Desvendando crimes na cidade de Boston

Kauê Cabrera Rosalem
Business School São Paulo
São Paulo / SP, Brasil
kauerosalem@yahoo.com.br

***Resumo*—Neste trabalho investigamos estatisticamente as frequências de crimes no distrito policial B2 da cidade de Boston, por hora do dia, dias da semana e dias do mês. Também estudamos os crimes por grau de risco através de um modelo preditivo pelo método de regressão logística. Por fim, analisamos os locais de crimes deste distrito e definimos agrupamentos através de um modelo descritivo pelo método de k-médias.**

***Abstract*— In this work we statistically investigated the crimes frequency in Boston's B2 police district, by hour of the day, days of the week and days of the month. We also studied crimes by risk degree taking a predictive model by the logistic regression method. Finally, we analyzed the crime locations in this district and defined clusters taking a descriptive model by the k-means method.**

Palavras-chave—ciência de dados, mineração de dados, regressão logística, matriz de confusão, curva de Elbow, agrupamento por k-médias, crimes, Boston.

I. INTRODUÇÃO

A polícia da cidade de Boston deseja otimizar o uso de recursos humanos e contratou seu departamento para auxiliar na tomada de decisões para alocação de seus policiais no distrito B2 e traçar novas estratégias pautadas em inteligência de dados. Por contar com uma estrutura moderna de monitoramento, a corporação formulou uma base de dados com mais de 65 mil registros de ações ofensivas. Através destes registros, este trabalho foi desenvolvido em três fases de intervenção.

A fase 1 consistiu em preparar um plano de ação indicando quais os principais turnos e horários em que a corporação deve se atentar para ocorrência de crimes. Para compreensão dos dados, a frequência de ocorrências criminosas foram estudas, e as médias de crimes foram calculadas por horas do dia, dias da semana e meses do ano. Além disto, verificamos os momentos de maior e menor criminalidade nestes intervalos de tempo para o distrito B2.

A fase 2 teve como objetivo otimizar as operações policiais do distrito B2, com ações mais assertivas para as equipes que serão deslocadas pelas ruas da cidade. Desta forma, a base de dados foi utilizada para prever os crimes que ocorreram e que afetaram apenas o distrito B2 (9823 registros). Esta etapa consistiu em identificar e implementar uma estratégia que fosse capaz de identificar novos crimes que sejam de alto risco (1) ou de baixo risco (0) com base nas características apresentadas. Nesta etapa, é proposto um modelo preditivo baseado em registros binários, o qual é chamado por regressão logística.

A fase 3 foi proposta uma nova ferramenta de análise de dados e de identificação de possíveis organizações criminosas que afetam o distrito B2. Atualmente, a corporação utiliza um conjunto de regiões distintas para tentar agrupar os crimes que ocorrem por localização (latitude e longitude) da ocorrência. No entanto, a estratégia identificada pela equipe policial é bastante exaustiva e a corporação deseja encontrar formas automatizadas de gerenciar a demarcação territorial para os crimes que ocorreram. Um local de ocorrência pode ser definido por um par coordenado, que geograficamente sem nenhum outro contexto, é formado por latitude e longitude. Logo, para facilitar esta visualização geográfica de dados, o método de agrupamento por k-médias é proposto e aplicado aos locais em que há ocorrências de crimes.

II. MATERIAIS E MÉTODOS

A. *Técnicas Utilizadas*

Na fase 1 do projeto, utilizou-se medidas centrais estatísticas para a análise de dados, por exemplo, medidas de média aritmética para o número de crimes por horas do dia, dias da semana e meses do ano. Em estatística, uma tendência central é um valor central ou valor típico para uma distribuição de probabilidade.

Na fase 2 do projeto, estudou-se os crimes por grau de risco através de um modelo preditivo com o uso de um algoritmo com base no método de regressão logística. A regressão logística é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento diante de um conjunto de variáveis explanatórias.

Na fase 3 do projeto, analisou-se os locais de crimes do distrito B2 através de um modelo descritivo com o uso de um algoritmo com base no método de agrupamento por k-médias. Em mineração de dados, agrupamento por k-médias é um método de *clustering* que objetiva particionar um número de observações dentre um número de grupos, onde cada observação pertence ao grupo mais próximo da média. O algoritmo deste método treina um modelo para agrupar observações semelhantes. Para isso, ele mapeia cada observação no conjunto de dados de entrada para um ponto no espaço de n dimensões, ou seja, o número de atributos da observação.

B. *Estratégias de Validação*

Para validar os valores apresentados na base de dados, apurou-se os o número e o formato dos registros em cada atributo, verificou-se a existência de valores nulos e inconsistentes com os demais conjuntos de dados. Portanto, os dados foram pré-processados e reorganizados antes de se realizar as análises exploratórias. As medidas centrais estatísticas de dados possibilitaram a compreensão de registros com maior ou menor destaque para um dado atributo.

C. *Premissas*

Este trabalho apresenta as seguintes premissas: identificar um conjunto de requisitos associados ao estudo de mineração

de dados em um contexto de crimes urbanos; definir um conjunto de métricas estatísticas para apoiar uma análise estruturada em mineração de dados; especificar uma técnica de classificação binárias que seja adequada ao contexto de estudo; especificar uma técnica de agrupamento que permita a formação de *clusters* baseados em geolocalização; aplicar técnicas de mineração de dados, pré-processamento e de análise estatística em uma base de dados estruturada; aplicar técnicas de classificação binárias em uma base de dados estruturada; aplicar técnicas de agrupamento em uma base de dados estruturada; analisar os principais resultados obtidos das aplicações de técnicas de mineração de dados; compor um relatório condensando os principais resultados obtidos ao longo de suas produções.

D. Resultados Pretendidos

Na fase inicial deste trabalho, pretendeu-se verificar o mês, o dia da semana e o horário de maior concentração de crimes através de histogramas. Ainda nesta etapa do trabalho, pretendeu-se calcular a média de crimes que ocorrem no distrito B2 quando comparado com a amostragem total da cidade de Boston.

Na fase intermediária, pretendeu-se estudar as características da base de dados que se relacionam com o fator de risco dos crimes. Por se tratar de dados binários, utilizou-se o método de regressão logística, ou seja, um modelo preditivo. Com base nestes resultados uma matriz de confusão foi proposta para estimar a ocorrência de crimes por grau de risco (zero ou um).

Na fase final, pretendeu-se calcular o número de *clusters* adequados para otimizar a visualização geográfica de crimes pertencentes ao distrito B2. Através do cálculo da curva de Elbow e do agrupamento por k-médias, foi possível particionar o território da cidade de Boston que este distrito policial deve atender.

III. RESULTADOS OBTIDOS

Todos os resultados apresentados nesta seção foram obtidos através da linguagem de programação Python. Os dados foram tratados e processados nesta mesma linguagem, para que posteriormente fosse realizadas as análises exploratórias. As modelagens de dados, preditiva e descritiva, foram obtidas por bibliotecas e *frameworks* compatíveis com a linguagem de programação Python, a qual foi escrita e comentada em três etapas ao se utilizar o Jupyter Notebook.

Na Figura 1, através de um gráfico do tipo histograma, mostramos o número de crimes por meses do ano, em 2018, que ocorreram na região do distrito policial B2, da cidade de Boston. Neste mesmo gráfico, é indicado por uma linha tracejada a média aritmética de crimes ao longo do ano de 2018, no distrito B2.

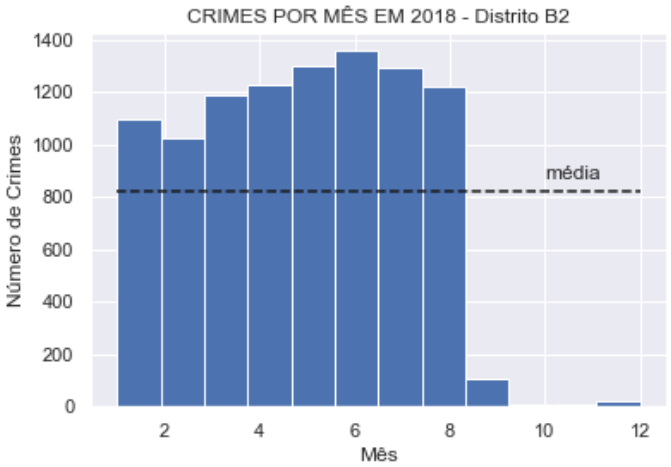

Figura 1: Crimes por mês em 2018 pertencentes ao distrito policial B2 da cidade de Boston.

Na Figura 2, através de um gráfico do tipo histograma, mostramos o número de crimes por meses do ano, em 2018, que ocorreram em toda a cidade de Boston, ou seja, ao considerar todos os distritos. Neste mesmo gráfico, é indicado por uma linha tracejada a média aritmética de crimes ao longo do ano de 2018, na cidade de Boston.

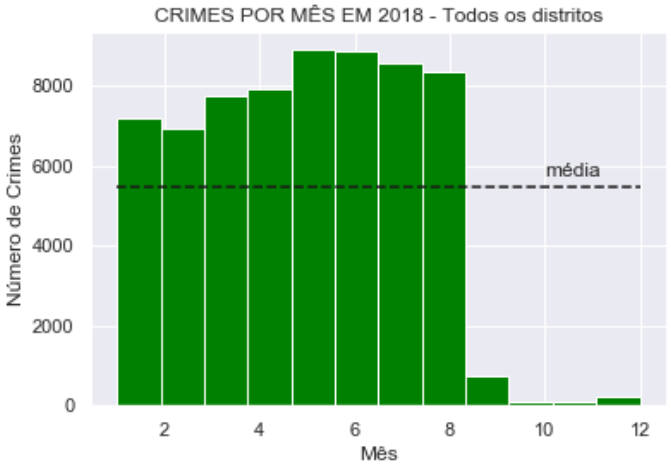

Figura 2: Crimes por mês em 2018 pertencentes à todos os distritos policiais da cidade de Boston.

Na Figura 3, através de um gráfico de linha, mostramos a porcentagem de crimes por mês ao comparar o distrito B2 aos demais distritos da cidade de Boston.

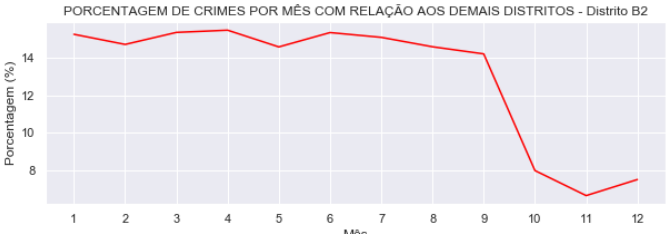

Figura 3: Porcentagem de crimes por mês em 2018 com relação aos demais distritos policiais da cidade de Boston.

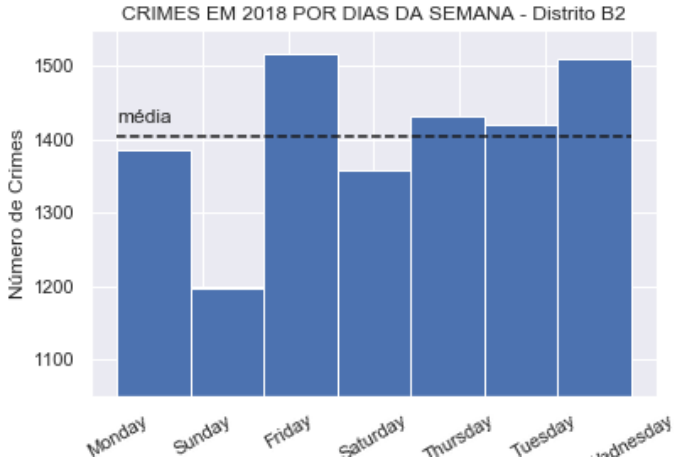

Figura 4: Crimes por dias da semana em 2018 pertencentes ao distrito policial B2 da cidade de Boston.

Na Figura 4, através de um gráfico do tipo histograma, mostramos o número de crimes por dias da semana, de 2018, que ocorreram na região do distrito policial B2, da cidade de Boston. Neste mesmo gráfico, é indicado por uma linha tracejada a média aritmética de crimes ao longo de uma semana, no distrito B2.

Na Figura 5, através de um gráfico do tipo histograma, mostramos o número de crimes por dias da semana, de 2018, que ocorreram na cidade de Boston, ou seja, ao considerar todos os distritos. Neste mesmo gráfico, é indicado por uma linha tracejada a média aritmética de crimes ao longo de uma semana, na cidade de Boston.

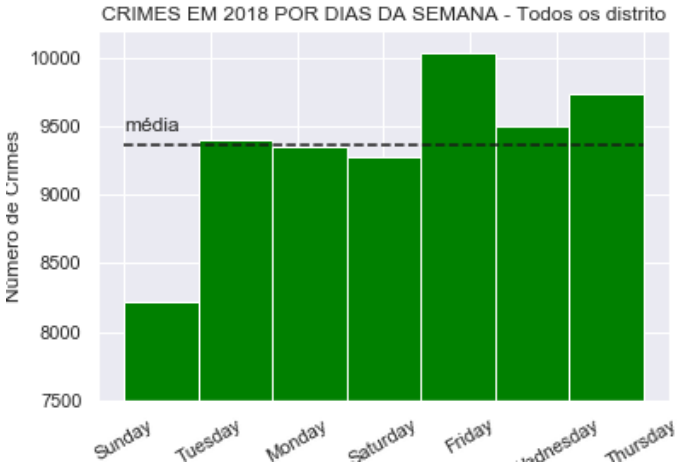

Figura 5: Crimes por dias da semana em 2018 pertencentes à todos os distritos policiais da cidade de Boston.

Na Figura 6, através de um gráfico de setores, mostramos a porcentagem de crimes por dias da semana, de 2018, que ocorreram na cidade de Boston, ou seja, ao considerar todos os distritos.

Na Figura 7, através de um gráfico de linha, mostramos a porcentagem de crimes por mês ao comparar o distrito B2 aos demais distritos da cidade de Boston.

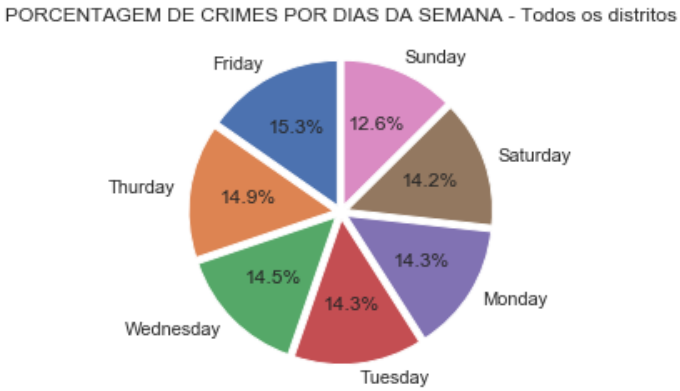

Figura 6: Porcentagem de crimes por dias da semana em 2018 pertencentes à todos os distritos policiais da cidade de Boston.

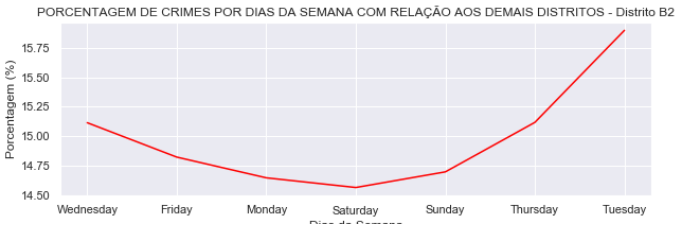

Figura 7: Porcentagem de crimes por dias da semana em 2018 com relação aos demais distritos policiais da cidade de Boston.

Na Figura 8, através de um gráfico do tipo histograma, mostramos o número de crimes por horas do dia, em 2018, que ocorreram na região do distrito policial B2 da cidade de Boston. Neste mesmo gráfico, é indicado por uma linha tracejada a média aritmética de crimes ao longo de um dia, no distrito B2.

Na Figura 9, através de um gráfico do tipo histograma, mostramos o número de crimes por horas do dia que ocorreram na cidade de Boston, ou seja, ao considerar todos os distritos. Neste mesmo gráfico, é indicado por uma linha tracejada a média aritmética de crimes ao longo de um dia, na cidade de Boston.

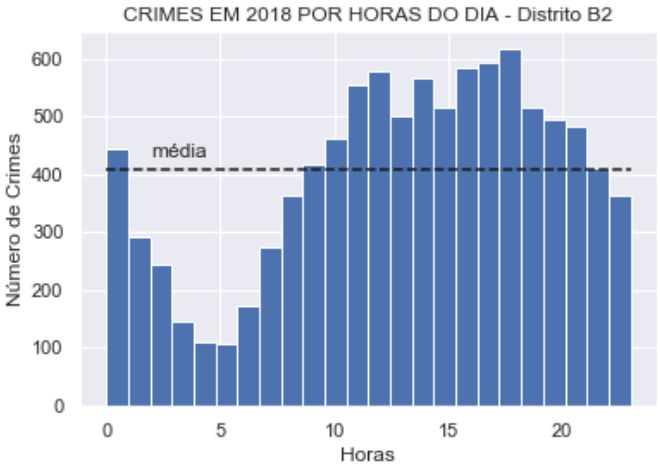

Figura 8: Crimes por horas do dia em 2018 pertencentes ao distrito policial B2 da cidade de Boston.

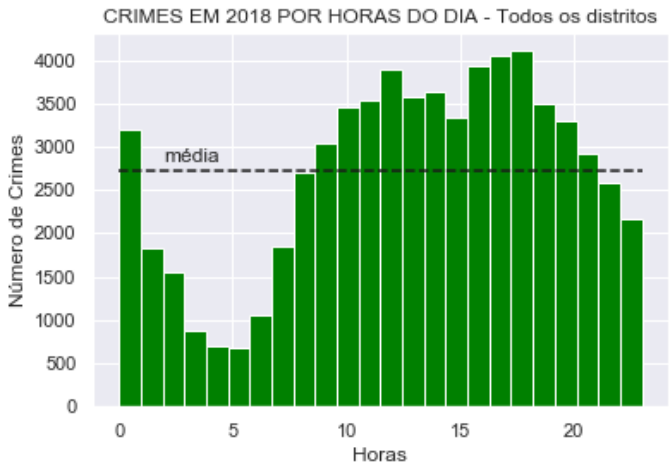

Figura 9: Crimes por horas do dia em 2018 pertencentes à todos os distritos policiais da cidade de Boston.

Na Figura 10, através de um gráfico de linha, mostramos a porcentagem de crimes por horas do dia ao comparar o distrito B2 aos demais distritos da cidade de Boston.

Na Figura 11, através de um gráfico de barras horizontais, mostramos o número de crimes por tipo que ocorreram no distrito policial B2, da cidade de Boston.

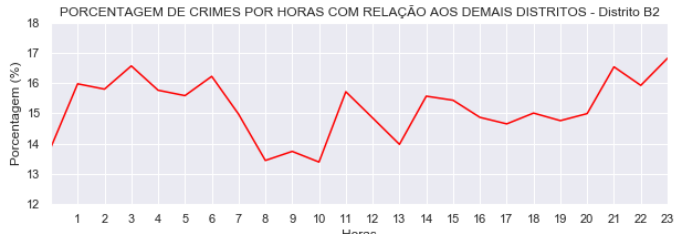

Figura 10: Porcentagem de crimes por horas do dia em 2018 com relação aos demais distritos policiais da cidade de Boston.

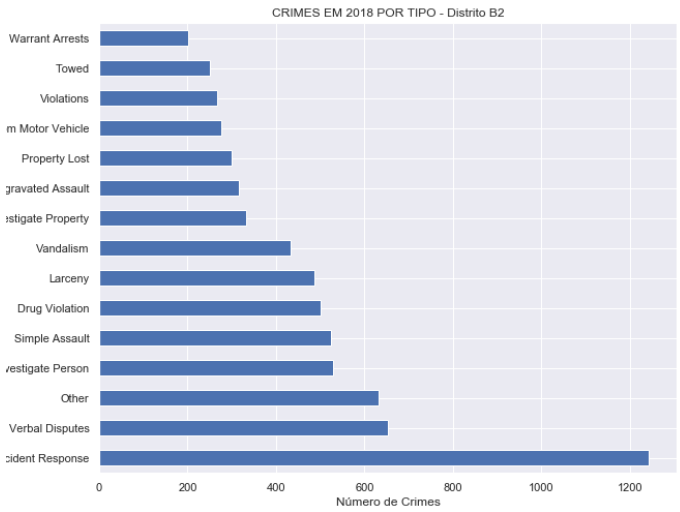

Figura 11: Crimes por tipo em 2018 pertencentes ao distrito policial B2 da cidade de Boston.

Na Figura 12, através de um gráfico de setores, mostramos a porcentagem de crimes por grau de risco que ocorreram na região do distrito policial B2, da cidade de Boston, sendo 0 (zero) associado ao baixo risco e 1 (um) associado ao alto risco.

Figura 12: Porcentagem de crimes por grau de risco em 2018 pertencentes ao distrito policial B2 da cidade de Boston.

Na Figura 13, através de um gráfico de barras, mostramos o número de crimes por grau de risco que ocorreram na região do distrito policial B2, da cidade de Boston, sendo 0 (zero) associado ao baixo risco e 1 (um) associado ao alto risco. Neste mesmo gráfico, mostramos a média aritmética de crimes para este conjunto de dados e a moda (valor mais frequente) do grau de risco.

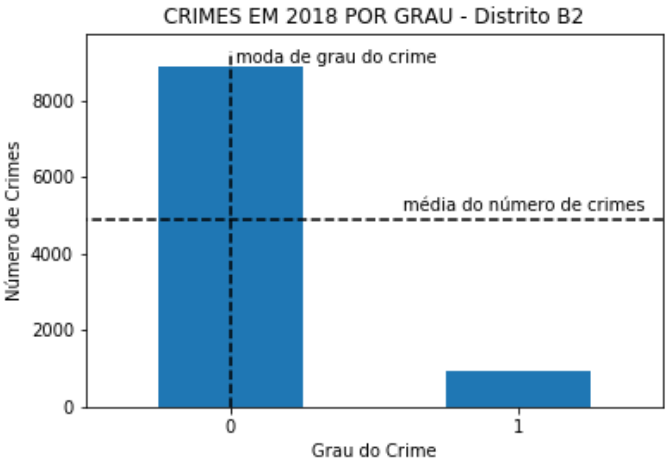

Figura 13: Crimes por grau de risco em 2018 pertencentes ao distrito policial B2 da cidade de Boston.

Na Figura 14, mostramos uma matriz de confusão não-normalizada, sendo 0 (zero) associado ao baixo risco e 1 (um) associado ao alto risco obtido através da técnica de regressão logística, ou seja, pela modelagem preditiva. Neste mesmo gráfico, o eixo das ordenadas indica valores verdadeiros, isto é, confirmados por análise, e o eixo das abscissas indica os valores previstos, isto é, predito pelo teste. A legenda de cores indica o número de crimes para cada classificação. Na figura 15, a mesma matriz é apresentada, porém com todos os valores normalizados.

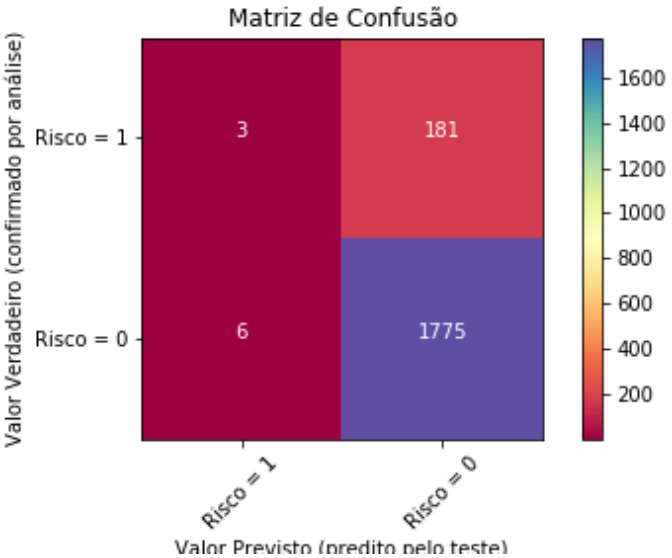

Figura 14: Matriz de confusão não-normalizada para o número de crimes por grau de risco (zero ou um) em 2018 pertencentes ao distrito policial B2 da cidade de Boston.

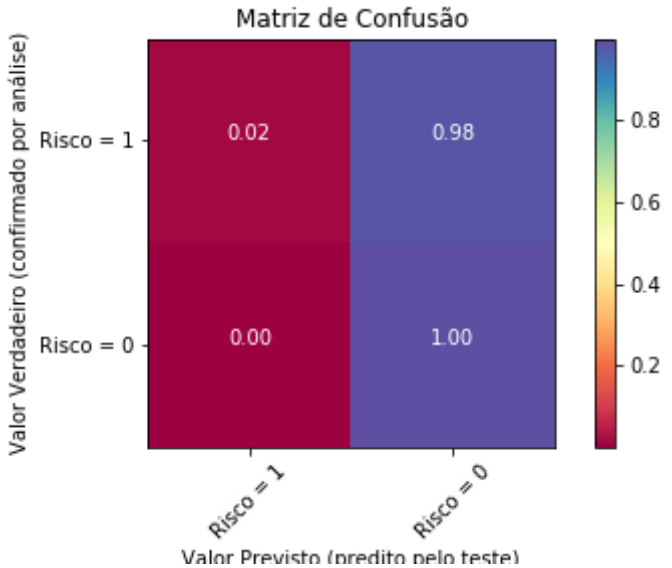

Figura 15: Matriz de confusão normalizada para o número de crimes por grau de risco (zero ou um) em 2018 pertencentes ao distrito policial B2 da cidade de Boston.

Na Figura 16, através de uma distribuição de pontos no espaço geográfico (latitude versus longitude), mostramos as posições geográficas de crimes que ocorreram na região do distrito policial B2 da cidade de Boston. Neste mesmo gráfico, apresentamos por uma escala de cores o grau (zero ou um) de risco que estes crimes pertencem.

Na Figura 17, através de um gráfico de linha, apresentamos a curva de Elbow, a qual nos auxilia na determinação do número de *clusters* a ser utilizado no modelo preditivo. Esta curva foi calculada para se determinar o número adequado de agrupamentos ao se aplicar o conjunto de dados no método de k-médias. Optamos em utilizar cinco *clusters* para facilitar as regiões de crimes e por não haver variações significativas de pontos (score) acima deste valor.

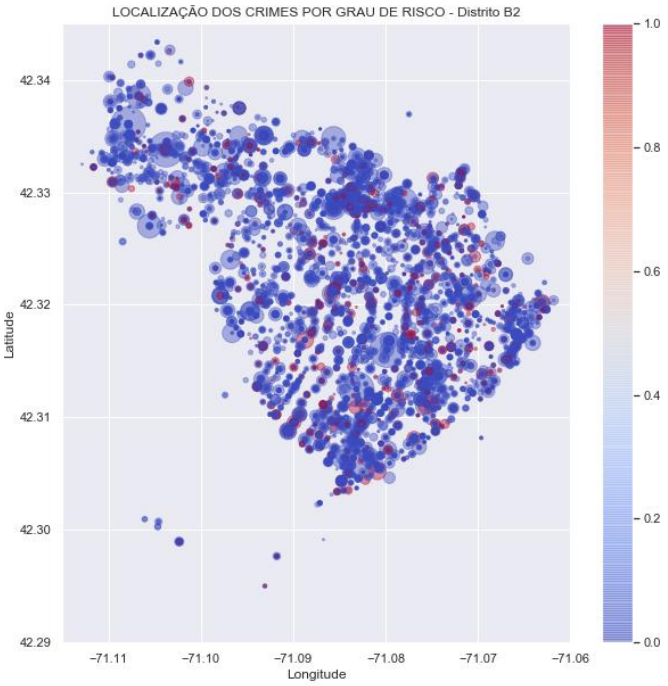

Figura 16: Mapa de coordenadas geográficas com os locais de crimes e grau de risco (escala de cores) em 2018 pertencentes ao distrito policial B2 da cidade de Boston.

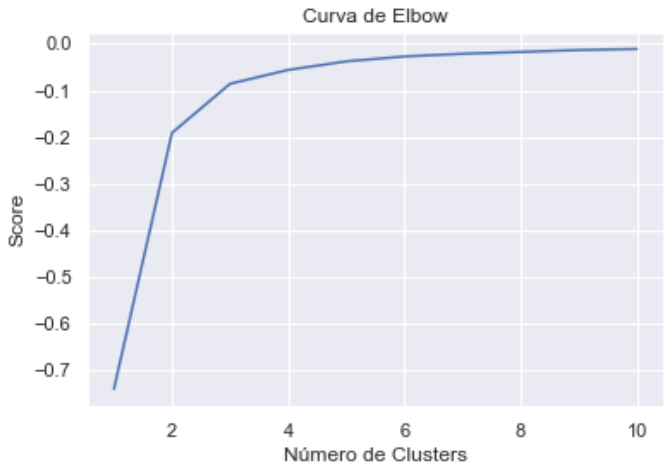

Figura 17: Curva de Elbow para a determinação do número de *clusters* (agrupamentos) que descrevem regiões de crimes no distrito policial B2 da cidade de Boston.

Na Figura 18, através de uma distribuição de pontos no espaço geográfico (latitude versus longitude), mostramos as posições geográficas de crimes por agrupamento do distrito policial B2 da cidade de Boston. Neste mesmo gráfico, apresentamos por uma escala de cores indicando o número do *cluster* ao qual o crime pertence, ou seja, um determinado crime deve pertencer ao *cluster* de número 0, 1, 2, 3 ou 4.

Na Figura 19, através de um mapa, mostramos as localizações dos crimes que ocorreram na região do distrito policial B2, da cidade de Boston. Neste mesmo gráfico, através de uma aproximação (*zoom*) na região de um determinado crime, é possível verificar o número do *cluster* ao qual pertence o crime.

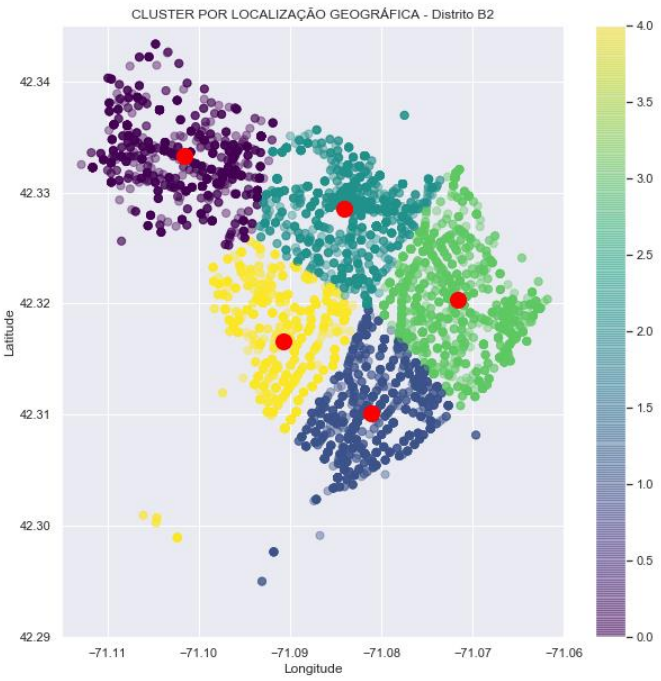

Figura 18: Mapa de coordenadas geográficas de cinco *clusters* (diferenciados pela escala de cores) de crimes pertencentes ao distrito policial B2 da cidade de Boston.

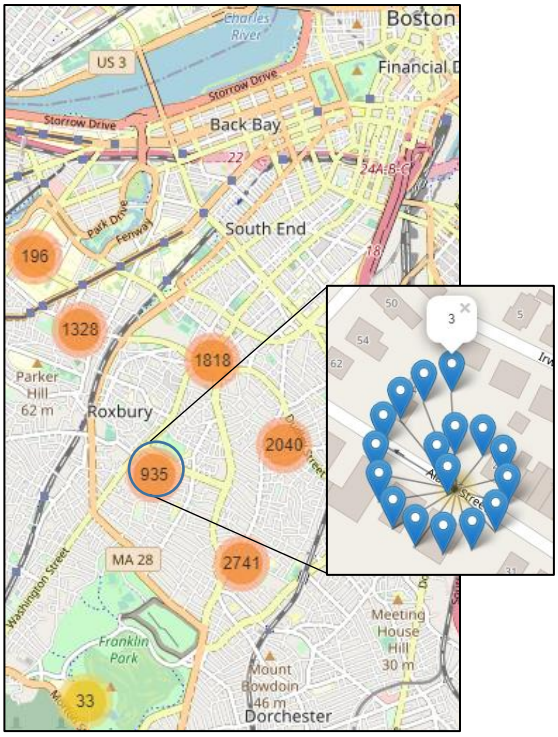

Figura 19: Mapa geográfico os locais de crimes em 2018 pertencentes ao distrito policial B2 da cidade de Boston. O zoom da imagem permite verificar o *cluster* (de zero a cinco) ao qual pertence o crime.

IV. ANÁLISE DE RESULTADOS

Na fase 1 do projeto, verificou-se o número de crimes e suas porcentagens pelos meses do ano de 2018, pelos dias da semana, pelas horas do dia e pelo tipo de crime. Os números de crimes foram apresentados através de gráficos do tipo histograma para os diferentes intervalos de tempo (meses, dias da semana e horas do dia) ao considerar todos os distritos e, especificamente, o distrito B2.

Para uma referência estatística adequada, apresentou-se os valores médios em todos os histogramas. Comparou-se os resultados referentes ao Distrito B2 aos demais distritos policiais através da porcentagem relativa mostradas nos gráficos de linha. Através dos resultados, verificou-se que no ano de 2018 o mês de junho se apresenta com o maior índice de crimes no distrito B2. Ao considerar todas as semanas do ano de 2018, sexta-feira é o dia da semana com o maior índice de crimes no distrito B2. Ao considerar todos os dias do ano de 2018, o maior índice de crimes ocorre às 19 horas no distrito B2.

De forma geral, o Distrito B2 representa, aproximadamente, 15% do total de crimes de nossa base de dados. Por fim, observou-se que os três principais acidentes no Distrito B2 estão classificados da seguinte forma: resposta a acidentes de automóvel; assistência médica; disputas verbais.

Na fase 2 do projeto, verificou-se inicialmente o número de crimes por grau de risco e a porcentagem relativa entre as classificações zero e um. Para uma análise estatística, apresentou-se no gráfico de barras a média de crimes e a moda referente ao grau de risco, a qual é dada pelo grau zero. Um modelo preditivo, dado pela regressão logística, foi aplicado à base de dados do distrito B2, a qual é composta pelos seguintes atributos: código do crime; área reportada; teve tiro; ano; mês; hora; alto risco. Para análise deste modelo, utilizou-se uma matriz de confusão com referência ao grau de risco dos crimes, o qual se apresenta com valores binários (zero e um).

A matriz de confusão foi obtida de forma não-normalizada e normalizada para uma completa análise dos dados. De acordo com a matriz de confusão: o modelo classificou 3 crimes como grau 1 (um) e que realmente eram de grau 1 (um); o modelo classificou 181 crimes como grau 0 (zero) que na verdade eram de grau 1 (um); o modelo classificou 6 crimes como grau 1 (um) que na verdade eram de grau 0 (zero); o modelo classificou 1775 crimes como grau 0 (zero) que realmente eram de grau 0 (zero).

Na fase 3 do projeto, verificou-se inicialmente a localização geográfica dos crimes que ocorrem no distrito B2 com indicação, através de um gráfico de cores, do grau de risco (zero ou um) destes crimes. Posteriormente, calculou-se a curva de Elbow para a determinação do número de agrupamentos adequados ao conjunto de dados. Através deste procedimento, identificou-se a necessidade do uso de 5 agrupamentos (*clusters*).

Os agrupamentos foram obtidos pela técnica de K-médias pelo uso do algoritmo "KMeans" da biblioteca "sklearn.cluster". Através dos conjuntos de dados rotulados pelo número de agrupamento, mostrou-se a distribuição dos crimes através de um gráfico de pontos. Com o objetivo de facilitar a visualização dos crimes e do *cluster* ao qual o crime pertence, geramos um mapa através da biblioteca "folium". Este recurso possibilita a aproximação e afastamento da área geográfica dos locais de crime e a consulta do *cluster* ao qual um determinado crime pertence.

V. CONCLUSÕES

Os objetivos deste trabalho foram alcançados através do pré-processamento de dados, de estudos estatísticos e de aplicações de modelos preditivos e descritivos. O pré-processamento de dados ocorreu através da seleção de atributos, que melhor descrevem as fases deste projeto, e do tratamento de registros nulos que compõem estes atributos. Através das medidas centrais estatísticas foi possível verificar as tendências de criminalidade por horas do dia, dias da semana e meses do ano, de forma a facilitar a compreensão do conjunto de dados. Com o uso da técnica de regressão logística foi possível prever o comportamento dos dados quanto ao grau de risco dos crimes que ocorrem no distrito B2, sendo que a moda de crimes é do grau zero. A curva de Elbow e o agrupamento por k-médias possibilitaram estimar um número de *clusters* e suas localizações geográficas destes *clusters*, dentro dos quais ocorrem os crimes de responsabilidade do distrito B2. A visualização dos crimes através de agrupamentos específico pode facilitar a ação de agentes policiais e de uma possível subdivisão de atendimentos deste distrito policial. Além disto, verificamos a maior ocorrência de crimes no mês de junho, às sextas-feiras e às 19 horas no distrito policial B2. Este distrito apresenta, aproximadamente, 15% do total de crimes da cidade de Boston. Através da matriz de confusão, podemos dizer que o modelo preditivo classificou 3 crimes como grau 1 (um) e 1775 crimes como grau 0 (zero) corretamente, e 181 crimes como grau 0 (zero) e 6 crimes como grau 1 (um) erroneamente. Sugere-se como trabalhos futuros, a compreensão dos tipos de crimes por localização geográfica, inclusive a análise de quais tipos crimes são mais frequentes em cada agrupamento.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] IBM. (Produtor). Introdução à classificação [Arquivo de vídeo], 2018.

[2] IBM. (Produtor). Introdução à regressão logística [Arquivo de vídeo], 2018.

[3] IBM. (Produtor). Regressão logística vs. linear [Arquivo de vídeo], 2018.

[4] IBM. (Produtor). Introdução a agrupamentos [Arquivo de vídeo], 2018.

[5] IBM. (Produtor). Agrupamento por k-médias [Arquivo de vídeo], 2018.

[6] IBM. (Produtor). Mais conteúdo sobre k-médias [Arquivo de vídeo], 2018.

[7] PIGMAN, Michael et al. Analyzing Inventory Data Using K-Means Clustering. In: The International Conference on Data Science (ICDATA), p. 117-122, 2018. Disponivel em: https://csce.ucmss.com/cr/books/2018/LFS/CSREA2018/ICD8072.pdf