Capítulo 1

Introdução

1.1 A Ascendência dos Dados

Vivemos em um mundo que está se afogando em dados. Os sites rastreiam o clique de cada usuário. Seu *smartphone* está construindo um registro de sua localização e velocidade a cada segundo diariamente. A turma *fitness* das redes sociais dispõem de seus *gadgets* que estão sempre registrando seus batimentos cardíacos, hábitos de movimento, dieta, padrões de sono, etc. Carros inteligentes coletam hábitos de direção. Casas inteligentes coletam hábitos de vida. Profissionais de marketing coletam hábitos de compra... A própria internet representa um colossal gráfico de conhecimento que contém (entre muitas coisas) uma enorme enciclopédia com referências cruzadas: bancos de dados específicos sobre filmes, músicas, esportes, finanças, *videogames*, *memes*, estatísticas governamentais e por aí vai...

Figura 1.1: Big data e suas tecnologias no cotidiano.

Enterrados nesses amontoados de dados estão as respostas para inúmeras perguntas que ninguém jamais pensou em fazer. Ao longo deste curso, aprenderemos como encontrar algumas delas!

1.2 O que é Ciência de Dados?

Ciência de dados (em inglês: *data science*) é uma área interdisciplinar, localizada entre a estatística e a ciência da computação, que se utiliza de processos, algoritmos e sistemas, para extrair conhecimento e tomar decisões a partir de dados dos diversos tipos, sendo eles ruidosos, nebulosos, estruturados ou não-estruturados.

É uma área voltada para o estudo e a análise organizada de dados científicos e mercadológicos, financeiros, sociais, geográficos, históricos, biológicos, psicológicos, dentre muitos outros. Visa, deste modo, a extração de conhecimento, detecção de padrões e/ou obtenção de *insights*[1] para possíveis tomadas de decisão. A ciência de dados enquanto área de estudo existe há 30 anos, porém ganhou mais destaque nos últimos anos devido a alguns fatores como a popularização de várias tecnologias digitais como o *big data*, a internet das coisas (IoT) e a inteligência artificial (AI).

> 💬 Resumindo: cientista de dados é alguém que extrai *insights* de um amontoado de números confusos e/ou bagunçados.

Cientistas de Dados podem trabalhar no setor privado, por exemplo, transformando grande quantidade de dados brutos em *insights* nos negócios, auxiliando empresas em tomadas de decisões para atingir melhores resultados ou na área acadêmica e governamental como pesquisadores interdisciplinares. Podemos listar algumas aplicações reais:

[1] *Insight* é a compreensão de correlações específicas dentro de um contexto particular.

Figura 1.2: Exemplos de aplicações de Ciência de Dados.

Alguns cientistas de dados ocasionalmente usam suas habilidades para o bem - usando dados para tornar o governo mais eficaz, ajudar sem-tetos e melhorar a saúde pública. Mas certamente não prejudicará sua carreira se você gosta de descobrir a melhor maneira de fazer as pessoas clicarem em anúncios e impactar um negócio de forma positiva.

1.3 Motivação Hipotética: Facedata

Parabéns! Você acaba de ser contratado para liderar o time de ciência de dados na *Facedata*, a rede social para cientistas de dados.

Figura 1.3: A logo da hipotética rede social *Facedata*.

Apesar de ter como foco os cientistas de dados, a *Facedata* nunca investiu na construção de sua própria prática de ciência de dados. Esse será o seu trabalho! Ao longo do curso, aprenderemos sobre alguns conceitos de ciência de dados resolvendo problemas que você encontrará no seu dia-a-dia laboral. Às vezes, analisamos dados fornecidos explicitamente pelos usuários, às vezes, dados gerados por meio de suas interações com o site e, às vezes, até dados de experimentos que projetamos.

Ao longo dessa jornada, construiremos nossas ferramentas do zero. Ao fim, você terá uma compreensão sólida dos fundamentos da ciência de dados. Daí, você estará pronto para aplicar suas habilidades em uma empresa, em um órgão governamental ou para qualquer outro problema que lhe interesse.

Bem-vindo(a) e boa sorte!

É seu primeiro dia de trabalho na *Facedata*. Temos inúmeras perguntas para responder sobre os usuários dessa rede, como podemos gerar informação através de seus dados e muito mais... No entanto, vamos seguindo como na vida real: ao longo do curso, nosso aprendizado será por meio de tarefas que (supostamente) surgiriam no cotidiano laboral da *Facedata*.

⚠ Neste Capítulo iremos trabalhar somente com raciocínio lógico. Ao longo do curso, você irá começar a desenvolver habilidades para otimizar as ideias aqui representadas.

1.3.1 Mapeamento da Rede

Tarefa 1 *Visão geral da rede.*

- *quem está conectado a quem?*
- *quem são os usuários mais conectados?*

Para realizar esta tarefa, os dados sobre usuários e de suas conexões nos são fornecidos – Na vida real, as pessoas não entregam os dados de que você precisa e será necessários 'caçá-los'. A base de dados foi separada em 2 blocos:

(1) Uma lista contendo o *id* de usuário (que é um número) e o nome (que, por pura coincidência, rima com seu *id*):

```
"id"   "nome"
  0     Nero
  1     Atum
  2     Bois
  3     Alvares
  4     Teatro
  5     Zinco
  6     Ameis
  7     Bete
  8     Biscoito
  9     Love
```

(2) Uma tabela demonstrando as conexões dos usuários. Se há conexão entre usuários, teremos *1*, caso contrário, *0*:

	ID_0	ID_1	ID_2	ID_3	ID_4	ID_5	ID_6	ID_7	ID_8	ID_9
ID_0	0	1	1	0	0	0	0	0	0	0
ID_1	1	0	1	1	0	0	0	0	0	0
ID_2	1	1	0	1	0	0	0	0	0	0
ID_3	0	1	1	0	1	0	0	0	0	0
ID_4	0	0	0	1	0	1	0	0	0	0
ID_5	0	0	0	0	1	0	1	1	0	0
ID_6	0	0	0	0	0	1	0	0	1	0
ID_7	0	0	0	0	0	1	0	0	1	0
ID_8	0	0	0	0	0	0	1	1	0	1
ID_9	0	0	0	0	0	0	0	0	1	0

Olhar para uma tabela preenchida por *0*'s e *1*'s não é nada agradável e dificulta a compreensão. Além disso, sempre pense no contexto de *big data*: imagine essa mesma tabela com mais de 100 mil usuários! Seria inviável tanto computacionalmente (precisaríamos de um super computador para lidar com ela) quanto intelectualmente (dificilmente conseguiríamos enxergar 100 mil linhas e 100 mil colunas ao mesmo tempo). Assim, a Figura 1.4 poderia representar de forma muito mais agradável:

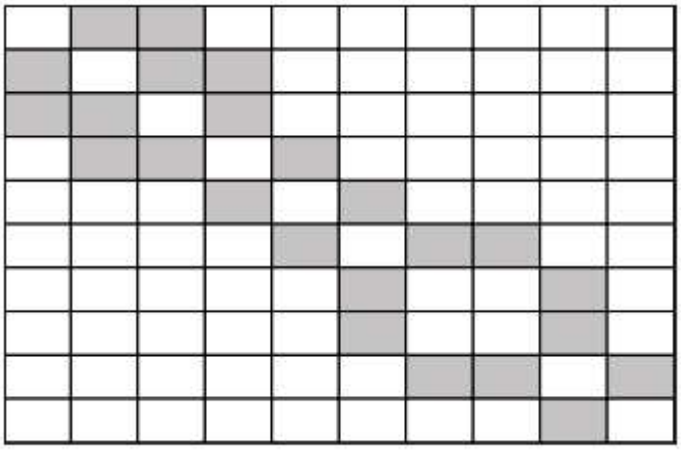

Figura 1.4: Exemplos de aplicações de ciência de dados.

Por exemplo, a afirmação na primeira linha e segunda coluna indica que os usuários com `id 0` (*Nero*) e com `id 1` (*Atum*) são amigos. Ainda assim, note que existe uma simetria na Figura 1.4. A partir da diagonal principal, temos uma cópia acima e abaixo dela. Como a nossa rede é *esparsa*[2], podemos simplificar toda essa informação a partir de uma lista de pares dos `ids`:

```
"pares_amizade"
 (0, 1) (0, 2)
 (1, 2) (1, 3)
 (2, 3)
 (3, 4)
 (4, 5)
 (5, 6) (5, 7)
 (6, 8)
 (7, 8)
 (8, 9)
```

Tendo a mesma interpretação anterior mas computacionalmente simplificada: `id 0` é amigo de `id 1`; `id 0` é amigo de `id 2`; `id 1` é amigo de `id 2` e assim sucessivamente... A rede é melhor ilustrada na Figura 1.5.

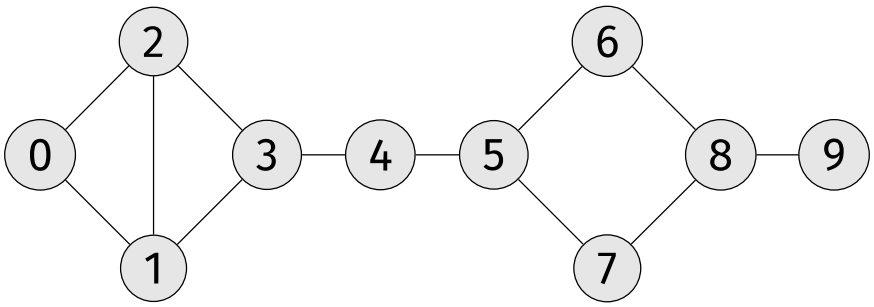

Figura 1.5: A rede *Facedata*.

Ter amizades representadas como uma lista de pares não é a maneira mais fácil de trabalhar com elas. Para encontrar todas as amizades para o usuário com `id 0`, você deve *iterar* (ou seja, fazer *loops*) sobre cada par procurando pares contendo `0`. Se você houvesse muitos pares, isso levaria muito tempo!

[2] não compacta, espalhada aqui e ali (diz-se geralmente de algo contável e pouco numeroso); dispersa, escassa, rala.

> 💬 Tenha como princípio o equilíbrio entre a praticidade computacional e a interpretação pelos humanos.

Tarefa 2 *Número de conexões.*

- *qual é o número médio de conexões?*
- *quem são os usuários menos/mais conectados?*

Primeiro, encontramos o número total de conexões somando cada linha da tabela (ou contando quantas vezes aparece determinado *id* nos *pares_amizade*):

```
     ID_0  ID_1  ID_2  ID_3  ID_4  ID_5  ID_6  ID_7  ID_8  ID_9    SOMA
ID_0    0     1     1     0     0     0     0     0     0     0       2
ID_1    1     0     1     1     0     0     0     0     0     0       3
ID_2    1     1     0     1     0     0     0     0     0     0       3
ID_3    0     1     1     0     1     0     0     0     0     0       3
ID_4    0     0     0     1     0     1     0     0     0     0       2
ID_5    0     0     0     0     1     0     1     1     0     0       3
ID_6    0     0     0     0     0     1     0     0     1     0       2
ID_7    0     0     0     0     0     1     0     0     1     0       2
ID_8    0     0     0     0     0     0     1     1     0     1       3
ID_9    0     0     0     0     0     0     0     0     1     0       1

SOMA    2     3     3     3     2     3     2     2     3     1      24


MEDIA 24 / 10 = 2.4
```

Agora, podemos destacar:

- Número médio de conexões: 2.4 (o que isso significa?).
- Usuários menos conectados: *id* 9.
- Usuários mais conectados: *id*s *1*, *2*, *3*, *5* e *8*.

Será que poderíamos incentivar mais conexões entre os usuários?

1.3.2 Sistema de Recomendação

Tarefa 3 *Sugestão de amigos.*

- *como sugerir amigos aos usuários?*

Uma primeira ideia é sugerir que os usuários possam conhecer os amigos de seus amigos.

> ⚠ Chegamos em um grande impasse: daria um enorme trabalho fazer isso de forma manual e precisaríamos automatizar de alguma forma.

Então, vamos para uma segunda opção que seria encontrar usuários com interesses semelhantes. Durante o cadastro na rede social, cada usuário lista 2 habilidades em que tem interesse:

```
"id"  "nome"      "habilidades1"        "habilidades2"
  0    Nero       big data              python
  1    Atum       python                marketing
  2    Bois       python                analise de dados
  3    Alvares    estatistica           aprendizado de maquina
  4    Teatro     analise descritiva    classificacao
  5    Zinco      python                big data
  6    Ameis      pandas                economia
  7    Bete       big data              python
  8    Biscoito   python                analise de dados
  9    Love       analise descritiva    aprendizado de maquina
```

Agora, podemos calcular quais são as habilidades mais ou menos comuns:

```
analise de dados        2
analise descritiva      2
aprendizado de maquina  2
big data                3
classificacao           1
economia                1
estatistica             1
marketing               1
pandas                  1
python                  6
```

Vemos que *Python* é a habilidade mais comum ao usuários. Curiosamente, é essa linguagem que iremos aprender nos próximos Capítulos. Aliado com a informação dos tópicos, ainda precisaríamos saber quem tem os mesmos interesses e ainda não são amigos. Postergaremos essa tarefa por enquanto.

1.3.3 Salários e Experiência

Tarefa 4 *Progressão salarial.*

- *os salários acompanham a experiência?*
- *quanto maior a experiência, maior o salário?*

Experiência é tudo mas o dinheiro tem a sua importância. A informação salarial (em reais) e da experiência (em anos) no mercado de trabalho são gentilmente cedidas pelos usuários de nossa rede social:

```
"id"  "nome"   "experiencia" "salario"
  0   Nero              2.5       7500
  1   Atum              4.2       7875
  2   Bois              8.1      11000
  3   Alvares           7.5       9500
  4   Teatro            1.9       6000
  5   Zinco             0.7       6000
  6   Ameis             6.5       8625
  7   Bete             10        10375
  8   Biscoito          6         9500
  9   Love              8.7      10375
```

O primeiro passo é visualizar os dados. Você pode ver os resultados na Figura 1.6.

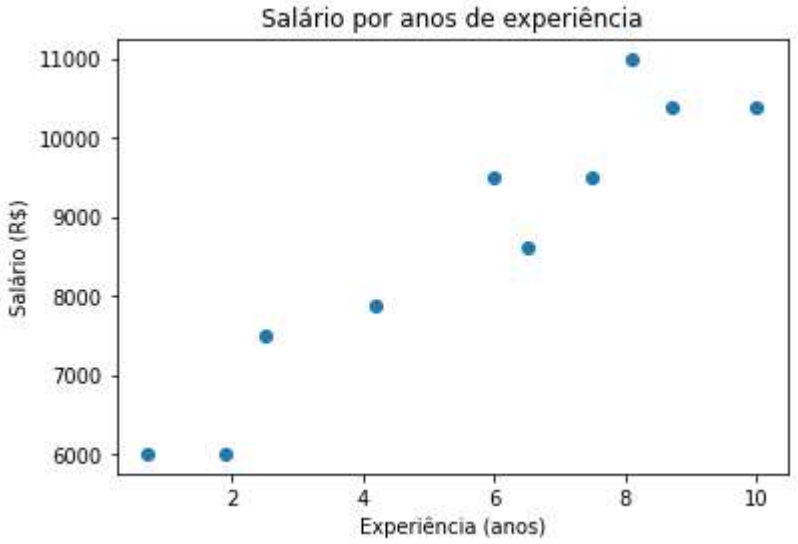

Figura 1.6: Salário por anos de experiência.

Parece claro que as pessoas com mais experiência tendem a ganhar mais. Como você pode transformar isso em um fato interessante? Vamos reagrupar as experiências pelas seguintes faixas:

```
"id"  "nome"  "experiencia" "salario"       "faixa experiencia"
  0      Nero           2.5      7500         entre 2 e 5 anos
  1      Atum           4.2      7875         entre 2 e 5 anos
  2      Bois           8.1     11000          mais que 5 anos
  3      Alvares        7.5      9500          mais que 5 anos
  4      Teatro         1.9      6000        menos que 2 anos
  5      Zinco          0.7      6000        menos que 2 anos
  6      Ameis          6.5      8625          mais que 5 anos
  7      Bete          10       10375          mais que 5 anos
  8      Biscoito       6        9500          mais que 5 anos
  9      Love           8.7     10375          mais que 5 anos
```

E, finalmente, calcular o salário médio para cada grupo:

```
"faixa experiencia"  "salario medio"
  menos que 2 anos          6000.00
  entre 2 e 5 anos          7687.50
   mais que 5 anos          9895.83
```

E você tem sua frase de efeito: *'Cientistas de dados com mais de cinco anos de experiência ganham 65% a mais do que cientistas de dados com pouca ou nenhuma experiência!'*

No entanto, escolhemos as faixas salariais de maneira arbitrária. O que realmente gostaríamos é de fazer uma afirmação sobre o impacto (em média) de ter mais um ano de experiência no salário. Além de ser um fato ainda mais interessante, isso nos permite fazer previsões sobre salários que ainda não conhecemos.

1.3.4 Usuários Premium

A Facedata dispõe de 2 tipos de contas: gratuita e paga. Os usuários que pagam pelo seu perfil são chamados de '*premium*' e têm acesso ilimitado à conteúdo extra.

Tarefa 5 *Captação de novos usuários premium.*

- *quais são as características dos usuários premium?*
- *como predizer um potencial usuário pagante?*

Vamos iniciar observando uma relação existente entre anos de experiência e contas pagas:

```
"experiencia"     "conta"
          0.7       gratuita
          1.9       gratuita
          2.5       gratuita
          4.2       gratuita
          6.0       paga
          6.5       gratuita
          7.5       paga
          8.1       gratuita
          8.7       paga
         10.0       paga
```

Usuários com mais experiência tendem a pagar pela conta *premium*. Com essa informação, você cria um modelo que prediz o tipo de conta da seguinte forma:

- 'gratuita' para usuários com experiência maior ou igual à 6 anos;
- 'paga' para os demais casos:

```
"experiencia"     "conta"      "previsao conta"
       0.7        gratuita          gratuita
       1.9        gratuita          gratuita
       2.5        gratuita          gratuita
       4.2        gratuita          gratuita
       6.0        paga              paga
       6.5        gratuita          paga
       7.5        paga              paga
       8.1        gratuita          paga
       8.7        paga              paga
      10.0        paga              paga
```

Apesar de muito intuitivo, sabemos que em uma grande base de dados, poderíamos errar bastante com este tipo de previsão e esse modelo de previsão não é o melhor a ser feito. Com mais dados (e mais matemática), poderíamos construir um modelo que prevê a probabilidade de um usuário pagar por uma conta com base em seus anos de experiência. Investigaremos esse tipo de problema (e muitos outros) adiante no curso.

1.3.5 O que vem Adiante...

Foi um primeiro dia de sucesso! Parabéns!

Fique tranquilo(a), tudo irá passar a fazer mais sentido e ficar cada vez mais interessante nos próximos capítulos!