

ANÁLISE EXPLORATÓRIA PELO PANDAS PROFILING

Base “airlines_data”:

A base possui apenas duas colunas (código e nome da companhia aérea) e não possui dados faltantes ou nulos. Logo, não foi preciso tratamento dos dados.

Base “airports_data”:

A base possui 7 colunas (código, nome, cidade, estado, país, latitude e longitude do aeroporto). Apresenta alguns dados faltantes, mas em colunas que foram, posteriormente, não selecionadas para análise. Logo, não foi preciso tratamento dos dados.

Base “flights_data”:

Após analisar os dados a partir do relatório gerado pelo Pandas Profiling, essa base apresentou alguns problemas em seus dados. Portanto, algumas colunas foram removidas por ter muitos dados faltantes ou por serem irrelevantes para os gráficos e relatórios mais tarde gerados.

- **COLONAS REMOVIDAS**

1. **Com 98.5% de dados faltantes:** 'CANCELLATION_REASON'
2. **Com 81.7% de dados faltantes:** 'AIR_SYSTEM_DELAY', 'SECURITY_DELAY', 'AIRLINE_DELAY', 'LATE_AIRCRAFT_DELAY' e 'WEATHER_DELAY'
3. **Colunas definidas como irrelevantes para análise:** 'YEAR', 'FLIGHT_NUMBER', 'TAIL_NUMBER', 'TAXI_OUT', 'TAXI_IN', 'WHEELS_OFF', 'WHEELS_ON', 'DIVERTED', 'CANCELLED'

- **DICIONÁRIO DAS COLUNAS RESTANTES**

MONTH: Mês do voo

DAY: Dia do mês do voo

DAY_OF_WEEK: Dia da semana do voo

AIRLINE: Código IATA da companhia aérea

ORIGIN_AIRPORT: Código IATA do aeroporto de origem

DESTINATION_AIRPORT: Código IATA do aeroporto de destino

SCHEDULED_DEPARTURE: Hora programada de partida (HHMM)

DEPARTURE_TIME: Hora efetiva de partida (HHMM)

DEPARTURE_DELAY: Minutos de atraso na partida

SCHEDULED_TIME: Tempo programado de voo (minutos)

ELAPSED_TIME: Tempo total de voo (minutos)

AIR_TIME: Tempo de voo no ar (minutos)

DISTANCE: Distância do voo (milhas)

SCHEDULED_ARRIVAL: Hora programada de chegada (HHMM)

ARRIVAL_TIME: Hora efetiva de chegada (HHMM)

ARRIVAL_DELAY: Minutos de atraso na chegada (valor negativo: minutos antecipados // valor positivo: minutos atrasados)