Instituto Atlântico

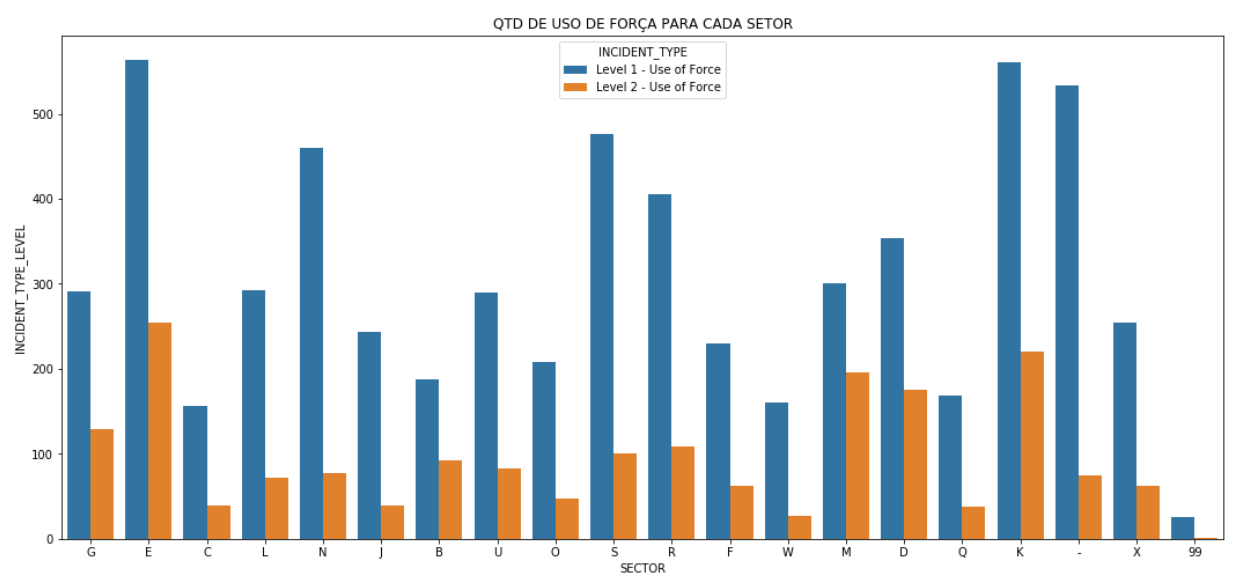
Felipe Rodrigues Dieb

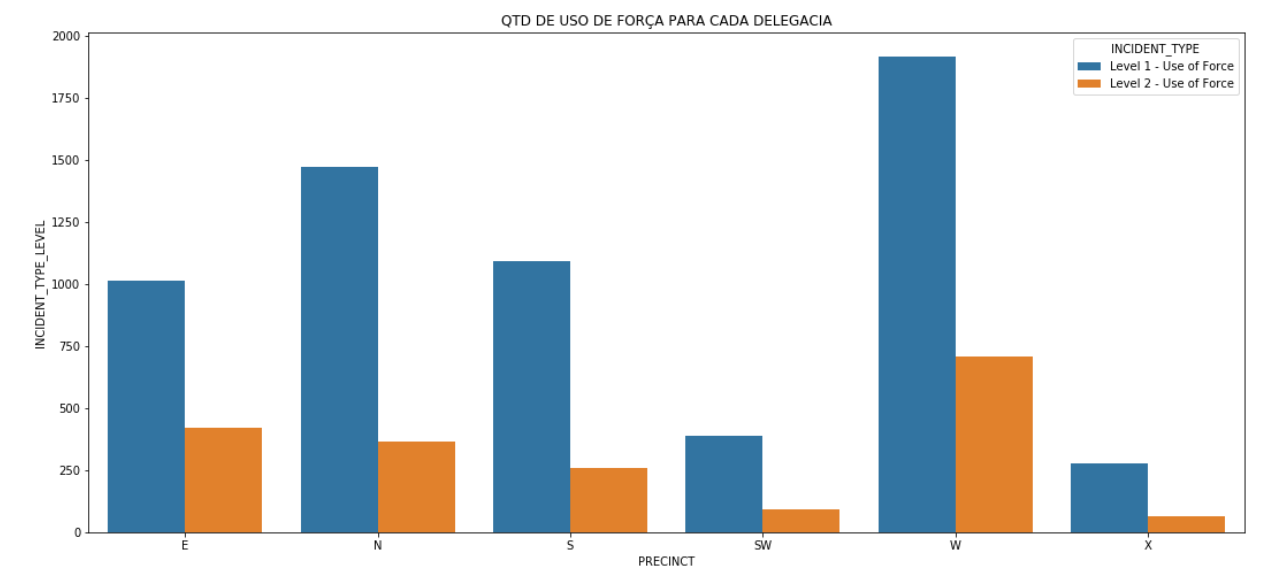
**Questão 01**

Como é a distribuição do uso de força dentre as delegacias e os setores? Em cada setor, qual o BEAT com maior número de incidentes? Apresente também o ranking dos setores segundo o percentual de incidentes “Level 2” em relação ao total de incidentes do respectivo setor.

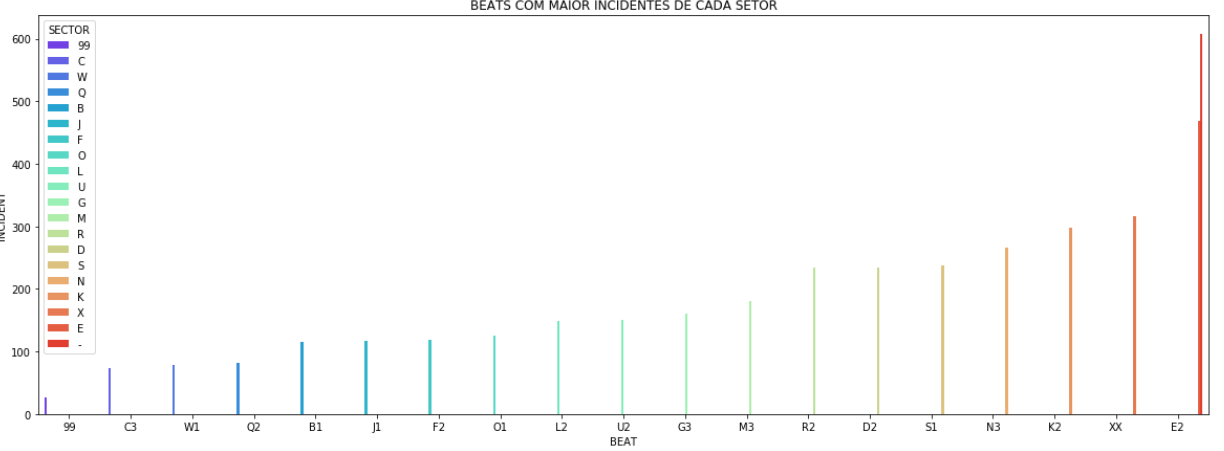
Comentário:

Os gráficos exibidos abaixo mostram a distribuição do uso de força para cada setor e para cada delegacia. Através dessa visualização, nota-se que o uso de força Nível 1 é o mais frequente em todos os cenários.

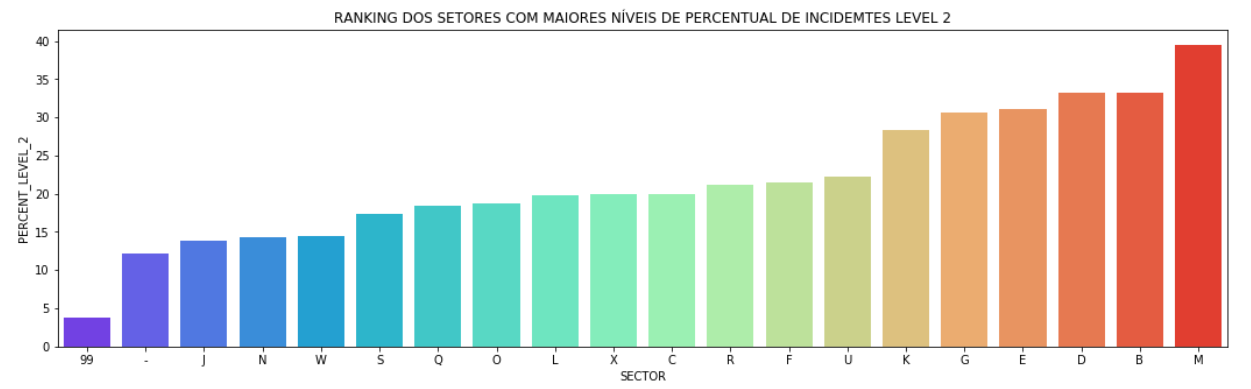




No gráfico abaixo são apresentados os BEATS que possuem maior incidência em cada setor, podendo verificar a quantidade de incidente por cada BEAT selecionado.



O gráfico a seguir descreve o nível de percentual de incidentes ‘Level 2’ de cada setor. A partir desta análise, verifica-se que o setor M possui o maior percentual.



**Questão 02**

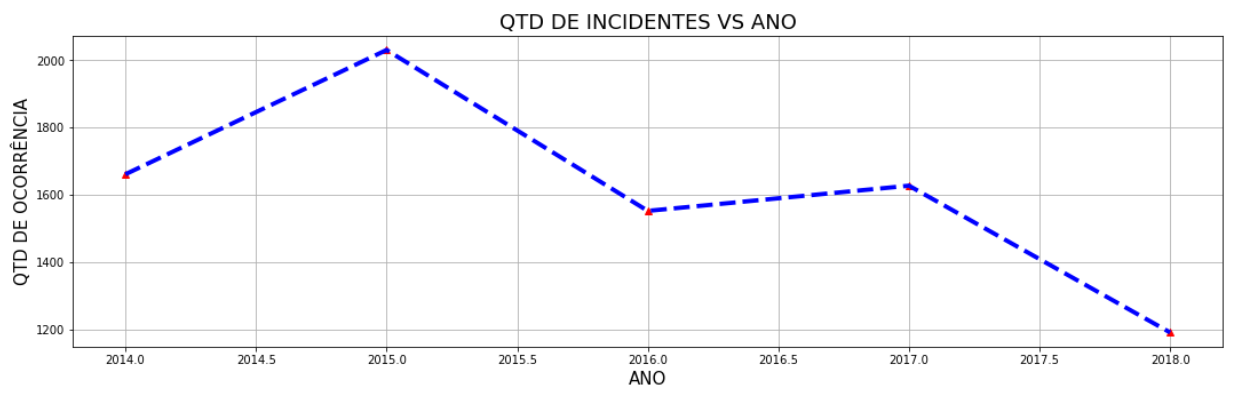
Com relação à distribuição dos incidentes no tempo, é possível encontrar picos ou linhas de tendência dentro dos dias, dos meses, das semanas ou dos anos?

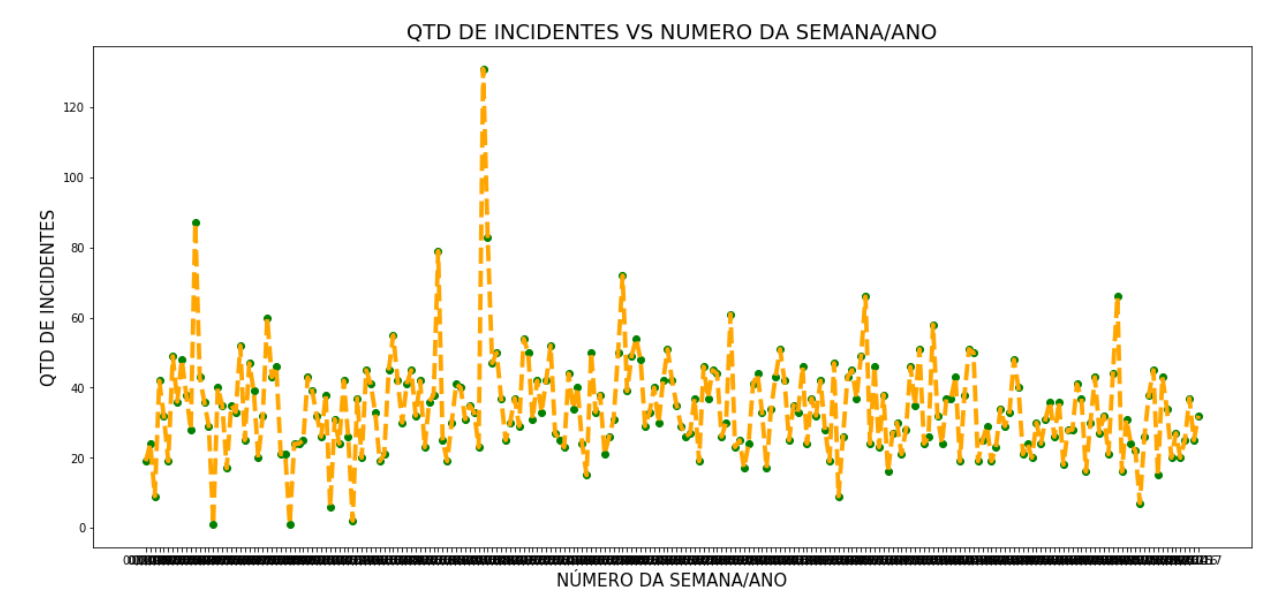
Comentário:

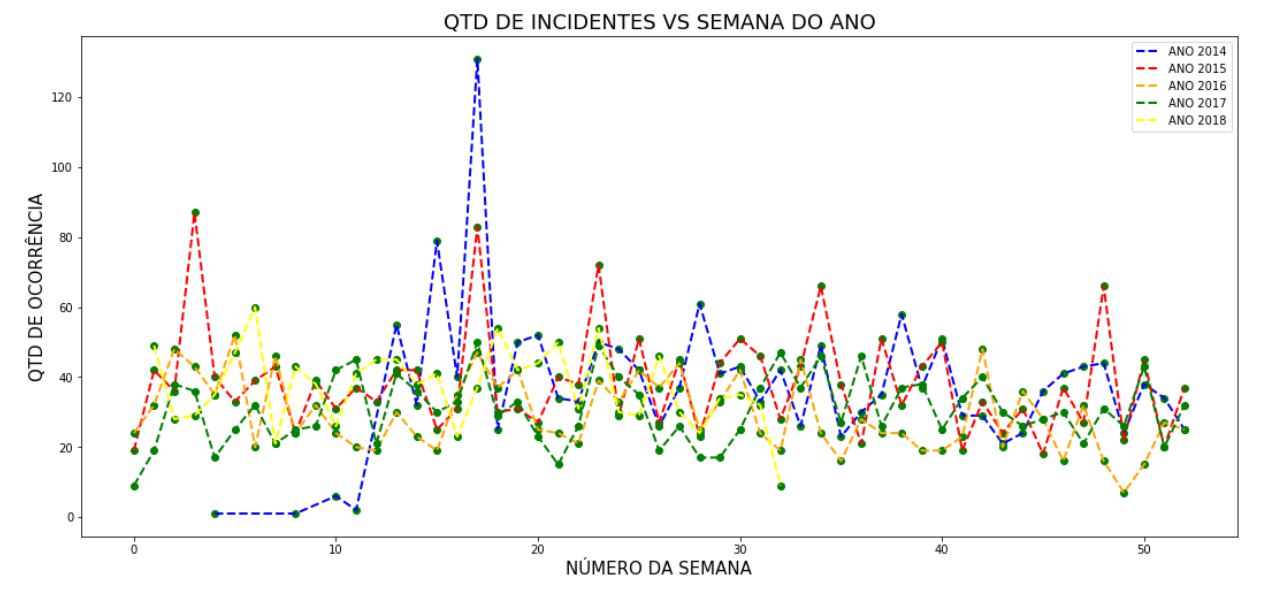
Foi feita uma análise através do gráfico de linha, utilizando a quantidade de incidentes por ano, e observou-se uma tendência decrescente. Isso significa que, a cada ano, ocorre uma redução no número de incidentes. De 2015 para 2018 a redução chega a aproximadamente 41%, passando de 2031 incidentes para 1190.

Ao analisar a semana de cada ano, confirma-se uma pequena tendência de redução no decorrer dos anos, sendo possível observar a existência de um outlier.

Os gráficos abaixo foram desenvolvidos para a análise dessa questão:







**Questão 03**

A polícia deseja dar início a uma investigação interna para verificar se existem policiais excessivamente violentos. No entanto, o prazo para o término desta investigação é bastante limitado. Elabore um script capaz de elencar os policiais em ordem decrescente de chance de violência excessiva com base no número de incidentes dos quais eles participaram.

Comentário:

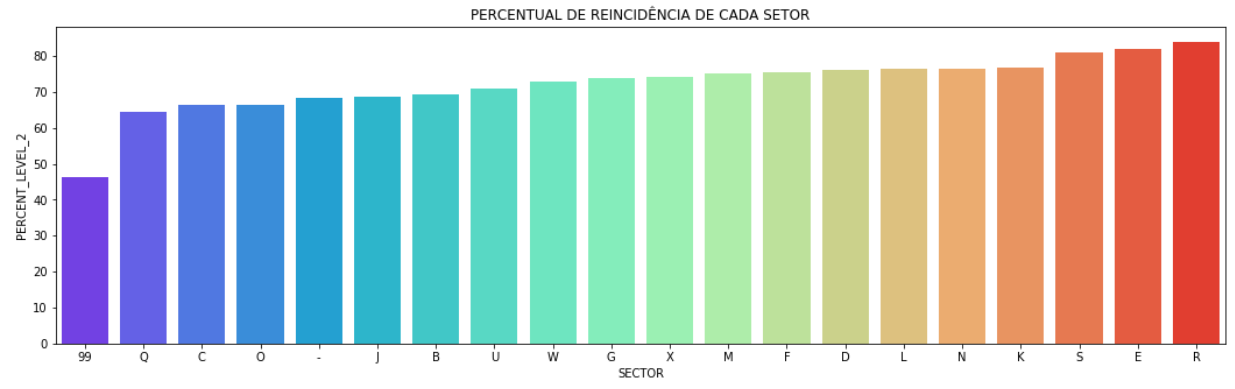
Script está no código fonte.

**Questão 04**

Uma métrica interessante para a polícia é o grau de reincidência por parte dos civis. Apresente o percentual de casos reincidentes em relação ao total de incidentes em cada setor e verifique se há correlação entre esta métrica e o percentual de incidentes "Level 2" calculado na questão 1. Que interpretação pode ser dada a este resultado?

Comentário:

O gráfico desenvolvido auxilia na análise da quantidade de reincidentes em cada setor. Através dele é possível observar que o setor 99, o qual possui menor número de reincidentes, apresenta uma taxa muito alta, de aproximadamente 45%.



**Questão 05**

A liderança do Departamento de Polícia de Seattle manifestou o interesse em uma aplicação que classifica os incidentes em "Level 1" ou "Level 2" com base em outras colunas da tabela e lhe requisitou um parecer sobre esta proposta. Descreva os desafios envolvidos, enumere fatores que fomentem a criação deste classificador e sugira um modelo estatístico para executar esta tarefa, justificando a sua escolha.

Alguns desafios encontrados foi analisar quais colunas fariam parte do treinamento, mas ao ser analisado um baixo ganho de informação nas colunas, tendo o valor mais alto de 30%, foi decidido treinar com todas as informações possíveis para evitar que se perca qualidade no modelo. Após a seleção das colunas foi analisado quais algoritmos seriam escolhidos, por conta da distribuição não linear dos dados, foi decidido utilizar os algoritmos *RandomForest* e *RegressionLogistic*. Sendo que eles se adaptam bem há dados não lineares.

O modelo RandomForest, se baseia no algoritmo de DecisionTree. Ele utiliza um conjunto de árvores de decisões tendo como objetivo reduzir a possibilidade de overfitting. Esse modelo possui duas técnicas importantes, o BootStrap e o Bagging.

O bootstrap é um método de geração de amostras que serão utilizados nas possíveis arvores do modelo, sendo que, cada amostra possui a mesma quantidade de dados. O Bagging, se utiliza das amostras separadas pelo bootstrap para o treinamento das arvores, com isso, seu objetivo é fazer a média dos resultados de cada arvore para calcular a predição final do modelo.

A regressão logística, é um modelo onde seu resultado consiste em probabilidade, ou seja, o valor de saída e entre 0 e 1. Isso ocorre por conta da função que é utilizada, uma função de achatamento, onde o resultado se dar por meio de probabilidade, essa função é chamada de função logística ou sigmoide.

O modelo que mais se adequou ao problema, foi o RandomForest, obtendo 77% de f1-score, e 80% de acurácia, sendo melhor avaliado nas duas métricas em relação ao Regressão Logistica. Além disso, ao analisar a matriz de confusão dos dois modelos, a regrassão logística cometeu bastante erro para a classificação de Level 2, já que a base de dados consiste em sua maior parte nas informações de Level 1 possibilitou uma melhor avaliação para o modelo.

Ao analisar cada modelo comparando as métricas com a base de treinamento e teste, é visto que o RandomForest tende a ter um *overfitting a medida que à profundidade aumenta, sendo que, o modelo tem uma tendência de melhoria na base de treinamento e uma tendência de piora na base de validação*.

