Dados, Informação e a Arquitetura de Dados:

Em busca do conhecimento através da IA

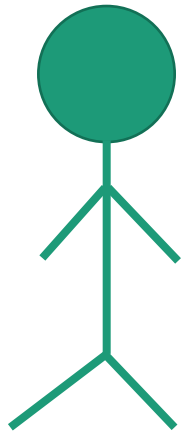
Prof. Dr. Dieval Guizelini

Primeiro exercício:
“Abstração, a essência da computação”

O que está “representado” no quadro abaixo?

Primeiro exercício:
“Abstração, a essência da computação”

O que está “representado” no quadro abaixo?



Uma pessoa, qual ou quem?

Capacidade de desenhar e a capacidade de imaginar ou associar esse boneco a alguém

Conseguimos dizer quem é?
Podemos afirmar, estimar ou inferir alguma coisa?

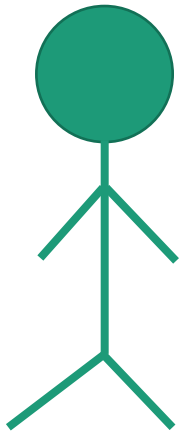
Ex: será que essa pessoa pode engravidar?

abstrato

Concreto ou real

Primeiro exercício:
“Abstração, a essência da computação”

O que está “representado” no quadro abaixo?



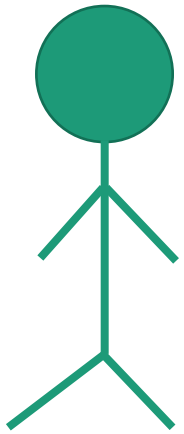
Uma pessoa, qual ou quem?

- Mulher
- Data de nascimento: 30/03/1993
- Idade: 28 anos
- Signo: ariana
- Altura: 1,62m
- Peso: 67kg
- Natural: Rio de Janeiro
- Profissão: ?

Essas informações pertencem ou descrevem uma pessoa,
mas são o suficiente para identificar uma pessoa ou um grupo específico de pessoas?

Primeiro exercício:
“Abstração, a essência da computação”

O que está “representado” no quadro abaixo?

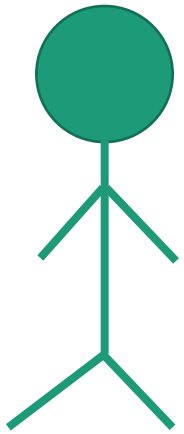


Uma pessoa, qual ou quem?

- Mulher
- Data de nascimento: 30/03/1993
- Idade: 28 anos
- Signo: ariana
- Altura: 1,62m
- Peso: 67kg
- Natural: Rio de Janeiro
- Profissão: Cantora,
  iniciou a carreira em 2010

Primeiro exercício:
“Abstração, a essência da computação”

O que está “representado” no quadro abaixo?



Uma pessoa, qual ou quem?

- Mulher
- Data de nascimento: 30/03/1993
- Idade: 28 anos
- Signo: ariana
- Altura: 1,62m
- Peso: 67kg
- Natural: Rio de Janeiro
- Profissão: Cantora,
  iniciou a carreira em 2010

- Nome: **Larissa de Macedo Machado**

Provavelmente um grupo de pessoas já associaram o nome a pessoa,
mas também é certo que muitas pessoas conhecem a cantora por seu nome artístico.

De igual forma, podemos adicionar muito mais “dados” para descrever essa pessoa e eles ainda podem não ter relação com o objetivo da aplicação.

Primeiro exercício:
“Abstração, a essência da computação”

O que está “representado” no quadro abaixo?

Uma pessoa, qual ou quem?

- Mulher
- Data de nascimento: 30/03/1993
- Idade: 28 anos
- Signo: ariana
- Altura: 1,62m
- Peso: 67kg
- Natural: Rio de Janeiro
- Profissão: Cantora, iniciou a carreira em 2010

- Nome: **Larissa de Macedo Machado**
- nome artístico: **Anittta**

Conjunto de atributos, características, campos e dados

Uma entidade, uma pessoa, um objeto ou **uma instância**

Definição da abstração

- Matemática:
  abstração é o processo de extrair a essência fundamental de um conceito matemático, removendo qualquer dependência do mundo real

- Filosofia:
  isolar um elemento à exclusão de outros

- Computação:
  Abstração é a habilidade de concentrar nos aspectos essenciais de um contexto qualquer, ignorando características menos importantes ou acidentais.

Dado e informação no mundo e na computação

- Dicionários (adjetivo):
  1. Informações que identificam o individuo
  2. Representação de fatos, conceitos e instruções, por meio de sinais, de maneira formalizada, possível de ser transmitida ou processada pelo homem ou por máquinas.
  3. fatos ou informações, especialmente quando examinados e usados para descobrir coisas ou para tomar decisões

- Na computação:
  1. Um estado atribuído a uma variável
  2. Uma representação/abstração de uma informação
  3. Um valor “atômico” e desprovido de referencial (ou dimensão)

Dado e informação no mundo e na computação

- Para a filosofia, a “Informação é a base de dados, é aquilo que a gente coleta, para construir o conhecimento” (Mario Sergio Cortella, 2020)

- Informação é o conjunto de dados e conhecimentos organizados, que possam constituir referências sobre um determinado acontecimento, fato ou fenômeno.

- Informação é representada ou descrita como um conjunto de dados relacionados e sistematizado.

- Denominamos informação o conjunto de dados uteis para processar (calcular) alguma coisa.

Gerenciando e organizando a Informação...

- Sistemas de informação: é o modelo de processos responsáveis por coletar, manter e transmitir dados que sejam úteis ao desenvolvimento de produtos ou serviços das empresas, organizações e de demais projetos.

- Os sistemas podem ser especializados: sistemas de monitoramento do volume de água de uma represa, e os sistemas podem se tornar componentes ou tecnologias para outros sistemas, tais como os sistema de gerenciamento de banco de dados.

- A Tecnologia da Informação atualmente é referenciada como área, mas devemos lembrar que são "recursos" e "meios" para aquisição, manutenção, recuperação e transmissão da informação.
  Fora dos computadores: livros, cadernos são exemplos de tecnologias...
  Computadores, a internet são exemplos de tecnologias

- As tecnologias, normalmente, acrescentam características de qualidade a informação.

Ao longo da história...

Entre 1950 e 1970	Entre 1970 e 1990	Entre 1990 e 2010	Tempo atual
Dados contabilizados e manipulados por hardwares específicos para finalidades específicas...	O software entra em evidência e manipula **informação**. Observe que o conjunto de instruções do hardware sofre pouca modificação, mas o software combina as instruções e amplia o escopo de utilização dos dados	Internet, informação distribuída e a "conexão" em escala... De quase tudo com quase tudo...	A era do **conhecimento**.
Profissionais da **área**			
Cuidavam das máquinas e manipulava os dados: Ciência da Computação Área: **computação**	Criaram programas e definiram os sistemas Processamento de dados / Analista de Sistemas / Programador Área: **computação / informática**	Web, organizar volumes de dados em ambientes distribuídos Área: TIC, Ciência de Dados, Eng. de dados...	

Mas, o que é o “conhecimento” no contexto da Inteligência Artificial?

- O conhecimento é um modelo matemático, normalmente, construído a partir de um conjunto de observações descritas na forma de “padrões” ou instâncias.

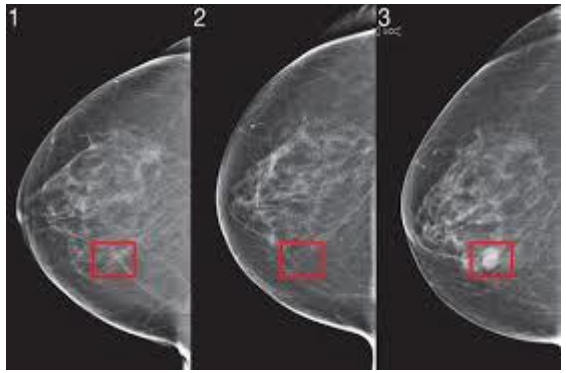
- O conhecimento modelado na forma matemática/computacional nem sempre pode ser descrito ou explicado, mas pode e deve ser útil, no sentido de poder ser aplicado para classificar ou predizer algo.

A motivação para “coletar” os dados...

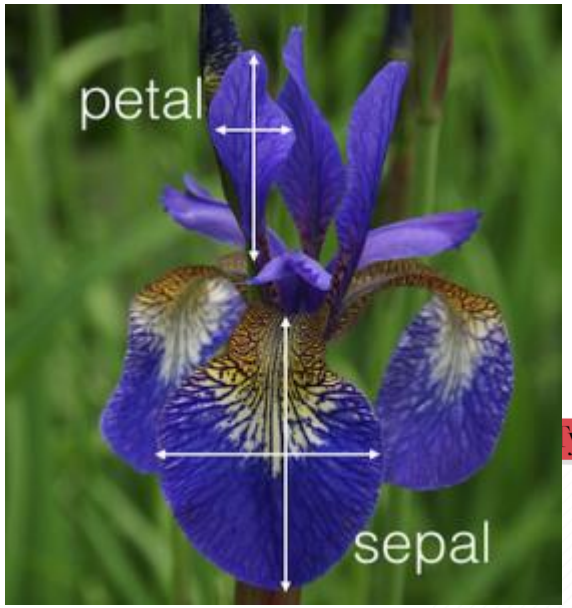
problema | Preparação dos dados | análises | Aplicações de IA

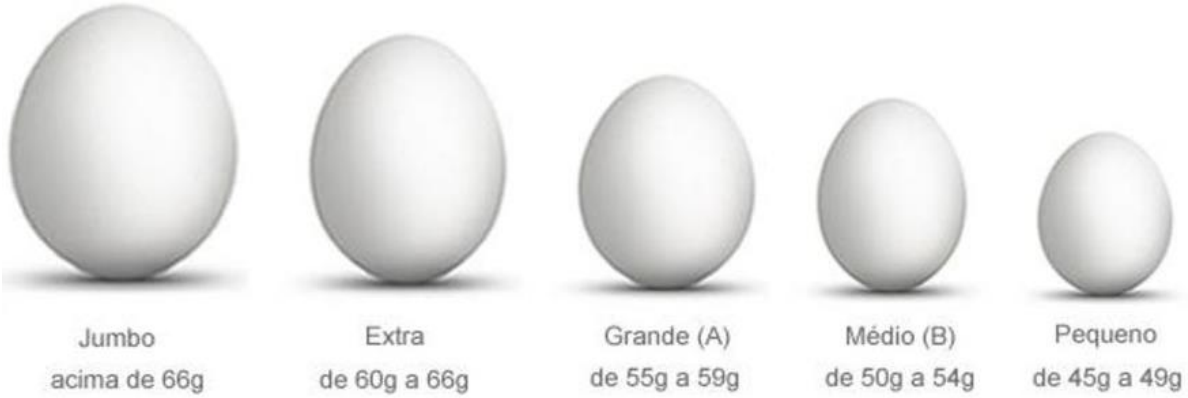
- Problema: a motivação para organizar a informação, para buscar a explicação, para descrever uma situação...
- Definir o escopo (limites, definições, contexto, público...)
- Prever e descrever um resultado esperado
- Tentar prever como avaliar os resultados

Quando não temos “o problema”, mas temos “observações”...

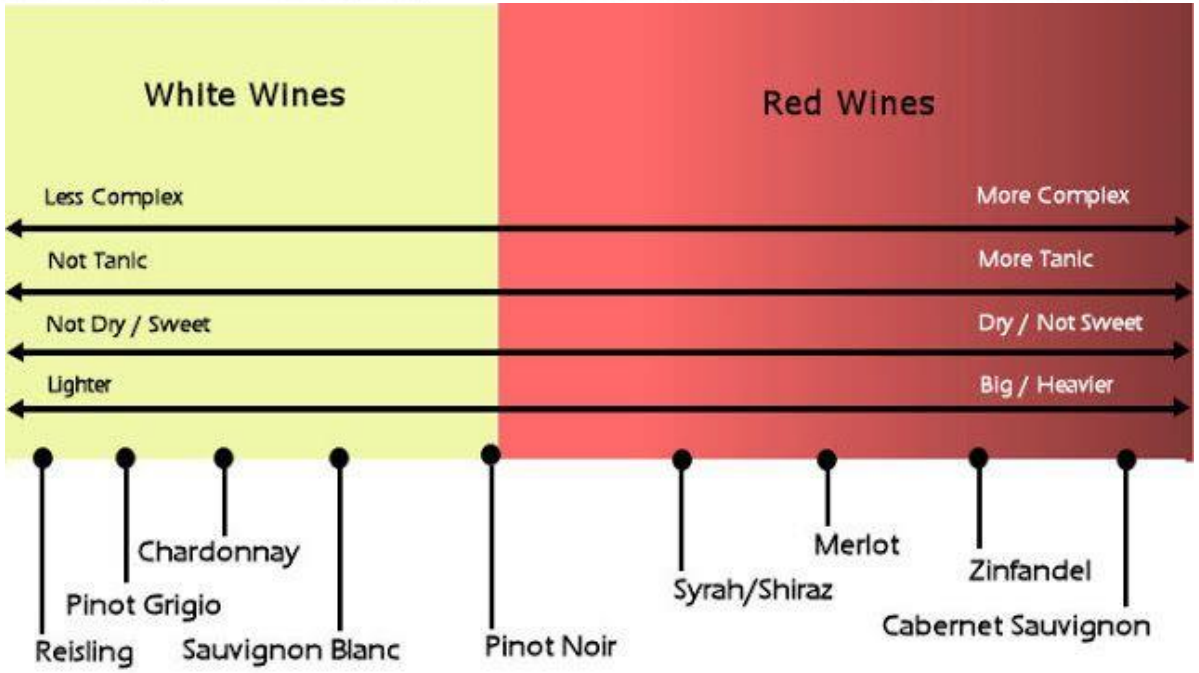


Alguns problemas são mais fáceis de “identificar” as características necessárias…





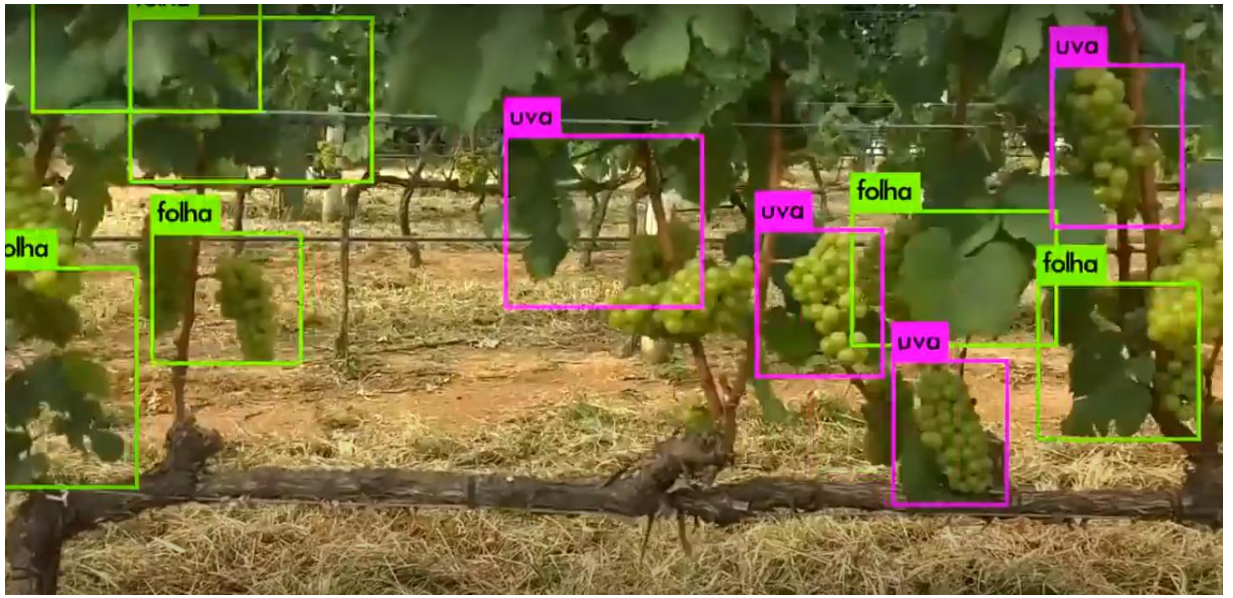
Sparkling	• Champagne • Prosecco • Cava	• Sekt • Cremant • Rosé	• American Sparkling Wine • Moscato d'Asti • Lambrusco
Dry White	• Pinot Grigio • Albariño • Grüner Veltliner	• Sauvignon Blanc • Chardonnay (Unoaked) • Muscadet	• Gewürtztraminer • Riesling • Vihno Verde
Sweet White	• Moscato d'Asti • Moscato • Gewurztraminer	• Moscatel • Ice Wine • Riesling	• Late Harvest Wine • Tokaji • Sauternes
Rich White	• Vourvay • Pinot Gris • Viognier	• Chardonnay (Oaked) • White Rhone • Semillon	• Chenin Blanc • White Rioja
Light Red	• Gamay • Cinsault • Lambrusco	• Nebbiolo • Primitivo • Pinot Noir	
Medium Red	• Grenache • Carmenere • Cabernet Franc	• Sangiovese • Negroamaro • Rhone Blend	• Merlot • Montepulciano • Zinfandel
Bold Red	• Tempranillo • Malbec • Bordeaux Blend	• Cabernet Sauvignon • Syrah / Shiraz • Mourvedre	• Pinotage • Petite Sirah • Tannat
Dessert / Liqueur	• Moscato d'Asti • Muscat • Sauternes	• Madeira • Sherry • Port	• Vinsanto • Tokaji



Outros problemas precisamos ser “inventivos”...

Fig. 1: Different samples of handwritten digits in MNIST

Imagens estática e dinâmica...



https://www.youtube.com/watch?v=YgZbTca1hl8

A origem do dado...

problema | Preparação dos dados | análises | Aplicações de IA

- Preparação dos dados:
  - Extrair características que descrevem o objeto, a situação, as entidades...
  - Medições (escala, precisão, proporção, relação...)
  - Como medir, como reproduzir...
  - Como coletar, quantidade de amostras etc

A coleta da informação...

- Meios digitais...
- Cadastros físicos, fichas, cartões...
- Derivados de outros dados...
- Proveniente de equipamentos...

- O tipo da informação:
  texto, som, imagem, sinal etc

A origem do dado...

problema | Preparação dos dados | análises | Aplicações de IA

- Preparação dos dados:

  E quando os dados já foram coletados?

  - Combinar e derivar os atributos/características
  - Sumarizar os dados
  - Olhar/identificar os erros
  - Transformar os dados ou a representação dos dados
  - Segmentar os dados

A origem do dado...

problema | Preparação dos dados | análises | Aplicações de IA

- Análises dos atributos / campos / característica:

Metadado (esquema)	Tipo de dado (texto, número, data, hora etc Tamanho (fixo, variável, tamanhos mínimos e máximos) Operações possíveis/derivadas Valores indefinidos ou ausentes Alguma regra de validação/geração (séries) Notação/padronização
Dados (instância, padrão, observação, fato, evento...)	Amostra, um exemplo, um dado específico Associado a uma dimensão ou referência Um valor adimensional

A origem do dado...

problema | Preparação dos dados | análises | Aplicações de IA

Como explorar os atributos e o contexto:

- Visualização gráfica
- Sumarização dos dados
- Explorar as relações entre os atributos
- Agrupar dados
- Identificar fatos não triviais, padrões e tendências
- Construir modelos de regressão
- Construir modelos de classificação

Estatística como ferramenta de descrição

- Estimativas de localização:
  médias, medianas, modas, mínimos e máximos...

- Estimativas de variabilidade:
  desvio padrão, variância...

- Distribuição dos dados:
  percentis, boxplots, tabelas de frequências, histogramas...

- Correlação, entropia...

A tabela (coleção de atributos e instâncias)

Variáveis do problema

Atributo 1	Atributo 2	Atributo 3	Atributo ...	Atributo N	Classe
Valor11	Valor12	Valor	...	Valor1n	Rótulo A
Valor21	Valor22	Valor	...	Valor2n	Rótulo B
Valor...	...	...	...		Rótulo ...
Valorn1	Valorn2	...	...	Valornn	Rótulo N

dado (Valor11 … Valorn1)

dado (Valor1n … Valornn)

A tabela (coleção de atributos e instâncias)

Esquema (descrição do registro, definição da estrutura, definição da classe...)

variáveis

Atributo 1	Atributo 2	Atributo 3	Atributo ...	Atributo N	Classe
Valor11	Valor12	Valor	...	Valor1n	Rótulo A
Valor21	Valor22	Valor	...	Valor2n	Rótulo B
Valor...	...	...	...		Rótulo ...
Valorn1	Valorn2	...	...	Valornn	Rótulo N

dado (Atributo 1) **dado** (Atributo N)

Instâncias (banco de dados)

A tabela (coleção de atributos e instâncias)

	Atributo 1	Atributo 2	Atributo 3	Atributo ...	Atributo N	Classe
Padrão 1	Valor11	Valor12	Valor	...	Valor1n	Rótulo A
Padrão 2	Valor21	Valor22	Valor	...	Valor2n	Rótulo B
Padrão ...	Valor...	...	...	...		Rótulo ...
Padrão n	Valorn1	Valorn2	...	...	Valornn	Rótulo N

Padrão = Instâncias (banco de dados)

A tabela (coleção de atributos e instâncias)

Formato do padrão para uso nas duas estratégias de aprendizado (modelo supervisionado e não supervisionado)

	Atributo 1	Atributo 2	Atributo 3	Atributo ...	Atributo N	Classe
Padrão 1	Valor11	Valor12	Valor	...	Valor1n	Rótulo A
Padrão 2	Valor21	Valor22	Valor	...	Valor2n	Rótulo B
Padrão ...	Valor...	...	...	...		Rótulo ...
Padrão n	Valorn1	Valorn2	...	...	Valornn	Rótulo N

Formato do padrão durante a fase de treinamento Para o modelo de aprendizado supervisionado

A origem do dado...

problema | Preparação dos dados | análises | Aplicações de IA

- Análises dos atributos / dados (algoritmos/estratégias):

Alvo	Supervisionado	Não supervisionado
Instância / padrão / tuplas	Estratificação	Embaralhar (shuffle)
	Reamostragem	Remover Duplicados
	Balanciamento	Remover falores frequentes
		Remover não classificados
		Remover faixas
		Remover com valores
Atributos	Ordem de classe	Produto cartesiano
	Seleção de atributos	Derivar data
	Discretização	Normalização
		Análise de Componentes Principal

Descrevendo os dados

1. Observações e variáveis

Nome	Consumo MPG	Cilindradas	autonomia	peso	Potência	aceleração	Ano modelo	Origem
Chevrolet Chevelle Malibu	18	8	307	3504	130	23	70	America
Buick Skylark 320	15	8	350	3693	165	11,5	70	America
Plymouth Satellite	18	8	318	3436	150	11	70	America
AMC Rebel SST	16	8	304	3433	150	12	70	America
Ford Torino	17	8	302	3449	140	10,5	70	America

MPG: miles per gallon

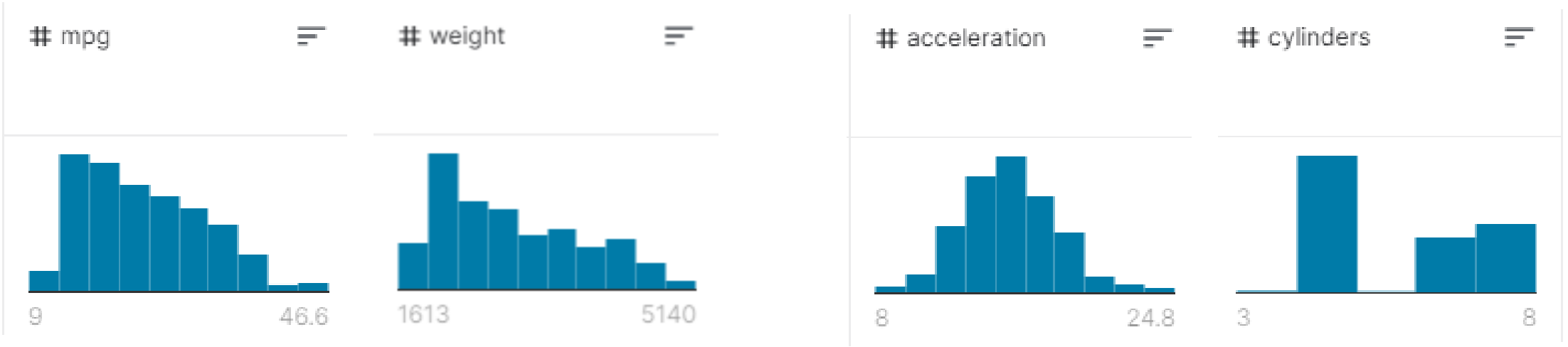
https://www.rpubs.com/dksmith01/cars

Tendência central

média, mediana, moda...

```
##        ID              mpg          cylinders      displacement
##  Min.   :  1.0   Min.   : 9.00   Min.   :3.000   Min.   : 68.0
##  1st Qu.:100.2   1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.2
##  Median :199.5   Median :23.00   Median :4.000   Median :148.5
##  Mean   :199.5   Mean   :23.51   Mean   :5.455   Mean   :193.4
##  3rd Qu.:298.8   3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0
##  Max.   :398.0   Max.   :46.60   Max.   :8.000   Max.   :455.0
##
##    horsepower      weight      acceleration       model
##  150    : 22   Min.   :1613   Min.   : 8.00   Min.   :70.00
##  90     : 20   1st Qu.:2224   1st Qu.:13.82   1st Qu.:73.00
##  88     : 19   Median :2804   Median :15.50   Median :76.00
##  110    : 18   Mean   :2970   Mean   :15.57   Mean   :76.01
##  100    : 17   3rd Qu.:3608   3rd Qu.:17.18   3rd Qu.:79.00
##  75     : 14   Max.   :5140   Max.   :24.80   Max.   :82.00
##  (Other):288
##      origin                  car_name       price
##  Min.   :1.000   ford pinto     :  6   Min.   : 1598
##  1st Qu.:1.000   amc matador    :  5   1st Qu.:23110
##  Median :1.000   ford maverick  :  5   Median :30000
##  Mean   :1.573   toyota corolla:  5   Mean   :29684
##  3rd Qu.:2.000   amc gremlin    :  4   3rd Qu.:36430
##  Max.   :3.000   amc hornet     :  4   Max.   :53746
##                  (Other)        :369
```

Distribuição de frequências, histogramas...

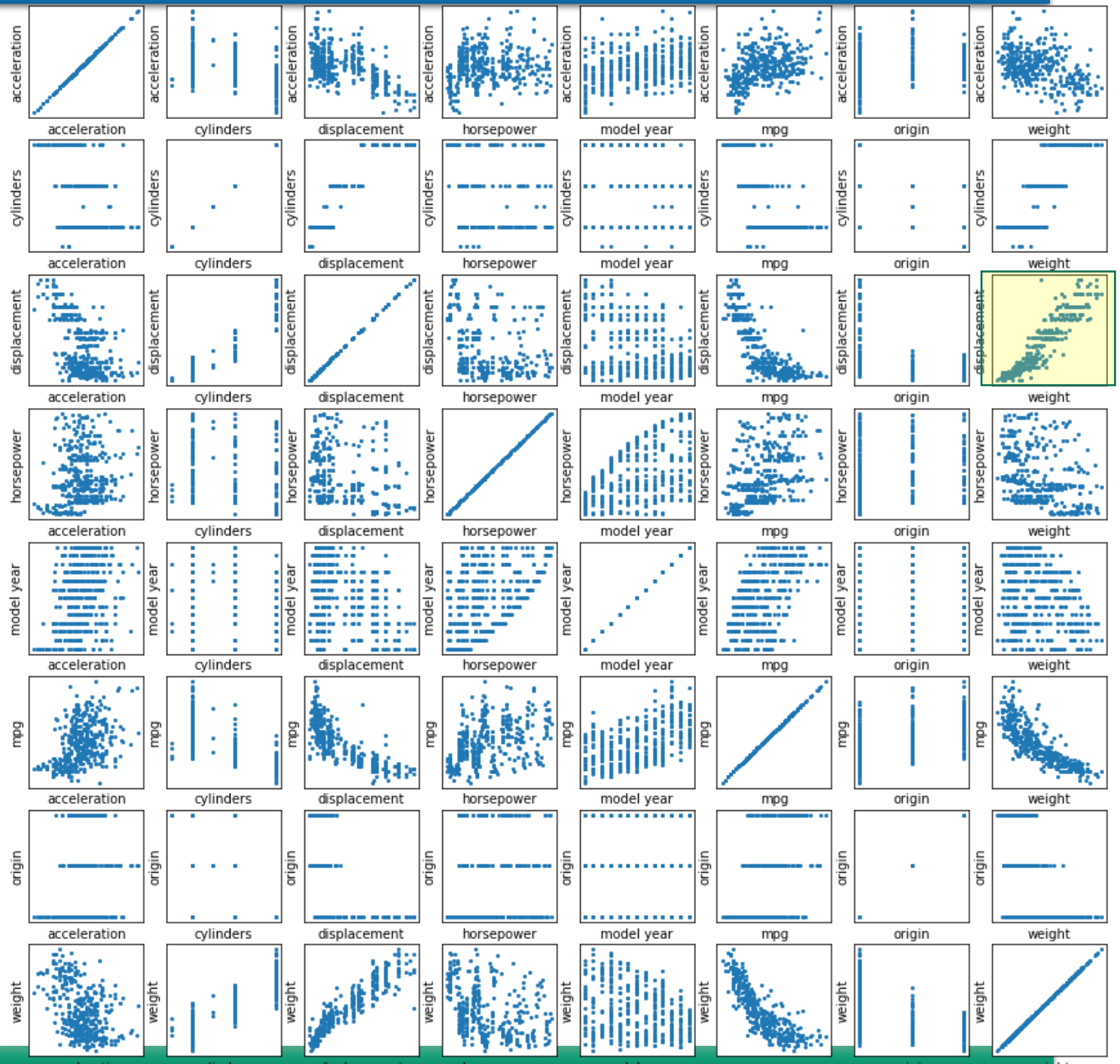


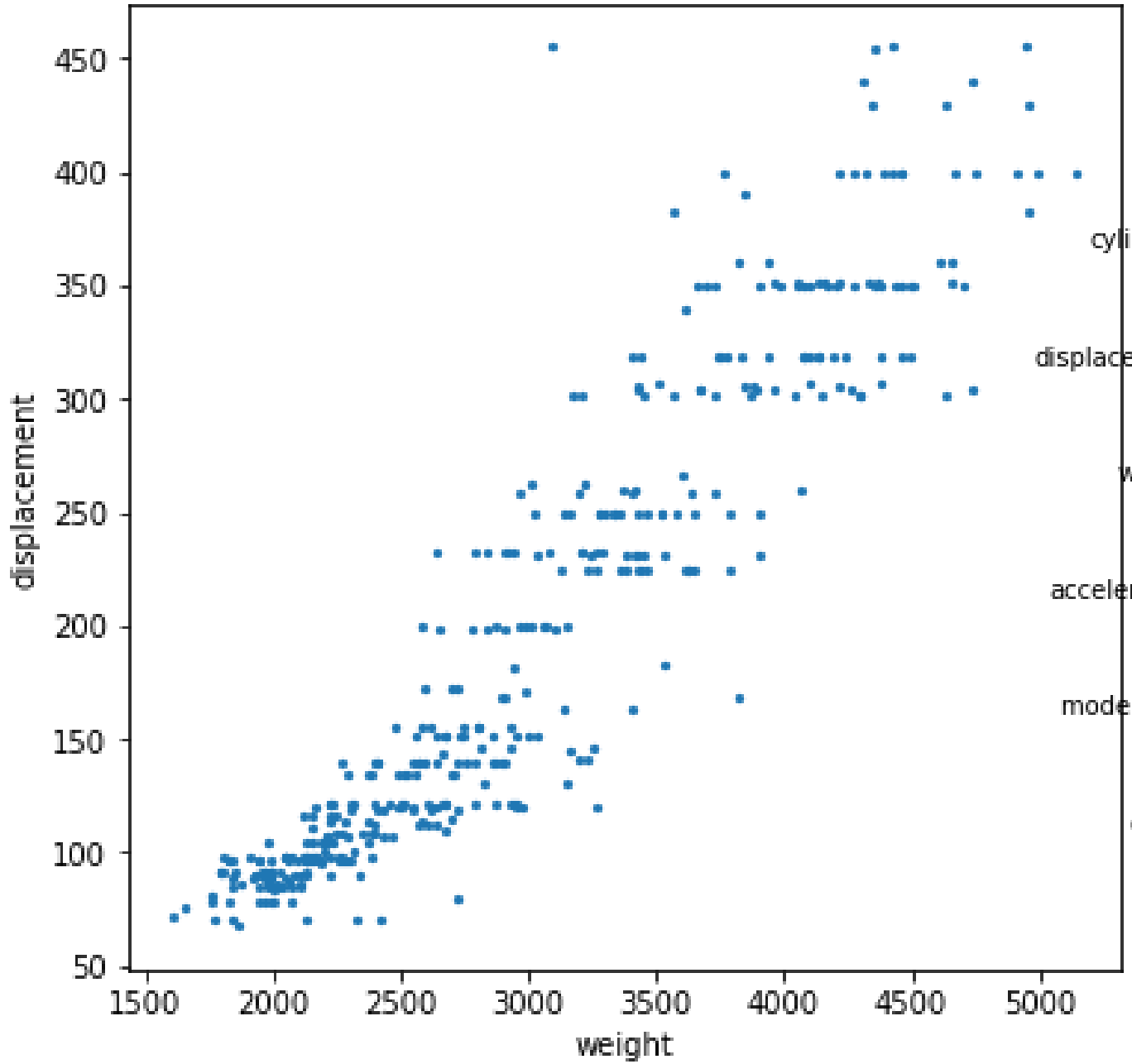
Fonte: https://www.kaggle.com/uciml/autompg-dataset?select=auto-mpg.csv

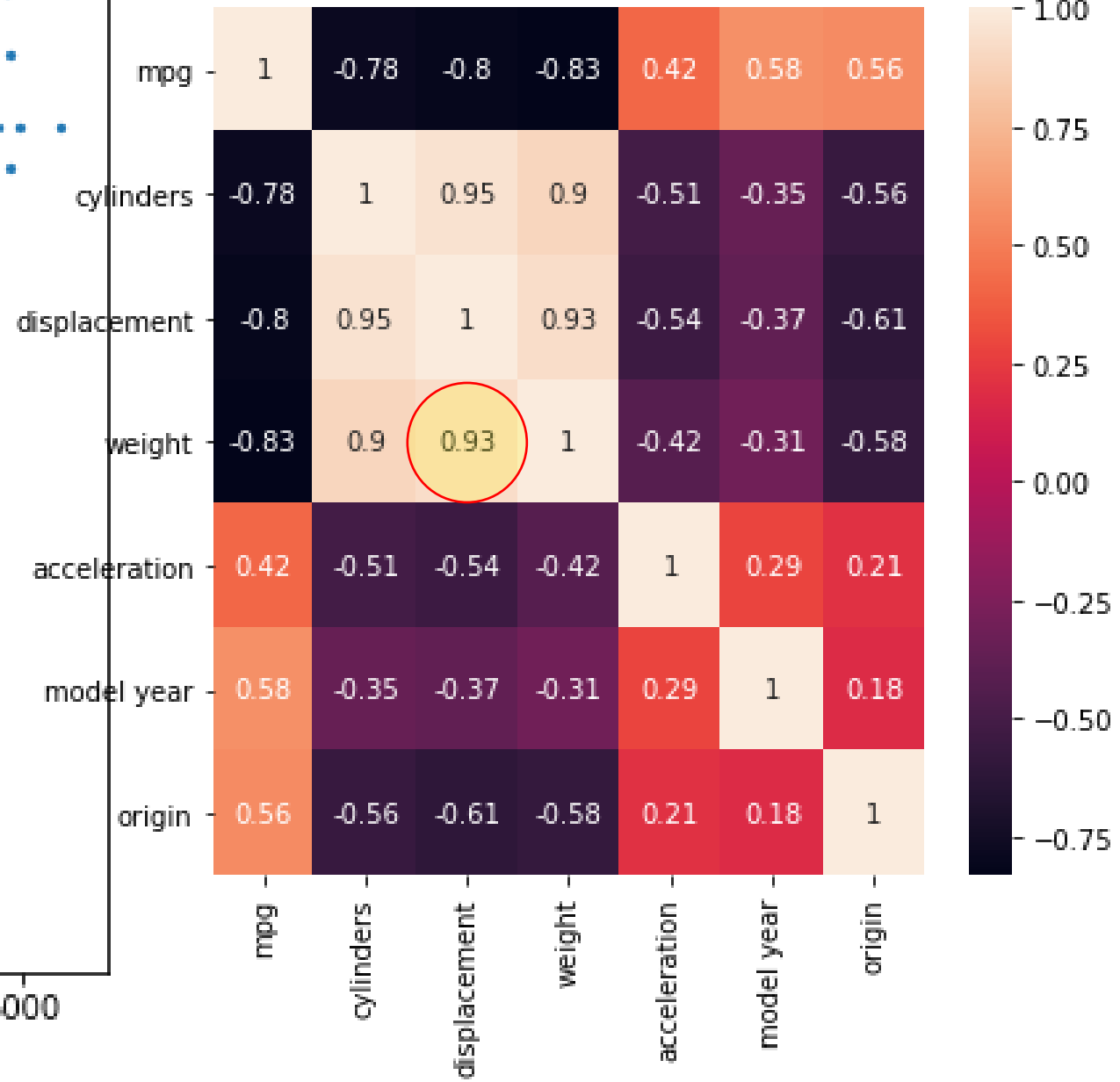
Preparando as tabelas de dados

- Limpeza dos dados
- Remoção de observações e variáveis
- Consistências das escalas entre as variáveis
- Conversão de texto para representação numérica/codificação
- Converter dados contínuos para categorizados
- Combinação de variáveis
- Criação de grupos

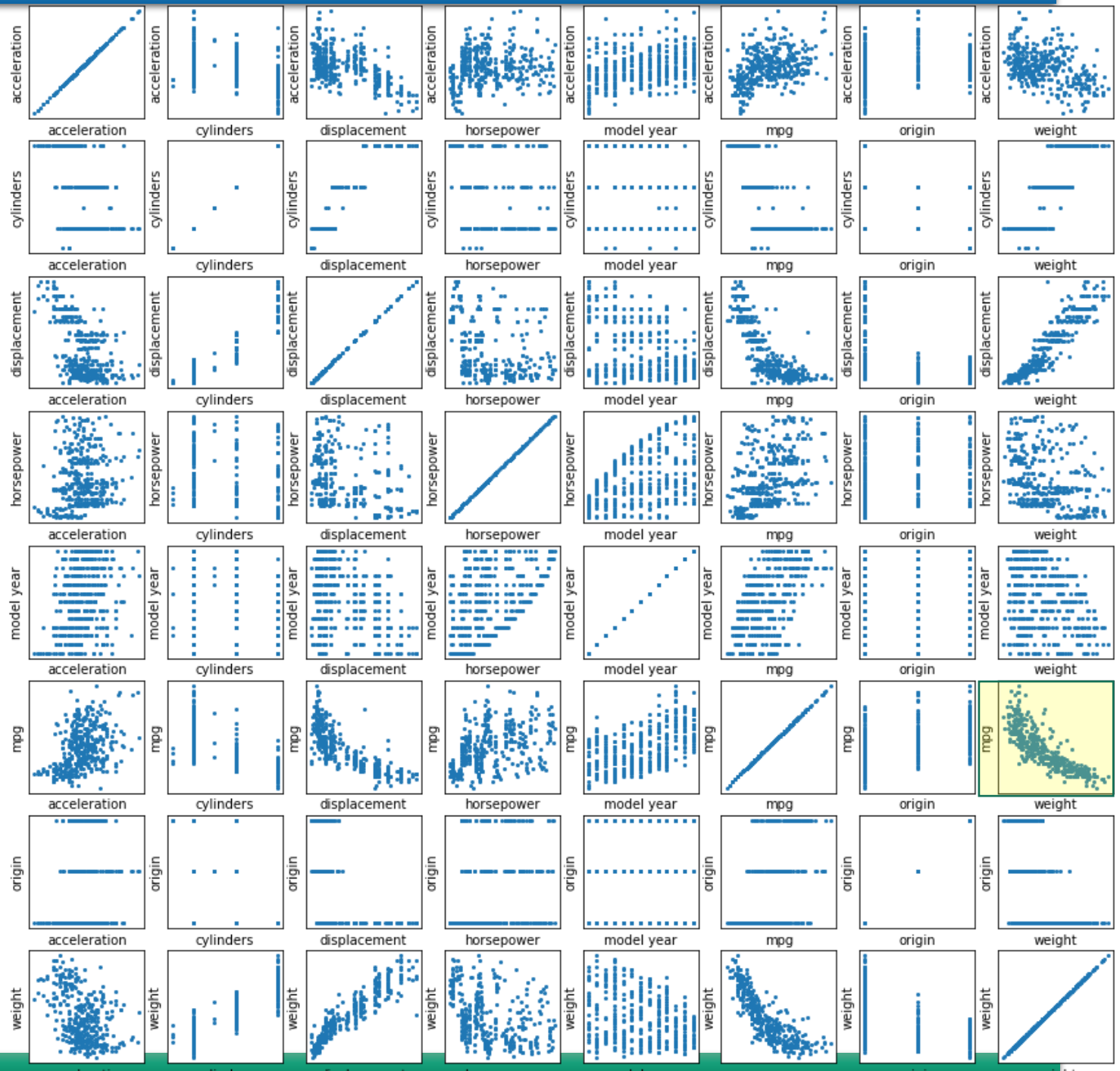
Estudo das relações entre atributos

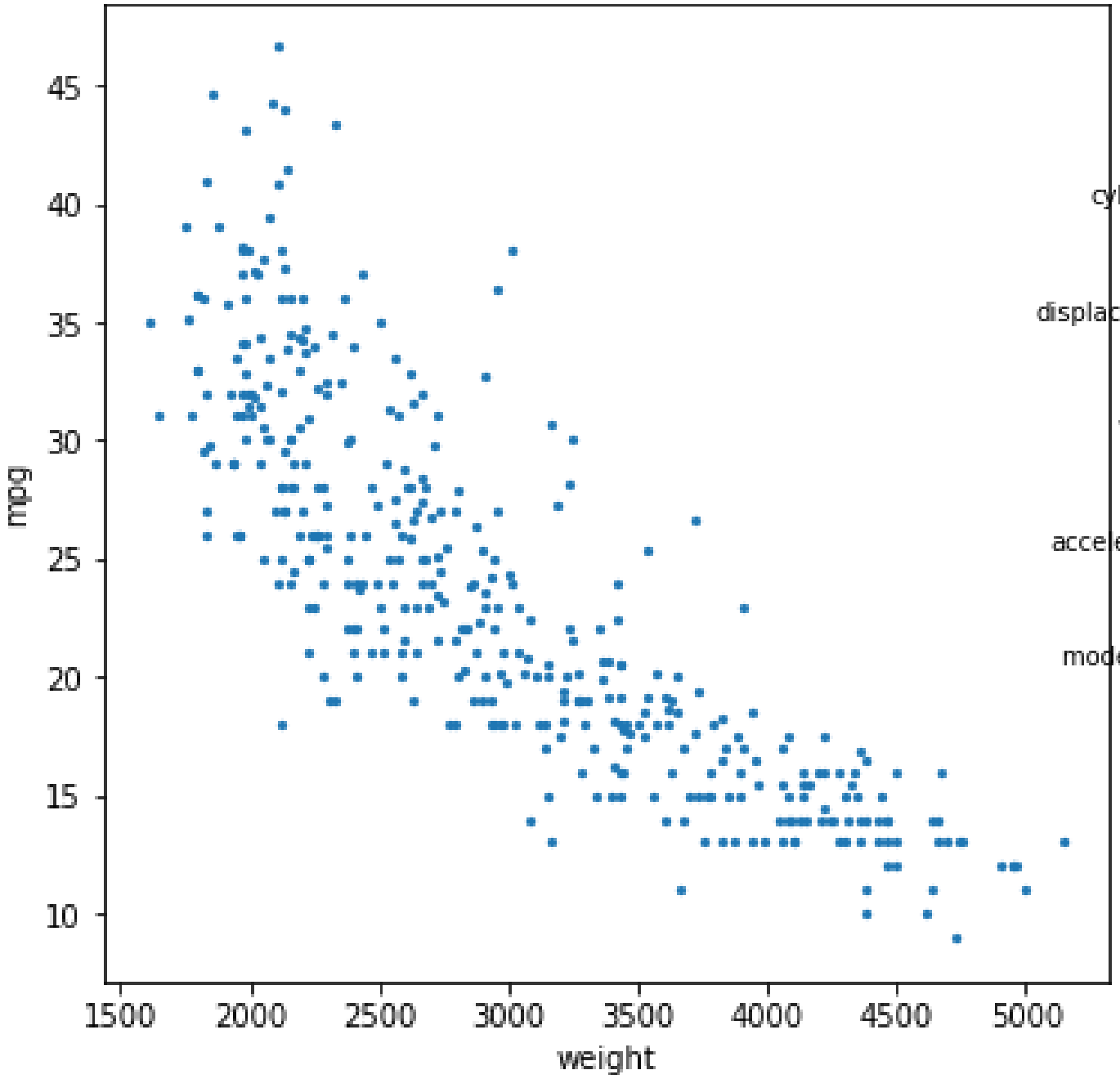


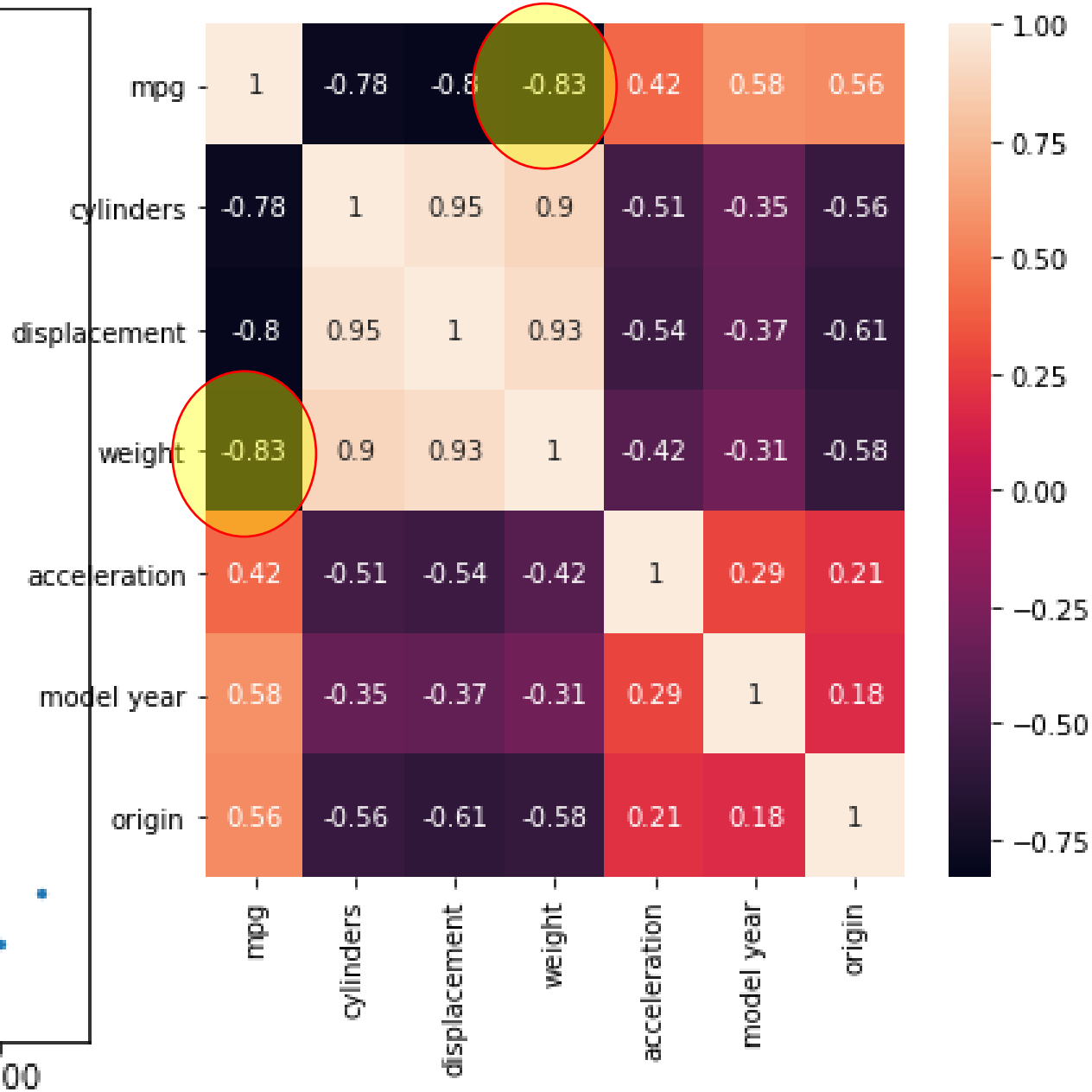

displacement
weight
450
400
350
300
250
200
150
100
50
1500
2000
2500
3000
3500
4000
4500
5000


mpg
cylinders
displacement
weight
acceleration
model year
origin
1 -0.78 -0.8 -0.83 0.42 0.58 0.56
-0.78 1 0.95 0.9 -0.51 -0.35 -0.56
-0.8 0.95 1 0.93 -0.54 -0.37 -0.61
-0.83 0.9 0.93 1 -0.42 -0.31 -0.58
0.42 -0.51 -0.54 -0.42 1 0.29 0.21
0.58 -0.35 -0.37 -0.31 0.29 1 0.18
0.56 -0.56 -0.61 -0.58 0.21 0.18 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Estudo das relações entre atributos




mpg
weight
45
40
35
30
25
20
15
10
1500
2000
2500
3000
3500
4000
4500
5000


mpg
cylinders
displacement
weight
acceleration
model year
origin
1 -0.78 -0.8 -0.83 0.42 0.58 0.56
-0.78 1 0.95 0.9 -0.51 -0.35 -0.56
-0.8 0.95 1 0.93 -0.54 -0.37 -0.61
-0.83 0.9 0.93 1 -0.42 -0.31 -0.58
0.42 -0.51 -0.54 -0.42 1 0.29 0.21
0.58 -0.35 -0.37 -0.31 0.29 1 0.18
0.56 -0.56 -0.61 -0.58 0.21 0.18 1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Arquitetura de Dados

Os principais objetivos da arquitetura de dados são:

1. Extração de características/atributos

2. Preparar o conjunto de dados de entrada adequado, compatível com os requisitos do algoritmo de aprendizado de máquina.

3. Melhorar o desempenho dos modelos de aprendizado de máquina.

Sobre a extração de características...

- A inspiração para obter uma ou várias característica que permita obter o padrão para o desenvolvimento de um modelo, está intrinsecamente relacionada a capacidade do analista de olhar a amostra e conhecer o problema para inferir a medida.

No colab

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sn

csv = pd.read_csv('/content/auto-mpg.csv')
colunas = list( csv.head() )
colunas = colunas[0:-1]
nomeColunas = sorted( colunas )
numColunas = len(nomeColunas)
fig, ax = plt.subplots( numColunas, numColunas, figsize=(15, 15), \
constrained_layout=False )
for lin in range(0,numColunas):
  for col in range(0,numColunas):
    ax[lin,col].scatter(x=csv[nomeColunas[col]],y=csv[nomeColunas[lin]],s=4 )
    ax[lin,col].set(xticks=[], yticks=[],xlabel=nomeColunas[col], \
    ylabel=nomeColunas[lin])
plt.show()
```

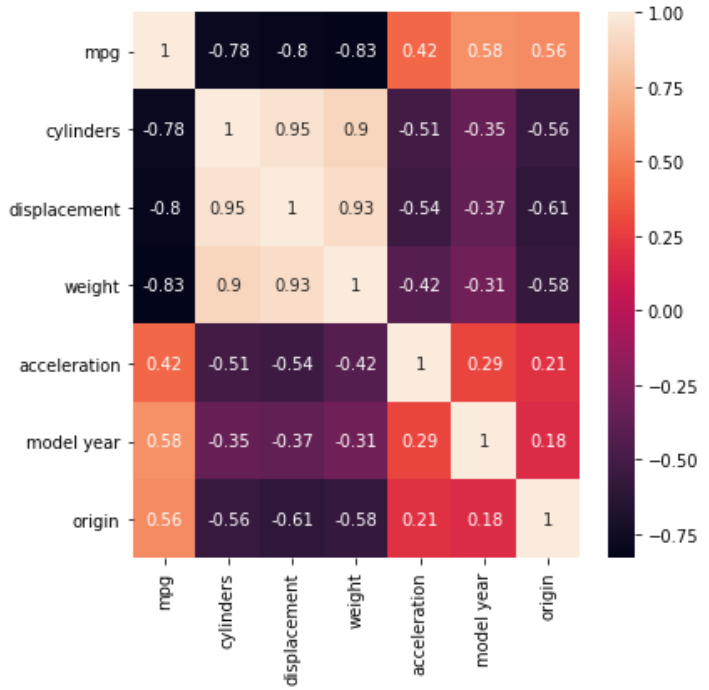
No colab

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sn

pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', -1)
corr = csv.corr()
print( corr )
plt.subplots( figsize=(6, 6) )
sn.heatmap(corr, annot=True)
plt.show()
```

```
                    mpg  cylinders  displacement
mpg            1.000000 -0.775396  -0.804203
cylinders     -0.775396  1.000000   0.950721
displacement  -0.804203  0.950721   1.000000
weight        -0.831741  0.896017   0.932824
acceleration   0.420289 -0.505419  -0.543684
model year     0.579267 -0.348746  -0.370164
origin         0.563450 -0.562543  -0.609409

              model year    origin
mpg             0.579267  0.563450
cylinders      -0.348746 -0.562543
```



Referências

- PATTERN RECOGNITION APPLIED TO FORMATTED INPUT OF HANDWRITTEN DIGITS
  https://www.sciencedirect.com/science/article/pii/S1474667017509325
- Exemplo do MNIST
  https://github.com/mbornet-hl/MNIST
  https://www.kaggle.com/c/digit-recognizer
- Informação mutua
  https://www.kaggle.com/ryanholbrook/mutual-information
- Informação e Conhecimento (Cortella)
  https://www.youtube.com/watch?v=fuGAcDLmHZU
  Base de Carros
  https://www.rpubs.com/dksmith01/cars