# Math156-Project

Huy Nguyen

5/22/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
```

## Load data

```
train <- suppressMessages(suppressWarnings(read_csv("/Users/huynguyen/Downloads/Math 156/train.csv"))) 
train_no_id <- train[,-c(1,2)] #rid of index and id
test <- suppressMessages(suppressWarnings(read_csv("/Users/huynguyen/Downloads/Math 156/test.csv"))) #t
test_no_id <- train[,-c(1,2)]

#deal with categorical variables
train_no_id$Gender <- factor(train_no_id$Gender)
train_no_id$`Customer Type` <- factor(train_no_id$`Customer Type`)
train_no_id$`Type of Travel` <- factor(train_no_id$`Type of Travel`)
train_no_id$Class <- factor(train_no_id$Class)
train_no_id$`Inflight wifi service` <- factor(train_no_id$`Inflight wifi service`)
train_no_id$`Departure/Arrival time convenient` <- factor(train_no_id$`Departure/Arrival time convenient
train_no_id$`Ease of Online booking` <- factor(train_no_id$`Ease of Online booking`)
train_no_id$`Gate location` <- factor(train_no_id$`Gate location`)
train_no_id$`Food and drink` <- factor(train_no_id$`Food and drink`)
train_no_id$`Online boarding` <- factor(train_no_id$`Online boarding`)
train_no_id$`Seat comfort` <- factor(train_no_id$`Seat comfort`)
train_no_id$`Inflight entertainment` <- factor(train_no_id$`Inflight entertainment`)
train_no_id$`On-board service` <- factor(train_no_id$`On-board service`)
train_no_id$`Leg room service` <- factor(train_no_id$`Leg room service`)
train_no_id$`Baggage handling` <- factor(train_no_id$`Baggage handling`)
train_no_id$`Checkin service` <- factor(train_no_id$`Checkin service`)
train_no_id$`Inflight service` <- factor(train_no_id$`Inflight service`)
train_no_id$Cleanliness <- factor(train_no_id$Cleanliness)
```

```
test_no_id$Gender <- factor(test_no_id$Gender)
test_no_id$`Customer Type` <- factor(test_no_id$`Customer Type`)
test_no_id$`Type of Travel` <- factor(test_no_id$`Type of Travel`)
test_no_id$Class <- factor(test_no_id$Class)
test_no_id$`Inflight wifi service` <- factor(test_no_id$`Inflight wifi service`)
test_no_id$`Departure/Arrival time convenient` <- factor(test_no_id$`Departure/Arrival time convenient`)
test_no_id$`Ease of Online booking` <- factor(test_no_id$`Ease of Online booking`)
test_no_id$`Gate location` <- factor(test_no_id$`Gate location`)
test_no_id$`Food and drink` <- factor(test_no_id$`Food and drink`)
test_no_id$`Online boarding` <- factor(test_no_id$`Online boarding`)
test_no_id$`Seat comfort` <- factor(test_no_id$`Seat comfort`)
test_no_id$`Inflight entertainment` <- factor(test_no_id$`Inflight entertainment`)
test_no_id$`On-board service` <- factor(test_no_id$`On-board service`)
test_no_id$`Leg room service` <- factor(test_no_id$`Leg room service`)
test_no_id$`Baggage handling` <- factor(test_no_id$`Baggage handling`)
test_no_id$`Checkin service` <- factor(test_no_id$`Checkin service`)
test_no_id$`Inflight service` <- factor(test_no_id$`Inflight service`)
test_no_id$Cleanliness <- factor(test_no_id$Cleanliness)
```

## Preliminary Plots

```
length(is.na(train_no_id)) # Number of NAs in train
```

```
## [1] 2389792
```

```
length(is.na(test_no_id)) #Number f NAs in test
```

```
## [1] 2389792
```

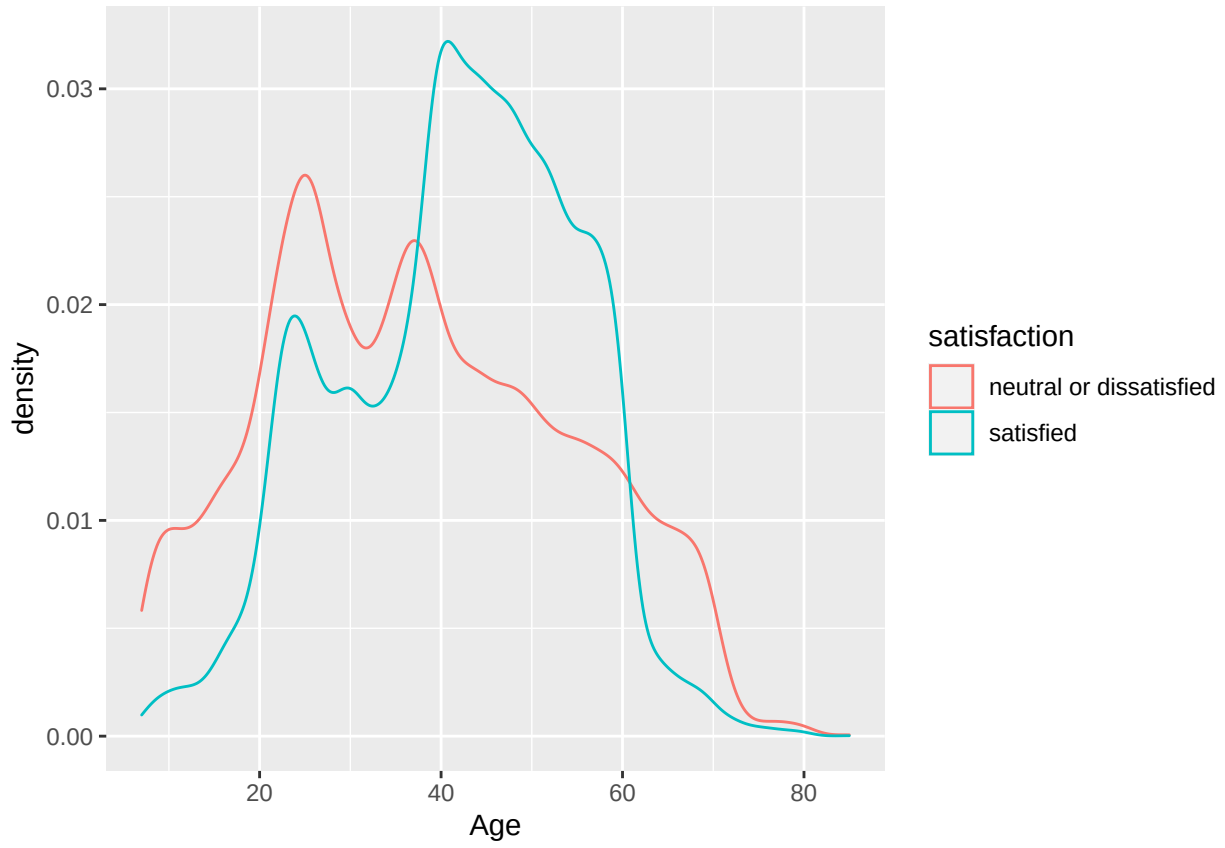
```
set.seed(123) #Set seed to 123
#plot()
```

```
# Numerical variables
library(gridExtra)
```

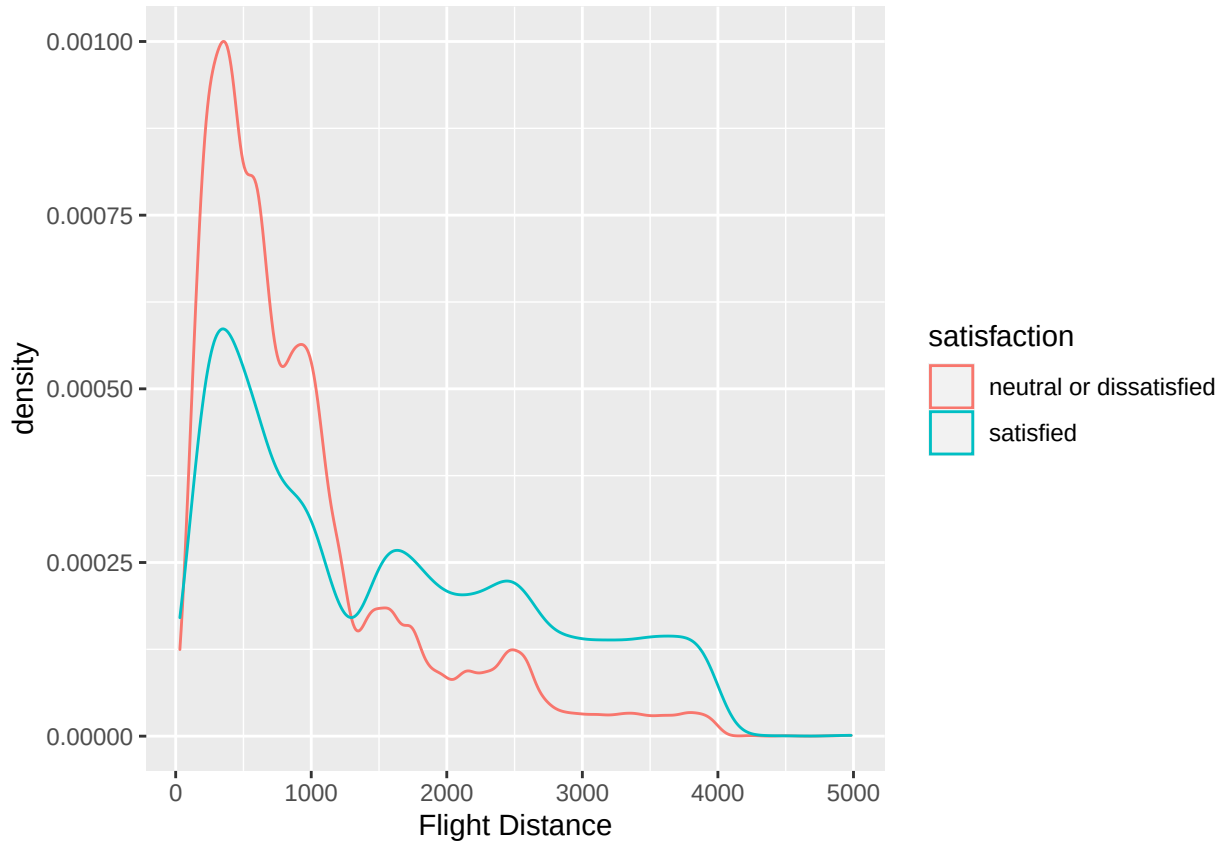
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

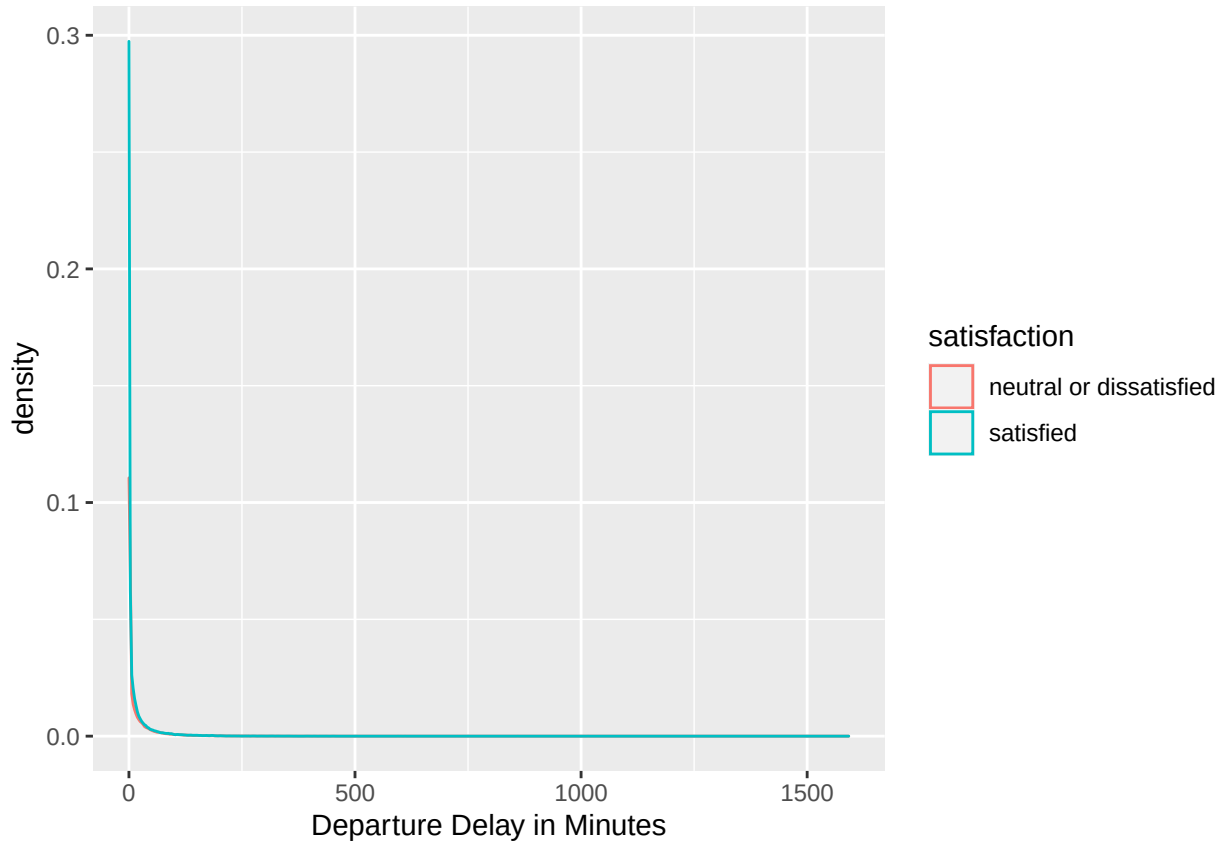
```
g1<-ggplot(train_no_id, aes(x = Age))+geom_density(aes(color = satisfaction)) #definitely in model
g2<-ggplot(train_no_id, aes(x = `Flight Distance`))+geom_density(aes(color = satisfaction)) # may not i
g3<-ggplot(train_no_id, aes(x = `Departure Delay in Minutes`))+geom_density(aes(color = satisfaction))
g4<-ggplot(train_no_id, aes(x = `Arrival Delay in Minutes`))+geom_density(aes(color = satisfaction)) #
#grid.arrange(g1,g2,ncol=2)
g1
```


0.03
0.02
0.01
0.00
density
20
40
60
80
Age
satisfaction
neutral or dissatisfied
satisfied

```
g2
```

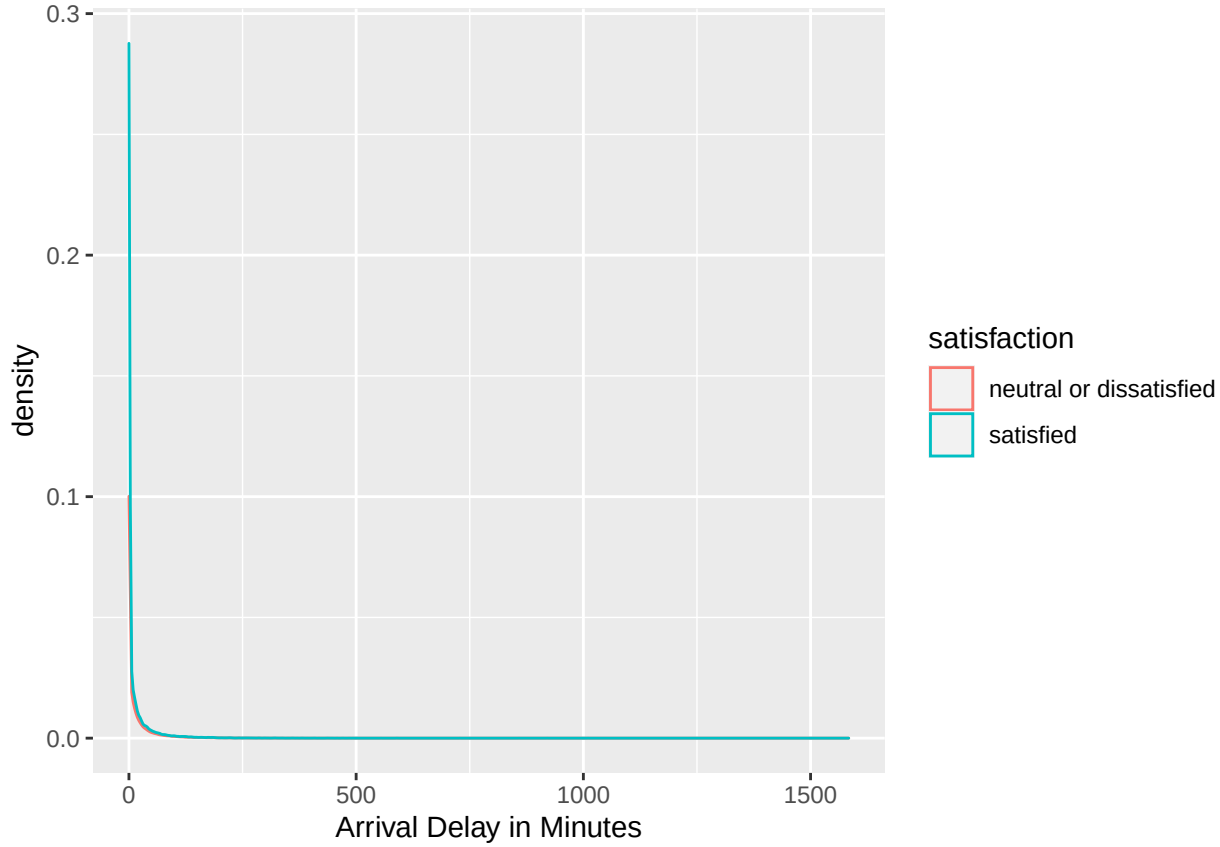

0.00100
0.00075
0.00050
0.00025
0.00000
density
0
1000
2000
3000
4000
5000
Flight Distance
satisfaction
neutral or dissatisfied
satisfied

```
g3
```

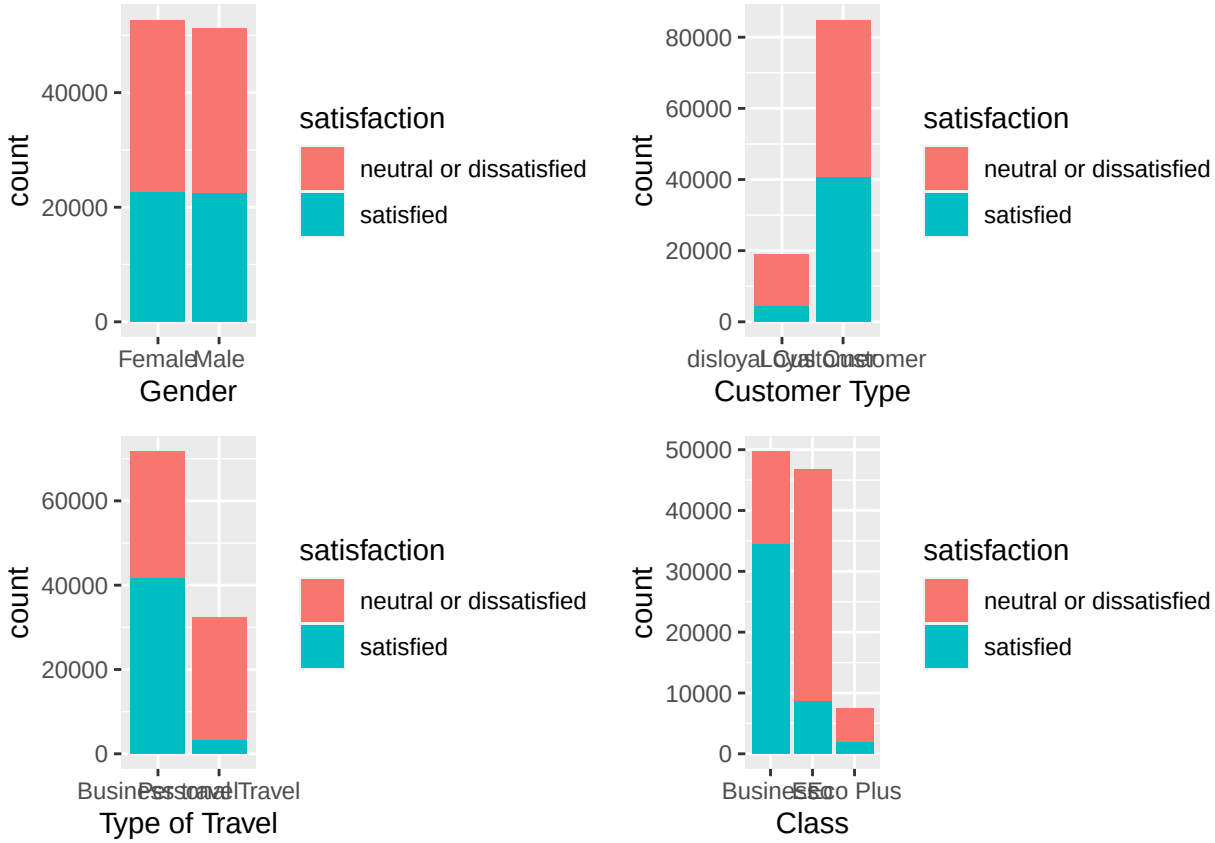



```
g4
```

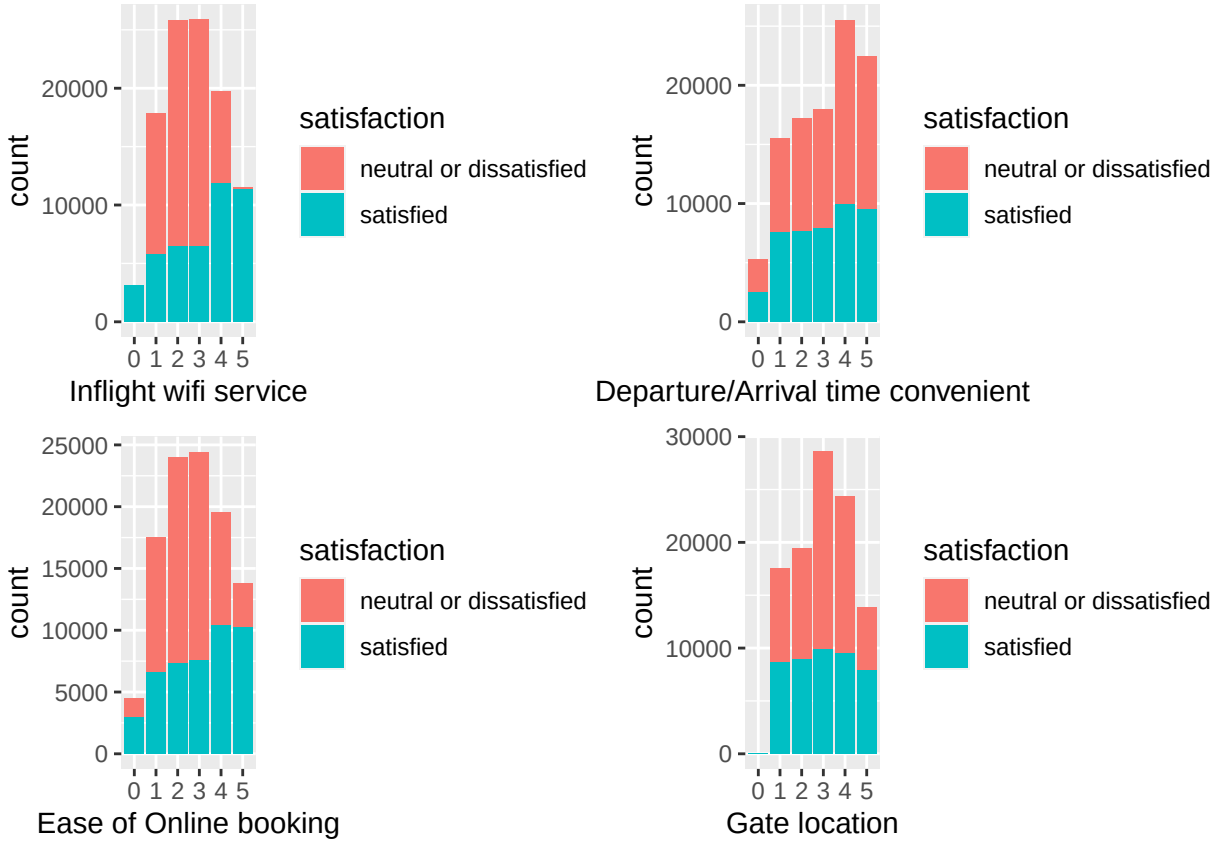
```
## Warning: Removed 310 rows containing non-finite values (stat_density).
```



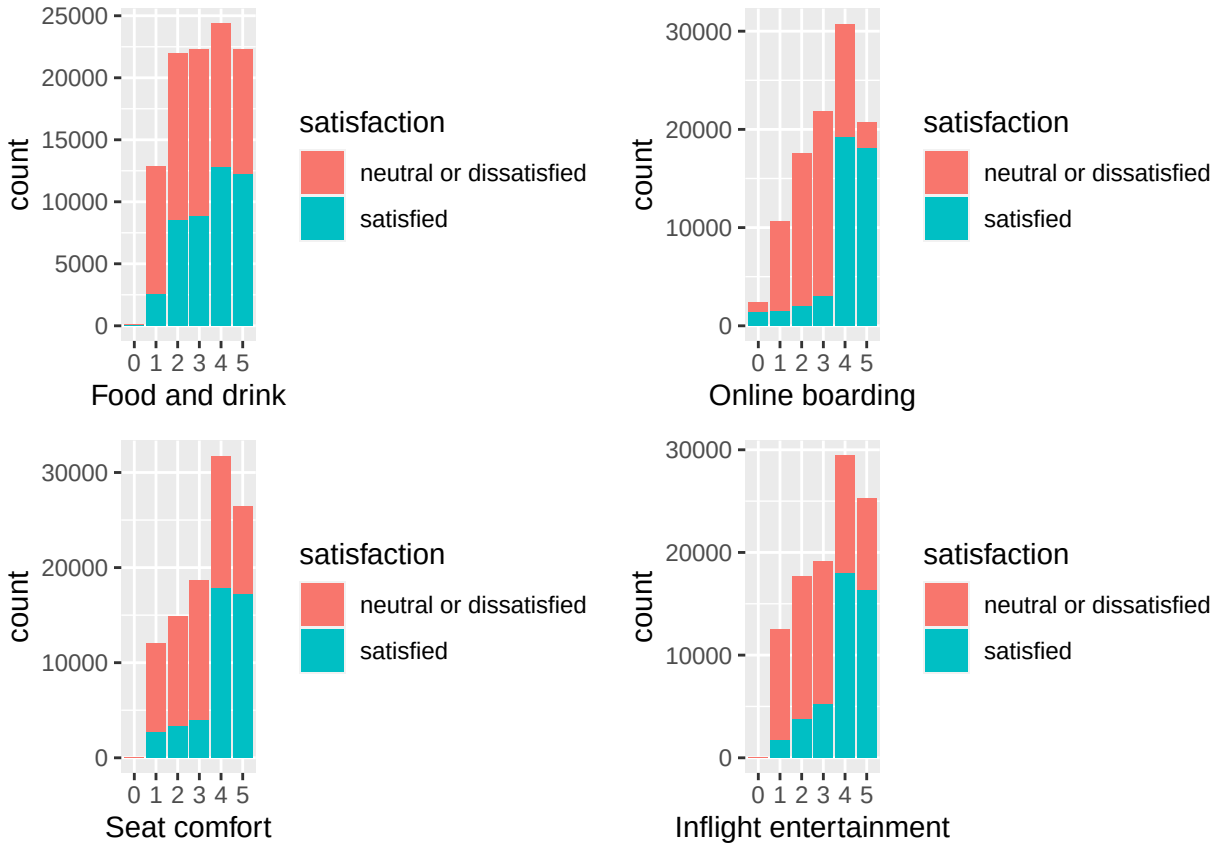

```
# categorical variable
g5<-ggplot(train_no_id, aes(x=Gender,fill=satisfaction)) + geom_bar() # not in model
g6<-ggplot(train_no_id, aes(x=`Customer Type`,fill=satisfaction)) + geom_bar() # in model
g7<-ggplot(train_no_id, aes(x=`Type of Travel`,fill=satisfaction)) + geom_bar() # in model
g8<-ggplot(train_no_id, aes(x=Class,fill=satisfaction)) + geom_bar() # in model
g9<-ggplot(train_no_id, aes(x=`Inflight wifi service`,fill=satisfaction)) + geom_bar() # in model
g10<-ggplot(train_no_id, aes(x=`Departure/Arrival time convenient`,fill=satisfaction)) + geom_bar() # i
g11<-ggplot(train_no_id, aes(x=`Ease of Online booking`,fill=satisfaction)) + geom_bar() # in model
g12<-ggplot(train_no_id, aes(x=`Gate location`,fill=satisfaction)) + geom_bar() # in model
g13<-ggplot(train_no_id, aes(x=`Food and drink`,fill=satisfaction)) + geom_bar() # in model
g14<-ggplot(train_no_id, aes(x=`Online boarding`,fill=satisfaction)) + geom_bar() # in model
g15<-ggplot(train_no_id, aes(x=`Seat comfort`,fill=satisfaction)) + geom_bar() # in model
g16<-ggplot(train_no_id, aes(x=`Inflight entertainment`,fill=satisfaction)) + geom_bar() # in model
g17<-ggplot(train_no_id, aes(x=`On-board service`,fill=satisfaction)) + geom_bar() # in model
g18<-ggplot(train_no_id, aes(x=`Leg room service`,fill=satisfaction)) + geom_bar() # in model
g19<-ggplot(train_no_id, aes(x=`Baggage handling`,fill=satisfaction)) + geom_bar() # in model
g20<-ggplot(train_no_id, aes(x=`Checkin service`,fill=satisfaction)) + geom_bar() # in model
g21<-ggplot(train_no_id, aes(x=`Inflight service`,fill=satisfaction)) + geom_bar() # in model
g22<-ggplot(train_no_id, aes(x=Cleanliness,fill=satisfaction)) + geom_bar() # in model
grid.arrange(g5,g6,g7,g8,ncol=2)
```



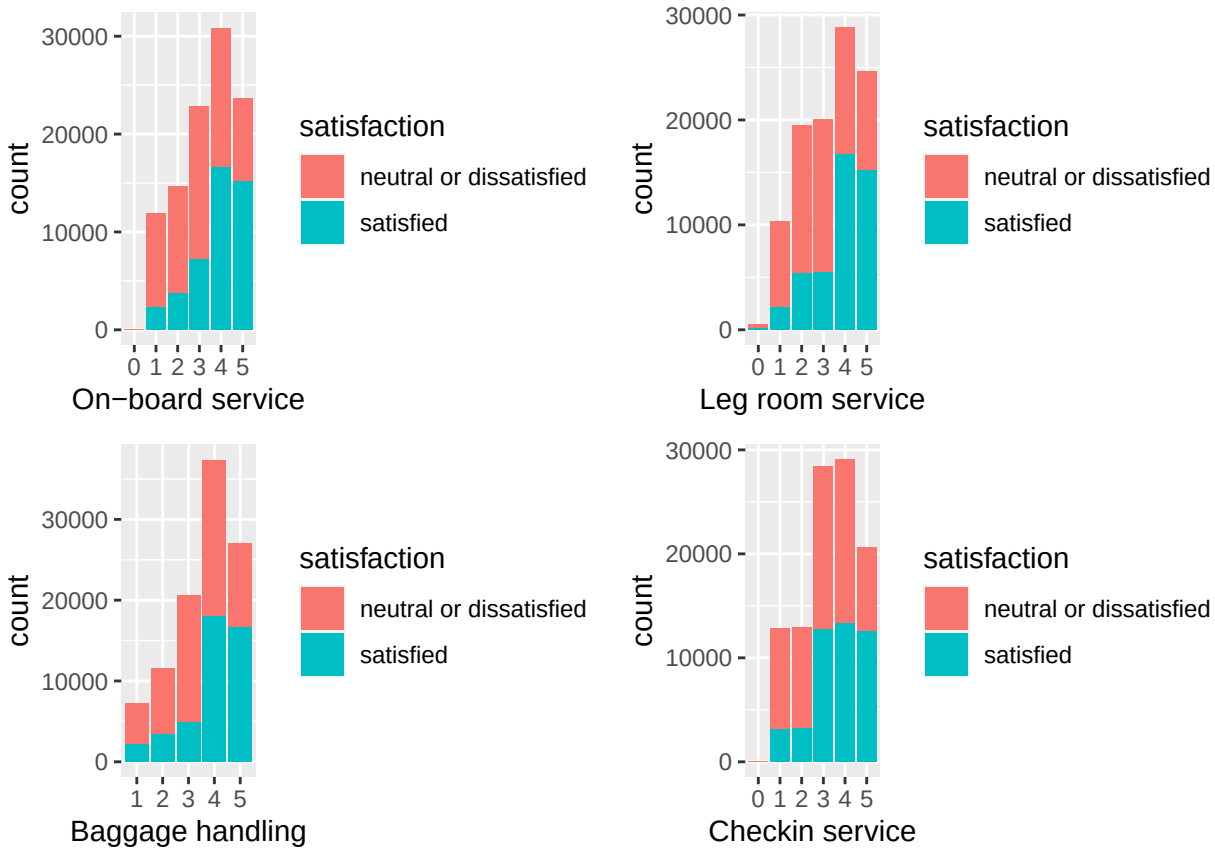

```
grid.arrange(g9,g10,g11,g12,ncol=2)
```



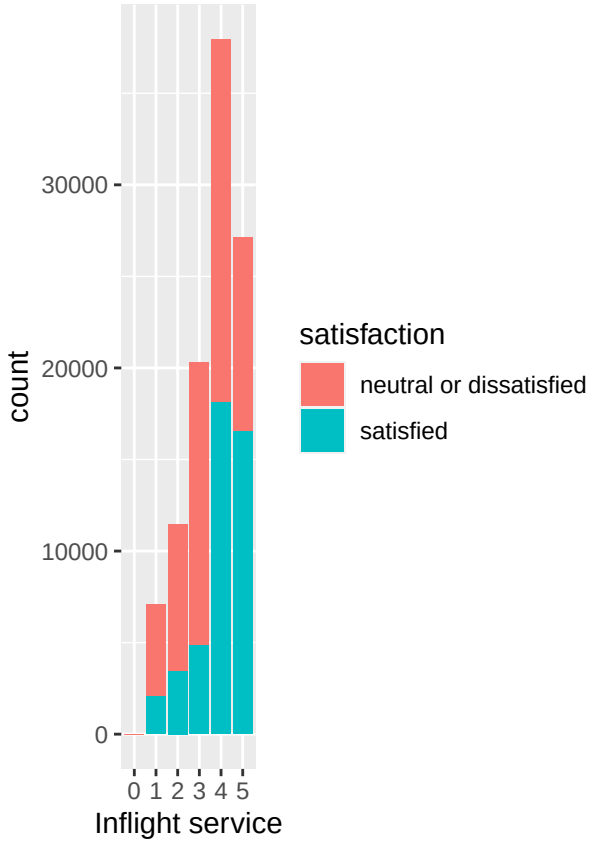

```
grid.arrange(g13,g14,g15,g16,ncol=2)
```

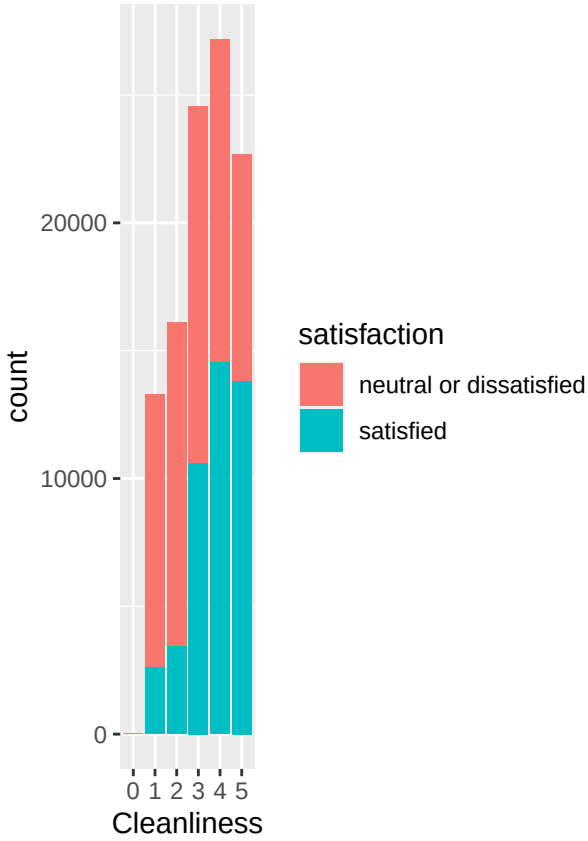



```
grid.arrange(g17,g18,g19,g20,ncol=2)
```




```
grid.arrange(g21,g22,ncol=2)
```




## Logistic Model