**Resumo Projeto Diego – Linhas Aéreas**

Correlação de Colunas importantes:

Para criar a coluna Chave, que é que faz correlação entre as viagens, segue abaixo a relação:

(Sempre no padrão 🡪 Caa – NavPass) 🡪 Essa será nossa chave principal.

**CallSign Flight No – Call Sign**

Representa o indicativo de chamada do voo, uma identificação única para o voo em questão. É usado em comunicações entre a aeronave e os controladores de tráfego aéreo, frequentemente composto pelo número do voo e um identificador da companhia aérea.

**Aircraft Registration No – Aircraft Registration**

Refere-se ao número de registro da aeronave, que é uma identificação única atribuída a cada aeronave, similar a uma placa de veículo. Este número é utilizado para rastrear informações específicas sobre a aeronave, como propriedade e especificações técnicas.

**From Airport Code ICAO – Origin ICAO Code**

Este código de quatro letras representa o aeroporto de origem do voo de acordo com o padrão ICAO (International Civil Aviation Organization). Por exemplo, "KSFO" seria o código ICAO do Aeroporto Internacional de San Francisco.

**To Airport Code ICAO – Destination ICAO Code**

Este código de quatro letras identifica o aeroporto de destino do voo, também seguindo o padrão ICAO.

**Aircraft Type Code ICAO – Aircraft Model ICAO Code**

Representa o código do modelo da aeronave no formato ICAO, que identifica o tipo e a família da aeronave. Por exemplo, "B738" pode ser usado para o Boeing 737-800.

Com elas todas juntas, conseguimos definir basicamente um ID de correlação único de cada viagem. Mas quem dera fosse tão perfeito assim, esses IDs na verdade não são únicos, eles podem sim ter diversas correlações. E é aí que entram as colunas de Data/Hora, que são:

**Entry Time – Fir Started**

**Exit Time – Fir Ended**

(horários de entrada/saída de uma determinada zona aérea)

Diferentemente do que estava no outro documento (o enviado no Workana com as explicações), eu retirei com minhas análises que, o dataset da NavPass ele é um pouco mais completo do que a do Caa, digo isso até mesmo em termos de início/fim de viagem.

No dataset NavPass temos as informações de Flight Takeoff e Flight Landing (que são os horários de decolagem e pouso do avião, diferentemente do dataset Caa, que apenas possui o Entry Time, Exit Time e Departure Time que são, respectivamente:

- Momento/localização em que o avião entra em uma região de controle aéreo (FIR - Flight Information Region).

- Momento/localização em que o avião sai dessa FIR.

- **Horário programado ou real de partida do voo.**

E para conseguirmos interligar ambos datasets, não podemos simplesmente pegar os horários de decolagem e pouso do avião, mas sim da entrada da zona aérea (FIR), e é aí que surge a correlação citada acima.

Agora com esses dados em mãos é só juntar, comparar e teremos os resultados, correto? Errado!

Como são empresas diferentes, imagino que elas realmente não tem obrigação de colocarem o mesmo horário nas planilhas, por isso podem errar por segundos ou minutos de diferença, então nosso papel é encontrar as menores diferenças de tempo possíveis que fazem sentido serem parte da mesma análise (mesmo vôo).

Desenvolvi uma análise por “prioridade”, vamos se dizer. Criei uma coluna a parte no dataset chamada “priority”, onde segue as regras abaixo:

**- Nível 0:** Quando ambas datas/horas Start Date, Fir Started e Fir Ended batem em ambos datasets;

**- Nível 1:** Quando ambas datas/horas Start Date e Fir Started batem em ambos datasets;

**- Nível 2:** Quando ambas datas/horas Start Date, e Fir Ended batem em ambos datasets;

**- Nível 3:** Quando apenas a data/hora de Start Date bate em ambos datasets;

**- Nível 4:** Quando ambas datas/horas Fir Started e Fir Ended batem em ambos datasets;

**- Nível 5:** Quando apenas a data/hora de Fir Started bate em ambos datasets;

**- Nível 6:** Quando apenas a data/hora de Fir Ended bate em ambos datasets;

**- Nível 7:** Quando nenhuma data/hora bate em ambos datasets.

E por que eu desenvolvi isso? Simples, quantos mais dados de horas batem nos datasets, melhor, e se não se encaixar nos níveis mais baixos, ele vai aumentando até se encaixar em algum.

O nível 7 é quando nenhum dado de hora bate no dataset (é mais normal do que parece, porque pra bater precisa ter 100% de acurácia até nos segundos).

Então, agora, com os resultados que tiverem obtido nível 7 na análise de prioridade, faremos a seguinte matemática:

Encontraremos a menor diferença de data/hora possível que se encaixa na chave criada tanto do CAA quanto do NavPass, então por exemplo:

**Chave aleatória:** GTR874TRDF5f6589G 🡪 Tem em ambos datasets, porém em NavPass é repetido (supondo)

**CAA:**

16/12/2024 01:00:00

**NavPass:**

16/12/2024 01:12:00

16/12/2024 01:30:00

Qual a data que tem menor diferença para bater com o CAA? Obviamente, a primeira data de NavPass, 16/12/2024 01:12:00. E será essa análise que faremos para tentar encontrar os correspondentes em ambos datasets (o batimento é vice-versa).

Após fazermos todas essas análises acima, chegamos no seguinte resultado:

**Quantidade de registros (NÃO DUPLICADOS) tabela CAA:** 13529

**Quantidade de registros (DUPLICADOS) tabela CAA:** 5965

**Quantidade de registros encontrados do CAA no NavPass:** 7807

**Quantidade de registros NÃO encontrados do CAA no NavPass:** 5722

**Quantidade de registros (NÃO DUPLICADOS) tabela NavPass:** 19531

**Quantidade de registros (DUPLICADOS) tabela NavPass:** 3

**Quantidade de registros encontrados do NavPass no CAA:** 10143

**Quantidade de registros NÃO encontrados do NavPass no CAA:** 9388

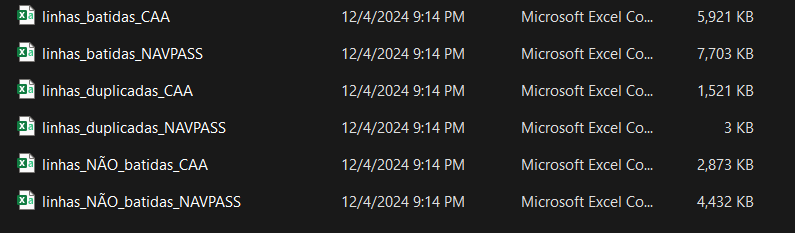
**Quantidade de registros NÃO duplicados:** Total de registros na tabela que não tem dados duplicados (no caso estamos considerando duplicados pela chave criada acima (dados dos aviões/aeroportos + Fir Started).

**Quantidade de registros duplicados:** Total de registros duplicados na tabela (mesma regra acima).

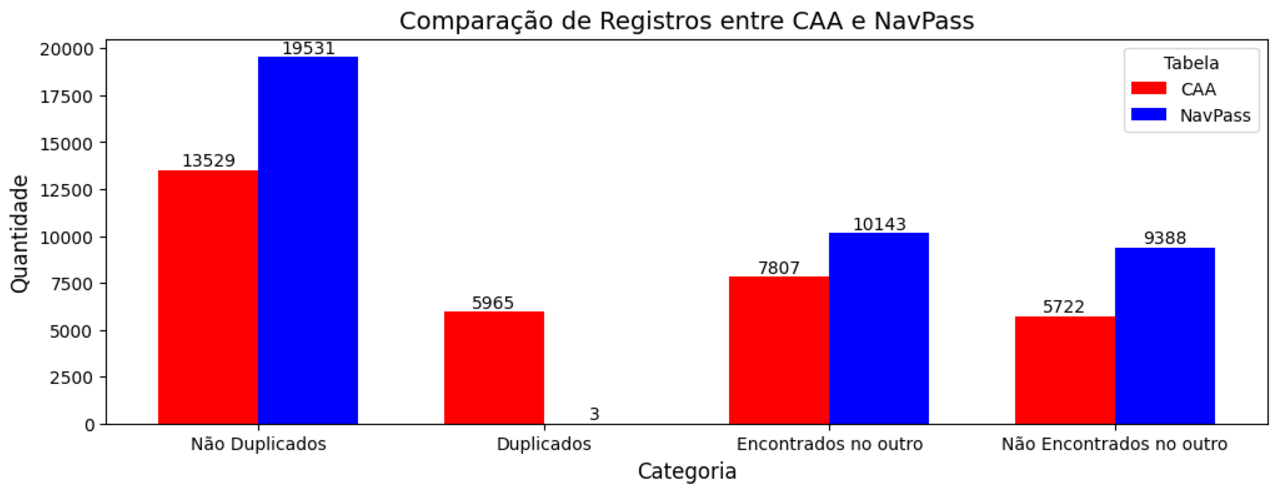
**Quantidade de registros encontrados do X no Y:** Quantidade de registros que estão na tabela X e que foram encontrados pelas chaves na tabela Y.

**Quantidade de registros NÃO encontrados do X no Y:** Quantidade de registros que estão na tabela X e que NÃO foram encontrados pelas chaves na tabela Y.

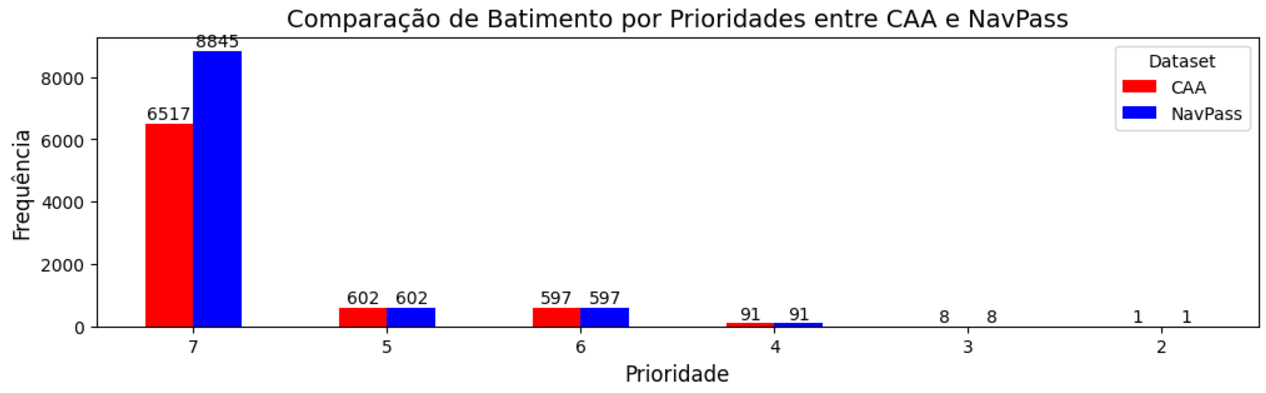
Esses registros não encontrados em ambas tabelas são os casos que apenas com a chave principal não encontrou nada (sem o dado de data).

Todos os dados foram salvos respectivamente em: 

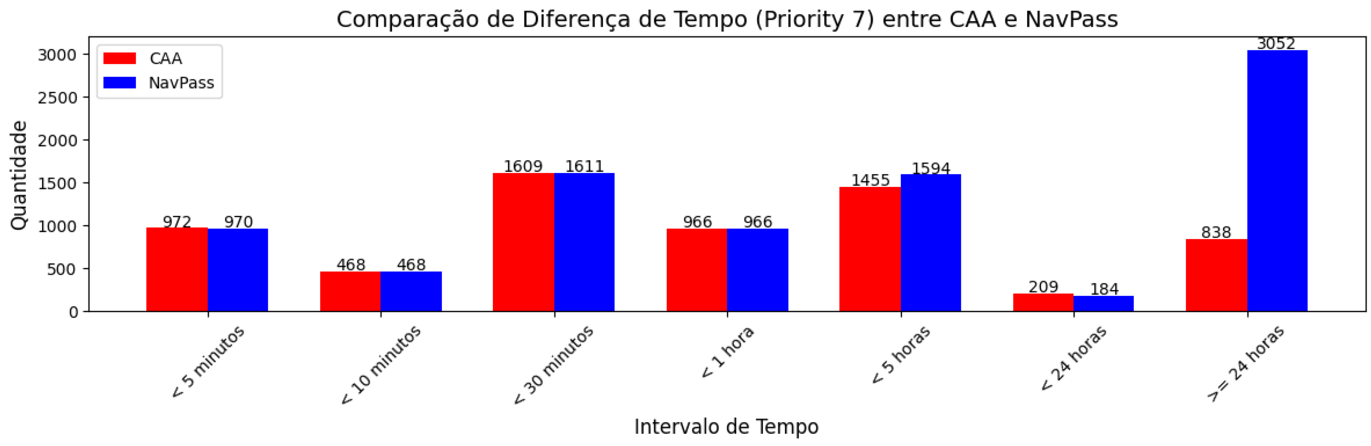
**Resultados obtidos diante todas as comparações:**



**Comparações de prioridades obtidas em ambos datasets (siga a lista citada mais acima para saber o que significa cada nível):**



**Com os dados classificados como prioridade 7, a busca de dados na outra tabela houve correspondência significante de:**



Esse gráfico acima representa o batimento dos dados não encontrados nas 2 tabelas pela chave com data (nível 7) – ou seja, tentamos encontrar a menor diferença possível entre as datas para conseguirmos interligar, e o resultado foi o acima, onde no eixo X você pode ver a diferença em minutos/horas.