A Computational Study on the Airline Passenger Reviews

Passenger reviews by British Airways Passengers

December 10, 2023

Introduction:

With the huge amount of airline review data that is generated every second, passenger satisfaction plays an important role in the Airline industry. As the saying goes, *"It's easy to lie with statistics but it's hard to tell the truth without statistics"*, bringing out the best from the multitude of problems that data can address, it is crucial to not just understand the patterns in data but also making meaningful predictions and build accurate models for analysis.

This very thought triggered me to find some meaningful results using the "Airline Reviews" data set for British Airways.

As part of this project I would like to find solutions the following questions,

(1) Based on the airline review data, what can we conclude about the Overall satisfaction of the passengers using British Airways.

(2) Using existing information about the ratings for each aircraft type(such as A230, A330, Airbus 380, Boeing 747,etc) of British Airways, how can we predict the Overall rating that a future passenger might provide for a certain flight/aircraft.

(3) Based on the airline reviews given by passengers, should British Airways be a recommended choice for passengers?

(4) Make an informed decision about which type of the passengers should the airline focus for future improvements by looking at the existing dataset and finding the average rating given by passengers of a specific seat type(Business class, Economy class, Premium Economy, First Class).

- Why a data-driven, computational approach?

A data-driven and computational approach, will not just make the analysis simple and visually straightforward to comprehend but will prove the results with strong evidence based statistical methods.

Additionally, customer satisfaction is a core KPI in the airline industry and any decision made will impact the entirety of the airline on a large scale. Therefore, decisions have to be justified by data and data alone.

Each research question can be backed up using core statistical concepts such as Law of Large numbers, Monte Carlo Simulations, Predictions, Model fitting, Testing of Hypothesis among many others.

Computational techniques will be helpful right from data wrangling to inference, including data organization, exploration, summarization, and analysis which will be useful in investigating and find solutions for current and pressing future real-world issues.

Dataset Description

The dataset for current project contains customer feedback for British Airways, which has been extracted from AirlineQuality(https://www.airlinequality.com/) through web scraping. This dataset serves as a rich resource for sentiment analysis to explore and understand the sentiments expressed by British Airways passengers.

The dataset consists of the Overall ratings given by customers on the following factors from their flight experience ranging from 2015 to 2023.

The raw dataset consists of 19 columns as stated below

OverallRating : The overall rating given by the customer.

ReviewHeader : The header or title of the customer's review.

Name : The name of the customer providing the feedback.

Datetime : The date and time when the feedback was posted.

VerifiedReview : Indicates whether the review is verified or not.

ReviewBody : The detailed body of the customer's review.

TypeOfTraveller : The type of traveler (e.g., Business, Leisure).

SeatType : Class of the traveler (e.g. Business, Economy).

Route : The flight route taken by the customer.

DateFlown : The date when the flight was taken.

SeatComfort : Rating for seat comfort.

CabinStaffService : Rating for cabin staff service.

GroundService : Rating for ground service.

ValueForMoney : Rating for the value for money.

Recommended : Whether the customer recommends British Airways.

Aircraft : The aircraft used for the flight.

Food&Beverages : Rating for food and beverages.

InflightEntertainment : Rating for inflight entertainment.

Wifi&Connectivity : Rating for onboard wifi and connectivity.

Data Wrangling

For the chosen dataset, we will have to perform the following data wrangling steps to make it ready for subsequent analysis.

(1) Importing the data from the .csv file using the *'read_csv()'* function from the *'readr'* package belonging to the core *'tidyverse'* package.

(2) Identifying the columns/ attributes that can be used for analysis and removing the columns that are not in use.

(3) As part of tidying the data,

- We split the "Datatime" column to Date, Month and Year for ease of calculation.
- We convert the symbols such as '-' to 'to' in the Route column and split the "Route" column to Source and Destination
- Change the type of the Year column from character to numeric.
- Remove special characters from the column names such as '&' in Food&Beverages and Wifi&Connectivity.
- Arrange/Sort the data according to Year and Month.

(4) We then select the columns needed for computation and drop the columns which do not contribute towards analysis.

(5) Finally we remove "null" from the data using the *'omit()'* function.

(6) Manually inspected the rows causing warnings and decide on the appropriate action.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(dplyr)
library(readr)
```

```
ba_airreview<-read_csv("BA_AirlineReviews.csv")
```

```
## New names:
## * ‘‘ -> ‘...1‘
```

```
## Warning: One or more parsing issues, call ‘problems()‘ on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 3701 Columns: 20
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): ReviewHeader, Name, Datetime, ReviewBody, TypeOfTraveller, SeatTyp...
## dbl  (9): ...1, OverallRating, SeatComfort, CabinStaffService, GroundService...
## lgl  (1): VerifiedReview
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
# consider the original dataset to a different variable
airline_reviews <- ba_airreview

#Tidy the Route Column
airline_reviews <- airline_reviews |>
  mutate(Route = gsub("-", " to ", Route))

# Split the Datetime column into two columns: date, month, year
airline_reviews <- airline_reviews |>
  separate(Datetime, into = c("Date", "Month", "Year"), sep = " ")

# Split the Route column into two columns: Source and Destination
airline_reviews <- airline_reviews |>
  separate(Route, into = c("Source", "Destination"), sep = " to ",extra = "merge",
           fill = "right")

# Convert the column from character to numeric
airline_reviews$Year <- as.numeric(airline_reviews$Year)

# Sort the data based on the Year column
airline_reviews <- airline_reviews |>
  arrange(Year,Month)

# Rearrange columns based on the desired order
```

```r
desired_order<-c("Year","Month","Source","Destination","Aircraft","SeatType",
                 "OverallRating","Recommended","SeatComfort","CabinStaffService",
                 "GroundService","ValueForMoney","Food&Beverages",
                 "InflightEntertainment","Wifi&Connectivity")

airline_review<-airline_reviews[,c("Year","Month","Source","Destination",
                              "Aircraft","SeatType","OverallRating","Recommended",
                              "SeatComfort","CabinStaffService","GroundService",
                              "ValueForMoney","Food&Beverages",
                              "InflightEntertainment","Wifi&Connectivity")]

#Omit the null values
airline_reviews_a<-na.omit(airline_review)

#Renaming columns
airline_reviews_a <- airline_reviews_a |>
  rename("FoodBeverage"="Food&Beverages","WifiConnectivity"="Wifi&Connectivity")

cat("After data wrangling, the data looks like below\n")
```

```
## After data wrangling, the data looks like below
```

```r
head(airline_reviews_a,10)
```

```
## # A tibble: 10 x 15
##     Year Month   Source  Destination Aircraft SeatType OverallRating Recommended
##    <dbl> <chr>   <chr>   <chr>       <chr>    <chr>            <dbl> <chr>
##  1  2011 October Chicago Kansas City A320, B~ Economy~             5 yes
##  2  2011 October London~ New York J~ Default  Premium~             4 no
##  3  2011 October SIN     LHR         Default  First C~             4 no
##  4  2012 August  LHR     BKK         A230     Economy~             8 yes
##  5  2012 August  LHR     HKG         Default  Economy~             4 no
##  6  2012 August  LHR     BKK         A230     Economy~             4 no
##  7  2012 August  LHR     JFK to LAX~ Default  Economy~             4 no
##  8  2012 August  LHR     JTR         Default  Busines~             6 yes
##  9  2012 August  YYZ     LHR         Default  Premium~             8 yes
## 10  2012 August  LHR     JTR         Default  Busines~             5 yes
## # i 7 more variables: SeatComfort <dbl>, CabinStaffService <dbl>,
## #   GroundService <dbl>, ValueForMoney <dbl>, FoodBeverage <dbl>,
## #   InflightEntertainment <dbl>, WifiConnectivity <dbl>
```

Exploratory analyses and modeling techniques

To address the research questions, I have used,

- Multi-variate regression analysis
- Monte-Carlo Simulation technique
- Prediction modeling using Contingency table(eg: Confusion Matrix)

Metrics used to evaluate the quality of the data analysis:

To evaluate the quality of data analysis, we use metrics such as Accuracy of the fitted/predicted model, p-Value for association between the exploratory and response variables, closeness of the fit to the actual model.

Data Analysis and Results

(1) Based on the airline review data, what can we conclude about the Overall satisfaction of the passengers using British Airways.

Multivariate Analysis:

To derive a conclusion to this question we reframe the question in statistical terms like,

Research Question: How do the ratings for "SeatComfort", "CabinStaffService", "GroundService", "ValueForMoney", "Food&Beverages", "InflightEntertainment", "Wifi&Connectivity" collectively influence the Overall rating?

To answer this we use multi-variate regression analysis and fit a model between the Overall Rating and "SeatComfort", "CabinStaffService", "GroundService", "ValueForMoney", "Food&Beverages", "InflightEntertainment", "Wifi&Connectivity".

```
# Fit a multivariate regression model
model <- lm(OverallRating ~ SeatComfort*CabinStaffService*GroundService*            ValueForMone
            WifiConnectivity, data = airline_reviews_a)
summary_info<-summary(model)
#adjusted r square
summary_info[9]
```

```
## $adj.r.squared
## [1] 0.8767275
```

```
#r square
summary_info[8]
```

```
## $r.squared
## [1] 0.8833612
```

```
# Extract minimum and maximum coefficients
min_coefficient <- min(summary_info$coefficients[, "Estimate"])
max_coefficient <- max(summary_info$coefficients[, "Estimate"])
cat("Minimum Coefficient:", min_coefficient, "\n")
```

```
## Minimum Coefficient: -5.365225
```

```
cat("Maximum Coefficient:", max_coefficient, "\n")
```

```
## Maximum Coefficient: 7.290639
```

```
# p-value: < 2.2e-16
```

From the analysis, the adjusted r square value is 0.87672748367711. From this we can infer how well the independent variables determine the dependent variable('OverallRating').

Also the R square value is 0.883361216547028 This high value of r square suggests that a larger proportion of the variability in the Overall Rating is explained by the influencing factors.

The p-value (2.2e-16) is less than 0.05 which indicates that the factors such as SeatComfort, CabinStaffService, GroundService, ValueForMoney, Food&Beverages, InflightEntertainment, Wifi&Connectivity significantly impact the overall rating given by British Airways passengers.

(2) Using existing information about the ratings for each aircraft type(such as A230, A330, Airbus 380, Boeing 747,etc) of British Airways, how can we predict that a future passenger recommends a certain flight/aircraft.

##Monte Carlo Simulation

To predict the possible future rating that a passenger using a certain flight might provide, we use Monte Carlo Simulations on the existing data and predict if a certain flight is to be recommended or not by a passenger flying in a particular aircraft might.

For the purpose of this project I am considering aircraft "A083".

```
# Define function to predict recommendation
predict_recommendation <- function(aircraft_number) {

  # Define threshold for recommendation
  recommendation_threshold <- 1

  # Filter data for the specified aircraft
 filtered_data<-airline_reviews_a[airline_reviews_a$Aircraft==aircraft_number, ]

  # Calculate average OverallRating for the aircraft
  average_overall_rating <- mean(filtered_data$OverallRating)

  # Predict recommendation based on average OverallRating
  if (average_overall_rating >= recommendation_threshold) {
    return("Recommend")
  }
  else {
    return("Not Recommend")
  }
}

# User input for aircraft number
aircraft_number <- "A083"
```

```
# Predict recommendation
pred_recommendation <- predict_recommendation(aircraft_number)

# Print the prediction
cat("Passenger flying in ",aircraft_number, "is likely to", pred_recommendation,
    " the flight.\n")
```

```
## Passenger flying in  A083 is likely to Recommend  the flight.
```

Based on this we can say that a future passenger is likely to recommend A083 flight for other passengers.

(3) Based on the airline reviews given by passengers, should British Airways be a recommended choice for passengers?

To determine the recommendation of British airlines from the data we create a confusion matrix and predict the recommendation choices to see how accurate is the chosen data for this research.

##Confusion matrix

```
# Assuming 'airline_reviews' is your data frame with relevant columns

# Convert 'Since Recommended' to a factor variable (if not already)
airline_reviews_a$Recommended <- as.factor(airline_reviews_a$Recommended)

# Create a confusion matrix
confusion_matrix <- table(Actual = airline_reviews_a$Recommended, Predicted = cut(airline_review
    c("Not Recommended", "Recommended")))

# Print the confusion matrix
print(confusion_matrix)
```

```
##       Predicted
## Actual Not Recommended Recommended
##    no             1089         136
##    yes              19         136
```

```
# Calculate accuracy, precision, recall, and F1 score
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
a<-confusion_matrix[2,2]
b<-sum(confusion_matrix[2, ])
precision <- a/b
recall <- confusion_matrix[2,2] / sum(confusion_matrix[, 2])
f1_score <- 2 * (precision * recall) / (precision + recall)

# Print performance metrics
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.8876812
```

```r
cat("Precision:", precision, "\n")
```

```
## Precision: 0.8774194
```

```r
cat("Recall:", recall, "\n")
```

```
## Recall: 0.5
```

```r
cat("F1 Score:", f1_score, "\n")
```

```
## F1 Score: 0.6370023
```

True Positive (TP): Reviews that were correctly predicted as "Recommended."

True Negative (TN): Reviews that were correctly predicted as "Not Recommended."

False Positive (FP): Reviews that were predicted as "Recommended" but were actually "Not Recommended."

False Negative (FN): Reviews that were predicted as "Not Recommended" but were actually "Recommended."

The model has a high overall accuracy, indicating its effectiveness in predicting recommendations. The precision rightly justifies the proportion of instances predicted as recommended that are actually recommended.

From the above metrics we can say that the dataset is appropriate for our research and the

The high number of false positives, suggests that the model might be overfitting or too sensitive to certain features. This could lead to recommending airlines that are not well-received by customers.

The recall for "Recommended" airlines is high, indicating that the model is good at identifying airlines with positive feedback.

(4) Make an informed decision about which type of the passengers should the airline focus for future improvements by looking at the existing dataset and finding the average rating given by passengers of a specific seat type(Business class, Economy class, Premium Economy, First Class).

To find which type of the passengers should the airline focus for future improvements, we look at the existing dataset and finding the average rating given by passengers of a specific seat type(Business class, Economy class, Premium Economy, First Class).

For the analysis we use the box plot to see where the average lies for each seattype.

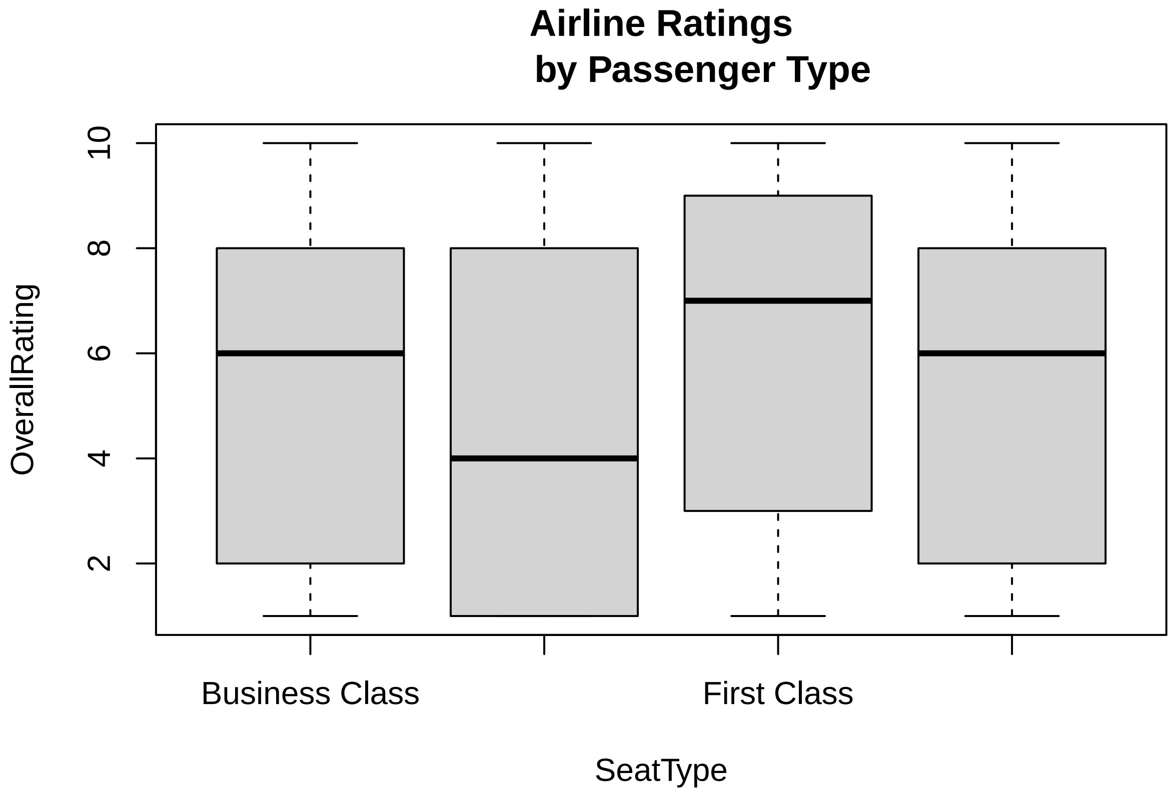
Rating by seat type:

```r
# Assuming 'airline_reviews' is your data frame and 'PassengerType' is a variable indicating pas
#Overall Average rating
ratings_by_passenger_type<-aggregate(OverallRating~SeatType,data=airline_reviews_a,
                                     FUN = mean)

ratings_by_passenger_type <- ratings_by_passenger_type |>
  rename("Average_Overall_Rating" = "OverallRating")
print(ratings_by_passenger_type)
```

```
##          SeatType Average_Overall_Rating
## 1  Business Class               5.476427
## 2   Economy Class               4.632943
## 3     First Class               6.430657
## 4 Premium Economy               5.360360
```

```r
# Example of a box plot
boxplot(OverallRating ~ SeatType, data = airline_reviews_a, main = "Airline Ratings
        by Passenger Type")
```



From the plot we can clearly say that passengers of economy class have generally given a lower Overall rating for British airways and conversely First class passengers seem to be satisfied with the services provided by British airways.

This clearly indicates that the management should focus more on the services provided for Economy class passengers.

Conclusion

The analysis sufficiently answers the research questions and we can draw the following valid conclusions from the research carried out.

- Various factors influence the satisfaction levels of a air passenger and SeatComfort, CabinStaffService, GroundService, ValueForMoney, Food&Beverages, InflightEntertainment, Wifi&Connectivity are some of the significant contributors to their feedback.
- In this research we build a predictive model which will determine if a passenger is likely to recommend a certain aircraft to a future traveller using flight A083 as an example.Based on this study a passenger with similar interests can be filtered to receive a suggestion from the model if the flight is recommended or not.
- We also have drawn meaningful conclusions on which areas of the seatype(Airline class) should British airways focus for better reach of customer satisfaction. Similarly various other factors can be assessed to identify the impact it can make on the overall rating.

Scope and generalizability of the analysis:

The scope of this project lies within the airline industry but the methods used for analysis can be implemented across any which deals with customer service and hospitality. The conclusions drawn deal with sentiments of the people therefore we can generalize the results as a whole for airlines with similar history.

Potential limitations and possibilities for improvement:

During the course of my study, I faced roadblocks in computing solutions for more thumping problems due to drawing conclusions from second hand data and lack of access to first-hand data.

A detailed research could have been carried to build a machine learning model which can access if a review provided by a passenger is positive or negative and classify any future reviews for ease business decisions from the given dataset but this involves NLP and other higher coding techniques.

This project deals with only a certain Airline which is both a boon and a bane as it is restricted to limited population size.

Dataset References:

- Airline Reviews dataset