Airline passengers satisfaction prediction

Süleyman Erim, Giacomo Schiavo, Mattia Varagnolo

Introduction

Objective 🎯

predict whether a passenger will be satisfied or dissatisfied with the services offered by an airline company

The dataset 📑

It contains insights into airline passengers' satisfaction levels and preferences, including demographic details, travel experiences, and perceptions of various services

Variables (1)

- **Satisfaction**: Airline satisfaction level (Satisfaction, Neutral, or Dissatisfaction).
- **Gender**: The gender of the passengers (Female, Male).
- **Customer Type**: The type of customer (Loyal customer, disloyal customer).
- **Type of Travel**: The purpose of the flight (Personal Travel, Business Travel).
- **Class**: The travel class in the plane (Business, Eco, Eco Plus).
- **Age**: The actual age of the passengers.
- **Flight Distance**: The distance of the flight journey.
- **Arrival Delay in Minutes**: Number of minutes delayed during arrival.
- **Departure Delay in Minutes**: Number of minutes delayed during departure.

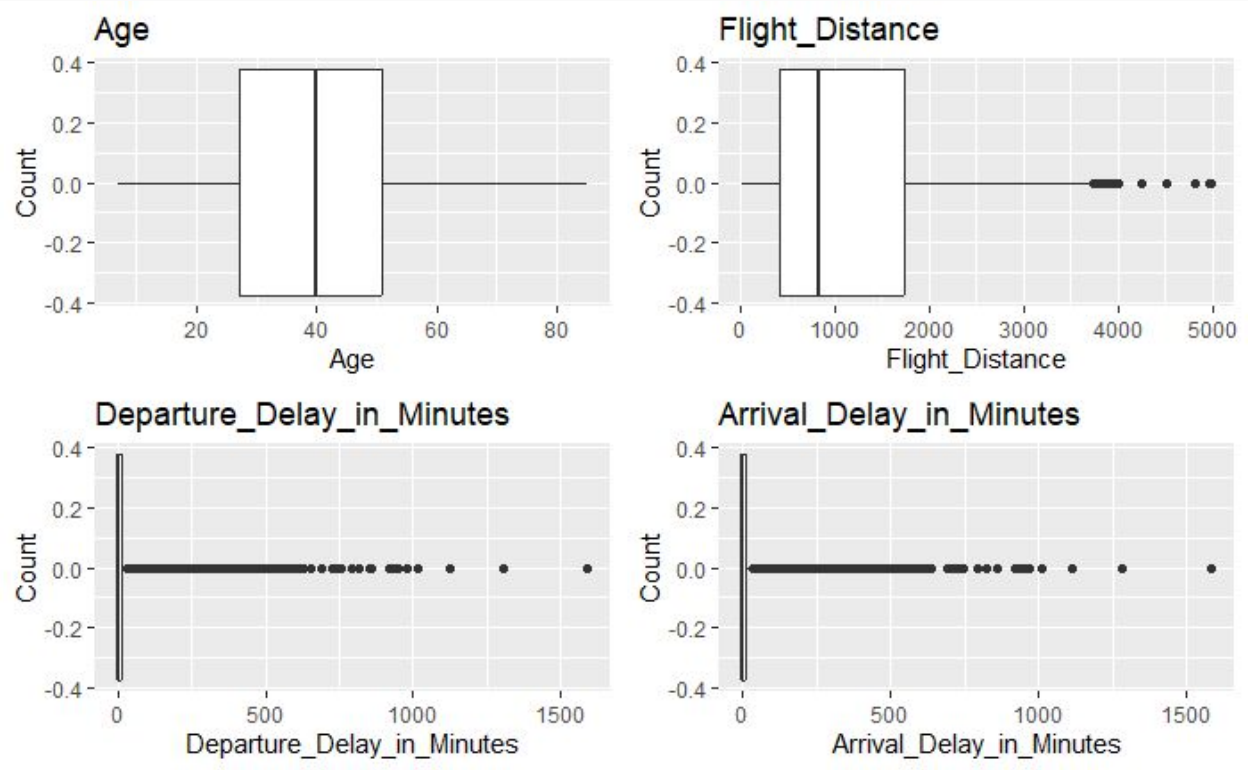
Variables (2)

- **Inflight WiFi Service**: Satisfaction level with the inflight WiFi service.
- **Departure/Arrival Time Convenience**: Satisfaction level with the convenience of departure and arrival times.
- **Ease of Online Booking**: Satisfaction level with the online booking process.
- **Gate Location**: Satisfaction level with the gate location.
- **Food and Drink**: Satisfaction level with the food and drink provided.
- **Online Boarding**: Satisfaction level with the online boarding process.
- **Seat Comfort**: Satisfaction level with the comfort of the seats.
- **Inflight Entertainment**: Satisfaction level with the inflight entertainment options.
- **On-board Service**: Satisfaction level with the service provided onboard.
- **Leg Room Service**: Satisfaction level with the legroom space.
- **Baggage Handling**: Satisfaction level with the handling of baggage.
- **Check-in Service**: Satisfaction level with the check-in service.
- **Inflight Service**: Satisfaction level with the inflight service.
- **Cleanliness**: Satisfaction level with the cleanliness of the aircraft.

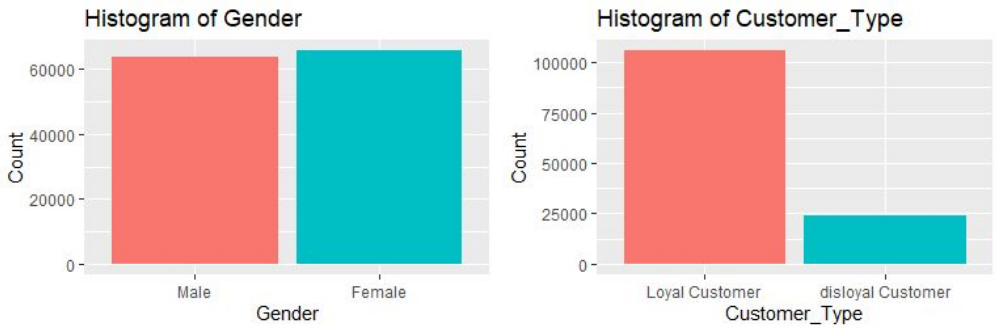
Data preprocessing

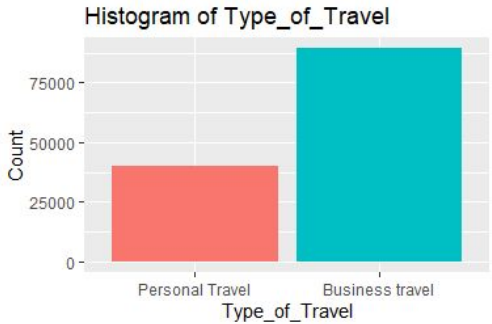
1. Renaming columns
2. Dropping unnecessary columns
3. Converting categorical variables to factors:
    - "Gender", "Customer Type", "Type of Travel" and "Class" and all the rating features
4. Handling NA values in Arrival_Delay_in_Minutes
    - only 3% of the entire dataset

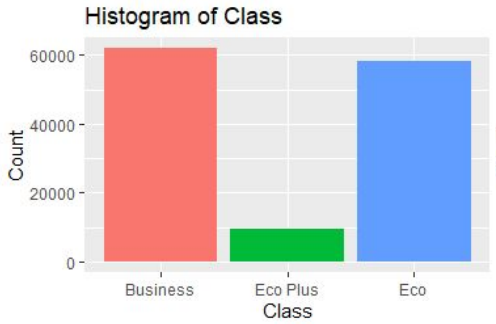
Distributions

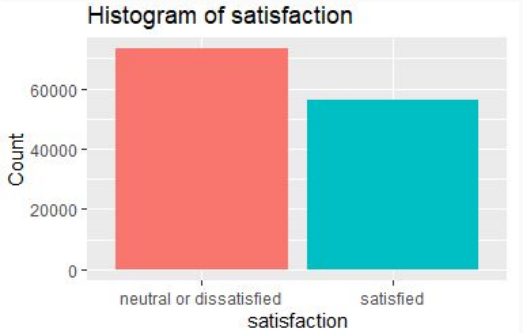


Checking for outliers

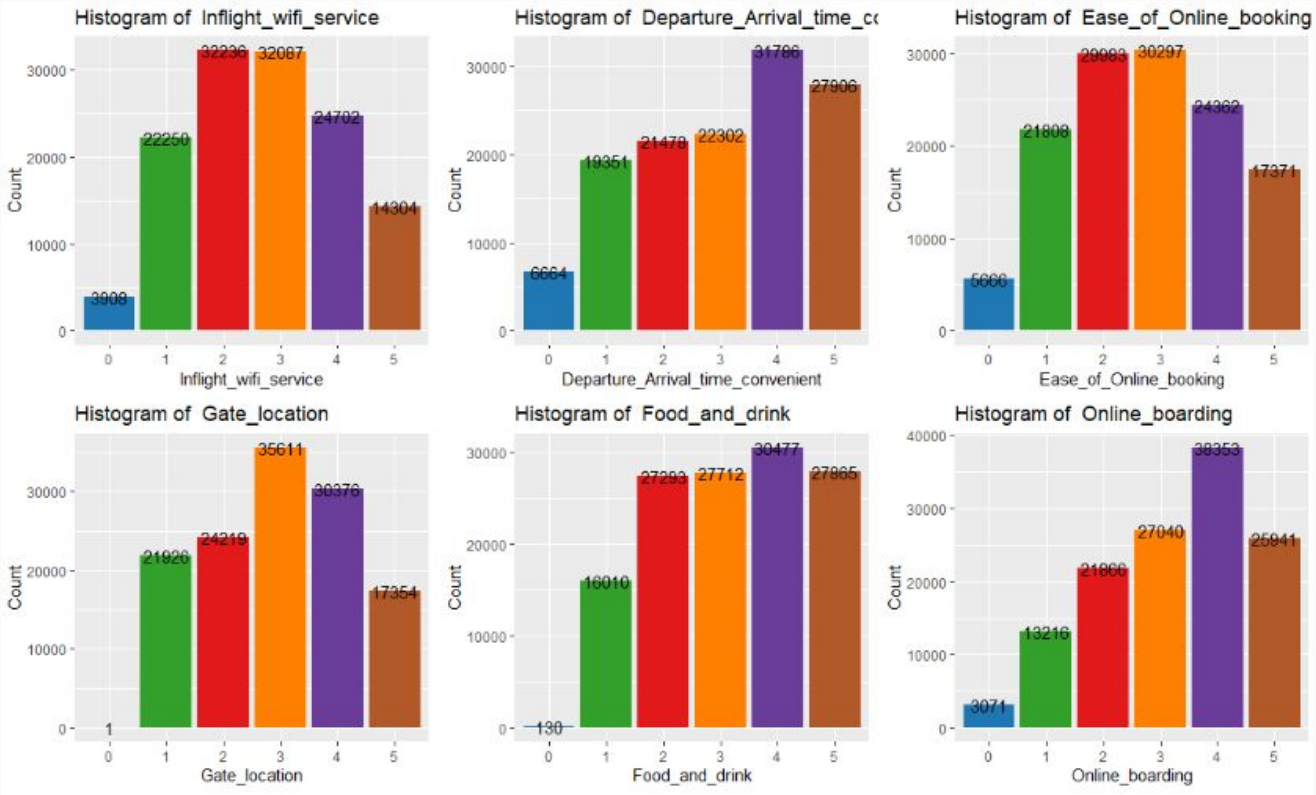




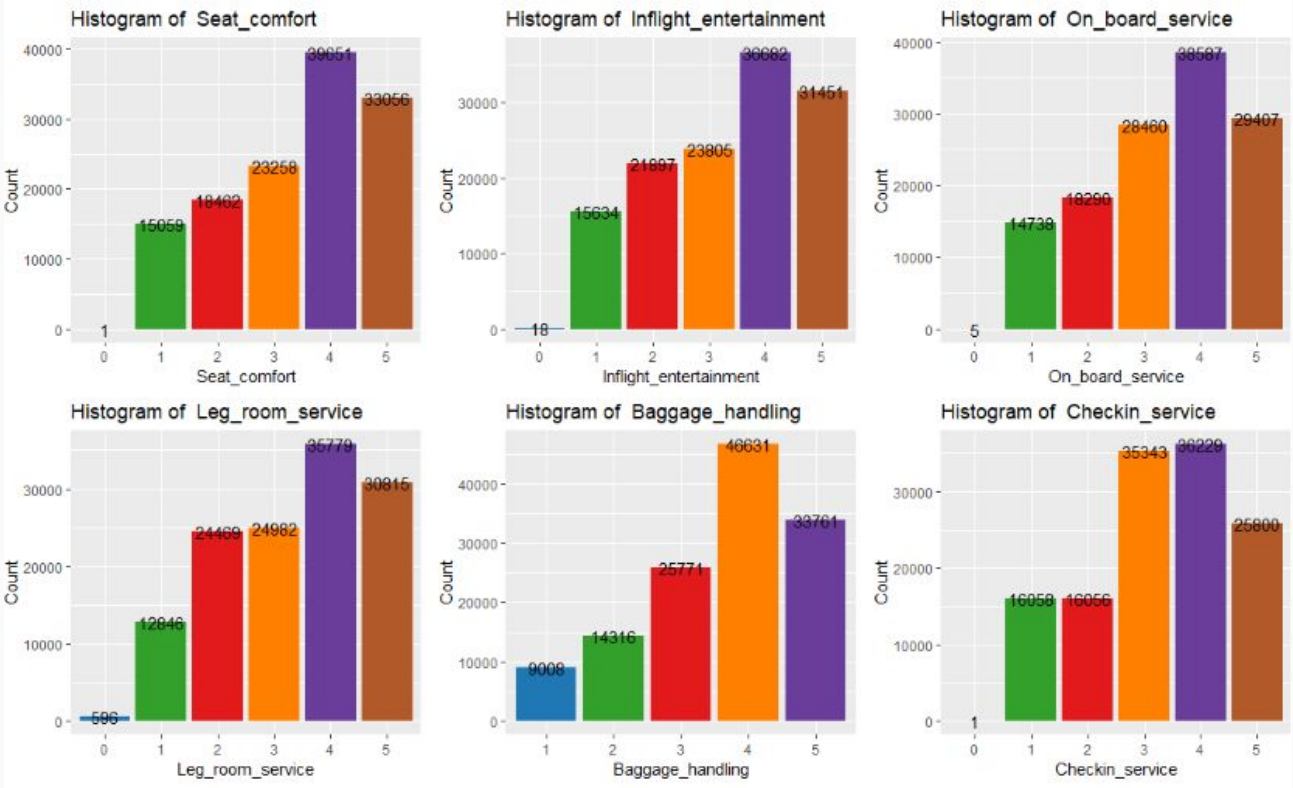




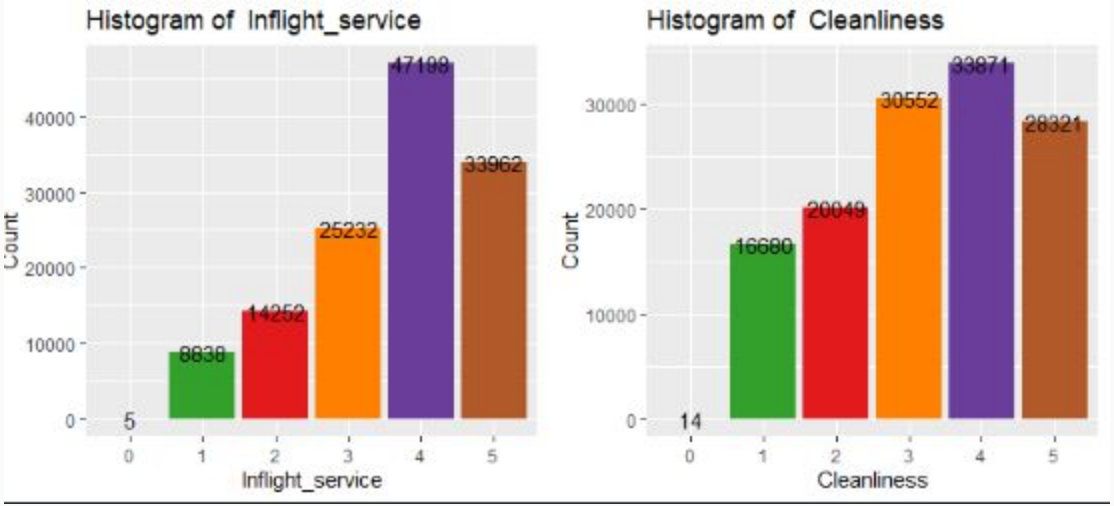
Distribution of nominal features



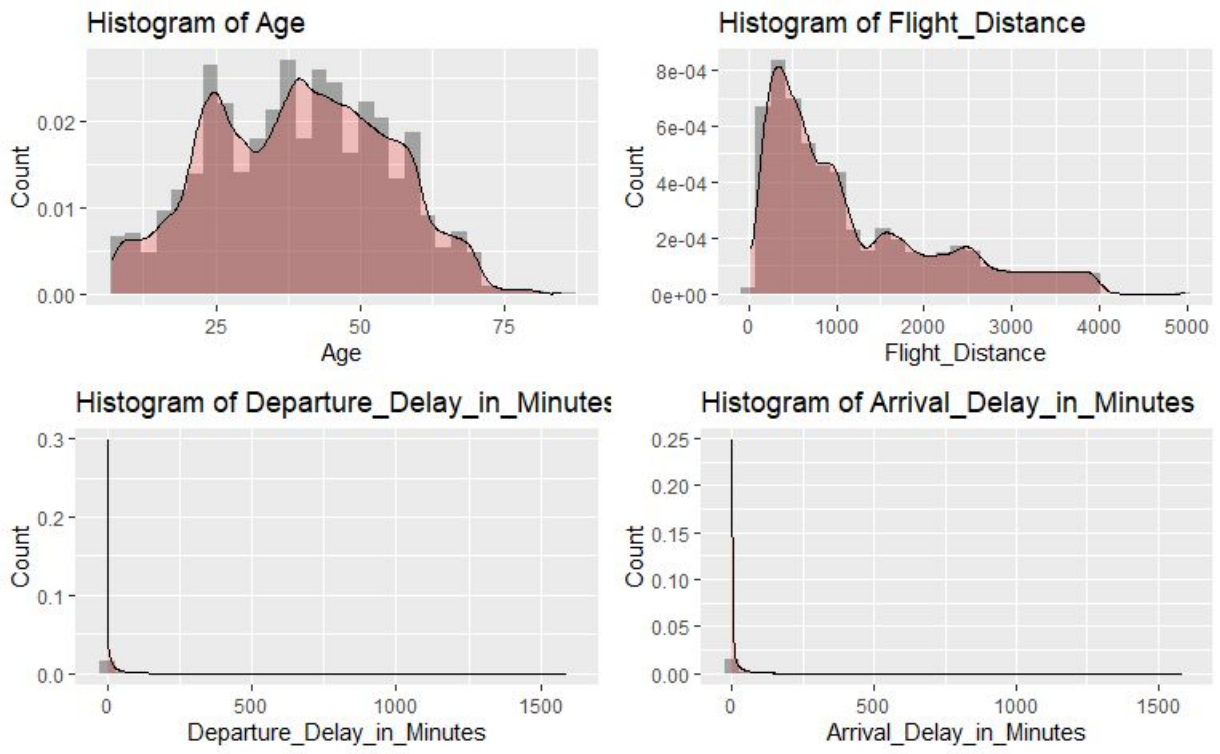
Distribution of the rating features (1)



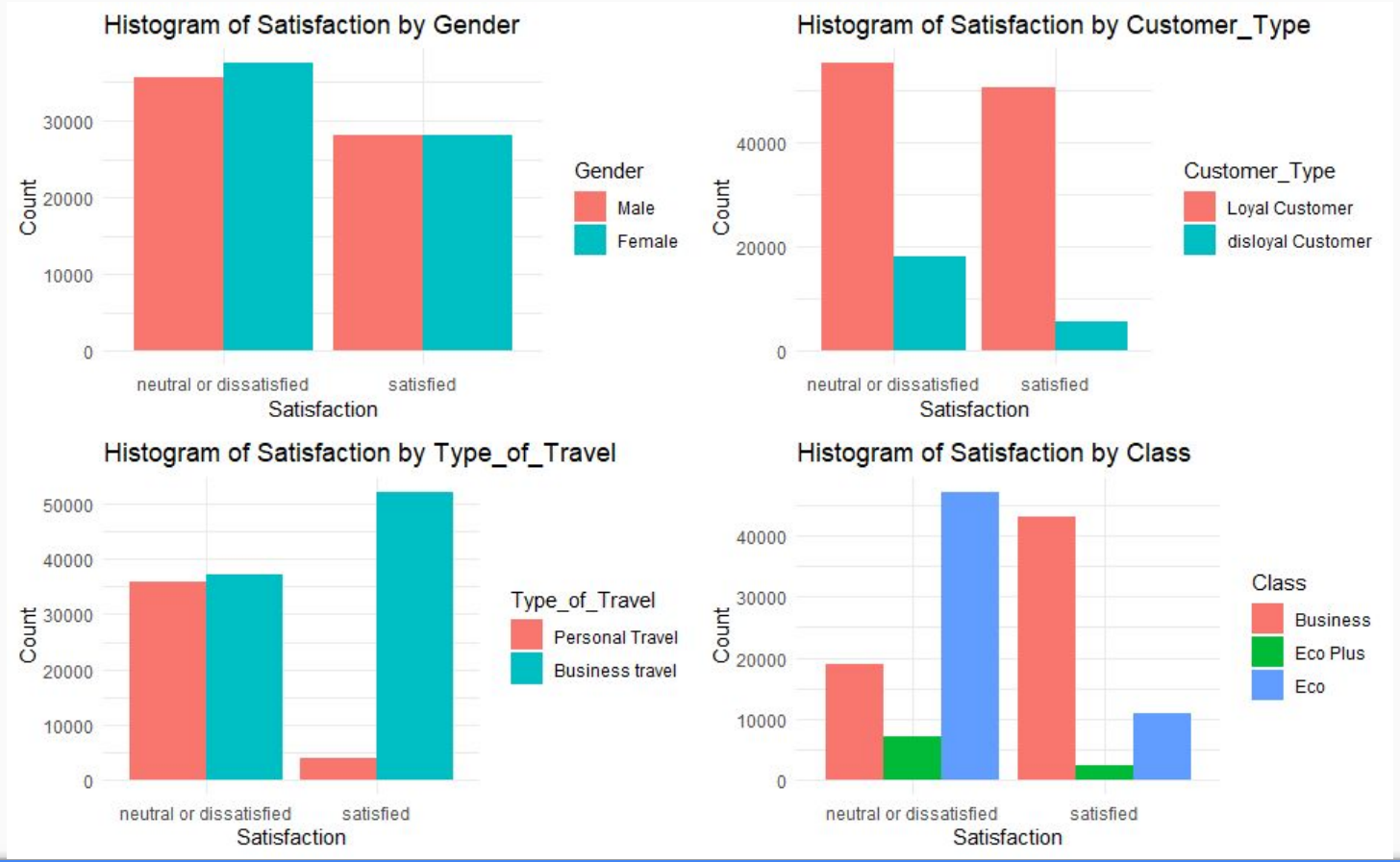
Distribution of the rating features (2)



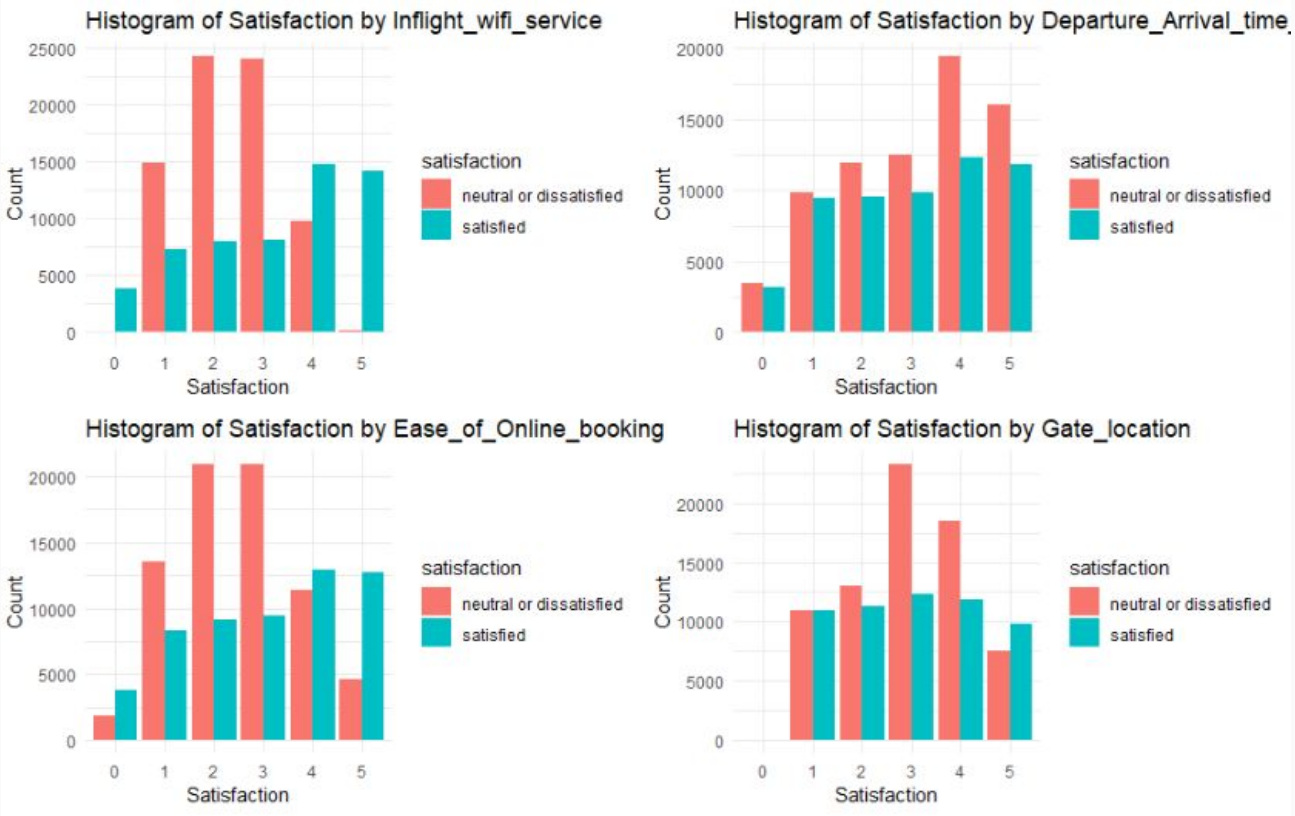
Distribution of the rating features (3)



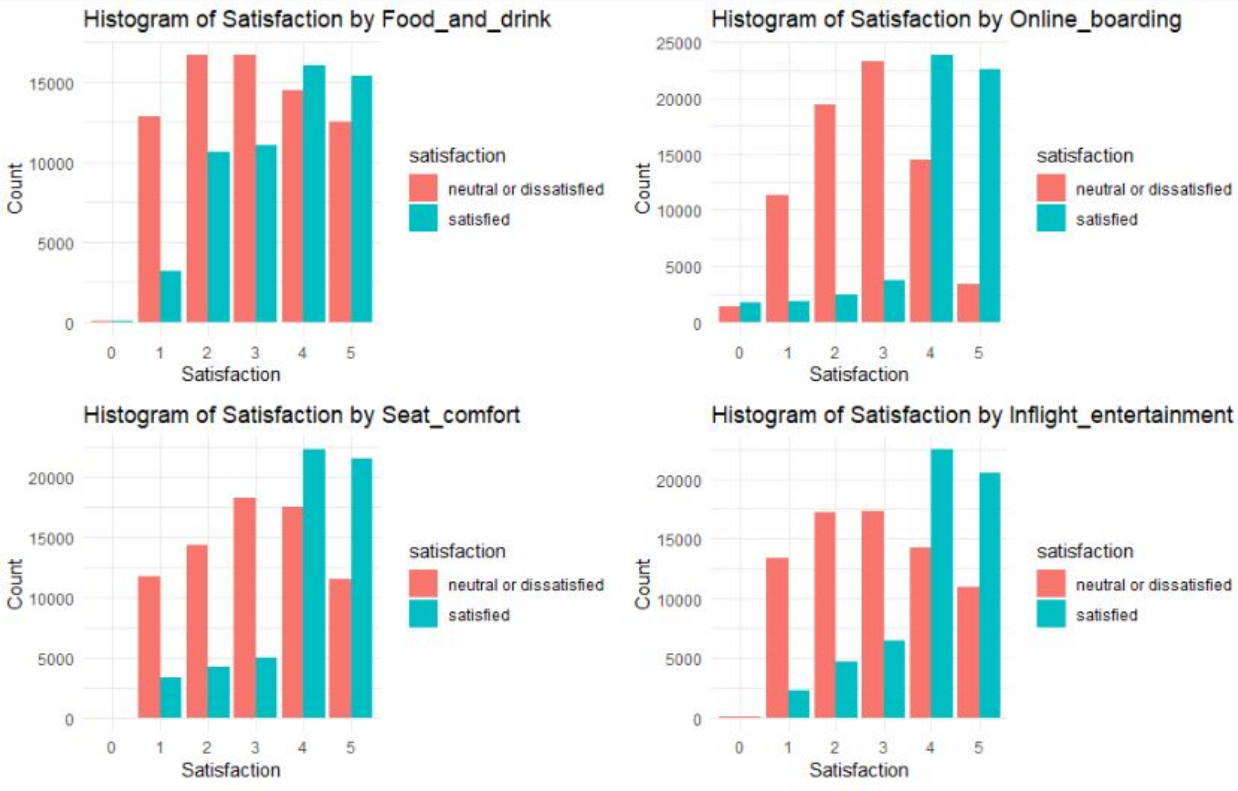
Numerical features



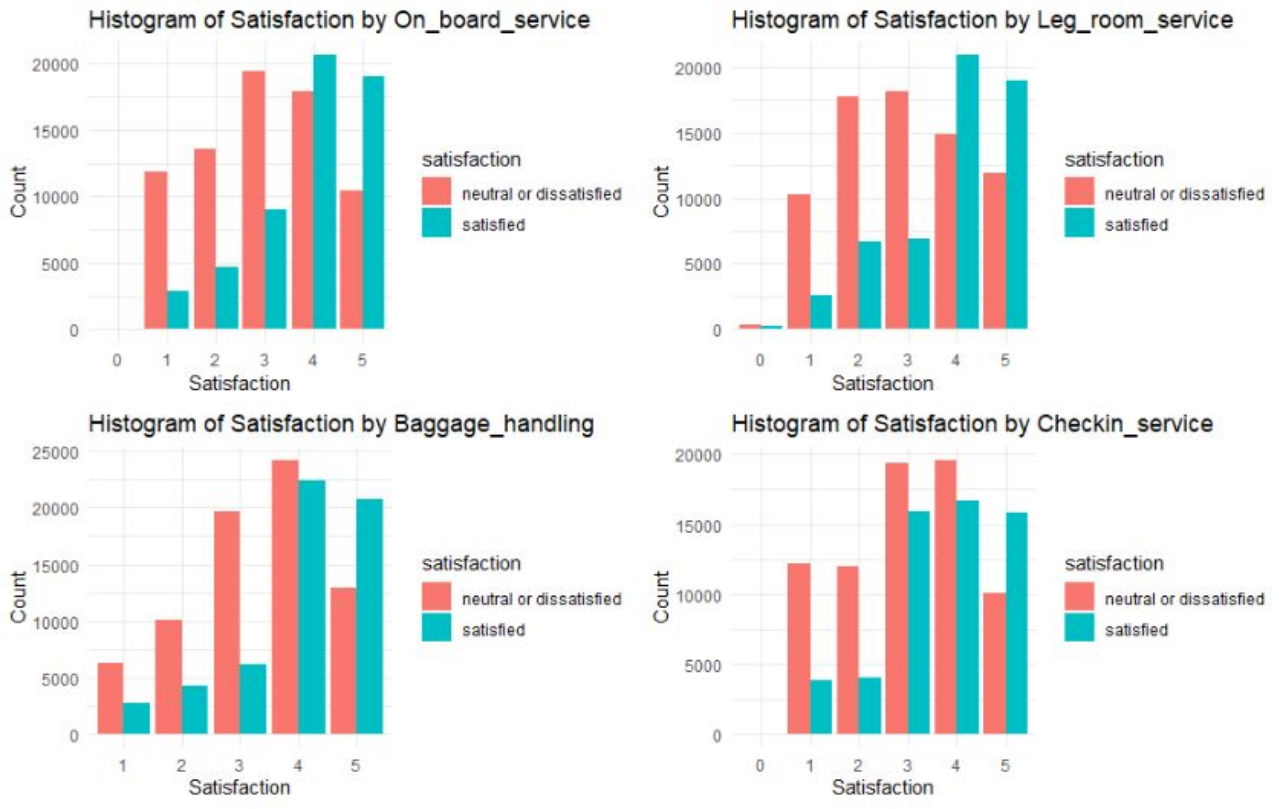
Distribution of nominal features vs satisfaction



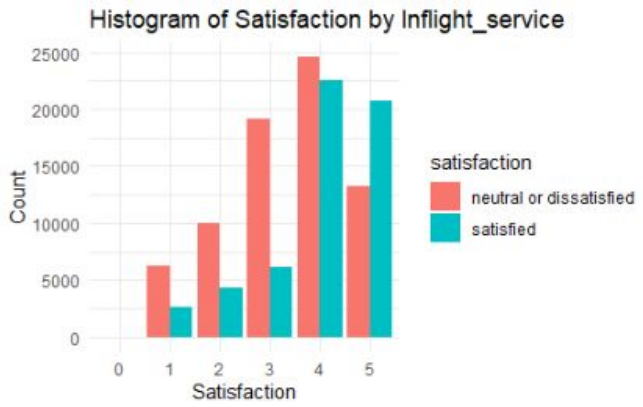
Distribution of rating features vs satisfaction (1)

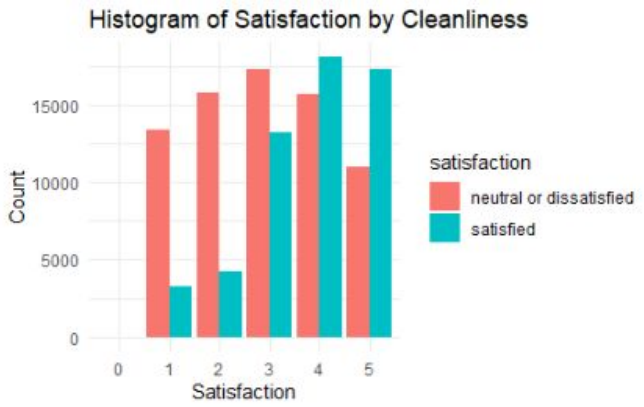


Distribution of rating features vs satisfaction (2)

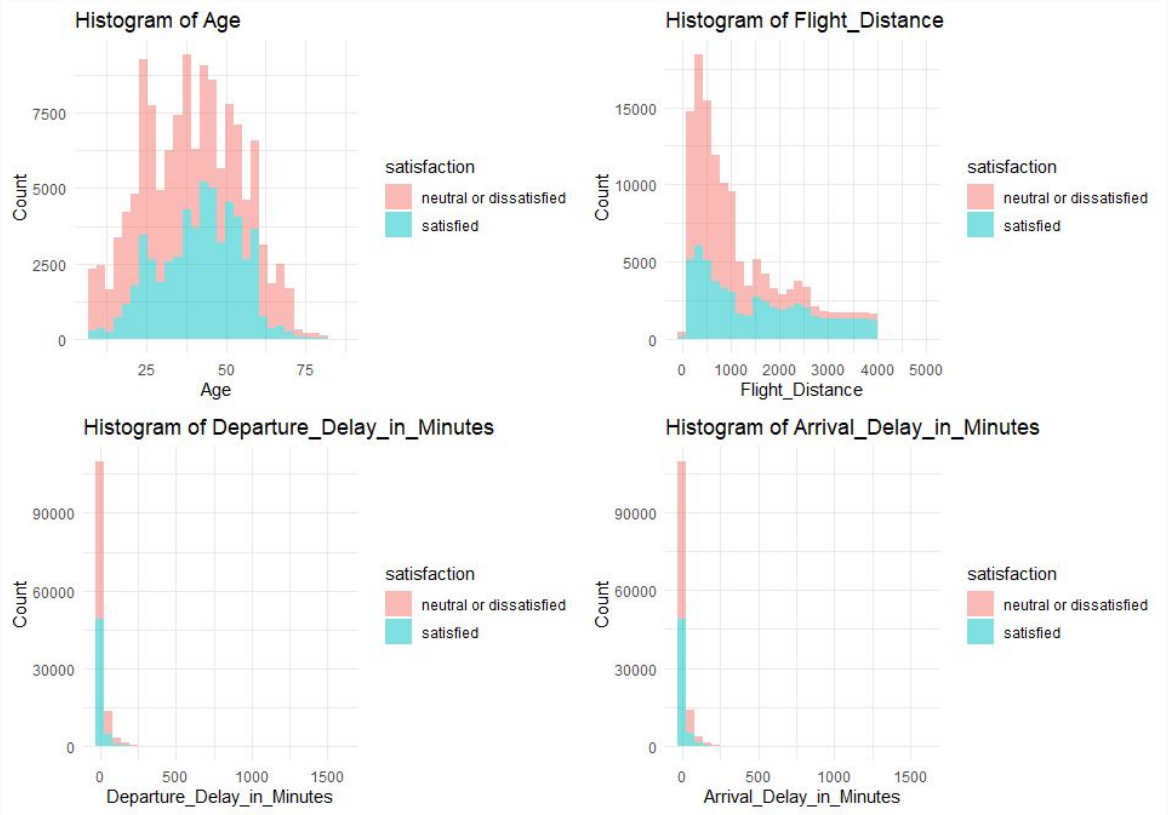


Distribution of rating features vs satisfaction (3)

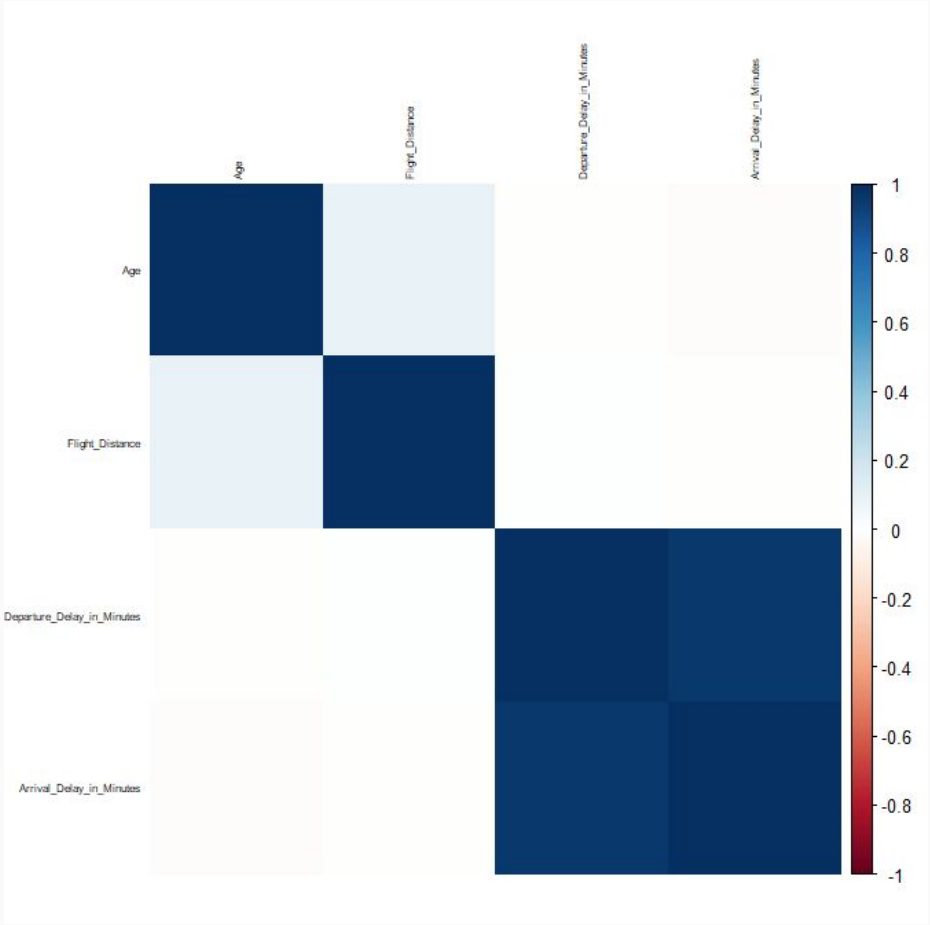




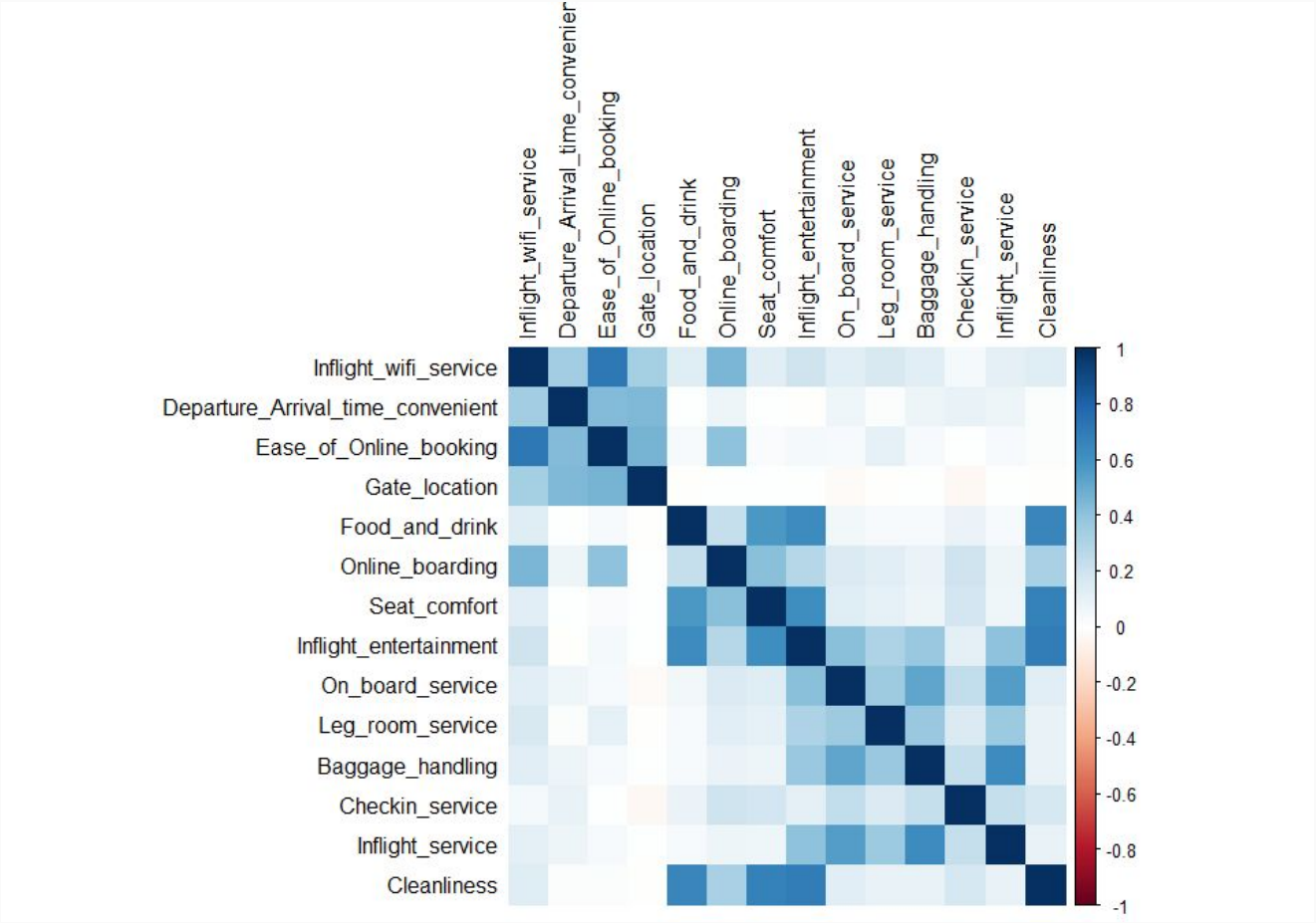
Distribution of rating features vs satisfaction (4)



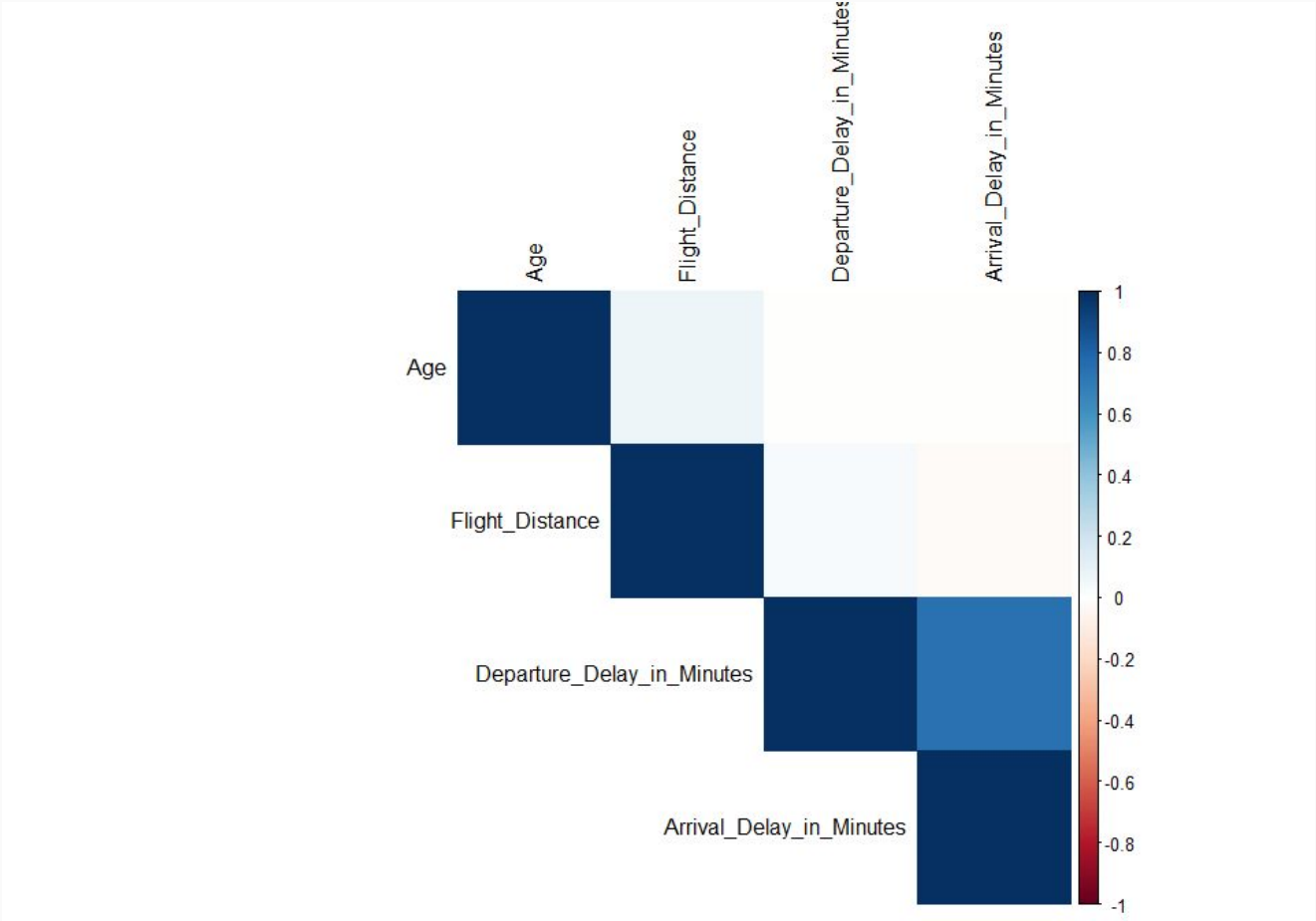
Distribution of numerical features vs satisfaction



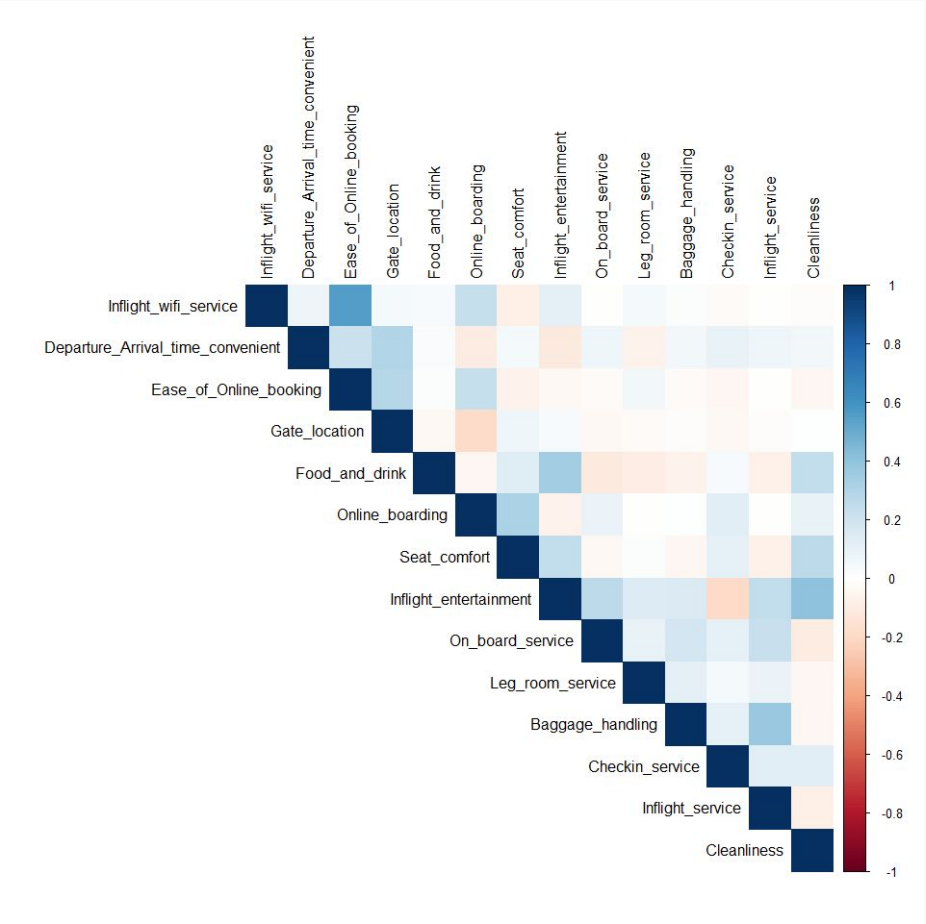
Correlation matrix for numeric variables



Correlation matrix between rating features



Partial correlation between numerical features



Partial correlation between rating features

Categorical to numerical

- Gender
  - Male = 0, Female = 1
- Customer type
  - Loyal customer = 0, disloyal customer = 1
- Type of travel
  - Personal travel = 0, Business travel = 1
- Class
  - Business = 0, Eco = 1, Eco plus = 2
- Satisfaction
  - neutral or dissatisfied = 0, satisfied = 1

Training and test set

Train Test Split:

- Data is split into training and testing sets using a random seed for reproducibility.
- 80% of the data is allocated for training, and 20% is set aside for testing.
- This ensures we have distinct datasets to build and evaluate our model.

Features and Outputs:

- The input features are extracted from the dataset, excluding the 'satisfaction' column.
- The target variable, 'satisfaction', is separated for further analysis.

Number of Samples:

- The number of samples in the training data: 103589.
- The number of samples in the test data: 25898.

Data Balance:

- Proportion of satisfied and dissatisfied customers in the training data:
  - Satisfied: 56.7%
  - Dissatisfied: 43.3%
- Proportion of satisfied and dissatisfied customers in the test data:
  - Satisfied: 56%
  - Dissatisfied: 44%

Classification models

Methodology

- **Parametric Approaches**
  - Logistic Classifier
  - Basic Logistic Regression
  - Logistic Regression with Backward Variable Selection
  - Logistic Regression with Shrinkage Method
  - Naive Bayes
- **Non-Parametric Approach**
  - K-Nearest Neighbors (KNN)

Basic Logistic Regression

VIF : 2.045321

```
##
## Call:
## glm(formula = satisfaction ~ ., family = "binomial", data = train_data)
##
## Coefficients:
##                                    Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                       -8.662e+00  8.467e-02 -102.293  < 2e-16 ***
## Gender                             6.747e-02  1.947e-02    3.465  0.00053 ***
## Customer_Type                     -2.013e+00  2.969e-02  -67.789  < 2e-16 ***
## Age                               -7.945e-03  7.109e-04  -11.176  < 2e-16 ***
## Type_of_Travel                     2.749e+00  3.131e-02   87.802  < 2e-16 ***
## Class                             -3.491e-01  1.281e-02  -27.266  < 2e-16 ***
## Flight_Distance                   -1.281e-06  1.116e-05   -0.115  0.90861
## Inflight_wifi_service              3.988e-01  1.148e-02   34.745  < 2e-16 ***
## Departure_Arrival_time_convenient -1.333e-01  8.178e-03  -16.301  < 2e-16 ***
## Ease_of_Online_booking            -1.535e-01  1.134e-02  -13.541  < 2e-16 ***
## Gate_location                      2.288e-02  9.178e-03    2.493  0.01266 *
## Food_and_drink                    -2.982e-02  1.070e-02   -2.785  0.00534 **
## Online_boarding                    6.243e-01  1.028e-02   60.734  < 2e-16 ***
## Seat_comfort                       5.763e-02  1.120e-02    5.145 2.68e-07 ***
## Inflight_entertainment             5.710e-02  1.425e-02    4.007 6.14e-05 ***
## On_board_service                   3.088e-01  1.017e-02   30.375  < 2e-16 ***
## Leg_room_service                   2.547e-01  8.530e-03   29.856  < 2e-16 ***
## Baggage_handling                   1.332e-01  1.141e-02   11.670  < 2e-16 ***
## Checkin_service                    3.288e-01  8.560e-03   38.417  < 2e-16 ***
## Inflight_service                   1.224e-01  1.204e-02   10.165  < 2e-16 ***
## Cleanliness                        2.294e-01  1.212e-02   18.933  < 2e-16 ***
## Departure_Delay_in_Minutes         5.019e-03  9.851e-04    5.095 3.48e-07 ***
## Arrival_Delay_in_Minutes          -9.839e-03  9.731e-04  -10.111  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 141746  on 103588  degrees of freedom
## Residual deviance:  69302  on 103566  degrees of freedom
## AIC: 69348
##
## Number of Fisher Scoring iterations: 6
```

Logistic Regression with Backward Variable Selection

	Variable	VIF
1	Arrival_Delay_in_Minutes	14.214669
2	Departure_Delay_in_Minutes	14.179713
3	Inflight_entertainment	3.256814
4	Ease_of_Online_booking	2.605357
5	Cleanliness	2.466629
6	Inflight_wifi_service	2.227467
7	Seat_comfort	2.050528
8	Food_and_drink	2.019669
9	Inflight_service	2.011267
10	Type_of_Travel	1.846157
11	Baggage_handling	1.818810
12	Departure_Arrival_time_convenient	1.716303
13	On_board_service	1.635092
14	Customer_Type	1.593352
15	Class	1.580339
16	Gate_location	1.523870
17	Online_boarding	1.492857
18	Flight_Distance	1.321514
19	Leg_room_service	1.218411
20	Checkin_service	1.208941
21	Age	1.180740
22	Gender	1.007157

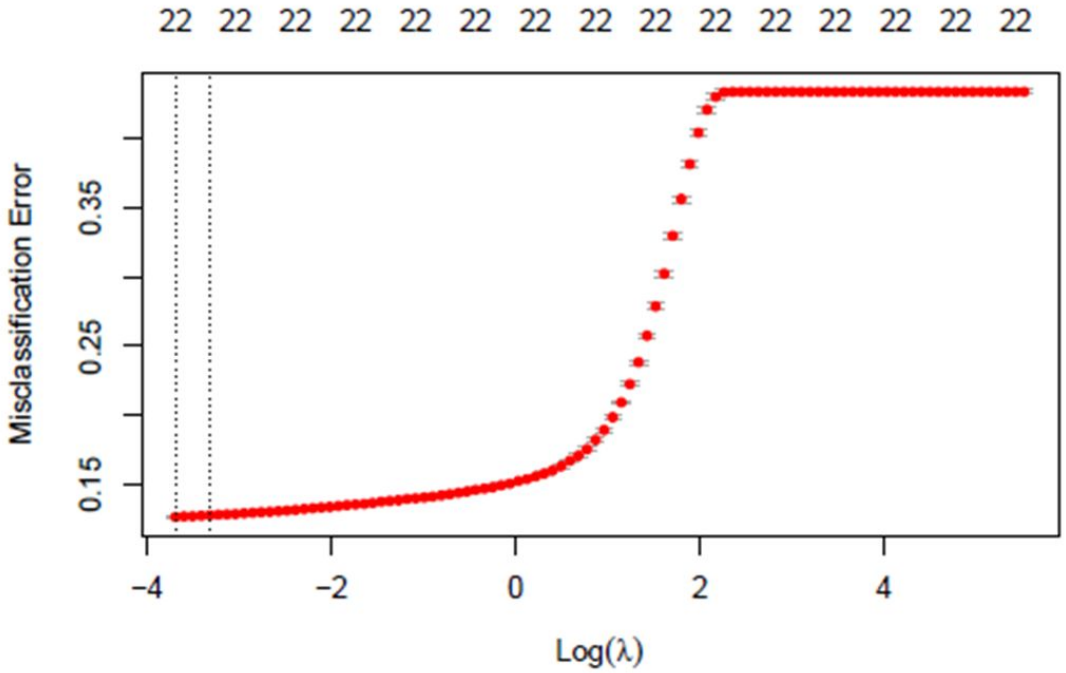
after 4 iterations

	Variable	VIF
1	Inflight_service	1.881686
2	Baggage_handling	1.787458
3	Type_of_Travel	1.776955
4	Departure_Arrival_time_convenient	1.587989
5	Seat_comfort	1.585879
6	Class	1.563525
7	Inflight_wifi_service	1.551518
8	On_board_service	1.534811
9	Customer_Type	1.512988
10	Food_and_drink	1.431732
11	Online_boarding	1.408212
12	Gate_location	1.401531
13	Flight_Distance	1.321957
14	Leg_room_service	1.202873
15	Age	1.166393
16	Checkin_service	1.164245
17	Departure_Delay_in_Minutes	1.013281
18	Gender	1.004409

Logistic Regression with Backward Variable Selection Model Output

```
##
## Call:
## glm(formula = satisfaction ~ . - Arrival_Delay_in_Minutes - Inflight_entertainment -
##     Ease_of_Online_booking - Cleanliness, family = "binomial",
##     data = train_data)
##
## Coefficients:
##                                  Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                    -8.753e+00  8.187e-02 -106.909  < 2e-16 ***
## Gender                          7.017e-02  1.929e-02    3.638 0.000275 ***
## Customer_Type                  -2.047e+00  2.906e-02  -70.452  < 2e-16 ***
## Age                            -8.049e-03  7.022e-04  -11.463  < 2e-16 ***
## Type_of_Travel                  2.757e+00  3.074e-02   89.667  < 2e-16 ***
## Class                          -3.243e-01  1.264e-02  -25.653  < 2e-16 ***
## Flight_Distance                -2.802e-06  1.103e-05   -0.254 0.799509
## Inflight_wifi_service           3.202e-01  9.479e-03   33.776  < 2e-16 ***
## Departure_Arrival_time_convenient -1.682e-01  7.817e-03  -21.520  < 2e-16 ***
## Gate_location                  -1.325e-02  8.707e-03   -1.522 0.128025
## Food_and_drink                  8.632e-02  8.915e-03    9.683  < 2e-16 ***
## Online_boarding                 6.124e-01  9.978e-03   61.379  < 2e-16 ***
## Seat_comfort                    1.865e-01  9.774e-03   19.080  < 2e-16 ***
## On_board_service                3.241e-01  9.740e-03   33.278  < 2e-16 ***
## Leg_room_service                2.540e-01  8.403e-03   30.228  < 2e-16 ***
## Baggage_handling                1.522e-01  1.116e-02   13.629  < 2e-16 ***
## Checkin_service                 3.329e-01  8.317e-03   40.023  < 2e-16 ***
## Inflight_service                1.445e-01  1.149e-02   12.575  < 2e-16 ***
## Departure_Delay_in_Minutes     -4.336e-03  2.619e-04  -16.553  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 141746  on 103588  degrees of freedom
## Residual deviance:  70165  on 103570  degrees of freedom
## AIC: 70203
##
## Number of Fisher Scoring iterations: 5
```

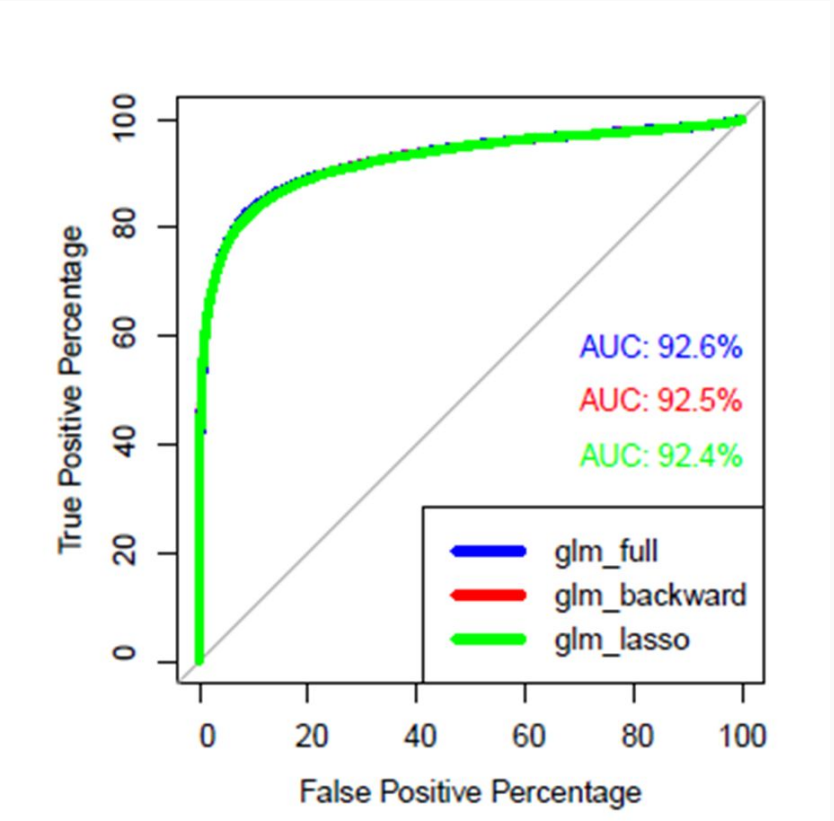
Logistic Regression with Shrinkage Method Lasso Regression




```
# We identify th best lambda value
best_lambda <- glm_lasso$lambda.min
best_lambda
```

```
## [1] 0.02491896
```

Comparison of Models - ROC CURVE



Logistic Regression Model Selection

Threshold	Accuracy	F1_Score	Precision	Recall
0.4	0.8586	0.8711	0.8892	0.8538
0.5	0.8705	0.8862	0.8725	0.9004
0.6	0.8722	0.8914	0.8508	0.9360
0.7	0.8628	0.8871	0.8224	0.9628

Logistic Regression with Backward Variable Selection

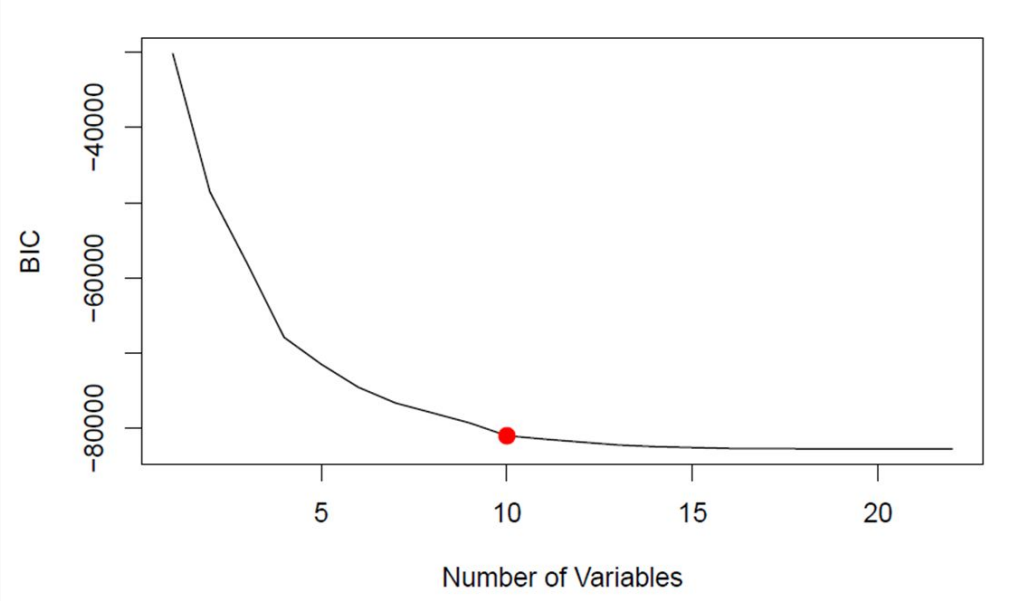
Threshold	Accuracy	F1_Score	Precision	Recall
0.4	0.8616	0.8740	0.8913	0.8574
0.5	0.8727	0.8881	0.8745	0.9021
0.6	0.8737	0.8926	0.8521	0.9371
0.7	0.8637	0.8877	0.8241	0.9619

Logistic Regression with Shrinkage Method

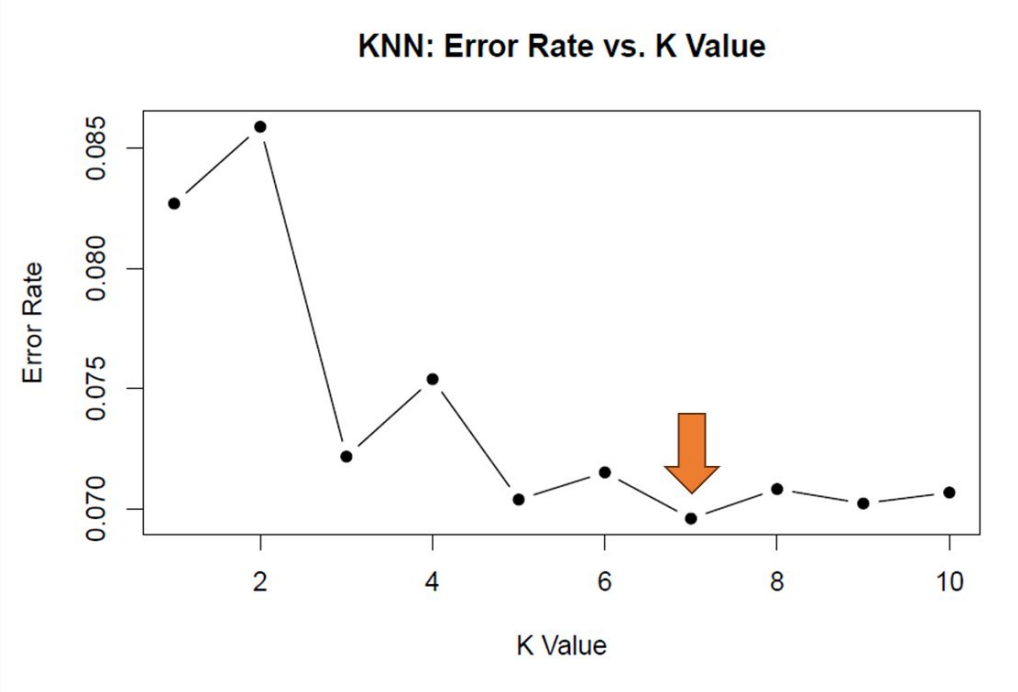
Threshold	Accuracy	F1_Score	Precision	Recall
0.4	0.8566	0.8690	0.8900	0.8489
0.5	0.8707	0.8868	0.8696	0.9047
0.6	0.8711	0.8916	0.8428	0.9463
0.7	0.8527	0.8808	0.8051	0.9723

Naïve Bayes

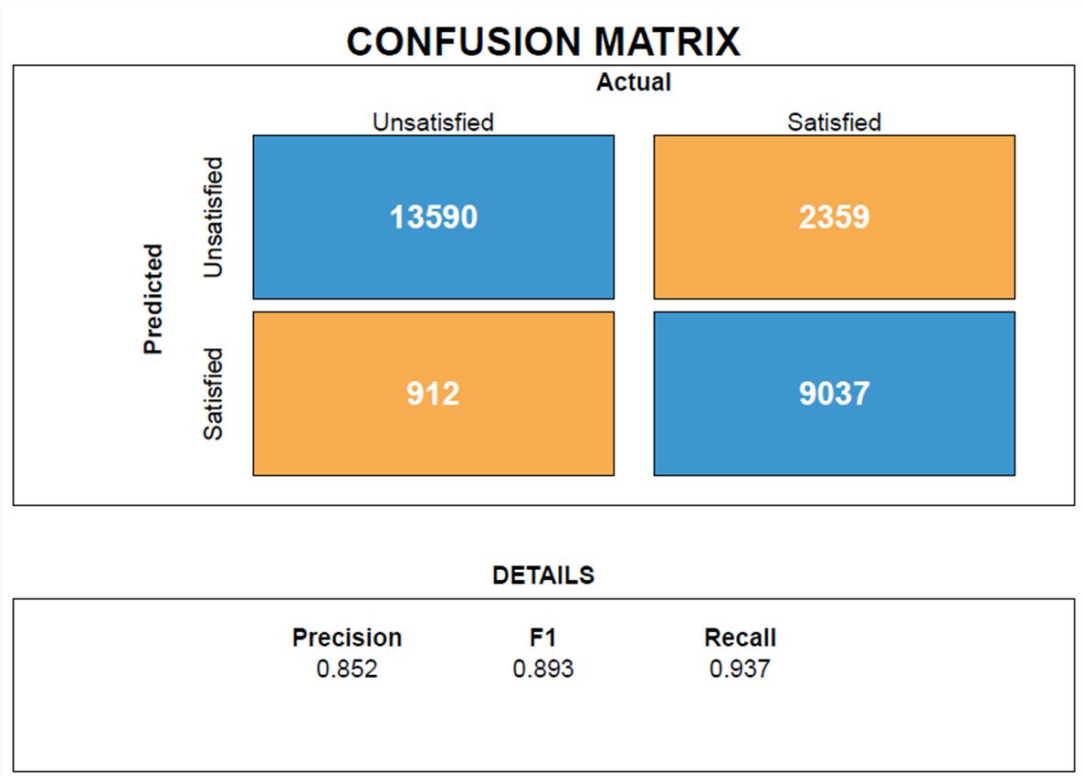
- Choose the model with 10 variables



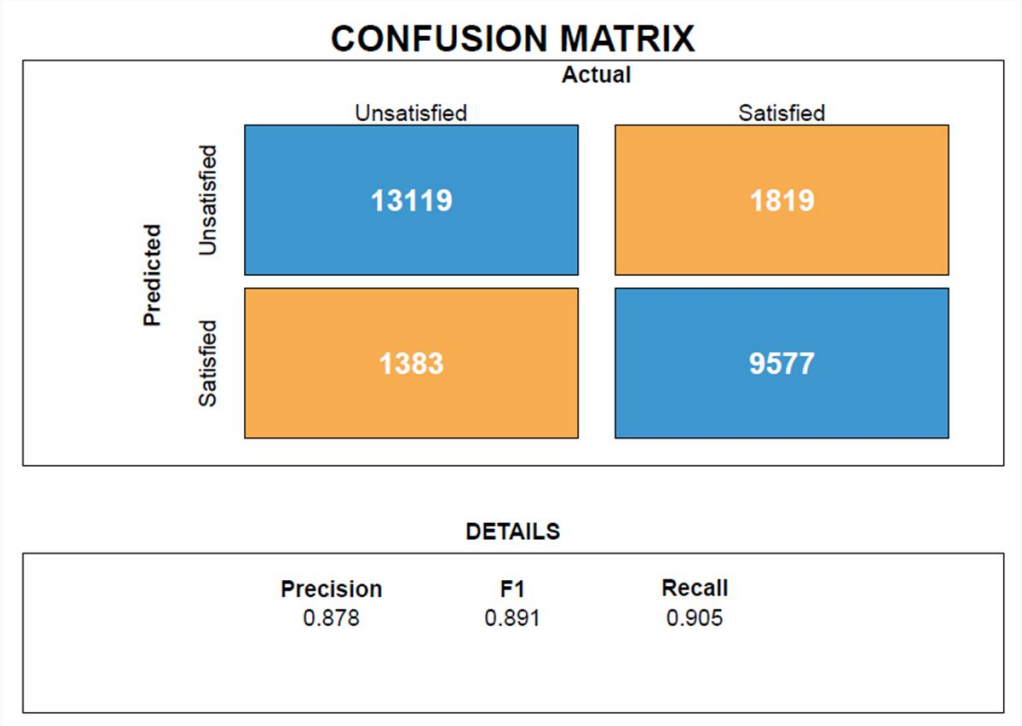
KNN with Cross Validation



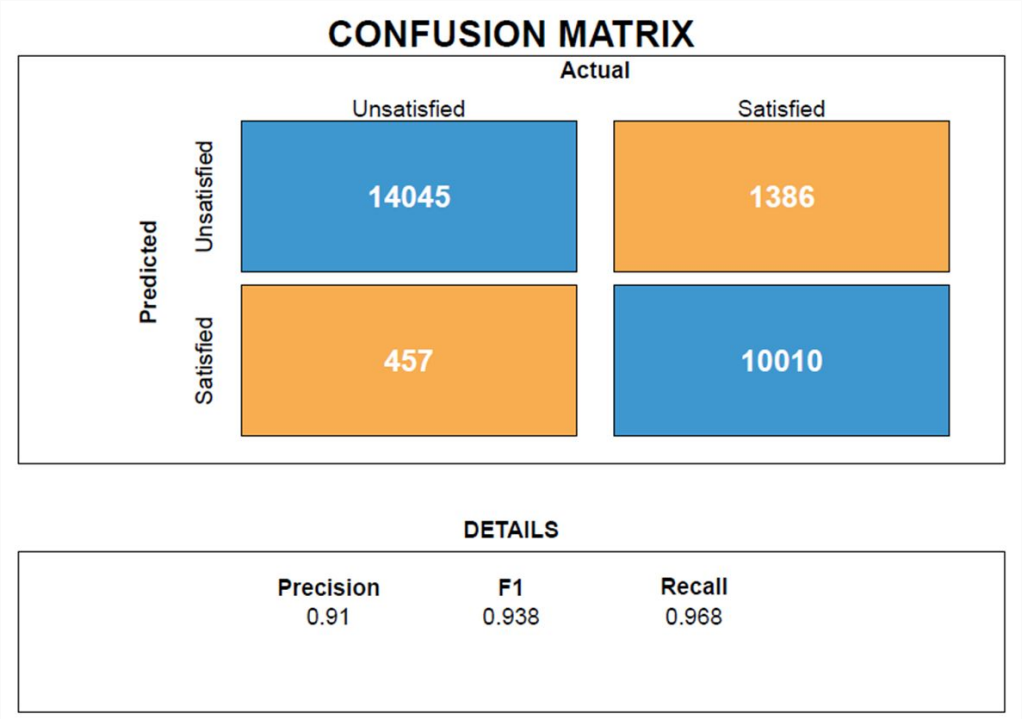
Classification results - logistic regression



Classification Results – Naïve Bayes



Classification Results - KNN



Thank you for your attention