



# Programa de Pós-graduação em Sistemas de Informação

SIN5007 - Reconhecimento de Padrões (2023)

**Atividade 1:** Análise exploratória de dados

MSc. Leonardo Cunha dos Santos  
Gabriel Francisco dos Santos Silva

São Paulo / 2023

# Agenda

- ➊ Introdução
- ➋ Conjunto de dados
- ➌ Pré-processamento de dados
- ➍ Análise exploratória
- ➎ Considerações finais

### GERAL

# Estudantes de TI têm alto índice de abandono da universidade

Cerca de 60% dos alunos dos cursos como ciência da computação, design de games ou sistema de informação não se formam.




Dino - Divulgador de Noticias

14 de Agosto de 2023 às 09:41  
Atualizado 14/08/2023 09:41:16

<https://www.folhavoria.com.br/geral/noticia/08/2023/estudantes-de-ti-tem-alto-indice-de-abandono-da-universidade>

### JORNAL DA USP

 PORTAL DA USP —  FALE CONOSCO —  WHATSAPP —  ENVIE UMA PAUTA —  NEWSLETTER —  PODCASTS —  RÁDIO USP


ATUALIDADES ▾ CIÊNCIAS ▾ CULTURA ▾ DIVERSIDADE ▾ EDUCAÇÃO INSTITUCIONAL ▾ RÁDIO USP ▾ TECNOLOGIA UNIVERSIDADE ▾  BUSCA

Início > Institucional > Evasão na graduação da USP é de 17%, mas número varia por curso



## Evasão na graduação da USP é de 17%, mas número varia muito por curso

Pesquisa investigou fatores relacionados à evasão na turma de 2018, a primeira com cotas sociais e raciais

 Publicado: 15/08/2023

Texto: Silvana Salles

Arte: Carolina Borin\*

- Os microdados do Inep reúnem informações detalhadas sobre pesquisas do INEP;
- As estatísticas produzidas pelo Inep visam fornecer os subsídios para a formulação e implementação de políticas voltadas para a melhoria contínua da educação no país;
- Os formatos de apresentação foram reestruturados de acordo com a Lei Geral de Proteção de Dados Pessoais (LGPD).

<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>

# Conjunto de dados

Microdados do Censo da Educação Superior

## Cadastro de IES

- Localização
- Quadro de funcionários
- Infraestrutura
- Estatísticas sobre professores

## Cadastro de Cursos

- Grau acadêmico
- Modalidade
- Estatísticas sobre estudantes

# Pré-processamento de dados

## Dicionário de dados

	Alteração de nomenclatura
	Variável nova
	Descontinuidade

### Cadastro\_IES

N	Nome da Variável	Descrição da Variável	Tipo	Tam.	Categoria
1	NU_ANO_CENSO	Ano de referência do Censo da Educação Superior	Num	4	
<b>DADOS DA INSTITUIÇÃO DE ENSINO SUPERIOR (IES) - SEDE ADMINISTRATIVA/REITORIA</b>					
2	NO_REGIAO_IES	Nome da região geográfica da sede administrativa ou reitoria da IES	Char	20	
3	CO_REGIAO_IES	Código da região geográfica da sede administrativa ou reitoria da IES	Num	2	
4	NO_UF_IES	Nome da Unidade da Federação da sede administrativa ou reitoria da IES	Char	50	
5	SG_UF_IES	Sigla da Unidade da Federação da sede administrativa ou reitoria da IES	Char	2	
6	CO_UF_IES	Código da Unidade da Federação da sede administrativa ou reitoria da IES	Num	2	
7	NO_MUNICIPIO_IES	Nome do Município da sede administrativa ou reitoria da IES	Char	150	
8	CO_MUNICIPIO_IES	Código do Município da sede administrativa ou reitoria da IES	Num	7	
9	IN_CAPITAL_IES	Informa se a sede administrativa ou reitoria da IES está localizada na capital da Unidade da Federação	Num	2	0. Não 1. Sim
10	NO_MESORREGIAO_IES	Nome da Mesorregião da sede administrativa ou reitoria da IES	Char	100	
11	CO_MESORREGIAO_IES	Código da Mesorregião da sede administrativa ou reitoria da IES	Num	4	
12	NO_MICRORREGIAO_IES	Nome da Microrregião da sede administrativa ou reitoria da IES	Char	100	
13	CO_MICRORREGIAO_IES	Código da Microrregião da sede administrativa ou reitoria da IES	Num	5	
14	TP_ORGANIZACAO_ACADEMICA	Tipo de Organização Acadêmica da IES	Num	1	1. Universidade 2. Centro Universitário 3. Faculdade 4. Instituto Federal de Educação, Ciência e Tecnologia 5. Centro Federal de Educação Tecnológica

# Pré-processamento de dados

Microdados 2021 - Variáveis

## Cadastro de IES

- Categóricas: 32 + 1 (ID:ANO)
- Numéricas: 48
- Total: 81

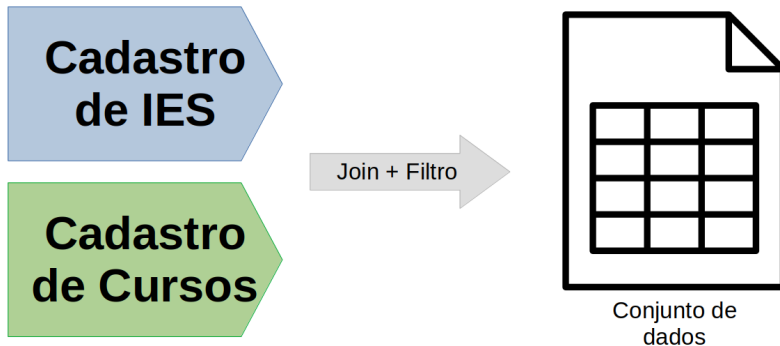
## Cadastro de CURSOS

- Categóricas: 27 + 1 (ID:ANO)
- Numéricas: 172
- Total: 200



# Pré-processamento de dados

Cursos de tecnologia



Variáveis numéricas: 05 | Variáveis categóricas: 22 + 1 (ID:ANO)  
34 cursos | Instâncias: 19158 | 558 cursos distintos

# Pré-processamento de dados

Depara de cursos



**EACH** | campus capital  
Escola de Artes, Ciências e Humanidades  
Universidade de São Paulo

NO CURSO DEPARA	NO CURSO
Agrocomputação	Agrocomputação
Análise e Desenvolvimento de Sistemas	Administração Em Sistemas E Serviços De Saúde Análise De Infraestrutura De Redes E Sistemas Computacionais Análise De Sistemas Análise E Desenvolvimento De Sistemas Desenvolvimento De Sistemas Sistemas Para Internet
Ciências da Computação	Abi - Ciência Da Computação Ciência Da Computação Ciências Da Computação Ciências De Computação Computação Computação E Informática Computação E Robótica Educativa Computação Em Nuvem Computação Gráfica Internet Das Coisas E Computação Em Nuvem
Engenharia da Computação	Engenharia Da Computação Engenharia De Computação Engenharia De Computação - Ênfase Sistemas Corporativos Engenharia De Computação E Informação
Engenharia de Sistemas	Engenharia De Automação E Sistemas Engenharia De Produção E Sistemas Engenharia De Sistemas Engenharia De Sistemas Ciber Físicos
Engenharia Elétrica - Ênfase Em Computação	Engenharia Elétrica - Ênfase Em Computação Engenharia Elétrica - Ênfase Em Eletrônica E Sistemas Computacionais Engenharia Eletrônica E De Computação
Matemática Aplicada e Computação Científica	Interdisciplinar Em Matemática E Computação E Suas Tecnologias Matemática Aplicada Com Habilitação Em Sistemas E Controle Matemática Aplicada E Computação Científica Matemática Aplicada E Computacional Com Habilitação Em Sistemas E Controle
Sistemas de Computação	Sistemas De Computação
Sistemas de Informação	Sistemas De Informação

# Análise exploratória

## Método info da biblioteca Pandas para variáveis numéricas

```
RangeIndex: 19158 entries, 0 to 19157
```

```
Data columns (total 28 columns):
```

#	Column	Non-Null Count	Dtype
0	NU_ANO_CENSO	19158 non-null	object
1	NO_REGIAO	19158 non-null	object
2	NO_UF	19158 non-null	object
3	SG_UF	19158 non-null	object
4	NO_MUNICIPIO	19158 non-null	object
5	CO_MUNICIPIO	19158 non-null	object
6	IN_CAPITAL_DEPARA	19158 non-null	object
7	CO_IES	19158 non-null	object
8	SG_IES	19158 non-null	object
9	NO_IES	19158 non-null	object
10	CO_MANTENEDORA	19158 non-null	object
11	NO_MANTENEDORA	19158 non-null	object
12	NO_CURSO	19158 non-null	object

13	NO_CURSO_DEPARA	19158 non-null	object
14	CO_CURSO	19158 non-null	object
15	TP_GRAU_ACADEMICO_DEPARA	19155 non-null	object
16	IN_GRATUITO_DEPARA	19158 non-null	object
17	TP_MODALIDADE_ENSINO_DEPARA	19158 non-null	object
18	TP_NIVEL_ACADEMICO_DEPARA	19158 non-null	object
19	TP_DIMENSAO_DEPARA	19158 non-null	object
20	TP_ORGANIZACAO_ACADEMICA_DEPARA	19158 non-null	object
21	TP_CATEGORIA_ADMINISTRATIVA_DEPARA	19158 non-null	object
22	TP_REDE_DEPARA	19158 non-null	object
23	QT_V6_TOTAL	19158 non-null	int64
24	QT_INSCRITO_TOTAL	19158 non-null	int64
25	QT_ING	19158 non-null	int64
26	QT_MAT	19158 non-null	int64
27	QT_CONC	19158 non-null	int64

```
dtypes: int64(5), object(23)
```

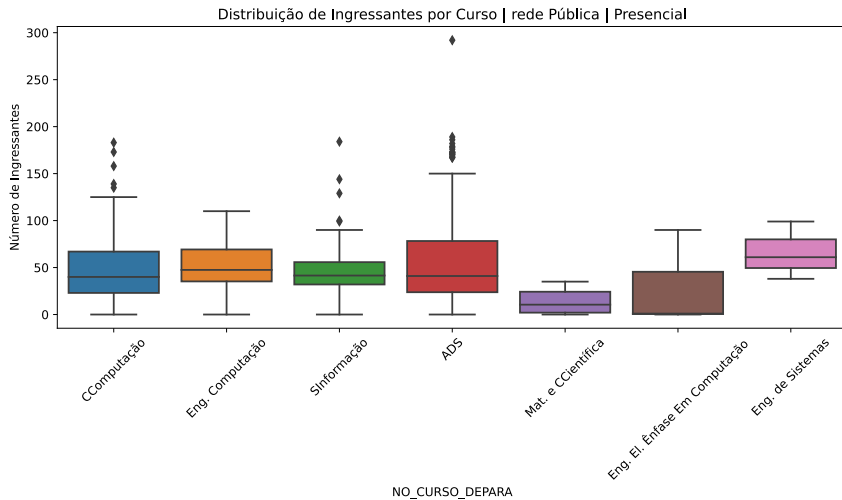
```
memory usage: 4.1+ MB
```

# Análise exploratória

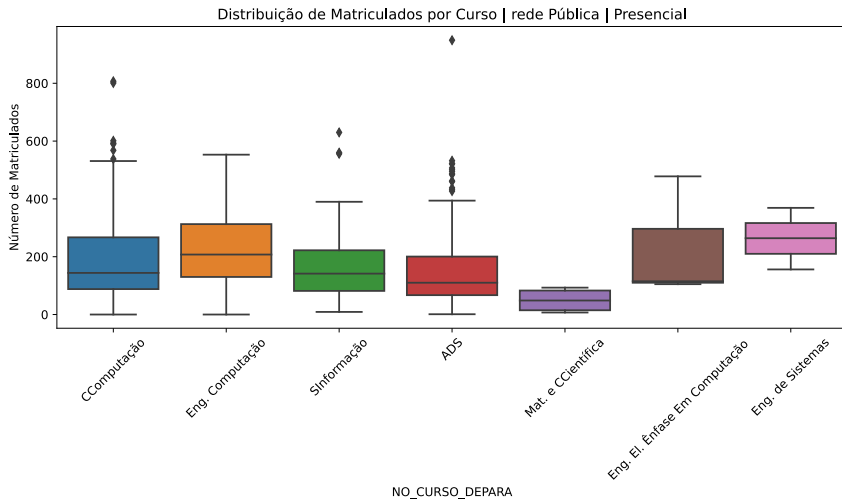
Método describe para variáveis numéricas (base total)

	QT_VG_TOTAL	QT_INSCRITO_TOTAL	QT_ING	QT_MAT	QT_CONC
<b>count</b>	19158	19158	19158	19158	19158
<b>mean</b>	42,84	36,20	10,22	19,61	2,22
<b>std</b>	754,03	468,54	37,86	68,31	10,61
<b>min</b>	0	0	0	0	0
<b>25%</b>	0	0	1	1	0
<b>50%</b>	0	0	2	2	0
<b>75%</b>	0	0	6	8	1
<b>max</b>	73280	32024	1794	3028	608

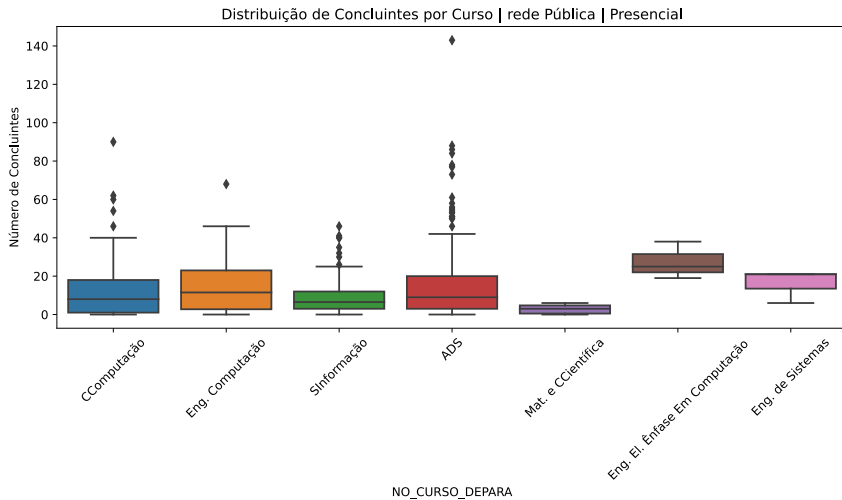
# Análise exploratória



# Análise exploratória

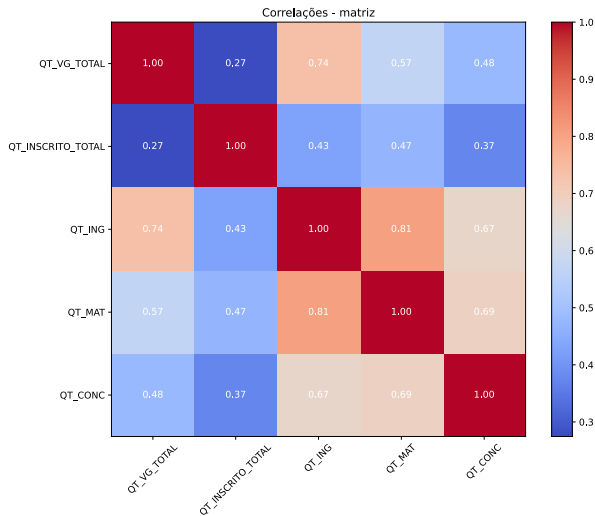


# Análise exploratória



# Análise exploratória

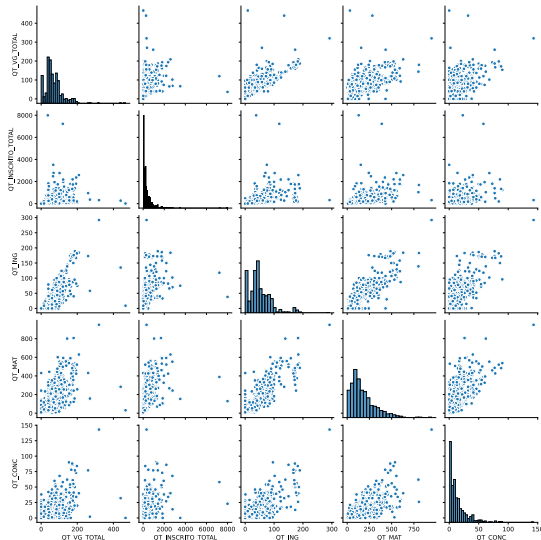
Correlações - matriz | rede Pública | Presencial





# Análise exploratória

Correlações - Pairplot | rede Pública | Presencial



# Considerações finais

Com este conjunto de dados, podemos:

- regionalizar (geograficamente) as análises;
- analisar cursos, modalidade, grau acadêmico e rede de ensino separadamente;
- ampliar o conjunto de dados com informações não utilizadas na primeira versão ou complementar com dados de anos anteriores.

# Obrigado!

Thanks! / ¡Gracias!

**Leonardo Cunha dos Santos**

`lattes.cnpq.br/5620610314140397`

`leonardo.cunha.santos@usp.br`

**Gabriel Francisco dos Santos Silva**

`gabfssilva@gmail.com`