Institute of Technology of Cambodia
Department of Applied Mathematics and Statistic

Group: I3-AMS-03

Subject: Advance Probability

Topic: Analyzing the factors that affect customer satisfaction in the airline industry using ordinal logistic regression

Name:	ID:
• Tang Piseth	e20201634
• Kry Senghort	e20200706
• Sao Samarth	e20200084
• Hon Rathana	e20201053
• Pheng Sothea	e20201734
• Ek Vong Panharith	e20200877

Table of contents:

I.) Project Description and Objectives

Airline Industry is the field of journey tourism or visiting to any places by using airline for travelling from one place to another one. On the other hand, We choose this topic beacause we want to use logistic linear regression to interpret and make prediction the result of satisfaction or disatifaction of passenger on airline industry. We will study on the variable that make effect to satisfaction or disatisfaction of customer.

2). Dataset and Variables Description

There are two data sets such as : test.csv and train.csv. Firstly, we test on file train.csv for EDA in which existing 25976 rows with 24 columns. There exists 24 variables as following:

- id (int64) (int64)
- Gender (object)
- Customer Type (object)
- Age (int64)
- Type of Travel (object)
- Class (object)
- Flight Distance (int64)
- Inflight wifi service (int64)
- Departure/Arrival time convenient (int64)
- Ease of Online booking (int64)
- Gate location (int64)
- Food and drink (int64)
- Online boarding (int64)
- Seat comfort (int64)
- Inflight entertainment (int64)
- On-board service (int64)
- Leg room service (int64)
- Baggage handling (int64)
- Checkin service (int64)
- Inflight service (int64)
- Cleanliness (int64)
- Departure Delay in Minutes (int64)
- Arrival Delay in Minutes (float)
- Satisfaction (object)

3). Data Preparation

After checking to our dataset we observe that there exist some missing values and also we decide to handle it by substituting with the mean values for numerical variable and mode for the categorical variables. As the result, feature "Arrival_Delay_in_Minutes" existing 393 missing values.

```
id                                      0
Gender                                  0
Customer_Type                           0
Age                                     0
Type_of_Travel                          0
Class                                   0
Flight_Distance                         0
Inflight_wifi_service                   0
Departure/Arrival_time_convenient       0
Ease_of_Online_booking                  0
Gate_location                           0
Food_and_drink                          0
Online_boarding                         0
Seat_comfort                            0
Inflight_entertainment                  0
On-board_service                        0
Leg_room_service                        0
Baggage_handling                        0
Checkin_service                         0
Inflight_service                        0
Cleanliness                             0
Departure_Delay_in_Minutes              0
Arrival_Delay_in_Minutes              393
satisfaction                            0
dtype: int64
```

Descriptive Statistics Before Handle Missing

	count	mean	std	min	25%	50%	75%	max
id	129880.0	64940.500000	37493.270818	1.0	32470.75	64940.5	97410.25	129880.0
Age	129880.0	39.427957	15.119360	7.0	27.00	40.0	51.00	85.0
Flight_Distance	129880.0	1190.316392	997.452477	31.0	414.00	844.0	1744.00	4983.0
Inflight_wifi_service	129880.0	2.728696	1.329340	0.0	2.00	3.0	4.00	5.0
Departure/Arrival_time_convenient	129880.0	3.057599	1.526741	0.0	2.00	3.0	4.00	5.0
Ease_of_Online_booking	129880.0	2.756876	1.401740	0.0	2.00	3.0	4.00	5.0
Gate_location	129880.0	2.976925	1.278520	0.0	2.00	3.0	4.00	5.0
Food_and_drink	129880.0	3.204774	1.329933	0.0	2.00	3.0	4.00	5.0
Online_boarding	129880.0	3.252633	1.350719	0.0	2.00	3.0	4.00	5.0
Seat_comfort	129880.0	3.441361	1.319289	0.0	2.00	4.0	5.00	5.0
Inflight_entertainment	129880.0	3.358077	1.334049	0.0	2.00	4.0	4.00	5.0
On-board_service	129880.0	3.383023	1.287099	0.0	2.00	4.0	4.00	5.0
Leg_room_service	129880.0	3.350878	1.316252	0.0	2.00	4.0	4.00	5.0
Baggage_handling	129880.0	3.632114	1.180025	1.0	3.00	4.0	5.00	5.0
Checkin_service	129880.0	3.306267	1.266185	0.0	3.00	3.0	4.00	5.0
Inflight_service	129880.0	3.642193	1.176669	0.0	3.00	4.0	5.00	5.0
Cleanliness	129880.0	3.286326	1.313682	0.0	2.00	3.0	4.00	5.0
Departure_Delay_in_Minutes	129880.0	14.713713	38.071126	0.0	0.00	0.0	12.00	1592.0
Arrival_Delay_in_Minutes	129487.0	15.091129	38.465650	0.0	0.00	0.0	13.00	1584.0

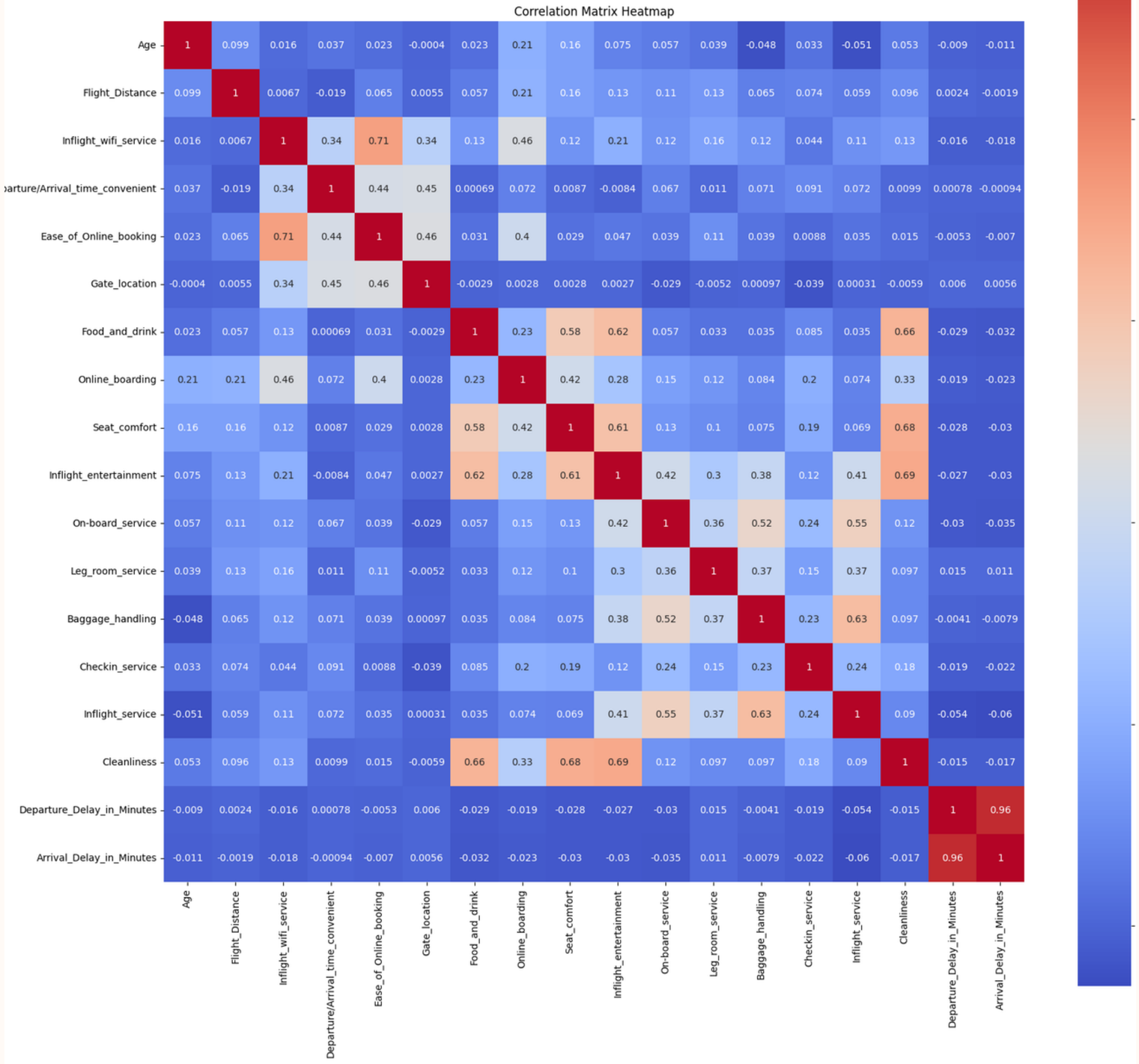
4). Exploratory Data Analysis (EDA)

Statistic Table

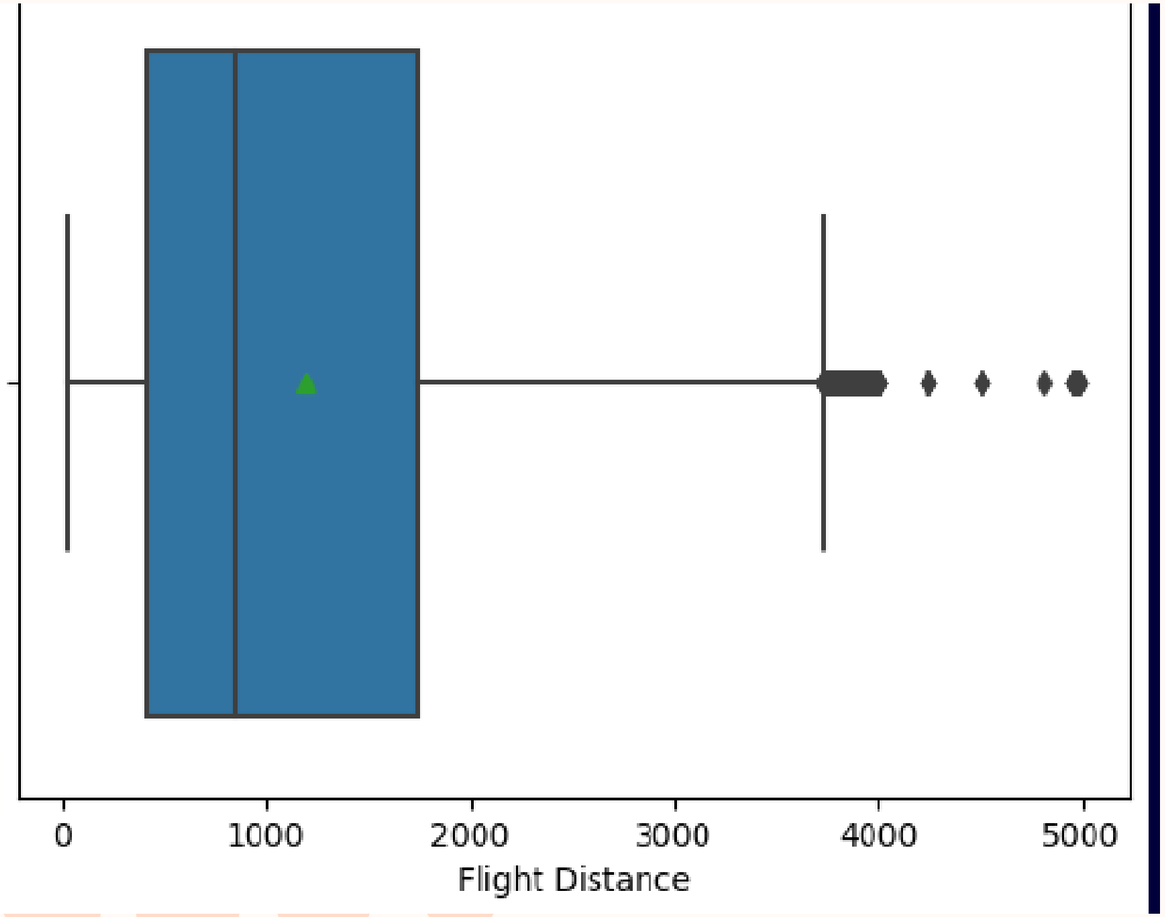
	id	Age	Flight_Distance	Inflight_wifi_service	Departure/Arrival_time_convenient	Ease_of_Online_booking	Gate_location	Food_and_drink
count	129570.000000	129570.000000	129570.000000	129570.000000	129570.000000	129570.000000	129570.000000	129570.000000
mean	64955.104754	39.428679	1190.219989	2.728749	3.057421	2.756942	2.977039	3.204777
std	37490.942449	15.117899	997.573150	1.329373	1.526871	1.401750	1.278603	1.329829
min	1.000000	7.000000	31.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	32488.250000	27.000000	414.000000	2.000000	2.000000	2.000000	2.000000	2.000000
50%	64966.500000	40.000000	844.000000	3.000000	3.000000	3.000000	3.000000	3.000000
75%	97411.750000	51.000000	1744.000000	4.000000	4.000000	4.000000	4.000000	4.000000
max	129880.000000	85.000000	4983.000000	5.000000	5.000000	5.000000	5.000000	5.000000

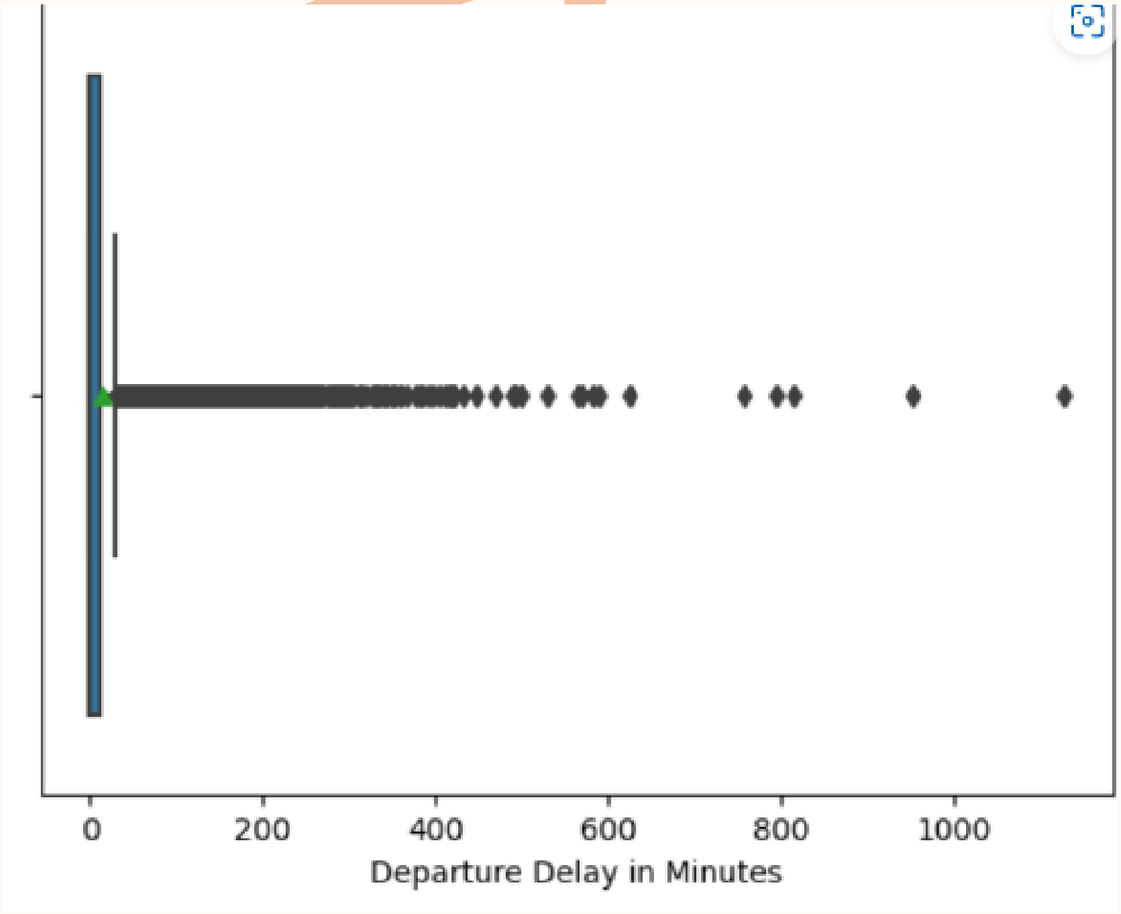
Baggage_handling	Checkin_service	Inflight_service	Cleanliness	Departure_Delay_in_Minutes	Online_boarding	Seat_comfort	Inflight_entertainment	On-board_service	Leg_room_service
129570.000000	129570.000000	129570.000000	129570.000000	129570.000000	129570.000000	129570.000000	129570.000000	129570.000000	129570.000000
3.631998	3.306298	3.642464	3.286363	14.659358	3.252736	3.441661	3.358223	3.383221	3.351154
1.180141	1.266187	1.176623	1.313623	37.978973	1.350661	1.319136	1.334083	1.287040	1.316097
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3.000000	3.000000	3.000000	2.000000	0.000000	2.000000	2.000000	2.000000	2.000000	2.000000
4.000000	3.000000	4.000000	3.000000	0.000000	3.000000	4.000000	4.000000	4.000000	4.000000
5.000000	4.000000	5.000000	4.000000	12.000000	4.000000	5.000000	4.000000	4.000000	4.000000
5.000000	5.000000	5.000000	5.000000	1592.000000	5.000000	5.000000	5.000000	5.000000	5.000000

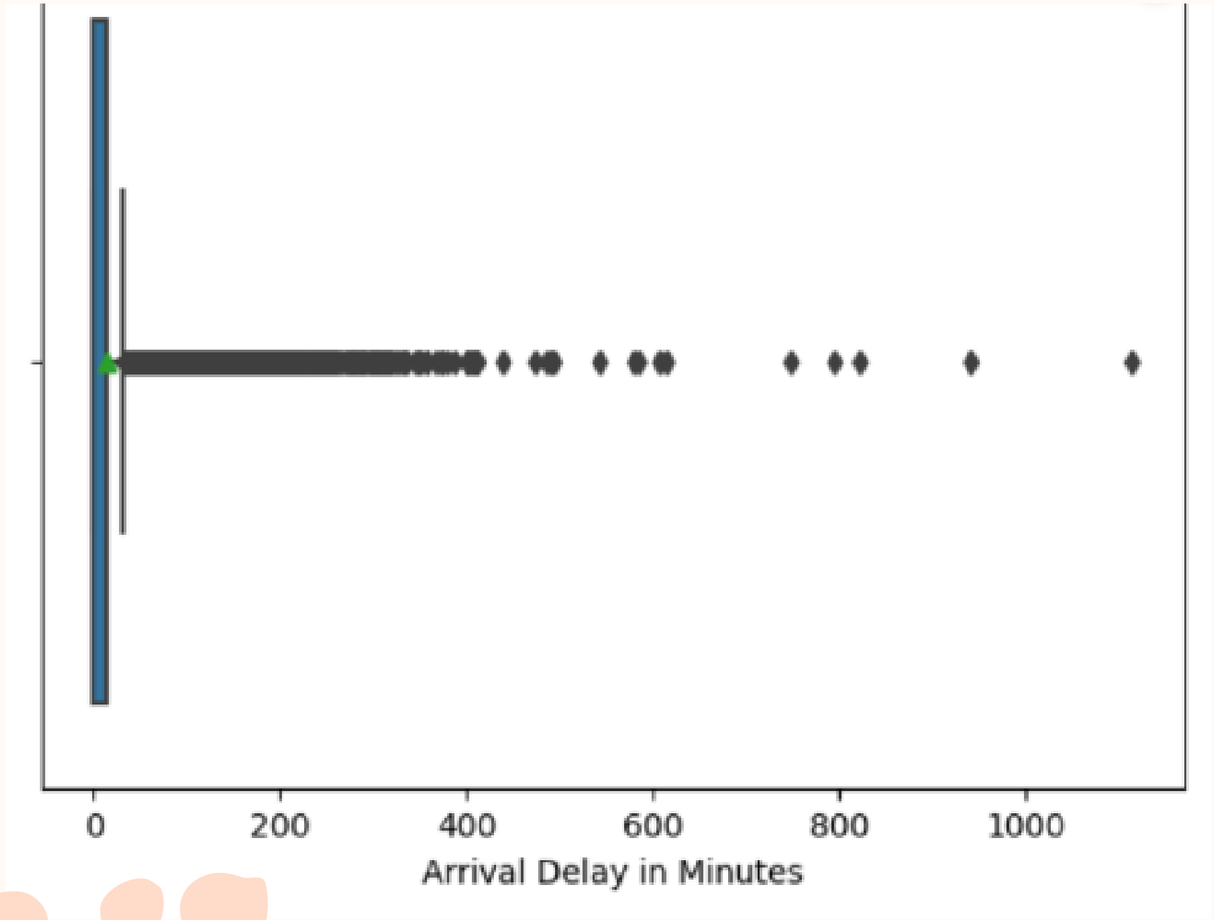
Heat map


Correlation Matrix Heatmap
Age 1 0.099 0.016 0.037 0.023 -0.0004 0.023 0.21 0.16 0.075 0.057 0.039 -0.048 0.033 -0.051 0.053 -0.009 -0.011
Flight_Distance 0.099 1 0.0067 -0.019 0.065 0.0055 0.057 0.21 0.16 0.13 0.11 0.13 0.065 0.074 0.059 0.096 0.0024 -0.0019
Inflight_wifi_service 0.016 0.0067 1 0.34 0.71 0.34 0.13 0.46 0.12 0.21 0.12 0.16 0.12 0.044 0.11 0.13 -0.016 -0.018
arture/Arrival_time_convenient 0.037 -0.019 0.34 1 0.44 0.45 0.00069 0.072 0.0087 -0.0084 0.067 0.011 0.071 0.091 0.072 0.0099 0.00078 -0.00094
Ease_of_Online_booking 0.023 0.065 0.71 0.44 1 0.46 0.031 0.4 0.029 0.047 0.039 0.11 0.039 0.0088 0.035 0.015 -0.0053 -0.007
Gate_location -0.0004 0.0055 0.34 0.45 0.46 1 -0.0029 0.0028 0.0028 0.0027 -0.029 -0.0052 0.00097 -0.039 0.00031 -0.0059 0.006 0.0056
Food_and_drink 0.023 0.057 0.13 0.00069 0.031 -0.0029 1 0.23 0.58 0.62 0.057 0.033 0.035 0.085 0.035 0.66 -0.029 -0.032
Online_boarding 0.21 0.21 0.46 0.072 0.4 0.0028 0.23 1 0.42 0.28 0.15 0.12 0.084 0.2 0.074 0.33 -0.019 -0.023
Seat_comfort 0.16 0.16 0.12 0.0087 0.029 0.0028 0.58 0.42 1 0.61 0.13 0.1 0.075 0.19 0.069 0.68 -0.028 -0.03
Inflight_entertainment 0.075 0.13 0.21 -0.0084 0.047 0.0027 0.62 0.28 0.61 1 0.42 0.3 0.38 0.12 0.41 0.69 -0.027 -0.03
On-board_service 0.057 0.11 0.12 0.067 0.039 -0.029 0.057 0.15 0.13 0.42 1 0.36 0.52 0.24 0.55 0.12 -0.03 -0.035
Leg_room_service 0.039 0.13 0.16 0.011 0.11 -0.0052 0.033 0.12 0.1 0.3 0.36 1 0.37 0.15 0.37 0.097 0.015 0.011
Baggage_handling -0.048 0.065 0.12 0.071 0.039 0.00097 0.035 0.084 0.075 0.38 0.52 0.37 1 0.23 0.63 0.097 -0.0041 -0.0079
Checkin_service 0.033 0.074 0.044 0.091 0.0088 -0.039 0.085 0.2 0.19 0.12 0.24 0.15 0.23 1 0.24 0.18 -0.019 -0.022
Inflight_service -0.051 0.059 0.11 0.072 0.035 0.00031 0.035 0.074 0.069 0.41 0.55 0.37 0.63 0.24 1 0.09 -0.054 -0.06
Cleanliness 0.053 0.096 0.13 0.0099 0.015 -0.0059 0.66 0.33 0.68 0.69 0.12 0.097 0.097 0.18 0.09 1 -0.015 -0.017
Departure_Delay_in_Minutes -0.009 0.0024 -0.016 0.00078 -0.0053 0.006 -0.029 -0.019 -0.028 -0.027 -0.03 0.015 -0.0041 -0.019 -0.054 -0.015 1 0.96
Arrival_Delay_in_Minutes -0.011 -0.0019 -0.018 -0.00094 -0.007 0.0056 -0.032 -0.023 -0.03 -0.03 -0.035 0.011 -0.0079 -0.022 -0.06 -0.017 0.96 1
Age
Flight_Distance
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
Departure_Delay_in_Minutes
Arrival_Delay_in_Minutes

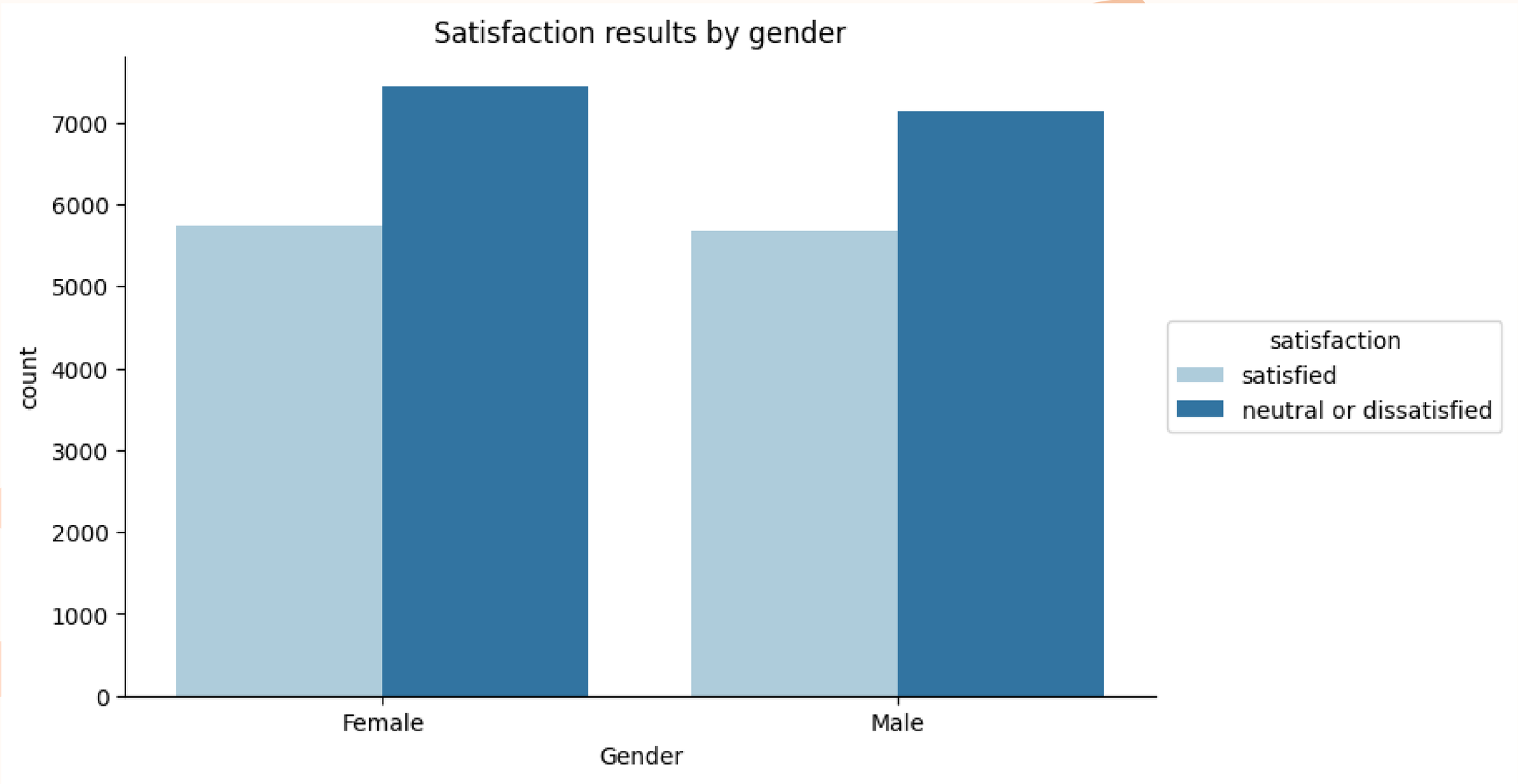
Box plot

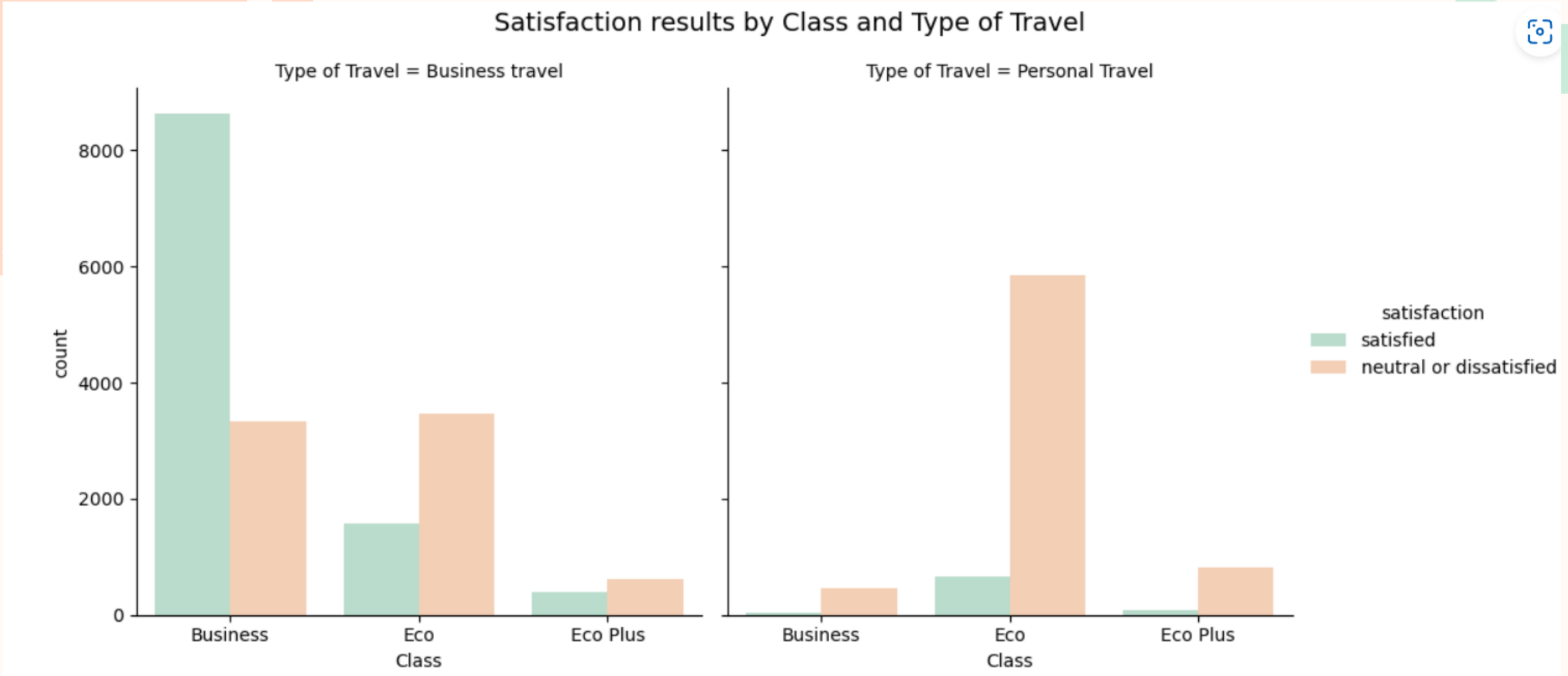

0
1000
2000
3000
4000
5000
Flight Distance

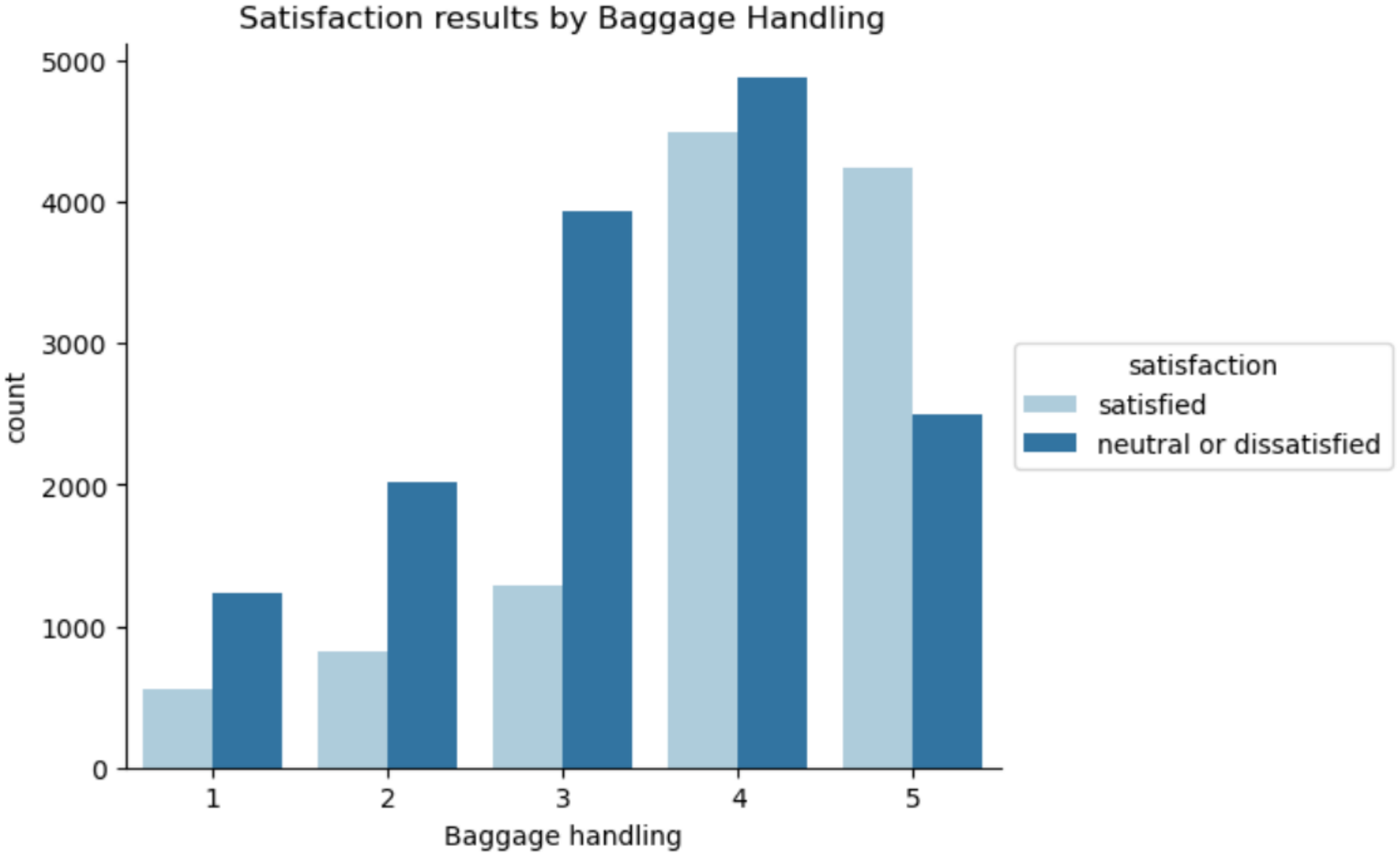

0
200
400
600
800
1000
Departure Delay in Minutes

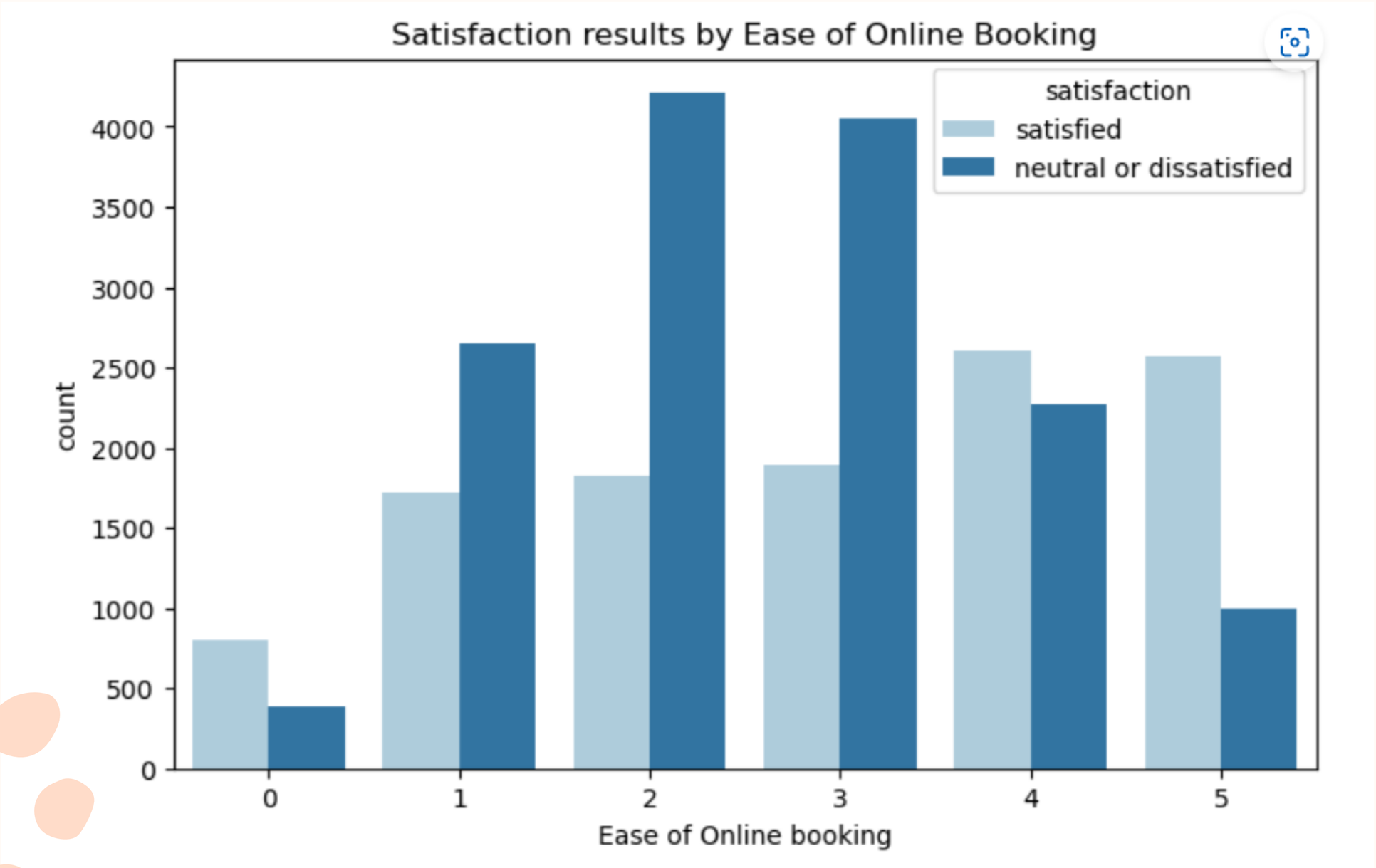

0
200
400
600
800
1000
Arrival Delay in Minutes

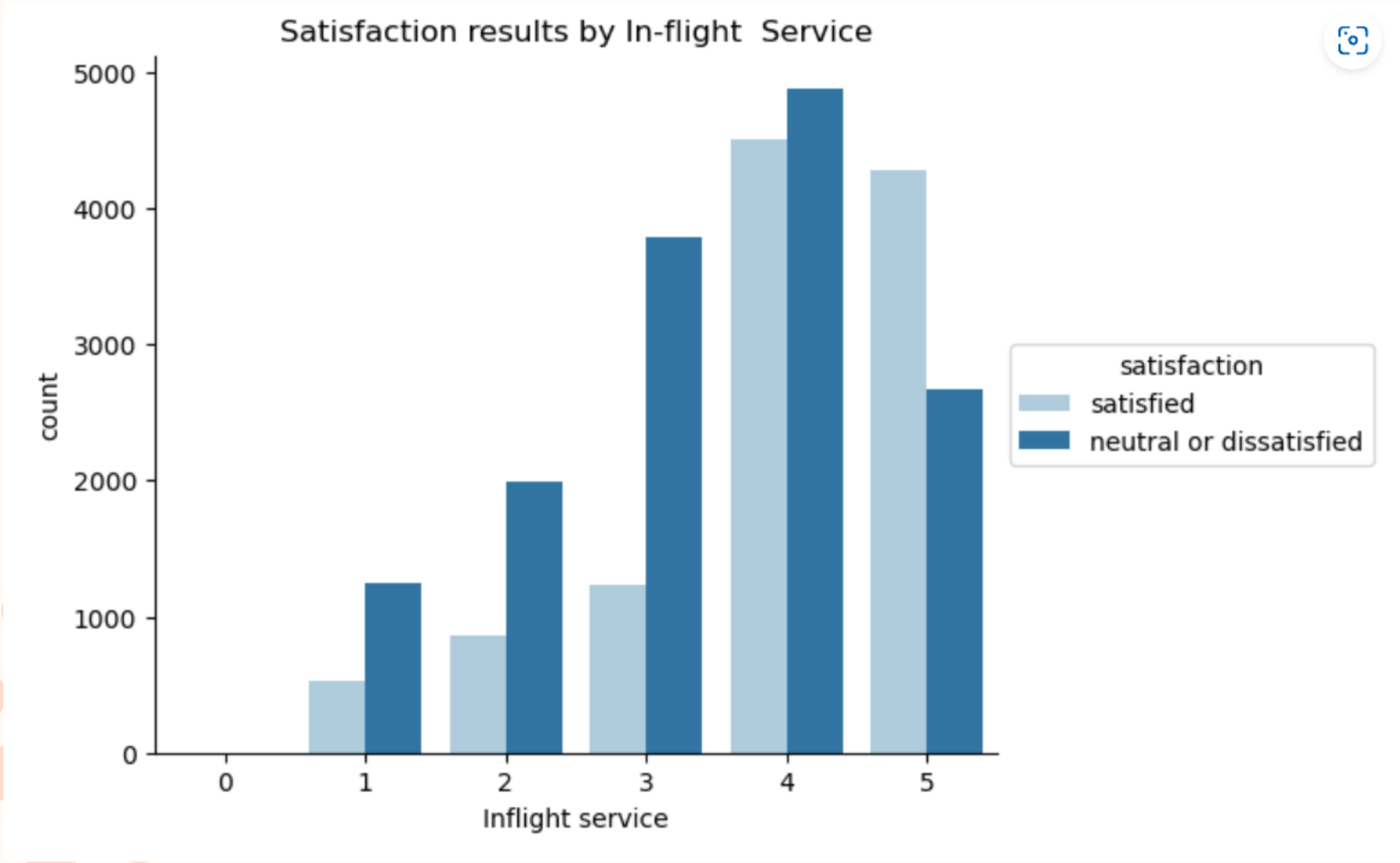
Bar graph

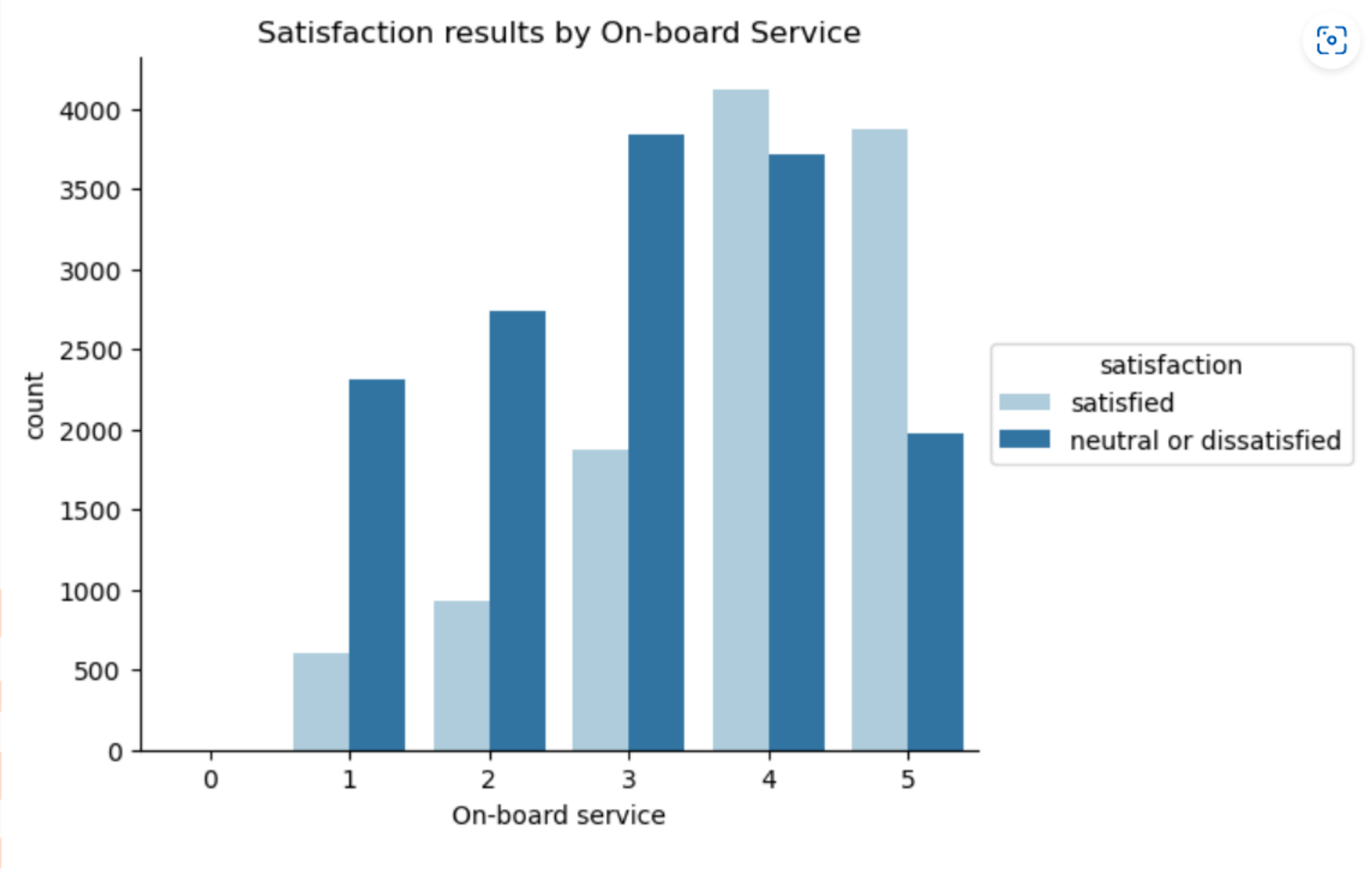

Satisfaction results by gender
count
7000
6000
5000
4000
3000
2000
1000
0
Female
Male
Gender
satisfaction
satisfied
neutral or dissatisfied

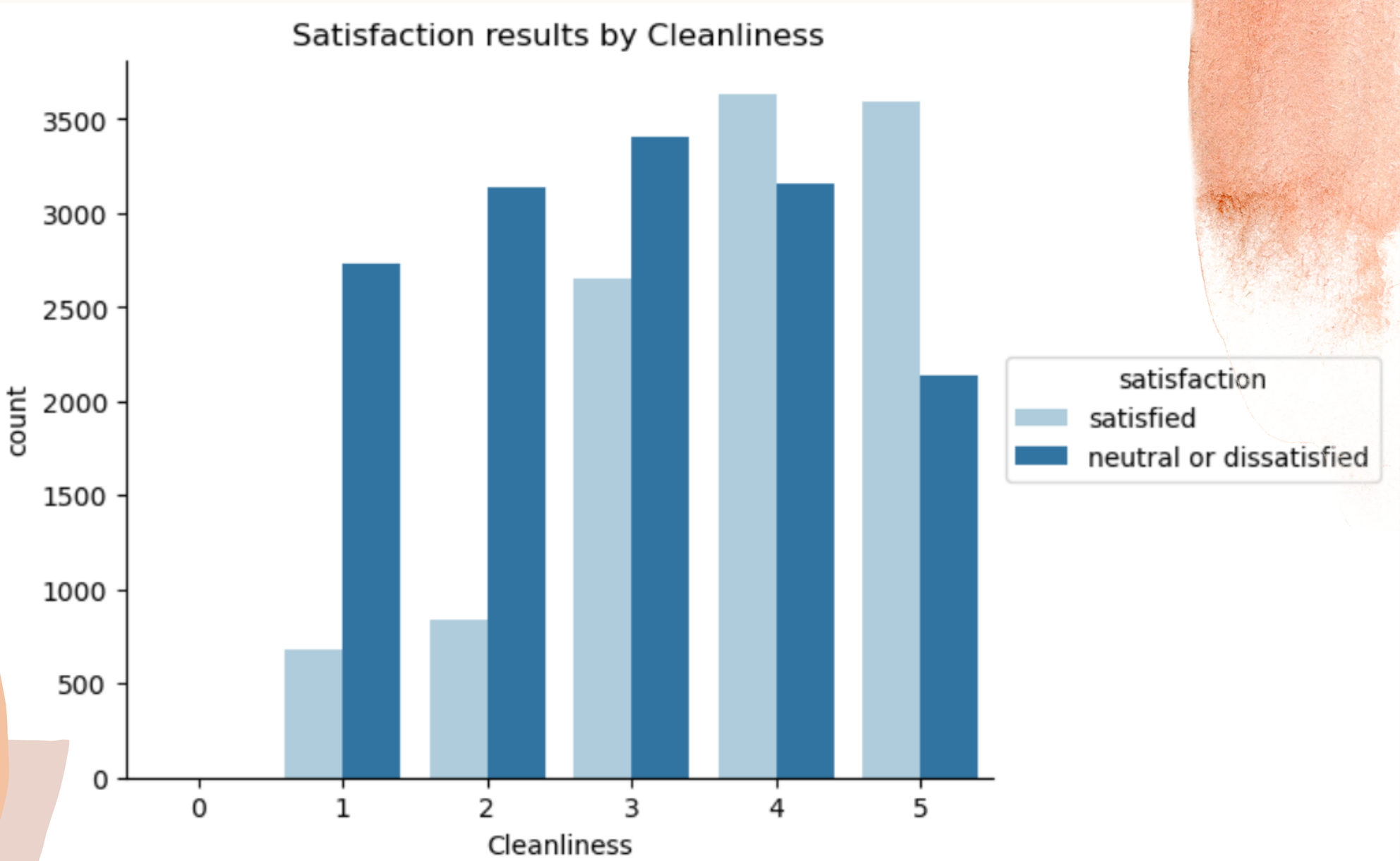

Satisfaction results by Class and Type of Travel
Type of Travel = Business travel
Type of Travel = Personal Travel
count
8000
6000
4000
2000
0
Business
Eco
Eco Plus
Class
satisfaction
satisfied
neutral or dissatisfied

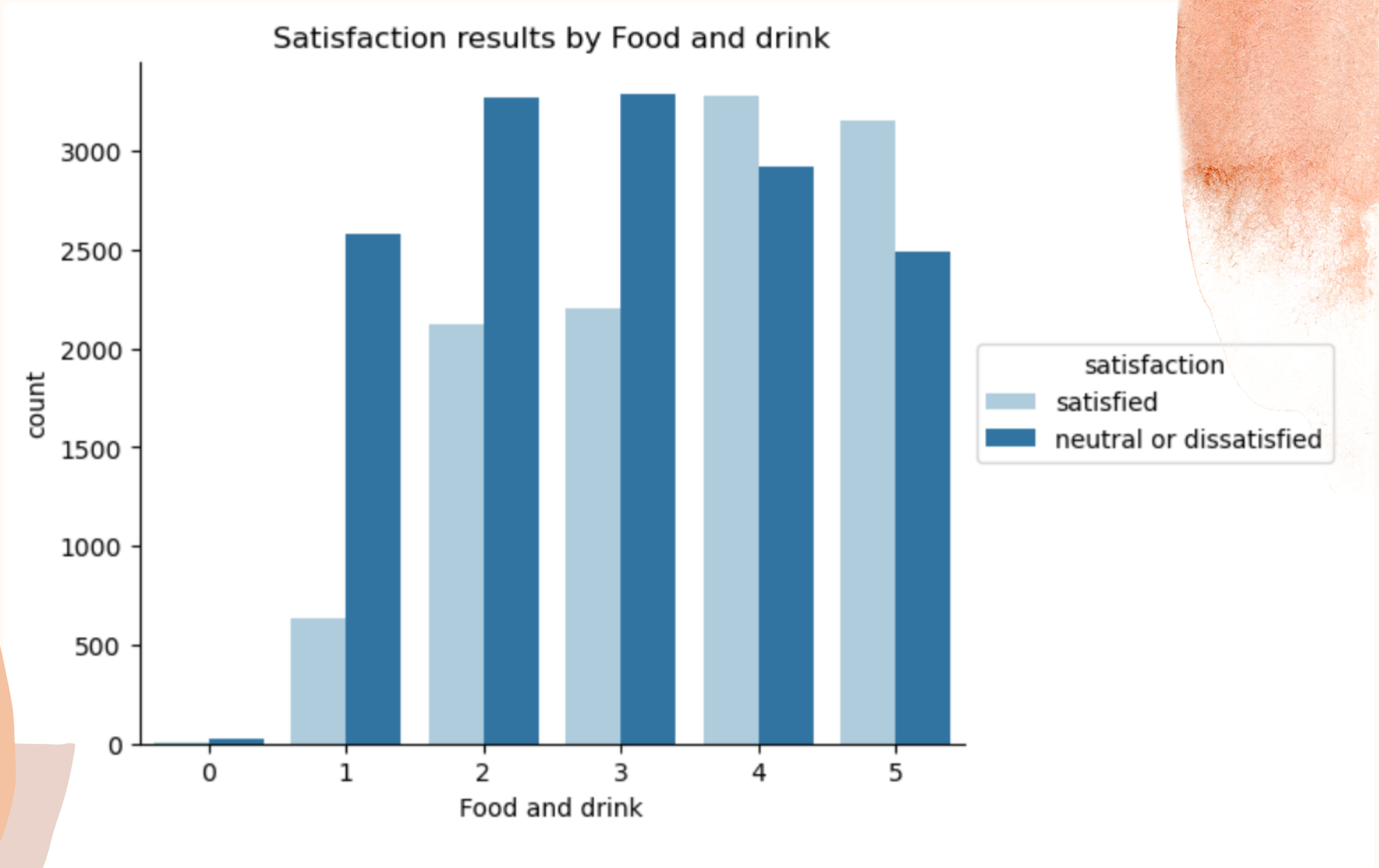

Satisfaction results by Baggage Handling
5000
4000
3000
2000
1000
0
count
1
2
3
4
5
Baggage handling
satisfaction
satisfied
neutral or dissatisfied


Satisfaction results by Ease of Online Booking
satisfaction
satisfied
neutral or dissatisfied
count
0
500
1000
1500
2000
2500
3000
3500
4000
0
1
2
3
4
5
Ease of Online booking

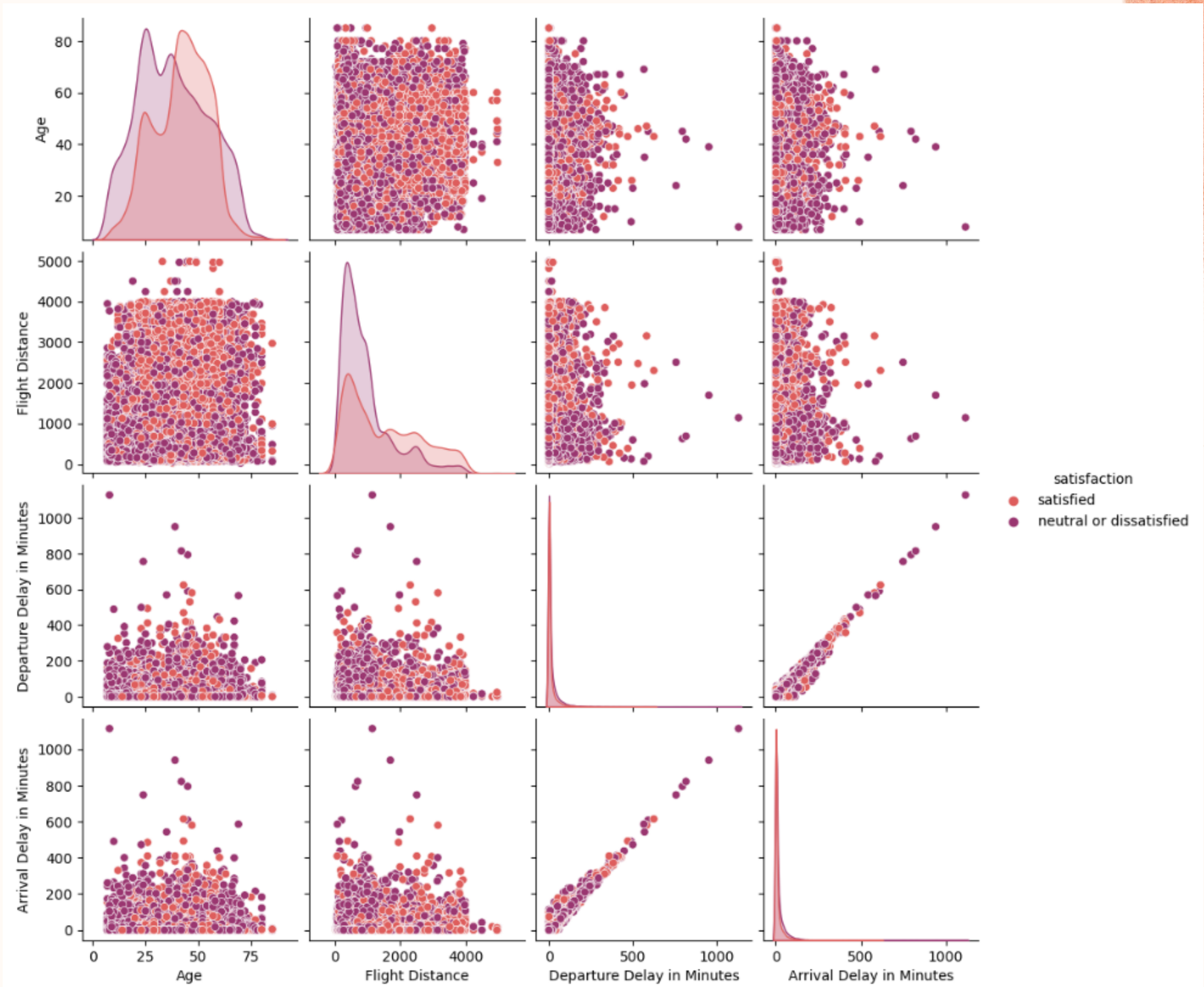

Satisfaction results by In-flight Service
5000
4000
3000
2000
1000
0
count
0
1
2
3
4
5
Inflight service
satisfaction
satisfied
neutral or dissatisfied


Satisfaction results by On-board Service
count
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
On-board service
satisfaction
satisfied
neutral or dissatisfied

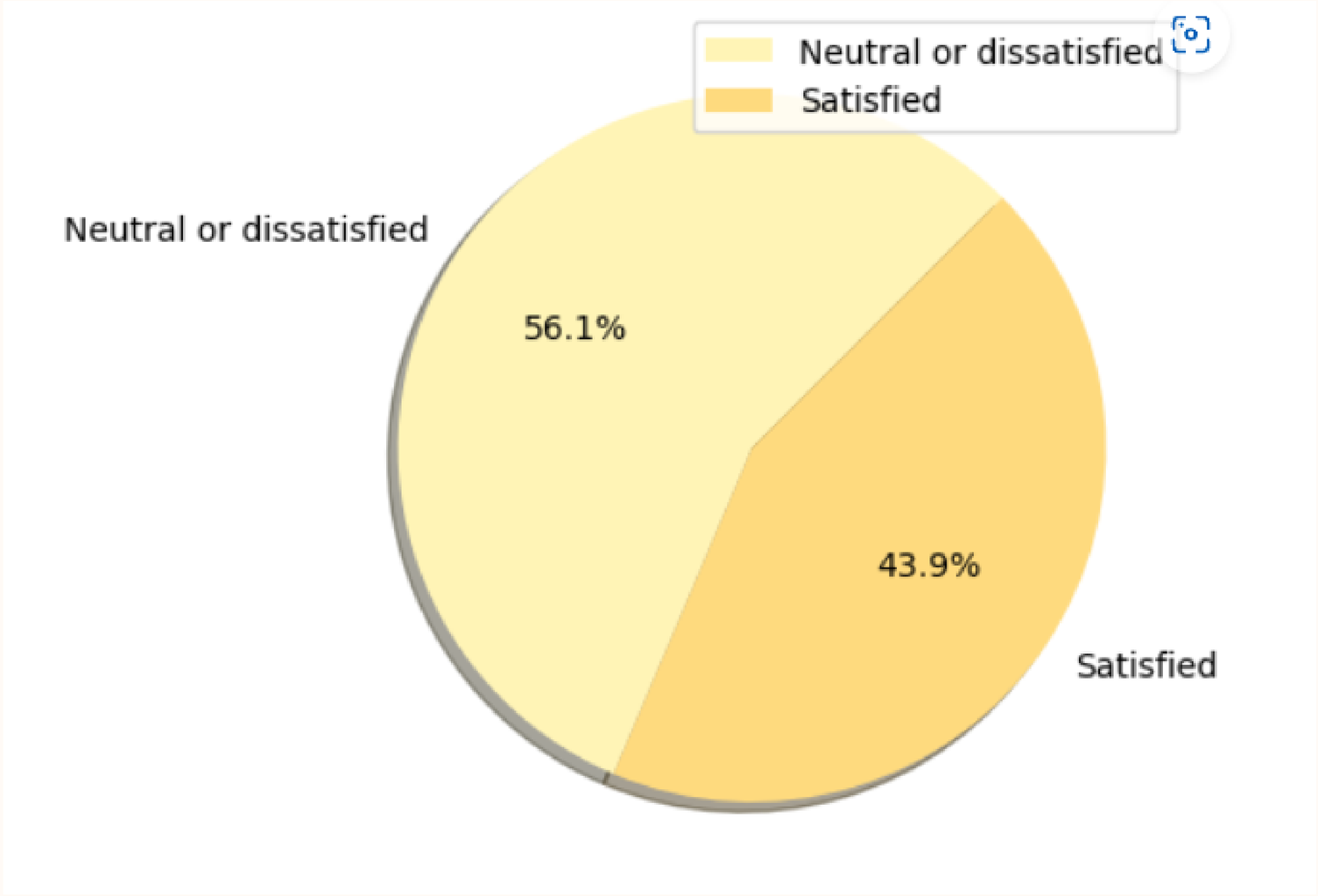

Satisfaction results by Cleanliness
count
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
Cleanliness
satisfaction
satisfied
neutral or dissatisfied

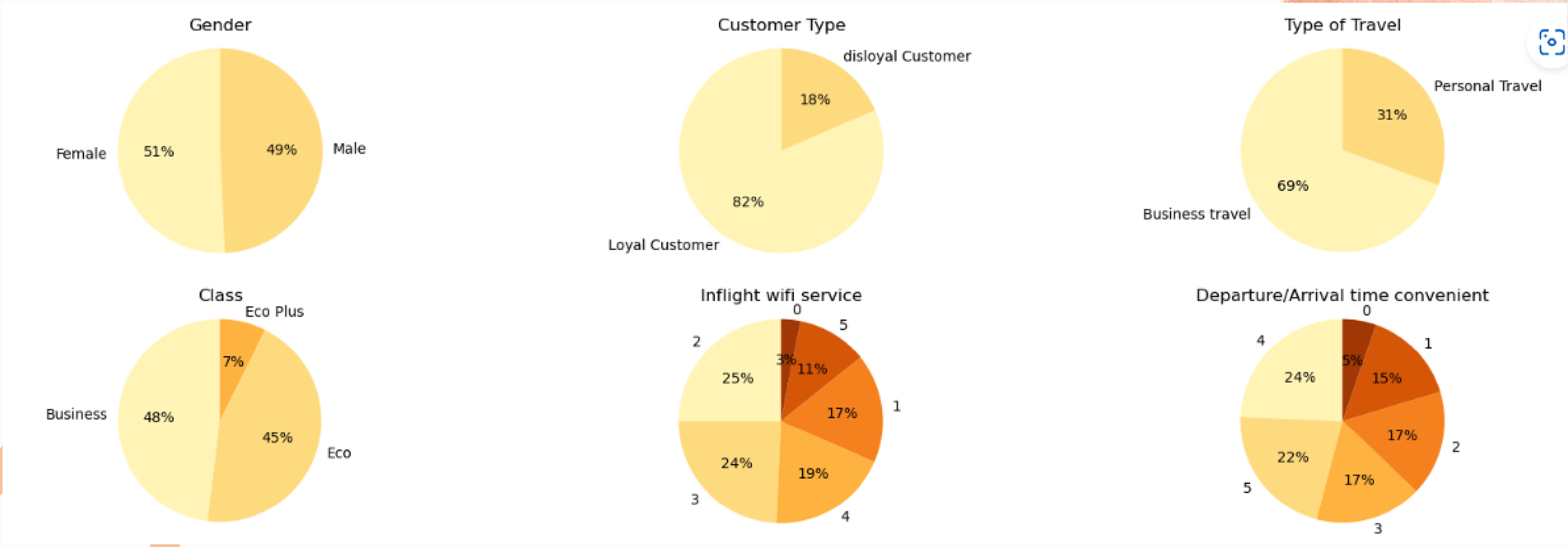

Satisfaction results by Food and drink
count
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
Food and drink
satisfaction
satisfied
neutral or dissatisfied

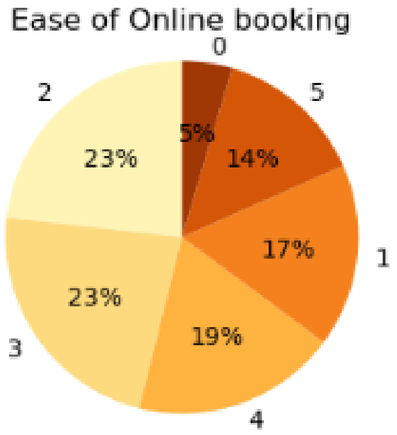
Pair pot

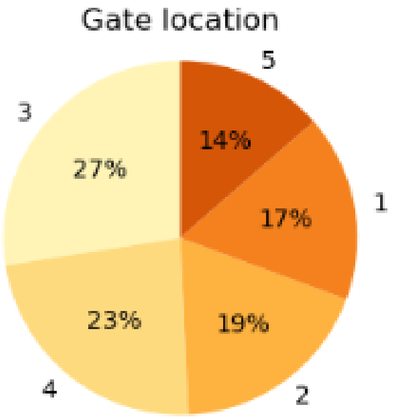

Age
Flight Distance
Departure Delay in Minutes
Arrival Delay in Minutes
satisfaction
satisfied
neutral or dissatisfied

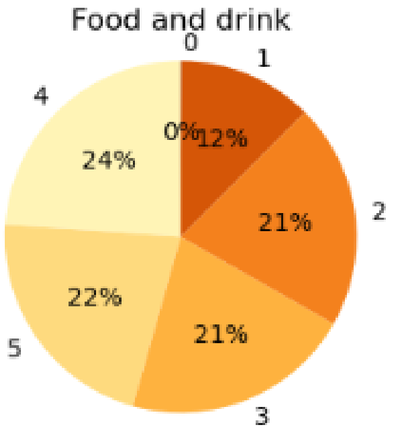
Pie chat

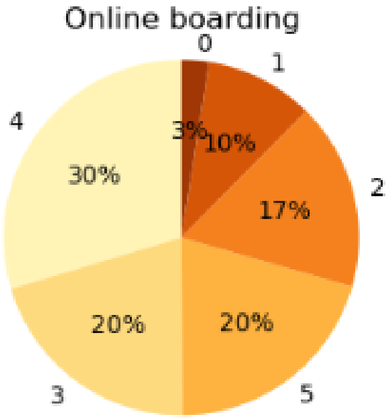

Neutral or dissatisfied
Satisfied
Neutral or dissatisfied
56.1%
43.9%
Satisfied

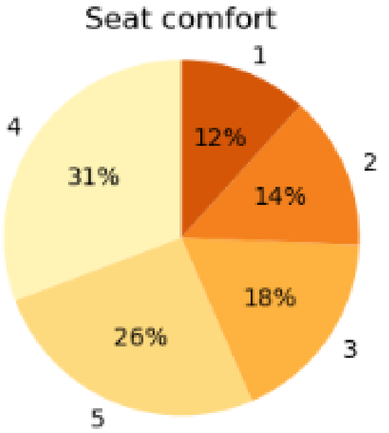

Gender
Female
51%
49%
Male
Customer Type
disloyal Customer
18%
82%
Loyal Customer
Type of Travel
Personal Travel
31%
69%
Business travel
Class
Eco Plus
7%
Business
48%
45%
Eco
Inflight wifi service
0
5
2
3%
11%
25%
17%
1
24%
19%
3
4
Departure/Arrival time convenient
0
1
4
5%
15%
24%
17%
2
22%
17%
5
3

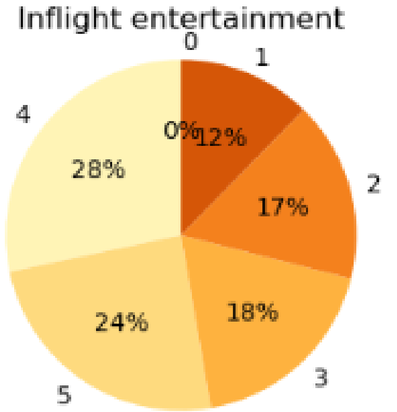

Ease of Online booking
0
5
1
2
3
4
5%
14%
17%
23%
19%
23%

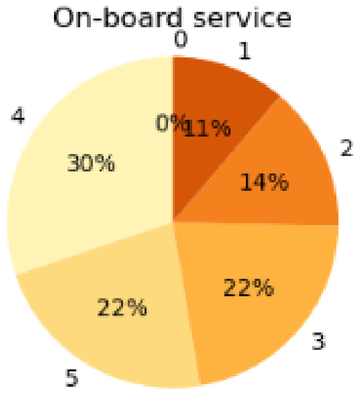

Gate location
5
1
2
3
4
14%
17%
19%
27%
23%

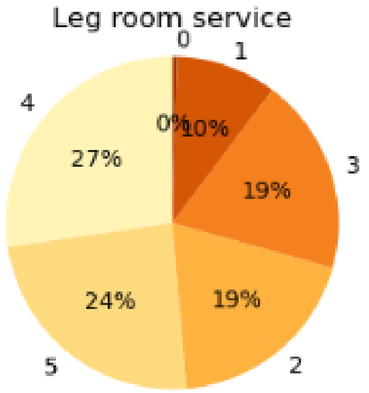

Food and drink
0
1
2
3
4
5
0%
12%
21%
24%
21%
22%

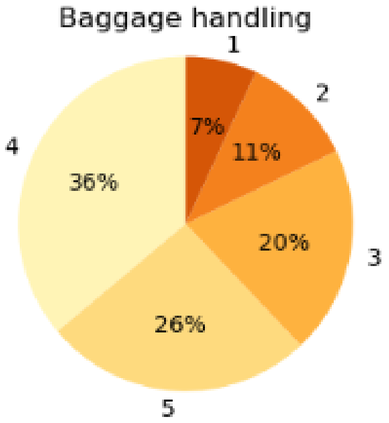

Online boarding
0
1
2
3
4
5
3%
10%
17%
30%
20%
20%

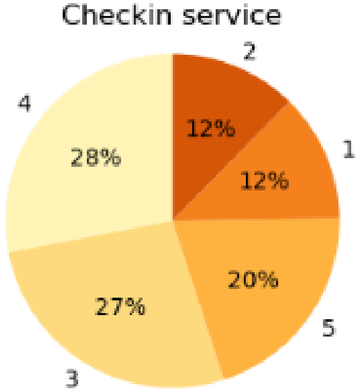

Seat comfort
1
2
3
4
5
12%
14%
18%
31%
26%

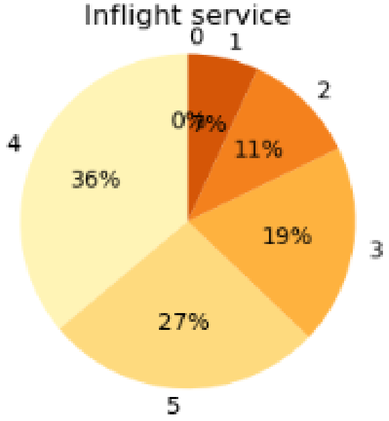

Inflight entertainment
0
1
2
3
4
5
0%
12%
17%
28%
24%
18%

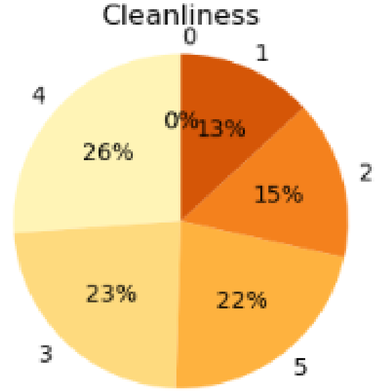

On-board service
0
1
2
3
4
5
0%
11%
14%
22%
22%
30%


Leg room service
0
1
2
3
4
5
0%
10%
19%
19%
24%
27%


Baggage handling
1
2
3
4
5
7%
11%
20%
26%
36%


Checkin service
2
1
5
3
4
12%
12%
20%
27%
28%


Inflight service
0
1
2
3
4
5
0%
7%
11%
19%
27%
36%


Cleanliness
0
1
2
3
5
0%
13%
15%
22%
23%
26%

5). Feature Engineering

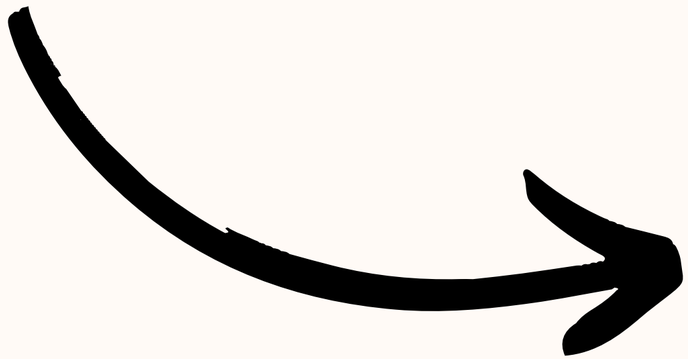
5.1). Encoding Categorical Variables

In classification problems, the goal is to predict the category of a new observation based on a set of training data. However, machine learning algorithms cannot directly work with categorical variables. They can only work with numerical data. Therefore, it is necessary to encode categorical variables into numerical values before they can be used in a classification model.

```python
from sklearn.preprocessing import LabelEncoder

def encode_dataframe(df):
    encoding_mapping = {}

    # Iterate over each column
    for column in df.columns:
        # Check if the column dtype is object (i.e., categorical)
        if df[column].dtype == 'object':
            # Create a LabelEncoder object
            label_encoder = LabelEncoder()
            # Fit and transform the column to obtain encoded values
            encoded_values = label_encoder.fit_transform(df[column])
            # Create a dictionary mapping original values to encoded values
            encoding_mapping[column] = dict(zip(df[column], encoded_values))
            # Replace the column values with encoded values
            df[column] = encoded_values

    return df, encoding_mapping
```

```python
encoded_df, encoding_mapping = encode_dataframe(df)

encoded_df
```

Gender	Customer_Type	Age	Type_of_Travel	Class	satisfaction
Male	Loyal Customer	13	Personal Travel	Eco Plus	neutral or dissatisfied
Male	disloyal Customer	25	Business travel	Business	neutral or dissatisfied
Female	Loyal Customer	26	Business travel	Business	satisfied
Female	Loyal Customer	25	Business travel	Business	neutral or dissatisfied
Male	Loyal Customer	61	Business travel	Business	satisfied



Gender	Customer_Type	Age	Type_of_Travel	Class	satisfaction
1	0	13	1	2	0
1	1	25	0	0	0
0	0	26	0	0	1
0	0	25	0	0	0
1	0	61	0	0	1
...	...	...	...	...	...
1	1	34	0	0	0
1	0	23	0	0	1
0	0	17	1	1	0
1	0	14	0	0	1
0	0	42	1	1	0

2.2). Feature Selection and Feature Importance

Setting our categorical & target variable

```
categorical_features = X[[
        'Gender', 'Customer_Type','Type_of_Travel','Class', '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']].copy()
y_target_cat = y
```

Feature Selection by Chi-Sqaure Test

```
selector = SelectKBest(chi2, k=10)
X_10_best = selector.fit_transform(categorical_features, y)

# Get the mask of selected features
selected_features_mask = selector.get_support()

# Get the names of selected features
chi_10_selected_features = categorical_features.columns[selected_features_mask]

print("The 10 most important features selected by chi-square test:")
print(chi_10_selected_features)
```

```
The 10 most important features selected by chi-square test:
Index(['Customer_Type', 'Type_of_Travel', 'Class', 'Inflight_wifi_service',
       'Online_boarding', 'Seat_comfort', 'Inflight_entertainment',
       'On-board_service', 'Leg_room_service', 'Cleanliness'],
      dtype='object')
```

```
Online_boarding                    0.163768
Inflight_wifi_service              0.145169
Type_of_Travel                     0.099629
Class                              0.097405
Inflight_entertainment             0.053983
Seat_comfort                       0.047860
Flight_Distance                    0.035487
Customer_Type                      0.035108
Ease_of_Online_booking             0.034178
On-board_service                   0.033440
Leg_room_service                   0.033235
Age                                0.030020
id                                 0.029826
Cleanliness                        0.026593
Checkin_service                    0.023300
Inflight_service                   0.022528
Baggage_handling                   0.021653
Departure/Arrival_time_convenient  0.016849
Gate_location                      0.015630
Arrival_Delay_in_Minutes           0.010873
Food_and_drink                     0.010354
Departure_Delay_in_Minutes         0.009383
Gender                             0.003728
dtype: float64
```

```python
from sklearn.feature_selection import SelectFromModel

selector = SelectFromModel(rf(n_estimators=100, random_state=0))
selected_features = selector.fit_transform(X, y)
support = selector.get_support()
wrapper_features = X.loc[:, support].columns.tolist()
print(wrapper_features)

model = rf(n_estimators=100, random_state=0)
model.fit(X, y)
feature_importances = pd.Series(model.feature_importances_, index=X.columns)
print(feature_importances)
```

Feature Importance by Wrapper Method (random forest)

Feature Importance by Permutation Method

```
import warnings
warnings.filterwarnings("ignore")
import eli5
from eli5.sklearn import PermutationImportance

perm = PermutationImportance(rf(n_estimators=100, random_state=0).fit(X,y),random_state=1).fit(X,y)
eli5.show_weights(perm, feature_names = X.columns.tolist())
```

Weight	Feature
0.1466 ± 0.0008	Inflight_wifi_service
0.1350 ± 0.0016	Type_of_Travel
0.0532 ± 0.0006	Customer_Type
0.0411 ± 0.0011	Online_boarding
0.0339 ± 0.0005	Class
0.0259 ± 0.0004	Checkin_service
0.0186 ± 0.0005	Seat_comfort
0.0179 ± 0.0003	Baggage_handling
0.0158 ± 0.0002	Inflight_service
0.0151 ± 0.0006	Cleanliness
0.0146 ± 0.0007	id
0.0108 ± 0.0004	Age
0.0101 ± 0.0005	On-board_service
0.0084 ± 0.0003	Leg_room_service
0.0082 ± 0.0003	Flight_Distance
0.0074 ± 0.0003	Inflight_entertainment
0.0073 ± 0.0001	Arrival_Delay_in_Minutes
0.0060 ± 0.0002	Ease_of_Online_booking
0.0043 ± 0.0003	Gate_location
0.0040 ± 0.0002	Departure_Delay_in_Minutes
… 3 more …	

Now, let's analyze the result:

1. Inflight_wifi_service: This feature has the highest weight (0.1466) and is considered the most important feature according to Permutation Importance. It means that if you randomly permute the values of this feature, the model's performance would decrease significantly.

2. Type_of_Travel: This feature has the second-highest weight (0.1350). It is another important feature for the model's predictions.

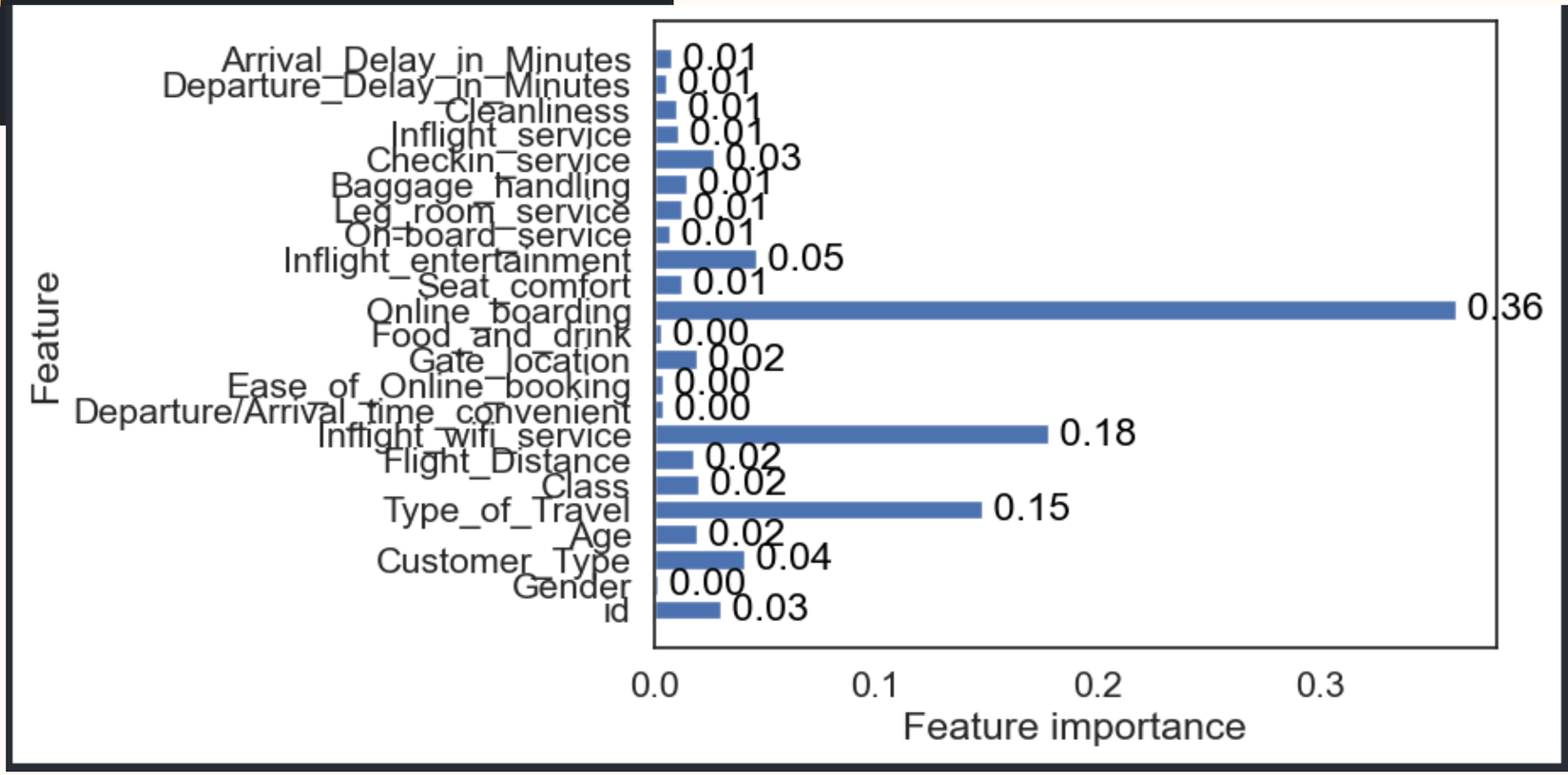
3. Customer_Type: This feature has a weight of 0.0532, indicating its importance but to a lesser extent than the previous two features.

4. Online_boarding: This feature has a weight of 0.0411, suggesting it has some impact on the model's predictions.

5. Class: This feature has a weight of 0.0339, indicating its importance but to a lesser extent than the previous features.

```python
def plot_feature_importances_cancer(model, feature_names):
    n_features = len(feature_names)
    feature_importances = model.feature_importances_
    plt.barh(range(n_features), feature_importances, align='center')
    plt.yticks(np.arange(n_features), feature_names)
    plt.xlabel("Feature importance")
    plt.ylabel("Feature")
    for i, v in enumerate(feature_importances):
        plt.text(v, i, f" {v:.2f}", color='black', va='center')
# Assuming you have the following variables: tree, X_train_tree, y_train_tree, and df
tree = DecisionTreeClassifier(random_state=1)
tree.fit(X_train_tree, y_train_tree)
# Call the plot_feature_importances_cancer function
plot_feature_importances_cancer(tree, X.columns)
plt.show()
```

Feature Importance by Decision Tree



```python
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE

# Scale the data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Create a logistic regression model
logreg = LogisticRegression(max_iter=1000)

# Apply Recursive Feature Elimination (RFE)
rfe = RFE(logreg, n_features_to_select=10)
rfe.fit(X_scaled, y)

selected_features_rfe = X.columns[rfe.support_].tolist()
selected_features_rfe
```

Feature Importance by Recursive Feature Elimination

```
['Customer_Type',
 'Type_of_Travel',
 'Class',
 'Inflight_wifi_service',
 'Ease_of_Online_booking',
 'Online_boarding',
 'On-board_service',
 'Leg_room_service',
 'Checkin_service',
 'Cleanliness']
```

```python
def backward_selection(X, y, k_features=10):
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)

    estimator = LogisticRegression()
    selector = RFE(estimator, n_features_to_select=k_features, step=1)
    selector = selector.fit(X_scaled, y)
    selected_features = X.columns[selector.support_].tolist()
    return selected_features

def forward_selection(X, y, k_features=10):
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)

    estimator = LogisticRegression()
    selector = SequentialFeatureSelector(estimator, n_features_to_select=k_features, direction='forward')
    selector = selector.fit(X_scaled, y)
    selected_features = X.columns[selector.support_].tolist()
    return selected_features

def stepwise_selection(X, y, k_features=10):
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)

    estimator = LogisticRegression()
    selector = SequentialFeatureSelector(estimator, k_features=k_features, forward=True, floating=True, scoring='accuracy', cv=5)
    selector = selector.fit(X_scaled, y)
    selected_features = X.columns[list(selector.k_feature_idx_)].tolist()
    return selected_features
```

✓ 0.0s

forward_features

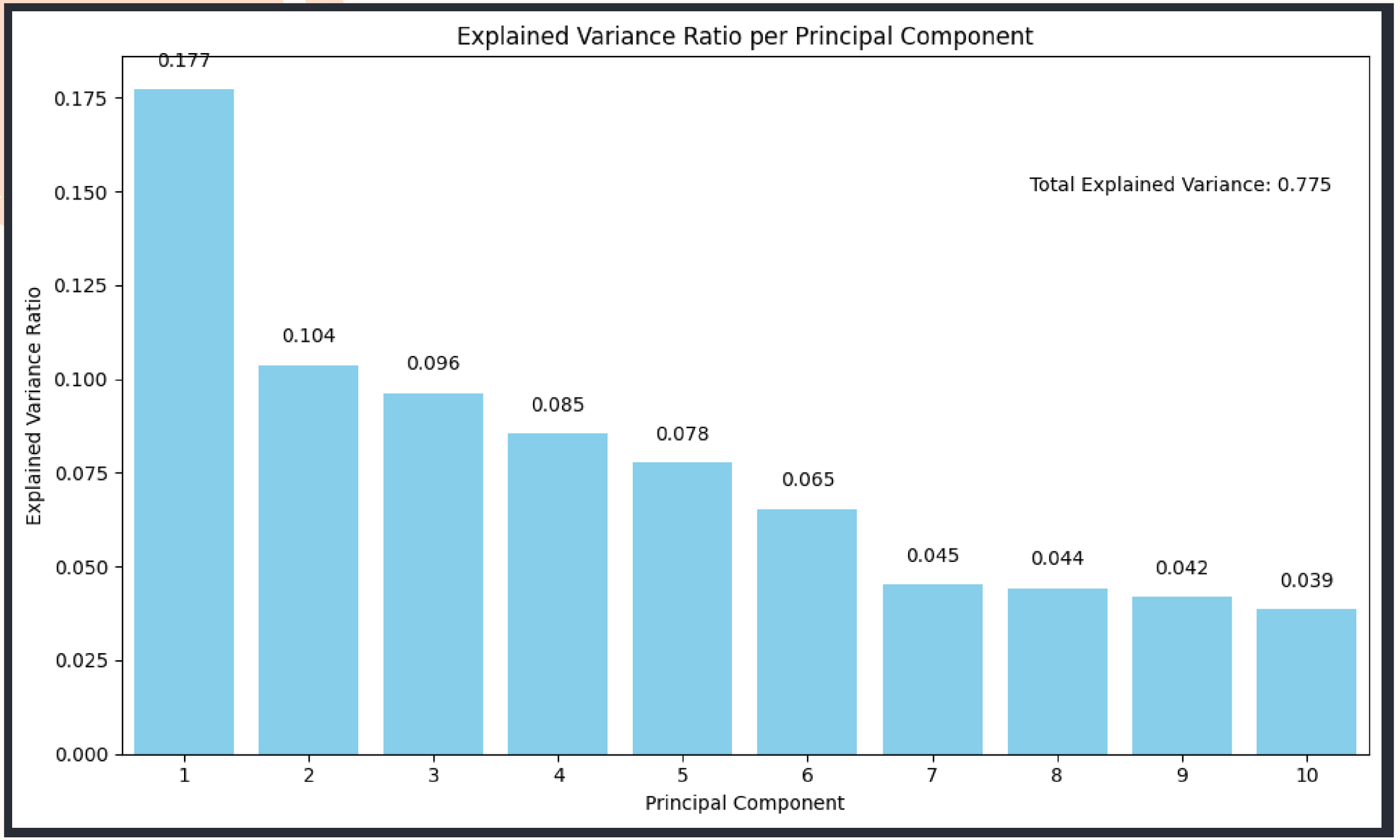
```
['Customer_Type',
 'Type_of_Travel',
 'Inflight_wifi_service',
 'Gate_location',
 'Online_boarding',
 'Inflight_entertainment',
 'On-board_service',
 'Leg_room_service',
 'Checkin_service',
 'Cleanliness']
```

backward_features

✓ 0.0s

```
['Customer_Type',
 'Type_of_Travel',
 'Class',
 'Inflight_wifi_service',
 'Ease_of_Online_booking',
 'Online_boarding',
 'On-board_service',
 'Leg_room_service',
 'Checkin_service',
 'Cleanliness']
```

2.3).Dimentionality Reduction




```
Full_Set.shape
```

```
(129880, 25)
```

```
Projected_data.shape
```

```
(129880, 10)
```

Perform Model Algorithm

6). Model Selection

Model1: Normal Logistic Regression
(all features)

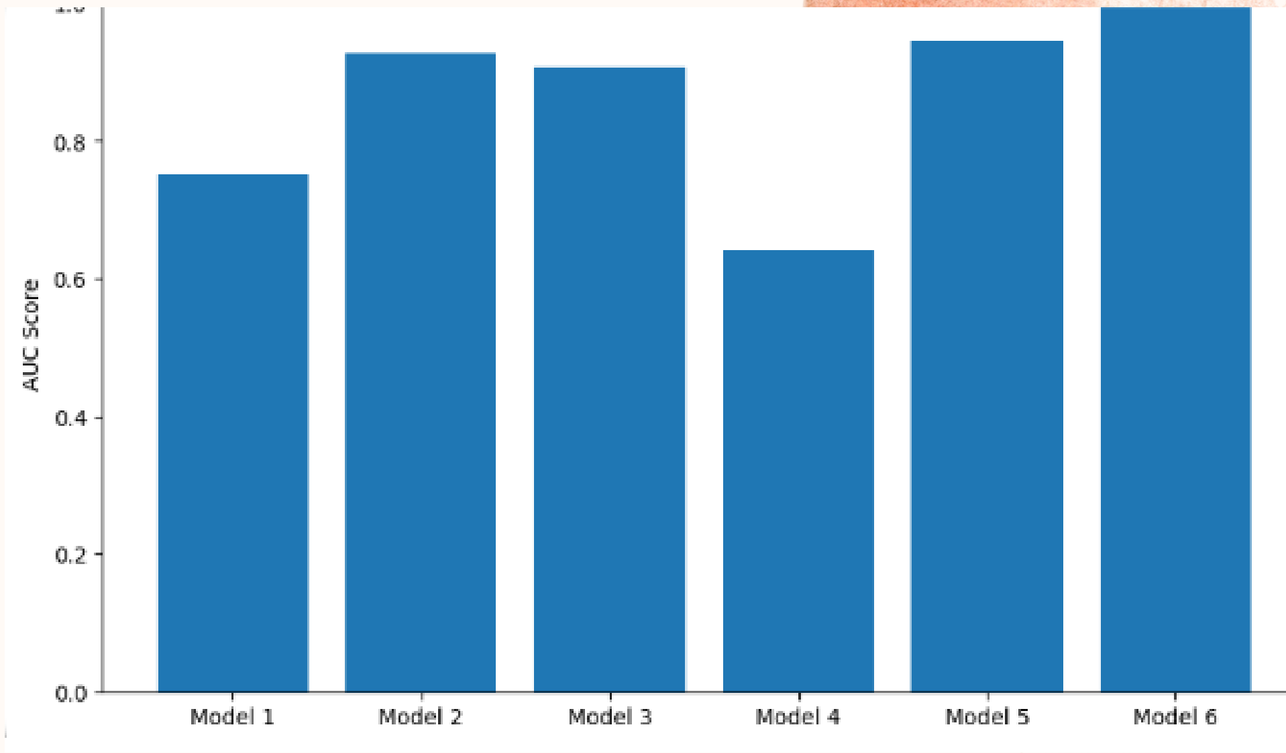
Model2: Logistic Regression Penalized with Lasso
(regularization-strength = 0.1)

Model3: Logistic Regression Penalized with Ridge
(L2 penalty = 50%)

Model4: Logistic Regression Penalized with Elastic Net (L1 penalty = 50%,
L2 penalty = 50%)

Model5: Decision Tree

Medel6: Random Forest




```
Model 1: AUC = 0.7506861718066697
Model 2: AUC = 0.9267654577702052
Model 3: AUC = 0.9079386048955072
Model 4: AUC = 0.6412692385246368
Model 5: AUC = 0.9459782563100856
Model 6: AUC = 0.9935930212658386
```

7). Decision Making

In terms of algorithmic complexity, it is best that we go we random forest classifier. However, in terms of logistic regression model, we select the mentioned features.

Thank you