Passos seguidos durante a execução do projeto

Equipe: Ana Clara Fontes, Heitor Negromonte e Matheus Felipe
Base de dados: https://www.kaggle.com/datasets/usdot/flight-delays?ref=hackernoon.com
Tecnologias utilizadas: VS Code, Mysql Workbench, SQL Server e Power BI.

Definindo relatório e gráficos

- **RELATÓRIOS**
    1. Eficiência de voo e tempo programado por companhia aérea
    2. Quantidade de voos atrasados por companhia aérea
    3. Quantidade de voos atrasados por mês
    4. Quantidade de voos que chegaram antecipadamente por dia da semana

- **GRÁFICOS:**
    1. Quantidade de voos por companhia aérea (gráfico de segmentação + cartão)
    2. Correlação entre a distância e minutos de atraso na chegada (gráfico de dispersão)
    3. Média de atrasos em minutos por distâncias agrupadas (histograma)
    4. Contagem de voos por distância (histograma)
    5. Correlação entre atraso na partida e atraso na chegada (gráfico de dispersão)
    6. Taxa de voos com/sem atraso (gráfico pizza)

Definindo estrutura do modelo dimensional

- **Tabelas Dimensão:**
    1. dim_airline: id, código e nome da companhia aérea
    2. dim_airport: id, código, nome, cidade, estado do aeroporto
    3. dim_date: id da data, mês, dia e dia da semana
    4. dim_flight: id do voo, aeroporto de origem, aeroporto de destino, id da companhia, hora programada de partida, hora programada de chegada, hora efetiva de partida e hora efetiva de chegada

- **Tabela de Fatos (atraso do voo):** id do fato, id da data, id do voo, id do aeroporto de partida, id do aeroporto de destino, tempo de atraso na chegada, tempo de voo, tempo programado de voo, distância, minutos de atraso na partida

Processo de ETL

Posterior à escolha das fontes de dados que seriam utilizadas (“airlines_data”, “aiports_data” e “flights_data”) e à análise exploratória dos dados, foi feito o processo de tratamento dos dados contidos na base “flights_data”, no qual consistiu em, por ser uma base com grande volume de linhas (+ de 500.000), remover as que tinham valores nulos e algumas colunas que se mostraram ter muitos dados faltantes ou eram irrelevantes.

Além disso, ao tentar realizar o processo de carga, percebemos a necessidade de diminuir o volume de dados para minimizar problemas com as consultas SQL. Para isso,

primeiramente foram embaralhadas entre si as linhas da base, pois estavam ordenadas mensalmente e para cada mês tinham milhares de dados. Fazer isso evitou que, ao diminuir a quantidade de linhas, ficassem somente os dados de alguns meses. A seguir, diminuímos o volume dos dados, restando 8.212 das mais de 500.000 linhas que a base principal continha.

Finalmente, avançando para a etapa de carga dos dados no Data Warehouse, feita no MySQL Workbench, utilizamos o script de criação de cada tabela de dimensão e fato e inserimos os dados dentro de cada uma delas extraindo das bases“airlines_data”, “aiports_data” e “flights_data”.

Análise do BI desenvolvido

O Schema criado no MySQL foi migrado para o SQL Server para facilitar a importação no Power BI e, por fim, as tabelas de dimensão e fato foram importadas para o Power BI para gerar os gráficos e relatórios previamente definidos.

Análise do relatório 1:

Eficiência de voo e tempo programado por companhia aérea

airline_name	Soma de air_time	Soma de scheduled_time	diferenca	ranking	flights_efficiency	ranking_efficiency
American Eagle Airlines Inc.	27155	39298	-12143	7	69,10%	1
Atlantic Southeast Airlines	55676	75479	-19803	6	73,76%	2
Skywest Airlines Inc.	59291	80176	-20885	5	73,95%	3
Delta Air Lines Inc.	133205	170252	-37047	2	78,24%	4
Southwest Airlines Co.	180641	223505	-42864	1	80,82%	5
American Airlines Inc.	135917	167232	-31315	3	81,27%	6
Hawaiian Airlines Inc.	7974	9729	-1755	14	81,96%	7
US Airways Inc.	40851	49743	-8892	9	82,12%	8
United Air Lines Inc.	106734	129052	-22318	4	82,71%	9
JetBlue Airways	57572	69193	-11621	8	83,20%	10
Frontier Airlines Inc.	15992	19113	-3121	12	83,67%	11
Spirit Air Lines	17049	20294	-3245	11	84,01%	12
Alaska Airlines Inc.	40237	46343	-6106	10	86,82%	13
Virgin America	18766	21429	-2663	13	87,57%	14
Total	**897060**	**1120838**	**-223778**	**1**	**80,03%**	**5**

Esse relatório mostra a relação entre o tempo programado para voo e o real tempo de voo por companhia aérea. Foram criadas algumas medidas para fazer essas relações: “diferenca”, ”ranking”, ”flights_efficiency” e “raking_efficiency”.

A medida “diferenca” foi criada para calcular a diferença entre os dois valores analisados, retornando valores positivos para quando voo ultrapassasse o tempo programado (air_time > scheduled_time) e valores negativos para quando o tempo de voo fosse menor que o tempo programado (air_time < scheduled_time). É possível observar que todos os valores obtidos foram negativos, portanto o tempo de todos os voos foi menor que o esperado.

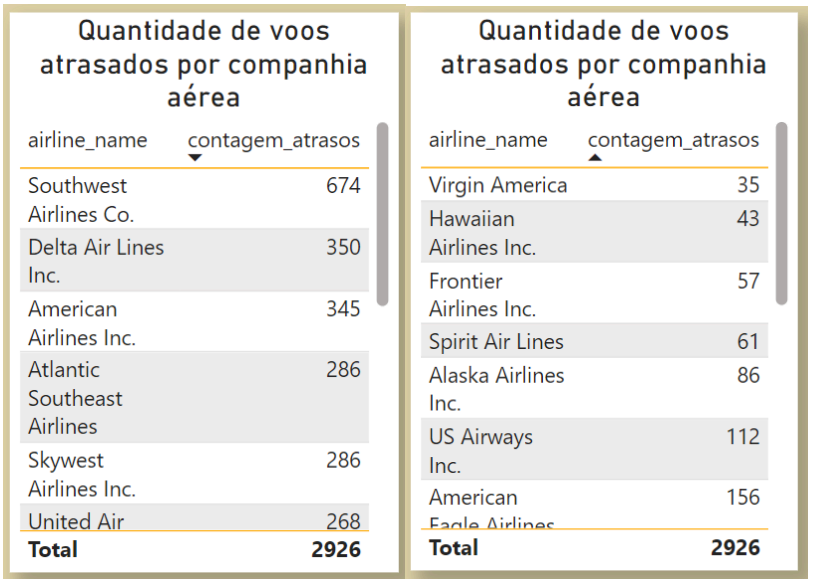
A medida “ranking”, classifica com o ícone verde companhias que mais se anteciparam em relação ao tempo de voo programado; com o amarelo as que foram medianas; e com vermelho, as que chegaram mais perto do tempo programado de voo.

A medida “flights_efficiency” faz uma divisão entre os dois valores, mostrando em quantos por cento do tempo programado o voo realmente aconteceu. Refletindo a eficiência obtida em todo o tempo de voo de cada companhia.

A medida “ranking_efficiency”, classifica com o ícone verde companhias que mais foram eficientes; com o amarelo as que foram medianas; e com vermelho, as que foram menos eficientes.

Insight interessante: Ter o menor valor na diferença, não significa ser a companhia com os voos mais eficientes. Por exemplo, apesar da companhia Southwest Airlines Co. ter sido classificada como a melhor no ranking da diferença, seus voos foram feitos, no geral, em 80.82% do tempo programado, então ela teve pouco mais que 19% de eficiência em relação a todo seu tempo de voo. Já a companhia American Eagle Airlines Inc., mesmo tendo sido classificada como mediana em relação à diferença, ela foi a mais eficiente em seu tempo total de voo, tendo pouco mais de 31% de eficiência.

Análise do relatório 2:



Quantidade de voos atrasados por companhia aérea

airline_name	contagem_atrasos
Southwest Airlines Co.	674
Delta Air Lines Inc.	350
American Airlines Inc.	345
Atlantic Southeast Airlines	286
Skywest Airlines Inc.	286
United Air	268
Total	**2926**

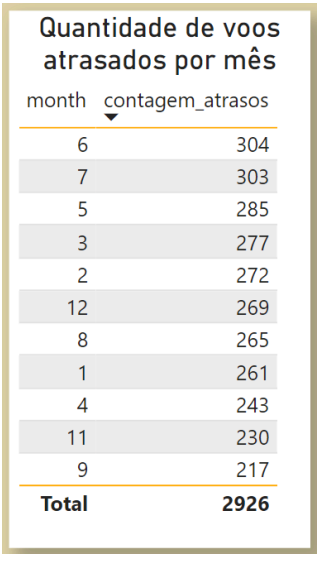
Quantidade de voos atrasados por companhia aérea

airline_name	contagem_atrasos
Virgin America	35
Hawaiian Airlines Inc.	43
Frontier Airlines Inc.	57
Spirit Air Lines	61
Alaska Airlines Inc.	86
US Airways Inc.	112
American Eagle Airlines	156
Total	**2926**

Foi criada uma medida (contagem_atrasos) para calcular a quantidade voos que chegaram atrasados. Tal quantidade foi obtida através do cálculo de, mediante todos os voos, quais tiveram o valor de atraso na chegada (arrival_delay) > 0, pois são os valores positivos que representam, em minutos, quanto o voo atrasou.

Insight: Através deste relatório, podemos concluir que a companhia aérea com mais voos atrasados foi a Southwest Airlines Co. (674) e que a com menos, foi a Virgin America (35).

Análise do relatório 3:



Quantidade de voos atrasados por mês

month	contagem_atrasos
6	304
7	303
5	285
3	277
2	272
12	269
8	265
1	261
4	243
11	230
9	217
Total	**2926**

Insight: Com a mesma medida que foi utilizada no relatório anterior, este relatório mostra que o mês em que mais houveram voos atrasados foi o mês 6 (Junho).

Análise do relatório 4:

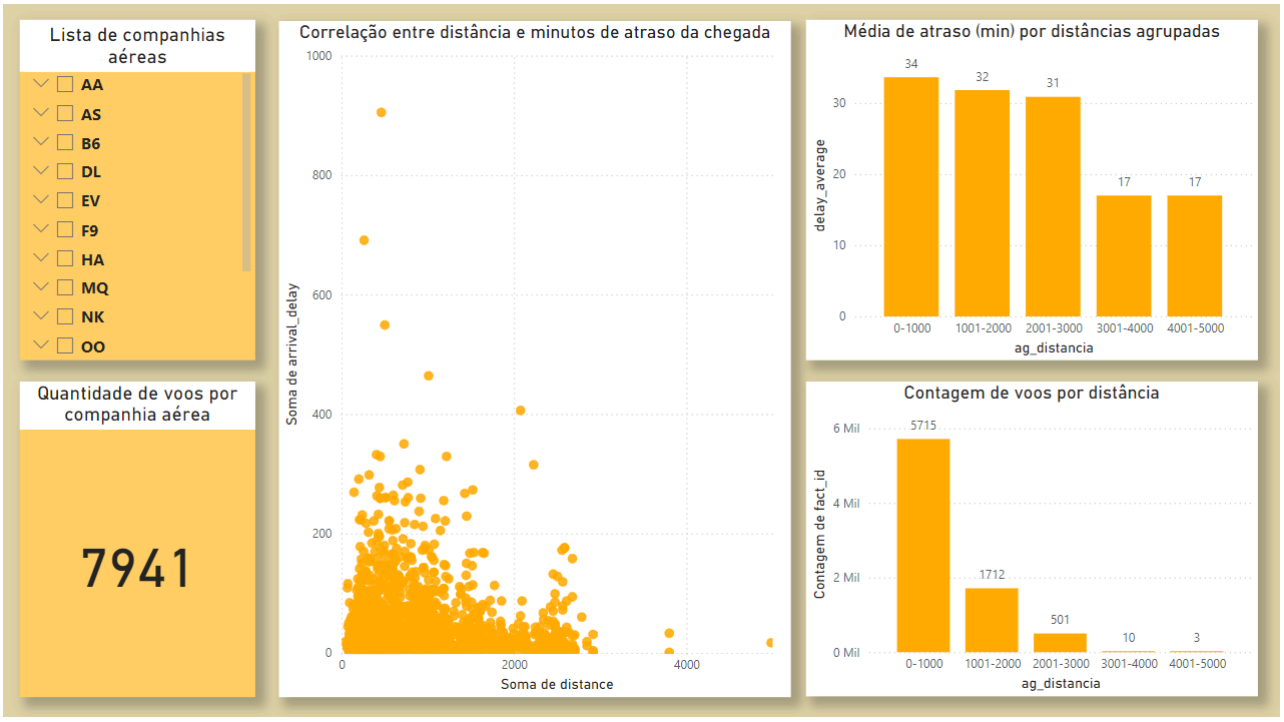
Quantidade de voos com chegada antecipada por dia da semana

day_of_week	contagem_sucessos
3	743
1	716
2	702
5	686
4	681
6	658
7	650
Total	**4836**

Para este relatório, foi criada uma medida que reflete o oposto da medida “contagem_atrasos”, trazendo a quantidade de voos que se anteciparam em sua chegada. Tal quantidade foi obtida através do cálculo de, mediante todos os voos, quais tiveram o valor de atraso na chegada (arrival_delay) < 0, pois são os valores negativos que representam, em minutos, quanto o voo se antecipou.

Insight: Através deste relatório, podemos concluir que o dia da semana com mais chegadas antecipadas foi o dia 3 (terça-feira) e o com menos, foi o dia 7 (sábado).

Análise do gráfico 1:



O gráfico 1, sendo o de segmentação (Lista de companhias aéreas) e o cartão, reflete a quantidade de voos por companhia aérea ao selecionar uma dada companhia na lista. O cartão, por padrão, mostra a quantidade total de voos.

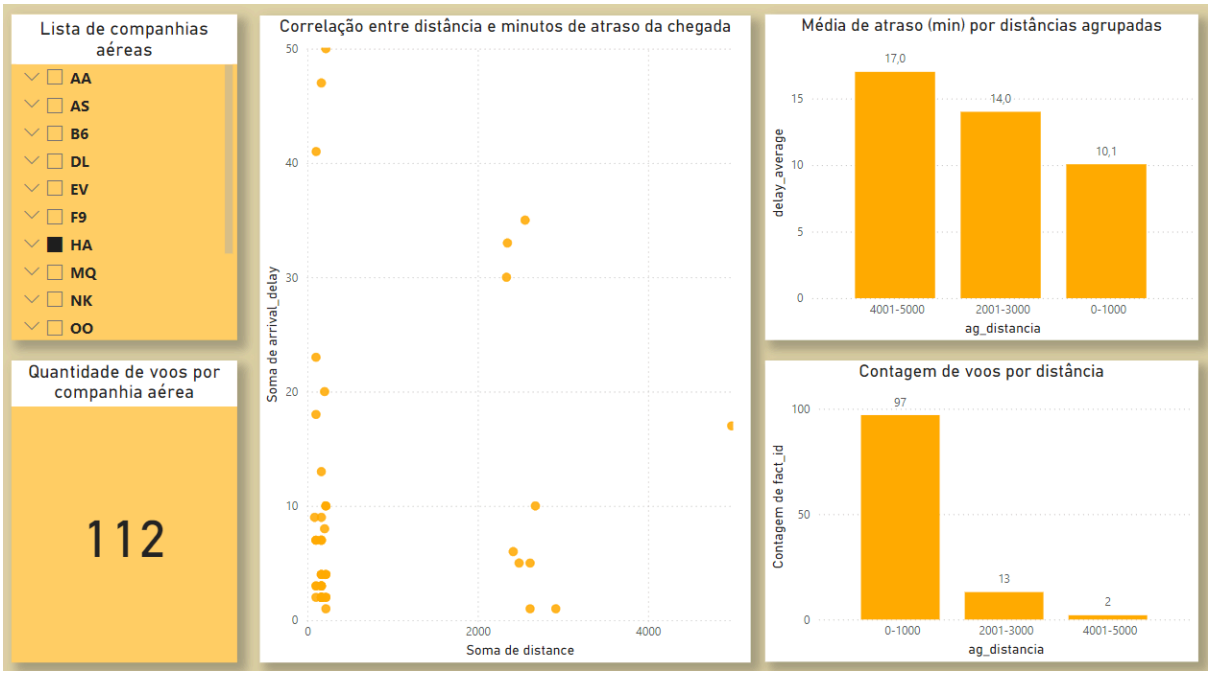
Análise do gráfico 2:
No gráfico 2, onde reflete a correlação entre a distância e os minutos de atraso na chegada dos voos, podemos observar que não há correlação e que os valores estão concentrados entre 0 e 3000 milhas, devido ao fato de haver uma quantidade maior de voos nesse intervalo (observação feita no gráfico 4).

Análise do gráfico 3:
Para entender melhor os atrasos em relação à distância dos voos, foi criado o gráfico 3. A medida “delay_avarage” retorna a média dos valores da coluna “arrival_delay” que forem > 0 (minutos atrasados). E para agrupar a distância em alguns intervalos, criamos uma coluna com uma função que agrupa os valores das distâncias e classifica cada valor de distância em um dos grupos, para cada classificação, uma linha da coluna. Colocando a medida e a nova coluna criada em conjunto, conseguimos ver a média de atraso por distância.

Análise do gráfico 4:
O gráfico 4 foi criado para explicitar a proporcionalidade entre a quantidade de voos e a distância.

Insight: Podemos inferir deste histograma, em conjunto com os gráficos 2 e 3, que terão mais atrasos dentro do intervalo de 0-3000 por terem mais voos e que a média de atrasos será maior nesses intervalos pelo mesmo motivo.

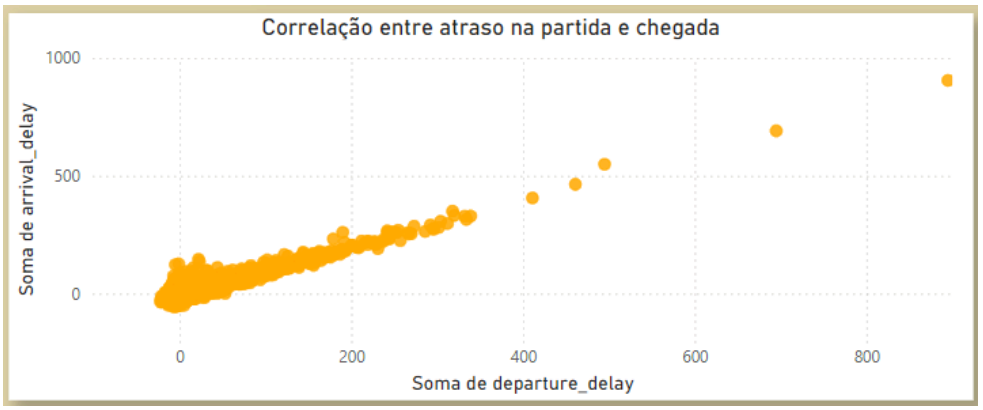
Insight interessante analisando os gráficos 1, 2, 3 e 4 com companhia HA:



A companhia aérea Hawaiian Airlines Inc. fez 112 voos (gráfico 1) e não houve correlação entre a distância e os minutos atrasados na chegada (gráfico 2). Dentre os voos feitos, 2 estavam entre 4001-5000 milhas de distância (gráfico 4) e os mesmos tiveram a média mais

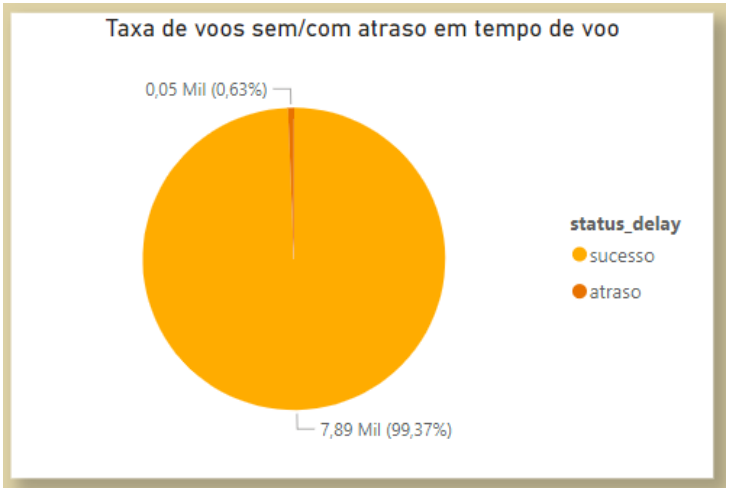
alta de atraso (gráfico 3). Inferimos que, nesse caso, os voos com maior distância foram os que mais atrasaram e que os de menor distância, atrasaram bem pouco.

Análise do gráfico 5:



Neste gráfico, podemos observar que há uma correlação significativa entre o atraso na partida e chegada do voo. Ou seja, se o voo atrasar em sua partida, muito provavelmente ele vai atrasar em sua chegada.

Análise do gráfico 6:



Este gráfico mostra a taxa de de voos no geral que tiveram sucesso ou atrasaram em relação ao tempo real de voo e o tempo programado. Para fazê-lo, foi criada uma coluna com a condição de que caso a divisão entre esses dois valores fosse < 0, o voo seria classificado como “sucesso” e caso contrário, “atraso”. É possível observar que uma porcentagem bem pequena de voos esteve atrasado.