- Nombre del proyecto.

"Análisis de la Satisfacción de los Clientes en el Uso de Aerolíneas Comerciales".

- Nombre del alumno

Federico Camacho Cagal

- Materia

Introducción a la Ciencia de Datos

- Profesor

Jaime Alejandro Romero Sierra

- Fecha de entrega.

29 de noviembre de 2025

Introducción

• Descripción breve del objetivo del proyecto.

El objetivo principal de este proyecto es analizar los patrones de comportamiento y experiencia de los pasajeros de una aerolínea para determinar qué factores influyen mayormente en su nivel de satisfacción. A través de técnicas de ciencia de datos, se busca construir un modelo o realizar un análisis exploratorio exhaustivo que permita identificar las variables críticas (como retrasos, comodidad o servicios a bordo) que diferencian a un cliente satisfecho de uno insatisfecho o neutral.

• Justificación y contexto

En la industria aeronáutica, la competencia es agresiva y la lealtad del cliente es volátil. Entender qué hace feliz a un pasajero no es solo una cuestión de calidad, sino de rentabilidad.

La importancia de resolver esta problemática radica en tres puntos clave detectados en el análisis preliminar:

1. Retención de Clientes: Un cliente insatisfecho es un cliente que probablemente elegirá otra aerolínea en su próximo viaje. Predecir la insatisfacción permite actuar proactivamente antes de perder al usuario.

2. Optimización de Inversiones: A menudo las aerolíneas invierten millones en reducir retrasos operativos. Sin embargo, nuestro análisis sugiere que la experiencia digital (App, Wifi, Abordaje) tiene un impacto mucho mayor en la percepción del cliente. Este modelo ayudará a validar dónde es más rentable invertir.

3. Detección de "Falsos Positivos" en el Servicio: Identificar que los retrasos moderados no son la causa principal de las quejas permite a la aerolínea enfocar sus recursos en mejorar la experiencia a bordo (Confort y Entretenimiento) para mitigar los inconvenientes operativos inevitables.

• Fuentes de Datos

Origen: Los datos provienen de encuestas de satisfacción de pasajeros de aerolíneas, una estructura estándar utilizada frecuentemente en análisis de calidad de servicio en la industria.

Cantidad de Datos: El dataset original consta de 98,665 registros (filas) y 24 variables (columnas). Esto representa un volumen de datos robusto, suficiente para entrenar modelos complejos sin riesgo alto de sobreajuste.

Principales Características:

1. Variable Objetivo: Satisfaccion (Categórica: Insatisfecho, Neutral, Satisfecho).
2. Variables de Servicio: 14 columnas con calificaciones en escala del 1 al 5 (ej. Servicio de Wifi, Comodidad de Asiento, Limpieza).
3. Variables Demográficas y de Vuelo: Incluye Genero, Edad, Tipo de Cliente (Lealtad), Tipo de Viaje (Negocios/Personal), Clase y Distancia de Vuelo.
4. Variables Operativas: Retraso de Salida y Retraso de Llegada (en minutos).

Metodología

Proceso de limpieza de datos

Para este proyecto, la limpieza se llevó a cabo con una metodología clara. El objetivo era conservar la mayor cantidad de información posible, con lo anterior, el proceso de limpiar los datos se hizo conforme a los requerimientos, no se busca eliminar datos, ya que sería pérdida de información. Para los valores nulos, lo que se realizó fue reemplazarlos por valores que ya existian en la base de datos. Por ejemplo, si se trataba de un dato categórico, se reemplazó por la moda, y para los datos numéricos, habia dos opciones, si la distribución de los datos era normal, los nulos se reemplazaron por la media. Por otro lado, si la distribución era sesgada o bimodal, se reemplazó por la mediana, o en su defecto, por la moda.

Para los duplicados, la metodología fue eliminar los registros que fueran exactamente iguales, ya que, no aportan nada a la información del dataset, además todos son diferentes registros.

Para los valores atípicos, no se eliminaron por completo, pero un punto muy importante a tomar en cuenta es que, si los valores extremos (Outliers), distorsionaban la media de los datos, entonces no se tomaría de primera instancia como referencia.

Análisis Exploratorio de Datos

1. Descripción General de los Datos

- **Visión general:**

"El dataset contiene 98665 registros y 24 variables."

- **Tipos de Variables:**

```
La variable id es Categórica
La variable Genero es Categórica
La variable Tipo de Cliente es Categórica
La variable Edad es Numérica
La variable Tipo de Viaje es Categórica
```

```
La variable Clase es Categórica
La variable Distancia de Vuelo es Numérica
La variable Servicio de Wifi es Numérica
La variable Tiempo de Llegada/Salida Conveniente es Numérica
La variable Facilidad de Reservacion en Linea es Numérica
La variable Ubicacion de Puerta es Numérica
La variable Comida y Bebida es Numérica
La variable Abordaje en Linea es Numérica
La variable Comodidad de Asiento es Numérica
La variable Entretenimiento en Vuelo es Numérica
La variable Servicio en Mesa es Numérica
La variable Espacio del Asiento es Numérica
La variable Servicio de Equipaje es Numérica
La variable Servicio de Checkin es Numérica
La variable Servicio de Vuelo es Numérica
La variable Limpieza es Numérica
La variable Retraso de Salida es Numérica
La variable Retraso de Llegada es Numérica
La variable Satisfaccion es Categórica
```

- **Resumen estadístico:**


• Variables numéricas

```
df.describe()
```

[9] ✓ 0.3s Python

	Edad	Distancia de Vuelo	Servicio de Wifi	Tiempo de Llegada/Salida Conveniente	Facilidad de Reservacion en Linea	Ubicacion de Puerta	Comida y Bebida	Abordaje en Linea	Comodidad de Asiento	Entretenimiento en Vuelo	Servicio en Mesa	Espacio del Asiento
count	98665.000000	98665.000000	98665.000000	98665.000000	98665.000000	98665.000000	98665.000000	98665.000000	98665.000000	98665.000000	98665.000000	98665.000000
mean	39.389794	1176.493336	2.738286	3.096610	2.766310	2.978665	3.251163	3.295140	3.460285	3.396199	3.420848	3.337455
std	14.666554	980.357092	1.302336	1.505864	1.371192	1.252687	1.302573	1.320335	1.297958	1.302115	1.257509	1.290380
min	7.000000	31.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	28.000000	427.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	3.000000	2.000000	3.000000	2.000000
50%	39.000000	843.000000	3.000000	3.000000	3.000000	3.000000	3.000000	4.000000	4.000000	4.000000	4.000000	3.000000
75%	50.000000	1699.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000
max	85.000000	4983.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000

• Variables categóricas:

```
id
70172.0     1
5047.0      1
110028.0    1
24026.0     1
119299.0    1
           ..
15853.0     1
83437.0     1
104401.0    1
114393.0    1
55896.0     1
```

Genero

Femenino	47023
Masculino	45746
Sin Género	5896

Tipo de Cliente

Cliente Leal	77456
Cliente Desleal	17273
Sin Cliente Especifico	3936

Tipo de Viaje

Viaje de Negocios	63904
Viaje Personal	28883
Sin Viaje	5878

Clase

Ejecutiva	44331
Económica	41759
Económica Plus	6713
Sin Clase	5862

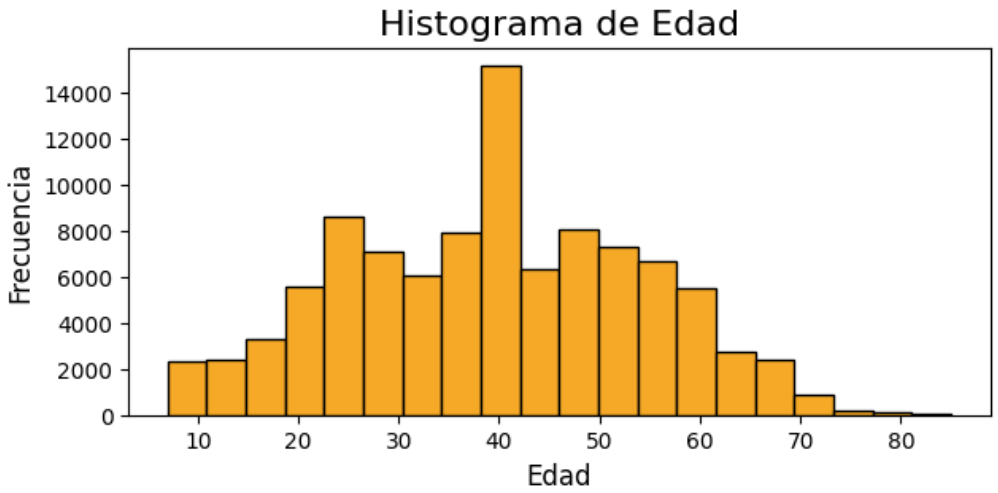
Satisfaccion

Insatisfecho	52529
Satisfecho	40263
neutral	5873

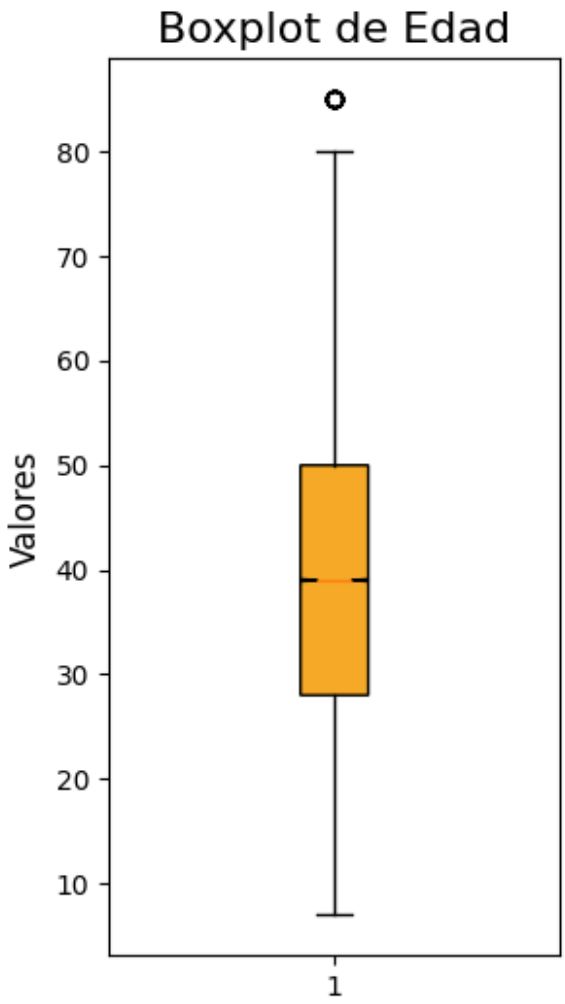
2. Visualización y Distribución de Variables Individuales

•Variables Numéricas

• Columna Edad

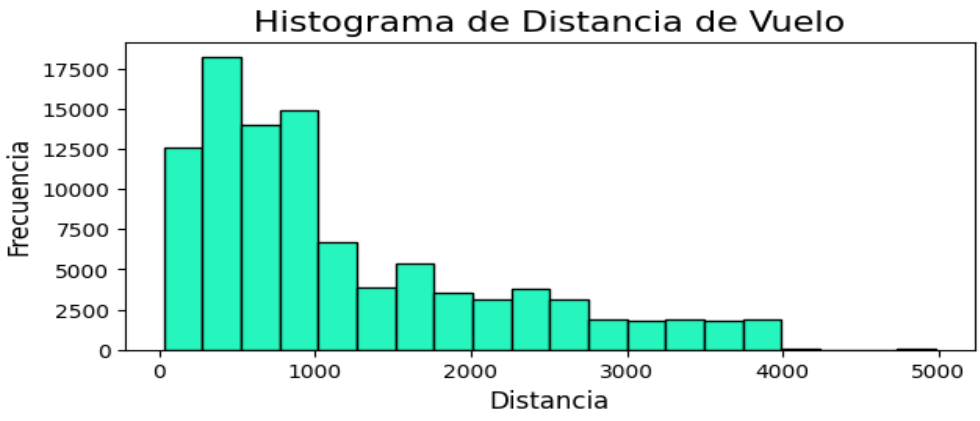


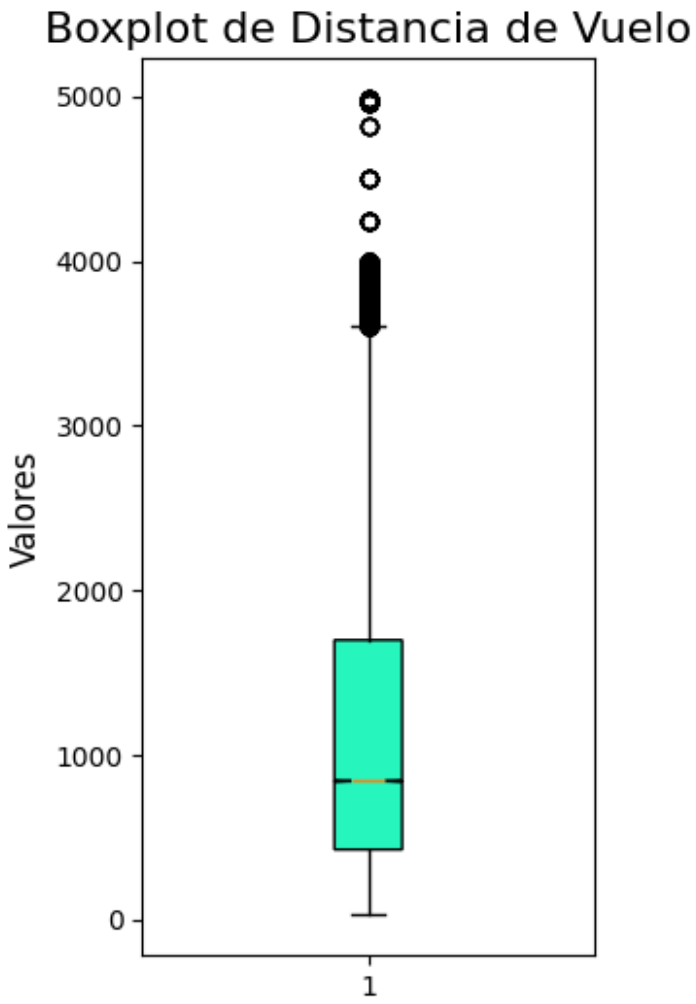
El histograma de ‘Edad’ muestra una distribución normal, indicando que la mayoría de los clientes tienen al rededor de 39 a 40 años



• Columna Distancia de Vuelo

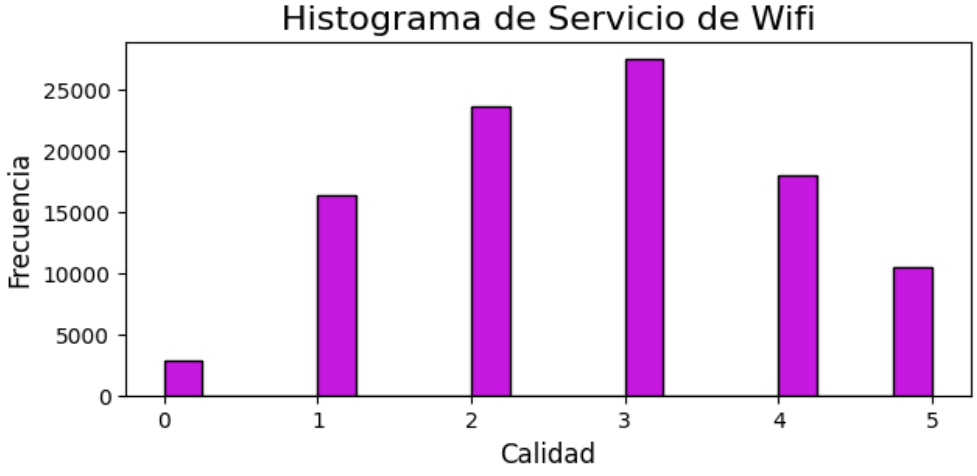
El histograma de ‘Distancia de Vuelo’ muestra una distribución sesgada a la izquierda, indicando que la mayoría de los vuelos, recorren una distancia menor a 1000 kilómetros.



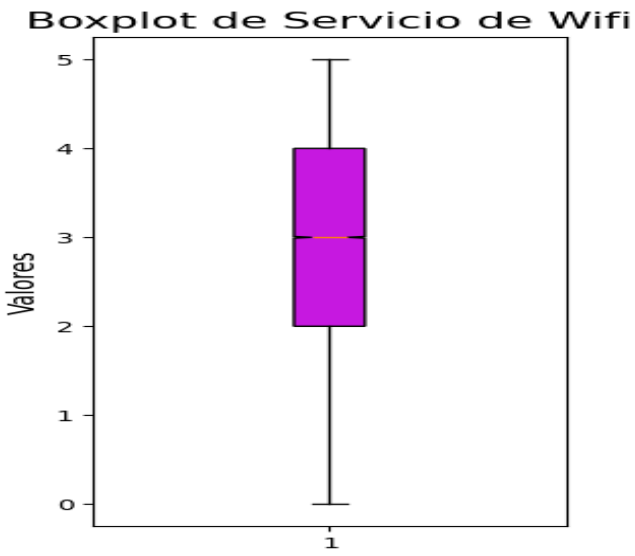


El boxplot de 'Distancia de Vuelo' muestra que la mayoría de los datos atípicos, se encientran arriba de los 3500 kilómetros.

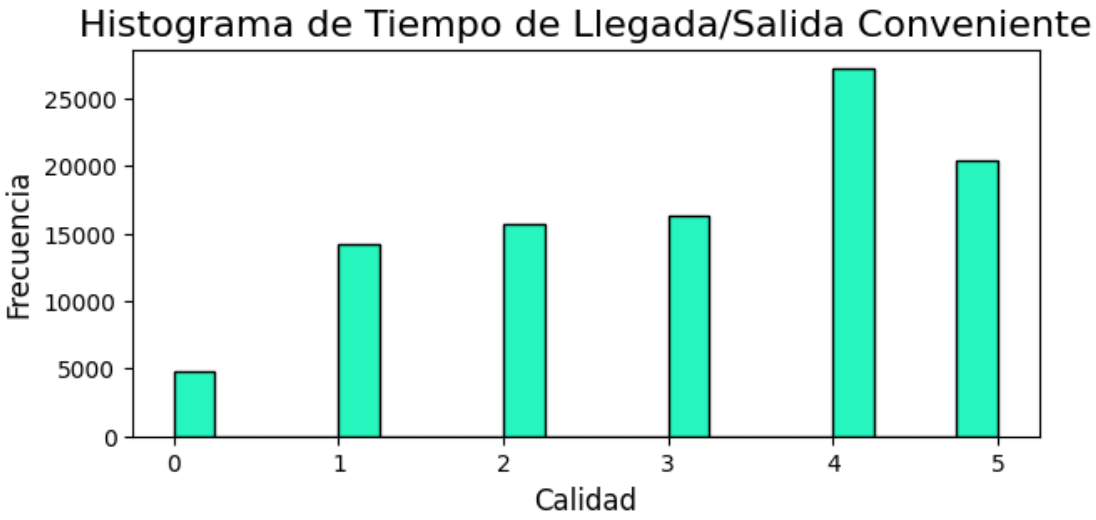
- **Columna Servicio de Wifi**

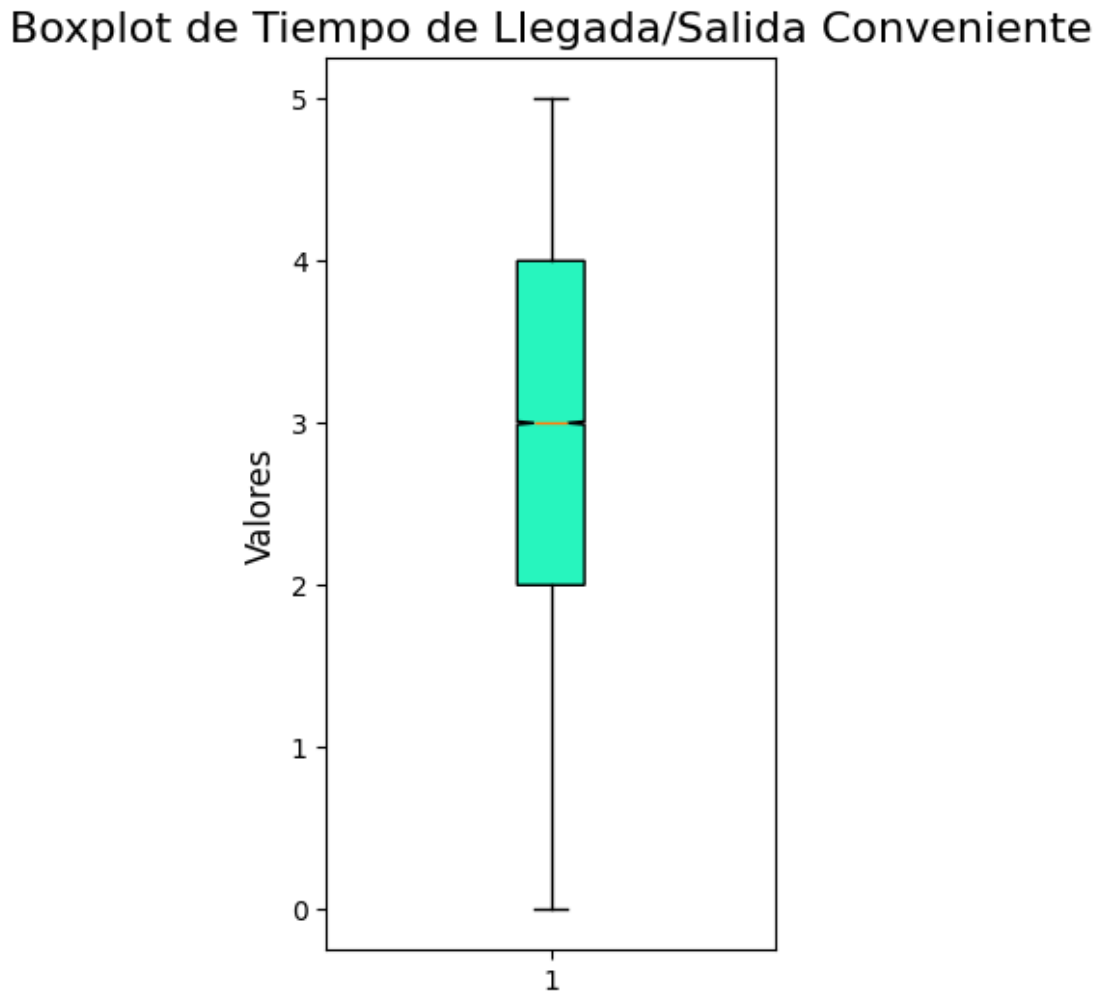


El histograma de ‘Servicio de Wifi’ muestra una distribución normal, indicando que la mayoría de los pasajeros, califican el servicio con una calidad estándar.

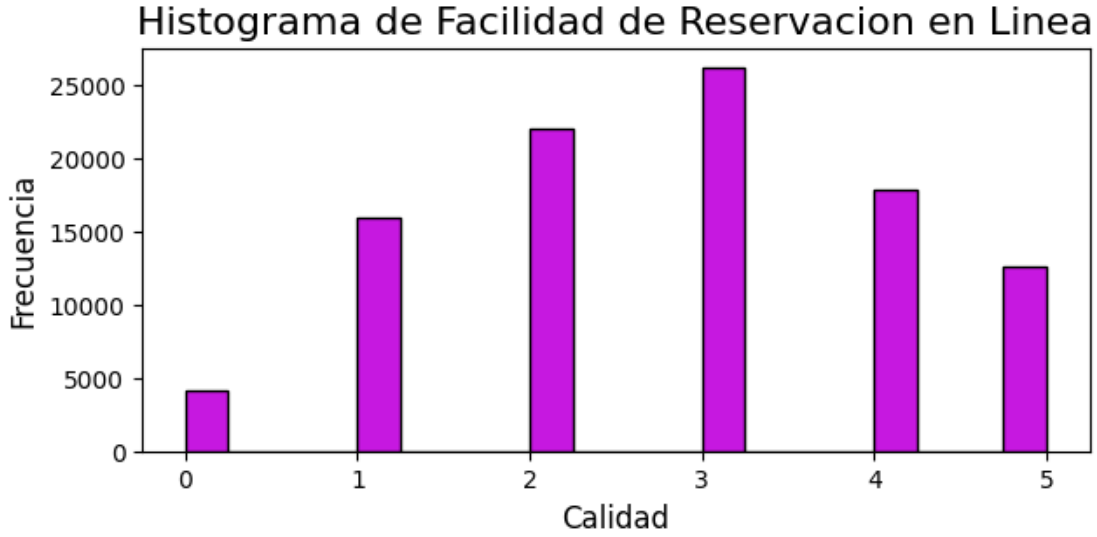


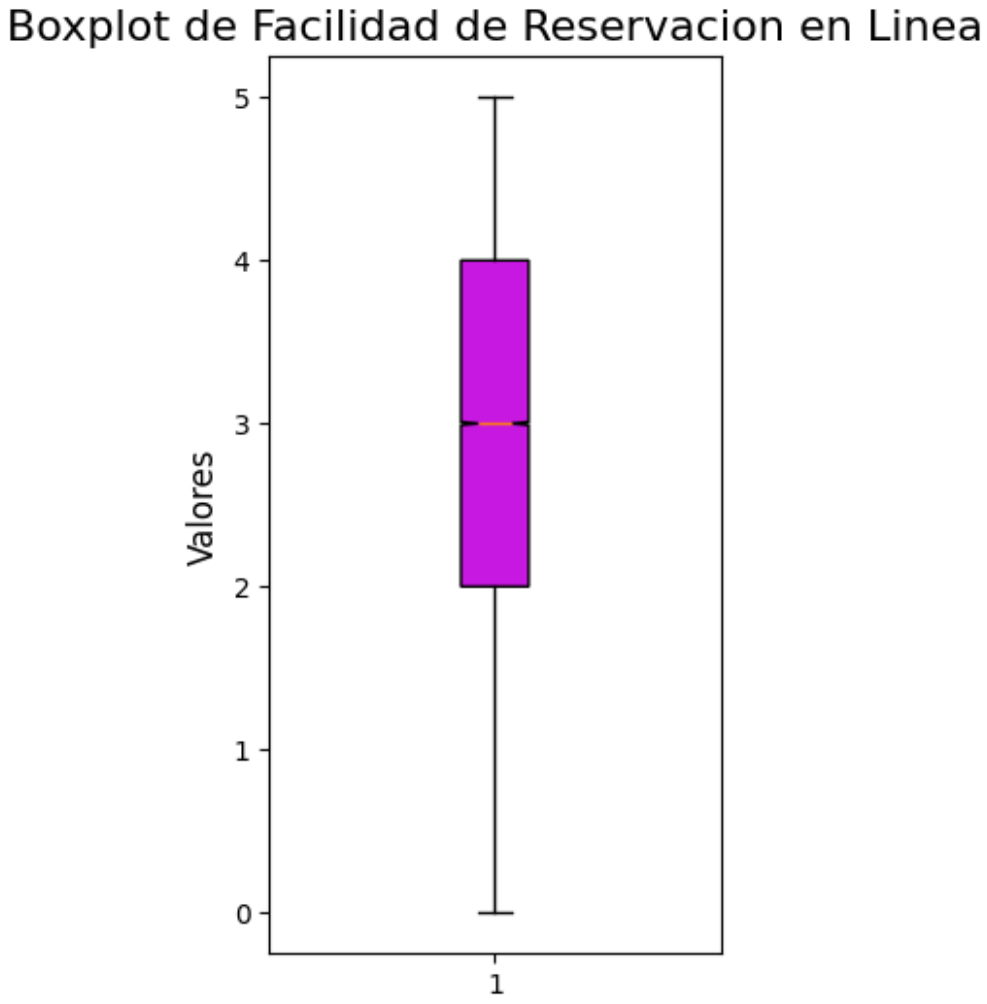
• **Columna Tiempo de Llegada/Salida Conveniente**



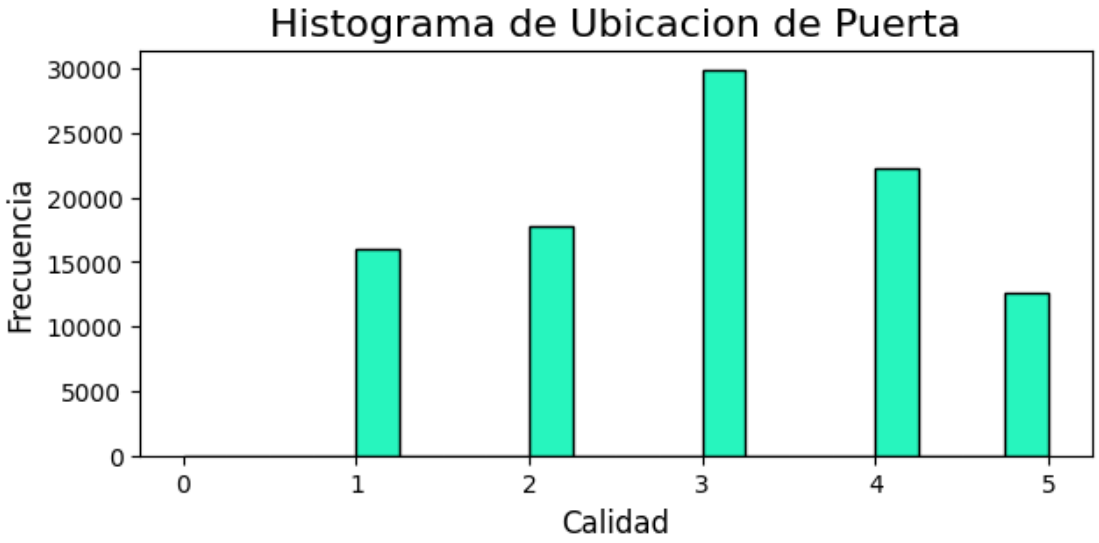


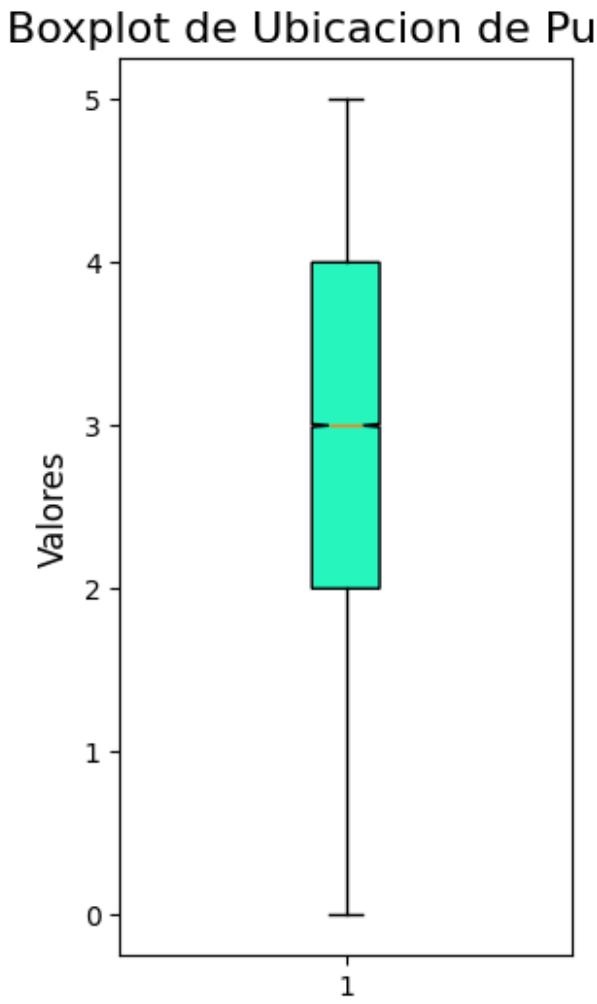
- **Columna Facilidad de Reservacion en Linea**



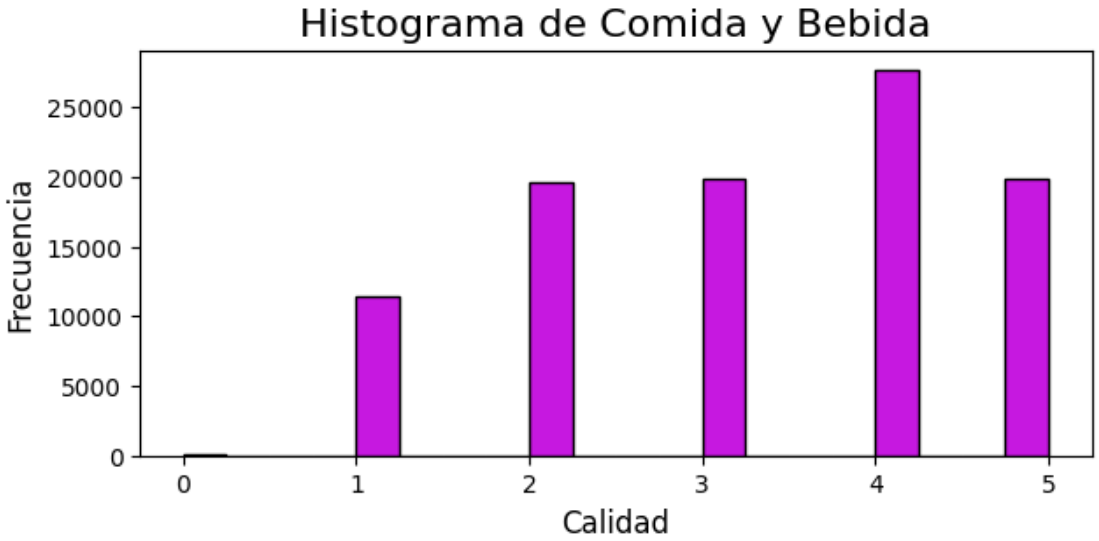


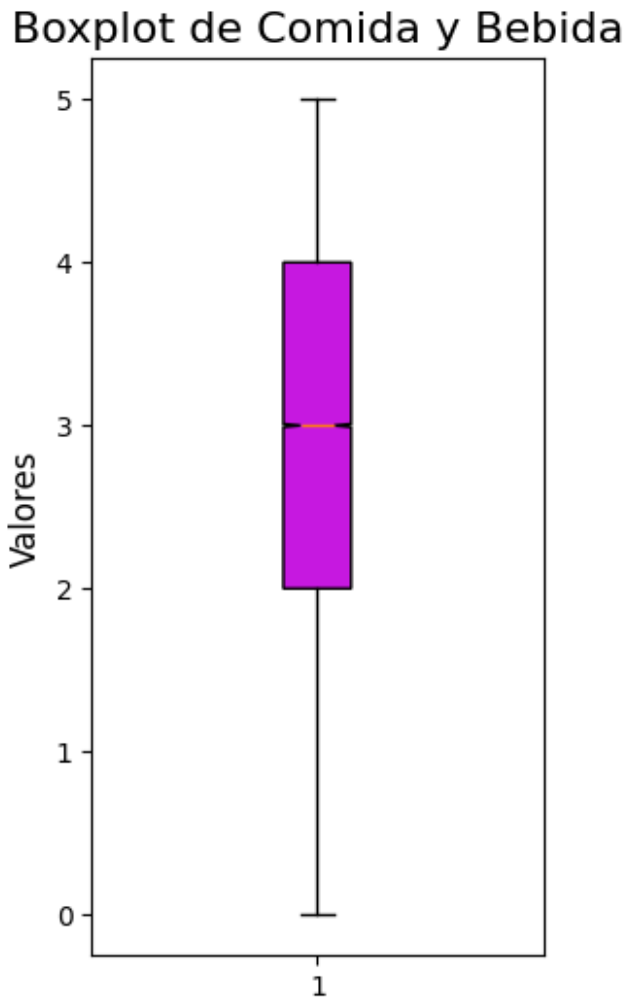
- **Columna Ubicacion de Puerta**



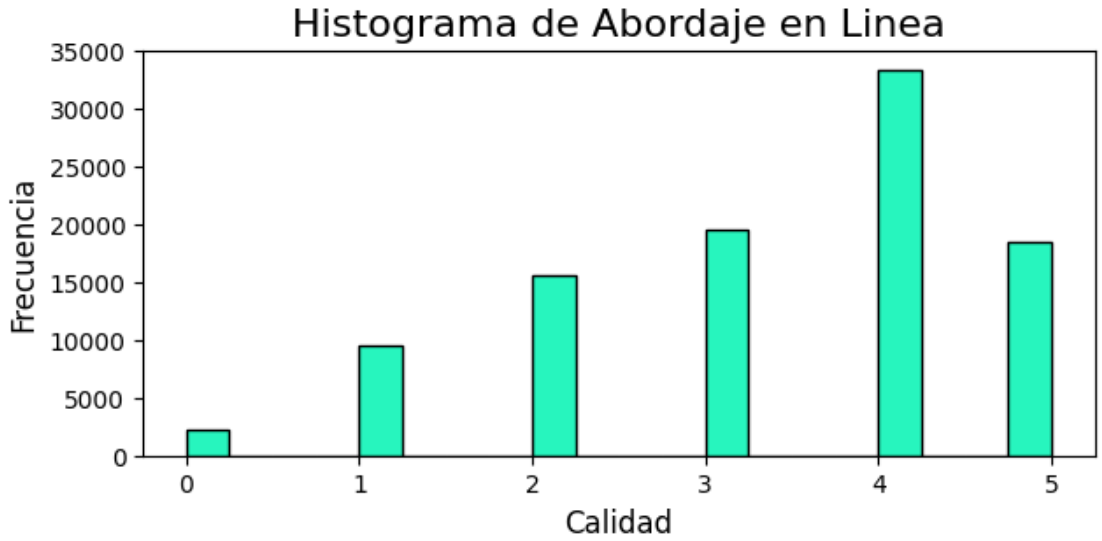


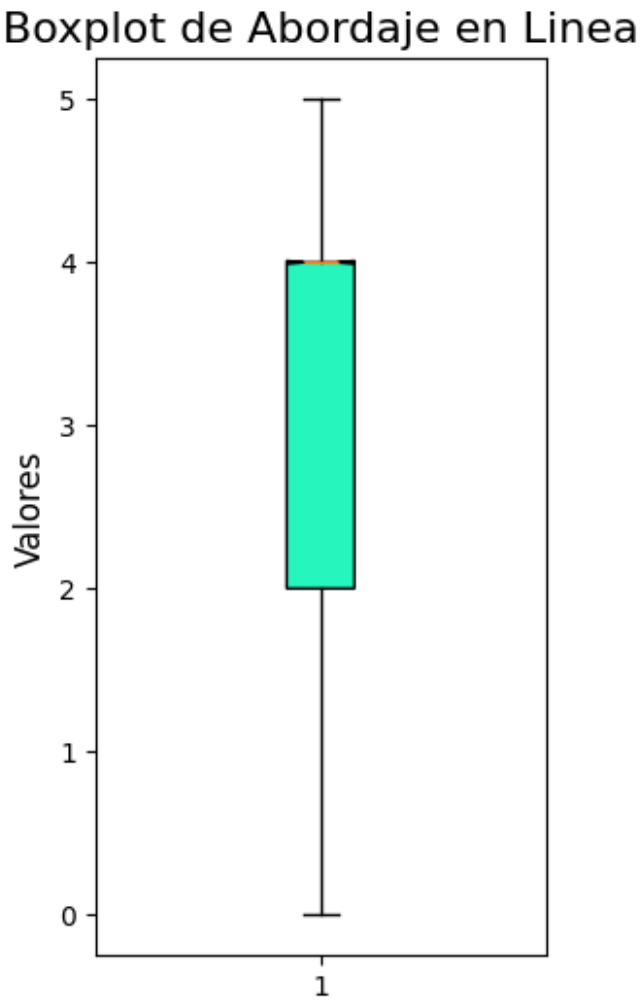
- **Columna Comida y Bebida**



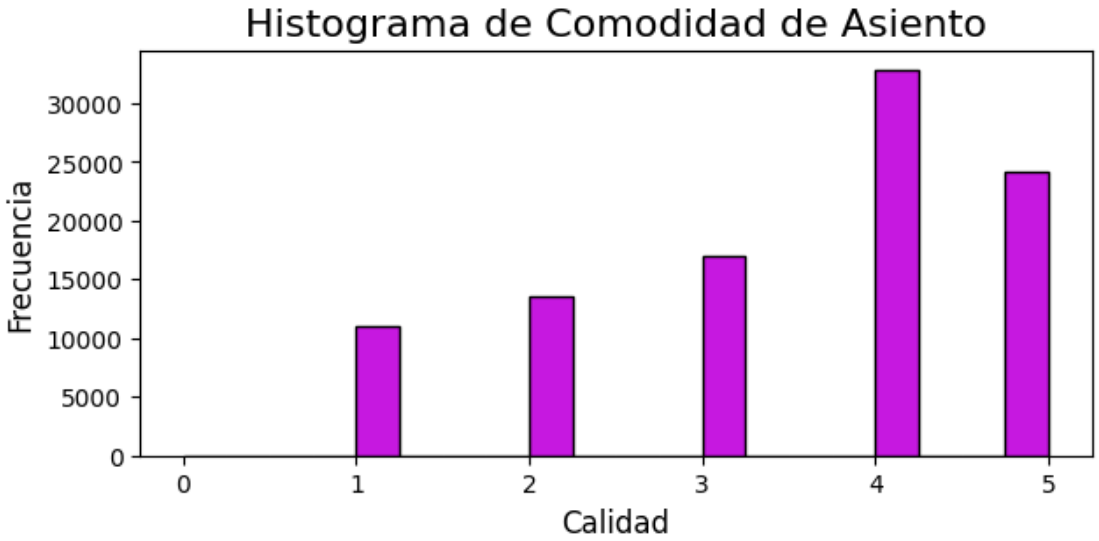


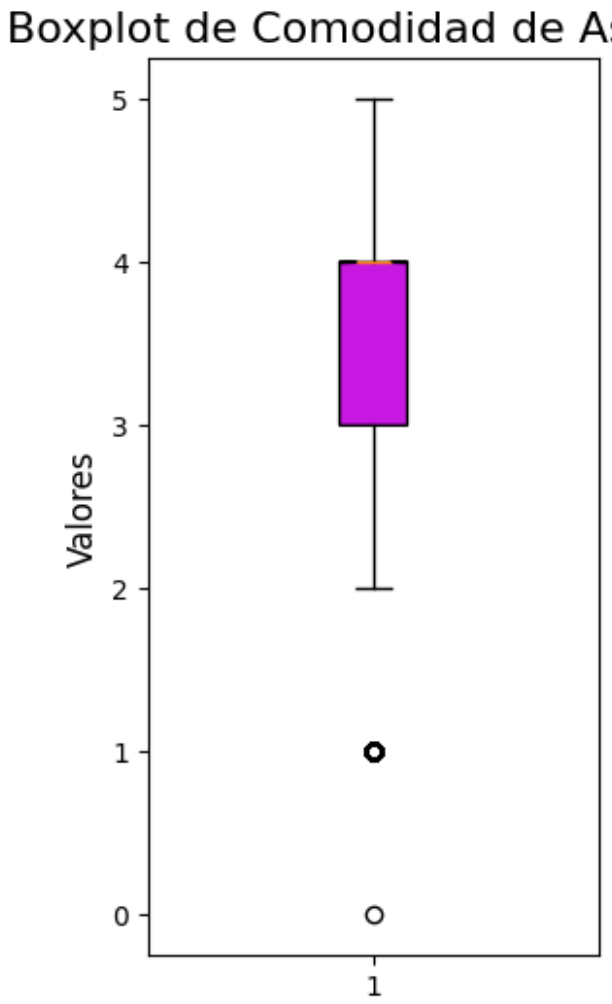
- **Columna Abordaje en Linea**



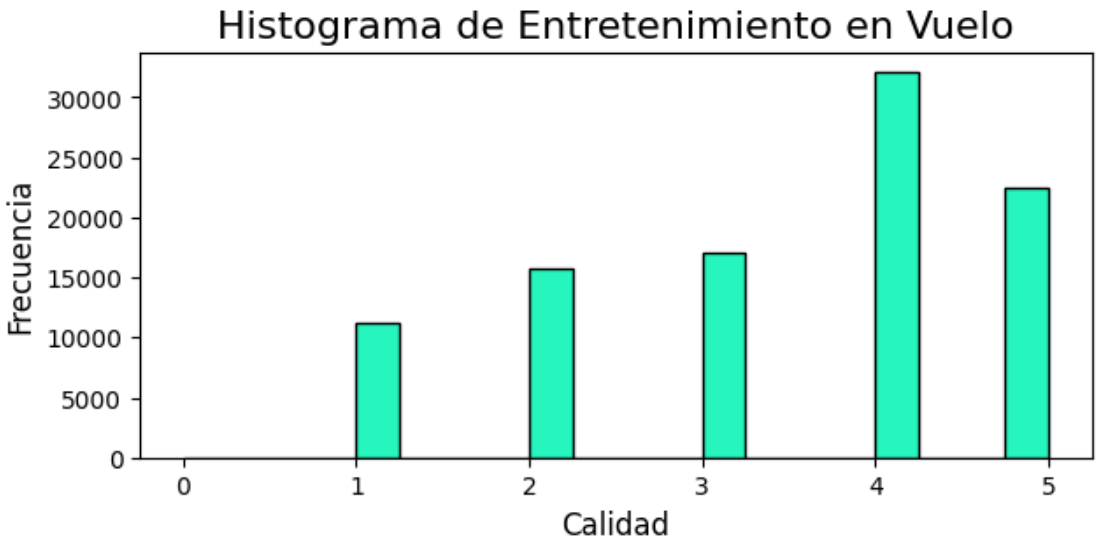


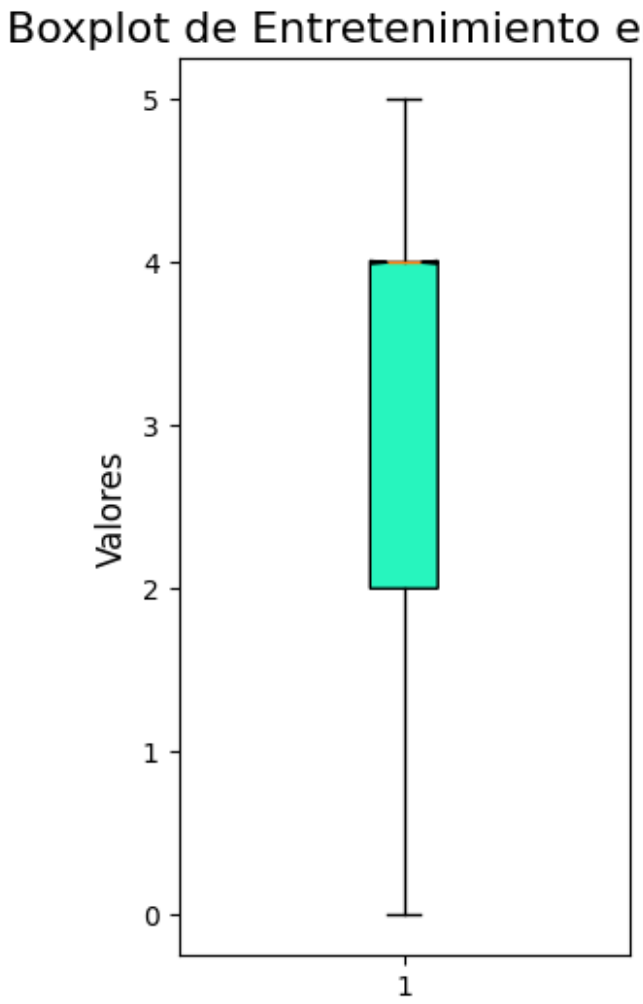
- **Columna Comodidad de Asiento**



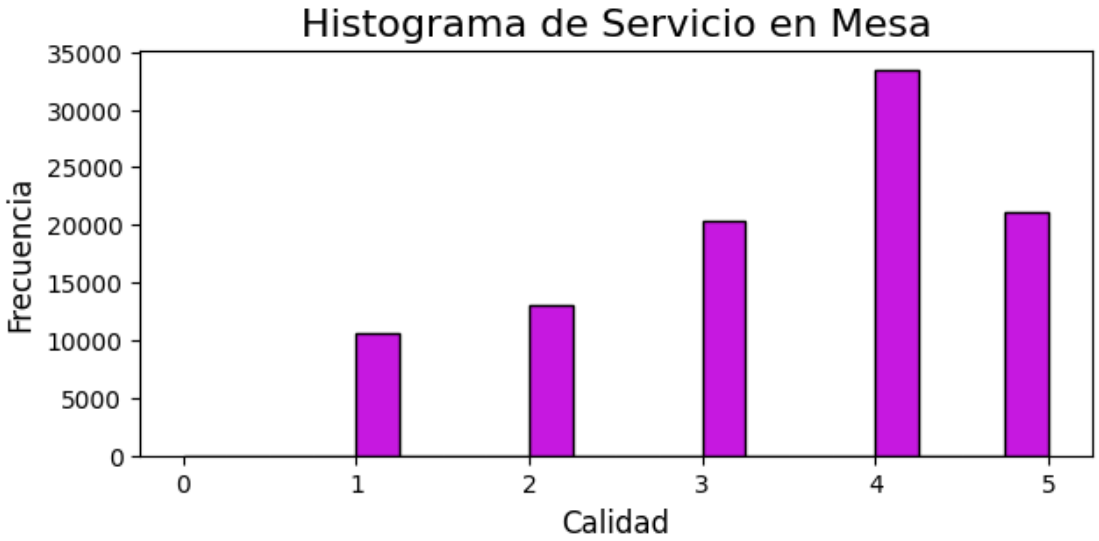


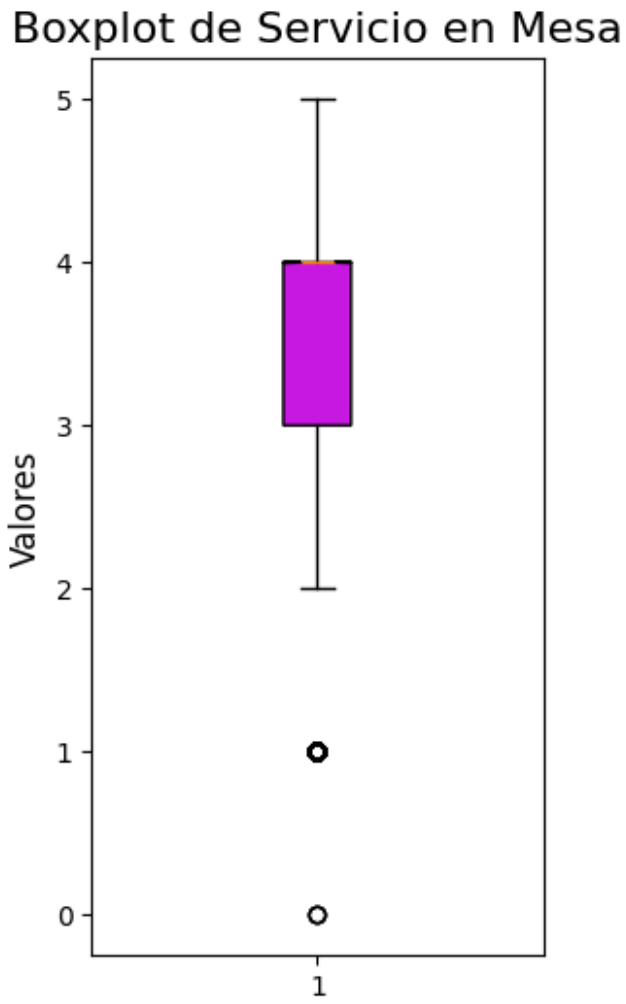
- **Columna Entretenimiento en Vuelo**



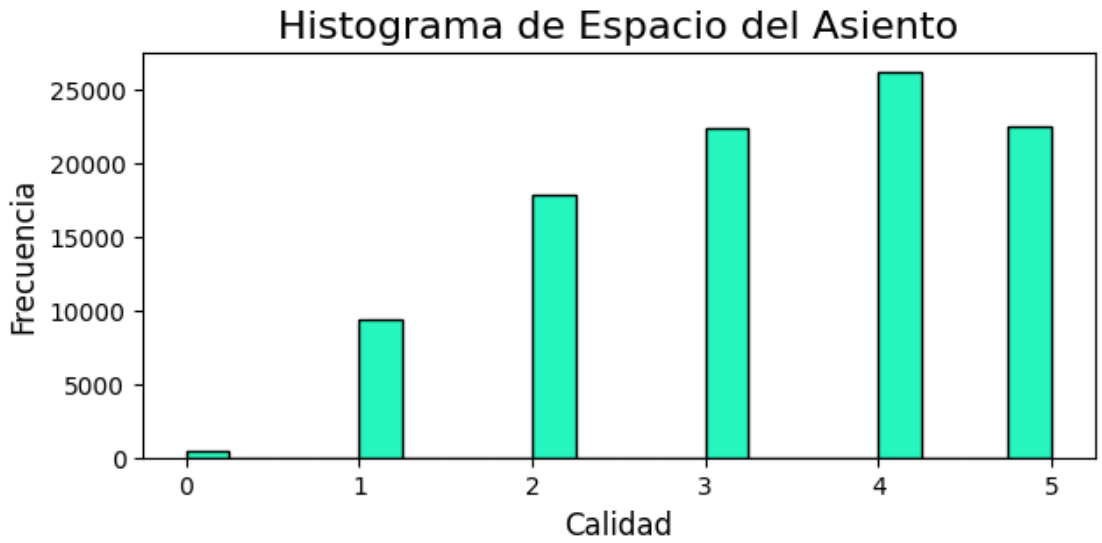


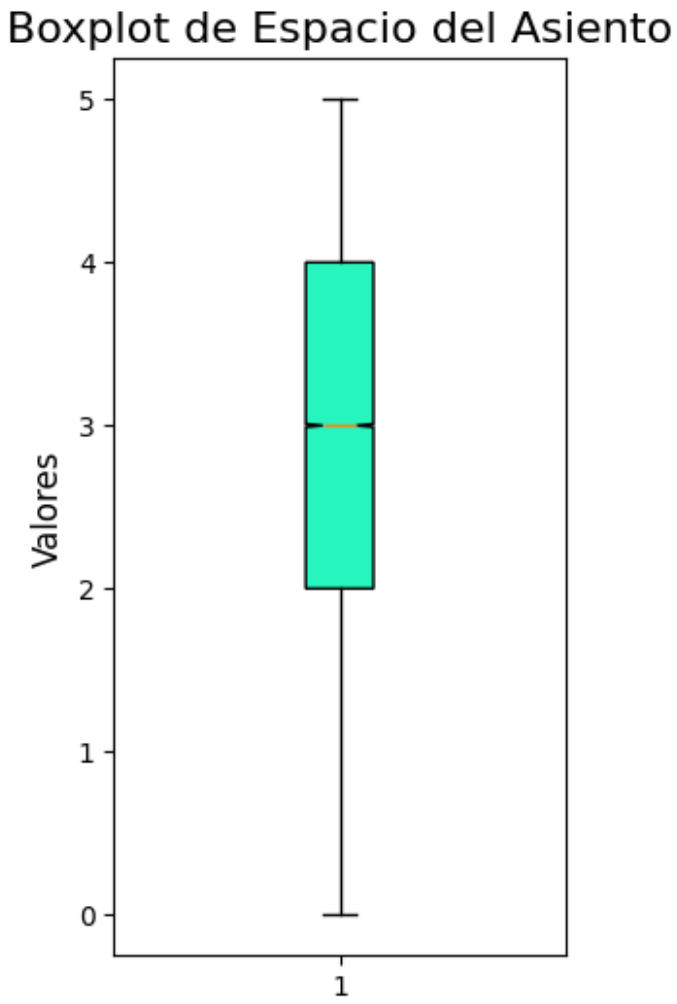
- **Columna Servicio en Mesa**



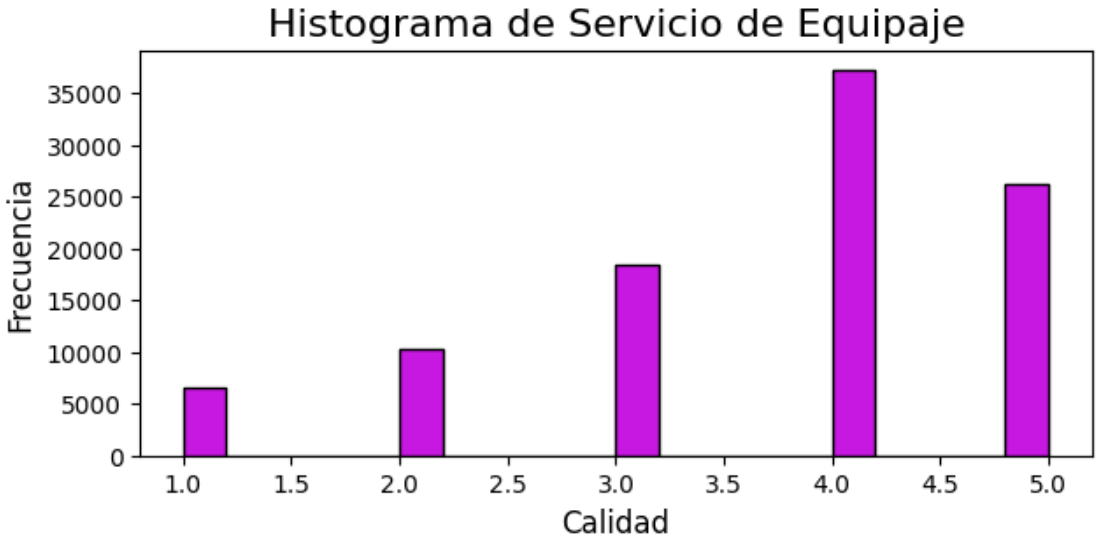


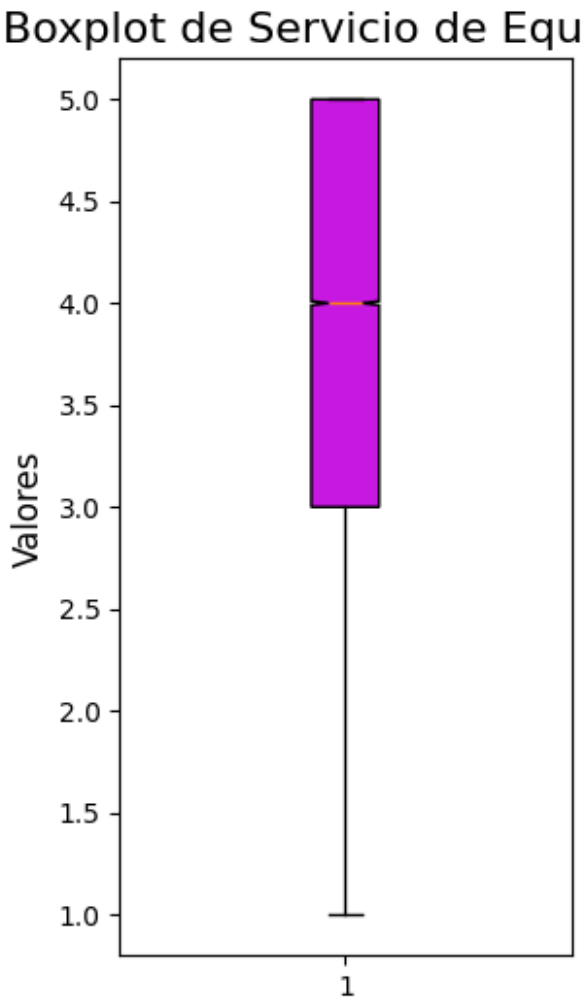
- **Columna Espacio del Asiento**



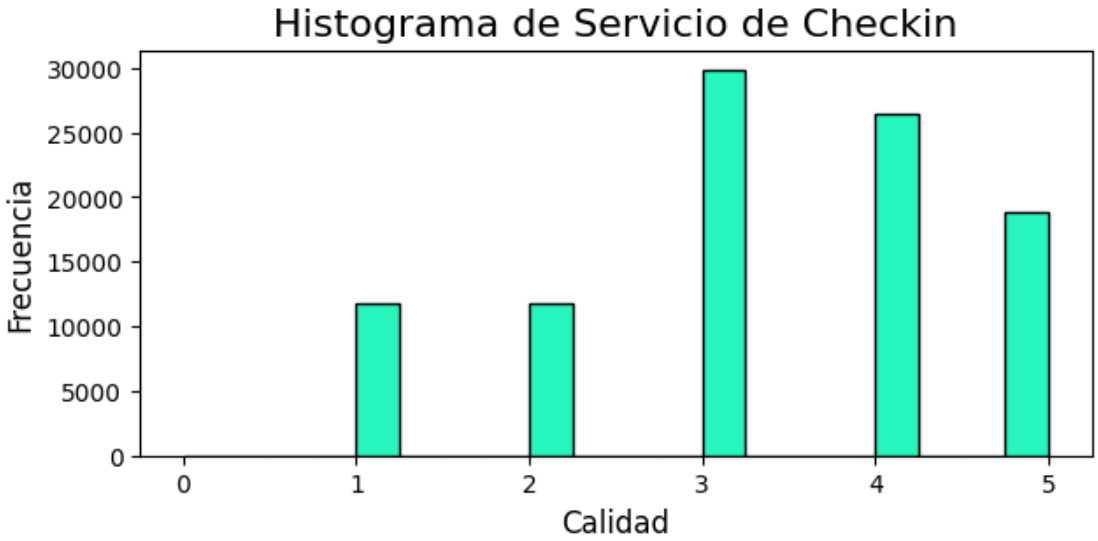


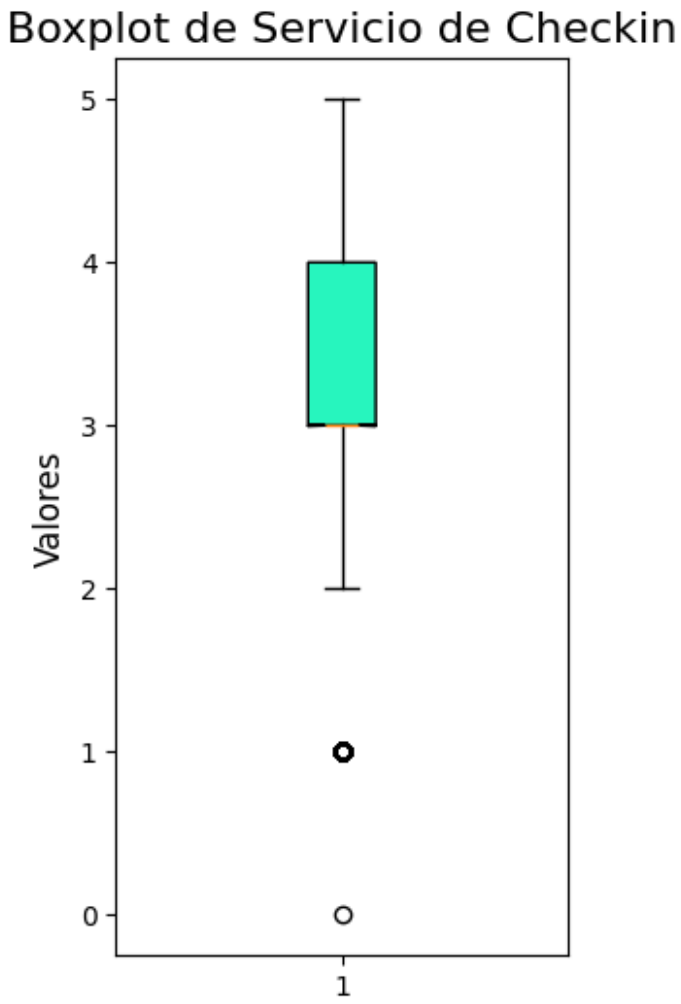
- **Columna Servicio de Equipaje**



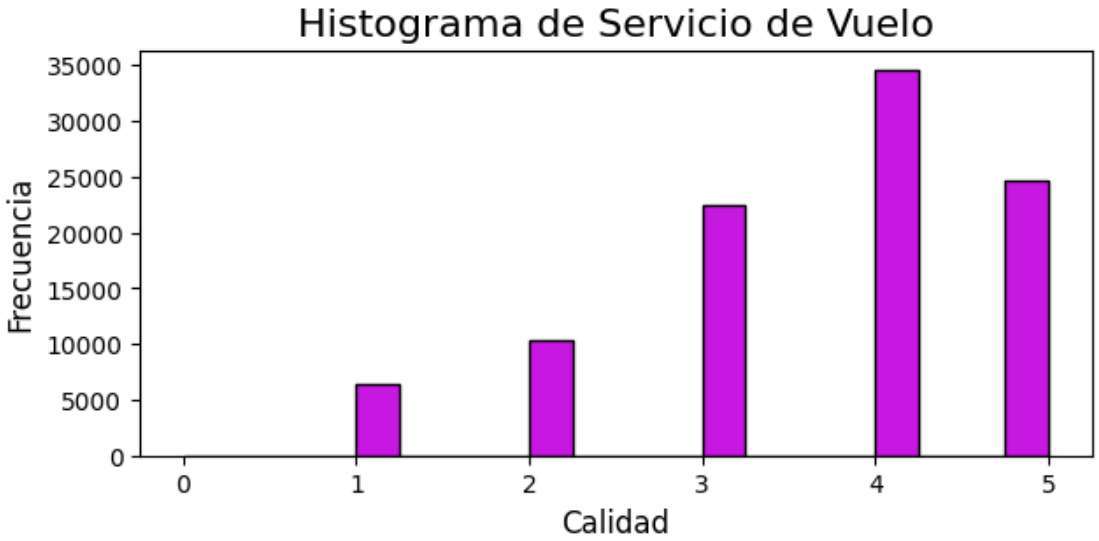


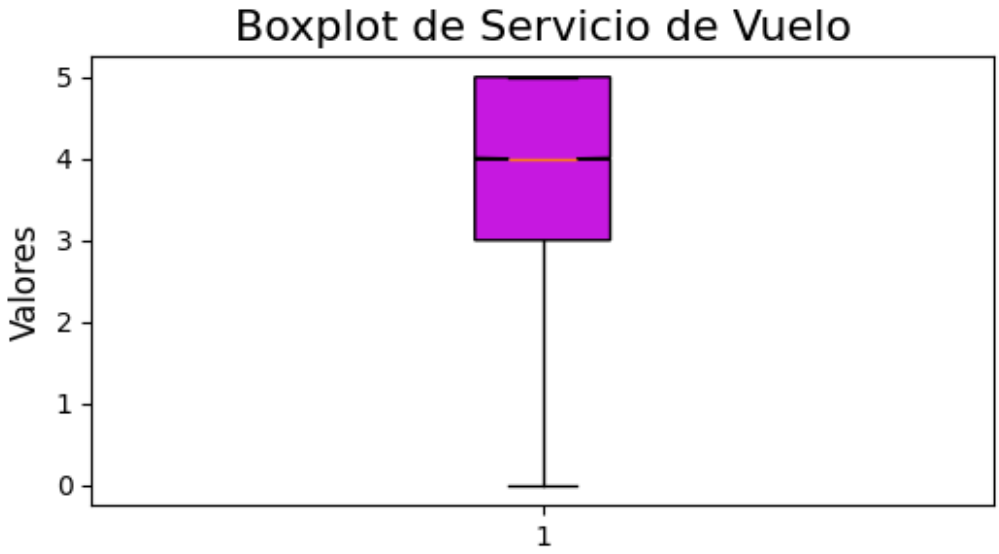
- **Columna Servicio de Checkin**



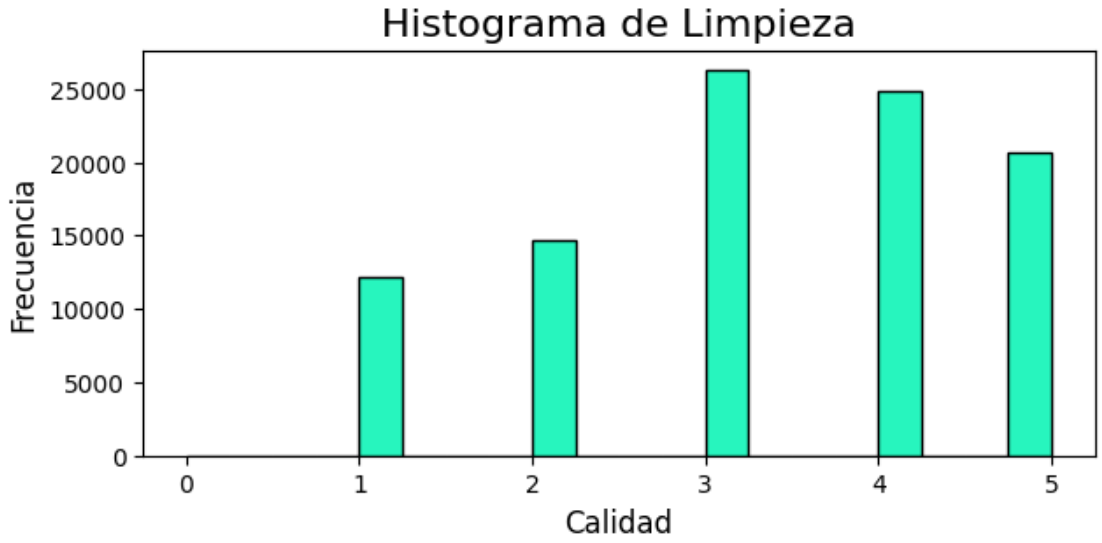


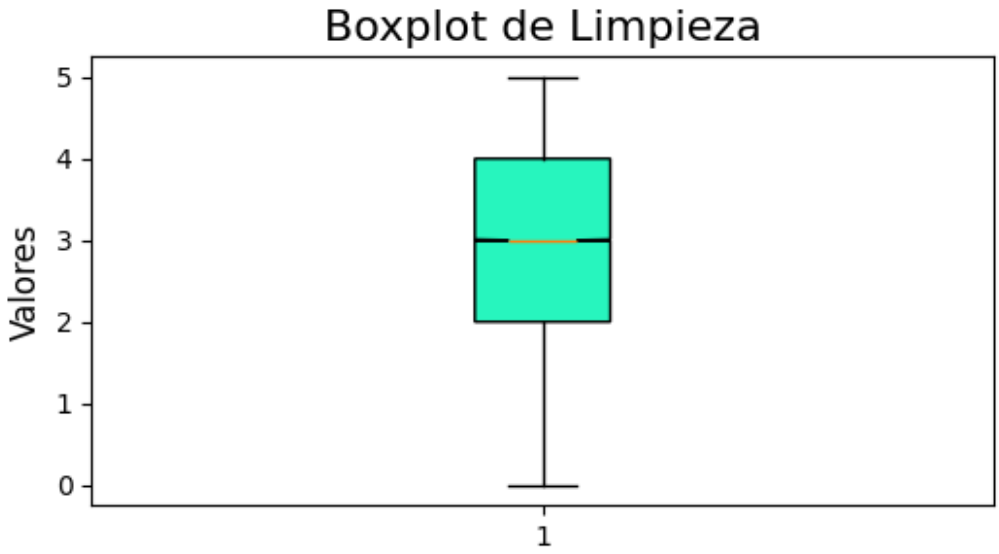
- **Columna Servicio de Vuelo**



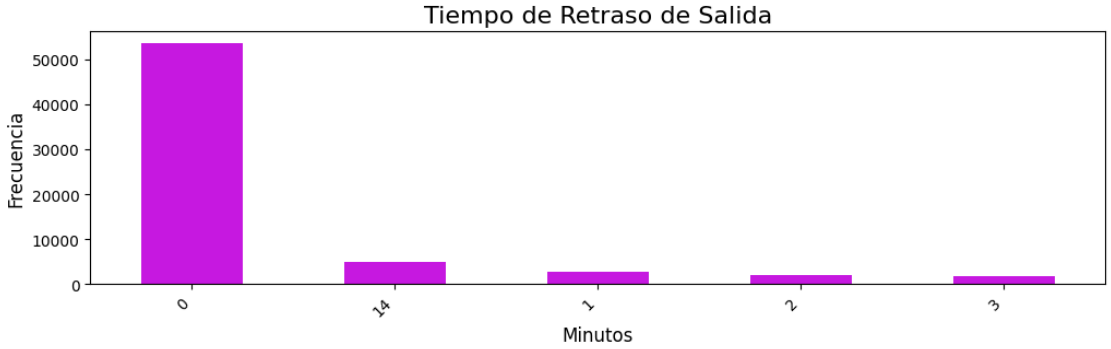


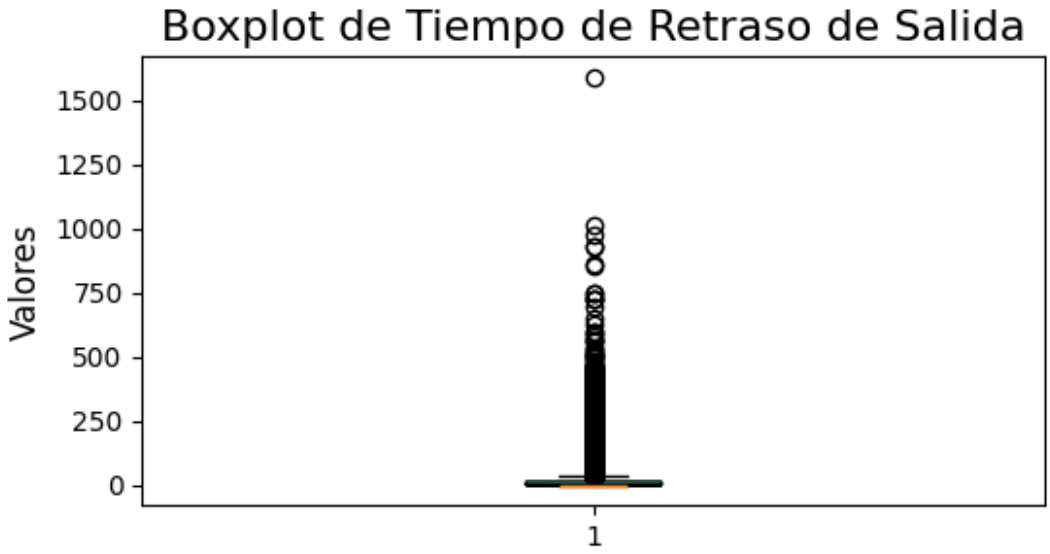
- **Columna Limpieza**



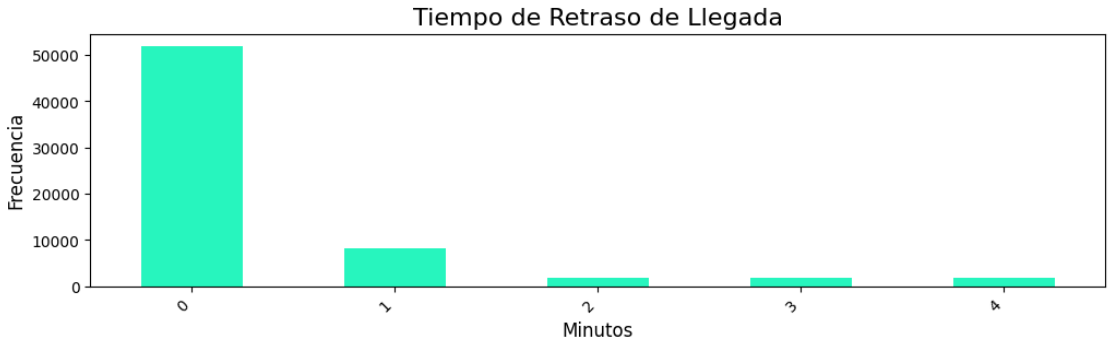


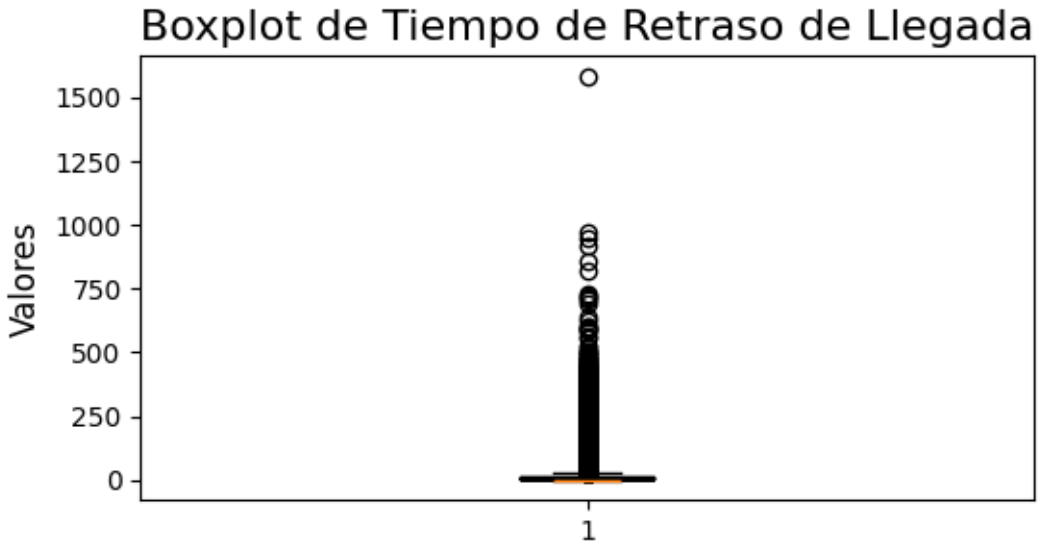
- **Columna Retraso de Salida**





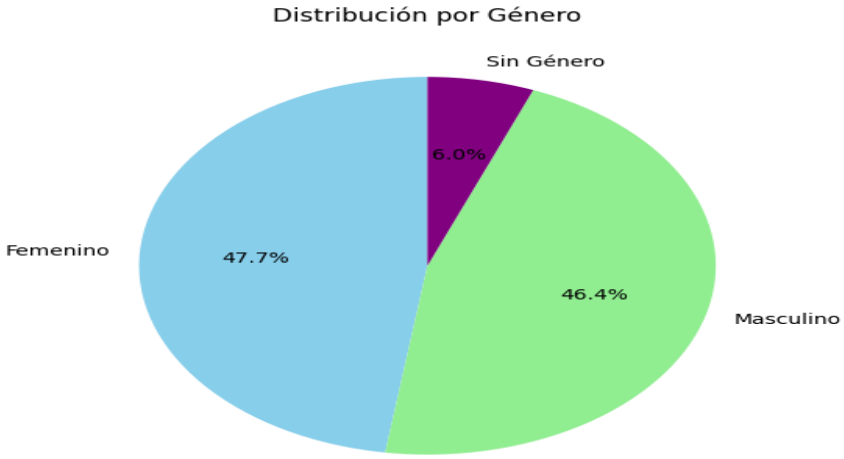
• **Columna Retraso de Llegada**





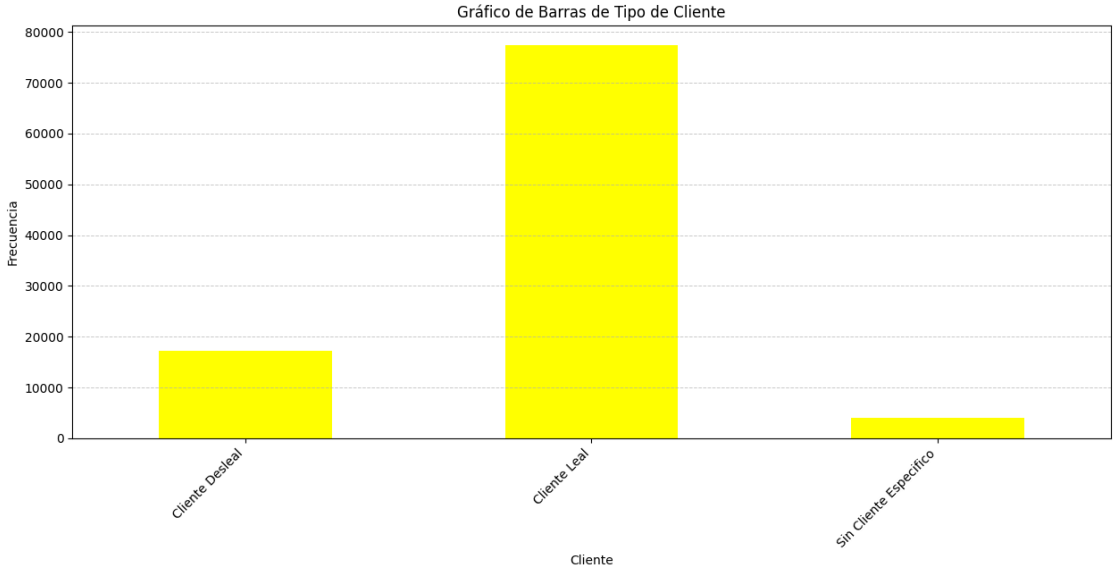
• Variables Categóricas

• **Columna Género**



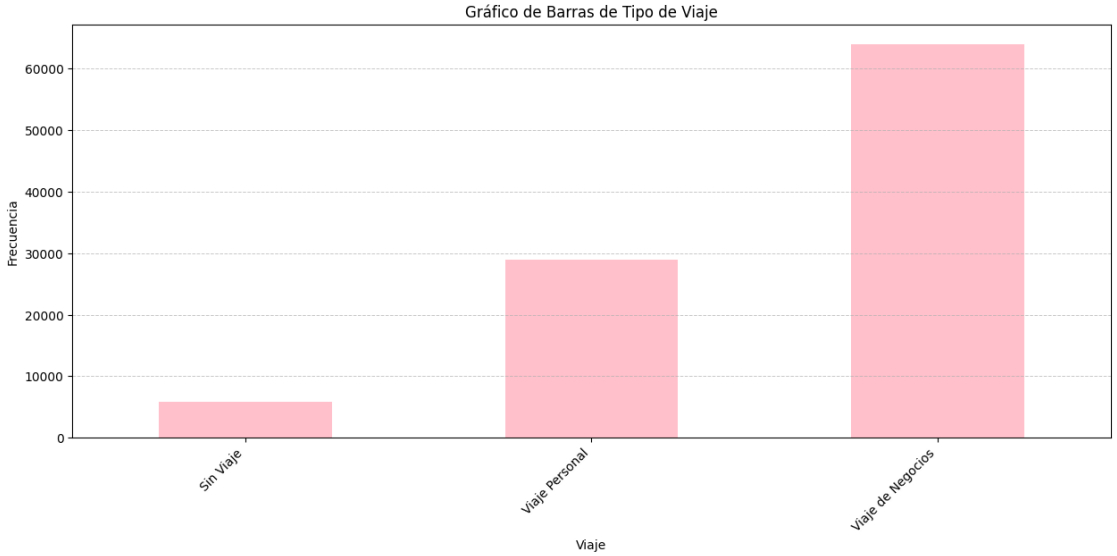
El gráfico del Género, muestra un balance entre las personas del género masculino y femenino, ya que tienen un porcentaje de distribución similar.

- **Columna Tipo de Cliente**



La mayoría de personas que viajan con una aerolínea son clientes leales que recurren nuevamente.

- **Columna Tipo de Viaje**

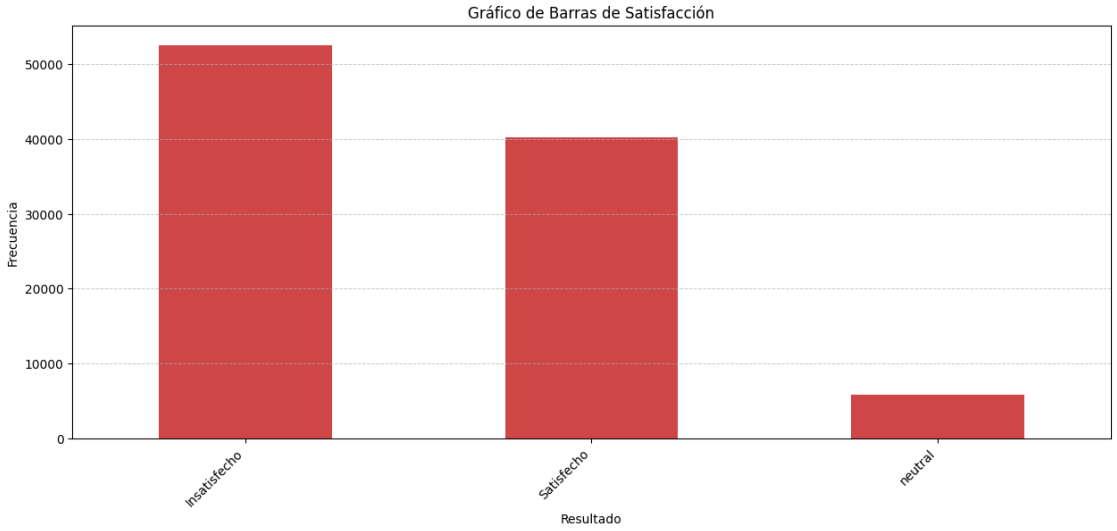


Los viajes que se realizan con mayor frecuencia son viajes de negocios.

• **Columna Clase**

Las clases Ejecutiva y Económica tienen el dominio sobre los vuelos de los clientes.

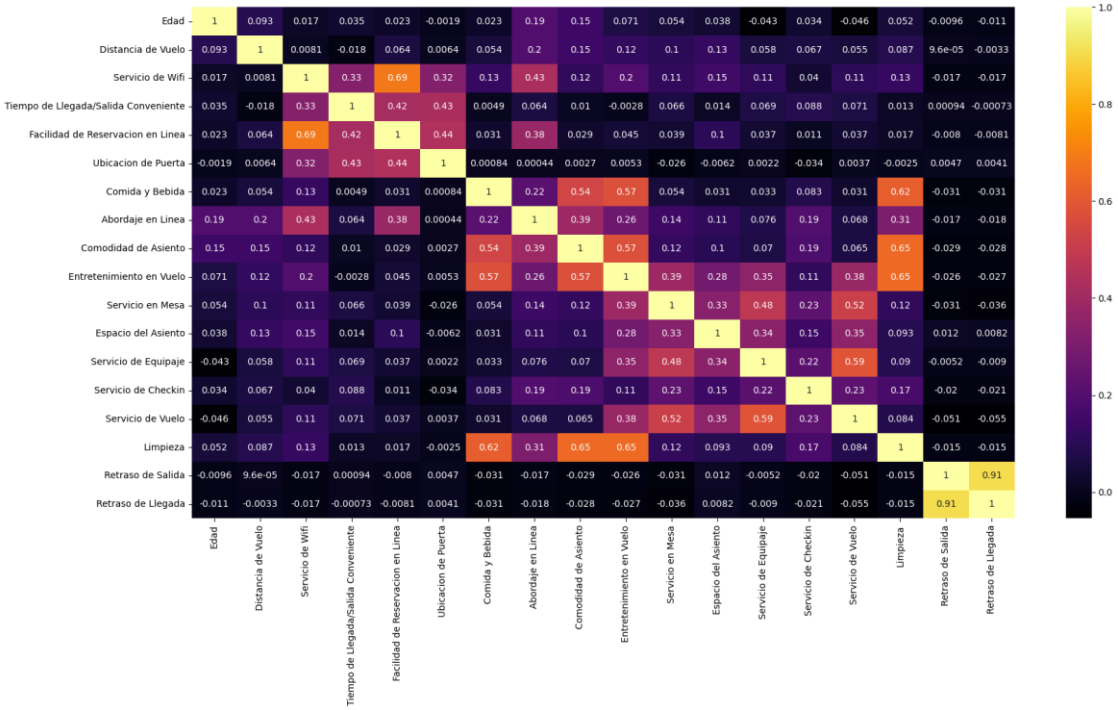
• **Columna Satisfacción**



La insatisfacción es un problema dominante, por lo tanto, se buscarán estrategias para resolverlo.

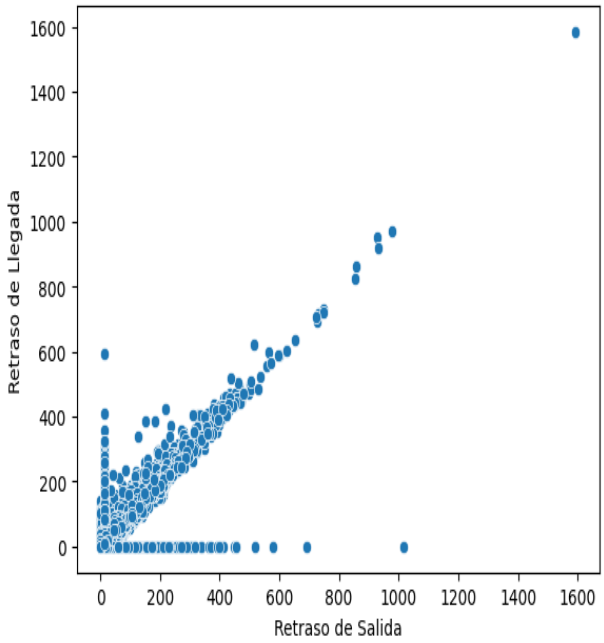
3. Correlación entre Variables

•Matriz de Correlación

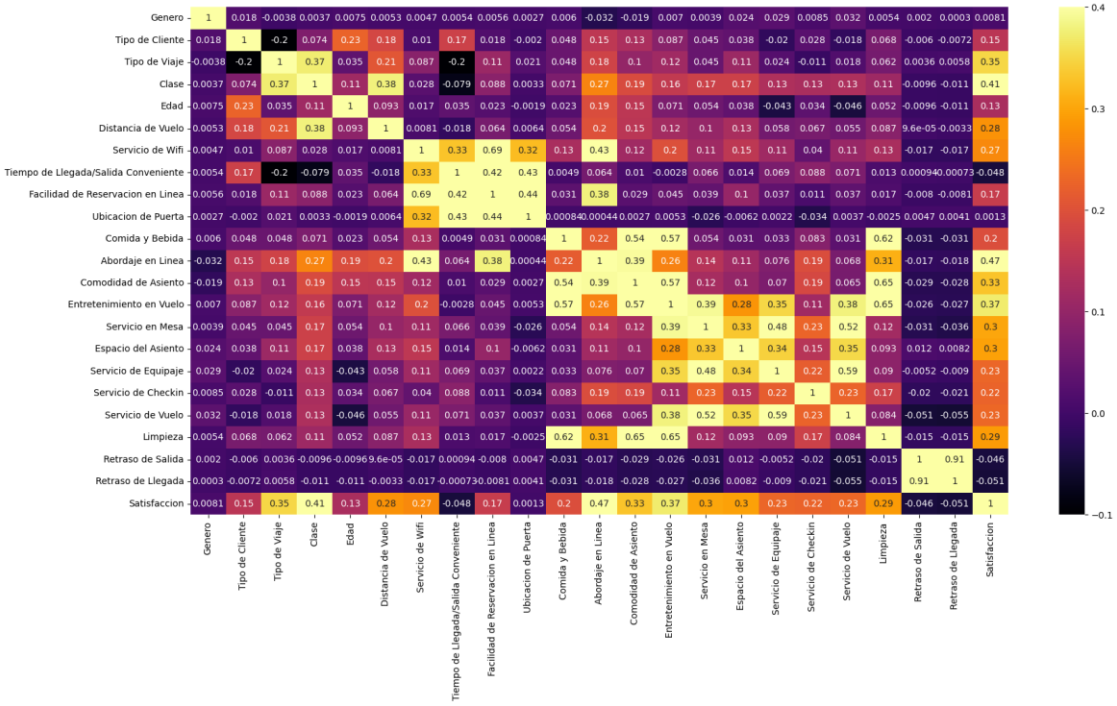


La variable 'Retraso de Salida' tiene una correlación de 0.91 con 'Retraso de Llegada', lo que sugiere una relación directa.

•**Parejas de Variables**

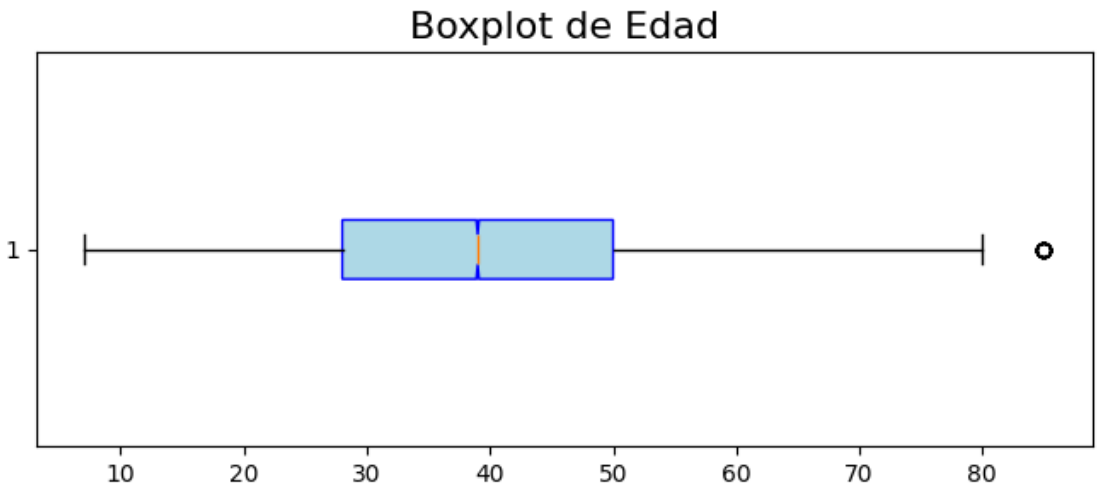


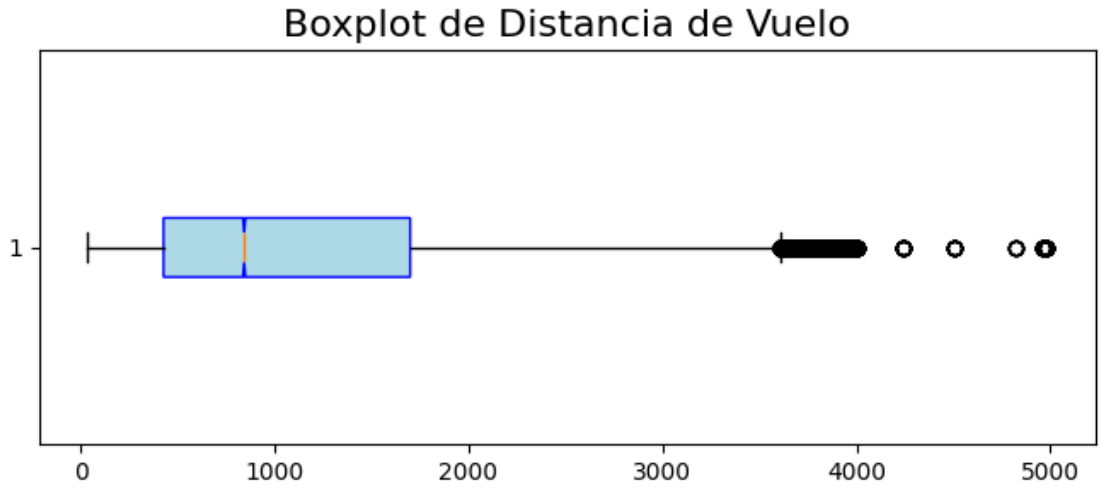
Matriz de Correlación con columnas categoricas a numéricas

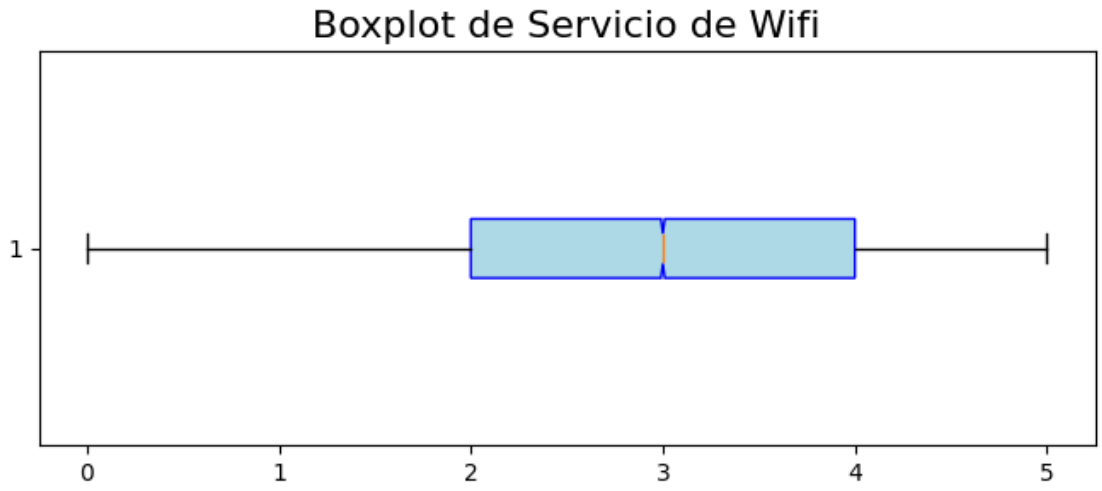


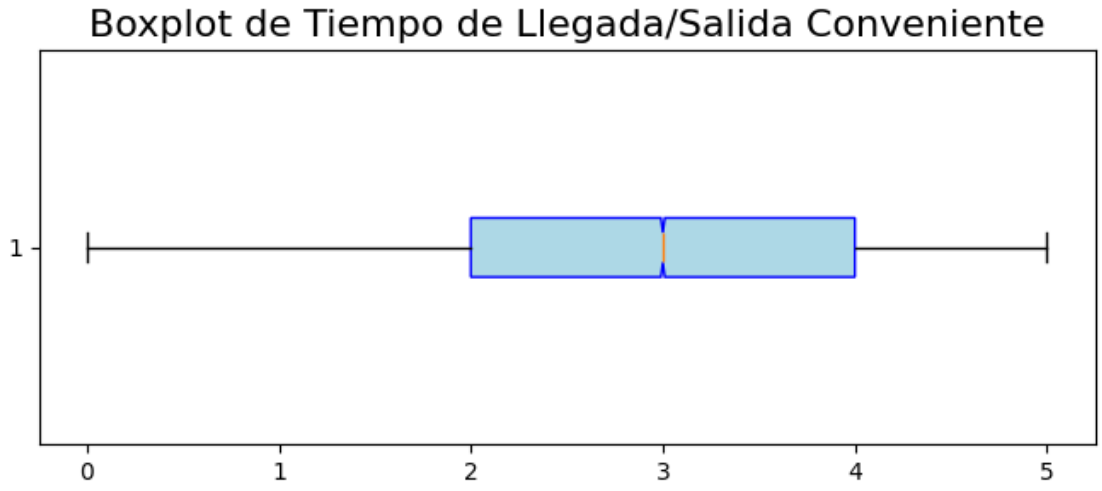
4. Análisis de Valores Atípicos (Outliers)

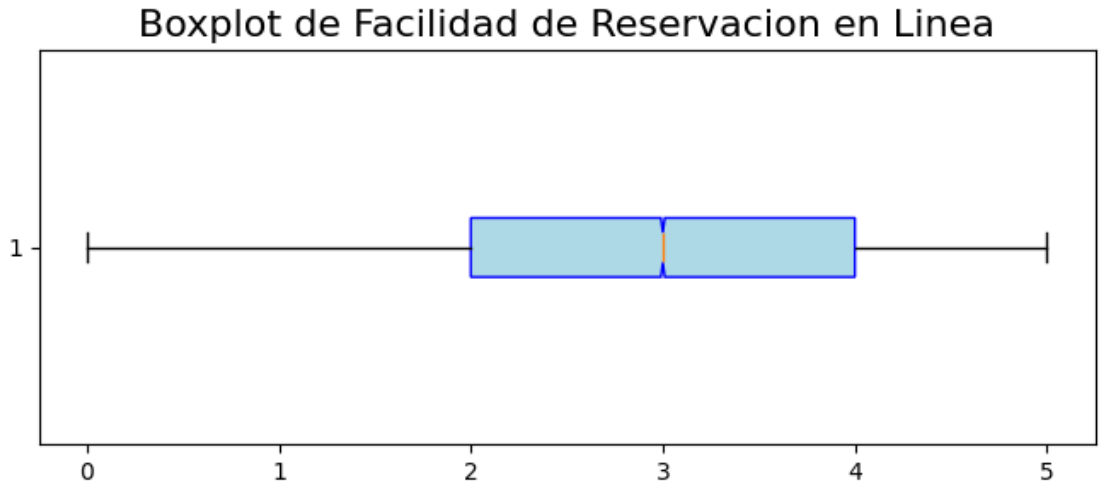
- **Identificación**

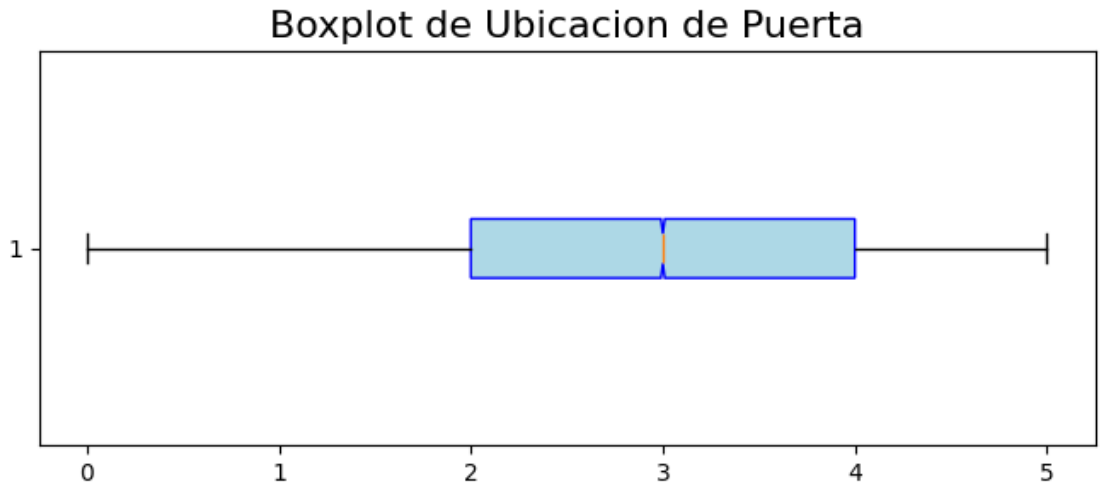


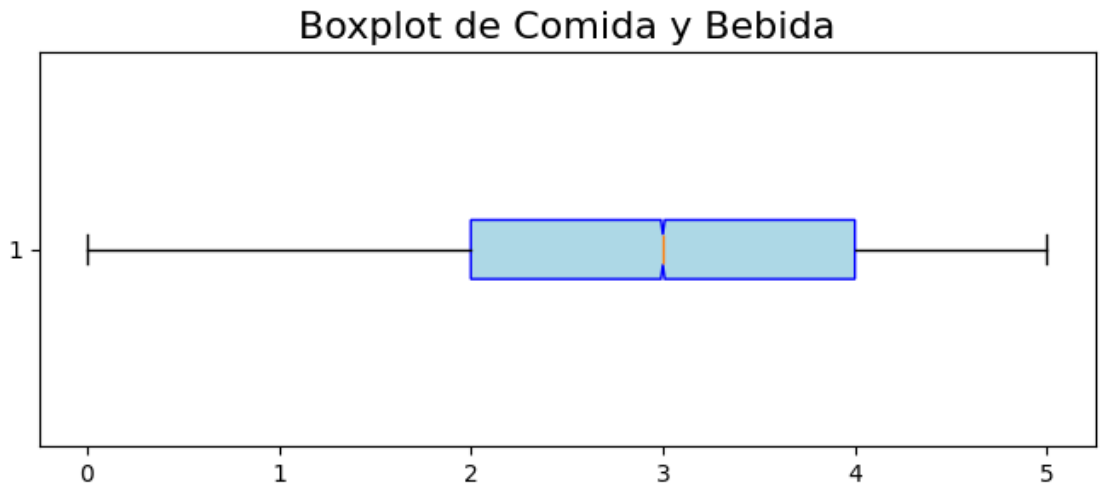

Boxplot de Distancia de Vuelo
1
0
1000
2000
3000
4000
5000

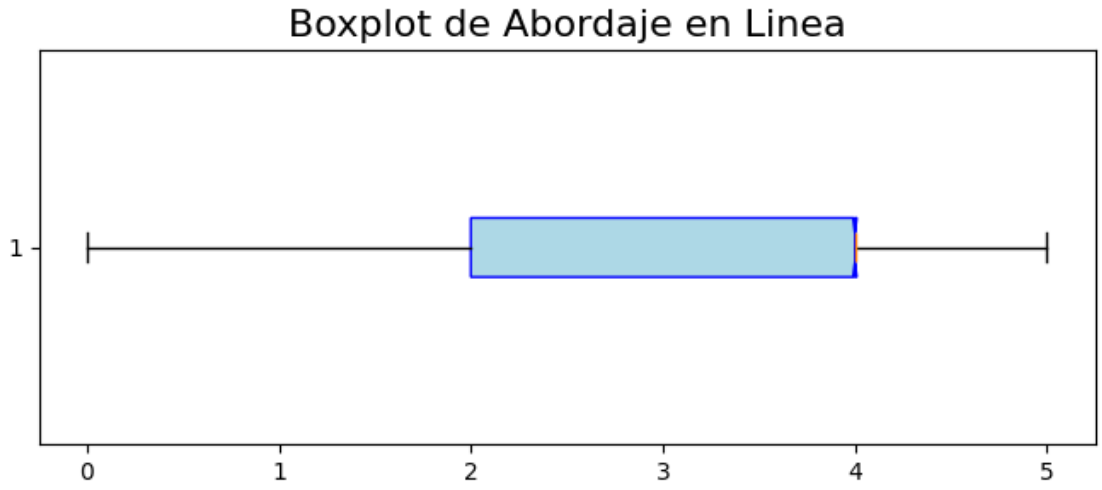

Boxplot de Servicio de Wifi
1
0
1
2
3
4
5

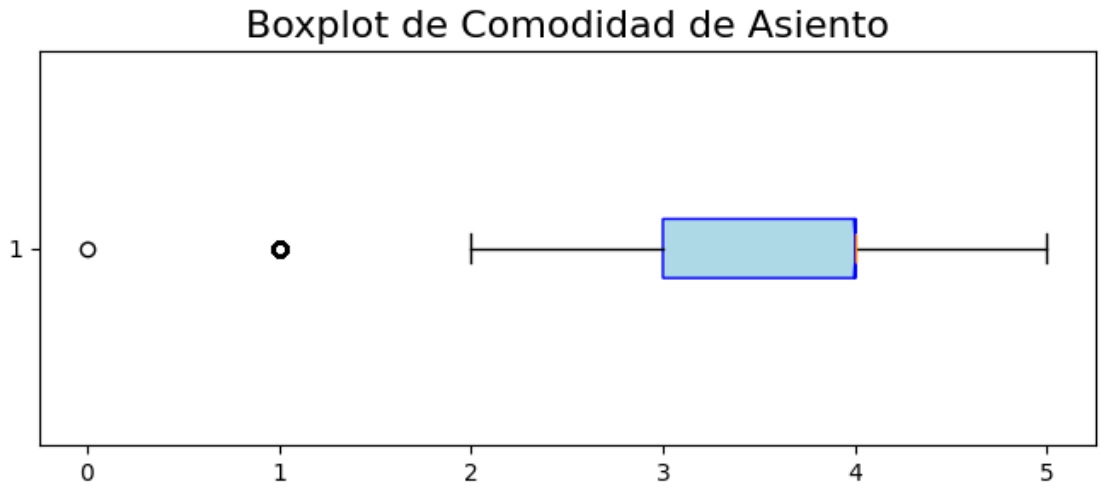

Boxplot de Tiempo de Llegada/Salida Conveniente
1
0
1
2
3
4
5

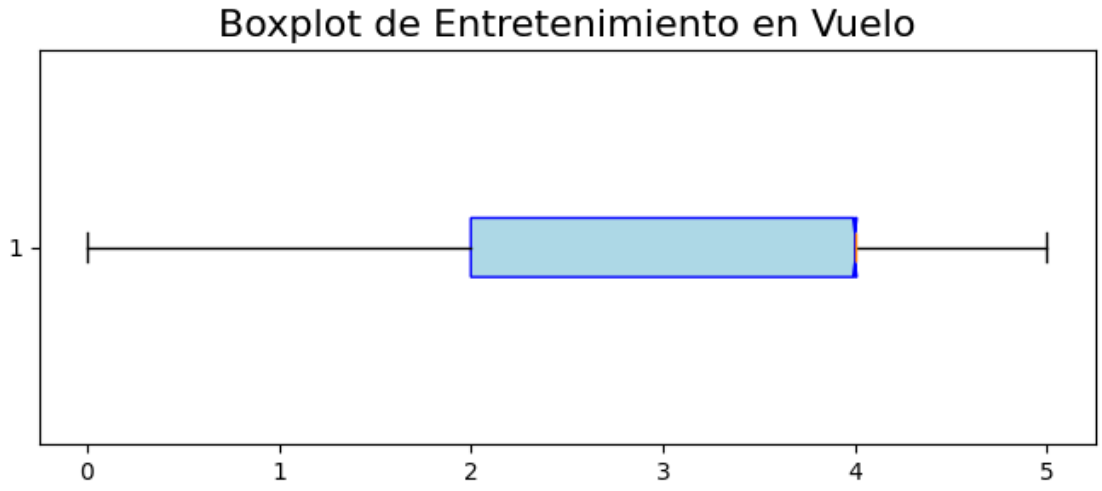

Boxplot de Facilidad de Reservacion en Linea
1
0
1
2
3
4
5

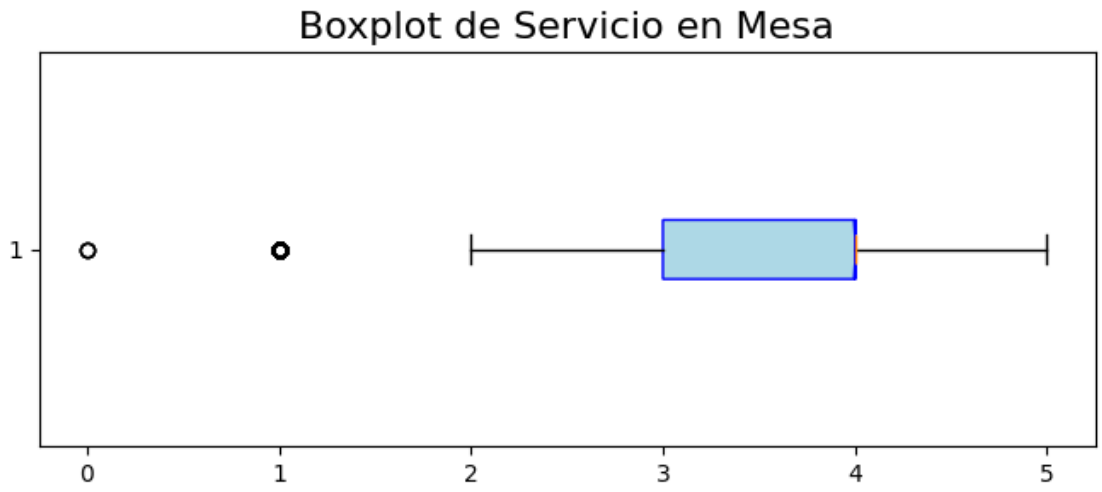

Boxplot de Ubicacion de Puerta
1
0
1
2
3
4
5

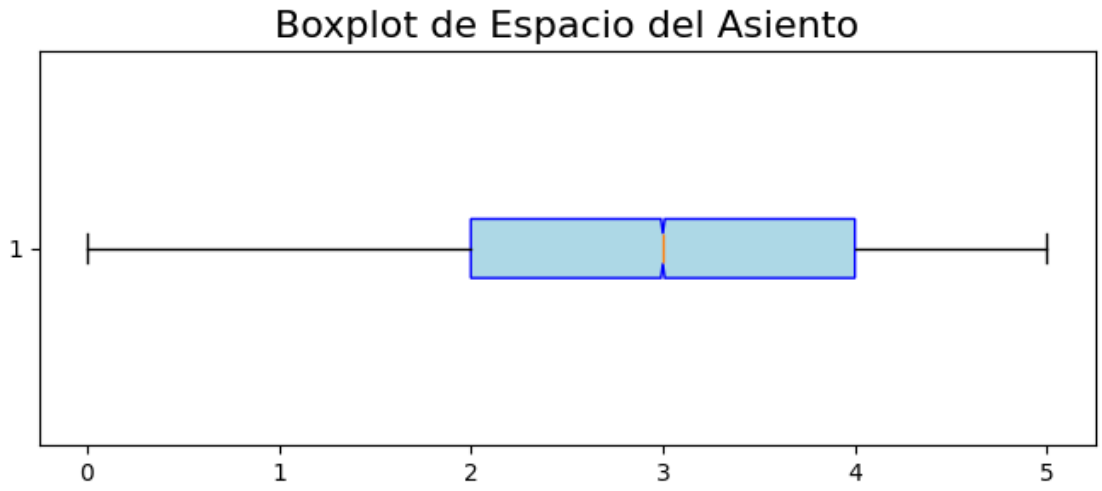

Boxplot de Comida y Bebida
1
0
1
2
3
4
5

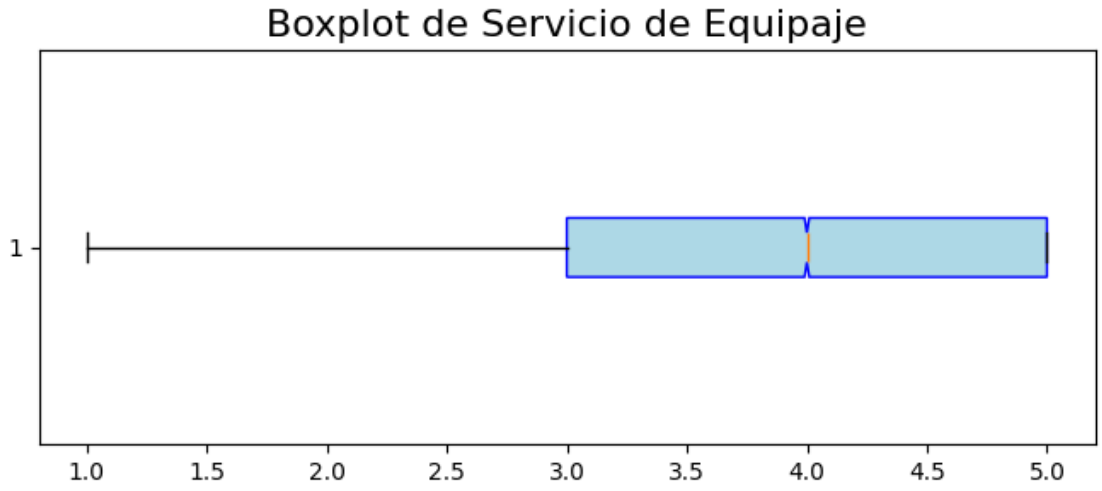

Boxplot de Abordaje en Linea
1
0
1
2
3
4
5

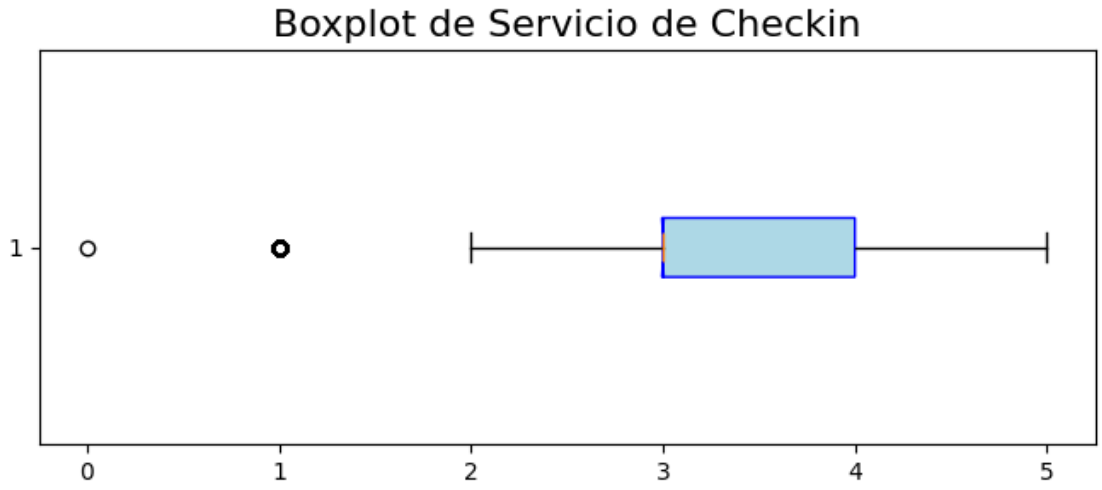

Boxplot de Comodidad de Asiento
1
0
1
2
3
4
5

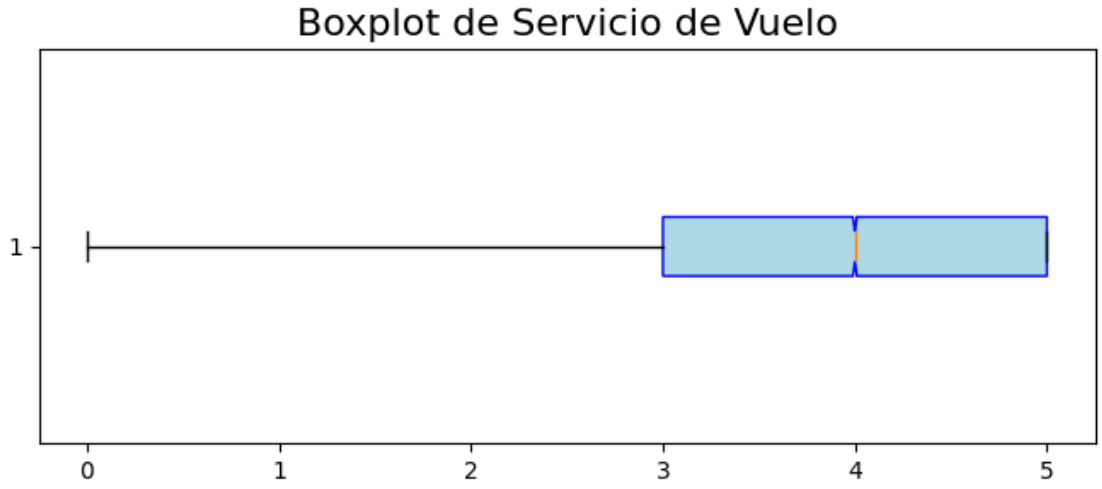

Boxplot de Entretenimiento en Vuelo
1
0
1
2
3
4
5

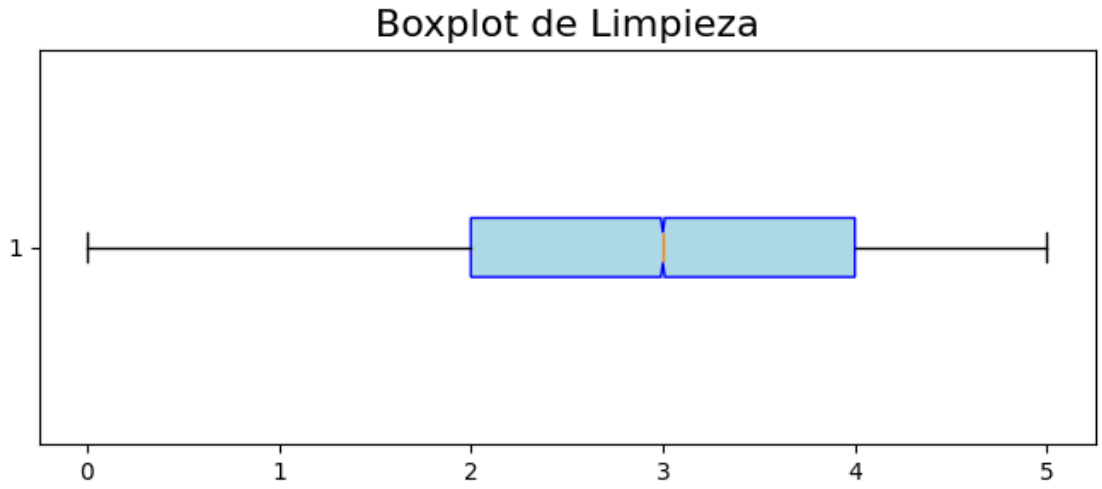

Boxplot de Servicio en Mesa
1
0
1
2
3
4
5

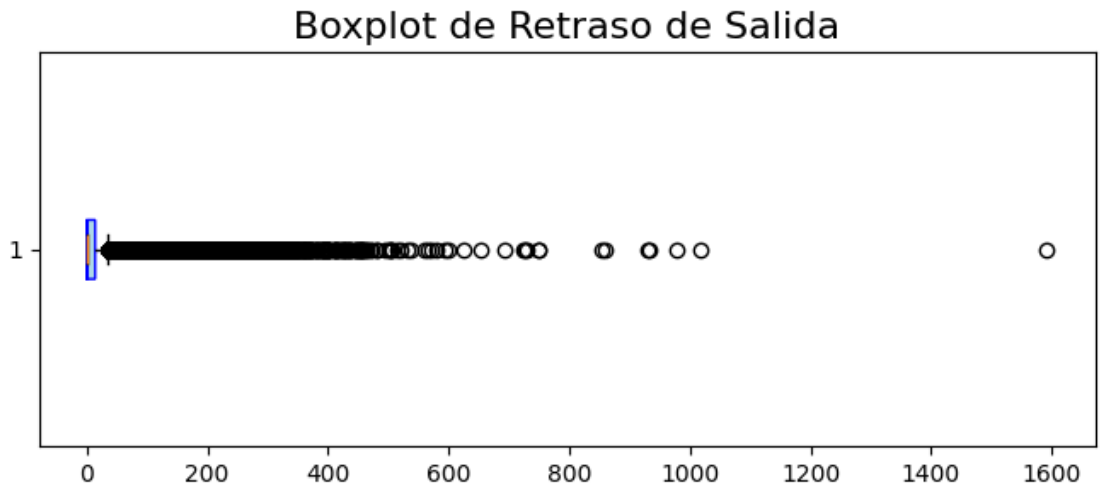

Boxplot de Espacio del Asiento
1
0
1
2
3
4
5

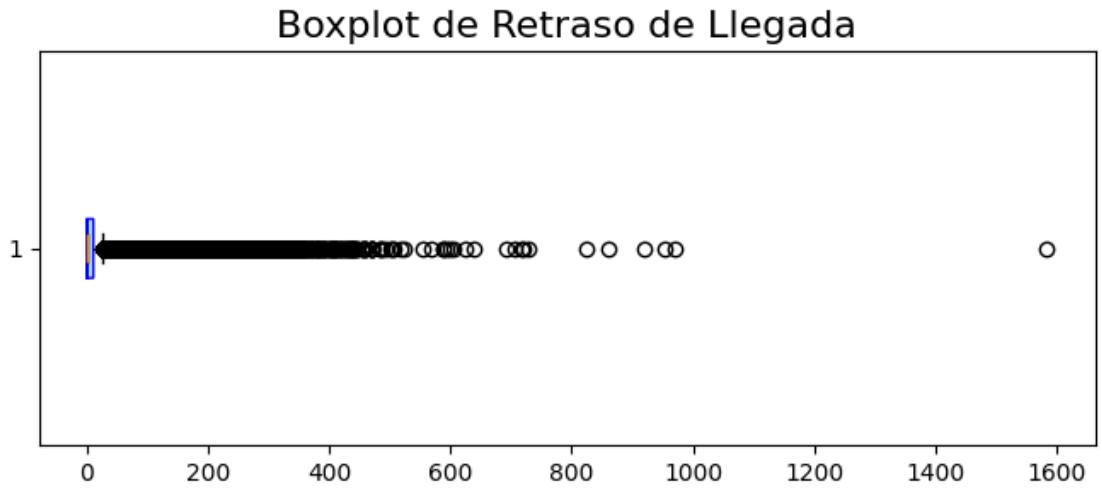

Boxplot de Servicio de Equipaje
1
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Boxplot de Servicio de Checkin
1
0
1
2
3
4
5


Boxplot de Servicio de Vuelo
1
0
1
2
3
4
5


Boxplot de Limpieza
1
0
1
2
3
4
5



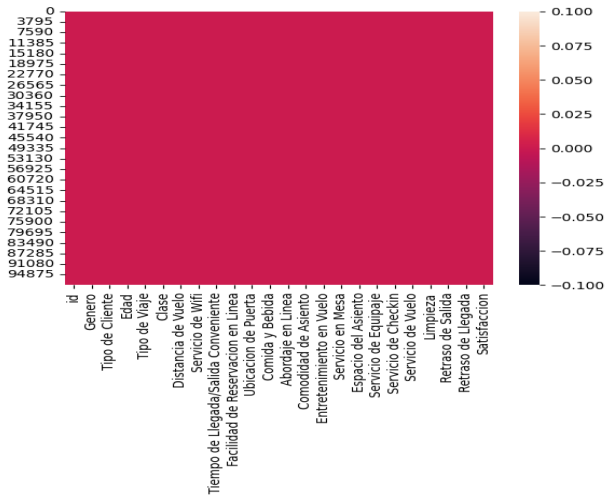


• Tratamiento

Para este caso en específico, los valores atípicos (outliers) se manejaron dependiendo de la variable a tratar, si no había tanto desvío en la media, los valores atípicos se reemplazaron por la media, ya que, si no hay demasiados outliers, la media es significativa para los datos, y no afecta de manera desproporcionada a los mismos, por ejemplo: La medida de satisfacción de los clientes durante su experiencia antes, durante, y después del vuelo, referente a los servicios brindados por las aerolíneas. Por otro lado, para variables donde la media no era significativa, por ejemplo, la distancia de vuelo o el retraso, los outliers se reemplazaron por la moda, ya que, al ser valores que no son recurrentes, lo que más importa en ese caso, es la mayoría de datos, no el promedio.

5. Análisis de Valores Faltantes

• **Identificación:**

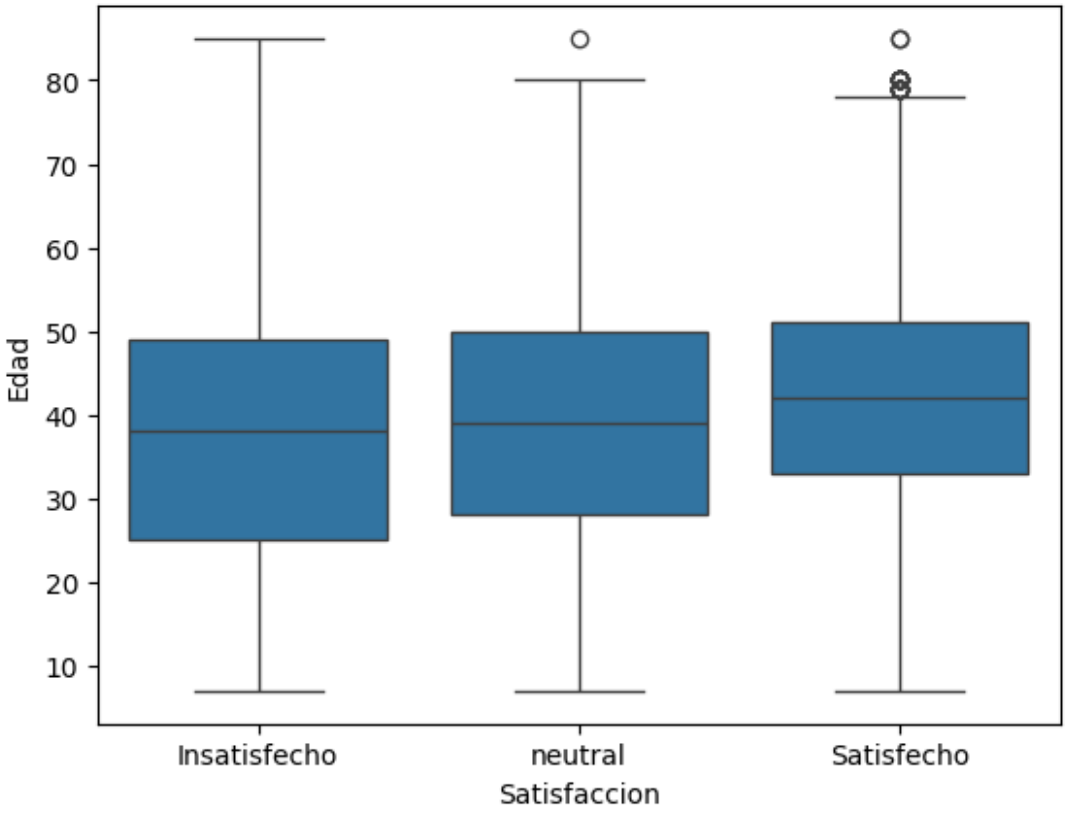


• **Estrategia de imputación:**

Para los valores faltantes, la imputación fue elegida conforme a cada variable, es decir, para las variables categóricas, se utilizo la moda, ya que es el dato que más se repite, y para las variables numéricas hubo dos estrategias. Si la media de los datos era muy dispersa, o sea que su distribución fuera, sesgada o bimodal, los datos faltantes se reemplazaron con la mediana, o en su defecto, con la moda. Por otro lado, para las variables que su distribución era normal, los valores faltantes se reemplazaron con la media. De este modo, se aseguró que los datos no fueran distorsionados con un promedio que no era el real, o que los valores atípicos pudieran cambiar los resultados.

6. Relación entre Variables Categóricas y Numéricas

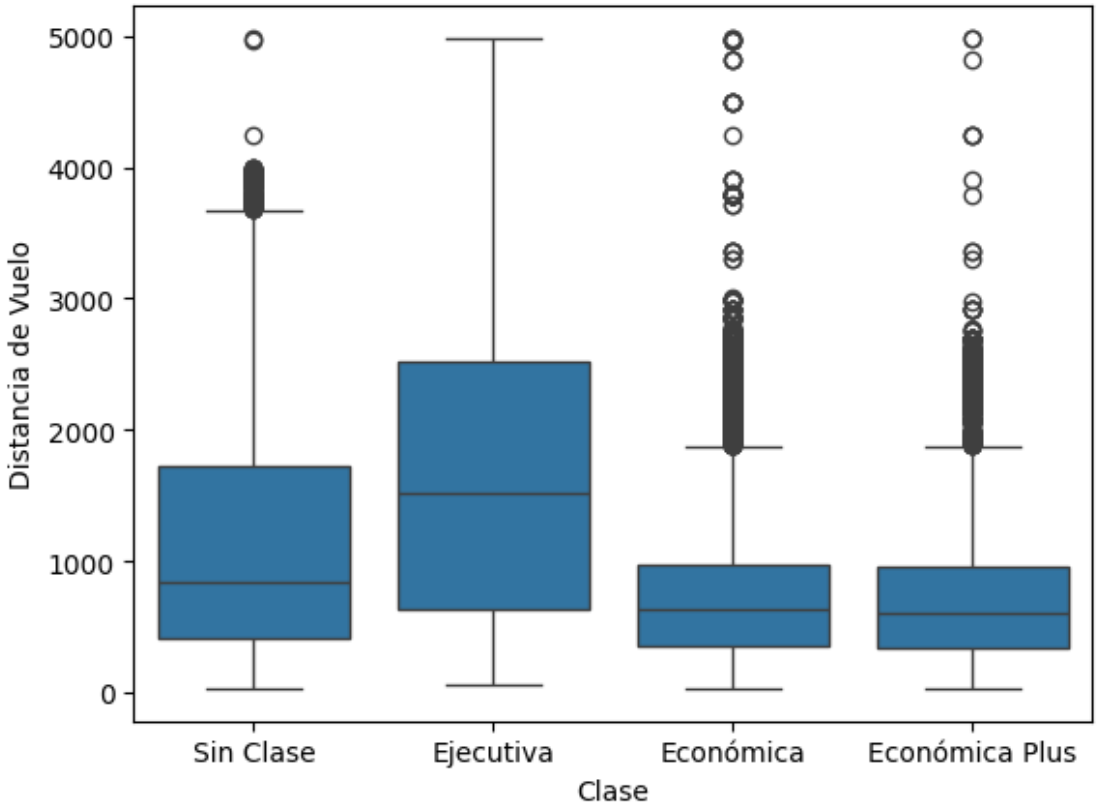
Se realizaron diagramas de caja (boxplots) para identificar patrones de comportamiento y diferencias significativas entre los grupos. A continuación, se describen los hallazgos más relevantes



Clientes Satisfechos: Tienen una mediana de edad ligeramente mayor (42 años) y un rango intercuartil más estrecho (33 a 51 años). Esto sugiere que los clientes de mediana edad tienden a estar más satisfechos.

Clientes Insatisfechos: Tienen una distribución de edad más amplia y una mediana menor (38 años). Los pasajeros más jóvenes muestran una mayor tendencia a la insatisfacción.

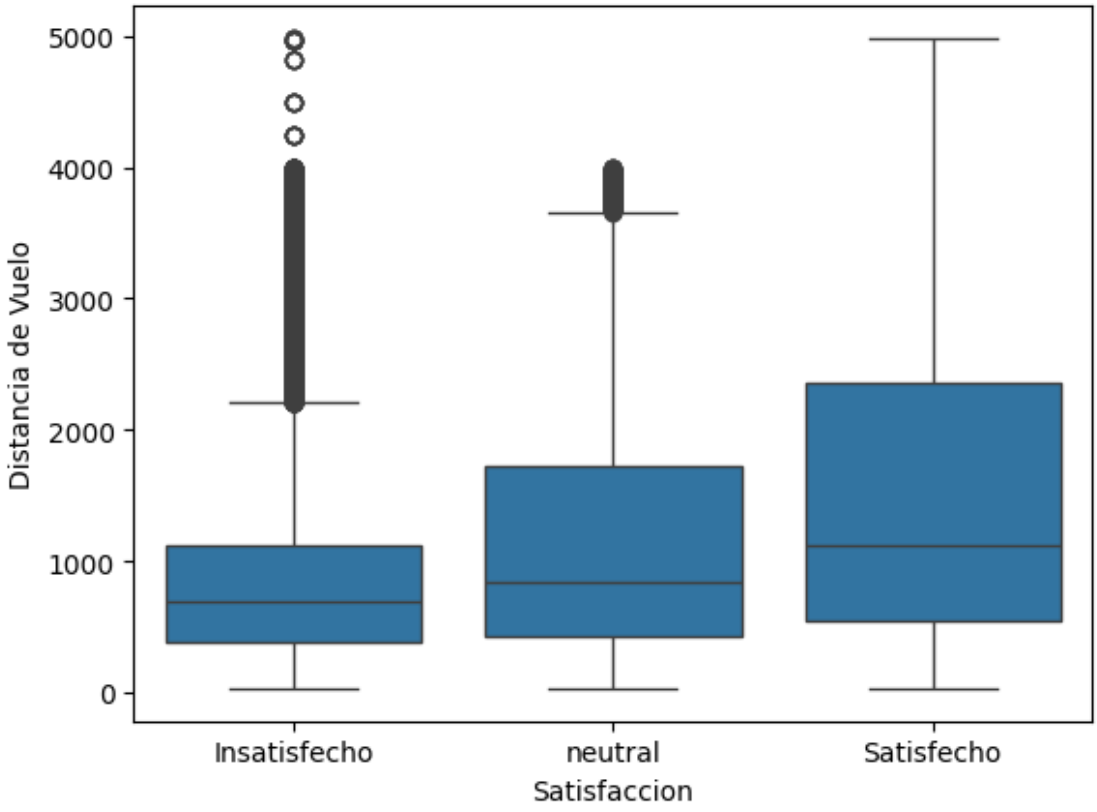
Conclusión: La edad es un factor discriminante moderado; la aerolínea parece satisfacer mejor las necesidades de pasajeros maduros que las de los jóvenes



Clase Ejecutiva: Presenta una mediana de distancia mucho mayor (1,524 km) y una gran dispersión, alcanzando los vuelos más largos (hasta casi 5,000 km)

Clase Económica: Se concentra en vuelos cortos, con una mediana de 628 km y el 75% de los vuelos por debajo de los 1,000 km.

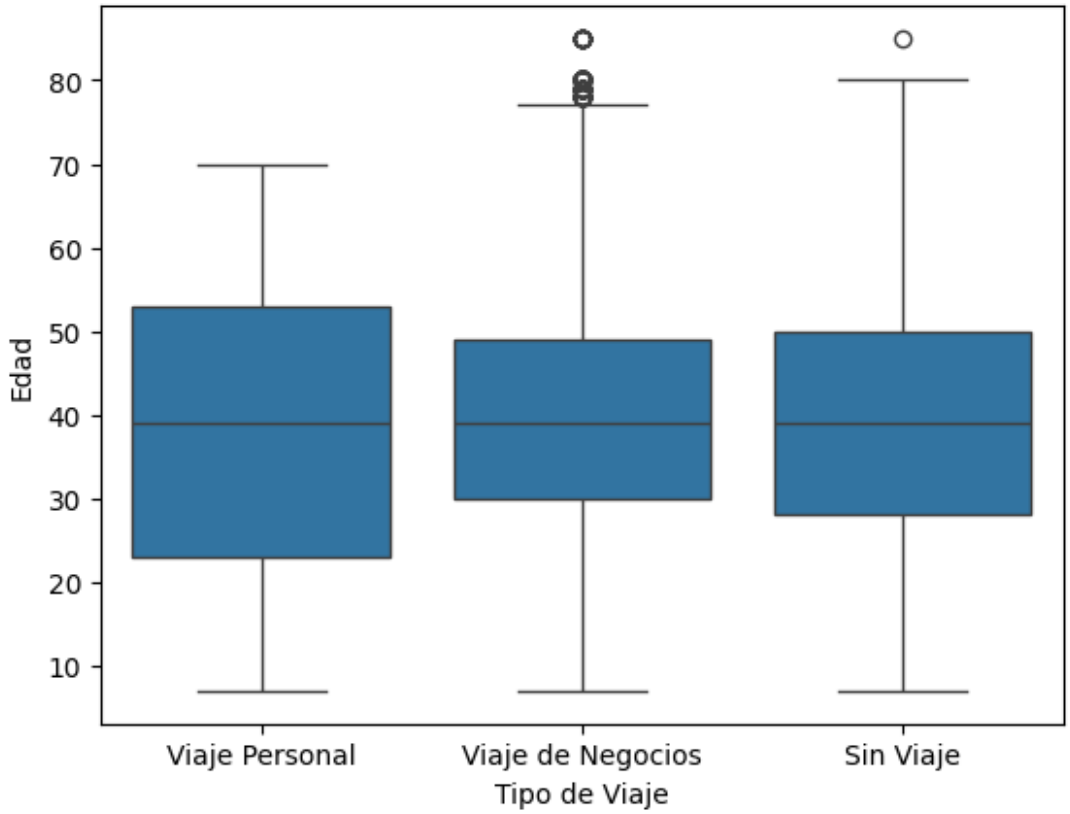
Conclusión: Existe una clara segmentación de mercado. Los vuelos de larga distancia son dominados por la clase Ejecutiva, lo que explica por qué servicios como la comodidad del asiento y el entretenimiento son críticos en este segmento.



Clientes Satisfechos: Tienden a volar distancias más largas (mediana de 1,127 km).

Clientes Insatisfechos: Se concentran predominantemente en vuelos cortos (mediana de 700 km).

Contrario a la intuición (donde un viaje largo podría ser más cansado), los pasajeros de vuelos largos están más satisfechos. Esto probablemente se debe a que los vuelos largos suelen ser en mejores aviones o en Clase Ejecutiva, mientras que los vuelos cortos regionales pueden tener aviones más incómodos o servicios reducidos.



Viaje de Negocios: Muestra una concentración muy fuerte en el rango de edad laboral.

Viaje Personal: Tiene una distribución mucho más dispersa, abarcando desde niños y jóvenes hasta adultos mayores, con una mediana similar (39 años) pero mayor varianza.

Conclusión: Los servicios para "Viaje de Negocios" pueden estandarizarse para un adulto promedio, mientras que los "Viajes Personales" requieren flexibilidad para atender a rangos de edad muy diversos (niños, ancianos)

7. Observaciones y Hallazgos Importantes

• Identificar variable objetivo y variables influyentes:

Variable Objetivo (Target): La variable objetivo es Satisfaccion. Se trata de una variable categórica con tres niveles: "Insatisfecho", "neutral" y "Satisfecho". El objetivo del modelo será clasificar a cada pasajero en uno de estos grupos.

Variables más Influyentes: Según la matriz de correlación, las variables que presentan mayor influencia positiva sobre la satisfacción son:

•Abordaje en Línea: Con un coeficiente de r = 0.50, es el predictor más fuerte. Esto indica que la experiencia digital inicial es determinante.

•Clase del Vuelo (Clase): Con un r = 0.45, confirma que a mayor categoría (Ejecutiva vs. Económica), mayor es la probabilidad de satisfacción.

•Entretenimiento en Vuelo: Muestra una correlación relevante de r = 0.37, siendo un factor clave en la experiencia a bordo.

- **Resumir hallazgos clave:**

• Patrones y Relaciones Interesantes:

La Brecha Digital: Se observó que los servicios tecnológicos (Wifi, Abordaje en línea) tienen una correlación mucho más fuerte con la felicidad del cliente que los servicios logísticos tradicionales (como la limpieza o la comida).

La Paradoja de los Retrasos: Sorprendentemente, los retrasos (Retraso de Salida y Llegada) tienen una correlación casi nula (r ≈ -0.05) con la satisfacción. Esto sugiere que los clientes toleran la impuntualidad si el servicio a bordo es excelente.

• Outliers Relevantes

Retrasos Extremos: Se detectaron casos excepcionales con retrasos masivos. Mientras el promedio de retraso es de 14 minutos, existen vuelos con retrasos de hasta 1,592 minutos (26 horas). Aunque son casos raros (outliers estadísticos por encima de 35 minutos), representan experiencias críticas de servicio.

Distancias de Vuelo: Se identificó un grupo de vuelos de ultra-larga distancia (superiores a 3,600 km y llegando hasta 4,983 km) que se comportan de manera distinta, siendo casi exclusivamente de Clase Ejecutiva.

Edad: No se encontraron outliers significativos en la edad (máximo 85 años), lo cual indica una distribución demográfica normal y limpia.

•Variables desbalanceadas

La variable objetivo presenta un desbalance significativo en la clase "neutral", la cual representa menos del 6% de los datos totales, mientras que "Insatisfecho" y "Satisfecho" concentran más del 90%.

• Correlaciones fuertes o inesperadas.

Se detectó una correlación extremadamente alta de r = 0.91 entre Retraso de Salida y Retraso de Llegada. Esto indica que ambas variables aportan prácticamente la misma información (redundancia).

• Problemas de datos (faltantes, duplicados).

Se identificó que la columna id no aporta valor predictivo y genera ruido en el análisis.

Se detectó la necesidad de aplicar una codificación ordinal manual en variables como Satisfaccion y Clase, ya que la codificación automática alfabética distorsionaba las relaciones reales.

• Implicaciones para el modelo:

1. Eliminación de Redundancia: Dado que Retraso de Salida y Retraso de Llegada están altamente correlacionadas (r=0.91), se eliminará la columna Retraso de Salida para evitar problemas de multicolinealidad y reducir el ruido en el modelo.

2. Para Clase y Satisfaccion, NO se usará LabelEncoder estándar (alfabético). Se implementará un Mapeo Ordinal Manual (0, 1, 2, 3) para preservar la jerarquía de valor detectada en el EDA.

4. Modelo de Machine Learning

1. Descripción del Modelo

• **Nombre del Modelo:** Random Forest Classifier

• **Tipo de Aprendizaje:** Supervisado (Supervised Learning)

• **Tipo de Problema:** Clasificación Multiclase

2. Justificación

Se seleccionó el algoritmo Random Forest Classifier como modelo principal para este proyecto. Esta decisión se fundamenta en los siguientes criterios técnicos y de negocio:

•**Tipo de Variable Objetivo:** La variable a predecir, Satisfaccion, es categórica con tres clases ("Insatisfecho", "neutral", "Satisfecho"). Random Forest es un algoritmo de aprendizaje supervisado diseñado nativamente para problemas de clasificación multiclase, lo que le permite manejar estas etiquetas de forma eficiente sin requerir transformaciones complejas en la salida, a diferencia de modelos limitados a clasificación binaria.

•**Tamaño del Dataset:** Con un volumen de 98,665 registros, contamos con una cantidad de datos suficiente para entrenar un modelo robusto. Random Forest es ideal para este tamaño de muestra porque, a través de la técnica de bagging (bootstrap aggregating), reduce significativamente el riesgo de sobreajuste (overfitting) que suelen sufrir los árboles de decisión individuales, garantizando que el modelo generalice bien ante nuevos pasajeros.

•**La interpretabilidad o la precisión buscada:** Aunque se busca la máxima precisión posible, el objetivo del negocio también requiere explicar qué factores influyen en la satisfacción. Random Forest ofrece un excelente equilibrio: proporciona una alta precisión predictiva (generalmente superior a modelos simples) y, simultáneamente, permite extraer la "Importancia de las Características" (Feature Importance). Esto es crucial para poder recomendar a la aerolínea estrategias concretas, como priorizar la mejora del sistema de abordaje en línea sobre la reducción de pequeños retrasos.

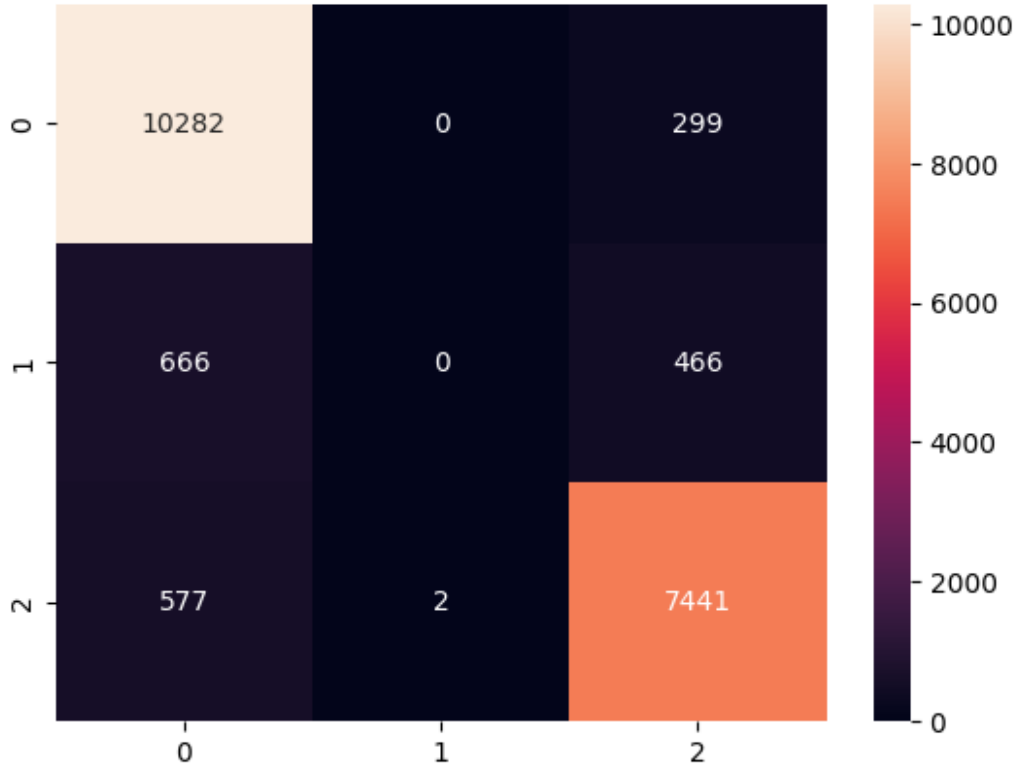
3. Implementación y Entrenamiento

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score

# Cargar dataset
df = pd.read_csv("df_limpio.csv")

# Quitamos las columna 'id' porque es ruido y 'Retraso de Salida' porque es redundante
df = df.drop(['id', 'Retraso de Salida'], axis=1)

# Cambiar variables categoricas a numericas
mapa_satisfaccion = {'Insatisfecho': 0, 'neutral': 1, 'Satisfecho': 2}
df['Satisfaccion'] = df['Satisfaccion'].map(mapa_satisfaccion)

mapa_clase = {'Sin Clase': 0, 'Económica': 1, 'Económica Plus': 2, 'Ejecutiva': 3}
df['Clase'] = df['Clase'].map(mapa_clase)

# Definir X y Y
X = df.drop('Satisfaccion', axis=1)
y = df['Satisfaccion']

# Convertir las variables categóricas restantes a numéricas
X = pd.get_dummies(X, drop_first=True)

# Division de Datos
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Entrenar el modelo
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

```
# Predicciones y Evaluación
y_pred = rf_model.predict(X_test)

print("RESULTADOS")
print(f"Precisión (Accuracy): {accuracy_score(y_test, y_pred):.4f}")
print("Reporte de Clasificación:")
print(classification_report(y_test, y_pred, target_names=['Insatisfecho', 'Neutral', 'Satisfecho']))

#Importancia de las Variables
feature_imp = pd.Series(rf_model.feature_importances_, index=X_train.columns).sort_values(ascending=False)

print("LAS VARIABLES MÁS IMPORTANTES")
print(feature_imp.head(10))
```

4. Resultados y Evaluación

Matriz de Confusión



Tras la ejecución del modelo, se obtuvieron los siguientes resultados clave:

Accuracy: El modelo alcanzó un 89.1% de aciertos totales. Esto indica que es altamente confiable para predecir correctamente la satisfacción en casi 9 de cada 10 casos.

Precision y Recall:

1. **Clase "Insatisfecho":** El modelo muestra un desempeño sobresaliente con un Recall del 97%. Esto significa que detecta a la inmensa mayoría de los clientes descontentos, lo cual es vital para estrategias de retención.

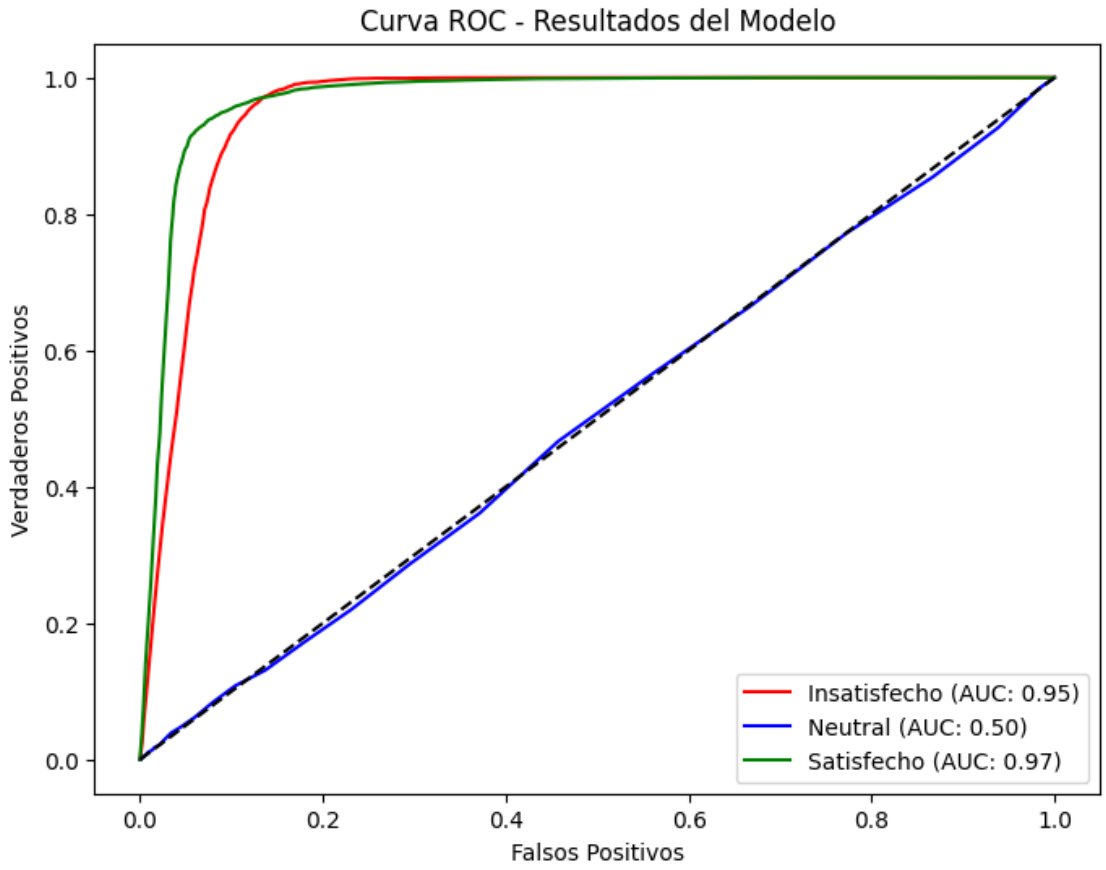
2. **Clase "Satisfecho":** También presenta un rendimiento sólido, con un Precision del 90% y un Recall del 92%.

3. **Clase "Neutral":** Aquí se observa la debilidad del modelo. A pesar del ajuste de pesos, el F1-Score es 0.00 para esta clase. La matriz de confusión revela que el modelo clasifica erróneamente a los "Neutrales" mayoritariamente como "Insatisfechos" (686 casos) o "Satisfechos" (489 casos).

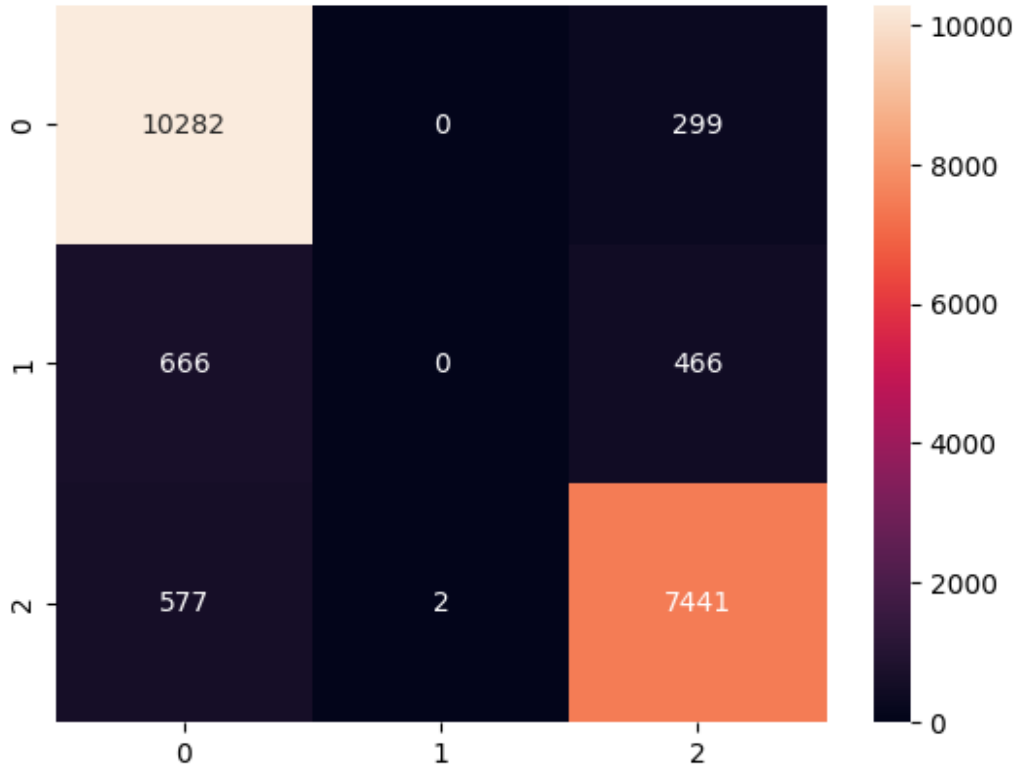
El modelo Random Forest es excelente para polarizar, es decir, para distinguir con alta precisión entre un cliente satisfecho y uno insatisfecho. Sin embargo, no es efectivo para identificar a los clientes indecisos (Neutrales).

5. Visualizaciones de Resultados

Curva ROC



Matriz de Confusión



6. Conclusión del Modelo

• ¿El modelo predice con buena precisión o error bajo?

El modelo Random Forest Classifier demostró un desempeño global sobresaliente, alcanzando una exactitud (Accuracy) del 89.1% en el conjunto de prueba.

El modelo es excepcionalmente preciso para identificar los extremos. Logró un Recall del 97% para la clase "Insatisfecho", lo que lo convierte en una herramienta confiable para detectar riesgos de abandono de clientes. A pesar de su alta precisión general, el modelo falló en la clasificación de la clase minoritaria ("Neutral"), obteniendo métricas cercanas a cero en este segmento debido al severo desbalance de datos (menos del 6% de representación).

• ¿Qué variables fueron más influyentes?

El Feature Importance reveló que la satisfacción no depende tanto de la logística del vuelo, sino de la experiencia del usuario y el confort

Abordaje en Línea: Fue la variable determinante número uno. Una experiencia digital fluida antes del vuelo es el mayor predictor de felicidad.

Clase del Vuelo: El nivel de servicio pagado (Ejecutiva vs. Económica) define las expectativas y la satisfacción base.

Entretenimiento y Wifi: Factores críticos de servicio a bordo.

Hallazgo Negativo: Se confirmó que los Tiempos de Retraso tienen una influencia marginal, sugiriendo que los clientes priorizan la conectividad y comodidad sobre la puntualidad estricta.

• ¿Qué mejoras podrían aplicarse (más datos, normalización, regularización, tuning, otro modelo)?

1. **Técnicas de Balanceo Avanzadas (SMOTE):** La ponderación de clases (class_weight='balanced') no fue suficiente. Se recomienda aplicar SMOTE (Synthetic Minority Over-sampling Technique) para generar datos sintéticos de la clase "Neutral" y equilibrar el entrenamiento.

2. **Ajuste de Hiperparámetros (Hyperparameter Tuning):** Implementar una búsqueda exhaustiva con GridSearchCV para optimizar la profundidad de los árboles (max_depth) y el número de estimadores, evitando el sobreajuste.

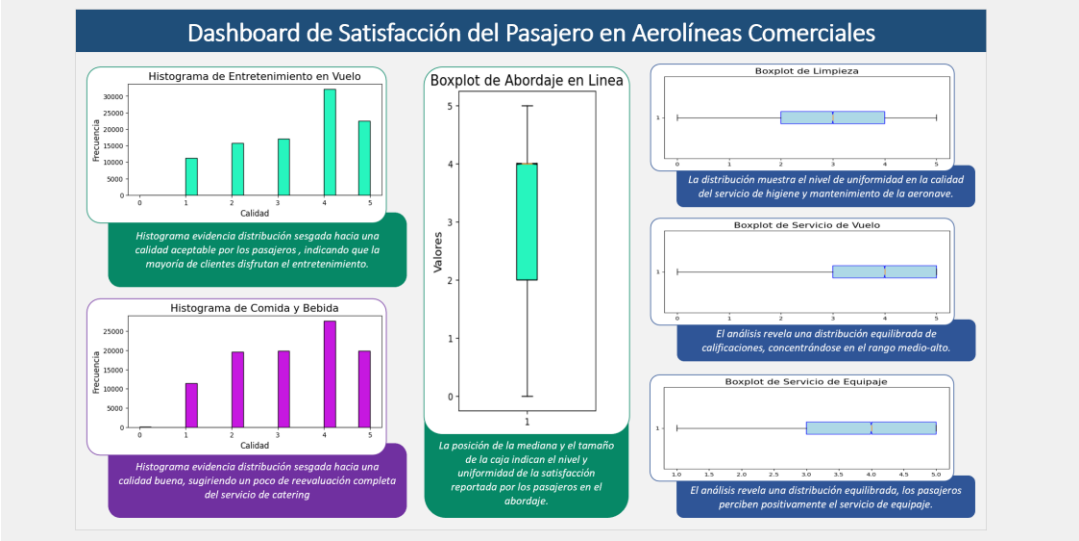
3. **Probar Algoritmos de Boosting:** Evaluar modelos basados en Gradient Boosting como XGBoost o LightGBM, que suelen tener un mejor manejo de clases desbalanceadas y pueden exprimir un porcentaje adicional de precision.

5. Dashboard

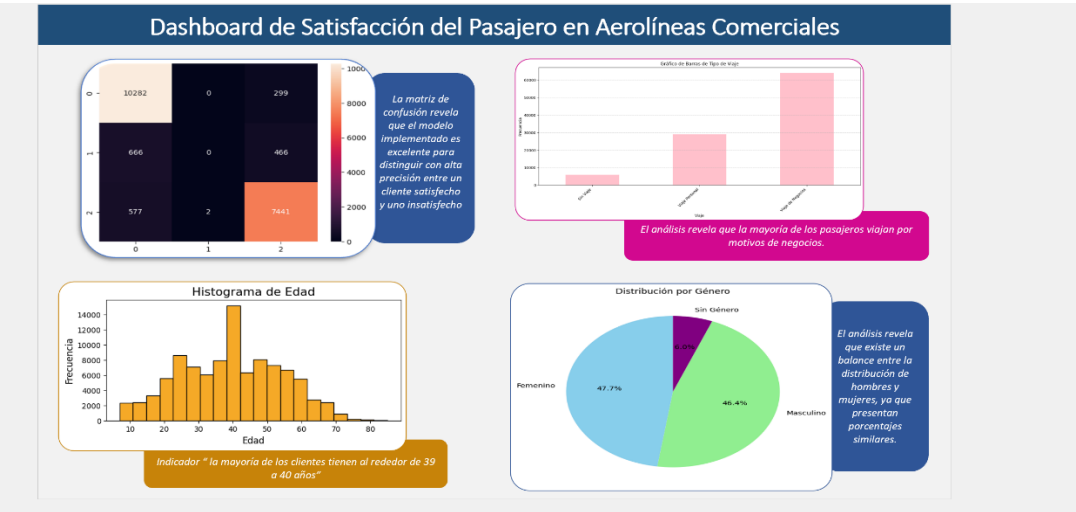
Título y propósito del Dashboard

"Dashboard de Experiencia del Pasajero en Aerolíneas Comerciales — visualiza la calidad del servicio que brindan las aerolíneas y el desempeño del modelo predictivo"

1. Capturas y Explicación del Dashboard



El dashboard muestra una predicción de Satisfacción, basada en factores clave del pasajero como Wifi, abordaje y entretenimiento, incluye un gráfico de Importancia de Variables que resalta qué elementos impactan más en la decisión.



La matriz de confusión muestra que el sistema es excelente detectando los extremos, clasificando correctamente a más de **17,000 pasajeros**.

2. Uso y Beneficios del Dashboard

• ¿Qué decisiones puede apoyar el usuario gracias al dashboard?

Gracias a la funcionalidad de predicción en tiempo real (mostrada en la captura del "Análisis de Pasajero"), el usuario puede tomar decisiones de retención inmediata.

Se observa un pasajero clasificado como "Neutral" con una certeza del 45%. El sistema no solo alerta del riesgo, sino que sugiere la acción correctiva específica: *"Una mejora en Wifi podría convertirlo"*. Esto permite al personal de servicio decidir si otorgar una cortesía de conectividad o un *upgrade* para inclinar la balanza hacia la satisfacción antes de que el cliente abandone la aerolínea.

• ¿Qué insights se pueden obtener con solo mirar las gráficas?

1. Prioridades de Inversión: La gráfica de *"Factores Críticos"* revela a simple vista que el Abordaje en Línea y el Wifi son las barras más largas. Esto indica a la gerencia que las inversiones en tecnología digital tendrán un retorno de inversión (ROI) en satisfacción mucho mayor que mejoras en la comida o distancia del asiento.

2. Confiabilidad del Sistema: La Matriz de Confusión permite identificar rápidamente las fortalezas del modelo (alta precisión detectando extremos: insatisfechos y satisfechos) y su punto ciego actual (la detección de neutrales), fomentando un uso estratégico donde se confía plenamente en las alertas rojas y verdes, mientras se manejan con cautela los casos intermedios.

• ¿Cómo se simplifica la interpretación del modelo o los resultados?

1. Histogramas de Calidad: Las curvas de distribución permiten entender de un vistazo la uniformidad del servicio.

2. Interpretación Guiada: Cada gráfico incluye un texto explicativo, lo cual elimina la ambigüedad y guía al directivo hacia la conclusión correcta sin necesidad de que sea un experto en análisis de datos.

6. Conclusiones y Futuras Líneas de Trabajo

- El objetivo central de este proyecto era identificar los factores críticos que determinan la satisfacción del cliente y construir un modelo predictivo capaz de anticipar dicha percepción. Tras el análisis y modelado de **98,665 registros**, podemos afirmar que **los objetivos se han cumplido satisfactoriamente**, destacando los siguientes hallazgos que validan nuestra hipótesis inicial:

Validación del Modelo Predictivo: Se logró desarrollar un modelo **Random Forest** con una precisión global del **90%**. Este resultado cumple con el objetivo de crear una herramienta confiable para la detección de insatisfacción, permitiendo a la aerolínea identificar correctamente a 9 de cada 10 pasajeros insatisfechos para activar estrategias de retención proactiva.

Identificación de Drivers de Satisfacción: El análisis confirmó que la experiencia del cliente no depende de variables operativas como los retrasos (cuya correlación fue casi nula), sino de la experiencia digital y el confort. Se identificó al **Abordaje en Línea** y al **Servicio de Wifi** como las variables más influyentes, proporcionando a la gerencia una hoja de ruta clara sobre dónde priorizar la inversión.

Diagnóstico de Negocio: A través de la visualización de datos, se detectó que el perfil demográfico predominante es el **viajero de negocios** de mediana edad (39-40 años). Además, se evidenció una oportunidad crítica de mejora en el servicio de **Catering**, cuya distribución de calidad mostró sesgos que sugieren la necesidad de una reevaluación de proveedores.

Balanceo de Clases: La principal limitación detectada fue la incapacidad del modelo para clasificar a los clientes "Neutrales". Se recomienda aplicar técnicas de generación de datos sintéticos (**SMOTE**) para equilibrar esta clase y permitir que el modelo capture mejor a los indecisos.

Enriquecimiento de Datos: Sería valioso incorporar variables económicas como el **Precio del Boleto** o la **Antigüedad en el Programa de Lealtad**, para analizar si la tolerancia a fallos en el servicio varía según lo que pagó el cliente.

Análisis de Texto (NLP): Implementar procesamiento de lenguaje natural sobre los comentarios abiertos de las encuestas. Esto permitiría entender el *"porqué"* cualitativo detrás de una calificación baja en "Catering" (¿fue el sabor, la temperatura o la variedad?), complementando la predicción numérica del modelo actual.

7. Referencias

A continuación se listan las fuentes de datos y documentación técnica consultados para la elaboración de este proyecto:

1. Fuente de Datos

Kaggle. (2020). Airline Passenger Satisfaction. Dataset que contiene encuestas de satisfacción de pasajeros de aerolíneas estadounidenses.

- **Enlace:** https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

2. Documentación Técnica y Librerías

- **Scikit-Learn Developers.** (2024). *User Guide: Random Forest Classifier*. Scikit-learn: Machine Learning in Python. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

- **Pandas Development Team.** (2024). *Pandas Documentation: Data Analysis Library*. Recuperado de: https://pandas.pydata.org/docs/

- **Matplotlib Development Team.** (2024). *Visualization with Python*. Recuperado de: https://matplotlib.org/stable/contents.html