

# Airline Passenger Satisfaction Analysis

DA56 - Thủy Triều - Đại Việt - Ngọc Linh



# Topic Overview

Chủ đề	Airline - Customer Satisfaction
Nội dung phân tích	Thông tin khách hàng và các đánh giá về dịch vụ của hãng bay
Mục đích phân tích	Tìm ra các yếu tố ảnh hưởng đến sự hài lòng của khách hàng khi sử dụng dịch vụ bay
Công cụ sử dụng	Python 
Mục tiêu & Ý nghĩa	Tìm hiểu những yếu tố có tác động trực tiếp đến trải nghiệm của khách hàng. Giúp hãng bay nói riêng nhận diện chính xác dịch vụ cần cải thiện, tăng lượng khách hàng hài lòng với hãng bay
Link Dataset	<a href="https://www.kaggle.com/uciml/airline-passenger-satisfaction">Airline Passenger Satisfaction (kaggle.com)</a>

# Agenda



**O1**

## Overview

Tổng quan về vấn đề

**O2**

## Workflow

Các bước xử lý và thực hiện

**O3**

## Customer Segment

Tìm hiểu phân khúc khách hàng

**O4**

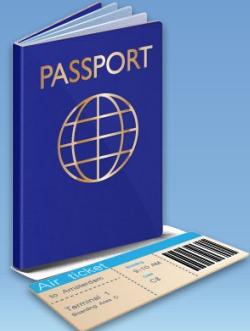
## Prediction

Mô hình và dự đoán

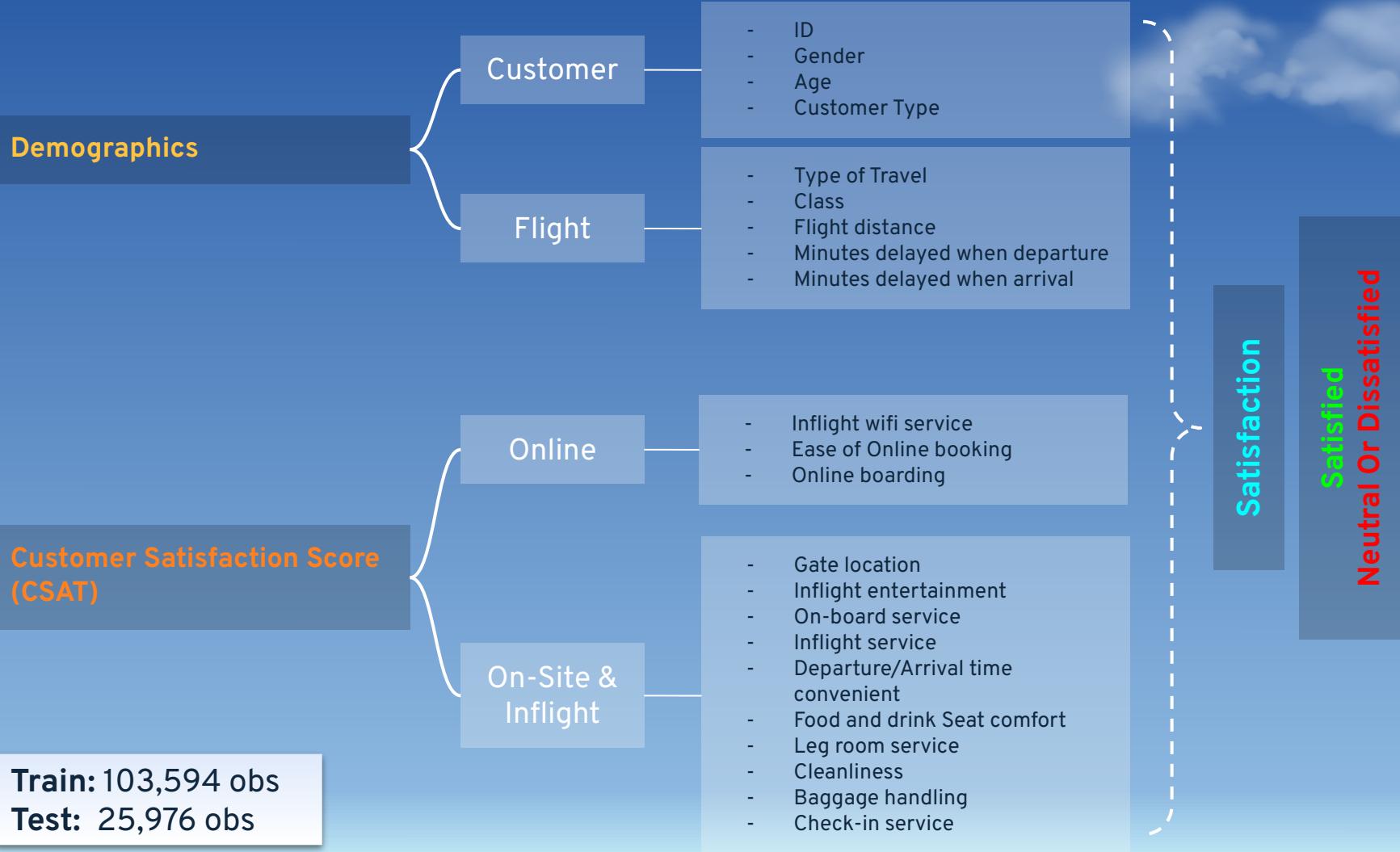


01

# Overview



# AIRLINE PASSENGER SATISFACTION





# Problem Statement



**~56% Customer Neutral or Dissatisfied**

Tỷ lệ khách hàng trung lập và không hài lòng



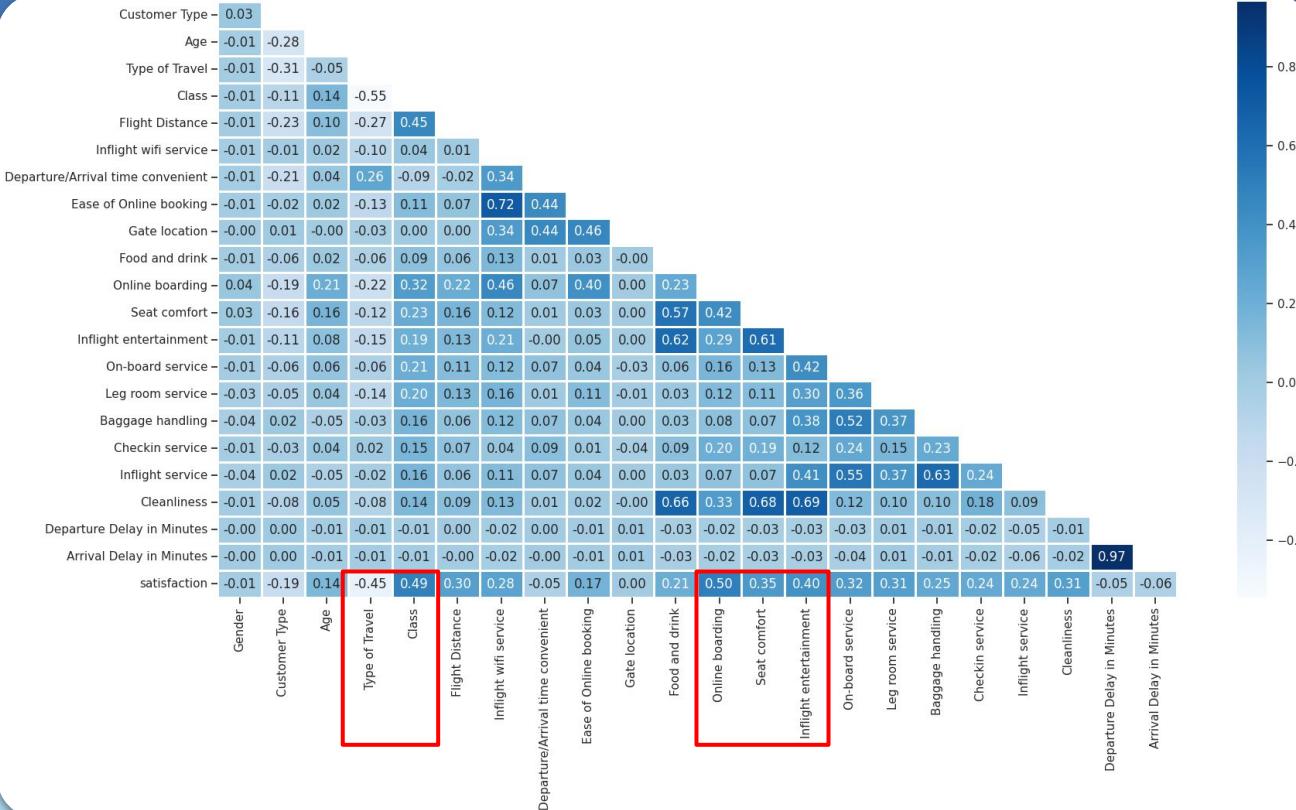
**~3.2 out of 5 - Average service score**

Điểm trung bình đánh giá của các dịch vụ hàng không cung cấp

## ***OBJECTIVE :***

- *Nâng cao tỷ lệ khách hàng hài lòng*
- *Nhận diện các yếu tố cần duy trì và các yếu tố cần cải thiện thêm*

# Correlation between Customer & Satisfaction



Dữ liệu đáng quan tâm:  
Demographics

- Class
- Type of Travel

CSAT score

- Online boarding
- Inflight entertainment
- Seat comfort

Dữ liệu ít ý nghĩa:

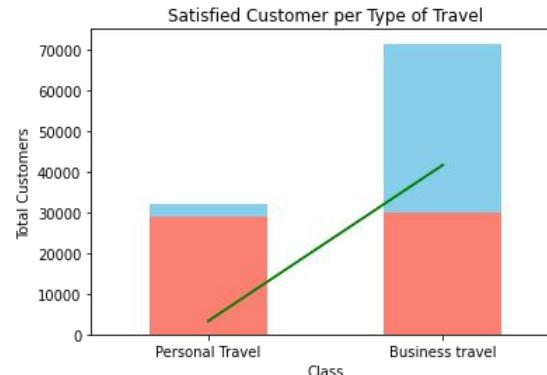
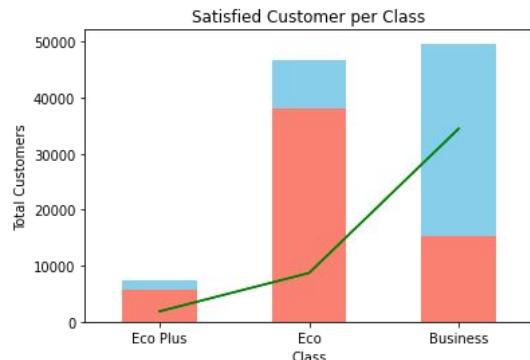
- Gender
- Departure/Arrival Delay in Minutes



# Demographics

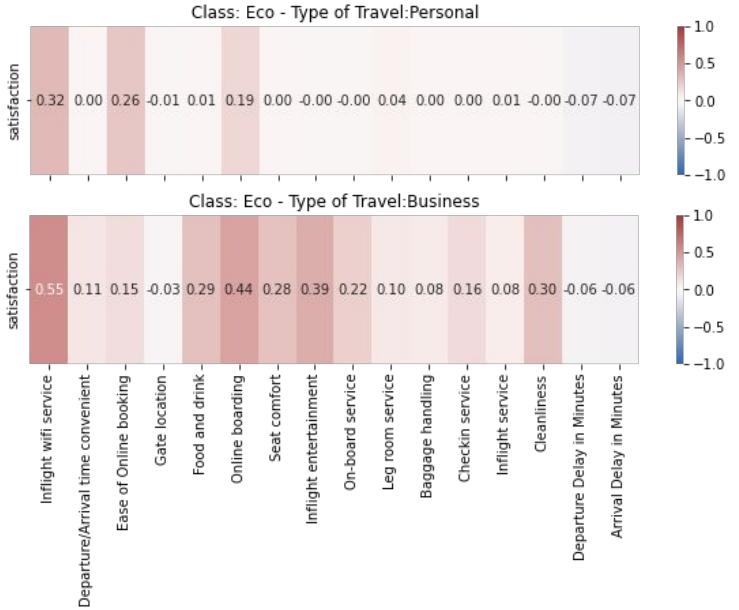
- Khách hàng hạng ghế “Phổ thông” đánh giá trải nghiệm khách hàng khá kém
- Với các khách di chuyển với mục đích thoả mái hơn (không vì công tác) đánh giá khá thấp về dịch vụ hãng bay

— Satisfied trend  
— neutral or dissatisfied  
— satisfied





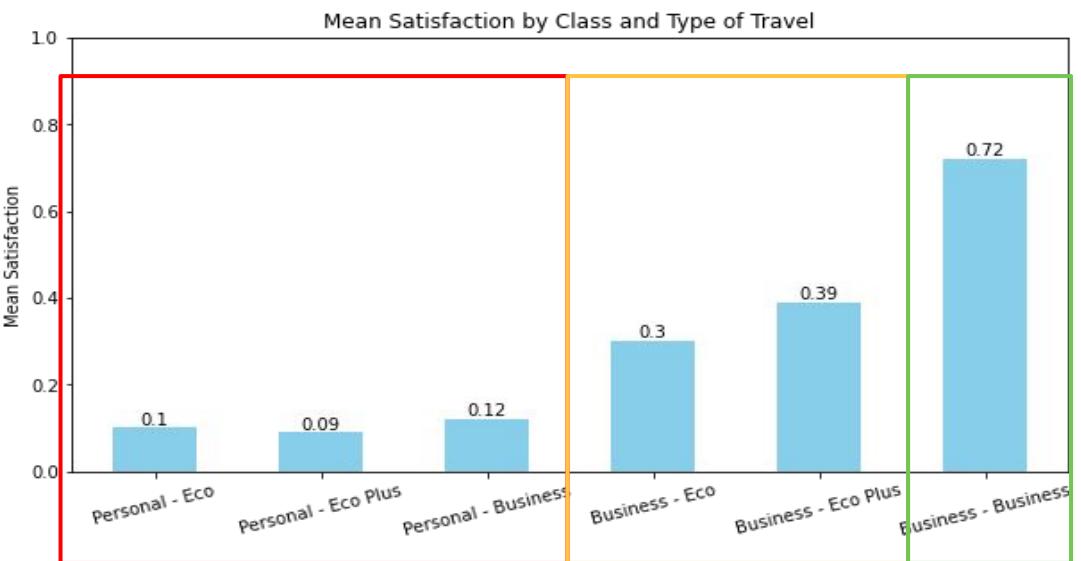
# Demographics



- Các khách hàng Personal quan tâm chủ yếu liên quan đến Wifi, Online boarding
- Nhóm khách Business ảnh hưởng nhiều bởi các trải nghiệm trực tiếp với hãng bay (On site - Inflight)



# Demographics



Personal

- Wifi
- Online

10%



Bus - Eco

- Wifi
- Online
- Entertainment

31%



Bus - Bus

- On site - Inflight

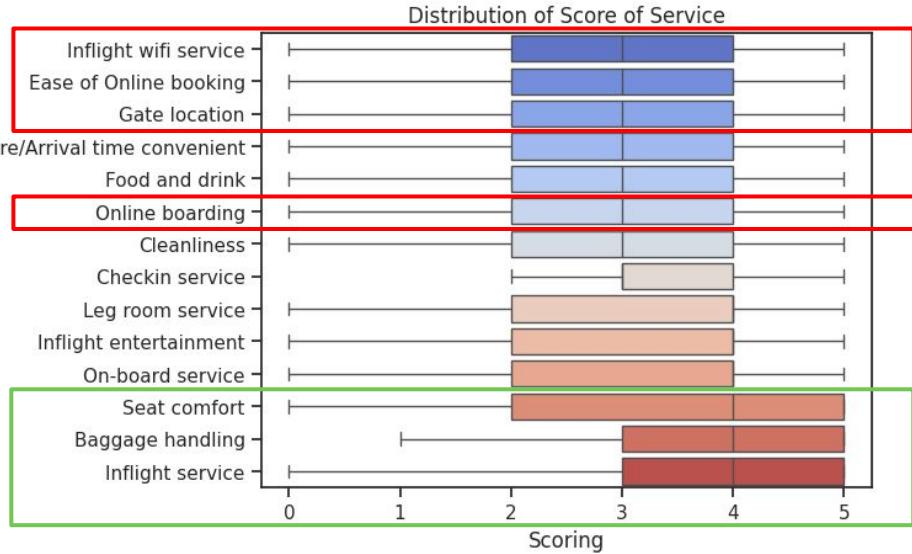
72%



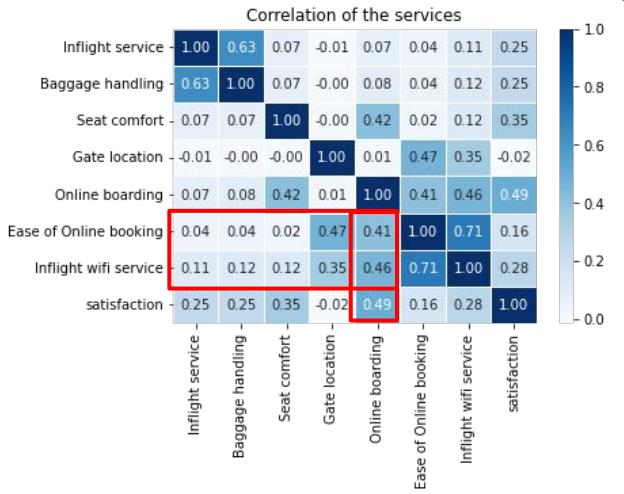
# Worldwide Airline

## - CSAT -

Services



Trung bình: 3.24 / 5



- Dịch vụ liên quan đến công nghệ thông tin của Airline này không được đánh giá cao
- Dịch vụ wifi ảnh hưởng lớn đến CSAT nhưng được đánh giá thấp điểm nhất

02

# Workflow



# Workflow description



## Cleaning & Encoding

- Xử lý missing value.
- Xử lý Outlier của tập dữ liệu.
- Encoding data các dữ liệu dạng chuỗi.



## Scaling & Balancing

- Xử lý imbalanced Data.
- Scaling Data.
- Xử lý giảm chiều dữ liệu bằng PCA trong phân tích Clustering.



## Customer Segment

- 4 Clusters
- Tìm hiểu thêm về sự quan tâm dịch vụ và hành vi của các cụm khách hàng nổi bật.
- Đưa ra một số đề xuất dựa vào hành vi khách hàng.



## Modeling & Prediction

- Thủ nghiệm với 6 model khác nhau.
- Tìm ra các yếu tố dịch vụ ảnh hưởng nổi bật đến khách hàng với Importance Feature.
- Đưa ra kết luận thêm về góc nhìn khi áp dụng mô hình học máy.

# Preprocessing Data

# Dataset information

#	Column	Non-Null Count	Dtype	Unnamed: 0	0
0	Unnamed: 0	103904	non-null	int64	0
1	id	103904	non-null	int64	0
2	Gender	103904	non-null	object	0
3	Customer Type	103904	non-null	object	0
4	Age	103904	non-null	int64	0
5	Type of Travel	103904	non-null	object	0
6	Class	103904	non-null	object	0
7	Flight Distance	103904	non-null	int64	0
8	Inflight wifi service	103904	non-null	int64	0
9	Departure/Arrival time convenient	103904	non-null	int64	0
10	Ease of Online booking	103904	non-null	int64	0
11	Gate location	103904	non-null	int64	0
12	Food and drink	103904	non-null	int64	83
13	Online boarding	103904	non-null	int64	0
14	Seat comfort	103904	non-null	int64	0
15	Inflight entertainment	103904	non-null	int64	0
16	On-board service	103904	non-null	int64	0
17	Leg room service	103904	non-null	int64	0
18	Baggage handling	103904	non-null	int64	0
19	Checkin service	103904	non-null	int64	0
20	Inflight service	103904	non-null	int64	0
21	Cleanliness	103904	non-null	int64	0
22	Departure Delay in Minutes	103904	non-null	int64	0
23	Arrival Delay in Minutes	103594	non-null	float64	318
24	satisfaction	103904	non-null	object	0

```
def clean_data(df):
    df.drop(['Unnamed: 0'], axis = 1, inplace = True)
    df = df.dropna(subset=['Arrival Delay in Minutes'])
    return df

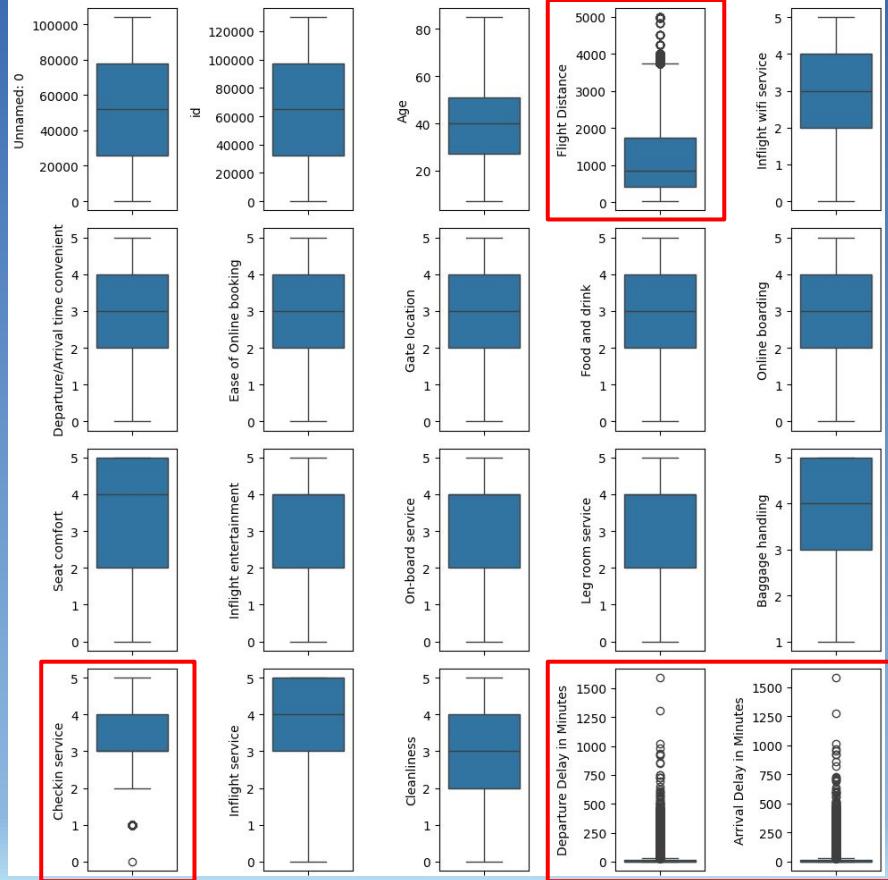
train_data = clean_data(train_data)
test_data = clean_data(test_data)
```



train shape : (103594, 24)  
test shape : (25893, 24)

Missing value xuất hiện ở tập train và test,  
tuy nhiên với số lượng ít nên xử lý drop để  
làm sạch dữ liệu

# Preprocessing Data



```
## Create function to drop outliers
def drop_outlier(df: pd.DataFrame, column: list):
    for col in column:
        Q1 = df[col].quantile(0.25)
        Q3 = df[col].quantile(0.75)
        IQR = Q3 - Q1
        lower_bound = Q1 - 1.5 * IQR
        upper_bound = Q3 + 1.5 * IQR
        outlier = df[(df[col] < lower_bound) | (df[col] > upper_bound)].index
        df.drop(outlier, axis = 0, inplace = True)
    return df

drop_outlier(train_data, ['Flight Distance'])
drop_outlier(train_data, ['Checkin service'])

drop_outlier(train_data, ['Arrival Delay in Minutes'])
drop_outlier(train_data, ['Departure Delay in Minutes'])

train_data.reset_index()
```

train shape : (66139, 24)  
test shape : (25893, 24)

- Một số Outlier ở delay có giá trị tương đối vô lý khi thời gian từ 750 lên đến 1500 phút → Xử lý Drop outlier bằng phương pháp IQR
- Tập train giảm 37.755 quan sát sau khi xử lý Outlier

# Data Encoding

```
## create function convert object to numeric

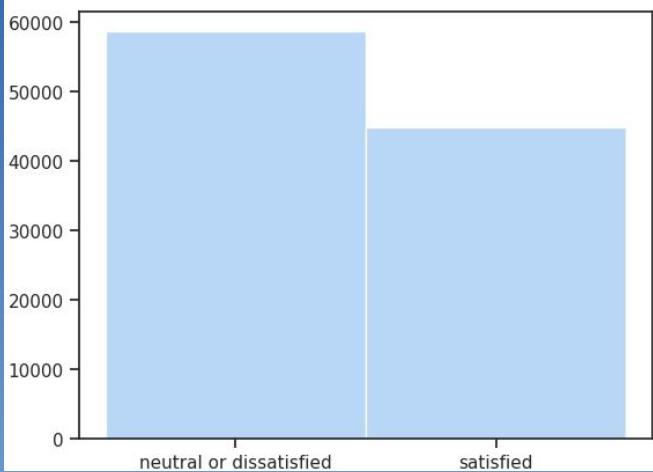
def label_encode(df):
    df['Gender'] = df['Gender'].map({'Male':0,'Female':1})
    df['satisfaction'] = df['satisfaction'].map({'neutral or dissatisfied':0,'satisfied':1})
    df['Customer Type'] = df['Customer Type'].map({'Loyal Customer':1,'disloyal Customer':2})
    df['Type of Travel'] = df['Type of Travel'].map({'Business travel':1,'Personal Travel':2})
    df['Class'] = df['Class'].map({'Eco':1,'Eco Plus':2,'Business':3})
    return df

## Encoding for train table
label_encode(df_train)

## Encoding for test table
label_encode(df_test)
```

Map các feature chuỗi sang các giá trị 0-3 để thực hiện các bước mô hình hóa tiếp theo

# Imbalance Data



- Sử dụng **SMOTE** để xử lý thiên lệch dữ liệu của cột kết quả
- Chia tập dữ liệu train thành tập 2 thành tập nhỏ train và valid với tỷ lệ 80 - 20
- **Mục đích:** So sánh kết quả khi predict giữa 2 tập valid và tập test thực tế để xác định mô hình có xảy ra over/under fitting hay không.

```
## Gán biến data train
X_train = df_train.drop(['satisfaction'],axis =1)
y_train = df_train['satisfaction']

## Gán biến data test
X_test = df_test.drop(['satisfaction'],axis =1)
y_test = df_test['satisfaction']

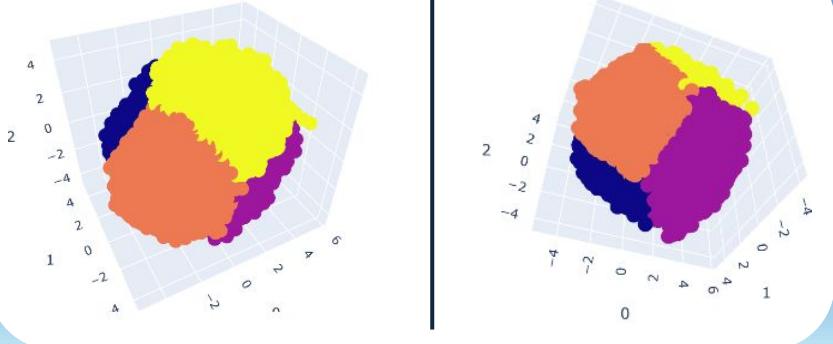
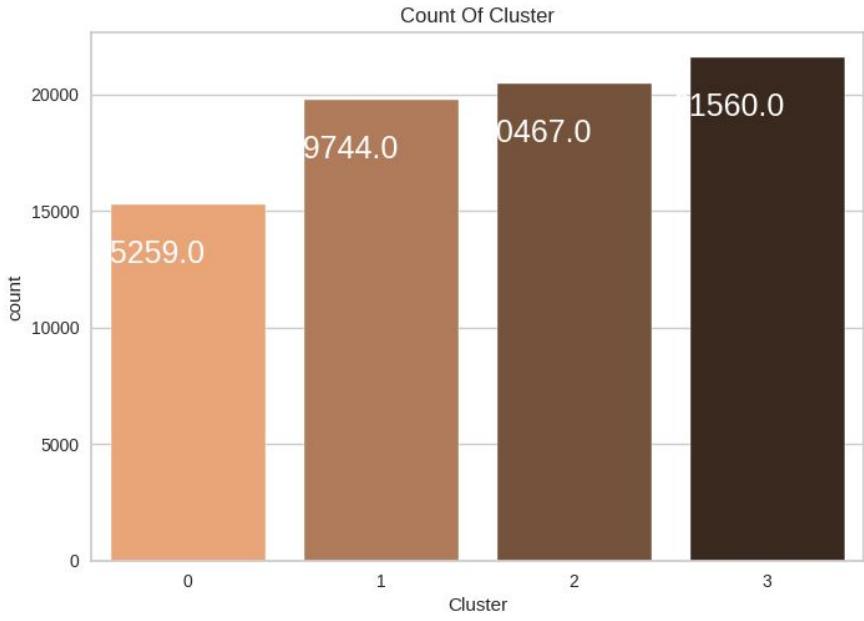
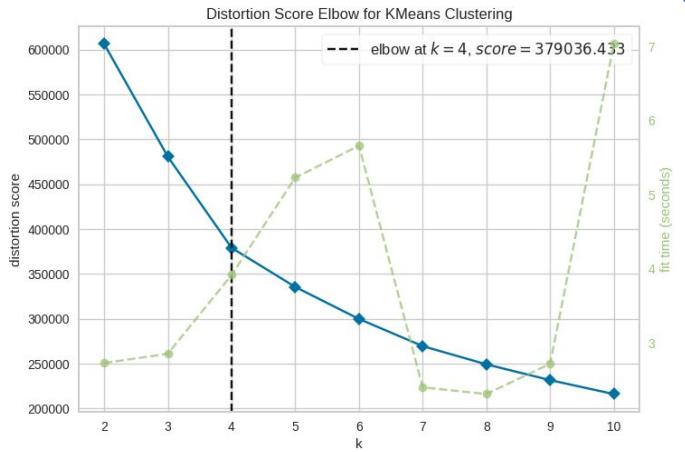
## Resample data
from imblearn.over_sampling import SMOTE
sm = SMOTE(random_state=42)
X_train, y_train = sm.fit_resample(X_train, y_train)
```

03



# Customer Segmentation

# Customer Segmentation



Dựa vào **KElbow** cùng giảm chiều dữ liệu bằng **PCA**, ta thu được **4 Clusters**

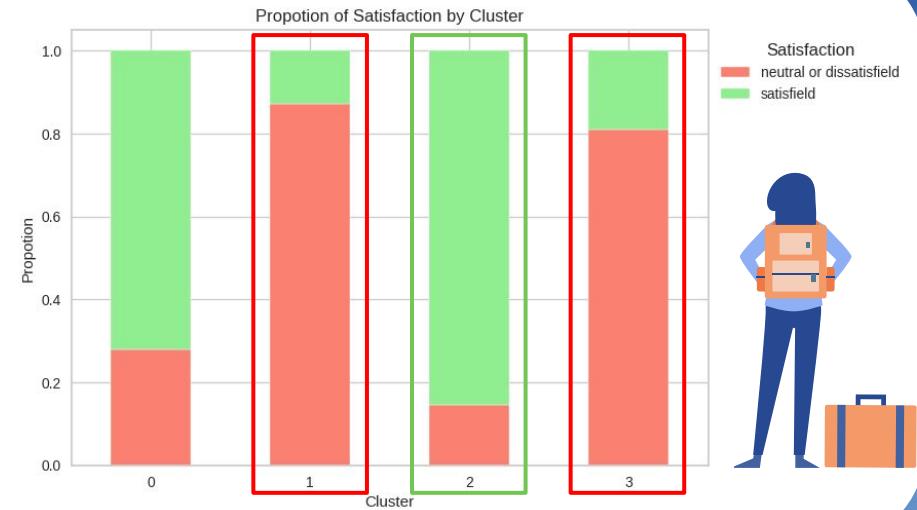


# Customer Segmentation

Nhìn vào target “Neutral or Dissatisfaction”

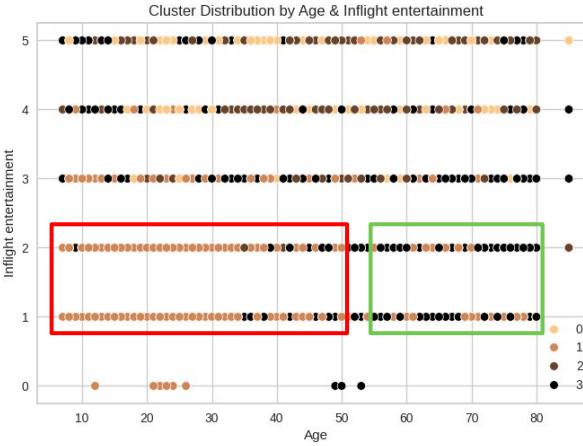
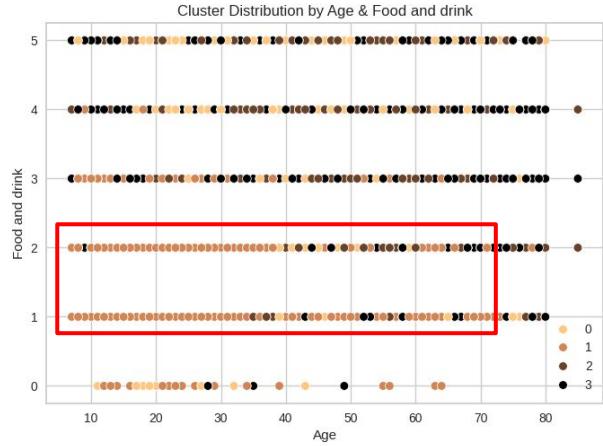
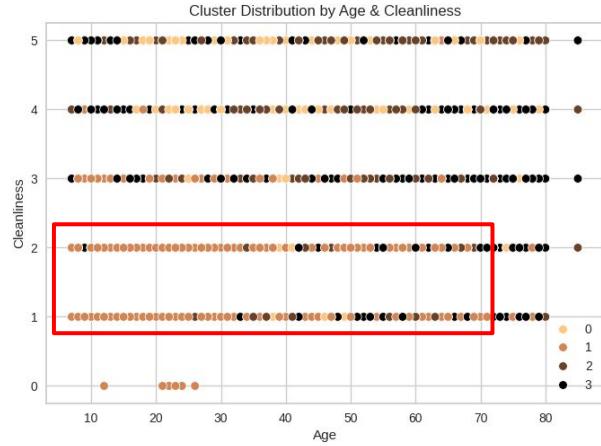
Cluster 1,3 → Nhóm **Không hài lòng** nhiều nhất

Cluster 2 → Nhóm **Hài lòng** nhiều nhất



		Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	mean score of all service
Cluster	satisfaction															
0	0	2.112564	2.664176	1.812704	2.423400	4.230967	2.250584	4.148762	4.438580	4.045072	3.480149	4.284447	3.713218	4.353807	4.249883	3.443451
1	0	2.254440	3.139002	2.434591	2.985432	1.902786	2.221416	2.011724	2.023680	3.257226	3.236042	3.731225	3.149855	3.762391	1.896982	2.714771
2	0	3.495327	4.276368	3.731308	3.492323	3.926569	3.729306	3.992323	4.273364	4.035714	3.624166	4.268692	3.632844	4.323097	3.927570	3.909212
3	0	2.433429	3.052776	2.623125	3.005839	3.580824	3.001202	3.641214	3.150544	2.392444	2.502061	2.682770	2.742931	2.691872	3.490899	2.927995
0	1	1.618930	1.615742	1.580760	1.695910	3.609547	3.942425	4.175913	4.303453	4.172816	4.094561	4.237770	3.745923	4.252528	3.909720	3.354000
1	1	3.026253	2.863564	2.944710	2.856802	1.649562	2.765712	1.727526	1.853222	3.505967	3.319411	3.866746	3.391408	3.883850	1.638425	2.806654
2	1	4.172743	3.913457	4.040753	3.844084	3.664186	4.270391	4.135653	4.255223	4.098678	4.033827	4.190945	3.714842	4.209376	3.903211	4.031955
3	1	3.110024	2.712958	2.649633	2.799022	3.693643	3.895110	3.917115	3.078240	2.197066	2.383863	2.207824	3.057946	2.201956	3.792665	2.978362

# Customer Behavior



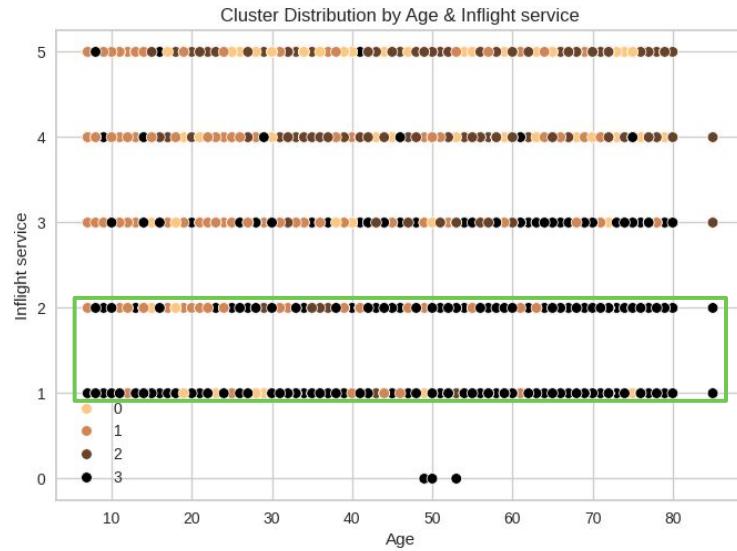
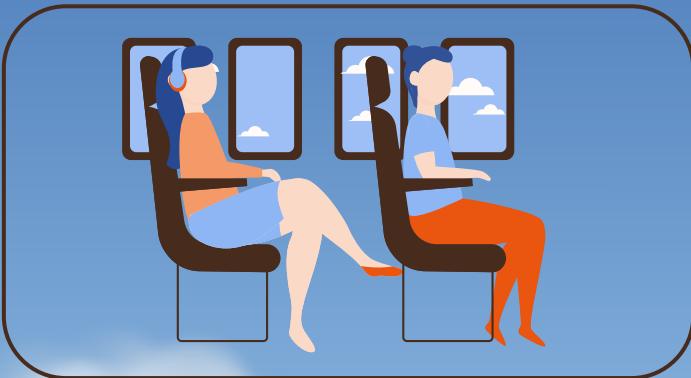
**Cluster 1 :** Tập trung vào các dịch vụ về giải trí, ăn uống và vệ sinh trên chuyến bay.

**Cluster 3 :** Cũng có dấu hiệu cho thấy nhóm này chú ý nhiều đến dịch vụ giải trí khi bay.



# Customer Behavior

**Cluster 3 →** Khẳng định thêm về nhóm khách hàng này có sự tập trung rất nhiều vào các dịch vụ được cung cấp trên chuyến bay

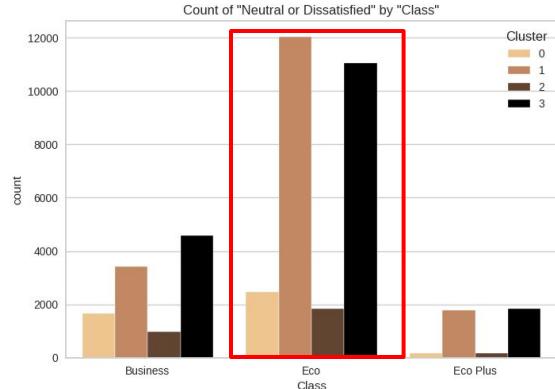
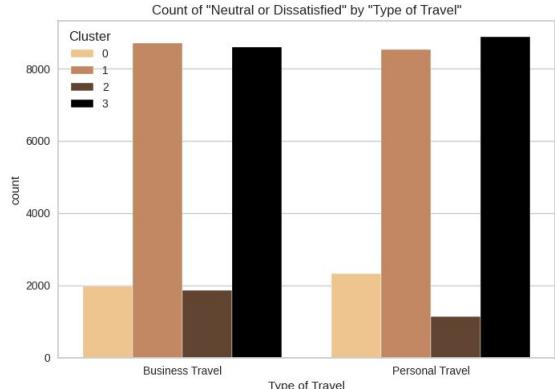
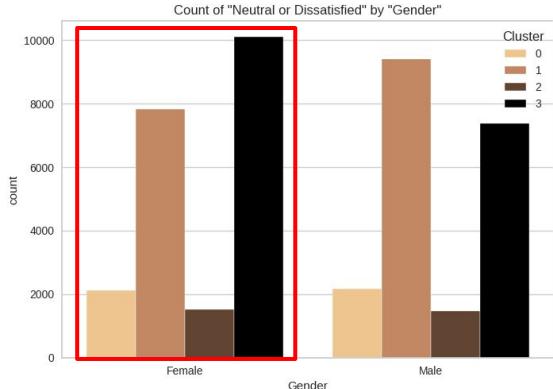
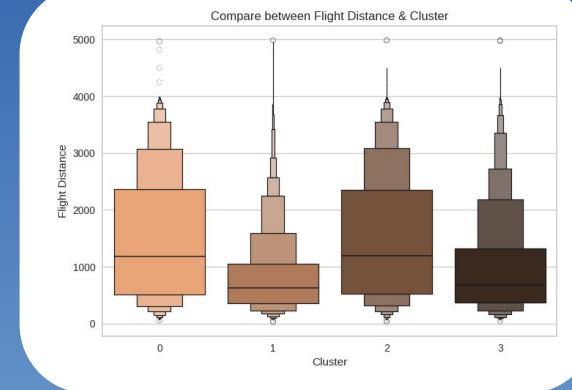
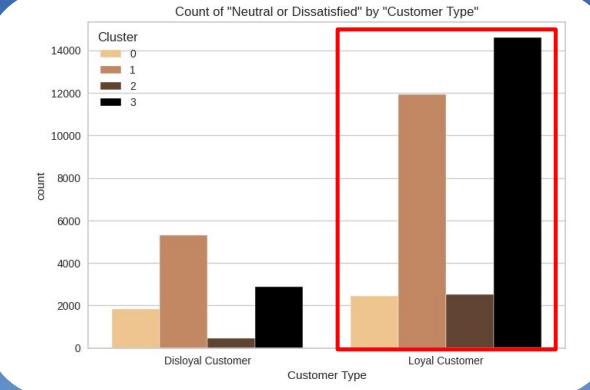


# Customer Demographics

Dựa vào target “Dissatisfaction”

Cluster 1,3 → Có một số đặc điểm như sau:

- Chủ yếu là khách hàng loyal
- Hơi thiên về nữ giới.
- Thường bay hạng Eco.



# How to improve?

## Về dịch vụ :

- Nâng cấp các phương tiện giải trí
- Cải thiện các dịch vụ cung cấp khi bay
- Chăm chút về cơ sở vật chất và vệ sinh

## Về khách hàng :

- Xem lại cách chăm sóc với khách hàng Eco Class
- Kiểm tra các ưu đãi với khách hàng trung thành
- Tăng khảo sát khách hàng để hiểu thêm về nhu cầu cần thiết của khách đi bay



04

# Modeling & Prediction



# Best Model Selection

```
# Here are some algorithms that will be tested to determine the best model:  
from sklearn.linear_model import LogisticRegression  
from sklearn.naive_bayes import GaussianNB  
from sklearn.neighbors import KNeighborsClassifier  
from sklearn.ensemble import AdaBoostClassifier, RandomForestClassifier  
from sklearn.tree import DecisionTreeClassifier  
  
# Metric Evaluation  
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix, recall_score  
  
logistic = LogisticRegression(random_state=0)  
bayes = GaussianNB()  
knn = KNeighborsClassifier()  
adaboost = AdaBoostClassifier(random_state=0)  
randomForest = RandomForestClassifier(random_state=0)  
dt = DecisionTreeClassifier(random_state=0)  
  
modellist = {"Logistic": logistic,  
             "Naive Bayes": bayes,  
             "KNN": knn,  
             "AdaBoost": adaboost,  
             "Random Forest": randomForest,  
             "Decision Tree": dt}
```

```
from sklearn.model_selection import cross_validate
from sklearn.metrics import f1_score

scoring = ['precision_macro', 'recall_macro', 'f1_macro']

for name_model, model in modellist.items():
    print(f"Running model {name_model}")
    scores = cross_validate(model, X_train, y_train, scoring=scoring, cv = 5)
    print("F1 Score %.2f accuracy with a standard deviation of %.2f"
          % (scores['test_f1_macro'].mean(), scores['test_f1_macro'].std()))
    print("=====")

Running model Logistic Regression
F1 Score 0.78 accuracy with a standard deviation of 0.01
=====
Running model Naive Bayes
F1 Score 0.81 accuracy with a standard deviation of 0.01
=====
Running model KNN
F1 Score 0.67 accuracy with a standard deviation of 0.01
=====
Running model AdaBoost
F1 Score 0.92 accuracy with a standard deviation of 0.01
=====
Running model Random Forest
F1 Score 0.95 accuracy with a standard deviation of 0.01
=====

Running model Decision Tree
F1 Score 0.93 accuracy with a standard deviation of 0.01
=====
```

- Thử nghiệm model với 6 mô hình bao gồm: **Logistic, Naive Bayes, KNN, Adaboost, Random Forest, Decision Tree.**
  - Kết quả cho thấy mô hình **Random Forest** cho kết quả tốt nhất với **Accuracy 95%** và **độ sai lệch 0.01**

# Random Forest Model

## Evaluation

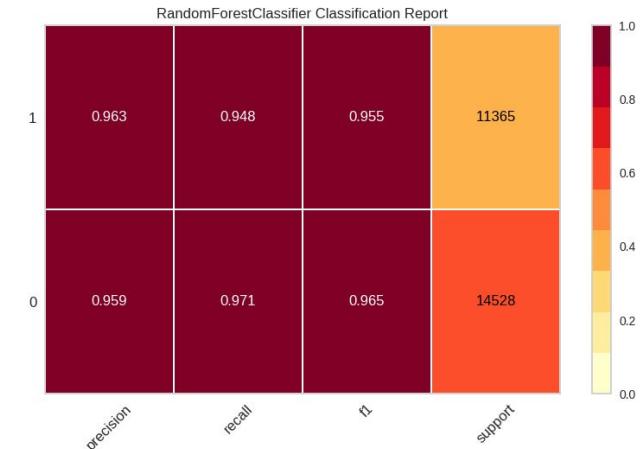
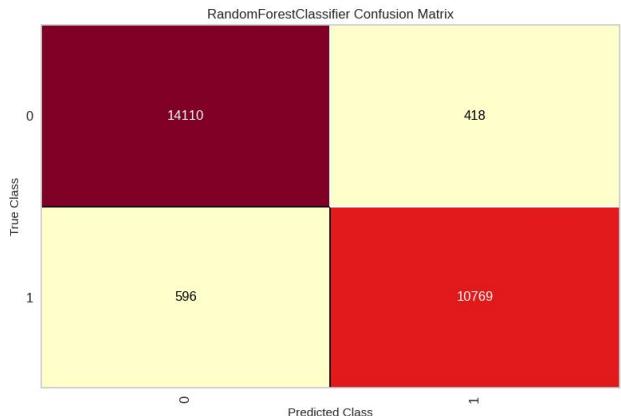


F1 Score: ~96.5%

Accuracy score: ~96%

Precision score: ~95.9%

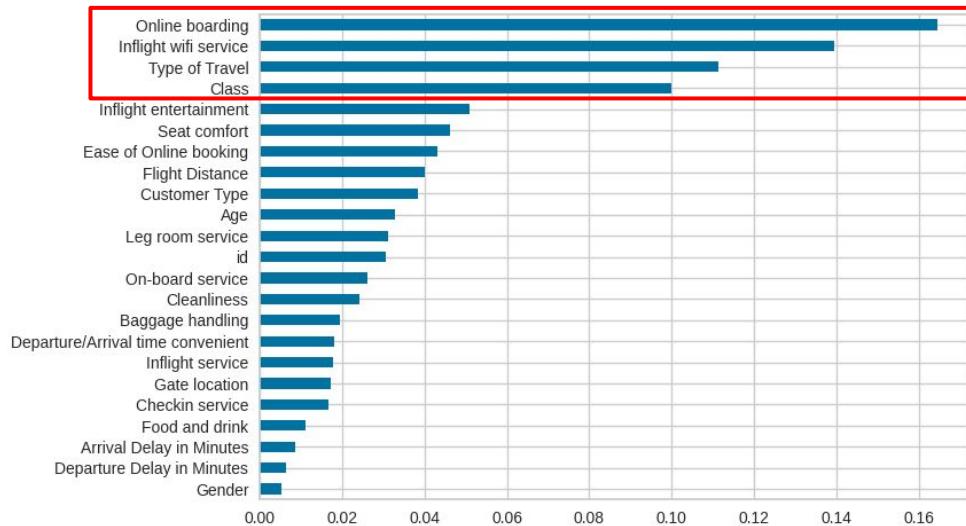
Recall score: ~97.1%



# Feature Importance

```
pd.Series(randomForest.feature_importances_, index=list(X_train.columns)).sort_values().plot(kind = 'barh')
```

<Axes: >



Đây là top những dịch vụ tạo ra ảnh hưởng đến sự hài lòng của khách hàng nhất

1  
Online Boarding



2  
Inflight Wifi Service



3  
Type of Travel



4  
Customer Class



# Conclusion

**URGENT  
IMPROVEMENT**

## Cải thiện Cơ sở vật chất

- Wifi
- Online services (Booking, Boarding)
- Gate location



**FURTHER  
IMPROVEMENT**

## Duy trì phát triển các hạng mục chủ chốt

- Cleanliness
- Entertainment



- Hợp tác với công ty mạng và nhà cung cấp để nâng cấp đường truyền mạng của máy bay
- Kết hợp với nhà tư vấn phát triển phần mềm để hoàn thiện các ứng dụng online Booking/Boarding của hãng

- Đảm bảo duy trì và nâng cấp chất lượng vệ sinh hãng bay
- Phát triển các dịch vụ giải trí mới dành cho đa dạng đối tượng khách hàng



# Functions and responsibilities of the role



## Project objective

- Tìm hiểu về nhu cầu của khách hàng và phản ứng thông qua điểm đánh giá đối với dịch vụ của hãng bay cung cấp.
- Ứng dụng Machine Learning để phân tích, tìm hiểu và dự đoán các yếu tố ảnh hưởng đến sự hài lòng, không hài lòng của khách hàng.



## Teamwork

- **Data Preprocessing:** All
- **EDA :** Linh
- **Clustering :** Việt
- **Modeling & Predict :** Triều
- **Slide & Presenting :** Linh - Việt - Triều



# Thank you!

Do you have any questions?

