Project_1_classification

Muhammad Zubair

10/14/2021

Link to dataset: https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction

I apologize in advance for serval pages full of warnings, I was unable to get a correlation mattrix with the regular function.

Reading in the data

```
df <- read.csv("airline_passenger_satisfaction.csv")
```

Data Exploration

```
# Dimensions of our data satisfies the assigment requirements
dim(df)
```

Our dataset have about 130k rows and 24 attributes

```
## [1] 129880     24
```

```
# Information about our data, and column types
str(df)
```

The columns will be converted into factor data types to enahnce the process for building model

```
## 'data.frame':    129880 obs. of  24 variables:
##  $ X                               : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Gender                          : chr  "Male" "Male" "Female" "Female" ...
##  $ customer_type                   : chr  "Loyal Customer" "disloyal Customer" "Loyal Customer" "Loy
##  $ age                             : int  13 25 26 25 61 26 47 52 41 20 ...
##  $ type_of_travel                  : chr  "Personal Travel" "Business travel" "Business travel" "Bus
##  $ customer_class                  : chr  "Eco Plus" "Business" "Business" "Business" ...
```

```
##  $ flight_distance                  : int  460 235 1142 562 214 1180 1276 2035 853 1061 ...
##  $ inflight_wifi_service            : int  3 3 2 2 3 3 2 4 1 3 ...
##  $ departure_arrival_time_convenient: int  4 2 2 5 3 4 4 3 2 3 ...
##  $ ease_of_online_booking           : int  3 3 2 5 3 2 2 4 2 3 ...
##  $ gate_location                    : int  1 3 2 5 3 1 3 4 2 4 ...
##  $ food_and_drink                   : int  5 1 5 2 4 1 2 5 4 2 ...
##  $ online_boarding                  : int  3 3 5 2 5 2 2 5 3 3 ...
##  $ seat_comfort                     : int  5 1 5 2 5 1 2 5 3 3 ...
##  $ inflight_entertainment           : int  5 1 5 2 3 1 2 5 1 2 ...
##  $ onboard_service                  : int  4 1 4 2 3 3 3 5 1 2 ...
##  $ leg_room_service                 : int  3 5 3 5 4 4 3 5 2 3 ...
##  $ baggage_handling                 : int  4 3 4 3 4 4 4 5 1 4 ...
##  $ checkin_service                  : int  4 1 4 1 3 4 3 4 4 4 ...
##  $ inflight_service                 : int  5 4 4 4 3 4 5 5 1 3 ...
##  $ cleanliness                      : int  5 1 5 2 3 1 2 4 2 2 ...
##  $ departure_delay_in_minutes       : int  25 1 0 11 0 0 9 4 0 0 ...
##  $ arrival_delay_in_minutes         : num  18 6 0 9 0 0 23 0 0 0 ...
##  $ satisfaction                     : chr  "neutral or dissatisfied" "neutral or dissatisfied" "satis
```

```
# Viewing first 5 rows of data
head(df)
```

Getting a general idea of how the data looks like

```
##   X Gender     customer_type age  type_of_travel customer_class flight_distance
## 1 0   Male    Loyal Customer  13 Personal Travel       Eco Plus             460
## 2 1   Male disloyal Customer  25 Business travel       Business             235
## 3 2 Female    Loyal Customer  26 Business travel       Business            1142
## 4 3 Female    Loyal Customer  25 Business travel       Business             562
## 5 4   Male    Loyal Customer  61 Business travel       Business             214
## 6 5 Female    Loyal Customer  26 Personal Travel            Eco            1180
##   inflight_wifi_service departure_arrival_time_convenient
## 1                     3                                 4
## 2                     3                                 2
## 3                     2                                 2
## 4                     2                                 5
## 5                     3                                 3
## 6                     3                                 4
##   ease_of_online_booking gate_location food_and_drink online_boarding
## 1                      3             1              5               3
## 2                      3             3              1               3
## 3                      2             2              5               5
## 4                      5             5              2               2
## 5                      3             3              4               5
## 6                      2             1              1               2
##   seat_comfort inflight_entertainment onboard_service leg_room_service
## 1            5                      5              4                3
## 2            1                      1              1                5
## 3            5                      5              4                3
## 4            2                      2              2                5
## 5            5                      3              3                4
```

```
## 6            1                     1               3               4
##   baggage_handling checkin_service inflight_service cleanliness
## 1                4               4                 5           5
## 2                3               1                 4           1
## 3                4               4                 4           5
## 4                3               1                 4           2
## 5                4               3                 3           3
## 6                4               4                 4           1
##   departure_delay_in_minutes arrival_delay_in_minutes            satisfaction
## 1                         25                       18 neutral or dissatisfied
## 2                          1                        6 neutral or dissatisfied
## 3                          0                        0               satisfied
## 4                         11                        9 neutral or dissatisfied
## 5                          0                        0               satisfied
## 6                          0                        0 neutral or dissatisfied
```

```
# Viewing the last 5 rows of data
tail(df)
```

```
##             X Gender     customer_type age  type_of_travel customer_class
## 129875 129874 Female disloyal Customer  36 Business travel            Eco
## 129876 129875   Male disloyal Customer  34 Business travel       Business
## 129877 129876   Male    Loyal Customer  23 Business travel       Business
## 129878 129877 Female    Loyal Customer  17 Personal Travel            Eco
## 129879 129878   Male    Loyal Customer  14 Business travel       Business
## 129880 129879 Female    Loyal Customer  42 Personal Travel            Eco
##        flight_distance inflight_wifi_service departure_arrival_time_convenient
## 129875             432                     1                                 5
## 129876             526                     3                                 3
## 129877             646                     4                                 4
## 129878             828                     2                                 5
## 129879            1127                     3                                 3
## 129880             264                     2                                 5
##        ease_of_online_booking gate_location food_and_drink online_boarding
## 129875                      1             3              4                1
## 129876                      3             1              4                3
## 129877                      4             4              4                4
## 129878                      1             5              2                1
## 129879                      3             3              4                4
## 129880                      2             5              4                2
##        seat_comfort inflight_entertainment onboard_service leg_room_service
## 129875            4                      4                5                2
## 129876            4                      4                3                2
## 129877            4                      4                4                5
## 129878            2                      2                4                3
## 129879            4                      4                3                2
## 129880            2                      1                1                2
##        baggage_handling checkin_service inflight_service cleanliness
## 129875                5               2                 3           4
## 129876                4               4                 5           4
## 129877                5               5                 5           4
## 129878                4               5                 4           2
## 129879                5               4                 5           4
## 129880                1               1                 1           1
```

```
##        departure_delay_in_minutes arrival_delay_in_minutes
## 129875                          0                        0
## 129876                          0                        0
## 129877                          0                        0
## 129878                          0                        0
## 129879                          0                        0
## 129880                          0                        0
##                   satisfaction
## 129875 neutral or dissatisfied
## 129876 neutral or dissatisfied
## 129877               satisfied
## 129878 neutral or dissatisfied
## 129879               satisfied
## 129880 neutral or dissatisfied
```

```
# statiscal metrics of numeric variables in data
summary(df)
```

These values help us interept what kind of data we're dealing with. For example, most people that travel on a plane are about 30-40 years old and avergae distance of flight is about 1190 miles. We can further use these values to find pattern in the data.

```
##        X             Gender          customer_type           age
##  Min.   :     0   Length:129880      Length:129880      Min.   : 7.00
##  1st Qu.: 32470   Class :character   Class :character   1st Qu.:27.00
##  Median : 64940   Mode  :character   Mode  :character   Median :40.00
##  Mean   : 64940                                         Mean   :39.43
##  3rd Qu.: 97409                                         3rd Qu.:51.00
##  Max.   :129879                                         Max.   :85.00
##
##  type_of_travel     customer_class     flight_distance inflight_wifi_service
##  Length:129880      Length:129880      Min.   :  31    Min.   :0.000
##  Class :character   Class :character   1st Qu.: 414    1st Qu.:2.000
##  Mode  :character   Mode  :character   Median : 844    Median :3.000
##                                        Mean   :1190    Mean   :2.729
##                                        3rd Qu.:1744    3rd Qu.:4.000
##                                        Max.   :4983    Max.   :5.000
##
##  departure_arrival_time_convenient ease_of_online_booking gate_location
##  Min.   :0.000                     Min.   :0.000          Min.   :0.000
##  1st Qu.:2.000                     1st Qu.:2.000          1st Qu.:2.000
##  Median :3.000                     Median :3.000          Median :3.000
##  Mean   :3.058                     Mean   :2.757          Mean   :2.977
##  3rd Qu.:4.000                     3rd Qu.:4.000          3rd Qu.:4.000
##  Max.   :5.000                     Max.   :5.000          Max.   :5.000
##
##  food_and_drink  online_boarding  seat_comfort   inflight_entertainment
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :3.000   Median :4.000   Median :4.000
##  Mean   :3.205   Mean   :3.253   Mean   :3.441   Mean   :3.358
```

```
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##
##  onboard_service leg_room_service baggage_handling checkin_service
##  Min.   :0.000   Min.   :0.000    Min.   :1.000    Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.000    1st Qu.:3.000    1st Qu.:3.000
##  Median :4.000   Median :4.000    Median :4.000    Median :3.000
##  Mean   :3.383   Mean   :3.351    Mean   :3.632    Mean   :3.306
##  3rd Qu.:4.000   3rd Qu.:4.000    3rd Qu.:5.000    3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000    Max.   :5.000    Max.   :5.000
##
##  inflight_service  cleanliness    departure_delay_in_minutes
##  Min.   :0.000    Min.   :0.000   Min.   :   0.00
##  1st Qu.:3.000    1st Qu.:2.000   1st Qu.:   0.00
##  Median :4.000    Median :3.000   Median :   0.00
##  Mean   :3.642    Mean   :3.286   Mean   :  14.71
##  3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:  12.00
##  Max.   :5.000    Max.   :5.000   Max.   :1592.00
##
##  arrival_delay_in_minutes satisfaction
##  Min.   :   0.00          Length:129880
##  1st Qu.:   0.00          Class :character
##  Median :   0.00          Mode  :character
##  Mean   :  15.09
##  3rd Qu.:  13.00
##  Max.   :1584.00
##  NA's   :393
```

```
# Checking for null values in data set
sapply(df, function(x) sum(is.na(x)))
```

One attribute have 393 missing values.

```
##                                  X                             Gender
##                                  0                                  0
##                      customer_type                                age
##                                  0                                  0
##                     type_of_travel                     customer_class
##                                  0                                  0
##                    flight_distance              inflight_wifi_service
##                                  0                                  0
## departure_arrival_time_convenient            ease_of_online_booking
##                                  0                                  0
##                      gate_location                     food_and_drink
##                                  0                                  0
##                    online_boarding                       seat_comfort
##                                  0                                  0
##             inflight_entertainment                    onboard_service
##                                  0                                  0
##                   leg_room_service                   baggage_handling
##                                  0                                  0
```

```
##                   checkin_service                   inflight_service
##                                 0                                 0
##                       cleanliness       departure_delay_in_minutes
##                                 0                                 0
##         arrival_delay_in_minutes                      satisfaction
##                               393                                 0
```

Data cleaning

```
# Dropping the X columns, as it is used to number the rows and wont have effect on algorithm
df <- subset(df, select = -c(X))
```

```
# chaging all the columns of type char to factor and integer to numeruc, so we can do Exploraitry data
df[sapply(df, is.character)] <- lapply(df[sapply(df, is.character)], as.factor)
df[sapply(df, is.integer)] <- lapply(df[sapply(df, is.integer)], as.numeric)
str(df)
```

```
## 'data.frame':    129880 obs. of  23 variables:
##  $ Gender                          : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 1 2 1 1 2 ...
##  $ customer_type                   : Factor w/ 2 levels "disloyal Customer",..: 2 1 2 2 2 2 2 2 2 1
##  $ age                             : num  13 25 26 25 61 26 47 52 41 20 ...
##  $ type_of_travel                  : Factor w/ 2 levels "Business travel",..: 2 1 1 1 1 2 2 1 1 1 .
##  $ customer_class                  : Factor w/ 3 levels "Business","Eco",..: 3 1 1 1 1 2 2 1 1 2 ..
##  $ flight_distance                 : num  460 235 1142 562 214 ...
##  $ inflight_wifi_service           : num  3 3 2 2 3 3 2 4 1 3 ...
##  $ departure_arrival_time_convenient: num  4 2 2 5 3 4 4 3 2 3 ...
##  $ ease_of_online_booking          : num  3 3 2 5 3 2 2 4 2 3 ...
##  $ gate_location                   : num  1 3 2 5 3 1 3 4 2 4 ...
##  $ food_and_drink                  : num  5 1 5 2 4 1 2 5 4 2 ...
##  $ online_boarding                 : num  3 3 5 2 5 2 2 5 3 3 ...
##  $ seat_comfort                    : num  5 1 5 2 5 1 2 5 3 3 ...
##  $ inflight_entertainment          : num  5 1 5 2 3 1 2 5 1 2 ...
##  $ onboard_service                 : num  4 1 4 2 3 3 3 5 1 2 ...
##  $ leg_room_service                : num  3 5 3 5 4 4 3 5 2 3 ...
##  $ baggage_handling                : num  4 3 4 3 4 4 4 5 1 4 ...
##  $ checkin_service                 : num  4 1 4 1 3 4 3 4 4 4 ...
##  $ inflight_service                : num  5 4 4 4 3 4 5 5 1 3 ...
##  $ cleanliness                     : num  5 1 5 2 3 1 2 4 2 2 ...
##  $ departure_delay_in_minutes      : num  25 1 0 11 0 0 9 4 0 0 ...
##  $ arrival_delay_in_minutes        : num  18 6 0 9 0 0 23 0 0 0 ...
##  $ satisfaction                    : Factor w/ 2 levels "neutral or dissatisfied",..: 1 1 2 1 2 1 1
```

```
# Droppping the rows that conatined NA values because there are only 393 rows compared to our datastet
df <- na.omit(df)
sapply(df, function(x) sum(is.na(x)))
```

```
##                             Gender                      customer_type
##                                  0                                  0
##                                age                     type_of_travel
##                                  0                                  0
```

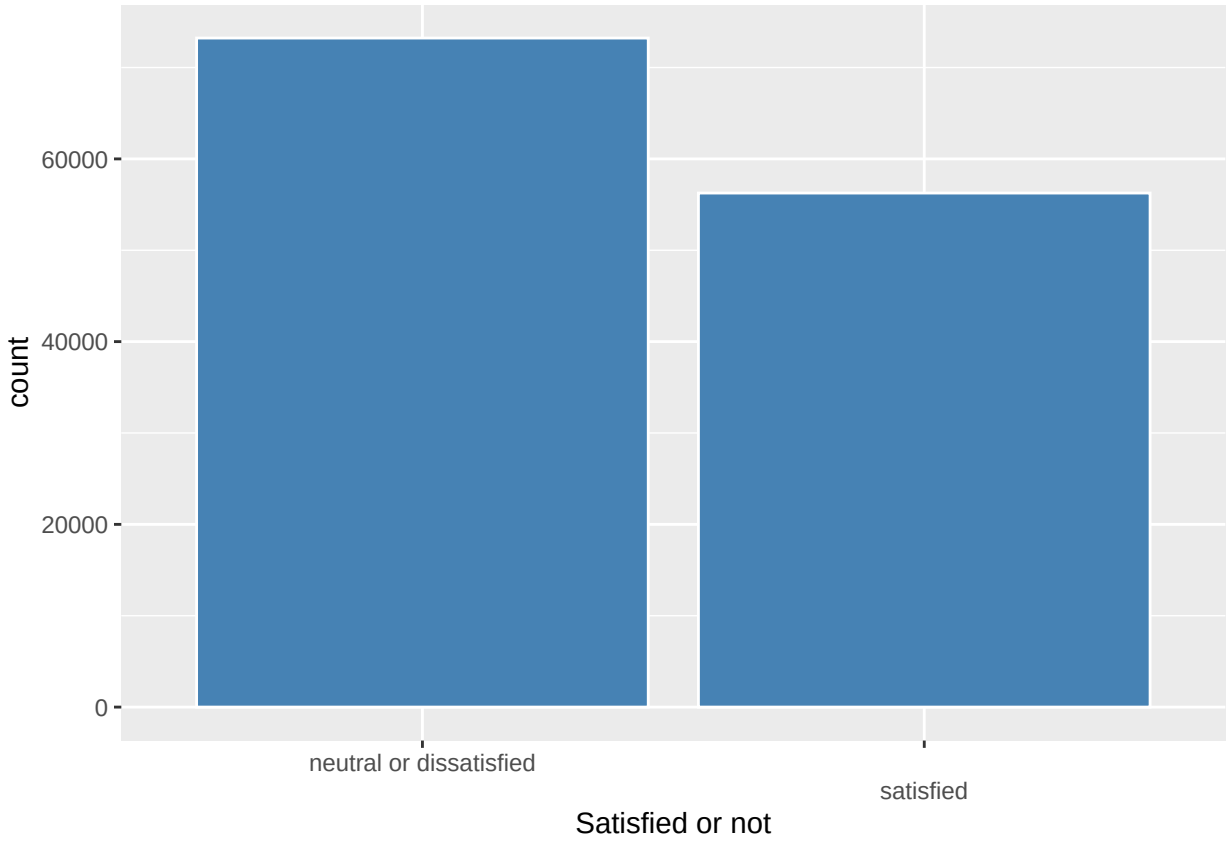
```
##                    customer_class                     flight_distance
##                                 0                                   0
##             inflight_wifi_service departure_arrival_time_convenient
##                                 0                                   0
##            ease_of_online_booking                       gate_location
##                                 0                                   0
##                    food_and_drink                    online_boarding
##                                 0                                   0
##                      seat_comfort             inflight_entertainment
##                                 0                                   0
##                   onboard_service                   leg_room_service
##                                 0                                   0
##                  baggage_handling                    checkin_service
##                                 0                                   0
##                  inflight_service                        cleanliness
##                                 0                                   0
##        departure_delay_in_minutes          arrival_delay_in_minutes
##                                 0                                   0
##                      satisfaction
##                                 0
```

Plots

```
# Checking to see if the dataset is balanced
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
ggplot(df, aes(x = satisfaction)) + geom_bar(color="white", fill = "steelblue") + scale_x_discrete(guid
```




```
library(tidyverse)
```

Columns like flight distance, deapprture/arrival time convinent, gate location, departure delay in minutes and arrival delay in minutes will be removed as they have low correlation with satisfaction.

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

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

```
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'purrr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## Warning: package 'forcats' was built under R version 4.0.5

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

```
library(lsr)
```

```
## Warning: package 'lsr' was built under R version 4.0.5
```

```
data <- df[sapply(df, is.numeric)]

# Randomly getting only 10% of the data to speed up process for EDA, and avoid error of "cannot allocat
set.seed(123)
index <- sample(1:nrow(data), 0.01*nrow(data), replace = FALSE)

small_df_numeric <- data[index,]

# function to get chi square p value and Cramers V
f = function(x,y) {
    tbl = df %>% select(x,y) %>% table()
    chisq_pval = round(chisq.test(tbl)$p.value, 4)
    cramV = round(cramersV(tbl), 4)
    data.frame(x, y, chisq_pval, cramV) }

# create unique combinations of column names
# sorting will help getting a better plot (upper triangular)
df_comb = data.frame(t(combn(sort(names(df)), 2)), stringsAsFactors = F)

# apply function to each variable combination
df_res = map2_df(df_comb$X1, df_comb$X2, f)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(x)` instead of `x` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.

## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(y)` instead of `y` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

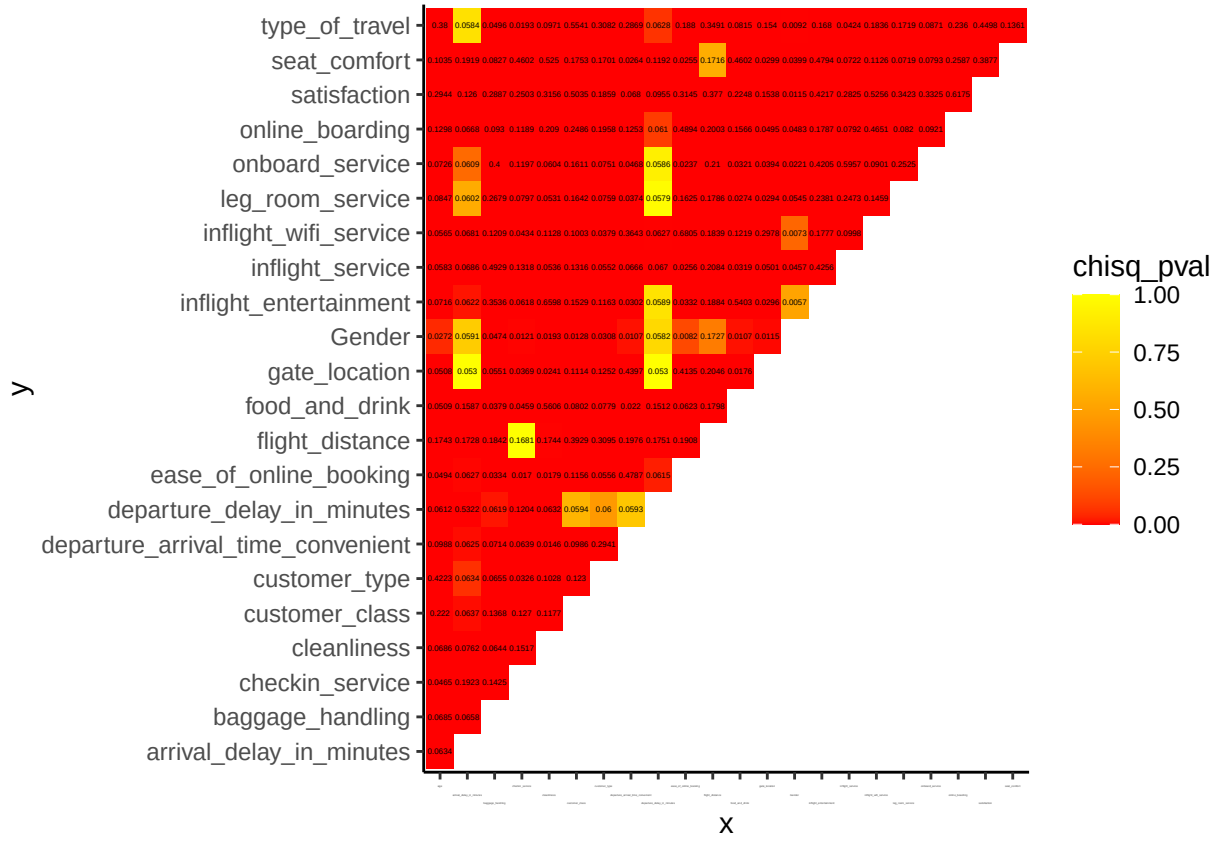
```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect

## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect

## Warning in stats::chisq.test(...): Chi-squared approximation may be incorrect
```

```
# plot results
df_res %>%
  ggplot(aes(x,y,fill=chisq_pval))+
  geom_tile()+
  geom_text(size = 1, aes(x,y,label=cramV))+
  scale_fill_gradient(low="red", high="yellow")+
  theme_classic() + scale_x_discrete(guide = guide_axis(n.dodge=3)) +
  theme(axis.text.x = element_text(size = 1))
```




```
# Removing the columns
df <- subset(df, select = -c(arrival_delay_in_minutes, arrival_delay_in_minutes, gate_location, departu
```

Building the model

```
# Splitting the data into trian/test data
set.seed(1234)
i <- sample(1:nrow(df), 0.75*nrow(df), replace = FALSE)
train <- df[i,]
test <- df[-i,]
```

Logistic regression

```
glm1 <- glm(satisfaction~., data = train, family = "binomial")
summary(glm1)
```

We can see that all of these observations are good predictors for satisfaction because of the three * next to them.**

```
##
## Call:
## glm(formula = satisfaction ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8375  -0.4982  -0.1812   0.3943   3.9901
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -7.788e+00  7.850e-02 -99.208  < 2e-16 ***
## GenderMale                     7.482e-02  1.998e-02   3.744 0.000181 ***
## customer_typeLoyal Customer    1.988e+00  3.034e-02  65.509  < 2e-16 ***
## age                           -9.043e-03  7.286e-04 -12.411  < 2e-16 ***
## type_of_travelPersonal Travel -2.831e+00  3.177e-02 -89.095  < 2e-16 ***
## customer_classEco             -7.071e-01  2.633e-02 -26.854  < 2e-16 ***
## customer_classEco Plus        -8.173e-01  4.264e-02 -19.169  < 2e-16 ***
## flight_distance               -1.956e-05  1.161e-05  -1.685 0.092084 .
## inflight_wifi_service          3.925e-01  1.166e-02  33.648  < 2e-16 ***
## ease_of_online_booking        -2.158e-01  1.004e-02 -21.482  < 2e-16 ***
## food_and_drink                -3.073e-02  1.097e-02  -2.802 0.005074 **
## online_boarding                6.140e-01  1.021e-02  60.121  < 2e-16 ***
## seat_comfort                   5.714e-02  1.151e-02   4.965 6.88e-07 ***
## inflight_entertainment         6.377e-02  1.462e-02   4.363 1.28e-05 ***
## onboard_service                2.883e-01  1.048e-02  27.507  < 2e-16 ***
## leg_room_service               2.623e-01  8.739e-03  30.016  < 2e-16 ***
## baggage_handling               1.203e-01  1.170e-02  10.280  < 2e-16 ***
## checkin_service                3.211e-01  8.785e-03  36.549  < 2e-16 ***
## inflight_service               1.248e-01  1.234e-02  10.111  < 2e-16 ***
## cleanliness                    2.282e-01  1.241e-02  18.399  < 2e-16 ***
## departure_delay_in_minutes    -4.731e-03  2.731e-04 -17.321  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 132939  on 97114  degrees of freedom
## Residual deviance:  65745  on 97094  degrees of freedom
## AIC: 65787
##
## Number of Fisher Scoring iterations: 5
```

```
library(caret)
```

Logistic regression performed fairly well, now lets try some other models and evaulate their results.

```
## Warning: package 'caret' was built under R version 4.0.5
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
# Evaluating the model on test data
probs <- predict(glm1, newdata=test, type="response")
pred <- ifelse(probs>0.5,"satisfied", "neutral or dissatisfied" )
# Accuracy score
acc <- mean(pred==(test$satisfaction))
print(paste("accuracy = ", acc))
```

```
## [1] "accuracy =  0.874953663659953"
```

```
confusionMatrix(as.factor(pred), test$satisfaction)
```

```
## Confusion Matrix and Statistics
##
##                           Reference
## Prediction                 neutral or dissatisfied satisfied
##   neutral or dissatisfied                   16488      2266
##   satisfied                                  1782     11836
##
##                Accuracy : 0.875
##                  95% CI : (0.8713, 0.8785)
##     No Information Rate : 0.5644
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7447
##
##  Mcnemar's Test P-Value : 3.163e-14
##
##             Sensitivity : 0.9025
##             Specificity : 0.8393
##          Pos Pred Value : 0.8792
##          Neg Pred Value : 0.8691
##              Prevalence : 0.5644
##          Detection Rate : 0.5093
##    Detection Prevalence : 0.5793
##       Balanced Accuracy : 0.8709
##
##        'Positive' Class : neutral or dissatisfied
##
```

Naive Bayes

```
# Naive Bayes model
library(e1071)
```

Naive bayes did worser than logisitc regression as we got an accuracy score of 0.85, comparted to 0.87. However, Naive bayes still did not perform bad on this dataset.

```
## Warning: package 'e1071' was built under R version 4.0.5
```

```
nb1 <- naiveBayes(satisfaction ~., data = train)
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
## neutral or dissatisfied               satisfied
##               0.5658755               0.4341245
##
## Conditional probabilities:
##                           Gender
## Y                            Female      Male
##   neutral or dissatisfied 0.5123465 0.4876535
##   satisfied               0.5008776 0.4991224
##
##                           customer_type
## Y                          disloyal Customer Loyal Customer
##   neutral or dissatisfied         0.2474570      0.7525430
##   satisfied                       0.1006167      0.8993833
##
##                           age
## Y                             [,1]     [,2]
##   neutral or dissatisfied 37.63585 16.42700
##   satisfied               41.74357 12.82347
##
##                           type_of_travel
## Y                          Business travel Personal Travel
##   neutral or dissatisfied      0.50950778      0.49049222
##   satisfied                    0.92656546      0.07343454
##
##                           customer_class
## Y                            Business        Eco   Eco Plus
##   neutral or dissatisfied 0.26044946 0.64278046 0.09677008
##   satisfied               0.76399431 0.19459203 0.04141366
##
##                           flight_distance
```

```
## Y                              [,1]      [,2]
##   neutral or dissatisfied  930.5045  791.1273
##   satisfied               1532.1378 1128.6311
##
##                          inflight_wifi_service
## Y                             [,1]      [,2]
##   neutral or dissatisfied 2.396998 0.9642376
##   satisfied               3.153985 1.5899266
##
##                          ease_of_online_booking
## Y                             [,1]     [,2]
##   neutral or dissatisfied 2.549959 1.212129
##   satisfied               3.024715 1.575855
##
##                          food_and_drink
## Y                             [,1]     [,2]
##   neutral or dissatisfied 2.960058 1.347340
##   satisfied               3.524336 1.235029
##
##                          online_boarding
## Y                             [,1]     [,2]
##   neutral or dissatisfied 2.662870 1.149971
##   satisfied               4.021371 1.198208
##
##                          seat_comfort
## Y                             [,1]     [,2]
##   neutral or dissatisfied 3.042289 1.303296
##   satisfied               3.962453 1.144109
##
##                          inflight_entertainment
## Y                             [,1]     [,2]
##   neutral or dissatisfied 2.892366 1.322524
##   satisfied               3.959132 1.082614
##
##                          onboard_service
## Y                             [,1]     [,2]
##   neutral or dissatisfied 3.018961 1.282100
##   satisfied               3.851471 1.130074
##
##                          leg_room_service
## Y                             [,1]     [,2]
##   neutral or dissatisfied 2.985297 1.304722
##   satisfied               3.815370 1.176746
##
##                          baggage_handling
## Y                             [,1]     [,2]
##   neutral or dissatisfied 3.373742 1.175804
##   satisfied               3.962192 1.104190
##
##                          checkin_service
## Y                             [,1]     [,2]
##   neutral or dissatisfied 3.042362 1.282945
##   satisfied               3.643359 1.159554
##
```

```
##                           inflight_service
## Y                              [,1]     [,2]
##   neutral or dissatisfied 3.386207 1.177601
##   satisfied                3.967244 1.095477
##
##                           cleanliness
## Y                              [,1]     [,2]
##   neutral or dissatisfied 2.934801 1.325047
##   satisfied                3.743904 1.145833
##
##                           departure_delay_in_minutes
## Y                              [,1]     [,2]
##   neutral or dissatisfied 16.39975 40.27640
##   satisfied                12.45247 35.23794
```

```
library(caret)
# Evaluating on the test data:
p1 <- predict(nb1, newdata = test, type = "class")
confusionMatrix(p1, test$satisfaction)
```

```
## Confusion Matrix and Statistics
##
##                           Reference
## Prediction                 neutral or dissatisfied satisfied
##   neutral or dissatisfied                   15844      2372
##   satisfied                                  2426     11730
##
##                Accuracy : 0.8518
##                  95% CI : (0.8479, 0.8556)
##     No Information Rate : 0.5644
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6987
##
##  Mcnemar's Test P-Value : 0.4442
##
##             Sensitivity : 0.8672
##             Specificity : 0.8318
##          Pos Pred Value : 0.8698
##          Neg Pred Value : 0.8286
##              Prevalence : 0.5644
##          Detection Rate : 0.4894
##    Detection Prevalence : 0.5627
##       Balanced Accuracy : 0.8495
##
##        'Positive' Class : neutral or dissatisfied
##
```

Decision Tree

```
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```
tree_airline <- tree(satisfaction ~., data = train)
summary(tree_airline)
```

```
##
## Classification tree:
## tree(formula = satisfaction ~ ., data = train)
## Variables actually used in tree construction:
## [1] "online_boarding"        "inflight_wifi_service"  "customer_class"
## [4] "inflight_entertainment" "customer_type"          "type_of_travel"
## Number of terminal nodes:  12
## Residual mean deviance:  0.49 = 47580 / 97100
## Misclassification error rate: 0.101 = 9808 / 97115
```

```
# Evaluating on test data
pred_tree <- predict(tree_airline, newdata = test, type = "class")
confusionMatrix(pred_tree, test$satisfaction)
```

Decision tree in classification performed much more efficiently than logistic regression and naive bayes, as seen from the high accuracy score of 0.90.

```
## Confusion Matrix and Statistics
##
##                          Reference
## Prediction                neutral or dissatisfied satisfied
##   neutral or dissatisfied                   17061      1960
##   satisfied                                  1209     12142
##
##                Accuracy : 0.9021
##                  95% CI : (0.8988, 0.9053)
##     No Information Rate : 0.5644
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7997
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9338
##             Specificity : 0.8610
##          Pos Pred Value : 0.8970
##          Neg Pred Value : 0.9094
##              Prevalence : 0.5644
##          Detection Rate : 0.5270
##    Detection Prevalence : 0.5876
##       Balanced Accuracy : 0.8974
```

```
##
##        'Positive' Class : neutral or dissatisfied
##
```

```
### Cross validation
##### We will prune the tree to 7 terminal nodes because we want to avoid overfitting by pruning it to
cv_tree <- cv.tree(tree_airline)
plot(cv_tree$size, cv_tree$dev, type='b')
```

Pruning the tree to eliminate overfitting


```
tree_pruned <- prune.tree(tree_airline, best=7)
pred_pruned <- predict(tree_pruned, newdata=test, type = "class")
confusionMatrix(pred_pruned, test$satisfaction)
```

In this case, the pruning did not improve results on test data because we got a higher accuracy score on the unpruned Tree.

```
## Confusion Matrix and Statistics
```

```
##
##                          Reference
## Prediction                neutral or dissatisfied satisfied
##   neutral or dissatisfied                   15849      1594
##   satisfied                                  2421     12508
##
##                Accuracy : 0.876
##                  95% CI : (0.8723, 0.8795)
##     No Information Rate : 0.5644
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7494
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8675
##             Specificity : 0.8870
##          Pos Pred Value : 0.9086
##          Neg Pred Value : 0.8378
##              Prevalence : 0.5644
##          Detection Rate : 0.4896
##    Detection Prevalence : 0.5388
##       Balanced Accuracy : 0.8772
##
##        'Positive' Class : neutral or dissatisfied
##
```

Results Analysis

Accuracy score for these algortihms:

Decision Tree: 0.90

Pruned Decision Tree: 0.876

Logistic regression: 0.875

Naive Bayes: 0.85

Summary:

Even in classification Decision tree gave us the most accurate results on the test data. Our decision tree was accurate 90% of the time, compared to 85% accuracy for Naive Bayes and 87.5% for logistic regression. Decision Tree worked well on this data because there are a lot of predictors in our data set, which is easy for decision trees to handle because they bisect the space into smaller and smaller regions. Unlike, Logistic regression, which divided the data into 2 classes and Naive Bayes, which calculated the likelihood of each event occurring. This data set also had non-linearity among predictors, which meant that a non-parametric algorithm like decision tree would perform better. We also pruned the Decision Tree and still got a higher accuracy score of 87.6% as compared to Naive Bayes and logistic regression. Furthermore, we can also use this script on new data to help airlines consider the factors that can satisfy a person and lead to a more comfortable trip.