Este projeto envolve uma análise comparativa entre os dados de voo fornecidos pela Autoridade de Aviação Civil da Somália (CAA) e os dados internos da NavPass para identificar e analisar discrepâncias, visando aprimorar a precisão e abrangência da coleta de dados.

**Principais pontos do projeto:**

**Objetivos**

1. **Identificar dados ausentes:** Determinar quais voos estão no dataset da CAA, mas não no da NavPass.
2. **Descobrir voos adicionais:** Verificar voos registrados pela NavPass, mas ausentes no dataset da CAA.
3. **Analisar discrepâncias:** Identificar padrões ou causas para discrepâncias, como horários, operadores, rotas ou localizações.
4. **Melhorar coleta de dados:** Usar os achados para otimizar os métodos e processos de coleta de dados da NavPass.

**Escopo do trabalho**

1. **Comparação de dados:** Realizar um mapeamento detalhado entre os dois datasets usando identificadores únicos.
2. **Identificação de lacunas:** Listar voos ausentes em cada dataset.
3. **Análise de características:** Avaliar padrões nas discrepâncias, como horários, companhias aéreas, modelos de aeronaves, rotas e localizações.
4. **Uso de dados externos:** Incorporar fontes como FlightAware, RadarBox e pesquisas na web para validar ou complementar a análise.
5. **Solicitações de dados:** Identificar e requisitar dados internos adicionais, caso necessário.

**Entregáveis**

1. **Relatório de Achados Iniciais:** Resumo das primeiras discrepâncias encontradas e insights preliminares.
2. **Relatório de Análise Detalhada:**
   * Análise aprofundada das lacunas.
   * Visualizações de padrões e tendências.
   * Características comuns dos dados discrepantes.
3. **Documento de Recomendações:**
   * Sugestões para resolver as lacunas identificadas.
   * Estratégias para melhorar coleta e validação de dados.
   * Métodos de validação contínua usando fontes externas.
4. **Arquivos de Dados Suplementares:**
   * Listas de voos ausentes em cada dataset.
   * Dados anotados com observações sobre discrepâncias e validações.

**Propósito e importância**

* **Precisão dos dados:** Manter padrões elevados de confiabilidade.
* **Eficiência operacional:** Identificar possíveis lacunas que afetem decisões.
* **Insights estratégicos:** Informar futuras estratégias de aquisição de dados.
* **Confiança:** Demonstrar à CAA que todos os voos estão sendo capturados.

**Expectativas**

* **Colaboração:** Trabalho conjunto com a equipe principal.
* **Comunicação:** Atualizações regulares sobre o progresso.
* **Garantia de qualidade:** Validação rigorosa de achados e documentação completa.
* **Pontualidade:** Entregas iterativas para feedback e ajustes.

Os datasets fornecidos contêm informações detalhadas sobre voos registrados pelo NavPass para os meses de julho, agosto e setembro de 2024. Alguns destaques sobre as colunas e dados disponíveis:

**Estrutura dos Dados**

1. **Identificadores e Datas**
   * Bill: Identificador único para o voo.
   * Fir Start Date: Data e hora de início do voo.
   * Datas associadas a cobrança, contabilidade e vencimento (Bill Date, Accounting Date, Due Date).
2. **Informações Financeiras**
   * Total e Outstanding: Valores financeiros em USD.
   * Moeda utilizada: Total Currency e Outstanding Currency.
3. **Localização e Tempo**
   * Country: Indica que os voos estão relacionados à Somália.
   * Informações de decolagem, aterrissagem, horários programados (Flight Takeoff, Flight Landing, Departure Scheduled Utc, Arrival Scheduled Utc).
4. **Características do Voo**
   * Curacao e distância do voo: Flight Segment Duration Seconds, Flight Duration Seconds, Flight Length Km.
   * Quantidade de pontos observados e pontos registrados na rota: Flight Path Observed Point Count, Flight Path Point Count.
5. **Outras Observações**
   * Indicador de atualização manual: Manually Updated (geralmente "False").

**Meus próximos Passos**

**Preparação dos Dados**

* + Consolidar os três datasets em um único dataframe para facilitar a análise.
  + Verificar dados duplicados ou ausentes.

**Definir Métricas para Comparação**

* + Utilizar colunas como Bill e Fir Start Date como chaves principais para comparar com os dados da CAA.

**Análise Comparativa**

* + Identificar voos ausentes ou adicionais em relação aos dados da CAA.
  + Explorar padrões nas discrepâncias, considerando tempos, rotas, operadores, etc.

**Estrutura dos Dados CAA**

1. **Identificação de Voos**
   * **Operador**: Nome e código legado do operador (Operator Name, Operator EF Legacy Code).
   * **Registro da Aeronave**: Número de registro da aeronave (Aircraft Registration No).
   * **CallSign** e **Flight No**: Identificadores relacionados ao voo.
2. **Informações da Aeronave**
   * **Tipo de aeronave**: Código ICAO (Aircraft Type Code ICAO).
   * **Peso máximo de decolagem**: Incluindo unidade de medida (Max Take Off Weight, Max Take Off Weight UNIT).
3. **Localizações**
   * Aeroportos de origem e destino (códigos ICAO) (From Airport Code ICAO, To Airport Code ICAO).
   * Pontos de entrada e saída na FIR (Flight Information Region) (Entry Point, Exit Point).
4. **Horários**
   * Datas e horários do voo (Flight Date Time).
   * Horários de entrada e saída na FIR (Entry Time, Exit Time).
5. **Rota e Distância**
   * Informações sobre a rota do voo (Flight Routing).
   * Distância percorrida e unidade de medida (Distance, Distance UNIT).
6. **Informações Financeiras**
   * Taxas cobradas por voo (Charge Unit Rate, Over-flight Total Amount).
   * Moeda utilizada (Currency Code).

**Preparação para Comparação**

* + Padronizar colunas e formatos entre os dados da CAA e da NavPass, como:
    - Datas e horários.
    - Identificadores de voos (CallSign, Flight No, Aircraft Registration No).
  + Consolidar datasets da CAA em um único dataframe para análise unificada.

**Definir Métricas de Comparação**

* + Usar Aircraft Registration No, CallSign, e horários como chaves primárias para comparação com os dados da NavPass.

**Realizar Análise Comparativa**

* + Identificar voos que aparecem apenas nos dados da CAA ou da NavPass.
  + Analisar padrões de discrepância, como:
    - Diferenças em aeroportos, horários, ou rotas.
    - Presença de operadores ou aeronaves específicas.

Algumas correlações adicionais que identifiquei manualmente entre os títulos de NAVPASS e CAA:

1. Aircraft Type Code ICAO (CAA) ↔ Aircraft Model ICAO Code (NAVPASS)  
   Ambos referem-se ao código ICAO do tipo ou modelo da aeronave.
2. From Airport Code ICAO (CAA) ↔ Origin ICAO Code (NAVPASS)  
   Referem-se ao aeroporto de origem pelo código ICAO.
3. To Airport Code ICAO (CAA) ↔ Destination ICAO Code (NAVPASS)  
   Referem-se ao aeroporto de destino pelo código ICAO.
4. Flight Date Time (CAA) ↔ Fir Start Date (NAVPASS)  
   Ambos indicam a data e hora de início do voo ou do segmento FIR.
5. Max Take Off Weight (CAA) ↔ Aircraft Model MTOW (kg) (NAVPASS)  
   Ambos referem-se ao peso máximo de decolagem da aeronave.
6. Operator EF Legacy Code (CAA) ↔ Operator Ref ID (NAVPASS)  
   Possível referência cruzada de identificadores de operadores.
7. Entry Point (CAA) ↔ Fir Started (NAVPASS)  
   Indicam o ponto de entrada ou início do FIR.
8. Exit Point (CAA) ↔ Fir Ended (NAVPASS)  
   Indicam o ponto de saída ou término do FIR.

Discrepancies Between CAA and NavPass Data for July

|  | **CallSign** | **Aircraft Registration** | **From Airport Code ICAO** |
| --- | --- | --- | --- |
| 0 | ETH690 | ETBAI | HAAB |
| 1 | ETH690 | ETBAI | HAAB |
| 2 | UAE725 | A6ECU | OMDB |
| 3 | IGO1854 | VTIIZ | HKJK |

discrepâncias identificadas, incluindo diferenças em horários de voo, distâncias e outras métricas, está disponível para análise.

**Passos realizados na comparação**

1. **Seleção de colunas relevantes:**
   * Do lado **NavPass**, selecionei colunas como:
     + Call Sign (indicativo de voo),
     + Aircraft Registration (registro da aeronave),
     + Origin ICAO Code e Destination ICAO Code (aeroportos de origem e destino),
     + Horários de decolagem e pouso (Flight Takeoff e Flight Landing),
     + Distância e duração do voo (Flight Length Km e Flight Duration Seconds).
   * Do lado **CAA**, usei colunas equivalentes, como:
     + CallSign Flight No,
     + Aircraft Registration No,
     + From Airport Code ICAO,
     + To Airport Code ICAO,
     + Horários do voo (Flight Date Time, Arrival Time),
     + Distância (Distance) e peso máximo de decolagem (Max Take Off Weight).
2. **Renomeação para alinhamento:**
   * As colunas foram renomeadas para criar correspondência direta entre os dois datasets, facilitando a fusão e a comparação.
3. **Fusão dos datasets:**
   * Ambos os conjuntos de dados foram combinados (merge) utilizando as chaves comuns, como Call Sign, Aircraft Registration, e os códigos ICAO de origem e destino.
4. **Identificação de discrepâncias:**
   * Foram identificadas diferenças em campos como horários de voo, duração, distância, etc., comparando valores correspondentes entre os dois datasets.

**Exemplo 1: Voo ETH690**

* **NavPass:**
  + Decolagem: 2024-07-11 20:28:00 UTC
  + Pouso: 2024-07-12 02:40:00 UTC
  + Distância: 458,967 km
* **CAA:**
  + Horários de voo: **não encontrados**
  + Distância: **não encontrada**

**Discrepância:** O voo aparece no NavPass, mas não foi registrado nos dados da CAA.

**Exemplo 2: Voo UAE725**

* **NavPass:**
  + Decolagem: 2024-07-30 05:59:00 UTC
  + Distância: 418,785 km
* **CAA:**
  + Decolagem: 30/07/2024 (formato diferente)
  + Distância: 1839,18 km
* **Discrepância:** A distância registrada é significativamente menor no CAA. Isso pode indicar um erro nos dados de uma das fontes ou diferentes interpretações do segmento do voo.

**Exemplo 3: Voo IGO1854**

* **NavPass:**
  + Decolagem: 2024-07-30 03:57:00 UTC
  + Distância: 464,536 km
* **CAA:**
  + Decolagens registradas em 01/07/2024 e 08/07/2024 (datas conflitantes).
  + Distância: 2438,69 km

**Discrepância:** Diferença de datas e distâncias. O voo tem várias ocorrências no CAA que não batem com os registros da NavPass.

Analise Julho

* **5.184 voos** estão presentes tanto nos dados da CAA quanto nos da NavPass.
* **2.485 voos** estão presentes nos dados da CAA, mas ausentes nos da NavPass.
* **1.520 voos** estão presentes nos dados da NavPass, mas ausentes nos da CAA.