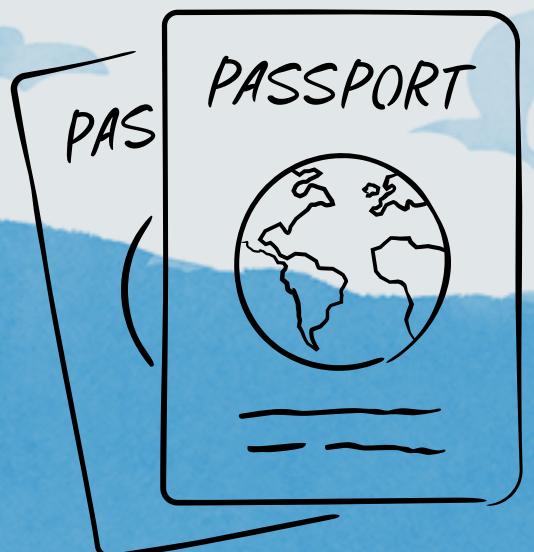


AIRLINE CUSTOMER SATISFACTION

GROUP 1



Nguyễn Thanh Phong
Nhữ Văn Tiến
Nguyễn Vũ Hoàng Nguyên
Mentor: Nguyễn Nhật Quang



MỤC LỤC

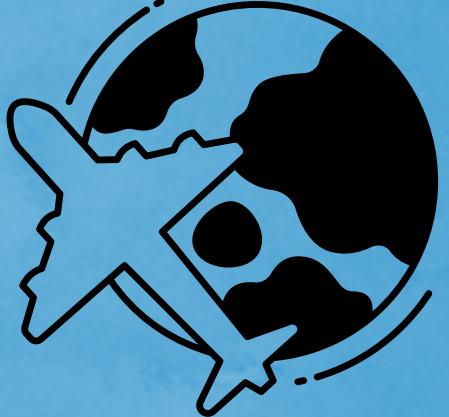
I. DATASET OVERVIEW

II. EDA

III. MODELING & EVALUATE



I. DATASET OVERVIEW



Bộ dữ liệu được cung cấp bởi một tổ chức hàng không tên **Invistico**. Bộ dữ liệu bao gồm thông tin chi tiết về những khách hàng đã bay cùng họ. Phản hồi của khách hàng về các bối cảnh khác nhau và dữ liệu chuyến bay của họ đã được tổng hợp.

Bộ dữ liệu gồm có **23 cột** và **129.881 dòng**



Mục đích chính đó là **dự đoán** rằng liệu khách hàng trong tương lai **có hài lòng** với dịch vụ của họ hay không từ đó cải thiện chất lượng dịch vụ và giảm thiểu số lượng khách hàng rời bỏ của họ

I. DATASET OVERVIEW

Ý nghĩa các cột thông tin:

Cột	Ý nghĩa	Cột	Ý nghĩa
1. satisfaction	Mức độ hài lòng	13. Inflight entertainment	Giải trí trên chuyến bay
2. Gender	Giới tính	14. Online support	Hỗ trợ trực tuyến
3. Customer Type	Loại khách hàng	15. Ease of Online booking	Dễ dàng trong việc đặt vé trực tuyến
4. Age	Tuổi	16. On-board service	Dịch vụ trên tàu
5. Type of Travel	Loại chuyến đi	17. Leg room service	Dịch vụ chỗ để chân
6. Class	Hạng ghế	18. Baggage handling	Xử lý hành lý
7. Flight Distance	Khoảng cách bay	19. Checkin service	Dịch vụ làm thủ tục lên máy bay
8. Seat comfort	Sự thoải mái của ghế ngồi	20. Cleanliness	Sự sạch sẽ của chuyến bay
9. Departure/Arrival time convenient	Thời gian khởi hành/đến có thuận tiện	21. Online boarding	Làm thủ tục lên máy bay trực tuyến
10. Food and drink	Thức ăn và đồ uống	22. Departure Delay in Minutes	Số phút chậm trễ khi khởi hành
11. Gate location	Vị trí cổng	23. Arrival Delay in Minutes	Số phút chậm trễ khi đến
12. Inflight wifi service	Dịch vụ wifi trên chuyến bay		

*Từ cột 8 đến cột 21 được dùng để khảo sát độ hài lòng với thang điểm được ghi từ 0 đến 5.

II. EXPLORE DATA ANALYSIS (EDA)

KIỂM TRA DỮ LIỆU NULL

satisfaction	0
Gender	0
Customer Type	0
Age	0
Type of Travel	0
Class	0
Flight Distance	0
Seat comfort	0
Departure/Arrival time convenient	0
Food and drink	0
Gate location	0
Inflight wifi service	0
Inflight entertainment	0
Online support	0
Ease of Online booking	0
On-board service	0
Leg room service	0
Baggage handling	0
Checkin service	0
Cleanliness	0
Online boarding	0
Departure Delay in Minutes	0
Arrival Delay in Minutes	393

Nhìn chung thì bộ dữ liệu khá đẹp
chỉ null khoảng 0.3% ở cột Arrival
Delay in Minutes

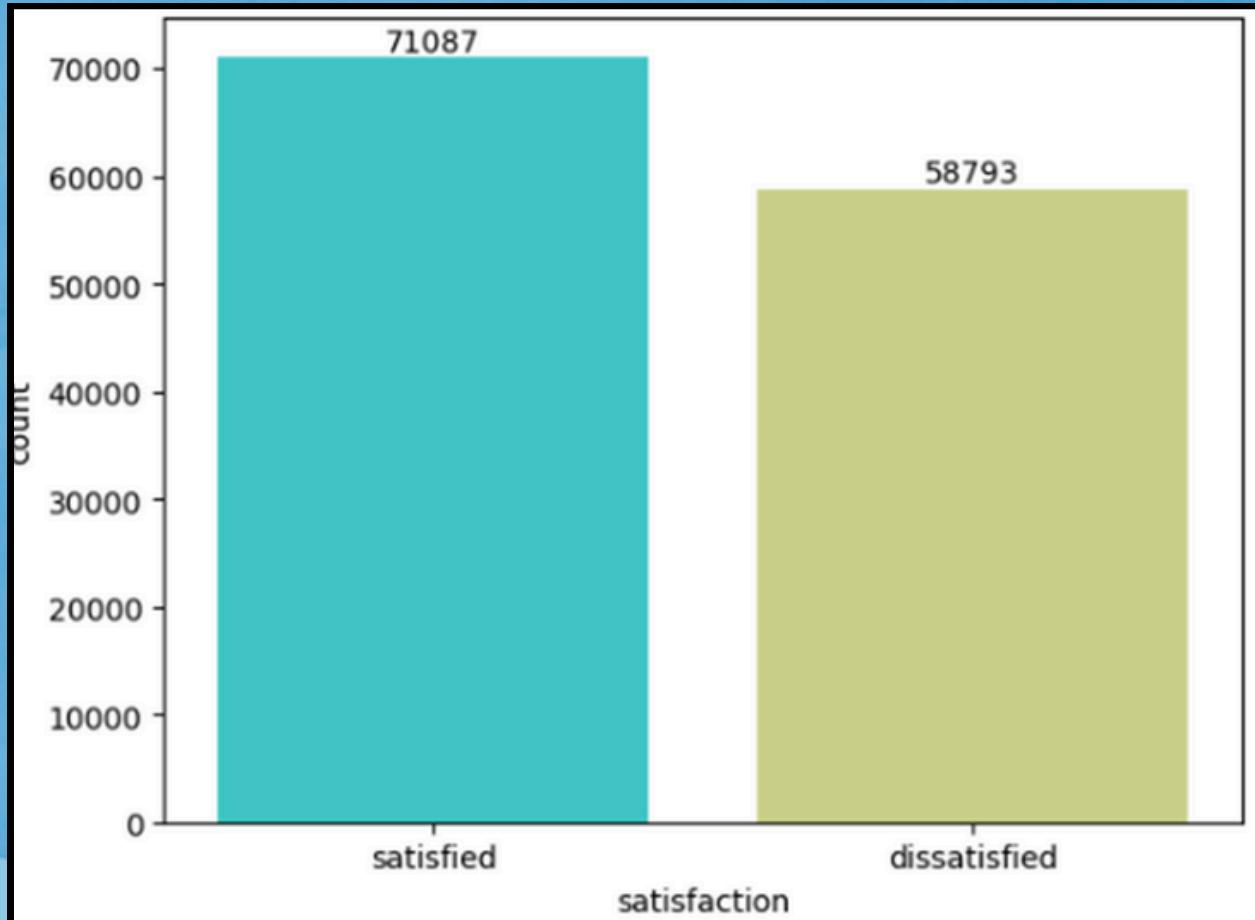
--> **Điền null bằng giá trị mode**

ĐỒNG NHẤT TÊN CÁC CỘT

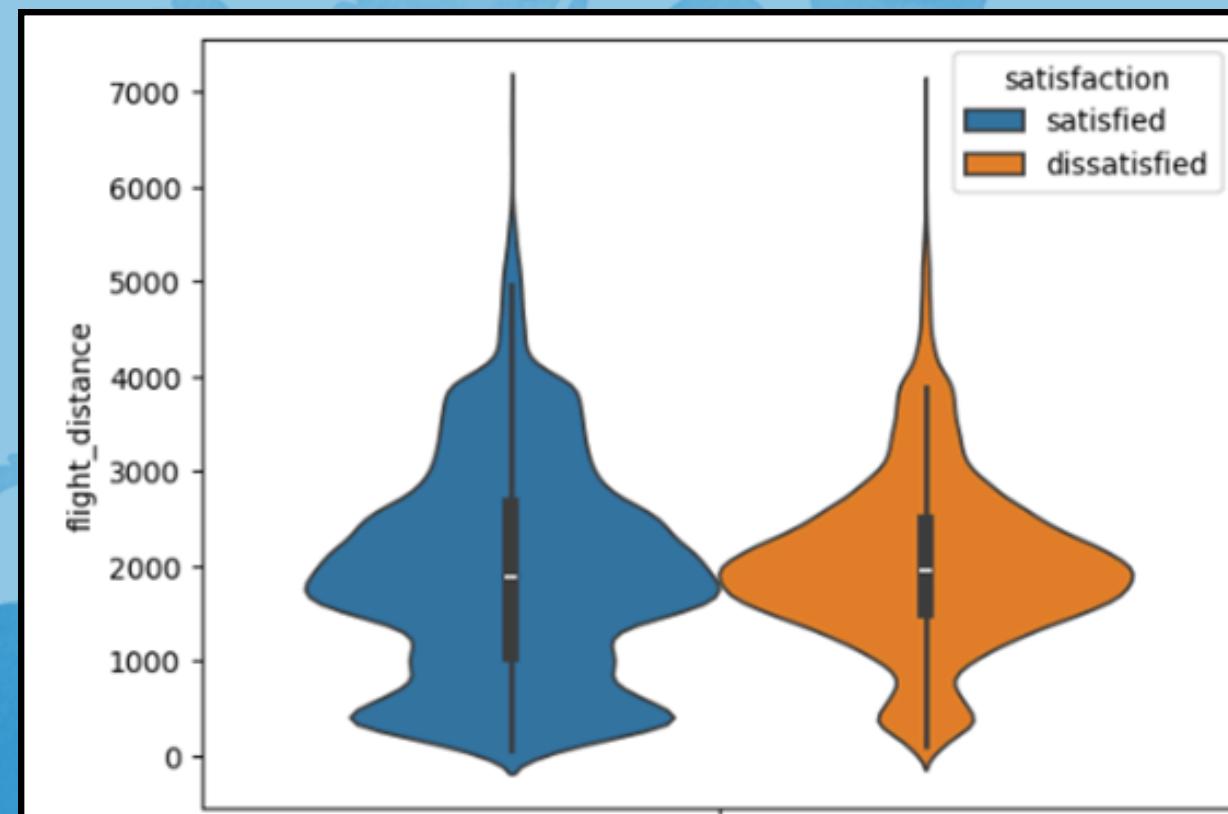
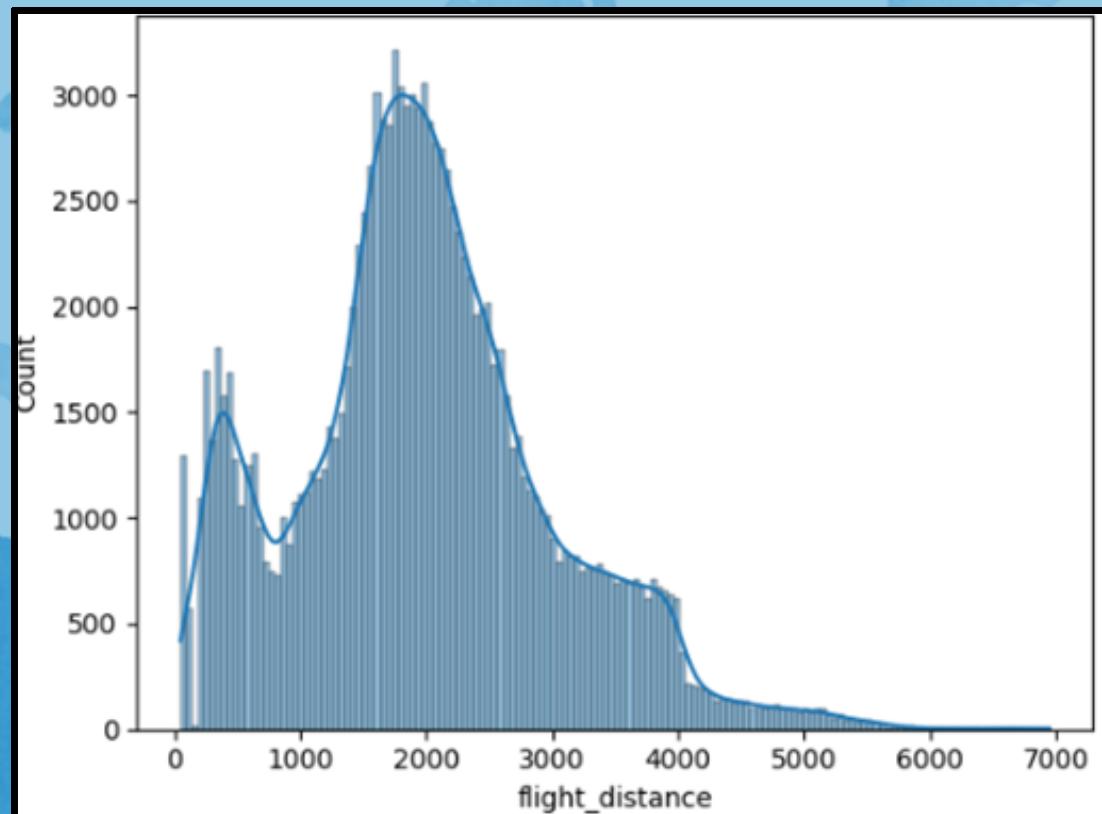
```
Index(['satisfaction', 'Gender', 'Customer Type', 'Age', 'Type of Travel',  
       'Class', 'Flight Distance', 'Seat comfort',  
       'Departure/Arrival time convenient', 'Food and drink', 'Gate location',  
       'Inflight wifi service', 'Inflight entertainment', 'Online support',  
       'Ease of Online booking', 'On-board service', 'Leg room service',  
       'Baggage handling', 'Checkin service', 'Cleanliness', 'Online boarding',  
       'Departure Delay in Minutes', 'Arrival Delay in Minutes'],  
      dtype='object')
```

```
Index(['satisfaction', 'gender', 'customer_type', 'age', 'type_of_travel',  
       'class', 'flight_distance', 'seat_comfort',  
       'departure_or_arrival_time_convenient', 'food_and_drink',  
       'gate_location', 'inflight_wifi_service', 'inflight_entertainment',  
       'online_support', 'ease_of_online_booking', 'on_board_service',  
       'leg_room_service', 'baggage_handling', 'checkin_service',  
       'cleanliness', 'online_boarding', 'departure_delay_in_minutes',  
       'arrival_delay_in_minutes'],
```

II. EXPLORE DATA ANALYSIS (EDA)

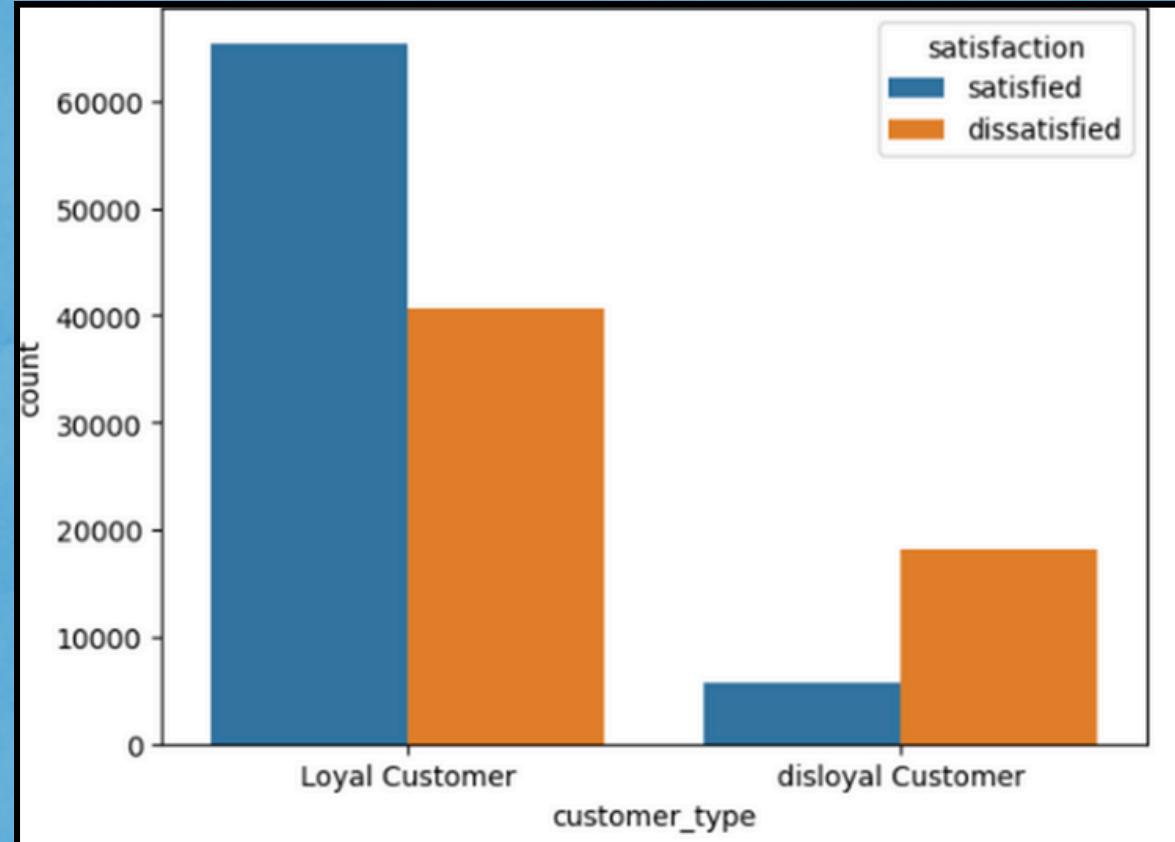


Nhìn chung thì số lượng hàng khách hài lòng và không hài lòng chiếm tỷ trọng khá tương đồng nhau.

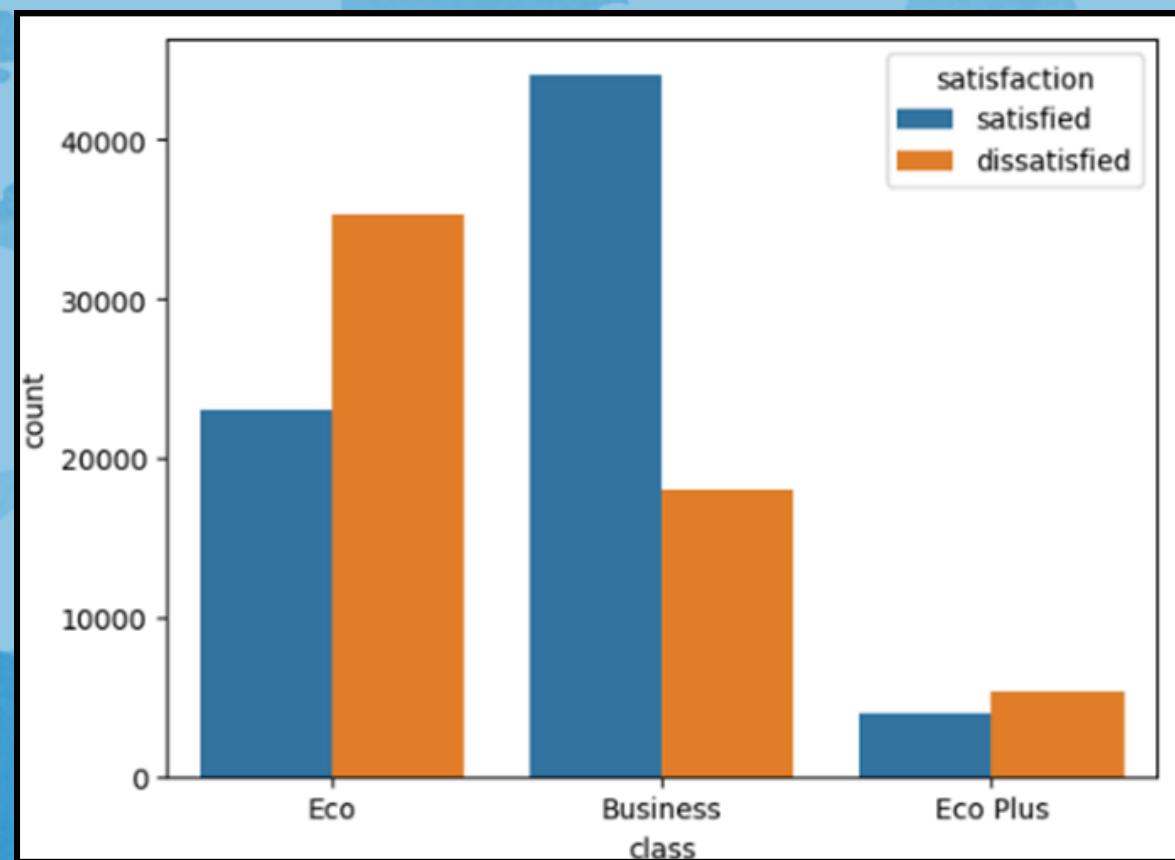


Một số lượng lớn khách hàng không hài lòng có quãng đường bay **từ 1500 đến 2500 km.**

II. EXPLORE DATA ANALYSIS (EDA)



Nhóm khách hàng trung thành cũng có tỷ lệ không hài lòng
rất lớn (chiếm hơn khoảng **50%**)

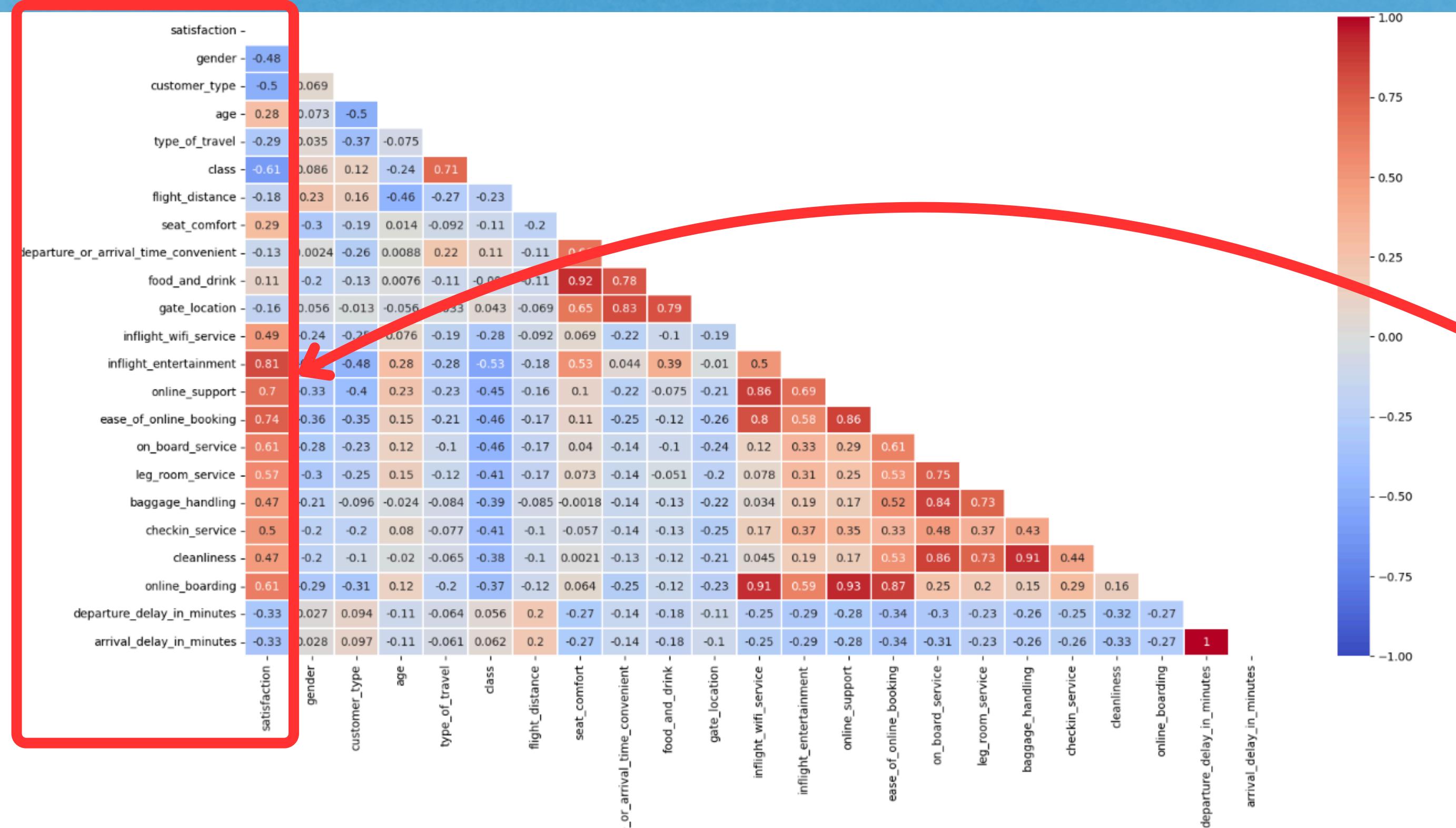


Nhóm **Eco** có số lượng khách hàng không hài lòng cao nhất
vì đây là hạng phổ thông với ít dịch vụ và ưu tiên.

III. MODELING & EVALUATE

1. ENCODING: Sử dụng phương pháp Label Encoding (Vì hầu hết các cột object đều chỉ số 2-3 giá trị duy nhất)

2. CORRELATION:



Nhìn vào tương quan giữa các cột với biến target (satisfaction) có thể rằng có những cột không có quá nhiều tương quan với nhau

--> Có thể bỏ hoặc thử 2 trường hợp

III. MODELING & EVALUATE

3. STANDARDIZATION:

Sử dụng phương pháp MinMax Scale

	satisfaction	gender	customer_type	class	inflight_wifi_service	inflight_entertainment	online_support	ease_of_online_booking	on_board_service	leg_room_serv	cancel_possibility
0	1	0	0	1		0.4	0.8	0.4	0.6	0.6	
1	1	1	0	0		0.0	0.4	0.4	0.6	0.8	
2	1	0	0	1		0.4	0.0	0.4	0.4	0.6	
3	1	0	0	1		0.6	0.8	0.6	0.2	0.2	
4	1	0	0	1		0.8	0.6	0.8	0.4	0.4	
...
129875	1	0	1	1		0.4	1.0	0.4	0.4	0.6	
129876	0	1	1	0		0.4	0.2	0.2	0.6	0.4	
129877	0	1	1	1		0.6	0.4	0.4	0.8	0.8	
129878	0	1	1	1		0.6	0.4	0.4	0.6	0.6	
129879	0	0	1	1		0.6	0.6	0.6	0.8	1.0	

Chia tập dữ liệu được theo tỷ lệ **train:test** là **7:3** (90916 dòng train - 38964 dòng test)

III. MODELING & EVALUATE

4. LỰA CHỌN MÔ HÌNH:

- **Logistic Regression** (Thử với 2 trường hợp)

TH1: Lựa chọn 10 biến có tương quan, ảnh hưởng nhiều đến biến Satisfaction

inflight_wifi_service, inflight_entertainment, online_support, ease_of_online_booking, on_board_service, leg_room_service, baggage_handling, checkin_service, cleanliness, online_boarding

TH2: Lấy hết tất cả các biến

- **XGBoost**
- **AdaBoost** (Dựa trên Decision Tree)
- **Gradient Boost** (Dựa trên Decision Tree)

Sau đó sử dụng metrics: Accuracy, Precision, Recall, F1-score để đánh giá mô hình.

III. MODELING & EVALUATE

5. ĐÁNH GIÁ MÔ HÌNH VÀ KẾT LUẬN

Số biến	Model	Accuracy	Precision		Recall		F1-Score	
			Yes	No	Yes	No	Yes	No
10 biến	Logistic Regression	82%	80%	84%	80%	84%	80%	84%
Tất cả các biến	Logistic Regression	83%	81%	85%	82%	85%	82%	85%
	Xgboost	96%	94%	97%	96%	95%	95%	96%
	Adaboost	95%	94%	96%	95%	95%	94%	95%
	Gradient boost	95%	94%	95%	94%	95%	94%	95%

Không có quá nhiều sự chênh lệch khi train mô hình **Logistic Regression** trên tập dữ liệu chỉ 10 biến và tập đủ tất cả các biến

Khi huấn luyện với tập dữ liệu có tất cả các biến:

- **Logistic Regression** cho ra kết quả tương đối ổn (Accuracy khoảng 85%) nhưng lại thấp nhất so với các mô hình có sử dụng kỹ thuật Boosting
- **Khi áp dụng kỹ thuật Boosting** thì Accuracy đạt tương đối cao (lên đến khoảng 95-96%) và Recall (Yes) cũng tăng từ 85 lên đến 95%



Có thể lựa chọn 1 trong 3 mô hình thuộc phương pháp Boosting trên để dự đoán sự hài lòng hay không



THANKS

For Your Watching

