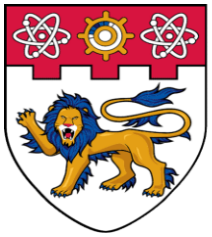


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MH3511 Data Analysis with Computer

Project Report

Group 8

Title: Helping British Airways make better data-driven decisions

Name	Matriculation Number
Chong Jinsheng	U2340252J
Haidar Prayata Wirasana	U2340701B
Low Yi Yin	U2340424B
Nasywa Hanan Huriyah	U2340940H
Tan Pang Boon	U2340262G

2024/2025 Semester 2

1. Introduction

With the rapid growth of the airline industry, understanding customer booking behaviour has become crucial for airlines to maximize their revenue. Identifying key factors, such as travel preferences, purchasing behaviour, and route characteristics that influence booking completion rate can provide valuable insights. These insights can be used by airlines to develop effective marketing strategies to attract or retain more customers. Hence, our objective of this study to explore the relationship between booking completion rates, and the various influencing factors.

In this project, we chose British Airways, the top-ranked airline in the United Kingdoms, and analysed their customer booking dataset. The dataset contains flight booking details, including variables such as the number of passengers, sales channel, booking origin, trip type, flight timings, and additional services requested, as well as whether a booking was successfully completed.

Through this analysis, we aim to answer the following two key questions:

1. Do factors such as preferences, sales channel, purchase lead time, flight duration, booking origin, and length of stay **show differences** in distribution or associations with customer's booking completion? **If an association is identified**, is it a positive or negative correlation?

2. Do factors like sales channel, length of stay, purchase lead time, and flight duration stay **show differences** in distribution or associations with customer preferences (e.g., in-flight meals, extra baggage, preferred seating)? **If an association is identified**, is it a positive or negative correlation?

This report will explore these questions through data visualization, statistical analysis, and predictive modelling using R language. We will draw appropriate conclusions based on our findings, using the most suitable methods supported by clear explanations and detailed elaborations. Finally, we will use these insights to offer actionable recommendations for British Airways.

2. Data Preparation

2.1 The Customer_booking Dataset

Our dataset is sourced from Kaggle and contains customer booking details for British Airways. You can download the data set from:

https://www.kaggle.com/datasets/minnikeswarrao/british-airways-customer-booking.

Attribute	Description	Data Type
num_passengers	Number of passengers travelling	Discrete
purchase_lead	Number of days between travel date and booking date	Discrete
length_of_stay	Number of days spent at destination	Discrete
flight_duration	Total duration of flight (in hours)	Continuous
flight_hour	Hour of flight departure	Ordinal
flight_day	Day of week of flight departure	Ordinal
sales_channel	Sales channel booking was made on	Nominal
trip_type	Trip Type (Round Trip, One Way, Circle Trip)	Nominal
route	Destination flight route	Nominal
booking_origin	Country from where booking was made	Nominal
wants_extra_baggage	If customer wanted extra baggage in the booking	Nominal
wants_preferred_seat	If customer wanted a preferred seat in the booking	Nominal
wants_in_flight_meals	If customer wanted in-flight meals in the booking	Nominal
booking_complete	Flag indicating if the customer completed	Nominal

Table 1: Description of Attributes

2.2 Attribution Selection

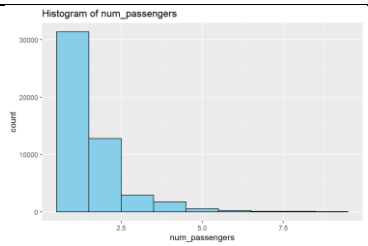
To answer our two key questions, we will first select the relevant attributes for further analysis and remove the "route", “flight day” and "flight hour" attributes.

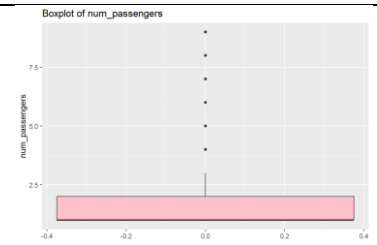
2.3 Data Analysis

We will begin by visualizing the attributes using a histogram or boxplot to identify any anomalies or outliers, which will be addressed in **Section 2.4: Data Cleaning.**

2.3.1 Summary statistics for the main variable of num_passengers

Plots of *num_passengers*

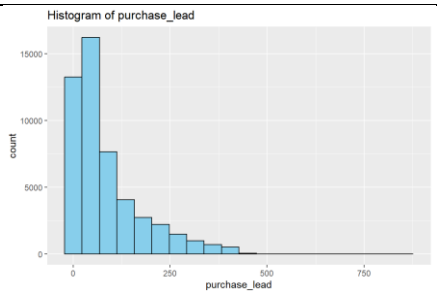


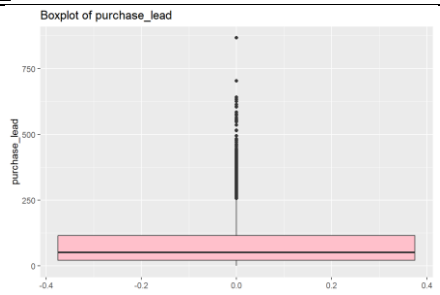



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   1.000   1.000   1.591   2.000   9.000
```

2.3.2 Summary statistics for the main variable of purchase_lead

Plots of *purchase_lead*





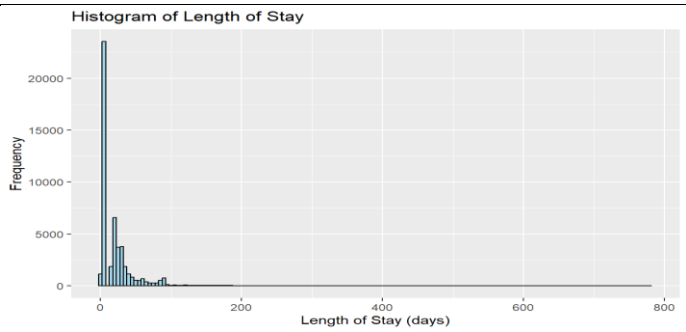

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00   21.00   51.00   84.94  115.00  867.00
```

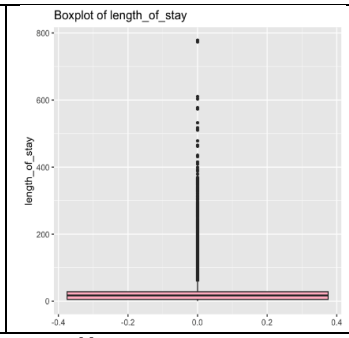
Anomaly Observed

The distribution of the dataset "Purchase_lead" is clearly right skewed. The highest purchase lead exceeds 750 days, meaning a customer completed their booking more than two years before the flight. This is extremely unusual, and we will address this anomaly later in the **Section 2.4**.

2.3.3 Summary statistics for the main variable of length_of_stay

Plots of *length_of_stay*



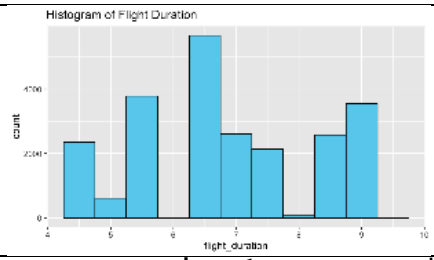
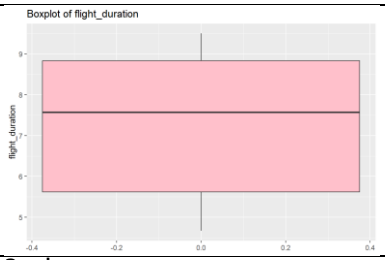



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    5.00   17.00   23.04   28.00  778.00
```

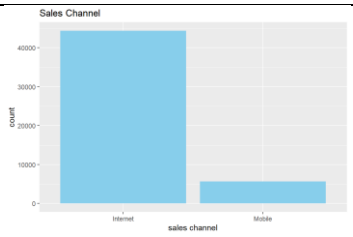
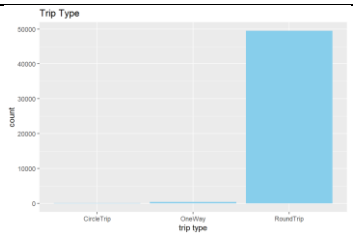
Anomaly Observed

In the histogram of Length of Stay, there is an unusual gap in the distribution. Upon zooming in, we observed that this gap is due to missing data in our dataset, which we will address in **Section 2.4**

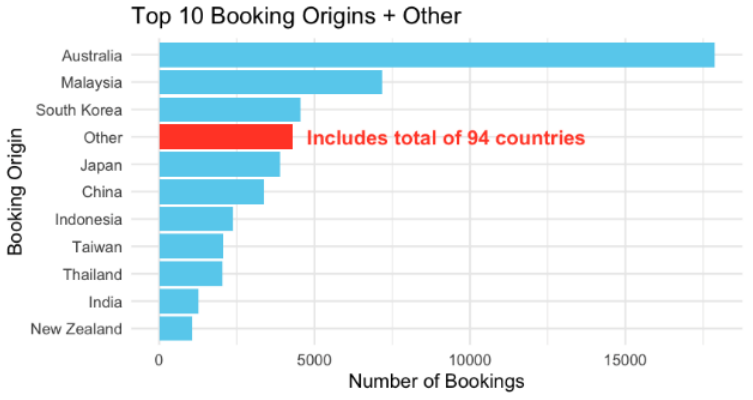
2.3.4 Summary statistics for the main variable of flight_duration

Bar plot of *flight_duration*	
	
Min. 1st Qu. Median Mean 3rd Qu. Max. 4.670 5.620 7.570 7.278 8.830 9.500	

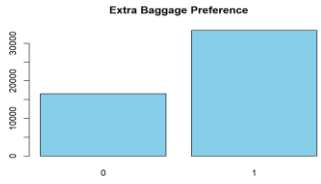
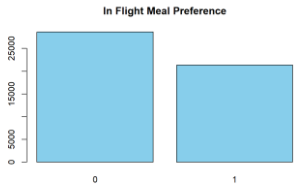
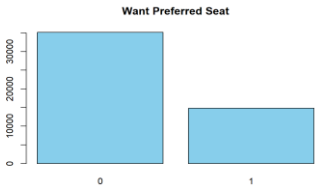
2.3.5 Summary statistics for the main variable of sales_channel and trip_type

Bar chart of *sales_channel*	Bar chart of *trip_type*
	
Internet: 44382 Mobile: 5618	CircleTrip 116 OneWay 387 RoundTrip 49497
Anomaly Observed Unusual distribution, heavily on RoundTrip	

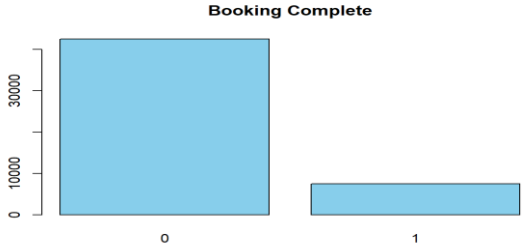
2.3.6 Summary statistics for the main variable of booking_origin

Bar chart of *booking_origin*	
	Australia:17872 Malaysia: 7174 South Korea: 4559 Other:4303 Japan: 3885 China:3387 Indonesia: 2369 Taiwan: 2077 Thailand: 2030 India: 1270 New Zealand:1074
Note: As there are 104 countries in the Booking Origin variable, it may not be feasible to plot all of them in a bar chart. Therefore, we combined the bottom 10 into a single group for easier illustration. However, for further analysis, we will select only a certain number of top countries and provide our justification in **Section 2.4.**	

2.3.7 Summary statistics for the main variable of perferences

Bar chart of *wants_extra_baggage*	Bar chart of *wants_in_flight_meal*	Bar plot of *wants_preferred_seat*
		
0 (No): 16562, 1 (Yes): 33439	0 (No): 28643, 1 (Yes): 21357	0 (No): 35152, 1 (Yes): 14848

2.3.8 Summary statistics for the main variable of booking_complete

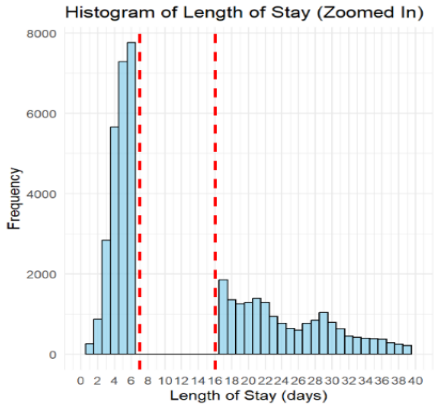
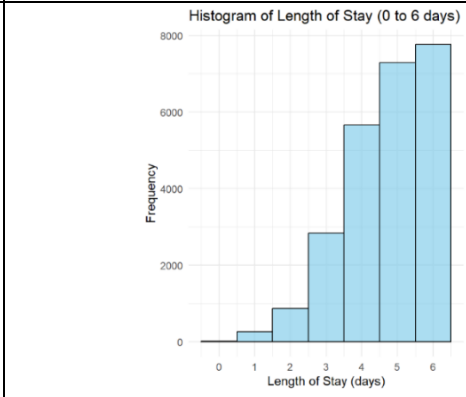
Bar chart of *booking_complete*

0: (No) 42522 , 1: (Yes) 7478

2.4 Data Cleaning

With a better understanding of the data distribution in our 10 attributes. We proceed to clean the data based on the analysis in the previous section by resolving missing data issue, removing outliers, and filtering the data to include only relevant entries. The code can be found in Appendix A.

2.4.1 Missing Data:

In the *length_of_stay* dataset, we observed a significant gap from the histogram. Upon closer inspection, there is missing data for *length_of_stay* values between 7 and 15(inclusive). The corresponding figure is shown below.

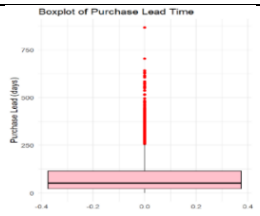
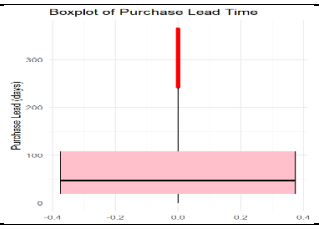
Before	After
	

This gap is significant in the context of length of stay for travellers, so we decided to filter out records with *length_of_stay* > 6.

After this filtering, 24673 records remained, which sufficiently addresses any sample size concerns. Additionally, research shows that 60% of people are more impulsive to buy plane tickets when there is a deal (Loo, 2024). Hence, by setting our target audience to short-stay travellers, we aim to uncover insights that may improve booking completion rates and ultimately drive higher revenue for the company. Aside from the *length_of_stay* attribute, there are no missing values in any of the other attributes in our dataset.

2.4.2 Outlier Removal:

In the *purchase_lead dataset*, we observed a significant number of outliers, as shown in the figure below.

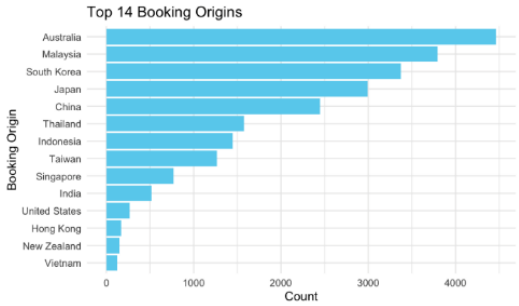
Before	After
	

We observed that outliers span from 250 days up to as long as 790 days. It is rare for travellers to book their flights two years in advance, especially since British Airways only publishes flight schedules for up to one year in advance (British Airways, n.d.). We believe that travellers that who purchase tickers more than a year in advance deviate from usual customer behaviour. Therefore, we chose to filter out records with *purchase_lead* greater than 365 days, **focusing only on records within a one-year purchase period**.

Insights derived from this dataset will be more intuitive and easily quantifiable for the company's stakeholders.

2.4.3 Filtering of countries:

In the *booking_origin* dataset, there were over 100 discrete booking origins, with some having significantly lower booking completions. Therefore, we decided to filter for the top 14 countries with a booking completion rate of at least 100, reducing the number of origins in our analysis and ensuring greater accuracy. The filtered *booking_origin* dataset is represented in the bar chart below.



2.4.3 Filtering of Trip Type:

In the *Trip_Type* dataset, more than 99% of the dataset consist of RoundTrip, while the remaining 1% comprising CircleTrip and OneWay. This imbalance may be problematic in our analysis. As such, we will not be using this attribute.

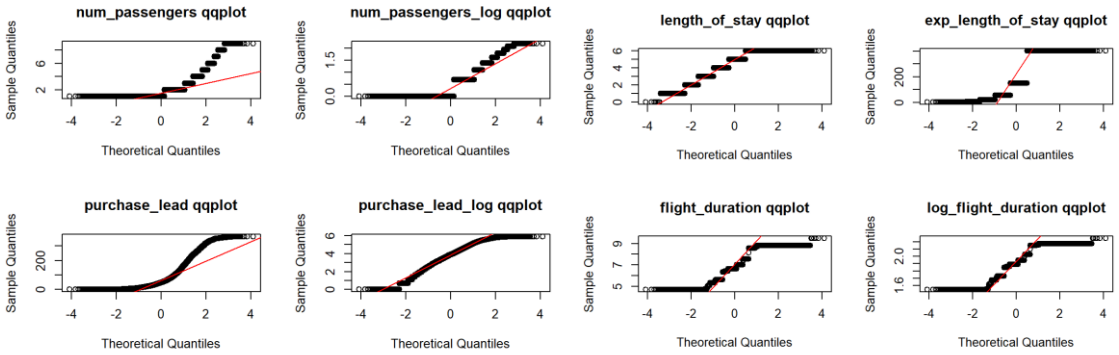
2.5 Data Normalization

The next step will be normalizing our numerical data. We apply the QQ plot to assess the normality of the dataset. If the data is not normally distributed, we will apply the appropriate log transformation and check for normality again. The process of this can be shown in Table 2 below.

Attribute	Before Transformation	Transformation Function	After Transformation
num_passengers	Not Normal	$\log(x)$	Not Normal
purchase_lead	Not Normal	$\log(x+1)$	Not Normal
length_of_stay	Not Normal	e^x	Not Normal
flight_duration	Not Normal	$\log(x)$	Not Normal

Table 2: Normalization Transformations

The illustration of the QQ plot, and the transformation of the plot before and after transformation, is shown below. Please refer to Appendix A for our code.



2.5 Final Dataset for Analysis

Based on the data preparation process, the dataset was reduced from 50,000 observations with 14 variables to 23,350 observations with 10 variables.

3. Data Analysis and Testing

In this section, we will perform the appropriate statistical test using the cleaned dataset and answer our two key questions:

3.1: Do factors such as preferences, sales channel, purchase lead time, flight duration, booking origin, and length of stay show differences in distribution or associations with customer's booking completion? If an association is identified, is it a positive or negative correlation?

3.1.1: Methodology

To answer this question. We will conduct the following three-step approach:

1. **Chi-Square Test:** We first examine the relationship between one of the factors and their booking completion rate using Chi-Square test at a 0.05 significance level.
   H_0: There is no association between the two variables.
   H_1: There is association between the two variables.

   If we obtain a p-value that is lesser than 0.05, then we reject the H_0, indicating a significant association between them. However, the test does not reveal the nature of the relationship as it could be either a positive or negative correlation.

2. **Proportion Table:** To understand the correlation better, we examine the row-wise proportions to assess whether there is a positive correlation.

3. **Logistic Regression Analysis:** We further analyse the relationship using Logistic Regression Analysis, and to understand the odds ratio in each unit factor increase if there is a significant association at a 0.05 significance level.
   H_0: The coefficient β_1 is equal to zero*, meaning there is no significant relationship between the variable and the response variable.*

H_1: The coefficient β_1 is not equal to zero, meaning there is a significant relationship between the variable and the response variable.

For the variables flight duration, length of stay and purchase lead time, which are numerical data, we adopt a different two-step approach:

1. **Wilcoxon Rank Sum Test:** To compare whether there is a statistically significant difference in the distribution of each numerical variables between those who completed their booking and those who did not.

2. **Logistic Regression Analysis:** We further analyse the relationship using Logistic Regression Analysis, and to understand the odds ratio in each unit factor increase.

3.1.2: Outcomes

The codes to achieve the outcomes can be found in Appendix B

3.1.2.1: Selecting an Extra Baggage and Booking Completion

1. Based on the Chi-Square Test, there is a significant association between selecting an extra baggage and completing a booking.

2. Based on the Proportion Table, we observe a positive correlation between Booking Completion and Selecting an Extra Baggage.

	0	1
Extra baggage No	0.8656999	0.1343001
Extra baggage Yes	0.7602156	0.2397844

wants_extra_baggage vs booking completion row wise proportion table

3. In our Logistic Regression Analysis, we concluded that booking completion will happen twice as likely if an extra baggage is selected.

Repeating Tests for Preferred Seating and In-Fight Meals

1. We observed that there is a significant association between selecting preferred seats or in-flight meals and completing a booking.
2. We also examined the row-wise proportions again observe yet again another positive correlation.

	0	1
Preferred Seat No	0.8287532	0.1712468
Preferred Seat Yes	0.7558348	0.2441652

wants_preferred_seat vs booking completion row wise proportion table

	0	1
In-Flight Meals No	0.8304730	0.1695270
In-Flight Meals Yes	0.7687719	0.2312281

wants_in_flight_meals vs booking completion row wise proportion table

3. Based on the result of the Logistic Regression analysis. Customers who select a preferred seat are 1.56 times more likely to complete their booking. Similarly, for in-flight meals, customers who indicate a preference are 1.4 times more likely to complete their booking compared to those who do not.

Findings:
There is a positive association between Customer preferences and customer's booking completion.

Among all the three preferences analysed, select extra baggage has the highest influence on customer's booking completion. This could be because if a customer is willing to pay for add-ons like baggage, it would mean that they are more committed to their travel plans, or they perceive higher value in British Airways' offerings compared to those of other airlines.

3.1.2.2: Sales Channel, Booking Origin and Booking Completion

1. Based on the Chi-Square Test, there is a significant association between Sales Channel, Booking Origin and completing a booking.

2. Based on the Proportion Table, we observed higher booking completion for bookings made via Internet than Mobile. Additionally, we found that the top three countries with the highest booking completion rates are Malaysia, Vietnam, and Singapore.

	0	1
Internet	0.7983277	0.2016723
Mobile	0.8643767	0.1356233

Sales channel vs booking completion row wise proportion table

	0	1
Malaysia	0.60812236	0.39187764
Vietnam	0.61904762	0.38095238
Singapore	0.71409922	0.28590078
Full table can be found in appendix		

Booking Origin vs booking completion row wise proportion table

3. In our Logistic Regression analysis for Sales Channel, it shows that the Internet channel is statistically significant with the P value of <2e-16, with odds ratio of e^0.476 = 1.61, meaning customers who select Internet are about 1.61 times more likely to complete their booking.

   For Booking Origin, it can be observed that all the variables except New Zealand is statistically significant, and the highest 3 odds ratio of e^ 2.22, e^ 2.17, e^1.74 for Malaysia, Vietnam, Singapore respectively (Australia as the reference category). Therefore, the findings in this analysis of Sales Channel and Booking Origin align with our observations with the row-wise proportion table.

Findings:

There is a positive association between Sales Channel or Booking Origin and customer's booking completion.

For Sales Channel, customers using the internet channel are more likely to complete their bookings, and this is likely because the internet platform offers a larger display, allowing customers to easily review more details about the flight and compare deals across different products. In contrast, customers using mobile users may be less serious and distracted, resulting in higher chance of abandoning the booking process before completion.

For Booking Origin, the countries with the highest booking completion rates are Malaysia, Vietnam and Singapore. This could be reasons such as there are more individuals travelling to Europe for business related matters.

3.1.2.3: Purchase lead time and Booking Completion

Since the distribution of purchase_lead is not normal and log transformation does not normalize the data, we use a non-parametric test: **Wilcoxon Rank-Sum Test**.

1. Based on the results of the Wilcoxon Rank Sum Test, with the p value of 0.475, we fail to reject H_0, thus, there is insufficient evidence to conclude that purchase lead and booking completion has statistically significant difference in their distribution. Therefore, no further tests are required.

3.1.2.4: Flight Duration, Length of Stay and Booking Completion

1. Based on the results of the Wilcoxon Rank Sum Test, the distribution of Flight Duration, Length of Stay of those who completed their booking compared to those who do not are different.

2. Based on the results of the Logistic Regression analysis, with each unit increase in Length of Stay, the odds of booking completion increase by 5.5%, whilst with each unit increase in Flight Duration, the odds of booking completion decrease by about 10.3%.

	Estimate	Pr(>|z|)	Odds ratio
Flight Duration:	-0.10825	<2e-16	0.897
Length of Stay:	0.05417	0.000136	1.055

Findings:

The distributions of Flight Duration and Length of Stay differ significantly between users who completed their bookings and those did not. This suggest that Flight Duration and Length of Stay could potentially influence customer's likelihood to complete their bookings.

The Logistic Regression results suggest that Length of Stay impacts on customer's likelihood to complete their booking ─ the longer you stay, the higher the likelihood of completion. This

could be because a longer stay likely indicates a more committed travel plan, and people who are stay longer may be more organized and committed to completing their bookings.

On the other hand, the results suggest the opposite for Flight Duration, where the longer your flight is, the lower likelihood of completing booking.. This is likely because long flights are less appealing to customers and possibly more expensive, making them less likely to complete their bookings.

3.2. Do factors like sales channel, length of stay, purchase lead time, and flight duration stay show differences in distribution or associations with customer preferences (e.g., in-flight meals, extra baggage, preferred seating)? If an association is identified, is it a positive or negative correlation?

3.2.1: Methodology

To answer this question. We will first conduct the following three-step approach for sales channel:

1. **Chi-Square Test:** We first examine the relationship between one of the factors and their booking completion rate using Chi-Square test at a 0.05 significance level .
   H_0: There is no association between the two variables.
   H_1: There is association between the two variables.

   If we obtain a p-value that is lesser than 0.05, then we reject the H_0, indicating a significant association between them. However, the test does not reveal the nature of the relationship.

2. **Proportion Table:** To understand the correlation better, we examine the row-wise proportions to assess whether there is a positive correlation or not.

3. **Logistic Regression Analysis:** We further analyse the relationship using Logistic Regression Analysis, to understand the odds ratio in each unit factor increase if there is a significant association at a 0.05 significance level.
   H_0: The coefficient β_1 is equal to zero*, meaning there is no significant relationship between the variable and the response variable.*
   H_1: The coefficient β_1 is not equal to zero, meaning there is a significant relationship between the variable and the response variable.

As for the *variables length of stay, purchase lead time and flight duration*. We will conduct the following two-step approach.

1. **Wilcoxon Rank Sum Test:** To compare whether there is a statistically significant difference in the distribution of each numerical variables between those who want to purchase additional services and those who do not.

2. **Logistic Regression Analysis:** We further analyse the relationship using Logistic Regression Analysis to understand the odds ratio in each unit factor increase.

3.2.2: Outcomes

The codes to achieve the outcomes can be found in Appendix C

3.2.2.1: Sales Channel and Customer Preferences

1. There is a statistically significant association between sales channel and each of customers preferences, namely: extra baggage, in flight meal, and preferred seat based on the result from Chi-Square Test. Below is the p-value from each test:

	Extra Baggage	Preferred Seat	In Flight Meal
Sales Channel	1.107e-05	1.234e-06	0.01839

2. Based on the Proportion Table, we observe that customers who bought their ticket via the internet are slightly more likely to purchase extra baggage or in-flight meal compared to those who use the mobile channel. On the other hand, those who bought their ticket via mobile channel are more likely to purchase preferred seat compared to those who use internet channel. However, it is important to note that these differences are less than 5%, which is very small in practice.

	wants_extra_baggage		*wants_in_flight_meal*		*wants_preferred_seat*	
	0	1	0	1	0	1
Internet	0.445952	0.554048	0.630751	0.369248	0.719671	0.280328
Mobile	0.486822	0.513177	0.652058	0.347942	0.678708	0.321291

sales_channel vs preference row wise proportion table

3. Based on the results of the Logistic Regression analysis, customers using the internet channel has a higher likelihood of selecting extra baggage than mobile channel. On the other hand, customers using the mobile channel has a higher likelihood of selecting in-flight meals and their preferred seat than internet channel.

	Estimate	Pr(>|z|)	Odds ratio
Internet$ExtraBaggage:	0.21704	<2e-16	(intercept)
Mobile$ExtraBaggage:	-0.16432	1.03e-05	0.8484
Internet$InFlightMeals:	-0.53544	<2e-16	(intercept)
Mobile$InFlightMeals:	-0.09266	0.0175	0.9115
Internet$PreferredSeat:	-0.94283	<2e-16	(intercept)
Mobile$PreferredSeat:	0.19499	1.14e-06	1.2152

Findings:

The channel used by customers to complete their bookings does have an association with customer preferences. Customers using the internet channel are more likely to purchase extra baggage, and this could be due to having more screen time to review baggage options and consider their travel needs in detail. On the other hand, customers using the mobile channel are more likely to purchase in flight meal and preferred seats, than Internet Channel. This could because mobile users, may be booking closer to departure, and they might have certain cravings for certain food or last-minute preferences for comfortable seats. However, it's

worth noting that the overall association between mobile bookings and in-flight meals is still negative.

3.2.2.2: Length of Stay and Customer Preferences

1. Based on the results of the Wilcoxon Rank Sum Test, the distribution of length of stay is statistically different for passengers who want extra baggage or in-flight meal compared to those who do not. However, there is insufficient evidence to conclude that there is a statistically difference in distributions for those who want preferred seats and those who do not. Below is the p-value from each test:

	Extra Baggage	In Flight Meal	Preferred Seat
Length of Stay	< 2.2e-16	3.367e-05	0.8217

2. Based on the result of the Logistic Regression analysis, with each additional day in length of stay, the odd of purchasing extra baggage increases by 28.8% and the odds of purchasing in-flight meals increases by 4.42%.

	Estimate	Pr(>|z|)	Odds ratio
ExtraBaggage:	0.25330	<2e-16	1.2882
InFlightMeals:	0.04333	<2e-16	1.0442

Findings:

The distributions of Length of Stay for those who select extra baggage or in-flight meals significantly differs from those who do not. This suggest that Length of Stay could potentially influence customer's likelihood of selecting an extra baggage or in-flight meals.

The Logistic Regression result suggest that Length of Stay has a significant impact on the customer's likelihood to choose extra baggage or in-flight meals. They longer the stay, the higher likelihood of customers choosing extra baggage or in-flight meals. This is likely because longer travel stays may require more luggage space due to extra packing needs, and customers traveling for longer periods, often for holidays, would be more inclined to purchase in-flight meals to enhance their comfort.

On the other hand, there appears to be no significant difference in the distributions between Length of Stay and the selection of preferred seats based on the Wilcoxon Rank Sum Test. This could suggest that seat preferences are driven by factors such as flight duration, personal comfort, or availability.

3.2.2.3: Purchase Lead and Customer Preferences

1. Based on the results of the Wilcoxon Rank Sum Test, the distribution of purchase lead is statistically different for passengers who want extra baggage or preferred seat compared to those who don't. However, there is insufficient evidence to conclude that there is a statistically difference in distributions for those who want in-flight meal and those who do not. Below is the p-value from each test:

	Extra Baggage	In-Flight Meal	Preferred Seat
Purchase Lead	< 2.2e-16	0.5432	0.04174

2. Based on the result of the Logistic Regression analysis, with each additional day in purchase lead, the odd of purchasing an extra baggage increases by 0.07% and the odd of selecting a preferred seat decreases by 0.06%.

	Estimate	Pr(>|z|)	Odds ratio
ExtraBaggage:	0.0007397	2.84e-06	1.0007
PreferredSeat:	-0.0006103	0.000519	0.9994

Findings:

The distribution of Purchase Lead for those who select extra baggage or preferred seats options significantly differs from those who do not. This suggest that Purchase Lead could potentially influence customer's likelihood of selecting an extra baggage or preferred seats.

The Logistic Regression result suggest that Purchase Lead has an impact on the customer's likelihood to choose extra baggage or preferred seats, although the impact is not very strong. This could be because customers who plan their trips earlier may prioritize budgeting or flexibility, while last-minute bookers might be more concerned with securing a comfortable seat quickly.

On the other hand, there appears to be no significant difference in the distributions between Purchase Lead and the selection of in-flight meals based on the Wilcoxon Rank Sum Test. This could suggest that meal preferences are likely influenced by other factors such as personal habits or flight duration.

3.2.2.4: Flight Duration and Customer Preferences

1. Based on the results of the Wilcoxon Rank Sum Test, the distribution of flight duration is statistically different for passengers who want preferred seat or in-flight meals compared to those who don't. However, there is insufficient evidence to conclude that there is a statistically difference in distributions for who want extra baggage and those who don't. Below is the p-value from each test:

	Extra Baggage	Preferred Seat	In Flight Meal
Purchase Lead	0.3206	< 2.2e-16	< 2.2e-16

2. Based on the results of the Logistic Regression analysis, with each unit increase in flight duration, the odd of purchasing in-flight meals increases by 18.3% and the odd of purchasing preferred seat increases by 13.4%.

	Estimate	Pr(>|z|)	Odds ratio
InFlightMeals:	0.16822	<2e-16	1.1832
PreferredSeat:	0.12578	<2e-16	1.1340

Findings:

The distributions of Flight Duration for those who select preferred seats or in-flight meals significantly differs from those who do not. This suggest that Flight Duration could potentially influence customer's likelihood of selecting preferred seats or in-flight meals.

The Logistic Regression results suggest that flight duration has a significant impact on the customer's likelihood to choose inflight meals or their preferred seating. The longer the flight duration is, the higher likelihood of customer choosing inflight meals or their preferred seats. is likely because customers see a bigger need for comfort and convenience in longer flights, and they would prefer to secure better seats and more inclined to purchase meals to stay nourished during the journey.

On the other hand, there appears to be no significant difference in the distributions between Flight Duration and the selection of extra baggage based on the Wilcoxon Rank Sum Test. This could suggest that baggage decisions are likely influenced by other factors such as trip length, packing habits or individual travel needs.

4. Conclusion and Discussion

Our project revealed several key findings through a series of statistical tests, aimed at answering two main questions. The first question focused on identifying factors associated with customer's booking completion, while the second explored the factors influencing customer preferences.

Factors Associated with Booking Completion

1. **Customer Preferences**: Select an extra baggage had the strongest positive association with booking completion.
2. **Sales Channel and Booking Origin:** Customers who booked their flights via the internet channel were more likely to complete bookings. The top countries with the highest bookings completion rates were Malaysia, Vietnam, and Singapore.
3. **Trip Details:** Longer length of stay increases the likelihood of booking completion, while longer flight durations had a negative Association.

Factors Associated with Customer Preferences

1. **Sales Channel:** Users who book their flights via Internet Channel are more likely to purchase extra baggage, while users who book their flights via Mobile Channel were more likely to select in-flight meals and preferred seats.
2. **Length of Stay:** There is positive association with extra baggage and in-flight meals, but there is no significant difference in the distribution of the length of stay of those who want and do not want preferred seats.
3. **Purchase Lead:** There is a slight positive association with extra baggage, and slight negative with preferred seats, but there is no significant difference in the distribution of the length of stay of those who want and do not want inflight meals.
4. **Flight Duration:** There is positive association with in-flight meals and preferred seats, but there is no significant difference in the distribution of length of stay and those who want and do not want extra baggage.

Based on these insights, British Airways can leverage on this information to develop effective marketing strategies to attract or retain more customers by:

1. British Airways can highlight the selection of extra baggage more prominently during the booking process. They could also introduce limited-time bundle deals that include extra baggage at a discounted fee, especially for customers with longer purchase lead times
2. For long haul flights, discounted prices for in-flight meals and preferred seats can be offered to encourage additional purchases.
3. Since internet booking show higher completion rates, British airways should focus on improving their website by making the booking process simpler, faster, and more intuitive.
4. Lastly, the airline can target the three markets like Malaysia, Vietnam, and Singapore by offering region-specific deals or promotions tailored to their local preferences.

Applying statistical test concepts to a real-world challenge was a valuable and insightful learning experience. Although we may not have the opportunity to formally present our findings to British Airways, we hope these insights demonstrate the power of data-driven decision-making and inspire smarter and effective strategies within the airline industry.

5. Appendix A

```
# MH3511 Project Chapter 2

#Load data
getwd()
setwd("C:/Users/BC2404/Desktop/MH3511/Labs")
df <- read.csv('customer_booking.csv')

#convert binary categorical variables into factor (it was classified as int before)
df$wants_extra_baggage <- factor(df$wants_extra_baggage, levels = c(0, 1))
df$wants_preferred_seat <- factor(df$wants_preferred_seat, levels = c(0, 1))
df$wants_in_flight_meals <- factor(df$wants_in_flight_meals, levels = c(0, 1))

#library
library(ggplot2)
library(dplyr)
##############################################################################
#########

# 2.3.1 Summary statistics for the main variable of num_passengers

# Histogram
ggplot(df, aes(x = num_passengers)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  ggtitle("Histogram of num_passengers")

#original data is heavily right skewed

# Boxplot
ggplot(df, aes(y = num_passengers)) +
  geom_boxplot(fill = "pink") +
  ggtitle("Boxplot of num_passengers")

# Summary
summary(df$num_passengers)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1.000   1.000   1.000   1.591   2.000   9.000

##############################################################################
#########

# 2.3.2 Summary statistics for the main variable of purchase_lead

# Histogram
ggplot(df, aes(x = purchase_lead)) +
  geom_histogram(binwidth = 45, fill = "skyblue", color = "black") +
  ggtitle("Histogram of purchase_lead")

# Boxplot
ggplot(df, aes(y = purchase_lead)) +
  geom_boxplot(fill = "pink") +
  ggtitle("Boxplot of purchase_lead")
```

```
# Summary
summary(df$purchase_lead)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.00   21.00   51.00   84.94  115.00  867.00

##############################################################################
#########

# 2.3.3 Summary statistics for the main variable of length_of_stay

# Histogram
ggplot(df, aes(x = length_of_stay)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  ggtitle("Histogram of length_of_stay")

# Boxplot
ggplot(df, aes(y = length_of_stay)) +
  geom_boxplot(fill = "pink") +
  ggtitle("Boxplot of length_of_stay")

# Summary
summary(df$length_of_stay)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.00    5.00   17.00   23.04   28.00  778.00

##############################################################################
#########

# 2.3.4 Summary statistics for the main variable of flight_duration

# Histogram
ggplot(df, aes(x = flight_duration)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  ggtitle("Histogram of flight_duration")

# Boxplot
ggplot(df, aes(y = flight_duration)) +
  geom_boxplot(fill = "pink") +
  ggtitle("Boxplot of flight_duration")

# Summary
summary(df$flight_duration)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 4.670   5.620   7.570   7.278   8.830   9.500

##############################################################################
#########

# 2.3.5 Summary statistics for the main variable of sales_channel and trip_type

# Bar plot of sales_channel
ggplot(df, aes(x = sales_channel)) +
  geom_bar(fill = "skyblue") +
```

```
  labs(title = "Sales Channel", x = "sales channel", y = "count")

# Summary of sales_channel
table(df$sales_channel)
# Internet   Mobile
# 44382     5618

# Bar plot of trip_type
ggplot(df, aes(x = trip_type)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Trip Type", x = "trip type", y = "count")

# Summary of trip_type
table(df$trip_type)
# CircleTrip     OneWay  RoundTrip
# 116        387      49497

##########################################################################
#########

# 2.3.6 Summary statistics for the main variable of booking_origin

top10 <- df %>%
  count(booking_origin, sort = TRUE) %>%
  slice_head(n = 10) %>%
  pull(booking_origin)

df1 <- df %>% mutate(origin_grouped = ifelse(booking_origin %in% top10, booking_origin,
"Other"))

origin_summary <- df1 %>%
  count(origin_grouped) %>%
  arrange(desc(n))

n_other_countries <- df1 %>%
  filter(!(booking_origin %in% top10)) %>%
  summarise(n = n_distinct(booking_origin)) %>%
  pull(n)

ggplot(origin_summary, aes(x = reorder(origin_grouped, n), y = n,
                           fill = ifelse(origin_grouped == "Other", "Other", "Top 10"))) +
  geom_bar(stat = "identity") +
  geom_text(data = origin_summary %>% filter(origin_grouped == "Other"),
            aes(label = paste0("Includes total of ", n_other_countries, " countries")),
            hjust = -0.05, size = 4, fontface = "bold", color = "tomato") +
  scale_fill_manual(values = c("Top 10" = "skyblue", "Other" = "tomato")) +
  coord_flip(clip = "off") +
  labs(
    title = "Top 10 Booking Origins + Other",
    x = "Booking Origin",
    y = "Number of Bookings",
    fill = "Group"
  ) +
```

```
  theme_minimal() +
  theme(plot.margin = margin(5.5, 50, 5.5, 5.5))  # extra right margin for label

print(origin_summary)
# Top 10 + Others
# origin_grouped     n
# 1       Australia 17872
# 2        Malaysia  7174
# 3     South Korea  4559
# 4           Japan  3885
# 5           China  3387
# 6       Indonesia  2369
# 7          Taiwan  2077
# 8        Thailand  2030
# 9           Other  1816
# 10          India  1270
# 11    New Zealand  1074

##########################################################################
#########

# 2.3.7 Summary statistics for the main variable of Preferences

# Bar plot of wants_extra_baggage
barplot(table(df$wants_extra_baggage), main= "Extra Baggage Preference", col =
"skyblue")

# Summary of wants_extra_baggage
table(df$wants_extra_baggage)
#    0     1
# 16561 33439

# Bar plot of wants_in_flight_meal
barplot(table(df$wants_in_flight_meals), main= "In Flight Meal Preference", col =
"skyblue")

# Summary of wants_in_flight_meal
table(df$wants_in_flight_meals)
#    0     1
# 28643 21357

# Bar plot of wants_preferred_seat
barplot(table(df$wants_preferred_seat), main= "In Flight Meal Preference", col =
"skyblue")

# Summary of wants_preferred_seat
table(df$wants_preferred_seat)
#     0     1
#  35152 1484

##########################################################################
#########
```

```
# 2.3.8 Summary statistics for the main variable of booking_complete

# Bar plot of booing_complete
barplot(table(df$booking_complete), main= "Booking Complete", col = "skyblue")

# Summary of booking_complete
table(df$booking_complete)
#    0     1
# 42522  7478

##########################################################################
#########

##########################################################################
#########

#2.4.1 Missing Data: length_of_stay

# Histogram of length_of_stay with zoomed-in view (Before/Uncleaned)
ggplot(df, aes(x = length_of_stay)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black", alpha = 0.7) +
  scale_x_continuous(limits = c(0, 40), breaks = seq(0, 40, by = 2)) +  # Adjust limits for
zoom
  labs(title = "Histogram of Length of Stay (Zoomed In)",
       x = "Length of Stay (days)",
       y = "Frequency") +
  theme_minimal() +
  geom_vline(xintercept = c(7, 16), color = "red", linetype = "dashed", size = 1)

# Histogram of length_of_stay (After/Cleaned)
df1 <- df %>%
  filter(length_of_stay >= 0 & length_of_stay <= 6)

ggplot(df1, aes(x = length_of_stay)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black", alpha = 0.7) +
  scale_x_continuous(breaks = seq(0, 6, by = 1)) +  # Set breaks
  labs(title = "Histogram of Length of Stay (0 to 6 days)",
       x = "Length of Stay (days)",
       y = "Frequency") +
  theme_minimal()

total_count <- nrow(df1)
print(total_count)
# 24673 records remained after deletion

##########################################################################
#########

# 2.4.2 Outlier Removal: purchase_lead
# Box plot of purchase_lead (Before/Uncleaned)
ggplot(df, aes(y = purchase_lead)) +
  geom_boxplot(fill = "pink", color = "black", outlier.color = "red") +
  labs(title = "Boxplot of Purchase Lead Time",
```

```
      y = "Purchase Lead (days)") +
  theme_minimal()

# Box plot of purchase_lead (After/Cleaned)
df1 <- df %>%
  filter(length_of_stay >= 0 & length_of_stay <= 6) %>%  # short stays only
  filter(purchase_lead <= 365)

ggplot(df1, aes(y = purchase_lead)) +
  geom_boxplot(fill = "pink", color = "black", outlier.color = "red") +
  labs(title = "Boxplot of Purchase Lead Time",
       y = "Purchase Lead (days)") +
  theme_minimal()

##########################################################################
#########

# 2.4.3 Filtering of Countries

top14 <- df1 %>%
  count(booking_origin, sort = TRUE) %>%
  slice_head(n = 14)

ggplot(top14, aes(x = n, y = reorder(booking_origin, n))) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Top 14 Booking Origins",
       x = "Count",
       y = "Booking Origin") +
  theme_minimal()

##########################################################################
#########

# Finalized Cleaning
df <- df %>%
  filter(length_of_stay >= 0 & length_of_stay <= 6) %>%  # short stays (0-6days)only
  filter(purchase_lead <= 365) %>%                  # remove extreme purchase lead
  select(-route, -flight_hour)                    # drop unused columns

# get top 14 countries
top14 <- df %>% count(booking_origin, sort = TRUE) %>% slice_head(n = 14)
df <- df %>% filter(booking_origin %in% top14$booking_origin)


##########################################################################
#########

##########################################################################
#########

#2.5 Num_passengers
# qqplot
qqnorm(df$num_passengers, main = "num_passengers qqplot")
```

```
qqline(df$num_passengers, col = "red")

# log transformation
df$num_passengers_log <- log(df$num_passengers)

# log transformed histogram
ggplot(df, aes(x = num_passengers_log)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  ggtitle("Histogram of num_passengers_log")

# log transformed boxplot
ggplot(df, aes(y = num_passengers_log)) +
  geom_boxplot(fill = "pink") +
  ggtitle("Boxplot of num_passengers_log")

# log transformed qqplot
qqnorm(df$num_passengers_log, main = "num_passengers_log qqplot")
qqline(df$num_passengers_log, col = "red")

#still not normal after log transformation

##########################################################################
#########

# 2.5 purchase_lead
#qqplot
qqnorm(df$purchase_lead, main = "purchase_lead qqplot")
qqline(df$purchase_lead, col = "red")

# log transformation
df$purchase_lead_log <- log(df$purchase_lead+1)

#log transformed histogram
ggplot(df, aes(x = purchase_lead_log)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  ggtitle("Histogram of purchase_lead_log")

# log transformed boxplot
ggplot(df, aes(y = purchase_lead_log)) +
  geom_boxplot(fill = "pink") +
  ggtitle("Boxplot of purchase_lead_log")

#log transformed qqplot
qqnorm(df$purchase_lead_log, main = "purchase_lead_log qqplot")
qqline(df$purchase_lead_log, col = "red")


##########################################################################
#########

# 2.5 length_of_stay
# qqplot
qqnorm(df$length_of_stay, main = "length_of_stay qqplot")
```

```
qqline(df$length_of_stay, col = "red")
df$exp_length_of_stay <- exp(df$length_of_stay)

# exp transformed histogram
ggplot(df, aes(x = exp_length_of_stay)) +
  geom_histogram(binwidth = 0.2, fill = "skyblue", color = "black") +
  ggtitle("Histogram of exp_length_of_stay")

# exponential transformation
df$exp_length_of_stay <- exp(df$length_of_stay)

# exp transformed boxplot
ggplot(df, aes(y = exp_length_of_stay)) +
  geom_boxplot(fill = "pink") +
  ggtitle("Boxplot of exp_length_of_stay")

# exp transformed qqplot
qqnorm(df$exp_length_of_stay, main = "exp_length_of_stay qqplot")
qqline(df$exp_length_of_stay, col = "red")


##############################################################################
#########

# 2.5 flight duration

#qqplot
qqnorm(df$flight_duration,main = "flight_duration qqplot" )
qqline(df$flight_duration, col = "red")
df$log_flight_duration <- log(df$flight_duration)

# log transformation
df$log_flight_duration <- log(df$flight_duration)

# log transformed histogram
ggplot(df, aes(x = log_flight_duration)) +
  geom_histogram(binwidth = 0.15, fill = "skyblue", color = "black") +
  ggtitle("Histogram of log_flight_duration")

# log transformed boxplot
ggplot(df, aes(y = log_flight_duration)) +
  geom_boxplot(fill = "pink") +
  ggtitle("Boxplot of log_flight_duration")

# log transformed qqplot
qqnorm(df$log_flight_duration,main = "log_flight_duration qqplot" )
qqline(df$log_flight_duration, col = "red")

##############################################################################
#########
```

Appendix B

```
# MH3511 Project Chapter 3.1

#library
library(ggplot2)
library(dplyr)

#Load data
getwd()
setwd("C:/Users/BC2404/Desktop/MH3511/Labs")
df <- read.csv('customer_booking.csv')
df <- df %>% filter(length_of_stay >= 0 & length_of_stay <= 6) %>%  # short stays only
  filter(purchase_lead <= 365) %>%                  # remove extreme leads
  select(-route, -flight_hour, -flight_day, -trip_type) # drop unused columns
top14 <- df %>% count(booking_origin, sort = TRUE) %>% slice_head(n = 14)
df <- df %>% filter(booking_origin %in% top14$booking_origin)
str(df)

#convert binary categorical variables into factor (it was classified as int before)
df$wants_extra_baggage <- factor(df$wants_extra_baggage, levels = c(0, 1))
df$wants_preferred_seat <- factor(df$wants_preferred_seat, levels = c(0, 1))
df$wants_in_flight_meals <- factor(df$wants_in_flight_meals, levels = c(0, 1))

##############################################################################

# 3.1.2.1  Selecting an Extra Baggage and Booking Completion
table_baggage <- table(df$wants_extra_baggage, df$booking_complete)
dimnames(table_baggage) <- list( "Extra Baggage" = c("No", "Yes"),"Booking Completed"
= c("No", "Yes"))
print(table_baggage)
chisq.test(table_baggage)
# H□: The row and column variables are independent.
# H□: They are associated (not independent).
# X-squared = 423.32, df = 1, p-value < 2.2e-16
# p < 0.05: Reject H□ → there is a significant association between the
wants_extra_baggage and booking completion.

# Customers who select extra baggage are significantly more or less likely to complete
their booking compared to those who don't.
prop.table(table_baggage, 1)
# Booking Completed
# Extra Baggage        No       Yes
#          No  0.8656999 0.1343001
#          Yes 0.7602156 0.2397844
# The Chi-squared test shows a statistically significant relationship between extra
baggage and booking completion.
# Customers who select extra baggage are more likely to complete their bookings (13.4%)
compared to those who did not select extra baggage (24.0%).



model <- glm(booking_complete ~ wants_extra_baggage, data = df, family = "binomial")
summary(model)
# Coefficients:
```

```
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)        -1.86346    0.02855  -65.27   <2e-16 ***
# wants_extra_baggage  0.70960    0.03527   20.12   <2e-16 ***
# Customers who select extra baggage (compared to those who don't) have significantly higher odds of completing the booking.
# Specifically, selecting extra baggage increases the log odds of booking completion by 0.70960,
# which translates to an odds ratio of e^0.7096 = 2.033. meaning customers who select baggage are about 2 times more likely to complete their booking.

##########################################################################

# Repeating Tests for Preferred Seating
table_seat <- table(df$wants_preferred_seat, df$booking_complete)
dimnames(table_seat) <- list("Preferred Seat" = c("No", "Yes"),"Booking Completed" = c("No", "Yes"))
print(table_seat)
chisq.test(table_seat)
# H□: The row and column variables are independent.
# H□: They are associated (not independent).
# X-squared = 162.96, df = 1, p-value < 2.2e-16
# p < 0.05: Reject H□ → there is a significant association between the wants_preferred_seat and booking completion.


# Customers who select preferred seating are significantly more or less likely to complete their booking compared to those who don't.
prop.table(table_seat, 1)
# Booking Completed
# Preferred Seat       No       Yes
#            No  0.8287532 0.1712468
#            Yes 0.7558348 0.2441652
# The Chi-squared test shows a statistically significant relationship between preferred seating and booking completion.
# Customers who select preferred seating are more likely to complete their bookings (17.1%) compared to those who did not select preferred seating (24.4%).

model_seat <- glm(booking_complete ~ wants_preferred_seat, data = df, family = "binomial")
summary(model_seat)
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)        -1.57682    0.02056  -76.69   <2e-16 ***
# wants_preferred_seat  0.44684    0.03512   12.72   <2e-16 ***
# Customers who select preferred seating (compared to those who don't) have significantly higher odds of completing the booking.
# Specifically, selecting preferred seating increases the log odds of booking completion by 0.44684,
# which translates to an odds ratio of e^0.44684 = 1.563. meaning customers who select preferred seating are about 1.56 times more likely to complete their booking.

##########################################################################

# Repeating Tests for In-Fight Meals
table_meal <- table(df$wants_in_flight_meals, df$booking_complete)
```

```
dimnames(table_meal) <- list(  "In-flight Meals" = c("No", "Yes"),  "Booking Completed" = c("No", "Yes"))
print(table_meal)
chisq.test(table_meal)
# H□: The row and column variables are independent.
# H□: They are associated (not independent).
# X-squared = 132.53, df = 1, p-value < 2.2e-16
# p < 0.05: Reject H□ → there is a significant association between th wants_in_flight_meals and booking completion.

# Customers who select flight meals are significantly more or less likely to complete their booking compared to those who don't.
prop.table(table_meal, 1)
# Booking Completed
# In-flight Meals        No       Yes
#              No  0.8304730 0.1695270
#              Yes 0.7687719 0.2312281
# The Chi-squared test shows a statistically significant relationship between flight meals and booking completion.
# Customers who select flight meals are more likely to complete their bookings (17.0%) compared to those who did not select extra baggage (23.1%).

model_meal <- glm(booking_complete ~ wants_in_flight_meals, data = df, family = "binomial")
summary(model_meal)
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)           -1.58898    0.02191  -72.53   <2e-16 ***
# wants_in_flight_meals1  0.38759    0.03373   11.49   <2e-16 ***
# Customers who select wants_in_flight_meals (compared to those who don't) have significantly higher odds of completing the booking.
# Specifically, selecting flight meals increases the log odds of booking completion by 0.38759,
# which translates to an odds ratio of e^0.38759 = 1.47. meaning customers who select flight meals are about 1.47 times more likely to complete their booking.

#########################################################################

model_prefernce <- glm(booking_complete ~ wants_extra_baggage +wants_in_flight_meals+wants_preferred_seat, data = df, family = "binomial")
summary(model_prefernce)

# Estimate Std. Error z value Pr(>|z|)
# wants_extra_baggage    0.62420    0.03634  17.179  < 2e-16 ***
# wants_in_flight_meals  0.20803    0.03601   5.777 7.60e-09 ***
# wants_preferred_seat   0.24302    0.03770   6.446 1.15e-10 ***

#########################################################################

#########################################################################

# 3.1.2.2: Sales Channel, Booking Origin and Booking Completion
table_channel <- table(df$sales_channel, df$booking_complete)
```

```
dimnames(table_channel) <- list("Sales Channel" = unique(df$sales_channel),"Booking Completed" = c("No", "Yes"))
print(table_channel)
chisq.test(table_channel)
# H□: The row and column variables are independent.
# H□: They are associated (not independent).
# X-squared = 80.764, df = 1, p-value < 2.2e-16
# p < 0.05: Reject H□ → there is a significant association between the sales_channel and booking completion.

prop.table(table_channel, 1)
#               Booking Completed
# Sales Channel        No       Yes
# Internet 0.7983277 0.2016723
# Mobile   0.8643767 0.1356233
# The Chi-squared test shows a statistically significant relationship between sales channel and booking completion.
# Customers who book via Internet (web/desktop) are more likely to complete their bookings (20.1%) compared to those using Mobile (13.6%).

model_channel <- glm(booking_complete ~ sales_channel, data = df, family = "binomial")
summary(model_channel)
# Coefficients:
#   Estimate Std. Error z value Pr(>|z|)
# (Intercept)        -1.37588    0.01763 -78.022   <2e-16 ***
# sales_channelMobile -0.47625    0.05326  -8.941   <2e-16 ***

# Customers who book through internet (compared to those who book through mobile) have significantly higher odds of completing the booking.
# Specifically, booking through internet increases the log odds of booking completion by -0.47625,
# which translates to an odds ratio of e^ -0.47625 = 0.621. meaning customers who book through internet are about 1/0.621 = 1.61 times more likely to complete their booking.

#############################################################################

# booking_origin
table_origin <- table(df$booking_origin, df$booking_complete)
chisq.test(table_origin)
# H□: The row and column variables are independent.
# H□: They are associated (not independent).
# X-squared = 1882.3, df = 13, p-value < 2.2e-16
# p < 0.05: Reject H□ → there is a significant association between the booking_origin and booking completion.

prop.table(table_origin, 1)
#                0         1
# Australia     0.93440788 0.06559212
# China         0.78349673 0.21650327
# Hong Kong     0.74850299 0.25149701 (5th)
# India         0.88416988 0.11583012
# Indonesia     0.71972318 0.28027682 (4th)
# Japan         0.87027750 0.12972250
# Malaysia      0.60812236 0.39187764 (1st)
```

```
# New Zealand   0.90540541 0.09459459
# Singapore     0.71409922 0.28590078 (3rd)
# South Korea   0.88348651 0.11651349
# Taiwan        0.86977111 0.13022889
# Thailand      0.75427486 0.24572514
# United States 0.79087452 0.20912548
# Vietnam       0.61904762 0.38095238 (2nd)

model_origin <- glm(booking_complete ~ booking_origin, data = df, family = "binomial")
summary(model_origin)
# Coefficients:
#   Estimate Std. Error z value Pr(>|z|)
#   (Intercept)               -2.65646    0.06044 -43.955  < 2e-16 ***
#   booking_originChina         1.37030    0.07785  17.602  < 2e-16 ***
#   booking_originHong Kong     1.56581    0.18831   8.315  < 2e-16 ***
#   booking_originIndia         0.62393    0.15001   4.159 3.19e-05 ***
#   booking_originIndonesia     1.71337    0.08416  20.358  < 2e-16 ***
#   booking_originJapan         0.75304    0.08133   9.259  < 2e-16 ***
#   booking_originMalaysia      2.21703    0.06899  32.137  < 2e-16 ***
#   booking_originNew Zealand   0.39768    0.28730   1.384    0.166
#   booking_originSingapore     1.74108    0.10023  17.370  < 2e-16 ***
#   booking_originSouth Korea   0.63059    0.08082   7.802 6.10e-15 ***
#   booking_originTaiwan        0.75752    0.10306   7.351 1.97e-13 ***
#   booking_originThailand      1.53491    0.08408  18.255  < 2e-16 ***
#   booking_originUnited States  1.32625    0.16322   8.125 4.46e-16 ***
#   booking_originVietnam       2.17095    0.19315  11.240  < 2e-16 ***

# Align with Prop table
# For Booking Origin, it can be observed that all the variables except New Zealand is
statistically significant, and
# the highest 3 odds ratio of e^ 2.22, e^ 2.17, e^1.74 for Malaysia, Vietnam, Singapore
respectively (Australia as the reference category)

#########################################################################

# 3.1.2.2: Purchase lead time
ggplot(df, aes(x = purchase_lead)) +
  geom_histogram(binwidth = 10, fill = "skyblue", color = "black") +
  facet_wrap(~ booking_complete, ncol = 1) +
  ggtitle("Distribution of Purchase Lead by Booking Completion")

qqnorm(df$purchase_lead[df$booking_complete == 1], main = "QQ Plot - Completed")
qqline(df$purchase_lead[df$booking_complete == 1], col = "red")

qqnorm(df$purchase_lead[df$booking_complete == 0], main = "QQ Plot - Not
Completed")
qqline(df$purchase_lead[df$booking_complete == 0], col = "red")

shapiro.test(df$purchase_lead[df$booking_complete == 1])
shapiro.test(df$purchase_lead[df$booking_complete == 0])

# try log Add small constant to avoid log(0) issues
df$purchase_lead_log <- log(df$purchase_lead + 1)
```

```
# Shapiro-Wilk test (for smaller group only)
shapiro.test(df$purchase_lead_log[df$booking_complete == 1])

qqnorm(df$purchase_lead_log[df$booking_complete == 0])
qqline(df$purchase_lead_log[df$booking_complete == 0], col = "red")

# still not normal, Use the non-parametric test:  Wilcoxon rank-sum test in R
wilcox.test(purchase_lead ~ booking_complete, data = df)
# H□: distribution of purchase_lead is the same for customers who completed their
booking and those who didn't.
# H□: distribution of purchase_lead is different between the two groups.
# W = 42602098, p-value = 0.4746
# not sufficient evidence to reject H□, meaning we found no evidence that purchase lead
time differs between those who complete their booking and those who don't.
# No evidence that lead time affects booking completion

model_purchase_lead <- glm(booking_complete ~ purchase_lead, data = df, family =
"binomial")
summary(model_purchase_lead)
# Coefficients:
#   Estimate Std. Error z value Pr(>|z|)
#   (Intercept)   -1.433e+00  2.291e-02 -62.552   <2e-16 ***
#   purchase_lead -4.268e-05  1.986e-04  -0.215     0.83
# p-value > 0.83, there is no sufficient evidence to reject H□. meaning we found no
evidence that purchase lead time differs between those who complete their booking and
those who don't.
# No evidence that purchase lead time affects booking completion

#########################################################################

# Repeating Tests for Flight Duration
wilcox.test(flight_duration ~ booking_complete, data = df)
# H□: distribution of flight duration is the same for customers who completed their booking
and those who didn't.
# H□: distribution of flight duration is different between the two groups.
# W = 45453984, p-value = 6.81e-15
# p-value < 0.05 , we reject H□, Significant difference in flight duration between
completers and non-completers

model_flight_duration <- glm(booking_complete ~ flight_duration, data = df, family =
"binomial")
summary(model_flight_duration)
#   Estimate Std. Error z value Pr(>|z|)
#   (Intercept)     -0.69899    0.08429  -8.292   <2e-16 ***
#   flight_duration -0.10825    0.01224  -8.841   <2e-16 ***
# Customers who booked flights with higher flight duration (compared to those who have
lower flight duration) have significantly lower odds of completing the booking.
# Specifically, those customers decreases the log odds of booking completion by  -
0.10825 ,
# which translates to an odds ratio of e^ -0.10825 = 0.8974. meaning customers who has
higher flight hour are about 10.2% times less likely to complete their booking.

#########################################################################
```

```
# Repeating Tests for Length of Stay
wilcox.test(length_of_stay ~ booking_complete, data = df)
# H□: distribution of length of stay is the same for customers who completed their booking and those who didn't.
# H□: distribution of length of stay is different between the two groups.
# W = 40674354, p-value = 2.857e-05
# p-value < 0.05, we reject H□, Significant difference in length_of_stay between completers and non-completers

model_length_of_stay <- glm(booking_complete ~ length_of_stay, data = df, family = "binomial")
summary(model_length_of_stay)
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)    -1.69220    0.06942 -24.377  < 2e-16 ***
#   length_of_stay  0.05417    0.01420   3.816 0.000136 ***
# Customers who booked flights with to have longer length of stay (compared to those who has lower length of stay) have significantly higher odds of completing the booking.
# Specifically, those customers decreases the log odds of booking completion by 0.05417,
# which translates to an odds ratio of e^ 0.05417  = 1.055. meaning customers who has higher flight hour are about 1.055 times less likely to complete their booking
```

Appendix C

```
#Load data
getwd()
setwd("C:/Users/BC2404/Desktop/MH3511/Labs")
df <- read.csv('customer_booking.csv')
str(df)
any(is.na(df)) # FALSE if no NAs are present.

#convert binary categorical variables into factor (it was classified as int before)
df$wants_extra_baggage <- factor(df$wants_extra_baggage, levels = c(0, 1))
df$wants_preferred_seat <- factor(df$wants_preferred_seat, levels = c(0, 1))
df$wants_in_flight_meals <- factor(df$wants_in_flight_meals, levels = c(0, 1))

#library
library(ggplot2)
library(dplyr)

# Cleaning
df <- df %>%
  filter(length_of_stay >= 0 & length_of_stay <= 6) %>%  # short stays only
  filter(purchase_lead <= 365) %>%                  # remove extreme leads
  select(-route, -flight_hour)                    # drop unused columns

# get top 14 countries
top14 <- df %>% count(booking_origin, sort = TRUE) %>% slice_head(n = 14)
df <- df %>% filter(booking_origin %in% top14$booking_origin)

unique(df$booking_origin)
str(df)

##############################################################################

##############################################################################

# 3.2.2.1 Preference vs Sales Channel

# 3.2.2.1a Extra Baggage vs Sales Channel
sales_baggage <- table(df$sales_channel, df$wants_extra_baggage)
chisq.test(sales_baggage)
# H□: The row and column variables are independent.
# H□: They are associated (not independent).
# X-squared = 19.316, df = 1, p-value = 1.107e-05
# p < 0.05: Reject H□ → there is a significant association between the
wants_extra_baggage and sales_channel.

prop.table(sales_baggage, 1)
#              0         1
# Internet 0.4459520 0.5540480
# Mobile   0.4868226 0.5131774
# Buyers who purchase their tickets via the internet are slightly more likely to purchase
extra baggage compared to those who use the mobile channel. Although this relationship
is statistically significant, the practical difference is relatively small—approximately 4.09%.
```

```
model <- glm(wants_extra_baggage ~ sales_channel, data = df, family = binomial)
summary(model)
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)         0.21704    0.01424  15.247  < 2e-16 ***
# sales_channelMobile -0.16432    0.03726  -4.411 1.03e-05 ***

# Customers who have book through mobile (compared to those who book via internet)
have significantly lower odds of selecting extra baggage.
# Specifically, those customers book via internet decreases the log odd of selecting extra
baggage by 0.16432,
# which translates to an odds ratio of e^-0.16432 = 0.8484 . meaning customers who has
higher length of stay are about 0.1515 times less likely to selecting extra baggage.

#############################################################################

# 3.2.2.1b Preferred Seat vs Sales Channel
sales_seat <- table(df$sales_channel, df$wants_preferred_seat)
chisq.test(sales_seat)
# H□: The row and column variables are independent.
# H□: They are associated (not independent).
# X-squared = 23.523, df = 1, p-value = 1.234e-06
# p < 0.05: Reject H□ → there is a significant association between th
ewants_preferred_seat and sales_channel.

prop.table(sales_seat, 1)
#                0         1
# Internet 0.7196716 0.2803284
# Mobile   0.6787089 0.3212911

model <- glm(wants_preferred_seat ~ sales_channel, data = df, family = binomial)
summary(model)
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)         -0.94283    0.01575 -59.849  < 2e-16 ***
# sales_channelMobile  0.19499    0.04008   4.865 1.14e-06 ***

# Customers who have book through mobile (compared to those who book via internet)
have significantly lower odds of selecting extra baggage.
# Specifically, those customers book via internet increases the log odd of selecting extra
baggage by 0.19499,
# which translates to an odds ratio of e^0.19499 = 1.2152 . meaning customers who has
higher length of stay are about 1.2152 times more likely to selecting extra baggage.


#############################################################################

# 3.2.2.1c In Flight Meal vs Sales Channel
sales_meal <- table(df$sales_channel, df$wants_in_flight_meals)
chisq.test(sales_meal)
# H□: The row and column variables are independent.
# H□: They are associated (not independent).
# X-squared = 5.5587, df = 1, p-value = 0.01839
```

```
# p < 0.05: Reject H□ → there is a significant association between the
wants_extra_baggage and sales_channel.

prop.table(sales_meal, 1)
#                 0         1
# Internet 0.6307515 0.3692485
# Mobile   0.6520580 0.3479420

model <- glm(wants_in_flight_meals ~ sales_channel, data = df, family = binomial)
summary(model)

# Estimate Std. Error z value Pr(>|z|)
# (Intercept)         -0.53544    0.01466 -36.519   <2e-16 ***
# sales_channelMobile -0.09266    0.03899  -2.376   0.0175 *

# Customers who have book through mobile (compared to those who book via internet)
have significantly lower odds of selecting extra baggage.
# Specifically, those customers book via internet decreases the log odd of selecting extra
baggage by 0.09266,
# which translates to an odds ratio of e^-0.09266 = 0.3958 . meaning customers who has
higher length of stay are about 0.604 times less likely to selecting extra baggage.

##########################################################################

##########################################################################

# 3.2.2.2 Preference vs Length of Stay

# 3.2.2.2a Extra Baggage vs Length of Stay
wilcox.test(length_of_stay ~ wants_extra_baggage, data = df)
# H□: distribution of length of stay is the same for customers who select
wants_extra_baggage and those who didn't.
# H□: distribution of length of stay is different between the two groups.
# W = 56599314, p-value < 2.2e-16
# p-value < 0.05, we reject H□, Significant difference in length_of_stay between those
who select wants extra baggage and those who don't.

model <- glm(wants_extra_baggage ~ length_of_stay, data = df, family = binomial)
summary(model)
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)    -0.99474    0.05466  -18.20   <2e-16 ***
# length_of_stay  0.25330    0.01132   22.38   <2e-16 ***
# Customers who have longer length of stay (compared to those who has lower length of
stay) have significantly higher odds of selecting extra baggage.
# Specifically, those customers increase the log odd of selecting extra baggage by
0.25330,
# which translates to an odds ratio of e^ 0.25330 = 1.2882. meaning customers who has
higher length of stay are about 1.2882 times more likely to selecting extra baggage.

##########################################################################

# 3.2.2.2b Flight meals vs Length of Stay
wilcox.test(length_of_stay ~ wants_in_flight_meals, data = df)
```

```
# H□: distribution of length of stay is the same for customers who select
wants_in_flight_meals and those who didn’t.
# H□: distribution of length of stay is different between the two groups.
# W = 61285443, p-value = 3.367e-05
# p-value < 0.05, we reject H□, Significant difference in length_of_stay between those
who select wants flight meals and those who don't.

model <- glm(wants_in_flight_meals ~ length_of_stay, data = df, family = binomial)
#######################################
summary(model)
# (Intercept)    -0.75271    0.05592 -13.460  < 2e-16 ***
# length_of_stay  0.04333    0.01150   3.768 0.000165 ***
# Customers who have longer length of stay (compared to those who has lower length of
stay) have significantly higher odds of selecting flight meals.
# Specifically, those customers increase the log odds of booking completion by 0.04333,
# which translates to an odds ratio of e^ 0.04333 = 1.0442. meaning customers who has
higher length of stay are about 1.0442 times more likely to selecting flight meals.

#############################################################################

# 3.2.2.2c Preferred seats vs Length of Stay
wilcox.test(length_of_stay ~ wants_preferred_seat, data = df)
# H□: distribution of length of stay is the same for customers who select
wants_preferred_seat and those who didn’t.
# H□: distribution of length of stay is different between the two groups.
# W = 55596609, p-value = 0.8217
# p-value > 0.05, we do not sufficient evidence to reject H□,there is no significant
difference in length_of_stay between those who select wants preferred seat and those
who don't.

#############################################################################

#############################################################################

#3.2.2.3 Preference vs Purchase Lead

# 3.2.2.3a Extra Baggage vs Purchase lead
wilcox.test(purchase_lead ~ wants_extra_baggage, data = df)
# H□: distribution of purchase lead is the same for customers who select
wants_extra_baggagee and those who didn’t.
# H□: distribution of purchase lead is different between the two groups.
# W = 62598857, p-value < 2.2e-16
# p-value < 0.05, we reject H□, Significant difference in purchase_lead between those
who select wants extra baggage and those who don't.

model <- glm(wants_extra_baggage ~ purchase_lead, data = df, family = binomial)
summary(model)
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)   0.1344706  0.0181334   7.416 1.21e-13 ***
#  purchase_lead 0.0007397  0.0001580   4.682 2.84e-06 ***
# Customers who have longer purchase lead (compared to those who has lower purchase
lead) have significantly higher odds of selecting extra baggage.
# Specifically, those customers increase the log odds of selecting extra baggage by
0.0007397,
```

```
# which translates to an odds ratio of e^ 0.0007397 = 1.0007. meaning customers who has higher purchase lead are about 1.0007 times more likely to selecting extra baggage.

##############################################################################

# 3.2.2.3b Flight meals vs Purchase lead
wilcox.test(purchase_lead ~ wants_in_flight_meals, data = df)
# H□: distribution of purchase lead is the same for customers who select wants_flight_meals and those who didn't.
# H□: distribution of purchase lead is different between the two groups.
# W = 63571656, p-value = 0.5432
# p-value > 0.05, we do not sufficient evidence to reject H□,there is no significant difference in purchase lead between those who select wants flight meals and those who don't.

##############################################################################

# 3.2.2.3c Preferred seat vs Purchase lead
wilcox.test(purchase_lead ~ wants_preferred_seat, data = df)
# H□: distribution of purchase lead is the same for customers who select wants_preferred_seat and those who didn't.
# H□: distribution of purchase lead is different between the two groups.
# W = 56645707, p-value = 0.04174
# p-value < 0.05, we reject H□,there is significant difference in purchase lead between those who select wants preferred seat and those who don't.

model <- glm(wants_preferred_seat ~ purchase_lead, data = df, family = binomial)
summary(model)
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)   -0.8656295  0.0199155 -43.465  < 2e-16 ***
# purchase_lead -0.0006103  0.0001758  -3.471 0.000519 ***

# Customers who have longer purchase lead (compared to those who has lower purchase lead) have significantly lower odds of selecting preferred seats.
# Specifically, those customers decrease the log odds of selecting preferred seats by 0.0006103,
# which translates to an odds ratio of e^-0.0006103 = 0.9994. meaning customers who has higher purchase lead are about 0.0006 times less likely to selecting extra baggage.


##############################################################################

##############################################################################


# 3.2.2.4 Preference vs Flight Duration

# 3.2.2.4a Extra Baggage vs Flight Duration
wilcox.test(flight_duration ~ wants_extra_baggage, data = df)
# H□: distribution of flight duration is the same for customers who select wants_extra_baggage and those who didn't.
# H□: distribution of flight duration is different between the two groups.
# W = 68027369, p-value = 0.3206
```

```
# p-value > 0.05, we do not sufficient evidence to reject H□,there is no significant
difference in flight duration between those who select wants_extra-baggage and those
who don't.

##########################################################################


# 3.2.2.4b Preferred seat vs Flight Duration
wilcox.test(flight_duration ~ wants_preferred_seat, data = df)
# H□: distribution of flight duration is the same for customers who select
wants_preferred_seat and those who didn’t.
# H□: distribution of flight duration is different between the two groups.
# W = 50000686, p-value < 2.2e-16
# p-value < 0.05, we reject H□,there is significant difference in flight duration between
those who select wants preferred seat and those who don't.

model <- glm(wants_preferred_seat ~ flight_duration, data = df, family = binomial)
summary(model)
# Estimate Std. Error z value Pr(>|z|)
# (Intercept)     -1.78440    0.07568  -23.58   <2e-16 ***
# flight_duration  0.12578    0.01065   11.81   <2e-16 ***
# Customers who have longer flight duration (compared to those who has lower flight
duration) have significantly lower odds of selecting preferred seats.
# Specifically, those customers increase the log odds of selecting preferred seats by
0.12578,
# which translates to an odds ratio of e^0.12578 = 1.134. meaning customers who has
higher flight duration are about 1.134 times more likely to selecting preferred seats.

##########################################################################

# 3.2.2.4c Flight meals vs Flight Duration
wilcox.test(flight_duration ~ wants_in_flight_meals, data = df)
# H□: distribution of flight duration is the same for customers who select
wants_in_flight_meals and those who didn’t.
# H□: distribution of flight duration is different between the two groups.
# W = 55327694, p-value < 2.2e-16
# p-value < 0.05, we reject H□,there is significant difference in flight duration between
those who select wants flight meals and those who don't.

model <- glm(wants_in_flight_meals ~ flight_duration, data = df, family = binomial)
##########################################
summary(model)
#                Estimate Std. Error z value Pr(>|z|)
# (Intercept)     -1.71192    0.07123  -24.04   <2e-16 ***
# flight_duration  0.16822    0.01005   16.74   <2e-16 ***
# Customers who have longer flight duration (compared to those who has lower flight
duration) have significantly more odds of selecting flight meals.
# Specifically, those customers increase the log odds of selecting flight meals by 0.16822,
# which translates to an odds ratio of e^0.16822 = 1.183. meaning customers who has
higher flight duration are about 1.183 times more likely to selecting flight meals.

##########################################################################
```

6. References

Loo, J. (2017, November). *The future of travel: New consumer behavior and the technology giving it flight*. Think with Google. https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/new-consumer-travel-assistance/

British Airways. (n.d.)**.** *Flight schedules*. https://www.britishairways.com/travel/schedules/public/en_sg