PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

MARCELO AUGUSTO DIAS GARRIDO

IMPACTO DA PANDEMIA DE COVID-19 NAS EMPRESAS NACIONAIS

Belo Horizonte
2021

MARCELO AUGUSTO DIAS GARRIDO

IMPACTO DA PANDEMIA DE COVID-19 NAS EMPRESAS NACIONAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2021

SUMÁRIO

1. Introdução 4
1.1. Contextualização 4
1.2. O problema proposto 5
2. Coleta de Dados 6
3. Processamento/Tratamento de Dados 9
4. Análise e Exploração dos Dados 23
5. Criação de Modelos de Machine Learning 34
6. Apresentação dos Resultados 34
7. Links 35
REFERÊNCIAS 36
APÊNDICE 36

1. Introdução

1.1. Contextualização

A pandemia do COVID-19 atingiu um nível mundial, com força de destruição atípica e diversa, e não temos como vislumbrar uma redução substantiva e a extinção da doença com assertividade no curto prazo. A falta de conhecimento sobre o tema e seus impactos destruidores nas organizações e nas companhias provocaram danos imensuráveis e irremediáveis nas regiões e nas sociedades de todo o planeta.

A pandemia teve seu primeiro caso detectado no Brasil no final de fevereiro de 2020 e passou a ser comunitária no país em março do mesmo ano, mês da primeira morte pela doença.

Em abril, após a adoção do isolamento social, foram adotadas medidas pelo Governo Federal para mitigar os efeitos na economia, como linhas de crédito e empréstimos, medidas fiscais, auxílio emergencial para as populações mais vulneráveis, trabalhistas e outras. Apesar de todo esforço, o estrago causado na economia, principalmente nas Micro e Pequenas Empresas, que, segundo informações do SEBRAE[1], respondem por mais de um quarto do Produto Interno Bruto (PIB) brasileiro, foi devastador.

De acordo com os resultados da Pesquisa Pulso Empresa: Impacto da Covid-19 nas Empresas, realizada pelo IBGE (Instituto Brasileiro de Geografia e Estatística)[2], do total de 1,3 milhão de empresas que fecharam (temporária ou definitivamente) desde o início de 2020 até a primeira quinzena de junho, 522,6 mil (40%) encerraram suas atividades por causa da pandemia do novo coronavírus.

Em janeiro de 2021, após esforços da comunidade científica internacional, o Brasil iniciou o processo de vacinação em massa da população, medida que tem diminuído consideravelmente o número de novos casos e óbitos em todo o planeta.

Todavia, o aparecimento de novas cepas, fruto da mutação do vírus, tem aterrorizado e mantido em estado de alerta todo o planeta. Além desse receio, o fantasma do aparecimento de novos vírus semelhantes ao COVID-19 ou ainda mais devastadores, ainda pairam no ar e tiram o sono dos governantes mundiais.

1.2. O problema proposto

No decorrer deste estudo, a meta principal será a realização de uma análise exploratória dos dados coletados para identificar o impacto do COVID-19 nas empresas, bem como as relações da pandemia com as demais informações econômicas e sociais. Adicionalmente, a criação de um modelo preditivo (utilizando técnicas de Machine Learning) do impacto pandemia na continuidade das empresas.

Para um melhor entendimento do problema e sua resolução, utilizaremos o método dos 5-Ws:

(Why?) A identificação do impacto da pandemia do COVID-19 no país, principalmente na economia, e a relação deste com as demais variáveis de uma empresa, podem ser de extrema importância na adoção de políticas sociais e econômicas no enfrentamento da questão atual (ainda com altos índices e possibilidade de novas ondas), bem como no aparecimento de pandemias similares.

(Who?) Foram utilizadas fontes do Governo Federal, entidades da Administração Pública Federal e Organizações não Governamentais, elencadas abaixo:

1. **Receita Federal Brasil:** dados cadastrais públicos do Cadastro Nacional das Pessoas Jurídicas (CNPJ).

2. **Tesouro Nacional:** informações do “Código Município SIAFI – Sistema Integrado de Administração Financeira”, disponíveis no site Tesouro Transparente (TT).

3. **Ministério da Saúde:** dados obtidos no site do Governo Federal – Ministério da Saúde – demonstrando o acumulado de casos de contaminação e mortes por COVID-19 no Brasil.

4. **Atlas do Desenvolvimento Humano no Brasil:** dados sobre o IDH (Índice de Desenvolvimento Humano) no site Atlas do Desenvolvimento Humano no Brasil, o qual realiza uma compilação das informações geradas pelo Programa das Nações Unidas para o Desenvolvimento, órgão componente da Organização das Nações Unidas.

5. **IBGE:** dados do indicador econômico PIB (Produto Interno Bruto, o qual apresenta a soma de todos os bens e serviços produzidos em uma área geográfica em um determinado período (podendo ser um ano ou um trimestre). As informações foram obtidas no site do IBGE (Instituto Brasileiro de Geografia e Estatística).

6. **CONCLA (Comissão Nacional de Classificação):** informações sobre o CNAE, classificação de atividades econômicas oficialmente adotada pelo Sistema Estatístico Nacional e pelos órgãos gestores de cadastros e registros da Administração Pública do país. A CNAE é uma classificação hierarquizada em cinco níveis – seções, divisões, grupos, classes e subclasses. Os dados foram carregados do site do IBGE.

(What?): O objetivo deste estudo é a realização de uma análise exploratória na base de dados de CNPJ´s da Receita Federal, buscando identificar o impacto da pandemia de COVID-19 no fechamento das empresas e suas relações com as demais informações socioeconômicas obtidas de outras fontes. Além disso, utilizando modelos de “Machine Learning”, se tentara prever a continuidade das operações das empresas em decorrência da pandemia.

(Where?): Serão manipulados dados de todas as empresas (CNPJ´s) que realizaram transações no Cadastro Nacional de Pessoas Jurídicas da Receita Federal do Brasil.

(When?): Para o estudo será analisado um período de 28 meses, conforme descrito abaixo:

- Período 20-21 - COM COVID - Início: 01/04/2020 - Término: 31/05/2021

- Período 19-20 - SEM COVID - Início: 01/02/2019 - Término: 31/03/2020

2. Coleta de Dados

2.1 - Dados Públicos do Cadastro Nacional da Pessoa Jurídica (CNPJ)

Os dados foram obtidos no site da Receita Federal do Brasil[3]. Devido ao tamanho, as informações foram fracionadas pela Receita Federal em 20 arquivos, sendo 10 deles com informações sobre a EMPRESA e outros 10 sobre o

ESTABELECIMENTO. Os dados estão em formato texto e acessados em junho/2021.

Após a junção dos 20 arquivos iniciais em dois grupos distintos, restaram:

Dataset EMPRESA: 46.232.299 registros

CAMPO	DESCRIÇÃO	TIPO
CNPJ BÁSICO	NÚMERO BASE DE INSCRIÇÃO NO CNPJ (OITO PRIMEIROS DÍGITOS DO CNPJ).	object
RAZÃO SOCIAL / NOME EMPRESARIA	NOME EMPRESARIAL DA PESSOA JURÍDICA	object
NATUREZA JURÍDICA	CÓDIGO DA NATUREZA JURÍDICA	object
QUALIFICAÇÃO DO RESPONSÁVEL	QUALIFICAÇÃO DA PESSOA FÍSICA RESPONSÁVEL PELA EMPRESA	object
CAPITAL SOCIAL DA EMPRESA	CAPITAL SOCIAL DA EMPRESA	object
PORTE DA EMPRESA	CÓDIGO DO PORTE DA EMPRESA: 01 – NÃO INFORMADO 02 - MICRO EMPRESA 03 - EMPRESA DE PEQUENO PORTE 05 - DEMAIS	object
ENTE FEDERATIVO RESPONSÁVEL	O ENTE FEDERATIVO RESPONSÁVEL É PREENCHIDO PARA OS CASOS DE ÓRGÃOS E ENTIDADES DO GRUPO DE NATUREZA JURÍDICA 1XXX. PARA AS DEMAIS NATUREZAS, ESTE ATRIBUTO FICA EM BRANCO.	object

Dataset ESTABELECIMENTO: 48.854.518 registros

CAMPO	DESCRIÇÃO	TIPO
CNPJ BÁSICO	NÚMERO BASE DE INSCRIÇÃO (OITO PRIMEIROS DÍGITOS DO CNPJ)	object
CNPJ ORDEM	NÚMERO DO ESTABELECIMENTO DE INSCRIÇÃO (DO NONO AO DÉCIMO SEGUNDO DÍGITO DO CNPJ)	object
CNPJ DV	DÍGITO VERIFICADOR DO NÚMERO DE INSCRIÇÃO (DOIS ÚLTIMOS DÍGITOS DO CNPJ)	object
IDENTIFICADORMATRIZ/FILIAL	CÓDIGO DO IDENTIFICADOR MATRIZ/FILIAL: 1 – MATRIZ 2 – FILIAL	object
NOME FANTASIA	CORRESPONDE AO NOME FANTASIA	object
SITUAÇÃO CADASTRAL	CÓDIGO DA SITUAÇÃO CADASTRAL: 01 – NULA 02 – ATIVA 03 – SUSPENSA 04 – INAPTA 08 – BAIXADA	object
DATA SITUAÇÃO CADASTRAL	DATA DO EVENTO DA SITUAÇÃO CADASTRAL	object
MOTIVO SITUAÇÃO CADASTRAL	CÓDIGO DO MOTIVO DA SITUAÇÃO CADASTRAL	object
NOME DA CIDADE NO EXTERIOR	NOME DA CIDADE NO EXTERIOR	object
PAIS	CÓDIGO DO PAIS	object
DATA DE INÍCIO ATIVIDADE	DATA DE INÍCIO DA ATIVIDADE	object
CNAE FISCAL PRINCIPAL	CÓDIGO DA ATIVIDADE ECONÔMICA PRINCIPAL DO ESTABELECIMENTO	object
CNAE FISCAL SECUNDÁRIA	CÓDIGO DA(S) ATIVIDADE(S) ECONÔMICA(S) SECUNDÁRIA(S) DO ESTABELECIMENTO	object
TIPO DE LOGRADOURO	DESCRIÇÃO DO TIPO DE LOGRADOURO	object
LOGRADOURO	NOME DO LOGRADOURO ONDE SE LOCALIZA O ESTABELECIMENTO	object
NÚMERO	NÚMERO ONDE SE LOCALIZA O ESTABELECIMENTO. QUANDO NÃO HOUVER , 'S/N'	object
COMPLEMENTO	COMPLEMENTO PARA O ENDEREÇO DE LOCALIZAÇÃO DO ESTABELECIMENTO	object
BAIRRO	BAIRRO ONDE SE LOCALIZA O ESTABELECIMENTO	object
CEP	CÓDIGO DE ENDEREÇAMENTO POSTAL REFERENTE AO LOGRADOURO	object
UF	SIGLA DA UNIDADE DA FEDERAÇÃO EM QUE SE ENCONTRA O ESTABELECIMENTO	object
MUNICÍPIO	CÓDIGO DO MUNICÍPIO DE JURISDIÇÃO ONDE SE ENCONTRA O ESTABELECIMENTO	object

2.2 - Dados do Tesouro Nacional

Os dados foram acessados no site do Tesouro Nacional[4] e contêm informações sobre o Código SIAFI - Sistema Integrado de Administração Financeira. O arquivo foi disponibilizado em formato ".xls" e posteriormente salvo em formato ".csv" para tratamento pelo notebook. O dataset possui 5589 registros. Dados obtidos em junho/2021.

CAMPO	DESCRIÇÃO	TIPO
codigo_municipio_siafi	Código do Município no Cadastro SIAFI	object
cnpj_base	CNPJ básico do Município	object
descricao_municipio	Descrição do Município	object
uf	Unidade Federativa do Município	object
codigo_ibge	Código do Município no Cadastro do IBGE	object

2.3 - Dados do Ministério da Saúde

As informações acerca da pandemia de COVID-19 do site do Ministério da Saúde[5] foram disponibilizadas em 3 arquivos ".csv" conforme abaixo:

- HIST_PAINEL_COVIDBR_2020_Parte1_31mai2021
- HIST_PAINEL_COVIDBR_2020_Parte2_31mai2021
- HIST_PAINEL_COVIDBR_Parte3_31mai2021

Posteriormente, foi realizada a junção dos 3 arquivos e criado o dataset "covid_brasil_todas" com 2.422.659 registros. Dados obtidos em junho/2021.

CAMPO	DESCRIÇÃO	TIPO
regiao	Região do Município	object
estado	Estado do Município	object
municipio	Nome do Município	object
coduf	Código da Unidade Federativa	object
codmun	Código do Município	object
codRegiaoSaude	Código da Região do Ministério da Saúde	object
nomeRegiaoSaude	Nome da Região do Ministério da Saúde	object
data	Data do Evento	object
semanaEpi	Semana Epidemiológica	object
populacaoTCU2019	População do Município em 2019	object
casosAcumulado	Casos Acumulados de COVID-19	object
casosNovos	Casos Novos de COVID-19	object
obitosAcumulado	Óbitos Acumulados de COVID-19	object
obitosNovos	Óbitos Novos de COVID-19	object
Recuperadosnovos	Número de Novos Recuperados	object
emAcompanhamentoNovos	Número de Novos Casos em Acompanhamento	object
interior/metropolitana	Indicador se Município da Região Metropolitana ou Interior	object

2.4 - Dados do IDH

Os dados do IDH – Índice de Desenvolvimento Humano disponibilizado pelo site Atlas do Desenvolvimento Humano no Brasil[6], contendo 5.565 registros. O arquivo apresenta-se no formato ".xls" e posteriormente convertido para ".cvs". Dados obtidos em junho/2021.

CAMPO	DESCRIÇÃO	TIPO
Territorialidade	Nome do Município	object
CODIGO IBGE	Código IBGE do Município	object
Posicao IDHM	Posição na Classificação do IDH Nacional	object
IDHM	Índice de Desenvolvimento Humano do Município	object
Posicao IDHM Educacao	Posição na Classificação do IDHM Educação	object
IDHM Educacao	Índice de Desenvolvimento Humano Educação do Município	object
Posicao IDHM Longevidade	Posição na Classificação do IDHM Longevidade	object
IDHM Longevidade	Índice de Desenvolvimento Longevidade do Município	object
Posicao IDHM Renda	Posição na Classificação do IDHM Renda	object
IDHM Renda	Índice de Desenvolvimento Renda do Município	object

2.5 - Dados do PIB

As informações do PIB (Produto Interno Bruto) foram adquiridas no site do IBGE (Instituto Brasileiro de Geografia e Estatística[7]. O arquivo foi disponibilizado no formato ".xls" e convertido para ".csv" para tratamento no notebook. Dados obtidos em junho/2021, contendo 5.570 registros.

CAMPO	DESCRIÇÃO	TIPO
Sigla da Unidade da Federação	Sigla do Estado do Município	object
Código do Município	Código do Município	object
Nome do Município	Nome do Município	object
PIB	Produto Interno Bruto do Município	object
PIB PER CAPITA	Produto Interno Bruto per Capita do Município	object
Ano	Ano da Informação	object

2.6 - Dados do CNAE

Os dados do CNAE dos CNPJ´s estudados foram obtidos no site do IBGE (Instituto Brasileiro de Geografia e Estatística[8] em junho/2021. Como não existia a disponibilidade de um arquivo contendo as informações condensadas das seções e suas divisões, foi criado um arquivo ".csv" com os dados disponíveis no site do IBGE. Foram importados 87 registros.

CAMPO	DESCRIÇÃO	TIPO
cnae_divisao	Número da Divisão da Classificação CNAE	object
cnae_secao	Letra da Seção da Classificação CNAE	object
cnae_descricao	Descrição da Classificação CNAE	object

3. Processamento/Tratamento de Dados

Nesta seção serão descritas as etapas para tratamento dos datasets utilizados na etapa anterior à análise e exploração dos dados. O resultado será o arquivo "dataset_final.csv".

Infelizmente, devido ao tamanho dos datasets, alguns não foram disponibilizados no repositório do trabalho de conclusão do curso em sua totalidade. Todavia, para exemplificar o conteúdo, um exemplo de cada um deles foi gerado com uma quantidade reduzida de registros. Caso haja necessidade de execução dos notebooks, a geração deverá ser realizada através dos respectivos notebooks.

3.1 Processamento dos Dados

Neste tópico, será descrito todo processamento dos dados obtidos nos diversos datasets. As informações relevantes para o estudo serão concatenadas em um único arquivo para tratamento em etapa posterior:

3.1.1 Datasets de CNPJ´s

As informações referentes aos CNPJ´s das empresas foram disponibilizados pela Receita Federal em dois conjuntos: EMPRESA e ESTABELECIMENTO, sendo cada um deles dividido em 10 arquivos.

a. **Dataset "ESTABELECIMENTO"**

Devido à quantidade de registros dos arquivos de CNPJ´s disponibilizados pela Receita Federal, os dados do ESTABELECIMENTO foram tratados em dois notebooks:

- “0001_Tratar_Tabelas_Estabelecimento_01”

- “0002_Tratar_Tabelas_Estabelecimento_02”

Nestes notebooks, foi realizada a leitura de 10 arquivos “.csv”, a eliminação de colunas devido ao grande número de registros sem informações ou redundância para o estudo e a seleção dos registros conforme data de corte e código da situação cadastral do CNPJ, segundo os seguintes critérios:

Eliminação das variáveis:

- Devido ao grande número de registros sem informações, o número de informações únicas para cada CNPJ e a sua irrelevância para o para o projeto em estudo: nome_fantasia, nome_cidade_exterior, codigo_pais, tipo_logradouro, logradouro, numero, complemento, bairro, cep, uf, ddd1, telefone1, ddd2, telefone2, dddfax, fax, e-mail e situacao_especial.

- A coluna "motivo_situacao_cadastral" traz o motivo pelo qual a Situação Cadastral do CNPJ foi alterada. No nosso estudo, como o TARGET é a Situação Cadastral "02" ou "08", indenpendentemente do seu motivo, a coluna será eliminada.

- A coluna "cnae_secundario" traz um detalhamento excessivo para o CNAE da empresa. No nosso estudo, para diminuirmos a amplitude das informações, iremos trabalhar com a classificação da Secão e e da Divisão, além do CNAE principal. Por esse motivo, a coluna "cnae_secundario" será eliminada.

Seleção dos registros:

- Data de Corte:

  - COM COVID - Início: 01/04/2020 - Término: 31/05/2021

- SEM COVID - Início: 01/02/2019 - Término: 31/03/2020

- Situação Cadastral:

- 02 (ATIVA)

- 08 (BAIXADA)

Como resultado, os arquivos “cnpj_estabelecimento_todas_parte01.csv” e “cnpj_estabelecimento_todas_parte02.csv” foram criados.

b. Dataset EMPRESA

No notebook “0003 - Criar Tabela Final de CNPJ.ipynb”, foi realizado o tratamento dos arquivos EMPRESA, momento em que algumas foram eliminadas, segundo os seguintes critérios:

- Razão Social: Traz o nome do estabelecimento. Por tratar-se de uma informação praticamente única para cada CNPJ, a informação é irrelevante para o estudo.

- Natureza Jurídica e Qualificação do Responsável: Traz informações sobre a característica da empresa. Como essas informações são complementares ao código CNAE, optamos pela eliminação.

- Ente Federativo Responsável: Informação redundante da informação do código SIAFI que será concatenado a seguir da tabela de Estabelecimentos.

- Porte da Empresa: conforme códigos do cadastro, a esmagadora maioria refere-se aos códigos 1 e 5, sem relevância para o estudo. Os números podem ser observados na pesquisa abaixo:

01 – NÃO INFORMADO

02 - MICRO EMPRESA

03 - EMPRESA DE PEQUENO PORTE

05 - DEMAIS

```
[ ] porte_empresa = cnpj_empresa_todas.groupby('porte_empresa')['porte_empresa'].count()

[ ] porte_empresa

    porte_empresa
    01    31649224
    03     1168801
    05    13354392
    Name: porte_empresa, dtype: int64
```

c. Dataset "CNPJ FINAL"

Finalmente, também no notebook "0003 - Criar Tabela Final de CNPJ.ipynb", foi realizado a concatenação dos dois datasets (EMPRESA e ESTABELECIMENTO) e criado um dataset chamado "cnpj_final.CSV" com as seguintes características:

```
[16] cnpj_final.shape

     (11157257, 10)

[17] cnpj_final.dtypes

     cnpj_basico                  object
     cnpj_ordem                   object
     cnpj_dv                      object
     matriz_filial                object
     codigo_situacao_cadastral    object
     data_situacao_cadastral      object
     data_inicio_atividade        object
     cnae_principal               object
     codigo_municipio_siafi       object
     capital_social               object
     dtype: object
```

3.1.2 Demais datasets

Os dados dos demais datasets integrantes deste estudo foram processados no notebook "0004 - Processamento e Tratamento dos Dados.ipynb". A seguir, descrição detalhada de cada etapa.

a. Dataset SIAFI

Este arquivo, disponibilizado pelo Tesouro Nacional, foi utilizado para a concatenação da informação do "código IBGE" ao dataset de CNPJ Final. O dataset de CNPJ da Receita Federal identifica o município pelo "código SIAFI" e os demais datasets utilizam o "código IBGE". Portanto, através do dataset SIAFI, que possuí as duas informações, foi incluída a coluna "código do IBGE" ao dataset CNPJ Final. Além dessa informação, também foi acrescentada, ao

dataset CNPJ Final, as colunas “descrição do município” e uma coluna contendo o “código do IBGE ajustado”. Essa coluna é necessária para o tratamento, a seguir, das informações de COVID-19 pois este dataset possuí o “código do IBGE” sem o dígito verificador.

b. Dataset COVID

A partir deste dataset, disponibilizado pelo Ministério da Saúde, incluiremos as informações da pandemia de COVID-19 ao dataset de CNPJ Final. Os dados foram concatenados através da variável “código_ibge_data” (junção da data da situação cadastral do evento e o código do IBGE). As seguintes informações serão acrescidas:

- informações sobre o município: região, estado, populacaoTCU2019 e interior_metropolitana (“0” para município do interior e “1” para região metropolitana);

- casosAcumulado_covid: número acumulado de casos de COVID no Município até a data de encerramento do CNPJ ou da data de corte em 31/05/2021 para as empresas ainda abertas;

- obitosAcumulado_covid: número acumulado de óbitos por COVID no Município até a data de encerramento do CNPJ ou da data de corte em 31/05/2021 para as empresas ainda abertas;

- casos_acumulados_porcentagem_covid: percentual do acumulado de casos de COVID pelo total de habitantes do Município. Coluna criada para possibilitar a comparação relativa dos casos de COVID entre os municípios;

- obitos_acumulados_porcentagem_covid: percentual do acumulado de casos de óbitos por COVID pelo total de habitantes do Município. Coluna criada para possibilitar a comparação relativa dos casos de óbitos por COVID entre os municípios.

Neste momento também foram eliminadas as colunas:

- municipio: informação reduntante do SIAFI;

- coduf: utilizado o nome do Estado;

- codRegiaoSaude e nomeRegiaoSaude: utilizada a região do IBGE;

- semanaEpi: corresponde a semana epideomológica. Utilizada da data de ocorrência;

- casosNovos e obitosNovos: utilizado o valor acumulado;

- Recuperadosnovos e emAcompanhamentoNovos: utilizado o valor acumulado de novos casos e óbitos.

c. Dataset IDH

Neste dataset iremos buscar as informações do IDH DO MUNICIPIO, IDHM EDUCACAO DO MUNICIPIO, IDHM LONGEVIDADE DO MUNICIPIO e IDHM RENDA DO MUNICIPIO.

- IDH DO MUNICÍPIO: O Índice de Desenvolvimento Humano (IDH) é uma unidade de medida utilizada para aferir o grau de desenvolvimento de uma determinada sociedade nos quesitos de educação, saúde e renda. O IDH é uma referência numérica que varia entre 0 e 1. Quanto mais próximo de zero, menor é o indicador para os quesitos de saúde, educação e renda. Quanto mais próximo de 1, melhores são as condições para esses quesitos.

- IDHM EDUCACAO DO MUNICÍPIO: O indicador educação refere-se à quantidade média de anos de estudo de uma população. Entende-se que, quanto maior for o tempo de permanência de uma população na escola, melhores serão as chances de desenvolvimento para esse país.

- IDHM LONGEVIDADE DO MUNICÍPIO: Na variável saúde, avalia-se basicamente a taxa de expectativa de vida dos cidadãos de cada país participante. Entende-se que, quanto maior for essa taxa, melhores serão as condições de vida de seus habitantes.

- IDHM RENDA DO MUNICÍPIO: No quesito renda, mede-se o valor médio do rendimento dos cidadãos com base na média do Produto

Interno Bruto (PIB), que é a soma de toda a riqueza produzida por um país em determinado período (normalmente anual) dividida pelo número de habitantes.

Neste momento também foram eliminadas as colunas:

- Territorialidade: nome do município. Informação reduntante do SIAFI

- Posição IDHM, Posição IDHM Educação, Posição IDHM Longevidade, Posição IDHM Renda: informação classificatória e única da posição do município no IDH. Como está sendo utilizado o valor absoluto de cada índice, optou-se pela eliminação destas informações.

d. Dataset PIB

Neste dataset iremos acessar os dados do PIB e PIB PER CAPITA do município de cada CNPJ´s estudado.

PIB é a sigla para Produto Interno Bruto, que, em linhas gerais, é um indicador econômico bastante utilizado na Macroeconomia (ramo das Ciências Econômicas) que apresenta a soma de todos os bens e serviços produzidos em uma área geográfica em um determinado período (podendo ser um ano ou um trimestre). Sendo assim, o PIB representa a dinâmica econômica do lugar, apontando o possível crescimento da economia. As seguintes informações serão acrescidas:

- PIB DO MUNICÍPIO: Valor total do PIB do município.

- PIB PER CAPITA: Valor do PIB pelo número de habitantes da localidade.

Neste momento também foram eliminadas as colunas:

- Sigla da Unidade da Federação, Código do Município e Nome do Município: informações redundantes. Já foram incorporadas de outros arquivos;

- Ano: ano em que a pesquisa foi realizada. Valor idêntico para todos os municípios.

e. Dataset CNAE

A CNAE é a classificação de atividades econômicas oficialmente adotada pelo Sistema Estatístico Nacional e pelos órgãos gestores de cadastros e registros da Administração Pública do país.

A CNAE é uma classificação hierarquizada em cinco níveis – seções, divisões, grupos, classes e subclasses. O quinto nível, o de subclasses, corresponde ao detalhamento usado para a identificação econômica das unidades de produção, normalmente constituídas como pessoa jurídica ou profissionais autônomos, em cadastros e registros da Administração Pública, nas três esferas de governo. As seguintes informações serão acrescidas:

- cnae_divisao : representa a divisão na classificação CNAE;

- cnae_secao : representa a seção na classificação CNAE a qual pertence o CNPJ. É composto por 21 seções, nível mais elevado da classificação CNAE, o qual será considerado no nosso estudo para observar o impacto do COVID em setores da economia;

- cnae_descricao : descrição da seção em análise.

3.1.3 Preparação Final do dataset “cnpj_final.csv”

Neste momento, as seguintes ações adicionais foram tomadas:

- Criação da coluna alvo do modelo de Machine Learning: coluna "situacao_cadastral_target" contendo a seguinte informação:

  CNPJ´s FECHADOS (situação cadastral "8"): valor um (1)

  CNPJ´s ABERTOS (situação cadastral "2"): valor zero (0)

- Criação da coluna contendo somente o ANO e MÊS da Situação Cadastral: utilizada na exploração visual dos dados.

3.2 Tratamento dos Dados

Nesta etapa será realizado o tratamento dos dados processados na etapa anterior. Os dados também foram tratados no notebook “0004 - Processamento e Tratamento dos Dados.ipynb”. A seguir, descrição detalhada de cada etapa.

3.2.1 Verificação e tratamento dos Missing Values

Após a análise dos registros que apresentam “missing values”, as informações abaixo foram coletadas e as seguintes ações foram tomadas:

```
[83] cnpj_final.isna().sum()

cnpj_basico                          0
cnpj_ordem                           0
cnpj_dv                              0
matriz_filial                        0
codigo_situacao_cadastral            0
data_situacao_cadastral              0
data_inicio_atividade                0
cnae_principal                       0
codigo_municipio_siafi               0
capital_social                       0
descricao_municipio              10329
codigo_ibge                      10329
regiao                           10329
estado                           10329
populacaoTCU2019                 10329
interior_metropolitana           10329
casosAcumulado_covid           4882546
obitosAcumulado_covid          4882546
casos_acumulados_porcentagem_covid 4882546
obitos_acumulados_porcentagem_covid 4882546
idh_municipio                    12487
idh_educacao                     12487
idh_longevidade                  12487
idh_renda                        12487
pib                              10329
pib_per_capita                   10329
cnae_divisao                         0
cnae_secao                           0
cnae_descricao                       0
situacao_cadastral_target            0
data_situacao_cadastral_ano_mes      0
dtype: int64
```

a. Informações do Município

Como observado abaixo, as informações com valores nulos para os campos analisados pertencem ao Município código SIAFI "9707". Isso ocorreu pelo fato do código "9707" indicar empresa estabelecida no exterior e não ter correlação com código IBGE. Logo, os CNPJ´s com essa condição serão eliminados da base de dados pois não existe maneira de referenciar as demais informações para os CNPJ´s nessa situação.

```
[84] codigo_municipio_siafi_NaN = cnpj_final[cnpj_final['codigo_ibge'].isnull()]
```

```
[86] codigo_municipio_siafi_NaN.groupby('codigo_municipio_siafi')['codigo_municipio_siafi'].count()

codigo_municipio_siafi
9707    10329
Name: codigo_municipio_siafi, dtype: int64
```

```
[87] cnpj_final = cnpj_final[cnpj_final.codigo_municipio_siafi != 9707]  #EMPRESAS DO EXTERIOR
```

b. Informações COVID-19

Os casos de nulidade nos campos de "COVID" são os pertencentes a empresas em que a data da situação cadastral se encontra fora dos 14 meses de ocorrência de COVID (Período 19-20 - SEM COVID (14 meses) - Início: 01/02/2019 - Término: 31/03/2020). Por essa razão, imputaremos o valor "0" nesses campos pois não houve casos de COVID-19 relatados neste período.

c. Informações IDH

Os municípios listados abaixo não fizeram parte do estudo realizado nacionalmente para determinação do IDH para o ano de 2010. Por esse motivo, para evitar incoerências nas análises, os CNPJ´s desses municípios serão eliminados da base de dados.

```
[90] idh_municipios_NaN = cnpj_final[cnpj_final['idh_municipio'].isnull()]

[91] idh_municipios_NaN.groupby('codigo_ibge')['codigo_ibge'].count()

     codigo_ibge
     1504752    332
     4212650    461
     4220000    945
     4314548    159
     5006275    261
     Name: codigo_ibge, dtype: int64

[92] cnpj_final = cnpj_final[cnpj_final.codigo_ibge != 1504752]  #MUNICÍPIOS SEM INFORMAÇÕES DE IDH
     cnpj_final = cnpj_final[cnpj_final.codigo_ibge != 4212650]
     cnpj_final = cnpj_final[cnpj_final.codigo_ibge != 4220000]
     cnpj_final = cnpj_final[cnpj_final.codigo_ibge != 4314548]
     cnpj_final = cnpj_final[cnpj_final.codigo_ibge != 5006275]
```

3.2.2 Verificação e tratamento dos CNAE´s com alto índice de Abertura e Fechamento de CNPJ´s

Ao serem analisados os gráficos com o agrupamento por seção das empresas fechadas no período, identificamos duas categorias com alto número de ocorrências: "COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS" e "OUTRAS ATIVIDADES DE SERVIÇOS". Analisando detalhadamente os dois grupos, observações incoerências no último. Conforme demonstrado na tabela de agrupamento de CNAE´s e no gráfico desse agrupamento, o CNAE Principal "9492800" (Atividades de organizações políticas) será eliminado pois não caracteriza empresas "DE FATO", objeto deste trabalho. O

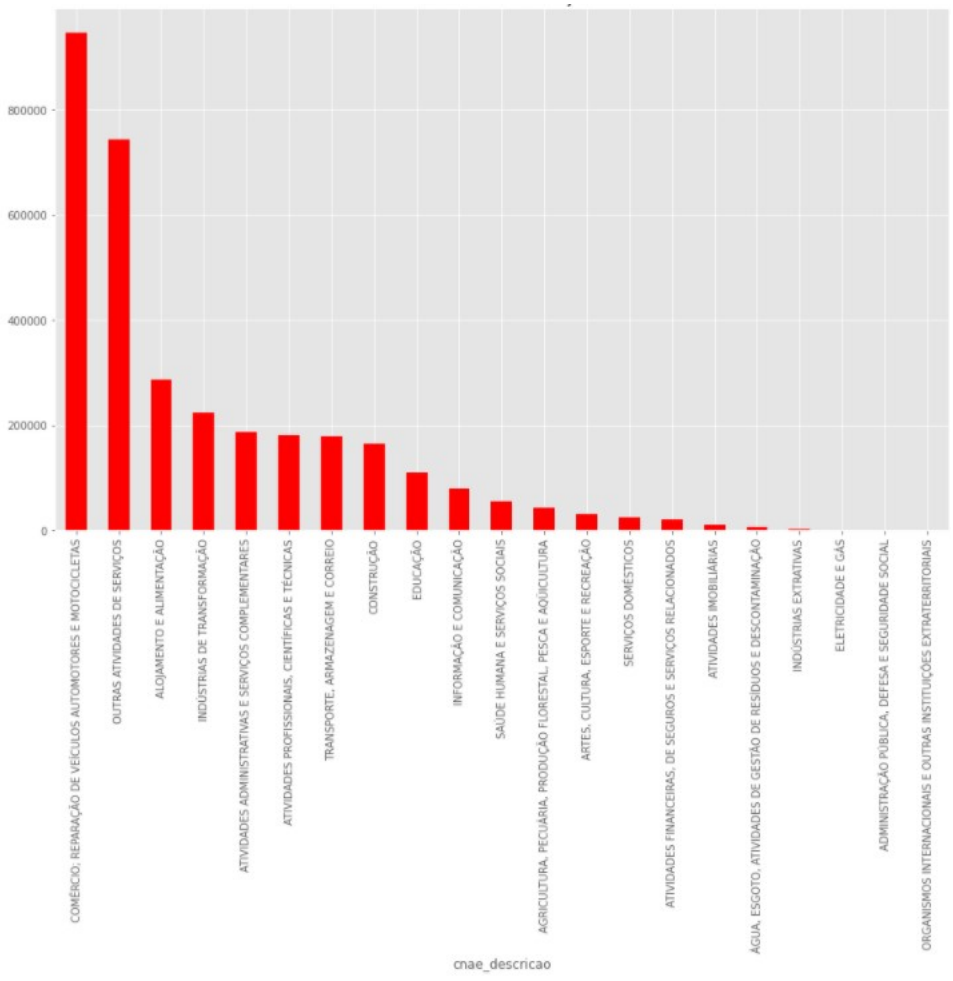
segundo CNAE com maior número de fechamentos (4781400), representa o setor "Comércio varejista de artigos do vestuário e acessórios", será mantido pois é alvo deste estudo. Por outro lado, conforme analisado abaixo, não houve incoerências nos CNAE´s das empresas abertas no período, pois o CNAE com maior abertura foi o CNAE Principal "4781400" (Comércio varejista de artigos do vestuário e acessórios), alvo deste estudo.

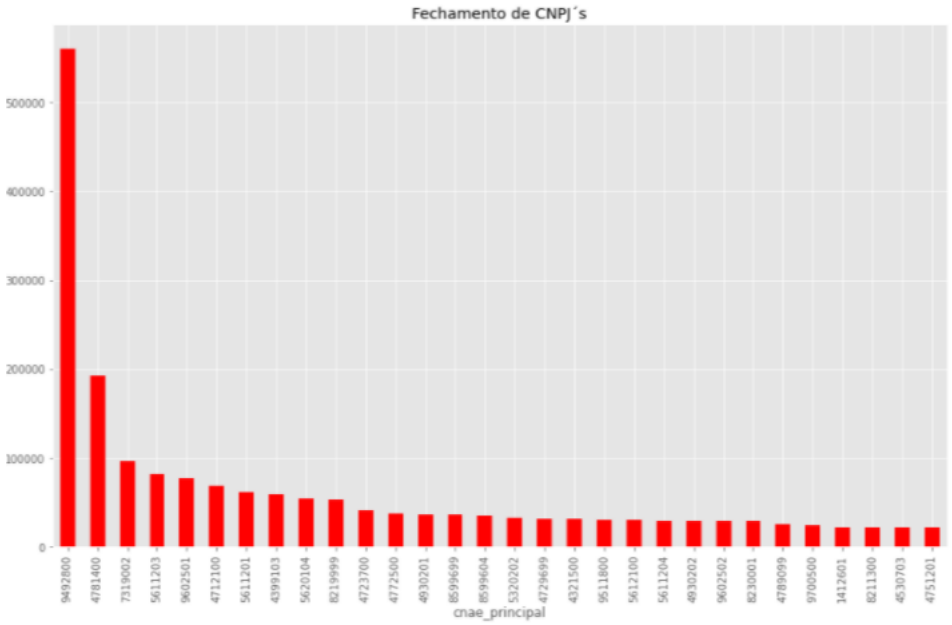
a. Empresas fechadas no período

```
[95] cnae = cnae.sort_values(ascending=False)

[96] cnae
```

```
cnae_descricao
COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS          946155
OUTRAS ATIVIDADES DE SERVIÇOS                                       742907
ALOJAMENTO E ALIMENTAÇÃO                                            287017
INDÚSTRIAS DE TRANSFORMAÇÃO                                         224336
ATIVIDADES ADMINISTRATIVAS E SERVIÇOS COMPLEMENTARES                187746
ATIVIDADES PROFISSIONAIS, CIENTÍFICAS E TÉCNICAS                    180446
TRANSPORTE, ARMAZENAGEM E CORREIO                                   180295
CONSTRUÇÃO                                                          165246
EDUCAÇÃO                                                            110086
INFORMAÇÃO E COMUNICAÇÃO                                             79247
SAÚDE HUMANA E SERVIÇOS SOCIAIS                                      54945
AGRICULTURA, PECUÁRIA, PRODUÇÃO FLORESTAL, PESCA E AQÜICULTURA       42783
ARTES, CULTURA, ESPORTE E RECREAÇÃO                                  30799
SERVIÇOS DOMÉSTICOS                                                  23991
ATIVIDADES FINANCEIRAS, DE SEGUROS E SERVIÇOS RELACIONADOS           21185
ATIVIDADES IMOBILIÁRIAS                                              11303
ÁGUA, ESGOTO, ATIVIDADES DE GESTÃO DE RESÍDUOS E DESCONTAMINAÇÃO      6702
INDÚSTRIAS EXTRATIVAS                                                1774
ELETRICIDADE E GÁS                                                    682
ADMINISTRAÇÃO PÚBLICA, DEFESA E SEGURIDADE SOCIAL                     445
ORGANISMOS INTERNACIONAIS E OUTRAS INSTITUIÇÕES EXTRATERRITORIAIS      10
Name: cnae_descricao, dtype: int64
```



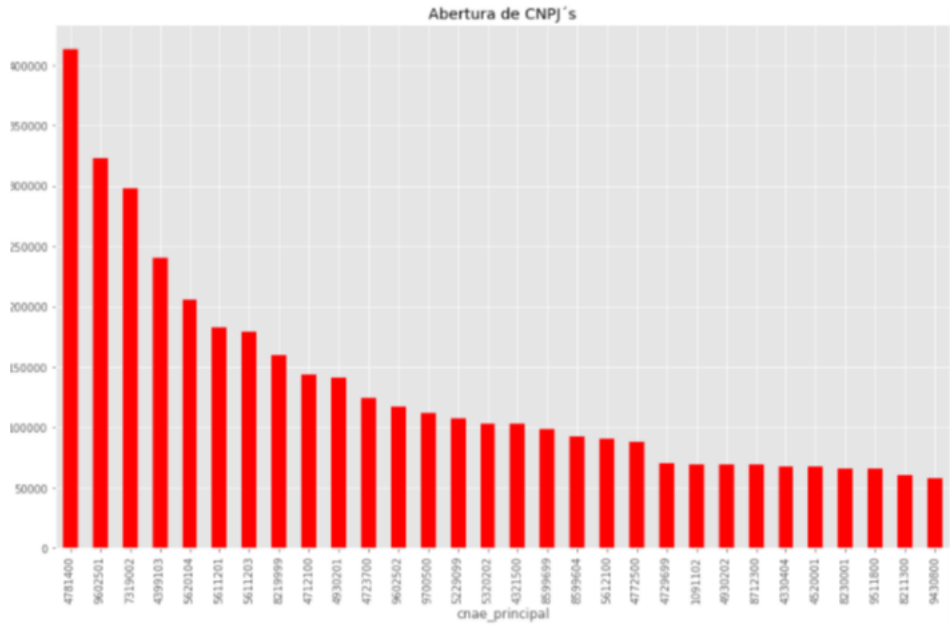


b. Empresas abertas no período

```
[103] cnae = cnae.sort_values(ascending=False)

[104] cnae
```

```
cnae_descricao
COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS          2156387
OUTRAS ATIVIDADES DE SERVIÇOS                                       814499
ALOJAMENTO E ALIMENTAÇÃO                                            801742
INDÚSTRIAS DE TRANSFORMAÇÃO                                         662576
TRANSPORTE, ARMAZENAGEM E CORREIO                                   614371
CONSTRUÇÃO                                                          613958
ATIVIDADES PROFISSIONAIS, CIENTÍFICAS E TÉCNICAS                    535251
ATIVIDADES ADMINISTRATIVAS E SERVIÇOS COMPLEMENTARES                487917
EDUCAÇÃO                                                            290710
SAÚDE HUMANA E SERVIÇOS SOCIAIS                                     216471
INFORMAÇÃO E COMUNICAÇÃO                                            191935
AGRICULTURA, PECUÁRIA, PRODUÇÃO FLORESTAL, PESCA E AQÜICULTURA      119487
SERVIÇOS DOMÉSTICOS                                                 111522
ATIVIDADES FINANCEIRAS, DE SEGUROS E SERVIÇOS RELACIONADOS           85160
ARTES, CULTURA, ESPORTE E RECREAÇÃO                                  69712
ATIVIDADES IMOBILIÁRIAS                                              53615
ÁGUA, ESGOTO, ATIVIDADES DE GESTÃO DE RESÍDUOS E DESCONTAMINAÇÃO     21573
INDÚSTRIAS EXTRATIVAS                                                 4809
ELETRICIDADE E GÁS                                                    4609
ADMINISTRAÇÃO PÚBLICA, DEFESA E SEGURIDADE SOCIAL                     2832
ORGANISMOS INTERNACIONAIS E OUTRAS INSTITUIÇÕES EXTRATERRITORIAIS        21
Name: cnae_descricao, dtype: int64
```



3.2.3 Criação de colunas com faixas de valores

Devido à alta cardinalidade/valores únicos (número de elementos deste conjunto) dos dados, as colunas abaixo serão em uma coluna com faixas de valores, conforme descrito abaixo:

a. Tratamento da variável CAPITAL SOCIAL

INICIAL	FINAL	FAIXA
R$ -	R$ 5.000,00	1
R$ 5.001,00	R$ 10.000,00	2
R$ 10.001,00	R$ 30.000,00	3
R$ 30.001,00	R$ 50.000,00	4
R$ 50.001,00	R$ 70.000,00	5
R$ 70.001,00	R$ 100.000,00	6
R$ 100.001,00	R$ 150.000,00	7
R$ 150.001,00	R$ 200.000,00	8
R$ 200.001,00	R$ 500.000,00	9
R$ 500.001,00	R$ -	10

```
[111] cnpj_final['capital_social'].describe()

      count     11144770
      unique       54293
      top         1000.00
      freq        1994135
      Name: capital_social, dtype: object
```

```
[112] cnpj_final['capital_social'] = cnpj_final['capital_social'].astype(float)
```

```
[113] cnpj_final['capital_social_faixas']=pd.cut(

          cnpj_final['capital_social'],

          bins=[-1, 5000, 10000, 30000, 50000, 70000, 100000, 150000, 200000, 500000, sys.maxsize],

          labels=['1', '2', '3', '4', '5','6','7','8','9','10']
      )
```

b. Tratamento das variáveis de porcentagem de casos acumulados e óbitos por COVID-19

PORCENTAGEM DE CASOS DE COVID		
INICIAL	FINAL	FAIXA
0,000	0,020	0
0,021	0,060	1
0,061	0,100	2
0,101	0,140	3
0,141	0,180	4
0,181	0,220	5
0,221	0,260	6
0,261	0,300	7
0,301	0,340	8
0,341	0,380	9
0,381	0,420	10
0,421	0,460	11
0,461	0,500	12
0,501	0,540	13
0,541	0,580	14
0,581	0,620	15
0,621	0,660	16
0,661	0,700	17
0,701	0,740	18
0,741	0,780	19
0,781	------	20

PORCENTAGEM DE ÓBITOS POR COVID		
INICIAL	FINAL	FAIXA
0,000	0,100	0
0,101	0,200	1
0,201	0,300	2
0,301	0,400	3
0,401	0,500	4
0,501	1,000	5
1,001	1,500	6
1,501	2,000	7
2,001	2,500	8
2,501	3,000	9
3,001	3,500	10
3,501	4,000	11
4,001	4,500	12
4,501	5,000	13
5,001	6,000	14
6,001	7,000	15
7,001	8,000	16
8,001	9,000	17
9,001	12,000	18
12,001	15,000	19
15,001	------	20

```
[114] cnpj_final['casos_acumulados_porcentagem_covid_faixas']=pd.cut(
        cnpj_final['casos_acumulados_porcentagem_covid'],
        bins=[-1, 0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 6.0, 7.0, 8.0, 9.0, 12.0, 15.0, sys.maxsize],
          labels=['0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20'])

     cnpj_final['obitos_acumulados_porcentagem_covid_faixas']=pd.cut(
        cnpj_final['obitos_acumulados_porcentagem_covid'],
        bins=[-1, 0.02, 0.06, 0.1, 0.14, 0.18, 0.22, 0.26, 0.3, 0.34, 0.38, 0.42, 0.46, 0.5, 0.54, 0.58, 0.62, 0.66, 0.7, 0.74, 0.78, sys.maxsize],
          labels=['0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20'])
```

c. Tratamento das variáveis de IDH

INICIAL	FINAL	FAIXA
0,000	0,100	0
0,101	0,200	1
0,201	0,300	2
0,301	0,400	3
0,401	0,500	4
0,501	0,600	5
0,601	0,700	6
0,701	0,800	7
0,801	0,900	8
0,901	0,999	9

```
115] cnpj_final['idh_municipio'] = cnpj_final['idh_municipio'].astype(float)
     cnpj_final['idh_educacao'] = cnpj_final['idh_educacao'].astype(float)
     cnpj_final['idh_longevidade'] = cnpj_final['idh_longevidade'].astype(float)
     cnpj_final['idh_renda'] = cnpj_final['idh_renda'].astype(float)
```

```
116] cnpj_final['idh_municipios_faixas']=pd.cut(
        cnpj_final['idh_municipio'],
        bins=[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, sys.maxsize],
          labels=['0', '1', '2', '3', '4', '5', '6', '7', '8', '9'])

     cnpj_final['idh_educacao_municipios_faixas']=pd.cut(
        cnpj_final['idh_educacao'],
        bins=[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, sys.maxsize],
          labels=['0', '1', '2', '3', '4', '5', '6', '7', '8', '9'])

     cnpj_final['idh_longevidade_municipios_faixas']=pd.cut(
        cnpj_final['idh_longevidade'],
        bins=[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, sys.maxsize],
          labels=['0', '1', '2', '3', '4', '5', '6', '7', '8', '9'])

     cnpj_final['idh_renda_municipios_faixas']=pd.cut(
        cnpj_final['idh_renda'],
        bins=[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, sys.maxsize],
          labels=['0', '1', '2', '3', '4', '5', '6', '7', '8', '9'])
```

d. Tratamento das variáveis de PIB PER CAPITA

INICIAL	FINAL	FAIXA
R$ -	R$ 5.000,00	1
R$ 5.001,00	R$ 10.000,00	2
R$ 10.001,00	R$ 20.000,00	3
R$ 20.001,00	R$ 30.000,00	4
R$ 30.001,00	R$ 40.000,00	5
R$ 40.001,00	R$ 50.000,00	6
R$ 50.001,00	R$ 60.000,00	7
R$ 60.001,00	R$ 70.000,00	8
R$ 70.001,00	R$ 80.000,00	9
R$ 80.001,00	R$ 90.000,00	10
R$ 90.001,00	R$ 100.000,00	11
R$ 100.001,00	R$ 150.000,00	12
R$ 150.001,00	R$ 200.000,00	13
R$ 200.001,00	R$ 300.000,00	14
R$ 300.001,00	R$ -	15

```
[117] cnpj_final['pib_per_capita'] = cnpj_final['pib_per_capita'].astype(float)
```

```
[118] cnpj_final['pib_per_capita_faixas']=pd.cut(

        cnpj_final['pib_per_capita'],

        bins=[-1, 5000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 150000, 200000, 300000, sys.maxsize],

        labels=['1', '2', '3', '4', '5','6','7','8','9','10','11','12','13','14','15']
    )
```

3.2.3 Gravação do dataset “dataset_final.csv”

Neste momento foi gerado o dataset “dataset_final.csv”. Este arquivo será base para a etapa seguinte de Exploração dos Dados, contendo 10.534.042 registros.

```
[125] cnpj_final.shape

      (10534042, 39)

[126] cnpj_final.dtypes

      cnpj_basico                                object
      cnpj_ordem                                 object
      cnpj_dv                                    object
      matriz_filial                              object
      capital_social                            float64
      capital_social_faixas                    category
      codigo_situacao_cadastral                  object
      data_inicio_atividade                      object
      data_situacao_cadastral                    object
      data_situacao_cadastral_ano_mes            object
      situacao_cadastral_target                  object
      codigo_ibge                                 int64
      codigo_municipio_siafi                      int64
      descricao_municipio                        object
      regiao                                     object
      estado                                     object
      interior_metropolitana                      int64
      casosAcumulado_covid                      float64
      obitosAcumulado_covid                     float64
      casos_acumulados_porcentagem_covid        float64
      obitos_acumulados_porcentagem_covid       float64
      casos_acumulados_porcentagem_covid_faixas category
      obitos_acumulados_porcentagem_covid_faixas category
      idh_municipio                             float64
      idh_educacao                              float64
      idh_longevidade                           float64
      idh_renda                                 float64
      idh_municipios_faixas                    category
      idh_educacao_municipios_faixas           category
      idh_longevidade_municipios_faixas        category
      idh_renda_municipios_faixas              category
      populacaoTCU2019                            int64
      pib                                        object
      pib_per_capita                            float64
      pib_per_capita_faixas                    category
      cnae_principal                             object
      cnae_divisao                                int64
      cnae_secao                                 object
      cnae_descricao                             object
      dtype: object
```

4. Análise e Exploração dos Dados

Esta etapa do trabalho foi realizada no notebook “0005 - Exploração Visual dos Dados.ipynb”.

A análise dos dados foi dividida em dois tópicos, conforme descrito abaixo.

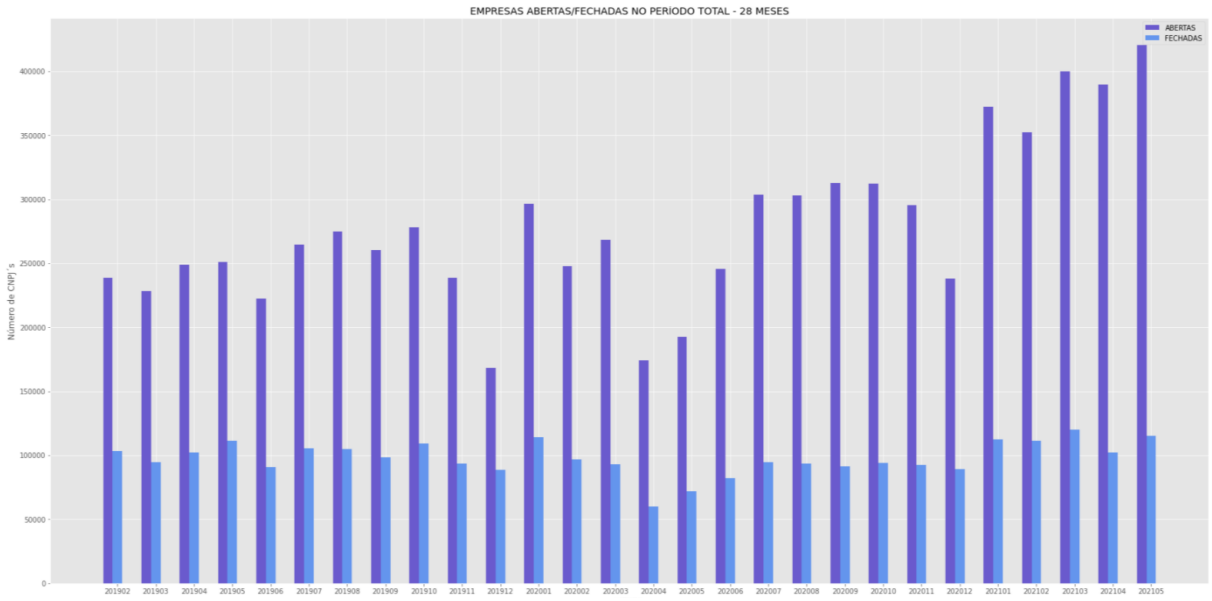
4.1 Exploração Visual dos Dados

Neste momento foi analisado, a partir das variáveis coletadas, o impacto que cada uma delas teve na abertura ou fechamento das empresas. O intuito, é a busca da confirmação, amplamente divulgada pelos meios de comunicação e órgãos

oficiais, de que a pandemia contribuiu significativamente para o fechamento das empresas.

4.1.1 Análise do Fechamento e Abertura de Empresas no período

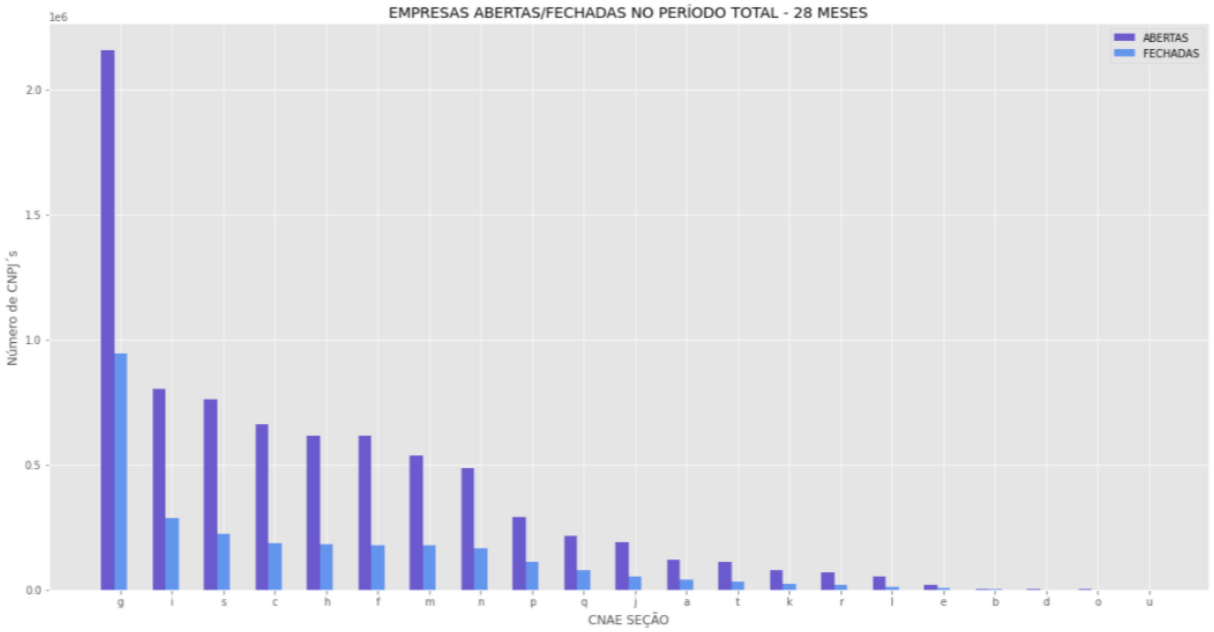
Visualização do total de empresas abertas e fechadas durante os 28 meses (01/02/2019 à 31/05/2021) estudados neste trabalho.



4.1.2 Análise do Fechamento e Abertura de Empresas por CNAE

Visualização do total de empresas abertas e fechadas por setor econômico. A intenção é a identificação dos setores da economia que mais sentiram os impactos da pandemia. Foi considerada a seção do CNAE, conforme tabela abaixo.

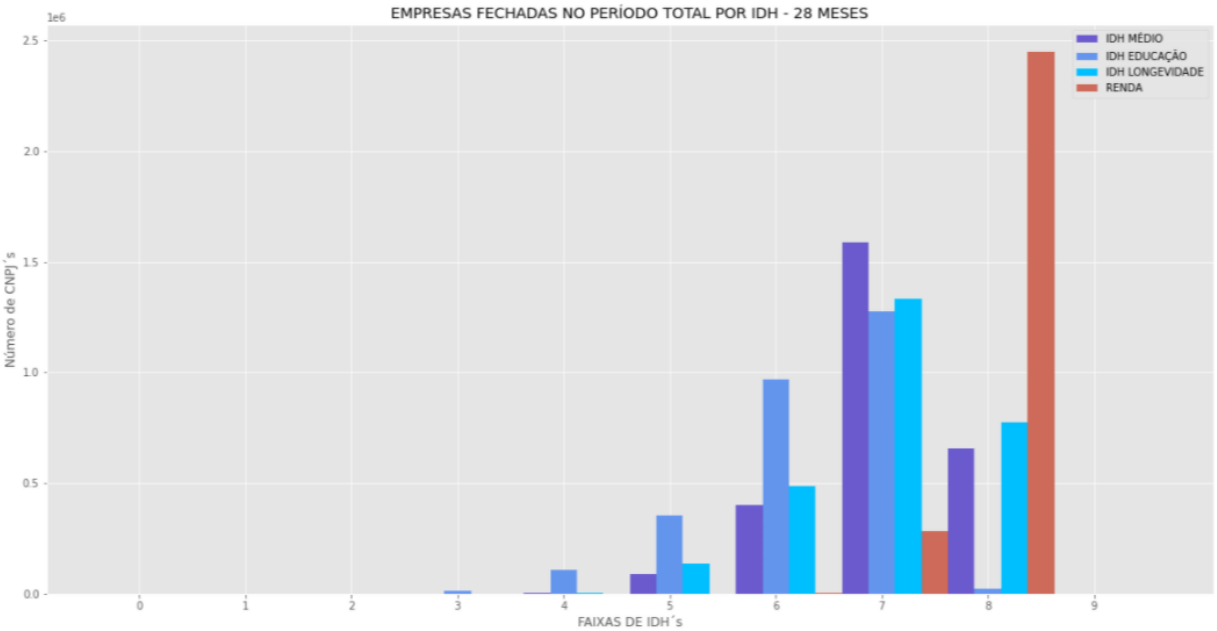
Seção	Divisões	Denominação
A	01 .. 03	AGRICULTURA, PECUÁRIA, PRODUÇÃO FLORESTAL, PESCA E AQÜICULTURA
B	05 .. 09	INDÚSTRIAS EXTRATIVAS
C	10 .. 33	INDÚSTRIAS DE TRANSFORMAÇÃO
D	35 .. 35	ELETRICIDADE E GÁS
E	36 .. 39	ÁGUA, ESGOTO, ATIVIDADES DE GESTÃO DE RESÍDUOS E DESCONTAMINAÇÃO
F	41 .. 43	CONSTRUÇÃO
G	45 .. 47	COMÉRCIO; REPARAÇÃO DE VEÍCULOS AUTOMOTORES E MOTOCICLETAS
H	49 .. 53	TRANSPORTE, ARMAZENAGEM E CORREIO
I	55 .. 56	ALOJAMENTO E ALIMENTAÇÃO
J	58 .. 63	INFORMAÇÃO E COMUNICAÇÃO
K	64 .. 66	ATIVIDADES FINANCEIRAS, DE SEGUROS E SERVIÇOS RELACIONADOS
L	68 .. 68	ATIVIDADES IMOBILIÁRIAS
M	69 .. 75	ATIVIDADES PROFISSIONAIS, CIENTÍFICAS E TÉCNICAS
N	77 .. 82	ATIVIDADES ADMINISTRATIVAS E SERVIÇOS COMPLEMENTARES
O	84 .. 84	ADMINISTRAÇÃO PÚBLICA, DEFESA E SEGURIDADE SOCIAL
P	85 .. 85	EDUCAÇÃO
Q	86 .. 88	SAÚDE HUMANA E SERVIÇOS SOCIAIS
R	90 .. 93	ARTES, CULTURA, ESPORTE E RECREAÇÃO
S	94 .. 96	OUTRAS ATIVIDADES DE SERVIÇOS
T	97 .. 97	SERVIÇOS DOMÉSTICOS
U	99 .. 99	ORGANISMOS INTERNACIONAIS E OUTRAS INSTITUIÇÕES EXTRATERRITORIAIS

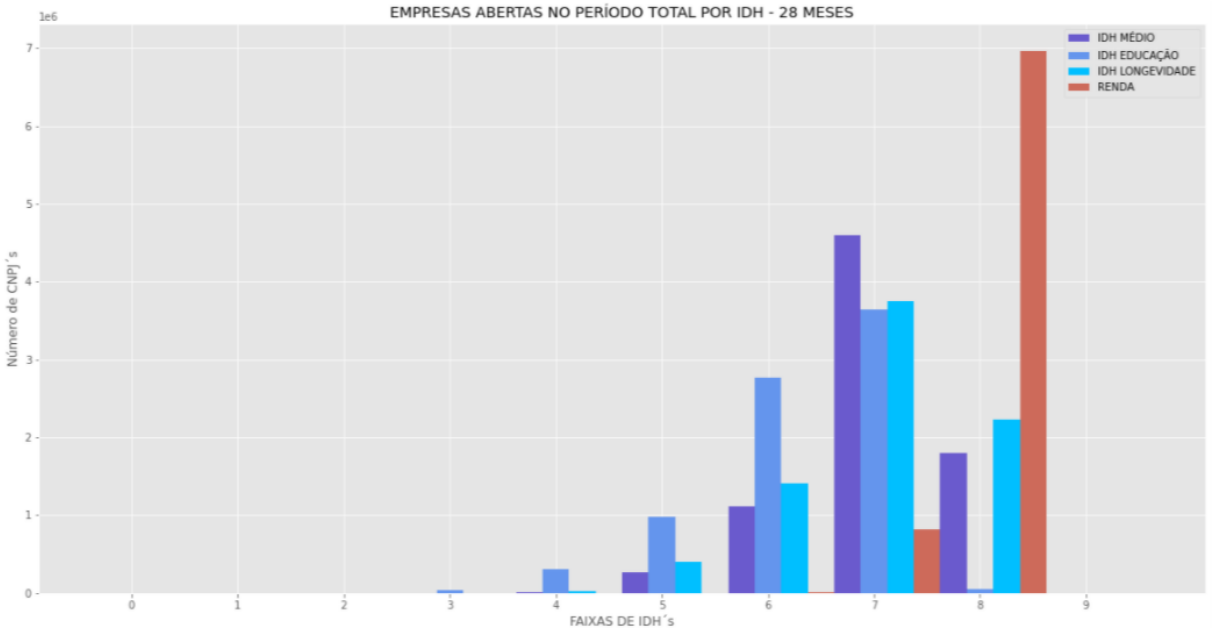


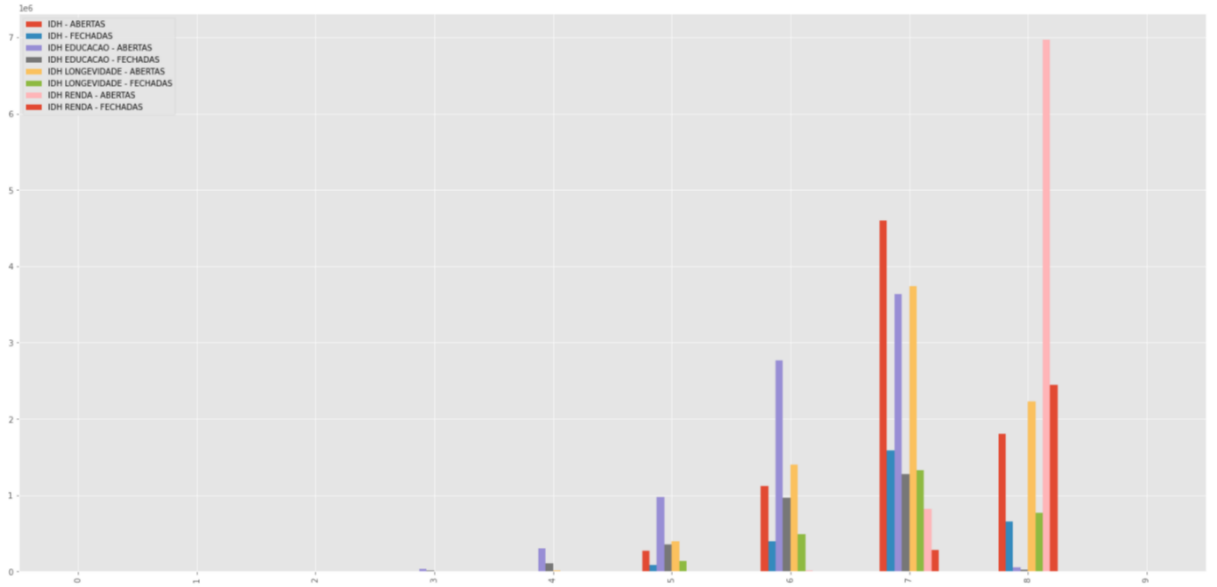
4.1.3 Análise do Fechamento e Abertura de Empresas por IDH

A análise neste momento tem como objetivo demonstrar a relação entre a abertura e fechamento das empresas considerando o índice de desenvolvimento humano do município. Utilizaremos os valores das faixas dos IDH´s criados em momento anterior.

INICIAL	FINAL	FAIXA
0,000	0,100	0
0,101	0,200	1
0,201	0,300	2
0,301	0,400	3
0,401	0,500	4
0,501	0,600	5
0,601	0,700	6
0,701	0,800	7
0,801	0,900	8
0,901	0,999	9


EMPRESAS FECHADAS NO PERÍODO TOTAL POR IDH - 28 MESES
1e6
Número de CNPJ´s
FAIXAS DE IDH´s
IDH MÉDIO
IDH EDUCAÇÃO
IDH LONGEVIDADE
RENDA

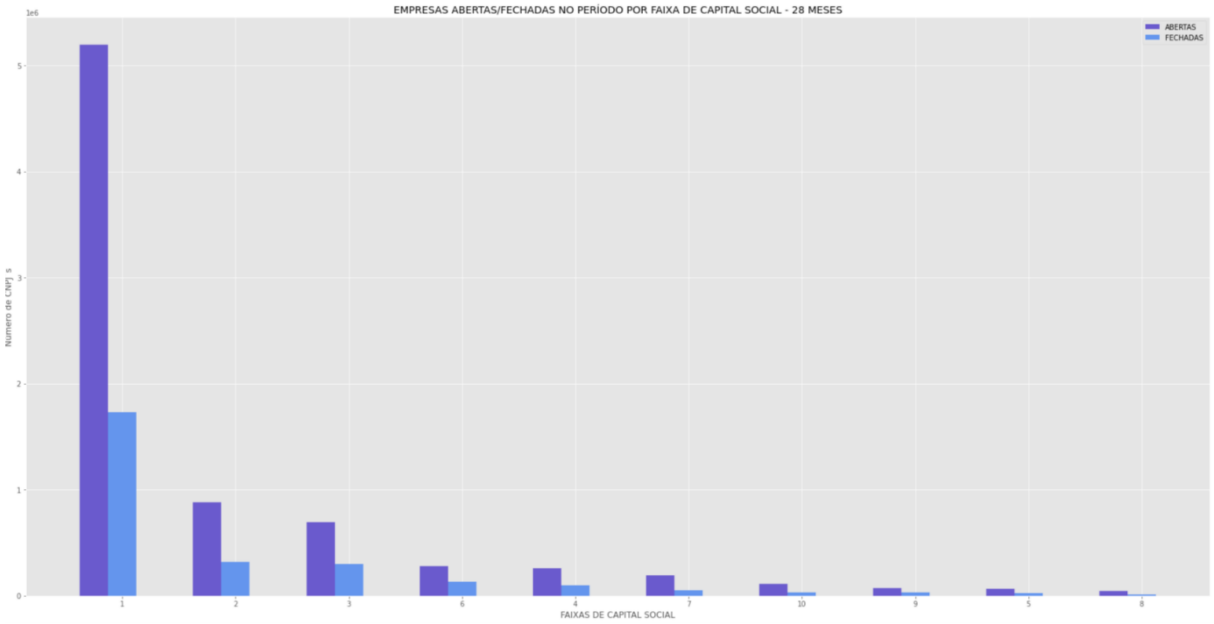

EMPRESAS ABERTAS NO PERÍODO TOTAL POR IDH - 28 MESES
1e6
Número de CNPJ´s
FAIXAS DE IDH´s
IDH MÉDIO
IDH EDUCAÇÃO
IDH LONGEVIDADE
RENDA


1e6
IDH - ABERTAS
IDH - FECHADAS
IDH EDUCACAO - ABERTAS
IDH EDUCACAO - FECHADAS
IDH LONGEVIDADE - ABERTAS
IDH LONGEVIDADE - FECHADAS
IDH RENDA - ABERTAS
IDH RENDA - FECHADAS

4.1.4 Análise do Fechamento e Abertura de Empresas por CAPITAL SOCIAL

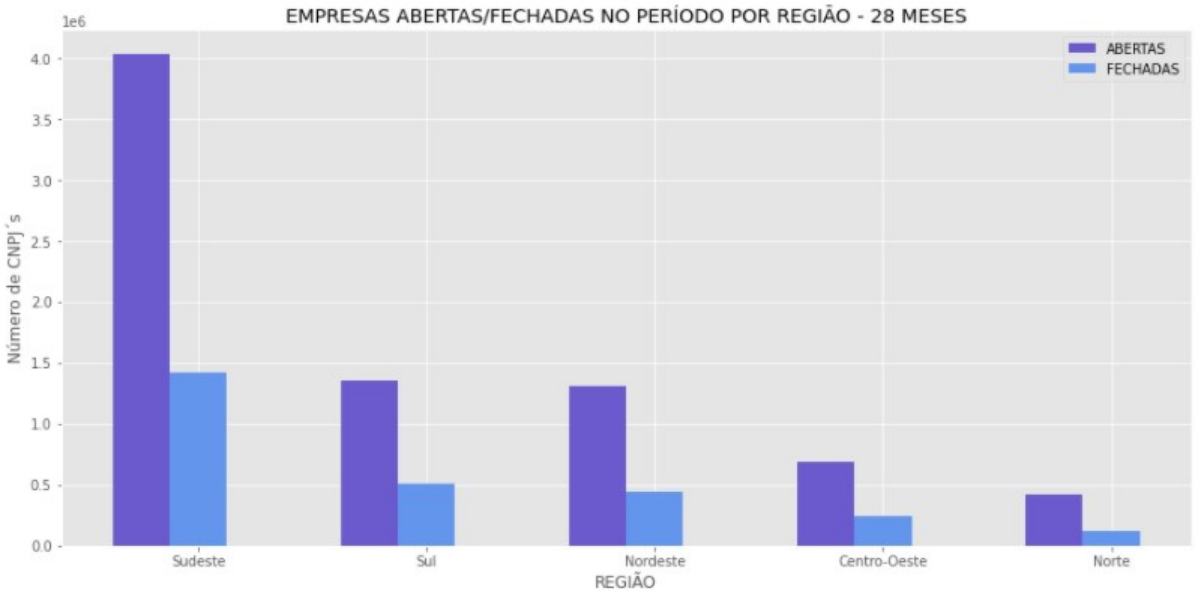
Considerado, neste gráfico, o número de empresas pelo valor das faixas do capital da empresa, conforme tabela abaixo.

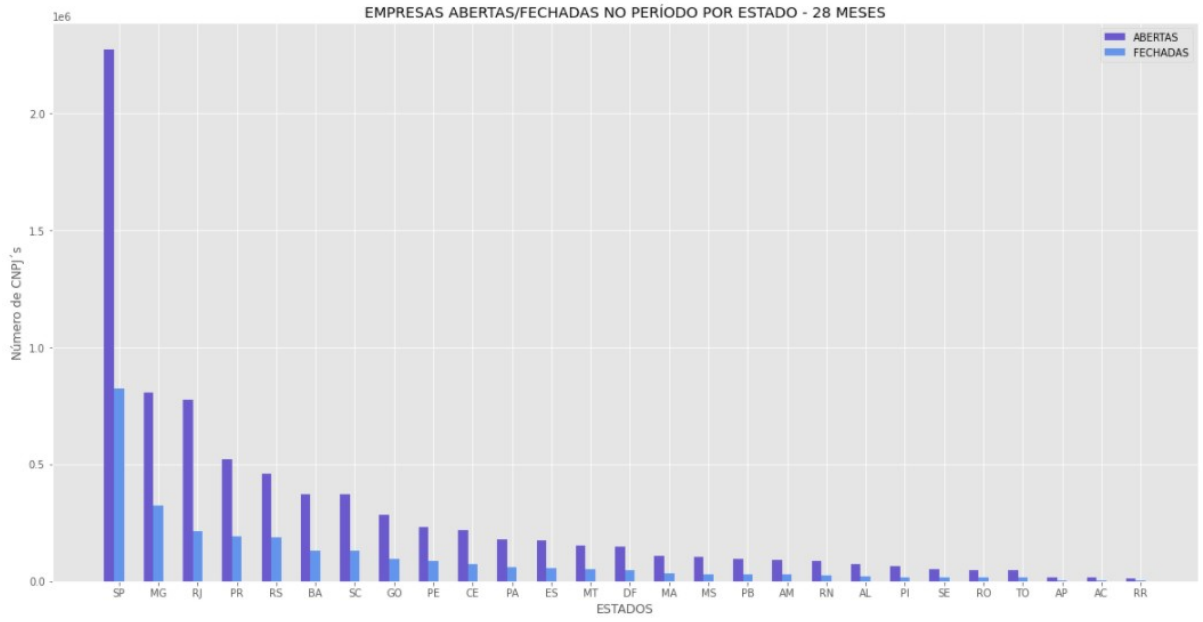
INICIAL	FINAL	FAIXA
R$ -	R$ 5.000,00	1
R$ 5.001,00	R$ 10.000,00	2
R$ 10.001,00	R$ 30.000,00	3
R$ 30.001,00	R$ 50.000,00	4
R$ 50.001,00	R$ 70.000,00	5
R$ 70.001,00	R$ 100.000,00	6
R$ 100.001,00	R$ 150.000,00	7
R$ 150.001,00	R$ 200.000,00	8
R$ 200.001,00	R$ 500.000,00	9
R$ 500.001,00	R$ -	10

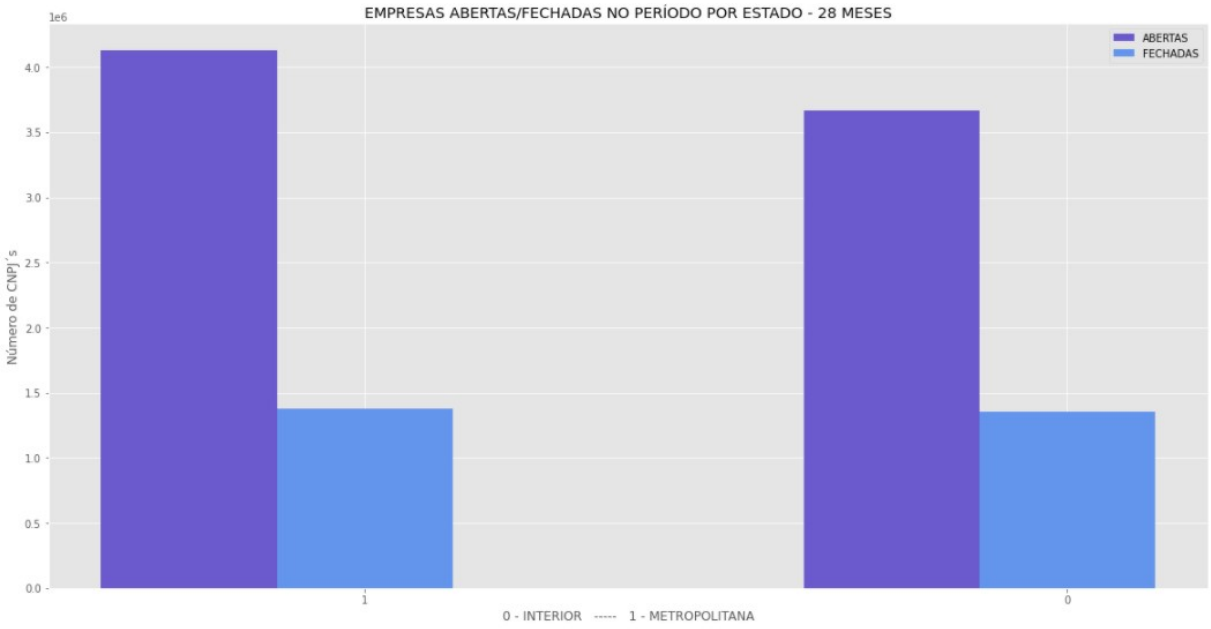


4.1.5 Análise do Fechamento e Abertura de Empresas por REGIÃO, ESTADO E INTERIOR/METROPOLITANA

Visualização do total de empresas abertas e fechadas considerando a localização física.


EMPRESAS ABERTAS/FECHADAS NO PERÍODO POR REGIÃO - 28 MESES
1e6
ABERTAS
FECHADAS
Número de CNPJ´s
Sudeste
Sul
Nordeste
Centro-Oeste
Norte
REGIÃO


EMPRESAS ABERTAS/FECHADAS NO PERÍODO POR ESTADO - 28 MESES
1e6
ABERTAS
FECHADAS
Número de CNPJ´s
SP MG RJ PR RS BA SC GO PE CE PA ES MT DF MA MS PB AM RN AL PI SE RO TO AP AC RR
ESTADOS

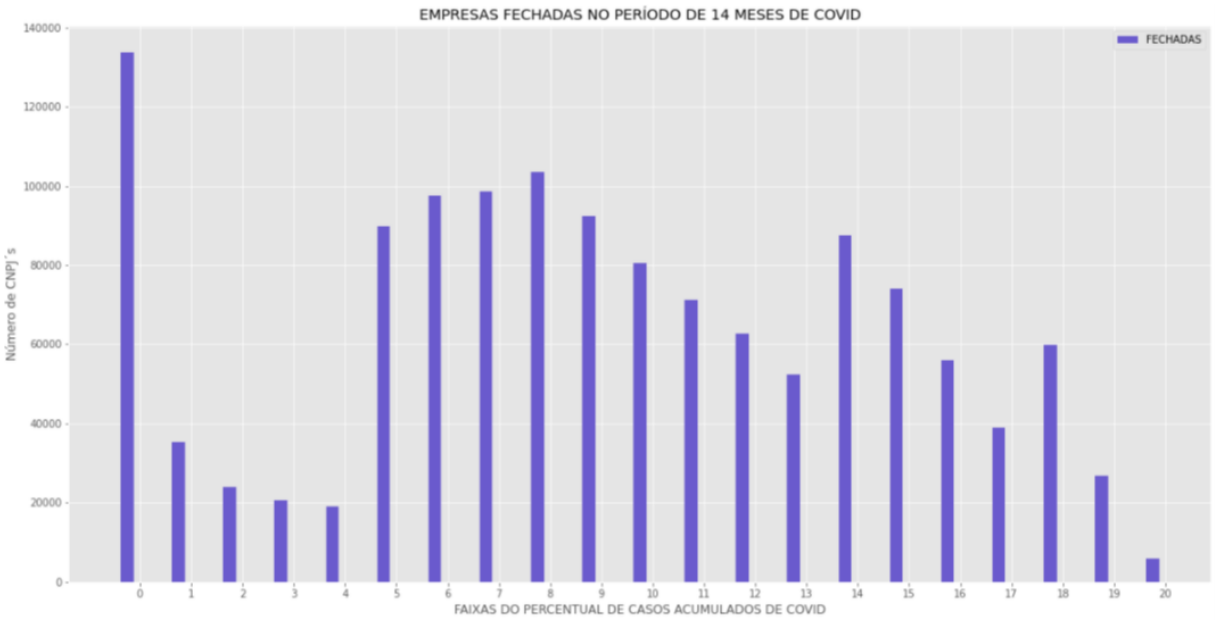

EMPRESAS ABERTAS/FECHADAS NO PERÍODO POR ESTADO - 28 MESES
1e6
ABERTAS
FECHADAS
Número de CNPJ´s
1
0
0 - INTERIOR ----- 1 - METROPOLITANA

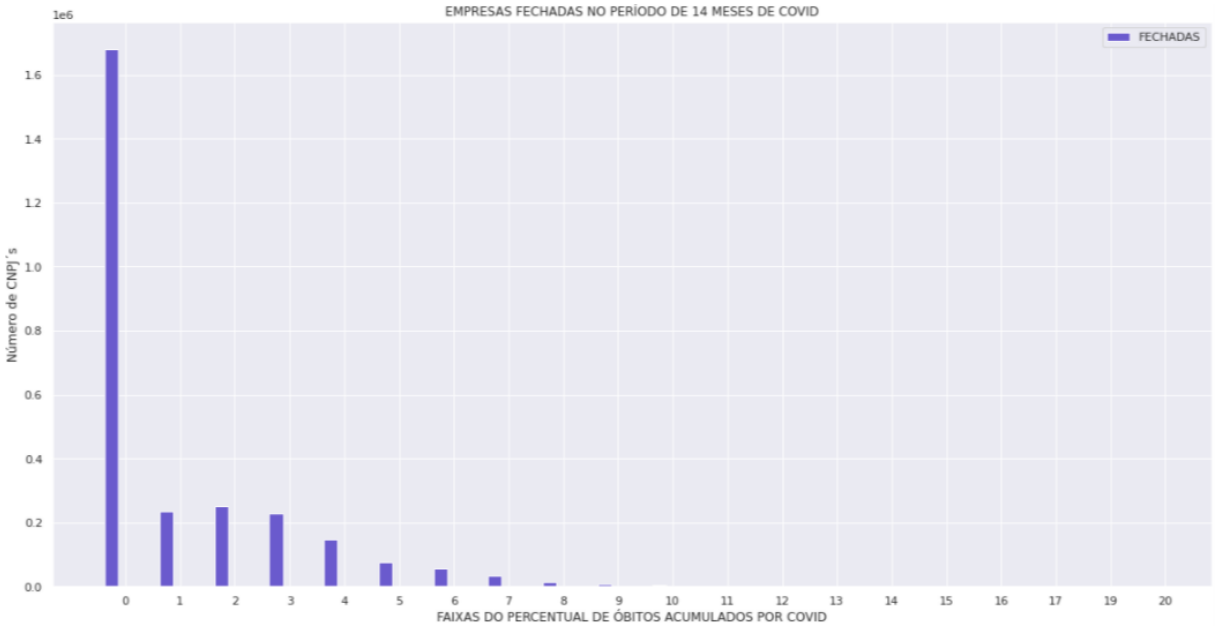
4.1.6 Análise do Fechamento e Abertura de Empresas por COVID

O gráfico mostra o número de empresas que abriram ou fecharam em decorrência da pandemia de COVID-19. Considerado as faixas percentuais dos números de casos e óbitos, conforme tabela.

PORCENTAGEM DE CASOS DE COVID		
INICIAL	FINAL	FAIXA
0,000	0,020	0
0,021	0,060	1
0,061	0,100	2
0,101	0,140	3
0,141	0,180	4
0,181	0,220	5
0,221	0,260	6
0,261	0,300	7
0,301	0,340	8
0,341	0,380	9
0,381	0,420	10
0,421	0,460	11
0,461	0,500	12
0,501	0,540	13
0,541	0,580	14
0,581	0,620	15
0,621	0,660	16
0,661	0,700	17
0,701	0,740	18
0,741	0,780	19
0,781	------	20

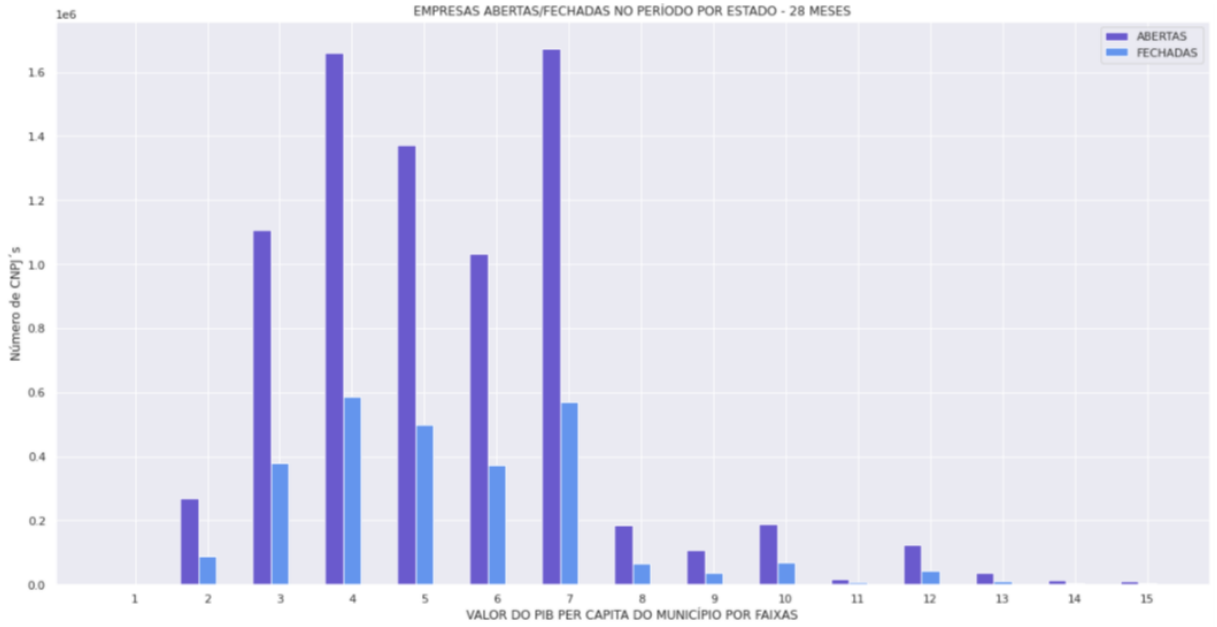
PORCENTAGEM DE ÓBITOS POR COVID		
INICIAL	FINAL	FAIXA
0,000	0,100	0
0,101	0,200	1
0,201	0,300	2
0,301	0,400	3
0,401	0,500	4
0,501	1,000	5
1,001	1,500	6
1,501	2,000	7
2,001	2,500	8
2,501	3,000	9
3,001	3,500	10
3,501	4,000	11
4,001	4,500	12
4,501	5,000	13
5,001	6,000	14
6,001	7,000	15
7,001	8,000	16
8,001	9,000	17
9,001	12,000	18
12,001	15,000	19
15,001	------	20





4.1.7 Análise do Fechamento e Abertura de Empresas por PIB PER CAPITA

Visualização do total de empresas abertas e fechadas considerando o valor do PIB PER CAPITA do município de localização da empresa.



4.2 Análise Exploratória das variáveis Categóricas e Qualitativas

Nesta etapa será realizada a separação e análise das várias em categóricas e quantitativas. Além dessa ação, foi feita a eliminação de variáveis desnecessárias para a modelagem e utilizadas somente para a visualização gráfica feita na etapa anterior.

4.2.1 Análise Exploratória das variáveis Quantitativas

Foram elaboradas estatísticas descritivas das variáveis quantitativas e gerado uma de calor das colunas fortemente correlacionadas.

```
cnpj_final_quantitativas.nunique()
```

capital_social	54293
casosAcumulado_covid	32081
obitosAcumulado_covid	4641
casos_acumulados_porcentagem_covid	16422
obitos_acumulados_porcentagem_covid	531
idh_municipio	349
idh_educacao	466
idh_longevidade	390
idh_renda	220
pib_per_capita	5562
capital_social_faixas	10
casos_acumulados_porcentagem_covid_faixas	21
obitos_acumulados_porcentagem_covid_faixas	21
idh_municipios_faixas	5
idh_educacao_municipios_faixas	7
idh_longevidade_municipios_faixas	6
idh_renda_municipios_faixas	3
pib_per_capita_faixas	15

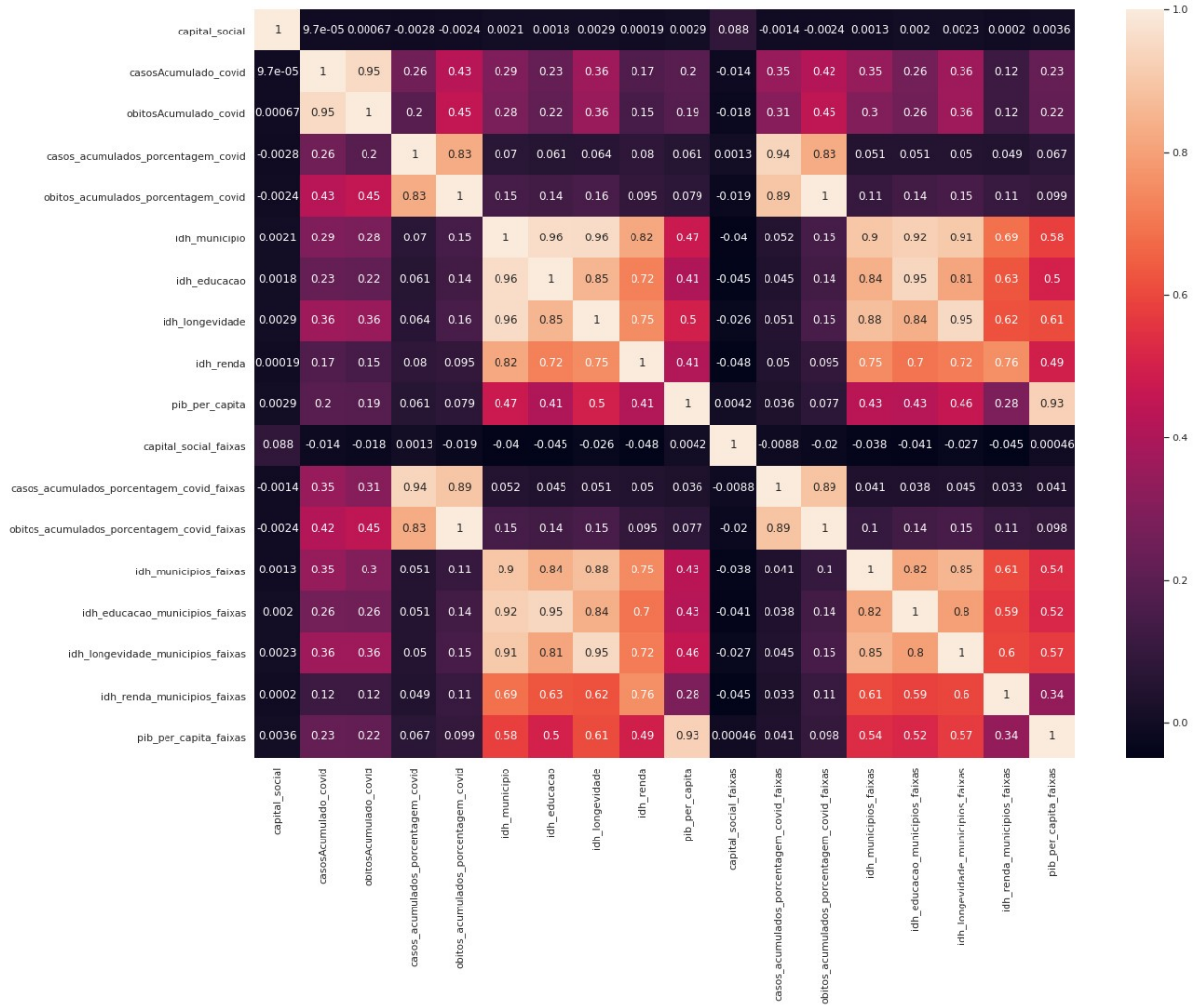
```
cnpj_final_quantitativas.median()
```

capital_social	3,000.00
casosAcumulado_covid	485.00
obitosAcumulado_covid	10.00
casos_acumulados_porcentagem_covid	1.22
obitos_acumulados_porcentagem_covid	0.04
idh_municipio	0.76
idh_educacao	0.70
idh_longevidade	0.75
idh_renda	0.84
pib_per_capita	36,445.75
capital_social_faixas	1.00
casos_acumulados_porcentagem_covid_faixas	6.00
obitos_acumulados_porcentagem_covid_faixas	1.00
idh_municipios_faixas	7.00
idh_educacao_municipios_faixas	6.00
idh_longevidade_municipios_faixas	7.00
idh_renda_municipios_faixas	8.00
pib_per_capita_faixas	5.00

```
cnpj_final_quantitativas.describe()
```

	capital_social	casosAcumulado_covid	obitosAcumulado_covid	casos_acumulados_porcentagem_covid	obitos_acumulados_porcentagem_covid	idh_municipio	idh_educacao	idh_longevidade	idh_renda	pib_per_capita	capital_social_faixas
count	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00
mean	29,693,229.55	61,062.54	2,491.26	3.86	0.11	0.75	0.67	0.75	0.84	40,086.92	2.01
std	1,397,425,337.23	164,444.57	6,984.40	4.63	0.13	0.06	0.08	0.08	0.03	26,093.01	1.94
min	0.00	0.00	0.00	0.00	0.00	0.42	0.21	0.40	0.67	4,788.18	1.00
25%	1,000.00	0.00	0.00	0.00	0.00	0.72	0.63	0.71	0.82	22,965.66	1.00
50%	3,000.00	485.00	10.00	1.22	0.04	0.76	0.70	0.75	0.84	36,445.75	1.00
75%	10,000.00	24,621.00	735.00	7.29	0.23	0.80	0.72	0.82	0.85	54,426.08	2.00
max	438,002,148,502.00	783,191.00	30,596.00	42.41	3.13	0.86	0.82	0.89	0.89	583,171.85	10.00

	casos_acumulados_porcentagem_covid_faixas	obitos_acumulados_porcentagem_covid_faixas	idh_municipios_faixas	idh_educacao_municipios_faixas	idh_longevidade_municipios_faixas	idh_renda_municipios_faixas	pib_per_capita_faixas
count	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00	10,534,042.00
mean	7.68	2.79	7.02	6.26	7.00	7.89	5.40
std	7.89	3.22	0.72	0.87	0.83	0.32	2.09
min	0.00	0.00	4.00	2.00	3.00	6.00	1.00
25%	0.00	0.00	7.00	6.00	7.00	8.00	4.00
50%	6.00	1.00	7.00	6.00	7.00	8.00	5.00
75%	16.00	6.00	7.00	7.00	8.00	8.00	7.00
max	20.00	20.00	8.00	8.00	8.00	8.00	15.00



4.2.2 Análise Exploratória das variáveis Categóricas

Nesta etapa foram a analisadas as variáveis categóricas e o tratamento para utilização nos modelos de Machine Learning.

Levando em consideração a cardinalidade (quantidade de valores únicos assumidos pelas variáveis categóricas), ou seja, a quantidade de valores únicos de cada variável, concluímos:

```
cnpj_final_categoricas.nunique()
```

```
matriz_filial              2
regiao                     5
interior_metropolitana     2
estado                    27
cnae_divisao              87
cnae_secao                21
dtype: int64
```

- as variáveis “matriz_filial”, “regiao” e “interior_metropolitana”, possuem menos que 10 valores únicos. Dessa forma, são consideradas variáveis de baixa cardinalidade. O tratamento correto para variáveis categóricas de baixa cardinalidade é a criação de dummies.

Criação de variáveis dummy

```
[56] cnpj_final_dummy = pd.get_dummies(cnpj_final,
                                       columns = ['matriz_filial',
                                                  'regiao',
                                                  'interior_metropolitana'],
                                        drop_first = True,
                                        prefix = ['matriz_filial',
                                                  'regiao',
                                                  'interior_metropolitana'],
                                        prefix_sep = '_' )
```

- por outro lado, as variáveis “estado”, “cnae_divisao” e “cnae_secao” têm mais que 10 ocorrências únicas. Portanto, são consideradas de alta cardinalidade. O tratamento correto para variáveis de alta cardinalidade é a criação de label encoders.

Criação de variáveis LabelEncoder

```
[ ] from sklearn.preprocessing import LabelEncoder

[ ] le = LabelEncoder()

[ ] cnpj_final_dummy.reset_index(inplace=True, drop=True)

[ ] le_estado = pd.DataFrame((le.fit_transform(cnpj_final_dummy['estado'])), columns=['le_estado'])
    le_cnae_divisao = pd.DataFrame((le.fit_transform(cnpj_final_dummy['cnae_divisao'])), columns=['le_cnae_divisao'])
    le_cnae_secao = pd.DataFrame((le.fit_transform(cnpj_final_dummy['cnae_secao'])), columns=['le_cnae_secao'])

[ ] cnpj_final_tratado = pd.concat([cnpj_final_dummy, le_estado,le_cnae_divisao,le_cnae_secao], axis=1)
```

4.3 Tratamento e Geração do Dataset antes dos Modelos de Machine Learning

Nesta etapa será realizado o tratamento final e a geração do dataset a ser utilizado nos modelos de Machine Learning.

As colunas “estado”, “cnae_divisao” e “cnae_secao”, transformadas durante o processo de Label Encoder forma eliminadas.

O dataset “cnpj_final_tratado.csv” gerado conforme especificações abaixo:

```
cnpj_final_tratado.dtypes
```

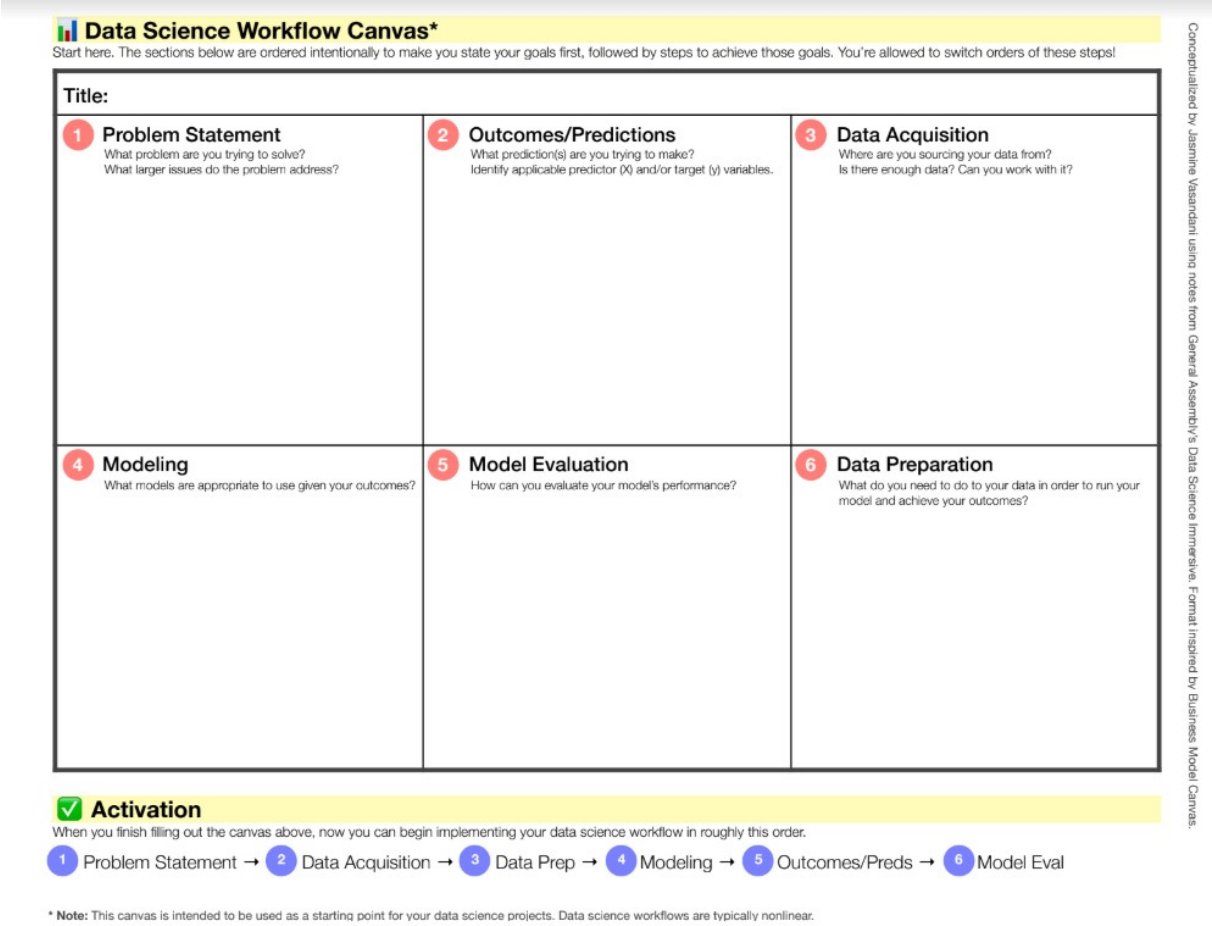
```
capital_social                              float64
capital_social_faixas                         int64
data_situacao_cadastral_ano_mes               int64
situacao_cadastral_target                     int64
casosAcumulado_covid                        float64
obitosAcumulado_covid                       float64
casos_acumulados_porcentagem_covid          float64
obitos_acumulados_porcentagem_covid         float64
casos_acumulados_porcentagem_covid_faixas     int64
obitos_acumulados_porcentagem_covid_faixas    int64
idh_municipio                               float64
idh_educacao                                float64
idh_longevidade                             float64
idh_renda                                   float64
idh_municipios_faixas                         int64
idh_educacao_municipios_faixas                int64
idh_longevidade_municipios_faixas             int64
idh_renda_municipios_faixas                   int64
pib_per_capita                              float64
pib_per_capita_faixas                         int64
idh_municipios_faixas_graf                 category
idh_educacao_municipios_faixas_graf        category
idh_longevidade_municipios_faixas_graf     category
idh_renda_municipios_faixas_graf           category
matriz_filial_2                               uint8
regiao_Nordeste                               uint8
regiao_Norte                                  uint8
regiao_Sudeste                                uint8
regiao_Sul                                    uint8
interior_metropolitana_1                      uint8
le_estado                                     int64
le_cnae_divisao                               int64
le_cnae_secao                                 int64
dtype: object
```

5. Criação de Modelos de Machine Learning

Conforme o documento de instruções para o TCC, essa etapa é obrigatória. Nessa seção você irá descrever as ferramentas e algoritmos utilizados. Se você utilizou ferramentas visuais como Knime e Rapid Miner, coloque aqui um print do seu modelo e a descrição detalhada do workflow do seu modelo. Caso você tenha escrito scripts em Python, por exemplo, coloque aqui o seu script. Explique as *features* utilizadas, faça a comparação entre diferentes algoritmos/modelos, justifique a escolha por determinado modelo, os parâmetros utilizados, etc. Por fim, salienta-se que embora você possa utilizar ferramentas como KNIME e RapidMiner para testar protótipos do seu modelo de ML, encorajamos você a fazer seus modelos em Python ou R.

6. Apresentação dos Resultados

Nessa seção você deve apresentar os resultados obtidos. Apresente gráficos, *dashboards*, conte a sua história de forma bastante criativa. Aqui você pode utilizar os modelos de Canvas propostos por Dourard (clique aqui) ou por Vasandani (clique aqui).



Data Science Workflow Canvas*

Start here. The sections below are ordered intentionally to make you state your goals first, followed by steps to achieve those goals. You're allowed to switch orders of these steps!

Title:

1 **Problem Statement** What problem are you trying to solve? What larger issues do the problem address?	2 **Outcomes/Predictions** What prediction(s) are you trying to make? Identify applicable predictor (X) and/or target (y) variables.	3 **Data Acquisition** Where are you sourcing your data from? Is there enough data? Can you work with it?
4 **Modeling** What models are appropriate to use given your outcomes?	5 **Model Evaluation** How can you evaluate your model's performance?	6 **Data Preparation** What do you need to do to your data in order to run your model and achieve your outcomes?

Conceptualized by Jasmine Vasandani using notes from General Assembly's Data Science Immersive. Format inspired by Business Model Canvas.

Activation

When you finish filling out the canvas above, now you can begin implementing your data science workflow in roughly this order.

1 Problem Statement → 2 Data Acquisition → 3 Data Prep → 4 Modeling → 5 Outcomes/Preds → 6 Model Eval

* **Note:** This canvas is intended to be used as a starting point for your data science projects. Data science workflows are typically nonlinear.

7. Links

Aqui você deve disponibilizar os links para o vídeo com sua apresentação de 5 minutos e para o repositório contendo os dados utilizados no projeto, scripts criados, etc.

Link para o vídeo: youtube.com/...

Link para o repositório: github.com/...

REFERÊNCIAS

1. https://www.sebrae.com.br/sites/PortalSebrae/ufs/mt/noticias/micro-e-pequenas-empresas-geram-27-do-pib-do-brasil,ad0fc70646467410VgnVCM2000003c74010aRCRD
2. https://covid19.ibge.gov.br/pulso-empresa/
3. https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/cadastros/consultas/dados-publicos-cnpj
4. https://www.tesourotransparente.gov.br/ckan/dataset/lista-de-municipios-do-siafi
5. https://covid.saude.gov.br/
6. http://www.atlasbrasil.org.br/ranking
7. https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html?t=downloads&c=1100023
8. https://cnae.ibge.gov.br/?option=com_cnae&view=estrutura&Itemid=6160&chave=&tipo=cnae&versao_classe=7.0.0&versao_subclasse=9.1.0

APÊNDICE

O apêndice contendo os códigos utilizados foi disponibilizado à Puc Minas no Repositório do Github.