



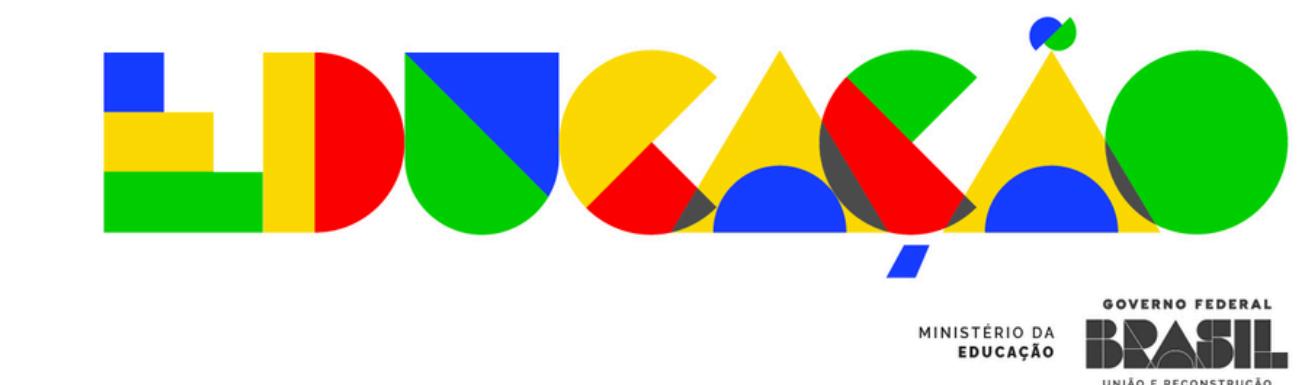
## CIÊNCIAS DE DADOS (IMD1151)

# PROJETO 01: CURSOS DE GRADUAÇÃO NO BRASIL

Francisco Willem Romão Moreira

Profº. Dr Heitor Medeiros Florencio

Profº. Dr Daniel Sabino Amorim de Araujo





# DESCRIÇÃO DO PROBLEMA

# DESCRIÇÃO DO PROBLEMA - Contexto

- O ensino superior no Brasil desempenha um papel de extrema importância no desenvolvimento econômico, social e cultural do país.
- Com uma ampla diversidade de cursos e instituições, o setor enfrenta desafios que necessitam de análises para promover melhorias.

# DESCRIÇÃO DO PROBLEMA - Objetivo Principal

- Este projeto se propõe a realizar uma análise abrangente dos cursos de graduação no Brasil.

# DESCRIÇÃO DO PROBLEMA - Questões a Serem Respondidas

- Como estão distribuídos os cursos de acordo com a região do país?
- De acordo com a modalidade de ensino (Presencial e EaD) como estão distribuídos os cursos em termos:
  - Regionais?
  - De Grau (Bacharelados, Tecnólogos, etc...)?
  - Situação (Ativos, Em Extinção e Extintos)?
- E os cursos de áreas correlacionadas a Tecnologia da Informação, como estão?



# DATASET

# DATASET - Origem dos Dados

Portal de  
**Dados Abertos**  
DO MINISTÉRIO DA EDUCAÇÃO

Buscar no portal 

Dados abertos do Governo Federal | Portal MEC

PÁGINA INICIAL > INDICADORES SOBRE ENSINO SUPERIOR > CURSOS DE GRADUAÇÃO DO BRASIL

Conheça o Plano de  
Dados Abertos do  
MEC para o biênio  
2020-2022

## Cursos de Graduação do Brasil

Publicado: Quinta, 29 de Dezembro de 2022, 15h17 | Última atualização em Quinta, 29 de Dezembro de 2022, 20h13 | Acessos: 4716

 Tweetar

 Compartilhar

### CONJUNTOS DE DADOS

Bolsa Formação

Brasil na Escola

EMTI

EPT

FIES

ID Estudantil

URL: [https://dadosabertos.mec.gov.br/images/conteudo/lnd-ensino-superior/2022//PDA\\_Dados\\_Cursos\\_Graduacao\\_Brasil.csv](https://dadosabertos.mec.gov.br/images/conteudo/lnd-ensino-superior/2022//PDA_Dados_Cursos_Graduacao_Brasil.csv) 

Detalhamento do quantitativo de Cursos de Graduação (Licenciatura, Bacharelado, Tecnológico, Sequencial e ABI - Área Básica de Ingresso) no Brasil por: código da Instituição de Educação Superior (IES); nome da IES; categoria da IES; organização acadêmica; código do curso; nome do curso; grau; área OCDE; modalidade de ensino (presencial ou EaD); situação do curso (ativo ou inativo); vagas autorizadas; carga horária; segmentadas por código do município (IBGE); município; UF; região.

Fonte: Portal de Dados Abertos MEC

# DATASET - Volume dos Dados



`df.shape`



`(902676, 18)`

# DATASET - Descrição das Colunas

```
▶ df.columns  
→ Index(['CODIGOIES', 'NOMEIES', 'CATEGORIAADMINISTRATIVA',  
        'ORGANIZACAOACADEMICA', 'CODIGOCURSO', 'NOMECURSO', 'GRAU',  
        'AREAOCDE', 'MODALIDADE', 'SITUACAOCURSO', 'QT_VAGAS_AUTORIZADAS',  
        'CARGAHORARIA', 'CODIGOAREAOCDECINE', 'AREAOCDECINE',  
        'CODIGOMUNICIPIO', 'MUNICIPIO', 'UF', 'REGIAO'],  
       dtype='object')
```

# DATASET - Qualidade dos Dados

▶ df.info()

```
→ <class 'pandas.core.frame.DataFrame'>
Index: 902068 entries, 0 to 902675
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   NOME_IES          902068 non-null   object  
 1   CATEGORIA_ADMINISTRATIVA 902068 non-null   object  
 2   ORGANIZACAO_ACADEMICA    902068 non-null   object  
 3   NOME_CURSO          902068 non-null   object  
 4   GRAU                902068 non-null   object  
 5   MODALIDADE          902068 non-null   object  
 6   SITUACAO_CURSO      902068 non-null   object  
 7   QT_VAGAS_AUTORIZADAS 902068 non-null   int64  
 8   CARGA_HORARIA        902068 non-null   int64  
 9   MUNICIPIO           902068 non-null   object  
 10  UF                  902068 non-null   object  
 11  REGIAO              902068 non-null   object  
dtypes: int64(2), object(10)
memory usage: 89.5+ MB
```

▶ df.isnull().sum()

```
→ NOME_IES                      0
    CATEGORIA_ADMINISTRATIVA     0
    ORGANIZACAO_ACADEMICA       0
    NOME_CURSO                  0
    GRAU                         0
    MODALIDADE                   0
    SITUACAO_CURSO              0
    QT_VAGAS_AUTORIZADAS        0
    CARGA_HORARIA                0
    MUNICIPIO                     0
    UF                           0
    REGIAO                        0
dtype: int64
```



# PRÉ-PROCESSAMENTO

# PRÉ-PROCESSAMENTO - Seleção de Variáveis

▶ df.nunique()

```
→  CODIGOIES          3706  
    NOMEIES            3672  
    CATEGORIAADMINISTRATIVA 6  
    ORGANIZACAOACADEMICA   6  
    CODIGOCURSO         86239  
    NOMECURSO           2206  
    GRAU                5  
    AREAOCDE             408  
    MODALIDADE            2  
    SITUACAOCURSO        3  
    QT_VAGAS_AUTORIZADAS 857  
    CARGA_HORARIA         3435  
    CODIGOAREAOCDECINE   423  
    AREAOCDECINE          519  
    CODIGOMUNICIPIO       3439  
    MUNICIPIO              3325  
    UF                   28  
    REGIAO               6  
    dtype: int64
```

- Note a diferença que existe na quantidade de dados únicos entre a coluna CODIGOIES e a coluna NOMEIES, explorando o dataset percebi que a coluna CODIGOIES, algumas vezes mostrava códigos diferentes quando se tratava de uma mesma instituição.
  - Dado essa situação, e também pela natureza do trabalho não vi o porque de utilizar a coluna CODIGOIES.
- A mesma situação acontece com as colunas CODIGO\_MUNICIPIO e MUNICIPIO. CODIGO\_CURSO e NOME\_CURSO. Indicando possíveis erros no preenchimento do dataset.

# PRÉ-PROCESSAMENTO - Seleção de Variáveis

- A coluna AREA\_OCDE e AREA\_OCDE\_CINE consistia em vários dados muito diversos e que não representava com abrangência a área no sentido geral, e sim em sentido mais específico.
    - **Exemplo:** ao invés da área ser educação, eles colocam Pedagogia, assim como, Educação Física ao invés de ser da área saúde, é a própria educação física. Então para fins de avaliação não me ajudará a responder as perguntas de minhas análises futuras.

```
df['AREA_OCDE'].unique()  
array(['Agronomia',  
       'Formação de professor de língua/literatura vernácula (português)',  
       'Formação de professor de geografia', 'Educação física', nan,  
       'Análise e Desenvolvimento de Sistemas (Tecnólogo)',  
       'Gestão da informação', 'Formação de professor de dança',  
       'Ciências contábeis', 'Administração',  
       'Formação de professor de artes visuais',  
       'Saúde e segurança no trabalho',  
       'Formação de professor de língua/literatura estrangeira moderna',  
       'Formação de professor de língua/literatura vernácula e língua estrangeira moderna',  
       'Engenharia de produção',  
       'Enfermagem', 'Física', 'Química', 'Biol
```

```
df['AREA_OCDE_CINE'].unique()  
array(['Agronomia',  
       'Formação de professor de língua/literatura vernácula (português)',  
       'Formação de professor de geografia', 'Educação física',  
       'Serviços penais', 'Gestão comercial',  
       'Análise e Desenvolvimento de Sistemas (Tecnólogo)',  
       'Gestão da informação', 'Formação de professor de dança',  
       'Agrocomputação', 'Ciências contábeis', 'Administração',  
       'Formação de professor de artes visuais',  
       'Educação infantil', 'Educação primária', 'Educação secundária'])
```

# PRÉ-PROCESSAMENTO - Seleção de Variáveis

```
▶ df = df.drop(columns=['CODIGOIES',
                         'CODIGOCURSO',
                         'AREAOCDE',
                         'CODIGOAREAOCDECINE',
                         'AREAOCTCINE',
                         'CODIGOMUNICIPIO'])
```

```
▶ df.shape
```

```
→ (902676, 12)
```

```
▶ df.columns
```

```
→ Index(['NOMEIES', 'CATEGORIAADMINISTRATIVA', 'ORGANIZACAOACADEMICA',
          'NOMECURSO', 'GRAU', 'MODALIDADE', 'SITUACAOCURSO',
          'QT_VAGAS_AUTORIZADAS', 'CARGA_HORARIA', 'MUNICIPIO', 'UF', 'REGIAO'],
          dtype='object')
```

# PRÉ-PROCESSAMENTO - Tipificação de Dados

```
df.dtypes
```

```
NOMEIES          object
CATEGORIAADMINISTRATIVA    object
ORGANIZACAOACADEMICA      object
NOMECURSO                 object
GRAU                       object
MODALIDADE                 object
SITUACAOCURSO              object
QT_VAGAS_AUTORIZADAS       int64
CARGA_HORARIA               int64
MUNICIPIO                  object
UF                          object
REGIAO                      object
dtype: object
```

- Os tipos de dados estão adequados para o que sua respectiva coluna representa.

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

- A limpeza feita serviu apenas para análises dos cursos de TI a despeito da quantidade de vagas e da carga horária.

```
▶ df_ea = df[df['SITUACAO_CURSO'] == 'Em atividade'] # df_ea = dataframe apenas com cursos em atividades
```



Home/Notícias/Explorando os cursos de graduação na área de tecnologia: conceito e opções

## NOTÍCIAS

### EXPLORANDO OS CURSOS DE GRADUAÇÃO NA ÁREA DE TECNOLOGIA: CONCEITO E OPÇÕES

02/08/2023

- Com base nessa fonte filtrei os cursos da área de TI mais procurados e com maior relevância no Brasil.

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

- A limpeza feita serviu apenas para análises dos cursos de TI a despeito da quantidade de vagas e da carga horária.

```
cursos_ti = [  
    'TECNOLOGIA DA INFORMAÇÃO',  
    'CIÊNCIA DA COMPUTAÇÃO',  
    'CIÊNCIAS DA COMPUTAÇÃO',  
    'ENGENHARIA DA COMPUTAÇÃO',  
    'ENGENHARIA DE COMPUTAÇÃO',  
    'ENGENHARIA DE SOFTWARE',  
    'ANÁLISE E DESENVOLVIMENTO DE SISTEMAS',  
    'SISTEMAS DE INFORMAÇÃO',  
    'GESTÃO DA TECNOLOGIA DA INFORMAÇÃO',  
    'ENGENHARIA DE TELECOMUNICAÇÕES',  
    'REDES DE COMPUTADORES',  
    'ENGENHARIA MECATRÔNICA',  
    'ENGENHARIA DE CONTROLE E AUTOMAÇÃO',  
    'SISTEMAS PARA INTERNET',  
    'BANCO DE DADOS',  
    'JOGOS DIGITAIS'  
]
```

```
dfti = df_ea[df_ea['NOME_CURSO'].isin(cursos_ti)] # dfti é o dataframe com os cursos de TI Em Atividade listados acima
```

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

- Analisando com unique todos os cursos, percebi que alguns cursos como o de Ciência da Computação e Engenharia da Computação divergia no nome, então apliquei um replace para todos esses casos estabelecendo o padrão:

```
▶ dfti['NOME_CURSO'] = dfti['NOME_CURSO'].replace('CIÊNCIAS DA COMPUTAÇÃO', 'CIÊNCIA DA COMPUTAÇÃO')
dfti['NOME_CURSO'] = dfti['NOME_CURSO'].replace('CIÊNCIAS DE COMPUTAÇÃO', 'CIÊNCIA DA COMPUTAÇÃO')

dfti['NOME_CURSO'] = dfti['NOME_CURSO'].replace('ENGENHARIA DA COMPUTAÇÃO', 'ENGENHARIA DE COMPUTAÇÃO')

dfti['NOME_CURSO'] = dfti['NOME_CURSO'].replace('GESTÃO EM TECNOLOGIA DA INFORMAÇÃO', 'GESTÃO DA TECNOLOGIA DA INFORMAÇÃO')
dfti['NOME_CURSO'] = dfti['NOME_CURSO'].replace('GESTÃO DE TECNOLOGIA DA INFORMAÇÃO', 'GESTÃO DA TECNOLOGIA DA INFORMAÇÃO')
```

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

- Primeiro analisei possíveis ruídos e outliers na coluna QT\_VAGAS\_AUTORIZADAS.
- Boxplot QT\_VAGAS\_AUTORIZADAS

```
▶ dfti[['QT_VAGAS_AUTORIZADAS', 'NOME_CURSO', 'NOMEIES', 'MODALIDADE', 'MUNICIPIO']].sort_values(by=['QT_VAGAS_AUTORIZADAS'], ascending=False).head(10)
```

	QT_VAGAS_AUTORIZADAS	NOME_CURSO	NOMEIES	MODALIDADE	MUNICIPIO
489344	73260	REDES DE COMPUTADORES	UNIVERSIDADE PAULISTA	Educação a Distância	Itapetininga
801603	73260	REDES DE COMPUTADORES	UNIVERSIDADE PAULISTA	Educação a Distância	Niterói
81190	73260	GESTÃO DA TECNOLOGIA DA INFORMAÇÃO	UNIVERSIDADE PAULISTA	Educação a Distância	Tailândia
364854	73260	REDES DE COMPUTADORES	UNIVERSIDADE PAULISTA	Educação a Distância	Guarantã do Norte
550990	73260	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE PAULISTA	Educação a Distância	Cruzeiro do Sul
364790	73260	REDES DE COMPUTADORES	UNIVERSIDADE PAULISTA	Educação a Distância	Ibaté
458034	73260	GESTÃO DA TECNOLOGIA DA INFORMAÇÃO	UNIVERSIDADE PAULISTA	Educação a Distância	Manicoré
803113	73260	REDES DE COMPUTADORES	UNIVERSIDADE PAULISTA	Educação a Distância	Serra Negra
81734	73260	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE PAULISTA	Educação a Distância	Sumaré
81870	73260	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE PAULISTA	Educação a Distância	Londrina

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

```
▶ dfti[['QT_VAGAS_AUTORIZADAS', 'NOME_CURSO', 'NOMEIES', 'MODALIDADE', 'MUNICIPIO']].sort_values(by=['QT_VAGAS_AUTORIZADAS'], ascending=True).head(10)
```

QT_VAGAS_AUTORIZADAS	NOME_CURSO	NOMEIES	MODALIDADE	MUNICIPIO
896478	0 ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	Universidade do Estado de Mato Grosso Carlos A...	Educação Presencial	Nova Mutum
508215	0 SISTEMAS DE INFORMAÇÃO	UNIVERSIDADE DE SANTA CRUZ DO SUL	Educação Presencial	Sobradinho
440217	0 ENGENHARIA DE COMPUTAÇÃO	UNIVERSIDADE DO ESTADO DO AMAZONAS	Educação Presencial	Manaus
127965	0 SISTEMAS DE INFORMAÇÃO	UNIVERSIDADE DE SANTA CRUZ DO SUL	Educação Presencial	Venâncio Aires
213479	0 REDES DE COMPUTADORES	Fatec Cruzeiro - Prof. Waldomiro May	Educação Presencial	Cruzeiro
5134	0 CIÊNCIA DA COMPUTAÇÃO	UNIVERSIDADE ESTADUAL DO CEARÁ	Educação Presencial	Iguatu
611222	7 ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE DO ESTADO DO AMAZONAS	Educação Presencial	Presidente Figueiredo
509922	10 ENGENHARIA DE CONTROLE E AUTOMAÇÃO	Instituto Universitário Una de Catalão	Educação Presencial	Catalão
895604	20 ENGENHARIA DE TELECOMUNICAÇÕES	UNIVERSIDADE LA SALLE	Educação Presencial	Canoas
841735	20 ENGENHARIA DE COMPUTAÇÃO	UNIVERSIDADE ESTADUAL DO RIO GRANDE DO SUL	Educação Presencial	Guaíba

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

- Pesquisando sobre a distribuição da quantidade de vagas em cursos no Brasil, não achei nenhuma informação.
- Apesar de serem cursos EaD, não posso confirmar se isso é possível.
- Portanto, faço o seguinte tratamento:

```
▶ dfti_limpo_qtv = dfti[
    (dfti['QT_VAGAS_AUTORIZADAS'] > dfti['QT_VAGAS_AUTORIZADAS'].quantile(0.001)) &
    (dfti['QT_VAGAS_AUTORIZADAS'] < dfti['QT_VAGAS_AUTORIZADAS'].quantile(0.600))]
# dfti_limpo_qtv = dataframe ti em atividade limpo para quantidade de vagas
```

```
[ ] dfti_limpo_qtv['QT_VAGAS_AUTORIZADAS'] = np.where(
    dfti_limpo_qtv['QT_VAGAS_AUTORIZADAS'] >= dfti_limpo_qtv['QT_VAGAS_AUTORIZADAS'].quantile(0.600),
    dfti_limpo_qtv[(dfti_limpo_qtv['QT_VAGAS_AUTORIZADAS'] > dfti_limpo_qtv['QT_VAGAS_AUTORIZADAS'].quantile(0.001)) & (dfti_limpo_qtv['QT_VAGAS_AUTORIZADAS'])]
```

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

- Resultado do tratamento:

```
▶ dfti_limpo_qtv[['QT_VAGAS_AUTORIZADAS', 'NOME_CURSO', 'NOMEIES', 'MODALIDADE', 'MUNICIPIO']].sort_values(by=['QT_VAGAS_AUTORIZADAS'], ascending=False).head()
```

	QT_VAGAS_AUTORIZADAS	NOME_CURSO	NOMEIES	MODALIDADE	MUNICIPIO
740049	990.0	SISTEMAS DE INFORMAÇÃO	UNIVERSIDADE DE TAUBATÉ	Educação a Distância	São José dos Campos
265250	990.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE DE TAUBATÉ	Educação a Distância	Recife
252289	990.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE DE TAUBATÉ	Educação a Distância	Paraisópolis
104473	990.0	SISTEMAS DE INFORMAÇÃO	UNIVERSIDADE DE TAUBATÉ	Educação a Distância	Piquete
483939	990.0	SISTEMAS DE INFORMAÇÃO	UNIVERSIDADE DE TAUBATÉ	Educação a Distância	São Bento do Sapucaí

```
[ ] dfti_limpo_qtv[['QT_VAGAS_AUTORIZADAS', 'NOME_CURSO', 'NOMEIES', 'MODALIDADE', 'MUNICIPIO']].sort_values(by=['QT_VAGAS_AUTORIZADAS'], ascending=True).head()
```

	QT_VAGAS_AUTORIZADAS	NOME_CURSO	NOMEIES	MODALIDADE	MUNICIPIO
320254	32.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNO...	Educação Presencial	Venâncio Aires
413964	32.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE DO ESTADO DO AMAZONAS	Educação Presencial	Maués
732415	32.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE DO ESTADO DO AMAZONAS	Educação Presencial	Boca do Acre
145728	32.0	SISTEMAS DE INFORMAÇÃO	INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNO...	Educação Presencial	Colatina
412958	32.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE DO ESTADO DO AMAZONAS	Educação Presencial	Humaitá

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

- Ruídos e outliers em CARGA\_HORARIA.
- Boxplot CARGA\_HORARIA

```
▶ dfti[['CARGA_HORARIA', 'NOME_CURSO', 'NOMEIES', 'MODALIDADE', 'MUNICIPIO']].sort_values(by=['CARGA_HORARIA'], ascending=False).head()
```

	CARGA_HORARIA	NOME_CURSO	NOMEIES	MODALIDADE	MUNICIPIO
311124	8600	ENGENHARIA DE CONTROLE E AUTOMAÇÃO	FACULDADE DE TECNOLOGIA PENTÁGONO	Educação Presencial	Santo André
517409	8000	ENGENHARIA DE COMPUTAÇÃO	CENTRO UNIVERSITÁRIO MAURÍCIO DE NASSAU DE BAR...	Educação Presencial	Barreiras
677004	7760	CIÊNCIA DA COMPUTAÇÃO	Centro Universitário Vértice	Educação Presencial	Matipó
62208	7560	ENGENHARIA DE CONTROLE E AUTOMAÇÃO	CENTRO UNIVERSITÁRIO RITTER DOS REIS	Educação Presencial	Canoas
848008	7560	ENGENHARIA DE COMPUTAÇÃO	CENTRO UNIVERSITÁRIO RITTER DOS REIS	Educação Presencial	Porto Alegre

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

```
▶ dfti[['CARGA_HORARIA', 'NOME_CURSO', 'NOMEIES', 'MODALIDADE', 'MUNICIPIO']].sort_values(by=['CARGA_HORARIA'], ascending=True).head()
```

	CARGA_HORARIA	NOME_CURSO	NOMEIES	MODALIDADE	MUNICIPIO
453746	0	SISTEMAS PARA INTERNET	Faculdade de Tecnologia Senac Curitiba Portão	Educação Presencial	Curitiba
58041	0	BANCO DE DADOS	Faculdade de Tecnologia Senac Curitiba Portão	Educação Presencial	Curitiba
5134	0	CIÊNCIA DA COMPUTAÇÃO	UNIVERSIDADE ESTADUAL DO CEARÁ	Educação Presencial	Iguatu
48389	0	REDES DE COMPUTADORES	Faculdade de Tecnologia Senac Curitiba Portão	Educação Presencial	Curitiba
262468	1700	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	FACULDADE VISCONDE DE CAIRÚ	Educação a Distância	Salvador

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos



Cursos de graduação ▾

Carreira

Enem e Vestibular

Materiais e Testes Gratuitos

INSCREVA-SE



**Carga horária da faculdade: como funciona e o que o MEC exige?**



- Existem ruídos nesses dados.
- Conforme o site [Unopar](#) a carga horária de uma faculdade, em geral, tem duração mínima exigida pelo Ministério da Educação de:
  - 2.400 horas para bacharelados,
  - para tecnólogos entre 1.600 a 2.000 horas,
  - enquanto graduações mais longas como Medicina e Odontologia, podem chegar a 7.200 horas

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

- Levando em conta que os cursos mais longos de TI são as Engenharias, e elas tem em média 3800 horas, então estabeleci esse padrão.

```
▶ dfti_limpo_ch = dfti[
    (dfti['CARGA_HORARIA'] > dfti['CARGA_HORARIA'].quantile(0.001)) &
    (dfti['CARGA_HORARIA'] < dfti['CARGA_HORARIA'].quantile(0.985))]

# dfti_limpo_ch = dataframe ti em atividade limpo para carga horária

[ ] dfti_limpo_ch['CARGA_HORARIA'] = np.where(
    dfti_limpo_ch['CARGA_HORARIA'] >= dfti_limpo_ch['CARGA_HORARIA'].quantile(0.985),
    dfti_limpo_ch[(dfti_limpo_ch['CARGA_HORARIA'] > dfti_limpo_ch['CARGA_HORARIA'].quantile(0.001))
    & (dfti_limpo_ch['CARGA_HORARIA'] < dfti_limpo_ch['CARGA_HORARIA'].quantile(0.985))]['CARGA_HORARIA'].mean(),
    dfti_limpo_ch['CARGA_HORARIA'])
```

# PRÉ-PROCESSAMENTO - Tratamento de Outliers e Ruídos

```
▶ dfti_limpo_ch[['CARGA_HORARIA', 'NOME_CURSO', 'NOMEIES', 'MODALIDADE', 'MUNICIPIO']].sort_values(by=['CARGA_HORARIA'], ascending=False).head()
```

	CARGA_HORARIA	NOME_CURSO		NOMEIES	MODALIDADE	MUNICIPIO
664823	3894.0	ENGENHARIA DE COMPUTAÇÃO	UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL	Educação Presencial	Campo Grande	
895535	3885.0	ENGENHARIA DE COMPUTAÇÃO	UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE	Educação Presencial		Natal
174548	3885.0	ENGENHARIA DE TELECOMUNICAÇÕES	UNIVERSIDADE FEDERAL DE SANTA MARIA	Educação Presencial		Santa Maria
849511	3885.0	ENGENHARIA DE COMPUTAÇÃO	PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JAN...	Educação Presencial		Rio de Janeiro
197260	3884.0	ENGENHARIA DE CONTROLE E AUTOMAÇÃO	INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNO...	Educação Presencial		Itumbiara

```
[ ] dfti_limpo_ch[['CARGA_HORARIA', 'NOME_CURSO', 'NOMEIES', 'MODALIDADE', 'MUNICIPIO']].sort_values(by=['CARGA_HORARIA'], ascending=True).head()
```

	CARGA_HORARIA	NOME_CURSO		NOMEIES	MODALIDADE	MUNICIPIO
233590	2001.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE PRESBITERIANA MACKENZIE	Educação a Distância		Bauru
338841	2001.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE PRESBITERIANA MACKENZIE	Educação a Distância		Ribeirão Preto
86236	2001.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE PRESBITERIANA MACKENZIE	Educação a Distância		São Luís
573130	2001.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE PRESBITERIANA MACKENZIE	Educação a Distância		São Paulo
311331	2001.0	ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	UNIVERSIDADE PRESBITERIANA MACKENZIE	Educação a Distância		Aracaju

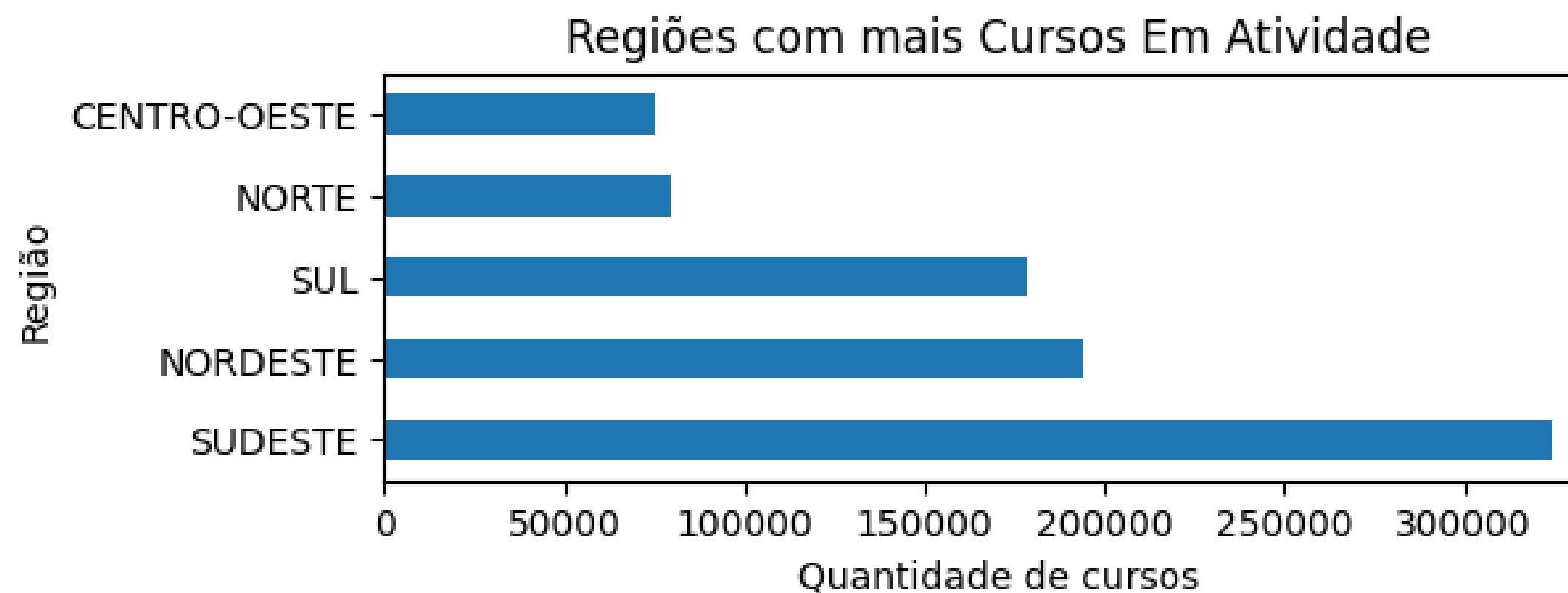
# PRÉ-PROCESSAMENTO - Transformação

- Sobre técnicas de transformação em si como discretização, binarização, codificação, normalização e etc... até o momento não vi como necessário.
- A depender do algoritmo de Machine Learning que irei aplicar no futuro, é muito provável que farei uso de alguma técnica de transformação.

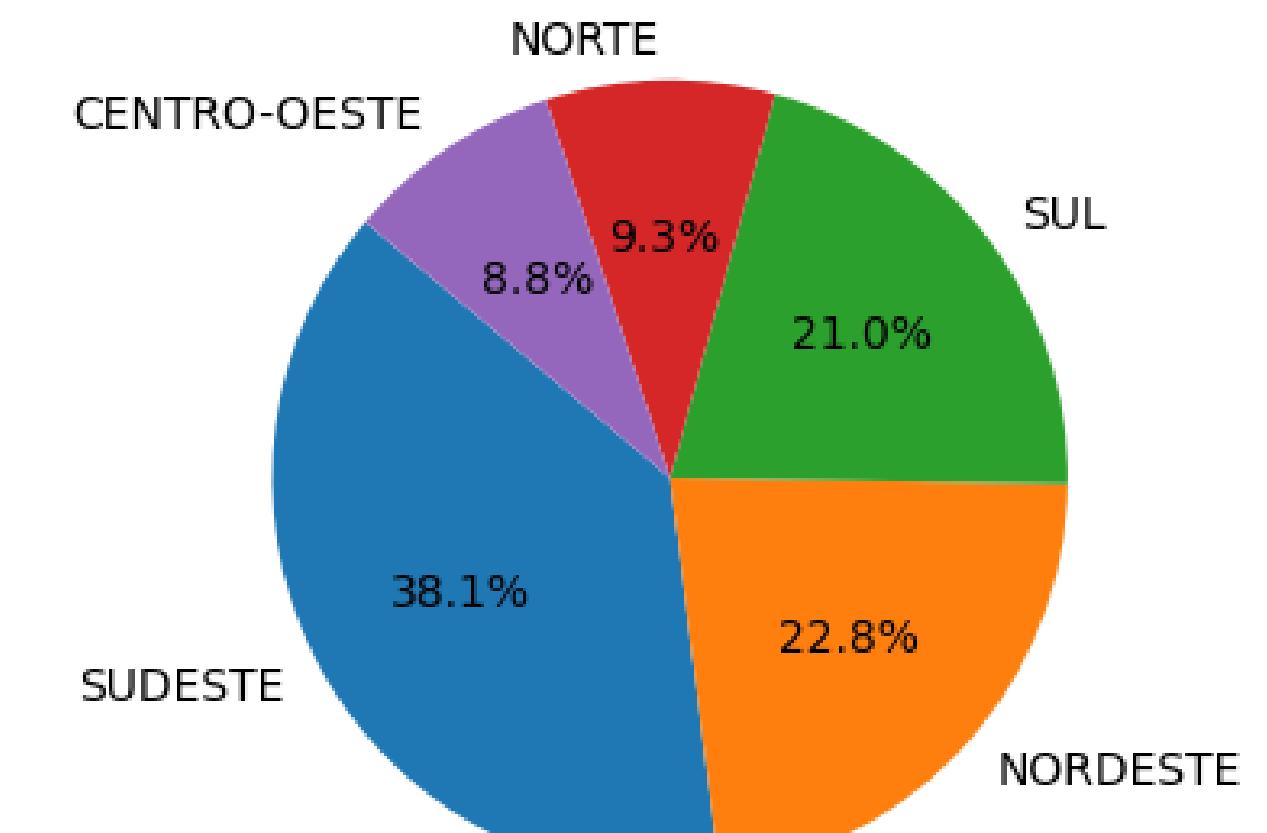


# ANÁLISE EXPLORATÓRIA DE DADOS

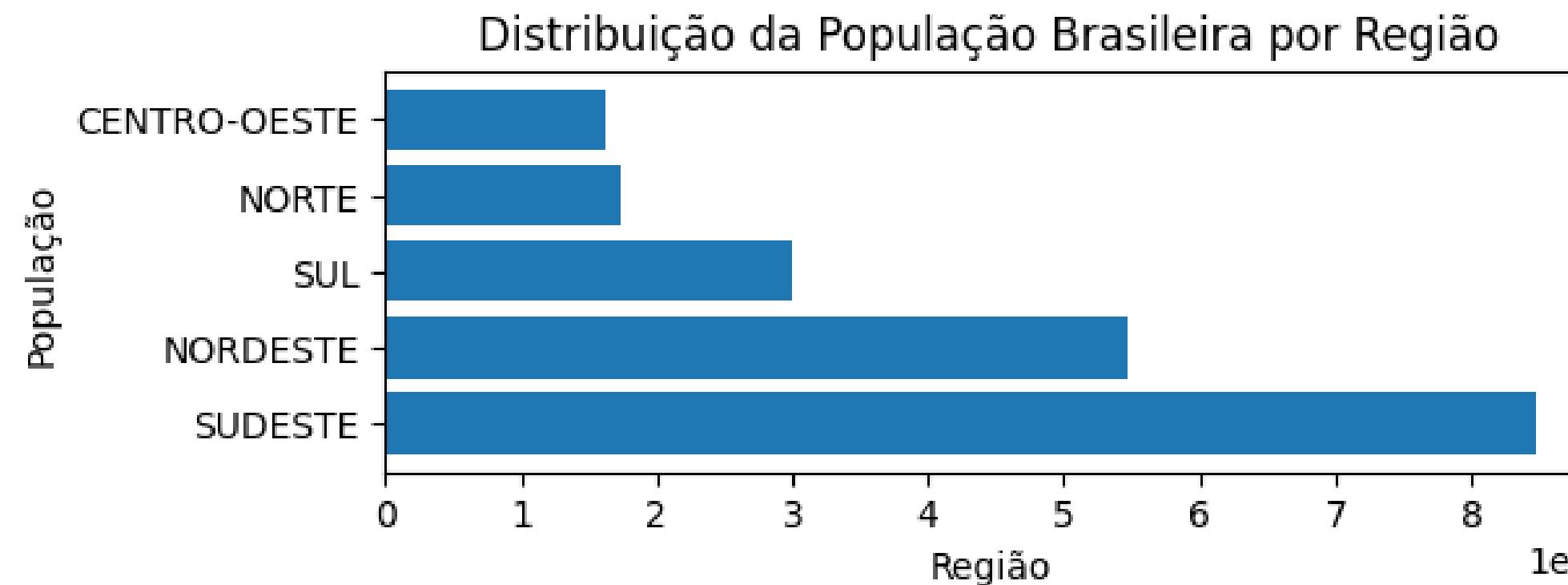
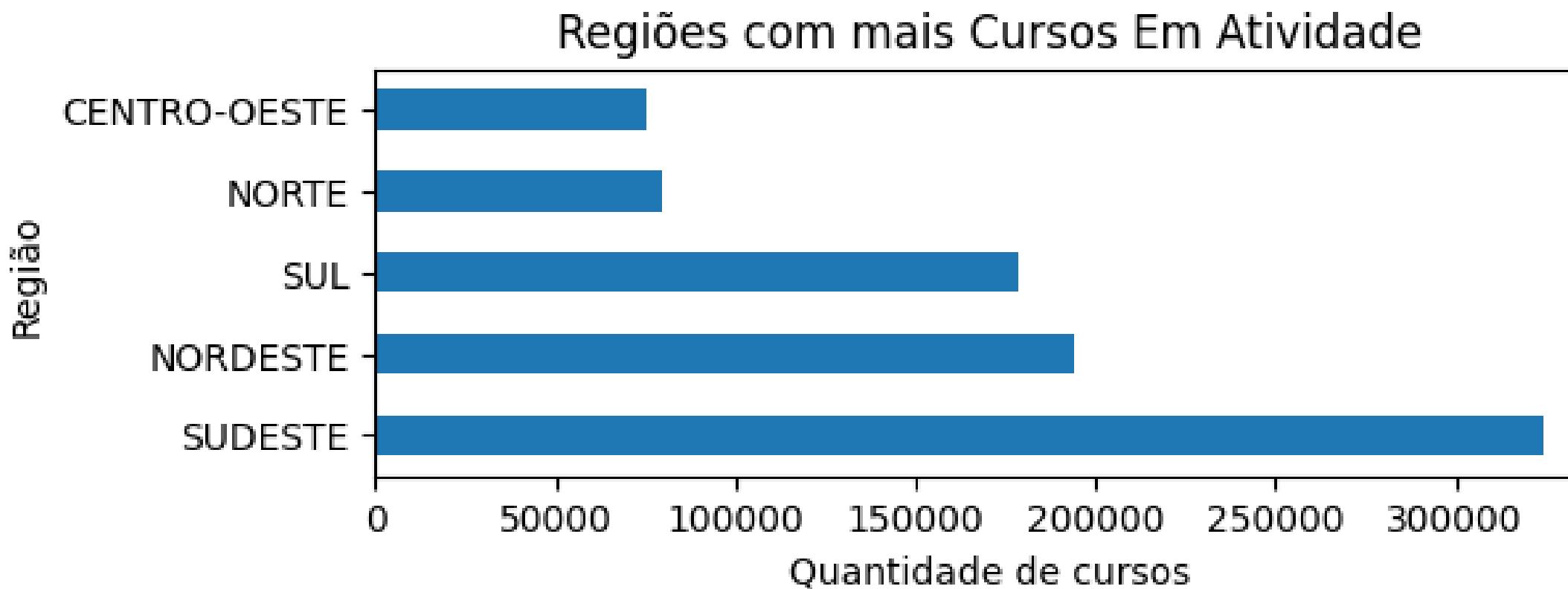
# ANÁLISE EXPLORATÓRIA DE DADOS - Distribuição Geográfica dos Cursos



Distribuição dos Cursos Em Atividade por Região

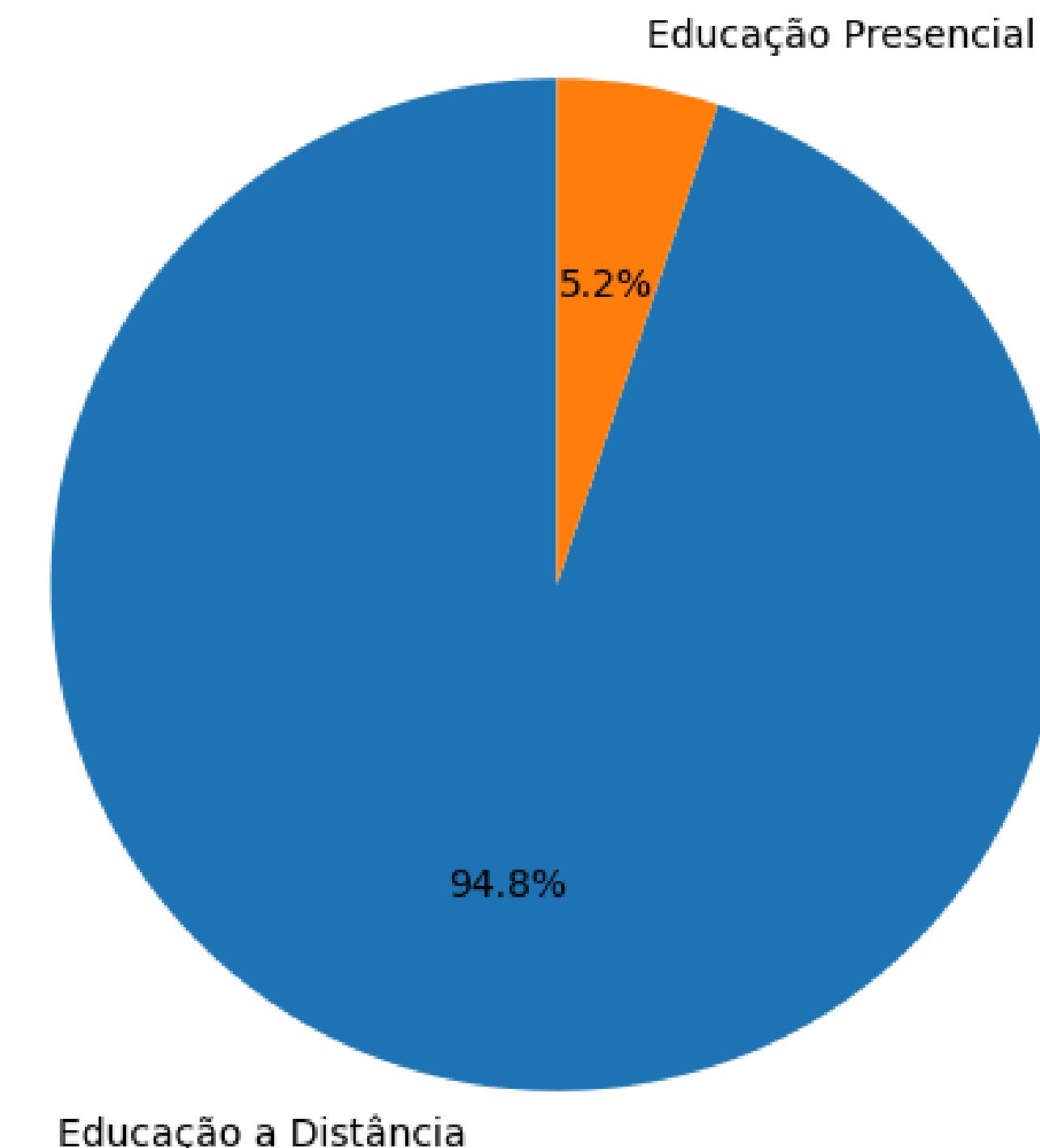


# ANÁLISE EXPLORATÓRIA DE DADOS - Distribuição Geográfica dos Cursos



# ANÁLISE EXPLORATÓRIA DE DADOS - Modalidade de Ensino

Distribuição dos Cursos Em Atividade por Modalidade de Ensino

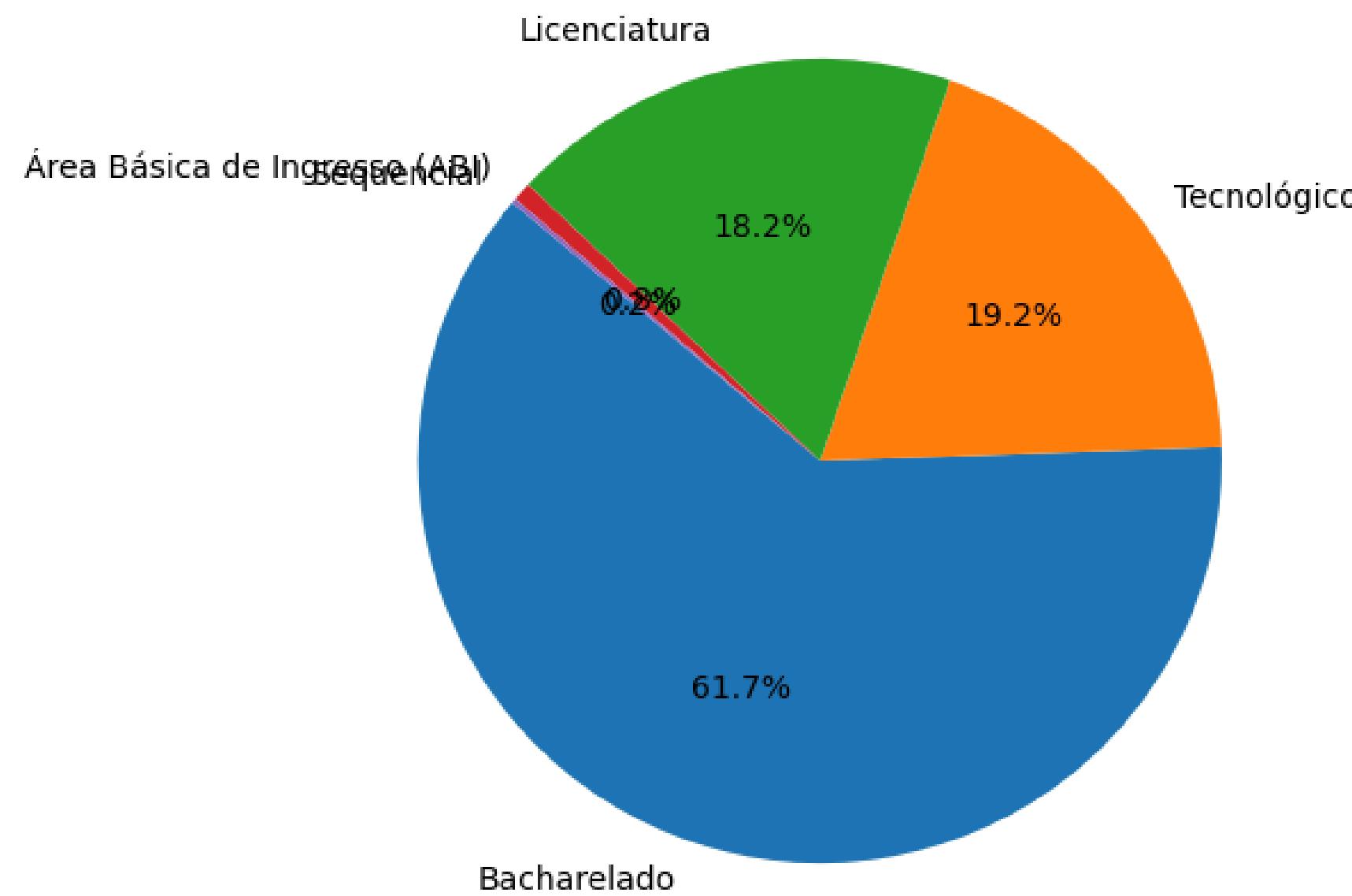


```
▶ df_ea['MODALIDADE'].value_counts()
```

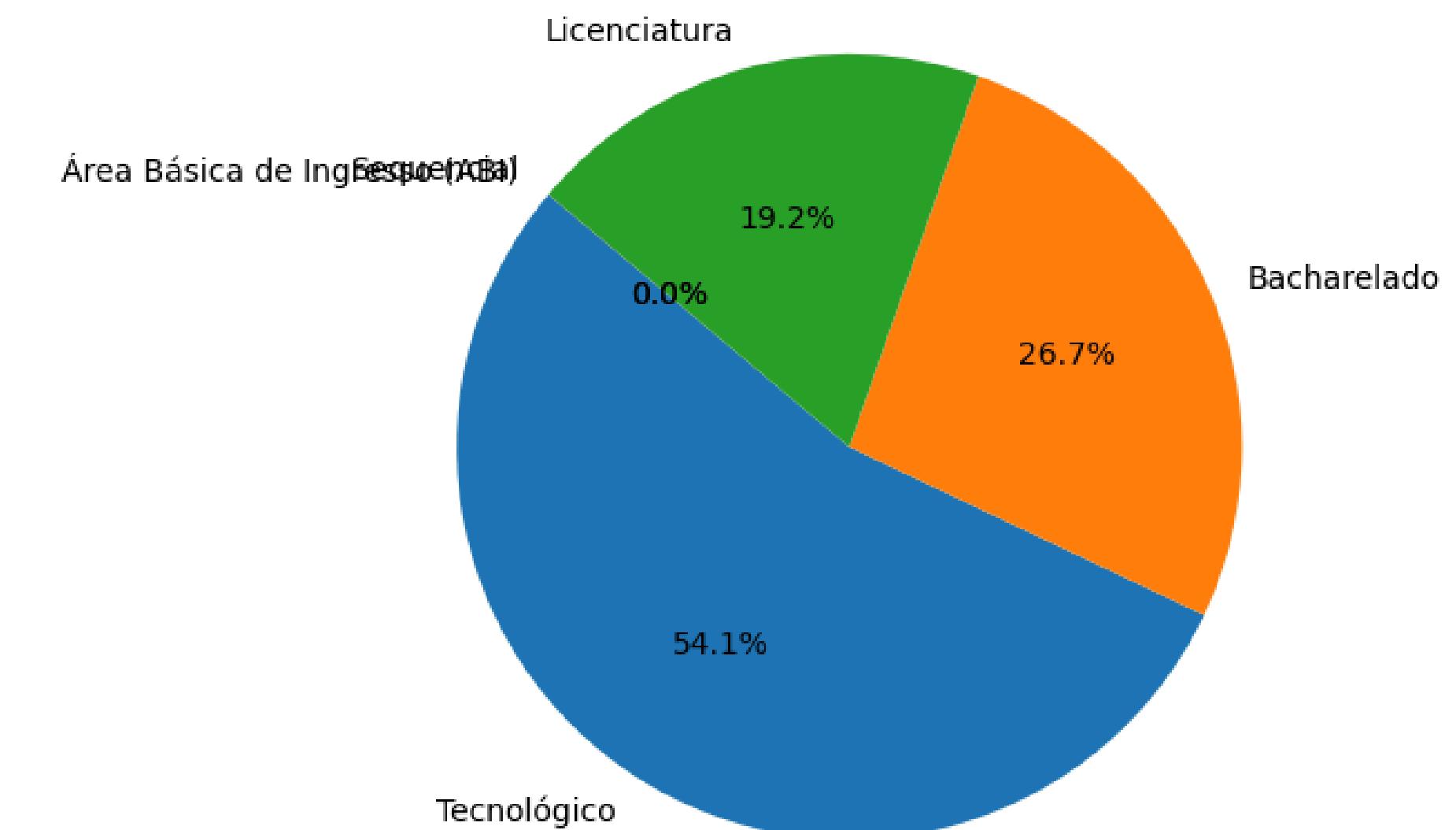
```
→ MODALIDADE
  Educação a Distância    808360
  Educação Presencial      43968
  Name: count, dtype: int64
```

# ANÁLISE EXPLORATÓRIA DE DADOS - Modalidade de Ensino

Distribuição de Cursos Presenciais por Grau



Distribuição de Cursos EaD por Grau



# ANÁLISE EXPLORATÓRIA DE DADOS - Modalidade de Ensino

- Quais cursos são majoritários para cada modalidade?

```
▶ df_ea[df_ea['MODALIDADE'] == 'Educação Presencial']['NOME_CURSO'].value_counts().head(30)
```

NOME_CURSO	count
ADMINISTRAÇÃO	2275
DIREITO	1965
PEDAGOGIA	1884
EDUCAÇÃO FÍSICA	1650
CIÊNCIAS CONTÁBEIS	1494
ENFERMAGEM	1376
PSICOLOGIA	1260
ENGENHARIA CIVIL	1181
FISIOTERAPIA	983
CIÊNCIAS BIOLÓGICAS	896
ENGENHARIA DE PRODUÇÃO	879
NUTRIÇÃO	812
FARMÁCIA	793
GESTÃO DE RECURSOS HUMANOS	783
ARQUITETURA E URBANISMO	752
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	709
BIOMEDICINA	696
MATEMÁTICA	685
ODONTOLOGIA	659
ENGENHARIA MECÂNICA	638
ENGENHARIA ELÉTRICA	608
LOGÍSTICA	558
MEDICINA VETERINÁRIA	533
HISTÓRIA	507
SISTEMAS DE INFORMAÇÃO	475
SERVIÇO SOCIAL	459
ESTÉTICA E COSMÉTICA	436
CIÊNCIA DA COMPUTAÇÃO	434
QUÍMICA	432
GEOGRAFIA	427

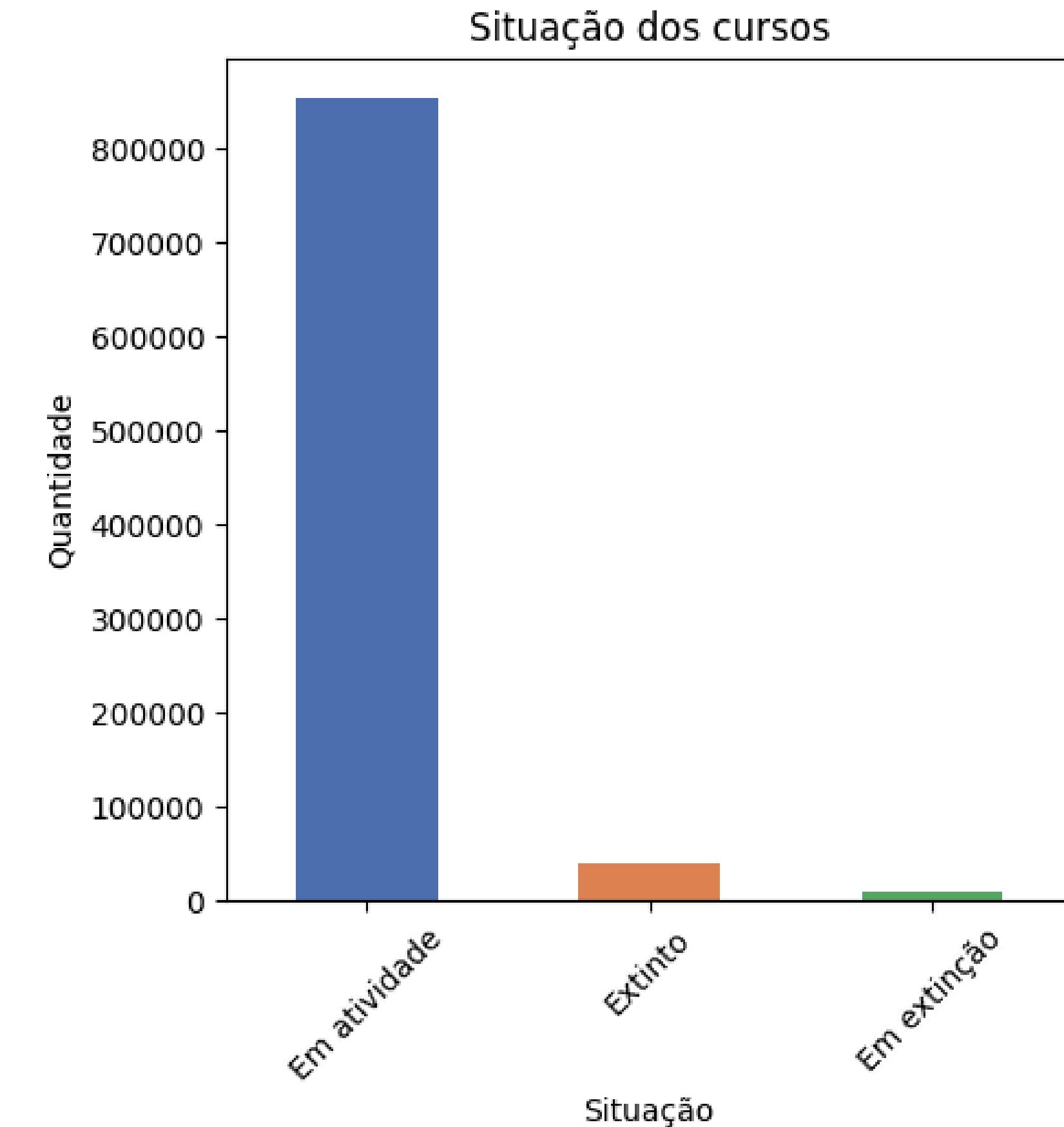
```
▶ df_ea[df_ea['MODALIDADE'] == 'Educação a Distância']['NOME_CURSO'].value_counts().head(30)
```

NOME_CURSO	count
ADMINISTRAÇÃO	17426
PEDAGOGIA	17303
EDUCAÇÃO FÍSICA	16363
CIÊNCIAS CONTÁBEIS	15092
GESTÃO DE RECURSOS HUMANOS	14884
HISTÓRIA	14035
MATEMÁTICA	13346
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS	13087
GESTÃO PÚBLICA	12943
LOGÍSTICA	12861
PROCESSOS GERENCIAIS	12851
GESTÃO FINANCEIRA	12762
GEOGRAFIA	12134
GESTÃO COMERCIAL	12105
CIÊNCIAS BIOLÓGICAS	11759
MARKETING	11510
GESTÃO AMBIENTAL	11418
SERVIÇO SOCIAL	11281
FILOSOFIA	10167
ENGENHARIA DE PRODUÇÃO	9992
GESTÃO DA TECNOLOGIA DA INFORMAÇÃO	9868
GESTÃO HOSPITALAR	9589
ARTES VISUAIS	8917
COMÉRCIO EXTERIOR	8663
GESTÃO DA QUALIDADE	8341
CIÊNCIAS ECONÔMICAS	8253
QUÍMICA	7295
DESIGN DE INTERIORES	7248
SEGURANÇA PÚBLICA	7193
NEGÓCIOS IMOBILIÁRIOS	7181

# ANÁLISE EXPLORATÓRIA DE DADOS - Situação dos Cursos

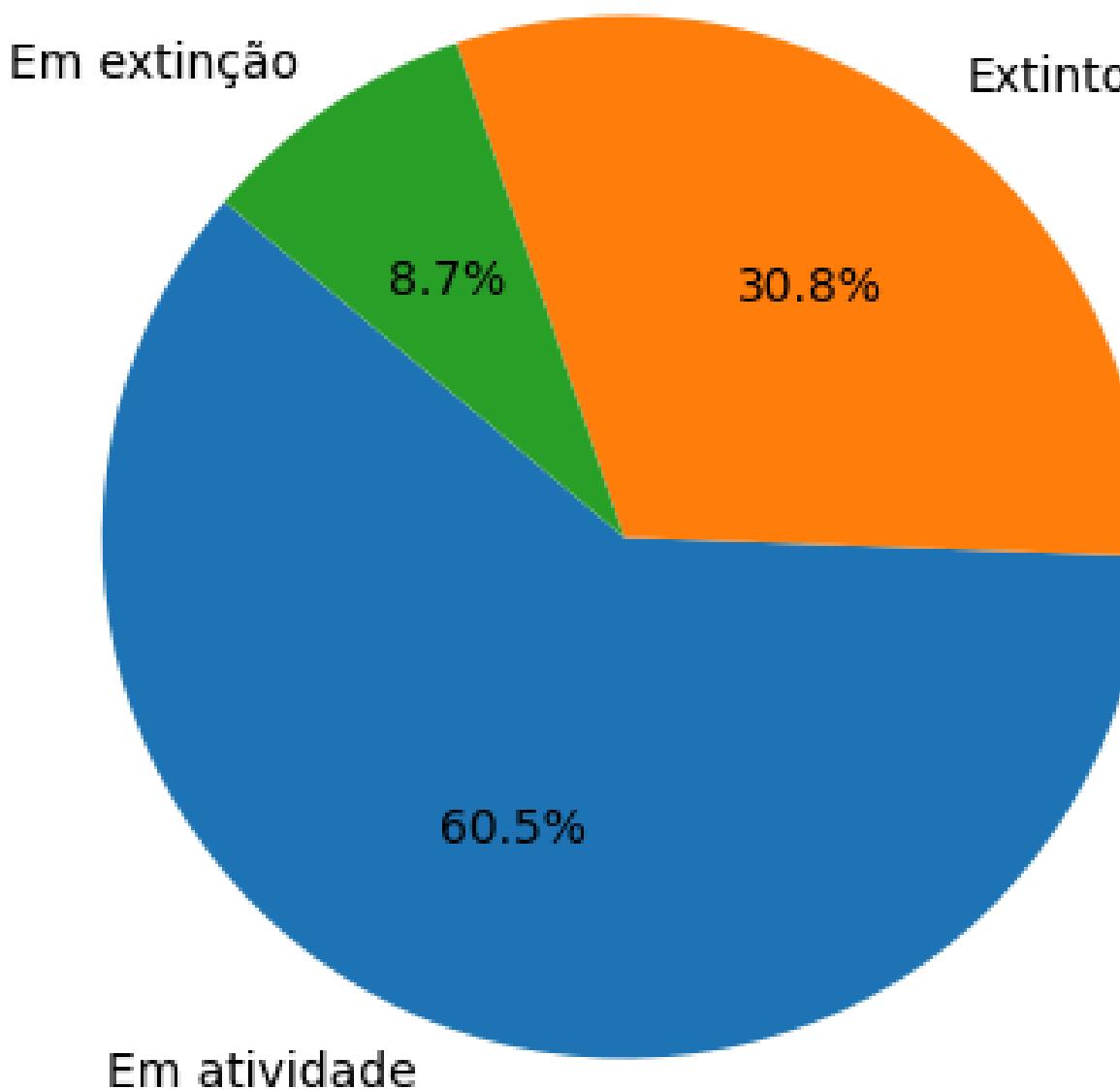
```
▶ df['SITUACAO_CURSO'].value_counts()
```

```
→ SITUACAO_CURSO
  Em atividade      852328
  Extinto          39479
  Em extinção     10261
  Name: count, dtype: int64
```

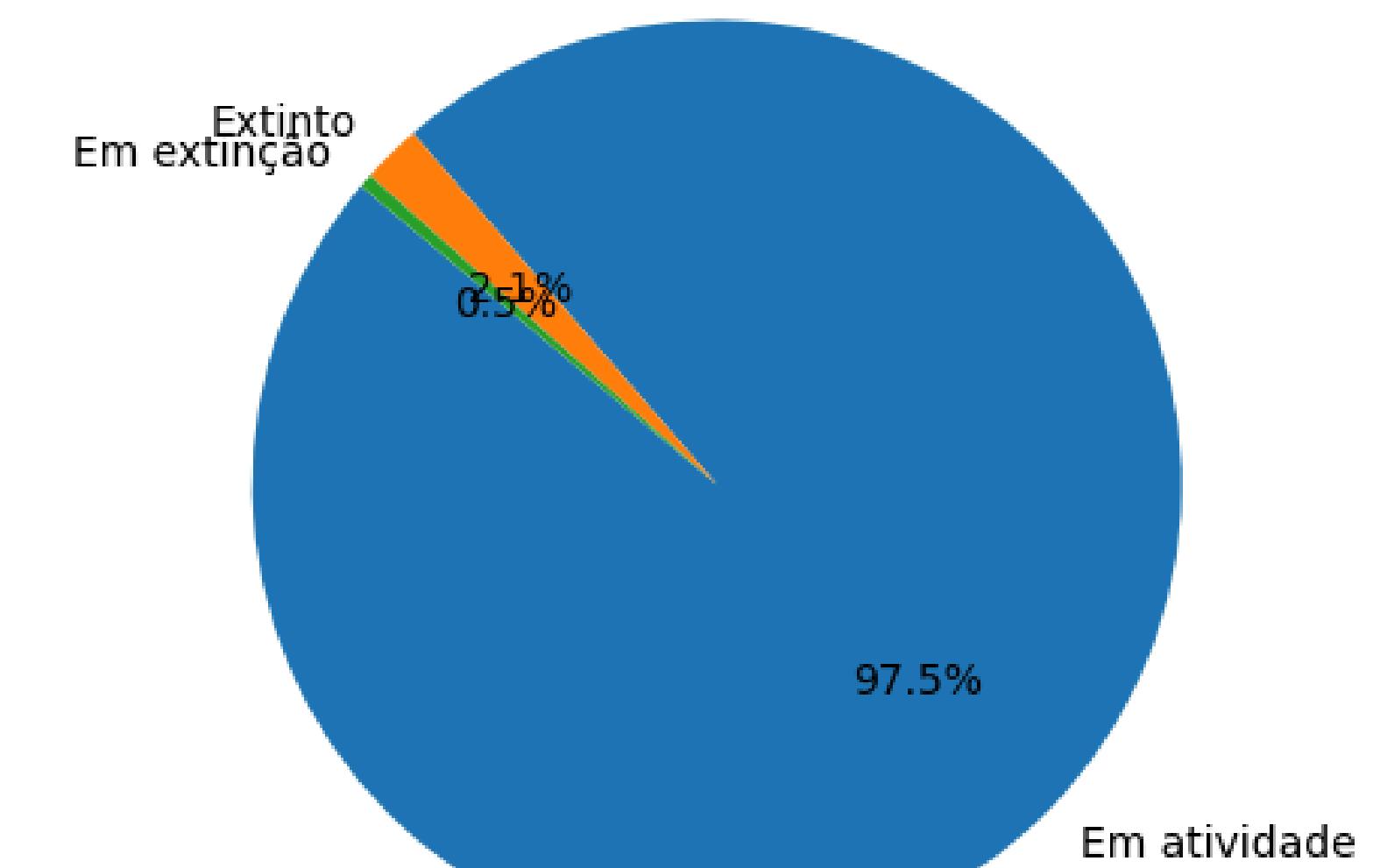


# ANÁLISE EXPLORATÓRIA DE DADOS - Situação dos Cursos

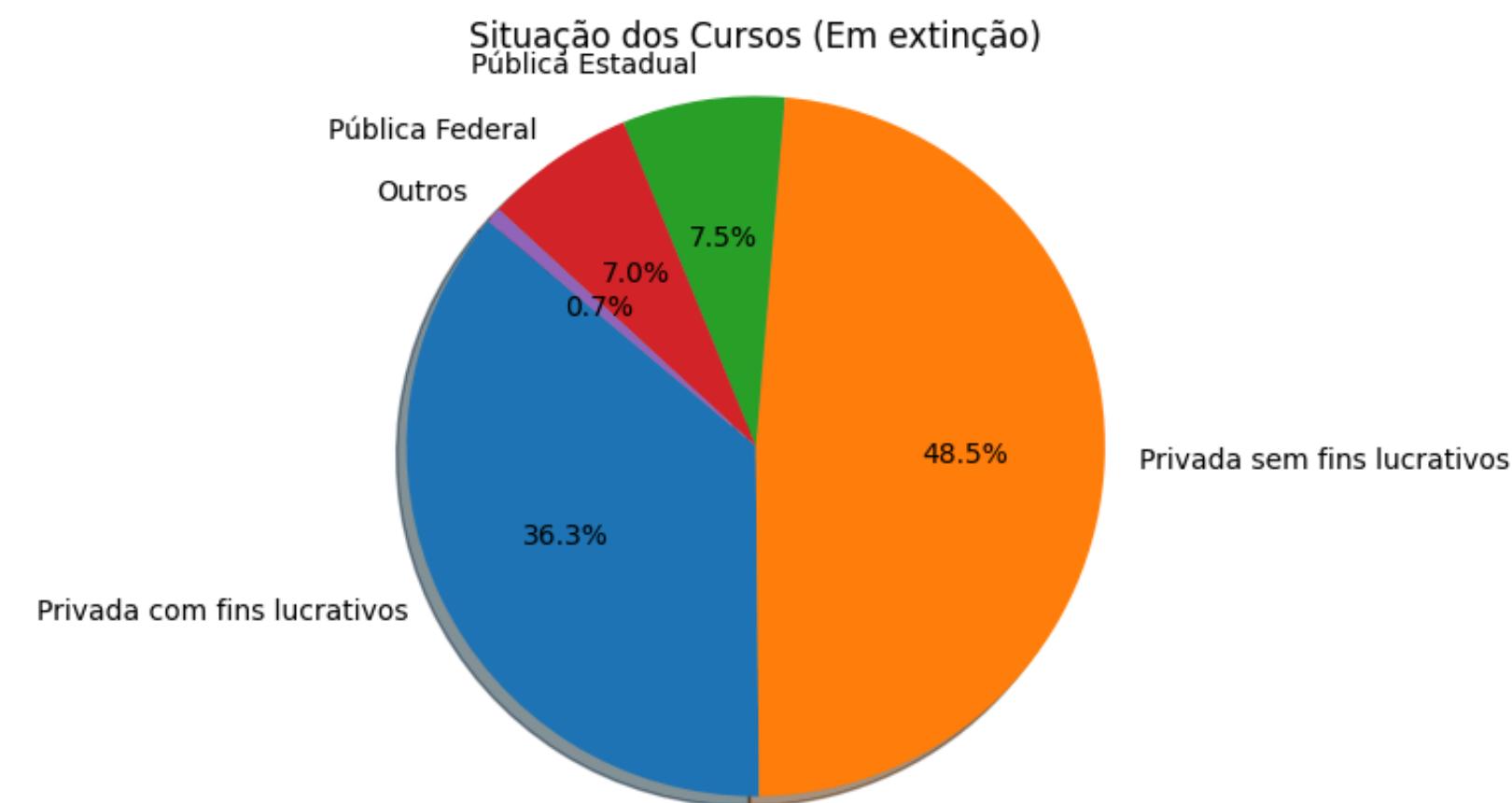
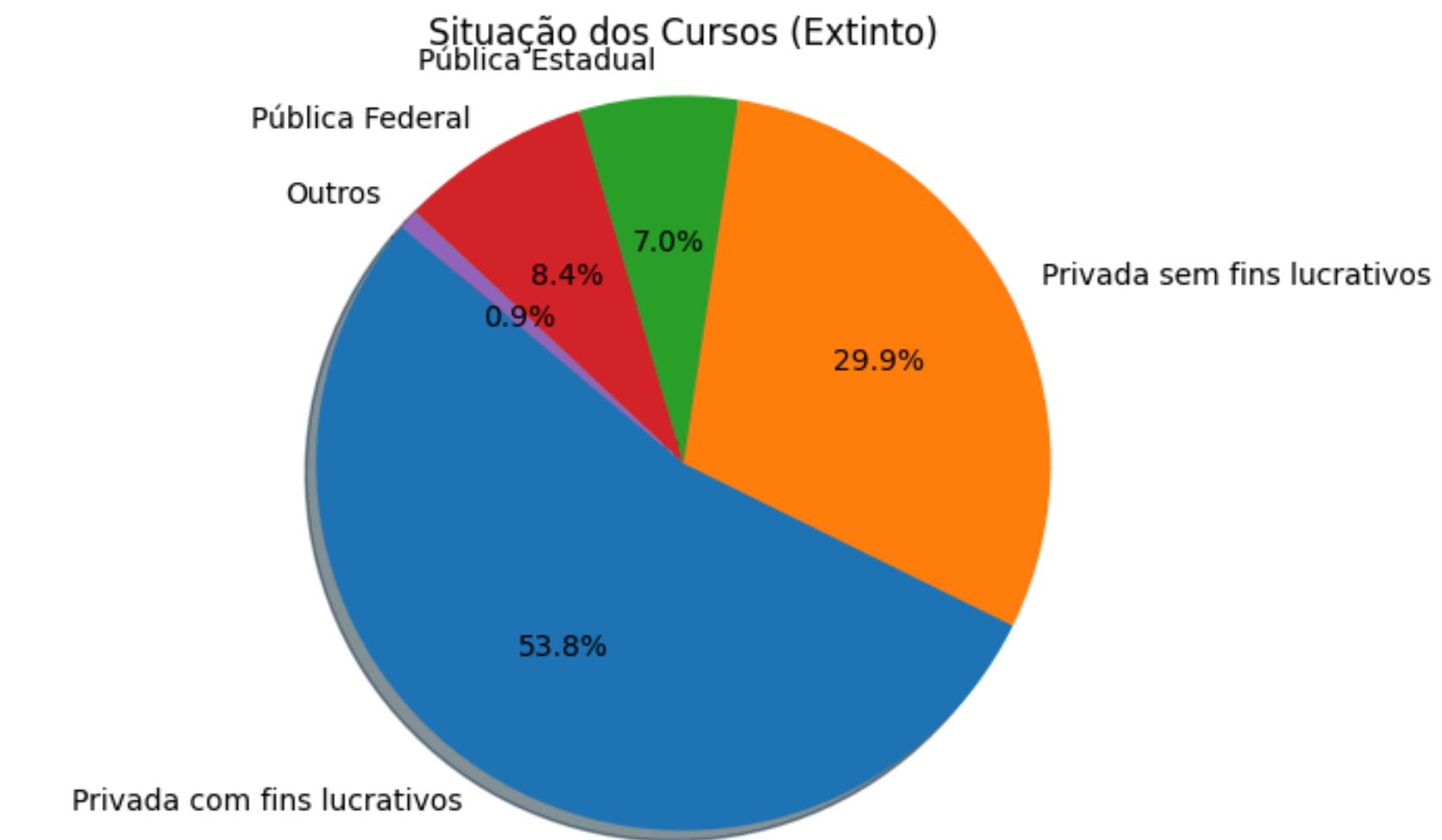
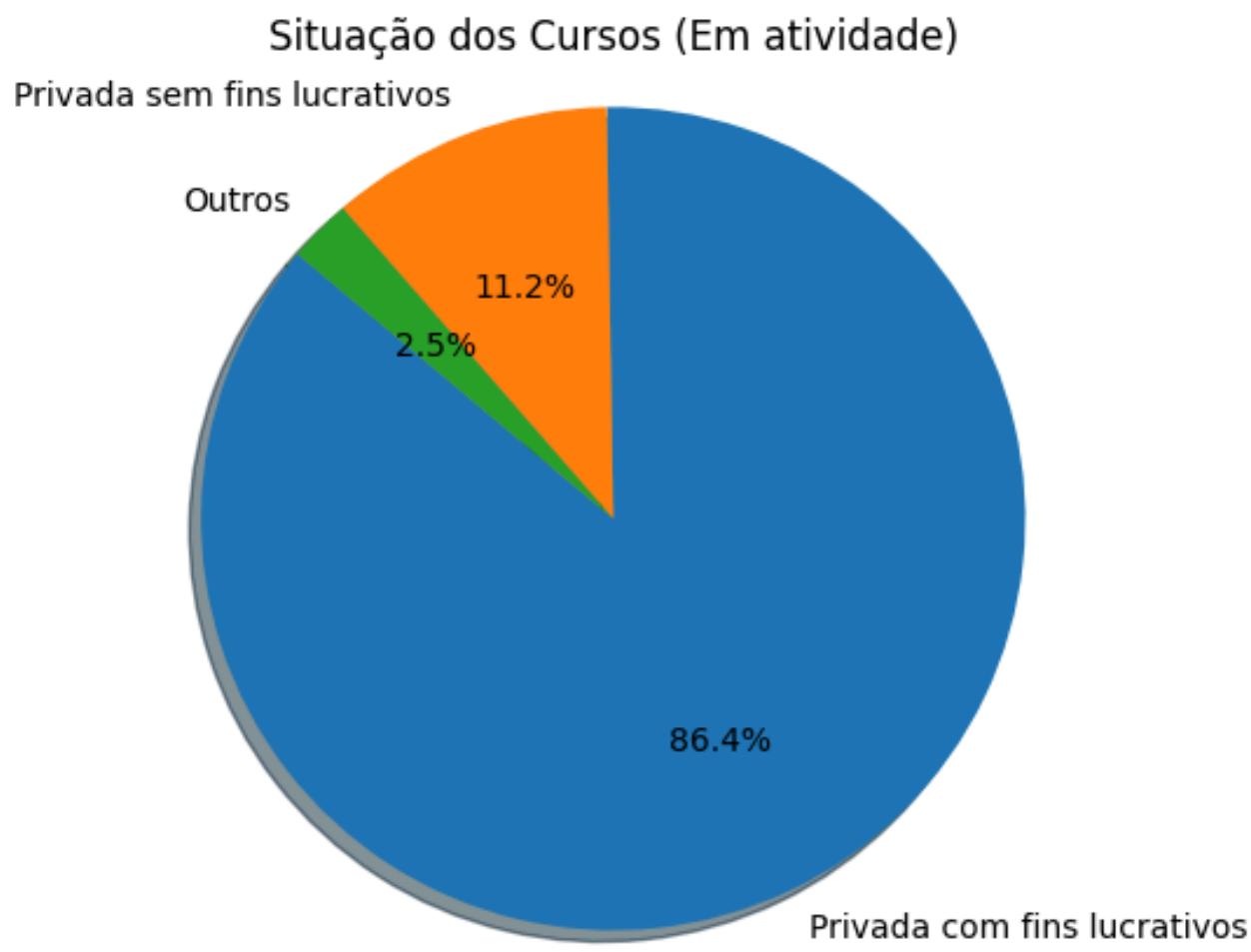
Distribuição da Situação dos Cursos de Educação Presencial



Distribuição da Situação dos Cursos de Educação a Distância



# ANÁLISE EXPLORATÓRIA DE DADOS - Situação dos Cursos



# ANÁLISE EXPLORATÓRIA DE DADOS - Cursos de TI

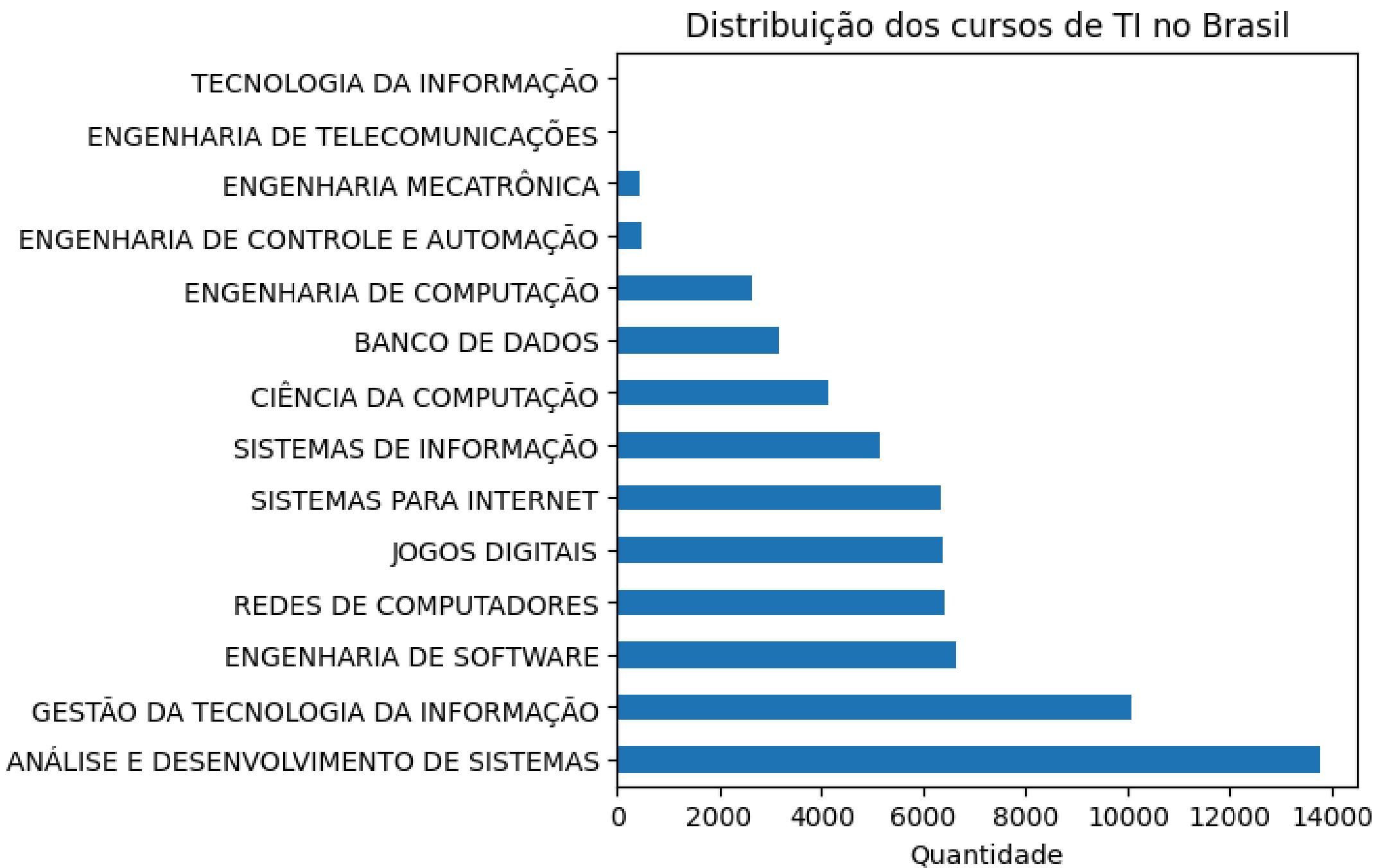
```
▶ dfti[(dfti['NOMEIES'] == 'UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE') & (dfti['NOME_CURSO'] == 'TECNOLOGIA DA INFORMAÇÃO')]
```

	NOMEIES	CATEGORIA_ADMINISTRATIVA	ORGANIZACAO_ACADEMICA	NOME_CURSO	GRAU	MODALIDADE	SITUACAO_CURSO	QT_VAGAS_AUTORIZADAS	CARGA_HORARIA	MUNICIPIO	UF	REGIAO
295004	UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE	Pública Federal	Universidade	TECNOLOGIA DA INFORMAÇÃO	Bacharelado	Educação Presencial	Em atividade	300	2600	Natal	RN	NORDESTE

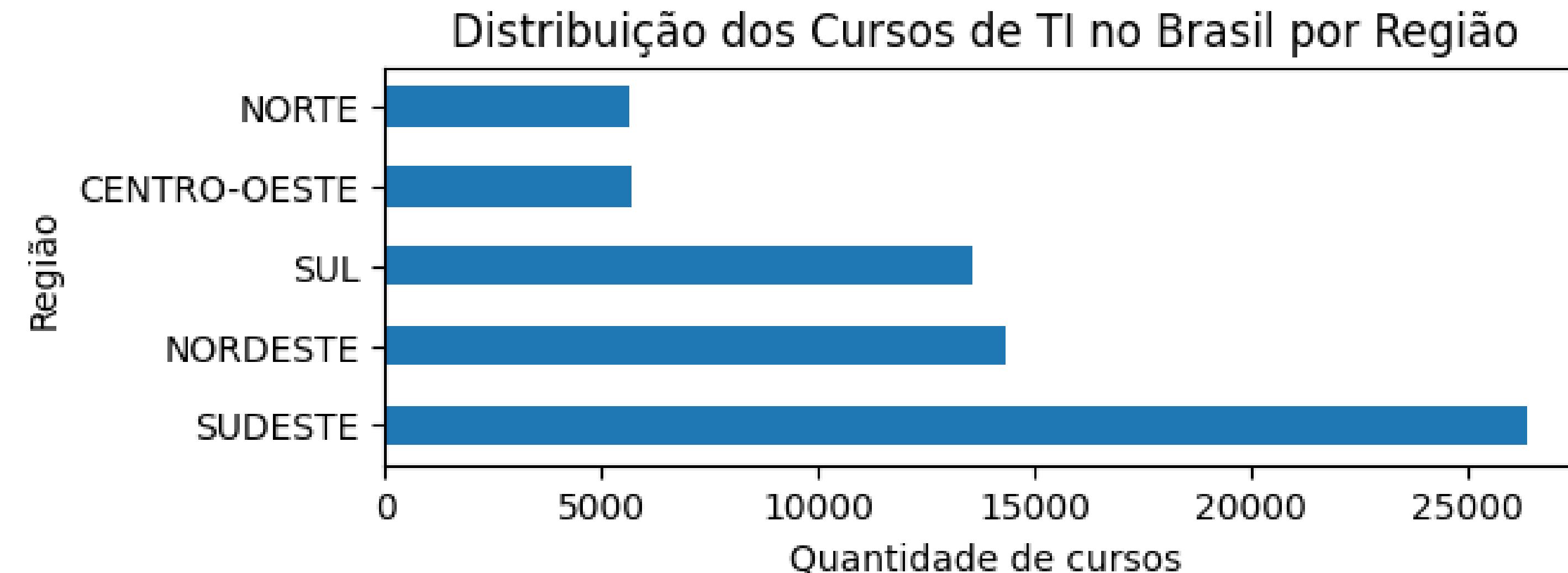
# ANÁLISE EXPLORATÓRIA DE DADOS - Cursos de TI

```
▶ dfti['NOME_CURSO'].value_counts()
```

```
→ NOME_CURSO
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS      13796
GESTÃO DA TECNOLOGIA DA INFORMAÇÃO           10096
ENGENHARIA DE SOFTWARE                      6635
REDES DE COMPUTADORES                      6399
JOGOS DIGITAIS                            6385
SISTEMAS PARA INTERNET                     6324
SISTEMAS DE INFORMAÇÃO                     5140
CIÊNCIA DA COMPUTAÇÃO                      4136
BANCO DE DADOS                            3164
ENGENHARIA DE COMPUTAÇÃO                   2625
ENGENHARIA DE CONTROLE E AUTOMAÇÃO          457
ENGENHARIA MECATRÔNICA                     433
ENGENHARIA DE TELECOMUNICAÇÕES             25
TECNOLOGIA DA INFORMAÇÃO                  14
Name: count, dtype: int64
```

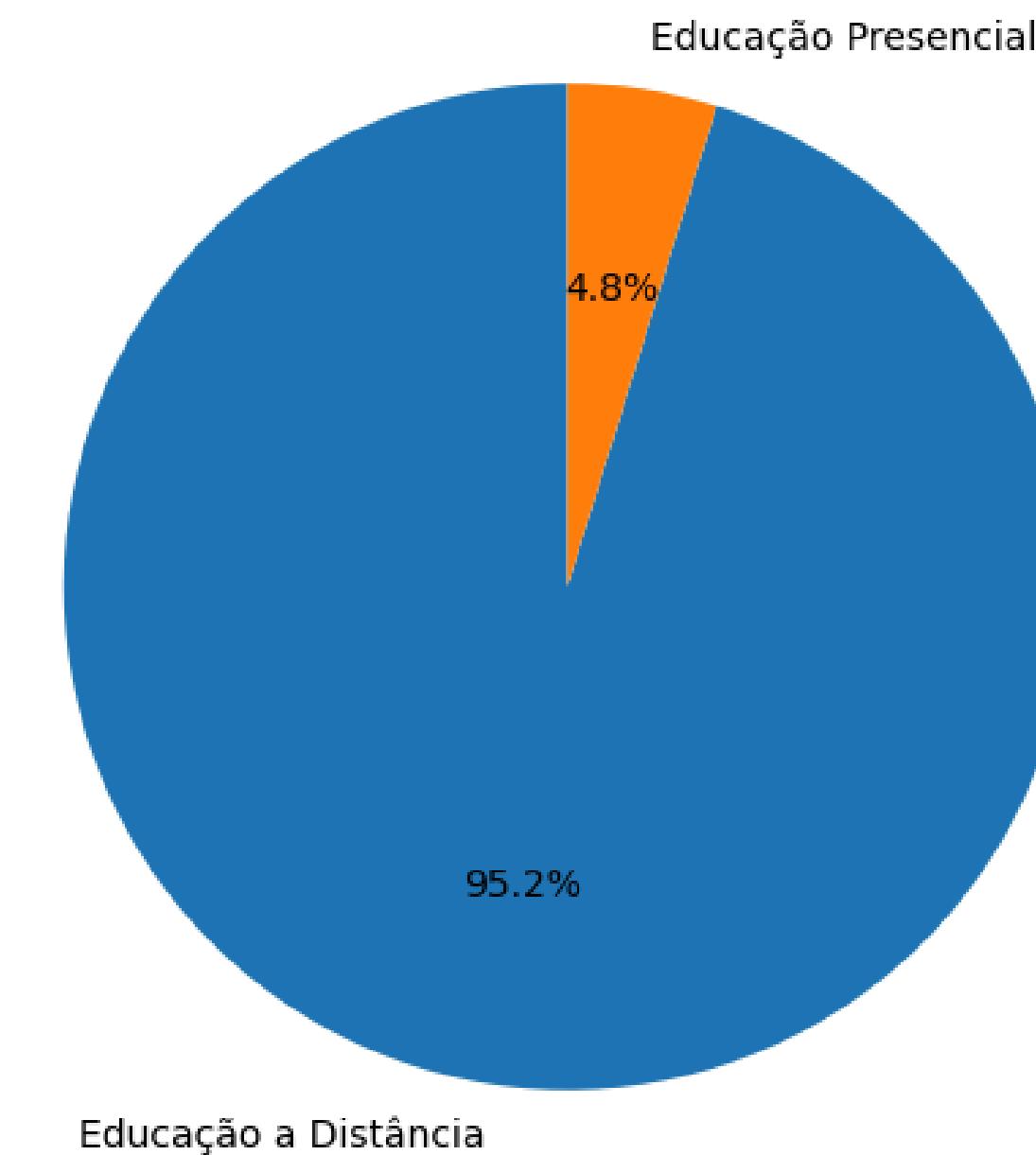


# ANÁLISE EXPLORATÓRIA DE DADOS - Cursos de TI

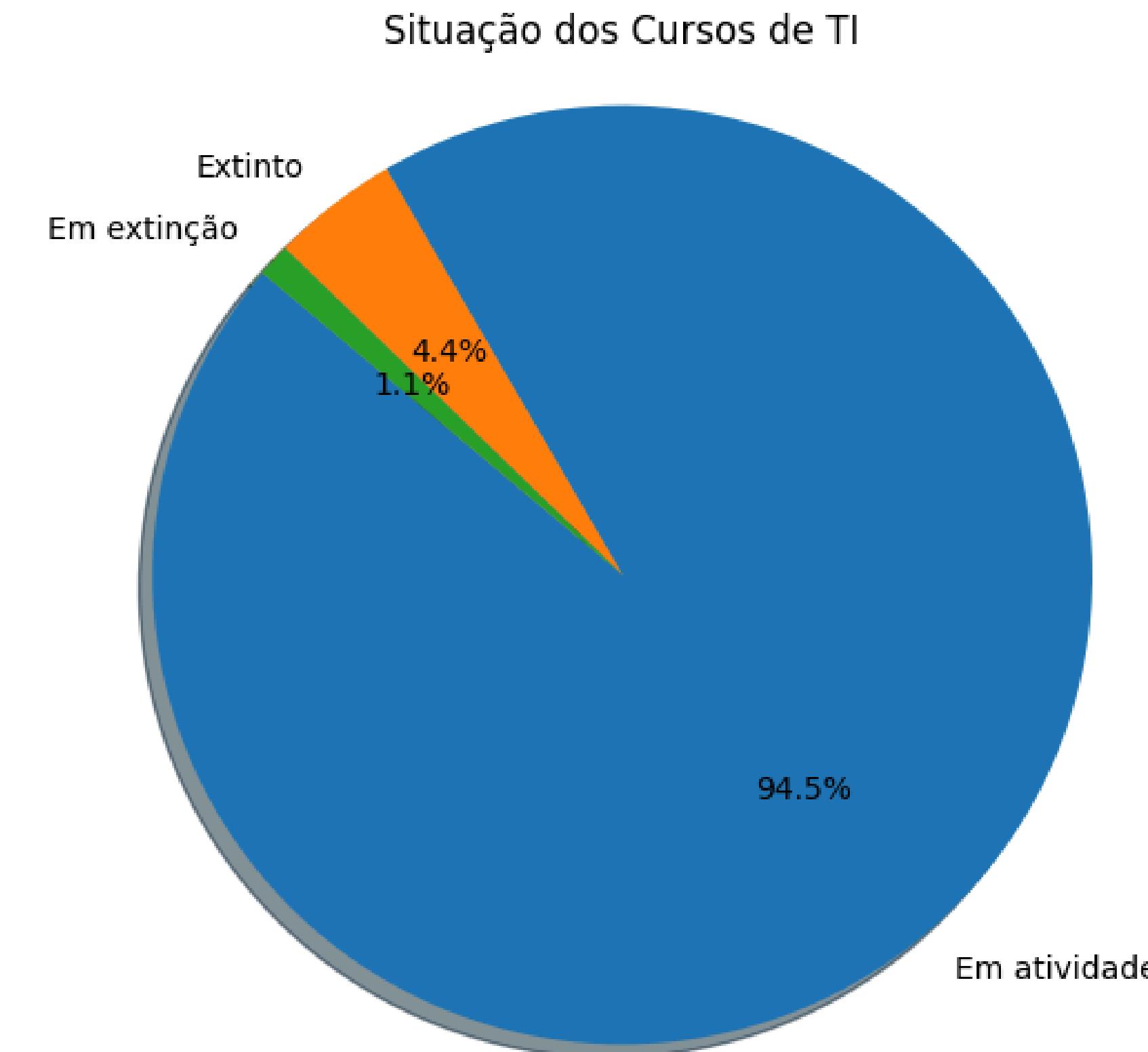


# ANÁLISE EXPLORATÓRIA DE DADOS - Cursos de TI

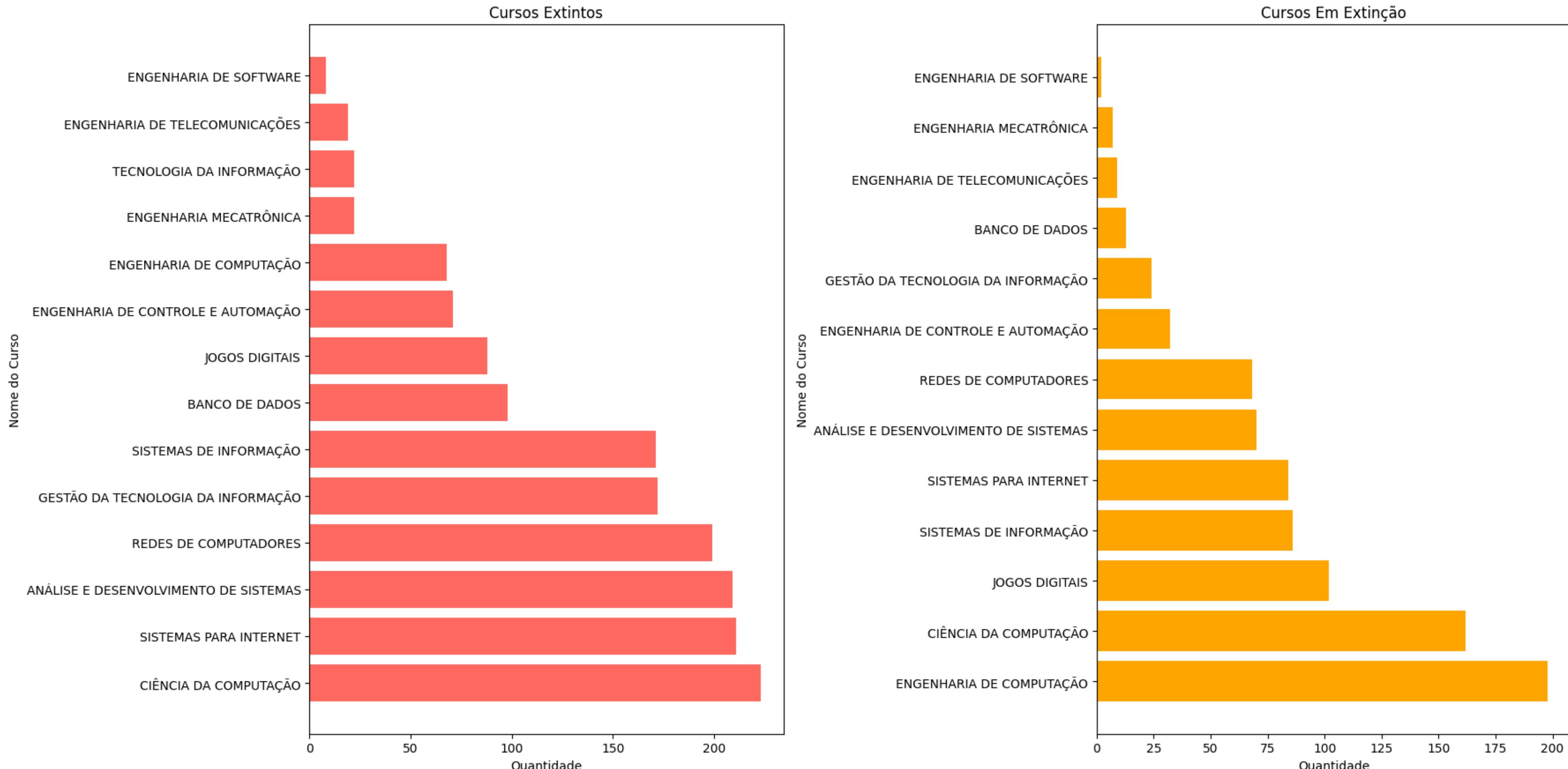
Distribuição dos Cursos de TI Em Atividade por Modalidade de Ensino



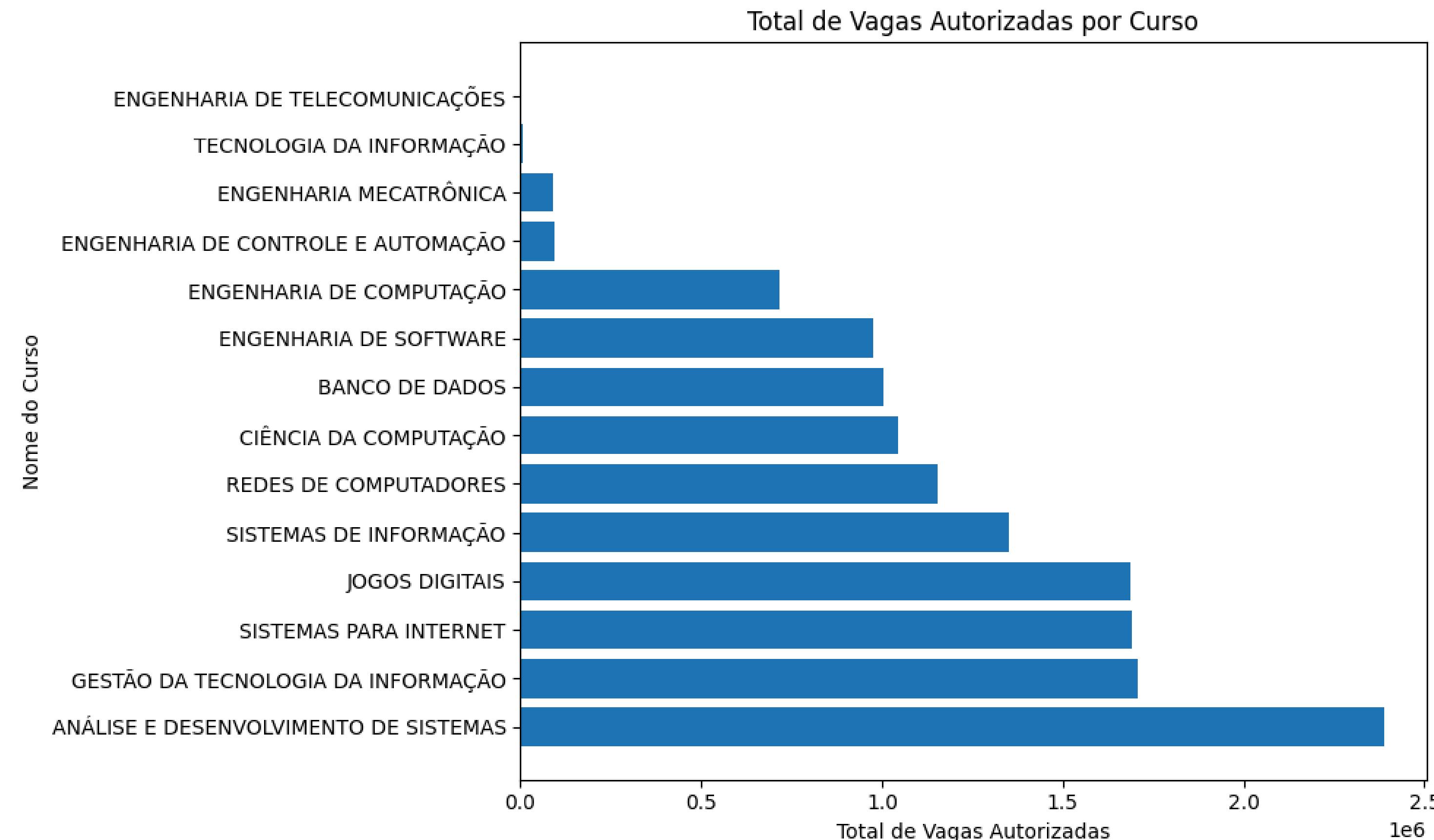
# ANÁLISE EXPLORATÓRIA DE DADOS - Cursos de TI



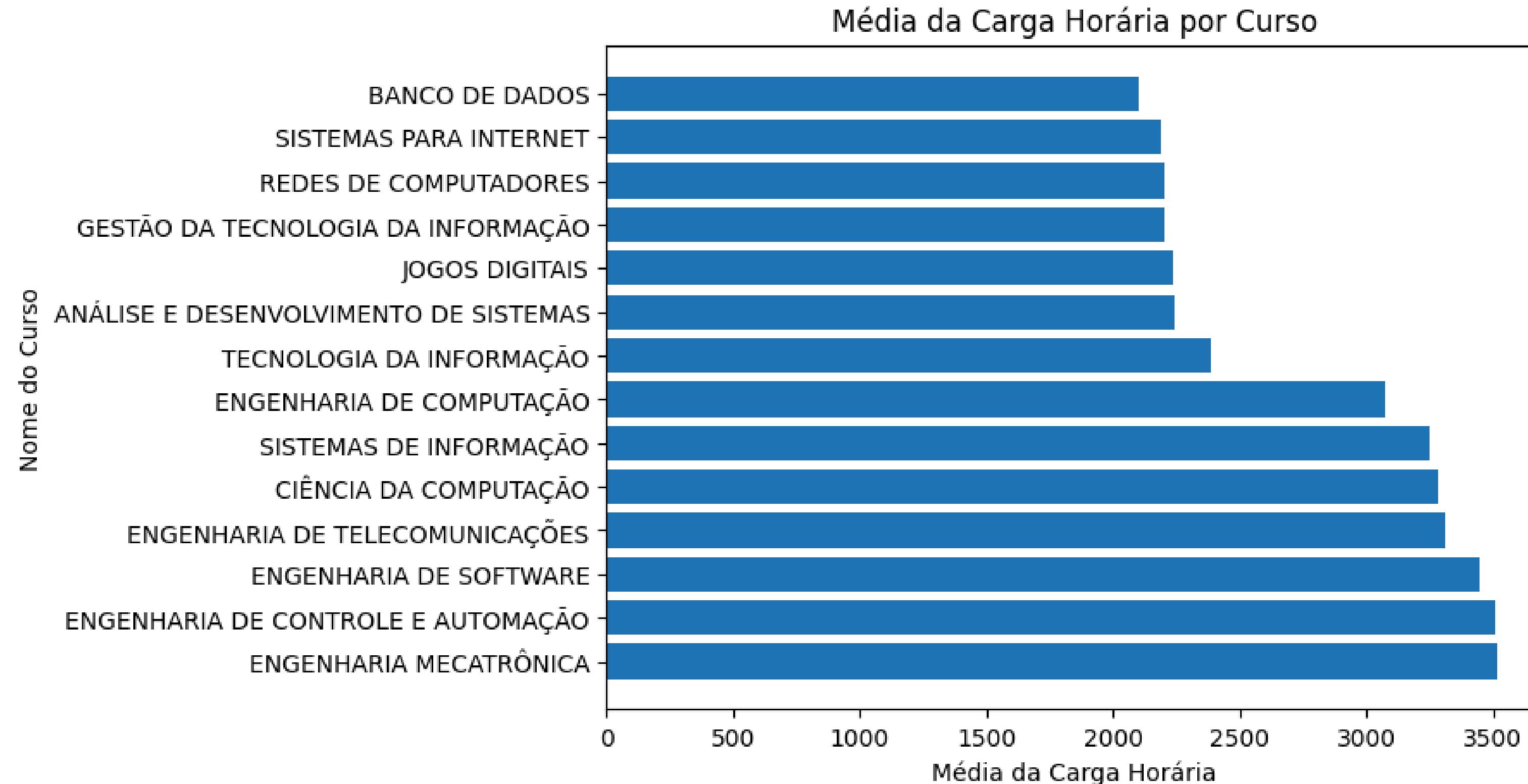
# ANÁLISE EXPLORATÓRIA DE DADOS - Cursos de TI



# ANÁLISE EXPLORATÓRIA DE DADOS - Cursos de TI



# ANÁLISE EXPLORATÓRIA DE DADOS - Cursos de TI



# REFERÊNCIAS

MEC. Cursos de Graduação do Brasil. 2022. Disponível em: <https://dadosabertos.mec.gov.br/indicadores-sobre-ensino-superior/item/183-cursos-de-graduacao-do-brasil>.

INSPER. Explorando os Cursos de Graduação na Área de Tecnologia: Conceitos e Opções. 2023. Disponível em: <https://www.insper.edu.br/noticias/cursos-de-graduacao-na-area-de-tecnologia/>.

UNOPAR. Carga horária da faculdade: como funciona e o que o MEC exige? 2022. Disponível em: <https://blog.unopar.com.br/carga-horaria-faculdade/>.

MUNDO EDUCAÇÃO. População do Brasil. 2022. Disponível em: <https://mundoeducacao.uol.com.br/geografia/populacao-brasileira.htm>.

INEP. DIRETORIA DE ESTATÍSTICAS EDUCACIONAIS. 2016. Disponível em: [https://www.uff.br/sites/default/files/paginas-internas-orgaos/modulo\\_ies\\_2016.pdf](https://www.uff.br/sites/default/files/paginas-internas-orgaos/modulo_ies_2016.pdf).



## CIÊNCIAS DE DADOS (IMD1151)

### PROJETO 01: CURSOS DE GRADUAÇÃO NO BRASIL

Profº. Dr Heitor Medeiros Florencio

Profº. Dr Daniel Sabino Amorim de Araujo

