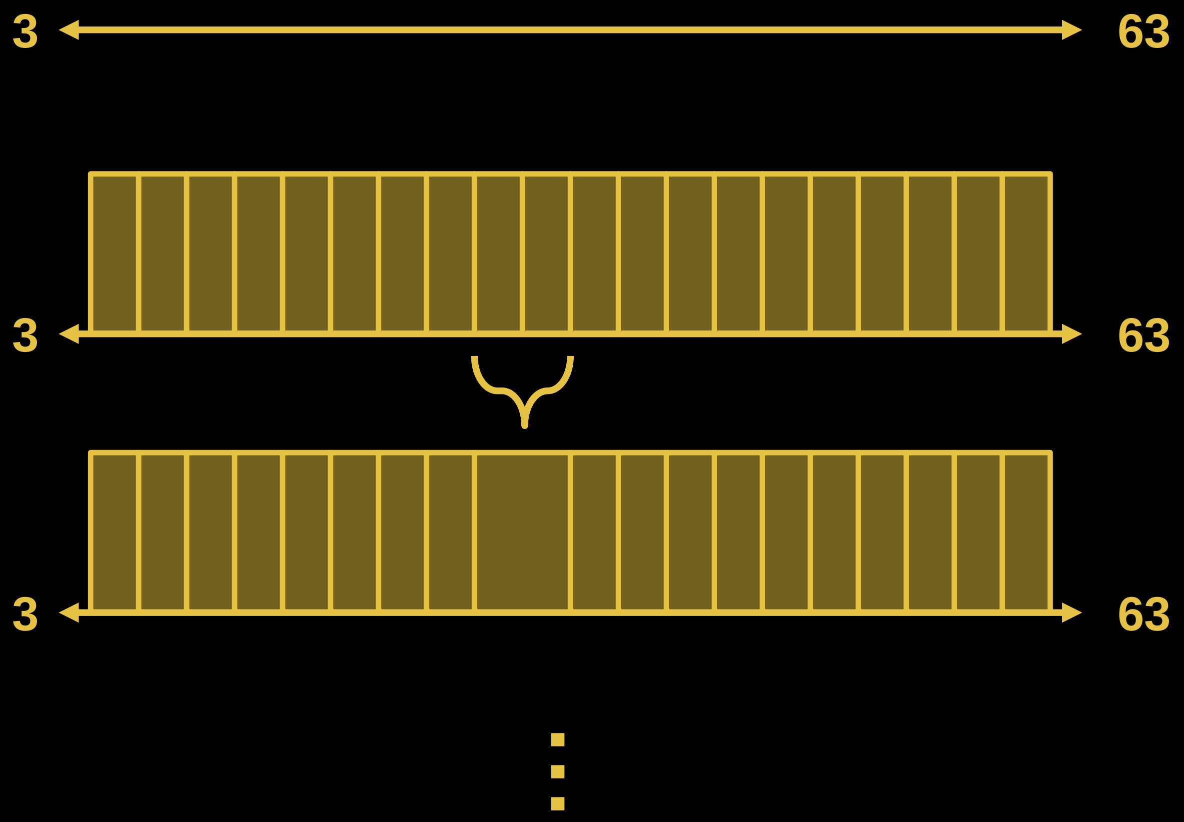
Initial Characteristic Analysis – SAS



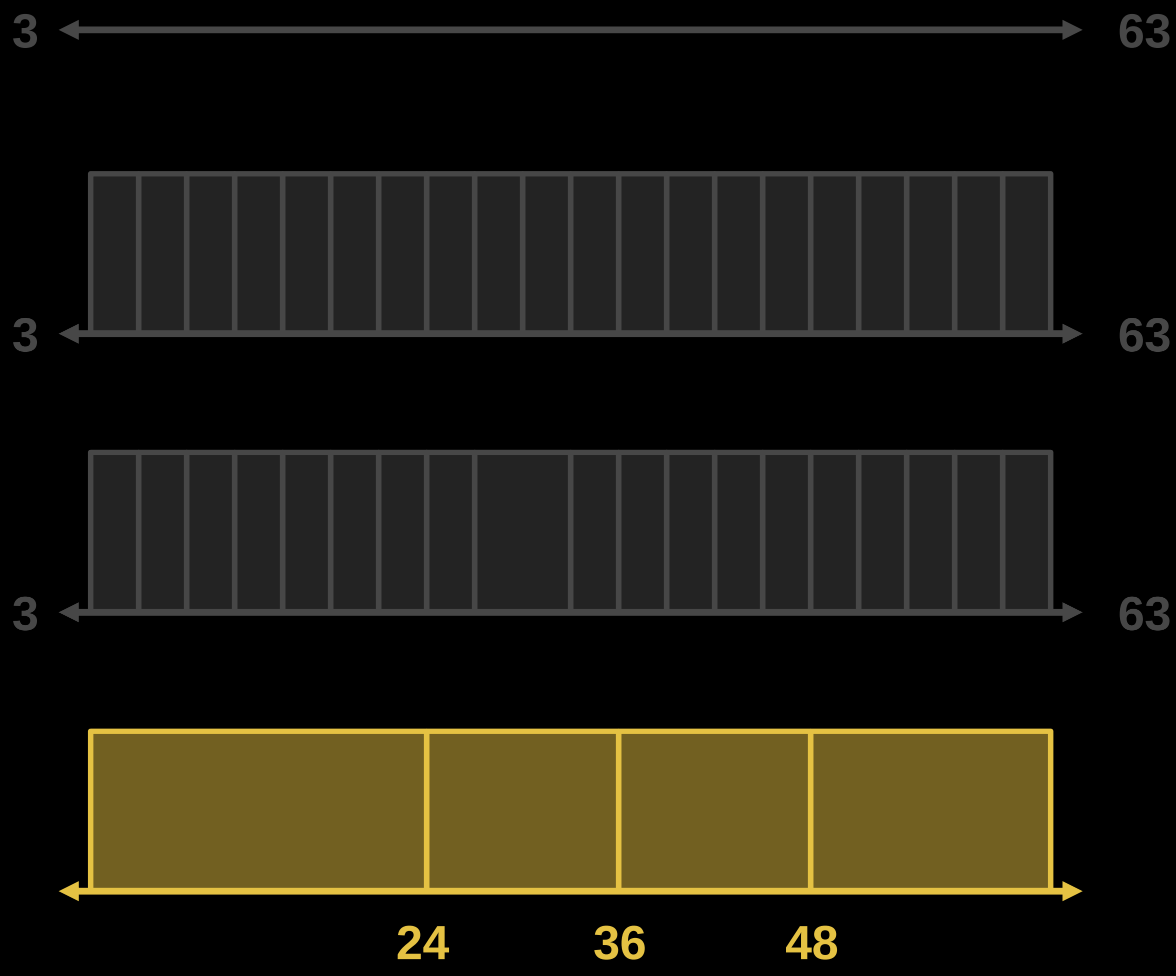
Initial Characteristic Analysis – SAS



Initial Characteristic Analysis – SAS



Initial Characteristic Analysis – SAS



Initial Characteristic Analysis – R

- The package (and function) “smbinning” uses a different approach than SAS.

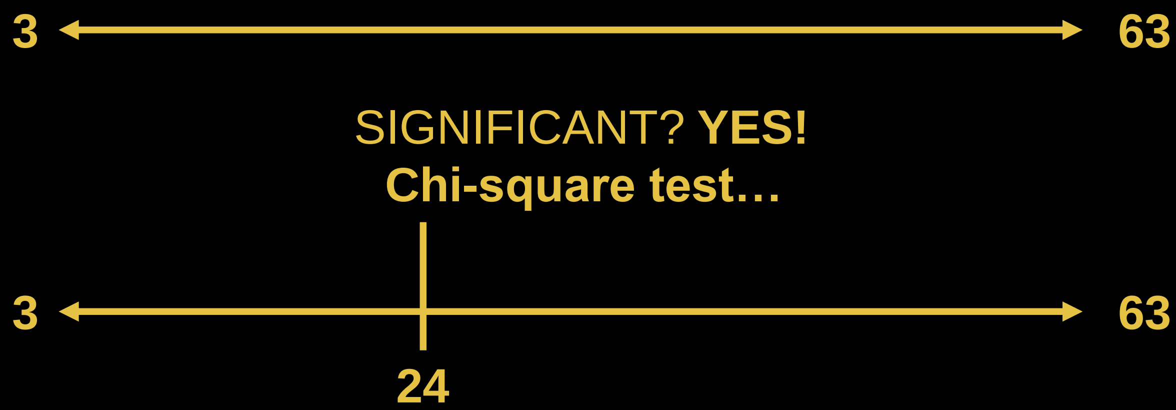
- Conditional Inference Trees:
  - CART methods have inherent bias – variables with more levels → more likely to be split on if split on Gini and Entropy.
  - CIT method adds extra statistical step before splits occur – statistical tests of significance.
  - What is **MOST** significant variable? → What is the best split (Chi-square) on **THIS** variable? → **REPEAT**.

Initial Characteristic Analysis – R

3 ⟷ **63**

SIGNIFICANT?

Initial Characteristic Analysis – R



Initial Characteristic Analysis – R

3 **63**

SIGNIFICANT? **YES!**

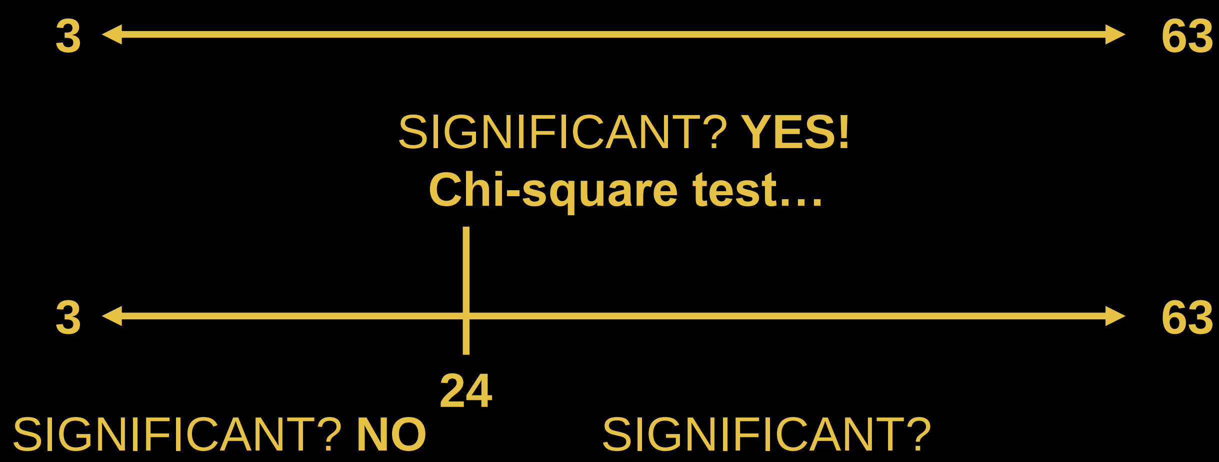
Chi-square test…

3 **63**

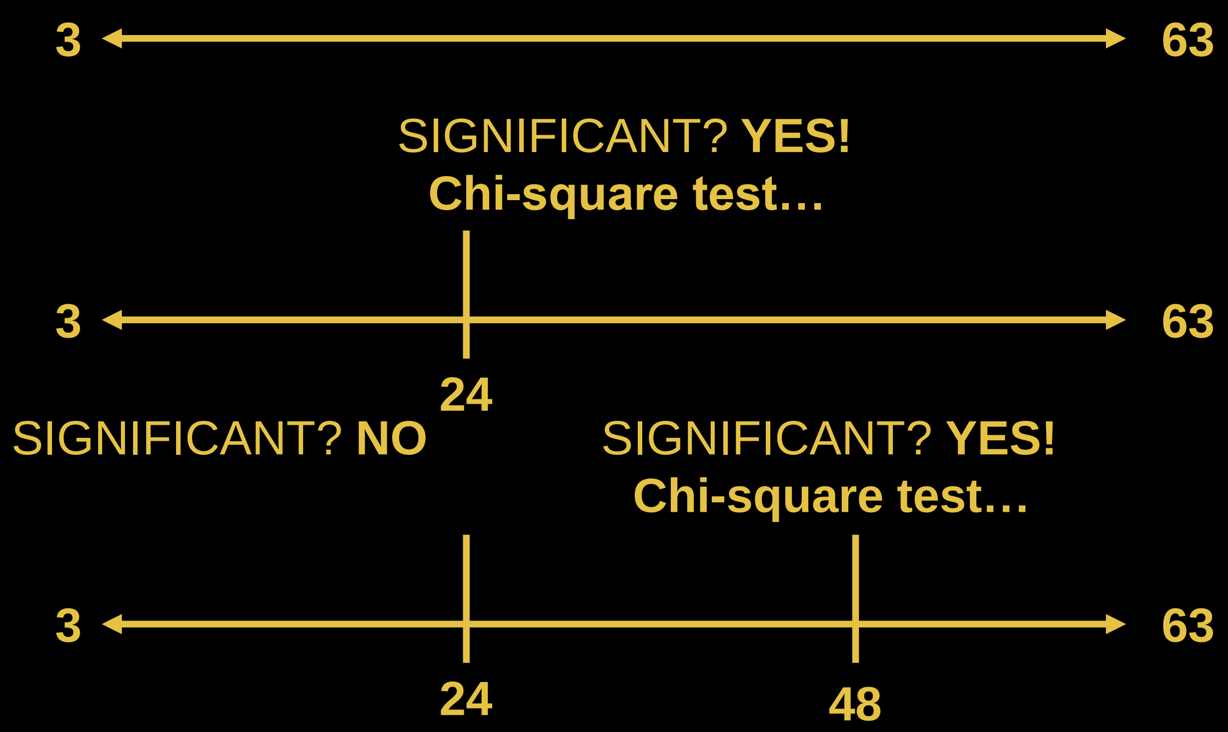
24

SIGNIFICANT?

Initial Characteristic Analysis – R



Initial Characteristic Analysis – R



Initial Characteristic Analysis – R

Initial Characteristic Analysis

- Cut-offs may be rough from decision tree combining.
- **Optional to override** automatically generated groups to conform to business rules.
- Overrides may make groups suboptimal.

Group Definition
Missing
< $35,200
$35,200 - $60,000
$60,000 - $85,000
$85,000 - $110,000
$110,000 - $142,530
> $142,530

Initial Characteristic Analysis

- Cut-offs may be rough from decision tree combining.
- **Optional to override** automatically generated groups to conform to business rules.
- Overrides may make groups suboptimal.

Group Definition	Override
Missing	Missing
< $35,200	< $35,000
$35,200 - $60,000	$35,000 - $60,000
$60,000 - $85,000	$60,000 - $85,000
$85,000 - $110,000	$85,000 - $110,000
$110,000 - $142,530	$110,000 - $140,000
> $142,530	> $140,000

Initial Characteristic Analysis

- Cut-offs may be rough from decision tree combining.
- **Optional to override** automatically generated groups to conform to business rules.
- Overrides may make groups suboptimal.

Group Definition	Override
Missing	Missing
< $35,200	**< $35,000**
$35,200 - $60,000	**$35,000** - $60,000
$60,000 - $85,000	$60,000 - $85,000
$85,000 - $110,000	$85,000 - $110,000
$110,000 - $142,530	$110,000 - **$140,000**
> $142,530	**> $140,000**

Initial Characteristic Analysis

- Calculate and examine the key assessment metrics:
  - **Weight of Evidence** (WOE) – how well attributes discriminate for each given characteristic
  - **Information Value** (IV) – evaluate a characteristic's overall predictive power
  - **Gini Statistic** – alternate to IV for selecting characteristics for final model.

WEIGHT OF EVIDENCE

Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.
- WOE is based on comparing the proportion of goods to bads at each attribute level (levels of the predictor variable).

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.
- WOE is based on comparing the proportion of goods to bads at each attribute level (levels of the predictor variable).

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

$$Dist.Good_i = \frac{Number\ Good\ in\ group\ i}{Total\ Number\ Good}$$

Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.
- WOE is based on comparing the proportion of goods to bads at each attribute level (levels of the predictor variable).

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

$$Dist.Bad_i = \frac{Number\ Bad\ in\ group\ i}{Total\ Number\ Bad}$$

Weight of Evidence (WOE)

- What are we looking for?
  - Looking for “big” differences in WOE between groups.
  - Monotonic changes within an attribute for **interval variables** (not always required).

- Why monotonic increases?
  - Oscillation back and forth of positive to negative values of WOE typically sign of variable that has trouble separating good vs. bad.
  - Not always required **if makes business sense** – credit card utilization for example.

WOE – Example

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 610	120	622	
2	610 – 630	96	646	
3	630 – 653	159	1,373	
4	653 – 662	60	684	
5	662 – 685	118	2,242	
6	685 – 708	98	2,213	
7	708 – 723	45	1,501	
8	723 – 763	50	3,045	
9	> 763	18	2,399	
10	MISSING	73	703	
Total		**837**	**15,428**	

WOE – Example

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 610	120	622	
2	610 – 630	96	646	
3	630 – 653	159	1,373	
4	653 – 662	60	684	
5	662 – 685	118	2,242	
6	685 – 708	98	2,213	
7	708 – 723	45	1,501	
8	723 – 763	50	3,045	
9	> 763	18	2,399	
10	MISSING	73	703	
Total		**837**	**15,428**	

$$Dist.Good_1 = \frac{622}{15428}$$

$$= 0.040$$

WOE – Example

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 610	120	622	
2	610 – 630	96	646	
3	630 – 653	159	1,373	
4	653 – 662	60	684	
5	662 – 685	118	2,242	
6	685 – 708	98	2,213	
7	708 – 723	45	1,501	
8	723 – 763	50	3,045	
9	> 763	18	2,399	
10	MISSING	73	703	
Total		**837**	**15,428**	

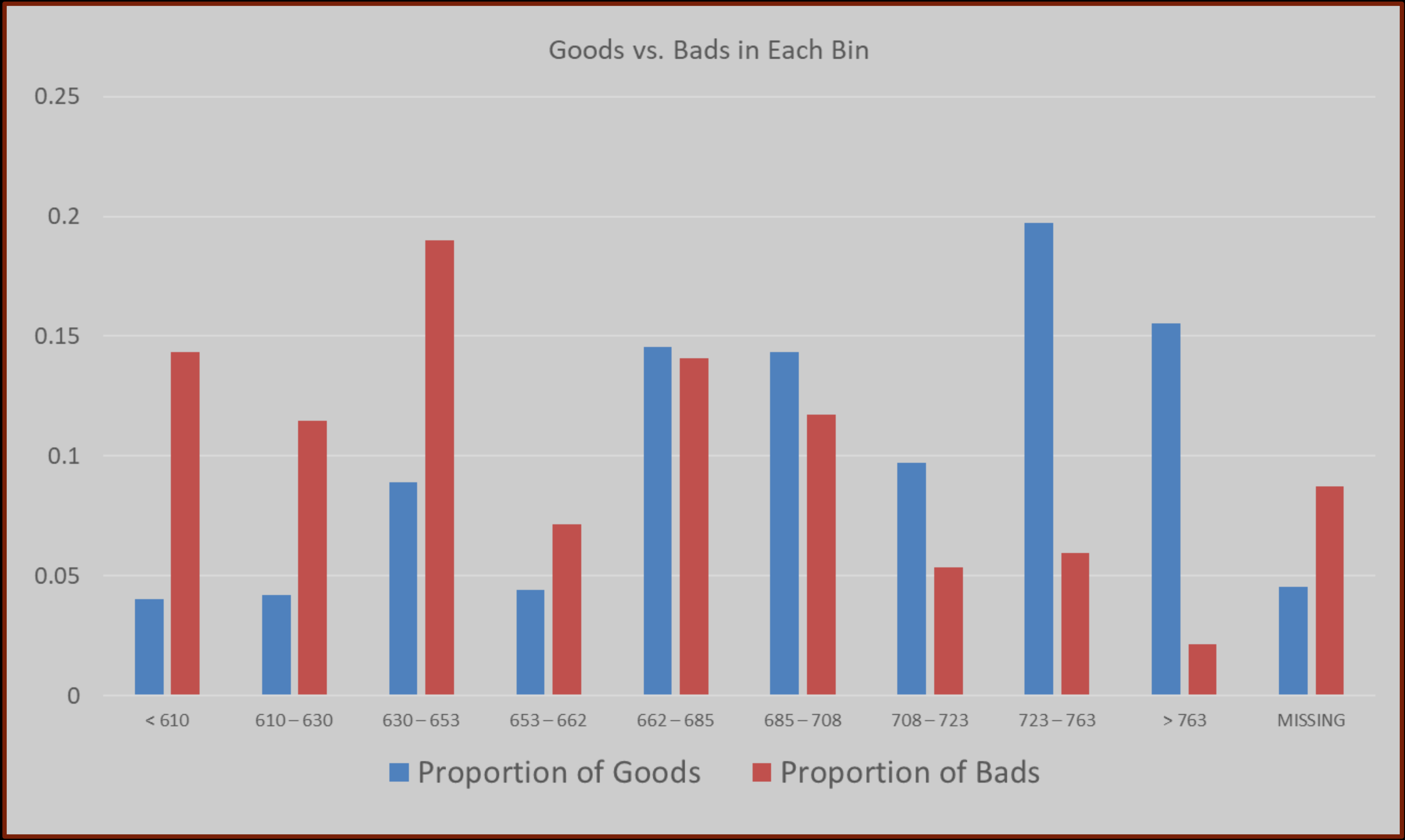
$$Dist.Good_1 = \frac{622}{15428}$$

$$= 0.040$$

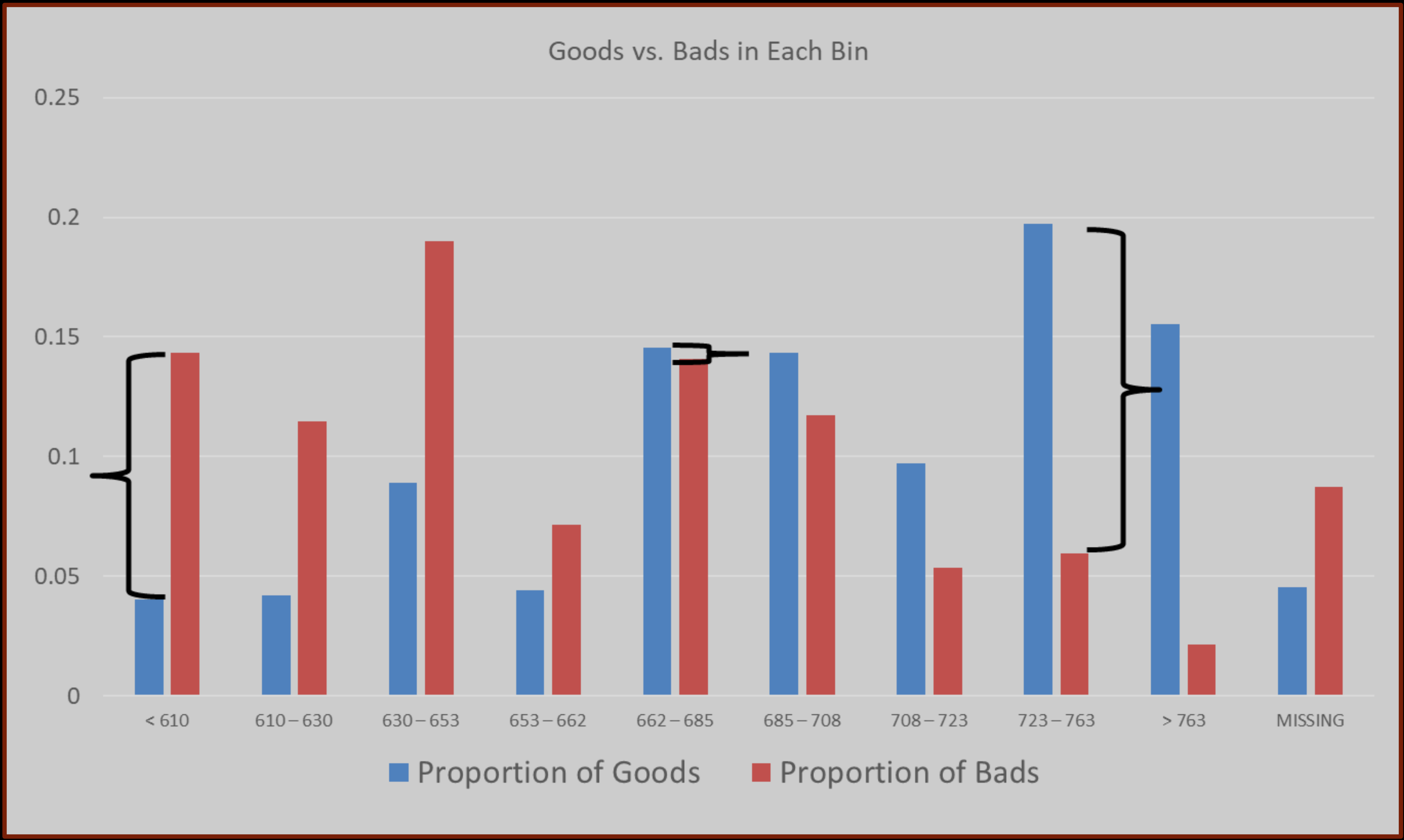
$$Dist.Bad_1 = \frac{120}{837}$$

$$= 0.143$$

WOE – Example


Goods vs. Bads in Each Bin
0.25
0.2
0.15
0.1
0.05
0
< 610
610 – 630
630 – 653
653 – 662
662 – 685
685 – 708
708 – 723
723 – 763
> 763
MISSING
Proportion of Goods
Proportion of Bads

WOE – Example


Goods vs. Bads in Each Bin
0.25
0.2
0.15
0.1
0.05
0
< 610
610 – 630
630 – 653
653 – 662
662 – 685
685 – 708
708 – 723
723 – 763
> 763
MISSING
Proportion of Goods
Proportion of Bads

WOE – Example

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 610	120	622	-1.269
2	610 – 630	96	646	
3	630 – 653	159	1,373	
4	653 – 662	60	684	
5	662 – 685	118	2,242	
6	685 – 708	98	2,213	
7	708 – 723	45	1,501	
8	723 – 763	50	3,045	
9	> 763	18	2,399	
10	MISSING	73	703	
Total		**837**	**15,428**	

$$Dist.Good_1 = \frac{622}{15428}$$

$$= 0.040$$

$$Dist.Bad_1 = \frac{120}{837}$$

$$= 0.143$$

$$WOE_1 = \log\left(\frac{0.040}{0.143}\right)$$

$$= -1.269$$

WOE – Example

WOE for Bureau Score				
Group	Values	Event Count	Non-event Count	WOE
1	< 610	120	622	**-1.269**
2	610 – 630	96	646	**-1.008**
3	630 – 653	159	1,373	**-0.758**
4	653 – 662	60	684	**-0.481**
5	662 – 685	118	2,242	**0.030**
6	685 – 708	98	2,213	**0.203**
7	708 – 723	45	1,501	**0.593**
8	723 – 763	50	3,045	**1.195**
9	> 763	18	2,399	**1.978**
10	MISSING	73	703	**-0.649**
Total		**837**	**15,428**	

Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

- WOE approximately zero implies what?

Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.

$$WOE_i = \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

- WOE approximately zero implies % good approximately equal to % bad so group doesn’t separate good vs. bad well.

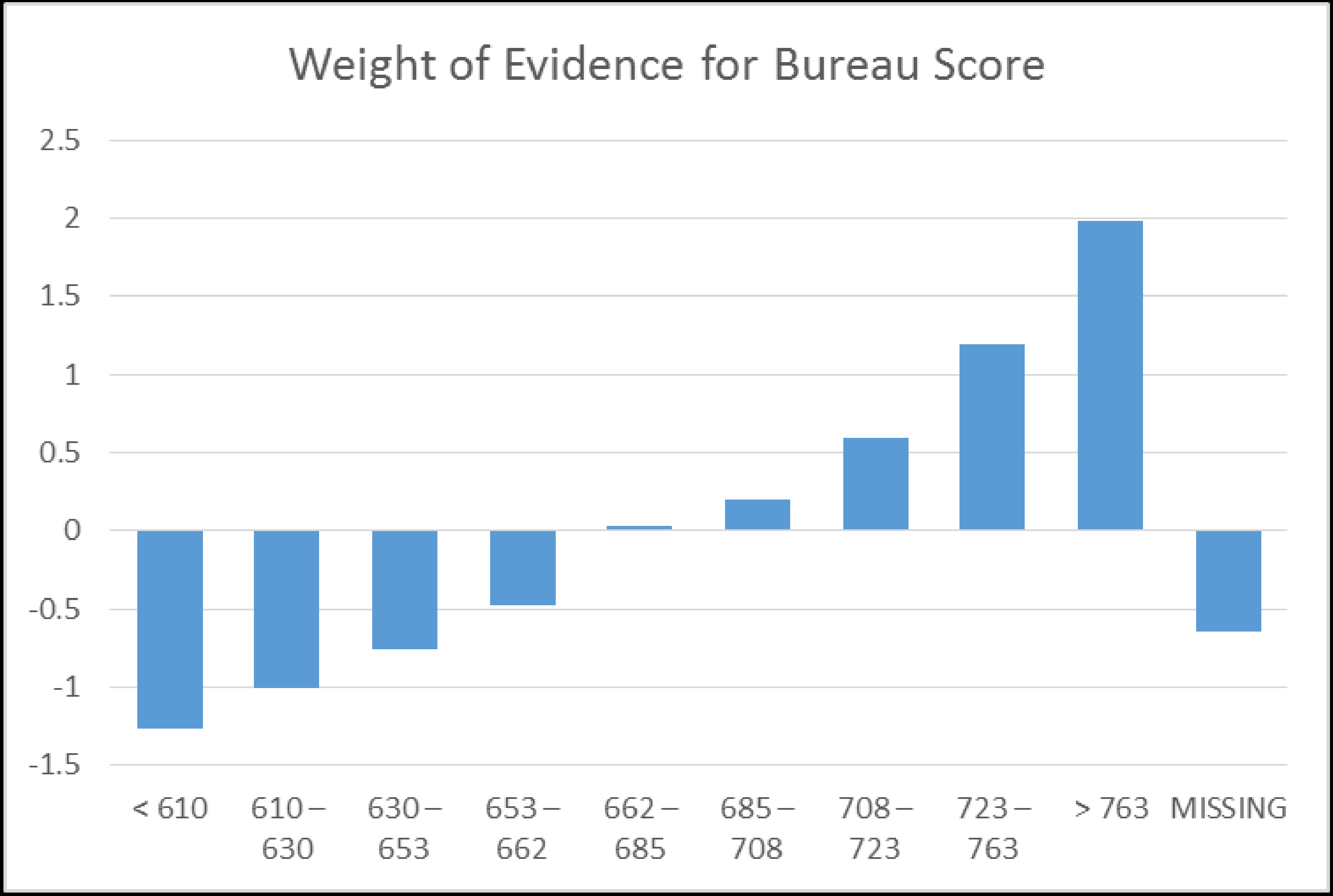
Weight of Evidence (WOE)

- WOE measures the strength of the attributes of a characteristic in **separating good and bad accounts**.

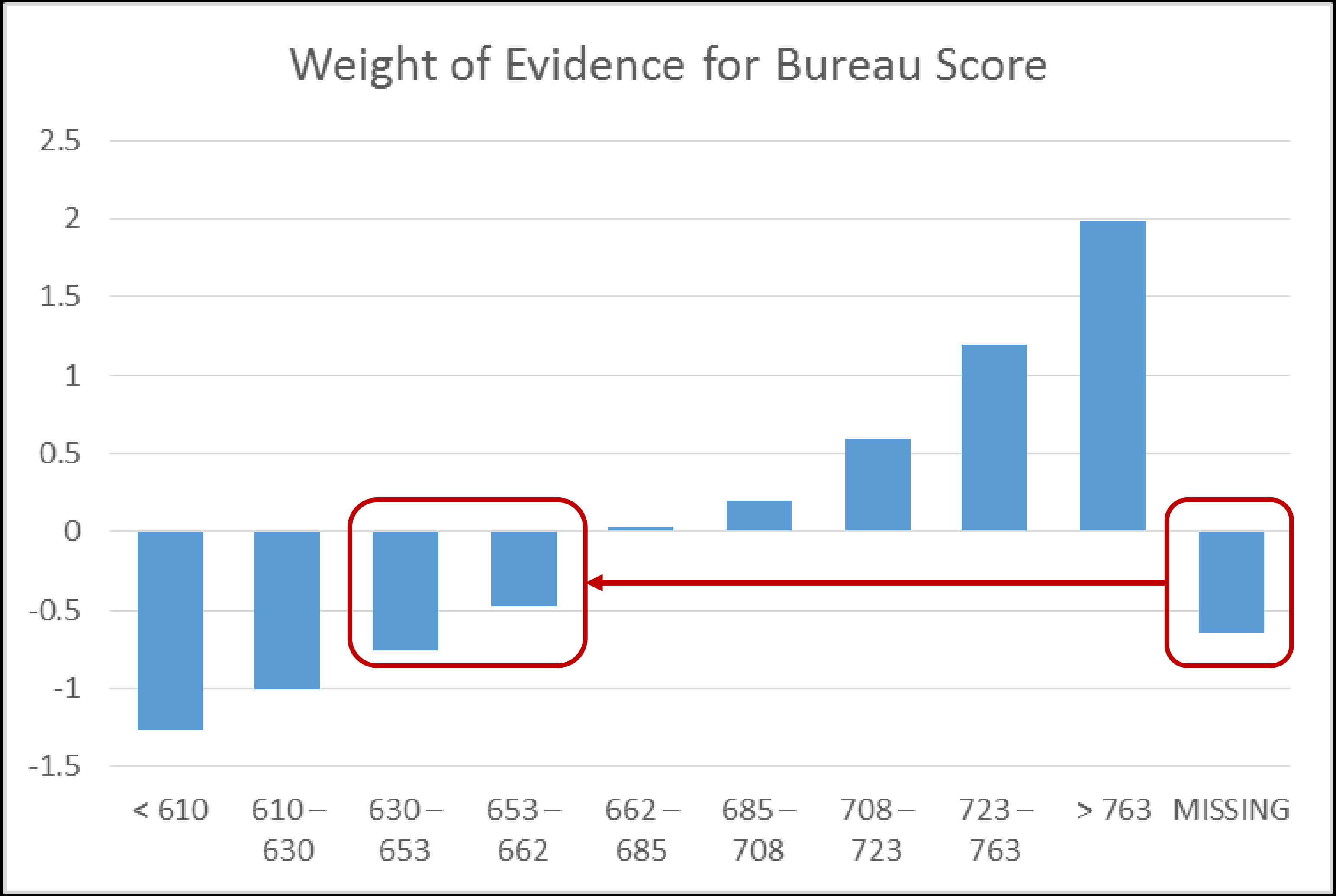
$$WOE_i = \log\left(\frac{Dist.\,Good_i}{Dist.\,Bad_i}\right)$$

- WOE approximately zero implies % good approximately equal to % bad so group doesn't separate good vs. bad well.
- WOE positive implies group identifies people who are good.
- WOE negative implies group identifies people who are bad.

WOE – Example


Weight of Evidence for Bureau Score
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
< 610
610 – 630
630 – 653
653 – 662
662 – 685
685 – 708
708 – 723
723 – 763
> 763
MISSING

WOE – Example


Weight of Evidence for Bureau Score
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
< 610
610 – 630
630 – 653
653 – 662
662 – 685
685 – 708
708 – 723
723 – 763
> 763
MISSING

WOE – Example – smbinning

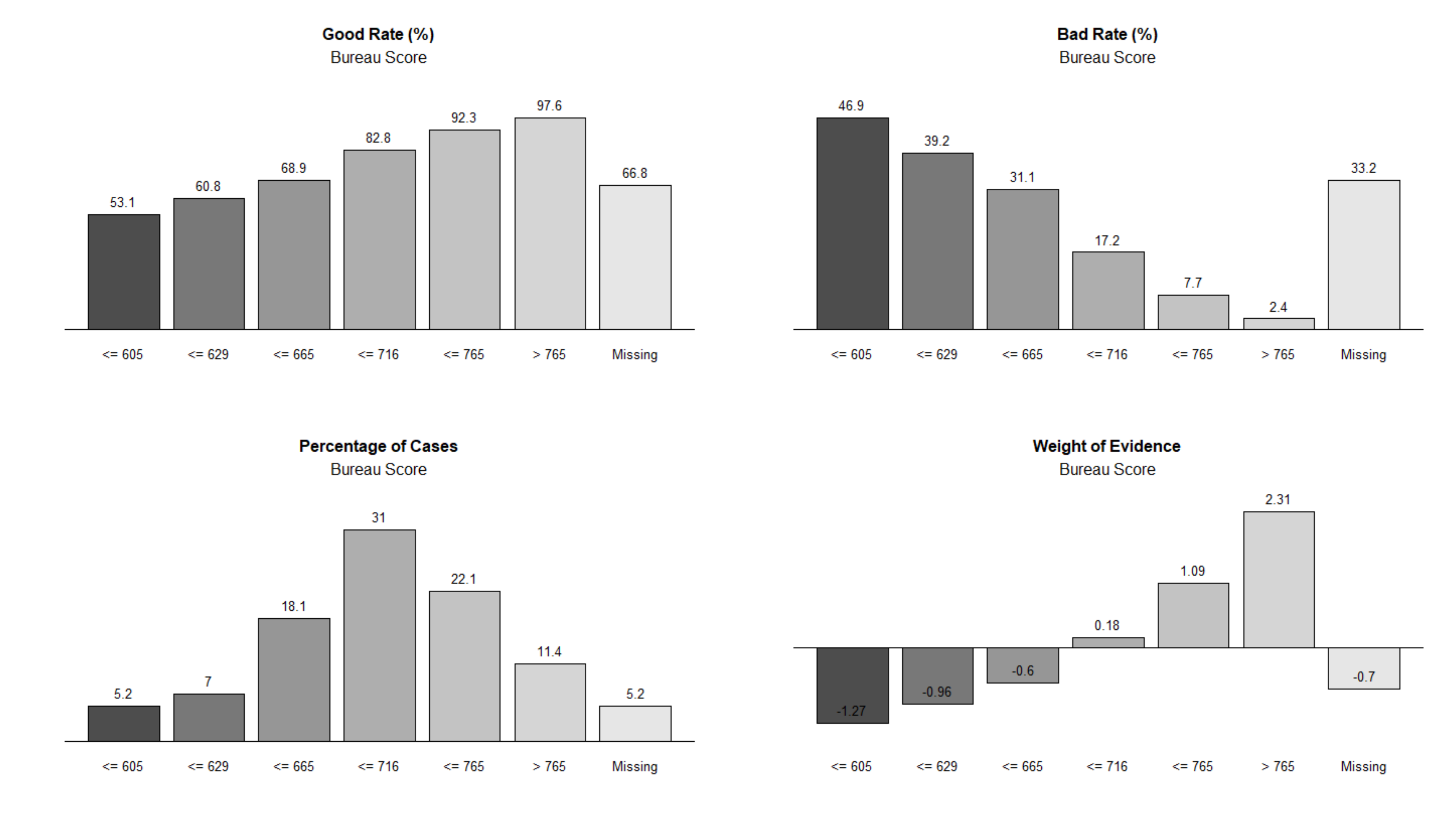
```
# Binning of Continuous Variables #
result <- smbinning(df = train, y = "good", x = "bureau_score")
result$ivtable
result$cut
result$iv

smbinning.plot(result,option="dist",sub="Bureau Score")
smbinning.plot(result,option="goodrate",sub="Bureau Score")
smbinning.plot(result,option="badrate",sub="Bureau Score")
smbinning.plot(result,option="WoE",sub="Bureau Score")
```

WOE – Example – ivtable

```
> # Binning of Continuous Variables #
> result <- smbinning(df = train, y = "good", x = "bureau_score")
> result$ivtable
  Cutpoint CntRec CntGood CntBad CntCumRec CntCumGood CntCumBad PctRec GoodRate BadRate    Odds LnOdds     WoE     IV
1   <= 605    228     121    107       228        121       107 0.0521   0.5307  0.4693  1.1308 0.1230 -1.2739 0.1131
2   <= 629    306     186    120       534        307       227 0.0699   0.6078  0.3922  1.5500 0.4383 -0.9586 0.0817
3   <= 665    791     545    246      1325        852       473 0.1807   0.6890  0.3110  2.2154 0.7955 -0.6014 0.0770
4   <= 716   1358    1125    233      2683       1977       706 0.3103   0.8284  0.1716  4.8283 1.5745  0.1776 0.0093
5   <= 765    965     891     74      3648       2868       780 0.2205   0.9233  0.0767 12.0405 2.4883  1.0914 0.1841
6    > 765    500     488     12      4148       3356       792 0.1142   0.9760  0.0240 40.6667 3.7054  2.3085 0.2891
7  Missing    229     153     76      4377       3509       868 0.0523   0.6681  0.3319  2.0132 0.6997 -0.6972 0.0306
8    Total   4377    3509    868        NA         NA        NA 1.0000   0.8017  0.1983  4.0426 1.3969  0.0000 0.7849
> 
```

WOE – Example – smbinning.plot



?

INFORMATION VALUE

Information Value (IV)

- How big is a "big" difference when looking across groups for WOE?
- IV measures the ability of the characteristic to separate goods vs. bads.

$$IV = \sum_{i=1}^{L} (Dist.Good_i - Dist.Bad_i) \times \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

Information Value (IV)

- How big is a "big" difference when looking across groups for WOE?
- IV measures the ability of the characteristic to separate goods vs. bads.

$$IV = \sum_{i=1}^{L} (Dist.Good_i - Dist.Bad_i) \times \underbrace{\boxed{\log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)}}_{\text{Weight of Evidence!}}$$

Information Value (IV)

- How big is a “big” difference when looking across groups for WOE?
- IV measures the ability of the characteristic to separate goods vs. bads.

$$IV = \sum_{i=1}^{L} (Dist.Good_i - Dist.Bad_i) \times \log\left(\frac{Dist.Good_i}{Dist.Bad_i}\right)$$

- Used to select characteristics with strong predictive value.

Information Value (IV)

- Characteristics of IV:
  - $IV \geq 0$
  - Bigger is Better!

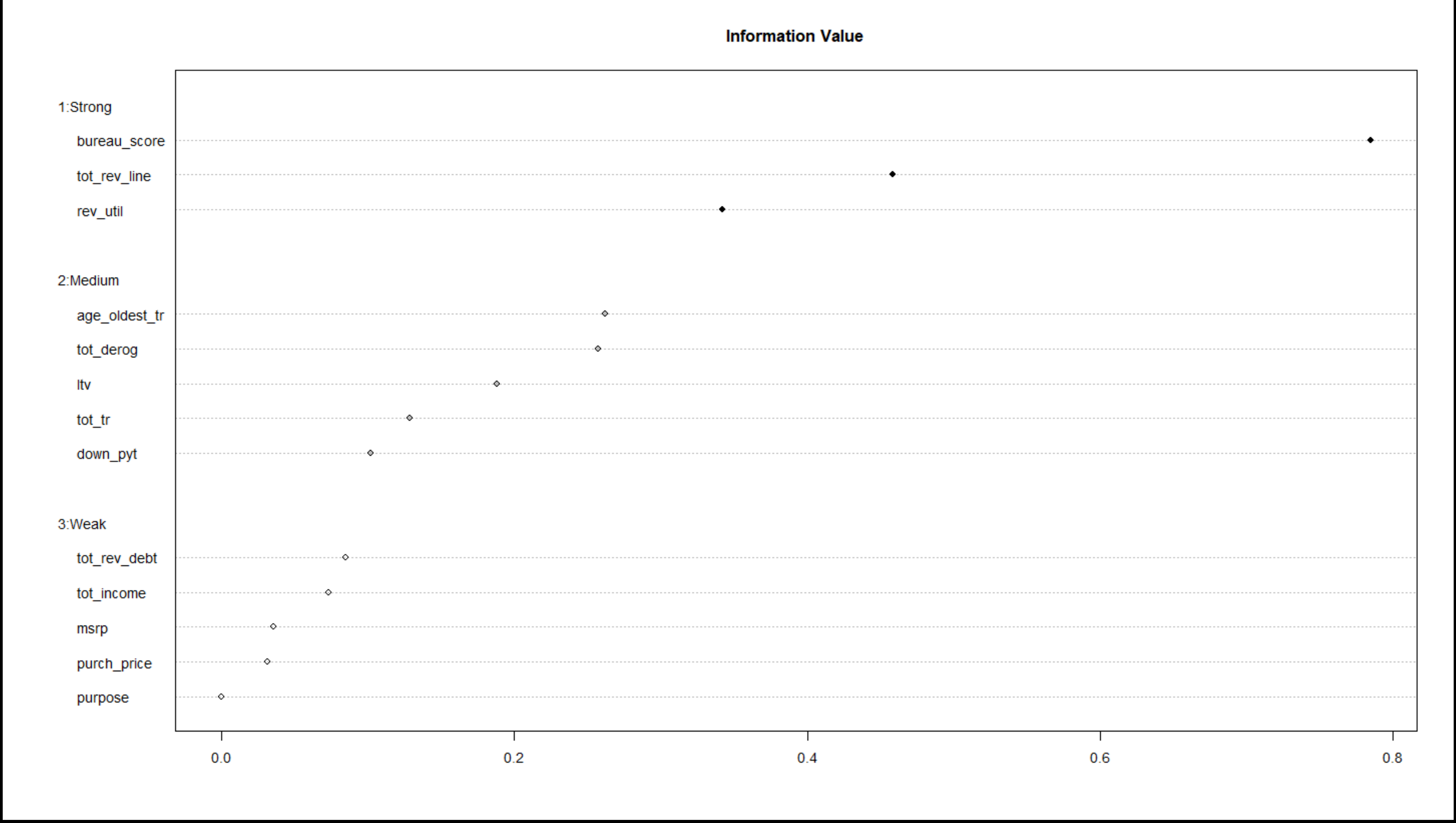
- Rules of Thumb:
  - IV < 0.02 – Not predictive
  - 0.02 < IV < 0.1 – Weak predictor
  - 0.1 < IV < 0.25 – Medium predictor
  - 0.25 < IV – Strong predictor

Information Value (IV) – Example

```
# Information Value for Each Variable #
iv_summary <- smbinning.sumiv(df = train, y = "good")

smbinning.sumiv.plot(iv_summary)
iv_summary
```

Information Value (IV) – Example



Information Value (IV) – Example

```
> smbinning.sumiv.plot(iv_summary)
> iv_summary
            Char     IV                 Process
12  bureau_score 0.7849     Numeric binning OK
10  tot_rev_line 0.4582     Numeric binning OK
11      rev_util 0.3420     Numeric binning OK
6  age_oldest_tr 0.2624     Numeric binning OK
4      tot_derog 0.2575     Numeric binning OK
19           ltv 0.1881     Numeric binning OK
5         tot_tr 0.1286     Numeric binning OK
15      down_pyt 0.1022     Numeric binning OK
9   tot_rev_debt 0.0853     Numeric binning OK
20    tot_income 0.0731     Numeric binning OK
14          msrp 0.0356     Numeric binning OK
13   purch_price 0.0314     Numeric binning OK
16       purpose 0.0000      Factor binning OK
1     bankruptcy     NA     Uniques values < 5
2            bad     NA     Uniques values < 5
3         app_id     NA No significant splits
7    tot_open_tr     NA No significant splits
8     tot_rev_tr     NA No significant splits
17     loan_term     NA No significant splits
18      loan_amt     NA No significant splits
21      used_ind     NA     Uniques values < 5
22        weight     NA     Uniques values < 5
> 
```

Information Value (IV)

- Characteristics of IV:
  - $IV \geq 0$
  - Bigger is Better!

- Rules of Thumb:
  - IV < 0.02 – Not predictive
  - 0.02 < IV < 0.1 – Weak predictor
  - 0.1 < IV < 0.25 – Medium predictor
  - 0.25 < IV < 0.5 – Strong predictor
  - IV > 0.5 – Over-predicting?

Information Value (IV)

- Rules of Thumb:
  - IV < 0.02 – Not predictive
  - 0.02 < IV < 0.1 – Weak predictor
  - 0.1 < IV < 0.25 – Medium predictor
  - 0.25 < IV < 0.5 – Strong predictor
  - IV > 0.5 – Over-predicting?
- Over-predicting Example:
  - All previous mortgage decisions have been made only on bureau score so of course bureau score is highly predictive – becomes only significant variable!
  - Create two models – one with bureau score, one without bureau score and **ensemble**.

?

GINI STATISTIC

Gini Statistic

- **Gini statistic** is optional technique that tries to answer the same question as Information Value – which variables are strong enough to enter the scorecard model?
- IV is more in line with WOE calculation and used more often.
- Characteristics:
  - Range is 0 to 100.
  - Bigger is Better.

Gini Statistic Calculation

- More complicated technique for trying to evaluate how characteristics separate good from bad.
- Majority of the time Gini and IV agree, but could be different on the borderline cases.
- Calculation:
  - Sort *L* groups of variable by descending order of the proportion of all events.

$$Gini = \left(1 - \frac{\left(2\sum_{i=2}^{L}\left(n_{i,E} \times \sum_{i=1}^{i-1} n_{i,NE}\right) + \sum_{i=1}^{L}\left(n_{i,E} \times n_{i,NE}\right)\right)}{N_E \times N_{NE}}\right) \times 100$$

Gini Statistic Calculation

- More complicated technique for trying to evaluate how characteristics separate good from bad.
- Majority of the time Gini and IV agree, but could be different on the borderline cases.
- Calculation:
  - Sort *L* groups of variable by descending order of the proportion of all events.

$$Gini = \left(1 - \frac{\left(2\sum_{i=2}^{L}\left(n_{i,E} \times \sum_{i=1}^{i-1} n_{i,NE}\right) + \sum_{i=1}^{L}\left(n_{i,E} \times n_{i,NE}\right)\right)}{N_E \times N_{NE}}\right) \times 100$$

Number of events in group *i*

Gini Statistic Calculation

- More complicated technique for trying to evaluate how characteristics separate good from bad.
- Majority of the time Gini and IV agree, but could be different on the borderline cases.
- Calculation:
  - Sort *L* groups of variable by descending order of the proportion of all events.

$$Gini = \left(1 - \frac{\left(2\sum_{i=2}^{L}\left(n_{i,E} \times \sum_{i=1}^{i-1} n_{i,NE}\right) + \sum_{i=1}^{L}\left(n_{i,E} \times n_{i,NE}\right)\right)}{N_E \times N_{NE}}\right) \times 100$$

Number of non-events in group *i*

Gini Statistic Calculation

- More complicated technique for trying to evaluate how characteristics separate good from bad.
- Majority of the time Gini and IV agree, but could be different on the borderline cases.
- Calculation:
  - Sort *L* groups of variable by descending order of the proportion of all events.

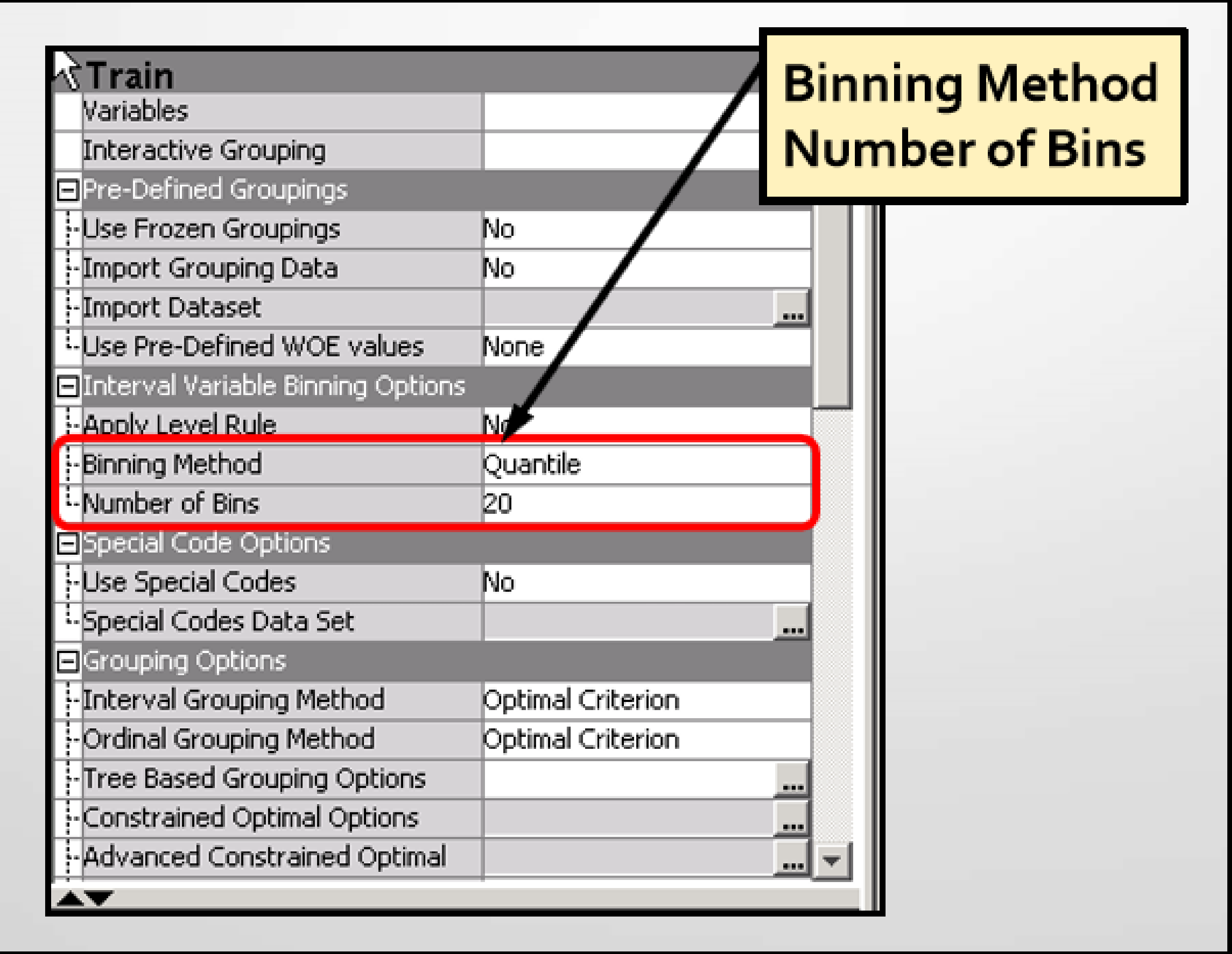
$$Gini = \left(1 - \frac{\left(2\sum_{i=2}^{L}\left(n_{i,E} \times \sum_{i=1}^{i-1} n_{i,NE}\right) + \sum_{i=1}^{L}\left(n_{i,E} \times n_{i,NE}\right)\right)}{N_E \times N_{NE}}\right) \times 100$$

Total number of events and non-events

?

INTERACTIVE GROUPING NODE IN SAS EM

Pre-Binning of the Interval Variables



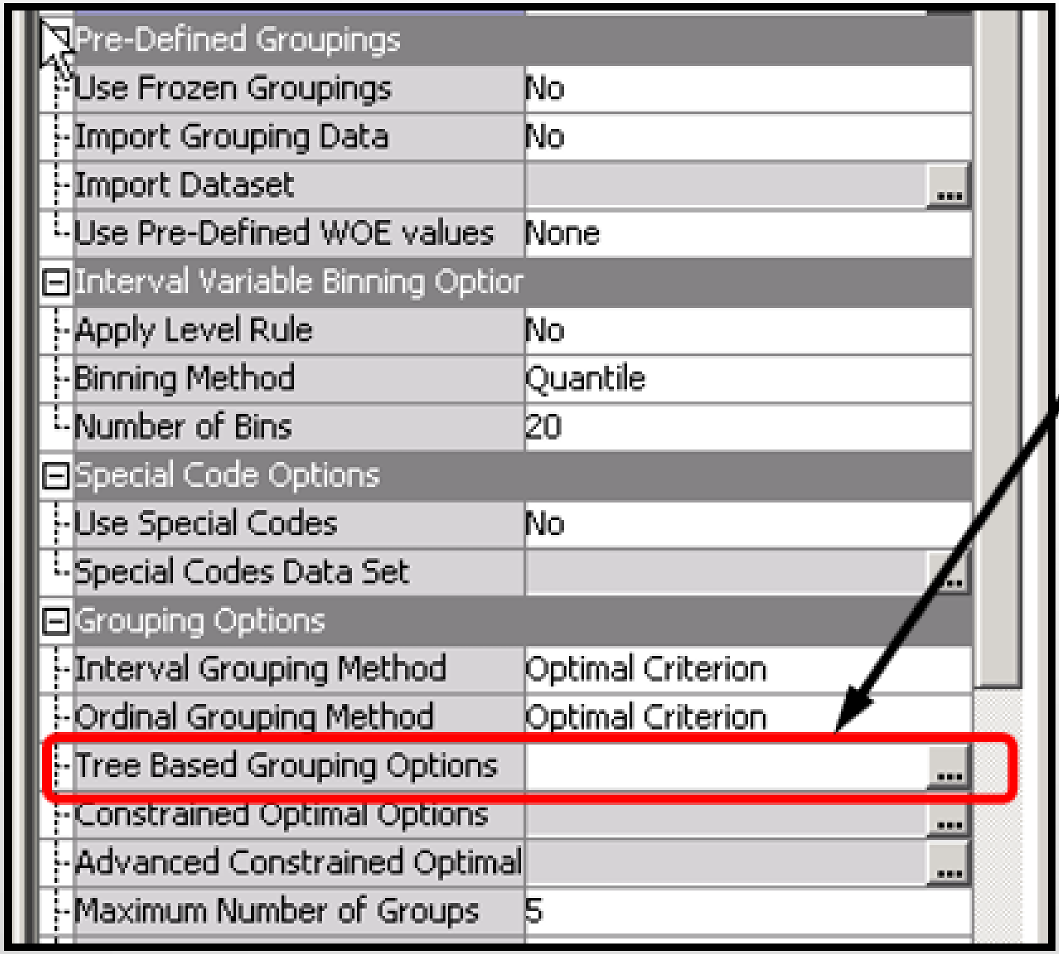
Grouping Options

Pre-Defined Groupings	
Use Frozen Groupings	No
Import Grouping Data	No
Import Dataset	...
Use Pre-Defined WOE values	None
Interval Variable Binning Optior	
Apply Level Rule	No
Binning Method	Quantile
Number of Bins	20
Special Code Options	
Use Special Codes	No
Special Codes Data Set	...
Grouping Options	
Interval Grouping Method	Optimal Criterion
Ordinal Grouping Method	Optimal Criterion
Tree Based Grouping Options	...
Constrained Optimal Options	...
Advanced Constrained Optimal	...
Maximum Number of Groups	5
Significant Digits	2
Apply Restrictions	Yes
Type	Percent
Percent	5.0
Count	.
Adjust WOE	Yes
Adjustment Factor	0.5

Pre-Defined Groupings

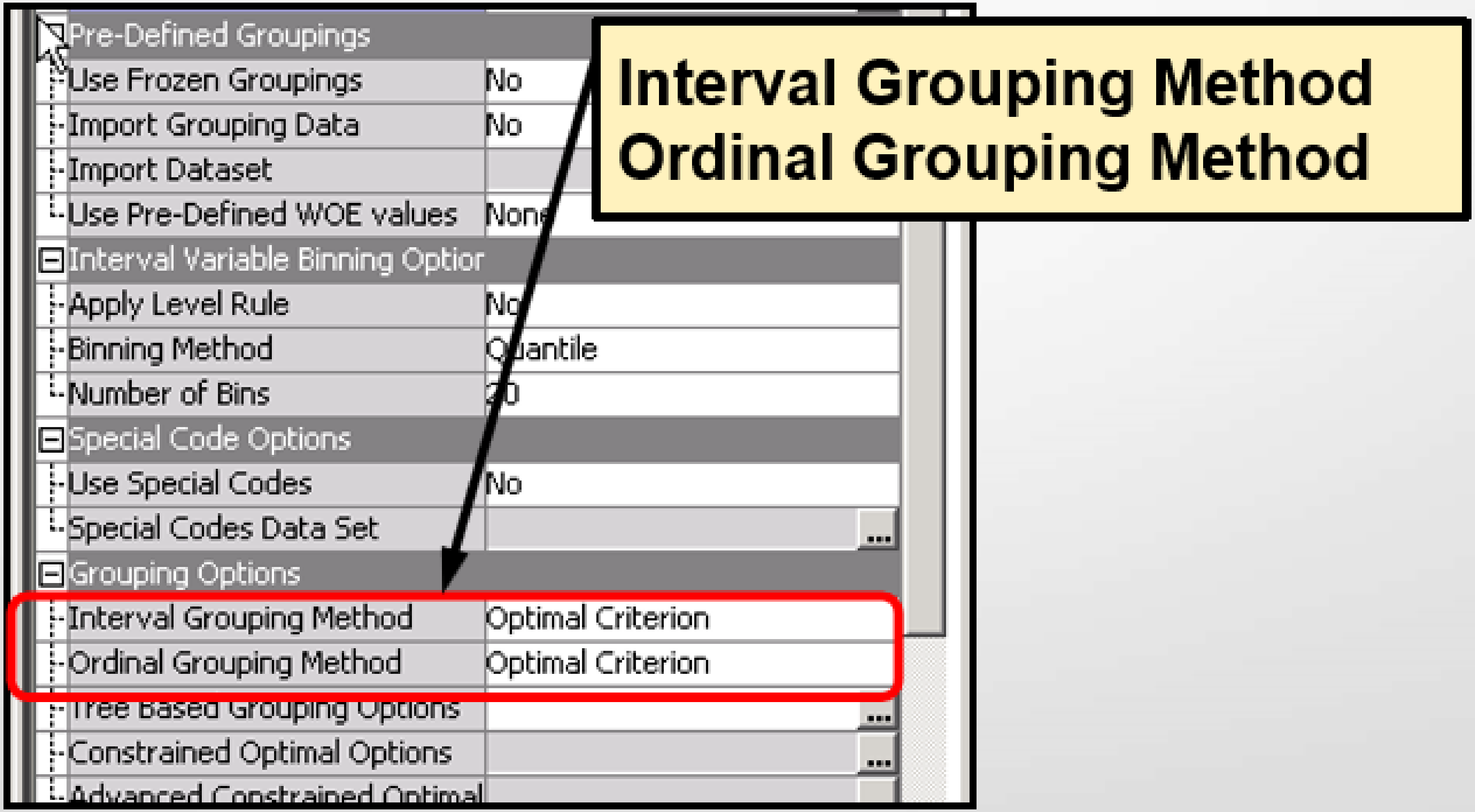
Grouping Options

Grouping Options: Tree Criteria

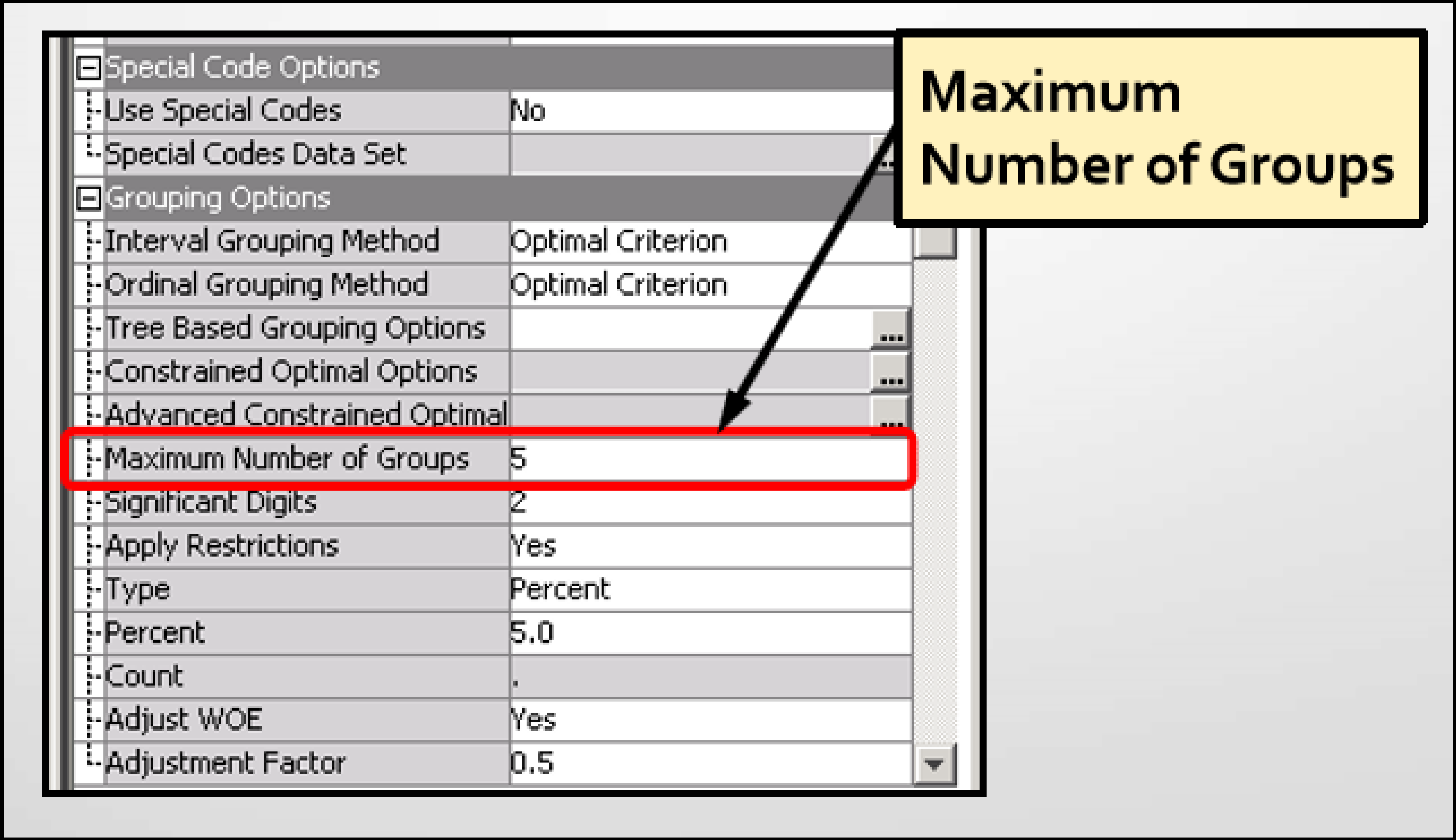


Control tree criteria
for grouping:
Split Criterion
Missing Values
Minimum Group Size

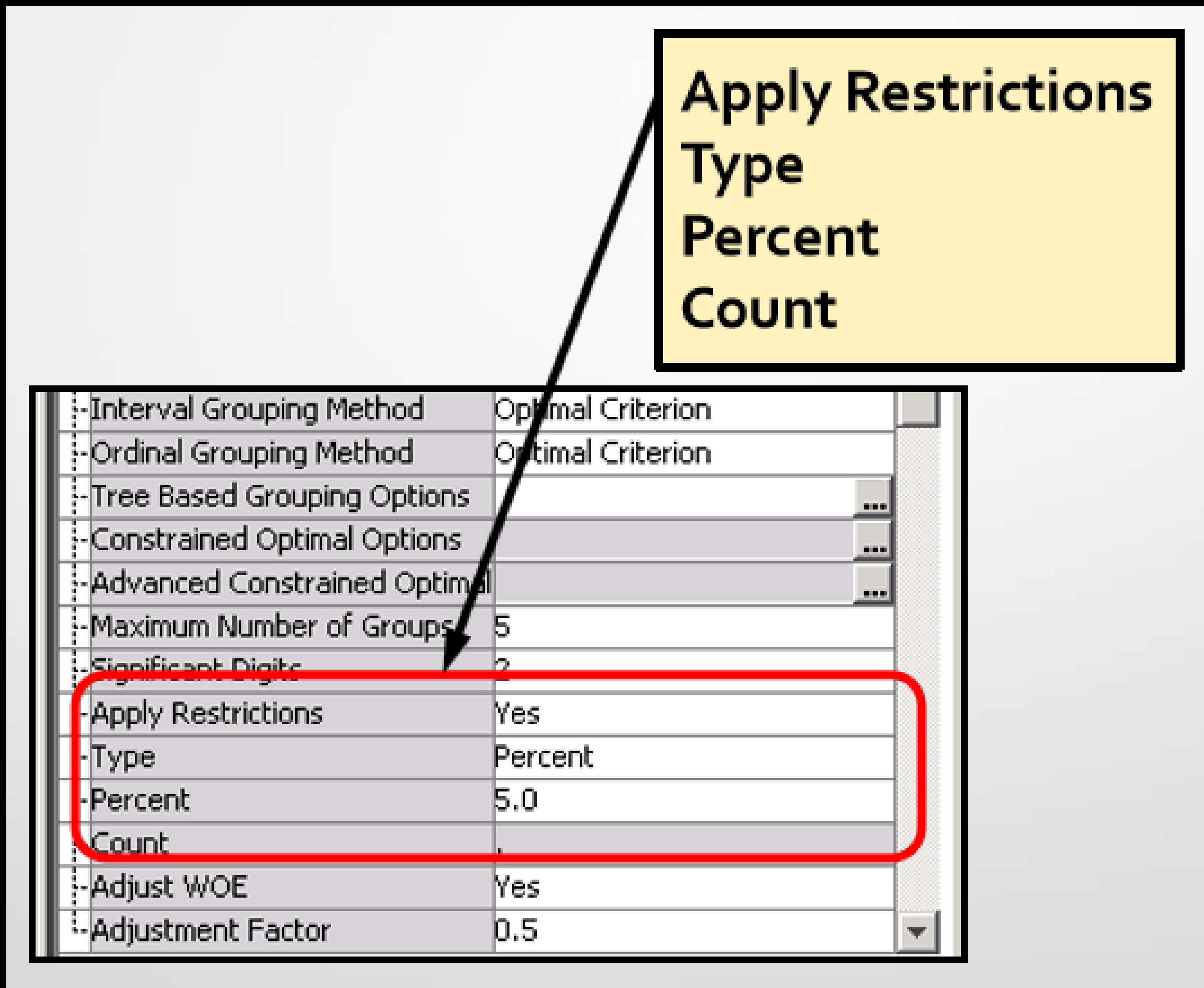
Grouping Options: Interval vs. Ordinal



Grouping Options: Number of Groups



Grouping Options: Stopping Rules



Grouping Options: WOE Adjustments