Plots and tables

Jialiang Hua

2022-05-12

```
library(tidyverse)
library(skimr)
library(caret)
library(visdat)
library(corrplot)
library(AppliedPredictiveModeling)
library(pROC)
library(rpart.plot)
library(vip)
library(ranger)
library(tidytext)
library(pdp)
library(lime)


ctrl <- trainControl(method = "cv",
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE)

knitr::opts_chunk$set(
  fig.width = 6,
  out.width = "80%",
  fig.align = "center"
  )
```

Data pre-process

```
# Import data
dat_raw <- read.csv("airline.csv")

# find unique value of each column
# sapply(dat_raw, function(x) length(unique(x)))

# Check missing value
# sapply(dat_raw, function(x) sum(is.na(x)))

# Have a glance of the data
skimr::skim_without_charts(dat_raw)
```

Table 1: Data summary

Name	dat_raw

Table 1: Data summary

Number of rows	129880
Number of columns	24
Column type frequency:	
character	5
numeric	19
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Gender	0	1	4	6	0	2	0
customer_type	0	1	14	17	0	2	0
type_of_travel	0	1	15	15	0	2	0
customer_class	0	1	3	8	0	3	0
satisfaction	0	1	9	23	0	2	0

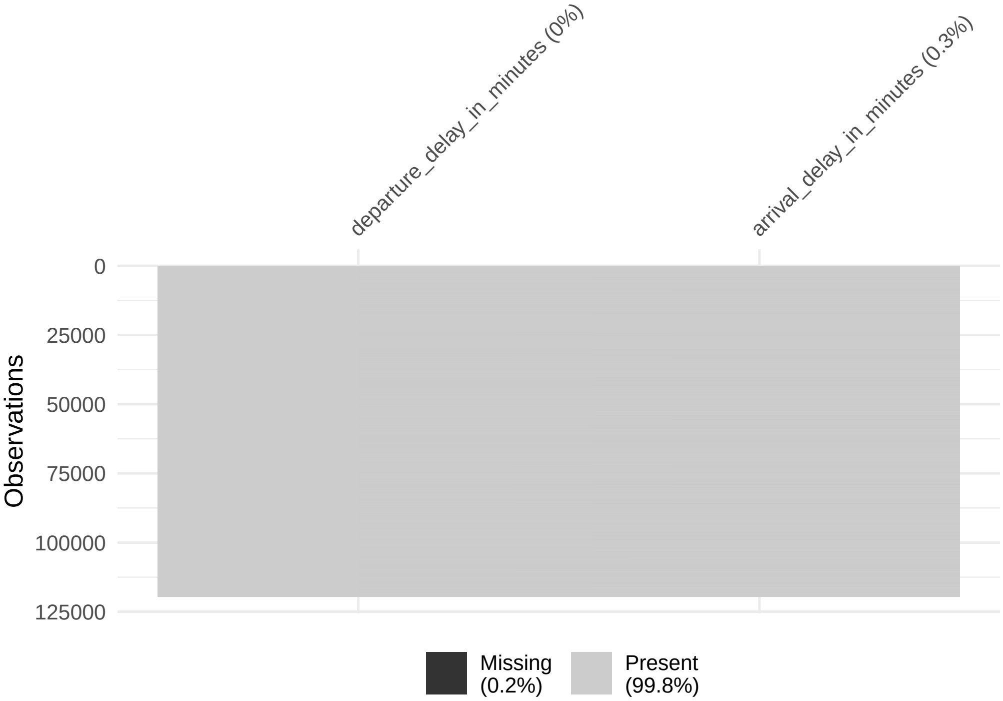
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
X	0	1	64939.50	37493.27	0	32469.75	64939.5	97409.25	129879
age	0	1	39.43	15.12	7	27.00	40.0	51.00	85
flight_distance	0	1	1190.32	997.45	31	414.00	844.0	1744.00	4983
inflight_wifi_service	0	1	2.73	1.33	0	2.00	3.0	4.00	5
departure_arrival_time_convenient	0	1	3.06	1.53	0	2.00	3.0	4.00	5
ease_of_online_booking	0	1	2.76	1.40	0	2.00	3.0	4.00	5
gate_location	0	1	2.98	1.28	0	2.00	3.0	4.00	5
food_and_drink	0	1	3.20	1.33	0	2.00	3.0	4.00	5
online_boarding	0	1	3.25	1.35	0	2.00	3.0	4.00	5
seat_comfort	0	1	3.44	1.32	0	2.00	4.0	5.00	5
inflight_entertainment	0	1	3.36	1.33	0	2.00	4.0	4.00	5
onboard_service	0	1	3.38	1.29	0	2.00	4.0	4.00	5
leg_room_service	0	1	3.35	1.32	0	2.00	4.0	4.00	5
baggage_handling	0	1	3.63	1.18	1	3.00	4.0	5.00	5
checkin_service	0	1	3.31	1.27	0	3.00	3.0	4.00	5
inflight_service	0	1	3.64	1.18	0	3.00	4.0	5.00	5
cleanliness	0	1	3.29	1.31	0	2.00	3.0	4.00	5
departure_delay_in_minutes	0	1	14.71	38.07	0	0.00	0.0	12.00	1592
arrival_delay_in_minutes	393	1	15.09	38.47	0	0.00	0.0	13.00	1584

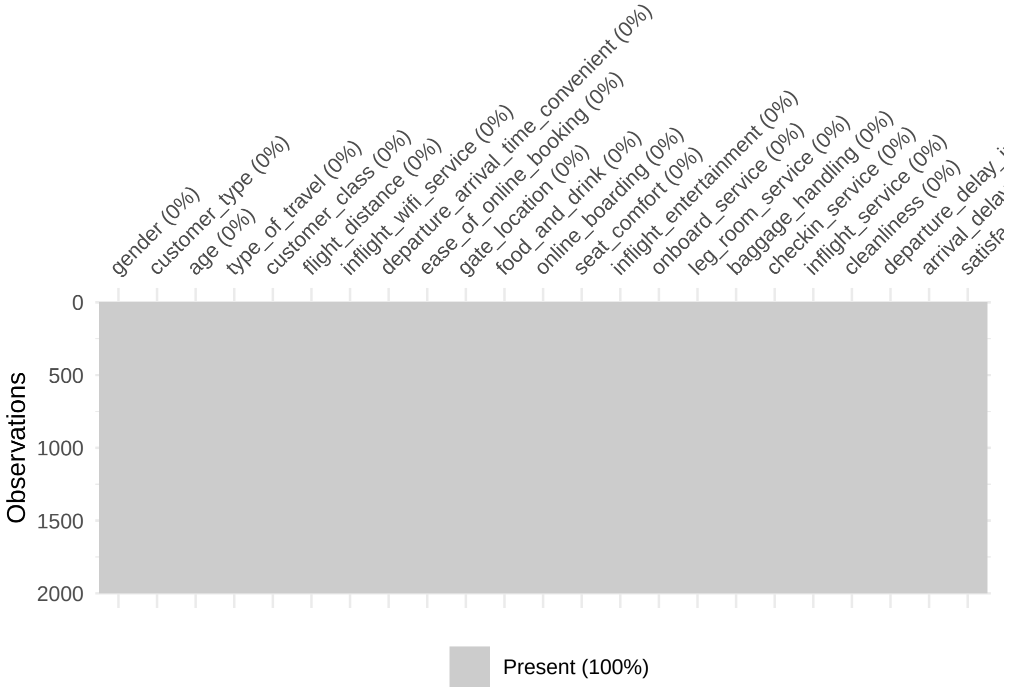
```r
# data clean
dat <- dat_raw %>%
  janitor::clean_names() %>%
  select(-1) %>%
  mutate(satisfaction = recode(satisfaction,
                         "satisfied" = "yes",
                         "neutral or dissatisfied" = "no")) %>%
  filter_at(vars(7:20), all_vars(. > 0.5))

# deal with missing values
deal_mis <- dat[, 21:22]
bagImp = preProcess(deal_mis, method = "bagImpute")
dat = predict(bagImp, dat)
vis_miss(deal_mis)
```

```
## Warning: `gather_()` was deprecated in tidyr 1.2.0.
## Please use `gather()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```



```
# sample data
set.seed(1234)
dat <- dat[sample(1:nrow(dat), 2000, replace = FALSE), ]
vis_miss(dat) ## check
```



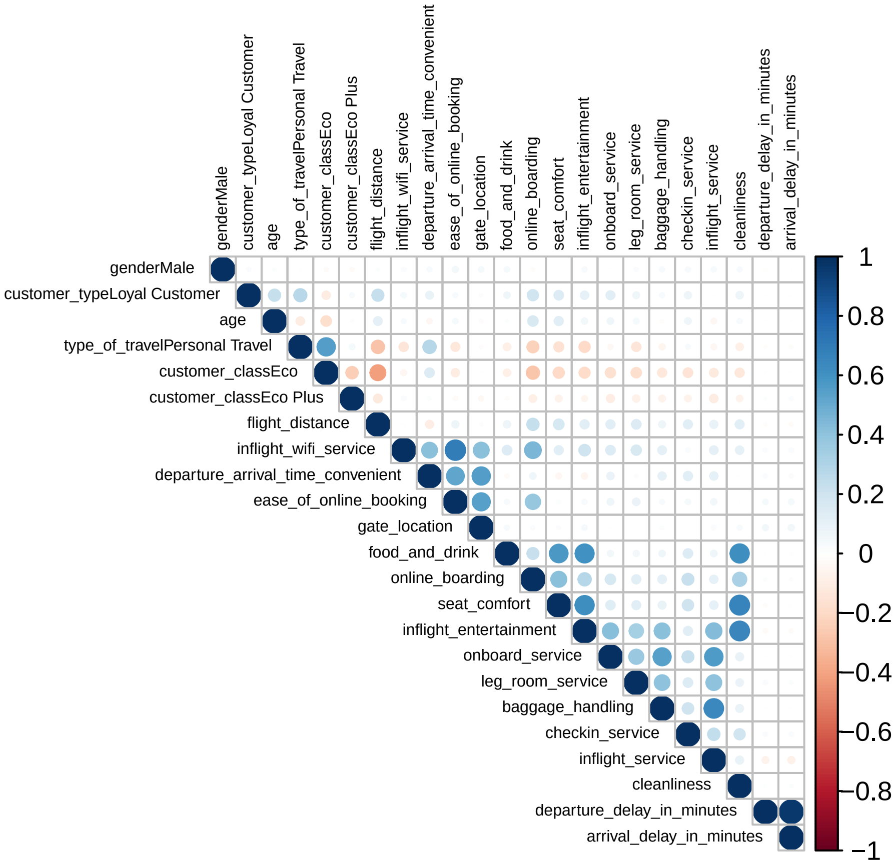
```r
# --- Split data ---
set.seed(1234)
trRow <- createDataPartition(dat$satisfaction, p = 0.8, list = F)

# Train data
train <- dat[trRow, ]
x_train <- model.matrix(satisfaction ~., train)[,-1]
y_train <- train$satisfaction

# Test data
test <- dat[-trRow, ]
x_test <- model.matrix(satisfaction ~., test)[,-1]
y_test <- test$satisfaction
```

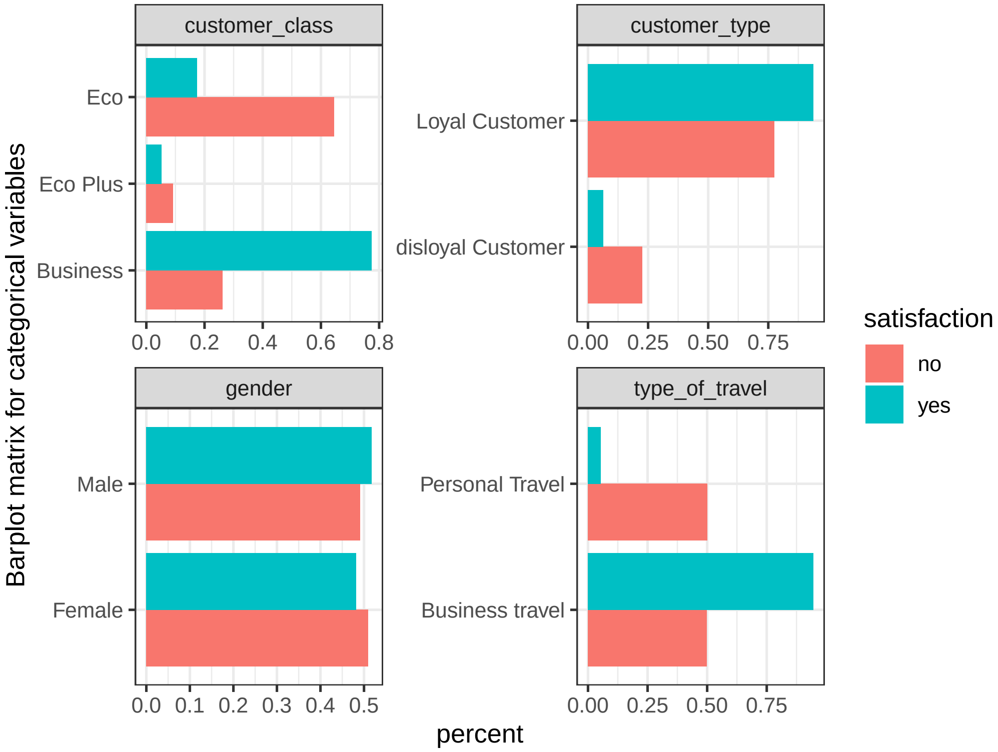
EDA

```r
# Correlation plot
corrplot(cor(x_train),
         method = "circle",
         type = "upper",
         tl.col = "black",
         tl.cex = 0.5)
```



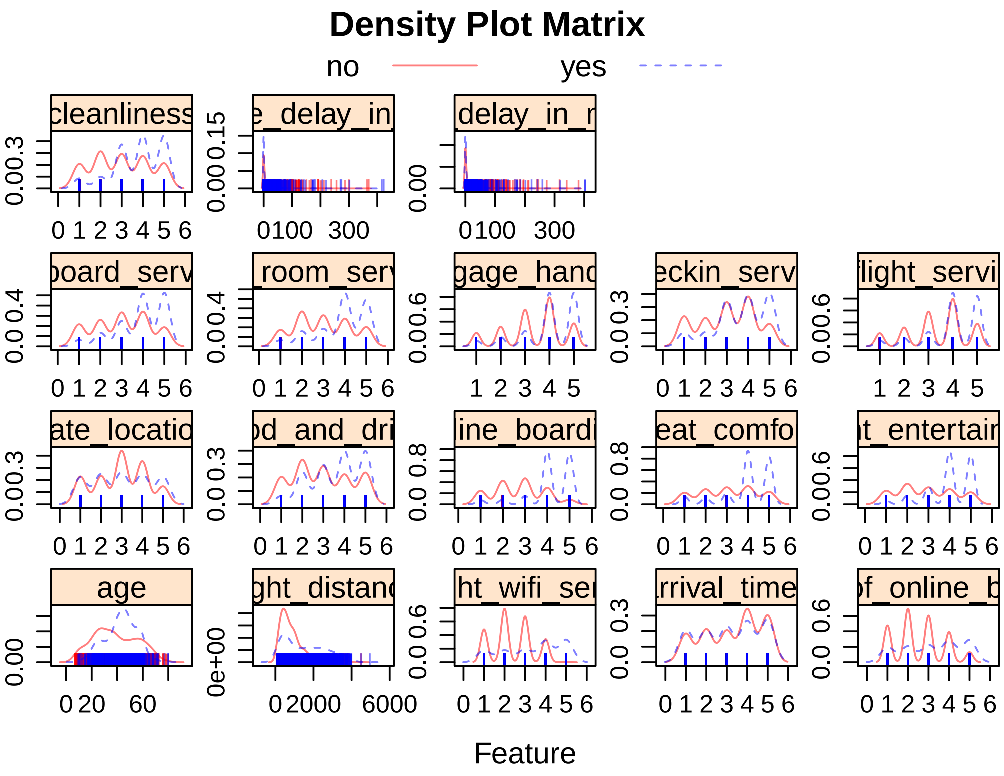
```
# Barplot matrix for categorical variables
train %>%
  select(1:2, 4:5, 23) %>%
  pivot_longer(-5,
               names_to = "variable",
               values_to = "value") %>%
  group_by(variable, value, satisfaction) %>%
  summarize(num = n()) %>%
  ungroup() %>%
  group_by(variable, satisfaction) %>%
  mutate(percent = num / sum(num),
         indicator = case_when(value == "Eco" ~ 3,
                               value == "Eco Plus" ~ 2,
                               value == "Business" ~ 1,
                               TRUE ~ 0)) %>%
  ggplot(aes(x = reorder_within(value, indicator, variable),
             y = percent, fill = satisfaction)) +
  geom_col(position = "dodge") +
  xlab("Barplot matrix for categorical variables") +
  coord_flip() +
  scale_x_reordered() +
  facet_wrap(~ variable, scales = "free") + theme_bw()
```

```
## `summarise()` has grouped output by 'variable', 'value'. You can override using
## the `.groups` argument.
```



```
# Density plot matrix
theme1 <- transparentTheme(trans = .5)
trellis.par.set(theme1)

plt_feature <-
  featurePlot(x = x_train[ , c(-1, -2, -4, -5, -6)],
              y = as.factor(y_train),
              plot = "density",
              scales = list(x = list(relation = "free"),
                            y = list(relation = "free")),
              pch = "|", auto.key = list(columns = 2))
update(plt_feature, main = "Density Plot Matrix")
```


Density Plot Matrix
no
yes
cleanliness
e_delay_in
delay_in_r
board_serv
room_serv
gage_hanc
eckin_serv
flight_servi
ate_locatio
od_and_dri
line_boardi
eat_comfo
t_entertain
age
ght_distanc
ht_wifi_ser
rrival_time
of_online_b
0.0 0.3
0.00 0.15
0.00
0.0 0.4
0.0 0.4
0.0 0.6
0.0 0.3
0.0 0.6
0.0 0.3
0.0 0.3
0.0 0.8
0.0 0.8
0.0 0.6
0.00
0e+00
0.0 0.6
0.0 0.3
0.0 0.6
0 1 2 3 4 5 6
0 100 300
1 2 3 4 5
0 20 60
0 2000 6000
Feature