T.Y. B.SC. APPLIED STATISTICS AND DATA ANALYTICS (HONS.)
ACADEMIC YEAR 2023-24

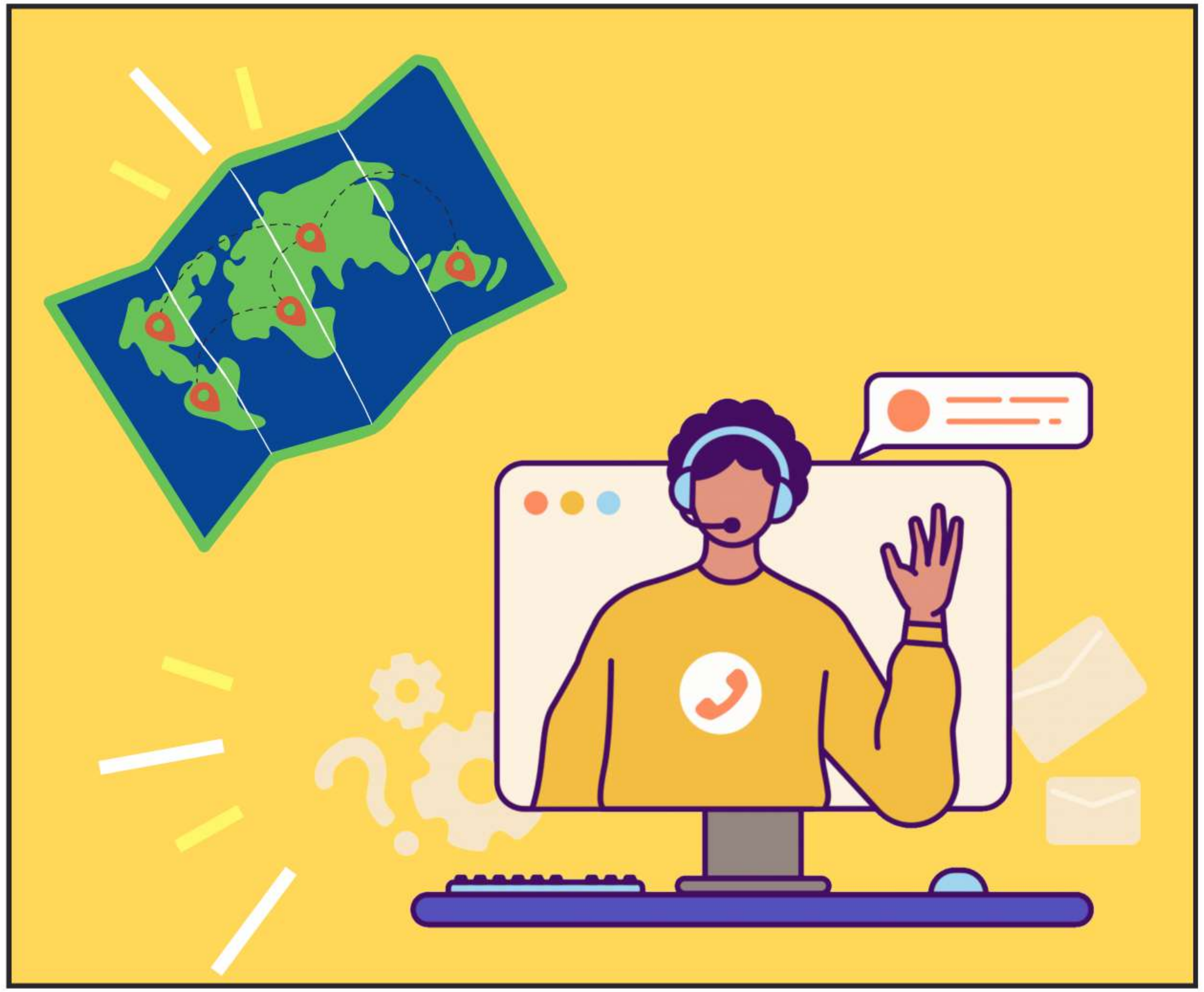
R FOR DATA SCIENCE

ICA II PROJECT

GROUP 3

Submitted By:
Senorita Chettiar, D004
Grihit Salian, D013
Omkar Telkar, D019
Revati Vitthal, D022
Yash Bhanushali, D024

Submitted To:
Mrs. Ashwini Ghanekar,
Ms. Gazala Sayed
Department of Statistics
Mithibai College(Mumbai)



Customer Satisfaction Analysis & Modelling

Understanding Key Drivers of Customer Contentment

INTRODUCTION

INTRODUCTION

- 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.

PURPOSE

The purpose of this analysis is to explore how airlines can identify the key factors contributing to passenger satisfaction.

By understanding the importance of each feature, airlines can enhance their brand image, increase customer loyalty, and boost engagement by effectively meeting and exceeding customer expectations.

Problem Statement

PROBLEM STATEMENT

- AIRLINES NEED TO IDENTIFY THE KEY FACTORS THAT DRIVE PASSENGER SATISFACTION TO IMPROVE THEIR BRAND IMAGE AND CUSTOMER LOYALTY.

- THIS ANALYSIS AIMS TO DETERMINE WHICH FEATURES MOST INFLUENCE CUSTOMER SATISFACTION, PROVIDING ACTIONABLE INSIGHTS FOR ENHANCING THE OVERALL CUSTOMER EXPERIENCE.

IMPACT OF THE PROBLEM ON THE BUSINESS AND WHY IT NEEDS TO BE SOLVED.

CUSTOMER RETENTION

Failure to identify key satisfaction factors can lead to losing customers to competitors.

BRAND REPUTATION

Brand Reputation reflects how customers and the public perceive an airline. It is a critical indicator of trust, service quality, and customer satisfaction. A strong brand reputation can drive passenger loyalty, influence customer choices, and enhance competitive advantage in the market, making it essential for sustained business success.

MARKET SHARE & REVENUE

Poor satisfaction leads to decreased market share and profitability.

IMPORTANCE OF SOLVING

Understanding satisfaction drivers allows airlines to tailor services, boosting loyalty and engagement.

BUSINESS IMPACT

Enhanced customer experience strengthens competitive position and drives higher profitability.

OBJECTIVE

Objectives

Key Factor Analysis

Identify and analyze the key factors that influence passenger satisfaction in the airline industry.

Customer Engagement

Leverage customer expectations to boost engagement.

Objectives

Actionable Insights

Provide actionable insights for airlines to enhance their brand image and improve customer loyalty.

Business Benefits

Highlight the benefits of improving customer satisfaction.

DATASET OVERVIEW

Overview

Data summary

Name	airline_satisfaction
Number of rows	103904
Number of columns	25
________________________	
Column type frequency:	
character	5
numeric	20
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gender	0	1	4	6	0	2	0
customer_type	0	1	14	17	0	2	0
type_of_travel	0	1	15	15	0	2	0
class	0	1	3	8	0	3	0
satisfaction	0	1	9	23	0	2	0

Overview

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p5
sr	0	1	51951.50	29994.65	0	25975.75	51951.
id	0	1	64924.21	37463.81	1	32533.75	64856.
age	0	1	39.38	15.11	7	27.00	40.
flight_distance	0	1	1189.45	997.15	31	414.00	843.
inflight_wifi_service	0	1	2.73	1.33	0	2.00	3.
departure_arrival_time_convenient	0	1	3.06	1.53	0	2.00	3.
ease_of_online_booking	0	1	2.76	1.40	0	2.00	3.
gate_location	0	1	2.98	1.28	0	2.00	3.
food_and_drink	0	1	3.20	1.33	0	2.00	3.
online_boarding	0	1	3.25	1.35	0	2.00	3.
seat_comfort	0	1	3.44	1.32	0	2.0	
inflight_entertainment	0	1	3.36	1.33	0	2.0	
on_board_service	0	1	3.38	1.29	0	2.00	4.
leg_room_service	0	1	3.35	1.32	0	2.00	4.
baggage_handling	0	1	3.63	1.18	1	3.00	4.
checkin_service	0	1	3.30	1.27	0	3.00	3.
inflight_service	0	1	3.64	1.18	0	3.00	4.
cleanliness	0	1	3.29	1.31	0	2.00	3.
departure_delay_in_minutes	0	1	14.82	38.23	0	0.00	0.
arrival_delay_in_minutes	310	1	15.18	38.70	0	0.00	0.

- Five(5) Variables are character classes, we will convert them to factors.
- The label is consists of two(2) classes “Yes” and “No”.
- Only arrival_delay_in_minutes variable has missing values(310).
- The age range is between 7 and 85 years.

Data Cleaning & Processing

```r
airline_satisfaction <- airline_satisfaction %>%
  mutate(across(where(is.character), ~ as.factor(str_squish(str_to_title(.))))) %>%
  mutate(
    satisfaction = str_replace_all(satisfaction, "Neutral Or Dissatisfied", replacement = "No"),
    satisfaction = str_replace_all(satisfaction, "Satisfied", replacement = "Yes"),
    satisfaction = factor(satisfaction, levels = c("Yes", "No")),
    arrival_delay_in_minutes = as.numeric(str_replace_na(
      arrival_delay_in_minutes,
      mean(arrival_delay_in_minutes,
           na.rm = TRUE)
    ))
  )

glimpse(airline_satisfaction)
```

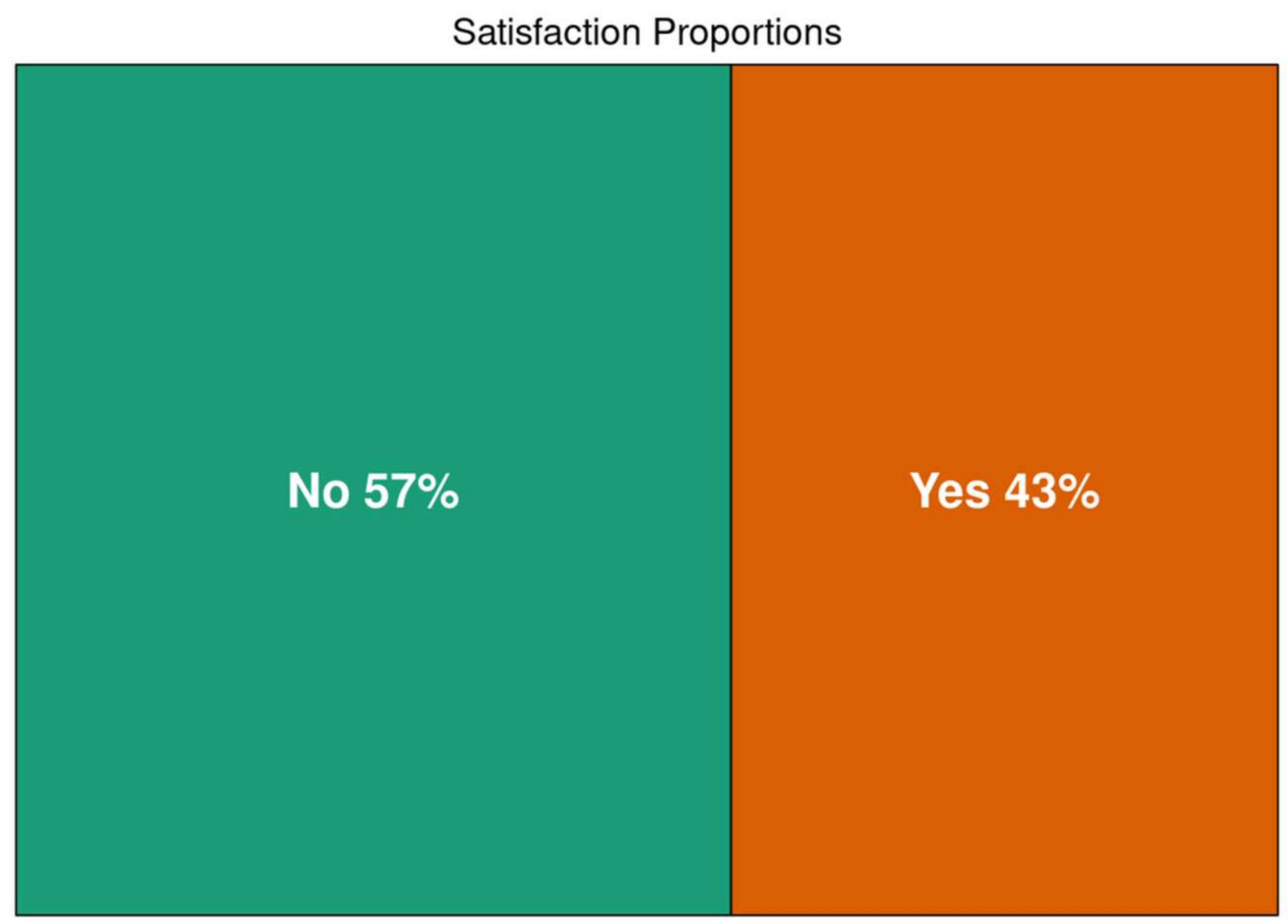
- **Replace the label with Yes and No**

neutral or dissatisfied to be "No"

- **satisfied to be "Yes"**
- **In addition to that, we have to ensure that the level of "yes" is 1 and "No" is 2, and then to replace the missing values in arrival_delay_in_minutes with the mean of itself, and convert the character variables to a factor for building a model.**

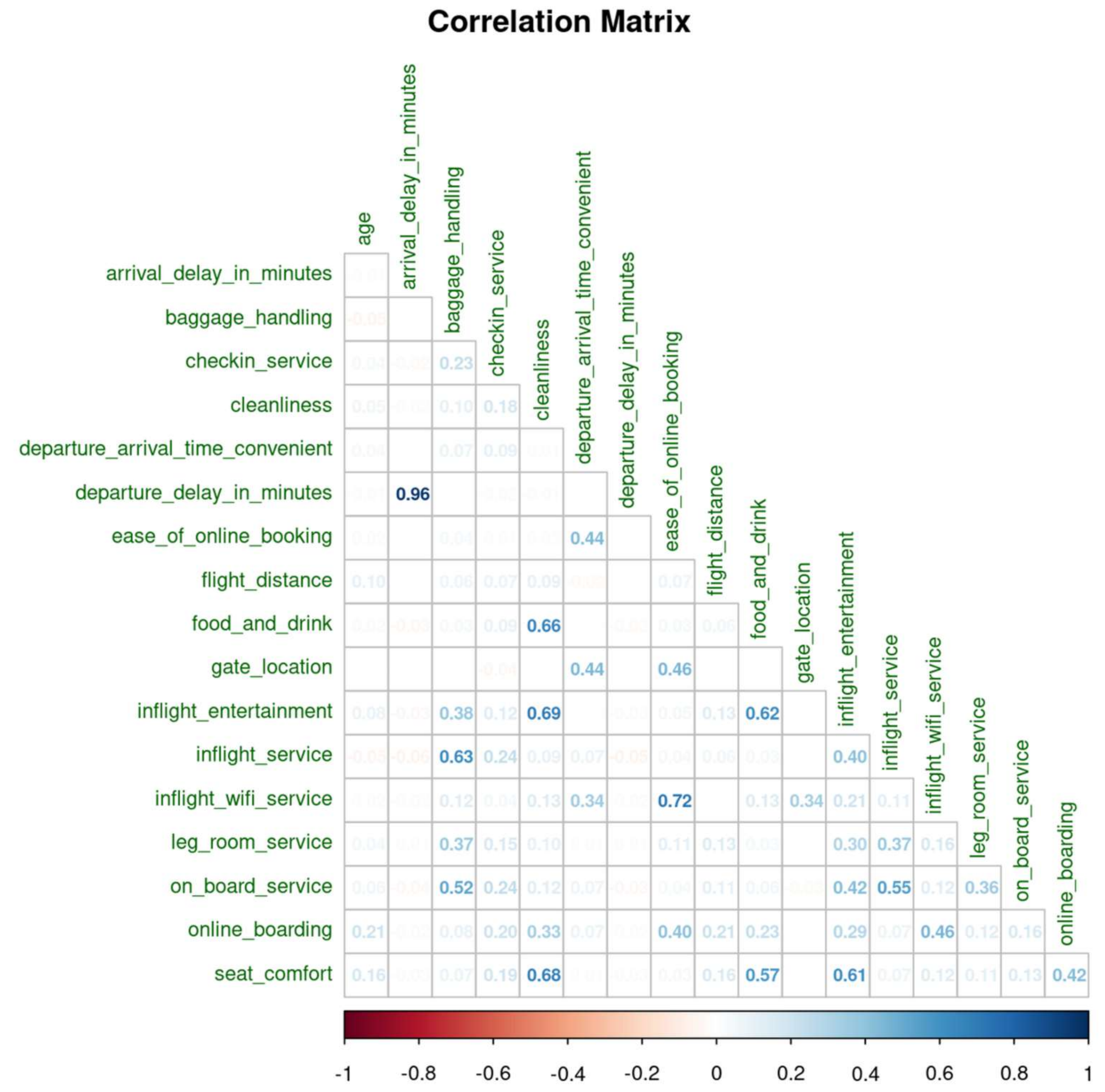
EXPLORATORY DATA ANALYSIS

- The chart shows that 57% of participants are not satisfied, while 43% are satisfied.
- This indicates a majority of dissatisfaction among the surveyed individuals.



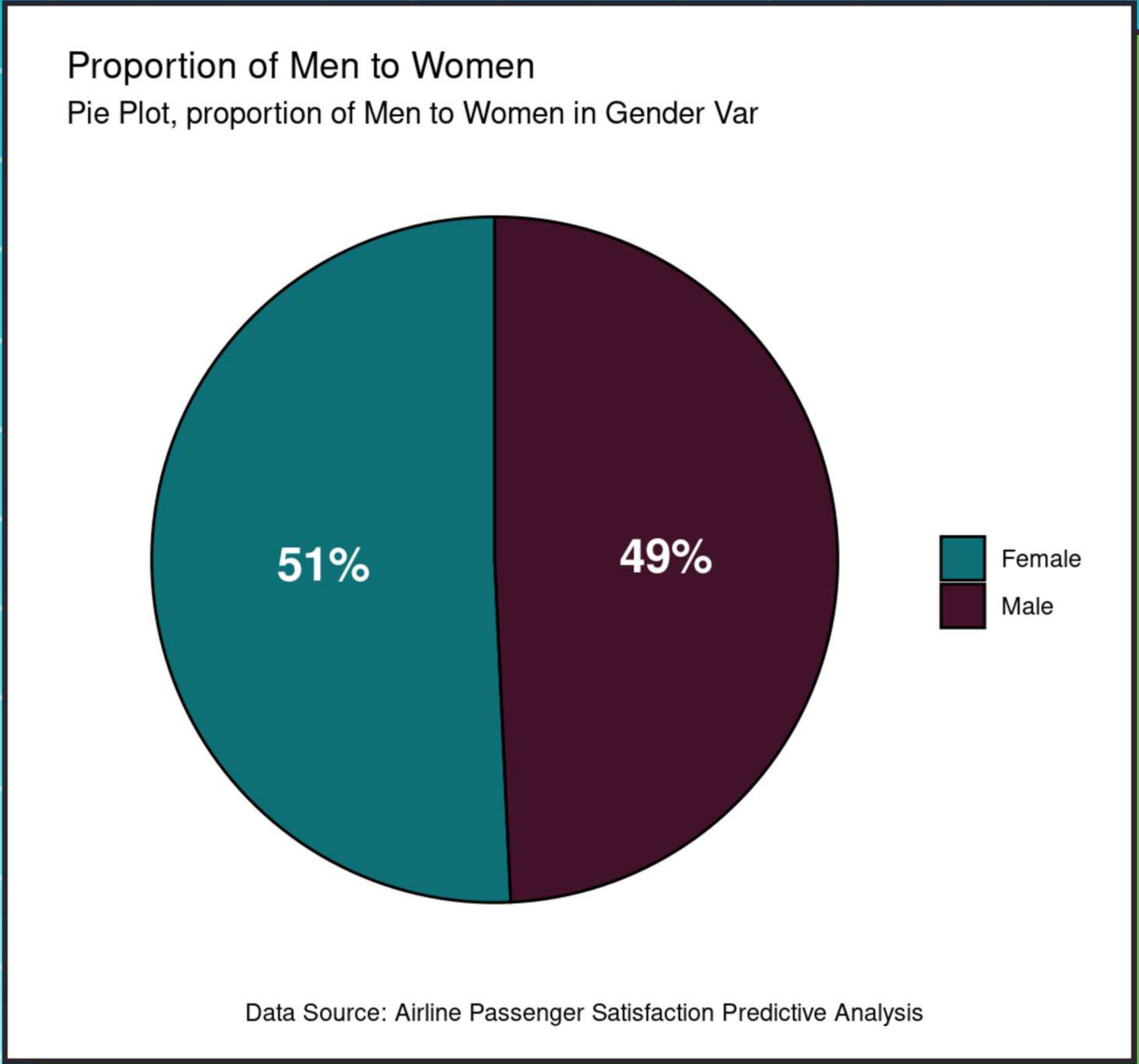
Correlation Matrix

- This matrix displays the correlations between different variables related to airline services.
- For example, a high positive correlation (0.96) exists between "arrival delay in minutes" and "departure delay in minutes," suggesting that delays are closely related.

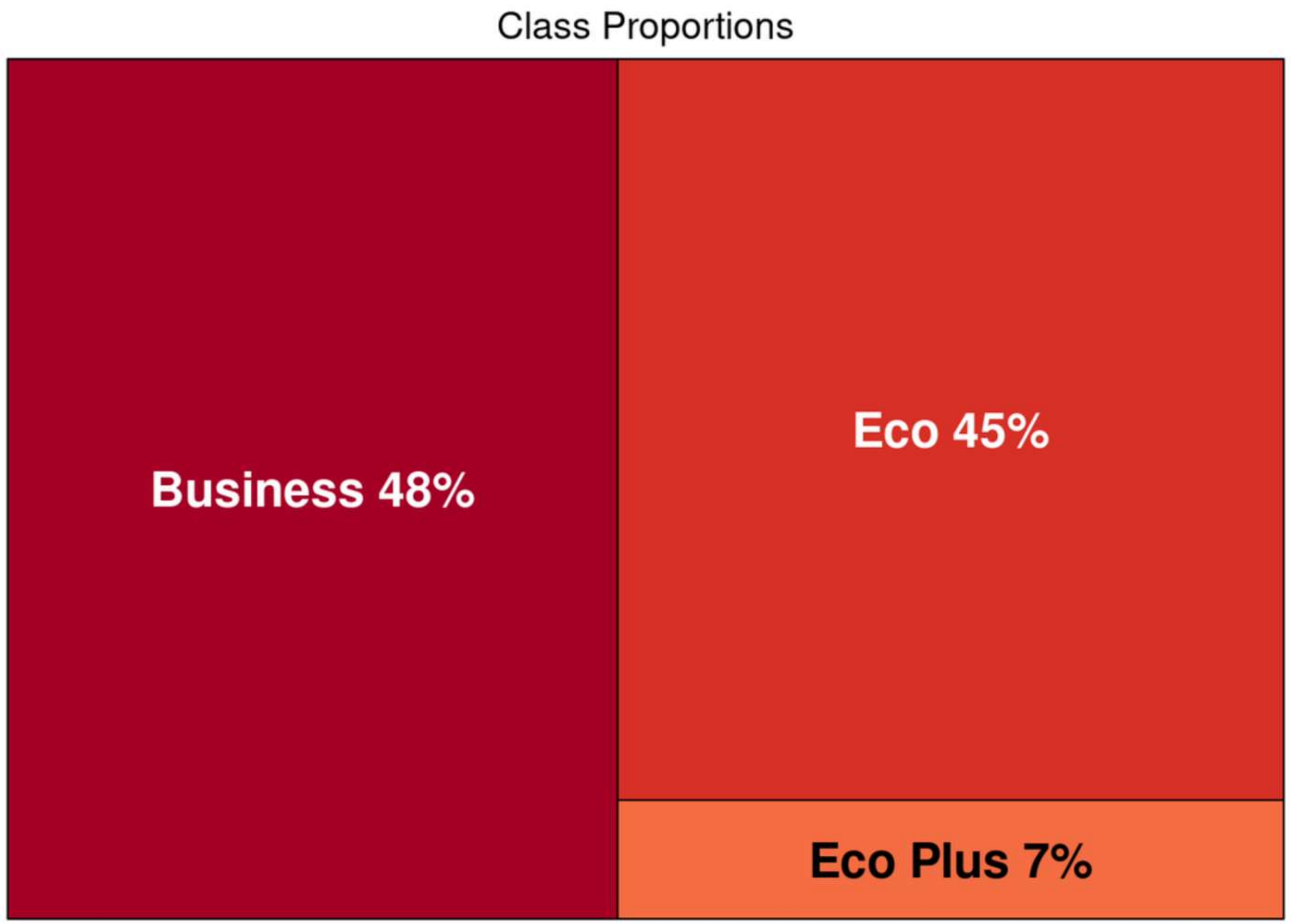


Pie Chart

- The pie chart reveals a nearly even gender distribution among the passengers, with 51% being male and 49% being female.

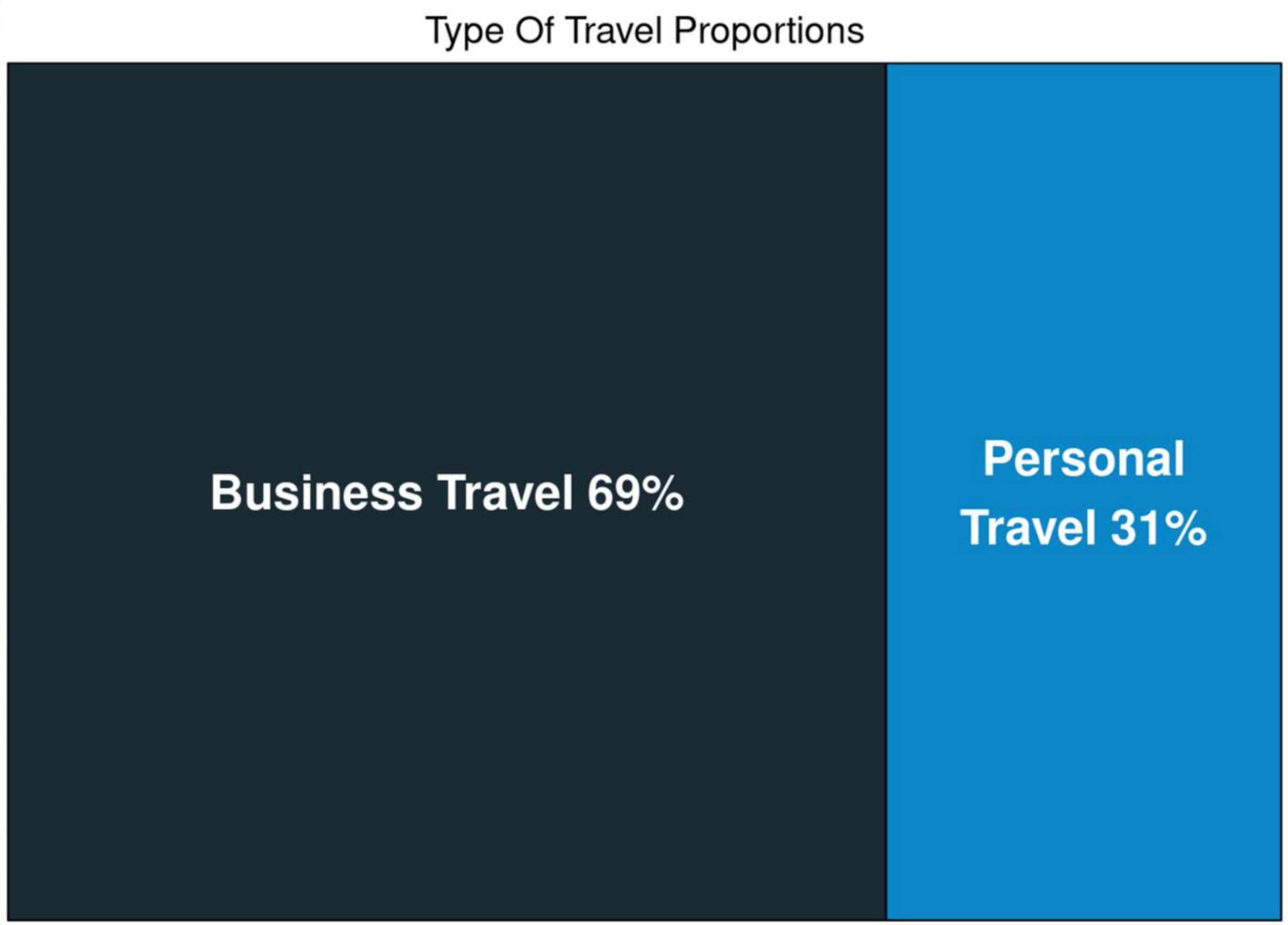


- **Business Class** has the highest proportion of travelers at 48% that is nearly half of the travelers flew in Business Class.
- **Economy Class** is close behind with 45% which shows a significant portion of travelers, slightly less than those in Business Class, flew in Economy.
- **Economy Plus** is the least popular, with only 7% of travelers choosing this class shows that a smaller number of travelers chose Economy Plus.

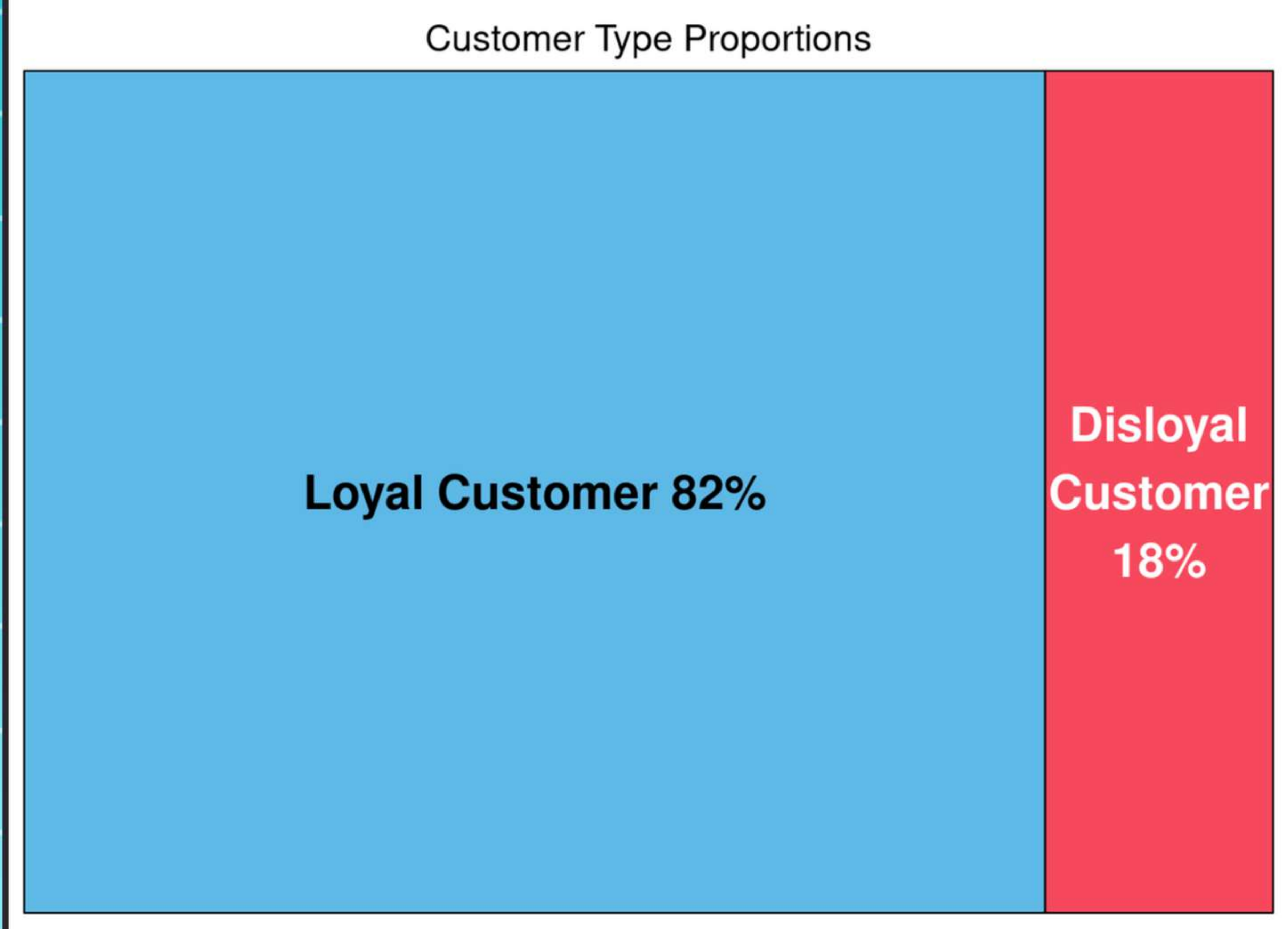


- A significant majority (69%) of the travel is for business purposes.

- Personal travel accounts for a smaller portion, with 31% of the total travel

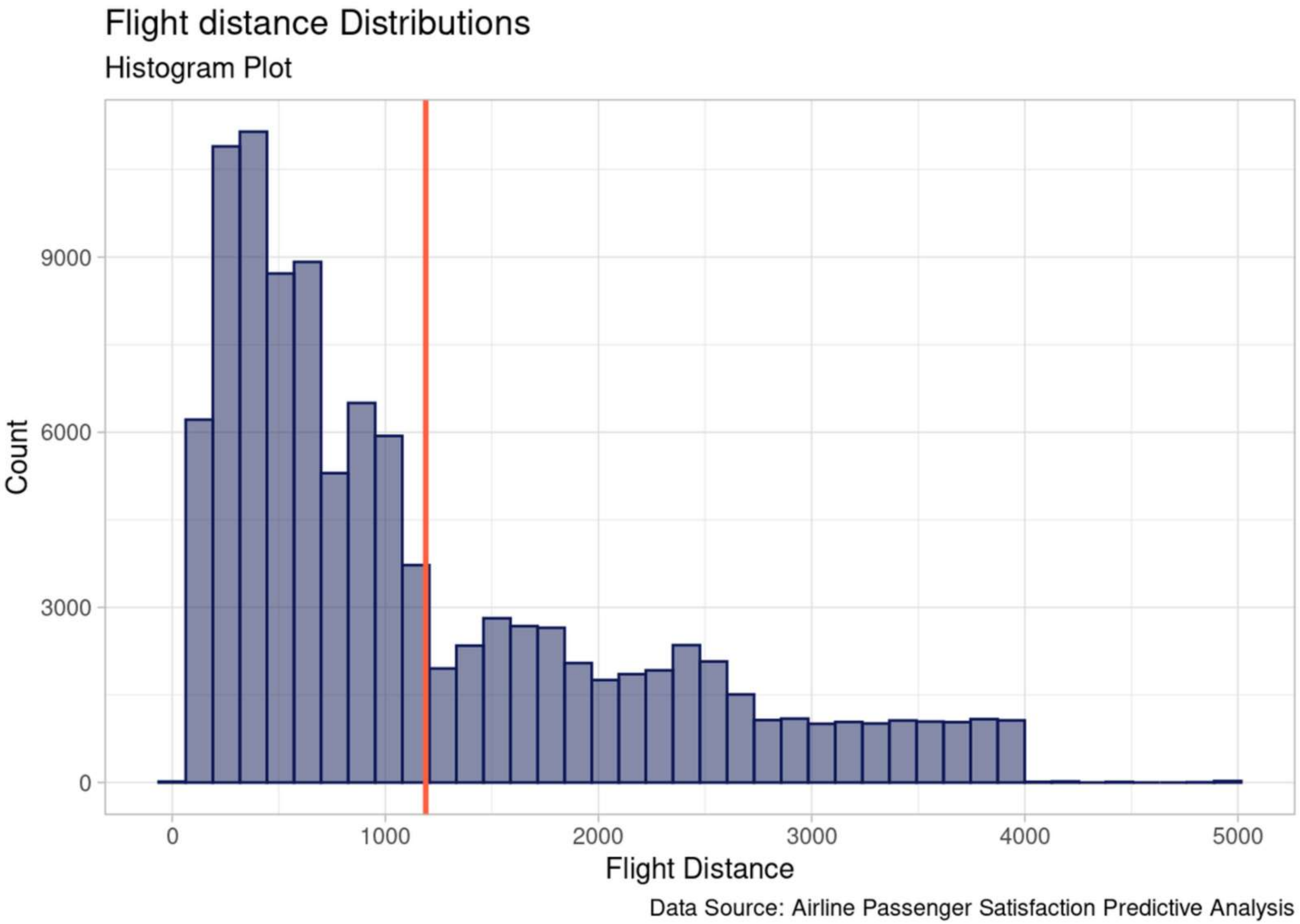


- A majority (82%) of the customers are loyal.
- Disloyal customer accounts for a smaller portion, with 18% of the total customer type proportions.



Histogram

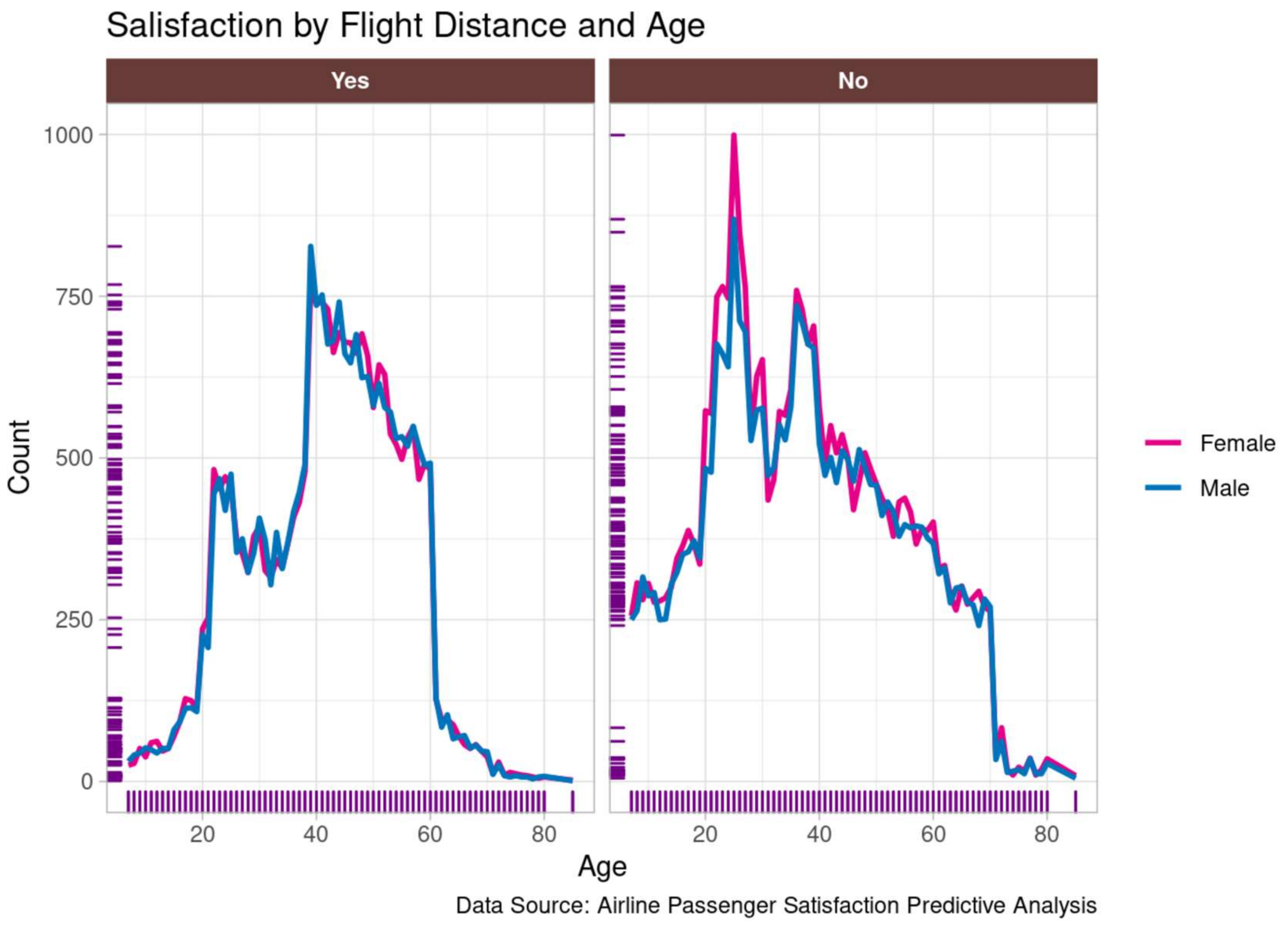
- The histogram is right-skewed, meaning most flights are on the shorter side, with fewer flights as the distance increases.
- The majority of flights seem to have distances less than 1000 units. There’s a noticeable peak between 500 and 1000, indicating that most flights in the dataset are short-haul flights.
- After 1000 units, the number of flights gradually decreases, with some flights ranging between 1000 to 4000 units.
- The red line is positioned at approximately 1000 units on the x-axis. It likely represents a threshold or a specific point of interest (e.g., distinguishing between short-haul and long-haul flights).
- In summary, the plot shows that most passengers in the dataset travel on shorter flights, with the frequency of flights declining as the distance increases.



Density Plot Combined With Rug Plot

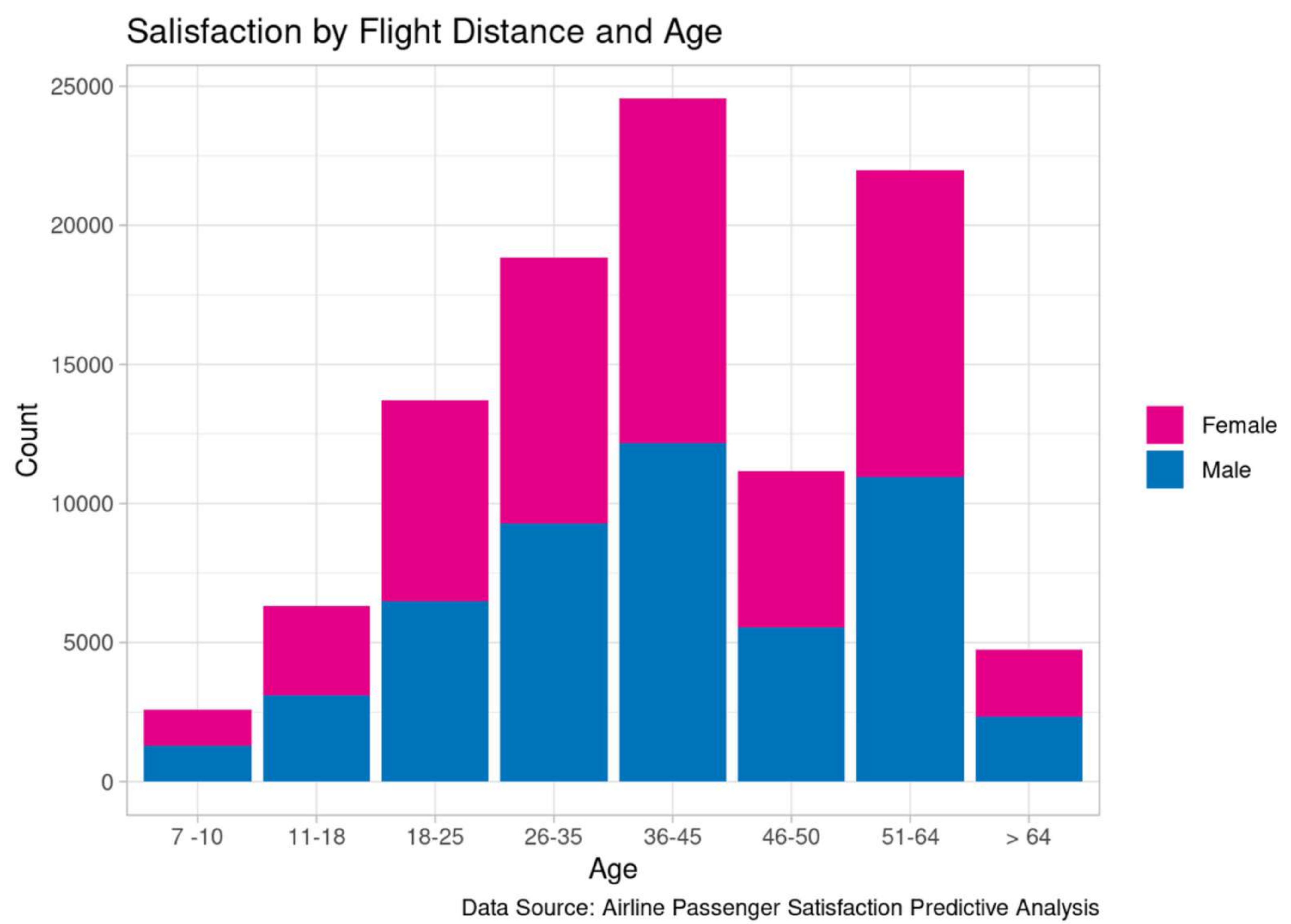
- The plot shows that passengers aged 20-40, especially males, are more likely to be dissatisfied, while satisfaction increases among those aged 40-60.

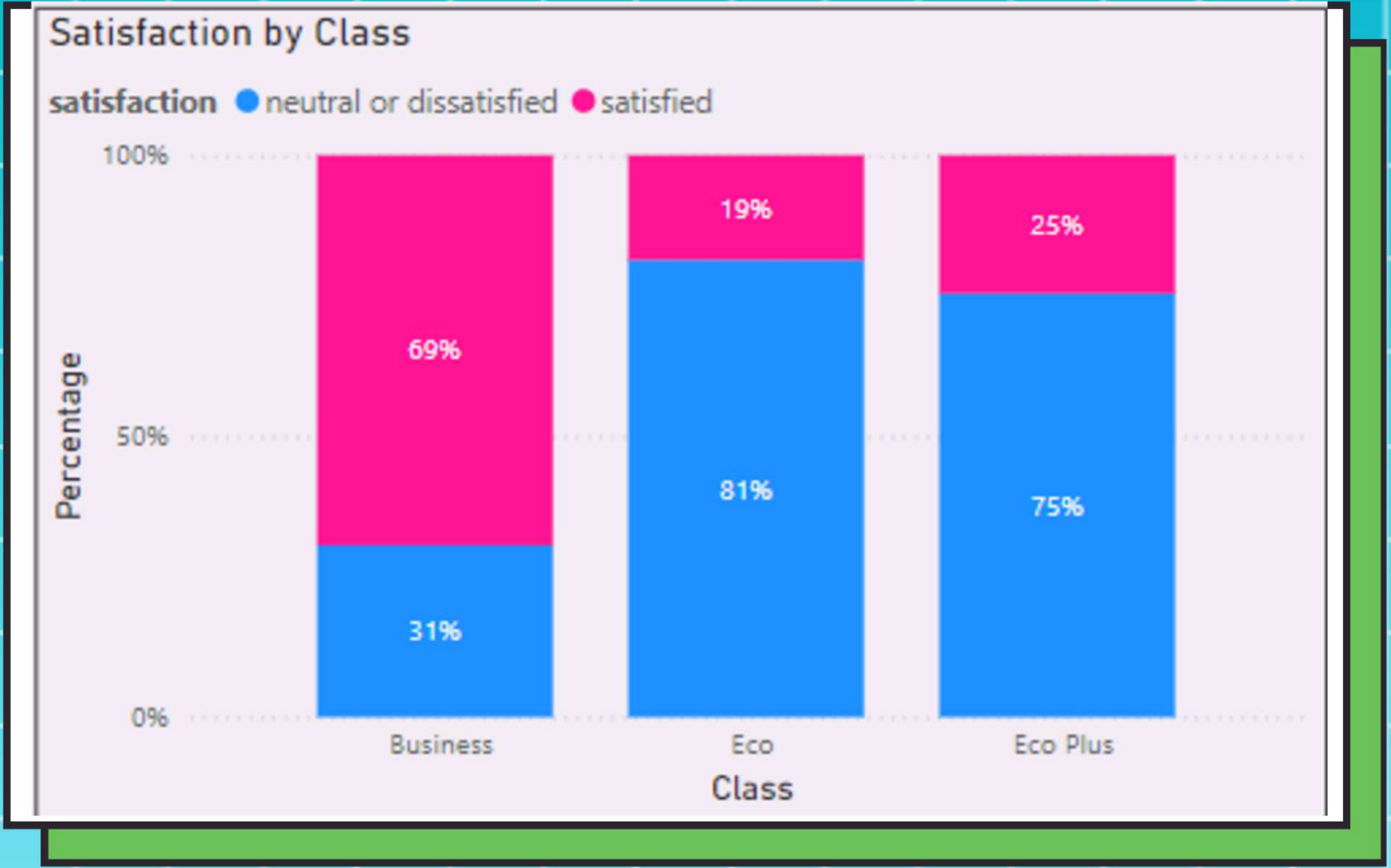
- Satisfaction declines after age 60, with similar trends for both genders.



Stacked Bar Plot

- The chart suggests that the majority of passengers in the dataset are between 26 and 64 years old, with the 36-45 age group being the most significant.
- Females slightly outnumber males in most age groups, especially in the higher-count categories (26-35, 36-45, 51-64).

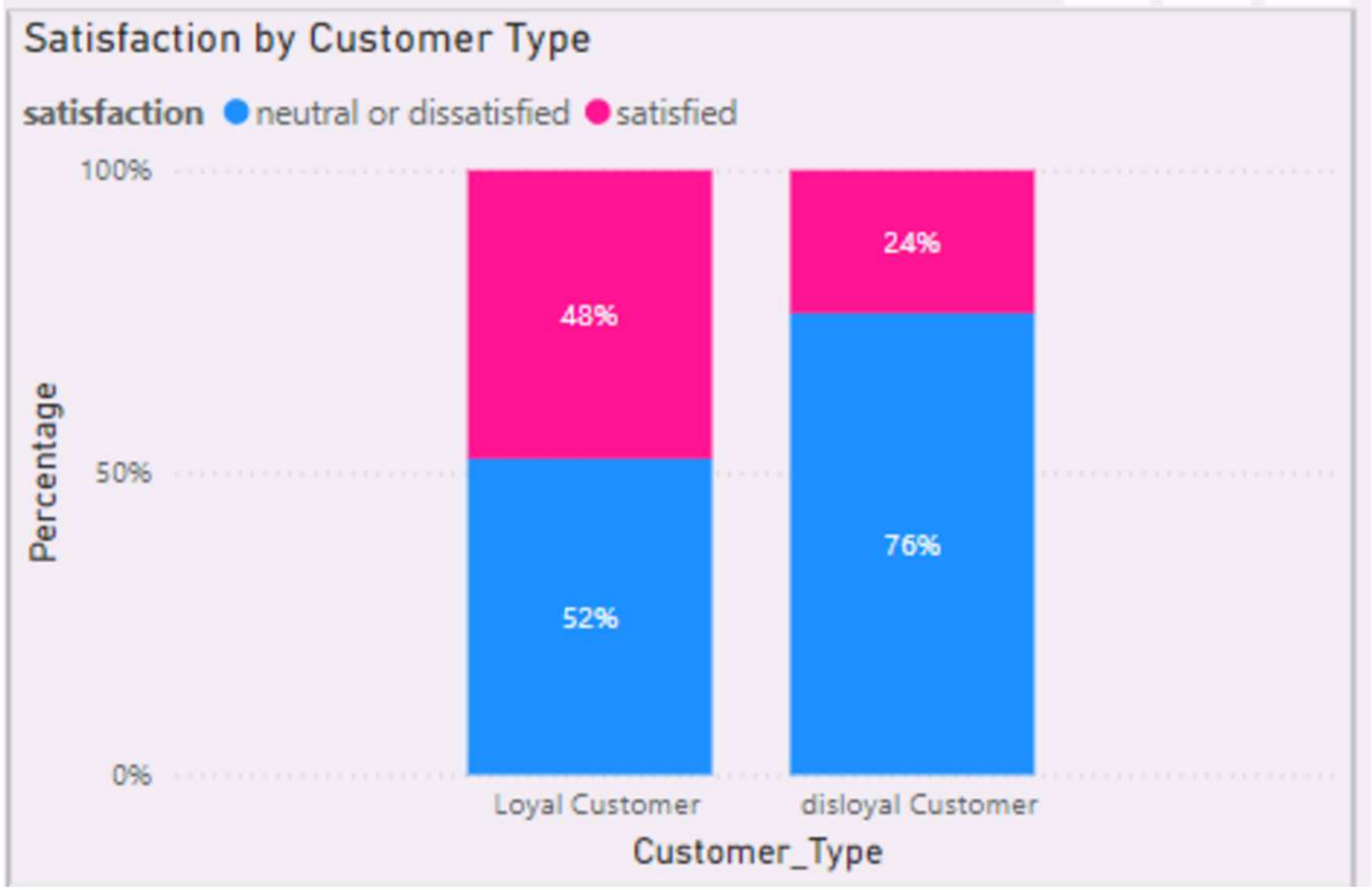




- **Business class:** The highest percentage of customers (69%) are satisfied with their experience in Business class. Only 31% are neutral or dissatisfied.
- **Eco class:** A lower percentage of customers (75%) are satisfied with their experience in Eco class. 25% are neutral or dissatisfied.
- **Eco Plus class:** The satisfaction level in Eco Plus class falls between Business and Eco class, with 81% of customers satisfied and 19% neutral or dissatisfied.
- Overall, the chart indicates that customer satisfaction is generally higher in Business class compared to Eco and Eco Plus classes.

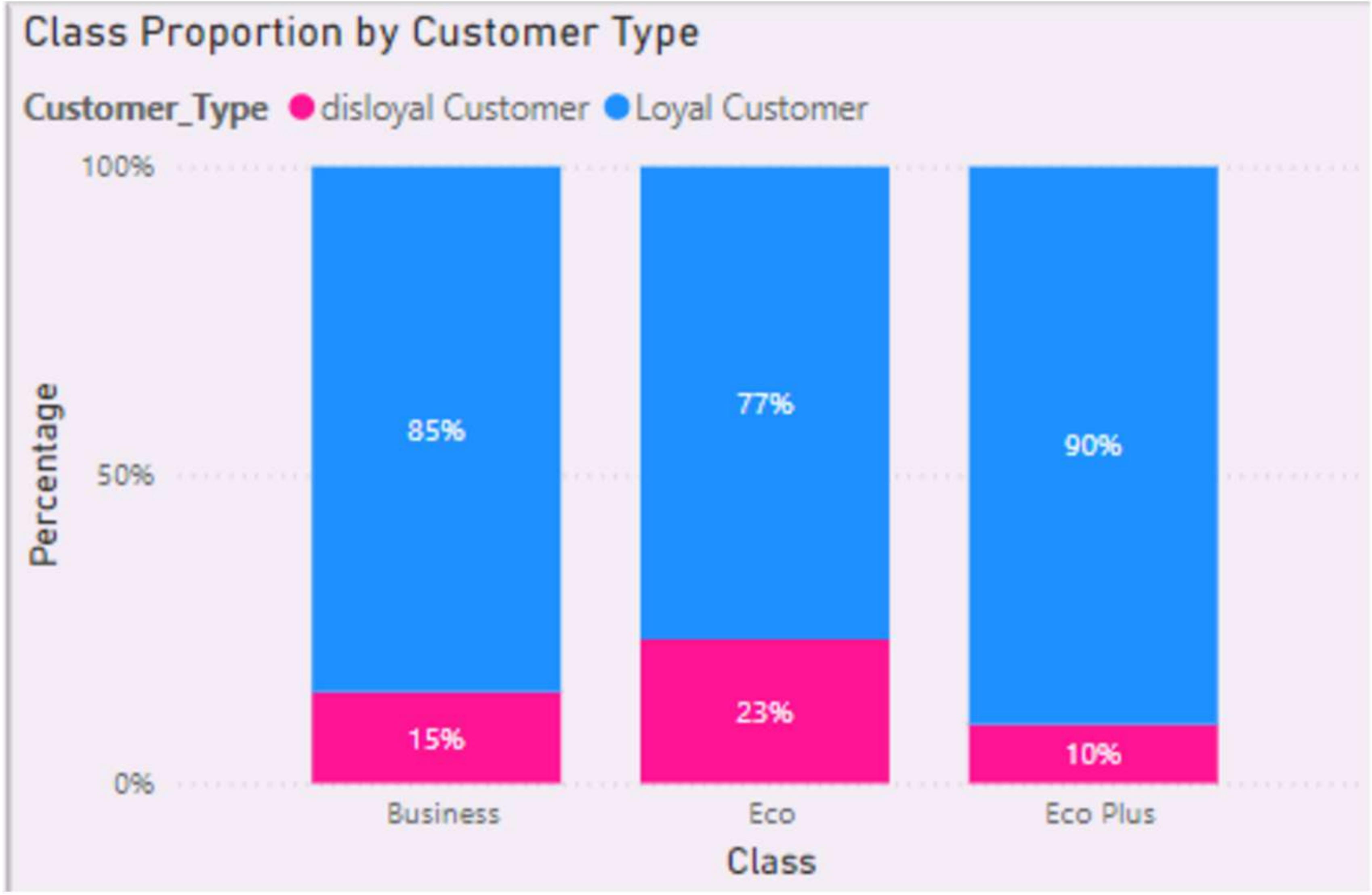
100% Stacked Bar Plot

- **Loyal Customer:** A higher percentage of Loyal Customers (52%) are not satisfied with their experience. 48% are satisfied.
- **Disloyal Customer:** A significantly lower percentage of Disloyal Customers (24%) are satisfied. 76% are neutral or dissatisfied.
- Overall, the chart indicates that customer satisfaction is significantly higher among Loyal Customers compared to Disloyal Customers.



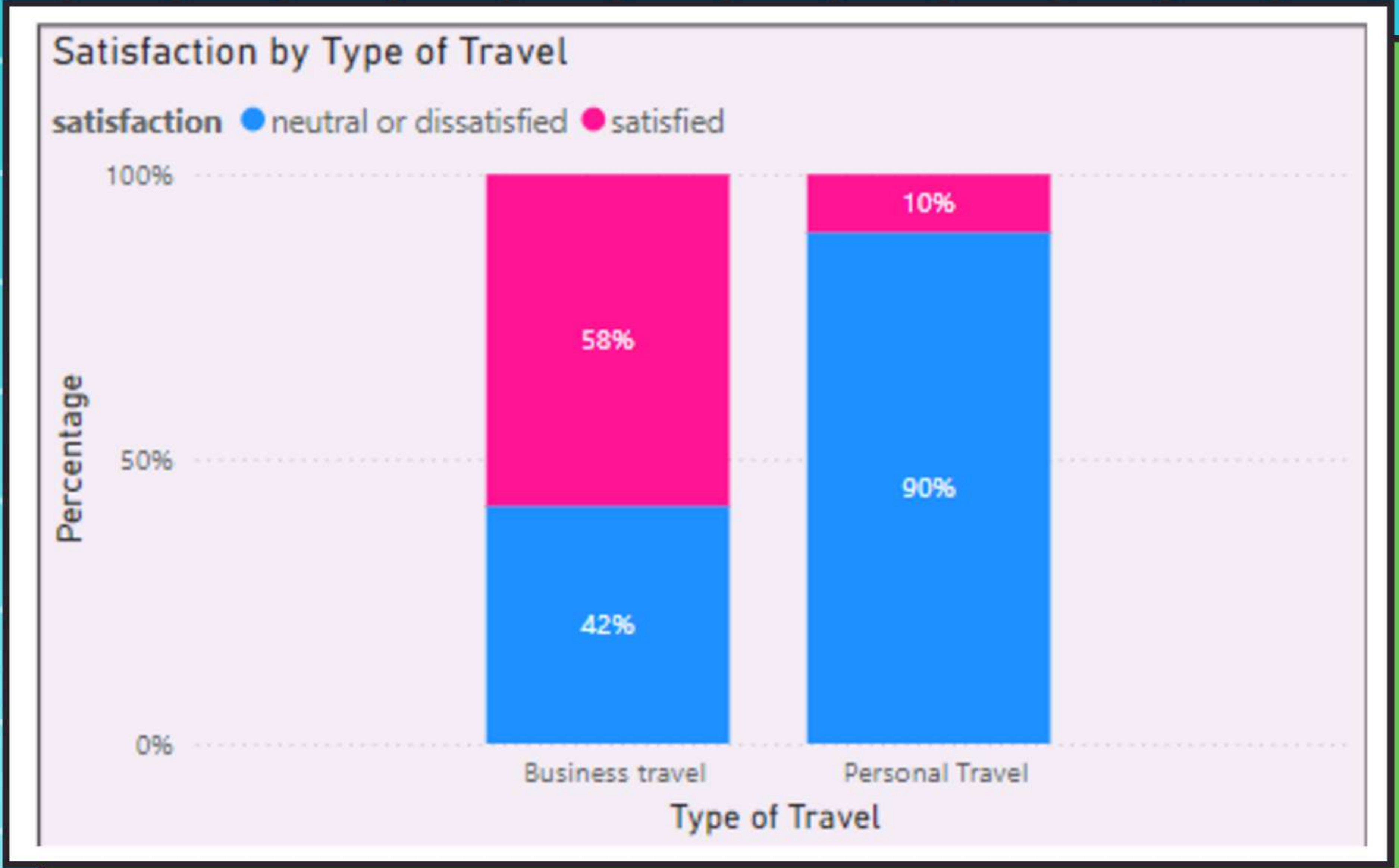
100% Stacked Bar Plot

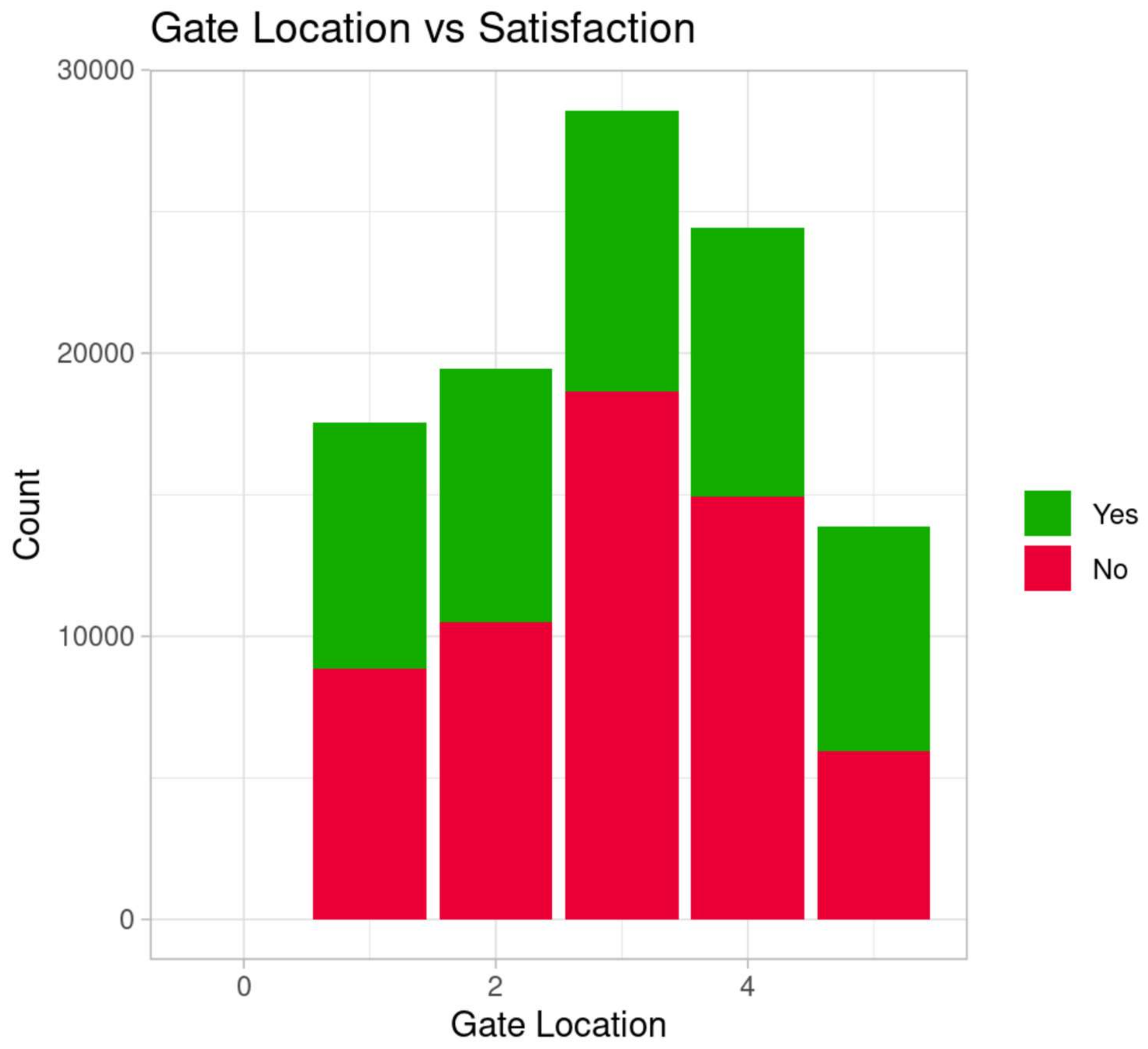
- The chart shows that loyal customers dominate across all classes, with Eco Plus having the highest loyalty at 90%, followed by Business at 85%, and Eco at 77%.
- The Eco class has the highest proportion of disloyal customers at 23%, indicating a potential area for improvement in customer satisfaction.



100% Stacked Bar Plot

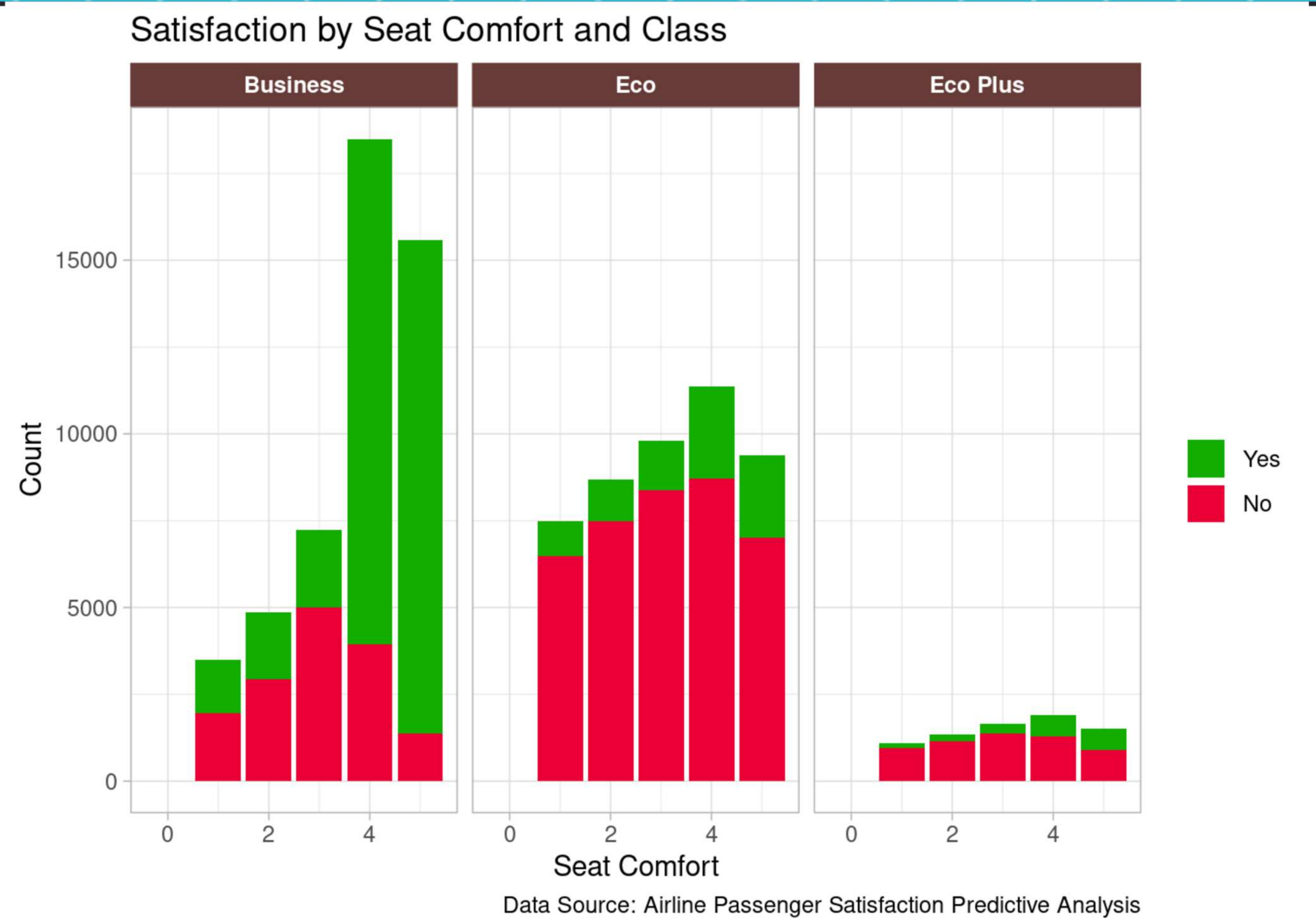
- **Business travel:** A lower percentage of customers (42%) are satisfied with their business travel experience. 58% are neutral or dissatisfied.
- **Personal travel:** A much higher percentage of customers (90%) are unsatisfied with their personal travel experience. Only 10% are satisfied
- Overall, the chart indicates that customer satisfaction is significantly higher for business travel compared to personal travel.

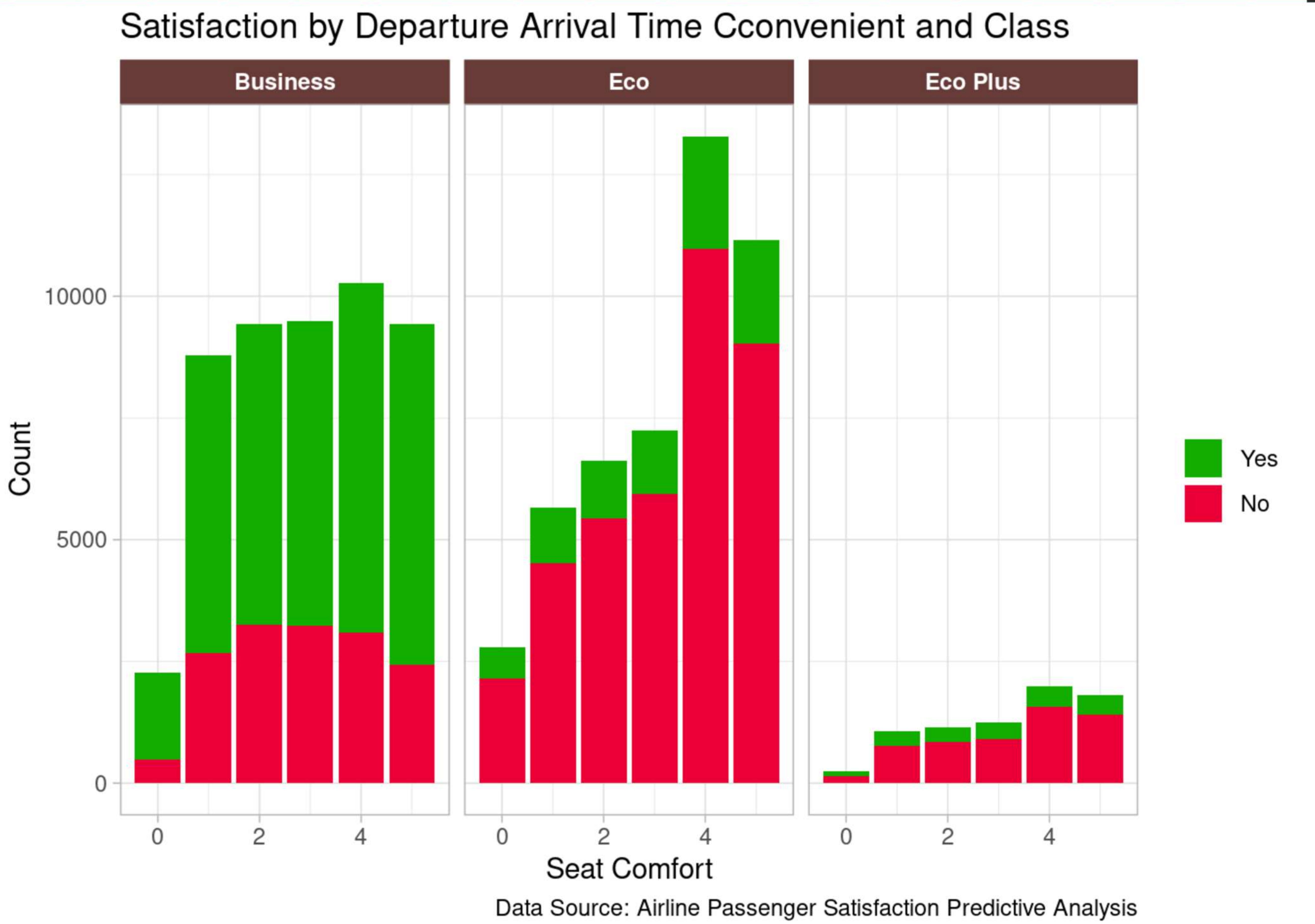




Data Source: Airline Passenger Satisfaction Predictive Analysis

	gate_location	satisfaction	n
1	0	Yes	1
2	1	Yes	8703
3	1	No	8859
4	2	Yes	8965
5	2	No	10494
6	3	Yes	9922
7	3	No	18655
8	4	Yes	9490
9	4	No	14936
10	5	Yes	7944
11	5	No	5935


Satisfaction by Seat Comfort and Class
Business
Eco
Eco Plus
Count
15000
10000
5000
0
0
2
4
Seat Comfort
Yes
No
Data Source: Airline Passenger Satisfaction Predictive Analysis



Satisfaction by Departure Arrival Time Cconvenient and Class
Business
Eco
Eco Plus
Count
10000
5000
0
0
2
4
Seat Comfort
Yes
No
Data Source: Airline Passenger Satisfaction Predictive Analysis

Data Preparation For Modeling

Eliminating id and sr variables

```r
airline_satisfaction_for_model <- airline_satisfaction %>%
  select(-c(
    id,
    sr
  ))
```

Splitting The Dataset & Generating 10 folds

```r
set.seed(31967)

airline_satisfaction_split <- initial_split( airline_satisfaction_for_model, prop = 3/4, strata = satisfaction)

train_data <- training(airline_satisfaction_split)
test_data  <- testing(airline_satisfaction_split)
```

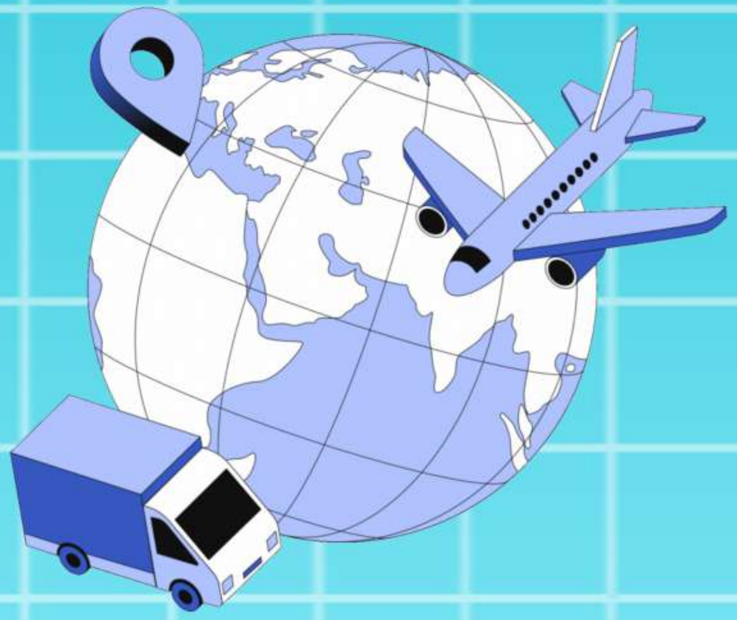
Viewing the proportion of both instances under training and testing datasets

```r
train_data %>%
  count(satisfaction) %>%
  mutate(prop = n/sum(n))
```

```
## # A tibble: 2 × 3
##   satisfaction     n  prop
##   <fct>        <int> <dbl>
## 1 Yes          33768 0.433
## 2 No           44159 0.567
```

```r
test_data  %>%
  count(satisfaction) %>%
  mutate(prop = n/sum(n))
```

```
## # A tibble: 2 × 3
##   satisfaction     n  prop
##   <fct>        <int> <dbl>
## 1 Yes          11257 0.433
## 2 No           14720 0.567
```



```r
set.seed(31967)
fold_cv <- vfold_cv(train_data, times = 10, apparent = TRUE)
```

Setting the seed ensures reproducibility and fold_cv performs 10-fold cross-validation on the dataset.

Logistic Regression

Building the model

```
logis_mod <-
  logistic_reg(penalty = tune(), mixture = 1) %>%
  set_engine("glmnet")
```

- We Define a model with a penalty parameter (penalty = tune()) to be tuned later.
- The mixture = 1 indicates that this is a Lasso regression (L1 regularization). set_engine("glmnet") specifies that the model will be trained using the glmnet package, which is commonly used for fitting generalized linear models.

Creating The Recipe

```
logis_recipe <-
  recipe(satisfaction ~ ., data = train_data) %>%
  step_dummy(all_nominal_predictors(), -all_outcomes()) %>%
  step_zv(all_numeric()) %>%
  step_normalize(all_numeric()) %>%
  step_corr(all_numeric_predictors(), threshold = .7)
```

- **By specifying a recipe, we ensure that the same preprocessing steps are applied to both training and testing data, which is crucial for model consistency. Removing highly correlated features helps prevent multicollinearity, ensuring that each predictor adds unique information to the model.**

Creating the workflow

```r
logis_workflow <-
  workflow() %>%
  add_model(logis_mod) %>%
  add_recipe(logis_recipe)
```

- A workflow object bundles together all preprocessing steps and the model itself, making the modeling process more streamlined and less error-prone. It ensures that the preprocessing steps are applied consistently whenever the model is trained or used for predictions.

Grid for Hyperparameter tuning

```
logis_reg_grid <- tibble(penalty = 10^seq(-3, -1, length.out = 60))
logis_reg_grid %>% top_n(-5)
```

```
## # A tibble: 5 × 1
##   penalty
##     <dbl>
## 1 0.001
## 2 0.00108
## 3 0.00117
## 4 0.00126
## 5 0.00137
```

```
logis_reg_grid %>% top_n(5)
```

```
## # A tibble: 5 × 1
##   penalty
##     <dbl>
## 1  0.0732
## 2  0.0791
## 3  0.0855
## 4  0.0925
## 5  0.1
```

- This creates a tibble (data frame) containing a sequence of values for the penalty parameter (from 10^-3 to 10^-1) to be used for hyperparameter tuning. Testing a range of penalty values helps in finding the optimal amount of regularization for the model.

- A well-chosen penalty can improve model performance by balancing the trade-off between bias and variance.

Training and Tuning

```
set.seed(31967)
logis_res <-
  logis_workflow %>%
  tune_grid(fold_cv,
            grid = logis_reg_grid,
            control = control_grid(save_pred = TRUE),
            metrics = metric_set(roc_auc))
```

- Hyperparameter tuning with cross-validation helps identify the model configuration that provides the best performance. The ROC AUC metric is chosen because it measures the ability of the model to discriminate between classes

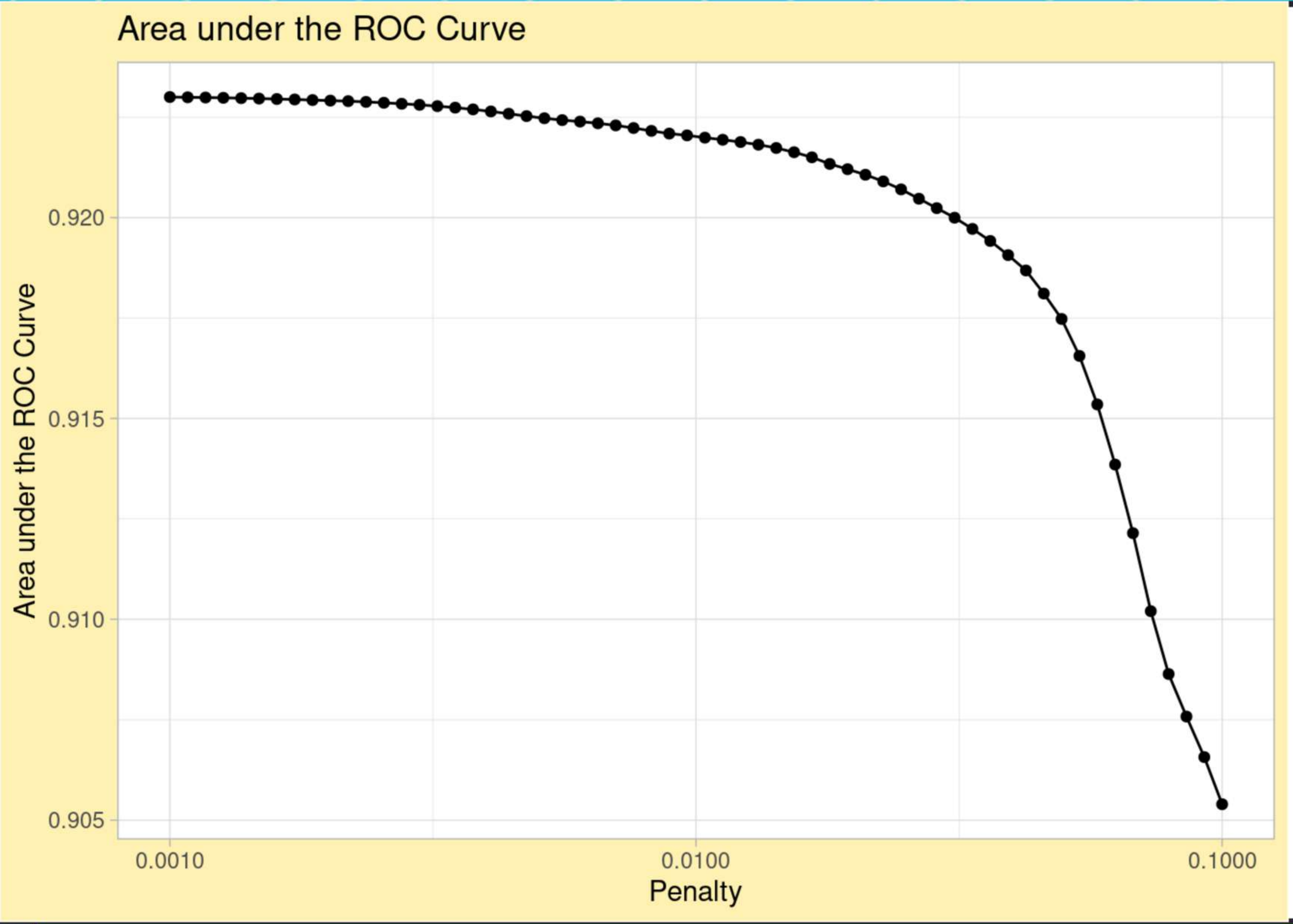
Plotting the Area Under the ROC Curve vs. Penalty Value

```r
logis_plot <-
  logis_res %>%
  collect_metrics() %>%
  ggplot(aes(x = penalty, y = mean)) +
  geom_point() +
  geom_line() +
  scale_x_log10(labels = scales::label_number()) +
   theme(
        plot.background = element_rect(fill = "#FFF1B2")) +
  labs(
    title = "Area under the ROC Curve",
    x= "Penalty",
    y = "Area under the ROC Curve"

  )

logis_plot
```

- **Plotting the performance metric against the penalty values helps to visually identify the penalty value that maximizes the model performance.**

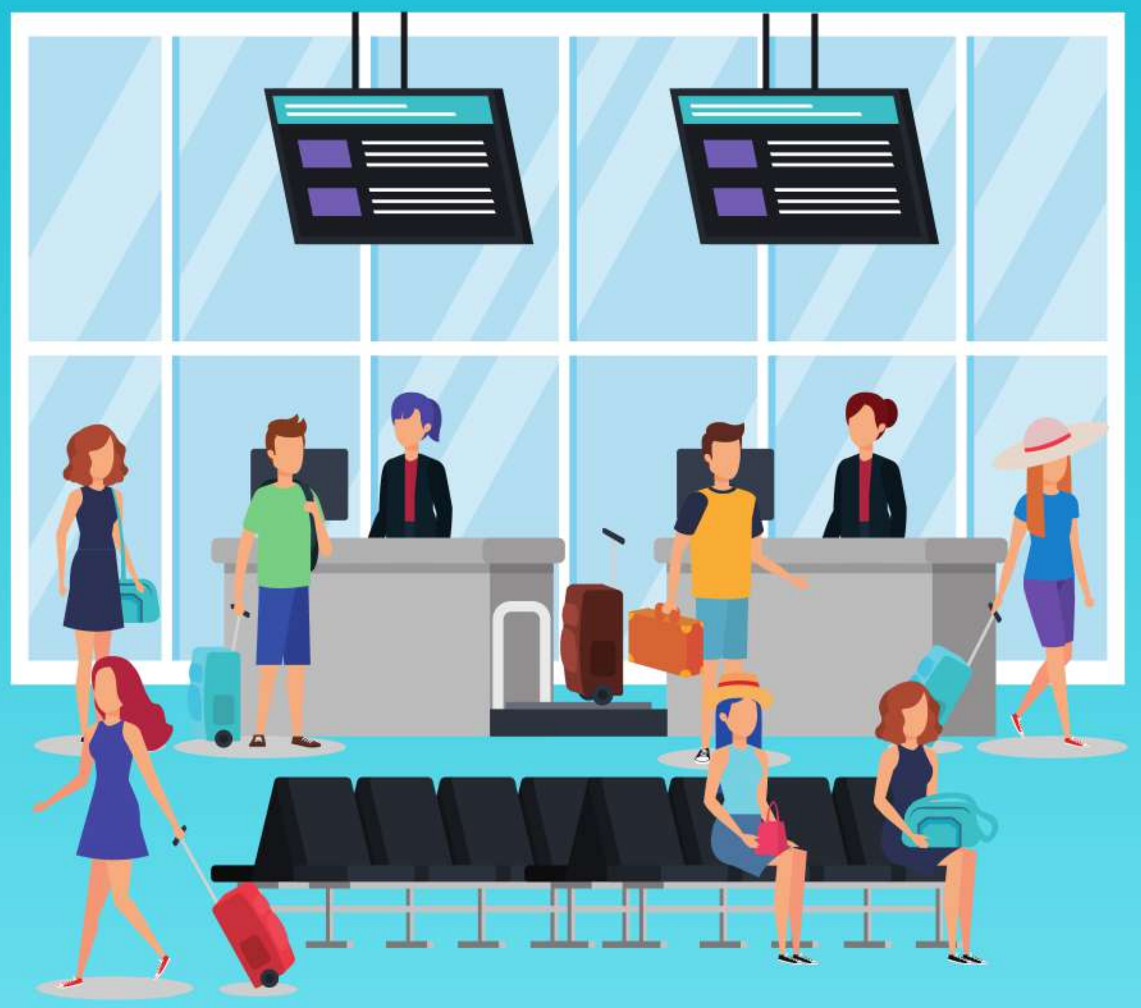
ROC CURVE

- Shows the relationship between the penalty parameter and the Area Under the ROC Curve (AUC).
- As the penalty increases, the AUC gradually decreases, indicating a drop in model performance.
- Initially, the AUC remains relatively stable, but it starts declining sharply as the penalty becomes larger, suggesting that over-penalizing the model harms its ability to discriminate between classes effectively.



Setting the best model

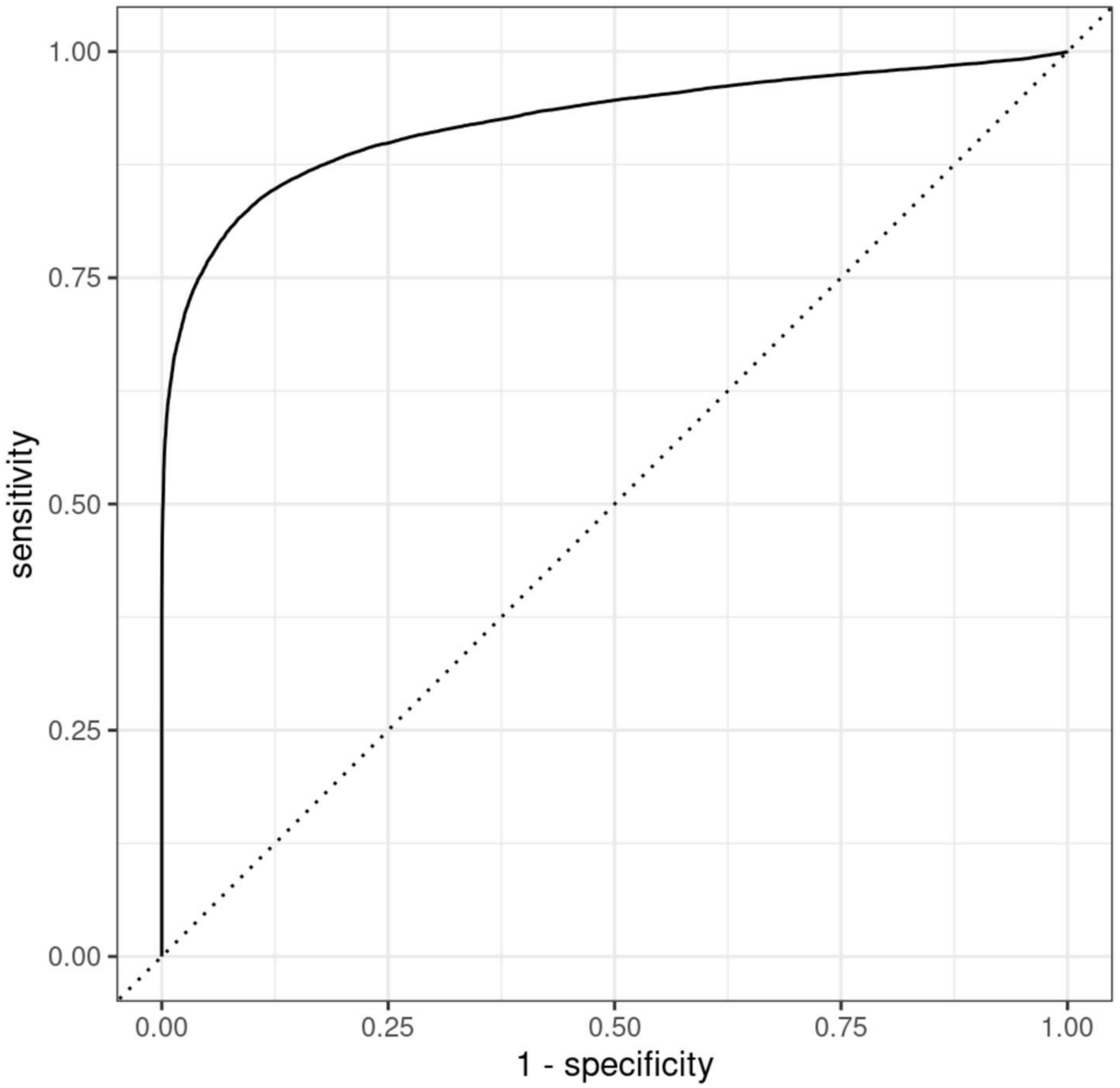
```r
logis_best <-
  logis_res %>%
  show_best() %>%
  arrange(desc(mean)) %>%
  dplyr::slice(2)

logis_best
```



```
## # A tibble: 1 × 7
##   penalty .metric .estimator  mean     n std_err .config
##     <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.00108 roc_auc binary     0.923    10 0.00124 Preprocessor1_Model02
```

ROC Curve

- It illustrates the performance of a binary classifier.
- The curve plots sensitivity (true positive rate) against 1-specificity (false positive rate).
- The closer the curve is to the top left corner, the better the model's performance, as it indicates high sensitivity and specificity.
- The dotted diagonal line represents a model with no discriminatory power (random guessing).



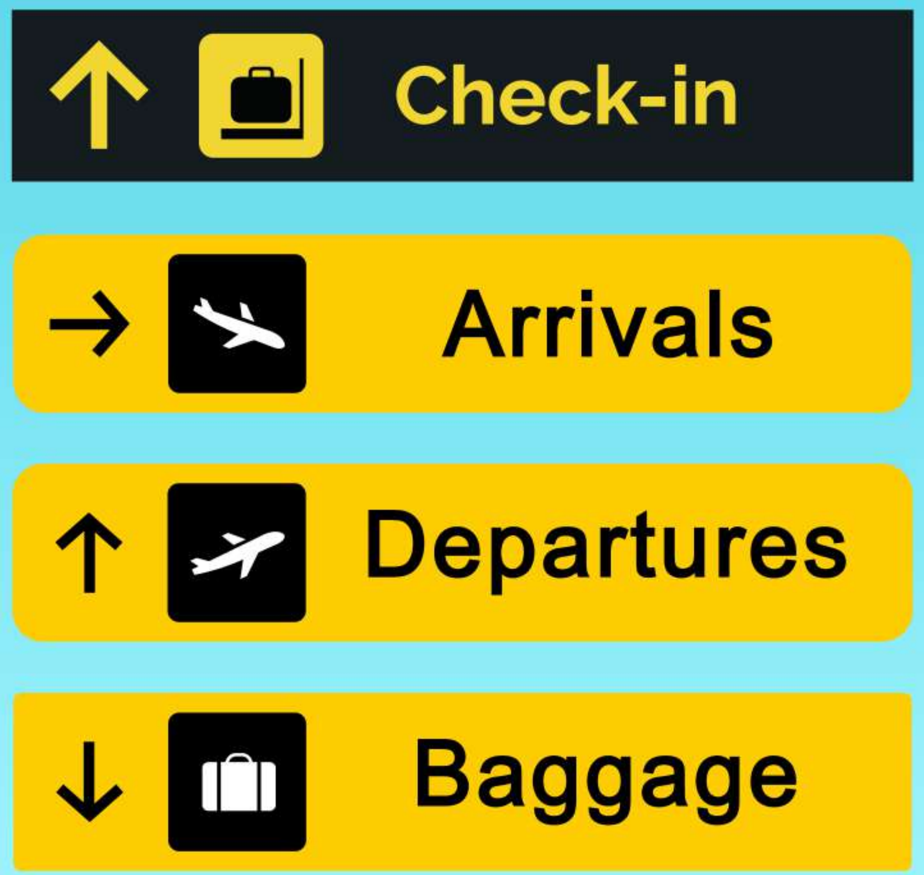
Final Fit

- The code finalizes the workflow with the best model parameters and fits it on the full training data.

- The model is then evaluated on the test set to assess its final performance.

```r
set.seed(31967)
final_logis_res <-
  logis_workflow %>%
  finalize_workflow(logis_best) %>%
  last_fit(airline_satisfaction_split)

final_logis_res
```

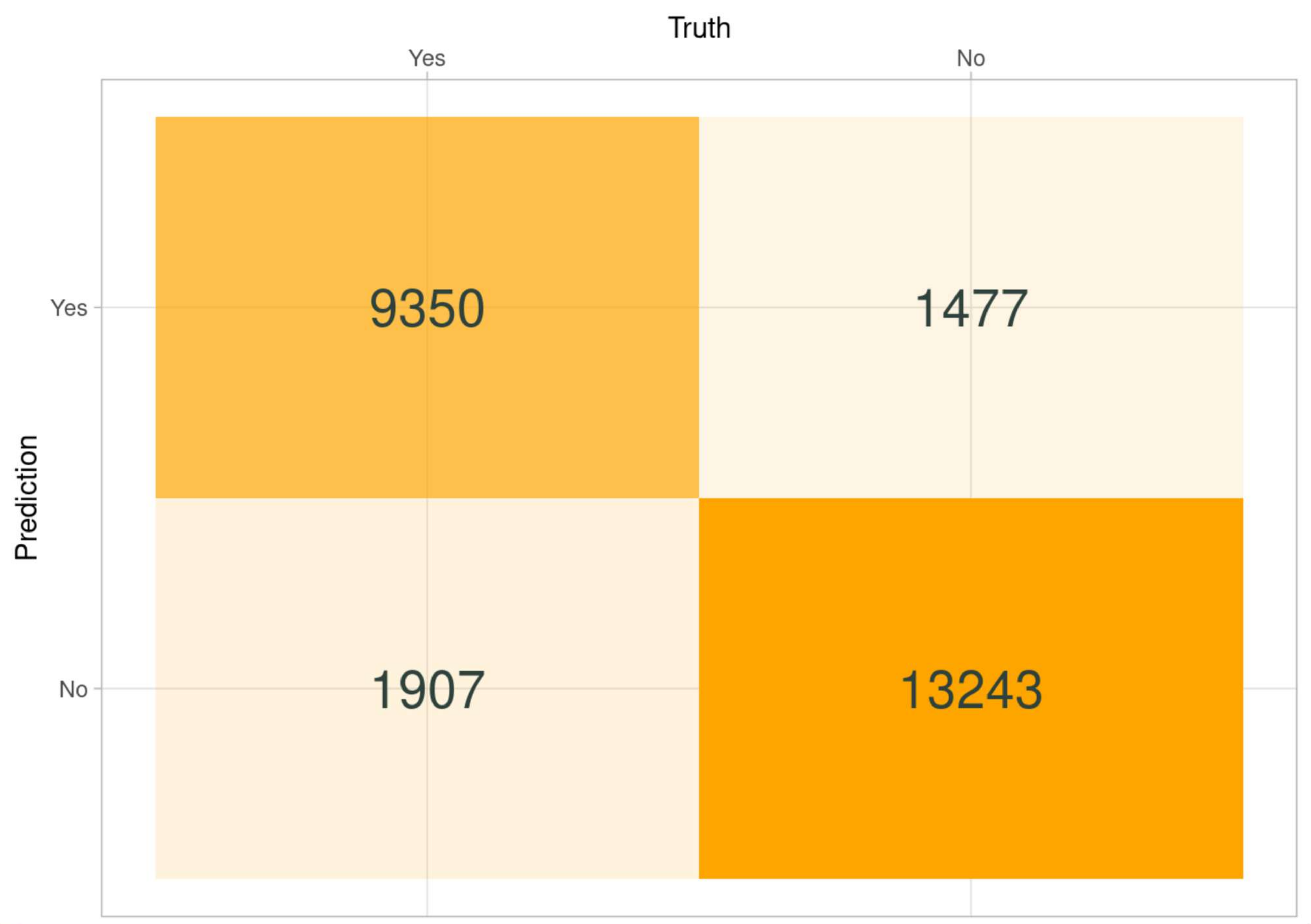
```
## # Resampling results
## # Manual resampling
## # A tibble: 1 × 6
##   splits                id                .metrics  .notes .predictions .workflow
##   <list>                <chr>             <list>    <list> <list>       <list>
## 1 <split [77927/25977]> train/test split <tibble … <tibb… <tibble [25… <workflo…
```



CONFUSION MATRIX

- This image is a confusion matrix, commonly used in evaluating the performance of a classification model. The matrix shows the following:

- **True Positives (Yes-Yes)**: 9350
- **False Positives (Yes-No)**: 1477
- **False Negatives (No-Yes)**: 1907
- **True Negatives (No-No)**: 13243
- The model has a higher number of true positives and true negatives, indicating it performs well, but there are still some misclassifications, as shown by the false positives and false negatives.



Results

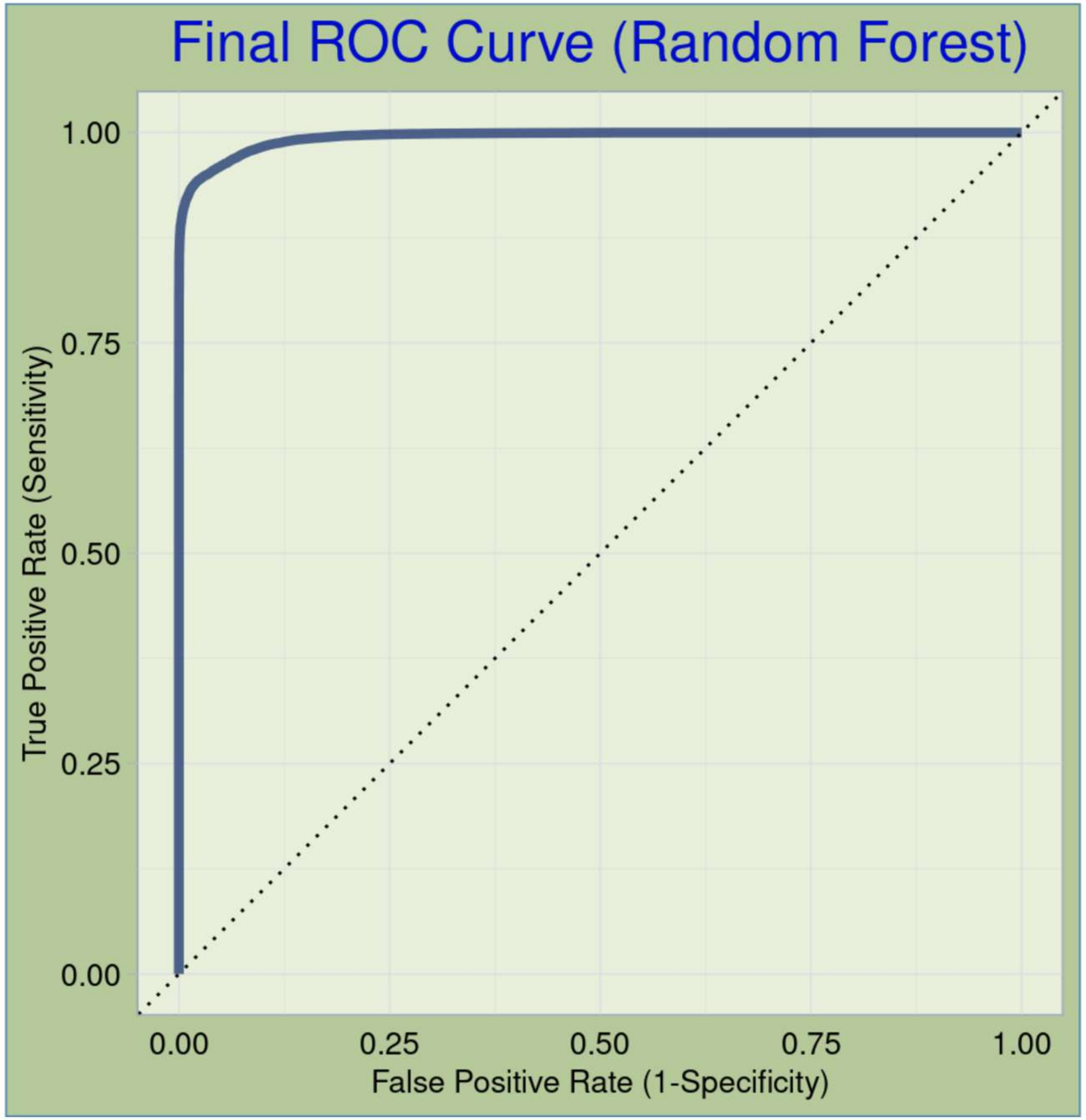
```
collect_metrics(final_logis_res)
```

```
## # A tibble: 2 × 4
##   .metric  .estimator .estimate .config
##   <chr>    <chr>          <dbl> <chr>
## 1 accuracy binary         0.870 Preprocessor1_Model1
## 2 roc_auc  binary         0.923 Preprocessor1_Model1
```

Accuracy is 87% and auc is 92%

ROC Curve

- **Curve Shape:** The ROC curve ideally should be as close to the top-left corner as possible. This indicates that the model can accurately predict both positive and negative cases with high sensitivity and specificity.
- **Diagonal Line:** The diagonal line represents a random classifier. A model that performs worse than random would have a curve below this line.
- **Area Under the Curve (AUC):** The AUC measures the overall performance of the model. A higher AUC indicates better performance.
- In this case, the ROC curve for the Random Forest model appears to be close to the top-left corner, suggesting good performance in terms of both sensitivity and specificity.



Random Forest Model

Preprocessing

```
rf_rec <-
  recipe(satisfaction ~ ., data = train_data) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_numeric()) %>%
  step_normalize(all_numeric())
```

The random forest model requires a different preprocessing recipe that suits its nature, as it handles correlated features and scaling differently compared to logistic regression.

Defining the Random Forest Model

```r
rf_mod <-
  rand_forest(mtry = tune(),
              min_n = tune(),
              trees = tune()
              ) %>%
  set_engine("ranger",  importance = "impurity") %>%
  set_mode("classification")
```

Tuning mtry (number of predictors sampled for splitting), min_n (minimum number of data points in a node), and trees (number of trees) allows finding the optimal random forest configuration for the data.

Creating and Tuning the Random Forest Workflow

```r
rf_workflow <-
  workflow() %>%
  add_model(rf_mod) %>%
  add_recipe(rf_rec)
```

```r
set.seed(31967)
rf_res <-
  rf_workflow %>%
  tune_grid(fold_cv,
            grid = 10,
            control = control_grid(save_pred = TRUE),
            metrics = metric_set(roc_auc))
```

Combines the random forest model and preprocessing recipe into a workflow and performs hyperparameter tuning using cross-validation.

Finding the best Random Forest Model

```
rf_res %>%
  show_best(metric = "roc_auc")
```

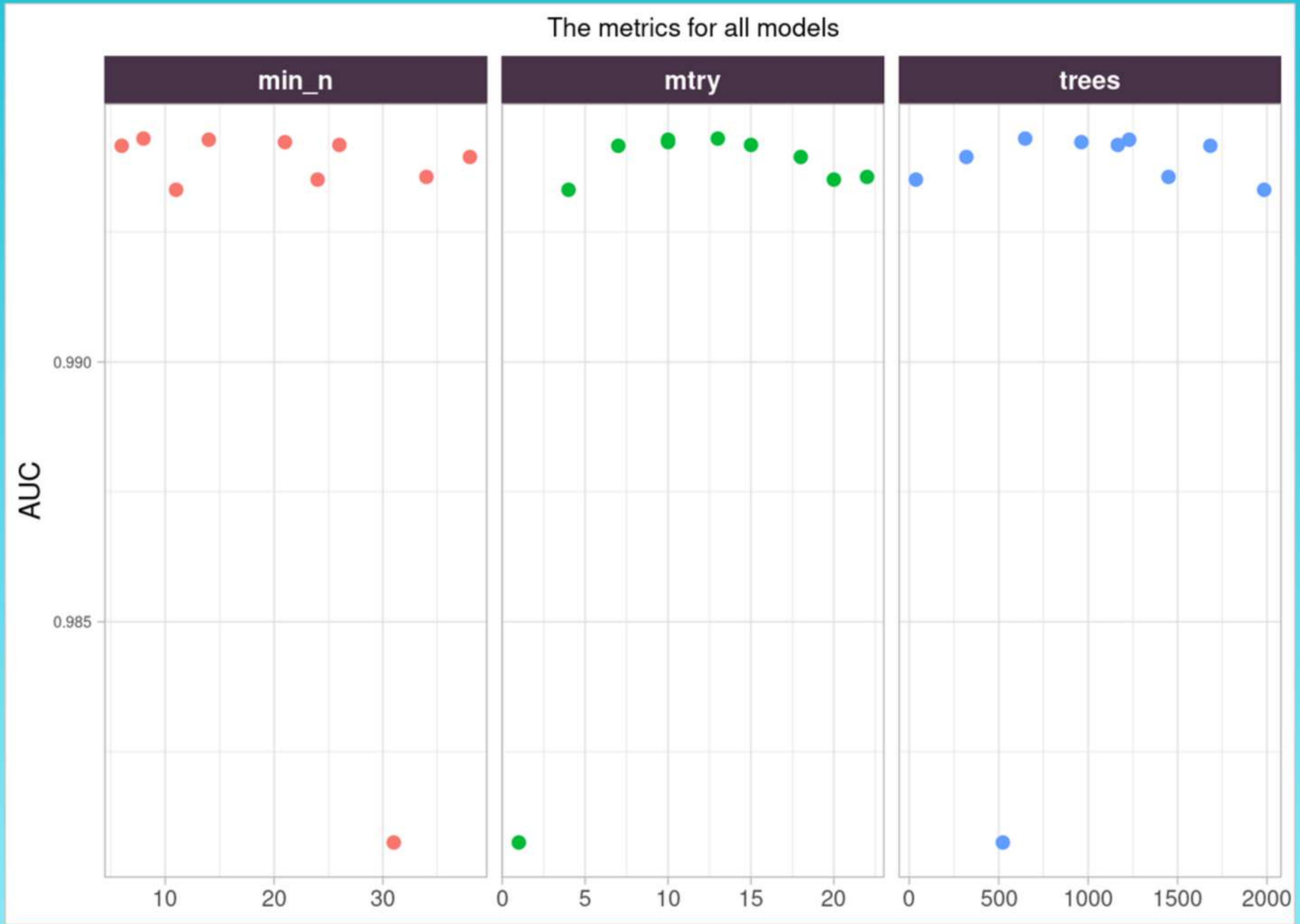
```
## # A tibble: 5 × 9
##    mtry trees min_n .metric .estimator  mean     n  std_err .config
##   <int> <int> <int> <chr>   <chr>      <dbl> <int>    <dbl> <chr>
## 1    13   648     8 roc_auc binary     0.994    10 0.000190 Preprocessor1_Model…
## 2    10  1229    14 roc_auc binary     0.994    10 0.000195 Preprocessor1_Model…
## 3    10   962    21 roc_auc binary     0.994    10 0.000192 Preprocessor1_Model…
## 4    15  1166    26 roc_auc binary     0.994    10 0.000195 Preprocessor1_Model…
## 5     7  1683     6 roc_auc binary     0.994    10 0.000200 Preprocessor1_Model…
```

We can see for different combinations of the hyperparameters which has the least std error which concludes that value will be the optimal value !

ROC AUC Plot VS mtry, min_n and trees

```
rf_best <-
  rf_res %>%
  select_best(metric = "roc_auc")
rf_best
```

```
## # A tibble: 1 × 4
##    mtry trees min_n .config
##   <int> <int> <int> <chr>
## 1    13   648     8 Preprocessor1_Model02
```



The graph and the code output align perfectly because the code has selected the hyperparameters (mtry = 13, trees = 648, min_n = 8) that correspond to the highest AUC scores on the graph. This visualization helps validate that the chosen model parameters are indeed optimal for maximizing the ROC AUC, which is the goal of hyperparameter tuning

Finalise workflow

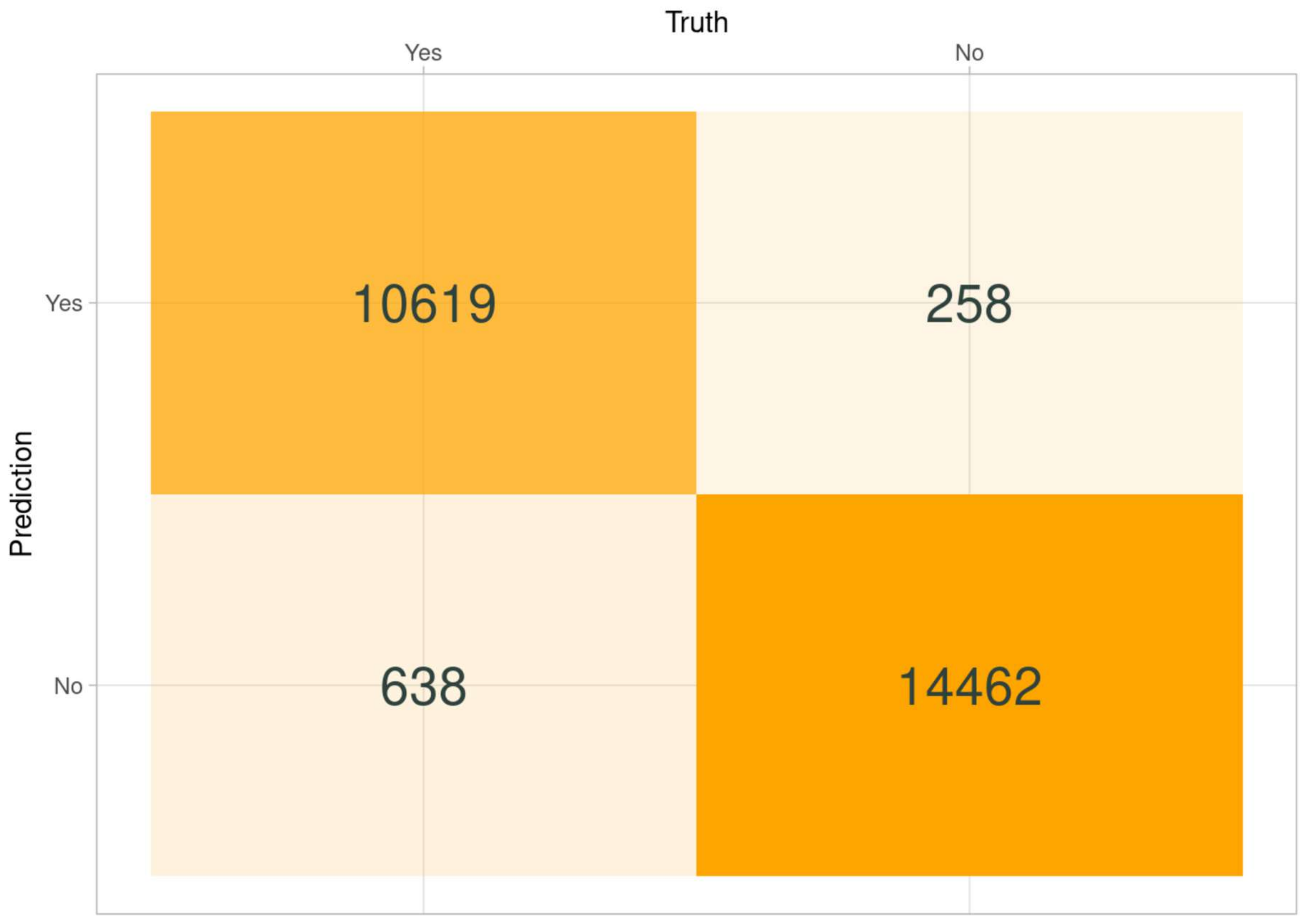
```
## == Workflow ==================================================================
## Preprocessor: Recipe
## Model: rand_forest()
##
## -- Preprocessor ----------------------------------------------------------------
## 3 Recipe Steps
##
## • step_dummy()
## • step_zv()
## • step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Random Forest Model Specification (classification)
##
## Main Arguments:
##   mtry = 13
##   trees = 648
##   min_n = 8
##
## Engine-Specific Arguments:
##   importance = impurity
##
## Computational engine: ranger
```

Confusion Matrix

- This confusion matrix shows improved model performance compared to the previous one:

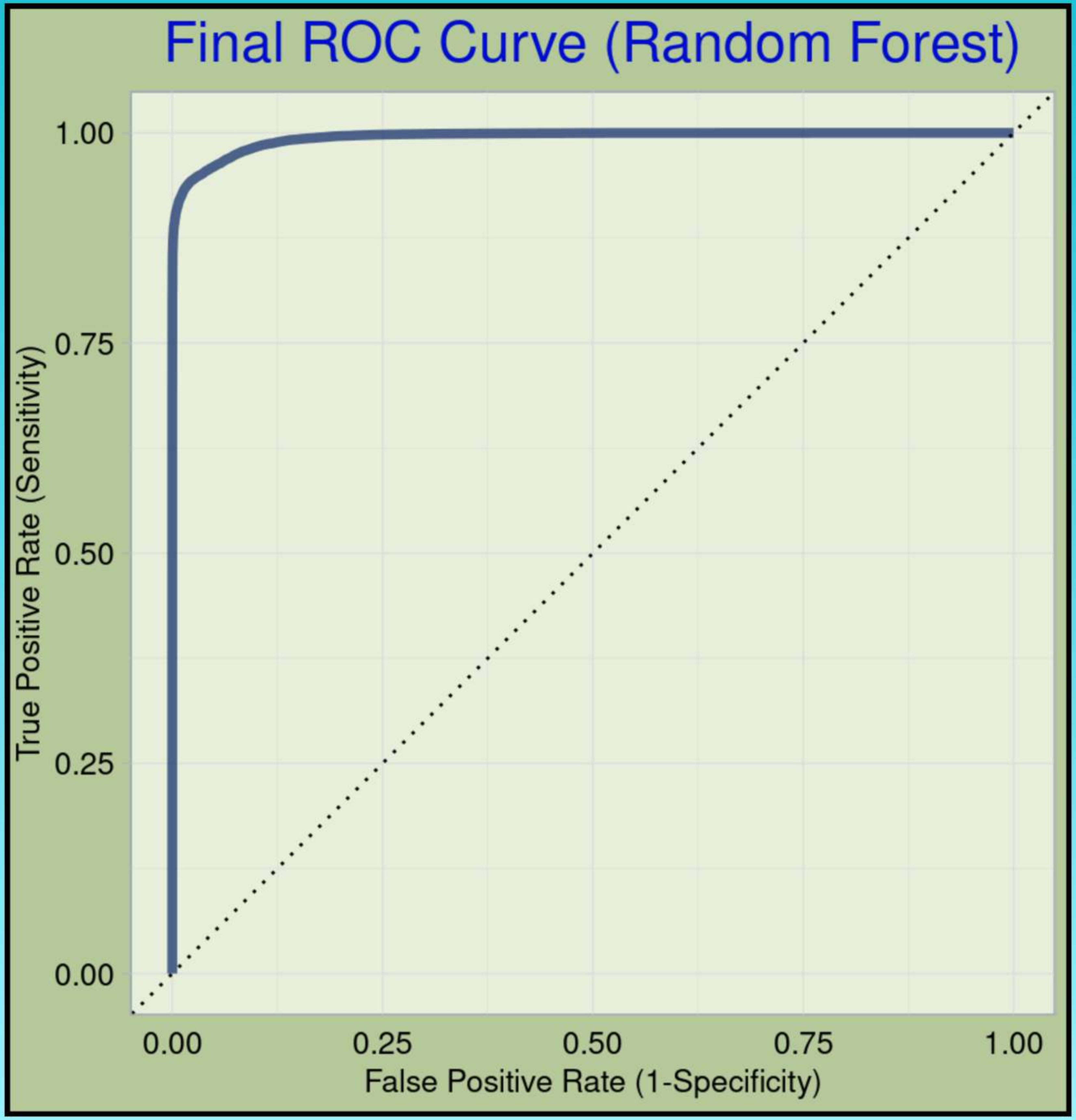
- **True Positives (Yes-Yes):** 10,619
- **False Positives (Yes-No):** 258
- **False Negatives (No-Yes):** 638
- **True Negatives (No-No):** 14,462

- The model correctly predicts a higher number of instances (both positive and negative) with fewer misclassifications, indicating better accuracy.



Final RF ROC curve

Hence, random forest is a better fit for our data with 96% accuracy

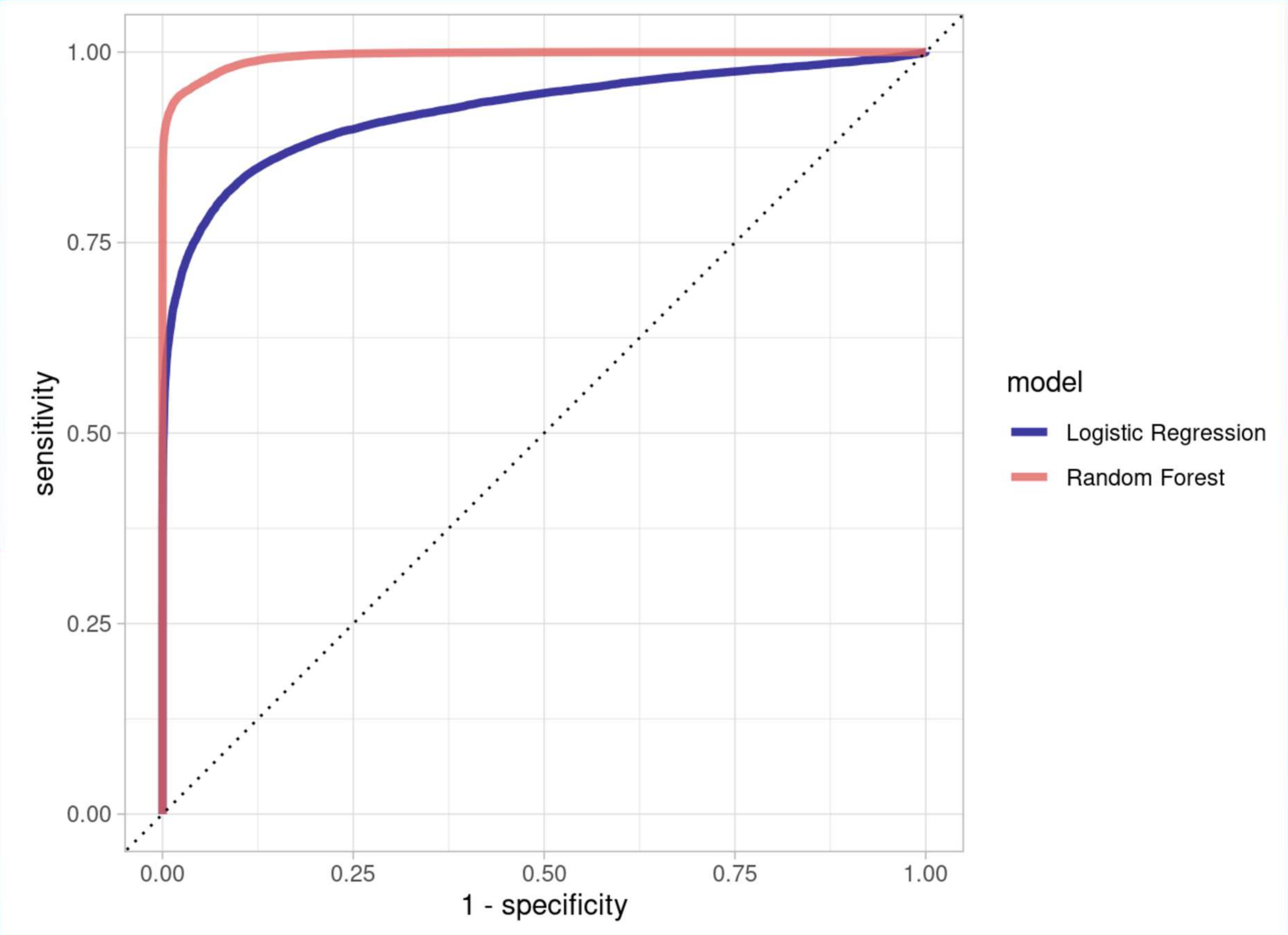


ROC Curve logistic regression vs RF

The ROC curve shows the trade-off between sensitivity and specificity.
Classifiers that give curves closer to the top-left corner indicate a better performance.
The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

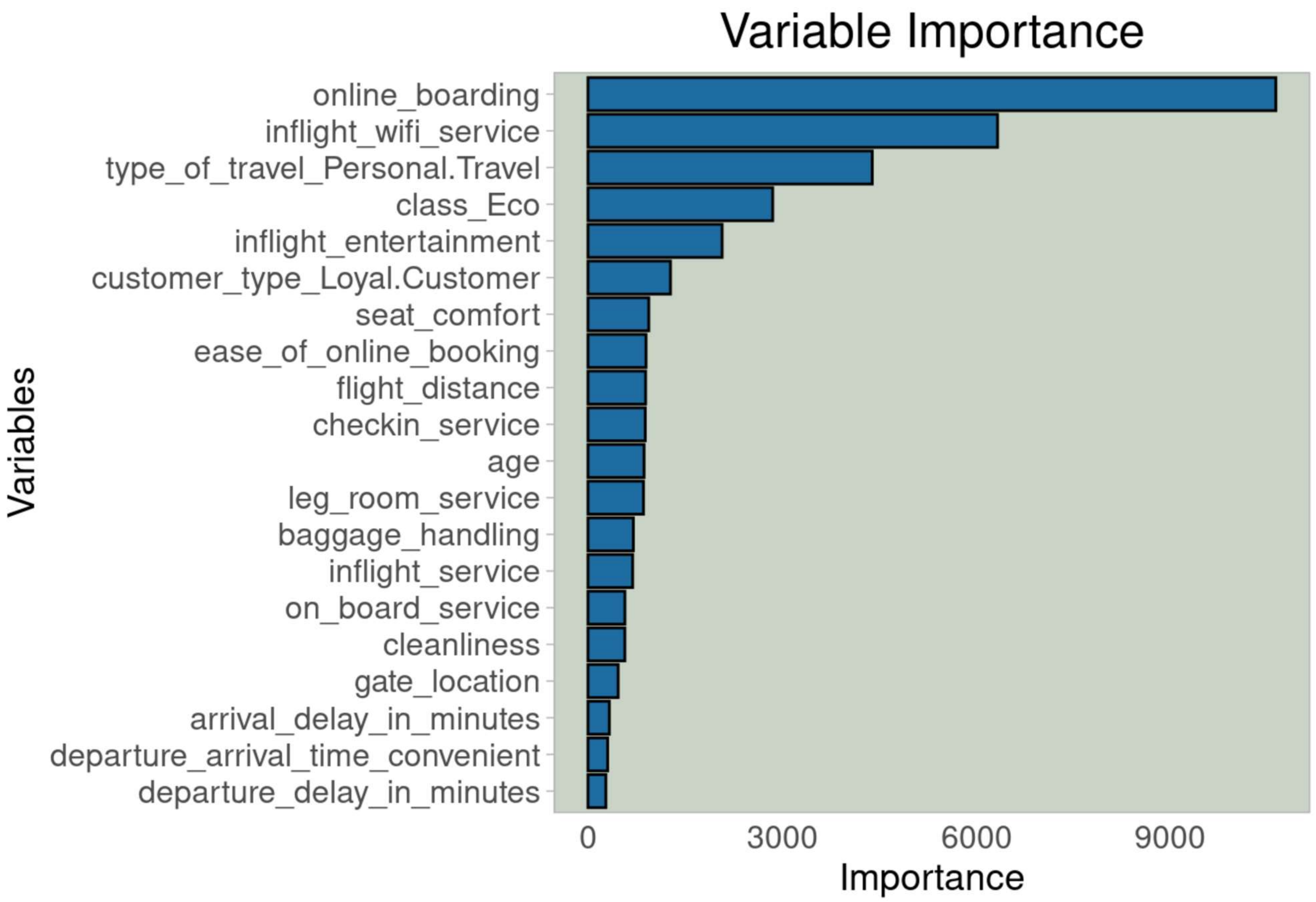
```
collect_metrics(final_rf_fit)
```

```
## # A tibble: 2 × 4
##   .metric  .estimator .estimate .config
##   <chr>    <chr>          <dbl> <chr>
## 1 accuracy binary         0.966 Preprocessor1_Model1
## 2 roc_auc  binary         0.995 Preprocessor1_Model1
```



FEATURE IMPORTANCE

- Online boarding and In-flight Wi-Fi service have the highest importance score, suggesting it's the most influential factor and indicates its significance in the model.
- Type of travel (Personal Travel): This categorical variable is also considered important.
- Departure delay in minutes: This variable has the lowest importance score, suggesting it has little impact on the model's predictions.



CONCLUSION

Learning Outcomes

- Conclusion for the Presentation
- Enhancing passenger satisfaction isn't just about the primary features; secondary services play a crucial role.
- By prioritizing and improving aspects like legroom service, baggage handling, in-flight service, cleanliness, onboard service, and food and drink, airlines can significantly elevate the overall customer experience.
- These enhancements not only contribute to higher satisfaction levels but also bolster customer loyalty, leading to a stronger brand reputation and competitive edge in the market.

THANK YOU

BOM - DXB

SAFE FLIGHT | HAVE A FUN DAY | NAMASTE

0 24563 84926 54 2