Projeto CTTU

Felipe Bormann[1]

Recife, Brazil

Abstract

Este é o relatório final para o projeto de Ciências dos dados na Universidade de Federal de Pernambuco. O tema deste trabalho é a análise e exploração de dados de acidentes, infrações e monitoramento de trânsito na cidade do Recife.

Keywords: `dados`, CTTU, trânsito, data science, geodata

1. Introdução

Dados. Os dados utilizados neste projeto foram todos extraídos do site de dados públicos da Prefeitura do Recife. Busquei utilizar somente os dados da entidade CTTU.

Objetivo. O objetivo deste projeto é explorar as relações e características dos dados da CTTU para extrair conhecimento, principalmente em relação aos acidentes e infrações cometidas por condutores. Os datasets possuem informações das seguintes entidades:

- Acidentes
- Infrações
- Equipamentos de Monitoramento
- Semafóros
- Equipamentos de Vigilância

[☆] Fully documented templates are available in the elsarticle package on CTAN.
[1] Since 1880.

Tecnologia. A tecnologia usada durante o projeto foi a linguagem Python e as seguintes bibliotecas do seu sistema: Pandas, GeoPandas, Matplotlib, Bokeh e Shapely. Além da criação de um notebook jupyter para a execução do projeto.

2. Limpeza dos dados

Acidentes. A limpeza dos dados dos acidentes foi relacionado ao fato de que houve uma transformação na forma como os dados eram armazenados entre o ano de 2014 e 2015, sendo 2016 bem parecido com 2015.

Em relação a datas, enquanto 2014 seguia o formato mês/dia/ano, os anos de 2015e 2016 seguiam o formato dia/mês/ano.

Outra diferença gritante estava no nome de colunas que representavam o mesmo dado, como: "tipo" e "tipo de abertura" ou "quantidade_vitimas" e "quantidade de vitimas".

Nesta variável, "quantidade de vítimas", foi necessário remover dados inválidos como "-", "VT" ou "F", como não havia especialista e a quantidade de dados errados não era tão alta, decidi remover todos estes valores ao invés de atribuir 0 por pura intuição.

Outra ocasião foi a junção de diferentes valores da variável "tipo de ocorrência" para um valor único, como o conjunto "ATROPELAMENTO", "atropelamento" e "Atropelamento", todos se tornando um único valor: "atropelamento", facilitando o resultado.

Infrações. A primeira ação em relação às infrações foi detectar quais colunas eram do tipo datetime e que podiam ser convertidas para este tipo em Pandas. Além da extração da criação da coluna "numerosemaforo" para todos os campos que possuiam o número do semafóro associado em sua descrição.

Para as infrações, que não possuiam um ponto geográfico(latitude, longitude) associado, utilizei da informação de ter um semafóro associado à infração, desta forma reduzi o dataset de 1.6milhões para 890mil, e associar à posição geográfica do semáforo à infração, dessa forma concendendo um espaço para

novas inferências baseadas em granularidades maiores e geográficas como Setor censitário, bairro, etc.

Semáforos. A única coluna que possuia dados inválidos em semafóro foi a do tipo de funcionamento, que possuia erros de digitação mas nenhum valor nulo.

3. Análise Exploratória

Acidentes. Em relação à quantidade de acidentes, consegui algumas evidências de que menos acidentes acontecem ao serem instalados equipamentos de monitoramento. Em números absolutos: somente 52 dos 4344 acidentes registrados foram à 100m de distância de um equipamento de monitoramento.

E a figura 1 mostra que há um aumento drástico de acidentes próximo aos equipamentos de monitoramento durante os finais de semana, que são justamente quando estes são desligados.

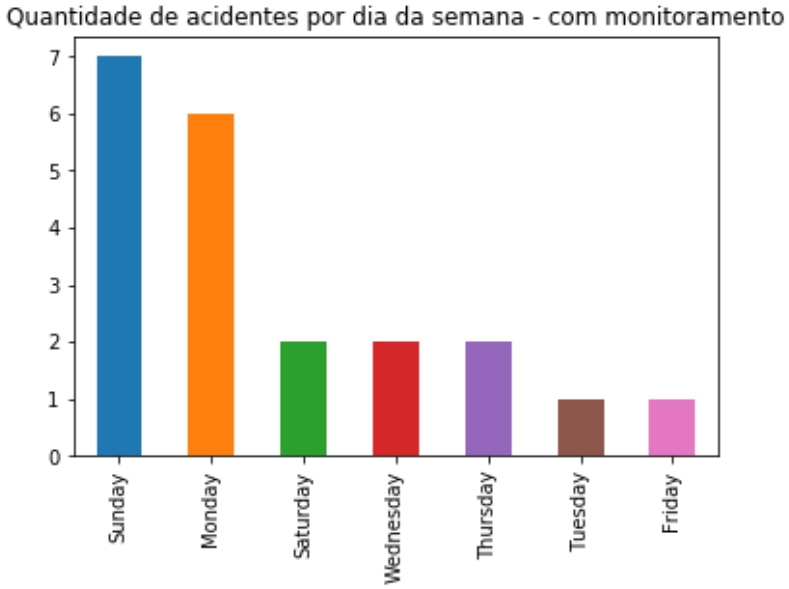

Figure 1: Quantidade de acidentes próximos à equipamentos de monitoramento

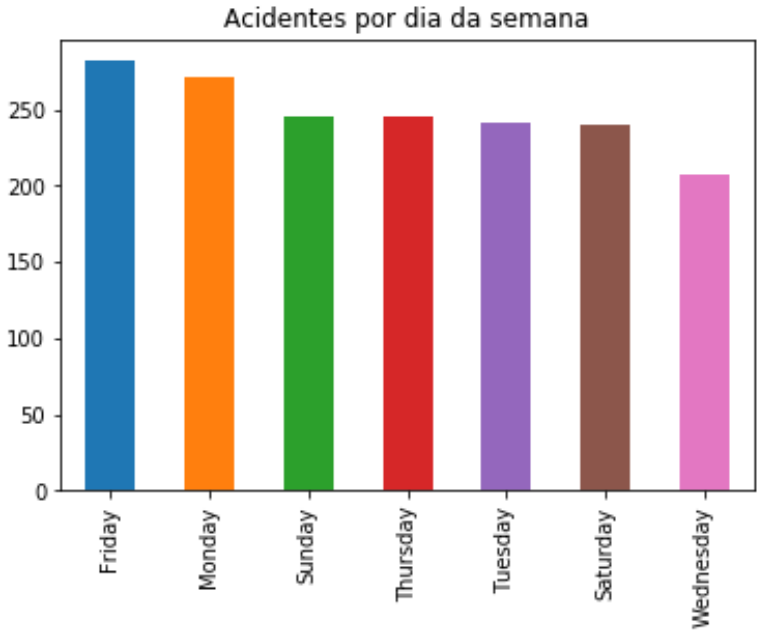

Figure 2: Quantidade de acidentes por dia da semana

Infrações. Procurei por variáveis que fossem interessantes e tentei relacioná-las com as variás dimensões do tempo em alguns casos ou explorar informações e insights simples que o dataframe poderia dar, como na Figura 3, onde fui capaz de tentar começar a investigar, apesar de sem sucesso, se onde haviam mais multas onde a velocidade estava acima entre 20 e 50 porcento acima do permitido causava mais mortes do que "somente" 20 por cento acima.

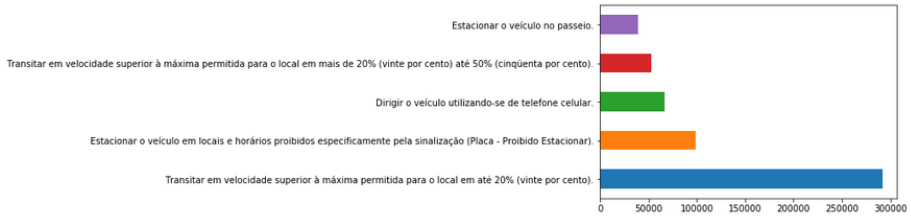

Figure 3: Tipo de Infrações mais frequentes

Infrações por semana. Outra descoberta foi, ao tentar correlacionar a semana do ano entre acidentes e infrações, não cheguei à resposta desejada, dado que não há correlação entre as duas curvas, como mostra, os dois gráficos 4 e 5 mas consegui descobrir que há um aumento crescente até à 43ª semana do ano, em outubro, onde o número de multas vem de uma crescente desde da 39ª semana. Não consegui confirmar a partir de fontes oficiais o motivo de tal incentivo mas é notório o crescimento como mostra a figura 4.

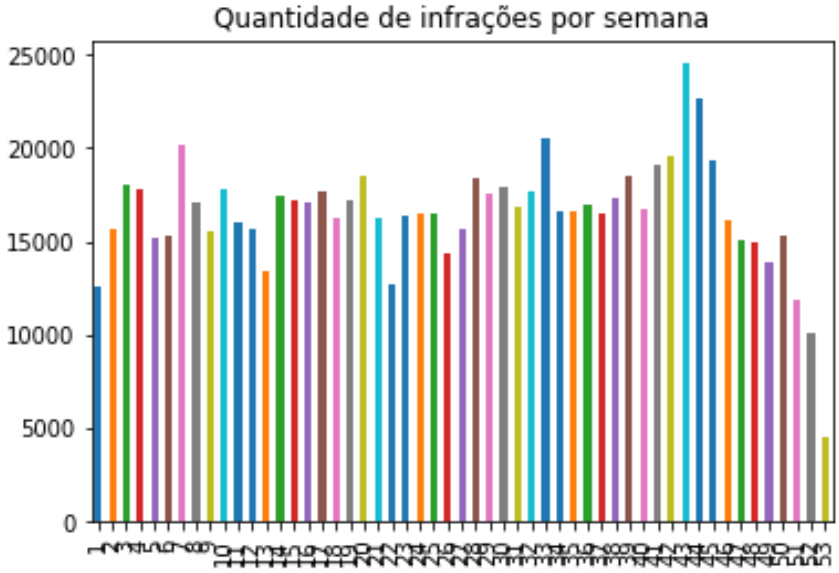

Figure 4: Quantidade de infrações por semana no ano

4. Conclusão

Consegui alcançar o objetivo de gerar alguns insights e visualizações que extraiam alguma informação, contudo, a tarefa se tornou mais complicada do ponto de vista de pré-processamento e análise de dados geográficos, ao ter de aprender sobre visões geográficas, distâncias que fizessem sentido e à busca de informações não tão fáceis do governo, como o sistema de bonificação dos funcionários da CTTU, por exemplo.

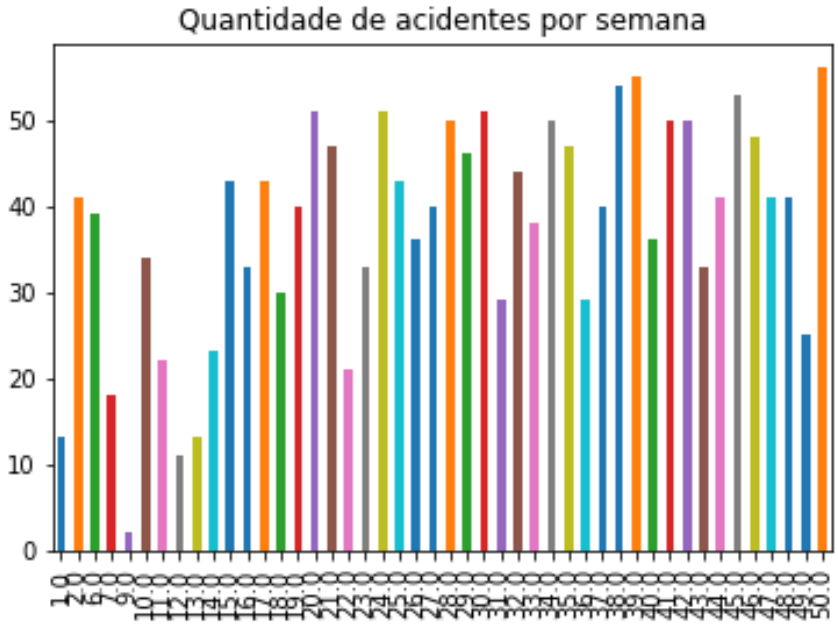

Figure 5: Quantidade de Acidentes por semana do ano

Objetivos Futuros.

- Gerar novas visualizações a partir dos dados de acidentes
- Gerar modelo de predição para a variável "houve um vítima?" em um acidente
- Correlacionar novas informações sobre a influência de infrações na ocorrência de acidentes