Análise de dados: Acidentes PRF 2024

1. Sobre o projeto

Como primeiro projeto do meu portfólio, escolhi fazer uma análise dos registros de acidentes ocorridos em 2024.Para essa análise usei como principais ferramentas o Excel, já que estou buscando me aprofundar no uso da ferramenta, Python, em especial a biblioteca Pandas e o Power BI para visualização.

Objetivo da análise:
Meu objetivo foi trazer visualizações claras das métricas dos acidentes ocorridos, como percentual de acidentes por estados, tipo de veículo e pessoa, perfil dos condutores e causas mais comuns de acidentes, assim como trazer visualizações que talvez não sejam tão claras, como os tipos de acidentes que ocorrem, as condições meteorológicas e quantidade por hora e dia da semana.
Com ele, busco que tenhamos uma visão mais clara dos motivos pelos quais ocorrem os acidentes no país e gerar insights de como contorná-los.

2. Fonte dos Dados

A base foi coletada do site de dados abertos da PRF, os arquivos foram exportados no formato CSV no dia 27/02/2025, na base esses dados são atualizados uma vez por mês, portanto a análise foi feita com os dados disponibilizados até essa data.
Segue o link das bases disponibilizadas:
Dados Abertos da PRF — Polícia Rodoviária Federal

3. Tecnologias Utilizadas

Fiz a importação dos arquivos CSV e a maior parte dos tratamentos no EXCEL, visto que estou buscando maior aprimoramento nesta ferramenta e não tinham tantas inconsistências nos dados que me levasse a tratar com outra ferramenta. Decidi por trabalhar os dados com ID's para ligação das tabelas dimensão com a fato, dessa forma usei o Python e a biblioteca pandas para atribuição dos ID's corretos para cada registro, explico como realizar essa tarefa logo a baixo.
Para uma boa visualização dos dados optei por utilizar o Power BI já que ela é um dos meus objetos de estudo do momento.

4. Metodologia

Foram importadas duas tabelas do portal da PRF, uma contém registros dos acidentes ocorridos, sendo um registro por acidente tendo por volta de 73 mil registros, a outra tabela, agrupa os acidentes por pessoas, sendo que muitas pessoas podem estar presente no mesmo acidente, sendo assim o ID acidente se repete, essa tabela carrega por volta de 196 mil registros e foi usada como a tabela fato.
Comecei o tratamento analisando as colunas e verificando quais seriam pertinentes para a análise e também de que forma elas se relacionam. Identifiquei que em muitas havia o

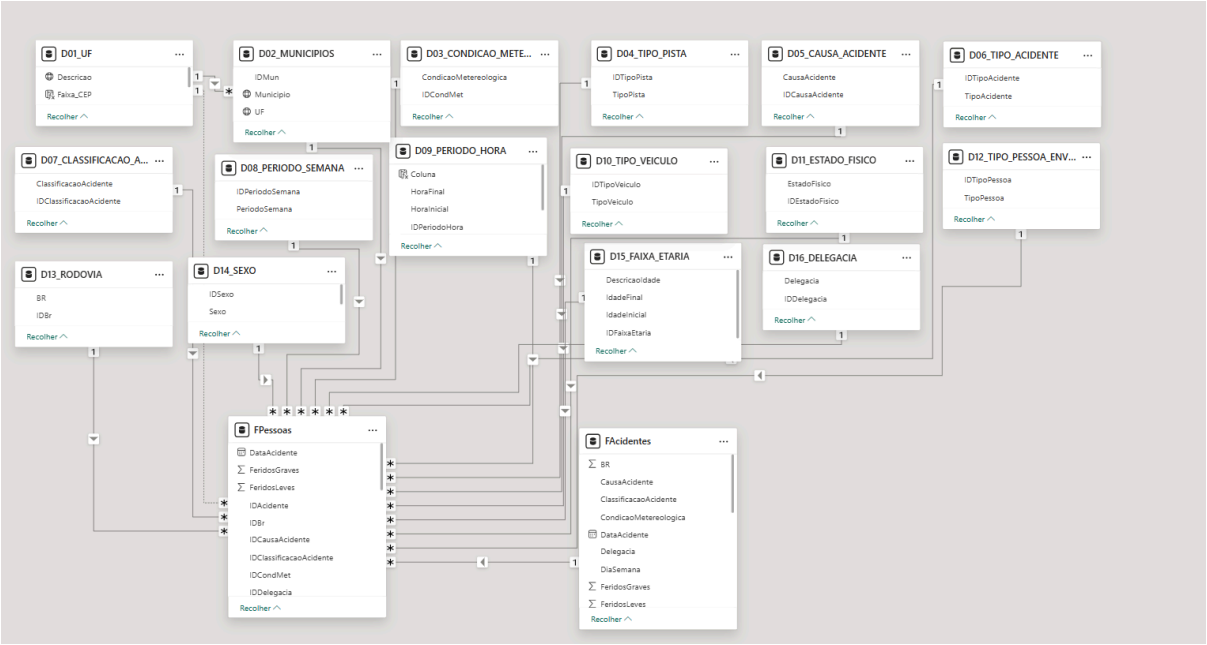
campo de “Não Informado”, dessa forma optei por registrar isso nos campos em branco.Logo depois removi as colunas que não fariam sentido para minha análise como colunas de latitude e longitude por exemplo. Em seguida criei planilhas que se tornariam as tabelas dimensões, para as maiorias das colunas presentes, fiz isso eliminando as duplicatas das colunas individualmente e passando para outra planilha com um campo de ID. Com todas as “tabelas” dimensões prontas, usei a biblioteca Pandas para atribuir os ID’s corretos aos registros na planilha de pessoas, no código eu passei a planilha dimensão e a planilha de pessoas como parâmetro,atribui como chave o campo em comum que tinham e acrescentei o campo de ID. Dessa forma ele localizou quais registros possuíam id 1, 2 e assim por diante.

AtribuiIDS.py > ...

```python
import pandas as pd

# Carregar os arquivos
planilha1 = pd.read_excel(r"C:\Users\Power BI\Desktop\Pessoal\ProjetoPrf\Tabelas dimensão\D08_PERIODO_SEMANA.xlsx")  # Contém ID,
planilha2 = pd.read_excel(r"C:\Users\Power BI\Desktop\Pessoal\ProjetoPrf\IDSAtribuidos\PessoaIdade.xlsx")  # Precisa receber os IDs

#Passar colunas pertinenetes da primeira planilha
planilha1 = planilha1[['IDPeriodoSemana','PeriodoSemana']]

#Faz o merge adicionando a coluna de ID através de uma coluna em comum

#Passar coluna "chave" a qual o codigo vais buscar para atribuir o ID
planilha2 = planilha2.merge(planilha1, on=['PeriodoSemana'], how='left')


# Salvar o resultado
planilha2.to_excel("PessoaPeriodoSemana.xlsx", index=False)


# Salvar no caminho desejado
caminho_saida = r"C:/Users/Power BI/Desktop/Pessoal/ProjetoPrf/IDSAtribuidos/PessoaPeriodoSemana.xlsx"
planilha2.to_excel(caminho_saida, index=False)

print(f"Arquivo salvo em: {caminho_saida}")
print("Processo concluído! A nova planilha foi salva")
```

Esse código foi adaptado para momentos em que os IDs não eram atribuídos a registros únicos mas faixas de valores, como por exemplo período de horas ou faixa etária. Ao final do processo fiquei com 16 tabelas dimensão e uma tabela fato de pessoas.
Gostaria de ressaltar que optei por não utilizar a planilha de acidentes, visto que tudo que ela carregava também estava presente na planilha de pessoas, apesar de ter importado ela para o PBI por precaução e acabei realmente não a utilizando.

Ligação das Tabelas:
Optei por transformar a tabela de pessoas na tabela fato, ela tinha dois campos principais, o de *IDAcidente* e *IDPessoas*, eles garantiram a integridade dos dados e foram essenciais para a análise quando tive que trazer contagem de dados únicos. Transformei a maioria das colunas dessa planilha em tabelas dimensões com registros que continham seus respectivos IDs, depois fiz a atribuição dos IDs na tabela fato para permitir a ligação entre elas feita no Power BI, todas elas tiveram a ligação de 1 para muitos.



5. Visualização e Insights

Para a visualização trabalhei majoritariamente com gráficos, planejei duas abas, a primeira de visualização geral dos acidentes, e a segunda uma visualização relacionada aos dados das pessoas envolvidas nos acidentes. Descreverei meu raciocínio por trás da seleção dos elementos que eu considero principais.
A primeira ideia que tive assim que pensei na estrutura do dashboard foi uma relação da quantidade de acidentes por estado, trouxe essa informação como um gráfico de barras empilhadas pois com ele a distinção da quantidade de acidentes é mais visivel.Logo abaixo coloquei um gráfico de linhas dividido em dia da semana e faixa de horários, para que se pudesse ter noção dos dias e horários onde acontecem maior taxa de acidentes, também coloquei um gráfico de rosca que traz a quantidade de acidentes por classificação, optei por ele pois eram poucas segmentações e traria uma visualização bem clara.Na visão de pessoas achei essencial trazer a comparação de sexo e faixa etária, além do estado fisico das pessoas depois dos acidentes, para ambas visualizações optei por gráficos de barras.Outro parâmetro que gostei bastante de trazer foi a posição que essas pessoas estavam no aicdentes, se eram condutores, passageiros etc.

5.1 Insights

Dessa análise pode se ter alguns insights valiosos, como saber que o estado de Minas Gerais há o maior índice de acidentes do país, com essa informação pode se buscar entender o motivo, talvez rodovias e ruas que há maior dificuldade na direção ou por conta do alto índice de turismo no estado acarretando em maior quantidade de pessoas.
Além disso é legal analisar que referente ao tamanho do país a quantidade de acidentes é relativamente baixa.Outra análise pertinente na página de acidentes é o fato do índice de acidentes com vítimas fatais ser o menor entre todos, podemos buscar entender se o motivo

disso é a busca das pessoas por carros mais seguros talvez, ou o uso de cinto de segurança ter crescido. São situações que podem estar relacionadas.

Já na página de pessoas, logo de cara se percebe que o maior índice de pessoas envolvidas em acidentes são de homens adultos, a maioria deles condutores, o oposto das mulheres,por exemplo, que em sua maioria são passageiras. Esses dados unidos podem explicar o motivo de o valor dos seguros de automóveis ser mais baixo para mulheres, por exemplo.

6. Desafios e Soluções

Logo na importação da planilha percebi que havia uma quantidade de dados em branco em alguma colunas, assim como a descrição de “Não Informado”, optei por deixar todos os dados em brancos com essa padronização visto que era menos de 1 % dos dados e os trouxe inclusive no dashboard.
Na análise de faixa etárias, me deparei com uma situação muito parecida, onde em torno de 1.000 linhas estavam preenchidas com idades discrepantes da realidade como 120 ou 130 anos. A solução adotada foi substituir esses dados para 999 permitindo desconsiderá-los na análise.
Finalizando o tópico, tive algumas dificuldades pontuais no momento de cruzar alguns dados nos gráficos, mas uma rápida pesquisa foi capaz de me auxiliar a resolver.

7. Conclusão e Próximos Passos

Esse projeto apesar de pequeno, me permitiu ter o primeiro contato com uma análise de dados e com a plataforma do Power BI. Com ele aprendi um pouco mais sobre a relação das tabelas fatos e dimensão, como apresentar um dado corretamente e resolver problemas pontuais das análises, como o caso das idades.
Futuramente quando aprimorar essa parte dos meus estudos, pretendo configurá-lo para atualizar mensalmente, além disso penso em recriar a análise trazendo a visualização de outros anos.
Particularmente achei um projeto muito divertido de realizar e estou satisfeita com o meu progresso até aqui,com ele pude aprimorar meus conhecimentos no EXCEL, colocar em prática o que estudei agora em python e entender um pouco mais sobre o Python que é uma ferramenta relativamente nova pra mim.