Análise de fatores de sucesso de pequenos negócios (Startups)

Analogia com análise de dados de assuntos abstratos/intangíveis

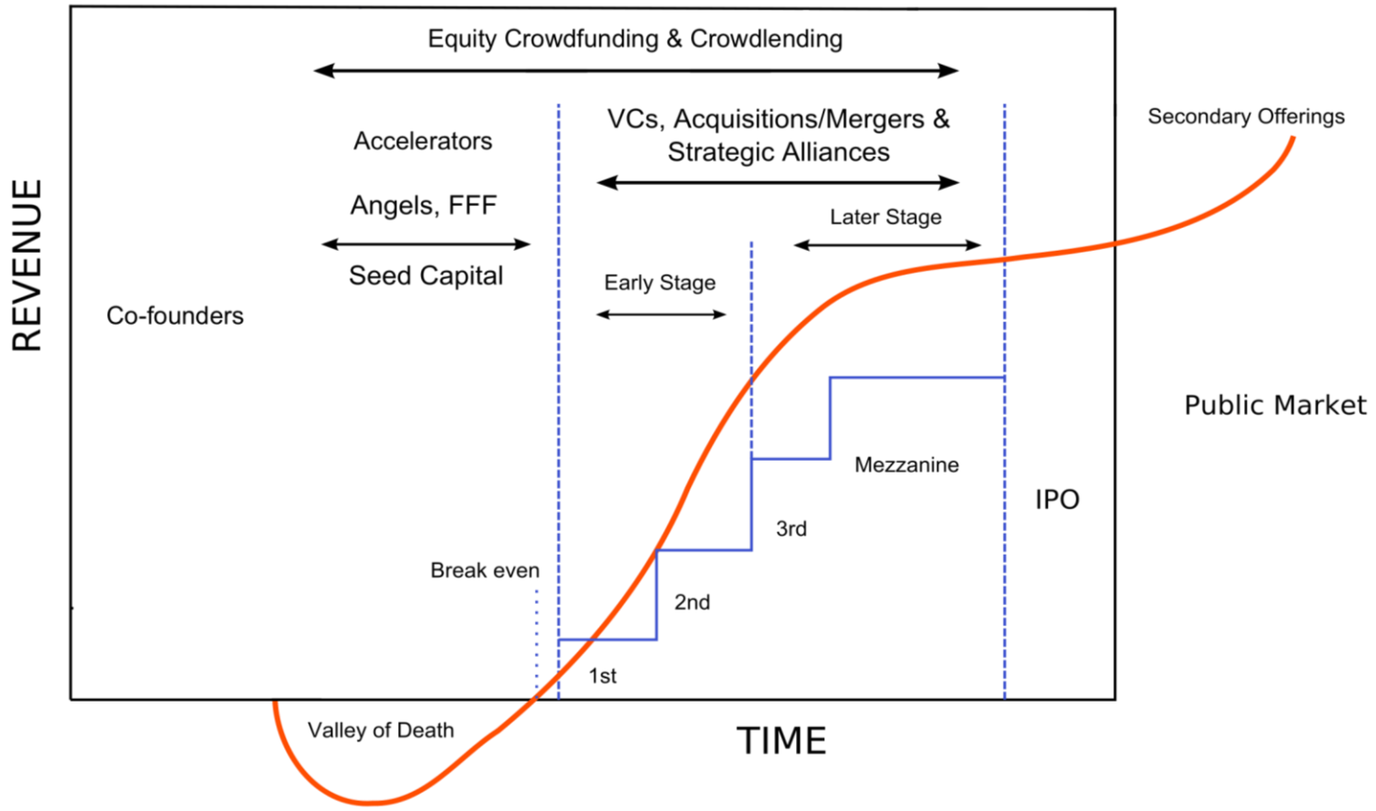


Imagem: https://commons.wikimedia.org/wiki/File:Startup_Financing_Cycle.png
Valley, N. (2018). Startup life cycle.

Professor: Alex Pereira

Apresentação Pessoal

2000 a 2004 – Graduação em Engenharia de Computação no ITA

2005 a 2008 – Mestrado em Eng. de Computação e Eletrônica no ITA

2009 a 2015 – Doutorado em Eng. de Computação e Eletrônica no ITA

2004 a 2010 – Empreendedor, sócio em empresa de base tecnológica

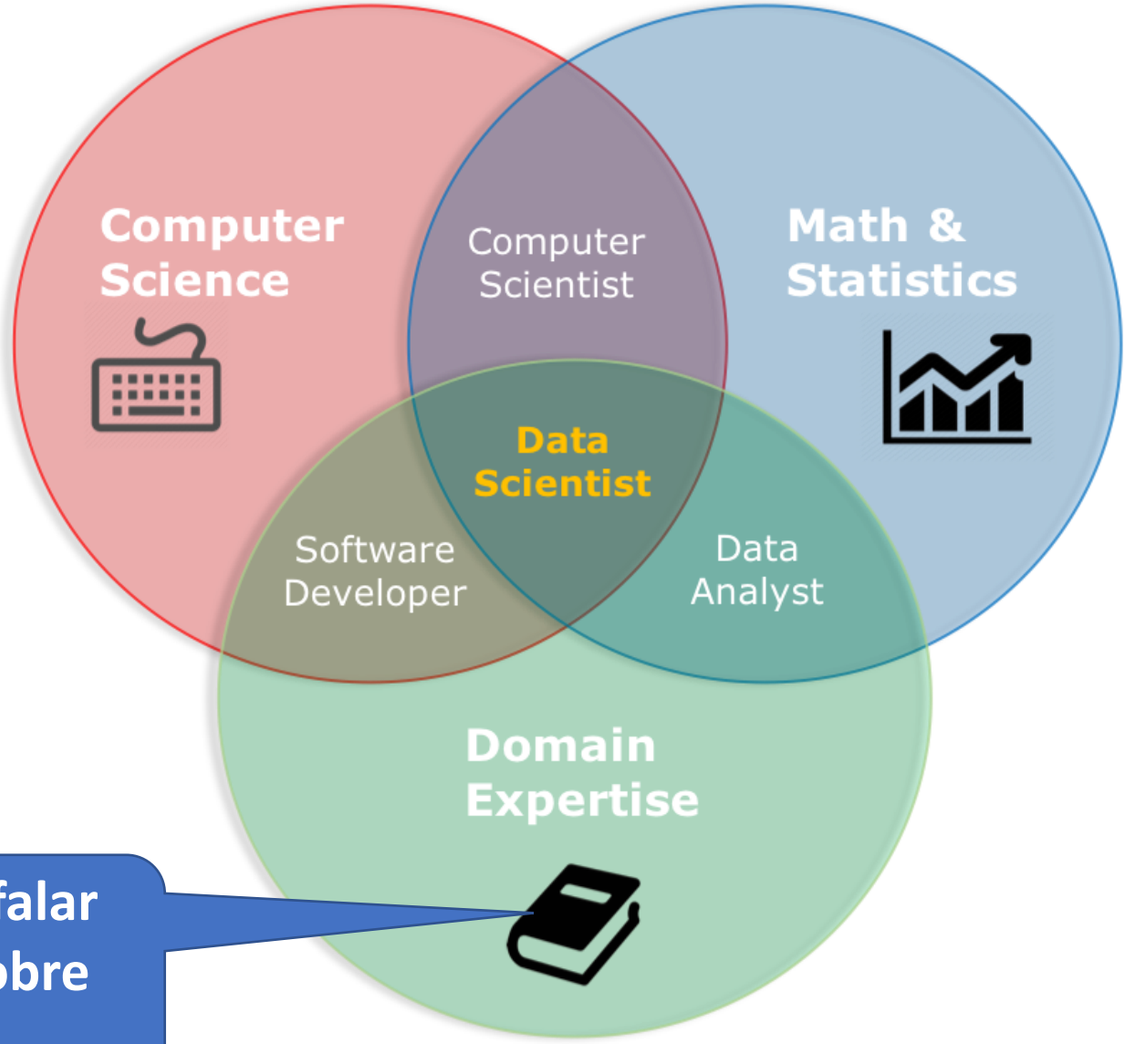
2013 a 2017 – Censipam / Ministério da Defesa

2016 / 2018 – Professor

2017 – Ministério da Economia

2020 – Fintech Empréstimo P2P

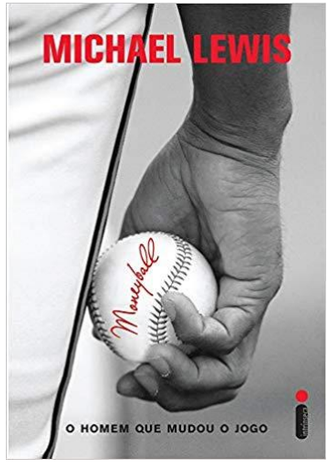
Conceituação de Ciência de Dados



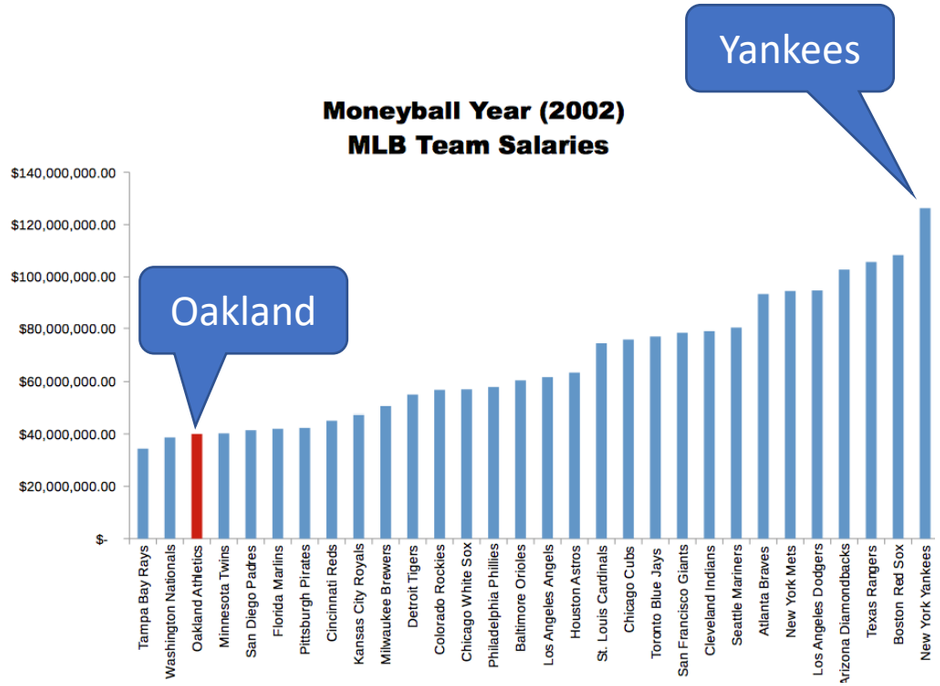
Hoje vamos falar um pouco sobre Startups

Professor: Alex Pereira

Moneyball: The Art of Winning an Unfair Game



- Billy Beane: Jovem promessa que não se viabilizou.
  - Criou método para o recrutamento de jogadores;
  - Contra tradição baseada na intuição dos olheiros,
    - ✓ Baseou seu método em análise estatística
      - para avaliar o potencial dos novos jogadores.
- Unfair game
  - Recursos financeiros escassos,
  - Metas desafiadoras
    - ✓ impossíveis sob o mesmo paradigma
- Solução
  - Inovação
    - ✓ Excelente desempenho e quebra de recorde



Experts VS Algorithms: https://towardsdatascience.com/experts-vs-algorithms-a-view-ea0951211f8f

Medir -> Analisar -> Melhorar -> Medir ...

- The Data Coach (o treinador embasado em dados)
  - Podcast do Michael Lewis
    - ✓ https://atrpodcast.com/episodes/the-data-coach-s1!68200
- High Speed Camera (Edgertronic.com)
  - "We provide the clarity and insights to tell de difference between expectations and what is really going on".

Parametrizando o Desempenho

- A câmera mede a velocidade do braço do lançador
  - Essa métrica tem grande correlação com a velocidade da bola
    - ✓ quando a velocidade do braço é alta, comparada com a de outros atletas
    - ✓ mas a velocidade da bola é baixa,
    - ✓ há uma ineficiência mecânica que pode ser investigada e corrigida

Gong.io: NLP aplicado a melhoria do desempenho de vendedores

- Coletam as ligações telefônicas dos vendedores de uma empresa
  - E os seus respectivos desempenhos
- Parametrizam os fatores de sucesso e fracasso nas vendas
  - em termos de uso da linguagem
    - ✓ Buscando por padrões em ligações bem sucedidas ou mal sucedidas

Métricas analisadas pelo Gong: escala de A a F

- Porcentagem de tempo de fala do vendedor
  - 46% é o ideal
- 13 questões é a quantidade ótima
  - Muito mais perguntas as pessoas ficam impacientes;
    - ✓ Pare, pense e responda.
  - Muito menos signifca que você está falando muito.
- Seu maior monólogo
  - Monólogo do cliente é um bom sinal
    - ✓ do vendedor, não.
- Interrupção do interlocutor
- Presença Filler Words
  - não tem correlação com o sucesso/fracasso de uma ligação
    - ✓ eh, ah, hum, like, you know, I mean, okay, so, actually, basically, and right

Objection Handeling: https://www.youtube.com/watch?v=DBpLFAS2s3w

Conclusão

Fatorar (quebrar em partes menores) o seu problema a partir de dados, pode:

1) aumentar a sua consciência sobre o real problema; e
2) Gerar insights de melhoria.

Algoritmos vs Experts – Discussão acadêmica desde 1954

- Famosa monografia de Paul Meehl em 1954
  - Comparou a acurácia de previsões feitas por
    - ✓ juízes humanos (psicólogos clínicos) e
    - ✓ as previsões realizadas por modelos estatísticos simples.
  - Os modelos usaram um subconjunto das informações disponíveis aos psicólogos
  - Ainda assim foram mais acurados em quase todos os casos.
- Meehl acreditava que a inferioridade do julgamento clínico devia-se em parte a erros sistemáticos,
  - tais como negligenciar consistentemente a taxa-base (base rate)

Condições para a boa qualidade de um julgamento intuitivo (Daniel Kahneman)

- Condição 1:
  - Previsibilidade/Validade do ambiente no qual o julgamento é realizado.

- Condição 2:
  - Oportunidade para os indivíduos aprenderem as regularidades deste ambiente.

Fonte: Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526. https://doi.org/10.1037/a0016755

Ocupações com bom e mal desempenho nas intuições

- Shanteau (1992) encontrou expertise em
  - livestock judges (juízes de gado), astrônomos, pilotos de teste, juízes de solo, mestres de xadrez, médicos, matemáticos, contadores, inspetores de grãos, foto intérpretes e analistas de seguro.

- E encontrou baixo desempenho em profissionais experientes
  - investidores da bolsa de valores, psicólogos clínicos, psiquiatras, oficiais de admissão de escolas, juízes de corte, selecionadores de pessoal e analistas de inteligência.

Fonte: Shanteau, J. (1992). Competence in experts: The role of task characteristics. Organizational Behavior and Human Decision Processes, 53,252–26

Porque a intuição pode falhar

- Sobrecarga cognitiva (*Cognitive overload*)
  - A ciência produziu muito mais informação do que uma pessoa consegue processar a cada decisão
    - ✓ Resultado: usamos atalhos (heurísticas) sujeitas a vieses

- Vieses cognitivos (erros sistemáticos). Por exemplo,
  - Excesso de confiança,
  - Viés de confirmação,
  - Ignorar base rates (probabilidade a priori).

Apgar Score (algoritmo) da Dra. Virginia Apgar – Reduziu significativamente os erros humanos

- Avaliação imediata do recém-nascido
  - no 1o e no 5o minutos de vida;
  - 5 critérios pontuados de 0 a 2 e somados
    - ✓ Cor da pele;
    - ✓ Pulsação arterial;
    - ✓ Irritabilidade Reflexa (caretas);
    - ✓ Atividade (tônus muscular);
    - ✓ Esforço respiratório;
- Critérios objetivos e simples de serem avaliados por não experts
  - Dividir um problema em partes menos e menos complexas.
- Interpretação dos resultados
  - de 0 a 3 – Asfixia grave
  - de 4 a 6 – Asfixia moderada
  - de 7 a 10 – Boa vitalidade, boa adaptação

O checklist segundo Atul Gawande

- Atul Gawande
  - Escritor e Professor da faculdade de medicina de Harvard
  - Membro do Conselho Consultivo COVID-19 de Joe Biden.
    - ✓ Esp. em reduzir erros e melhorar a eficiência dos procedimentos cirúrgicos.
- Possíveis causas dos erros humanos
  - Ignorância - podemos errar porque a ciência nos deu apenas uma compreensão parcial do mundo e de como ele funciona.
  - Inépcia (estupidez, desleixo) - o conhecimento existe, mas deixamos de aplicá-lo corretamente.
    - ✓ "Pela primeira vez na história os erros por inépcia são mais frequentes do que os erros por ignorância" (Atul Gawande)
    - ✓ Tem uma carga emocional - não perdoamos erros por inépcia
- Os checklists (algoritmos) podem reduzir os erros por inépcia
  - causados por vieses cognitivos e pela sobrecarga cognitiva.

Livro - The Checklist Manifesto: How To Get Things Right

- Atul Gawande (2009)
  - Inspirado no checklist de hospital (caso da menina afogada)
    - ✓ e nos checklists da aviação
  - Desenvolveu uma estratégia para evitar os erros por inépcia
    - ✓ Pegar o know-how duramente conquistado por pessoas altamente treinadas e habilidosas
      - Eliminando falhas comuns, persistentes e evitáveis
  - Ou seja, prescrevendo um checklist construído por experts
    - ✓ que leve em conta nossas falhas de julgamentos
- Generalização do checklist da Dra. Virginia Apgar
- Não se coloca todas as etapas num checklist. Exemplos a checar
  - Os membros de uma equipe de cirurgia se apresentarem;
  - Há bolsa de sangue disponível em caso de hemorragia...

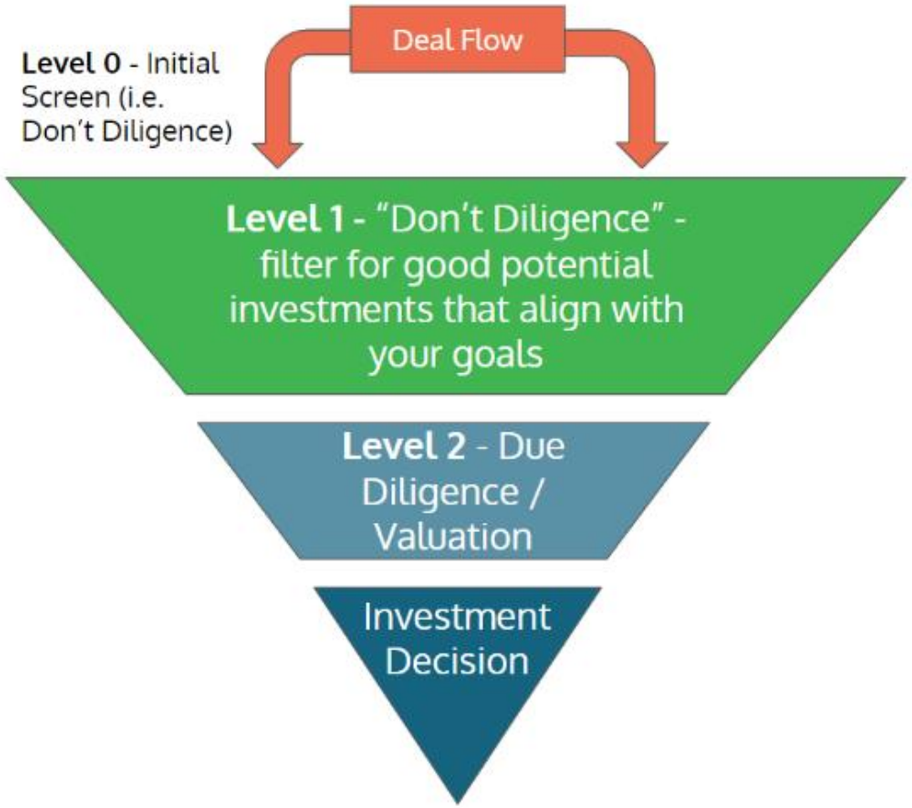
Checklists de outras áreas

- Checklist de processo seletivo;
  - por Daniel Kahneman;
  - por Alex Pereira.
- Checklist de resolução de questões de prova;
  - por Alex Pereira.
- Venture Capital (capital de risco) e investimento em Startups.
  - por Atul Gawande;
  - pela organização Crowdwise,
    - ✓ e outros investidores de VC

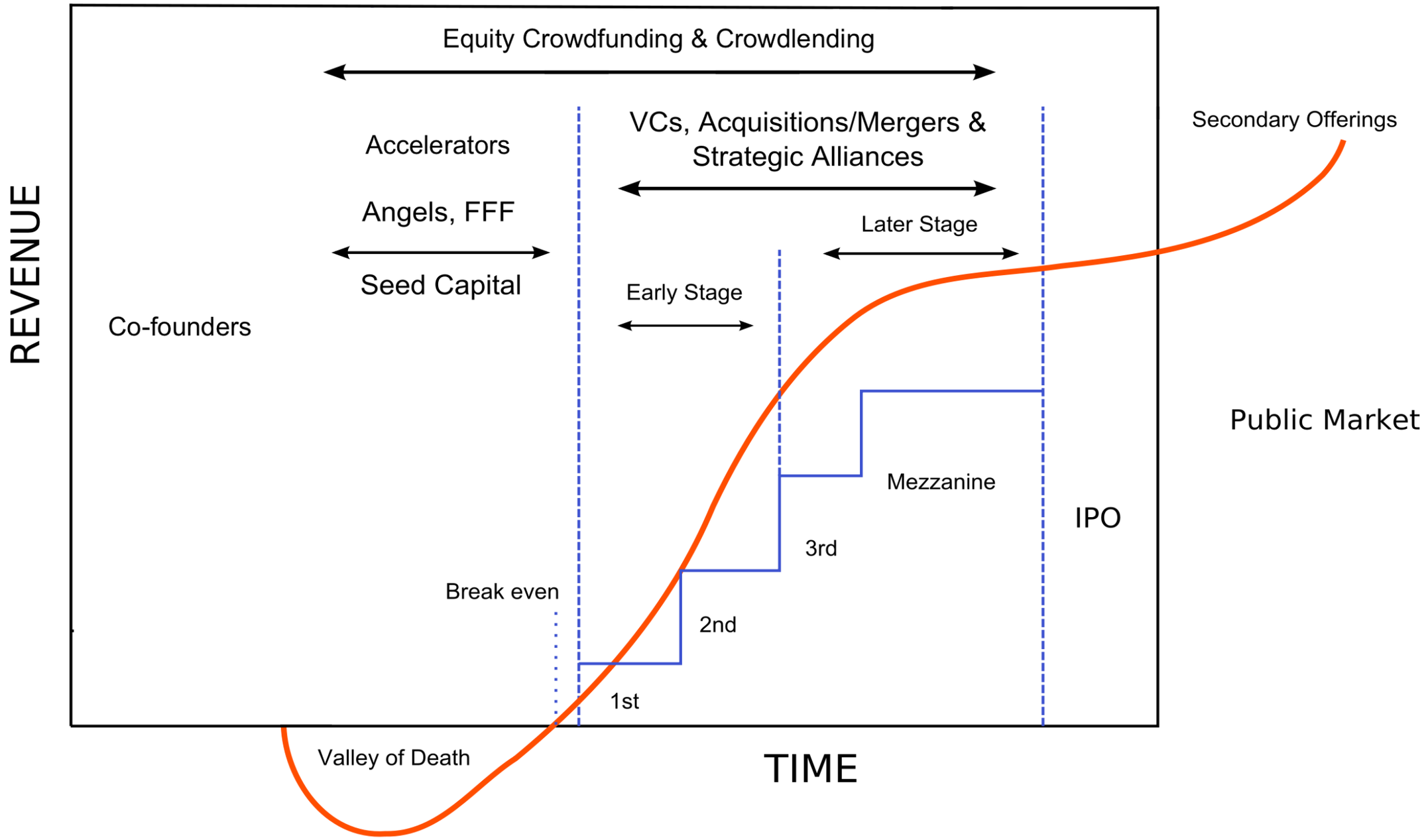
Checklist de Venture Capital (VC)

- Crowdwise
  - https://crowdwise.org/books/gut-vs-checklists-which-is-better-for-startup-investment-decisions/

- Embasados por dados, vamos adicionar
  - itens ao Checklist de VC ?



Ciclo de Investimento das Startups



Ciclo de Vida do Financiamento

- Pre-seed
  - ocorre nos estágios iniciais do ciclo de vida de uma startup,
  - o dinheiro é normalmente coletado por meio de FFFs
    - ✓ FFF = Family, Friends and Fools
- Seed
  - quando a startup tem um MVP, ela busca financiamento com investidores com maior potencial financeiro
    - ✓ aceleradores, anjos (angels) mais ricos e Venture Capital (VC) em estágio inicial
- Series
  - Assim que a startup começa a crescer, ela levanta dinheiro em uma série de rodadas de investimento (Series A até Series G).
    - ✓ A primeira rodada Serie A e geralmente é a primeira rodada significativa de financiamento de capital de risco.

Fonte: Startup Success Prediction in the Dutch Startup Ecosystem.
https://repository.tudelft.nl/islandora/object/uuid%3A1adc2972-db09-4583-b2da-05fd4e462941

Exemplos de métricas de sucesso de Startups

- Quantidade de Empregados
  - Total Full Employment – TFE
- Rodada de Financiamento
  - Seed, Series A, Series B, ...
- Lucratividade
- Total de Financiamento
- Crescimento sustentável
  - Scaleup: uma empresa que tem um retorno médio anualizado de pelo menos 20% nos últimos 3 anos
    - ✓ com pelo menos 10 funcionários no início do período

Fonte: Startup Success Prediction in the Dutch Startup Ecosystem.
https://repository.tudelft.nl/islandora/object/uuid%3A1adc2972-db09-4583-b2da-05fd4e462941

Pandas – Biblioteca Python para Ciência de Dados

- Adota o estilo idiomático de computação baseada em arrays
  - do NumPy
    - ✓ Preferência por processar dados sem loops
- NumPy armazena dados homogêneos
  - Pandas foi projetado para trabalhar com dados tabulares e heterogêneos
- Se tornou open source em 2010
  - conta com mais de 800 colaboradores

Fontes de Dados

	Valores Ausentes
• dealroom.co	16%
• crunchbase.com ○ webscrapy	57%
• mattermark.com	Não avaliado

Python/Pandas

```
shape_dfd = dfd.shape
shape_dfc = dfc.shape
pct_null_dfd = 100*np.sum(dfd.isnull().sum())/(shape_dfd[0]*shape_dfd[1])
pct_null_dfc = 100*np.sum(dfc.isnull().sum())/(shape_dfc[0]*shape_dfc[1])
```

Selenium

- Selenium automates browsers. That's it!
  - What you do with that power is entirely up to you.
- Selenium automatiza os Navegadores
  - O que você faz com esse poder é por sua conta.
- Qualquer interação do usuário com o Navegador
  - pode ser automatizada pelo Selenium. Exemplos:
    - ✓ Cliques;
    - ✓ Digitar texto;
    - ✓ Minimizar/Fechar janela;
    - ✓ Mouse over (passar o mouse sobre um elemento html)

História do Selenium: https://www.selenium.dev/history/

Webscrapy com Selenium

- Como fazer o Selenium abrir uma página no seu navegador

```
driver = webdriver.Firefox()
url = 'http://books.toscrape.com/'
driver.get(url)
print(driver)
```

- Dependendo do sistema operacional, você precisará especificar
  - o path do driver com o argumento executable_path
    - ✓ seja do driver do Firefox ou do Chrome

Webscrapy

- Coleta em larga escala de dados contidos em páginas web
- Legalidade
  - tem sido bastante discutida nos EUA no caso
    - ✓ https://en.wikipedia.org/wiki/HiQ_Labs_v._LinkedIn
  - A primeira decisão da suprema corte americana foi
    - ✓ Não criminalizar o acesso a dados não protegidos por login e senha

- Tenha
  - Bom senso, cuidado e responsabilidade
    - ✓ "With great power, comes great responsibility" (Uncle Bem/Stan Lee)
      - Pequenos loops, limitado a uma quantidade que valide seu experimento.

Webscrapy do crunchbase.com – Download de Arquivo

- Salvar arquivo automaticamente
  - Sem pedir pra escolher um diretório de destino

```
fp = webdriver.FirefoxProfile()
fp.set_preference("browser.download.folderList", 2)
fp.set_preference("browser.download.manager.showWhenStarting", False)
fp.set_preference("browser.download.dir", self.download_dir)
fp.set_preference("browser.helperApps.neverAsk.saveToDisk",
                  """text/plain, application/octet-stream, application/binary,
                   text/csv, application/csv, application/excel,
                   text/comma-separated-values, text/xml, application/xml""")
fp.set_preference("pdfjs.disabled", True)
self.driver = webdriver.Firefox(firefox_profile=fp)
```

Preencher a busca incremental (parametrizada)

```
dates_founded = pd.date_range(start=start_founded, end=end_founded, freq="3M")
status = wait_element(self.driver, "//search-date//input[@type='search']", by=By.XPATH,
ins_search = self.driver.find_elements_by_xpath("//search-date//input[@type='search']")
for sa, ea in zip(dates_founded[0:-1], dates_founded[1:]):
    in_start_announced = ins_search[0]
    in_start_announced.clear()
    in_start_announced.send_keys(sa.strftime("%Y/%m/%d"))
    in_end_announced = ins_search[1]
    in_end_announced.clear()
    in_end_announced.send_keys(ea.strftime("%Y/%m/%d"))
    wait_element(self.driver, '//button[@aria-label="Search"]', by=By.XPATH, to_sleep=2)
    button_search = self.driver.find_element_by_xpath('//button[@aria-label="Search"]')
    button_search.click()
```

Webscrapy do crunchbase.com – Exportar e renomear

- Renomear e mover o arquivo

```
button_export = self.driver.find_element_by_xpath('//export-csv-button//button')
button_export.click()
today_str = datetime.datetime.today().strftime("%m-%d-%Y")
filename = f"{self.download_dir}/{self.search_name}-{today_str}.csv"
while not os.path.exists(filename):
    time.sleep(1)
dest_file = f"""{self.dest_dir}/{self.search_name}_
                {sa.strftime('%Y%m%d')}_{ea.strftime('%Y%m%d')}.csv"""
shutil.move(filename, dest_file)
```

Webscrapy do crunchbase.com – Download de Arquivo

- Salvar arquivo automaticamente
  - Sem pedir pra escolher um diretório de destino

```
def concat_files(files_path, drop_duplicates_by=[]):
    #Exemplo: files_path="./*.csv"
    import glob
    import datetime
    pieces = []
    result = None
    for f in glob.glob(files_path):
        frame = pd.read_csv(f, engine='python')
        pieces.append(frame)
    result = pd.concat(pieces, ignore_index=True)
    if drop_duplicates_by:
        result.drop_duplicates(subset=drop_duplicates_by, inplace=True)
    return result
```

Pre-visualização de uma amostra dos dados

id	Organization Name	Industries	Founded Date	Number of Founders	Number of Funding Rounds
0	Sounds	Apps, Digital Entertainmen	16/04/2014	1.0	2
1	Responsival	Advertising, Consulting, E-(	16/04/2014	2.0	1
2	FALCON Agency	Advertising, Digital Media	04/04/2014	2.0	1
3	Single Case	Consulting	24/04/2014	1.0	1
4	FromLabs	Education, Enterprise Softv	01/11/2012	2.0	1
5	Getup Cloud	Cloud Computing, PaaS, So	20/11/2012	1.0	2
6	MistLayer	Artificial Intelligence, Clou	10/11/2012	1.0	1
7	Vested	Financial Services, FinTech	01/08/2018	3.0	5

Calculando o tempo médio das rodadas de investimento

- Salvar arquivo automaticamente
  - Sem pedir pra escolher um diretório de destino

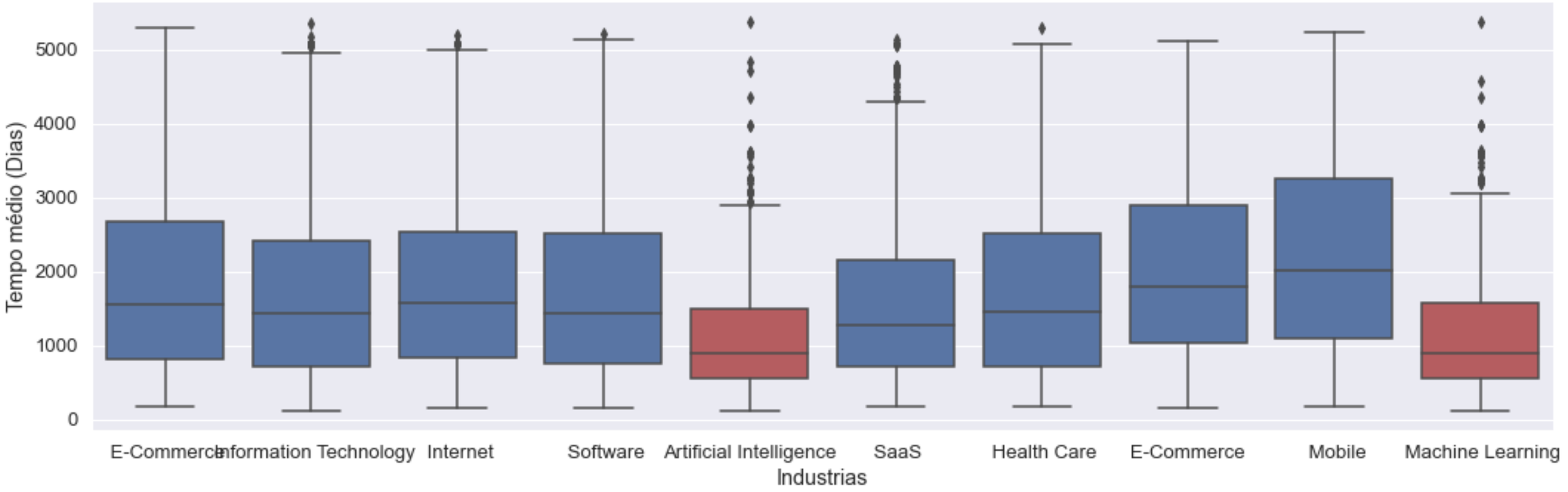
```
df_old['inter_funding_period'] = datetime.today() - df_old['Founded Date']
df_old['inter_funding_days'] = df_old['inter_funding_period'].apply(
                                    lambda x: x/np.timedelta64(1, 'D'))
df_old['average_if_period'] = df_old['inter_funding_days']
                                  /df_old['Number of Funding Rounds']
```

Seaborn (Gráficos com Python/Pandas)

- Interface de alto nível
  - para o Matplotlib
- Assume valores padrão
  - facilitando a vida do usuário
- Foca em tipos de problemas
  - e não em tipos de gráficos

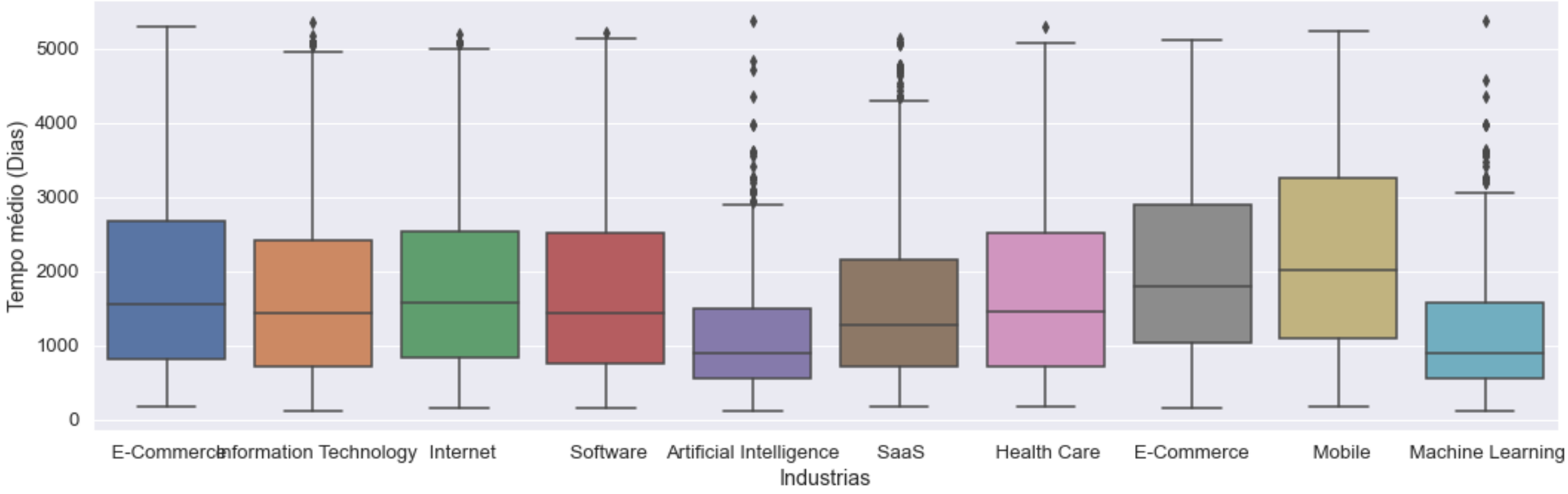
Tempo médio das rodadas de investimento X Indústria

- IA / Machine Learning está na "crista da onda"
  - Elas têm as menores medianas
    - ✓ e as outras medidas de tendência central também

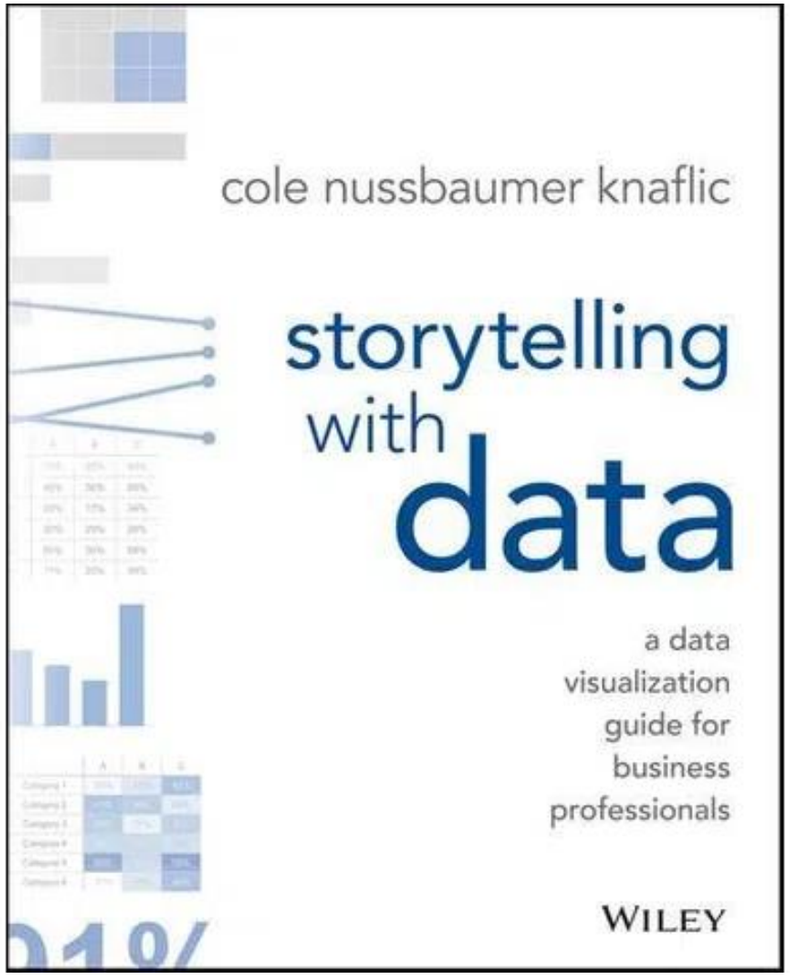


Tempo médio das rodadas de investimento X Indústria

- IA / Machine Learning está na "crista da onda"
  - Elas têm as menores medianas
    - ✓ e as outras medidas de tendência central também



Análise Explanatória e Storytelling



Conte os números 3

75639506847 3

6586630375 76

8603726586 02

8465891078 30

FIGURE 4.2 Count the 3s example

Conte os números 3

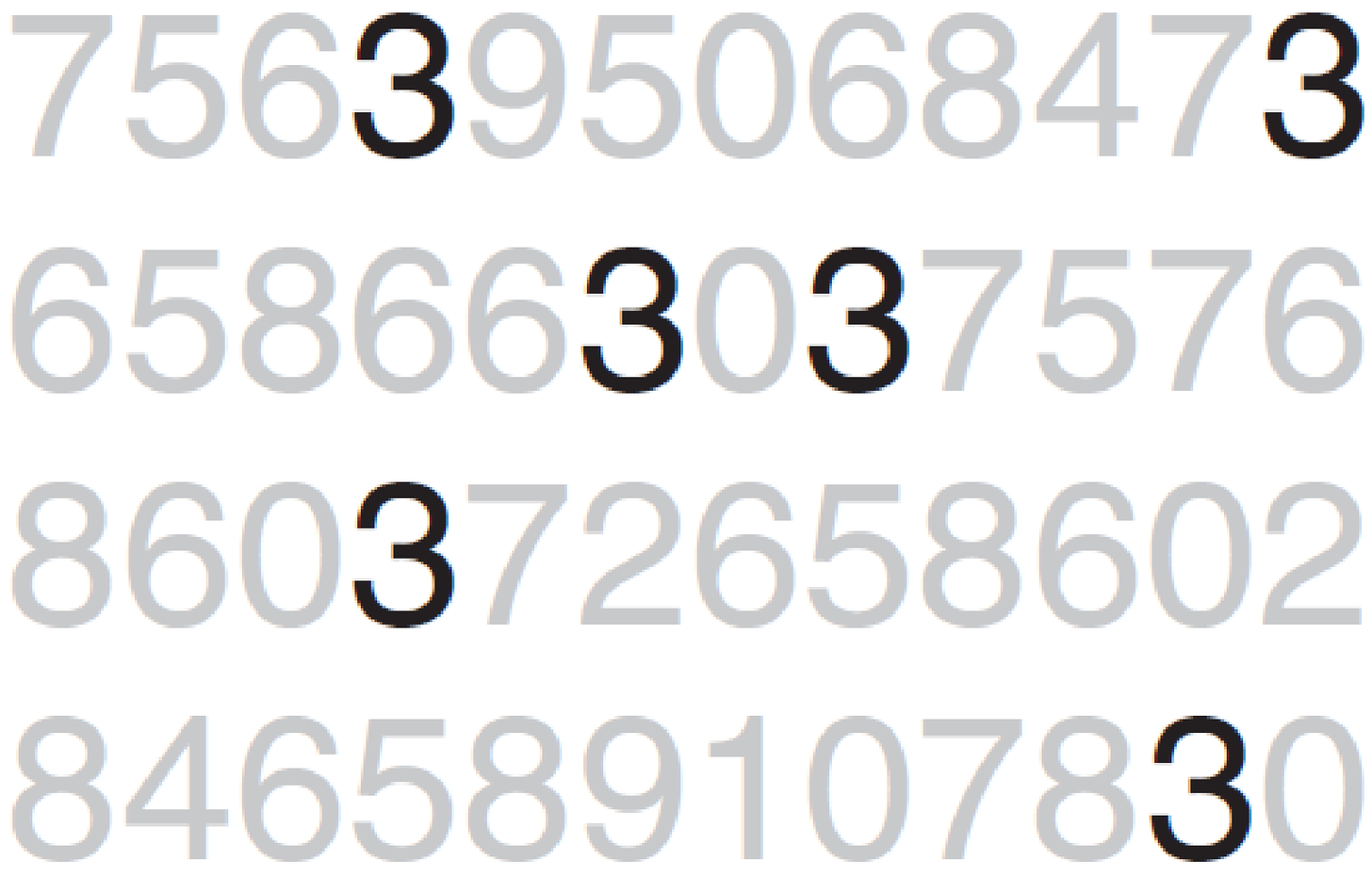


FIGURE 4.3 Count the 3s example with preattentive attributes

Tempo médio das rodadas de investimento X Indústria

- Código para gerar o gráfico com o Seaborn

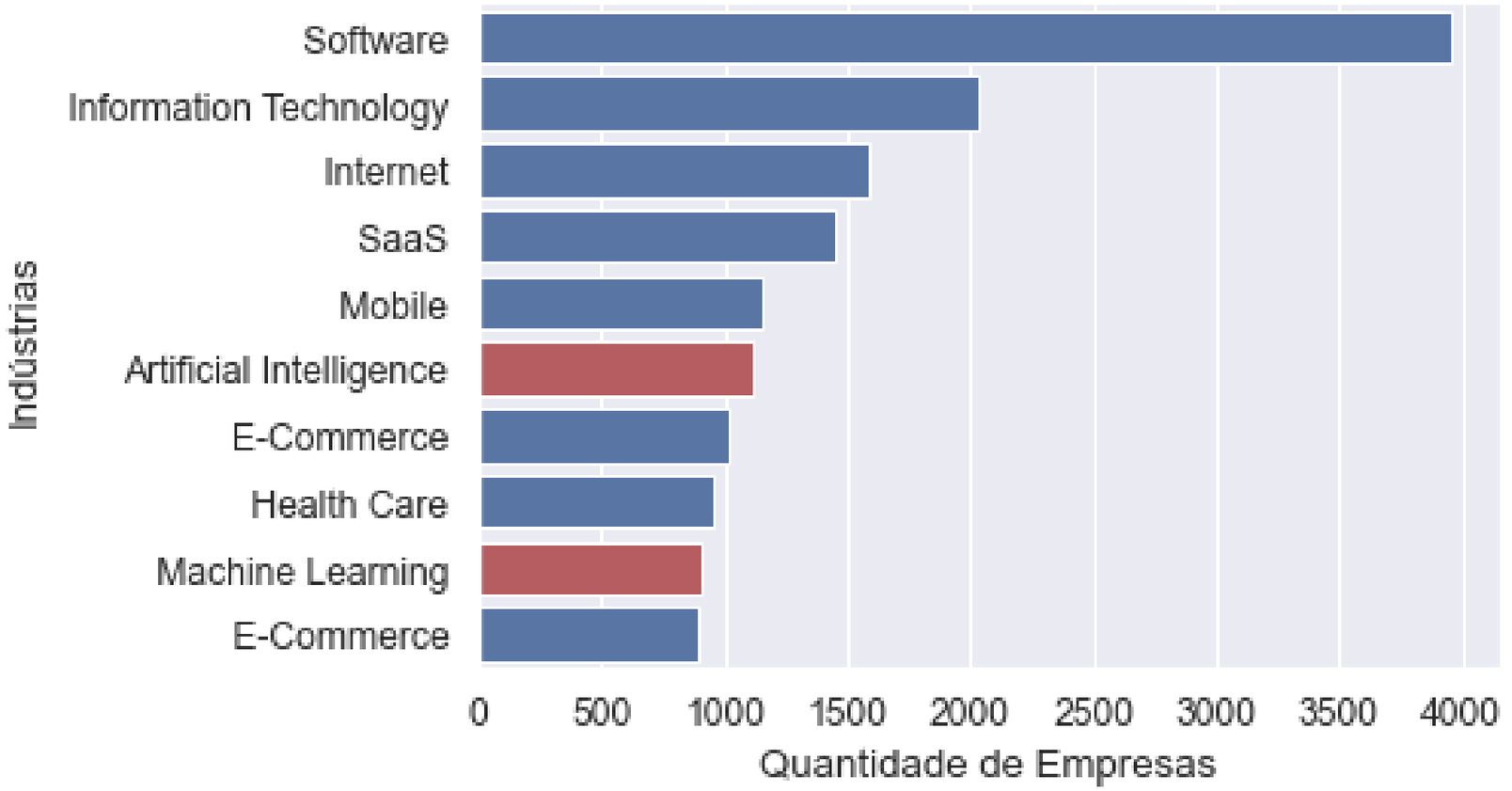
```
sns.set(font_scale=1.2)
my_pal = {ind: "r" if ind in ['Artificial Intelligence', ' Machine Learning']
                      else "b" for ind in df_ind["Industries_x"].unique()}
ax = sns.catplot(y="average_if_period", data=df_ind, kind="box",
                 x="Industries_x", aspect=3, palette=my_pal,
                 order=pd.value_counts(df_ind['Industries_x']).head(10).index)
ax.set(xlabel='Industrias', ylabel="Tempo médio (Dias)")
```

Tempo médio das rodadas de investimento X Indústria

- Como transforma uma coluna multi-valorada em várias colunas?

```
ind = df_old[['Industries', 'id']].assign(
        Industries=df_old.['Industries'].str.split(',')).explode('Industries')
df_ind = ind.merge(right=df_old, on='id', how='left')
```

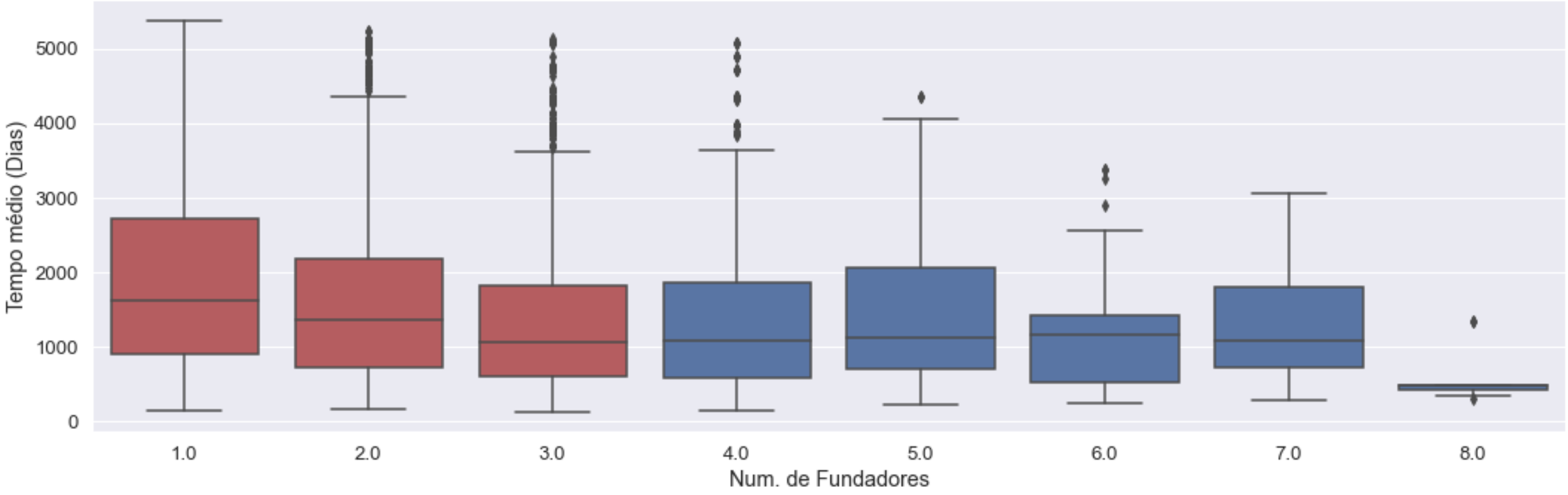
Quantidade de Empresas por Indústria




```
my_pal = {ind: "r" if ind in ['Artificial Intelligence', ' Machine Learning']
                   else "b" for ind in df_rel_ind["Industries_x"].unique()}
ax = sns.countplot(y="Industries_x", data=df_rel_ind, palette=my_pal,
                   order=pd.value_counts(df_ind['Industries_x']).iloc[:10].index)
ax.set(xlabel='Quantidade de Empresas', ylabel="Indústrias")
```

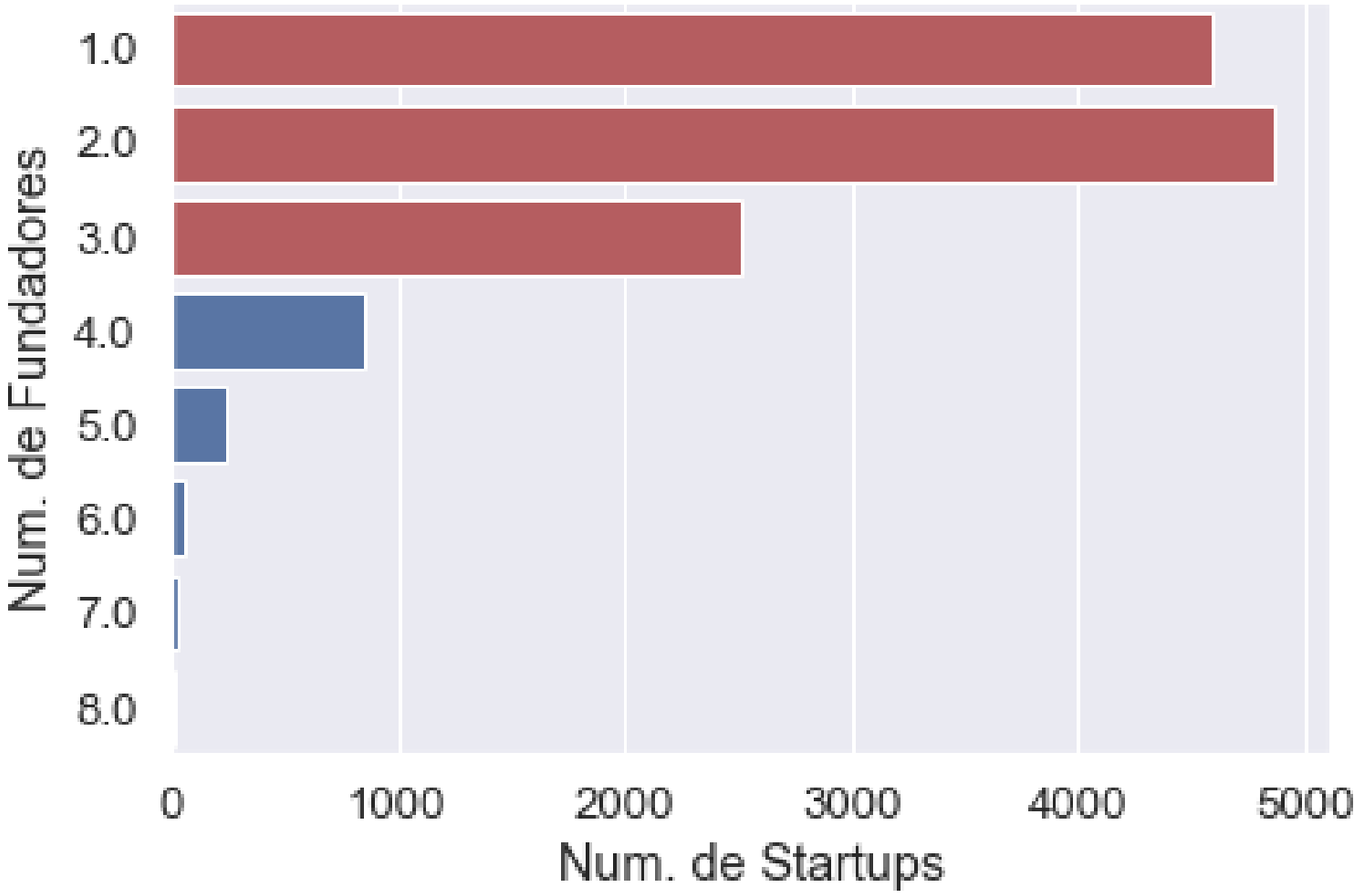
Tempo médio das rodadas de investimento X Quantidade de Fundadores

- O consenso faz a diferença



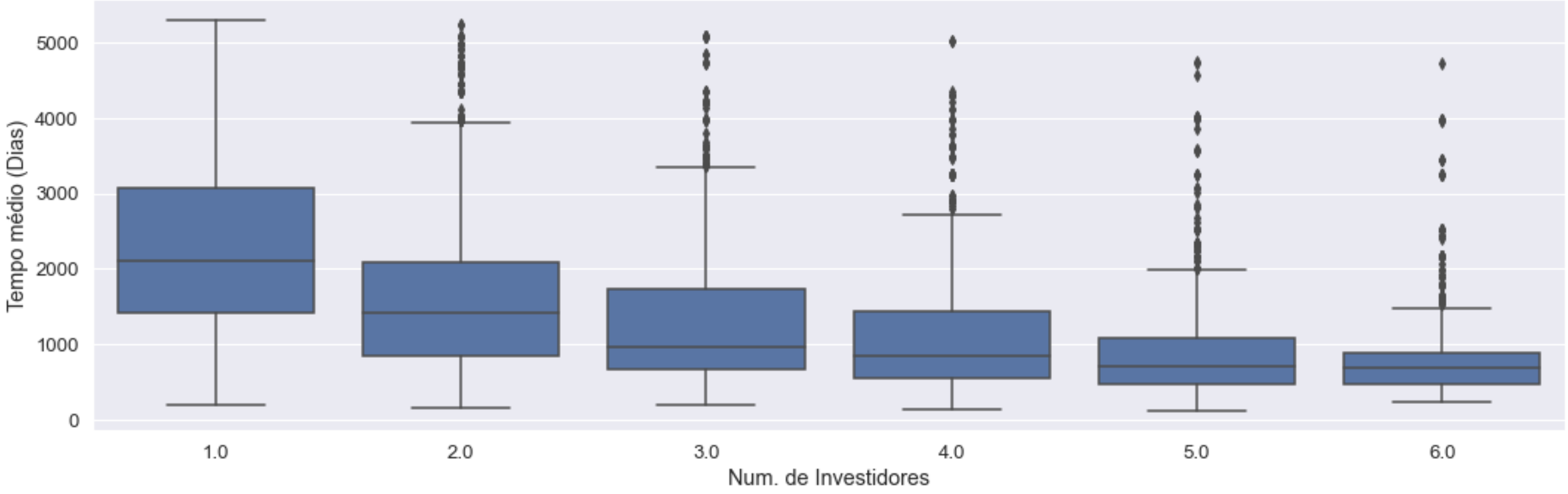
Tempo médio das rodadas de investimento X Quantidade de Fundadores

- O consenso faz a diferença



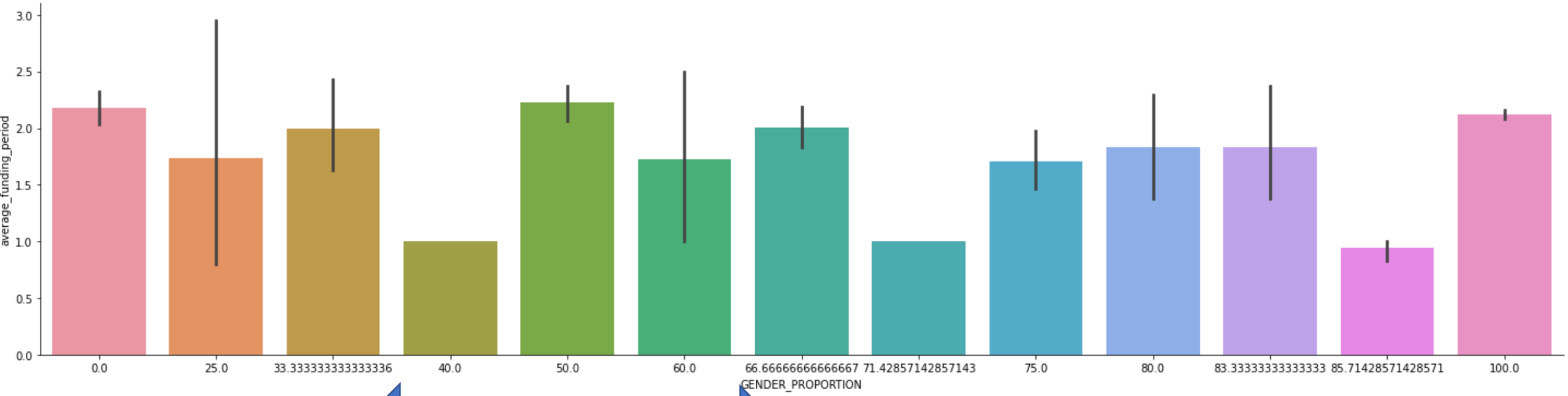
Tempo médio das rodadas de investimento X Quantidade de Investidores

- Se você convence uma maior quantidade de investidores
  - tem maiores chances de sucesso



Efeito do gênero sobre o tempo médio das rodadas de investimento

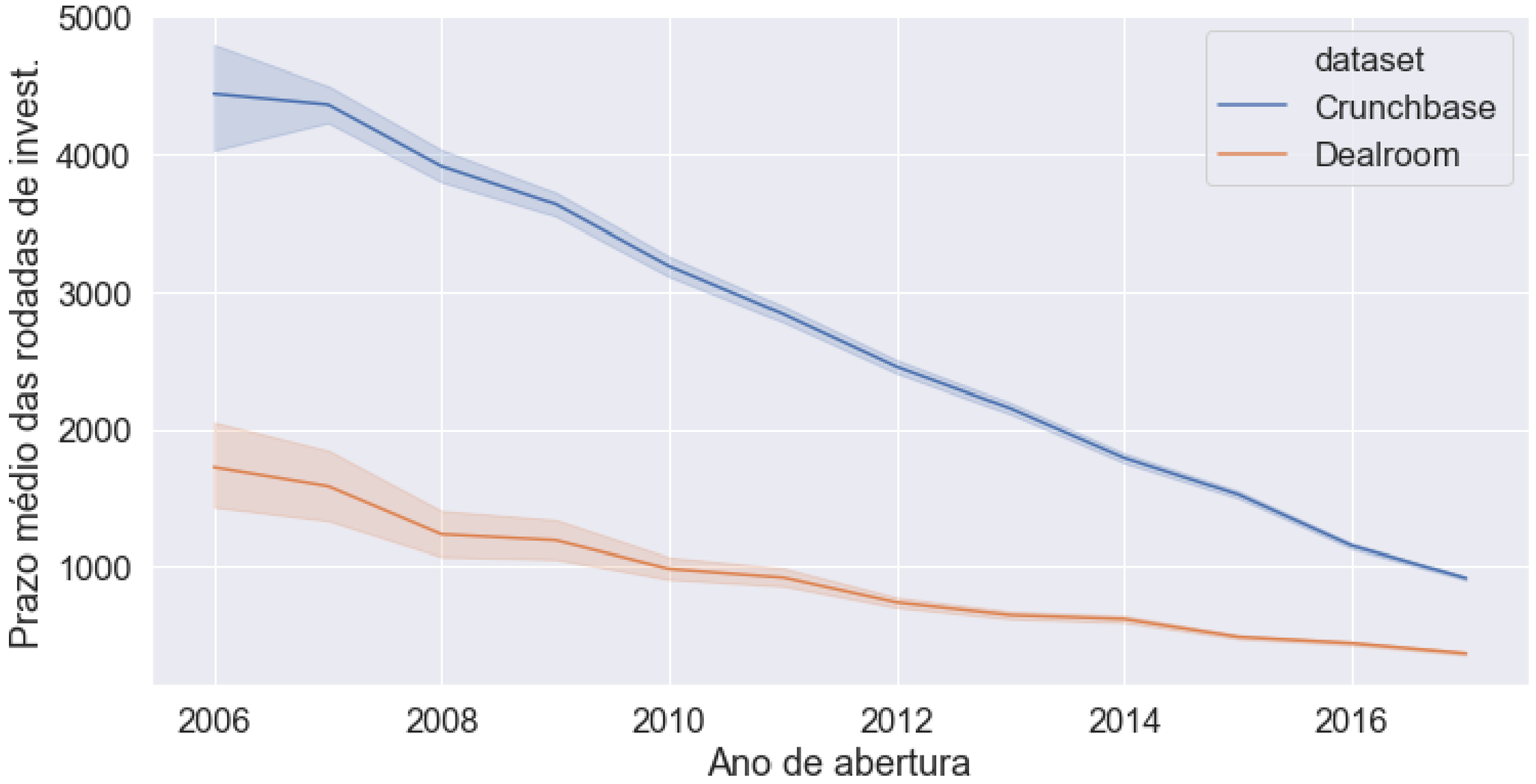
- Não há efeito do gênero sobre o sucesso das startups
  - No eixo vertical: Número médio de anos em cada rodada



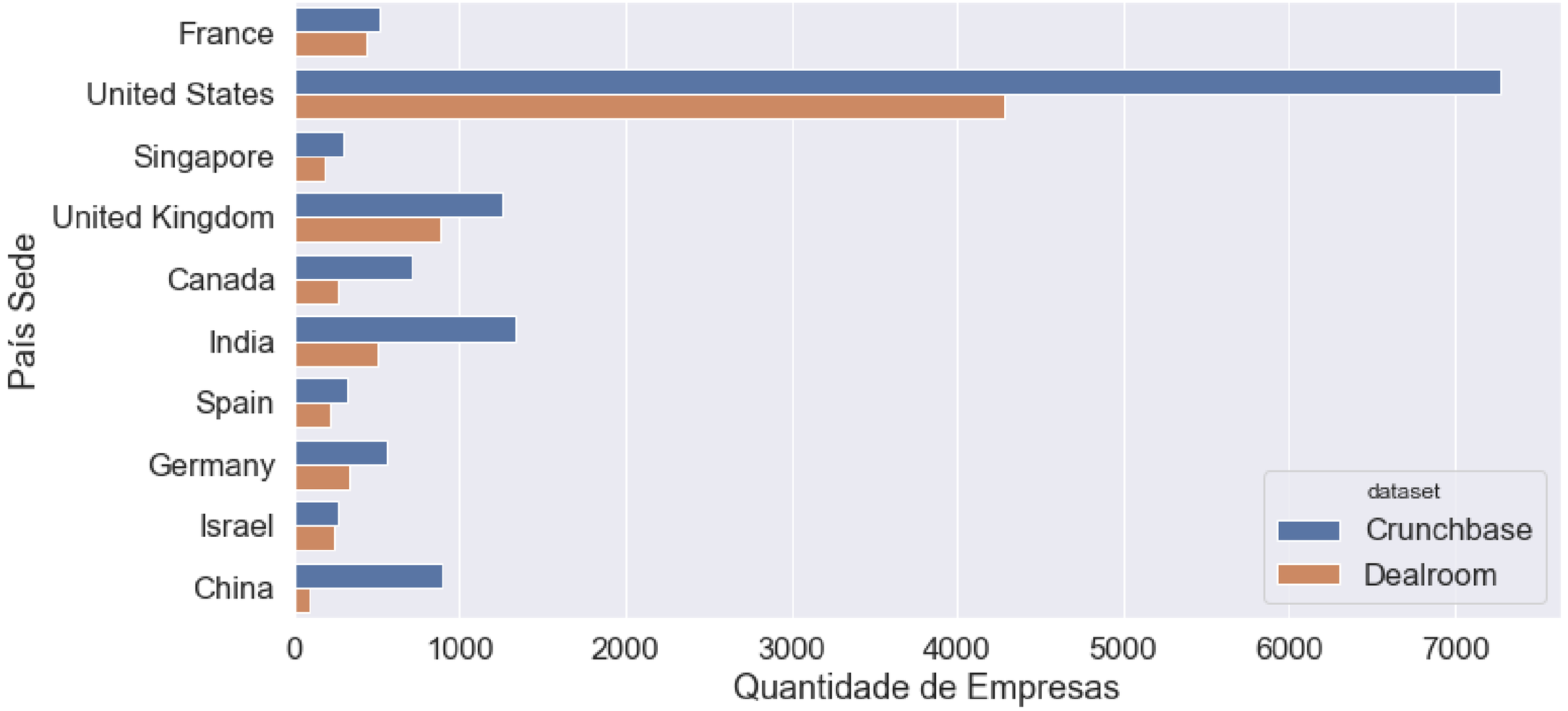
Mais mulheres

Mais homens

O prazo médio das rodadas de investimento tem diminuído



Análise da Quantidade de Empresas por País



FIM

Obrigado e boa sorte com a

Ciência de Dados

Código fonte dos notebooks da aula de hoje:
https://github.com/alexlopespereira/companies