Universidade de Brasília
Departamento de Estatística

Plano de Dados Abertos da Polícia Federal
Apreensão de Drogas

Thays Alves do Prado Cruz

Trabalho de conclusão de curso do primeiro semestre de 2021 pelo Departamento de Estatística da Universidade de Brasília.

Brasília
2021

Thays Alves do Prado Cruz

Plano de Dados Abertos da Polícia Federal
Apreensão de Drogas

Orientador: Leandro Tavares Correia
Coorientadora: Ana Maria Nogales Vasconcelos

Trabalho de conclusão de curso do primeiro semestre de 2021 pelo Departamento de Estatística da Universidade de Brasília.

Brasília
2021

Agradecimentos

Primeiramente agradeço à minha mãe Flavia por ter dado sempre o melhor de si para eu e meus irmãos termos a melhor educação apesar de todas as adversidades. Também agradeço por todo seu apoio durante minha vida acadêmica e não hesitar em proporcionar sempre o melhor para o meu desenvolvimento e nunca sair do meu lado mesmo nos meus piores momentos, incluindo os meus de problemas de saúde.

Agradeço à todos da Corregedoria-Geral de Polícia Federal, local onde estagiei durante meus últimos semestres na universidade. Em especial ao estatístico da Polícia Federal Raucelio Coelho Cardoch Valdes que esteve ao meu lado em todo período de minha montagem do meu trabalho de conclusão de curso e me ensinou grande parte do que realizei no trabalho. Também à Maria Aparecida dos Santos Moretti estatística da polícia federal que me auxiliou na compreensão das bases de dados utilizadas nesse trabalho. Ao meu supervisor e delegado de polícia federal Rafael Dall Agnol e o delegado de polícia federal Raphael Baggio de Luca que me insipiraram na escolha do tema do trabalho.

Ao professor Leandro Tavares Correia, meu orientador que sempre me auxiliou na montagem do trabalho. Por fim, a professora Ana Maria Nogales Vasconcelos que sempre foi uma inspiração para mim na minha vida acadêmica.

Resumo

O presente estudo é baseado no Plano de Dados Abertos da Polícia Federal tendo como dados de referência as bases de dados da Coordenação-Geral de Repressão a Drogas, Armas e Facções Criminosas (CGPRE/DICOR/PF) dos anos de 2019 e 2020. Para o estudo foi-se necessário inicialmente uma estruturação e limpeza dos bancos de dados para então a implementação das técnicas de *text mining*, a criação do algoritmo de verificação dos valores lançados, a análise descritiva e por fim, a criação do *dashboard*. A criação do algoritmo se deu em partes e foi necessário a criação de 5 funções, sendo todas essas criadas no *software* R. Os objetivos do estudo são através do algoritmo criado recuperar os valores das quantidades apreendidas através do texto cadastrado no sistema e assim, reduzir os erros que ocorrem no lançamento desses. O outro objetivo do estudo é a análise descritiva das quantidades apreendidas pela Polícia Federal para entender como ocorre o comportamento dessas apreensões. Por fim, o último objetivo é a criação de um *dashboard* auto informativo acerca das análises das apreensões de drogaque sirva de exemplo para a própria PF implementar em seus trabalhos futuros. Ao fim do algoritmo e com a implentação desse reduz em 99,31% o trabalho que os servidores da PF possuem de verificar em bases de dados muito grandes cada observação separadamente. Com isso, o tempo gasto com a verificação dos dados para depois a criação das análises estatísticas importantes para a PF seria muito menor possibilitando talvez até em estudos mais importantes. A outra contribuição do estudo são as análises temporais e regionais sobre as apreensões de maconha e cocaína pela PF nos anos de 2019 e 2020 para observar os comportamentos dessas apreensões (em números e quantidade apreendida em kg) para um possível estudo PF sobre essas análises.

Palavras-chaves: *text mining*, mineração dos dados, algoritmo, análise descritiva, apreensões, *dashboard*.

Lista de Tabelas

Lista de Figuras

Siglas

CGPRE/DICOR/PF Coordenação-Geral de Polícia de Repressão à Drogas, Armas e Facções Criminosas. 9, 11–13, 25, 36, 60

DICOR Diretoria de Combate ao Crime Organizado. 11

EI Extração da Informação. 15

FLA Flagrante. 13, 26

IPL Inquérito Policial. 13, 26, 28, 30, 60

LAI Lei de Acesso à informação. 8, 9

NC Notícia Crime. 13, 26

NCV Notícia Crime em Verificação. 13

PDA Plano de Dados Abertos. 8, 9, 13, 42

PF Polícia Federal. 9, 11–13, 25, 26, 29, 30, 36–38, 40–43, 46, 48, 51, 52, 57, 58, 60

RDF Registro de Fato. 13

RI Recuperação da Informação. 14

TC Termo Circunstânciado. 13

Sumário

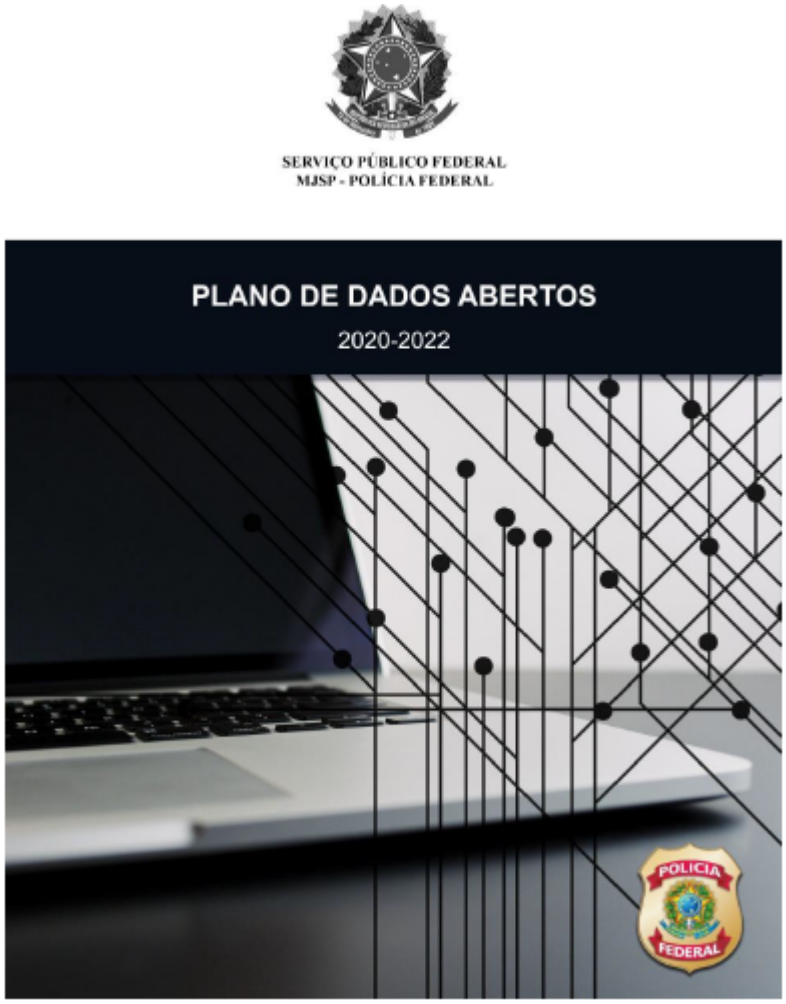

Figura 1: Plano de dados abertos Polícia Federal 2020 - 2022

Disponível em: https://www.gov.br/pf/pt-br/acesso-a-informacao/dados-abertos/pda-pf-2021-23

1 Introdução

O Plano de Dados Abertos da Polícia Federal (PDA) é um tema que vem merecendo muita atenção nos últimos anos devido ao empenho da instituição em atender às necessidades da sociedade no âmbito da transparência da informação, delimitado pela Lei nº 12.527 [1], de 18 de novembro de 2011, denominada Lei de Acesso à Informação (LAI), e pelo Decreto nº 7.724 [2], de 6 de maio de 2012, que a regulamentou. (BRASIL, 2011)

[1] https://www2.camara.leg.br/legin/fed/lei/2011/lei-12527-18-novembro-2011-611802-publicacaooriginal-134287-pl.html

[2] https://www.gov.br/ouvidorias/pt-br/ouvidorias/legislacao/decretos/decreto-no-7-724-de-16-de-maio-de-2012-1/view

A LAI é uma importante ferramenta para a democracia participativa, onde ocorre a aproximação do estado e sociedade, ampliando o interesse e acesso do cidadão às informações públicas. A disponibilização de informações públicas em atendimento a solicitações específicas de um interessado, e a divulgação de informações de interesse coletivo ou geral pelo setor público independente de requisição são as duas vertentes da LAI.

O PDA relaciona-se à Política de Dados Abertos e Espaciais no âmbito do Ministério da Justiça e Segurança Pública, instituída pela Portaria nº 1.378, de 20 de agosto de 2014 [3], a qual permite maior transparência e reutilização dos dados públicos pela sociedade.

Em 14 de janeiro de 2020, a Instrução Normativa nº 153-DG/PF [4] determina então, que o presente PDA se torne instrumento de planejamento e coordenação das ações de disponibilização de dados pela PF, onde todas as diretorias da Instituição se tornam responsáveis pela sua elaboração, monitoramento, atualização e avaliação seguindo as seguintes premissas: (BRASIL, 2020)

a. Desenvolvimento da cultura de transparência e da participação social no acompanhamento dos dados da PF;

b. Transparência do dado público, resultando em amplo acesso perante progressiva divulgação ativa;

c. Maior disponibilidade de informações cujo acesso não seja restringido por ato, legislação ou regramento específico;

d. Estímulo do uso de novas tecnologias para gestão e prestação de serviços públicos;

e. Utilização dos meios de comunicação dispostos pela tecnologia da informação;

f. Atualização periódica dos dados, garantindo a qualidade destes e seu valor para a sociedade.

Dentro do escopo do trabalho é desenvolvido o tema sobre Apreensão de Drogas, o qual trata as bases de dados disponibilizadas pela Coordenação-Geral de Polícia de Repressão a Drogas, Armas e Facções Criminosas (CGPRE/DICOR/PF), unidade da PF responsável por este tema.

[3] https://dspace.mj.gov.br/handle/1/309

[4] https://dspace.mj.gov.br/bitstream/1/2066/2/PRT$_S$E$_2$020$_1$297.*html*

Figura 2: Apreensão de Drogas MT

Disponível em: http://www.mt.gov.br/-/4794125-acao-conjunta-resulta-na-apreensao-de-300-quilos-de-droga-na-fronteira

A Apreensão de Drogas é um tema de grande relevância para a sociedade atual, pois apresenta dados que podem ser analisados de diversas formas e em inúmeros contextos, como por exemplo, análises temporais e regionais das apreensões.

O uso de drogas tem efeitos prejudiciais sobre a saúde de seus consumidores e apresenta reflexos negativos para a família e sociedade. Compromete o desenvolvimento, gera tensões sociais, desagrega a família e causa violência, pois estão relacionadas à origem dos altos níveis de criminalidade.

Além disso, tem também um enfoque político, pois o tráfico e o uso de drogas repercutem nas relações internacionais do país. Alguns países recebem críticas formais por não adotarem medidas eficazes de enfrentamento ao tráfico de drogas em seus territórios. Em outras vezes são adotadas até represálias, como pressões econômicas e até ameaças de intervenção contra o país.

E por fim, o tráfico de drogas tem um enfoque econômico para o país, pois gera uma economia clandestina e ilegal que provoca concorrência desleal e desorganização do sistema financeiro. Também provoca o aumento das despesas do Estado com os programas de saúde e recuperação de dependentes, compromete o sistema de segurança pública e fomenta a corrupção nas agências de controle e fiscalização do Estado.

As ações de combate ao tráfico ilícito de drogas exigem esforços concentrados e harmônicos das agências que constituem o sistema de repressão às atividades de organizações criminosas, operando em diversos locais e de diversas formas.

A PF tem buscado ferramentas adequadas para enfrentar o crescimento do tráfico de drogas no país tomando medidas preventivas e repressivas para combater o narcotráfico transnacional. Para isso, vem utilizando mecanismos de cooperação institucionalizados em convenções, acordos, convênios, entre outros.

Dentro da PF existem unidades que se dividem para execução de um trabalho de organização e planejamento nas suas áreas de atuação. Uma delas é a CGPRE/DICOR/PF.

Essa unidade, no âmbito da Diretoria de Combate ao Crime Organizado (DICOR), é responsável pela repressão ao tráfico de drogas e também pela administração do Canil Central onde são adestrados cães farejadores de drogas. Ela disponibiliza anualmente relatórios informativos a partir dos bancos de dados da PF.

Estes relatórios têm como objetivo mostrar um panorama do esforço brasileiro no combate de produção e tráfico de drogas. Sua organização compreende as seguintes partes: (CGPRE, 2021)

1. Introdução: discute-se a organização do devido relatório, como a repressão a drogas se insere na estrutura da DICOR, a importância da CGPRE/DICOR/PF e sua missão no âmbito do Sistema Nacional Antidrogas e da Polícia Nacional antidrogas;

2. Atividades de Repressão: mostra o marco legal em que estão inseridas as atividades de repressão no país, detalhando seus programas de trabalho e seus conceitos. Também aborda a questão da cooperação internacional e a adesão do Brasil às recomendações da *International Drug Enforcement Conference* (IDEC);

3. Conclusões: mostra o contexto regional e o contexto nacional das drogas e como afetam as tendências do fenômeno no Brasil. A análise dessas tendências inclui dados sobre produção, tráfico e consumo de drogas e também informações sobre as principais operações realizadas e os resultados alcançados. Metas para o próximo ano também são destacadas;

4. Estatísticas gerais: são apresentadas as estatísticas, um banco de imagens e um glossário com alguns termos utilizados na publicação. Essas informações servem como referência rápida sobre os resultados do esforço brasileiro de repressão à drogas nos últimos anos.

Para a elaboração de seus relatórios, primeiramente é necessário a organização das bases de dados que são dispostas para a CGPRE/DICOR/PF. Isso ocorre de forma manual e demanda bastante tempo dos servidores da PF responsáveis por esses serviços, pois esses bancos de dados ainda apresentam alguns erros na sua construção. Dentre esses erros podemos citar os mais importantes que são:

1. Dados faltantes, que como visto, atrapalham nos resultados e estatísticas finais;
2. Divergência entre as variáveis: como por exemplo, no banco de dados utilizado divergências entre as variáveis Material Observação e Quantidade Apreendida.

Nos últimos anos a PF vem buscado otimizar essas tarefas deixando-as mais rápidas e menos trabalhosas. A partir de agosto de 2019, a PF passou a contar com a primeira unidade a gerir e movimentar todos os seus inquéritos policiais por meio eletrônico. (BRASIL, 2019)

Com isso as apurações passam a ser criadas no sistema de inquéritos eletrônico da PF, o ePol, e enviados ao Judiciário ou Ministério Público também em meio digital. Com isso, evita-se a impressão de documentos e o translado físico de expedientes.

Além disso o ePol permite que os dados gerados pela PF se tornem mais visíveis e de mais fácil acesso para os gestores, para caso necessário, sejam feitas correções e mudanças em suas bases de dados. Inclusive, para essas correções já estão sendo estudadas técnicas para deixá-las mais automatizadas e assim, demandar menos tempo da PF na elaboração de boletins informativos.

Entretanto essas técnicas para automatizar as correções e melhorar os bancos de dados ainda não são reais. No estudo, um dos objetivos é a criação de algumas dessas técnicas com base no banco de dados disposto pela CGPRE/DICOR/PF.

Para isso alguns procedimentos já estão sendo implementados para o estudo, são eles:

1. Compreensão das variáveis: entender o que a variável significa no contexto da PF, compreender suas informações, como ela influencia ou é influenciada pelas demais variáveis e sua relevância no banco de dados. Variáveis que não tenham grande relevância para o estudo podem então, ser retiradas sem que acarrete problemas aos dados;
2. Restrição do período de estudo: foram disponibilizados pela CGPRE/DICOR/PF bancos de dados dos anos de 2014 até 2020. Entretanto, foi verificado que os bancos

de dados com informações mais corretas e melhor dispostas, foram os dos anos de 2019 e 2020. Assim, no estudo serão utilizados os dados apenas destes anos;

3. Restrição do tipo de droga: historicamente os tipos de droga mais apreendidos pela PF são maconha e cocaína e isso se verifica na base de dados utilizada. Assim, foram considerados apenas maconha e cocaína para o estudo;

4. Restrição do tipo de procedimento policial: para a polícia existem diferentes tipos de procedimento. Na base de dados verifica-se entre eles: Inquérito Policial (IPL), Notícia-Crime (NC), Flagrante (FLA), Notícia-Crime em Verificação (NCV), Registro de Fato (RDF) e Termo circustanciado (TC). Entretanto, o procedimento mais recorrente é o de IPL e por isso, foi o único considerado para o estudo.

O objetivo desse estudo é auxiliar a PF no aumento e melhora da qualidade dos seus dados, tanto os que serão publicados no PDA quanto os para análises internas, mostrando a eles os erros que ainda ocorrem em suas bases de dados, mas que esses podem sim ser resolvidos.

Esse aumento da qualidade dos dados possibilitará a eles pesquisas mais informativas para a sociedade e também para eles próprios. E para divulgação dessas informações e análises, o *dashboard* proposto no estudo pode ser útil. Para que isso se concretize são necessários que ocorram:

1. Limpeza e estruturação do banco de dados para aumentarmos a qualidade dos dados/estastísticas que são obtidos através desses dados;

2. Análise descritiva à cerca da base de dados de apreensões de drogas de 2019 e 2020 disponibilizada pela CGPRE/DICOR/PF;

3. Após a limpeza para diminuição ao máximo dos erros que podem ocorrer com os bancos de dados, uma divulgação maior e mais correta desses dados e estatísticas obtidas através do *dashboard*.

2 Referencial Teórico

2.1 Text Mining

A técnica de *Text Mining* (mineração de texto) é um processo em que textos não estruturados são transformados em importantes informações, sendo capaz de simplificar a análise de dados brutos em grande escala. Com isso, o trabalho que antes era manual, se torna muito mais rápido e também mais barato para quem utiliza.

Segundo Morais (2007) a área de *Text Mining* consiste em duas formas de mineração de textos:

1. Descoberta de Conhecimento em Textos (DCT): ajuda a explorar conhecimento armazenado em meios textuais, podendo ser definida como o processo de extrair padrões ou conhecimento, interessantes e não-triviais;
2. Descoberta do Conhecimento em Base de Dados (DCBD): consiste na mineração em base de dados estruturados que contenham textos.

Dentre os métodos mais comuns de *Text Mining* destacam-se a recuperação da informação, extração da informação e por fim, indexação automática.

2.1.1 Recuperação da Informação (RI)

"A RI tem como objetivo localizar os documentos que contém informações definidas pelo usuário em uma consulta. Para agilizar, utiliza-se a indexação, extraindo assim os termos mais significativos e excluindo os que não tem importância."(UBER, 2004)

De acordo com Uber (2004), a RI está preocupada com a organização e recuperação da informação em grande número de documentos textuais. O seu problema consiste em localizar documentos pertinentes com as palavras chaves que o usuário deseja encontrar. Para localizar essas informações utiliza-se a indexação.

A indexação é considerada um filtro capaz de selecionar e identificar as características de um documento, extraindo os seus termos mais significativos e excluindo aqueles que não são importantes. Segundo Yates apud Silva (2002), realiza-se de três formas:

1. Tradicional: os termos descritivos dos documentos são selecionados manualmente, especificando quais farão parte do índice;

2. *Full-text:* os termos que compõem o documento são usados como parte do índice;

3. *Tags:* os termos são selecionados automaticamente.

2.1.2 Extração da Informação (EI)

Dixon (1997) define a Extração da Informação (EI) como a identificação de itens (características, palavras), relevantes nos documentos, devendo ser extraídos e convertidos em dados.

Ainda de acordo com Uber (2004), a EI consiste nas principais técnicas:

1. Sumarização: abstração das partes mais importantes do conteúdo do texto e com isso, produzir um resumo do texto original;

2. *Clustering:* alocação de documentos que tenham assuntos similares e assim, criação de novos grupos para cada elemento distinto. É utilizado no processo de classificação, pois facilita a definição de classes;

3. Classificação ou categorização: também denominada de aprendizado supervisionado, pois a entrada e a saída desejadas são fornecidas previamente por um supervisor externo Fausett (1994). Suas principais técnicas são: estatística, aprendizagem de máquina simbólica e redes neurais.

2.1.3 Indexação Automática

Segundo Riloff, apud Loh, Wives e Oliveira (2000), a indexação automática consiste de quatro etapas:

1. Identificação de termos: identifica as palavras importantes do texto, ignorando símbolos e caracteres de controle de arquivo ou de formatação.,

2. Remoção de *stopwords*;

3. Normalização e padronização do vocabulário;

4. Seleção de termos relevantes: descoberta da importância das palavras em um texto, utilizando a frequência com que elas aparecem.

2.2 Expressões Regulares

Uma expressão regular é um método formal de se especificar um padrão de texto e com ela podemos lidar com situações de procura, substituição, validação de formatos e filtragem de informações. De acordo com Eis (2016), a expressão regular é apenas uma representação formada por símbolos onde cada símbolo representa um tipo de informação.

2.2.1 Clase de caracteres

Uma classe de caracteres permite procurar qualquer símbolo de um determinado conjunto de caracteres podendo encontrar vários caracteres ao mesmo tempo. Exemplos:

1. [a-z]: reconhece todas as letras minúsculas;
2. [A-Z]: reconhece todas as letras maiúsculas;
3. [A-z]: reconhece todas as letras maiúsculas e minúsculas;
4. [A-Z0-9]: reconhece todas as letras maiúsculas e números.

Existem também atalhos para as classes mais comuns que são:

1. \w: recupera todos caracteres alpha numericos, ou seja, letras e números, mas não acentos ou caracteres especiais. Equivale a [a-zA-Z_0-9];
2. \W: pega todos os caracteres que não seja alpha numericos, ou seja, pontuações e espaços;
3. \d: equivale a [0-9].

Por fim temos as classes de negação onde temos as expressões de classe. Exemplos:

1. [$\wedge y$]: reconhece qualquer caracter, exceto o y;
2. [$\wedge a - e$]: reconhece qualquer caracter, exceto a, b, c, d e;
3. [$\wedge \diagdown d$]: reconhece qualquer caracter, exceto 0, 1, 2, 3, 4, 5, 6, 7, 8 e 9.

Múltiplos padrões

Caso queira encontrar dois padrões diferentes de caracteres basta usar o símbolo | (pipe). Este símbolo fará a expressão regular reconhecer um ou outro padrão. Por exemplo: real |cor.

Âncoras

Servem para recuperar a posição entre os caracteres, mas não os caracteres em si. Por exemplo: a expressão real$ recupera as palavras real que estiverem no final da linha e a expressão ^real recupera as palavras real que estiverem no início da linha.

Modos

Servem para pegar uma sequência que contenha um termo parecido, mas que possa estar com algumas letras maiúsculas ou minúsculas. A representação (?i) deve estar antes do termo a ser buscado. Por exemplo: para as palavras Real, reAl, real basta apenas usar a expressão (?i)real.

2.3 Estatística Descritiva

Huot (2002) define estatística descritiva como o conjunto das técnicas e das regras que resumem a informação recolhida sobre uma amostra ou uma população, e isso sem distorção nem perda de informação.

A estatística descritiva é uma das áreas da estatística que aplica gráficos, tabulações e medidas descritivas para analisar, descrever e resumir um conjunto de dados, estudando o comportamento geral dos dados observados e assim, facilitar a resolução de problemas.

As análises descritivas geralmente são o primeiro passo de um estudo quantitativo. Ela consiste em descrever as principais tendências nos dados existentes e observar as situações que levam a outros fatos. Para isso são necessários a coleta de dados, organização desses, tabulação e a descrição dos resultados obtidos.

2.3.1 Interpretação dos dados

Na interpretação dos dados, de acordo com Morais (2010), deve-se produzir um resumo verbal ou numérico, ou então, usar métodos gráficos para descrição das principais características.

O método mais apropriado para isso depende da natureza dos dados, sendo eles:

1. Dados qualitativos: "Os dados qualitativos representam a informação que identifica alguma qualidade, categoria ou característica, não susceptível de medida, mas de classificação, assumindo várias modalidades."(MORAIS, 2010). Eles podem ser tanto nominais, como por exemplo, estado civil, quanto ordinais, por exemplo, desempenho do aluno.

2. Dados quantitativos: complementam o uso de dados qualitativos que usam descrições, adjetivos e elementos linguísticos para descrever objetos e imagens, pois representam a informação resultante de características susceptíveis de serem medidas.

   Também de acordo com Morais (2010), os dados quantitativos apresentam-se com diferentes intensidades, e podem ser discretos, por exemplo, quantidade de habitantes de um Estado, ou contínuos, por exemplo, peso e altura.

2.3.2 Organização e apresentação dos dados

A utilidade dos dados estatísticos depende, muitas vezes, de sua organização e apresentação. Essa apresentação é realizada principalmente através de tabelas, gráficos e distribuições de frequência.

Os gráficos são ferramentas de apresentação de dados tanto em relatórios, quanto em exposições para o público. Gráficos bem feitos conseguem apresentar uma grande quantidade de informação de uma forma mais simples e acessível.

É muito importante determinar qual o tipo de gráfico a ser usado no estudo e tomar cuidado na sua escolha, pois quando usado de forma incorreta ao invés de deixar as informações mais claras, acabam deixando-as mais confusas.

Uma vez escolhido o tipo apropriado, o gráfico ainda precisa ser projetado atentamente para que seja atingido o efeito desejado. Guedes et al. (2010) define que os principais gráficos, e mais utilizados, são:

1. Gráfico de colunas: compara frequências, especialmente quando se usam escalas ou hierarquias, e também, resultados observados com metas fixadas, mostrando o percentual obtido;

2. Gráfico de barras: possui as mesmas características do gráfico de colunas e pode ser utilizado em seu lugar. Quando os títulos são longos ou complexos, recomenda-se utilizá-los;

3. Gráfico de setores: também conhecido como gráfico de pizza, é próprio para apresentações simples de alguns valores. É utilizado para dados qualitativos nominais;

4. Gráfico de linhas: próprio para dados contínuos ao longo do tempo, especialmente em séries temporais ou outros conjuntos de dados que em gráficos de colunas ou de barras pareçam sobrecarregados;

5. Gráfico de dispersão: é um gráfico de duas dimensões que exibe como pares os resultados equivalentes de dois conjunto de dados. O acréscimo de uma regressão linear ao gráfico pode indicar a direção e a natureza do relacionamento entre os resultados.

2.3.3 Tabulação dos dados

Depois da coleta, a tabulação dos dados é a etapa mais importante antes da tomada de decisão. Ela consiste, basicamente, em organizar os dados obtidos de um estudo em uma só planilha ou *dataframe*, para facilitar o uso desses ao realizar análises comparativas, montar gráficos e etc.

Guedes et al. (2010) define que basicamente a tabulação pode ser:

1. Simples: quando ocorre a simples contagem do número de casos que ocorram em cada uma das variáveis analisadas;

2. Cruzada: quando os resultados estão relacionados com duas ou mais variáveis analisadas.

2.4 Testes de Hipóteses

2.4.1 Testes de Normalidade

Os testes de normalidade são usados para determinar se um conjunto de dados de uma dada variável aleatória segue uma distribuição normal.

A suposição de normalidade dos dados amostrais é uma condição exigida para a realização de muitas inferências válidas a respeito de parâmetros populacionais. Também muitos métodos de estimação e testes de hipóteses foram formulados sob a suposição de normalidade da amostra aleatória.

As hipóteses dos testes que testão normalidade são para ambos:

H_0) Os dados seguem uma distribuição Normal;

H_1) Os dados não seguem uma distribuição Normal.

Teste de Shapiro Wilk

De acordo com Canteli (2020), o teste de Shapiro Wilk é possivelmente o teste mais utilizado para verificar se um determinado conjunto de dados independentes segue a distribuição Normal e tem um maior poder de decisão do que outros testes. Também de acordo com Canteli (2020) o teste tem as seguintes vantagens e desvantagens:

Vantagens do teste:

1. É específico para a distribuição Normal;
2. Tem alto poder de decisão.

Desvantagens do teste:

1. É limitado a amostras com tamanho entre 3 e 50 observações;
2. É sensível a valores iguais ou muito próximos;
3. Para amostras pequenas dificilmente rejeita a hipótese nula de normalidade;
4. Quando os dados apresentam pequenas discrepâncias quanto a normalidade dos dados em amostras muito grandes o teste tende a rejeitar a hipótese nula de normalidade.

Estatística do teste:

$$W = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Onde:

y_i: variável aleatória observada;

a_i: coeficientes tabelados.

2.4.2 Teste de Kruskal Wallis

O teste de Kruskal Wallis é um teste não paramêtrico para a análise de variância de três ou mais amostras sendo a alternativa não paramétrica para o ANOVA. O teste pressupõe as seguintes condições para o seu uso adequado:

1. Comparação de três ou mais amostras independentes;
2. Não pode ser usado para testar diferenças numa única amostra de respondentes mensurados mais de uma vez;
3. Dados cujo nível de mensuração seja no mínimo ordinal;
4. Cada amostra deve ter tamanho mínimo de n = 6.

De acordo com Junior (2020), o mesmo é utilizado quando se deseja testar se várias amostras têm a mesma distribuição se baseando nos postos (*ranks*) das observações em cada grupo e tem como as seguintes hipóteses:

H_0) Os grupos têm a mesma distribuição de valores;

H_1) Os grupos não têm a mesma distribuição de valores.

Estatística do teste

$$H = [\frac{12}{N(N+1)}][\frac{\sum R_1^2}{n_1} + \frac{\sum R_2^2}{n_2}] + \frac{\sum R_3^2}{n_3}] - 3(N-1)$$

Onde:

N: número de dados em todos os grupos;

n: número de sujeitos em cada grupo;

$\sum R$: somatória dos postos (*ranks*) em cada grupo.

2.5 *Dashboards*

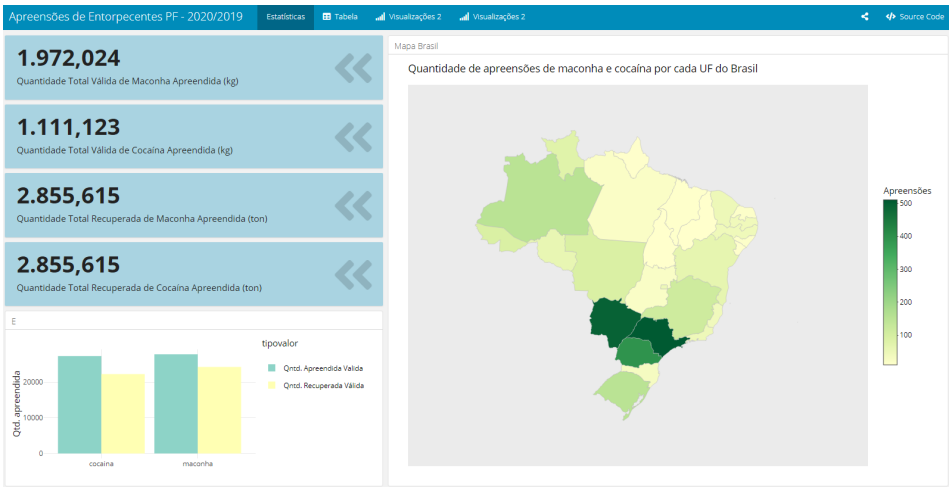

Figura 3: Exemplo de *dashboard*

Segundo Gomes (2017), *dashboard* é um painel visual que contêm informações, indicadores e métricas de um objeto de estudo e tem como objetivo auxiliar na tomada de decisões pois, é a forma mais eficiente de acompanhar múltiplas fontes de dados.

Isto acontece, pois fornecem em tempo real e em um único local, todas as informações necessárias, como por exemplo as estatísticas através de gráficos, números e tabulações para averiguar o desempenho de uma empresa ou órgão público. Eles são customisáveis de acordo com às demandas solicitadas, mas para que funcione de forma efetiva deve estar conectado aos servidores da empresa e/ou instituição.

Também de acordo com Gomes (2017), o principal objetivo para a contrução de um bom *dashboard* é que ele responda às perguntas dos negócios que precisam de respostas.

Eles são desenvolvidos para análises rápidas e atenção a informações importantes. As perguntas que o *dashboard* responderá depende da indústria, departamento, processo ou negócio que demanda a sua construção.

Todos os diversos tipos de dashboards têm como objetivo principal facilitar o acompanhamento eficiente das operações de uma empresa e/ou orgão.

Por fim, de acordo com Batista (2018), entre os diversos tipos, os principais são: *dashboard* analítico, *dashboard* operacional, *dashboard* tático e *dashboard* estratégico. As definições à seguir são baseadas nas definições deste.

2.5.1 *Dashboard* analítico

O *dashboard* analítico é usado para ajudar a identificar tendências e padrões de comportamento. Por meio de informações detalhadas, ele ajuda a avaliar se processos ou projetos estão evoluindo de acordo com o esperado.

Por exemplo, uma redução do quadro de funcionários de uma empresa e suas consequências.

2.5.2 *Dashboard* operacional

O *dashboard* operacional é utilizado por equipes que realmente participam das operações. O painel auxilia no acompanhamento do fluxo de trabalho e, assim, orienta as melhores decisões para momentos distintos.

Por exemplo, um vendedor precisa constantemente de informações à respeito dos produtos para venda em estoque.

2.5.3 *Dashboard* tático

O *dashboard* tático ajuda no planejamento e execução relacionados aos recursos. Por exemplo, o dono de uma loja precisa saber qual ou quais produtos estão sendo mais vendidos e quais estão sendo menos vendidos, para então, tomar providências.

2.5.4 *Dashboard* estratégico

O *dashboard* estratégico mostra os dados mais gerais dos negócios, instituições, etc. Com esses dados, é possível fazer comparações do cenário atual com os de outros

tempos. Por exemplo, uma análise do crescimento de uma empresa desde sua criação.

O principal motivo para utilizar um *dashboard* é conseguir acessar, interagir e analisar dados atualizados em tempo real para poder tomar melhores decisões. Assim, os principais benefícios da utilização deles são:

1. Aumento da eficiência com acesso rápido à *BI (Business Intelligence)*;
2. Melhora do processo de tomada de decisão;
3. Melhora do alinhamento entre as áreas da empresa, órgão ou instituição;
4. Dados mais visíveis.

3 Metodologia

A PF disponibiliza para a população dados e estatísticas que envolvem todas suas diretrizes e diversos temas quando essas não são sigilosas. Para a construção dessas estatísticas é necessário um trabalho de seus servidores nos bancos de dados que são dispostos pela PF através de seus sistemas internos.

Para o presente estudo, os bancos de dados disponibilizados foram dos dados sobre Apreensão de Drogas dos anos de 2019 e 2020. Estes bancos de dados disponibilizados pela CGPRE/DICOR/PF dizem respeito às apreensões de drogas de 2019 à 2020.

Esse estudo é constituído de uma pesquisa científica quantitativa centrada nas questões à respeito das apreensões de drogas cujo banco de dados é disponibilizado pela PF. Essas análises são muito importantes para verificação da qualidade dos dados que são dispostos por eles para sociedade e possíveis melhoras, caso necessário.

Isso foi feito através da junção das base de dados ofertada pela CGPRE/DICOR/PF sobre apreensão de drogas e seguindo os seguintes passos:

1. Estruturação e limpeza dos erros do banco de dados;
2. Implementação das técnicas de *text mining* e algoritmos de especificação para refinamento, melhora na qualidade dos dados e diminuição de erros em suas elaborações;
3. Análise descritiva das variáveis e seus respectivos cruzamentos, verificando como são influenciadas e influenciam entre si;
4. Uso de técnicas estatísticas como por exemplo, análises temporais, para obtenção de resultados que auxiliarão a PF em seus estudos;
5. Criação de dashboards para auxiliar a PF na divulgação de suas análises, estudos e pesquisas.

3.1 Estruturação e limpeza dos erros do banco de dados

3.1.1 Escolha das variáveis

As bases de dados dos anos de 2019 e 2020 apresentam estruturações diferentes uma da outra. Enquanto a base de dados do ano de 2019 apresenta 10 variáveis, a do ano de 2020 apresenta 34 variáveis.

Então, para a junção das bases de dados em uma única foi necessário a retirada de algumas variáveis da base de dados de 2020 para se igualar a de 2019 e conseguirmos realizar o estudo.

Para realizar o estudo, as variáveis da base de dados de 2020 que foram consideradas foram: UF, GestãoBens Item Data Apreensão, Mês, GestãoBens Item Material, GestãoBens Item Quantidade Apreendida, Valores lançados, GestãoBens Item Material Observação e GestãoBens Item Unidade Material. As restantes foram todas desconsideradas por serem sigilosas ou por não terem importância para o estudo.

3.1.2 Tratativa do banco de dados

As seguintes etapas realizadas para tratativa dos dados foram todas realizadas utilizando o *software* estatístico R.

Uma variável apresentava um dado importante para o estudo, o tipo de procedimento (IPL, NC, FLA e etc) mas com ela outras informações sigilosas. Então, por ser considerada sigilosa, foi necessário a retirada apenas da informação o tipo de procedimento desconsiderando as outras informações presentes e a criação de uma nova variável. Essa variável criada tem o nome de Tipo Procedimento.

Na primeira etapa para a tratativa dos dados foi necessário igualar o nome das variáveis de ambas as bases de dados. Assim, os nomes considerados para todas as variáveis dos bancos de dados foram respectivamente: UF, Data Apreensão, Mês, Material, Quantidade Apreendida, Valores lançados, Observação, Unidade Material e Tipo Procedimento.

Assim, depois dessa tratativa, o estudo parte de um cruzamento dos dados das seguintes variáveis as quais servirão para devida análise de correlações entre si:

1. UF: variável qualitativa que descreve a Unidade Federativa da apreensão;
2. Data Apreensão: variável qualitativa que descreve a data da apreensão;
3. Mês: variável qualitativa que descreve o mês da apreensão;
4. Material: variável qualitativa que descreve o tipo de droga apreendido pela PF;
5. Quantidade Apreendida: variável quantitativa a qual apresenta a quantidade de droga apreendida (em gramas, unidade, comprimido, miligramas ou líquido);

6. Valores lançados: variável quantitativa que representa o valor cadastrado no sistema pelo escrivão sobre a quantidade apreendida descrita na variável Material Observação;

7. Observação: corresponde ao texto cadastrado no sistema pelo escrivão sobre a apreensão;

8. Unidade Material: variável qualitativa que representa a unidade da droga apreendida que foi cadastrada;

9. Tipo Procedimento: variável qualitativa que representa o tipo de procedimento policial realizado na apreensão.

Na base de dados a mesma droga é referenciada de várias formas. No banco de dados foram detectadas 65 formas diferentes para referenciar uma mesma droga, entre elas letras em maiúsculo, todas as letras em minúsculo entre outras várias formas. A segunda etapa de tratativa dos dados foi realizada para solucionar esse problema e foram necessários os seguintes passos:

1. Uniformização das formas de escrita para onde todas as letras estão em forma minúscula;

2. Retirada das indicações de unidade na descrição do tipo de droga, sendo elas: (GR), (UN), (LI), (CO). Por exemplo: maconha (GR);

3. Retirada dos termos "comprimidos"e "líquidos"da descrição da droga;

4. Uniformização da escrita de algumas droga como por exemplo: "skank princípio da maconha"para apenas "skank", "ecstasi"para "esctasy"e "mudas de maconha"para "pés de maconha", entre outras maneiras.

Por fim, na variável Tipo Procedimento, da mesma forma que ocorre com os tipos de drogas apreendidos, o mesmo procedimento estar escrito de formas diferentes. Assim, a terceira e última etapa da tratativa dos dados foi a uniformização desses procedimentos, retirando alguns espaçamentos que ocorrem e por isso os diferenciam.

3.1.3 Filtragem dos dados a serem usados no estudo

Figura 4: Nuvem de palavras da base de dados de apreensões de drogas 2019 e 2020 acerca do texto cadastrado da apreensão

Ao início do estudo, após a junção das bases de dados e uniformização de todas as formas de escrita dos textos, possuía-se 12930 observações. Entretanto, percebe-se como demonstrado na Figura 4 quais as informações que mais se repetem na base de dados e que seriam mais viáveis de trabalhar. Dessa forma, foram aplicados filtros para o trabalho nessa base de dados.

O primeiro filtro foi para a variável 'tipo procedimento' filtrando apenas IPL e então reduziu-se para 8189 dados.

O segundo filtro aplicado foi para a variável 'ano', uma variável criada à partir da variável 'Data Apreensão', para apenas os anos 2019 e 2020 (pois mesmo as bases de dados sendo apenas desses anos, alguns equívocos de outros anos registrados ocorreram). Assim, restaram 8179 dados.

O terceiro filtro aplicado foi para a variável 'Material', filtrando apenas os tipos de drogas maconha e cocaína, restando 5403 dados.

O quarto e último filtro aplicado foi para a variável, também criada, chamada 'Alteração'. Essa variável foi criada após uma primeira análise da base de dados, feita de forma manual para identificar se havia alguma alteração ou não nos valores registrados pela PF e o texto descrito. Possui três categorias: 'sim', 'não' e 'inconclusivo'. Como inconclusivo foi considerado todo dado onde não havia nenhuma forma de confirmar algum erro ou não no lançamento do valor, pois nenhum valor foi registrado no texto.

Esses dados registrados como inconclusivos foram também retirados para a implementação do algoritmo e com isso, restaram apenas 3363 dados para análise da nossa base de dados inicial de 12930.

Em suma, foram realizadas 4 etapas na tratativa e essas são representadas na Tabela 1.

Tabela 1: Etapas da trativa dos dados

Etapa	Descrição
1	Uniformização dos nomes das variáveis
2	Uniformização dos nomes dos tipos de drogas;
3	Uniformização dos nomes dos tipos de procedimentos;
4	Filtragem dos dados da base.

3.2 *Text mining* e algoritmos de especificação

A estruturação dos bancos de dados disponibilizados pela PF para seus servidores realizarem suas análises e transpor elas à população ainda necessita de muitos ajustes e melhoras. Para isso, a criação de algoritmos e uso das técnicas de *text mining* são implementadas no estudo para melhora da qualidade dos dados dispostos de forma automática nos sistemas da PF. Essas técnicas e o algoritmo criado serão sugeridas para a PF implementar em todas suas bases de dados para a melhora de todas elas.

Dentre os diversos erros que ocorrem em suas elaborações podemos citar como por exemplo, valores divergentes da variável Quantidade apreendida com relação às variáveis Valores lançados e Observação.

Quantidade Apreendida	Valores lançados	Material Observação	Unidade Material
20.000.000,00	20,00	Aproximadamente mais de 20 toneladas de substância análoga a MACONHA (de acordo com a estimativa do policial federal condutor do flagrante, mais de 20 toneladas de substância análoga à maconha, cujo volume exato deverá ser quantificado quando for ocorrer a sua incineração)	GR
	20,00	Aproximadamente mais de 20 toneladas de substância análoga a MACONHA (de acordo com a estimativa do policial federal condutor do flagrante, mais de 20 toneladas de substância análoga à maconha, cujo volume exato deverá ser quantificado quando for ocorrer a sua incineração)	GR
703.950,00		650 tabletes, envoltos em fita adesiva, de substância em pó, de coloração branca, popularmente denominda cocaína.	GR
670.000,00		Uma certa quantidade de substância entorpecente aparentando ser MACONHA, com peso bruto aproximado de 657200 gr, todos devidamente lacrados em 18 (dezoito) sacos sob os números: 45306, 45303, 45354, 45332, 45380, 45330, 45304, 45319, 45377, 45397, 45320, 45301, 45363, 45351, 45355, 45369, 45311 e 45307;	GR

Figura 5: Exemplo de erros do banco de dados de apreensões de drogas da PF dos anos de 2019 e 2020

De acordo com a análise inicial feita na base de dados observamos que 47% das observações correspondem à maconha e cocaína e que o procedimento que mais ocorre nas apreensões é o procedimento de IPL. Assim, para o trabalho e a montagem do algoritmo em questão foram considerados apenas essas observações da base de dados.

Então, o uso de *text mining* no trabalho acontece para conseguirmos "minerar"o banco de dados, principalmente a variável Observação, que como visto anteriormente é uma das variáveis que mais influencia nas demais. Assim, o trabalho da montagem do banco de dados será mais rápida, fácil e com menos erros.

Após a junção das bases de dados estruturando-as e limpando-as parcialmente deu-se início a mineração dos dados para uma maior limpeza dos dados. Para isso foi utilizado o *software* R e dentre outros, o seu principal pacote utilizado foi o pacote *stringr*.

"O pacote *stringr* integra uma coleção de pacotes projetados para a ciência de dados, o *tidyverse*. Combinado ao pacote *stringi*, você terá acesso a praticamente

todas as possíveis funções necessárias para o processamento de *strings* em mais alto nível."(MOREIRA, 2020)

De acordo com Moreira (2020) existem quatro famílias principais de funções nesse pacote que são:

1. Manipulação de caracteres: permitem manipular caracteres indivuais dentro de sequência de caracteres;
2. Ferramentas de espaço em branco que permitem adicionar, remover e manipular espaços;
3. Operações sensíveis à localização geográfica que variam para cada local;
4. Funções de correspondência de valores onde o mais comum são as expressões regulares.

O processo de mineração dos dados e a montagem do algoritmo ocorreram em partes sendo todas elas muito importantes para o resultado final. Todas essas partes do processo estão descritas na Tabela 2 e foram realizadas na variável Observação.

Tabela 2: Etapas do processo de text mining no estudo

Etapa	Descrição
1	Unificação dos textos
2	Reconhecimento da quantidade apreendida
3	Lançamento correto dos valores apreendidos

3.2.1 Unificação dos textos

O processo de unificação dos textos é necessário para o estudo, pois o valor apreendido expresso no texto cadastrado pode acontecer de diferentes formas.

Então, esse processo de unificação dos textos da variável ocorreu em algumas partes as quais são:

1. Transformação de todos os caracteres em minúsculos;
2. Retirada de todos os acentos das palavras e mudança da letra 'ç' para apenas c;
3. Retirada de todos os excessos de espaços;
4. Retirada dos sinais -, ", ' e \;
5. Unificação dos modelos das unidades de medida para as seguintes: kg, gr, ton, co ou un.

```
library(dplyr); library(lubridate); library(tidytext)
library(stopwords); library(stringi); library(stringr)

base <- mutate(dados, anoApreensao = year(dataApreensao),
               observacaoPadrao = str_to_lower(observacao),
               observacaoPadrao = str_squish(observacaoPadrao),
               # padroniza kg
               observacaoPadrao = str_replace(observacaoPadrao,'(\\b(quil).
                   *|\\b(qu)[ul].*)', 'kg'),
               observacaoPadrao = str_replace(observacaoPadrao, 'kgs |kkg
                   |\\bk\\b|kilograma|kgsubst?ncia|kilos?', 'kg'),
               observacaoPadrao = str_replace_all(observacaoPadrao, '
                   ([0-9])\\s?kg\\b', '\\1 kg'),
               # padroniza ton
               observacaoPadrao = str_replace(observacaoPadrao, '\\b(ton).*
                   ', 'ton'),
               # padroniza gr
               observacaoPadrao = str_replace_all(observacaoPadrao, '\\b(
                   gram).*', 'gr'),
               observacaoPadrao = str_replace_all(observacaoPadrao, '
                   ([0-9])g\\b|([0-9])gramas\\b|([0-9])gr\\b', '\\1 gr'),
               observacaoPadrao = str_replace_all(observacaoPadrao, '(\\bg
                   \\b)', 'gr'))%>%
       filter(materialApreendido %in% c('maconha', 'cocaina'),
              tipoProcedimento == "IPL")
```

3.2.2 Algoritmo final

Para o reconhecimento da quantidade apreendida correta escrita na variável observação foi necessário a abertura e quebra do texto da variável Observação onde apenas as informações importantes foram consideradas. Essas informações são: valor apreendido e unidade de medida. Isso foi feito através das funções criadas chamada apreensaoUnidade e apreensaoValor.

Todas as funções criadas à seguir foram criadas através do *software* R.

```
apreensaoUnidade <- function(x)
{
  unidade <-  ifelse (str_detect (x,'ton|gr|kg' ),
                      str_extract(x,'ton|gr|kg' ), NA)
}

apreensaoValor <- function(x)
{
  valor <- str_remove_all ( x, '[[:alpha:].())]\\s*')
  valor <- str_replace_all(valor , ',', '\\.')
  valor <- as.numeric(valor)
  return(valor)
}
```

Após a extração dos valores e unidades de medida do texto foi criada então uma função para que transformasse todos os valores para a unidade de medida gramas. Para isso foi criada a função valorGrama que quando a unidade do número for identificada como 'kg' irá multiplicar esse valor por 1000, e quando for identificado como 'ton' será multiplicado por 1000000.

```
valorGrama <- function(valor , unidade)
{

  valorGrama <- ifelse(unidade ==  'gr' , valor ,
                       ifelse(unidade ==  'kg' , valor * (10^3),
                              ifelse(unidade ==  'ton', valor * (10^6) , NA
                                  )))
return(valorGrama)
}
```

Por fim, a etapa mais complexa da criação desse algoritmo foi a leitura dos números escritos de forma extensa na variável Observação. Para isso foi necessário primeiramente a adequação desses valores de quantidade apreendida escritos por extenso banco de dados. Foi necessária a leitura desses números escritos por extenso e então sua transformação para número para então a extração do valor. Essa etapa foi realizada através de funções criadas denominadas "extrai_numero"e "extrai_valor".

```
extrai_numero <- function(texto){

tokens <- str_split(texto, '\\s')

numero <- vector ("character", length = length(tokens))
aux <- vector ("character", length = length(tokens))

for(a in 1:length(tokens))
{

cont = 1
aux = NA

for ( i in tokens[[a]])
{
  for (j in c(numeroExtenso,numeroComplemento))
  {

    if (i == j) {
      aux [cont] <- j
    }
  }
  cont = cont + 1
}

aux <- na.omit(aux)
attributes(aux) <- NULL

numero[a] <- paste(aux, collapse = ' ')
}
return(numero)
}
```

Para criação dessa função foram criadas as funções numeroExtenso, que retirou do texto a forma como o número estava escrito por extenso, e numeroComplemento que retirou as unidades de medida. Foi possível essa retirada do número correspondente à quantidade apreendida e dos complementos pois essas funções contêm a forma usual de escrita do número por extenso na linguagem português brasileiro, retirada de um documento disponibilizado pelo estatístico de PF, Raucelio Valdes, para a realização do trabalho. Neste documento temos representações dos números escritos por extenso como por exemplo: "cento e vinte mil", onde "cento"e "vinte"são retirados pela função numeroExtenso e "e"e "mil"pela função numeroComplemento.

Ao fim da criação dessas funções foi então possível a implementação do algoritmo final desejado para o estudo. Com a criação desse algoritmo os erros de lançamento de valores no sistema das bases de dados da PF irá reduzir significativamente, pois não será necessário para o escrivão registrar o valor da quantidade apreendida, pois esse valor irá ser automaticamente retirado do texto escrito por ele na variável Observação.

Através da base de dados usada no estudo foi percebido que através do algoritmo mais de 80% dos valores recuperados coincidem com os valores corrigidos e lançados manualmente pelos servidores da CGPRE/DICOR/PF e isso representa um grande avanço na recuperação desses dados.

Temos então na Tabela 3 o resumo das funções criadas ao longo da criação do algoritmo.

Tabela 3: Funções criadas para implementação do algoritmo

Fase	Nome função
1	apreensaoUnidade
2	valorApreensao
3	valorGrama
4.1	extrai_numero
4.2	extrai_valor

3.3 Validação do estudo

Para a validaçao dos dados do estudo a análise de correlação será importante para conseguirmos estudar, analisar e verificar o quanto o algoritmo está sendo eficaz na recuperação dos dados.

De acordo com Peternelli (2004) a correlação serve para estudar o comportamento conjunto de duas variáveis quantitativas destintas, medindo a associação entre duas variáveis aleatórias.

Portanto, a análise de correlação no estudo será para avaliar se há uma correlação entre as variáveis quantidade apreendida (quantidade apreendida lançada já corrigida pela PF) e quantidade recuperada, que é a quantidade apreendida que o algoritmo conseguiu recuperar.

3.3.1 Erro absoluto e relativo

> "O erro absoluto é definido como o valor absoluto da diferença entre o valor medido e o valor real de uma medição e geralmente é dado como o erro máximo possível, dado o grau de precisão de uma ferramenta de medição. O erro absoluto tem as mesmas unidades da medida. O erro relativo é definido como o erro absoluto em relação ao tamanho da medição e depende tanto do erro absoluto quanto do valor medido. O erro relativo é grande quando o valor medido é pequeno ou quando o erro absoluto é grande. O erro relativo não possui unidades."(RODRIGO, 2020)

Para a validação do estudo será considerado o erro absoluto que será dado por:

$$erro\% = \frac{ValorObservado - ValorRecuperado}{ValorObservado} \times 100$$

Onde:

- ValorObservado: valor da quantidade apreendida (em kg) já validada pela PF;

- ValorRecuperado: valor recuperado da quantidade apreendida pelo algoritmo (em kg).

Para a análise de correlação no estudo foram pegos apenas os erros absolutos em relação ao dado observado que sejam menores ou iguais à 99%, pois foi observado que 81% dos dados não possuem erro na recuperação do algoritmo (erro absoluto = 0%), e 93% da base está nesse intervalo (<= 99%).

Um erro relativo com valor igual à 99% significa dizer que o valor recuperado pelo algoritmo é aproximadamente 2 vezes maior que o valor registrado no sistema pela PF.

Na Figura 6 representamos melhor o que foi descrito acima, onde o eixo X representa o percentual da base de dados com determinado erro e Y o valor desse erro absoluto em percentual.

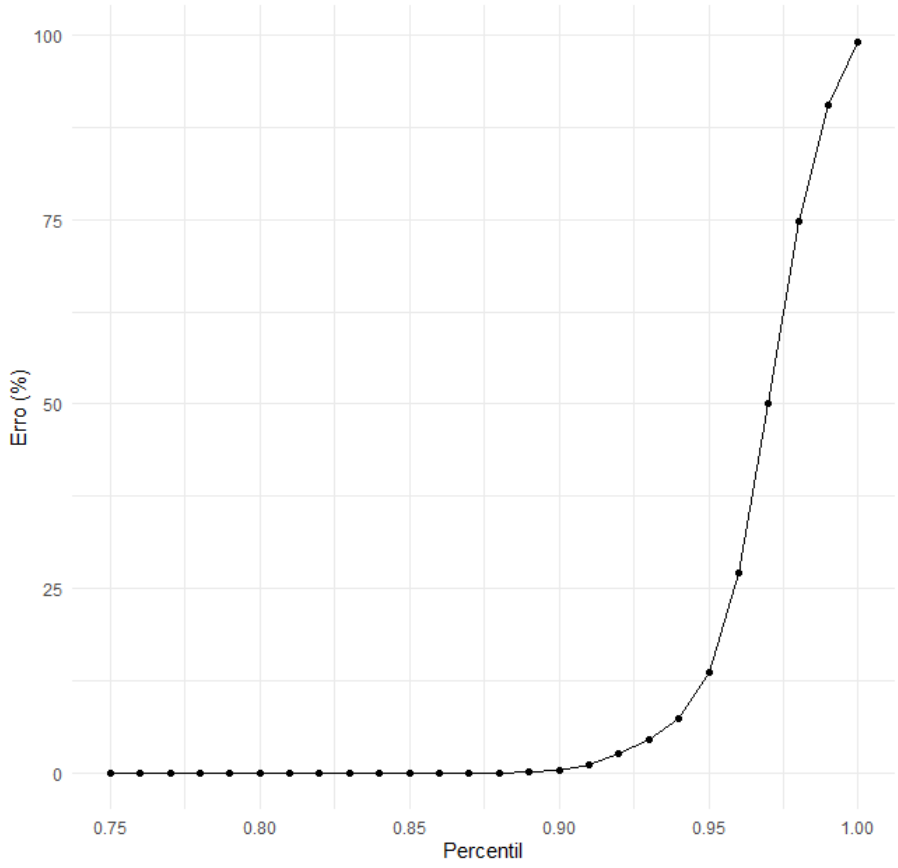

Figura 6: Distribuição dos erros absolutos do algoritmo ao recuperar o valor da quantidade apreendida de drogas (em kg) na base de dados da PF

3.3.2 Teste de correlação

Foi realizado o teste de correlação para confirmar se existe ou não correlação entre as variáveis. Para o teste foram consideradas as seguintes hipóteses:

H_0) Não existe correlação entre as variáveis do estudo ($\rho = 0$);

H_1) Exite correlação entre as variáveis do estudo ($\rho \neq 0$).

Considerando (X,Y) com distribuição normal bidimensional, temos que o coeficiente de correlação linear de Pearson (ρ) é dado por:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Considerando uma amostra aleatória de tamanho n o coeficiente amostral é dado por:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2(Y_i - \bar{Y})^2}}$$

Sendo:

X_i e Y_i as i-ésimas observações da amostra, $i = 1, ..., n$.

$\bar{X}$ e $\bar{Y}$ suas respectivas médias.

A estatística do teste utilizada para essa análise é dada por:

$$t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \sim \ T_{n-2}$$

Onde t tem distribuição t de Student com n-2 graus de liberdade, sendo esse o teste usado para testar a significância de coeficientes de regressão, confirmando ou não se a variável que está sendo usada está realmente contribuindo para a estimativa.

Utilizou-se o coeficiente de correlação de Pearson pois ele avalia a relação linear entre duas variáveis contínuas ou ordinais, o que se adequa ao estudo.

Para essa análise foram usados gráficos de dispersão para analisar essas duas variáveis e também para comprovarmos se o coeficiente de correlação é significativo, usou-se a estatística do teste através do coeficiente de correlação de Pearson.

3.4 Análise dos dados

Após a mineração dos dados e validação desses, as análises descritivas servem para compreendermos e analisarmoss o comportamento dessas apreensões de forma temporal e regional através de gráficos e medidas estatísticas importantes. As análises feitas nesse trabalho são todas baseadas nos valores disponibilizados pela PF da quantidade apreendida já validadas por seus servidores de forma manual.

Um dos maiores problemas enfrentados atualmente pela PF para construção de dados confiáveis é a qualidade desses e como são dispostos. Usualmente ocorrem erros no lançamento dos valores de suas apreensões, principalmente nas de drogas, e são esses erros que atrapalham a qualidade das estatísticas dispostas pela PF.

Historicamente, maconha e cocaína sempre foram as drogas mais apreendidas no Brasil. Na base de dados usada para o trabalho aproximadamente 42% das apreensões dizem respeito às essas drogas sendo 3247 cocaína e 2163 maconha.

Para essas duas drogas a unidade de medida da quantidade apreendida usual adotada pela PF é a de valores em gramas, no entanto no lançamento desses valores ainda ocorrem algumas divergências quando o servidor dispõe esse dado no sistema.

Apreensões de toneladas de maconha e/ou cocaína são comuns pela PF, apesar de não serem muito recorrentes. Uma apreensão de tonelada quando registrada no sistema, dado que a unidade usual de medida é gramas, teria o valor descrito no texto multiplicado por 10^5, mas isso muitas vezes não acontece.

As apreensões de maconha e cocaína realizadas pela PF divergem bastante, pois as apreensões de maconha em kg são maiores do que as de cocaína. Verifica-se que as apreensões de maconha são maiores em determinadas UFs e as de cocaína em outros. Também verifica-se que para ambas as drogas determinados meses dos anos de 2019 e 2020 ocorrem mais apreensões do que em outros.

Então, será utilizado no trabalho o teste de Kruskal Wallis para analisarmos esses comportamentos dessas apreensões tanto pelas UFs quanto pelos meses dos anos individualmente.

3.4.1 Teste de Shapiro Wilk

O teste de Shapiro Wilk no estudo é utilizado para testar a normalidade da quantidade apreendida de maconha e cocaína. As hipóteses do teste utilizadas no trabalho são:

H_0) As quantidades apreendidas de maconha ou cocaína seguem uma distribuição normal;

H_1) As quantidades apreendidas de maconha ou cocaína não seguem uma distribuição normal.

3.4.2 Teste de Kruskal Wallis

O teste de Kruskal Wallis é o teste de análise de variância não paramétrico utilizado no estudo para comparar as amostras de maconha e cocaína por suas UFs e pelos meses dos anos 2019 e 2020. Esse teste é utilizado devido a não normalidade dos dados do estudo.

As hipóteses do teste utilizadas no trabalho para a análise geográfica por UF de cada tipo de droga apreendida serão:

H_0) As quantidades apreendidas em kg em todas as UFs são semelhantes;

H_1) Em ao menos uma das UFs a quantidade apreendida em kg difere das demais.

E para a análise temporal as hipóteses do teste utilizadas no trabalho serão:

H_0) O número de apreensões em todos os meses dos anos são semelhantes;

H_1) Em ao menos um mês dos anos o número de apreensões difere dos demais.

Em ambas as análises o teste será feito para cada tipo de droga apreendido separadamente e será considerado um $\alpha = 0{,}05$ para ambas.

3.5 *Dashboards*

Os dados presentes no estudo são dados restritos aos servidores da PF, mas as estatísticas e análises feitas a partir deles são públicas e de amplo acesso à população, o

que segue a política do PDA. Isso permite ao servidor então, a verificação dos resultados obtidos e os métodos empregados para sua aquisição.

Esse estudo é uma importante ferramenta para a PF conseguir uma divulgação de dados mais precisos, tornando-os mais visíveis para a população e também, para os próprios servidores. Uma das ferramentas utilizadas para isto então, é a criação de *dashboards*.

Assim, a criação de *dashboard* no trabalho ocorre para uma melhora na divulgação e apresentação dos dados da PF através do seu PDA. Isso acontece porque os *dashboards* conseguem deixar as informações obtidas com os dados mais visíveis e claros, tornando a apresentação mais interativa para o público.

Será utilizado no trabalho um *dashboard* analítico que como já foi descito, é utilizado para ajudar a identificar tendências e padrões de comportamento, que para o estudo são as apreensões de maconha e cocaína nos anos de 2019 e 2020 e por cada UF do Brasil. Todos os gráficos usados são gráficos interativos, onde é possível verificar valores extremos, analisar padrões, tendências e detectar o comportamento dessas apreensões por região e durante os anos.

4 Resultados

4.1 Validação do estudo

Após a implementação do algoritmo no estudo foram recuperados 86% dos valores das quantidades apreendidas descritas no texto da variável Observação. Essa análise se deu através da comparação com os valores da quantidade apreendida lançada no sistema pelos servidores da PF.

Os erros que ocorrem na recuperação de valores pelo algoritmo ocorrem na maior parte das vezes pela divergência na forma de escrita no texto do valor apreendido e a forma usual de escrita desses valores quando são escritos por extenso.

Um dos outros motivos para a ocorrência dos erros, quando ocorrem, é devido à informações divergentes. Em alguns casos, ocorre de no texto cadastrado no sistema possuir o valor da apreensão em 'massa líquida' e 'massa bruta', como denominado por eles. O que acontece nesses casos é que o valor cadastrado por eles no sistema pode estar como o valor em 'massa líquida' e o valor recuperado pelo algoritmo em 'massa bruta' ou vice-versa.

As análises de correlação também foram feitas de forma individual, pois as apreensões de maconha e cocaína divergem bastante em quantidades apreendidas. Para ambas as análises foi considerado um valor de erro absoluto menor ou igual à 99%, pois verificou-se que a maior parte dos erros possuíam esse comportamento.

4.1.1 Maconha

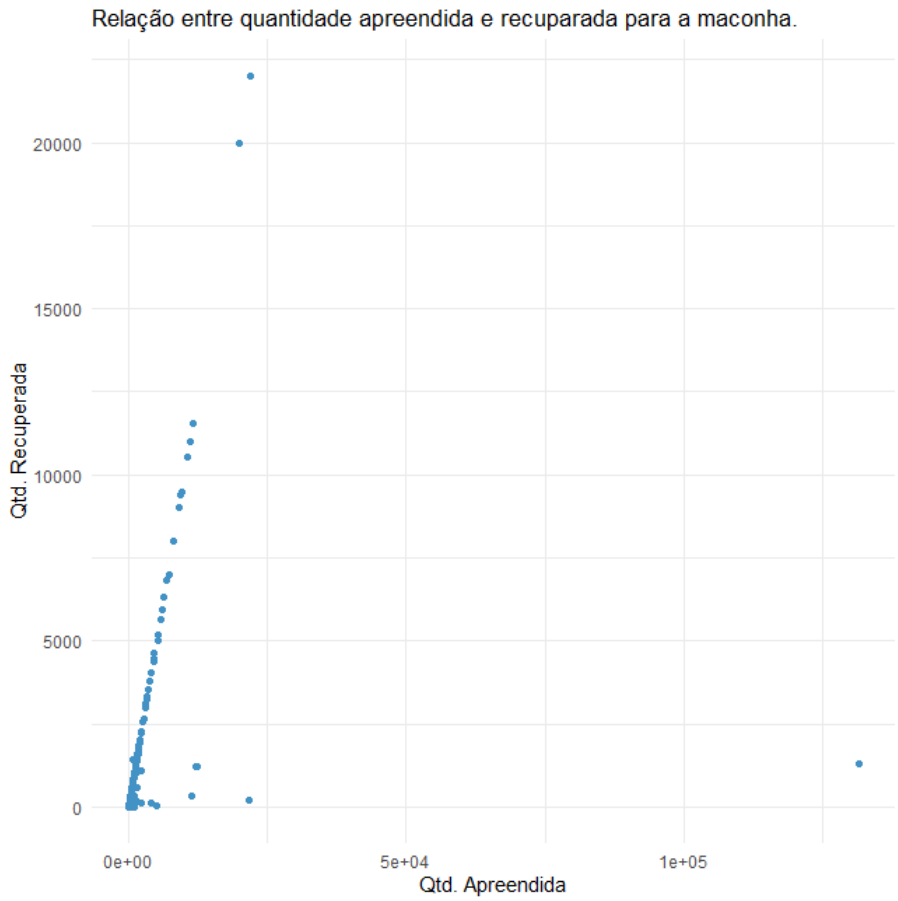

Figura 7: Gráfico de correlação da quantidade apreendida de maconha lançada no sistema da PF e quantidade recuperada pelo algoritmo (em kg)

Através da análise da Figura 7 vemos que a correlação entre as variáveis quase não existe. Isso ocorre pois alguns valores discrepantes muito altos influenciam nesse valor.

Para o estudo da análise de correlação foram considerados também os valores sem erro de recuperação (erro absoluto = 0%). A estatística t do teste tem valor igual à 14,957. O coeficiente de correlação de Pearson obtido na análise de correlação das variáveis quantidade apreendida e quantidade recuperada corresponde à 0,3784. Com esse valor para o coeficiente de correlação linear e p-valor do teste $< 2,2 \times 10^{-16}$, a hipótese nula do teste de que não existe correlação entre as variáveis é rejeitada. Apesar de baixo, vemos que há uma correlação entre as duas variáveis.

4.1.2 Cocaína

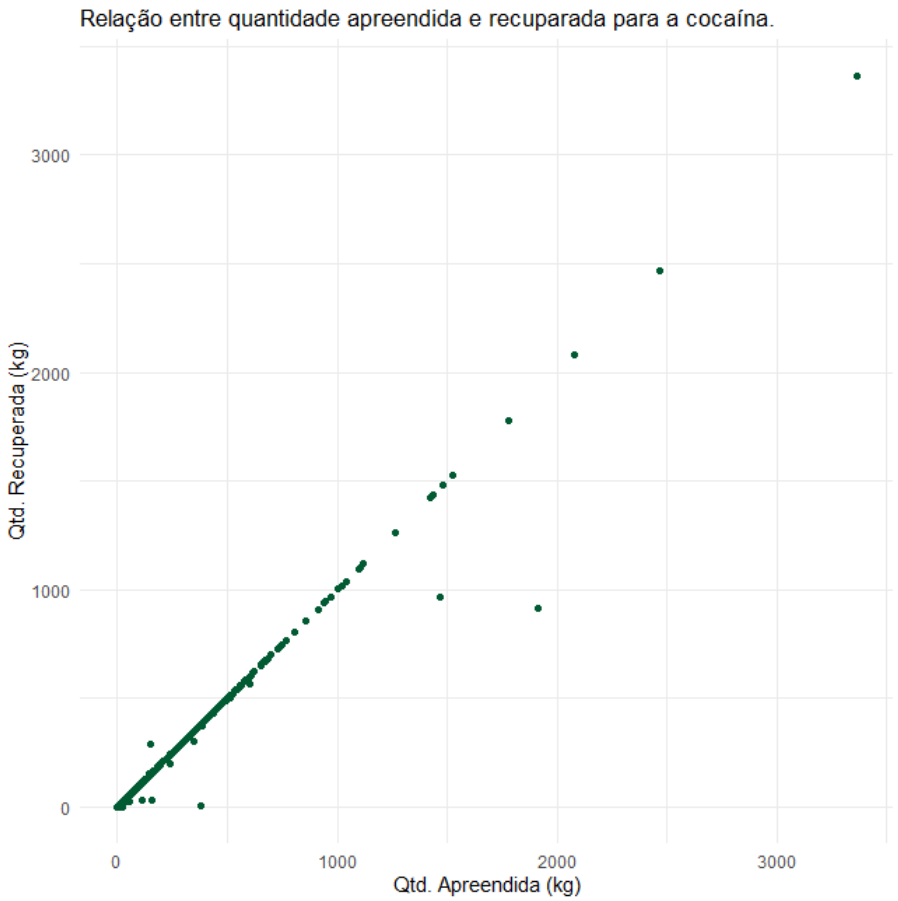

Figura 8: Gráfico de correlação da quantidade apreendida de cocaína lançada no sistema da PF e quantidade recuperada pelo algoritmo (em kg)

Na Figura 8, assim como em maconha, também conseguimos verificar uma correlação entre a quantidade apreendida e a quantidade recuperada pelo algoritmo. Entretanto, aqui percebemos que essa correlação é muito maior do que a da primeira análise de correlação das apreensões de maconha.

A estatística do teste t tem valor igual à 236,96. O coeficiente de correlação de Pearson obtido na análise de correlação das variáveis quantidade apreendida e quantidade recuperada corresponde à 0,9900248. Com esse valor para o coeficiente de correlação e p-valor do teste igual à $< 2,2 \times 10^{-16}$, a hipótese nula do teste de que não existe correlação entre as variáveis é rejeitada.

Esse valor obtido com o coeficiente nos mostra uma correlação quase perfeita, representando que a quantidade recuperada pelo algoritmo para as apreensões de cocaína tem uma alta confiabilidade.

4.2 Análise dos dados

Com isso verifica-se que as apreensões de maconha e cocaína pela PF apresentam um comportamento diferente e assim, as análises descritivas serão feitas de forma individual.

4.2.1 Apreensões de Maconha

As apreensões de maconha pela PF costumam ser em valores mais altos do que os valores das apreensões de cocaína. Isso ocorre pois a maconha que entra no Brasil geralmente vem de portos e por isso não seria viável para o contrabando enviar poucas quantidades. Isso será visto mais à seguir. Na Tabela 4 temos as principais estatísticas sobre essas apreensões.

Tabela 4: Medidas estatísticas sobre as apreensões de maconha (em kg)

Mínimo	1º Quartil	Mediana	Média	3º quartil	Máximo
0,001	0,670	11,855	213,888	52,255	21.606

Esses valores mostram que a quantidade apreendida de maconha nos anos de 2019 e 2020 pela PF tem valor mínimo muito baixo (0,001 kg) e um valor máximo muito alto (21.606 kg), o que representa uma alta variação nessas apreensões. Isso ocorre pois as grandes apreensões de maconha ocorrem, mas não são tão recorrentes quanto as de apreensões menores. Conseguimos verificar também que a quantidade média dessas apreensões é 213,888 kg.

Na Figura 9 vemos as apreensões de maconha por quantidade apreendida. Conseguimos verificar que a maior parte das apreensões (mais de 450) correspondem à pequenas quantidades. As apreensões com mais de 100 kg ocorrem, mas poucas vezes.

Teste de Shapiro Wilk

O teste de Shapiro Wilk foi utilizado para testar a normalidade das quantidades apreendidas de maconha. Obtemos que a estatística do teste (W) tem valor igual 0,148 e o p-valor igual a $< 2,2 \times 10^{-16}$. Assim, rejeitamos a hipótese nula do teste de que as quantidades apreendidas de maconha seguem uma distribuição normal.

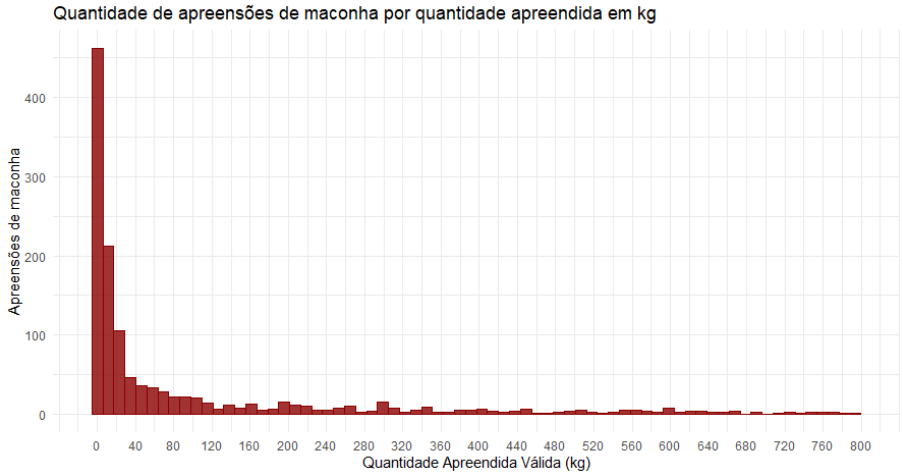

Figura 9: Apreensões de maconha por quantidade apreendida (em kg) nos anos de 2019 e 2020

4.2.1.1 - Análise Temporal

A análise temporal das apreensões de maconha será feita pelos meses dos anos de 2019 e 2020. Os meses são representados pelos números, sendo 1 janeiro, 2 fevereiro e assim por diante.

Aqui temos nas Figuras 10 a quantidade de apreensões de maconha por meses do ano e na 11 e a quantidade apreendida (em kg) de maconha por mês, ambas para os anos de 2019 e 2020. A distribuição das quantidades apreendidas é representada por um gráfico de barras e a quantidade apreendida através dos boxplots que conseguem representar melhor as medidas estatísticas desses valores.

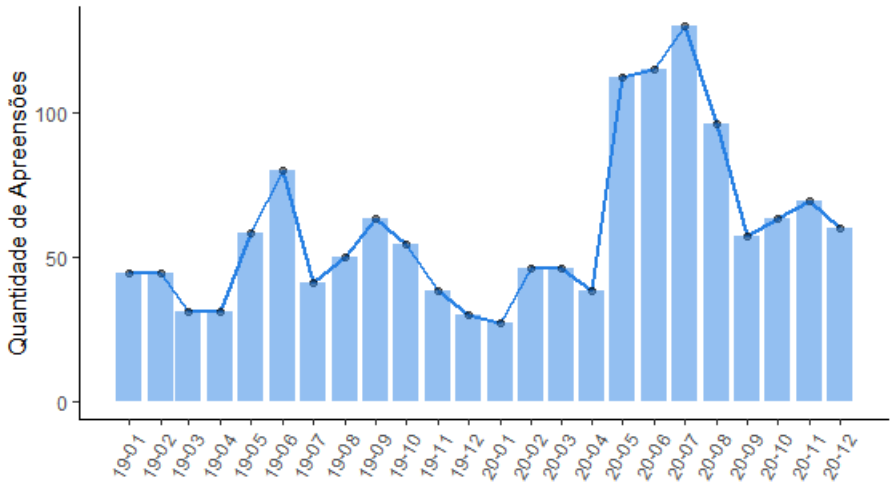

Figura 10: Número de apreensões de maconha pela PF nos meses dos anos de 2019 e 2020

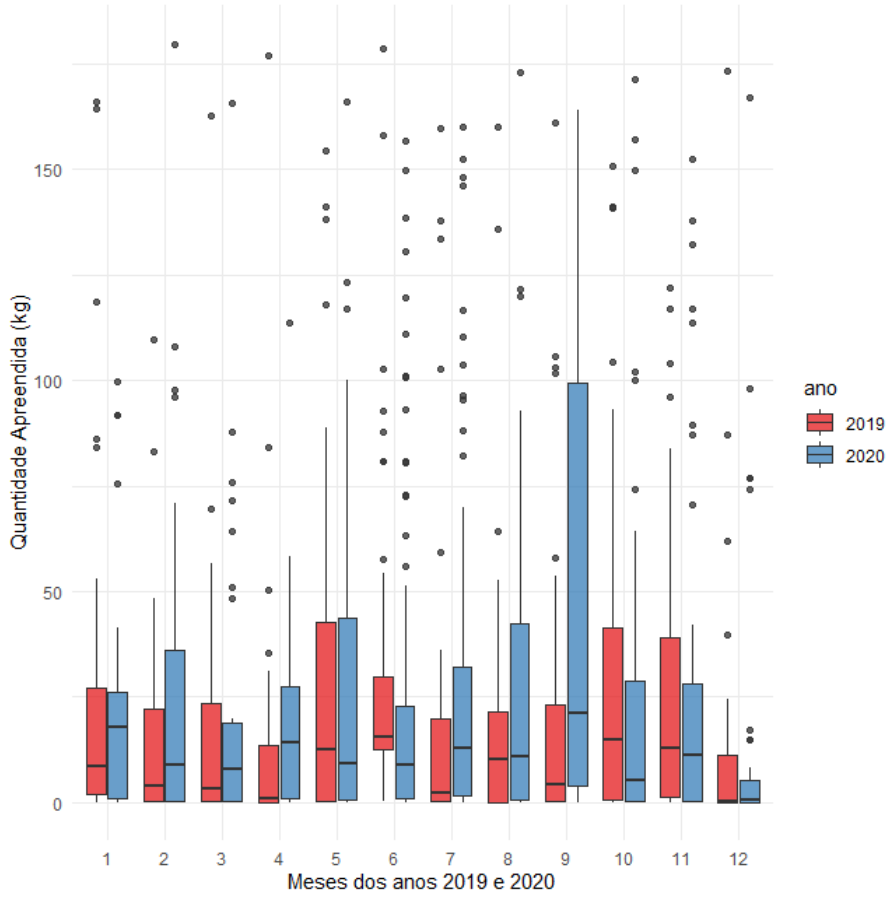

Figura 11: Boxplots da quantidade apreendida (em kg) de maconha pela PF por meses dos anos de 2019 e 2020

Conseguimos verificar na Figura 10 melhor o comportamento do número de apreensões de maconha durante os meses dos anos e também que nos meses de julho a setembro do ano de 2019 são os meses com maiores apreensões. No ano de 2020 os meses de maio até o mês de agosto. Isso ocorre pois historicamente as opreações da PF nos meses iniciais dos anos realmente diminuem.

Já com a Figura 11 conseguimos verificar que em ambos os meses de ambos os anos ocorrem muitos valores discrepantes nas quantidades apreendidas (em kg) e temos a média e a mediana dessas quantidades como valores baixos.

Teste de Kruskal-Wallis

O teste de Kruskal-Wallis foi utilizado para testarmos se o número de apreensões de maconha durante os meses dos anos de 2019 e 2020 são semelhantes ou não.

Obtemos que a estatística do teste tem valor igual a 35,275 e o p-valor do teste igual a $2,863 \times 10^{-9}$. Assim, rejeitamos a hipótese nula do teste de que o número de apreensões de maconha são semelhantes nos 12 meses dos anos.

4.2.1.2 - Análise Geográfica

A análise geográfica das apreensões de maconha será feita para todas as Unidades Federativas do Brasil para tentarmos identificar o comportamento dessas apreensões ao longo de todo o país.

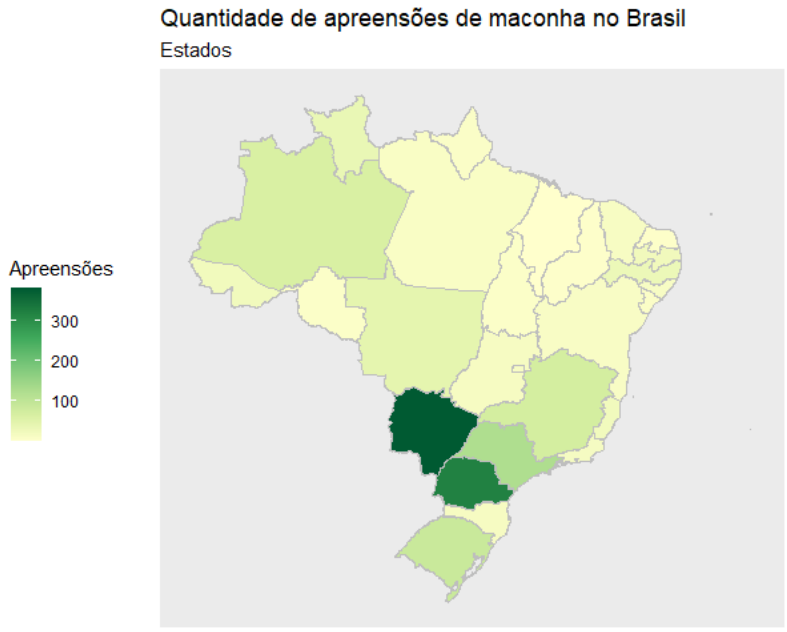

Figura 12: Mapa de calor do número de apreensões de maconha pela PF por cada UF do Brasil nos anos de 2019 e 2020

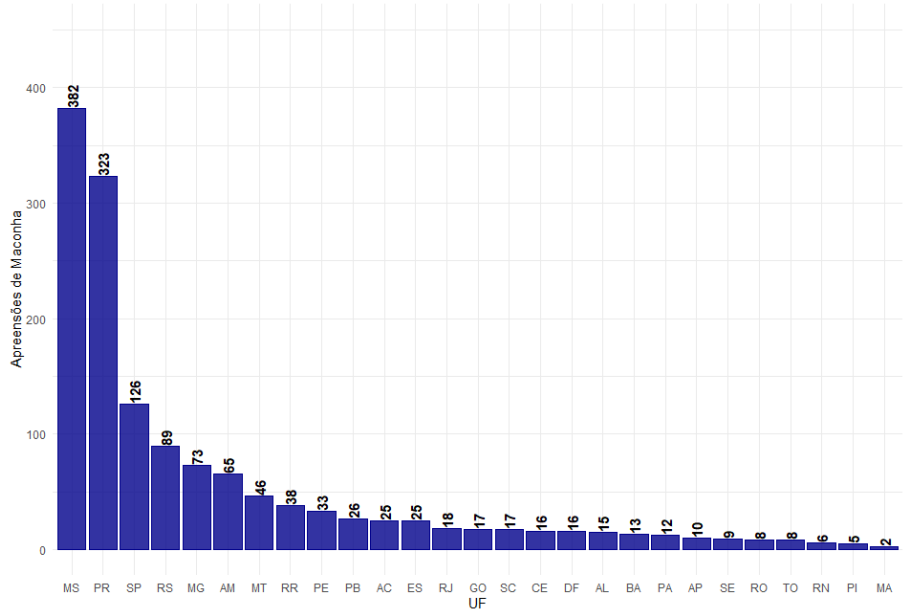

Figura 13: Número de apreensões de maconha pela PF por UF do Brasil nos anos de 2019 e 2020

Inicialmente, nas Figuras 12 e 13 vemos a distribuição do número de apreensões por cada UF do país. Conseguimos verificar aqui uma maior concentração de apreensões de maconha no Paraná e no Mato Grosso do Sul. Isso ocorre pois grande parte da maconha que é apreendida no Brasil vem do Paraguai e esses são os estados que fazem fronteira com o país. Assim, as apreensões nessas duas UFs são de valores muito altos pois para o contrabando não é viável enviar pequenas quantidades de maconha para o Brasil.

Por fim, na Figura 14 conseguimos observar a quantidade apreendida de maconha (em kg) para cada UF do Brasil. Conseguimos confirmar que as UFs onde ocorrem as maiores apreensões (em número de apreensão) também são as que ocorrem as maiores quantidades apreendidas (em kg). Aqui verificamos ainda o mesmo que ocorre nas análises temporais onde existem muitos valores discrepantes e valores baixos para média e mediana de cada UF.

Teste de Kruskal-Wallis

O teste de Kruskal-Wallis foi utilizado para testarmos se as quantidades apreendidas de maconha em kg são significativamente divergentes de uma UF para outra.

Obtemos que a estatística do teste tem valor igual a 199,83 e o p-valor do teste igual a $< 2,2 \times 10^{-16}$. Assim, rejeitamos a hipótese nula do teste de que as quantidades apreendidas de maconha em todas as UFs são semelhantes.

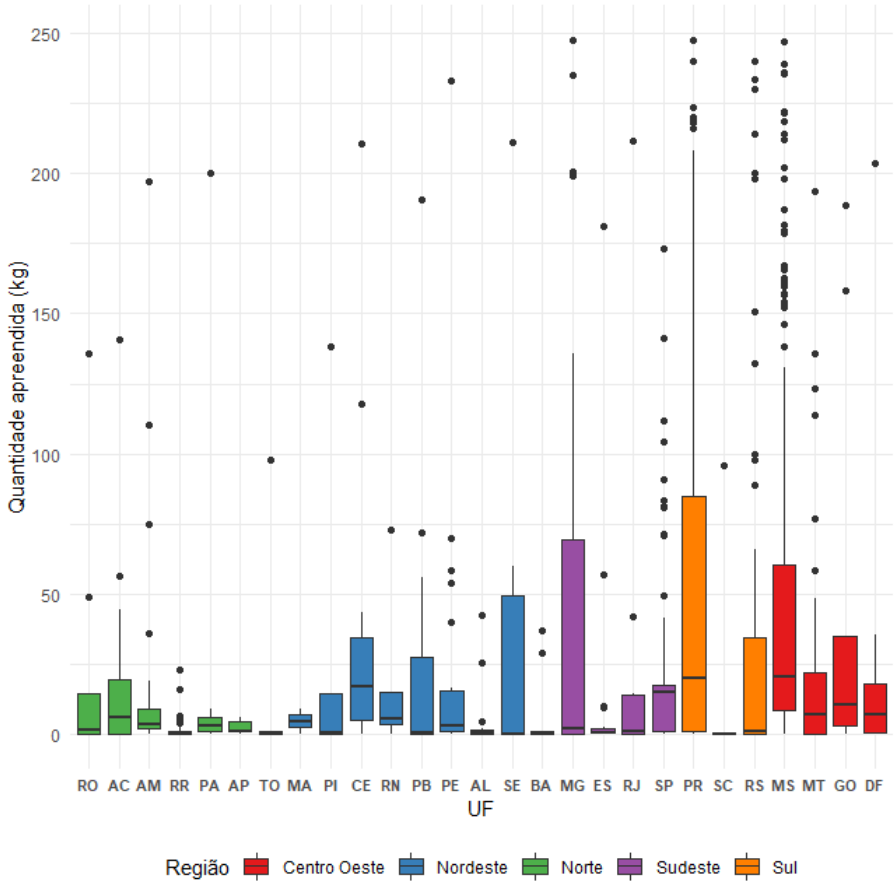

Figura 14: Boxplots da quantidade apreendida (em kg) de maconha pela PF por UF do Brasil nos anos de 2019 e 2020

4.2.2 Apreensões de cocaína

A segunda análise é sobre as apreensões de cocaína pela PF nos anos de 2019 e 2020. Como dito na seção acima as apreensões de cocaína realmente são em valores menores e isso se dá pois a maior parte da cocaína que chega até o Brasil vem de aeroportos e então as quantidades não são tão altas quanto às de maconha.

Tabela 5: Medidas estatísticas sobre as apreensões de cocaína

Mínimo	1º Quartil	Mediana	Média	3º quartil	Máximo
0,0001	0,5170	2,9150	21,0202	10,1477	1347,0000

Vemos na Tabela 5 a menor quantidade apreendida de cocaína é 0,0001 kg e a maior 1347,00 kg. A quantidade média de apreensões de cocaína tem o valor de 21,0202 kg. Com esses dados conseguimos ver também aqui a variação alta nos valores das apreensões.

Na Figura 15, assim como vemos em maconha, temos as apreensões de cocaína por quantidade apreendida. Conseguimos verificar aqui que também a maior parte das

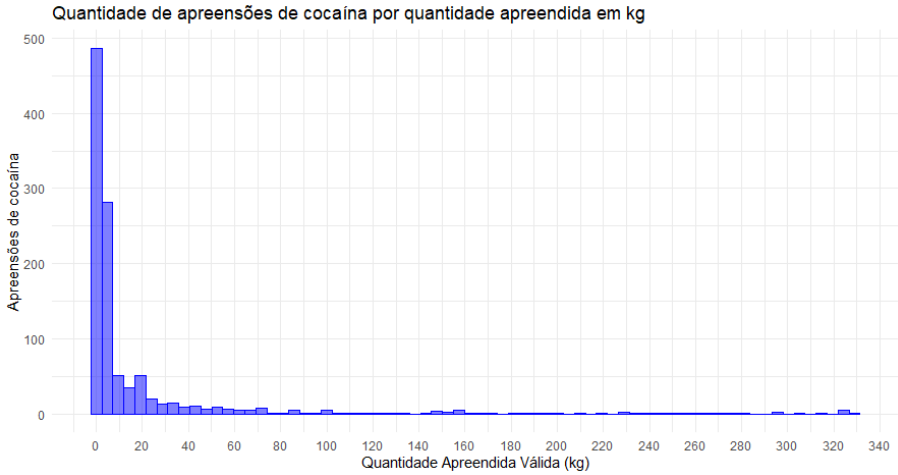

Figura 15: Apreensões de cocaína por quantidade apreendida (em kg) nos anos de 2019 e 2020

apreensões (quase 500) correspondem à pequenas quantidades. Entretanto, o que vemos aqui nas apreensões de cocaína é que as quantidades apreendidas (em kg) são bem menores dos que as de maconha.

Teste de Shapiro Wilk

O teste de Shapiro Wilk foi utilizado para testar a normalidade das quantidades apreendidas de cocaína. Obtemos que a estatística do teste (W) tem valor igual 0,175 e o p-valor igual a $< 2,2 \times 10^{-16}$. Assim, rejeitamos a hipótese nula do teste de que as quantidades apreendidas de cocaína seguem uma distribuição normal.

4.2.2.1 - Análise Temporal

A análise temporal das apreensões de cocaína também será feita pelos meses dos anos de 2019 e 2020 assim como a de apreensões de maconha. Os meses são representados pelos números, sendo 1 janeiro, 2 fevereiro e assim por diante.

Nas Figuras 16 e 17 temos as representações gráficas da quantidade de apreensões de cocaína e da quantidade apreendida (em kg) por mês dos anos de 2019 e 2020. Também aqui, a distribuição da quantidade de apreensões é representada por um gráfico de barras e a quantidade apreendida através dos boxplots que conseguem representar melhor as medidas estatísticas desses valores.

Conseguimos verificar na Figura 16 que aqui, como ocorre com maconha, os meses iniciais do ano também são os meses com menores apreensões de cocaína e isso também ocorre pelos mesmo motivos que ocorre com maconha, a redução de operações pela PF

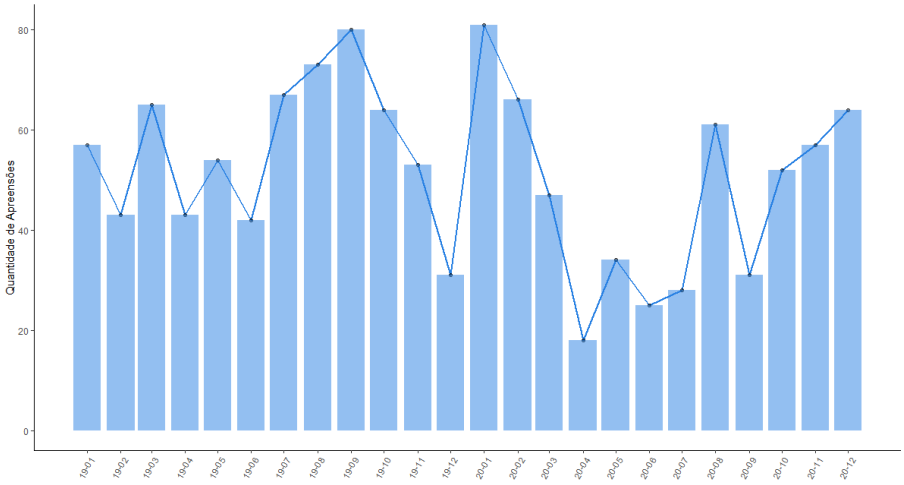

Figura 16: Quantidade de apreensões de cocaína por meses dos anos 2019 e 2020

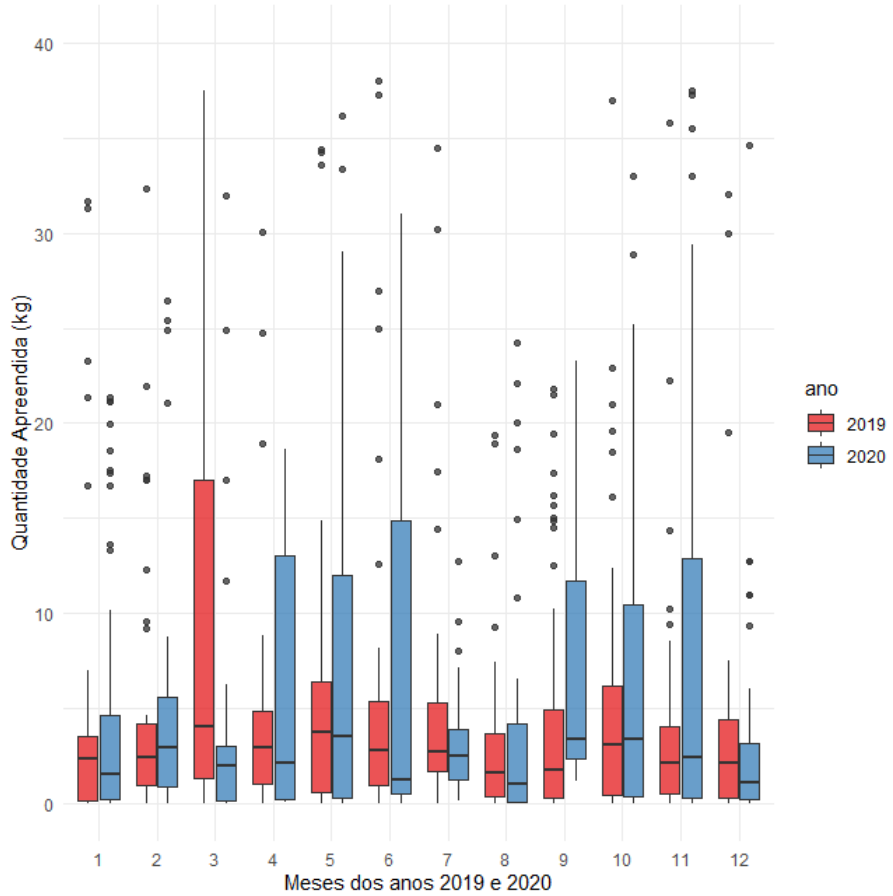

Figura 17: Boxplots da quantidade apreendida (em kg) de cocaína por meses dos anos 2019 e 2020

nos meses iniciais do ano.

Com a Figura 17 verificamos que em ambos os meses dos anos de 2019 e 2020 ocorrem muitos valores discrepantes ns quantidades apreendidas (em kg) e temos a média e a mediana dessas quantidades como valores baixos.

Teste de Kruskal-Wallis

O teste de Kruskal-Wallis foi utilizado para testarmos se o número de apreensões de cocaína durante os meses dos anos de 2019 e 2020 são semelhantes ou não.

Obtemos que a estatística do teste tem valor igual a 35,273 e o p-valor do teste igual a $2,866 \times 10^{-9}$. Assim, rejeitamos a hipótese nula do teste de que o número de apreensões de cocaína são semelhantes nos 12 meses dos anos.

4.2.2.2 - Análise geográfica

A análise geográfica das apreensões de cocaína também será feita para todas as Unidades Federativas do Brasil para tentarmos identificar o comportamento dessas apreensões ao longo de todo o país.

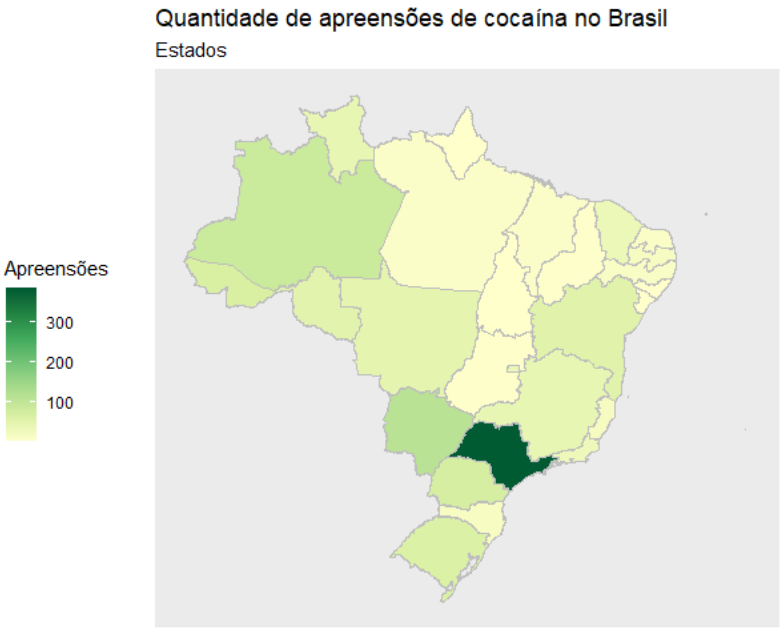

Figura 18: Mapa de calor do número de apreensões de cocaína pela PF por cada UF do Brasil nos anos de 2019 e 2020

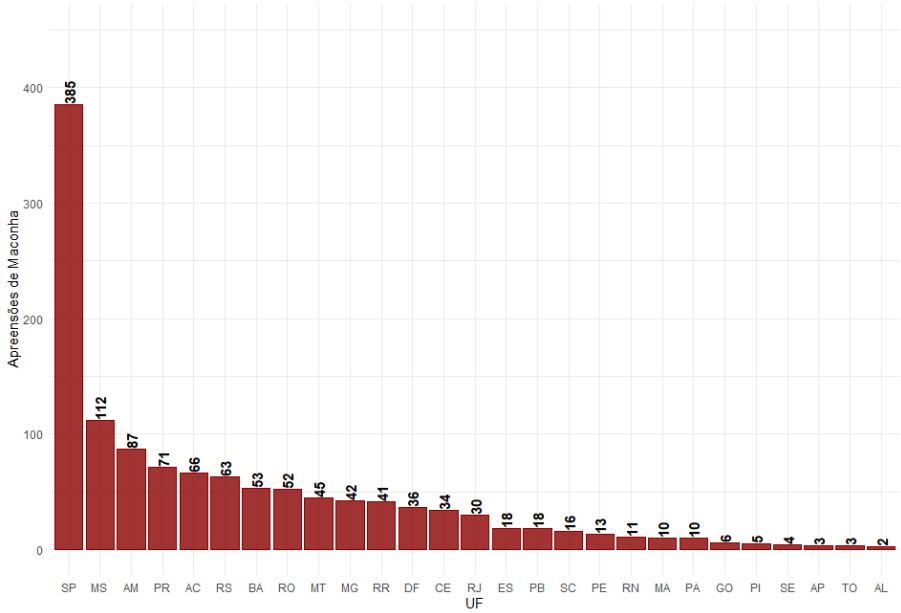

Figura 19: Número de apreensões de cocaína pela PF por UF do Brasil nos anos de 2019 e 2020

Inicialmente, nas Figuras 18 e 19 vemos a distribuição do número de apreensões de cocaína por cada UF do país. Conseguimos verificar aqui uma maior concentração de apreensões de cocaína em São Paulo e em Mato Grosso do Sul. O número de apreensões nessas duas UF do país são historicamente mais altas pois a cocaína que chega ao Brasil, diferentemente da maconha, chega em menores quantidades e por isso, um dos principais meios de chegada ao país é pelos aeroportos, onde ocorrem as maiores apreensões.

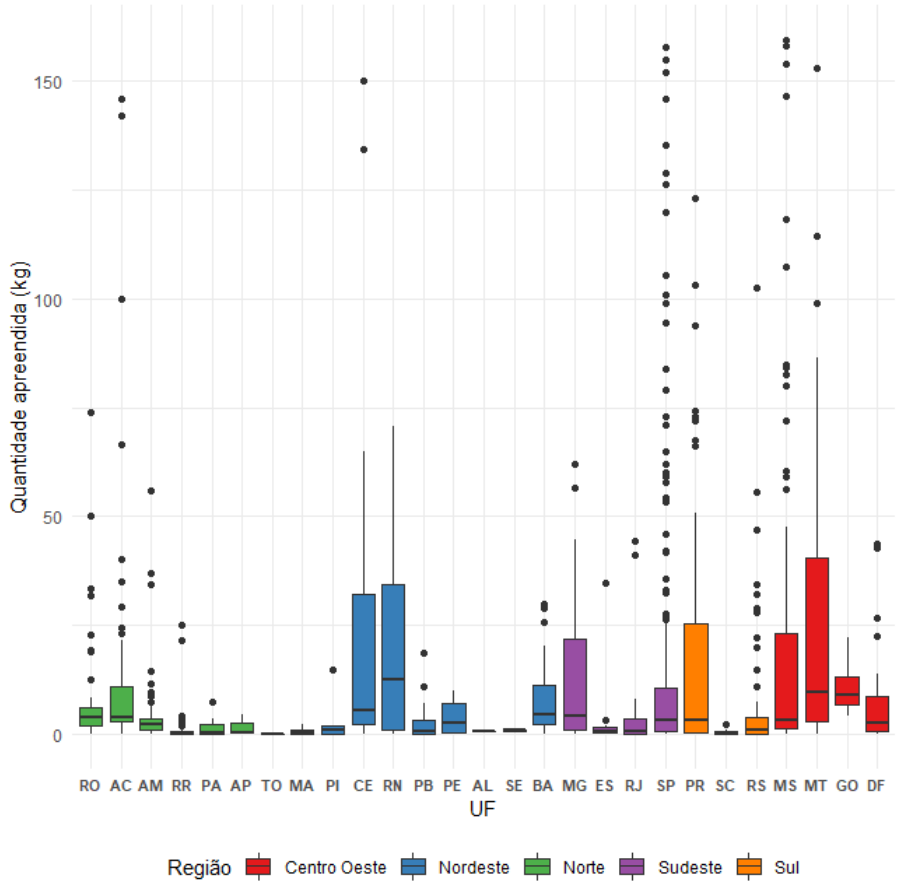

Figura 20: Boxplots da quantidade apreendida de cocaína (em kg) pela PF por UF do Brasil nos anos de 2019 e 2020

Por fim, na Figura 20 conseguimos observar a quantidade apreendida de cocaína (em kg) para cada UF do Brasil nos anos de 2019 e 2020. Novamente conseguimos confirmar que as UFs onde ocorrem as maiores apreensões (em número de apreensão) também são as que ocorrem as maiores quantidades apreendidas (em kg). Aqui verificamos ainda o mesmo que ocorre nas análises temporais onde existem muitos valores discrepantes e valores baixos para média e mediana de cada UF.

Teste de Kruskal-Wallis

O teste de Kruskal-Wallis foi utilizado para testarmos se as quantidades apreendidas de cocaína em kg são significativamente divergentes de uma UF para outra.

Obtemos que a estatística do teste tem valor igual a 152,08 e o p-valor do teste igual a $< 2,2 \times 10^{-16}$. Assim, rejeitamos a hipótese nula do teste de que as quantidades apreendidas de cocaína em todas as UFs são semelhantes.

4.3 *Dashboard*

O *dashboard* analítico desenvolvido no estudo é interativo apresentando 4 telas de visualização e todos seus gráficos e sua tabela também interativos. O objetivo desse *dashboard* criado é disponibilizar para a PF um exemplo para os *dashboards* já criados por eles onde há a divulgação de estatísticas importantes de uma maneira de mais fácil compreensão e visualização.

Na tela inicial representada pela Figura 21 é apresentado a análise das apreensões por cada UF através do mapa do Brasil. Também são disponibilizadas informações acerca das quantidades apreendidas lançadas pela PF e a quantidade apreendida recuperada pelo algoritmo.

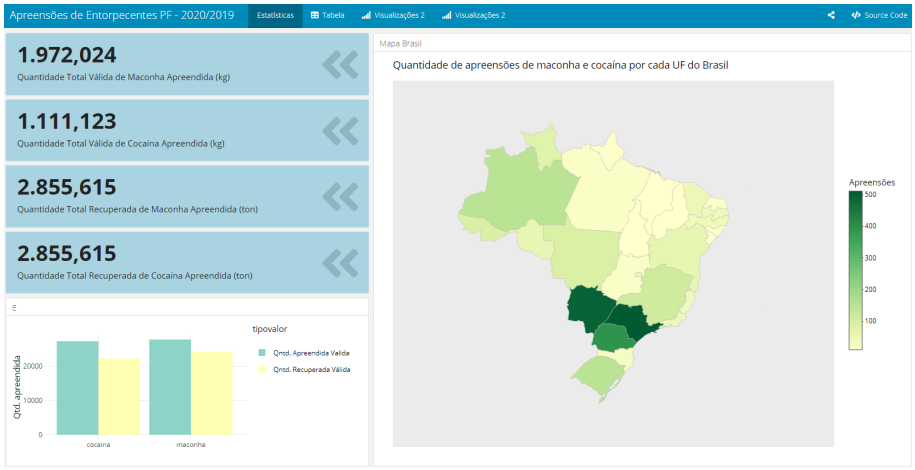

Figura 21: Tela inicial do *dashboard* criado acerca das apreensões de maconha e cocaína nos anos de 2019 e 2020 pela PF

Na segunda tela do *dashboard* representada pela Figura 22 será possível realizar a análise de correlação através dos gráficos interativos, onde cada ponto da correlação poderá ser analisado individualmente podendo observar o motivo daquele valor. Também é apresentado uma tabela interativa que apresenta a quantidade de apreensões, quantidade apreendida e a quantidade recuperada para ambas as drogas, sendo possível aplicar filtros sobre UF, mês da apreensão e ano da apreensão.

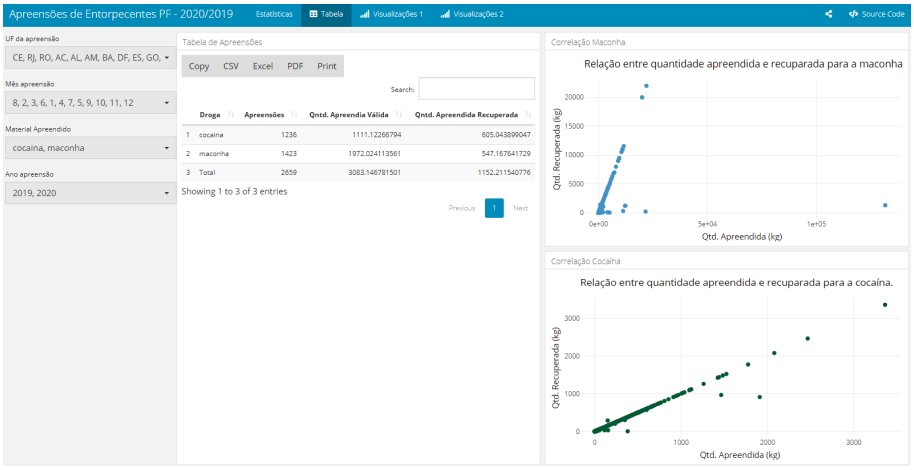

Figura 22: Segunda tela do *dashboard* criado acerca das apreensões de maconha e cocaína nos anos de 2019 e 2020 pela PF

Na terceira tela do *dashboard* representada pela Figura 23 as análises de quantidades de apreensões são realizadas através dos gráficos de barras apresentados. Os histogramas disponibilizados permitem verificar a quantidade de apreensões por cada valor de quantidade apreendida. O intuito das representações gráficas aqui é deixar mais visível e mais compreensível as quantidades de apreensões realizadas pela PF.

Figura 23: Terceira tela do *dashboard* criado acerca das apreensões de maconha e cocaína nos anos de 2019 e 2020 pela PF

Por fim, na última tela do nosso *dashboard* representada pela Figura 24 a análise será acerca das apreensões de maconha e cocaína pela data de apreensão. Aqui conseguimos verificar o comportamento das apreensões durante os meses dos anos de 2019 e 2020 o que pode possibilitar para a PF um estudo do comportamento das quantidades de apreensões durante os anos.

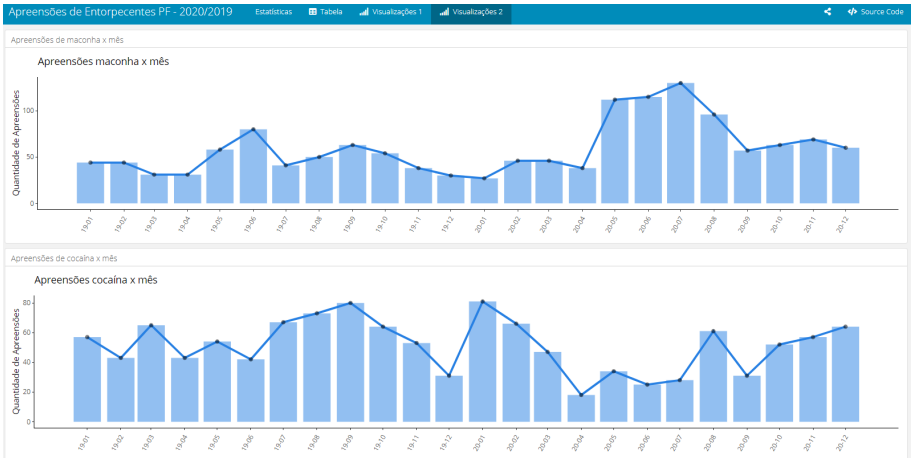

Figura 24: Quarta tela *dashboard* acerca das apreensões de maconha e cocaína nos anos de 2019 e 2020 pela PF

5 Considerações finais

Em 18 de novembro de 2011, foi publicada a Lei nº 12.527 (Lei de Acesso a Informação - LAI), que regula o acesso a informações. Conforme previsão no Art. 6º da referida lei:

> "Art. 6º Cabe aos órgãos e entidades do poder público, observadas as normas e procedimentos específicos aplicáveis, assegurar a:
>
> I - gestão transparente da informação, propiciando amplo acesso a ela e sua divulgação;
>
> II - proteção da informação, garantindo-se sua disponibilidade, autenticidade e integridade; e
>
> III - proteção da informação sigilosa e da informação pessoal, observada a sua disponibilidade, autenticidade, integridade e eventual restrição de acesso."(BRASIL, 2011)

O presente estudo baseou-se no Plano de Dados Abertos da Polícia Federal que tem como referência a LAI. O intuito do estudo foi auxiliar a PF para uma maior qualidade dos dados que serão divulgados para a população. Mas os dados que serão divulgados à população precisam ser dados seguros e de alta qualidade para não ocasionar em análises distorcidas.

Para que esses dados sejam cada dia mais seguros é necessário uma alta tratativa desses dados antes de suas divulgações. Entretanto esse serviço de trativa dos dados ainda não ocorre de forma automática, prejudicando um pouco a qualidade desses e tomando um tempo maior de seus servidores para fazer essa tratativa de forma manual.

Ao início desse trabalho possuíamos uma base de dados com 12930 observações

onde não havia nenhum filtro acerca do tipo de drogas e do tipo de procedimento. Durante o processo para a criação do algoritmo foram necessárias algumas alterações, filtrar alguns dados e retirada da base de dados aqueles que não possuíam nenhuma informação (quantidade apreendida, por exemplo) para que o algoritmo consiga recuperar os valores apreendidos .

Os filtros aplicados na base de dados foram os seguintes: tipo de procedimento apenas IPL, tipo de droga apenas maconha e cocaína e anos apenas 2019 e 2020 (isso porque mesmo a base de dados sendo desses anos ainda ocorreram algumas observações de anos divergentes).

Portanto, ao início desse trabalho possuíamos uma base de dados com 12930 observações, onde todas são verificadas manualmente pelos servidores da CGPRE/DICOR/PF. Ao fim do trabalho com a implementação do algoritmo apenas 334 observações não foram recuperadas. Apesar de ainda não ser 100% eficaz, o algoritmo reduz em 99,31% das observações que serão necessárias que os servidores as tratem manualmente.

A tratativa de bases de dados antes de iniciar um estudo estatístico é um dos maiores trabalhos que os servidores de empresas e órgãos públicos, pesquisadores entre outros enfrentam, pois muitas vezes esse trabalho é feito de forma manual e demanda um certo tempo.

A criação desse algoritmo usado no trabalho tem como intuito auxiliar a PF nas trativas de seus bancos de dados de forma automática melhorando cada dia mais a qualidade desses não só para serem divulgados para a população, mas também para a criação de estatísticas de uso próprio da PF. O *dashboard* construído no trabalho apresenta um modelo de como essas informações acerca do trabalho realizado pela PF, como por exemplo as apreensões que são o objeto de estudo do presente trabalho, poderão ser dispostas de maneira mais informativa para seus servidores.

Ao fim desse trabalho chegamos a conclusão de um algoritmo criado que irá agilizar o processo de tratativa das bases de dados e de um *dashboard* para divulgação das análises estatísticas temporais e geográficas também feitas durante o estudo. Esse estudo tem o intuito de auxiliar a PF em futuras análises com tratativas de dados mais ágeis, monitoramento de seus *dashboards* e até inspiração para os já criados por eles.

Referências

BATISTA, A. *Afinal, o que é dashboard e para que serve?* 2018. Disponível em: ⟨https://blog.hariken.co/afinal-o-que-e-dashboard-e-para-que-serve/⟩.

BRASIL. *Lei nº 12.527, de 18 de novembro de 2011 - Lei de Acesso à Informação – LAI.* 2011. Disponível em: ⟨http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm⟩.

BRASIL. *Polícia Federal tem sua primeira delegacia com todos os inquéritos policiais em meio digital.* 2019. Disponível em: ⟨https://www.gov.br/pf/pt-br/assuntos/noticias/2019/09/policia-federal-tem-sua-primeira-delegacia-com-todos-os-inqueritos-policiais-em-meio-digital⟩.

BRASIL, P. F. *Plano de Dados Abertos 2020-2022.* 2020.

CANTELI, A. *Teste de Shapiro-Wilk.* 2020.

CGPRE, C. G. de Repressão a E. [S.l.], 2021.

DIXON, M. An overview of document mining technology. 1997. Disponível em: ⟨http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dixm97_dm.ps⟩.

EIS, D. *O básico sobre Expressões Regulares. Desmistificando as Expressões Regulares.* 2016. Disponível em: ⟨https://tableless.com.br/o-basico-sobre-expressoes-regulares/⟩.

FAUSETT, L. V. *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications.* [S.l.: s.n.], 1994.

GOMES, P. C. T. *O QUE É UM DASHBOARD? O GUIA COMPLETO E DEFINITIVO!* 2017. Disponível em: ⟨https://www.opservices.com.br/o-que-e-um-dashboard/⟩.

GUEDES, T. et al. *Estatística descritiva: Projeto de Ensino-Aprender Fazendo Estatística.* [S.l.]: EACH-USP, 2010.

HUOT, R. *Métodos quantitativos para as ciências humanas (tradução de Maria Luísa Figueiredo).* [S.l.: s.n.], 2002.

JUNIOR, G. de B. V. *Estatística: Teste de Kruskal-Wallis.* 2020.

LOH, S.; WIVES, L. K.; OLIVEIRA, J. P. M. de. *Descoberta proativa de conhecimento em coleções textuais: iniciando sem hipóteses.* 2000.

MORAIS, C. Descrição, análise e interpretação de informação quantitativa. *Escalas de medida, estatística descritiva e inferência estatística. Escola Superior de Educação-Instituto Politécnico de Bragança-2012*, 2010.

MORAIS, E. A. M. Mineração de textos. 2007.

MOREIRA, D. *Text as Data* para ciências sociais - guia prático com aplicações. 2020.

PETERNELLI, L. A. CapÍtulo 9 - regressão linear e correlação. 2004. Disponível em: ⟨http://www.dpi.ufv.br/~peternelli/inf162.www.16032004/materiais/CAPITULO9.pdf⟩.

RODRIGO. *Erro absoluto e relativo: definição e fórmula*. 2020. Disponível em: ⟨https://pt.estudyando.com/erro-absoluto-e-relativo-definicao-e-formula⟩.

SILVA, E. M. *Descoberta de conhecimento com o uso de text mining: cruzando o abismo de Moore*. 2002.

UBER, J. L. *Descoberta de conhecimento com o uso de text mining aplicada ao SAC*. 2004.