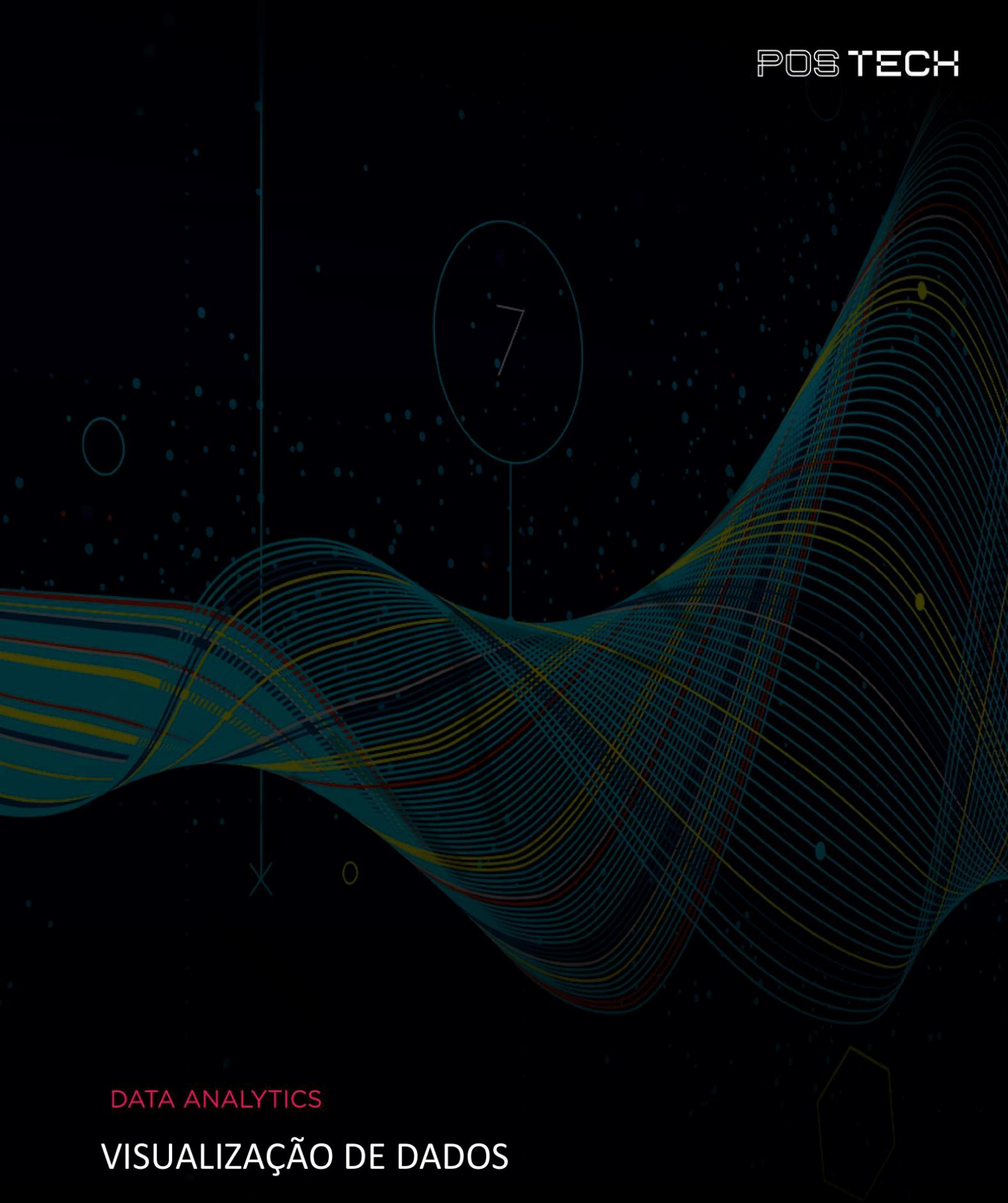


DATA ANALYTICS

VISUALIZAÇÃO DE DADOS

AULA 01

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

SUMÁRIO

FIAP

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

O QUE VEM POR AÍ?

Olá, jovem padawan!

Você está na segunda disciplina do curso! Aqui começaremos mais um ponto da jornada em que você está trilhando.

Até aqui, você teve os primeiros contatos com o Pandas e a como manipular visualizações básicas do seu Dataframe.

A fim de complementar suas análises, essa aula vai te mostrar como ler dados de fontes diferentes e teremos uma introdução à união de Dataframes com pequenas manipulações e tratamentos.

A base de dados você encontra aqui <https://github.com/alura-cursos/agendamento-hospitalar/blob/main/dados/estimativa_dou_2020.xls?raw=true>.

Agora, vamos para cima deste maravilhoso mundo dos dados!

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

CONHEÇA SOBRE O ASSUNTO

Em âmbito prático e na realidade, é comum trabalharmos com situações em que temos que buscar dados de diversas fontes como csv, excel, html, pdf e muito mais!

A questão aqui é: como trazer esses dados e por onde eu posso lê-los?

Estamos acostumados a trabalhar com arquivos que estão no mesmo diretório do nosso notebook ou código Python, mas você sabia que as opções de “read” do Pandas aceitam que você leia arquivos que estejam na internet?

Uau... E como começar a ler desta forma?

De certa forma, é bem simples. Imagine que você quer ler um arquivo excel de um repositório github, dessa forma você pode simplesmente codar:

```
ibge_estimativa = pd.read_excel("https://github.com/alura-cursos/agendamento-hospitalar/blob/main/dados/estimativa_dou_2020.xls?raw=true")
```

Código 1 - Exemplo – ler arquivos com a url
Fonte: Elaborado pelo autor (2023)

Veja que agora o parâmetro não é o “caminho” de um arquivo excel em minha máquina local, mas um “caminho” dentro da web. O importante aqui é o caminho passado onde terá que existir esse arquivo excel.

Outra coisa muito comum é surgir a seguinte pergunta: posso só copiar e colar de uma página web o conteúdo de uma tabela direto para o Python?

A reposta é: sim! Porém, não é recomendado, a não ser que seja uma última instância de desespero. Já que, se você precisa ler dados de páginas web que não estão contidos em arquivos, você tem o read_html( ) do Pandas ou até mesmo bibliotecas auxiliares com requests, urllib e BeautifulSoap, que fazem um web scraping (cenas fases futuras) que é uma beleza!.

Uma maneira que trouxemos para você ver, de forma simplificada, utilizando StringIO do python está contida na primeira parte da nossa super videoaula!

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

Outro ponto interessante é olharmos para um universo dos tratamentos de dados de maneira mais incisiva.

Não é comum na realidade seus dados virem do jeitinho que você imagina, com os campos certos, nenhum valor nulo, campos com o seu “tipo” correto.

Por isso, existem funções que nos ajudam, como por exemplo a função dropna( ) que dropa tudo o que é nulo (aconselho a examinar com calma, porque ela deleta registros caso algo nulo seja identificado, podendo tirar mais do que deveria).

Outro caso é a função .info( ) que nos traz uma descrição do Dataframe, nos datando a certidão de nascimento desse objeto, mostrando as colunas e seus tipos nativos.

A partir daí é função que não acaba mais...

Na nossa videoaula foram feitas algumas codificações:

FIAP

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

```
    ibge_estimativa = pd.read_excel("https://github.com/alura-
cursos/agendamento-
hospitalar/blob/main/dados/estimativa_dou_2020.xls?raw=true")
    ibge_estimativa.head()
```

```
dados_da_populacao = """Posição Unidade federativa   População  % da
pop. total País comparável
(habitantes)

1     São Paulo 46 649 132 21,9% Flag of Spain.svg Espanha (46 439
864)
2     Minas Gerais   21 411 923 10,1%  Sri Lanka (20 675 000)
3     Rio de Janeiro 17 463 349 8,2%   Países Baixos (16 922 900)
4    Bahia Bahia     14 985 284 7,1%   Chade (14 037 000)
5     Paraná    11 597 484 5,4%   Bolívia (11 410 651)
6     Rio Grande do Sul    11 466 630 5,4%   Bélgica (11 250 659)
7     Pernambuco     9 674 793  4,5%   Bielorrússia (9 485 300)
8     Ceará     9 240 580  4,3%   Emirados Árabes Unidos (9 157 000)
9    Pará Pará  8 777 124  4,1%   Áustria (8 602 112)
10    Santa Catarina 7 338 473  3,4%   Sérvia (7 114 393)
11    Goiás     7 206 589  3,4%   Paraguai (7 003 406)
12    Maranhão  7 153 262  3,4%   Paraguai (7 003 406)
13    Amazonas  4 269 995  2,0%   Líbano (4 168 000)
14    Espírito Santo 4 108 508  1,9%   Líbano (4 168 000)
15    Paraíba   4 059 905  1,9%   Líbano (4 168 000)
16    Mato Grosso    3 567 234  1,7%   Uruguai (3 415 866)
17    Rio Grande do Norte  3 560 903  1,7%   Uruguai (3 415 866)
18    Alagoas   3 365 351  1,6%   Uruguai (3 415 866)
19    Piauí     3 289 290  1,6%   Kuwait (3 268 431)
20    Distrito Federal     3 094 325  1,4%   Lituânia (2 900 787)
21    Mato Grosso do Sul   2 839 188  1,3%   Jamaica (2 717 991)
22    Sergipe   2 338 474  1,1%   Namíbia (2 280 700)
23    Rondônia  1 815 278  0,8%   Gabão (1 725 000)
24    Tocantins 1 607 363  0,7%   Bahrein (1 359 800)
25    Acre 906 876    0,4%   Fiji (859 178)
26    Amapá     877 613    0,4%   Fiji (859 178)
27    Roraima   652 713    0,3%   Luxemburgo (562 958)"""


# fonte
https://pt.wikipedia.org/wiki/Lista_de_unidades_federativas_do_Brasi
l_por_popula%C3%A7%C3%A3o#cite_note-IBGE_POP-1
# fonte indireta IBGE
```

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

Na estrutura exibida, na primeira parte lemos a base de dados em Excel, enquanto na segunda parte estamos preocupados em trazer essa base de dados em forma de string.

Mas, para fazer essa ação, devemos saber manipular isto. Então, em seguida podemos realizar:

```
from io import StringIO

dados_da_populacao_io = StringIO(dados_da_populacao)

novos_dados = pd.read_csv(dados_da_populacao_io, sep="\t")
novos_dados = novos_dados.dropna()
novos_dados.head()
```

```
populacao.columns = ["posicao", "uf", "populacao", "porcentagem",
"pais_comparavel"]
populacao["populacao"] = populacao["populacao"].str.replace(" ",
"").astype(int)
populacao = populacao[["uf", "populacao"]]
populacao.head()
```

```
populacao.info()
populacao.describe()
```

```
display(gastos_do_mais_recente.head())
display(populacao.head())
```

```
populacao = populacao.set_index("uf")
populacao.head()
```

```
populacao.index = populacao.index.str.strip()
populacao.head()
```

```
gastos_do_mais_recente.index = gastos_do_mais_recente.index.str[3:]
gastos_do_mais_recente.head()
```

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

```
for estado in gastos_do_mais_recente.index:
  populacao.index = populacao.index.str.replace(f"{estado} {estado}", estado)
populacao.index
```

```
gastos_e_populacao = populacao.join(gastos_do_mais_recente)
gastos_e_populacao.head()
```

Notebook Aula 1 – Dados IBGE
Fonte: Elaborado pelo autor (2023)

Dica de leitura:

Um bom jeito de entrar no mundo do Analytics é acompanhar as principais páginas de artigos e blogs da área.

Que tal dar uma lida no **artigo** <https://www.analyticsvidhya.com/blog/2021/06/data-manipulation-using-pandas-essential-functionalities-of-pandas-you-need-to-know/#:~:text=Pandas%20is%20an%20open%2Dsource,work%20on%2C%20DataFrame%20and%20Series> do blog Analytics Vidhya? Aqui ele te dá uma visão geral do Pandas e algumas manipulações importantes.

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

HANDS ON

Agora, chegou o momento de ver, na prática, como começar a importar nossos dados e trabalhar com eles via programação. A ideia é não se limitar ao código explícito no hands on, então é sempre bom procurar a documentação das bibliotecas, explorar novas funcionalidades e muito mais!

Uma dica valiosa neste momento é que você utilize a base dessa aula para replicar os conhecimentos para ler dados a partir de strings. Experimente pegar sites, procure por tabelas, copie o conteúdo e cole na sua string para poder treinar essas formatações.

Já os notebooks das aulas se encontram aqui <https://github.com/alura-tech/pos-datascience-introducao-a-visualizacao/archive/refs/heads/aula1.zip>.

FIAP

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

O QUE VOCÊ VIU NESTA AULA?

Como comparar os gastos em relação à determinado estado; como fazer a leitura de dados no formato Excel com Pandas; como tornar uma string em (com formato de tabela) legível em Pandas usando o String.IO e como usar Join em Pandas. Isso tudo foi o que você viu nesta aula!

Daqui para a frente, é importante que você replique os conhecimentos adquiridos para fortalecer ainda mais suas bases e conhecimentos.

IMPORTANTE: não esqueça de praticar com o desafio da disciplina, para que assim você possa aprimorar os seus conhecimentos!

Você não está sozinho ou sozinha nesta jornada! Te esperamos no Discord e nas lives com os especialistas, onde você poderá tirar dúvidas, compartilhar conhecimentos e estabelecer conexões!

FIAP

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

REFERÊNCIAS

DOCUMENTAÇÃO PANDAS. **<https://pandas.pydata.org/>.** Acesso em: 08 fev. 2023.

GOOGLE COLAB. **<https://colab.research.google.com/>.** Acesso em: 08 fev. 2023.

IBGE. **<https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.html?=&t=resultados>.** Acesso em: 08 fev. 2023.

TABNET. **<https://datasus.saude.gov.br/informacoes-de-saude-tabnet/>.** Acesso em: 08 fev. 2023.

FIAP

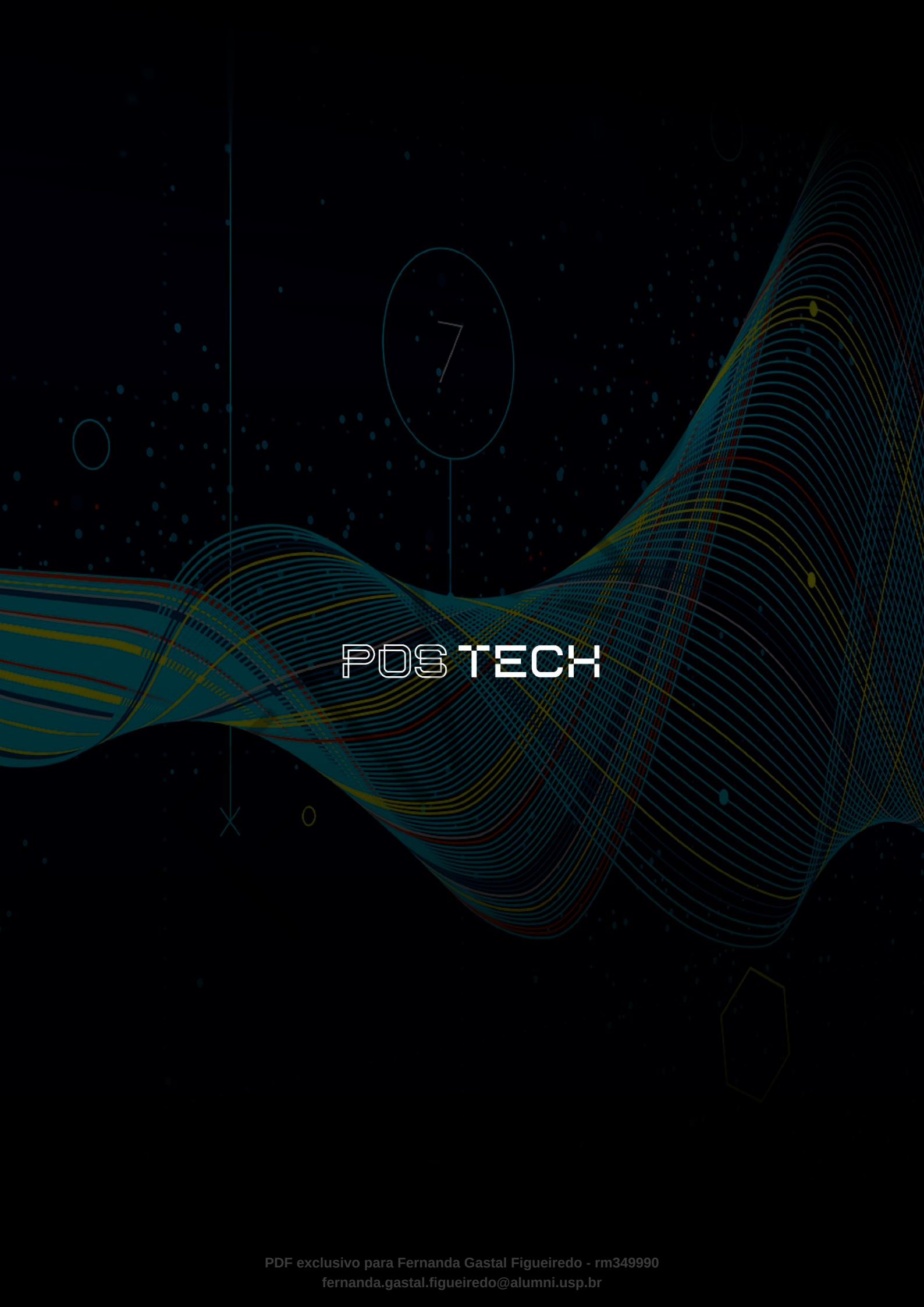
PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br

PALAVRAS-CHAVE

Python. Pandas. Dataframe.

FIAP

PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br



PDF exclusivo para Fernanda Gastal Figueiredo - rm349990
fernanda.gastal.figueiredo@alumni.usp.br