ITBA
Instituto Tecnológicc
de Buenos Aires

82.05 | Análisis Predictivo

FINAL

AGUSTINA GONZALEZ CRESPO

Airline Passenger Satisfaction

contiene información sobre una encuesta realizada a distintos pasajeros de una aerolínea

Objetivo: predecir si el cliente va a estar satisfecho con el servicio de la aerolínea o le va a ser indiferente/no estar satisfecho

Hipótesis:

- Cuanto más grande sea la distancia del vuelo, mayor es la probabilidad de estar insatisfecho
- Las clases influyen en la satisfacción del cliente

Fuente: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?select=train.csv

Análisis Exploratorio

VARIABLES

Pasajero

- **Gender:** género del pasajero
- **Customer type:** tipo de cliente (regular o no)
- **Age:** edad del pasajero
- **Type of travel:** propósito del viaje (personal o negocio)
- **Class:** business, economy, economy plus

Vuelo

- **Departure delay in minutes:** cantidad de minutos de atraso en la partida del vuelo
- **Arrival delay in minutes:** cantidad de minutos de atraso en el arribo del vuelo
- **Flight distance:** distancia del vuelo

Ratings

- **Inflight wifi service:** satisfacción con el servicio de Wi-fi en el avión (0-5)
- **Departure/Arrival time convenient:** conveniencia del horario de partida/arribo (0-5)
- **Ease of Online booking:** satisfacción con la facilidad del online booking (0-5)
- **Gate location:** satisfacción con la ubicación de la puerta de embarque (0-5)
- **Food and drink:** satisfacción con la comida y bebida (0-5)
- **Online boarding:** satisfacción con el online boarding (0-5)
- **Seat comfort:** satisfacción con el asiento (0-5)
- **Inflight entertainment:** satisfacción con el entretenimiento abordo(0-5)
- **On-board service:** satisfacción con el servicio previo al vuelo(0-5)
- **Leg room service**: satisfacción con el espacio para las piernas (0-5)
- **Baggage handling:** satisfacción con el servicio de valijas (0-5)
- **Checkin service:** satisfacción con el check-in (0-5)
- **Inflight service:** satisfacción con el servicio durante el vuelo (0-5)
- **Cleanliness:** satisfacción con la limpieza (0-5)

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	103904 non-null	int64
1	id	103904 non-null	int64
2	Gender	103904 non-null	object
3	Customer Type	103904 non-null	object
4	Age	103904 non-null	int64
5	Type of Travel	103904 non-null	object
6	Class	103904 non-null	object
7	Flight Distance	103904 non-null	int64
8	Inflight wifi service	103904 non-null	int64
9	Departure/Arrival time convenient	103904 non-null	int64
10	Ease of Online booking	103904 non-null	int64
11	Gate location	103904 non-null	int64
12	Food and drink	103904 non-null	int64
13	Online boarding	103904 non-null	int64
14	Seat comfort	103904 non-null	int64
15	Inflight entertainment	103904 non-null	int64
16	On-board service	103904 non-null	int64
17	Leg room service	103904 non-null	int64
18	Baggage handling	103904 non-null	int64
19	Checkin service	103904 non-null	int64
20	Inflight service	103904 non-null	int64
21	Cleanliness	103904 non-null	int64
22	Departure Delay in Minutes	103904 non-null	int64
23	Arrival Delay in Minutes	103594 non-null	float64
24	satisfaction	103904 non-null	object

Base de datos:

- 103,904 registros
- 25 columnas

Variables numéricas:

- 20 variables

Variables categóricas:

- 5 variables

#	Column	Non-Null Count	Dtype
0	Gender	103904 non-null	object
1	Customer_Type	103904 non-null	object
2	Age	103904 non-null	int64
3	Type_of_Travel	103904 non-null	object
4	Class	103904 non-null	object
5	FlightDistance	103904 non-null	int64
6	Inflight-wifiService	103904 non-null	int64
7	Departure/Arrival_TimeConvenience	103904 non-null	int64
8	OnlineBooking_Ease	103904 non-null	int64
9	GateLocation	103904 non-null	int64
10	Food/Drink	103904 non-null	int64
11	OnlineBoarding	103904 non-null	int64
12	SeatComfort	103904 non-null	int64
13	InflightEntertainment	103904 non-null	int64
14	On-boardService	103904 non-null	int64
15	Leg-roomService	103904 non-null	int64
16	BaggageHandling	103904 non-null	int64
17	CheckinService	103904 non-null	int64
18	Inflight_service	103904 non-null	int64
19	Cleanliness	103904 non-null	int64
20	DepartureDelay	103904 non-null	int64
21	ArrivalDelay	103594 non-null	float64
22	satisfaction	103904 non-null	object

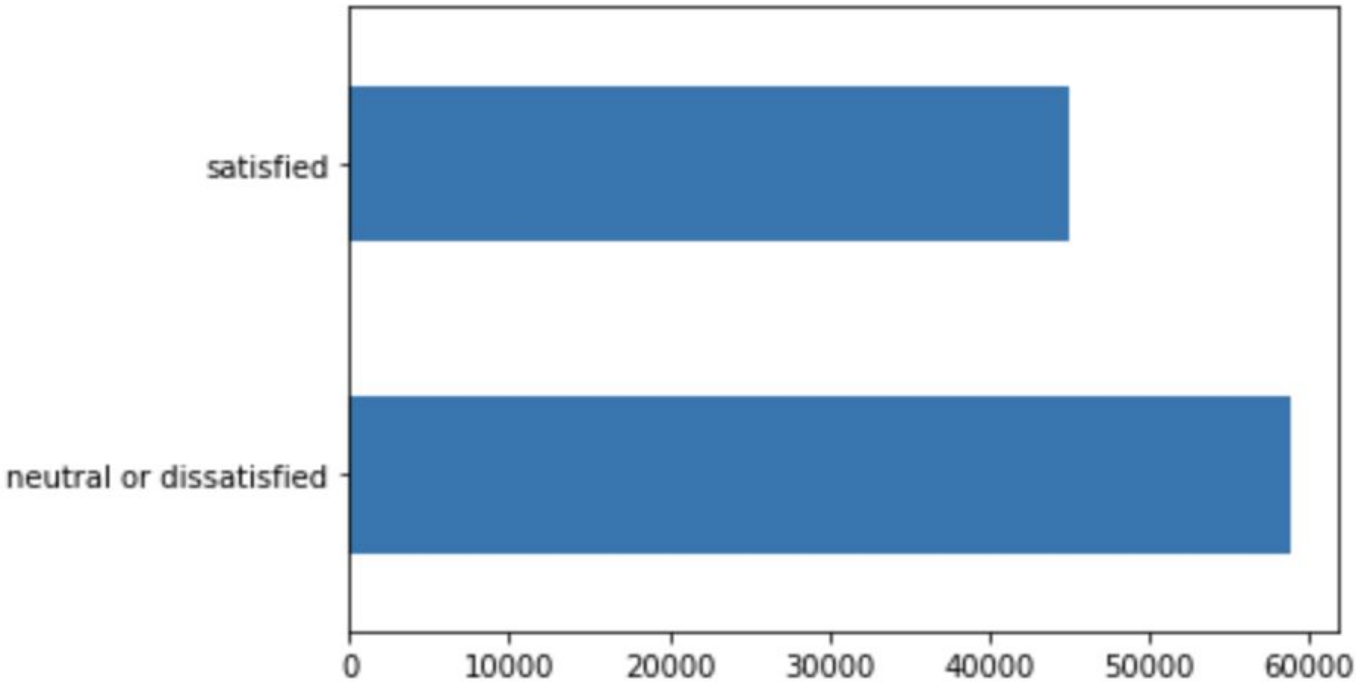
Modificación:

- se reemplazaron los espacios por _

Eliminación:

- unnamed:0
- id

Distribución de la variable target



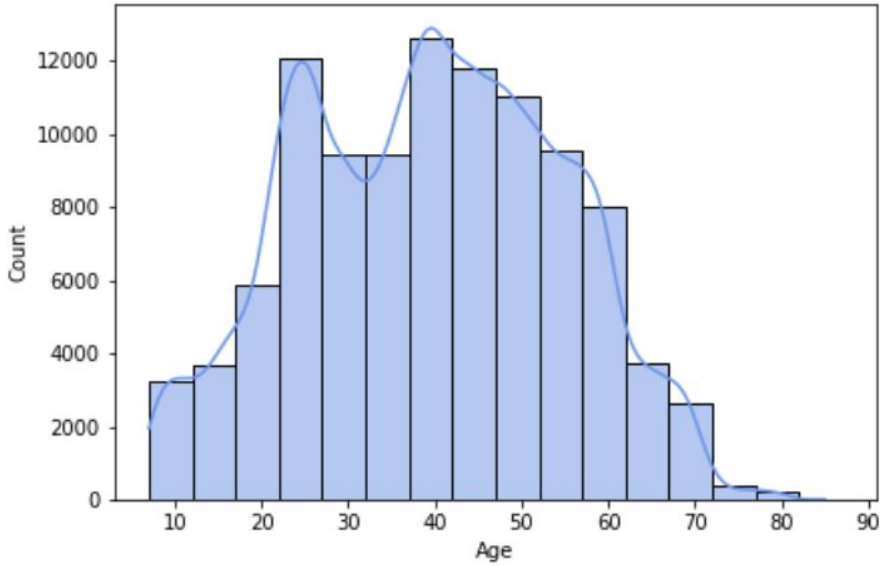
satisfaction:

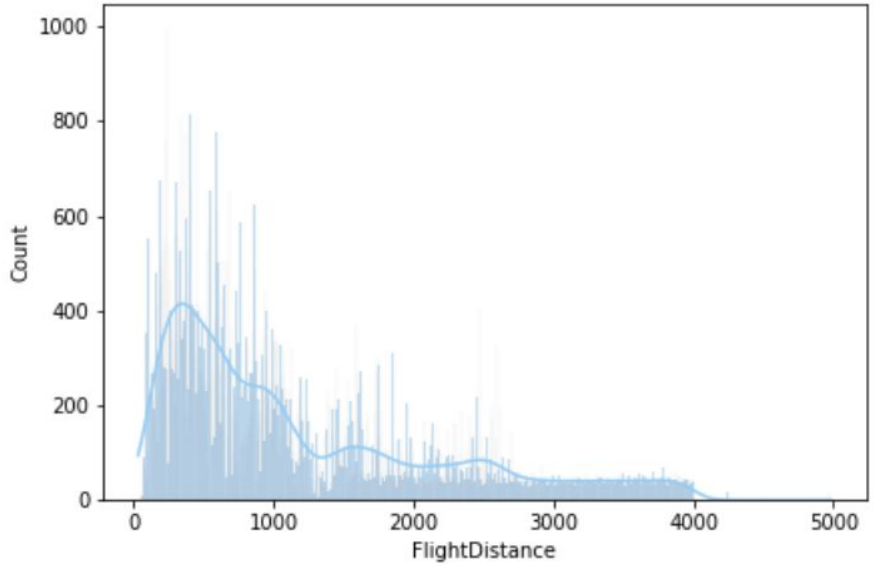
- satisfied **(43,3%)**
- neutral/dissatisfaction **(56,7%)**

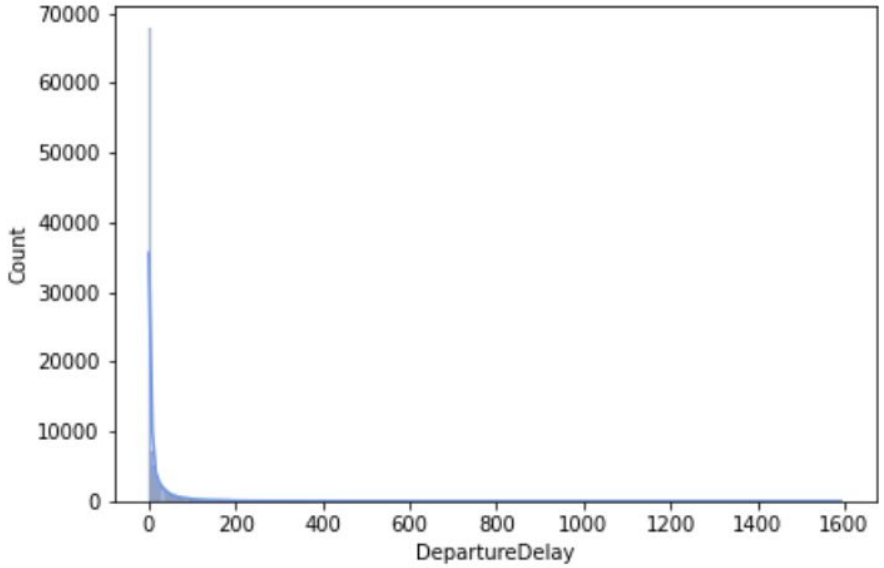
Variables Numéricas

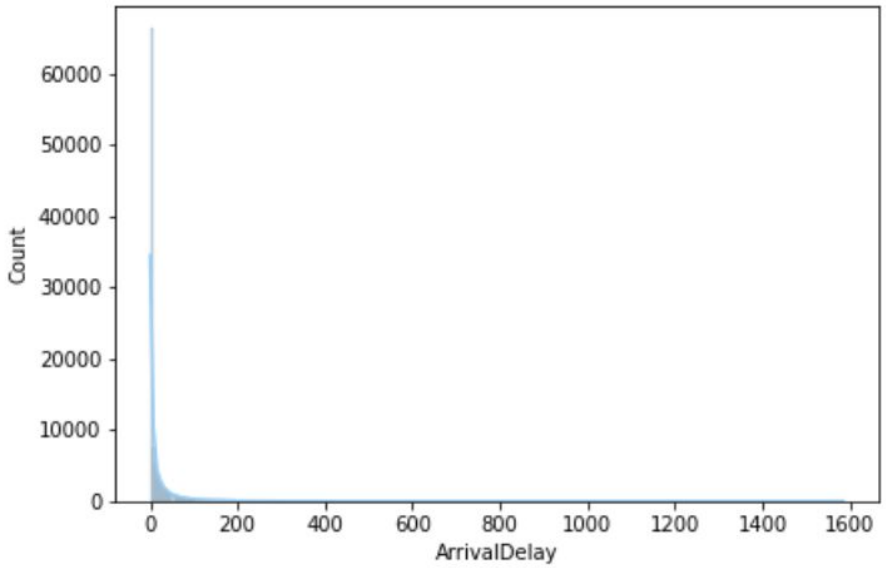
	count	mean	std	min	25%	50%	75%	max
Age	103904.0	39.379706	15.114964	7.0	27.0	40.0	51.0	85.0
FlightDistance	103904.0	1189.448375	997.147281	31.0	414.0	843.0	1743.0	4983.0
Inflight-wifiService	103904.0	2.729683	1.327829	0.0	2.0	3.0	4.0	5.0
Departure/Arrival_TimeConvenience	103904.0	3.060296	1.525075	0.0	2.0	3.0	4.0	5.0
OnlineBooking_Ease	103904.0	2.756901	1.398929	0.0	2.0	3.0	4.0	5.0
GateLocation	103904.0	2.976883	1.277621	0.0	2.0	3.0	4.0	5.0
Food/Drink	103904.0	3.202129	1.329533	0.0	2.0	3.0	4.0	5.0
OnlineBoarding	103904.0	3.250375	1.349509	0.0	2.0	3.0	4.0	5.0
SeatComfort	103904.0	3.439396	1.319088	0.0	2.0	4.0	5.0	5.0
InflightEntertainment	103904.0	3.358158	1.332991	0.0	2.0	4.0	4.0	5.0
On-boardService	103904.0	3.382363	1.288354	0.0	2.0	4.0	4.0	5.0
Leg-roomService	103904.0	3.351055	1.315605	0.0	2.0	4.0	4.0	5.0
BaggageHandling	103904.0	3.631833	1.180903	1.0	3.0	4.0	5.0	5.0
CheckinService	103904.0	3.304290	1.265396	0.0	3.0	3.0	4.0	5.0
Inflight_service	103904.0	3.640428	1.175663	0.0	3.0	4.0	5.0	5.0
Cleanliness	103904.0	3.286351	1.312273	0.0	2.0	3.0	4.0	5.0
DepartureDelay	103904.0	14.815618	38.230901	0.0	0.0	0.0	12.0	1592.0
ArrivalDelay	103594.0	15.178678	38.698682	0.0	0.0	0.0	13.0	1584.0

Distribución de las variables









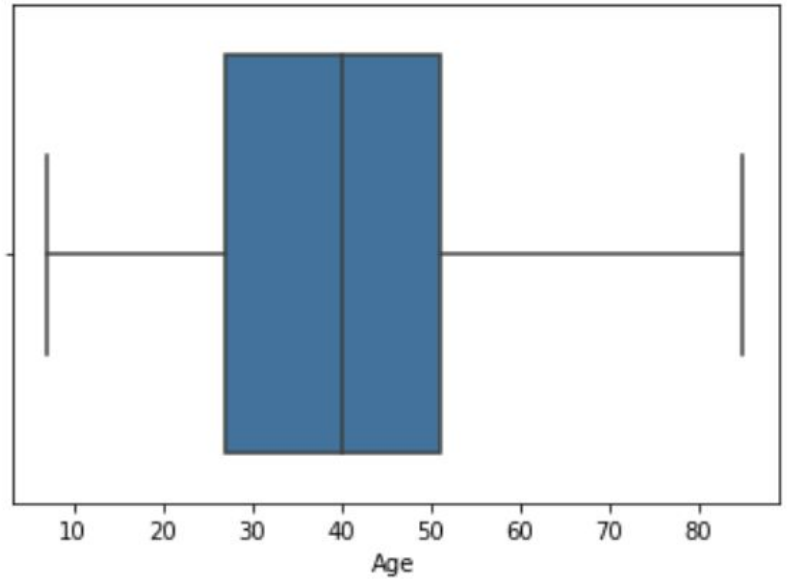
Missings

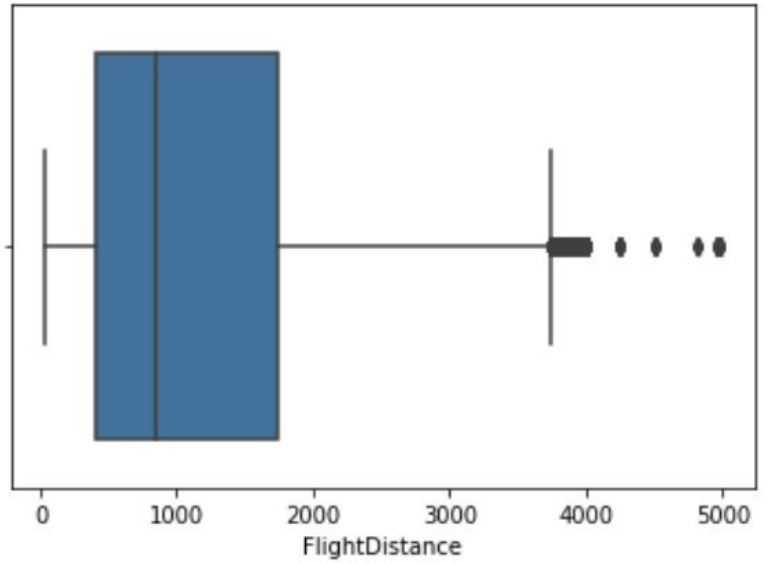
Gender	0
Customer_Type	0
Age	0
Type_of_Travel	0
Class	0
FlightDistance	0
Inflight-wifiService	0
Departure/Arrival_TimeConvenience	0
OnlineBooking_Ease	0
GateLocation	0
Food/Drink	0
OnlineBoarding	0
SeatComfort	0
InflightEntertainment	0
On-boardService	0
Leg-roomService	0
BaggageHandling	0
CheckinService	0
Inflight_service	0
Cleanliness	0
DepartureDelay	0
ArrivalDelay	310
satisfaction	0

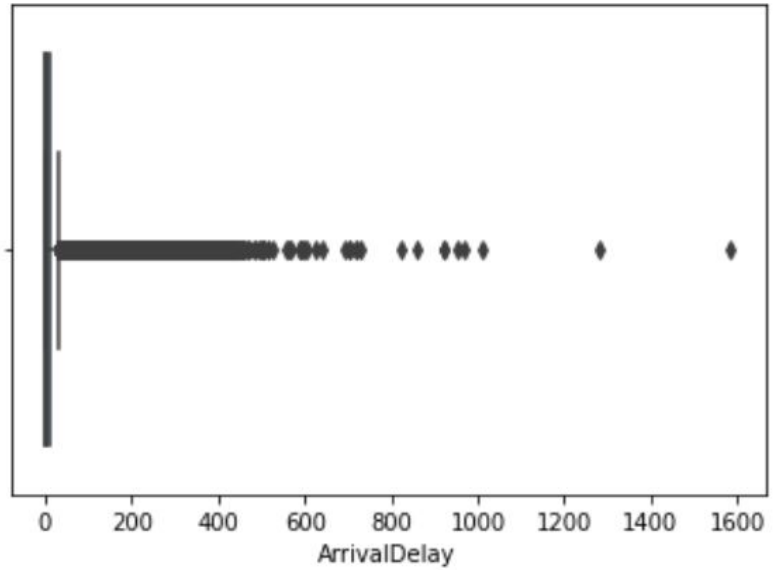
ArrivalDelay:

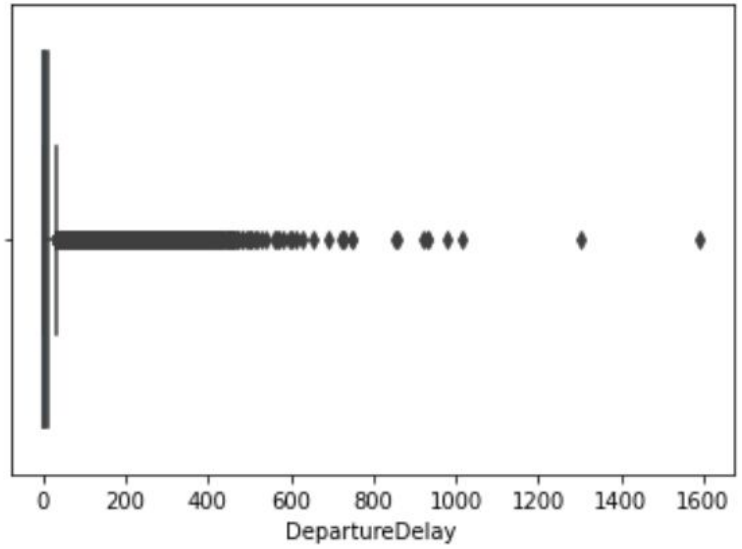
- se imputaron con la mediana

Outliers


10
20
30
40
50
60
70
80
Age

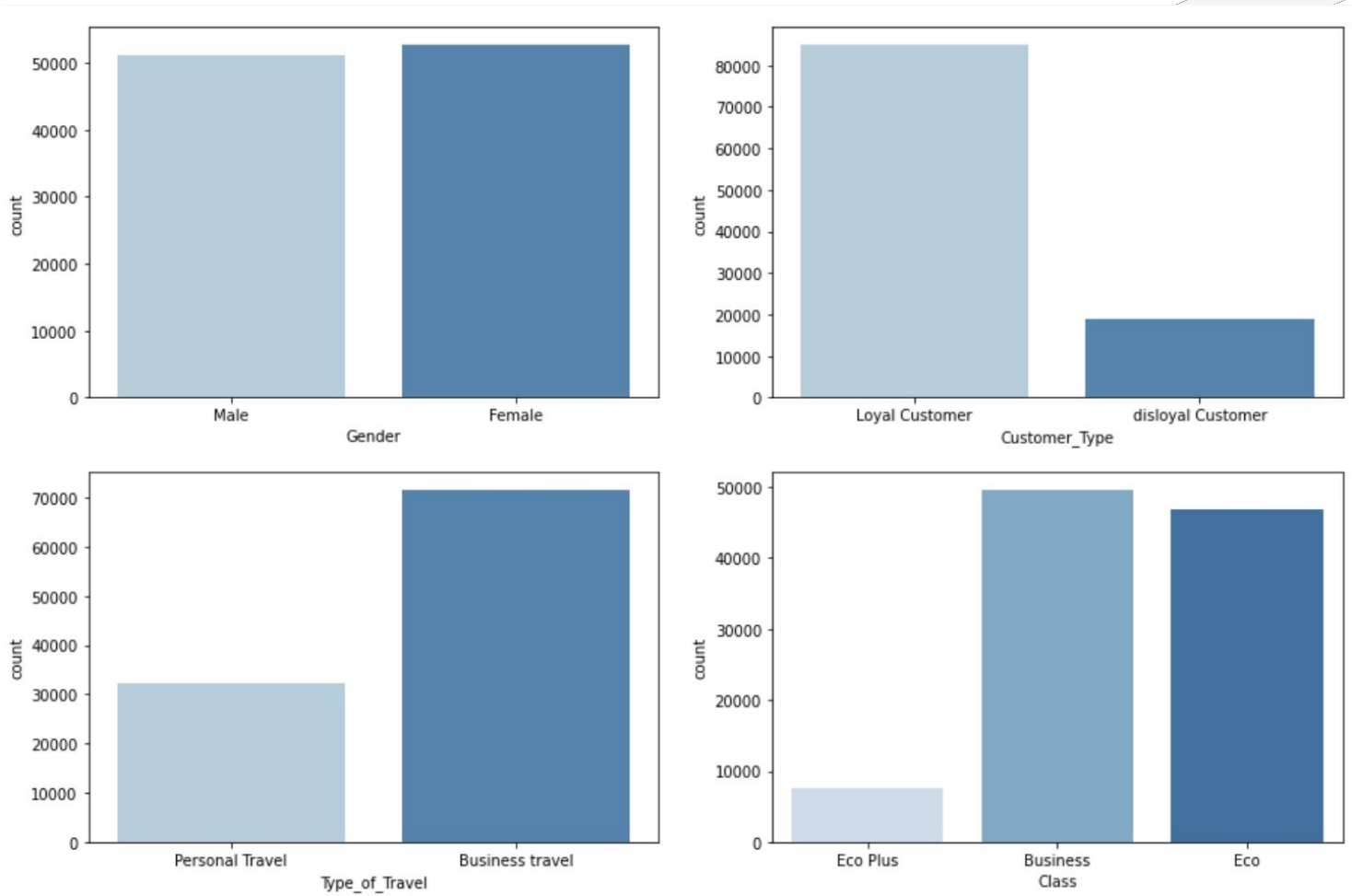

0
1000
2000
3000
4000
5000
FlightDistance


0
200
400
600
800
1000
1200
1400
1600
ArrivalDelay


0
200
400
600
800
1000
1200
1400
1600
DepartureDelay

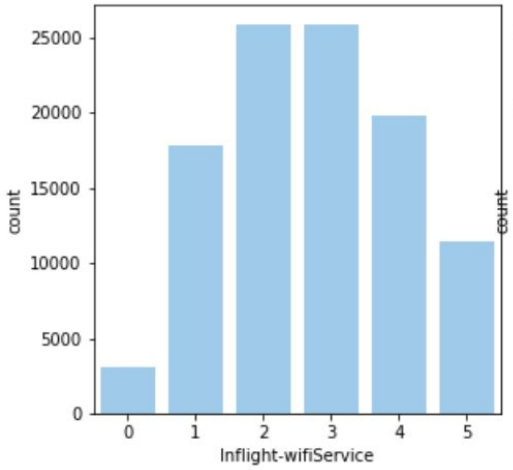
Variables Categóricas

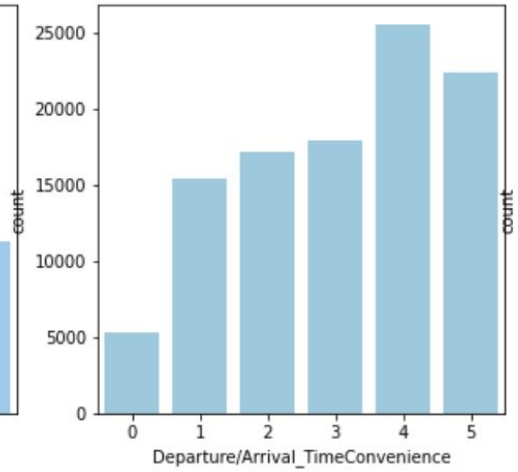
	count	unique	top	freq
Gender	103904	2	Female	52727
Customer_Type	103904	2	Loyal Customer	84923
Type_of_Travel	103904	2	Business travel	71655
Class	103904	3	Business	49665
satisfaction	103904	2	neutral or dissatisfied	58879

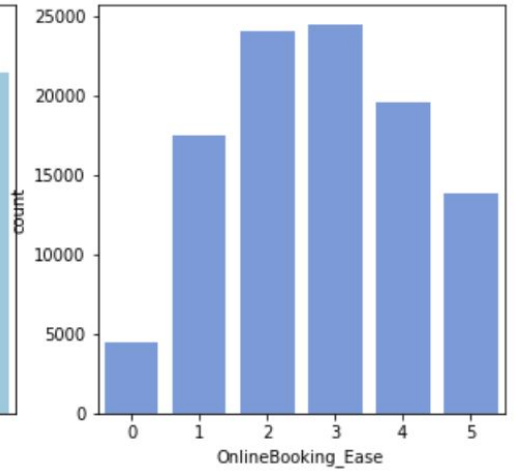


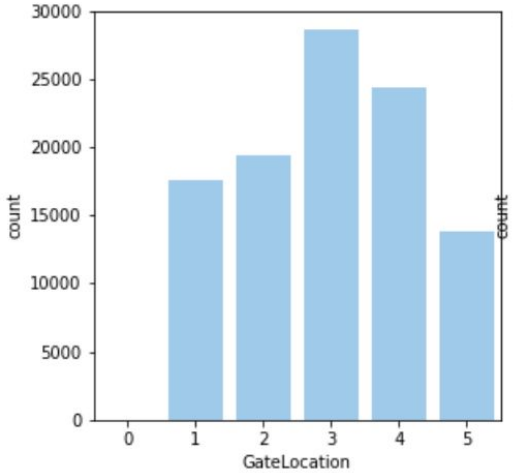
Ratings

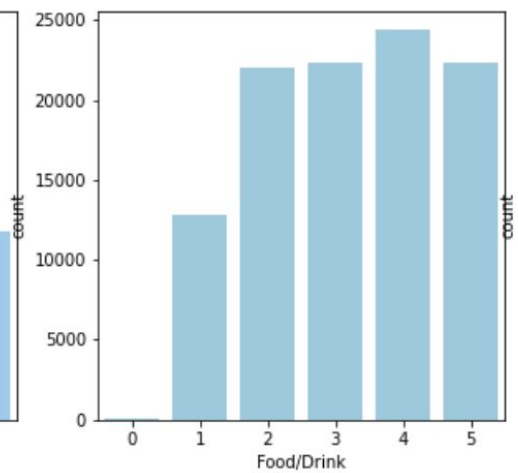
Distribución de las variables

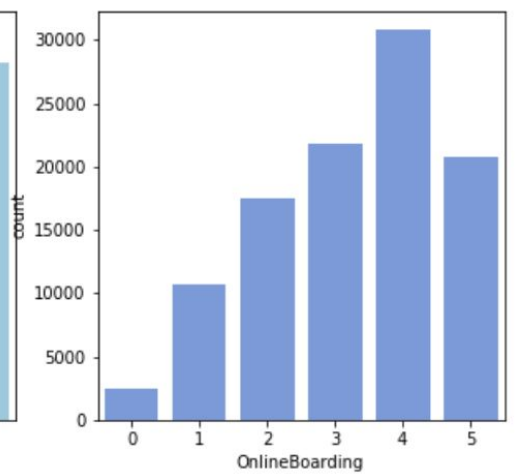


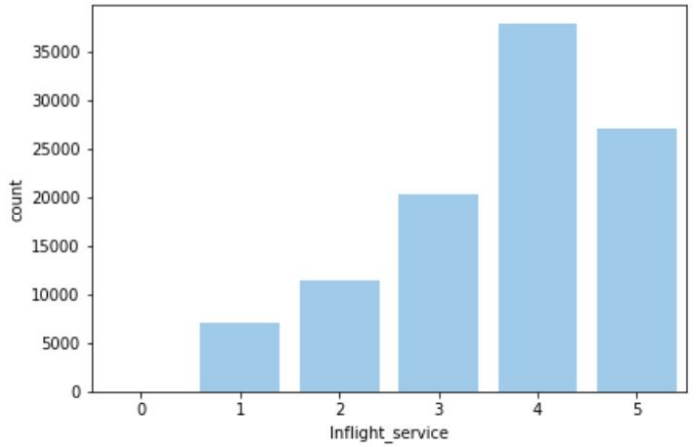


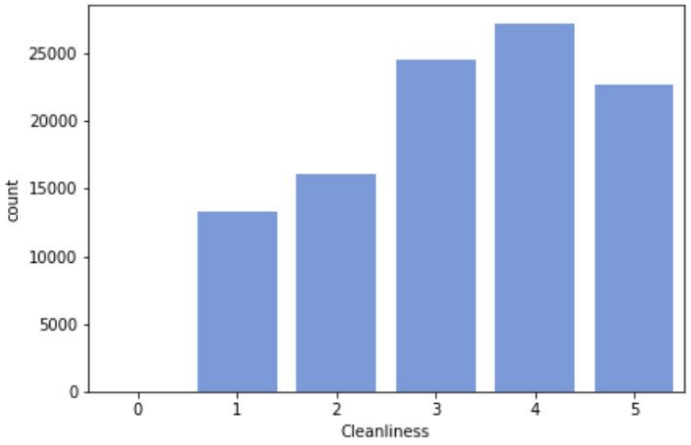




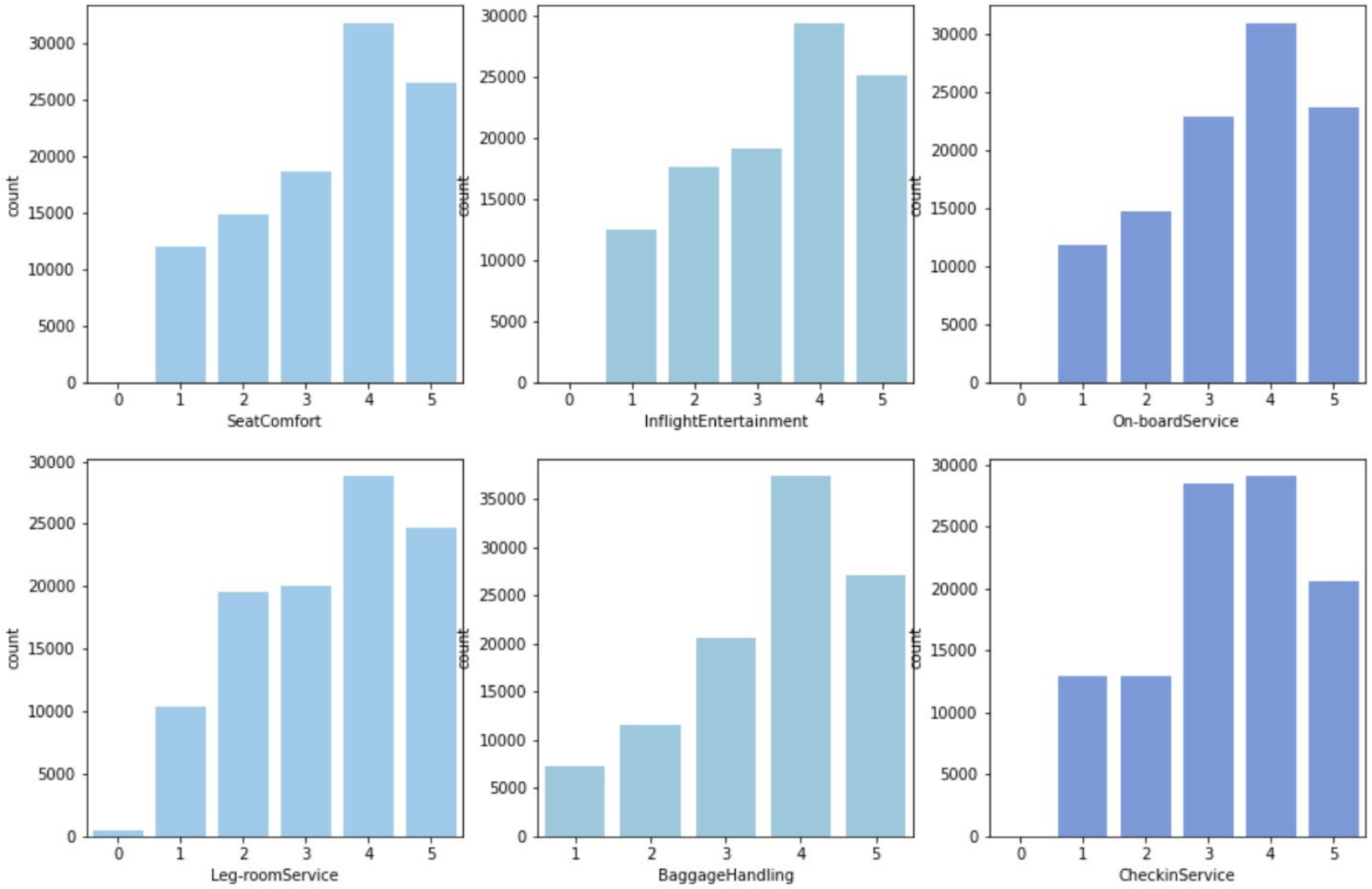




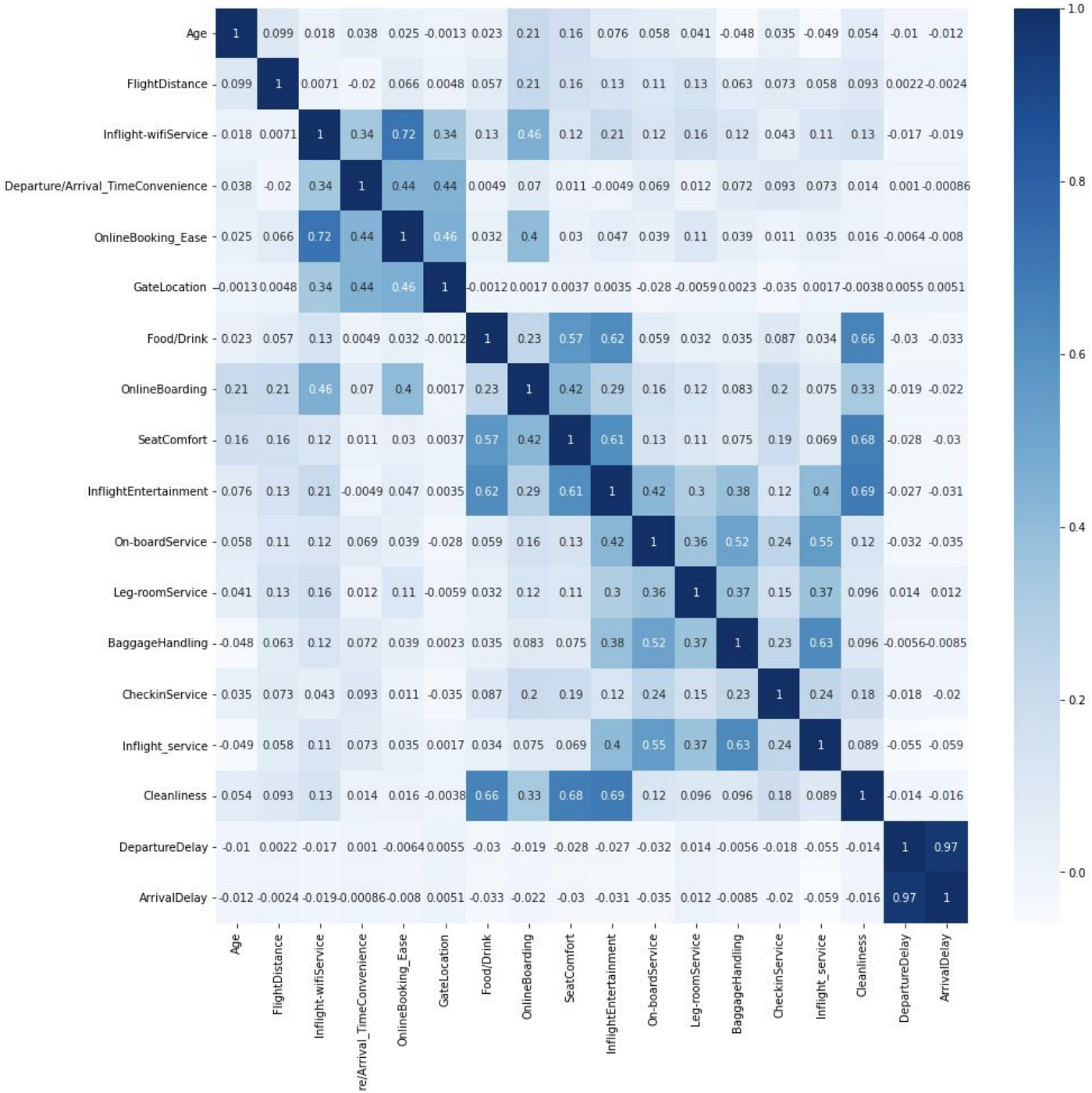




Distribución de las variables

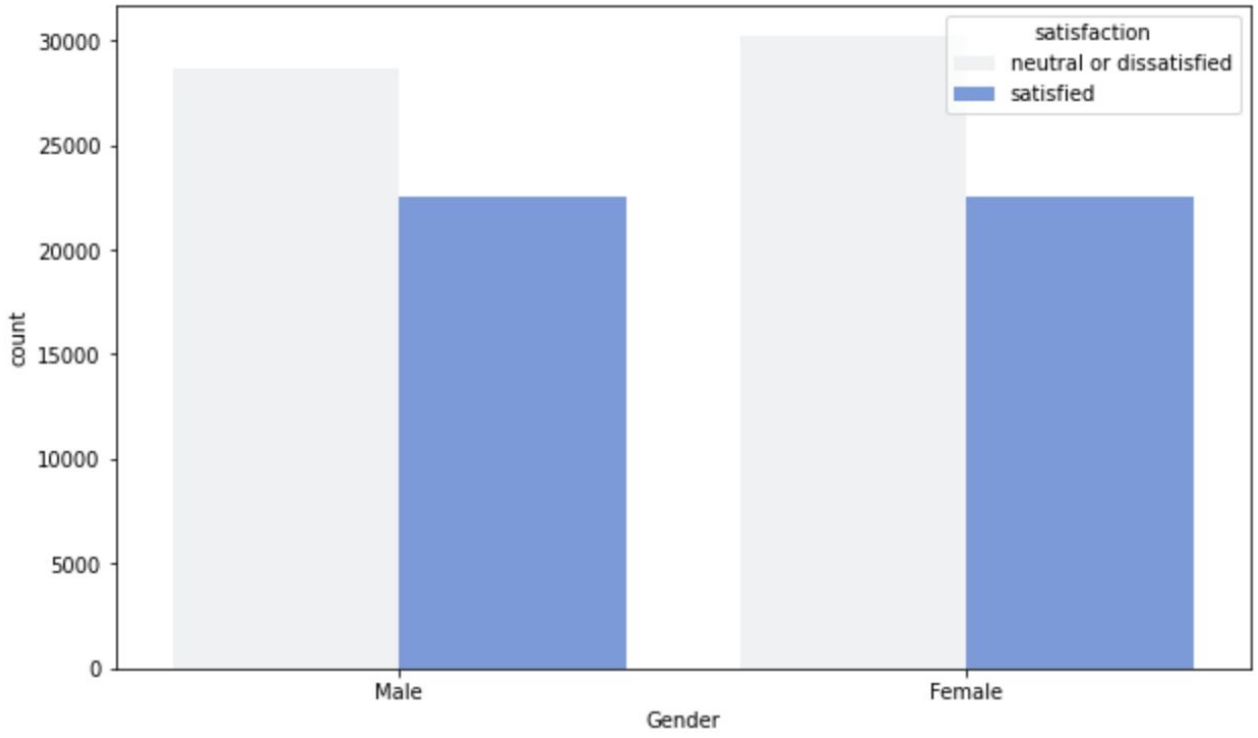


Correlaciones

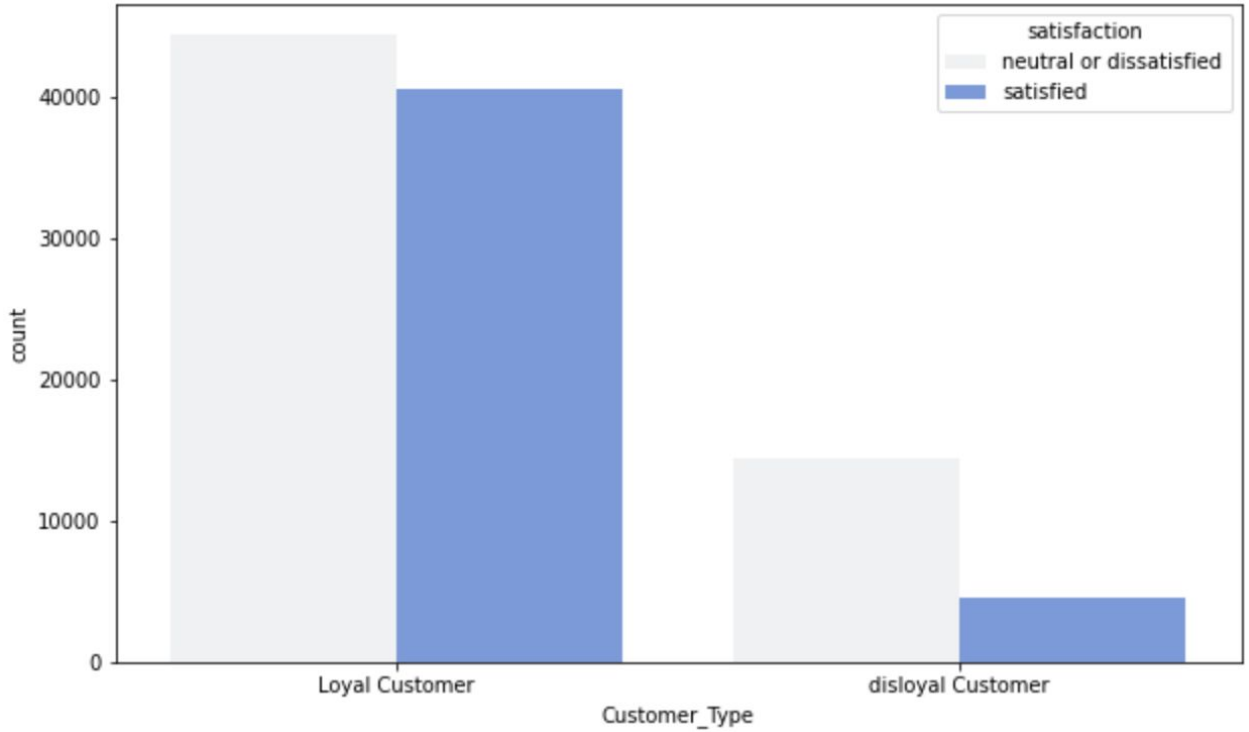

Age 1 0.099 0.018 0.038 0.025 -0.0013 0.023 0.21 0.16 0.076 0.058 0.041 -0.048 0.035 -0.049 0.054 -0.01 -0.012
FlightDistance 0.099 1 0.0071 -0.02 0.066 0.0048 0.057 0.21 0.16 0.13 0.11 0.13 0.063 0.073 0.058 0.093 0.0022 -0.0024
Inflight-wifiService 0.018 0.0071 1 0.34 0.72 0.34 0.13 0.46 0.12 0.21 0.12 0.16 0.12 0.043 0.11 0.13 -0.017 -0.019
Departure/Arrival_TimeConvenience 0.038 -0.02 0.34 1 0.44 0.44 0.0049 0.07 0.011 -0.0049 0.069 0.012 0.072 0.093 0.073 0.014 0.001 -0.00086
OnlineBooking_Ease 0.025 0.066 0.72 0.44 1 0.46 0.032 0.4 0.03 0.047 0.039 0.11 0.039 0.011 0.035 0.016 -0.0064 -0.008
GateLocation -0.0013 0.0048 0.34 0.44 0.46 1 -0.0012 0.0017 0.0037 0.0035 -0.028 -0.0059 0.0023 -0.035 0.0017 -0.0038 0.0055 0.0051
Food/Drink 0.023 0.057 0.13 0.0049 0.032 -0.0012 1 0.23 0.57 0.62 0.059 0.032 0.035 0.087 0.034 0.66 -0.03 -0.033
OnlineBoarding 0.21 0.21 0.46 0.07 0.4 0.0017 0.23 1 0.42 0.29 0.16 0.12 0.083 0.2 0.075 0.33 -0.019 -0.022
SeatComfort 0.16 0.16 0.12 0.011 0.03 0.0037 0.57 0.42 1 0.61 0.13 0.11 0.075 0.19 0.069 0.68 -0.028 -0.03
InflightEntertainment 0.076 0.13 0.21 -0.0049 0.047 0.0035 0.62 0.29 0.61 1 0.42 0.3 0.38 0.12 0.4 0.69 -0.027 -0.031
On-boardService 0.058 0.11 0.12 0.069 0.039 -0.028 0.059 0.16 0.13 0.42 1 0.36 0.52 0.24 0.55 0.12 -0.032 -0.035
Leg-roomService 0.041 0.13 0.16 0.012 0.11 -0.0059 0.032 0.12 0.11 0.3 0.36 1 0.37 0.15 0.37 0.096 0.014 0.012
BaggageHandling -0.048 0.063 0.12 0.072 0.039 0.0023 0.035 0.083 0.075 0.38 0.52 0.37 1 0.23 0.63 0.096 -0.0056 -0.0085
CheckinService 0.035 0.073 0.043 0.093 0.011 -0.035 0.087 0.2 0.19 0.12 0.24 0.15 0.23 1 0.24 0.18 -0.018 -0.02
Inflight_service -0.049 0.058 0.11 0.073 0.035 0.0017 0.034 0.075 0.069 0.4 0.55 0.37 0.63 0.24 1 0.089 -0.055 -0.059
Cleanliness 0.054 0.093 0.13 0.014 0.016 -0.0038 0.66 0.33 0.68 0.69 0.12 0.096 0.096 0.18 0.089 1 -0.014 -0.016
DepartureDelay -0.01 0.0022 -0.017 0.001 -0.0064 0.0055 -0.03 -0.019 -0.028 -0.027 -0.032 0.014 -0.0056 -0.018 -0.055 -0.014 1 0.97
ArrivalDelay -0.012 -0.0024 -0.019 -0.00086 -0.008 0.0051 -0.033 -0.022 -0.03 -0.031 -0.035 0.012 -0.0085 -0.02 -0.059 -0.016 0.97 1
Age
FlightDistance
Inflight-wifiService
re/Arrival_TimeConvenience
OnlineBooking_Ease
GateLocation
Food/Drink
OnlineBoarding
SeatComfort
InflightEntertainment
On-boardService
Leg-roomService
BaggageHandling
CheckinService
Inflight_service
Cleanliness
DepartureDelay
ArrivalDelay
1.0
0.8
0.6
0.4
0.2
0.0

EDA

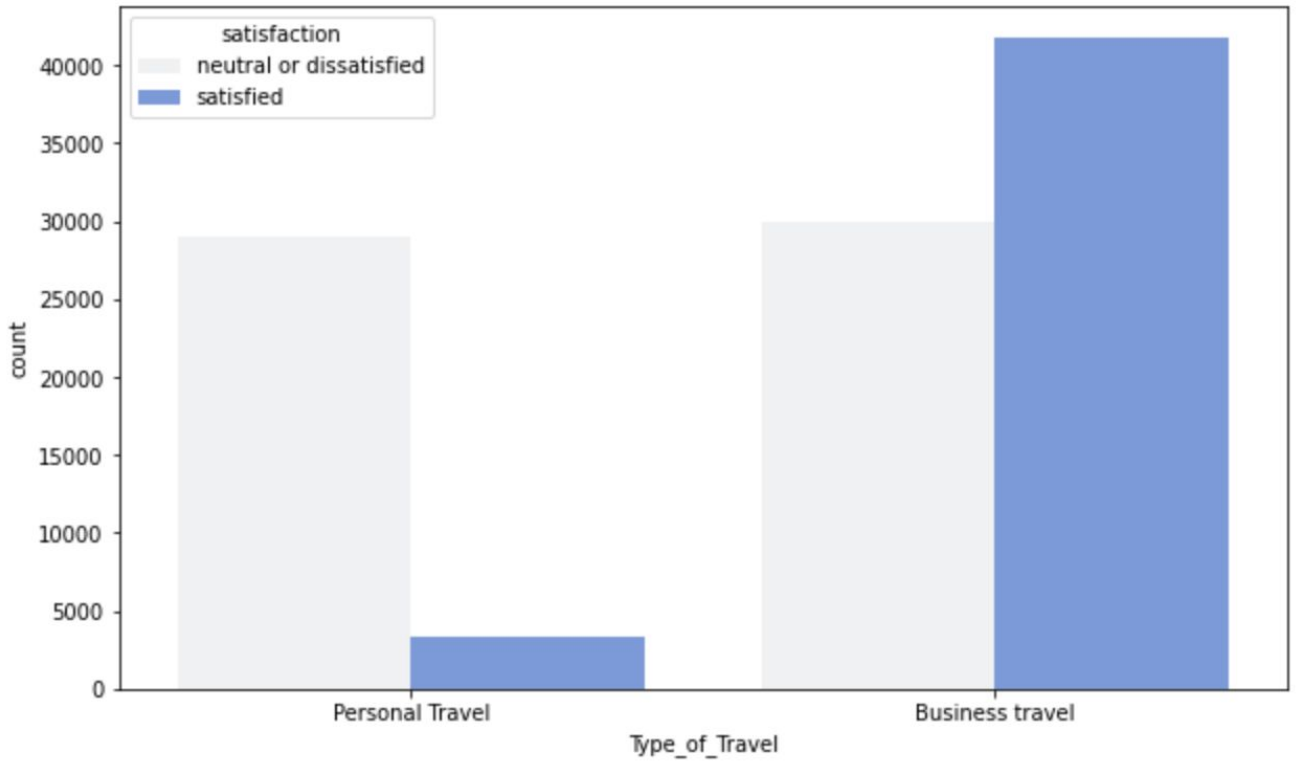
Género



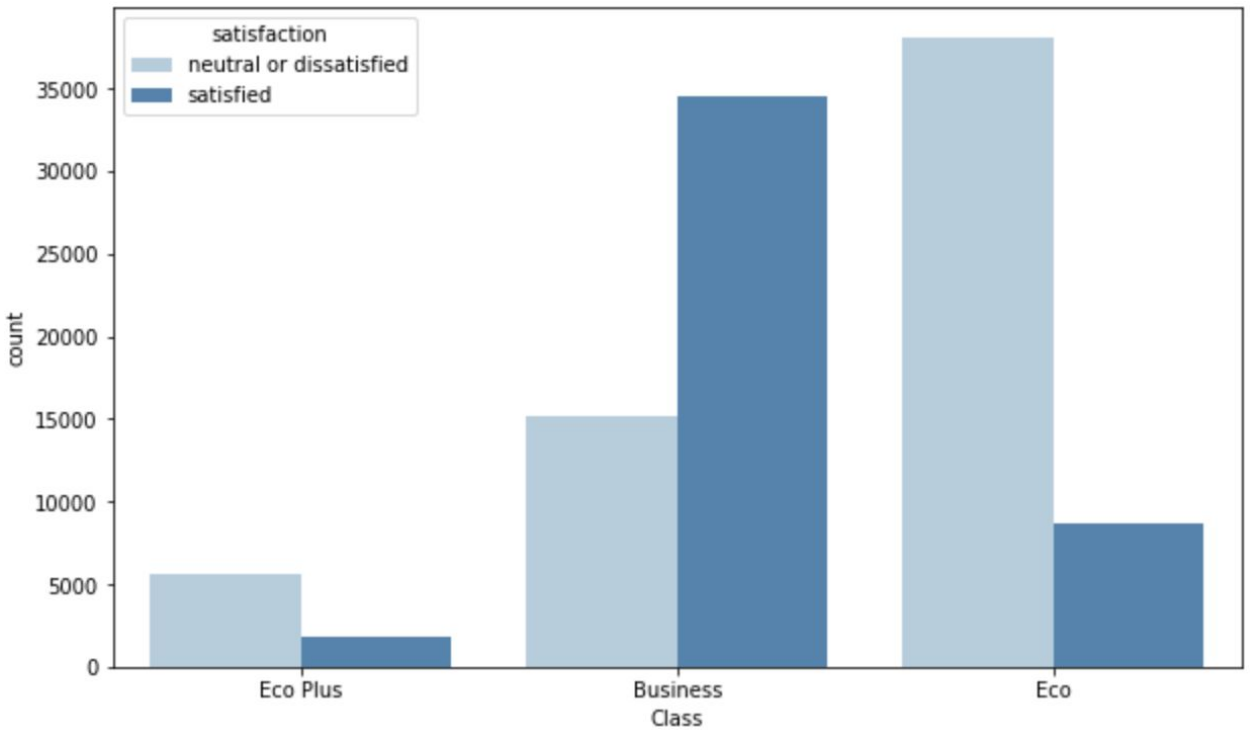
Tipo de cliente



Tipo de viaje



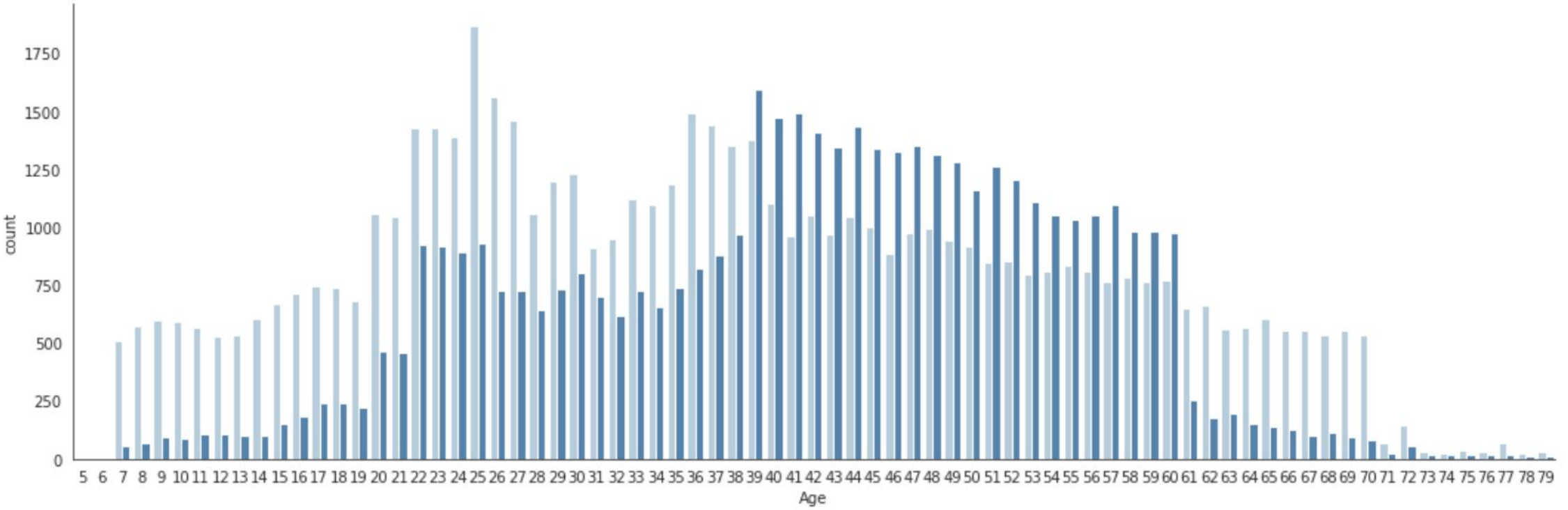
Clase



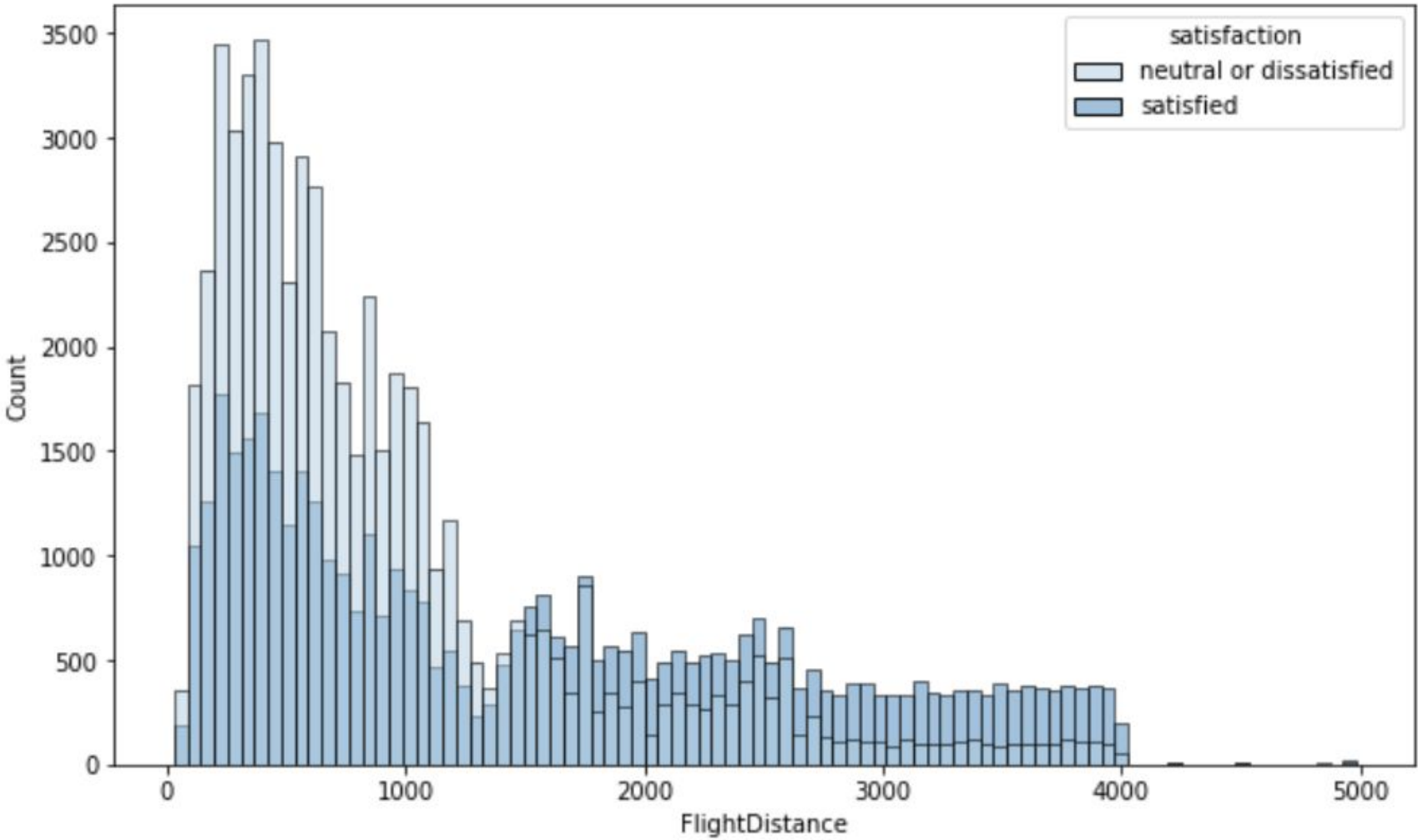
Promedios de los ratings:

Class	Business	Eco	Eco Plus
Age	41.574328	37.164253	38.654524
FlightDistance	1675.976925	743.439748	747.125567
Inflight-wifiService	2.775315	2.675067	2.767948
Departure/Arrival_TimeConvenience	2.905910	3.199123	3.217507
OnlineBooking_Ease	2.913964	2.605241	2.661996
GateLocation	2.982926	2.971954	2.967574
Food/Drink	3.323165	3.086277	3.122631
OnlineBoarding	3.716541	2.812985	2.889245
SeatComfort	3.760858	3.138838	3.183747
InflightEntertainment	3.635437	3.098256	3.141713
On-boardService	3.679472	3.120355	3.047638
Leg-roomService	3.644498	3.085720	3.061382
BaggageHandling	3.842907	3.450551	3.363758
CheckinService	3.519178	3.122002	3.017214
Inflight_service	3.844579	3.463921	3.388444
Cleanliness	3.477600	3.108097	3.130771
DepartureDelay	14.398067	15.160509	15.431545
ArrivalDelay	14.577272	15.672183	16.088645

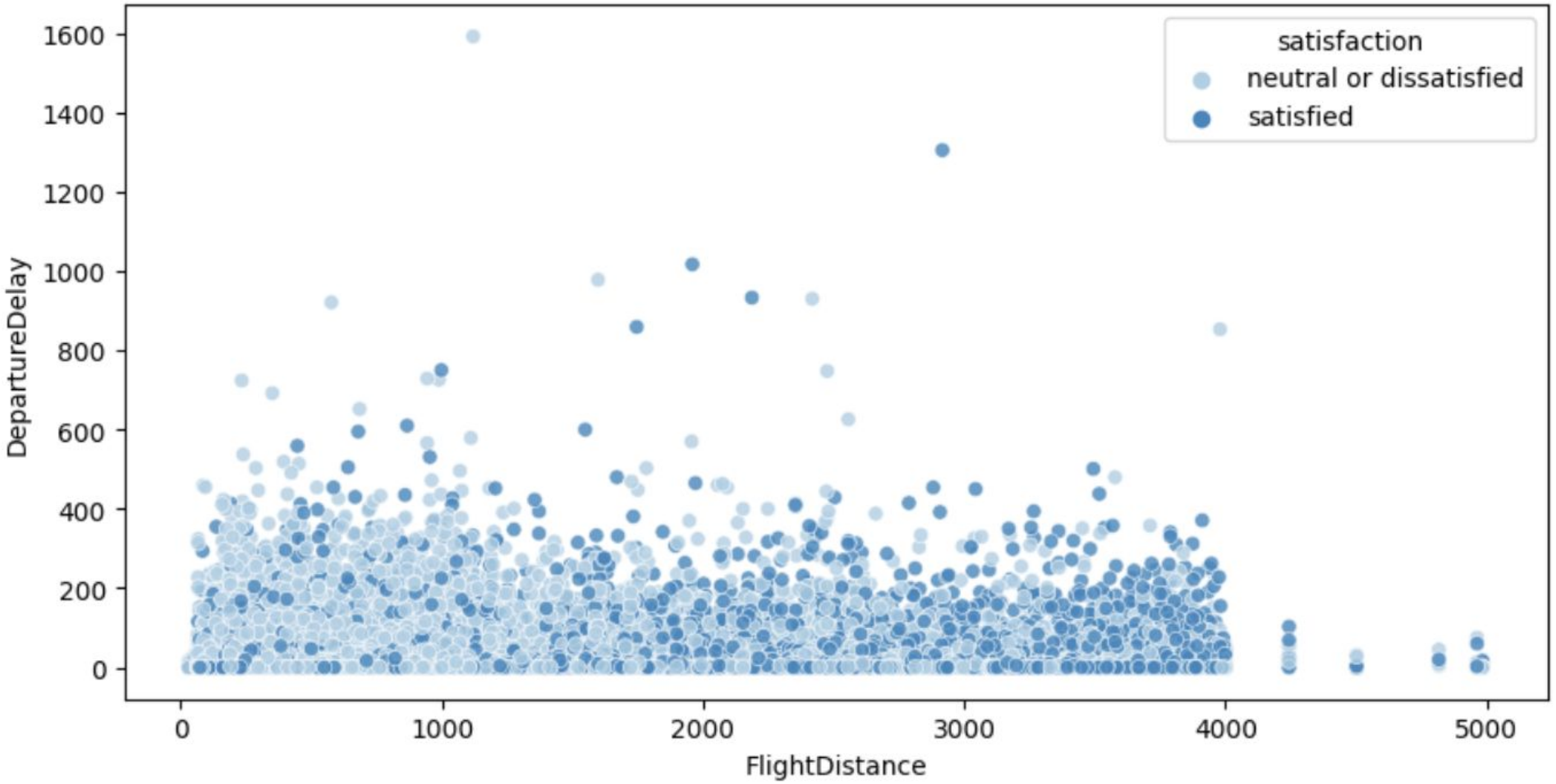
EDAD

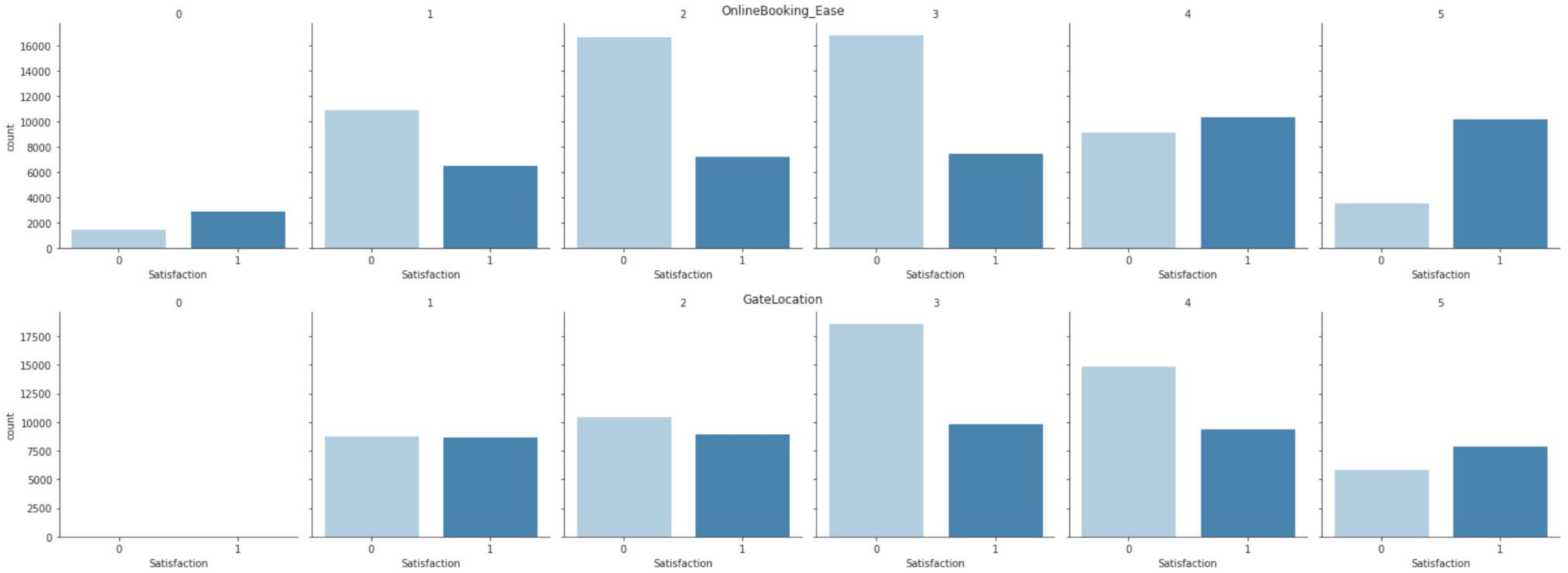

count
1750
1500
1250
1000
750
500
250
0
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79
Age
satisfaction
neutral or dissatisfied
satisfied

Distancia del vuelo

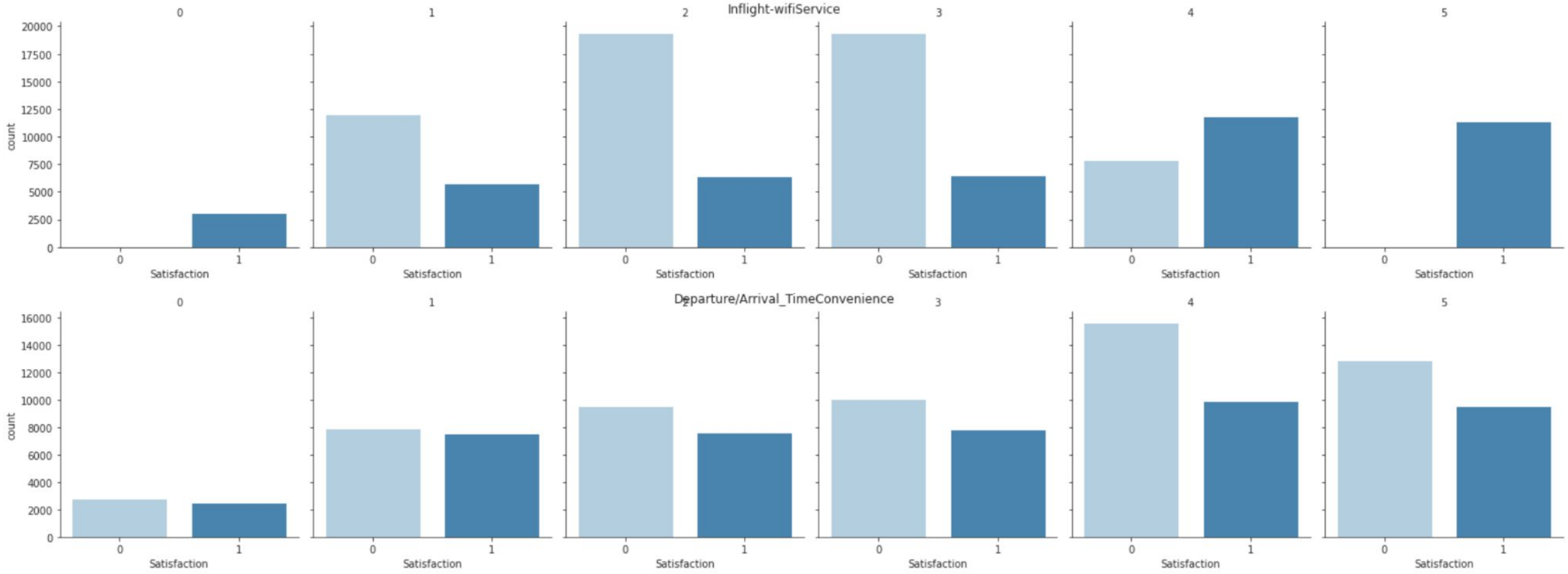


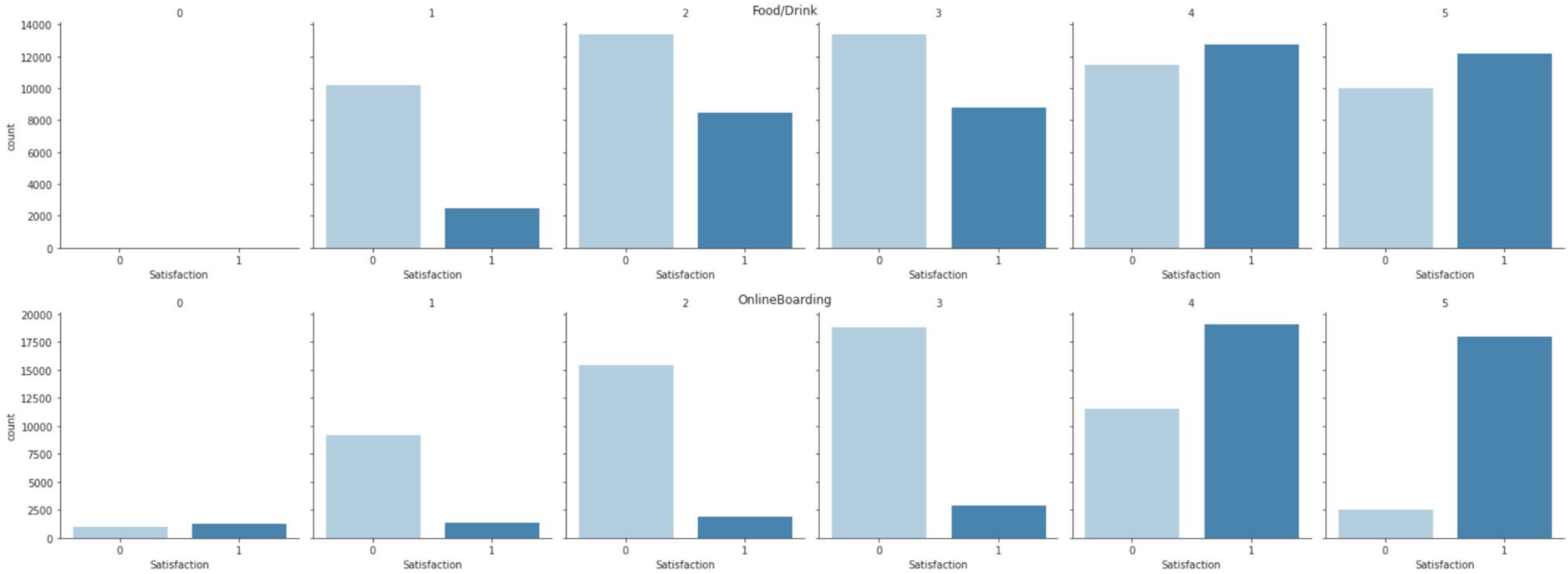
Distancia del vuelo - Demora en la partida

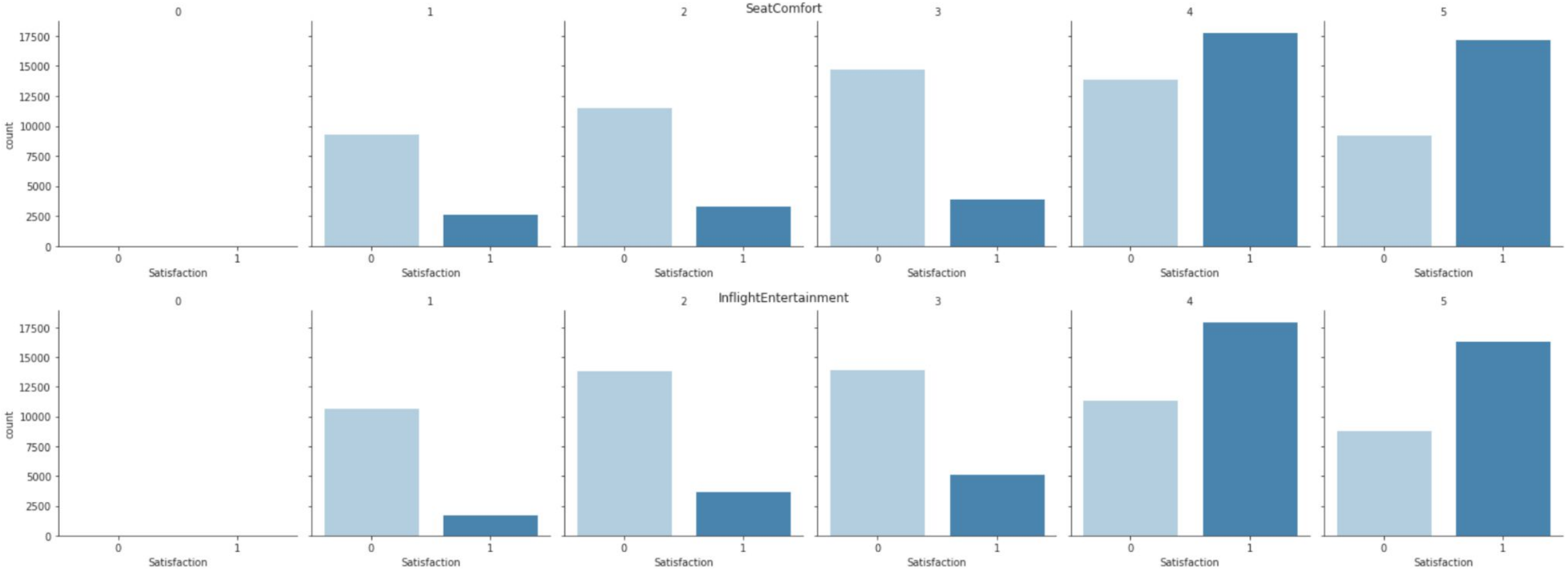


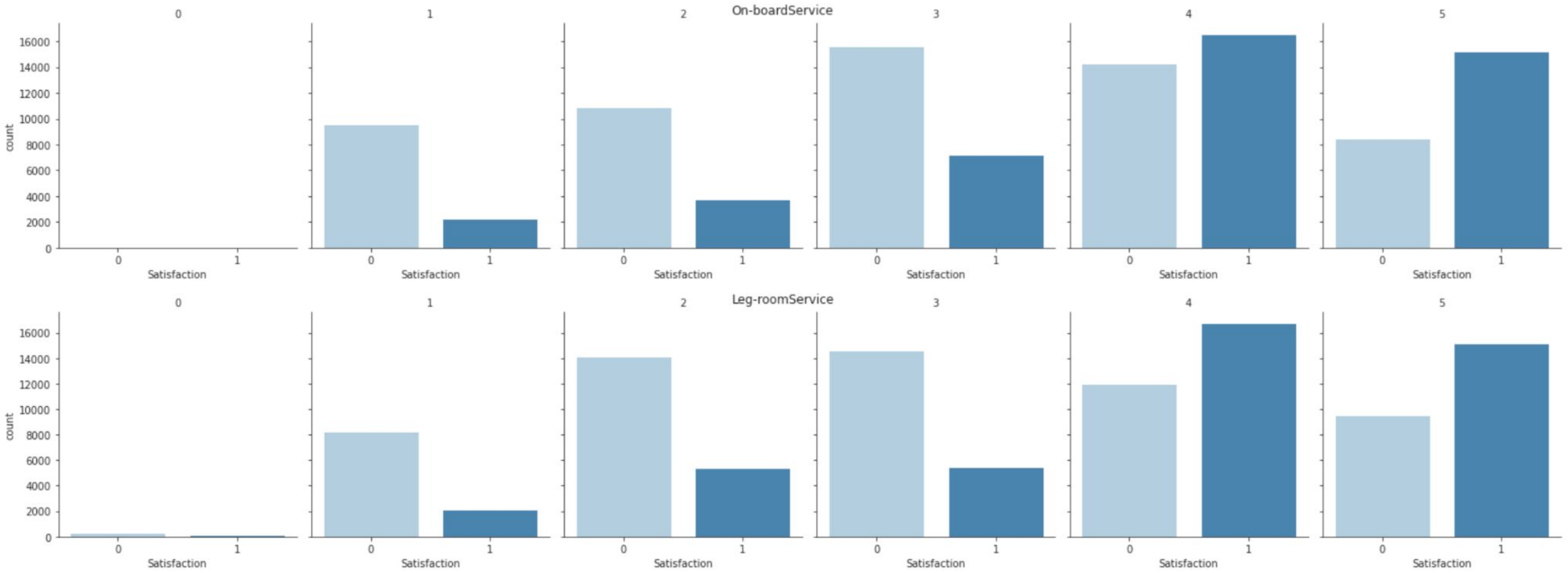

OnlineBooking_Ease
0
1
2
3
4
5
count
Satisfaction
GateLocation

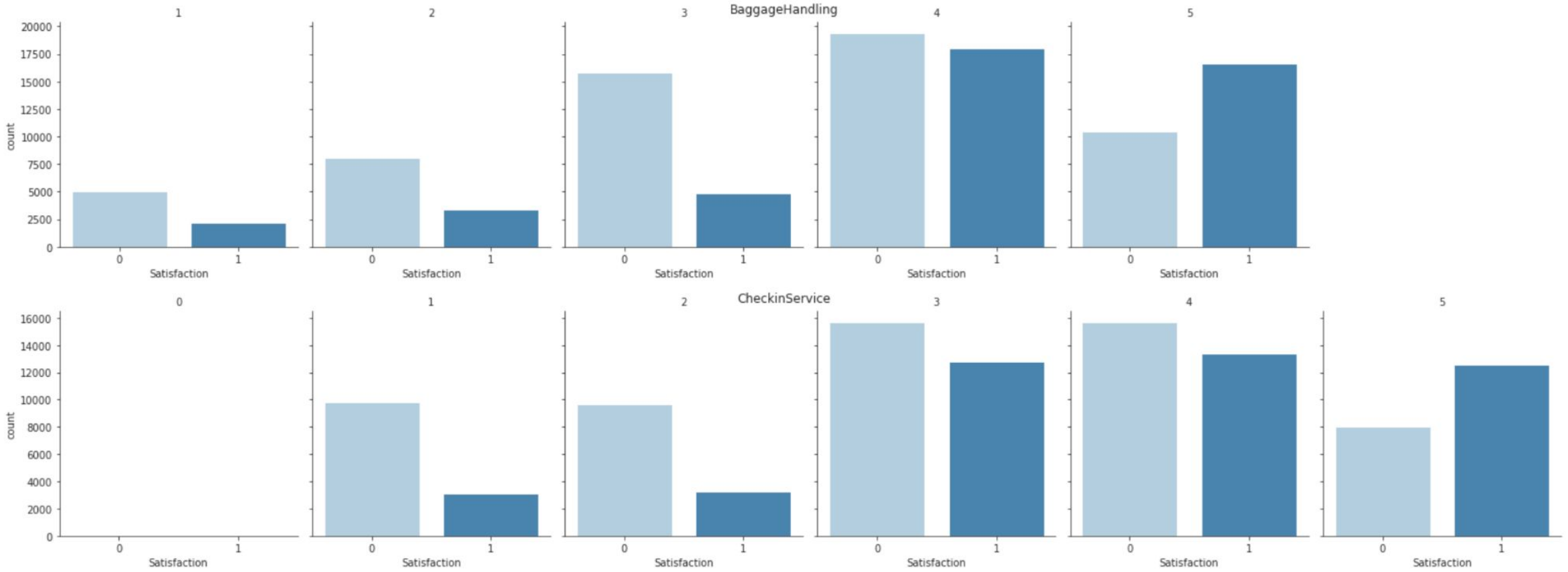
0 = neutral/dissatisfied
1=satisfied

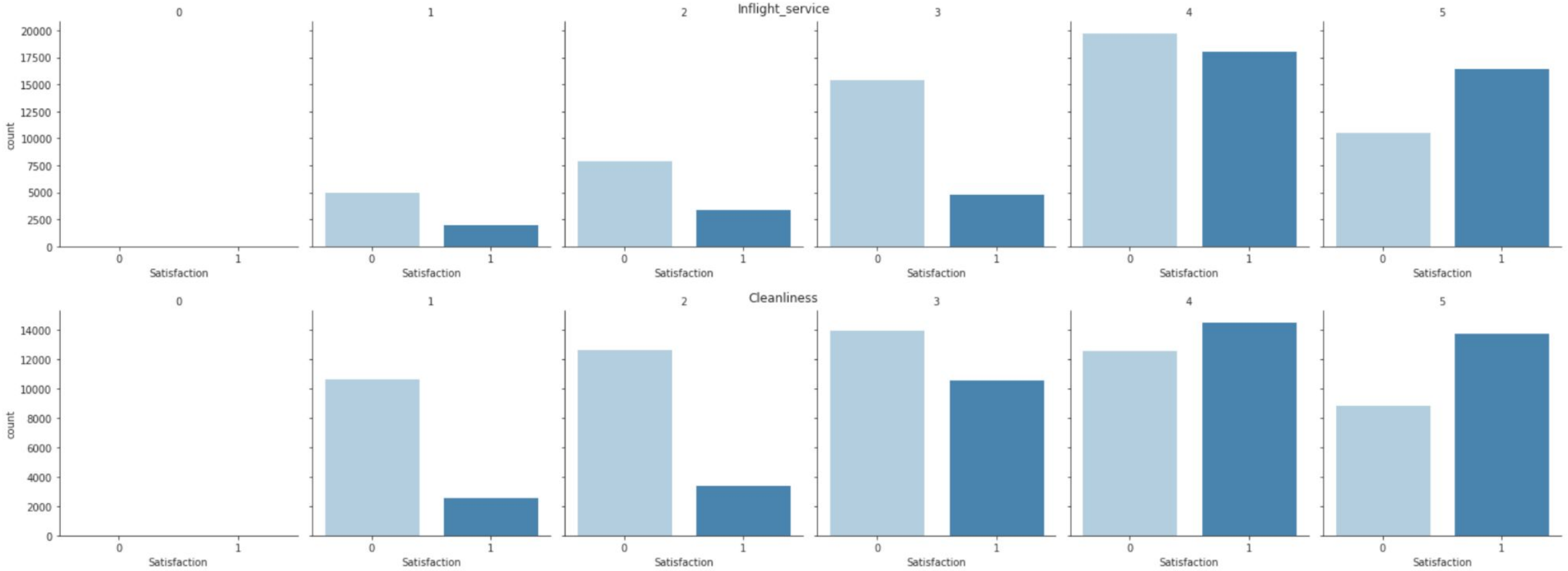



Food/Drink
0
1
2
3
4
5
count
Satisfaction
OnlineBoarding


SeatComfort
0
1
2
3
4
5
count
Satisfaction
InflightEntertainment


On-boardService
0
1
2
3
4
5
count
16000
14000
12000
10000
8000
6000
4000
2000
0
0
1
Satisfaction
Leg-roomService


BaggageHandling
1
2
3
4
5
count
Satisfaction
CheckinService
0
1
2
3
4
5


Inflight_service
0
1
2
3
4
5
count
Satisfaction
Cleanliness

Label encoding

satisfaction

neutral or dissatisfied	0
satisfied	1

Travel type

business travel	0
personal travel	1

Gender

female	0
male	1

Customer type

loyal customer	0
disloyal customer	1

Creación de variables: variables dummies para Class (Eco, Eco Plus y Business)

Train: **83123** registros

Test: **20781** registros

Estandarización

Modelos

Modelos utilizados:

- Logistic Regression
- K-neighbors
- MLP
- Extra Trees
- Decision Tree
- Random Forest
- XG Boosting
- AdaBoost
- Hist Gradient Boosting

Random Forest

Accuracy = 0.9568836918338868
ROC Area under Curve = 0.9543664053933957

	precision	recall	f1-score	support
0	0.95108	0.97376	0.96229	11739
1	0.96485	0.93497	0.94967	9042
accuracy			0.95688	20781
macro avg	0.95796	0.95437	0.95598	20781
weighted avg	0.95707	0.95688	0.95680	20781

Logistic Regression

Accuracy = 0.8733458447620422
ROC Area under Curve = 0.8687113542825646

	precision	recall	f1-score	support
0	0.87548	0.90442	0.88972	11739
1	0.87035	0.83300	0.85127	9042
accuracy			0.87335	20781
macro avg	0.87292	0.86871	0.87049	20781
weighted avg	0.87325	0.87335	0.87299	20781

Extra Trees

Accuracy = 0.9233434387180598
ROC Area under Curve = 0.9156970031609312

	precision	recall	f1-score	support
0	0.89832	0.97461	0.93491	11739
1	0.96296	0.85678	0.90677	9042
accuracy			0.92334	20781
macro avg	0.93064	0.91570	0.92084	20781
weighted avg	0.92644	0.92334	0.92267	20781

Decision Tree

Accuracy = 0.9458158895144603
ROC Area under Curve = 0.9448113845075312

	precision	recall	f1-score	support
0	0.95158	0.95255	0.95206	11739
1	0.93832	0.93707	0.93769	9042
accuracy			0.94582	20781
macro avg	0.94495	0.94481	0.94488	20781
weighted avg	0.94581	0.94582	0.94581	20781

XG Boosting

Accuracy = 0.9623213512343005
ROC Area under Curve = 0.9599798012206772

	precision	recall	f1-score	support
0	0.95627	0.97802	0.96702	11739
1	0.97060	0.94194	0.95605	9042
accuracy			0.96232	20781
macro avg	0.96344	0.95998	0.96154	20781
weighted avg	0.96251	0.96232	0.96225	20781

MLP

Accuracy = 0.9453346807179636
ROC Area under Curve = 0.942340039861683

	precision	recall	f1-score	support
0	0.93948	0.96541	0.95227	11739
1	0.95343	0.91927	0.93604	9042
accuracy			0.94533	20781
macro avg	0.94646	0.94234	0.94415	20781
weighted avg	0.94555	0.94533	0.94521	20781

Hist Gradient Boosting

Accuracy = 0.9640055820220393
ROC Area under Curve = 0.9611656520924897

	precision	recall	f1-score	support
0	0.95459	0.98305	0.96861	11739
1	0.97711	0.93928	0.95782	9042
accuracy			0.96401	20781
macro avg	0.96585	0.96117	0.96321	20781
weighted avg	0.96438	0.96401	0.96391	20781

AdaBoost

Accuracy = 0.9287810981184736
ROC Area under Curve = 0.9264429906086671

	precision	recall	f1-score	support
0	0.93051	0.94446	0.93743	11739
1	0.92646	0.90843	0.91736	9042
accuracy			0.92878	20781
macro avg	0.92848	0.92644	0.92739	20781
weighted avg	0.92875	0.92878	0.92870	20781

K-neighbors

Accuracy = 0.9248351859871998
ROC Area under Curve = 0.9231790390337326

	precision	recall	f1-score	support
0	0.93134	0.93594	0.93363	11739
1	0.91630	0.91042	0.91335	9042
accuracy			0.92484	20781
macro avg	0.92382	0.92318	0.92349	20781
weighted avg	0.92479	0.92484	0.92481	20781

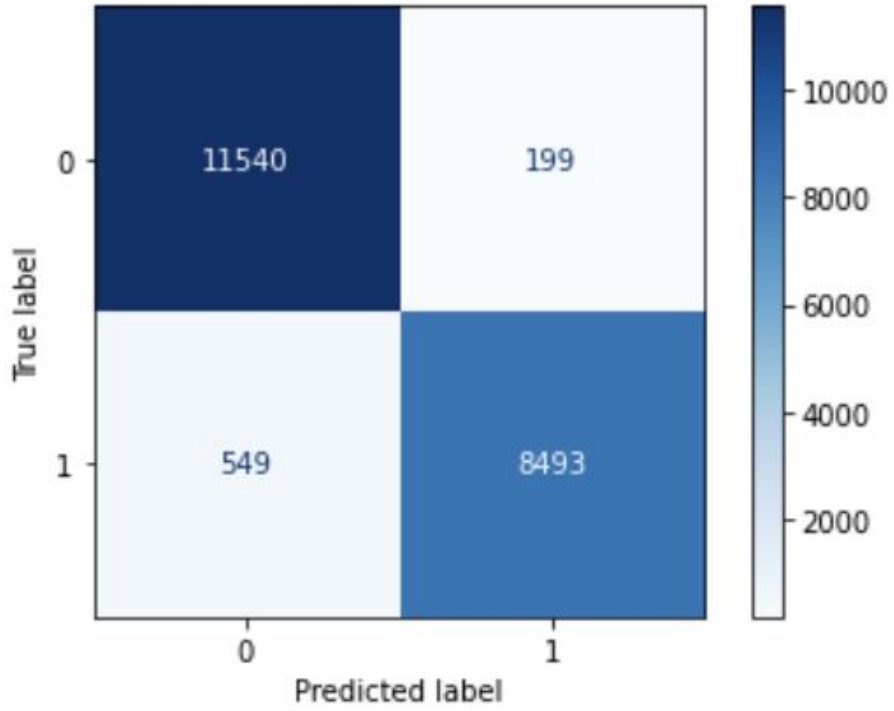
Modelo elegido

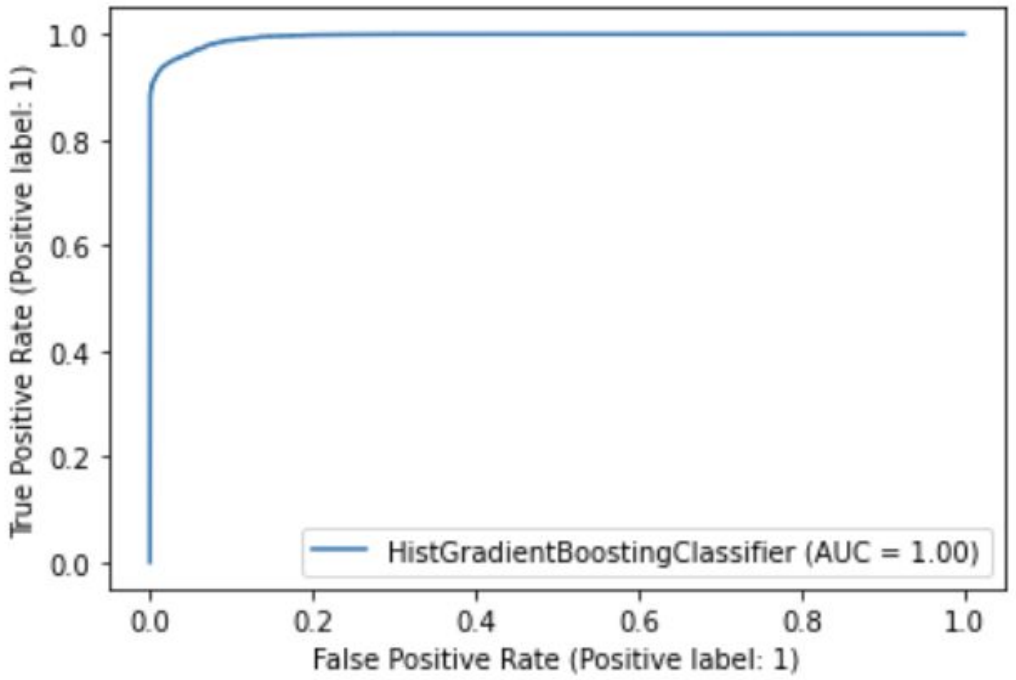
HistGradientBoostingClassifier

model_hgb =

HistGradientBoostingClassifier(max_bins=150,max_iter=250,max_depth=25,learning_rate=0.1,max_leaf_nodes=55)

- mayor accuracy
- mayor ROC-AUC
- menor tiempo





Conclusiones

Conclusiones

- Se puede predecir la satisfacción del cliente
- Es posible determinar la tendencia de las personas
- No solo el servicio es importante sino que también el tipo de cliente
    - sectorizar a los clientes para brindar un servicio especializado

Hipótesis:

- La distancia del vuelo no tiene un efecto negativo en la satisfacción
- La categoría del cliente es importante para la satisfacción del cliente

ITBA
Instituto Tecnológicc
de Buenos Aires

¡Muchas Gracias!