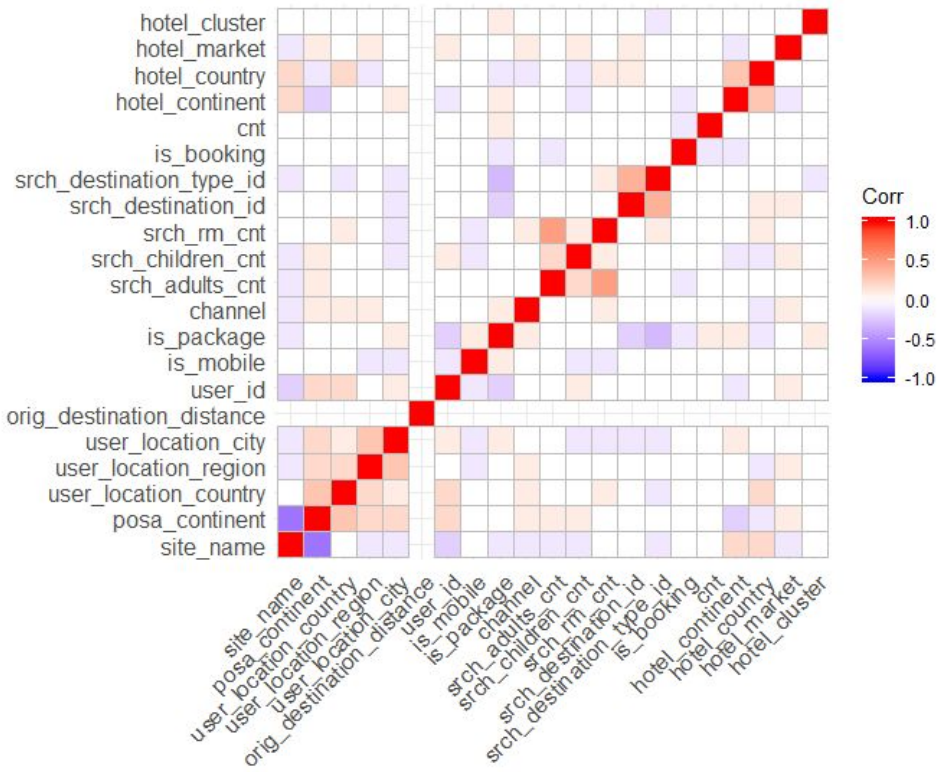
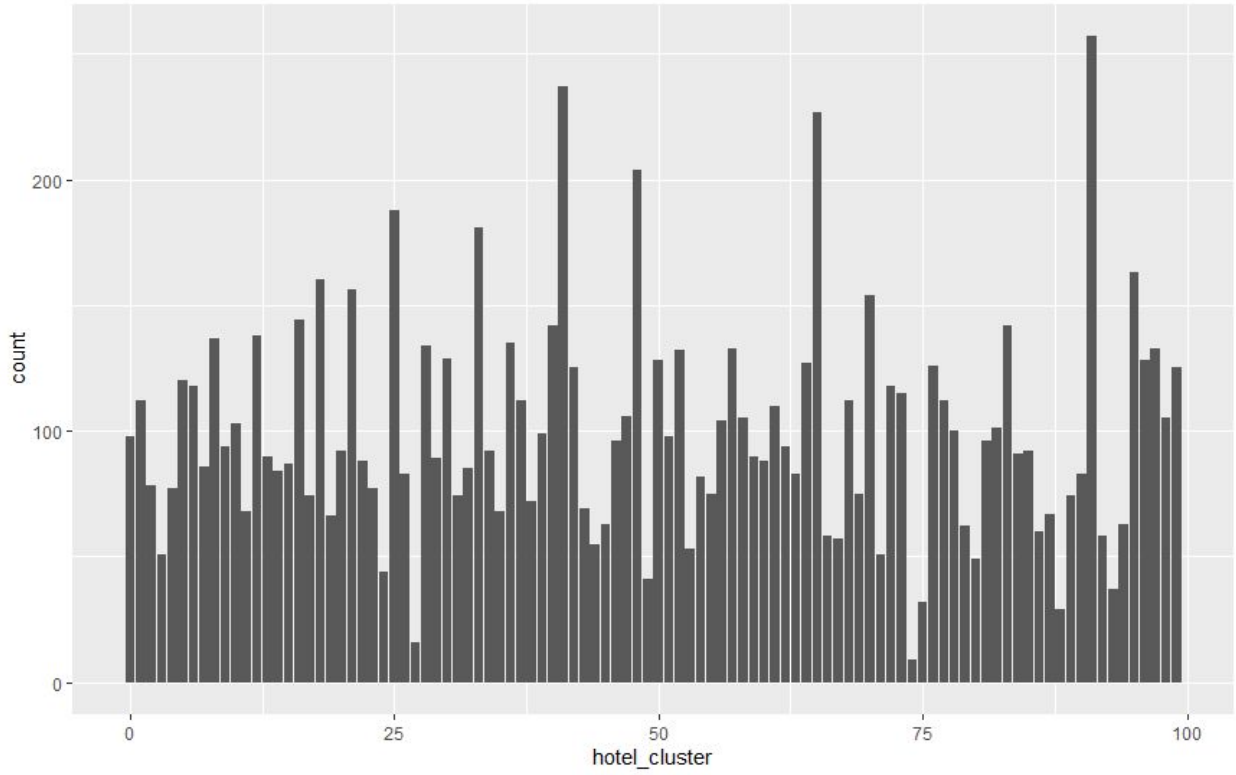
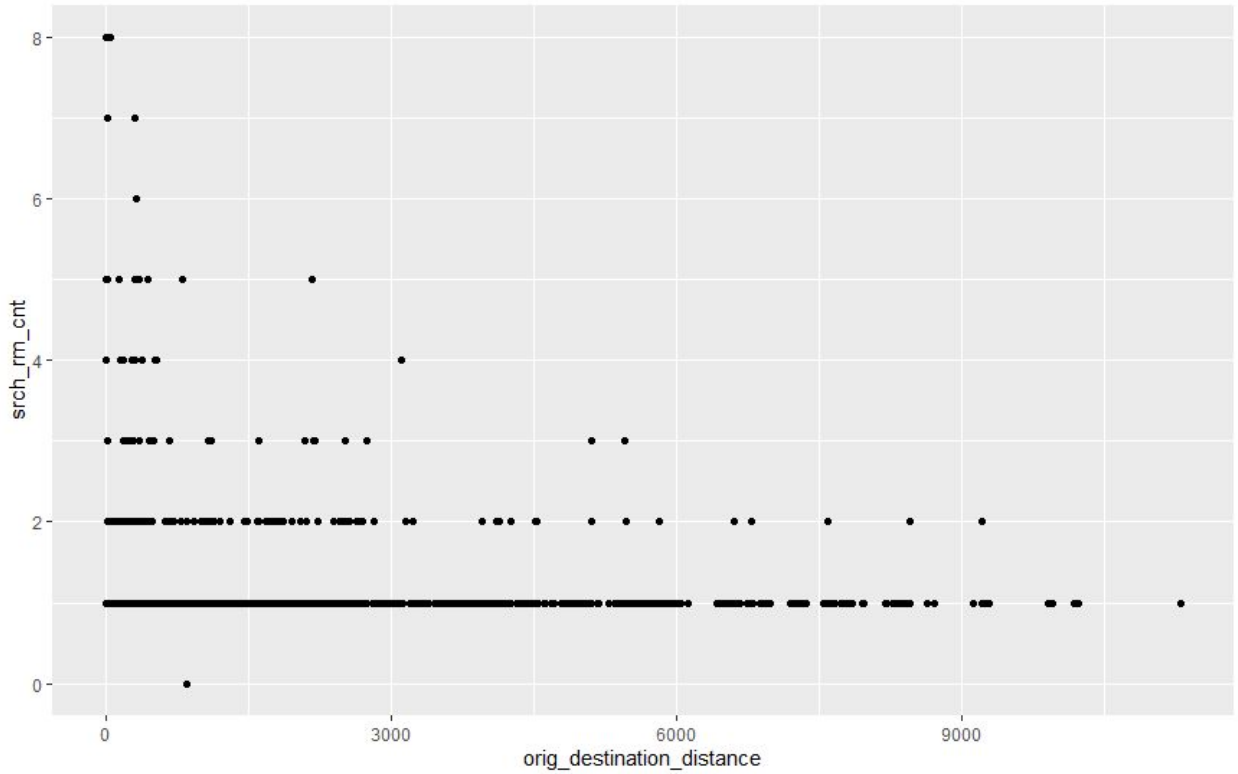
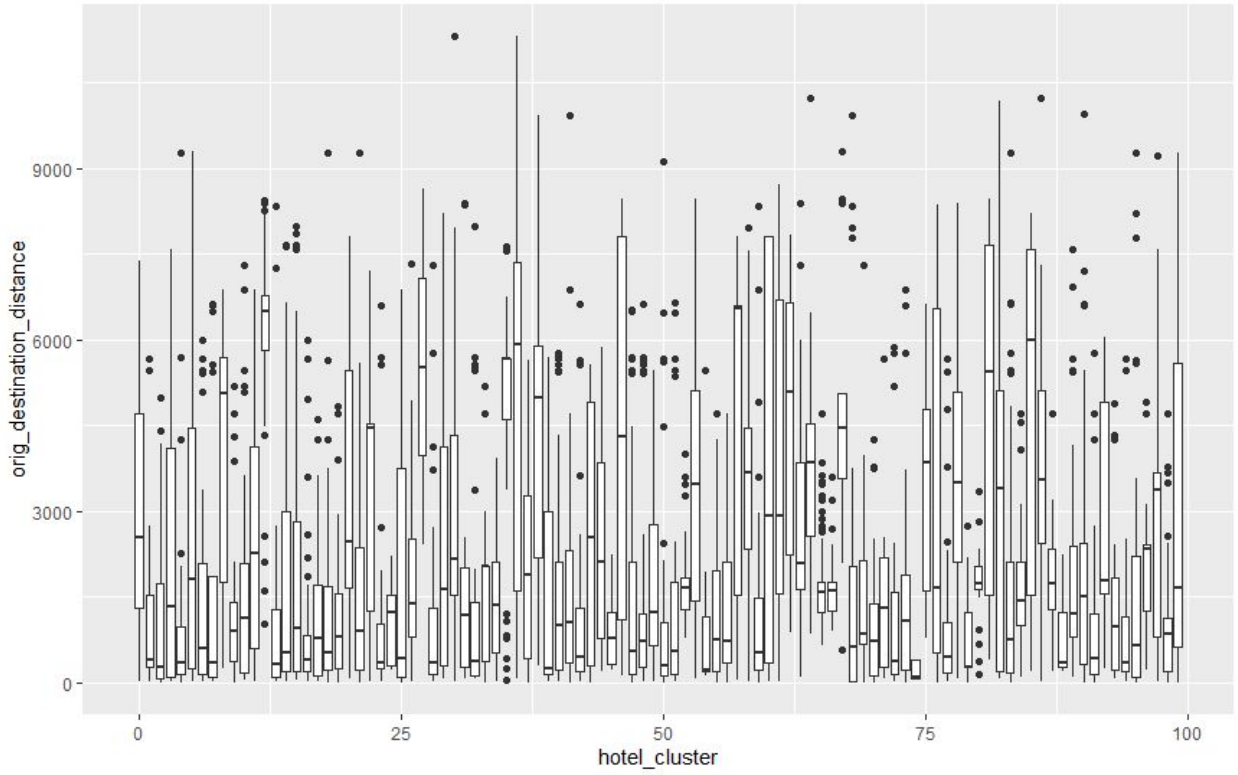
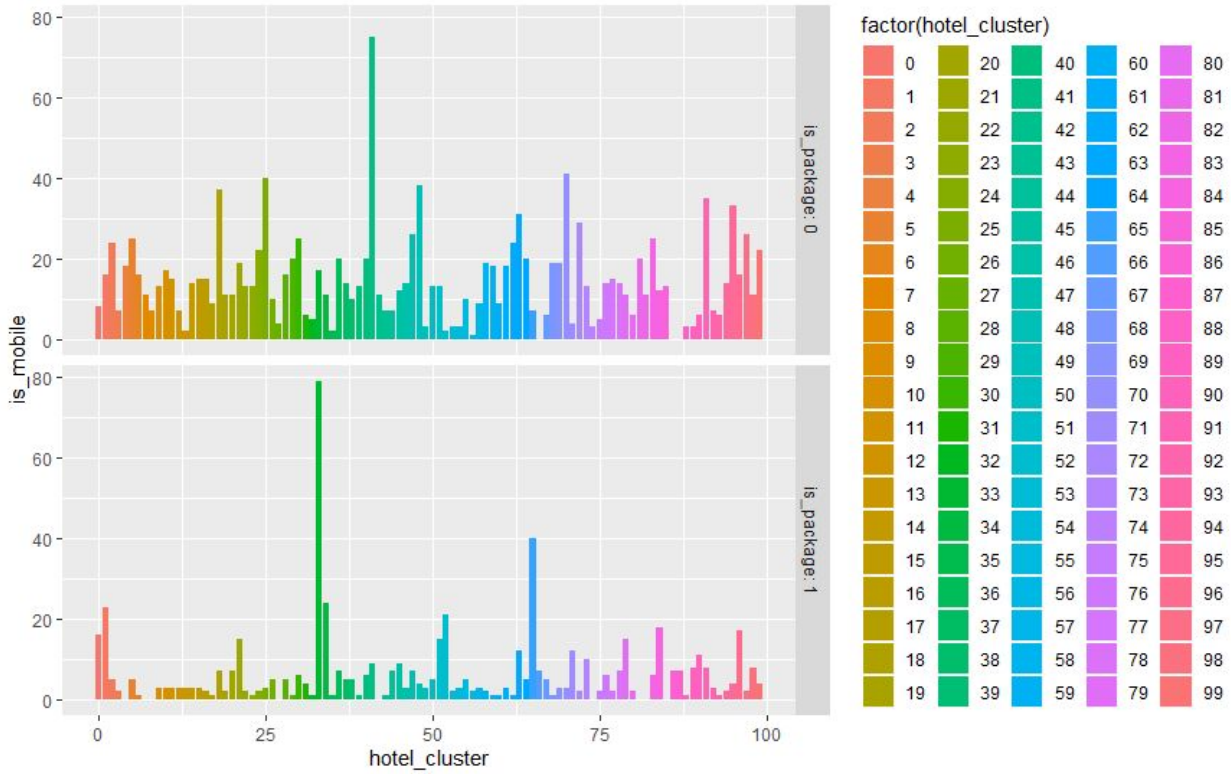
```
library(ggplot2)
library(plotly)
library(ggcorrplot)
library(dplyr)

df1 <- read.csv("C:/Master/Semester 4/expedia-hotel-recommendations/train.csv",
nrows=10000)

# Compute correlation Matrix
df1 %>% select_if(is.numeric)-> nums
corr <- round(cor(nums), 1)
head(corr[, 1:6])
ggcorrplot(corr)

#Histogram to show distribution of hotel clusters
ggplot(data = df1) +
  geom_bar(mapping = aes(x = hotel_cluster))

# Scatterplot srch_rm_cnt & orig_destination_distance
ggplot(data = df1, mapping = aes(x = orig_destination_distance, y = srch_rm_cnt)) +
  geom_point()

# Boxplot origination_destination_distance by hotel_cluster
ggplot(data = df1, mapping = aes(x = hotel_cluster, y = orig_destination_distance, group =
hotel_cluster)) +
  geom_boxplot()

# check relation between is_package, is_mobile and hotel_cluster.
ggplot(df1, aes(x = hotel_cluster, y = is_mobile)) +
  geom_bar(aes(fill=factor(hotel_cluster)), stat = 'identity') +
  facet_grid(is_package~., labeller = label_both)
```


hotel_cluster
hotel_market
hotel_country
hotel_continent
cnt
is_booking
srch_destination_type_id
srch_destination_id
srch_rm_cnt
srch_children_cnt
srch_adults_cnt
channel
is_package
is_mobile
user_id
orig_destination_distance
user_location_city
user_location_region
user_location_country
posa_continent
site_name
Corr
1.0
0.5
0.0
-0.5
-1.0


count
200
100
0
0
25
50
75
100
hotel_cluster


srch_rm_cnt
8
6
4
2
0
0
3000
6000
9000
orig_destination_distance


orig_destination_distance
9000
6000
3000
0
0
25
50
75
100
hotel_cluster


is_package:0
is_package:1
is_mobile
hotel_cluster
0
25
50
75
100
0
20
40
60
80
factor(hotel_cluster)
0 20 40 60 80
1 21 41 61 81
2 22 42 62 82
3 23 43 63 83
4 24 44 64 84
5 25 45 65 85
6 26 46 66 86
7 27 47 67 87
8 28 48 68 88
9 29 49 69 89
10 30 50 70 90
11 31 51 71 91
12 32 52 72 92
13 33 53 73 93
14 34 54 74 94
15 35 55 75 95
16 36 56 76 96
17 37 57 77 97
18 38 58 78 98
19 39 59 79 99