# Análisis de Satisfacción de Pasajeros

Agustín D'Alessandro

# Índice

**1. Metadata** **2**

**2. Hipótesis** **4**
2.1. Hipótesis 1: Clase - Satisfacción . . . . . 4
2.2. Hipótesis 2: Distancia - Satisfacción . . . . . 4
2.3. Hipótesis 3: Edad - Satisfacción . . . . . 6
2.4. Hipótesis 4: Puntajes de Satisfacción - Satisfacción . . . . . 8

**3. Insights** **9**

**Resumen**

En el siguiente informe se analiza una encuesta realizada a los clientes de una aerolínea con el fin de conocer el grado de satisfacción de los mismos. El dataset correspondiente a la encuesta cuenta originalmente con 25 columnas, de las cuales una corresponde al grado de satisfacción de los usuarios y 22 de las restantes columnas refieren a los diversos aspectos tenidos en cuenta durante la encuesta. Las 22 columnas abarcan un amplio abanico de factores que influyen en el vuelo: desde la edad del pasajero hasta el espacio de los asientos o la limpieza del avión, por ejemplo. Mediante la implementación de algoritmos de *Machine Learnign* se busca poder predecir si un cliente estará o no satisfecho en base a sus respuestas en la encuesta. Esto permitirá a la aerolínea saber qué aspectos mejorar, que promociones ofrecer y con qué usuarios tomar mayores recaudos para lograr aumentar el número de clientes satisfechos, en busca de mejorar las ganancias de la empresa.

# 1. Metadata

El dataset original consta de 25 columnas con 103904 registros. La variable objetivo a analizar es *'satisfaction'*, variable categórica cuyos valores son *'neutral or dissatisfied'* o *satisfied*. De las demás columnas la primera es eliminada por tratarse de un duplicado del índice de la tabla y se agrega una nueva columna calculada como el promedio de un subconjunto de las columnas del dataset. A continuación se enumeran las variables:

1. *Unnamed: 0*:
   Columna de tipo numérico que es eliminada por ser un duplicado del índice de la tabla.

2. *id*:
   Columna de tipo numérico que registra identifica a cada usuario de manera única.

3. *Gender*:
   Columna de tipo *object*, categórica, cuyos valores son *'male'* y *'female'*.

4. *Customer Type*:
   Columna de tipo *object*, categórica, cuyos valores son *'Loyal Customer'* y *'disloyal Customer'*.

5. *Age*:
   Columna de tipo numérica con los rangos de edades de los usuarios, que varían entre 7 y 85 años.

6. *Type of Travel*:
   Columna de tipo *object*, categórica, cuyos valores son *'Business Travel'* y *'Personal Travel'*.

7. *Class*:
   Columna de tipo *object*, categórica, cuyos valores son *'Eco'*, *'Eco Plus'* y *'Business Class'*.

8. *Flight Distance*:
   Columna de tipo numérica que registra la distancia de los vuelos.

9. *Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, Cleanliness*:
   Columnas numéricas cuyos valores varían entre 0 y 5, por lo que, a pesar del tipo de dalto, resultan ser variables del tipo categóricas ordinales. Todas indican el grado de satisfacción respecto a cada uno de los aspectos tenidos en cuenta: Servicio de WiFi, Espacio, Limpieza, etc.

10. *Mean Satisfaction*:
    Columna de tipo numérico calculada y agregada al dataset. Recopila el promedio de las respuestas de cada usuario acerca de las variables ordinales mencionadas previamente.

11. *Departure Delay in Minutes*:
    Columna de tipo numérico que mide el tiempo de demora en la salida de cada vuelo.

12. *Arrival Delay in Minutes*:

Columna de tipo numérico que mide el tiempo de demora en la llegada de cada vuelo. Cuenta con 310 datos nulos que fueron reemplazados con la media de la columna.

# 2. Hipótesis

## 2.1. Hipótesis 1: Clase - Satisfacción

Se supone que la clase en la que viajan los usuarios es un aspecto influyente en la satisfacción final de los mismos.

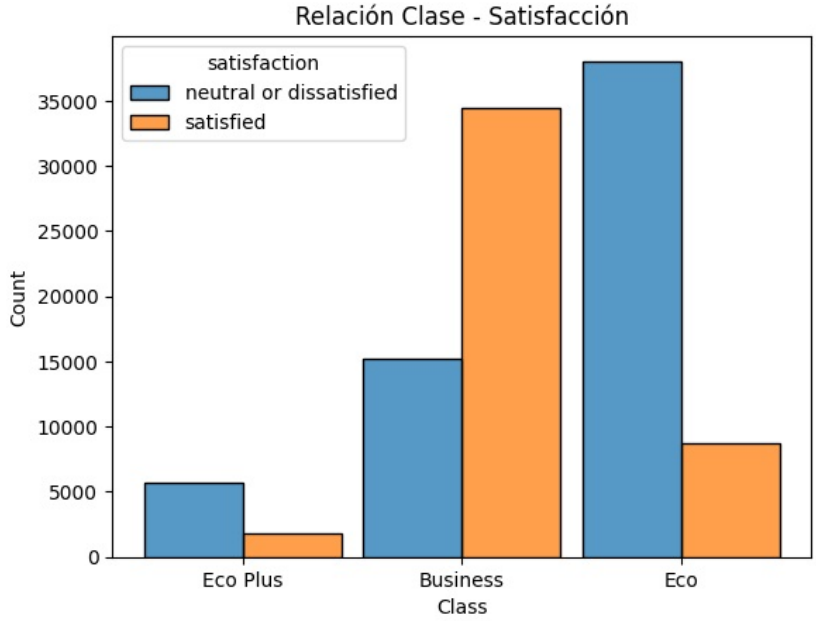


Fig. 1. Relación Clase - Satisfacción.

El gráfico de barras indica que la clase *Business* es la que concentra la mayor cantidad de opiniones positivas, mientras que la clase *Eco* muestra un grado de satisfacción muy bajo.

## 2.2. Hipótesis 2: Distancia - Satisfacción

Se supone originalmente que a mayor distancia la satisfacción de los usuarios será menor.

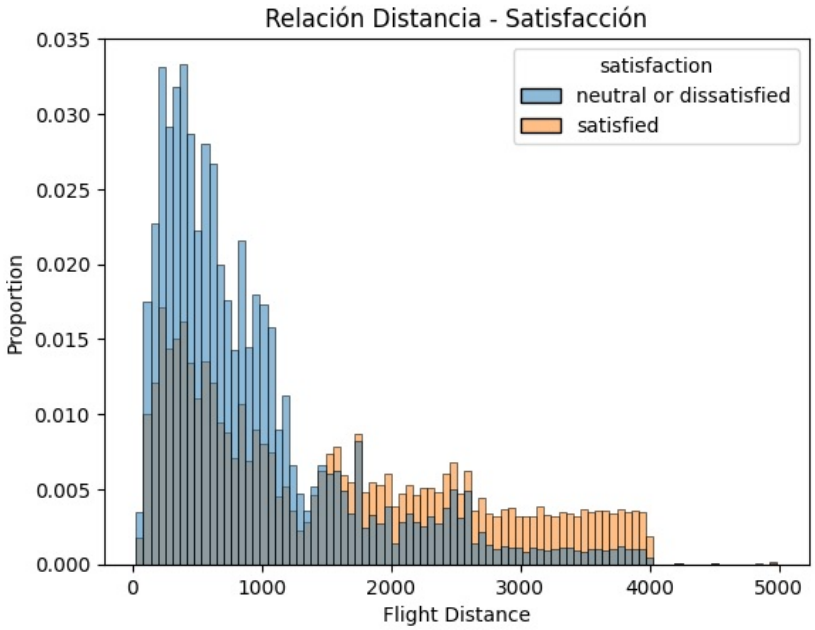


Fig. 2. Relación Distancia - Satisfacción.

El histograma refuta la hipótesis propuesta. Si bien es cierto que la cantidad de vuelos disminuye al aumentar la distancia viajada, las proporciones en la satisfacción pasan a invertirse: a menor distancia, mayor insatisfacción y a mayor distancia, mayor satisfacción. Eso lleva al planteamiento de una nueva hipótesis: **la distancia del vuelo conlleva una elección de la clase a volar, lo que afecta a la satisfacción (a mayor distancia, mejor clase y por lo tanto más satisfacción)**.

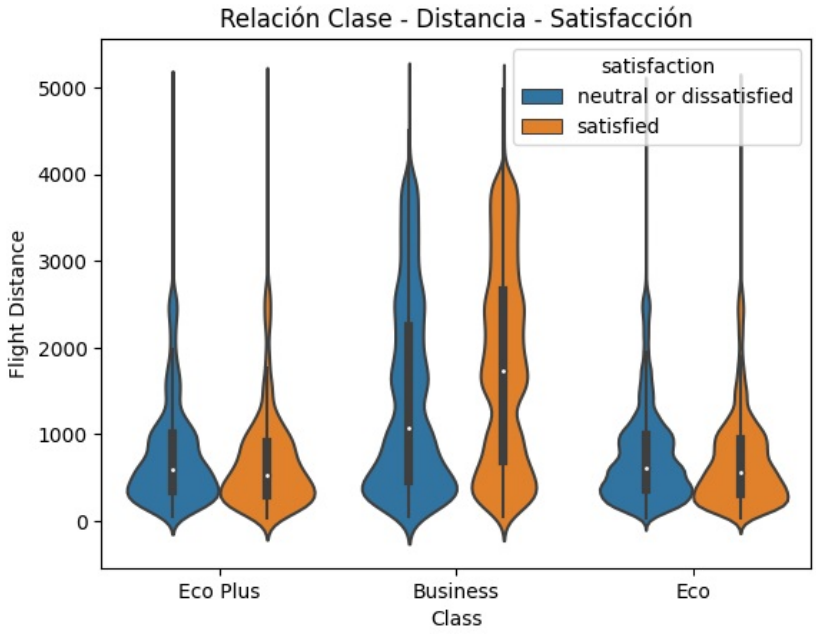


Fig. 3. Relación Distancia - Clase - Satisfacción.

Se puede apreciar que el violín central, correspondiente a la clase *'Business Class'* es la que concentra la mayoría de los vuelos de largas distancias y presenta mayor grosor que el violín que indica la cantidad de gente insatisfecha en esa clase.

## 2.3. Hipótesis 3: Edad - Satisfacción

Se hace la suposición que la edad es un aspecto influyente en la satisfacción de los clientes.

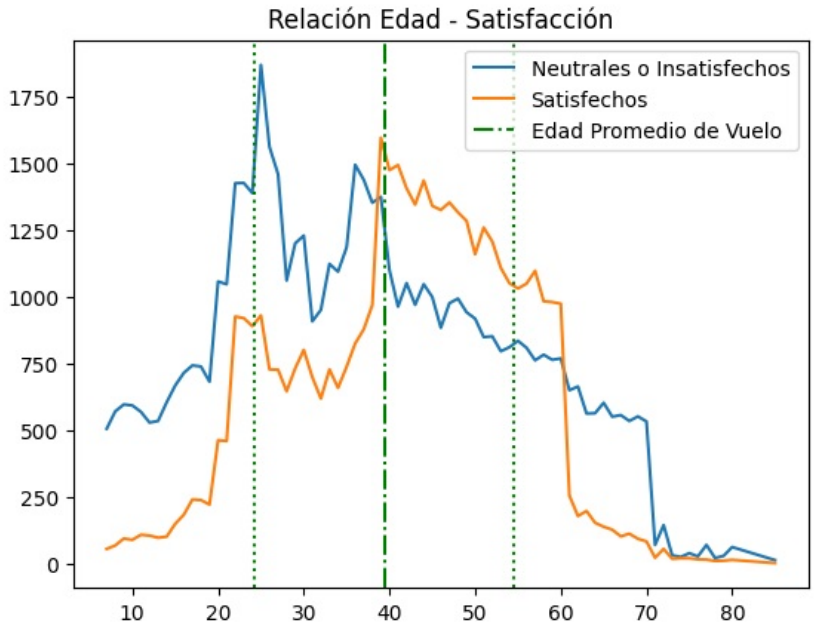


Fig. 4. Relación Edad - Satisfacción.

Se puede apreciar que la cantidad de usuarios insatisfechos tiene un pico en la franja de edades entre 20 y 30 años, donde la cantidad de personas insatisfechas supera a las satisfechas. La cantidad de pasajeros satisfechos alcanza su pico a los 40 años, donde supera a la cantidad de gente insatisfecha. Para explicar este fenómeno se plantea la hipótesis: **las personas de mayor edad pueden acceder a vuelos en clases superiores que la gente más joven. Esto puede deberse a la diferencia económica entre los pasajeros en edad laboral y aquellos que todavía, por ser jóvenes, no han comenzado a trabajar o por ser incluso menores de edad que viajan junto a sus padres.**

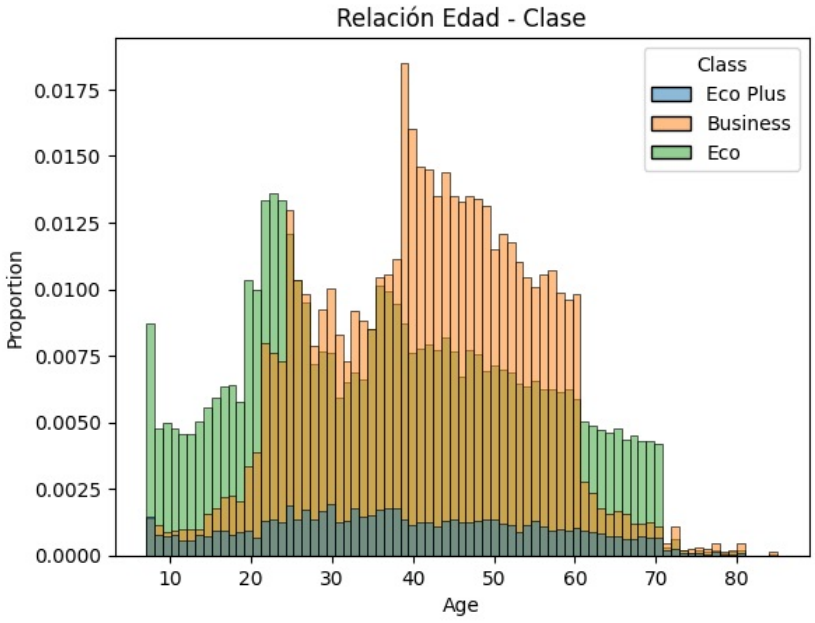


Fig. 5. Relación Edad - Clase.

En el gráfico queda plasmado que en torno a los 20 y 30 años se concentra la mayor cantidad de clientes de la clase *Eco*, que es la que tiene mayor grado de insatisfacción, mientras que entre los 40 y 50 la mayoría de los usuarios elige viajar en clase *Business*, que es la clase con mayor cantidad de pasajeros satisfechos.

## 2.4. Hipótesis 4: Puntajes de Satisfacción - Satisfacción

Se estima que los usuarios satisfechos han dado mayores puntajes a los servicios del vuelo y obtenido una mayor satisfacción promedio que los pasajeros que no quedaron satisfechos.

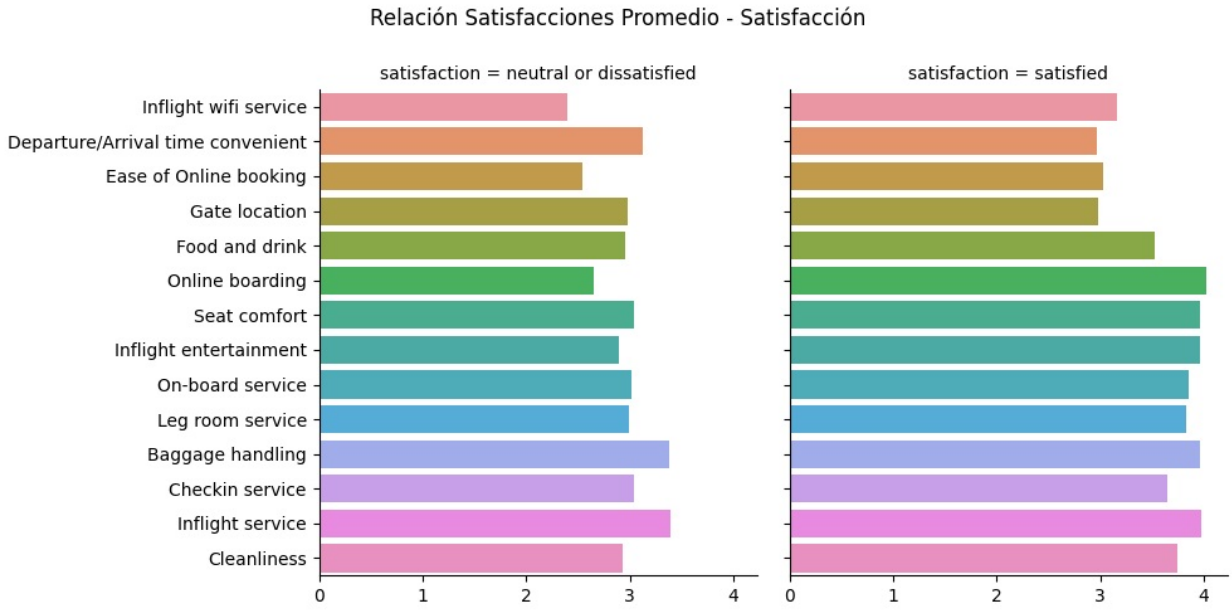


Fig. 6. Relación Puntajes - Satisfacción.

Los graficos de barras muestran que en los usuarios insatisfechos, el puntaje promedio de cada servicio del avión es más bajo que cada puntaje promedio entre la gente que sí ha quedado satisfecha.

# 3. Insights

A partir de una primera exploración de los datos y de revisión de diferentes gráficos se puede concluir que hay ciertas variables que presentan una correlación con la variable objetivo a estudiar, *'satisfaction'*. Como era de esperarse, la clase en la que las personas viajan influye en la satisfacción. De las tres opciones que presenta la aerolínea, la clase más cara (*Business Class*) es la que presenta mayor cantidad de pasajeros satisfechos. No sólo es la clase con más cantidad de pasajeros satisfechos, sino que es la única clase en la que hay más usuarios satisfechos que insatisfechos. La mayoría de los viajeros de clases económicas, en cambio, se muestran insatisfechos con sus vuelos.

Distintas variables se correlacionan con la clase que eligen las personas. Esta correlación lleva, por transitividad, a que las variables como *'Flight Distance'* y *'Age'* afecten a la satisfacción de los usuarios. En el caso de las dos variables mencionadas, la correlación entre la distancia y la satisfacción es directa, ya que, para afrontar un viaje largo, los usuarios eligen viajar en una clase más cómoda y, por ende, resultan más satisfechos. La correlación entre la edad y la satisfacción, principalmente, también directa. Es de suponerse que las personas con mayor edad pertenecen a un sector de la población en plena edad laboral. Más específicamente, a los 40 años, donde se registran los picos de satisfacción, la gente no sólo se encuentra en edad laboral, sino que cuenta con ya varios años de experiencia laboral, que pueden implicar mejores cargos, con mejores sueldos e incluso aumentos debido a la antigüedad laboral. Esto le permite a los usuarios que se encuentran en la franja etaria entre los 40 y 50 años acceder a vuelos en mejores clases. En contraste, aquellas personas menores a 30 años o mayores de 60 viajan principalmente en clases económicas y se muestran menos satisfechas. El viajar en clases más económicas podría explicarse debido a ser usuarios en edad infantil o jóvenes adultos sin independencia económica o con cargos laborales con poca experiencia y sueldos bajos (para la franja de menor edad) o por los pocos ingresos que posee la mayoría de las personas jubiladas.

Por último, los puntajes obtenidos en los servicios del vuelo son un buen indicador de la satisfacción de los usuarios. Aquellos que han estado satisfechos con sus vuelos han plasmado su satisfacción mediante buenos puntajes, mientras que las personas insatisfechas han dado puntajes más bajos en los servicios.

En vista del análisis realizado, se estima que al querer armar un modelo de machine learning para predecir la satisfacción de los usuarios, las columnas que se han contemplado en las hipótesis serán relevantes.