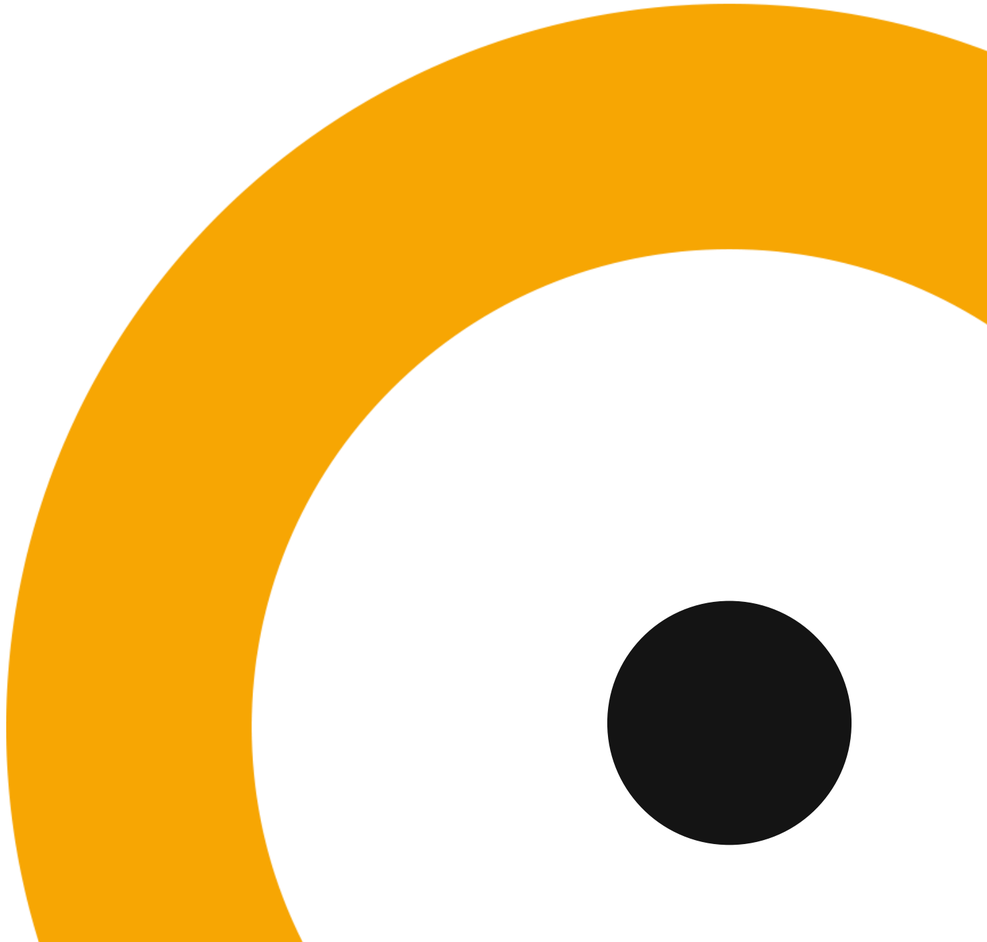
Mestrado Data Science

Data Transformation

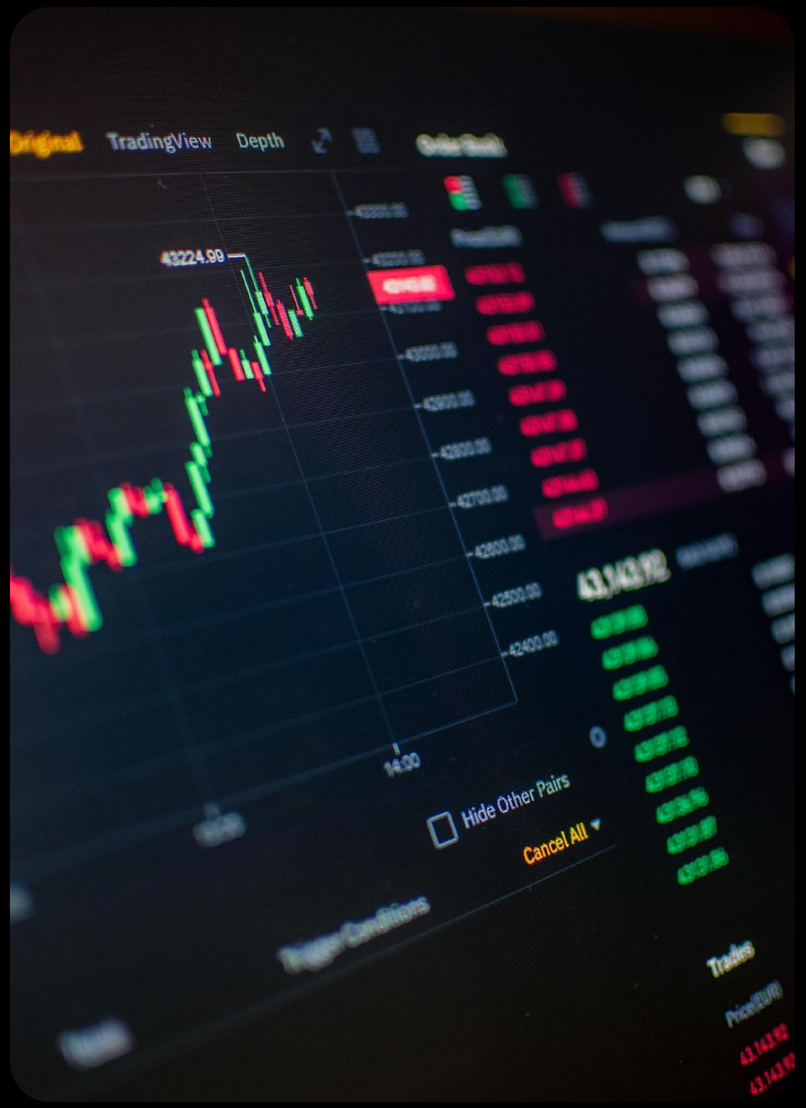


Módulo: Data Cleaning and Transformation in Python

01

Motivação

Porque é importante transformarmos os dados?



Novas variáveis construidas por nós

A criatividade humana torna-nos insubstituíveis: Rácios, Subtrações, etc

Por vezes há algoritmos muito lentos com variáveis numéricas

Temos que as discretizar.

Discretizações podem trazer mais estabilidade nos modelos

Devemos conseguir construir modelos com boa performance ao longo do tempo.

Alguns algoritmos não conseguem lidar com variáveis categóricas

Temos que codificar em formato numérico.

Certos algoritmos dão importancia á escala da variável

O salário de uma pessoa e o número de maçãs consumidas por dia têm escalas completamente diferentes.

Conteúdo

Sessão nº 4

02

Criação de novas variáveis

Que variáveis seriam interessantes para prever se um cliente merece um empréstimo?

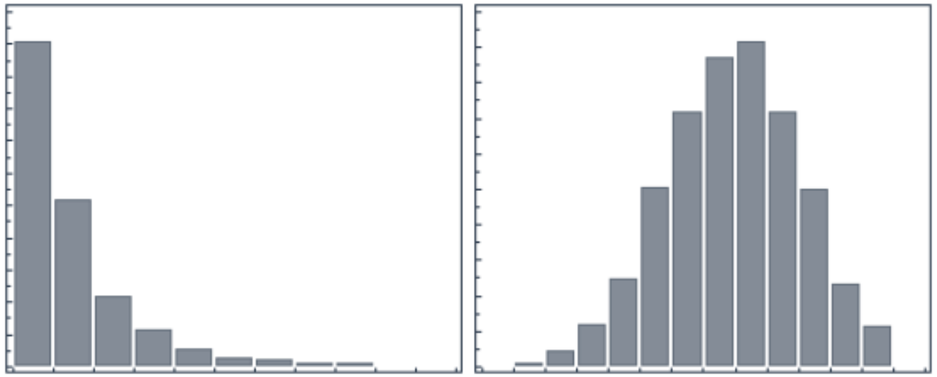
E neste exemplo, que variáveis podemos criar?

Variável	Exemplo	
Tipo de crédito	Automóvel	
Localidade	Rio de Janeiro	
Salário mensal		10.000
Outros ganhos mensais		1.000
Valor do empréstimo que pretende contrair		100.000
Mensalidade do empréstimo que pretende contrair		1.000
Valor de outros empréstimos		30.000
Total mensalidades de outros empréstimos		500
Estado civil	Solteiro	
Nº de créditos		3
Nº filhos		2

Transformação logarítmica

Quando devo fazer?

$$X' = a*\log_b(X)$$



A escala é muito grande

Pode ajudar a melhorar a visualização destas variáveis

O modelo que queremos usar precisa de uma determinada distribuição

Podemos conseguir encontrar mais simetría na distribuição.

03

Discretização de variáveis numéricas

Que métodos conhecem?

Métodos mais conhecidas

Não supervisionados

- Bins com mesmo tamanho;
- Bins com mesmo volume

Supervisionados

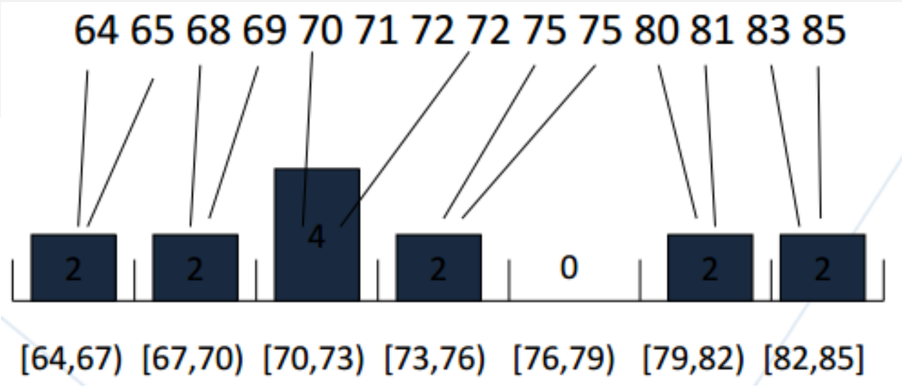
- Chi-square

Bins com o mesmo tamanho

Como fazer?

Dividir o intervalo por N

No nosso exemplo: (85-64)/7=3



Vantagens:

1. Simples e fácil de implementar;
2. Produz uma boa abstração dos dados;

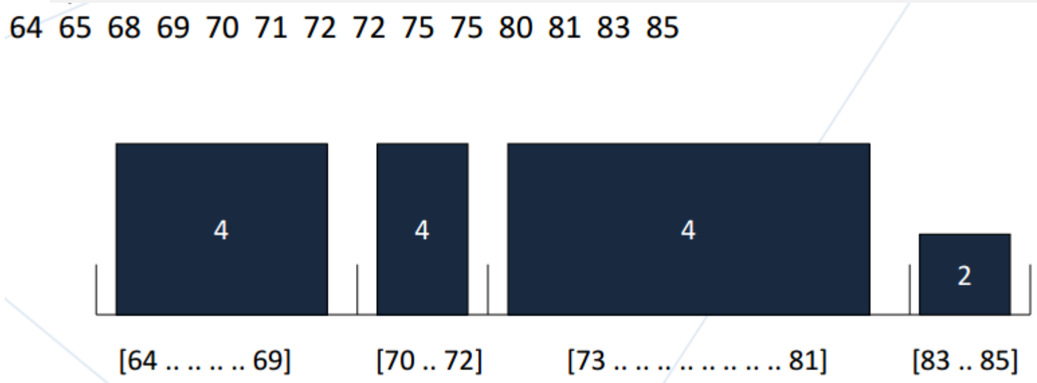
Desvantagens:

1. Sensível a outliers!
2. Não é supervisionado (não ótimo para ML);
3. Como escolher bem o N?

Bins com o mesmo volume

Como fazer?

Divide o intervalo em N intervalos com o mesmo volume.



Vantagens:

1. Simples e fácil de implementar;
2. Evita aglomeração de dados;

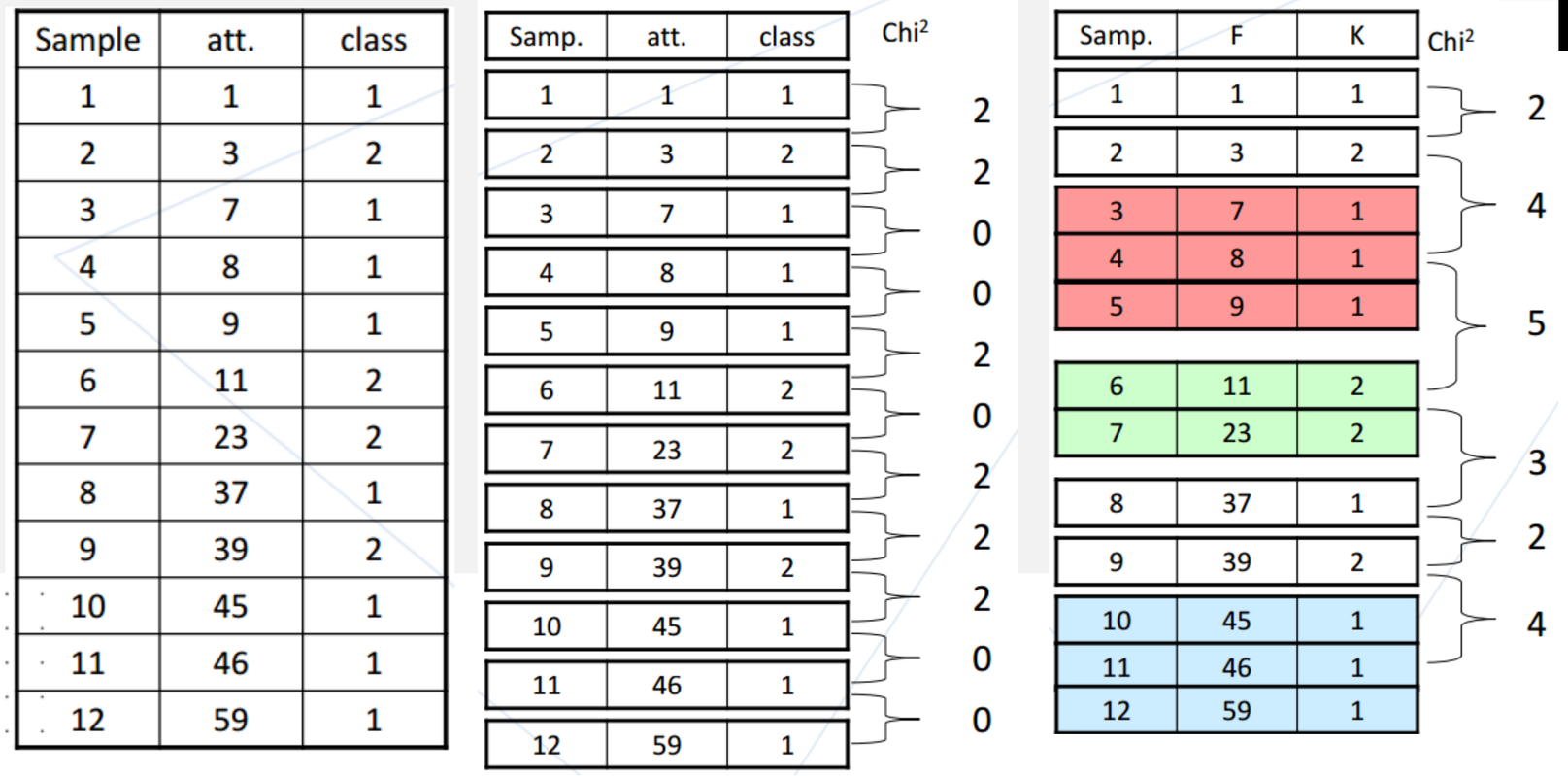
Desvantagens:

1. Não é supervisionado (não ótimo para ML);
2. Como escolher bem o N?

Chi-square

Como fazer?

1. Ordena a variável;
2. Calcula o teste chi-square em cada ponto de divisão;
3. Junta intervalos com o valor mais pequeno;
4. Repete até não ter valores menores que X;



Sample	att.	class
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

Samp.	att.	class	Chi^2
1	1	1	2
2	3	2	2
3	7	1	0
4	8	1	0
5	9	1	2
6	11	2	0
7	23	2	2
8	37	1	2
9	39	2	2
10	45	1	0
11	46	1	0
12	59	1	

Samp.	F	K	Chi^2
1	1	1	2
2	3	2	
3	7	1	4
4	8	1	
5	9	1	5
6	11	2	
7	23	2	3
8	37	1	
9	39	2	2
10	45	1	4
11	46	1	
12	59	1	

04

Encodings de variáveis categóricas

Como transformariam uma variável categórica em numérica?

Métodos mais conhecidas

Não supervisionados

- One Hot Encoding;
- Ordinal Encoding

Supervisionados

- Target Encoding

Outros Encoders: https://contrib.scikit-learn.org/category_encoders/targetencoder.html

One Hot Encoding

Como funciona?

1. Cria uma coluna para cada categoria;
2. A linha terá um 1 se representava essa categoria, 0 caso contrário.

Qualidade do produto	Target
Boa	1
Má	0
Boa	0
Boa	1
Média	0

Qualidade má	Qualidade média	Qualidade boa	Target
0	0	1	1
1	0	0	0
0	0	1	0
0	0	1	1
0	1	0	0

Ordinal Encoding

Como funciona?

1. Mapeia cada categoria a um numero (de forma crescente);
2. Neste caso:

Má: 0; Média: 1; Boa: 2

Qualidade do produto	Target
Boa	1
Má	0
Boa	0
Boa	1
Média	0

Encoding	Target
2	1
0	0
2	0
2	1
1	0

***Nota**: deve ser utilizada em variáveis ordinais.*

Target Encoding

Como funciona?

1. Mapeia cada categoria com a média da target;
2. Neste caso:

Má: 0; Média: 1; Boa: 2

Qualidade do produto	Target
Boa	1
Má	0
Boa	0
Boa	1
Média	0

Encoding	Target
2/3	1
0	0
2/3	0
2/3	1
0	0

***Nota**: devemos garantir que não estamos a utilizar dados do futuro!*

05

Escalonamento dos dados

Quem já ouviu falar disto?

Para métodos baseados em distância, o escalonamento ajuda a evitar que variáveis com intervalos grandes superem variáveis com intervalos pequenos.

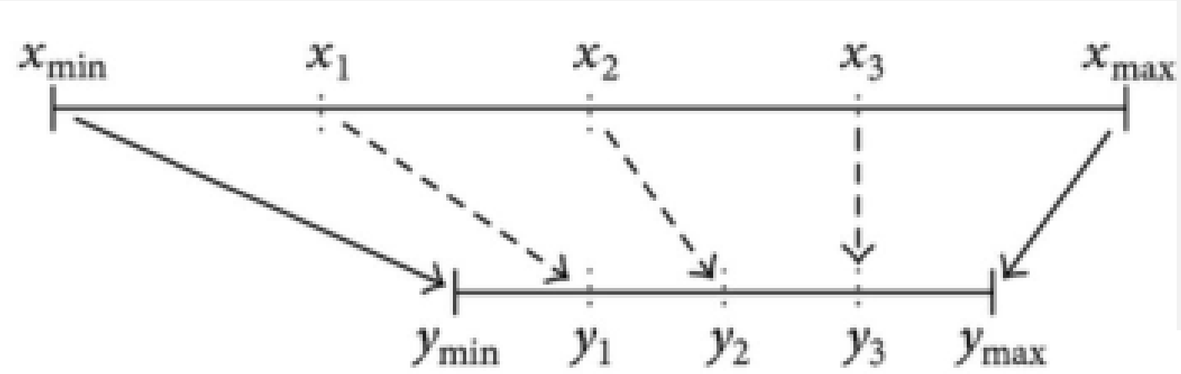
Min-max Scaler

Como funciona?

Podemos colocar todas as variáveis entre x e y.

Por exemplo para x=0 e y=1:

$$X'_i = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$



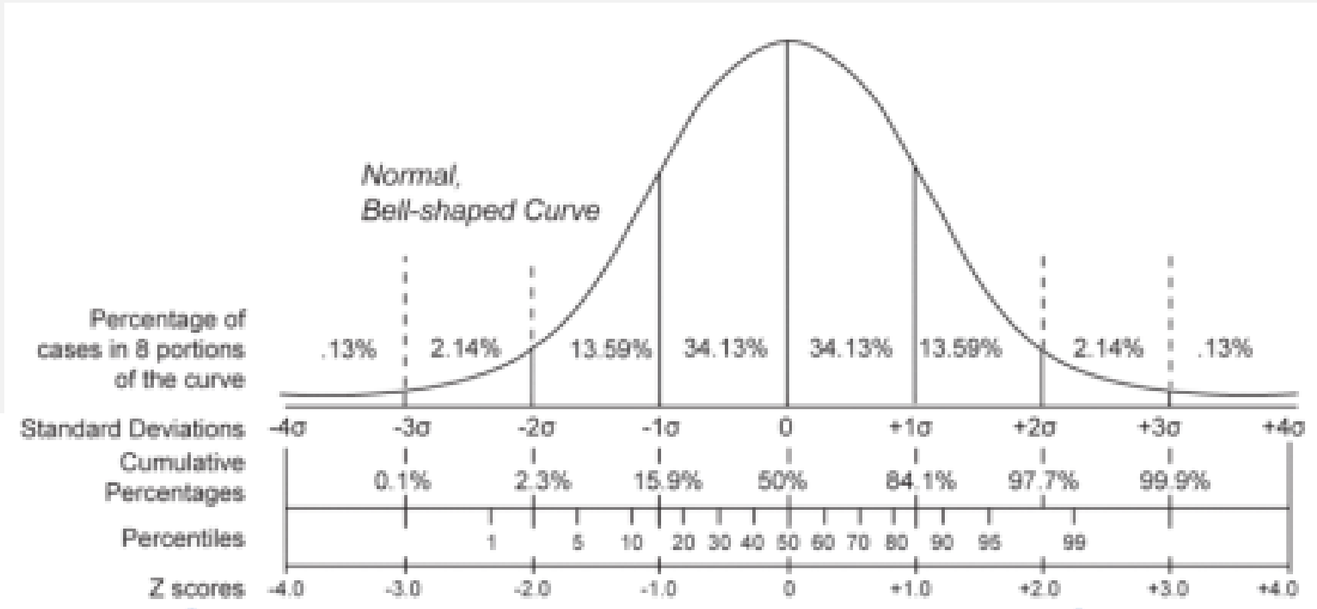
Standard Scaler

Como funciona?

Torna as variáveis com média 0 e desvio padrão 1.

Transformação:

$$X'_i = \frac{X_i - \bar{X}}{\sigma}$$



06

Possíveis perguntas numa entrevista

Muito obrigado pela sua atenção.

Fábio Ferreira
Endereço de e-mail:
fabio7jnferreira@gmail.com

NUCLIO DIGITAL SCHOOL