Data Analysis Assignment 4

Yuanjing Zhu

11/15/2022

Introduction

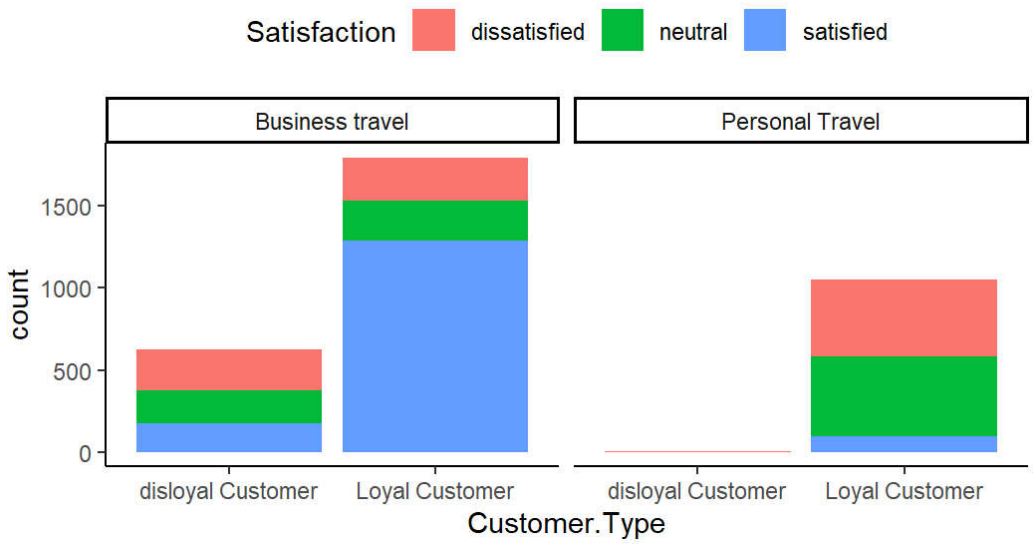
In this research, we looked into the customer satisfaction survey from those who had recently traveled with LaneAir. The goal of this analysis is to better understand the factors that influence customer satisfaction to help LaneAir to better spend its funding for customer satisfaction.

In the survey, customers rated their overall satisfaction as dissatisfied, neutral, or satisfied. The dataset contains demographic information, flight details of customers, and their satisfaction level upon various aspects of the flight. The summary statistics table and plots are in the *Appendix*. We can see that loyal customers are more satisfied than those who are disloyal. In terms of age, we can see that customers who are satisfied with peak around 40-60 years old. Before that, there is only a slight peak on 25 years old. However, dissatisfied/neutral customers peak around 25 to 35 years old and most of them are between 20 and 40 years old. In other words, customers are more dissatisfied at younger age. About 60% of customers who travel for business are satisfied with their flights whereas over 80% of personal travel are either neutral or dissatisfied. We can also spot a big difference among business, eco, and eco plus customers. Most business-class customers are satisfied with their flights while that's the opposite in economy class (only about 20% of them are satisfied). Therefore, we can infer that a substantial portion of LaneAir's loyal clients fly business class for professional reasons and LaneAir performed a wonderful job of upholding the contentment of business-class customers.

Methods

To come up with the relationship of customer satisfaction and various flight services, I applied a ordinal regression model to the data set. I chose ordinal regression model because: 1) our outcome variable (Satisfaction) has 3 categories: dissatisfied, neutral, satisfied and they are "ordered" as such; 2) we have multiple predictors; 3) ordinal regression analysis inferences a dependence relationship between independent variables and dependent variable so that we can know which services to work on to increase customer satisfaction according to the coefficients of fitted model.

Results



For the investments that LaneAir is considering, purpose of the passenger's flight, whether they are loyal, satisfaction of inflight wifi service, inflight service, on-board service, leg-room service, cleanliness, arrival delay are related to customer's overall satisfaction. Considering the difficulties of implementing them, improving customer loyalty and appealing to different customer type would be the most effective strategy. From the plot, we can see that most loyal business travelers are satisfied or at least neutral to their flight experience while personal travelers are more likely to be dissatisfied even they are loyal customers. Therefore, we can implement marketing strategies to attract more business travelers.

Conclusion

Considering the difficulties of implementing investments, here are my recommendations for LaneAir to improve customer satisfaction with best investment (ranked from most easy to least easy to implement)

1. Implementing marketing strategies to increase customer loyalty such as creating a point system, rewarding customers, offering discounts, encouraging referrals, etc., and develop promotion intitatives to appeal to customers who fly for business needs.
2. Increase technology investment on inflight wifi service to deliver a great wifi passenger experience.
3. Hire more flight attendants or other staff to improve inflight service and on-board service

I would also recommend improving online boarding and checkin service. For example, improvement such as cooperating with US TSA to shave off a few minutes from the waiting time, improving LaneAir mobile app to make online boarding more smoothly, adding more lanes to reduce wait time when checking in.

I won't recommend getting new plane models to improve reliability to minimize delays or change newer,larger seats to improve seat comfort and leg room because they are hard to implement and contribute little (even not related) to improving customer satisfaction.

But there are some limitations that we should be aware of.

1. The majority of clients who respond—between 70 and 80 percent—are probably happy with your service. Research shows that most clients will offer either a perfect or very good rating when presented with a 10-point satisfaction scale or the 11-point likeliness-to-recommend scale. The remaining 20% of customers are either unsatisfied or neutral. Therefore, there would be a higher proportion of unsatisfied customers and we are not able to know why and which part they are unsatisfied with.
2. Although we crafted the survey questions carefully about satisfaction level into 5 categories: 1-5, customers view surveys differently than we do. If a person's viewpoint was optimistic, he/she will respond positively to each organized question. The opposite is also true. Additionally, customers probably swiftly answer all of the questions in the survey without giving it any thought, e.g. giving the same top 1 or 2 box scores. Therefore, the 5-level rating doesn't provide much of insight.

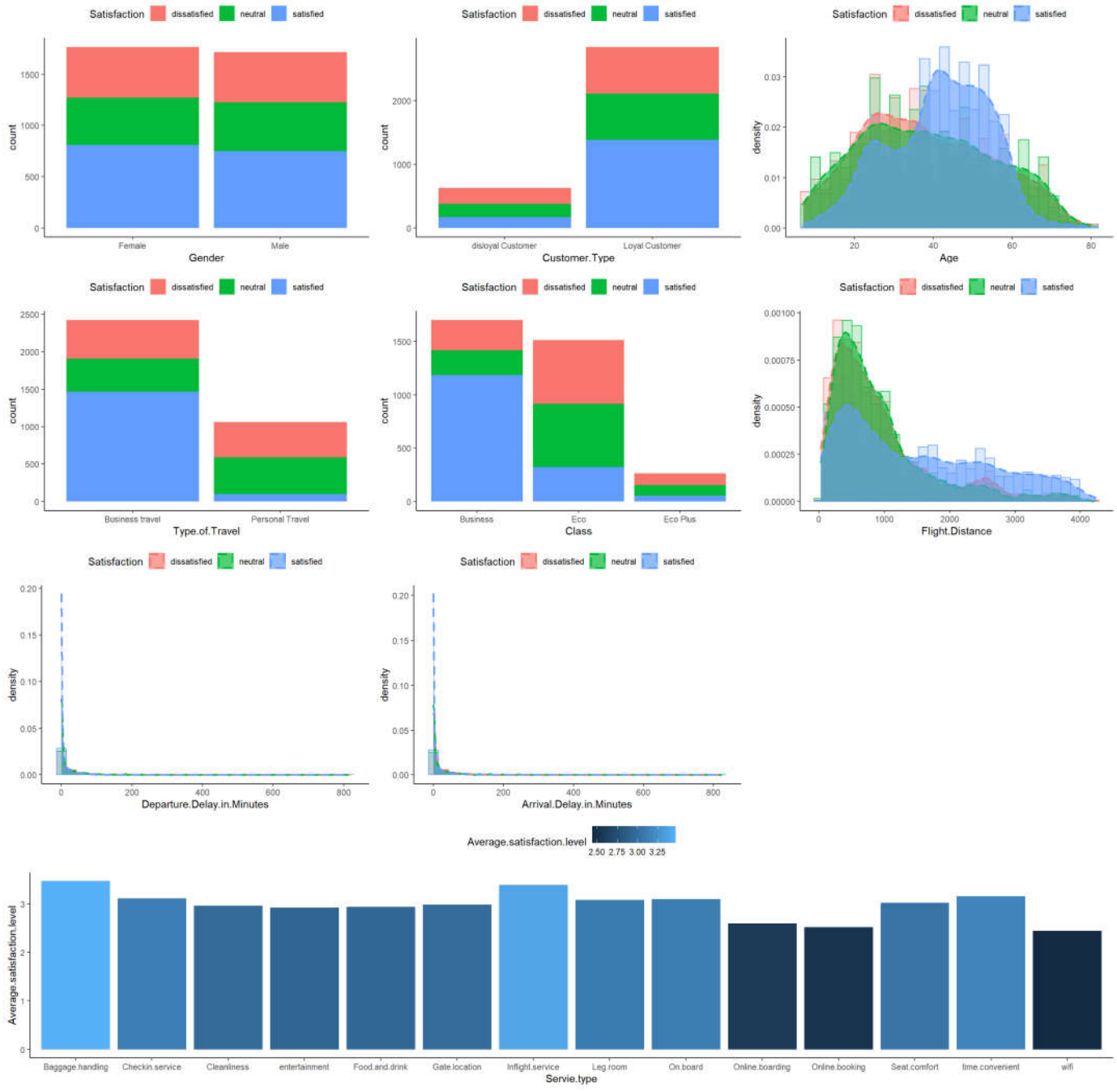
Appendix

Summary statistics of dataset

	dissatisfied (N=988)	neutral (N=932)	satisfied (N=1558)	Overall (N=3478)
Gender				
Female	497 (50.3%)	459 (49.2%)	809 (51.9%)	1765 (50.7%)
Male	491 (49.7%)	473 (50.8%)	749 (48.1%)	1713 (49.3%)
Customer.Type				
disloyal Customer	253 (25.6%)	202 (21.7%)	174 (11.2%)	629 (18.1%)
Loyal Customer	735 (74.4%)	730 (78.3%)	1384 (88.8%)	2849 (81.9%)
Age				
Mean (SD)	38.0 (16.5)	38.4 (16.9)	41.5 (12.7)	39.7 (15.1)
Median [Min, Max]	36.0 [7.00, 80.0]	37.0 [7.00, 80.0]	42.0 [8.00, 80.0]	40.0 [7.00, 80.0]
Type.of.Travel				
Business travel	514 (52.0%)	441 (47.3%)	1464 (94.0%)	2419 (69.6%)
Personal Travel	474 (48.0%)	491 (52.7%)	94 (6.0%)	1059 (30.4%)
Class				
Business	283 (28.6%)	233 (25.0%)	1186 (76.1%)	1702 (48.9%)
Eco	597 (60.4%)	599 (64.3%)	319 (20.5%)	1515 (43.6%)
Eco Plus	108 (10.9%)	100 (10.7%)	53 (3.4%)	261 (7.5%)
Flight.Distance				
Mean (SD)	939 (799)	928 (765)	1500 (1110)	1190 (985)
Median [Min, Max]	687 [31.0, 4240]	690 [56.0, 3990]	1200 [67.0, 4000]	846 [31.0, 4240]
Inflight.wifi.service				
Mean (SD)	2.43 (0.951)	2.35 (0.937)	3.15 (1.61)	2.73 (1.34)
Median [Min, Max]	2.00 [1.00, 5.00]	2.00 [1.00, 5.00]	4.00 [0, 5.00]	3.00 [0, 5.00]
Departure.Arrival.time.convenient				
Mean (SD)	3.09 (1.52)	3.21 (1.50)	2.95 (1.58)	3.06 (1.55)
Median [Min, Max]	3.00 [0, 5.00]	4.00 [0, 5.00]	3.00 [0, 5.00]	3.00 [0, 5.00]
Ease.of.Online.booking				
Mean (SD)	2.62 (1.23)	2.54 (1.19)	2.99 (1.60)	2.76 (1.41)
Median [Min, Max]	3.00 [0, 5.00]	2.00 [0, 5.00]	3.00 [0, 5.00]	3.00 [0, 5.00]
Gate.location				
Mean (SD)	3.02 (1.23)	3.04 (1.19)	2.94 (1.38)	2.99 (1.29)
Median [Min, Max]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]
Food.and.drink				
Mean (SD)	2.96 (1.36)	2.92 (1.33)	3.56 (1.24)	3.22 (1.33)
Median [Min, Max]	3.00 [0, 5.00]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	3.00 [0, 5.00]
Online.boarding				
Mean (SD)	2.65 (1.12)	2.64 (1.14)	4.00 (1.24)	3.25 (1.36)

	dissatisfied (N=988)	neutral (N=932)	satisfied (N=1558)	Overall (N=3478)
Median [Min, Max]	3.00 [0, 5.00]	3.00 [0, 5.00]	4.00 [0, 5.00]	4.00 [0, 5.00]
Seat.comfort				
Mean (SD)	3.02 (1.33)	3.04 (1.31)	3.99 (1.13)	3.46 (1.33)
Median [Min, Max]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]
Inflight.entertainment				
Mean (SD)	2.92 (1.34)	2.83 (1.32)	4.01 (1.06)	3.38 (1.34)
Median [Min, Max]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]
On.board.service				
Mean (SD)	3.06 (1.27)	2.98 (1.28)	3.89 (1.12)	3.41 (1.28)
Median [Min, Max]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]
Leg.room.service				
Mean (SD)	3.01 (1.34)	2.94 (1.33)	3.82 (1.18)	3.36 (1.34)
Median [Min, Max]	3.00 [0, 5.00]	3.00 [0, 5.00]	4.00 [0, 5.00]	4.00 [0, 5.00]
Baggage.handling				
Mean (SD)	3.43 (1.11)	3.34 (1.21)	4.00 (1.08)	3.66 (1.17)
Median [Min, Max]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]
Checkin.service				
Mean (SD)	3.06 (1.27)	3.10 (1.31)	3.63 (1.16)	3.32 (1.26)
Median [Min, Max]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	3.00 [1.00, 5.00]
Inflight.service				
Mean (SD)	3.37 (1.17)	3.34 (1.22)	3.99 (1.11)	3.64 (1.20)
Median [Min, Max]	4.00 [1.00, 5.00]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	4.00 [1.00, 5.00]
Cleanliness				
Mean (SD)	2.92 (1.35)	2.91 (1.32)	3.76 (1.15)	3.29 (1.32)
Median [Min, Max]	3.00 [1.00, 5.00]	3.00 [1.00, 5.00]	4.00 [1.00, 5.00]	3.00 [1.00, 5.00]
Departure.Delay.in.Minutes				
Mean (SD)	16.0 (38.0)	19.8 (53.3)	10.4 (26.7)	14.5 (38.8)
Median [Min, Max]	0 [0, 565]	0 [0, 815]	0 [0, 352]	0 [0, 815]
Arrival.Delay.in.Minutes				
Mean (SD)	17.2 (39.1)	20.2 (53.2)	10.9 (27.6)	15.2 (39.4)
Median [Min, Max]	1.00 [0, 586]	0 [0, 822]	0 [0, 350]	0 [0, 822]

Figure 1: Exploratory data analysis



Data overview and analysis plan

The plot shows the distribution of the outcome variable. Abouth 44.8% of our customers are satisfied, 26.8% of them are neutral and 28.4% of them are dissatisfied. Before building the model, I drop the data where the satisfaction level is 0, which is not applicable. The total number of rows reduced from 3478 to 3187, so we didn't lose much of information by dropping the missing values.

In this analysis, I used ordinal regression model because the outcome variable has three categories: dissatisfied, neutral, satisfied, and the order matters. In order to use ordinal regression, I made the assumption that odds ratio is the same regardless of cumulative probability, i.e. slope is the same among each category of a predictor. The link function is cumulative logit.

Model results

Ordinal regression model results

	Value	Std. Error	t value	p
GenderMale	-0.05	0.08	-0.57	0.57
Customer.TypeLoyal Customer	1.86	0.12	15.40	0
Age	-0.003	0.003	-1.12	0.26
Type.of.TravelPersonal Travel	-1.97	0.12	-16.90	0
ClassEco	-0.23	0.10	-2.45	0.01
ClassEco Plus	-0.54	0.15	-3.58	0.0003
Flight.Distance	0.0001	0.0001	1.29	0.20
Inflight.wifi.service	0.39	0.05	7.58	0
Departure.Arrival.time.convenient	-0.08	0.04	-2.02	0.04
Ease.of.Online.booking	-0.18	0.05	-3.51	0.0004
Gate.location	-0.04	0.04	-1.11	0.27
Food.and.drink	-0.07	0.04	-1.67	0.10
Online.boarding	0.52	0.04	11.93	0
Seat.comfort	0.04	0.05	0.91	0.36
Inflight.entertainment	0.09	0.06	1.64	0.10

On.board.service	0.18	0.04	4.68	0.0000
Leg.room.service	0.17	0.03	5.10	0.0000
Baggage.handling	0.06	0.05	1.25	0.21
Checkin.service	0.20	0.03	5.95	0
Inflight.service	0.12	0.05	2.70	0.01
Cleanliness	0.11	0.05	2.21	0.03
Departure.Delay.in.Minutes	0.01	0.004	1.45	0.15
Arrival.Delay.in.Minutes	-0.01	0.004	-2.04	0.04
dissatisfied| neutral	4.05	0.05	84.90	0
neutral| satisfied	6.07	0.08	76.87	0

Ordinal regression model confidence interval

	OR	2.5 %	97.5 %
GenderMale	0.96	0.82	1.12
Customer.TypeLoyal Customer	6.41	5.02	8.20
Age	1.00	0.99	1.00
Type.of.TravelPersonal Travel	0.14	0.11	0.18
ClassEco	0.79	0.64	0.98
ClassEco Plus	0.59	0.43	0.80
Flight.Distance	1.00	1.00	1.00
Inflight.wifi.service	1.47	1.33	1.63
Departure.Arrival.time.convenient	0.93	0.86	1.00
Ease.of.Online.booking	0.84	0.76	0.92
Gate.location	0.96	0.89	1.03
Food.and.drink	0.93	0.85	1.02
Online.boarding	1.68	1.54	1.83
Seat.comfort	1.04	0.95	1.14
Inflight.entertainment	1.10	0.98	1.23
On.board.service	1.20	1.11	1.30
Leg.room.service	1.18	1.11	1.26
Baggage.handling	1.06	0.97	1.16
Checkin.service	1.22	1.14	1.31
Inflight.service	1.13	1.03	1.24
Cleanliness	1.12	1.01	1.24
Departure.Delay.in.Minutes	1.01	1.00	1.01
Arrival.Delay.in.Minutes	0.99	0.98	1.00

According to p-value, type of travel, class, customer type, inflight wifi service, ease of online booking, inflight service, online boarding, on-board service, leg room service, departure-arrival time convenient, cleanliness checkin service, and arrival-delay(min) are statistically significant. Here are the interpretations of some compelling results:

Controlling other variables constant, comparing personal purpose with business purpose (baseline), the odds of being neutral or satisfied is 0.14 times of being dissatisfied. We are 95% confident that the true odds ratio of more & less satisfied categories is between 0.11 and 0.18.

Controlling other variables constant, comparing loyal with disloyal customer, the odds of being neutral or satisfied is 6.41 times of being dissatisfied. We are 95% confident that the true odds ratio of more & less satisfied categories is between 5.02 and 8.20.

Controlling other variables constant, for every unit increase in satisfaction level of online boarding, the odds of being neutral or satisfied is 1.68 times of being dissatisfied. We are 95% confident that the true odds ratio of more & less satisfied catogories is between 1.53 and 1.83.

Controlling other variables constant, for every unit increase in satisfaction level of inflight wifi service, the odds of being neutral or satisfied is 1.47 times of being dissatisfied. We are 95% confident that the true odds ratio of more & less satisfied catogories is between 1.33 and 1.63.

Controlling other variables constant, for every unit increase in satisfaction level of check-in service, the odds of being neutral or satisfied is 1.22 times of being dissatisfied. We are 95% confident that the true odds ratio of more & less satisfied catogories is between 1.14 and 1.31.

Controlling other variables constant, for every unit increase in satisfaction level of on-board service, the odds of being neutral or satisfied is 1.20 times of being dissatisfied. We are 95% confident that the true odds ratio of more & less satisfied catogories is between 1.11 and 1.30.

Controlling other variables constant, for every unit increase in satisfaction level of leg room, the odds of being neutral or satisfied is 1.18 times of being dissatisfied. We are 95% confident that the true odds ratio of more & less satisfied catogories is between 1.11 and 1.26.

Model assessment

Ordinal regression model confusion matrix

	dissatisfied	neutral	satisfied
dissatisfied	471	462	13
neutral	324	305	104
satisfied	126	105	1,277

Multinormial regression model confusion matrix

	dissatisfied	neutral	satisfied
dissatisfied	489	408	78
neutral	314	363	19
satisfied	118	101	1,297

According to the confusion matrix, the accuracy of this ordinal regression model is 64.42%, which is not very high. The key assumption that is unique to this model is that odds ratio remains the same regardless of cumulative probability. Then I used the predictors Gender and Customer type to compare predictions between ordinal regression model and multinormial regression model. I created a new data set with 4 observations for each Gender, Customer type pair while keeping all other variables constant(numerical variables as mean and categorical variables as mode). The predicted probabilities of ordinal regression and multinormial regression are not very similar, indicating the assumption of ordinal regression might be violated. Finally, I presented the confusion matrix of the two prediction models. We can see that the accuracy of multinormial model is 67.43%, which is higher than that in ordinal regression (64.42%). So multinormial is more “precise” because it has different slope for each category of categorical variables.

Conclusion

The validity of this analysis relies on the assumption that we’ve made about ordinal regression. Compared multinormial regression and ordinal regression model predictions using Gender and Customer type, the predicted probabilities are different for each pair, so the assumption might not hold. Therefore, we would

recommend using multinormial regression model for better model accuracy.

```
h1, h4 {
  text-align: center;
}

table, td, th {
  border: none;
  padding-left: 1em;
  padding-right: 1em;
  margin-left: auto;
  margin-right: auto;
  margin-top: 1em;
  margin-bottom: 1em;
}
```

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
library(tidyr)
library(dplyr)
library(ggplot2)
library(brew)
library(stargazer)
library(patchwork)
library(corrplot)
library(caret)
library(kableExtra)
library(broom)
library(car)
library(leaps)
library(MASS)
library(grid)
library(gridExtra)
library(pROC)
library(PerformanceAnalytics)
library(foreign)
library(nnet)
library(table1)
# Data
load("airline_survey")

head(airline)
dim(airline)
str(airline)

# Data cleaning
any(is.na(airline))    # no missing value
summary(airline)

# table1
table1(~ Gender + Customer.Type + Age + Type.of.Travel + Class + Flight.Distance + Inflight.w
ifi.service  + Departure.Arrival.time.convenient + Ease.of.Online.booking + Gate.location + F
ood.and.drink + Online.boarding + Seat.comfort + Inflight.entertainment + On.board.service +
 Leg.room.service + Baggage.handling + Checkin.service + Inflight.service + Cleanliness + Dep
arture.Delay.in.Minutes + Arrival.Delay.in.Minutes | Satisfaction, data = airline, caption =
"Summary statistics of dataset")


ggplot(airline, aes(x = Customer.Type, fill = Satisfaction)) +
```

```
  geom_bar() +
  theme_classic() +
#  labs(title = "Customer satisfaction by Customer type")+
  theme(legend.position="top") + facet_wrap("Type.of.Travel")

# table1
table1(~ Gender + Customer.Type + Age + Type.of.Travel + Class + Flight.Distance + Inflight.w
ifi.service  + Departure.Arrival.time.convenient + Ease.of.Online.booking + Gate.location + F
ood.and.drink + Online.boarding + Seat.comfort + Inflight.entertainment + On.board.service +
 Leg.room.service + Baggage.handling + Checkin.service + Inflight.service + Cleanliness + Dep
arture.Delay.in.Minutes + Arrival.Delay.in.Minutes | Satisfaction, data = airline, caption =
"Summary statistics of dataset")


gender <- ggplot(airline, aes(x=Gender, fill=Satisfaction)) +
  geom_bar() +
  theme_classic() +
#  labs(title = "Customer satisfaction by gender")+
  theme(legend.position="top")

customerType <- ggplot(airline, aes(x = Customer.Type, fill = Satisfaction)) +
  geom_bar() +
  theme_classic() +
#  labs(title = "Customer satisfaction by Customer type")+
  theme(legend.position="top")

age <-ggplot(airline, aes(x=Age, fill=Satisfaction, color=Satisfaction)) +
  geom_histogram(aes(y=..density..), position="identity", alpha=0.2)+
  geom_density(alpha=0.6, size = 1, linetype = "dashed")+
 # labs(title="Age distribution by sex based on satisfaction") +
  theme_classic() +
  theme(legend.position="top")


typeOfTravel <- ggplot(airline, aes(x = Type.of.Travel, fill = Satisfaction)) +
  geom_bar() +
  theme_classic() +
 # labs(title = "Customer satisfaction by Type of travel")+
  theme(legend.position="top")

Class <- ggplot(airline, aes(x = Class, fill = Satisfaction)) +
  geom_bar() +
  theme_classic() +
#labs(title = "Customer satisfaction by Type of travel")+
  theme(legend.position="top")

flightDistance <- ggplot(airline, aes(x=Flight.Distance, fill=Satisfaction, color=Satisfactio
n)) +
  geom_histogram(aes(y=..density..), position="identity", alpha=0.2)+
  geom_density(alpha=0.6, size = 1, linetype = "dashed")+
  #labs(title="Customer satisfaction by Flight distance") +
  theme_classic()+
  theme(legend.position="top")

departureDelay <- ggplot(airline, aes(x=Departure.Delay.in.Minutes, fill=Satisfaction, color=
Satisfaction)) +
```

```
  geom_histogram(aes(y=..density..), position="identity", alpha=0.2)+
  geom_density(alpha=0.6, size = 1, linetype = "dashed")+
  #labs(title="Customer satisfaction by Departure.Delay.in.Minutes") +
  theme_classic()+
  theme(legend.position="top")

Arrival.Delay.in.Minutes <- ggplot(airline, aes(x=Arrival.Delay.in.Minutes, fill=Satisfactio
n, color=Satisfaction)) +
  geom_histogram(aes(y=..density..), position="identity", alpha=0.2)+
  geom_density(alpha=0.6, size = 1, linetype = "dashed")+
  #labs(title="Customer satisfaction by Arrival.Delay.in.Minutes") +
  theme_classic()+
  theme(legend.position="top")

disatisfied <- airline %>% dplyr::select(Class, Inflight.wifi.service, Ease.of.Online.bookin
g, Inflight.service, Online.boarding, Inflight.entertainment, Food.and.drink, Seat.comfort, O
n.board.service, Leg.room.service, Departure.Arrival.time.convenient, Baggage.handling, Gate.
location, Cleanliness, Checkin.service, Satisfaction) %>% filter((Satisfaction == "dissatisfi
ed") & (Class == "Eco"))
Means <- colMeans(disatisfied[, -c(1,16)])

Servie.type <- c('wifi', 'Online.booking', 'Inflight.service', 'Online.boarding', 'entertainm
ent', 'Food.and.drink', 'Seat.comfort', 'On.board', 'Leg.room', 'time.convenient', 'Baggage.h
andling', 'Gate.location', 'Cleanliness', 'Checkin.service')
Average.satisfaction.level <- c(2.440536,  2.514238, 3.383585,2.589615,2.924623,2.936348,3.02
0101, 3.092127, 3.078727, 3.157454, 3.469012, 2.979899, 2.958124, 3.110553)
average_service_satisfaction <- data.frame(Servie.type, Average.satisfaction.level)

serviceSatisfaction <- ggplot(average_service_satisfaction, aes(x=Servie.type, fill=Average.s
atisfaction.level, y = Average.satisfaction.level)) +
  geom_bar(stat="identity") +
  theme_classic() +
#  labs(title = "Customer satisfaction by gender")+
  theme(legend.position="top")

ggrid1 <- grid.arrange(gender, customerType, age, typeOfTravel, Class, flightDistance, depart
ureDelay, Arrival.Delay.in.Minutes, nrow = 3)
grid.arrange(ggrid1, serviceSatisfaction, nrow = 2, top = textGrob("Figure 1: Exploratory dat
a analysis",gp=gpar(fontsize=20,font=3)), heights = c(3/4, 1/4))
ggplot(airline, aes(x=`Satisfaction`, fill = Satisfaction))+
  geom_bar(alpha=0.6)+
  geom_text(aes(label=scales::percent((..count..)/sum(..count..))), stat="count", vjust = -0.
2)+
  stat_count(aes(y=..count..,label=paste0("n=",..count..)),geom="text",vjust=1.2,color="gray3
5")+
  labs(x="Satisfaction", title = "Distribution of Customer Satisfaction")+
  theme()

#par(mfrow=c(1,2))
# numerical variables
#num_cols <- cor(subset(airline, select = c(Age, Flight.Distance,Departure.Delay.in.Minutes,A
rrival.Delay.in.Minutes)))
#corrplot(num_cols, na.label = " ", method="color", tl.col = "black", tl.cex = 1)

# categorical variables
#cat_cols <- cor(select_if(subset(airline, select=-c(id, Age, Flight.Distance,Departure.Dela
```

```
y.in.Minutes,Arrival.Delay.in.Minutes)), is.numeric))
#corrplot(cat_cols, na.label=" ", tl.cex=1, tl.col="black", method="color", type = 'lower')

# remove 0 values
airline_new <- airline[apply(airline, 1, function(row) all(row !=0 )), ]
airline_new <- airline_new[, -1]

# Ordinal distribution
ordinal <- polr(Satisfaction ~ ., data = airline_new, Hess = TRUE)
#summary(ordinal)

# calculate p value
p <- pnorm(-abs(summary(ordinal)$coef[, "t value"])) * 2
ctable <- cbind(summary(ordinal)$coef, p)
table_ordinal <- exp(cbind(OR = coef(ordinal), confint(ordinal)))

cm_ordinal <- confusionMatrix(predict(ordinal), airline_new$Satisfaction)

stargazer(ctable, type = 'html', digits = 2, title = "Ordinal regression model results", out
 = 'ctable.html')
stargazer(table_ordinal, type = 'html', digits = 2, title = "Ordinal regression model confide
nce interval", out = 'table_ordinal.html')
ConfMat <- as.data.frame.matrix(cm_ordinal$table)
stargazer(ConfMat, type = 'html', out = 'cm_ordinal.html', title = "Ordinal regression model
 confusion matrix", digits = 2, summary = FALSE)
# multinormial distribution
multi <- multinom(Satisfaction ~ ., data = airline_new)
summary(multi)

# calculate p value
z <- summary(multi)$coefficients/summary(multi)$standard.errors
p <- (1-pnorm(abs(z)))*2
cm_multi <- confusionMatrix(predict(multi), airline_new$Satisfaction)

# using the predictor: Gender & Customer type

#define function to calculate mode
find_mode <- function(x) {
  u <- unique(x)
  tab <- tabulate(match(x, u))
  u[tab == max(tab)]
}

newdat <- data.frame(Age = mean(airline_new$Age),
                     Gender = c('Male', "Female", "Male", "Female"),
                     Type.of.Travel = find_mode(airline_new$Type.of.Travel),
                     Class = find_mode(airline_new$Class),
                     Customer.Type = c("Loyal Customer", "disloyal Customer", "disloyal Custo
mer", "Loyal Customer"),
                     Flight.Distance = mean(airline_new$Flight.Distance),
                     Inflight.wifi.service = mean(airline_new$Inflight.wifi.service),
                     Ease.of.Online.booking = mean(airline_new$Ease.of.Online.booking),
                     Inflight.service = mean(airline_new$Inflight.service),
                     Online.boarding = mean(airline_new$Online.boarding),
                     Inflight.entertainment = mean(airline_new$Inflight.entertainment),
                     Food.and.drink = mean(airline_new$Food.and.drink),
```

```
                    Seat.comfort = mean(airline_new$Seat.comfort),
                    On.board.service = mean(airline_new$On.board.service),
                    Leg.room.service = mean(airline_new$Leg.room.service),
                    Departure.Arrival.time.convenient = mean(airline_new$Departure.Arrival.t
ime.convenient),
                    Baggage.handling = mean(airline_new$Baggage.handling),
                    Gate.location = mean(airline_new$Gate.location),
                    Cleanliness = mean(airline_new$Cleanliness),
                    Checkin.service = mean(airline_new$Checkin.service),
                    Departure.Delay.in.Minutes = mean(airline_new$Departure.Delay.in.Minute
s),
                    Arrival.Delay.in.Minutes = mean(airline_new$Arrival.Delay.in.Minutes)
)

predict(ordinal, newdat, type = "probs")

predict(multi, newdat, type = 'probs')

cm_multi <- as.data.frame.matrix(cm_multi$table)
stargazer(cm_multi, type = 'html', out = 'multi_cm.html', title = "Multinormial regression mo
del confusion matrix", digits = 2, summary = FALSE)
```