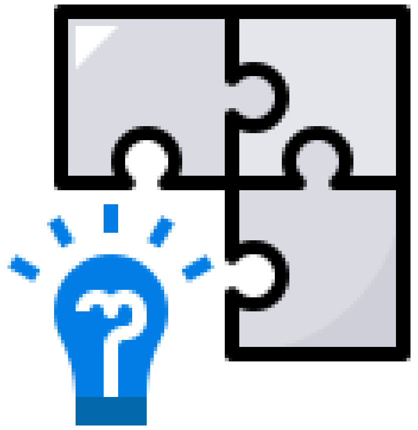
DLI Accelerated Data Science Teaching Kit

Lecture 3.2 - Data Cleaning and Statistical Preprocessing

Data Cleaning

Data cleaning is the preprocessing of missing values and removal data noise.

- **Pre-processing missing values**



- **Removing noise**

Why the Need to Handle Missing Values?

Missing values are generated by collection errors or missing observations.

Missing values are the most common problem in data analytics.

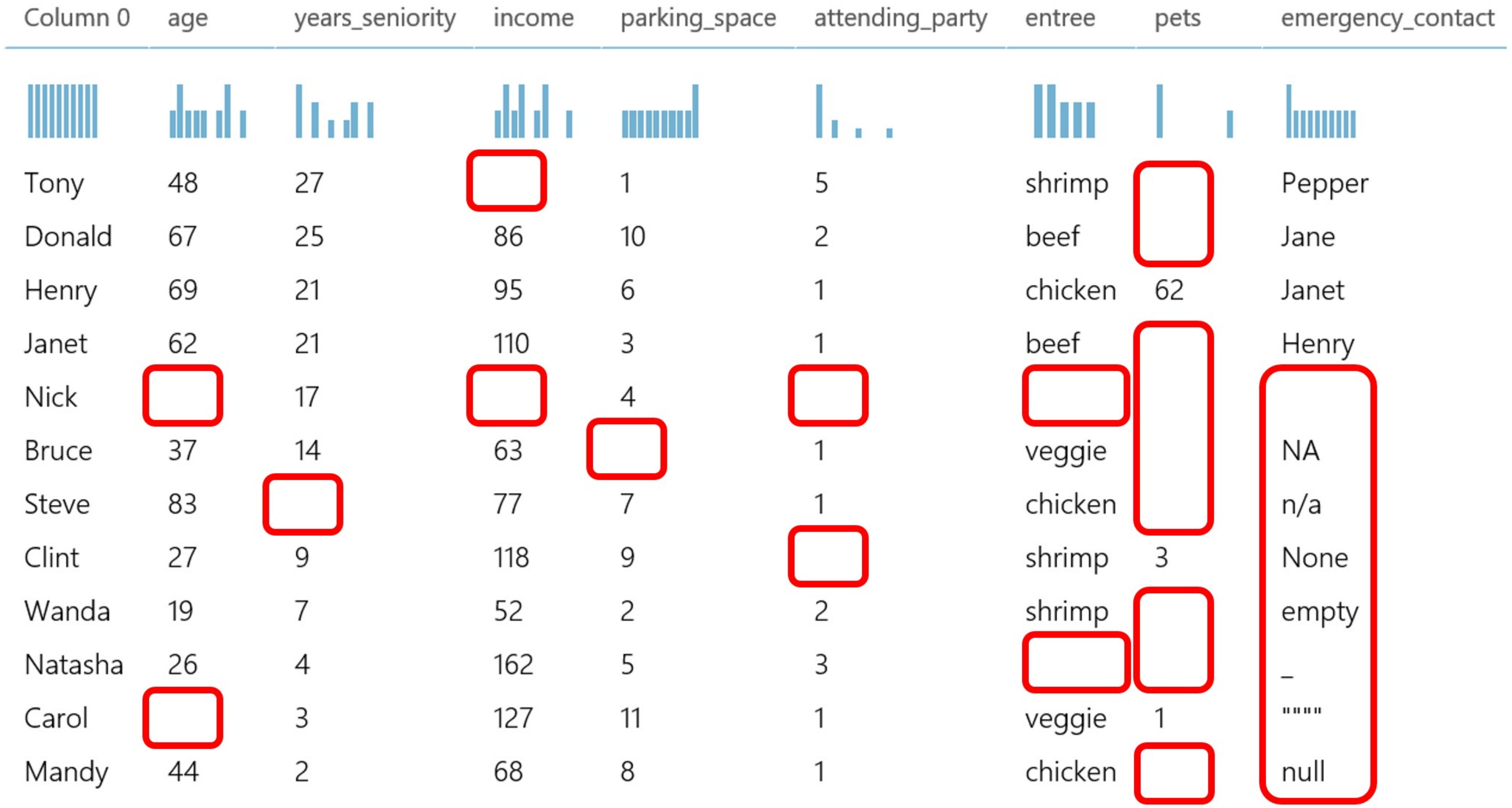
When missing values come out, certain model performance is decreased.

Even for models that can handle missing values, they might be sensitive to it.

How to deal with missing values?

Typical methods to process missing values:

- Deletion
- Dummy substitution
- Mean substitution
- Frequent substitution
- Regression substitution



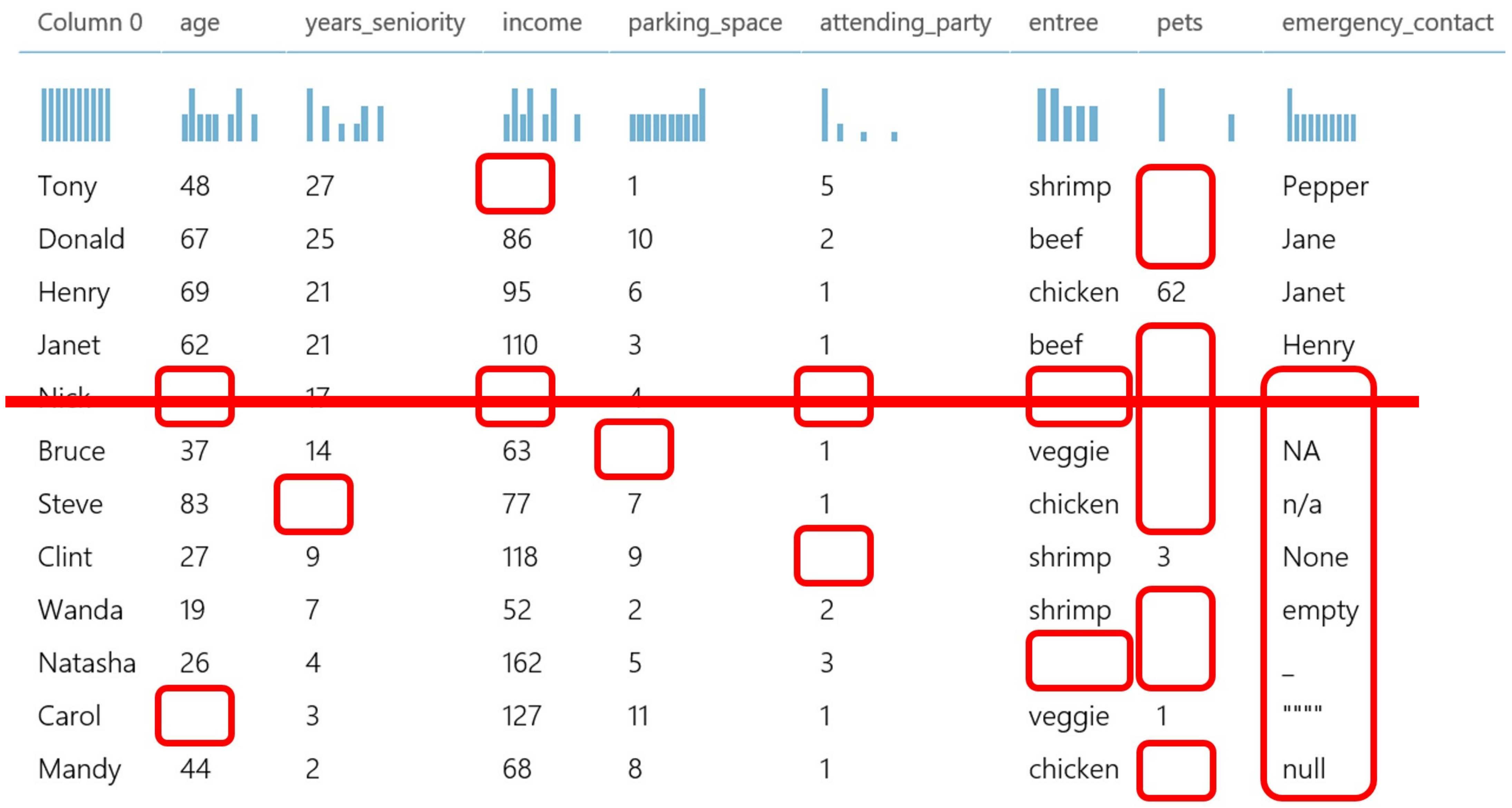
Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Source: https://gallery.azure.ai/Experiment/Methods-for-handling-missing-values-1

How to deal with missing values?

Typical methods to process missing values:

- **Deletion:** removing records with missing values



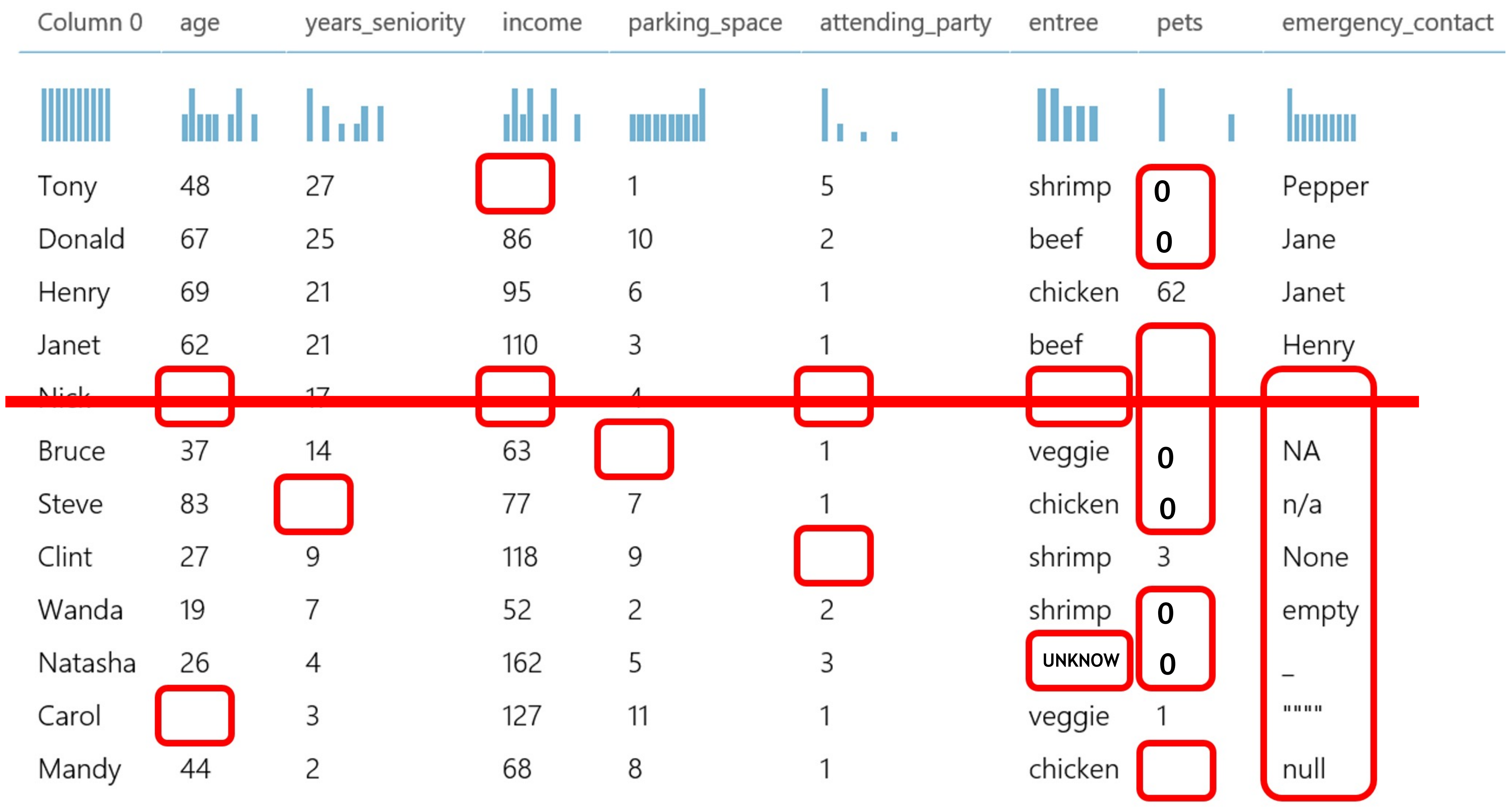
Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
~~Nick~~		~~17~~		~~4~~				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Source: https://gallery.azure.ai/Experiment/Methods-for-handling-missing-values-1

How to deal with missing values?

Typical methods to process missing values:

- **Dummy substitution:** Replace missing values with a dummy value e.g. **UNKNOWN** for category values or **0** for numerical values



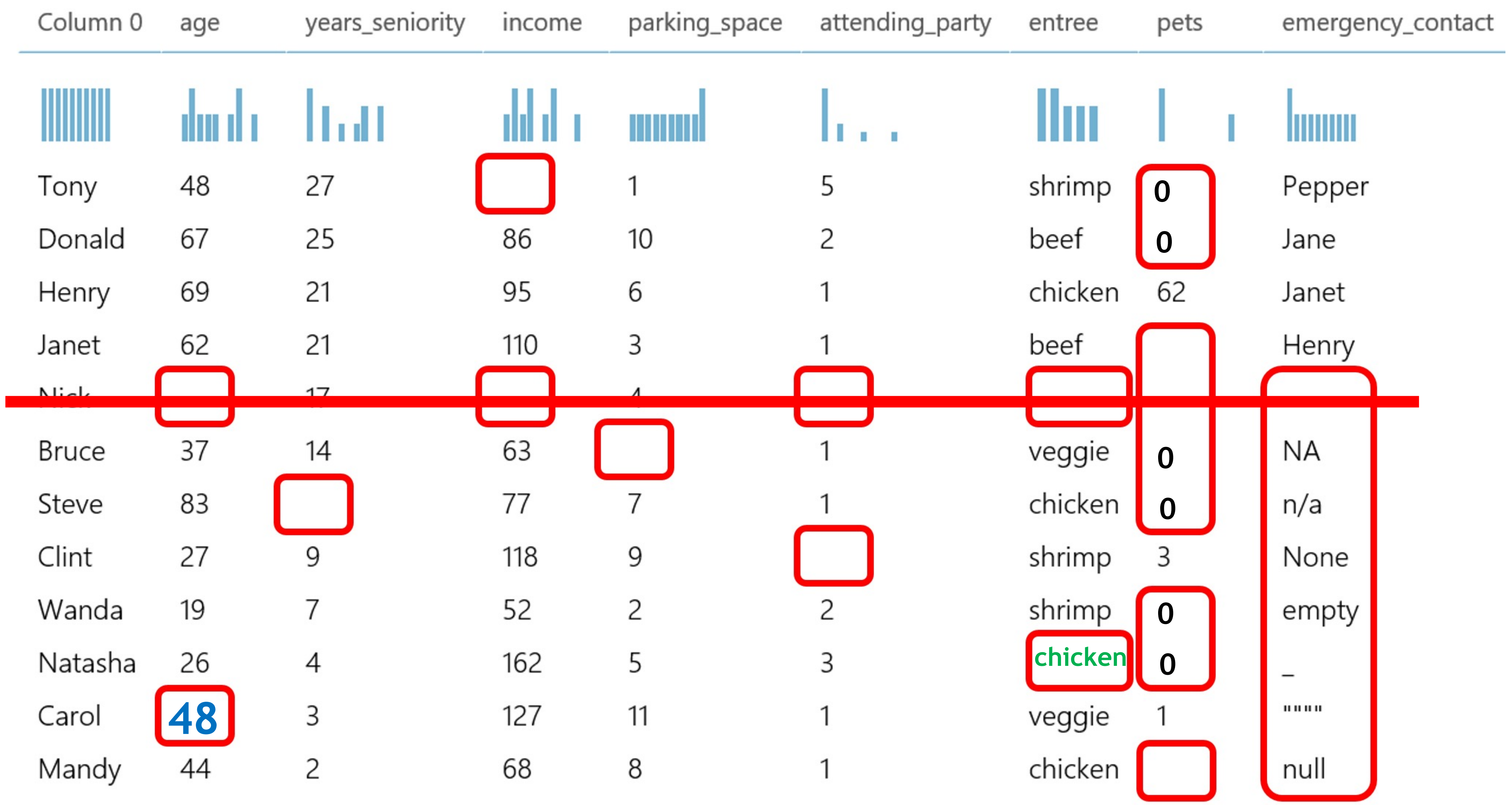
Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp	0	Pepper
Donald	67	25	86	10	2	beef	0	Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
~~Nick~~		~~17~~		~~4~~				
Bruce	37	14	63		1	veggie	0	NA
Steve	83		77	7	1	chicken	0	n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp	0	empty
Natasha	26	4	162	5	3	UNKNOW	0	_
Carol		3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Source: https://gallery.azure.ai/Experiment/Methods-for-handling-missing-values-1

How to deal with missing values?

Typical methods to process missing values:

- **Frequent substitution:** If the missing values are categorical, replace the missing values with the most frequent item.



Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp	0	Pepper
Donald	67	25	86	10	2	beef	0	Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
~~Nick~~		~~17~~		~~4~~				
Bruce	37	14	63		1	veggie	0	NA
Steve	83		77	7	1	chicken	0	n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp	0	empty
Natasha	26	4	162	5	3	chicken	0	_
Carol	48	3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Source: https://gallery.azure.ai/Experiment/Methods-for-handling-missing-values-1

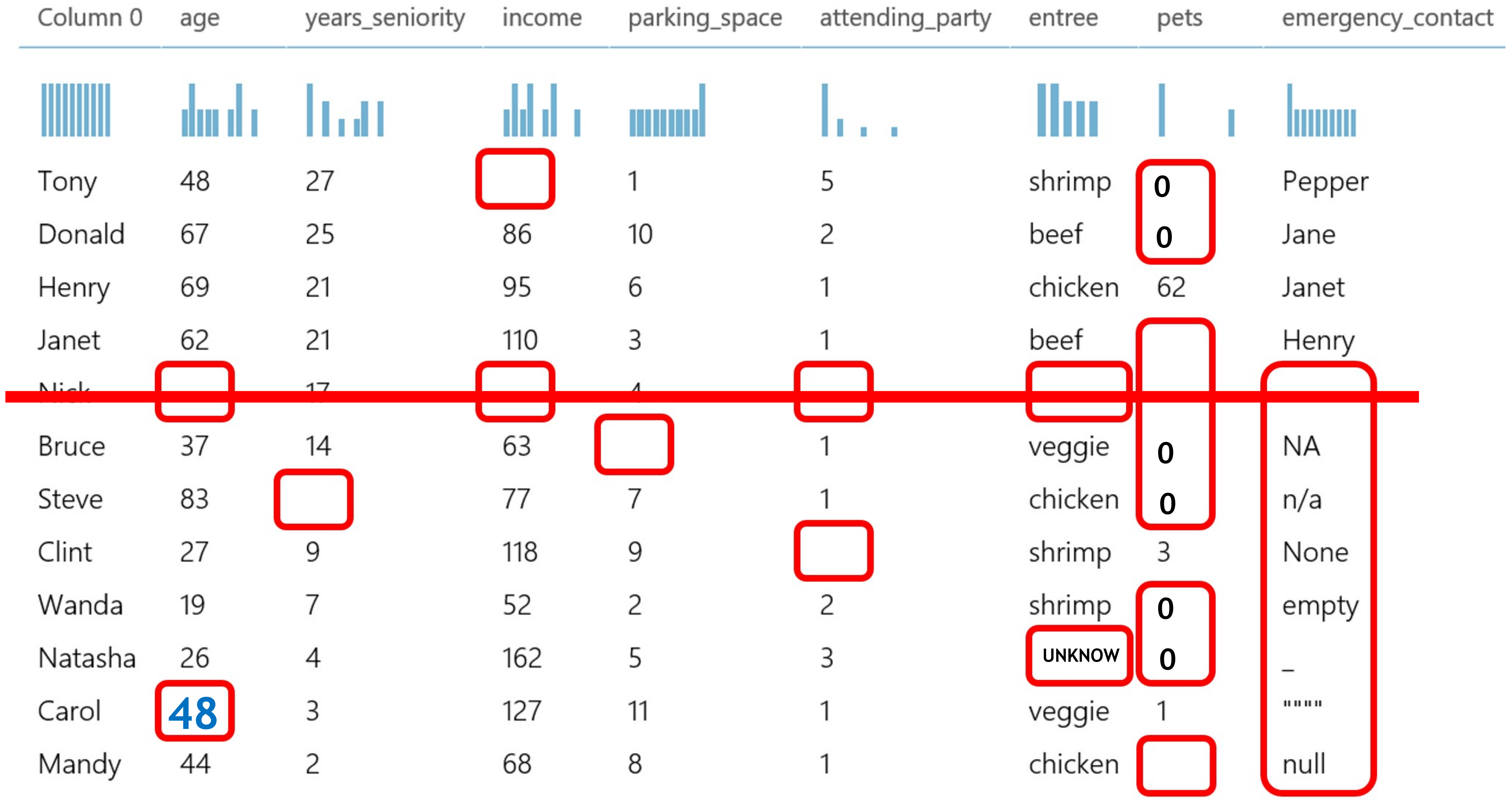
How to deal with missing values?

Typical methods to process missing values:

- **Mean substitution:**

$$x_{mean} = \sum_{i=1}^{n} x_i / n$$

$\boldsymbol{n}$: number of values



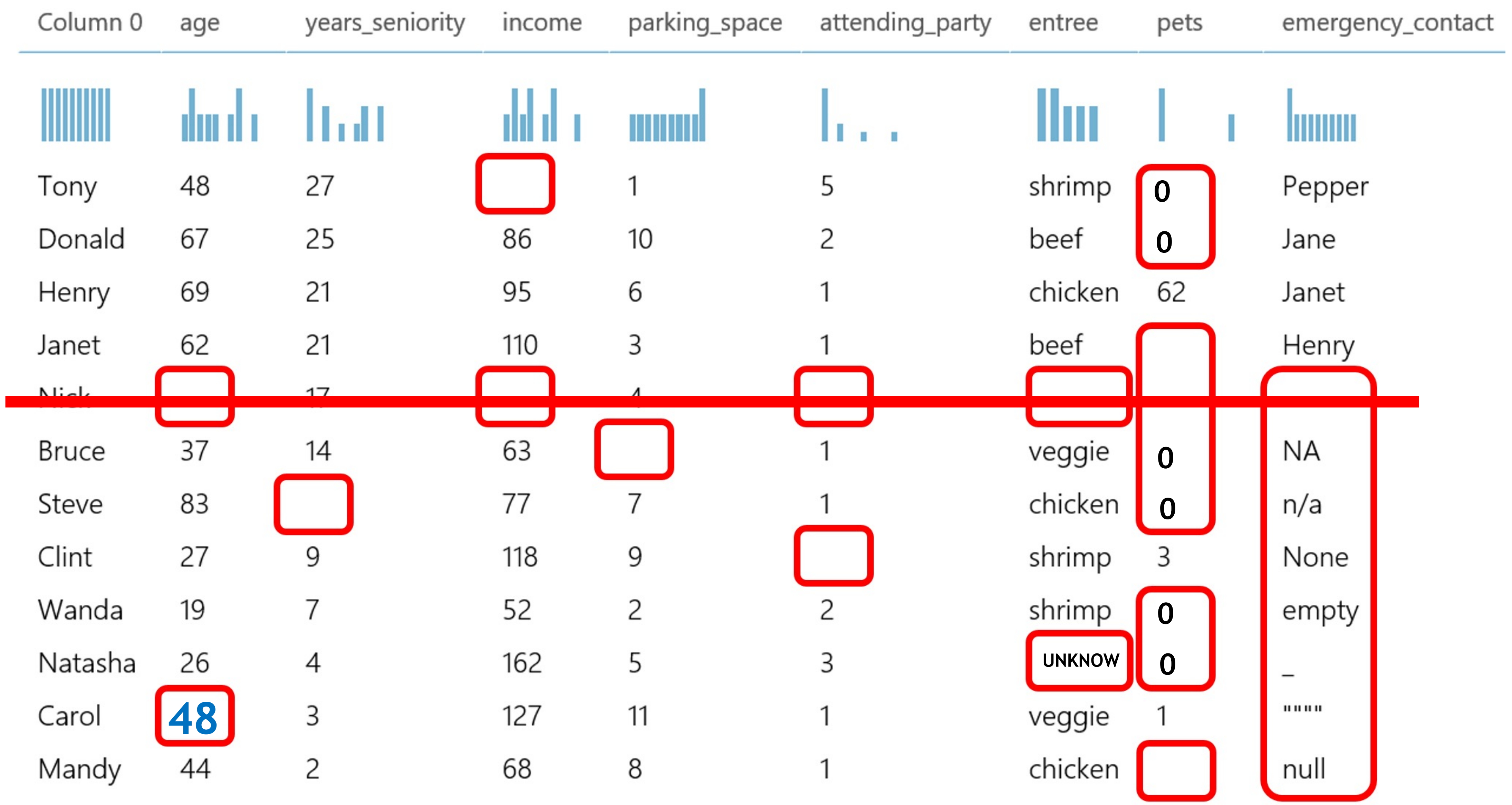
Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp	0	Pepper
Donald	67	25	86	10	2	beef	0	Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
~~Nick~~		~~17~~		~~4~~				
Bruce	37	14	63		1	veggie	0	NA
Steve	83		77	7	1	chicken	0	n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp	0	empty
Natasha	26	4	162	5	3	UNKNOW	0	_
Carol	48	3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

Source: https://gallery.azure.ai/Experiment/Methods-for-handling-missing-values-1

How to deal with missing values?

Typical methods to process missing values:

- **Regression substitution:** Use a regression method to replace the missing values with regressed values.
  - Time-series data
  - **ARIMA** [1] (autoregressive integrated moving average)



Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp	0	Pepper
Donald	67	25	86	10	2	beef	0	Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
~~Nick~~		~~17~~		~~4~~				
Bruce	37	14	63		1	veggie	0	NA
Steve	83		77	7	1	chicken	0	n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp	0	empty
Natasha	26	4	162	5	3	UNKNOW	0	_
Carol	48	3	127	11	1	veggie	1	""""
Mandy	44	2	68	8	1	chicken		null

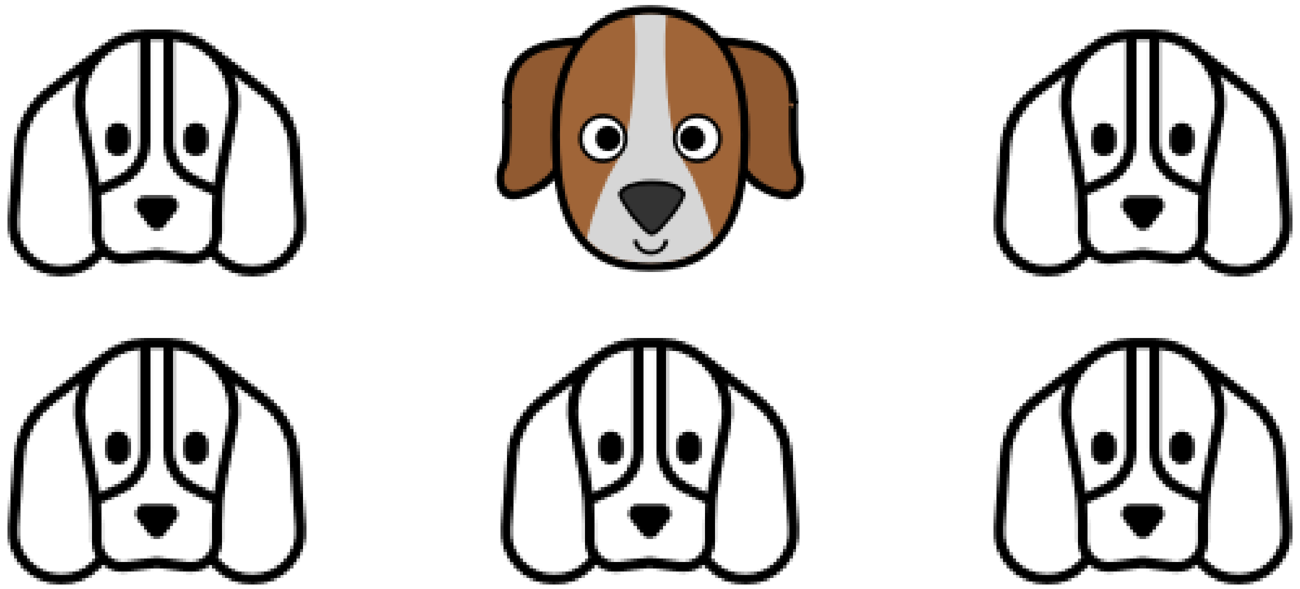
[1] Contreras, J., Espinola, R., Nogales, F. J., & Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. IEEE transactions on power systems, 18(3), 1014-1020.c

Source: https://gallery.azure.ai/Experiment/Methods-for-handling-missing-values-1

Outliers

An outlier is a data point that differs significantly from other observations.

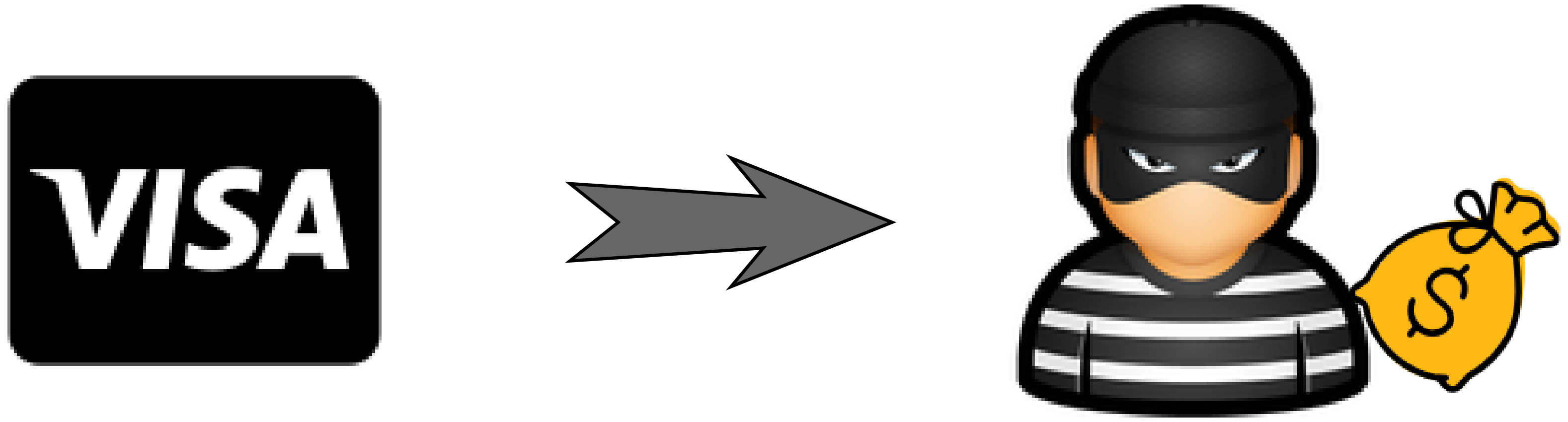
- Outliers can be very common in multidimensional data.
- Outliers can be results of bad data collection.
- Outliers would distort the models.
- Some models are sensitive to outliers.
- Sometimes outliers are the interesting data points.



How to deal with outliers?

It depends on how to generate the outliers.

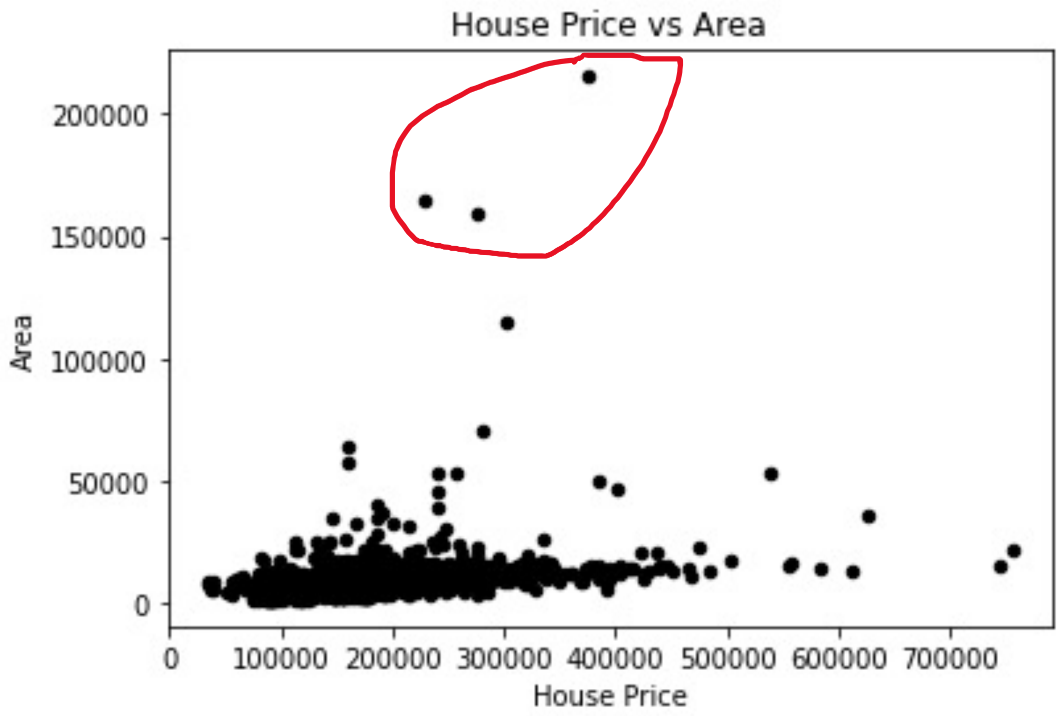
- **Keep outliers**
  - We should pay more attention to the outliers since they may be genuine observations in the collected data.
  - In many applications, outliers provide crucial information for data analytics.



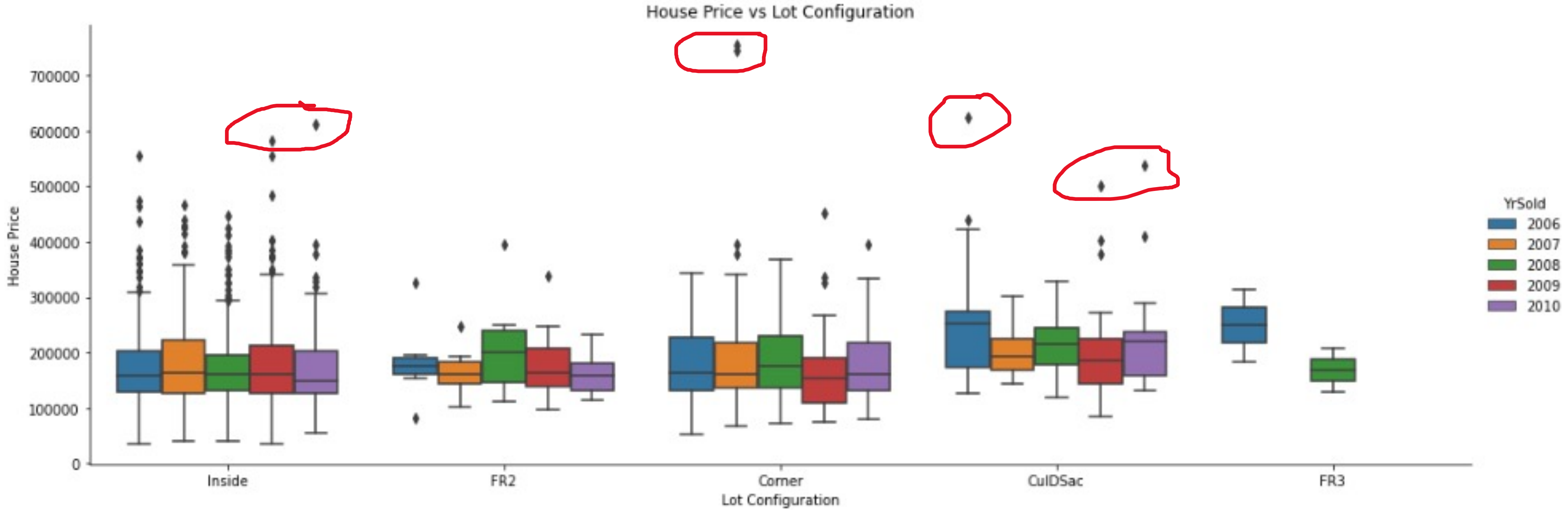
How to detect outliers?

Two popular methods:

- **Scatter Plot**



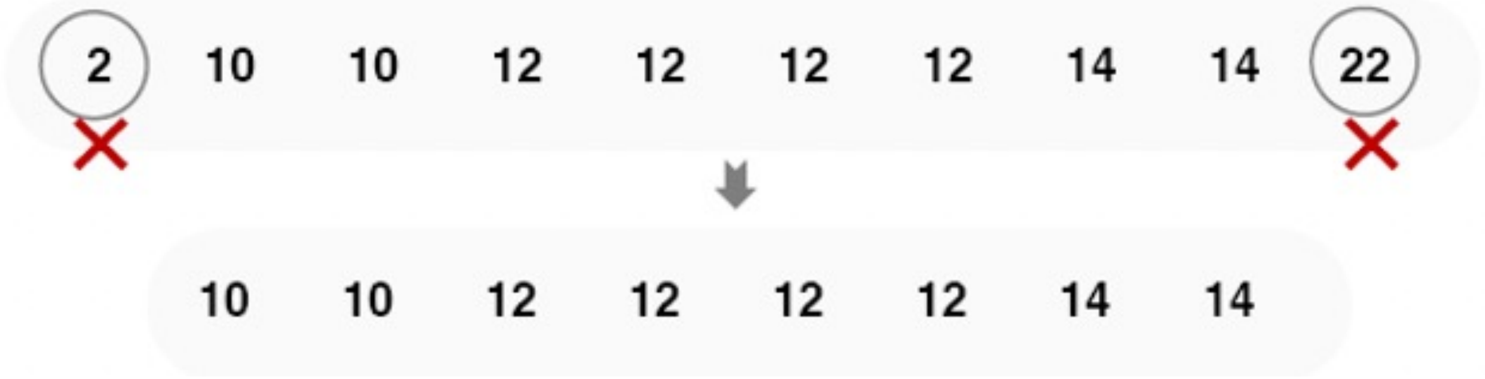
- **Box Plot**



How to deal with outliers?

It depends on how to generate the outliers.

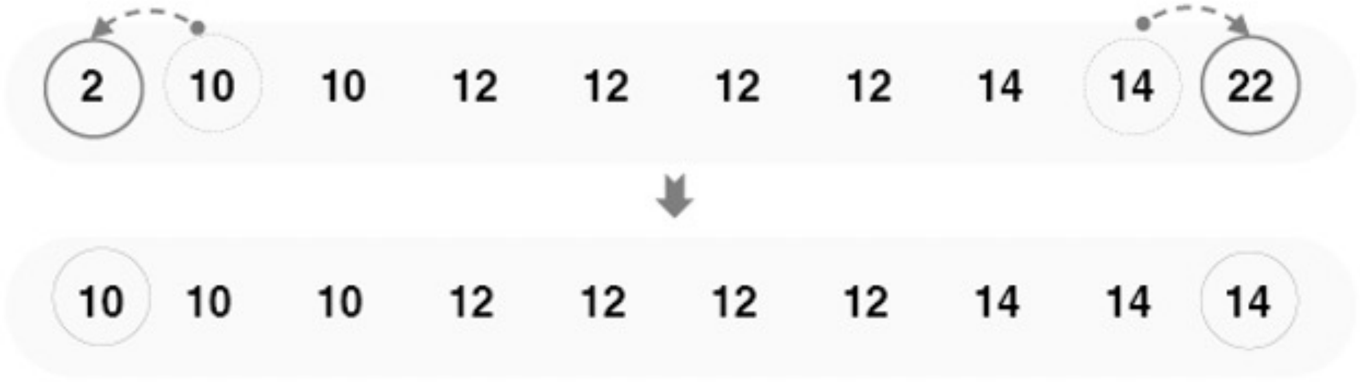
- **Exclude outliers**
  - Trimming: discarding the outliers.



2 10 10 12 12 12 12 14 14 22

10 10 12 12 12 12 14 14

  - Replacement: replacing the outliers with the nearest “normal” data point



2 10 10 12 12 12 12 14 14 22

10 10 10 12 12 12 12 14 14 14

Data Normalization

Data normalization is the rescaling of numerical values to a specific range.

Two popular methods:

- **Min-Max Normalization:**

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- **Z-score Normalization (or Standardization)**

$$\boldsymbol{X_{Z-score}} = \frac{\boldsymbol{X - \mu}}{\boldsymbol{\delta}}$$

$\boldsymbol{\mu}$: mean of X

δ: standard deviation of X

Data Down-Sampling

Data down-sampling is reducing large data to a smaller and more manageable size.

- **Record down-sampling (clustering):** select the records and only choose the representative subset from the data.

- **Attribute down-sampling (Feature selection):** select only a subset of the most important attributes from the data.

DLI Accelerated Data Science Teaching Kit

Thank You