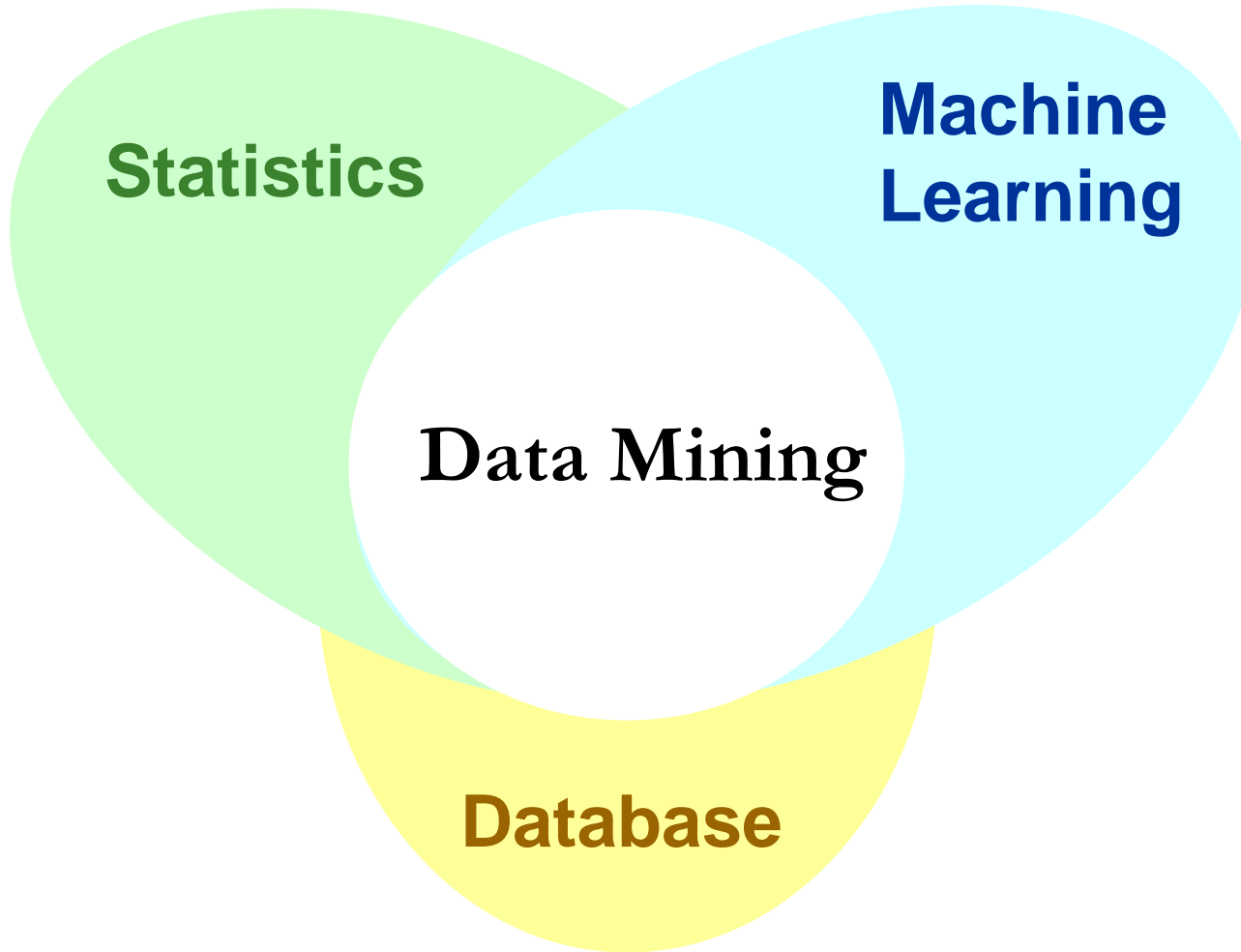


---

# Data Exploration

---

Saed Sayad



---

# Data Mining Steps

- Define the problem
- Prepare data
- Explore data
- Build model
- Evaluate model
- Deploy model

---

# Dataset

- Credit Scoring Dataset
  - 900 cases
  - 20 variables + 1 target
    1. checking\_status
    2. duration
    3. credit\_history
    4. purpose
    5. credit\_amount
    6. savings\_status
    7. employment
    8. ...
- target : credit\_assessment

---

# Data Exploration

## ■ Univariate Analysis

- Categorical variables (e.g., Housing)
- Numeric variables (e.g., Age)

## ■ Bivariates Analysis

- Categorical-Categorical variables (e.g., Gender and Housing)
- Numeric-Numeric variables (e.g., Age and Credit Amount)
- Numeric-Categorical variables (e.g., Age and Housing)

## ■ Multivariates Analysis

- Multi-dimensional analysis (e.g., Age, Housing and Credit Amount)

---

# Univariate Analysis

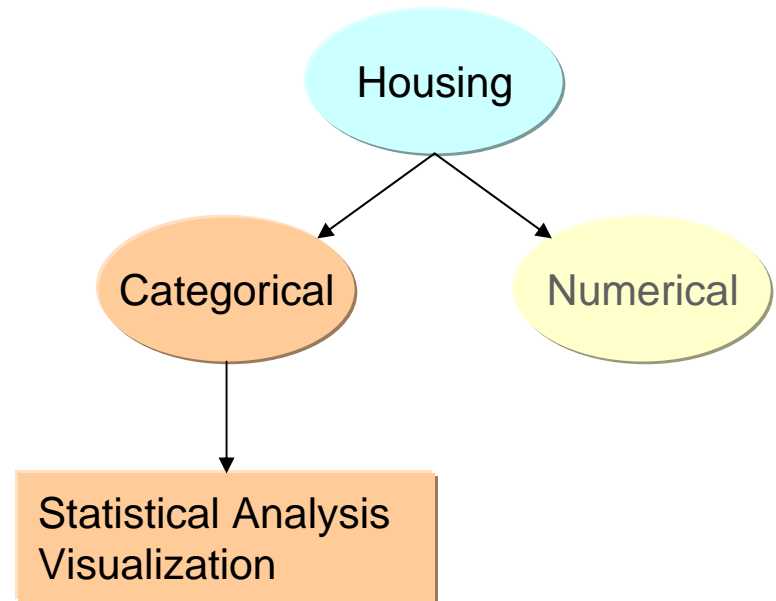
# Univariate Analysis – Categorical

- **Statistical Analysis**

- Count, Count%
- Missing Values
- Invalid Values
- Numerization

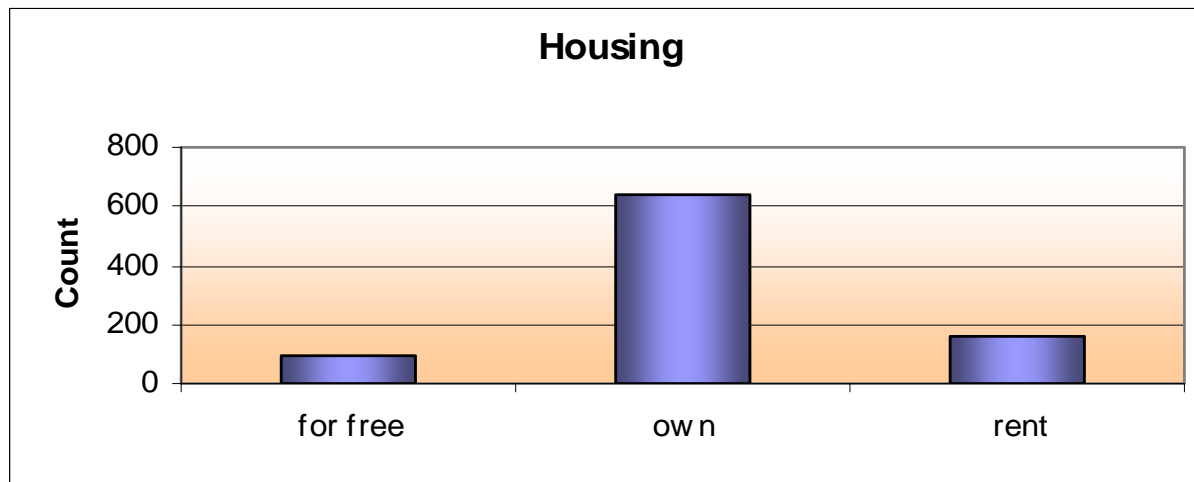
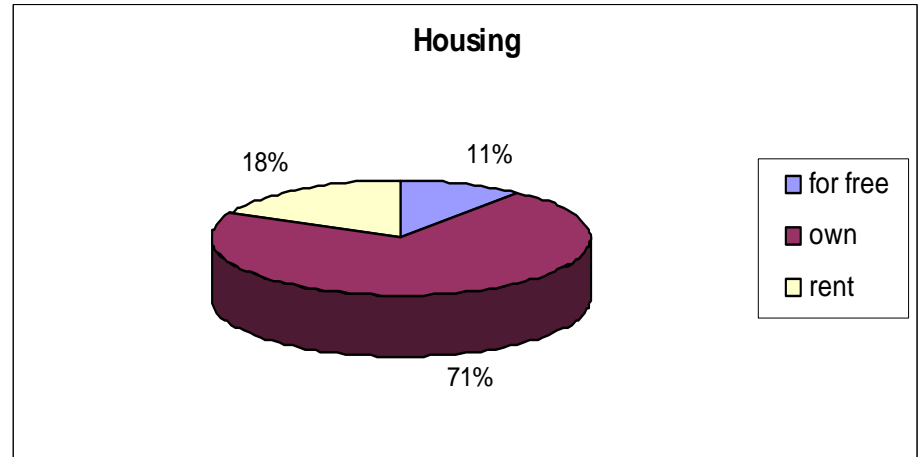
- **Visualization**

- Bar Chart, Pie Chart, ...

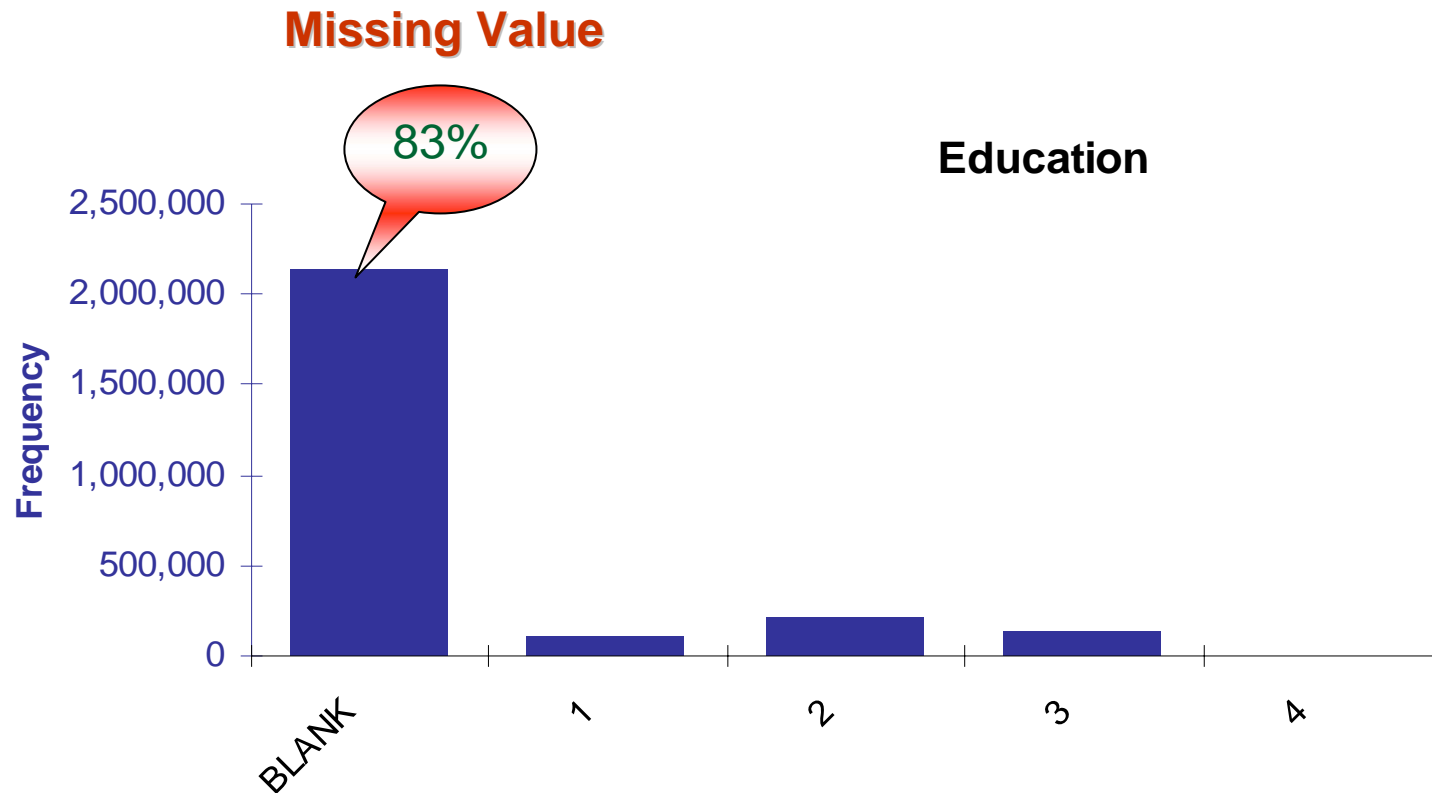


# Univariate Analysis – Categorical

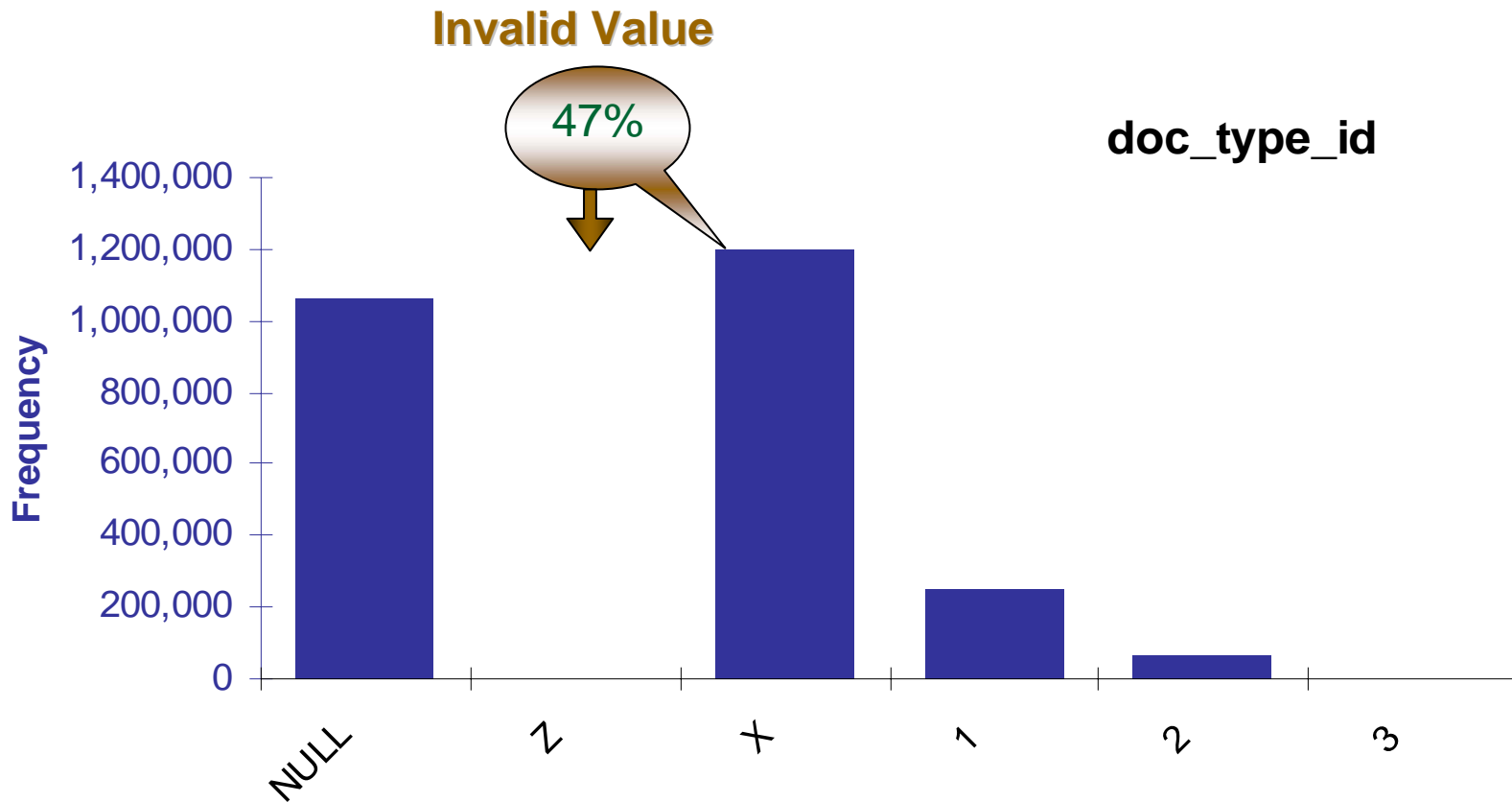
housing	Count	Count%
for free	96	10.67%
own	641	71.22%
rent	163	18.11%



# Univariate Analysis – Categorical

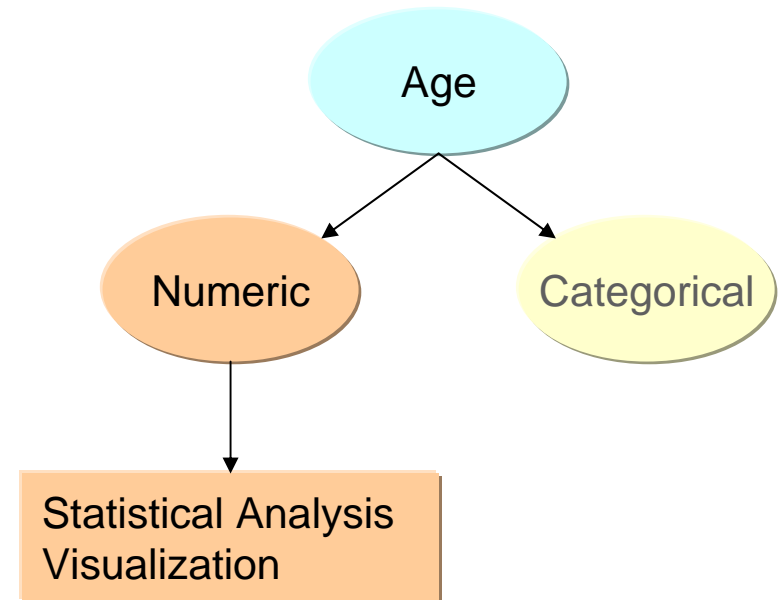


# Univariate Analysis – Categorical



# Univariate Analysis – Numeric

- Statistical Analysis
  - Point Estimation
    - Count, Min, Max, Average, Median, Mode
  - Dispersion
    - Range, StDev, Var, CV
    - Skewness, Kurtosis
  - Missing Value
  - Outliers
  - Binning
- Visualization
  - Histogram, Box Plot, ...

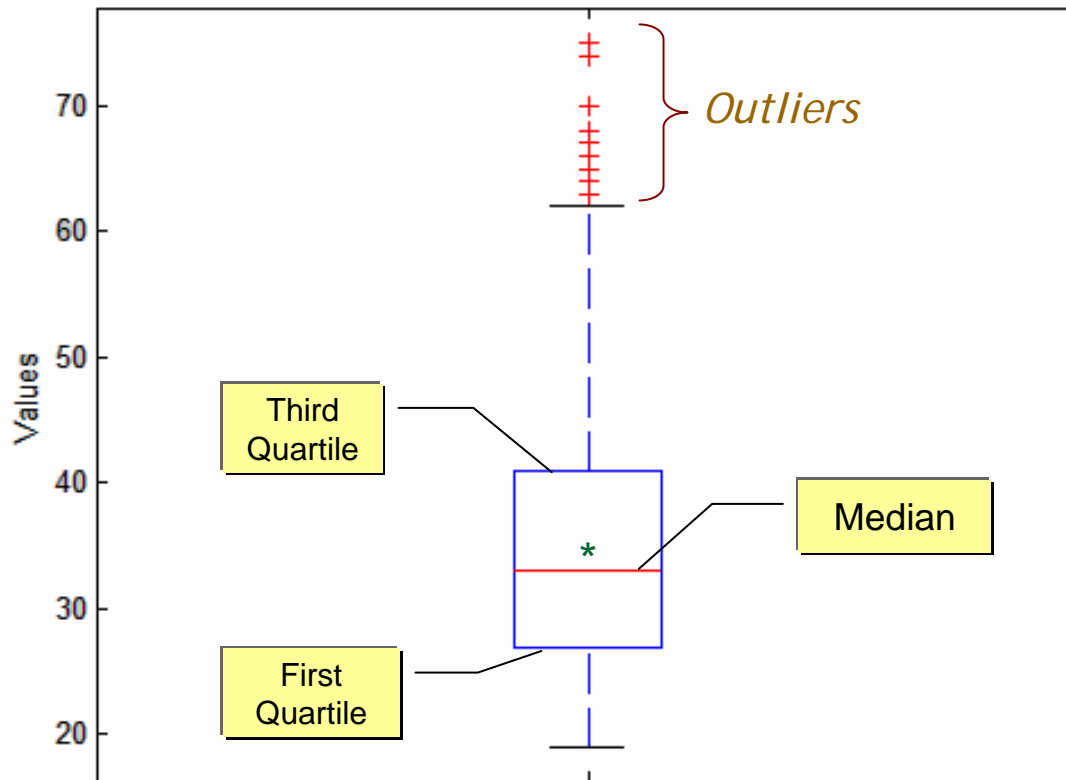


# Univariate Analysis – Numeric

Age					
Count	<b>900</b>	Average	<b>35.25</b>	StDev	<b>11.20</b>
Min	<b>19</b>	Median	<b>33</b>	Variance	<b>125.37</b>
Maximum	<b>75</b>	Mode	<b>27</b>	CV	<b>32%</b>
Range	<b>56</b>	Skewness	<b>1.09</b>		
Missing	<b>0</b>	Kurtosis	<b>0.88</b>		

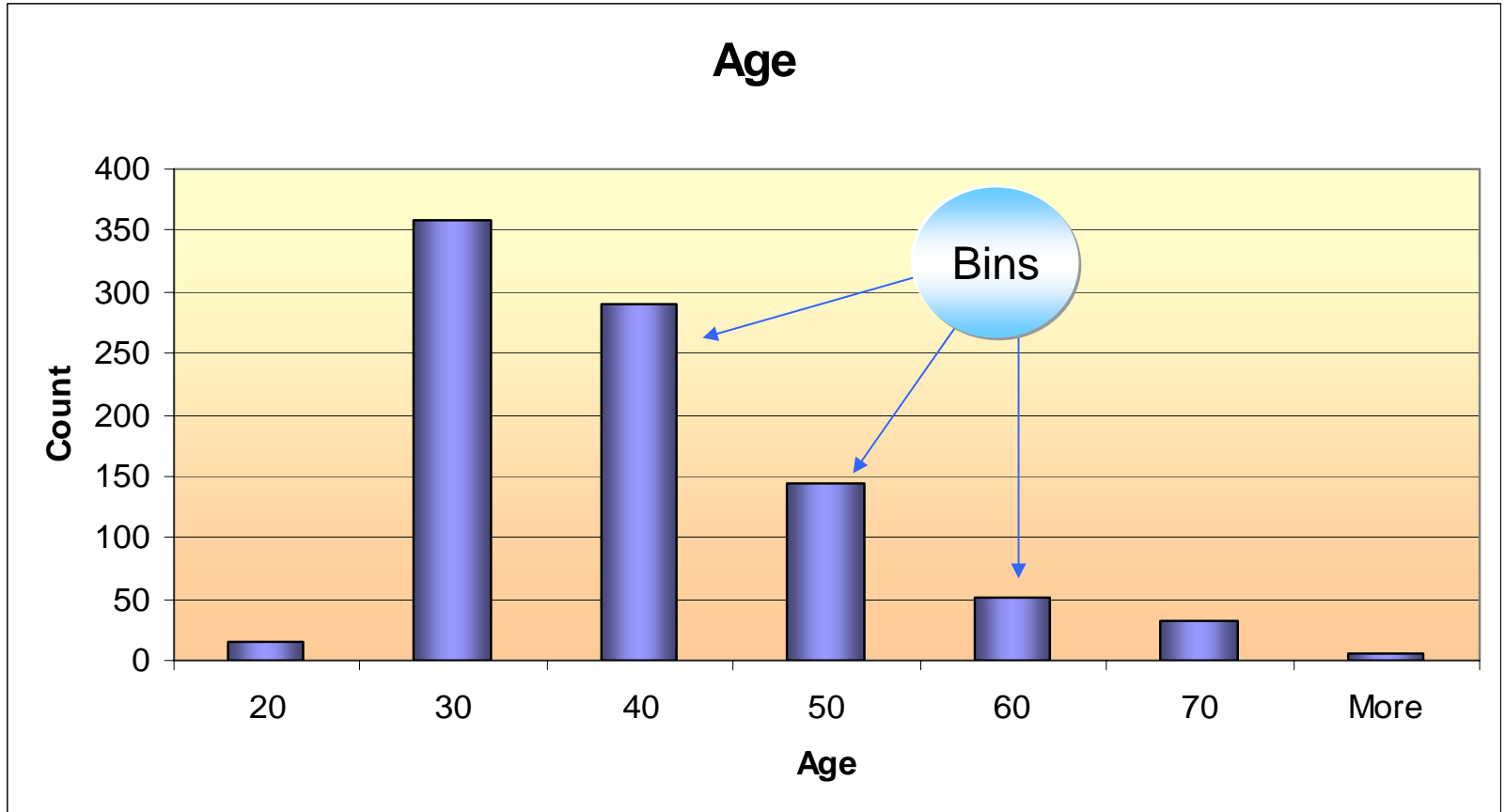
# Univariate Analysis – Numeric

Box Plot



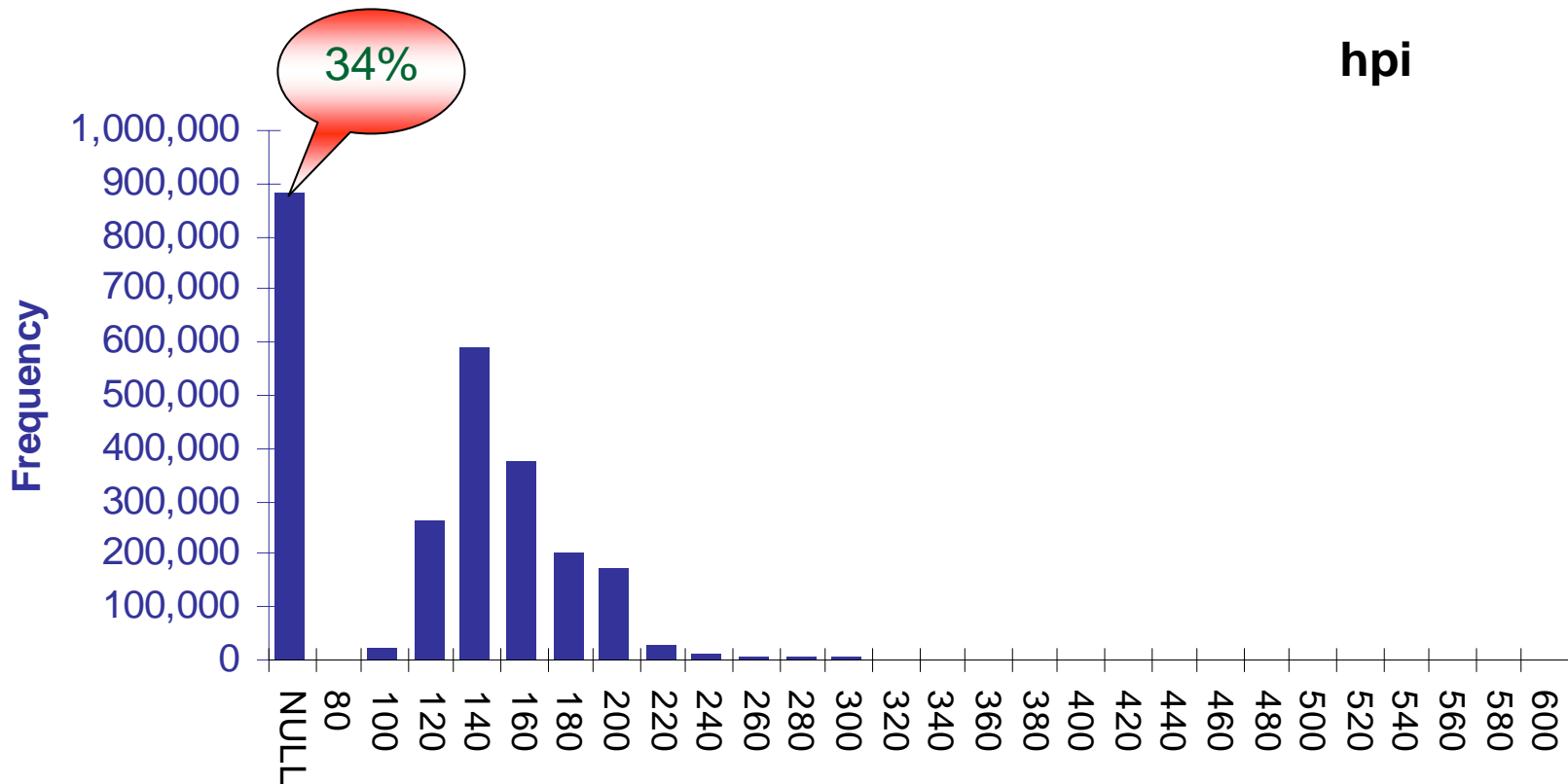
# Univariate Analysis – Numeric

Histogram



# Univariate Analysis – Numeric

Missing Value



# Univariate Analysis - Challenges

Variable	
Categorical	Numeric
Missing Values	Missing Values
Invalid Values	Outliers
Numerization	Binning

---

# Missing Data

- Missing data may be due to
  - Data entry error
  - Data processing error
  - Certain data may not be available at the time of entry
  - ...

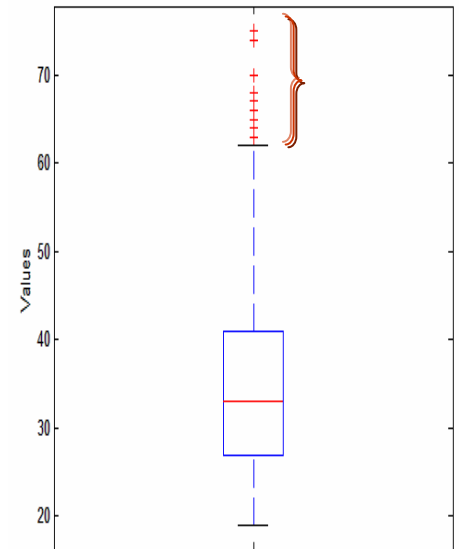
---

# How to Handle Missing Data?

- Fill in missing values manually
- Ignore the records with missing data
- Fill in it automatically:
  - a global constant (e.g., “?”)
  - the variable mean
  - Inference-based methods such as Bayes’ rule, decision tree, or EM algorithm

# Outliers

- Data points inconsistent with the majority of data
- Different outliers
  - Valid: CEO's salary
  - Noisy: One's age = 200, widely deviated points
- Removal methods
  - Box plot
  - Clustering
  - Curve-fitting
  - Hypothesis-testing with a given model



---

# Binning

- **Binning:**
  - is the process of transferring continuous variables into categorical counterparts.
- **Binning methods:**
  - Equal-width
  - Equal-frequency
  - Entropy-based methods



---

# Numerization

- **Numerization:**
  - is the process of transferring categorical variables into numerical counterparts.
  
- **Numerization methods:**
  - Binary method
  - Ordinal Method
  - ...

# Numerization

- **Variable values** (e.g., Housing):
  - for free, own, rent
- **Binary method:**
  - for free: 1, 0, 0
  - own: 0, 1, 0
  - rent: 0, 0, 1
- **Ordinal method:**
  - own: 5
  - for free: 3
  - rent: 1

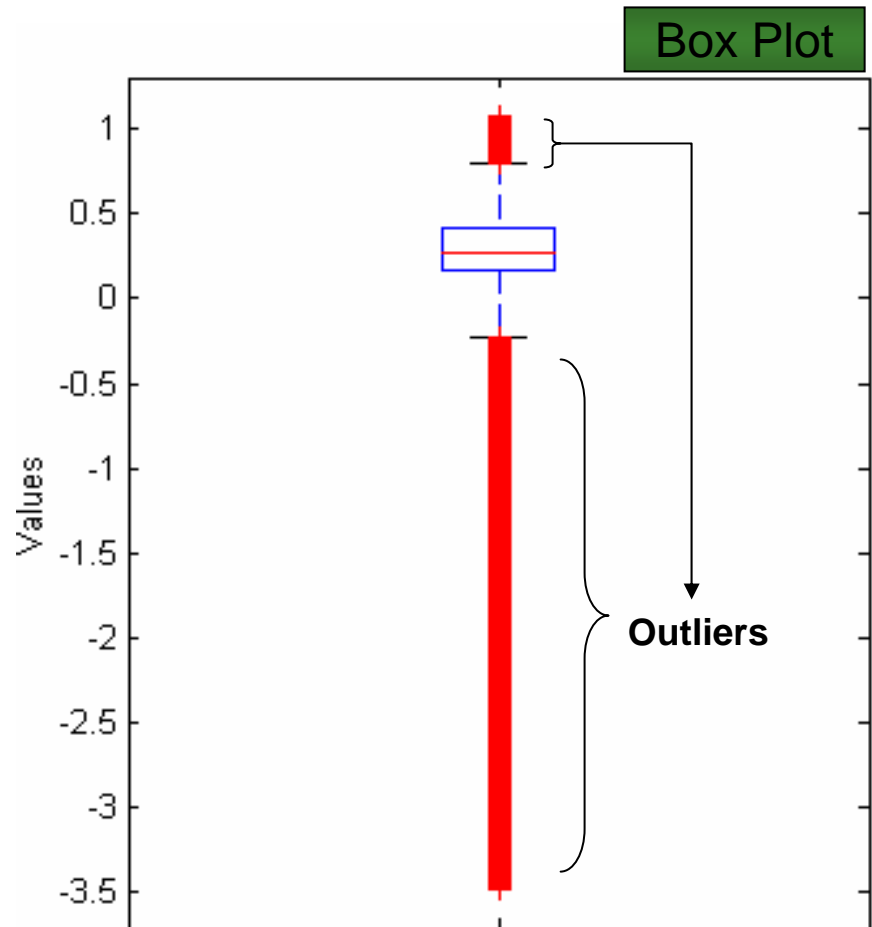
---

# An example from a real world project

- Multi-billions dollars market
- 100+ millions of records
- 500+ variables

# Sale Price Index: Univariate Analysis

SPI	
Count	2576584
Min	-151.25 ←
Max	3681.17 ←
Mean	0.38
Median	0.30
StDev	6.70 ←
Missing	4013 ←



# Sale Price Index : Outliers

Value	Count	Count%
< 0	11064	0.4%
> 1	43845	1.7%

## Modification:

IF SPI < 0 THEN SPI = 0

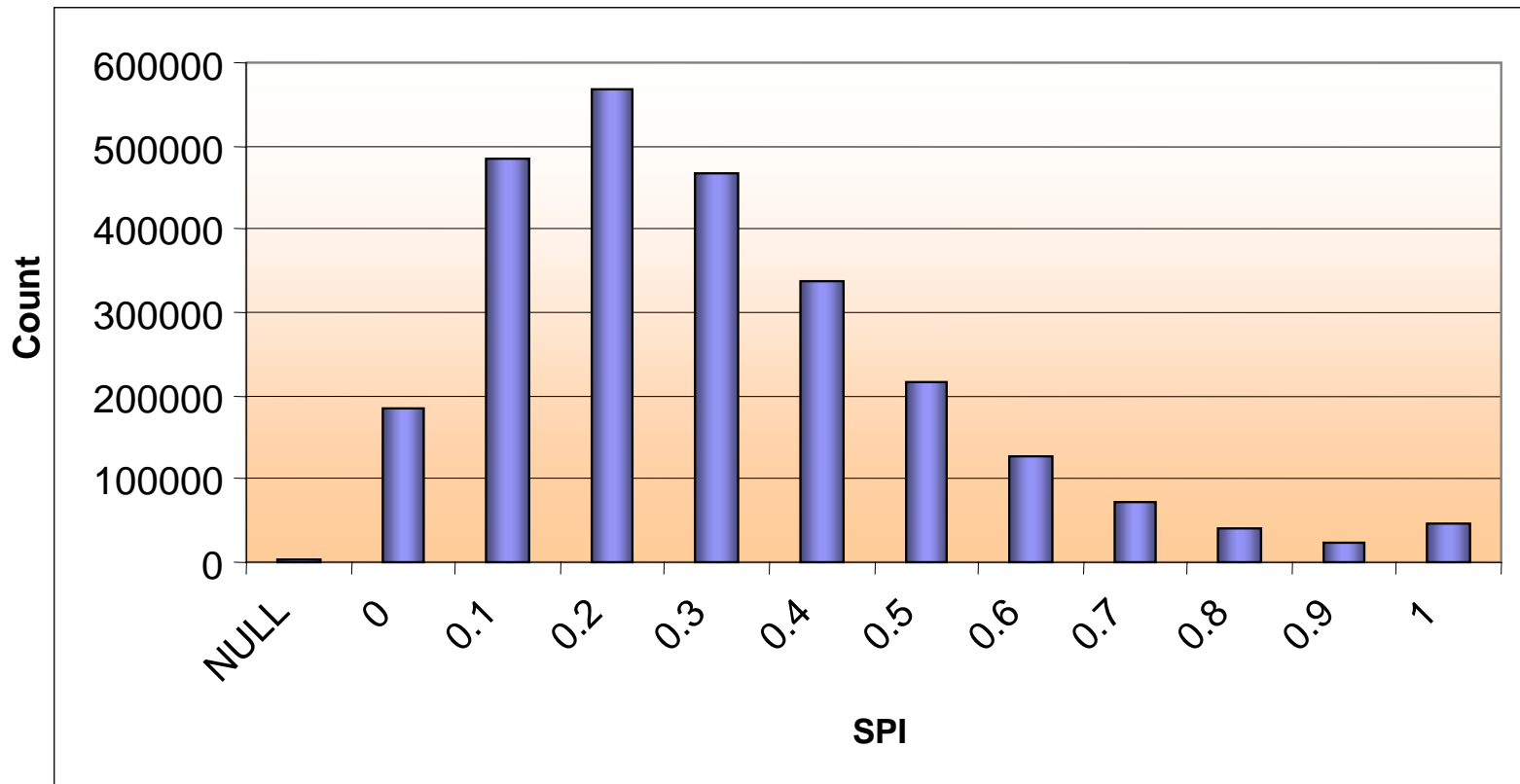
IF SPI > 1 THEN SPI = 1

SPI	Target Rate%
< 0	0.005%
> 1	0.038%
0 to 1	1.814%

# Sale Price Index : after modification


	SPI	SPI*
Count	2576584	2576584
Min	-151.25	0.0 ←
Max	3681.17	1.0 ←
Mean	0.38	0.35
Median	0.31	0.31
StDev	6.70	0.21 ←
Missing	4013	4013 ←

# Sale Price Index : Histogram



# Sale Price Index : Missing values


<i>SPI</i>	Total	Target Rate%
<b>Null</b>	0.16%	0.003%
Not Null	99.84%	1.856%

<i>Target</i>	Likelihood (SPI = <b>Null</b> ) 
0	0.0016
1	0.0017

1. Do nothing!
2. Delete the records with a missing value.
3. The modeling toolbox replaces the missing values.
4. **Replace the missing values with a known value (e.g., 0.5)**

# Sale Price Index : Z test

$$Z = \frac{\mu_1 - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}}$$


 Z	P(Z)
Target	<b>18.56</b> <b>0.0</b>

- The average SPI for the subset with target=1 is **significantly** different from the average SPI for the subset with target=0.
- It shows SPI has some predictive value and should be in the modeling variable list.

# Sale Price Index : Chi<sup>2</sup> test

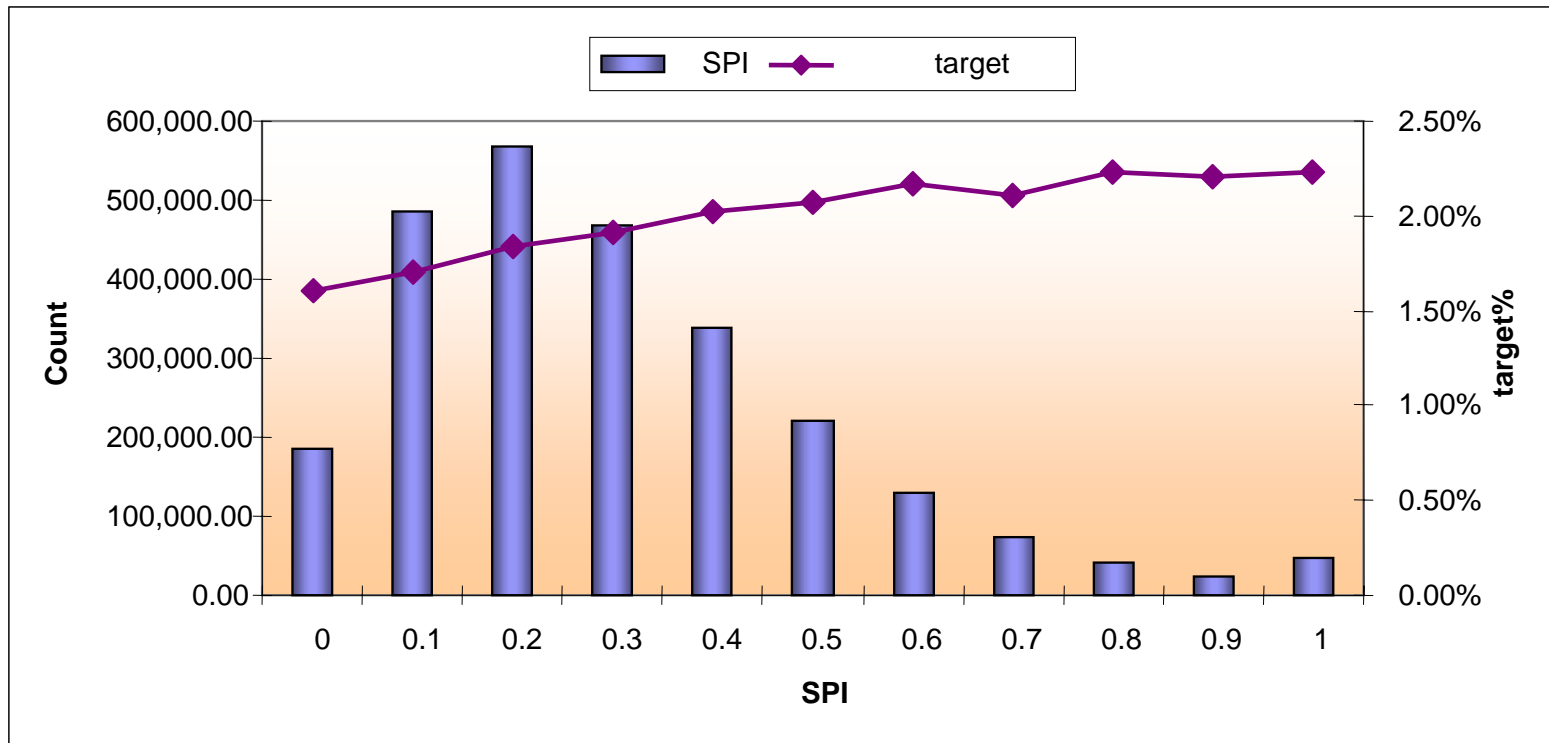
$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

$$df = (r - 1)(c - 1)$$

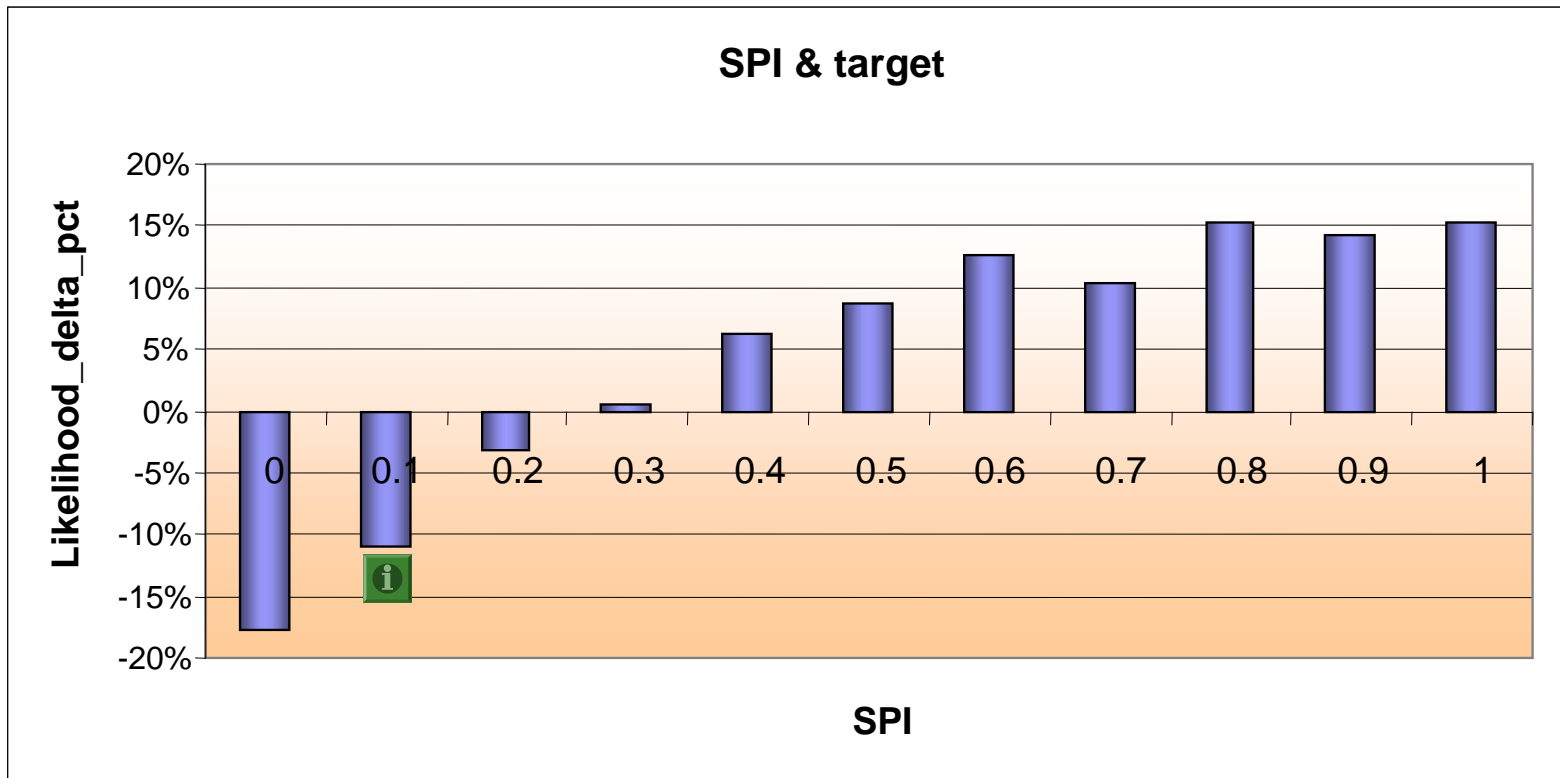
	Chi <sup>2</sup>	P(Chi <sup>2</sup> )
Target	351.85	0.0

- The SPI distribution for the subset with target=1 is **significantly** different from the one with target=0.
- It shows discretized (bin) version of SPI has some predictive value and should be in the modeling variable list.

# Sale Price Index & Target



# Sale Price Index & Target



---

# Thank You!

# Z test : Example

<b>Refi</b>	<b>0</b>	<b>1</b>
Count ( $n$ )	<b>2528670</b>	<b>47914</b>
Mean ( $\mu$ )	<b>0.3494</b>	<b>0.3678</b>
Var ( $\sigma$ )	<b>0.0442</b>	<b>0.0463</b>

$$Z = \frac{\mu_1 - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}} \quad Z = 18.56$$
$$P(Z) = 0$$

# Chi<sup>2</sup> test : Example

<i>Mover</i>	Observed		Expected		<i>Total</i>
	0	1	0	1	
0	184548	975	184394.13	1128.87	185523
0.1	482657	3096	482797.29	2955.71	485753
0.2	563788	3807	564141.30	3453.70	567595
0.3	465564	3016	465728.78	2851.22	468580
0.4	335066	2009	335023.97	2051.03	337075
0.5	219758	1232	219645.32	1344.68	220990
0.6	127329	684	127234.07	778.93	128013
0.7	71696	406	71663.27	438.73	72102
0.8	40430	190	40372.84	247.16	40620
0.9	23308	95	23260.60	142.40	23403
1	46762	168	46644.44	285.56	46930
<b>Total</b>	<b>2560906</b>	<b>15678</b>			<b>2576584</b>

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

$$df = (r-1)(c-1)$$

$$\chi^2 = 176.08$$

$$P = 0.0$$

$$df = 10$$

# Likelihood :Example

$$\begin{aligned}\text{Likelihood} &= P(\text{SPI} = \text{null} \mid \text{target} = 1) \\ &= \frac{\text{freq}(\text{SPI} = \text{null} \ \& \ \text{target} = 1)}{\text{freq}(\text{target} = 1)}\end{aligned}$$

$$\text{freq}(\text{SPI} = \text{null} \ \& \ \text{target} = 1) = 80$$

$$\text{freq}(\text{target} = 1) = 47914$$

$$\text{Likelihood} = \frac{80}{47914} = 0.0017$$

# Likelihood\_delta\_pct : Example

$$\text{Likelihood}_0 = \frac{\text{freq}(\text{SPI} = 0.1 \ \& \ \text{target} = 0)}{\text{freq}(\text{target} = 0)}$$

$$\text{Likelihood}_1 = \frac{\text{freq}(\text{SPI} = 0.1 \ \& \ \text{target} = 1)}{\text{freq}(\text{target} = 1)}$$

$$\text{Likelihood\_delta\_pct} = \frac{(\text{Likelihood}_1 - \text{Likelihood}_0)}{\text{Likelihood}_1}$$

$$\text{Likelihood}_0 = \frac{477591}{2528670} = 0.18887$$

$$\text{Likelihood}_1 = \frac{8162}{47914} = 0.170347$$

$$\text{Likelihood\_delta\_pct} = \frac{(0.170347 - 0.18887)}{0.170347} = -0.10874 = -11\%$$