Relatório do Projeto – Tópicos de Big Data com Python

Aluno:
Italo Rodrigues Bezerra
Analista de Dados / Estudante de Ciências da Computação

Dataset escolhido:
Arquivos (CSV/JSON) – Segurança Operacional – Ocorrências Aeronáuticas

Fonte:
Portal Brasileiro de Dados Abertos – Governo Federal
Link: https://dados.gov.br/dados/conjuntos-dados/ocorrencias-aeronauticas

---

1. Introdução

A análise de dados tem se tornado uma ferramenta essencial para compreender fenômenos complexos e apoiar a tomada de decisões em setores estratégicos. No contexto da aviação civil, o estudo de ocorrências aeronáuticas é fundamental para a segurança operacional, uma vez que incidentes e acidentes impactam diretamente passageiros, tripulações, empresas aéreas e órgãos reguladores.

Este projeto tem como foco a análise de um conjunto de dados públicos sobre ocorrências aeronáuticas no Brasil, utilizando ferramentas e técnicas de Big Data, como Python, Pandas, PySpark e SQL. O objetivo é extrair informações relevantes, identificar padrões temporais e geográficos e desenvolver habilidades práticas em limpeza, transformação, análise e visualização de dados.

---

2. Justificativa da Escolha e Descrição do Dataset

A escolha deste conjunto de dados sobre ocorrências aeronáuticas (acidentes e incidentes aéreos) se deve à relevância do tema para a segurança do transporte aéreo, um setor altamente regulado e essencial para o deslocamento de pessoas e cargas. A análise desses registros permite compreender fatores que influenciam falhas operacionais, contribuindo para melhorias nos processos de fiscalização, prevenção e tomada de decisão por órgãos reguladores e empresas do setor.

O dataset utilizado reúne informações oficiais do Governo Federal, contendo dados sobre data, local, tipo e classificação da ocorrência, fase do voo, características da aeronave e fatores contribuintes. Por se tratar de um banco de dados amplo e multidimensional, ele possibilita investigações sobre padrões regionais, categorias de operação, perfis de risco e a frequência das ocorrências ao longo do tempo.

Além disso, o conjunto de dados apresenta diversos erros, valores ausentes e inconsistências, o que torna o trabalho mais didático ao exigir etapas de limpeza, padronização e tratamento dos dados. Esse aspecto reforça sua utilidade acadêmica, pois simula desafios comuns encontrados em bases reais e permite a aplicação prática de técnicas de preparação e análise exploratória de dados.

---

3. Objetivos do Projeto

Objetivo Geral

Analisar dados de ocorrências aeronáuticas no Brasil utilizando técnicas de Big Data para identificar padrões relevantes relacionados à segurança operacional.

Objetivos Específicos

- Realizar limpeza e tratamento de dados reais e inconsistentes
- Aplicar consultas SQL para filtragem e agregação
- Criar variáveis derivadas utilizando PySpark
- Analisar ocorrências por estado, região e período
- Desenvolver visualizações gráficas para interpretação dos dados
- Consolidar aprendizados práticos em análise de dados

---

4. Metodologia e Atividades Desenvolvidas

O desenvolvimento do projeto seguiu uma abordagem estruturada, sendo conduzido integralmente pelo autor, que atuou como responsável por todas as etapas do processo.

Inicialmente, foi realizada a leitura e compreensão do dataset utilizando Python e Pandas. Em seguida, iniciou-se o processo de limpeza e preparação dos dados, que incluiu:

- remoção de colunas duplicadas e irrelevantes
- tratamento de valores nulos e strings literais “null”
- correção de inconsistências na coluna UF (remoção de valores inválidos como “Substancial” e “Leve”)

Posteriormente, foram realizadas consultas SQL para:

- filtragem de ocorrências por estado (SP)
- agrupamentos por município e classificação
- análise temporal por período específico
- identificação de registros com dados ausentes

Com o uso do PySpark, foi criada a coluna *Região*, classificando cada UF em uma das cinco regiões do Brasil, possibilitando análises geográficas mais amplas. Também foram aplicados filtros específicos, como a identificação de ocorrências relacionadas a colisões com aves.

A etapa de visualização de dados envolveu a criação de gráficos de barras, pizza, linha, histograma e dispersão, permitindo uma análise exploratória clara e objetiva dos padrões identificados.

---

5. Análise dos Resultados

Os resultados mostraram que o estado de São Paulo concentra o maior número de ocorrências, o que pode ser explicado pelo elevado volume de operações aéreas. Em nível regional, o Sudeste se destacou como a região com mais registros.

A análise temporal revelou que 2019 foi o ano com maior número de ocorrências, evidenciado tanto pelo gráfico de linha quanto pelo histograma anual. Esse pico pode indicar aumento no tráfego aéreo, melhoria nos registros ou maior rigor na fiscalização.

No Nordeste, o gráfico de dispersão mostrou que:

- Ceará apresentou 7 ocorrências
- Bahia 6 ocorrências
- Pernambuco 4 ocorrências

Além disso, a análise da classificação das ocorrências indicou que a categoria **“Acidente”** representa cerca de 70% dos registros, evidenciando a gravidade dos eventos analisados.

6. Conclusão

A análise dos dados de ocorrências aeronáuticas permitiu identificar padrões relevantes relacionados à segurança da aviação no Brasil. Observou-se a concentração de registros em estados e regiões com maior atividade aérea, bem como um aumento significativo no número de ocorrências no ano de 2019. A predominância da classificação “Acidente” reforça a importância de estudos contínuos voltados à prevenção e mitigação de riscos operacionais.

De forma geral, o projeto demonstrou como técnicas de limpeza, transformação e visualização de dados são fundamentais para extrair informações confiáveis a partir de bases reais, contribuindo para análises mais precisas e embasadas.

---

7. Aprendizados e Desenvolvimento Pessoal

Este projeto possibilitou o fortalecimento de conhecimentos técnicos em Python, Pandas, SQL, PySpark e visualização de dados. Além disso, proporcionou aprendizado prático sobre a importância da qualidade dos dados, evidenciando que grande parte do esforço analítico está concentrada nas etapas de preparação e limpeza.

Atuando como responsável por todo o projeto, foi possível desenvolver habilidades de organização, planejamento e liderança, assumindo a condução completa das decisões técnicas, definição da metodologia e validação dos resultados. O trabalho também contribuiu para uma compreensão mais profunda do domínio da aviação e da aplicação prática da análise de dados em problemas reais.

---

8. Referências e Bibliotecas Utilizadas

- Pandas
- NumPy
- PySpark
- Matplotlib
- SQL
- Google Colab
- Portal Brasileiro de Dados Abertos – Governo Federal