Capgemini
ACELERAÇÃO
PySpark

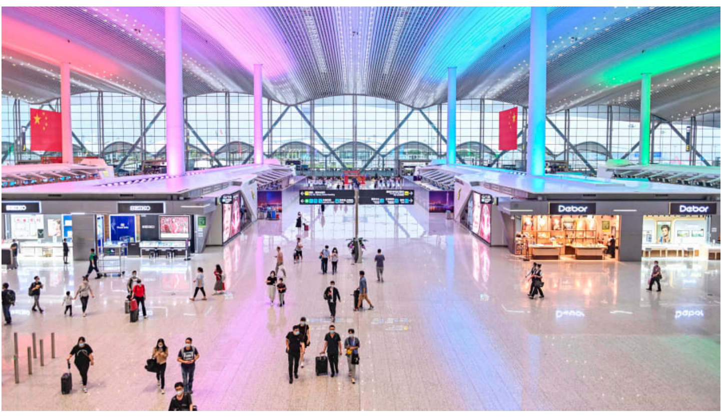
Qualidade

Antes de realizarmos transformações nos dados, é importante estabelecermos processos que apontem erros de qualidade nos dados que estamos trabalhado, dessa forma é possível ter clareza das inconsistências comuns e assim criar formas de melhorar a qualidade dos dados. Pensando nisso, a primeira atividade planejada é criarmos uma coluna adicional reportando o tipo de inconsistência que encontramos nos datasets.

Conteúdo

- Airports Dataset
  - Dicionário
  - Perguntas
- Planes Dataset
  - Dicionário
  - Perguntas
- Flights Dataset
  - Dicionário
  - Perguntas

Airports Dataset



Dicionário

- **faa** (string): Identificador do aeroporto determinado pela Federal Aviation Administration. Formato: 3-5 caracteres alfanuméricos.
- **name** (string): Nome do aeroporto.
- **lat** (float): Latitude do aeroporto. Intervalo de valores $[-180, 180]$.
- **lon** (float): Longitude do aeroporto Intervalo de valores $[-180, 180]$.
- **alt** (int): Altitude do aeroporto. Unidade de medida em pés. Intervalo de valores $[0, +\infty)$.
- **tz** (float): Fuso horário baseado no deslocamento de horas a partir de UTC/GMT. Intervalo de valores $[-11, +14]$. Pode ser fuso fracionário [1]
- **dst** (category): Horário de verão. Descrição dos possíveis valores [2]:
  - E (Europe)
  - A (US/Canada)
  - S (South America)
  - O (Australia)
  - Z (New Zealand)
  - N (None)
  - U (Unknown)

faa	name	lat	lon	alt	tz	dst
04G	Lansdowne Airport	41.130472	-80.619583	1044	-5	A
06A	Moton Field Municipal Airport	32.460572	-85.680028	264	-5	A
06C	Schaumburg Regional	41.989341	-88.101243	801	-6	A
06N	Randall Airport	41.431912	-74.391561	523	-5	A
09J	Jekyll Island Airport	31.074472	-81.427778	11	-4	A

Perguntas

Considere o dataset `airports.csv` para realizar as seguintes tarefas:

1. Crie a coluna `qa_faa` e aponte inconsistências da coluna `faa` de acordo com as regras abaixo.
   - `M`: Indica que está com dado faltante.
   - `F`: Indica que não respeita o formator de 3-5 caracteres alfanuméricos.

2. Crie a coluna `qa_name` e aponte inconsistências da coluna `name` de acordo com as regras abaixo.
   - `M`: Indica que está com dado faltante.

3. Crie a coluna `qa_lat` e aponte inconsistências da coluna `lat` de acordo com as regras abaixo.
   - `M`: Indica que está com dado faltante.
   - `I`: Indica que o valor excede o intervalo $[-180, 180]$.
   - `A`: Indica que o valor é alfanumérico.

4. Crie a coluna `qa_lon` e aponte inconsistências da coluna `lon` de acordo com as regras abaixo.
   - `M`: Indica que está com dado faltante.
   - `I`: Indica que o valor excede o intervalo $[-180, 180]$.
   - `A`: Indica que o valor é alfanumérico.

5. Crie a coluna `qa_alt` e aponte inconsistências da coluna `alt` de acordo com as regras abaixo.
   - `M`: Indica que está com dado faltante.
   - `I`: Indica que o valor excede o intervalo $[0, +\infty)$.
   - `A`: Indica que o valor é alfanumérico.

6. Crie a coluna `qa_tz` e aponte inconsistências da coluna `tz` de acordo com as regras abaixo.
   - `M`: Indica que está com dado faltante.
   - `I`: Indica que o valor excede o intervalo $[-11, +14]$.
   - `A`: Indica que o valor é alfanumérico.

7. Crie a coluna `qa_dst` e aponte inconsistências da coluna `dst` de acordo com as regras abaixo.
   - `M`: Indica que está com dado faltante.
   - `C`: Indica que o valor não pertence a nenhuma das categorias esperadas: E, A, S, O, Z, N, U
   - `N`: Indica que o valor é numérico.

Planes Dataset

Dicionário

- **tailnum** (string): Identificação do avião. Formato "N-Number", é composto por 5-6 caracteres.
  - Primeira letra é sempre "N".
  - De 1 a 4 digitos seguidos por 1 letra (ex. N1234Z).
  - De 1 a 3 digitos seguidos por 2 letras (ex. N123AZ).
  - Não deve conter 0 (zero) como primeiro digito, e não deve conter as letras "I" ou "O".
- **year** (int): Ano de fabricação do avião. Intervalo de valores $[1950, +\infty]$
- **type** (string): Tipo do avião.
- **manufacturer** (string): Nome do fabricante.
- **model** (string): Modelo do avião
- **engines** (int): Número de motores. Intervalo de valores $[1, 4]$
- **seats** (int): Número de assentos. Intervalo de valores $[2, 500]$
- **speed** (int): Velocidade média de cruseiro. Unidade de medida em milhas. Intervalo de valores

$[50, 150]$

- **engine** (category): Tipo de motor.

tailnum	year	type	manufacturer	model	engines	seats	speed	engine
N102UW	1998	Fixed wing multi engine	AIRBUS INDUSTRIE	A320-214	2	182		Turbo-fan
N103US	1999	Fixed wing multi engine	AIRBUS INDUSTRIE	A320-214	2	182		Turbo-fan
N104UW	1999	Fixed wing multi engine	AIRBUS INDUSTRIE	A320-214	2	182		Turbo-fan
N105UW	1999	Fixed wing multi engine	AIRBUS INDUSTRIE	A320-214	2	182		Turbo-fan
N107US	1999	Fixed wing multi engine	AIRBUS INDUSTRIE	A320-214	2	182		Turbo-fan

Perguntas

Considere o dataset `planes.csv` para realizar as seguintes tarefas:

1. Crie a coluna `qa_tailnum` e aponte inconsistências da coluna `tailnum` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `S` : Indica que não tem exatamente 5 caracteres.
    - `F` : Indica que não respeita o formato esperado (ex. N1234Z ou N123AZ).
    - `FN` : Indica que não inicia com a letra "N".
    - `FE` : Indica que contém caracteres inválidos ("I", "O", ou 0 como primeiro digito).

2. Crie a coluna `qa_year` e aponte inconsistências da coluna `year` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `I` : Indica que o valor excede o intervalo $[1950, +\infty)$.

3. Crie a coluna `qa_type` e aponte inconsistências da coluna `type` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `C` : Indica que o valor não pertence a nenhuma categoria esperada:
        - Fixed wing multi engine
        - Fixed wing single engine
        - Rotorcraft

4. Crie a coluna `qa_manufacturer` e aponte inconsistências da coluna `manufacturer` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `C` : Indica que o valor não pertence a nenhuma categoria esperada:
        - AIRBUS
        - BOEING
        - BOMBARDIER
        - CESSNA
        - EMBRAER
        - SIKORSKY
        - CANADAIR
        - PIPER
        - MCDONNELL DOUGLAS
        - CIRRUS
        - BELL
        - KILDALL GARY
        - LAMBERT RICHARD
        - BARKER JACK
        - ROBINSON HELICOPTER
        - GULFSTREAM
        - MARZ BARRY

5. Crie a coluna `qa_model` e aponte inconsistências da coluna `model` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `F` : Indica que não respeita o formato esperado
        - Modelos AIRBUS devem começar com "A"
        - Modelos BOEING devem começar com "7"
        - Modelos BOMBARDIER e CANADAIR devem começar com "CL"
        - Modelos MCDONNELL DOUGLAS devem começar com "MD" ou "DC"

6. Crie a coluna `qa_engines` e aponte inconsistências da coluna `engines` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `I` : Indica que o valor excede o intervalo $[1, 4]$.
    - `A` : Indica que o valor é alfanumérico.

7. Crie a coluna `qa_seats` e aponte inconsistências da coluna `seats` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `I` : Indica que o valor excede o intervalo $[2, 500]$.
    - `A` : Indica que o valor é alfanumérico.

8. Crie a coluna `qa_speed` e aponte inconsistências da coluna `speed` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `I` : Indica que o valor excede o intervalo $[50, 150]$.
    - `A` : Indica que o valor é alfanumérico.
9. Crie a coluna `qa_engine` e aponte inconsistências da coluna `engine` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `C` : Indica que o valor não pertence a nenhuma categoria esperada:
        - Turbo-fan
        - Turbo-jet
        - Turbo-prop
        - Turbo-shaft
        - 4 Cycle

Flights Dataset

Dicionário

- **year** (int), **month** (int), **day** (int): Ano, Mês, Dia de partida.
- **hour** (int), **minute** (int): Hora e Minuto agendada para partida.
- **dep_time** (string), **arr_time** (string): Horario real de partida/chegada do voo no horário local. Formato: HHMM ou HMM.
- **dep_delay** (int), **arr_delay** (int): Atraso de partida/chegada do voo em minutos. Valores negativos representam partidas/chegadas antecipadas.
- **carrier** (string): Identificador da empresa aérea.
- **tailnum** (string): Identificador do avião. Veja dataset planes.
- **flight** (string): Identificador do vôo. Formato: 4 dígitos (preenchidos com zero a esquerda caso necessário).
- **origin** (string), **dest** (string): Identificadores dos aeroportos de origem e destino. Veja dataset airports.
- **air_time** (int): Tempo de vôo. Unidade de medida em minutos. Intervalo de dados $[20, 500]$.
- **distance** (int): Distancia entre aeroportos. Unidade de medida em milhas. Intervalo de valores $[50, 3000]$.

year	month	day	hour	minute	dep_time	dep_delay	arr_time	arr_delay	carrier	tailnum	flight	origin	de
2014	12	8	6	58	658	-7	935	-5	VX	N846VA	1780	SEA	LA
2014	1	22	10	40	1040	5	1505	5	AS	N559AS	851	SEA	HN
2014	3	9	14	43	1443	-2	1652	2	VX	N847VA	755	SEA	SF
2014	4	9	17	5	1705	45	1839	34	WN	N360SW	344	PDX	SJ
2014	3	9	7	54	754	-1	1015	1	AS	N612AS	522	SEA	BU

Perguntas

Considere o dataset `flights.csv` para realizar as seguintes tarefas:

1. Crie a coluna `qa_year_month_day` e aponte inconsistências das colunas `year`, `month`, `day` de acordo com as regras abaixo.
    - `MY` : Indica que está com dado faltante no ano.
    - `MM` : Indica que está com dado faltante no mes.
    - `MD` : Indica que está com dado faltante no dia.
    - `IY` : Indica que o valor excede o intervalo $[1950, +\infty)$ no ano.
    - `IM` : Indica que o valor excede o intervalo $[1, 12]$ no mês.
    - `ID` : Indica que o valor excede o intervalo $[1, 31]$ no dia. No mês de Fevereiro o intervalo é $[1, 29]$

.

2. Crie a coluna `qa_hour_minute` e aponte inconsistencias das colunas `hour` e `minute` de acordo com as regras abaixo.
    - `MH` : Indica que está com dado faltante na hora.
    - `MM` : Indica que está com dado faltante no minuto.
    - `IH` : Indica que o valor excede o intervalo $[0, 24]$ na hora.
    - `IM` : Indica que o valor excede o intervalo $[0, 59]$ no minuto.

3. Crie a coluna `qa_dep_arr_time` e aponte inconsistências da coluna `dep_time` e `arr_time` de acordo com as regras abaixo.
    - `MD` : Indica que está com dado faltante no `dep_time` .
    - `MA` : Indica que está com dado faltante no `arr_time` .
    - `FD` : Indica que não respeita o formato esperado (HHMM ou HMM) no `dep_time` .
    - `FA` : Indica que não respeita o formato esperado (HHMM ou HMM) no `arr_time` .

4. Crie a coluna `qa_dep_arr_delay` e aponte inconsistências da coluna `dep_delay` e `arr_delay` de acordo com as regras abaixo.
    - `MD` : Indica que está com dado faltante no `dep_delay` .
    - `MA` : Indica que está com dado faltante no `arr_delay` .

5. Crie a coluna `qa_carrier` e aponte inconsistências da coluna `carrier` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `F` : Indica que não respeita o formato esperado (2 caracteres alfanuméricos).

6. Crie a coluna `qa_tailnum` e aponte inconsistências da coluna `tailnum` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `S` : Indica que não tem o número de caracteres esperado.
    - `F` : Indica que não respeita o formato esperado (ex. N1234Z ou N123AZ).
    - `FN` : Indica que não inicia com a letra "N".
    - `FE` : Indica que contém caracteres inválidos ("I", "O", ou 0 como primeiro digito).

7. Crie a coluna `qa_flight` e aponte inconsistências da coluna `flight` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `F` : Indica que não respeita o formato esperado (4 caracteres numéricos).

8. Crie a coluna `qa_origin_dest` e aponte inconsistências da coluna `origin` , `dest` de acordo com as regras abaixo.
    - `MO` : Indica que está com dado faltante no `origin` .
    - `MD` : Indica que está com dado faltante no `dest` .
    - `FO` : Indica que não respeita o formato esperado (3 caracteres alfanuméricos) no `origin` .
    - `FD` : Indica que não respeita o formato esperado (3 caracteres alfanuméricos) no `dest` .

9. Crie a coluna `qa_air_time` e aponte inconsistencias da coluna `air_time` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `I` : Indica que o valor excede o intervalo $[20, 500]$

10. Crie a coluna `qa_distance` e aponte inconsistências da coluna `distance` de acordo com as regras abaixo.
    - `M` : Indica que está com dado faltante.
    - `I` : Indica que o valor excede o intervalo $[50, 3000]$.

11. Crie a coluna `qa_distance_airtime` e aponte inconsistências entre as colunas `distance` e `air_time` de acordo com as regras abaixo.
    - `M` : Indica que está com `distance` ou `air_time` faltante.
    - `TL` : Indica que a viagem é longa de acordo com a condição: $\text{air\_time} >= \text{distance } \times 0.1 + 30$.
    - `TS` : Indica que a viagem é curta de acordo com a condição: $\text{air\_time} <= \text{distance } \times 0.1 + 10$.
    - `TR` : Indica que a viagem é normal caso as duas anteriores não sejam verdade.