2025

Análisis de la satisfacción de pasajeros en aerolíneas

Isaac Avila

Universidad Tecnológica de Panamá

7-4-2025

Universidad Tecnológica de Panamá

Facultad de Ingeniería en Sistemas y Computación

Maestría en analítica de datos

Asignatura

Modelos predictivos

Proyecto Final

Análisis de la satisfacción de pasajeros en aerolíneas

Estudiante

Isaac Aldair Avila

Cedula

8-954-1813

Profesor

PhD Juan Marcos Castillo

Grupo

Año

2025

Introducción

La industria aérea cambia constantemente, y uno de los aspectos más importantes hoy en día es la experiencia del pasajero. Las aerolíneas ya no solo compiten por ser puntuales, sino por ofrecer un buen servicio en cada etapa del viaje.

Este proyecto busca analizar qué tan satisfechos están los pasajeros usando técnicas de clasificación, a partir de datos sobre el vuelo, el servicio y el perfil del cliente. Para eso, se usa el dataset “Airline Passenger Satisfaction” disponible en Kaggle, que ha sido bastante utilizado por la comunidad.

El desarrollo principal del modelo se hace en KNIME, una herramienta que permite trabajar con los datos de forma visual y estructurada. También se usa Weka para probar algunos algoritmos y Jupyter Notebook para explorar, limpiar y graficar la información. Todo esto forma parte del curso de Modelos Predictivos y además tiene aplicación directa en mi entorno laboral.

Contenido

Introducción 3
Justificación 5
Antecedentes 6
Definición del problema 7
Avance de análisis predictivos 8
1. Tamaño del dataset 8
2. Descripción del dataset 8
3. Variable objetivo: Satisfaction 9
4. Estadísticas generales del dataset 9
Estadísticas descriptivas 10
Graficas de visualización 14
KNIME Para procesamiento de la data 14
Correlación entre variables 15
Jupyter Notebook para gráficos 17
Weka para gráficos 22
Bibliográfica 23

Justificación

Este proyecto nace del interés por aplicar lo aprendido en clase a una situación real. Trabajo en el centro de operaciones de una aerolínea y veo todos los días cómo decisiones operativas pueden influir en la experiencia del pasajero, desde los retrasos hasta el trato del personal. Por eso quise usar datos para entender mejor qué factores tienen más peso en la satisfacción del cliente.

Elegí una base de datos disponible en Kaggle, ya validada por la comunidad, porque ofrece una buena variedad de variables relacionadas con el viaje. Decidí trabajar principalmente en KNIME, que me permite manejar todo el flujo de análisis de forma visual y ordenada. También utilicé Weka y Python (Jupyter Notebook) para explorar los datos, hacer gráficas y probar modelos complementarios.

Este proyecto no solo cumple con los objetivos del curso, sino que también deja ideas que podrían aplicarse en la empresa, especialmente en temas de mejora del servicio al cliente desde una perspectiva basada en datos.

Antecedentes

La satisfacción del pasajero en aerolíneas ha sido tema de estudio en muchas investigaciones, especialmente en áreas como puntualidad, atención del personal, comodidad en el vuelo y servicios a bordo. Con el avance de la analítica de datos, se ha vuelto común aplicar modelos de aprendizaje automático para analizar o predecir el nivel de satisfacción de los clientes, usando variables tanto operativas como de percepción.

En sitios como Kaggle hay varios trabajos que usan el dataset “Airline Passenger Satisfaction” para este propósito. Aunque no se tiene una fuente oficial del dataset, ha sido ampliamente utilizado por la comunidad y cuenta con buena documentación, lo que lo hace una referencia confiable para estudios de este tipo.

Este proyecto busca ir más allá de la aplicación técnica. La idea es conectar los resultados con situaciones reales dentro de una aerolínea, especialmente en áreas donde las decisiones operativas pueden tener un impacto directo en cómo el pasajero evalúa su experiencia. Eso le da al análisis un enfoque más aplicado y útil.

Definición del problema

El objetivo principal de este proyecto es responder a la pregunta:

¿Qué variables permiten predecir si un pasajero está satisfecho con su experiencia de vuelo?

La variable objetivo es la columna “Satisfaction”, que clasifica a los pasajeros en dos grupos: “Satisfied” y “Neutral or Unsatisfied”. La meta es construir un modelo de clasificación que pueda predecir correctamente esta variable, a partir de información sobre el perfil del pasajero, el tipo de viaje, aspectos operativos del vuelo y evaluaciones del servicio.

- Para lograr esto, el proyecto se enfoca en los siguientes pasos:
- Preparar y limpiar la base de datos para su uso en KNIME.
- Hacer un análisis descriptivo con apoyo de Jupyter Notebook.
- Identificar patrones, correlaciones y posibles outliers.
- Probar distintos modelos de clasificación (como árboles de decisión, random forest, naive bayes, entre otros) y comparar su rendimiento.
- Evaluar qué variables tienen mayor peso en la predicción y qué se puede interpretar a partir de eso.

El enfoque no es solo técnico, también busca generar conclusiones que puedan tener sentido práctico dentro de una aerolínea, especialmente en la toma de decisiones orientadas al cliente.

Avance de análisis predictivos

Antes de aplicar modelos de clasificación, es importante conocer cómo está compuesto el dataset y explorar los datos para entender mejor qué información aporta cada variable. Este análisis preliminar ayuda a identificar patrones, posibles errores, valores atípicos y relaciones entre variables, lo cual es clave para el proceso de modelado.

1. Tamaño del dataset

El dataset contiene un total de 129.880 registros (filas), donde cada uno representa a un pasajero individual. La cantidad de datos es suficiente para entrenar y validar modelos predictivos con buena representatividad. Además, incluye un conjunto variado de columnas que permiten analizar múltiples aspectos de la experiencia del pasajero.

2. Descripción del dataset

El dataset utilizado en esta investigación se llama “Airline Passenger Satisfaction” y está compuesto por información de pasajeros de aerolínea, incluyendo características personales, aspectos operativos del vuelo y evaluaciones subjetivas sobre distintos servicios. Cada fila representa a un pasajero y las variables se organizan de la siguiente forma:

- ID: Identificador único del pasajero.
- Gender: Género del pasajero (Female / Male).
- Age: Edad del pasajero.
- Customer Type: Tipo de cliente (First-time / Returning).
- Type of Travel: Motivo del viaje (Business / Personal).
- Class: Clase de asiento en la que viajó el pasajero (Economy, Business, Economy Plus).
- Flight Distance: Distancia del vuelo en millas.
- Departure Delay: Minutos de demora en la salida del vuelo.
- Arrival Delay: Minutos de demora en la llegada del vuelo.

El resto de las variables corresponden a evaluaciones del pasajero sobre distintos servicios del proceso de viaje. Todas estas están medidas en una escala del 1 (muy insatisfecho) al 5 (muy satisfecho), con valor 0 cuando el ítem no aplica para el pasajero:

- Departure and Arrival Time Convenience: Nivel de satisfacción con la conveniencia de los horarios de salida y llegada.
- Ease of Online Booking: Facilidad para hacer la reserva online.
- Check-in Service: Satisfacción con el proceso de Check-in.
- Online Boarding: Experiencia con el embarque en línea.

- Gate Location: Evaluación de la ubicación de la puerta de embarque.
- On-board Service: Servicio recibido antes de abordar.
- Seat Comfort: Comodidad del asiento durante el vuelo.
- Leg Room Service: Espacio para las piernas en el asiento.
- Cleanliness: Limpieza del avión.
- Food and Drink: Calidad de la comida y bebida ofrecidas.
- In-flight Service: Servicio durante el vuelo.
- In-flight Wi-Fi Service: Calidad del servicio de wifi a bordo.
- In-flight Entertainment: Opciones de entretenimiento disponibles durante el vuelo.
- Baggage Handling: Evaluación del manejo del equipaje por parte de la aerolínea.

Finalmente, la columna Satisfaction es la variable objetivo de este análisis, y clasifica la satisfacción general del pasajero en dos categorías:

- Satisfied
- Neutral or unsatisfied

Este dataset combina variables numéricas, categóricas y ordinales, lo que lo hace muy útil para el desarrollo de modelos de clasificación. Además, incluye tanto factores operativos como percepciones personales, lo que permite explorar una variedad de relaciones e impactos sobre la satisfacción final del pasajero.

3. Variable objetivo: Satisfaction

La columna Satisfaction representa el nivel general de satisfacción del pasajero con la aerolínea. Es una variable categórica que toma dos valores:

- Satisfied
- Neutral or unsatisfied

Esta variable será utilizada como etiqueta en los modelos de clasificación. El objetivo es predecir correctamente a cuál de estas dos categorías pertenece un pasajero, en base a las demás variables del dataset.

4. Estadísticas generales del dataset

Se realizó un análisis estadístico de todas las variables numéricas y ordinales presentes en el dataset. A continuación, se destacan algunos puntos relevantes a partir de los valores mínimos, máximos, medias, desviaciones estándar, sesgos y curtosis:

Estadísticas descriptivas

Column	Min	Max	Mean	Std. deviation	Variance	Skewness	Kurtosis	Overall sum
Age	7.00	85.00	39.43	15.12	228.60	0.00	-0.72	5120903.00
Flight Distance	31.00	4983.00	1190.32	997.45	994911.44	1.11	0.27	154598293.00
Departure Delay	0.00	1592.00	14.71	38.07	1449.41	6.82	100.64	1911017.00
Arrival Delay	0.00	1584.00	15.05	38.42	1475.82	6.68	95.36	1954105.00
Departure and Arrival Time Convenience	0.00	5.00	3.06	1.53	2.33	-0.33	-1.04	397121.00
Ease of Online Booking	0.00	5.00	2.76	1.40	1.96	-0.02	-0.91	358063.00
Check-in Service	0.00	5.00	3.31	1.27	1.60	-0.37	-0.83	429418.00
Online Boarding	0.00	5.00	3.25	1.35	1.82	-0.46	-0.70	422452.00
Gate Location	0.00	5.00	2.98	1.28	1.63	-0.06	-1.03	386643.00
On-board Service	0.00	5.00	3.38	1.29	1.66	-0.42	-0.89	439387.00
Seat Comfort	0.00	5.00	3.44	1.32	1.74	-0.49	-0.92	446964.00
Leg Room Service	0.00	5.00	3.35	1.32	1.73	-0.35	-0.98	435212.00
Cleanliness	0.00	5.00	3.29	1.31	1.73	-0.30	-1.01	426828.00
Food and Drink	0.00	5.00	3.20	1.33	1.77	-0.16	-1.15	416236.00
In-flight Service	0.00	5.00	3.64	1.18	1.38	-0.69	-0.36	473048.00
In-flight Wi-Fi Service	0.00	5.00	2.73	1.33	1.77	0.04	-0.85	354403.00
In-flight Entertainment	0.00	5.00	3.36	1.33	1.78	-0.37	-1.06	436147.00
Baggage Handling	1.00	5.00	3.63	1.18	1.39	-0.68	-0.38	471739.00
delayed	0.00	1.00	0.46	0.50	0.25	0.17	-1.97	59498.00

Edad (Age): El rango va de 7 a 85 años, con una media de aproximadamente 39. La distribución es ligeramente simétrica (skewness ≈ 0), aunque tiende un poco hacia la izquierda (curtosis -0.72).

Distancia del vuelo (Flight Distance): Va desde vuelos muy cortos (31 millas) hasta vuelos largos de casi 5.000 millas. El sesgo positivo (1.11) indica que hay una cantidad significativa de vuelos cortos, pero también unos pocos vuelos muy largos que elevan el promedio.

Demoras (Departure Delay y Arrival Delay): Ambas variables presentan un promedio bajo (~15 minutos), pero una desviación alta y valores máximos muy elevados (más de 1500 minutos), lo que muestra la presencia de outliers significativos. Además, tienen un alto sesgo positivo (>6) y curtosis muy alta (>95), lo que confirma una distribución fuertemente asimétrica con valores extremos.

Variables de satisfacción (escala 0-5):

En general, las medias de las variables de satisfacción están entre 2.7 y 3.6, lo cual indica una percepción moderada.

Algunas variables como In-flight Service, Seat Comfort y On-board Service tienen medias más altas, lo que sugiere una mejor valoración.

In-flight Wi-Fi Service y Ease of Online Booking son de las más bajas (≈ 2.7), lo que podría indicar oportunidades de mejora.

En casi todas estas variables, la distribución es ligeramente sesgada hacia la izquierda (sesgo negativo), lo cual implica que hay más pasajeros evaluando con puntajes altos (4 y 5), pero también una proporción considerable de bajas calificaciones y ceros (no aplica).

Variable delayed: Es una variable binaria que indica si el vuelo tuvo alguna demora. Su media es 0.46, es decir, aproximadamente el 46% de los vuelos tuvieron alguna demora, lo cual es una proporción significativa a tener en cuenta en el análisis de satisfacción.

Modelo de regresión lineal

Variable	Coeff.	Std. Err.	t-value	P>|t|
Gender=Male	0.01	0.00	5.07	0.00
Age	0.00	0.00	-14.97	0.00
Customer Type=Returning	0.31	0.00	103.62	0.00
Type of Travel=Personal	-0.38	0.00	-133.13	0.00
Class=Economy	-0.13	0.00	-46.27	0.00
Class=Economy Plus	-0.14	0.00	-34.23	0.00
Flight Distance	0.00	0.00	1.00	0.32
Departure Delay	0.00	0.00	5.12	0.00
Arrival Delay	0.00	0.00	-11.15	0.00
Departure and Arrival Time Convenience	-0.02	0.00	-21.99	0.00
Ease of Online Booking	-0.04	0.00	-37.96	0.00
Check-in Service	0.04	0.00	46.49	0.00
Online Boarding	0.08	0.00	81.98	0.00
Gate Location	0.00	0.00	3.64	0.00
On-board Service	0.03	0.00	36.49	0.00
Seat Comfort	0.01	0.00	7.14	0.00
Leg Room Service	0.03	0.00	40.69	0.00
Cleanliness	0.03	0.00	21.87	0.00
Food and Drink	0.00	0.00	-4.43	0.00
In-flight Service	0.01	0.00	12.78	0.00
In-flight Wifi Service	0.07	0.00	62.98	0.00
In-flight Entertainment	0.02	0.00	11.32	0.00
Baggage Handling	0.02	0.00	15.47	0.00
Intercept	-0.49	0.01	-73.59	0.00

R-Squared: 0.5519
Adjusted R-Squared: 0.5518

Se aplicó un modelo de regresión lineal para predecir el nivel de satisfacción del cliente (variable is_satisfied) en función de múltiples características relacionadas con el vuelo, el perfil del pasajero y la calidad del servicio. El modelo obtuvo un R-cuadrado de 0.5519, lo cual indica que aproximadamente el 55.2% de la variabilidad en la satisfacción del cliente puede ser explicada por las variables independientes consideradas.

Entre los predictores más influyentes se destacan:

Customer Type = Returning con un coeficiente positivo (0.31), indicando que los clientes recurrentes tienden a estar significativamente más satisfechos.

Type of Travel = Personal y Class = Economy / Economy Plus presentan coeficientes negativos, lo que sugiere que los viajeros por motivos personales y quienes viajan en clases más económicas tienen menor nivel de satisfacción.

Las variables relacionadas con la experiencia de vuelo como Check-in Service, Online Boarding, Leg Room Service, In-flight Wi-Fi Service y On-board Service muestran fuertes asociaciones positivas con la satisfacción, lo que evidencia la importancia de la calidad del servicio en la percepción del cliente.

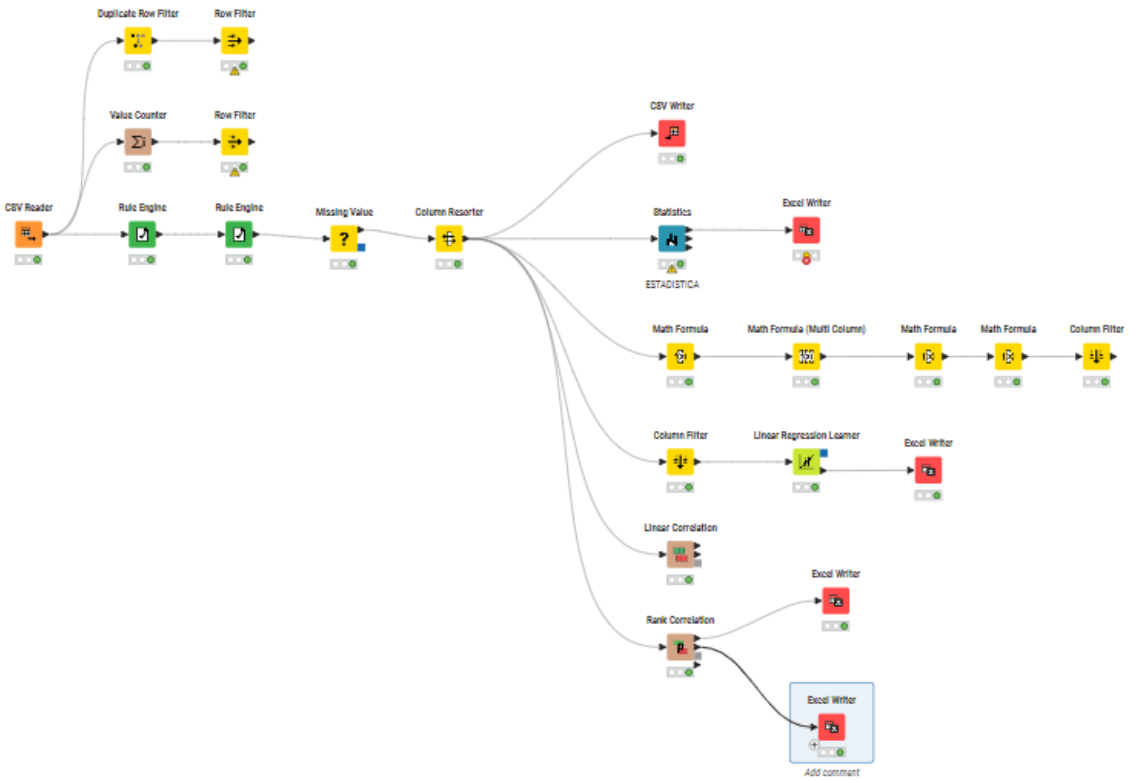
Por otro lado, Arrival Delay y Departure and Arrival Time Convenience se asocian negativamente con la satisfacción, lo que es consistente con el impacto negativo de las demoras y la incomodidad en los horarios.

La mayoría de las variables tienen una significancia estadística alta ($p < 0.05$), indicando que sus efectos son relevantes dentro del modelo. No obstante, Flight Distance no resultó significativa ($p = 0.3162$), por lo que su impacto sobre la satisfacción es poco claro en este contexto.

Graficas de visualización

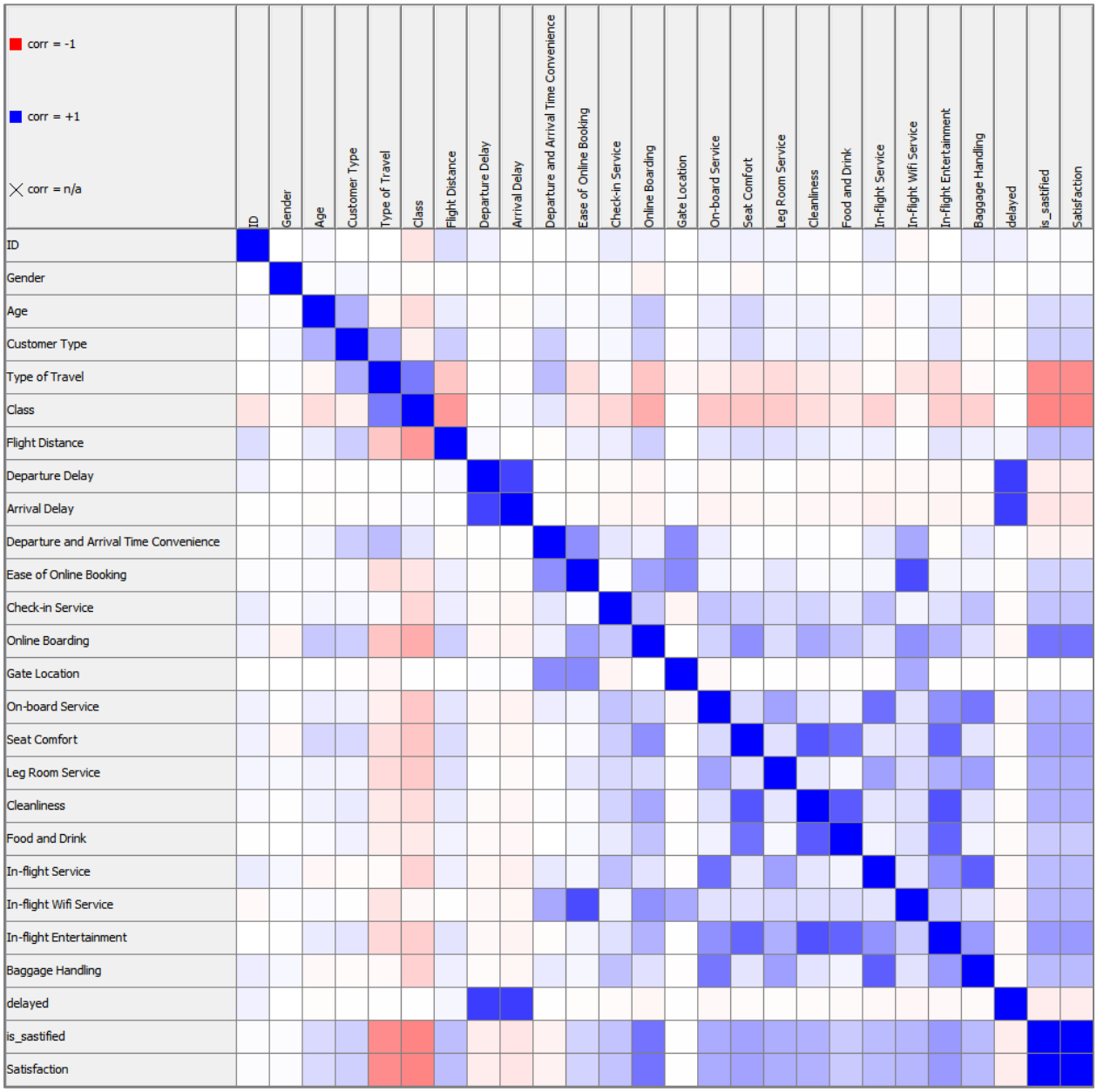
KNIME Para procesamiento de la data

He utilizado la herramienta KNIME para la limpieza y procesamiento de la data, y así identificar Insights.



Correlación entre variables

Utilizando Rank Correlation de KNIME obtenemos



Se generó una matriz de correlación para analizar la relación entre las distintas variables del conjunto de datos. En el gráfico, los colores azules indican correlaciones positivas y los rojos correlaciones negativas, siendo más intensos cuanto mayor es la magnitud. Se observa una fuerte correlación positiva entre las variables de servicios a bordo (como Seat Comfort, Wi-Fi, Cleanliness) y las variables de satisfacción (is_satisfied y Satisfaction), lo cual refuerza su relevancia en el modelo predictivo. Asimismo, las

variables relacionadas con demoras muestran correlaciones negativas con la satisfacción.

First column name	Second column name	Correlation value	p value
Departure Delay	Arrival Delay	74%	0%
Ease of Online Booking	In-flight Wi-Fi Service	71%	0%
Cleanliness	In-flight Entertainment	68%	0%
Seat Comfort	Cleanliness	67%	0%
Cleanliness	Food and Drink	65%	0%
In-flight Service	Baggage Handling	63%	0%
Food and Drink	In-flight Entertainment	61%	0%
Seat Comfort	In-flight Entertainment	60%	0%

En el análisis de correlación, se identificaron los pares de variables con mayor asociación lineal. El valor de correlación más alto se da entre Departure Delay y Arrival Delay (74%), lo cual es esperable, ya que una demora en salida suele implicar una llegada tardía. También se observa una fuerte correlación entre Ease of Online Booking y In-flight Wifi Service (71%), posiblemente reflejando una percepción general positiva del servicio tecnológico ofrecido por la aerolínea.

Las variables relacionadas con la comodidad y limpieza presentan correlaciones notables: Cleanliness se asocia fuertemente con In-flight Entertainment (68%) y Food and Drink (65%), mientras que Seat Comfort también se relaciona con Cleanliness (67%) y In-flight Entertainment (60%). Estos patrones indican que los pasajeros que valoran una dimensión del confort o la experiencia a bordo tienden a calificar bien otras similares.

Todos los pares tienen valores-p del 0%, lo que confirma que estas correlaciones son estadísticamente significativas.

Jupyter Notebook para gráficos

Relación satisfacción con el genero

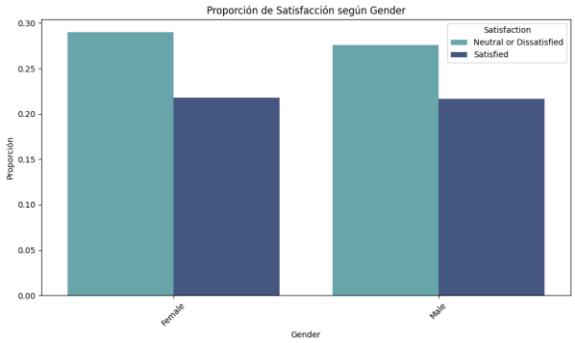


Ilustración 1 Satisfacción según Gender

La distribución de satisfacción es similar entre hombres y mujeres, sin diferencias significativas. Esto indica que el género no parece tener un peso importante en la percepción general del servicio.

Relación satisfacción por edad

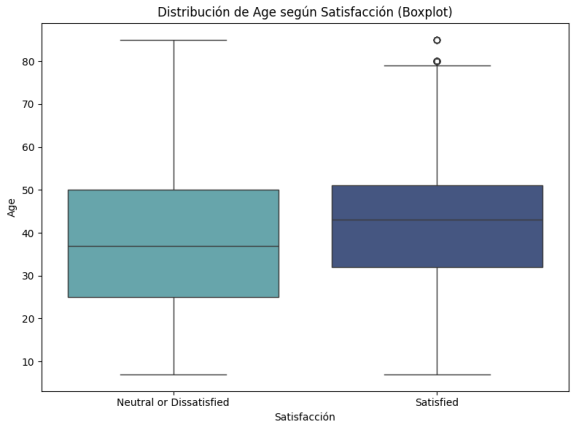


Ilustración 2 Distribución de edad según satisfacción

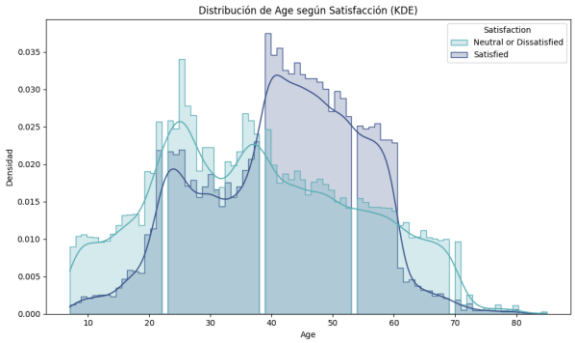


Ilustración 3 Distribución de edad según satisfacción

Los pasajeros satisfechos tienden a tener una edad promedio ligeramente mayor. También presentan una menor dispersión en comparación con los pasajeros neutrales o insatisfechos.

Satisfacción según tipo de viaje

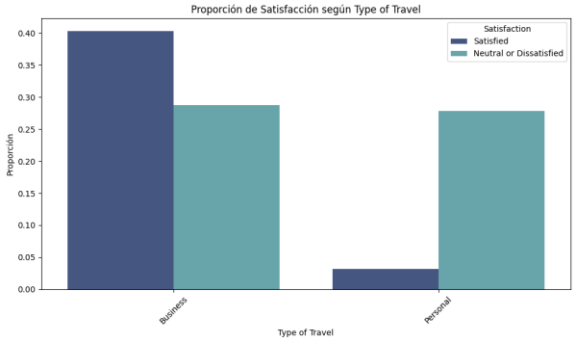


Ilustración 4 Satisfacción según tipo de viaje

Los pasajeros que viajan por negocios muestran un nivel de satisfacción claramente más alto que los que viajan por motivos personales. Esta variable marca una diferencia notable en la percepción del servicio.

Satisfacción según clase

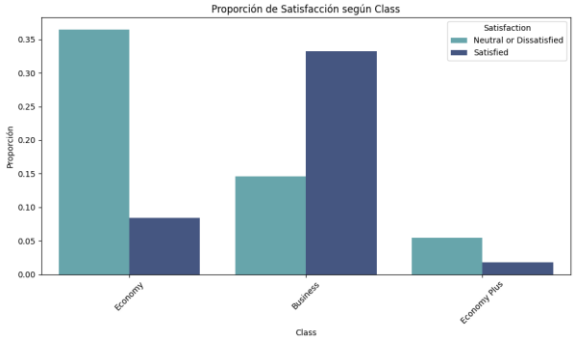


Ilustración 5 Satisfacción según clase

La clase del asiento influye directamente en la satisfacción: los pasajeros en clase Business reportan mayor satisfacción, seguidos por Economy Plus. Los de clase Economy son los menos satisfechos.

Satisfacción según distancia del vuelo

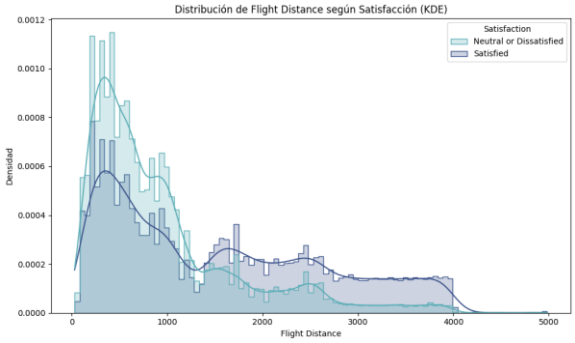


Ilustración 6 Satisfacción según distancia del vuelo

Aunque no hay una relación lineal fuerte, los vuelos de mayor distancia tienden a estar asociados con un porcentaje ligeramente mayor de pasajeros satisfechos.

Check-in Service y satisfacción

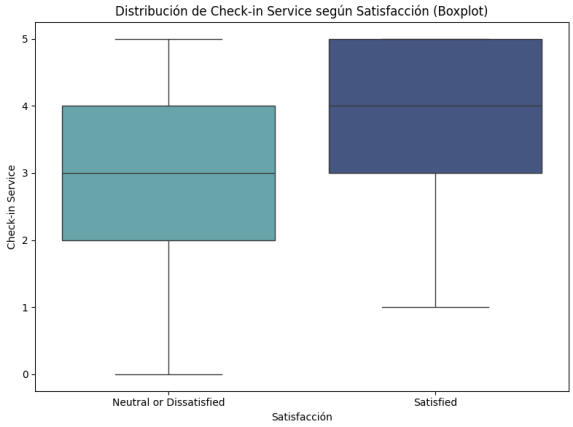


Ilustración 7 Distribución de satisfacción según Check Service

A mayor valoración del servicio de Check-in, mayor es el nivel de satisfacción. Este es uno de los puntos de contacto iniciales con el pasajero y su impacto es visible.

Online Boarding y satisfacción

Ilustración 8 Distribución de satisfacción según Online Boarding

La experiencia con el embarque en línea también tiene un impacto positivo en la satisfacción. Los pasajeros que calificaron bien este servicio suelen estar más satisfechos en general.

On-board Service y satisfacción

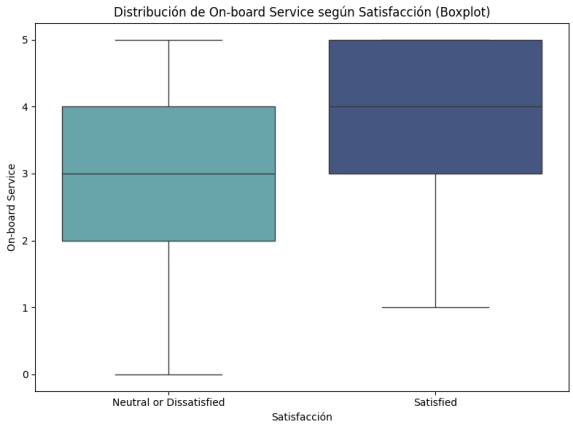


Ilustración 9 Distribución de satisfacción según On-board Service

El servicio recibido a bordo influye notablemente en la percepción final del pasajero. Puntajes altos en esta variable se asocian con una mayor proporción de satisfacción.

Seat Comfort y satisfacción

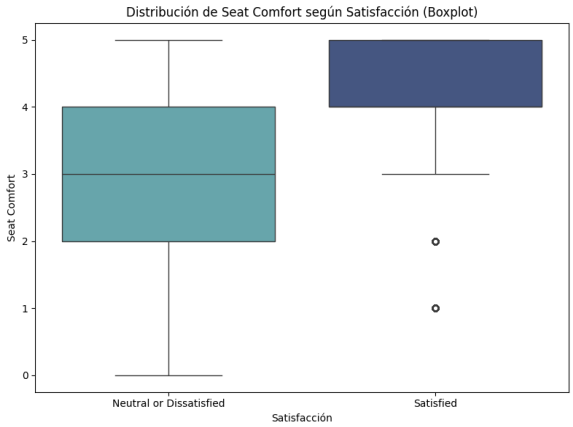
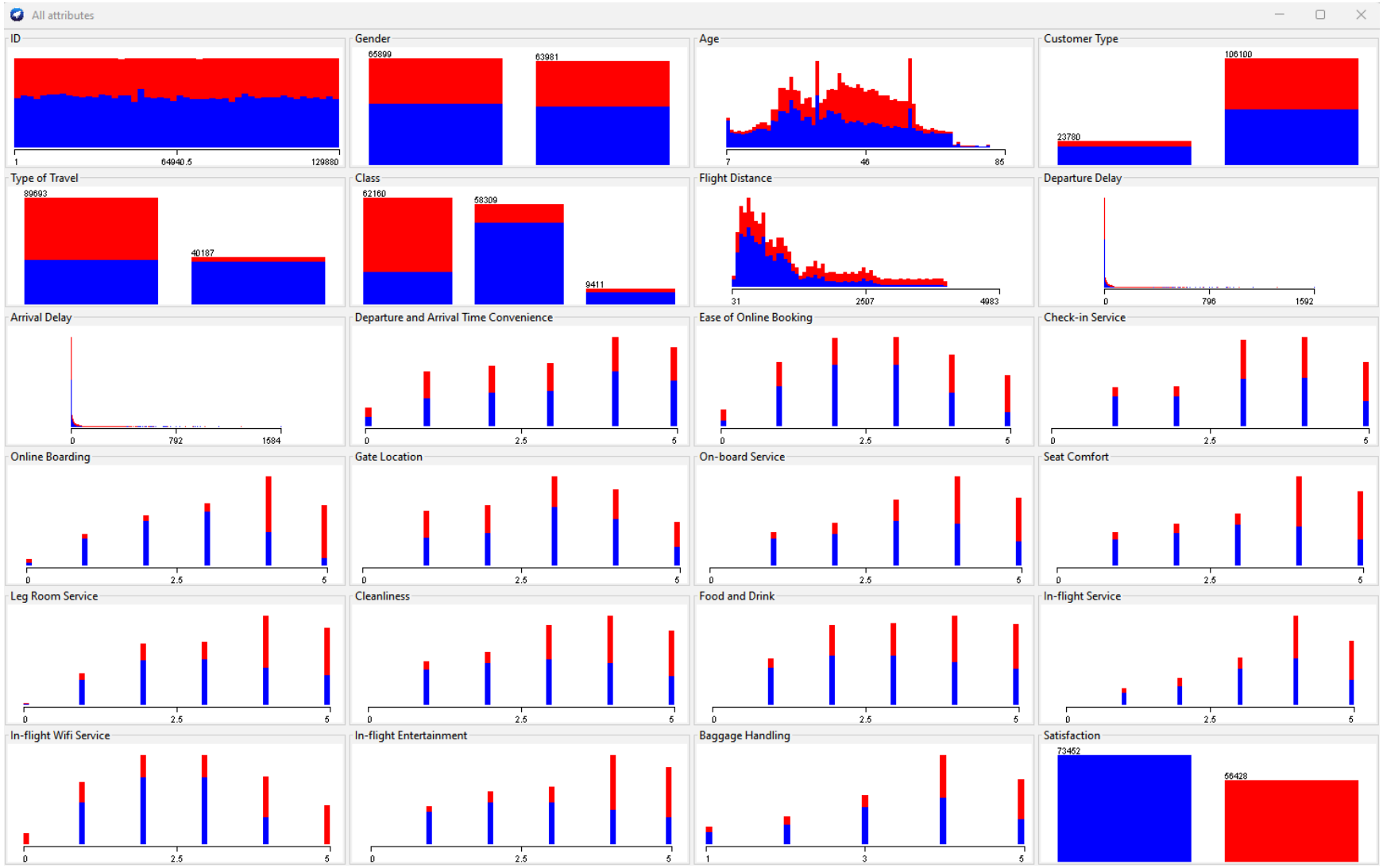


Ilustración 10 Distribución de satisfacción según Seat Comfort

La comodidad del asiento es una de las variables con mayor peso visual. Los pasajeros que calificaron alto este aspecto tienden a estar significativamente más satisfechos.

Weka para gráficos

La herramienta Weka nos permite una visualización de la relación entre Satisfacción y las demás variables



Bibliográfica

Prestwich, O. (2020, 15 febrero). *people in airplane during daytime*. Unsplash. https://unsplash.com/photos/people-in-airplane-during-daytime-FCM4k7LcggU