Airbnb project Naive Bayes

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(ggplot2)
library(naniar)
library(dplyr)
library(class)
library(AER)
```

```
## Cargando paquete requerido: car
## Cargando paquete requerido: carData
##
## Adjuntando el paquete: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
##
## Cargando paquete requerido: lmtest
## Cargando paquete requerido: zoo
##
## Adjuntando el paquete: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Cargando paquete requerido: sandwich
## Cargando paquete requerido: survival
```

```r
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
library(visualize)
library(caret)
```

```
## Cargando paquete requerido: lattice
##
## Adjuntando el paquete: 'caret'
##
## The following object is masked from 'package:survival':
##
##     cluster
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(FNN)
```

```
##
## Adjuntando el paquete: 'FNN'
##
## The following objects are masked from 'package:class':
##
##     knn, knn.cv
```

```r
# don't show scientific notation
options(scipen = 999)
#read data
zurich <- read_csv("C:\\Users\\amaia\\OneDrive\\Escritorio\\Data Mining\\Assignments\\Group\\zur
ich_listings.csv")
```

```
## Rows: 2819 Columns: 75
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (29): listing_url, last_scraped, source, name, neighborhood_overview, pi...
## dbl (36): id, scrape_id, host_id, host_listings_count, host_total_listings_c...
## lgl (10): description, host_is_superhost, host_has_profile_pic, host_identit...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Read the data in R and browsed through data

```
colSums(is.na(zurich))
```

```
##                                              id
##                                               0
##                                     listing_url
##                                               0
##                                       scrape_id
##                                               0
##                                    last_scraped
##                                               0
##                                          source
##                                               0
##                                            name
##                                               0
##                                     description
##                                            2819
##                           neighborhood_overview
##                                            1526
##                                     picture_url
##                                               0
##                                         host_id
##                                               0
##                                        host_url
##                                               0
##                                       host_name
##                                               0
##                                      host_since
##                                               0
##                                   host_location
##                                             503
##                                      host_about
##                                            1181
##                              host_response_time
##                                               0
##                              host_response_rate
##                                               0
##                            host_acceptance_rate
##                                               0
##                               host_is_superhost
##                                               2
##                              host_thumbnail_url
##                                               0
##                                host_picture_url
##                                               0
##                              host_neighbourhood
##                                            2791
##                             host_listings_count
##                                               0
##                       host_total_listings_count
##                                               0
##                              host_verifications
##                                               0
##                            host_has_profile_pic
##                                               0
```

```
##                         host_identity_verified
##                                              0
##                                  neighbourhood
##                                           1526
##                         neighbourhood_cleansed
##                                              0
##                   neighbourhood_group_cleansed
##                                              0
##                                       latitude
##                                              0
##                                      longitude
##                                              0
##                                  property_type
##                                              0
##                                      room_type
##                                              0
##                                   accommodates
##                                              0
##                                      bathrooms
##                                           2819
##                                 bathrooms_text
##                                              1
##                                       bedrooms
##                                           2819
##                                           beds
##                                             74
##                                      amenities
##                                              0
##                                          price
##                                            447
##                                 minimum_nights
##                                              0
##                                 maximum_nights
##                                              0
##                         minimum_minimum_nights
##                                              0
##                         maximum_minimum_nights
##                                              0
##                         minimum_maximum_nights
##                                              0
##                         maximum_maximum_nights
##                                              0
##                         minimum_nights_avg_ntm
##                                              0
##                         maximum_nights_avg_ntm
##                                              0
##                               calendar_updated
##                                           2819
##                               has_availability
##                                            447
##                                availability_30
##                                              0
```

```
##                                availability_60
##                                              0
##                                availability_90
##                                              0
##                               availability_365
##                                              0
##                           calendar_last_scraped
##                                              0
##                              number_of_reviews
##                                              0
##                          number_of_reviews_ltm
##                                              0
##                         number_of_reviews_l30d
##                                              0
##                                   first_review
##                                            731
##                                    last_review
##                                            731
##                           review_scores_rating
##                                            730
##                         review_scores_accuracy
##                                            730
##                      review_scores_cleanliness
##                                            730
##                          review_scores_checkin
##                                            730
##                    review_scores_communication
##                                            730
##                         review_scores_location
##                                            730
##                            review_scores_value
##                                            730
##                                        license
##                                           2819
##                               instant_bookable
##                                              0
##                 calculated_host_listings_count
##                                              0
##   calculated_host_listings_count_entire_homes
##                                              0
## calculated_host_listings_count_private_rooms
##                                              0
##   calculated_host_listings_count_shared_rooms
##                                              0
##                              reviews_per_month
##                                            731
```

Chcking how many missing values in each of the columns

```
b <- unique(zurich$amenities)
print(b)
```

```
## [1] "[]"
```

It appears amenities has only [], no actual values in it

```
# remove cols with no value at all
zurich1 <- zurich[,-c(7,36,38,40,50,69)]
colSums(is.na(zurich1))
```

```
##                                   id
##                                    0
##                          listing_url
##                                    0
##                            scrape_id
##                                    0
##                         last_scraped
##                                    0
##                               source
##                                    0
##                                 name
##                                    0
##                neighborhood_overview
##                                 1526
##                          picture_url
##                                    0
##                              host_id
##                                    0
##                             host_url
##                                    0
##                            host_name
##                                    0
##                           host_since
##                                    0
##                        host_location
##                                  503
##                           host_about
##                                 1181
##                   host_response_time
##                                    0
##                   host_response_rate
##                                    0
##                 host_acceptance_rate
##                                    0
##                    host_is_superhost
##                                    2
##                   host_thumbnail_url
##                                    0
##                     host_picture_url
##                                    0
##                   host_neighbourhood
##                                 2791
##                  host_listings_count
##                                    0
##            host_total_listings_count
##                                    0
##                   host_verifications
##                                    0
##                 host_has_profile_pic
##                                    0
##               host_identity_verified
##                                    0
```

```
##                   neighbourhood
##                            1526
##          neighbourhood_cleansed
##                               0
##    neighbourhood_group_cleansed
##                               0
##                        latitude
##                               0
##                       longitude
##                               0
##                   property_type
##                               0
##                       room_type
##                               0
##                    accommodates
##                               0
##                  bathrooms_text
##                               1
##                            beds
##                              74
##                           price
##                             447
##                  minimum_nights
##                               0
##                  maximum_nights
##                               0
##          minimum_minimum_nights
##                               0
##          maximum_minimum_nights
##                               0
##          minimum_maximum_nights
##                               0
##          maximum_maximum_nights
##                               0
##          minimum_nights_avg_ntm
##                               0
##          maximum_nights_avg_ntm
##                               0
##                has_availability
##                             447
##                 availability_30
##                               0
##                 availability_60
##                               0
##                 availability_90
##                               0
##                availability_365
##                               0
##           calendar_last_scraped
##                               0
##               number_of_reviews
##                               0
```

```
##                          number_of_reviews_ltm
##                                              0
##                         number_of_reviews_l30d
##                                              0
##                                   first_review
##                                            731
##                                    last_review
##                                            731
##                           review_scores_rating
##                                            730
##                         review_scores_accuracy
##                                            730
##                      review_scores_cleanliness
##                                            730
##                          review_scores_checkin
##                                            730
##                    review_scores_communication
##                                            730
##                         review_scores_location
##                                            730
##                            review_scores_value
##                                            730
##                               instant_bookable
##                                              0
##                 calculated_host_listings_count
##                                              0
##  calculated_host_listings_count_entire_homes
##                                              0
## calculated_host_listings_count_private_rooms
##                                              0
##  calculated_host_listings_count_shared_rooms
##                                              0
##                              reviews_per_month
##                                            731
```

Now the data is saved in new df zurich1 with only cols that have values, removed- bedrooms, bathrooms, calender_update, license, description. In addition removing amenities, which has no actual values, only []

```
test_cases <- complete.cases(zurich1)
l <- sum(test_cases)
percentage <- (l/nrow(zurich1))*100
cat("percentage  and l", percentage, l)
```

```
## percentage  and l 0.1773679 5
```

```
library(naniar)
missing_var <- miss_var_summary(zurich1)
print(missing_var)
```

```
## # A tibble: 69 × 3
##    variable                   n_miss pct_miss
##    <chr>                       <int>    <num>
##  1 host_neighbourhood           2791     99.0
##  2 neighborhood_overview        1526     54.1
##  3 neighbourhood                1526     54.1
##  4 host_about                   1181     41.9
##  5 first_review                  731     25.9
##  6 last_review                   731     25.9
##  7 reviews_per_month             731     25.9
##  8 review_scores_rating          730     25.9
##  9 review_scores_accuracy        730     25.9
## 10 review_scores_cleanliness     730     25.9
## # ℹ 59 more rows
```

Viewing percentage of values missing per each column

```
# Review_score_value converted to factor and new variable is Review_value
zurich1$review_scores_value[is.na(zurich1$review_scores_value)] <- 0
summary(zurich1$review_scores_value)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    4.58    3.43    4.83    5.00
```

```
bins= c(-Inf,3.5,4.5,5)
zurich1$review_value <- cut(zurich1$review_scores_value, breaks= bins, labels= c("poor_review
s","moderate_reviews", "good_reviews" ))
table(zurich1$review_value)
```

```
##
##     poor_reviews moderate_reviews     good_reviews
##              787              517             1515
```

review_scores_value has 730 NA’s, imputing missing values with mean(4.629) or median (4.71) may not be appropriate as NA means it must not be reviewed yet as the listing is new or no one has lived there to review.

```
# Review_score_rating converted to factor and new variable is Review_rating

zurich1$review_scores_rating[is.na(zurich1$review_scores_rating)] <- 0
quantile_edges <- quantile(zurich1$review_scores_rating, probs = c(0, 1/3, 2/3, 1), na.rm = TRU
E)
zurich1$review_rating <- cut(zurich1$review_scores_rating,
                             breaks = quantile_edges,
                             labels = c("poor_reviews", "moderate_reviews", "good_reviews"),
                             include.lowest = TRUE)

table(zurich1$review_rating)
```

```
##
##     poor_reviews moderate_reviews     good_reviews
##              942              945              932
```

```
summary(zurich1$review_rating)
```

```
##     poor_reviews moderate_reviews     good_reviews
##              942              945              932
```

Followed similar process for rating

```
# Review_score_rating converted to factor and new variable is Review_rating
zurich1$review_scores_accuracy[is.na(zurich1$review_scores_accuracy)] <- 0
summary(zurich1$review_scores_accuracy)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   4.800   3.547   5.000   5.000
```

```
bins= c(-Inf,0,4,5)
zurich1$review_accuracy <- cut(zurich1$review_scores_accuracy, breaks= bins, labels= c("no_revie
ws", "poor_reviews","good_reviews" ))
table(zurich1$review_accuracy)
```

```
##
##   no_reviews poor_reviews good_reviews
##          730           99         1990
```

Followe similar process for accuracy

```
# Review_score_cleanliness  converted to factor and new variable is Review_cleanliness
zurich1$review_scores_cleanliness [is.na(zurich1$review_scores_cleanliness )] <- 0
summary(zurich1$review_scores_cleanliness )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   4.750   3.521   4.980   5.000
```

```
bins= c(-Inf,0,4,5)
zurich1$review_cleanliness  <- cut(zurich1$review_scores_cleanliness , breaks= bins, labels= c
("no_reviews", "poor_reviews","good_reviews" ))
table(zurich1$review_cleanliness )
```

```
##
##   no_reviews poor_reviews good_reviews
##          730          121         1968
```

Followed similar process for cleanliness

```
# Review_score_checkin  converted to factor and new variable is Review_checkin
zurich1$review_scores_checkin [is.na(zurich1$review_scores_checkin )] <- 0
summary(zurich1$review_scores_checkin )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    4.87    3.59    5.00    5.00
```

```
bins= c(-Inf,0,4,5)
zurich1$review_checkin  <- cut(zurich1$review_scores_checkin , breaks= bins, labels= c("no_revie
ws", "poor_reviews","good_reviews" ))
table(zurich1$review_checkin )
```

```
##
##   no_reviews poor_reviews good_reviews
##          730           70         2019
```

Followed similar process for checkin

```
# Review_score_communication  converted to factor and new variable is Review_communication
zurich1$review_scores_communication [is.na(zurich1$review_scores_communication )] <- 0
summary(zurich1$review_scores_communication )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   4.860   3.587   5.000   5.000
```

```
bins= c(-Inf,0,4,5)
zurich1$review_communication  <- cut(zurich1$review_scores_communication , breaks= bins, labels=
c("no_reviews", "poor_reviews", "good_reviews" ))
table(zurich1$review_communication )
```

```
##
##   no_reviews poor_reviews good_reviews
##          730           68         2021
```

Followed same process for communication

```
# Review_score_location   converted to factor and new variable is Review_location
zurich1$review_scores_location  [is.na(zurich1$review_scores_location  )] <- 0
summary(zurich1$review_scores_location  )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    4.78    3.55    4.97    5.00
```

```
bins= c(-Inf,0,4,5)
zurich1$review_location   <- cut(zurich1$review_scores_location  , breaks= bins, labels= c("no_r
eviews", "poor_reviews","good_reviews"))
table(zurich1$review_location  )
```

```
##
##   no_reviews poor_reviews good_reviews
##          730           87         2002
```

Followed same process for location

```
# Removing redundant review scores

zurich1 <- zurich1[,-c(57,58,59,60,61,62,62)]
dim(zurich1)
```

```
## [1] 2819   70
```

Removed 7 columns

```
# EXTRACTING PRICE & BATHS
zurich2 <- zurich1 %>%
          mutate(NumPrice=as.numeric(gsub("[$,]","",zurich1$price))) %>%
          mutate(baths=case_when(
            grepl("(half).*",  zurich1$bathrooms_text, ignore.case = TRUE) ~0.5,
            TRUE ~ as.numeric(gsub("[^0-9.]+","",zurich1$bathrooms_text))
          ))
head(zurich2$NumPrice)
```

```
## [1] 100  60 200  78 500  NA
```

```
colSums(is.na(zurich2))
```

```
##                                    id
##                                     0
##                           listing_url
##                                     0
##                             scrape_id
##                                     0
##                          last_scraped
##                                     0
##                                source
##                                     0
##                                  name
##                                     0
##                neighborhood_overview
##                                  1526
##                           picture_url
##                                     0
##                               host_id
##                                     0
##                              host_url
##                                     0
##                             host_name
##                                     0
##                            host_since
##                                     0
##                         host_location
##                                   503
##                            host_about
##                                  1181
##                    host_response_time
##                                     0
##                    host_response_rate
##                                     0
##                  host_acceptance_rate
##                                     0
##                     host_is_superhost
##                                     2
##                    host_thumbnail_url
##                                     0
##                      host_picture_url
##                                     0
##                    host_neighbourhood
##                                  2791
##                   host_listings_count
##                                     0
##             host_total_listings_count
##                                     0
##                    host_verifications
##                                     0
##                  host_has_profile_pic
##                                     0
##                host_identity_verified
##                                     0
```

```
##                                       neighbourhood
##                                                1526
##                              neighbourhood_cleansed
##                                                   0
##                        neighbourhood_group_cleansed
##                                                   0
##                                            latitude
##                                                   0
##                                           longitude
##                                                   0
##                                       property_type
##                                                   0
##                                           room_type
##                                                   0
##                                        accommodates
##                                                   0
##                                      bathrooms_text
##                                                   1
##                                                beds
##                                                  74
##                                               price
##                                                 447
##                                      minimum_nights
##                                                   0
##                                      maximum_nights
##                                                   0
##                              minimum_minimum_nights
##                                                   0
##                              maximum_minimum_nights
##                                                   0
##                              minimum_maximum_nights
##                                                   0
##                              maximum_maximum_nights
##                                                   0
##                              minimum_nights_avg_ntm
##                                                   0
##                              maximum_nights_avg_ntm
##                                                   0
##                                    has_availability
##                                                 447
##                                     availability_30
##                                                   0
##                                     availability_60
##                                                   0
##                                     availability_90
##                                                   0
##                                    availability_365
##                                                   0
##                               calendar_last_scraped
##                                                   0
##                                   number_of_reviews
##                                                   0
```

```
##                        number_of_reviews_ltm
##                                            0
##                       number_of_reviews_l30d
##                                            0
##                                 first_review
##                                          731
##                                  last_review
##                                          731
##                          review_scores_value
##                                            0
##                             instant_bookable
##                                            0
##               calculated_host_listings_count
##                                            0
##  calculated_host_listings_count_entire_homes
##                                            0
## calculated_host_listings_count_private_rooms
##                                            0
##  calculated_host_listings_count_shared_rooms
##                                            0
##                            reviews_per_month
##                                          731
##                                 review_value
##                                            0
##                                review_rating
##                                            0
##                              review_accuracy
##                                            0
##                           review_cleanliness
##                                            0
##                               review_checkin
##                                            0
##                         review_communication
##                                            0
##                              review_location
##                                            0
##                                     NumPrice
##                                          447
##                                        baths
##                                            1
```

```
zurich2$baths <- ifelse(is.na(zurich2$baths), 1,zurich2$baths) # imputing the last na value
```

Extracting price in chr columnn and converting to interger and extracting baths and creating a ratio variable guests per bath

```
#***To me removed later***
# Imputing missing price values
u_room_type<- unique(zurich2$room_type)
u_property_type <- unique(zurich2$property_type)
print(u_room_type)
```

```
## [1] "Entire home/apt" "Private room"    "Hotel room"      "Shared room"
```

```
print(u_property_type)
```

```
##  [1] "Entire rental unit"                 "Private room in rental unit"
##  [3] "Private room in home"               "Entire loft"
##  [5] "Entire condo"                       "Entire home"
##  [7] "Private room in castle"             "Private room in condo"
##  [9] "Private room in townhouse"          "Entire serviced apartment"
## [11] "Private room in hut"                "Private room in guesthouse"
## [13] "Private room in villa"              "Tiny home"
## [15] "Room in boutique hotel"             "Private room in loft"
## [17] "Private room in bed and breakfast"  "Entire townhouse"
## [19] "Entire guest suite"                 "Entire villa"
## [21] "Shared room in hostel"              "Room in serviced apartment"
## [23] "Shared room in rental unit"         "Room in bed and breakfast"
## [25] "Room in hotel"                      "Entire guesthouse"
## [27] "Private room in serviced apartment" "Barn"
## [29] "Private room in cabin"              "Private room in casa particular"
## [31] "Private room"                       "Private room in chalet"
## [33] "Entire vacation home"               "Camper/RV"
## [35] "Casa particular"                    "Shared room in home"
## [37] "Shared room in hotel"
```

Unique property types

```
#***To me removed later***
test_u_property_type <- zurich2 %>% filter(property_type=="Casa particular")
test_u_room_type <- zurich2 %>% filter(room_type=="Hotel room")
test_u_beds <- unique(zurich2$beds)
table(test_u_beds)
```

```
## test_u_beds
##  1  2  3  4  5  6  7  8  9 10 18 32
##  1  1  1  1  1  1  1  1  1  1  1  1
```

```
summary(test_u_beds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    3.75    6.50    8.75    9.25   32.00       1
```

```
c<- mode(test_u_beds)
print(c)
```

```
## [1] "numeric"
```

test

```
#Imputing missing values in 'beds'
zurich2$beds <- ifelse(is.na(zurich2$beds) & zurich2$room_type == "Shared room", zurich2$accommo
dates,
                       ifelse(is.na(zurich2$beds) & zurich2$room_type %in% c("Private room", "En
tire home/apt") & zurich2$accommodates %in% 1:2, 1,
                              ifelse(is.na(zurich2$beds) & zurich2$room_type %in% c("Private roo
m", "Entire home/apt") & zurich2$accommodates %in% 3:8, ceiling(zurich2$accommodates/2),
                       zurich2$beds)))

colSums(is.na(zurich2))
```

```
##                                   id
##                                    0
##                          listing_url
##                                    0
##                            scrape_id
##                                    0
##                         last_scraped
##                                    0
##                               source
##                                    0
##                                 name
##                                    0
##                neighborhood_overview
##                                 1526
##                          picture_url
##                                    0
##                              host_id
##                                    0
##                             host_url
##                                    0
##                            host_name
##                                    0
##                           host_since
##                                    0
##                        host_location
##                                  503
##                           host_about
##                                 1181
##                   host_response_time
##                                    0
##                   host_response_rate
##                                    0
##                 host_acceptance_rate
##                                    0
##                    host_is_superhost
##                                    2
##                   host_thumbnail_url
##                                    0
##                     host_picture_url
##                                    0
##                   host_neighbourhood
##                                 2791
##                  host_listings_count
##                                    0
##            host_total_listings_count
##                                    0
##                   host_verifications
##                                    0
##                 host_has_profile_pic
##                                    0
##               host_identity_verified
##                                    0
```

```
##                     neighbourhood
##                              1526
##             neighbourhood_cleansed
##                                 0
##       neighbourhood_group_cleansed
##                                 0
##                          latitude
##                                 0
##                         longitude
##                                 0
##                     property_type
##                                 0
##                         room_type
##                                 0
##                      accommodates
##                                 0
##                    bathrooms_text
##                                 1
##                              beds
##                                 0
##                             price
##                               447
##                    minimum_nights
##                                 0
##                    maximum_nights
##                                 0
##            minimum_minimum_nights
##                                 0
##            maximum_minimum_nights
##                                 0
##            minimum_maximum_nights
##                                 0
##            maximum_maximum_nights
##                                 0
##            minimum_nights_avg_ntm
##                                 0
##            maximum_nights_avg_ntm
##                                 0
##                  has_availability
##                               447
##                   availability_30
##                                 0
##                   availability_60
##                                 0
##                   availability_90
##                                 0
##                  availability_365
##                                 0
##             calendar_last_scraped
##                                 0
##                 number_of_reviews
##                                 0
```

```
##                             number_of_reviews_ltm
##                                                 0
##                            number_of_reviews_l30d
##                                                 0
##                                      first_review
##                                               731
##                                       last_review
##                                               731
##                               review_scores_value
##                                                 0
##                                  instant_bookable
##                                                 0
##                    calculated_host_listings_count
##                                                 0
##  calculated_host_listings_count_entire_homes
##                                                 0
## calculated_host_listings_count_private_rooms
##                                                 0
##  calculated_host_listings_count_shared_rooms
##                                                 0
##                                 reviews_per_month
##                                               731
##                                      review_value
##                                                 0
##                                     review_rating
##                                                 0
##                                   review_accuracy
##                                                 0
##                               review_cleanliness
##                                                 0
##                                    review_checkin
##                                                 0
##                              review_communication
##                                                 0
##                                   review_location
##                                                 0
##                                          NumPrice
##                                               447
##                                             baths
##                                                 0
```

Imputed missing beds values such based on room type and accommodates combinations

```
# Creating new variabales guests per bath and bed
zurich_FE <- zurich2 %>% mutate(guestsPerBath= zurich2$accommodates/zurich2$baths) %>% mutate(gu
estsPerBed = zurich2$accommodates/zurich2$beds)
head(zurich_FE,2)
```

id	listing_url	scrape_id	last_scraped	source
<dbl>	<chr>	<dbl>	<chr>	<chr>
73282	https://www.airbnb.com/rooms/73282	20231200000000	12/28/2023	previous scrape
178448	https://www.airbnb.com/rooms/178448	20231200000000	12/27/2023	city scrape

2 rows | 1-5 of 74 columns

```
zurich_FE<- zurich_FE %>%
  mutate(guestsPerBath= ifelse(baths==0, 0, guestsPerBath))
```

Created new variables guests per bath and guests per bed

```
#***REDUNDANT STEP, REMOVE LATER****
zurich2_price_nonas<- zurich2 %>% filter(!is.na(NumPrice))
zurich2_price_nas<- zurich2 %>% filter(is.na(NumPrice))
zurich2_beds_nonas <- zurich2 %>% filter(is.na(beds))
price_imputing_mlr_model <- lm(NumPrice~neighbourhood_cleansed + neighbourhood_group_cleansed  +
room_type +accommodates+beds, zurich2_price_nonas)
step_mlr <- step(price_imputing_mlr_model,  method= "backward")
```

```
## Start:  AIC=29304.18
## NumPrice ~ neighbourhood_cleansed + neighbourhood_group_cleansed +
##     room_type + accommodates + beds
##
##
## Step:  AIC=29304.18
## NumPrice ~ neighbourhood_cleansed + room_type + accommodates +
##     beds
##
##                          Df Sum of Sq       RSS   AIC
## - neighbourhood_cleansed 33  10763366 543114654 29286
## - room_type               3    294358 532645646 29300
## - beds                    1     42266 532393554 29302
## <none>                                532351288 29304
## - accommodates            1   2472270 534823558 29313
##
## Step:  AIC=29285.66
## NumPrice ~ room_type + accommodates + beds
##
##                Df Sum of Sq       RSS   AIC
## - room_type     3    445725 543560379 29282
## - beds          1     16501 543131154 29284
## <none>                      543114654 29286
## - accommodates  1   3052382 546167035 29297
##
## Step:  AIC=29281.61
## NumPrice ~ accommodates + beds
##
##                Df Sum of Sq       RSS   AIC
## - beds          1     10425 543570804 29280
## <none>                      543560379 29282
## - accommodates  1   3929337 547489716 29297
##
## Step:  AIC=29279.66
## NumPrice ~ accommodates
##
##                Df Sum of Sq       RSS   AIC
## <none>                      543570804 29280
## - accommodates  1   7756058 551326862 29311
```

```
summary(price_imputing_mlr_model)
```

```
##
## Call:
## lm(formula = NumPrice ~ neighbourhood_cleansed + neighbourhood_group_cleansed +
##     room_type + accommodates + beds, data = zurich2_price_nonas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -508.7   -58.6   -21.8    15.8 19747.7
##
## Coefficients: (11 not defined because of singularities)
##                                            Estimate Std. Error t value
## (Intercept)                                  30.814     65.694   0.469
## neighbourhood_cleansedAlbisrieden           -14.065     93.498  -0.150
## neighbourhood_cleansedAlt-Wiedikon           32.276     74.445   0.434
## neighbourhood_cleansedAltstetten            155.730     73.898   2.107
## neighbourhood_cleansedCity                  420.080    147.525   2.848
## neighbourhood_cleansedEnge                  162.771     79.564   2.046
## neighbourhood_cleansedEscher Wyss            68.964     97.372   0.708
## neighbourhood_cleansedFluntern               51.915    103.277   0.503
## neighbourhood_cleansedFriesenberg            48.414     92.513   0.523
## neighbourhood_cleansedGewerbeschule          48.134     82.806   0.581
## neighbourhood_cleansedHard                   22.705     81.541   0.278
## neighbourhood_cleansedHirslanden             36.788     91.978   0.400
## neighbourhood_cleansedHirzenbach             29.681    123.827   0.240
## neighbourhood_cleansedHochschulen            57.704    109.939   0.525
## neighbourhood_cleansedHöngg                  14.547     85.787   0.170
## neighbourhood_cleansedHottingen              51.964     81.159   0.640
## neighbourhood_cleansedLangstrasse            60.276     71.918   0.838
## neighbourhood_cleansedLeimbach               11.163    163.490   0.068
## neighbourhood_cleansedLindenhof             448.859    105.073   4.272
## neighbourhood_cleansedMühlebach              44.172     80.796   0.547
## neighbourhood_cleansedOberstrass             65.644     85.403   0.769
## neighbourhood_cleansedOerlikon                8.975     74.708   0.120
## neighbourhood_cleansedRathaus                76.059     76.222   0.998
## neighbourhood_cleansedSaatlen                26.971    222.601   0.121
## neighbourhood_cleansedSchwamendingen-Mitte   16.539    151.527   0.109
## neighbourhood_cleansedSeebach                16.062     86.445   0.186
## neighbourhood_cleansedSeefeld               101.083     82.486   1.225
## neighbourhood_cleansedSihlfeld               23.419     73.139   0.320
## neighbourhood_cleansedUnterstrass            50.952     78.414   0.650
## neighbourhood_cleansedWeinegg                40.207     89.724   0.448
## neighbourhood_cleansedWerd                   51.340     93.024   0.552
## neighbourhood_cleansedWipkingen              41.327     85.723   0.482
## neighbourhood_cleansedWitikon                12.330     96.295   0.128
## neighbourhood_cleansedWollishofen            49.326     81.044   0.609
## neighbourhood_group_cleansedKreis 10             NA         NA      NA
## neighbourhood_group_cleansedKreis 11             NA         NA      NA
## neighbourhood_group_cleansedKreis 12             NA         NA      NA
## neighbourhood_group_cleansedKreis 2              NA         NA      NA
## neighbourhood_group_cleansedKreis 3              NA         NA      NA
## neighbourhood_group_cleansedKreis 4              NA         NA      NA
## neighbourhood_group_cleansedKreis 5              NA         NA      NA
```

```
## neighbourhood_group_cleansedKreis 6               NA          NA       NA
## neighbourhood_group_cleansedKreis 7               NA          NA       NA
## neighbourhood_group_cleansedKreis 8               NA          NA       NA
## neighbourhood_group_cleansedKreis 9               NA          NA       NA
## room_typeHotel room                           -2.524     166.120   -0.015
## room_typePrivate room                        -14.739      26.311   -0.560
## room_typeShared room                        -114.511     112.739   -1.016
## accommodates                                  30.492       9.264    3.292
## beds                                           4.778      11.102    0.430
##                                             Pr(>|t|)
## (Intercept)                                  0.63908
## neighbourhood_cleansedAlbisrieden            0.88044
## neighbourhood_cleansedAlt-Wiedikon           0.66465
## neighbourhood_cleansedAltstetten             0.03519 *
## neighbourhood_cleansedCity                   0.00444 **
## neighbourhood_cleansedEnge                   0.04089 *
## neighbourhood_cleansedEscher Wyss            0.47886
## neighbourhood_cleansedFluntern               0.61524
## neighbourhood_cleansedFriesenberg            0.60080
## neighbourhood_cleansedGewerbeschule          0.56110
## neighbourhood_cleansedHard                   0.78070
## neighbourhood_cleansedHirslanden             0.68922
## neighbourhood_cleansedHirzenbach             0.81058
## neighbourhood_cleansedHochschulen            0.59973
## neighbourhood_cleansedHöngg                  0.86536
## neighbourhood_cleansedHottingen              0.52206
## neighbourhood_cleansedLangstrasse            0.40205
## neighbourhood_cleansedLeimbach               0.94557
## neighbourhood_cleansedLindenhof            0.0000202 ***
## neighbourhood_cleansedMühlebach              0.58463
## neighbourhood_cleansedOberstrass             0.44219
## neighbourhood_cleansedOerlikon               0.90439
## neighbourhood_cleansedRathaus                0.31845
## neighbourhood_cleansedSaatlen                0.90357
## neighbourhood_cleansedSchwamendingen-Mitte   0.91309
## neighbourhood_cleansedSeebach                0.85261
## neighbourhood_cleansedSeefeld                0.22053
## neighbourhood_cleansedSihlfeld               0.74885
## neighbourhood_cleansedUnterstrass            0.51590
## neighbourhood_cleansedWeinegg                0.65411
## neighbourhood_cleansedWerd                   0.58107
## neighbourhood_cleansedWipkingen              0.62978
## neighbourhood_cleansedWitikon                0.89812
## neighbourhood_cleansedWollishofen            0.54283
## neighbourhood_group_cleansedKreis 10              NA
## neighbourhood_group_cleansedKreis 11              NA
## neighbourhood_group_cleansedKreis 12              NA
## neighbourhood_group_cleansedKreis 2               NA
## neighbourhood_group_cleansedKreis 3               NA
## neighbourhood_group_cleansedKreis 4               NA
## neighbourhood_group_cleansedKreis 5               NA
## neighbourhood_group_cleansedKreis 6               NA
```

```
## neighbourhood_group_cleansedKreis 7                NA
## neighbourhood_group_cleansedKreis 8                NA
## neighbourhood_group_cleansedKreis 9                NA
## room_typeHotel room                           0.98788
## room_typePrivate room                         0.57541
## room_typeShared room                          0.30987
## accommodates                                  0.00101 **
## beds                                          0.66696
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 477.7 on 2333 degrees of freedom
## Multiple R-squared:  0.03442,    Adjusted R-squared:  0.01869
## F-statistic: 2.188 on 38 and 2333 DF,  p-value: 0.00004105
```

```
impute_price_preds <- predict(price_imputing_mlr_model,zurich2_price_nas)
```

Another way

```
# Impute Price
zurich_test <- zurich_FE %>% group_by(neighbourhood_group_cleansed,neighbourhood_cleansed, prope
rty_type,room_type, beds) %>% arrange(neighbourhood_group_cleansed,neighbourhood_cleansed, prope
rty_type,room_type, beds)

zurich_imputed <- zurich_test %>% mutate(NumPrice= ifelse(is.na(NumPrice), mean(NumPrice, na.rm=
TRUE), NumPrice))
colSums(is.na(zurich_imputed))
```

```
##                          id
##                           0
##                 listing_url
##                           0
##                   scrape_id
##                           0
##                last_scraped
##                           0
##                      source
##                           0
##                        name
##                           0
##       neighborhood_overview
##                        1526
##                 picture_url
##                           0
##                     host_id
##                           0
##                    host_url
##                           0
##                   host_name
##                           0
##                  host_since
##                           0
##               host_location
##                         503
##                  host_about
##                        1181
##          host_response_time
##                           0
##          host_response_rate
##                           0
##        host_acceptance_rate
##                           0
##           host_is_superhost
##                           2
##          host_thumbnail_url
##                           0
##            host_picture_url
##                           0
##          host_neighbourhood
##                        2791
##         host_listings_count
##                           0
##   host_total_listings_count
##                           0
##          host_verifications
##                           0
##        host_has_profile_pic
##                           0
##      host_identity_verified
##                           0
```

```
##                                   neighbourhood
##                                            1526
##                          neighbourhood_cleansed
##                                               0
##                    neighbourhood_group_cleansed
##                                               0
##                                        latitude
##                                               0
##                                       longitude
##                                               0
##                                   property_type
##                                               0
##                                       room_type
##                                               0
##                                    accommodates
##                                               0
##                                  bathrooms_text
##                                               1
##                                            beds
##                                               0
##                                           price
##                                             447
##                                  minimum_nights
##                                               0
##                                  maximum_nights
##                                               0
##                          minimum_minimum_nights
##                                               0
##                          maximum_minimum_nights
##                                               0
##                          minimum_maximum_nights
##                                               0
##                          maximum_maximum_nights
##                                               0
##                          minimum_nights_avg_ntm
##                                               0
##                          maximum_nights_avg_ntm
##                                               0
##                                had_availability
##                                             447
##                                 availability_30
##                                               0
##                                 availability_60
##                                               0
##                                 availability_90
##                                               0
##                                availability_365
##                                               0
##                           calendar_last_scraped
##                                               0
##                               number_of_reviews
##                                               0
```

```
##                        number_of_reviews_ltm
##                                            0
##                       number_of_reviews_l30d
##                                            0
##                                 first_review
##                                          731
##                                  last_review
##                                          731
##                           review_scores_value
##                                            0
##                             instant_bookable
##                                            0
##               calculated_host_listings_count
##                                            0
##  calculated_host_listings_count_entire_homes
##                                            0
## calculated_host_listings_count_private_rooms
##                                            0
##  calculated_host_listings_count_shared_rooms
##                                            0
##                            reviews_per_month
##                                          731
##                                 review_value
##                                            0
##                                review_rating
##                                            0
##                              review_accuracy
##                                            0
##                           review_cleanliness
##                                            0
##                               review_checkin
##                                            0
##                         review_communication
##                                            0
##                              review_location
##                                            0
##                                     NumPrice
##                                           43
##                                        baths
##                                            0
##                                 guestsPerBath
##                                            0
##                                  guestsPerBed
##                                            0
```

```
# Removing redundant price and bathrooms_text as new variales NumPrice and baths are created
zurich_imputed <- zurich_imputed[,-c(35,37)]
head(zurich_imputed,2)
```

id	listing_url	scrape
<dbl>	<chr>	<c
1013580000000000000	https://www.airbnb.com/rooms/1013584589601563807	20231200000
49201447	https://www.airbnb.com/rooms/49201447	20231200000

2 rows | 1-4 of 72 columns

Still we have 43 missing values for Price

```
test_values <- unique(zurich1$host_neighbourhood)
print(test_values)
```

```
##  [1] NA                      "El Gòtic"              "Leopoldstadt"
##  [4] "Roquebrune-Cap-Martin" "São Paulo"             "Vila Mariana"
##  [7] "Jacumã"                "Upper Sukhumvit"       "Leme"
## [10] "Wilmersdorf"           "Daille"                "Friedrichshain"
## [13] "Covent Garden"         "Klong Toey"            "South Kensington"
## [16] "Am Hart"               "Morningside Heights"   "Isle of Dogs"
## [19] "Indre By"              "Dreta de l'Eixample"   "Clapham Common"
## [22] "Wan Chai"              "Ipanema"               "Copacabana"
## [25] "Nation"
```

Seeing what values does host-neighborhood have

```
seerow<- zurich1[465,]
print(seerow)
```

```
## # A tibble: 1 × 70
##         id listing_url scrape_id last_scraped source name  neighborhood_overview
##      <dbl> <chr>           <dbl> <chr>        <chr>  <chr> <chr>
## 1 20621446 https://ww…   2.02e13 12/28/2023   city … Loft… Shops, cafes, bars, …
## # ℹ 63 more variables: picture_url <chr>, host_id <dbl>, host_url <chr>,
## #   host_name <chr>, host_since <chr>, host_location <chr>, host_about <chr>,
## #   host_response_time <chr>, host_response_rate <chr>,
## #   host_acceptance_rate <chr>, host_is_superhost <lgl>,
## #   host_thumbnail_url <chr>, host_picture_url <chr>, host_neighbourhood <chr>,
## #   host_listings_count <dbl>, host_total_listings_count <dbl>,
## #   host_verifications <chr>, host_has_profile_pic <lgl>, …
```

To copy

```
cleansed<- unique(zurich1$neighbourhood_cleansed)
grp_cleansed <- unique(zurich1$neighbourhood_group_cleansed)
print(cleansed)
```

```
##  [1] "Sihlfeld"             "Enge"                 "Höngg"
##  [4] "Wollishofen"          "Escher Wyss"          "Wipkingen"
##  [7] "Lindenhof"            "Rathaus"              "Hard"
## [10] "Hochschulen"          "Oerlikon"             "Werd"
## [13] "Alt-Wiedikon"         "Friesenberg"          "Seebach"
## [16] "Schwamendingen-Mitte" "Gewerbeschule"        "Langstrasse"
## [19] "Mühlebach"            "Unterstrass"          "Hirzenbach"
## [22] "Weinegg"              "Fluntern"             "Hottingen"
## [25] "Altstetten"           "Hirslanden"           "Oberstrass"
## [28] "Seefeld"              "Witikon"              "Affoltern"
## [31] "City"                 "Saatlen"              "Albisrieden"
## [34] "Leimbach"
```

```
print(grp_cleansed)
```

```
##  [1] "Kreis 3"  "Kreis 2"  "Kreis 10" "Kreis 5"  "Kreis 1"  "Kreis 4"
##  [7] "Kreis 11" "Kreis 12" "Kreis 8"  "Kreis 6"  "Kreis 7"  "Kreis 9"
```

The above neighborhoods and circles are in Zurich

```
zurich_baths <- zurich2 %>% filter(is.na(baths))

zurich2$baths <- as.numeric(gsub("(half).*", "0.5", zurich2$baths, ignore.case = TRUE))
```

##NAIVE BAYES

```
zurich_naive <- zurich_imputed

zurich_naive <- zurich_naive %>%
  select(c(host_response_time, host_is_superhost, host_has_profile_pic, host_identity_verified,
property_type, accommodates, number_of_reviews, review_rating, NumPrice, guestsPerBath, guestsPe
rBed))%>%
  mutate(across(where(is.character), as.factor))
```

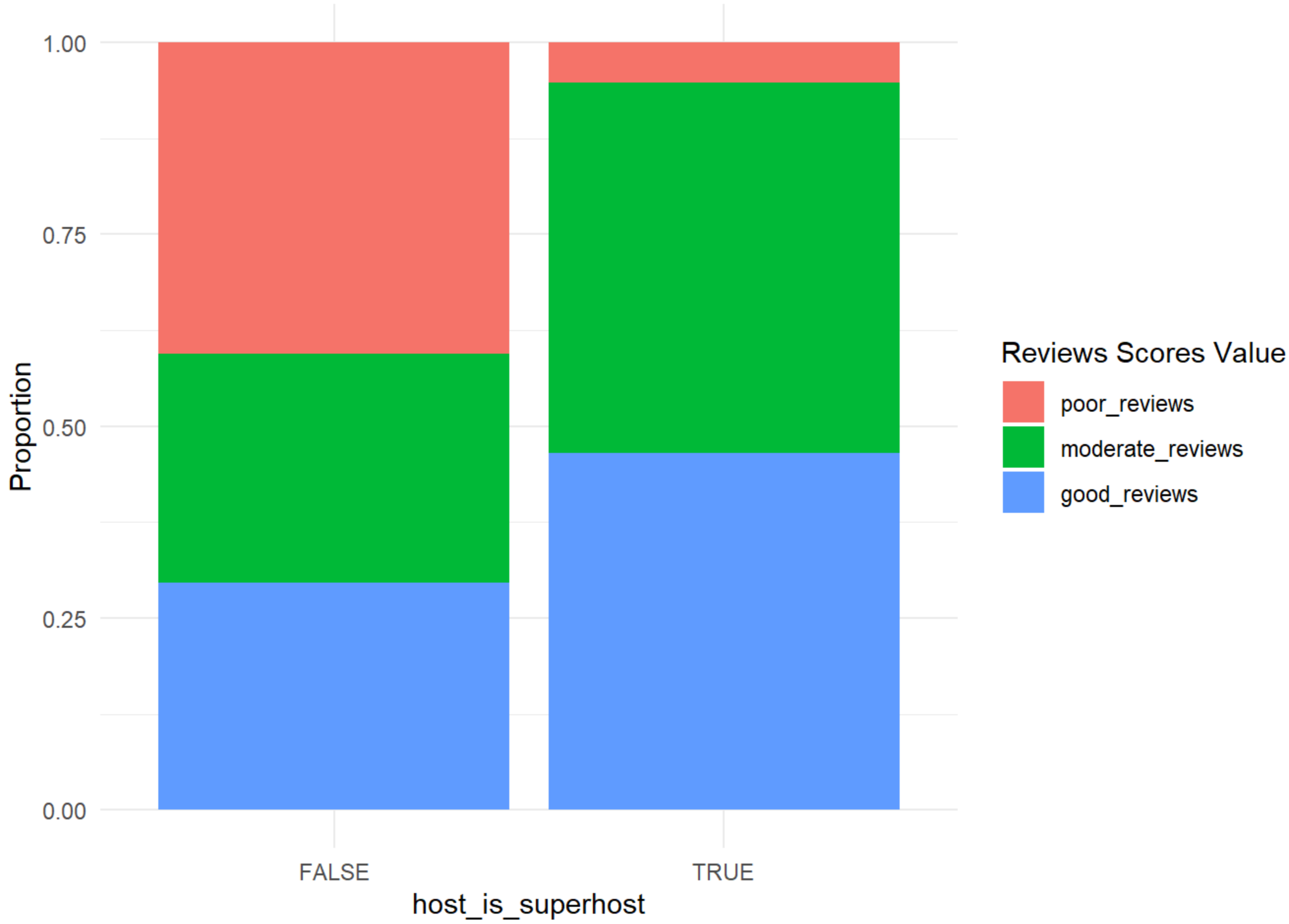
```
## Adding missing grouping variables: `neighbourhood_group_cleansed`,
## `neighbourhood_cleansed`, `room_type`, `beds`
```

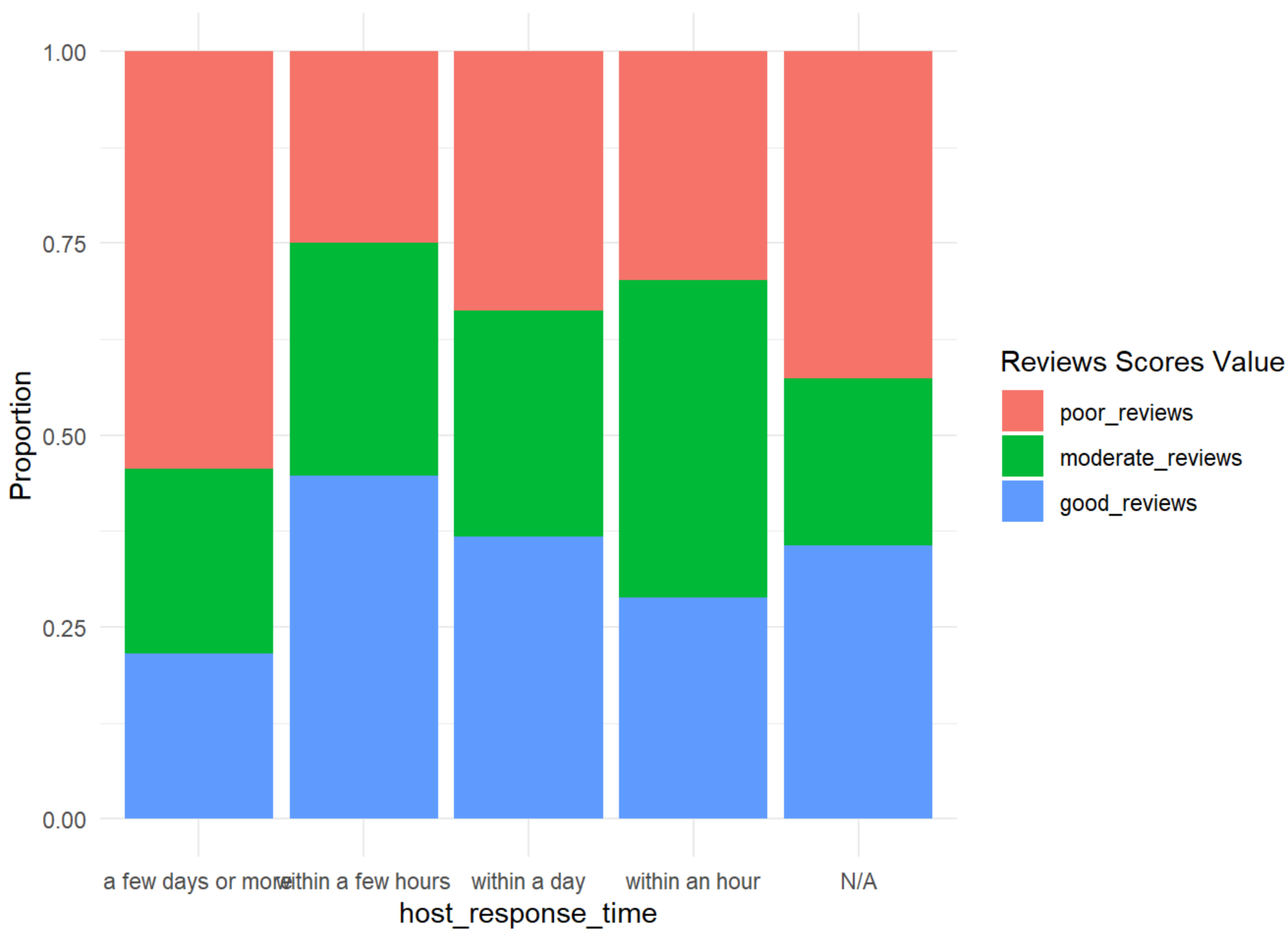
```
zurich_naive$host_is_superhost[c(593, 1299)] <- "FALSE"

quantile_edges <- quantile(zurich_naive$NumPrice, probs = c(0, 1/3, 2/3, 1), na.rm = TRUE)
zurich_naive$price_category <- cut(zurich_naive$NumPrice,
                                   breaks = quantile_edges,
                                   labels = c("Low", "Medium", "High"),
                                   include.lowest = TRUE)
zurich_naive <- zurich_naive %>%
  select(-c(NumPrice))

zurich_naive <- na.omit(zurich_naive, subset = c("price_category"))

q_accomodates <- quantile(zurich_naive$accommodates, probs = c(0, 1/3, 2/3, 1), na.rm = TRUE)
zurich_naive$accommodates_group <- cut(zurich_naive$accommodates,
                                   breaks = q_accomodates,
                                   labels = c("Low", "Medium", "High"),
                                   include.lowest = TRUE)
zurich_naive <- zurich_naive %>%
  select(-c(accommodates))

q_beds <- c(min(zurich_naive$guestsPerBed), 1.5, 3, max(zurich_naive$guestsPerBed))
zurich_naive$beds_group <- cut(zurich_naive$guestsPerBed,
                                   breaks = q_beds,
                                   labels = c("Low", "Medium", "High"),
                                   include.lowest = TRUE)
zurich_naive <- zurich_naive %>%
  select(-c(guestsPerBed))

q_number_reviews <- quantile(zurich_naive$number_of_reviews, probs = c(0, 1/3, 2/3, 1), na.rm =
TRUE)
zurich_naive$number_reviews_group <- cut(zurich_naive$number_of_reviews,
                                   breaks = q_number_reviews,
                                   labels = c("Low", "Medium", "High"),
                                   include.lowest = TRUE)
zurich_naive <- zurich_naive %>%
  select(-c(number_of_reviews))

q_baths <- quantile(zurich_naive$guestsPerBath, probs = c(0, 1/3, 2/3, 1), na.rm = TRUE)
zurich_naive$baths_group <- cut(zurich_naive$guestsPerBath,
                                   breaks = q_baths,
                                   labels = c("Low", "Medium", "High"),
                                   include.lowest = TRUE)
zurich_naive <- zurich_naive %>%
  select(-c(guestsPerBath))
```

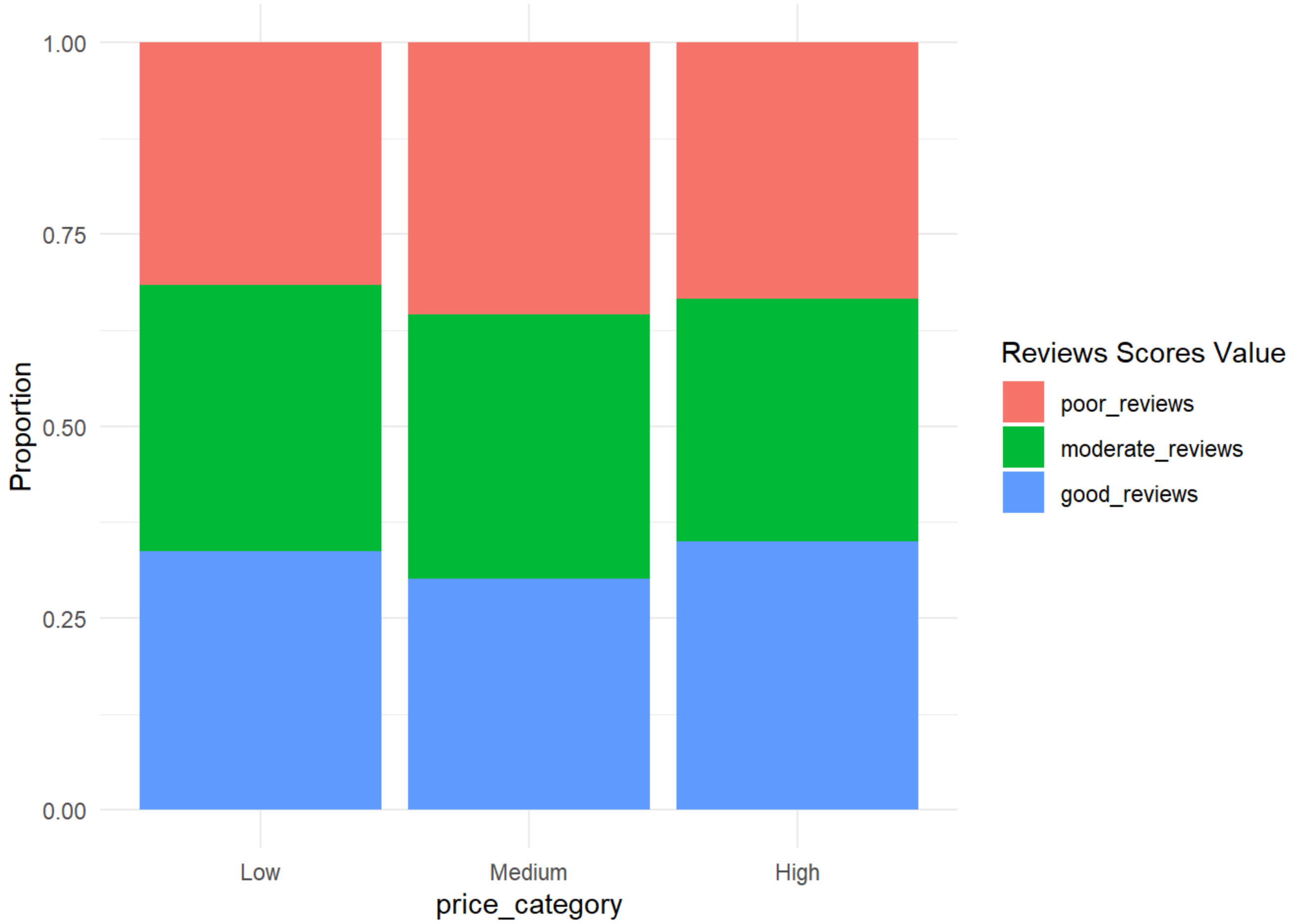
Barplot

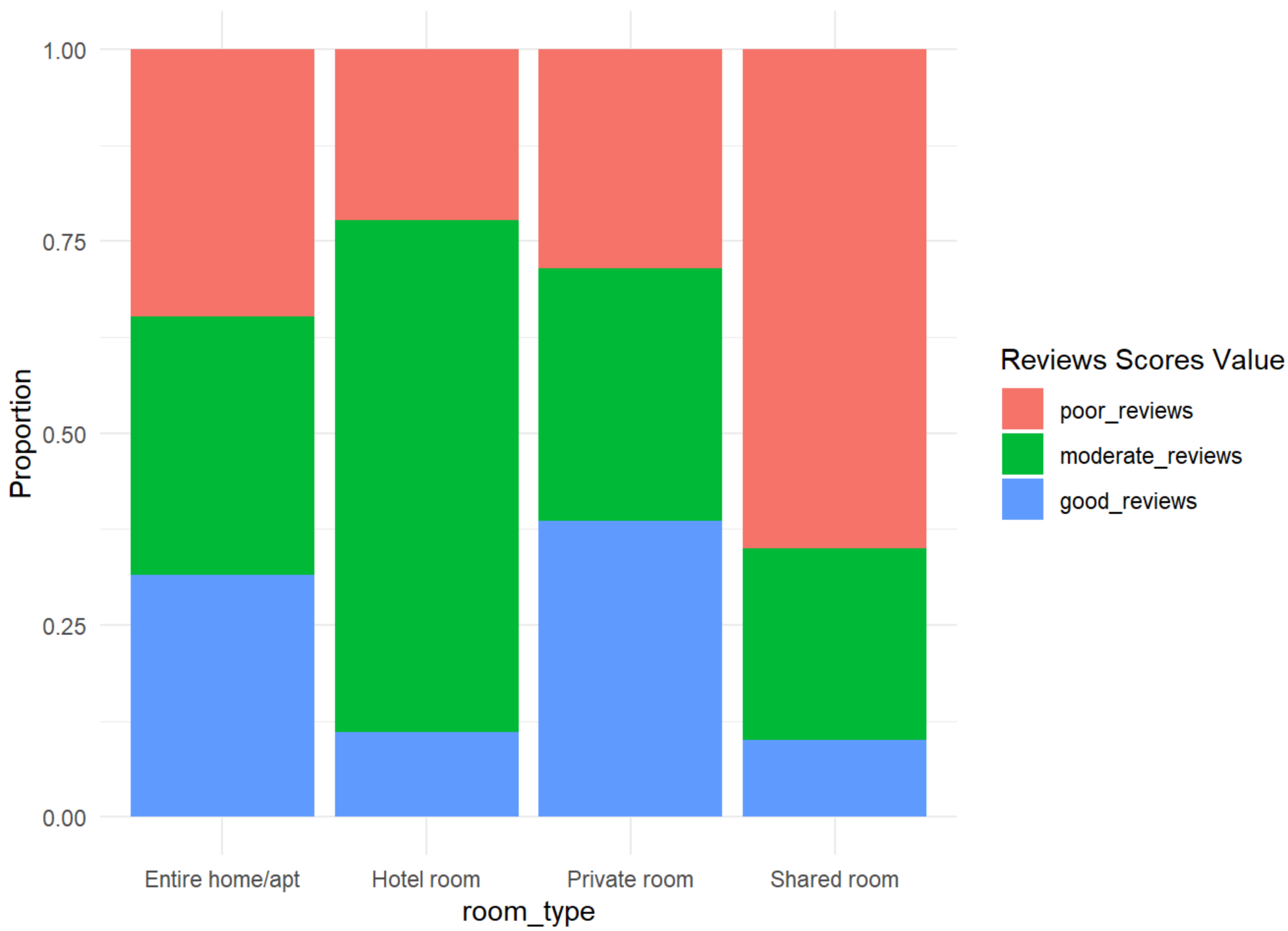
```
categorical_vars <- c("host_is_superhost", "host_response_time", "price_category", "room_type",
"host_has_profile_pic", "neighbourhood_group_cleansed","neighbourhood_cleansed", "room_type", "b
eds", "host_identity_verified", "property_type", "accommodates_group", "beds_group", "number_rev
iews_group", "baths_group")
plot_list <- list()

for (var in categorical_vars) {
  zurich_naive[[var]] <- as.factor(zurich_naive[[var]])
  zurich_naive[["review_rating"]] <- as.factor(zurich_naive[["review_rating"]])


  plot_list[[var]] <- ggplot(zurich_naive, aes_string(x = var, fill = "review_rating")) +
    geom_bar(position = "fill") +
    labs(y = "Proportion", x = var, fill = "Reviews Scores Value") +
    theme_minimal()

  print(plot_list[[var]])
}
```

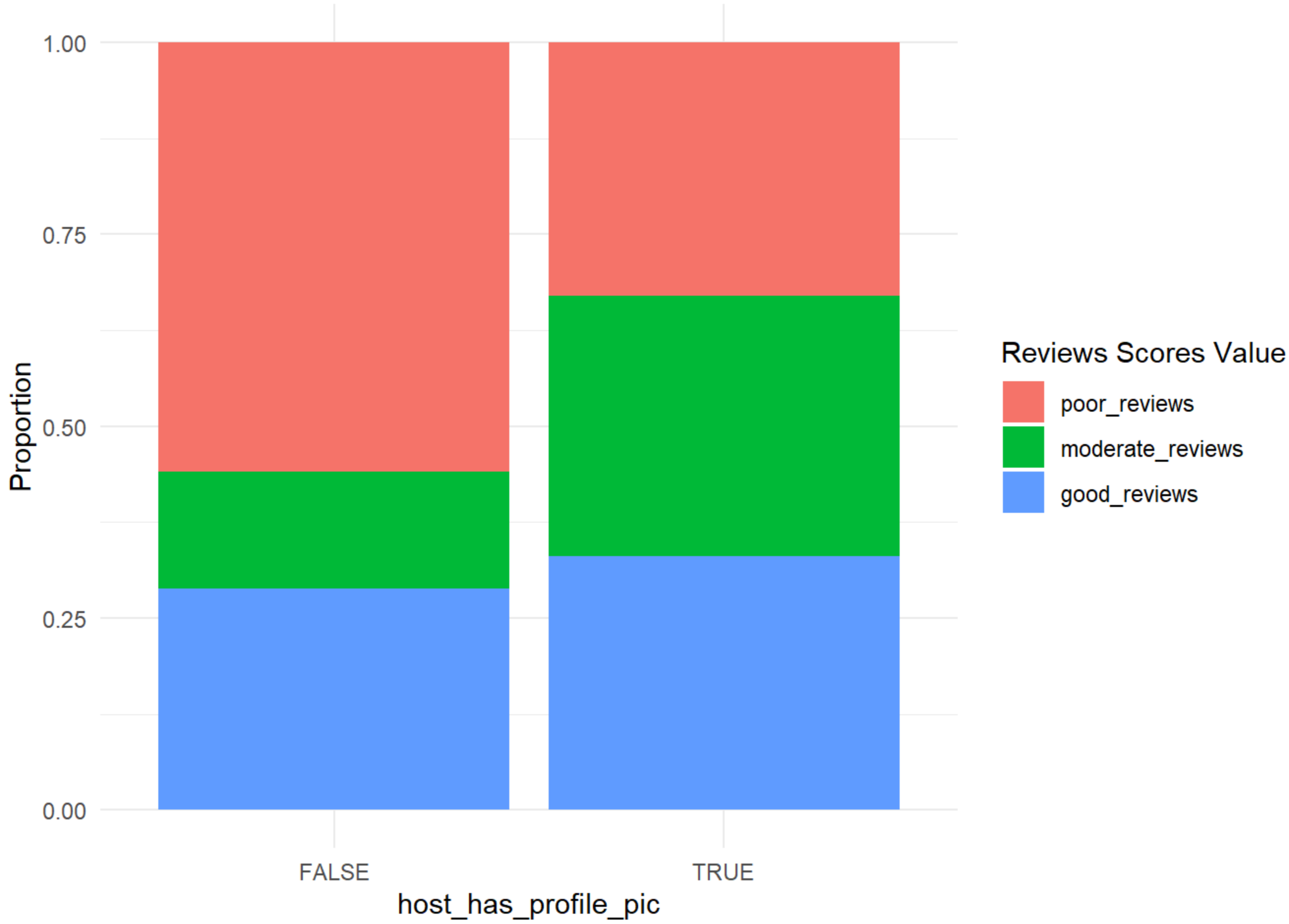
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## ℹ Please use tidy evaluation idioms with `aes()`.
## ℹ See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

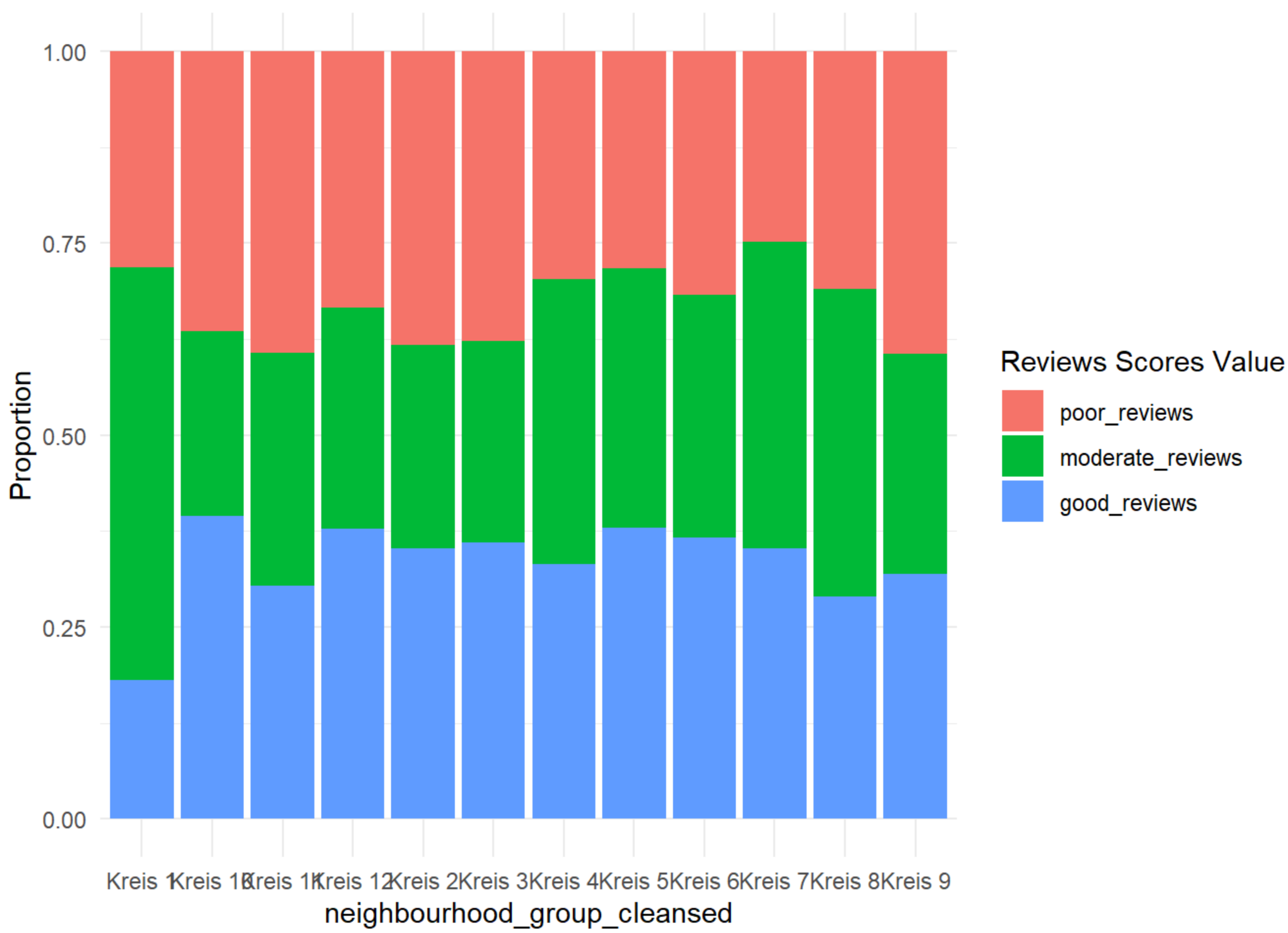

Proportion
1.00
0.75
0.50
0.25
0.00
FALSE
TRUE
host_is_superhost
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews

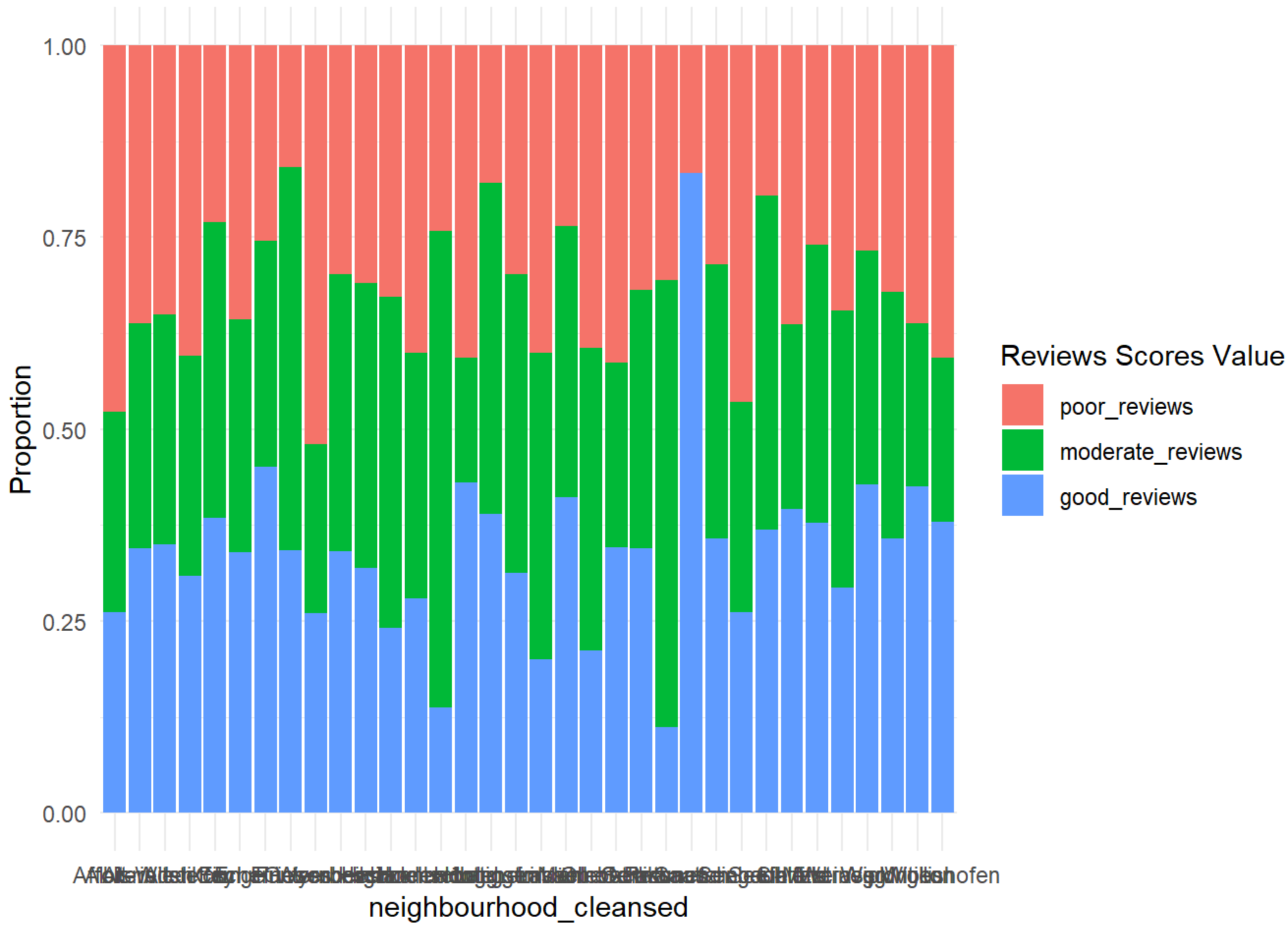

Proportion
1.00
0.75
0.50
0.25
0.00
a few days or more
within a few hours
within a day
within an hour
N/A
host_response_time
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews

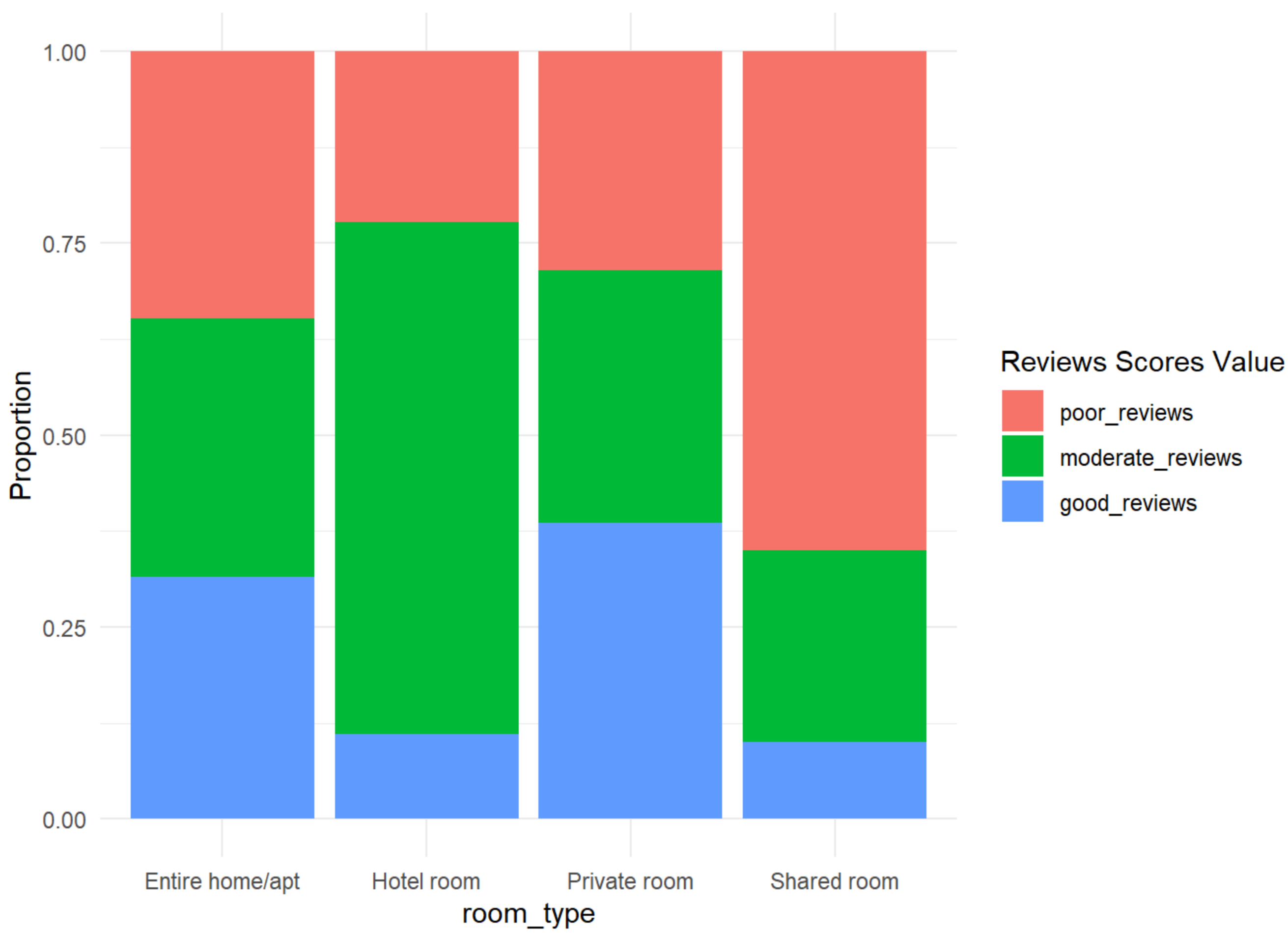

1.00
0.75
0.50
0.25
0.00
Proportion
Low
Medium
High
price_category
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews

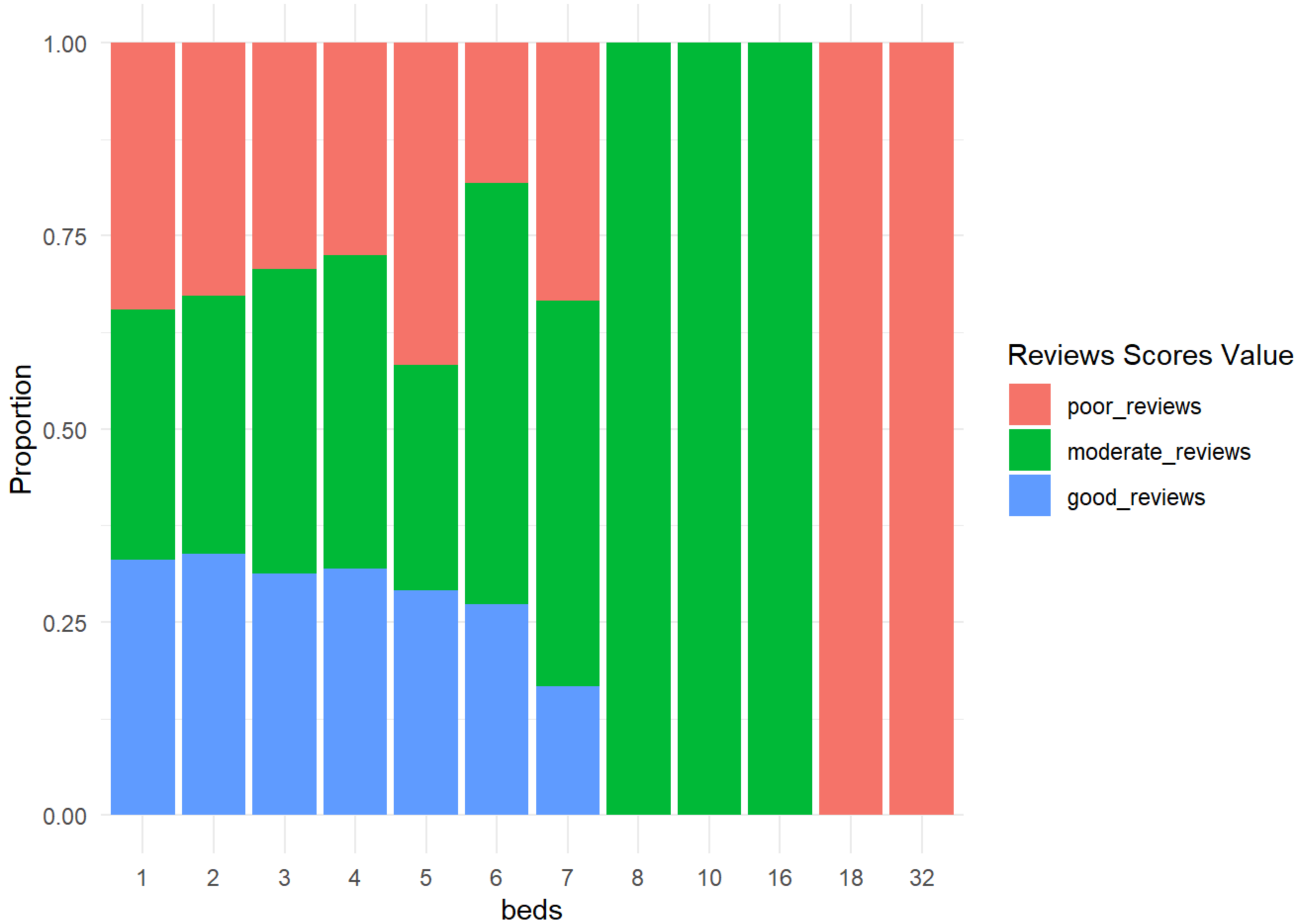

1.00
0.75
0.50
0.25
0.00
Proportion
Entire home/apt
Hotel room
Private room
Shared room
room_type
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews

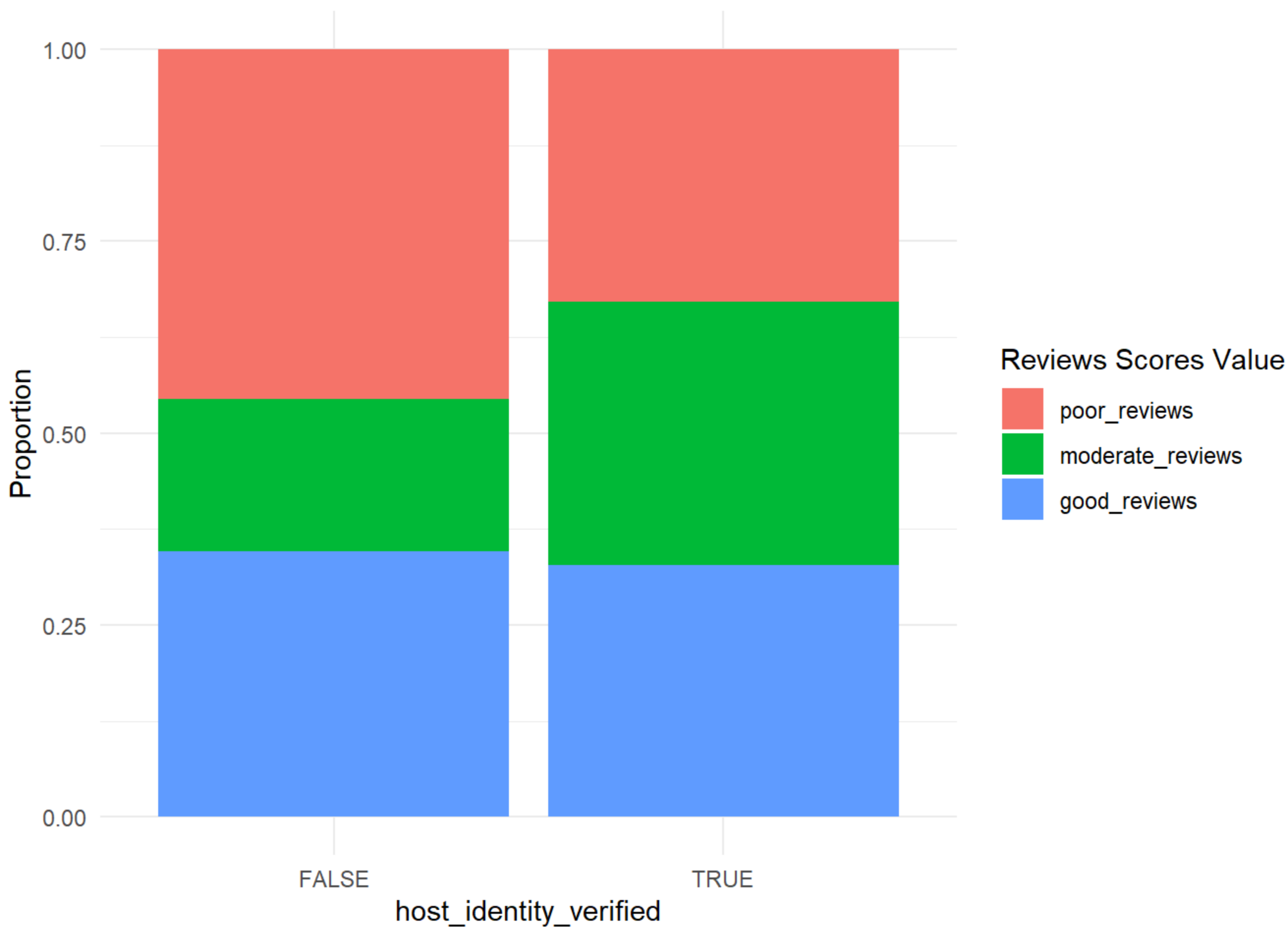

1.00
0.75
0.50
0.25
0.00
Proportion
FALSE
TRUE
host_has_profile_pic
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews

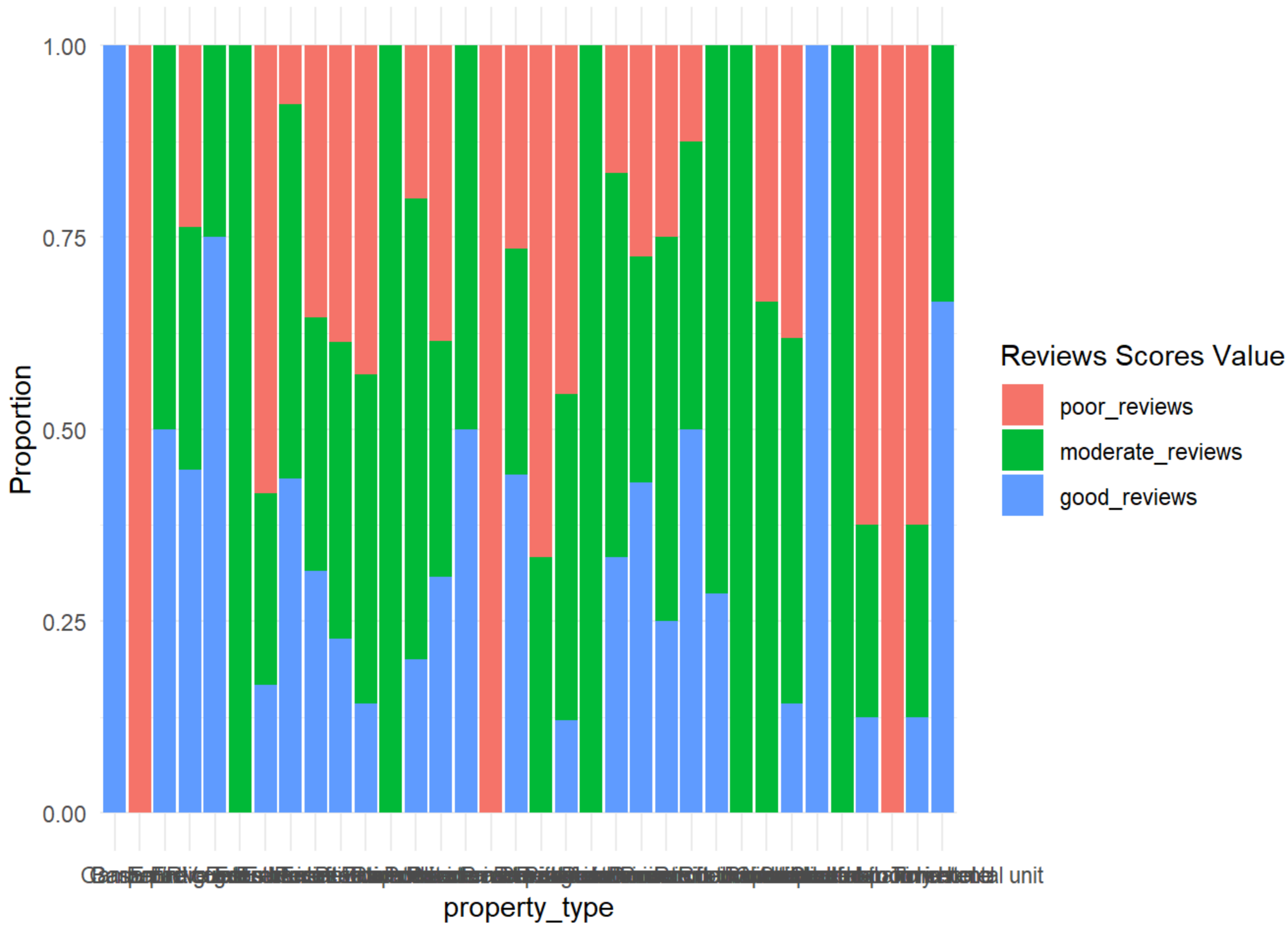

1.00
0.75
0.50
0.25
0.00
Proportion
Kreis 1
Kreis 10
Kreis 11
Kreis 12
Kreis 2
Kreis 3
Kreis 4
Kreis 5
Kreis 6
Kreis 7
Kreis 8
Kreis 9
neighbourhood_group_cleansed
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews

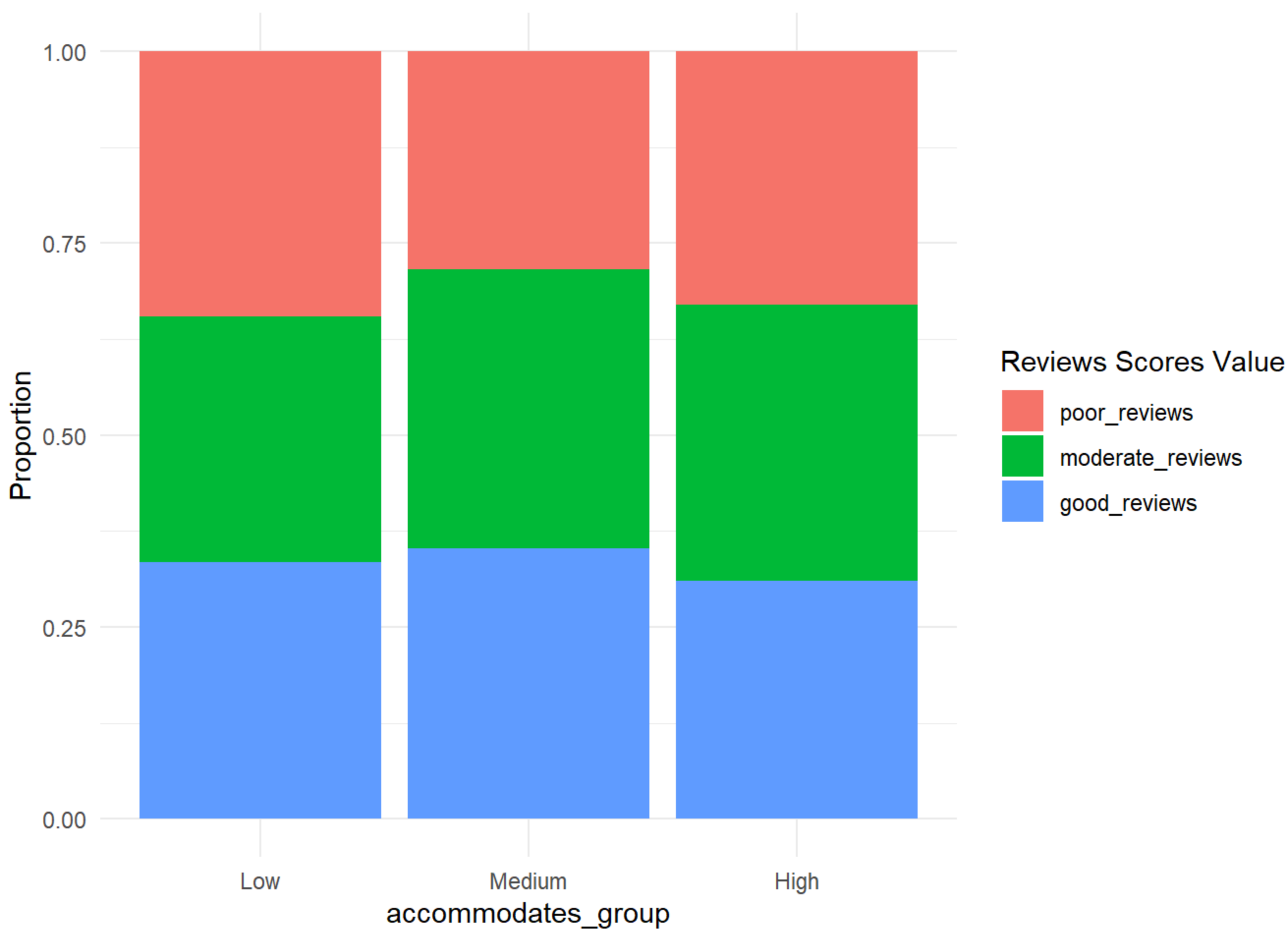

1.00
0.75
0.50
0.25
0.00
Proportion
neighbourhood_cleansed
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews

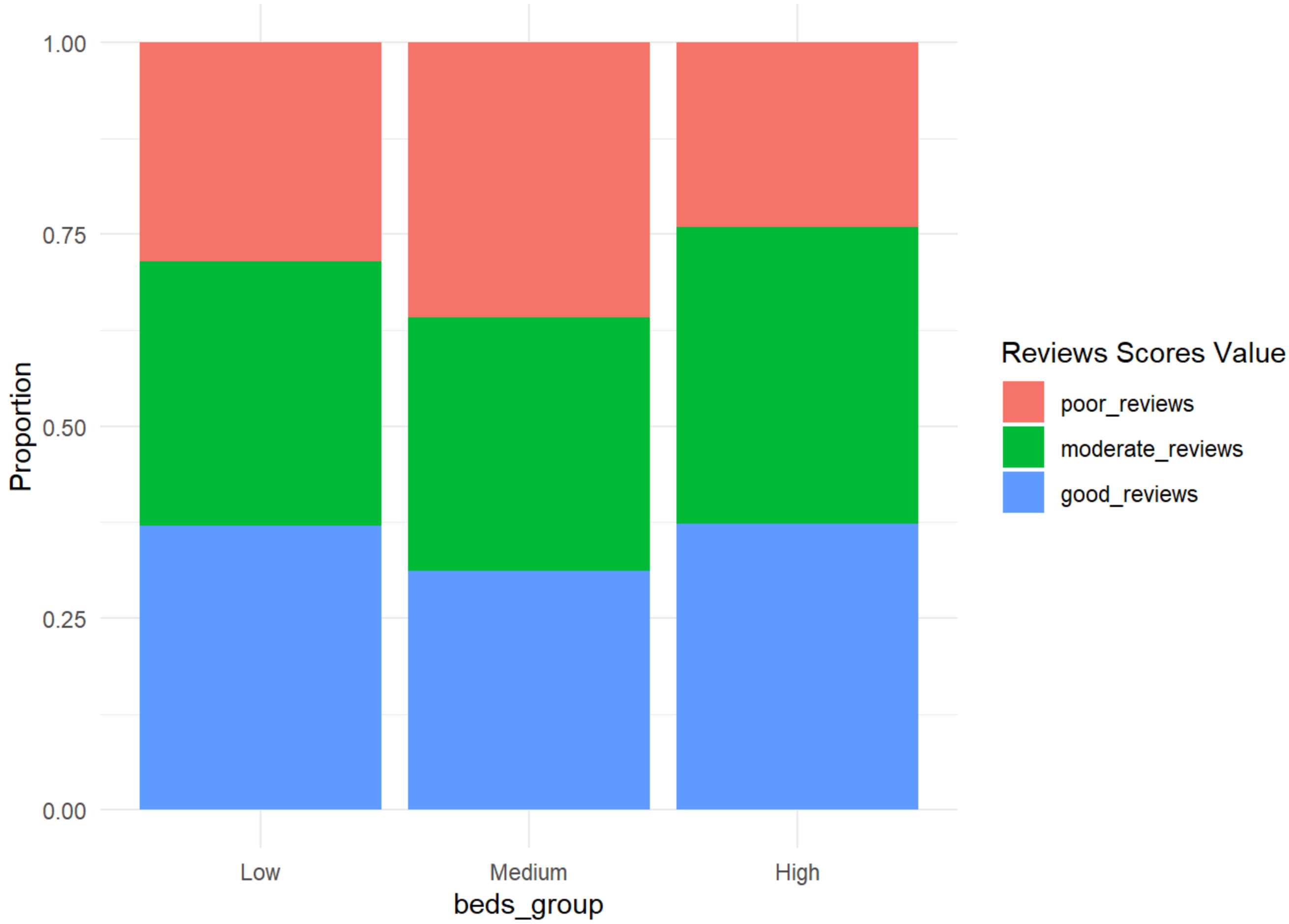

1.00
0.75
0.50
0.25
0.00
Proportion
Entire home/apt
Hotel room
Private room
Shared room
room_type
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews

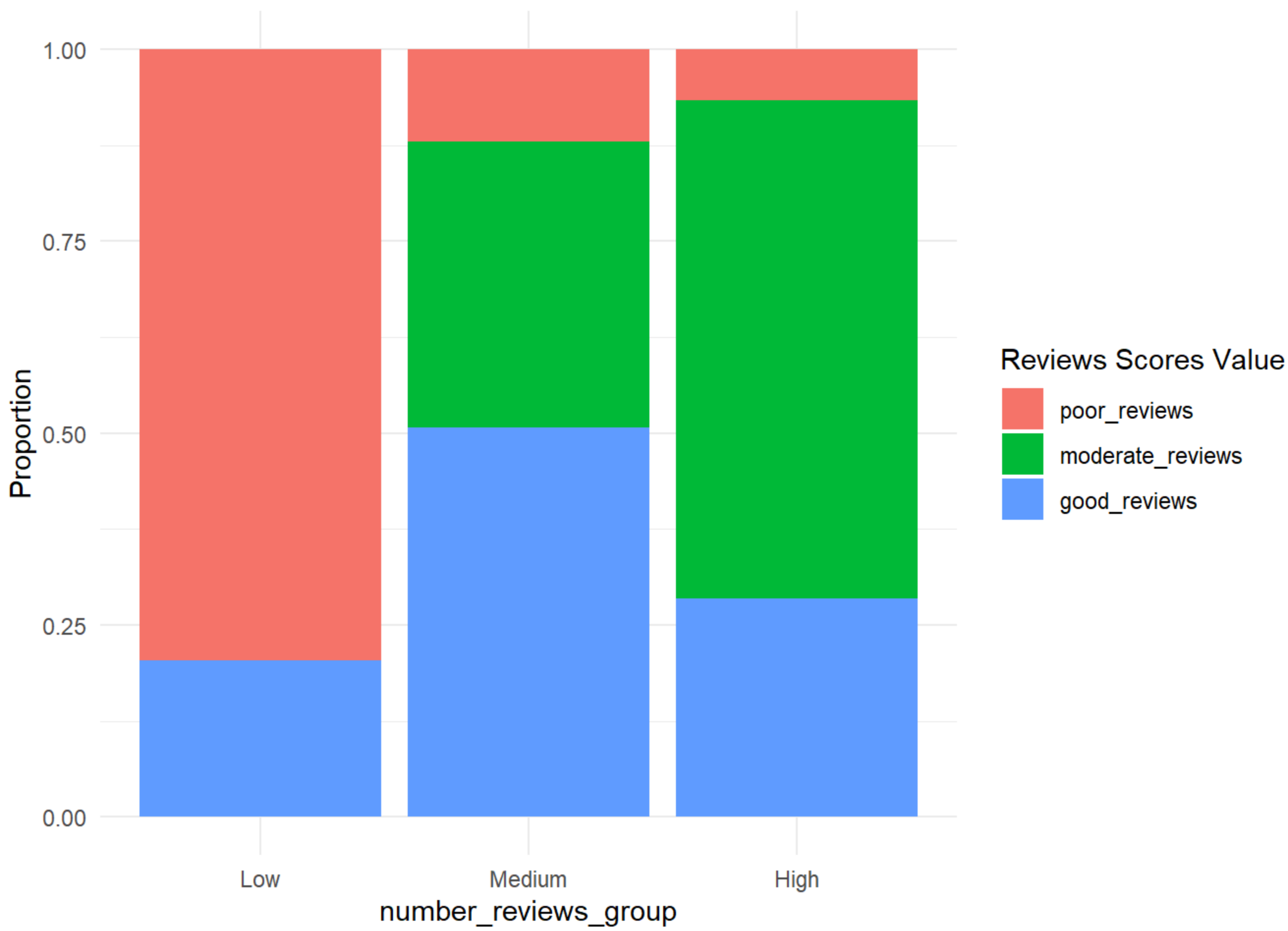

1.00
0.75
0.50
0.25
0.00
Proportion
1
2
3
4
5
6
7
8
10
16
18
32
beds
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews

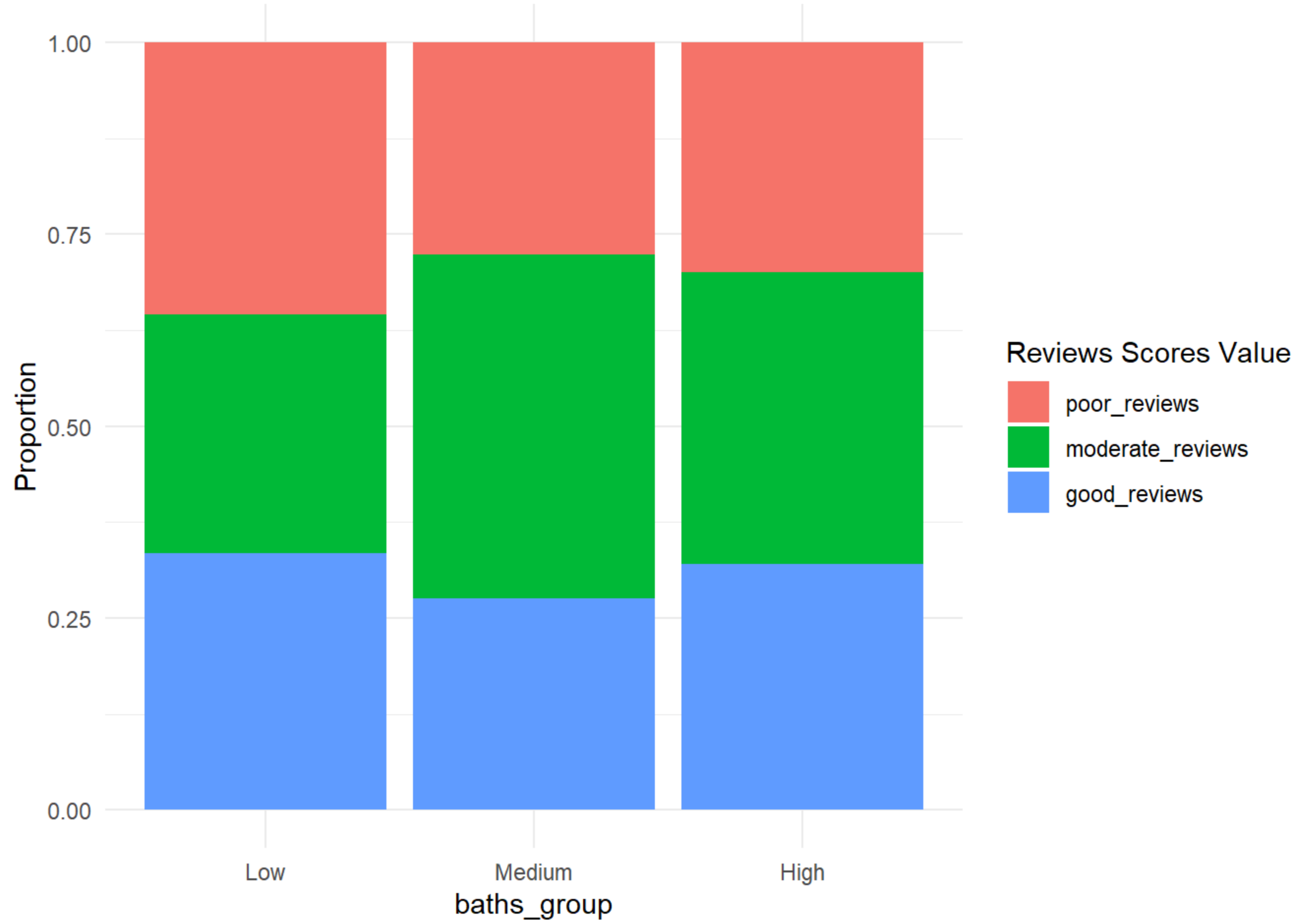

1.00
0.75
0.50
0.25
0.00
Proportion
FALSE
TRUE
host_identity_verified
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews


Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews
Proportion
1.00
0.75
0.50
0.25
0.00
property_type


Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews
Proportion
1.00
0.75
0.50
0.25
0.00
Low
Medium
High
accommodates_group


1.00
0.75
0.50
0.25
0.00
Proportion
Low
Medium
High
beds_group
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews


1.00
0.75
0.50
0.25
0.00
Proportion
Low
Medium
High
number_reviews_group
Reviews Scores Value
poor_reviews
moderate_reviews
good_reviews



Naive Model

```
zurich_naive <- zurich_naive %>%
  select(-c(price_category, accommodates_group, beds_group, baths_group, host_identity_verifie
d))

library(e1071)
zurich_naive <- zurich_naive %>%
  mutate(across(where(is.character), as.factor))

set.seed(70)
idx <- createDataPartition(zurich_naive$review_rating, p=0.6, list=FALSE)
training_set <- zurich_naive[idx,]
validation_set <- zurich_naive[-idx,]

nb_model <- naiveBayes(review_rating ~., data = training_set)
nb_model
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     poor_reviews moderate_reviews     good_reviews
##        0.3349340        0.3355342        0.3295318
##
## Conditional probabilities:
##                   neighbourhood_group_cleansed
## Y                     Kreis 1   Kreis 10   Kreis 11   Kreis 12    Kreis 2
##   poor_reviews     0.05913978 0.08243728 0.12365591 0.02329749 0.09139785
##   moderate_reviews 0.12701252 0.04293381 0.09481216 0.01610018 0.06261181
##   good_reviews     0.04553734 0.07285974 0.10382514 0.01639344 0.07650273
##                   neighbourhood_group_cleansed
## Y                     Kreis 3    Kreis 4    Kreis 5    Kreis 6    Kreis 7
##   poor_reviews     0.14516129 0.11290323 0.04480287 0.07706093 0.05913978
##   moderate_reviews 0.11091234 0.14132379 0.05187835 0.07513417 0.09481216
##   good_reviews     0.15482696 0.14389800 0.06010929 0.07468124 0.09471767
##                   neighbourhood_group_cleansed
## Y                     Kreis 8    Kreis 9
##   poor_reviews     0.07526882 0.10573477
##   moderate_reviews 0.10733453 0.07513417
##   good_reviews     0.07832423 0.07832423
##
##                   neighbourhood_cleansed
## Y                    Affoltern Albisrieden Alt-Wiedikon  Altstetten        City
##   poor_reviews     0.037634409 0.026881720  0.050179211 0.078853047 0.003584229
##   moderate_reviews 0.017889088 0.017889088  0.055456172 0.057245081 0.005366726
##   good_reviews     0.020036430 0.025500911  0.065573770 0.052823315 0.005464481
##                   neighbourhood_cleansed
## Y                        Enge Escher Wyss    Fluntern Friesenberg
##   poor_reviews     0.044802867 0.010752688 0.007168459 0.026881720
##   moderate_reviews 0.032200358 0.010733453 0.019677996 0.012522361
##   good_reviews     0.040072860 0.027322404 0.012750455 0.018214936
##                   neighbourhood_cleansed
## Y                  Gewerbeschule        Hard  Hirslanden  Hirzenbach
##   poor_reviews       0.034050179 0.028673835 0.019713262 0.016129032
##   moderate_reviews   0.041144902 0.035778175 0.023255814 0.010733453
##   good_reviews       0.032786885 0.032786885 0.014571949 0.009107468
##                   neighbourhood_cleansed
## Y                  Hochschulen       Höngg   Hottingen Langstrasse    Leimbach
##   poor_reviews     0.005376344 0.048387097 0.019713262 0.068100358 0.003584229
##   moderate_reviews 0.023255814 0.012522361 0.041144902 0.082289803 0.003577818
##   good_reviews     0.007285974 0.043715847 0.047358834 0.078324226 0.001821494
##                   neighbourhood_cleansed
## Y                    Lindenhof   Mühlebach  Oberstrass    Oerlikon     Rathaus
##   poor_reviews     0.010752688 0.032258065 0.034050179 0.048387097 0.039426523
##   moderate_reviews 0.010733453 0.041144902 0.017889088 0.053667263 0.087656530
```

```
##   good_reviews     0.018214936 0.023679417 0.027322404 0.061930783 0.014571949
##                   neighbourhood_cleansed
## Y                      Saatlen Schwamendingen-Mitte     Seebach     Seefeld
##   poor_reviews     0.001792115          0.005376344 0.037634409 0.016129032
##   moderate_reviews 0.000000000          0.005366726 0.023255814 0.046511628
##   good_reviews     0.003642987          0.003642987 0.021857923 0.038251366
##                   neighbourhood_cleansed
## Y                     Sihlfeld Unterstrass     Weinegg        Werd   Wipkingen
##   poor_reviews     0.068100358 0.043010753 0.026881720 0.016129032 0.034050179
##   moderate_reviews 0.042933810 0.057245081 0.019677996 0.023255814 0.030411449
##   good_reviews     0.071038251 0.047358834 0.016393443 0.032786885 0.029143898
##                   neighbourhood_cleansed
## Y                      Witikon Wollishofen
##   poor_reviews     0.012544803 0.043010753
##   moderate_reviews 0.010733453 0.026833631
##   good_reviews     0.020036430 0.034608379
## 
##                   room_type
## Y                  Entire home/apt  Hotel room Private room Shared room
##   poor_reviews         0.784946237 0.003584229  0.197132616 0.014336918
##   moderate_reviews     0.765652952 0.000000000  0.228980322 0.005366726
##   good_reviews         0.715846995 0.001821494  0.282331512 0.000000000
## 
##                   beds
## Y                            1           2           3           4           5
##   poor_reviews     0.654121864 0.243727599 0.066308244 0.021505376 0.008960573
##   moderate_reviews 0.613595707 0.248658318 0.078711986 0.042933810 0.005366726
##   good_reviews     0.612021858 0.269581056 0.071038251 0.036429872 0.007285974
##                   beds
## Y                            6           7           8          10          16
##   poor_reviews     0.001792115 0.001792115 0.000000000 0.000000000 0.000000000
##   moderate_reviews 0.005366726 0.001788909 0.001788909 0.000000000 0.001788909
##   good_reviews     0.003642987 0.000000000 0.000000000 0.000000000 0.000000000
##                   beds
## Y                           18          32
##   poor_reviews     0.000000000 0.001792115
##   moderate_reviews 0.000000000 0.000000000
##   good_reviews     0.000000000 0.000000000
## 
##                   host_response_time
## Y                  a few days or more within a few hours within a day
##   poor_reviews             0.04301075         0.07706093   0.11290323
##   moderate_reviews         0.02146691         0.08944544   0.10017889
##   good_reviews             0.02367942         0.16575592   0.12021858
##                   host_response_time
## Y                  within an hour        N/A
##   poor_reviews         0.44265233 0.32437276
##   moderate_reviews     0.63148479 0.15742397
##   good_reviews         0.45719490 0.23315118
## 
##                   host_is_superhost
## Y                       FALSE       TRUE
```

```
##   poor_reviews     0.96236559 0.03763441
##   moderate_reviews 0.71914132 0.28085868
##   good_reviews     0.69763206 0.30236794
## 
##                 host_has_profile_pic
## Y                      FALSE        TRUE
##   poor_reviews     0.026881720 0.973118280
##   moderate_reviews 0.007155635 0.992844365
##   good_reviews     0.014571949 0.985428051
## 
##                 property_type
## Y                       Barn   Camper/RV Casa particular Entire condo
##   poor_reviews     0.000000000 0.000000000     0.000000000  0.016129032
##   moderate_reviews 0.000000000 0.000000000     0.001788909  0.021466905
##   good_reviews     0.000000000 0.000000000     0.000000000  0.029143898
##                 property_type
## Y                 Entire guest suite Entire guesthouse Entire home Entire loft
##   poor_reviews           0.000000000       0.000000000 0.014336918 0.000000000
##   moderate_reviews       0.001788909       0.000000000 0.008944544 0.025044723
##   good_reviews           0.001821494       0.000000000 0.001821494 0.020036430
##                 property_type
## Y                 Entire rental unit Entire serviced apartment
##   poor_reviews           0.684587814               0.064516129
##   moderate_reviews       0.633273703               0.069767442
##   good_reviews           0.624772313               0.034608379
##                 property_type
## Y                 Entire townhouse Entire villa Private room
##   poor_reviews         0.005376344  0.000000000  0.001792115
##   moderate_reviews     0.001788909  0.000000000  0.005366726
##   good_reviews         0.000000000  0.000000000  0.001821494
##                 property_type
## Y                 Private room in bed and breakfast
##   poor_reviews                          0.003584229
##   moderate_reviews                      0.001788909
##   good_reviews                          0.003642987
##                 property_type
## Y                 Private room in casa particular Private room in chalet
##   poor_reviews                        0.000000000            0.003584229
##   moderate_reviews                    0.001788909            0.000000000
##   good_reviews                        0.001821494            0.000000000
##                 property_type
## Y                 Private room in condo Private room in guesthouse
##   poor_reviews              0.012544803                0.000000000
##   moderate_reviews          0.014311270                0.001788909
##   good_reviews              0.020036430                0.000000000
##                 property_type
## Y                 Private room in home Private room in hut
##   poor_reviews             0.014336918         0.000000000
##   moderate_reviews         0.016100179         0.000000000
##   good_reviews             0.007285974         0.000000000
##                 property_type
## Y                 Private room in loft Private room in rental unit
```

```
##   poor_reviews              0.001792115                      0.143369176
##   moderate_reviews          0.001788909                      0.146690519
##   good_reviews              0.003642987                      0.231329690
##                   property_type
## Y                  Private room in serviced apartment Private room in townhouse
##   poor_reviews                                0.000000000               0.001792115
##   moderate_reviews                            0.003577818               0.001788909
##   good_reviews                                0.001821494               0.003642987
##                   property_type
## Y                  Private room in villa Room in bed and breakfast
##   poor_reviews               0.000000000               0.000000000
##   moderate_reviews           0.012522361               0.000000000
##   good_reviews               0.003642987               0.000000000
##                   property_type
## Y                  Room in boutique hotel Room in hotel
##   poor_reviews                0.003584229   0.014336918
##   moderate_reviews            0.001788909   0.019677996
##   good_reviews                0.000000000   0.003642987
##                   property_type
## Y                  Room in serviced apartment Shared room in home
##   poor_reviews                    0.000000000         0.000000000
##   moderate_reviews                0.000000000         0.000000000
##   good_reviews                    0.001821494         0.000000000
##                   property_type
## Y                  Shared room in hostel Shared room in hotel
##   poor_reviews               0.001792115          0.003584229
##   moderate_reviews           0.003577818          0.000000000
##   good_reviews               0.000000000          0.000000000
##                   property_type
## Y                  Shared room in rental unit   Tiny home
##   poor_reviews                    0.008960573 0.000000000
##   moderate_reviews                0.001788909 0.001788909
##   good_reviews                    0.000000000 0.003642987
##
##                   number_reviews_group
## Y                          Low     Medium       High
##   poor_reviews     0.82437276 0.12186380 0.05376344
##   moderate_reviews 0.00000000 0.37030411 0.62969589
##   good_reviews     0.20400729 0.49362477 0.30236794
```

Confusion matrix

```
#training
confusionMatrix(predict(nb_model, newdata=training_set), training_set$review_rating)
```

```
## Confusion Matrix and Statistics
## 
##                   Reference
## Prediction         poor_reviews moderate_reviews good_reviews
##   poor_reviews              439                7          106
##   moderate_reviews           62              375          165
##   good_reviews               57              177          278
## 
## Overall Statistics
## 
##                Accuracy : 0.6555          
##                  95% CI : (0.6321, 0.6783)
##     No Information Rate : 0.3355          
##     P-Value [Acc > NIR] : < 0.00000000000000022
## 
##                   Kappa : 0.4831          
## 
##  Mcnemar's Test P-Value : 0.0000000000009652
## 
## Statistics by Class:
## 
##                      Class: poor_reviews Class: moderate_reviews
## Sensitivity                       0.7867                  0.6708
## Specificity                       0.8980                  0.7949
## Pos Pred Value                    0.7953                  0.6229
## Neg Pred Value                    0.8932                  0.8271
## Prevalence                        0.3349                  0.3355
## Detection Rate                    0.2635                  0.2251
## Detection Prevalence              0.3313                  0.3613
## Balanced Accuracy                 0.8424                  0.7329
##                      Class: good_reviews
## Sensitivity                       0.5064
## Specificity                       0.7905
## Pos Pred Value                    0.5430
## Neg Pred Value                    0.7652
## Prevalence                        0.3295
## Detection Rate                    0.1669
## Detection Prevalence              0.3073
## Balanced Accuracy                 0.6484
```

```
#validation
confusionMatrix(predict(nb_model, newdata=validation_set), validation_set$review_rating)
```

```
## Confusion Matrix and Statistics
##
##                   Reference
## Prediction         poor_reviews moderate_reviews good_reviews
##   poor_reviews              286                7           85
##   moderate_reviews           53              245           98
##   good_reviews               33              120          183
##
## Overall Statistics
##
##                Accuracy : 0.6432
##                  95% CI : (0.6143, 0.6715)
##     No Information Rate : 0.3351
##     P-Value [Acc > NIR] : < 0.00000000000000022
##
##                   Kappa : 0.4647
##
##  Mcnemar's Test P-Value : 0.0000000000004823
##
## Statistics by Class:
##
##                      Class: poor_reviews Class: moderate_reviews
## Sensitivity                       0.7688                  0.6586
## Specificity                       0.8753                  0.7954
## Pos Pred Value                    0.7566                  0.6187
## Neg Pred Value                    0.8825                  0.8221
## Prevalence                        0.3351                  0.3351
## Detection Rate                    0.2577                  0.2207
## Detection Prevalence              0.3405                  0.3568
## Balanced Accuracy                 0.8221                  0.7270
##                      Class: good_reviews
## Sensitivity                       0.5000
## Specificity                       0.7944
## Pos Pred Value                    0.5446
## Neg Pred Value                    0.7636
## Prevalence                        0.3297
## Detection Rate                    0.1649
## Detection Prevalence              0.3027
## Balanced Accuracy                 0.6472
```

Prediction

```
fictional_rental <- data.frame(
  neighbourhood_group_cleansed= "Kreis 1",
  neighbourhood_cleansed = "City",
  room_type = "Entire home/apt",
  beds = factor("3", levels= levels(zurich_naive$beds)),
  host_response_time = factor("within a few hours", levels = levels(zurich_naive$host_response_t
ime)),
  host_is_superhost = factor("TRUE", levels = c("FALSE", "TRUE")),
  host_has_profile_pic = TRUE,
  property_type = "Entire rental unit",
  review_rating = factor("Moderate_reviews", levels = c("poor_reviews", "Moderate_reviews", "goo
d_reviews")),
  number_reviews_group = factor("High", levels = c("Low", "Medium", "High"))
)

predicted_bin <- predict(nb_model, newdata = fictional_rental, type = "class")
print(predicted_bin)
```

```
## [1] moderate_reviews
## Levels: poor_reviews moderate_reviews good_reviews
```

The objective of this project is to utilize a Naive Bayes classification model to predict guest satisfaction levels for Airbnb rentals in Zurich, focusing on how much value guests perceive from their stay. This approach aimed to categorize their experiences into three distinct levels of satisfaction: poor, moderate, and good reviews.

During the data preparation phase, we strategically selected features that could significantly impact a guest's experience, such as 'host_is_superhost' and 'host_response_time', while excluding less impactful variables like URLs and geolocation data. This careful selection helped streamline our model, focusing on variables most likely to affect guest satisfaction. The Naive Bayes classifier was trained using these chosen features, with 'review_scores_value' being divided into three balanced categories. This categorization facilitated a more effective learning process for the model, enabling it to distinguish between different levels of guest reviews more accurately.

We evaluated the model using key performance metrics such as sensitivity, specificity, and the positive predictive value for each review category. The final statistics indicated an overall accuracy of 65.55% in the training phase and 64.32% in validation, demonstrating the model's capability to consistently predict guest satisfaction across different datasets. The Kappa statistic of 0.4831 further validated the model's effectiveness beyond chance, highlighting its reliability in classifying reviews accurately.

To illustrate the model's practical utility, we crafted a fictional rental scenario and successfully predicted it as 'moderate_reviews'. This demonstration not only confirmed the model's operational effectiveness but also its potential for real-world application.

In conclusion, while the Naive Bayes model proved to be a valuable tool for gauging and predicting guest satisfaction, there is potential for further enhancement. Future improvements could include more sophisticated feature engineering and the integration of additional data sources to enrich the model's understanding and predictive power, thereby refining its accuracy and broadening its applicability in real-world scenarios.