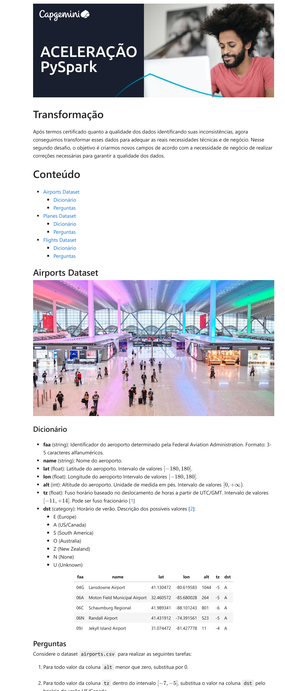


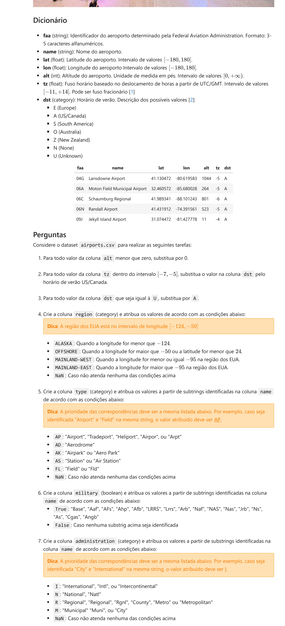
Transformação

Após termos certificado quanto a qualidade dos dados identificando suas inconsistências, agora conseguimos transformar esses dados para adequar as reais necessidades técnicas e de negócio. Nesse segundo desafio, iremos trabalhar 4 criarmos novos campos de acordo com a necessidade de negócio de realizar correções necessárias para garantir a qualidade dos dados.

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, +∞].
- **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. Para todo valor da coluna `alt` menor que zero, substitua por 0.
2. Para todo valor da coluna `tz` dentro do intervalo [−7, −5], substitua o valor na coluna `dst` pelo horário de verão US/Canada.
3. Para todo valor da coluna `dst` que seja igual à `U`, substitua por `A`.
4. Crie a coluna `region` (category) e atribua os valores de acordo com as condições abaixo:

   Dica: A região dos EUA está no intervalo de longitude [−124, −50]

   - `ALASKA` : Quando a longitude for menor que −124.
   - `OFFSHORE` : Quando a longitude for maior que −50 ou a latitude for menor que 24.
   - `MAINLAND-WEST` : Quando a longitude for menor ou igual −95 na região dos EUA.
   - `MAINLAND-EAST` : Quando a longitude for maior que −95 na região dos EUA.
   - `N/A` : Caso não atenda nenhuma das condições acima
5. Crie a coluna `type` (category) e atribua os valores a partir de subtrings identificadas na coluna `name` de acordo com as condições abaixo:

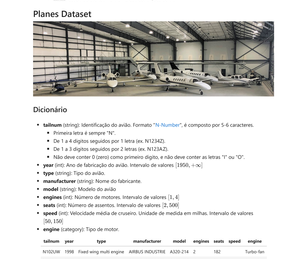
   Dica: A prioridade das correspondências deve ser a mesma listada abaixo. Por exemplo, caso seja identificada "Airport" e "Field" na mesma string, o valor atribuído deve ser AP.

   - `AP` : "Airport", "Tradeport", "Heliport", "Airpor", ou "Arpt"
   - `AD` : "Aerodrome"
   - `AE` : "Airpark" ou "Aero Park"
   - `AS` : "Station" ou "Air Station"
   - `FL` : "Field" ou "Fld"
   - `NAN` : Caso não atenda nenhuma das condições acima
6. Crie a coluna `military` (boolean) e atribua os valores a partir de subtrings identificadas na coluna `name` de acordo com as condições abaixo:
   - `True` : "Base", "Aaf", "Afs", "Afrc", "Usfs", "Lrrs", "Arb", "Naf", "NAS", "Nas", "Jrb", "Ns", "Ac", "Cgas", "Angb"
   - `False` : Caso nenhuma subtring acima seja identificada
7. Crie a coluna `administration` (category) e atribua os valores a partir de subtrings identificadas na coluna `name` de acordo com as condições abaixo:

   Dica: A prioridade das correspondências deve ser a mesma listada abaixo. Por exemplo, caso seja identificada "City" e "International" na mesma string, o valor atribuído deve ser I.

   - `I` : "International", "Intl", ou "Intercontinental"
   - `N` : "National", "Natl"
   - `R` : "Regional", "Regional", "Rgnl", "County", "Metro" ou "Metropolitan"
   - `M` : "Municipal", "Muni", ou "City"
   - `NAN` : Caso não atenda nenhuma das condições acima

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 dígitos seguidos por 1 letra (ex. N123AZ).
  - De 1 a 3 dígitos seguidos por 2 letras (ex. N123AZ).
  - Não devem conter 0 (zero) como primeiro dígito, e não deve conter as letras "I" ou "O".
- **year** (int): Ano de fabricação do avião. Intervalo de valores [1950, +∞]
- **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 cruzeiro. 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 `tailchar` (string) e atribua os caracteres alfabéticos de `tailnum` (exceto o "N")
2. Na coluna `year`, substitua todo valor igual a 0 por 1996.
3. Na coluna `year`, impute todo valor nulo de acordo com a sequencia de regras abaixo:
   - Ordene por `manufacturer`, `model` e `year`.
   - Use o valor da primeira linha anterior que compartilhe os mesmos valores das seguintes colunas, em prioridade:
     1. `manufacturer` e `model`
     2. `manufacturer`.
4. Crie a coluna `age` (int) e atribua a idade do avião com base na coluna `year` e o ano atual.
5. Para todo valor da coluna `type` substitua o valor na coluna `type` de acordo com as regras abaixo:
   - "Fixed wing multi engine": `MULTI_ENG`.
   - "Fixed wing single engine": `SINGLE_ENG`
   - "Rotorcraft": `ROTORCRAFT`
6. Para todo valor da coluna `manufacturer` substitua o valor na coluna `manufacturer` de modo que exista somente os valores únicos listados abaixo:
   - AIRBUS
   - BOEING
   - BOMBARDIER
   - CESSNA
   - EMBRAER
   - SIKORSKY
   - CANADAIR
   - PIPER
   - MCDONNELL DOUGLAS
   - CIRRUS
   - BELL
   - KILDALL GARY
   - LAMBERT RICHARD
   - BARKER JACK
   - ROBINSON HELICOPTER
   - GULFSTREAM
   - MARZ BARRY
7. Remova todos caracteres entre parenteses da coluna `model` mantendo somente os valores fora do parenteses.

   Dica: Os valores alterados não devem ficar com espaço no fim e no início da string ou com parenteses.

8. Impute os valores nulos da coluna `speed` com base na fórmula $\left\lfloor \frac{mean}{10} \right\rfloor * 10$
9. Crie a coluna `engine_type` (category) e atribua os valores a partir de substrings identificadas na coluna `engine` de acordo com as condições abaixo:
   - `FAN` : "Turbo-fan"
   - `JET` : "Turbo-jet"
   - `PROP` : "Turbo-prop"
   - `SHAFT` : "Turbo-shaft"
   - `CYCLE` : "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 MMM.
- **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 voo. 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 voo. 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	dest
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	HF
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	N613AS	522	SEA	BL

Perguntas

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

1. Impute os valores nulos das colunas `hour` e `minute` por 0.
2. Substitua os valores iguais a 24 da coluna `hour` por 0.
3. Crie a coluna `dep_datetime` (datetime) usando as colunas `year`, `month`, `day`, `hour`, `minute` no formato `YYYY-MM-DD HH:MM:00`.
4. Impute os valores nulos da coluna `dep_time` usando as colunas `hour`, `minute` de acordo com o formato esperado pela coluna `dep_time`.
5. Impute os valores nulos da coluna `dep_delay` por 0.
6. Impute os valores nulos da coluna `arr_delay` por 0.
7. Remova as colunas `year`, `month`, `day`, `hour`, `minute`.
8. Crie a coluna `air_time_projected` (int) de acordo com a fórmula distance `* 0.1 + 20`
9. Crie a coluna `air_time_expected` (int) de acordo com a média de valores dos voos com mesma origem e destino
10. Impute os valores da coluna `air_time` de acordo com a regra **max(air_time_projected , air_time_expected)**
11. Impute os valores nulos da coluna `arr_time` de acordo com a formula `dep_time` + `air_time`
12. Crie a coluna `haul_duration` (category) com base na coluna `air_time` de acordo com as regras abaixo
    - `SHORT-HAUL` : 20 min - 3 horas
    - `MEDIUM-HAUL` : 3 horas - 6 horas
    - `LONG-HAUL` : 6 horas+
13. Crie a coluna `dep_season` (category) com base na coluna `dep_datetime` de acordo com as regras abaixo:
    - `SPRING` : De 21 de Dez às 09:48 PM até 20 de Mar às 03:33 PM
    - `SUMMER` : De 20 de Mar às 03:33 PM até 21 de Jun às 10:14 AM
    - `FALL` : De 21 de Jun às 10:14 AM até 22 de Set às 10:04 AM
    - `WINTER` : De 23 de Set às 02:04 AM até 21 de Dez às 09:48 PM
14. Crie a coluna `dep_delay_category` (category) com base na coluna `dep_delay` de acordo com as regras abaixo:
    - `ANTECIPATED` : Menor que 0.
    - `INTIME` : Igual a 0
    - `MINOR` : Maior que 0 e menor que 60.
    - `MAJOR` : Maior ou igual a 60.