

Big Data and Cloud Computing

CMP 4011

Phase 2

|  |  |  |
| --- | --- | --- |
| **Team 5** | | |
| **Name** | **Section** | **BN** |
| **Aya Ahmed Musad** | **1** | **14** |
| **Salma Ragab Hassan** | **1** | **30** |
| **Shredan Abdallah** | **1** | **32** |
| **Nada Osman** | **2** | **30** |

# Problem statement

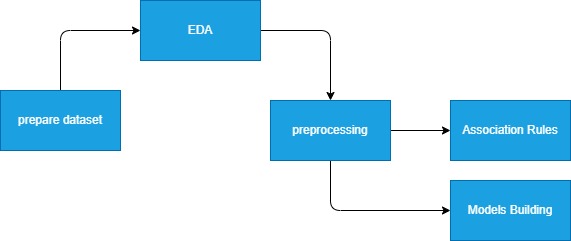
Due to aggressive competition in the airline industry, airline companies need to focus on passenger satisfaction. And customer feedback is critical since it is a consequence measurement for business performance.

The purpose of this notebook is to examine how Airlines can aim to determine the importance of each feature with regards to their contribution to passenger satisfaction, and in turn, boost their brand image and their customers’ loyalty and engagement by leveraging customer expectation.

# Dataset

|  |  |
| --- | --- |
| Link | <https://www.kaggle.com/code/khsamaha/customer-satisfaction-analysis-modelling-r/input> |
| Number of features | 25 unique features |
| Number of records | 103,904 |
| Size | 11.99 MB |

# Project Pipeline



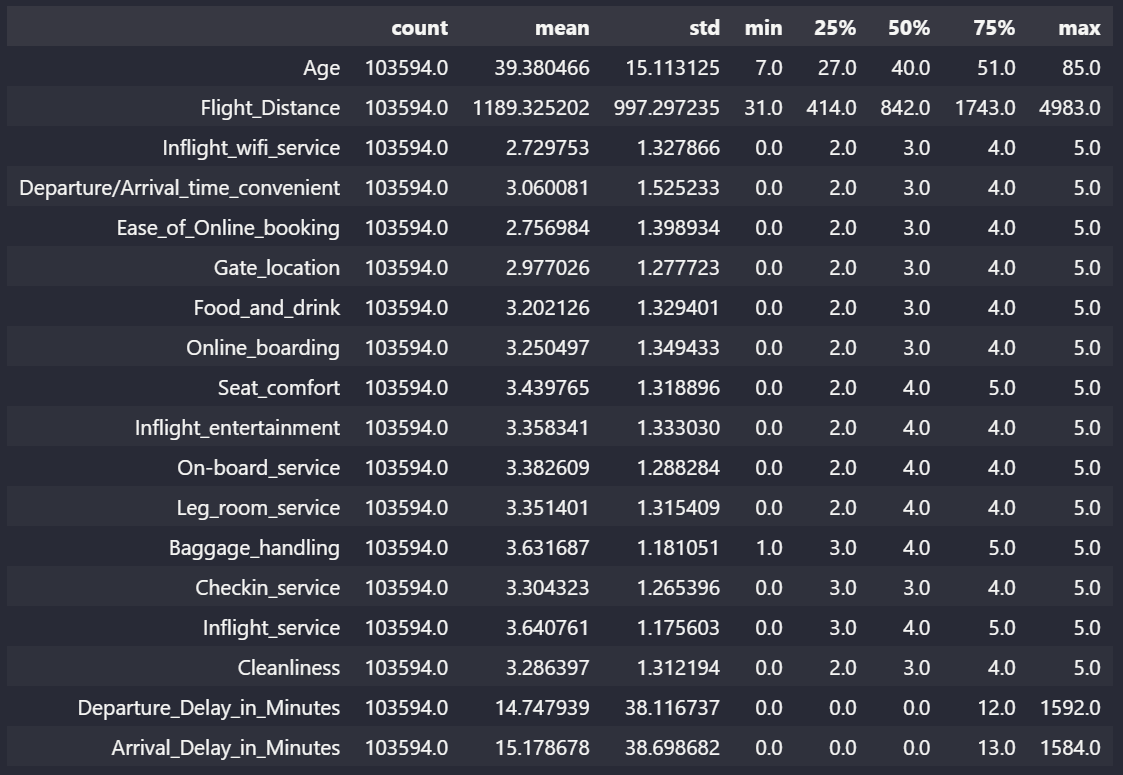
# Analysis and Solution of the Problem

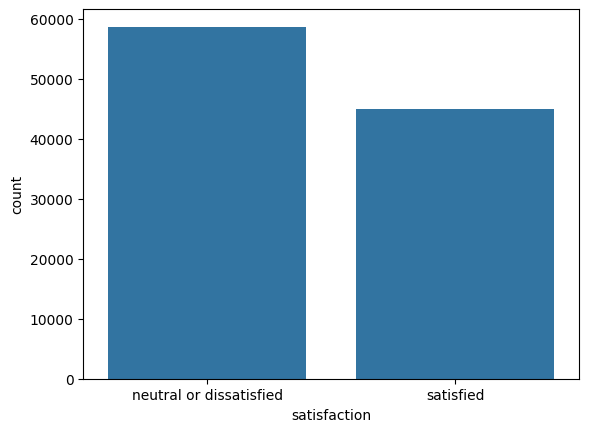
* **Prepare dataset**

1. Collecting Dataset .
2. Replacing Nan and null values from columns with column mean .

* **EDA , Observations and Insights**

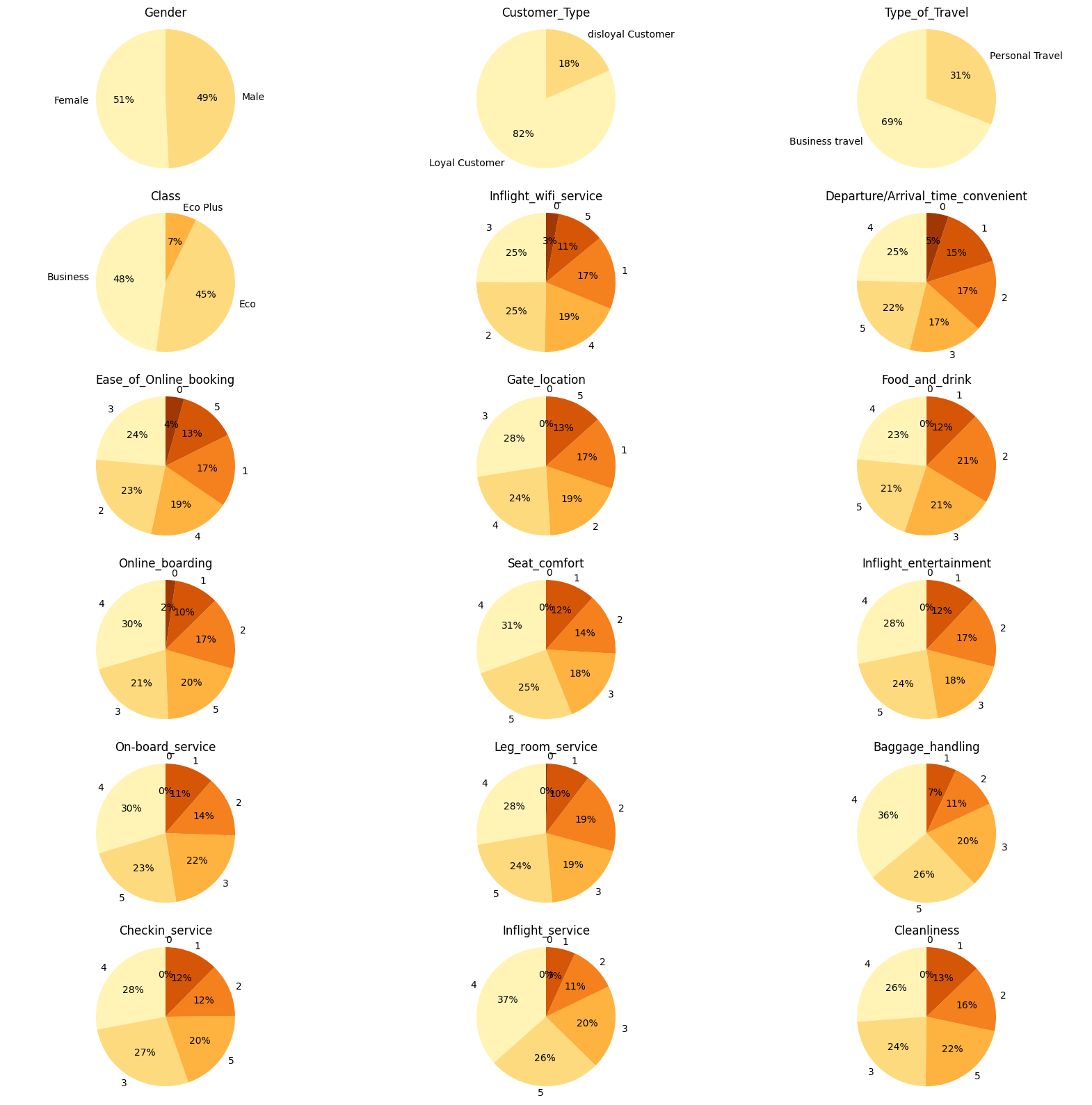
1. Data Description:



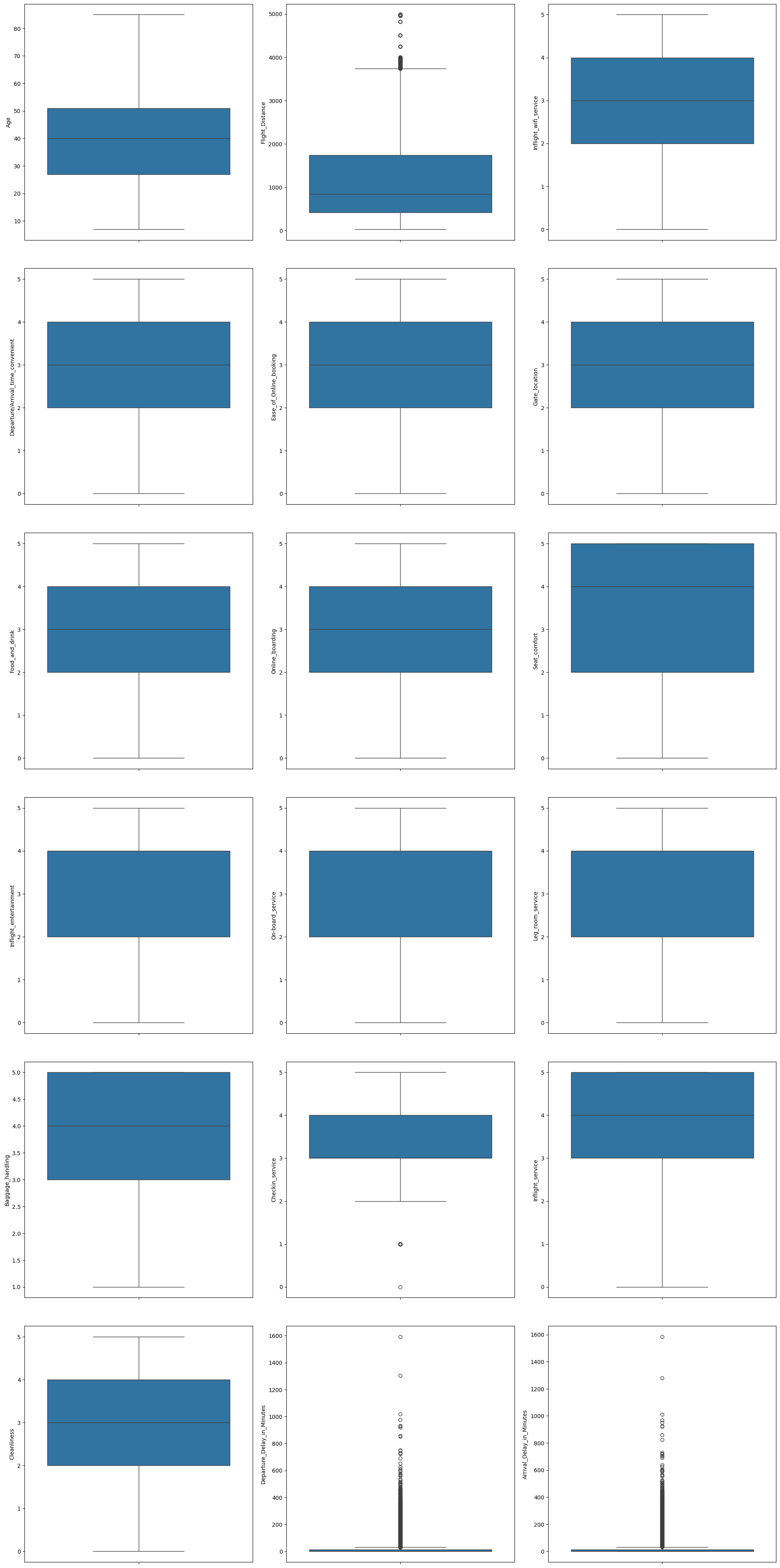


from this appears that we have to improve our services in order to make customers more satisfied

1. Pie chart for each feature to know its distribution:



1. Box Plot for each feature:

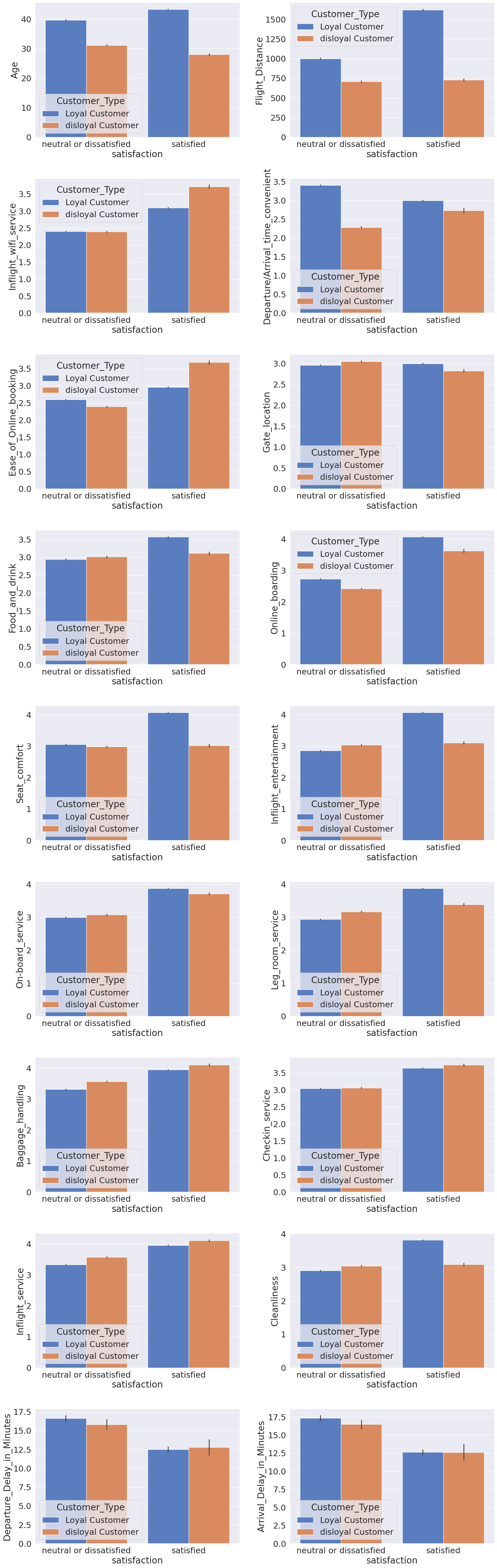
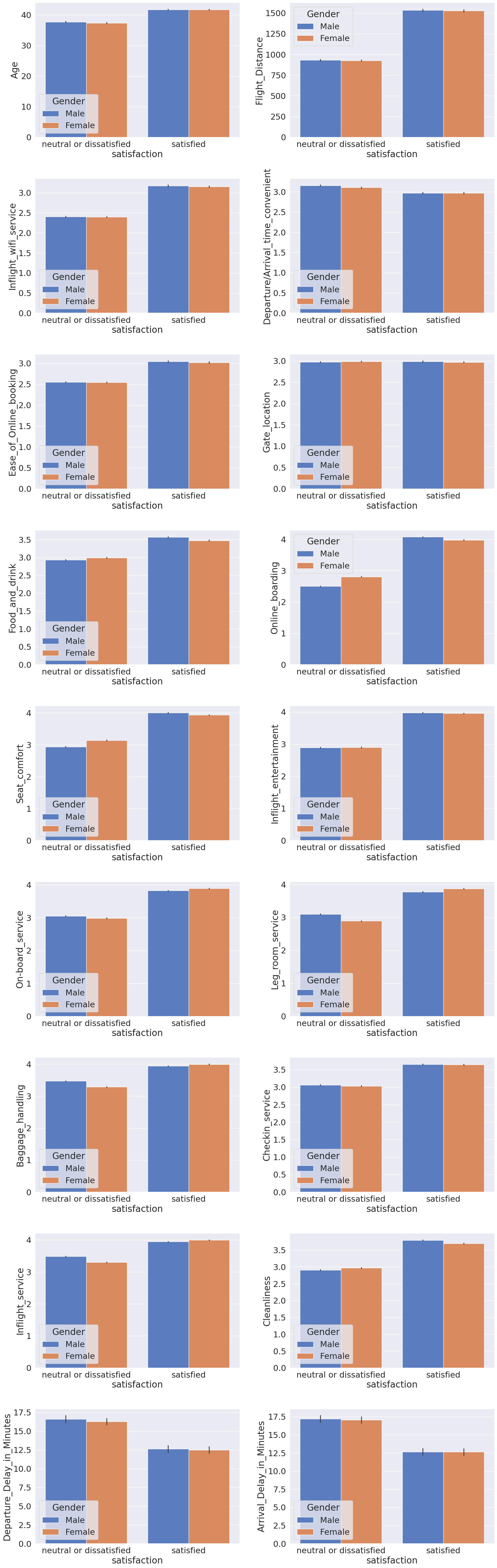


It can be seeing that the delay time minutes exists so many outliers.

Generally, passenger satisfy with the baggage handling and inflight service where the score ranges from 3 to 5 while others mainly stay within 2 to 4 scores.

1. BarPlot based on some important features:

We compare customer features with services of air line features in order to know level of satisfaction of customer type about our sevice .



- The long flight distance make them more likely to satisfy the trip.

But in average of delay time, they may not satisfy when time is 12.5 mins above.

Between the male and female, there are no significant patterns.

so this column or the information of if the customer male or female is not important as both have almost same pattern .

- For loyal customer:

Age around 40:

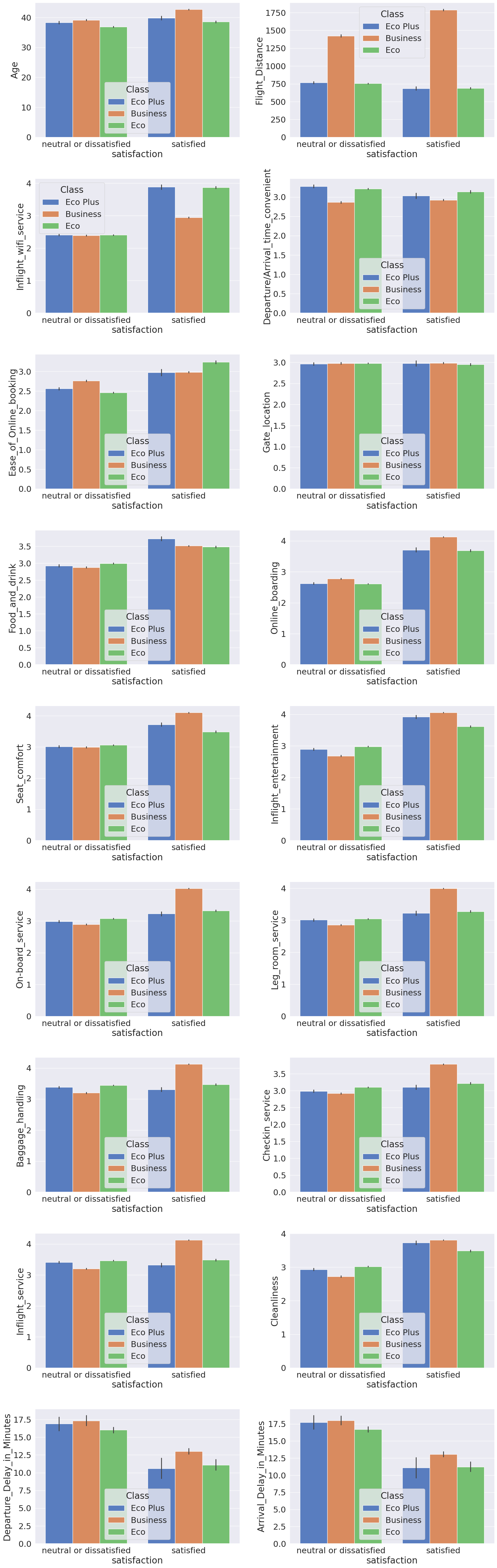
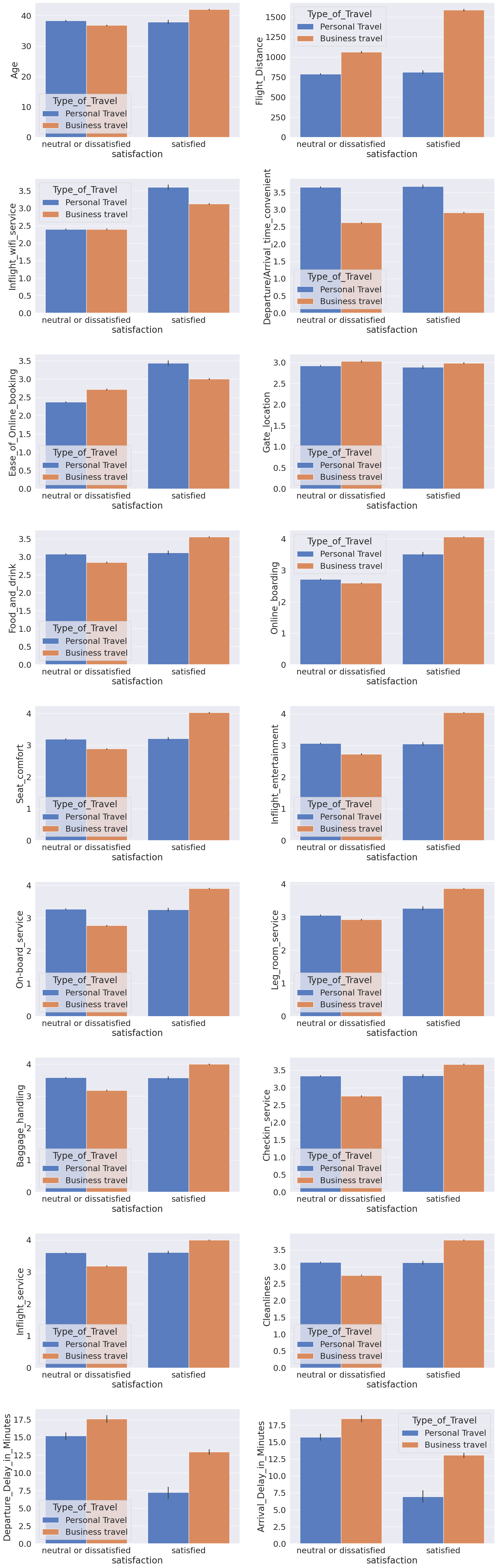
Satisfy on high flight distance,seat comfort, inflight entertainment, cleanliness,leg room service,chekin service,baggaging service

For disloyal customer:

Age around 30:

Satisfy on Inflight wifi service, ease of online booking,online boarding,on board service ,chekin service

Again, they generally feel neutral or dissatisfied when the delay time minutes is 12.5 or above .



- Business travel:

satisfaction increase on long flight distance,online boarding,seat comfort,inflight entertaiment,on board service,food and drink almost all features except gate location

they generally feel neutral or dissatisfied when the delay time minutes is 12.5 or above

Personal travel:

satisfaction increase on inflight wifi service,ease of online booking,online boarding

they generally feel neutral or dissatisfied when the delay time minutes is almost below 7.5 or above .

- Business Class:

satisfaction increase on long flight distance,inflight wifi service,inflight service,online boarding,seat comfort,inflight entertaiment,on board service,food and drink almost all features except gate location,ease of online booking and arrival time

they generally feel neutral or dissatisfied when the delay time minutes is 12.5 or above

Eco:

all features except long flight distance,gate location,checkin services,baggage handling,inflight service and arrival time

they generally feel neutral or dissatisfied when the delay time minutes is 11.25 or above

Eco Pluse:

satisfaction increase on inflight wifi service,online boarding,seat comfort,inflight entertaiment,on board service,food and drink ,ease of online booking .

they generally feel neutral or dissatisfied when the delay time minutes is 11 or above .



child:

they generally feel neutral or dissatisfied when the delay time minutes is 4 or above

teenager:

they generally feel neutral or dissatisfied when the delay time minutes is 8 or above

youth:

they generally feel neutral or dissatisfied when the delay time minutes is 12.5 or above

middle age:

satisfied most by long flight deistance more tha 1000

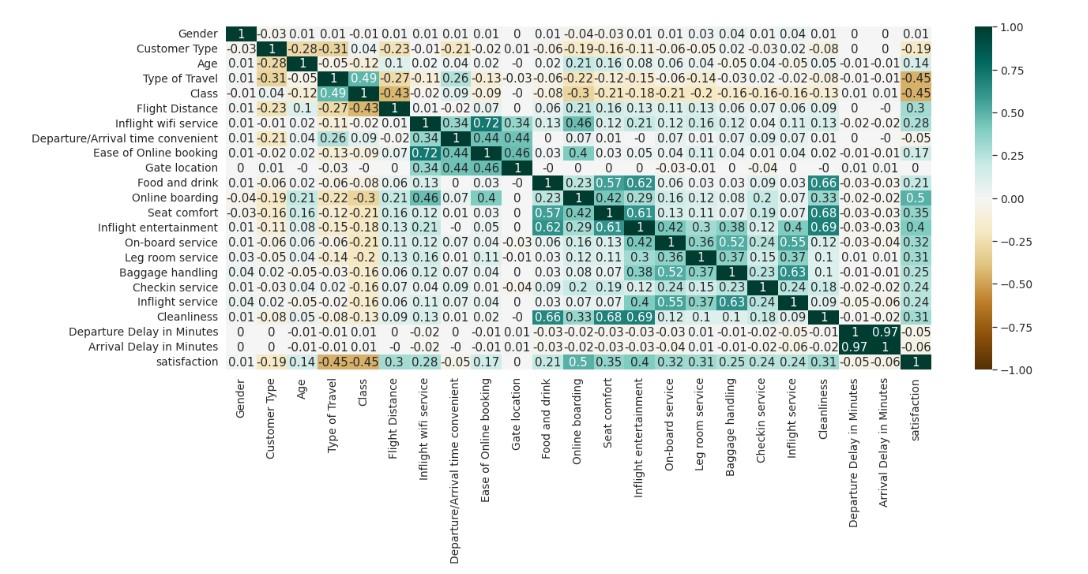
they generally feel neutral or dissatisfied when the delay time minutes is 13 or above

old:

they generally feel neutral or dissatisfied when the delay time minutes is 10 or above

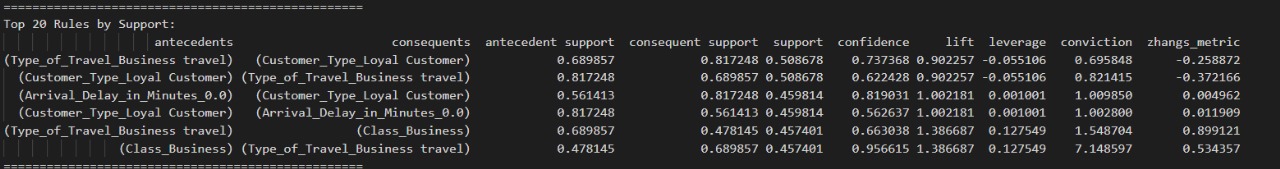
for all satisfaction increase on long flight distance,inflight wifi service,online boarding,seat comfort,inflight entertaiment,on board service,food and drink almost all features except gate location and arrival time

1. Correlation Matrix

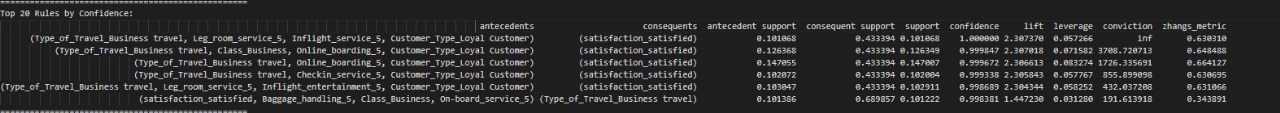


* **Assosiation Rules**

1. Items with the highest support

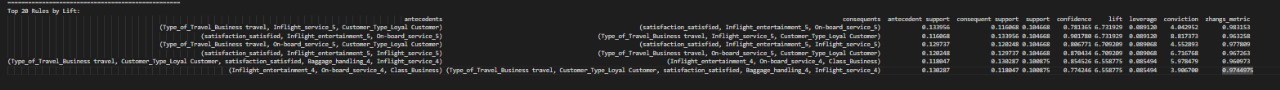


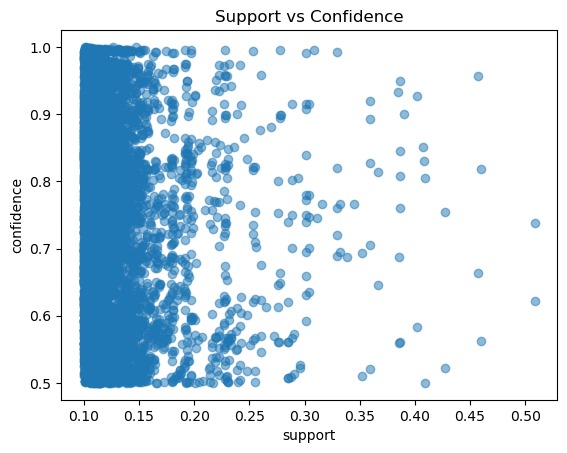
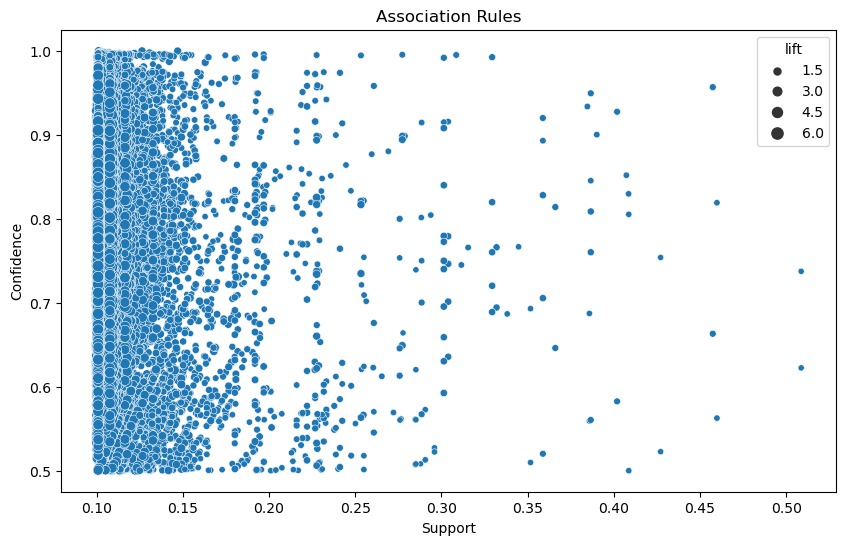
1. Items with the highest confidence

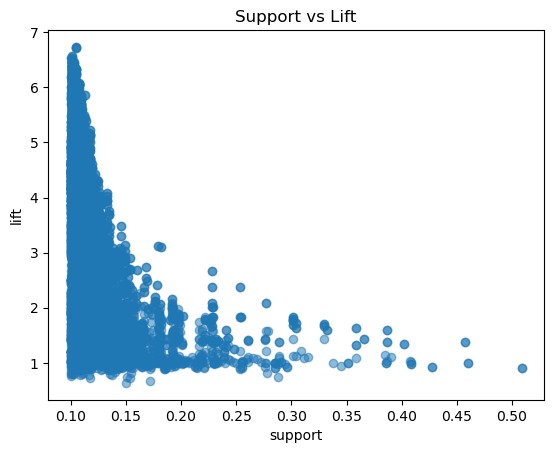


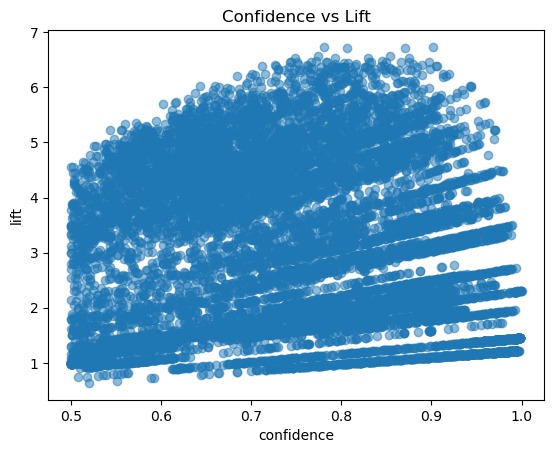
1. Items with the highest lift

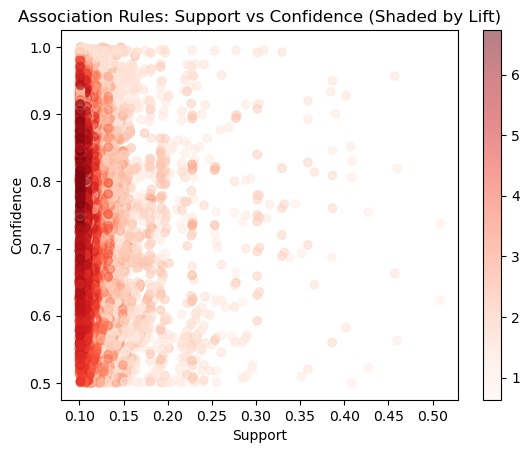
NB: photo in the notebook if it is not clear here











# Model Training and Evaluation

**Naive Bayes Classifier**:

**Class 0**:

Prior Probability: 0.5674241786268541

Mean of Features: [-0.00878688 0.16366464 -0.12188124 0.39356015 -0.26187529 -0.24593171

0.04247212 -0.14749529 0.00063657 -0.18320678 -0.44208835 -0.30718172

-0.34848612 -0.28006531 -0.27012803 -0.21747635 -0.2046829 -0.21379959

-0.26828008 0.04635538 0.0525161 -0.44460024 0.39859175]

Variance of Features: [0.9996606 1.24202348 1.18720436 1.16773101 0.62795616 0.52834782

0.97063841 0.74130625 0.87540419 1.02254278 0.71765473 0.97385293

0.98208467 0.99565776 0.98118768 0.99368769 1.02476791 0.99695796

1.01730422 1.12425615 1.11800969 0.76314879 0.92142839]

**Class 1**:

Prior Probability: 0.4325758213731458

Mean of Features: [ 0.01611528 -0.2131559 0.15400927 -0.51324866 0.34169377 0.32344607

-0.05946848 0.19645956 0.00113706 0.23921903 0.57370001 0.39741848

0.45106775 0.37000193 0.35886552 0.28087 0.27151055 0.2781707

0.34431756 -0.05862983 -0.06701757 0.57780177 -0.51646811]

Variance of Features: [1.00022116 0.60446846 0.71391 0.31584065 1.28034887 1.43447049

1.03972065 1.2717667 1.15956903 0.86238893 0.78475332 0.75321597

0.65615493 0.76242413 0.795827 0.8696001 0.83760966 0.86436733

0.76097647 0.8627543 0.87081374 0.71706577 0.62920851]

Accuracy: 0.857754679755546

Classification Report:

precision recall f1-score support

0 0.86 0.89 0.88 11713

1 0.85 0.82 0.83 9068

accuracy 0.86 20781

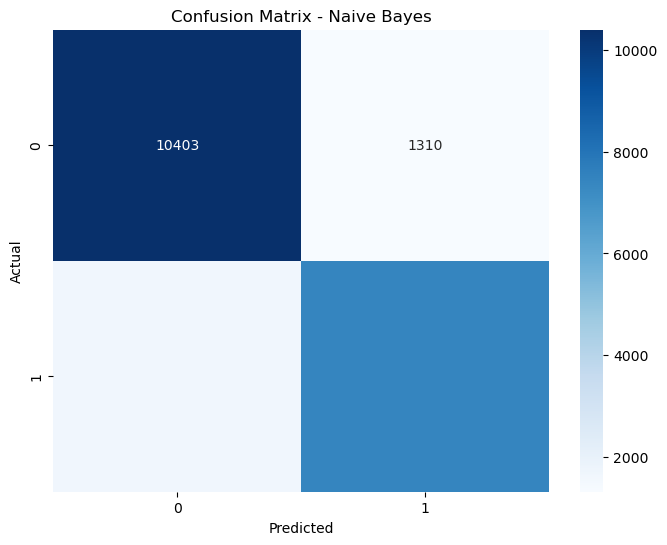
macro avg 0.86 0.85 0.85 20781

weighted avg 0.86 0.86 0.86 20781

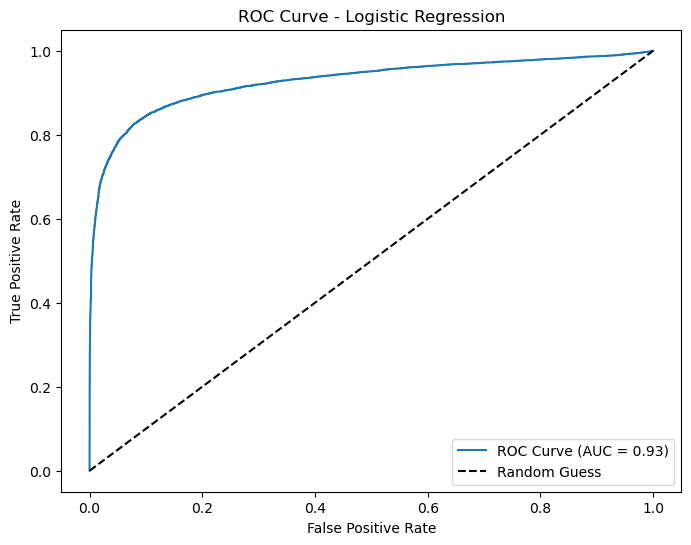
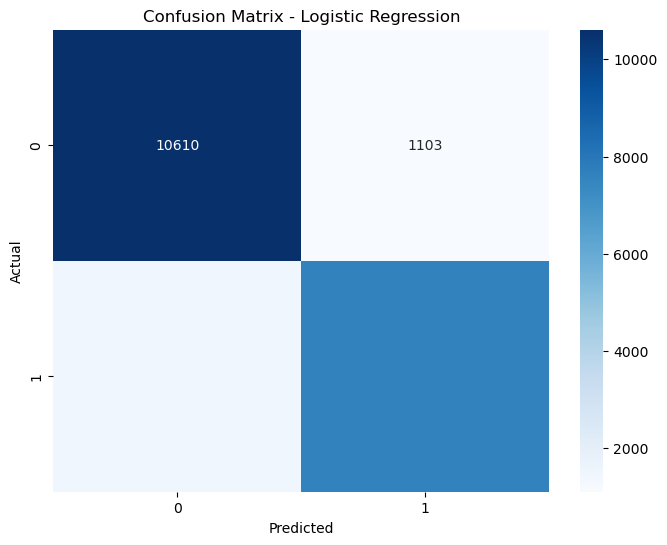
Misclassification Rate for Class 0 (Negative Class): 0.14224532024445408

Misclassification Rate for Class 1 (Positive Class): 0.14224532024445408

Misclassification Rate for both classes is 0.14 which is not so high , so accurancy gives us a good indication here about the model , and f1 score approximately equals accurancy .



**Logistic Regression** :



Logistic Regression Summary:

Accuracy: 0.8765699436985708

Classification Report:

precision recall f1-score support

0 0.88 0.91 0.89 11713

1 0.87 0.84 0.86 9068

accuracy 0.88 20781

macro avg 0.88 0.87 0.87 20781

weighted avg 0.88 0.88 0.88 20781

Area Under the Curve (AUC): 0.9274453797222206

Each point on the ROC (Receiver Operating Characteristic) graph represents a different threshold for classifying a binary classification model's output as positive or negative. The ROC graph plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

Model Coefficients: [ 0.02618162 -0.7809889 -0.12543714 -1.25570493 -0.01714768 0.51587839

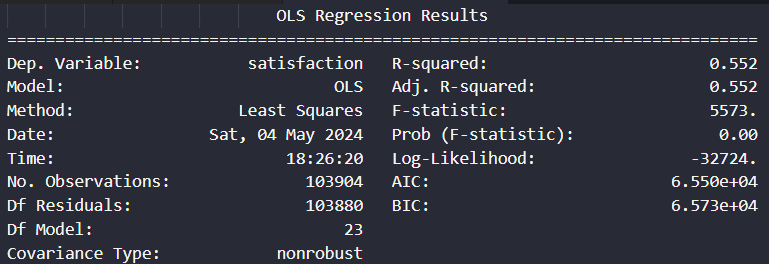
-0.17846334 -0.20658948 0.0361048 -0.02892152 0.8298964 0.08883714

0.07775466 0.39641602 0.333328 0.15684452 0.40657153 0.13765398

0.28366956 0.16240454 -0.34908739 0.44206753 0.06556975]

Model Intercept: -0.5075160746265843

**Linear Regression** :



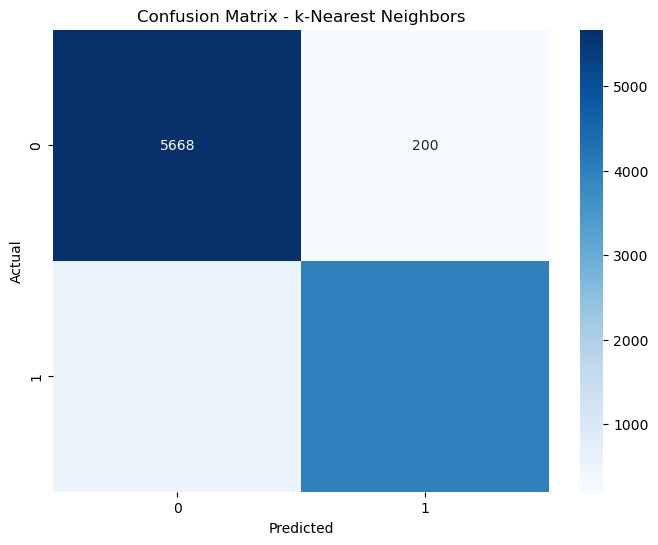


Mean Squared Error (MSE): 0.1099219394433318

**KNN Pyspark Implementation :**

Accuracy of KNN with K=5: 92.79%

**KNN SKlearn Implementation :**



Accuracy: 93.04205562506015

It is to be noticed that accurancies are so approximate .

# Unsuccessful Trials

1. Tried to remove 3 correlated features before classification : Gate\_location , Departure\_Delay\_in\_Minutes , Ease\_of\_Online\_booking but R-squared , mse of linear regression has dropped , also tried to remove the first 2 of them but got same results .