INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA BAHIA
Campus Santo Antônio de Jesus
Curso Superior de Tecnologia em Análise
e Desenvolvimento de Sistemas

Alex Gondim Lima

Data Science e Analytics como ferramenta essencial para implementar políticas de Segurança Pública no Brasil.

Prof. Me. Ikaro Campos de Araújo.

Santo Antônio de Jesus (BA) – 2023

Sumário

RESUMO 03

1. INTRODUÇÃO 04

2. FUNDAMENTAÇÃO TEÓRICA. 07

3. METODOLOGIA 10

3.1 Ferramentas Tecnológicas Utilizadas 11

3.2 Definição do Probelma 11

3.3 Coleta de dados 12

3.4 Descrição dos dados 12

3.5 Preparação dos dados ou Pré -Processamento 13

3.6 Exploração dos dados 14

3.7 Visualização dos Dados 15

4. RESULTADOS 16

4.1 Análise Descritiva 16

4.1.2 Exploração e Visualização de Dados 20

4.2 Análise Diagnóstica 24

4.2 .1 A expansão para Análise de Séries Temporais 24

5. DISCUSSÃO 28

6. CONCLUSÃO E TRABALHOS FUTUROS 31

REFERENCIAS BIBLIOGRÁFICAS 33

RESUMO

A Ciência de Dados é uma área multidisciplinar que combina métodos matemáticos e estatísticos com técnicas computacionais para resolver problemas em diversos campos. Na área de Segurança Pública, a Ciência de Dados pode ser utilizada para identificar proativamente novos comportamentos criminosos e entender em quais condições eles têm maior probabilidade de ocorrer. Para isso, é necessário coletar e analisar dados sobre crimes ao longo do tempo, utilizando técnicas de machine learning e análise de dados. Os principais resultados de uma amostragem mostram que a ciência de dados pode ajudar a identificar padrões que possam indicar ações criminosas e, assim, permitir que as forças policiais possam prevenir crimes antes que aconteçam. Além disso, o uso de dados pode permitir às autoridades concentrar melhor seus recursos e ter um impacto positivo na sociedade. Em resumo, a Ciência de Dados oferece ferramentas e conjuntos de dados cada vez mais avançados para entender e combater o crime, melhorar a eficácia das operações policiais e reduzir a criminalidade.

1. INTRODUÇÃO

Segundo Sebastian-Coleman (2013), dados são representações abstratas de características selecionadas de objetos do mundo real, eventos e conceitos, expressos e compreendidos através de convenções explicitamente definidos relacionados com o seu significado, coleta e armazenamento. Em suma, as escolhas de como codificar os dados são influenciadas pelo modo como os dados serão criados, utilizados, armazenados e acessados. A interpretação dos dados depende de um processo de decodificação.

Os últimos anos tem vivenciado uma revolução na forma como os dados são produzidos e como circulam entre pessoas, empresas e governos. As novas tecnologias existentes potencializaram a importância dos dados para a humanidade, e o mesmo ocorreu nos governos ao redor do mundo. Essa nova realidade passou a exigir novas competências daqueles que trabalham nos governos hoje. Ao mesmo tempo, há um claro *gap* entre as competências necessárias e a capacidade existente nos seus recursos humanos.

No que tange a ciência de dados, ela é uma ferramenta possível para desenvolver modelos de previsão e classificação, que são fundamentais para a tomada de decisões estratégicas, principalmente em ambientes dinâmicos, que possuem grandes volumes de dados em ritmo acelerado. Além disso, a ciência de dados também pode ser usada para melhorar processos e otimizar recursos, detectar fraudes e, até mesmo, para entender melhor o comportamento de indivíduos e pessoas através de técnicas de Análise Temporal dos Dados e Machine Learning.

O crime pode assumir muitas formas e está em constante mudança. Lutar contra o crime é como lutar contra um monstro de muitas cabeças já que, quando cortamos uma cabeça, duas outras brotam, que são mais inteligentes e ferozes do que antes. Assim sendo, para combater o crime, precisamos ser flexíveis e criativos. Em um futuro próximo, cada vez mais automatizado, sentimos que o campo da ciência de dados nos dará o poder e a flexibilidade para lidar com crimes, prevenção e investigação.

Por isso é de fundamental importância discutir o uso da ciência de dados, visando compreender a função do profissional da área, no intuito de desenvolver e utilizar ferramentas de análise de dados no planejamento da atividade policial para o combate efetivo ao crime. Dessa forma, é possível que muitos dos insights gerados após as análises dos dados fornecidos por tais ferramentas, possam contribuir para novos debates de implementação de melhorias nas políticas públicas relativas ao tema em questão.

Por existir uma vasta quantidade de dados, estruturados e não estruturados, que são coletados pela polícia e por outras forças de segurança pública diariamente, quando não são bem organizados e analisados com as ferramentas adequadas, tendem a gerar desafios analíticos significativos. O que interfere diretamente na organização policial e estruturas de políticas públicas de segurança que atendam às demandas.

Para uma melhor compreensão, dados estruturados são aqueles que possuem bem definidas, pensadas antes mesmo da existência do dado que será carregado, não permitindo que um dado de diferente estrutura, daquela pré-definida, seja carregado. Já os dados não estruturados não possuem essa estrutura pré-definida, podendo ser compostos por diversos elementos diferentes dentro de um todo. A Figura 1 define de forma ilustrativa o que são dados estruturados e não estruturados .

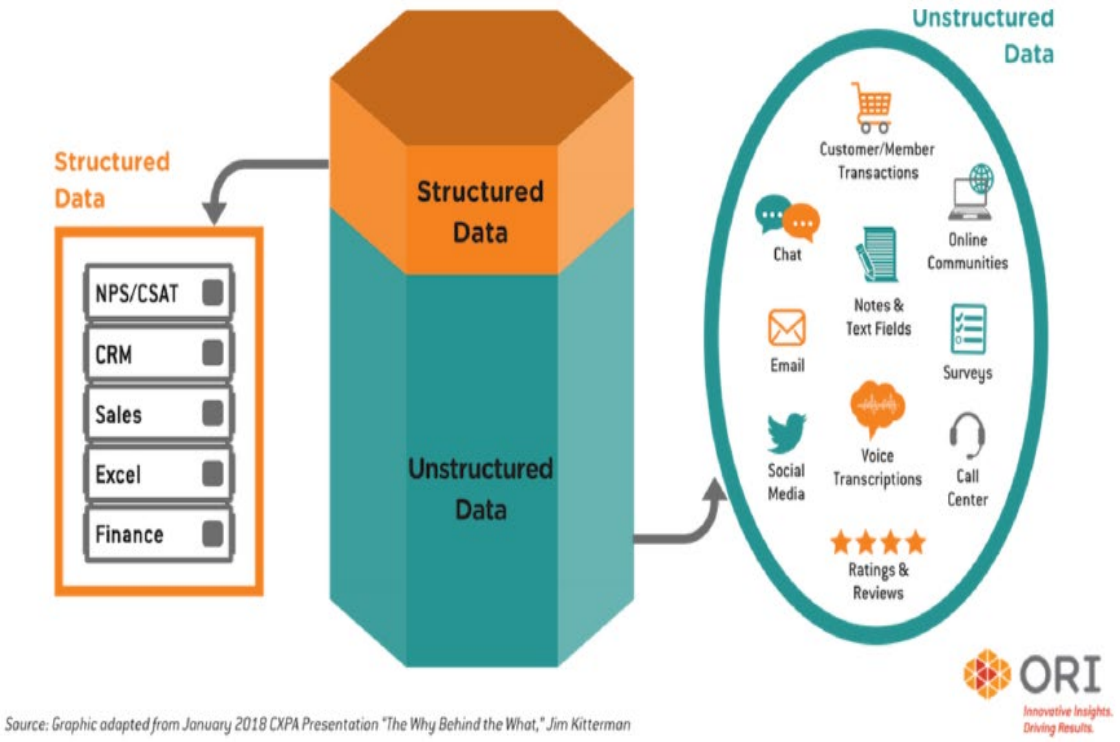

Figura 1 – Coleta de Dados

Abordamos neste estudo conceitos práticos e teóricos a fim de caracterizar técnicas de análise descritiva, análise diagnóstica, e abordagens inteligentes sobre dados de ocorrências criminais no Brasil. No que tange a análise descritiva consiste em uma forma de estudo que visa obter informações sobre os dados de ocorrências criminais, através da descrição e interpretação dos dados coletados. Esta técnica busca verificar a incidência de crimes em diferentes regiões do país, bem como o tipo de crimes que ocorrem em cada local.

Com isso, é possível descobrir quais são os locais onde é necessário maior atenção e, consequentemente, aplicar medidas de prevenção. Análise diagnóstica é uma técnica usada para identificar e diagnosticar problemas em um sistema ou organização. É usado para avaliar o desempenho, identificar áreas de melhoria e identificar as causas de problemas. Além disso, no presente estudo se propõe padronizar a coleta de dados de ocorrências criminais através da criação de um Big Data. Essa iniciativa envolveria a coleta, armazenamento e análise de uma grande quantidade de dados de ocorrências criminais de diferentes fontes, em um formato padronizado e acessível.

A criação de um Big Data para o crime envolveria a identificação das fontes de dados disponíveis, como registros policiais, registros de tribunais, pesquisas de vítimas e testemunhas, e outras fontes relevantes. Os dados seriam coletados e armazenados em um formato padronizado, que permitisse a análise em larga escala.

Uma vez que os dados tenham sido coletados e padronizados, uma série de análises poderia ser realizada para identificar padrões e tendências no crime. Por exemplo, pode-se analisar a distribuição geográfica do crime em diferentes regiões do país, identificar os tipos de crimes mais comuns em determinadas áreas, e comparar a taxa de crimes em diferentes estados ou cidades. Os resultados dessas análises poderiam ser compartilhados com as autoridades governamentais, organizações da sociedade civil e outras partes interessadas, a fim de informar políticas e práticas relacionadas à prevenção e combate ao crime e traria maior confiabilidade e precisão aos dados para tomada de decisões estratégicas.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção apresentaremos as principais conceituações teóricas necessárias para a construção deste trabalho. Nesse sentido, alguns trabalhos secundários relacionados ao tema deste artigo foram encontrados.

Em Xing et al. (2018) é implementado um método para obter a representação semântica de artigos jurídicos utilizando *Keywords Group* e redes neurais de convolução. É uma ferramenta que gera uma coleção de palavras que contém algumas palavras-chave em evidência, como homicídio, assassinato, estupro, roubo, são bastante valiosas para os juízes classificarem as evidências.

Em Noor et al. (2015) os autores apresentaram uma revisão sistemática, sob o título "*Análise Inteligente de Dados Aplicada a Dados Governamentais Relacionados a Incidentes Criminais: Uma Revisão Sistemática e abrangente sobre um framework de suporte à tomada de decisão para classificação na prevenção de crime*". Nesta revisão, os autores analisaram quarenta e quatro artigos e classificou-os em duas categorias de crimes - crimes violentos e crimes contra a propriedade - e seis classes de técnicas de mineração de dados - predição, classificação, visualização, regressão, clusterização e detecção de anomalias). Os pesquisadores propuseram um *framework* de classificação utilizando *data mining* e concluíram que as principais técnicas de mineração de dados utilizadas foram *Bayesian*, *Neural Network* e *Nearest Neighbor*.

Já em Rewari; Singh, et al. (2017) os autores realizaram revisões consistentemente nas variadas técnicas, tipos de *datasets*, desafios e problemas encontrados na análise de dados criminais volumosos em estudos existentes, fornecendo aos pesquisadores um resumo do status na área e suas lacunas. Os autores concluíram nesse estudo que a *"revolução da big data"* trouxe uma melhor perspectiva de que, as armas contra os crimes não são mais balas ou bombas, e sim os grandes dados que aprimoram a análise criminal. Segundo Pereira e Galvão (2014) e Berwanger *et al*. (2007), uma revisão sistemática relevante deve apresentar resultados consistentes ou a causa de heterogeneidade deve ser explorada. A maneira mais elaborada de resumir e divulgar os dados é por meio de metanálise, ou seja, uma soma estatística dos resultados de cada estudo.

Segundo a pesquisa realizada por Hinkle et al. (2018), a implementação do sistema CompStat em cidades dos Estados Unidos mostrou-se eficaz para prevenir e combater a criminalidade. O estudo analisou a implementação do sistema "CompStat" em algumas cidades dos Estados Unidos e verificou que a coleta e análise de dados criminais em tempo real permitiu que as forças policiais se antecipassem a tendências criminais e adotassem medidas preventivas.

Conforme apresentado nesta fundamentação teórica, este artigo tem como objetivo explorar a aplicação das técnicas de ciência de dados na análise de dados coletados em diversos formatos. Por meio da utilização dessas técnicas, busca-se transformar os dados brutos em informações confiáveis e valiosas, que possam gerar conhecimento e insights relevantes para o usuário final. Para tanto, serão apresentadas as etapas do ciclo de vida da ciência de dados, que compreendem desde a definição do problema até a visualização dos dados, utilizando exemplos de aplicação prática para cada uma dessas etapas. É importante ressaltar que o ciclo de vida da ciência de dados é um processo iterativo, ou seja, as etapas podem ser repetidas diversas vezes até que se alcance um resultado satisfatório. Ademais, as etapas foram adaptadas de acordo com as necessidades específicas do projeto. Seguem abaixo as definições de cada etapa do ciclo de vida da ciência de dados abordadas no presente estudo:

1. Definição do Problema: Nesta etapa, o objetivo é identificar e definir o problema que se deseja resolver por meio da análise de dados. É importante que o problema seja claramente definido, para que as próximas etapas do ciclo possam ser conduzidas de maneira eficiente.
2. Coleta de Dados: Nesta etapa, o objetivo é coletar os dados necessários para a análise. Os dados podem ser coletados de diversas fontes, como bancos de dados, planilhas, arquivos de texto ou fontes externas como APIs ou web scraping. É importante garantir que os dados sejam relevantes e confiáveis.
3. Descrição dos Dados: Nesta etapa, o objetivo é entender e descrever os dados coletados. Isso envolve a identificação de quais variáveis estão presentes nos dados, qual é a distribuição dos valores, quais são as tendências e padrões observados, bem como quaisquer valores ausentes ou outliers.
4. Preparação dos Dados: Nesta etapa, o objetivo é preparar os dados para análise. Isso pode envolver a limpeza dos dados, preenchimento de valores ausentes, remoção de outliers, normalização ou transformação dos dados, e outras técnicas para garantir que os dados estejam prontos para a análise.

5. Exploração dos Dados: Nesta etapa, o objetivo é explorar os dados para identificar padrões, tendências e relacionamentos entre as variáveis. Isso pode ser feito por meio de técnicas de estatística descritiva, modelagem exploratória ou outras técnicas de análise exploratória de dados.
6. Visualização dos Dados: Nesta etapa, o objetivo é apresentar os dados de forma clara e visualmente atraente. Isso pode envolver a criação de gráficos, tabelas ou outras visualizações que ajudem a comunicar os resultados da análise de forma efetiva.

3. METODOLOGIA

Como forma de alcançar o objetivo proposto neste estudo, iremos implementar técnicas de *análise de dados* em 6(seis) fases para uma melhor perspectiva do problema. Estas seis fases são:

Figura 2 – Ciclo de Vida da Ciênica de Dados

Pipeline mais detalhado da Ciência de Dados demonstrado na Figura:

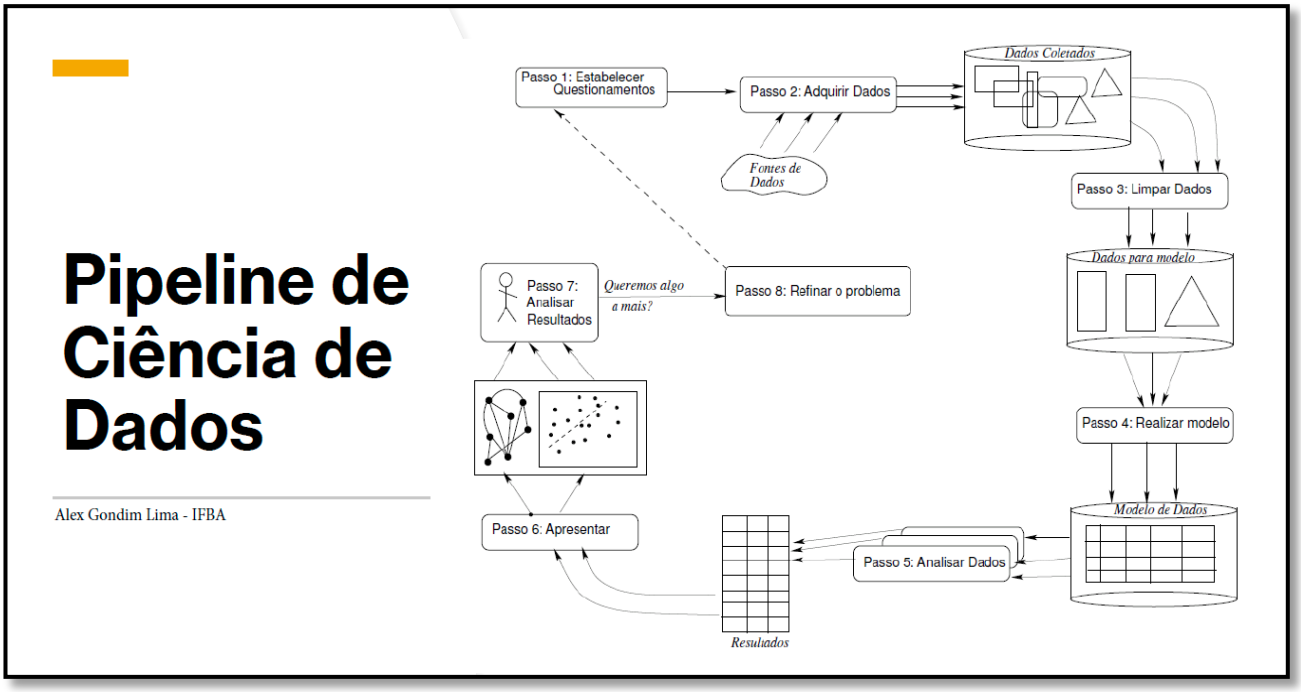

Figura 3 – Pipeline Ciência de Dados

A definição das questões de pesquisa, seu escopo, estratégia de busca e os critérios de seleção serão descritos nas subseções seguintes.

3.1 Ferramentas Utilizadas

No que tange ao ambiente de programação, para obter os resultados esperados, foi utilizado o *Colaboratory* que, como define o *Google*, "é um produto do *Google Research*, área de pesquisas científicas do Google. O Colab permite que qualquer pessoa escreva e execute código Python arbitrário pelo navegador e, é especialmente adequado para aprendizado de máquina, análise de dados e educação".

Para melhor entendimento de quais ferramentas foram utilizadas em cada uma das fases descritas acima, listamos a seguir:

1- Coleta de Dados:, Python , Pandas, Numpy.
2- Modelagem de dados : Pandas, Numpy.
3- Visualização de dados: Matplotlib , Seaborn.
4- Comuniação : Colabority, Jupyter.

3.2 Definição do Problema

Esta etapa tem como objetivo analisar grandes volume de dados estruturados e não estruturados coletados, por diversas fontes e de diferentes formatos, pelos órgãos e forças de segurança pública do Brasil, no que tange aos registros de ocorrências criminais, é o problema a ser solucionado. Além disso, para agravamento dos problemas, muitas vezes os dados coletados nem sempre são totalmente confiáveis. Isto ocorre porquê nem todos os crimes são denunciados e, por isso, a dimensão do problema tende a ser sub-representada.

Diante do exposto, a presente proposta tem em seu escopo a aplicação de Data Science e Analytics, com o intuito de encontrar valores atípicos, sobre o rol de informações criminais, disponíveis nos bancos de dados abertos governamentais implementando técnicas de Análise de Dados Descritiva e Diagnóstica conforme o fluxo a seguir para que assim possamos extrair informações mais confiáveis.

3.3 Coleta de Dados

Foram utilizados 13(treze) Datasets(arquivos de dados) no formato **".CSV"** disponibilizados pelo Sistema Nacional de Estatísticas de Segurança Pública e Justiça Criminal (SINESPJC) entre os anos de 2004 a 2017. Salientando que para cada ano existe um Dataset correspondente.

Além disso o Dataset agrupado contém um total de **375.445** registros e tem a seguinte estrutura de atributos:

a) Regiao: Regiões do Brasil;
b) UF: Sigla da Unidade da Federação ;
c) Estado: Estado da Federação;
d) Crime: Tipo de crime como: Homicídio doloso;
e) Mes: mês de referência dos registros ;
f) Municipio: Espaço público administrado por uma prefeitura;
g) Ano: Ano da ocorrência registrada;
h) Ocorrencias: quantidade de ocorrências registradas;
i) codigo_IBGE : Esse atributo, **codigo_IBGE** não foi interessante para esta análise, portanto foi excluido.

3.4 Descrição dos Dados

Nesta etapa, analisamos os tipos de dados, tuplas e atributos que estamos manipulando e, a propriedade dos dados coletados para resolver o problema. É de fundamental importância entender as características e formatos de cada variável e como identificar as variáveis que possivelmente não serão necessárias na modelagem dos dados.

O tipo de conjunto de dados se encontra no formato string(object) e decimal(float), todos estruturados em tabelas que compõe os dataframes gerados para análise descritiva do presente artigo. Vide Figura 3.1 abaixo:

```
[140] data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 375445 entries, 0 to 66817
Data columns (total 8 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   regiao        375445 non-null  object
 1   uf            375445 non-null  object
 2   estado        375445 non-null  object
 3   municipio     375445 non-null  object
 4   crime         375445 non-null  object
 5   ano           375445 non-null  object
 6   mes           375445 non-null  object
 7   ocorrencias   375445 non-null  float64
dtypes: float64(1), object(7)
memory usage: 25.8+ MB
```

Figura 3.1 – Pipeline Ciência de Dados

O método info(), aqui implementado, nos fornece algumas informações sobre o o conjunto de dados que estamos trabalhando como números de colunas, de linhas. a memória que esse dataframe está utilizando e os tipos de dados utilizados.

3.5 Preparação dos Dados ou Pré-processamento

Nesta etapa, a coleta de dados foi realizada para evitar valores fora do intervalo, combinações de dados impossíveis foram tratados, valores ausentes e redundâncias também foram tratados durante a etapa de pré-processamento de dados, resultando em um conjunto de dados mais confiável e relevante, adequado para a descoberta de conhecimento. Na área marcada na Figura 7 podemos demonstrar exatamente essa fase.

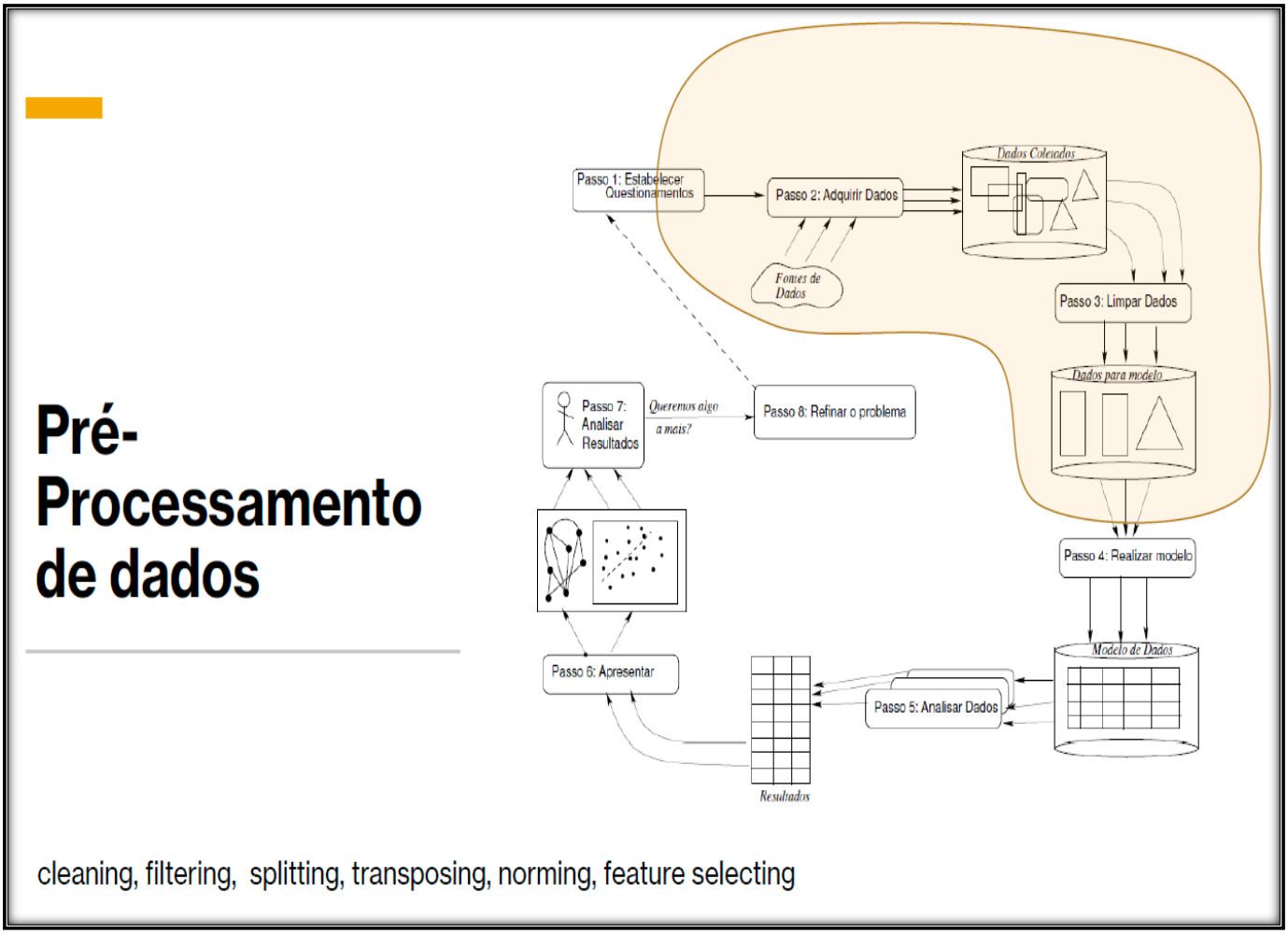

Figura 7 – Pipeline Pré- Processamento de dados

Transformar os dados em formatos adequados para um determinado problema de aprendizado de máquina é uma consideração essencial no início do projeto. A presença de informações irrelevantes, redundantes, ruidosas e não confiáveis afeta significativamente os resultados do modelo e a descoberta de conhecimento, tornando a fase mais difícil.

As etapas de preparação e filtragem de dados levam tempo para uma análise mais precisa, já que devem incluir a limpeza, seleção de instância, normalização, transformação, extração de recursos e seleção. A partir daí temos como produto do pré-processamento de dados o conjunto de treinamento.

3.6 Exploração de Dados

Esta etapa visa conduzir inicialmente uma análise exploratória dos dados, com o objetivo de qualificar a qualidade geral dos dados selecionados para análise. Com os resultados desta qualificação, algumas ações podem ser consideradas, como eliminação de registros de má qualidade ou imputação de valores faltantes. Em suma, a **exploração de dados** significa a análise profunda dos dados em busca de novos insights.

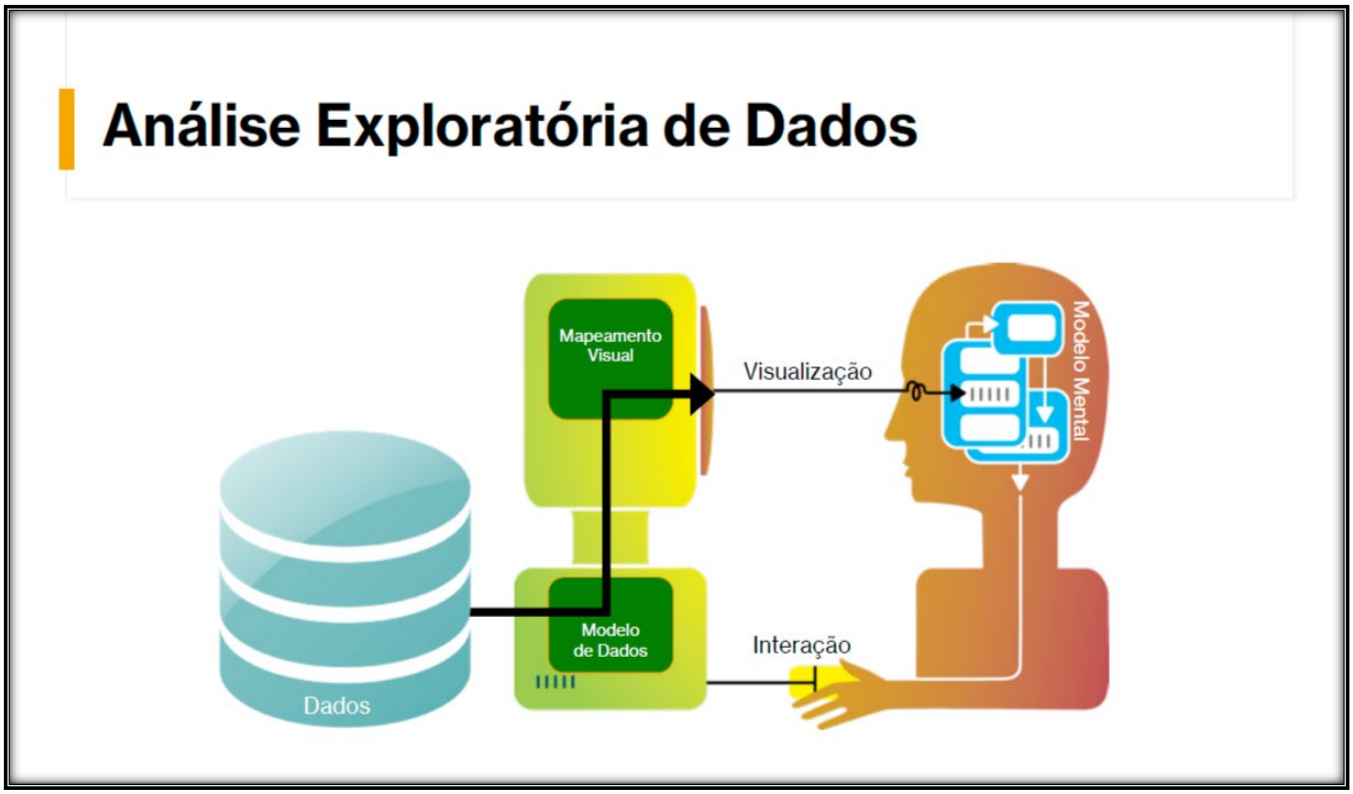

Figura 8 – Pipeline Análise Exploratótia de dados

3.7 Visualização de Dados

Uma das principais vantagens de se utilizar a visualização de dados é que, ela facilita o processo para os usuários entenderem os dados e detectar padrões escondidos. Além disso, a visualização de dados pode ajudar a melhorar a eficiência da tomada de decisão pois, fornece uma representação visual dos dados que pode ser facilmente compreendida.

Nesta fase é realizada uma avaliação dos gráficos e dashboards para verificar relação com problema proposto. Nesse sentido, buscou-se identificar se os dados estão íntegros e consistentes para a próxima etapa, caso necessário como (aplicação de algoritmos de *Machine Learning*).

4. RESULTADOS

Apresentamos neste capítulo a produção bibliográfica para compor este artigo. Como pontos em comum, estas publicações advêm dos trabalhos relacionados e utilizam modelos de processos de análise de dados apresentados na Metodologia, tal como as contribuições da ciência de dados para a segurança pública.

Será feita uma demonstração que utiliza os conceitos gerais de Análise de Dados realizando a simulação de um processo de aquisição e normalização de dados que foram sintetizados por meio de gráficos para apresentar os resultados das análises que tornam os dados mais limpos e confiáveis.

Todo resultado desta análise, foi realizada a partir dos dados disponibilizados pelo Sistema Nacional de Estatísticas de Segurança Pública e Justiça Criminal (SINESPJC) do ano de 2004 a 2017 está armazenada no meu repositório no github, no endereço disponibilizado na seção de Referências Bibliográficas.

Descreveremos a seguir os resultados obtidos, seguindo cada uma das fases que correspondem à ferramenta utilizada na análise dos dados.

4.1 Análise Descritiva

Foi feita uma coleta dos dados brutos, carregando o script desenvolvido na função **read_year_data(year),** abrindo os arquivos de dados no formato ".csv" referente aos períodos de 2004 a 2017 que correspondem a 13(treze) arquivos. Através das funções setadas no scritpt se torna possível analisar cada ano separadamente e, já formata-lo automaticamente quando o arquivo for carregado, através da função format_Columns(data) no intuito de facilitar a análise como resultado mostrado na figura 9.

```python
import glob
import pandas as pd
import numpy as np


def read_year_data(year):
    file = pd.read_csv('./drive/MyDrive/dados/ocorrenciasmun-brasil{}.csv'.format(year), encoding='latin-1',sep=';',skiprows=4)
    return format_columns(file)

def format_columns(data):
    data.columns=[
    "regiao",
    "uf",
    "estado",
    "codigo_IBGE",
    "municipio",
    "crime",
    "ano",
    "mes",
    "ocorrencias"
    ]
    return data
```

Figura 9– Script para carregar os dados brutos

Como o objetivo deste estudo é o de expandir e gerar conhecimento mais abrangente, neste mesmo scritpt consignamos a função **read_all_data()** que carrega todos os dados dos arquivos ".csvs" de ocorrências do Brasil demonstrado na figura 10:

```python
def read_all_data():
    '''A função carrega todos os dados dos csvs de ocorrencias do Brasil
    '''
    files = glob.glob('dados/*.csv')
    years = []
    for file in files:
        data = pd.read_csv(file, encoding='latin-1',sep=';',skiprows=4)
        years.append(data)
    return format_columns(pd.concat(years))

def format_legend(legend_handless):
    legend = []
    for label in legend_handless[1]:
            legend.append(label.split(",")[1][:-1])
    return legend

def create_month_xlegend(months):
    xlegend = ['']
    for month in months:
        if(int(month[1].split('/')[0]) % 2 != 0):
            xlegend.append(month[1])
    return xlegend

def create_year_xlegend(years):
    xlegend = ['']
    for year in years:
        if(int(year) % 2 == 0):
            xlegend.append(year)
    return xlegend
```

Figura 10 – Script para carregar os dados brutos

Nesta etapa utilizamos a função **year.head()** que gerou um dataframe com os dados extraidos do arquivo “ocorrenciasnum-brasil2016.csv” com o intuito de conhecer a estrutura desses dados demonstrado na figura 11 .

Nesta primeira etapa apenas um ano será tratado, para depois expandir a análise para todos os dados disponíveis. O ano escolhido para ser analisado é o ano 2016.

```
[4] year = read_year_data(2016)
```

Carregando o dados do arquivo 'ocorrenciasmun-brasil2016.csv' e gerando Dataframe com os dados formatados para visualização

```
[58] year.head()
```

	regiao	uf	estado	municipio	crime	ano	mes	ocorrencias
0	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	1	01/2016	1.0
1	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	4	04/2016	8.0
2	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	5	05/2016	8.0
3	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	6	06/2016	2.0
4	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	7	07/2016	4.0

Figura 11 – Carregar dados e Apresentar o DataFrame

Os dados tratados neste trabalho foram dados numéricos e dados categóricos, tais como, tipo de Crime, Ocorrências, que podem ser representados por números ou textos. Alguns dados categóricos foram formatados e outros excluídos por não serem relevantes ao objeto desta pequisa.

Inicialmente foi tratado apenas os dados do ano de 2016 e, logo após a pesquisa, foi expandida para tratar todos os registros criminiais disponíveis. Aqui foi realizada algumas limpezas dos dados como a exclusão da **coluna codigo_IBGE,** algumas funções foram construídas para melhor visualização do dataset, para ser reconhecido através da função **head()** conforme a figura 12.

Esse dado catégorico, **codigo_IBGE** não é interessante para esta análise, portanto será excluida

```
[6] del(year['codigo_IBGE'])
```

Vamos carregar o dataframe novamente para visualizar as alterações

```
year.head(5)
```

	regiao	uf	estado	municipio	crime	ano	mes	ocorrencias
0	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	1	01/2016	1.0
1	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	4	04/2016	8.0
2	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	5	05/2016	8.0
3	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	6	06/2016	2.0
4	CENTRO-OESTE	DF	Distrito Federal	Brasília	Estupro	7	07/2016	4.0

Figura 12 –Limpar Dados e Apresentar Dataframe

Logo após a estrutura básica ser conhecida, foram verificadas informações sobre as estatísticas gerais destes dados que é possível pela implementação da função **describe()** que, por padrão, escolhe apenas os dados numéricos, retornando a contagem de linhas (count), o cálculo da média (mean) e desvio padrão (std) e identifica o valor mínimo (min), os quartis (25%, 50% e 75%) e o valor máximo (max), conforme figura abaixo:

```
[8] year.describe()
```

	ano	ocorrencias
count	66818.000000	66818.000000
mean	6.514786	7.195771
std	3.430935	30.081783
min	1.000000	1.000000
25%	4.000000	1.000000
50%	7.000000	2.000000
75%	9.000000	4.000000
max	12.000000	874.000000

Figura 13 – Utilização da função describe()

4.1.2 Exploração e Visualização de Dados

Com estas estatísticas engendradas na fase de Preparação dos Dados, fez-se necessário elaborar as perguntas cruciais para prosseguir com as análises. Ao observar as informações geradas, surgiram as seguintes indagações:

- Qual o tipo de crime tem a maior quantidade de ocorrência no ano de 2016?
- Quais meses têm a maior taxa de cada um dos crimes no conjunto de dados?
- Quais cidades apresentaram as maiores quantidades de cada tipo de crimes presentes nos dados?

A partir daqui que começa a fase exploratória dos dados, que será analisada, pelo conjunto de dados disponíveis, os crimes mais frequentes do ano de 2016 e depois será feita a normalização(tratamento) desses dados. Utilizamos aqui uma função **Groupby()** para agrupar os demais dados pelo atributo 'crime', e a soma das ocorrências associadas ao agrupamento como se vê na figura 14.

```
crime = year.groupby(['crime'])
occurrences = crime.agg({'ocorrencias':sum})
occurrences
```

crime	ocorrencias
Estupro	38,363.00
Furto de veículo	209,129.31
Homicídio doloso	46,461.00
Lesão corporal seguida de morte	715.00
Roubo de veículo	184,030.71
Roubo seguido de morte (latrocínio)	2,108.00

Figura 14 – Utilização da função groupby()

Nessa análise tivemos uma noção geral da quantidade de ocorrências para cada tipo de crime, que nos traz uma perspectiva mais ampla de quais foram os crimes mais cometidos durante esse período.

Agora a análise será feita as taxas de crimes (2016) com seus respectivos meses. Para obter esta informação, os dados serão agrupados primeiro, levando em consideração os meses e, então, a contagem das ocorrências para cada tipo de crime conforme a figura 15.

```
[92] crimes_per_month = year.groupby(['mes','crime']).agg({'ocorrencias':sum})
     crimes_per_month.head()
```

		ocorrencias
mes	**crime**	
01/2016	**Estupro**	2,669.00
	Furto de veículo	13,461.00
	Homicídio doloso	3,983.00
	Lesão corporal seguida de morte	58.00
	Roubo de veículo	13,054.63

Figura 15 – Utilização da função Groupby()

No caso desses dados existem muitos valores, o que gerou a necessidade de criar um “plot” que facilite a visualização e entendimento dos dados onde utilizamos a função “Plot” da biblioteca do Pandas gerar o gráfico de Barras conforme a figura 16 e figura 17.

```
[75] crime_per_month_bar_graph = crimes_per_month.unstack(level=0)
     crime_per_month_bar_graph.plot(kind="barh", figsize=(10,25), subplots=True, legend=False, title="Criminalidade por mês em 2016")
```

Figura 16 – Utilização da função Plot – Apresentação de Gráfico

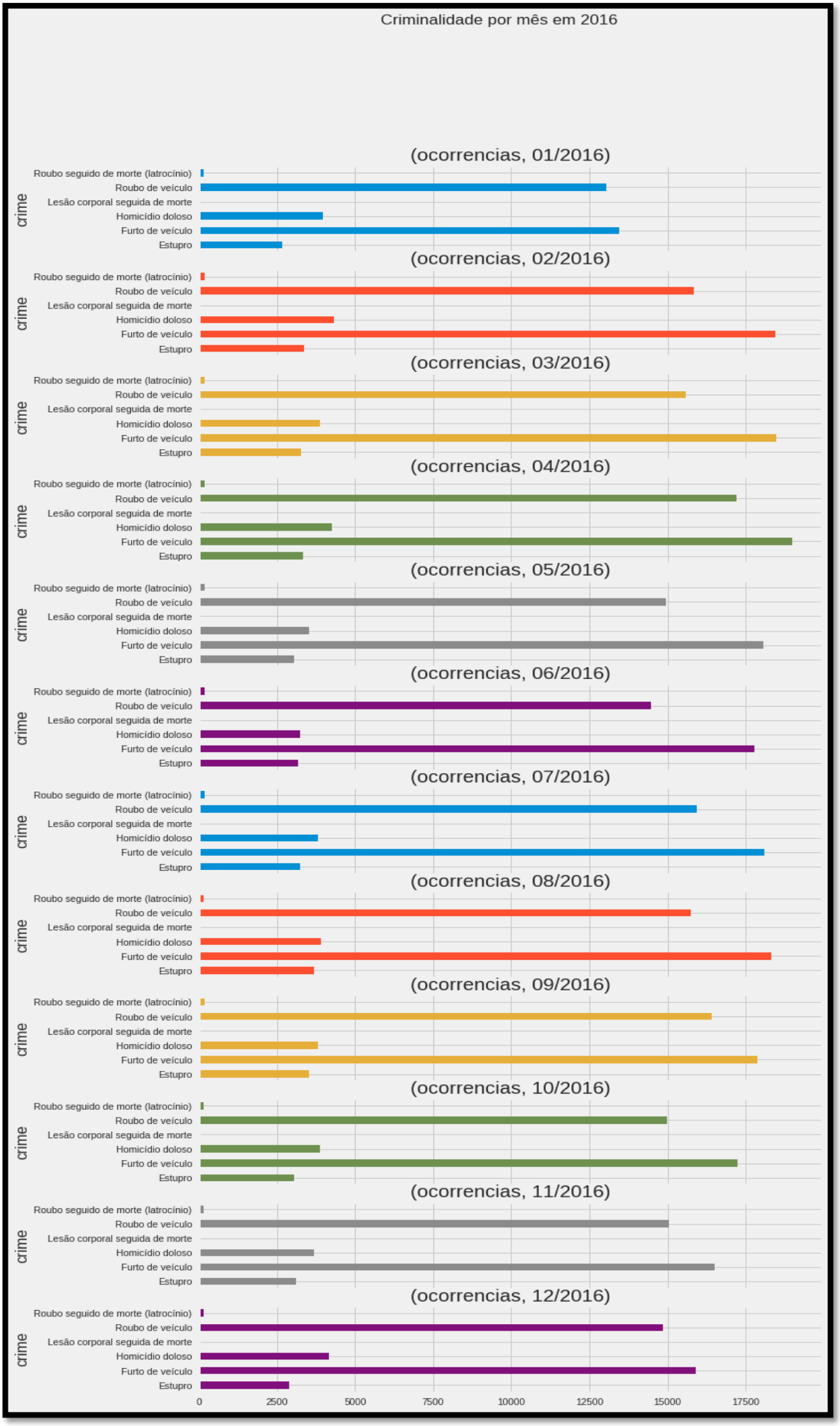

Figura 17 – Utilização da função Plot – Apresentação de Gráfico

Analisamos também quais cidades tiveram os maiores índices de cada um dos tipos de crime, no ano de 2016. Mais uma vez trabalhamos com funções estátisticas da biblioteca do Pandas da Linguagem Python demonstrada por **dataframe e por gráficos** conforme figura 18 e figura 19.

```
[14] crimes_per_city = year.groupby(['municipio','crime'], as_index=False).agg({'ocorrencias':sum})
     crime_per_city_max = crimes_per_city.groupby(['crime']).agg({'ocorrencias':max})
     criteria = ['ocorrencias','crime']
     crime_per_city_graph = pd.merge(crime_per_city_max,crimes_per_city,on=criteria)
     crime_per_city_graph
```

	ocorrencias	crime	municipio
0	2131.0	Estupro	São Paulo
1	7553.0	Furto de veículo	Belo Horizonte
2	1277.0	Homicídio doloso	Salvador
3	36.0	Lesão corporal seguida de morte	Natal
4	8128.0	Roubo de veículo	Porto Alegre
5	107.0	Roubo seguido de morte (latrocínio)	São Paulo

Figura 18 – Utilização da função Groupby –DataFrame

```
[54] crime_per_city_graph.plot(kind='barh',y='ocorrencias',x='municipio',  figsize=(10,5))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa0fd76f970>
```

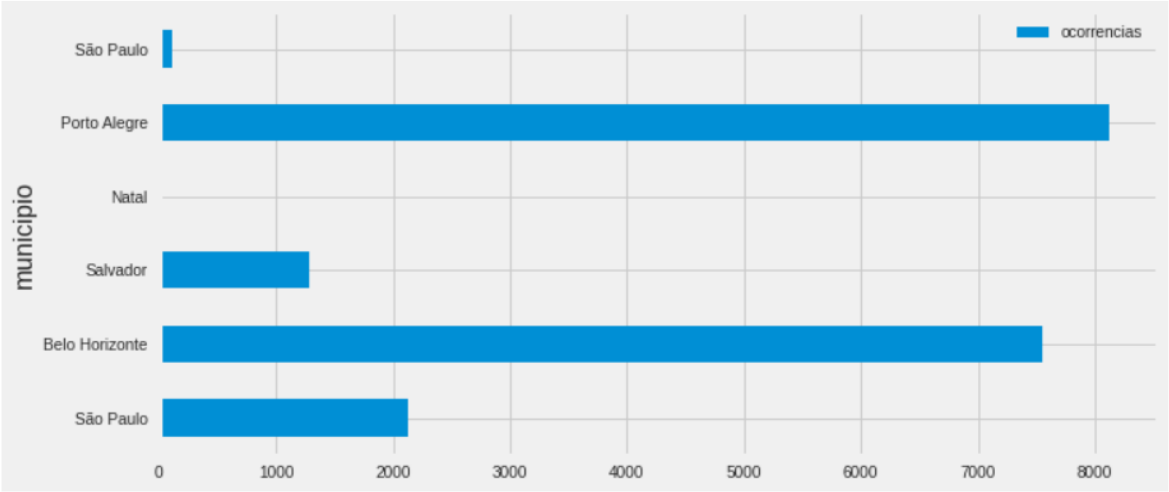

Figura 19 – Utilização da função Plot – Apresentação de Gráfico

4.2 Análise Diagnóstica

Aqui foi feita uma abordagem importante para identificar quais regiões necessitam de maior atenção e investimento em medidas de prevenção e segurança, contribuindo para a redução da criminalidade nos fornecendo a possibilidade de diagnosticar os problemas relacionados à incidência de crimes em diferentes regiões do Brasil.

4.2.1 A expansão para Análise de Séries Temporais

Diante aos dados de um único ano, obtivemos diversas informações importantes e descobrimos alguns **insghts.** A análise agora é no intuito de entender os fenômenos dos dados ao decorrer dos anos. Esta avaliação será possível já que o **SINESPJC** disponibiliza uma base de dados de 2004 até 2017.

Utilizaremos a função **read_all_data()** para carregar todos os dados. Essa função está encapsulada no script inicial desse Projeto.

```
[17] data = read_all_data()
```

Como explicado anteriormente, a coluna codigo_IBGE não é necessária, portanto será excluída novamente

```
[18] del(data['codigo_IBGE'])
```

Vizualisando novamente a sua estrutura depois de alterada. Este passo é importante para o entendimento dos dados.

```
[19] data.head()
```

	regiao	uf	estado	municipio	crime	ano	mes	ocorrencias
0	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	1	01/2004	1.0
1	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	3	03/2004	2.0
2	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	5	05/2004	4.0
3	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	6	06/2004	2.0
4	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	7	07/2004	1.0

Figura 20– Função Carrega os 13 arquivos ".csv" – Apresentação em unico DataFrame

Quando carregamos os dados, é fácil notar que eles não possuem uma coluna com o valor dos anos, portanto esta foi criada a partir da coluna mês com o comando, como se vê na figura 21.

```
[21] data['ano'] = data['mes'].apply(lambda element: element.split("/")[1])
```

Para confirmar que a coluna foi alterada os dados foram visualizados novamente

```
[22] data.head()
```

	regiao	uf	estado	municipio	crime	ano	mes	ocorrencias
0	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	2004	01/2004	1.0
1	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	2004	03/2004	2.0
2	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	2004	05/2004	4.0
3	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	2004	06/2004	2.0
4	SUDESTE	ES	Espírito Santo	Cachoeiro De Itapemirim	Estupro	2004	07/2004	1.0

Figura 21 – Cruzamento de dados – Criação de Colunas

Observando o DataFrame nota-se que há uma variação muito grande nos meses e anos e, ainda, que a estrutura base é a mesma dos dados vistos para um único ano (2016).

Nesta fase realizamos uma nova etapa de análises na qual foi necessário criar uma linha do tempo, variando com a quantidade total de ocorrências em cada um dos anos presentes no conjunto de dados, onde foi aplicado apenas uma função para gerar uma coluna com apenas os valores dos anos e, os dados foram agrupados pelos anos para realizar a contagem conforme Figura 22.

```
[22] crime_occurrences = data.groupby(['ano'],as_index=False).agg({'ocorrencias':sum})
```

```
crime_occurrences_xlegend = create_year_xlegend(crime_occurrences.ano)
crime_occurrences.plot(kind="line", x='ano', y='ocorrencias', figsize=(10,6)).set_xticklabels(crime_occurrences_xlegend,rotation=0)
crime_occurrences.plot
```

```
<pandas.plotting._core.PlotAccessor object at 0x7fa10704a880>
```

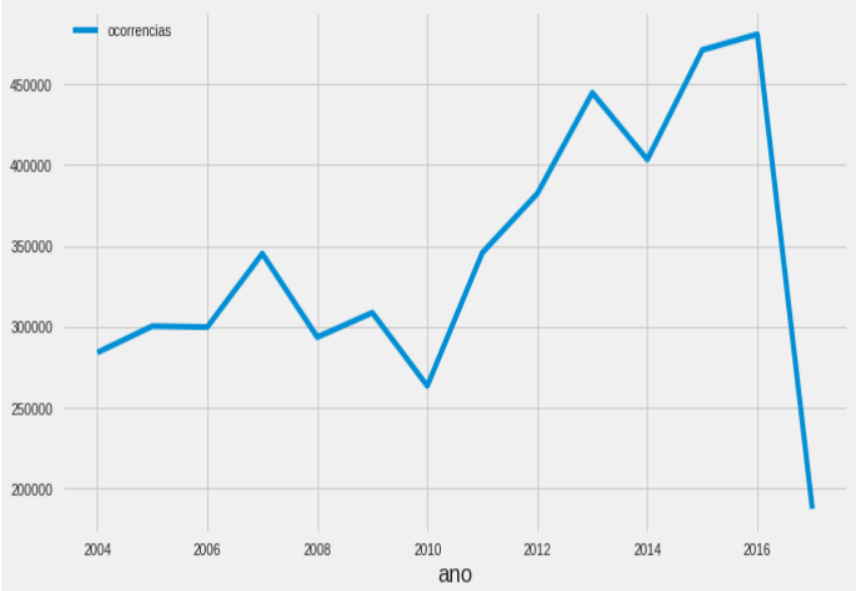

Figura 22 – Agrupamento de todos os conjuntos de dados

Depois da análise anterior evoluímos para uma análise mais abrangente sendo que foi preciso implementar novamente a função Groupby() que nos permite conseguir visualizar a variação no tempo de cada um dos tipos de crimes através de um “plot” conforme se vê na Figura 23.

```
[55] crimes_in_history = data.groupby(['crime','ano']).agg({'ocorrencias':sum})
     crimes_in_history_line_graph = crimes_in_history.unstack(0).plot(kind="line",legend=True, subplots=False,figsize=(10,5))
     legend_history = format_legend(crime_per_month_line_graph.get_legend_handles_labels())
     crimes_in_history_line_graph.legend(legend_history,loc="lower left", bbox_to_anchor=(1, 0.5))
     crimes_in_history_xlabel = create_year_xlegend(crimes_in_history.unstack(0).index)
     crimes_in_history_line_graph.set_xticklabels(crimes_in_history_xlabel)
     crimes_in_history_line_graph.plot
```

```
<bound method Axes.plot of <matplotlib.axes._subplots.AxesSubplot object at 0x7fa0fd759430>>
```

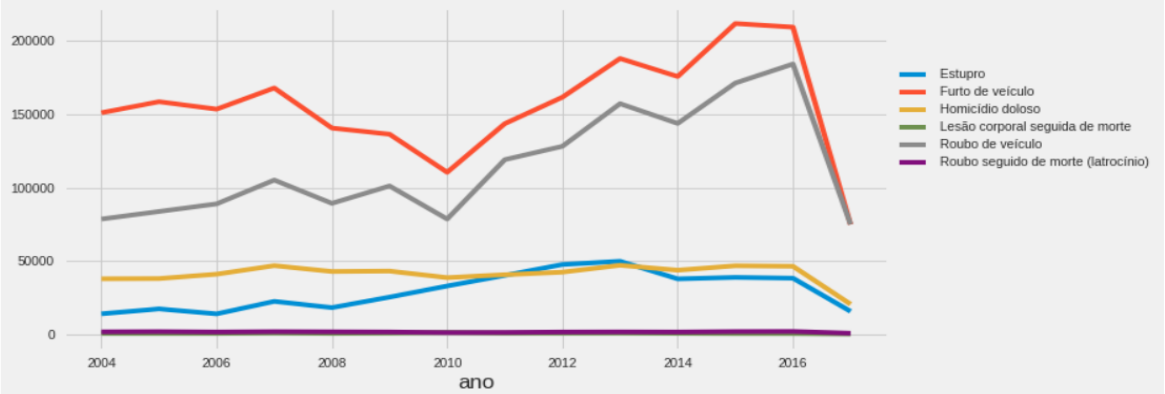

Figura 23 – Agrupamento de Dados e Visuzalização

Para visualizar um cenário mais amplo, geramos um gráfico com os anos que apresentaram os maiores números de ocorrências para cada um dos crimes, conforme a Figura 24.

```
[25] max_occurrences_by_year = data.groupby(['crime','ano']).agg({"ocorrencias":sum}).unstack(1).sum()

max_occurrences_by_year.plot(kind="barh", figsize=(10,6), color = color)

<matplotlib.axes._subplots.AxesSubplot at 0x7fa1029a9d00>
```

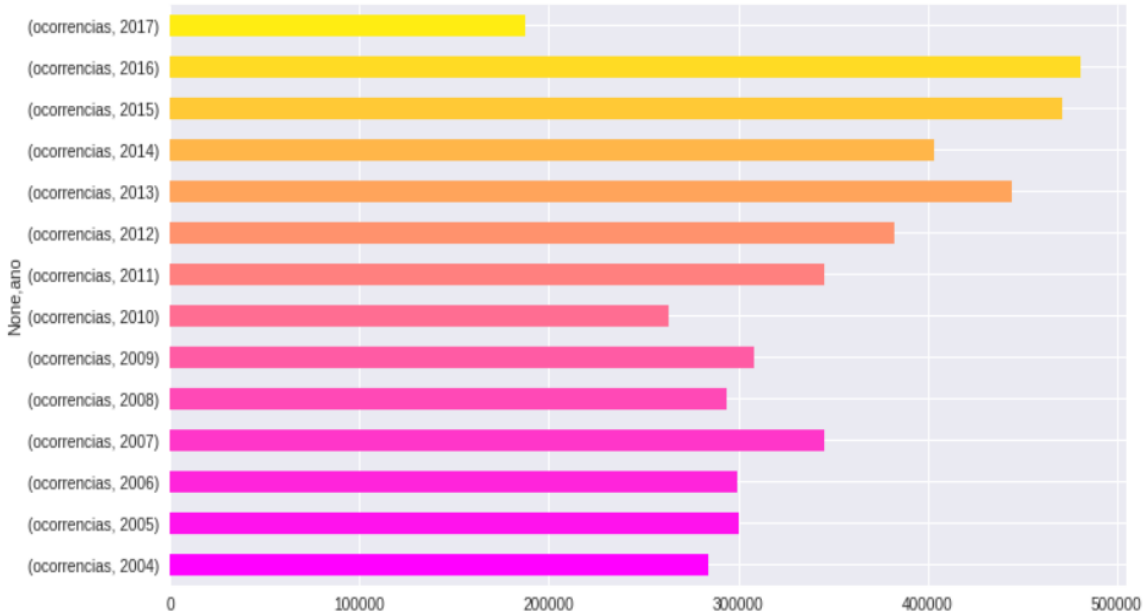

Figura 24 – Agrupamento de dados – Visualização

5. DISCUSSÃO

Diante da análise, os resultados mostram que ao mesmo tempo em que temos um grande volume de dados gerados, boa parte desses dados não recebem o devido tratamento, embora as tecnologias emergidas pela ciência de dados tenham potencial suficiente para lidar com muitos dados, assim como demonstramos com os resultados aqui apresentados.

A partir da demonstração ora apresentada neste estudo, podemos afirmar que, essas tecnologias exigem cada vez mais capacidade de processamento para a execução dessas tarefas. Nesse sentido iremos abordar alguns pontos dos resultados obtidos.

Os primeiros passos foram tratar, limpar e normalizar os dados. Em seguida dentro dos 13 arquivos correspondentes a cada ano no conjunto de dados, entre o período de 2004 a 2017, escolhemos o ano de 2016 para darmos início a primeira análise propriamente dita, no qual as funções estatísticas implementadas nos apresentaram um quadro em que nos mostra que neste ano o furto de veículos foi muito alto, o Roubo de Veículos ficou na segunda posição e na terceira posição o crime de Homicídio doloso, conforme a Figura 25. Sendo que a menor ocorrência foi Lesão Corporal seguida de morte.

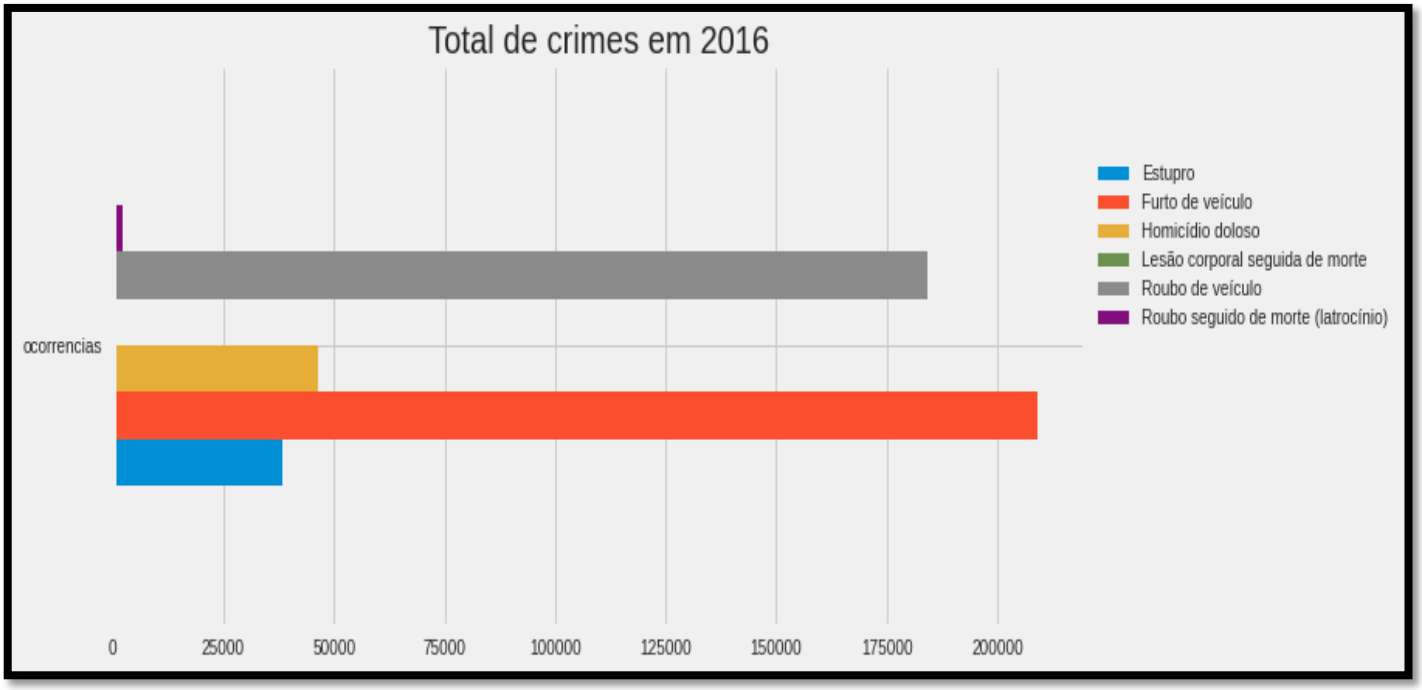

Figura 25 – Total de crimes x tipo de crime

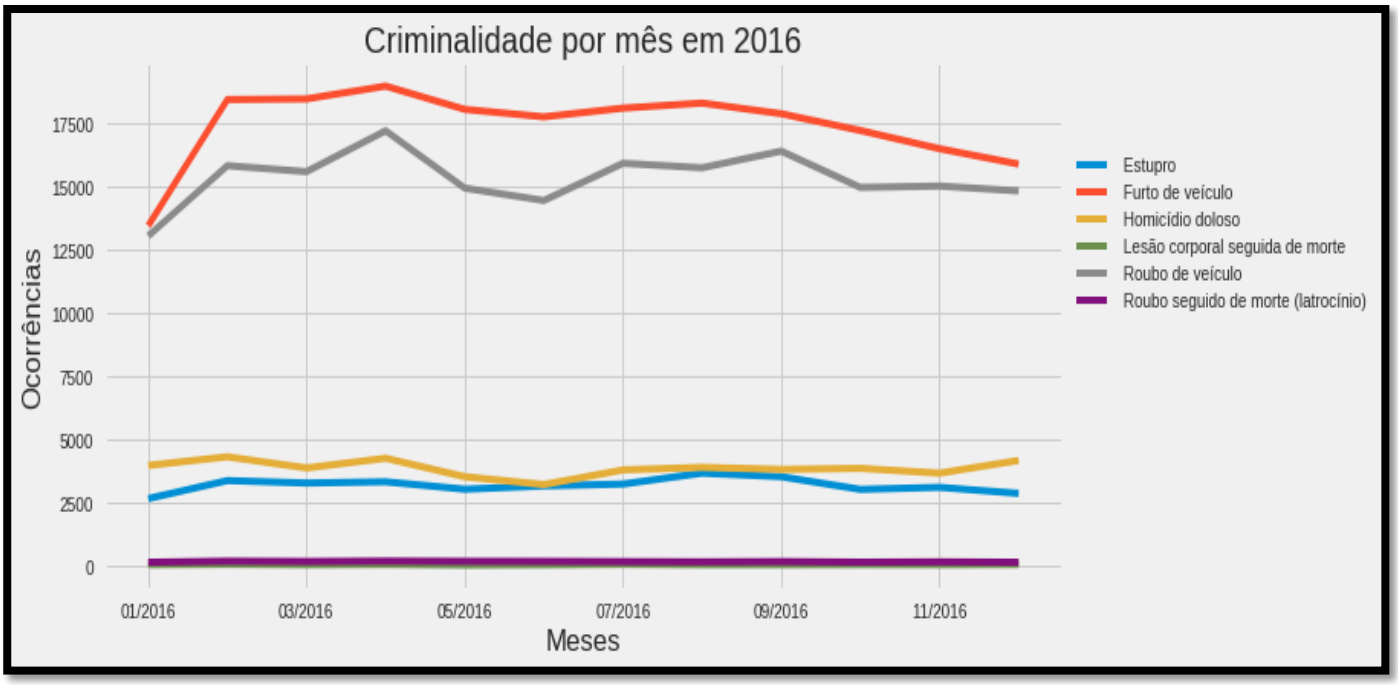

Figura 26 – Quantidade de crimes ao longo dos meses x tipo de crime

O gráfico nos mostra também que existe padrão na distribuição dos dados, onde certos tipos de crimes apresentam valores muito próximos. É possível perceber, por exemplo, que o furto de carro ao longo de todo o ano de 2016 foi muito alto, além de crimes de Latrocínio e estupro, que mantiveram as praticamente as mesmas quantidades de ocorrências durante quase o ano todo.

No terceiro passo foi analisado quais cidades tiveram os maiores índices para cada crime, no ano de 2016. Porto Alegre foi a cidade que registrou maiores ocorrênicas totais. A cidade com maior registro de estupro foi São Paulo e, o furto de veículo teve as maiores ocorrências em Belo Horizonte. A cidade de Salvador ficou em primeiro lugar em registro de crime de Homicídio Doloso.

No quarto passo no que tange aos estados, São Paulo é o estado que mais registrou crimes, seguido por Minas Gerais, Paraná e Bahia . Isso indica que essas regiões brasileiras são as mais vulneráveis à violência, como se vê no gráfico abaixo :

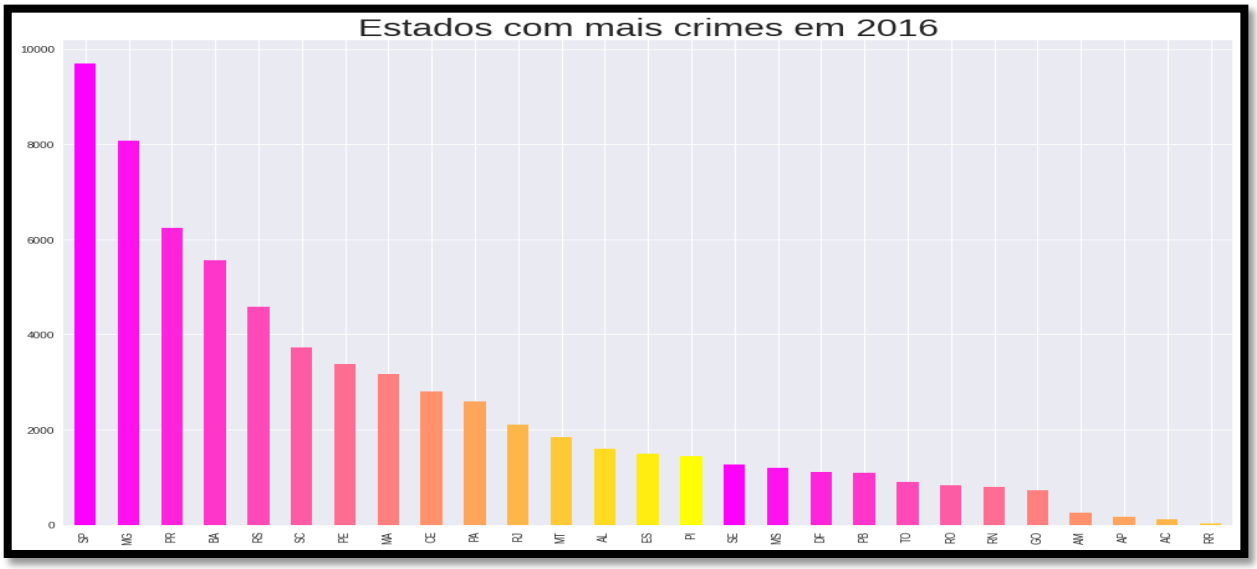

Figura 27 – Varição de Crimes por Estado da Federação.

No quinto passo, partimos para uma expansão da análise consginada na seção anterior na subseção 4.5. Os resultados obtidos nos mostram que há uma varição grande nos meses e anos e, ainda, que a estrutura base é a mesma dos dados vistos para um único ano (2016).

Observa-se através do gráfico na figura 24, que para todos os crimes, os maiores índices estão entre 2011 e 2016, o que demonstra que este período de tempo representaram muito mais violência que qualquer outro desde 2004, tendo uma queda em 2017. A queda das ocorrênicas em 2017, pode estar vinculada com os dados que estão subnotificados.

No sexto passo conseguimos visualizar a variação no tempo de cada um dos tipos de crimes e a variação ao longo dos anos, sendo os que mais variaram ao longo do tempo analisado foram o Furto, Roubo, Estupro e Homicídio doloso, conforme a figura 28.

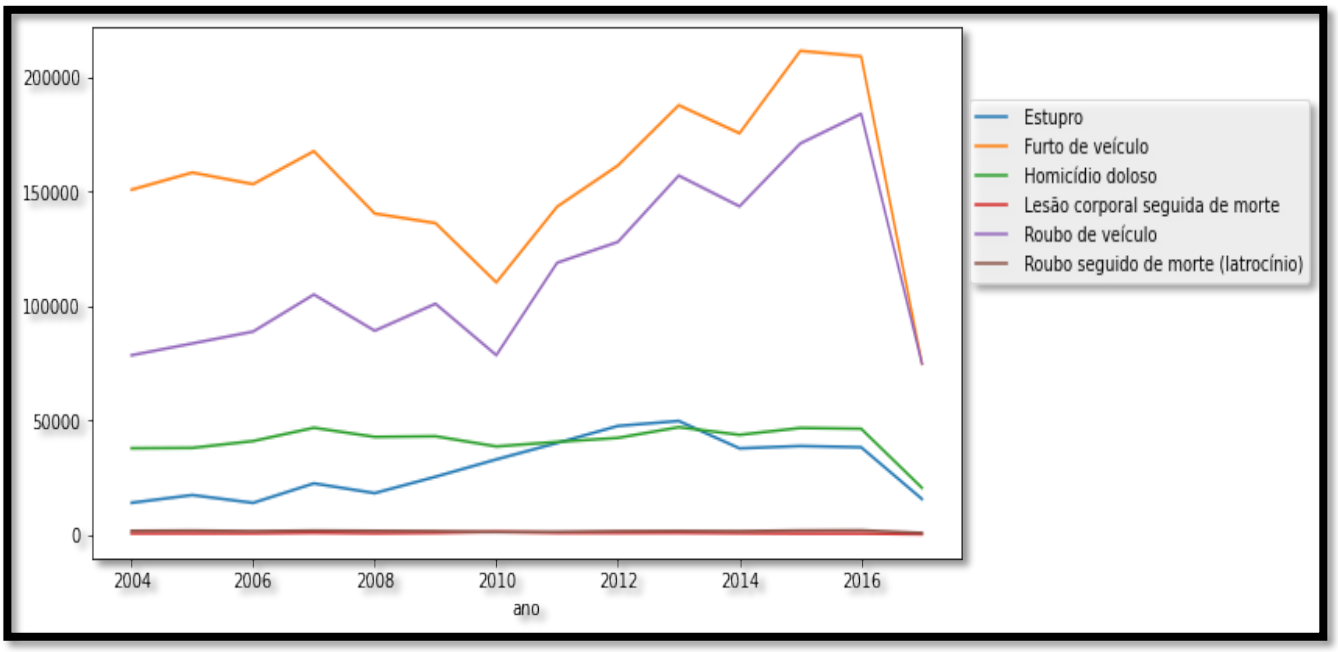

Figura 28 – Variação da quantidade de crimes ao longo dos anos.

É importante nos atentarmos aos crimes de lesão corporal e latrocínio, que em todos os resultados apareceram em baixa, este pode ser um dado real ou pode estar ocorrendo uma falta de dados, o que é gerado pela falta de padronização da coleta de dados pelos órgãos de segurança pública. Além do que, esses dados podem ser manipulados para esconder o tamanho real do problema. É importante que haja mais transparência e padronização na coleta de dados sobre esses tipos de crimes, para que possamos avaliar o tamanho do problema e tomar as medidas necessárias para combatê-lo.

6. Conclusão e Trabalhos Futuros

Este artigo descreve como a ciência de dados pode ser usada para auxiliar as forças de segurança publica no combate ao crime. Com tecnologia como o sistema de Análise de Dados, organizações criminosas podem ser prejudicadas e políticas públicas mais eficazes podem ser implementadas, melhorando assim os esforços de combate ao crime. Além disso, as ferramentas que o campo da ciênica de dados oferece, tem um alto potencial para subjugar a concentração da ocorrência de crimes em áreas infestadas de crimes específicos. Assim sendo, tendo uma detecção de crimes, o processo de normalização de dados agora pode ser automatizado para se tornar mais eficiente e eficaz.

Tendo essas informações podemos comparar com a conjuntura das políticas de segurança, que foram efetivas ou não, nos anos relacionados nesse conjunto de dados. Essas conjunturas estão resumidas nos relatórios do Atlas da Violência que é um portal que reúne, organiza e disponibiliza informações sobre violência no Brasil, bem como reúne publicações do Ipea sobre violência e segurança pública ao longo dos anos. Assim, poderemos compreender melhor o aumento ou diminuição de cada crime, conforme políticas acertadas ou erradas. O plano de melhoria futura para o Sistema de Análise de dados, perpassa em buscar meios mais eficazes na coleta dos registros criminais no Brasil pelos órgão de segurança pública, para que assim possamos expandir os recursos de análise, não somente a dados númericos, mas também dados de textos, imagens e/ou vídeos.

Atualmente, as formas de análise e os dados divulgados divergem, o que pode comprometer a tomada de decisões e a otimização do trabalho das forças policiais. Um exemplo de divergência na coleta de dados criminais no Brasil pode ser observado na contagem de homicídios. O país possui duas fontes de informações oficiais sobre o tema: o Sistema de Informações sobre Mortalidade (SIM), do Ministério da Saúde, e o Anuário Brasileiro de Segurança Pública, do Fórum Brasileiro de Segurança Pública. Essas duas fontes de dados apresentam números diferentes sobre o número de homicídios no país, o que dificulta a obtenção de informações precisas e pode comprometer a eficácia das políticas públicas de segurança. Enquanto o SIM contabilizou cerca de 50 mil homicídios em 2019, o Anuário Brasileiro de Segurança Pública registrou mais de 41 mil casos no mesmo período.

Essa diferença pode ser explicada por questões metodológicas e pela forma como as informações são coletadas e processadas em cada fonte de dados. O SIM se baseia em dados fornecidos por cartórios de registro civil, enquanto o Anuário Brasileiro de Segurança Pública se apoia em informações repassadas pelas secretarias de segurança pública dos estados.

Por isso, é importante implementar um Sistema Nacional de Coleta de Dados de Registros Criminais no Brasil, para garantir a padronização da coleta de informações sobre crimes e criminosos em todo o país. Com esse sistema, os estados e municípios poderiam coletar informações sobre crimes e acusados de forma padronizada, permitindo o cruzamento de dados e a compilação em tempo real. Além disso, com o advento do big data, o Sistema Nacional de Coleta de Dados de Registros Criminais poderia se tornar uma poderosa ferramenta para a análise e previsão de tendências criminais, auxiliando no desenvolvimento de estratégias mais eficazes de combate à criminalidade. Ademais, um aplicativo com recursos como mapas e visualização de dados poderia aprimorar ainda mais a qualidade das informações e facilitar a disseminação dos dados coletados. Em suma, é imprescindível investir em uma coleta de dados eficiente e confiável para aprimorar o combate à criminalidade em todo o país e aproveitar todo o potencial da big data para esse fim.

REFERÊNCIAS BIBLIOGRÁFICAS

BECKER, J. L. **Sem dados, combate ao crime é 'míope', diz professor da FGV,2021.** [Entrevista concedida a] Anais Fernandes. Valor Econômico, São Paulo, ano 2021. Disponível em < **https://valor.globo.com/brasil/noticia/2021/05/24/sem-dados-combate-ao-crime-e-miope-diz-professor-da-fvg.ghtml**>. Acesso em: 14/09/2021.

GAMAGE, P. **New development: Leveraging 'big data' analytics in the public sector**. Public Money & Management, [s.l.], v. 36, n. 5, p.385-390, 10 jun. 2016. Informa UK Limited. Disponível em: <http://dx.doi.org/10.1080/09540962.2016.1194087>. Acesso em: 15 abril. 2021.

LAURA SEBASTIAN-COLEMAN, **in Measuring Data Quality for Ongoing Improvement**, 2013. Citada na página 4.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis . Pearson Prentice Hall, 2008.**

NETO, J. V. et al. Boxplot: **um recurso gráfico para a análise e interpretação de dados .quantitativos [Boxplot: A graphic resource for the analysis and interpretation of quantitative data**]. Revista Odontológica do Brasil Central, v. 26, n. 76, 2017.

PRADO, K. H. d. J.; COLAÇO JÚNIOR, M. **Data science applied to crime analysis based on brazilian open government data.** *Journal of Applied Security Research***,** Taylor & Francis, 2020.

RICCIO, V. *et al*. **A utilização das estatísticas criminais no planejamento da ação policial: cultura e contexto organizacional como elementos centrais à sua compreensão.**Scielo Brasil, 2012. Disponível em:< https://www.scielo.br /j/ci/a/VrGbrXwwf3V7GBxC9FBWqpQ/?lang=pt#/>. Acesso em: 14/09/2021.

ROUSSEEUW, P. J. **"Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis"**. Computational and Applied Mathematics. 20: 53-65. Disponível em: https://www.sciencedirect.com/science/article/pii/0377042787901257?via%3Dihub . Acesso em 14 de set. 2021.

DONOHO, D. **50 Years of Data Science. Journal of Computational and Graphical Statistics, Taylor & Francis, v. 26, n. 4, p. 745–766, 2017.ISSN 15372715.**

PEREIRA, M. G.; GALVÃO, T. F. **Heterogeneidade e viés de publicação em revisões sistemáticas. Epidemiologia e Serviços de Saúde, SciELO Public Health, v. 23, p. 775–778, 2014.** Citado na página 7.

HINKLE, J. C., WEISBURD, D., FAMEGA, C., READY, J., & MADENSEN, T. (2018). **How does CompStat produce organizational change? A study of three police departments. Crime & Delinquency, 64(1), 59-85**. https://doi.org/10.1177/0011128716684377.Citado na página 8.