expediaEDA2

Anandan Balaji

13 May 2016

## Expedia Hotel Recommendation - Data Anyalysis

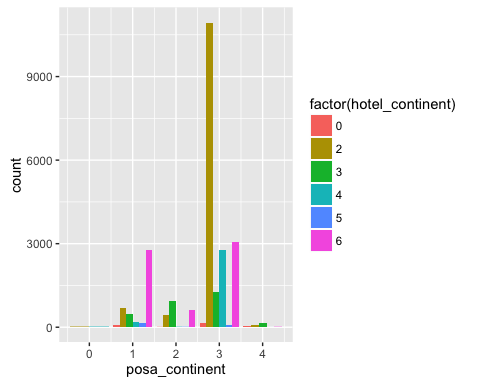
This document shows the charts related to different fields in the training data set.

# read 25 K  
train\_df <- read.csv(file = "train.25000.csv")  
  
library(ggplot2)

## posa\_continent (user location) and hotel\_continent (desitnation location)

The below chart shows that the people from continent #3 travel a lot and prefer visiting continent #2.

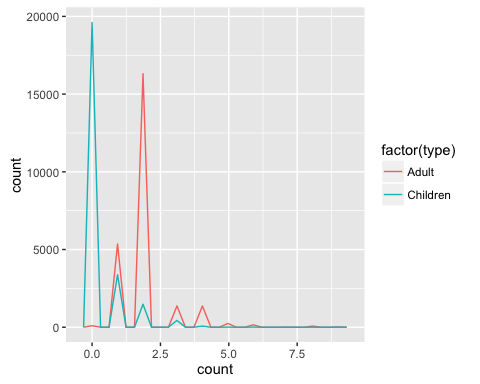
ggplot(train\_df, aes(posa\_continent, fill = factor(hotel\_continent))) + geom\_bar(position = "dodge")



## srch\_adults\_cnt, srch\_children\_cnt

keys <- c("date\_time", "user\_id")  
  
adult\_df <- train\_df[keys]  
adult\_df$count <- train\_df$srch\_adults\_cnt  
adult\_df$type <- "Adult"  
  
child\_df <- train\_df[keys]  
child\_df$count <- train\_df$srch\_children\_cnt  
child\_df$type <- "Children"  
  
adult\_child\_df <- rbind(adult\_df, child\_df)  
  
#ggplot(adult\_child\_df, aes(x = user\_id, y = count, col = factor(type))) +  
# geom\_line()  
  
# how to change the scale in x-axis to 0, 1, 2, 3 etc.?  
ggplot(adult\_child\_df, aes(count, col = factor(type))) + geom\_freqpoly()

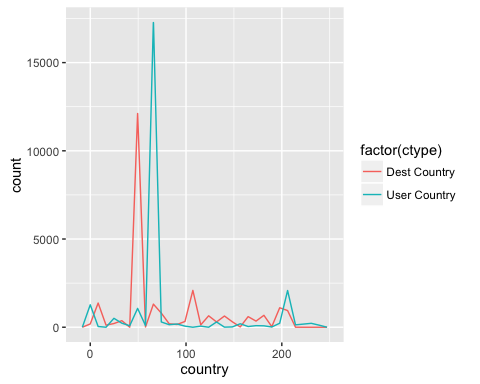
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## source and destination country

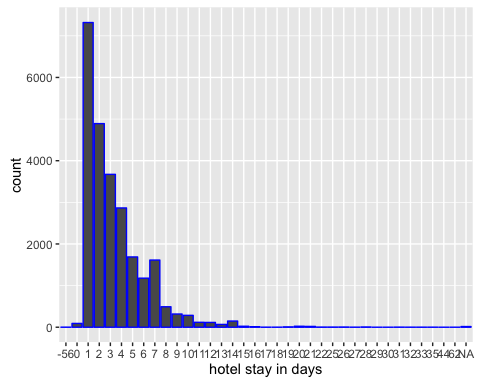
src\_df <- train\_df[keys]  
src\_df$country <- train\_df$user\_location\_country  
src\_df$ctype <- "User Country"  
  
dest\_df <- train\_df[keys]  
dest\_df$country <- train\_df$hotel\_country  
dest\_df$ctype <- "Dest Country"  
  
country\_df <- rbind(src\_df, dest\_df)  
  
ggplot(country\_df, aes(country, col = factor(ctype))) +   
 geom\_freqpoly()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



## number of days stay : srch\_co - srch\_ci (checkout - checkin date)

ggplot(train\_df,   
 aes(factor(as.Date(as.character(train\_df$srch\_co), format = "%Y-%m-%d") - as.Date(as.character(train\_df$srch\_ci), format = "%Y-%m-%d")))) +  
 geom\_bar(col = "blue") +  
 xlab("hotel stay in days")



## hotel\_continent & hotel\_cluster

There are more number of hotel\_clusters in the continent #2 and #6.

ggplot(train\_df,   
 aes(x = hotel\_cluster, y = hotel\_continent, col = factor(hotel\_continent))) + # to avoid over plotting   
 geom\_jitter(alpha = 0.7) +  
 geom\_smooth(method = "lm", se = F)

