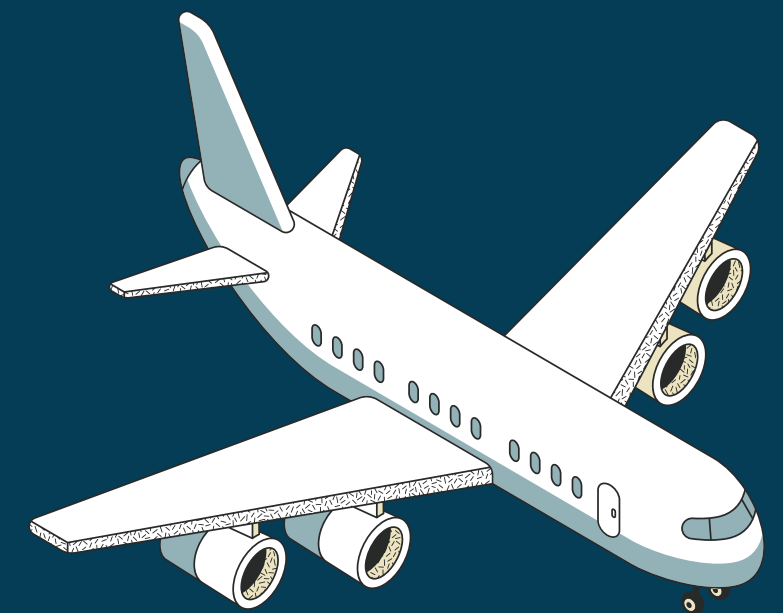


FINAL ANÁLISIS PREDICTIVO

Sofía González del Solar

Introducción

- Base de datos elegida: Airline_Passenger_Satisfaction
- Origen: Kaggle
- Variable a predecir: satisfaction (Satisfied/ Neutral or Dissatisfied)
- Objetivos: lograr un entendimiento profundo de la base elegida y utilizar la información descubierta en el EDA para lograr la mejor predicción posible de la variable *satisfaction*.
- Hipótesis: Se conseguirá un alto accuracy en la predicción debido a las variables que contiene el dataset. Este accuracy será mayor en la categoría neutral or dissatisfied



	A	B	C	D	E
1	Gender	Customer Type	Age	Type of Travel	Class
2	Male	Loyal Customer	13	Personal Travel	Eco Plus
3	Male	disloyal Customer	25	Business travel	Business
4	Female	Loyal Customer	26	Business travel	Business
5	Female	Loyal Customer	25	Business travel	Business
6	Male	Loyal Customer	61	Business travel	Business
7	Female	Loyal Customer	26	Personal Travel	Eco
8	Male	Loyal Customer	47	Personal Travel	Eco
9	Female	Loyal Customer	52	Business travel	Business
10	Female	Loyal Customer	41	Business travel	Business
11	Male	disloyal Customer	20	Business travel	Eco
12	Female	disloyal Customer	24	Business travel	Eco
13	Female	Loyal Customer	12	Personal Travel	Eco Plus
14	Male	Loyal Customer	53	Business travel	Eco
15	Male	Loyal Customer	33	Personal Travel	Eco
16	Female	Loyal Customer	26	Personal Travel	Eco
17	Male	disloyal Customer	13	Business travel	Eco
18	Female	Loyal Customer	26	Business travel	Business
19	Male	Loyal Customer	41	Business travel	Business
20	Female	Loyal Customer	45	Business travel	Business
21	Male	Loyal Customer	38	Personal Travel	Eco
22	Male	Loyal Customer	9	Business travel	Eco
23	Female	Loyal Customer	17	Personal Travel	Eco
24	Female	Loyal Customer	43	Personal Travel	Eco
25	Female	Loyal Customer	58	Personal Travel	Eco
26	Female	disloyal Customer	23	Business travel	Eco
27	Male	Loyal Customer	57	Personal Travel	Eco
28	Female	Loyal Customer	22	Business travel	Business

DATASET

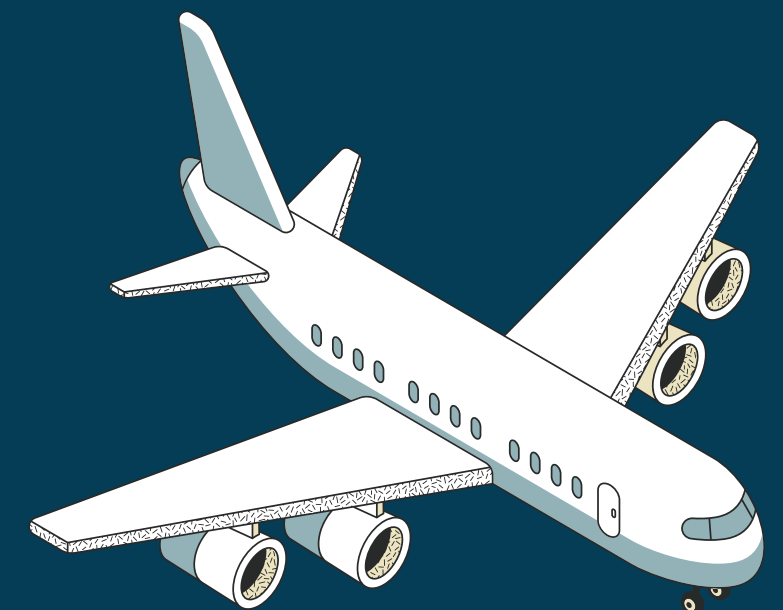
AIRLINE_PASSENGER_SATISFACTION

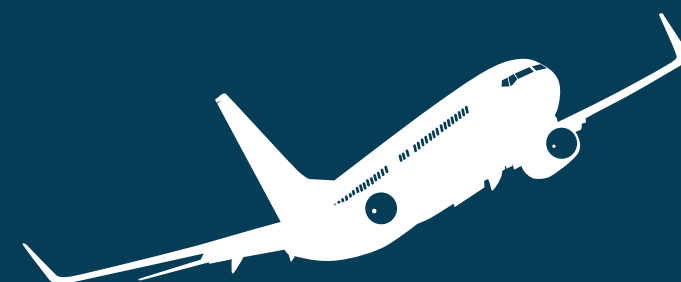
- Cantidad de Variables Totales: 23
- Cantidad de variables numéricas: 18
- Cantidad de variables categóricas: 5
- Cantidad de variables score: 14
- Cantidad de Registros: 103.905
- Frecuencia de Actualización: nunca

CASO DE NEGOCIO

Se busca predecir la variable categórica llamada satisfacción.

El objetivo es que las aerolíneas puedan predecir si el cliente quedó satisfecho o neutral/desatisfecho para así poder mejorar el servicio o remediar una mala experiencia del cliente.

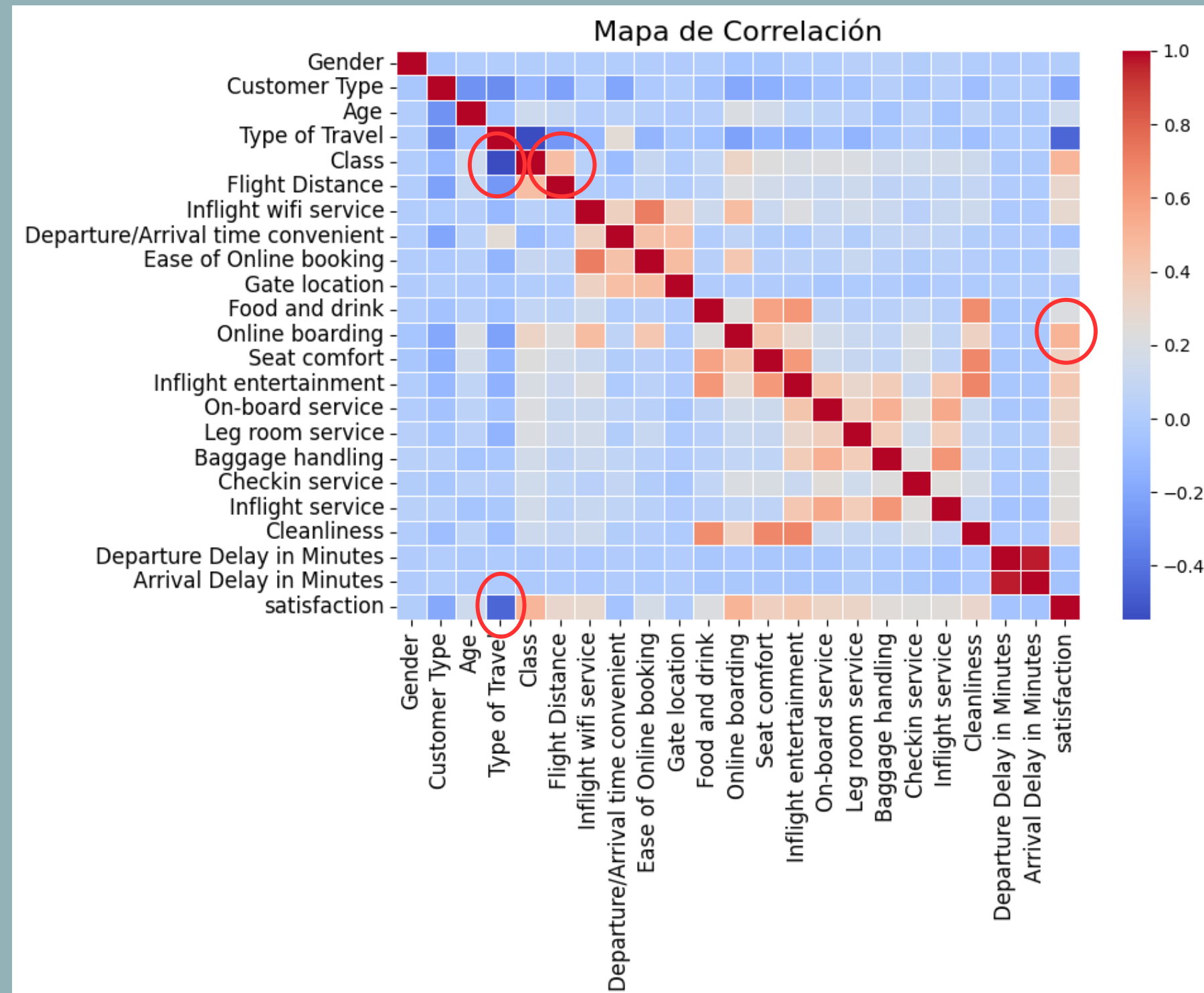




EDA

ITBA

Mapa de correlación

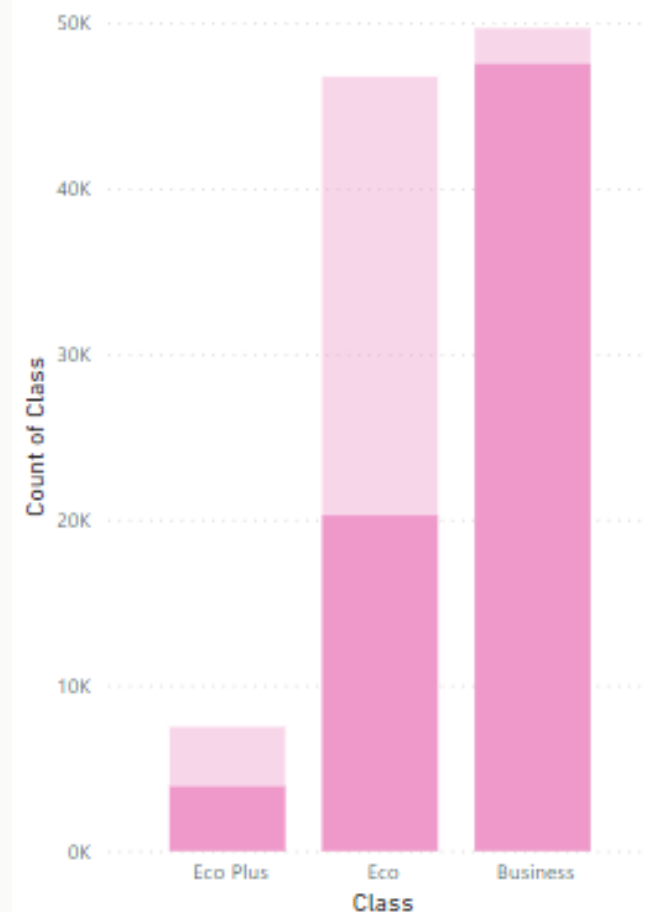


Class y Type of Travel

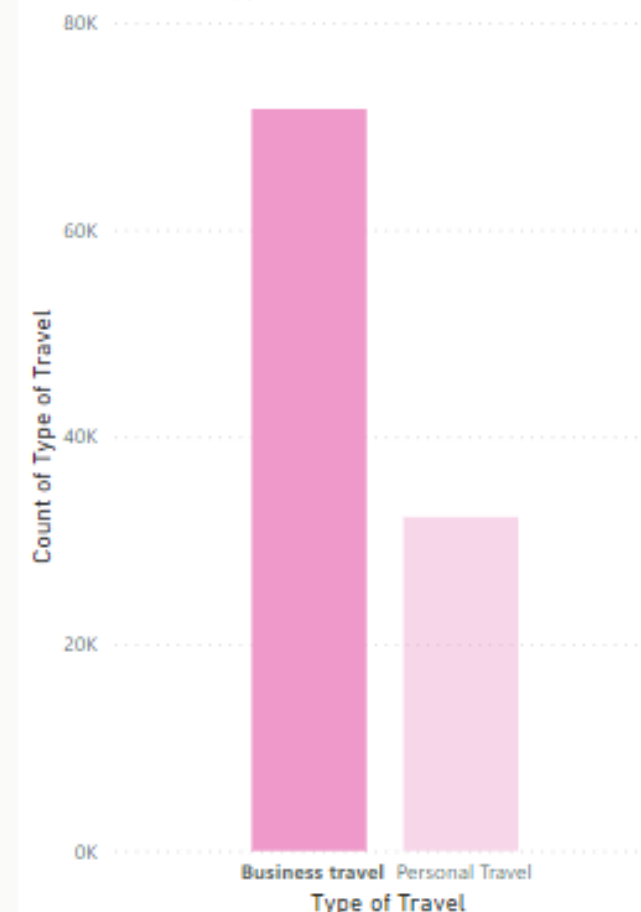
Tamaño del efecto: 0.5540544

Relación entre Type of Travel y Class

Frecuencia de Class



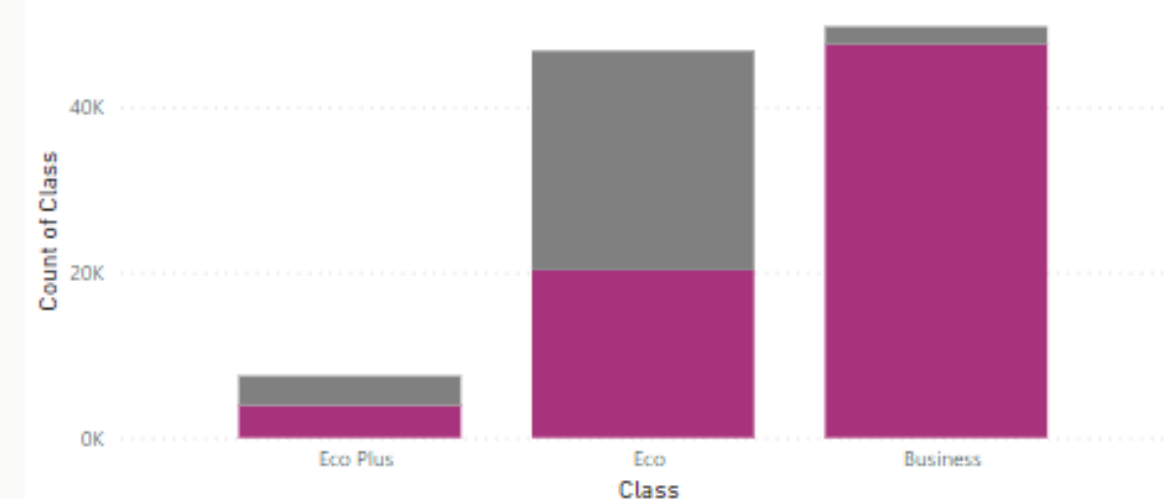
Frecuencia de Type of Travel



Class	Business travel	Personal Travel
Business	47508	2157
Eco	20257	26488
Eco Plus	3890	3604

Class y Type of Travel

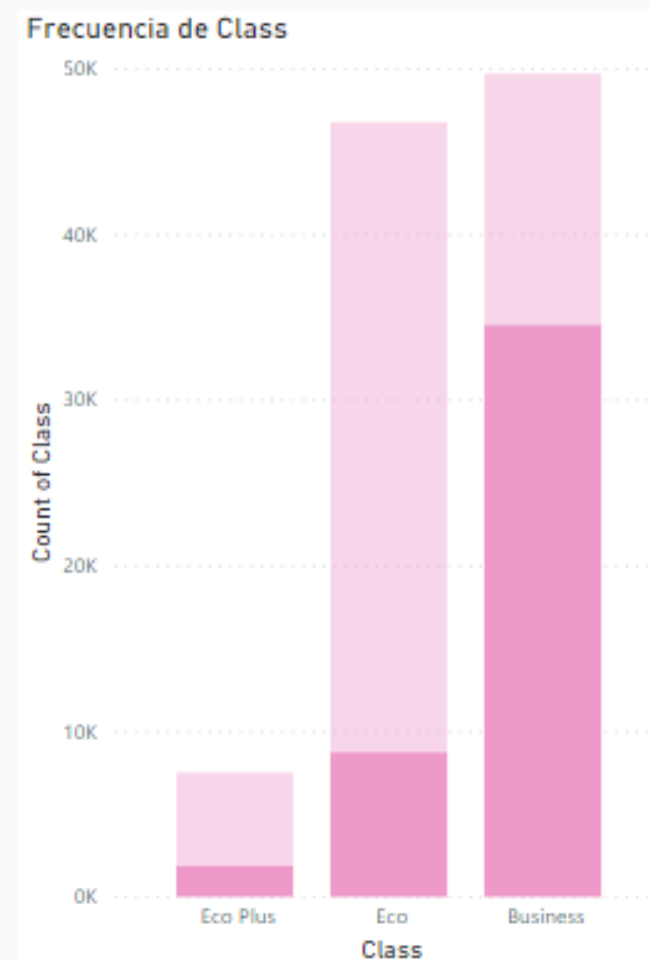
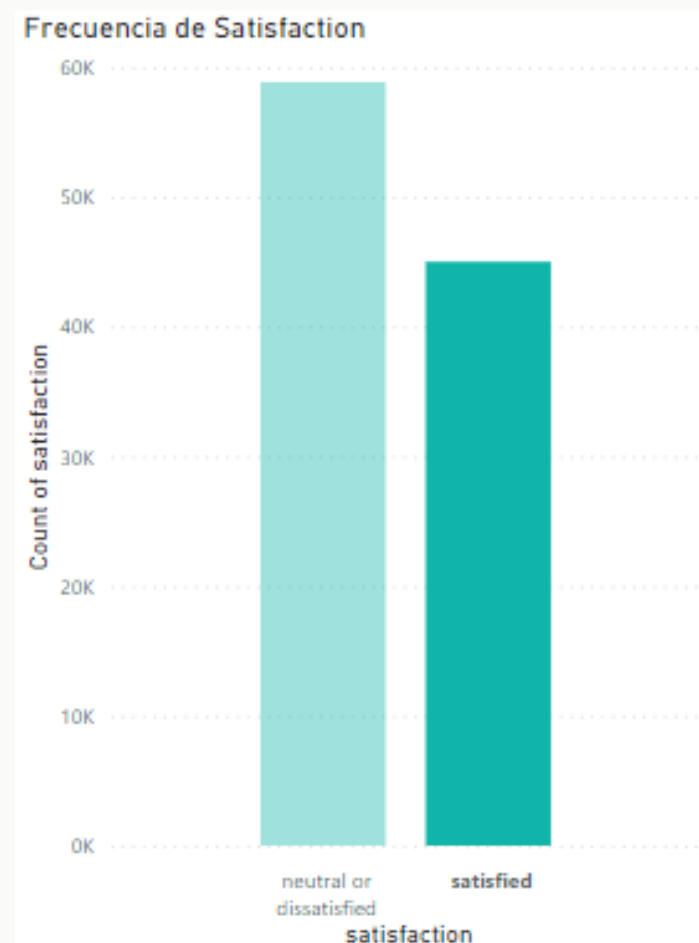
Type of Travel ● Business travel ● Personal Travel



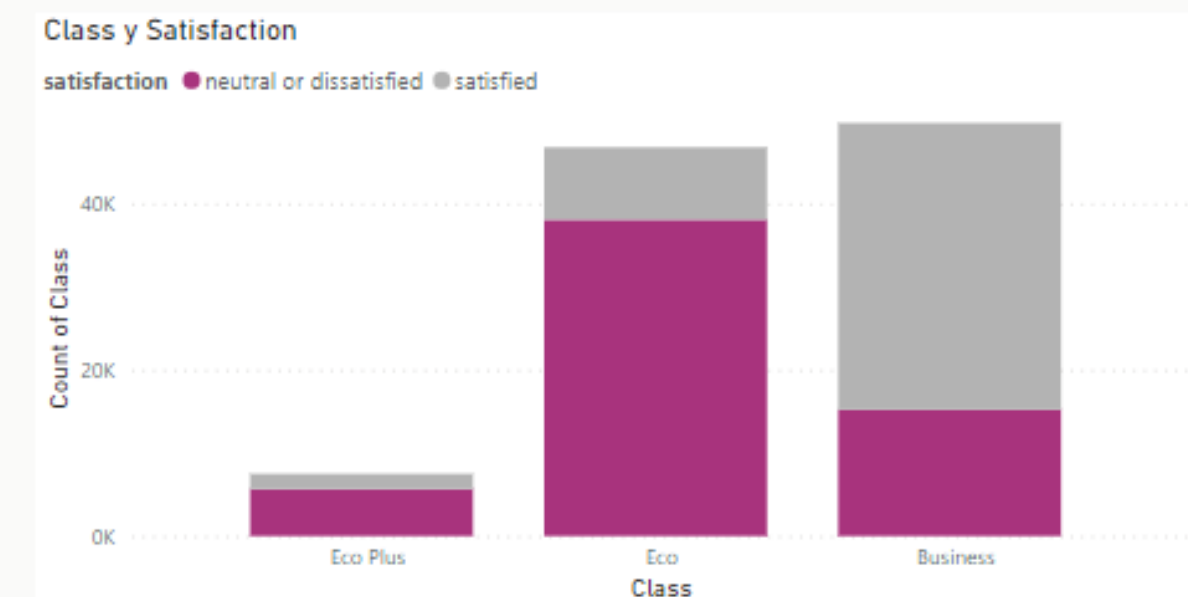
Class y Satisfaction

Tamaño del efecto: **0.5047498**

Relación entre Satisfaction y Class



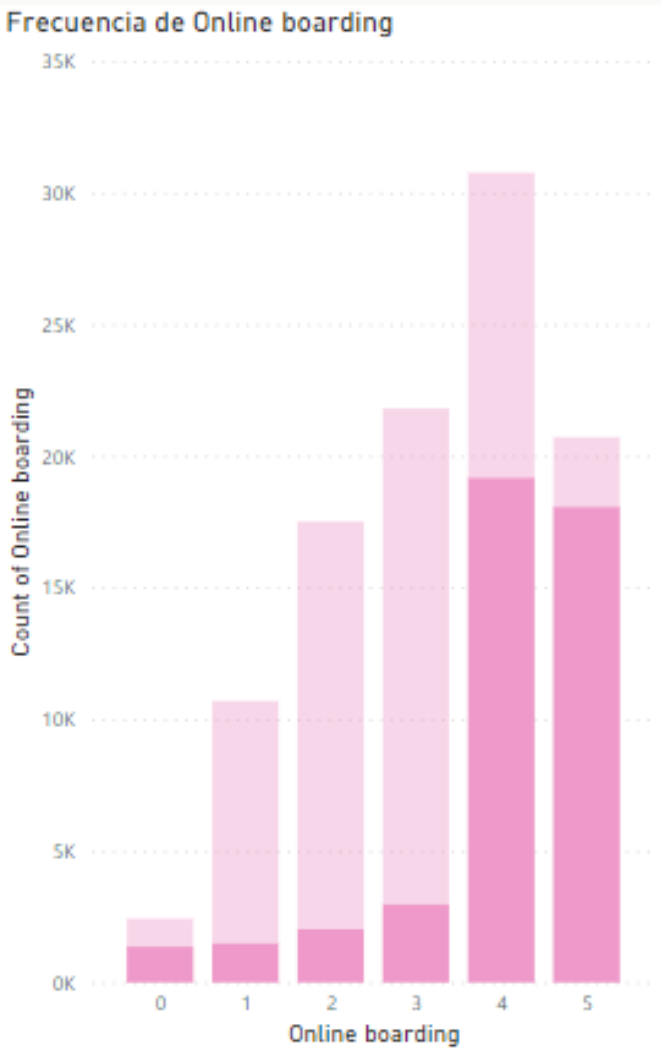
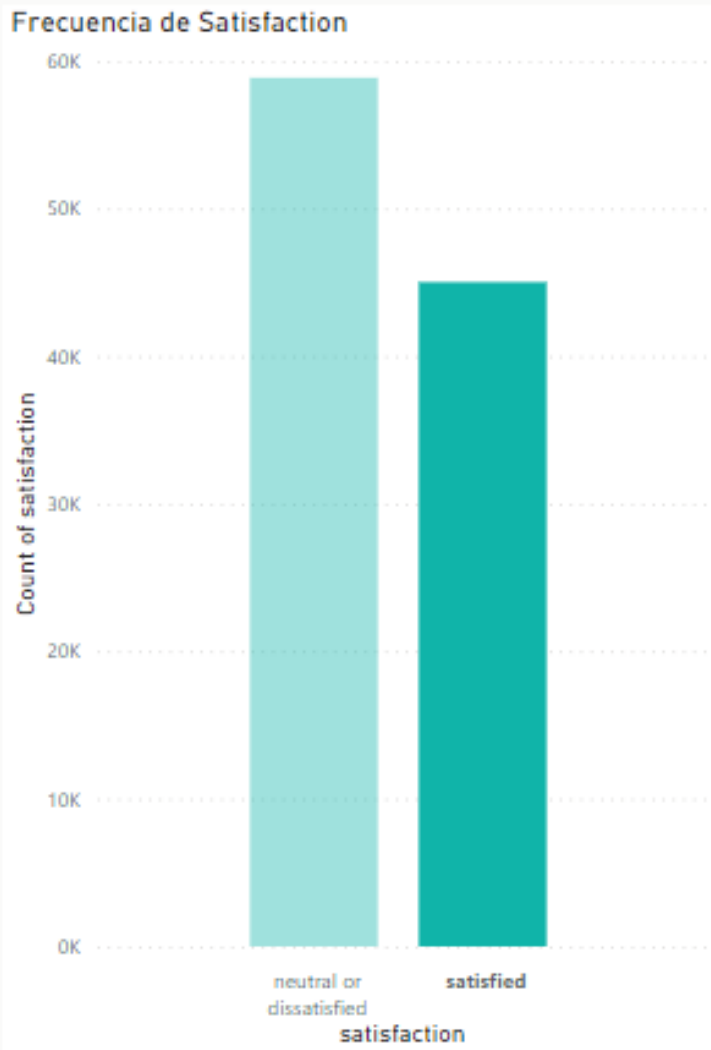
Class	neutral or dissatisfied	satisfied
Business	15185	34480
Eco	38044	8701
Eco Plus	5650	1844



Ease of Online Boarding y Satisfaction

Tamaño del efecto: 0.6185259

Relación entre Satisfaction y Online Booking



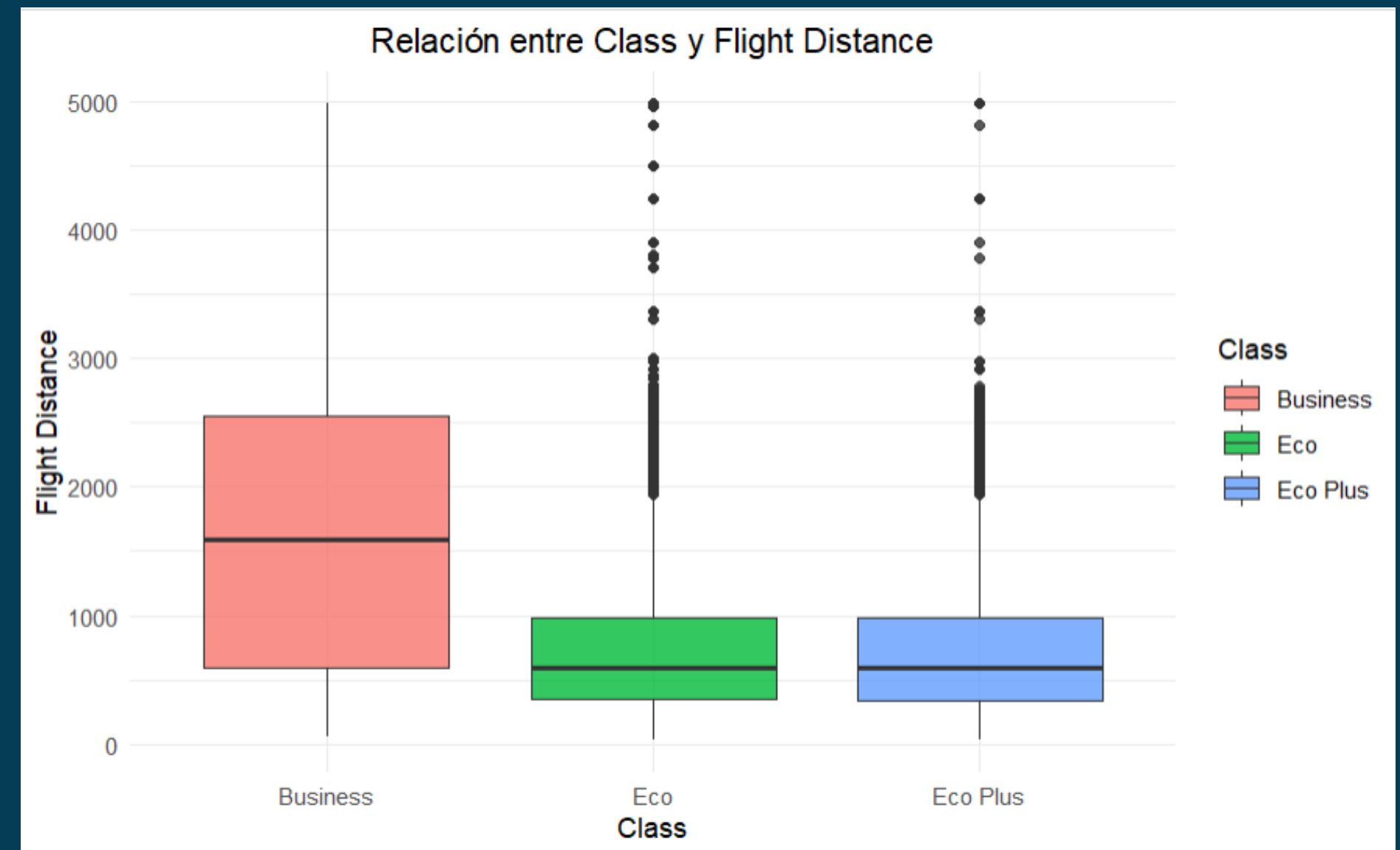
Online boarding	neutral or dissatisfied	satisfied
0	1077	1351
1	9219	1473
2	15486	2019
3	18845	2959
4	11596	19166
5	2656	18057



Flight distance y Class



Kruskal-wallis chi-squared
0.1707619



Métricas Generales

Departure Delay in Minutes

Mínimo	Máximo	Varianza	Mediana	Promedio
0	1592	1461.59	0	14.82

Age

Mínimo	Máximo	Varianza	Mediana	Promedio
7	85	228.46	40	39.38

Arrival Delay in Minutes

Mínimo	Máximo	Varianza	Mediana	Promedio
0	1584	1497.57	0	15.18

Flight Distance

Mínimo	Máximo	Varianza	Mediana	Promedio
31	4983	994293.13	843	1189.45

ANÁLISIS DE MISSINGS

- Arrival Delay in Minutes: 310

Departure Delay in Minutes	Arrival Delay in Minutes	Atraso/Adelanto
25	18	7
1	6	-5
0	0	0
11	9	2
0	0	0
0	0	0
9	23	-14
4	0	4
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
28	8	20
0	0	0
43	35	8
1	0	1
49	51	-2
0	10	-10
7	5	2
17	18	-1
0	4	-4

Arrival Delay in Minutes - Departure Delay in Minutes

Promedio =0



ANÁLISIS DE OUTLIERS

Flight Distance: 2.291

Departure Delay: 14.529

Arrival Delay: 13.954

[illegible]

Departure.Delay.in.Minutes	Arrival.Delay.in.Minutes
1592	1584
1305	1280
1017	1011
978	970
933	920
930	952
921	924
859	860
853	823
750	729
748	720
729	717
726	691
724	705
692	702
652	638
626	604
610	593





TRATAMIENTO DE LA BASE



¿La variable ID tiene valores duplicados?

COMO LA VARIABLE ID NO TIENE DUPLICADOS SE
ELIMINÓ



Se agregaron las variables Arrival Delay Indicator, Departure Delay Indicator

```
#Creo dos nuevas variables que marquen si hubo atraso en el aterrizaje o en el despegue  
  
df['Arrival Delay Indicator'] = df['Arrival Delay in Minutes'].apply(lambda x: 1 if x > 0 else 0)  
df['Departure Delay Indicator'] = df['Departure Delay in Minutes'].apply(lambda x: 1 if x > 0 else 0)
```




Se reemplazó a la Variable Arrival Delay in Minutes por Atraso/Adelanto

```
#Reemplazo la variable de arrival delay por la resta entre departure delay y arrival  
df['Arrival Delay in Minutes']=df['Departure Delay in Minutes'] - df['Arrival Delay in Minutes']  
print(df['Arrival Delay in Minutes'])  
  
#Agrego otra variable que indique si el vuelo duró menos o igual a lo esperado  
df['Atraso/Adelanto Indicator'] = df['Atraso/Adelanto'].apply(lambda x: 1 if x >= 0 else 0)
```

Atraso...



7

-5

0

2

0

0

-14

4

0

0

0

0

20

0



Nueva variable Grupo Edad

5 categorías:

- Grupo 0 → 0-12 años
- Grupo 2 → 13-19 años
- Grupo 4 → 20-40 años
- Grupo 6 → 41-65 años
- Grupo 8 → 65-100 años



Encoding de variables categóricas

1- Creación de Dummies

```
# Paso la variable Gender, costumer type, Type of travel a una dummy
df['Gender'] = pd.get_dummies(df['Gender'], prefix='Gender', drop_first=True).astype(int)
df['Customer Type'] = pd.get_dummies(df['Customer Type'], prefix='Customer Type', drop_first=True).astype(int)
df['Type of Travel'] = pd.get_dummies(df['Type of Travel'], prefix='Type of Travel', drop_first=True).astype(int)
```

2- Encoding Ordinal

```
# Definir el diccionario de codificación ordinal
encoding_dict = {'Eco': 1, 'Business': 3, 'Eco Plus': 2}
# Aplicar el encoding ordinal a la columna 'Class'
df['Class'] = df['Class'].map(encoding_dict)
```



Creación de la variable Promedio Ponderado segun grupo etario

Inflight wifi se	Departure/Arr	Ease of Online	Gate location	Food and drink	Online boardi	Seat comfort	Inflight entert	On-board serv	Leg room serv	Baggage hand	Checkin servic	Inflight service	Cleanliness
3	4	3	1	5	3	5	5	4	3	4	4	5	5
3	2	3	3	1	3	1	1	1	5	3	1	4	1
2	2	2	2	5	5	5	5	4	3	4	4	4	5
2	5	5	5	2	2	2	2	2	5	3	1	4	2
3	3	3	3	4	5	5	3	3	4	4	3	3	3
3	4	2	1	1	2	1	1	3	4	4	4	4	1
2	4	2	3	2	2	2	2	3	3	4	3	5	2
4	3	4	4	5	5	5	5	5	5	5	4	5	4
1	2	2	2	4	3	3	1	1	2	1	4	1	2
3	3	3	4	2	3	3	2	2	3	4	4	3	2
4	5	5	4	2	5	2	2	3	3	5	3	5	2
2	4	2	2	1	2	1	1	1	2	5	5	5	1
1	4	4	4	1	1	1	1	1	1	3	4	4	1
4	2	4	3	4	4	4	4	4	5	2	2	2	4
3	2	3	2	2	3	2	2	4	3	2	2	1	2
2	1	2	3	4	2	1	4	2	1	4	1	3	4
3	3	3	3	4	4	4	4	5	3	4	5	4	4



Creación de la variable Promedio Ponderado segun grupo etario

Niños

0,Inflight wifi service	0.221155...
0,Departure/Arrival time c...	-0.01906...
0,Ease of Online booking	0.167323...
0,Gate location	-0.00805...
0,Food and drink	0.070595...
0,Online boarding	0.217978...
0,Seat comfort	0.055102...
0,Inflight entertainment	0.068722...
0,On-board service	0.045976...
0,Leg room service	0.007050...
0,Baggage handling	0.025957...
0,Checkin service	0.044818...
0,Inflight service	0.033806...
0,Cleanliness	0.068622...

Jóvenes

4,Inflight wifi service	0.096004...
4,Departure/Arrival time c...	-0.00063...
4,Ease of Online booking	0.062184...
4,Gate location	-0.00208...
4,Food and drink	0.081625...
4,Online boarding	0.164681...
4,Seat comfort	0.102525...
4,Inflight entertainment	0.108990...
4,On-board service	0.077032...
4,Leg room service	0.050157...
4,Baggage handling	0.051214...
4,Checkin service	0.064589...
4,Inflight service	0.050192...
4,Cleanliness	0.093518...

Adultos

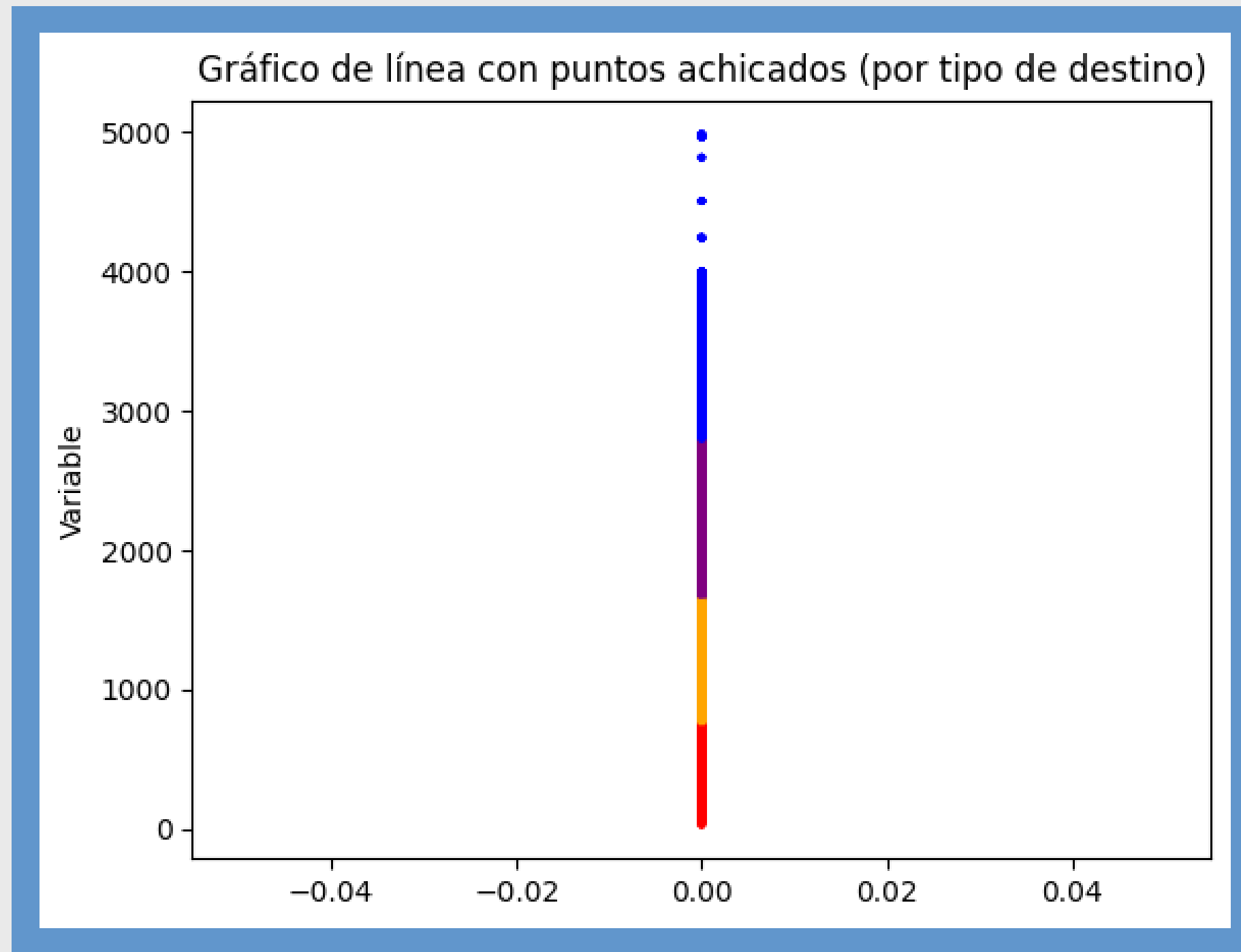
6,Inflight wifi service	0.062273...
6,Departure/Arrival time c...	-0.02200...
6,Ease of Online booking	0.030558...
6,Gate location	0.001755...
6,Food and drink	0.042858...
6,Online boarding	0.116642...
6,Seat comfort	0.093390...
6,Inflight entertainment	0.117437...
6,On-board service	0.103180...
6,Leg room service	0.119927...
6,Baggage handling	0.090629...
6,Checkin service	0.070419...
6,Inflight service	0.090390...
6,Cleanliness	0.082539...

Ancianos

8,Inflight wifi service	0.150080...
8,Departure/Arrival time c...	-0.03929...
8,Ease of Online booking	0.078931...
8,Gate location	-0.00331...
8,Food and drink	0.041008...
8,Online boarding	0.078048...
8,Seat comfort	0.038356...
8,Inflight entertainment	0.127561...
8,On-board service	0.095794...
8,Leg room service	0.149386...
8,Baggage handling	0.089004...
8,Checkin service	0.046342...
8,Inflight service	0.090221...
8,Cleanliness	0.057874...



Agrego la variable TipoDestino

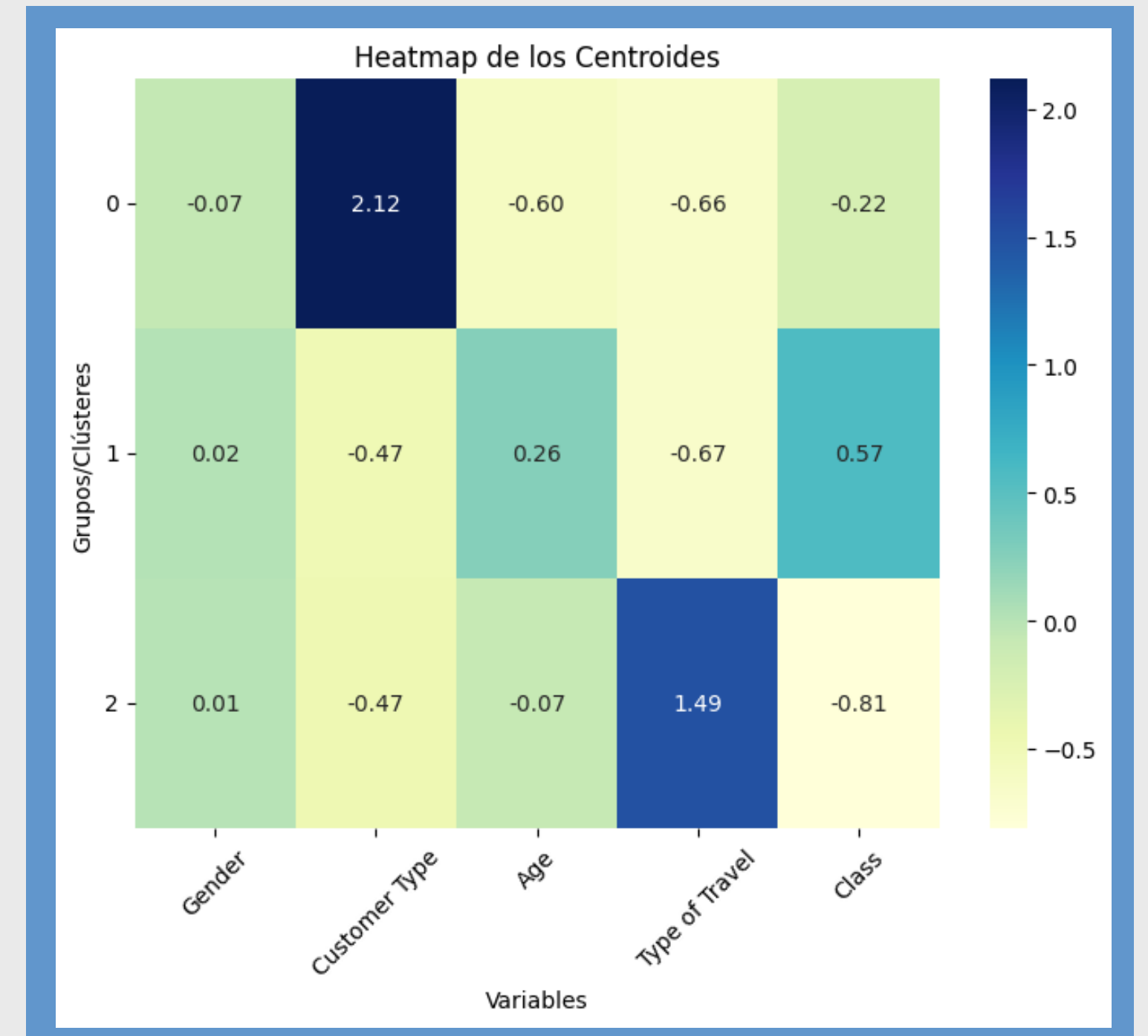




Agrego la variable TipoPasajero

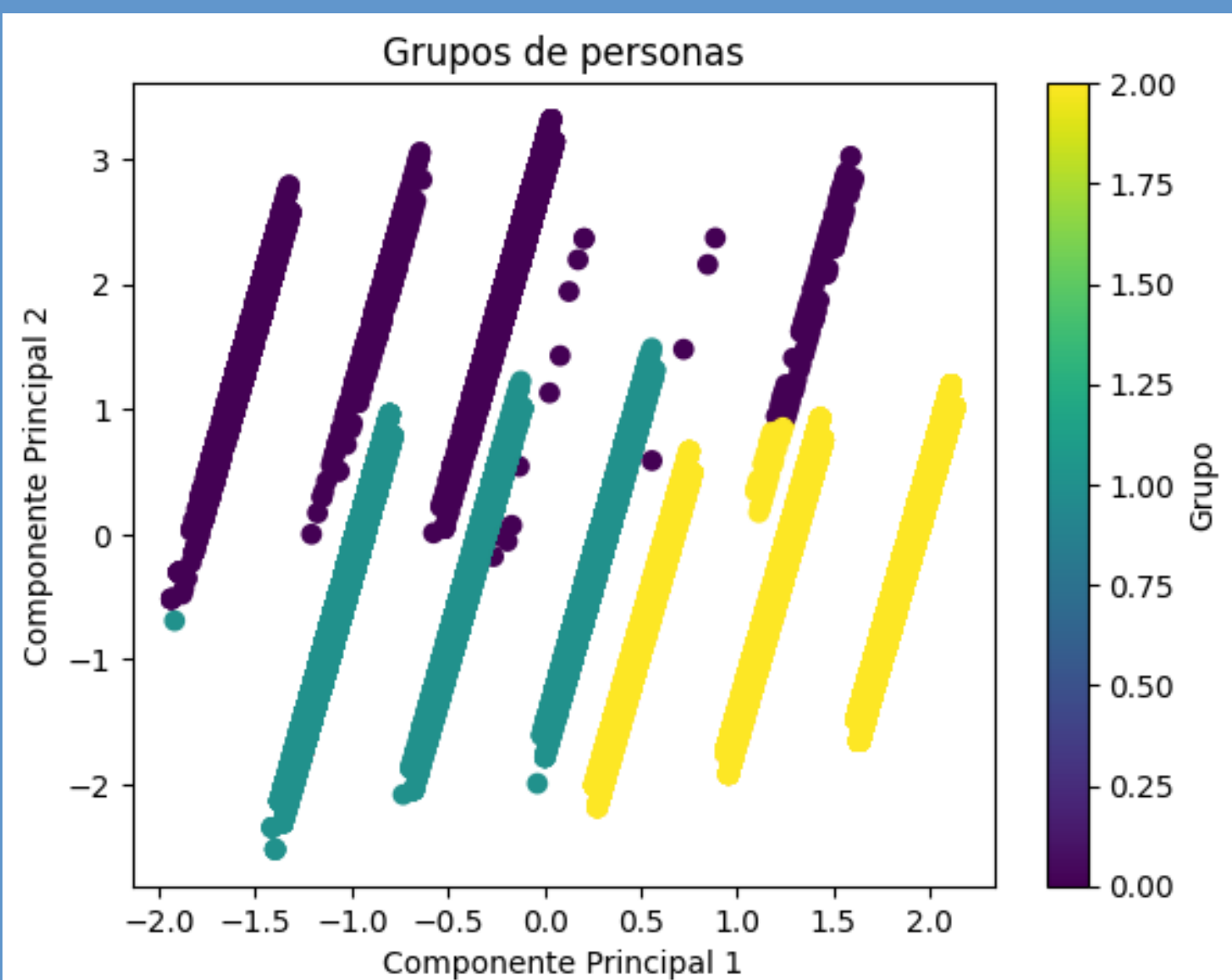
```
# Crear un objeto KMeans con el número deseado de grupos
kmeans = KMeans(n_clusters=3)

# Ajustar el modelo a los datos estandarizados
kmeans.fit(data_scaled)
```

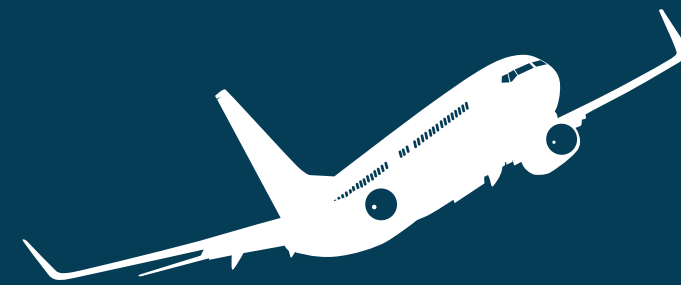




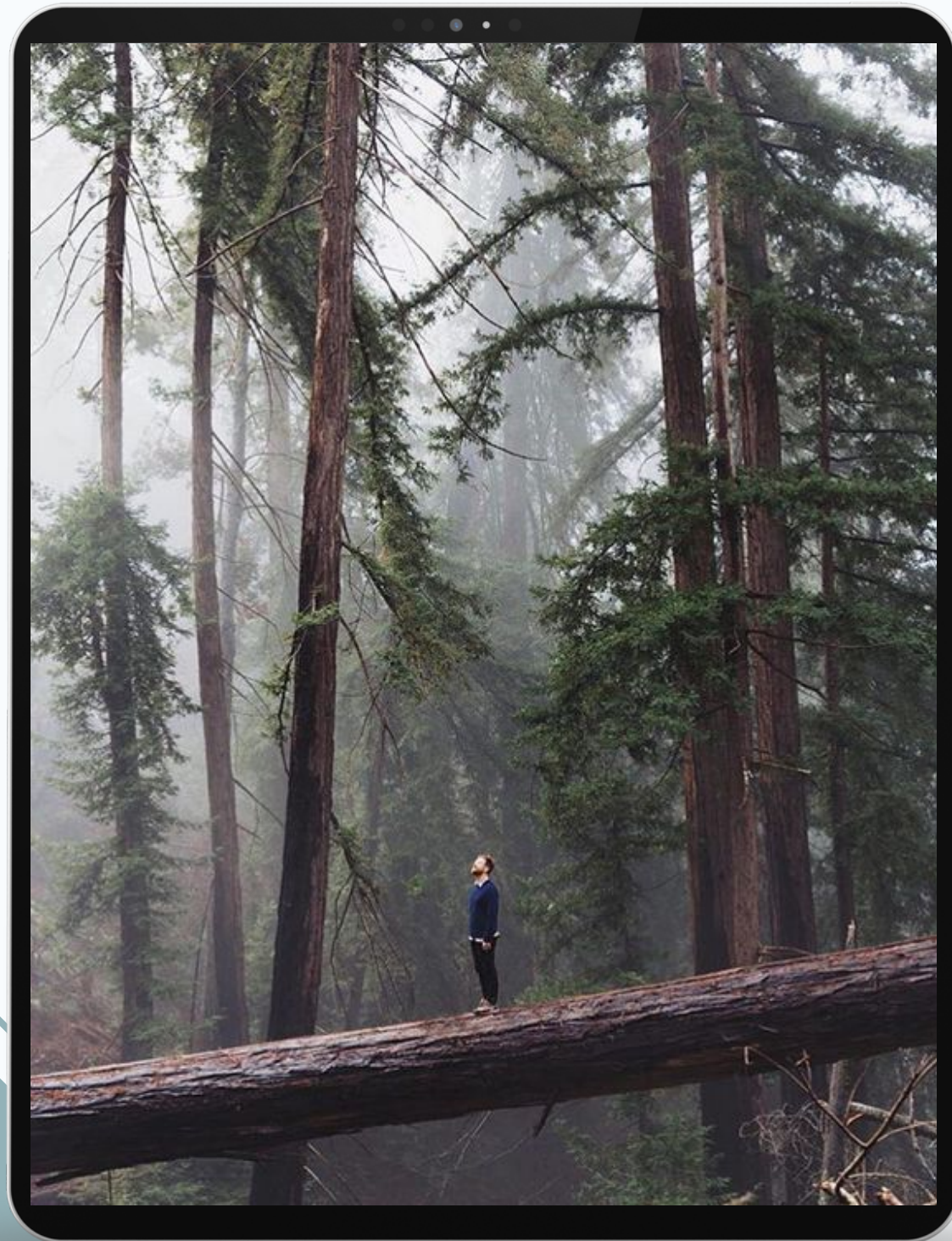
Agrego la variable TipoPasajero



	<i>index</i>	↓ Componente Principal 1	Componente Principal 2
	3 Type of Travel	0.7187308457	-0.1372490682
	0 Gender	0.0082997733	-0.0872834151
	1 Customer Type	-0.2017592637	0.7064088597
	4 Class	-0.6547030205	-0.2536657629
	2 Age	-0.1183817544	-0.6404564797



MODELOS PREDICTORES



Modelos Testeados

- Random Forest
- Regresión Logística
- Catboost
- Naive Bayes
- Extra Trees

Tipo de Partición:



Random Forest

Grid Search!

Hiperparámetros utilizados:

- min_samples_split=5
- n_estimators=300

AUC-ROC: 0.99

Reporte de Clasificación:

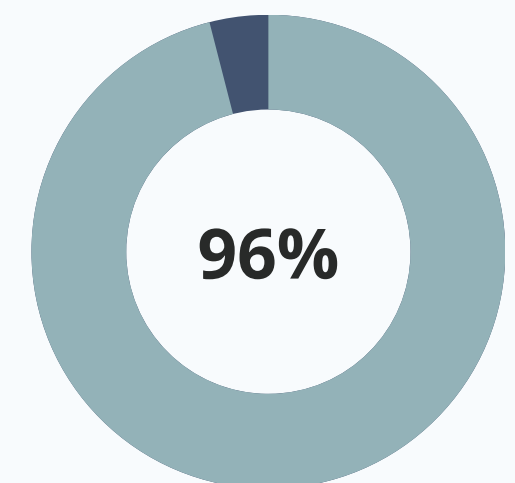
Reporte de clasificación:

	precision	recall	f1-score
0	0.97	0.94	0.96
1	0.96	0.98	0.97

Matriz de confusión

↺ ↻		0	1
	↯		
0		8515	553
1		249	11464

Precisión del modelo:



Extra Trees

Grid Search!

Hiperparámetros utilizados:


- min_samples_split=5
- n_estimators=300

AUC-ROC: 0.99

Reporte de Clasificación:

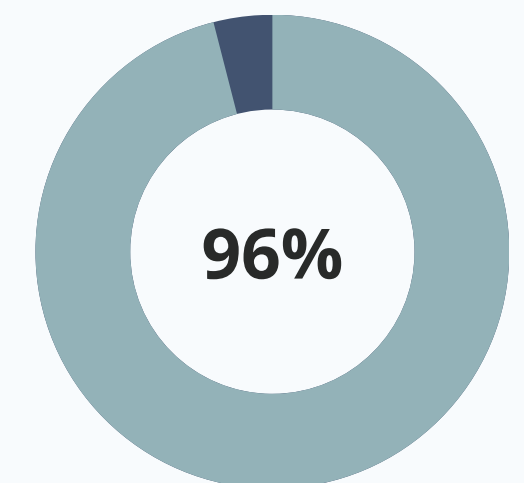
```
Classification Report:
              precision    recall  f1-score
0               0.97       0.94       0.95
1               0.95       0.98       0.97
```

Matriz de confusión



	0	1
0	8510	558
1	253	11460

Precisión del modelo:



Catboost

Grid Search!

Hiperparámetros utilizados:

- iterations=200
- depth=8
- learning_rate=0.1

AUC-ROC: 0.99

Reporte de Clasificación:

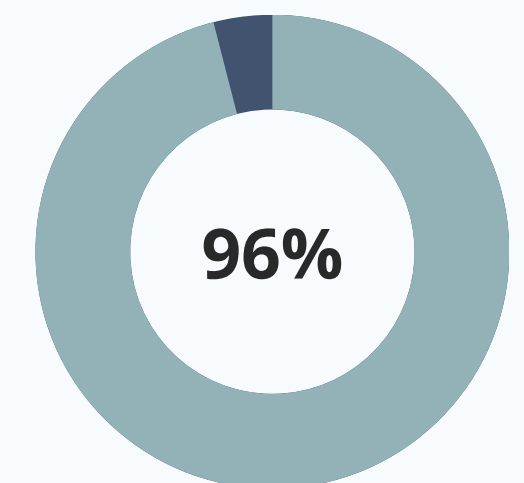
Classification Report:

	precision	recall	f1-score
0	0.98	0.94	0.96
1	0.96	0.98	0.97

Matriz de confusión

	0	1
0	8549	519
1	213	11500

Precisión del modelo:



Conclusiones:

- Los modelos muestran un rendimiento sólido y altamente preciso.
- La matriz de confusión revela que el modelo logra una alta cantidad de predicciones correctas con un número relativamente bajo de predicciones incorrectas.
- El reporte de clasificación muestra altos valores de precisión, recall y F1-score para ambas clases, lo que indica una capacidad confiable para clasificar correctamente las instancias positivas y negativas. El recall de la clase 1 es especialmente alto, lo cual ayuda al objetivo principal del trabajo.
- Además, el valor del AUC-ROC de 0.99, el modelo es capaz de mantener un buen equilibrio entre la sensibilidad (recall) y la especificidad (tasa de verdaderos negativos)

Análisis Predictivo



Profesores:

- Ezequiel Martín Eliano Sombory
- Leonardo Andrés Caravaggio
- Francisco Valentini



1er cuatrimestre - 2023
Trabajo final



Sofía Gonzalez del Solar