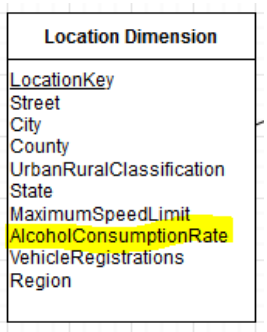
Grupo 17 /
João Ferreira / 54121 /
Mariana Santos / 49525 /
Tiago Pereira / 49174 /
Vasco Leitão / 49455 /

Dúvida: Em que moldes se deve proceder à discretização de valores contínuos em atributos das dimensões?



Exemplo:

- Atributo sobre consumo de álcool per capita em cada estado dos EUA

  Insere-se num intervalo de valores onde cada estado contém o seu valor de consumo (por *gallon* de álcool).

	State	rate
28	New Hampshire	4.74
7	Delaware	3.60
27	Nevada	3.43
33	North Dakota	3.15
25	Montana	3.11
44	Vermont	3.10
48	Wisconsin	2.99
11	Idaho	2.92
40	South Dakota	2.86
5	Colorado	2.85
18	Maine	2.84
1	Alaska	2.81
22	Minnesota	2.77

(...)

Possibilidades de discretização pensadas:

- Intuição: não nos regemos por qualquer característica estatística para definir valores de separação, apenas procurando criar categorias representativas.

```
alcohol_consumption.loc[alcohol_consumption['rate'].apply(lambda x : x <= 2.0), 'discretizeRate'] = 'Low'
alcohol_consumption.loc[alcohol_consumption['rate'].apply(lambda x : (x > 2.0) & (x <= 2.5)), 'discretizeRate'] = 'Medium'
alcohol_consumption.loc[alcohol_consumption['rate'].apply(lambda x : (x > 2.5) & (x <= 3.5)), 'discretizeRate'] = 'High'
alcohol_consumption.loc[alcohol_consumption['rate'].apply(lambda x : x > 3.5), 'discretizeRate'] = 'Very High'
```

- Discretização baseada em quantis: Função *qcut()* faz a criação de *buckets* com que são definidos pelo número de quantis definidos.

```
alcohol_consumption['qcut_rate'] = pd.qcut(alcohol_consumption['rate'], q=4)
print(alcohol_consumption['qcut_rate'].value_counts())
```

Acontece que embora a discretização seja feita ao nível destes dados, pode acontecer que depois a distribuição de registos em cada categoria fica muito desnivelada.

	State	rate	discretizeRate	qcut_rate
28	New Hampshire	4.74	Very High	(2.768, 4.74]
7	Delaware	3.60	Very High	(2.768, 4.74]
27	Nevada	3.43	High	(2.768, 4.74]
33	North Dakota	3.15	High	(2.768, 4.74]
25	Montana	3.11	High	(2.768, 4.74]
44	Vermont	3.10	High	(2.768, 4.74]
48	Wisconsin	2.99	High	(2.768, 4.74]
11	Idaho	2.92	High	(2.768, 4.74]
40	South Dakota	2.86	High	(2.768, 4.74]
5	Colorado	2.85	High	(2.768, 4.74]
18	Maine	2.84	High	(2.768, 4.74]
1	Alaska	2.81	High	(2.768, 4.74]
22	Minnesota	2.77	High	(2.768, 4.74]
36	Oregon	2.76	High	(2.34, 2.768]
49	Wyoming	2.68	High	(2.34, 2.768]
10	Hawaii	2.65	High	(2.34, 2.768]
8	Florida	2.63	High	(2.34, 2.768]

(...)

Também temos esta questão a nível de dados meteorológicos. Se encontramos alguma referência na qual sejam definidas categorias para dados como a temperatura, velocidade do vento, etc., devemos seguir esses dados?

Obrigado.