expediaEDA\_ext

Anandan Balaji

18 May 2016

## Expedia Hotel Recommendation - Data Anyalysis

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

# read 25 K  
#system.time(train\_df <- read.csv(file = "train.25000.csv"))  
  
# read complete data  
library(data.table)

## Warning: package 'data.table' was built under R version 3.2.3

system.time(train\_df <- fread("train.csv"))

##   
Read 0.0% of 37670293 rows  
Read 1.0% of 37670293 rows  
Read 1.5% of 37670293 rows  
Read 2.1% of 37670293 rows  
Read 2.8% of 37670293 rows  
Read 3.7% of 37670293 rows  
Read 4.8% of 37670293 rows  
Read 6.2% of 37670293 rows  
Read 7.8% of 37670293 rows  
Read 8.8% of 37670293 rows  
Read 9.7% of 37670293 rows  
Read 11.6% of 37670293 rows  
Read 12.1% of 37670293 rows  
Read 13.9% of 37670293 rows  
Read 15.1% of 37670293 rows  
Read 16.8% of 37670293 rows  
Read 18.2% of 37670293 rows  
Read 18.7% of 37670293 rows  
Read 20.6% of 37670293 rows  
Read 22.4% of 37670293 rows  
Read 23.2% of 37670293 rows  
Read 25.0% of 37670293 rows  
Read 26.7% of 37670293 rows  
Read 28.5% of 37670293 rows  
Read 28.9% of 37670293 rows  
Read 30.5% of 37670293 rows  
Read 32.2% of 37670293 rows  
Read 33.8% of 37670293 rows  
Read 35.4% of 37670293 rows  
Read 36.0% of 37670293 rows  
Read 37.6% of 37670293 rows  
Read 39.2% of 37670293 rows  
Read 41.0% of 37670293 rows  
Read 42.9% of 37670293 rows  
Read 44.7% of 37670293 rows  
Read 45.3% of 37670293 rows  
Read 47.1% of 37670293 rows  
Read 48.9% of 37670293 rows  
Read 50.7% of 37670293 rows  
Read 52.5% of 37670293 rows  
Read 54.2% of 37670293 rows  
Read 56.0% of 37670293 rows  
Read 57.3% of 37670293 rows  
Read 59.1% of 37670293 rows  
Read 60.8% of 37670293 rows  
Read 62.5% of 37670293 rows  
Read 64.2% of 37670293 rows  
Read 65.9% of 37670293 rows  
Read 67.6% of 37670293 rows  
Read 69.3% of 37670293 rows  
Read 71.0% of 37670293 rows  
Read 72.6% of 37670293 rows  
Read 73.5% of 37670293 rows  
Read 75.2% of 37670293 rows  
Read 76.8% of 37670293 rows  
Read 78.4% of 37670293 rows  
Read 80.1% of 37670293 rows  
Read 81.8% of 37670293 rows  
Read 83.4% of 37670293 rows  
Read 85.0% of 37670293 rows  
Read 86.6% of 37670293 rows  
Read 88.3% of 37670293 rows  
Read 89.9% of 37670293 rows  
Read 91.5% of 37670293 rows  
Read 93.1% of 37670293 rows  
Read 94.1% of 37670293 rows  
Read 95.8% of 37670293 rows  
Read 97.7% of 37670293 rows  
Read 99.3% of 37670293 rows  
Read 37670293 rows and 24 (of 24) columns from 3.791 GB file in 00:03:00

## user system elapsed   
## 169.029 10.547 197.427

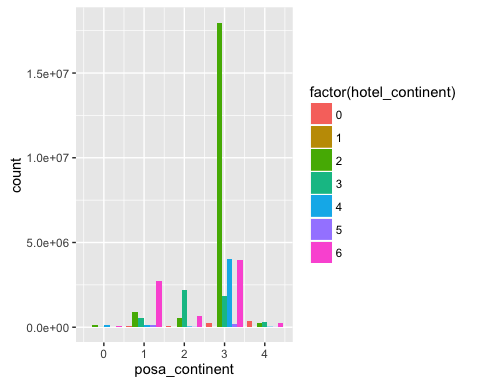
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.2.4

## 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")  
setkeyv(train\_df, keys)  
  
#adult\_df <- train\_df[keys]  
adult\_df <- train\_df[keys]  
adult\_df$count <- train\_df$srch\_adults\_cnt

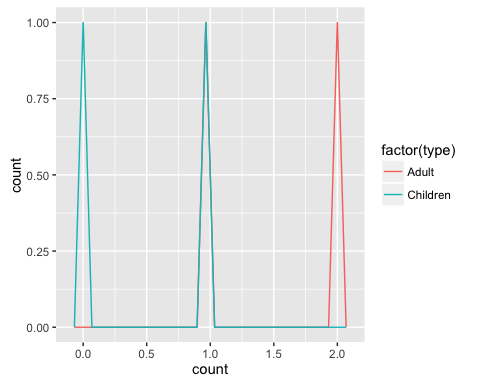
## Warning in `[<-.data.table`(x, j = name, value = value): Supplied 37670293  
## items to be assigned to 2 items of column 'count' (37670291 unused)

adult\_df$type <- "Adult"  
  
child\_df <- train\_df[keys]  
child\_df$count <- train\_df$srch\_children\_cnt

## Warning in `[<-.data.table`(x, j = name, value = value): Supplied 37670293  
## items to be assigned to 2 items of column 'count' (37670291 unused)

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

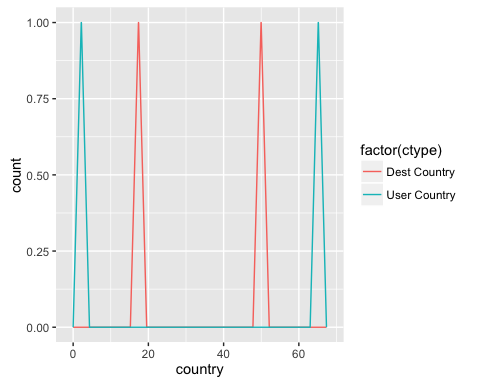
## Warning in `[<-.data.table`(x, j = name, value = value): Supplied 37670293  
## items to be assigned to 2 items of column 'country' (37670291 unused)

src\_df$ctype <- "User Country"  
  
dest\_df <- train\_df[keys]  
dest\_df$country <- train\_df$hotel\_country

## Warning in `[<-.data.table`(x, j = name, value = value): Supplied 37670293  
## items to be assigned to 2 items of column 'country' (37670291 unused)

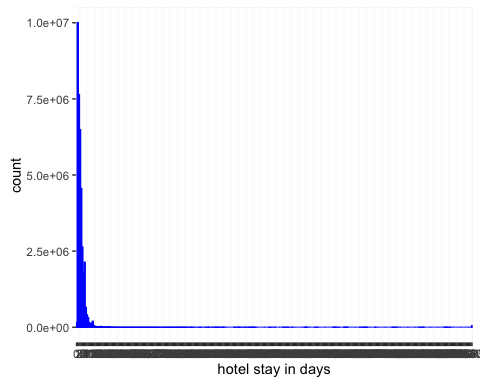
dest\_df$ctype <- "Dest Country"  
  
country\_df <- rbind(src\_df, dest\_df)  
  
ggplot(country\_df, aes(country, col = factor(ctype))) +   
 geom\_freqpoly()

## `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(abs(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

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)

