Universidade Federal de Viçosa – Campus UFV-Florestal
Ciência da Computação – Introdução à Ciência de Dados
Professor: Fabrício A. Silva
Aluno: Arthur Teodoro Borges - Matrícula: 4672
Aluno: Gabriel Benez Duarte Costa- Matrícula: 4701
Aluno: Matheus Kauan Passos de Souza- Matrícula: 5093
Aluno: Lucas Fonseca Sabino Lara- Matrícula: 5105
Projeto de Desenvolvimento

Relatório Etapa 1

Introdução:

Neste projeto, desenvolvido no âmbito da disciplina CCF425 - Introdução à Ciência dos Dados, buscamos aplicar os conceitos teóricos aprendidos em sala de aula para analisar conjuntos de dados reais e desafiadores. O trabalho está dividido em cinco etapas, cada uma com objetivos específicos, culminando em uma apresentação final que destaca as principais descobertas.

Nesta primeira etapa, nosso foco é o entendimento inicial dos dados e sua preparação. Trabalharemos com um dos conjuntos de dados disponíveis: Dados de criminalidade de SP (SPSafe) ou Dados demográficos dos municípios brasileiros (BrStats), dependendo da turma prática. Nosso grupo ficou responsável pelo conjunto de dados do SPSafe. O objetivo é explorar os atributos presentes, identificar possíveis ruídos, tratar informações ausentes e formular pelo menos 10 perguntas que guiarão nossas análises futuras. Além disso, realizaremos ajustes na formatação dos dados, enriquecimento com atributos externos e outras atividades necessárias para garantir a qualidade da base de dados.

A entrega desta etapa inclui a criação de um projeto no GitHub, onde disponibilizaremos um arquivo README com os integrantes do grupo, as perguntas elaboradas, o código utilizado e um relatório parcial documentando as decisões tomadas e suas justificativas. Cada integrante contribuirá ativamente, e suas ações serão registradas no histórico de commits e na documentação.

Esta fase é fundamental para estabelecer as bases do projeto, garantindo que os dados estejam prontos para as análises exploratórias e técnicas mais avançadas que serão aplicadas nas etapas subsequentes. Com um trabalho bem estruturado desde o início, estaremos preparados para enfrentar os desafios típicos da Ciência de Dados e extrair insights valiosos dos conjuntos de dados selecionados.

Perguntas escolhidas:

Como dito anteriormente, formulamos 10 perguntas que serão nossas guias ao longo do percurso. A seguir estão as perguntas escolhidas, o que elas buscam entender e como estes questionamentos serão úteis nas análises.

1) *Houve picos de criminalidade em eventos específicos?*
   Buscaremos conhecer se a incidência de crimes aumenta em datas comemorativas ou em eventos com grande número de pessoas. Estes dados podem esclarecer sobre a efetividade da segurança pública em dias não rotineiros e/ou festivos.

2) *Quais municípios com menos de 50 mil habitantes possuem média de crimes acima da média estadual?*
   Uma estatística importante para entender a situação de pequenas cidades, como, por exemplo, aquelas situadas no interior.

3) *Quais cidades pequenas apresentam média de crimes violentos acima do padrão estadual?*
   É possível descobrir se cidades de pequeno porte sofrem com o descaso na segurança pública? Investigaremos se municípios com baixa população têm índices de violência elevados, e, se sim, quais são elas?

4) *Quais tipos de ocorrência (acidentes, furtos, etc.) têm as maiores taxas de resolução?*
   Hipótese: Acidentes de trânsito têm maior taxa de solução. Existe uma diferença notável na resolução dos crimes com relação à sua natureza?

5) *Existe associação entre cor da pele da vítima e crimes de violência física/verbal?*
   Estudaremos fatores sociodemográficos nos apegando totalmente a dados estatísticos e evitando conflitos de viés interpretativo.

6) *Quantos registros são iniciados mas não finalizados? Qual o percentual?*
   A quantidade de BO´s não finalizados pode ser considerada alta? Há uma burocratização do processo de registro de queixa ou o sistema é eficaz?

7) *Qual porcentagem de BOs é registrada diretamente pelas vítimas?*
   Questionamento que visa entender o quão acessível é o sistema policial.

8) *Quais horários/dias da semana têm maior incidência de invasões domiciliares?*
   As invasões tendem a ocorrer mais em horários noturnos ou em horários comerciais? E quanto aos dias da semana, é possível traçar um padrão temporal?

9) *A classificação da via (avenida, rua, praça) correlaciona-se com a frequência de crimes?*
   Os crimes possuem incidência maior em regiões centrais ou em zonas periféricas? Buscamos entender se há uma relação notável entre local e crime.

10) *Quais regiões apresentam maior taxa de tentativas de crimes não consumados?*
   Um questionamento essencial para o entendimento da efetividade do sistema público de segurança. De maneira oposta, estes dados também nos ajudam a identificar pontos onde há maior carência de vigilância.

Atributos externos adicionados

Para responder todas estas questões será necessário um pouco mais que os dados disponibilizados no Sp Safe, por este motivo, adicionaremos mais informações que serão essenciais para a análise e entendimento dos dados. As novas adições foram:

1) Dias da semana:

Como buscamos traçar um padrão de comportamento baseado em períodos da semana, foi necessário transcrever as datas dos ocorridos para seus respectivos dias da semana. Para isso utilizamos *"df['DATA_OCORRENCIA'].dt.day_name()"*, que vai extrair da data da ocorrência o seu dia da semana, possibilitando fazer a análise de um possível padrão semanal.

2) Eventos atípicos:

Entender se há uma relação entre crimes ocorridos e dias "não rotineiros" faz parte dos inúmeros objetivos deste projeto, logo, precisamos conhecer as datas de grandes eventos que movimentaram a população no ano de 2022. Para isso adicionamos um conjunto de datas marcantes no ano de 2022, como, por exemplo, feriados relevantes e jogos importantes. Usando estes dados exteriores podemos identificar crimes que ocorreram em datas e eventos movimentados e entender como se correlacionam.

3) Incidentes:

Com a finalidade de agrupar os crimes juntos, podemos observar que no banco de dados pode ser que um crime tenha ocorrido como uma invasão de propriedade e tenha ocorrido um furto. Esses dois crimes são considerados distintos, um sendo de invasão de propriedade e o outro de furto. Então, é interessante criar um novo atributo que identifique incidentes. Se houver informações iguais em outros atributos, significa que provavelmente são do mesmo incidente, ocorreram ao mesmo tempo e com a mesma pessoa.

Com o tratamento de dados atual, que ainda apresenta muitos registros em branco, foi possível associar um número de incidente a 20.556 BOs (Boletins de Ocorrência). No entanto, o dataset completo contém 720.446 registros. Essa diferença evidencia a necessidade de complementar as informações faltantes, pois só assim será possível determinar se um

determinado BO ou crime faz parte de um conjunto de ocorrências relacionadas a um mesmo incidente.

Tratamento de ausência de dados

Para responder as perguntas escolhidas precisaríamos da completude de alguns campos, o que nem sempre ocorre. Veja algumas destas demandas e suas soluções abaixo:

1) Falta de nomes de cidades:

Para resolver este problema, usamos a biblioteca "*geopy"* para recuperar o nome da cidade onde o crime descrito ocorreu usando sua latitude e longitude. Essa biblioteca pega os valores de latitude e longitude, e por meio de uma API, recupera o endereço onde essas coordenadas batem, retornando a informação faltante no campo *"cidade"*.

2) Ausência de períodos:

Alguns registros de BO não possuem seus períodos registrados de acordo seus respectivos horários, apresentando valores nulos (NaN). Com base nas informações do SPSafe, o turnos do periodo foram adicionados a partir desses horários registrados, foi feito, primeiramente, filtrando o dataframe para exibir as linhas de dados em que esse fator ocorre e em seguida foi aplicada a lógica para adição dos turnos. Após o tratamentos desses dados nulos, foram acrescentadas mais de 6 mil novas linhas sem dados nulos na coluna de PERIODO_OCORRENCIA.

Ruídos detectados:

1) Erro na digitação ou formatação dos dados de latitude e longitude.

Foi detectado alguns erros na digitação dos dados do campo "latitude" e "longitude" do dataframe. Um desses erros foi o valor numérico que não corresponde ao padrão usado para identificar as coordenadas, e deslocamento da vírgula.

2) BO's com idade igual a 0:

Foi detectado também alguns BO's em que o autor/vítima possui 0 anos de idade.

Papéis de cada integrante

Todos os integrantes tiveram papel fundamental no decorrer do trabalho até aqui, seja contribuindo com as ideias e perguntas, ou seja contribuindo com desenvolvimento dos códigos. A seguir detalharemos o papel de cada integrante:

- *Arthur Teodoro:* Participou ativamente no desenvolvimento e decisão das das perguntas, na elaboração do relatório e foi responsável pela implementação da adição dos dados externos de “Datas Atípicas”.

- *Gabriel Benez:* Participou na criação e decisão de quais perguntas selecionar, na elaboração do relatório e também foi responsável pela a implementação do código que preenche os dados vazios do campo “Cidade”, e detecção de alguns ruídos.

- *Lucas Fonseca:* Participou ativamente no desenvolvimento e decisão das das perguntas, na elaboração do relatório e foi responsável pela implementação da adição dos dados externos de “Incidentes”.

- *Matheus Kauan: Responsável pela criação do repositório e adição dos colaboradores. Participou ativamente no desenvolvimento e nas escolhas das perguntas e também trabalhou na adição/criação de uma nova coluna no dataframe inserindo os dias da semana correspondente a data.*