



Data Science & Business Analytics

Machine Learning Models

David Issá

davidribeiro.issa@gmail.com

1. Pré-processamento de dados

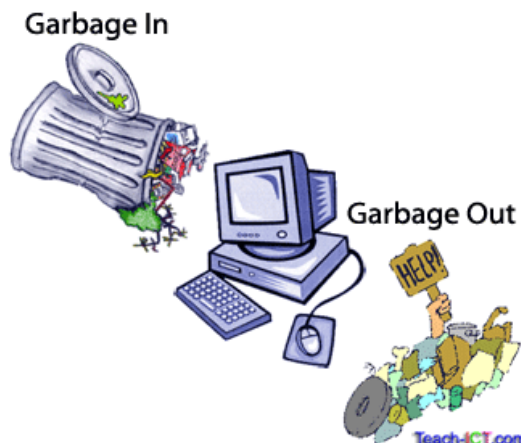
1. Pré-processamento de dados

Processo de **transformação de dados** em formato bruto **para um estado** que pode ser **facilmente utilizado por um algoritmo**.

“Garbage in, garbage out”

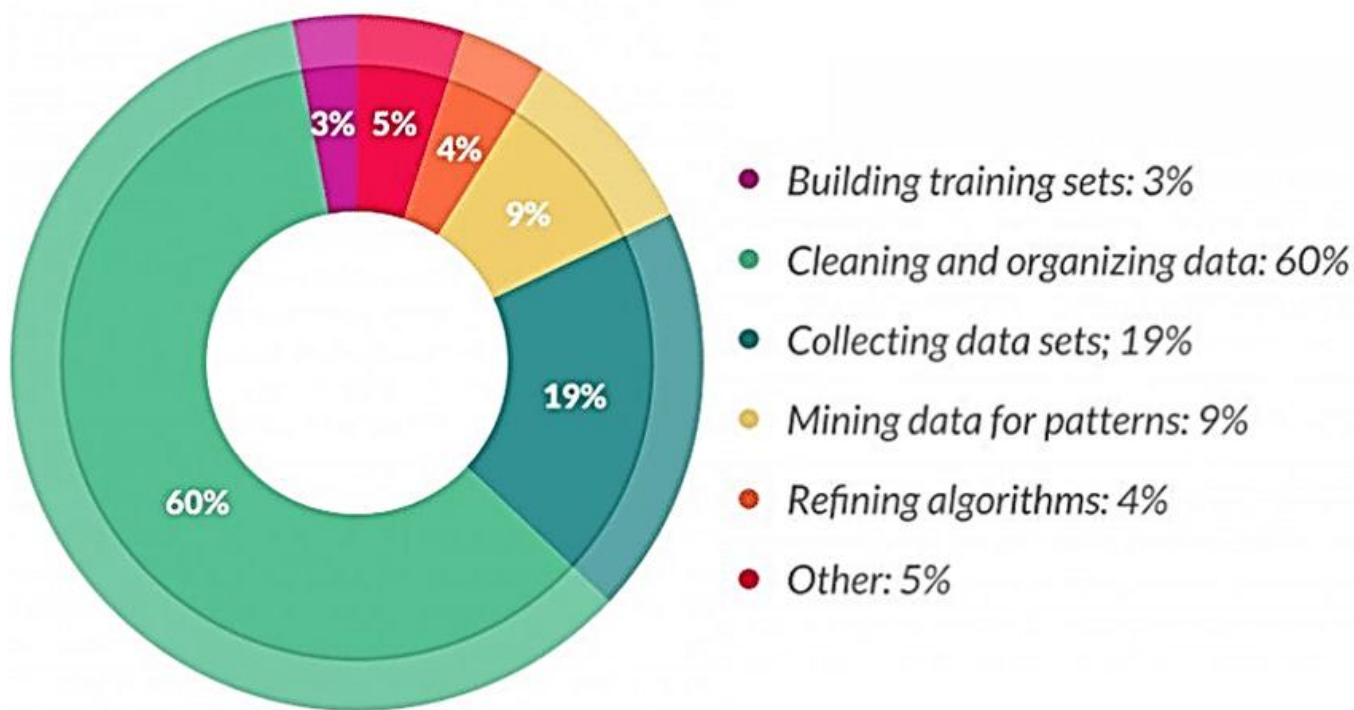


Your analysis is as good as your data.



Objetivo desta etapa: **minimizar GIGO**. Corresponde a cerca de **60% do trabalho de um Data Scientist**.

1. Pré-processamento de dados



1.1 Data Cleaning

1.1 Data Cleaning

ID	Código Postal	Género	Rendimento	Idade	Estado Civil	Montante Transacionado
1	10048	M	75.000	C	M	5.000
2	J2S7K7	F	-40.000	40	W	4.000
3	90210		10.000.000	45	S	7.000
4	6269	M	50.000	0	S	1.000
5	55101	F	99.999	36	D	3.000

Código postal diferentes?

- Nem todos os países utilizam o mesmo formato ,
- 10048 (EUA) vs J2S7K7(Canada),

Rendimento = -40.000\$

- Erro: valor além dos limites esperado para o rendimento.
- Causado por um erro de introdução de dados?

O campo Idade contém “C”

- Outros registos têm valores numéricos.
- O valor deve ser resolvido.

Código postal com 4 dígitos?

- Adicionar zero à esquerda? (6269 vs 06269)
- Campo da base de dados como numérico (apaga o zero inicial)

Missing value**Rendimento = 10.000.000\$**

- Pode ser considerado um outlier (valor extremo).

Rendimento = 99.999\$

- Outros valores arredondados para os \$5.000 mais próximos.
- O valor pode ser totalmente válido.
- Pode representar o código utilizado para indicar valores em falta.

Idade = 0

- Valor zero utilizado para indicar valor em falta/desconhecido?
- Campo de idade pode tornar-se obsoleto.
- Utilizar a data de nascimento e depois derivar idade.

ID	Código Postal	Género	Rendimento	Idade	Estado Civil	Montante Transacionado
1	10048	M	75.000	C	M	5.000
2	J2S7K7	F	-40.000	40	W	4.000
3	90210		10.000.000	45	S	7.000
4	6269	M	50.000	0	S	1.000
5	55101	F	99.999	36	D	3.000

O campo do estado civil contém “S”

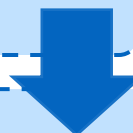
- O que é que este símbolo significa?
- O “S” significa solteiro ou separado?

1.1 Data Cleaning – Missing Values



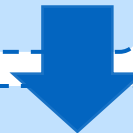
O que são?

Observação sem o valor preenchido numa determinada variável.



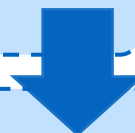
Porque nos devemos preocupar?

Reduzem a qualidade dos dados, levando a modelos pouco fiáveis e enviados.



Como detetar?

Sumário descritivo dos dados.



O que podemos fazer?

Remover, preencher com constante, com média/moda, aplicar um modelo preditivo, etc.

1.1 Data Cleaning – Missing Values

1. Missing completely at random (MCAR)

As causas dos dados em falta não estão relacionadas com os restantes dados.

Nem os valores em falta da variável nem as outras variáveis do conjunto de dados prevêm se um valor estará em falta.

Por exemplo: valores laboratoriais em falta porque um lote de amostras foi processado incorretamente.

2. Missing at random (MAR)

Outras variáveis (mas não a própria variável com valores em falta) no conjunto de dados podem ser utilizadas para prever a ausência de dados.

Por exemplo: Os homens podem ter mais probabilidades de se recusarem a responder a algumas perguntas do que as mulheres.

3. Missing not at random (MNAR)

O valor não observado da variável em falta está relacionado com a razão pela qual está em falta.

Por exemplo: Os indivíduos com rendimentos muito elevados têm mais probabilidades de não responder a perguntas sobre o seu próprio rendimento.

1.1 Data Cleaning – Missing Values

Estratégia #1: Apagar observações com valores em falta

ID	Rendimento	Idade	Estado Civil	Origem
1	1.370	27	Single	US
2		36	Divorced	Europe
3	1.590	42	Married	US
4	0	12		France
5	1.370	44	Divorced	Japan

Devemos eliminar os registos que contêm valores em falta?

- Não é necessariamente a melhor abordagem;
- O padrão dos valores em falta pode ser sistemático;
- A eliminação de registos pode criar um subconjunto enviesado;
- Perde-se informação valiosa noutros campos.

1.1 Data Cleaning – Missing Values

Estratégia #2: Preencher com constante

ID	Rendimento	Idade	Estado Civil	Origem
1	1.370	27	Single	US
2	999	36	Divorced	Europe
3	1.590	42	Married	US
4	0	12	Missing	France
5	1.370	44	Divorced	Japan

- Valores numéricos em falta substituídos por 999;
- Valores categóricos em falta substituídos por “Missing”.

1.1 Data Cleaning – Missing Values

Estratégia #2: Preencher com estatística da variável

ID	Rendimento	Idade	Estado Civil	Origem
1	1.370	27	Single	US
2	1028.5	36	Divorced	Europe
3	1.590	42	Married	US
4	0	12	Divorced	France
5	1.370	44	Divorced	Japan

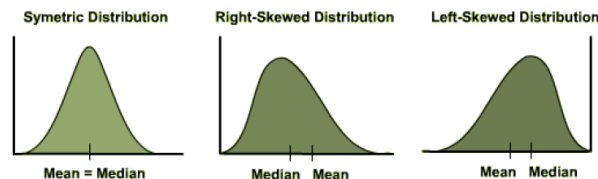
- Valores numéricos em falta substituídos pela média (1082.5);
- Valores categóricos em falta substituídos pela moda (Divorced).

1.1 Data Cleaning – Missing Values

Estratégia #2: Preencher com estatística da variável

ID	Rendimento	Idade	Estado Civil	Origem
1	1.370	27	Single	US
2	1028.5	36	Divorced	Europe
3	1.590	42	Married	US
4	0	12	Divorced	France
5	1.370	44	Divorced	Japan

- Valores numéricos em falta substituídos pela média (1082.5);
- Valores categóricos em falta substituídos pela moda (Divorced);
- A média nem sempre é a melhor escolha para o valor típico. Pode fazer mais sentido usar a mediana.



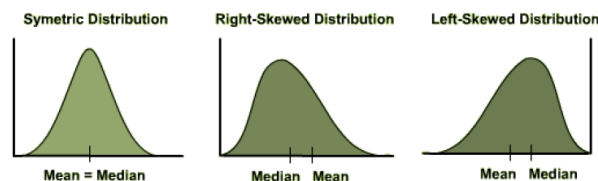
- **Mean:** 1082.5
- **Median:** 1370

1.1 Data Cleaning – Missing Values

Estratégia #3: Preencher com valor aleatório (dos valores disponíveis na variável)

ID	Rendimento	Idade	Estado Civil	Origem
1	1.370	27	Single	US
2	1.370	36	Divorced	Europe
3	1.590	42	Married	US
4	0	12	Single	France
5	1.370	44	Divorced	Japan

- Valores retirados aleatoriamente da distribuição subjacente;
- Método superior ao da substituição pela média da variável.



- Mean: 1082.5

- Median: 1370

1.1 Data Cleaning – Missing Values

Estratégia #4: Aplicar modelo preditivo

ID	Rendimento	Idade	Estado Civil	Origem
1	1.370	27	Single	US
2	1.590	36	Divorced	Europe
3	1.590	42	Married	US
4	0	12		France
5	1.370	44	Divorced	Japan

KNN (K-Nearest Neighbors) Imputer

- Faz corresponder um ponto com os seus k vizinhos mais próximos num espaço multidimensional;
- Suponhamos que vamos utilizar apenas a idade para estimar o Rendimento;
- O vizinho mais próximo de {36, Missing} é {42, 1590};
- Preenchemos o valor de Rendimento em falta com 1590.

Podemos fazer o mesmo na variável Estado Civil (depois de fazer encoding dos dados categóricos – veremos como mais à frente...)

1.1 Data Cleaning – Incoerências

ID	Rendimento	Idade	Estado Civil	Origem
1	1.370	27	Single	US
2	1028.5	36	Divorced	Europe
3	1.590	42	Married	US
4	0	12	Divorced	France
5	1.370	44	Divorced	Japan

Verificar se os valores são válidos e coerentes!

Divorciado aos 12 anos?

- Não há garantia de que estes dados façam sentido;
- Os métodos alternativos esforçam-se por substituir os missing values mais precisamente

Europa , França , EUA , EUA

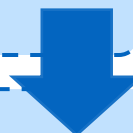
- Registos classificados de forma inconsistente no que diz respeito à origem do cliente
- Manter a coerência: USA & US → North America e France → Europe

1.1 Data Cleaning – Outliers



O que são?

Observação a uma distância anormal de outros valores. Não são necessariamente erros!



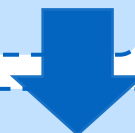
Porque nos devemos preocupar?

Distorcem a distribuição dos dados. Modelos estatísticos são sensíveis às distorções.



Como detetar?

Summary statistics e visualizações (histogramas, boxplots, etc.)

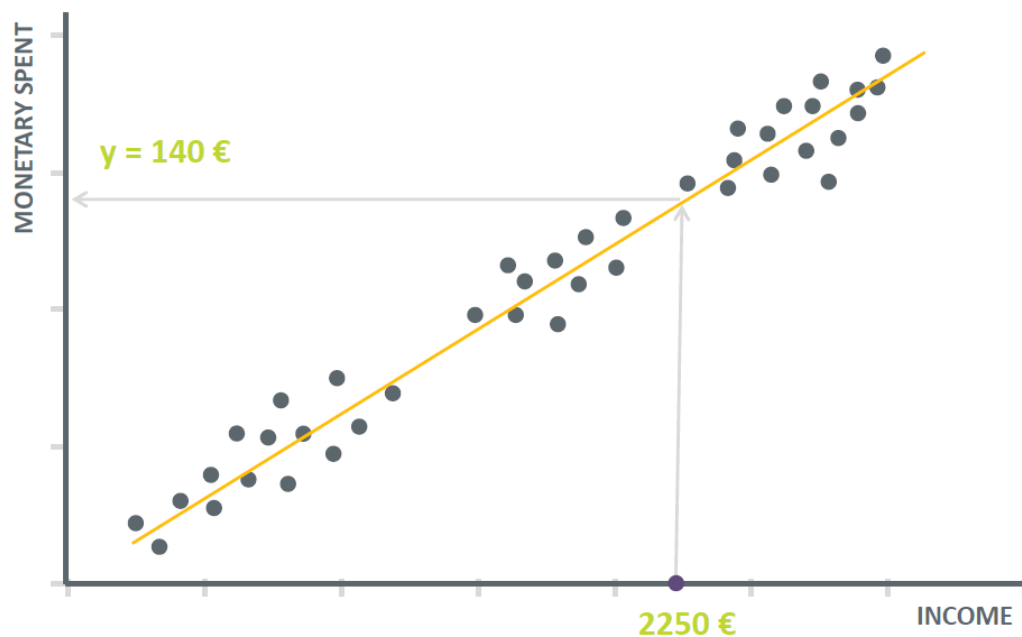


O que podemos fazer?

Remover, clip, etc.

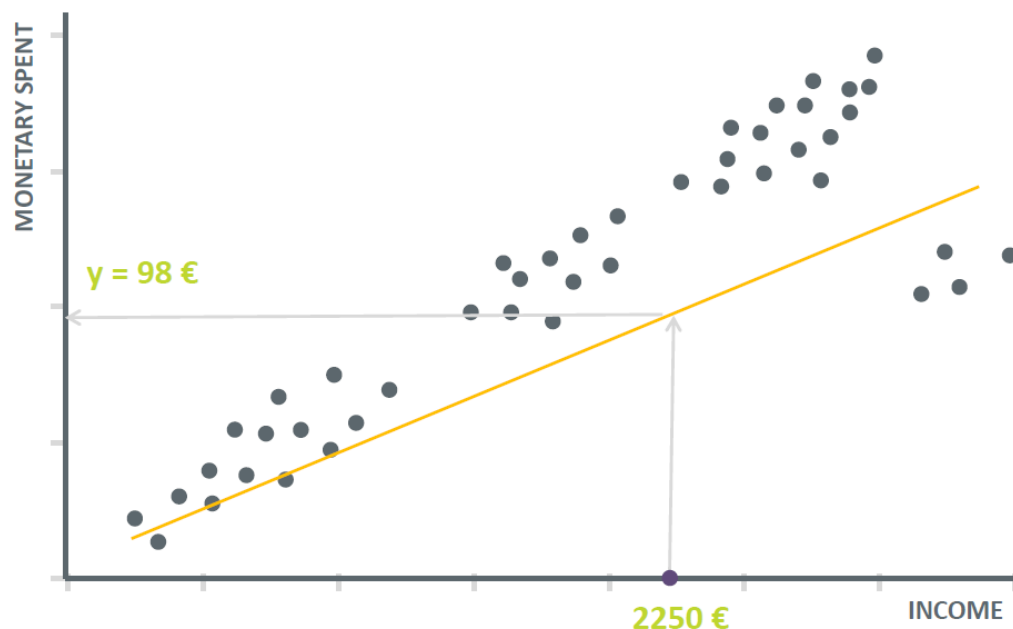
1.1 Data Cleaning – Outliers

Pergunta #1: Quanto é que um cliente com um rendimento de 2250 euros irá gastar na minha loja?



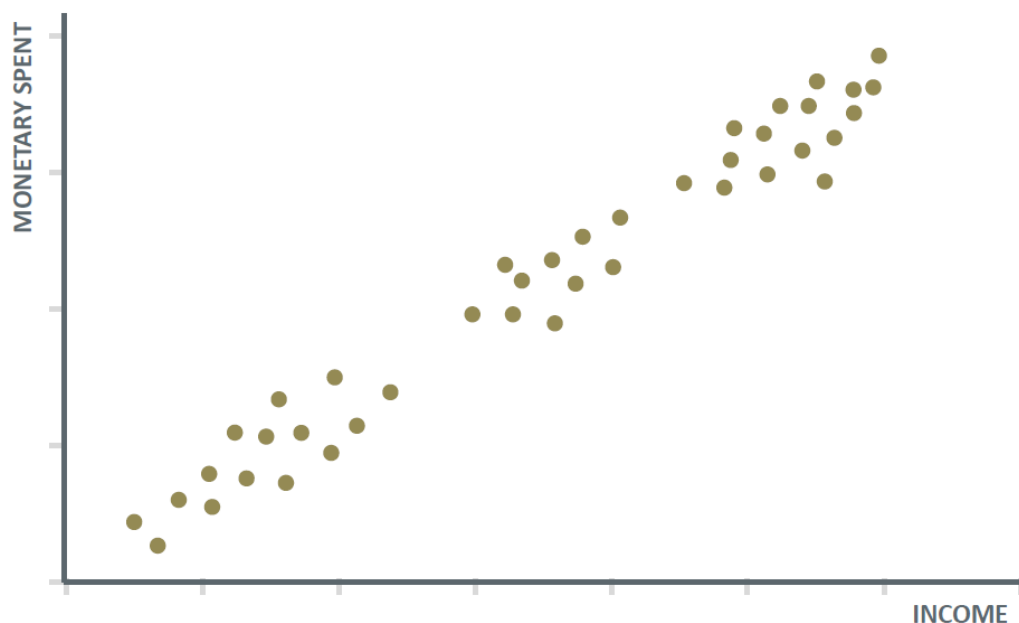
1.1 Data Cleaning – Outliers

Pergunta #1: Quanto é que um cliente com um rendimento de 2250 euros irá gastar na minha loja?



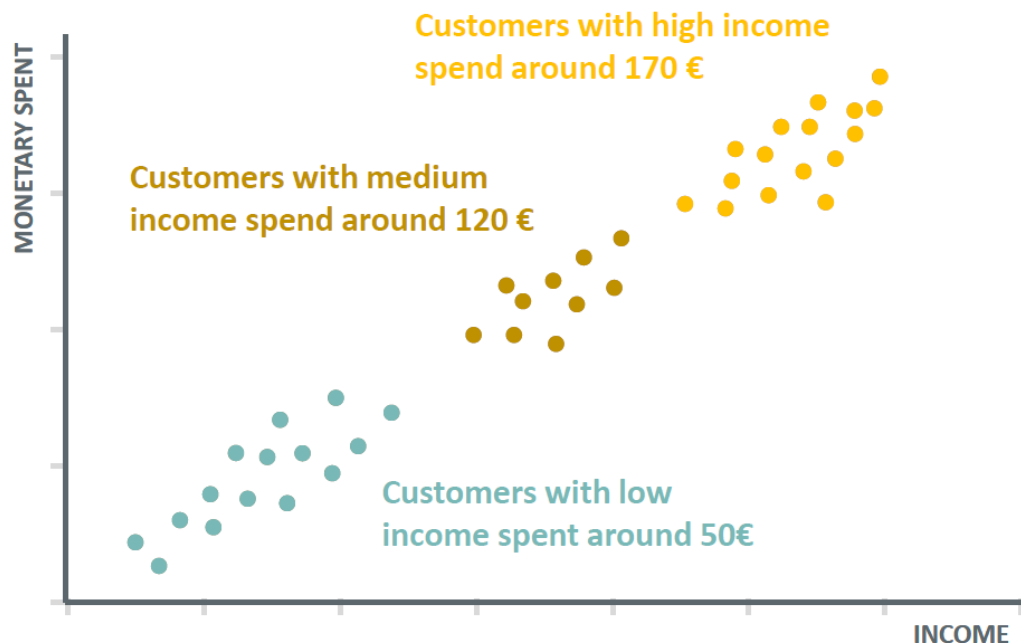
1.1 Data Cleaning – Outliers

Pergunta #2.1: Quantos grupos distintos de clientes tenho? Qual é o perfil de cada um deles?



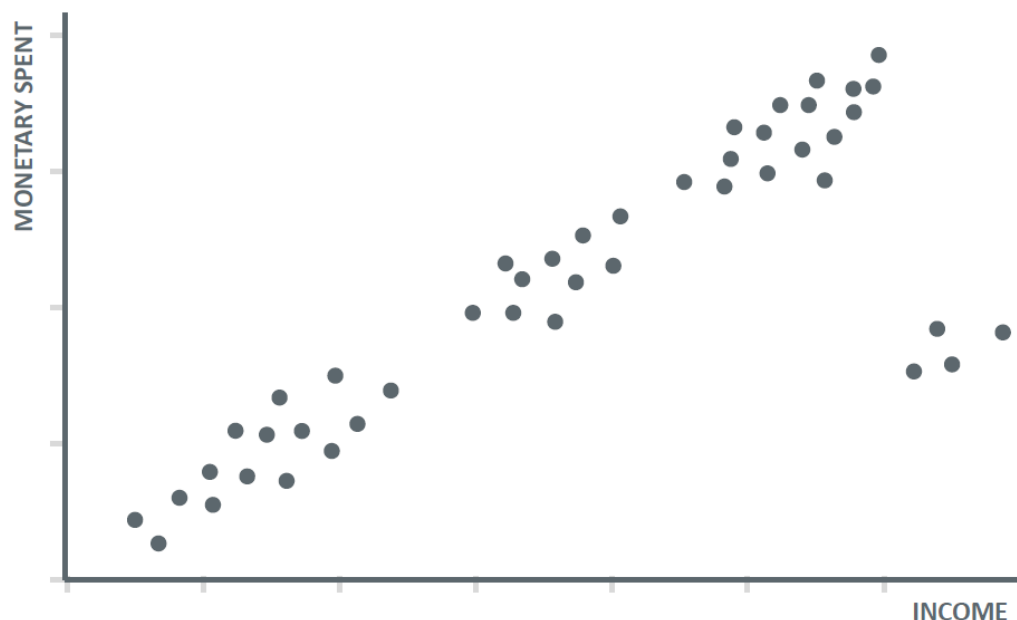
1.1 Data Cleaning – Outliers

Pergunta #2.1: Quantos grupos distintos de clientes tenho? Qual é o perfil de cada um deles?



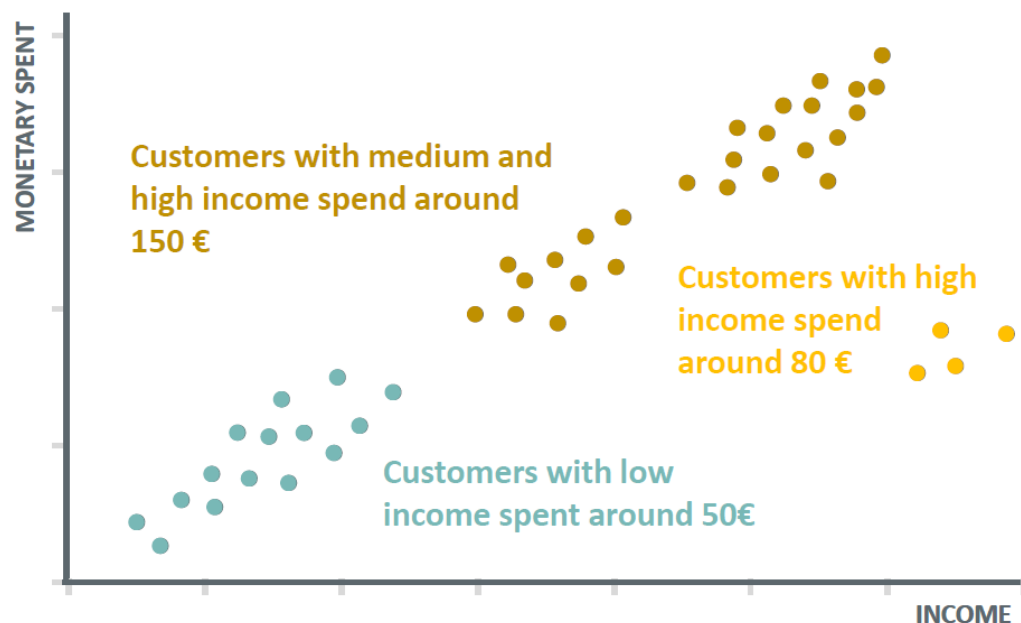
1.1 Data Cleaning – Outliers

Pergunta #2.2: Quantos grupos distintos de clientes tenho? Qual é o perfil de cada um deles?



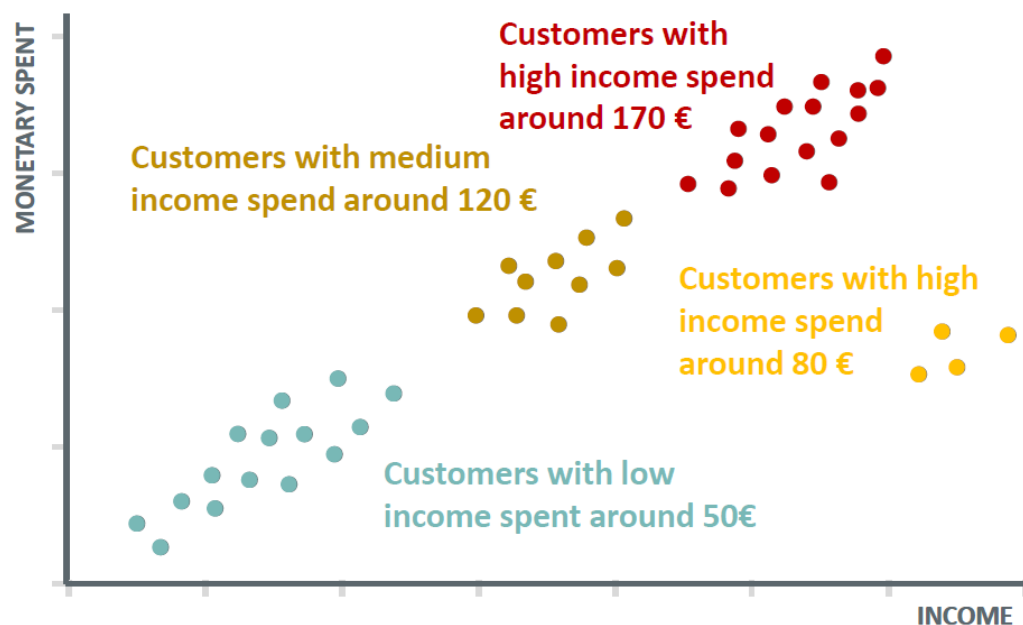
1.1 Data Cleaning – Outliers

Pergunta #2.2: Quantos grupos distintos de clientes tenho? Qual é o perfil de cada um deles?



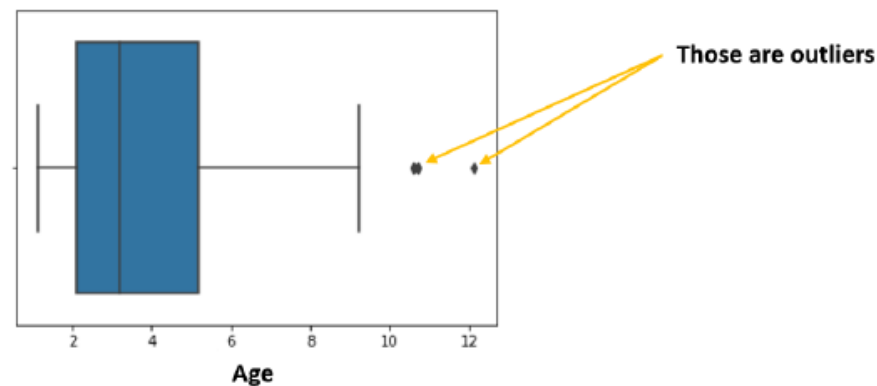
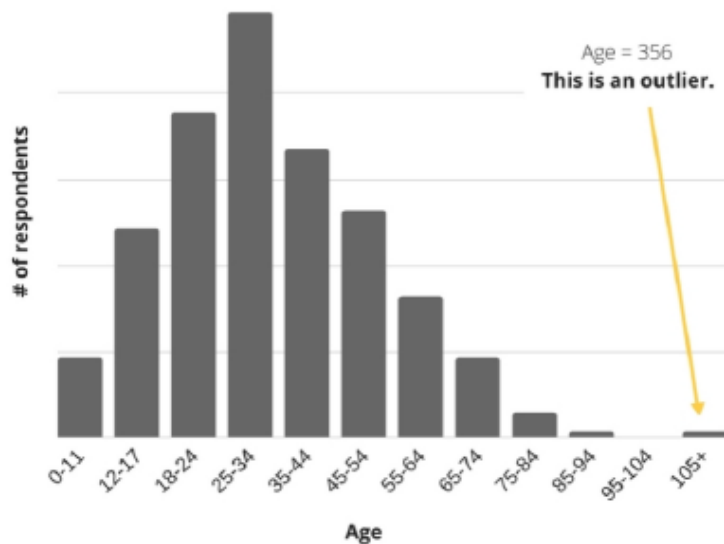
1.1 Data Cleaning – Outliers

Pergunta #2.2: Quantos grupos distintos de clientes tenho? Qual é o perfil de cada um deles?



1.1 Data Cleaning – Outliers

Como detetar outliers?



1.1 Data Cleaning – Outliers

Como tratar outliers detetados?

Remover

- Apenas os mais extremos.
- Regra geral: Não apagar mais do que 3% das observações. Se for mais, tentar as outras abordagens para os outliers.

Clipping

- Envolve limitar os valores dos outliers a um limite superior ou inferior definido. Em vez de remover, são substituídos por um valor limite.
- Podemos escolher estes valores de limite inferior e superior utilizando os percentis da variável.

Atribuir um novo valor

- Se um outlier parecer dever-se a um erro nos dados, tente imputar um valor (média /mediana / modelo preditivo ...)

Transformara a variável

- Por exemplo, criar uma versão categórica agrupada por percentis da variável original.

1.2 Data Transformation

1.2 Data Transformation - Scaling

As **variáveis** tendem a ter **intervalos diferentes umas das outras**.

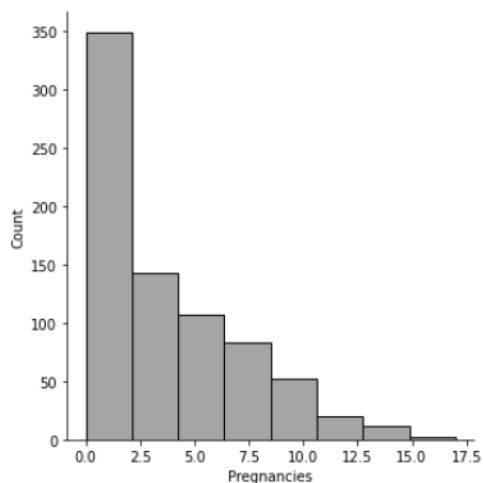
Alguns algoritmos (principalmente os que medem distâncias) são afectados negativamente por diferenças nos intervalos das variáveis:

- Ex: Se uma característica varia de 0 a 1 (por exemplo, “probabilidade”) e outra varia de 1.000 a 100.000 (por exemplo, “preço”), o modelo dará muito mais peso à característica maior, a menos que seja transformada para a mesma escala.

Assim, regra geral, **os valores dos campos numéricos devem ser normalizados**. Os métodos de normalização mais comuns são:

- Normalização min-max
- Normalização Z-score

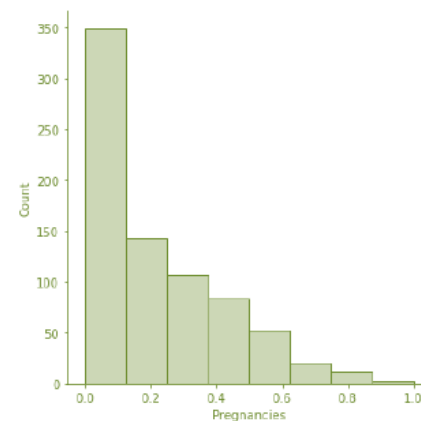
1.2 Data Transformation - Scaling



Normalização Min-Max

A variável passa a estar compreendida entre 0 e 1

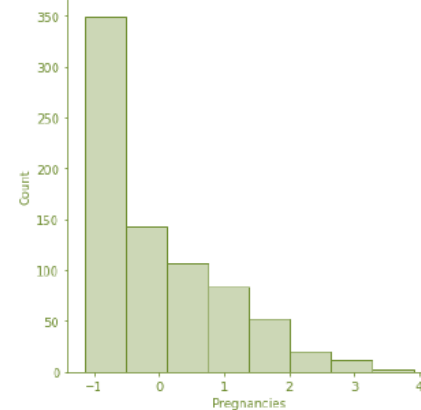
$$v' = \frac{v - \min_x}{\max_x - \min_x}$$



Normalização Z-score

A variável passa a ter média 0 e desvio padrão 1

$$v' = \frac{v - \mu_x}{\sigma_x}$$

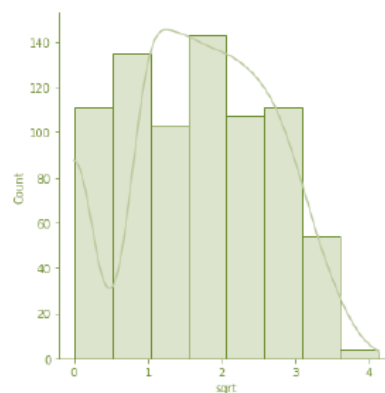
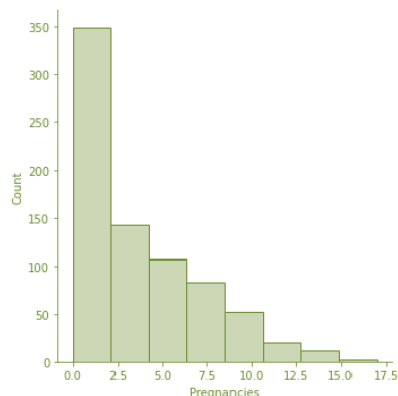


1.2 Data Transformation – Transformações power

Muitas vezes, as **variáveis tendem a ter distribuições enviadas** (skewed), e/ou apresentam um elevado número de outliers.

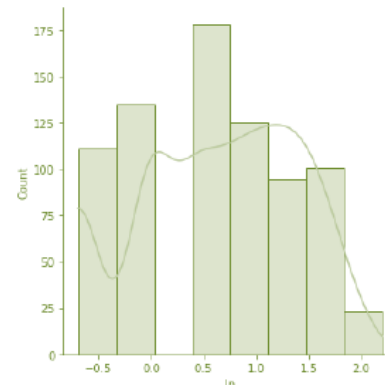
Assim, nestes casos, é normal proceder a uma “transformação power”, de forma a:

- **Reduzir o envizamento** e tornar a **distribuição mais simétrica** (mais próxima da normal);
- **Comprimir os outliers** para os aproximar dos valores centrais.



Raíz quadrada

$$x' = \sqrt{x}$$



Logaritmo


$$x' = \log(x)$$

1.2 Data Transformation – Variáveis dummy

A maior parte dos modelos de machine learning requerem variáveis numéricas.

Assim, é **necessário recodificar os valores categóricos numa ou mais variáveis dummy**, dependendo do número de categorias.

As variáveis dummy são uma **variáveis binárias para cada categoria da variável categórica original, assumindo apenas dois valores: 0 ou 1**.



Género
M
F
M
M
F


Género_M	Género_F
1	0
0	1
1	0
1	0
0	1

1.2 Data Transformation – Variáveis dummy

A maior parte dos modelos de machine learning requerem variáveis numéricas.

Assim, é **necessário recodificar os valores categóricos numa ou mais variáveis dummy**, dependendo do número de categorias.

As variáveis dummy são uma **variáveis binárias para cada categoria da variável categórica original, assumindo apenas dois valores: 0 ou 1**.




Género		Género_M	Género_F
M		1	0
F		0	1
M		1	1
M		1	0
F		0	1

De forma a evitar redundância, **são apenas geradas k-1 variáveis dummy** (onde k é o número de categorias da variável categórica)

1.2 Data Transformation – Ordinal encoding

A maior parte dos modelos de machine learning requerem variáveis numéricas.

Além de variáveis dummy, **podemos codificar uma variável categórica através de ordinal encoding**, caso as categorias da variável apresentem uma “ordem”.




Satisfação	
Muito Satisfeito	
Insatisfeito	
Satisfeito	
Insatisfeito	
Pouco Satisfeito	

Satisfação	Satisfação_enc
Muito Satisfeito	4
Insatisfeito	2
Satisfeito	3
Insatisfeito	2
Pouco Satisfeito	1

1.2 Data Transformation – Ordinal encoding

A maior parte dos modelos de machine learning requerem variáveis numéricas.

Além de variáveis dummy, **podemos codificar uma variável categórica através de ordinal encoding**, caso as categorias da variável apresentem uma “ordem”.



Satisfação		Satisfação	Satisfação_enc
Muito Satisfeito		Muito Satisfeito	4
Insatisfeito		Insatisfeito	2
Satisfeito		Satisfeito	3
Insatisfeito		Insatisfeito	2
Pouco Satisfeito		Pouco Satisfeito	1

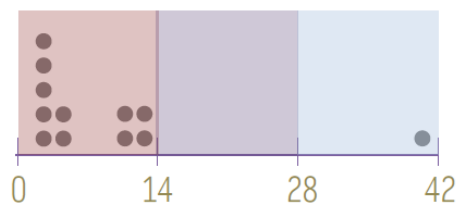
De forma a evitar redundância, no modelo, não incluímos a variável original.

1.2 Data Transformation – Binning

Apesar da maior parte das vezes convertermos variáveis categóricas para numéricas, o inverso também pode ser necessário.

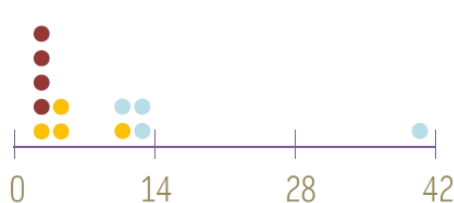
Alguns algoritmos preferem variáveis categóricas em vez de contínuas (nomeadamente árvores de decisão).

Assim, podemos criar partições (categorias) de variáveis numéricas, transformando-as em variáveis categóricas:



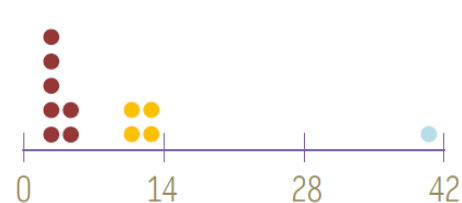
Partição por intervalo

Divide as observações em k categorias de intervalo iguais.



Partição por frequência

Divide as observações em k categorias, cada uma com k/n observações.



Clustering K-Means

Algoritmo de clustering, calculando as partições ideais.

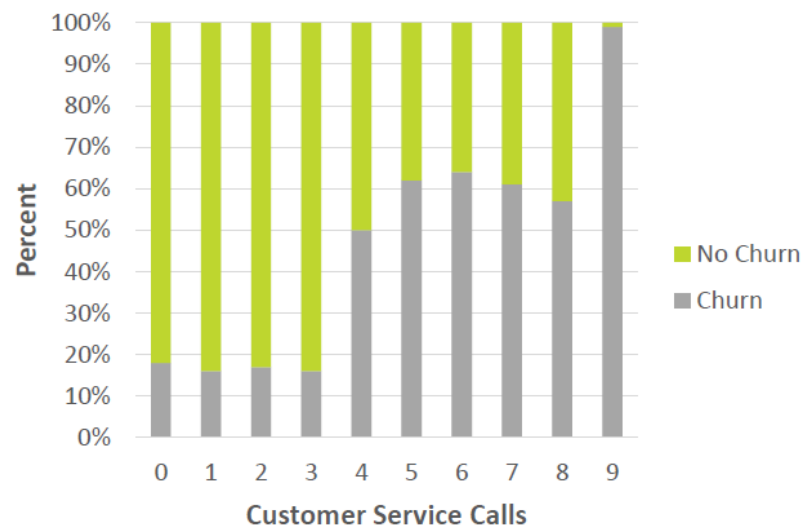
1.2 Data Transformation – Binning

Binning com base na variável a prever (para supervised learning)

- Divide a variável numérica com base no efeito que cada partição tem no valor da variável alvo.

Os clientes com < 4 chamadas para o serviço de apoio ao cliente tiveram uma % churn menor do que os clientes que fizeram ≥ 4 chamadas.

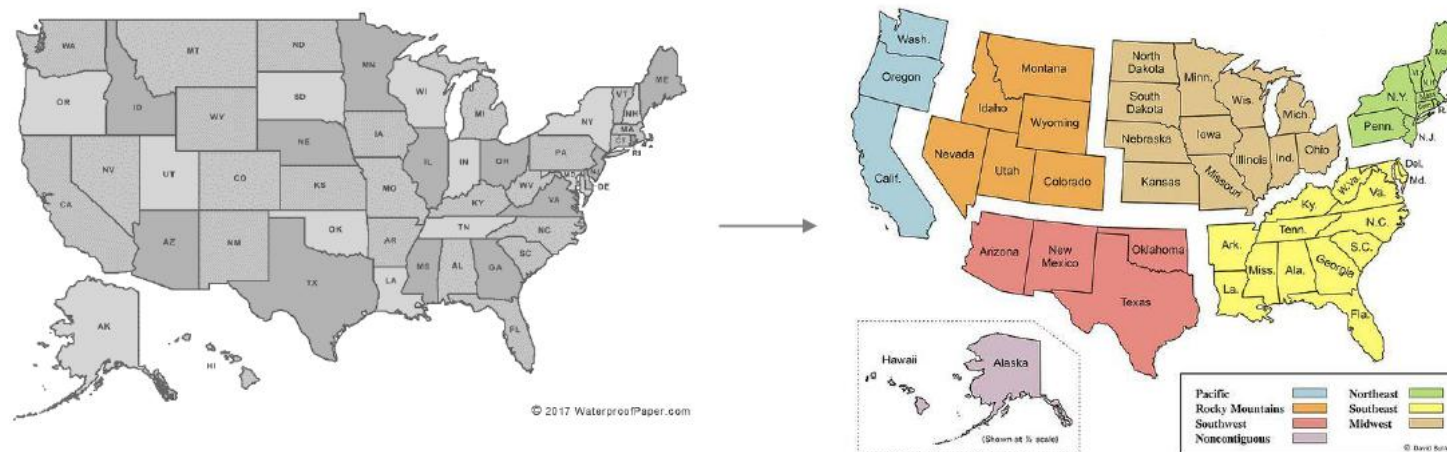
- Dividir a variável em duas classes:** Baixa (menos de quatro chamadas) e Alta (quatro ou mais).



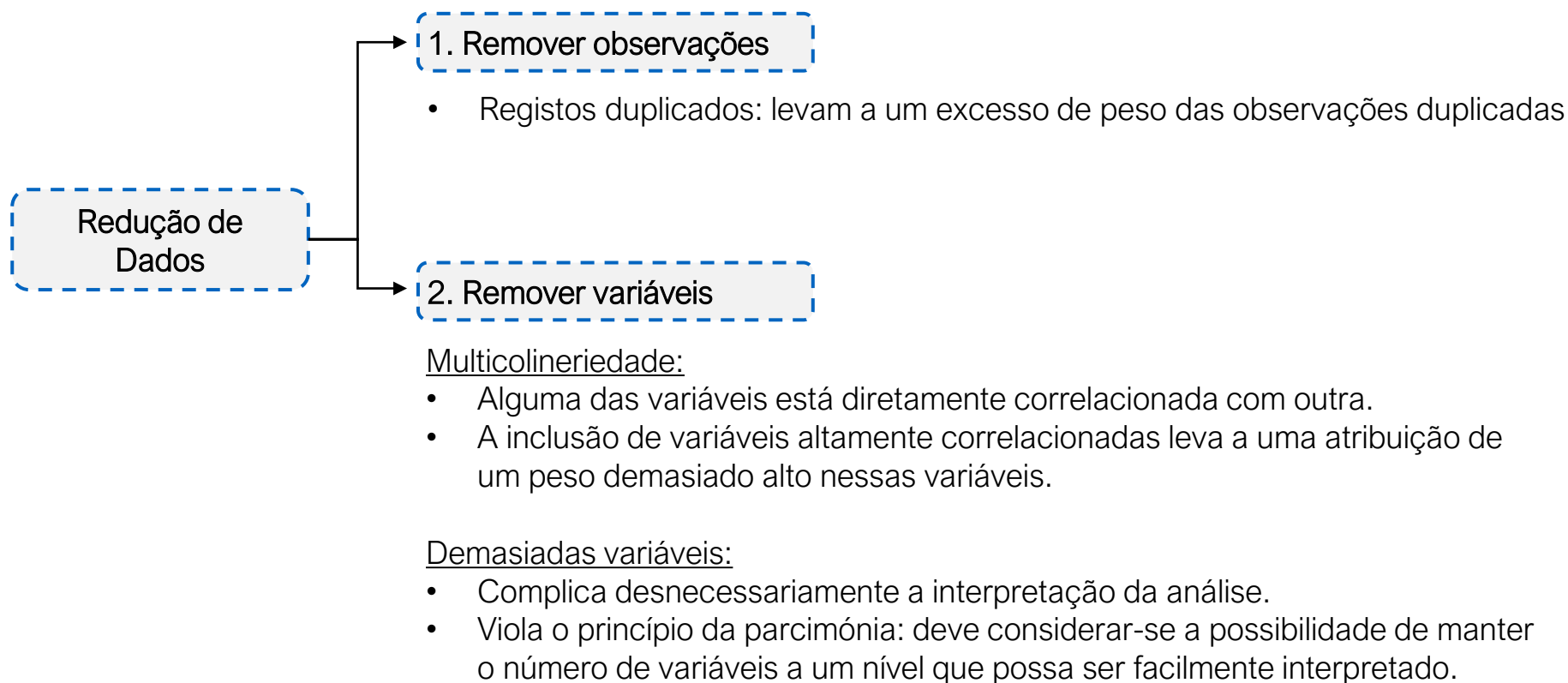
1.2 Data Transformation – Reclassificar

É o equivalente do binning de variáveis numéricas, mas para variáveis categóricas.

É normalmente aplicado a variáveis categóricas **cardinalidade elevada** (com muitas categorias), i) reduzindo o ruído, ii) ajudando o modelo a focar-se nas distinções mais significativas e iii) aumentando a interpretabilidade dos resultados.



1.2 Data Transformation – Reduzir dados



1.2 Data Transformation – Reduzir dados

ID	Idade	Ano de Nascimento	Género	Doação €
1	23	2001	F	
2	28	1996	F	
3	37	1987	F	
4	23	2001	F	
4	23	2001	F	
5	32	1992	F	
6	65	1959	F	75
7	34	1990	F	
8	46	1980	F	
9	31	1993	F	

1.2 Data Transformation – Reduzir dados

Problema de multicolinearidade: Remover 1 das variáveis

ID	Idade	Ano de Nascimento	Gênero	Doação €
1	23	2001	F	
2	28	1996	F	
3	37	1987	F	
4	23	2001	F	
4	23	2001	F	
5	32	1992	F	
6	65	1959	F	75
7	34	1990	F	
8	46	1980	F	
9	31	1993	F	

Observações duplicadas:
Remover duplicados

Variáveis para as quais 90% ou mais dos valores estão em falta: Remover

Mas...

Pode apresentar um padrão na ausência de valores:

- quem faz muitos donativos comunica os seus donativos
- quem não doa ou doa um valor baixo pode não responder a esta pergunta
- criar uma variável sinalizadora (o padrão na falta de resposta pode vir a ter poder preditivo)

Variáveis unárias ou quase unárias. Remover?
Unário: sim! Quase unária: veja variável alvo.

Obrigado!