Nombre completo

Federico Camacho Cagal

- **Materia**

Introducción a la Ciencia de Datos

- **Nombre del profesor**

Jaime Alejandro Romero Sierra

- **Fecha de entrega**

20 de octubre de 2025

- Link al repositorio de GitHub

https://github.com/fedcamc/ProyectoDS

Descripción inicial de la base de datos

- **Fuente o contexto de la base de datos.**

En la actualidad, la industria del transporte aéreo se ha consolidado como uno de los pilares más importantes del comercio, el turismo y la movilidad global. Cada día, millones de personas utilizan los servicios de distintas aerolíneas para trasladarse entre ciudades, países y continentes, lo que ha generado una competencia cada vez más intensa entre las compañías. En este escenario, la satisfacción del pasajero se ha convertido en un elemento estratégico clave para garantizar la fidelización de los clientes, fortalecer la reputación de la marca y asegurar la sostenibilidad del negocio a largo plazo.

El aumento de la oferta de vuelos, junto con la aparición de aerolíneas de bajo costo y la expansión de los servicios digitales, ha transformado la forma en que los pasajeros perciben su experiencia de viaje. Ya no basta con ofrecer un transporte eficiente; los usuarios esperan comodidad, puntualidad, atención personalizada y servicios tecnológicos que faciliten todo el proceso, desde la reserva hasta el aterrizaje. Esto ha impulsado a las empresas del sector a recopilar y analizar datos sobre la percepción de sus clientes, con el fin de comprender mejor sus necesidades, expectativas y niveles de satisfacción.

La medición de la satisfacción de los pasajeros permite a las aerolíneas identificar los factores que más influyen en la experiencia de vuelo, como la puntualidad, el trato del personal, la comodidad de los asientos, la calidad de

los alimentos, el entretenimiento a bordo o la eficiencia en el manejo del equipaje. A partir de esta información, las compañías pueden diseñar estrategias más efectivas para mejorar la calidad del servicio, optimizar los procesos operativos y ofrecer una experiencia más coherente con las expectativas del mercado actual.

En un contexto globalizado, la satisfacción del cliente también tiene un impacto directo en la competitividad y sostenibilidad de las aerolíneas. Las evaluaciones y comentarios de los pasajeros se difunden con rapidez a través de plataformas digitales, influyendo en la decisión de compra de otros usuarios y afectando la imagen pública de las empresas. Por ello, comprender y gestionar adecuadamente la experiencia del cliente no solo contribuye a incrementar la lealtad, sino que también representa una ventaja competitiva en un entorno cada vez más exigente y transparente.

El análisis de datos sobre la satisfacción de los pasajeros se ha convertido, además, en una herramienta fundamental dentro de la ciencia de datos aplicada al transporte aéreo. Mediante técnicas estadísticas y modelos predictivos, es posible detectar patrones de comportamiento, segmentar perfiles de usuarios y anticipar posibles áreas de mejora. Esto permite transformar grandes volúmenes de información en conocimiento útil para la toma de decisiones estratégicas, orientadas a elevar la calidad del servicio y, en última instancia, la experiencia del viajero.

En suma, estudiar la satisfacción de los pasajeros en vuelos comerciales no solo aporta valor desde la perspectiva del cliente, sino que también representa una oportunidad para la innovación, la eficiencia y la sostenibilidad en la industria

aérea. A través del análisis de datos, las aerolíneas pueden comprender mejor a sus usuarios, adaptarse a las nuevas demandas del mercado y consolidar su posición en un sector donde la experiencia del pasajero es, cada vez más, el centro de toda estrategia empresarial.

• Descripción general del contenido

La base de datos analizada contiene información detallada sobre la **experiencia y satisfacción de pasajeros que viajan en una aerolínea comercial**. En ella se combinan datos demográficos, características del viaje, evaluaciones de distintos aspectos del servicio y variables relacionadas con el rendimiento operativo de los vuelos. En conjunto, esta información permite realizar un análisis integral del comportamiento del cliente y de los factores que determinan su nivel de satisfacción general.

En primer lugar, el conjunto de datos incluye una serie de **variables demográficas y de identificación del pasajero**, como el *género*, la *edad* y el *tipo de cliente*. Estas características son relevantes porque permiten segmentar a los usuarios según su perfil y analizar si existen diferencias en la percepción del servicio entre grupos específicos, como por ejemplo entre viajeros frecuentes y pasajeros ocasionales. También se registra un número de identificación individual para cada pasajero, lo cual facilita la organización y trazabilidad de los registros sin comprometer la privacidad de los participantes.

El segundo grupo de variables se relaciona con las **características del viaje**. Aquí se incluyen campos como el *tipo de viaje* (personal o de negocios), la *clase*

en la que viaja el pasajero (económica, ejecutiva o sin clase) y la *distancia de vuelo*, expresada en millas o kilómetros. Estas variables permiten contextualizar la experiencia del cliente, ya que factores como la duración del vuelo o el propósito del viaje influyen directamente en las expectativas y niveles de exigencia del pasajero.

Posteriormente, el conjunto de datos incorpora una amplia gama de **indicadores de servicio y experiencia del cliente**, los cuales representan la evaluación que los pasajeros otorgan a distintos aspectos del servicio aéreo. Entre ellos se encuentran la *facilidad de reservación en línea*, el *servicio de wifi*, la *comodidad del asiento*, la *limpieza de la aeronave*, el *servicio del personal de vuelo*, el *manejo del equipaje*, el *entretenimiento a bordo* y la *calidad de los alimentos y bebidas*. Cada uno de estos elementos está calificado con valores numéricos que reflejan la percepción del usuario sobre la calidad del servicio recibido. Estas calificaciones son fundamentales para determinar qué aspectos influyen de forma más significativa en la satisfacción general.

Además, la base incluye variables operativas como el *retraso de salida* y el *retraso de llegada*, expresadas en minutos, que reflejan la eficiencia logística y puntualidad de la aerolínea. Estas métricas son especialmente relevantes, ya que los retrasos son una de las causas más comunes de insatisfacción entre los pasajeros.

Finalmente, la variable **“Satisfacción”** resume la opinión general del pasajero respecto a su experiencia total de vuelo, clasificándola en categorías como *satisfecho*, *neutral* o *insatisfecho*. Esta columna representa el punto central del

análisis, pues permite evaluar cómo las diferentes características del viaje y los factores de servicio se relacionan con la satisfacción global.

En conjunto, la base de datos ofrece una visión completa del proceso de viaje desde la perspectiva del cliente. Su estructura permite aplicar técnicas de análisis estadístico y modelos de aprendizaje automático para **identificar patrones, correlaciones y predictores clave de satisfacción**, con el fin de proponer estrategias de mejora y optimización en la experiencia del pasajero.

- **Significado de cada columna.**

Nombre de la columna	Descripción / Significado
id	Identificador único asignado a cada pasajero o registro dentro del conjunto de datos.
Género	Indica el sexo del pasajero, ya sea masculino o femenino.
Tipo de Cliente	Clasificación del pasajero según su relación con la aerolínea: cliente leal, desleal o sin tipo específico.
Edad	Edad del pasajero al momento del vuelo.

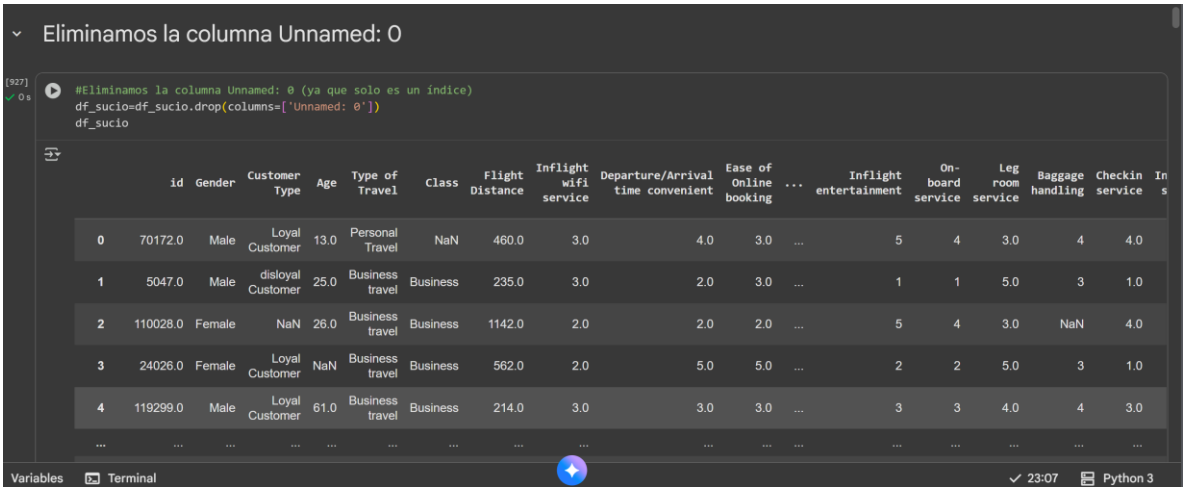
Tipo de Viaje	Motivo principal del viaje: puede ser de negocios o personal.
Clase	Tipo de asiento o nivel de servicio adquirido: económica, ejecutiva o sin clase.
Distancia de Vuelo	Longitud del trayecto en kilómetros entre el aeropuerto de origen y el de destino.
Servicio de Wifi	Evaluación del pasajero sobre la calidad del servicio de internet a bordo.
Tiempo de Llegada/Salida Conveniente	Nivel de satisfacción respecto a la puntualidad y conveniencia de los horarios de vuelo.
Facilidad de Reservación en Línea	Opinión del pasajero sobre la simplicidad y eficiencia del sistema de reserva en línea.
Ubicación de Puerta	Evaluación de la conveniencia o accesibilidad de la puerta de embarque asignada.

Comida y Bebida	Calificación otorgada a la calidad y variedad del servicio de alimentos y bebidas.
Abordaje en Línea	Satisfacción respecto al proceso de abordaje digital o electrónico.
Comodidad de Asiento	Opinión sobre el confort físico del asiento durante el vuelo.
Entretenimiento en Vuelo	Evaluación de la calidad y disponibilidad de opciones de entretenimiento a bordo.
Servicio en Mesa	Percepción sobre la atención y servicio brindado por el personal de cabina durante el vuelo.
Espacio del Asiento	Opinión acerca del espacio disponible para las piernas y comodidad general del área del pasajero.
Servicio de Equipaje	Nivel de satisfacción respecto al manejo, entrega y cuidado del equipaje.

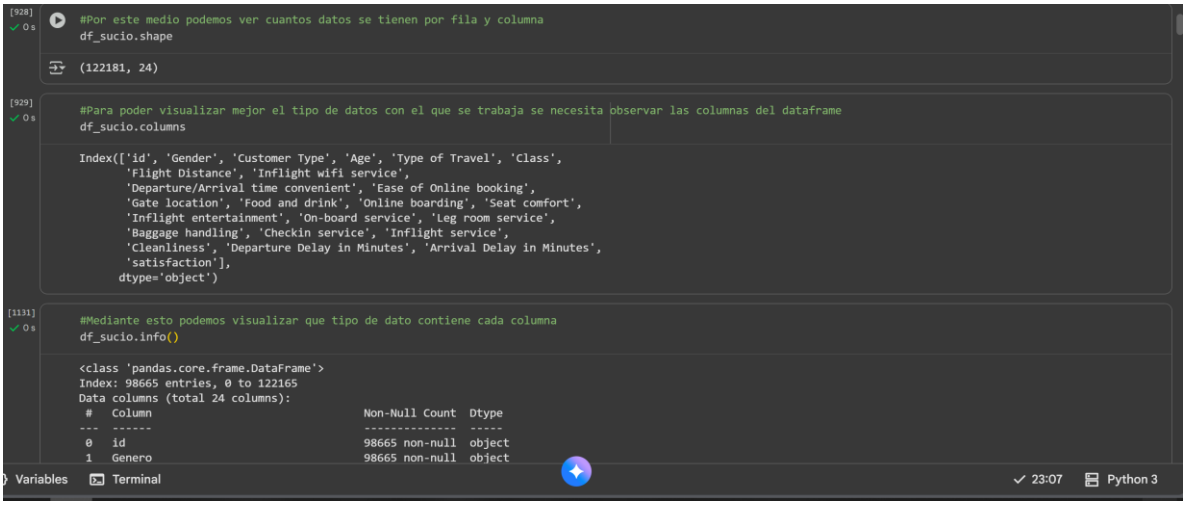
Servicio de Check-in	Evaluación sobre la eficiencia y facilidad del proceso de registro o facturación antes del vuelo.
Servicio de Vuelo	Calificación global del servicio brindado por la tripulación durante el viaje.
Limpieza	Opinión sobre la limpieza y presentación general de la aeronave y las instalaciones.
Retraso de Salida	Tiempo (en minutos) de retraso en el despegue del vuelo respecto al horario programado.
Retraso de Llegada	Tiempo (en minutos) de retraso en la llegada del vuelo a su destino.
Satisfacción	Valor final que representa el nivel general de satisfacción del pasajero (Satisfecho, Neutral o Insatisfecho).

Proceso de Limpieza

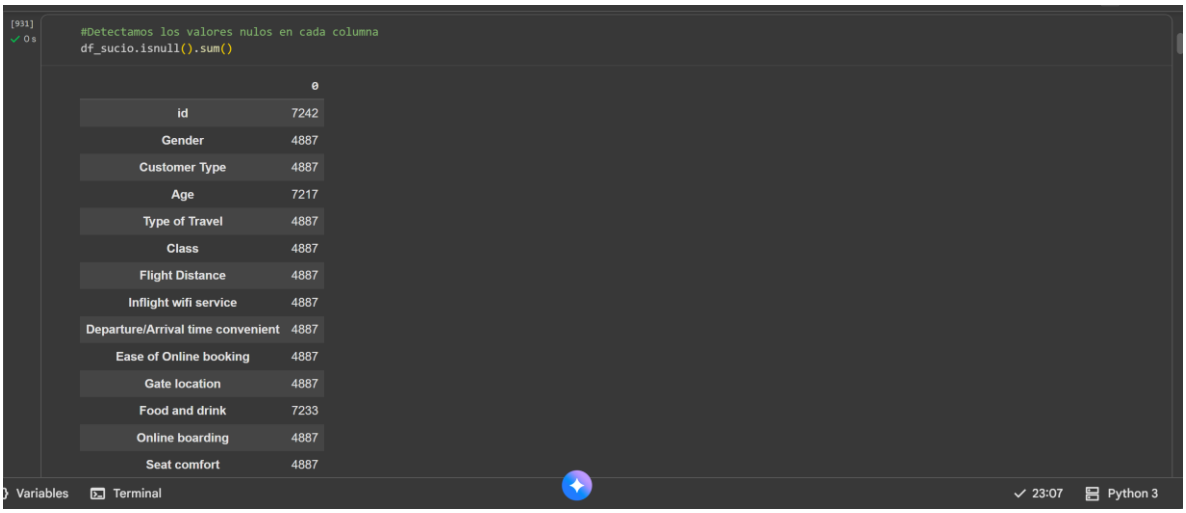
En primera instancia, se elimina la columna “Unnamed: 0” ya que solo representa un índice en los datos



Despues visualizamos las dimensiones de la base datos, además de las columnas y el tipo de dato que contienen



Luego, observamos los valores nulos



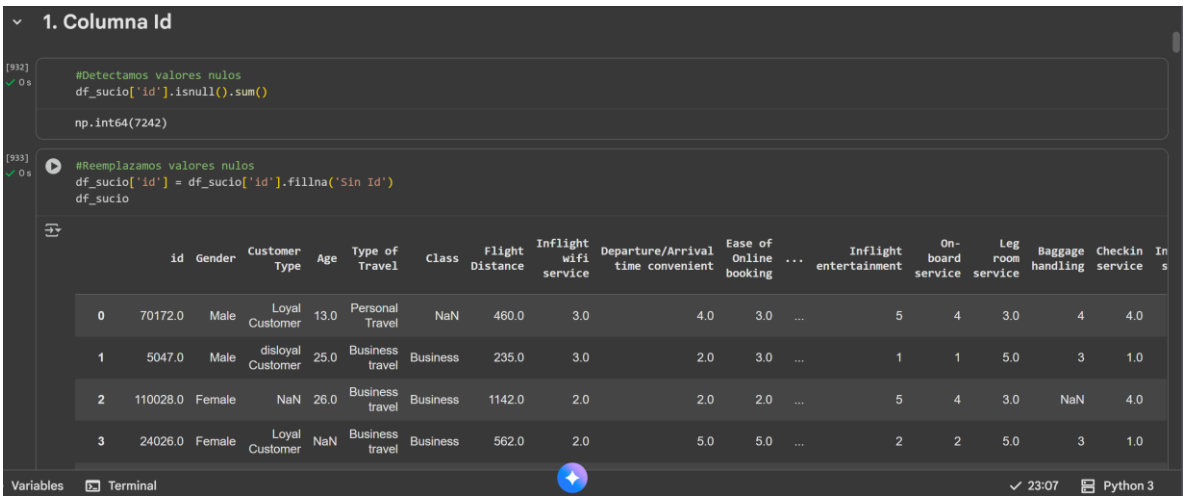
```
#Detectamos los valores nulos en cada columna
df_sucio.isnull().sum()
```

	0
id	7242
Gender	4887
Customer Type	4887
Age	7217
Type of Travel	4887
Class	4887
Flight Distance	4887
Inflight wifi service	4887
Departure/Arrival time convenient	4887
Ease of Online booking	4887
Gate location	4887
Food and drink	7233
Online boarding	4887
Seat comfort	4887

Variables Terminal 23:07 Python 3

Limpieza en cada columna

Algo fundamental en la limpieza, es detectar valores nulos (NaN), para cuando se identifiquen, lo ideal seria buscar un valor para reemplazar ese dato, ya que eliminarlos siginifica pérdida de información, como se puede observar aquí



1. Columna Id

```
#Detectamos valores nulos
df_sucio['id'].isnull().sum()
```

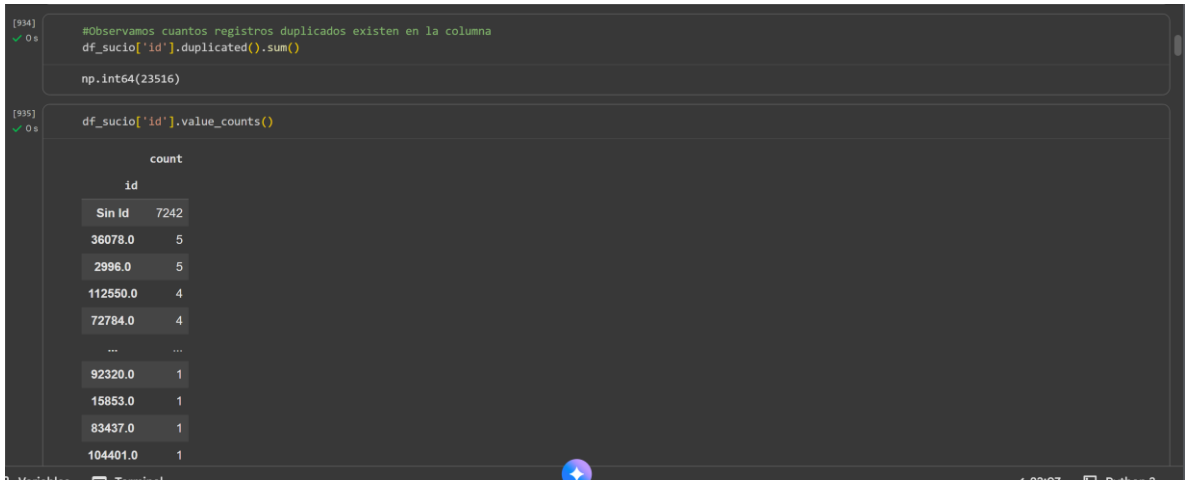
```
np.int64(7242)
```

```
#Reemplazamos valores nulos
df_sucio['id'] = df_sucio['id'].fillna('Sin Id')
df_sucio
```

	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	In s
0	70172.0	Male	Loyal Customer	13.0	Personal Travel	NaN	460.0	3.0	4.0	3.0	...	5	4	3.0	4	4.0	
1	5047.0	Male	disloyal Customer	25.0	Business travel	Business	235.0	3.0	2.0	3.0	...	1	1	5.0	3	1.0	
2	110028.0	Female	NaN	26.0	Business travel	Business	1142.0	2.0	2.0	2.0	...	5	4	3.0	NaN	4.0	
3	24026.0	Female	Loyal Customer	NaN	Business travel	Business	562.0	2.0	5.0	5.0	...	2	2	5.0	3	1.0	

Variables Terminal 23:07 Python 3

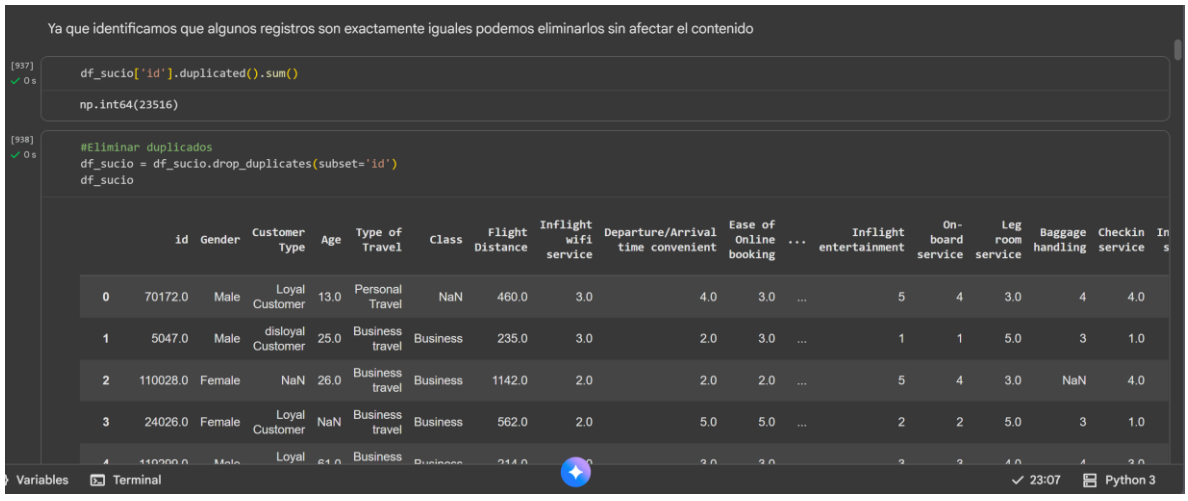
Además de también identificar datos duplicados



```
[934] ✓ 0 s
#Observamos cuantos registros duplicados existen en la columna
df_sucio['id'].duplicated().sum()

np.int64(23516)

[935] ✓ 0 s
df_sucio['id'].value_counts()
```

id	count
Sin Id	7242
36078.0	5
2996.0	5
112550.0	4
72784.0	4
...	...
92320.0	1
15853.0	1
83437.0	1
104401.0	1

Como los registros duplicados son exactamente iguales entre si, podemos eliminarlos sin tener una pérdida de información



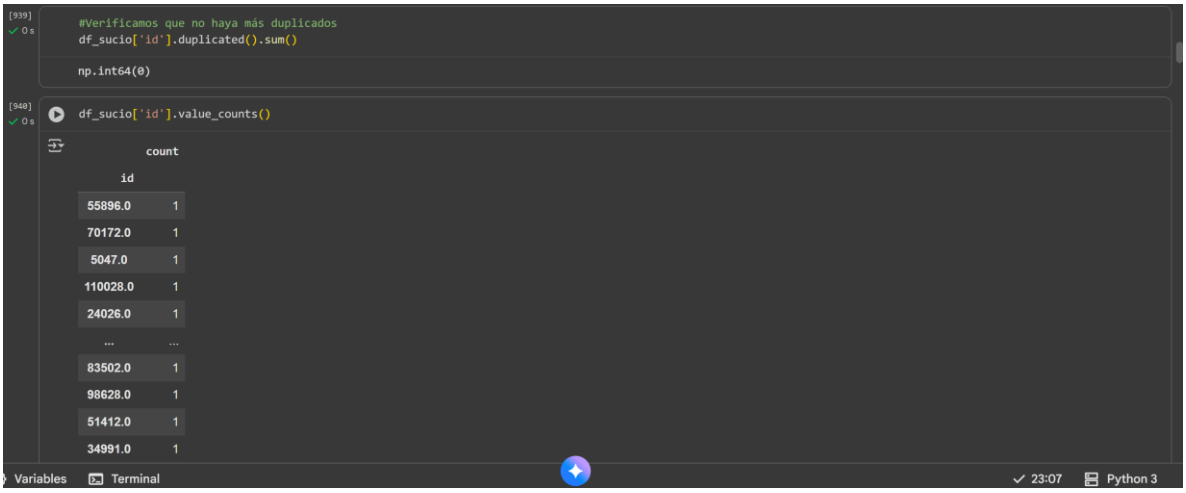
Ya que identificamos que algunos registros son exactamente iguales podemos eliminarlos sin afectar el contenido

```
[937] ✓ 0 s
df_sucio['id'].duplicated().sum()

np.int64(23516)

[938] ✓ 0 s
#Eliminar duplicados
df_sucio = df_sucio.drop_duplicates(subset='id')
df_sucio
```

	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	In s
0	70172.0	Male	Loyal Customer	13.0	Personal Travel	NaN	460.0	3.0	4.0	3.0	...	5	4	3.0	4	4.0	
1	5047.0	Male	disloyal Customer	25.0	Business travel	Business	235.0	3.0	2.0	3.0	...	1	1	5.0	3	1.0	
2	110028.0	Female	NaN	26.0	Business travel	Business	1142.0	2.0	2.0	2.0	...	5	4	3.0	NaN	4.0	
3	24026.0	Female	Loyal Customer	NaN	Business travel	Business	562.0	2.0	5.0	5.0	...	2	2	5.0	3	1.0	
4	119299.0	Male	Loyal	61.0	Business	Business	214.0	3.0	3.0	3.0		3	3	4.0	4	3.0	

Variables Terminal ✓ 23:07 Python 3

Verificamos que no hayan más duplicados



```
#Verificamos que no haya más duplicados
df_sucio['id'].duplicated().sum()
```

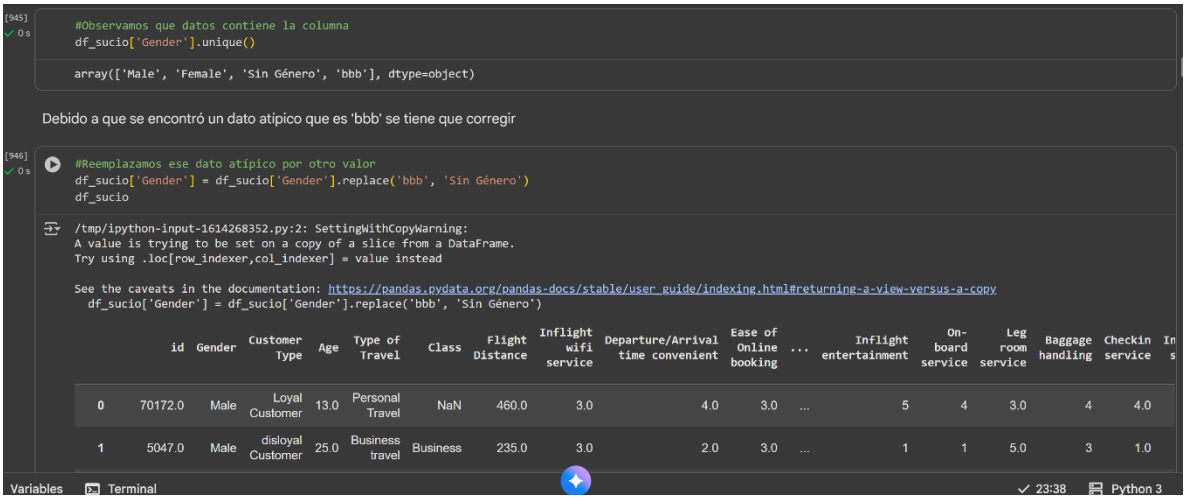
```
np.int64(0)
```

```
df_sucio['id'].value_counts()
```

id	count
55896.0	1
70172.0	1
5047.0	1
110028.0	1
24026.0	1
...	...
83502.0	1
98628.0	1
51412.0	1
34991.0	1

Variables Terminal 23:07 Python 3

Para datos atípicos como cadenas de texto, primero se tienen que identificar, y después reemplazar para que tampoco haya una pérdida de información



```
#Observamos que datos contiene la columna
df_sucio['Gender'].unique()
```

```
array(['Male', 'Female', 'Sin Género', 'bbb'], dtype=object)
```

Debido a que se encontró un dato atípico que es 'bbb' se tiene que corregir

```
#Reemplazamos ese dato atípico por otro valor
df_sucio['Gender'] = df_sucio['Gender'].replace('bbb', 'Sin Género')
df_sucio
```

```
/tmp/ipython-input-1614268352.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_sucio['Gender'] = df_sucio['Gender'].replace('bbb', 'Sin Género')
```

	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	In s
0	70172.0	Male	Loyal Customer	13.0	Personal Travel	NaN	460.0	3.0	4.0	3.0	...	5	4	3.0	4	4.0	
1	5047.0	Male	disloyal Customer	25.0	Business travel	Business	235.0	3.0	2.0	3.0	...	1	1	5.0	3	1.0	

Variables Terminal 23:38 Python 3

En el caso de que los datos numéricos estén dados por otro formato (tipo) se debe cambiar al tipo de dato conveniente

```
[958]
✓ 0 s
#Convertimos los valores numéricos a tipo entero
df_sucio['Age']=df_sucio['Age'].astype(int)
df_sucio
```

```
/tmp/ipython-input-3523648331.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_sucio['Age']=df_sucio['Age'].astype(int)
```

	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	In s
0	70172.0	Male	Loyal Customer	13	Personal Travel	NaN	460.0	3.0	4.0	3.0	...	5	4	3.0	4	4.0	
1	5047.0	Male	disloyal Customer	25	Business travel	Business	235.0	3.0	2.0	3.0	...	1	1	5.0	3	1.0	
2	110028.0	Female	Sin Cliente Especifico	26	Business travel	Business	1142.0	2.0	2.0	2.0	...	5	4	3.0	NaN	4.0	
3	24026.0	Female	Loyal Customer	39	Business travel	Business	562.0	2.0	5.0	5.0	...	2	2	5.0	3	1.0	
4	119299.0	Male	Loyal Customer	61	Business travel	Business	214.0	3.0	3.0	3.0	...	3	3	4.0	4	3.0	

Además se debe traducir el nombre de las columnas y algunos datos presentes en la base

Traducción para algunos datos

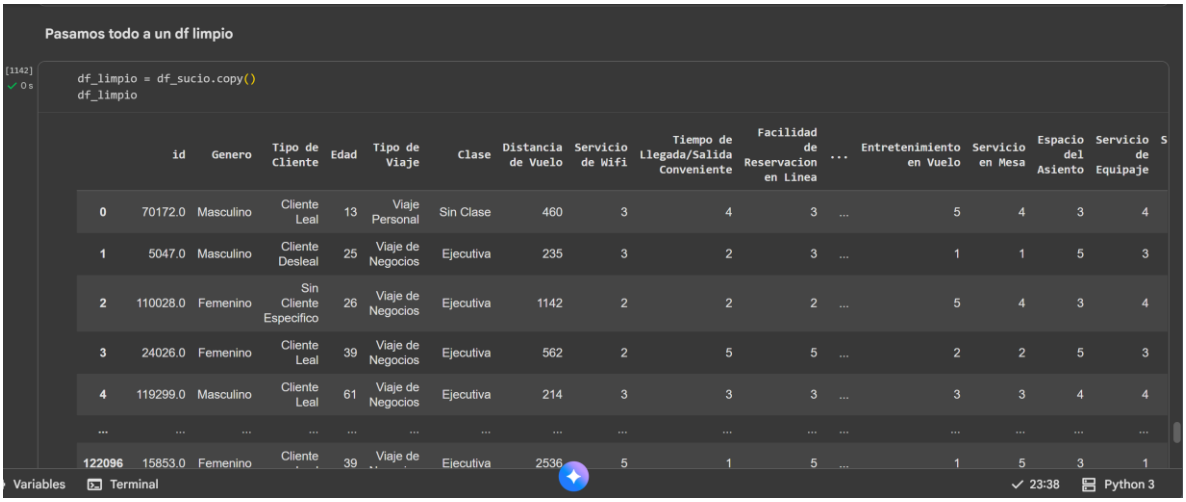
```
[1136]
✓ 0 s
#Columna Genero
df_sucio['Genero'] = df_sucio['Genero'].replace({
    'Female': 'Femenino',
    'Male': 'Masculino'
})

#Columna Tipo de Cliebre
df_sucio['Tipo de Cliente'] = df_sucio['Tipo de Cliente'].replace({
    'Loyal Customer': 'Cliente Leal',
    'disloyal Customer': 'Cliente Desleal'
})

#Columna Tipo de Viaje
df_sucio['Tipo de Viaje'] = df_sucio['Tipo de Viaje'].replace({
    'Business travel': 'Viaje de Negocios',
    'Personal Travel': 'Viaje Personal'
})

#Columna Clase
df_sucio['Clase'] = df_sucio['Clase'].replace({
    'Business': 'Ejecutiva',
    'Eco': 'Económica',
    'Eco Plus': 'Económica Plus'
})
```

Variables Terminal ✓ 23:38 Python 3

```
dtype='object')
```

```
[1138]
✓ 0 s
# Cambiar nombres de columnas
df_sucio = df_sucio.rename(columns={
    'Gender': 'Genero',
    'Customer Type': 'Tipo de Cliente',
    'Age': 'Edad',
    'Type of Travel': 'Tipo de Viaje',
    'Class': 'Clase',
    'Flight Distance': 'Distancia de Vuelo',
    'Inflight wifi service': 'Servicio de Wifi',
    'Departure/Arrival time convenient': 'Tiempo de Llegada/Salida Conveniente',
    'Ease of Online booking': 'Facilidad de Reservacion en Linea',
    'Gate location': 'Ubicacion de Puerta',
    'Food and drink': 'Comida y Bebida',
    'Online boarding': 'Abordaje en Linea',
    'Seat comfort': 'Comodidad de Asiento',
    'Inflight entertainment': 'Entretenimiento en Vuelo',
    'On-board service': 'Servicio en Mesa',
    'Leg room service': 'Espacio del Asiento',
    'Baggage handling': 'Servicio de Equipaje',
    'Checkin service': 'Servicio de Checkin',
    'Inflight service': 'Servicio de Vuelo',
    'Cleanliness': 'Limpieza',
```

Teniendo eso se pasa a un dataframe limpio



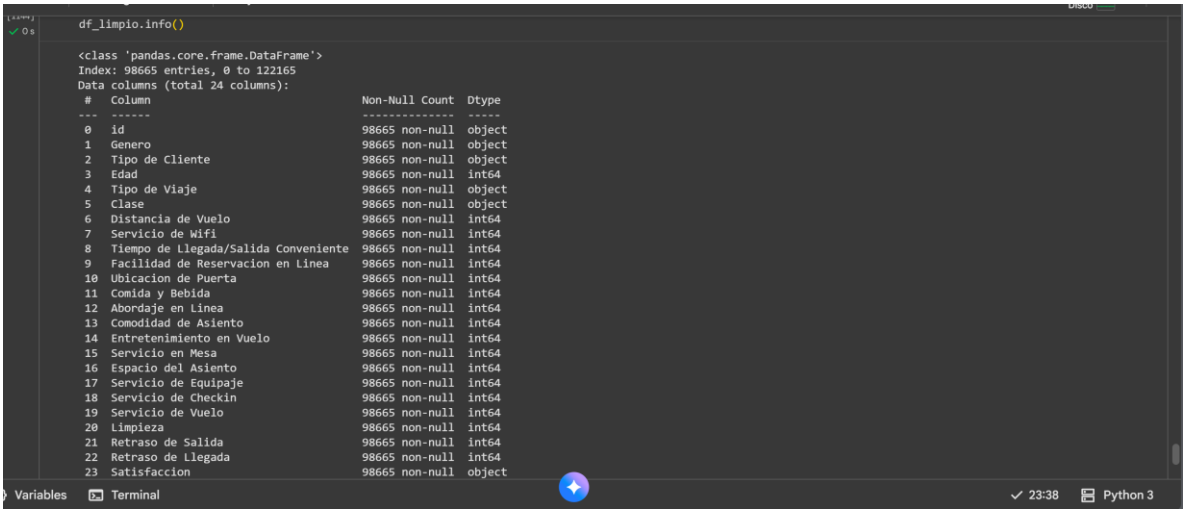
Pasamos todo a un df limpio

```
[1142]
0 s
df_limpio = df_sucio.copy()
df_limpio
```

	id	Genero	Tipo de Cliente	Edad	Tipo de Viaje	Clase	Distancia de Vuelo	Servicio de Wifi	Tiempo de Llegada/Salida Conveniente	Facilidad de Reservacion en Linea	...	Entretenimiento en Vuelo	Servicio en Mesa	Espacio del Asiento	Servicio de Equipaje	S
0	70172.0	Masculino	Cliente Leal	13	Viaje Personal	Sin Clase	460	3	4	3	...	5	4	3	4	
1	5047.0	Masculino	Cliente Desleal	25	Viaje de Negocios	Ejecutiva	235	3	2	3	...	1	1	5	3	
2	110028.0	Femenino	Sin Cliente Especifico	26	Viaje de Negocios	Ejecutiva	1142	2	2	2	...	5	4	3	4	
3	24026.0	Femenino	Cliente Leal	39	Viaje de Negocios	Ejecutiva	562	2	5	5	...	2	2	5	3	
4	119299.0	Masculino	Cliente Leal	61	Viaje de Negocios	Ejecutiva	214	3	3	3	...	3	3	4	4	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
122096	15853.0	Femenino	Cliente	39	Viaje de	Ejecutiva	2536	5	1	5	...	1	5	3	1	

Variables Terminal 23:38 Python 3

Finalmente, se hace la verificación de todo el proceso



```
0 s
df_limpio.info()

<class 'pandas.core.frame.DataFrame'>
Index: 98665 entries, 0 to 122165
Data columns (total 24 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   id                                98665 non-null  object
 1   Genero                            98665 non-null  object
 2   Tipo de Cliente                   98665 non-null  object
 3   Edad                              98665 non-null  int64
 4   Tipo de Viaje                     98665 non-null  object
 5   Clase                             98665 non-null  object
 6   Distancia de Vuelo                98665 non-null  int64
 7   Servicio de Wifi                  98665 non-null  int64
 8   Tiempo de Llegada/Salida Conveniente  98665 non-null  int64
 9   Facilidad de Reservacion en Linea  98665 non-null  int64
 10  Ubicacion de Puerta               98665 non-null  int64
 11  Comida y Bebida                   98665 non-null  int64
 12  Abordaje en Linea                 98665 non-null  int64
 13  Comodidad de Asiento              98665 non-null  int64
 14  Entretenimiento en Vuelo          98665 non-null  int64
 15  Servicio en Mesa                  98665 non-null  int64
 16  Espacio del Asiento               98665 non-null  int64
 17  Servicio de Equipaje              98665 non-null  int64
 18  Servicio de Checkin               98665 non-null  int64
 19  Servicio de Vuelo                 98665 non-null  int64
 20  Limpieza                          98665 non-null  int64
 21  Retraso de Salida                 98665 non-null  int64
 22  Retraso de Llegada                98665 non-null  int64
 23  Satisfaccion                      98665 non-null  object
```

Variables Terminal 23:38 Python 3

```
[1145]
✓ 0 s
#Verificacion de valores nulos
df_limpio.isnull().sum()
```

	0
id	0
Genero	0
Tipo de Cliente	0
Edad	0
Tipo de Viaje	0
Clase	0
Distancia de Vuelo	0
Servicio de Wifi	0
Tiempo de Llegada/Salida Conveniente	0
Facilidad de Reservacion en Linea	0
Ubicacion de Puerta	0
Comida y Bebida	0
Abordaje en Linea	0
Comodidad de Asiento	0

Variables Terminal 23:38 Python 3

```
[1148]
✓ 0 s
#Verificación de datos atipicos con un ciclo for
#Ciclo for para detectar valores atípicos (por ejemplo texto)
lista_col=df_limpio.columns
for i in lista_col:
    print(f"En la columna {i} los bbb son: {df_sucio[df_sucio[i] == 'bbb'].shape[0]}")
```

```
En la columna id los bbb son: 0
En la columna Genero los bbb son: 0
En la columna Tipo de Cliente los bbb son: 0
En la columna Edad los bbb son: 0
En la columna Tipo de Viaje los bbb son: 0
En la columna Clase los bbb son: 0
En la columna Distancia de Vuelo los bbb son: 0
En la columna Servicio de Wifi los bbb son: 0
En la columna Tiempo de Llegada/Salida Conveniente los bbb son: 0
En la columna Facilidad de Reservacion en Linea los bbb son: 0
En la columna Ubicacion de Puerta los bbb son: 0
En la columna Comida y Bebida los bbb son: 0
En la columna Abordaje en Linea los bbb son: 0
En la columna Comodidad de Asiento los bbb son: 0
En la columna Entretenimiento en Vuelo los bbb son: 0
En la columna Servicio en Mesa los bbb son: 0
En la columna Espacio del Asiento los bbb son: 0
En la columna Servicio de Equipaje los bbb son: 0
En la columna Servicio de Checkin los bbb son: 0
En la columna Servicio de Vuelo los bbb son: 0
En la columna Limpieza los bbb son: 0
En la columna Retraso de Salida los bbb son: 0
En la columna Retraso de Llegada los bbb son: 0
En la columna Satisfaccion los bbb son: 0
```

Variables Terminal 23:38 Python 3

Y por último se guarda en un csv limpio

```
En la columna Servicio de Vuelo los bbb son: 0
En la columna Limpieza los bbb son: 0
En la columna Retraso de Salida los bbb son: 0
En la columna Retraso de Llegada los bbb son: 0
En la columna Satisfaccion los bbb son: 0
```

```
[1146]
✓ 0 s
#Verificación de duplicados
df_limpio.duplicated().sum()
```

```
np.int64(0)
```

```
[1147]
✓ 0 s
#Guardar la base de datos en un archivo csv
df_limpio.to_csv('df_limpio.csv', index=False)
```

Variables Terminal 23:38 Python 3

4. Conclusiones

• Qué problemas principales presentaba la base.

Antes del proceso de depuración y limpieza, la base de datos presentaba diversos problemas que dificultaban su análisis y afectaban la calidad de la información. Uno de los principales inconvenientes era la **presencia de valores nulos o faltantes** en varias columnas, especialmente en aquellas relacionadas con la evaluación de servicios o características del pasajero. Estos valores ausentes generaban inconsistencias en los cálculos estadísticos y podían sesgar los resultados del análisis de satisfacción.

Otro problema importante fue la **inconsistencia en los tipos de datos**. Algunas variables numéricas, como el identificador de pasajero o los retrasos de vuelo, aparecían registradas como texto o con formato decimal incorrecto, lo que impedía realizar operaciones matemáticas y análisis cuantitativos adecuados. Asimismo, se detectaron **errores de formato y escritura** en variables categóricas, como diferencias en mayúsculas, espacios innecesarios o etiquetas mal escritas (por ejemplo, “Satisfecho” y “satisfecho” tratadas como valores distintos).

También se observaron **valores atípicos o poco realistas**, especialmente en variables como edad, distancia de vuelo y tiempos de retraso, lo que indicaba errores de registro o medición. Estos datos extremos podían distorsionar los resultados del análisis y generar conclusiones erróneas si no eran tratados correctamente.

Finalmente, existían **duplicados de registros**, lo que aumentaba el tamaño del dataset sin aportar nueva información y afectaba la representatividad de los resultados. Todos estos problemas hicieron necesario aplicar un proceso riguroso de limpieza, estandarización y validación de los datos para garantizar la precisión y fiabilidad del análisis posterior.

• Qué técnicas aplicaste para solucionarlos.

Se aplicó la limpieza mediante la utilización de Python y Pandas, con las cuales se pudo corregir errores como datos nulos, duplicados, texto incoherente, etc.

• Qué aprendiste del proceso.

Generalmente, se pueden tomar muchos métodos y decisiones para optimizar un proceso mediante ciencia de datos, todo este proceso sirve para saber en que momento es adecuado, tomar una decisión, o en todo caso, mejorar y optimizar procesos, sin duda, esto es de gran ayuda dentro del ámbito de la tecnología y los negocios, mostrando un resalte dentro de otras disciplinas. En conclusión, todas estas herramientas y procesos ayudan a la humanidad para alcanzar un futuro mejor.