Air BnB Amenities

Braeden Norman

2021-11-08

```
library(tidyverse)
library(testthat)
listings <- read.csv("data/Listings.csv")
```

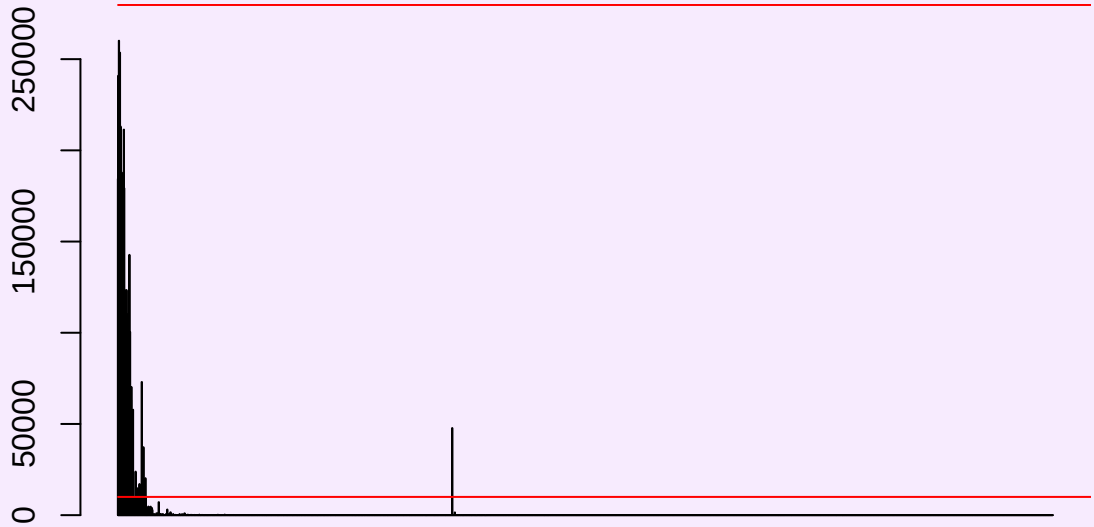
```
listOfAllamenities <- vector()
amenitiesCount <- integer(length(listOfAllamenities))
for (ii in 1:length(listings$amenities)) {#length(listings$amenities)
  splitString <- strsplit(str_replace_all(listings$amenities[ii], '[^( |")[:alnum:]]', ""), '"')
  # progress tracker
  # if (ii %% 10000 == 0) print(ii)
  if (!is_empty(splitString[[1]])) {
      for (nn in 1:length(splitString[[1]])) {
      if ( (nn %% 2) == 0 && !(splitString[[1]][nn] %in% listOfAllamenities)) {
        listOfAllamenities <- c(listOfAllamenities, splitString[[1]][nn])
        amenitiesCount <- c(amenitiesCount, 0)
      }
        index <- match(splitString[[1]][nn], listOfAllamenities)
        amenitiesCount[index] <- amenitiesCount[index] + 1
    }
  }

}
# remove space
amenitiesCount <- amenitiesCount[-match(" ", listOfAllamenities)]
listOfAllamenities <- listOfAllamenities[-match(" ", listOfAllamenities)]
head(listOfAllamenities)
```

```
## [1] "Heating"                "Kitchen"
## [3] "Washer"                 "Wifi"
## [5] "Long term stays allowed" "Shampoo"
```

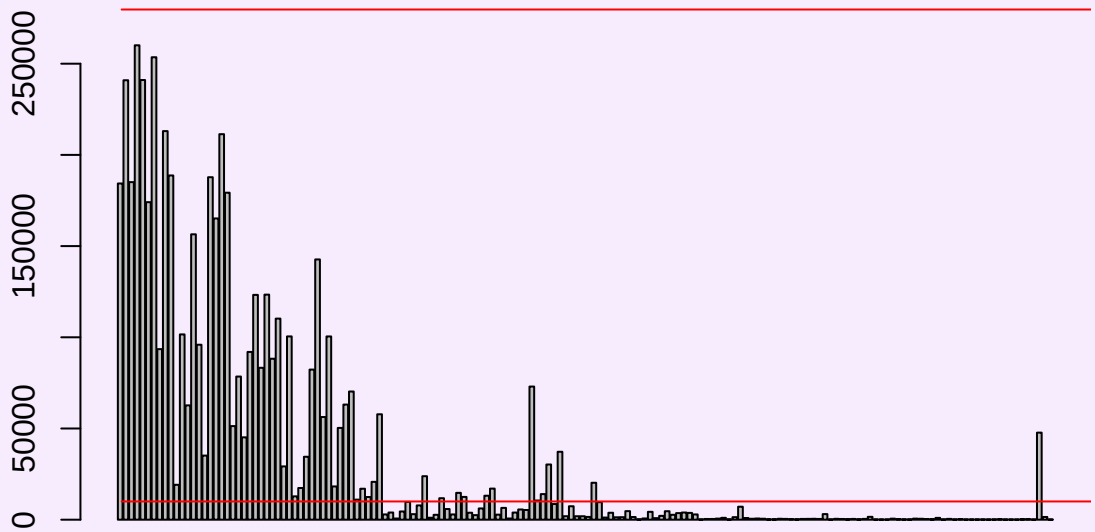
```
head(amenitiesCount)
```

```
## [1] 184327 240923 185073 260090 241054 174082
```

```r
barplot(amenitiesCount, ylim = c(0, 280000))
lines((integer(279712) + 1)*279712, col = "red")
lines((integer(279712) + 1)*10000, col = "red")
```



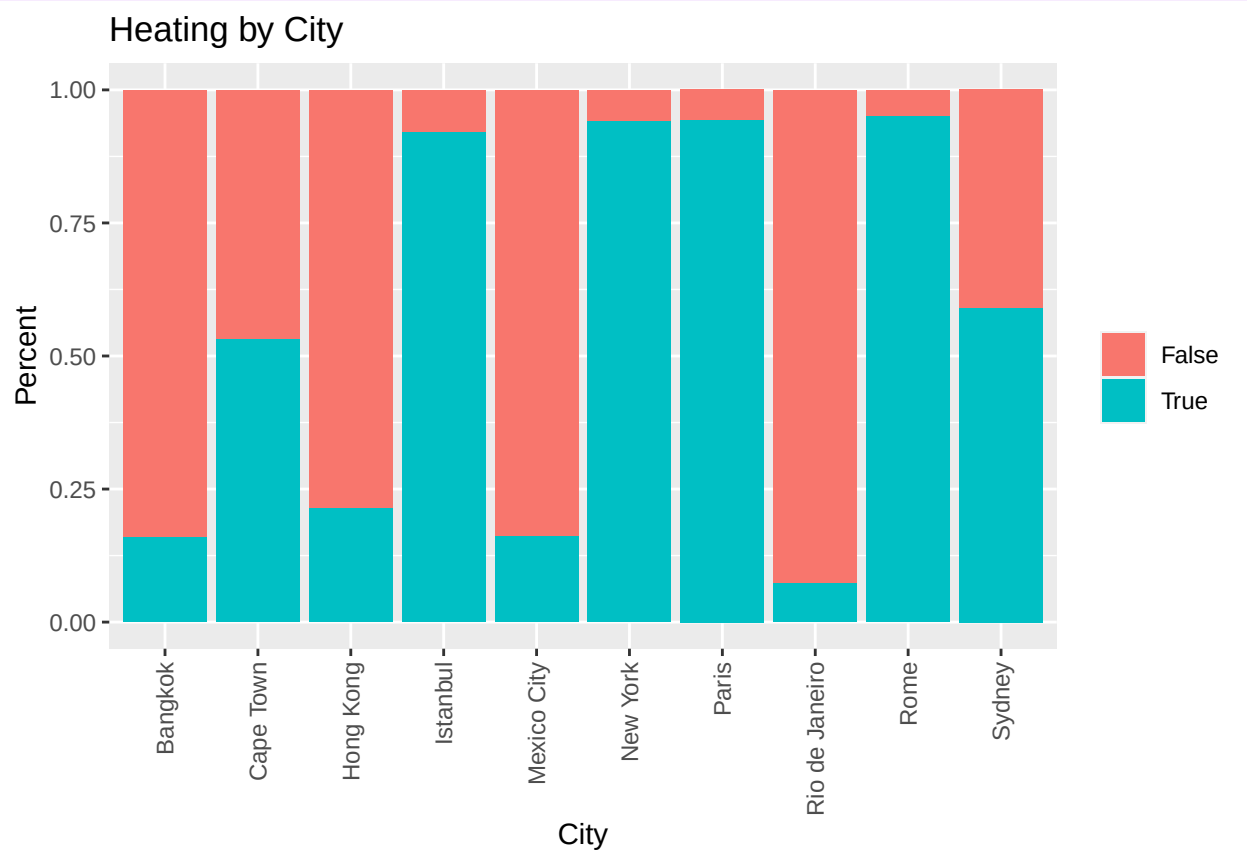

```r
# barplot(round(log2(amenitiesCount), digits = 0), ylim =  c(0, log2(280000)) ,ylab = "Log of count")
# lines((integer(279712) + 1)*log2(279712), col = "red")
# lines((integer(279712) + 1)*log2(10000), col = "red")

noOnes <- amenitiesCount[-which( amenitiesCount < 100)]

barplot(noOnes, ylim = c(0, 280000))
lines((integer(279712) + 1)*279712, col = "red")
lines((integer(279712) + 1)*10000, col = "red")
```

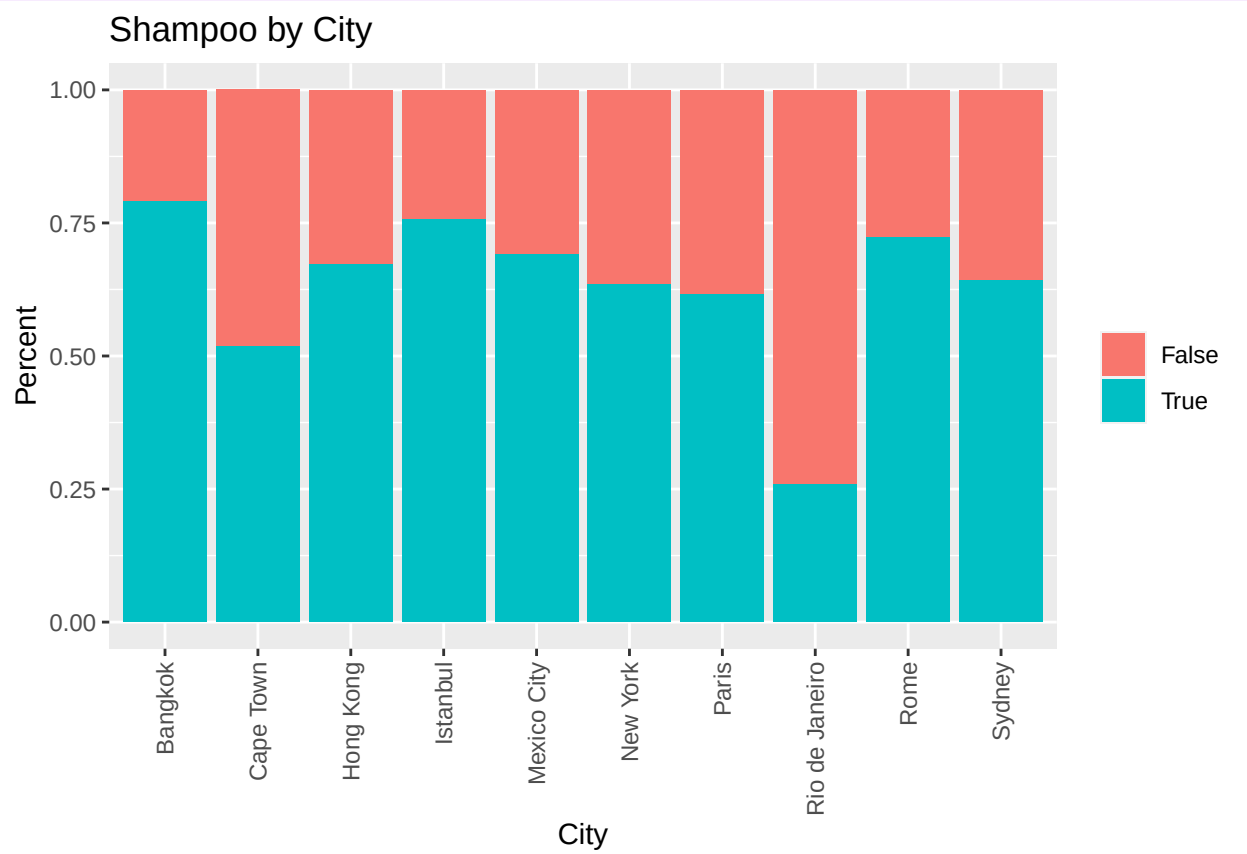


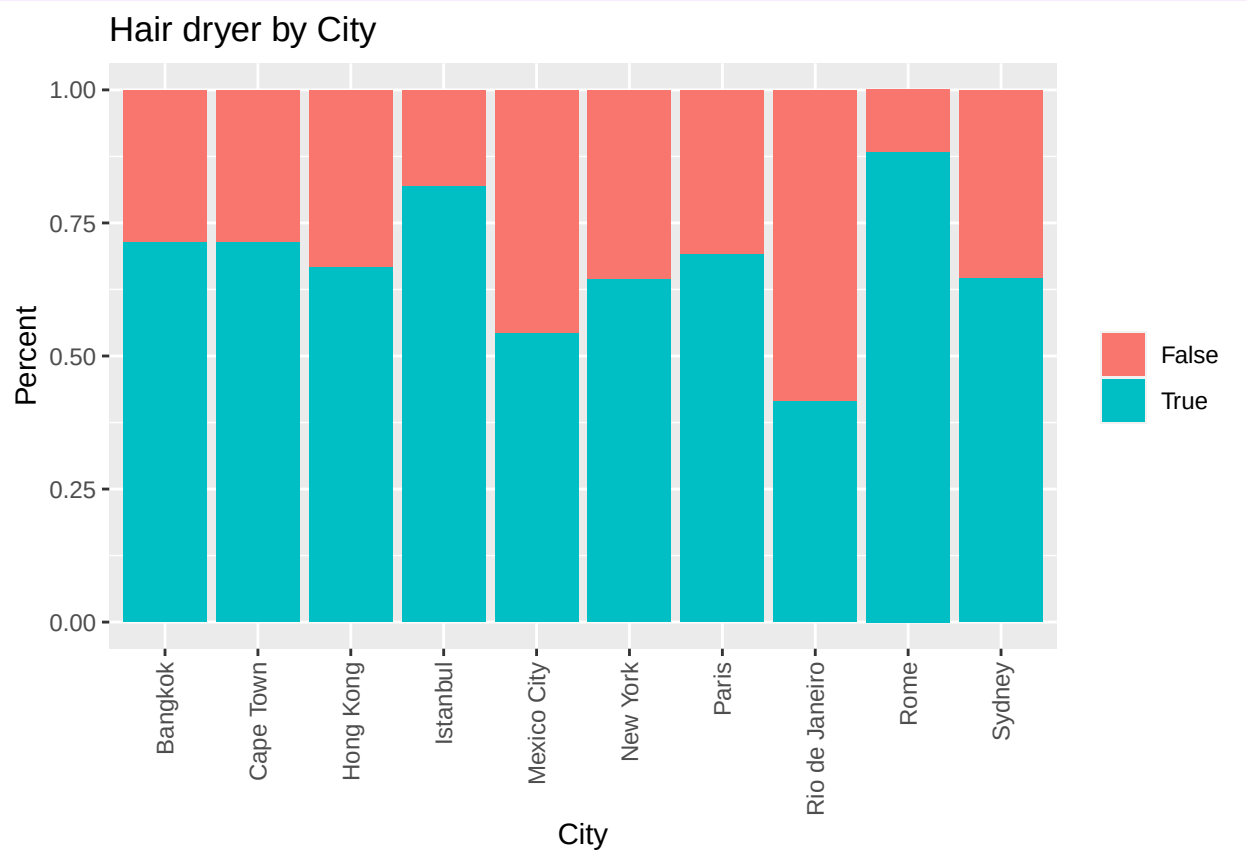

```
sum(amenitiesCount > 10000)
```

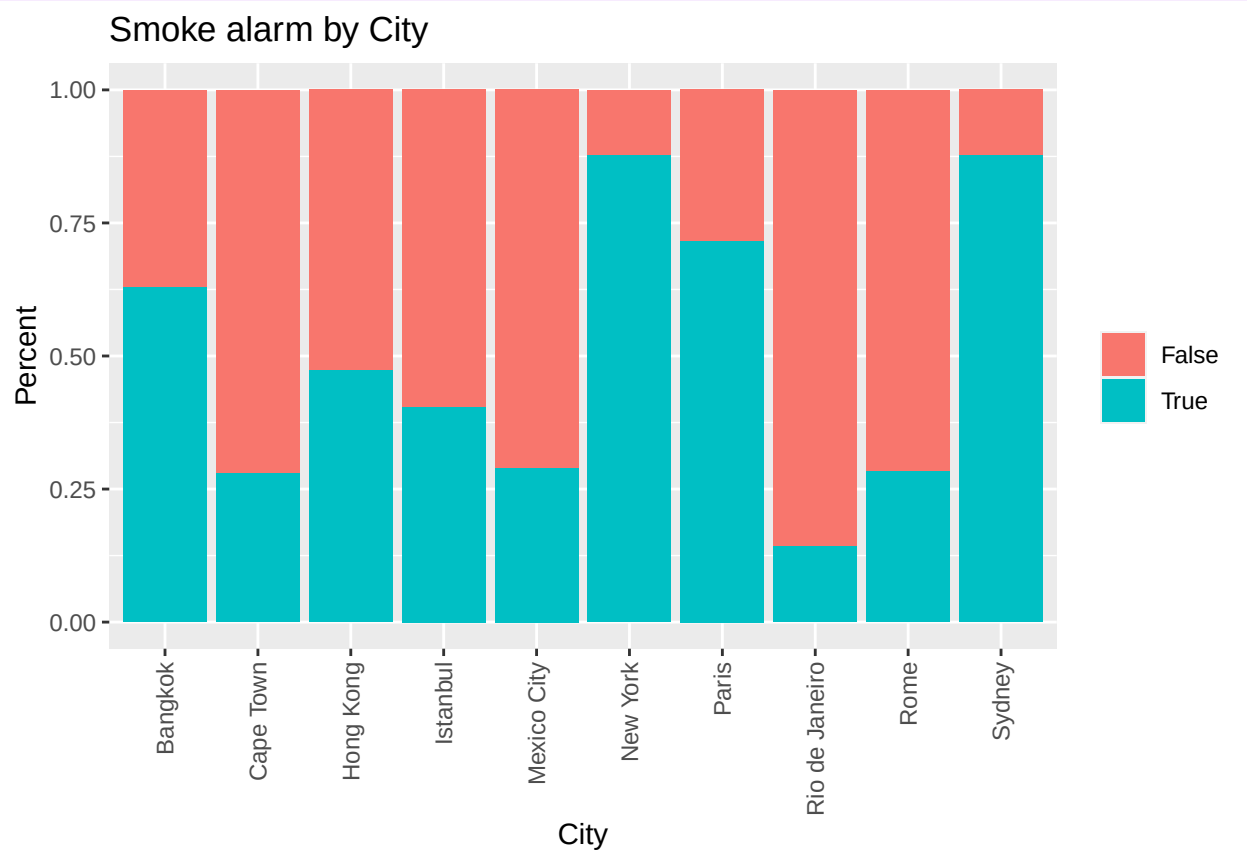
```
## [1] 60
```

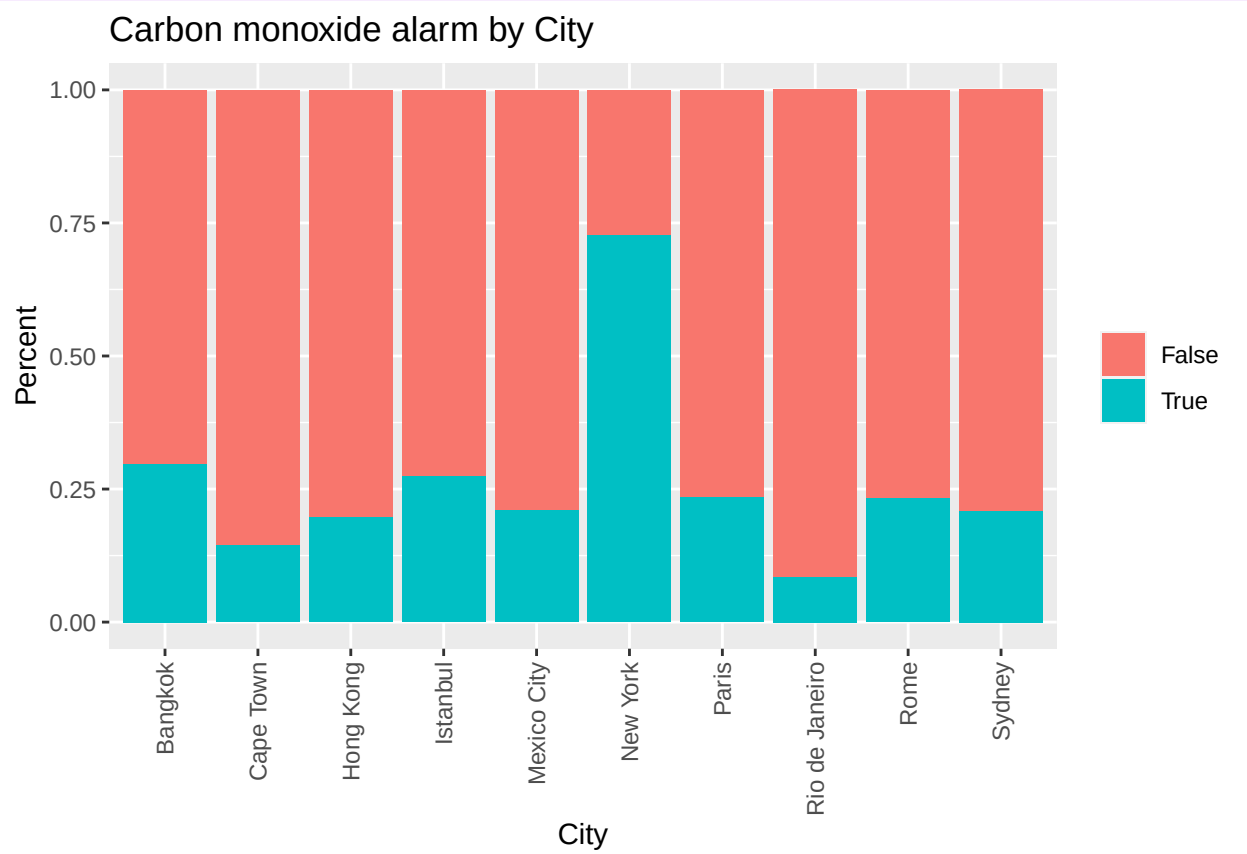
```r
reducedAmenities <- listOfAllamenities[which(amenitiesCount > 10000)]
reducedAmenitiesCount <- amenitiesCount[which(amenitiesCount > 10000)]

newListings <- listings


newData <- array(0, c(length(reducedAmenities),nrow(listings)))

for (ii in 1:nrow(listings)) {
  # progress tracker
  # if (ii %% 10000 == 0) print(ii)
  splitString <- strsplit(str_replace_all(listings$amenities[ii], '[^( |")[:alnum:]]', ""), '"')
  if (!is_empty(splitString[[1]])) {
    for (nn in 1:length(splitString[[1]])) {
      index <- match(splitString[[1]][nn], reducedAmenities)
      if (!is.na(index)) {
        newData[index,ii] = 1
      }
    }
  }
}

# add to new dataframe
for (ii in 1:length(reducedAmenities)) {
  newListings[,reducedAmenities[ii]] <- newData[ii,]
}


test_that("Check that count in new dataframe is same as originally calculated", {
    for (ii in 1:length(reducedAmenities)) {
      expect_equal(sum(newListings[,reducedAmenities[ii]]), reducedAmenitiesCount[ii])
    }
  })



## Test passed


# write.csv(newListings,"data/Listings_updated.csv")



selected <- c(1,6,10,14,22,36,54,58)
for (ii in selected) { #1:length(reducedAmenities)) {
   plt <- ggplot(newListings, aes(x = city, fill = factor(newListings[,reducedAmenities[ii]],
                         levels = c(0, 1), labels = c("False", "True")))) +
    geom_bar(position = "fill") +
    labs(y = "Percent", fill = "", x = "City", title =  paste(reducedAmenities[ii],"by City")) +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
    print(plt)
}
```

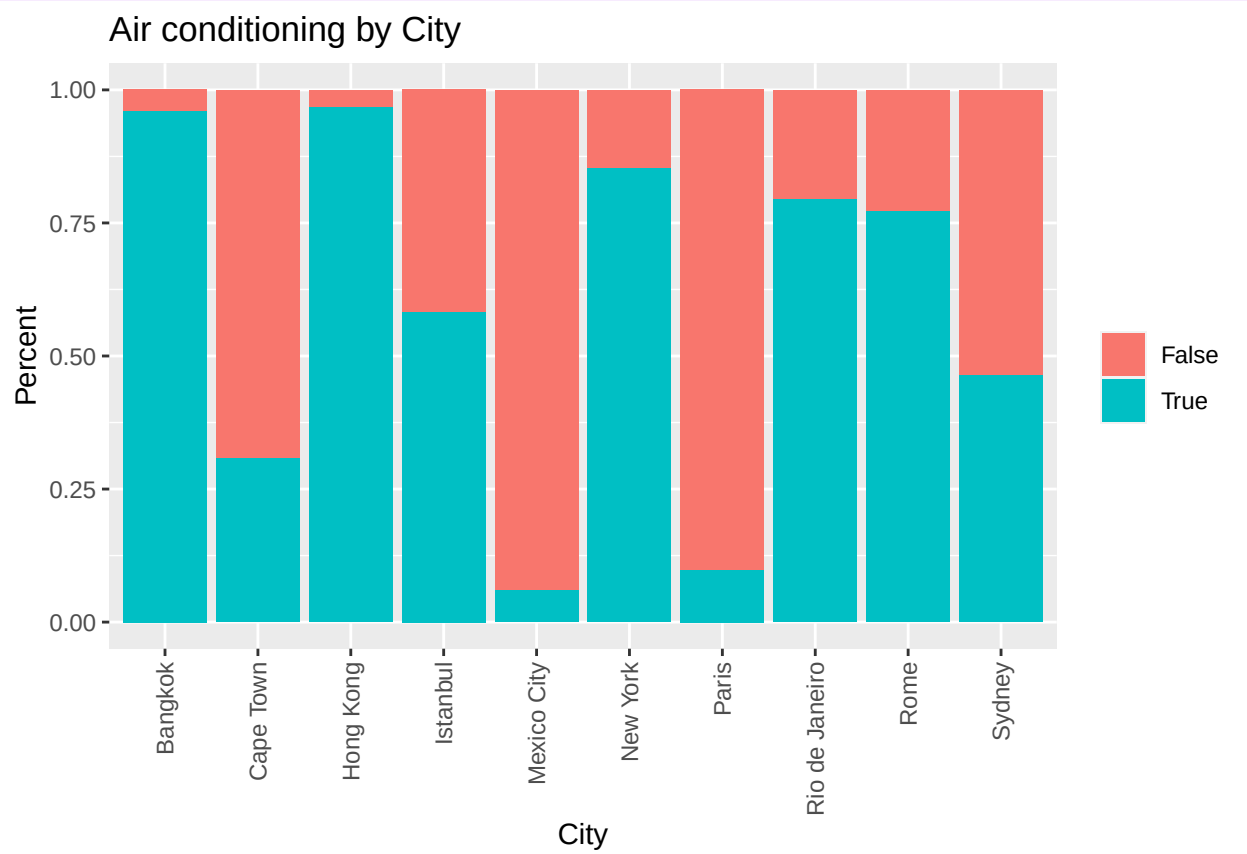

Heating by City
1.00
0.75
0.50
0.25
0.00
Percent
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True

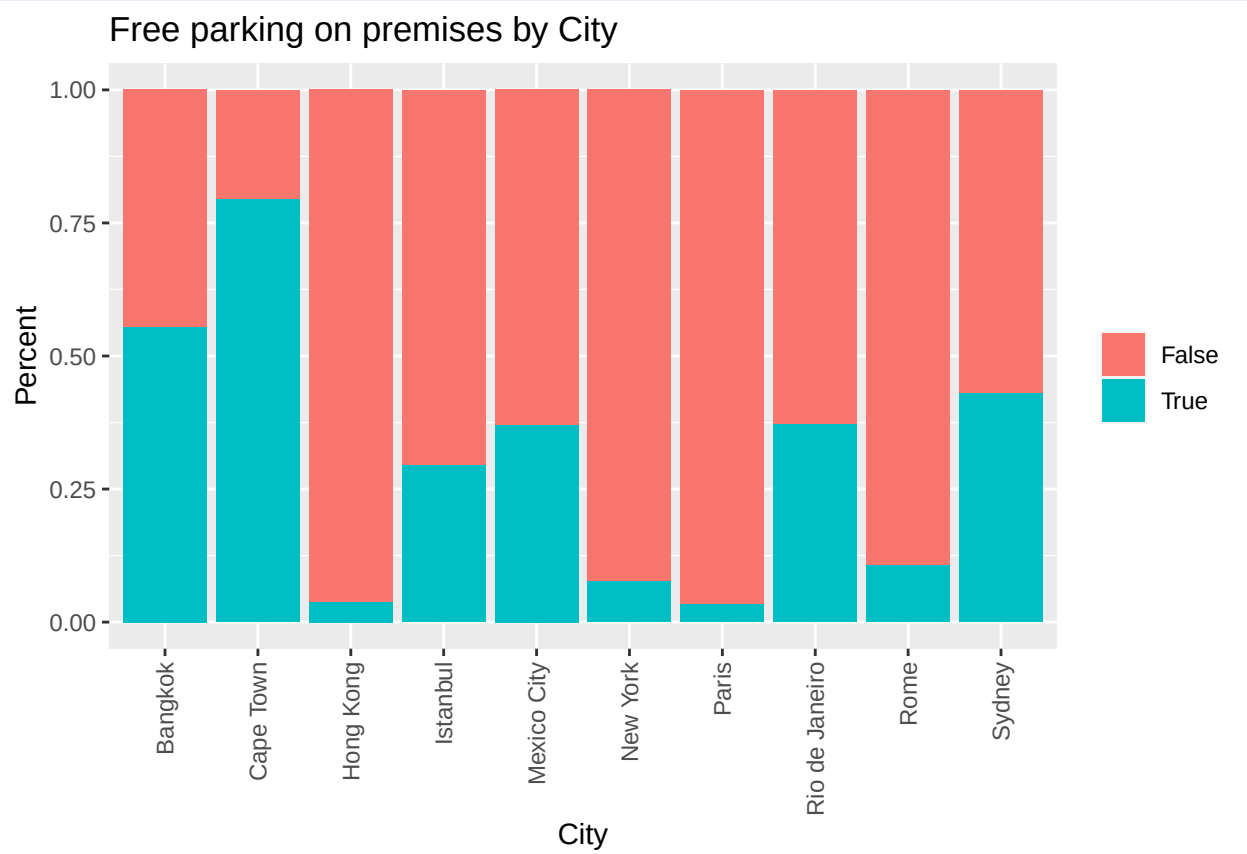

Shampoo by City
1.00
0.75
0.50
0.25
0.00
Percent
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True

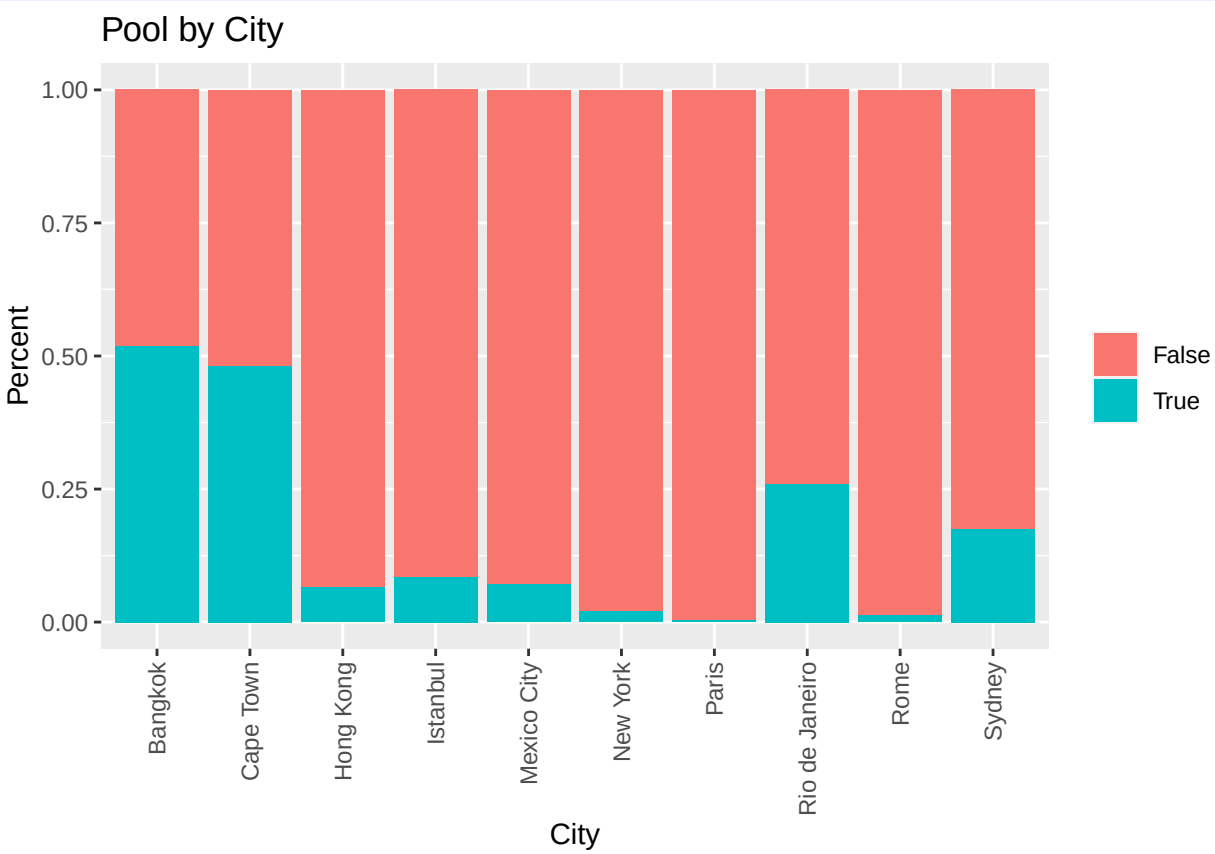

Hair dryer by City
1.00
0.75
0.50
0.25
0.00
Percent
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Smoke alarm by City
1.00
0.75
0.50
0.25
0.00
Percent
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Carbon monoxide alarm by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Air conditioning by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Free parking on premises by City
Percent
1.00
0.75
0.50
0.25
0.00
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True


Pool by City
1.00
0.75
0.50
0.25
0.00
Percent
Bangkok
Cape Town
Hong Kong
Istanbul
Mexico City
New York
Paris
Rio de Janeiro
Rome
Sydney
City
False
True