

Limpeza de Dados – Valores Faltantes



– Aula 14 –
Coleta, Preparação e
Análise de Dados

Prof. Me. Lucas R. C. Pessutto



PUCRS

Pontifícia Universidade Católica
do Rio Grande do Sul



Slides adaptados do material do Prof. Lucas Silveira
Kupssinskü e do Prof. Luan Fonseca Garcia






























Data Cleaning

- O processo de limpeza dos dados pode ser separado em dois processos:
 - Pré-processamento (ou imputação) de valores faltantes.
 - Remoção de Ruído

Dados faltantes






























- Dados faltantes podem ser ocasionados por diversos motivos
 - Problemas na coleta;
 - Observações que faltaram coletar;
 - Podem apresentar informação.
- Mesmo modelos robustos contra dados faltantes podem se beneficiar desse tipo de processamento.

O que fazer?

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










O que fazer?

- Deleção:
 - Remover toda a observação.

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		1				
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










O que fazer?

- Deleção:
 - Remover toda a observação.
- Substituição Dummy:
 - Substituir por um valor dummy (0 ou Unknown).

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	0	Henry
Nick		17		1				
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	Unknown	0	-
Carol		3	127	11	1	veggie	1	""
Mandy	44	2	68	8	1	chicken	0	null










O que fazer?

- Deleção:
 - Remover toda a observação.
- Substituição Dummy:
 - Substituir por um valor dummy (0 ou Unknown).
- Substituição pela Média:
 - Substituir pelo valor médio.

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	0	Henry
Nick		17		1				
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	Unknown	0	-
Carol	48	3	127	11	1	veggie	1	""
Mandy	44	2	68	8	1	chicken	0	null










O que fazer?

- Deleção:
 - Remover toda a observação.
- Substituição Dummy:
 - Substituir por um valor dummy (0 ou Unknown).
- Substituição pela Média:
 - Substituir pelo valor médio.
- Substituição por item frequente:
 - Substituir pela moda.

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	0	Henry
Nick		17		1				
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	0	null

O que fazer?

- Deleção:
 - Remover toda a observação.
- Substituição Dummy:
 - Substituir por um valor dummy (0 ou Unknown).
- Substituição pela Média:
 - Substituir pelo valor médio.
- Substituição por item frequente:
 - Substituir pela moda.
- Substituição por regressão/class.
 - Usar um método de regressão ou classificação para estimar o valor.

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
								
Tony	48	27	110	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	0	Henry
Nick		17		1				
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	0	null