Asignatura: Aprendizaje Automatico

David Morán GorgojoMiguel Sánchez Rodríguez

Pablo Ruíz Morán

ANALISIS EXPLORATORIO DE DATOS

EDA

Contents

[1. Introducción 2](#_Toc183126042)

[2. Información General del Dataset 2](#_Toc183126043)

[3. Distribuciones de Variables 3](#_Toc183126044)

[3.1 Distribución de Edad 3](#_Toc183126045)

[3.2 Distribución de la Distancia del Vuelo 3](#_Toc183126046)

[3.3 Distribución de la Satisfacción del Pasajero 4](#_Toc183126047)

[3.3.1 Distribución por genero 4](#_Toc183126048)

[3.4 Distribución de los Retrasos 5](#_Toc183126049)

[4. Visualización de datos 5](#_Toc183126050)

[4.1 Matriz de Correlación 5](#_Toc183126051)

[4.2 Gráficos 7](#_Toc183126052)

[4.2.1 Distribución de la distancia de vuelo: 7](#_Toc183126053)

[4.2.1 Distribución de la satisfacción respecto de ratings: 8](#_Toc183126054)

[5. Conclusiones 10](#_Toc183126055)

# 1. Introducción

En este documento se presenta el análisis exploratorio de datos (EDA) realizado para el proyecto sobre satisfacción de usuarios en una aerolínea. Este análisis tiene como objetivo comprender las características principales del dataset, identificar patrones, detectar valores atípicos y preparar los datos para su posterior modelado.

# 2. Información General del Dataset

* **Dimensiones del Dataset**:  
  El dataset contiene **103,904 filas(datos)** y **24 columnas(características)**, que representan características tanto categóricas como numéricas.
* **Columnas y Tipos de Datos**:  
  Incluye variables categóricas como *Gender*, *Customer Type* y *satisfaction*, además de variables numéricas como *Age*, *Flight Distance* y los servicios a bordo.
* **Relación entre distancia de vuelo y satisfacción:**

Los pasajeros satisfechos tienden a estar más asociados a vuelos de mayor distancia, mientras que los insatisfechos se concentran en vuelos más cortos.

* **Datos Nulos**:  
  La columna *Arrival Delay in Minutes* presenta algunos valores nulos que deben ser tratados en futuras etapas.
* **Distribución de valores en columnas categóricas:** Las características categóricas, como *Gender*, *Customer Type*, *Type of Travel*, *Class* y *satisfaction*, fueron analizadas para identificar la proporción de cada categoría. Por ejemplo:
  + La mayoría de los clientes pertenecen a la clase económica (*Eco*), seguidos por la clase *Business* y *Eco Plus*.
* **Relación entre características categóricas y satisfacción:**
  + Los gráficos de barras mostraron que los clientes en la clase *Business* tienen mayores niveles de satisfacción, mientras que los de la clase *Eco* presentaron una mayor proporción de insatisfacción.
  + Se identificaron patrones interesantes en columnas como *Customer Type*, donde los clientes leales mostraron consistentemente mayores niveles de satisfacción.

# 3. Distribuciones de Variables

## 3.1 Distribución de Edad

La mayoría de los pasajeros se encuentran en el rango de **25 a 50 años**, con un pico en la mediana de aproximadamente **40 años**. La distribución es aproximadamente simétrica. La edad de los pasajeros varía entre **7 años** y **85 años**, con una media de **39.41 años**.

A graph with a line

Description automatically generated

## 3.2 Distribución de la Distancia del Vuelo

La distancia de vuelo tiene una distribución sesgada hacia distancias cortas, con un número significativo de vuelos de menos de **1,000 km**. La distancia de vuelo promedio es de **1,188 km**, con un máximo de **4,983 km**.

A graph of a distancing

Description automatically generated

## 3.3 Distribución de la Satisfacción del Pasajero

La mayoría de los pasajeros se clasifican como *neutral o insatisfechos*, lo que podría indicar áreas de mejora para la aerolínea en términos de servicio.

A graph with blue squares

Description automatically generated

### 3.3.1 Distribución por genero

A red and blue circles with black text

Description automatically generated

* Podemos concluir, gracias a la gráfica, que la variación de satisfacción entre géneros es prácticamente nula siendo tan solo 1,2% de diferencia en el que las mujeres están mas insatisfechas. Por lo tanto apenas afecta el genero al resultado final.

## 3.4 Distribución de los Retrasos

Los retrasos en llegada y salida presentan distribuciones altamente sesgadas, con valores máximos de **1,280 minutos** y **1,305 minutos**, respectivamente.

A graph with blue and white bars

Description automatically generated

* Así podemos ver que la mayoría de retrasos son de como mucho 10mins, siendo los retrasos de mas de una hora muy reducidos. Siguiendo un descenso semejante a una exponencial inversa

# 4. Visualización de datos

Se han hecho diversas comparaciones de los clientes según la clase y otros factores categóricos. Para ello se utilizaron diversas técnicas de visualización para mejorar la comprensión de los datos:

## 4.1 Matriz de Correlación

Se observa una fuerte correlación positiva entre:

* *Inflight Entertainment* y *Overall Satisfaction*.
* *Online Boarding* y *Ease of Booking*.
* Departure Delay y Arrival Delay

La fuerte correlación positiva entre estas características, nos puede sugerir, la posibilidad de prescindir de una de ellas, a la hora de realizar el modelo, ya que pueden considerarse redundantes.

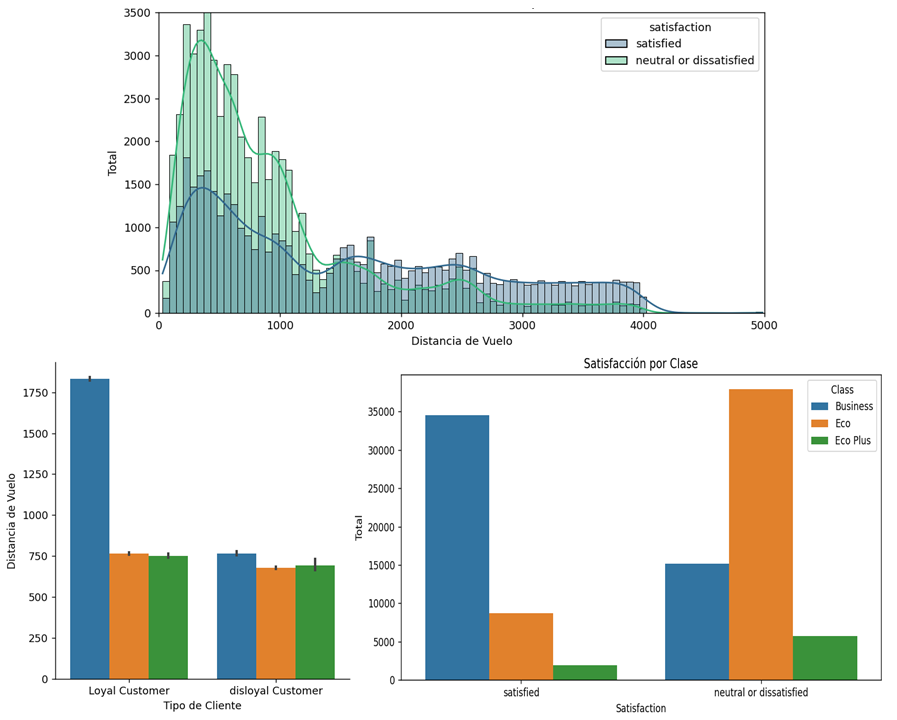
Por otro lado, las variables de retrasos (*Departure Delay* y *Arrival Delay*) muestran una correlación negativa con la satisfacción del pasajero, lo que refuerza la importancia de minimizar retrasos para mejorar la experiencia del cliente.

A graph with numbers and letters

Description automatically generated with medium confidence

## 4.2 Gráficos

## 4.2.1 Distribución de la distancia de vuelo:



Análisis de las gráficas:

**1. Distancia de Vuelo y Satisfacción (Gráfica Superior):**

* + Los pasajeros "neutral o insatisfechos" predominan en distancias cortas (<1000 km).
  + Esto sugiere que la distancia del vuelo podría influir en la satisfacción del cliente, siendo los vuelos más largos asociados con mayor satisfacción.

**2. Tipo de Cliente vs. Distancia Promedio (Gráfica Inferior Izquierda):**

* + Los clientes "**Loyal Customer**" tienen distancias de vuelo significativamente mayores, especialmente en la clase Business.
  + Los **“disloyal Customer**" presentan distancias más cortas en promedio, sin importar la clase.
  + Los pasajeros frecuentes (Loyal Customer) en la clase Business podrían estar más satisfechos debido a mejores servicios ofrecidos en vuelos más largos.

**3. Satisfacción por Clase (Gráfica Inferior Derecha):**

* + En la clase **Business**, la mayoría de los pasajeros están "satisfechos", lo que sugiere que esta clase proporciona una experiencia mejor. Lo que se relación a que estos suelen hacer vuelos mas largos con esas mejores prestaciones afectando así al resultado de las otras dos graficas
  + Las clases **Eco** y **Eco Plus** tienen un mayor número de pasajeros "neutral o insatisfechos", con Eco siendo la que acumula la mayor insatisfacción. Esto implica que las clases más económicas están asociadas con niveles más bajos de satisfacción, probablemente por diferencias en comodidad, servicio, u otros factores.

## 4.2.1 Distribución de la satisfacción respecto de ratings:

A group of blue and green bars

Description automatically generated

En esta tabla recogemos la relación de todas las características que los pasajeros han evaluado del 0 al 5, siendo 0 que no están contentos con esa característica y 5 que si, relacionándolo con su satisfacción. Por lo tanto el resultado esperado es que si un pasajero vota 5 es que le ha gustado, por lo tanto es esperable que este satisfecho. Si una característica no cumple con ello, puede ser debido a que no tiene suficiente peso para influir en la decisión de si esta satisfecho o no. Otra opción es que sean datos anómalos o que no tenga relación directa.

#### 4.2.1.1 Ejemplos de los casos

* **Caso Favorable**

A graph of different colored bars

Description automatically generated with medium confidence

En esta tabla podemos ver el caso que explicábamos como esperado. En el observamos como a mayor nota mayor es el numero de satisfechos y su vez en notas media el número de pasajeros neutrales es mayor y según sube la nota este número desciende.

* **Caso Anómalo**

A graph of a bar graph

Description automatically generated with medium confidenceEn este es al revés y vemos como no se nota una correlación. El numero de pasajeros satisfecho parece mantenerse más o menos constante a pesar de la nota y sin embargo el numero de neutrales o insatisfecho aumenta con la nota. Esto podría ser debido a que este no tiene relación o no es una característica que importe a los pasajeros.

* **Caso Medio**

A graph of a bar chart

Description automatically generated with medium confidence

Este seria un caso medio en el que si bien aparecen las características del caso favorable podemos ver que no tienen tanta fuerza y con notas altas la diferencia en la satisfacción no parece ser demasiada. Por lo tanto es una característica que se tiene en cuenta, pero tiene menos fuerza.

# 5. Conclusiones

El análisis exploratorio ha revelado información valiosa sobre el dataset:

* La satisfacción del pasajero está influenciada por factores como el entretenimiento a bordo, el embarque en línea, los retrasos en los vuelos y la distancia de vuelo.
* Existen valores nulos y distribuciones sesgadas en algunas variables que deberán abordarse en las etapas de limpieza y preparación de datos.
* El siguiente paso será realizar una limpieza de datos exhaustiva, ingeniería de características y dividir el dataset para el modelado predictivo.
* Observamos algunos valores nules y distribuciones sesgadas, por lo que un posible preprocesado, sería usar la mediana o la media, o incluso predecir el valor, usando alguna técnica de regresión o clasificación.
* Respecto a las variables categóricas, sería interesante, discretizar, y en aquellas con gran variedad de valores, reducirlas a intervalos.
* Algunas variables sesgadas, como los valores que se toman en la columna **Flight Distance**, podrían escalarse, haciendo uso del escalado minMax o el Standard.