Introdução

A mobilidade urbana é um dos principais problemas da sociedade atual, especialmente em grandes cidades. Assim sendo, a importância do papel desempenhado pelos aeroportos é maior do que nunca, assegurando a interligação entre cidades, países ou mesmo de continentes, impactando diretamente, tanto a economia, quanto a experiência dos passageiros. Problemas que levem a atrasos de voos, ou até mesmo cancelamentos, tais como atrasos em check-ins, sobrelotação dos aeroportos, condições climatéricas adversas, afetam não só os passageiros, mas também as companhias aéreas e aeroportos, quer de forma monetária, quer a sua reputação.

Compreender os fatores que contribuem para atrasos e cancelamentos de voos, as causas, e como estes influenciam os preços dos bilhetes, pode ajudar na procura de soluções mais eficientes.

**Objetivos do projeto**:

O projeto tem como objetivo analisar as possíveis causas dos atrasos e cancelamentos de voos e as suas possíveis causas com os preços dos mesmos.

* Determinar aeroportos com maior quantidade de atrasos e possíveis causas
* Determinar aeroportos com mais voos cancelados e possíveis causas
* Perceber quais as rotas mais utilizadas
* Analisar os preços dos bilhetes nos aeroportos mencionados
* Analisar tempo médio de atraso por aeroporto vs. distância média por origem

Metodologia

De forma a conseguir resultados mais consistentes, foram utilizados 2 datasets com dados de voos dos EUA. Um dos datasets contém dados de voos como, trimestre, origem do voo, destino, preço, etc., sendo estes dados do ano 2018. Como segunda fonte de dados, foi utilizado um dataset com detalhes de voos, atrasos e cancelamentos. Ambos os datasets foram encontrados na plataforma Kaggle.

No caso do dataset dos voos encontrámos um problema, com quase 10 milhões de linhas, o pySpark era capaz de ler o ficheiro CSV, porém quando pedíamos para fazer leituras de linhas específicas ou alterar certos dados, era devolvido um erro de falta de memória, este erro para além de parar a célula do notebook em questão também acabava por terminar a ligação criada inicialmente com o MongoDB. Por isso decidimos criar um script simples em Python para diminuir o tamanho desse dataset, primeiro seleciona apenas o terceiro trimestre de 2018 através da coluna “Quarter” e depois escolhe aleatoriamente 250 mil linhas, acabando por exportar para um novo ficheiro CSV.

Depois de feito este pequeno tratamento de dados acedemos à plataforma MongoDB Compass para importar os dados, utilizando a instância criada pelo container do Docker proporcionado para a realização do projeto. Dentro da aplicação criámos duas coleções uma para os voos outra para os atrasos e importámos o ficheiro correspondente a cada uma.

Já dentro do notebook, começamos por criar a ligação entre o pySpark e a instância de MongoDB hospedada no nosso computador, depois é só fazer a leitura das duas coleções para cada um dos dataframes que iremos analisar.

Após a leitura dos dados, foram selecionadas apenas as features consideradas úteis, em ambos os datasets, eliminando as restantes. No dataset dos atrasos, foi feita a seleção dos voos apenas de 2018, de forma a ir de encontro ao outro dataset. No caso do dataset dos voos, para além de apagar as colunas consideradas desnecessárias, normalizámos os títulos de todas as colunas para ficarem com letras minúsculas. Aplicámos filtros de limpeza em ambos os datasets, eliminando linhas duplicadas ou com valores nulos, acabando por verificar se resta algum. Por último, reiniciámos o index nos dois dataframes, isto tem de ser feito de forma diferente do que em Pandas, utilizando a função Window para tal.

No que toca à análise exploratória inicial, começámos por perceber quais os pontos mais interessantes para se estudar. Fazemos uma análise de atrasos, cancelamentos e preços de bilhetes, começando pela identificação dos aeroportos com mais atrasos. Primeiro, os dados de atrasos são agrupados por aeroporto, e a soma dos atrasos (arr\_del15) é calculada para cada um. Depois, os aeroportos são ordenados em ordem decrescente pelo total de atrasos, e os 5 principais aeroportos mais atrasados são selecionados. Com base nesses aeroportos, o dataframe de voos é filtrado para incluir apenas os voos com destino a eles, e calcula-se o preço médio das passagens para cada um desses destinos. Além disso, os dados de atrasos são agrupados novamente por aeroporto para calcular o total de voos, total de cancelamentos e a percentagem de voos cancelados. A partir disso, os 10 aeroportos com as maiores percentagens de cancelamentos são identificados. Por fim, o dataframe de voos é filtrado para incluir apenas os aeroportos mais cancelados, e calcula-se a média do preço das passagens para cada um desses destinos. Os resultados fornecem uma visão abrangente sobre atrasos, cancelamentos e preços médios de bilhetes nos aeroportos mais importantes.