CENTRO UNIVERSITÁRIO CARIOCA

FACULDADE CIÊNCIA DA

COMPUTAÇÃO

GUILHERME NUNES FARIAS

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS PARA ANALISAR LICITAÇÕES E CONTRATOS DO GOVERNO FEDERAL

Rio de Janeiro

2020

GUILHERME NUNES FARIAS

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS PARA ANALISAR LICITAÇÕES E CONTRATOS DO GOVERNO FEDERAL

Trabalho de Conclusão de Curso apresentado ao Centro Universitário Carioca, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. D.Sc Sérgio Assunção Monteiro

RIO DE JANEIRO

2020

GUILHERME NUNES FARIAS

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS PARA ANALISAR LICITAÇÕES E CONTRATOS DO GOVERNO FEDERAL

Trabalho de Conclusão de Curso apresentado ao Centro Universitário Carioca, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

BANCA EXAMINADORA

Prof. Sérgio Monteiro Assunção, D.Sc - Orientador
Centro Universitário Carioca

Prof. André Luiz Avelino Sobral, M.Sc
Centro Universitário Carioca

Centro Universitário Carioca

AGRADECIMENTOS

Agradeço em primeiro lugar a Deus, por me dar forças e condições para chegar. Agradeço aos meus pais que nunca desistiram e sempre acreditam em mim e a toda minha família e amigos por ter me apoiado. Agradeço ao meu orientador, DSc. Sérgio Assunção Monteiro, por confiar em mim e acreditar no meu potencial.

RESUMO

O governo federal do Brasil possui um portal de transparência chamado de Dados Abertos. Nesse portal, é possível fazer ter acesso aos dos mais diversos assuntos no âmbito público. O objetivo desse trabalho, é analisar os dados relacionados a licitações e contratos do governo federal do período de janeiro de 2020 a agosto de 2020. O trabalho vai se concentrar em aspectos técnicos, do tipo origem, tratamento e visualização dos dados através de uma aplicação desenvolvida em python.

Palavras chaves: Análise de dados, licitações, dataFrame

ABSTRACT

The federal government of Brazil has a transparency portal called Open Data. In this portal, it is possible to have access to the most diverse subjects in the public sphere. The objective of this work is to analyze the data related to bids and contracts from the federal government from January 2020 to August 2020. The work will focus on technical aspects, such as origin, treatment and visualization of data through an application developed in python.

Keywords: data analysis, bidding, dataFrame

LISTA DE ILUSTRAÇÕES

Figura 01 – Categoria de Licitações.....15
Figura 02 – Cálculo da média.....23
Figura 03 – Cálculo da mediana número ímpar.....23
Figura 04 – Cálculo da mediana número par.....23
Figura 05 – Cálculo do desvio médio.....24
Figura 06 – Cálculo da variância.....24
Figura 07 – Gráfico de linha.....25
Figura 08 – Gráfico de barra.....25
Figura 09 – Gráfico de setor.....26
Figura 10 – Gráfico de histograma.....26
Figura 11 – Palavras reservadas Python.....30
Figura 12 – Importando bibliotecas.....34
Figura 13 – Importando arquivos.....34
Figura 14 – Verificando tipos.....35
Figura 15 – Formatando valor.....36
Figura 16 – Removendo linhas e colunas.....36
Figura 17 – Removendo linhas e colunas demais tabelas.....36
Figura 18 – Verificando dados ausentes.....36
Figura 19 – Contagem dados ausentes.....37
Figura 20 – Tratando valores ausentes.....37
Figura 21 – Visualizando linhas duplicadas.....38
Figura 22 – Visualizando linhas duplicadas demais tabelas.....38
Figura 23 – Removendo duplicatas.....38
Figura 24 – Removendo espaços em brancos.....38
Figura 25 – Valor por mês.....39
Figura 26 – Tabela valor por mês.....39
Figura 27 – Total de licitações por mês.....40
Figura 28 – Código total de licitações por mês.....40
Figura 29 – Evolução das licitações.....41
Figura 30 – Evolução das modalidades.....41
Figura 31 – Modalidade x Valor.....42
Figura 32 – Código gráfico dispersão.....42
Figura 33 – Tabela valor por modalidade.....43
Figura 34 – Valor por modalidade.....44
Figura 35 – Maiores participantes no ano.....46

Figura 36 – Participantes com mais vitórias no ano.....46

LISTA DE ABREVIATURA E SIGLAS

EPP	Empresa de pequeno porte
ME	Microempresa
TCU	Tribunal de Contas da União
JVM	Java Virtual Machine
PVM	Python Virtual Machine

SUMÁRIO

1 INTRODUÇÃO.....11
1.1 Objetivos do Trabalho.....11
2 Licitações e Análise Estatística.....12
2.1 Licitações.....12
2.2 Princípios da licitação.....13
2.3 Leis que regem as licitações.....14
2.4 Modalidades de licitações.....15
2.5 Dados Abertos.....16
2.6 Análises Estatísticas.....17
2.61 Tipos de amostragem.....18
2.6.1.1 Técnicas de amostragem probabilística.....19
2.6.1.2 Técnicas de amostragem não probabilística.....20
2.6.2 Variáveis.....21
2.6.3 Distribuição de frequências.....21
2.6.4 Medidas de resumo.....22
2.6.5 Tipos de Gŕaficos.....24
2.7 Introdução ao Python.....27
2.7.1 Características.....27
2.7.2 Interpretador.....28
2.7.3 Estrutura.....29
3 Python.....30
3.1 Principais bibliotecas python.....31
3.1 Introdução ao Jupyter-Notebook.....32
3.2 Vantagens de trabalhar com notebook.....32
3.3 Instalação.....33
4 Estudo de Caso.....34
4.1 Leitura e tratamento dos dados.....34
4.2 Análise e Visualização dos dados.....38
5 Conclusão.....47
REFERÊNCIAS BIBLIOGRÁFICAS.....48

1 INTRODUÇÃO

O governo brasileiro tem sempre a necessidade da realização de obras, manutenção, compras de materiais, etc, e para isso são contratadas empresas para realizar tais serviços. A licitação é um processo administrativo que visa assegurar a contratação dessas empresas, permitindo que diversas empresas possam concorrer igualmente, atendendo aos requisitos, sem que haja fraude na contratação de uma determinada empresa.

Todos os meses o governo libera dados referentes a licitação, permitindo que possamos coletar esses dados e analisá los. E para tal tarefa utilizaremos a linguagem python, que vem crescendo cada vez mais em mineração de dados, em conjuntos com suas principais bibliotecas, como Pandas e NumPy, para todo o processo de análise de dados.

1.1 Objetivos do Trabalho

O objetivo central deste trabalho é utilizar a ferramenta jupyter-notebook, junto com as principais bibliotecas de python para análise de dados e realizar todo o trabalho de mineração de dados, realizado a coleta, tratamento, análise e visualização dos dados, para apresentar os dados referentes às licitações do período de janeiro a agosto de 2020 retiradas do site do governo da receita federal. Com Python em conjunto com suas ferramentas e bibliotecas, apresentaremos o porquê da linguagem ser utilizada e ganhar cada vez mais espaço para tratamento, mineração e análise de dados.

É importante esclarecer que este trabalho tem como objetivo apenas fazer a tratativa e apresentação dos dados sem tirar conclusões.

1.2 Organização do Trabalho

O trabalho está dividido em cinco capítulos, este primeiro capítulo introdutório descreve os objetivos e a organização do trabalho.

O segundo capítulo abordaremos os conceitos básicos sobre licitações e contratos. Falaremos também sobre a utilização dos dados abertos e uma introdução a análise estatística e, por fim, falaremos sobre a linguagem Python bem como algumas de suas características.

O terceiro capítulo continuará a falar sobre python, sua instalação, principais bibliotecas bem como a ferramenta que será utilizada nesse projeto que é o jupyter-notebook.

O quarto capítulo é um estudo de caso onde o python e suas bibliotecas são aplicadas para tratamento de dados, leitura de dados e para a análise de dados.

O quinto capítulo apresenta as conclusões que pudemos retirar dos dados e futuros trabalhos que podem continuar sendo feito.

2 Licitações e Análise Estatística

2.1 Licitações

Licitação é o processo administrativo utilizado pela Administração Pública e pelas demais pessoas indicadas pela lei com o objetivo de selecionar a melhor proposta, por meio de critérios objetivos e impessoais, para celebração de contratos. O art. 3.º da Lei 8.666/1993 elenca os objetivos da licitação, quais sejam: a) garantir a observância do princípio constitucional da isonomia, b) selecionar a proposta mais vantajosa para a Administração e c) promover o desenvolvimento nacional sustentável. (CARVALHO, 2015).

Art. 3º A licitação destina-se a garantir a observância do princípio constitucional da isonomia, a seleção da proposta mais vantajosa para a administração e a promoção do desenvolvimento nacional sustentável e será processada e julgada em estrita conformidade com os princípios básicos da legalidade, da impessoalidade, da moralidade, da igualdade, da publicidade, da probidade administrativa, da vinculação ao instrumento convocatório, do julgamento objetivo e dos que lhes são correlatos (Lei Nº 8666).

É um processo administrativo que visa assegurar a contratação de serviços ao governo e que antecede a assinatura de contrato entre a empresa contratante e a administração pública.

O governo deve comprar e contratar serviços seguindo as leis e de forma transparente, dessa forma, licitação é um processo formal onde essas contratações acontecem e os interessados devem competir para ganhar a licitação e dessa forma fornecer os serviços solicitados ao governo. Há uma necessidade da administração pública, como: serviços, obras a serem realizadas, entre outras, e com isso se inicia o planejamento, entrando na fase interna onde serão discutidas as regras de contratação, o que deve ser feito e o que comprar. Com tudo definido a licitação vai pra fase externa, com a licitação sendo publicada, onde as empresas que quiseram

oferecer o serviço deverão oferecer suas propostas. E termina com o contrato, onde a contratada deverá executar o serviço que foi contratado e à administração fiscalizar a execução.

Cada licitação tem seu respectivo edital, e pode conter diversos processos, com diversos itens ou serviços referentes a diversos órgãos superiores onde estarão todas as regras, sendo a Lei interna da licitação. No edital tudo deve estar de acordo com as leis, não pode conter cláusulas ou condições que comprometam a competição. Também não podendo conter falta de informação, impreciso, genérico ou qualquer outra coisa que permita que a licitação seja burlada ou facilite para uma determinada empresa.

2.2 Princípios da licitação

A licitação, por ser um processo administrativo, pressupõe o atendimento dos princípios constitucionais aplicáveis à Administração Pública, notadamente aqueles expressamente previstos no art. 37, caput, da CRFB (legalidade, impessoalidade, moralidade, publicidade e eficiência) (CARVALHO, 2015).

Além do princípio constitucional, as licitações possuem outros principios que devem ser respeitados, sendo eles:

1. **Princípio da competitividade:** Busca pela melhor proposta para Administração, buscando também maior quantidade de propostas.
2. **Princípio da isonomia:** A administração deve permitir a participação de qualquer concorrente(que atende as especificações do edital da licitação), isto é, não pode haver discriminação. Um princípio que se relaciona com a da competitividade pois, não havendo restrições a concorrentes o número de propostas seria maior.
3. **Princípio da vinculação ao instrumento convocatório:** As regras do edital da licitação deve ser seguida rigorosamente, pois o não cumprimento poderá acarretar na ilegabilidade da licitação. Lei esta que deve ser respeitada pelos licitantes e Poder Público.
4. **Princípio do procedimento formal:** É a fidelidade as normas contidas na licitação. Exemplos: quando todos os licitantes forem inabilitados ou todas as propostas forem desclassificadas, a Administração poderá fixar prazo para que os licitantes apresentem nova documentação ou outras propostas (CARVALHO, 2015).
5. **Princípio do julgamento objetivo:** As propostas devem ser avaliadas segundo os objetivos listados na licatação.

2.3 Leis que regem as licitações

A lei Federal é a lei 8666, sendo esta uma lei nacional que deve ser observada pela União, Estados e Municípios.

Com o surgimento do Pregão em 2002, surgiu a Lei 10.520 que rege os pregões, mas às vezes é necessário se recorrer a Lei 8666 quando a Lei do Pregão não conseguir responder.

Ambas as leis permitem que os governos façam seus Regulamentos próprios, isso facilita para os governos adequarem às regras gerais as particularidades de cada administração. Também existe a Lei complementar 123 que traz orientações para licitações quando as empresas forem EPP ou ME. Art. 44.

Nas licitações será assegurada, como critério de desempate, preferência de contratação para as microempresas e empresas de pequeno porte (Lei complementar 123, 2006).

E em 2016 foi criada a Lei das Estatais, que é a Lei 13.303/2016. Onde as regras são direcionadas para empresas públicas de economia mista (empresas que o Estado tem controle acionário e que compõem a Administração indireta) e de suas subsidiárias.

Art. 31. As licitações realizadas e os contratos celebrados por empresas públicas e sociedades de economia mista destinam-se a assegurar a seleção da proposta mais vantajosa, inclusive no que se refere ao ciclo de vida do objeto, e a evitar operações em que se caracterize sobrepreço ou superfaturamento, devendo observar os princípios da impessoalidade, da moralidade, da igualdade, da publicidade, da eficiência, da probidade administrativa, da economicidade, do desenvolvimento nacional sustentável, da vinculação ao instrumento convocatório, da obtenção de competitividade e do julgamento objetivo (Lei 13.303, 2016).

Leis, Decretos, Instruções Normativas e Regulamentos devem ser lidos e compreendidos com profundidade para todos que desejam participar da competição e que atendam as leis e seus princípios da constituição. Tornando a licitação justa. E para o governo, pois ele licitará corretamente cumprindo seu dever de aplicar a legislação e pela supremacia dos interesses públicos.

2.4 Modalidades de licitações

Modalidade de licitação é a forma como o processo de compra de produtos e serviços de órgãos públicos é conduzido (ZUCCO, 2018).

São 5 modalidades de licitações, dentre outros modelos, sendo o pregão a 6ª modalidade criada em 2002 pela lei 10.520.

Figura 01 – Categoria de Licitações

Concorrência: Pode participar qualquer interessado que na fase de habilitação preliminar satisfaçam os requisitos mínimos de qualificação exigidos no edital.

Tomada de preços: Realizada entre interessados devidamente cadastrados ou que atenderem a todas as condições exigidas para cadastramento.

Concurso: No concurso há a instituição de prêmio ou remuneração aos vencedores, que possuirá caráter incentivo e não de pagamento aos serviços prestados. O autor do projeto se obriga a ceder os direitos relativos ao seu trabalho à Administração, que poderá utilizá-lo para o fim previsto no Edital de licitação (JusBrasil).

Essa modalidade é destinada à escolha de trabalhos que exijam uma criação intelectual como trabalhos científicos, projetos arquitetônicos, entre outros. Um exemplo de concurso é o plano para a cidade de Brasília, no qual o vencedor foi Lúcio Costa. O autor do projeto deve ceder os direitos do trabalho à Administração.

Convite: O convite é a mais modalidade mais simples. Realizada entre interessados do ramo do qual se trata a licitação, escolhidos e convidados pela Administração. A divulgação deve ser feita em quadros de avisos do órgão em locais de ampla divulgação.

Leilão: Pode participar qualquer interessado. Nessa modalidade é onde o governo poderá vender bens móveis que não possuem mais utilidade e por isso podem ser colocados à venda para obtenção de renda. Assim como pode ocorrer a

venda de produtos apreendidos ou empenhados.

Pregão: Art. 1º Para aquisição de bens e serviços comuns, poderá ser adotada a licitação na modalidade de pregão, que será regida por esta Lei(Lei 10.520).

A disputa é feita por propostas e lances de maneira sucessiva em sessão pública, presencial ou eletrônica. Nessa modalidade não há limites de valores e o Pregão funciona de maneira inversa das outras licitações, sendo primeira a análise da proposta e depois a análise da documentação.

Existem ainda dois outros tipos de licitação, que é a dispensa de licitação e inexigibilidade de licitação.

Dispensa de licitação: Utilizada em casos especiais onde é necessária uma atitude rápida e eficaz. Existem muitos motivos para a dispensa de licitação, sendo um dos principais, casos de emergência ou calamidade pública onde é necessária uma rápida atitude, dispensando algumas burocracias da licitação.

Inexigibilidade de licitação: Nessa modalidade não ocorre necessariamente a competição entre os participantes por um dos concorrentes possuírem características e habilidades que o tornam único, automaticamente eliminando os outros candidatos e dispensando a licitação.

Há diferença entre inexigibilidade e dispensa de licitação é que no caso da dispensa licitação ainda ocorre um processo de licitação, o que não é o caso da inexigibilidade, pois um concorrente acaba inibindo os demais, tornando o processo de licitação desnecessário.

2.5 Dados Abertos

Dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras (Open Data Handbook).

Esta forma é bem parecida com a Gaussiana, uma das diferenças mais características é que a distribuição da possibilidade entre os valores abaixo e acima da média não são os mesmos.

A lei de acesso à informação pública foi sancionada em 18 de novembro de 2011 e regula todo o acesso a dados e informações de posse do governo. E os dados disponibilizados devem seguir as três leis de dados abertos.

Art. 3º Os procedimentos previstos nesta Lei destinam-se a assegurar o direito fundamental de acesso à informação e devem ser executados em conformidade com os princípios básicos da administração pública (Lei 12.527,

2011).

As chamadas três “leis” dos dados abertos não são leis no sentido literal, promulgadas por algum Estado. São, em suma, um conjunto de testes para avaliar se um dado pode, de fato, ser considerado aberto. (Portal Brasileiro de Dados Abertos)

Foram propostas por David Eaves, especialista em políticas públicas. São elas:

1. Se o dado não pode ser encontrado na web, não existe;
2. Se o dado não estiver disponível em formato compreensível por máquina, não pode ser reutilizado.
3. Se o dado não puder ser replicado por algum dispositivo legal, ele não é útil.

Como o próprio nome diz, são dados abertos ao público, podendo ser usado por qualquer pessoa para acessá-los, utilizá-los, modificá-los e compartilhá-los para qualquer finalidade, estando sujeito somente a exigências que visem preservar sua proveniência. Quanto à distribuição dos dados, existem alguns pontos importantes que devem ser considerados, como:

Completo: Dados públicos é qualquer informação eletronicamente gravada. E estes dados não podem estar sujeito a controle de acesso, privacidade, segurança.

Primário: Os dados devem ser publicados no formato que são coletados, sem nenhum tipo de tratamento, e da forma mais fiel possível.

Atual: Os dados devem ser disponibilizados o mais rápido possível.

Processável por máquina: Os dados devem ser disponibilizados de forma a poder ser lida por linguagem de máquina.

Sem formato de proprietário: Os dados devem estar disponíveis em formato que todos tenham acesso, sem controle de uma determinada entidade.

Disponibilidade e acesso: Os dados devem estar disponíveis, sendo preferencialmente podendo ser baixado pela internet. E tendo um custo razoável, preferencialmente de graça para reprodução. Também deve estar disponível a todos sem necessidade de identificação.

Reutilização e redistribuição: Deve ser possível reutilizar e redistribuir os dados, podendo até ser combinados com outros conjuntos de dados.

Participação Universal: Todos devem ser capazes de baixar os dados, não havendo discriminação por grupo de pessoas, áreas, pessoas.

O TCU incentiva às organizações públicas a disponibilizar seus dados, sendo os motivos para este incentivo o seguinte:

- Transparência na gestão pública;
- Contribuição da sociedade com serviços inovadores ao cidadão;
- Aprimoramento na qualidade dos dados governamentais;
- Viabilização de novos negócios;
- Obrigatoriedade por lei

2.6 Análises Estatísticas

Estatística é um conjunto de técnicas utilizadas para a coleta, organização, resumo, análise e interpretação de dados. O avanço da informática e a popularização dos computadores contribuíram para o uso de métodos estatísticos. (Valéria Ferreira, 2015).

Estatística é uma ciência exata que visa fornecer subsídios ao analista para coletar, organizar, resumir, analisar e apresentar dados. Trata de parâmetros extraídos da população, tais como média ou desvio padrão.

A estatística fornece-nos as técnicas para extrair informação de dados, os quais são muitas vezes incompletos, na medida em que nos dão informação útil sobre o problema em estudo, sendo assim, é objetivo da Estatística extrair informação dos dados para obter uma melhor compreensão das situações que representam.

A estatística divide-se em duas áreas: A descritiva (que descreve e analisa um conjunto de dados sem tirar conclusões) e a indutiva (que se baseia na análise e na interpretação de dados). Tal conhecimento torna-se parte fundamental de diversas áreas, principalmente da área de pesquisas científicas.

Estatística tem como base o estudo de uma população, e este estudo pode ser feito investigando todos os elementos de uma população ou por amostragem, ou seja, selecionando alguns elementos da população. Obviamente teria-se uma precisão muito superior se fosse analisado o grupo inteiro, a população, do que uma pequena parcela representativa.

População é o conjunto de todos os elementos ou resultados sob investigação. Amostra é qualquer subconjunto da população. (Morettin, 2017)

População: É o conjunto de informações de pelo menos uma característica em comum, cujo comportamento interessa analisar. Conjunto de todas as medidas e observações relativas ao estudo de determinado fenômeno que formam o universo de nosso estudo, podendo ser uma população finita ou infinita.

- **População finita:** Apresenta um número limitado de elementos. Podendo enumerar todos os componentes.
- **População infinita:** Apresenta um número ilimitado o que torna impossível a enumeração de todos os elementos.

Amostragem: É a coleta de informações de parte da população. Os dados de observação registrados na amostra fornecem informações sobre a população. O processo pelo qual são tiradas conclusões sobre a população, a partir da amostra, é inferência estatística. Existem muitos motivos para se trabalhar com amostragem em vez da população, quando uma população é tão grande que sua contagem se torna "infinita", se torna melhor utilizar uma amostra para representar a população.

2.6.1 Tipos de amostragem

Existem dois tipos de técnicas de amostragem, a probabilística e a não probabilística. Ao selecionar os elementos que farão parte da amostra, pode-se selecioná-los mais de uma vez, trabalhando com reposição, ou pode trabalhar sem reposição, nesse caso, removendo o elemento da população. Geralmente, a amostragem com reposição é a mais adequada, pois implica independência entre os elementos.

2.6.1.1 Técnicas de amostragem probabilística

O objetivo da amostragem probabilística é medir a precisão da amostra obtida, tendo como base o resultado da própria amostra. A seguir veremos as diferentes técnicas de amostragem empregadas.

Amostragem aleatória simples: Nesta técnica, os elementos que farão parte da amostra são sorteados aleatoriamente, e a quantidade de elementos varia de acordo com o tamanho da amostra. Para esse procedimento o ideal é que o tamanho da população seja finito, caso a população seja muito grande, deverá se utilizar de ferramentas para gerar números aleatórios e sortear os elementos. Para esse tipo de amostragem é importante que a população seja homogênea, ou seja, que os elementos sejam similares. Caso seja heterogênea, poderá se obter avaliações diferentes quanto à variável em estudo. Há duas maneiras de obtermos a amostra, por meio do método de sorteio, no qual são escolhidos um a um até que esteja completa a amostragem e a tabela de números aleatórios, na qual serão

sorteados até que seja satisfeita a solicitação.

Amostragem estratificada: Trabalhamos com está técnica quando temos uma população heterogênea. Dividimos a população em subgrupos mais homogêneos e após isso selecionamos os elementos que farão parte da amostra através de uma amostragem aleatória simples ou através de uma seleção proporcional ao tamanho de cada grupo. Por exemplo, para obter uma amostra estratificada de estudantes universitários, o pesquisador primeiro organizaria primeiro a população por semestre de graduação e então selecionar determinado número de representantes de calouros, pessoas que estão no meio do curso e formandos, por exemplo. Isso garante que o pesquisador tem quantidades adequadas de indivíduos de cada classe na amostra final.

Amostragem sistemática: Nesta técnica, utilizamos um sistema preestabelecido para a seleção de elementos. Primeiramente nós ordenamos os elementos da população numa lista, e assim, os elementos serão selecionados por intervalos regulares que ocorrem a partir do elemento inicial. Devemos apenas nos atentar para não surgir sequências periódicas ou cíclicas na ordenação dos elementos. Por exemplo, se a população do estudo contém 2000 estudantes do ensino fundamental e o pesquisador quer uma amostra de 100 estudantes. Estes poderiam ser colocados em uma lista e cada 20º estudante seria selecionado para inclusão na amostra.

Amostragem por conglomerado: Similar à amostragem estratificada, dividimos os elementos em subgrupos (conglomerados), mais dessa vez os elementos são heterogêneos, em seguida selecionamos aleatoriamente alguns conglomerados e escolhemos todos os elementos dos conglomerados selecionados para compor a amostra. Por exemplo, digamos que a população-alvo em um estudo seja membros de igrejas no Brasil. Não há uma lista de todos os membros de igrejas no país. O pesquisador poderia, nesse caso, criar uma lista de igrejas no Brasil, escolher uma amostra de igrejas e então obter listas de membros dessas igrejas.

2.6.1.2 Técnicas de amostragem não probabilística

Amostragem por conveniência: Escolhemos os elementos mais acessíveis para fazer parte da amostra, ou seja, selecionamos elementos que estejam prontamente disponíveis e não selecionado por um critério estático. Esta técnica representa uma maior facilidade operacional e baixo custo de amostragem, porém tem a incapacidade de fazer afirmações gerais com rigor estatístico. Por exemplo, se quisermos uma amostra sobre o número de imigrantes no país, buscamos por pessoas que conhecemos, ou seja, que seja fácil de nos.

Amostragem por cota: Os elementos selecionados para fazer parte da amostra são retirados da população por meio de cotas estabelecidas. É parecida com a técnica de amostragem estratificada, a diferença é que os elementos são selecionados por julgamento e não de maneira aleatória e depois é confirmada as características dos elementos. Essa técnica é muito utilizada em pesquisa de opinião e de mercado por possuir um baixo custo. Por exemplo, se estivermos conduzindo uma amostragem com base na distribuição da população do país, provavelmente precisaremos saber qual a proporção de mulheres e homens, ou a proporção de homens e mulheres por grupo etário ou escolaridade. Nesse caso devemos selecionar unidades com as mesmas proporções da população nacional.

Ao trabalharmos com estatística, precisamos nos atentar e executar bem um processo de coleta de amostra, mas mesmo assim, pode e provavelmente ocorreram alguns erros. Se você fizer uma amostra com outros elementos pode ser que encontre outra estimativa. Ou, ainda, pode obter uma amostra que gere resultados diferentes se trabalhasse com toda a população. A partir disso podemos esperar dois tipos de erros:

Erro amostral: Diferença entre o resultado da amostra e o resultado da população.

Erro não amostral: Quando os dados são coletados e analisados de maneira incorreta.

Se os elementos que farão parte da amostra são obtidos por meio de um processo probabilístico, é de esperar que sejam representativos da população, permitindo assim analisar o erro amostral, mas independente disso devemos ter cuidado para não gerar um erro não amostral, planejando e realizando uma boa coleta e análise dos dados.

2.6.2 Variáveis

Na análise estatística os dados coletados de uma população ou amostra são resultantes das variáveis em estudo. As variáveis são a característica de interesse no nosso estudo. Tomando como exemplo este estudo, podemos ter como interesse o valor gasto em licitação, os órgãos, etc. Essas variáveis podem ser classificadas em qualitativas ou quantitativas.

Variáveis qualititativas se dividem em ordinais e nominais. Se os dados podem ser ordenados naturalmente, como por exemplo, desempenho de um aluno (bom, regular, péssimo), essa variável é classificada como ordinal. Já as nominais não existem uma ordenação como, sexo, estado civil, etc.

Variáveis quantitativas podem ser classificadas como discretas ou contínuas.

As discretas são provenientes de uma operação de contagem e, por isso, só fazem sentido números inteiros. Quanto às contínuas, são números resultantes de medições, sendo assim, possui valores com casas decimais.

2.6.3 Distribuição de frequências

Para se conhecer melhor o comportamento das variáveis no estudo, nós agrupamos os dados em classes, de tal forma que contabilizamos o número de ocorrências em cada classe. Assim, podemos visualizar os dados de uma maneira mais resumida e que nos permita extrair informações sobre seu comportamento, analisando o número de vezes que cada dado ocorre e a porcentagem com que aparece. A seguir, apresentamos algumas definições necessárias à construção da distribuição de frequências.

Frequência absoluta: É o número de observações que dado aparece.

Frequência relativa: É o quociente entre a frequência absoluta e a soma das frequências. Podendo ser expressa em porcentagem.

Frequência acumulada: É o somatório de todas as frequências anteriores até a atual.

Dependendo da quantidade de dados que se esteja trabalhando, pode acontecer da tabela de distribuição de frequência ficar muito extensa e não conseguir resumir adequadamente o conjunto de dados. Neste caso, agrupamos os dados em intervalos de classes. Este intervalo possui os seguintes valores:

Limite inferior: Menor valor que uma variável pode assumir.

Limite superior: Maior valor que uma variável pode assumir.

Ponto médio: Média aritmética entre limite inferior e

superior. **Amplitude:** Diferença entre limite superior e inferior.

2.6.4 Medidas de resumo

Quando trabalhamos com estatística também é importante termos conhecimento de como resumir os dados de forma a encontrar o centro de como os dados se distribuem e a forma como estão dispersos.

Média: Um termo que está bastante presente em nossas vidas, sendo um cálculo simples de ser efetuado para ser encontrado, basta apenas somar os valores e dividir pelo total de valores.

Figura 02 – Cálculo da média

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Moda: É o conjunto que aparece com maior frequência. Pode acontecer de em um grupo de dados, nenhum conjunto de dados aparecer com mais frequência, nesse caso não teremos moda. Ou pode haver mais de um conjunto de dados que se repete com mais frequência, assim nosso conjunto pode ser bimodal ou multimodal. A moda também pode ser encontrada em uma variável qualitativa, diferente de outras medidas de posição.

Mediana: É a divisão do conjunto de dados no meio, de forma que ambas as partes tenham a mesma quantidade de valores. Caso o número de elementos for ímpar, a mediana é simples, será exatamente o valor do meio, ou seja:

Figura 03 – Cálculo da mediana número ímpar

$$\mathrm{Md} = \mathrm{x}_{\frac{n+1}{2}}$$

Caso o número de elementos seja par, então a mediana será a média dos dois valores do meio, sendo:

Figura 04 – Cálculo da mediana número par

$$\mathrm{Md} = \frac{\mathrm{x}_{\frac{n}{2}} + \mathrm{x}_{\frac{n}{2}+1}}{2}$$

Desvio médio: É a média dos valores absolutos dos desvios em relação à média.

Figura 05 – Cálculo do desvio médio

$$dm = \frac{\sum_{i=1}^{n} \left| x_i - \overline{x} \right|}{n}$$

Variância: É a média dos quadrados dos desvios.

Figura 06 – Cálculo da variância

$$s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2}{n-1}$$

2.6.5 Tipos de Gŕaficos

Para uma melhor apresentação e compreensão dos dados coletados, é interessante a utilização de gŕaficos. Devendo tomar cuidado em sua construção pois, um gŕafico mal elaborado pode levar a conclusões equivocadas e com isso pode se gerar atitudes equivocadas. Existem diversos tipos de gráficos que podem ser usados para a análise estatística, cada um tendo seu objetivo. A seguir será apresentado os principais gráficos utilizados.

Gráfico de linhas: Muito utilizado quando os dados estão distribuídos seguindo uma variável de tempo. Retratando as mudanças no decorrer do tempo. Muito eficiente quando queremos analisar se o conjunto de dados possui tendências. Exemplo: evolução do total de modalidades em três meses.

Figura 07 exibi como é um gráfico de histograma

Figura 07 – Gráfico de linha

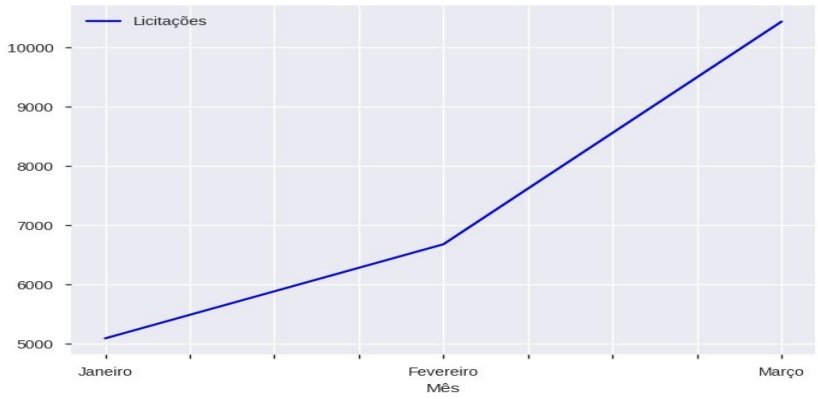


Gráfico de barras: Um tipo de gráfico para variáveis qualitativas, ou seja, dados que possam ser ordenados, podendo exibir o gráfico de forma horizontal ou vertical, também é possível aninhar as barras, permitindo comparar determinados dados. Como por exemplo, a frequência que uma modalidade se repete.

Figura 08 exibi como é um gráfico de barra.

Figura 08 – Gráfico de barra

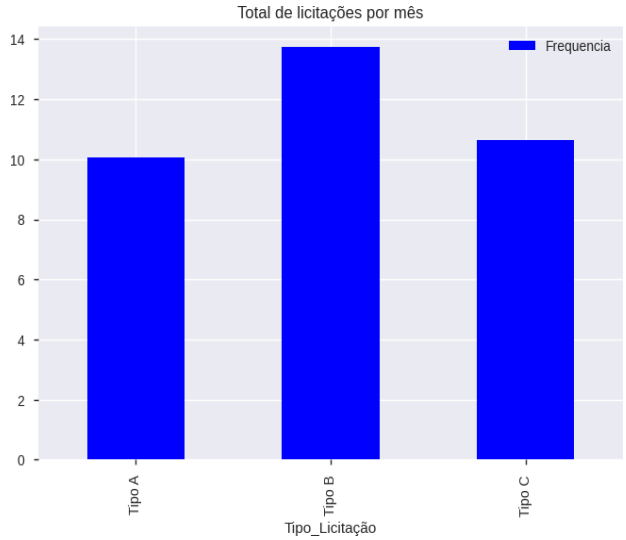
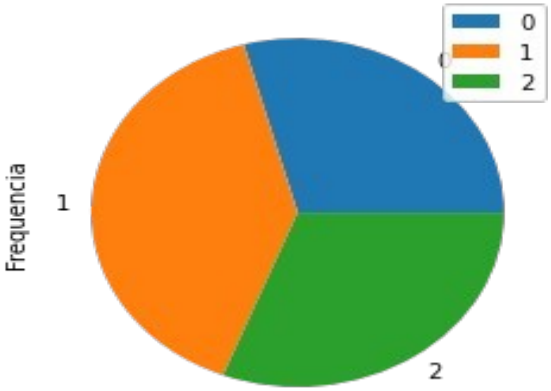


Gráfico de setores: Mais conhecido como gráfico de pizza , é bastante utilizado quando desejamos exibir variáveis qualitativas, e queremos visualizar a proporção de cada categoria. Utilizando os mesmos dados da tabela anterior, teríamos o seguinte gráfico.

Figura 09 exibi como é um gráfico de setor.

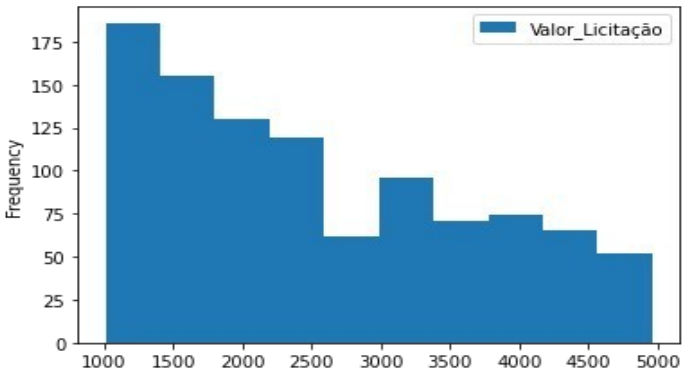
Figura 09 – Gráfico de setor



Histograma: O gráfico de histograma é semelhante ao gráfico de barras, porém é utilizado quando possuímos dados divididos em intervalos de classes, eficiente para visualizarmos a distribuição dos dados. Como exemplo faremos um gráfico para verificar a frequência do valor da licitação no mês de janeiro.

Figura 10 exibi como é um gráfico de histograma.

Figura 10 – Gráfico de histograma



Além desses gráficos existem diversos outros gráficos, como o gráfico de mapas, dispersão, bolhas, entre outros, cada um tendo sua finalidade e melhor aplicação. A escolha do gráfico para a visualização dos dados também é uma etapa muito importante. Devemos sempre escolher apresentar os dados de uma maneira simples, e fácil compreensão, tornando o gráfico autoexplicativo.

2.7 Introdução ao Python

Para muitas pessoas, a linguagem de programção Python tem um forte apelo. Desde o seu surgimento em 1991, Python se tornou uma das linguagens de programação interpretadas mais populares. (Wes, 2018)

Python é uma linguagem de alto nível que surgiu em 1991, sendo uma linguagem de programação de propósito geral, e com tipagem dinâmica. Python é multiparadigma com foco em orientação a objetos, procedural e funcional. Foi desenvolvida com base na linguagem ABC, trazendo suas melhores propriedades e acrescentando outros recursos importantes, como: Listas, dicionários, declaração básica e uso obrigatório, sendo ainda uma linguagem de script simples, tendo parte da sintaxe derivada de C combinada com recursos de sua biblioteca e módulos e frameworks desenvolvidos por terceiros.

Python é principalmente orientada a objetos, facilitando o controle sobre a estabilidade dos projetos ao se tomar grandes proporções. Mas além disso, python também permite que o usuário programe de forma procedural, ou com outro paradigma. Python também é altamente modular, ou seja, provavelmente alguém já fez parte ou todo do programa, economizando tempo e permitindo se concentrar no problema real.

A linguagem foi projetada com a filosofia de enfatizar a importância do esforço do programador sobre o esforço computacional. Prioriza a legibilidade do código sobre a velocidade ou expressividade. Segundo a pesquisa da revista IEEE Spectrum, python lidera no ranking das melhores linguagens de 2020.

2.7.1 Características

Python possui diversas características que a fazem se tornar uma linguagem de programação de alto nível. A seguir foram esclarecidas algumas dessas características e junto, uma breve descrição.

- **Fácil leitura e compreensão**

  Python possui uma sintaxe limpa e clara, assim como, possui um conjunto de bibliotecas bem estruturadas. Priorizando a legibilidade do código, possuindo uma sintaxe concisa e clara.

- **Fácil manutenção**

  Devido a simplicidade da sintática e da excelente estruturação da biblioteca, a manutenção do código é muito fácil e compreensível.

- **Multiparadigma**

  Python possui a capacidade de ser escrita em diversos paradigmas, sendo orientado a objetos, imperativo, funcional e procedural.

- **Interoperabilidade**

  Python também tem a capacidade de funcionar em conjunto com várias outras linguagens.

- **Interpretador interativo**

  Possibilita testar o código em tempo real, conforme se desenvolve o código, receberá o resultado, antes de iniciar a compilação.

- **Indentação**

  Diferente das outras linguagens, para facilitar a leitura e torna o código mais legível, python utiliza espaços em brancos e forçando uma indentação, ao invés de utilizar delimitadores como chaves.

Python possui diversas outras características além das citadas acima. Tendo também uma tipagem dinâmica e exigindo poucas linhas de código se comparado a outras linguagens.

2.7.2 Interpretador

Python é tanto uma linguagem interpretada quanto uma linguagem compilada. Um compilador traduz linguagem Python em linguagem de máquina – código, é traduzido em um código intermediário que deve ser executado por uma máquina virtual conhecida como PVM. Bastante similar ao do Java JVM.

O interpretador faz esta 'tradução' em tempo real para código de máquina, ou

seja, em tempo de execução. Já o compilador traduz o programa inteiro em código de máquina de uma só vez e então o executa, criando um arquivo que pode ser rodado (executável). O compilador gera um relatório de erros (casos eles existam) e o interpretador interrompe a tradução quando encontra um primeiro erro. Em geral, o tempo de execução de um código compilado é menor que um interpretado já que o compilado é inteiramente traduzido antes de sua execução. Enquanto o interpretado é traduzido instrução por instrução. Python é uma linguagem interpretada, mas, assim como Java, passa por um processo de compilação.

2.7.3 Estrutura

Assim como nas outras linguagens, as construções de python incluem estruturas de seleção (if, else, elif); estruturas de repetição (for e while); construção de classes se dá por (class) e a construção de sub-rotinas como métodos e funções (def).

Tipos de dados

Os valores e objetos têm tipos bem definidos, sendo uma linguagem de tipagem forte, ou seja, o mesmo dado não pode ser tratado como de outro tipo.

Tipo de dado	Descrição
Int	Número inteiro sem ponto flutuante
Float	Número com ponto flutuante (vírgula)
Boolean	True ou False
Complex	Número complexo (3 + 2k)
List	Lista heterogênea mutável
Tuple	Tupla imutável
Set, Fronzeset	Conjunto não ordenado sem elementos duplicados
Str, unicode	Cadeia de caracteres imutável
dict	Conjunto associativo

Palavras reservadas

Palavras reservadas definem as regras de sintaxe e estrutura da linguagem, e não podem ser usadas como nome de variáveis. Python tem mais de 30 palavras reservadas.

Figura 11 exibi os tipos de palavras reservadas em Python

Figura 11 – Palavras reservadas Python

and	as	assert	break	class	continue
def	del	elif	else	except	exec
finally	for	from	global	if	import
in	is	lambda	nonlocal	not	or
pass	raise	return	try	while	with
yield	True	False	None		

Outras linguagens de programação podem ser para a mesma finalidade e fatores pessoais também influenciam, mas devido as suas característica, bibliotecas de ciências de dados (NumPy, SciPy, Pandas, entre outras) e ferramentas como Jupyter-Notebook. Python acaba se tornando uma das linguagens favoritas para a análise de dados.

A princípio python não foi desenvolvida para análise de dados Python é uma linguagem de programação de amplo uso, para diversos fins, como: coleta de dados, análise de dados, construção de aplicativos web e muito mais.

3 Python

Esta parte do material visa explicar melhor sobre as bibliotecas e ferramentas da linguagem python e como elas podem ser usadas para uma análise exploratória de um conjunto de dados, de forma a encontrar padrões nos dados e extrair informações. Podendo haver duas maneiras de explorar um conjunto de dados: por meio de técnicas estatísticas e de visualização, por meio de gráficos.

Todos os dados utilizados nestre trabalho foram retirados do site do governo, portal brasileiro de dados abertos (dados.gov.br). Para este trabalho será utilizado o

jupyter-notebook que é um ambiente computacional onde seu código pode ser dividido por células e podendo até ser concatenado com textos, tornando o projeto organizado e explicativo. Faremos uso também das principais bibliotecas de Python utilizadas para analise de dados, como: Pandas e NumPy. Toda a documentação juntamente com o código se encontra disponivel no github por meio do site: https://github.com/Guilherme8/TCC.

3.1 Principais bibliotecas python

Assim, como em outras linguagens, python possui diversas bibliotecas que auxiliam no desenvolvimento de seus projetos, será mostrada as principais bibliotecas implementadas quando estamos trabalhando com ciência de dados.

NumPy: Ela oferece código aglutinador para as estruturas de dados , os algoritmos e a biblioteca necessários a maioria das aplicações cientificas que envolvam dados númericos em Python (Wes, 2018).

É uma das principais bibliotecas dessa área. Utilizada no processamento de grandes matrizes e matrizes multidimensionais, possui uma extensa coleção de funções matemáticas e métodos implementados permitem a execução de várias operações. Devido às suas correções e melhorias, agora algumas funções podem manipular arquivos de qualquer codificação.

Pandas: Oferece estrutura de dados de alto nível e funções, projetadadas para fazer com que trabalhar com dados estruturados ou tabulares seja rápido, fácil e expressivo (Wes, 2018).

Fornece uma grande variedade de ferramentas de análise. Permite traduzir operações complexas com dados em um ou dois comandos. Pandas que nos permite ler diferentes tipos de arquivos como: csv, xlxs, etc. E nos permite agrupar, filtrar, combinar dados.

SciPy: Sendo Baseado no NumPy, sua principal estrutura é uma matriz multidimensional, implementada pelo NumPy. Além disso, a biblioteca possui ferramentas para cálculos de álgebra linear, probabilidade, etc.

StatsModels: Utilizado para análise de dados estatísticos, estimação de modelos estatísticos, realização de testes estatísticos, etc. Permite a implementação de diversos métodos de aprendizado de máquina e diversas possibilidades de plotagem.

Matplotlib: Biblioteca Python mais popular para fazer plotagens e gerar

outras visualizações de dados bidimensionais (Wes, 2018).

É a biblioteca utilizada para criar gŕaficos bidimensionais. Com ela podemos construir diversos gráficos, sendo gráficos de histogramas, dispersão e coordenadas não cartesianas.

Plotly: Uma das bibliotecas mais populares que permite construir, de maneira fácil, gráficos sofisticados. Entre eles temos gráficos de contorno, ternários e 3D. E o Plotly também permite trabalhar em aplicativos web interativos.

Além dessas, Python possui muitas outras bibliotecas, utilizadas em algoritmos de machine learning, deep learning, processamento neural, etc. Bibliotecas como: Scikit-learn para tarefas de aprendizado de máquina e mineração de dados, TensorFlow, que na verdade é um framework, utilizado para trabalhar com redes neurais, entre diversas outras grandes bibliotecas.

3.1 Introdução ao Jupyter-Notebook

Jupyter é um ambiente de desenvolvimento interativo baseado na web para notebooks. Sendo muito prático para organizar e trabalhar com grandes volumes de dados em ciência de dados. O jupyter notebook permite criar, visualizar documentos que têm código ativo, equações, etc. Podendo usar limpeza, transformação de dados, visualização de dados, aprendizado de máquina e muito mais. Qual a vantagem de se trabalhar com notebooks? Em Data Science vamos executar trechos de códigos inúmeras vezes, já que estamos trabalhando com leitura e processamento de dados. Então executar scripts inteiros quando o nosso objetivo está em apenas um trecho do arquivo fica inviável, sem falar no tempo que será gasto. Desta maneira o Jupyter traz a computação interativa.

O Jupyter nos permite não somente desenvolver nosso código, como costumamos fazer com scripts, como também ele permite uma melhor organização e visualização tanto dos algoritmos codificados como também dos dados que estão sendo processados por meio da plotagem de gráficos.

3.2 Vantagens de trabalhar com notebook

Como já dito anteriormente, o grande diferencial que o Jupyter traz é a capacidade da computação de maneira interativa, de forma que ainda seja possível tanto uma melhor visualização dos seus dados bem como também a facilidade para compartilhar seus códigos (em formato de notebooks).

Diferente da programação “tradicional” que estamos habituados a criar scripts de códigos (pequenos arquivos de código com instruções significantes), com o Jupyter nós vamos trabalhar com notebooks, ou seja, arquivos que nos permitem organizar o nosso código dividido por partes que são colocadas em células permitindo assim sua execução individual. E como os valores são persistidos na memória após a sua execução eles se tornam acessíveis a outras células também, mesmo que de forma atemporal. Com o nosso código dividido em células nós vamos precisar apenas executar e re executar aquilo que for do nosso interesse, tornando o processo bem mais fácil, rápido e confortável.

Com o jupyter-notebook também é possível unir texto e código com isso é possível descrever a pesquisa, sua abordagem, narrando as etapas da análise. Todo o documento pode ser escrito de forma dinâmica, escrevendo trechos do código, exibindo os resultados no relatório e continuando a análise, acrescentando mais descrição e códigos, podendo assim gerar um documento completo de programação com a descrição da proposta, o código utilizado, e os resultados e quais conclusões podem ser tiradas com isso, abrindo espaço para uma discussão.

3.3 Instalação

A instalação do jupyter é super prática e simples, pode ser instalado usando conda ou pip.

Conda

Se você usar conda, pode instalar da seguinte maneira

```
conda install -c conda-forge jupyterlab
```

pip

Se você usar pip, pode instalar da seguinte maneira

```
pip install jupyterlab
```

Também tem a possibilidade de usar o jupyter via web, nesse caso, não

precisando instalar nada.

4 Estudo de Caso

Nosso trabalho consiste em analisar os dados abertos de licitações do governo federal no período de janeiro-agosto de 2020. E como já dito anteriormente faremos uso do python no jupyter notebook.

4.1 Leitura e tratamento dos dados

O processo de tratamento de dados é um dos mais demorados e custoso. Nesse processo o analista deve ser cuidadoso pois, é necessário identificar os dados que estão faltando e descobrir se existem valores que podem alterar o resultado. Sem o tratamento e limpeza adequado dos dados não é possível fazer uma análise correta.

Figura 12 exibi como importar as bibliotecas necessárias para nossa análise.

Figura 12 – Importando bibliotecas

```
import pandas as pd
import math as math
import numpy as np
import matplotlib.pyplot as plt
import plotly.offline as py
import  plotly.graph_objects  as  go
```

E agora passamos todos os arquivos que usaremos no projeto para o programa fazer a leitura. Como estamos trabalhando com diversos arquivos os exemplos ficariam muito grandes, portanto será mostrada apenas uma parte do código, senão as explicações ficariam muito extensa variando apenas os nomes das variáveis, os dataFrame referente às licitações terão o nome “lic01”, “lic02”, assim por diante, totalizando oito, e os dataFrame referente aos participantes das licitações terão o nome “partLic01”, e assim sucessivamente.

Figura 13 exibi como é feita a importação dos arquivos.

Figura 13 – Importando arquivos

```
lic01 = pd.read_csv('/media/guilherme/Arquivos/Guilherme/Estudos/Licitacoes/202001_Licitacoes/
202001_Licitacao.csv',sep=";",encoding="utf-8")
partLic01 = pd.read_csv('/media/guilherme/Arquivos/Guilherme/Estudos/Licitacoes/202001_Licitacoes/
202001_ParticipantesLicitacao.csv',sep=";",encoding="utf-8", low_memory=False)
```

Além de lermos o arquivos que desejamos também passamos outros comandos junto com a leitura, como:

Read_csv: É utilizado para ler um arquivo de extensão .csv.

Sep: É para definir como é dada a separação entre as colunas, vamos separar com “;”.

Encoding: É para declarar a codificação das fontes, nesse caso utf-8, assim ele reconhece caracteres especiais.

low_memory: Este comando na verdade não faz nada de diferente, é uma opção que deveria estar obsoleta, mas ainda não está. Em alguns arquivos você pode receber o erro “low memory” (mesmo não estando no limite de memória do seu notebook), é porque adivinha os tipos para cada coluna exige muita da memória. O pandas só determina o tipo de uma coluna depois que todo o arquivo for lido. Por isso passamos o comando low_memory, uma outra opção seria passar dentro do read_csv() o tipo da coluna, adicionando o parâmetro dtypes={‘Número Licitação’: int64}.

Figura 14 exibi como utilizar o comando dtypes para ver o tipo de cada coluna do arquivo.

Figura 14 – Verificando tipos

```
lic01.dtypes
```

```
Número Licitação          int64
Número Processo          object
Objeto                   object
Modalidade Compra        object
Situação Licitação       object
Código Órgão Superior     int64
Nome Órgão Superior      object
Código Órgão              int64
Nome Órgão               object
Código UG                 int64
Nome UG                  object
Município                object
Data Resultado Compra    object
Data Abertura            object
Valor Licitação          object
dtype: object
```

Como pode ser notado a coluna “Valor Licitação” foi lida de forma errada ficando como tipo object, para isso devemos fazer um tratamento e alterar o tipo para float.

Figura 15 exibi como formatar o valor.

Figura 15 – Formatando valor

```
lic01['Valor Licitação'] = lic01['Valor Licitação'].str.replace(',','.').astype(float)
```

Nem todas as colunas do dataFrame são importantes para nossa análise, como por exemplo, Código do Órgão Superior ou Código UG, como não está no nosso foco iremos removê-las pois, manter informações desnecessárias só será um gasto de memória e processamento desnecessária. Para isso utilizaremos o método drop() da biblioteca Pandas, que remove linhas e colunas.

Figura 16 exibi como aplicar o método

Figura 16 – Removendo linhas e colunas

```
lic01.drop(['Código Órgão Superior', 'Código Órgão', 'Código UG'], inplace=True, axis=1)
```

Figura 17 exibi que faremos a mesma coisa para dataFrame de participantes.

Figura 17 – Removendo linhas e colunas demais tabelas

```
partLic01.drop(['Código Órgão', 'Código UG'], inplace=True, axis=1)
```

Como pode ser notado, foi passado três parâmetros no método drop(). Entre colchetes passamos as colunas que desejamos excluir, o parâmetro inplace=True informa que as alterações devem ser feitas no dataFrame original, e o parâmetro axis=1, é para informar que exclusão deve ser aplicada na coluna especificada.

Outro ponto importante que devemos analisar e tratar são os dados ausentes (NaN). Para identificar valores NaN utilizamos o método isnull(). Mas esse método retorna o dataFrame inteiro, retornando False caso não possua valor NaN e True caso encontre valores do tipo NaN.

Figura 18 exibi como verificar valores ausentes.

Figura 18 – Verificando dados ausentes

```
lic01.isnull()
```

	Número Licitação	Número Processo	Objeto	Modalidade Compra	Situação Licitação	Nome Órgão Superior	Nome Órgão	Nome UG	Município	Data Resultado Compra	Data Abertura	Valor Licitação
0	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
5	False	False	False	False	False	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False	False	False	False	False	False

Para facilitar a visualização desses valores é utilizado em conjunto o método sum(), que retorna o total de valores NaN em cada coluna.

Figura 19 exibi como realizar a consulta.

Figura 19 – Contagem dados ausentes

```
lic01.isnull().sum()
```

```
Número Licitação            0
Número Processo             0
Objeto                      0
Modalidade Compra           0
Situação Licitação          0
Nome Órgão Superior         0
Nome Órgão                  0
Nome UG                     0
Município                   0
Data Resultado Compra       0
Data Abertura            3090
Valor Licitação             0
dtype: int64
```

Com a identificação de valores nulos, devemos agora fazer o devido tratamento, aplicando o método dropna(), será excluída toda linha que possui valor nulo. Mas, não é interessante excluirmos essas linhas pois afetaria nossa análise, então devemos fazer o preenchimento desses campos.

Para isso será utilizado o método fillna() e dentro dele passamos o parâmetro que substituirá os valores ausentes. É muito comum na substituição de valores ausentes se trabalharem com a média, mas nesse caso iremos substituir os valores ausentes da coluna “Data Abertura” pelos valores da coluna “Data Resultado Compra”.

A figura 20 mostra como realizar este tratamento

Figura 20 – Tratando valores ausentes

```
1
lic01.fillna(axis=1, method='ffill')
```

No processo de tratamento de dados também é importante verificar se existe valores duplicados, o método utilizado para isso é o duplicated(). O método retornará uma lista com todo o dataFrame, caso duas linhas sejam totalmente iguais terá o valor True. Para facilitar a visualização aplicarei o método sum() para retornar o total de valores duplicados.

Figura 21 exibi como realizar a consulta de linhas duplicadas.

Figura 21 – Visualizando linhas duplicadas

```
lic01.duplicated().sum()
```

```
0
```

Com isso verificamos que não existem linhas duplicadas. No dataFrame de participantes pode se verificar que existe colunas duplicadas.

A figura 22 exibi como realizar a consulta de linhas duplicadas.

Figura 22 – Visualizando linhas duplicadas demais tabelas

```
partLic01.duplicated().sum()
```

```
1409
```

Através da figura 23 explica como tratar essas duplicidades utilizando o método drop_duplicates().

Figura 23 – Removendo duplicatas

```
partLic01.drop_duplicates()
```

Um outro tratamento que devemos fazer para facilitar o trabalho com os dados é tirar os espaços em brancos dos nomes das colunas e substituir por “_”, utilizando o método str.replace().

Através da figura 24 podemos ver como fazer.

Figura 24 – Removendo espaços em brancos

```
lic01.columns = lic01.columns.str.replace(' ', '_')
```

Com isso a parte de limpeza e tratamentos de dados está concluída e podemos começar a analisar os dados.

4.2 Análise e Visualização dos dados

Com o devido tratamento dos dados podemos dar inicio na análises dos dados, em nosso dataFrame uma das principais colunas é a “valor_licitação”, a seguir é exibido uma tabela e um gráfico demonstrando o total gasto por mês em licitação.

As figuras 25 e 26 exibem o valor gasto por mês em licitações.

Figura 25 – Valor por mês

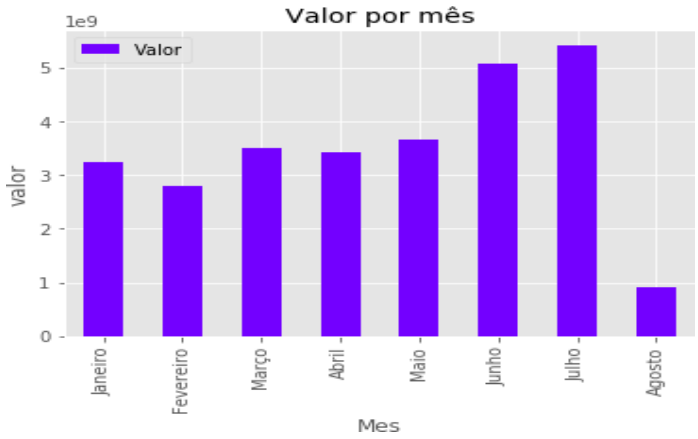


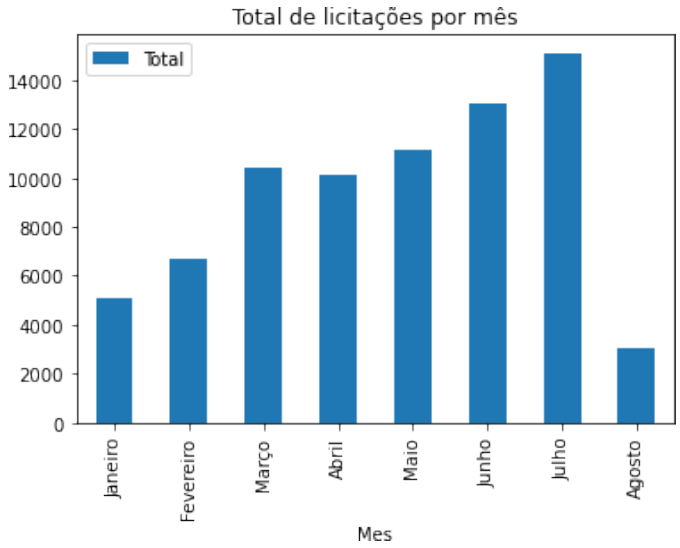
Figura 26 – Tabela valor por mês

Meses	Valor
Janeiro	R$ 3.235.405.819,22
fevereiro	R$ 2.789.293.218,96
março	R$ 3.499.439.861,96
abril	R$ 3.435.475.005,37
maio	R$ 3.653.134.190,33
junho	R$ 5.079.818.064,60
julho	R$ 5.431.214.194,69
agosto	R$ 898.611.516,69
Total	R$ 28.022.391.871,82

Através do gráfico podemos ver que os meses de junho e julho foram os de maiores gasto em licitações. Acompanhado da quantidade de licitações podemos ver que estes mês foram os que tiveram maior número de licitações tendo uma queda em agosto e apesar do mês de janeiro apresentar um gasto maior do que em fevereiro. Esta teve um maior número de licitações.

A figura 27 exibi o total de licitações por mês.

Figura 27 – Total de licitações por mês



A figura 28 exibi o código de como foi construido o gráfico de barras para o total de licitações por mês.

Figura 28 – Código total de licitações por mês

```
TotalLic.plot(kind='bar', x='Mes', y='Total')
plt.title("Total de licitações por mês")
plt.show()
```

Um gráfico simples de ser feito, onde TotalLic e o dataFrame que possui os dados que desejo utilizar, foi criado a partir da soma dos valores dos outros dataFrame e armazenado num novo. Kind é onde especificamos o tipo que gráfico que desejamos, sendo ‘bar’ para barra, ‘line’ para linhas, ‘pie’ para setores, etc. X será a variável do eixo X e y para variável do eixo Y. Title é o titulo do gráfico e por fim exibimos o gráfico com show().

A figura 29 demonstra a evolução das licitações no período de janeiro a agosto do ano de 2020, sendo um total de 74.662 licitações ocorridas.

Figura 29 – Evolução das licitações

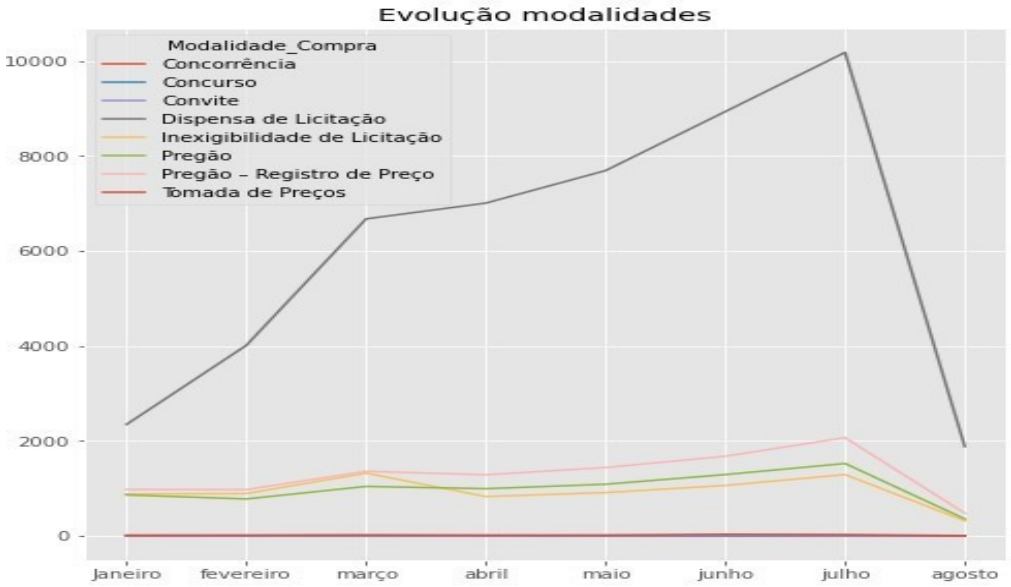
Evolução das Licitações (Jan / Ago)

Mês	Quantidade De propostas	Licitações	Valor Mediano
Janeiro	269242	5091	R$ 7.744,00
fevereiro	342319	6677	R$ 4.000,00
março	474727	10431	R$ 4.141,00
abril	478040	10141	R$ 5.000,00
maio	491655	11156	R$ 4.165,00
junho	595933	13014	R$ 4.502,78
julho	779257	15106	R$ 4.196,14
agosto	186191	3046	R$ 4.450,00
Média	452170,5	9332,75	R$ 4.774,87

Apesar do ano de 2020 estar sendo um ano atípico devido a pandemia, os dados indicam um aumento na quantidade de licitações. Através do método mode(), foi verificado que o tipo de licitação que mais ocorreu foi “dispensa de licitação” (o mesmo se repetindo em todos os meses), apresentando uma frequência acima de 50%, sendo a modalidade concurso a de menor frequência, ficando abaixo de 1%. Ao se observar a frequência das modalidades da maior para menor, temos respectivamente: Dispensa de Licitação (65,37%), Tomada de Preços (20,09%), Pregão – Registro de preço (13,74%), Pregão (10,63%), Inexigibilidade de licitação (10,05%), Concorrência (5,49%), Convite (2,95%), Concurso (0,27%).

A figura 30 explicita à evolução do total de licitações segundo as modalidades.

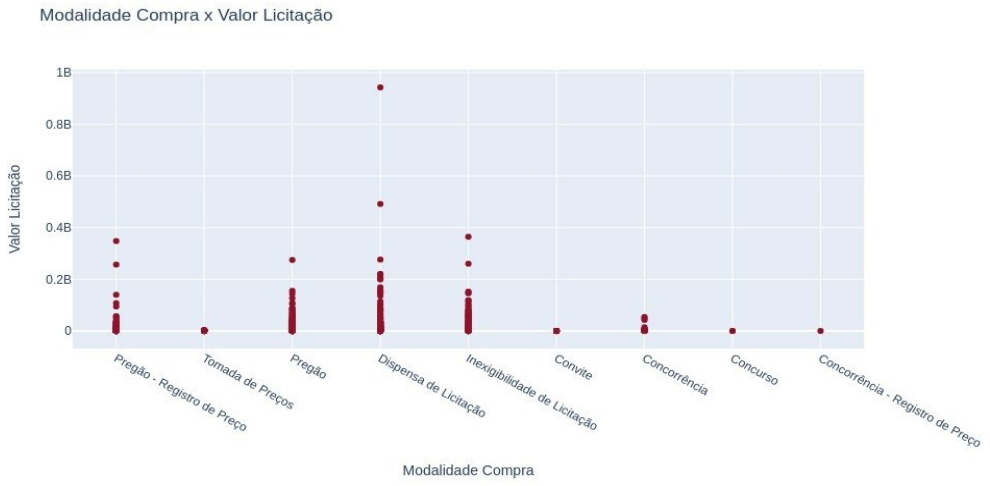
Figura 30 – Evolução das modalidades



É interessante analisarmos também os valores de cada licitação de acordo com a modalidade compra, verificando se existe algum valor que se destaca dos demais, ou se todos os valores costumam ter o mesmo preço. Podemos fazer isso utilizando um gráfico de dispersão, que é a representação de um conjunto de dados utilizando duas ou mais variáveis, muito utilizado quando se quer verificar se existe alguma relação de causa e efeito entre as variáveis.

Aqui iremos verificar se os valores da licitação mudam de acordo com a modalidade, sendo a modalidade da compra no eixo x e o valor da licitação no eixo y. A figura 31 a seguir exibe o gráfico de dispersão.

Figura 31 – Modalidade x Valor



A figura 32 a seguir exibi o código de como foi construido o gráfico de dispersão.

Figura 32 – Código gráfico dispersão

```
trace = go.Scatter(x=Ano20['Modalidade_Compra'],
                   y=Ano20['Valor_Licitação'],
                   mode='markers',
                   marker={'color':'#941229'})

data=[trace]

layout = go.Layout(title='Modalidade Compra x Valor Licitação',
                   yaxis={'title':'Valor Licitação'},
                   xaxis={'title': 'Modalidade Compra'})

fig = go.Figure(data=data, layout=layout)

py.iplot(fig)
```

Trace é a variável que irá armazenar o objeto do gráfico, que será criado pela função go.Scatter. Nesta função iremos passar os argumentos para geração do gráfico, o eixo X sendo a Modalidade_Compra e eixo Y Valor_Licitação, ambos proveniente do dataFrame “Ano20”, que é a junção de todos os dataFrame. No argumento mode se passa a forma como os pontos do gráfico será representado e no argumento marker definimos a cor do gŕafico.

A variável data armazena trace na forma de uma lista, pois pra se exibir um gráfico no Ploty, é necessário passar como lista. Pode passar a variável trace dentro de py.iplot() mas é boa prática passar dentro da variável data.

A função go.Layout é como definiremos a exibição do gráfico, sendo title o titulo do gráfico, yaxis o titulo do eixo y e xaxis o titulo do eixo x.

A função go.Figure irá armazenar o objeto que será exibido como gráfico, como argumento se passa os dados que quer plotar e o layout que quer utilizar.

De acordo com o gŕafico verificamos que em todas as modalidades os valores não se dispersam, sendo bem próximos uns dos outros, ficando abaixo dos duzentos mil, com exceção de alguns valores que se dispersam dos demais.

Em média, a licitação tem um valor de trezentos e setenta e cinco mil, trezentos e vinte e três reais e trinta e cinco centavos (R$375.323,35), sendo a amplitude máxima de novecentos e quarenta e dois mil e oitocentos reais (R$942.800).

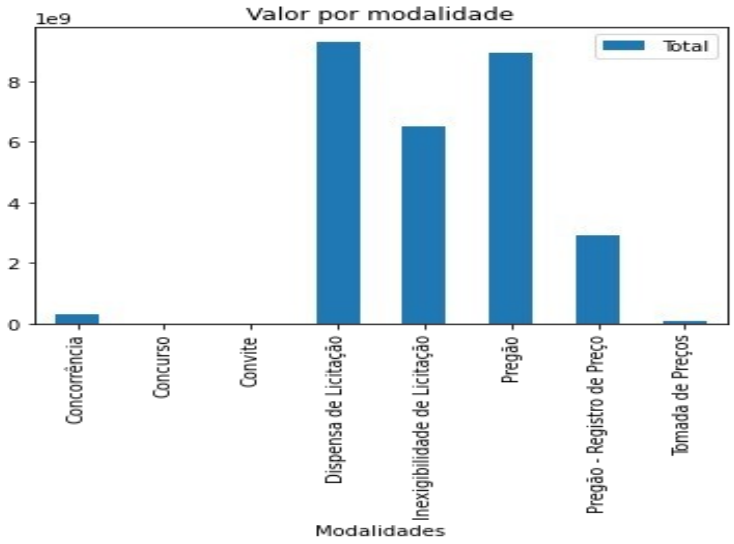
Sabemos até o momento que já foi gasto mais de vinte bilhões em licitações (R$ 28.022.391.871,82) e a modalidade mais frequente é a dispensa por licitação, vamos analisar qual o impacto dessa modalidade em cima desse valor.

As figuras 33 e 34 a seguir exibem o quanto foi gasto por modalidade

Figura 33 – Tabela valor por modalidade

Modalidade	Valor
Concorrência	R$ 281.950.472,15
Concurso	R$ 4.300,00
Convite	R$ 2.564.599,22
Dispensa de Licitação	R$ 9.320.541.233,05
Inexigibilidade de Licitação	R$ 6.491.257.010,23
Pregão	R$ 8.924.391.860,44
Pregão – Registro de preço	R$ 2.912.430.352,80
Tomada de Preços	R$ 89.252.043,93

Figura 34 – Valor por modalidade



Apesar da modalidade dispensa de licitação ter uma frequência 50% maior que a modalidade de pregão, podemos ver que os valores são bem próximos quanto a modalidade de pregão e até de Inexigibilidade de licitação. Um fator que pode explicar isso é que a dispensa por licitação é uma modalidade que um dos casos em que ela possa vir a ocorrer é em contratação de pequeno valor. Os valores desta modalidade tem uma média aproximadamente de cento e noventa e um mil reais (R$ 191.253,35), enquanto que as modalidades Pregão e Inexigibilidade de licitação têm em média aproximadamente, respectivamente, um milhão cento e vinte e quatro mil (R$ 1.124.120,40) e oitocentos e sessenta e quatro mil (R$ 864.348,47).

Até agora pudemos analisar que devido a pandemia o tipo de licitação que mais ocorreu foi dispensa de licitação, conforme o passar dos meses a quantidade de licitações nesta modalidade veio crescendo. Quanto ao valor das licitações, em todas as modalidades pode ser verificado que não há uma variância grande, e seus valores ficam em torno da média. Como a modalidade dispensa de licitação foi a que mais tendo uma frequência acima dos sessenta e cinco por cento, é interessante vermos a quem mais foi destinada às licitações.

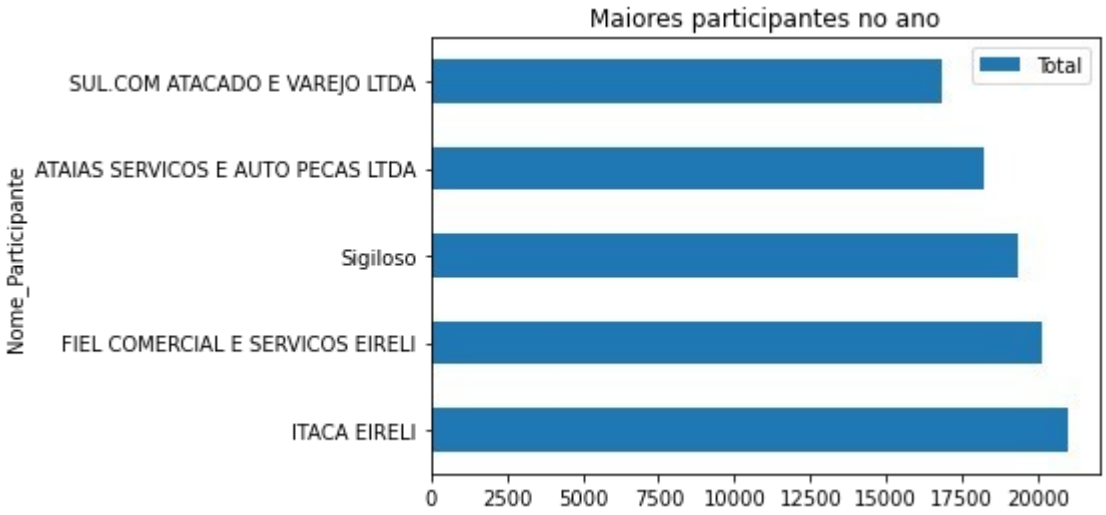
A tabela a seguir exibe os dez principais órgãos a qual foi destinado a modalidade de compra dispensa licitação.

Nome do Órgão Superior	Total de Processos
Ministério da Defesa	26540
Ministério da Educação	10954
Ministério da Economia	2647
Ministério da Saúde	2107
Ministério da Justiça e Segurança Pública	1607
Ministério da Agricultura, Pecuária e Abastec.	1532
Ministério da Ciência, Tecnologia, Inovações	1035
Ministério da Infraestrutura	383
Ministério do Turismo	331
Ministério de Minas e Energia	285

Para explicar o total de processos, devemos primeiro explicar uma coisa. Cada licitação pode ser referir a um órgão superior diferente e cada licitação pode conter, ser dividida, em diversos processos, e cada um destes processos, podem ser divididos em diversos itens, onde cada um destes representa o item/serviço que está sendo ofertado na licitação e ter seu respectivo valor. Por exemplo, uma determinada licitação pode ter dois processos, onde o processo A refere se ao Ministério da Defesa e possui dois itens, e cada um deles possui um determinado valor e o processo B refere se ao Ministério da Educação e possui apenas um item. Com isto explicado, retornando a tabela acima, podemos entender melhor ao que se refere o total de processos. Ou seja, dizer que o ministério da Educação teve dez mil novecentos e cinquenta e quatro (10954) licitações seria errado, mas sim 10954 processos pois, uma licitação pode conter diversos processos.

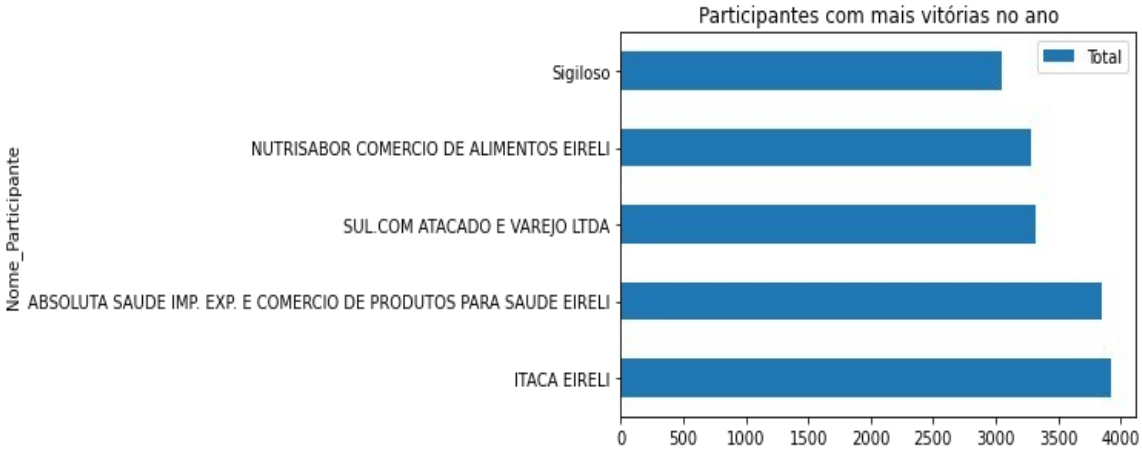
Cada um dos processos das licitações tem diversos participantes, logicamente, Conforme o número de licitações aumentou o número de participantes também se aumentou. O mesmo participante pode participar ,fazer propostas, em diversos processos. Neste ano teve se em média cerca de dezesseis mil quinhentos e setenta participantes (16570). A figura 35 apresenta os maiores participantes do ano.

Figura 35 – Maiores participantes no ano



Esses foram os maiores participantes, a figura 36 exibi os participantes com maior número de vitórias.

Figura 36 – Participantes com mais vitórias no ano



Observando os dois gráficos podemos ver que a ITACA EIRELI foi a que mais participou de processos de licitações e também a que mais venceu, mas a partir do 2 lugar já vemos alterações nos valores, como a ABSOLUTA SAUDE, que não aparece entre as 5 maiores participantes mas é a segunda que mais venceu em licitações.

5 Conclusão

Este trabalho teve por objetivo demonstrar as análises de licitações ocorridas durante o período de janeiro a agosto de 2020. Tendo como objetivo específico a análise e visualização dos dados.

Para uma compreensão melhor dos dados e do estudo de caso foi feita uma breve explicação sobre licitações, suas modalidades e suas regras. Bem como foi feita uma explicação sobre estatística, suas diferentes técnicas de análise numa população, e seus conceitos. Assim como foi feita uma explicação sobre a linguagem python e suas bibliotecas e ferramentas.

Com isso pudemos entender como python e análise estatística se relacionam e como a linguagem é utilizada e vem crescendo na área de mineração de dados.
No estudo de caso foi verificado um crescente número de licitações, e que mais de 60% é do tipo dispensa de licitação, e os valores são pouco dispersos, grande parte fazendo parte do conglomerado em torno da média.

Para estudo futuro pode se continuar a análise, aplicando técnicas de machine learning como aprendizado de máquina para uma análise preditiva, podendo se prever os futuros gastos com licitações, tão como uma aplicação de técnicas para detecção de casos suspeitos de fraudes em licitações.

REFERÊNCIAS BIBLIOGRÁFICAS

LEI Nº 8.666, Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm>. 1993

LEI Nº 10.520, Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/2002/l10520.htm#:~:text=LEI%20No%2010.520%2C%20DE%2017%20DE%20JULHO%20DE%202002.&text=Institui%2C%20no%20%C3%A2mbito%20da%20Uni%C3%A3o,comuns%2C%20e%20d%C3%A1%20outras%20provid%C3%Aancias>. 2002

LEI Nº 13.303, Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/lei/l13303.htm>. 2016

LEI COMPLEMENTAR Nº 123, Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp123.htm> 2006

Portal, Concilitação. **O que é licitação?** Disponível em: <https://portal.conlicitacao.com.br/o-que-e-licitacao/>

Open Data Handbook. O que são dados abertos? Disponível em: <https://opendatahandbook.org/guide/pt_BR/what-is-open-data/>

JusBrasil, **Modalidades de licitação** Disponível em: <https://triunfolegis.jusbrasil.com.br/artigos/403995892/modalidades-de-licitacao>

Zucco, Fabiano. **Entenda as noções básicas sobre licitações e contratos administrativos** Disponível em: <https://www.rcc.com.br/blog/modalidade-de-licitacao-2/> Rio de Janeiro, 2018

LEI Nº 12.527, Disponível em: <http://www.presidencia.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm> 2001

Portal brasileiro de dados abertos. **O que são dados abertos?** Disponível em: <https://dados.gov.br/pagina/dados-abertos#:~:text=Segundo%20a%20defini%C3%A7%C3%A3o%20da%20Open,sua%20proveni%C3%AAncia%20e%20sua%20abertura>

Mckinney, Wes. **Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy e Ipython**. Rio de Janeiro, 2018

Valéria Ferreira. **Estatística Básica**. Rio de Janeiro, 2015

Rafael Carvalho R. Oliveira. **Licitações e Contratos Administrativos: Teoria e prática.** 4 ed. 2015

Pedro A. Morettin. **Estatística básica.** 9 ed. 2017