Unlocking Customer Experience: A Data-Driven Analysis of Air France's Survey Results

Group Project Using R Programming and Data Analysis Techniques

MOKRANE Naïm

GOMEZ Diana

OHEIX Romain

ARAGON Monica

R Programming for Business
Serge Nyawa

November 10, 2024

Toulouse Business School

Table of Contents

1. Introduction

Overview

This project focuses on analyzing customer satisfaction data collected through a survey conducted by Air France's marketing department. The survey captures customer feedback on various services experienced at different stages of their journey—before, during, and after their flights. By applying data analysis techniques using R programming, the goal of this analysis is to uncover patterns in customer satisfaction and identify areas where service improvements can be made. These insights will help Air France enhance its operational strategies and optimize the overall customer experience.

Objectives and Research Questions

The primary objective of this analysis is to clean, explore, and analyze the customer satisfaction data to uncover key patterns and areas for improvement. Through this process, we aim to answer several research questions, such as:

- What factors influence customer satisfaction at different stages of the flight experience?
- How do demographic variables (such as age, travel purpose, and class of service) correlate with satisfaction levels?
- Are there any specific service areas that require targeted improvements to enhance overall satisfaction?

Data Source and Description

The dataset used in this project contains detailed responses from Air France customers, collected across 25 columns. These include ratings for various services, such as inflight Wi-Fi, seating comfort, and flight delays, as well as demographic information (e.g., type of travel, class of service). The dataset is structured with both quantitative and categorical variables, the data is diverse and includes multiple data types. This detailed data offers a comprehensive view of the customer journey and provides an ideal basis for uncovering meaningful insights to drive service improvements.

Variables in the Dataset

Customer Information:

- `id`: Customer ID
- `Gender`: Gender of the customer (Male / Female)
- `Customer Type`: Whether the customer is loyal or disloyal
- `Age`: Age of the customer (values ranging from 7 to 85, with a mean of 39.38)
- `Type of Travel`: Business or personal travel
- `Class`: Travel class (Eco Plus, Business,Eco)
- `Flight Distance`: Distance traveled in miles (values ranging from 31 to 4,983 miles, with a mean of 1,189)

Service Ratings:

These variables represent customer ratings on different aspects of the airline service. Ratings range from 0 to 5, where 0 represents the lowest satisfaction and 5 represents the highest. The ratings are as follows:

- `Inflight wifi service`
- `Seat comfort`
- `Food and drink`
- `Ease of online booking`
- `Departure/Arrival time convenient`
- `Gate Location`
- `Online Boarding`
- `Inflight entertainment`
- `On-board service`
- `Leg room service`
- `Baggage handling`
- `Check-in service`
- `Inflight service`
- `Cleanliness`

Flight Delays:

- `Departure Delay:` Delay time before the flight
- `Arrival Delay in Minutes`: Delay time after the flight

Customer Satisfaction:

- `Satisfaction`: Customer satisfaction (e.g., "satisfied", "neutral or dissatisfied")

Data Summary

The dataset provided by Airbus comprises 103,904 rows and 25 columns. It contains 19 numerical variables and 6 categorical variables.

- **Numerical variables** (`Age`, `Flight Distance`, `Departure Delay in Minutes`, `Arrival Delay in Minutes`, and service ratings like `Inflight wifi service`).
- **Categorical variables** (`Gender`, `Customer Type`, `Type of Travel`, `Class`, `satisfaction`).

The dataset has no missing values and no duplicated rows. However, a minor issue was identified with the ‘Arrival Delay in Minutes’ variable, which was mistakenly categorized as a character type rather than an integer.

Additionally, we standardized the letter case for all variables. For example, in the ‘Customer Type’ variable, we will change the "d" in "disloyal" to uppercase so that the classification is consistent and prevents potential issues in later stages, such as data visualization.

Bellow is a summary of all the variables included in the Dataset:

```
       X                id             Gender          Customer.Type          Age
 Min.   :     0   Min.   :     1   Length:103904      Length:103904      Min.   : 7.00
 1st Qu.: 25976   1st Qu.: 32534   Class :character   Class :character   1st Qu.:27.00
 Median : 51952   Median : 64857   Mode  :character   Mode  :character   Median :40.00
 Mean   : 51952   Mean   : 64924                                        Mean   :39.38
 3rd Qu.: 77927   3rd Qu.: 97368                                        3rd Qu.:51.00
 Max.   :103903   Max.   :129880                                        Max.   :85.00
 Type.of.Travel       Class           Flight.Distance Inflight.wifi.service
 Length:103904      Length:103904      Min.   :  31    Min.   :0.00
 Class :character   Class :character   1st Qu.: 414    1st Qu.:2.00
 Mode  :character   Mode  :character   Median : 843    Median :3.00
                                       Mean   :1189    Mean   :2.73
                                       3rd Qu.:1743    3rd Qu.:4.00
                                       Max.   :4983    Max.   :5.00
 Departure.Arrival.time.convenient Ease.of.Online.booking Gate.location   Food.and.drink
 Min.   :0.00                      Min.   :0.000          Min.   :0.000   Min.   :0.000
 1st Qu.:2.00                      1st Qu.:2.000          1st Qu.:2.000   1st Qu.:2.000
 Median :3.00                      Median :3.000          Median :3.000   Median :3.000
 Mean   :3.06                      Mean   :2.757          Mean   :2.977   Mean   :3.202
 3rd Qu.:4.00                      3rd Qu.:4.000          3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :5.00                      Max.   :5.000          Max.   :5.000   Max.   :5.000
 Online.boarding  Seat.comfort   Inflight.entertainment On.board.service Leg.room.service
 Min.   :0.00    Min.   :0.000   Min.   :0.000          Min.   :0.000    Min.   :0.000
 1st Qu.:2.00    1st Qu.:2.000   1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.000
 Median :3.00    Median :4.000   Median :4.000          Median :4.000    Median :4.000
 Mean   :3.25    Mean   :3.439   Mean   :3.358          Mean   :3.382    Mean   :3.351
 3rd Qu.:4.00    3rd Qu.:5.000   3rd Qu.:4.000          3rd Qu.:4.000    3rd Qu.:4.000
 Max.   :5.00    Max.   :5.000   Max.   :5.000          Max.   :5.000    Max.   :5.000
 Baggage.handling Checkin.service Inflight.service  Cleanliness    Departure.Delay.in.Minutes
 Min.   :1.000    Min.   :0.000   Min.   :0.00     Min.   :0.000   Min.   :   0.00
 1st Qu.:3.000    1st Qu.:3.000   1st Qu.:3.00     1st Qu.:2.000   1st Qu.:   0.00
 Median :4.000    Median :3.000   Median :4.00     Median :3.000   Median :   0.00
 Mean   :3.632    Mean   :3.304   Mean   :3.64     Mean   :3.286   Mean   :  14.82
 3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:5.00     3rd Qu.:4.000   3rd Qu.:  12.00
 Max.   :5.000    Max.   :5.000   Max.   :5.00     Max.   :5.000   Max.   :1592.00
 Arrival.Delay.in.Minutes satisfaction
 Length:103904            Length:103904
 Class :character         Class :character
 Mode  :character         Mode  :character
```

2. Data cleaning and enrichment

Data Cleaning:

- **Missing Values and Duplicates**: We observed that there are no missing values or duplicated rows in the dataset.
- **Column Removal**: The first column was removed as it served only as a row index and duplicated the customer ID provided by Air France.

Feature Engineering:

- **New Columns Created**:
  - `Total Flight avg (Final Note)`: Average of ratings given by customers (0 to 5).

    - `Before flight Average Rate`: Average rating for categories related to pre-flight services.
    - `After flight Average Rate`: Average rating for categories related to post-flight services.

Variable Transformation:

- Arrival Delay In Minutes: Converted into an integer format.
- Satisfaction Variable Transformation: The original binary satisfaction variable ("Satisfied" / "Neutral or Dissatisfied") was enhanced to reflect five categories:
    - **Dissatisfied**: 0 to 2
    - **Neutral**: 2 to 3
    - **Satisfied**: 3 to 4
    - **Strongly Satisfied**: 4 to 5
- Satisfaction Column Removal: The original satisfaction column was deleted after the new variable was created.

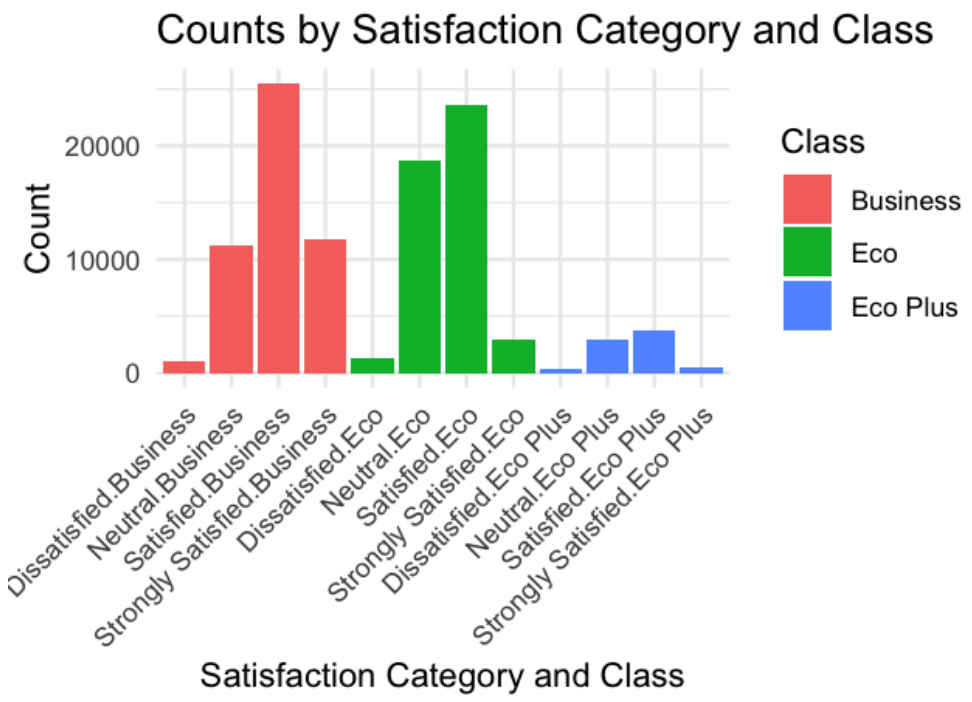
Customer Segmentation:

- Age Groups: The age variable was divided into categories to allow better customer segmentation:
    - **8 - 17**: Kids
    - **18 - 30**: Young Adults
    - **31 - 60**: Adults
    - **60+**: Senior

3. Exploratory Data Analysis

Initial Data Exploration: We conducted descriptive analysis and visualizations to uncover trends and distributions across demographic and service-related variables.

	Dissatisfied	Neutral	Satisfied	Strongly Satisfied
Business	1004	11269	25531	11722
Eco	1324	18734	23589	2946
Eco Plus	306	2877	3740	545



To identify the most influential variables affecting customer satisfaction, a linear regression model is employed to analyze the data. This approach allows us to determine which factors have the greatest impact on satisfaction levels.

Result of the Multinomial LM model :

```
Coefficients:
                  (Intercept) Range.AgeYoung Adults Range.AgeAdults Range.AgeSenior
Neutral             -32.62233             0.1699640       0.3628792       0.5886536
Satisfied           -83.44190             0.2442844       0.5707665       0.8029293
Strongly Satisfied -175.37899             0.4221840       0.2164661       0.3938470
                  Flight.Distance    ClassEco ClassEco Plus  GenderMale Inflight.wifi.service
Neutral             4.784521e-05 -0.04866475    -0.3466558 -0.12999943              1.557613
Satisfied           1.472851e-04 -0.05533123    -0.2662624 -0.19690943              3.048484
Strongly Satisfied  1.292605e-04  0.41420318     0.4859254  0.02405319              4.924180
                  Departure.Arrival.time.convenient Ease.of.Online.booking Gate.location
Neutral                                    1.250148               1.139628      1.188477
Satisfied                                  2.307503               2.369132      2.388726
Strongly Satisfied                         3.963948               4.019322      4.108102
                  Online.boarding Seat.comfort On.board.service Leg.room.service
Neutral                  1.088154     1.878012         1.324217         1.241728
Satisfied                2.078940     3.723096         2.778488         2.528702
Strongly Satisfied       3.787768     6.215536         4.657129         4.284701
                  Baggage.handling Checkin.service Inflight.service Cleanliness
Neutral                   1.244650        1.157624         1.309263    2.141869
Satisfied                 2.632006        2.118029         2.823029    4.761983
Strongly Satisfied        4.591604        3.667681         4.820211    7.651763
                  Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
Neutral                       -0.005810031              0.007827909
Satisfied                     -0.007259183              0.008771713
Strongly Satisfied            -0.006810378              0.003731265
```

```
Std. Errors:
                  (Intercept) Range.AgeYoung Adults Range.AgeAdults Range.AgeSenior
Neutral           0.0025505911            0.02133707      0.02161942     0.017482187
Satisfied         0.0016650948            0.01913220      0.01800405     0.018728997
Strong Satisfied  0.0009195478            0.01937816      0.02244883     0.005482742
                  Flight.Distance   ClassEco ClassEco Plus GenderMale Inflight.wifi.service
Neutral              4.283241e-05 0.03452572    0.03209385 0.02981097            0.02352933
Satisfied            4.718178e-05 0.02756847    0.03658166 0.02190454            0.02061995
Strong Satisfied     5.304888e-05 0.03542975    0.01200559 0.03198179            0.02718102
                  Departure.Arrival.time.convenient Ease.of.Online.booking Gate.location
Neutral                                  0.02115890             0.02366753    0.02090744
Satisfied                                0.02144718             0.02288070    0.02113140
Strong Satisfied                         0.02714642             0.02992533    0.02735266
                  Online.boarding Seat.comfort On.board.service Leg.room.service
Neutral                0.02035283   0.02172180       0.01737739       0.02006829
Satisfied              0.01932175   0.02029545       0.01550380       0.02056309
Strong Satisfied       0.02629883   0.02748548       0.01972068       0.02619723
                  Baggage.handling Checkin.service Inflight.service Cleanliness
Neutral                 0.01906029      0.03031381       0.01964652  0.02153464
Satisfied               0.01690771      0.03329684       0.01756518  0.01946412
Strong Satisfied        0.02350661      0.04092750       0.02414800  0.02471124
                  Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
Neutral                          0.003706895              0.003664244
Satisfied                        0.004116603              0.004066277
Strong Satisfied                 0.004916056              0.004850286
```

The most impactful variables when we take every class of flights and every age are :

- Inflight.wifi.service:
    - Coefficients for "Strongly Satisfied": ~4.92
    - This variable has one of the highest coefficients, suggesting a very strong positive impact on satisfaction. Good wifi service likely correlates with higher satisfaction ratings.

- Seat.comfort and On.board.service:
    - Coefficients for "Strongly Satisfied":
        - Seat.comfort: ~6.22
        - On.board.service: ~4.66
    - These service quality variables indicate that improvements in seat comfort and on-board service significantly elevate satisfaction levels

- ClassEco Plus:
    - Coefficient for "Strongly Satisfied": ~0.49
    - This positive coefficient shows that passengers in this class have a greater likelihood of reporting higher satisfaction compared to other classes.

- Departure.Arrival.time.convenient:
    - Coefficients for "Strongly Satisfied": ~4.01
    - Convenience in scheduling is another key factor, with a strong positive association with higher satisfaction.

4. Customer Segmentation and Targeted Analysis

The project has shifted its focus from analyzing the correlation between satisfaction categories and their individual average rates to a more insightful approach centered on customer demographics and flight-related characteristics. Rather than exploring the average rates of satisfaction categories, which inherently leads to biased results due to their composition; the goal is now to understand customer satisfaction by examining factors such as age, gender, customer type, type of travel, and class. By analyzing how satisfaction or dissatisfaction varies across these demographic and flight-related attributes, the project aims to identify patterns in customer behavior and attitudes.

Class-Based Segmentation

The primary focus will be on grouping customers based on their demographic information and flight data, such as age groups, gender differences, customer loyalty, business versus leisure travel, and class of service (eco, business, etc.). These segments will be analyzed to determine which groups are more prone to dissatisfaction and what factors contribute to their experiences. The aim is to create clusters or categories of customers who exhibit distinct levels of dissatisfaction.

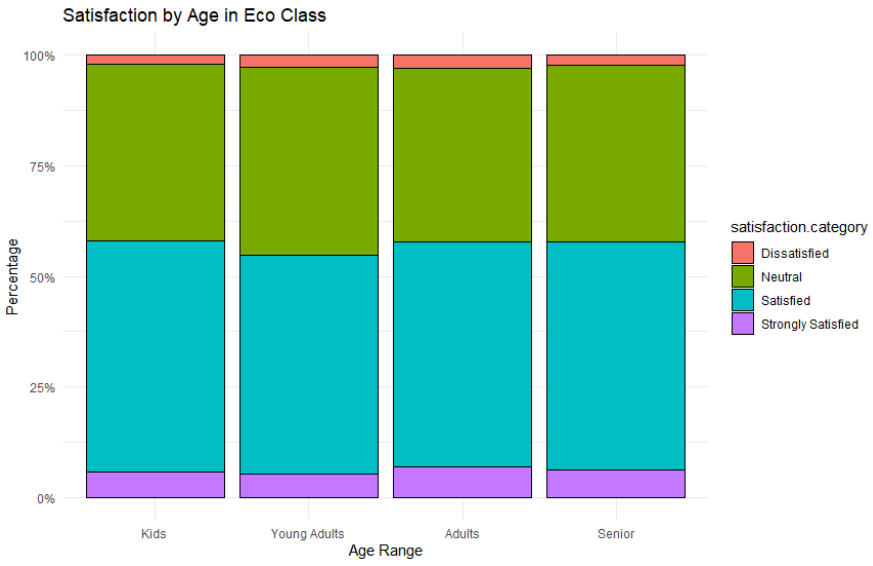
After analyzing the different demographic and flight-related information (Age, Gender, Customer Type, Class Type, and the Type of Travel) of customers by creating different plots, we identified that Class Type is the one that provides the most valuable insights for customer segmentation. Through data visualization (provided above), we observed that Business class passengers exhibit the highest satisfaction levels, both proportionally and in absolute terms. In contrast, Eco and Eco Plus classes show similar satisfaction levels proportionally, though Eco class has a larger customer base. Given its larger customer base and the potential value this volume represents for the company, the Eco segment will be our primary focus for further recommendations, rather than Eco Plus

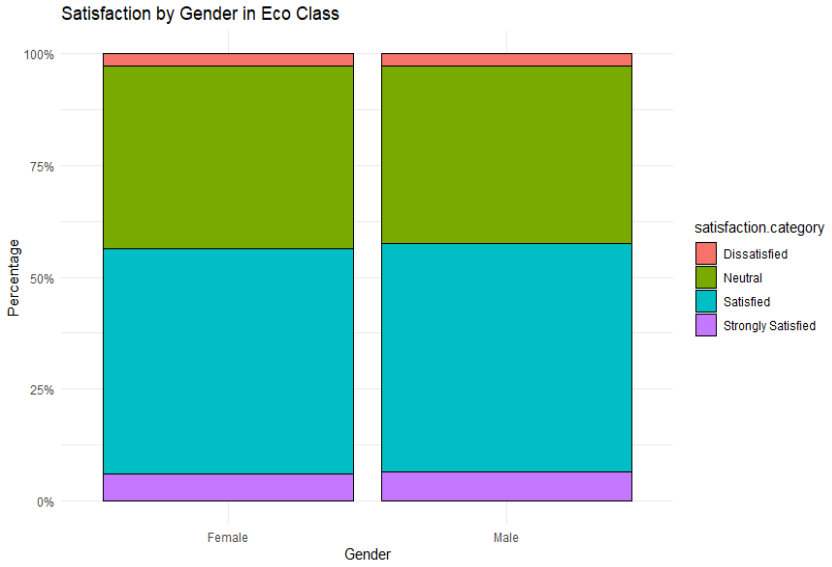
After selecting the Eco group for targeted recommendations, we created a new dataset d_eco that contains only the data from eco class since it is the segment we are focusing on. We delved further into segmentation as we believe it offers valuable potential for deeper analysis and insights. We created plots with three different variables (Age, Gender and Flight Distance), all in proportional values.

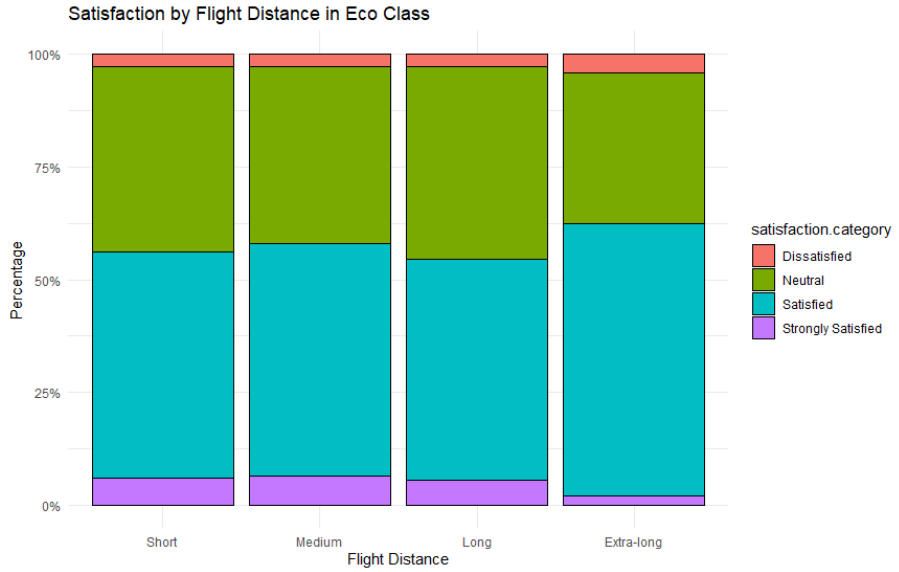
Flight Distance Analysis

In this phase, we also refined the flight distance variable; as part of this, we converted the flight distance variable from a numerical value into categories, similar to how we previously categorized Age and satisfaction. This approach enables a more structured analysis by standardizing the data for meaningful comparisons across segments. The new standard for Flight distance which was divided into four segments to assess satisfaction as follows:

- **Short**: 0 - 500 km
- **Medium**: 501 - 1500 km
- **Long**: 1501 - 3000 km
- **Extra-long**: 3001+ km







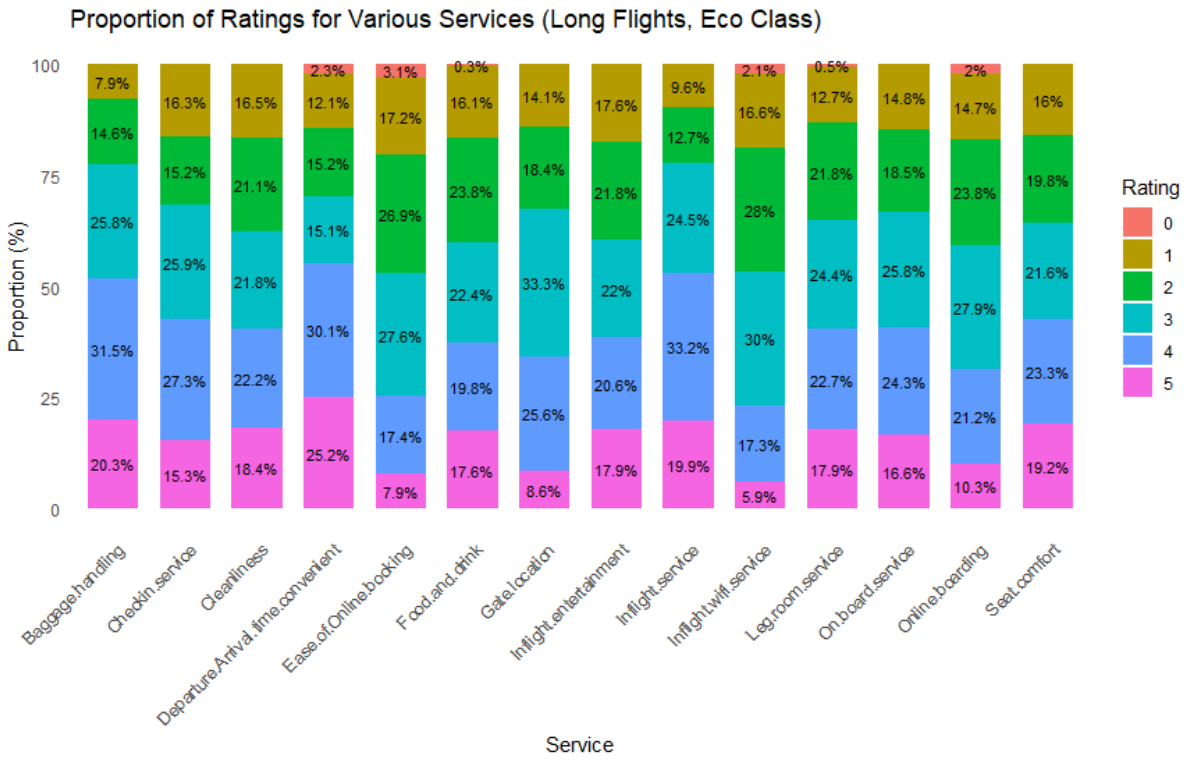
For the Age and Gender plots we see no significant insights, as there are no remarkable differences among the groups in either of the plots. Nor can we find any overall tendency from the visuals.

In contrast, we can find some insight from the Flight Distance plot compared to the previous ones. Here we can see the “Dissatisfied” category tends to go higher the longer the flight means, therefore, a good theory would be the customers are lacking good service and amenities during the flight. On the other end of the spectrum we find the Strongly Satisfied ratings also get smaller the longer the flight gets. As to the neutral and satisfied groups of ratings tend to be contradictory for the Extra Long flights.
Therefore, considering the overall picture we can regard the “Long” flights to be the worst performing ones, as the proportion of dissatisfaction and neutrality is the highest. We chose

to also include the neutral rating as an opportunity for improvement as well since it represents points between 2 and 3 out of 5.

Service Impact Analysis for Long Flights (Eco Class)

After finding a further level of segmentation, the idea is to find how each of the services offered by Air France are individually rated to make informed recommendations on which specific services must be improved. Therefore the last phase into our analysis consists of finding the proportion the ratings given in every service and see which services have the highest opportunity of improvement
To achieve this, we created a new dataset long_flights_eco where we use filters so it only contains the eco class and the “long” travels and a list that contains the different rating columns from our dataset. We transform the data into a long format, with each service rating in a single row then we count each rating for each service and we calculate the proportion of each rating in the long_flights_ratings object. At last, we used this object to build the following plot.



Our analysis is now complete, allowing us to extract key insights and recommendations from the final plot. This will help identify the most crucial services for enhancing the Eco class experience on long-haul flights.

5. Recommendations

Summary of Findings & Recommendations.

Along the creation of the project and analysis, we have identified several key insights and recommendations to enhance the Eco class experience on long-haul flights for Air France. The analysis shows that customer satisfaction is particularly low in areas such as “Gate Location” and “Inflight WiFi Service.” To address this, we recommend improving gate coordination to reduce long walks and inconvenient gate assignments and enhancing WiFi quality and accessibility to meet customer expectations on long flights.
Conversely, services like “Baggage Handling” and “On-board Service” received strong satisfaction ratings, suggesting these areas are currently performing well. Maintaining these high standards will be essential to continue meeting passenger expectations. Ratings for “Seat Comfort” and “Leg Room Service” were mixed, indicating variability in passenger experience, possibly due to personal preferences or seat configurations. Investing in seat upgrades or offering more legroom options on long-haul flights could help convert neutral or dissatisfied passengers into satisfied ones.
Further improvement opportunities exist in “Inflight Entertainment” and “Food and Drink,” which show moderate satisfaction but also a fair share of neutral and low ratings. Expanding food variety and entertainment options could boost satisfaction in these categories. Lastly, “Ease of Online Booking” and “Check-in Service” displayed moderate satisfaction levels with many neutral ratings, suggesting that streamlining these processes, perhaps with faster kiosks or digital assistance, could positively impact the customer experience.
In summary, our recommendations focus on prioritizing improvements in areas with high dissatisfaction, such as gate location, inflight WiFi, and seat comfort, while maintaining the high quality of baggage handling and on-board service. Incremental upgrades in food quality, entertainment, and check-in processes could further enhance satisfaction, creating a more positive and seamless experience for Eco class passengers on long flights.

6. Final Thoughts

This project successfully analyzed customer satisfaction by examining factors like demographics, flight characteristics, and service ratings. Key influences on satisfaction included inflight WiFi, seat comfort, onboard service, and convenient scheduling, especially impacting Eco class passengers on long flights. The project highlighted the connection between demographic factors, service quality, and customer satisfaction, offering airlines a clear path for data-driven improvements in the passenger experience. By using data wrangling, exploratory analysis, and segmentation, we’ve laid the groundwork for creating targeted strategies that can enhance satisfaction and foster customer loyalty.