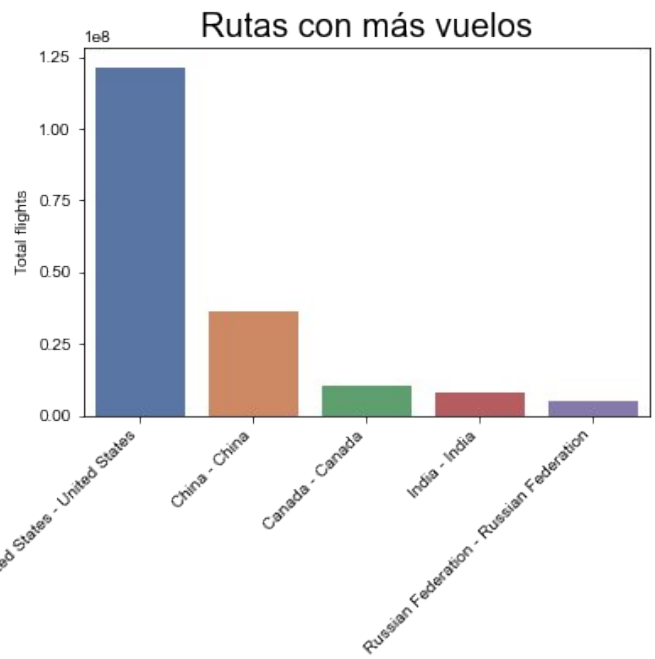
PRESENTACIÓN DEL ANÁLISIS DE DATOS

Después de analizar los datos e intentando averiguar los datos más importantes hemos destacado los siguiente:

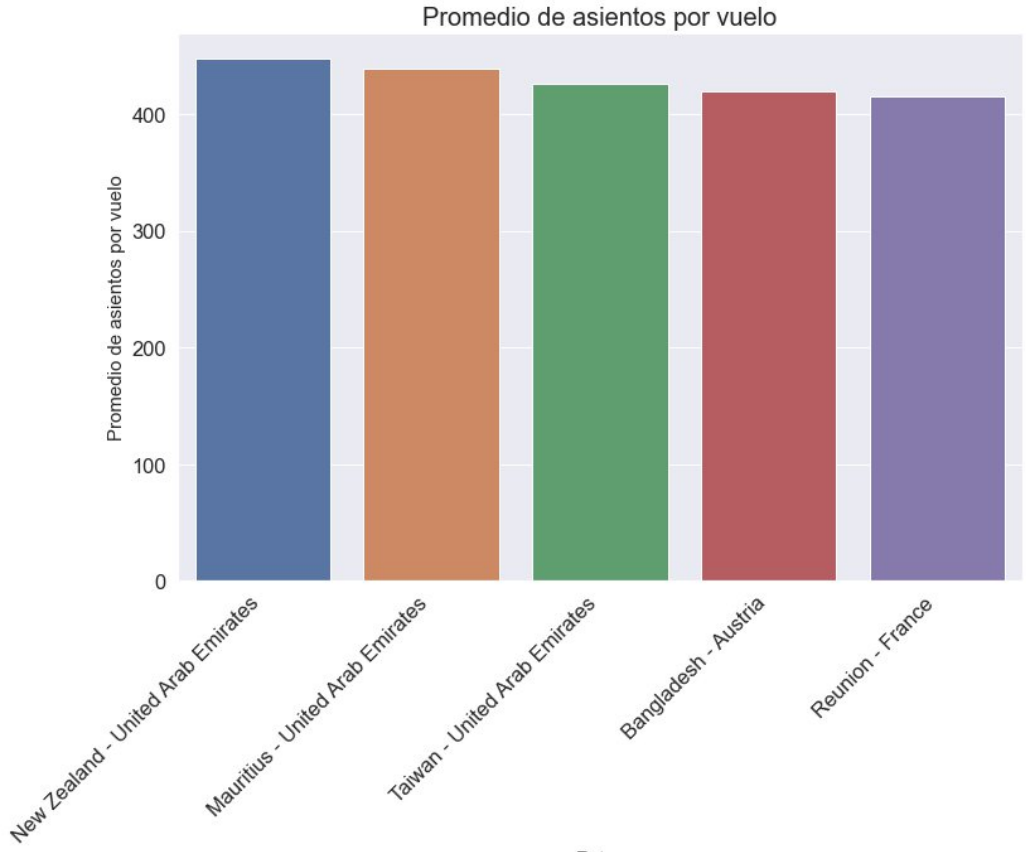
	Origin-Destination	Total flights	Total seats	Total ASKs
0	United States - United States	122110915	12266317361	16475458883996
1	China - China	37130654	5768921272	6717650384252
2	Canada - Canada	10914211	720474941	803921025232
3	India - India	8367460	1240511260	1109039277120
4	Russian Federation - Russian Federation	5508537	679629604	1209026112919
5	Spain - Spain	4764144	615578219	403124125708
6	United Kingdom - United Kingdom	4759346	431533744	180691189899
7	France - France	3884158	451429947	262076619199
8	Germany - Germany	3670973	472327348	202837152135
9	Italy - Italy	3632011	525509502	315303036798
10	United States - Canada	2603912	219983687	344944627632
11	Turkey - Turkey	2563617	437895627	263251943596
12	Canada - United States	2448285	208113780	326625073961
13	Thailand - Thailand	2199147	361405288	218356537784
14	South Africa - South Africa	1914423	215781347	194372265369



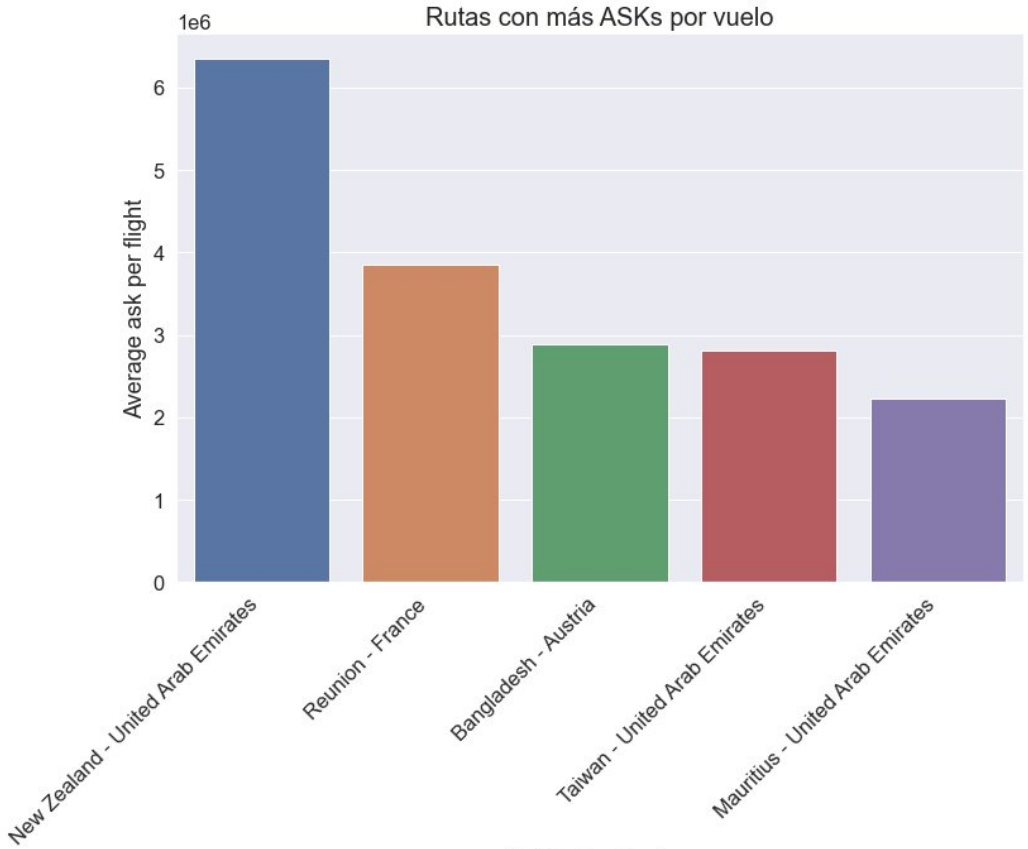
Podemos ver que las rutas internas en Estados Unidos es la ruta con más vuelos (multiplicando por más de tres la cantidad de vuelos internos en China).

Hasta la ruta número 11 no encontramos una ruta internacional por lo cual procedemos a comparar el número de rutas nacionales vs internacionales

Vemos cómo casi 3/4 de los vuelos son nacionales y 1/4 de los vuelos son internacionales.



En éste caso vemos cómo los vuelos con más promedia de asientos son por consiguiente los vuelos más largos, sin diferencias destacables entro los 5 primeros.



Proceso de machine learning:

Primero, se realizó una limpieza de los datos, eliminando valores faltantes, duplicados y corrigiendo valores incorrectos. Luego, se realizó un preprocesamiento, que incluyó la codificación de variables categóricas y la normalización de variables numéricas.

Posteriormente, se verificó que la muestra extraída de la base de datos era representativa. Luego, se probaron diferentes modelos de aprendizaje automático, árboles de decisión y random forest. Después de comparar los resultados de los diferentes modelos, se escogió RandomForestClassifier por tener el mejor desempeño.

Por último, se utilizó GridSearchcv para mejorar los parámetros del modelo, y se obtuvo un resultado final de 0.81. Este resultado representa la precisión del modelo, es decir, la proporción de predicciones correctas sobre el total de predicciones realizadas. Un resultado de 0.81 es considerado un buen desempeño, y sugiere que el modelo es capaz de predecir con una precisión alta el valor de la variable objetivo.

En resumen, se ha realizado un proceso exhaustivo para limpiar, preprocesar y modelar la base de datos, utilizando técnicas estadísticas y de aprendizaje automático. El resultado final es positivo, y sugiere que el modelo es útil para realizar predicciones precisas sobre la variable objetivo.