**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS  
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**William Sanches Lima**

**obtenção de insights dos dados do portal da transparência através de Análise exploratória e representação visual dos dados**

Belo Horizonte

2020

**William Sanches Lima**

**OBTENÇÃO DE INSIGHTS DOS DADOS DO PORTAL DA TRANSPARÊNCIA ATRAVÉS DE ANÁLISE EXPLORATÓRIA E REPRESENTAÇÃO VISUAL DOS DADOS**

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

2020 **SUMÁRIO**

[1. Introdução 4](#_Toc445198572)

[1.1. Contextualização 4](#_Toc445198573)

[1.1. O problema proposto 4](#_Toc445198574)

[2. Coleta de Dados 4](#_Toc445198576)

[3. Processamento/Tratamento de Dados 5](#_Toc445198582)

[4. Análise e Exploração dos Dados 5](#_Toc445198585)

[5. Criação de Modelos de Machine Learning 5](#_Toc445198586)

[6. Apresentação dos Resultados 5](#_Toc445198587)

[7. Links 6](#_Toc445198588)

[REFERÊNCIAS 7](#_Toc445198589)

# 1. Introdução

## 1.1. Contextualização

A Controladoria-Geral da União (CGU) acredita que a transparência é o melhor antídoto contra a corrupção, uma vez que ela é um importante mecanismo indutor para que os gestores públicos ajam com responsabilidade. Uma gestão pública transparente permite à sociedade, com informações, colaborar no controle das ações de seus governantes, com intuito de checar se os recursos públicos estão sendo usados como deveriam. Nesse contexto a CGU criou o programa Brasil Transparente para auxiliar Estados e Municípios na implementação das medidas de governo transparente previstas na Lei de Acesso à Informação (LAI). O objetivo é juntar esforços no incremento da transparência pública e na adoção de medidas de governo aberto. A cobrança por mais transparência tem crescido no Brasil em ritmo acelerado nos últimos anos. A promoção da transparência e do acesso à informação é considerada medida indispensável para o fortalecimento da democracia e para a melhoria da gestão pública. Nesse sentido foi lançado pelo Ministério da Transparência e Controladoria-Geral da União em 2004, o Portal da Transparência do Governo Federal é um site de acesso livre, no qual o cidadão pode encontrar informações sobre como o dinheiro público é utilizado, além de se informar sobre assuntos relacionados à gestão pública do Brasil. Com base nesses dados foi possível utilizar técnicas de Análise Exploratória de Dados (Exploratory Data Analysis – EDA) e Visualização de Dados para obtenção de “insights” (também conhecido como “data insight”).

## 1.2. O problema proposto

O problema proposto foi analisar os dados disponíveis no Portal da Transparência da Controladoria Geral da União para obtenção de “insights”, aplicando técnicas de Análise Exploratória de Dados e Visualização de Dados, com o objetivo de comprovar a eficácia dessas técnicas no apoio ao processo de seleção e contratação de fornecedores no âmbito da administração pública.

(Why?) Por que esse problema é importante? Desde 2014 o Brasil vem caindo no ranking do Índice de Percepção da Corrupção (IPC), elaborado pela organização Transparência Internacional que avalia a percepção da corrupção no setor público em 180 países. Em 2016, o Brasil ficou em 79º; em 2017 o país estava na 96ª colocação; em 2019 ocupou o 105º e agora em 2020 está em 106º lugar. Dessa forma entende-se que o problema em questão é de extrema relevância pois o “custo” da corrupção tem impacto direto na diminuição dos investimentos na saúde, na educação, em infraestrutura, segurança, habitação, entre outros direitos essenciais à vida.

(Who?) De quem são os dados analisados? De um governo? Um ministério ou secretaria? Dados de clientes? Os dados são do governo, do portal da transparência, do Ministério da Transparência e Controladoria-Geral da União, fruto do programa Brasil Transparente para auxiliar Estados e Municípios na implementação das medidas de governo transparente previstas na Lei de Acesso à Informação (LAI).

(What?): Quais os objetivos com essa análise? O que iremos analisar? O objetivo recuperar os dados do Portal da Transparência para realização de Análise Exploratória dos Dados e Visualização dos Dados afim de obter “insights”, para comprovar a eficácia da aplicação dessas técnicas no apoio ao processo de seleção e contratação de fornecedores no âmbito da administração pública. Os dados analisados serão do Cadastro de Empresas Inidôneas e Suspensas (CEIS), do Portal da Transparência da Controladoria Geral da União.

(Where?): Trata dos aspectos geográficos e logísticos de sua análise. Os dados tem abrangência nacional, mas possuem detalhes relacionados aos estados e munícipios, possibilitando a qualificação dos dados com essas informações.

# 2. Coleta de Dados

Os dados foram obtidos do portal da transparência, do Ministério da Transparência e Controladoria-Geral da União. Os dados divulgados no Portal são provenientes de diversas fontes de informação, entre as quais estão os grandes sistemas estruturadores do Governo Federal – como o Sistema Integrado de Administração Financeira do Governo Federal (Siafi) e o Sistema Integrado de Administração de Recursos Humanos (Siape) –, as bases de benefícios sociais, as faturas de Cartão de Pagamentos do Governo Federal, as bases de imóveis funcionais, entre diversas outras.

Os órgãos responsáveis por cada fonte de informação encaminham seus dados para a CGU, que recebe, reúne e disponibiliza as informações na ferramenta. A periodicidade de envio dos dados depende do assunto tratado, assim como a periodicidade de atualização das informações no Portal.

Uma vez carregadas no Portal, as informações são disponibilizadas para conhecimento do cidadão de diversas formas, como: painéis, consultas detalhadas, gráficos, dados abertos.

O acesso ao Portal não requer usuário nem senhas, sendo permitido a qualquer cidadão navegar pelas páginas de forma livre, bem como visualizar e utilizar os dados disponíveis da forma que melhor lhe convier.

O Portal da Transparência está atento aos princípios de Governo Eletrônico e sabe que os dados devem ser disponibilizados de formas diferentes a fim de atender aos diversos perfis de usuários. Para isso, além de consultas online e com visualizações que buscam transmitir, de forma simples, como o governo usa os recursos públicos, formas de acesso aos dados para desenvolvedores e engajados com a tecnologia da informação também estão disponíveis.

O acesso para desenvolvedores e engajados ocorre através de uma Interface de Programa de Aplicativos (do inglês, “Application Programming Interface”), ou simplesmente “API”. Com ela, é possível ter um serviço de consulta direta aos dados do Portal da Transparência sem precisar navegar pelo site ou utilizar robôs para a obtenção das informações de forma automática. Os dados disponíveis são os mesmos apresentados em tela, com a flexibilidade característica das APIs.

No Portal da Transparência, o serviço de consulta via API foi implementado em REST (Representational State Transfer). Entendemos que o REST é mais simples de ser utilizado, além de poder ser testado aqui mesmo no Portal. Além do mais, o REST costuma ser a escolha preferida dos desenvolvedores.

As consultas via API têm restrição quanto ao número de requisições por minuto, de modo a não sobrecarregar o Portal da Transparência e impactar no tempo de resposta das consultas realizadas por meio dos navegadores. No período de 6:00 às 23:59, o Portal aceita 90 requisições por minuto. Já no período das 00:00 às 5:59, são aceitas 300 requisições por minuto.

O acesso via API fornece toda uma flexibilidade para consultas pontuais. Para acesso ao conjunto completo de dados, no entanto, sugerimos utilizar as planilhas de dados abertos. O Portal da Transparência agrega uma enorme quantidade de dados e baixar as planilhas para uso local certamente fornecerá melhores resultados para grandes volumes de dados. O download de planilhas pode ser feito em seção específica.

A documentação do uso das APIs pode ser acessada em <http://www.transparencia.gov.br/swagger-ui.html>. Estão disponíveis as seguintes consultas: Bolsa Família, Programa de Erradicação do Trabalho Infantil (Peti), Garantia-Safra, Seguro Defeso, Cadastro de Expulsões da Administração Federal (CEAF), Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS), Cadastro Nacional de Empresas Punidas (CNEP), Contratos do Poder Executivo Federal, Convênios do Poder Executivo Federal, Despesas Públicas, Entidades Privadas sem Fins Lucrativos Impedidas (CEPIM), Licitações do Poder Executivo Federal, Servidores do Poder Executivo Federal e Viagens a Serviço.

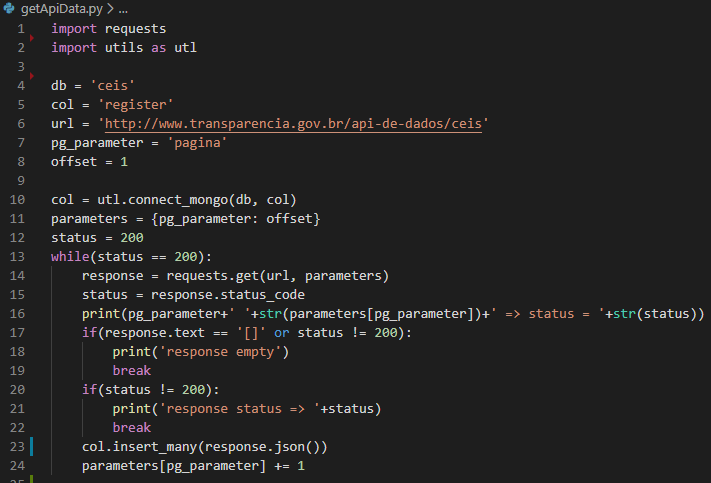
Para o desenvolvimento desse trabalho foram utilizadas as API’s do Cadastro Nacional de Empresas Inidôneas e Suspensa (CEIS), que apresenta um dataset maior e com mais atributos para as análises.

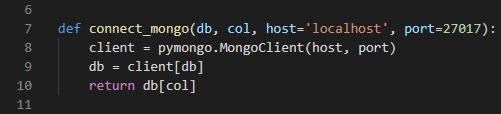
Os dados foram obtidos através da API de Consulta dos registros do CNEP por CNPJ ou CPF Sancionado/Órgão Sancionador/Período. Abaixo a documentação da API:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GET | /api-de-dados/ceis | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Notas de Implementação | | |  |  |  |  |  |  |
| Filtros mínimos: Página (padrão = 1); | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Classe de resposta (Status 200) | | | |  |  |  |  |  |
| OK |  |  |  |  |  |  |  |  |
| Example Value: | |  |  |  |  |  |  |  |
| [  {  "abrangenciaDefinidaDecisaoJudicial": "string",  "dataFimSancao": "string",  "dataInicioSancao": "string",  "dataOrigemInformacao": "string",  "dataPublicacaoSancao": "string",  "dataReferencia": "string",  "dataTransitadoJulgado": "string",  "detalhamentoPublicacao": "string",  "fonteSancao": {  "enderecoContato": "string",  "nomeExibicao": "string",  "telefoneContato": "string"  },  "id": 0,  "informacoesAdicionaisDoOrgaoSancionador": "string",  "legislacao": {  "descricaoFundamentacaoLegal": "string",  "fundamentacaoLegal": "string"  },  "linkPublicacao": "string",  "numeroProcesso": "string",  "orgaoSancionador": {  "nome": "string",  "poder": "string",  "siglaUf": "string"  },  "pessoa": {  "cnae": {  "classe": "string",  "codigoClasse": "string",  "codigoDivisao": "string",  "codigoGrupo": "string",  "codigoSecao": "string",  "codigoSubclasse": "string",  "divisao": "string",  "grupo": "string",  "secao": "string",  "subclasse": "string"  },  "codigoFormatado": "string",  "complementoEndereco": "string",  "dataAbertura": "string",  "descricaoLogradouro": "string",  "enderecoEletronico": "string",  "localidadePessoa": "string",  "municipio": {  "codigoIBGE": "string",  "nomeIBGE": "string",  "pais": "string",  "uf": {  "nome": "string",  "sigla": "string"  }  },  "naturezaJuridica": {  "codigo": "string",  "codigoTipo": "string",  "descricao": "string",  "descricaoTipo": "string"  },  "nome": "string",  "nomeBairro": "string",  "nomeFantasiaReceita": "string",  "numeroCEP": "string",  "numeroEndereco": "string",  "numeroInscricaoSocial": "string",  "numeroTelefone": "string",  "razaoSocialReceita": "string",  "tipoCodigo": "string",  "tipoPessoa": "string"  },  "sancionado": {  "codigoFormatado": "string",  "nome": "string"  },  "textoPublicacao": "string",  "tipoSancao": {  "descricaoPortal": "string",  "descricaoResumida": "string"  }  } ] | | | | | | | | |
|
|
|  |  |  |  |  |  |  |  |  |
| Parâmetros | |  |  |  |  |  |  |  |
| Parâmetro | | Descrição | | | | Tipo de parâmetro | | Tipo de dados |
|  | |  | | | |  | |  |
| cnpjSancionado | | CNPJ ou CPF do Sancionado | | | | query | | string |
| nomeSancionado | | Nome, nome fantasia ou razão social do Sancionado | | | | query | | string |
| orgaoSancionador | | Órgão Sancionador | | | | query | | string |
| dataInicialSancao | | Data Inicial da Sanção (DD/MM/AAAA) | | | | query | | string |
| dataFinalSancao | | Data Final da Sanção (DD/MM/AAAA) | | | | query | | string |
| **pagina (required)** | | **Página consultada** | | | | query | | integer |
| Mensagens de resposta | | |  |  |  |  |  |  |
| Código de status HTTP | | Razão | |  |  |  |  |  |
|  | |  | |  |  |  |  |  |
| 401 | | Unauthorized | |  |  |  |  |  |
| 403 | | Forbidden | |  |  |  |  |  |
| 404 | | Not Found | |  |  |  |  |  |

Tabela 1 - Documentação API CEIS

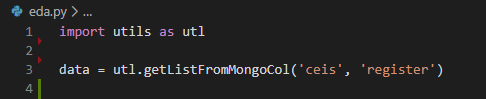
Para obtenção dos dados foi utilizada função “get” da biblioteca “requests” (<https://requests.readthedocs.io/en/master/>) - “requests.get(url, parameters)” - passando a URL da API no primeiro parâmetro e o número da página no segundo parâmetro. Para que não fosse necessário calcular o número total de páginas que seriam retornadas pela API foi implementado um laço que executa a função get e incrementa o número da página enquanto o status HTTP de retorno for igual 200 (<https://developer.mozilla.org/pt-BR/docs/Web/HTTP/Status/200>). A cada iteração os dados eram persistidos no Banco MongoDB e, para tanto, foi utilizada a biblioteca pymongo. Através da classe MongoClient, da biblioteca pymongo, obtem-se o objeto de conexão com o MongoDB local passando como parâmetro a coleção que será manipulada. Para essas regras foi implementada a classe “getApiData.py”. Essa classe importa, além da biblioteca “requests”, já mencionada anteriormente, a classe utils.py, importada com pseudônimo “utl”, onde está implementada a função de conexão com o MongoDB, chamada “connect\_mongo”.

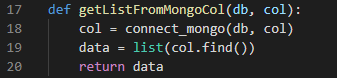




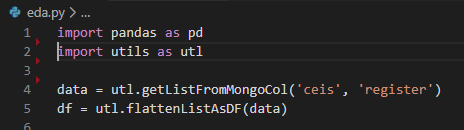
# 3. Processamento/Tratamento de Dados

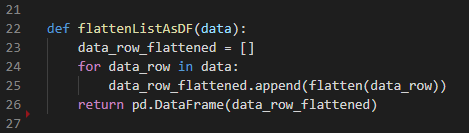
Após a persistência de todo conteúdo obtido através da API do Portal da Transparência no MongoDB, recuperou-se o dataset para o início da fase de validação e tratamento dos dados. Para recuperar o dataset foi criada a função “getListFromMongoCol”, implementada na classe utils.py.



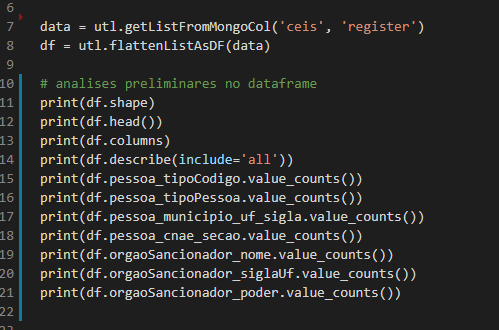


O primeiro ponto tratado é referente à estrutura dos dados. Conforme pode ser observado na Tabela 1, os dados possuem uma estrutura aninhada por categorias (ex: pessoa.municipio.uf.sigla). Como os dados foram persistidos “in natura” o MongoDB preservou essa estrutura. Na recuperação do dataset e posterior amazenamento em um “pandas DataFrame”, notou-se a dificuldade em obtenção de subsets e consultas agrupadas, porque os dados aninhados eram do tipo dict (dicionário), o que impossibilitava utilização de funções do “pandas DataFrame” e, para tanto, seriam necessárias diversas implementações em estruturas de repetição para efetuar as análises. Para contornar esse problema o dataset foi “achatado” e todo conteúdo convertido em “pandas DataFrame”. Para isso foi utilizada a biblioteca “flatten\_json” (<https://pypi.org/project/flatten-json/>). Foi criada uma função para executar essa instrução, chamada “flattenListAsDF” (implementada na classe utils.py).





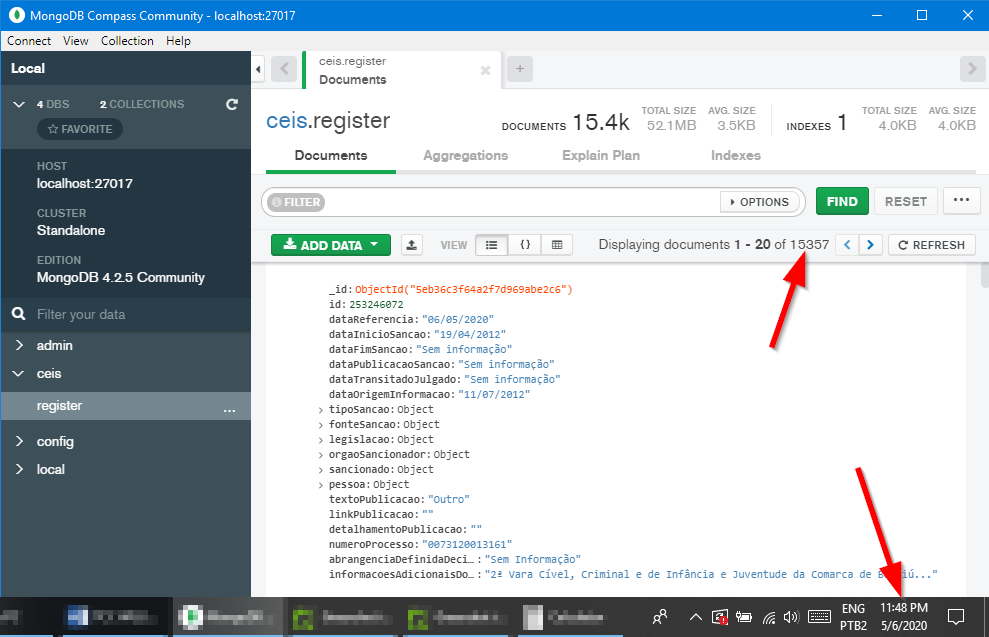
Após o armazenamento dos dados foram executadas análises preliminares no dataframe.

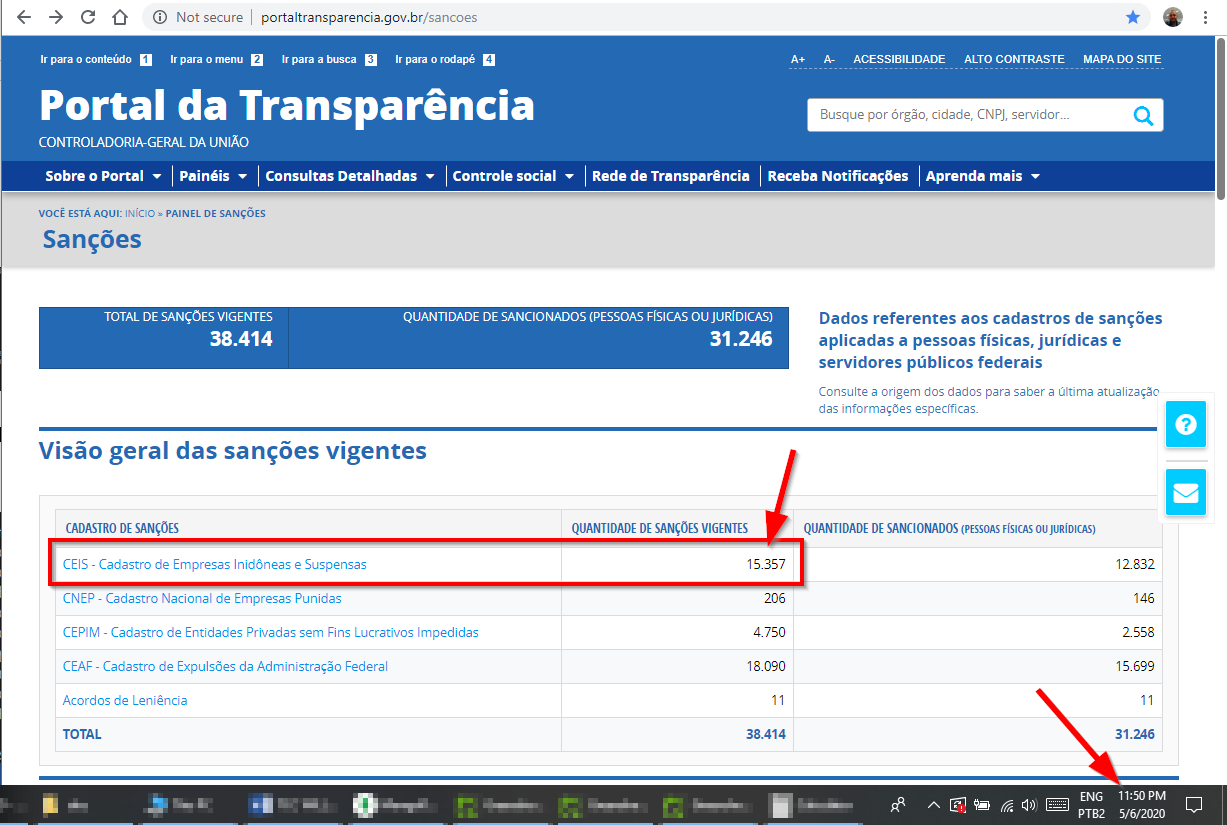


* Propriedade “df.shape”

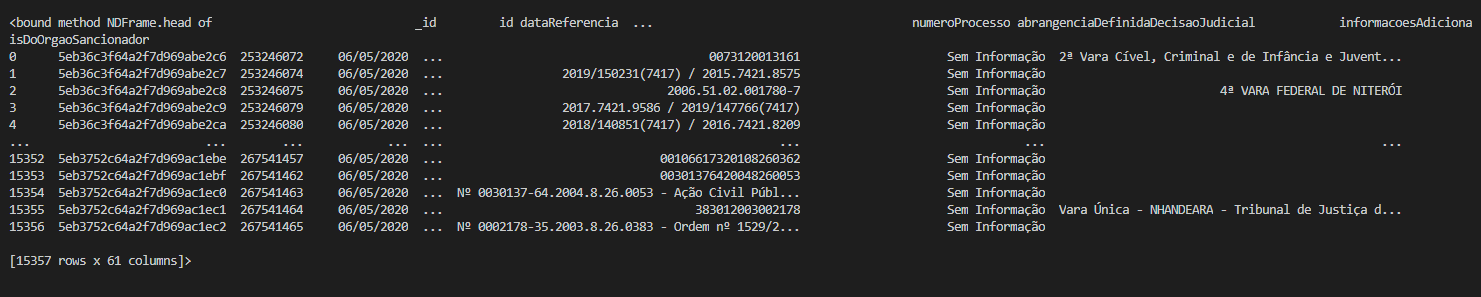


Informa que o frame tem 15.357 linhas e 61 colunas. A quantidade de linhas corrobora com o quantidade total de linhas informada pelo pelo MongoDB e também com o quantidade total de sanções informadas no Portal da Transparência na data em que as análises foram publicadas (06/05/2020). Isso demonstra que os dados analisados possuem um certo grau de confiabilidade em relação à sua origem.



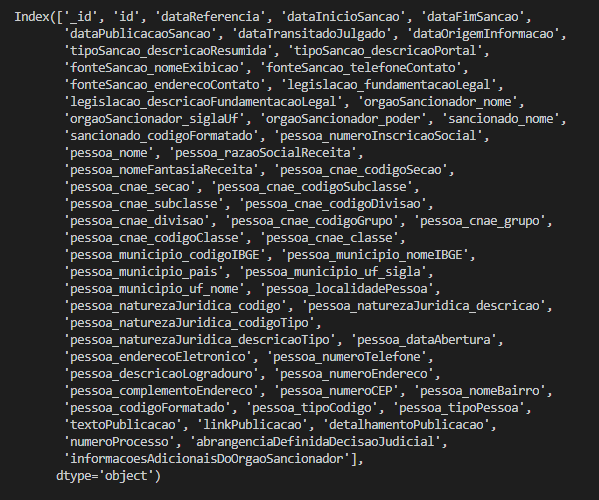


* Função “df.head()”



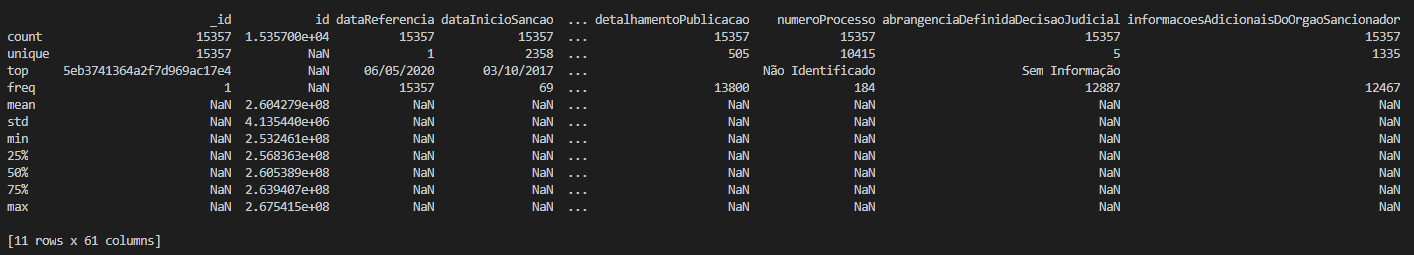
Exibe os 5 primeiros e 5 últimos itens do dataframe.

* Propriedade “df.columns”



Exibe o nome das colunas do dataframe.

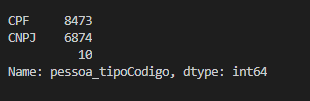
* Função “df.describe(include=’all’)”



Exibe alguns dados estatísticos sobre o dataframe.

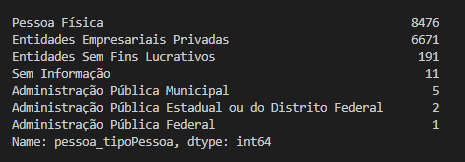
Com o intuito de buscar insights dos dados algumas séries (“pandas.Series”) foram eleitas para serem alvo da EDA (Exploratory Data Analysis) A série “pessoa\_tipoCodigo” poderia informar qual a quantidade total de sanções entre pessoas físicas e jurídicas. A série “pessoa\_tipoPessoa” poderia informar qual a categorização das entidades que mais incorrem nas sanções. A série “pessoa\_municipio\_uf\_sigla” poderia informar em quais estados estão os maiores números de entidades sancionadas. A série “pessoa\_cnae\_secao” poderia informar quais os tipos de atividade mais comuns das pessoas jurídicas sancionadas. A série “orgaoSancionador\_nome” poderia informar quais os órgãos aplicam mais sanções. A série “orgaoSancionador\_siglaUf” poderia informar em quais estados os órgãos aplicam mais sanções. A série “orgaoSancionador\_poder” poderia informar quais são os poderes que mais aplicam sanções. Outras séries poderiam ser incluídas no roteiro das análises conforme os insights obtidos.

* Função “df.pessoa\_tipoCodigo.value\_counts()”



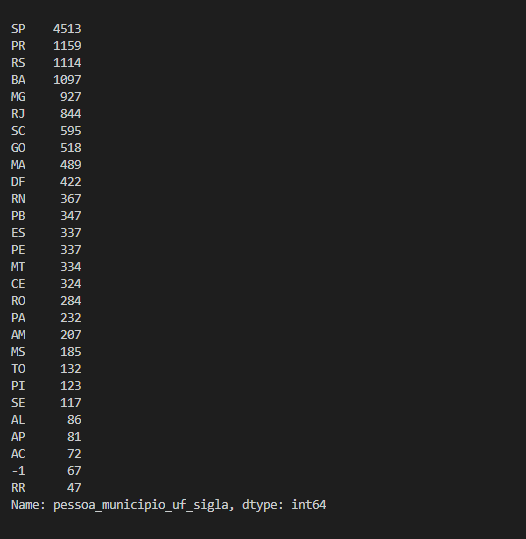
Contagem de valores da série “pessoa\_tipoPessoa”. Nota-se que a distribuição é balanceada entre pessoas físicas e jurídicas. Nota-se também um volume muito pequeno de “¿missing value?” nessa série, o que poderia, eventualmente, ser tratado mais adiante, caso se mostre necessário. Por ora não foi tratado por representar um valor irrelevante (cerca de 0,06% do total).

* Função “df.pessoa\_tipoPessoa.value\_counts()”



Contagem de valores da série “pessoa\_tipoPessoa”. Aqui observou-se que a quantidade de pessoas físicas permanece inalterada, porque não existe uma categorização para esse tipo, mas obteve-se um maior nível de detalhe na categorização de pessoas jurídicas. De qualquer forma foi possível verificar que, predominantemente, as sanções foram aplicadas à pessoas jurídicas categorizadas como Entidades Empresariais Privadas e que, apenas essa categoria, representa cerca de 43% do total do universo de entidades sancionadas.

* Função “df.pessoa\_municipio\_uf\_sigla.value\_counts()”



Efetua a contagem de valores da série “pessoa\_municipio\_uf\_sigla”.

* Função

Detalhes sobre a função/propriedade.

* Função

Detalhes sobre a função/propriedade.

* Função

Detalhes sobre a função/propriedade.

* Função

Detalhes sobre a função/propriedade.

* Função

Detalhes sobre a função/propriedade.

* Função

Detalhes sobre a função/propriedade.

* Função

Detalhes sobre a função/propriedade.

* Função

Detalhes sobre a função/propriedade.

* Função

Detalhes sobre a função/propriedade.

* Função

Detalhes sobre a função/propriedade.

* Normalizar o campo Estado para que o Power BI consiga entender
* Tratar campos nulos (quais?)

# 4. Análise e Exploração dos Dados

Nessa seção você deve mostrar como foi realizada a análise e exploração dos seus. Mostre as hipóteses levantadas durante essa etapa e os padrões e *insights* identificados.

# 5. Criação de Modelos de Machine Learning

Conforme o documento de instruções para o TCC, essa etapa não é obrigatória, mas é fortemente recomendada. Caso você crie modelos de *Machine Learn*ing em seu projeto, nessa seção você irá descrever as ferramentas utilizadas. Se você utilizou ferramentas visuais como Knime e Rapid Miner, coloque aqui um print do seu modelo. Caso você tenha escrito scripts em Python, por exemplo, coloque aqui o seu script. Explique as *features* utilizadas, justifique a escolha por determinado modelo, os parâmetros utilizados, etc.

# 6. Apresentação dos Resultados

Nessa seção você deve apresentar os resultados obtidos. Apresente gráficos, dahsboards, conte a sua história de forma bastante criativa. Aqui você pode utilizar os modelos de Canvas propostos por Dourard (clique [aqui](https://www.louisdorard.com/machine-learning-canvas)) ou por Vasandani (clique [aqui](https://towardsdatascience.com/a-data-science-workflow-canvas-to-kickstart-your-projects-db62556be4d0)).



# 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.

# REFERÊNCIAS

Um projeto de Ciência de Dados não requer revisão bibliográfica. Portanto, a inclusão das referências não é obrigatória. No entanto, caso você deseje incluir referências relacionadas às tecnologias ou às metodologias usadas em seu trabalho, relacione-as de acordo com o modelo a seguir.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.

SOBRENOME DO AUTOR, Nome do autor. **Título do livro ou artigo.** Cidade: Editora, ano.