Apresentação

Introdução

Na atualidade, trabalhamos com abundantes dados, mas sem conseguir tirar proveito disso, sem poder adquirir informação deles, para depois tomar boas decisões de negócios baseados na informação.
Como resposta disso, fiz uma análise dos dados do aeroporto de São Francisco e após entender eles, apliquei um algoritmo de Machine Learning que agrupe os voos com propriedades semelhantes, para conhecer quais são essas propriedades com maior influência e achar padrões nos voos com ligação ao aeroporto.

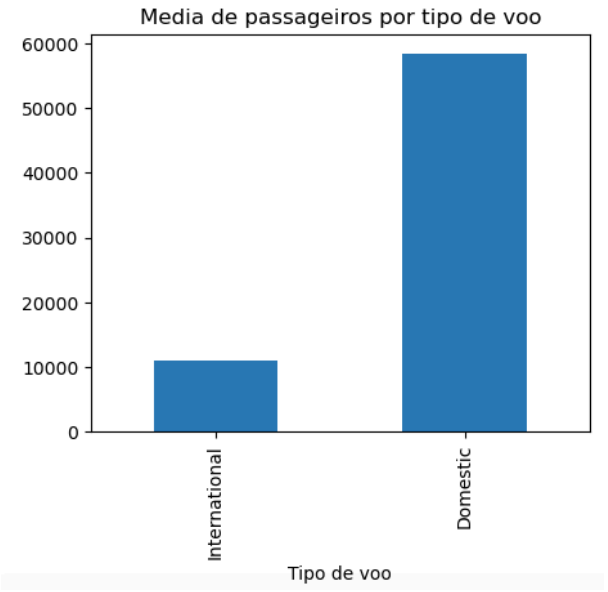
Objetivo

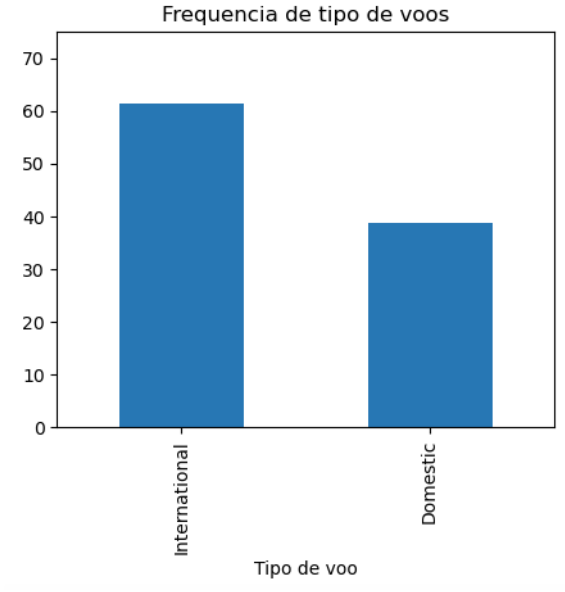
Esta apresentação visa conhecer os dados com os que trabalhamos diariamente, para conseguir achar padrões e caraterísticas relevantes através da aplicação de um modelo de aprendizagem não supervisionado.
Esperando que os resultados do modelo brindem informação para as diferentes equipas da empresa, e assim poder corrigir, melhorar ou mudar processos, tomando decisões de negócios em sintonia com os nossos clientes.

Com o objetivo em mente, vamos começar a conhecer os nossos dados, como a distribuição de algumas caraterísticas e a relação entre elas.

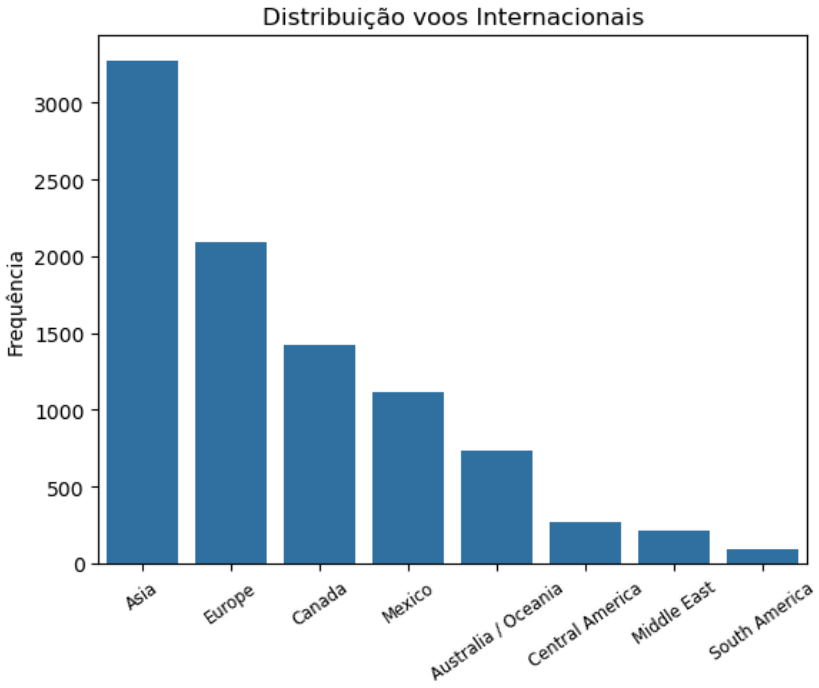
Conhecendo os dados

Para entender quais é o tipo de voos que tem maior número de passageiros e qual deles é mais frequente, temos os seguintes gráficos.



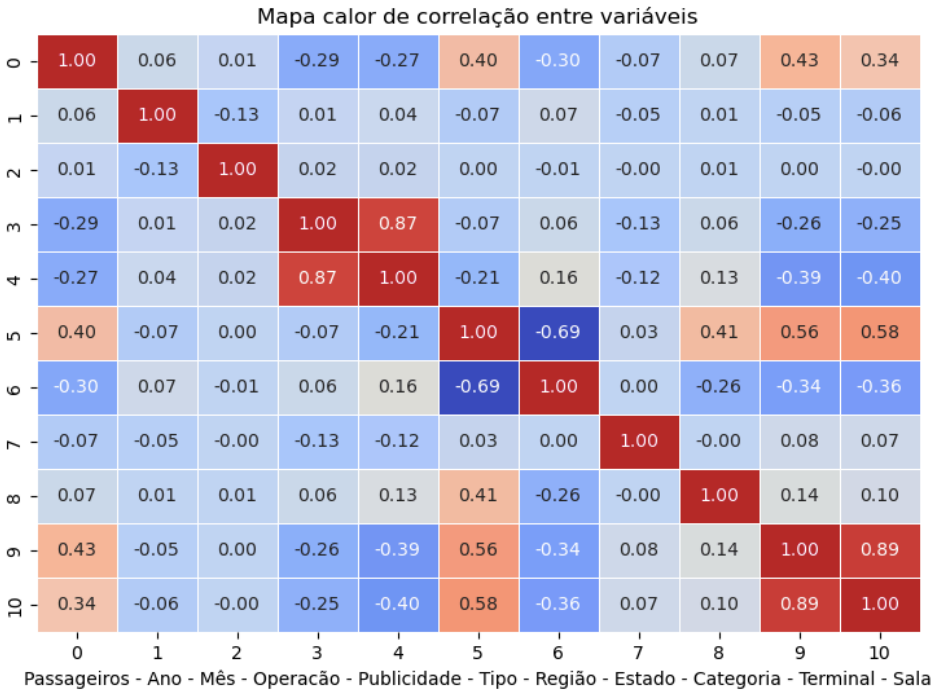


Podemos observar que no aeroporto de SF existe uma diferença notável na quantidade de passageiros entre voos Internacionais e Domésticos. E contrário a isso, a frequência dos voos Internacionais é maior, com mais dos 60% do total.
Sugerindo que o aeroporto de SF tem muitas rotas nacionais, que os locais utilizam frequentemente, e os estrangeiros fazem conexões nesse aeroporto, funcionando como porta de entrada para os Estados Unidos, devido a sua localização geográfica.



Continuando com a interpretação dos dados, e ligado à ideia anteriormente mencionada, dentro os voos internacionais, as rotas mais frequentes são as ligadas com Ásia, sugerindo também que este aeroporto oferece muitas conexões domésticas.
Também existem outros fatores, como a oferta turística desta cidade, diferente ao que existe no continente asiático, por exemplo. Ou variabilidade dos preços, onde podem ser mais baratos pela cercania.
Mas sim, fica mais claro que o aeroporto de São Francisco, é um aeroporto importante que existem múltiplas conexões.

Para poder explicar ainda melhor as relações entre os dados, vou demonstrar um mapa de calor onde o vermelho intenso demonstra uma forte relação direta entre caraterísticas, e a azul intenso demonstra o contrário.



Começando pela quantidade de passageiros, observamos que o tipo de voo (nacional ou internacional) é a caraterística que influencia nela, o que tem sentido pelo mencionado anteriormente, pela grande diferença entre estes voos. Consequente com isso, a região dos voos também afeta a quantidade de passageiros.
Ligado a isso, a terminal e sala de embarque também são propriedades que variam dependo da rota.

Observamos que o ano e mês não tem grande influência no conjunto de dados inteiro, sugerindo que neste aeroporto todo o ano têm uma demanda constante.

O tipo de voo além de ter uma relação com a quantidade de passageiros, vemos que a categoria do voo (baixo custo ou outros) sugere que maioritariamente, se o voo é doméstico também será de baixo custo, sugerindo que no país é mais frequente e mais barato fazer voos ‘low cost’.

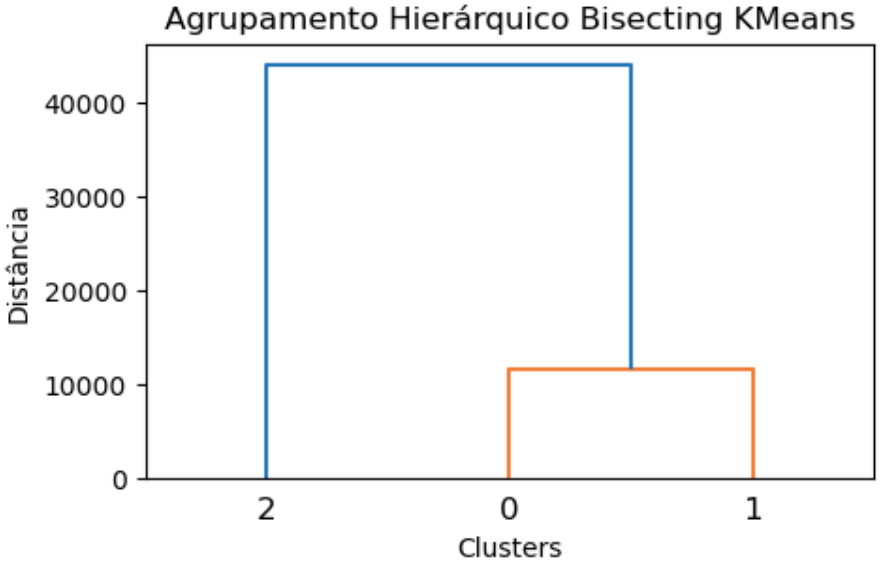
Existem relações diretas fortes entre as companhias operacionais e as companhias comerciais, algo que faz sentido já que na maioria das vezes as empresas que oferecem o serviço operacional é a mesma responsável pela comercialização desses serviços.

Após compreender os principais padrões e limitantes dos nossos dados, foi necessário aplicar um modelo de ML que auxilie no agrupamento de objetos com caraterísticas semelhantes.

Mas, antes de treinar o modelo e considerando que existe em enviesamento de passageiros nos voos domésticos, decidi filtrar os voos com ligações internacionais, e analisar eles. Considerando que esta análise pode ser aplicada aos voos domésticos num futuro.

Com base nestas necessidades, a melhor opção foi treinar um modelo de aprendizagem não supervisionado chamado BisectingKMeans, o qual se adequa à natureza do problema.

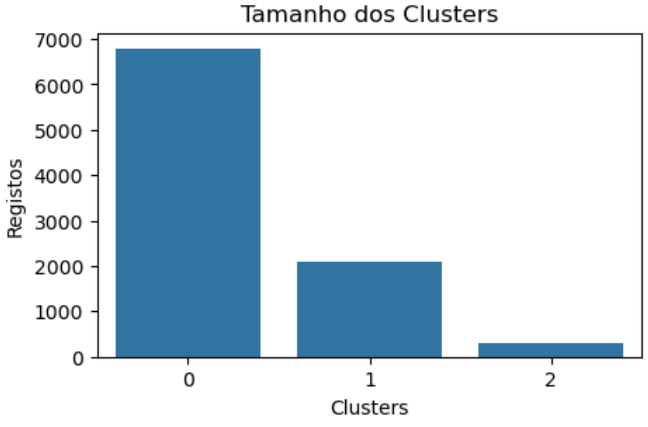
Este modelo agrupa hierarquicamente os dados, baseado na semelhança das propriedades deles, é dá uma hierarquia dentro dos agrupamentos. Na imagem podemos ver como foram separados os grupos criados pelo modelo.



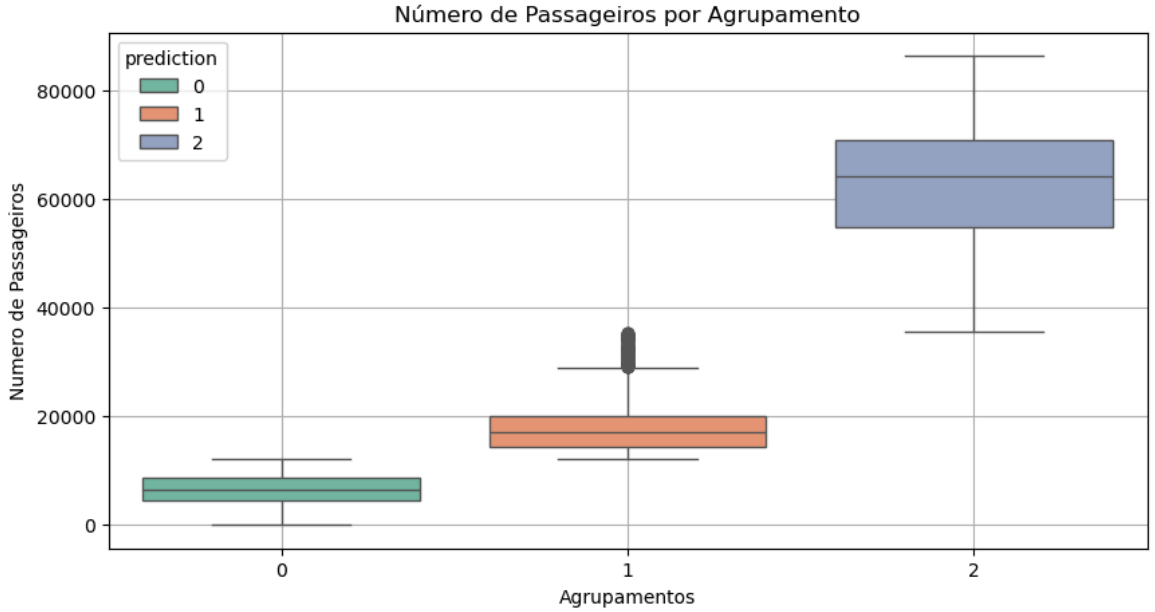
Como se vê no gráfico, o grupo 2 tem caraterísticas bem diferentes que os grupos 0 e 1, podendo dizer então que nesse grupo estão os registos mais atípicos, voos que foram com muitos passageiros, ou poucos, rotas ou companhias pouco frequentes, entre outras possibilidades.
Enquanto os grupos restantes partilham mais caraterísticas semelhantes, como rotas frequentes, número de passageiros, companhias, portas de embarque.

Foram feitos vários cálculos e análises para concluir que é importante analisar o impacto de fatores como sazonalidade, ocupação dos voos e distribuição das rotas aéreas. Pelo que a seguir, cada propriedade será explorada em maior detalhe para aprofundar a interpretação dos agrupamentos gerados pelo modelo.

Análise de resultados do modelo



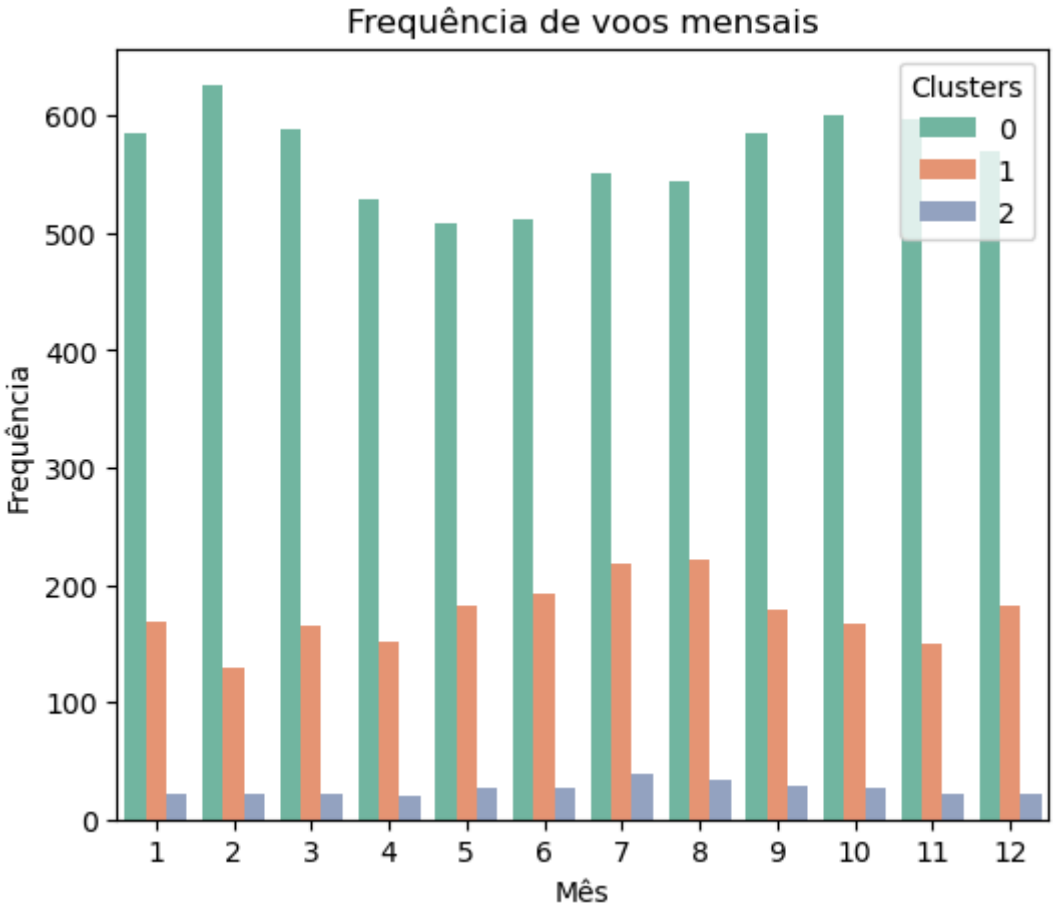
Esta primeira imagem nos permite observar a distribuição dos registos entre os grupos. Da para ver que o agrupamento 0 é o maior e por muita diferença. Isso pode sugerir que os agrupamentos estão separados por registos frequentes, menos frequentes e atípicos.



O modelo agrupou no **cluster nº0** os registos que partilham uma quantidade de passageiros relativamente constante, sem valores atípicos, e próximos à média. Considerando também a distribuição deste grupo, posso inferir que ele representa os voos mais frequentes, com uma ocupação estável ao longo do tempo.

No **cluster nº 1**, o algoritmo identificou e juntou valores atípicos, principalmente com uma dispersão significativa dos valores máximos. Isso indica que os voos que estão dentro desse grupo costumam estar lotados, possivelmente devido a períodos de alta demanda, como o verão e feriados.
Além disso, promoções regulares das companhias aéreas podem ter incentivado a compra de bilhetes, derivando em aviões com ocupação máxima.

Agora o **cluster nº 2** apresenta uma variabilidade de valores significativamente maior que os outros grupos. Somado a baixa quantidade de registos desse grupo, sugere que estão os registos atípicos do conjunto, como voos especiais, ligações não recorrentes ou promoções especiais de algumas companhias.

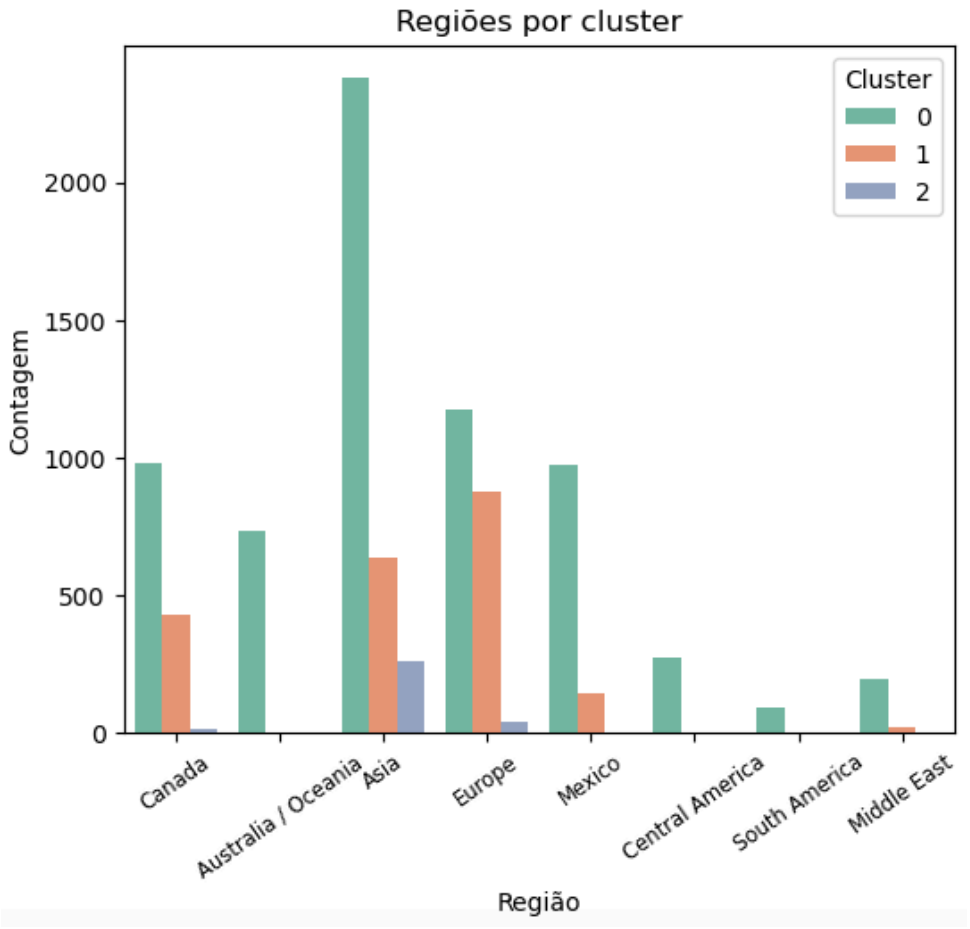


O gráfico representa a frequência mensal dos voos dentro de cada agrupamento.

No **cluster nº 0** o modelo agrupou os voos com maior frequência fora da temporada alta de turismo. Como esses registos representam os voos mais recorrentes, existe uma redução durante o verão, quando a demanda por outros tipos de voos aumenta.

O **cluster nº 1** concentra os voos cuja frequência cresce significativamente durante os períodos de maior turismo, como o verão e feriados. Esse padrão sugere uma forte influência sazonal nesses registos.

Já no **cluster nº 2** a distribuição dos voos se mantém relativamente constante, sugerindo que não tem influência sazonal, podendo estar relacionados a eventos particulares importantes ou datas comemorativas singulares.



Desenhando as distribuições das regiões nos agrupamentos, posso perceber como foi feita a segmentação pelo modelo.

No **cluster nº 0** predominam as rotas de Ásia, Europa, Canada e México. Isso sugere serem as rotas mais frequentes durante o ano.

A distribuição do **cluster nº1** tem ligações mais fortes com a Europa, Ásia e Canada. Tendo em consideração a influência da sazonalidade deste grupo, pode sugerir que as conexões durante o verão com estas regiões aumenta consideravelmente.

Por último, o modelo agrupou no **cluster nº2** rotas com Ásia, Europa e Canada, mas em pouca quantidade, indicando que dentro dessas rotas, existem voos atípicos graças a possíveis acontecimentos excecionais, promoções de companhias aéreas ou alguma data comemorativa específica.

As principais propriedades que influenciaram os agrupamentos foram:

1. **Quantidade de passageiros** – O algoritmo agrupou voos com ocupação próxima à média, outros com valores mais dispersos, mas ainda num intervalo razoável, e um terceiro grupo contendo voos com valores atípicos e maior variabilidade.

2. **Sazonalidade** – O modelo identificou picos de demanda no verão e no final do ano, diferenciando-os de voos cuja frequência se mantém alta fora dessas temporadas e de registos com distribuição homogénea ao longo do ano.

3. **Região** – Este fator teve um impacto significativo, destacando-se os voos para a Ásia, sendo os mais numerosos.

Com as análises cruzadas de diferentes variáveis e as observações obtidas, posso tirar conclusões relevantes dos agrupamentos e consequentemente, dos voos do aeroporto do São Francisco.

Conclusão dos Agrupamentos

A análise revelou padrões interessantes sobre as ligações internacionais do aeroporto. Identificamos que os voos mais frequentes, com uma média de passageiros relativamente constante, concentram-se principalmente nas rotas para a Ásia, Canadá e Europa. Além disso, essas operações ocorrem predominantemente fora da alta temporada turística.

Outro padrão relevante é o impacto significativo da sazonalidade na quantidade de passageiros por voo. Durante os períodos de maior demanda, observa-se um aumento no porte das aeronaves e na necessidade de mais pessoal para atender às operações do aeroporto.

Além disso, identificamos valores atípicos em algumas rotas frequentes ao longo de todo o ano, sem variações sazonais significativas. Esses picos de passageiros podem estar relacionados a estratégias comerciais das companhias aéreas, como promoções que incentivam a compra antecipada de bilhetes e resultam em voos mais lotados.

Fecho

Gostaria de fechar dizendo que conhecendo as caraterísticas importantes e alguns padrões interessantes, é possível aplicar este conhecimento em câmbios estruturais para continuar a crescer e receber cada ano, mais passageiros e criar incentivos para aumentar o uso de rotas menos frequentes.
Como também tomar outras decisões estratégicas consoante seja a área que quisermos melhorar.