Grupo: Fernanda, Célia e Rodney

Ao receber a tarefa, primeiramente importamos o dataset. Verificamos quantas linhas e colunas possui. Em seguida, estabelecemos objetivos à nossa análise, baseados no escopo principal da instituição que disponibilizou os dados. Dessa forma, buscamos relacionar as condições ambientais e as características das vítimas à ocorrência dos ataques.

Para isso, analisamos a que se refere cada coluna. De início, concentramos-nos apenas nas colunas “Case number”, “Year”, “Type”, “Country”, “Area”, “Location”, “Sex”, “Age”, “Fatal (Y/N), Time, Species, e geramos outra tabela que apresentasse somente esses dados. As demais foram desprezadas pelos seguintes motivos:

- Date: apresenta mesmos dados que o Case Number, porém em outra formatação;
- Name: cada vítima possui um nome diferente, e não é possível extrair nenhum dado dessa informação;
- Injury: cada vítima apresenta um caso diferente e peculiar, por isso optamos em nos concentrar apenas se o acidente foi ou não fatal;
- original order: não apresenta dados que se relacionam com os demais;
- href e href formula: apresentam dados que são diferentes entre si, porém optamos por desprezar, porque não nos interessa as informações que estão no pdf nesse momento;
- pdf: traz o nome do pdf correspondente à cada caso, e não nos interessa porque esse arquivo não está na nossa máquina;
- Unnamed 22 e Unnamed 23: apresentam todas as linhas, praticamente, nulas;
- Case N.1 e Case N.2: representam a coluna Case Number, mas com maior incidência de dados nulos;
- Activity: dados não possuem variância significativa para análise;
- Investigator or source: não é relevante para o escopo da nossa análise quem realizou a coleta.

Seguindo a nossa análise, estabelecemos que “Country” é um parâmetro relevante e que relaciona os demais dados. Então, nivelamos a tabela, retirando todos as linhas que apresentam valores nulos para coluna “Country”.

Assim, fomos em todas as colunas restantes observando quais resultados eram apresentados. Decidimos que todos os resultados NaN e incoerentes (como, por exemplo, “Sex”=”N”) substituiríamos por “Unknown”, com a finalidade de estabelecer um padrão, sem perder informações.

Dessa maneira, seguimos pelas colunas “Country”, “Sex”, “Fatal (Y/N)”, “Type”, “Time” e “Species”, olhando detalhadamente quais eram os dados e como poderíamos categorizá-los.

Em cima dos dados tratados, acreditamos ser possível realizar algumas análises iniciais, tanto qualitativas quanto quantitativas.