Cruzeiro do Sul Virtual
Educação a Distância

DIAGRAMAS DE DISPERSÃO

Prof. Ismar Frango

Relações entre duas variáveis

É bastante comum, em Ciência de Dados, que os *datasets* tragam mais de uma variável com dados sobre os indivíduos. Assim, é possível, em alguns casos, estabelecermos relações entre variáveis.

Vamos estudar aqui as relações entre duas variáveis quantitativas. Nos interessam aquelas relações em que as duas variáveis têm papéis diferentes, em que uma influencia a outra. Quando isto ocorre, dizemos que uma variável (chamada **variável explicativa**) serve como "explicação" para mudanças observadas em outra variável (chamada **variável resposta**).

Uma **variável resposta (dependente)** traz valores cuja mudança é influenciada por outra variável.

Uma **variável explicativa (independente)** influencia as mudanças observadas em uma variável resposta.

A forma mais comum de se mostrar a relação entre duas variáveis quantitativas é por meio de um **diagrama de dispersão**, que mostra essa relação entre variáveis, relativas aos mesmos indivíduos.

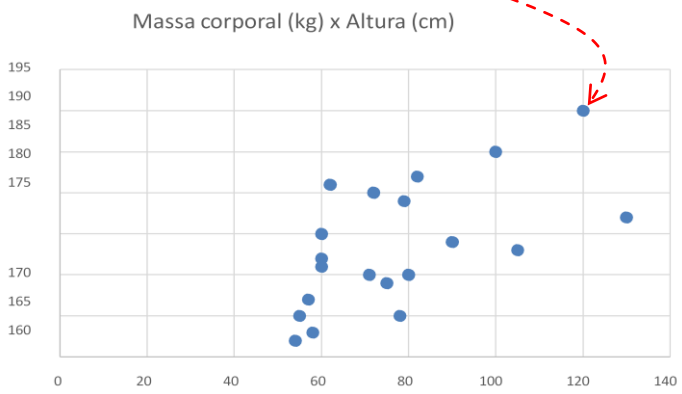
Em inglês, o termo utilizado para esses gráficos é *scatter graph* ou *scatter plot*.

Os valores da variável explicativa em geral aparecem no eixo das abscissas (horizontal) e os valores da variável resposta aparecem no eixo das ordenadas (vertical).

Quando estudamos a relação entre duas variáveis, em especial no que diz respeito à "direção" (positiva ou negativa) desta relação, usaremos a palavra **associação**. Atenção, pois os termos **associação** e **relação** são frequentemente tratadas como sinônimos na área de Análise de Dados.

Observe o exemplo com o seguinte *dataset*, com dados contendo a massa corporal (em kg) e as idades de um conjunto de indivíduos:

Massa corporal (kg)	72	80	60	90	100	120	82	79	78	55	71	75	130	105	60	54	58	57	60	62
Altura (cm)	180	170	175	174	185	190	182	179	165	165	170	169	177	173	172	162	163	167	171	181



Cada indivíduo nos dados aparece como um ponto no gráfico

Neste gráfico, a **massa corporal** está sendo considerada a variável **explicativa**, enquanto a **altura**, a variável **resposta**.

Isso faz sentido para você?

Não estamos aqui buscando relações de causa e efeito!

Não queremos dizer, com este gráfico, que a causa para que as pessoas sejam altas seria uma massa corporal maior. Em nem o oposto disso (de que pessoas mais altas teriam mais massa corporal).

O que os diagramas de dispersão mostram, a princípio, é a existência ou não de uma associação entre duas variáveis, apenas isso.

Associações positivas e negativas

Duas variáveis podem ser associadas de maneira positiva ou negativa. Isso depende unicamente do comportamento de ambas: se os valores de uma variável crescem à medida que os valores da outra também crescem (ainda que em taxas diferente), diz-se que essas variáveis têm uma associação **positiva**.

Quando, ao contrário, os valores de uma das variáveis decresce à medida que os valores da outra crescem, dizemos que elas mantêm uma associação **negativa**.

Cruzeiro do Sul Virtual
Educação a Distância

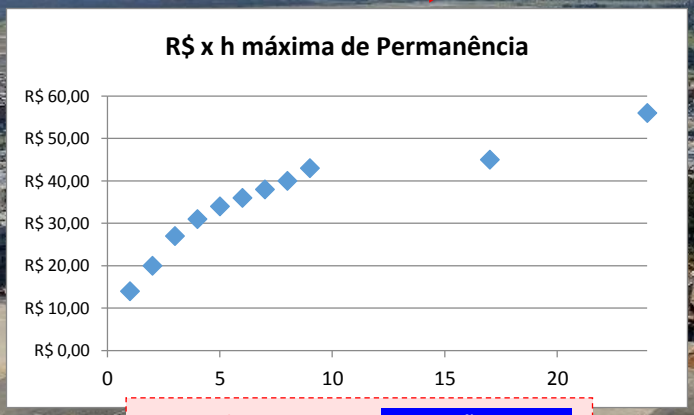
Veja o seguinte exemplo:

A tabela a seguir traz os valores do estacionamento do Aeroporto de Viracopos, em Campinas-SP

(Fonte: http://www.aeroportoviracopos.net)

Até 1 hora	R$ 14,00
Até 2 horas	R$ 20,00
Até 3 horas	R$ 27,00
Até 4 horas	R$ 31,00
Até 5 horas	R$ 34,00
Até 6 horas	R$ 36,00
Até 7 horas	R$ 38,00
Até 8 horas	R$ 40,00
Até 9 horas	R$ 43,00
de 10h até 17h	R$ 45,00
de 18h até 24 h	R$ 56,00

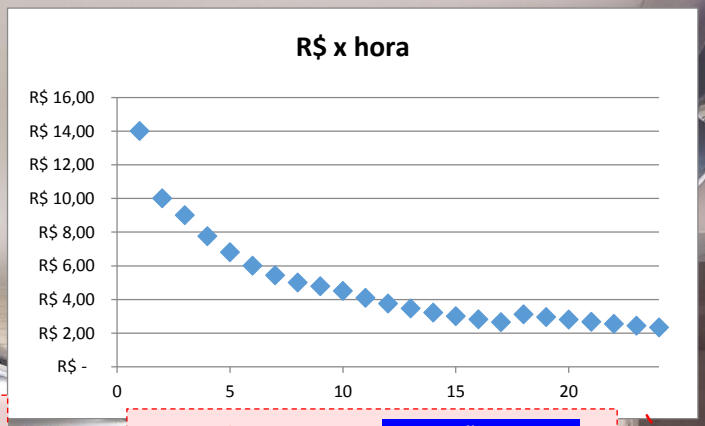
Neste gráfico, temos como variável explicativa a quantidade máxima de horas de permanência de um veículo no estacionamento, e como variável resposta, o custo final do estacionamento.



Temos claramente uma **associação positiva**

Permanência	R$ por h
1	R$ 14,00
2	R$ 10,00
3	R$ 9,00
4	R$ 7,75
5	R$ 6,80
6	R$ 6,00
7	R$ 5,43
8	R$ 5,00
9	R$ 4,78
10	R$ 4,50
11	R$ 4,09
12	R$ 3,75
13	R$ 3,46
14	R$ 3,21
15	R$ 3,00
16	R$ 2,81
17	R$ 2,65
18	R$ 3,11
19	R$ 2,95
20	R$ 2,80
21	R$ 2,67
22	R$ 2,55
23	R$ 2,43
24	R$ 2,33

Nesta outra tabela, porém, temos o custo por hora, de acordo com a quantidade máxima de horas (calculado a partir da tabela acima)



Temos claramente uma **associação negativa**

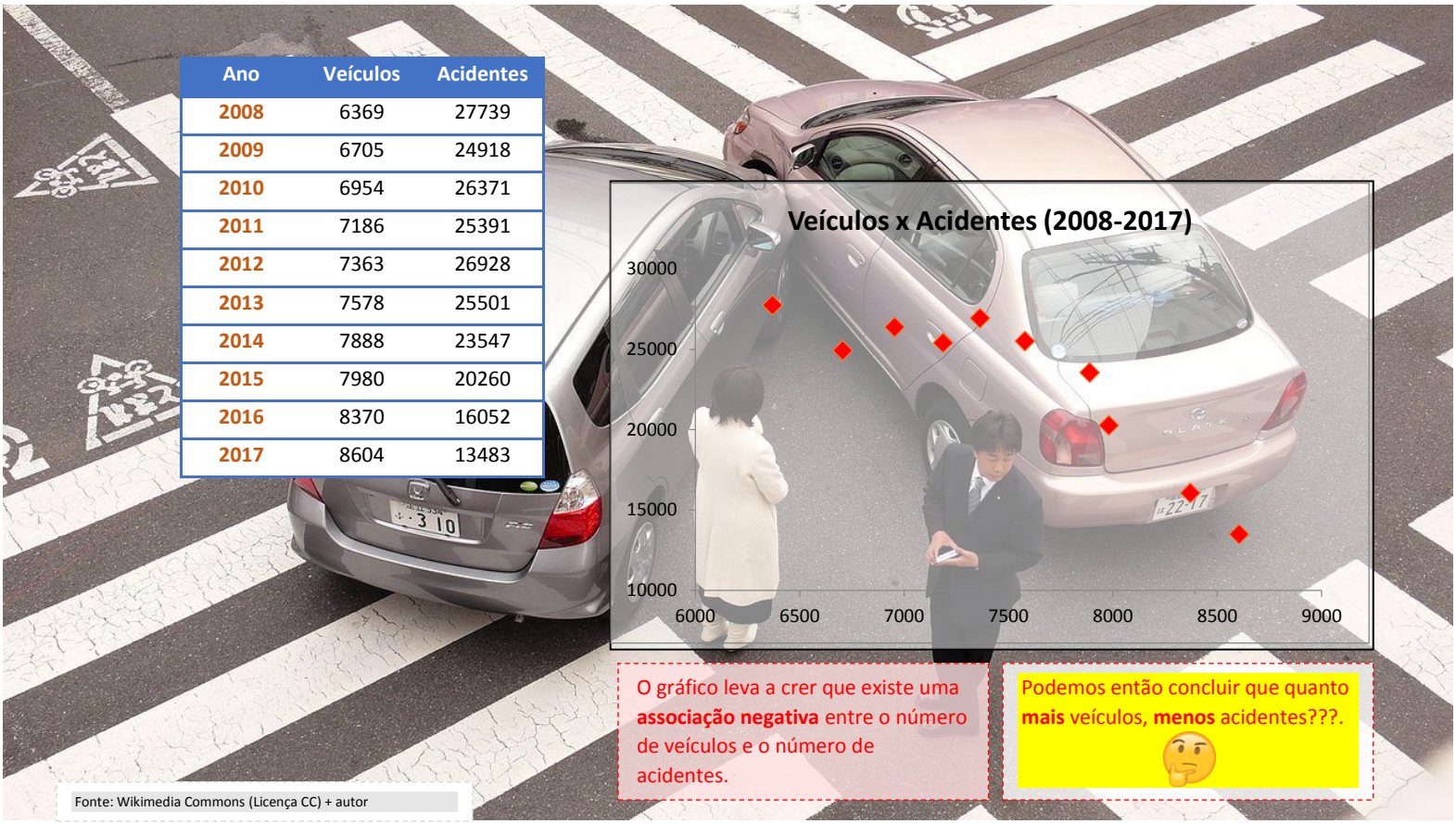
Ainda que tenha havido um acréscimo do valor de hora neste ponto, a associação é majoritariamente negativa.

Neste gráfico, temos como variável explicativa a quantidade máxima de horas de permanência de um veículo no estacionamento, e como variável resposta, o custo do estacionamento por hora.

Fonte: Wikimedia Commons / Pixabay (Licenças CC) + autor

Analisemos um outro exemplo, relativo ao crescimento do número de veículos da cidade de São Paulo e o número de acidentes registrado por ano (dados de 2008 a 2017; Fonte: CET-SP) http://www.cetsp.com.br/media/785452/Relatorio_anual_acidentes_transito_2017.pdf

Ano	Veículos	Acidentes
2008	6369	27739
2009	6705	24918
2010	6954	26371
2011	7186	25391
2012	7363	26928
2013	7578	25501
2014	7888	23547
2015	7980	20260
2016	8370	16052
2017	8604	13483



Fonte: Wikimedia Commons (Licença CC) + autor

Novamente aqui, estamos diante de uma relação entre duas variáveis em que temos que tomar bastante cuidado com os seguintes aspectos:

- **A associação não leva a causalidade**: não é razoável imaginar que o aumento do número de veículos seja o causador da redução de acidentes. As variáveis podem aparentar estar associadas sem que haja uma real relação entre elas.
- **Atenção com as variáveis escondidas**: há diversos fatores que podem ter impactado na redução do número de acidentes no período de 2008 a 2017, como: redução nas velocidades das vias; campanhas de educação no trânsito; veículos mais inteligentes; vias mais bem sinalizadas; aumento no número de radares; etc.

Cruzeiro do Sul Virtual
Educação a Distância

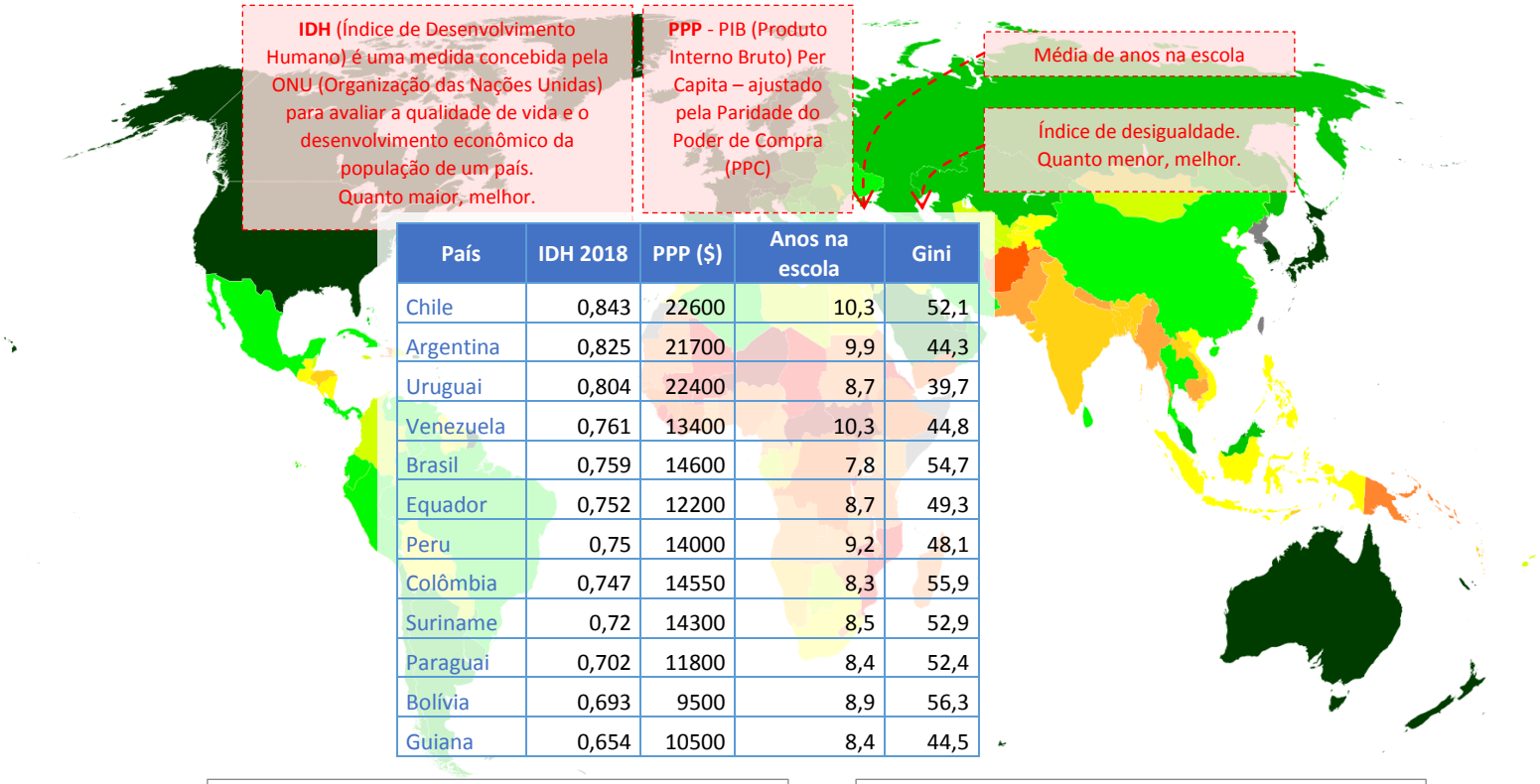
Vejamos mais um exemplo, com dados sobre os países da América do Sul:

IDH (Índice de Desenvolvimento Humano) é uma medida concebida pela ONU (Organização das Nações Unidas) para avaliar a qualidade de vida e o desenvolvimento econômico da população de um país. Quanto maior, melhor.

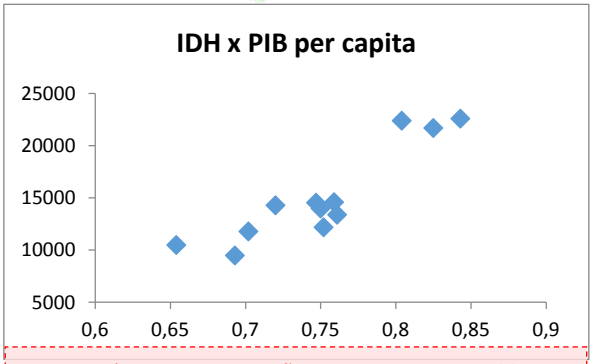
PPP - PIB (Produto Interno Bruto) Per Capita – ajustado pela Paridade do Poder de Compra (PPC)

Média de anos na escola

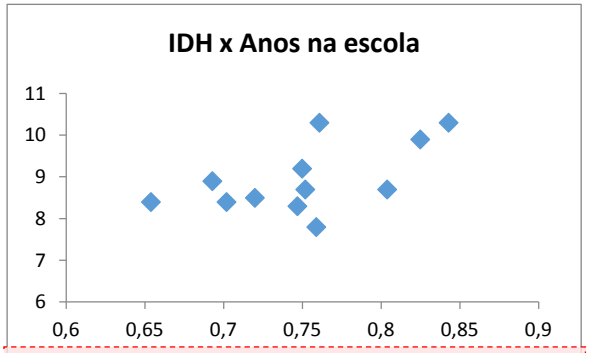
Índice de desigualdade. Quanto menor, melhor.



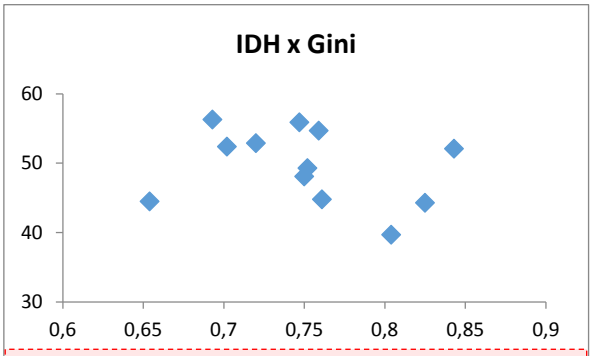
País	IDH 2018	PPP ($)	Anos na escola	Gini
Chile	0,843	22600	10,3	52,1
Argentina	0,825	21700	9,9	44,3
Uruguai	0,804	22400	8,7	39,7
Venezuela	0,761	13400	10,3	44,8
Brasil	0,759	14600	7,8	54,7
Equador	0,752	12200	8,7	49,3
Peru	0,75	14000	9,2	48,1
Colômbia	0,747	14550	8,3	55,9
Suriname	0,72	14300	8,5	52,9
Paraguai	0,702	11800	8,4	52,4
Bolívia	0,693	9500	8,9	56,3
Guiana	0,654	10500	8,4	44,5



Parece haver uma associação majoritariamente **positiva**, embora **alguns** pontos não a obedeçam



Parece haver uma associação que tende a ser **positiva**, embora **muitos** pontos não a obedeçam



Parece haver uma associação que tende a ser **negativa**, embora **muitos** pontos não a obedeçam (em especial, países com IDH alto e desigualdade alta)

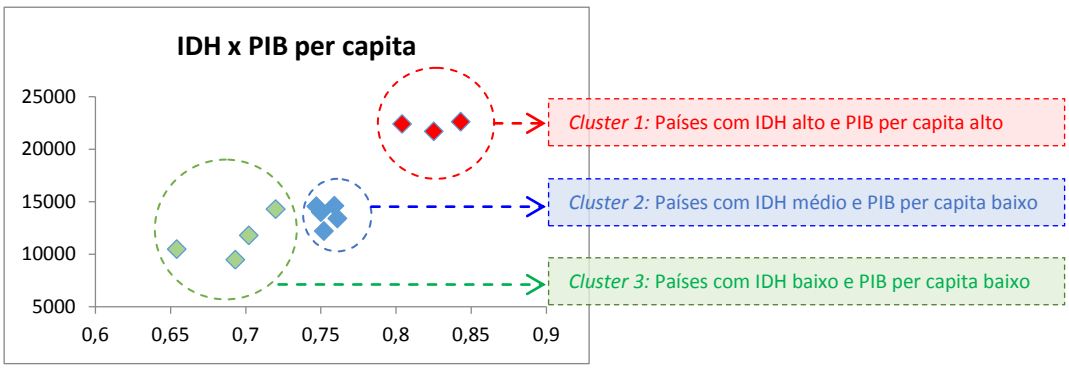
Não há uma associação clara.

Como vimos no exemplo anterior, nem sempre é possível estabelecer associações entre duas variáveis. Além disso, as classificações das associações em positivas e negativas tem sentido apenas para associações **lineares** (na qual os pontos se espalham ao redor de uma reta imaginária). Há associações que podem não ser lineares.

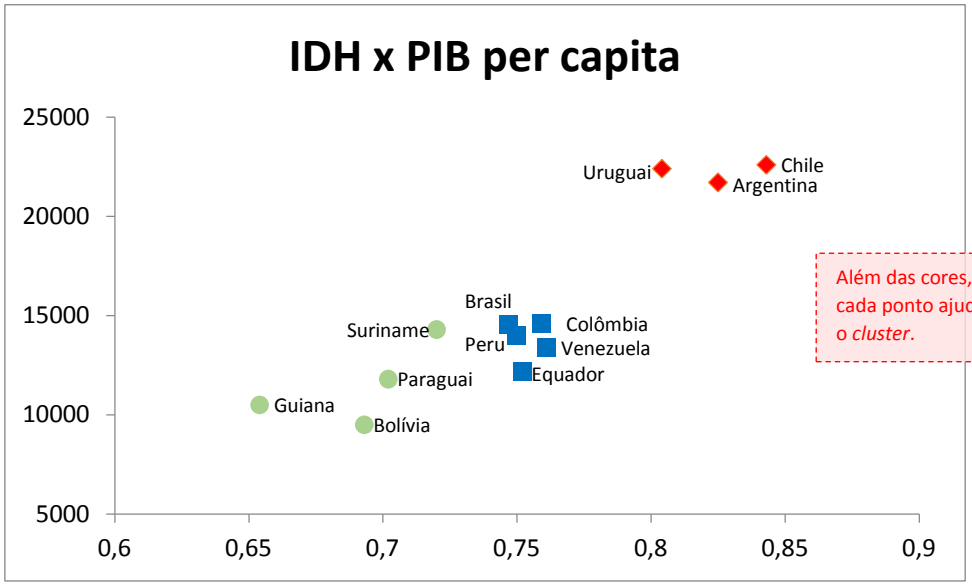
Agrupamentos (*clusters*)

Uma das aplicações importantes dos diagramas de dispersão é a detecção de agrupamentos (também chamados de aglomerados ou *clusters*) – são grupos de pontos que, por sua proximidade no gráfico, revelam as características de um conjunto específico de indivíduos.

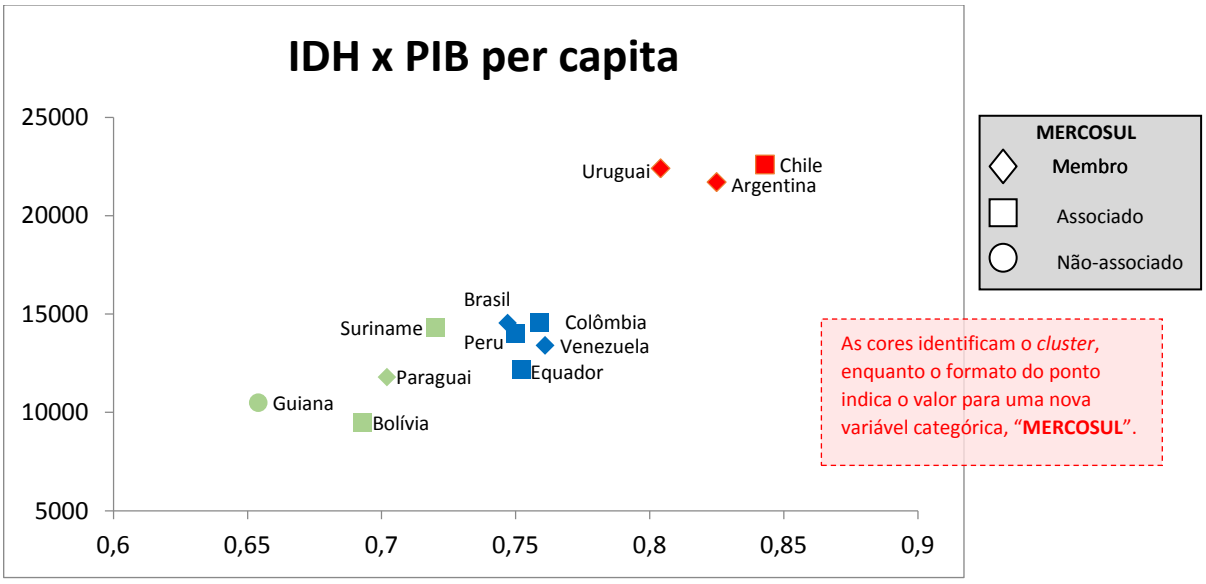
Vejamos um exemplo, baseado no *dataset* relacionado aos dados socioeconômicos da América do Sul, apresentado anteriormente.



Note que a definição do número de clusters e a distribuição dos indivíduos por cluster depende dos critérios de separação adotados no processo de Análise de Dados. Veja também que é possível incluir símbolos que diferenciem os clusters, bem como pode-se identificar os indivíduos no gráfico, se necessário, como mostrado a seguir:



Além de ajudarem a identificar clusters diferentes, símbolos e cores podem ser empregados também para a inclusão de variáveis categóricas em um gráfico de dispersão.



Cruzeiro do Sul Virtual
Educação a Distância

Para saber mais, leia o capítulo 4 do e-book:

MOORE, David S.; NOTZ, William I.; FLINGER, Michael A. A Estatística Básica e sua Prática. 7 ed. Rio de Janeiro: LTC, 2017 – Capítulo 4