Data_Analytics_Project

dhruv_raghav

```
knitr::opts_chunk$set(echo = TRUE)
```

Introduction

This document analyzes an Airbnb dataset with the goal of understanding user behavior and improving predictive modeling for destination booking. We will preprocess the data, visualize key characteristics, and apply several machine learning models.

Data Preprocessing & One-Hot Encoding

Question 1: How can we prepare the dataset for analysis?

```
remove(list=ls())

library(readr)
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
library(stringr)

set.seed(1)
df = read.csv("G:/semester_2/data_analytics/PROJECT/airbnb-recruiting-new-user-bookings/train_us
ers_2.csv/train_users_2.csv")
labels = df['country_destination']
head(df)
```

```
##          id date_account_created timestamp_first_active date_first_booking
## 1 gxn3p5htnn           2010-06-28           2.009032e+13
## 2 820tgsjxq7           2011-05-25           2.009052e+13
## 3 4ft3gnwmtx           2010-09-28           2.009061e+13         2010-08-02
## 4 bjjt8pjhuk           2011-12-05           2.009103e+13         2012-09-08
## 5 87mebub9p4           2010-09-14           2.009121e+13         2010-02-18
## 6 osr2jwljor           2010-01-01           2.010010e+13         2010-01-02
##      gender age signup_method signup_flow language affiliate_channel
## 1 -unknown-  NA      facebook           0       en            direct
## 2      MALE  38      facebook           0       en               seo
## 3    FEMALE  56         basic           3       en            direct
## 4    FEMALE  42      facebook           0       en            direct
## 5 -unknown-  41         basic           0       en            direct
## 6 -unknown-  NA         basic           0       en             other
##   affiliate_provider first_affiliate_tracked signup_app first_device_type
## 1             direct               untracked        Web       Mac Desktop
## 2             google               untracked        Web       Mac Desktop
## 3             direct               untracked        Web   Windows Desktop
## 4             direct               untracked        Web       Mac Desktop
## 5             direct               untracked        Web       Mac Desktop
## 6              other                     omg        Web       Mac Desktop
##   first_browser country_destination
## 1        Chrome                 NDF
## 2        Chrome                 NDF
## 3            IE                  US
## 4       Firefox               other
## 5        Chrome                  US
## 6        Chrome                  US
```

- **Objective:** Load the dataset and display the first few rows to understand its structure

```
df = df[-c(which(colnames(df) %in% c('date_first_booking')))]
```

- **Objective:** Remove columns that are not required for further analysis.

```
for(i in 1:ncol(df)){
  df[is.na(df[,i]), i] <- mean(df[,i], na.rm = TRUE)
}
```

```
## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA

## Warning in mean.default(df[, i], na.rm = TRUE): argument is not numeric or
## logical: returning NA
```

```
df$age <- as.integer(df$age)
```

- **Objective:** Fill missing values and ensure correct data types for analysis.

```
dac = as.data.frame(str_split_fixed(df$date_account_created, '-', 3))
df['dac_year'] = dac[,1]
df['dac_month'] = dac[,2]
df['dac_day'] = dac[,3]
df = df[,-c(which(colnames(df) %in% c('date_account_created')))]
```

- **Objective:** Extract year, month, and day from ‘date_account_created’ and remove the original column.

```
# Check if 'timestamp_first_active' exists before proceeding
if('timestamp_first_active' %in% colnames(df)) {
  df[,'tfa_year'] = substring(as.character(df[,'timestamp_first_active']), 1, 4)
  df['tfa_month'] = substring(as.character(df['timestamp_first_active']), 5, 6)
  df['tfa_day'] = substring(as.character(df['timestamp_first_active']), 7, 8)

  # Now you can safely remove 'timestamp_first_active'
  df = df[,-c(which(colnames(df) %in% c('timestamp_first_active')))]
} else {
  print("Column 'timestamp_first_active' does not exist in dataframe.")
}
```

- **Objective:** Check for the existence of 'timestamp_first_active', process it if present, and clean up the dataframe.

Rename One-Hot Encoded Columns for Clarity

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

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

```
ohe_feats = c('country_destination')
dummies <- dummyVars(~ country_destination, data = df)
df_all_ohe <- as.data.frame(predict(dummies, newdata = df))
df_combined <- cbind(df[,-c(which(colnames(df) %in% ohe_feats))],df_all_ohe)

colnames(df_combined)
```

```
##  [1] "id"                     "gender"
##  [3] "age"                    "signup_method"
##  [5] "signup_flow"            "language"
##  [7] "affiliate_channel"      "affiliate_provider"
##  [9] "first_affiliate_tracked" "signup_app"
## [11] "first_device_type"      "first_browser"
## [13] "dac_year"               "dac_month"
## [15] "dac_day"                "tfa_year"
## [17] "tfa_month"              "tfa_day"
## [19] "country_destinationAU"  "country_destinationCA"
## [21] "country_destinationDE"  "country_destinationES"
## [23] "country_destinationFR"  "country_destinationGB"
## [25] "country_destinationIT"  "country_destinationNDF"
## [27] "country_destinationNL"  "country_destinationother"
## [29] "country_destinationPT"  "country_destinationUS"
```

```
names(df_combined)[30]<-"US"
names(df_combined)[29]<-"PT"
names(df_combined)[28]<-"Other"
names(df_combined)[27]<-"NL"
names(df_combined)[26]<-"NDF"
names(df_combined)[25]<-"IT"
names(df_combined)[24]<-"GB"
names(df_combined)[23]<-"FR"
names(df_combined)[22]<-"ES"
names(df_combined)[21]<-"DE"
names(df_combined)[20]<-"CA"
names(df_combined)[19]<-"AU"
```

- **Objective:** Rename the columns generated from one-hot encoding to represent the respective country codes clearly. This aids in interpretability when using these variables in modeling and analysis.

Final Preparation of Dataset for Analysis

```
# Creating a final dataframe by excluding the original categorical column now represented by one
-hot encoding
df_final <- df_combined[-c(19:29)]
head(df_final)
```

```
##           id    gender age signup_method signup_flow language affiliate_channel
## 1 gxn3p5htnn -unknown-  49      facebook           0       en            direct
## 2 820tgsjxq7      MALE  38      facebook           0       en               seo
## 3 4ft3gnwmtx    FEMALE  56         basic           3       en            direct
## 4 bjjt8pjhuk    FEMALE  42      facebook           0       en            direct
## 5 87mebub9p4 -unknown-  41         basic           0       en            direct
## 6 osr2jwljor -unknown-  49         basic           0       en             other
##   affiliate_provider first_affiliate_tracked signup_app first_device_type
## 1             direct               untracked        Web       Mac Desktop
## 2             google               untracked        Web       Mac Desktop
## 3             direct               untracked        Web   Windows Desktop
## 4             direct               untracked        Web       Mac Desktop
## 5             direct               untracked        Web       Mac Desktop
## 6              other                     omg        Web       Mac Desktop
##   first_browser dac_year dac_month dac_day tfa_year tfa_month tfa_day US
## 1        Chrome     2010        06      28     2009        09      03  0
## 2        Chrome     2011        05      25     2009        09      03  0
## 3            IE     2010        09      28     2009        09      03  1
## 4       Firefox     2011        12      05     2009        09      03  0
## 5        Chrome     2010        09      14     2009        09      03  1
## 6        Chrome     2010        01      01     2010        09      03  1
```

- **Objective:** Remove redundant columns from the dataset now that we have the one-hot encoded columns, and display the first few rows of the cleaned dataframe to verify its readiness for the next steps in analysis and modeling.

Visualization

Question 2: What are the key characteristics of Airbnb users?

```
library(ggplot2)
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.3.3
```

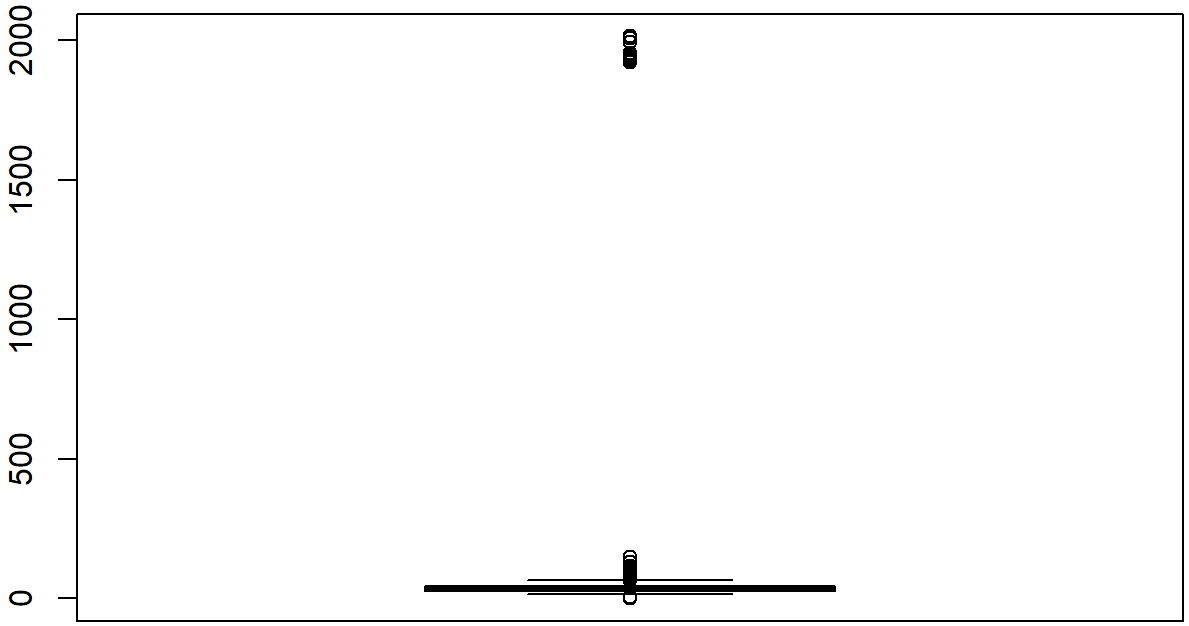
```
library(tidyverse)
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ purrr     1.0.2
## ✓ forcats   1.0.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::between()     masks data.table::between()
## ✖ dplyr::filter()      masks stats::filter()
## ✖ dplyr::first()       masks data.table::first()
## ✖ lubridate::hour()    masks data.table::hour()
## ✖ lubridate::isoweek() masks data.table::isoweek()
## ✖ dplyr::lag()         masks stats::lag()
## ✖ dplyr::last()        masks data.table::last()
## ✖ purrr::lift()        masks caret::lift()
## ✖ lubridate::mday()    masks data.table::mday()
## ✖ lubridate::minute()  masks data.table::minute()
## ✖ lubridate::month()   masks data.table::month()
## ✖ lubridate::quarter() masks data.table::quarter()
## ✖ lubridate::second()  masks data.table::second()
## ✖ purrr::transpose()   masks data.table::transpose()
## ✖ lubridate::wday()    masks data.table::wday()
## ✖ lubridate::week()    masks data.table::week()
## ✖ lubridate::yday()    masks data.table::yday()
## ✖ lubridate::year()    masks data.table::year()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
data <- fread("G:/semester_2/data_analytics/PROJECT/airbnb-recruiting-new-user-bookings/train_us
ers_2.csv/train_users_2.csv")
data <- as.data.frame(data)
boxplot(data$age)
```

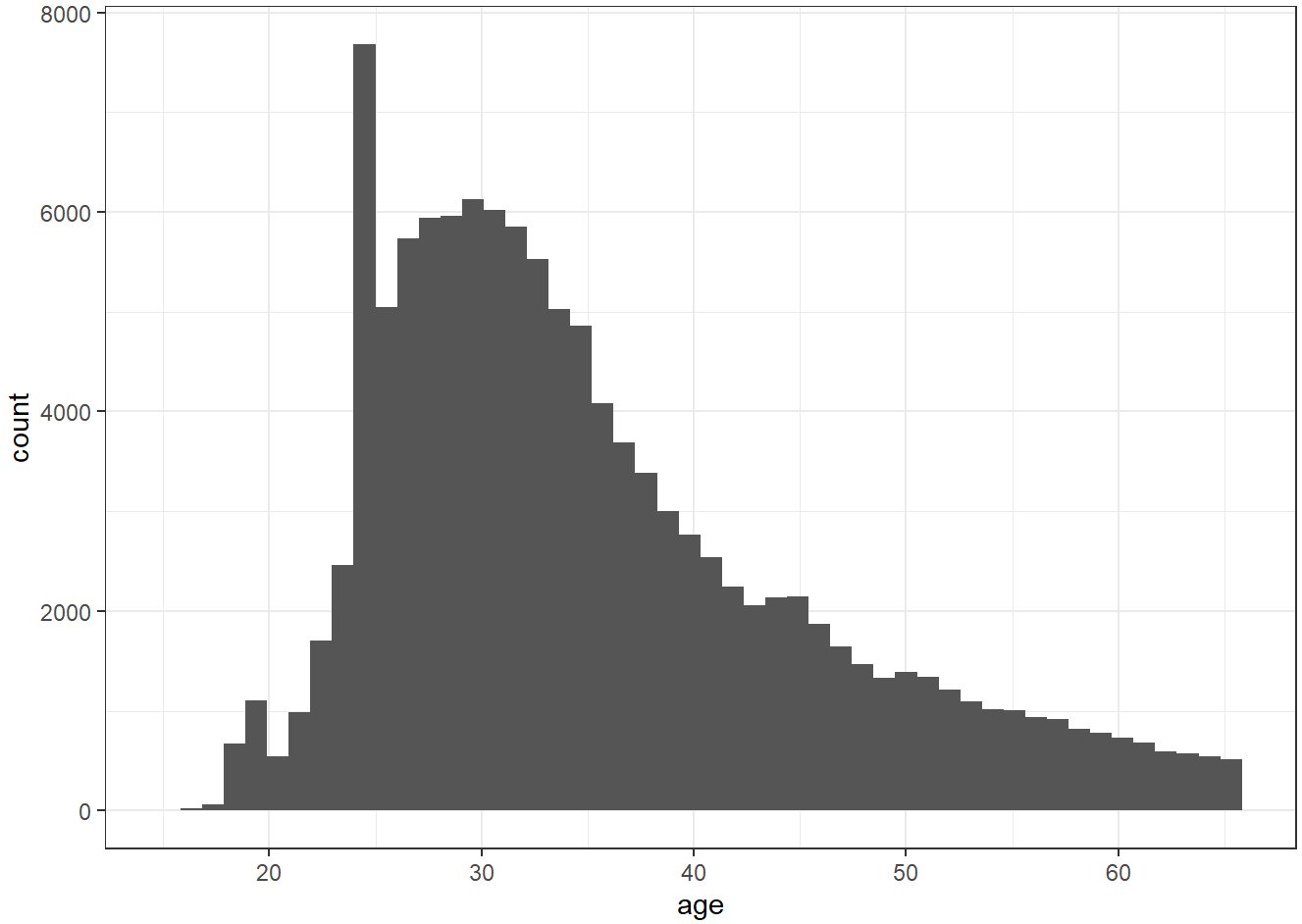



```
#outlier
remove_outliers <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x, na.rm = na.rm)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}
data$age <- remove_outliers(data$age)
```

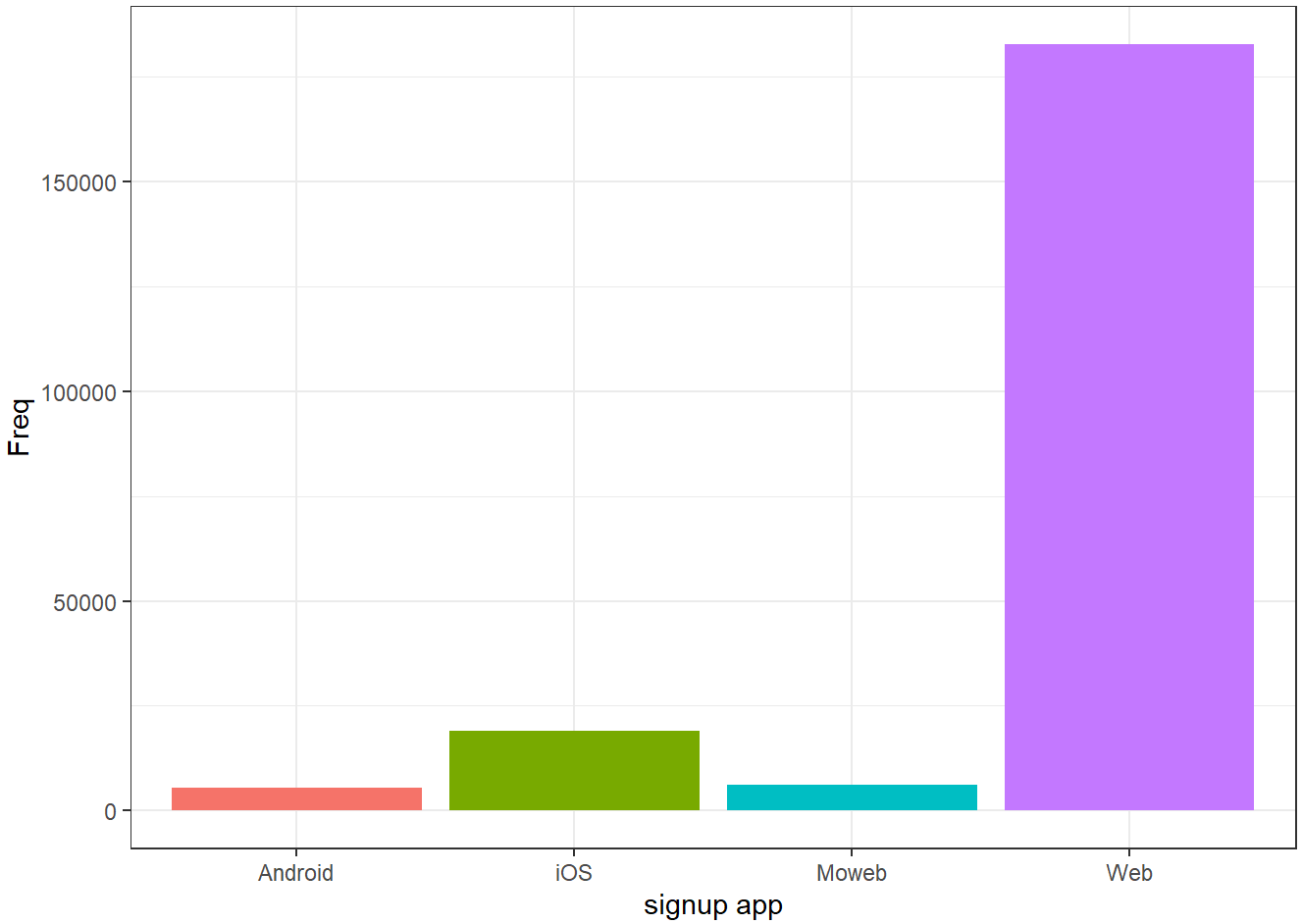
- **Objective:** Visualize the age distribution and manage outliers.

```
ggplot(data,aes(age))+
  geom_histogram(bins=50)+
  theme_bw()
```

```
## Warning: Removed 93584 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



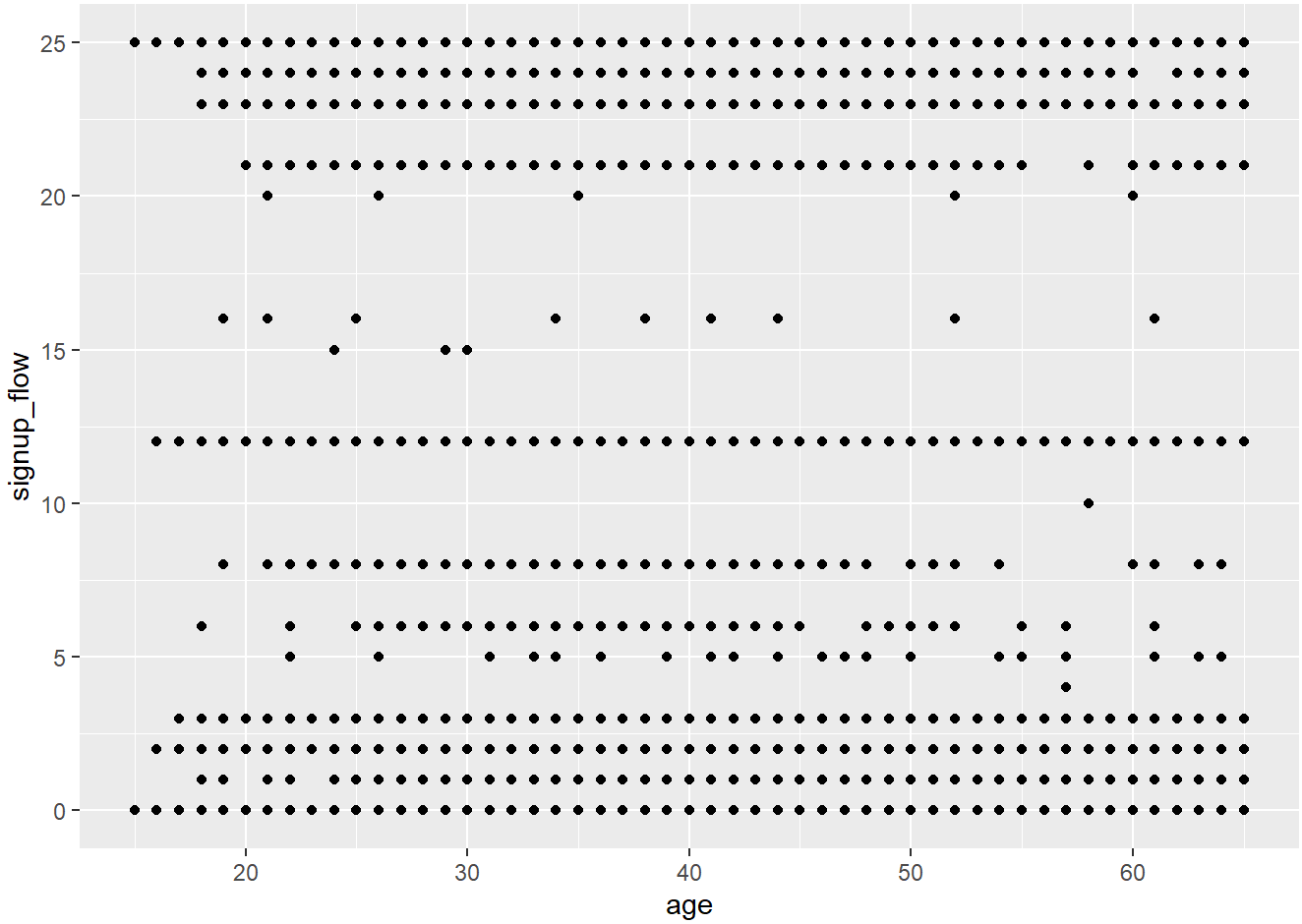

```
# app
signup_app <- table(data$signup_app) %>%
  as.data.frame()
ggplot(signup_app,aes(Var1,Freq,fill=Var1))+
  geom_bar(stat="identity")+
  theme_bw()+
  theme(legend.position = "none")+
  xlab("signup app")
```



- **Objective:** Display the usage distribution of different signup apps.

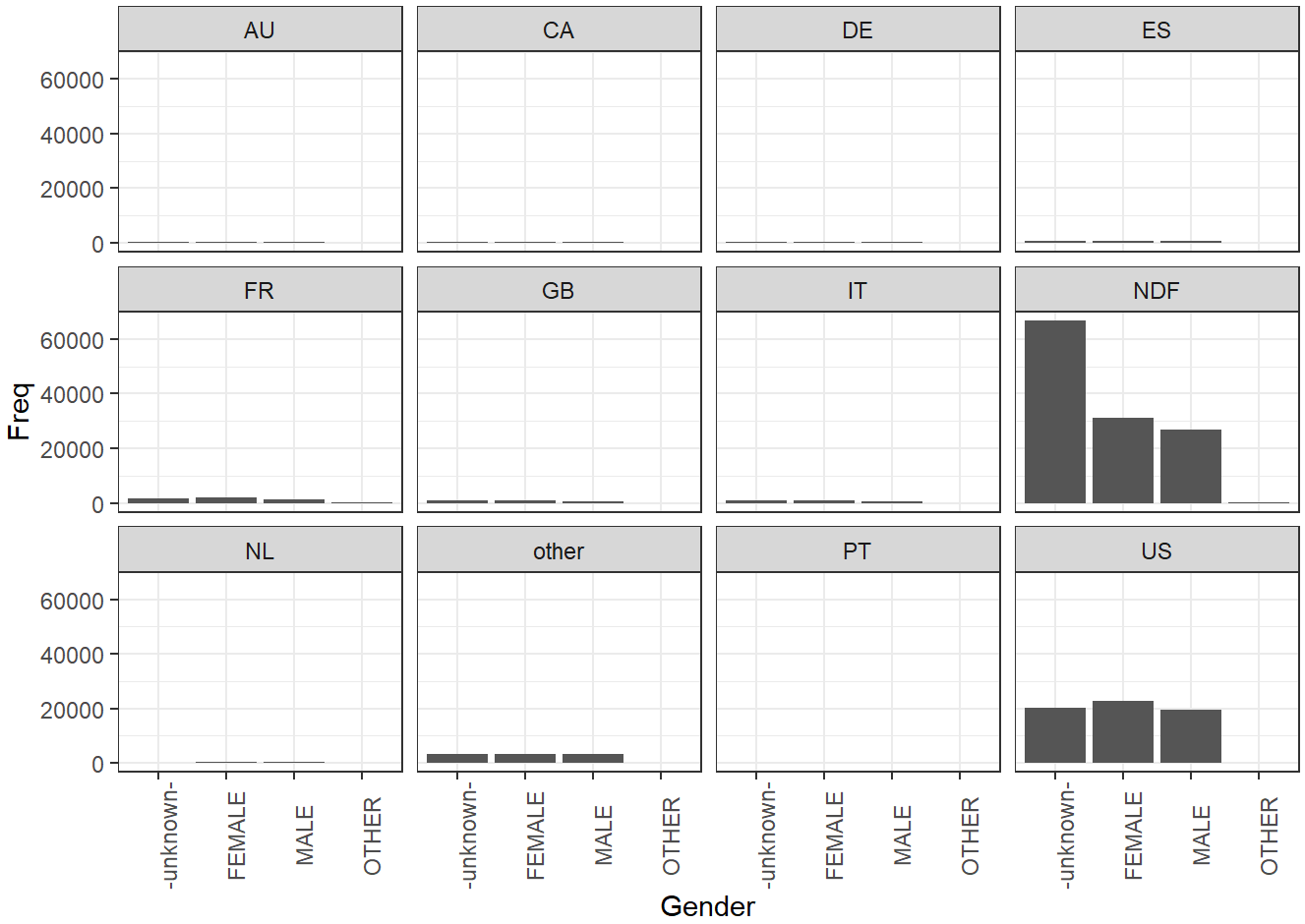
```
ggplot(data,aes(age,signup_flow))+
  geom_point()
```

```
## Warning: Removed 93584 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
gender <- table(data$country_destination,data$gender) %>%
  as.data.frame()

ggplot(gender,aes(Var2,Freq))+
  geom_bar(stat="identity")+
  facet_wrap(~Var1)+
  theme_bw()+
  theme(axis.text.x = element_text(angle=90))+
  xlab("Gender")
```



- **Objective:** Explore the relationship between gender and destination choices across different countries.

Machine Learning Models

Question 3: How accurately can we predict a user's destination booking within the US using machine learning

models?

K-Nearest Neighbors (KNN) Model

```
Airbnb <-(df_final)
                colClasses=c("id"="factor","gender"="factor","age"="factor",
                                "signup_method"="factor","signup_flow"="factor","language"="fact
or",
                                "affiliate_channel"="factor","affiliate_provider"="factor","firs
t_affiliate_tracked"="factor",
                                "signup_app"="factor","first_device_type"="factor","first_browse
r"="factor",
                                "dac_year"="factor","dac_month"="factor","dac_day"="factor",
                                "tfa_year"="factor","tfa_month"="factor","tfa_day"="factor","U
S"="factor")

index <-seq(1, nrow(Airbnb), by = 3)

test <- Airbnb[index,]
train <- Airbnb[-index,]
dim(Airbnb)
```

```
## [1] 213451     19
```

```
dim(test)
```

```
## [1] 71151    19
```

```
dim(train)
```

```
## [1] 142300     19
```

```
head(Airbnb)
```

```
##          id    gender age signup_method signup_flow language affiliate_channel
## 1 gxn3p5htnn -unknown-  49      facebook           0       en            direct
## 2 820tgsjxq7      MALE  38      facebook           0       en               seo
## 3 4ft3gnwmtx    FEMALE  56         basic           3       en            direct
## 4 bjjt8pjhuk    FEMALE  42      facebook           0       en            direct
## 5 87mebub9p4 -unknown-  41         basic           0       en            direct
## 6 osr2jwljor -unknown-  49         basic           0       en             other
##   affiliate_provider first_affiliate_tracked signup_app first_device_type
## 1             direct               untracked        Web       Mac Desktop
## 2             google               untracked        Web       Mac Desktop
## 3             direct               untracked        Web   Windows Desktop
## 4             direct               untracked        Web       Mac Desktop
## 5             direct               untracked        Web       Mac Desktop
## 6              other                     omg        Web       Mac Desktop
##   first_browser dac_year dac_month dac_day tfa_year tfa_month tfa_day US
## 1        Chrome     2010        06      28     2009        09      03  0
## 2        Chrome     2011        05      25     2009        09      03  0
## 3            IE     2010        09      28     2009        09      03  1
## 4       Firefox     2011        12      05     2009        09      03  0
## 5        Chrome     2010        09      14     2009        09      03  1
## 6        Chrome     2010        01      01     2010        09      03  1
```

```
head(train)
```

```
##          id    gender age signup_method signup_flow language affiliate_channel
## 2 820tgsjxq7      MALE  38      facebook           0       en               seo
## 3 4ft3gnwmtx    FEMALE  56         basic           3       en            direct
## 5 87mebub9p4 -unknown-  41         basic           0       en            direct
## 6 osr2jwljor -unknown-  49         basic           0       en             other
## 8 0d01nltbrs    FEMALE  47         basic           0       en            direct
## 9 a1vcnhxeij    FEMALE  50         basic           0       en             other
##   affiliate_provider first_affiliate_tracked signup_app first_device_type
## 2             google               untracked        Web       Mac Desktop
## 3             direct               untracked        Web   Windows Desktop
## 5             direct               untracked        Web       Mac Desktop
## 6              other                     omg        Web       Mac Desktop
## 8             direct                     omg        Web       Mac Desktop
## 9          craigslist               untracked        Web       Mac Desktop
##   first_browser dac_year dac_month dac_day tfa_year tfa_month tfa_day US
## 2        Chrome     2011        05      25     2009        09      03  0
## 3            IE     2010        09      28     2009        09      03  1
## 5        Chrome     2010        09      14     2009        09      03  1
## 6        Chrome     2010        01      01     2010        09      03  1
## 8        Safari     2010        01      03     2010        09      03  1
## 9        Safari     2010        01      04     2010        09      03  1
```

```
head(test)
```

```
##           id    gender age signup_method signup_flow language
## 1  gxn3p5htnn -unknown-  49      facebook           0       en
## 4  bjjt8pjhuk    FEMALE  42      facebook           0       en
## 7  lsw9q7uk0j    FEMALE  46         basic           0       en
## 10 6uh8zyj2gn -unknown-  46         basic           0       en
## 13 k6np330cm1 -unknown-  49         basic           0       en
## 16 v4d5rl22px    FEMALE  33         basic           0       en
##    affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1             direct             direct               untracked        Web
## 4             direct             direct               untracked        Web
## 7              other         craigslist               untracked        Web
## 10             other         craigslist                     omg        Web
## 13            direct             direct                                Web
## 16            direct             direct               untracked        Web
##    first_device_type first_browser dac_year dac_month dac_day tfa_year
## 1        Mac Desktop        Chrome     2010        06      28     2009
## 4        Mac Desktop       Firefox     2011        12      05     2009
## 7        Mac Desktop        Safari     2010        01      02     2010
## 10       Mac Desktop       Firefox     2010        01      04     2010
## 13     Other/Unknown     -unknown-     2010        01      05     2010
## 16   Windows Desktop        Chrome     2010        01      07     2010
##    tfa_month tfa_day US
## 1         09      03  0
## 4         09      03  0
## 7         09      03  1
## 10        09      03  1
## 13        09      03  0
## 16        09      03  0
```

```
str(train)
```

```
## 'data.frame':    142300 obs. of  19 variables:
##  $ id                     : chr  "820tgsjxq7" "4ft3gnwmtx" "87mebub9p4" "osr2jwljor" ...
##  $ gender                 : chr  "MALE" "FEMALE" "-unknown-" "-unknown-" ...
##  $ age                    : int  38 56 41 49 47 50 36 47 37 36 ...
##  $ signup_method          : chr  "facebook" "basic" "basic" "basic" ...
##  $ signup_flow            : int  0 3 0 0 0 0 0 0 0 0 ...
##  $ language               : chr  "en" "en" "en" "en" ...
##  $ affiliate_channel      : chr  "seo" "direct" "direct" "other" ...
##  $ affiliate_provider     : chr  "google" "direct" "direct" "other" ...
##  $ first_affiliate_tracked: chr  "untracked" "untracked" "untracked" "omg" ...
##  $ signup_app             : chr  "Web" "Web" "Web" "Web" ...
##  $ first_device_type      : chr  "Mac Desktop" "Windows Desktop" "Mac Desktop" "Mac Desktop"
...
##  $ first_browser          : chr  "Chrome" "IE" "Chrome" "Chrome" ...
##  $ dac_year               : chr  "2011" "2010" "2010" "2010" ...
##  $ dac_month              : chr  "05" "09" "09" "01" ...
##  $ dac_day                : chr  "25" "28" "14" "01" ...
##  $ tfa_year               : chr  "2009" "2009" "2009" "2010" ...
##  $ tfa_month              : chr  "09" "09" "09" "09" ...
##  $ tfa_day                : chr  "03" "03" "03" "03" ...
##  $ US                     : num  0 1 1 1 1 1 1 0 0 0 ...
```

```
# After setting preference, you can run your kknn model
library(kknn)
```

```
## Warning: package 'kknn' was built under R version 4.3.3
```

```
##
## Attaching package: 'kknn'
```

```
## The following object is masked from 'package:caret':
##
##     contr.dummy
```

```
predict <- kknn(factor(US)~gender+age+signup_method+signup_flow+affiliate_channel+first_affiliat
e_tracked+signup_app+first_device_type+dac_year+dac_month+dac_day+tfa_year, train, test, kernel
="rectangular", k=10)
fit <- fitted(predict)
table(kknn=fit, test$US)
```

```
##
## kknn     0     1
##    0 44523 16026
##    1  5993  4609
```

```
knn_error_rate = sum(fit != test$US) / length(test$US)
print(knn_error_rate)
```

```
## [1] 0.3094686
```

```
accuracy <- (1 - knn_error_rate) * 100
accuracy
```

```
## [1] 69.05314
```

- **Objective and Result:** Implement a KNN model to predict whether a booking destination is the US and evaluate its accuracy.

Random Forest Model

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

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

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
Airbnb <- df_final
set.seed(123)
head(Airbnb)
```

```
##          id    gender age signup_method signup_flow language affiliate_channel
## 1 gxn3p5htnn -unknown-  49      facebook           0       en            direct
## 2 820tgsjxq7      MALE  38      facebook           0       en               seo
## 3 4ft3gnwmtx    FEMALE  56         basic           3       en            direct
## 4 bjjt8pjhuk    FEMALE  42      facebook           0       en            direct
## 5 87mebub9p4 -unknown-  41         basic           0       en            direct
## 6 osr2jwljor -unknown-  49         basic           0       en             other
##   affiliate_provider first_affiliate_tracked signup_app first_device_type
## 1             direct               untracked        Web       Mac Desktop
## 2             google               untracked        Web       Mac Desktop
## 3             direct               untracked        Web   Windows Desktop
## 4             direct               untracked        Web       Mac Desktop
## 5             direct               untracked        Web       Mac Desktop
## 6              other                     omg        Web       Mac Desktop
##   first_browser dac_year dac_month dac_day tfa_year tfa_month tfa_day US
## 1        Chrome     2010        06      28     2009        09      03  0
## 2        Chrome     2011        05      25     2009        09      03  0
## 3            IE     2010        09      28     2009        09      03  1
## 4       Firefox     2011        12      05     2009        09      03  0
## 5        Chrome     2010        09      14     2009        09      03  1
## 6        Chrome     2010        01      01     2010        09      03  1
```

```
#dsn2<-na.omit(dsn)
#set.seed(123)

index <-seq(1, nrow(Airbnb), by = 3)
Airbnb <- Airbnb[index,]

index <-seq(1, nrow(Airbnb), by = 5)
test <- Airbnb[index,]
train <- Airbnb[-index,]
train<- train[c(2,3,13,16,19)]
test<-test[c(2,3,13,16,19)]

#train$Reverse <- as.character(train$Reverse)
#train$US <- as.factor(train$US)

x<-randomForest(factor(US)~.,data=test, importance=TRUE, ntree=1000)
importance(x)
```

```
##                 0         1 MeanDecreaseAccuracy MeanDecreaseGini
## gender   107.84299 -100.41476             96.90977        135.76610
## age      108.97213  -22.57780            110.66594        507.83063
## dac_year  32.75144  -31.40911             31.21961         44.03119
## tfa_year  28.99789  -29.31313             24.28155         46.70048
```

```
varImpPlot(x)
```

x


```
Prediction<-predict(x,test)
table(actual=test[,5],Prediction)
```

```
##       Prediction
## actual    0    1
##      0 9889  300
##      1 3545  497
```

```
wrong<-(test[,5]!=Prediction)
error_rate<-sum(wrong)/length(wrong)
error_rate
```

```
## [1] 0.2701848
```

```
accuracy <- (1-error_rate)*100
accuracy
```

```
## [1] 72.98152
```

- **Objective and Result:** Train a Random Forest model, assess feature importance, predict the 'US' booking, and calculate model accuracy.

```
Airbnb <-(df_final)
                colClasses=c("id"="factor","gender"="factor","age"="factor",
                                "signup_method"="factor","signup_flow"="factor","language"="fact
or",
                                "affiliate_channel"="factor","affiliate_provider"="factor","firs
t_affiliate_tracked"="factor",
                                "signup_app"="factor","first_device_type"="factor","first_browse
r"="factor",
                                "dac_year"="factor","dac_month"="factor","dac_day"="factor",
                                "tfa_year"="factor","tfa_month"="factor","tfa_day"="factor","U
S"="factor")
```

```
index <-seq(1, nrow(Airbnb), by = 3)

test <- Airbnb[index,]
train <- Airbnb[-index,]
dim(Airbnb)
```

```
## [1] 213451     19
```

```
dim(test)
```

```
## [1] 71151    19
```

```
dim(train)
```

```
## [1] 142300     19
```

```
head(Airbnb)
```

```
##           id    gender age signup_method signup_flow language affiliate_channel
## 1 gxn3p5htnn -unknown-  49      facebook           0       en            direct
## 2 820tgsjxq7      MALE  38      facebook           0       en               seo
## 3 4ft3gnwmtx    FEMALE  56         basic           3       en            direct
## 4 bjjt8pjhuk    FEMALE  42      facebook           0       en            direct
## 5 87mebub9p4 -unknown-  41         basic           0       en            direct
## 6 osr2jwljor -unknown-  49         basic           0       en             other
##   affiliate_provider first_affiliate_tracked signup_app first_device_type
## 1             direct               untracked        Web       Mac Desktop
## 2             google               untracked        Web       Mac Desktop
## 3             direct               untracked        Web   Windows Desktop
## 4             direct               untracked        Web       Mac Desktop
## 5             direct               untracked        Web       Mac Desktop
## 6              other                     omg        Web       Mac Desktop
##   first_browser dac_year dac_month dac_day tfa_year tfa_month tfa_day US
## 1        Chrome     2010        06      28     2009        09      03  0
## 2        Chrome     2011        05      25     2009        09      03  0
## 3            IE     2010        09      28     2009        09      03  1
## 4       Firefox     2011        12      05     2009        09      03  0
## 5        Chrome     2010        09      14     2009        09      03  1
## 6        Chrome     2010        01      01     2010        09      03  1
```

```
head(train)
```

```
##           id    gender age signup_method signup_flow language affiliate_channel
## 2 820tgsjxq7      MALE  38      facebook           0       en               seo
## 3 4ft3gnwmtx    FEMALE  56         basic           3       en            direct
## 5 87mebub9p4 -unknown-  41         basic           0       en            direct
## 6 osr2jwljor -unknown-  49         basic           0       en             other
## 8 0d01nltbrs    FEMALE  47         basic           0       en            direct
## 9 a1vcnhxeij    FEMALE  50         basic           0       en             other
##   affiliate_provider first_affiliate_tracked signup_app first_device_type
## 2             google               untracked        Web       Mac Desktop
## 3             direct               untracked        Web   Windows Desktop
## 5             direct               untracked        Web       Mac Desktop
## 6              other                     omg        Web       Mac Desktop
## 8             direct                     omg        Web       Mac Desktop
## 9          craigslist               untracked        Web       Mac Desktop
##   first_browser dac_year dac_month dac_day tfa_year tfa_month tfa_day US
## 2        Chrome     2011        05      25     2009        09      03  0
## 3            IE     2010        09      28     2009        09      03  1
## 5        Chrome     2010        09      14     2009        09      03  1
## 6        Chrome     2010        01      01     2010        09      03  1
## 8        Safari     2010        01      03     2010        09      03  1
## 9        Safari     2010        01      04     2010        09      03  1
```

```
head(test)
```

```
##           id    gender age signup_method signup_flow language
## 1  gxn3p5htnn -unknown-  49      facebook           0       en
## 4  bjjt8pjhuk    FEMALE  42      facebook           0       en
## 7  lsw9q7uk0j    FEMALE  46         basic           0       en
## 10 6uh8zyj2gn -unknown-  46         basic           0       en
## 13 k6np330cm1 -unknown-  49         basic           0       en
## 16 v4d5rl22px    FEMALE  33         basic           0       en
##    affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1             direct             direct               untracked        Web
## 4             direct             direct               untracked        Web
## 7              other         craigslist               untracked        Web
## 10             other         craigslist                     omg        Web
## 13            direct             direct                                Web
## 16            direct             direct               untracked        Web
##    first_device_type first_browser dac_year dac_month dac_day tfa_year
## 1        Mac Desktop        Chrome     2010        06      28     2009
## 4        Mac Desktop       Firefox     2011        12      05     2009
## 7        Mac Desktop        Safari     2010        01      02     2010
## 10       Mac Desktop       Firefox     2010        01      04     2010
## 13     Other/Unknown     -unknown-     2010        01      05     2010
## 16   Windows Desktop        Chrome     2010        01      07     2010
##    tfa_month tfa_day US
## 1         09      03  0
## 4         09      03  0
## 7         09      03  1
## 10        09      03  1
## 13        09      03  0
## 16        09      03  0
```

```
library(naivebayes)
```

```
## Warning: package 'naivebayes' was built under R version 4.3.3
```

```
## naivebayes 1.0.0 loaded
```

```
## For more information please visit:
```

```
## https://majkamichal.github.io/naivebayes/
```

```
##
## Attaching package: 'naivebayes'
```

```
## The following object is masked from 'package:data.table':
##
##     tables
```

```
# Split the data into training and testing sets
set.seed(123)
index <- seq(1, nrow(Airbnb), by = 3)
test <- Airbnb[index,]
train <- Airbnb[-index,]
head(train)
```

```
##           id    gender age signup_method signup_flow language affiliate_channel
## 2 820tgsjxq7      MALE  38      facebook           0       en               seo
## 3 4ft3gnwmtx    FEMALE  56         basic           3       en            direct
## 5 87mebub9p4 -unknown-  41         basic           0       en            direct
## 6 osr2jwljor -unknown-  49         basic           0       en             other
## 8 0d01nltbrs    FEMALE  47         basic           0       en            direct
## 9 a1vcnhxeij    FEMALE  50         basic           0       en             other
##   affiliate_provider first_affiliate_tracked signup_app first_device_type
## 2             google               untracked        Web       Mac Desktop
## 3             direct               untracked        Web   Windows Desktop
## 5             direct               untracked        Web       Mac Desktop
## 6              other                     omg        Web       Mac Desktop
## 8             direct                     omg        Web       Mac Desktop
## 9          craigslist               untracked        Web       Mac Desktop
##   first_browser dac_year dac_month dac_day tfa_year tfa_month tfa_day US
## 2        Chrome     2011        05      25     2009        09      03  0
## 3            IE     2010        09      28     2009        09      03  1
## 5        Chrome     2010        09      14     2009        09      03  1
## 6        Chrome     2010        01      01     2010        09      03  1
## 8        Safari     2010        01      03     2010        09      03  1
## 9        Safari     2010        01      04     2010        09      03  1
```

```
head(test)
```

```
##           id    gender age signup_method signup_flow language
## 1  gxn3p5htnn -unknown-  49      facebook           0       en
## 4  bjjt8pjhuk    FEMALE  42      facebook           0       en
## 7  lsw9q7uk0j    FEMALE  46         basic           0       en
## 10 6uh8zyj2gn -unknown-  46         basic           0       en
## 13 k6np330cm1 -unknown-  49         basic           0       en
## 16 v4d5rl22px    FEMALE  33         basic           0       en
##    affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1             direct             direct               untracked        Web
## 4             direct             direct               untracked        Web
## 7              other         craigslist               untracked        Web
## 10             other         craigslist                     omg        Web
## 13            direct             direct                                Web
## 16            direct             direct               untracked        Web
##    first_device_type first_browser dac_year dac_month dac_day tfa_year
## 1        Mac Desktop        Chrome     2010        06      28     2009
## 4        Mac Desktop       Firefox     2011        12      05     2009
## 7        Mac Desktop        Safari     2010        01      02     2010
## 10       Mac Desktop       Firefox     2010        01      04     2010
## 13     Other/Unknown     -unknown-     2010        01      05     2010
## 16   Windows Desktop        Chrome     2010        01      07     2010
##    tfa_month tfa_day US
## 1         09      03  0
## 4         09      03  0
## 7         09      03  1
## 10        09      03  1
## 13        09      03  0
## 16        09      03  0
```

```
# Fit Naive Bayes model
nb_model <- naive_bayes(factor(US)~age+gender+signup_method+language+dac_year+affiliate_channel+
first_device_type+signup_flow+signup_app+affiliate_provider, data = train)
```

```
## Warning: naive_bayes(): Feature language - zero probabilities are present.
## Consider Laplace smoothing.
```

```
## Warning: naive_bayes(): Feature affiliate_provider - zero probabilities are
## present. Consider Laplace smoothing.
```

```
# Make predictions on the test set
predictions <- predict(nb_model, test)
```

```
## Warning: predict.naive_bayes(): more features in the newdata are provided as
## there are probability tables in the object. Calculation is performed based on
## features to be found in the tables.
```

```
# Evaluate the performance of the model
conf_matrix <- table(predictions, test$US)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
#print(conf_matrix)
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.7047266
```

```
accuracy_percentage <- accuracy*100
accuracy_percentage
```

```
## [1] 70.47266
```